20th Edition

HARRISON'S™

PRINCIPLES OF INTERNAL MEDICINE

JAMESON
FAUCI
KASPER
HAUSER
LONGO
LOSCALZO

Mc Graw Hill Education

Chapter 1: The Practice of Medicine

The Editors

FIGURE 1-1

ENDURING VALUES OF THE MEDICAL PROFESSION

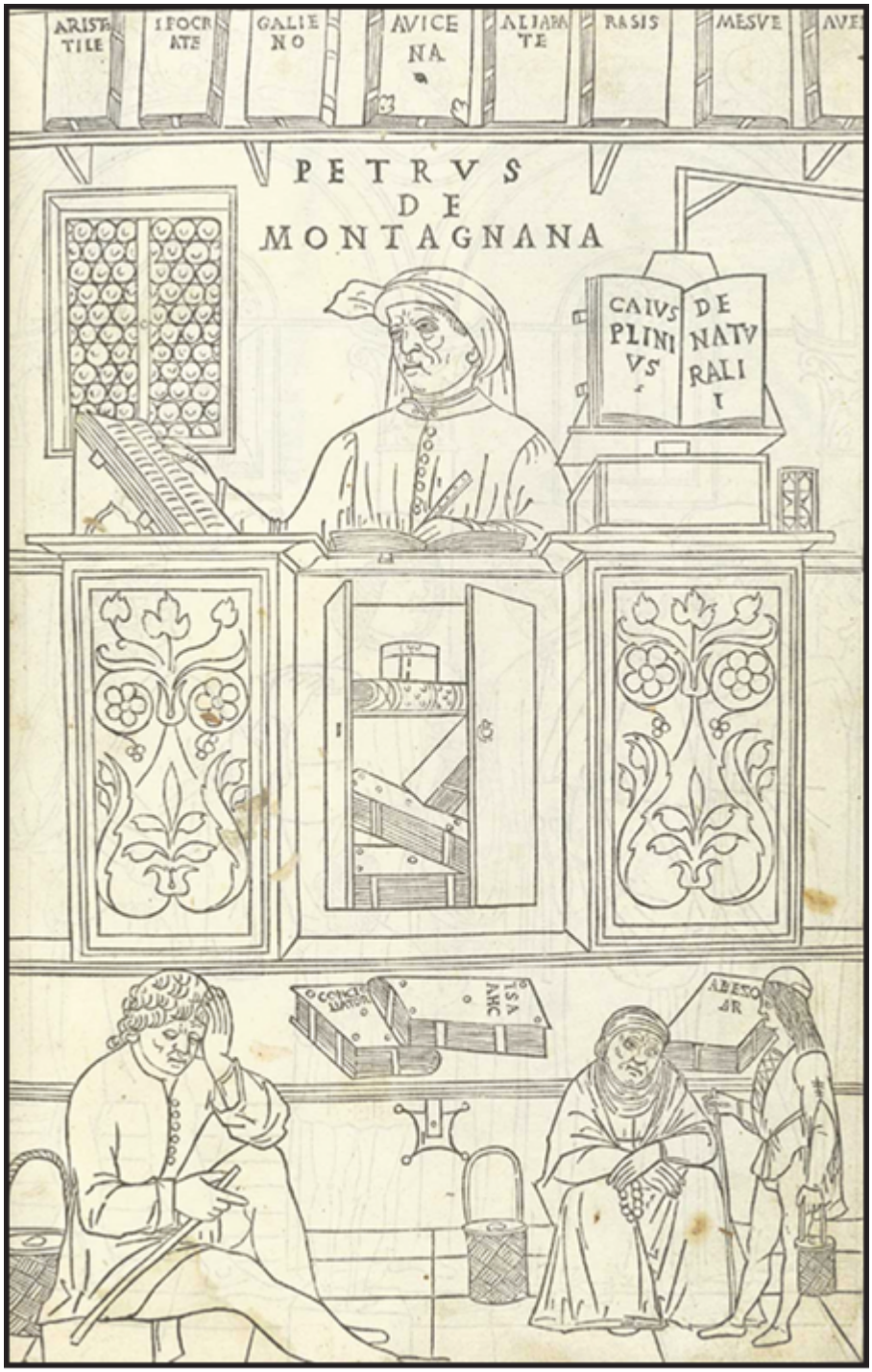
> *No greater opportunity, responsibility, or obligation can fall to the lot of a human being than to become a physician. In the care of the suffering, [the physician] needs technical skill, scientific knowledge, and human understanding.… Tact, sympathy, and understanding are expected of the physician, for the patient is no mere collection of symptoms, signs, disordered functions, damaged organs, and disturbed emotions. [The patient] is human, fearful, and hopeful, seeking relief, help, and reassurance.*
>
> —*Harrison's Principles of Internal Medicine,* 1950

The practice of medicine has changed in significant ways since the first edition of this book appeared in 1950. The advent of molecular genetics, sophisticated new imaging techniques, robotics, and advances in bioinformatics and information technology have contributed to an explosion of scientific information that has changed fundamentally the way physicians define, diagnose, treat, and attempt to prevent disease. This growth of scientific knowledge is ongoing and accelerating.

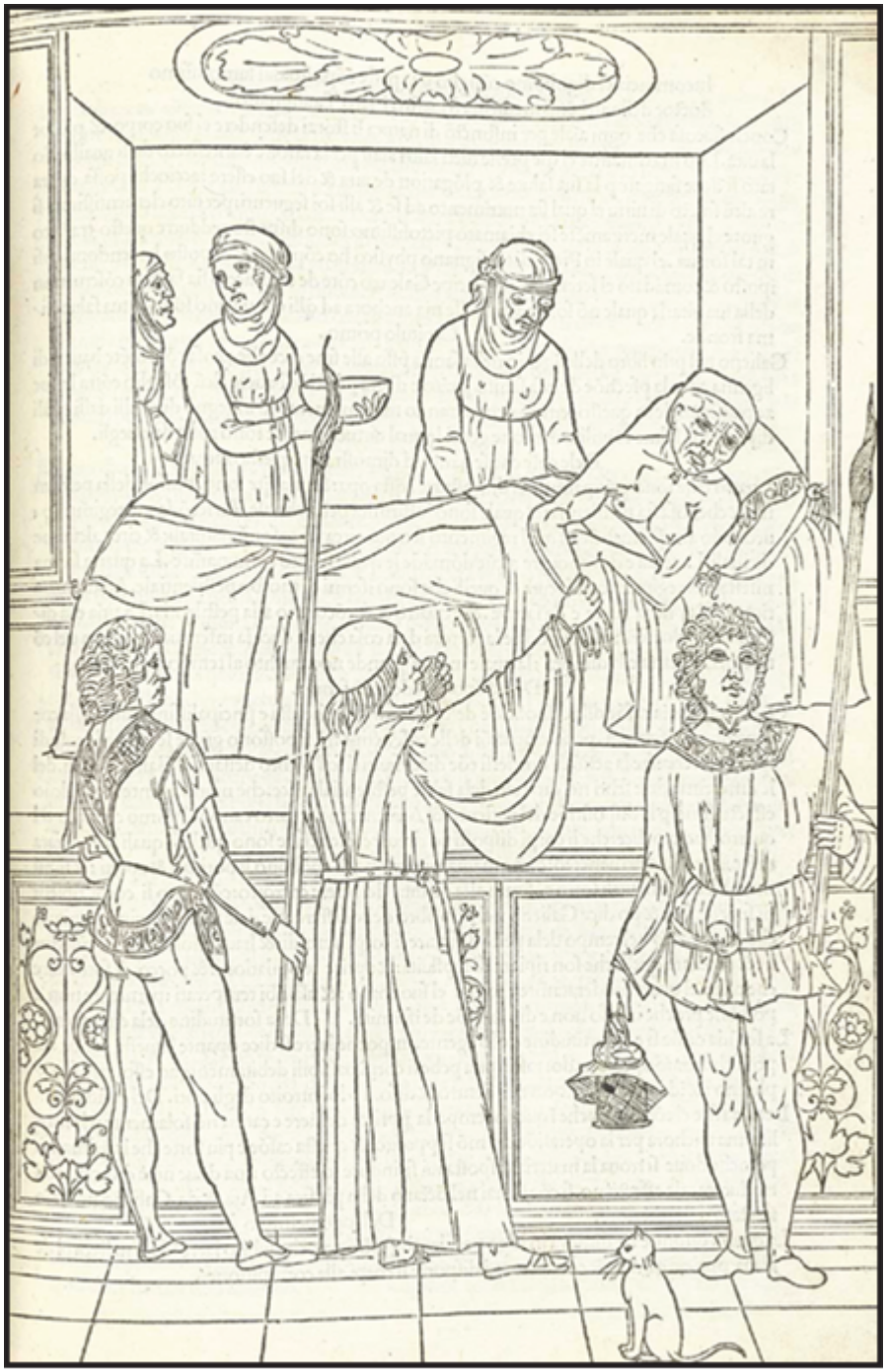
The widespread use of electronic medical records and the Internet have altered the way physicians access and exchange information as a routine part of medical practice **(Fig. 1-1)**. As today's physicians strive to integrate copious amounts of scientific knowledge into everyday practice, it is critically important to remember two things: first, the ultimate goal of medicine is to prevent disease and, when it occurs, to diagnose it early and provide effective treatment; and second, despite nearly 70 years of scientific advances since the first edition of this text, a trusting relationship between physician and patient still lies at the heart of successful patient care.

FIGURE 1-1

Woodcuts from Johannes de Ketham's *Fasciculus Medicinae*, the first illustrated medical text ever printed, show methods of information access and exchange in medical practice during the early Renaissance. Initially published in 1491 for use by medical students and practitioners, *Fasciculus Medicinae* appeared in six editions over the next 25 years. *Left:* Petrus de Montagnana, a well-known physician and teacher at the University of Padua and author of an anthology of instructive case studies, consults medical texts dating from antiquity up to the early Renaissance. *Right:* A patient with plague is attended by a physician and his attendants. *(Courtesy, U.S. National Library of Medicine.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE SCIENCE AND ART OF MEDICINE

Deductive reasoning and applied technology form the foundation for the solution to many clinical problems. Spectacular advances in biochemistry, cell biology, and genomics, coupled with newly developed imaging techniques, allow access to the innermost parts of the cell and provide a window into the most remote recesses of the body. Revelations about the nature of genes and single cells have opened a portal for formulating a new molecular basis for the physiology of systems. Increasingly, physicians are learning how subtle changes in many different genes can affect the function of cells and organisms. Researchers are

deciphering the complex mechanisms by which genes are regulated. Clinicians have developed a new appreciation of the role of stem cells in normal tissue function, in the development of cancer and other disorders, and in the treatment of certain diseases. Entirely new areas of research, including studies of chronobiology, the human microbiome, and epigenetics, have become important for understanding both health and disease. Information technology enables the interrogation of medical records from millions of individuals, yielding new insights into the etiology, characteristics, and stratification of many diseases. The knowledge gleaned from the ***science of medicine*** continues to enhance the understanding by physicians of complex pathologic processes and to provide new approaches to disease prevention, diagnosis, and treatment. Yet skill in the most sophisticated applications of laboratory technology and in the use of the latest therapeutic modality alone does not make a good physician.

When a patient poses challenging clinical problems, an effective physician must be able to identify the crucial elements in a complex history and physical examination; order the appropriate laboratory, imaging, and diagnostic tests; and extract the key results from densely populated computer screens to determine whether to treat or to "watch." As the number of tests increases, so does the likelihood that some incidental finding, completely unrelated to the clinical problem at hand, will be uncovered. Deciding whether a clinical clue is worth pursuing or should be dismissed as a "red herring" and weighing whether a proposed test, preventive measure, or treatment entails a greater risk than the disease itself are essential judgments that a skilled clinician must make many times each day. This combination of medical knowledge, intuition, experience, and judgment defines the ***art of medicine***, which is as necessary to the practice of medicine as is a sound scientific base.

CLINICAL SKILLS

History-Taking

The recorded history of an illness should include all the facts of medical significance in the life of the patient. Recent events should be given the most attention. Patients should, at some early point, have the opportunity to tell their own story of the illness without frequent interruption and, when appropriate, should receive expressions of interest, encouragement, and empathy from the physician. Any event related by a patient, however trivial or seemingly irrelevant, may provide the key to solving the medical problem. A methodical review of systems is important to elicit features of an underlying disease that might not be mentioned in the patient's narrative. In general, patients who feel comfortable with the physician will offer more complete information; thus, putting the patient at ease contributes substantially to obtaining an adequate history.

An informative history is more than an orderly listing of symptoms. By listening to patients and noting the way in which they describe their symptoms, physicians can gain valuable insight. Inflections of voice, facial expression, gestures, and attitude (i.e., "body language") may offer important clues to patients' perception of their symptoms. Because patients vary considerably in their medical sophistication and ability to recall facts, the reported medical history should be corroborated whenever possible. The social history also can provide important insights into the types of diseases that should be considered and can identify practical considerations for subsequent management. The family history not only identifies rare Mendelian disorders

but often reveals risk factors for common disorders, such as coronary heart disease, hypertension, autoimmunity, and asthma. A thorough family history may require input from multiple relatives to ensure completeness and accuracy. An experienced clinician can usually formulate a relevant differential diagnosis from the history alone, using the physical examination and diagnostic tests to narrow the list or reveal unexpected findings that lead to more focused inquiry.

The very act of eliciting the history provides the physician with an opportunity to establish or enhance a unique bond that forms the basis for a good patient–physician relationship. This process helps the physician develop an appreciation of the patient's view of the illness, the patient's expectations of the physician and the health care system, and the financial and social implications of the illness for the patient. Although current health care settings may impose time constraints on patient visits, it is important not to rush the encounter. A hurried approach may lead patients to believe that what they are relating is not of importance to the physician, and thus they may withhold relevant information. The confidentiality of the patient–physician relationship cannot be overemphasized.

Physical Examination

The purpose of the physical examination is to identify physical signs of disease. The significance of these objective indications of disease is enhanced when they confirm a functional or structural change already suggested by the patient's history. At times, however, physical signs may be the only evidence of disease and may not have been suggested by the history.

The physical examination should be methodical and thorough, with consideration given to the patient's comfort and modesty. Although attention is often directed by the history to the diseased organ or part of the body, the examination of a new patient must extend from head to toe in an objective search for abnormalities. The results of the examination, like the details of the history, should be recorded at the time they are elicited—not hours later, when they are subject to the distortions of memory. Physical examination skills should be learned under direct observation of experienced clinicians. Even highly experienced clinicians can benefit from ongoing coaching and feedback. Simulation laboratories and standardized patients play an increasingly important role in the development of clinical skills. Although the skills of physical diagnosis are acquired with experience, it is not merely technique that determines success in identifying signs of disease. The detection of a few scattered petechiae, a faint diastolic murmur, or a small mass in the abdomen is not a question of keener eyes and ears or more sensitive fingers, but of a mind alert to those findings. Because physical findings can change with time, the physical examination should be repeated as frequently as the clinical situation warrants.

Given the many highly sensitive diagnostic tests now available (particularly imaging techniques), it may be tempting to place less emphasis on the physical examination. Indeed, many patients are seen by consultants after a series of diagnostic tests have been performed and the results are known. This fact should not deter the physician from performing a thorough physical examination since important clinical findings may have escaped detection. The act of examining (touching) the patient also offers an opportunity for communication and may have reassuring effects that foster the patient–physician relationship.

Diagnostic Studies

Physicians rely increasingly on a wide array of laboratory and imaging tests to make diagnoses and ultimately to solve clinical problems. However, accumulated results do not relieve the physician from the responsibility of carefully observing and examining the patient. It is also essential to appreciate the limitations of diagnostic tests. By virtue of their apparent precision, these tests often gain an aura of certainty regardless of the fallibility of the tests themselves, the instruments used in the tests, and the individuals performing or interpreting the tests. Physicians must weigh the expense involved in laboratory procedures against the value of the information these procedures are likely to provide.

Single laboratory tests are rarely ordered. Instead, physicians generally request “batteries” of multiple tests, which often prove useful and can be performed with a single specimen at relatively low cost. For example, abnormalities of hepatic function may provide the clue to nonspecific symptoms such as generalized weakness and increased fatigability, suggesting a diagnosis of chronic liver disease. Sometimes a single abnormality, such as an elevated serum calcium level, points to a particular disease, such as hyperparathyroidism or an underlying malignancy.

The thoughtful use of screening tests (e.g., measurement of low-density lipoprotein cholesterol) may allow early intervention to prevent disease **(Chap. 4)**. Screening tests are most informative when they are directed toward common diseases and when their results indicate whether other useful—but often costly—tests or interventions are needed. On the one hand, biochemical measurements, together with simple laboratory determinations such as routine serum chemistries, blood counts, and urinalysis, often provide a major clue to the presence of a pathologic process. On the other hand, the physician must learn to evaluate occasional screening-test abnormalities that do not necessarily connote significant disease. An in-depth workup after the report of an isolated laboratory abnormality in a person who is otherwise well is often wasteful and unproductive. Because so many tests are performed routinely for screening purposes, it is not unusual for one or two values to be slightly abnormal. Nevertheless, even if there is no reason to suspect an underlying illness, tests yielding abnormal results ordinarily are repeated to rule out laboratory error. If an abnormality is confirmed, it is important to consider its potential significance in the context of the patient’s condition and other test results.

There is almost continual development of technically improved imaging studies with greater sensitivity and specificity. These tests provide remarkably detailed anatomic information that can be pivotal in informing medical decision-making. Ultrasonography, CT, MRI, a variety of isotopic scans, and positron emission tomography (PET) have supplanted older, more invasive approaches and opened new diagnostic vistas. In light of their capabilities and the rapidity with which they can lead to a diagnosis, it is tempting to order a battery of imaging studies. All physicians have had experiences in which imaging studies revealed findings that led to an unexpected diagnosis. Nonetheless, patients must endure each of these tests, and the added cost of unnecessary testing is substantial. Furthermore, investigation of an unexpected abnormal finding may be associated with risk and/or expense and may lead to the diagnosis of an irrelevant or incidental problem. A skilled physician must learn to use these powerful diagnostic tools judiciously, always considering whether the results will alter management and benefit the patient.

MANAGEMENT OF PATIENT CARE

Team-Based Care

Medical practice has long involved teams, particularly physicians working with nurses. Advances in medicine have increased our ability to manage very complex clinical situations (e.g., intensive care units [ICUs], bone marrow transplantation) and have shifted the burden of disease toward chronic illnesses. Because an individual patient may have multiple chronic diseases, he or she may be cared for by different specialists as well as a primary care physician. In the inpatient setting, care may involve multiple consultants along with the primary admitting physician. Communication through the medical record is necessary but not sufficient, particularly when patients have complex medical problems or when difficult decisions need to be made about the optimal management plan. Physicians should willingly meet face-to-face or by phone to ensure clear communication and thoughtful planning. It is important to note that patients often receive or perceive different messages from various care providers; attempts should be made to provide consistency among these messages to the patient. Management plans and treatment options should be outlined succinctly and clearly for the patient.

Another dimension of team-based care involves allied health professions. It is not unusual for a hospitalized patient to encounter physical therapists, pharmacists, respiratory therapists, radiology technicians, social workers, dieticians, and transport personnel (among others) in addition to physicians and nurses. Each of these individuals contributes to clinical care as well as to the patient's experience with the health care system. In the outpatient setting, disease screening and chronic disease management are often carried out by nurses, physician assistants, or other allied health professionals.

The growth of team-based care has important implications for medical culture, student and resident training, and the organization of health care systems. Despite diversity in training, skills, and responsibilities among health care professionals, common values need to be espoused and reinforced. Many medical schools have incorporated interprofessional teamwork into their curricula. Effective communication is inevitably the most challenging aspect of implementing team-based care. While communication can be aided by electronic devices, including medical records, apps, or text messages, it is vitally important to balance efficiency with taking the necessary time to speak directly with colleagues.

The Dichotomy of Inpatient and Outpatient Internal Medicine

The hospital environment has experienced sweeping changes over the last few decades. Emergency departments and critical care units have evolved to manage critically ill patients, allowing them to survive formerly fatal conditions. In parallel, there is increasing pressure to reduce the length of stay in the hospital and to manage complex disorders in the outpatient setting. This transition has been driven not only by efforts to reduce costs but also by the availability of new outpatient technologies, such as imaging and percutaneous infusion catheters for long-term antibiotics or nutrition, minimally invasive surgical procedures, and evidence that outcomes often are improved by reducing inpatient hospitalization.

In addition to traditional medical beds, hospitals now encompass multiple distinct levels of care, such as the emergency department, procedure rooms, overnight observation units, critical care units, and palliative care units. A consequence of this differentiation has been the emergence of new specialties (e.g., emergency medicine and end-of-life care) and the provision of in-hospital care by hospitalists and intensivists. Most *hospitalists* are board-certified internists who bear primary responsibility for the care of hospitalized patients and whose work is limited entirely to the hospital setting. The shortened length of hospital stay means that most patients receive only acute care while hospitalized; the increased complexities of inpatient medicine make the presence of an internist with specific training, skills, and experience in the hospital environment extremely beneficial. *Intensivists* are board-certified physicians who are further certified in critical care medicine and who direct and provide care for very ill patients in critical care units. Clearly, an important challenge in internal medicine today is to ensure the continuity of communication and information flow between a patient's primary care physician and those who are in charge of the patient's hospital care. Maintaining these channels of communication is frequently complicated by patient "handoffs"—i.e., transitions from the outpatient to the inpatient environment, from the critical care unit to a general medicine floor, from a medical to a surgical service and vice versa, and from the hospital to the outpatient environment.

The involvement of many care providers in conjunction with these transitions can threaten the traditional one-to-one relationship between patient and primary care physician. Of course, patients can benefit greatly from effective collaboration among a number of health care professionals; however, *it is the duty of the patient's principal or primary physician to provide cohesive guidance through an illness*. To meet this challenge, primary care physicians must be familiar with the techniques, skills, and objectives of specialist physicians and allied health professionals who care for their patients in the hospital. In addition, primary care physicians must ensure that their patients benefit from scientific advances and the expertise of specialists, both in and out of the hospital. Primary care physicians should explain the role of these specialists to reassure patients that they are in the hands of physicians best trained to manage an acute illness. However, the primary care physician should assure patients and their families that decisions are being made in consultation with these specialists. The evolving concept of the "medical home" incorporates team-based primary care with subspecialty care in a cohesive environment that ensures smooth transitions of care.

Mitigating the Stress of Acute Illness

Few people are prepared for a new diagnosis of cancer or anticipate the occurrence of a myocardial infarction, stroke, or major accident. The care of a frightened or distraught patient is confounded by these understandable responses to life-threatening events. The physician and other health providers can reduce the shock of life-changing events by providing information in a clear, calm, consistent, and reassuring manner. Often, information and reassurance need to be repeated. Caregivers should also recognize that, for outsiders, hospital emergency rooms, operating rooms, ICUs, and general medical floors represent an intimidating environment. Hospitalized patients find themselves surrounded by air jets, buttons, and glaring lights; invaded by tubes and wires; and beset by the numerous members of the health care team—

hospitalists, specialists, nurses, nurses' aides, physicians' assistants, social workers, technologists, physical therapists, medical students, house officers, attending and consulting physicians, and many others. They may be transported to special laboratories and imaging facilities replete with blinking lights, strange sounds, and unfamiliar personnel; they may be left unattended at times; and they may be obligated to share a room with other patients who have their own health problems. It is little wonder that patients may be stressed by this environment. Physicians who appreciate the hospital experience from the patient's perspective and who make an effort to guide the patient through this experience may make a stressful situation more tolerable and enhance the patient's chances for an optimal recovery.

Medical Decision-Making

Medical decision-making is a fundamental responsibility of the physician and occurs at each stage of the diagnostic and therapeutic process. The decision-making process involves the ordering of additional tests, requests for consultations, decisions about treatment, and predictions concerning prognosis. This process requires an in-depth understanding of the pathophysiology and natural history of disease. Formulating a differential diagnosis requires not only a broad knowledge base but also the ability to assess the relative probabilities of various diseases for a given patient. Application of the scientific method, including hypothesis formulation and data collection, is essential to the process of accepting or rejecting a particular diagnosis. Analysis of the differential diagnosis is an iterative process. As new information or test results are acquired, the group of disease processes being considered can be contracted or expanded appropriately. Whenever possible, decisions should be evidence-based, taking advantage of rigorously designed clinical trials or objective comparisons of different diagnostic tests. *Evidence-based medicine* is in sharp contrast to anecdotal experience, which is often biased. Unless attuned to the importance of using larger, objective studies for making decisions, even the most experienced physicians can be influenced to an undue extent by recent encounters with selected patients. Evidence-based medicine has become an increasingly important part of routine medical practice and has led to the publication of many useful practice guidelines.

Despite the importance of evidence-based medicine, much medical decision-making still relies on good clinical judgment, an attribute that is difficult to quantify or even to assess qualitatively. Physicians must use their knowledge and experience as a basis for weighing known factors, along with the inevitable uncertainties, and then making a sound judgment; this synthesis of information is particularly important when a relevant evidence base is not available. Several quantitative tools may be invaluable in synthesizing the available information, including diagnostic tests, Bayes' theorem, and multivariate statistical models. Diagnostic tests serve to reduce uncertainty about an individual's diagnosis or prognosis and help the physician decide how best to manage that individual's condition. The battery of diagnostic tests complements the history and the physical examination. The accuracy of a particular test is ascertained by determining its sensitivity (true-positive rate) and specificity (true-negative rate) as well as the predictive value of a positive and a negative result. **See Chap. 3 for a more thorough discussion of decision-making in clinical medicine.**

Practice Guidelines

Many professional organizations and government agencies have developed formal clinical-practice guidelines to aid physicians and other caregivers in making diagnostic and therapeutic decisions that are evidence-based, cost-effective, and most appropriate to a particular patient and clinical situation. As the evidence base of medicine increases, guidelines can provide a useful framework for managing patients with particular diagnoses or symptoms. Clinical guidelines can protect patients—particularly those with inadequate health care benefits—from receiving substandard care. These guidelines also can protect conscientious caregivers from inappropriate charges of malpractice and society from the excessive costs associated with the overuse of medical resources. There are, however, caveats associated with clinical-practice guidelines since they tend to oversimplify the complexities of medicine. Furthermore, groups with different perspectives may develop divergent recommendations regarding issues as basic as the need for screening of women by mammography or of men by serum prostate-specific antigen (PSA). Finally, guidelines, as the term implies, do not—and cannot be expected to—account for the uniqueness of each individual and his or her illness. The physician's challenge is to integrate into clinical practice the useful recommendations offered by experts without accepting them blindly or being inappropriately constrained by them.

Precision Medicine

The concept of *precision* or *personalized medicine* reflects the growing recognition that diseases once lumped together can be further stratified on the basis of genetic, biomarker, phenotypic, and/or psychosocial characteristics that distinguish a given patient from other patients with similar clinical presentations. Inherent in this concept is the goal of targeting therapies in a more specific way to improve clinical outcomes for the individual patient and minimize unnecessary side effects for those less likely to respond to a particular treatment. In some respects, precision medicine represents the evolution of clinical practice guidelines, which are usually developed for populations of patients or a particular diagnosis (e.g., hypertension, thyroid nodule). As the pathophysiology, prognosis, and treatment responses of subgroups within these diagnoses become better understood, the relevant clinical guidelines incorporate progressively more refined recommendations for individuals within these subgroups. The role of precision medicine is particularly important for cancers in which genetic testing is able to predict responses (or the lack thereof) to targeted therapies (Chap. 69). One can anticipate similar applications of precision medicine in pharmacogenomics, immunologic disorders, and diseases in which biomarkers better predict treatment responses.

Evaluation of Outcomes

Clinicians generally use *objective* and readily measurable parameters to judge the outcome of a therapeutic intervention. These measures may oversimplify the complexity of a clinical condition as patients often present with a major clinical problem in the context of multiple complicating background illnesses. For example, a patient may present with chest pain and cardiac ischemia, but with a background of chronic obstructive pulmonary disease and renal insufficiency. For this reason, outcome measures such as mortality, length of hospital stay, or readmission rates are typically risk-adjusted. An important point to remember is

that patients usually seek medical attention for *subjective* reasons; they wish to obtain relief from pain, to preserve or regain function, and to enjoy life. The components of a patient's health status or quality of life can include bodily comfort, capacity for physical activity, personal and professional function, sexual function, cognitive function, and overall perception of health. Each of these important domains can be assessed through structured interviews or specially designed questionnaires. Such assessments provide useful parameters by which a physician can judge patients' subjective views of their disabilities and responses to treatment, particularly in chronic illness. The practice of medicine requires consideration and integration of both objective and subjective outcomes.

Many health systems use survey and patient feedback data to assess qualitative features such as patient satisfaction, access to care, and communication with nurses and physicians. In the United States, HCAHPS (Hospital Consumer Assessment of Healthcare Providers and Systems) surveys are used by many systems and are publically reported. Social media is also being used to assess feedback in real time as well as to share patient experiences with health care systems.

Errors in the Delivery of Health Care

A series of reports from the Institute of Medicine (now the National Academy of Medicine [NAM]) called for an ambitious agenda to reduce medical error rates and improve patient safety by designing and implementing fundamental changes in health care systems. It is the responsibility of hospitals and health care organizations to develop systems to reduce risk and ensure patient safety. Medication errors can be reduced through the use of ordering systems that rely on electronic processes or, when electronic options are not available, that eliminate misreading of handwriting. Whatever the clinical situation, it is the physician's responsibility to use powerful therapeutic measures wisely, with due regard for their beneficial actions, potential dangers, and cost. Implementation of infection control systems, enforcement of hand-washing protocols, and careful oversight of antibiotic use can minimize the complications of nosocomial infections. Central-line infection rates have been dramatically reduced at many centers by careful adherence of trained personnel to standardized protocols for introducing and maintaining central lines. Rates of surgical infection and wrong-site surgery can likewise be reduced by the use of standardized protocols and checklists. Falls by patients can be minimized by judicious use of sedatives and appropriate assistance with bed-to-chair and bed-to-bathroom transitions. Taken together, these and other measures are saving thousands of lives each year.

Electronic Medical Records

Both the growing reliance on computers and the strength of information technology now play central roles in medicine, including efforts to reduce medical errors. Laboratory data are accessed almost universally through computers. Many medical centers now have electronic medical records (EMRs), computerized order entry, and bar-coded tracking of medications. Some of these systems are interactive, sending reminders or warning of potential medical errors.

EMRs offer rapid access to information that is invaluable in enhancing health care quality and patient safety, including relevant data, historical and clinical information, imaging studies, laboratory results, and medication records. These data can be used to monitor and reduce unnecessary variations in care and to provide real-time information about processes of care and clinical outcomes. Ideally, patient records are easily transferred across the health care system. However, technological limitations and concerns about privacy and cost continue to limit broad-based use of EMRs in many clinical settings.

For all of the advantages of EMRs, they can create distance between the physician and patient if care is not taken to preserve face-to-face contact. EMRs also require training and time for data entry. Many providers spend significant time entering information to generate structured data and to meet billing requirements. They may feel pressured to take short cuts, such as "cutting and pasting" parts of earlier notes into the daily record, thereby increasing the risk of errors. EMRs also structure information in a manner that disrupts the traditional narrative flow across time and among providers. These features, which may be frustrating for some providers, must be weighed against the advantages of ready access to past medical history, imaging, laboratory data, and consultant notes.

It is important to emphasize that information technology is merely a tool and can never replace the clinical decisions that are best made by the physician. Clinical knowledge and an understanding of a patient's needs, supplemented by quantitative tools, still represent the best approach to decision-making in the practice of medicine.

THE PATIENT–PHYSICIAN RELATIONSHIP

> *The significance of the intimate personal relationship between physician and patient cannot be too strongly emphasized, for in an extraordinarily large number of cases both the diagnosis and treatment are directly dependent on it. One of the essential qualities of the clinician is interest in humanity,* ***for the secret of the care of the patient is in caring for the patient****.*
>
> —Francis W. Peabody, October 21, 1925, Lecture at Harvard Medical School

Physicians must never forget that patients are individuals with problems that all too often transcend their physical complaints. They are not "cases" or "admissions" or "diseases." Patients do not fail treatments; treatments fail to benefit patients. This point is particularly important in this era of high technology in clinical medicine. Most patients are anxious and fearful. Physicians should instill confidence and offer reassurance, but they must never come across as arrogant or patronizing. A professional attitude, coupled with warmth and openness, can do much to alleviate anxiety and to encourage patients to share all aspects of their medical history. Empathy and compassion are the essential features of a caring physician. The physician needs to consider the setting in which an illness occurs—in terms not only of patients themselves but also of their familial, social, and cultural backgrounds. The ideal patient–physician relationship is based on thorough knowledge of the patient, mutual trust, and the ability to communicate.

Informed Consent

The fundamental principles of medical ethics require physicians to act in the patient's best interest and to respect the patient's autonomy. These requirements are particularly relevant to the issue of informed consent. Patients are required to sign consent forms for most diagnostic or therapeutic procedures. Many patients possess limited medical knowledge and must rely on their physicians for advice. Communicating in a clear and understandable manner, physicians must fully discuss the alternatives for care and explain the risks, benefits, and likely consequences of each alternative. The physician is responsible for ensuring that the patient thoroughly understands these risks and benefits; encouraging questions is an important part of this process. It may be necessary to go over certain issues with the patient more than once. This is the very definition of *informed consent*. Complete, clear explanation and discussion of the proposed procedures and treatment can greatly mitigate the fear of the unknown that commonly accompanies hospitalization. Often the patient's understanding is enhanced by repeatedly discussing the issues in an unthreatening and supportive way, answering new questions that occur to the patient as they arise. Clear communication can also help alleviate misunderstandings in situations where complications of intervention occur.

Special care should be taken to ensure that a physician seeking a patient's informed consent has no real or apparent conflict of interest.

Approach to Grave Prognoses and Death

No circumstance is more distressing than the diagnosis of an incurable disease, particularly when premature death is inevitable. What should the patient and family be told? What measures should be taken to maintain life? What can be done to optimize quality of life?

Transparency of information, delivered in an appropriate manner, is essential in the face of a terminal illness. Even patients who seem unaware of their medical circumstances, or whose family members have protected them from diagnoses or prognoses, often have keen insights into their condition. They may also have misunderstandings that can lead to additional anxiety. The patient must be given an opportunity to talk with the physician and ask questions. A wise and insightful physician uses such open communication as the basis for assessing what the patient wants to know and when he or she wants to know it. On the basis of the patient's responses, the physician can assess the right tempo for sharing information. Ultimately, the patient must understand the expected course of the disease so that appropriate plans and preparations can be made. The patient should participate in decision-making with an understanding of the goal of treatment (palliation) and its likely effects. The patient's religious beliefs should be taken into consideration. Some patients may find it easier to share their feelings about death with their physician, nurses, or members of the clergy than with family members or friends.

The physician should provide or arrange for emotional, physical, and spiritual support and must be compassionate, unhurried, and open. In many instances, there is much to be gained by the laying on of hands. Pain should be controlled adequately, human dignity maintained, and isolation from family and close friends avoided. These aspects of care tend to be overlooked in hospitals, where the intrusion of life-sustaining equipment can detract from attention to the whole person and encourage concentration instead on the life-threatening disease, against which the battle ultimately will be lost in any case. In the face of

terminal illness, the goal of medicine must shift from *cure* to *care* in the broadest sense of the term. *Primum succurrere*, first hasten to help, is a guiding principle. In offering care to a dying patient, a physician should be prepared to provide information to family members and deal with their grief and sometimes their feelings of guilt or even anger. It is important for the physician to assure the family that everything reasonable is being done. A substantial challenge in these discussions is that the physician often does not know how to gauge the prognosis. In addition, various members of the health care team may offer different opinions. Good communication among providers is essential so that consistent information is provided to patients. This is especially important when the best path forward is uncertain. Advice from experts in palliative and terminal care should be sought whenever appropriate to ensure that clinicians are not providing patients with unrealistic expectations. **For a more complete discussion of end-of-life care, see Chap. 9.**

Maintaining Humanism and Professionalism

Many trends in the delivery of health care tend to make medical care impersonal. These trends, some of which have been mentioned already, include (1) vigorous efforts to reduce the escalating costs of health care; (2) the growing number of managed-care programs, which are intended to reduce costs but in which the patient may have little choice in selecting a physician; (3) increasing reliance on technological advances and computerization; and (4) the need for numerous physicians and other health professionals to be involved in the care of most patients who are seriously ill.

In light of these changes in the medical care system, it is a major challenge for physicians to maintain the *humane* aspects of medical care. The American Board of Internal Medicine, working together with the American College of Physicians–American Society of Internal Medicine and the European Federation of Internal Medicine, has published a *Charter on Medical Professionalism* that underscores three main principles in physicians' contract with society: (1) the primacy of patient welfare, (2) patient autonomy, and (3) social justice. While medical schools appropriately place substantial emphasis on professionalism, a physician's personal attributes, including integrity, respect, and compassion, also are extremely important. In the United States, the Gold Humanism Society recognizes individuals who are exemplars of humanistic patient care and serve as role models for medical education and training.

Availability to the patient, expression of sincere concern, willingness to take the time to explain all aspects of the illness, and a nonjudgmental attitude when dealing with patients whose cultures, lifestyles, attitudes, and values differ from those of the physician are just a few of the characteristics of a humane physician. Every physician will, at times, be challenged by patients who evoke strongly negative or positive emotional responses. Physicians should be alert to their own reactions to such situations and should consciously monitor and control their behavior so that the patient's best interest remains the principal motivation for their actions at all times.

Another important aspect of patient care involves an appreciation of the patient's "quality of life," a subjective assessment of what each patient values most. This assessment requires detailed, sometimes intimate knowledge of the patient, which usually can be obtained only through deliberate, unhurried, and

often repeated conversations. Time pressures will always threaten these interactions, but they should not diminish the importance of understanding and seeking to fulfill the priorities of the patient.

EXPANDING FRONTIERS IN MEDICAL PRACTICE

The Era of "Omics"

In the spring of 2003, announcement of the complete sequencing of the human genome officially ushered in the genomic era. However, even before that landmark accomplishment, the practice of medicine had been evolving as a result of insights into both the human genome and the genomes of a wide variety of microbes. The clinical implications of these insights are illustrated by the complete genome sequencing of H1N1 influenza virus in 2009 and the rapid identification of H1N1 influenza as a potentially fatal pandemic illness, leading to the swift development and dissemination of an effective protective vaccine. Today, gene expression profiles are being used to guide therapy and inform prognosis for a number of diseases, and genotyping is providing a new means to assess the risk of certain diseases as well as variations in response to a number of drugs. Despite these advances, the use of complex genomics in the diagnosis, prevention, and treatment of disease is still in its early stages. The task of physicians is complicated by the fact that phenotypes generally are determined not by genes alone but by the interplay of genetic and environmental factors.

Rapid progress is also being made in other areas of molecular medicine. *Epigenetics* is the study of alterations in chromatin and histone proteins and methylation of DNA sequences that influence gene expression (**Chap. 471**). Every cell of the body has identical DNA sequences; the diverse phenotypes a person's cells manifest are the result of epigenetic regulation of gene expression. Epigenetic alterations are associated with a number of cancers and other diseases. *Proteomics*, the study of the entire library of proteins made in a cell or organ and the complex relationship of these proteins to disease, is enhancing the repertoire of the 23,000 genes in the human genome through alternate splicing, posttranslational processing, and posttranslational modifications that often have unique functional consequences. The presence or absence of particular proteins in the circulation or in cells is being explored for diagnostic and disease-screening applications. *Microbiomics* is the study of the resident microbes in humans and other mammals, which together compose the microbiome. The human haploid genome has ~23,000 genes, whereas the microbes residing on and in the human body encompass more than 3–4 million genes; these resident microbes are likely to be of great significance with regard to health status. Ongoing research is demonstrating that the microbes inhabiting human mucosal and skin surfaces play a critical role in maturation of the immune system, in metabolic balance, and in disease susceptibility. A variety of environmental factors, including the use and overuse of antibiotics, have been tied experimentally to substantial increases in disorders such as obesity, metabolic syndrome, atherosclerosis, and immune-mediated diseases in both adults and children. *Metagenomics*, of which microbiomics is a part, is the genomic study of environmental species that have the potential to influence human biology directly or indirectly. An example is the study of exposures to microorganisms in farm environments that may be responsible for the lower incidence of asthma among children raised on farms. *Metabolomics* is the study of the range of metabolites in cells or

organs and the ways they are altered in disease states. The aging process itself may leave telltale metabolic footprints that allow the prediction (and possibly the prevention) of organ dysfunction and disease. It seems likely that disease-associated patterns will be found in lipids, carbohydrates, membranes, mitochondria, and other vital components of cells and tissues. *Exposomics* is the study of the exposome—i.e., the environmental exposures such as smoking, sunlight, diet, exercise, education, and violence that together have an enormous impact on health. All of this new information represents a challenge to the traditional reductionist approach to medical thinking. The variability of results in different patients, together with the large number of variables that can be assessed, creates challenges in identifying preclinical disease and defining disease states unequivocally. Accordingly, the tools of *systems biology* and *network medicine* are being applied to the enormous body of information now obtainable for every patient and may eventually provide new approaches to classifying disease. **For a more complete discussion of a complex systems approach to human disease, see Chap. 476.**

The rapidity of these advances may seem overwhelming to practicing physicians. However, physicians have an important role to play in ensuring that these powerful technologies and sources of new information are applied judiciously to patient care. Since "omics" are evolving so rapidly, physicians and other health care professionals must engage in continuous learning so that they can apply this new knowledge to the benefit of their patients' health and well-being. Genetic testing requires wise counsel based on an understanding of the value and limitations of the tests as well as the implications of their results for specific individuals. **For a more complete discussion of genetic testing, see Chap. 457.**

The Globalization of Medicine

Physicians should be cognizant of diseases and health care services beyond local boundaries. Global travel has implications for disease spread, and it is not uncommon for diseases endemic to certain regions to be seen in other regions after a patient has traveled to and returned from those regions. The outbreak of Zika virus infections in the Americas is a cogent example of this phenomenon. In addition, factors such as wars, the migration of refugees, and climate change are contributing to changing disease profiles worldwide. Patients have broader access to unique expertise or clinical trials at distant medical centers, and the cost of travel may be offset by the quality of care at those distant locations. As much as any other factor influencing global aspects of medicine, the Internet has transformed the transfer of medical information throughout the world. This change has been accompanied by the transfer of technological skills through telemedicine and international consultation—for example, interpretation of radiologic images and pathologic specimens. **For a complete discussion of global issues, see Chap. 460.**

Medicine on the Internet

On the whole, the Internet has had a positive effect on the practice of medicine; through personal computers, a wide range of information is available to physicians and patients almost instantaneously at any time and from anywhere in the world. This medium holds enormous potential for the delivery of current information, practice guidelines, state-of-the-art conferences, journal content, textbooks (including this text), and direct communications with other physicians and specialists, expanding the depth and breadth of

information available to the physician regarding the diagnosis and care of patients. Medical journals are now accessible online, providing rapid sources of new information. By bringing them into direct and timely contact with the latest developments in medical care, this medium also serves to lessen the information gap that has hampered physicians and health care providers in remote areas.

Patients, too, are turning to the Internet in increasing numbers to acquire information about their illnesses and therapies and to join Internet-based support groups. Patients often arrive at a clinic visit with sophisticated information about their illnesses. In this regard, physicians are challenged in a positive way to keep abreast of the latest relevant information while serving as an "editor" as patients navigate this seemingly endless source of information, the accuracy and validity of which are not uniform.

A critically important caveat is that virtually anything can be published on the Internet, with easy circumvention of the peer-review process that is an essential feature of academic publications. Both physicians and patients who search the Internet for medical information must be aware of this danger. Notwithstanding this limitation, appropriate use of the Internet is revolutionizing information access for physicians and patients and in this regard represents a remarkable resource that was not available to practitioners a generation ago.

Public Expectations and Accountability

The general public's level of knowledge and sophistication regarding health issues has grown rapidly over the last few decades. As a result, expectations of the health care system in general and of physicians in particular have risen. Physicians are expected to master rapidly advancing fields (the *science* of medicine) while considering their patients' unique needs (the *art* of medicine). Thus, physicians are held accountable not only for the technical aspects of the care they provide but also for their patients' satisfaction with the delivery and costs of care.

In many parts of the world, physicians increasingly are expected to account for the way in which they practice medicine by meeting certain standards prescribed by federal and local governments. The hospitalization of patients whose health care costs are reimbursed by the government and other third parties is subjected to utilization review. Thus, a physician must defend the cause for and duration of a patient's hospitalization if it falls outside certain "average" standards. Authorization for reimbursement increasingly is based on documentation of the nature and complexity of an illness, as reflected by recorded elements of the history and physical examination. A growing "pay-for-performance" movement seeks to link reimbursement to quality of care. The goal of this movement is to improve standards of health care and contain spiraling health care costs. In many parts of the United States, managed (capitated) care contracts with insurers have replaced traditional fee-for-service care, placing the onus of managing the cost of all care directly on the providers and increasing the emphasis on preventive strategies. In addition, physicians are expected to give evidence of their current competence through mandatory continuing education, patient record audits, maintenance of certification, and relicensing.

Medical Ethics and New Technologies

The rapid pace of technological advances has profound implications for medical applications that go far beyond the traditional goals of disease prevention, treatment, and cure. Cloning, genetic engineering, gene therapy, human–computer interfaces, nanotechnology, and use of targeted therapies have the potential to modify inherited predispositions to disease, select desired characteristics in embryos, augment “normal” human performance, replace failing tissues, and substantially prolong life span. Given their unique training, physicians have a responsibility to help shape the debate on the appropriate uses of and limits placed on these new techniques and to consider carefully the ethical issues associated with the implementation of such interventions. As medicine becomes more complex, shared decision-making is increasingly important, particularly in areas such as genetic counseling and end-of-life care, but also in most instances of considering diagnostic and treatment options.

Learning Medicine

More than a century has passed since the publication of the Flexner Report, a seminal study that transformed medical education and emphasized the scientific foundations of medicine as well as the acquisition of clinical skills. In an era of burgeoning information and access to medical simulation and informatics, many schools are implementing new curricula that emphasize lifelong learning and the acquisition of competencies in teamwork, communication skills, system-based practice, and professionalism. The tools of medicine also change continuously, necessitating formal training in the use of EMRs, large datasets, ultrasound, robotics, and new imaging techniques. These and other features of the medical school curriculum provide the foundation for many of the themes highlighted in this chapter and are expected to allow physicians to progress, with experience and learning over time, from competency to proficiency to mastery.

At a time when the amount of information that must be mastered to practice medicine continues to expand, increasing pressures both within and outside of medicine have led to the implementation of restrictions on the amount of time a physician-in-training can spend in the hospital and in clinics. Because the benefits associated with continuity of medical care and observation of a patient’s progress over time were thought to be outstripped by the stresses imposed on trainees by long hours and by fatigue-related errors, strict limits were set on the number of patients that trainees could be responsible for at one time, the number of new patients they could evaluate in a day on call, and the number of hours they could spend in the hospital. In 1980, residents in medicine worked in the hospital more than 90 hours per week on average. In 1989, their hours were restricted to no more than 80 per week. Resident physicians’ hours further decreased by ~10% between 1996 and 2008, and in 2010 the Accreditation Council for Graduate Medical Education further restricted (i.e., to 16 hours per shift) consecutive in-hospital duty hours for first-year residents. The impact of these changes is still being assessed, but the evidence that medical errors have decreased as a consequence is sparse. An unavoidable by-product of fewer hours at the bedside is an increase in the number of “handoffs” of patient responsibility from one physician to another. These transfers often involve a transition from a physician who knows the patient well, having evaluated that individual on admission, to a physician who knows the patient less well. It is imperative that these transitions of responsibility be handled with care and thoroughness, with all relevant information exchanged and acknowledged.

The Physician as Perpetual Student

From the time physicians graduate from medical school, it becomes all too apparent that this milestone is symbolic and that they must embrace the role of a "perpetual student." This realization is at the same time exhilarating and anxiety-provoking. It is exhilarating because physicians can apply constantly expanding knowledge to the treatment of their patients; it is anxiety-provoking because physicians realize that they will never know as much as they want or need to know. Ideally, physicians will translate the latter feeling into energy through which they can continue to improve and reach their potential. It is the physician's responsibility to pursue new knowledge continually by reading, attending conferences and courses, and consulting colleagues and the Internet. This is often a difficult task for a busy practitioner; however, a commitment to continued learning is an integral part of being a physician and must be given the highest priority.

The Physician as Citizen

Being a physician is a privilege. The capacity to apply one's skills for the benefit of fellow human beings is a noble calling. The physician–patient relationship is inherently unbalanced in the distribution of power. In light of their influence, physicians must always be aware of the potential impact of what they do and say and must always strive to strip away individual biases and preferences to find what is best for their patients. To the extent possible, physicians should also act within their communities to promote health and alleviate suffering. Meeting these goals begins by setting a healthy example and continues in taking action to deliver needed care even when personal financial compensation may not be available.

Research, Teaching, and the Practice of Medicine

The word *doctor* is derived from the Latin *docere*, "to teach." As teachers, physicians should share information and medical knowledge with colleagues, students of medicine and related professions, and their patients. The practice of medicine is dependent on the sum total of medical knowledge, which in turn is based on an unending chain of scientific discovery, clinical observation, analysis, and interpretation. Advances in medicine depend on the acquisition of new information through research, and improved medical care requires the transmission of that information. As part of their broader societal responsibilities, physicians should encourage patients to participate in ethical and properly approved clinical investigations if these studies do not impose undue hazard, discomfort, or inconvenience. Physicians engaged in clinical research must be alert to potential conflicts of interest between their research goals and their obligations to individual patients. The best interests of the patient must always take priority.

> *To wrest from nature the secrets which have perplexed philosophers in all ages, to track to their sources the causes of disease, to correlate the vast stores of knowledge, that they may be quickly available for the prevention and cure of disease—these are our ambitions.*
>
> —William Osler, 1849–1919

FURTHER READING

Cheston CC et al: Social media use in medical education: A systematic review. Acad Med 88:893, 2013. [PubMed: 23619071]

Cooke M et al: American medical education 100 years after the Flexner report. N Engl J Med 355:1339, 2006. [PubMed: 17005951]

Institute of Medicine: Dying in America: Improving quality and honoring individual preferences near the end of life. Washington, DC, National Academies Press, 2015.

Institute of Medicine: Improving diagnosis in health care. Washington, DC, National Academies of Sciences, Engineering, and Medicine, 2015.

Jameson JL, Longo DL: Precision medicine—personalized, problematic, and promising. N Engl J Med 372:2229, 2015. [PubMed: 26014593]

Reeves S et al: Interprofessional education: Effects on professional practice and healthcare outcomes (update). Cochrane Database Syst Rev 28:CD02213, 2013.

Stern DT, Papadakis M: The developing physician—becoming a professional. N Engl J Med 355:1794, 2006. [PubMed: 17065641]

Vickrey BG et al: How neurologists think: A cognitive psychology perspective on missed diagnoses. Ann Neurol 67:425, 2010. [PubMed: 20437577]

West P et al: Intervention to promote physician well-being, job satisfaction, and professionalism. A randomized clinical trial. JAMA Intern Med 174:527, 2014. [PubMed: 24515493]

Chapter 2: Promoting Good Health

Donald M. Lloyd-Jones; Kathleen M. McKibbin

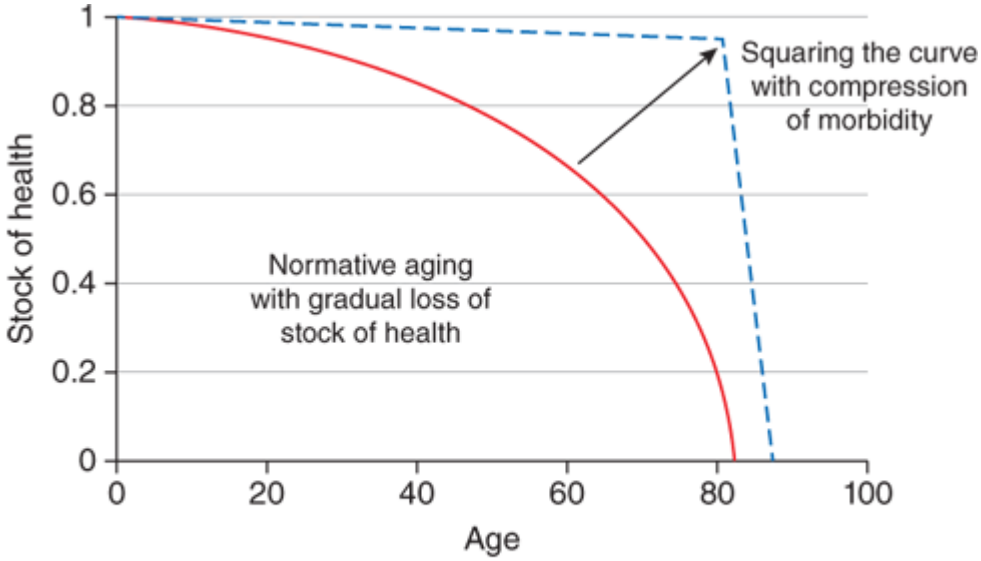
FIGURE 2-1

GOALS AND APPROACHES TO PREVENTION

Prevention of acute and chronic diseases before their onset has been recognized as one of the hallmarks of excellent medical practice for centuries, and is now used as a metric for highly functioning healthcare systems. The ultimate goal of preventive strategies is to avoid premature death. However, as longevity has increased dramatically worldwide over the last century (largely as a result of public health practices), increasing emphasis is placed on prevention for the purpose of preserving quality of life and extending the healthspan, not just the lifespan. Given that all patients will eventually die, the goal of prevention ultimately becomes compression of morbidity toward the end of the lifespan; that is, reduction of the amount of burden and time spent with disease prior to dying. As shown in Fig. 2-1, normative aging tends to involve a steady decline in the stock of health, with accelerating decline over time. Successful prevention offers the opportunity both to extend life and to extend healthy life, thus “squaring the curve” of health loss during aging.

FIGURE 2-1

Loss of health with aging. Representation of normative aging with loss of the full stock of health with which individuals are born (indicating gain of morbidity), contrasted with a squared curve with greater longevity and fuller stock of health (less morbidity) until shortly before death. The “squared curve” represents the likely ideal situation for most patients.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prevention strategies have been characterized as tertiary, secondary, primary, and primordial. *Tertiary prevention* requires rapid action to prevent imminent death in the setting of acute illness, such as through percutaneous coronary intervention in the setting of ST-segment elevation myocardial infarction. *Secondary prevention* strategies focus on avoiding the recurrence of disease and death in an individual who is already affected. For example, tamoxifen is recommended for women with surgically treated early-stage, estrogen-receptor-positive breast cancer, because it reduces the risk of recurrent breast cancer (including in the contralateral breast) and death. *Primary prevention* attempts to reduce the risk of incident disease among individuals with a risk factor. Treatment of elevated blood pressure in individuals who have not yet experienced cardiovascular disease represents one example of primary prevention that has proven effective in reducing the incidence of stroke, heart failure, and coronary heart disease.

Primordial prevention is a more recent concept (first introduced in 1979) which focuses on prevention of the development of *risk factors* for disease, not just prevention of disease. Primordial prevention strategies emphasize upstream determinants of risk for chronic diseases, such as eating patterns, physical activity, and environmental and social determinants of health. It therefore encompasses medical treatment strategies for individuals as well as a strong reliance on public health and social policy. It is increasingly clear that primordial prevention represents the ultimate means for reducing the burden of chronic diseases of aging. Once risk factors develop, it is difficult to restore risk to the low level of someone who never developed the risk factor. The time spent with adverse levels of the risk factor often causes irreversible damage that precludes complete restoration of low risk. For example, individuals with hypertension who are treated back to optimal levels (<120/<80 mmHg) do have a lower risk compared with untreated patients with hypertension, but they still have twice the risk of cardiovascular events as those who maintained optimal blood pressure without medications. Patients with elevated blood pressure that is subsequently treated have greater left ventricular mass index, worse renal function, and more evidence of atherosclerosis and other target organ damage as a result of the time spent with elevated blood pressure; such damage cannot be fully reversed despite efficacious therapy with antihypertensive medications. Conversely, as described below in greater detail, individuals who maintain optimal levels of all major cardiovascular risk factors into middle age

through primordial prevention essentially abolish their lifetime risk of developing cardiovascular disease while also living substantially longer and having a lower burden of other comorbid illnesses (compression of morbidity).

Prevention strategies should be distinguished from disease screening strategies. Screening attempts to detect evidence of disease at its earliest stages, when treatment is likely to be more efficacious than for advanced disease (Chap. 4). Screening can be performed in service of prevention, especially if it aids in identifying pre-clinical markers associated with elevated disease risk.

HEALTH PROMOTION

In recent decades, medical practice has increasingly focused on public health approaches to promote health, and not just prevent disease. Prevention of disease is a worthy individual and societal goal in and of itself, but it does not necessarily guarantee health. Health is a broader construct encompassing more than just absence of disease. It includes biological, physiological, and psychological domains (among others) in a continuum, rather than occurring as a dichotomous trait. Health is therefore somewhat subjective, but attempts have been made to use more objective criteria to define health in order to raise awareness, prevent disease, and promote healthy longevity.

For example, in 2010 the American Heart Association (AHA) defined a new construct of “cardiovascular health” based on evidence of associations with longevity, disease avoidance, healthy longevity, and quality of life. The definition of cardiovascular health is based on seven health behaviors and health factors (eating pattern, physical activity, body mass, smoking status, and levels of blood pressure, blood cholesterol, and blood glucose) and includes a spectrum from poor to ideal. Individuals with optimal levels of all seven metrics simultaneously are considered to have ideal cardiovascular health. The state of cardiovascular health for an individual or a population can be assessed with simple scoring by counting the number of ideal metrics (out of 7) or applying 0 points for each poor metric, 1 point for each intermediate metric, and 2 points for each ideal metric, thus creating a composite cardiovascular health score ranging from 0 to 14 points. Higher cardiovascular health scores in younger and middle ages have been associated with greater longevity, lower incidence of cardiovascular disease, lower incidence of other chronic diseases of aging (including dementia, cancer, and more), compression of morbidity, greater quality of life, and lower healthcare costs, achieving both individual and societal goals for healthy aging, and further establishing the critical importance of primordial prevention and cardiovascular health promotion.

Focusing on health promotion, rather than just disease prevention, may also provide greater motivation for patients to pursue lifestyle changes or adhere to clinician recommendations. Extensive literature suggests that providing patients solely with information regarding disease risk, or risk reduction with treatment, is unlikely to motivate desired behavior change. Empowering patients with strategies to achieve positive health goals after discussing risks can provide more effective adherence and better long-term outcomes. In the case of smoking cessation, enumerating only the risks of smoking can lead to patient inertia and therapeutic nihilism, and has proven an ineffective approach, whereas strategies that incorporate positive health

messaging, support and feedback, with appropriate use of evidence-based therapies, have proven far more effective.

PRIORITIZING PREVENTION STRATEGIES

In secondary prevention, the patient already has manifest clinical disease, and is therefore at high risk for progression. The approach should be to work with the patient to implement all evidence-based strategies that will help to prevent recurrence or progression. This will typically include drug therapy as well as therapeutic lifestyle changes to control ongoing risk factors which may have caused disease in the first place. Juggling priorities can be difficult, and barriers to implementation are many, including costs, time, patient health literacy, and patient and caregiver capacity to organize the regimen. Addressing these potential barriers with the patient can help to forge a therapeutic bond and may improve adherence; ignoring them will likely lead to therapeutic failure. Numerous studies demonstrate that, even in high-functioning health systems, only ~50% of patients are taking recommended, evidence-based secondary prevention medications, such as statins, by 1 year after a myocardial infarction.

In patients who are eligible for primary prevention strategies, it is important to frame the discussion around the overall evidence base as well as an individual patient's likelihood of benefit from a given preventive intervention. A first step is to understand the patient's estimated absolute risk for disease in the foreseeable future, or during their remaining lifespan. However, absolute risk estimation and presentation of those risks is generally insufficient to motivate behavior change. It is critical to assess the patient's understanding and tolerance of the risk, their readiness to implement lifestyle changes or adhere to drug therapy, and their overall preferences regarding use of drug therapy to prevent an event (e.g., cancer, myocardial infarction, stroke). The clinician can help the patient by informing them of the risks for disease and potential for absolute benefits (and harms) from the available evidence-based choices. This may take more than one conversation, but given that diseases, such as cancer and cardiovascular disease, are the leading causes of premature death and disability, the time is well spent.

Partnering with the patient through motivational interviewing may assist in the process of selecting initial approaches to prevention. Selecting an area that the patient feels they are ready to change can lead to better adherence and greater achievement of success in the short and longer term. If the patient is uncertain what course to choose, prudence would dictate focusing on control of risk factors that may lead to the most rapid reduction in risk for acute events. For example, blood pressure is both a chronic risk factor and an acute trigger for cardiovascular events. Thus, if a patient has both significant elevations in blood pressure and dyslipidemia, it would be appropriate to focus initial efforts on blood pressure control. Likewise, focus on smoking cessation can lead to more rapid reductions in risk for acute events than some other lifestyle interventions.

PREVENTION AND HEALTH PROMOTION ACROSS THE LIFE COURSE

Periodic Health Evaluations

The “routine annual physical” has in many ways become an expected part of the patient-physician relationship in primary care practice. However, evidence for the efficacy of the periodic health evaluation in asymptomatic adults unselected for risk factors or disease is mixed, and depends on the outcome. Systematic reviews and meta-analyses of published trials have consistently observed lack of benefit (and also lack of harm) in terms of total mortality in association with periodic health evaluations. Data are more heterogeneous but overall suggest no benefit for cancer- or cardiovascular-specific mortality, with the potential for either benefit or harm depending on number of evaluations and patient-level factors. Well-designed studies on non-fatal clinical events and morbidity have been sparsely reported but there appear to be no large effects.

Periodic health evaluations do appear to lead to greater diagnosis of certain conditions such as hypertension and dyslipidemia, as expected. Likewise, periodic health examinations also improve the delivery of recommended preventive services, such as gynecologic examinations and Papanicolaou smears, fecal occult blood testing, and cholesterol screening. The benefits and risks associated with screening tests are discussed in detail in Chap. 4. Risks of routine evaluations include inappropriate or over-testing, or false-positive findings that require follow-up and induce patients to worry. Periodic health examinations appear to be associated with less patient worry. On balance, given the lack of convincing evidence of harm and the potential for better delivery of appropriate screening, counseling, and preventive services, periodic health evaluations appear reasonable for general populations at average risk for chronic conditions.

It is important to note that routine annual comprehensive physical examinations of asymptomatic adult patients have very low yield and may take an inordinate amount of time in a wellness visit. Such time may be better spent on assessing and counseling the patient on other aspects of their health, as discussed below. Evidence-based components that should be included in periodic evaluations focused on health and prevention include a number of age-appropriate screening tests for chronic disease and risk factors, preventive interventions including immunizations and chemoprevention for at-risk individuals, and preventive counseling. The United States Preventive Services Task Force publishes its *Guide to Clinical Preventive Services,* which contains evidence-based recommendations from the Task Force on preventive services for which there is a high degree of certainty that the service provides at least moderate net clinical benefit (i.e., benefits outweigh harms significantly and to a reasonable magnitude).

Healthy Behaviors and Lifestyles

Owing to the paucity of evidence, the heterogeneity of study designs and the diverse nature of interventions studied, many clinicians are uncertain as to how to deliver advice regarding healthy behaviors and lifestyles. Nevertheless, adverse behaviors and lifestyles contribute to more than 75% of premature, preventable deaths and disability. Estimates from the US National Health and Nutrition Survey indicate that fewer than 1% of Americans achieve an optimal heart-healthy eating pattern. Thus, whereas there are many demands on time during a typical patient-clinician encounter, few things may have more impact on longevity, health and quality of life for asymptomatic patients than an efficient approach to assessing, documenting, and improving patients’ health behaviors. Indeed, the mere act of assessing health behaviors has been shown to

affect patient's health behaviors. Facility with tools for assessment of lifestyle and with strategies for counseling are therefore of paramount importance.

Healthy Eating Patterns

Despite the existence of numerous "fad" diets, and seemingly inconsistent recommendations on dietary composition, there is remarkable agreement about what should constitute a healthy eating pattern for the broad population to avoid nutritional deficits (i.e., vitamin deficiency) and excesses (i.e., excessive caloric intake) and to maximize potential health **(**Table 2-1, **see** Chap. 325**)**. Optimal eating patterns consist of whole fruits and vegetables, whole grains, lean proteins, healthy oils, and allow for non-fat or low-fat dairy intake. They tend to exclude frequent ingestion of foods high in refined sugars and starches, saturated fat, and sodium. Since sodium and refined sugars and starches are the hallmark of much of the processed/packaged food supply, a simple rule of thumb is to provide/cook the majority of one's own meals starting from whole foods and emphasizing fruits and vegetables. Likewise, foods prepared outside of the home tend to have higher fat and sodium content, so special attention to menu choices focused on fruits, vegetables, lean proteins, and whole grains, while minimizing sauces and dressings can help most individuals follow healthier eating patterns. In all cases, sugar-sweetened beverages and non-nutritious snack foods should be minimized. If snacks are included, small amounts of healthy nuts and seeds, or more fruits and vegetables, should be encouraged.

TABLE 2-1

Guidelines and Key Recommendations from the *Dietary Guidelines for Americans*, 2015–2020

Guidelines	Key Recommendations
1. **Follow a healthy eating pattern across the lifespan.** All food and beverage choices matter. Choose a healthy eating pattern at an appropriate calorie level to help achieve and maintain a healthy body weight, support nutrient adequacy, and reduce the risk of chronic disease. 2. **Focus on variety, nutrient density, and amount.** To meet nutrient needs within calorie limits, choose a variety of nutrient-dense foods across and within all food groups in recommended amounts. 3. **Limit calories from added sugars and saturated fats and reduce sodium intake.** Consume an eating pattern low in added sugars, saturated fats, and sodium. Cut back on foods and beverages higher in these components to amounts that fit within healthy eating patterns. 4. **Shift to healthier food and beverage choices.** Choose nutrient-dense foods and beverages across and within all food groups in place of less healthy choices. Consider cultural and personal preferences to make these shifts easier to accomplish and maintain.	The Dietary Guidelines' Key Recommendations for healthy eating patterns should be applied in their entirety, given the interconnected relationship that each dietary component can have with others. **Consume a healthy eating pattern that accounts for all foods and beverages within an appropriate calorie level.** **A healthy eating pattern includes:** • A variety of vegetables from all of the subgroups—dark green, red and orange, legumes (beans and peas), starchy, and other • Fruits, especially whole fruits • Grains, at least half of which are whole grains • Fat-free or low-fat dairy, including milk, yogurt, cheese, and/or fortified soy beverages • A variety of protein foods, including seafood, lean meats and poultry, eggs, legumes (beans and peas), and nuts, seeds, and soy products • Oils **A healthy eating pattern limits:** • Saturated fats and trans fats, added sugars, and sodium Key Recommendations that are quantitative are provided for several components of the diet that should be limited. These components are of particular public health concern in the United States, and the specified limits can help individuals achieve healthy eating patterns within calorie limits: • Consume <10% of calories per day from added sugars • Consume <10% of calories per day from saturated fats • Consume <2300 milligrams (mg) per day of sodium • If alcohol is consumed, it should be consumed in moderation—up to one drink per day for women and up to two drinks per day for men—and only by adults of legal drinking age. In tandem with the recommendations above, Americans of all ages—children, adolescents, adults, and older adults—should meet the *Physical Activity Guidelines for Americans* to help promote health

Guidelines	Key Recommendations
5. **Support healthy eating patterns for all.** Everyone has a role in helping to create and support healthy eating patterns in multiple settings nationwide, from home to school to work to communities.	and reduce the risk of chronic disease. Americans should aim to achieve and maintain a healthy body weight. The relationship between diet and physical activity contributes to calorie balance and managing body weight. As such, the Dietary Guidelines includes a Key Recommendation to: **Meet the US Department of Health and Human Services' *Physical Activity Guidelines for Americans***

Source: Adapted from the *Dietary Guidelines for Americans, 2015–2020*. Washington, DC: US Department of Agriculture and US Department of Health and Human Services; 2015. Available at *https://health.gov/dietaryguidelines/2015/guidelines/*.

Specific conditions and diseases, such as diabetes, other metabolic disorders, allergies, and gastrointestinal disorders, may require tailored approaches to diet. In counseling most patients, the general approach should focus on whole foods, eating patterns and appropriate calorie balance, rather than on specific micronutrients such as electrolytes or selected vitamins. It should be remembered that most patients have difficulty understanding nutritional labels on packaged foods, with the attendant demands on numeracy and health literacy.

Dietary guidelines are published by the US Department of Agriculture (USDA) and US Department of Health and Human Services every 5 years, and these guidelines have undergone substantial evolution over time. The current US Dietary Guidelines and Key Recommendations for 2015–2020 are summarized in Table 2-1 and emphasize eating patterns with nutrient-dense (rather than calorie-dense) whole foods, and appropriate caloric intake to achieve and maintain healthy weight. The USDA Guidelines focus on the concept of a healthy plate (rather than the prior food pyramid) for ease of counseling and adoption. Fifty percent of the plate should consist of fruits and vegetables, with remaining portions for whole grains and lean protein foods. When using fat for cooking, it should be done by sauteing in healthier oils (e.g., canola oil), and addition of judicious amounts of healthy raw oils (e.g., olive oil) to dishes is appropriate.

The USDA Guidelines focus on specific healthy eating patterns that adhere to these broad recommendations, and are appropriate for ~97% of the general population. They identify a "Healthy US-Style Eating Pattern" that adheres closely to the evidence-based Dietary Approaches to Stop Hypertension (DASH) eating pattern. Alternative patterns, which vary more in emphasis than in content, include a "Healthy Mediterranean-Style Eating Pattern" and a "Healthy Vegetarian Eating Pattern."

AGE- AND SEX-SPECIFIC RECOMMENDATIONS

Current dietary recommendations are generally similar for all life stages from ages ≥2 years, but recommended levels of caloric intake (and hence amounts of foods) differ by age, sex, and physical activity

level. For example, recommended caloric intake ranges from 1000 calories/d for sedentary 2-year-old children to as high as 3200 calories/d for active 16- to 18-year-old young men. Recommended caloric intakes peak in the early twenties for men and women and gradually decrease over ensuing decades.

As with all lifestyle counseling aimed at behavior change, dietary approaches that partner with the patient and utilize motivational interviewing strategies and shared goals and commitments tend to work best, as described below (see Approach to the Patient).

Physical Activity

Similar to the approach to counseling regarding healthy eating patterns, recommendations on participation in physical activity emphasize the point that any physical activity is better than none. A simple rule of thumb for patients is: “If you are doing nothing, do something; and if you are doing something, do more, every day.” The evidence base for physical activity indicates that the marginal benefits from physical activity are greatest in advancing from no activity to low levels of moderate activity. With increasing duration and intensity of activity, there is a continued curvilinear increase in health benefits, but the marginal gains for each additional minute of moderate-to-vigorous activity slowly diminish. Thus, for adults, the optimal amount of physical activity recommended is 150 min of moderate-intensity or 75 min of vigorous intensity aerobic activity per week, performed in episodes of at least 10 min, and preferably spread throughout the week. Additional health benefits can be realized by engaging in physical activity beyond this amount, and/or by adding muscle-strengthening activities that involve all major muscle groups 2 or more days per week.

In counseling patients regarding physical activity, it is important to note that sedentary time (e.g., seated at work, or at home in front of electronic screens) has adverse health consequences independent of the lack of physical activity during these episodes. Therefore, even modest efforts like standing at the desk and doing gentle stretching for periods during the day may be beneficial. It is also important to emphasize that participating in a variety of aerobic activities (biking, swimming, walking, jogging, rowing, elliptical training, stair-climbing, etc.) can be beneficial and may help to avoid overuse injuries and boredom with the exercise regimen. If patients choose to participate in muscle-strengthening activities for health improvement, emphasis should be placed on weights that allow more repetitions (e.g., 3 sets of 15–20 repetitions that can be performed comfortably, with a rest period in between) and on avoiding breath-holding and straining against a closed glottis.

SUDDEN CARDIAC DEATH RISK

Patients may express concerns regarding the risk of sudden cardiac death during exercise. Whereas the risk of sudden death during exercise does increase directly with the amount of time spent exercising, this association is substantially mitigated by training effects. Thus, patients embarking on an exercise program should be encouraged to increase the duration of aerobic exercise gradually as tolerated, aiming for episodes of at least 30 min 5 times a week as an ideal. Once a comfortable duration is reached, incorporating interval training periods of more intensive activity interspersed during the exercise can provide greater fitness gains.

EXTREME ENDURANCE ACTIVITIES

As with other forms of exercise, extreme endurance activities such as triathlons and marathons should be undertaken only with appropriate and graded training. Such activities tend to take a greater toll on the musculoskeletal system over time than less extreme activities, and they are also associated with measurable damage to the myocardium and greater risks for other organ damage. Athletes participating in endurance activities routinely have elevations in cardiac troponin (a specific circulating marker of myocardial cell damage and death) at the end of the race, although elevations are lower in those who are well trained. Patients and clinicians should consider the patient's overall health, specific limitations, potential for injury, and ability to train in decision-making regarding participation in endurance events.

AGE-SPECIFIC RECOMMENDATIONS

The US Department of Health and Human Services' *Physical Activity Guidelines for Americans* (Table 2-2) recommend that children and adolescents aged 6–17 years should participate in ≥60 min of physical activity daily, most of which should be moderate- or vigorous-intensity aerobic activity, including vigorous activity at least 3 days a week. As noted above, adults aged 18–64 years are recommended to pursue at least 150 min of moderate-intensity or 75 min of vigorous-intensity aerobic activity per week (or equivalent combinations). Adults aged ≥65 years should follow the adult guidelines, or be as active as possible as abilities and conditions allow. Special emphasis is also placed on exercises to improve balance in those at risk for falling.

TABLE 2-2

Recommendations from *Physical Activity Guidelines for Americans*

Age	Recommendations
6–17 years	Children and adolescents should do 60 min (1 h) or more of physical activity daily. • Aerobic: Most of the ≥60 min a day should be either moderate[a] or vigorous-intensity[b] aerobic physical activity, and should include vigorous-intensity physical activity at least 3 days a week. • Muscle-strengthening:[c] As part of their ≥60 min of daily physical activity, children and adolescents should include muscle-strengthening physical activity on at least 3 days of the week. • Bone-strengthening:[d] As part of their ≥60 min of daily physical activity, children and adolescents should include bone-strengthening physical activity on at least 3 days of the week. • It is important to encourage young people to participate in physical activities that are appropriate for their age, that are enjoyable, and that offer variety.
18–64 years	• All adults should avoid inactivity. Some physical activity is better than none, and adults who participate in any amount of physical activity gain some health benefits. • For substantial health benefits, adults should do at least 150 min (2 h and 30 min) a week of moderate-intensity, or 75 min (1 h and 15 min) a week of vigorous-intensity aerobic physical activity, or an equivalent combination of moderate- and vigorous-intensity aerobic activity. Aerobic activity should be performed in episodes of at least 10 min, and preferably, it should be spread throughout the week. • For additional and more extensive health benefits, adults should increase their aerobic physical activity to 300 min (5 h) a week of moderate-intensity, or 150 min a week of vigorous-intensity aerobic physical activity, or an equivalent combination of moderate- and vigorous-intensity activity. Additional health benefits are gained by engaging in physical activity beyond this amount. • Adults should also include muscle-strengthening activities that involve all major muscle groups on ≥2 days a week.
≥65 years	• Older adults should follow the adult guidelines. When older adults cannot meet the adult guidelines, they should be as physically active as their abilities and conditions will allow. • Older adults should do exercises that maintain or improve balance if they are at risk of falling. • Older adults should determine their level of effort for physical activity relative to their level of fitness. • Older adults with chronic conditions should understand whether and how their conditions affect their ability to do regular physical activity safely.

[a]Moderate-intensity physical activity: Aerobic activity that increases a person's heart rate and breathing to some extent. On a scale relative to a person's capacity, moderate-intensity activity is usually a 5 or 6 on a 0 to 10 scale. Brisk walking, dancing, swimming, or bicycling on a level terrain are examples. [b]Vigorous-intensity physical activity: Aerobic activity that greatly increases a person's heart rate and breathing. On a scale relative to a person's capacity, vigorous-intensity activity is usually a 7 or 8 on a 0 to 10 scale. Jogging, singles tennis, swimming continuous laps, or bicycling uphill are examples. [c]Muscle-strengthening activity: Physical activity, including exercise that increases skeletal muscle strength, power, endurance, and mass. It includes strength training, resistance training, and muscular strength and endurance exercises. [d]Bone-strengthening activity: Physical activity that produces an impact or tension force on bones, which promotes bone growth and strength. Running, jumping rope, and lifting weights are examples.

Source: Adapted from U.S. Department of Health and Human Services. *2008 Physical Activity Guidelines for Americans.* Washington, DC: U.S. Department of Health and Human Services; 2008. Available at *http://www.health.gov/paguidelines*.

Sleep Hygiene

Sleeping between 7 and 9 h per night appears to be optimal for health in adults aged ≥18 years. Sleeping <7 h is associated with adverse outcomes, including obesity, diabetes, elevated blood pressure, cardiovascular disease, depression, and all-cause mortality, as well as physiologic disturbances such as impaired immune function, increased pain sensitivity, and impaired cognitive performance. Conversely, achieving appropriate levels of sleep is associated with more success in weight loss, better blood pressure control among patients with hypertension, and improved mental health and performance. Regular sleep more than 9 h per night is appropriate for children and adolescents, or individuals recovering from sleep deprivation or illness, but for most individuals the effects on health are uncertain.

Patients often express concerns about the quantity and quality of their sleep. With aging, both aspects of sleep tend to decline, even without overt sleep disorders. Documentation of sleep using a sleep log may assist in understanding different types of insomnia and sleep disorders. Encouraging daily activity to promote fatigue, avoidance of eating and drinking alcohol too close to bedtime, and regular daily sleep habits may help patients achieve better sleep. Regular use of sedative medications should generally be discouraged given the high potential for dependence, addiction, and altered sleep quality.

DISORDERS OF SLEEP

The prevalence of sleep-related breathing disorders, including obstructive sleep apnea (OSA), is poorly documented. Based on data from the 1990s, the prevalence of diagnosed mild OSA in the US population was ~10%, and of moderate to severe apnea was ~5%. However, the increasing prevalence of obesity, a major risk factor for OSA, suggests that the prevalence may have increased. The prevalence of asymptomatic or undiagnosed sleep apnea is unknown. Patients with persistent complaints of poor sleep quality, excessive daytime somnolence, or with witnessed apneic spells may benefit from screening for sleep disorders, prior to consideration of a formal sleep study. A number of clinical tools have been developed to screen for sleep

apnea, including the Epworth Sleepiness Scale, the STOP (Snoring, Tiredness, Observed apnea, high blood Pressure) Questionnaire, and the STOP-Bang Questionnaire (STOP plus assessment of body mass index, age, neck circumference, and gender), among others. The US Preventive Services Task Force found that current evidence is insufficient to assess the balance of benefits and harms of screening for OSA in asymptomatic adults owing to a lack of validation data in primary care settings. Nonetheless, the high prevalence and significant health consequences of sleep apnea suggest that clinicians should be alert for its potential presence, particularly in patients who are obese with symptoms of excessive daytime somnolence or witnessed apnea episodes. Other sleep disorders, such as restless leg syndrome, may be identified with simple history.

Weight Management

Overweight and obesity are prevalent in epidemic proportions in the US and other industrialized nations (**Chaps. 394, 395**). Since 1985, the prevalence of obesity in the United States has increased from ~10% to almost 35%, and the prevalence of overweight is now ~40%. Overweight and obesity disproportionately affect individuals in lower socio-economic strata, and in many underserved minority populations, including African Americans, Latino Americans, and American Indians. In all race-ethnic groups, both overweight and obesity are associated with adverse health consequences, including diabetes, certain cancers, cardiovascular diseases, and degenerative joint disease. Eating disorders such as anorexia and bulimia are much less common but pose major health consequences for affected patients, and should be suspected particularly in younger women with history of rapid weight shifts or underweight status.

Weight loss is one of the most difficult preventive interventions to achieve and sustain over time. However, several key factors can assist the patient and clinician, and early referral to a dietician can be very helpful. The first therapeutic goal is to aim for weight stabilization. Many of the risks of overweight and obesity are driven more strongly by continued weight gain, rather than overweight/obese status per se. Working with the patient to find initial strategies for weight maintenance can be a successful initial step with success for many patients. For those who can progress to considering weight loss, it is critical to help the patient understand that there is no standard solution. Experimentation and documentation are key. Tools to assist patients can include food and weight logs, activity logs, and smart phone apps. Some patients respond best to structured commercial dietary programs where meals are provided to them. Any of these approaches can be tried with or without social group supports.

The key construct for weight loss is, of course, negative calorie balance. This is achieved through a combination of reduced caloric intake and increased physical activity. Patients may already understand, from prior weight loss attempts, what combination works best for them to achieve this. Some patients find that they cannot lose weight without increasing their exercise. For many, reduction of caloric intake is most efficient. Encouraging the patient to find what works for them is most important. The same principle holds for dietary content. Well done feeding studies indicate that weigh loss is dependent far more on the reduction of caloric intake than on the relative composition of fat, protein and carbohydrate in the diet.

There may be other medical reasons to choose one approach over another, but if not, encouraging the patient to pick one approach and document the results is an important start.

Tobacco Cessation

Escaping nicotine dependence is another major, but critical, challenge to prevention and wellness efforts **(see Chap. 448)**. The addictive effects of nicotine have been well documented, with effects that can last for years after successful cessation. Assessing a patient's past history of cessation attempts and current readiness for change are key first steps in forging a successful approach. Frequent follow-up and reinforcement, as well as use of nicotine replacement therapy and other cessation-promoting medications are additional critical elements. Recidivism is the rule, and patients should expect to resume smoking and attempt again as they journey to tobacco cessation.

MENTAL HEALTH AND ADDICTION

Assessment for depression and cognitive impairment are important to address when patients exhibit symptoms, or they or their family members express concerns. Both of these common conditions play a major role in reducing quality of life and are high on patients' lists of concerns, even if not clearly expressed. Screening tools for depression are reviewed in **Chap. 444**. Cognitive function decline with aging or comorbid illness, including depression, should be anticipated. Assessment tools such as the General Practitioner Assessment of Cognition or the Mini-Cog™ test are widely available and effective rapid assessment tools.

Alcohol and Opioids

Alcohol dependence and abuse are common and underdiagnosed **(see Chaps. 445, 446)**. Rapid screening tools have proven efficacy for identifying patients with alcohol problems. In a systematic review, the CAGE (cut down, guilty, annoyed, eye opener) questionnaire was most effective at identifying alcohol abuse and dependence, with reasonable sensitivity and high specificity. The present opioid epidemic in the United States presents a new and substantial public health challenge given the high potential for dependency and abuse of these drugs. Rapid screening tools are being developed and validated to assist clinicians in screening for opioid dependence.

ACCIDENTS AND SUICIDE

Regular assessment of patient safety through simple questions about seat belt use, domestic violence, and gun safety in the home continue to be important parts of health promotion and wellness. Longstanding recommendations for assessment of suicidal ideation among patients with depression or a history of suicide attempts also continue to be relevant.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

In the context of a clinical visit focused on health assessment, health promotion, and prevention, the basic skills of history taking are of paramount importance. Much of the evaluation, counseling, and management that focus on health promotion and prevention also require engagement and buy-in from the patient in order to assist with recognition of contributing behaviors and to promote adherence to therapeutic plans. Therefore, in addition to standard history-taking, additional skills such as motivational interviewing and eliciting patient commitments and contracting may prove of significant value. The availability of additional tools to assist with screening and chronic management, both online and through mobile health technologies, is rapidly expanding, with uncertain implications for the future. Major research gaps exist in our understanding of how best to employ these newer technologies to improve health outcomes. Concepts of behavioral economics are being explored to better understand the psychology of decision-making and incentives as a means to improve lifestyle choices and adherence to treatment plans **(Chap. 468)**.

The limited time available to clinicians and patients during a wellness visit or periodic health examination (not driven by specific patient issues) makes it important to prioritize assessment and counseling for factors that affect longevity, healthspan, and quality of life over approaches that may have low yield, such as the annual comprehensive physical examination in an asymptomatic patient. Setting clear expectations for the content of a wellness visit may be a first step, and scheduling follow-up visits for findings or to continue indicated counseling are important steps to achieving better health outcomes.

FURTHER READING

Boulware LE et al: Systematic review: The value of the periodic health evaluation. Ann Intern Med 146:289, 2007. [PubMed: 17310053]

Dietary Guidelines for Americans, 2015–2020. Washington, DC: US Department of Agriculture and US Department of Health and Human Services; 2015. Available at *https://health.gov/dietaryguidelines/2015/guidelines/*.

Irish LA et al: The role of sleep hygiene in promoting public health: A review of empirical evidence. Sleep Med Rev 22:23, 2015. [PubMed: 25454674]

Krogsboll LT et al: General health checks in adults for reducing morbidity and mortality from disease: Cochrane systematic review and meta-analysis. BMJ 345:e7191, 2012. [PubMed: 23169868]

U.S. Department of Health and Human Services: *2008 Physical Activity Guidelines for Americans.* Washington DC: U.S. Department of Health and Human Services; 2008. Available at *http://www.health.gov/paguidelines*.

U.S. Preventive Services Task Force: The Guide to Clinical Preventive Services, 2014. Washington, DC: U.S. Department of Health and Human Services, Agency for Healthcare Research and Quality; 2014. Available at *https://www.ahrq.gov/sites/default/files/publications/files/cpsguide.pdf*.

Chapter 3: Decision-Making in Clinical Medicine

Daniel B. Mark; John B. Wong

INTRODUCTION

Sir William Osler's familiar quote "Medicine is a science of uncertainty and an art of probability" captures well the complex nature of clinical medicine. Although the science of medicine is often taught as if the mechanisms of the human body operate with Newtonian predictability, every aspect of medical practice is infused with an element of irreducible uncertainty that the clinician ignores at her peril. Clinical medicine has deep roots in science, but it is an imprecise science. More than 100 years after the practice of medicine took its modern form, it remains at its core a craft, to which individual doctors bring varying levels of skill and understanding. With the exponential growth in medical literature and other technical information and an ever increasing number of testing and treatment options, twenty-first century physicians who seek excellence in their craft must master a more diverse and complex set of skills than any of the generations that preceded them. This chapter provides an introduction to three of the pillars upon which the craft of modern medicine rests: (1) expertise in clinical reasoning (what it is and how it can be developed); (2) rational diagnostic tests, use and interpretation; and (3) integration of the best available research evidence with clinical judgment in the care of individual patients (*evidence-based medicine* or *EBM* and the tools of EBM).

BRIEF INTRODUCTION TO CLINICAL REASONING

Clinical Expertise

Defining "clinical expertise" remains surprisingly difficult. Chess has an objective ranking system based on skill and performance criteria. Athletics, similarly, have ranking systems to distinguish novices from Olympians. But in medicine, after physicians complete training and pass the boards (or get recertified), no tests or benchmarks are used to identify those who have attained the highest levels of clinical performance. Physicians often consult a few "elite" clinicians for their "special problem-solving prowess" when particularly difficult or obscure cases have baffled everyone else. Yet despite their skill, even such master clinicians typically cannot explain their exact processes and methods, thereby limiting the acquisition and dissemination of the expertise used to achieve their impressive results. Furthermore, clinical virtuosity appears not to be generalizable, e.g., an expert on hypertrophic cardiomyopathy may be no better (and possibly worse) than a first-year medical resident at diagnosing and managing a patient with neutropenia, fever, and hypotension.

Broadly construed, clinical expertise includes not only cognitive dimensions involving the integration of disease knowledge with verbal and visual cues and test interpretation but also potentially the complex fine-motor skills necessary for invasive procedures and tests. In addition, “the complete package” of expertise in medicine requires effective communication and care coordination with patients and members of the medical team. Research on medical expertise remains sparse overall and mostly centered on diagnostic reasoning, so in this chapter, we focus primarily on the cognitive elements of clinical reasoning.

Because clinical reasoning occurs in the heads of clinicians, objective study of the process is difficult. One research method used for this area asks clinicians to “think out loud” as they receive increments of clinical information in a manner meant to simulate a clinical encounter. Another research approach focuses on how doctors should reason diagnostically to identify remediable “errors” rather than on how they actually do reason. Much of what is known about clinical reasoning comes from empirical studies of nonmedical problem-solving behavior. Because of the diverse perspectives contributing to this area, with important contributions from cognitive psychology, medical education, behavioral economics, sociology, informatics, and decision sciences, no single integrated model of clinical reasoning exists, and not infrequently, different terms and reasoning models describe similar phenomena.

Intuitive Versus Analytic Reasoning

A useful contemporary model of reasoning, dual-process theory distinguishes two general systems of cognitive processes. *Intuition* (System 1) provides rapid effortless judgments from memorized associations using pattern recognition and other simplifying “rules of thumb” (i.e., heuristics). For example, a very simple pattern that could be useful in certain situations is “African-American women plus hilar adenopathy equals sarcoid.” Because no effort is involved in recalling the pattern, typically, the clinician is unable to say how those judgments were formulated. In contrast, *Analysis* (System 2), the other form of reasoning in the dual-process model, is slow, methodical, deliberative, and effortful. A student might read about lymph nodes in the lung and from that list (e.g., **Chap. 62**), identify diseases more common in African-American women or examine the patient for skin or eye findings that may occur with sarcoid. These dual processes, of course, represent two exemplars taken from the cognitive continuum. They provide helpful descriptive insights but very little guidance in how to develop expertise in clinical reasoning. How these idealized systems interact in different decision problems, how experts use them differently from novices, and when their use can lead to errors in judgment remain the subject of study and considerable debate.

Pattern recognition, an important part of System 1 reasoning, is a complex cognitive process that appears largely effortless. One can recognize people’s faces, the breed of a dog, an automobile model, or a piece of music from just a few notes within milliseconds without necessarily being able to articulate the specific features that prompted the recognition. Analogously, experienced clinicians often recognize familiar diagnosis patterns very quickly. The key here is having a large library of stored patterns that can be rapidly accessed. In the absence of an extensive stored repertoire of diagnostic patterns, students (as well as more experienced clinicians operating outside their area of expertise and familiarity) often must use the more

laborious System 2 analytic approach along with more intensive and comprehensive data collection to reach the diagnosis.

The following three brief scenarios of a patient with hemoptysis illustrate three distinct patterns that experienced clinicians recognize without effort:

A 46-year-old man presents to his internist with a chief complaint of hemoptysis. An otherwise healthy, nonsmoker, he is recovering from an apparent viral bronchitis. This presentation pattern suggests that the small amount of blood-streaked sputum is due to acute bronchitis, so that a chest x-ray provides sufficient reassurance that a more serious disorder is absent.

In the second scenario, a 46-year-old patient who has the same chief complaint but with a 100-pack-year smoking history, a productive morning cough, with blood-streaked sputum, and weight loss fits the pattern of carcinoma of the lung. Consequently, along with the chest x-ray, the clinician obtains a sputum cytology examination and refers this patient for a chest CT scan.

In the third scenario, the clinician hears a soft diastolic rumbling murmur at the apex on cardiac auscultation in a 46-year-old patient with hemoptysis who immigrated from a developing country and orders an echocardiogram as well, because of possible pulmonary hypertension from suspected rheumatic mitral stenosis.

Pattern recognition by itself is not, however, sufficient for secure diagnosis. Without deliberative systematic reflection, pattern recognition can result in premature closure: mistakenly jumping to the conclusion that one has correct diagnosis before all the relevant data are in. A critical second step, even when the diagnosis seems obvious, is *diagnostic verification*: considering whether the diagnosis adequately accounts for the presenting symptoms and signs and can explain all the ancillary findings. An example of premature closure is contained in the following case, modified from a real clinical encounter. A 45-year-old man presents with a 3-week history of a “flulike” upper respiratory infection (URI) including dyspnea and a productive cough. The Emergency Department (ED) clinician pulled out a “URI assessment form” which defines and standardizes the information gathered. After quickly acquiring the requisite structured examination components and noting in particular the absence of fever and a clear chest examination, the physician prescribed a cough suppressant for acute bronchitis and reassured the patient that his illness was not serious. Following a sleepless night at home with significant dyspnea, the patient developed nausea and vomiting and collapsed. He was brought back to the ED in cardiac arrest and was unable to be resuscitated. His autopsy showed a posterior wall myocardial infarction (MI) and a fresh thrombus in an atherosclerotic right coronary artery. What went wrong? Presumably, the ED clinician felt that the patient was basically healthy (one can be misled by the way the patient appears on examination—a patient that does not “appear sick” may be incorrectly assumed to have an innocuous illness). So in this case, the physician, upon hearing the overview of the patient from the triage nurse, elected to use the URI assessment protocol even before starting the history, closing consideration of the broader range of possibilities and associated tests required to confirm or refute these possibilities. In particular, by concentrating on the abbreviated and focused URI protocol, the clinician

failed to elicit the full dyspnea history, which was precipitated by exertion and accompanied by chest heaviness and relieved by rest, suggesting a far more serious disorder.

Heuristics or rules of thumb are a part of the intuitive system. These cognitive shortcuts provide a quick and easy path to reaching conclusions and making choices, but when used improperly they can lead to errors. Two major research programs have studied heuristics in a mostly non-medical context and have reached very different conclusions about the value of these cognitive tools. The “heuristics and biases” program focuses on how relying on heuristics can lead to cognitive biases and incorrect judgments. Over 100 different cognitive biases have been described. So far, however, there is little evidence that educating physicians and other decision makers to watch for these cognitive biases has any effect on the rate of diagnostic errors. In contrast, the “fast and frugal heuristics” research program explores how and when relying on simple heuristics can produce good decisions. Although many heuristics have relevance to clinical reasoning, only four will be mentioned here.

When diagnosing patients, clinicians usually develop diagnostic hypotheses based on the similarity of that patient’s symptoms, signs and other data to their mental representations (memorized patterns) of the disease possibilities. In other words, clinicians pattern match to identify the diagnoses which share the most similar findings to the patient at hand. This cognitive shortcut is called the representativeness heuristic. Consider a patient with hypertension and headache, palpitations, and diaphoresis. Based on the representativeness heuristic, clinicians might judge pheochromocytoma to be quite likely given this classic presenting symptom triad suggesting pheochromocytoma. Doing so however, would be incorrect given that other causes of hypertension are much more common than pheochromocytoma and this triad of symptoms can occur in patients who do not have it. Thus, clinicians using the representativeness heuristic may overestimate the likelihood of a particular disease based on its representativeness by failing to recognize the low underlying prevalence (i.e., the prior, or pretest, probabilities). Conversely, atypical presentations of common diseases may lead to underestimating the likelihood of a particular disease. Thus, inexperience with a specific disease and with the breadth of its presentations may also lead to diagnostic delays or errors, e.g., diseases that affect multiple organ systems, such as sarcoid or tuberculosis, may be particularly challenging to diagnose because of the many different patterns they may manifest.

A second commonly used cognitive shortcut, the availability heuristic, involves judgments based on how easily prior similar cases or outcomes can be brought to mind. For example, a clinician may recall a case from a morbidity and mortality conference in which an elderly patient presented with painless dyspnea of acute onset and was evaluated for a pulmonary cause, but eventually found to have acute MI with the diagnostic delay likely contributing to the development of ischemic cardiomyopathy. If the case was associated with a malpractice accusation, such examples may be even more memorable. Errors with the availability heuristic arise from several sources of recall bias. Rare catastrophes are likely to be remembered with a clarity and force disproportionate to their likelihood for future diagnosis—for example, a patient with a sore throat eventually found to have leukemia or a young athlete with leg pain subsequently found to have a sarcoma—and those publicized in the media or recent experience are, of course, easier to recall and therefore more influential on clinical judgments.

The third commonly used cognitive shortcut, the anchoring heuristic (also called conservatism or stickiness), involves insufficiently adjusting the initial probability of disease up (or down) following a positive (or negative test) when compared with Bayes' theorem, i.e., sticking to the initial diagnosis. For example, a clinician may still judge the probability of coronary artery disease (CAD) to be high despite a negative exercise perfusion test and go on to cardiac catheterization (see "Measures of Disease Probability and Bayes' Rule," below).

The fourth heuristic states that clinicians should use the simplest explanation possible that will adequately account for the patient's symptoms and findings (Occam's razor or alternatively the simplicity heuristic). Although this is an attractive and often used principle, it is important to remember that no biologic basis for it exists. Errors from the simplicity heuristic include premature closure leading to the neglect of unexplained significant symptoms or findings.

For complex or unfamiliar diagnostic problems, clinicians typically resort to analytic reasoning processes (System 2) and proceed methodically using the *hypothetico-deductive model of reasoning*. Based on the stated reasons for seeking medical attention, clinicians develop an initial list of diagnostic possibilities in *hypothesis generation*. During the history of the present illness, the initial hypotheses evolve in *diagnostic refinement* as emerging information is tested against the mental models of the diseases being considered with diagnoses increasing and decreasing in likelihood or even being dropped from consideration as the working hypotheses of the moment. These mental models often generate additional questions that distinguish the diagnostic possibilities from one another. The focused physical examination contributes further distinguishing the working hypotheses. Is the spleen enlarged? How big is the liver? Is it tender? Are there any palpable masses or nodules? *Diagnostic verification* involves testing the adequacy (whether the diagnosis accounts for all symptoms and signs) and coherency (whether the signs and symptoms are consistent with the underlying pathophysiological causal mechanism) of the diagnosis. For example, if the enlarged and quite tender liver felt on physical examination is due to acute hepatitis (the hypothesis), then certain specific liver function tests will be markedly elevated (the prediction). Should the tests come back normal, the hypothesis may have to be discarded or substantially modified.

Although often neglected, negative findings are as important as positive ones because they reduce the likelihood of the diagnostic hypotheses under consideration. Chest discomfort that is not provoked or worsened by exertion and not relieved by rest in an active patient reduces the likelihood that chronic ischemic heart disease is the underlying cause. The absence of a resting tachycardia and thyroid gland enlargement reduces the likelihood of hyperthyroidism in a patient with paroxysmal atrial fibrillation.

The acuity of a patient's illness may override considerations of prevalence and the other issues described above. "Diagnostic imperatives" recognize the significance of relatively rare but potentially catastrophic diagnoses if undiagnosed and untreated. For example, clinicians should consider aortic dissection routinely as a possible cause of acute severe chest discomfort. Although the typical presenting symptoms of dissection differ from that of MI, dissection may mimic MI, and because it is far less prevalent and potentially fatal if mistreated, diagnosing dissection remains a challenging diagnostic imperative **(Chap. 274)**. Clinicians taking care of acute, severe chest pain patients should explicitly and routinely inquire about symptoms suggestive

of dissection, measure blood pressures in both arms for discrepancies, and examine for pulse deficits. When these are all negative, clinicians may feel sufficiently reassured to discard the aortic dissection hypothesis. If, however, the chest x-ray shows a possible widened mediastinum, the hypothesis should be reinstated and an appropriate imaging test ordered (e.g., thoracic computed tomography [CT] scan or transesophageal echocardiogram). In non-acute situations, the prevalence of potential alternative diagnoses should play a much more prominent role in diagnostic hypothesis generation.

Cognitive scientists studying the thought processes of expert clinicians have observed that clinicians group data into packets, or "chunks," that are stored in short-term or "working memory" and manipulated to generate diagnostic hypotheses. Because short-term memory is limited (classically humans can accurately repeat a list of 7±2 numbers read to them), the number of diagnoses that can be actively considered in hypothesis-generating activities is similarly limited. For this reason, cognitive shortcuts discussed above play a key role in the generation of diagnostic hypotheses, many of which are discarded as rapidly as they are formed, thereby demonstrating that the distinction between analytic and intuitive reasoning is an arbitrary and simplistic, but nonetheless useful, representation of cognition.

Research into the hypothetico-deductive model of reasoning has had difficulty identifying the elements of the reasoning process that distinguish experts from novices. This has led to a shift from examining the problem-solving process of experts to analyzing the organization of their knowledge for pattern matching as exemplars, prototypes, and illness scripts. For example, diagnosis may be based on the resemblance of a new case to patients seen previously (exemplars). As abstract mental models of disease, prototypes incorporate the likelihood of various disease features. Illness scripts include risk factors, pathophysiology, and symptoms and signs. Experts have a much larger store of exemplar and prototype cases, an example of which is the visual long-term memory of experienced radiologists. However, clinicians do not simply rely on literal recall of specific cases but have constructed elaborate conceptual networks of memorized information or models of disease to aid in arriving at their conclusions (illness scripts). That is, expertise involves an enhanced ability to connect symptoms, signs, and risk factors to one another in meaningful ways; relate those findings to possible diagnoses; and identify the additional information necessary to confirm the diagnosis.

No single theory accounts for all the key features of expertise in medical diagnosis. Experts have more knowledge about presenting symptoms of diseases and a larger repertoire of cognitive tools to employ in problem solving than non-experts. One definition of expertise highlights the ability to make powerful distinctions. In this sense, expertise involves a working knowledge of the diagnostic possibilities and those features that distinguish one disease from another. Memorization alone is insufficient, e.g., photographic memory of a medical textbook would not make one an expert. But having access to detailed case-specific relevant information is critically important. In the past, clinicians primarily acquired clinical knowledge through their patient experiences, but now clinicians have access to a plethora of information sources (see Evidence-Based Medicine [EBM] below). Clinicians of the future will be able to leverage the experiences of large numbers of other clinicians using electronic tools, but, as with the memorized textbook, the data alone will be insufficient for becoming an expert. Nonetheless, availability of these data removes one barrier for acquiring experience with connecting symptoms, signs, and risk factors to the possible diagnoses and

identifying the additional distinguishing information necessary to confirm the diagnosis, thereby potentially facilitating the development of the working knowledge necessary for becoming an expert.

Despite all of the research seeking to understand expertise in medicine and other disciplines, it remains uncertain whether any didactic program can actually accelerate the progression from novice to expert or from experienced clinician to master clinician. Deliberate effortful practice (over an extended period of time, sometimes said to be 10 years or 10,000 practice hours) and personal coaching are two strategies that are often used outside medicine (e.g., music, athletics, chess) to promote expertise. Their use in developing medical expertise and maintaining or enhancing it has not yet been adequately explored. Some studies in medicine suggest that didactic education exposing students to both the signs and symptoms of specific diseases and, in addition, the diseases that may present with specific symptoms and signs may be beneficial. Developing a personal learning system (e.g., metacognition) through for example EBM processes below and follow-up to identify diagnoses and treatments for patients that you have cared for provide active learning opportunities.

DIAGNOSTIC VERSUS THERAPEUTIC DECISION-MAKING

The modern ideal of medical therapeutic decision making is to “personalize” treatment recommendations. In the abstract, personalizing treatment involves combining the best available evidence about what works with an individual patient’s unique features (e.g., risk factors, genomics and co-morbidities) and his or her preferences and health goals to craft an optimal treatment recommendation with the patient. Operationally, two different and complementary levels of personalization are possible: individualizing the risk of harm and benefit for the options being considered based on the specific patient characteristics (precision medicine), and personalizing the therapeutic decision process by incorporating the patient’s preferences and values for the possible health outcomes. This latter process is sometimes referred to as shared decision-making, and typically involves clinicians sharing their knowledge about the options and the associated consequences and tradeoffs, and patients sharing their health goals, e.g., avoiding a short-term risk of dying from coronary artery bypass grafting to see their grandchild get married in a few months.

Individualizing the evidence about therapy **does not** mean relying on physician impressions of benefit and harm from their personal experience. Because of small sample sizes and rare events, the chance of drawing erroneous causal inferences from one’s own clinical experience is very high. For most chronic diseases, therapeutic effectiveness is only demonstrable statistically in large patient populations. It would be incorrect to infer with any certainty, for example, that treating a hypertensive patient with angiotensin-converting enzyme (ACE) inhibitors necessarily prevented a stroke from occurring during treatment, or that an untreated patient would definitely have avoided their stroke had they been treated. For many chronic diseases, a majority of patients will remain event free regardless of treatment choices; some will have events regardless of which treatment is selected; and those who avoided having an event through treatment cannot be individually identified. Blood pressure lowering, a readily observable surrogate endpoint, does not have a tightly coupled relationship with strokes prevented. Consequently, in most situations demonstrating

therapeutic effectiveness cannot rely simply on observing the outcome of an individual patient but should instead be based on large groups of patients carefully studied and properly analyzed.

Therapeutic decision-making, therefore, should be based on the best available evidence from clinical trials and well done outcome studies. Trustworthy clinical practice guidelines that synthesize such evidence offer normative guidance for many testing and treatment decisions. However, all guidelines recognize that “one size fits all” recommendations may not apply to individual patients. Increased research into the heterogeneity of treatment effects seeks to understand how best to adjust group level clinical evidence of treatment harms and benefits to account for the absolute level of risks faced by subgroups and even by individual patients, using, for example, validated clinical risk scores.

NON-CLINICAL INFLUENCES ON CLINICAL DECISION-MAKING

More than three decades of research on variations in clinician practice patterns has identified important non-clinical forces that shape clinical decisions. These factors can be grouped conceptually into three overlapping categories: (1) factors related to individual physicians practice, (2) factors related to practice setting, and (3) factors related to payment systems.

Factors Related to Practice Style

To ensure that necessary care is provided at a high level of quality, physicians fulfill a key role in medical care by serving as the patient’s advocate. Factors that influence performance in this role include the physician’s knowledge, training, and experience. Clearly, physicians cannot practice evidence-based medicine if they are unfamiliar with the evidence. As would be expected, specialists generally know the evidence in their field better than do generalists. Beyond published evidence and practice guidelines, a major set of influences on physician practice can be subsumed under the general concept of “practice style.” The practice style serves to define norms of clinical behavior. Beliefs about effectiveness of different therapies and preferred patterns of diagnostic test use are examples of different facets of a practice style. The physician beliefs that drive these different practice styles may be based on training, personal experience, and medical evidence. For example, in heart failure patients, heart failure specialists have more familiarity than general internists with the target doses of ACE inhibitor therapy as defined by large clinical trials and the specific drugs (including adverse effects), and are less likely to overreact to foreseeable problems in therapy such as a rise in creatinine levels or asymptomatic hypotension. Not surprisingly, the specialists are much more likely than generalists to achieve target doses of ACE inhibitor therapy. By contrast, perhaps due to specialization, cardiologists may overestimate the benefit and underestimate the harm of coronary revascularization relative to general internists.

Beyond the patient’s welfare, physician perceptions about the risk of a malpractice suit resulting from either an erroneous decision or a bad outcome may drive clinical decisions and create a practice referred to as defensive medicine. This practice involves using tests and therapies with very small marginal benefits, ostensibly to preclude future criticism should an adverse outcome occur. With conscious or unconscious

awareness of a connection to the risk of litigation or to payment, however, over time such patterns of care may become accepted as part of the practice norm, thereby perpetuating their overuse, e.g., annual cardiac exercise testing in asymptomatic patients.

Practice Setting Factors

Factors in this category relate to work systems including tasks and workflow (interruptions, inefficiencies, workload), technology (poor design or implementation, errors in use, failure, misuse), organizational characteristics (e.g., culture, leadership, staffing, scheduling), and the physical environment (e.g., noise, lighting, layout). Physician-induced demand is a term that refers to the repeated observation that once medical facilities and technologies become available to physicians, they will use them. Other environmental factors that can influence decision-making include the local availability of specialists for consultations and procedures; “high-tech” advanced imaging or procedure facilities such as MRI machines and proton beam therapy centers; and fragmentation of care.

Payment Systems

Economic incentives are closely related to the other two categories of practice-modifying factors. Financial issues can exert both stimulatory and inhibitory influences on clinical practice. Historically, physicians are paid on a fee-for-service, capitation, or salary basis. In fee-for-service, physicians who do more get paid more, thereby encouraging overuse, consciously or unconsciously. When fees are reduced (discounted reimbursement), clinicians tend to increase the number of services provided to maintain revenue. Capitation, in contrast, provides a fixed payment per patient per year to encourage physicians to consider a global population budget in managing individual patients and ideally reducing the use of interventions with small marginal benefit. To discourage volume-based excessive utilization, fixed salary compensation plans pay physicians the same regardless of the clinical effort expended, but may provide an incentive to see fewer patients. In recognition of the non-sustainability of continued growth in medical expenditures and the opportunity costs associated with that (funds that might be more beneficially applied to education, energy, social welfare or defense), current efforts seek to transition to a value-based payment system to reduce overuse and to reflect benefit. Work to define how to actually tie payment to value has mostly focused so far on “pay for performance” models. High quality clinical trial evidence for the effectiveness of these models is still mostly lacking.

INTERPRETATION OF DIAGNOSTIC TESTS

Despite impressive technological advances in medicine over the last century, uncertainty still abounds and challenges all aspects of medical decision-making. Compounding this challenge, massive information overload characterizes modern medicine. Clinicians on average subscribe to seven journals, presenting them with over 2500 new articles each year, and need access to 2 million pieces of information to practice medicine. Of course, to be useful, this information must be sifted for quality and examined for applicability for integration into patient-specific care. Although computers appear to offer an obvious solution both for

information management and for quantification of medical care uncertainties, many practical problems must be solved before computerized decision support can be routinely incorporated into the clinical reasoning process in a way that demonstrably improves the quality of care. For the present, understanding the nature of diagnostic test information can help clinicians become more efficient users of such data. The next section reviews concepts related to diagnostic testing.

DIAGNOSTIC TESTING: MEASURES OF TEST ACCURACY

The purpose of performing a test on a patient is to reduce uncertainty about the patient's diagnosis or prognosis in order to facilitate appropriate management. Although diagnostic tests commonly refer to laboratory (e.g., blood count) or imaging tests or procedures (e.g., colonoscopy or bronchoscopy), any information that changes a provider's understanding of the patient's problem qualifies as a diagnostic test. Thus, even the history and physical examination should be considered as diagnostic tests. In clinical medicine, it is common to reduce the results of a test to a dichotomous outcome, such as positive or negative, normal or abnormal. Although this simplification ignores useful information (such as the degree of abnormality), it facilitates illustrating some important principles of test interpretation which are described below.

The accuracy of any diagnostic test is assessed relative to a "gold standard," where a positive gold standard test defines the patients who have disease and a negative test rules out disease (Table 3-1). Characterizing the diagnostic performance of a new test requires identifying an appropriate population (ideally, patients representative of those in whom the new test would be used) and applying both the new and the gold standard tests to all subjects. Biased estimates of test performance occur when diagnostic accuracy is defined using an inappropriate population or one in which gold standard determination of disease status is incomplete. The accuracy of the new test in distinguishing disease from health is determined relative to the gold standard results and summarized in four estimates. The sensitivity or true-positive rate of the new test reflects how well the new test identifies patients with disease. It is the proportion of patients with disease (defined by the gold standard) who have a positive test. The proportion of patients with disease who have a negative test is the false-negative rate, calculated as 1 – sensitivity. The specificity, or true-negative rate reflects how well the new test correctly identifies patients without disease. It is the proportion of patients without disease (defined by the gold standard) who have a negative test. The proportion of patients without disease who have positive test is the false-positive rate, calculated as 1 – specificity. In theory, a perfect test would be one with a sensitivity of 100% and a specificity of 100% and would completely distinguish patients with disease from those without it. A useful mnemonic is the following: a *negative* high sensitivity (*Sn*) test helps rule *out* disease (Negative SnOut), and a *positive* high specificity (*Sp*) test helps rule *in* disease (Positive SpIn).

TABLE 3-1

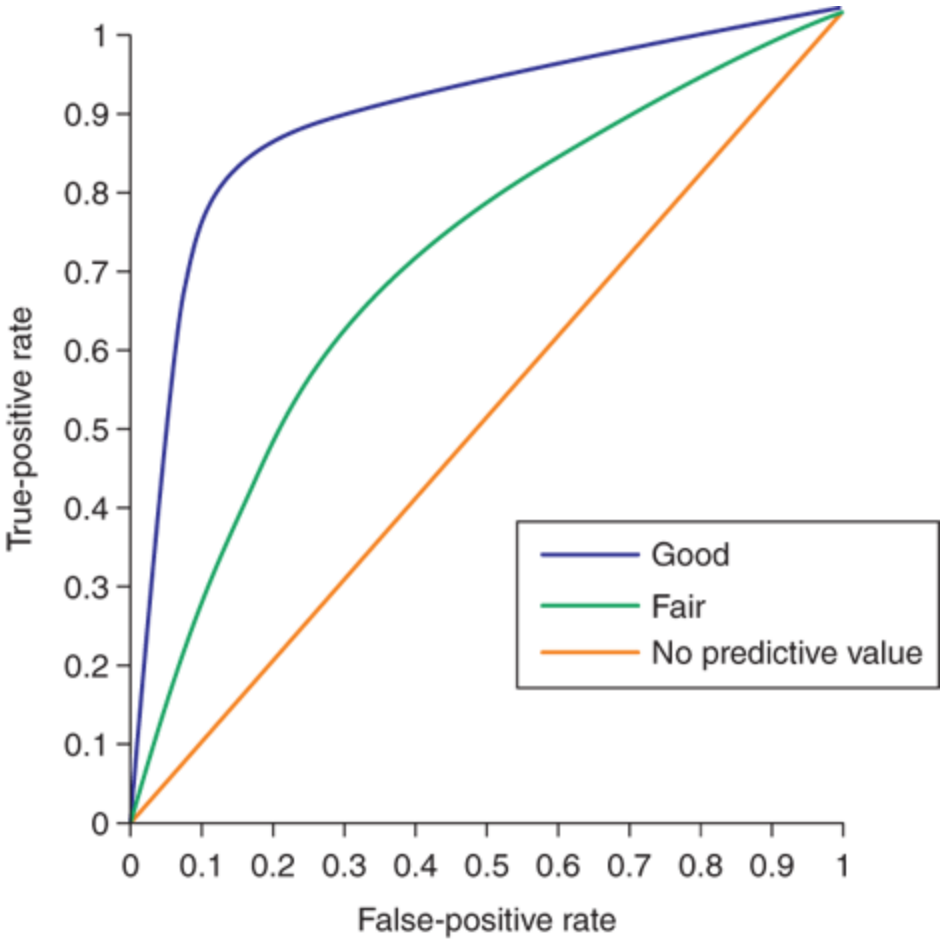
Measures of Diagnostic Test Accuracy

Test Result	Disease Status	
	Present	**Absent**
Positive	True positives (TP)	False positives (FP)
Negative	False negatives (FN)	True Negatives (TN)
Test Characteristics in Patients with Disease		
True-positive rate (sensitivity) = TP/(TP + FN)		
False-negative rate = FN/(TP + FN) = 1 – true positive rate		
Test Characteristics in Patients without Disease		
True-negative rate (specificity) = TN/(TN + FP)		
False-positive rate = FP/(TN + FP) = 1 – true-negative rate		

Calculating sensitivity and specificity requires selection of a threshold value or cut point above which the test is considered "positive." Making the cut point "stricter" (e.g., raising it) lowers sensitivity but improves specificity, while making it "laxer" (e.g., lowering it) raises sensitivity but lowers specificity. This dynamic trade-off between more accurate identification of subjects with disease versus those without disease is often displayed graphically as a receiver operating characteristic (ROC) curve **(Fig. 3-1)** by plotting sensitivity (*y* axis) versus 1 – specificity (*x* axis). Each point on the curve represents a potential cut point with an associated sensitivity and specificity value. The area under the ROC curve often is used as a quantitative measure of the information content of a test. Values range from 0.5 (no diagnostic information from testing at all; the test is equivalent to flipping a coin) to 1.0 (perfect test). The choice of cut point should in theory reflect the relative harms and benefits of treatment for those without versus those with disease. For example, if treatment was safe with substantial benefit, then choosing a high sensitivity cut point (upper right of the ROC curve) for a low risk test may be appropriate (e.g., phenylketonuria in newborns), but if treatment had substantial risk for harm, then choosing a high specificity cut point (lower left of the ROC curve) may be appropriate (e.g., chemotherapy for cancer). The choice of cut point may also depend on the likelihood of disease with low likelihoods placing a greater emphasis on the harms of false positive tests (e.g., HIV testing in marriage applicants) or the harms of false-negative tests (e.g., HIV testing in blood donors).

Figure 3-1

Each receiver operating characteristic (ROC curve) illustrates a trade-off that occurs between improved test sensitivity (accurate detection of patients with disease) and improved test specificity (accurate detection of patients without disease), as the test value defining when the test turns from “negative” to “positive” is varied. A 45° line would indicate a test with no predictive value (sensitivity = specificity at every test value). The area under each ROC curve is a measure of the information content of the test. Thus, a larger ROC area signifies increased diagnostic accuracy.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MEASURES OF DISEASE PROBABILITY AND BAYES’ RULE

In the absence of perfect tests, the true disease state of the patient remains uncertain after every test. Bayes’ rule provides a way to quantify the revised uncertainty using simple probability mathematics (and thereby avoid anchoring bias). It calculates the *posttest probability* or likelihood of disease after a test result, from three parameters: the pretest probability of disease, the test sensitivity, and the test specificity. The *pretest probability* is a quantitative estimate of the likelihood of the diagnosis before the test is performed and is usually estimated from the prevalence of the disease in the underlying population (if known) or clinical context (e.g., age, sex and type of chest pain). For some common conditions, such as CAD, existing nomograms and statistical models generate estimates of pretest probability that account for history, physical examination, and test findings. The posttest probability (also called the predictive value of the test, see

below) is a recalibrated statement of the likelihood of the diagnosis, accounting for both pretest probability and test results. For the likelihood of disease following a positive test (i.e., positive predictive value), Bayes' rule is calculated as:

$$\text{Posttest probability} = \frac{\text{Pretest probability} \times \text{test sensitivity}}{\text{Pretest probability} \times \text{test sensitivity} + (1 - \text{Pretest probability}) \times \text{test false-positive rate}}$$

For example, consider a 64-year-old woman with atypical chest pain who has a pretest probability of 0.50 and a "positive" diagnostic test result (assuming test sensitivity = 0.90 and specificity = 0.90).

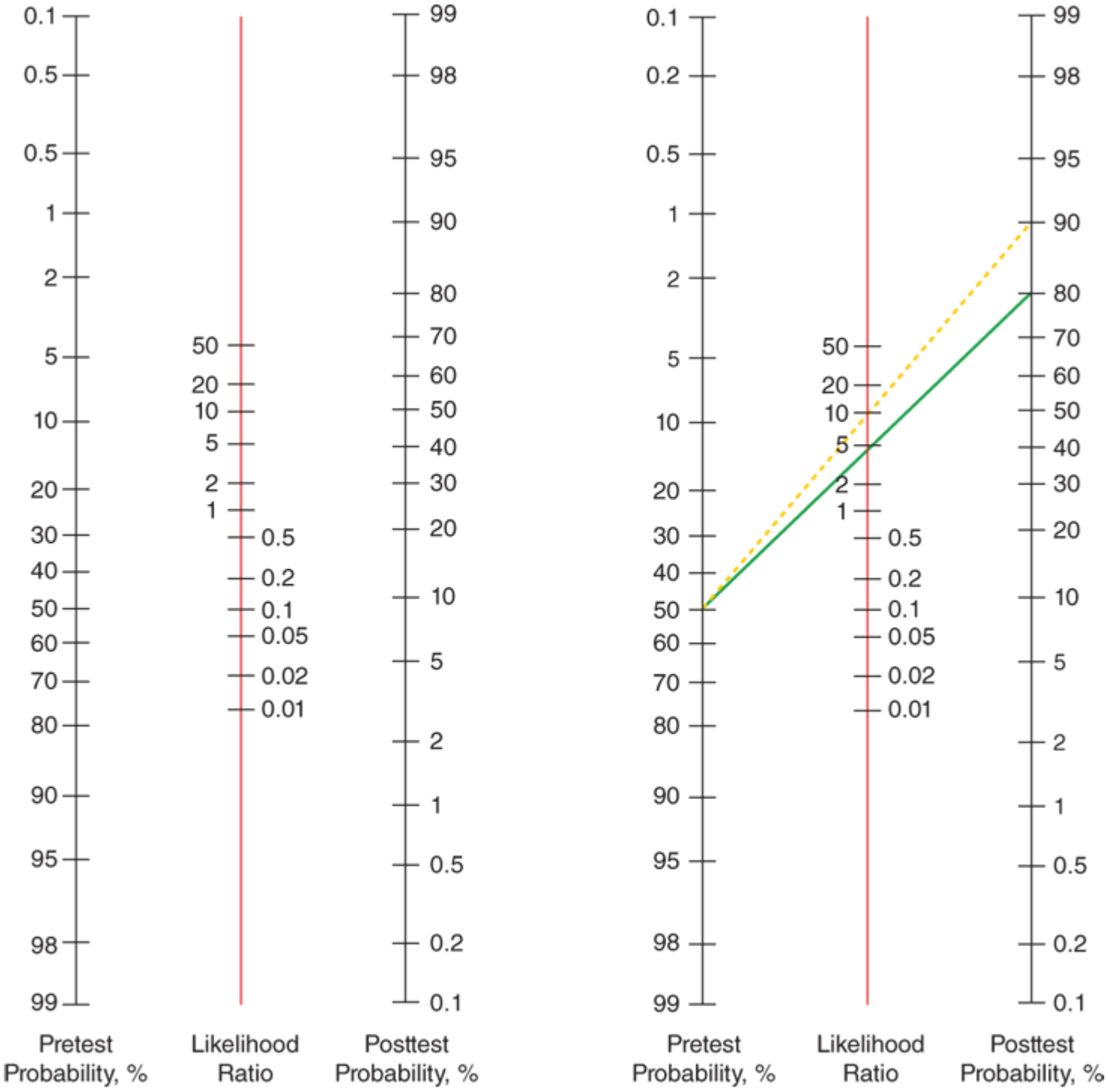
$$\text{Posttest probability} = \frac{(0.50)(0.90)}{(0.50)(0.90) + (0.50)(0.10)} = 0.90$$

The term predictive value has often been used as a synonym for the posttest probability. Unfortunately, clinicians commonly misinterpret reported predictive values as intrinsic measures of test accuracy rather than calculated probabilities. Studies of diagnostic test performance compound the confusion by calculating predictive values from the same sample used to measure sensitivity and specificity. Such calculations are misleading unless the test is applied subsequently to populations with exactly the same disease prevalence. For these reasons, the term predictive value is best avoided in favor of the more descriptive posttest probability following a positive or a negative test result.

The nomogram version of Bayes' rule **(Fig. 3-2)** helps us to understand at a conceptual level how it estimates the posttest probability of disease. In this nomogram, the impact of the diagnostic test result is summarized by the likelihood ratio, which is defined as the ratio of the probability of a given test result (e.g., "positive" or "negative") in a patient with disease to the probability of that result in a patient without disease, thereby providing a measure of how well the test distinguishes those with from those without disease.

Figure 3-2

Nomogram version of Bayes' theorem used to predict the posttest probability of disease (right-hand scale) using the pretest probability of disease (left-hand scale) and the likelihood ratio for a positive test (middle scale). See text for information on calculation of likelihood ratios. To use, place a straightedge connecting the pretest probability and the likelihood ratio and read off the posttest probability. The right-hand part of the figure illustrates the value of a positive exercise treadmill test (likelihood ratio 4, *green line*) and a positive exercise thallium single-photon emission CT perfusion study (likelihood ratio 9, broken *yellow line*) in a patient with a pretest probability of coronary artery disease of 50%. *(Adapted from Centre for Evidence-Based Medicine: Likelihood ratios. Available at http://www.cebm.net/likelihood-ratios/.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The *likelihood ratio for a positive test* is calculated as the ratio of the true-positive rate to the false-positive rate (or sensitivity/[1 – specificity]). For example, a test with a sensitivity of 0.90 and a specificity of 0.90 has a likelihood ratio of 0.90/(1 – 0.90), or 9. Thus, for this hypothetical test, a "positive" result is 9 times more likely in a patient with the disease than in a patient without it. Most tests in medicine have likelihood ratios for a positive result between 1.5 and 20. Higher values are associated with tests that more substantially increase the posttest likelihood of disease. A very high likelihood ratio positive (>10) usually implies high specificity, so a positive high specificity test helps "rule in" disease. If sensitivity is excellent but specificity is less so, the

likelihood ratio will be reduced substantially (e.g., with a 90% sensitivity but a 55% specificity, the likelihood ratio positive is 2.0).

The corresponding *likelihood ratio for a negative test* is the ratio of the false-negative rate to the true-negative rate (or [1 – sensitivity]/specificity). Lower likelihood ratio negative values more substantially lower the posttest likelihood of disease. A very low likelihood ratio negative (falling below 0.10) usually implies high sensitivity, so a negative high sensitivity test helps "rule out" disease. The hypothetical test considered above with a sensitivity of 0.9 and a specificity of 0.9 would have a likelihood ratio for a negative test result of (1 – 0.9)/0.9, or 0.11, meaning that a negative result is about one-tenth as likely in patients with disease than in those without disease (or about ten times more likely in those without disease than in those with disease).

APPLICATIONS TO DIAGNOSTIC TESTING IN CAD

Consider two tests commonly used in the diagnosis of CAD: an exercise treadmill and an exercise single-photon emission CT (SPECT) myocardial perfusion imaging test **(Chap. 236)**. Meta-analysis has shown that a positive treadmill ST-segment response has an average sensitivity of 60% and an average specificity of 75%, yielding a likelihood ratio positive of 2.4 (0.60/[1 – 0.75]) (consistent with modest discriminatory ability because it falls between 2 and 5). For a 41-year-old man with nonanginal pain and a 10% pretest probability of CAD, the posttest probability of disease after a positive result rises to only about 30%. For a 60-year-old woman with typical angina and a pretest probability of CAD of 80%, a positive test result raises the posttest probability of disease to about 95%.

In contrast, exercise SPECT myocardial perfusion test is more accurate for diagnosis of CAD. For simplicity, assume that the finding of a reversible exercise-induced perfusion defect has both a sensitivity and a specificity of 90% (a bit higher than reported), yielding a likelihood ratio for a positive test of 9.0 (0.90/[1 – 0.90]) (consistent with intermediate discriminatory ability because it falls between 5 and 10). For the same 10% pretest probability patient, a positive test raises the probability of CAD to 50% (Fig. 3-2). However, despite the differences in posttest probabilities between these two tests (30 versus 50%), the more accurate test may not improve diagnostic likelihood enough to change patient management (e.g., decision to refer to cardiac catheterization) because the more accurate test has only moved the physician from being fairly certain that the patient did not have CAD to a 50:50 chance of disease. In a patient with a pretest probability of 80%, exercise SPECT test raises the posttest probability to 97% (compared with 95% for the exercise treadmill). Again, the more accurate test does not provide enough improvement in posttest confidence to alter management, and neither test has improved much on what was known from clinical data alone.

In general, positive results with an accurate test (e.g., likelihood ratio positive 10) when the pretest probability is low (e.g., 20%) do not move the posttest probability to a range high enough to rule in disease (e.g., 80%). In screening situations, pretest probabilities are often particularly low because patients are asymptomatic. In such cases, specificity becomes particularly important. For example, in screening first-time female blood donors without risk factors for HIV, a positive test raised the likelihood of HIV to only 67% despite a specificity of 99.995% because the prevalence was 0.01%. Conversely, with a high pretest

probability, a negative test may not rule out disease adequately if it is not sufficiently sensitive. Thus, the largest change in diagnostic likelihood following a test result occurs when the clinician is most uncertain (i.e., pretest probability between 30 and 70%). For example, if a patient has a pretest probability for CAD of 50%, a positive exercise treadmill test will move the posttest probability to 80% and a positive exercise SPECT perfusion test will move it to 90% (Fig. 3-2).

As presented above, Bayes' rule employs a number of important simplifications that should be considered. First, few tests provide only "positive" or "negative" results. Many tests have multi-dimensional outcomes (e.g., extent of ST-segment depression, exercise duration, and exercise-induced symptoms with exercise testing). Although Bayes' theorem can be adapted to this more detailed test result format, it is computationally more complex to do so. Similarly, when multiple sequential tests are performed, the posttest probability may be used as the pretest probability to interpret the second test. However, this simplification assumes conditional independence—that is, that the results of the first test do not affect the likelihood of the second test result—and this is often not true.

Finally, many texts assert that sensitivity and specificity are prevalence-independent parameters of test accuracy. This statistically useful assumption, however, is clinically simplistic. A treadmill exercise test, for example, has a sensitivity of ~30% in a population of patients with 1-vessel CAD, whereas its sensitivity in patients with severe 3-vessel CAD approaches 80%. Thus, the best estimate of sensitivity to use in a particular decision may vary, depending on the severity of disease in the local population. A hospitalized, symptomatic, or referral population typically has a higher prevalence of disease and, in particular, a higher prevalence of more advanced disease than does an outpatient population. Consequently, test sensitivity will likely be higher in hospitalized patients, and test specificity higher in outpatients.

STATISTICAL PREDICTION MODELS

Bayes' rule, when used as presented above, is useful in studying diagnostic testing concepts but may prove too simplistic for use in actual patient care decisions. Predictions based on multivariable statistical models can more accurately address these more complex problems by simultaneously accounting for additional relevant patient characteristics. In particular, these models explicitly account for multiple, even possibly overlapping, pieces of patient-specific information and assign a relative weight to each on the basis of its unique independent contribution to the prediction in question. For example, a logistic regression model to predict the probability of CAD ideally considers all the relevant independent factors from the clinical examination and diagnostic testing and their relative importance instead of the limited data that clinicians can manage in their heads or with Bayes' rule. However, despite this strength, prediction models are usually too complex computationally to use without a calculator or computer. Guideline-driven treatment recommendations based on statistical prediction models available online, e.g., the ACC/AHA risk calculator for primary prevention with statins and the CHA_2DS_2-VASC calculator for anticoagulation for atrial fibrillation have generated more widespread usage. Whether the adoption of electronic health records will promote more use of predictive models in clinical practice and increase their impact on clinical encounters and outcomes remains unclear.

One reason for limited clinical use is that, to date, only a handful of prediction models have been validated properly (for example, Wells' criteria for pulmonary embolism, **see Table 3-2**). The importance of independent validation in a population separate from the one used to develop the model cannot be overstated. An unvalidated prediction model should be viewed with the skepticism appropriate for any new drug or medical device that has not had rigorous clinical trial testing.

TABLE 3-2

Wells Clinical Prediction Rule for Pulmonary Embolism

Clinical Feature	**Points**
Clinical signs of deep-vein thrombosis	3
Alternative diagnosis is less likely than PE	3
Heart rate >100 beats per min	1.5
Immobilization ≥3 d or surgery in previous 4 weeks	1.5
History of deep-vein thrombosis or pulmonary embolism	1.5
Hemoptysis	1
Malignancy (with treatment within 6 months) or palliative	1
Interpretation	
Score >6.0	High
Score 2.0–6.0	Intermediate
Score <2.0	Low

When statistical survival models in cancer and heart disease have been compared directly with clinicians' predictions, the survival models have been found to be more consistent, as would be expected but not always more accurate. On the other hand, comparison of clinicians with websites and apps that generate lists of possible diagnoses to help patients with self-diagnosis found that physicians outperformed the currently available programs. For students and less-experienced clinicians, the biggest value of diagnostic decision support may be in extending diagnostic possibilities and triggering "rational override" but their impact on knowledge, information-seeking, and problem-solving needs additional research.

FORMAL DECISION SUPPORT TOOLS

DECISION SUPPORT SYSTEMS

Over the last 40 years, many attempts have been made to develop computer systems to aid clinical decision-making and patient management. Conceptually, computers offer several levels of potentially useful support for clinicians. At the most basic level, they provide ready access to vast reservoirs of information, which may, however, be quite difficult to sort through to find what is needed. At higher levels, computers can support care management decisions by making accurate predictions of outcome, or can simulate the whole decision process, and provide algorithmic guidance. Computer-based predictions using Bayesian or statistical regression models inform a clinical decision but do not actually reach a “conclusion” or “recommendation.” Machine learning methods are being applied to pattern recognition tasks such as the examination of skin lesions and the interpretation of x-rays. Artificial intelligence systems attempt to simulate or replace human reasoning with a computer-based analogue. To date, such approaches have achieved only limited success. Reminder or protocol-directed systems do not make predictions but use existing algorithms, such as guidelines or appropriate utilization criteria, to direct clinical practice. In general, however, decision support systems have had little impact on practice. Reminder systems built into electronic health records have shown the most promise, particularly in correcting drug dosing and promoting adherence to guidelines. Checklists may also help avoid or reduce errors.

DECISION ANALYSIS

Compared with the decision support methods above, decision analysis represents a normative prescriptive approach to decision-making in the face of uncertainty. Its principal application is in complex decisions. For example, public health policy decisions often involve *trade-offs* in length versus quality of life, benefits versus resource use, population versus individual health, and *uncertainty* regarding efficacy, effectiveness, and adverse events as well as *values* or preferences regarding mortality and morbidity outcomes.

One recent analysis using this approach involved the optimal screening strategy for breast cancer, which has remained controversial, in part because a randomized controlled trial to determine when to begin screening and how often to repeat screening mammography is impractical. In 2016, the National Cancer Institute sponsored Cancer Intervention and Surveillance Network (CISNET) examined eight strategies differing by whether to initiate mammography screening at age 40, 45, or 50 years and whether to screen annually, biennially, or annually for women in their forties and biennially thereafter (hybrid). The six simulation models found biennial strategies to be the most efficient for average-risk women. Biennial screening for 1000 women from age 50 to 74 years versus no screening avoided seven breast cancer deaths. Screening annually from age 40 to 74 years avoided three additional deaths but required 20,000 additional mammograms and yielded 1988 more false-positive results. Factors that influenced the results included patients with a 2–4-fold higher risk for developing breast cancer in whom annual screening from 40 to 74 yielded similar benefits as biennial screening from age 50 to 74. For average-risk patients with moderate or severe co-morbidities, screening could be stopped earlier at ages 66–68 years.

This analysis involved six models that reproduced epidemiologic trends and a screening trial result, accounted for digital technology and treatments advances, and considered quality of life, risk factors, breast density, and comorbidity. It provided novel insights into a public health problem in the absence of a randomized clinical trial and helped weigh the pros and cons of such a health policy recommendation. Although such models have been developed for selected clinical problems, their benefit and application to individual real-time clinical management has yet to be demonstrated.

DIAGNOSIS AS AN ELEMENT OF QUALITY OF CARE

High quality medical care begins with accurate diagnosis. The incidence of diagnostic errors has been estimated by a variety of methods including postmortem examinations, medical record reviews, and medical malpractice claims, with each yielding complementary but different estimates of this quality of care patient-safety problem. In the past, diagnostic errors tended to be viewed as a failure of individual clinicians. The modern view is that they are mostly system of care deficiencies. Current estimates suggest that nearly everyone will experience at least one diagnostic error in their lifetime, leading to mortality, morbidity, unnecessary tests and procedures, costs, and anxiety.

Solutions to the “diagnostic errors as a system of care problem” have focused on system-level approaches, such as decision support and other tools integrated into electronic medical records. The use of checklists has been proposed as a means of reducing some of the cognitive errors discussed earlier in the chapter, such as premature closure. While checklists have been shown useful in certain medical contexts, such as the ORs and ICUs, their value in preventing diagnostic errors that lead to patient adverse events remains to be shown.

EVIDENCE-BASED MEDICINE

Clinical medicine is defined traditionally as a practice combining medical knowledge (including scientific evidence), intuition, and judgment in the care of patients **(Chap. 1)**. Evidence-based medicine (EBM) updates this construct by placing much greater emphasis on the processes by which clinicians gain knowledge of the most up-to-date and relevant clinical research to determine for themselves whether medical interventions alter the disease course and improve the length or quality of life. The meaning of practicing EBM becomes clearer through an examination of its four key steps:

1. Formulating the management question to be answered
2. Searching the literature and online databases for applicable research data
3. Appraising the evidence gathered with regard to its validity and relevance
4. Integrating this appraisal with knowledge about the unique aspects of the patient (including the patient’s preferences about the possible outcomes)

The process of searching the world's research literature and appraising the quality and relevance of studies can be time-consuming and requires skills and training that most clinicians do not possess. Thus, identifying recent systematic overviews of the problem in question (Table 3-3) may offer the best starting point for most EBM searches. However, the medical literature is now being flooded with systematic reviews of varying quality and clinical utility. Therefore, systematic reviews should be used in conjunction with selective reading of some of the best empirical studies.

TABLE 3-3

Selected Tools for Finding the Evidence in Evidence-Based Medicine

Name	Description	Web Address	Availability
Evidence-Based Medicine Reviews	Comprehensive electronic database that combines and integrates: 1. The Cochrane Database of Systematic Reviews 2. ACP Journal Club 3. The Database of Abstracts of Reviews of Effectiveness	*www.ovid.com*	Subscription required. Available through medical center libraries and other institutions.
Cochrane Library	Collection of EBM databases, including the Cochrane Database of Systematic Reviews—full text articles reviewing specific health care topics	*www.cochrane.org*	Subscription required. Abstracts of systematic reviews available free online. Some countries have funding to provide free access to all residents.
ACP Journal Club	Collection of summaries of original studies and systematic reviews. Published bimonthly. All data since 1991 available on website, updated yearly.	*www.acpjc.org*	Subscription required.
Clinical Evidence	Monthly updated directory of concise overviews of common clinical interventions.	*www.clinicalevidence.com*	Subscription required. Free access for United Kingdom and developing countries.
MEDLINE	National Library of Medicine database with citations back to 1966.	*www.nlm.nih.gov*	Free via Internet.

Generally, the EBM tools listed in Table 3-3 provide access to research information in one of two forms. The first, primary research reports, is the original peer-reviewed research work that is published in medical

journals and accessible through MEDLINE in abstract form. However, without training in using MEDLINE, locating reports quickly and efficiently that are on point in a sea of irrelevant or unhelpful citations remains difficult, and important studies are easily missed. Systematic reviews, the second form, are regarded by some as the highest level of evidence in the hierarchy because they are intended to comprehensively summarize the available evidence on a particular topic. To avoid the potential biases found in review articles, predefined reproducible explicit search strategies and inclusion and exclusion criteria seek to find all of the relevant scientific research and grade its quality. The prototype for this kind of resource is the Cochrane Database of Systematic Reviews. When appropriate, a meta-analysis is used to quantitatively summarize the systematic review findings. The next two sections explicate the major types of clinical research reports available in the literature and the process of aggregating those data into meta-analyses.

SOURCES OF EVIDENCE: CLINICAL TRIALS AND REGISTRIES

The notion of learning from observation of patients is as old as medicine itself. Over the last 50 years, physicians' understanding of how best to turn raw observation into useful evidence has evolved considerably. Case reports, personal anecdotal experience, and small single-center case series are now recognized as having severe limitations in validity and generalizability, and although they may generate hypotheses or be the first reports of adverse events or therapeutic benefit, they have no role in formulating modern standards of practice. The major tools used to develop reliable evidence consist of the randomized clinical trial and the large observational registry. A registry or database typically is focused on a disease or syndrome (e.g., different types of cancer, acute or chronic CAD, pacemaker capture or chronic heart failure), a clinical procedure (e.g., bone marrow transplantation, coronary revascularization), or an administrative process (e.g., claims data used for billing and reimbursement).

By definition, in observational data, the investigator does not control patient care. Carefully collected prospective observational data, however, can at times achieve a level of evidence quality approaching that of major clinical trial data. At the other end of the spectrum, data collected retrospectively (e.g., chart review) are limited in form and content to what previous observers recorded and may not include the specific research data being sought (e.g., claims data). Advantages of observational data include the inclusion of a broader population as encountered in practice than is typically represented in clinical trials because of their restrictive inclusion and exclusion criteria. In addition, observational data provide primary evidence for research questions when a randomized trial cannot be performed. For example, it would be difficult to randomize patients to test diagnostic or therapeutic strategies that are unproven but widely accepted in practice, and it would be unethical to randomize based on sex, racial/ethnic group, socioeconomic status, or country of residence or to randomize patients to a potentially harmful intervention, such as smoking or deliberately overeating to develop obesity.

A well-done prospective observational study of a particular management strategy differs from a well-done randomized clinical trial most importantly by its lack of protection from treatment selection bias. The use of observational data to compare diagnostic or therapeutic strategies assumes that sufficient uncertainty and heterogeneity exists in clinical practice to ensure that similar patients will be managed differently by diverse

physicians. In short, the analysis assumes that a sufficient element of randomness (in the sense of disorder rather than in the formal statistical sense) exists in clinical management. In such cases, statistical models attempt to adjust for important imbalances to "level the playing field" so that a fair comparison among treatment options can be made. When management is clearly not random (e.g., all eligible left main CAD patients are referred for coronary bypass surgery), the problem may be too confounded (biased) for statistical correction, and observational data may not provide reliable evidence.

In general, the use of concurrent controls is vastly preferable to that of historical controls. For example, comparison of current surgical management of left main CAD with medically treated patients with left main CAD during the 1970s (the last time these patients were routinely treated with medicine alone) would be extremely misleading because "medical therapy" has substantially improved in the interim.

Randomized controlled clinical trials include the careful prospective design features of the best observational data studies but also include the use of random allocation of treatment. This design provides the best protection against measured and unmeasured confounding due to treatment selection bias (a major aspect of internal validity). However, the randomized trial may not have good external validity (generalizability) if the process of recruitment into the trial resulted in the exclusion of many potentially eligible subjects or if the nominal eligibility for the trial describe a very heterogeneous population.

Consumers of medical evidence need to be aware that randomized trials vary widely in their quality and applicability to practice. The process of designing such a trial often involves many compromises. For example, trials designed to gain U.S. Food and Drug Administration (FDA) approval for an investigational drug or device must fulfill regulatory requirements (such as the use of a placebo control) that may result in a trial population and design that differs substantially from what practicing clinicians would find most useful.

META-ANALYSIS

The Greek prefix *meta* signifies something at a later or higher stage of development. Meta-analysis is research that combines and summarizes the available evidence quantitatively. Although it is used to examine nonrandomized studies, meta-analysis is most useful for summarizing all randomized trials examining a particular therapy. Ideally, unpublished trials should be identified and included to avoid publication bias (i.e., missing "negative" trials which may not be published). Furthermore, the best meta-analyses obtain and analyze individual patient-level data from all trials rather than using only the summary data from published reports. Nonetheless, not all published meta-analyses yield reliable evidence for a particular problem, so their methodology should be scrutinized carefully to ensure proper study design and analysis. The results of a well-done meta-analysis are likely to be most persuasive if they include at least several large-scale, properly performed randomized trials. Meta-analysis can especially help detect benefits when individual trials are inadequately powered (e.g., the benefits of streptokinase thrombolytic therapy in acute MI demonstrated by ISIS-2 in 1988 were evident by the early 1970s through meta-analysis). However, in cases in which the available trials are small or poorly done, meta-analysis should not be viewed as a remedy for deficiencies in primary trial data or trial design.

Meta-analyses typically focus on summary measures of relative treatment benefit, such as odds ratios or relative risks. Clinicians also should examine what absolute risk reduction (ARR) can be expected from the therapy. A summary metric of absolute treatment benefit is the number needed to treat (NNT) to prevent one adverse outcome event (e.g., death, stroke). NNT is simply 1/ARR. For example, if a hypothetical therapy reduced mortality rates over a 5-year follow-up by 33% (the relative treatment benefit) from 12% (control arm) to 8% (treatment arm), the absolute risk reduction would be 12% – 8% = 4% and the NNT would be 1/.04, or 25. Thus, it would be necessary to treat 25 patients for 5 years to prevent 1 death. If the hypothetical treatment was applied to a lower-risk population, say, with a 6% 5-year mortality, the 33% relative treatment benefit would reduce absolute mortality by 2% (from 6 to 4%), and the NNT for the same therapy in this lower-risk group of patients would be 50. Although not always made explicit, comparisons of NNT estimates from different studies should account for the duration of follow-up used to create each estimate. In addition, the NNT concept assumes a homogeneity in response to treatment that may not be accurate. The NNT is simply another way of summarizing the absolute treatment difference and does not provide any unique information.

CLINICAL PRACTICE GUIDELINES

According to the 1990 Institute of Medicine definition, clinical practice guidelines are “systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances.” This definition emphasizes several crucial features of modern guideline development. First, guidelines are created by using the tools of EBM. In particular, the core of the development process is a systematic literature search followed by a review of the relevant peer-reviewed literature. Second, guidelines usually are focused on a clinical disorder (e.g., diabetes mellitus, stable angina pectoris) or a health care intervention (e.g., cancer screening). Third, the primary objective of guidelines is to improve the quality of medical care by identifying care practices which should be routinely implemented, based on high quality evidence and high benefit to harm ratios for the interventions. Guidelines are intended to “assist” decision-making, not to define explicitly what decisions should be made in a particular situation, in part because guideline level evidence alone is never sufficient for clinical decision-making (e.g., deciding whether to intubate and administer antibiotics for pneumonia in a terminally ill individual, in an individual with dementia, or in an otherwise healthy 30-year-old mother).

Guidelines are narrative documents constructed by expert panels whose composition often is determined by interested professional organizations. These panels vary in expertise and in the degree to which they represent all relevant stakeholders. The guideline documents consist of a series of specific management recommendations, a summary indication of the quantity and quality of evidence supporting each recommendation, an assessment of the benefit to harm ratio for the recommendation, and a narrative discussion of the recommendations. Many recommendations simply reflect the expert consensus of the guideline panel because literature-based evidence is insufficient or absent. The final step in guideline construction is peer review, followed by a final revision in response to the critiques provided. To improve the reliability and trustworthiness of guidelines, the National Academy of Medicine (formerly Institute of Medicine) has made methodological recommendations for guideline development.

Guidelines are closely tied to the process of quality improvement in medicine through their identification of evidence-based best practices. Such practices can be used as quality indicators. Examples include the proportion of acute MI patients who receive aspirin upon admission to a hospital and the proportion of heart failure patients with a depressed ejection fraction treated with an ACE inhibitor.

CONCLUSIONS

In this era of EBM, it is tempting to think that all the difficult decisions practitioners face have been or soon will be solved and digested into practice guidelines and computerized reminders. However, EBM provides practitioners with an ideal rather than a finished set of tools with which to manage patients. Moreover, even with such evidence, it is always worth remembering that the response to therapy of the “average” patient represented by the summary clinical trial outcomes may not be what can be expected for the specific patient sitting in front of a provider in the clinic or hospital. In addition, meta-analyses cannot generate evidence when there are no adequate randomized trials, and most of what clinicians confront in practice will never be thoroughly tested in a randomized trial. For the foreseeable future, excellent clinical reasoning skills and experience supplemented by well-designed quantitative tools and a keen appreciation for the role of individual patient preferences in their health care will continue to be of paramount importance in the practice of clinical medicine.

FURTHER READING

Croskerry P: A universal model of diagnostic reasoning. Acad Med 84:1022, 2009. [PubMed: 19638766]

Dhaliwal G, Detsky AS: The evolution of the master diagnostician. JAMA 310:579, 2013. [PubMed: 23942674]

Hunink MGM et al: *Decision Making in Health and Medicine: Integrating Evidence and Values*, 2nd ed. Cambridge, Cambridge University Press, 2014.

Kahneman D: *Thinking Fast and Slow*. New York, Farrar, Straus and Giroux, 2013.

Kassirer JP et al: *Learning Clinical Reasoning*, 2nd ed. Baltimore, Lippincott Williams & Wilkins, 2009.

Mandelblatt JS et al: Collaborative modeling of the benefits and harms of associated with different U.S. breast cancer screening strategies. Ann Intern Med 164:215, 2016. [PubMed: 26756606]

McDonald KM et al: Patient safety strategies targeted at diagnostic errors. A systematic review. Ann Intern Med 158:381, 2013. [PubMed: 23460094]

National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care. Washington, DC, The National Academies Press, 2015.

Norman GR et al: The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Acad Med 92:23, 2017. [PubMed: 27782919]

Saposnik G et al: Cognitive biases associated with medical decisions: A systematic review. BMC Med Inform Decis Mak 16:138, 2016. [PubMed: 27809908]

Chapter 4: Screening and Prevention of Disease

Katrina A. Armstrong; Gary J. Martin

INTRODUCTION

A primary goal of health care is to prevent disease or detect it early enough that intervention will be more effective. Tremendous progress has been made toward this goal over the last 50 years. Screening tests are available for many common diseases and encompass biochemical (e.g., cholesterol, glucose), physiologic (e.g., blood pressure, growth curves), radiologic (e.g., mammogram, bone densitometry), and cytologic (e.g., Pap smear) approaches. Effective preventive interventions have resulted in dramatic declines in mortality from many diseases, particularly infections. Preventive interventions include counseling about risk behaviors, vaccinations, medications, and, in some relatively uncommon settings, surgery. Preventive services (including screening tests, preventive interventions, and counseling) are different than other medical interventions because they are proactively administered to healthy individuals instead of in response to a symptom, sign, or diagnosis. Thus, the decision to recommend a screening test or preventive intervention requires a particularly high bar of evidence that testing and intervention are both practical and effective.


Because population-based screening and prevention strategies must be extremely low risk to have an acceptable benefit-to-harm ratio, the ability to target individuals who are more likely to develop disease could enable the application of a wider set of potential approaches and increase efficiency. Currently, there are many types of data that can predict disease incidence in an asymptomatic individual. Genomic data have received the most attention to date, at least in part because mutations in high-penetrance genes have clear implications for preventive care **(Chap. 457)**. Women with mutations in either *BRCA1* or *BRCA2*, the two major breast cancer susceptibility genes identified to date, have a markedly increased risk (five- to twentyfold) of breast and ovarian cancer. Screening and prevention recommendations include prophylactic oophorectomy and breast magnetic resonance imaging (MRI), both of which are considered to incur too much harm for women at average cancer risk. Some women opt for prophylactic mastectomy to dramatically reduce their breast cancer risk. Although the proportion of common disease explained by high-penetrance genes appears to be relatively small (5–10% of most diseases), mutations in rare, moderate-penetrance genes, and variants in low-penetrance genes, also contribute to the prediction of disease risk. The advent of affordable whole exome/whole genome sequencing is likely to speed the dissemination of these tests into clinical practice and may transform the delivery of preventive care.

Other forms of “omic” data also have the potential to provide important predictive information, including proteomics and metabolomics. These fields are earlier in development and have yet to move into clinical practice. Imaging and other clinical data may also be integrated into a risk-stratified paradigm as evidence grows about the predictive ability of these data and the feasibility of their collection. Of course, all of these data may also be helpful in predicting the risk of harms from screening or prevention, such as the risk of a false-positive mammogram. To the degree that this information can be incorporated into personalized screening and prevention strategies, it could also improve delivery and efficiency.

In addition to advances in risk prediction, there are several other factors that are likely to promote the importance of screening and prevention in the near term. New imaging modalities are being developed that promise to detect changes at the cellular and subcellular levels, greatly increasing the probability that early detection improves outcomes. The rapidly growing understanding of the biologic pathways underlying initiation and progression of many common diseases has the potential to transform the development of preventive interventions, including chemoprevention. Furthermore, screening and prevention offer the promise of both improving health and sparing the costs of disease treatment, an issue that has gained national attention with the relatively high proportion of Gross Domestic Product spent on health care in the United States.

This chapter will review the basic principles of screening and prevention in the primary care setting. Recommendations for specific disorders such as cardiovascular disease, diabetes, and cancer are provided in the chapters dedicated to those topics.

BASIC PRINCIPLES OF SCREENING

The basic principles of screening populations for disease were published by the World Health Organization in 1968 **(Table 4-1)**.

TABLE 4-1

Principles of Screening

The condition should be an important health problem.
There should be a treatment for the condition.
Facilities for diagnosis and treatment should be available.
There should be a latent stage of the disease.
There should be a test or examination for the condition.
The test should be acceptable to the population.
The natural history of the disease should be adequately understood.
There should be an agreed policy on whom to treat.
The cost of finding a case should be balanced in relation to overall medical expenditure.

In general, screening is most effective when applied to relatively common disorders that carry a large disease burden (Table 4-2). The five leading causes of mortality in the United States are heart diseases, malignant neoplasms, chronic obstructive pulmonary disease, accidents, and cerebrovascular diseases. Thus, many screening strategies are targeted at these conditions. From a global health perspective, these conditions are priorities, but malaria, malnutrition, AIDS, tuberculosis, and violence also carry a heavy disease burden (Chap. 460).

TABLE 4-2

Lifetime Cumulative Risk

Breast cancer for women	10%
Colon cancer	6%
Cervical cancer for women[a]	2%
Domestic violence for women	Up to 15%
Hip fracture for white women	16%

[a]Assuming an unscreened population.


Having an effective treatment for early disease has proven challenging for some common diseases. For example, although Alzheimer's disease is the sixth leading cause of death in the United States, there are no curative treatments and no evidence that early treatment improves outcomes. Lack of facilities for diagnosis and treatment is a particular challenge for developing countries and may change screening strategies, including the development of "see and treat" approaches such as those currently used for cervical cancer screening in some countries. A long latent or preclinical phase where early treatment increases the chance of cure is a hallmark of many cancers; for example, polypectomy prevents progression to colon cancer. Similarly, early identification of hypertension or hyperlipidemia allows therapeutic interventions that reduce the long-term risk of cardiovascular or cerebrovascular events. In contrast, lung cancer screening has historically proven more challenging because most tumors are not curable by the time they can be detected on a chest x-ray. However, the length of the preclinical phase also depends on the level of resolution of the screening test, and this situation changed with the development of chest computed tomography (CT). Low-dose chest CT scanning can detect tumors earlier and has been demonstrated to reduce lung cancer mortality by 20% in individuals who had at least a 30-pack-year history of smoking. The short interval between the ability to detect disease on a screening test and the development of incurable disease also contributes to the limited effectiveness of mammography screening in reducing deaths from some forms of breast cancer. Similarly, the early detection of prostate cancer may not lead to a difference in the mortality rate because the disease is often indolent and competing morbidities, such as coronary artery disease, may ultimately cause mortality (Chap. 66). This uncertainty about the natural history is also reflected in the controversy about treatment of prostate cancer, further contributing to the challenge of screening in this disease. Finally, screening programs can incur significant economic costs that must be considered in the context of the available resources and alternative strategies for improving health outcomes.

METHODS OF MEASURING HEALTH BENEFITS

Because screening and preventive interventions are recommended to asymptomatic individuals, they are held to a high standard for demonstrating a favorable risk-benefit ratio before implementation. In general, the principles of evidence-based medicine apply to demonstrating the efficacy of screening tests and preventive interventions, where randomized controlled trials (RCTs) with mortality outcomes are the gold standard. However, because RCTs are often not feasible, observational studies, such as case-control designs, have been used to assess the effectiveness of some interventions such as colorectal cancer screening. For some strategies, such as Pap smear screening for cervical cancer, the only data available are ecologic data demonstrating dramatic declines in mortality.

Irrespective of the study design used to assess the effectiveness of screening, it is critical that disease incidence or mortality is the primary endpoint rather than length of disease survival. This is important because lead time bias and length time bias can create the appearance of an improvement in disease survival from a screening test when there is no actual effect. Lead time bias occurs because screening identifies a case before it would have presented clinically, thereby creating the perception that a patient lived longer after diagnosis simply by moving the date of diagnosis earlier rather than the date of death later. Length time bias occurs because screening is more likely to identify slowly progressive disease than rapidly progressive disease. Thus, within a fixed period of time, a screened population will have a greater proportion of these slowly progressive cases and will appear to have better disease survival than an unscreened population.

A variety of endpoints are used to assess the potential gain from screening and preventive interventions.

1. *The absolute and relative impact of screening on disease incidence or mortality*. The absolute difference in disease incidence or mortality between a screened and nonscreened group allows the comparison of size of the benefit across preventive services. A meta-analysis of Swedish mammography trials (ages 40–70) found that ~1.2 fewer women per 1000 would die from breast cancer if they were screened over a 12-year period. By comparison, ~3 lives per 1000 would be saved from colon cancer in a population (aged 50–75) screened with annual fecal occult blood testing (FOBT) over a 13-year period. Based on this analysis, colon cancer screening may actually save more women's lives than does mammography. However, the relative impact of FOBT (30% reduction in colon cancer death) is similar to the relative impact of mammography (14–32% reduction in breast cancer death), emphasizing the importance of both relative and absolute comparisons.

2. *The number of subjects screened to prevent disease or death in one individual*. The inverse of the absolute difference in mortality is the number of subjects who would need to be screened or receive a preventive intervention to prevent one death. For example, 731 women aged 65–69 would need to be screened by dual-energy x-ray absorptiometry (DEXA) (and treated appropriately) to prevent one hip fracture from osteoporosis.

3. *Increase in average life expectancy for a population*. Predicted increases in life expectancy for various screening and preventive interventions are listed in **Table 4-3**. It should be noted, however, that the increase in life expectancy is an average that applies to a population, not to an individual. In reality, the

vast majority of the population does not derive any benefit from a screening test or preventive intervention. A small subset of patients, however, will benefit greatly. For example, Pap smears do not benefit the 98% of women who never develop cancer of the cervix. However, for the 2% who would have developed cervical cancer, Pap smears may add as much as 25 years to their lives. Some studies suggest that a 1-month gain of life expectancy is a reasonable goal for a population-based screening or prevention strategy.

TABLE 4-3

Estimated Average Increase in Life Expectancy for a Population

Screening or Preventive Intervention	Average Increase
Mammography: Women, 40–50 years Women, 50–70 years	 0–5 days 1 month
Pap smears, age 18–65	2–3 months
Getting a 35-year-old smoker to quit	3–5 years
Beginning regular exercise for a 40-year-old man (30 min, 3 times a week)	9 months–2 years

ASSESSING THE HARMS OF SCREENING AND PREVENTION

Just as with most aspects of medical care, screening and preventive interventions also incur the possibility of adverse outcomes. These adverse outcomes include side effects from preventive medications and vaccinations, false-positive screening tests, overdiagnosis of disease from screening tests, anxiety, radiation exposure from some screening tests, and discomfort from some interventions and screening tests. The risk of side effects from preventive medications is analogous to the use of medications in therapeutic settings and is considered in the U.S. Food and Drug Administration (FDA) approval process. Side effects from currently recommended vaccinations are primarily limited to discomfort and minor immune reactions. However, the concern about associations between vaccinations and serious adverse outcomes continues to limit the acceptance of many vaccinations despite the lack of data supporting the causal nature of these associations.

The possibility of a false-positive test occurs with nearly all screening tests, although the definition of what constitutes a false-positive result often varies across settings. For some tests such as screening mammography and screening chest CT, a false-positive result occurs when an abnormality is identified that is not malignant, requiring either a biopsy diagnosis or short-term follow-up. For other tests such as Pap smears, a false-positive result occurs because the test identifies a wide range of potentially premalignant

states, only a small percentage of which would ever progress to an invasive cancer. This risk is closely tied to the risk of overdiagnosis in which the screening test identifies disease that would not have presented clinically in the patient's lifetime. Assessing the degree of overdiagnosis from a screening test is very difficult given the need for long-term follow-up of an unscreened population to determine the true incidence of disease over time. Recent estimates suggest that as much as 15–40% of breast cancers identified by mammography screening and 15–37% of prostate cancers identified by prostate-specific antigen testing may never have presented clinically. Screening tests also have the potential to create unwarranted anxiety, particularly in conjunction with false-positive findings. Although multiple studies have documented increased anxiety through the screening process, there are few data suggesting this anxiety has long-term adverse consequences, including subsequent screening behavior. Screening tests that involve radiation (e.g., mammography, chest CT) add to the cumulative radiation exposure for the screened individual. The absolute amount of radiation is very small from any of these tests, but the overall impact of repeated exposure from multiple sources is still being determined. Some preventive interventions (e.g., vaccinations) and screening tests (e.g., mammography) may lead to discomfort at the time of administration, but again, there is little evidence of long-term adverse consequences.

WEIGHING THE BENEFITS AND HARMS

The decision to implement a population-based screening and prevention strategy requires weighing the benefits and harms, including the economic impact of the strategy. The costs include not only the expense of the intervention but also time away from work, downstream costs from false-positive results or adverse events, and other potential harms. Cost-effectiveness is typically assessed by calculating the cost per year of life saved, with adjustment for the quality of life impact of different interventions and disease states (i.e., quality-adjusted life-year). Typically, strategies that cost $50,000–100,000 per quality-adjusted year of life saved are considered "cost-effective" **(Chap. 3)**.

The U.S. Preventive Services Task Force (USPSTF) is an independent panel of experts in preventive care that provides evidence-based recommendations for screening and preventive strategies based on an assessment of the benefit-to-harm ratio **(Tables 4-4 and 4-5)**. Because there are multiple advisory organizations providing recommendations for preventive services, the agreement among the organizations varies across the different services. For example, all advisory groups support screening for hyperlipidemia and colorectal cancer, whereas consensus is lower for breast cancer screening among women in their forties and for prostate cancer screening. Because the guidelines are only updated periodically, differences across advisory organizations may also reflect the data that were available when the guideline was issued. For example, the recommendations about lung cancer screening among heavy smokers varied across organizations after the results of the National Lung Screening Trial (NLST) were published in 2011 based upon how quickly the screening guidelines were updated.

TABLE 4-4

Screening Tests Recommended by the U.S. Preventive Services Task Force for Average-Risk Adults

Disease	Test	Population	Frequency	Chapter
Abdominal aortic aneurysm	Ultrasound	Men 65–75 who have ever smoked	Once	
Alcohol misuse	Alcohol Use Disorders Identification Test	All adults	Unknown	445
Breast cancer	Mammography with or without clinical breast examination	Women 50–75	Every 2 years	
Cervical cancer	Pap smear	Women 21–65	Every 3 years	66
	Pap smear and HPV testing	Women 30–65	Every 5 years if HPV negative	
Chlamydia/gonorrhea	Nucleic acid amplification test on urine or cervical swab	Sexually active women <25	Unknown	184
Colorectal cancer	Fecal occult blood testing	50–75	Every year	66, 77
	Sigmoidoscopy	50–75	Every 5 years	
	Colonoscopy (or occult blood testing combined with sigmoidoscopy)	50–75	Every 10 years	
Depression	Screening questions	All adults	Periodically	
Diabetes	Fasting blood glucose or HgbA1c	Adults overweight, obese or with hypertension	Every 3 years	396
Hepatitis C	Anti-HCV antibody followed by confirmatory PCR	Adults born between 1945 and 1965	Once	

Disease	Test	Population	Frequency	Chapter
HIV	Reactive immunoassay or rapid HIV followed by confirmatory test	15–65	At least once	
Hyperlipidemia	Cholesterol	40–75	Unknown	400
Hypertension	Blood pressure	All adults	Periodically	271
Intimate partner violence	Screening questions	Women of childbearing age	Unknown	
Obesity	Body mass index	All adults	Unknown	
Osteoporosis	DEXA	Women >65 or >60 with risk factors	Unknown	404

Abbreviations: DEXA, dual-energy x-ray absorptiometry; HCV, hepatitis C virus; HPV, human papillomavirus; PCR, polymerase chain reaction.

Source: Adapted from the U.S. Preventive Services Task Force 2017. *www.uspreventiveservicestaskforce.org/Page/Name/uspstf-a-and-b-recommendations/.*

TABLE 4-5

Preventive Interventions Recommended for Average-Risk Adults

Intervention	Disease	Population	Frequency	Chapter
Adult immunization				118, 119
Tetanus-diphtheria		>18	Every 10 years	
Varicella		Susceptibles only, >18	Two doses	
Measles-mumps-rubella		Women, childbearing age	One dose	
Pneumococcal		>64	13 followed by 23 valent	
Influenza		>18	Yearly	
Human papillomavirus		Up to age 27	If not done prior	
Zoster		>60	Once	
Chemoprevention				
Aspirin	Cardiovascular disease	Aged 50 to 59 years with a ≥10% 10-year CVD risk		
Folic acid	Neural tube defects in baby	Women planning or capable of pregnancy		
Tamoxifen/raloxifene	Breast cancer	Women at high risk for breast cancer		
Vitamin D	Fracture/falls	>64 at increased risk for falls		

For many screening tests and preventive interventions, the balance of benefits and harms may be uncertain for the average-risk population but more favorable for individuals at higher risk for disease. Although age is

the most commonly used risk factor for determining screening and prevention recommendations, the USPSTF also recommends some screening tests in populations based upon the presence of other risk factors for the disease. In addition, being at increased risk for the disease often supports initiating screening at an earlier age than that recommended for the average-risk population. For example, when there is a significant family history of colon cancer, it is prudent to initiate screening 10 years before the age at which the youngest family member was diagnosed with cancer.

Although informed consent is important for all aspects of medical care, shared decision-making may be a particularly important approach to decisions about preventive services when the benefit-to-harm ratio is uncertain for a specific population. For example, many expert groups, including the American Cancer Society, recommend an individualized discussion about prostate cancer screening, because the decision-making process is complex and relies heavily on personal issues. Some men may decline screening, whereas others may be more willing to accept the risks of an early detection strategy. Recent analysis suggests that many men may be better off not screening for prostate cancer because watchful waiting was the preferred strategy when quality-adjusted life-years were considered. Another example of shared decision-making involves the choice of techniques for colon cancer screening **(Chap. 66)**. In controlled studies, the use of annual FOBT reduces colon cancer deaths by 15–30%. Flexible sigmoidoscopy reduces colon cancer deaths by ~40–60%. Colonoscopy appears to offer a greater benefit than flexible sigmoidoscopy with a reduction in risk of ~70%, but its use incurs additional costs and risks. These screening procedures have not been compared directly in the same population, but models suggest that appropriate frequencies of each technique may be associated with similar numbers of lives saved and cost to society per life saved ($10,000–25,000). Thus, although one patient may prefer the ease of preparation, less time disruption, and the lower risk of flexible sigmoidoscopy, others may prefer the sedation, thoroughness and time interval of colonoscopy.

COUNSELING ON HEALTHY BEHAVIORS

In considering the impact of preventive services, it is important to recognize that tobacco and alcohol use, diet, and exercise constitute the vast majority of factors that influence preventable deaths in developed countries. Perhaps the single greatest preventive health care measure is to help patients quit smoking **(Chap. 448)**. However, efforts in these areas frequently require behavior changes (e.g., weight loss, exercise) or the management of addictive conditions (e.g., tobacco and alcohol use) that are often recalcitrant to intervention. Although these are challenging problems, evidence strongly supports the role of counseling by health care providers **(Table 4-6)** in effecting health behavior change. Educational campaigns, public policy changes, and community-based interventions have also proven to be important parts of a strategy for addressing these factors in some settings. Although the USPSTF found that the evidence was conclusive to recommend a relatively small set of counseling activities, counseling in areas such as physical activity and injury prevention (including seat belts and bicycle and motorcycle helmets) has become a routine part of primary care practice.

TABLE 4-6

Preventive Counseling Recommended by the U.S. Preventive Services Task Force (USPSTF)

Topic	Chapter Reference
Alcohol and drug use	445, 446, 447
Genetic counseling for *BRCA1/2* testing among women at increased risk for deleterious mutations	75, 457
Nutrition and diet	325, 326
Sexually transmitted infections	131, 197
Sun exposure	57
Tobacco use	448

IMPLEMENTING DISEASE PREVENTION AND SCREENING

The implementation of disease prevention and screening strategies in practice is challenging. A number of techniques can assist physicians with the delivery of these services. An appropriately configured electronic health record can provide reminder systems that make it easier for physicians to track and meet guidelines. Some systems give patients secure access to their medical records, providing an additional means to enhance adherence to routine screening. Systems that provide nurses and other staff with standing orders are effective for immunizations. The USPSTF has developed flow sheets and electronic tools to assist clinicians (*https://www.uspreventiveservicestaskforce.org/Page/Name/tools-and-resources-for-better-preventive-care*). Many of these tools use age categories to help guide implementation. Age-specific recommendations for screening and counseling are summarized in **Table 4-7**.

TABLE 4-7

Age-Specific Causes of Mortality and Corresponding Preventive Options

Age Group	Leading Causes of Age-Specific Mortality	Screening Prevention Interventions to Consider for Each Specific Population
15–24	1. Accident 2. Homicide 3. Suicide 4. Malignancy 5. Heart disease	• Counseling on routine seat belt use, bicycle/motorcycle/ATV helmets (1) • Counseling on diet and exercise (5) • Discuss dangers of alcohol use while driving, swimming, boating (1) • Assess and update vaccination status (tetanus, diphtheria, hepatitis B, MMR, rubella, varicella, meningitis, HPV) • Ask about gun use and/or gun possession (2,3) • Assess for substance abuse history including alcohol (2,3) • Screen for domestic violence (2,3) • Screen for depression and/or suicidal/homicidal ideation (2,3) • Pap smear for cervical cancer screening after age 21 (4) • Discuss skin, breast awareness, and testicular self-examinations (4) • Recommend UV light avoidance and regular sunscreen use (4) • Measurement of blood pressure, height, weight, and body mass index (5) • Discuss health risks of tobacco use, consider emphasis on cosmetic and economic issues to improve quit rates for younger smokers (4,5) • Chlamydia and gonorrhea screening and contraceptive counseling for sexually active females, discuss STD prevention • Hepatitis B, and syphilis testing if there is high-risk sexual behavior(s) or any prior history of sexually transmitted disease • HIV testing • Continue annual influenza vaccination

Age Group	Leading Causes of Age-Specific Mortality	Screening Prevention Interventions to Consider for Each Specific Population
25–44	1. Accident 2. Malignancy 3. Heart disease 4. Suicide 5. Homicide 6. HIV	*As above plus consider the following:* • Readdress smoking status, encourage cessation at every visit (2,3) • Obtain detailed family history of malignancies and begin early screening/prevention program if patient is at significant increased risk (2) • Assess all cardiac risk factors (including screening for diabetes and hyperlipidemia) and consider primary prevention with aspirin for patients at >3% 5-year risk of a vascular event (3) • Assess for chronic alcohol abuse, risk factors for viral hepatitis, or other risks for development of chronic liver disease • Consider individualized breast cancer screening with mammography at age 40 (2)
45–64	1. Malignancy 2. Heart disease 3. Accident 4. Diabetes mellitus 5. Cerebrovascular disease 6. Chronic lower respiratory disease 7. Chronic liver disease and cirrhosis 8. Suicide	• Consider prostate cancer screen with annual PSA and digital rectal examination at age 50 (or possibly earlier in African Americans or patients with family history) (1) • Begin colorectal cancer screening at age 50 with fecal occult blood testing, flexible sigmoidoscopy, or colonoscopy (1) • Reassess and update vaccination status at age 50 and vaccinate all smokers against *S. pneumoniae* at age 50 (6) • Consider screening for coronary disease in higher-risk patients (2,5) • Consider screening for hepatitis C in adults born between 1945 and 1965 (7) • Zoster vaccination at age 60 • Begin mammography screening by age 50

Age Group	Leading Causes of Age-Specific Mortality	Screening Prevention Interventions to Consider for Each Specific Population
≥65	1. Heart disease 2. Malignancy 3. Cerebrovascular disease 4. Chronic lower respiratory disease 5. Alzheimer's disease 6. Influenza and pneumonia 7. Diabetes mellitus 8. Kidney disease 9. Accidents 10. Septicemia	*As above plus consider the following:* • Readdress smoking status, encourage cessation at every visit (1,2,3,4) • One-time ultrasound for AAA in men 65–75 who have ever smoked • Consider pulmonary function testing for all long-term smokers to assess for development of chronic obstructive pulmonary disease (4,6) • Screen all postmenopausal women (and all men with risk factors) for osteoporosis • Continue annual influenza vaccination and vaccinate against *S. pneumoniae* at age 65 (4,6) • Screen for visual and hearing problems, home safety issues, and elder abuse (9)

Note: The numbers in parentheses refer to areas of risk in the mortality column affected by the specified intervention.

Abbreviations: AAA, abdominal aortic aneurysm; ATV, all-terrain vehicle; HPV, human papillomavirus; MMR, measles-mumps-rubella; PSA, prostate-specific antigen; STD, sexually transmitted disease; UV, ultraviolet.

Many patients see a physician for ongoing care of chronic illnesses, and this visit provides an opportunity to include a "measure of prevention" for other health problems. For example, a patient seen for management of hypertension or diabetes can have breast cancer screening incorporated into one visit and a discussion about colon cancer screening at the next visit. Other patients may respond more favorably to a clearly defined visit that addresses all relevant screening and prevention interventions. Because of age or comorbidities, it may be appropriate with some patients to abandon certain screening and prevention activities, although there are fewer data about when to "sunset" these services. For many screening tests, the benefit of screening does not accrue until 5–10 years of follow-up, and there are generally few data to support continuing screening for most diseases past age 75. In addition, for patients with advanced diseases and limited life expectancy, there is considerable benefit from shifting the focus from screening procedures to the conditions and interventions more likely to affect quality and length of life.

FURTHER READING

Bretthauer M et al: America, We Are Confused: The Updated U.S. Preventive Services Task Force Recommendation on Colorectal Cancer Screening. Ann Intern Med 166:139, 2017. [PubMed: 27820949]

Hayes JH et al: Observation versus initial treatment for men with localized, low-risk prostate cancer: A cost-effectiveness analysis. Ann Intern Med 158:853, 2013. [PubMed: 23778902]

Hugosson J et al: Mortality results from the Goteborg randomized population-based prostate-cancer screening trial. Lancet Oncol 11:725, 2010. [PubMed: 20598634]

Oeffinger KC et al: Breast Cancer Screening for Women at Average Risk 2015. Guideline Update from the American Cancer Society. JAMA 314:1599, 2015. [PubMed: 26501536]

US Preventive Services Task Force: Screening for Colorectal Cancer. US Preventive Services Task Force Recommendation Statement. JAMA 315:2564, 2016. [PubMed: 27304597]

Chapter 5: Health Care Systems in Developed Countries

Richard B. Saltman

INTRODUCTION

Health care systems are highly complex organizations, with many interdependent components. Traditionally, health systems in the developed world have been classified by their type of financing—i.e., either predominantly tax-funded (such as the National Health Service in England and publicly operated regional care systems in the four European Nordic countries) or predominantly statutory social health insurance (SHI)-funded (such as in Germany, the Netherlands, and France). Over the last decade, however, there has been structural convergence in the technical characteristics of both funding arrangements, and also in the associated delivery systems, making analytic observations about the differences across national systems more difficult.

A second confounding factor has been that former Soviet Bloc countries have, since 1991, replaced their former Soviet-style Semashko models (a top-down, national government–controlled structure with a parallel Communist Party apparatus) with various hybrid arrangements built on government-run SHI financing. Distinctions across health systems, especially in Europe, have been further compressed by the continuing negative impact of the 2008 global financial crisis on public revenues in many countries.

This chapter focuses on the individual patient care system: on the financing and delivery of individual clinical and preventive services. The individual patient care system is composed of the financing and delivery of necessary services to prevent death or serious harm (“rule of rescue”); to maintain quality of life; and to manage, reduce, and/or prevent the burden of illness on individual patients. While the technical dimensions of most clinical services are similar across countries, their organizational, social, and economic characteristics differ markedly. Health systems in different countries exhibit substantial differences, for example, in access to care; in the design and reliance on quality assurance and provider payment mechanisms; in the relationship of primary care to hospital services; in the coordination of health care with home care and nursing home services; in the design and use of provider management strategies; in the way physicians work and are paid; in the decision-making roles of politically elected officials and of national, regional and municipal governments; and in participation of both citizens and patients. These differences reflect differing country contexts (geographical, social, economic and political), differences in national

culture (consisting of prioritized norms and values), and substantial variation in how health sector institutions are structured.

FINANCING INDIVIDUAL PATIENT CARE SERVICES

Funding for individual care services in developed countries comes from the particular national mix among four possible sources of revenue: national, regional and/or municipal taxes; mandatory social health insurance; private health insurance; and out-of-pocket payments. Most countries have one preponderant payer, which then defines its funding arrangements and serves to frame the structure of its delivery system as well.

The Organization for Economic Co-operation and Development (OECD) data from 2015 (adjusted for purchasing power parities) show that total health care expenditures in developed countries vary across a considerable range, tied to health system structure as well as national history and culture. Total health expenditures in tax-funded health systems in Western Europe ranged from a low of 9.0% of GDP in Spain and 9.6% in Finland to a high of 10.6% in Denmark and 11.1% in Sweden. In SHI-funded systems in Western Europe, the range was about 1% higher, running from 10.4% in Belgium and 10.8% in the Netherlands to a high of 11.1% in Germany and 11.5% in Switzerland. Central European health care systems, reflecting the economic and health system consequences of their pre-1991 Soviet Bloc history, spend considerably lower percentages of their GDP on health care: from a low of 5.6% in Latvia and 6.1% in Estonia to 7.7% in the Czech Republic and 8.4% in Slovenia. In Asia, total health expenditures ranged widely from 4.9% in 2014 in Singapore to 7.1% in 2015 in South Korea to 11.4% in 2015 in Japan. Total health expenditures in the United States in 2015 were 16.9% of GDP.

Tax-Funded Systems

In the United Kingdom, 79% of all health care funding was furnished through general tax revenues allocated by the national government in its annual budget process (all figures from OECD 2016). In Sweden, all public taxes combined raised 83.7% of total health care spending in 2015. Sweden's 21 regional level elected governments provide approximately 70% of that 83.7%, with the remaining 13.7% of total health spending raised by national and municipal taxes. In Canada, 71% of total health spending was raised by tax revenues, with 66% of that 71% coming from provincial or territorial taxes, while 5% came from national and local government taxes.

Social Insurance–Funded Systems

In Western Europe, SHI funds have traditionally been organized on a private not-for-profit basis, but with statutory responsibilities under national law. When former Soviet Bloc countries in Eastern Europe regained their independence in 1991, they returned to pre-WWII SHI models, but since there was no remaining organizational infrastructure, these post-1991 arrangements typically became a single SHI fund, run as an arm of the national government. In the United States, the Medicare social insurance system for citizens over age 65, enacted in 1965, is organized as a single fund tied to the national Social Security (public pension)

Administration, an independent agency within the national government, with reimbursement arrangements supervised by the Centers for Medicare & Medicaid Services inside the Department of Health and Human Services. Medicare covers inpatient hospital care plus limited post-hospital nursing home services (Medicare Part A). Supplemental private insurance policies are bought by covered individuals to help pay for outpatient physician visits (Medicare Part B) and for outpatient pharmaceuticals (Medicare Part D).

In Germany, 85% of the population is enrolled in 132 not-for-profit private social health insurance funds (in addition, 11% of the population are enrolled in private health insurance, and 4% in sector-specific public programs such as the military). Since 2009, all SHI members pay a flat tax on gross monthly income as a contribution (8.2% in 2013), which is sent by their SHI fund to a national pool. In addition, employers send 7.3% of each employee's salary to the same national pool. Special arrangements exist for payments from self-employed, retired, and unemployed workers. Since 1995 there has been a separate mandatory social insurance fund for long-term care (LTC), with an annual premium of 1.95% of each adult's gross monthly income, split 50%–50% with their employer. Pensioners since 2004 are required to pay the full amount from their pensions. Childless SHI enrollees pay a surcharge of 0.25% of monthly gross income. Overall, 85% of all health care expenditures in Germany were paid from public and/or mandatory private SHI sources.

In the Netherlands since 2006, all adult citizens pay a fixed premium (about 1290 euros in 2015) to their choice among 35 private health insurers (not-for-profit and for-profit). In addition, employers pay 6.95% of salary below 51,400 euro for each employee into a national health insurance fund. Self-employed individuals pay 4.85% into the national fund for taxable income up to the same limit. Retired and unemployed individuals also make payments. In addition to the individual premiums paid to their choice of private insurance fund, payments from the national health insurance fund, adjusted by individual age, sex, and health characteristics, also are made to the individual's chosen insurer. The Netherlands also has a separate mandatory social insurance fund for LTC (the ABWZ, since 2015 the WLZ, and now only for residential nursing home care) to which each employee pays 9.5% of taxable income beneath 33,600 euros every year. Self-employed, unemployed, and retired individuals also are required to pay premiums to the WLZ. Overall, including SHI revenues, public spending provided 87% of total health expenditures in 2014.

In Estonia, a former Soviet Republic that re-established a social health insurance system in 1991 upon regaining its independence, there is one national social health insurance fund that is an arm of the national government. This fund collects mandatory payments of 13% from salaried workers and 20% from self-employed individuals, covering both health care and retirement pensions. Overall, including SHI revenues, public spending accounted for 78.8% of total health expenditures.

Singapore, Japan, South Korea, and Taiwan have predominantly SHI systems of funding for individual care services. In these Asian countries (except Japan) there is one SHI fund that typically is operated as an arm of the national government.

In Singapore, starting in 1983, all employees up to age 50 have been required to place 20% of their income (employers add 16% more) into a health savings account to pay for direct health care costs, managed in their name by the Singapore government, called a Medisave account. Medisave accounts have a maximum

amount, are tax-exempt, and receive interest payments (currently set at 4%). Consistent with a Confucian emphasis on family, the funds that accumulate in the Medisave account can be spent on health care for family members as well. If the accumulated funds are not spent on health care during the insured's life, they become part of the individual's personal estate and are distributed as an inheritance to his/her designated heirs. In addition, Singaporean citizens are also automatically enrolled into a second government-run health insurance plan called MediShield that pays for supplemental catastrophic, chronic, and long-term care. While citizens can opt out, 90% of citizens remain in the program. The Singapore government also operates a third, wholly tax-funded payer called Medifund that, with approval of a local neighborhood committee, will pay hospital costs for 3–4% of the population who are recognized as indigent. In part reflecting the high level of mandatory individual saving, tax funds provided only 41.7% of total health expenditures in 2014.

In South Korea, a state-run social health insurance system was established in 1977, which in 1990 covered 30.9% of total health care costs. This percentage paid by the SHI system rose to 43.6% of total costs in 2011, leaving out-of-pocket costs at 35.2% of total costs. In 2000, three types of public SHI funds were merged into a single national state-run fund. As of 2011, 5.64% of an employee's salary must be paid as a social insurance contribution into this fund, with employees and employers each paying 50% of that amount. In 2008, an additional SHI fund was introduced to pay for LTC, operated by the main state-run SHI fund to reduce administrative costs. Contributions to the LTC fund are set at 6.55% of the individual's regular SHI contribution, coupled with 20% copayments for institutional care and 15% copayments for home care services.

There is no single preponderant source of health care spending in the United States. The source of health care revenues is fairly evenly divided among (1) national, state, county, and municipal taxes at 20% of all health spending in 2011 (for Medicaid, Children's Health Insurance Program [CHIP], the Veteran's Administration Hospitals, the Public Health Service, and the Indian Health Service); (2) mandatory social health insurance (for Medicare for all citizens over 65) at 23% of all spending; and (3) private health insurance (company and individual) at 35% of all spending. Out-of-pocket payments make up the remaining 14%. The World Bank, combining tax and social insurance funding, sets public funding in the United States at 48.3% of total health expenditures in 2014.

In 2010, the passage of the Affordable Care Act (ACA) extended privately provided but heavily regulated and federally subsidized health insurance to a number of low- and middle-income uninsured individuals and families. Since the same act reduced the availability of existing individually purchased private health insurance, the total increase in the number of newly covered individuals was less than expected. Insurance premium increases for 2017 have risen from 20% to over 100%, depending on the particular state, with additional increases in up-front deductible requirements, raising serious questions about the long-term sustainability of the ACA initiative. The current Republican administration has sought to repeal major financial and tax elements of the ACA (using congressional budget reconciliation rules) and to replace existing subsidy arrangements with a system of refundable tax credits toward the establishment of individual health savings accounts and/or purchase of private health insurance on open cross-state markets.

DELIVERING INDIVIDUAL PATIENT CARE SERVICES

Hospital Services

In Europe, hospitals in both tax-funded and SHI-funded health systems are mostly publicly owned and operated by regional or municipal governments. In tax-funded health systems, most hospital-based physicians are civil servants, employed on a negotiated salary basis (often by a physician labor union), and subject to most of the usual advantages and disadvantages of being a public sector employee. There are somewhat more private hospitals in SHI-funded health systems. However, most larger hospitals are public institutions operated by local governments, and most hospital physicians (with the notable exception of the Netherlands, where they are private contractors organized in private group practices) are, like those in tax-funded systems, public sector employees. In most tax-funded European countries (but not continental SHI-funded countries), few specialist physicians have office-based practices, and in both tax- and SHI-funded systems, office-based specialists do not have admitting privileges to publicly operated hospitals.

Most public hospitals in both tax-funded and SHI-funded health systems are single free-standing institutions that can be classified into three broad categories by complexity of patients admitted and number of specialties available: (1) district hospitals (four specialties: internal medicine, general surgery, obstetrics, and psychiatry); (2) regional hospitals (20 specialties); and (3) university hospitals (>40 specialties). In addition, many countries have a number of small, 15- to 20-bed, freestanding, private (typically for-profit) clinics. Recently, some countries have begun to merge district and regional hospitals in an effort to improve the quality of care and create financial efficiencies (for example, Norway; planned for Finland starting in 2019). Institutional mergers can be difficult to negotiate among publicly operated hospitals, due to the role that these large institutions play as important care providers and as large employers in smaller cities and towns, especially given political and union concerns about maintaining current employment levels. In the United States, financial and reimbursement pressures triggered by the implementation of the 2010 ACA have generated a number of private sector hospital mergers into larger hospital groups.

In tax-funded health systems, publicly funded patients who are admitted for an elective procedure cannot choose their specialist physician (except private-pay patients in “pay beds” in NHS hospitals in England). Specialists are assigned by the clinic to a patient based on availability, with both junior and senior doctors placed in rotation.

Capital costs (buildings, large medical equipment) are publicly funded in all tax-funded systems and in most traditional SHI systems. For example, in Germany capital costs for all public hospitals are paid for by the regional governments. As a result, new capital investment is often allocated politically, according to location and political priorities. In Finland, local politicians in the 1980s would say that it “takes 10 years to build a hospital,” meaning that it took that long to become a political priority for the regional government that controlled capital expenditures. As a result, local politicians would regularly overbuild when they got their one opportunity to obtain new capital. Because capital was not depreciated on the operating budget, such investment was perceived to be “free.” As a result, new equipment often was not properly serviced or kept in

use, as maintenance costs came from the operating budget, which was held by a different level political organization (municipalities in Finland).

Recently, efforts have been made to make public hospitals more responsible for their use of capital. In the Netherlands, public hospitals were shifted into private not-for-profit entities that are expected either to fund new capital from operating surplus or to borrow the funds from a bank with a viable business plan. In England, more than 100 hospitals have been built using the Public Finance Initiative (PFI) program, in which private developers build turn-key facilities (thus taking capital costs off the public borrowing limit), and then rent these facilities back to the NHS and/or the relevant NHS Foundation Trust.

In Singapore or South Korea, both of which are SHI funded, larger hospitals are publicly operated. However, there are a substantial number of smaller private clinics typically owned by specialist physicians. In the United States, the passage of the 2010 ACA has triggered the selling of many private specialist group practices to hospital groups, transforming previously independent practicing physicians into hospital employees.

Primary Care Services

Most primary health care in SHI-funded health systems, and also in an increasing number of tax-funded health systems (except in low-income areas of some large cities), is delivered by independent private general practitioners (GPs), working either individually or in small privately owned group practices. Recent changes in tax-funded health systems include Norway, where most primary care moved from municipally employed physicians to private-practice GPs in 2003, and Sweden, where, following a 2010 change in national reimbursement requirements, new privately owned not-for-profit and for-profit GP practices were established and now deliver 50% of all primary care visits. In Finland, where public primary health care centers used to provide most primary care visits, delays in getting public health center appointments have pushed up to 40% of all visits to a parallel occupational health system, as well as to publicly employed primary care physicians working privately in the afternoons, seeing patients who are partly reimbursed by Finland's separate Social Insurance Institution (known as KELA).

In England, most primary care physicians are private GPs who are contractors to the NHS, working either independently or in small group practices. These private GPs own their own practices, which they can sell when they retire. However, as part of the original agreement establishing the NHS in 1948 (which most physicians strongly opposed), private GPs also receive a national government pension upon retirement. In the inner cities in England, there are some larger primary health clinics.

In 2001, England's private primary care doctors were organized into geographically based Primary Care Trusts (PCTs). These PCTs were allocated 80% of the total NHS budget to contract for elective hospital services required by their patients with both NHS hospital trusts as well as private hospitals. In 2013, PCTs were restructured into Clinical Commissioning Groups with similar contracting responsibilities.

In 2004, the Quality Outcomes Framework (QOF) was introduced as a quality of care–tied approach to providing additional income for NHS GPs. This regulatory mechanism in 2010 set 134 different standards for

best practice primary care in four main domains: 86 clinical, 36 organizational, 4 preventive service, and 3 patient experience. GPs income grew on average by 25% through the introduction of the QOF, with general practices averaging 96% of possible QOF points. Total spending on QOF in 2014 in England consumed 15% of all primary care expenditures.

In Central European countries like Poland and Estonia that were formerly within the Soviet Bloc, primary care provision had to be newly established after independence was regained in 1991, since first-line care in the former Semashko model was provided in specialist polyclinics. Primary care doctors rapidly emerged as almost entirely private for-profit GPs working on contract from the national SHI fund. Private GPs in most Central European countries now are paid on a per-visit basis, in an amount set by the national SHI fund. This arrangement was heavily influenced by the structure of primary care in Germany's SHI-based health system.

In Asian countries such as Singapore, South Korea, and Japan, most primary care is provided by private for-profit GPs working independently or in small group practices. Private GPs are reimbursed at a set per-service fee by the national SHI fund(s).

Developed countries have varying policies regarding access to individual preventive services. Health systems in most countries provide vaccinations and mammography as part of funded health care services. In the United States, most insured individuals—and in Canada, most covered residents—automatically receive an annual physical exam including full blood profiles. In Norway and Denmark, adult physical exams are provided only upon special request by the individual, and in Sweden adult physical exams are provided only to pregnant women. In Sweden, adults who wish to know their cholesterol or PSA levels have begun to purchase blood tests out-of-pocket from private laboratories. Lack of physical exams and accompanying blood profiles may contribute to lower health care expenditures in the Nordic region.

Access to Elective Specialist Care

Approximately half of all European health care systems have a gatekeeping system that requires referrals from primary care physicians to book specialist visits (for publicly paid visits). In most tax-funded health systems (although not in most SHI systems), there are substantial waiting times, typically several months or more, for elective specialist appointments and high-tech diagnostic procedures, especially for cancer and other elective surgical or high-demand services. In England, a patient who requires a further consultation with a second specialist typically has to return to their primary care physician for a second referral, and then has to wait in the regular patient queue for that second appointment. In Finland, middle class families purchase separate private health insurance for their children to enable them to skip the long waiting times for primary and secondary pediatric health care services. More than 400,000 Finnish children have privately purchased policies.

There is also substantial waiting time for radiologic imaging services in most tax-funded systems. In Malta, the tax-funded health system's recent efforts to prioritize elective MRI investigations have succeeded in reducing waiting times from 18 months to 4 months. In both Alberta and British Columbia Provinces in

Canada, waiting time in 2016 for a publicly funded elective MRI is approximately 6 months, whereas privately paid MRIs are available in both provinces within 1 week.

This issue of waiting times in tax-funded health systems reflects a combination of growing demand (including increasing clinical indications), financial constraints, and insufficient capacity, including inadequate physician working hours. In the 1980s, when several surgical procedures for the elderly became more routine practice (e.g., hip replacement, coronary artery bypass graft, corneal lens implantation), the waiting list problem worsened. It had been mitigated somewhat by the early 2000s, only to return as a growing policy challenge once public sector financial resources became constrained after 2008. Timely cancer diagnosis and care have been a particularly sensitive issue, with tax-funded systems often taking several months for a patient to see an oncologist and then months more to begin treatment. In Sweden, a newspaper journalist set off a political storm in 2013 when he wrote extensively about women patients in one large county council (Malmo) who had to wait 47 days to receive the results from their breast cancer biopsy.

In response to patient anger in the early 2000s, a number of tax-funded health care systems introduced maximum waiting times for elective hospital procedures. (Most Western European SHI systems do not have long waiting times or treatment guarantees.) These maximum waiting times typically include initial primary care visits as well as specialist evaluations and treatment. In Denmark, a patient has the right to go to a different Danish public hospital for care after waiting 30 days without treatment. In Sweden, under the 2005 “waiting time guarantee,” an untreated patient’s local county council is required to pay for care in another county’s hospital after 180 days. Beginning in 1997, the European Union Court of Justice has slowly expanded the right of all EU citizens to travel to another EU country to receive “timely” care, with their home country health system required to pay for that care.

Long-Term Care Services

LTC (consisting of residential and home-based services) consumes a relatively small but increasing proportion of gross domestic product (GDP) in developed countries. In Sweden, LTC consumes 3.6% of GDP, mostly from public funds, whereas in Switzerland LTC services consume 2.1% of GDP, with only 0.8% of GDP coming from public funding. In the United States, total LTC expenditures represent 1.0% of total GDP, with 0.6% of GDP representing public funds, mostly from state-based Medicaid programs, which typically spend about 40% of their total funding on nursing home services. (Note that these figures do not include emergency, inpatient, or outpatient hospital costs generated by elderly patients.)

Since nursing home care is far more expensive than home care (nursing home care requiring the provision of housing, food, and around-the-clock care providers), government policymakers seek to keep the elderly and the chronically ill out of nursing homes for as long as feasible. Moreover, in developed countries like Sweden and Norway, some 70% of all home care services come from informal caregivers: spouses, children (typically daughters), neighbors, and friends. While some SHI systems (e.g., Germany) make available cash payments for LTC that can be used to compensate informal caregivers, most policymakers work hard to not monetize what is a large amount of essentially free care. Indeed, they actively seek to encourage those providing these services to continue to do so as long as possible, trying to postpone caregiver burnout by providing support

services such as free respite care, special call-in lines for caregiving advice, pension points toward retirement for the informal caregiver (Nordic countries), and free day-care center services.

In most tax-funded and SHI-funded European countries, home care services are organized at the municipal government level. In tax-funded systems, these services are also delivered mostly by municipal employees, working according to union-negotiated protocols. In some European SHI systems, and recently in tax-funded Sweden and Finland, private companies also provide home care services on contract to municipal governments. In combination with national legislation, these municipal systems also provide important support for informal caregivers, since the financial costs of caring for adults in their own home are substantially less than providing housing, food, and caregiver support in publicly funded homes for the aged or in nursing homes.

A high proportion of nursing homes in European tax-funded and SHI-funded health systems are publicly owned facilities operated by municipal governments; in some instances in SHI-funded systems (Israel, Netherlands), they are operated by private not-for-profit organizations. Recently, in some tax-funded systems (e.g., Sweden), private for-profit chains have begun to open nursing homes that are funded on a contract basis with local municipal governments. Costs for nursing home care can be expensive: in Norway, the cost per patient is often over $100,000 per year in a publicly funded home, with the patient responsible for paying up to 80% depending on the family's economic status. In Sweden, patients living in publicly funded nursing homes in Stockholm County pay a relatively small official fee, but they also pay room rent and up to 2706 Swedish Krona (SEK) per month (about $350 USD) for food out of their pensions.

In 2012, in an effort to reduce demand for expensive hospital and nursing home services, Norway and Denmark both began a number of elderly care reforms that shifted service delivery as well as funding responsibilities to municipal governments. Among innovations in Norway, municipalities are required to establish a municipal acute bed unit (MAU) to treat stable elderly patients and provide observation beds for evaluation. Partial funding for these units is provided by the four regional health care administrations. Some municipalities have also embedded primary care units inside their regional hospital to arrange discharge and to coordinate care for the chronically ill elderly. Norwegian municipalities are also responsible through their contracted (mostly private) primary care physicians to implement the National Pathways Program, which established treatment protocols for cross-sector conditions such as diabetes and cardiovascular conditions.

A differently configured structural innovation to better integrate LTC for the chronically ill elderly with clinical individual health services has been to consolidate both social and health care services within the same public administrative organization. In proposed 2019 health reforms in Finland, as well as a pilot decentralization program in England for 2.8 million people in Greater Manchester, social and health care programs are to be administered by a single responsible agency.

In the SHI-funded system in the Netherlands, almost 7% of the population live in a residential home. National government legislation revised the structure of nursing home funding and care in 2015. Three acts restructured the separate public LTC SHI fund, which requires mandatory payments by 100% of Dutch adults, and introduced delivery-related reforms that reduced the number and overall cost of nursing home patients

paid for by the fund. Determination of eligibility for public payment for nursing home care is now made by an independent national assessment body (the Centre for Needs Assessment). Moreover, municipal governments now play a stronger role in funding and delivering home care services. The reforms created social care teams that hold “kitchen table talks” to steer the elderly first toward seeking care from family, neighbors, churches, and other local community organizations before they qualify for publicly paid in-home care. In 2012, some 1.5 million people (12% of total population) provided informal care to ill or disabled persons, averaging 22 hours per week of care per person.

Home care recipients in the Netherlands can choose to set up a “personal budget,” using their public funding allocation to select their preferred individual care personnel (either publicly employed or publicly approved private providers). This arrangement also enables these home care recipients to determine the particular mix of services they want, as well as to augment the allocated public funds with personal funds. A number of innovative not-for-profit nursing homes have been created to provide additional services to elderly living in their neighborhood (primary care home visits), as well as terminal hospice care (e.g., the Saffier De Residentie Groep residences in The Haag).

In the United States, nursing home and home care are funded and delivered in a variety of different ways. For individuals who have minimal financial assets, nursing home costs are paid by a joint federal-regional (state) welfare program called Medicaid. Most state government Medicaid programs pay out more than 40% of their total budget for nursing home care. In the past, Medicaid did not pay for home care services. However, some states have programs with private for-profit and not-for-profit providers that provide home care as a way to forestall the need for the more expensive nursing home care.

Many private individuals take out private LTC insurance, typically from commercial insurance companies. These policies require individuals to make premium payments for years in advance (often 20 or more) before the individual learns whether they will, in fact, require home or nursing home care. Some private insurers have also raised premiums after individuals have paid in for many years and canceled policies if the new higher rate is not affordable. The 2010 ACA contained a new public LTC insurance program. However, the program was designed to be voluntary, and U.S. Department of Health and Human Services administrators decided not to implement that portion of the law.

In addition to the tax-funded Medicaid program, and privately purchased LTC insurance, many middle-class families pay for care from savings, by selling the elderly person’s home, or by direct contribution from children and other family members. Expenses can reach more than $60,000 per year depending on the location of a facility and who operates it.

Nursing home care in the United States is provided by a wide mix of private not-for-profit and for-profit providers, ranging from church-owned single-site homes to large stock market–listed companies. Many of these homes are purpose-built as assisted-living residences. There also are special units and facilities designed to care for the memory impaired. Home care services are delivered by a mix of private and not-for-profit and for-profit providers.

In Japan, a national LTC insurance fund was introduced in 2000. Although the new fund applies uniformly across the country, the program is administered by municipal governments and the premium level differs across municipalities, with an average monthly premium of 3000 yen (about $30 USD). In South Korea, an SHI fund for LTC is funded by mandatory contributions of 4.78% of a person's regular national health insurance contribution, with an additional 20% of total LTC expenditures provided by national government funds. The client copayment for home care is set at 15% of expenses and at 20% for residential care.

PHARMACEUTICALS

Pharmaceutical expenditures in developed countries (inpatient and outpatient combined) vary widely across different health system types, as well as between different countries within those different institutional types. OECD figures for 2014 show drug expenditures in tax-funded countries in Western Europe ranging from 6.7% of total health expenditures (THE) in Denmark to 12.2% of THE in the United Kingdom and 17.9% of THE in Spain. In SHI-funded Western European systems, pharmaceuticals absorbed 7.6% of THE in the Netherlands, while in Germany that figure was 14.5%. In the hybrid tax-funded SHI systems of Central Europe, pharmaceuticals were much higher: 18% of THE in Estonia to 30.2% of THE in Hungary. In Asian SHI systems, pharmaceuticals consumed 20.6% of THE in South Korea and 21% of THE (in 2012) in Japan. The OECD's 2014 figures for pharmaceutical spending in North America are 12.3% of THE in the United States and 17.2% in Canada.

Contributing factors to this wide-ranging variation are (1) the ratio problem (relatively fixed level of pharmaceutical costs due to international prices—the numerator—divided by a greatly varying per-capita health expenditure cost in different developed country health systems); (2) the range and type of pharmaceutical price controls in each country; and (3) the degree of limitation placed on pharmaceutical supply, tied to formularies and/or explicit forms of drug rationing.

Most European health systems have tight national controls on the cost and, in some tax-based countries, on the availability of pharmaceuticals. Most European countries also use a number of different regulatory measures to limit prices and/or availability of both inpatient and outpatient drugs, including mandatory generic prescribing, reference pricing, patient copays (sometimes with an annual ceiling, after which copayments are no longer required), and (particularly in tax-funded systems) national formularies tied to clinical effectiveness. (Norway, for example, allows only about 2300 different preparations—including dosage, delivery method, and box size—to be stocked by pharmacies.) Prices for drugs can vary considerably across different European countries, tied to economic development and domestic pricing patterns. One consequence of these differential national pricing controls has been the development of a parallel import market, in which drug wholesalers and pharmacists in the more expensive countries purchase supplies from a cheaper market elsewhere in Europe.

Access to expensive drugs has also been intentionally limited in some tax-funded health systems in Europe. One basis for rationing, as noted above, has been rationing tied to QALYs (quality-adjusted life-years). Rationing also reflects a clash between strained public drug budgets and public pressure. For example, in the

case of cancer drugs in England, the recommendation of NICE against funding the breast cancer drug trastuzumab (Herceptin) was subsequently overturned by the Minister of Health. Expensive cancer drugs continue to be rationed in England where the NHS Cancer Drug Fund, established to provide access on a case-by-case basis, ran out of funds and was forced to close down for 3 months to restructure its operations.

As part of the medical tradition in Asian countries, office-based physicians fill prescriptions as well as prescribing drugs to patients. These sales serve to supplement their income in the setting of relatively low per-visit payments from state-run SHI funds. Korea has now implemented restrictions on these office-based sales. Japan has attempted to reduce physician sale of pharmaceuticals by various changes in reimbursement rates, reducing the total percentage of physician-sold pharmaceuticals to 42.8% of all outpatient prescriptions.

GOVERNANCE AND REGULATION

Health care services in developed countries are steered, constrained, monitored, and (to varying degrees) assessed by governments and governmentally established and/or empowered bodies. Although these measures apply particularly to the financial efficiency of government-funded services, they also seek to promote patient and community safety, equity of access, and high-quality clinical outcomes. This oversight is often strongly focused on privately operated and contracted providers and insurers, although in principle it applies to publicly operated organizations as well.

Governance consists of macro national-level policy, meso institutional-level management, and micro clinic-level care decisions. This complex mix of governance decisions is often shared among different national, regional, and local governments, depending on the degree of centralization, decentralization, or, recently, recentralization (e.g., Norway). While most systems officially prioritize “good governance,” governance activities frequently comingle with political objectives as core policy concepts are developed and transformed into concrete organizational targets.

In Sweden, health system governance is shared among national, regional (county), and local municipal governments. The national government has responsibility to pass “frame” legislation, which establishes the basic structure of the system. To cite one example, until recently, the national government had limited an adult patient’s total copayments for outpatient physician (specialist and primary care) and pharmaceuticals to 2800 SEK (about $350 USD) for a 12-month period. The 20 regional governments, in turn, made policy decisions within that legislation, deciding how to apportion the specific copayments for each primary care and specialist outpatient visit. Since Swedes can self-refer to specialists, some counties double the copayment to hospital-based doctors to discourage unnecessary appointments. Similarly, fiscal policy normally is shared between the regional government, which raises about 70% of total health expenditures through its own county-set flat income tax, and the national government, which provides additional purpose-tied funds for national objectives such as consolidating open-heart surgery across county lines and balancing lower tax receipts in rural counties with smaller working populations. However, this normal funding relationship across governments can change. In the early 1990s, the national government placed a

"stop" on raising county taxes prior to Sweden's admission in 1995 to the European Union. In 2016, each of the 20 counties could set their own ceilings, which were almost all at 3300 SEK (about $370 USD).

In Spain's tax-funded health system (70.9% publicly funded), 17 regional "autonomous communities" were given full managerial responsibility for the provision of health services in a decentralization process, along with ownership of all publicly owned hospitals. The national government generates a substantial proportion of health care resources, which are included in the broad block grants it allocates to the regional governments, which then add regional tax revenue to make up the full public-sector budget. In a mechanism to further influence operating policy, the national Spanish government established a joint federal-regional council to review quality and performance data (through the 2003 Health System Cohesion and Quality Act). Italy's tax-funded health system (75.6% publicly funded) is similarly operated by 20 regional governments, which pay for the publicly operated system through a complicated mix of national, and nationally stipulated but regionally collected, taxes. Again like Spain, the national government established a federal-regional government council, which seeks to coordinate care standards and information among the regions and with national government agencies.

In Germany, where funding for the health system is formally the responsibility of 132 private not-for-profit sickness funds, governance decisions are shared among these private sector sickness funds and public-sector national, regional, and municipal governments. The sickness funds receive a risk-adjusted premium payment for each enrolled individual, determined by a national government–determined formula, and from a national government–run health insurance pool. Most hospitals are owned and operated by municipal governments, while investment capital for structural renovations and new building comes from the 16 regional Länder taken from their tax revenues. Payment frameworks and amounts for public hospitals are negotiated between associations of these municipally owned hospitals and associations of the private sickness funds, without formal government participation.

Regulation is an essential element of an effective health care system and a key component of overall health system governance. Regulation incorporates both broad standard requirements that affect all organizations that operate in a country (e.g., hiring, firing, and wage decisions) as well as specific health sector–related regulations (e.g., proper handling, use, and disposal of low-grade nuclear waste from radiation treatments). Recent examples of health sector regulation in England, for example, include the following:

1. Requiring all cancer drugs adopted for use in the NHS to cost no more than £30,000/QALY;
2. Requiring in their employment contract that junior doctors in hospitals work a specific number of Sundays; and
3. Requiring that all emergency department patients receive care within 4 h of their arrival.

A powerful tool that has the force of law, regulation can have substantial negative as well as positive effects. A well-known political science corollary of regulatory power is that "the right to regulate is also the right to destroy." For example, in the United States, the federal Environmental Protection Agency, as part of its pursuit of cleaner air, issued wide-ranging regulatory orders setting performance standards that resulted in

the closing of many West Virginia coal mines, resulting in the loss of tens of millions of dollars of productive capacity and thousands of high-paying jobs. Similarly, in some tax-funded European systems, such as those in Sweden and England, there is growing pressure from public health advocates to prohibit the making of a profit from publicly paid funds. In Sweden, the national government's Reepalu report honored a pledge made by the Social Democratic government to its Left (socialist) Party ally by calling for a legislated ban on profit-making in the provision of publicly funded health care services. The Report's publication resulted in substantial divestment of existing investor-owned primary care, nursing home, and home care companies.

FUTURE CHALLENGES

Health systems in developed countries face serious challenges in the coming years. These include financial, organizational, and policy dilemmas for which institutionally viable, financially sustainable, and politically supportable solutions will be complicated to develop and difficult to implement. On the delivery side, a key question is whether privately structured GP-based primary care is more efficient and effective than various clinic-based forms of primary care services. Recent movement in Northern and Central Europe toward more private GPs, along with continued private office–based primary care in much of Canada, the United States, and economically developed countries in Asia, raises complex policy issues for international organizations like the World Health Organization (WHO), as well as national policymakers. In the hospital sector, existing levels of clinical quality and patient responsiveness in publicly operated command-and-control institutions will increasingly have to compete with those of semi-autonomous public hospitals, as well as various types of private, sometimes very innovative providers. In the financing arena, continued pressure on publicly raised health system revenues is likely to erode longtime commitments in some tax-funded health systems to minimal patient copayments and low out-of-pocket funding.

An additional set of challenges will arise from recent commitments by international organizations like WHO to restructure health systems in developed countries to better address the social determinants of health. This new, incomplete strategy calls for a dramatic expansion of health sector responsibility to include a wide range of existing institutional arrangements in housing, education, work-life, and social and political decision-making. The influential 2010 Strategic Review of Health Inequalities in England entitled "Fair Society, Healthy Lives," led by Sir Michael Marmot, a British epidemiologist, called for the elimination of all "inequities in power, money, and resources." Separate from the political dimensions of this proposed new paradigm, how such fundamental societal change will be funded has yet to be addressed.

Looking forward, among the most essential challenges to national decision-makers in the coming period will be four specific health system imperatives:

1. **Finding a more sustainable balance between ethics and funding.**

   Policymakers in publicly funded health systems face a growing gap between patient expectations of high-quality clinical care, staff expectations of better compensation, and the economic imperative of no new taxes. While the present solidaristic foundation for raising collective revenues is insufficient, available non-solidaristic tools (copayments, supplemental insurance, private pay) inevitably contribute to overall

inequality. But what then are the realistic policy alternatives? The minimalist new policy goal necessarily will have to become one of raising new revenues while doing the least economic and social harm.

2. **Developing better strategies to steer provider diversity.**

   Health systems in developed countries are becoming more diverse with more and different types of public owners: hospital trusts, state enterprises, and mixed public-private hospital owners/managers. There also are more and different types of private providers: not-for-profit community groups, foundations, and cooperatives, as well as for-profit small local entrepreneurs, large international companies, and risk capital funds (venture capital). Furthermore, new innovative delivery models are reorganizing traditional service boundaries: not-for-profit private nursing homes in the Netherlands also provide outpatient primary care to neighborhood elderly patients, as well as hospice care; Israeli technology companies combine high-tech home-based patient monitoring with standard medical and custodial home care services. Public pressure from citizens for more choice and better outcomes will pressure policymakers toward new, more accommodative health system arrangements.

3. **Ensuring better coordination between social and health services.**

   Tax-funded and SHI-funded systems alike are under intense policy pressure to develop better strategies to integrate services for the chronically ill elderly, as a way to improve the quality of services that these patients receive and to keep them at home healthier and longer, reducing expensive acute visits to hospitals and emergency departments. The clear delivery system goal will increasingly be to keep the elderly out of nursing homes and acute care facilities for as long as possible.

4. **Building labor unions into provider innovation.**

   In many developed countries, health sector staff, including hospital physicians, are members of labor unions. Effective policymaking will require finding mechanisms to build these personnel unions into accelerated health system restructuring processes. This process will necessarily involve integrating unions into more innovative, flexible, fiscally sustainable organizational arrangements with contracts that reward active participation in organizational change, contracts that pay incentives to more productive employees, quicker reassignment and redundancy procedures (firing health sector workers can take a year or longer in some European health systems), and establishing profit-sharing payments to teams/unions, also in public sector organizations.

While the structure and complexity of resolving these specific organizational challenges will vary depending on a country's cultural and institutional context, the commonality of these problems suggests that health systems in the developed world will require a new, broader range of targeted policy strategies and solutions.

FURTHER READING

Figueras J, McKee M (eds): *Health Systems, Health, Wealth, and Societal Well-Being: Assessing the Case for Investing in Health Systems*. Maidenhead, Open University Press/McGraw-Hill Education, 2011. *www.euro.who.int/__data/assets/pdf_file/0007/164383/e96159.pdf*.

Haseltine W: *Affordable Excellence: The Singapore Health Story*. Washington, Brookings Institution Press, 2013. *www.brookings.edu/wp-www.brookings.edu/wp-content/uploads/2016/07/AffordableExcellencePDF.pdf*.

Kuhlmann E et al (eds): *The Palgrave International Handbook on Healthcare Policy and Governance*. London, Palgrave MacMillan, 2015.

Oliver E, Mossialos E: European health systems reforms: Looking backward to see forward? J Health Polit Policy Law 30:7, 2005. [PubMed: 15943385]

Rice T et al: *United States of America: Health System Review*. Health in Transition (HiT) Series 15 (3). Brussels, European Observatory on Health Systems and Policies, 2013. *www.euro.who.int/__data/assets/pdf_file/0019/215155/HiT-United-States-of-America.pdf*.

Chapter 6: The Safety and Quality of Health Care

David W. Bates

INTRODUCTION

Safety and quality are two of the central dimensions of health care. In recent years it has become easier to measure safety and quality, and it is increasingly clear that performance in both dimensions could be much better. The public is—with good justification—demanding measurement and accountability, and payment for services will increasingly be based on performance in these areas. Thus, physicians must learn about these two domains, how they can be improved, and the relative strengths and limitations of the current ability to measure them.

Safety and quality are closely related but do not completely overlap. The Institute of Medicine has suggested in a seminal series of reports that safety is the first part of quality and that the health care system must first and foremost guarantee that it will deliver safe care, although quality is also pivotal. In the end, it is likely that more net clinical benefit will be derived from improving quality than from improving safety, though both are important and safety is in many ways more tangible to the public. The first section of this chapter will address issues relating to the safety of care and the second will cover quality of care.

SAFETY IN HEALTH CARE

Safety Theory and Systems Theory

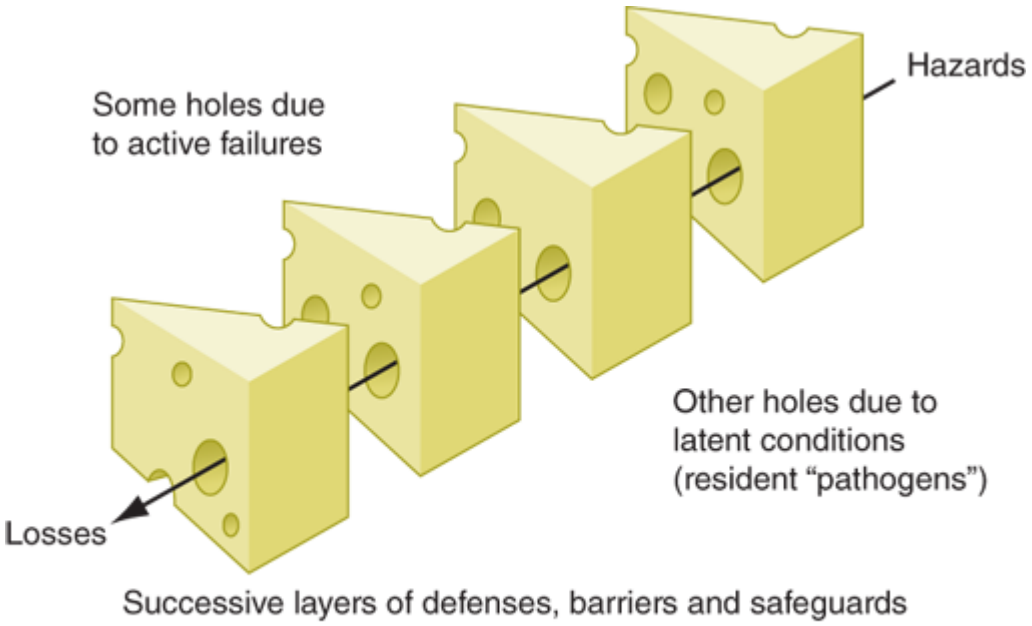
Safety theory clearly points out that individuals make errors all the time. Think of driving home from the hospital: you intend to stop and pick up a quart of milk on the way home but find yourself entering your driveway without realizing how you got there. Everybody uses low-level, semiautomatic behavior for many activities in daily life; this kind of error is called a *slip*. Slips occur often during care delivery—e.g., when people intend to write an order but forget because they have to complete another action first. *Mistakes*, by contrast, are errors of a higher level; they occur in new or nonstereotypic situations in which conscious decisions are being made. An example would be dosing of a medication with which a physician is not familiar. The strategies used to prevent slips and mistakes are often different.

Systems theory suggests that most accidents occur as the result of a series of small failures that happen to line up in an individual instance so that an accident can occur **(Fig. 6-1)**. It also suggests that most individuals in an industry such as health care are trying to do the right thing (e.g., deliver safe care) and that most

accidents thus can be seen as resulting from defects in systems. Systems should be designed both to make errors less likely and to identify those that do inevitably occur.

FIGURE 6-1

"Swiss cheese" diagram. Reason argues that most accidents occur when a series of "latent failures" are present in a system and happen to line up in a given instance, resulting in an accident. Examples of latent failures in the case of a fall might be that the unit is unusually busy and the floor happens to be wet. *(Adapted from J Reason: BMJ 320:768, 2000; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Factors that Increase the Likelihood of Errors

Many factors ubiquitous in health care systems can increase the likelihood of errors, including fatigue, stress, interruptions, complexity, and transitions. The effects of fatigue in other industries are clear, but its effects in health care have been more controversial until recently. For example, the accident rate among truck drivers increases dramatically if they work over a certain number of hours in a week, especially with prolonged shifts. A recent study of house officers in the intensive care unit demonstrated that they were about one-third more likely to make errors when they were on a 24-h shift than when they were on a schedule that allowed them to sleep 8 h the previous night. The American College of Graduate Medical Education has moved to address this issue by putting in place the 80-h workweek. Although this stipulation is a step forward, it does not address the most important cause of fatigue-related errors: extended-duty shifts. High levels of stress and heavy workloads also can increase error rates. Thus, in extremely high-pressure situations, such as cardiac arrests, errors are more likely to occur. Strategies such as using protocols in these settings can be helpful, as can simple recognition that the situation is stressful.

Interruptions also increase the likelihood of error and occur frequently in health care delivery. It is common to forget to complete an action when one is interrupted partway through it by a page, for example.

Approaches that may be helpful in this area include minimizing interruptions and setting up tools that help define the urgency of an interruption.

Complexity represents a key issue that contributes to errors. Providers are confronted by streams of data (e.g., laboratory tests and vital signs), many of which provide little useful information but some of which are important and require action or suggest a specific diagnosis. Tools that emphasize specific abnormalities or combinations of abnormalities may be helpful in this area.

Transitions between providers and settings are also common in health care, especially with the advent of the 80-h workweek, and generally represent points of vulnerability. Tools that provide structure in exchanging information—for example, when transferring care between providers—may be helpful.

The Frequency of Adverse Events in Health Care

Most large studies focusing on the frequency and consequences of adverse events have been performed in the inpatient setting; some data are available for nursing homes, but much less information is available about the outpatient setting. The Harvard Medical Practice Study, one of the largest studies to address this issue, was performed with hospitalized patients in New York. The primary outcome was the adverse event: an injury caused by medical management rather than by the patient's underlying disease. In this study, an event either resulted in death or disability at discharge or prolonged the length of hospital stay by at least 2 days. Key findings were that the adverse event rate was 3.7% and that 58% of the adverse events were considered preventable. Although New York is not representative of the United States as a whole, the study was replicated later in Colorado and Utah, where the rates were essentially similar. Since then, other studies using analogous methodologies have been performed in various developed nations, and the rates of adverse events in these countries appear to be ~10%. Rates of safety issues appear to be even higher in developing and transitional countries; thus, this is clearly an issue of global proportions. The World Health Organization has focused on this area, forming the World Alliance for Patient Safety.

In the Harvard Medical Practice Study, adverse drug events (ADEs) were most common, accounting for 19% of all adverse events, and were followed in frequency by wound infections (14%) and technical complications (13%). Almost half of adverse events were associated with a surgical procedure. Among nonoperative events, 37% were ADEs, 15% were diagnostic mishaps, 14% were therapeutic mishaps, 13% were procedure-related mishaps, and 5% were falls.

ADEs have been studied more than any other error category. Studies focusing specifically on ADEs have found that they appear to be much more common than was suggested by the Harvard Medical Practice Study, although most other studies use more inclusive criteria. Detection approaches in the research setting include chart review and the use of a computerized ADE monitor, a tool that explores the database and identifies signals that suggest an ADE may have occurred. Studies that use multiple approaches find more ADEs than does any individual approach, and this discrepancy suggests that the true underlying rate in the population is higher than would be identified by a single approach. About 6–10% of patients admitted to U.S. hospitals experience an ADE.

Injuries caused by drugs are also common in the outpatient setting. One study found a rate of 21 ADEs per every 100 patients per year when patients were called to assess whether they had had a problem with one of their medications. The severity level was lower than in the inpatient setting, but approximately one-third of these ADEs were preventable.

The period immediately after a patient is discharged from the hospital appears to be very risky. A recent study of patients hospitalized on a medical service found an adverse event rate of 19%; about one-third of those events were preventable, and another one-third were ameliorable (i.e., they could have been made less severe). ADEs were the single leading error category.

Prevention Strategies

Most work on strategies to prevent adverse events has targeted specific types of events in the inpatient setting, with nosocomial infections and ADEs having received the most attention. Nosocomial infection rates have been reduced greatly in intensive care settings, especially through the use of checklists. For ADEs, several strategies have been found to reduce the medication error rate, although it has been harder to demonstrate that they reduce the ADE rate overall, and no studies with adequate power to show a clinically meaningful reduction have been published.

Implementation of checklists to ensure that specific actions are carried out has had a major impact on rates of catheter-associated bloodstream infection and ventilator-associated pneumonia, two of the most serious complications occurring in intensive care units. The checklist concept is based on the premise that several specific actions can reduce the frequency of these issues; when these actions are all taken for every patient, the result has been an extreme reduction in the frequency of the associated complication. These practices have been disseminated across wide areas, in particular in the state of Michigan.

Computerized physician order entry (CPOE) linked with clinical decision support reduces the rate of serious medication errors, defined as those that harm someone or have the potential to do so. In one study, CPOE, even with limited decision support, decreased the serious medication error rate by 55%. CPOE can prevent medication errors by suggesting a default dose, ensuring that all orders are complete (e.g., that they include dose, route, and frequency), and checking orders for allergies, drug–drug interactions, and drug–laboratory issues. In addition, clinical decision support can suggest the right dose for a patient, tailoring it to level of renal function and age. In one study, patients with renal insufficiency received the appropriate dose only one-third of the time without decision support, whereas that fraction increased to approximately two-thirds with decision support; moreover, with such support, patients with renal insufficiency were discharged from the hospital half a day earlier. As of 2017, over 90% of U.S. hospitals have implemented CPOE, though the decision support often is still limited.

Another technology that can improve medication safety is bar coding linked with an electronic medication administration record. Bar coding can help ensure that the right patient gets the right medication at the right time. Electronic medication administration records can make it much easier to determine what medications a patient has received. Studies to assess the impact of bar coding on medication safety are under way, and

the early results are promising. Another technology to improve medication safety is “smart pumps.” These pumps can be set according to which medication is being given and at what dose; the health care professional will receive a warning if too high a dose is about to be administered.

The National Safety Picture

Several organizations, including the National Quality Forum and the Joint Commission, have made recommendations for improving safety. In particular, the National Quality Forum has released recommendations to U.S. hospitals about what practices will most improve the safety of care, and all hospitals are expected to implement these recommendations. Many of these practices arise frequently in routine care. One example is “readback,” the practice of recording all verbal orders and immediately reading them back to the physician to verify the accuracy of what was heard. Another is the consistent use of standard abbreviations and dose designations; some abbreviations and dose designations are particularly prone to error (e.g., 7U may be read as 70).

Measurement of Safety

Measuring the safety of care is difficult and expensive, since adverse events are, fortunately, rare. Most hospitals rely on spontaneous reporting to identify errors and adverse events, but the sensitivity of this approach is very low, with only ~1 in 20 ADEs reported. Promising research techniques involve searching the electronic record for signals suggesting that an adverse event has occurred. These methods are not yet in wide use but will probably be used routinely in the future. Claims data have been used to identify the frequency of adverse events; this approach works much better for surgical care than for medical care and requires additional validation. The net result is that, except for a few specific types of events (e.g., falls and nosocomial infections), hospitals have little idea about the true frequency of safety issues.

Nonetheless, all providers have the responsibility to report problems with safety as they are identified. All hospitals have spontaneous reporting systems, and, if providers report events as they occur, those events can serve as lessons for subsequent improvement.

Conclusions about Safety

It is abundantly clear that the safety of health care can be improved substantially. As more areas are studied closely, more problems are identified. Much more is known about the epidemiology of safety in the inpatient setting than in outpatient settings. A number of effective strategies for improving inpatient safety have been identified and are increasingly being applied. Some effective strategies are also available for the outpatient setting. Transitions appear to be especially risky. The solutions to improving care often entails the consistent use of systematic techniques such as checklists and often involves leveraging of information technology. Nevertheless, solutions will also include many other domains, such human factors techniques, team training, and a culture of safety.

QUALITY IN HEALTH CARE

Assessment of quality of care has remained somewhat elusive, although the tools for this purpose have increasingly improved. Selection of health care and measurement of its quality are components of a complex process.

Quality Theory

Donabedian has suggested that quality of care can be categorized by type of measurement into structure, process, and outcome. *Structure* refers to whether a particular characteristic is applicable in a particular setting—e.g., whether a hospital has a catheterization laboratory or whether a clinic uses an electronic health record. *Process* refers to the way care is delivered; examples of process measures are whether a Pap smear was performed at the recommended interval or whether an aspirin was given to a patient with a suspected myocardial infarction. *Outcome* refers to what actually happens—e.g., the mortality rate in myocardial infarction. It is important to note that good structure and process do not always result in a good outcome. For instance, a patient may present with a suspected myocardial infarction to an institution with a catheterization laboratory and receive recommended care, including aspirin, but still die because of the infarction.

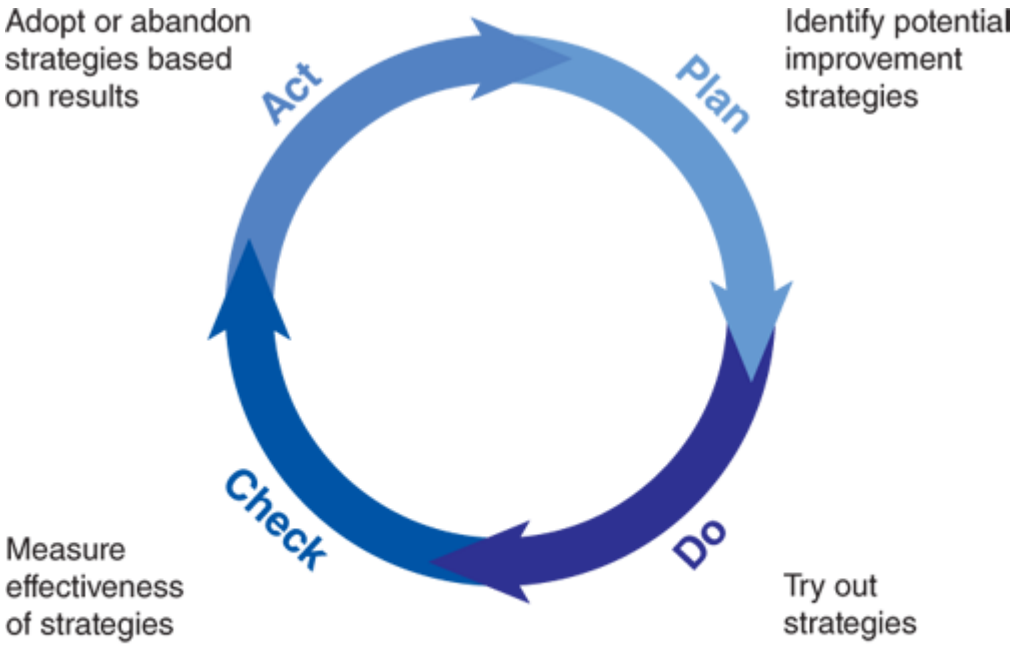
Quality theory also suggests that overall quality will be improved more in the aggregate if the performance level of all providers is raised rather than if a few poor performers are identified and punished. This view suggests that systems changes are especially likely to be helpful in improving quality, since large numbers of providers may be affected simultaneously.

The theory of *continuous quality improvement* suggests that organizations should be evaluating the care they deliver on an ongoing basis and continually making small changes to improve their individual processes. This approach can be very powerful if embraced over time.

A number of specific tools have been developed to help improve process performance. One of the most important is the Plan-Do-Check-Act cycle **(Fig. 6-2)**. This approach can be used for "rapid cycle" improvement of a process—e.g., the time that elapses between a diagnosis of pneumonia and administration of antibiotics to the patient. Specific statistical tools, such as control charts, are often used in conjunction to determine whether progress is being made. Because most medical care includes one or many processes, this tool is especially important for improvement.

FIGURE 6-2

Plan-Do-Check-Act cycle. This approach can be used to improve a specific process rapidly. First, planning is undertaken, and several potential improvement strategies are identified. Next, these strategies are evaluated in small "tests of change." "Checking" entails measuring whether the strategies have appeared to make a difference, and "acting" refers to acting on the results.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Factors Relating to Quality

Many factors can decrease the level of quality, including stress to providers, high or low levels of production pressure, and poor systems. Stress can have an adverse effect on quality because it can lead providers to omit important steps, as can a high level of production pressure. Low levels of production pressure sometimes can result in worse quality, as providers may be bored or have little experience with a specific problem. Poor systems can have a tremendous impact on quality, and even extremely dedicated providers typically cannot achieve high levels of performance if they are operating within a poor system.

Data about the Current State of Quality

A study published by the RAND Corporation in 2006 provided the most complete picture of quality of care delivered in the United States to date. The results were sobering. The authors found that, across a wide range of quality parameters, patients in the United States received only 55% of recommended care overall; there was little variation by subtype, with scores of 54% for preventive care, 54% for acute care, and 56% for care of chronic conditions. The authors concluded that, in broad terms, the chances of getting high-quality care in the United States were little better than those of winning a coin flip.

Work from the Dartmouth Atlas of Health Care evaluating geographic variation in use and quality of care demonstrates that, despite large variations in utilization, there is no positive correlation between the two variables at the regional level. An array of data demonstrate, however, that providers with larger volumes for specific conditions, especially for surgical conditions, do have better outcomes.

Strategies for Improving Quality and Performance

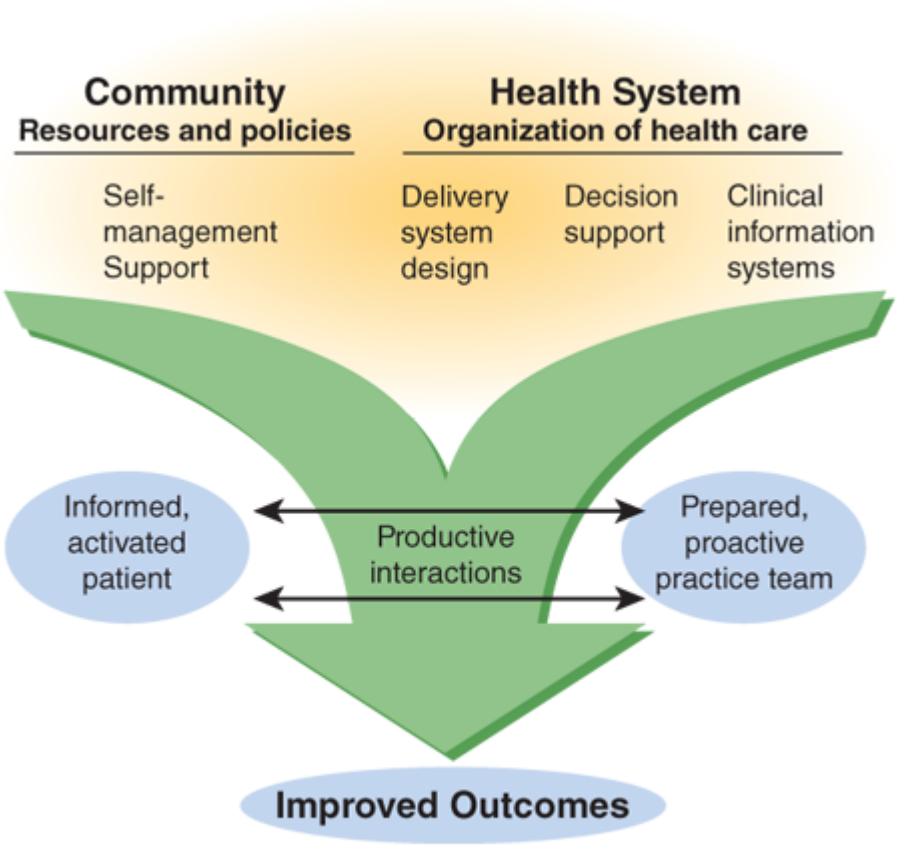
A number of specific strategies can be used to improve quality at the individual level, including rationing, education, feedback, incentives, and penalties. *Rationing* has been effective in some specific areas, such as persuading physicians to prescribe within a formulary, but it generally has been resisted. *Education* is effective in the short run and is necessary for changing opinions, but its effect decays fairly rapidly with time. *Feedback* on performance can be given at either the group or the individual level. Feedback is most effective if it is individualized and is given in close temporal proximity to the original events. *Incentives* can be effective, and many believe that they will prove to be a key to improving quality, especially if pay-for-performance with sufficient incentives is broadly implemented (see below). *Penalties* produce provider resentment and are rarely used in health care.

Another set of strategies for improving quality involves changing the systems of care. An example would be introducing reminders about which specific actions needed to be taken at a visit for a specific patient—a strategy that has been demonstrated to improve performance in certain situations, such as the delivery of preventive services. Another approach that has been effective is the development of "bundles" or groups of quality measures that can be implemented together with a high degree of fidelity. A number of hospitals have implemented a bundle for ventilator-associated pneumonia in the intensive care unit that includes five measures (e.g., ensuring that the head of the bed is elevated). These hospitals have been able to improve performance substantially.

Perhaps the most pressing need is to improve the quality of care for chronic diseases. The Chronic Care Model has been developed by Wagner and colleagues **(Fig. 6-3)**; it suggests that a combination of strategies is necessary (including self-management support, changes in delivery system design, decision support, and information systems) and that these strategies must be delivered by a practice team composed of several providers, not just a physician.

FIGURE 6-3

The Chronic Care Model, which focuses on improving care for chronic diseases, suggests that (1) delivery of high-quality care requires a range of strategies that must closely involve and engage the patient and (2) team care is essential. *(From EH Wagner et al: Eff Clin Pract 1:2, 1998.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Available evidence about the relative efficacy of strategies in reducing hemoglobin A1c (HbA1c) in outpatient diabetes care supports this general premise. It is especially notable that the outcome was the HbA1c level, as it has generally been much more difficult to improve outcome measures than process measures (such as whether HbA1c was measured). In this meta-analysis, a variety of strategies were effective, but the most effective ones were the use of team changes and the use of a case manager. When cost-effectiveness is considered in addition, it appears likely that an amalgam of strategies will be needed. However, the more expensive strategies, such as the use of case managers, probably will be implemented widely only if pay-for-performance takes hold.

National State of Quality Measurement

In the inpatient setting, quality measurement is now being performed by a very large proportion of hospitals for several conditions, including myocardial infarction, congestive heart failure, pneumonia, and surgical infection prevention; 20 measures are included in all. This is the result of the Hospital Quality Initiative, which represents a collaboration among many entities, including the Hospital Quality Alliance, the Joint Commission, the National Quality Forum, and the Agency for Healthcare Research and Quality. The data are housed at the Center for Medicare and Medicaid Services, which publicly releases performance data on the measures on a website called *Hospital Compare* (*www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalCompare.html*). These data are reported voluntarily and are available for a very high proportion of the nation's hospitals. Analyses demonstrate substantial regional variation in quality and important differences among hospitals. Analyses by the Joint Commission

for similar indicators reveal that performance on measures by hospitals has improved over time and that, as might be hoped, lower performers have improved more than higher performers.

Public Reporting

Overall, public reporting of quality data is becoming increasingly common. There are now commercial websites that have quality-related data for most regions of the United States, and these data can be accessed for a fee. Similarly, national data for hospitals are available. The evidence to date indicates that patients have not made much use of such data, but that the data have had an important effect on provider and organization behavior. Instead, patients have relied on provider reputation to make choices, partly because little information was available until very recently and the information that was available was not necessarily presented in ways that were easy for patients to access. Many authorities think that, as more information about quality becomes available, it will become increasingly central to patients' choices about where to access care.

Pay-for-Performance

Currently, providers in the United States get paid exactly the same amount for a specific service, regardless of the quality of care delivered. The pay-for-performance theory suggests that, if providers are paid more for higher-quality care, they will invest in strategies that enable them to deliver that care. The current key issues in the pay-for-performance debate relate to (1) how effective it is, (2) what levels of incentives are needed, and (3) what perverse consequences are produced. The evidence on effectiveness is fairly limited, although a number of studies are ongoing. With respect to incentive levels, most quality-based performance incentives have accounted for merely 1–2% of total payment in the United States to date. In the United Kingdom, however, 40% of general practitioners' salaries have been placed at risk according to performance across a wide array of parameters; this approach has been associated with substantial improvements in reported quality performance, although it is still unclear to what extent this change represents better performance versus better reporting. The potential for perverse consequences exists with any incentive scheme. One problem is that, if incentives are tied to outcomes, there may be a tendency to transfer the sickest patients to other providers and systems. Another concern is that providers will pay too much attention to quality measures with incentives and ignore the rest of the quality picture. The validity of these concerns remains to be determined. Nonetheless, it appears likely that, under health care reform, the use of various pay-for-performance schemes is likely to increase.

CONCLUSIONS

The safety and quality of care in the United States could be improved substantially. A number of available interventions have been shown to improve the safety of care and should be used more widely; others are undergoing evaluation or soon will be. Quality also could be dramatically better, and the science of quality improvement continues to mature. Implementation of value-based approaches such as accountable care which include pay-for-performance related to safety and quality should make it much easier for

organizations to justify investments in improving safety and quality parameters, including health information technology. However, many improvements will also require changing the structure of care—e.g., moving to a more team-oriented approach and ensuring that patients are more involved in their own care. Payment reform focusing on value seems very likely to progress and will likely include both positive incentives and penalties related to safety and quality performance. Measures of safety are still relatively immature and could be made much more robust; it would be particularly useful if organizations had measures they could use in routine operations to assess safety at a reasonable cost, and substantial research is addressing this. Although the quality measures available are more robust than those for safety, they still cover a relatively small proportion of the entire domain of quality, and more measures need to be developed. The public and payers are demanding better information about safety and quality as well as better performance in these areas. The clear implication is that these domains will have to be addressed directly by providers.

FURTHER READING

Bates DW et al: Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. JAMA 280:1311, 1998. [PubMed: 9794308]

Berwick DM: Era 3 for medicine and health care. JAMA 315:1329, 2016. [PubMed: 26940610]

Brennan TA et al: Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med 324:370, 1991. [PubMed: 1987460]

Chertow GM et al: Guided medication dosing for inpatients with renal insufficiency. JAMA 286:2839, 2001. [PubMed: 11735759]

Institute of Medicine. Crossing the quality chasm: A new health system for the 21st century. 2001. http://nationalacademies.org/hmd/~/media/Files/Report%20Files/2001/Crossing-the-Quality-Chasm/Quality%20Chasm%202001%20%20report%20brief.pdf.

Institute of Medicine. Report: To err is human: Building a safer health system. 1999. http://www.nationalacademies.org/hmd/~/media/Files/Report%20Files/1999/To-Err-is-Human/To%20Err%20is%20Human%201999%20%20report%20brief.pdf.

Landrigan C et al: Effect of reducing interns' work hours on serious medical errors in intensive care units. N Engl J Med 351:1838, 2004. [PubMed: 15509817]

McGlynn et al: The quality of health care delivered to adults in the United States. N Engl J Med 348:2635, 2003. [PubMed: 12826639]

Pronovost P et al: An intervention to decrease catheter-related bloodstream infections in the ICU. N Engl J Med 355:2725, 2006. Erratum in: N Engl J Med 356:2660, 2007. [PubMed: 17192537]

Starmer AJ et al: Rates of medical errors and preventable adverse events among hospitalized children following implementation of a resident handoff bundle. JAMA 310:2262, 2013. [PubMed: 24302089]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 7: Racial and Ethnic Disparities in Health Care

Joseph R. Betancourt; Alexander R. Green

INTRODUCTION

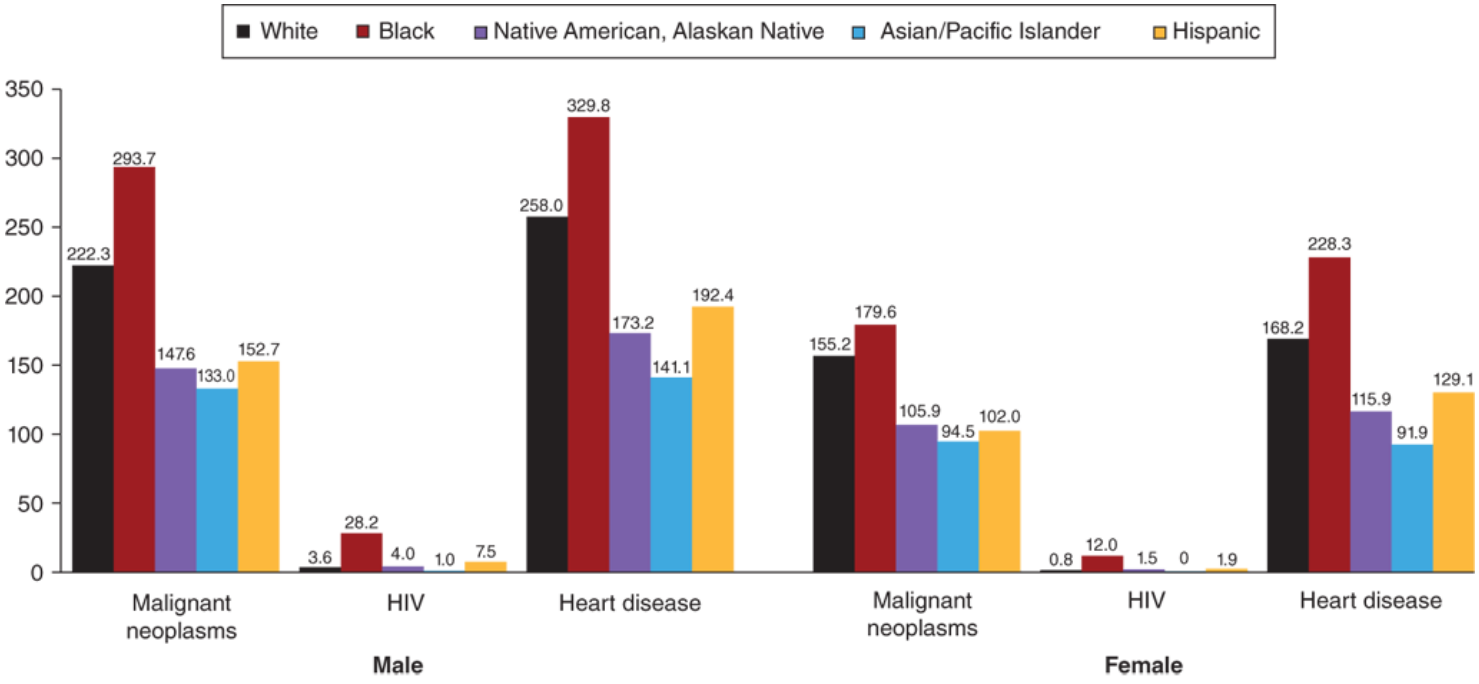
Over the course of its history, the United States has experienced dramatic improvements in overall health and life expectancy, largely as a result of initiatives in public health, health promotion, disease prevention, and chronic care management. Our ability to prevent, detect, and treat diseases in their early stages has allowed us to target and reduce rates of morbidity and mortality. Despite interventions that have improved the overall health of the majority of Americans, racial and ethnic minorities (blacks, Hispanics/Latinos, Native Americans/Alaskan Natives, Asian/Pacific Islanders) have benefited less from these advances than whites and have suffered poorer health outcomes from many major diseases, including cardiovascular disease, cancer, and diabetes. These disparities highlight the importance of recognizing and addressing the *social determinants of health*, which contribute enormously to health outcomes. Research has revealed that minorities may receive less care and lower-quality care than whites, even when confounders such as stage of presentation, comorbidities, and health insurance are controlled. These differences in quality are called *racial and ethnic disparities in health care*. These health care disparities have taken on greater importance with the significant transformation of the U.S. health care system and value-based purchasing. The shift toward creating financial incentives and disincentives to achieve quality goals makes focusing on those who receive lower-quality care more important than ever before. This chapter will provide an overview of racial and ethnic disparities in health and health care, identify root causes, and provide key recommendations to address these disparities at both the clinical and health system levels.

NATURE AND EXTENT OF DISPARITIES

Minority Americans have poorer health outcomes than whites from preventable and treatable conditions such as cardiovascular disease, diabetes, asthma, cancer, and HIV/AIDS **(Fig. 7-1)**. Multiple factors contribute to these racial and ethnic disparities in health. First and foremost, social determinants —such as lower socioeconomic status (SES; e.g., lower income, less wealth, and lower educational attainment), inadequate and unsafe housing, and racism—are strongly linked to poor health outcomes. These factors disproportionately impact minority populations. In fact, SES has consistently been found to be one of the strongest predictors of health outcomes. While the mechanisms are complex (i.e., does poverty cause poor health, or does poor health cause poverty?), it is clear that low-SES populations experience disparities in health and that low SES is a major factor in racial/ethnic disparities.

FIGURE 7-1

Age-adjusted death rates for selected causes by race and ethnic origin, 2005. *(From the U.S. Census Bureau, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

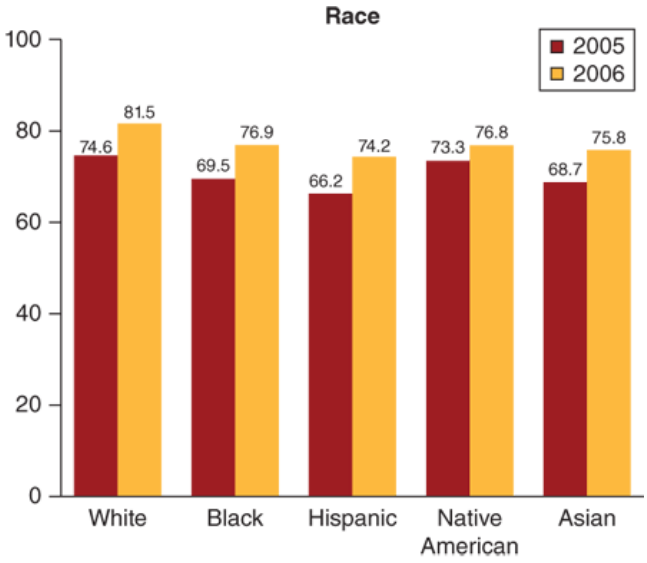
Racial/ethnic disparities are documented globally, although their assessment has centered more on the comparison of individuals by SES in other countries than in the United States. Similar to the U.S. pattern, low-SES residents of other nations tend to have poorer health outcomes. It is noteworthy that results are mixed when the health status of nations is compared by SES. High-SES nations such as the United States do not necessarily have health outcomes that correlate with their high expenditures for health care. For example, as of 2016, the United States ranks 27th in the world—just behind Serbia—on basic public health measures such as infant mortality. This ranking may be due in part to the correlation between wealth distribution and SES rather than just absolute SES. This area of active research is outside the scope of this chapter.

Racism has more recently been shown to predict poorer health outcomes. The physiologic impact of the stress imposed by racism (and poverty), including increased cortisol levels, can lead to chronic adverse effects on health. Lack of access to care also takes a significant toll. Uninsured individuals are less likely to have a regular source of care and are more likely to delay seeking care and to go without needed care; this limited access results in avoidable hospitalizations, emergency hospital care, and adverse health outcomes.

In addition to racial and ethnic disparities in *health*, there are racial and ethnic disparities in the *quality of care* for persons with access to the health care system. For instance, disparities have been found in the treatment of pneumonia **(Fig. 7-2)** and congestive heart failure, with blacks receiving less optimal care than whites when hospitalized for these conditions. Moreover, blacks with end-stage renal disease are referred less often to the transplant list than are their white counterparts **(Fig. 7-3)**. Disparities have been found, for example, in the use of cardiac diagnostic and therapeutic procedures (with blacks being referred less often than whites for cardiac catheterization and bypass grafting), prescription of analgesia for pain control (with blacks and Latinos receiving less pain medication than whites for long-bone fractures and cancer), and surgical treatment of lung cancer (with blacks receiving less curative surgery than whites for non-small-cell lung cancer). Again, many of these disparities have occurred even when variations in factors such as insurance status, income, age, comorbid conditions, and symptom expression are taken into account. However, one additional factor—disparities in the quality of care provided at the sites where minorities tend to receive care—has been shown to be an important contributor to overall disparities.

FIGURE 7-2

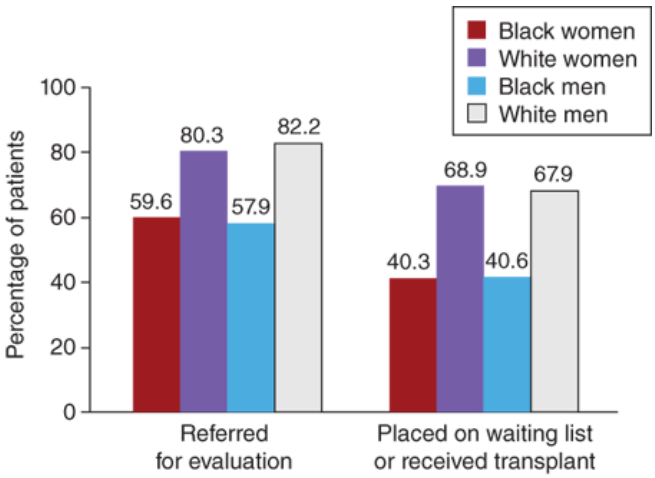
Recommended hospital care received by Medicare patients with pneumonia, by race/ethnicity, 2006. The reference population consisted of Medicare beneficiaries with pneumonia who were hospitalized. The composite was calculated by averaging the percentage of the population that received each of the five incorporated components of care. *(Adapted from Agency for Healthcare Research and Quality: The 2008 National Health Care Disparities Report.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 7-3

Referral for evaluation at a transplantation center or placement on a waiting list/receipt of a renal transplant within 18 months after the start of dialysis among patients who wanted a transplant, according to race and sex. The reference population consisted of 239 black women, 280 white women, 271 black men, and 271 white men. Racial differences were statistically significant among both the women and the men ($p<.0001$ for each comparison). *(From JZ Ayanian et al: N Engl J Med 341:1661, 1999.)*

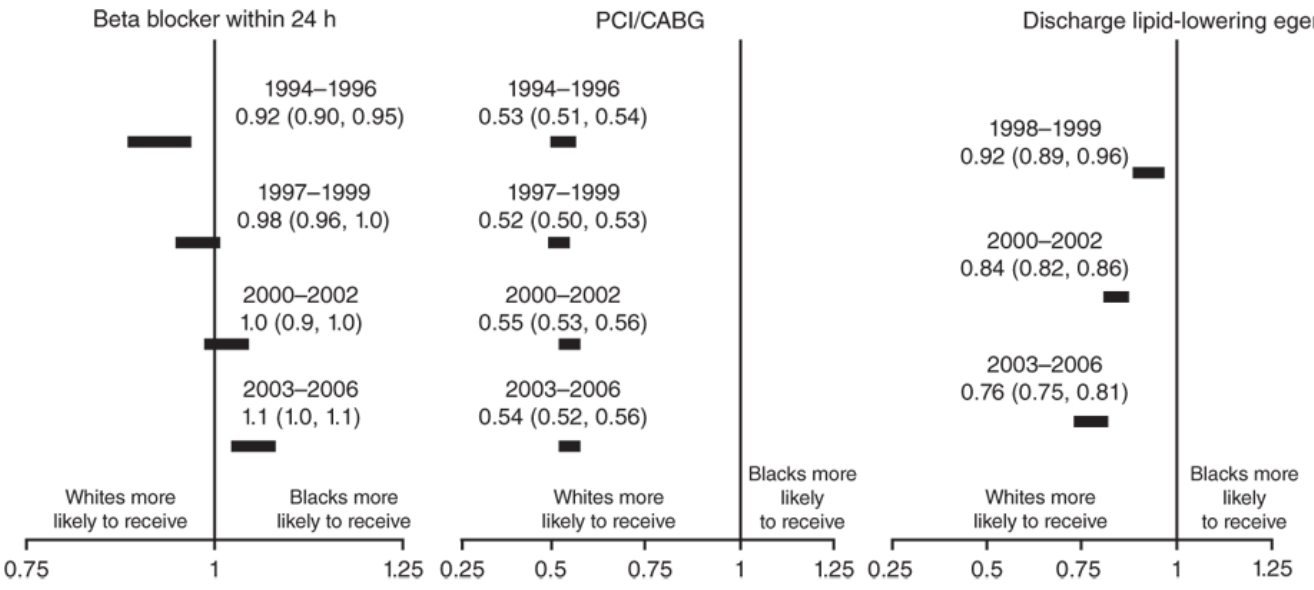


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Little progress has been made in addressing racial/ethnic disparities in cardiovascular procedures and other advanced surgical procedures, whereas some progress has been made in eliminating disparities in primary-care process measures. Data from the National Registry of Myocardial Infarction found evidence of continued disparities in guideline-based admission, procedural, and discharge therapy use from 1994 to 2006. Black patients were less likely than white patients to receive percutaneous coronary intervention/coronary artery bypass grafting (PCI/CABG), a disparity that has improved little since 1994. Further, compared with whites, black patients were less likely to receive lipid-lowering medications at discharge, with a gap that has widened since 1998 (Fig. 7-4). A 2009 study showed that blacks had worse post–myocardial infarction outcomes than whites, but that the difference could be explained by site of care and patient factors (such as socioeconomic status and comorbid conditions).

FIGURE 7-4

Racial differences in guideline-based treatments for acute myocardial infarction (AMI). The reference population consisted of 2,515,106 patients with AMI admitted to U.S. hospitals between July 1990 and December 2006. CABG, coronary artery bypass grafting; PCI, percutaneous coronary intervention. *(From ED Peterson et al: Am Heart J 156:1045, 2008.)*

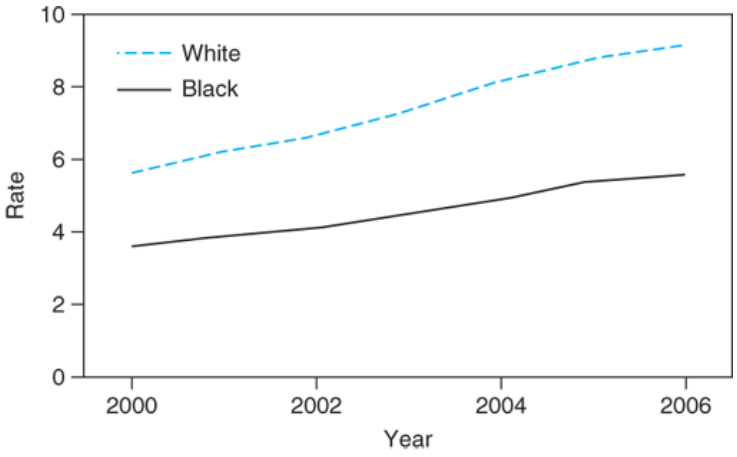


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The Centers for Disease Control and Prevention (CDC) analyzed national and state rates of total knee replacement (TKR) for Medicare enrollees for the period 2000–2006, with patients stratified by sex, age, and black or white race. TKR rates overall in the United States increased 58%, with similar increases among whites (61%) and blacks (56%). However, the TKR rate for blacks was 37% lower than the rate for whites in 2000 and 39% lower in 2006; i.e., the disparity not only did not improve but even worsened slightly (Fig. 7-5). More recent data (up to 2010) show no apparent change in these figures. Data for enrollees in Medicare managed-care plans provide evidence for a narrowing in racial disparities between 2006 and 2011 in several "report card" preventive care measures, such as mammography and glucose and cholesterol testing. However, racial disparities in more complex measures, such as glucose control in diabetic patients and cholesterol levels in patients after a heart attack, did not improve during that interval.

FIGURE 7-5

Racial trends in age-adjusted total knee replacement in Medicaid enrollees from 2000 to 2006. The reference population consisted of Medicaid part A enrollees ≥65 years of age who were not members of a managed-care plan. *(From the Centers for Disease Control and Prevention, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The 2015 National Healthcare Quality and Disparities Report, released by the Agency for Healthcare Research and Quality, found few improvements in disparities for a wide range of quality measures between 2001 and 2013. In fact, for minority groups and poor people, disparities in the vast majority of core quality measures either stayed the same or got worse, including measures of effectiveness, patient safety, and timeliness of care. For example, for black vs white individuals, disparities in quality worsened for 102 measures and persisted for 116 measures out of 248 total measures, and disparities were eliminated in no measured area. This annual report is particularly important, given that most studies of disparities have not been repeated with the same methodology used to document possible trends. Some studies have tracked disparities using specific disease and treatment registries. For example, by 2008, the use of acute and discharge medications for myocardial infarction had largely been equalized among racial and ethnic groups; however, African American and Hispanic patients still experienced longer delays before reperfusion, with door-to-balloon times of <90 min for 83% of white patients as opposed to 75 and 76% of black and Hispanic patients, respectively. A recent review of disparities for kidney transplant recipients over two decades showed similar mixed findings with trends for some measures improving while others worsened.

ROOT CAUSES OF DISPARITIES

The National Academy of Medicine (formerly, the Institute of Medicine, IOM) report *Unequal Treatment*, released in March 2002, remains the preeminent study of racial and ethnic disparities in health care in the United States. The IOM was charged with assessing the extent of racial/ethnic differences in health care that are not otherwise attributable to known factors, such as access to care. To provide recommendations regarding interventions aimed at eliminating health care disparities, the IOM studied health system, provider, and patient factors. The study found the following:

Racial and ethnic disparities in health care exist and, because they are associated with worse health outcomes, are unacceptable.

Racial and ethnic disparities in health care occur in the context of (1) broader historic and contemporary social and economic inequality and (2) evidence of persistent racial and ethnic discrimination in many sectors of American life.

Many sources—including health systems, health care providers, patients, and utilization managers—may contribute to racial and ethnic disparities in health care.

Bias, stereotyping, prejudice, and clinical uncertainty on the part of health care providers may contribute to racial and ethnic disparities in health care.

A small number of studies suggest that minority patients may be more likely to refuse treatments, yet these refusal rates are generally small and do not fully explain health care disparities.

Unequal Treatment went on to identify a set of root causes that included the following:

Health system factors: These include issues related to the complexity of the health care system, the difficulty that minority patients may have in navigating this complex system, and the lack of availability of interpreter services to assist patients with limited English proficiency. In addition, health care systems are generally ill prepared to identify and address disparities.

Provider-level factors: These include issues related to the health care provider, including stereotyping, the impact of race/ethnicity on clinical decision-making, and clinical uncertainty due to poor communication.

Patient-level factors: These include patients' mistrust of the health care system leading to refusal of services, poor adherence to treatment, and delay in seeking care.

A more detailed analysis of these root causes is presented below.

Health System Factors

HEALTH SYSTEM COMPLEXITY

Even among persons who are insured and educated and who have a high degree of health literacy, navigating the U.S. health care system can be complicated and confusing. Some individuals may be at higher risk for receiving substandard care because of their difficulty navigating the system's complexities. These individuals may include those from cultures unfamiliar with the Western model of health care delivery, those with limited English proficiency, those with low health literacy, and those who are mistrustful of the health care system. These individuals may have difficulty knowing how and where to go for a referral to a specialist; how to prepare for a procedure such as a colonoscopy; or how to follow up on an abnormal test result such as a mammogram. Since people of color in the United States tend to be overrepresented among the groups listed above, the inherent complexity of navigating the health care system has been seen as a root cause for racial/ethnic disparities in health care.

OTHER HEALTH SYSTEM FACTORS

Racial/ethnic disparities are due not only to differences in care provided within hospitals but also to where and from whom minorities receive their care; i.e., certain specific providers, geographic regions, or hospitals are lower-performing on certain aspects of quality. For example, one study showed that 25% of hospitals cared for 90% of black Medicare patients in the United States and that these hospitals tended to have lower performance scores on certain quality measures than other hospitals. That said, health systems generally are not well prepared to measure, report, and intervene to reduce disparities in care. Few hospitals or health plans stratify their quality data by race/ethnicity or language to measure disparities, and even fewer use data of this type to develop disparity-targeted interventions. Similarly, despite regulations concerning the need for professional interpreters, research demonstrates that many health care organizations and providers fail to routinely provide this service for patients with limited English proficiency. Despite the link between limited English proficiency and health-care quality and safety, few providers or institutions monitor performance for patients in these areas.

Provider-Level Factors

PROVIDER–PATIENT COMMUNICATION

Significant evidence highlights the impact of sociocultural factors, race, ethnicity, and limited English proficiency on health and clinical care. Health care professionals frequently care for diverse populations with varied perspectives, values, beliefs, and behaviors regarding health and well-being. The differences include variations in the recognition of symptoms, thresholds for seeking care, comprehension of management strategies, expectations of care (including preferences for or against diagnostic and therapeutic procedures), and adherence to preventive measures and medications. In addition, sociocultural differences between patient and provider influence communication and clinical decision-making and are especially pertinent: evidence clearly links provider–patient communication to improved patient satisfaction, regimen adherence, and better health outcomes **(Fig. 7-6)**. Thus, when sociocultural differences between patient and provider are not appreciated, explored, understood, or communicated effectively during the medical encounter, patient dissatisfaction, poor adherence, poorer health outcomes, and racial/ethnic disparities in care may result.

FIGURE 7-6

The link between effective communication and patient satisfaction, adherence, and health outcomes. *(From the Institute of Medicine: Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC, National Academy Press, 2002.)*

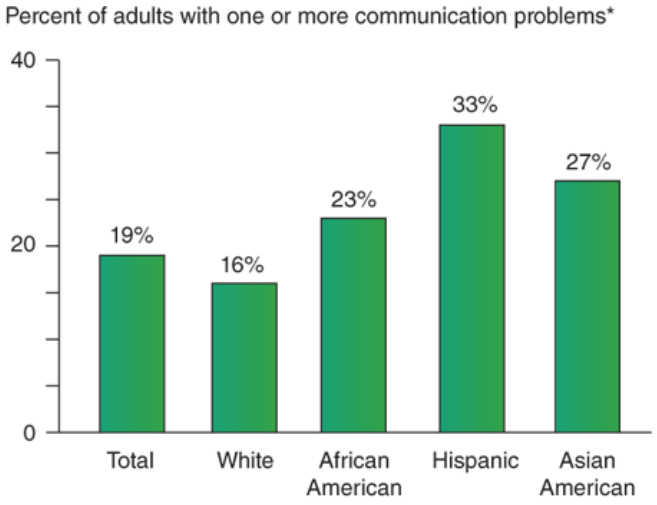


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A survey of 6722 Americans ≥18 years of age is particularly relevant to this important link between provider–patient communication and health outcomes. Whites, African Americans, Hispanics/Latinos, and Asian Americans who had made a medical visit in the past 2 years were asked whether they had trouble understanding their doctors; whether they felt the doctors did not listen; and whether they had medical questions they were afraid to ask. The survey found that 19% of all patients experienced one or more of these problems, yet whites experienced them 16% of the time as opposed to 23% of the time for African Americans, 33% for Hispanics/Latinos, and 27% for Asian Americans **(Fig. 7-7)**.

FIGURE 7-7

Communication difficulties with physicians, by race/ethnicity. The reference population consisted of 6722 Americans ≥18 years of age who had made a medical visit in the previous 2 years and were asked whether they had had trouble understanding their doctors, whether they felt that the doctors had not listened, and whether they had had medical questions they were afraid to ask. *(From the Commonwealth Fund Health Care Quality Survey, 2001.)*



Base: Adults with health care visit in past two years
*Problems include understanding doctor, feeling doctor listened, had questions but did not ask.

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In addition, in the setting of even a minimal language barrier, provider–patient communication without an interpreter is recognized as a major challenge to effective health care delivery. These communication barriers for patients with limited English proficiency lead to frequent misunderstanding of diagnosis, treatment, and follow-up plans; inappropriate use of medications; lack of informed consent for surgical procedures; high rates of adverse events with more serious clinical consequences; and a lower-quality health care experience than is provided to patients who speak fluent English. Physicians who have access to trained interpreters report a significantly higher quality of patient–physician communication than physicians who use other methods. Communication issues related to discordant language disproportionately affect minorities and likely contribute to racial/ethnic disparities in health care.

CLINICAL DECISION-MAKING

Theory and research suggest that variations in clinical decision-making may contribute to racial and ethnic disparities in health care. Two factors are central to this process: clinical uncertainty and stereotyping.

First, a doctor's decision-making process is nested in *clinical uncertainty*. Doctors depend on inferences about severity based on what they understand about illness and the information obtained from the patient. A doctor caring for a patient whose symptoms he or she has difficulty understanding and whose "signals"—the set of clues and indications that physicians rely on to make clinical decisions—are hard to read may make a decision different from the one that would be made for another patient who presents with exactly the same clinical condition. Given that the expression of symptoms may differ among cultural and racial groups, doctors—the overwhelming majority of whom are white—may understand symptoms best when expressed by patients of their own racial/ethnic groups. The consequence is that white patients may be treated differently from minority patients. Differences in clinical decisions can arise from this mechanism even when the doctor has the same regard for each patient (i.e., is not prejudiced).

Second, the literature on social cognitive theory highlights how natural tendencies to stereotype may influence clinical decision-making. *Stereotyping* can be defined as the way in which people use social categories (e.g., race, gender, age) in acquiring, processing, and recalling information about others. Faced with enormous information loads and the need to make many decisions, people often subconsciously simplify the decision-making process and lessen cognitive effort by using "categories" or "stereotypes" that bundle information into groups or types that can be processed more quickly. Although functional, stereotyping can be systematically biased, as people are automatically classified into social categories based on dimensions such as *race*, *gender*, and *age*. Many people may not be aware of their attitudes, may not consciously endorse specific stereotypes, and paradoxically may consider themselves egalitarian and not prejudiced.

Stereotypes may be strongly influenced by the messages presented consciously and unconsciously in society. For instance, if the media and our social/professional contacts tend to present images of minorities as being less educated, more violent, and nonadherent to health care recommendations, these impressions may generate stereotypes that unnaturally and unjustly impact clinical decision-making. As signs of racism, classism, gender bias, and ageism are experienced (consciously or unconsciously) in our society, stereotypes may be created that impact the way doctors

manage patients from these groups. On the basis of training or practice location, doctors may develop certain perceptions about race/ethnicity, culture, and class that may evolve into stereotypes. For example, many medical students and residents are trained—and minorities cared for—in academic health centers or public hospitals located in socioeconomically disadvantaged areas. As a result, doctors may begin to equate certain races and ethnicities with specific health beliefs and behaviors (e.g., "these patients" engage in risky behaviors, "those patients" tend to be noncompliant) that are more associated with the social environment (e.g., poverty) than with a patient's racial/ethnic background or cultural traditions. This "conditioning" phenomenon may also be operative if doctors are faced with certain racial/ethnic patient groups who frequently do not choose aggressive forms of diagnostic or therapeutic intervention. The result over time may be that doctors begin to believe that "these patients" do not like invasive procedures; thus they may not offer these procedures as options. A wide range of studies have documented the potential for provider biases to contribute to racial/ethnic disparities in health care. For example, one study measured physicians' unconscious (or implicit) biases and showed that these were related to differences in decisions to provide thrombolysis for a hypothetical black or white patient with a myocardial infarction.

It is important to differentiate stereotyping from prejudice and discrimination. *Prejudice* is a conscious prejudgment of individuals that may lead to disparate treatment, and *discrimination* is conscious and intentional disparate treatment. All individuals *stereotype* subconsciously, yet, if left unquestioned, these subconscious assumptions may lead to lower-quality care for certain groups because of differences in clinical decision-making or differences in communication and patient-centeredness. For example, one study tested physicians' unconscious racial/ethnic biases and showed that patients perceived more biased physicians as being less patient-centered in their communication. What is particularly salient is that stereotypes tend to be activated most in environments where the individual is stressed, multitasking, and under time pressure—the hallmarks of the clinical encounter. In fact, in a survey of close to 16,000 physicians, 42% admitted that bias—including by race and ethnicity—impacted their clinical decision-making. Interestingly, emergency medicine physicians, who worked in environments of stress, time pressure, risk, and where they are multitasking, topped the list by discipline at 62%.

Patient-Level Factors

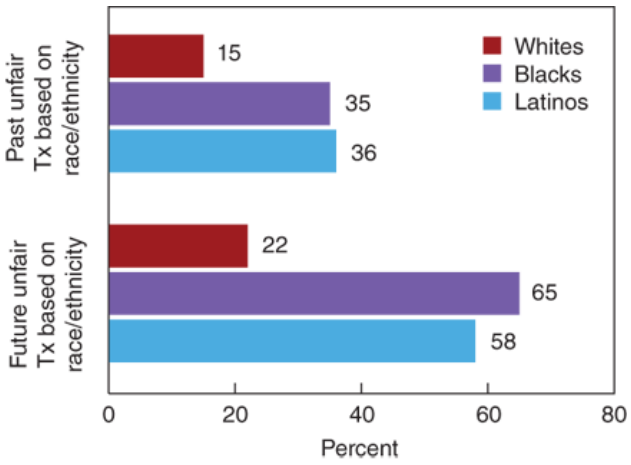
Lack of trust has become a major concern for many health care institutions today. For example, an IOM report, *To Err Is Human: Building a Safer Health System*, documented alarming rates of medical errors that made patients feel vulnerable and less trustful of the U.S. health care system. The increased media and academic attention to problems related to quality of care (and of disparities themselves) has clearly diminished trust in doctors and nurses.

Trust is a crucial element in the therapeutic alliance between patient and health care provider. It facilitates open communication and is directly correlated with adherence to the physician's recommendations and the patient's satisfaction. In other words, patients who mistrust their health care providers are less satisfied with the care they receive, and mistrust of the health care system greatly affects patients' use of services. Mistrust can also result in inconsistent care, "doctor-shopping," self-medication, and an increased demand by patients for referrals and diagnostic tests.

On the basis of historic factors such as discrimination, segregation, and medical experimentation, blacks may be especially mistrustful of providers. The exploitation of blacks by the U.S. Public Health Service during the Tuskegee syphilis study from 1932 to 1972 left a legacy of mistrust that persists even today among this population. Other populations, including Native Americans/Alaskan Natives, Hispanics/Latinos, and Asian Americans, also harbor significant mistrust of the health care system. A national survey conducted by the Kaiser Family Foundation found that there is significant mistrust for the health care system among minority populations. Of the 3884 individuals surveyed, 36% of Hispanics and 35% of blacks (compared to 15% of whites) felt they were treated unfairly in the health care system in the past based on their race and ethnicity. Perhaps even more alarming—65% of blacks and 58% of Hispanics (compared to 22% of whites) were afraid of being treated unfairly in the future based on their race/ethnicity **(Fig. 7-8)**.

FIGURE 7-8

Patient perspectives regarding unfair treatment (Tx) based on race/ethnicity. The reference population consisted of 3884 individuals surveyed about how fairly they had been treated in the health care system in the past and how fairly they felt they would be treated in the future on the basis of their race/ethnicity. *(From Race, Ethnicity & Medical Care: A Survey of Public Perceptions and Experiences. Kaiser Family Foundation, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

This mistrust may contribute to wariness in accepting or following recommendations, undergoing invasive procedures, or participating in clinical research, and these choices, in turn, may lead to misunderstanding and the perpetuation of stereotypes among health professionals.

KEY RECOMMENDATIONS TO ADDRESS RACIAL/ETHNIC DISPARITIES IN HEALTH CARE

The publication *Unequal Treatment* provides a series of recommendations to address racial and ethnic disparities in health care, focusing on a broad set of stakeholders. These recommendations include *health system interventions*, *provider interventions*, *patient interventions*, and *general recommendations*, which are described in more detail below.

Health System Interventions

COLLECTION AND REPORTING OF DATA ON HEALTH CARE ACCESS AND USE, BY PATIENTS' RACE/ETHNICITY

Unequal Treatment found that the appropriate systems to track and monitor racial and ethnic disparities in health care are lacking and that less is known about the disparities affecting minority groups other than African Americans (Hispanics, Asian Americans, Pacific Islanders, Native Americans, and Alaskan Natives). For instance, only in the mid-1980s did the Medicare database begin to collect data on patient groups outside the standard categories of "white," "black," and "other." Federal, private, and state-supported data-collection efforts are scattered and unsystematic, and many health care systems and hospitals still do not collect data on the race, ethnicity, or primary language of enrollees or patients. A survey by the Institute for Diversity in Health Management and the Health Research and Educational Trust in 2015 found that 98% of 1083 U.S. hospitals collected information on race, 95% collected data on ethnicity, and 94% collected data on primary language. However, only 45% collected data on race, 40% collected data on ethnicity, and 38% collected data on primary language to benchmark gaps in care. A survey by America's Health Insurance Plans Foundation in 2008 and 2010 showed that the proportion of enrollees in plans that collected race/ethnicity data of some type increased from 75 to 79%; however, the total percentage of plan enrollees whose race/ethnicity and language are recorded is still much lower than these figures.

ENCOURAGEMENT OF THE USE OF EVIDENCE-BASED GUIDELINES AND QUALITY IMPROVEMENT

Unequal Treatment highlights the subjectivity of clinical decision-making as a potential cause of racial and ethnic disparities in health care by describing how clinicians—despite the existence of well-delineated practice guidelines—may offer (consciously or unconsciously) different diagnostic and therapeutic options to different patients on the basis of their race or ethnicity. Therefore, the widespread adoption and implementation of evidence-based guidelines is a key recommendation in eliminating disparities. For instance, evidence-based guidelines are now available for the management of diabetes, HIV/AIDS, cardiovascular diseases, cancer screening and management, and asthma—all areas where significant disparities exist. As part of ongoing quality-improvement efforts, particular attention should be paid to the implementation of evidence-based guidelines for all patients, regardless of their race and ethnicity.

SUPPORT FOR THE USE OF LANGUAGE INTERPRETATION SERVICES IN THE CLINICAL SETTING

As described previously, a lack of efficient and effective interpreter services in a health care system can lead to patient dissatisfaction, to poor comprehension and adherence, and thus to ineffective/lower-quality care for patients with limited English proficiency. *Unequal Treatment*'s recommendation to support the use of interpretation services has clear implications for delivery of quality health care by improving doctors' ability to communicate effectively with these patients.

INCREASES IN THE PROPORTION OF UNDERREPRESENTED MINORITIES IN THE HEALTH CARE WORKFORCE

Data for 2014 from the Association of American Medical Colleges indicate that, of the 72.4% of U.S. physicians whose race and ethnicity are known, Hispanics make up 4.1%, blacks 4.1%, and Native American and Alaskan Natives 0.4%. Furthermore, U.S. national data show that minorities (excluding Asians) compose just 7.1 % of full-time medical school faculty. In addition, minority faculty in 2007 were more likely to be at or below the rank of assistant professor, while whites composed the highest proportion of full professors. Similarly, a 2012 study found that both Hispanic and Black faculty were promoted at lower rates than their white counterparts. Despite representing ~26% of the U.S. population (a number projected to almost double by 2050), minority students are still underrepresented in medical schools. In 2016, matriculates to U.S. medical schools were 6.1% Latino, 6.6% African American, 0.1% Native Hawaiian or Other Pacific Islander, and 0.3% Native American or Alaskan Native. These percentages have decreased or remained the same since 2007. It will be difficult to develop a diverse health-care workforce that can meet the needs of an increasingly diverse population without dramatic changes in the racial and ethnic composition of medical student bodies.

Provider Interventions

INTEGRATION OF CROSS-CULTURAL EDUCATION INTO THE TRAINING OF ALL HEALTH CARE PROFESSIONALS

The goal of cross-cultural education is to improve providers' ability to understand, communicate with, and care for patients from diverse backgrounds. Such education focuses on enhancing awareness of sociocultural influences on health beliefs and behaviors and on building skills to facilitate understanding and management of these factors in the medical encounter. Cross-cultural education includes curricula on health care disparities, use of interpreters, and effective communication and negotiation across cultures. These curricula can be incorporated into health-professions training in medical schools, residency programs, nursing schools, and other health professions programs, and can be offered as a component of continuing education. Despite the importance of this area of education and the attention it has attracted from medical education accreditation bodies, a national survey of senior resident physicians by Weissman and colleagues found that up to 28% felt unprepared to deal with cross-cultural issues, including caring for patients who have religious beliefs that may affect treatment, patients who use complementary medicine, patients who have health beliefs at odds with Western medicine, patients who mistrust the health care system, and new immigrants. In a study at one medical school, 70% of fourth-year students felt inadequately prepared to care for patients with limited English proficiency. Efforts to incorporate cross-cultural education into medical education will contribute to improving communication and to providing a better quality of care for all patients.

INCORPORATION OF TEACHING ON THE IMPACT OF RACE, ETHNICITY, AND CULTURE ON CLINICAL DECISION-MAKING

Unequal Treatment and more recent studies found that stereotyping by health care providers can lead to disparate treatment based on a patient's race or ethnicity. The Liaison Committee on Medical Education, which accredits medical schools, issued a directive that medical education should include instruction on how a patient's race, ethnicity, and culture might unconsciously impact communication and clinical decision-making.

Patient Interventions

Difficulty navigating the health care system and obtaining access to care can be a hindrance to all populations, particularly to minorities. Similarly, lack of empowerment or involvement in the medical encounter by minorities can be a barrier to care. Patients need to be educated on how to navigate the health care system and how best to access care. Interventions should be used to increase patients' participation in treatment decisions.

General Recommendations

INCREASE AWARENESS OF RACIAL/ETHNIC DISPARITIES IN HEALTH CARE

Efforts to raise awareness of racial/ethnic health care disparities have done little for the general public but have been fairly successful among physicians, according to a Kaiser Family Foundation report. In 2006, nearly 6 in 10 people surveyed believed that blacks received the same quality of care as whites, and 5 in 10 believed that Latinos received the same quality of care as whites. These estimates are similar to findings in a 1999 survey. Despite this lack of awareness, most people believed that all Americans deserve quality care, regardless of their background. In contrast, the level of awareness among physicians has risen sharply. In 2002, the majority (69%) of physicians said that the health care system "rarely or never" treated people unfairly on the basis of their racial/ethnic background. In 2005, less than one-quarter (24%) of physicians disagreed with the statement that "minority patients generally receive lower-quality care than white patients." More recently, a survey by WedMD showed that 42% of 16,000 physicians admitted that their own personal biases impact their clinical decision-making, including on characteristics such as race and ethnicity. Increasing awareness of racial and ethnic health disparities, and their root causes, among health care professionals and the public is an important first step in addressing these disparities. The ultimate goals are to generate discourse and to mobilize action to address disparities at multiple levels, including health policy makers, health systems, and the community.

CONDUCT FURTHER RESEARCH TO IDENTIFY SOURCES OF DISPARITIES AND PROMISING INTERVENTIONS

While the literature that formed the basis for the findings reported and recommendations made in *Unequal Treatment* provided significant evidence for racial and ethnic disparities, additional research is needed in several areas. First, most of the literature on disparities focuses on black-versus-white differences; much less is known about the experiences of other minority groups. Improving the ability to collect racial and ethnic patient data should

facilitate this process. However, in instances where the necessary systems are not yet in place, racial and ethnic patient data may be collected prospectively in the setting of clinical or health services research to more fully elucidate disparities for other populations. Second, much of the literature on disparities to date has focused on defining areas in which these disparities exist, but less has been done to identify the multiple factors that contribute to the disparities or to test interventions to address these factors. There is clearly a need for research that identifies promising practices and solutions to disparities.

IMPLICATIONS FOR CLINICAL PRACTICE

Individual health care providers can do several things in the clinical encounter to address racial and ethnic disparities in health care.

Be Aware that Disparities Exist

Increasing awareness of racial and ethnic disparities among health care professionals is an important first step in addressing disparities in health care. Only with greater awareness can care providers be attuned to their behavior in clinical practice and thus monitor that behavior and ensure that all patients receive the highest quality of care, regardless of race, ethnicity, or culture.

Practice Culturally Competent Care

Previous efforts have been made to teach clinicians about the attitudes, values, beliefs, and behaviors of certain cultural groups—the key practice "dos and don'ts" in caring for "the Hispanic patient" or the "Asian patient," for example. In certain situations, learning about a particular local community or cultural group, with a goal of following the principles of community-oriented primary care, can be helpful; when broadly and uncritically applied, however, this approach can actually lead to stereotyping and oversimplification of culture, without respect for its complexity.

Cultural competence has thus evolved from merely learning information and making assumptions about patients on the basis of their backgrounds to focusing on the development of skills that follow the principles of patient-centered care. *Patient-centeredness* encompasses the qualities of compassion, empathy, and responsiveness to the needs, values, and expressed preferences of the individual patient. *Cultural competence* aims to take things a step further by expanding the repertoire of knowledge and skills classically defined as "patient-centered" to include those that are especially useful in cross-cultural interactions (and that, in fact, are vital in all clinical encounters). This repertoire includes effectively using interpreter services, eliciting the patient's understanding of his or her condition, assessing decision-making preferences and the role of family, determining the patient's views about biomedicine versus complementary and alternative medicine, recognizing sexual and gender issues, and building trust. For example, while it is important to understand all patients' beliefs about health, it may be particularly crucial to understand the health beliefs of patients who come from a different culture or have a different health care experience. With the individual patient as teacher, the physician can adjust his or her practice style to meet the patient's specific needs.

Avoid Stereotyping

Several strategies can allow health care providers to counteract, both systemically and individually, the normal tendency to stereotype. For example, when racially/ethnically/culturally/socially diverse teams in which each member is given equal power are assembled and are tasked to achieve a common goal, a sense of camaraderie develops and prevents the development of stereotypes based on race/ethnicity, gender, culture, or class. Thus, health care providers should aim to gain experiences working with and learning from a diverse set of colleagues. In addition, simply being aware of the operation of social cognitive factors allows providers to actively check up on or monitor their behavior. Physicians can constantly reevaluate to ensure that they are offering the same things, in the same ways, to all patients. Understanding one's own susceptibility to stereotyping—and how disparities may result—is essential in providing equitable, high-quality care to all patients.

Work to Build Trust

Patients' mistrust of the health care system and of health care providers impacts multiple facets of the medical encounter, with effects ranging from decreased patient satisfaction to delayed care. Although the historic legacy of discrimination can never be erased, several steps can be taken to build trust with patients and to address disparities. First, providers must be aware that mistrust exists and is more prevalent among minority populations, given the history of discrimination in the United States and other countries. Second, providers must reassure patients that they come first, that everything possible will be done to ensure that they always get the best care available, and that their caregivers will serve as their advocates. Third, interpersonal skills and communication techniques that demonstrate honesty, openness, compassion, and respect on the part of the health care provider are essential tools in dismantling mistrust. Finally, patients indicate that trust is built when there is shared, participatory decision-making and the provider makes a concerted effort to understand the patient's background. When the doctor–patient relationship is reframed as one of solidarity, the patient's sense of vulnerability can be transformed into one of trust. The successful elimination of disparities requires trust-building interventions and strengthening of this relationship.

CONCLUSION

The issue of racial and ethnic disparities in health care has gained national prominence, both with the release of the IOM report *Unequal Treatment* and with more recent articles that have confirmed their persistence and explored their root causes. Furthermore, another influential IOM report, *Crossing the Quality Chasm*, has highlighted the importance of equity—i.e., no variations in quality of care due to personal characteristics, including race and ethnicity—as a central principle of quality. Current efforts in health care reform and transformation, including a greater focus on value (high-quality care and cost-control), will sharpen the nation's focus on the care of populations who experience low-quality, costly care. Addressing disparities will become a major focus, and there will be many obvious opportunities for interventions to eliminate them. Greater attention to addressing the root causes of disparities will improve the care provided to all patients, not just those who belong to racial and ethnic minorities.

ACKNOWLEDGMENTS

The authors thank Marina Cervantes and Andrea Madu for their contributions to this chapter.

FURTHER READING

Ayanian JZ et al: Racial and ethnic disparities among enrollees in Medicare Advantage plans. N Engl J Med 371:2288, 2014. [PubMed: 25494268]

Hausmann LR et al: Racial and ethnic disparities in pneumonia treatment and mortality. Med Care 47:1009, 2009. [PubMed: 19648832]

Medscape: Medscape Lifestyle Report 2016: Bias and Burnout. *http://www.medscape.com/features/slideshow/lifestyle/2016/public/overview*. Accessed February 21, 2017.

Nunez-Smith M et al: Institutional variation in the promotion of racial/ethnic minority faculty at US medical schools. Am J Public Health 102:852, 2012. [PubMed: 22420820]

Patzer RE et al: Neighborhood poverty and racial disparities in kidney transplant waitlisting. J Am Soc Nephrol 20:1333, 2009. [PubMed: 19339381]

Rhee CM et al: Impact of age, race and ethnicity on dialysis patient survival and kidney transplantation disparities. Am J Nephrol 39:183, 2014. [PubMed: 24556752]

Taber DJ et al: Twenty years of evolving trends in racial disparities for adult kidney transplant recipients. Kidney Int 90:878, 2016. [PubMed: 27555121]

Yancy CW et al: Quality of care of and outcomes for African Americans hospitalized with heart failure: findings from the OPTIMIZE-HF (Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients With Heart Failure) registry. J Am Coll Cardiol 51:1675, 2008. [PubMed: 18436120]

Chapter 8: Ethical Issues in Clinical Medicine

Bernard Lo; Christine Grady

INTRODUCTION

Twenty-first-century physicians face novel ethical dilemmas that can be perplexing and emotionally draining. For example, electronic medical records, handheld personal devices, and provision of care by interdisciplinary teams all hold the promise of more coordinated and comprehensive care, but also raise new concerns about confidentiality, appropriate boundaries of the doctor–patient relationship, and responsibility. Chapter 1 puts the practice of medicine into a professional and historical context. The current chapter presents approaches and principles that physicians can use to address the ethical issues they encounter in their work. Physicians make ethical judgments about clinical situations every day. Traditional professional codes and ethical principles provide instructive guidance for physicians but need to be interpreted and applied to each situation. Physicians need to be prepared for lifelong learning about ethical issues and dilemmas as well as about new scientific and clinical developments. When struggling with difficult ethical issues, physicians may need to reevaluate their basic convictions, tolerate uncertainty, and maintain their integrity while respecting the opinions of others. Discussing perplexing ethical issues with other members of the health care team, ethics consultation services, or the hospital ethics committee can clarify issues and reveal strategies for resolution, including improving communication and dealing with strong or conflicting emotions.

APPROACHES TO ETHICAL PROBLEMS

Several approaches may be useful for resolving ethical issues. Among these approaches are those based on ethical principles, virtue ethics, professional oaths, and personal values. These various sources of guidance encompass precepts that may conflict in a particular case, leaving the physician in a quandary. In a diverse society, different individuals may turn to different sources of moral guidance. In addition, general moral precepts often need to be interpreted and applied in the context of a particular clinical situation. When facing an ethical challenge, physicians should articulate their concerns and reasoning, discuss and listen to the views of others involved in the case, and call on available resources as needed. Through these efforts, physicians can gain deeper insight into the ethical issues they face and often can reach mutually acceptable resolutions to complex problems.

ETHICAL PRINCIPLES

Ethical principles can serve as general guidelines to help physicians determine the right thing to do.

Respecting Patients

Physicians should always treat patients with respect, which entails understanding patients' goals, communicating effectively, obtaining informed and voluntary consent, respecting informed refusals, and protecting confidentiality. Different clinical goals and approaches are often feasible, and interventions result in both benefit and harm. Individuals differ in how they value health and medical care and how they weigh the benefits and risks of medical interventions. Generally, the values and informed choices of patients should be respected.

GOALS AND TREATMENT DECISIONS

Physicians should discuss the goals of care with patients, as well as relevant and accurate information about diagnosis, current clinical circumstances, likely trajectory and prognosis, and treatment options. Physicians may be tempted to withhold a serious diagnosis, misrepresent it by using ambiguous terms, or limit discussions of prognosis or risks for fear that patients will become anxious or depressed. Providing honest information about clinical situations preserves patients' autonomy and trust and promotes sound communication with patients and colleagues. To help patients cope with bad news, doctors can adjust the pace of disclosure, offer empathy and hope, provide emotional support, and call on other resources such as spiritual care or social work. However, patients may choose not to receive such information or ask surrogates to make decisions on their behalf, as is common with serious diagnoses in some traditional cultures.

OBTAINING INFORMED CONSENT

Physicians should discuss with patients the nature of proposed care, alternatives, and the risks, benefits, and likely consequences of each option. Informed consent involves more than obtaining signatures on consent forms. Physicians should promote shared decision-making by educating patients, answering their questions, checking that they understand key issues, making recommendations, and helping them to deliberate. Patients can be overwhelmed by medical jargon, needlessly complicated explanations, or the provision of too much information at once. Patients can make informed decisions only if they receive honest and understandable information. Competent, informed patients may refuse recommended interventions and choose among reasonable alternatives. If patients cannot give consent in an emergency and if delay of treatment while surrogates are contacted will place their lives or health in peril, treatment can be given without informed consent. People are presumed to want such emergency care unless they have previously indicated otherwise.

Respect for patients does not entitle patients to insist on any care they want. Physicians are not obligated to provide interventions that have no physiologic rationale, that have already failed, or that are contrary to evidence-based practice recommendations or good clinical judgment. Public policies and laws also dictate certain decisions—e.g., allocation of cadaveric organs for transplantation and physician aid-in-dying.

CARING FOR PATIENTS WHO LACK DECISION-MAKING CAPACITY

Many patients are not able to make informed decisions because of unconsciousness, dementia, delirium, or other medical conditions. Although only courts have the legal authority to determine that a patient is legally

incompetent, in practice, physicians determine when patients lack the capacity to make particular health care decisions and arrange for authorized surrogates to make decisions for them, without involving the courts. Patients with decision-making capacity can express a choice and appreciate their medical situation, the nature of proposed care, alternatives, and the risks, benefits, and consequences of each alternative. Patient choices should be consistent with their values and not the result of delusions or hallucinations. Physicians should use available assessment tools, other resources such as psychiatry consultation, and clinical judgment to ascertain whether individuals have the capacity to consent and make decisions for themselves. It should not be automatically assumed that a patient who disagrees with a recommendation or refuses treatment lacks capacity, but such decisions should be probed to be sure the patient has the capacity for an informed decision and that there are no misunderstandings. When impairments are fluctuating or reversible, decisions should be postponed if possible until the patient recovers decision-making capacity.

If a patient lacks decision-making capacity, physicians should seek the appropriate surrogate, and ask what the patient would have wanted done. Patients may designate a health care proxy or a durable power of attorney for health care in advance; such choices should be respected. (See **Chap. 9** for further details about advance care planning.) If a patient without decision-making capacity has not previously designated a health care proxy, physicians usually ask family members to serve as surrogates. Many patients want family members as surrogates, and family members generally have the patient's best interests at heart. Statutes in most U.S. states delineate a prioritized list of relatives who may serve as surrogates if the patient has not designated a proxy. Surrogates' decisions should be guided by the patient's values, goals, and previously expressed preferences. However, it may be appropriate to override previous preferences in favor of the patient's current best interests if an intervention is likely to provide a significant benefit, if previous statements do not fit the situation well, or if the patient indicated that the surrogate should have leeway in decisions.

MAINTAINING CONFIDENTIALITY

Maintaining confidentiality is essential in respecting patients' autonomy and privacy; it encourages them to seek treatment and to discuss problems candidly, and helps to prevent discrimination. However, confidentiality may be overridden to prevent serious harm to third parties or to the patient. Exceptions to confidentiality are justified if the risk is serious and probable, there are no less restrictive measures by which to avert risk, and the adverse effects of overriding confidentiality are minimized and deemed acceptable by society. For example, laws require physicians to report cases of tuberculosis, sexually transmitted infection, elder or child abuse, and domestic violence.

Beneficence or Acting in Patients' Best Interests

The principle of *beneficence* requires physicians to act for the patient's benefit. Patients typically lack medical expertise, and illness may make them vulnerable. They rely on and trust physicians to treat them with compassion, provide sound recommendations and promote their well-being. Physicians encourage such trust and have a fiduciary duty to act in the best interests of the patient, which should prevail over physicians' self-interest or the interests of third parties such as hospitals or insurers. Physicians' fiduciary obligations contrast sharply with business relationships, which are characterized by "buyer beware," and not

by reliance and trust. A related principle, “first do no harm,” obliges physicians to prevent unnecessary harm by recommending interventions that maximize benefit and minimize harm, and forbids physicians from providing known ineffective interventions or acting without due care. Although often cited, this precept alone provides limited guidance because many beneficial interventions pose serious risks.

Physicians increasingly provide care with a multidisciplinary team. Team members contribute different types of expertise to the provision of comprehensive, high-quality care for patients. Physicians should collaborate with and respect the contributions of the various members of the multidisciplinary team. Physicians should also initiate and participate in regular communication and planning to avoid diffusion of responsibility and ensure accountability for quality patient care.

INFLUENCES ON PATIENTS’ BEST INTERESTS

Conflicts can arise when patients’ refusal or request of interventions thwarts their own goals for care, causes serious harm, or conflicts with their best medical interests. For example, simply accepting refusal of mechanical ventilation for reversible respiratory failure by a young adult with asthma, in the name of respecting autonomy, is morally constricted. Physicians should elicit patients’ expectations and concerns, correct their misunderstandings, and try to persuade them to accept beneficial therapies. If disagreements persist after such efforts, patients’ informed choices and views of their own best interests should prevail.

Physicians should appreciate that patients, who face increasing co-payments and out-of-pocket expenses, may not be able to afford tests and interventions that are ordered. Physicians should follow up with patients who don’t fill prescriptions or skip doses, discuss alternative drugs, and when possible, prescribe medications that are affordable to the patient.

Organizational policies may sometimes conflict with patients’ best interests. For example, limitations on work-hours could lead to a shift-worker mentality that undermines physician’s dedication to patient’s well-being and sense of responsibility for decisions. Forced handoffs might actually tend to increase the risk of errors unless other measures are taken. Patients’ best interests may be served by flexibility in work-hour limits in some cases, especially if there is rapport with the patient or family that is not easily transferred to another provider. For example, a resident may want to discuss decisions about life-sustaining interventions or comfort a family member over a patient’s death (Chap. 9). Physicians, residents, and medical students should take responsibility for helping to design and improve work-hour schedules based on empirical evidence.

Patients’ interests are also served by improvements in overall quality of care resulting from the increasing use of evidence-based practice guidelines and performance benchmarking. However, practice guideline recommendations may not serve the interests of each individual patient, especially when another plan of care may provide substantially greater benefits. In such situations, physicians should prioritize their duty to act in the patient’s best interests. Physicians should be familiar with relevant practice guidelines, be able to recognize situations in which exceptions might be reasonable, and be prepared to justify an exception.

Acting Justly

The principle of *justice* provides guidance to physicians about how to ethically treat patients and make decisions about allocating important resources, including their own time. *Justice* in a general sense means fairness: people should receive what they deserve. In addition, it is important to act consistently in cases that are similar in ethically relevant ways, in order to avoid arbitrary, biased, and unfair decisions. Justice forbids discrimination in health care based on race, religion, gender, sexual orientation, or other personal characteristics **(Chap. 7)**.

ALLOCATION OF RESOURCES

Justice also requires that limited health care resources be allocated fairly. Universal access to medically needed health care remains an unrealized moral aspiration in the United States and much of the rest of the world. Patients without health insurance often cannot afford health care and lack access to safety-net services. Even among insured patients, insurers may deny coverage for interventions recommended by the physician. In this situation, physicians should advocate for patients and try to help them obtain needed care. Doctors might consider—or patients might request—the use of lies or deception to obtain such benefits. For example, a physician might sign a disability form for a patient who does not meet disability criteria. Although motivated by a desire to help the patient, such deception breaches a basic ethical guideline and undermines physicians' credibility and trustworthiness.

Allocation of health care resources is unavoidable because resources are limited. Many allocation decisions are made at the level of public policy, with physician input. For example, the United Network for Organ Sharing (*https://unos.org/*) provides criteria for allocating scarce organs. *Ad hoc* resource allocation by the physician at the bedside is problematic because it may be inconsistent, unfair, and ineffective. Physicians do have an important role, however, in avoiding unnecessary interventions. Evidence-based lists of tests and procedures that physicians and patients should question and discuss are available through *Choosing Wisely* (*http://www.choosingwisely.org/*). At the bedside, physicians should act as patient advocates within constraints set by society, reasonable insurance coverage, and evidence-based practice. For example, if a patient's insurer has a higher copayment for nonformulary drugs, it still may be reasonable for physicians to advocate for nonformulary products for good reasons (e.g., when the formulary drugs are less effective or not tolerated).

VIRTUE ETHICS

Virtue ethics focuses on physicians' character and qualities, with the expectation that doctors will cultivate such virtues as compassion, trustworthiness, intellectual honesty, humility, and integrity. Proponents argue that, if such characteristics become ingrained, they help guide physicians in unforeseen situations. Moreover, following ethical precepts or principles without any of these virtues could lead to uncaring doctor–patient relationships.

PROFESSIONAL OATHS AND CODES

Professional oaths and codes are useful guides for physicians. Most physicians take oaths at medical school white-coat ceremonies and graduations, and many are members of professional societies that have

professional codes. Physicians pledge to the public and to their patients that they will be guided by the principles and values in these oaths or codes. Oaths and codes—including the Hippocratic tradition—focus on ethical ideals rather than on daily pragmatic concerns, and have been criticized for lack of patient or public input and the limited role given to patients in making decisions.

PERSONAL VALUES

Personal values, cultural traditions, and religious beliefs are important sources of personal morality that help physicians address ethical issues and cope with the moral distress they may experience in practice. While essential, personal morality alone is a limited ethical guide in clinical practice. Physicians have role-specific ethical obligations that go beyond their obligations as good people, including the duties to obtain informed consent and maintain confidentiality discussed earlier. Furthermore, in a culturally and religiously diverse world, it is not uncommon for patients and colleagues to have personal moral beliefs that differ from those of their physicians.

ETHICALLY COMPLEX PROFESSIONAL ISSUES FOR PHYSICIANS

CLAIMS OF CONSCIENCE

Some physicians have conscientious objections to providing, or referring patients for, certain treatments such as contraception. Although physicians should not be asked to violate deeply held moral beliefs or religious convictions, patients need medically appropriate, timely care. Institutions such as clinics and hospitals have a collective duty to provide care that patients need while making reasonable attempts to accommodate health care workers' conscientious objections—for example, when possible by arranging for another professional to provide the service in question. Patients seeking a relationship with a doctor or health care institution should be notified in advance of any conscientious objections to the provision of specific interventions. Since patients commonly must select providers for insurance purposes, switching providers for a specific service can be burdensome. There are also important limits on claims of conscience. Health care workers may not insist that patients receive unwanted medical interventions and may not refuse to treat patients because of their race, ethnicity, national origin, gender, or religion. Such discrimination is illegal and violates the physician's duty to respect patients. While legally more controversial, refusal to treat patients because of their sexual orientation or gender identity is ethically inappropriate because it falls short of helping patients in need and respecting them as persons.

OCCUPATIONAL RISKS

Some health care workers, fearing fatal occupational infections, have refused to care for certain patients, such as those with HIV infection, Ebola virus disease, or severe acute respiratory syndrome (SARS). Such fears about personal safety need to be acknowledged. Health care institutions should reduce occupational risk by providing proper training, protective equipment, and supervision. To fulfill their mission of helping patients,

physicians should provide appropriate care within their clinical expertise, despite sometimes considerable personal risk.

MORAL DISTRESS

Health care providers, including residents and medical students, may experience moral distress when they feel that the ethically appropriate action to take in a particular situation is hindered by institutional policies, limited resources, decision-making hierarchies, or other reasons. Moral distress can lead to anger, anxiety, frustration, fatigue, and work dissatisfaction. Discussing complex or unfamiliar clinical situations with colleagues and seeking assistance with difficult decisions helps to alleviate moral distress, as does a healthy work environment characterized by open communication and mutual respect. In addition, physicians should take good care of their own well-being, and be aware of the personal and system factors associated with stress, burnout, and depression. A physician's health can affect how he or she cares for patients.

CONFLICTS OF INTEREST

Acting in patients' best interests may sometimes conflict with the physician's self-interest or the interests of third parties such as insurers or hospitals. From an ethical viewpoint, patients' interests are paramount. Even the appearance of a conflict of interest may undermine trust in the profession.

FINANCIAL INCENTIVES

Health care providers may be offered financial incentives to improve the quality or efficiency of care. Such pay-for-performance incentives, however, could lead physicians to avoid sicker patients with more complicated cases or to focus on benchmarked outcomes even when such a focus is not in the best interests of an individual patient. In contrast, fee-for-service payments might encourage physicians to order more interventions than may be necessary or to refer patients to laboratory or imaging facilities in which they have a financial stake. Regardless of financial incentives, physicians should recommend available care that is in the patient's best interests—no more and no less.

RELATIONSHIPS WITH PHARMACEUTICAL COMPANIES

Financial relationships between physicians and industry are increasingly scrutinized. Gifts from drug and device companies may create an inappropriate risk of undue influence, induce subconscious feelings of reciprocity, impair public trust, and increase the cost of health care. Many academic medical centers have banned drug-company gifts of branded pens and notepads and meals to physicians. The federal Open Payments website provides public information on the payments and amounts that drug and device companies give to individual physicians by name. The challenge is to distinguish payments for scientific consulting and research contracts—which are consistent with professional and academic missions and should be encouraged—from those for promotional speaking and consulting whose goal is to increase sales of company products.

LEARNING CLINICAL SKILLS

Not all conflicts of interest are financial. Medical students, residents, and physicians' interests in learning, which fosters the long-term goal of benefiting future patients, may conflict with the short-term goal of providing optimal care to current patients. When trainees are learning procedures on patients, they lack the proficiency of experienced physicians, and patients may experience inconvenience, discomfort, longer procedures, or increased risk. Seeking patients' consent for trainee participation in their care is always important, and particularly important for intimate examinations, such as pelvic, rectal, breast, and testicular examinations, and for invasive procedures. Patients should be told who is providing care and how trainees are supervised. Failing to introduce students or not telling patients that trainees will be performing procedures undermines trust, may lead to more elaborate deception, and makes it difficult for patients to make informed choices about their care. Most patients, when informed, allow trainees to play an active role in their care.

RESPONSE TO MEDICAL ERRORS

Errors are inevitable in clinical medicine, and some errors cause serious adverse events that harm patients. Most errors are caused by lapses of attention or flaws in the system of delivering health care; only a small number result from blameworthy individual behavior (**Chaps. 3 and 6**). Physicians and students may fear that disclosing errors will damage their careers. However, patients are owed an explanation, and appreciate being told when an error occurs, receiving an apology, and being informed about efforts to prevent similar errors in the future. Physicians and health care institutions show respect for patients by disclosing errors, offering appropriate compensation for harm done, and using errors as opportunities to improve the quality of care. Overall, patient safety is more likely to be improved through a quality improvement rather than a punitive approach to errors, except in cases of gross incompetence, physician impairment, boundary violations, or repeated violations of standard procedures.

PHYSICIAN IMPAIRMENT

Physicians may hesitate to intervene when colleagues impaired by alcohol abuse, drug abuse, or psychiatric or medical illness place patients at risk. However, society relies on physicians to regulate themselves. If colleagues of an impaired physician do not take steps to protect patients, no one else may be in a position to do so.

USE OF SOCIAL MEDIA

Increasingly, physicians use social and electronic media to share information with patients and other providers. Social networking may be especially useful in reaching young or otherwise hard-to-access patients. However, the use of social media, including blogs, social networks, and websites, raises ethical challenges and should be approached prudently to avoid harmful consequences for patients. Injudicious use of social media can pose risks to patient confidentiality, cross professional boundaries, and jeopardize therapeutic relationships. Internet and social networking postings are usually permanent and may be

accessible to the public, physicians' employers, and their patients. Unprofessional posts can lead to adverse consequences for a provider's reputation, safety, or even employment, especially if they express frustration or anger over work incidents, disparage patients or colleagues, use offensive or discriminatory language, reveal highly personal information, or picture a physician intoxicated, using illegal drugs, or in sexually suggestive poses. Physicians should separate professional from personal websites, social networking accounts and blogs, and should follow guidelines developed by institutions and professional societies on using social media to communicate with patients.

ETHICAL ISSUES IN CLINICAL RESEARCH

Clinical research is essential to translate scientific discoveries into beneficial tests and therapies for patients. However, clinical research raises ethical concerns because participants face inconvenience and risks in research, which is not designed specifically to benefit them but rather to advance scientific knowledge. Ethical guidelines for researchers require them to rigorously design research, minimize risk to participants, and obtain informed and voluntary consent from participants and approval from an institutional review board (IRB). IRBs determine that risks to participants are acceptable and have been minimized, and recommend appropriate additional protections when research includes vulnerable participants. Physicians may be involved as clinical research investigators or may be in a position to refer or recommend clinical trial participation to their patients. Physician-investigators may feel an inherent tension between conducting research and providing health care. Awareness of this tension, familiarity with the ethics of research, collaboration with others on the research and clinical teams, and utilizing research ethics consultation can help to mitigate the tension. Before starting clinical research, investigators should receive training in the ethics of clinical research. Courses and guidance on the ethics of clinical research are widely available.

Physicians should be critical consumers of clinical research results and keep up with the expanding scope of research and advances that change standards of practice. Precision medicine initiatives aim to individualize clinical care by sometimes combining clinical information from electronic health records, genomic sequencing of leftover biomaterials originally obtained for clinical care, and data from personal mobile devices. Furthermore, physicians and health care institutions are analyzing data routinely collected and available in electronic health records in order to improve the quality of care in real-world clinical settings; these efforts may be through quality improvement, comparative effectiveness research, or learning health care systems. These new types of research raise important issues about informed consent, privacy, and risk.

GLOBAL CONSIDERATIONS

International Research

 Clinical research is increasingly conducted at multiple sites and across national borders. Societal, legal, and cultural norms and perspectives about research may vary and there are many ethical challenges. Physician-investigators involved in international research should be familiar with international guidelines, such as the Declaration of Helsinki, the Council for International Organizations of Medical Sciences (CIOMS)

guidelines, and the International Council on Harmonisation Good Clinical Practice guidelines, as well as national and local laws where the research is taking place. Partnering with local researchers and communities is essential not only to demonstrate respect but also to facilitate successful clinical research.

Global Health Field Experiences

Many physicians and trainees choose to gain valuable experience by providing patient care in international settings. Such arrangements, however, raise ethical challenges—for example, as a result of differences in beliefs about health and illness, expectations regarding health care and the physician's role, standards of clinical practice, resource limitations, and norms for disclosure of serious diagnoses. Additional dilemmas arise if visiting physicians and trainees take on responsibilities beyond their expertise or if donated drugs and equipment are not appropriate to local needs. Visiting physicians and trainees should receive training and mentoring and seek information regarding cultural and clinical practices in the host community, respect local customs and values, work closely with local professionals and team members, and be explicit about their skills, knowledge, and limits. Leaders of global health field experiences should ensure that participating physicians receive training on ethical and cultural issues, mentoring, backup, and debriefing and that plans for evacuation are in place in case they are needed.

FURTHER READING

Beauchamp T, Childress J: *Principles of Biomedical Ethics*, 7th ed. New York, Oxford University Press, 2013.

Bilimoria KY et al: National Cluster-Randomized Trial of Duty-Hour Flexibility in Surgical Training. N Engl J Med 374:713, 2016. [PubMed: 26836220]

Emanuel EJ et al: What makes clinical research ethical? JAMA 283:2701, 2000. [PubMed: 10819955]

JAMA Conflict of Interest Theme Issue. JAMA 317:1707, 2017.

Lewis-Newby M et al: An official American Thoracic Society policy statement: Managing conscientious objections in intensive care medicine. Am J Respir Crit Care Med 191:219, 2015. [PubMed: 25590155]

Chapter 9: Palliative and End-of-Life Care

Ezekiel J. Emanuel

FIGURE 9-1

EPIDEMIOLOGY

CAUSES OF DEATH

In 2015, 2,712,630 individuals died in the United States (Table 9-1). Approximately 73% of these deaths occurred in those aged >65 years. The epidemiology of death has changed significantly since 1900 and even since 1980. In 1900, heart disease caused ~8% of all deaths and cancer accounted for <4% of all deaths. In 1980, heart disease accounted for 38.2% of all deaths, cancer 20.9%, and cerebrovascular disease 8.6% of all deaths. By 2014, there had been a dramatic drop in deaths from cardiovascular and cerebrovascular diseases. In 2014, 23.4% of all deaths were from cardiovascular disease and just 5.1% from cerebrovascular disease. Deaths attributable to cancer, however, had increased to 22.5%. The proportions of deaths due to chronic lower respiratory disease, diabetes, Alzheimer's, and suicides have also increased. Interestingly, in 2014, HIV/AIDS accounted for <0.26% of all U.S. deaths.

TABLE 9-1

Ten Leading Causes of Death in the United States and Britain

Cause of Death	United States (2014)		England and Wales (2015)	
	Number of Deaths, All Ages (%)	Number of Deaths, People ≥65 Years of Age	Number of Deaths, All Ages (%)	Number of Deaths, People ≥65 Years of Age
All deaths	2,626,418	1,922,271	529,655	449,409
Heart disease	614,348 (23.4)	489,722 (25.5)	114,345 (21.6)	99,029 (22.0)
Malignant neoplasms	591,699 (22.5)	413,885 (21.5)	144,330 (27.2)	115,302 (25.7)
Chronic lower respiratory diseases	147,101 (5.6)	124,693 (6.5)	30,368 (5.7)	27,674 (6.2)
Accidents	136,053 (5.2)	48,295 (2.5)	13,871 (2.6)	8,214 (1.8)
Cerebrovascular diseases	133,103 (5.1)	113,308 (5.9)	34,883 (6.6)	32,212 (7.2)
Alzheimer's disease	93,541 (3.6)	92,604 (4.8)	14,323 (2.7)	14,222 (3.2)
Diabetes mellitus	76,488 (2.9)	54,161 (2.8)	5,582 (1.1)	4,843 (1.1)
Influenza and pneumonia	55,227 (2.1)	44,836 (2.3)	29,885 (5.6)	27,982 (6.2)
Nephritis, nephritic syndrome, nephrosis	48,146 (1.8)	39,957 (2.1)	3,537 (0.7)	3,312 (0.7)
Intentional self-harm	42,773 (1.6)	-	4,150 (0.8)	727 (0.2)

Source: National Center for Health Statistics (United States, 2014), *http://www.cdc.gov/nchs;* National Statistics (Great Britain, 2015), *http://www.statistics.gov.uk*.

This change in the epidemiology of death is also reflected in the costs of illness. In the United States, ~84% of all health care spending goes to patients with chronic illnesses, and ~12% of total personal health care spending—slightly <$400 billion in 2015—goes to the 0.83% of the population in the last year of their lives.

In developed countries, an estimated ~70% of all deaths are preceded by a disease or condition, making it reasonable to plan for dying in the foreseeable future. Cancer has served as the paradigm for terminal care, but it is not the only type of illness with a recognizable and predictable terminal phase. Since heart failure, chronic obstructive pulmonary disease (COPD), chronic liver failure, dementia, and many other conditions have recognizable terminal phases, a systematic approach to end-of-life care should be part of all medical specialties. Many patients with illness-related suffering also can benefit from palliative care regardless of prognosis. Ideally, palliative care should be considered part of comprehensive care for all chronically ill patients. Reviews of the recent literature have found strong evidence that palliative care can be improved by coordination between caregivers, doctors, and patients for advance care planning, as well as dedicated teams of physicians, nurses, and other providers.

SITE OF DEATH

Where patients die varies by country **(Table 9-2)**. In Belgium and Canada, for instance, over half of all cancer patients still die in the hospital. The last few decades have seen a steady shift, both in the United States and other countries like the Netherlands, out of the hospital, as patients and their families list their own homes as the preferred site of death. In the early 1980s, about 70% of American cancer patients died in the hospital. Today that percentage is ~20% **(Fig. 9-1)**. A recent report shows that even since 2000, there has been a shift in the United States from inpatient to home deaths, especially for patients with cancer, COPD, and dementia. For instance, 30.1% of deaths due to cancer in 2000 occurred in acute care hospitals; by 2009, this figure had dropped to 22.1%.

TABLE 9-2

Elements of Communicating Bad News—The P-SPIKES Approach

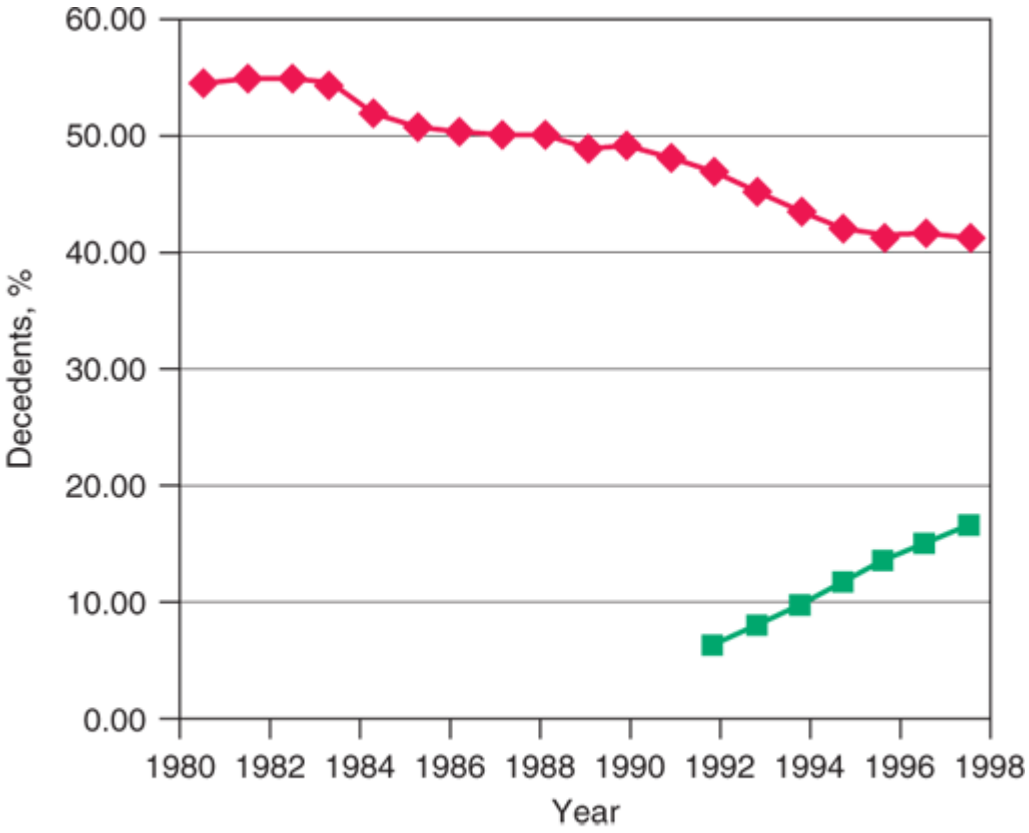
Acronym	Steps	Aim of the Interaction	Preparations, Questions, or Phrases
P	Preparation	Mentally prepare for the interaction with the patient and/or family.	Review what information needs to be communicated. Plan how you will provide emotional support. Rehearse key steps and phrases in the interaction.
S	Setting of the interaction	Ensure the appropriate setting for a serious and potentially emotionally charged discussion.	Ensure that patient, family, and appropriate social supports are present. Devote sufficient time. Ensure privacy and prevent interruptions by people or beeper. Bring a box of tissues.
P	Patient's perception and preparation	Begin the discussion by establishing the baseline and whether the patient and family can grasp the information. Ease tension by having the patient and family contribute.	Start with open-ended questions to encourage participation. Possible phrases to use: *What do you understand about your illness?* *When you first had symptom X, what did you think it might be?* *What did Dr. X tell you when he or she sent you here?* *What do you think is going to happen?*
I	Invitation and information needs	Discover what information needs the patient and/or family have and what limits they want regarding the bad information.	Possible phrases to use: *If this condition turns out to be something serious, do you want to know?* *Would you like me to tell you all the details of your condition? If not, who would you like me to talk to?*

Acronym	Steps	Aim of the Interaction	Preparations, Questions, or Phrases
K	Knowledge of the condition	Provide the bad news or other information to the patient and/or family sensitively.	Do not just dump the information on the patient and family. Check for patient and family understanding. Possible phrases to use: *I feel badly to have to tell you this, but…* *Unfortunately, the tests showed…* *I'm afraid the news is not good…*
E	Empathy and exploration	Identify the cause of the emotions—e.g., poor prognosis. Empathize with the patient and/or family's feelings. Explore by asking open-ended questions.	Strong feelings in reaction to bad news are normal. Acknowledge what the patient and family are feeling. Remind them such feelings are normal, even if frightening. Give them time to respond. Remind patient and family you won't abandon them. Possible phrases to use: *I imagine this is very hard for you to hear.* *You look very upset. Tell me how you are feeling.* *I wish the news were different.* *We'll do whatever we can to help you.*
S	Summary and planning	Delineate for the patient and the family the next steps, including additional tests or interventions.	It is the unknown and uncertain that can increase anxiety. Recommend a schedule with goals and landmarks. Provide your rationale for the patient and/or family to accept (or reject). If the patient and/or family are not ready to discuss the next steps, schedule a follow-up visit.

Source: Adapted from R Buckman: *How to Break Bad News: A Guide for Health Care Professionals.* Baltimore, Johns Hopkins University Press, 1992.

FIGURE 9-1

Graph showing trends in the site of death in the last two decades. ◆, percentage of hospital inpatient deaths; ■, percentage of decedents enrolled in a hospice.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Paradoxically, while deaths in acute care hospitals have declined in the United States since 2000, both hospitalizations in the last 90 days of life and—even more troublingly—admission to the ICU in the last 30 days have actually increased. Recent data show that >40% of cancer patients in the United States are admitted to the ICU in their last 6 months of life, and >25% of cancer patients are admitted to the hospital in the last 30 days.

The shift in deaths out of the hospital has been accompanied by an increase in the use of hospice in the United States. In 2000, 21.6% of all decedents used hospice at the time of death; by 2009, 42.2% were using hospice. Among cancer patients, ~60% were using hospice at the time of death. Hospice is also increasingly being used by non-cancer patients. Today, cancer patients constitute <40% of hospice users. About 79% of patients receiving hospice care die out of the hospital, and around 41% of those receiving hospice care die in a private residence.

In 2008, for the first time, the American Board of Medical Specialties (ABMS) offered certification in hospice and palliative medicine. With the shortening of hospital stays, many serious conditions are now being treated at home or on an outpatient basis. Consequently, providing optimal palliative and end-of-life care requires ensuring that appropriate services are available in a variety of settings, including non-institutional settings.

HOSPICE AND THE PALLIATIVE CARE FRAMEWORK

Central to this type of care is an interdisciplinary team approach that typically encompasses pain and symptom management, spiritual and psychological care for the patient, and support for family caregivers during the patient's illness and the bereavement period.

One of the more important changes in this field is beginning palliative care many months before death in order to focus on symptom relief, and then switching to hospice in the patient's last few months. This approach avoids leaving hospice until the very end by introducing palliative care earlier, thereby allowing patients and families time to transition. Phasing palliative care into end-of-life care means that patients will often receive palliative interventions long before they are formally diagnosed as terminally ill, or likely to die within 6 months.

Fundamental to ensuring quality palliative and end-of-life care is a focus on four broad domains: (1) physical symptoms; (2) psychological symptoms; (3) social needs that include interpersonal relationships, caregiving, and economic concerns; and (4) existential or spiritual needs.

ASSESSMENT AND CARE PLANNING

Comprehensive Assessment

Standardized methods for conducting a comprehensive assessment focus on evaluating the patient's condition in all four domains affected by the illness: physical, psychological, social, and spiritual.

A comprehensive assessment should follow a modified version of the traditional medical history and physical examination, and should emphasize both physical and mental symptoms. Questions should aim to elucidate symptoms, discern sources of suffering, and gauge how much those symptoms interfere with the patient's quality of life. Standardized and repeated assessments to evaluate the effectiveness of interventions are critical. Thus, clinicians should use shorter, validated instruments, such as: (1) The revised Edmonton Symptom Assessment Scale; (2) Condensed Memorial Symptom Assessment Scale (MSAS); (3) MD Anderson Brief Symptom Inventory; (4) Rotterdam Symptom Checklist; (5) Symptom Distress Scale; (6) Patient-Reported Outcomes Measurement Information System; and (7) The Interactive Symptom Assessment and Collection (ISAAC) tool.

Mental Health: With respect to mental health, many practices use the Patient Health Questionnaire-9 (PHQ-9) to screen for depression and the Generalized Anxiety Disorder-7 (GAD-7) to screen for anxiety. Using such tools ensures that the assessment is comprehensive and does not focus excessively on only pain.

Invasive Tests: Invasive tests are best avoided in end-of-life care, and even minimally invasive tests should be evaluated carefully for their benefit-to-burden ratio for the patient. Aspects of the physical examination that are uncomfortable and unlikely to yield useful information should be omitted.

Social Needs: Health care providers should also assess the status of important relationships, financial burdens, caregiving needs, and access to medical care. Relevant questions will include the following: *How often is there someone to feel close to? How has this illness been for your family? How has it affected your*

relationships? How much help do you need with things like getting meals and getting around? How much trouble do you have getting the medical care you need?

Existential Needs: To determine a patient's existential needs, providers should assess distress, the patient's sense of emotional and existential well-being, and whether the patient believes he or she has found purpose or meaning. Helpful assessment questions can include the following: *How much are you able to find meaning since your illness began? What things are most important to you at this stage?*

Perception of Care: In addition, it can be helpful to ask how the patient perceives his or her care: *How much do you feel your doctors and nurses respect you? How clear is the information from us about what to expect regarding your illness? How much do you feel that the medical care you are getting fits with your goals?* If concern is detected in any of these areas, deeper evaluative questions are warranted.

Communication

Particularly when an illness is life-threatening, there exists the potential for many emotionally charged and potentially conflict-creating moments—collectively called "bad news" situations—in which empathic and effective communication skills are essential. Those moments include the sharing of a terminal diagnosis with the patient and/or family, the discussion of patient's prognosis and any treatment failures, the consideration of deemphasizing efforts to cure and prolong life while focusing more on symptom management and palliation, advance care planning, and the patient's actual death. Although these conversations can be difficult, research indicates that end-of-life discussions can lead to earlier hospice referrals, rather than overly aggressive treatment, ultimately benefiting quality of life for patients and improving the bereavement process for families.

Just as surgeons prepare for major operations and investigators rehearse a presentation of research results, physicians and health care providers caring for patients with significant or advanced illnesses should develop a standardized approach for sharing important information and planning interventions. In addition, physicians must be aware that families often care not only about how prepared the physician was to deliver bad news, but also the setting in which it was delivered. For instance, one study found that 27% of families making critical decisions for patients in an intensive care unit (ICU) desired better and more private physical space to communicate with physicians.

One structured seven-step procedure for communicating bad news goes by the acronym P-SPIKES: (1) ***p***repare for the discussion, (2) ***s***et up a suitable environment, (3) begin the discussion by finding out what the ***p***atient and/or family understand, (4) determine how they will comprehend new ***i***nformation best and how much they want to know, (5) provide needed new ***k***nowledge accordingly, (6) allow for ***e***motional responses, and (7) ***s***hare plans for the next steps in care. Table 9-2 provides a summary of these steps, along with suggested phrases and underlying rationales for each one.

Continuous Goal Assessment

Major barriers to providing high-quality palliative and end-of-life care include the difficulty in determining both an accurate prognosis, and the emotional resistance of patients and their families to accepting the implications of a poor prognosis. A practical solution to these barriers is to integrate palliative care interventions or home visits from a palliative care visiting nurse months before the estimated final 6 months of life. Under this approach, palliative care no longer conveys the message of failure, having no more treatments, or "giving up hope." The transition from palliative to end-of-life care or hospice also feels less hasty and unexpected to the family. Fundamental to integrating palliative care with curative therapy is the inclusion of a continuous goal assessment as part of the routine patient reassessments that occur at most patient-physician encounters.

Goals for care are numerous, ranging from curing a specific disease, to prolonging life, to relieving a particular symptom, to adapting to a progressive disability without disrupting the family, to finding peace of mind or personal meaning, to dying in a manner that leaves loved ones with positive memories. Discerning a patient's goals for care can be approached through a seven-step protocol: (1) ensure that medical and other information is as complete as reasonably possible and is understood by all relevant parties (see above); (2) explore what the patient and/or family is hoping for, while also identifying relevant and realistic goals; (3) share all the options with the patient and family; (4) respond with empathy as they adjust to changing expectations; (5) make a plan that emphasizes what can be done to achieve the realistic goals; (6) follow through with the plan; and (7) periodically review the plan and consider at every encounter whether the goals of care should be revised with the patient and/or family. Each of these steps need not be followed in rote order, but together they provide a helpful framework for interactions with patients and their families regarding their goals for care. Such interactions can be especially challenging if a patient or family member has difficulty letting go of an unrealistic goal. In such cases, the provider should help them refocus on more realistic goals, and should also suggest that while it is fine to hope for the best, it is still prudent to plan for other outcomes as well.

Advance Care Planning

PRACTICES

Advance care planning is the process of planning for future medical care in case the patient becomes incapable of making medical decisions. A 2010 study of adults aged ≥60 who died between 2000 and 2006 found that while 42% of adults were required to make treatment decisions in their final days of life, 70% lacked decision-making capacity. Among those lacking decision-making capacity, approximately one-third did not have advance planning directives. Ideally, such planning would occur before a health care crisis or the terminal phase of an illness. Unfortunately, diverse barriers prevent this. Approximately 80% of Americans endorse advance care planning and living wills. However, according to a Pew survey only 35% of adults have written down their end-of-life wishes. Other studies report even fewer Americans—with some estimates as low as 26% of adults—having filled out advance care directives. Larger numbers of adults, between 50 and 70%, claim to have talked with someone about their treatment wishes.

Effective advance care planning should follow six key steps: (1) introducing the topic, (2) structuring a discussion, (3) reviewing plans that have been discussed by the patient and family, (4) documenting the

plans, (5) updating them periodically, and (6) implementing the advance care directives (Table 9-3). Two of the main barriers to advance care planning are problems in raising the topic and difficulty in structuring a succinct discussion. Raising the topic can be done efficiently as a routine matter, noting that it is recommended for all patients, analogous to purchasing insurance or estate planning. Many of the most difficult cases have involved unexpected, acute episodes of brain damage in young individuals.

TABLE 9-3

Steps in Advance Care Planning

Step	Goals to be Achieved and Measures to Cover	Useful Phrases or Points to Make
Introducing advance care planning	Ask the patient what he or she knows about advance care planning and if he or she has already completed an advance care directive.	*I'd like to talk with you about something I try to discuss with all my patients. It's called advance care planning. In fact, I feel that this is such an important topic that I have done this myself. Are you familiar with advance care planning or living wills?*
	Indicate that you as a physician have completed advance care planning.	*Have you thought about the type of care you would want if you ever became too sick to speak for yourself? That is the purpose of advance care planning.*
	Indicate that you try to perform advance care planning with all patients regardless of prognosis.	*There is no change in health that we have not discussed. I am bringing this up now because it is sensible for everyone, no matter how well or ill, old or young.*
	Explain the goals of the process as empowering the patient and ensuring that you and the proxy understand the patient's preferences.	*Have many copies of advance care directives available, including in the waiting room, for patients and families.*
	Provide the patient relevant literature, including the advance care directive that you prefer to use.	*Know resources for state-specific forms (available at www.nhpco.org).*

Step	Goals to be Achieved and Measures to Cover	Useful Phrases or Points to Make
	Recommend the patient identify a proxy decision-maker who should attend the next meeting.	
Structured discussion of scenarios and patient	Affirm that the goal of the process is to follow the patient's wishes if the patient loses decision-making capacity.	*Use a structured worksheet with typical scenarios.*
	Elicit the patient's overall goals related to health care. Elicit the patient's preferences for specific interventions in a few salient and common scenarios. Help the patient define the threshold for withdrawing and withholding interventions. Define the patient's preference for the role of the proxy.	*Begin the discussion with persistent vegetative state and consider other scenarios, such as recovery from an acute event with serious disability, asking the patient about his or her preferences regarding specific interventions, such as ventilators, artificial nutrition, and CPR, and then proceeding to less invasive interventions, such as blood transfusions and antibiotics.*
Review the patient's preferences	After the patient has made choices of interventions, review them to ensure they are consistent and the proxy is aware of them.	

Step	Goals to be Achieved and Measures to Cover	Useful Phrases or Points to Make
Document the patient's preferences	Formally complete the advance care directive and have a witness sign it.	
	Provide a copy for the patient and the proxy.	
	Insert a copy into the patient's medical record and summarize in a progress note.	
Update the directive	Periodically, and with major changes in health status, review the directive with the patient and make any modifications.	
Apply the directive	The directive goes into effect only when the patient becomes unable to make medical decisions for himself or herself.	
	Reread the directive to be sure about its content.	
	Discuss your proposed actions based on the directive with the proxy.	

Abbreviation: CPR, cardiopulmonary resuscitation.

Structuring a focused discussion is an important communication skill. To do so, a provider must first identify the health care proxy and recommend his or her involvement in the advance care planning process. Next, a worksheet must be selected that has been demonstrated to produce reliable and valid expressions of patient preferences, and the patient and proxy must be oriented to it. Such worksheets exist for both general and disease-specific situations. The provider should then discuss with the patient and proxy one example scenario to demonstrate how to think about the issues. It is often helpful to begin with a scenario in which the patient is likely to have settled preferences for care, such as being in a persistent vegetative state. Once the patient's preferences for interventions in this scenario are determined, the provider should suggest that the patient and proxy discuss and complete the worksheet for each other. If appropriate, the patient and proxy should consider involving other family members in the discussion. During a subsequent return visit, the provider should go over the patient's preferences, checking and resolving any inconsistencies. After having the patient and proxy sign the document, the provider should place the document in the patient's medical chart and make sure that copies are provided to relevant family members and care sites. Since patients' preferences can change, these documents must be reviewed periodically.

TYPES OF DOCUMENTS

There are two broad types of advance care planning documents. The first includes living wills, also known as instructional directives; these are advisory documents that describe the types of decisions that should direct a patient's care. Some are more specific, delineating different scenarios and interventions for the patient to choose from. Among these, some are for general use and others are designed for use by patients with a specific type of disease, such as cancer, renal failure, or HIV. Less specific directives can be general statements, such as not wanting life-sustaining interventions, or forms that describe the values that should guide specific discussions about terminal care. The second type of advance directive allows the designation of a health care proxy (sometimes also referred to as a durable attorney for health care), an individual selected by the patient to make decisions. The choice is not either/or; a combined directive that includes a living will and designates a proxy is often used, and the directive should indicate clearly whether the specified patient preferences or the proxy's choice takes precedence if they conflict. Some states have begun to put into practice a "Physician Orders for Life-Sustaining Treatment (POLST)" directive, which builds on communication between providers and patients by including guidance for end-of-life care in a color-coordinated form that follows the patient across treatment settings. The procedures for completing advance care planning documents vary according to state law.

A potentially misleading distinction relates to statutory, as opposed to advisory, documents. Statutory documents are drafted to fulfill relevant state laws. Advisory documents are drafted to reflect the patient's wishes. Both are legal, the first under state law and the latter under common or constitutional law.

LEGAL ASPECTS

As of 2017, 48 states and the District of Columbia had enacted living will legislation. Massachusetts and Michigan are the two states without living will legislation. Indiana has a life-prolonging procedures declaration. States differ in the requirements for advanced directives, whether they need to be witnessed, by how many witnesses, or notarized. Importantly, in 26 states, the laws state that the living will is not valid if a woman is pregnant. All states except Alaska have enacted durable power of attorney for health care laws that

permits patients to designate a proxy decision-maker with authority to terminate life-sustaining treatments. Only in Alaska does the law prohibit proxies from terminating life-sustaining treatments for pregnant women.

The U.S. Supreme Court has ruled that patients have a constitutional right to decide any issues related to refusing or terminating medical interventions, including life-sustaining interventions, and that mentally incompetent patients can exercise this right by providing “clear and convincing evidence” of their preferences. Since advance care directives permit patients to provide such evidence, commentators agree that they are constitutionally protected. Most commentators believe that a state is required to honor any clear advance care directive, regardless of whether it is written on an “official” form. Many states have enacted laws for the explicit purpose of honoring out-of-state directives. If a patient is not using a statutory form, it may be advisable to attach a statutory form to the advance care directive being used. State-specific forms are readily available free of charge for health care providers, patients, and families through the website of the National Hospice and Palliative Care Organization (*http://www.nhpco.org*).

Reimbursement: As of January 1, 2016, the Center for Medicare and Medicaid Services amended the physician fee schedule to reimburse discussions of advance care planning (ACP) under CPT codes 99497 and 99498. The session must be voluntary and include an explanation of advance care planning, but need not include a completed advance care document. There can be multiple bills for the discussion if it extends over several encounters.

INTERVENTIONS

PHYSICAL SYMPTOMS AND THEIR MANAGEMENT

Great emphasis has been placed on addressing dying patients’ pain. In order to emphasize its importance, pain assessment has frequently been included as the fifth vital sign. Heightened consideration of pain has been advocated by large health care systems such as the Veterans’ Administration and accrediting bodies such as the Joint Commission on the Accreditation of Health Care Organizations (JCAHO). Although this embrace of pain has been symbolically important, available data suggests that making pain the fifth vital sign does not lead to improved pain management practices. In light of the growing opioid crisis in the United States, the emphasis on pain management has begun to be re-examined. For instance, in 2017 draft standards, the JCAHO recommends nonpharmacological pain treatment as well as identification of psychosocial risk factors for addiction. Importantly, good palliative care requires much more than good pain management. The frequency of symptoms varies by disease and other factors. The most common physical and psychological symptoms among all terminally ill patients include pain, fatigue, insomnia, anorexia, dyspnea, depression, anxiety, nausea, and vomiting. In the last days of life, terminal delirium is also common. Assessments of patients with advanced cancer have shown that patients experienced an average of 11.5 different physical and psychological symptoms **(Table 9-4)**.

TABLE 9-4

Common Physical and Psychological Symptoms of Terminally Ill Patients

Physical Symptoms	Psychological Symptoms
Pain	Anxiety
Fatigue and weakness	Depression
Dyspnea	Hopelessness
Insomnia	Meaninglessness
Dry mouth	Irritability
Anorexia	Impaired concentration
Nausea and vomiting	Confusion
Constipation	Delirium
Cough	Loss of libido
Swelling of arms or legs	
Itching	
Diarrhea	
Dysphagia	
Dizziness	
Fecal and urinary incontinence	
Numbness/tingling in hands/feet	

In the vast majority of cases, evaluations to determine the etiology of these symptoms should be limited to the history and physical examination. In some cases, radiologic or other diagnostic examinations will provide sufficient benefit in directing optimal palliative care to warrant the risks, potential discomfort, and inconvenience, especially to a seriously ill patient. Only a few of the common symptoms that present difficult

management issues will be addressed in this chapter. **Additional information on the management of other symptoms, such as nausea and vomiting, insomnia, and diarrhea, can be found in Chaps. 41, 65, 27, and 42, respectively.**

Pain

FREQUENCY

The frequency of pain among terminally ill patients varies significantly. Cancer (~85%), CHF (~75%), and AIDS have been associated with a higher prevalence of pain compared to other advanced illnesses, such as COPD (~45%), CKD (~40%), and dementia (~40%). One meta-analysis of adults with advanced or terminal illness found pain prevalence of 30–94% in patients with cancer, compared to 21–77% for COPD, 14–78% for CHF, 11–83% for ESRD, 14–63% for dementia, and 30–98% for AIDS.

ETIOLOGY

There are two types of pain: nociceptive and neuropathic. Nociceptive pain is further divided into somatic or visceral pain. *Somatic pain* is the result of direct mechanical or chemical stimulation of nociceptors and normal neural signaling to the brain. It tends to be localized, aching, throbbing, and cramping. The classic example is bone metastases. *Visceral pain* is caused by nociceptors in gastrointestinal (GI), respiratory, and other organ systems. It is a deep or colicky type of pain classically associated with pancreatitis, myocardial infarction, or tumor invasion of viscera. *Neuropathic pain* arises from disordered nerve signals. It is described by patients as burning, electrical, or shock-like pain. Classic examples are post-stroke pain, tumor invasion of the brachial plexus, and herpetic neuralgia.

ASSESSMENT

Pain is a subjective experience. Depending on the patient's circumstances, perspective, and physiologic condition, the same physical lesion or disease state can produce different levels of reported pain and need for pain relief. Systematic assessment includes eliciting the following: (1) type: throbbing, cramping, burning, etc.; (2) periodicity: continuous, with or without exacerbations, or incident; (3) location; (4) intensity; (5) modifying factors; (6) effects of treatments; (7) functional impact; and (8) impact on patient. Several validated pain assessment measures may be used, including the Visual Analogue Scale (VAS), the Brief Pain Inventory (BPI), or the Numerical Pain Rating Scale (NRS-11). Other scales have been developed for neuropathic pain, such as the Neuropathic Pain Scale and the DN4 Questionnaire. Frequent reassessments on a consistent scale are essential to assess the impact of and need to readjust interventions.

INTERVENTIONS

Interventions for pain must be tailored to each individual, with the goal of preempting chronic pain and relieving breakthrough pain. At the end of life, there is rarely reason to doubt a patient's report of pain. With the opioid crisis in the United States there is more emphasis on making opioids one component of multimodal analgesia. Nevertheless, at the end of life, pain medications, especially opioids, remain the cornerstone of management. If they are failing and nonpharmacologic interventions—including radiotherapy and anesthetic or neurosurgical procedures such as peripheral nerve blocks or epidural medications—are required, a pain consultation is appropriate.

Pharmacologic interventions still largely follow the World Health Organization three-step, “analgesic ladder” approach, which involves non-opioid analgesics, “mild” opioids, and “strong” opioids, with or without adjuvants **(Chap. 10)**. Nonopioid analgesics, especially nonsteroidal anti-inflammatory drugs (NSAIDs), are the initial treatments for mild pain. They work primarily by inhibiting peripheral prostaglandins and reducing inflammation, but may also have central nervous system (CNS) effects. Additionally, NSAIDs have a ceiling effect. Ibuprofen, up to 2400 mg/d qid, has a minimal risk of causing bleeding and renal impairment and is a good initial choice. In patients with a history of severe GI or other bleeding, however, ibuprofen should be avoided. In patients with a history of mild gastritis or gastroesophageal reflux disease (GERD), acid-lowering therapy, such as a proton pump inhibitor, should be used. Acetaminophen is an alternative in patients with a history of GI bleeding and can be used safely at up to 4 g/d qid. In patients with liver dysfunction due to metastases or other causes, and in patients with heavy alcohol use, doses should be reduced.

If nonopioid analgesics are insufficient, opioids should be introduced. Opioids primarily work by interacting with μ opioid receptors to activate pain-inhibitory neurons in the CNS, although they also interact variably with δ and κ receptors. Receptor agonists, such as morphine, codeine, and fentanyl, produce analgesia by activating pain-inhibitory neurons in the CNS. Partial agonists, such as buprenorphine, have a ceiling effect for analgesia and a lower potential for abuse. They are useful for post-acute pain, but should not be used for chronic pain in end-of-life care. Pure antagonists, such as naloxone and methylnaltrexone, are used for reversal of opioid effects.

Traditionally, “weak” opioids such as codeine were used first. If they failed to relieve pain after dose escalation, “strong” opioids like morphine were used in doses of 5–10 mg every 4 h. However, this breakdown between “weak” and “strong” opioids is no longer commonly accepted, with smaller doses of “stronger” opioids frequently being preferred over similar or larger doses of “weaker” opioids, and different pain syndromes having different preferred therapies. Regardless, nonopioid analgesics should be combined with opioids, as they potentiate the effect of opioids.

For continuous pain, opioids should be administered on a regular, around-the-clock basis consistent with their duration of analgesia. They should not be provided only when the patient experiences pain; the goal is to prevent patients from experiencing pain. Patients should also be provided rescue medication, such as liquid morphine, for breakthrough pain, generally at 20% of the baseline dose. Patients should be informed that using the rescue medication does not obviate the need to take the next standard dose of pain medication. If the patient’s pain remains uncontrolled after 24 h and recurs before the next dose, requiring the patient to utilize the rescue medication, the daily opioid dose can be increased by the total dose of rescue medications used by the patient, or by 50% of the standing opioid daily dose for moderate pain and 100% for severe pain.

It is inappropriate to start with extended-release preparations. Instead, an initial focus on using short-acting preparations to determine how much is required in the first 24–48 h will allow clinicians to determine opioid needs. Once pain relief is obtained using short-acting preparations, the switch should be made to extended-release preparations. Even with a stable extended-release preparation regimen, the patient may experience incident pain, such as during movement or dressing changes. Short-acting preparations should be taken

before such predictable episodes. Although less common, patients may have “end-of-dose failure” with long-acting opioids, meaning that they develop pain after 8 h in the case of an every-12-h medication. In these cases, a trial of giving an every-12-h medication every 8 h is appropriate.

Due to differences in opioid receptors, cross-tolerance among opioids is incomplete, and patients may experience different side effects with different opioids. Therefore, if a patient is not experiencing pain relief or is experiencing too many side effects, a change to another opioid preparation is appropriate. When switching, one should begin with 50–75% of the published equianalgesic dose of the new opioid.

Unlike NSAIDs, opioids have no ceiling effect; therefore, there is no maximum dose, no matter how many milligrams the patient is receiving. The appropriate dose is the dose needed to achieve pain relief. This is an important point for clinicians to explain to patients and families. Addiction or excessive respiratory depression is extremely unlikely in the terminally ill; fear of these side effects should neither prevent escalating opioid medications when the patient is experiencing insufficient pain relief, nor justify using opioid antagonists.

Opioid side effects should be anticipated and treated preemptively. Nearly all patients experience constipation that can be debilitating (see below). Failure to prevent constipation often results in noncompliance with opioid therapy. The preferred treatment is prevention. Cathartics (senna 2 tables qHS), stool softeners (docusate 100 mg PO qd), and/or laxatives (laxtulose 30 mL qd) are considered first-line. For refractory cases, opioid antagonists or other therapies, such as lubiprostone, should be considered.

Methylnaltrexone is the best-studied opioid antagonist for use in refractory opioid-induced constipation. It reverses opioid-induced constipation by blocking peripheral opioid receptors, but not central receptors, for analgesia. In placebo-controlled trials, it has been shown to cause laxation within 24 h of administration. As with the use of opioids, about a third of patients using methylnaltrexone experience nausea and vomiting, but unlike with opioid usage, tolerance usually develops within a week. Therefore, when one is beginning opioids, an antiemetic such as metoclopramide or a serotonin antagonist is often prescribed prophylactically and stopped after 1 week. Olanzapine has also been shown to have anti-nausea properties and can be effective in countering delirium or anxiety, with the advantage of some weight gain.

Drowsiness, a common side effect of opioids, also usually abates within a week. For refractory or severe cases, pharmacologic therapy should be considered. The best-studied agents are the psychostimulants dextroamphetamine, methylphenidate, and modafinil, although evidence regarding their efficacy is weak. Modafinil has the advantage of once-a-day dosing compared to methyphenidate’s twice daily dosing.

Seriously ill patients who require chronic pain relief rarely become addicted. Suspicion of addiction should not be a reason to withhold pain medications from terminally ill patients. Nonetheless, patients and families may withhold prescribed opioids for fear of addiction or dependence. Physicians and health care providers should reassure patients and families that the patient will not become addicted to opioids if they are used as prescribed for pain relief; this fear should not prevent the patient from taking the medications around the clock. However, diversion of drugs for use by other family members or illicit sale may occur. It may be

necessary to advise the patient and caregiver about secure storage of opioids. Contract writing with the patient and family can help. If that fails, transfer to a safe facility may be necessary.

Tolerance describes the need to increase medication dosage for the same pain relief without a concurrent change in disease. In the case of patients with advanced disease, the need for increasing opioid dosage for pain relief usually is caused by disease progression rather than tolerance. Physical dependence is indicated by symptoms resulting from the abrupt withdrawal of opioids and should not be confused with addiction.

In recent years, the potential dangers of opioid drugs have become increasingly apparent. To help mitigate the risk of these powerful drugs, several strategies should be used to reduce the risk of aberrant drug use. To start, all patients should be assessed for their individual levels of risk. While there are multiple surveys available, including the Opioid Risk Tool, none have gained wide-spread use or validation. In general, however, it is important to screen for prior substance abuse and major psychiatric disorders.

For patients deemed to be high-risk, a multidisciplinary effort should be pursued to reduce the risk of adverse consequences, such as addiction and diversion. Prescribing strategies include selecting opioids with longer durations of action and lower street values, such as methadone, and prescribing smaller quantities with more frequent follow-up. Monitoring options include periodic urine screening and referral to pain specialists. In some cases, it may also be reasonable to consider not offering short-acting opioids for breakthrough pain. In no situation, however, should adequate pain-relief be withheld due to risk.

Adjuvant analgesic medications are nonopioids that potentiate the analgesic effects of opioids. They are especially important in the management of neuropathic pain. Gabapentin, an anticonvulsant initially studied in the setting of herpetic neuralgia, is now the first-line treatment for neuropathic pain resulting from a variety of causes. It is begun at 100–300 mg bid or tid, with 50–100% dose increments every 3 days. Usually 900–3600 mg/d in two or three doses is effective. The combination of gabapentin and nortriptyline may be more effective than gabapentin alone. One potential side effect of gabapentin to be aware of is confusion and drowsiness, especially in the elderly. Other effective adjuvant medications include pregabalin, which has the same mechanism of action as gabapentin, but is absorbed more efficiently from the GI tract. Lamotrigine is a novel agent whose mechanism of action is unknown, but has been shown to be effective. It is recommended to begin at 25–50 mg/d, increasing to 100 mg/d. Carbamazepine, a first-generation agent, has been proven effective in randomized trials for neuropathic pain. Other potentially effective anticonvulsant adjuvants include topiramate (25–50 mg qd or bid, rising to 100–300 mg/d) and oxcarbazepine (75–300 mg bid, rising to 1200 mg bid).

Glucocorticoids, preferably dexamethasone given once a day, can be useful in reducing inflammation that causes pain, while also elevating mood, energy, and appetite. Its main side effects include confusion, sleep difficulties, and fluid retention. Glucocorticoids are especially effective for bone pain and abdominal pain from distention of the GI tract or liver. Other drugs, including clonidine and baclofen, can be effective in providing pain relief. These drugs are adjuvants and generally should be used in conjunction with—not instead of—opioids. Methadone, carefully dosed because of its unpredictable half-life in many patients, has activity at the *N*-methyl-D-aspartamate (NMDA) receptor and is useful for complex pain syndromes and

neuropathic pain. It is generally reserved for cases in which first-line opioids (morphine, oxycodone, hydromorphone) are either ineffective or unavailable.

Radiation therapy can treat bone pain from single metastatic lesions. Bone pain from multiple metastases can be amenable to radiopharmaceuticals such as strontium 89 and samarium 153. Bisphosphonates, such as pamidronate (90 mg every 4 weeks) and calcitonin (200 IU intranasally once or twice a day), also provide relief from bone pain, but have multi-day onsets of action.

Constipation

FREQUENCY

Constipation is reported in up to 70–100% of patients requiring palliative care.

ETIOLOGY

Although hypercalcemia and other factors can cause constipation, it is most frequently a predictable consequence of the use of opioids for pain and dyspnea relief, and of the anticholinergic effects of tricyclic anti-depressants, as well as due to the inactivity and poor diets common among seriously ill patients. If left untreated, constipation can cause substantial pain and vomiting, and also is associated with confusion and delirium. Whenever opioids and other medications known to cause constipation are used, preemptive treatment for constipation should be instituted.

ASSESSMENT

Assessing constipation can be difficult, because people describe it differently. Four commonly used assessment scales are the Bristol Stool Form Scale, the Constipation Assessment Scale, the Constipation Visual Analogue Scale, and the Eton Scale Risk Assessment for Constipation. The Bowel Function Index can be used to quantify opioid induced constipation. The physician should establish the patient's previous bowel habits, as well as any changes in subjective and objective qualities such as bloating or decreased frequency. Abdominal and rectal examinations should be performed to exclude impaction or an acute abdomen. Radiographic assessments beyond a simple flat plate of the abdomen in cases in which obstruction is suspected are rarely necessary.

INTERVENTION

Any measure to address constipation during end-of-life care should include interventions to reestablish comfortable bowel habits and to relieve pain or discomfort. Although physical activity, adequate hydration, and dietary treatments with fiber can be helpful, each is limited in its effectiveness for most seriously ill patients, and fiber may exacerbate problems in the setting of dehydration or if impaired motility is the etiology. Fiber is contraindicated in the presence of opioid use. Stimulant and osmotic laxatives, stool softeners, fluids, and enemas are the mainstays of therapy **(Table 9-5)**. To prevent constipation from opioids and other medications, a combination of a laxative and a stool softener (such as senna and docusate) should be used. If after several days of treatment a bowel movement has not occurred, a rectal examination to remove impacted stool and place a suppository is necessary. For patients with impending bowel obstruction or gastric stasis, octreotide to reduce secretions can be helpful. For patients in whom the suspected mechanism is dysmotility, metoclopramide can be helpful.

TABLE 9-5

Medications for the Management of Constipation

Intervention	Dose	Comment
Stimulant laxatives		These agents directly stimulate peristalsis and may reduce colonic absorption of water.
Prune juice	120–240 mL/d	Work in 6–12 h.
Senna (Senokot)	2–8 tablets PO bid	
Bisacodyl	5–15 mg/d PO, PR	
Osmotic laxatives		These agents are not absorbed. They attract and retain water in the gastrointestinal tract.
Lactulose	15–30 mL PO q4–8h	Lactulose may cause flatulence and bloating.
Magnesium hydroxide (Milk of Magnesia)	15–30 mL/d PO	Lactulose works in 1 day, magnesium products in 6 h.
Magnesium citrate	125–250 mL/d PO	
Stool softeners		These agents work by increasing water secretion and as detergents, increasing water penetration into the stool.
Sodium docusate (Colace)	300–600 mg/d PO	Work in 1–3 days.
Calcium docusate	300–600 mg/d PO	
Suppositories and enemas		

Intervention	Dose	Comment
Bisacodyl	10–15 PR qd	
Sodium phosphate enema	PR qd	Fixed dose, 4.5 oz, Fleet's.

Nausea

FREQUENCY

Up to 70% of patients with advanced cancer have nausea, defined as the subjective sensation of wanting to vomit.

ETIOLOGY

Nausea and vomiting are both caused by stimulation at one of four sites: the GI tract, the vestibular system, the chemoreceptor trigger zone (CTZ), and the cerebral cortex. Medical treatments for nausea are aimed at receptors at each of these sites: The GI tract contains mechanoreceptors, chemoreceptors, and 5-hydroxytryptamine type 3 (5-HT3) receptors; the vestibular system probably contains histamine and acetylcholine receptors; and the CTZ contains chemoreceptors, dopamine type 2 receptors, and 5-HT3 receptors. An example of nausea that most likely is mediated by the cortex is anticipatory nausea before a dose of chemotherapy or other noxious stimuli.

Specific causes of nausea include metabolic changes (liver failure, uremia from renal failure, hypercalcemia), bowel obstruction, constipation, infection, GERD, vestibular disease, brain metastases, medications (including antibiotics, NSAIDs, proton pump inhibitors, opioids, and chemotherapy), and radiation therapy. Anxiety can also contribute to nausea.

INTERVENTION

Medical treatment of nausea is directed at the anatomic and receptor-mediated cause revealed by a careful history and physical examination. When no specific cause of nausea is identified, many advocate beginning treatment with either metoclopramide, a serotonin type 3 (5-HT3) receptor antagonist like ondansetron, granisetron, palonosetron, dolasetron, tropisetron, or ramosetron, or a dopamine antagonist such as chlorpromazine, haloperidol or prochlorperazine. When decreased motility is suspected, metoclopramide can be an effective treatment. When inflammation of the GI tract is suspected, glucocorticoids, such as dexamethasone, are an appropriate treatment. For nausea that follows chemotherapy and radiation therapy, one of the 5-HT3 receptor antagonists or neurokinin-1 antagonists, such as aprepitant or fosaprepitant, is recommended. Clinicians should attempt prevention of post-chemotherapy nausea, rather than simply providing treatment after the fact. Current clinical guidelines recommend tailoring the strength of treatments to the specific emetic risk posed by a specific chemotherapy drug. When a vestibular cause (such as "motion sickness" or labyrinthitis) is suspected, antihistamines, such as meclizine (whose primary side effect is

drowsiness), or anticholinergics, such as scopolamine, can be effective. In anticipatory nausea, patients can benefit from non-pharmacological interventions, such as biofeedback and hypnosis. The most common pharmacological intervention for anticipatory nausea is a benzodiazepine, such as lorazepam. As with antihistamines, drowsiness and confusion are the main side effects.

The use of medical marijuana or oral cannabinoids for palliative treatment of nausea is controversial, as there are no controlled trials showing its effectiveness for patients at the end of life. A 2015 meta-analysis showed “low-quality evidence suggesting that cannabinoids were associated with improvements in nausea and vomiting due to chemotherapy,” and such treatments are not as good as 5-HT3 receptor antagonists and can sometimes even cause cannabis hyperemesis syndrome. Older patients—the vast majority of dying patients— seem to tolerate cannabinoids poorly.

Dyspnea

FREQUENCY

Dyspnea is the subjective experience of being short of breath. Over 50%, and as many as 75%, of dying patients, especially those with lung cancer, congestive heart failure and COPD, experience dyspnea at some point near the end of life. Dyspnea is among the most distressing of physical symptoms and can be even more distressing than pain.

ASSESSMENT

As with pain, dyspnea is a subjective experience that may not correlate with objective measures of P_{O2}, P_{CO2}, or respiratory rate. Consequently, measurements of oxygen saturation through pulse oximetry or blood gases are rarely helpful in guiding therapy. Despite the limitations of existing assessment methods, physicians should regularly assess and document patients’ experience of dyspnea and its intensity. Guidelines recommend visual analogue dyspnea scales to assess the severity of symptoms and the effects of treatment. Potentially reversible or treatable causes of dyspnea include infection, pleural effusions, pulmonary emboli, pulmonary edema, asthma, and tumor encroachment on the airway. However, the risk-versus-benefit ratio of the diagnostic and therapeutic interventions for patients with little time left to live must be considered carefully before undertaking diagnostic steps. Frequently, the specific etiology cannot be identified, and dyspnea is the consequence of progression of the underlying disease that cannot be treated. The anxiety caused by dyspnea and the choking sensation can significantly exacerbate the underlying dyspnea in a negatively reinforcing cycle.

INTERVENTIONS

When reversible or treatable etiologies are diagnosed, they should be treated as long as the side effects of treatment, such as repeated drainage of effusions or anticoagulants, are less burdensome than the dyspnea itself. More aggressive treatments such as stenting a bronchial lesion may be warranted if it is clear that the dyspnea is due to tumor invasion at that site and if the patient and family understand the risks of such a procedure.

Usually, treatment will be symptomatic **(Table 9-6)**. Supplemental oxygen does not appear to be effective. “A systematic review of the literature failed to demonstrate a consistent beneficial effect of oxygen inhalation

over air inhalation for study participants with dyspnea due to end-stage cancer or cardiac failure." Therefore, oxygen may be no more than an expensive placebo. Low-dose opioids reduce the sensitivity of the central respiratory center and relieve the sensation of dyspnea. If patients are not receiving opioids, weak opioids can be initiated; if patients are already receiving opioids, morphine or other stronger opioids should be used. Controlled trials do not support the use of nebulized opioids for dyspnea at the end of life. Phenothiazines and chlorpromazine may be helpful when combined with opioids. Benzodiazepines can be helpful in treating dyspnea, but only if anxiety is present. Benzodiazepines should neither be used as first-line therapy nor if there is no anxiety. If the patient has a history of COPD or asthma, inhaled bronchodilators and glucocorticoids may be helpful. If the patient has pulmonary edema due to heart failure, diuresis with a medication such as furosemide is indicated. Excess secretions can be transdermally or intravenously dried with scopolamine. More general interventions that medical staff can perform include sitting the patient upright, removing smoke or other irritants like perfume, ensuring a supply of fresh air with sufficient humidity, and minimizing other factors that can increase anxiety.

TABLE 9-6

Medications for the Management of Dyspnea

Intervention	Dose	Comments
Weak opioids		For patients with mild dyspnea
Codeine (or codeine with 325 mg acetaminophen)	30 mg PO q4h	For opioid-naïve patients
Hydrocodone	5 mg PO q4h	
Strong opioids		For opioid-naïve patients with moderate to severe dyspnea
Morphine	5–10 mg PO q4h	For patients already taking opioids for pain or other symptoms
	30–50% of baseline opioid dose q4h	
Oxycodone	5–10 mg PO q4h	
Hydromorphone	1–2 mg PO q4h	
Anxiolytics		Give a dose every hour until the patient is relaxed, then provide a dose for maintenance
Lorazepam	0.5–2.0 mg PO/SL/IV qh then q4–6h	
Clonazepam	0.25–2.0 mg PO q12h	
Midazolam	0.5 mg IV q15min	

Fatigue

FREQUENCY

Fatigue is one of the most commonly reported symptoms of not only cancer treatment, but also of the palliative care of multiple sclerosis, COPD, heart failure, and HIV. More than 90% of terminally ill patients experience fatigue and/or weakness. Fatigue is frequently cited among the most distressing symptoms.

ETIOLOGY

The multiple causes of fatigue in the terminally ill can be categorized as resulting from the underlying disease; from disease-induced factors such as tumor necrosis factor and other cytokines; and from secondary factors such as dehydration, anemia, infection, hypothyroidism, and drug side effects. In addition to low caloric intake, loss of muscle mass and changes in muscle enzymes may play an important role in fatigue during terminal illness. The importance of changes in the CNS, especially the reticular activating system, have been hypothesized based on reports of fatigue in patients receiving cranial radiation, experiencing depression, or having chronic pain in the absence of cachexia or other physiologic changes. Finally, depression and other causes of psychological distress can contribute to fatigue.

ASSESSMENT

Like pain and dyspnea, fatigue is subjective, as it represents a patient's sense of tiredness and decreased capacity for physical work. Objective changes, even in body mass, may be absent. Consequently, assessment must rely on patient self-reporting. Scales used to measure fatigue, such as the Edmonton Functional Assessment Tool, the Fatigue Self-Report Scales, and the Rhoten Fatigue Scale, are usually appropriate for research, but not clinical purposes. In clinical practice, a simple performance assessment such as the Karnofsky Performance Status or the Eastern Cooperative Oncology Group's question "How much of the day does the patient spend in bed?" may be the best measure. In this 0–4 performance status assessment, 0 = normal activity; 1 = symptomatic without being bedridden; 2 = requiring some, but <50%, bed time; 3 = bedbound more than half the day; and 4 = bedbound all the time. Such a scale allows for assessment over time and correlates with overall disease severity and prognosis. A 2008 review by the European Association of Palliative Care also described several longer assessment tools that contained 9–20 items, including the Piper Fatigue Inventory, the Multidimensional Fatigue Inventory, and the Brief Fatigue Inventory (BFI).

INTERVENTIONS

Reversible causes of fatigue, such as anemia and infection, should be treated. However, at the end of life, it must be realistically acknowledged that fatigue will not be "cured." The goal is to ameliorate fatigue and help patients and families adjust expectations. Behavioral interventions should be utilized to avoid blaming the patient for inactivity and to educate both the family and the patient that the underlying disease causes physiologic changes that produce low energy levels. Understanding that the problem is physiologic and not psychological can help alter expectations regarding the patient's level of physical activity. Practically, this may mean reducing routine activities such as housework, cooking, and social events outside the house, and making it acceptable to receive guests while lying on a couch. At the same time, the implementation of exercise regimens and physical therapy can raise endorphins, reduce muscle wasting, and decrease the risk of depression. In addition, ensuring good hydration without worsening edema may help reduce fatigue. Discontinuing medications that worsen fatigue may help, including cardiac medications, benzodiazepines, certain antidepressants, or opioids if the pain is well-controlled. As end-of-life care proceeds into its final stages, fatigue may protect patients from further suffering, and continued treatment could be detrimental.

Only a few pharmacologic interventions target fatigue and weakness. Randomized controlled trials suggest glucocorticoids can increase energy and enhance mood. Dexamethasone (8 mg per d) is preferred for its once-a-day dosing and minimal mineralocorticoid activity. Benefit, if any, is usually seen within the first month. For fatigue related to anorexia, megestrol (480–800 mg) can be helpful. Psychostimulants such as dextroamphetamine (5–10 mg PO) and methylphenidate (2.5–5 mg PO) may enhance energy levels, although controlled trials have not shown these drugs to be effective for fatigue induced by mild to moderate cancer. Doses should be given in the morning and at noon to minimize the risk of counterproductive insomnia. Modafinil and armodafinil, developed for narcolepsy, have shown promise in the treatment of fatigue and have the advantage of once-daily dosing. Their precise role in fatigue at the end of life has not been documented, but may be worth trying if other interventions are not beneficial. Anecdotal evidence suggests that L-carnitine may improve fatigue, depression, and sleep disruption.

PALLIATIVE SEDATION

When patients experience severe symptoms, such as pain or dyspnea, that cannot be relieved by conventional interventions or experience acute catastrophic symptoms, such as uncontrolled seizures, then palliative sedation should be considered as an intervention of last resort. Palliative sedation is used in distressing situations that cannot be addressed in other ways. It can be abused if done to hasten death (which it usually does not), when at the request of the family, rather than the patient's wishes, or when there are other interventions that could still be tried. The use of palliative sedation in cases of extreme existential or spiritual distress remains controversial. Typically, palliative sedation should be introduced only after the patient and family have been assured that all other interventions have been tried, and after the patient and their loved ones have been able to "say goodbye."

Palliative sedation can be achieved by significantly increasing opioid doses until patients become unconscious, then putting them on a continuous infusion. Another commonly used medication for palliative sedation is midazolam at 1–5 mg IV every 5–15 min to calm the patient, followed by a continuous IV or subcutaneous infusion of 1 mg per h. In hospital settings, a continuous propofol infusion of 5 μg/kg per min can be used. There are also other, less commonly used medications for palliative sedation that include levomepromazine, chlorpromazine, and phenobarbital.

PSYCHOLOGICAL SYMPTOMS AND THEIR MANAGEMENT

Depression

FREQUENCY AND IMPACT

Depression at the end of life presents an apparently paradoxical situation. Many people believe that depression is normal among seriously ill patients, because they are dying. People frequently say, "Wouldn't you be depressed?" Although sadness, anxiety, anger, and irritability are normal responses to a serious condition, they are typically of modest intensity and transient. Persistent sadness and anxiety and the physically disabling symptoms that they can lead to are abnormal and suggestive of major depression. The precise number of terminally ill patients who are depressed is uncertain, primarily due to a lack of consistent

diagnostic criteria and screening. Careful follow-up of patients suggests that while as many as 75% of terminally ill patients experience depressive symptoms, ~25% of terminally ill patients have major depression. Depression at the end of life is concerning, because it can decrease the quality of life, interfere with closure in relationships and other separation work, obstruct adherence to medical interventions, and amplify the suffering associated with pain and other symptoms.

ETIOLOGY

Previous history of depression, family history of depression or bipolar disorder, and prior suicide attempts are associated with increased risk for depression among terminally ill patients. Other symptoms, such as pain and fatigue, are associated with higher rates of depression; uncontrolled pain can exacerbate depression, and depression can cause patients to be more distressed by pain. Many medications used in the terminal stages, including glucocorticoids, and some anticancer agents, such as tamoxifen, interleukin 2, interferon α, and vincristine, also are associated with depression. Some terminal conditions, such as pancreatic cancer, certain strokes, and heart failure, have been reported to be associated with higher rates of depression, although this is controversial. Finally, depression may be attributable to grief over the loss of a role or function, social isolation, or loneliness.

ASSESSMENT

Unfortunately, most studies suggest that depressed patients at the end of life are neither diagnosed, nor even properly treated if diagnosed. Diagnosing depression among seriously ill patients is complicated, as many of the vegetative symptoms in the DSM-V (*Diagnostic and Statistical Manual of Mental Disorders*) criteria for clinical depression—insomnia, anorexia and weight loss, fatigue, decreased libido, and difficulty concentrating—are associated with the process of dying itself. The assessment of depression in seriously ill patients therefore should focus on the dysphoric mood, helplessness, hopelessness, and lack of interest, enjoyment, and concentration in normal activities. It is now recommended that patients near the end of life should be screened either with the Patient Health Questionnarie-9 (PHQ-9) or the PHQ-2 which asks “Over the past two weeks, how often have you been bothered by any of the following problems? (1) Little interest or pleasure in doing things and (2) feeling down, depressed or hopeless.” The answer categories are: Not at all, Several days, More than half the days, Nearly every day. There are other possible diagnostic tools such as the short form of the Beck Depression Index or a visual analog scale.

Certain conditions may be confused with depression. Endocrinopathies, such as hypothyroidism and Cushing’s syndrome, electrolyte abnormalities, such as hypercalcemia, and akathisia, especially from dopamine-blocking antiemetics such as metoclopramide and prochlorperazine, can mimic depression and should be excluded.

INTERVENTIONS

Under-treatment of depressed, terminally ill patients is common. Physicians must treat any physical symptom, such as pain, that may be causing or exacerbating depression. Fostering adaptation to the many losses that the patient is experiencing can also be helpful. Unfortunately, there are few randomized trials to guide such interventions. Thus, treatment typically follows the treatment used for non-terminally ill depressed patients.

While there are no randomized controlled trials, nonpharmacologic interventions, including group or individual psychological counseling, and behavioral therapies such as relaxation and imagery can be helpful, especially in combination with drug therapy.

Pharmacologic interventions remain at the core of therapy. The same medications are used to treat depression in terminally ill as in non-terminally ill patients. Psychostimulants may be preferred for patients with a poor prognosis, or for those with fatigue or opioid-induced somnolence. Psychostimulants are comparatively fast-acting, working within a few days instead of the weeks required for selective serotonin reuptake inhibitors (SSRIs). Dextroamphetamine or methylphenidate should be started at 2.5–5.0 mg in the morning and at noon, the same starting doses used for treating fatigue. The doses can eventually be escalated up to 15 mg bid. Modafinil is started at 100 mg qd and can be increased to 200 mg if there is no effect at the lower dose. Pemoline is a nonamphetamine psychostimulant with minimal abuse potential. It is also effective as an antidepressant beginning at 18.75 mg in the morning and at noon. Because it can be absorbed through the buccal mucosa, it is preferred for patients with intestinal obstruction or dysphagia. If it is used for prolonged periods, liver function must be monitored. The psychostimulants can also be combined with more traditional antidepressants while waiting for the antidepressants to become effective, then tapered down after a few weeks if necessary. Psychostimulants have side effects, particularly initial anxiety, insomnia, and very rarely paranoia, which may necessitate lowering the dose or discontinuing treatment.

Mirtazapine, an antagonist at the postsynaptic serotonin receptors, is a promising psychostimulant. It should be started at 7.5 mg before bed and titrated up no more than once every 1–2 weeks to a maximal dose of 45 mg per d. It has sedating, antiemetic, and anxiolytic properties, with few drug interactions. Its side effect of weight gain may be beneficial for seriously ill patients; it is available in orally disintegrating tablets.

For patients with a prognosis of several months or longer, SSRIs, including fluoxetine, sertraline, paroxetine, escitalopram, and citalopram, and serotonin-noradrenaline reuptake inhibitors, such as venlafaxine and duloxetine, are the preferred treatments, due to their efficacy and comparatively few side effects. Because low doses of these medications may be effective for seriously ill patients, one should use half the usual starting dose as for healthy adults. The starting dose for fluoxetine is 10 mg once a day. In most cases, once-a-day dosing is possible. The choice of which SSRI to use should be driven by (1) the patient's past success or failure with the specific medication and (2) the most favorable side-effect profile for that specific agent. For instance, for a patient in whom fatigue is a major symptom, a more activating SSRI (fluoxetine) would be appropriate. For a patient in whom anxiety and sleeplessness are major symptoms, a more sedating SSRI (paroxetine) would be appropriate. Importantly, it can take up to 4 weeks for these drugs to have an effect.

Atypical antidepressants are recommended only in select circumstances, usually with the assistance of a specialty consultation. Trazodone can be an effective antidepressant, but is sedating and can cause orthostatic hypotension and, occasionally, priapism. Therefore, it should be used before bed and only when a sedating effect is desired, and is often used for patients with insomnia, at a dose starting at 25 mg. Bupropion can also be used. In addition to its antidepressant effects, bupropion is energizing, making it useful for depressed patients who experience fatigue. However, it can cause seizures, preventing its use for patients with a risk of CNS neoplasms or terminal delirium. Finally, alprazolam, a benzodiazepine, starting at

0.25–1.0 mg tid, can be effective in seriously ill patients who have a combination of anxiety and depression. Although it is potent and works quickly, it has many drug interactions and may cause delirium, especially among very ill patients, because of its strong binding to the benzodiazepine–γ-aminobutyric acid (GABA) receptor complex.

Unless used as adjuvants for the treatment of pain, tricyclic antidepressants are not recommended. While they can be effective, their therapeutic window and serious side effects typically limit their utility. Similarly, monoamine oxidase (MAO) inhibitors are not recommended because of their side effects and dangerous drug interactions.

Delirium

(See Chap. 24)

FREQUENCY

In the weeks or months before death, delirium is uncommon, although it may be significantly underdiagnosed. However, delirium becomes relatively common in the days and hours immediately before death. Up to 85% of patients dying from cancer may experience terminal delirium.

ETIOLOGY

Delirium is a global cerebral dysfunction characterized by alterations in cognition and consciousness. It is frequently preceded by anxiety, changes in sleep patterns (especially reversal of day and night), and decreased attention. In contrast to dementia, delirium has an acute onset, is characterized by fluctuating consciousness and inattention, and is reversible, although reversibility may be more theoretical than real for patients near death. Delirium may occur in a patient with dementia; indeed, patients with dementia are more vulnerable to delirium.

Causes of delirium include metabolic encephalopathy arising from liver or renal failure, hypoxemia, or infection; electrolyte imbalances such as hypercalcemia; paraneoplastic syndromes; dehydration; and primary brain tumors, brain metastases, or leptomeningeal spread of tumor. Among dying patients, delirium is commonly caused by side effects of treatments, including radiation for brain metastases and medications, such as opioids, glucocorticoids, anticholinergic drugs, antihistamines, antiemetics, benzodiazepines, and chemotherapeutic agents. The etiology may be multifactorial; e.g., dehydration may exacerbate opioid-induced delirium.

ASSESSMENT

Delirium should be recognized in any terminally ill patient exhibiting new onset of disorientation, impaired cognition, somnolence, fluctuating levels of consciousness, or delusions with or without agitation. Delirium must be distinguished from acute anxiety, depression, and dementia. The central distinguishing feature is altered consciousness, which usually is not noted in anxiety, depression, or dementia. Although “hyperactive” delirium, characterized by overt confusion and agitation, is probably more common, patients should also be assessed for “hypoactive” delirium, which is characterized by sleep-wake reversal and decreased alertness.

In some cases, use of formal assessment tools such as the Mini-Mental Status Examination (which does not distinguish delirium from dementia) and the Delirium Rating Scale (which does distinguish delirium from dementia) may be helpful in distinguishing delirium from other processes. The patient's list of medications must be evaluated carefully. Nonetheless, a reversible etiologic factor for delirium is found in fewer than half of all terminally ill patients. Given that most terminally ill patients experiencing delirium are very close to death and often at home, extensive diagnostic evaluations such as lumbar punctures and neuroradiologic examinations are inappropriate.

INTERVENTIONS

One of the most important objectives of terminal care is to provide terminally ill patients the lucidity to say goodbye to the people they love. Delirium, especially when in combination with agitation during the final days, is distressing to family and caregivers. A strong determinant of bereavement difficulties is witnessing a difficult death. Thus, terminal delirium should be treated aggressively.

At the first sign of delirium, such as day-night reversal with slight changes in mentation, the physician should let the family members know that it is time to be sure that everything they want to say has been said. The family should be informed that delirium is common just before death.

If medications are suspected of being a cause of the delirium, unnecessary agents should be discontinued. Other potentially reversible causes, such as constipation, urinary retention, and metabolic abnormalities, should be treated. Supportive measures aimed at providing a familiar environment should be instituted, including restricting visits only to individuals with whom the patient is familiar and eliminating new experiences; orienting the patient, if possible, by providing a clock and calendar; and gently correcting the patient's hallucinations or cognitive mistakes.

Pharmacologic management focuses on the use of neuroleptics and, in extreme cases, anesthetics **(Table 9-7)**. Haloperidol remains the first-line therapy. Usually, patients can be controlled with a low dose (1–3 mg/d), given every 6 h, although some may require as much as 20 mg/d. Haloperidol can be administered PO, SC, or IV. IM injections should not be used, except when this is the only way to address a patient's delirium. Olanzapine, an atypical neuroleptic, has shown significant effectiveness in completely resolving delirium in cancer patients. It also has other beneficial effects for terminally ill patients, including anti-nausea, antianxiety, and weight gain. Olanzapine is useful for patients with longer anticipated life expectancies, because it is less likely to cause dysphoria and has a lower risk of dystonic reactions. Additionally, because olanzapine is metabolized through multiple pathways, it can be used in patients with hepatic and renal dysfunction. Olanzapine has the disadvantage that it is only available orally and takes a week to reach steady state. The usual dose is 2.5–5 mg PO bid. Chlorpromazine (10–25 mg every 4–6 h) can be useful if sedation is desired and can be administered IV or PR in addition to PO. Dystonic reactions resulting from dopamine blockade are a side effect of neuroleptics, although they are reported to be rare when these drugs are used to treat terminal delirium. If patients develop dystonic reactions, benztropine should be administered. Neuroleptics may be combined with lorazepam to reduce agitation when the delirium is the result of alcohol or sedative withdrawal.

TABLE 9-7

Medications for the Management of Delirium

Interventions	Dose
Neuroleptics	
Haloperidol	0.5–5 mg q2–12h, PO/IV/SC/IM
Thioridazine	10–75 mg q4–8h, PO
Chlorpromazine	12.5–50 mg q4–12h, PO/IV/IM
Atypical neuroleptics	
Olanzapine	2.5–5 mg qd or bid, PO
Risperidone	1–3 mg q12h, PO
Anxiolytics	
Lorazepam	0.5–2 mg q1–4h, PO/IV/IM
Midazolam	1–5 mg/h continuous infusion, IV/SC
Anesthetics	
Propofol	0.3–2.0 mg/h continuous infusion, IV

If no response to first-line therapy is observed, a specialty consultation should be obtained with a goal to change to a different medication. If the patient fails to improve after a second neuroleptic, sedation with either an anesthetic such as propofol or continuous-infusion midazolam may be necessary. By some estimates, as many as 25% of patients at the very end of life who experience delirium, especially restless delirium with myoclonus or convulsions, may require sedation.

Physical restraints should be used with great reluctance and only when the patient's violence is threatening to himself or others. If restraints are used, their appropriateness should be frequently reevaluated.

Insomnia

FREQUENCY

Sleep disorders, defined as difficulty initiating sleep or maintaining sleep, sleep difficulty at least 3 nights a week, or sleep difficulty that causes impairment of daytime functioning, occurs in 19–63% of patients with advanced cancer. Some 30–74% of patients with other end-stage conditions, including AIDS, heart disease, COPD, and renal disease, experience insomnia.

ETIOLOGY

Patients with cancer may experience changes in sleep efficiency, such as an increase in stage I sleep. Insomnia may also coexist with both physical illnesses, like thyroid disease, and psychological illnesses, like depression and anxiety. Medications, including antidepressants, psychostimulants, steroids, and β agonists, are significant contributors to sleep disorders, as are caffeine and alcohol. Multiple over-the-counter medications contain caffeine and antihistamines, which can contribute to sleep disorders.

ASSESSMENT

Assessments should include specific questions concerning sleep onset, sleep maintenance, and early-morning wakening, as these will provide clues to both the causative agents and management of insomnia. Patients should be asked about previous sleep problems, screened for depression and anxiety, and asked about symptoms of thyroid disease. Caffeine and alcohol are prominent causes of sleep problems, and a careful history of the use of these substances should be obtained. Both excessive use and withdrawal from alcohol can be causes of sleep problems.

INTERVENTIONS

The mainstays of any intervention include improvement of sleep hygiene (encouragement of regular time for sleep, decreased nighttime distractions, elimination of caffeine and other stimulants and alcohol), interventions to treat anxiety and depression, and treatment for the insomnia itself. For patients with depression who have insomnia and anxiety, a sedating antidepressant such as mirtazapine can be helpful. In the elderly, trazodone, beginning at 25 mg at nighttime, is an effective sleep aid at doses lower than those which cause its antidepressant effect. Zolpidem may have a decreased incidence of delirium in patients compared with traditional benzodiazepines, but this has not been clearly established. When benzodiazepines are prescribed, short-acting ones (such as lorazepam) are favored over longer-acting ones (such as diazepam). Patients who receive these medications should be observed for signs of increased confusion and delirium.

SOCIAL NEEDS AND THEIR MANAGEMENT

Financial Burdens

FREQUENCY

Dying can impose substantial economic strains on patients and families, potentially causing distress. In the United States, which has one of the least comprehensive health insurance systems among developed countries, a quarter of families coping with end-stage cancer report that care was a major financial burden and a third used up most of their savings. Among Medicare beneficiaries, average out-of-pocket costs were >$8,000. Between 10 and 30% of families are forced to sell assets, use savings, or take out a mortgage to pay for the patient's health care costs.

The patient is likely to reduce hours worked, and eventually stop working altogether. In 20% of cases, a family member of the terminally ill patient also must stop working to provide care. The major underlying causes of economic burden are related to poor physical functioning and care needs, such as the need for housekeeping, nursing, and personal care. More debilitated patients and poor patients experience greater economic burdens.

INTERVENTION

The economic burden of end-of-life care should not be ignored as a private matter. It has been associated with a number of adverse health outcomes, including preferring comfort care over life-prolonging care, as well as consideration of euthanasia or physician-assisted suicide (PAS). Economic burdens increase the psychological distress of the families and caregivers of terminally ill patients, and poverty is associated with many adverse health outcomes. Importantly, recent studies have found that “patients with advanced cancer who reported having end-of-life conversations with physicians had significantly lower health care costs in their final week of life. Higher costs were associated with worse quality of death.” Assistance from a social worker, early on if possible, to ensure access to all available benefits may be helpful. Many patients, families, and health care providers are unaware of options for long-term care insurance, respite care, the Family Medical Leave Act (FMLA), and other sources of assistance. Some of these options (such as respite care) may be part of a formal hospice program, but others (such as the FMLA) do not require enrollment in a hospice program.

Relationships

FREQUENCY

Settling personal issues and closing the narrative of lived relationships are universal needs. When asked if sudden death or death after an illness is preferable, respondents often initially select the former, but soon change to the latter as they reflect on the importance of saying goodbye. Bereaved family members who have not had the chance to say goodbye often have a more difficult grief process.

INTERVENTIONS

Care of seriously ill patients requires efforts to facilitate the types of encounters and time spent with family and friends that are necessary to meet those needs. Family and close friends may need to be accommodated in hospitals and other facilities with unrestricted visiting hours, which may include sleeping near the patient, even in otherwise regimented institutional settings. Physicians and other health care providers may be able to facilitate and resolve strained interactions between the patient and other family members. Assistance for patients and family members who are unsure about how to create or help preserve memories, whether by providing materials such as a scrapbook or memory box, or by offering them suggestions and informational resources, can be deeply appreciated. Taking photographs and creating videos can be especially helpful to terminally ill patients who have younger children or grandchildren.

Family Caregivers

FREQUENCY

Caring for seriously ill patients places a heavy burden on families. Families are frequently required to provide transportation and homemaking, as well as other services. Typically, paid professionals, such as home health

nurses and hospice workers, supplement family care; only about a quarter of all caregiving consists of exclusively paid professional assistance. Over the last 40 years, there has been a significant decline in the United States of deaths occurring in hospitals, with a simultaneous increase in deaths in other facilities and at home. Over a third of deaths occur in patients' home. This increase in out-of-hospital deaths increases reliance on families for end-of-life care. Increasingly, family members are being called upon to provide physical care (such as moving and bathing patients) and medical care (such as assessing symptoms and giving medications) in addition to emotional care and support.

Three-quarters of family caregivers of terminally ill patients are women—wives, daughters, sisters, and even daughters-in-law. Since many are widowed, women tend to be able to rely less on family for caregiving assistance and may need more paid assistance. About 20% of terminally ill patients report substantial unmet needs for nursing and personal care. The impact of caregiving on family caregivers is substantial: both bereaved and current caregivers have a higher mortality rate than that of non-caregiving controls.

INTERVENTIONS

It is imperative to inquire about unmet needs and to try to ensure that those needs are met either through the family or by paid professional services when possible. Community assistance through houses of worship or other community groups often can be mobilized by telephone calls from the medical team to someone the patient or family identifies. Sources of support specifically for family caregivers should be identified through local sources or nationally through groups such as the National Family Caregivers Association (*www.nfcacares.org*), the American Cancer Society (*www.cancer.org*), and the Alzheimer's Association (*www.alz.org*).

EXISTENTIAL NEEDS AND THEIR MANAGEMENT

Frequency

Religion and spirituality are often important to dying patients. Nearly 70% of patients report becoming more religious or spiritual when they became terminally ill, and many find comfort in religious or spiritual practices such as prayer. However, ~20% of terminally ill patients become less religious, frequently feeling cheated or betrayed by becoming terminally ill. For other patients, the need is for existential meaning and purpose that is distinct from, and may even be antithetical to, religion or spirituality. When asked, patients and family caregivers frequently report wanting their professional caregivers to be more attentive to religion and spirituality.

ASSESSMENT

Health care providers are often hesitant about involving themselves in the religious, spiritual, and existential experiences of their patients because it may seem private or not relevant to the current illness. But physicians and other members of the care team should be able at least to detect spiritual and existential needs. Screening questions have been developed for a physician's spiritual history taking. Spiritual distress can amplify other types of suffering and even masquerade as intractable physical pain, anxiety, or depression. The screening questions in the comprehensive assessment are usually sufficient. Deeper evaluation and intervention are rarely appropriate for the physician unless no other member of a care team

is available or suitable. Pastoral care providers may be helpful, whether from the medical institution or from the patient's own community.

INTERVENTIONS

Precisely how religious practices, spirituality, and existential explorations can be facilitated and improve end-of-life care is not well established. What is clear is that for physicians, one main intervention is to inquire about the role and importance of spirituality and religion in a patient's life. This will help a patient feel heard and help physicians identify specific needs. In one study, only 36% of respondents indicated that a clergy member would be comforting. Nevertheless, the increase in religious and spiritual interest among a substantial fraction of dying patients suggests inquiring of individual patients how this need can be addressed. Some evidence supports specific methods of addressing existential needs in patients, ranging from establishing a supportive group environment for terminal patients to individual treatments emphasizing a patient's dignity and sources of meaning.

MANAGING THE LAST STAGES

PALLIATIVE CARE SERVICES: HOW AND WHERE

Determining the best approach to providing palliative care to patients will depend on patient preferences, the availability of caregivers and specialized services in close proximity, institutional resources, and reimbursement. Hospice is a leading, but not the only, model of palliative care services. In the United States, slightly more than a third—35.7%—of hospice care is provided in private residential homes. In 2014, 14.5% of hospice care was provided in nursing homes. In the United States, Medicare pays for hospice services under Part A, the hospital insurance part of reimbursement. Two physicians must certify that the patient has a prognosis of ≤6 months if the disease runs its usual course. Prognoses are probabilistic by their nature; patients are not required to die within 6 months but rather to have a condition from which half the individuals with it would not be alive within 6 months. Patients sign a hospice enrollment form that states their intent to forgo curative services related to their terminal illness, but can still receive medical services for other comorbid conditions. Patients also can withdraw enrollment and reenroll later; the hospice Medicare benefit can be revoked later to secure traditional Medicare benefits. Payments to the hospice are per diem (or capitated), not fee-for-service. Payments are intended to cover physician services for the medical direction of the care team; regular home care visits by registered nurses and licensed practical nurses; home health aide and homemaker services; chaplain services; social work services; bereavement counseling; and medical equipment, supplies, and medications. No specific therapy is excluded, and the goal is for each therapy to be considered for its symptomatic (as opposed to disease-modifying) effect. Additional clinical care, including services of the primary physician, is covered by Medicare Part B even while the hospice Medicare benefit is in place.

The Affordable Care Act directs the Secretary of Health and Human Services to gather data on Medicare hospice reimbursement with the goal of reforming payment rates to account for resource use over an entire episode of care. The legislation also requires additional evaluations and reviews of eligibility for hospice care

by hospice physicians or nurses. Finally, the Center for Medicare and Medicaid Innovation (CMMI) is testing concurrent hospice and palliative care services with curative treatment with ~120 providers.

By 2014, the mean length of enrollment in a hospice was 71 days, with the median being 17 days. Such short stays create barriers to establishing high-quality palliative services in patients' homes and also place financial strains on hospice providers since the initial assessments are resource intensive. Physicians should initiate early referrals to the hospice to allow more time for patients to receive palliative care.

In the United States, hospice care has been the main method for securing palliative services for terminally ill patients. However, as leading physicians have increasingly emphasized the need to introduce palliative care much earlier in patients' illness, efforts are being made to develop palliative care services that can be provided before the last 6 months of life and across a variety of settings. For instance, some companies and home health agencies are offering non-hospice palliative care services in patients' homes in an effort to increase quality of life and forestall hospitalizations. Similarly, palliative care services are increasingly available via consultation, rather than present only in hospital, day care, outpatient, and nursing home settings. Palliative care consultations for non-hospice patients can be billed as for other consultations under Medicare Part B. It is argued that using palliative care earlier in patients' illness allows patients and family members to become more acculturated to avoiding life-sustaining treatments, facilitating a smoother transition to hospice care closer to death.

WITHDRAWING AND WITHHOLDING LIFE-SUSTAINING TREATMENT

LEGAL ASPECTS

For centuries, it has been deemed ethical to withhold or withdraw life-sustaining interventions. The current legal consensus in the United States and most developed countries is that patients have a moral as well as constitutional or common law right to refuse medical interventions. American courts also have held that incompetent patients have a right to refuse medical interventions. For patients who are incompetent and terminally ill and who have not completed an advance care directive, next of kin can exercise that right, although this may be restricted in some states, depending on how clear and convincing the evidence is of the patient's preferences. Courts have limited families' ability to terminate life-sustaining treatments in patients who are conscious and incompetent, but not terminally ill. In theory, patients' right to refuse medical therapy can be limited by four countervailing interests: (1) preservation of life, (2) prevention of suicide, (3) protection of third parties such as children, and (4) preservation of the integrity of the medical profession. In practice, these interests almost never override the right of competent patients and incompetent patients who have left explicit wishes or advance care directives.

For incompetent patients who either appointed a proxy without specific indications of their wishes or never completed an advance care directive, three criteria have been suggested to guide the decision to terminate medical interventions. First, some commentators suggest that ordinary care should be administered but extraordinary care could be terminated. Because the ordinary/extraordinary distinction is too vague, courts and commentators widely agree that it should not be used to justify decisions about stopping treatment. Second, many courts have advocated the use of the substituted-judgment criterion, which holds that the

proxy decision-makers should try to imagine what the incompetent patient would do if he or she were competent. However, multiple studies indicate that many proxies, even close family members, cannot accurately predict what the patient would have wanted. Therefore, substituted judgment becomes more of a guessing game than a way of fulfilling the patient's wishes. Finally, the best-interests criterion holds that proxies should evaluate treatments by balancing their benefits and risks and select those treatments in which the benefits maximally outweigh the burdens of treatment. Clinicians have a clear and crucial role in this by carefully and dispassionately explaining the known benefits and burdens of specific treatments. Yet even when that information is as clear as possible, different individuals can have very different views of what is in the patient's best interests, and families may have disagreements or even overt conflicts. This criterion has been criticized because there is no single way to determine the balance between benefits and burdens; it depends on a patient's personal values. For instance, for some people being alive even if mentally incapacitated is a benefit, whereas for others it may be the worst possible existence. As a matter of practice, physicians rely on family members to make decisions that they feel are best and object only if those decisions seem to demand treatments that the physicians consider not beneficial.

PRACTICES

Withholding and withdrawing acutely life-sustaining medical interventions from terminally ill patients are now standard practice. More than 90% of American patients die without cardiopulmonary resuscitation (CPR), and just as many forgo other potentially life-sustaining interventions. For instance, in ICUs in the period 1987–1988, CPR was performed 49% of the time, but it was performed only 10% of the time in 1992–1993 and on just 1.8% of admissions from 2001 to 2008. On average, 3.8 interventions, such as vasopressors and transfusions, were stopped for each dying ICU patient. However, up to 19% of decedents in hospitals received interventions such as extubation, ventilation, and surgery in the 48 h preceding death. There is wide variation in practices among hospitals and ICUs, suggesting an important element of physician preferences rather than consistent adherence to professional society recommendations.

Mechanical ventilation may be the most challenging intervention to withdraw. The two approaches are *terminal extubation*, which is the removal of the endotracheal tube, and *terminal weaning*, which is the gradual reduction of the FIo_2 or ventilator rate. One-third of ICU physicians prefer to use the terminal weaning technique, and 13% extubate; the majority of physicians utilize both techniques. The American Thoracic Society's 2008 clinical policy guidelines note that there is no single correct process of ventilator withdrawal and that physicians use and should be proficient in both methods but that the chosen approach should carefully balance benefits and burdens as well as patient and caregiver preferences. Some recommend terminal weaning because patients do not develop upper airway obstruction and the distress caused by secretions or stridor; however, terminal weaning can prolong the dying process and not allow a patient's family to be with the patient unencumbered by an endotracheal tube. To ensure comfort for conscious or semiconscious patients before withdrawal of the ventilator, neuromuscular blocking agents should be terminated and sedatives and analgesics administered. Removing the neuromuscular blocking agents permits patients to show discomfort, facilitating the titration of sedatives and analgesics; it also permits interactions between patients and their families. A common practice is to inject a bolus of midazolam (2–4 mg) or lorazepam (2–4 mg) before withdrawal, followed by a bolus of 5–10 mg of morphine

and continuous infusion of morphine (50% of the bolus dose per hour) during weaning. In patients who have significant upper airway secretions, IV scopolamine at a rate of 100 µg/h can be administered. Additional boluses of morphine or increases in the infusion rate should be administered for respiratory distress or signs of pain. Higher doses will be needed for patients already receiving sedatives and opioids.

The median time to death after stopping of the ventilator is ~1 h. However, up to 10% of patients unexpectedly survive for 1 day or more after mechanical ventilation is stopped. Women and older patients tend to survive longer after extubation. Families need to be reassured about both the continuations of treatments for common symptoms, such as dyspnea and agitation, after withdrawal of ventilatory support and the uncertainty of length of survival after withdrawal of ventilatory support.

FUTILE CARE

Beginning in the late 1980s, some commentators argued that physicians could terminate futile treatments demanded by the families of terminally ill patients. Although no objective definition or standard of futility exists, several categories have been proposed. Physiologic futility means that an intervention will have no physiologic effect. Some have defined qualitative futility as applying to procedures that "fail to end a patient's total dependence on intensive medical care." Quantitative futility occurs "when physicians conclude (through personal experience, experiences shared with colleagues, or consideration of reported empiric data) that in the last 100 cases, a medical treatment has been useless." The term conceals subjective value judgments about when a treatment is "not beneficial." Deciding whether a treatment that obtains an additional 6 weeks of life or a 1% survival advantage confers benefit depends on patients' preferences and goals. Furthermore, physicians' predictions of when treatments are futile deviate markedly from the quantitative definition. When residents thought CPR was quantitatively futile, more than one in five patients had a >10% chance of survival to hospital discharge. Most studies that purport to guide determinations of futility are based on insufficient data and therefore cannot provide statistical confidence for clinical decision-making. Quantitative futility rarely applies in ICU settings.

Many commentators reject using futility as a criterion for withdrawing care, preferring instead to consider futility situations as ones that represent conflict that calls for careful negotiation between families and health care providers. The AMA and other professional societies have developed process-based approaches to resolving cases clinicians feel are futile. These process-based measures mainly suggest involving consultants and/or ethics committees when there are seemingly irresolvable differences. Some hospitals have enacted "unilateral DNR" policies to allow clinicians to provide a do-not-resuscitate order in cases in which consensus cannot be reached with families and medical opinion is that resuscitation would be futile if attempted. This type of a policy is not a replacement for careful and patient communication and negotiation but recognizes that agreement cannot always be reached.

In 1999 Texas enacted the so-called Futile Care Act. Other states, such as Virginia, Maryland, and California, have also enacted such laws that provide physicians a "safe harbor" from liability if they refuse a patient or family's request for life-sustaining interventions. For instance, in Texas when a disagreement about terminating interventions between the medical team and the family has not been resolved by an ethics

consultation, the physician is tasked with trying to facilitate transfer of the patient to an institution willing to provide treatment. If this fails after 10 days, the hospital and physician may unilaterally withdraw treatments determined to be futile. The family may appeal to a state court. Early data suggest that the law increases futility consultations for the ethics committee and that although most families concur with withdrawal, about 10–15% of families refuse to withdraw treatment. As of 2007, there had been 974 ethics committee consultations on medical futility cases and 65 in which committees ruled against families and gave notice that treatment would be terminated. In 2007 a survey of Texas hospitals showed that 30% of hospitals had used the futility law in 213 adult cases and 42 pediatric cases. Treatment was withdrawn for 27 of those patients, and the remainder transferred to other facilities or died while awaiting transfer.

EUTHANASIA AND PHYSICIAN-ASSISTED SUICIDE

Euthanasia and PAS are defined in Table 9-8. Terminating life-sustaining care and providing opioid medications to manage symptoms such as pain or dyspnea have long been considered ethical by the medical profession and legal by courts and should not be conflated with euthanasia or PAS.

TABLE 9-8

Definitions of Physician-Assisted Suicide and Euthanasia

Term	Definition	Legal Status
Voluntary active euthanasia	Intentionally administering medications or other interventions to cause the patient's death with the patient's informed consent	Netherlands, Belgium, Luxembourg, Canada, Colombia
Involuntary active euthanasia	Intentionally administering medications or other interventions to cause the patient's death when the patient was competent to consent but did not—e.g., the patient may not have been asked	Nowhere
Passive euthanasia	Withholding or withdrawing life-sustaining medical treatments from a patient to let him or her die (terminating life-sustaining treatments)	Everywhere
Physician-assisted suicide	A physician provides medications or other interventions to a patient with the understanding that the patient can use them to commit suicide	Netherlands, Belgium, Luxembourg, Canada, Colombia, Switzerland, Oregon, Washington, Montana, Vermont, California

Euthanasia and PAS are legal in the Netherlands, Belgium, Luxembourg, Colombia, and Canada. Euthanasia was legalized in the Northern Territory of Australia in 1996, but that legislation was repealed 9 months later in 1997. Under certain conditions, a layperson in Switzerland can legally elect assisted suicide. In the United States, PAS is legal in five states: Oregon, Washington State, Montana, Vermont, and California. No state in the United States has legalized euthanasia. In the United States, multiple criteria must be met for PAS: the patient must have a terminal condition of <6 months, and must be determined eligible through a process that includes a 15-day waiting period. In 2009, the state supreme court of Montana ruled that state law permits PAS for terminally ill patients. Many other countries, such as Australia, are actively debating the legalization of euthanasia and/or PAS.

PRACTICES

Fewer than 10–20% of terminally ill patients actually consider euthanasia and/or PAS for themselves. Use of euthanasia and PAS is relatively rare. In all countries, even the Netherlands and Belgium where these practices have been tolerated and legal for many years, fewer than 5% of death occur by euthanasia or PAS. As of the most recent data, the share of deaths attributable to euthanasia or PAS was 2.9% in the Netherlands (2010) and 4.6% in Belgium (2013). In 2015, 0.39% of all deaths in Oregon and 0.31% of all deaths in Washington State were reported to be by PAS, although these may be underestimates.

In the Netherlands, Belgium, Oregon, and Washington >70% of patients utilizing these interventions are dying of cancer; <10% of deaths by euthanasia or PAS involve patients with AIDS or amyotrophic lateral sclerosis.

Pain is not the primary motivator for patients' requests for or interest in euthanasia and/or PAS. Among the first patients to receive PAS in Oregon, only 1 of the 15 patients had inadequate pain control, compared with 15 of the 43 patients in a control group who experienced inadequate pain relief. Only 25% of patients in Oregon seeking PAS currently cite pain or fear of pain as their main reason for doing so. Conversely, depression and hopelessness are strongly associated with patient interest in euthanasia and PAS. Concerns about loss of dignity or autonomy or being a burden on family members appear to be more important factors motivating a desire for euthanasia or PAS. Losing autonomy (91% Oregon, 90% Washington), not being able to enjoy activities (89% OR, 89% WA), or fear of losing dignity (68% OR, 76% WA) are the most cited end of life concerns in both states. Over a third of patients seeking PAS note being a burden on family (41% OR, 53% WA). A study from the Netherlands showed that depressed terminally ill cancer patients were four times more likely to request euthanasia and confirmed that uncontrolled pain was not associated with greater interest in euthanasia.

Euthanasia and PAS are no guarantee of a painless, quick death. Data from the Netherlands indicate that in as many as 20% of euthanasia and PAS cases technical and other problems arose, including patients waking from coma, not becoming comatose, regurgitating medications, and experiencing a prolonged time to death. Data from Oregon indicate that between 1998 and 2015, 53% of cases had no complications, 44% of patients had no data on complications, and 2.4% of cases had regurgitation after taking the prescribed medicine as the only complication. In addition, six patients awakened and the reported range of time to death extended to 104 h. In Washington State between 2014 and 2015, 1.4% of cases had regurgitation, 1 patient had a

seizure, and the reported range of time to death extended to 30 h. In the Netherlands, problems were significantly more common in PAS, sometimes requiring the physician to intervene and provide euthanasia.

Regardless of whether they practice in a setting where euthanasia is legal or not, many physicians over the course of their careers will receive a patient request for euthanasia or PAS. In the United States, 18% of physicians have received a request for PAS and 11% have received a request for euthanasia. Three percent complied with a request for PAS, while 5% complied with a request for euthanasia. In the Netherlands, where the practices are legal, 77% of physicians have received a request for PAS or euthanasia and 60% have performed these interventions.

Competency in dealing with such a request is crucial. Although challenging, the request can also provide a chance to address intense suffering. After receiving a request for euthanasia and/or PAS, health care providers should carefully clarify the request with empathic, open-ended questions to help elucidate the underlying cause for the request, such as: “What makes you want to consider this option?” Endorsing either moral opposition or moral support for the act tends to be counterproductive, giving an impression of being judgmental or of endorsing the idea that the patient’s life is worthless. Health care providers must reassure the patient of continued care and commitment. The patient should be educated about alternative, less controversial options, such as symptom management and withdrawing any unwanted treatments and the reality of euthanasia and/or PAS, since the patient may have misconceptions about their effectiveness as well as the legal implications of the choice. Depression, hopelessness, and other symptoms of psychological distress as well as physical suffering and economic burdens are likely factors motivating the request, and such factors should be assessed and treated aggressively. After these interventions and clarification of options, most patients proceed with another approach, declining life-sustaining interventions, possibly including refusal of nutrition and hydration.

CARE DURING THE LAST HOURS

Most laypersons have limited experiences with the actual dying process and death. They frequently do not know what to expect of the final hours and afterward. The family and other caregivers must be prepared, especially if the plan is for the patient to die at home.

Patients in the last days of life typically experience extreme weakness and fatigue and become bedbound; this can lead to pressure sores. The issue of turning patients who are near the end of life, however, must be balanced against the potential discomfort that movement may cause. Patients stop eating and drinking with drying of mucosal membranes and dysphagia. Careful attention to oral swabbing, lubricants for lips, and use of artificial tears can provide a form of care to substitute for attempts at feeding the patient. With loss of the gag reflex and dysphagia, patients may also experience accumulation of oral secretions, producing noises during respiration sometimes called “the death rattle.” Scopolamine can reduce the secretions. Patients also experience changes in respiration with periods of apnea or Cheyne-Stokes breathing. Decreased intravascular volume and cardiac output cause tachycardia, hypotension, peripheral coolness, and livedo reticularis (skin mottling). Patients can have urinary and, less frequently, fecal incontinence. Changes in consciousness and neurologic function generally lead to two different paths to death.

Each of these terminal changes can cause patients and families distress, requiring reassurance and targeted interventions (Table 9-9). Informing families that these changes might occur and providing them with an information sheet can help preempt problems and minimize distress. Understanding that patients stop eating because they are dying, not dying because they have stopped eating, can reduce family and caregiver anxiety. Similarly, informing the family and caregivers that the “death rattle” may occur and that it is not indicative of suffocation, choking, or pain can reduce their worry from the breathing sounds.

TABLE 9-9

Managing Changes in the Patient's Condition during the Final Days and Hours

Changes in the Patient's Condition	Potential Complication	Family's Possible Reaction and Concern	Advice and Intervention
Profound fatigue	Bedbound with development of pressure ulcers that are prone to infection, malodor, and pain, and joint pain	Patient is lazy and giving up.	Reassure family and caregivers that terminal fatigue will not respond to interventions and should not be resisted. Use an air mattress if necessary.
Anorexia	None	Patient is giving up; patient will suffer from hunger and will starve to death.	Reassure family and caregivers that the patient is not eating because he or she is dying; not eating at the end of life does not cause suffering or death. Forced feeding, whether oral, parenteral, or enteral, does not reduce symptoms or prolong life.
Dehydration	Dry mucosal membranes (see below)	Patient will suffer from thirst and die of dehydration.	Reassure family and caregivers that dehydration at the end of life does not cause suffering because patients lose consciousness before any symptom distress. Intravenous hydration can worsen symptoms of dyspnea by pulmonary edema and peripheral edema as well as prolong dying process.

Changes in the Patient's Condition	Potential Complication	Family's Possible Reaction and Concern	Advice and Intervention
Dysphagia	Inability to swallow oral medications needed for palliative care		Do not force oral intake. Discontinue unnecessary medications that may have been continued, including antibiotics, diuretics, antidepressants, and laxatives. If swallowing pills is difficult, convert essential medications (analgesics, antiemetics, anxiolytics, and psychotropics) to oral solutions, buccal, sublingual, or rectal administration.
"Death rattle"—noisy breathing		Patient is choking and suffocating.	Reassure the family and caregivers that this is caused by secretions in the oropharynx and the patient is not choking. Reduce secretions with scopolamine (0.2–0.4 mg SC q4h or 1–3 patches q3d). Reposition patient to permit drainage of secretions. Do not suction. Suction can cause patient and family discomfort and is usually ineffective.
Apnea, Cheyne-Stokes respirations, dyspnea		Patient is suffocating.	Reassure family and caregivers that unconscious patients do not experience suffocation or air hunger. Apneic episodes are frequently a premorbid change. Opioids or anxiolytics may be used for dyspnea. Oxygen is unlikely to relieve dyspneic symptoms and may prolong the dying process.

Changes in the Patient's Condition	Potential Complication	Family's Possible Reaction and Concern	Advice and Intervention
Urinary or fecal incontinence	Skin breakdown if days until death Potential transmission of infectious agents to caregivers	Patient is dirty, malodorous, and physically repellent.	Remind family and caregivers to use universal precautions. Frequent changes of bedclothes and bedding. Use diapers, urinary catheter, or rectal tube if diarrhea or high urine output.
Agitation or delirium	Day/night reversal Hurt self or caregivers	Patient is in horrible pain and going to have a horrible death.	Reassure family and caregivers that agitation and delirium do not necessarily connote physical pain. Depending on the prognosis and goals of treatment, consider evaluating for causes of delirium and modify medications. Manage symptoms with haloperidol, chlorpromazine, diazepam, or midazolam.
Dry mucosal membranes	Cracked lips, mouth sores, and candidiasis can also cause pain. Odor	Patient may be malodorous, physically repellent.	Use baking soda mouthwash or saliva preparation q15–30 min. Use topical nystatin for candidiasis. Coat lips and nasal mucosa with petroleum jelly q60–90 min. Use ophthalmic lubricants q4h or artificial tears q30 min.

Families and caregivers may also feel guilty about stopping treatments, fearing that they are "killing" the patient. This may lead to demands for interventions, such as feeding tubes, that may be ineffective. In such cases, the physician should remind the family and caregivers about the inevitability of events and the palliative goals. Interventions may prolong the dying process and cause discomfort. Physicians also should emphasize that withholding treatments is both legal and ethical and that the family members are not the cause of the patient's death. This reassurance may have to be provided multiple times.

Hearing and touch are said to be the last senses to stop functioning. Whether this is the case or not, families and caregivers can be encouraged to communicate with the dying patient. Encouraging them to talk directly to the patient, even if he or she is unconscious, and hold the patient's hand or demonstrate affection in other ways can be an effective way to channel their urge "to do something" for the patient.

When the plan is for the patient to die at home, the physician must inform the family and caregivers how to determine that the patient has died. The cardinal signs are cessation of cardiac function and respiration; the pupils become fixed; the body becomes cool; muscles relax; and incontinence may occur. Remind the family and caregivers that the eyes may remain open even after the patient has died.

The physician should establish a plan for who the family or caregivers will contact when the patient is dying or has died. Without a plan, family members may panic and call 911, unleashing a cascade of unwanted events, from arrival of emergency personnel and resuscitation to hospital admission. The family and caregivers should be instructed to contact the hospice (if one is involved), the covering physician, or the on-call member of the palliative care team. They should also be told that the medical examiner need not be called unless the state requires it for all deaths. Unless foul play is suspected, the health care team need not contact the medical examiner either.

Just after the patient dies, even the best-prepared family may experience shock and loss and be emotionally distraught. They need time to assimilate the event and be comforted. Health care providers are likely to find it meaningful to write a bereavement card or letter to the family. The purpose is to communicate about the patient, perhaps emphasizing the patient's virtues and the honor it was to care for the patient, and to express concern for the family's hardship. Some physicians attend the funerals of their patients. Although this is beyond any medical obligation, the presence of the physician can be a source of support to the grieving family and provides an opportunity for closure for the physician.

Death of a spouse is a strong predictor of poor health, and even mortality, for the surviving spouse. It may be important to alert the spouse's physician about the death so that he or she is aware of symptoms that might require professional attention.

FURTHER READING

Emanuel E et al: Attitudes and practices of euthanasia and physician-assisted suicide in the United States, Canada, and Europe. JAMA 316:79, 2016. [PubMed: 27380345]

Kelley AS, Meier DE: Palliative care —A shifting paradigm. N Engl J Med 363:781, 2010. [PubMed: 20818881]

Kelley AS et al: Hospice enrollment saves money for medicare and improves care quality across a number of different lengths-of-stay. Health Affairs 32:552, 2012.

Kelley AS et al: Palliative care for the seriously ill. N Engl J Med 373:747, 2015. [PubMed: 26287850]

Mack JW et al: Associations between end-of-life discussion characteristics and care received near death: A prospective cohort study. J Clin Oncol 30:4387, 2012. [PubMed: 23150700]

Murray SA et al: Illness trajectories and palliative care. BMJ 330:1007, 2005. [PubMed: 15860828]

Neuman P et al: Medicare per capita spending by age and service: New data highlights oldest beneficiaries. Health Aff (Millwood) 34:335, 2015. [PubMed: 25588646]

Nicholas LH et al: Regional variation in the association between advance directives and end-of-life Medicare expenditures. JAMA 306:1447, 2011. [PubMed: 21972306]

Teno JM et al: Change in end-of-life care for medicare beneficiaries: Site of death, place of care, and health transitions in 2000, 2005, and 2009. JAMA 309:470, 2013. [PubMed: 23385273]

Van Den Beuken-Van Everdingen MH et al: Update on prevalence of pain in patients with cancer: Systematic review and meta-analysis. J Pain Symptom Manage 51:1070, 2016. [PubMed: 27112310]

WEBSITES

American Academy of Hospice and Palliative Medicine: *www.aahpm.org*

Center to Advance Palliative Care: *http://www.capc.org*

Education in Palliative and End of Life Care (EPEC): *http://www.epec.net*

End of Life—Palliative Education Resource Center: *https://library.tmc.edu/website/end-of-lifepalliative-education-resource-center-eperc/*

Family Caregiver Alliance: *http://www.caregiver.org*

The Medical Directive: *http://www.medicaldirective.org*

National Family Caregivers Association: *http://www.nfcacares.org/*

National Hospice and Palliative Care Organization (including state-specific advance directives): *http://www.nhpco.org*

NCCN: The National Comprehensive Cancer Network palliative care guidelines: *http://www.nccn.org*

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library

Silverchair

Chapter 10: Pain: Pathophysiology and Management

James P. Rathmell; Howard L. Fields

INTRODUCTION

The province of medicine is to preserve and restore health and to relieve suffering. Understanding pain is essential to both of these goals. Because pain is universally understood as a signal of disease, it is the most common symptom that brings a patient to a physician's attention. The function of the pain sensory system is to protect the body and maintain homeostasis. It does this by detecting, localizing, and identifying potential or actual tissue-damaging processes. Because different diseases produce characteristic patterns of tissue damage, the quality, time course, and location of a patient's pain lend important diagnostic clues. It is the physician's responsibility to assess each patient promptly for any remediable cause underlying the pain and to provide rapid and effective pain relief whenever possible.

THE PAIN SENSORY SYSTEM

Pain is an unpleasant sensation localized to a part of the body. It is often described in terms of a penetrating or tissue-destructive process (e.g., stabbing, burning, twisting, tearing, squeezing) and/or of a bodily or emotional reaction (e.g., terrifying, nauseating, sickening). Furthermore, any pain of moderate or higher intensity is accompanied by anxiety and the urge to escape or terminate the feeling. These properties illustrate the duality of pain: it is both sensation and emotion. When it is acute, pain is characteristically associated with behavioral arousal and a stress response consisting of increased blood pressure, heart rate, pupil diameter, and plasma cortisol levels. In addition, local muscle contraction (e.g., limb flexion, abdominal wall rigidity) is often present.

PERIPHERAL MECHANISMS

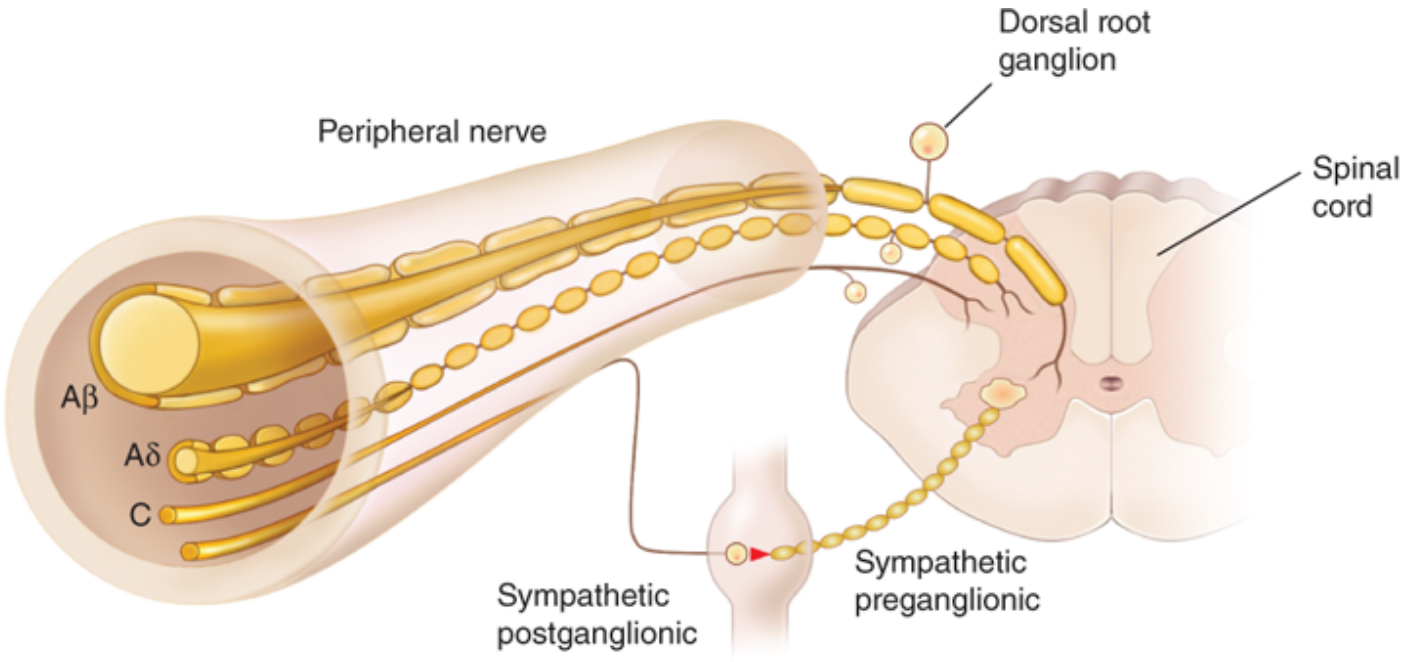
The Primary Afferent Nociceptor

A peripheral nerve consists of the axons of three different types of neurons: primary sensory afferents, motor neurons, and sympathetic postganglionic neurons **(Fig. 10-1)**. The cell bodies of primary sensory afferents are located in the dorsal root ganglia within the vertebral foramina. The primary afferent axon has two branches: one projects centrally into the spinal cord and the other projects peripherally to innervate tissues. Primary afferents are classified by their diameter, degree of myelination, and conduction velocity. The largest diameter afferent fibers, A-beta (Aβ), respond maximally to light touch and/or moving stimuli; they are present primarily in nerves that innervate the skin. In normal individuals, the activity of these fibers does not produce pain. There are two other classes of primary afferent nerve fibers: the small diameter myelinated A-delta (Aδ) and the

unmyelinated (C) axons (Fig. 10-1). These fibers are present in nerves to the skin and to deep somatic and visceral structures. Some tissues, such as the cornea, are innervated only by Aδ and C fiber afferents. Most Aδ and C fiber afferents respond maximally only to intense (painful) stimuli and produce the subjective experience of pain when they are electrically stimulated; this defines them as *primary afferent nociceptors* (*pain receptors*). The ability to detect painful stimuli is completely abolished when conduction in Aδ and C fiber axons is blocked.

FIGURE 10-1

Components of a typical cutaneous nerve. There are two distinct functional categories of axons: primary afferents with cell bodies in the dorsal root ganglion and sympathetic postganglionic fibers with cell bodies in the sympathetic ganglion. Primary afferents include those with large-diameter myelinated (Aβ), small-diameter myelinated (Aδ), and unmyelinated (C) axons. All sympathetic postganglionic fibers are unmyelinated.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Individual primary afferent nociceptors can respond to several different types of noxious stimuli. For example, most nociceptors respond to heat; intense cold; intense mechanical distortion, such as a pinch; changes in pH, particularly an acidic environment; and application of chemical irritants including adenosine triphosphate (ATP), serotonin, bradykinin (BK), and histamine. The transient receptor potential cation channel subfamily V member 1 (TrpV1), also known as the vanilloid receptor, mediates perception of some noxious stimuli, especially heat sensations, by nociceptive neurons; it is activated by acidic pH, endogenous mediators and by capsaicin, a component of hot chili peppers.

Sensitization

When intense, repeated, or prolonged stimuli are applied to damaged or inflamed tissues, the threshold for activating primary afferent nociceptors is lowered, and the frequency of firing is higher for all stimulus intensities. Inflammatory mediators such as BK, nerve-growth factor, some prostaglandins (PGs), and leukotrienes contribute to this process, which is called *sensitization*. Sensitization occurs at the level of the

peripheral nerve terminal (*peripheral sensitization*) as well as at the level of the dorsal horn of the spinal cord (*central sensitization*). Peripheral sensitization occurs in damaged or inflamed tissues, when inflammatory mediators activate intracellular signal transduction in nociceptors, prompting an increase in the production, transport, and membrane insertion of chemically gated and voltage-gated ion channels. These changes increase the excitability of nociceptor terminals and lower their threshold for activation by mechanical, thermal, and chemical stimuli. Central sensitization occurs when activity, generated by nociceptors during inflammation, enhances the excitability of nerve cells in the dorsal horn of the spinal cord. Following injury and resultant sensitization, normally innocuous stimuli can produce pain (termed *allodynia*). Sensitization is a clinically important process that contributes to tenderness, soreness, and *hyperalgesia* (increased pain intensity in response to the same noxious stimulus; e.g., pinprick causes severe pain). A striking example of sensitization is sunburned skin, in which severe pain can be produced by a gentle slap on the back or a warm shower.

Sensitization is of particular importance for pain and tenderness in deep tissues. Viscera are normally relatively insensitive to noxious mechanical and thermal stimuli, although hollow viscera do generate significant discomfort when distended. In contrast, when affected by a disease process with an inflammatory component, deep structures such as joints or hollow viscera characteristically become exquisitely sensitive to mechanical stimulation.

A large proportion of Aδ and C fiber afferents innervating viscera are completely insensitive in normal noninjured, noninflamed tissue. That is, they cannot be activated by known mechanical or thermal stimuli and are not spontaneously active. However, in the presence of inflammatory mediators, these afferents become sensitive to mechanical stimuli. Such afferents have been termed *silent nociceptors*, and their characteristic properties may explain how, under pathologic conditions, the relatively insensitive deep structures can become the source of severe and debilitating pain and tenderness. Low pH, PGs, leukotrienes, and other inflammatory mediators such as BK play a significant role in sensitization.

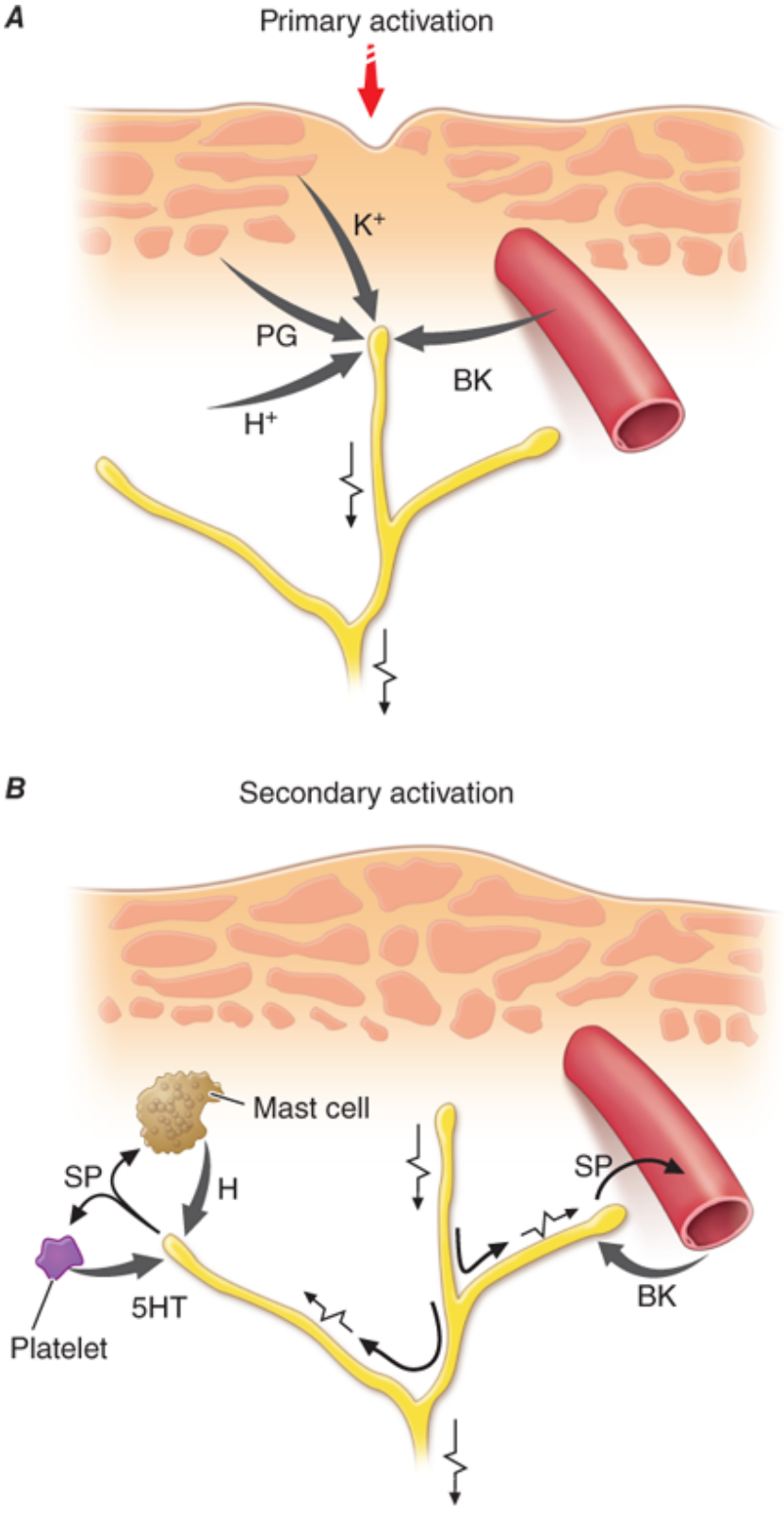
Nociceptor-Induced Inflammation

Primary afferent nociceptors also have a neuroeffector function. Most nociceptors contain polypeptide mediators that are released from their peripheral terminals when they are activated **(Fig. 10-2)**. An example is substance P, an 11-amino-acid peptide. Substance P is released from primary afferent nociceptors and has multiple biologic activities. It is a potent vasodilator, causes mast cell degranulation, is a chemoattractant for leukocytes, and increases the production and release of inflammatory mediators. Interestingly, depletion of substance P from joints reduces the severity of experimental arthritis. Primary afferent nociceptors are not simply passive messengers of threats to tissue injury but also play an active role in tissue protection through these neuroeffector functions.

FIGURE 10-2

Events leading to activation, sensitization, and spread of sensitization of primary afferent nociceptor terminals. ***A***. Direct activation by intense pressure and consequent cell damage. Cell damage induces lower pH (H^+) and leads to release of potassium (K^+) and to synthesis of prostaglandins (PGs) and bradykinin (BK). PGs increase the sensitivity of the terminal to BK and other pain-producing substances. ***B***. Secondary activation. Impulses

generated in the stimulated terminal propagate not only to the spinal cord but also into other terminal branches where they induce the release of peptides, including substance P (SP). Substance P causes vasodilation and neurogenic edema with further accumulation of BK. Substance P also causes the release of histamine (H) from mast cells and serotonin (5HT) from platelets.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

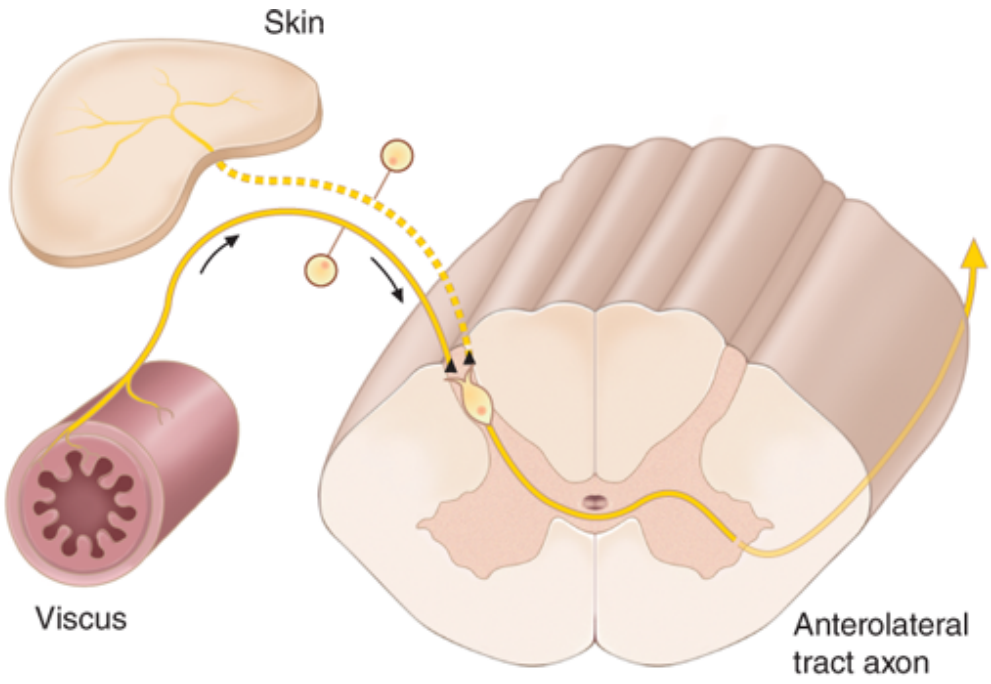
CENTRAL MECHANISMS

The Spinal Cord and Referred Pain

The axons of primary afferent nociceptors enter the spinal cord via the dorsal root. They terminate in the dorsal horn of the spinal gray matter **(Fig. 10-3)**. The terminals of primary afferent axons contact spinal neurons that transmit the pain signal to brain sites involved in pain perception. When primary afferents are activated by noxious stimuli, they release neurotransmitters from their terminals that excite the spinal cord neurons. The major neurotransmitter released is glutamate, which rapidly excites the second-order dorsal horn neurons. Primary afferent nociceptor terminals also release peptides, including substance P and calcitonin gene-related peptide, which produce a slower and longer-lasting excitation of the dorsal horn neurons. The axon of each primary afferent contacts many spinal neurons, and each spinal neuron receives convergent inputs from many primary afferents.

FIGURE 10-3

The convergence-projection hypothesis of referred pain. According to this hypothesis, visceral afferent nociceptors converge on the same pain-projection neurons as the afferents from the somatic structures in which the pain is perceived. The brain has no way of knowing the actual source of input and mistakenly "projects" the sensation to the somatic structure.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The convergence of sensory inputs to a single spinal pain-transmission neuron is of great importance because it underlies the phenomenon of referred pain. All spinal neurons that receive input from the viscera and deep musculoskeletal structures also receive input from the skin. The convergence patterns are determined by the spinal segment of the dorsal root ganglion that supplies the afferent innervation of a structure. For example, the afferents that supply the central diaphragm are derived from the third and fourth cervical dorsal root ganglia. Primary afferents with cell bodies in these same ganglia supply the skin of the shoulder and lower neck. Thus,

sensory inputs from both the shoulder skin and the central diaphragm converge on pain-transmission neurons in the third and fourth cervical spinal segments. *Because of this convergence and the fact that the spinal neurons are most often activated by inputs from the skin, activity evoked in spinal neurons by input from deep structures is mislocalized by the patient to a place that roughly corresponds with the region of skin innervated by the same spinal segment*. Thus, inflammation near the central diaphragm is often reported as shoulder discomfort. This spatial displacement of pain sensation from the site of the injury that produces it is known as *referred pain*.

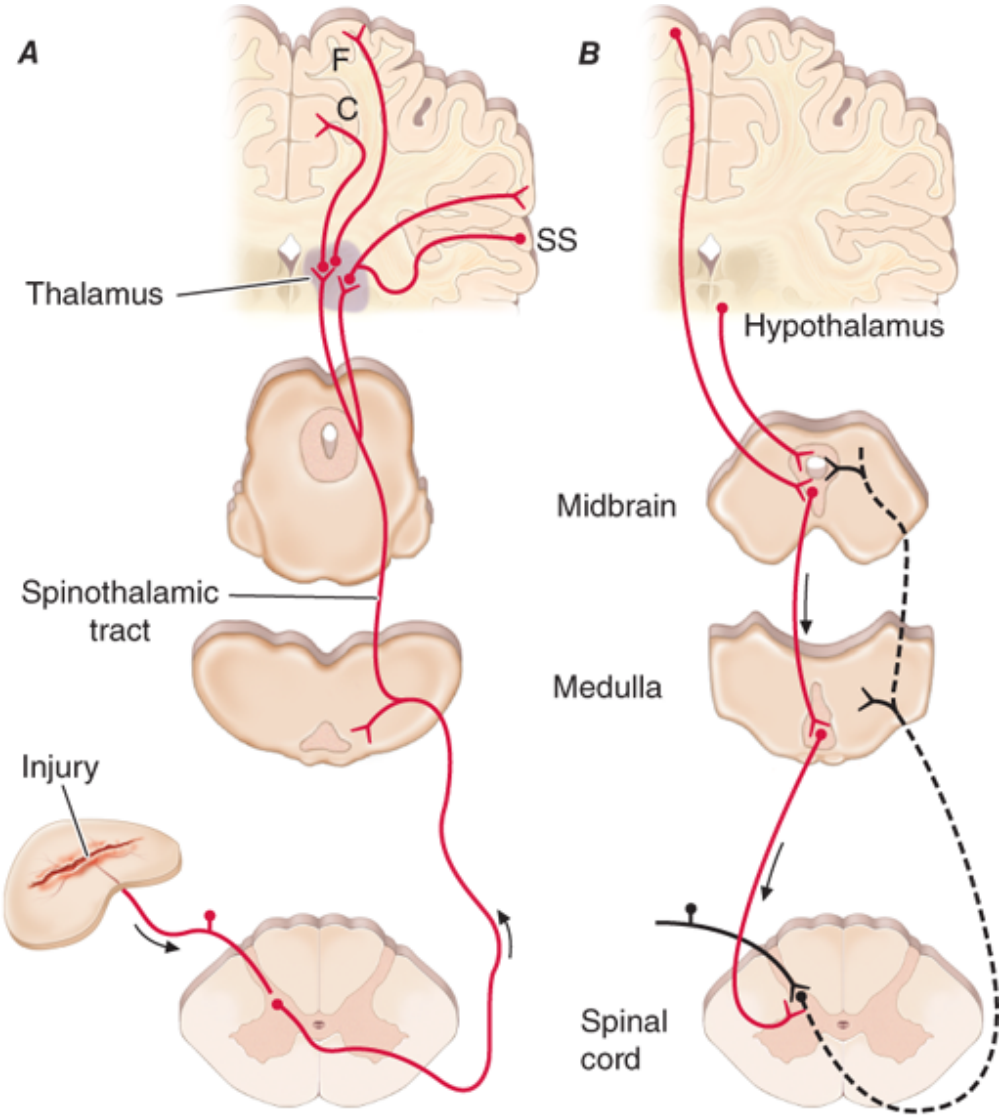
Ascending Pathways for Pain

A majority of spinal neurons contacted by primary afferent nociceptors send their axons to the contralateral thalamus. These axons form the contralateral spinothalamic tract, which lies in the anterolateral white matter of the spinal cord, the lateral edge of the medulla, and the lateral pons and midbrain. The spinothalamic pathway is crucial for pain sensation in humans. Interruption of this pathway produces permanent deficits in pain and temperature discrimination.

Spinothalamic tract axons ascend to several regions of the thalamus. There is tremendous divergence of the pain signal from these thalamic sites to several distinct areas of the cerebral cortex that subserve different aspects of the pain experience **(Fig. 10-4)**. One of the thalamic projections is to the somatosensory cortex. This projection mediates the purely sensory aspects of pain, i.e., its location, intensity, and quality. Other thalamic neurons project to cortical regions that are linked to emotional responses, such as the cingulate gyrus and other areas of the frontal lobes, including the insular cortex. These pathways to the frontal cortex subserve the affective or unpleasant emotional dimension of pain. This affective dimension of pain produces suffering and exerts potent control of behavior. Because of this dimension, fear is a constant companion of pain. As a consequence, injury or surgical lesions to areas of the frontal cortex activated by painful stimuli can diminish the emotional impact of pain while largely preserving the individual's ability to recognize noxious stimuli as painful.

FIGURE 10-4

Pain-transmission and modulatory pathways. ***A***. Transmission system for nociceptive messages. Noxious stimuli activate the sensitive peripheral ending of the primary afferent nociceptor by the process of transduction. The message is then transmitted over the peripheral nerve to the spinal cord, where it synapses with cells of origin of the major ascending pain pathway, the spinothalamic tract. The message is relayed in the thalamus to the anterior cingulate (C), frontal insular (F), and somatosensory cortex (SS). ***B***. Pain-modulation network. Inputs from frontal cortex and hypothalamus activate cells in the midbrain that control spinal pain-transmission cells via cells in the medulla.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PAIN MODULATION

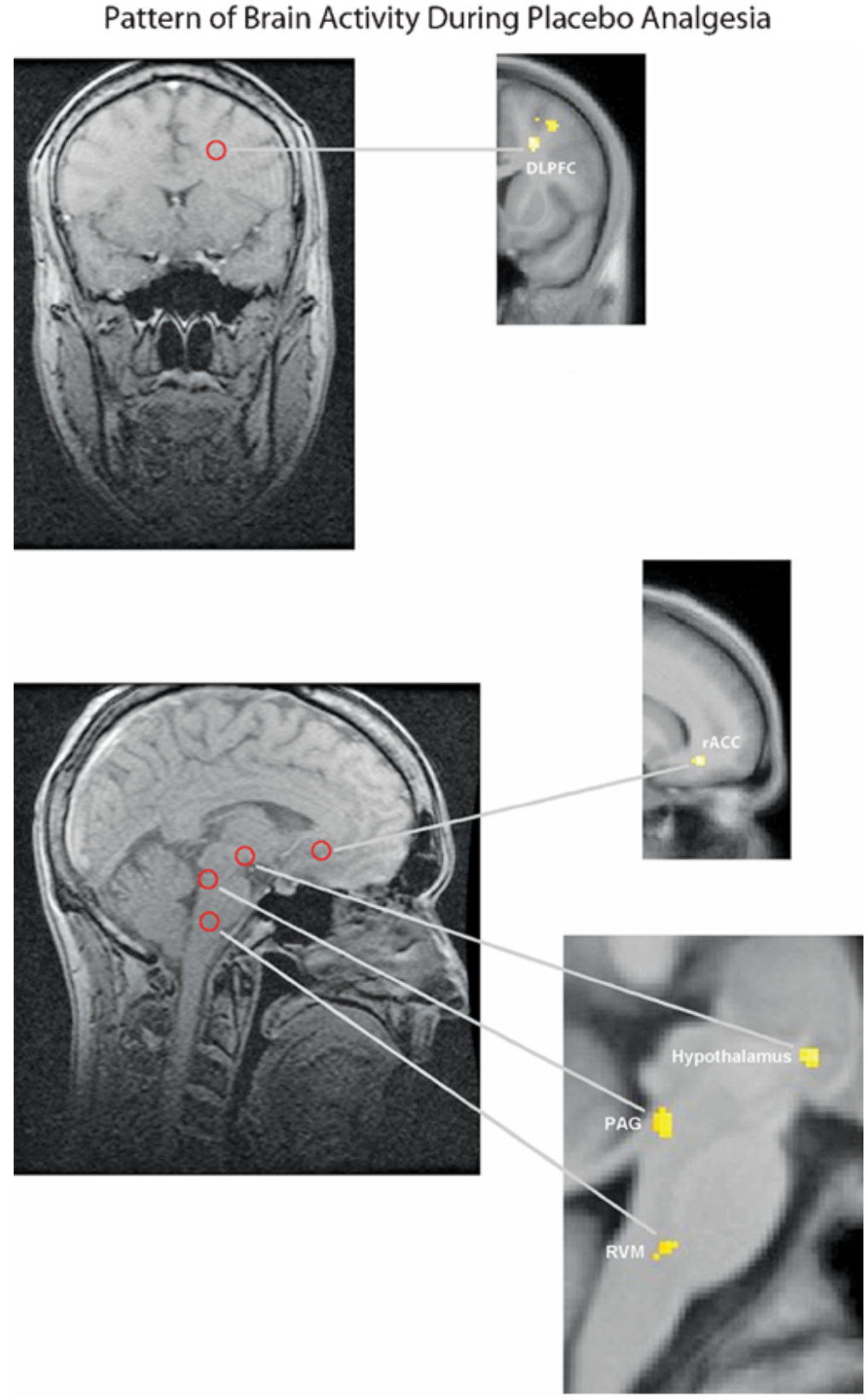
The pain produced by injuries of similar magnitude is remarkably variable in different situations and in different individuals. For example, athletes have been known to sustain serious fractures with only minor pain, and Beecher's classic World War II survey revealed that many soldiers in battle were unbothered by injuries that would have produced agonizing pain in civilian patients. Furthermore, even the suggestion that a treatment will relieve pain can have a significant analgesic effect (the *placebo effect*). On the other hand, many patients find even minor injuries such as venipuncture frightening and unbearable, and the expectation of pain can induce pain even without a noxious stimulus. The suggestion that pain will worsen following administration of an inert substance can increase its perceived intensity (the *nocebo effect*).

The powerful effect of expectation and other psychological variables on the perceived intensity of pain is explained by brain circuits that modulate the activity of the pain-transmission pathways. One of these circuits has links to the hypothalamus, midbrain, and medulla, and it selectively controls spinal pain-transmission neurons through a descending pathway (Fig. 10-4).

Human brain-imaging studies have implicated this pain-modulating circuit in the pain-relieving effect of attention, suggestion, and opioid analgesic medications **(Fig. 10-5)**. Furthermore, each of the component structures of the pathway contains opioid receptors and is sensitive to the direct application of opioid drugs. In animals, lesions of this descending modulatory system reduce the analgesic effect of systemically administered opioids such as morphine. Along with the opioid receptor, the component nuclei of this pain-modulating circuit contain endogenous opioid peptides such as the enkephalins and β-endorphin.

FIGURE 10-5

Functional magnetic resonance imaging (fMRI) demonstrates placebo-enhanced brain activity in anatomic regions correlating with the opioidergic descending pain control system. *Top panel:* Frontal fMRI image shows placebo-enhanced brain activity in the dorsal lateral prefrontal cortex (DLPFC). *Bottom panel:* Sagittal fMRI images show placebo-enhanced responses in the rostral anterior cingulate cortex (rACC), the rostral ventral medullae (RVM), the periaqueductal gray (PAG) area, and the hypothalamus. The placebo-enhanced activity in all areas was reduced by naloxone, demonstrating the link between the descending opioidergic system and the placebo analgesic response. *(Adapted with permission from F Eippert et al: Neuron 63:533, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The most reliable way to activate this endogenous opioid-mediated modulating system is by suggestion of pain relief or by intense emotion directed away from the pain-causing injury (e.g., during severe threat or an athletic competition). In fact, pain-relieving endogenous opioids are released following surgical procedures and in patients given a placebo for pain relief.

Pain-modulating circuits can enhance as well as suppress pain. Both pain-inhibiting and pain-facilitating neurons in the medulla project to and control spinal pain-transmission neurons. Because pain-transmission neurons can be activated by modulatory neurons, it is theoretically possible to generate a pain signal with no peripheral noxious stimulus. In fact, human functional imaging studies have demonstrated increased activity in this circuit during migraine headaches. A central circuit that facilitates pain could account for the finding that pain can be induced by suggestion or enhanced by expectation and provides a framework for understanding how psychological factors can contribute to chronic pain.

NEUROPATHIC PAIN

Lesions of the peripheral or central nociceptive pathways typically result in a loss or impairment of pain sensation. Paradoxically, damage to or dysfunction of these pathways can also produce pain. For example, damage to peripheral nerves, as occurs in diabetic neuropathy, or to primary afferents, as in herpes zoster infection, can result in pain that is referred to the body region innervated by the damaged nerves. Pain may also be produced by damage to the central nervous system (CNS), for example, in some patients following trauma or vascular injury to the spinal cord, brainstem, or thalamic areas that contain central nociceptive pathways. Such pains are termed *neuropathic* and are often severe and are typically resistant to standard treatments for pain.

Neuropathic pain typically has an unusual burning, tingling, or electric shock-like quality and may occur spontaneously, without any stimulus, or be triggered by very light touch. These features are rare in other types of pain. On examination, a sensory deficit is characteristically present in the area of the patient's pain. *Hyperpathia*, a greatly exaggerated pain response to innocuous or mild nociceptive stimuli, especially when applied repeatedly, is also characteristic of neuropathic pain; patients often complain that the very lightest moving stimulus evokes exquisite pain (allodynia). In this regard, it is of clinical interest that a topical preparation of 5% lidocaine in patch form is effective for patients with postherpetic neuralgia who have prominent allodynia.

A variety of mechanisms contribute to neuropathic pain. As with sensitized primary afferent nociceptors, damaged primary afferents, including nociceptors, become highly sensitive to mechanical stimulation and may generate impulses in the absence of stimulation. Increased sensitivity and spontaneous activity are due, in part, to an increased density of sodium channels in the damaged nerve fiber. Damaged primary afferents may also develop sensitivity to norepinephrine. Interestingly, spinal cord pain-transmission neurons cut off from their normal input may also become spontaneously active. Thus, both central and peripheral nervous system hyperactivity contribute to neuropathic pain.

Sympathetically Maintained Pain

Patients with peripheral nerve injury occasionally develop spontaneous pain in the region innervated by the nerve. This pain is often described as having a burning quality. The pain typically begins after a delay of hours to days or even weeks and is accompanied by swelling of the extremity, periarticular bone loss, and arthritic changes in the distal joints. The pain may be relieved by a local anesthetic block of the sympathetic innervation to the affected extremity. Damaged primary afferent nociceptors acquire adrenergic sensitivity and can be activated by stimulation of the sympathetic outflow. This constellation of spontaneous pain and signs of

sympathetic dysfunction following injury has been termed *complex regional pain syndrome* (CRPS). When this occurs after an identifiable nerve injury, it is termed CRPS type II (also known as posttraumatic neuralgia or, if severe, *causalgia*). When a similar clinical picture appears without obvious nerve injury, it is termed CRPS type I (also known as *reflex sympathetic dystrophy*). CRPS can be produced by a variety of injuries, including fractures of bone, soft tissue trauma, myocardial infarction, and stroke. CRPS type I typically resolves with symptomatic treatment; however, when it persists, detailed examination often reveals evidence of peripheral nerve injury. Although the pathophysiology of CRPS is poorly understood, the pain and the signs of inflammation, when acute, can be rapidly relieved by blocking the sympathetic nervous system. This implies that sympathetic activity can activate undamaged nociceptors when inflammation is present. Signs of sympathetic hyperactivity should be sought in patients with posttraumatic pain and inflammation and no other obvious explanation.

TREATMENT

TREATMENT

Acute Pain

The ideal treatment for any pain is to remove the cause; thus, while treatment can be initiated immediately, efforts to establish the underlying etiology should always proceed as treatment begins. Sometimes, treating the underlying condition does not immediately relieve pain. Furthermore, some conditions are so painful that rapid and effective analgesia is essential (e.g., the postoperative state, burns, trauma, cancer, or sickle cell crisis). Analgesic medications are a first line of treatment in these cases, and all practitioners should be familiar with their use.

ASPIRIN, ACETAMINOPHEN, AND NONSTEROIDAL ANTI-INFLAMMATORY AGENTS (NSAIDs)

These drugs are considered together because they are used for similar problems and may have a similar mechanism of action **(Table 10-1)**. All these compounds inhibit cyclooxygenase (COX), and, except for acetaminophen, all have anti-inflammatory actions, especially at higher dosages. They are particularly effective for mild to moderate headache and for pain of musculoskeletal origin.

Because they are effective for these common types of pain and are available without prescription, COX inhibitors are by far the most commonly used analgesics. They are absorbed well from the gastrointestinal tract and, with occasional use, have only minimal side effects. With chronic use, gastric irritation is a common side effect of aspirin and NSAIDs and is the problem that most frequently limits the dose that can be given. Gastric irritation is most severe with aspirin, which may cause erosion and ulceration of the gastric mucosa leading to bleeding or perforation. Because aspirin irreversibly acetylates platelet COX and thereby interferes with coagulation of the blood, gastrointestinal bleeding is a particular risk. Older age and history of gastrointestinal disease increase the risks of aspirin and NSAIDs. In addition to the well-known gastrointestinal toxicity of NSAIDs, nephrotoxicity is a significant problem for patients using these drugs on a chronic basis. Patients at risk for renal insufficiency, particularly those with significant contraction of their intravascular volume as occurs with chronic diuretic use or acute hypovolemia, should avoid NSAIDs. NSAIDs can also increase blood pressure in some individuals. Long-term treatment with NSAIDs requires regular blood pressure monitoring and

treatment if necessary. Although toxic to the liver when taken in high doses, acetaminophen rarely produces gastric irritation and does not interfere with platelet function.

The introduction of parenteral forms of NSAIDs, ketorolac and diclofenac, extends the usefulness of this class of compounds in the management of acute severe pain. Both agents are sufficiently potent and rapid in onset to supplant opioids for many patients with acute severe headache and musculoskeletal pain.

There are two major classes of COX: COX-1 is constitutively expressed, and COX-2 is induced in the inflammatory state. COX-2-selective drugs have similar analgesic potency and produce less gastric irritation than the nonselective COX inhibitors. The use of COX-2-selective drugs does not appear to lower the risk of nephrotoxicity compared to nonselective NSAIDs. On the other hand, COX-2-selective drugs offer a significant benefit in the management of acute postoperative pain because they do not affect blood coagulation. Nonselective COX inhibitors are usually contraindicated postoperatively because they impair platelet-mediated blood clotting and are thus associated with increased bleeding at the operative site. COX-2 inhibitors, including celecoxib (Celebrex), are associated with increased cardiovascular risk, including cardiovascular death, myocardial infarction, stroke, heart failure, or a thromboembolic event. It appears that this is a class effect of NSAIDs, excluding aspirin. These drugs are contraindicated in patients in the immediate period after coronary artery bypass surgery and should be used with caution in elderly patients and those with a history of or significant risk factors for cardiovascular disease.

OPIOID ANALGESICS

Opioids are the most potent pain-relieving drugs currently available. Of all analgesics, they have the broadest range of efficacy and provide the most reliable and effective method for rapid pain relief. Although side effects are common, most are reversible: nausea, vomiting, pruritus, and constipation are the most frequent and bothersome side effects. Respiratory depression is uncommon at standard analgesic doses, but can be life-threatening. Opioid-related side effects can be reversed rapidly with the narcotic antagonist naloxone. Many physicians, nurses, and patients have a certain trepidation about using opioids that is based on a fear of initiating addiction in their patients. In fact, there is a very small chance of patients becoming addicted to narcotics as a result of their appropriate medical use. For chronic pain, particularly chronic noncancer pain, the risk of addiction in patients taking opioids on a chronic basis remains small, but the risk does appear to increase with dose escalation. The physician should not hesitate to use opioid analgesics in patients with acute severe pain. Table 10-1 lists the most commonly used opioid analgesics.

Opioids produce analgesia by actions in the CNS. They activate pain-inhibitory neurons and directly inhibit pain-transmission neurons. Most of the commercially available opioid analgesics act at the same opioid receptor (μ-receptor), differing mainly in potency, speed of onset, duration of action, and optimal route of administration. Some side effects are due to accumulation of nonopioid metabolites that are unique to individual drugs. One striking example of this is normeperidine, a metabolite of meperidine. At higher doses of meperidine, typically >1 g/d, accumulation of normeperidine can produce hyperexcitability and seizures that are not reversible with naloxone. Normeperidine accumulation is increased in patients with renal failure.

The most rapid pain relief is obtained by intravenous administration of opioids; relief with oral administration is significantly slower. Because of the potential for respiratory depression, patients with any form of respiratory

compromise must be kept under close observation following opioid administration; an oxygen-saturation monitor may be useful, but only in a setting where the monitor is under constant surveillance. Opioid-induced respiratory depression is typically accompanied by sedation and a reduction in respiratory rate. A fall in oxygen saturation represents a critical level of respiratory depression and the need for immediate intervention to prevent life-threatening hypoxemia. Newer monitoring devices that incorporate capnography or pharyngeal air flow can detect apnea at the point of onset and should be used in hospitalized patients. Ventilatory assistance should be maintained until the opioid-induced respiratory depression has resolved. The opioid antagonist naloxone should be readily available whenever opioids are used at high doses or in patients with compromised pulmonary function. Opioid effects are dose-related, and there is great variability among patients in the doses that relieve pain and produce side effects. Synergistic respiratory depression is common when opioids are administered with other CNS depressants, most commonly the benzodiazepines. Because of this, initiation of therapy requires titration to optimal dose and interval. The most important principle is to provide adequate pain relief. This requires determining whether the drug has adequately relieved the pain and frequent reassessment to determine the optimal interval for dosing. *The most common error made by physicians in managing severe pain with opioids is to prescribe an inadequate dose. Because many patients are reluctant to complain, this practice leads to needless suffering*. In the absence of sedation at the expected time of peak effect, a physician should not hesitate to repeat the initial dose to achieve satisfactory pain relief.

A now standard approach to the problem of achieving adequate pain relief is the use of patient-controlled analgesia (PCA). PCA uses a microprocessor-controlled infusion device that can deliver a baseline continuous dose of an opioid drug as well as preprogrammed additional doses whenever the patient pushes a button. The patient can then titrate the dose to the optimal level. This approach is used most extensively for the management of postoperative pain, but there is no reason why it should not be used for any hospitalized patient with persistent severe pain. PCA is also used for short-term home care of patients with intractable pain, such as that caused by metastatic cancer.

It is important to understand that the PCA device delivers small, repeated doses to maintain pain relief; in patients with severe pain, the pain must first be brought under control with a loading dose before transitioning to the PCA device. The bolus dose of the drug (typically 1 mg of morphine, 0.2 mg of hydromorphone, or 10 µg of fentanyl) can then be delivered repeatedly as needed. To prevent overdosing, PCA devices are programmed with a lockout period after each demand dose is delivered (typically starting at 10 min) and a limit on the total dose delivered per hour. Although some have advocated the use of a simultaneous continuous or basal infusion of the PCA drug, this may increase the risk of respiratory depression and has not been shown to increase the overall efficacy of the technique.

The availability of new routes of administration has extended the usefulness of opioid analgesics. Most important is the availability of spinal administration. Opioids can be infused through a spinal catheter placed either intrathecally or epidurally. By applying opioids directly to the spinal or epidural space adjacent to the spinal cord, regional analgesia can be obtained using relatively low total doses. Indeed, the dose required to produce effective analgesia when using morphine intrathecally (0.1–0.3 mg) is a fraction of that required to produce similar analgesia when administered intravenously (5–10 mg). In this way, side effects such as sedation, nausea, and respiratory depression can be minimized. This approach has been used extensively

during labor and delivery and for postoperative pain relief following surgical procedures. Continuous intrathecal delivery via implanted spinal drug-delivery systems is now commonly used, particularly for the treatment of cancer-related pain that would require sedating doses for adequate pain control if given systemically. Opioids can also be given intranasally (butorphanol), rectally, and transdermally (fentanyl and buprenorphine), or through the oral mucosa (fentanyl), thus avoiding the discomfort of frequent injections in patients who cannot be given oral medication. The fentanyl and buprenorphine transdermal patches have the advantage of providing fairly steady plasma levels, which may improve patient comfort.

Recent additions to the armamentarium for treating opioid-induced side effects are the peripherally acting opioid antagonists alvimopan (Entereg) and methylnaltrexone (Rellistor). Alvimopan is available as an orally administered agent that is restricted to the intestinal lumen by limited absorption; methylnaltrexone is available in a subcutaneously administered form that has virtually no penetration into the CNS. Both agents act by binding to peripheral μ-receptors, thereby inhibiting or reversing the effects of opioids at these peripheral sites. The action of both agents is restricted to receptor sites outside of the CNS; thus, these drugs can reverse the adverse effects of opioid analgesics that are mediated through their peripheral receptors without reversing their analgesic effects. Alvimopan has proven effective in lowering the duration of persistent ileus following abdominal surgery in patients receiving opioid analgesics for postoperative pain control. Methylnaltrexone has proven effective for relief of opioid-induced constipation in patients taking opioid analgesics on a chronic basis.

Opioid and COX Inhibitor Combinations

When used in combination, opioids and COX inhibitors have additive effects. Because a lower dose of each can be used to achieve the same degree of pain relief and their side effects are nonadditive, such combinations are used to lower the severity of dose-related side effects. However, fixed-ratio combinations of an opioid with acetaminophen carry an important risk. Dose escalation as a result of increased severity of pain or decreased opioid effect as a result of tolerance may lead to ingestion of levels of acetaminophen that are toxic to the liver. Although acetaminophen-related hepatotoxicity is uncommon, it remains a significant cause for liver failure. Thus, many practitioners have moved away from the use of opioid-acetaminophen combination analgesics to avoid the risk of excessive acetaminophen exposure as the dose of the analgesic is escalated.

TABLE 10-1

Drugs for Relief of Pain

Generic Name	Dose, mg	Interval	Comments
Nonnarcotic Analgesics: Usual Doses and Intervals			
Acetylsalicylic acid	650 PO	q4h	Enteric-coated preparations available
Acetaminophen	650 PO	q4h	Side effects uncommon
Ibuprofen	400 PO	q4–6h	Available without prescription
Naproxen	250–500 PO	q12h	Naproxen is the common NSAID that poses the least cardiovascular risk; but it has a somewhat higher incidence of gastrointestinal bleeding
Fenoprofen	200 PO	q4–6h	Contraindicated in renal disease
Indomethacin	25–50 PO	q8h	Gastrointestinal side effects common
Ketorolac	15–60 IM/IV	q4–6h	Available for parenteral use
Celecoxib	100–200 PO	q12–24h	Useful for arthritis
Valdecoxib	10–20 PO	q12–24h	Removed from U.S. market in 2005
Generic Name	**Parenteral Dose, mg**	**PO Dose, mg**	**Comments**
Narcotic Analgesics: Usual Doses and Intervals			
Codeine	30–60 q4h	30–60 q4h	Nausea common
Oxycodone	—	5–10 q4–6h	Usually available with acetaminophen or aspirin

Generic Name	Dose, mg	Interval	Comments
Morphine	5 q4h	30 q4h	
Morphine sustained release	—	15–60 bid to tid	Oral slow-release preparation
Hydromorphone	1–2 q4h	2–4 q4h	Shorter acting than morphine sulfate
Levorphanol	2 q6–8h	4 q6–8h	Longer acting than morphine sulfate; absorbed well PO
Methadone	5–10 q6–8h	5–20 q6–8h	Due to long half-life, respiratory depression and sedation may persist after analgesic effect subsides; therapy should not be initiated with >40 mg/d, and dose escalation should be made no more frequently than every 3 days
Meperidine	50–100 q3–4h	300 q4h	Poorly absorbed PO; normeperidine is a toxic metabolite; routine use of this agent is not recommended
Butorphanol	—	1–2 q4h	Intranasal spray
Fentanyl	25–100 µg/h	—	72-h transdermal patch
Buprenorphine	5–20 µg/h		7-day transdermal patch
Buprenorphine	0.3 q6–8h		Parenteral administration
Tramadol	—	50–100 q4–6h	Mixed opioid/adrenergic action

Generic Name	Uptake Blockade		Sedative Potency	Anticholinergic Potency	Orthostatic Hypotension	Cardiac Arrhythmia	Ave. Dose, mg/d	Range, mg/d
	5-HT	NE						
Antidepressants[a]								

Generic Name		Dose, mg		Interval		Comments		
Doxepin	++	+	High	Moderate	Moderate	Less	200	75–400
Amitriptyline	++++	++	High	Highest	Moderate	Yes	150	25–300
Imipramine	++++	++	Moderate	Moderate	High	Yes	200	75–400
Nortriptyline	+++	++	Moderate	Moderate	Low	Yes	100	40–150
Desipramine	+++	++++	Low	Low	Low	Yes	150	50–300
Venlafaxine	+++	++	Low	None	None	No	150	75–400
Duloxetine	+++	+++	Low	None	None	No	40	30–60

Generic Name	PO Dose, mg	Interval	Generic Name	PO Dose, mg	Interval
Anticonvulsants and Antiarrhythmics[a]					
Phenytoin	300	daily/qhs	Clonazepam	1	q6h
Carbamazepine	200–300	q6h	Gabapentin[b]	600–1200	q8h
Oxcarbazepine	300	bid	Pregabalin	150–600	bid

[a]Antidepressants, anticonvulsants, and antiarrhythmics have not been approved by the U.S. Food and Drug Administration (FDA) for the treatment of pain. [b]Gabapentin in doses up to 1800 mg/d is FDA approved for postherpetic neuralgia.

Abbreviations: 5-HT, serotonin; NE, norepinephrine; NSAID, nonsteroidal anti-inflammatory agent.

CHRONIC PAIN

Managing patients with chronic pain is intellectually and emotionally challenging. Sensitization of the nervous system can occur without an obvious precipitating cause, e.g., fibromyalgia, or chronic headache. In many patients, chronic pain becomes a distinct disease unto itself. The pain-generating mechanism is often difficult or impossible to determine with certainty; such patients are demanding of the physician's time and often appear

emotionally distraught. The traditional medical approach of seeking an obscure organic pathology is usually unhelpful. On the other hand, psychological evaluation and behaviorally based treatment paradigms are frequently helpful, particularly in the setting of a multidisciplinary pain-management center. Unfortunately, this approach, while effective, remains largely underused in current medical practice.

There are several factors that can cause, perpetuate, or exacerbate chronic pain. First, of course, the patient may simply have a disease that is characteristically painful for which there is presently no cure. Arthritis, cancer, chronic daily headaches, fibromyalgia, and diabetic neuropathy are examples of this. Second, there may be secondary perpetuating factors that are initiated by disease and persist after that disease has resolved. Examples include damaged sensory nerves, sympathetic efferent activity, and painful reflex muscle contraction (spasm). Finally, a variety of psychological conditions can exacerbate or even cause pain.

There are certain areas to which special attention should be paid in a patient's medical history. Because depression is the most common emotional disturbance in patients with chronic pain, patients should be questioned about their mood, appetite, sleep patterns, and daily activity. A simple standardized questionnaire, such as the Beck Depression Inventory, can be a useful screening device. It is important to remember that major depression is a common, treatable, and potentially fatal illness.

Other clues that a significant emotional disturbance is contributing to a patient's chronic pain complaint include pain that occurs in multiple, unrelated sites; a pattern of recurrent, but separate, pain problems beginning in childhood or adolescence; pain beginning at a time of emotional trauma, such as the loss of a parent or spouse; a history of physical or sexual abuse; and past or present substance abuse.

On examination, special attention should be paid to whether the patient guards the painful area and whether certain movements or postures are avoided because of pain. Discovering a mechanical component to the pain can be useful both diagnostically and therapeutically. Painful areas should be examined for deep tenderness, noting whether this is localized to muscle, ligamentous structures, or joints. Chronic myofascial pain is very common, and, in these patients, deep palpation may reveal highly localized trigger points that are firm bands or knots in muscle. Relief of the pain following injection of local anesthetic into these trigger points supports the diagnosis. A neuropathic component to the pain is indicated by evidence of nerve damage, such as sensory impairment, exquisitely sensitive skin (allodynia), weakness, and muscle atrophy, or loss of deep tendon reflexes. Evidence suggesting sympathetic nervous system involvement includes the presence of diffuse swelling, changes in skin color and temperature, and hypersensitive skin and joint tenderness compared with the normal side. Relief of the pain with a sympathetic block supports the diagnosis, but once the condition becomes chronic, the response to sympathetic blockade is of variable magnitude and duration; the role for repeated sympathetic blocks in the overall management of CRPS is unclear.

A guiding principle in evaluating patients with chronic pain is to assess both emotional and organic factors before initiating therapy. Addressing these issues together, rather than waiting to address emotional issues after organic causes of pain have been ruled out, improves compliance in part because it assures patients that a psychological evaluation does not mean that the physician is questioning the validity of their complaint. Even when an organic cause for a patient's pain can be found, it is still wise to look for other factors. For example, a

cancer patient with painful bony metastases may have additional pain due to nerve damage and may also be depressed. Optimal therapy requires that each of these factors be looked for and treated.

TREATMENT

TREATMENT

Chronic Pain

Once the evaluation process has been completed and the likely causative and exacerbating factors identified, an explicit treatment plan should be developed. An important part of this process is to identify specific and realistic functional goals for therapy, such as getting a good night's sleep, being able to go shopping, or returning to work. A multidisciplinary approach that uses medications, counseling, physical therapy, nerve blocks, and even surgery may be required to improve the patient's quality of life. There are also some newer, minimally invasive procedures that can be helpful for some patients with intractable pain. These include image-guided interventions such as epidural injection of glucocorticoids for acute radicular pain and radiofrequency treatment of the facet joints for chronic facet-related back and neck pain. For patients with severe and persistent pain that is unresponsive to more conservative treatment, placement of electrodes within the spinal canal overlying the dorsal columns of the spinal cord (spinal cord stimulation) or implantation of intrathecal drug-delivery systems has shown significant benefit. The criteria for predicting which patients will respond to these procedures continue to evolve. They are generally reserved for patients who have not responded to conventional pharmacologic approaches. Referral to a multidisciplinary pain clinic for a full evaluation should precede any invasive procedure. Such referrals are clearly not necessary for all chronic pain patients. For some, pharmacologic management alone can provide adequate relief.

ANTIDEPRESSANT MEDICATIONS

The tricyclic antidepressants (TCAs), particularly nortriptyline and desipramine (Table 10-1), are useful for the management of chronic pain. Although developed for the treatment of depression, the TCAs have a spectrum of dose-related biologic activities that include analgesia in a variety of chronic clinical conditions. Although the mechanism is unknown, the analgesic effect of TCAs has a more rapid onset and occurs at a lower dose than is typically required for the treatment of depression. Furthermore, patients with chronic pain who are not depressed obtain pain relief with antidepressants. There is evidence that TCAs potentiate opioid analgesia, so they may be useful adjuncts for the treatment of severe persistent pain such as occurs with malignant tumors. **Table 10-2** lists some of the painful conditions that respond to TCAs. TCAs are of particular value in the management of neuropathic pain such as occurs in diabetic neuropathy and postherpetic neuralgia, for which there are few other therapeutic options.

The TCAs that have been shown to relieve pain have significant side effects (Table 10-1; **Chap. 444**). Some of these side effects, such as orthostatic hypotension, drowsiness, cardiac conduction delay, memory impairment, constipation, and urinary retention, are particularly problematic in elderly patients, and several are additive to the side effects of opioid analgesics. The selective serotonin reuptake inhibitors such as fluoxetine (Prozac) have fewer and less serious side effects than TCAs, but they are much less effective for relieving pain. It is of interest that venlafaxine (Effexor) and duloxetine (Cymbalta), which are nontricyclic antidepressants that block both

serotonin and norepinephrine reuptake, appear to retain most of the pain-relieving effect of TCAs with a side effect profile more like that of the selective serotonin reuptake inhibitors. These drugs may be particularly useful in patients who cannot tolerate the side effects of TCAs.

ANTICONVULSANTS AND ANTIARRHYTHMICS

These drugs are useful primarily for patients with neuropathic pain. Phenytoin (Dilantin) and carbamazepine (Tegretol) were first shown to relieve the pain of trigeminal neuralgia **(Chap. 433)**. This pain has a characteristic brief, shooting, electric shock-like quality. In fact, anticonvulsants seem to be particularly helpful for pains that have such a lancinating quality. Newer anticonvulsants, the calcium channel alpha-2-delta subunit ligands gabapentin (Neurontin) and pregabalin (Lyrica), are effective for a broad range of neuropathic pains. Furthermore, because of their favorable side effect profile, these newer anticonvulsants are often used as first-line agents.

CHRONIC OPIOID MEDICATION

The long-term use of opioids is accepted for patients with pain due to malignant disease. Although opioid use for chronic pain of nonmalignant origin is controversial, it is clear that, for many patients, opioids are the only option that produces meaningful pain relief. This is understandable because opioids are the most potent and have the broadest range of efficacy of any analgesic medications. Although addiction is rare in patients who first use opioids for pain relief, some degree of tolerance and physical dependence is likely with long-term use. Furthermore, studies suggest that long-term opioid therapy may worsen pain in some individuals, termed *opioid-induced hyperalgesia*. Therefore, before embarking on opioid therapy, other options should be explored, and the limitations and risks of opioids should be explained to the patient. It is also important to point out that some opioid analgesic medications have mixed agonist-antagonist properties (e.g., butorphanol and buprenorphine). From a practical standpoint, this means that they may worsen pain by inducing an abstinence syndrome in patients who are physically dependent on other opioid analgesics.

With long-term outpatient use of orally administered opioids, it may be desirable to use long-acting compounds such as levorphanol, methadone, sustained-release morphine, or transdermal fentanyl (Table 10-1). The pharmacokinetic profiles of these drug preparations enable the maintenance of sustained analgesic blood levels, potentially minimizing side effects such as sedation that are associated with high peak plasma levels, and reducing the likelihood of rebound pain associated with a rapid fall in plasma opioid concentration. Although long-acting opioid preparations may provide superior pain relief in patients with a continuous pattern of ongoing pain, others suffer from intermittent severe episodic pain and experience superior pain control and fewer side effects with the periodic use of short-acting opioid analgesics. Constipation is a virtually universal side effect of opioid use and should be treated expectantly. As noted above in the discussion of acute pain treatment, a recent advance for patients is the development of peripherally acting opioid antagonists that can reverse the constipation associated with opioid use without interfering with analgesia.

Soon after the introduction of a controlled-release oxycodone formulation (OxyContin) in the late 1990s, a dramatic rise in emergency department visits and deaths associated with oxycodone ingestion appeared, focusing public attention on misuse of prescription pain medications. The magnitude of prescription opioid abuse has grown over the last decade, leading the Centers for Disease Control and Prevention to classify

prescription opioid analgesic abuse as an epidemic. This appears to be due in large part to individuals using a prescription drug nonmedically, most often an opioid analgesic. Drug-induced deaths have rapidly risen and are now the second leading cause of death in Americans, just behind motor vehicle fatalities. In 2011, the Office of National Drug Control Policy established a multifaceted approach to address prescription drug abuse, including Prescription Drug Monitoring Programs (PDMPs) that allow practitioners to determine if patients are receiving prescriptions from multiple providers and use of law enforcement to eliminate improper prescribing practices. In 2016, the Centers for Disease Control (CDC) released the *CDC Guideline for Prescribing Opioids for Chronic Pain*, with recommendations for primary care clinicians who are prescribing opioids for chronic noncancer. The guideline is based on the best available scientific evidence and addresses (1) when to initiate or continue opioids for chronic pain; (2) opioid selection, dosage, duration, follow-up, and discontinuation; and (3) assessing risk and addressing harms of opioid use. The recent increase in scrutiny leaves many practitioners hesitant to prescribe opioid analgesics, other than for brief periods to control pain associated with illness or injury. For now, the choice to begin chronic opioid therapy for a given patient is left to the individual practitioner. Pragmatic guidelines for properly selecting and monitoring patients receiving chronic opioid therapy are shown in **Table 10-3;** a checklist for primary care clinicians prescribing opioids for noncancer pain is shown in **Table 10-4**.

TREATMENT OF NEUROPATHIC PAIN

It is important to individualize treatment for patients with neuropathic pain. Several general principles should guide therapy: the first is to move quickly to provide relief and the second is to minimize drug side effects. For example, in patients with postherpetic neuralgia and significant cutaneous hypersensitivity, topical lidocaine (Lidoderm patches) can provide immediate relief without side effects. Anticonvulsants (gabapentin or pregabalin; see above) or antidepressants (nortriptyline, desipramine, duloxetine, or venlafaxine) can be used as first-line drugs for patients with neuropathic pain. Systemically administered antiarrhythmic drugs such as lidocaine and mexiletine are less likely to be effective; although intravenous infusion of lidocaine can provide analgesia for patients with different types of neuropathic pain, the relief is usually transient, typically lasting just hours after the cessation of the infusion. The oral lidocaine congener mexiletine is poorly tolerated, producing frequent gastrointestinal adverse effects. There is no consensus on which class of drug should be used as a first-line treatment for any chronically painful condition. However, because relatively high doses of anticonvulsants are required for pain relief, sedation is very common. Sedation is also a problem with TCAs but is much less of a problem with serotonin/norepinephrine reuptake inhibitors (SNRIs; e.g., venlafaxine and duloxetine). Thus, in the elderly or in patients whose daily activities require high-level mental activity, these drugs should be considered the first line. In contrast, opioid medications should be used as a second- or third-line drug class. Although highly effective for many painful conditions, opioids are sedating, and their effect tends to lessen over time, leading to dose escalation and, occasionally, a worsening of pain. Drugs of different classes can be used in combination to optimize pain control. Repeated injection of botulinum toxin is an emerging approach that is showing some promise in treating focal neuropathic pain, particularly post-herpetic, trigeminal, and post-traumatic neuralgias.

It is worth emphasizing that many patients, especially those with chronic pain, seek medical attention primarily because they are suffering and because only physicians can provide the medications required for pain relief. A primary responsibility of all physicians is to minimize the physical and emotional discomfort of their patients.

Familiarity with pain mechanisms and analgesic medications is an important step toward accomplishing this aim.

TABLE 10-2

Painful Conditions That Respond to Tricyclic Antidepressants

Postherpetic neuralgia[a]
Diabetic neuropathy[a]
Fibromyalgia[a]
Tension headache[a]
Migraine headache[a]
Rheumatoid arthritis[a,b]
Chronic low back pain[b]
Cancer
Central poststroke pain

[a]Controlled trials demonstrate analgesia. [b]Controlled studies indicate benefit but not analgesia.

TABLE 10-3

Guidelines for Selecting and Monitoring Patients Receiving Chronic Opioid Therapy (COT) for the Treatment of Chronic, Noncancer Pain

Patient Selection
• Conduct a history, physical examination, and appropriate testing, including an assessment of risk of substance abuse, misuse, or addiction. • Consider a trial of COT if pain is moderate or severe, pain is having an adverse impact on function or quality of life, and potential therapeutic benefits outweigh potential harms. • A benefit-to-harm evaluation, including a history, physical examination, and appropriate diagnostic testing, should be performed and documented before and on an ongoing basis during COT.
Informed Consent and Use of Management Plans
• Informed consent should be obtained. A continuing discussion with the patient regarding COT should include goals, expectations, potential risks, and alternatives to COT. • Consider using a written COT management plan to document patient and clinician responsibilities and expectations and assist in patient education.
Initiation and Titration
• Initial treatment with opioids should be considered as a therapeutic trial to determine whether COT is appropriate. • Opioid selection, initial dosing, and titration should be individualized according to the patient's health status, previous exposure to opioids, attainment of therapeutic goals, and predicted or observed harms.
Monitoring
• Reassess patients on COT periodically and as warranted by changing circumstances. Monitoring should include documentation of pain intensity and level of functioning, assessments of progress toward achieving therapeutic goals, presence of adverse events, and adherence to prescribed therapies. • In patients on COT who are at high risk or who have engaged in aberrant drug-related behaviors, clinicians should periodically obtain urine drug screens or other information to confirm adherence to the COT plan of care. • In patients on COT not at high risk and not known to have engaged in aberrant drug-related behaviors, clinicians should consider periodically obtaining urine drug screens or other information to confirm adherence to the COT plan of care.

Source: Adapted with permission from R Chou et al: J Pain 10:113, 2009.

TABLE 10-4

Centers for Disease Control Checklist for Prescribing Opioids for Chronic Pain

For Primary Care Providers Treating Adults (18+) with Chronic Pain ≥3 months, Excluding Cancer, Palliative, and End-of-Life Care
CHECKLIST
WHEN CONSIDERING LONG-TERM OPIOID THERAPY • Set realistic goals for pain and function based on diagnosis (e.g., walk around the block). • Check that nonopioid therapies tried and optimized. • Discuss benefits and risks (e.g., addiction, overdose) with patient. • Evaluate risk of harm or misuse. ◦ Discuss risk factors with patient. ◦ Check prescription drug monitoring program (PDMP) data. ◦ Check urine drug screen. • Set criteria for stopping or continuing opioids. • Assess baseline pain and function (e.g., Pain, Enjoyment, General Activity [PEG] scale). • Schedule initial reassessment within 1–4 weeks. • Prescribe short-acting opioids using lowest dosage on product labeling; match duration to scheduled reassessment.
IF RENEWING WITHOUT A PATIENT VISIT • Check that return visit is scheduled ≤3 months from last visit.
WHEN REASSESSING AT A PATIENT VISIT • Continue opioids only after confirming clinically meaningful improvements in pain and function without significant risks or harm. • Assess pain and function (e.g., PEG); compare results to baseline. • Evaluate risk of harm or misuse: ◦ Observe patient for signs of oversedation or overdose risk. If yes: Taper dose. ◦ Check PDMP. ◦ Check for opioid use disorder if indicated (e.g., difficulty controlling use). If yes: Refer for treatment. • Check that nonopioid therapies optimized. Determine whether to continue, adjust, taper, or stop opioids. • Calculate opioid dosage morphine milligram equivalent (MME). ◦ If ≥50 MME/day total (≥50 mg hydrocodone; ≥33 mg oxycodone), increase frequency of follow-up; consider offering naloxone. ◦ Avoid ≥90 MME/day total (≥90 mg hydrocodone; ≥60 mg oxycodone), or carefully justify; consider specialist referral. • Schedule reassessment at regular intervals (≤3 months).

Source: Centers for Disease Control, Available at: https://stacks.cdc.gov/view/cdc/38025, accessed May 25, 2017 (Public Domain).

FURTHER READING

Dowell D et al: CDC guideline for prescribing opioids for chronic pain—United States, 2016. JAMA 315:1624, 2016. [PubMed: 26977696]

Finnerup NB et al: Pharmacotherapy for neuropathic pain in adults: A systematic review and meta-analysis. Lancet Neurol 14:162, 2015. [PubMed: 25575710]

Sun EC et al: Incidence of and risk factors for chronic opioid use among opioid-naive patients in the postoperative period. JAMA Intern Med 176:1286, 2016. [PubMed: 27400458]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 11: Chest Discomfort

David A. Morrow

INTRODUCTION

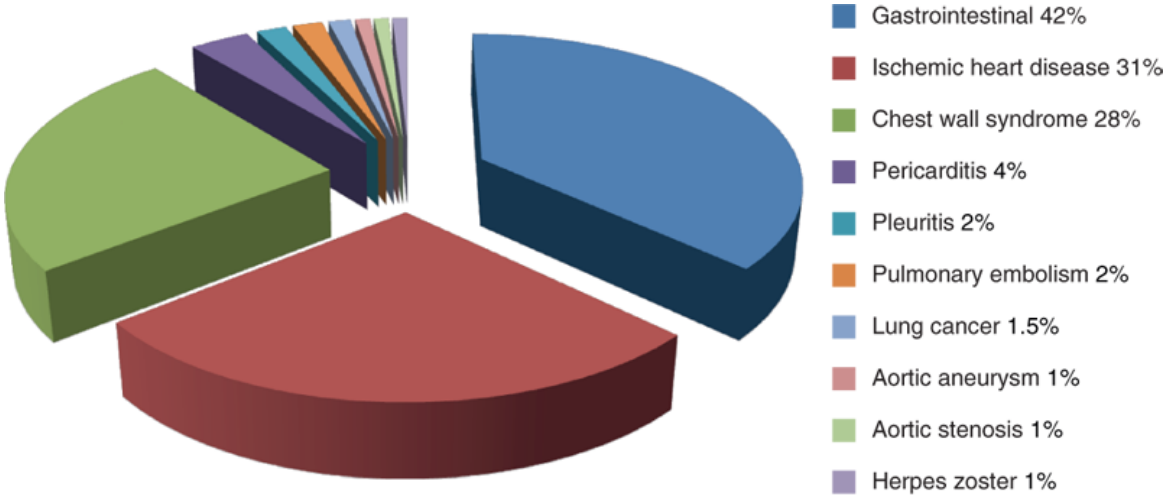
Chest discomfort is among the most common reasons for which patients present for medical attention at either an emergency department (ED) or an outpatient clinic. The evaluation of nontraumatic chest discomfort is inherently challenging owing to the broad variety of possible causes, a minority of which are life-threatening conditions that should not be missed. It is helpful to frame the initial diagnostic assessment and triage of patients with acute chest discomfort around three categories: (1) myocardial ischemia; (2) other cardiopulmonary causes (pericardial disease, aortic emergencies, and pulmonary conditions); and (3) non-cardiopulmonary causes. Although rapid identification of high-risk conditions is a priority of the initial assessment, strategies that incorporate routine liberal use of testing carry the potential for adverse effects of unnecessary investigations.

EPIDEMIOLOGY AND NATURAL HISTORY

Chest discomfort is the third most common reason for visits to the ED in the United States, resulting in 6 to 7 million emergency visits each year. More than 60% of patients with this presentation are hospitalized for further testing, and the remainder undergo additional investigation in the ED. As few as 15% of evaluated patients are eventually diagnosed with acute coronary syndrome (ACS), with rates of 10–20% in most series of unselected populations, and a rate as low as 5% in some studies. The most common diagnoses are gastrointestinal causes **(Fig. 11-1)**, and fewer than 10% are other life-threatening cardiopulmonary conditions. In a large proportion of patients with transient acute chest discomfort, ACS or another acute cardiopulmonary cause is excluded but the cause is not determined. Therefore, the resources and time devoted to the evaluation of chest discomfort *in the absence of a severe cause* are substantial. Nevertheless, a disconcerting 2% to 6% of patients with chest discomfort of presumed non-ischemic etiology who are discharged from the ED are later deemed to have had a missed myocardial infarction (MI). Patients with a missed diagnosis of MI have a 30-day risk of death that is double that of their counterparts who are hospitalized.

FIGURE 11-1

Distribution of final discharge diagnoses in patients with nontraumatic acute chest pain. *(Figure prepared from data in P Fruergaard et al: Eur Heart J 17:1028, 1996.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The natural histories of ACS, acute pericardial diseases, pulmonary embolism, and aortic emergencies are discussed in **Chaps. 265**, **268**, **269**, **273**, **and 274**, respectively. In a study of >350,000 patients with unspecified presumed non-cardiopulmonary chest discomfort, the mortality rate 1 year after discharge was <2% and did not differ significantly from age-adjusted mortality in the general population. The estimated rate of major cardiovascular events through 30 days in patients with acute chest pain who had been stratified as low risk was 2.5% in a large population-based study that excluded patients with ST-segment elevation or definite noncardiac chest pain.

CAUSES OF CHEST DISCOMFORT

The major etiologies of chest discomfort are discussed in this section and summarized in **Table 11-1**. Additional elements of the history, physical examination, and diagnostic testing that aid in distinguishing these causes are discussed in a later section (see "Approach to the Patient").

TABLE 11-1

Typical Clinical Features of Major Causes of Acute Chest Discomfort

System	Condition	Onset/Duration	Quality	Location	Associated Features
Cardiopulmonary					
Cardiac	Myocardial ischemia	*Stable angina:* Precipitated by exertion, cold, or stress; 2–10 min *Unstable angina:* Increasing pattern or at rest *Myocardial infarction:* Usually >30 min	Pressure, tightness, squeezing, heaviness, burning	Retrosternal; often radiation to neck, jaw, shoulders, or arms; sometimes epigastric	S_4 gallop or mitral regurgitation murmur (rare) during pain; S_3 or rales if severe ischemia or complication of myocardial infarction
	Pericarditis	Variable; hours to days; may be episodic	Pleuritic, sharp	Retrosternal or toward cardiac apex; may radiate to left shoulder	May be relieved by sitting up and leaning forward; pericardial friction rub
Vascular	Acute aortic syndrome	Sudden onset of unrelenting pain	Tearing or ripping; knifelike	Anterior chest, often radiating to back, between shoulder blades	Associated with hypertension and/or underlying connective tissue disorder; murmur of aortic insufficiency; loss of peripheral pulses
	Pulmonary embolism	Sudden onset	Pleuritic; may manifest as heaviness with massive pulmonary embolism	Often lateral, on the side of the embolism	Dyspnea, tachypnea, tachycardia, and hypotension
	Pulmonary hypertension	Variable; often exertional	Pressure	Substernal	Dyspnea, signs of increased venous pressure
Pulmonary	Pneumonia or pleuritis	Variable	Pleuritic	Unilateral, often localized	Dyspnea, cough, fever, rales, occasional rub
	Spontaneous pneumothorax	Sudden onset	Pleuritic	Lateral to side of pneumothorax	Dyspnea, decreased breath sounds on side of pneumothorax
Non-cardiopulmonary					
Gastrointenstinal	Esophageal reflux	10–60 min	Burning	Substernal, epigastric	Worsened by postprandial recumbency; relieved by antacids
	Esophageal spasm	2–30 min	Pressure, tightness, burning	Retrosternal	Can closely mimic angina
	Peptic ulcer	Prolonged; 60–90 min after meals	Burning	Epigastric, substernal	Relieved with food or antacids

System	Condition	Onset/Duration	Quality	Location	Associated Features
Cardiopulmonary					
	Gallbladder disease	Prolonged	Aching or colicky	Epigastric, right upper quadrant; sometimes to the back	May follow meal
Neuromuscular	Costochondritis	Variable	Aching	Sternal	Sometimes swollen, tender, warm over joint; may be reproduced by localized pressure on examination
	Cervical disk disease	Variable; may be sudden	Aching; may include numbness	Arms and shoulders	May be exacerbated by movement of neck
	Trauma or strain	Usually constant	Aching	Localized to area of strain	Reproduced by movement or palpation
	Herpes zoster	Usually prolonged	Sharp or burning	Dermatomal distribution	Vesicular rash in area of discomfort
Psychological	Emotional and psychiatric conditions	Variable; may be fleeting or prolonged	Variable; often manifests as tightness and dyspnea with feeling of panic or doom	Variable; may be retrosternal	Situational factors may precipitate symptoms; history of panic attacks, depression

MYOCARDIAL ISCHEMIA/INJURY

Myocardial ischemia causing chest discomfort, termed *angina pectoris*, is a primary clinical concern in patients presenting with chest symptoms. Myocardial ischemia is precipitated by an imbalance between myocardial oxygen requirements and myocardial oxygen supply, resulting in insufficient delivery of oxygen to meet the heart's metabolic demands. Myocardial oxygen consumption may be elevated by increases in heart rate, ventricular wall stress, and myocardial contractility, whereas myocardial oxygen supply is determined by coronary blood flow and coronary arterial oxygen content. When myocardial ischemia is sufficiently severe and prolonged in duration (as little as 20 min), irreversible cellular injury occurs, resulting in MI.

Ischemic heart disease is most commonly caused by atheromatous plaque that obstructs one or more of the epicardial coronary arteries. Stable ischemic heart disease **(Chap. 267)** usually results from the gradual atherosclerotic narrowing of the coronary arteries. *Stable angina* is characterized by ischemic episodes that are typically precipitated by a superimposed increase in oxygen demand during physical exertion and relieved upon resting. Ischemic heart disease becomes unstable most commonly when rupture or erosion of one or more atherosclerotic lesions triggers coronary thrombosis. Unstable ischemic heart disease is classified clinically by the presence or absence of detectable myocardial injury and the presence or absence of ST-segment elevation on the patient's electrocardiogram (ECG). When acute coronary atherothrombosis occurs, the intracoronary thrombus may be partially obstructive, generally leading to myocardial ischemia in the absence of ST-segment elevation. Marked by ischemic symptoms at rest, with minimal activity, or in an accelerating pattern, unstable ischemic heart disease is classified as *unstable angina* when there is no detectable myocardial injury and as *non–ST elevation MI* (NSTEMI) when there is evidence of myocardial necrosis **(Chap. 268)**. When the coronary thrombus is acutely and completely occlusive, transmural myocardial ischemia usually ensues, with ST-segment elevation on the ECG and myocardial necrosis leading to a diagnosis of *ST elevation MI* (STEMI, see **Chap. 269**).

Clinicians should be aware that unstable ischemic symptoms may also occur predominantly because of increased myocardial oxygen demand (e.g., during intense psychological stress or fever) or because of decreased oxygen delivery due to anemia, hypoxia, or hypotension. However, the term *acute coronary syndrome*, which encompasses unstable angina, NSTEMI, and STEMI, is in general reserved for ischemia precipitated by acute coronary atherothrombosis. In order to guide therapeutic strategies, a standardized system for classification of MI has been expanded to discriminate MI resulting from acute coronary thrombosis (type 1) from MI occurring secondary to other imbalances of myocardial oxygen supply and demand (type 2; see **Chap. 268**).

Other contributors to stable and unstable ischemic heart disease, such as endothelial dysfunction, microvascular disease, and vasospasm, may exist alone or in combination with coronary atherosclerosis and may be the dominant cause of myocardial ischemia in some patients. Moreover, non-atherosclerotic processes, including congenital abnormalities of the coronary vessels, myocardial bridging, coronary arteritis, and radiation-induced

coronary disease, can lead to coronary obstruction. In addition, conditions associated with extreme myocardial oxygen demand and impaired endocardial blood flow, such as aortic valve disease **(Chap. 274)**, hypertrophic cardiomyopathy, or idiopathic dilated cardiomyopathy **(Chap. 254)**, can precipitate myocardial ischemia in patients with or without underlying obstructive atherosclerosis.

Characteristics of Ischemic Chest Discomfort

The clinical characteristics of angina pectoris, often referred to simply as "angina," are highly similar whether the ischemic discomfort is a manifestation of stable ischemic heart disease, unstable angina, or MI; the exceptions are differences in the pattern and duration of symptoms associated with these syndromes (Table 11-1). Heberden initially described angina as a sense of "strangling and anxiety." Chest discomfort characteristic of myocardial ischemia is typically described as aching, heavy, squeezing, crushing, or constricting. However, in a substantial minority of patients, the quality of discomfort is extremely vague and may be described as a mild tightness, or merely an uncomfortable feeling, that sometimes is experienced as numbness or a burning sensation. The site of the discomfort is usually retrosternal, but radiation is common and generally occurs down the ulnar surface of the left arm; the right arm, both arms, neck, jaw, or shoulders may also be involved. These and other characteristics of ischemic chest discomfort pertinent to discrimination from other causes of chest pain are discussed later in this chapter (see "Approach to the Patient").

Stable angina usually begins gradually and reaches its maximal intensity over a period of minutes before dissipating within several minutes with rest or with nitroglycerin. The discomfort typically occurs predictably at a characteristic level of exertion or psychological stress. By definition, unstable angina is manifest by anginal chest discomfort that occurs with progressively lower intensity of physical activity or even at rest. Chest discomfort associated with MI is typically more severe, is prolonged (usually lasting ≥30 min), and is not relieved by rest.

Mechanisms of Cardiac Pain

The neural pathways involved in ischemic cardiac pain are poorly understood. Ischemic episodes are thought to excite local chemosensitive and mechanoreceptive receptors that, in turn, stimulate release of adenosine, bradykinin, and other substances that activate the sensory ends of sympathetic and vagal afferent fibers. The afferent fibers traverse the nerves that connect to the upper five thoracic sympathetic ganglia and upper five distal thoracic roots of the spinal cord. From there, impulses are transmitted to the thalamus. Within the spinal cord, cardiac sympathetic afferent impulses may converge with impulses from somatic thoracic structures, and this convergence may be the basis for referred cardiac pain. In addition, cardiac vagal afferent fibers synapse in the nucleus tractus solitarius of the medulla and then descend to the upper cervical spinothalamic tract, and this route may contribute to anginal pain experienced in the neck and jaw.

OTHER CARDIOPULMONARY CAUSES

Pericardial and Other Myocardial Diseases

Inflammation of the pericardium due to infectious or noninfectious causes can be responsible for acute or chronic chest discomfort **(See also Chap. 265)**. The visceral surface and most of the parietal surface of the pericardium are insensitive to pain. Therefore, the pain of pericarditis is thought to arise principally from associated pleural inflammation. Because of this pleural association, the discomfort of pericarditis is usually pleuritic pain that is exacerbated by breathing, coughing, or changes in position. Moreover, owing to the overlapping sensory supply of the central diaphragm via the phrenic nerve with somatic sensory fibers originating in the third to fifth cervical segments, the pain of pleural pericarditis is often referred to the shoulder and neck. Involvement of the pleural surface of the lateral diaphragm can lead to pain in the upper abdomen.

Acute inflammatory and other non-ischemic myocardial diseases can also produce chest discomfort. The symptoms of *Takotsubo (stress-related) cardiomyopathy* often start abruptly with chest pain and shortness of breath. This form of cardiomyopathy, in its most recognizable form, is triggered by an emotionally or physically stressful event and may mimic acute MI because of its commonly associated ECG abnormalities, including ST-segment elevation, and elevated biomarkers of myocardial injury. Observational studies support a predilection for women >50 years of age. The symptoms of acute myocarditis are highly varied. Chest discomfort may either originate with inflammatory injury of the myocardium or be due to severe increases in wall stress related to poor ventricular performance.

Diseases of the Aorta

Acute aortic dissection (Fig. 11-1) is a less common cause of chest discomfort but is important because of the catastrophic natural history of certain subsets of cases when recognized late or left untreated **(See also Chap. 274)**. Acute aortic syndromes encompass a spectrum of acute aortic diseases related to disruption of the media of the aortic wall. *Aortic dissection* involves a tear in the aortic intima, resulting in separation of the media and creation of a separate "false" lumen. A *penetrating ulcer* has been described as ulceration of an aortic atheromatous plaque that extends through the intima and into the aortic media, with the potential to initiate an intramedial dissection or rupture into the adventitia. *Intramural hematoma* is an aortic wall hematoma with no demonstrable intimal flap, no radiologically apparent intimal tear, and no false lumen. Intramural hematoma can occur due to either rupture of the vasa vasorum or, less commonly, a penetrating ulcer.

Each of these subtypes of acute aortic syndrome typically presents with chest discomfort that is often severe, sudden in onset, and sometimes described as “tearing” in quality. Acute aortic syndromes involving the *ascending* aorta tend to cause pain in the midline of the anterior chest, whereas *descending* aortic syndromes most often present with pain in the back. Therefore, dissections that begin in the ascending aorta and extend to the descending aorta tend to cause pain in the front of the chest that extends toward the back, between the shoulder blades. Proximal aortic dissections that involve the ascending aorta (type A in the Stanford nomenclature) are at high risk for major complications that may influence the clinical presentation, including (1) compromise of the aortic ostia of the coronary arteries, resulting in MI; (2) disruption of the aortic valve, causing acute aortic insufficiency; and (3) rupture of the hematoma into the pericardial space, leading to pericardial tamponade.

Knowledge of the epidemiology of acute aortic syndromes can be helpful in maintaining awareness of this relatively uncommon group of disorders (estimated annual incidence, 3 cases per 100,000 population). Nontraumatic aortic dissections are very rare in the absence of hypertension or conditions associated with deterioration of the elastic or muscular components of the aortic media, including pregnancy, bicuspid aortic disease, or inherited connective tissue diseases, such as Marfan and Ehlers-Danlos syndromes.

Although aortic aneurysms are most often asymptomatic, thoracic aortic aneurysms can cause chest pain and other symptoms by compressing adjacent structures. This pain tends to be steady, deep, and occasionally severe. Aortitis, whether of noninfectious or infectious etiology, in the absence of aortic dissection is a rare cause of chest or back discomfort.

Pulmonary Conditions

Pulmonary and pulmonary-vascular conditions that cause chest discomfort usually do so in conjunction with dyspnea and often produce symptoms that have a pleuritic nature.

PULMONARY EMBOLISM

Pulmonary emboli (annual incidence, ~1 per 1000) can produce dyspnea and chest discomfort that is sudden in onset **(See also Chap. 273)**. Typically pleuritic in pattern, the chest discomfort associated with pulmonary embolism may result from (1) involvement of the pleural surface of the lung adjacent to a resultant pulmonary infarction; (2) distention of the pulmonary artery; or (3) possibly, right ventricular wall stress and/or subendocardial ischemia related to acute pulmonary hypertension. The pain associated with small pulmonary emboli is often lateral and pleuritic and is believed to be related to the first of these three possible mechanisms. In contrast, massive pulmonary emboli may cause severe substernal pain that may mimic an MI and that is plausibly attributed to the second and third of these potential mechanisms. Massive or submassive pulmonary embolism may also be associated with syncope, hypotension, and signs of right heart failure. Other typical characteristics that aid in the recognition of pulmonary embolism are discussed later in this chapter (see “Approach to the Patient”).

PNEUMOTHORAX

Primary spontaneous pneumothorax is a rare cause of chest discomfort, with an estimated annual incidence in the United States of 7 per 100,000 among men and <2 per 100,000 among women **(See also Chap. 289)**. Risk factors include male sex, smoking, family history, and Marfan syndrome. The symptoms are usually sudden in onset, and dyspnea may be mild; thus, presentation to medical attention is sometimes delayed. *Secondary spontaneous pneumothorax* may occur in patients with underlying lung disorders, such as chronic obstructive pulmonary disease, asthma, or cystic fibrosis, and usually produces symptoms that are more severe. Tension pneumothorax is a medical emergency caused by trapped intrathoracic air that precipitates hemodynamic collapse.

Other Pulmonary Parenchymal, Pleural, or Vascular Disease

Most pulmonary diseases that produce chest pain, including pneumonia and malignancy, do so because of involvement of the pleura or surrounding structures **(See also Chaps. 277, 278, and 288)**. Pleurisy is typically described as a knifelike pain that is worsened by inspiration or coughing. In contrast, chronic pulmonary hypertension can manifest as chest pain that may be very similar to angina in its characteristics, suggesting right ventricular myocardial ischemia in some cases. Reactive airways diseases similarly can cause chest tightness associated with breathlessness rather than pleurisy.

NON-CARDIOPULMONARY CAUSES

Gastrointestinal Conditions

Gastrointestinal disorders are the most common cause of nontraumatic chest discomfort and often produce symptoms that are difficult to discern from more serious causes of chest pain, including myocardial ischemia **(See also Chap. 314)**. Esophageal disorders, in particular, may simulate angina in the character and location of the pain. Gastroesophageal reflux and disorders of esophageal motility are common and should be considered in the differential diagnosis of chest pain (Fig. 11-1 and Table 11-1). Acid reflux often causes a burning discomfort. The pain of esophageal spasm, in contrast, is commonly an intense, squeezing discomfort that is retrosternal in location and, like angina, may be relieved by nitroglycerin or dihydropyridine calcium channel antagonists. Chest pain can also result from injury to the esophagus, such as a Mallory-Weiss tear or even an esophageal rupture (Boerhaave syndrome) caused by severe vomiting. Peptic ulcer disease is most commonly epigastric in location but can radiate into the chest (Table 11-1).

Hepatobiliary disorders, including cholecystitis and biliary colic, may mimic acute cardiopulmonary diseases. Although the pain arising from these disorders usually localizes to the right upper quadrant of the abdomen, it is variable and may be felt in the epigastrium and radiate to the back and lower chest. This discomfort is sometimes referred to the scapula or may in rare cases be felt in the shoulder, suggesting diaphragmatic irritation. The pain is steady, usually lasts several hours, and subsides spontaneously, without symptoms between attacks. Pain resulting from pancreatitis is typically aching epigastric pain that radiates to the back.

Musculoskeletal and Other Causes

Chest discomfort can be produced by any musculoskeletal disorder involving the chest wall or the nerves of the chest wall, neck, or upper limbs **(See also Chap. 363)**. Costochondritis causing tenderness of the costochondral junctions (*Tietze's syndrome*) is relatively common. Cervical radiculitis may manifest as a prolonged or constant aching discomfort in the upper chest and limbs. The pain may be exacerbated by motion of the neck. Occasionally, chest pain can be caused by compression of the brachial plexus by the cervical ribs, and tendinitis or bursitis involving the left shoulder may mimic the radiation of angina. Pain in a dermatomal distribution can also be caused by cramping of intercostal muscles or by herpes zoster **(Chap. 188)**.

Emotional and Psychiatric Conditions

As many as 10% of patients who present to EDs with acute chest discomfort have a panic disorder or related condition (Table 11-1). The symptoms may include chest tightness or aching that is associated with a sense of anxiety and difficulty in breathing. The symptoms may be prolonged or fleeting.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Chest Discomfort

Given the broad set of potential causes and the heterogeneous risk of serious complications in patients who present with acute nontraumatic chest discomfort, the priorities of the initial clinical encounter include assessment of (1) the patient's clinical stability and (2) the probability that the patient has an underlying cause of the discomfort that may be life-threatening. The high-risk conditions of principal concern are acute cardiopulmonary processes, including ACS, acute aortic syndrome, pulmonary embolism, tension pneumothorax, and pericarditis with tamponade. Among non-cardiopulmonary causes of chest pain, esophageal rupture likely holds the greatest urgency for diagnosis. Patients with these conditions may deteriorate rapidly despite initially appearing well. The remaining population with non-cardiopulmonary conditions has a more favorable prognosis during completion of the diagnostic work-up. A rapid targeted assessment for a serious cardiopulmonary cause is of particular relevance for patients with acute ongoing pain who have presented for emergency evaluation. Among patients presenting in the outpatient setting with chronic pain or pain that has resolved, a general diagnostic assessment is reasonably undertaken (see "Outpatient Evaluation of Chest Discomfort," below). A series of questions that can be used to structure the clinical evaluation of patients with chest discomfort is shown in **Table 11-2**.

HISTORY

The evaluation of nontraumatic chest discomfort relies heavily on the clinical history and physical examination to direct subsequent diagnostic testing. The evaluating clinician should assess the quality, location (including radiation), and pattern (including onset and duration) of the pain as well as any provoking or alleviating factors. The presence of associated symptoms may also be useful in establishing a diagnosis.

Quality of Pain

The quality of chest discomfort alone is never sufficient to establish a diagnosis. However, the characteristics of the pain are pivotal in formulating an initial clinical impression and assessing the likelihood of a serious cardiopulmonary process (Table 11-1), including ACS in particular **(Fig. 11-2)**. Pressure or tightness is consistent with a typical presentation of myocardial ischemic pain. Nevertheless, the clinician must remember that some patients with ischemic chest symptoms deny any "pain" but rather complain of dyspnea or a vague sense of anxiety. The severity of the discomfort has poor diagnostic accuracy. It is often helpful to ask about the similarity of the discomfort to previous definite ischemic symptoms. It is unusual for angina to be sharp, as in knifelike, stabbing, or pleuritic; however, patients sometimes use the word "sharp" to convey the intensity of discomfort rather than the quality. Pleuritic discomfort is suggestive of a process involving the pleura, including pericarditis, pulmonary embolism, or pulmonary parenchymal processes. Less frequently, the pain of pericarditis or massive pulmonary embolism is a steady severe pressure or aching that can be difficult to discriminate from myocardial ischemia. "Tearing" or "ripping" pain is often described by patients with acute aortic dissection. However, acute aortic emergencies also present commonly with severe, knifelike pain. A burning quality can suggest acid reflux or peptic ulcer disease but may also occur with myocardial ischemia. Esophageal pain, particularly with spasm, can be a severe squeezing discomfort identical to angina.

Location of Discomfort

A substernal location with radiation to the neck, jaw, shoulder, or arms is typical of myocardial ischemic discomfort. Radiation to both arms has a particularly high association with MI as the etiology. Some patients present with aching in sites of radiated pain as their only symptoms of ischemia. However, pain that is highly localized—for example, that which can be demarcated by the tip of one finger—is highly unusual for angina. A retrosternal location should prompt consideration of esophageal pain; however, other gastrointestinal conditions usually present with pain that is most intense in

the abdomen or epigastrium, with possible radiation into the chest. Angina may also occur in an epigastric location. However, pain that occurs solely above the mandible or below the epigastrium is rarely angina. Severe pain radiating to the back, particularly between the shoulder blades, should prompt consideration of an acute aortic syndrome. Radiation to the trapezius ridge is characteristic of pericardial pain and does not usually occur with angina.

Pattern

Myocardial ischemic discomfort usually builds over minutes and is exacerbated by activity and mitigated by rest. In contrast, pain that reaches its peak intensity immediately is more suggestive of aortic dissection, pulmonary embolism, or spontaneous pneumothorax. Pain that is fleeting (lasting only a few seconds) is rarely ischemic in origin. Similarly, pain that is constant in intensity for a prolonged period (many hours to days) is unlikely to represent myocardial ischemia if it occurs in the absence of other clinical consequences, such as abnormalities of the ECG, elevation of cardiac biomarkers, or clinical sequelae (e.g., heart failure or hypotension). Both myocardial ischemia and acid reflux may have their onset in the morning.

Provoking and Alleviating Factors

Patients with myocardial ischemic pain usually prefer to rest, sit, or stop walking. However, clinicians should be aware of the phenomenon of "warm-up angina" in which some patients experience relief from angina as they continue at the same or even a greater level of exertion **(Chap. 267)**. Alterations in the intensity of pain with changes in position or movement of the upper extremities and neck are less likely with myocardial ischemia and suggest a musculoskeletal etiology. The pain of pericarditis, however, often is worse in the supine position and relieved by sitting upright and leaning forward. Gastroesophageal reflux may be exacerbated by alcohol, some foods, or by a reclined position. Relief can occur with sitting.

Exacerbation by eating suggests a gastrointestinal etiology such as peptic ulcer disease, cholecystitis, or pancreatitis. Peptic ulcer disease tends to become symptomatic 60–90 min after meals. However, in the setting of severe coronary atherosclerosis, redistribution of blood flow to the splanchnic vasculature after eating can trigger postprandial angina. The discomfort of acid reflux and peptic ulcer disease is usually diminished promptly by acid-reducing therapies. In contrast with its impact in some patients with angina, physical exertion is very unlikely to alter symptoms from gastrointestinal causes of chest pain. Relief of chest discomfort within minutes after administration of nitroglycerin is suggestive of but not sufficiently sensitive or specific for a definitive diagnosis of myocardial ischemia. Esophageal spasm may also be relieved promptly with nitroglycerin. A delay of >10 min before relief is obtained after nitroglycerin suggests that the symptoms either are not caused by ischemia or are caused by severe ischemia, such as during acute MI.

Associated Symptoms

Symptoms that accompany myocardial ischemia may include diaphoresis, dyspnea, nausea, fatigue, faintness, and eructations. In addition, these symptoms may exist in isolation as anginal equivalents (i.e., symptoms of myocardial ischemia other than typical angina), particularly in women and the elderly. Dyspnea may occur with multiple conditions considered in the differential diagnosis of chest pain and thus is not discriminative, but the presence of dyspnea is important because it suggests a cardiopulmonary etiology. Sudden onset of significant respiratory distress should lead to consideration of pulmonary embolism and spontaneous pneumothorax. Hemoptysis may occur with pulmonary embolism, or as blood-tinged frothy sputum in severe heart failure but usually points toward a pulmonary parenchymal etiology of chest symptoms. Presentation with syncope or pre-syncope should prompt consideration of hemodynamically significant pulmonary embolism or aortic dissection as well as ischemic arrhythmias. Although nausea and vomiting suggest a gastrointestinal disorder, these symptoms may occur in the setting of MI (more commonly inferior MI), presumably because of activation of the vagal reflex or stimulation of left ventricular receptors as part of the Bezold-Jarisch reflex.

Past Medical History

The past medical history is useful in assessing the patient for risk factors for coronary atherosclerosis and venous thromboembolism **(Chap. 273)** as well as for conditions that may predispose the patient to specific disorders. For example, a history of connective tissue diseases such as Marfan syndrome should heighten the clinician's suspicion of an acute aortic syndrome or spontaneous pneumothorax. A careful history may elicit clues about depression or prior panic attacks.

PHYSICAL EXAMINATION

In addition to providing an initial assessment of the patient's clinical stability, the physical examination of patients with chest discomfort can provide direct evidence of specific etiologies of chest pain (e.g., unilateral absence of lung sounds) and can identify potential precipitants of acute cardiopulmonary causes of chest pain (e.g., uncontrolled hypertension), relevant comorbid conditions (e.g., obstructive pulmonary disease), and complications of the presenting syndrome (e.g., heart failure). However, because the findings on physical examination may be normal in patients with unstable ischemic heart disease, an unremarkable physical examination is not definitively reassuring.

General

The patient's general appearance is helpful in establishing an initial impression of the severity of illness. Patients with acute MI or other acute cardiopulmonary disorders often appear anxious, uncomfortable, pale, cyanotic, or diaphoretic. Patients who are massaging or clutching their chests may describe their pain with a clenched fist held against the sternum (*Levine's sign*). Occasionally, body habitus is helpful—for example, in patients with Marfan syndrome or the prototypical young, tall, thin man with spontaneous pneumothorax.

Vital Signs

Significant tachycardia and hypotension are indicative of important hemodynamic consequences of the underlying cause of chest discomfort and should prompt a rapid survey for the most severe conditions, such as acute MI with cardiogenic shock, massive pulmonary embolism, pericarditis with tamponade, or tension pneumothorax. Acute aortic emergencies usually present with severe hypertension but may be associated with profound hypotension when there is coronary arterial compromise or dissection into the pericardium. Sinus tachycardia is an important manifestation of submassive pulmonary embolism. Tachypnea and hypoxemia point toward a pulmonary cause. The presence of low-grade fever is nonspecific because it may occur with MI and with thromboembolism in addition to infection.

Pulmonary

Examination of the lungs may localize a primary pulmonary cause of chest discomfort, as in cases of pneumonia, asthma, or pneumothorax. Left ventricular dysfunction from severe ischemia/infarction as well as acute valvular complications of MI or aortic dissection can lead to pulmonary edema, which is an indicator of high risk.

Cardiac

The jugular venous pulse is often normal in patients with acute myocardial ischemia but may reveal characteristic patterns with pericardial tamponade or acute right ventricular dysfunction (**Chaps. 234** and **265**). Cardiac auscultation may reveal a third or, more commonly, a fourth heart sound, reflecting myocardial systolic or diastolic dysfunction. Murmurs of mitral regurgitation or a ventricular-septal defect may indicate mechanical complications of STEMI. A murmur of aortic insufficiency may be a complication of proximal aortic dissection. Other murmurs may reveal underlying cardiac disorders contributory to ischemia (e.g., aortic stenosis or hypertrophic cardiomyopathy). Pericardial friction rubs reflect pericardial inflammation.

Abdominal

Localizing tenderness on the abdominal examination is useful in identifying a gastrointestinal cause of the presenting syndrome. Abdominal findings are infrequent with purely acute cardiopulmonary problems, except in the case of underlying chronic cardiopulmonary disease or severe right ventricular dysfunction leading to hepatic congestion.

Vascular pulse deficits may reflect underlying chronic atherosclerosis, which increases the likelihood of coronary artery disease. However, evidence of acute limb ischemia with loss of the pulse and pallor, particularly in the upper extremities, can indicate catastrophic consequences of aortic dissection. Unilateral lower-extremity swelling should raise suspicion about venous thromboembolism.

Musculoskeletal

Pain arising from the costochondral and chondrosternal articulations may be associated with localized swelling, redness, or marked localized tenderness. Pain on palpation of these joints is usually well localized and is a useful clinical sign, though deep palpation may elicit pain in the absence of costochondritis. Although palpation of the chest wall often elicits pain in patients with various musculoskeletal conditions, it should be appreciated that chest wall tenderness does not exclude myocardial ischemia. Sensory deficits in the upper extremities may be indicative of cervical disk disease.

ELECTROCARDIOGRAPHY

Electrocardiography is crucial in the evaluation of nontraumatic chest discomfort. The ECG is pivotal for identifying patients with ongoing ischemia as the principal reason for their presentation as well as secondary cardiac complications of other disorders. Professional society guidelines recommend that an ECG be obtained within 10 min of presentation, with the primary goal of identifying patients with ST-segment elevation diagnostic of MI who are candidates for immediate interventions to restore flow in the occluded coronary artery. ST-segment depression and symmetric T-wave inversions at least 0.2 mV in depth are useful for detecting myocardial ischemia in the absence of STEMI and are also indicative of higher risk of death or recurrent ischemia. Serial performance of ECGs (every 30–60 min) is recommended in the ED evaluation of suspected ACS. In addition, an ECG with right-sided lead placement should be considered in patients with clinically suspected ischemia and a nondiagnostic standard 12-lead ECG. Despite the value of the resting ECG, its sensitivity for ischemia is poor—as low as 20% in some studies.

Abnormalities of the ST segment and T wave may occur in a variety of conditions, including pulmonary embolism, ventricular hypertrophy, acute and chronic pericarditis, myocarditis, electrolyte imbalance, and metabolic disorders. Notably, hyperventilation associated with panic disorder can also lead to nonspecific ST and T-wave abnormalities. Pulmonary embolism is most often associated with sinus tachycardia but can also lead to rightward shift of the ECG axis, manifesting as an S-wave in lead I, with a Q-wave and T-wave in lead III (**Chaps. 235** and **273**). In patients with ST-segment elevation, the presence of diffuse lead involvement not corresponding to a specific coronary anatomic distribution and PR-segment depression can aid in distinguishing pericarditis from acute MI.

CHEST RADIOGRAPHY

(**See Chap. A12**) Plain radiography of the chest is performed routinely when patients present with acute chest discomfort and selectively when individuals who are being evaluated as outpatients have subacute or chronic pain. The chest radiograph is most useful for identifying pulmonary processes, such as pneumonia or pneumothorax. Findings are often unremarkable in patients with ACS, but pulmonary edema may be evident. Other specific findings include widening of the mediastinum in some patients with aortic dissection, Hampton's hump or Westermark's sign in patients with pulmonary embolism (**Chaps. 273** and **A12**), or pericardial calcification in chronic pericarditis.

CARDIAC BIOMARKERS

Laboratory testing in patients with acute chest pain is focused on the detection of myocardial injury. Such injury can be detected by the presence of circulating proteins released from damaged myocardial cells. Owing to the time necessary for this release, initial biomarkers of injury may be in the normal range, even in patients with STEMI. Because of superior cardiac tissue-specificity compared with creatine kinase MB, cardiac troponin is the preferred biomarker for the diagnosis of MI and should be measured in all patients with suspected ACS at presentation and repeated in 3–6 h. Testing after 6 h is required only when there is uncertainty regarding the onset of pain or when stuttering symptoms have occurred. It is not necessary or advisable to measure troponin in patients without suspicion of ACS unless this test is being used specifically for risk stratification (e.g., in pulmonary embolism or heart failure).

The development of cardiac troponin assays with progressively greater analytical sensitivity has facilitated detection of substantially lower blood concentrations of troponin than was previously possible. This evolution permits earlier detection of myocardial injury, enhances the overall accuracy of a diagnosis of MI, and improves risk stratification in suspected ACS. The greater negative predictive value of a negative troponin result with current-generation assays is an advantage in the evaluation of chest pain in the ED. Rapid rule-out protocols that use serial testing and changes in troponin concentration over as short a period as 1–2 h appear promising and have been adopted in some centers where high-sensitivity assays for troponin are used routinely. In patients presenting >2 h after symptom onset, a concentration of cardiac troponin below the limit of detection using a high-sensitivity assay may be sufficient to exclude MI with a negative predictive value >99% at the time of hospital presentation. However, with these advantages has come a trade-off: myocardial injury is detected in a larger proportion of patients who have non-ACS cardiopulmonary conditions than with previous, less sensitive assays. This evolution in testing for myocardial necrosis has rendered other aspects of the clinical evaluation critical to the practitioner's determination of the probability that the symptoms represent ACS. In addition, observation of a change in cardiac troponin concentration between serial samples is useful in discriminating acute causes of myocardial injury from chronic elevation due to underlying structural heart disease, end-stage renal disease, or interfering antibodies. The diagnosis of MI is reserved for acute myocardial injury that is marked by a rising and/or falling pattern—with at least one value exceeding the 99th percentile reference limit—and that is caused by ischemia. Other non-ischemic insults, such as myocarditis, may result in myocardial injury but should not be labeled MI.

Other laboratory assessments may include the D-dimer test to aid in exclusion of pulmonary embolism **(Chap. 273)**. Measurement of a B-type natriuretic peptide is useful when considered in conjunction with the clinical history and examination for the diagnosis of heart failure. B-type natriuretic peptides also provide prognostic information among patients with ACS and those with pulmonary embolism.

INTEGRATIVE DECISION-AIDS

Multiple clinical algorithms have been developed to aid in decision-making during the evaluation and disposition of patients with acute nontraumatic chest pain. Such decision-aids estimate either of two closely related but not identical probabilities: (1) the probability of a final diagnosis of ACS and (2) the probability of major cardiac events during short-term follow-up. Such decision-aids are used most commonly to identify patients with a low clinical probability of ACS who are candidates either for early provocative testing for ischemia or for discharge from the ED. Goldman and Lee developed one of the first such decision-aids, using only the ECG and risk indicators—hypotension, pulmonary rales, and known ischemic heart disease—to categorize patients into four risk categories ranging from a <1% to a >16% probability of a major cardiovascular complication. The Acute Cardiac Ischemia Time-Insensitive Predictive Instrument (ACI-TIPI) combines age, sex, chest pain presence, and ST-segment abnormalities to define a probability of ACS. More recently developed decision-aids are shown in **Fig. 11-3**. Elements common to each of these tools are (1) symptoms typical for ACS; (2) older age; (3) risk factors for or known atherosclerosis; (4) ischemic ECG abnormalities; and (5) elevated cardiac troponin levels. Although, because of very low specificity, the overall diagnostic performance of such decision-aids is poor (area under the receiver operating curve, 0.55–0.65), they can help identify patients with a very low probability of ACS (e.g., <1%). Nevertheless, no such decision-aid (or single clinical factor) is sufficiently sensitive and well validated to use as a sole tool for clinical decision-making.

Clinicians should differentiate between the algorithms discussed above and risk scores derived for stratification of prognosis (e.g., the TIMI and GRACE risk scores, **Chap. 269**) in patients *who already have an established diagnosis of ACS*. The latter risk scores were not designed to be used for *diagnostic* assessment.

PROVOCATIVE TESTING FOR ISCHEMIA

Exercise electrocardiography ("stress testing") is commonly employed for completion of risk stratification of patients who have undergone an initial evaluation that has not revealed a specific cause of chest discomfort and has identified them as being at low or selectively intermediate risk of ACS. Early exercise testing is safe in patients without high-risk findings after 8–12 h of observation and can assist in refining their prognostic assessment. For example, of low-risk patients who underwent exercise testing in the first 48 h after presentation, those without evidence of ischemia had a 2% rate of cardiac events through 6 months, whereas the rate was 15% among patients with either clear evidence of ischemia or an equivocal result. Patients who are unable to exercise may undergo pharmacological stress testing with either nuclear perfusion imaging or echocardiography. Notably, some experts have deemed the routine use of stress testing for low-risk patients unsupported by direct clinical evidence and a potentially unnecessary source of cost.

Professional society guidelines identify ongoing chest pain as a contraindication to stress testing. In selected patients with persistent pain and nondiagnostic ECG and biomarker data, resting myocardial perfusion images can be obtained; the absence of any perfusion abnormality substantially reduces the likelihood of coronary artery disease. In some centers, early myocardial perfusion imaging is performed as part of a routine strategy for

evaluating patients at low or intermediate risk of ACS in parallel with other testing. Management of patients with normal perfusion images can be expedited with earlier discharge and outpatient stress testing, if indicated. Those with abnormal rest perfusion imaging, which cannot discriminate between old or new myocardial defects, usually warrant additional in-hospital evaluation.

OTHER NONINVASIVE STUDIES

Other noninvasive imaging studies of the chest can be used selectively to provide additional diagnostic and prognostic information on patients with chest discomfort.

Echocardiography

Echocardiography is not necessarily routine in patients with chest discomfort. However, in patients with an uncertain diagnosis, particularly those with nondiagnostic ST elevation, ongoing symptoms, or hemodynamic instability, detection of abnormal regional wall motion provides evidence of possible ischemic dysfunction. Echocardiography is diagnostic in patients with mechanical complications of MI or in patients with pericardial tamponade. Transthoracic echocardiography is poorly sensitive for aortic dissection, although an intimal flap may sometimes be detected in the ascending aorta.

CT Angiography

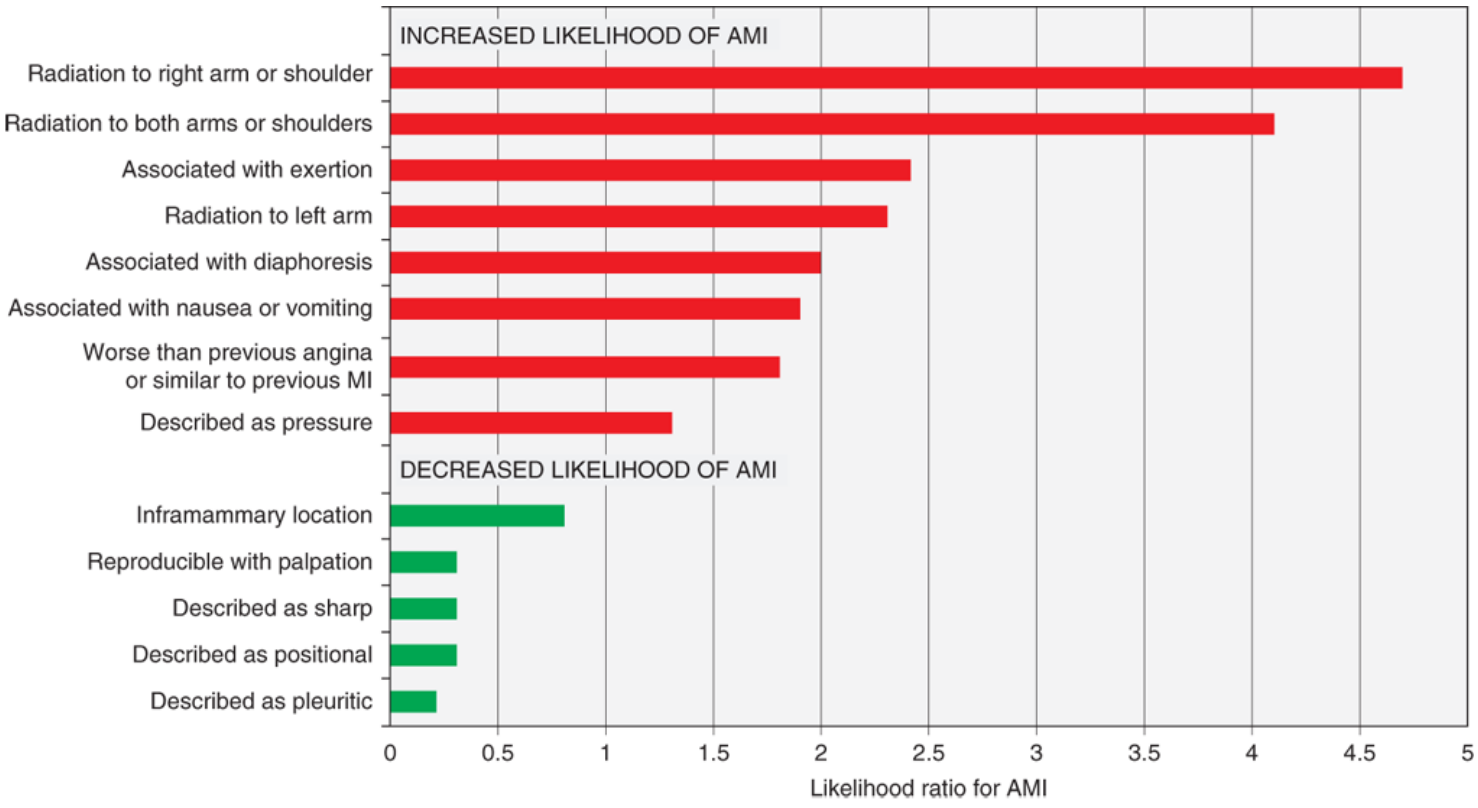
CT angiography is emerging as a modality for the evaluation of patients with acute chest discomfort **(See Chap. 236)**. Coronary CT angiography is a sensitive technique for detection of obstructive coronary disease, particularly in the proximal third of the major epicardial coronary arteries. CT appears to enhance the speed to disposition of patients with a low-intermediate probability for ACS; its major strength being the negative predictive value of a finding of no significant disease. In addition, contrast-enhanced CT can detect focal areas of myocardial injury in the acute setting. At the same time, CT angiography can exclude aortic dissection, pericardial effusion, and pulmonary embolism. Balancing factors in the consideration of the emerging role of coronary CT angiography in low-risk patients are radiation exposure and additional testing prompted by nondiagnostic abnormal results.

MRI

Cardiac magnetic resonance (CMR) imaging is an evolving, versatile technique for structural and functional evaluation of the heart and the vasculature of the chest **(See Chap. 236)**. CMR can be performed as a modality for pharmacologic stress perfusion imaging. Gadolinium-enhanced CMR can provide early detection of MI, defining areas of myocardial necrosis accurately, and can delineate patterns of myocardial disease that are often useful in discriminating ischemic from non-ischemic myocardial injury. Although usually not practical for the urgent evaluation of acute chest discomfort, CMR can be a useful modality for cardiac structural evaluation of patients with elevated cardiac troponin levels in the absence of definite coronary artery disease. CMR coronary angiography is in its early stages. MRI also permits highly accurate assessment for aortic dissection but is infrequently used as the first test because CT and transesophageal echocardiography are usually more practical.

FIGURE 11-2

Association of chest pain characteristics with the probability of acute myocardial infarction (AMI). Note that a subsequent larger study showed a non-significant association with radiation to the right arm. *(Figure prepared from data in CJ Swap, JT Nagurney: JAMA 294:2623, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 11-3

Examples of decision-aids used in conjunction with serial measurement of cardiac troponin for evaluation of acute chest pain. *(Figure prepared from data in SA Mahler et al: Int J Cardiol 168:795, 2013.)*

HEART Score		
History	Highly suspicious Moderately suspicious Slightly suspicious	2 1 0
ECG	Significant ST-depression Non-specific abnormality Normal	2 1 0
Age	≥65 y 45–<65 y <45 y	2 1 0
Risk factors	≥3 risk factors 1–2 risk factors None	2 1 0
Troponin (serial)	≥3 × 99th percentile 1–<3 × 99th percentile ≤99th percentile	2 1 0
	TOTAL	
	Low-risk: 0–3 **Not low risk:** ≥4	

North American Chest Pain Rule	
High Risk Criteria	**Y/N**
Typical symptoms for ischemia	
ECG: acute ischemic changes	
Age ≥50 y	
Known coronary artery disease	
Troponin (serial) >99th percentile	
Low-risk: All No **Not Low-risk:** Any Yes	



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 11-2

Considerations in the Assessment of the Patient with Chest Discomfort

1. Could the chest discomfort be due to an acute, potentially life-threatening condition that warrants urgent evaluation and management?			
Unstable ischemic heart disease	Aortic dissection	Pneumothorax	Pulmonary embolism
2. If not, could the discomfort be due to a chronic condition likely to lead to serious complications?			
Stable angina	Aortic stenosis	Pulmonary hypertension	
3. If not, could the discomfort be due to an acute condition that warrants specific treatment?			
Pericarditis	Pneumonia/pleuritis	Herpes zoster	
4. If not, could the discomfort be due to another treatable chronic condition?			
Esophageal reflux		Cervical disk disease	
Esophageal spasm		Arthritis of the shoulder or spine	
Peptic ulcer disease		Costochondritis	
Gallbladder disease		Other musculoskeletal disorders	
Other gastrointestinal conditions		Anxiety state	

Source: Developed by Dr. Thomas H. Lee for the 18th edition of *Harrison's Principles of Internal Medicine.*

CRITICAL PATHWAYS FOR ACUTE CHEST DISCOMFORT

Because of the challenges inherent in reliably identifying the small proportion of patients with serious causes of acute chest discomfort while not exposing the larger number of low-risk patients to unnecessary testing and extended ED or hospital evaluations, many medical centers have adopted critical pathways to expedite the assessment and management of patients with nontraumatic chest pain, often in dedicated chest pain units. Such pathways are generally aimed at (1) rapid identification, triage, and treatment of high-risk cardiopulmonary conditions (e.g., STEMI); (2) accurate identification of low-risk patients who can be safely observed in units with less intensive monitoring, undergo early exercise testing, or be discharged home; and (3) through more efficient and systematic accelerated diagnostic protocols, safe reduction in costs associated with overuse of testing and unnecessary hospitalizations. In some studies, provision of protocol-driven care in chest pain units has decreased costs and overall duration of hospital evaluation with no detectable excess of adverse clinical outcomes.

OUTPATIENT EVALUATION OF CHEST DISCOMFORT

Chest pain is common in outpatient practice, with a lifetime prevalence of 20–40% in the general population. More than 25% of patients with MI have had a related visit with a primary care physician in the previous month. The diagnostic principles are the same as in the ED. However, the pretest probability of an acute cardiopulmonary cause is significantly lower. Therefore, testing paradigms are less intense, with an emphasis on the history, physical examination, and ECG. Moreover, decision-aids developed for settings with a high prevalence of significant cardiopulmonary disease have lower positive predictive value when applied in the practitioner's office. However, in general, if the level of clinical suspicion of ACS is sufficiently high to consider troponin testing, the patient should be referred to the ED for evaluation.

FURTHER READING

Amsterdam EA et al: Testing of low-risk patients presenting to the emergency department with chest pain: A scientific statement from the American Heart Association. Circulation 122:1756, 2010.

Fanaroff AC et al: Does this patient with chest pain have acute coronary syndrome? JAMA 314:1955, 2015. [PubMed: 26547467]

Hermann LK et al: Yield of routine provocative cardiac testing among patients in an emergency department-based chest pain unit. JAMA Int Med 173:1128, 2013.

Mahler SA et al: The HEART Pathway randomized trial: Identifying emergency department patients with acute chest pain for early discharge. Circulation Cardiovasc Qual Outcomes 8:195, 2015.

Shah AS et al: High-sensitivity cardiac troponin I at presentation in patients with suspected acute coronary syndrome: A cohort study. Lancet 386:2481, 2016.

Chapter 12: Abdominal Pain

Danny O. Jacobs

INTRODUCTION

Correctly interpreting acute abdominal pain can be quite challenging. Few clinical situations require greater judgment, because the most catastrophic of events may be forecast by the subtlest of symptoms and signs. In every instance, the clinician must distinguish those conditions that require urgent intervention from those that do not and can best be managed nonoperatively. A meticulously executed, detailed history and physical examination are critically important for focusing the differential diagnosis and allowing the diagnostic evaluation to proceed expeditiously **(Table 12-1)**.

TABLE 12-1

Some Key Components of the Patient's History

Age
Time and mode of onset of the pain
Pain characteristics
Duration of symptoms
Location of pain and sites of radiation
Associated symptoms and their relationship to the pain
Nausea, emesis, and anorexia
Diarrhea, constipation, or other changes in bowel habits
Menstrual history

The etiologic classification in Table 12-2, although not complete, provides a useful framework for evaluating patients with abdominal pain.

TABLE 12-2

Some Important Causes of Abdominal Pain

Pain Originating in the Abdomen	
Parietal peritoneal inflammation Bacterial contamination Perforated appendix or other perforated viscus Pelvic inflammatory disease Chemical irritation Perforated ulcer Pancreatitis Mittelschmerz Mechanical obstruction of hollow viscera Obstruction of the small or large intestine Obstruction of the biliary tree Obstruction of the ureter	Vascular disturbances Embolism or thrombosis Vascular rupture Pressure or torsional occlusion Sickle cell anemia Abdominal wall Distortion or traction of mesentery Trauma or infection of muscles Distension of visceral surfaces, e.g., by hemorrhage Hepatic or renal capsules Inflammation Appendicitis Typhoid fever Neutropenic enterocolitis or “typhlitis”
Pain Referred from Extraabdominal Source	
Cardiothoracic Acute myocardial infarction Myocarditis, endocarditis, pericarditis Congestive heart failure Pneumonia (especially lower lobes) Pulmonary embolus	Pleurodynia Pneumothorax Empyema Esophageal disease, including spasm, rupture, or inflammation Genitalia Torsion of the testis
Metabolic Causes	
Diabetes Uremia Hyperlipidemia Hyperparathyroidism	Acute adrenal insufficiency Familial Mediterranean fever Porphyria C1 esterase inhibitor deficiency (angioneurotic edema)
Neurologic/Psychiatric Causes	

Pain Originating in the Abdomen	
Herpes zoster Tabes dorsalis Causalgia Radiculitis from infection or arthritis	Spinal cord or nerve root compression Functional disorders Psychiatric disorders
Toxic Causes	
Lead poisoning Insect or animal envenomation Black widow spider bites Snake bites	
Uncertain Mechanisms	
Narcotic withdrawal Heat stroke	

Any patient with abdominal pain of recent onset requires an early and thorough evaluation. The most common causes of abdominal pain on admission are nonspecific abdominal pain, acute appendicitis, pain of urologic origin, and intestinal obstruction. A diagnosis of “acute or surgical abdomen” is not acceptable because of its often misleading and erroneous connotations. Most patients who present with acute abdominal pain will have self-limited disease processes. However, it is important to remember that pain severity does not necessarily correlate with the severity of the underlying condition. And, the presence or absence of various degrees of “hunger” is unreliable as a sole indicator of the severity of intra-abdominal disease. The most obvious of “acute abdomens” may not require operative intervention, and the mildest of abdominal pains may herald an urgently correctable disease.

SOME MECHANISMS OF PAIN ORIGINATING IN THE ABDOMEN

Inflammation of the Parietal Peritoneum

The pain of parietal peritoneal inflammation is steady and aching in character and is located directly over the inflamed area, its exact reference being possible because it is transmitted by somatic nerves supplying the parietal peritoneum. The intensity of the pain is dependent on the type and amount of material to which the peritoneal surfaces are exposed in a given time period. For example, the sudden release of a small quantity of *sterile* acidic gastric juice into the peritoneal cavity causes much more pain than the same amount of grossly contaminated neutral feces. Enzymatically active pancreatic juice incites more pain and inflammation than does the same amount of sterile bile containing no potent enzymes. Blood is normally

only a mild irritant and the response to urine is also typically bland, so exposure of blood and urine to the peritoneal cavity may go unnoticed unless it is sudden and massive. Bacterial contamination, such as may occur with pelvic inflammatory disease or perforated distal intestine, causes low-intensity pain until multiplication causes a significant amount of inflammatory mediators to be released. Patients with perforated upper gastrointestinal ulcers may present entirely differently depending on how quickly gastric juices enter the peritoneal cavity, and its pH. Thus, the rate at which any inflammatory material irritates the peritoneum is important.

The pain of peritoneal inflammation is invariably accentuated by pressure or changes in tension of the peritoneum, whether produced by palpation or by movement such as with coughing or sneezing. The patient with peritonitis characteristically lies quietly in bed, preferring to avoid motion, in contrast to the patient with colic, who may be thrashing in discomfort.

Another characteristic feature of peritoneal irritation is tonic reflex spasm of the abdominal musculature, localized to the involved body segment. Its intensity depends on the integrity of the nervous system, the location of the inflammatory process, and the rate at which it develops. Spasm over a perforated retrocecal appendix or perforation into the lesser peritoneal sac may be minimal or absent because of the protective effect of overlying viscera. Catastrophic abdominal emergencies may be associated with minimal or no detectable pain or muscle spasm in obtunded, seriously ill, debilitated, immunosuppressed, or psychotic patients. A slowly developing process also often greatly attenuates the degree of muscle spasm.

Obstruction of Hollow Viscera

Intraluminal obstruction classically elicits intermittent or colicky abdominal pain that is not as well localized as the pain of parietal peritoneal irritation. However, the absence of cramping discomfort can be misleading because distention of a hollow viscus may also produce steady pain with only rare paroxysms.

Small-bowel obstruction often presents as poorly localized, intermittent periumbilical, or supraumbilical pain. As the intestine progressively dilates and loses muscular tone, the colicky nature of the pain may diminish. With superimposed strangulating obstruction, pain may spread to the lower lumbar region if there is traction on the root of the mesentery. The colicky pain of colonic obstruction is of lesser intensity, is commonly located in the infraumbilical area, and may often radiate to the lumbar region.

Sudden distention of the biliary tree produces a steady rather than colicky type of pain; hence, the term *biliary colic* is misleading. Acute distention of the gallbladder typically causes pain in the right upper quadrant with radiation to the right posterior region of the thorax or to the tip of the right scapula, but discomfort is also not uncommonly found near the midline. Distention of the common bile duct often causes epigastric pain that may radiate to the upper lumbar region. Considerable variation is common, however, so that differentiation between gallbladder or common ductal disease may be impossible.

Gradual dilatation of the biliary tree, as can occur with carcinoma of the head of the pancreas, may cause no pain or only a mild aching sensation in the epigastrium or right upper quadrant. The pain of distention of the

pancreatic ducts is similar to that described for distention of the common bile duct but, in addition, is very frequently accentuated by recumbency and relieved by the upright position.

Obstruction of the urinary bladder usually causes dull, low-intensity pain in the suprapubic region. Restlessness, without specific complaint of pain, may be the only sign of a distended bladder in an obtunded patient. In contrast, acute obstruction of the intravesicular portion of the ureter is characterized by severe suprapubic and flank pain that radiates to the penis, scrotum, or inner aspect of the upper thigh. Obstruction of the ureteropelvic junction manifests as pain near the costovertebral angle, whereas obstruction of the remainder of the ureter is associated with flank pain that often extends into the same side of the abdomen.

Vascular Disturbances

A frequent misconception is that pain due to intraabdominal vascular disturbances is sudden and catastrophic in nature. Certain disease processes, such as embolism or thrombosis of the superior mesenteric artery or impending rupture of an abdominal aortic aneurysm, can certainly be associated with diffuse, severe pain. Yet, just as frequently, the patient with occlusion of the superior mesenteric artery only has mild continuous or cramping diffuse pain for 2 or 3 days before vascular collapse or findings of peritoneal inflammation appear. The early, seemingly insignificant discomfort is caused by hyperperistalsis rather than peritoneal inflammation. Indeed, absence of tenderness and rigidity in the presence of continuous, diffuse pain (e.g., “pain out of proportion to physical findings”) in a patient likely to have vascular disease is quite characteristic of occlusion of the superior mesenteric artery. Abdominal pain with radiation to the sacral region, flank, or genitalia should always signal the possible presence of a rupturing abdominal aortic aneurysm. This pain may persist over a period of several days before rupture and collapse occur.

Abdominal Wall

Pain arising from the abdominal wall is usually constant and aching. Movement, prolonged standing, and pressure accentuate the discomfort and associated muscle spasm. In the relatively rare case of hematoma of the rectus sheath, now most frequently encountered in association with anticoagulant therapy, a mass may be present in the lower quadrants of the abdomen. Simultaneous involvement of muscles in other parts of the body usually serves to differentiate myositis of the abdominal wall from other processes that might cause pain in the same region.

REFERRED PAIN IN ABDOMINAL DISEASE

Pain referred to the abdomen from the thorax, spine, or genitalia may present a diagnostic challenge because diseases of the upper part of the abdominal cavity such as acute cholecystitis or perforated ulcer may be associated with intrathoracic complications. A most important, yet often forgotten, dictum is that the possibility of intrathoracic disease must be considered in every patient with abdominal pain, especially if the pain is in the upper abdomen.

Systematic questioning and examination directed toward detecting myocardial or pulmonary infarction, pneumonia, pericarditis, or esophageal disease (the intrathoracic diseases that most often masquerade as abdominal emergencies) will often provide sufficient clues to establish the proper diagnosis. Diaphragmatic pleuritis resulting from pneumonia or pulmonary infarction may cause pain in the right upper quadrant and pain in the supraclavicular area, the latter radiation to be distinguished from the referred subscapular pain caused by acute distention of the extrahepatic biliary tree. The ultimate decision as to the origin of abdominal pain may require deliberate and planned observation over a period of several hours, during which repeated questioning and examination will provide the diagnosis or suggest the appropriate studies.

Referred pain of thoracic origin is often accompanied by splinting of the involved hemithorax with respiratory lag and a decrease in excursion more marked than that seen in the presence of intraabdominal disease. In addition, apparent abdominal muscle spasm caused by referred pain will diminish during the inspiratory phase of respiration, whereas it persists throughout both respiratory phases if it is of abdominal origin. Palpation over the area of referred pain in the abdomen also does not usually accentuate the pain and, in many instances, actually seems to relieve it.

Thoracic disease and abdominal disease frequently coexist and may be difficult or impossible to differentiate. For example, the patient with known biliary tract disease often has epigastric pain during myocardial infarction, or biliary colic may be referred to the precordium or left shoulder in a patient who has suffered previously from angina pectoris. **For an explanation of the radiation of pain to a previously diseased area, see Chap. 10.**

Referred pain from the spine, which usually involves compression or irritation of nerve roots, is characteristically intensified by certain motions such as cough, sneeze, or strain and is associated with hyperesthesia over the involved dermatomes. Pain referred to the abdomen from the testes or seminal vesicles is generally accentuated by the slightest pressure on either of these organs. The abdominal discomfort experienced is of dull, aching character and is poorly localized.

METABOLIC ABDOMINAL CRISES

Pain of metabolic origin may simulate almost any other type of intraabdominal disease. Several mechanisms may be at work. In certain instances, such as hyperlipidemia, the metabolic disease itself may be accompanied by an intraabdominal process such as pancreatitis, which can lead to unnecessary laparotomy unless recognized. C1 esterase deficiency associated with angioneurotic edema is often associated with episodes of severe abdominal pain. Whenever the cause of abdominal pain is obscure, a metabolic origin always must be considered. Abdominal pain is also the hallmark of familial Mediterranean fever (Chap. 362).

The pain of porphyria and of lead colic is usually difficult to distinguish from that of intestinal obstruction, because severe hyperperistalsis is a prominent feature of both. The pain of uremia or diabetes is nonspecific, and the pain and tenderness frequently shift in location and intensity. Diabetic acidosis may be precipitated by acute appendicitis or intestinal obstruction, so if prompt resolution of the abdominal pain does not result from correction of the metabolic abnormalities, an underlying organic problem should be suspected. Black

widow spider bites produce intense pain and rigidity of the abdominal muscles and back, an area infrequently involved in intraabdominal disease.

IMMUNOCOMPROMISE

Evaluating and diagnosing causes of abdominal pain in immunosuppressed or otherwise immunocompromised patients is very difficult. This includes those who have undergone organ transplantation; who are receiving immunosuppressive treatments for autoimmune diseases, chemotherapy, or glucocorticoids; who have AIDS; and who are very old. In these circumstances, normal physiologic responses may be absent or masked. In addition, unusual infections may cause abdominal pain where the etiologic agents include cytomegalovirus, mycobacteria, protozoa, and fungi. These pathogens may affect all gastrointestinal organs, including the gallbladder, liver, and pancreas, as well as the gastrointestinal tract, causing occult or overtly symptomatic perforations of the latter. Splenic abscesses due to *Candida* or *Salmonella* infection should also be considered, especially when evaluating patients with left upper quadrant or left flank pain. Acalculous cholecystitis may be observed in immunocompromised patients or those with AIDS, where it is often associated with cryptosporidiosis or cytomegalovirus infection.

Neutropenic enterocolitis (typhlitis) is often identified as a cause of abdominal pain and fever in some patients with bone marrow suppression due to chemotherapy. Acute graft-versus-host disease should be considered in this circumstance. Optimal management of these patients requires meticulous follow-up including serial examinations to assess the need for more surgical intervention, for example, to address perforation.

NEUROGENIC CAUSES

Diseases that injure sensory nerves may cause causalgic pain. It has a burning character and is usually limited to the distribution of a given peripheral nerve. Stimuli that are normally not painful such as touch or a change in temperature may be causalgic and are often present even at rest. The demonstration of irregularly spaced cutaneous “pain spots” may be the only indication that an old nerve injury exists. Even though the pain may be precipitated by gentle palpation, rigidity of the abdominal muscles is absent, and the respirations are not usually disturbed. Distention of the abdomen is uncommon, and the pain has no relationship to food intake.

Pain arising from spinal nerves or roots comes and goes suddenly and is of a lancinating type (Chap. 14). It may be caused by herpes zoster, impingement by arthritis, tumors, a herniated nucleus pulposus, diabetes, or syphilis. It is not associated with food intake, abdominal distention, or changes in respiration. Severe muscle spasms, when present, are either relieved but are certainly not accentuated by abdominal palpation. The pain is made worse by movement of the spine and is usually confined to a few dermatomes. Hyperesthesia is very common.

Pain due to functional causes conforms to none of the aforementioned patterns. Mechanisms of disease are not clearly established. Irritable bowel syndrome (IBS) is a functional gastrointestinal disorder characterized by abdominal pain and altered bowel habits. The diagnosis is made on the basis of clinical criteria **(Chap. 320)** and after exclusion of demonstrable structural abnormalities. The episodes of abdominal pain may be brought on by stress, and the pain varies considerably in type and location. Nausea and vomiting are rare. Localized tenderness and muscle spasm are inconsistent or absent. The causes of IBS or related functional disorders are not yet fully understood.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Abdominal Pain

Few abdominal conditions require such urgent operative intervention that an orderly approach needs to be abandoned, no matter how ill the patient is. Only patients with exsanguinating intraabdominal hemorrhage (e.g., ruptured aneurysm) must be rushed to the operating room immediately, but in such instances, only a few minutes are required to assess the critical nature of the problem. Under these circumstances, all obstacles must be swept aside, adequate venous access for fluid replacement obtained, and the operation begun. Unfortunately, many of these patients may die in the radiology department or the emergency room while awaiting unnecessary examinations. *There are no absolute contraindications to operation when massive intraabdominal hemorrhage is present.* Fortunately, this situation is relatively rare. This statement does not necessarily apply to patients with intraluminal gastrointestinal hemorrhage, who can often be managed by other means **(Chap. 44)**. In these patients, obtaining a *detailed history when possible* can be extremely helpful even though it can be laborious and time-consuming. Decision-making regarding next steps is facilitated and a reasonably accurate diagnosis can be made before any further diagnostic testing is undertaken.

In cases of *acute* abdominal pain, a diagnosis can be readily established in most instances, whereas success is not so frequent in patients with *chronic* pain. IBS is one of the most common causes of abdominal pain and must always be kept in mind **(Chap. 320)**. The location of the pain can assist in narrowing the differential diagnosis **(Table 12-3)**; however, the *chronological sequence of events* in the patient's history is often more important than the pain's location. Careful attention should be paid to the extraabdominal regions. Narcotics or analgesics should *not* be withheld until a definitive diagnosis or a definitive plan has been formulated; obfuscation of the diagnosis by adequate analgesia is unlikely.

An accurate menstrual history in a female patient is essential. It is important to remember that normal anatomic relationships can be significantly altered by the gravid uterus. Abdominal and pelvic pain may occur during pregnancy due to conditions that do not require operation. Lastly, some otherwise noteworthy laboratory values (e.g., leukocytosis) may represent the normal physiologic changes of pregnancy.

In the examination, simple critical inspection of the patient, for example, of facies, position in bed, and respiratory activity, provides valuable clues. The amount of information to be gleaned is directly proportional to the *gentleness* and thoroughness of the examiner. Once a patient with peritoneal inflammation has been examined brusquely, accurate assessment by the next examiner becomes almost impossible. Eliciting rebound tenderness by sudden release of a deeply palpating hand in a patient with suspected peritonitis is cruel and unnecessary. The same information can be obtained by gentle percussion of the abdomen (rebound tenderness on a miniature scale), a maneuver that can be far more precise and localizing. Asking the patient to cough will elicit true rebound tenderness without the need for placing a hand on the abdomen. Furthermore, the forceful demonstration of rebound tenderness will startle and induce protective spasm in a nervous or worried patient in whom true rebound tenderness is not present. A palpable gallbladder will be missed if palpation is so aggressive that voluntary muscle spasm becomes superimposed on involuntary muscular rigidity.

As with history-taking, sufficient time should be spent in the examination. Abdominal signs may be minimal but nevertheless, if accompanied by consistent symptoms, may be exceptionally meaningful. Abdominal signs may be virtually or totally absent in cases of pelvic peritonitis, so careful *pelvic and rectal examinations are mandatory in every patient with abdominal pain.* Tenderness on pelvic or rectal examination in the absence of other abdominal signs can be caused by operative indications such as perforated appendicitis, diverticulitis, twisted ovarian cyst, and many others. Much attention has been paid to the presence or absence of peristaltic sounds, their quality, and their frequency. Auscultation of the abdomen is one of the least revealing aspects of the physical examination of a patient with abdominal pain. Catastrophes such as a strangulating small-intestinal obstruction or perforated appendicitis may occur in the presence of normal peristaltic sounds. Conversely, when the proximal part of the intestine above obstruction becomes markedly distended and edematous, peristaltic sounds may lose the characteristics of borborygmi and become weak or absent, even when peritonitis is not present. It is usually the severe chemical peritonitis of sudden onset that is associated with the truly silent abdomen.

Laboratory examinations may be valuable in assessing the patient with abdominal pain, yet, with few exceptions, they rarely establish a diagnosis. Leukocytosis should never be the single deciding factor as to whether or not operation is indicated. A white blood cell count >20,000/μL may be observed with perforation of a viscus, but pancreatitis, acute cholecystitis, pelvic inflammatory disease, and intestinal infarction may also be associated with marked leukocytosis. A normal white blood cell count is not rare in cases of perforation of abdominal viscera. A diagnosis of anemia may be more helpful than the white blood cell count, especially when combined with the history.

The urinalysis may reveal the state of hydration or rule out severe renal disease, diabetes, or urinary infection. Blood urea nitrogen, glucose, and serum bilirubin levels and liver function tests may be helpful. Serum amylase levels may be increased by many diseases other than pancreatitis, for example, perforated ulcer, strangulating intestinal obstruction, and acute cholecystitis; thus, elevations of serum amylase do not rule in or rule out the need for an operation.

Plain and upright or lateral decubitus radiographs of the abdomen have limited utility and may be unnecessary in some patients who have substantial evidence of some diseases such as acute appendicitis or strangulated external hernia. Where the indications for surgical or medical intervention are not clear, low dose computed tomography is preferred to abdominal radiography when evaluating non-traumatic acute abdominal pain.

Very rarely, barium or water-soluble contrast study of the upper part of the gastrointestinal tract are appropriate radiographic investigations and may demonstrate partial intestinal obstruction that may elude diagnosis by other means. If there is any question of obstruction of the colon, oral administration of barium sulfate should be avoided. On the other hand, in cases of suspected colonic obstruction (without perforation), a contrast enema may be diagnostic.

In the absence of trauma, peritoneal lavage has been replaced as a diagnostic tool by CT scanning and laparoscopy. Ultrasonography has proved to be useful in detecting an enlarged gallbladder or pancreas, the presence of gallstones, an enlarged ovary, or a tubal pregnancy. Laparoscopy is especially helpful in diagnosing pelvic conditions, such as ovarian cysts, tubal pregnancies, salpingitis, and acute appendicitis and other disease processes. Laparoscopy has a particular advantage over imaging in that the underlying etiologic condition can often be definitively addressed.

Radioisotopic hepatobiliary iminodiacetic acid scans (HIDAs) may help differentiate acute cholecystitis or biliary colic from acute pancreatitis. A CT scan may demonstrate an enlarged pancreas, ruptured spleen, or thickened colonic or appendiceal wall and streaking of the mesocolon or mesoappendix characteristic of diverticulitis or appendicitis.

Sometimes, even under the best circumstances with all available aids and with the greatest of clinical skill, a definitive diagnosis cannot be established at the time of the initial examination. And, in some cases, operation may be indicated based on clinical grounds alone. Should that decision be questionable, watchful waiting with repeated questioning and examination will often elucidate the true nature of the illness and indicate the proper course of action.

TABLE 12-3

Differential Diagnoses of Abdominal Pain by Location

Right Upper Quadrant	**Epigastric**	**Left Upper Quadrant**
Cholecystitis Cholangitis Pancreatitis Pneumonia/empyema Pleurisy/pleurodynia Subdiaphragmatic abscess Hepatitis Budd-Chiari syndrome	Peptic ulcer disease Gastritis GERD Pancreatitis Myocardial infarction Pericarditis Ruptured aortic aneurysm Esophagitis	Splenic infarct Splenic rupture Splenic abscess Gastritis Gastric ulcer Pancreatitis Subdiaphragmatic abscess
Right Lower Quadrant	**Periumbilical**	**Left Lower Quadrant**
Appendicitis Salpingitis Inguinal hernia Ectopic pregnancy Nephrolithiasis Inflammatory bowel disease Mesenteric lymphadenitis Typhlitis	Early appendicitis Gastroenteritis Bowel obstruction Ruptured aortic aneurysm	Diverticulitis Salpingitis Inguinal hernia Ectopic pregnancy Nephrolithiasis Irritable bowel syndrome Inflammatory bowel disease
Diffuse Nonlocalized Pain		
Gastroenteritis Mesenteric ischemia Bowel obstruction Irritable bowel syndrome Peritonitis Diabetes	Malaria Familial Mediterranean fever Metabolic diseases Psychiatric disease	

Abbreviation: GERD, gastroesophageal reflux disease.

ACKNOWLEDGMENT

We gratefully acknowledge the enormous contribution to this chapter and the approach it espouses to William Silen, who wrote this chapter for many editions.

FURTHER READING

Bhangu A et al: Acute appendicitis: Modern understanding of pathogenesis, diagnosis and management, Lancet 386:1278, 2015. [PubMed: 26460662]

Cartwright SL, Knudson MP: Diagnostic imaging of acute abdominal pain in adults. Am Fam Phys 91: 452, 2015.

Huckins DS et al: Diagnostic performance of a biomarker panel as a negative predictor for acute appendicitis in acute emergency department patients with abdominal pain. Available from *http://dx.doi.org.ezproxy1.library.usyd.edu.au/10.1016/j.ajem.2016.11.027*. Accessed November 2016.

Nayor J et al: Tracing the cause of abdominal pain. N Engl J Med 375:e8, 2016. [PubMed: 27509123]

Phillips MT: Clinical yield of computed tomography scans in the emergency department for abdominal pain. J Invest Med 64:542, 2016.

Silen W, Cope Z: *Cope's Early Diagnosis of the Acute Abdomen*, 22nd ed. New York, Oxford University Press, 2010.

Chapter 13: Headache

Peter J. Goadsby

FIGURE 13-1

INTRODUCTION

Headache is among the most common reasons patients seek medical attention, on a global basis being responsible for more disability than any other neurologic problem. Diagnosis and management are based on a careful clinical approach augmented by an understanding of the anatomy, physiology, and pharmacology of the nervous system pathways mediating the various headache syndromes. This chapter will focus on the general approach to a patient with headache; migraine and other primary headache disorders are discussed in **Chap. 422**.

GENERAL PRINCIPLES

A classification system developed by the International Headache Society (*www.ihs-headache.org/ichd-guidelines*) characterizes headache as primary or secondary **(Table 13-1)**. *Primary headaches* are those in which headache and its associated features are the disorder itself, whereas *secondary headaches* are those caused by exogenous disorders (Headache Classification Committee of the International Headache Society, 2018). Primary headache often results in considerable disability and a decrease in the patient's quality of life. Mild secondary headache, such as that seen in association with upper respiratory tract infections, is common but rarely worrisome. Life-threatening headache is relatively uncommon, but vigilance is required in order to recognize and appropriately treat such patients.

TABLE 13-1

Common Causes of Headache

Primary Headache		Secondary Headache	
Type	**%**	**Type**	**%**
Tension-type	69	Systemic infection	63
Migraine	16	Head injury	4
Idiopathic stabbing	2	Vascular disorders	1
Exertional	1	Subarachnoid hemorrhage	<1
Cluster	0.1	Brain tumor	0.1

Source: After J Olesen et al: *The Headaches*. Philadelphia, Lippincott Williams & Wilkins, 2005.

ANATOMY AND PHYSIOLOGY OF HEADACHE

Pain usually occurs when peripheral nociceptors are stimulated in response to tissue injury, visceral distension, or other factors (Chap. 10). In such situations, pain perception is a normal physiologic response mediated by a healthy nervous system. Pain can also result when pain-producing pathways of the peripheral or central nervous system (CNS) are damaged or activated inappropriately. Headache may originate from either or both mechanisms. Relatively few cranial structures are pain-producing; these include the scalp, meningeal arteries, dural sinuses, falx cerebri, and proximal segments of the large pial arteries. The ventricular ependyma, choroid plexus, pial veins, and much of the brain parenchyma are not pain-producing.

The key structures involved in primary headache appear to be the following:

The large intracranial vessels and dura mater and the peripheral terminals of the trigeminal nerve that innervate these structures

The caudal portion of the trigeminal nucleus, which extends into the dorsal horns of the upper cervical spinal cord and receives input from the first and second cervical nerve roots (the trigeminocervical complex)

Rostral pain-processing regions, such as the ventroposteromedial thalamus and the cortex

The pain-modulatory systems in the brain that modulate input from trigeminal nociceptors at all levels of the pain-processing pathways and influence vegetative functions, such as hypothalamus and brainstem

structures

The innervation of the large intracranial vessels and dura mater by the trigeminal nerve is known as the *trigeminovascular system*. Cranial autonomic symptoms, such as *lacrimation, conjunctival injection, nasal congestion, rhinorrhea, periorbital swelling, aural fullness*, and *ptosis*, are prominent in the trigeminal autonomic cephalalgias (TACs), including cluster headache and paroxysmal hemicrania, and may also be seen in migraine, even in children. These autonomic symptoms reflect activation of cranial parasympathetic pathways, and functional imaging studies indicate that vascular changes in migraine and cluster headache, when present, are similarly driven by these cranial autonomic systems. Moreover, they can often be mistaken for symptoms or signs of cranial sinus inflammation, which is thus overdiagnosed and inappropriately managed. Migraine and other primary headache types are not “vascular headaches”; these disorders do not reliably manifest vascular changes, and treatment outcomes cannot be predicted by vascular effects. Migraine is a brain disorder and is best understood and managed as such.

CLINICAL EVALUATION OF ACUTE, NEW-ONSET HEADACHE

The patient who presents with a new, severe headache has a differential diagnosis that is quite different from the patient with recurrent headaches over many years. In new-onset and severe headache, the probability of finding a potentially serious cause is considerably greater than in recurrent headache. Patients with recent onset of pain require prompt evaluation and appropriate treatment. Serious causes to be considered include meningitis, subarachnoid hemorrhage, epidural or subdural hematoma, glaucoma, tumor, and purulent sinusitis. When worrisome symptoms and signs are present (Table 13-2), rapid diagnosis and management are critical.

TABLE 13-2

Headache Symptoms That Suggest a Serious Underlying Disorder

Sudden-onset headache
First severe headache
“Worst” headache ever
Vomiting that precedes headache
Subacute worsening over days or weeks
Pain induced by bending, lifting, cough
Pain that disturbs sleep or presents immediately upon awakening
Known systemic illness
Onset after age 55
Fever or unexplained systemic signs
Abnormal neurologic examination
Pain associated with local tenderness, e.g., region of temporal artery

A careful neurologic examination is an essential first step in the evaluation. In most cases, patients with an abnormal examination or a history of recent-onset headache should be evaluated by a computed tomography (CT) or magnetic resonance imaging (MRI) study of the brain. As an initial screening procedure for intracranial pathology in this setting, CT and MRI methods appear to be equally sensitive. In some circumstances, a lumbar puncture (LP) is also required, unless a benign etiology can be otherwise established. A general evaluation of acute headache might include cranial arteries by palpation; cervical spine by the effect of passive movement of the head and by imaging; the investigation of cardiovascular and renal status by blood pressure monitoring and urine examination; and eyes by funduscopy, intraocular pressure measurement, and refraction.

The psychological state of the patient should also be evaluated because a relationship exists between head pain, depression, and anxiety. This is intended to identify comorbidity rather than provide an explanation for the headache, because troublesome headache is seldom simply caused by mood change. Although it is

notable that medicines with antidepressant actions are also effective in the preventive treatment of both tension-type headache and migraine, each symptom must be treated optimally.

Underlying recurrent headache disorders may be activated by pain that follows otologic or endodontic surgical procedures. Thus, pain about the head as the result of diseased tissue or trauma may reawaken an otherwise quiescent migraine syndrome. Treatment of the headache is largely ineffective until the cause of the primary problem is addressed.

Serious underlying conditions that are associated with headache are described below. Brain tumor is a rare cause of headache and even less commonly a cause of severe pain. The vast majority of patients presenting with severe headache have a benign cause.

SECONDARY HEADACHE

The management of secondary headache focuses on diagnosis and treatment of the underlying condition.

MENINGITIS

Acute, severe headache with stiff neck and fever suggests meningitis. LP is mandatory. Often there is striking accentuation of pain with eye movement. Meningitis can be easily mistaken for migraine in that the cardinal symptoms of pounding headache, photophobia, nausea, and vomiting are frequently present, perhaps reflecting the underlying biology of some of the patients.

Meningitis is discussed in Chaps. 133 and 134.

INTRACRANIAL HEMORRHAGE

Acute, maximal in <5 min, severe headache lasting >5 min with stiff neck but without fever suggests subarachnoid hemorrhage. A ruptured aneurysm, arteriovenous malformation, or intraparenchymal hemorrhage may also present with headache alone. Rarely, if the hemorrhage is small or below the foramen magnum, the head CT scan can be normal. Therefore, LP may be required to diagnose definitively subarachnoid hemorrhage.

Subarachnoid hemorrhage is discussed in Chap. 302, and intracranial hemorrhage in Chap. 421.

BRAIN TUMOR

Approximately 30% of patients with brain tumors consider headache to be their chief complaint. The head pain is usually nondescript—an intermittent deep, dull aching of moderate intensity, which may worsen with exertion or change in position and may be associated with nausea and vomiting. This pattern of symptoms results from migraine far more often than from brain tumor. The headache of brain tumor disturbs sleep in about 10% of patients. Vomiting that precedes the appearance of headache by weeks is highly characteristic of posterior fossa brain tumors. A history of amenorrhea or galactorrhea should lead one to question

whether a prolactin-secreting pituitary adenoma (or the polycystic ovary syndrome) is the source of headache. Headache arising de novo in a patient with known malignancy suggests either cerebral metastases or carcinomatous meningitis, or both. Head pain appearing abruptly after bending, lifting, or coughing can be due to a posterior fossa mass, a Chiari malformation, or low cerebrospinal fluid (CSF) volume.

Brain tumors are discussed in Chap. 86.

TEMPORAL ARTERITIS

(See also Chaps. 28 and 356) Temporal (giant cell) arteritis is an inflammatory disorder of arteries that frequently involves the extracranial carotid circulation. It is a common disorder of the elderly; its annual incidence is 77 per 100,000 individuals aged ≥50. The average age of onset is 70 years, and women account for 65% of cases. About half of patients with untreated temporal arteritis develop blindness due to involvement of the ophthalmic artery and its branches; indeed, the ischemic optic neuropathy induced by giant cell arteritis is the major cause of rapidly developing bilateral blindness in patients >60 years. Because treatment with glucocorticoids is effective in preventing this complication, prompt recognition of the disorder is important.

Typical presenting symptoms include headache, polymyalgia rheumatica **(Chap. 356)**, jaw claudication, fever, and weight loss. Headache is the dominant symptom and often appears in association with malaise and muscle aches. Head pain may be unilateral or bilateral and is located temporally in 50% of patients but may involve any and all aspects of the cranium. Pain usually appears gradually over a few hours before peak intensity is reached; occasionally, it is explosive in onset. The quality of pain is infrequently throbbing; it is almost invariably described as dull and boring, with superimposed episodic stabbing pains similar to the sharp pains that appear in migraine. Most patients can recognize that the origin of their head pain is superficial, external to the skull, rather than originating deep within the cranium (the pain site usually identified migraineurs). Scalp tenderness is present, often to a marked degree; brushing the hair or resting the head on a pillow may be impossible because of pain. Headache is usually worse at night and often aggravated by exposure to cold. Additional findings may include reddened, tender nodules or red streaking of the skin overlying the temporal arteries, and tenderness of the temporal or, less commonly, the occipital arteries.

The erythrocyte sedimentation rate (ESR) is often, although not always, elevated; a normal ESR does not exclude giant cell arteritis. A temporal artery biopsy followed by immediate treatment with prednisone 80 mg daily for the first 4–6 weeks should be initiated when clinical suspicion is high. The prevalence of migraine among the elderly is substantial, considerably higher than that of giant cell arteritis. Migraineurs often report amelioration of their headaches with prednisone; thus, caution must be used when interpreting the therapeutic response.

GLAUCOMA

Glaucoma may present with a prostrating headache associated with nausea and vomiting. The headache often starts with severe eye pain. On physical examination, the eye is often red with a fixed, moderately dilated pupil.

Glaucoma is discussed in Chap. 28.

PRIMARY HEADACHE DISORDERS

Primary headaches are disorders in which headache and associated features occur in the absence of any exogenous cause. The most common are migraine, tension-type headache, and the TACs, notably cluster headache. These entities are discussed in detail in **Chap. 422**.

CHRONIC DAILY OR NEAR-DAILY HEADACHE

The broad description of chronic daily headache (CDH) can be applied when a patient experiences headache on 15 days or more per month. CDH is not a single entity; it encompasses a number of different headache syndromes, both primary and secondary **(Table 13-3)**. In aggregate, this group presents considerable disability and is thus specially dealt with here. Population-based estimates suggest that about 4% of adults have daily or near-daily headache.

TABLE 13-3

Classification of Daily or Near-Daily Headache

Primary		
>4 h Daily	**<4 h Daily**	**Secondary**
Chronic migraine[a]	Chronic cluster headache[b]	Posttraumatic Head injury Iatrogenic Postinfectious
Chronic tension-type headache[a]	Chronic paroxysmal hemicrania	Inflammatory, such as Giant cell arteritis Sarcoidosis Behçet's syndrome
Hemicrania continua[a]	SUNCT/SUNA	Chronic CNS infection
New daily persistent headache[a]	Hypnic headache	Medication-overuse headache[a]

[a]May be complicated by medication overuse. [b]Some patients may have headache >4 h/d.

Abbreviations: CNS, central nervous system; SUNA, *s*hort-lasting *u*nilateral *n*euralgiform headache *a*ttacks with cranial autonomic symptoms; SUNCT, *s*hort-lasting *u*nilateral *n*euralgiform headache attacks with *c*onjunctival injection and *t*earing.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Chronic Daily Headache

The first step in the management of patients with CDH is to diagnose any secondary headache and treat that problem (Table 13-3). This can sometimes be a challenge where the underlying cause triggers a worsening of a primary headache. For patients with primary headaches, diagnosis of the headache type will guide therapy. Preventive treatments such as tricyclics, either amitriptyline or nortriptyline at doses up to 1 mg/kg, are very useful in patients with CDH arising from migraine or tension-type headache or where the secondary cause has activated the underlying primary headache. Tricyclics are started in low doses (10–25 mg) daily and may be given 12 h before the expected time of awakening in order to avoid excess morning sleepiness. Medicines

including topiramate, valproate, propranolol, flunarizine (not available in the United States), and candesartan are also useful in migraine.

MANAGEMENT OF MEDICALLY INTRACTABLE DISABLING PRIMARY CHRONIC DAILY HEADACHE

The management of medically intractable headache is difficult, although developments in therapy are at hand. Monoclonal antibodies to calcitonin gene-related peptide (CGRP) or its receptor have been reported to be effective and well-tolerated in chronic migraine in phase II/III randomized placebo-controlled trials. Non-invasive neuromodulatory approaches, such as single pulse transcranial magnetic stimulation and non-invasive vagal nerve stimulation, which appear to modulate thalamic processing or brainstem mechanisms, respectively, in migraine have, or are, entering clinical practice, respectively. Non-invasive vagal nerve stimulation has also shown promise in chronic cluster headache, chronic paroxysmal hemicrania, short-lasting unilateral neuralgiform headache attacks with cranial autonomic symptoms (SUNA), short-lasting unilateral neuralgiform headache attacks with conjunctival injection and tearing (SUNCT), and hemicrania continua (Chap. 422). Other modalities are discussed in Chap. 422.

MEDICATION-OVERUSE HEADACHE

Overuse of analgesic medication for headache can aggravate headache frequency, markedly impair the effect of preventive medicines, and induce a state of refractory daily or near-daily headache called *medication-overuse headache*. A proportion of patients who stop taking analgesics will experience substantial improvement in the severity and frequency of their headache. However, even after cessation of analgesic use, many patients continue to have headache, although they may feel clinically improved in some way, especially if they have been using opioids or barbiturates regularly. The residual symptoms probably represent the underlying primary headache disorder, and most commonly, this issue occurs in patients prone to migraine.

Management of Medication Overuse: Outpatients

For patients who overuse medications, it is often helpful that analgesic use be reduced and eliminated. One approach is to reduce the medication dose by 10% every 1–2 weeks. Immediate cessation of analgesic use is possible for some patients, provided there is no contraindication. Both approaches are facilitated by the use of a medication diary maintained during the month or two before cessation; this helps to identify the scope of the problem. A small dose of a nonsteroidal anti-inflammatory drug (NSAID) such as naproxen, 500 mg bid, if tolerated, will help relieve residual pain as analgesic use is reduced. NSAID overuse is not usually a problem for patients with daily headache when a NSAID with a longer half-life is taken once or twice daily; however, overuse problems may develop with more frequent dosing schedules or shorter acting NSAIDS. Once the patient has substantially reduced analgesic use, a preventive medication should be introduced, although another equally widely used approach is to commence the preventive at the same time as the analgesic reduction is started. It must be emphasized that *preventives often do not work in the presence of analgesic overuse*. The most common cause of unresponsiveness to treatment is the use of a preventive when analgesics continue to be used regularly. For some patients, discontinuing analgesics is very difficult; often the best approach is to inform the patient that some degree of pain is inevitable during this initial period.

Management of Medication Overuse: Inpatients

Some patients will require hospitalization for detoxification. Such patients have typically failed efforts at outpatient withdrawal or have a significant medical condition, such as diabetes mellitus or epilepsy, which would complicate withdrawal as an outpatient. Following admission to the hospital, acute medications are withdrawn completely on the first day, in the absence of a contraindication. Antiemetics and fluids are administered as required; clonidine is used for opioid withdrawal symptoms. For acute intolerable pain during the waking hours, aspirin, 1 g IV (not approved in United States), is useful. IM chlorpromazine can be helpful at night; patients must be adequately hydrated. Three to five days into the admission, as the effect of the withdrawn substance wears off, a course of IV dihydroergotamine (DHE) can be used. DHE, administered every 8 h for 5 consecutive days, can induce a significant remission that allows a preventive treatment to be established. Serotonin 5-HT_3 receptor antagonists, such as ondansetron or granisetron, or the neurokinin receptor antagonist, aprepitant, may be required with DHE to prevent significant nausea, and domperidone (not approved in the United States) orally or by suppository can be very helpful. Avoiding sedating or otherwise side effect–prone antiemetics is helpful.

NEW DAILY PERSISTENT HEADACHE

New daily persistent headache (NDPH) is a clinically distinct syndrome with important secondary causes; these are listed in **Table 13-4**.

Clinical Presentation

The patient with NDPH presents with headache on most if not all days, and the patient can clearly, and often vividly, recall the moment of onset. The headache usually begins abruptly, but onset may be more gradual; evolution over 3 days has been proposed as the upper limit for this syndrome. Patients typically recall the exact day and circumstances of the onset of headache; the new, persistent head pain does not remit. The first priority is to distinguish between a primary and a secondary cause of this syndrome. Subarachnoid hemorrhage is the most serious of the secondary causes and must be excluded either by history or appropriate investigation **(Chap. 302)**.

Secondary NDPH

Low Csf Volume Headache

In these syndromes, head pain is positional: it begins when the patient sits or stands upright and resolves upon reclining. The pain, which is occipitofrontal, is usually a dull ache but may be throbbing. Patients with chronic low CSF volume headache typically present with a history of headache from 1 day to the next that is generally not present on waking but worsens during the day. Recumbency usually improves the headache within minutes, and it can take only minutes to an hour for the pain to return when the patient resumes an upright position.

The most common cause of headache due to persistent low CSF volume is CSF leak following LP. Post-LP headache usually begins within 48 h but may be delayed for up to 12 days. Its incidence is between 10 and 30%. Beverages with caffeine may provide temporary relief. Besides LP, index events may include epidural injection or a vigorous Valsalva maneuver, such as from lifting, straining, coughing, clearing the eustachian tubes in an airplane, or multiple orgasms. Spontaneous CSF leaks are well recognized, and the diagnosis

should be considered whenever the headache history is typical, even when there is no obvious index event. As time passes from the index event, the postural nature may become less apparent; cases in which the index event occurred several years before the eventual diagnosis have been recognized. Symptoms appear to result from low volume rather than low pressure: although low CSF pressures, typically 0–50 mm CSF, are usually identified, a pressure as high as 140 mm CSF has been noted with a documented leak.

Postural orthostatic tachycardia syndrome (POTS; **Chap. 432**) can present with orthostatic headache similar to low CSF volume headache and is a diagnosis that needs consideration in this setting.

When imaging is indicated to identify the source of a presumed leak, an MRI with gadolinium is the initial study of choice **(Fig. 13-1)**. A striking pattern of diffuse meningeal enhancement is so typical that in the appropriate clinical context the diagnosis is established. Chiari malformations may sometimes be noted on MRI; in such cases, surgery to decompress the posterior fossa is *not* indicated and usually worsens the headache. Spinal MRI with T2 weighting may reveal a leak, and spinal MRI may demonstrate spinal meningeal cysts whose role in these syndromes is yet to be elucidated. The source of CSF leakage may be identified by spinal MRI with appropriate sequences, by CT, or increasingly by MR myelography. Less used now, ^{111}In-DTPA CSF studies in the absence of a directly identified site of leakage, may demonstrate early emptying of ^{111}In-DTPA tracer into the bladder or slow progress of tracer across the brain suggesting a CSF leak.

Initial treatment for low CSF volume headache is bed rest. For patients with persistent pain, IV caffeine (500 mg in 500 mL of saline administered over 2 h) can be very effective. An electrocardiogram (ECG) to screen for arrhythmia should be performed before administration. It is reasonable to administer at least two infusions of caffeine before embarking on additional tests to identify the source of the CSF leak. Because IV caffeine is safe and can be curative, it spares many patients the need for further investigations. If unsuccessful, an abdominal binder may be helpful. If a leak can be identified, an autologous blood patch is usually curative. A blood patch is also effective for post-LP headache; in this setting, the location is empirically determined to be the site of the LP. In patients with intractable headache, oral theophylline is a useful alternative; however, its effect is less rapid than caffeine.

Raised CSF Pressure Headache

Raised CSF pressure is well recognized as a cause of headache. Brain imaging can often reveal the cause, such as a space-occupying lesion. NDPH due to raised CSF pressure can be the presenting symptom for patients with idiopathic intracranial hypertension (pseudotumor cerebri) without visual problems, particularly when the fundi are normal. Persistently raised intracranial pressure can trigger chronic migraine. These patients typically present with a history of generalized headache that is present on waking and improves as the day goes on. It is generally worse with recumbency. Visual obscurations are frequent. The diagnosis is relatively straightforward when papilledema is present, but the possibility must be considered even in patients without funduscopic changes. Formal visual field testing should be performed even in the absence of overt ophthalmic involvement. Headache on rising in the morning or nocturnal headache is also characteristic of obstructive sleep apnea or poorly controlled hypertension.

Evaluation of patients suspected to have raised CSF pressure requires brain imaging. It is most efficient to obtain an MRI, including an MR venogram, as the initial study. If there are no contraindications, the CSF pressure should be measured by LP; this should be done when the patient is symptomatic so that both the pressure and the response to removal of 20–30 mL of CSF can be determined. An elevated opening pressure and improvement in headache following removal of CSF are diagnostic in the absence of fundal changes.

Initial treatment is with acetazolamide (250–500 mg bid); the headache may improve within weeks. If ineffective, topiramate is the next treatment of choice; it has many actions that may be useful in this setting, including carbonic anhydrase inhibition, weight loss, and neuronal membrane stabilization, likely mediated via effects on phosphorylation pathways. Severely disabled patients who do not respond to medical treatment require intracranial pressure monitoring and may require shunting.

Posttraumatic Headache

A traumatic event can trigger a headache process that lasts for many months or years after the event. The term *trauma* is used here in a very broad sense: headache can develop following an injury to the head, but it can also develop after an infectious episode, typically viral meningitis, a flulike illness, or a parasitic infection. Complaints of dizziness, vertigo, and impaired memory can accompany the headache. Symptoms may remit after several weeks or persist for months and even years after the injury. Typically the neurologic examination is normal and CT or MRI studies are unrevealing. Chronic subdural hematoma may on occasion mimic this disorder. Posttraumatic headache may also be seen after carotid dissection and subarachnoid hemorrhage and after intracranial surgery. The underlying theme appears to be that a traumatic event involving the pain-producing meninges can trigger a headache process that lasts for many years.

Other Causes

In one series, one-third of patients with NDPH reported headache beginning after a transient flulike illness characterized by fever, neck stiffness, photophobia, and marked malaise. Evaluation typically reveals no apparent cause for the headache. There is no convincing evidence that persistent Epstein-Barr virus infection plays a role in NDPH. A complicating factor is that many patients undergo LP during the acute illness; iatrogenic low CSF volume headache must be considered in these cases.

Treatment

Treatment is largely empirical and directed at the headache phenotype. Tricyclic antidepressants, notably amitriptyline, and anticonvulsants, such as topiramate, valproate, and gabapentin, have been used with reported benefit. The monoamine oxidase inhibitor phenelzine may also be useful in carefully selected patients. The headache usually resolves within 3–5 years, but it can be quite disabling.

TABLE 13-4

Differential Diagnosis of New Daily Persistent Headache

Primary	Secondary
Migrainous-type	Subarachnoid hemorrhage
Featureless (tension-type)	Low cerebrospinal fluid (CSF) volume headache
	Raised CSF pressure headache
	Posttraumatic headache[a]
	Chronic meningitis

[a]Includes postinfectious forms.

FIGURE 13-1

Magnetic resonance image showing diffuse meningeal enhancement after gadolinium administration in a patient with low cerebrospinal fluid (CSF) volume headache.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PRIMARY CARE AND HEADACHE MANAGEMENT

Most patients with headache will be seen first in a primary care setting. The task of the primary care physician is to identify the very few worrisome secondary headaches from the very great majority of primary and less troublesome secondary headaches (Table 13-2).

Absent any warning signs, a reasonable approach is to treat when a diagnosis is established. As a general rule, the investigation should focus on identifying worrisome causes of headache or on gaining confidence if no primary headache diagnosis can be made.

After treatment has been initiated, follow-up care is essential to identify whether progress has been made against the headache complaint. Not all headaches will respond to treatment, but, in general, worrisome headaches will progress and will be easier to identify.

When a primary care physician feels the diagnosis is a primary headache disorder, it is worth noting that >90% of patients who present to primary care with a complaint of headache will have migraine (Chap. 422).

In general, patients who do not have a clear diagnosis, have a primary headache disorder other than migraine or tension-type headache, or are unresponsive to two or more standard therapies for the considered headache type should be considered for referral to a specialist. In a practical sense, the threshold

for referral is also determined by the experience of the primary care physician in headache medicine and the availability of secondary care options.

ACKNOWLEDGMENT

The editors acknowledge the contributions of Neil H. Raskin to earlier editions of this chapter.

FURTHER READING

Headache Classification Committee of the International Headache Society: The International Classification of Headache Disorders, 3rd ed. Cephalalgia 33:629, 2018.

Kernick D, Goadsby PJ: *Headache: A Practical Manual*. Oxford: Oxford University Press, 2008.

Lance JW, Goadsby PJ: *Mechanism and Management of Headache*, 7th ed. New York, Elsevier, 2005.

Olesen J et al: *The Headaches*. Philadelphia, Lippincott, Williams & Wilkins, 2005.

Silberstein SD, Lipton RB, Dodick D: *Wolff's Headache and Other Head Pain*, 8th ed. New York, Oxford, 2008.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 14: Back and Neck Pain

John W. Engstrom

INTRODUCTION

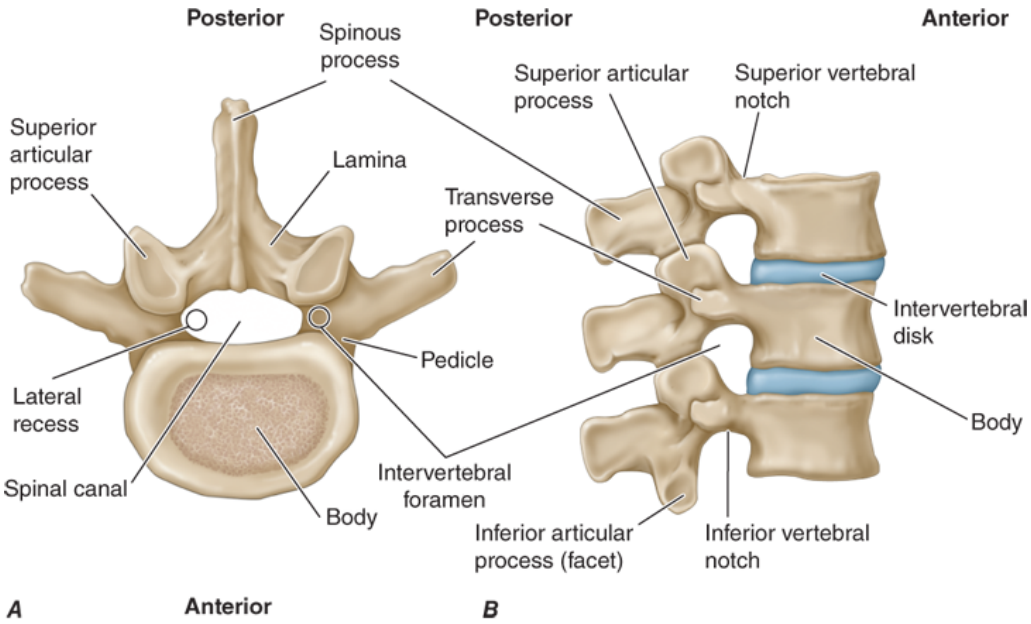
The importance of back and neck pain in our society is underscored by the following: (1) the cost of chronic back pain in the United States is estimated at $177 billion annually; approximately one-third of this cost is due to direct health care expenses and two-thirds are indirect costs resulting from loss of wages and productivity; (2) back symptoms are the most common cause of disability in individuals <45 years of age; (3) low back pain (LBP) is the second most common reason for visiting a physician in the United States; and (4) more than four out of five people will experience significant back pain at some point in their lives.

ANATOMY OF THE SPINE

The anterior spine consists of cylindrical vertebral bodies separated by intervertebral disks and held together by the anterior and posterior longitudinal ligaments. The intervertebral disks are composed of a central gelatinous nucleus pulposus surrounded by a tough cartilaginous ring, the annulus fibrosis. Disks are responsible for 25% of spinal column length and allow the bony vertebrae to move easily upon each other (Figs. 14-1 and 14-2). Desiccation of the nucleus pulposus and degeneration of the annulus fibrosus increase with age, resulting in loss of disk height. The disks are largest in the cervical and lumbar regions where movements of the spine are greatest. The anterior spine absorbs the shock of bodily movements such as walking and running and, with the posterior spine, protects the spinal cord and nerve roots in the spinal canal.

FIGURE 14-1

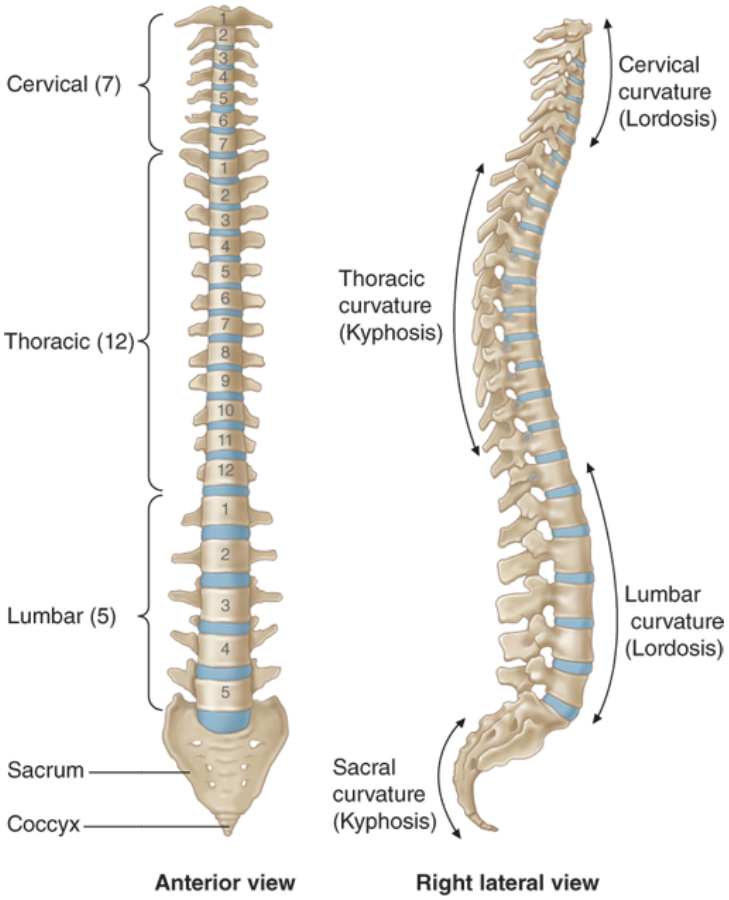
Vertebral anatomy. *(From A Gauthier Cornuelle, DH Gronefeld: Radiographic Anatomy Positioning. New York, McGraw-Hill, 1998; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 14-2

Spinal column. *(From A Gauthier Cornuelle, DH Gronefeld: Radiographic Anatomy Positioning. New York, McGraw-Hill, 1998; with permission.)*



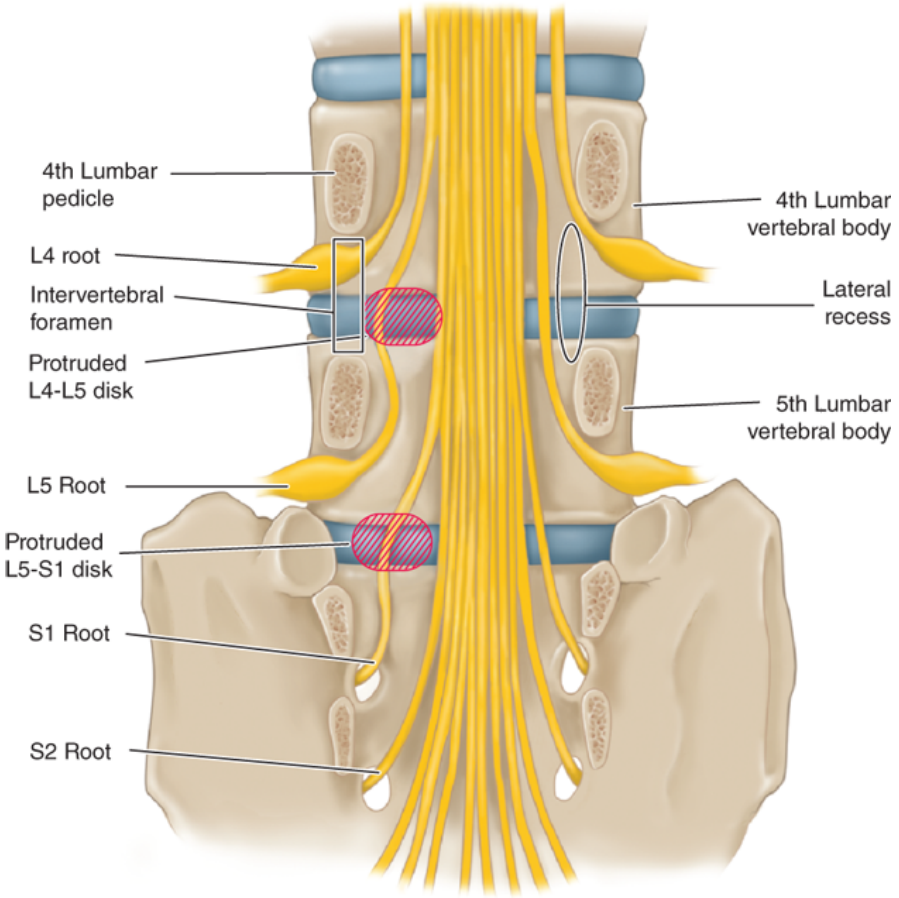
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The posterior spine consists of the vertebral arches and processes. Each arch consists of paired cylindrical pedicles anteriorly and paired lamina posteriorly. The vertebral arch also gives rise to two transverse processes laterally, one spinous process posteriorly, plus two superior and two inferior articular facets. The apposition of a superior and inferior facet constitutes a *facet joint*. The posterior spine provides an anchor for the attachment of muscles and ligaments. The contraction of muscles attached to the spinous and transverse processes and lamina works like a system of pulleys and levers that results in flexion, extension, and lateral bending movements of the spine.

Nerve root injury (*radiculopathy*) is a common cause of neck and arm, or low back and buttock or leg, pain **(see dermatomes in Figs. 22-2 and 22-3)**. The nerve roots exit at a level above their respective vertebral bodies in the cervical region (e.g., the C7 nerve root exits at the C6-C7 level) and below their respective vertebral bodies in the thoracic and lumbar regions (e.g., the T1 nerve root exits at the T1-T2 level). The cervical nerve roots follow a short intraspinal course before exiting. By contrast, because the spinal cord ends at the vertebral L1 or L2 level, the lumbar nerve roots follow a long intraspinal course and can be injured anywhere from the upper lumbar spine to the intervertebral foramen or extraforaminal space. For example, disk herniation at the L4-L5 level can produce L4 root compression laterally, but more often compression of the traversing L5 nerve root **(Fig. 14-3)**. The lumbar nerve roots are mobile in the spinal canal, but eventually pass through the narrow *lateral recess* of the spinal canal and *intervertebral foramen* (Figs. 14-2 and 14-3). Neuroimaging of the spine must include both sagittal and axial views to assess possible compression in either the lateral recess or intervertebral foramen.

FIGURE 14-3

Compression of L5 and S1 roots by herniated disks. *(From AH Ropper, MA Samuels: Adams and Victor's Principles of Neurology, 9th ed. New York, McGraw-Hill, 2009; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Beginning at the C3 level, each cervical (and the first thoracic) vertebral body projects a lateral bony process upward—the uncinate process. The uncinate process articulates with the cervical vertebral body above via the uncovertebral joint. The uncovertebral joint can hypertrophy with age and contribute to neural foraminal narrowing and radiculopathy in the cervical spine.

Pain-sensitive structures of the spine include the periosteum of the vertebrae, dura, facet joints, annulus fibrosus of the intervertebral disk, epidural veins and arteries, and the longitudinal ligaments. Disease of these diverse structures may explain many cases of back pain without nerve root compression. Under normal circumstances, the nucleus pulposus of the intervertebral disk is not pain sensitive.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Back Pain

TYPES OF BACK PAIN

Delineating the type of pain reported by the patient is the essential first step. Attention is also focused on identification of risk factors for a serious underlying etiology. The most frequent serious causes of back pain are radiculopathy, fracture, tumor, infection, or referred pain from visceral structures (Table 14-1).

Local pain is caused by injury to pain-sensitive structures that compress or irritate sensory nerve endings. The site of the pain is near the affected part of the back.

Pain referred to the back may arise from abdominal or pelvic viscera. The pain is usually described as primarily abdominal or pelvic, accompanied by back pain and usually unaffected by posture. The patient may occasionally complain of back pain only.

Pain of spine origin may be located in the back or referred to the buttocks or legs. Diseases affecting the upper lumbar spine tend to refer pain to the lumbar region, groin, or anterior thighs. Diseases affecting the lower lumbar spine tend to produce pain referred to the buttocks, posterior thighs, calves, or feet. Referred pain can explain pain syndromes that cross multiple dermatomes without evidence of nerve or nerve root injury.

Radicular pain is typically sharp and radiates from the low back to a leg within the territory of a nerve root (see "Lumbar Disk Disease," below). Coughing, sneezing or voluntary contraction of abdominal muscles (lifting heavy objects or straining at stool) may elicit or worsen the radiating pain. The pain may increase in postures that stretch the nerves and nerve roots. Sitting with the leg outstretched places traction on the sciatic nerve and L5 and S1 roots because the sciatic nerve passes posterior to the hip. The femoral nerve (L2, L3, and L4 roots) passes anterior to the hip and is not stretched by sitting. The description of the pain alone often fails to distinguish between referred pain and radiculopathy, although a burning or electric quality favors radiculopathy.

Pain associated with muscle spasm is commonly associated with many spine disorders. The spasms may be accompanied by an abnormal posture, tense paraspinal muscles, and dull or achy pain in the paraspinal region.

Knowledge of the circumstances associated with the onset of back pain is important when weighing possible serious underlying causes for the pain. Some patients involved in accidents or work-related injuries may exaggerate their pain for the purpose of compensation or for psychological reasons.

EXAMINATION

A physical examination that includes the abdomen and rectum is advisable. Back pain referred from visceral organs may be reproduced during palpation of the abdomen (pancreatitis, abdominal aortic aneurysm [AAA]) or percussion over the costovertebral angles (pyelonephritis).

The normal spine has a cervical and lumbar lordosis and a thoracic kyphosis. Exaggeration of these normal alignments may result in hyperkyphosis of the thoracic spine or hyperlordosis of the lumbar spine. Inspection may reveal a lateral curvature of the spine (scoliosis). An asymmetry in the prominence of the paraspinal muscles suggests muscle spasm. Spine pain reproduced by palpation over the spinous process reflects injury of the affected vertebrae or adjacent pain-sensitive structures.

Forward bending is often limited by paraspinal muscle spasm; the latter may flatten the usual lumbar lordosis. Flexion at the hips is normal in patients with lumbar spine disease, but flexion of the lumbar spine is limited and sometimes painful. Lateral bending to the side opposite the injured spinal element may stretch the damaged tissues, worsen pain, and limit motion. Hyperextension of the spine (with the patient prone or standing) is limited when nerve root compression, facet joint pathology, or other bony spine disease is present.

Pain from hip disease may mimic the pain of lumbar spine disease. Hip pain can be reproduced by passive internal and external rotation at the hip with the knee and hip in flexion or by percussing the heel with the examiner's palm with the leg extended (heel percussion sign).

The *straight leg-raising (SLR)* maneuver is a simple bedside test for nerve root disease. With the patient supine, passive straight leg flexion at the hip stretches the L5 and S1 nerve roots and the sciatic nerve; dorsiflexion of the foot during the maneuver adds to the stretch. In healthy individuals, flexion to at least 80° is normally possible without causing pain, although a tight, stretching sensation in the hamstring muscles is common. The SLR test is positive if the maneuver reproduces the patient's usual back or limb pain. Eliciting the SLR sign in both the supine and sitting positions can help determine if the finding is reproducible. The patient may describe pain in the low back, buttocks, posterior thigh, or lower leg, but the *key feature is reproduction of the patient's usual pain*. The *crossed SLR sign* is present when flexion of one leg reproduces the usual pain in the opposite leg or buttocks. In disk herniation, the crossed SLR sign is less sensitive but more specific than the SLR sign. The *reverse SLR sign* is elicited by standing the patient next to the examination table and passively extending each leg with the knee fully extended. This maneuver, which stretches the L2-L4 nerve roots, lumbosacral plexus, and femoral nerve, is considered positive if the patient's usual back or limb pain is reproduced. For all of these tests, the nerve or nerve root lesion is always on the side of the pain.

The neurologic examination includes a search for focal weakness or muscle atrophy, focal reflex changes, diminished sensation in the legs, or signs of spinal cord injury. The examiner should be alert to the possibility of breakaway weakness, defined as fluctuations in the maximum power generated during muscle testing. Breakaway weakness may be due to pain, inattention, or a combination of pain and underlying true weakness. Breakaway weakness without pain is usually due to a lack of effort. In uncertain cases, electromyography (EMG) can determine if true weakness due to nerve tissue injury is present. Findings with specific lumbosacral nerve root lesions are shown in **Table 14-2** and are discussed below.

LABORATORY, IMAGING, AND EMG STUDIES

Laboratory studies are rarely needed for the initial evaluation of nonspecific acute (<3 months in duration) low back pain (ALBP). Risk factors for a serious underlying cause and for infection, tumor, or fracture, in particular, should be sought by history and examination. If risk factors are present (Table 14-1), then laboratory studies (complete blood count [CBC], erythrocyte sedimentation rate [ESR], urinalysis) are indicated. If risk factors are absent, then management is conservative (see "Treatment," below).

Computed tomography (CT) scanning is superior to x-rays for detection of fractures involving posterior spine structures, craniocervical and cervicothoracic junctions, C1 and C2 vertebrae, bone fragments in the spinal canal, or misalignment. CT scans are increasingly used as a primary screening modality for moderate to severe acute trauma. Magnetic resonance imaging (MRI) or CT myelography is the radiologic test of choice for evaluation of most serious diseases involving the spine. MRI is superior for the definition of soft tissue structures, whereas CT myelography provides optimal imaging of the lateral recess of the spinal canal, defines bony abnormalities, and is tolerated by claustrophobic patients.

Population surveys in the United States suggest that patients with back pain report greater functional limitations in recent years, despite rapid increases in spine imaging, opioid prescribing, injections, and spine surgery. This suggests that more selective use of diagnostic and treatment modalities may be reasonable for many patients.

Spine imaging often reveals abnormalities of dubious clinical relevance that may alarm clinicians and patients alike and prompt further testing and unnecessary therapy. When imaging tests are reported, it is important to remember that degenerative findings are common in normal, pain-free individuals. Randomized trials and observational studies have suggested that imaging can have a “cascade effect”, creating a gateway to other unnecessary care. Based in part on such evidence, the American College of Physicians and the North American Spine Society have partnered to make parsimonious use of spine imaging a high priority in the “Choosing Wisely” campaign, aimed at reducing unnecessary spine care. Successful efforts to reduce unnecessary imaging have typically been multifaceted. Some include physician education and computerized decision support to identify prior imaging examinations and to require specific indications for approval of imaging tests. Other strategies have included audit and feedback of individual practitioners’ rates of ordering, and more rapid access to physical therapy or expert consultation for patients without imaging indications.

For example, educational tools for patients and the public have included “Five Things Physicians and Patients Should Question”: (1) Do not recommend advanced imaging (e.g., MRI) of the spine within the first 6 weeks in patients with nonspecific ALBP in the absence of red flags. (2) Do not perform elective spinal injections without imaging guidance, unless contraindicated. (3) Do not use bone morphogenetic protein (BMP) for routine anterior cervical spine fusion surgery. (4) Do not use EMG and nerve conduction studies (NCSs) to determine the cause of axial lumbar, thoracic or cervical spine pain. (5) Do not recommend bed rest for >48 h when treating LBP. In an observational study, application of this strategy was associated with lower rates of repeat imaging, opioid use, and referrals for physical therapy.

Electrodiagnostic studies can be used to assess the functional integrity of the peripheral nervous system **(Chap. 438)**. Sensory NCSs are normal when focal sensory loss confirmed by examination is due to nerve root damage because the nerve roots are proximal to the nerve cell bodies in the dorsal root ganglia. Injury to nerve tissue distal to the dorsal root ganglion (e.g., plexus or peripheral nerve) results in reduced sensory nerve signals. Needle EMG complements NCSs by detecting denervation or reinnervation changes in a myotomal (segmental) distribution. Multiple muscles supplied by different nerve roots and nerves are sampled; the pattern of muscle involvement indicates the nerve root(s) responsible for the injury. Needle EMG provides objective information about motor nerve fiber injury when clinical evaluation of weakness is limited by pain or poor effort. EMG and NCSs will be normal when sensory nerve root injury or irritation is the pain source.

TABLE 14-1

Acute Low Back Pain: Risk Factors for an Important Structural Cause

History
Pain worse at rest or at night
Prior history of cancer
History of chronic infection (especially lung, urinary tract, skin)
History of trauma
Incontinence
Age >70 years
Intravenous drug use
Glucocorticoid use
History of a rapidly progressive neurologic deficit
Examination
Unexplained fever
Unexplained weight loss
Palpation/percussion tenderness over the midline spine
Abdominal, rectal, or pelvic mass
Internal/external rotation of the leg at the hip; heel percussion sign
Straight leg- or reverse straight leg-raising signs
Progressive focal neurologic deficit

TABLE 14-2

Lumbosacral Radiculopathy: Neurologic Features

Lumbosacral Nerve Roots	Examination Findings			Pain Distribution
	Reflex	Sensory	Motor	
L2[a]	—	Upper anterior thigh	Psoas (hip flexors)	Anterior thigh
L3[a]	—	Lower anterior thigh	Psoas (hip flexors)	Anterior thigh, knee
		Anterior knee	Quadriceps (knee extensors)	
			Thigh adductors	
L4[a]	Quadriceps (knee)	Medial calf	Quadriceps (knee extensors)[b]	Knee, medial calf
			Thigh adductors	Anterolateral thigh
L5[c]	—	Dorsal surface—foot	Peronei (foot evertors)[b]	Lateral calf, dorsal foot, posterolateral thigh, buttocks
		Lateral calf	Tibialis anterior (foot dorsiflexors)	
			Gluteus medius (leg abductors)	
			Toe dorsiflexors	
S1[c]	Gastrocnemius/soleus (ankle)	Plantar surface—foot	Gastrocnemius/soleus (foot plantar flexors)[b]	Bottom foot, posterior calf, posterior thigh, buttocks
		Lateral aspect—foot	Abductor hallucis (toe flexors)[b]	
			Gluteus maximus (leg extensors)	

[a]Reverse straight leg-raising sign present—see "Examination of the Back." [b]These muscles receive the majority of innervation from this root. [c]Straight leg-raising sign present—see "Examination of the Back."

CAUSES OF BACK PAIN

Table 14-3

TABLE 14-3

Causes of Back or Neck Pain

Lumbar or Cervical Disk Disease
Degenerative Spine Disease
Lumbar spinal stenosis without or with neurogenic claudication Intervertebral foraminal or lateral recess narrowing Disk-osteophyte complex Facet or uncovertebral joint hypertrophy Lateral disk protrusion Spondylosis (osteoarthritis) and spondylolisthesis
Spine Infection
Vertebral osteomyelitis Spinal epidural abscess Septic disk (diskitis) Meningitis Lumbar arachnoiditis
Neoplasms—Metastatic, Hematologic, Primary Bone Tumors, Fractures
Trauma/falls, motor vehicle accidents Atraumatic fractures: osteoporosis, neoplastic infiltration, osteomyelitis
Minor Trauma
Strain or sprain Whiplash injury
Metabolic Spine Disease
Osteoporosis—hyperparathyroidism, immobility Osteosclerosis (e.g., Paget's disease)
Congenital/Developmental
Spondylolysis Kyphoscoliosis Spina bifida occulta Tethered spinal cord
Autoimmune Inflammatory Arthritis
Other Causes of Back Pain Referred pain from visceral disease (e.g., abdominal aortic aneurysm) Postural Psychiatric, malingering, chronic pain syndromes

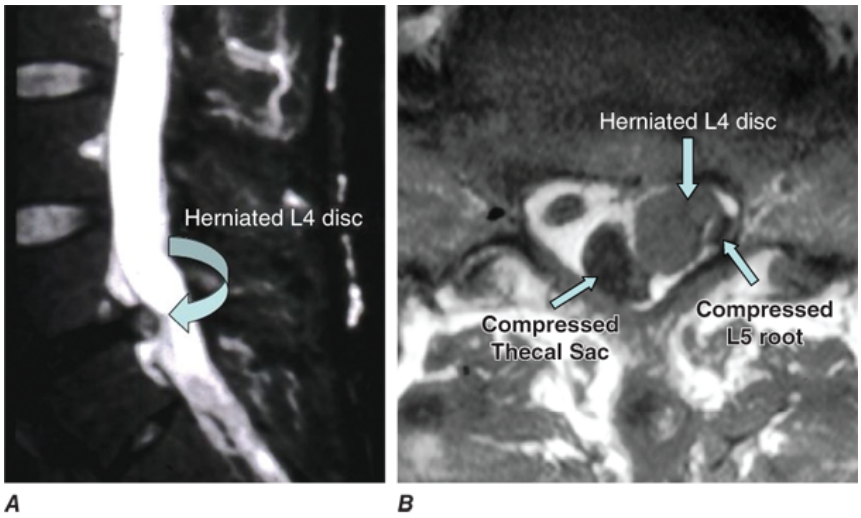
LUMBAR DISK DISEASE

This is a common cause of acute, chronic, or recurrent low back and leg pain (**Figs. 14-3** and **14-4**). Disk disease is most likely to occur at the L4-L5 or L5-S1 levels, but upper lumbar levels can also be involved. The cause is often unknown, but the risk is increased in overweight individuals. Disk herniation is unusual prior to age 20 years and is rare in the fibrotic disks of the elderly. Complex genetic factors may play a role in predisposition. The pain may be located in the low back only or referred to a leg, buttock, or hip. A sneeze, cough, or trivial movement may cause the nucleus pulposus to prolapse, pushing

the frayed and weakened annulus posteriorly. With severe disk disease, the nucleus can protrude through the annulus (herniation) or become extruded to lie as a free fragment in the spinal canal.

FIGURE 14-4

Left L5 radiculopathy. ***A.*** Sagittal T2-weighted image on the left reveals disk herniation at the L4-L5 level. ***B.*** Axial T1-weighted image shows paracentral disk herniation with displacement of the thecal sac medially and the left L5 nerve root posteriorly in the left lateral recess.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The mechanism by which intervertebral disk injury causes back pain is uncertain. The inner annulus fibrosus and nucleus pulposus are normally devoid of innervation. Inflammation and production of proinflammatory cytokines within a ruptured nucleus pulposus may trigger or perpetuate back pain. Ingrowth of nociceptive (pain) nerve fibers into the nucleus pulposus of a diseased disk may be responsible for some cases of chronic "diskogenic" pain. Nerve root injury (radiculopathy) from disk herniation is usually due to inflammation, but lateral herniation may produce compression in the lateral recess or at the intervertebral foramen.

A ruptured disk may be asymptomatic or cause back pain, limited spine motion (particularly flexion), a focal neurologic deficit, or radicular pain. A dermatomal pattern of sensory loss or a reduced or absent deep tendon reflex is more suggestive of a specific root lesion than is the pattern of pain. Motor findings (focal weakness, muscle atrophy, or fasciculations) occur less frequently than focal sensory or reflex changes. Symptoms and signs are usually unilateral, but bilateral involvement does occur with large central disk herniations that compress multiple roots or cause inflammation of nerve roots within the spinal canal. Clinical manifestations of specific nerve root lesions are summarized in Table 14-2.

The differential diagnosis covers a variety of serious and treatable conditions, including epidural abscess, hematoma, fracture, or tumor. Fever, constant pain uninfluenced by position, sphincter abnormalities, or signs of spinal cord disease suggest an etiology other than lumbar disk disease. Absence of ankle reflexes can be a normal finding in persons >60 years or a sign of bilateral S1 radiculopathy. An absent deep tendon reflex or focal sensory loss may indicate injury to a nerve root, but other sites of injury along the nerve must also be considered. For example, an absent knee reflex may be due to a femoral neuropathy or an L4 nerve root injury, and a loss of sensation over the foot and lateral lower calf may result from a peroneal or lateral sciatic neuropathy or an L5 nerve root injury. Focal muscle atrophy may reflect injury to the anterior horn cells of the spinal cord, a nerve root, peripheral nerve, or disuse.

A lumbar spine MRI scan or CT myelogram can often confirm the location and type of pathology. Spine MRIs yield exquisite views of intraspinal and adjacent soft tissue anatomy, whereas bony lesions of the lateral recess or intervertebral foramen are optimally visualized by CT myelography. The correlation of neuroradiologic findings to clinical symptoms, particularly pain, is not simple. Contrast-enhancing tears in the annulus fibrosus or disk protrusions are widely accepted as common sources of back pain; however, studies have found that many asymptomatic adults have similar findings. Entirely asymptomatic disk protrusions are also common, occurring in up to one-third of adults, and these may also enhance with contrast. Furthermore, in patients with known disk herniation treated either medically or surgically, persistence of the herniation 10 years later had no relationship to the clinical outcome. In summary, MRI findings of disk protrusion, tears in the annulus fibrosus, or hypertrophic facet joints are common incidental findings that, by themselves, should not dictate management decisions for patients with back pain.

The diagnosis of nerve root injury is most secure when the history, examination, results of imaging studies, and the EMG are concordant. There is often good correlation between CT and EMG for localization of nerve root injury.

Management of lumbar disk disease is discussed below.

Cauda equina syndrome (CES) signifies an injury of multiple lumbosacral nerve roots within the spinal canal distal to the termination of the spinal cord at L1-L2. LBP, weakness and areflexia in the legs, saddle anesthesia, or loss of bladder function may occur. The problem must be distinguished from disorders of the lower spinal cord (conus medullaris syndrome), acute transverse myelitis **(Chap. 434)**, and Guillain-Barré syndrome **(Chap. 439)**. Combined

involvement of the conus medullaris and cauda equina can occur. CES is most commonly due to a large ruptured lumbosacral intervertebral disk, but other causes include lumbosacral spine fracture, hematoma within the spinal canal (sometimes following lumbar puncture in patients with coagulopathy), and tumor or other compressive mass lesions. Treatment is surgical decompression, sometimes on an urgent basis in an attempt to restore or preserve motor or sphincter function, or radiotherapy for metastatic tumors **(Chap. 86)**.

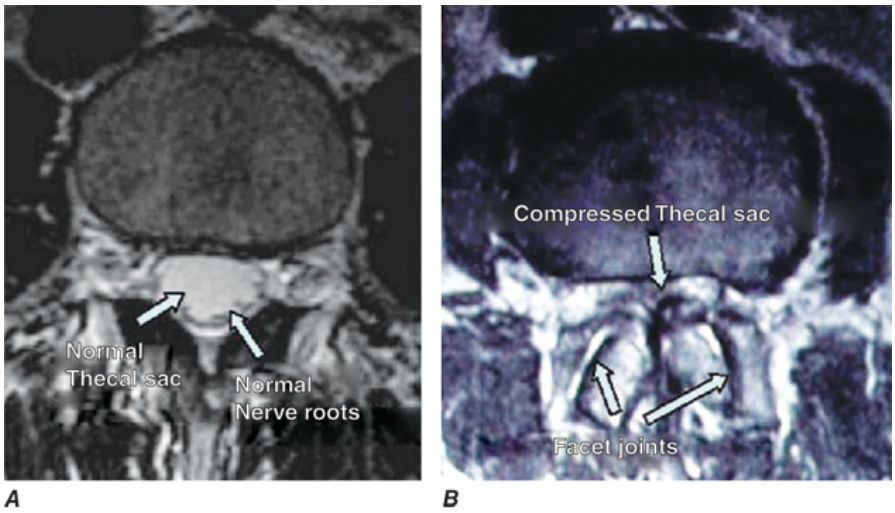
DEGENERATIVE CONDITIONS

Lumbar spinal stenosis (*LSS*) describes a narrowed lumbar spinal canal. *Neurogenic claudication* consists of pain, typically in the back and buttock or leg, that is brought on by walking or standing and relieved by sitting. Symptoms in the legs are usually bilateral. Unlike vascular claudication, symptoms are often provoked by standing without walking. Unlike lumbar disk disease, symptoms are usually relieved by sitting. Patients with neurogenic claudication can often walk much farther when leaning over a shopping cart and can pedal a stationary bike with ease while sitting. These flexed positions increase the anteroposterior spinal canal diameter and reduce intraspinal venous hypertension, producing pain relief. Focal weakness, sensory loss, or reflex changes may occur when spinal stenosis is associated with neural foraminal narrowing and radiculopathy. Severe neurologic deficits, including paralysis and urinary incontinence, occur only rarely.

LSS by itself is common (6–7% of adults) and is frequently asymptomatic. The correlation between the severity of symptoms and the degree of spinal canal stenosis is variable. LSS is most often acquired (75%), but can also be congenital or due to a mixture of both. Congenital forms (achondroplasia and idiopathic) are characterized by short, thick pedicles that produce both spinal canal and lateral recess stenosis. Acquired factors that contribute to spinal stenosis include degenerative diseases (spondylosis, spondylolisthesis, and scoliosis), trauma, spine surgery, metabolic or endocrine disorders (epidural lipomatosis, osteoporosis, acromegaly, renal osteodystrophy, and hypoparathyroidism), and Paget's disease. MRI provides the best definition of the abnormal anatomy **(Fig. 14-5)**.

FIGURE 14-5

Axial T2-weighted images of the lumbar spine. ***A***. The image shows a normal thecal sac within the lumbar spinal canal. The thecal sac is bright. The lumbar roots are dark punctate dots in the posterior thecal sac with the patient supine. ***B***. The thecal sac is not well visualized due to severe lumbar spinal canal stenosis, partially the result of hypertrophic facet joints.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

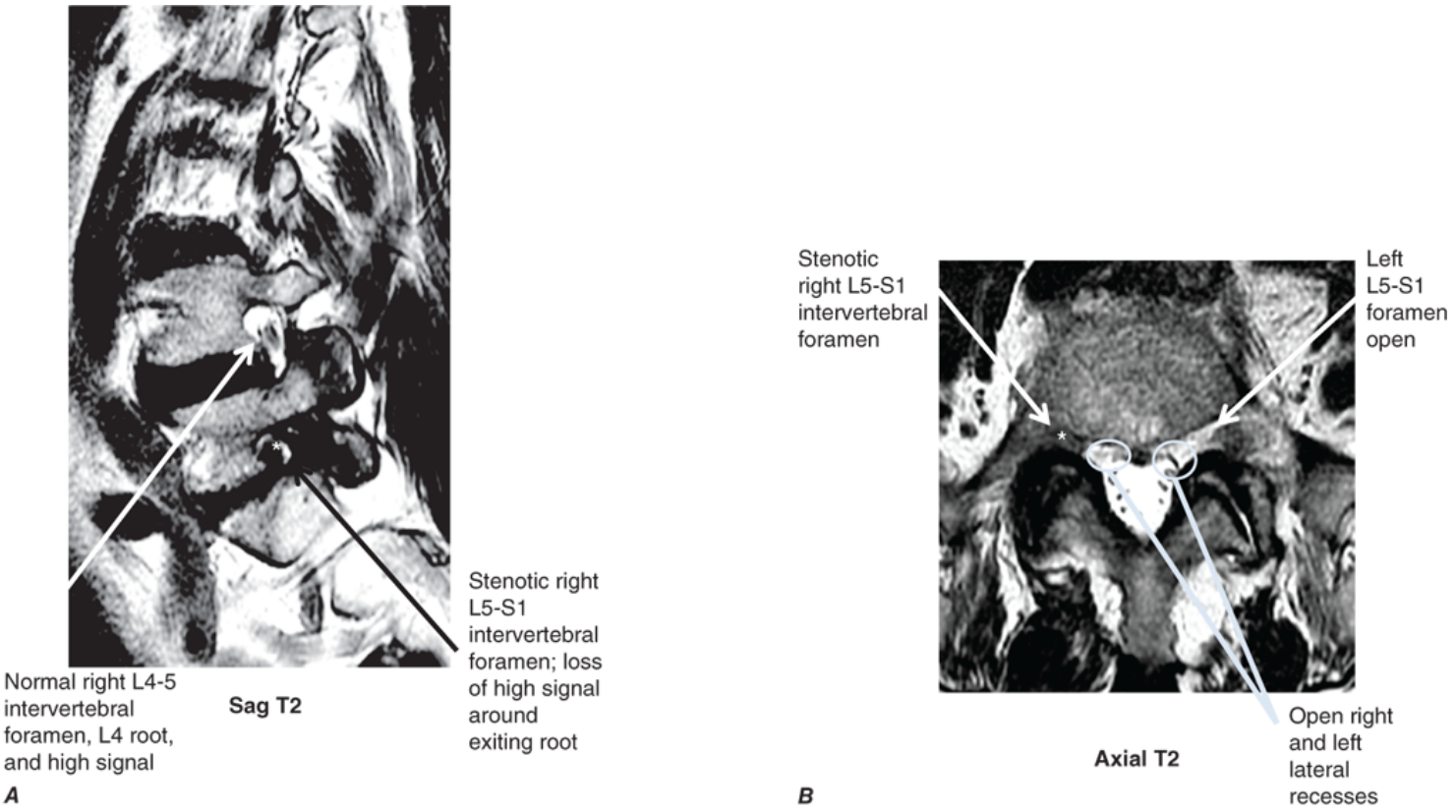
LSS accompanied by neurogenic claudication responds to surgical decompression of the stenotic segments. The same processes leading to LSS may cause lumbar foraminal or lateral recess narrowing resulting in coincident lumbar radiculopathy that may require treatment as well. A recent trial for LSS accompanied by leg pain did not show an overall benefit for epidural glucocorticoids plus lidocaine, but subgroup analysis showed a small improvement in disability scores at 6 weeks of uncertain clinical significance.

Conservative treatment of symptomatic LSS can include nonsteroidal anti-inflammatory drugs (NSAIDs), acetaminophen, exercise programs, and symptomatic treatment of acute pain episodes. There is insufficient evidence to support the routine use of epidural glucocorticoid injections. Surgical therapy is considered when medical therapy does not relieve symptoms sufficiently to allow for resumption of activities of daily living or when focal neurologic signs are present. Most patients with neurogenic claudication who are treated medically do not improve over time. Surgical management with laminectomy can produce significant relief of exertional back and leg pain, leading to less disability and improved functional outcome at 4 years. Laminectomy and fusion is usually reserved for patients with LSS and spondylolisthesis. Predictors of a poor surgical outcome include impaired walking preoperatively, depression, cardiovascular disease, and scoliosis. Up to one-quarter of surgically treated patients develop recurrent stenosis at the same spinal level or at an adjacent level within 7–10 years; recurrent symptoms usually respond to a second surgical decompression.

Neural foraminal narrowing with radiculopathy is a common consequence of osteoarthritic processes that cause LSS (Figs. 14-1 and 14-6), including osteophytes, lateral disk protrusion, calcified disk-osteophytes, facet joint hypertrophy, uncovertebral joint hypertrophy (in the cervical spine), congenitally shortened pedicles, or, frequently, a combination of these processes. Neoplasms (primary or metastatic), fractures, infections (epidural abscess), or hematomas are other less common causes. Most common is bony foraminal narrowing leading to nerve root ischemia and persistent symptoms, in contrast to the inflammation associated with a herniated disk and radiculopathy. These conditions can produce unilateral nerve root symptoms or signs due to compression at the intervertebral foramen or in the lateral recess; symptoms are indistinguishable from disk-related radiculopathy, but treatment may differ depending on the specific etiology. The history and neurologic examination alone cannot distinguish between these possibilities. Neuroimaging (CT or MRI) is required to identify the anatomic cause. Neurologic findings from the examination and EMG can help direct the attention of the radiologist to specific nerve roots, especially on axial images. For *facet joint hypertrophy,* surgical foraminotomy produces long-term relief of leg and back pain in 80–90% of patients. Facet joint blocks for back or neck pain are sometimes used to help determine the anatomic origin of back pain or for treatment, but there is a lack of clinical data to support their utility. Medical causes of lumbar or cervical radiculopathy unrelated to anatomic spine disease include infections (e.g., herpes zoster and Lyme disease), carcinomatous meningitis, and root avulsion or traction (trauma).

FIGURE 14-6

Right L5 radiculopathy. ***A.*** Sagittal T2-weighted image. There is normal high signal around the exiting right L4 nerve root in the right neural foramen at L4-L5; effacement of the high signal in the right L5-S1 foramen is present one level caudal on the right at L5-S1. ***B.*** Axial T2-weighted image. The lateral recesses are normal bilaterally; the intervertebral foramen is normal on the left, but severely stenotic on the right. *Severe right L5-S1 foraminal stenosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SPONDYLOSIS AND SPONDYLOLISTHESIS

Spondylosis, or osteoarthritic spine disease, typically occurs in later life and primarily involves the cervical and lumbosacral spine. Patients often complain of back pain that increases with movement, is associated with stiffness, and is better with inactivity. The relationship between clinical symptoms and radiologic findings is usually not straightforward. Pain may be prominent when x-ray, CT, or MRI findings are minimal, and prominent degenerative spine disease can be seen in asymptomatic patients. Osteophytes, combined disk-osteophytes, or thickened ligamentum flavum may cause or contribute to central spinal canal stenosis, lateral recess stenosis, or neural foraminal narrowing.

Spondylolisthesis is the anterior slippage of the vertebral body, pedicles, and superior articular facets, leaving the posterior elements behind. Spondylolisthesis can be associated with spondylolysis, congenital anomalies, degenerative spine disease, or other causes of mechanical weakness of the pars interarticularis (e.g., infection, osteoporosis, tumor, trauma, prior surgery). The slippage may be asymptomatic or may cause LBP and hamstring tightness, nerve root injury (the L5 root most frequently), symptomatic spinal stenosis, or CES in severe cases. A “step-off” on palpation or tenderness may be elicited near the segment that has “slipped” forward (most often L4 on L5 or occasionally L5 on S1). Focal anterolisthesis or retrolisthesis can occur at any cervical or lumbar level and be the source of neck or LBP. Plain x-rays of the neck or low back in flexion and extension will reveal movement at the

abnormal spinal segment. Surgery is performed for spinal instability (slippage 5–8 mm) and considered for pain symptoms that do not respond to conservative measures (e.g., rest, physical therapy), cases with progressive neurologic deficit, or scoliosis.

NEOPLASMS

Back pain is the most common neurologic symptom in patients with systemic cancer and is the presenting symptom in 20%. The cause is usually vertebral body metastasis (85–90%) but can also result from spread of cancer through the intervertebral foramen (especially with lymphoma), carcinomatous meningitis, or metastasis to the spinal cord. The thoracic spine is most often affected. Cancer-related back pain tends to be constant, dull, unrelieved by rest, and worse at night. By contrast, mechanical causes of LBP usually improve with rest. MRI, CT, and CT myelography are the studies of choice when spinal metastasis is suspected. Once a metastasis is found, imaging of the entire spine is essential, as it reveals additional tumor deposits in one-third of patients. MRI is preferred for soft tissue definition, but the most rapidly available imaging modality is best because the patient's condition may worsen quickly without intervention. Early diagnosis is crucial. A strong predictor of outcome is the baseline neurologic function prior to diagnosis. Half to three quarters of patients are nonambulatory at the time of diagnosis and few regain the ability to walk. **The management of spinal metastasis is discussed in detail in Chap. 86.**

INFECTIONS/INFLAMMATION

Vertebral osteomyelitis is most often caused by hematogenous seeding of staphylococci, but other bacteria or tuberculosis (Pott's disease) may be responsible. The primary source of infection is usually the skin or urinary tract; IV drug use, poor dentition, endocarditis, pulmonary disease, IV catheters, or post-operative wound sites may also be responsible. Back pain at rest, tenderness over the involved vertebra, and an elevated ESR or CRP are the most common findings in vertebral osteomyelitis. Fever or an elevated white blood cell count is found in a minority of patients. MRI and CT are sensitive and specific for early detection of osteomyelitis. The intervertebral disk can also be affected by infection (diskitis) and almost never by tumor. Extension of the infection posteriorly from the vertebra can produce a spinal epidural abscess.

Spinal epidural abscess **(Chap. 434)** presents with back pain (aggravated by movement or spinous process palpation), fever, radiculopathy, or signs of spinal cord compression. The subacute development of two or more of these findings should increase the index of suspicion for spinal epidural abscess. The abscess is best delineated by spine MRI and may track over multiple spinal levels.

Lumbar adhesive arachnoiditis with radiculopathy is due to fibrosis following inflammation within the subarachnoid space. The fibrosis results in nerve root adhesions and presents as back and leg pain associated with multifocal motor, sensory, or reflex changes. Causes of arachnoiditis include multiple lumbar operations (most common in the United States), chronic spinal infections (especially tuberculosis in the developing world), spinal cord injury, intrathecal hemorrhage, myelography (rare), intrathecal injections (glucocorticoids, anesthetics, or other agents), and foreign bodies. The MRI shows clumped nerve roots on axial views or loculations of cerebrospinal fluid within the thecal sac. Clumped nerve roots alone are not diagnostic and may also occur with demyelinating polyneuropathy or neoplastic infiltration. Treatment is usually unsatisfactory. Microsurgical lysis of adhesions, dorsal rhizotomy, dorsal root ganglionectomy, and epidural glucocorticoids have been tried, but outcomes have been poor. Dorsal column stimulation for pain relief has produced varying results.

TRAUMA

A patient complaining of back pain and an inability to move the legs may have a spine fracture or dislocation; with fractures above L1 the spinal cord is at risk for compression. Care must be taken to avoid further damage to the spinal cord or nerve roots by immobilizing the back or neck pending the results of radiologic studies. Vertebral fractures frequently occur in the absence of trauma in association with osteoporosis, glucocorticoid use, osteomyelitis, or neoplastic infiltration.

Sprains and Strains

The terms *low back sprain*, *strain*, and *mechanically induced muscle spasm* refer to minor, self-limited injuries associated with lifting a heavy object, a fall, or a sudden deceleration such as in an automobile accident. These terms are used loosely and do not clearly describe a specific anatomic lesion. The pain is usually confined to the lower back. Patients with paraspinal muscle spasm often assume unusual postures.

Traumatic Vertebral Fractures

Most traumatic fractures of the lumbar vertebral bodies result from injuries producing anterior wedging or compression. With severe trauma, the patient may sustain a fracture-dislocation or a "burst" fracture involving the vertebral body and posterior elements. Traumatic vertebral fractures are caused by falls from a height, sudden deceleration in an automobile accident, or direct injury. Neurologic impairment is common, and early surgical treatment is indicated. In victims of blunt trauma, CT scans of the chest, abdomen, or pelvis can be reformatted to detect associated vertebral fractures. Rules have been developed to avoid unnecessary spine imaging associated with low risk trauma, but these studies excluded patients aged >65—a group that can sustain fractures with minor trauma.

METABOLIC CAUSES

Osteoporosis and Osteosclerosis

Immobilization, osteomalacia, the postmenopausal state, renal disease, multiple myeloma, hyperparathyroidism, hyperthyroidism, metastatic carcinoma, or glucocorticoid use may accelerate osteoporosis and weaken the vertebral body, leading to compression fractures and pain. Up to two-thirds of compression fractures seen on radiologic imaging are asymptomatic. The most common nontraumatic vertebral body fractures are due to postmenopausal or senile osteoporosis **(Chap. 404)**. The risk of an additional vertebral fracture 1 year following a first vertebral fracture is 20%. The presence of fever, weight loss, fracture at a level above T4, any fracture in a young adult, or the predisposing conditions described above should increase suspicion for a cause other than senile osteoporosis. The sole manifestations of a compression fracture may be localized back or radicular pain exacerbated by movement and often reproduced by palpation over the spinous process of the affected vertebra.

Relief of acute pain can often be achieved with acetaminophen, NSAIDs, opioids, or a combination of these medications. Both pain and disability are improved with bracing. Antiresorptive drugs are not recommended in the setting of acute pain, but are the preferred treatment to prevent additional fractures. Less than one-third of patients with prior compression fractures are adequately treated for osteoporosis despite the increased risk for future fractures; even fewer at-risk patients without a history of fracture are adequately treated. The literature for percutaneous vertebroplasty (PVP) or kyphoplasty for osteoporotic compression fractures associated with debilitating pain is mixed, but meta-analyses do not support their utility.

Osteosclerosis, an abnormally increased bone density often due to Paget's disease, is readily identifiable on routine x-ray studies and can sometimes be a source of back pain. It may be associated with an isolated increase in alkaline phosphatase in an otherwise healthy older person. Spinal cord or nerve root compression can result from bony encroachment. The diagnosis of Paget's disease as the cause of a patient's back pain is a diagnosis of exclusion.

For further discussion of these bone disorders, see Chaps. 403, 404, and 405.

AUTOIMMUNE INFLAMMATORY ARTHRITIS

Autoimmune inflammatory disease of the spine can present with the insidious onset of low back, buttock, or neck pain. Examples include rheumatoid arthritis (RA) **(Chap. 351)**, ankylosing spondylitis, reactive arthritis, psoriatic arthritis, or inflammatory bowel disease **(Chaps. 319 and 355)**.

CONGENITAL ANOMALIES OF THE LUMBAR SPINE

Spondylolysis is a bony defect in the vertebral pars interarticularis (a segment near the junction of the pedicle with the lamina); the cause is usually a stress microfracture in a congenitally abnormal segment. It occurs in up to 6% of adolescents. The defect (usually bilateral) is best visualized on plain x-rays or CT scan and is frequently asymptomatic. Symptoms may occur in the setting of a single injury, repeated minor injuries, or during a growth spurt. Spondylolysis is the most common cause of persistent LBP in adolescents and is often associated with sports-related activities.

Scoliosis refers to an abnormal curvature in the coronal (lateral) plane of the spine. With *kyphoscoliosis,* there is, in addition, a forward curvature of the spine. The abnormal curvature may be congenital, due to abnormal spine development, acquired in adulthood due to degenerative spine disease, or occasionally progressive due to neuromuscular disease. The deformity can progress until ambulation or pulmonary function is compromised.

Spina bifida occulta (closed spinal dysraphism) is a failure of closure of one or several vertebral arches posteriorly; the meninges and spinal cord are normal. A dimple or small lipoma may overlie the defect, but the skin is intact. Most cases are asymptomatic and discovered incidentally during an evaluation for back pain.

Tethered cord syndrome usually presents as a progressive cauda equina disorder (see below), although myelopathy may also be the initial manifestation. The patient is often a child or young adult who complains of perineal or perianal pain, sometimes following minor trauma. MRI studies typically reveal a low-lying conus (below L1 and L2) and a short and thickened filum terminale.

REFERRED PAIN FROM VISCERAL DISEASE

Diseases of the thorax, abdomen, or pelvis may refer pain to the spinal segment that innervates the diseased organ. Occasionally, back pain may be the first and only manifestation. Upper abdominal diseases generally refer pain to the lower thoracic or upper lumbar region (eighth thoracic to the first and second lumbar vertebrae), lower abdominal diseases to the midlumbar region (second to fourth lumbar vertebrae), and pelvic diseases to the sacral region. Local signs (pain with spine palpation, paraspinal muscle spasm) are absent, and little or no pain accompanies routine movements.

Low Thoracic or Lumbar Pain with Abdominal Disease

Tumors of the posterior wall of the stomach or duodenum typically produce epigastric pain **(Chaps. 76 and 317)**, but back pain may occur if retroperitoneal extension is present. Fatty foods occasionally induce back pain associated with biliary or pancreatic disease. Pathology in retroperitoneal structures (hemorrhage, tumors, and pyelonephritis) can produce paraspinal pain that radiates to the lower abdomen, groin, or anterior thighs. A mass in the iliopsoas region can produce unilateral lumbar pain with radiation toward the groin, labia, or testicle. The sudden appearance of lumbar pain in a patient receiving anticoagulants suggests retroperitoneal hemorrhage.

Isolated LBP occurs in some patients with a contained rupture of an AAA. The classic clinical triad of abdominal pain, shock, and back pain occurs in <20% of patients. The diagnosis may be missed because the symptoms and signs can be nonspecific. Misdiagnoses include nonspecific back pain, diverticulitis, renal colic, sepsis, and myocardial infarction. A careful abdominal examination revealing a pulsatile mass (present in 50–75% of patients) is an important physical finding. Patients with suspected AAA should be evaluated with abdominal ultrasound, CT, or MRI **(Chap. 274)**.

Sacral Pain with Gynecologic and Urologic Disease

Pelvic organs rarely cause LBP. Uterine malposition (retroversion, descensus, and prolapse) may cause traction on the uterosacral ligament. The pain is referred to the sacral region, sometimes appearing after prolonged standing. Endometriosis or uterine cancers can invade the uterosacral ligaments. Pain associated with endometriosis is typically premenstrual and often continues until it merges with menstrual pain.

Menstrual pain with poorly localized, cramping pain can radiate down the legs. LBP that radiates into one or both thighs is common in the last weeks of pregnancy. Continuous and worsening pain unrelieved by rest or at night may be due to neoplastic infiltration of nerves or nerve roots.

Urologic sources of lumbosacral back pain include chronic prostatitis, prostate cancer with spinal metastasis **(Chap. 83)**, and diseases of the kidney or ureter. Infectious, inflammatory, or neoplastic renal diseases may produce ipsilateral lumbosacral pain, as can renal artery or vein thrombosis. Paraspinal lumbar pain may be a symptom of ureteral obstruction due to nephrolithiasis.

OTHER CAUSES OF BACK PAIN

Postural Back Pain

There is a group of patients with nonspecific chronic low back pain (CLBP) in whom no specific anatomic lesion can be found despite exhaustive investigation. Exercises to strengthen the paraspinal and abdominal muscles are sometimes helpful.

Psychiatric Disease

CLBP may be encountered in patients who seek financial compensation; in malingerers; or in those with concurrent substance abuse. Many patients with CLBP have a history of psychiatric illness (depression, anxiety states) or childhood trauma (physical or sexual abuse) that antedates the onset of back pain. Preoperative psychological assessment has been used to exclude patients with marked psychological impairments that predict a poor surgical outcome from spine surgery.

IDIOPATHIC

The cause of low back pain occasionally remains unclear. Some patients have had multiple operations for disk disease. The original indications for surgery may have been questionable, with back pain only, no definite neurologic signs, or a minor disk bulge noted on CT or MRI. Scoring systems based on neurologic signs, psychological factors, physiologic studies, and imaging studies have been devised to minimize the likelihood of unsuccessful surgery.

GLOBAL CONSIDERATIONS

 While many of the history and examination features described in this chapter apply to all patients, information regarding the global epidemiology and prevalence of LBP is limited. The Global Burden of Diseases Study 2010 reported that LBP ranked #6 overall as a cause of disability-related life years (DALYs), and was the #1 cause overall for total years lived with disability (YLD). These numbers increased substantially from 1990 estimates, and with the aging of the population worldwide, the numbers of individuals suffering from low back pain are expected to increase further in the future. Although rankings for low back pain generally were higher in developed regions of the world, this was not uniformly the case; for example, in North Africa and the Middle East low back pain ranked #2 for DALYs. Another area of uncertainty is the extent to which regional differences exist in terms of the specific etiologies of LBP and how these are managed. For example, the most common cause of arachnoiditis in developing countries is prior spine infection, but in developed countries is multiple lumbar spine surgeries. The longstanding history and acceptance of acupuncture in China may also explain the large number of studies from China regarding the efficacy of acupuncture in many pain settings.

TREATMENT

TREATMENT

Back Pain

Mounting evidence of morbidity from long-term opioid therapy (including overdose, dependency, addiction, falls, fractures, accident risk, and sexual dysfunction) has prompted efforts to reduce its use for chronic pain, including back pain **(Chap. 10)**. Safety may be improved with automated notices for high doses, early refills, prescriptions from multiple pharmacies, and overlapping opioid and benzodiazepine prescriptions. Greater access to alternative treatments for chronic pain, such as tailored exercise programs and cognitive-behavioral therapy (CBT), may also reduce opioid prescribing. Public concern in the United States resulted in passage of the Comprehensive Addiction and Recovery Act of 2016.

The high cost, wide geographic variations, and rapidly increasing rates of spinal fusion surgery have prompted scrutiny regarding the lack of standardization of appropriate indications. Some insurance carriers have begun to limit coverage for the most controversial indications, such as low back pain without radiculopathy. Finally, educating patients and the public about the risks of overtreatment may be necessary.

ALBP WITHOUT RADICULOPATHY

ALBP is defined as pain of <3 months in duration. Full recovery can be expected in >85% of adults with ALBP without leg pain. Most have purely "mechanical" symptoms (i.e., pain that is aggravated by motion and relieved by rest).

The initial assessment excludes serious causes of spine pathology that require urgent intervention, including infection, cancer, or trauma. Risk factors for a serious cause of ALBP are shown in Table 14-1. Laboratory and imaging studies are unnecessary if risk factors are absent. CT, MRI, or plain spine films are rarely indicated in the first month of symptoms unless a spine fracture, tumor, or infection is suspected.

The prognosis of ALBP is generally excellent, however episodes tend to recur, and as many as two-thirds of patients will experience a second episode within 1 year. Most patients do not seek medical care and improve on their own. Even among those seen in primary care, two-thirds report being substantially improved after 7 weeks. Spontaneous improvement can mislead clinicians and patients alike about the efficacy of treatment interventions unless subjected to rigorous prospective trials. Many treatments commonly used in the past are now known to be ineffective, including bed rest and lumbar traction.

Clinicians should reassure and educate patients that improvement is very likely and instruct them in self-care. Satisfaction and the likelihood of follow-up increase when patients are educated about prognosis, treatment methods, activity modifications, and strategies to prevent future exacerbations. Patients who report that they did not receive an adequate explanation for their symptoms are likely to request further diagnostic tests. In general, bed rest should be avoided for relief of severe symptoms or kept to a day or two at most. Several randomized trials suggest that bed rest does not hasten the pace of recovery. In general, the best activity recommendation is for early resumption of normal physical activity, avoiding only strenuous manual labor. Possible advantages of early ambulation for ALBP include maintenance of cardiovascular conditioning, improved bone, cartilage, and muscle strength, and increased endorphin levels. Specific back exercises or early vigorous exercise have not shown benefits for acute back pain. Use of heating pads or blankets is sometimes helpful.

Evidence-based guidelines recommend over-the-counter medicines such as NSAIDs and acetaminophen as first-line options for treatment of ALBP. In otherwise healthy patients, a trial of NSAIDs can be followed by acetaminophen for time-limited periods. In theory, the anti-inflammatory effects of NSAIDs might provide an advantage over acetaminophen to suppress inflammation that accompany many causes of ALBP, but in practice there is no clinical evidence to support the superiority of NSAIDs. The risk of renal and gastrointestinal toxicity with NSAIDs is increased in patients with preexisting medical comorbidities (e.g., renal insufficiency, cirrhosis, prior gastrointestinal hemorrhage, use of anticoagulants or glucocorticoids, heart failure). Some patients elect to take acetaminophen and a NSAID together in hopes of a more rapid benefit. Skeletal muscle relaxants, such as cyclobenzaprine or methocarbamol, may be useful, but sedation is a common side effect. Limiting the use of muscle relaxants to nighttime only may be an option for patients with back pain that interferes with sleep.

There is no good evidence to support the use of opioid analgesics or tramadol as first-line therapy for ALBP. Their use is best reserved for patients who cannot tolerate acetaminophen or NSAIDs and for those with severe refractory pain. As with muscle relaxants, these drugs are often sedating, so it may be useful to prescribe them at nighttime only. Side effects of short-term opioid use include nausea, constipation, and pruritus; risks of long-term opioid use include hypersensitivity to pain, hypogonadism, and dependency. Falls, fractures, driving accidents, and fecal impaction are other risks. Clinical efficacy of opioids for chronic pain beyond 16 weeks of use is unproven.

There is no evidence to support use of oral or injected glucocorticoids, antiepileptics, antidepressants, therapies for neuropathic pain such as gabapentin or herbal therapies. Commonly used nonpharmacologic treatments for ALBP are also of unproven benefit, including spinal manipulation, physical therapy, massage, acupuncture, laser therapy, therapeutic ultrasound, corsets, transcutaneous electrical nerve stimulation (TENS), special mattresses, or lumbar traction. Although important for chronic pain, back exercises for ALBP are generally not supported by clinical evidence. There is no convincing evidence regarding the value of ice or heat applications for ABLP; however, many patients report temporary symptomatic relief from ice or frozen gel packs, and heat may produce a short-term reduction in pain after the first week. Patients often report improved satisfaction with the care that they receive when they actively participate in the selection of symptomatic approaches.

CLBP WITHOUT RADICULOPATHY

CLBP is defined as pain lasting >12 weeks; it accounts for 50% of total back pain costs. Risk factors include obesity, female gender, older age, prior history of back pain, restricted spinal mobility, pain radiating into a leg, high levels of psychological distress, poor self-rated health, minimal physical activity, smoking, job dissatisfaction, and widespread pain. In general, the same treatments that are recommended for ALBP can be useful for patients with CLBP. In this setting, however, the benefit of opioid therapy or muscle relaxants is less clear. In general, activity tolerance is the primary goal, while pain relief is secondary.

Evidence supports the use of exercise therapy to alleviate pain symptoms and improve function. Exercise can be one of the mainstays of treatment for CLBP. Effective regimens have generally included a combination of core strengthening exercises, stretching, and gradually increasing aerobic exercise. A program of supervised exercise can improve compliance. Supervised intensive physical exercise or "work hardening" regimens have been effective in returning some patients to work, improving walking distance, and reducing pain. In addition, some forms of yoga have been evaluated in randomized trials and may be helpful for patients who are interested. A long-term benefit of spinal manipulation or massage for CLBP is unproven.

Medications for CLBP may include short courses of NSAIDs or acetaminophen. Tricyclic antidepressants can provide modest pain relief for some patients without evidence of depression. Trials do not support the efficacy of selective serotonin reuptake inhibitors (SSRIs) for CLBP. However, depression is common among patients with chronic pain and should be appropriately treated.

CBT is based on evidence that psychological and social factors, as well as somatic pathology, are important in the genesis of chronic pain and disability; CBT focuses on efforts to identify and modify patients' thinking about their condition. In one randomized trial, CBT reduced disability and pain in patients with CLBP. Such behavioral treatments appear to provide benefits similar in magnitude to exercise therapy.

Back pain is the most frequent reason for seeking complementary and alternative treatments, most commonly spinal manipulation, acupuncture, and massage. The value of these approaches remains unclear, however. Biofeedback has not been studied rigorously. There is no convincing evidence that either spinal manipulation, TENS, laser therapy, or ultrasound are effective in treating CLBP. Rigorous trials of acupuncture suggest that true acupuncture is not superior to sham acupuncture, but that both may offer an advantage over routine care. Whether this is due entirely to placebo effects provided even by sham acupuncture is uncertain. Some trials of massage therapy have been encouraging for short-term relief only.

Various injections, including epidural glucocorticoid injections, facet joint injections, and trigger point injections, have been used for treating CLBP. However, in the absence of radiculopathy, there is no clear evidence that these approaches are effective.

Injection studies are sometimes used diagnostically to help determine the anatomic source of back pain. Pain relief following a glucocorticoid and anesthetic injection into a facet is commonly used as evidence that the facet joint is the pain source; however, the possibility that the response was a placebo effect or due to systemic absorption of the glucocorticoids is difficult to exclude.

Another category of intervention for CLBP is electrothermal and radiofrequency therapy. Intradiskal therapy has been proposed using both types of energy to thermocoagulate and destroy nerves in the intervertebral disk, using specially designed catheters or electrodes. Current evidence does not support the use of discography to identify a specific disk as the pain source, or the use of intradiskal therapy for CLBP.

Radiofrequency denervation is sometimes used to destroy nerves that are thought to mediate pain, and this technique has been used for facet joint pain (with the target nerve being the medial branch of the primary dorsal ramus), for back pain thought to arise from the intervertebral disk (ramus communicans), and radicular back pain (dorsal root ganglia). A few small trials have produced conflicting results for facet joint and diskogenic pain. A trial in patients with chronic radicular pain found no difference between radiofrequency denervation of the dorsal root ganglia and sham treatment. These interventional therapies have not been studied in sufficient detail to draw firm conclusions regarding their value for CLBP.

Surgical intervention for CLBP without radiculopathy has been evaluated in a number of randomized trials. The case for fusion surgery for CLBP without radiculopathy is weak. While some studies have shown modest benefit, there has been no benefit when compared to an active medical treatment arm, often including highly structured, rigorous rehabilitation combined with CBT. The use of BMP instead of iliac crest graft for the fusion was shown to increase hospital costs and length of stay, but not improve clinical outcomes. Guidelines suggest that referral for an opinion on spinal fusion be considered for people who have completed an optimal nonsurgical treatment program (including combined physical and psychological treatment) and who have persistent severe back pain for which they would consider surgery.

Lumbar disk replacement with prosthetic disks is U.S. Food and Drug Administration approved for uncomplicated patients needing single-level surgery at the L3-S1 levels. The disks are generally designed as metal plates with a polyethylene cushion sandwiched in between. The trials that led to approval of these devices were not blinded. When compared to spinal fusion, the artificial disks were "not inferior." Serious complications are somewhat more likely with the artificial disk. This treatment remains controversial for CLBP.

Intensive multidisciplinary rehabilitation programs can include daily or frequent physical therapy, exercise, CBT, a workplace evaluation, and other interventions. For patients who have not responded to other approaches, such programs appear to offer some benefit. Systematic reviews suggest that the evidence is limited and benefits are limited.

Some observers have raised concerns that CLBP may often be overtreated. For CLBP without radiculopathy, multiple guidelines explicitly recommend against use of SSRIs, any type of injection, TENS, lumbar supports, traction, ulrtaradiofrequency facet joint denervation, intradiskal electrothermal therapy, or intradiskal radiofrequency thermocoagulation. On the other hand, exercise therapy and treatment of depression appear to be useful and underused.

LOW BACK PAIN WITH RADICULOPATHY

A common cause of back pain with radiculopathy is a herniated disk affecting the nerve root and producing back pain with radiation down the leg. The term sciatica is used when the leg pain radiates posteriorly in a sciatic or L5/S1 distribution. The prognosis for acute low back and leg pain with radiculopathy due to disk herniation is generally favorable, with most patients showing substantial improvement over months. Serial imaging studies suggest spontaneous regression of the herniated portion of the disk in two-thirds of patients over 6 months. Nonetheless, there are several important treatment options that provide symptomatic relief while the healing process unfolds.

Resumption of normal activity is recommended. Randomized trial evidence suggests that bed rest is ineffective for treating sciatica as well as back pain alone. Acetaminophen and NSAIDs are useful for pain relief, although severe pain may require short courses of opioid analgesics. Opioids are superior for acute pain relief in the emergency room.

Epidural glucocorticoid injections have a role in providing symptom relief for acute lumbar radiculopathy due to a herniated disk. However, there does not appear to be a benefit in terms of reducing subsequent surgical interventions. A brief course of high dose oral glucocorticoids for 5 days followed by a rapid taper >5 days can be helpful for some patients with acute disk-related radiculopathy, although this specific regimen has not been studied rigorously.

Diagnostic nerve root blocks have been advocated to determine if pain originates from a specific nerve root. However, improvement may result even when the nerve root is not responsible for the pain; this may occur as a placebo effect, from a pain-generating lesion located distally along the peripheral nerve, or from effects of systemic absorption.

Urgent surgery is recommended for patients who have evidence of CES or spinal cord compression, generally manifest as combinations of bowel or bladder dysfunction, diminished sensation in a saddle distribution, a sensory level on the trunk, and bilateral leg weakness or spasticity. Surgical intervention is also indicated for patients with progressive motor weakness due to nerve root injury demonstrated on clinical examination or EMG.

Surgery is also an important option for patients who have disabling radicular pain despite optimal conservative treatment. Because patients with a herniated disk and sciatica generally experience rapid improvement over weeks, most experts do not recommend considering surgery unless the patient has failed to respond to a minimum of 6–8 weeks of nonsurgical management. For patients who have not improved, randomized trials indicate that, compared to nonsurgical treatment, surgery results in more rapid pain relief. However, after 2 years of follow-up, patients appear to have similar pain relief and functional improvement with or without surgery. Thus, both treatment approaches are reasonable, and patient preferences and needs (e.g., rapid return to employment) strongly influence decision making. Some patients will want the fastest possible relief and find surgical risks acceptable. Others will be more risk-averse, more tolerant of symptoms and will choose watchful waiting, especially if they understand that improvement is likely in the end.

The usual surgical procedure is a partial hemilaminectomy with excision of the prolapsed disk (diskectomy). Minimally invasive techniques have gained in popularity in recent years, but preliminary evidence suggests they may be less effective than standard surgical techniques, with more residual back pain, leg pain, and higher rates of rehospitalization. Fusion of the involved lumbar segments should be considered only if significant spinal instability is present (i.e., degenerative spondylolisthesis). The costs associated with lumbar interbody fusion have increased dramatically in recent years. There are no large prospective, randomized trials comparing fusion to other types of surgical intervention. In one study, patients with persistent low back pain despite an initial diskectomy fared no better with spine fusion than with a conservative regimen of cognitive intervention and exercise. Artificial disks are used in Europe; their utility remains controversial in the United States.

PAIN IN THE NECK AND SHOULDER

Neck pain, which usually arises from diseases of the cervical spine and soft tissues of the neck, is common. Neck pain arising from the cervical spine is typically precipitated by movement and may be accompanied by focal tenderness and limitation of motion. Many of the prior comments made regarding causes of low back pain also apply to disorders of the cervical spine. The text below will emphasize differences. Pain arising from the brachial plexus, shoulder, or peripheral nerves can be confused with cervical spine disease **(Table 14-4)**, but the history and examination usually identify a more distal origin for the pain. When the site of nerve tissue injury is unclear, EMG studies can localize the lesion. Cervical spine trauma, disk disease, or spondylosis with intervertebral foraminal narrowing may be asymptomatic or painful and can produce a myelopathy, radiculopathy, or both. The same risk factors for serious causes of low back pain also apply to neck pain with the additional feature that neurologic signs of myelopathy (incontinence, sensory level, spastic legs) may also occur. Lhermitte's sign, an electrical shock down the spine with neck flexion, suggests involvement of the cervical spinal cord.

TABLE 14-4

Cervical Radiculopathy: Neurologic Features

Cervical Nerve Roots	Examination Findings			Pain Distribution
	Reflex	Sensory	Motor	
C5	Biceps	Lateral deltoid	Rhomboids[a] (elbow extends backward with hand on hip)	Lateral arm, medial scapula
			Infraspinatus[a] (arm rotates externally with elbow flexed at the side)	
			Deltoid[a] (arm raised laterally 30–45° from the side)	
C6	Biceps	Thumb/index finger;	Biceps[a] (arm flexed at the elbow in supination)	Lateral forearm, thumb/index fingers
		Dorsal hand/lateral forearm	Pronator teres (forearm pronated)	
C7	Triceps	Middle fingers	Triceps[a] (forearm extension, flexed at elbow)	Posterior arm, dorsal forearm, dorsal hand
		Dorsal forearm	Wrist/finger extensors[a]	
C8	Finger flexors	Palmar surface of little finger	Abductor pollicis brevis (abduction of thumb)	Fourth and fifth fingers, medial hand and forearm
		Medial hand and forearm	First dorsal interosseous (abduction of index finger)	
			Abductor digiti minimi (abduction of little finger)	
T1	Finger flexors	Axilla and medial arm	Abductor pollicis brevis (abduction of thumb)	Medial arm, axilla
			First dorsal interosseous (abduction of index finger)	
			Abductor digiti minimi (abduction of little finger)	

[a]These muscles receive the majority of innervation from this root.

TRAUMA TO THE CERVICAL SPINE

Trauma to the cervical spine (fractures, subluxation) places the spinal cord at risk for compression. Motor vehicle accidents, violent crimes, or falls account for 87% of cervical spinal cord injuries **(Chap. 434)**. Immediate immobilization of the neck is essential to minimize further spinal cord injury from movement of unstable cervical spine segments. The decision to obtain imaging should be based on the nature of the injury. The National Emergency X-Radiography Utilization Study (NEXUS) low-risk criteria established that normally alert patients without palpation tenderness in the midline; intoxication; neurologic deficits; or painful distracting injuries were very unlikely to have sustained a clinically significant traumatic injury to the cervical spine. The Canadian C-spine rule recommends that imaging should be obtained following neck region trauma if the patient is >65 years old or has limb paresthesias or if there was a dangerous mechanism for the injury (e.g., bicycle collision with tree or parked car, fall from height >3 feet or five stairs, diving accident). These guidelines are helpful but must be tailored to individual circumstances; for example, patients with advanced osteoporosis, glucocorticoid use, or cancer may warrant imaging after even mild trauma. A CT scan is the diagnostic procedure of choice for detection of acute fractures following severe trauma; plain x-rays can be used for lesser degrees of trauma. When traumatic injury to the vertebral arteries or cervical spinal cord is suspected, visualization by MRI with magnetic resonance angiography is preferred.

Whiplash injury is due to rapid flexion and extension of the neck, usually from automobile accidents. The exact mechanism of injury is unclear. This diagnosis should not be applied to patients with fractures, disk herniation, head injury, focal neurologic findings, or altered consciousness. Up to 50% of persons reporting whiplash injury acutely have persistent neck pain 1 year later. When personal compensation for pain and suffering was removed from the

Australian health care system, the prognosis for recovery at 1 year improved. Imaging of the cervical spine is not cost-effective acutely but is useful to detect disk herniations when symptoms persist for >6 weeks following the injury. Severe initial symptoms have been associated with a poor long-term outcome.

CERVICAL DISK DISEASE

Degenerative cervical disk disease is very common and usually asymptomatic. Herniation of a lower cervical disk is a common cause of pain or tingling in the neck, shoulder, arm, or hand. Neck pain, stiffness, and a range of motion limited by pain are the usual manifestations. Herniated cervical disks are responsible for ~25% of cervical radiculopathies. Extension and lateral rotation of the neck narrow the ipsilateral intervertebral foramen and may reproduce radicular symptoms (Spurling's sign). In young adults, acute nerve root compression from a ruptured cervical disk is often due to trauma. Cervical disk herniations are usually posterolateral near the lateral recess. Typical patterns of reflex, sensory, and motor changes that accompany cervical nerve root lesions are summarized in Table 14-4. Although the classic patterns are clinically helpful, there are numerous exceptions because (1) there is overlap in sensory function between adjacent nerve roots, (2) symptoms and signs may be evident in only part of the injured nerve root territory, and (3) the location of pain is the most variable of the clinical features.

CERVICAL SPONDYLOSIS

Osteoarthritis of the cervical spine may produce neck pain that radiates into the back of the head, shoulders, or arms, or may be the source of headaches in the posterior occipital region (supplied by the C2-C4 nerve roots). Osteophytes, disk protrusions, or hypertrophic facet or uncovertebral joints may alone or in combination compress one or several nerve roots at the intervertebral foramina; these causes together account for 75% of cervical radiculopathies. The roots most commonly affected are C7 and C6. Narrowing of the spinal canal by osteophytes, ossification of the posterior longitudinal ligament (OPLL), or a large central disk may compress the cervical spinal cord and produce signs of myelopathy alone or radiculopathy with myelopathy (myeloradiculopathy). When little or no neck pain accompanies cervical cord involvement, other diagnoses to be considered include amyotrophic lateral sclerosis **(Chap. 429)**, multiple sclerosis **(Chap. 436)**, spinal cord tumors, or syringomyelia **(Chap. 434)**. Cervical spondylotic myelopathy should be considered even when the patient presents with symptoms or spinal cord signs in the legs only. MRI is the study of choice to define soft tissues in the cervical region including the spinal cord, whereas plain CT is optimal to identify bone pathology including foraminal, lateral recess, or spinal canal stenosis. With spondylotic myelopathy focal enhancement by MRI, sometimes in a characteristic "pancake pattern", may be present at the site of maximal cord compression.

There is no evidence to support prophylactic surgery for asymptomatic cervical spinal stenosis unaccompanied by myelopathic signs or abnormal spinal cord findings on MR imaging, except in the setting of *dynamic instability* (see spondylolisthesis above). If the patient has postural neck pain, a prior history of whiplash or other spine/head injury, a Lhermitte sign, or preexisting listhesis at the stenotic segment on cervical MRI, or CT, then cervical spine flexion-extension x-rays are indicated to look for dynamic instability. Surgical intervention is not recommended for patients with listhesis alone, unaccompanied by dynamic instability.

OTHER CAUSES OF NECK PAIN

RA **(Chap. 351)** of the cervical facet joints produces neck pain, stiffness, and limitation of motion. Synovitis of the atlantoaxial joint (C1-C2; Fig. 14-2) may damage the transverse ligament of the atlas, producing forward displacement of the atlas on the axis (atlantoaxial subluxation). Radiologic evidence of atlantoaxial subluxation occurs in up to 30% of patients with RA and plain x-ray films of the neck should be routinely performed preoperatively to assess the risk of neck hyperextension in patients requiring intubation. The degree of subluxation correlates with the severity of erosive disease. When subluxation is present, careful assessment is important to identify early signs of myelopathy that could be a harbinger of life-threatening spinal cord compression. Surgery should be considered when myelopathy or spinal instability is present. *Ankylosing spondylitis* is another cause of neck pain and less commonly atlantoaxial subluxation.

Acute *herpes zoster* can present as acute posterior occipital or neck pain prior to the outbreak of vesicles. *Neoplasms* metastatic to the cervical spine, *infections* (osteomyelitis and epidural abscess), and *metabolic bone diseases* may be the cause of neck pain, as discussed above. Neck pain may also be referred from the heart with coronary artery ischemia (cervical angina syndrome).

THORACIC OUTLET SYNDROMES

The thoracic outlet contains the first rib, the subclavian artery and vein, the brachial plexus, the clavicle, and the lung apex. Injury to these structures may result in postural or movement-induced pain around the shoulder and supraclavicular region, classified as follows.

True neurogenic thoracic outlet syndrome (*TOS*) is an uncommon disorder resulting from compression of the lower trunk of the brachial plexus or ventral rami of the C8 or T1 nerve roots, caused most often by an anomalous band of tissue connecting an elongate transverse process at C7 with the first rib. Pain is mild or may be absent. Signs include weakness and wasting of intrinsic muscles of the hand and diminished sensation on the palmar aspect of the fifth digit. An anteroposterior cervical spine x-ray will show an elongate C7 transverse process (an anatomic marker for the anomalous cartilaginous band), and EMG and NCSs confirm the diagnosis. Treatment consists of surgical resection of the anomalous band. The weakness and wasting of intrinsic hand muscles typically does not improve, but surgery halts the insidious progression of weakness.

Arterial TOS results from compression of the subclavian artery by a cervical rib, resulting in poststenotic dilatation of the artery and in some cases secondary thrombus formation. Blood pressure is reduced in the affected limb, and signs of emboli may be present in the hand. Neurologic signs are absent. Ultrasound can confirm the diagnosis noninvasively. Treatment is with thrombolysis or anticoagulation (with or without embolectomy) and surgical excision of the cervical rib compressing the subclavian artery.

Venous TOS is due to subclavian vein thrombosis resulting in swelling of the arm and pain. The vein may be compressed by a cervical rib or anomalous scalene muscle. Venography is the diagnostic test of choice.

Disputed TOS accounts for 95% of patients diagnosed with TOS; chronic arm and shoulder pain are prominent and of unclear cause. The lack of sensitive and specific findings on physical examination or specific markers for this condition results in diagnostic uncertainty. The role of surgery in disputed TOS is controversial. Major depression, chronic symptoms, work-related injury, and diffuse arm symptoms predict poor surgical outcomes. Multidisciplinary pain management is a conservative approach, although treatment is often unsuccessful.

BRACHIAL PLEXUS AND NERVES

Pain from injury to the brachial plexus or peripheral nerves of the arm can occasionally mimic referred pain of cervical spine origin including cervical radiculopathy. Neoplastic infiltration of the lower trunk of the brachial plexus may produce shoulder or supraclavicular pain radiating down the arm, numbness of the fourth and fifth fingers or medial forearm, and weakness of intrinsic hand muscles innervated by the lower trunk and medial cord of the brachial plexus. Delayed radiation injury may produce weakness in the upper arm or numbness of the lateral forearm or arm due to involvement of the upper trunk and lateral cord of the plexus. Pain is less common and less severe than with neoplastic infiltration. A Pancoast tumor of the lung **(Chap. 74)** is another cause and should be considered, especially when a concurrent Horner's syndrome is present. *Suprascapular neuropathy* may produce severe shoulder pain, weakness, and wasting of the supraspinatus and infraspinatus muscles. *Acute brachial neuritis* is often confused with radiculopathy; the acute onset of severe shoulder or scapular pain is followed typically over days by weakness of the proximal arm and shoulder girdle muscles innervated by the upper brachial plexus. The onset may be preceded by an infection, vaccination, or minor surgical procedure. The long thoracic nerve may be affected, resulting in a winged scapula. Brachial neuritis may also present as an isolated paralysis of the diaphragm with or without involvement of other nerves of the upper limb. Recovery may take up to 3 years, and full functional recovery can be expected in the majority of patients.

Occasional cases of carpal tunnel syndrome produce pain and paresthesias extending into the forearm, arm, and shoulder resembling a C5 or C6 root lesion. Lesions of the radial or ulnar nerve can also mimic radiculopathy, at C7 or C8, respectively. EMG and NCSs can accurately localize lesions to the nerve roots, brachial plexus, or peripheral nerves.

For further discussion of peripheral nerve disorders, see Chap. 438.

SHOULDER

Pain arising from the shoulder can on occasion mimic pain from the spine. If symptoms and signs of radiculopathy are absent, then the differential diagnosis includes mechanical shoulder pain (tendonitis, bursitis, rotator cuff tear, dislocation, adhesive capsulitis, or rotator cuff impingement under the acromion) and referred pain (subdiaphragmatic irritation, angina, Pancoast tumor). Mechanical pain is often worse at night, associated with local shoulder tenderness and aggravated by passive abduction, internal rotation, or extension of the arm. Demonstrating normal passive full range of motion of the arm at the shoulder without worsening the usual pain can help exclude mechanical shoulder pathology as a cause of neck region pain. Pain from shoulder disease may radiate into the arm or hand, but focal neurologic signs (sensory, motor, or reflex changes) are absent.

GLOBAL CONSIDERATIONS


Many of the considerations described above for LBP also apply to neck pain. Neck pain was ranked #21 as a cause of DALYs in the Global Burden of Diseases Study 2010, accounting for ~40% of the total global DALYs due to LBP. In general, neck pain rankings were also higher in developed regions of the world.

TREATMENT

TREATMENT

Neck Pain without Radiculopathy

The evidence regarding treatment for neck pain is less comprehensive than that for low back pain, but the approach is remarkably similar in many respects. As with low back pain, spontaneous improvement is the norm for acute neck pain. The usual goals of therapy are to promote a rapid return to normal function and provide pain relief while healing proceeds.

Acute neck pain is often treated with a combination of NSAIDs, acetaminophen, cold packs, or heat while awaiting spontaneous recovery. For patients kept awake by symptoms, cyclobenzaprine (5–10 mg) at night can help relieve muscle spasm and promote drowsiness. For patients with neck pain unassociated with trauma, supervised exercise with or without mobilization appears to be effective. Exercises often include shoulder rolls and neck stretches. The

evidence in support of nonsurgical treatments for whiplash-associated disorders is generally of limited quality and neither supports nor refutes the common treatments used for symptom relief. Gentle mobilization of the cervical spine combined with exercise programs may be beneficial. Evidence is insufficient to recommend the use of cervical traction, TENS, ultrasound, electromagnetic therapy, trigger point injections, botulinum toxin injections, tricyclic antidepressants, and SSRIs for acute or chronic neck pain. Some patients obtain modest pain relief using a soft neck collar; there is little risk or cost. Massage can produce temporary pain relief.

For patients with chronic neck pain, supervised exercise programs can provide symptom relief and improve function. Acupuncture provided short-term benefit for some patients when compared to a sham procedure and is an option. Spinal manipulation alone has not been shown to be effective and carries a risk for injury. Surgical treatment for chronic neck pain without radiculopathy or spine instability is not recommended.

TREATMENT

TREATMENT

Neck Pain with Radiculopathy

The natural history of neck pain with acute radiculopathy due to disk disease is favorable, and many patients will improve without specific therapy. Although there are no randomized trials of NSAIDs for neck pain, a course of NSAIDs, acetaminophen, or both, with or without muscle relaxants, and avoidance of activities that trigger symptoms are reasonable as initial therapy. Gentle supervised exercise and avoidance of inactivity are reasonable as well. A short course of high dose oral glucocorticoids with a rapid taper, or epidural steroids administered under imaging guidance can be effective for acute or subacute disk-related cervical radiculopathy, but have not been subjected to rigorous trials. The risk of injection complications is higher in the neck than the low back; vertebral artery dissection, dural puncture, and embolism from injection particles in the vertebral arteries have all been reported. Opioid analgesics can be used in the emergency room and for short courses as an outpatient. Soft cervical collars can be modestly helpful by limiting spontaneous and reflex neck movements that exacerbate pain; hard collars are in general poorly tolerated.

If cervical radiculopathy is due to bony compression from cervical spondylosis with foraminal narrowing, periodic follow-up to assess for progression is indicated and consideration of surgical decompression is reasonable. Surgical treatment can produce rapid pain relief, although it is unclear whether long-term outcomes are improved over nonsurgical therapy. Indications for cervical disk surgery include a progressive motor deficit due to nerve root compression, functionally limiting pain that fails to respond to conservative management, or spinal cord compression.

Surgical treatments include anterior cervical diskectomy alone, laminectomy with diskectomy, or diskectomy with fusion. The risk of subsequent radiculopathy or myelopathy at cervical segments adjacent to a fusion is ~3% per year and 26% per decade. Although this risk is sometimes portrayed as a late complication of surgery, it may also reflect the natural history of degenerative cervical disk disease.

FURTHER READING

Agency for Healthcare Research and Quality (AHRQ): Noninvasive treatments for low back pain. AHRQ Publication No. 16-EHC004-EF. February 2016, *https://effectivehealthcare.ahrq.gov/ehc/products/553/2178/back-pain-treatment-report-160229.pdf.*

Benzon HT et al: Improving the safety of epidural steroid injections. JAMA 313:1713, 2015. [PubMed: 25822848]

Friedly JL et al: A randomized trial of epidural glucocorticoid injections for spinal stenosis. N Engl J Med 371:11, 2014. [PubMed: 24988555]

Goldberg H et al: Oral steroids for acute radiculopathy due to a herniated lumbar disk. JAMA 313:1915, 2015. [PubMed: 25988461]

Hoy DG et al: Reflecting on the global burden of musculoskeletal conditions: Lessons learnt from the global burden of disease 2010 study and the next steps forward. Ann Rheum Dis 74:4, 2015. [PubMed: 24914071]

Katz JN, Harris MB: Clinical practice. Lumbar spinal stenosis. N Engl J Med 358:818, 2008. [PubMed: 18287604]

Lamb SE et al: Group cognitive behavioural treatment for low-back pain in primary care: A randomised controlled trial and cost-effectiveness analysis. Lancet 375:916, 2010. [PubMed: 20189241]

Malmivaara A et al: The treatment of acute low back pain—Bed rest, exercises, or ordinary activity? N Engl J Med 332:351, 1995. [PubMed: 7823996]

Melanica J et al: Spinal stenosis. Handb Clin Neurol 109:541, 2014.

Serinken M et al: Comparison of intravenous morphine versus paracetamol in sciatica: A randomized placebo controlled trial. Acad Emerg Med 23:674, 2016. [PubMed: 26938140]

Zygourakis CC et al: Geographic and hospital variation in cost of lumbar laminectomy and lumbar fusion for degenerative conditions. Neurosurgery 81:331, 2017.

Chapter 15: Fever

Charles A. Dinarello; Reuven Porat

FIGURE 15-1

INTRODUCTION

Body temperature is controlled by the hypothalamus. Neurons in both the preoptic anterior hypothalamus and the posterior hypothalamus receive two kinds of signals: one from peripheral nerves that transmit information from warmth/cold receptors in the skin and the other from the temperature of the blood bathing the region. These two types of signals are integrated by the thermoregulatory center of the hypothalamus to maintain normal temperature. In a neutral temperature environment, the human metabolic rate produces more heat than is necessary to maintain the core body temperature in the range of 36.5–37.5°C (97.7–99.5°F).

A normal body temperature is ordinarily maintained despite environmental variations because the hypothalamic thermoregulatory center balances the excess heat production derived from metabolic activity in muscle and the liver with heat dissipation from the skin and lungs. According to studies of healthy individuals 18–40 years of age, the mean oral temperature is 36.8° ± 0.4°C (98.2° ± 0.7°F), with low levels at 6 A.M. and higher levels at 4–6 P.M. The maximal normal oral temperature is 37.2°C (98.9°F) at 6 A.M. and 37.7°C (99.9°F) at 4 P.M.; these values define the 99th percentile for healthy individuals. In light of these studies, *an A.M. temperature of >37.2°C (>98.9°F) or a P.M. temperature of >37.7°C (>99.9°F) would define a fever.* The normal daily temperature variation, also called the *circadian rhythm*, is typically 0.5°C (0.9°F). However, in some individuals recovering from a febrile illness, this daily variation can be as great as 1.0°C. During a febrile illness, the diurnal variation is usually maintained, but at higher, febrile levels. The daily temperature variation appears to be fixed in early childhood; in contrast, elderly individuals can exhibit a reduced ability to develop fever, with only a modest fever even in severe infections.

Rectal temperatures are generally 0.4°C (0.7°F) higher than oral readings. The lower oral readings are probably attributable to mouth breathing, which is a factor in patients with respiratory infections and rapid breathing. Lower-esophageal temperatures closely reflect core temperature. Tympanic membrane thermometers measure radiant heat from the tympanic membrane and nearby ear canal and display that absolute value (*unadjusted mode*) or a value automatically calculated from the absolute reading on the basis of nomograms relating the radiant temperature measured to actual core temperatures obtained in clinical studies (*adjusted mode*). These measurements, although convenient, may be more variable than directly determined oral or rectal values. Studies in adults show that readings are lower with unadjusted-mode than

with adjusted-mode tympanic membrane thermometers and that unadjusted-mode tympanic membrane values are 0.8°C (1.6°F) lower than rectal temperatures.

In women who menstruate, the A.M. temperature is generally lower during the 2 weeks before ovulation; it then rises by ~0.6°C (1°F) with ovulation and stays at that level until menses occur. During the luteal phase, the amplitude of the circadian rhythm remains the same.

FEVER VERSUS HYPERTHERMIA

Fever is an elevation of body temperature that exceeds the normal daily variation and occurs *in conjunction with an increase in the hypothalamic set point* (e.g., from 37°C to 39°C). This shift of the set point from "normothermic" to febrile levels very much resembles the resetting of the home thermostat to a higher level in order to raise the ambient temperature in a room. Once the hypothalamic set point is raised, neurons in the vasomotor center are activated and vasoconstriction commences. The individual first notices vasoconstriction in the hands and feet. Shunting of blood away from the periphery to the internal organs essentially decreases heat loss from the skin, and the person feels cold. For most fevers, body temperature increases by 1–2°C. Shivering, which increases heat production from the muscles, may begin at this time; however, shivering is not required if mechanisms of heat conservation raise blood temperature sufficiently. Nonshivering heat production from the liver also contributes to increasing core temperature. Behavioral adjustments (e.g., putting on more clothing or bedding) help raise body temperature by decreasing heat loss.

The processes of heat conservation (vasoconstriction) and heat production (shivering and increased nonshivering thermogenesis) continue until the temperature of the blood bathing the hypothalamic neurons matches the new "thermostat setting." Once that point is reached, the hypothalamus maintains the temperature at the febrile level by the same mechanisms of heat balance that function in the afebrile state. When the hypothalamic set point is again reset downward (in response to either a reduction in the concentration of pyrogens or the use of antipyretics), the processes of heat loss through vasodilation and sweating are initiated. Loss of heat by sweating and vasodilation continues until the blood temperature at the hypothalamic level matches the lower setting. Behavioral changes (e.g., removal of clothing) facilitate heat loss.

A fever of >41.5°C (>106.7°F) is called *hyperpyrexia*. This extraordinarily high fever can develop in patients with severe infections but most commonly occurs in patients with central nervous system (CNS) hemorrhages. In the preantibiotic era, fever due to a variety of infectious diseases rarely exceeded 106°F, and there has been speculation that this natural "thermal ceiling" is mediated by neuropeptides functioning as central antipyretics.

In rare cases, the hypothalamic set point is elevated as a result of local trauma, hemorrhage, tumor, or intrinsic hypothalamic malfunction. The term *hypothalamic fever* is sometimes used to describe elevated temperature caused by abnormal hypothalamic function. However, most patients with hypothalamic damage have *sub*normal, not *supra*normal, body temperatures.

Although most patients with elevated body temperature have fever, there are circumstances in which elevated temperature represents not fever but *hyperthermia* (*heat stroke*). Hyperthermia is characterized by an uncontrolled increase in body temperature that exceeds the body's ability to lose heat. The setting of the hypothalamic thermoregulatory center is unchanged. In contrast to fever in infections, hyperthermia does not involve pyrogenic molecules. Exogenous heat exposure and endogenous heat production are two mechanisms by which hyperthermia can result in dangerously high internal temperatures. Excessive heat production can easily cause hyperthermia despite physiologic and behavioral control of body temperature. For example, work or exercise in hot environments can produce heat faster than peripheral mechanisms can lose it. **For a detailed discussion of hyperthermia, see Chap. 455**.

It is important to distinguish between fever and hyperthermia since hyperthermia can be rapidly fatal and characteristically does not respond to antipyretics. In an emergency situation, however, making this distinction can be difficult. For example, in systemic sepsis, fever (hyperpyrexia) can be rapid in onset, and temperatures can exceed 40.5°C (104.9°F). Hyperthermia is often diagnosed on the basis of the events immediately preceding the elevation of core temperature—e.g., heat exposure or treatment with drugs that interfere with thermoregulation. In patients with heat stroke syndromes and in those taking drugs that block sweating, the skin is hot but dry, whereas in fever the skin can be cold as a consequence of vasoconstriction. Antipyretics do not reduce the elevated temperature in hyperthermia, whereas in fever—and even in hyperpyrexia—adequate doses of either aspirin or acetaminophen usually result in some decrease in body temperature.

PATHOGENESIS OF FEVER

PYROGENS

The term *pyrogen* (Greek *pyro*, "fire") is used to describe any substance that causes fever. *Exogenous* pyrogens are derived from outside the patient; most are microbial products, microbial toxins, or whole microorganisms (including viruses). The classic example of an exogenous pyrogen is the lipopolysaccharide (endotoxin) produced by all gram-negative bacteria. Pyrogenic products of gram-positive organisms include the enterotoxins of *Staphylococcus aureus* and the groups A and B streptococcal toxins, also called *superantigens*. One staphylococcal toxin of clinical importance is that associated with isolates of *S. aureus* from patients with toxic shock syndrome. These products of staphylococci and streptococci cause fever in experimental animals when injected intravenously at concentrations of 1–10 μg/kg. Endotoxin is a highly pyrogenic molecule in humans: when injected intravenously into volunteers, a dose of 2–3 ng/kg produces fever, leukocytosis, acute-phase proteins, and generalized symptoms of malaise.

PYROGENIC CYTOKINES

Cytokines are small proteins (molecular mass, 10,000–20,000 Da) that regulate immune, inflammatory, and hematopoietic processes. For example, the elevated leukocytosis seen in several infections with an absolute neutrophilia is attributable to the cytokines interleukin (IL) 1 and IL-6. Some cytokines also cause fever;

formerly referred to as *endogenous pyrogens*, they are now called *pyrogenic cytokines*. The pyrogenic cytokines include IL-1, IL-6, tumor necrosis factor (TNF), and ciliary neurotropic factor, a member of the IL-6 family. Fever is a prominent side effect of interferon α therapy. Each pyrogenic cytokine is encoded by a separate gene, and each has been shown to cause fever in laboratory animals and in humans. When injected into humans at low doses (10–100 ng/kg), IL-1 and TNF produce fever; in contrast, for IL-6, a dose of 1–10 μg/kg is required for fever production.

A wide spectrum of bacterial and fungal products induce the synthesis and release of pyrogenic cytokines. However, fever can be a manifestation of disease in the absence of microbial infection. For example, inflammatory processes such as pericarditis, trauma, stroke, and routine immunizations induce the production of IL-1, TNF, and/or IL-6; individually or in combination, these cytokines trigger the hypothalamus to raise the set point to febrile levels.

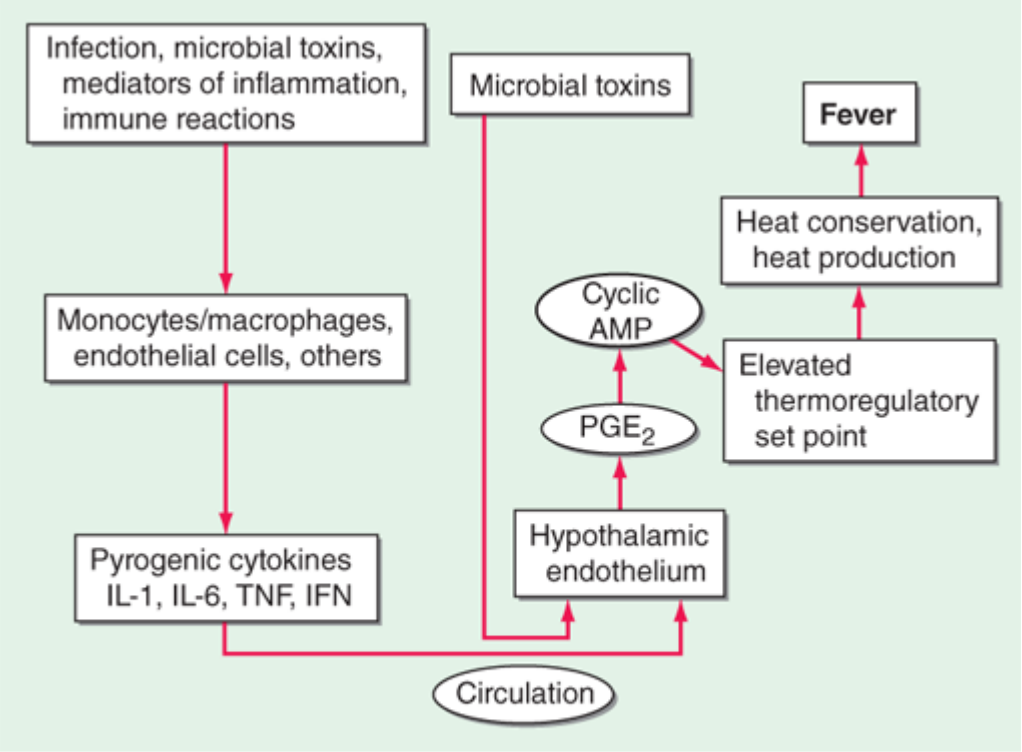
ELEVATION OF THE HYPOTHALAMIC SET POINT BY CYTOKINES

During fever, levels of prostaglandin E_2 (PGE_2) are elevated in hypothalamic tissue and the third cerebral ventricle. The concentrations of PGE_2 are highest near the circumventricular vascular organs (organum vasculosum of lamina terminalis)—networks of enlarged capillaries surrounding the hypothalamic regulatory centers. Destruction of these organs reduces the ability of pyrogens to produce fever. Most studies in animals have failed to show, however, that pyrogenic cytokines pass from the circulation into the brain itself. Thus, it appears that both exogenous pyrogens and pyrogenic cytokines interact with the endothelium of these capillaries and that this interaction is the first step in initiating fever—i.e., in raising the set point to febrile levels.

The key events in the production of fever are illustrated in **Fig. 15-1**. Myeloid and endothelial cells are the primary cell types that produce pyrogenic cytokines. Pyrogenic cytokines such as IL-1, IL-6, and TNF are released from these cells and enter the systemic circulation. Although these circulating cytokines lead to fever by inducing the synthesis of PGE_2, they also induce PGE_2 in peripheral tissues. The increase in PGE_2 in the periphery accounts for the nonspecific myalgias and arthralgias that often accompany fever. It is thought that some systemic PGE_2 escapes destruction by the lung and gains access to the hypothalamus via the internal carotid. However, it is the elevation of PGE_2 in the brain that starts the process of raising the hypothalamic set point for core temperature.

FIGURE 15-1

Chronology of events required for the induction of fever. AMP, adenosine 5′-monophosphate; IFN, interferon; IL, interleukin; PGE_2, prostaglandin E_2; TNF, tumor necrosis factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

There are four receptors for PGE_2, and each signals the cell in different ways. Of the four receptors, the third (EP-3) is essential for fever: when the gene for this receptor is deleted in mice, no fever follows the injection of IL-1 or endotoxin. Deletion of the other PGE_2 receptor genes leaves the fever mechanism intact. Although PGE_2 is essential for fever, it is not a neurotransmitter. Rather, the release of PGE_2 from the brain side of the hypothalamic endothelium triggers the PGE_2 receptor on glial cells, and this stimulation results in the rapid release of cyclic adenosine 5′-monophosphate (cAMP), which is a neurotransmitter. As shown in Fig. 15-1, the release of cAMP from glial cells activates neuronal endings from the thermoregulatory center that extend into the area. The elevation of cAMP is thought to account for changes in the hypothalamic set point either directly or indirectly (by inducing the release of neurotransmitters). Distinct receptors for microbial products are located on the hypothalamic endothelium. These receptors are called *Toll-like receptors* and are similar in many ways to IL-1 receptors. IL-1 receptors and Toll-like receptors share the same signal-transducing mechanism. Thus, the direct activation of Toll-like receptors or IL-1 receptors results in PGE_2 production and fever.

PRODUCTION OF CYTOKINES IN THE CNS

Cytokines produced in the brain may account for the hyperpyrexia of CNS hemorrhage, trauma, or infection. Viral infections of the CNS induce microglial and possibly neuronal production of IL-1, TNF, and IL-6. In experimental animals, the concentration of a cytokine required to cause fever is several orders of magnitude lower with direct injection into the brain substance or brain ventricles than with systemic injection. Therefore, cytokines produced in the CNS can raise the hypothalamic set point, bypassing the

circumventricular organs. CNS cytokines likely account for the hyperpyrexia of CNS hemorrhage, trauma, or infection.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Fever

PHYSICAL EXAMINATION

The chronology of events preceding fever, including exposure to other infected individuals or to vectors of disease, should be ascertained. Electronic devices for measuring oral, tympanic membrane, or rectal temperatures are reliable, but the same site should be used consistently to monitor a febrile disease. Moreover, physicians should be aware that newborns, elderly patients, patients with chronic hepatic or renal failure, and patients taking glucocorticoids or being treated with an anticytokine may have active infection in the absence of fever because of a blunted febrile response.

LABORATORY TESTS

The workup should include a complete blood count; a differential count should be performed manually or with an instrument sensitive to the identification of juvenile or band forms, toxic granulations, and Döhle bodies, which are suggestive of bacterial infection. Neutropenia may be present with some viral infections.

Measurement of circulating cytokines in patients with fever is not helpful since levels of cytokines such as IL-1 and TNF in the circulation often are below the detection limit of the assay or do not coincide with fever. However, in patients with low-grade fevers or with suspected occult disease, the most valuable measurements are the C-reactive protein (CRP) level and the erythrocyte sedimentation rate. These markers of inflammatory processes are particularly helpful in detecting occult disease. Measurement of circulating IL-6, which induces CRP, can be useful. However, whereas IL-6 levels may vary during a febrile disease, CRP levels remain elevated. **Acute-phase reactants are discussed in Chap. 297**.

FEVER IN PATIENTS RECEIVING ANTICYTOKINE THERAPY

Patients receiving long-term treatment with anticytokine-based regimens are at increased risk of infection because of lowered host defenses. For example, latent *Mycobacterium tuberculosis* infection can disseminate in patients receiving anti-TNF therapy. With the increasing use of anticytokines to reduce the activity of IL-1, IL-6, IL-12, IL-17, or TNF in patients with Crohn's disease, rheumatoid arthritis, or psoriasis, the possibility that these therapies blunt the febrile response should be kept in mind.

The blocking of cytokine activity has the distinct clinical drawback of lowering the level of host defenses against both routine bacterial and opportunistic infections such as *M. tuberculosis* and fungal infections. The use of monoclonal antibodies to reduce IL-17 in psoriasis increases the risk of systemic candidiasis.

In nearly all reported cases of infection associated with anticytokine therapy, fever is among the presenting signs. However, the extent to which the febrile response is blunted in these patients remains unknown.

Therefore, low-grade fever in patients receiving anticytokine therapies is of considerable concern. The physician should conduct an early and rigorous diagnostic evaluation in these cases. The febrile response is also blunted in patients receiving chronic glucocorticoid therapy or anti-inflammatory agents such as nonsteroidal anti-inflammatory drugs (NSAIDs).

TREATMENT

TREATMENT

Fever

THE DECISION TO TREAT FEVER

Most fevers are associated with self-limited infections, such as common viral diseases. The use of antipyretics is not contraindicated in these infections: no significant clinical evidence indicates either that antipyretics delay the resolution of viral or bacterial infections or that fever facilitates recovery from infection or acts as an adjuvant to the immune system. In short, treatment of fever and its symptoms with routine antipyretics does no harm and does not slow the resolution of common viral and bacterial infections.

However, in bacterial infections, the withholding of antipyretic therapy can be helpful in evaluating the effectiveness of a particular antibiotic, especially in the absence of positive cultures of the infecting organism, and the routine use of antipyretics can mask an inadequately treated bacterial infection. Withholding antipyretics in some cases may facilitate the diagnosis of an unusual febrile disease. Temperature–pulse dissociation (*relative bradycardia*) occurs in typhoid fever, brucellosis, leptospirosis, some drug-induced fevers, and factitious fever. As stated earlier, in newborns, elderly patients, patients with chronic liver or kidney failure, and patients taking glucocorticoids, fever may not be present despite infection. Hypothermia can develop in patients with septic shock.

Some infections have characteristic patterns in which febrile episodes are separated by intervals of normal temperature. For example, *Plasmodium vivax* causes fever every third day, whereas fever occurs every fourth day with *Plasmodium malariae*. Another relapsing fever is related to *Borrelia* infection, with days of fever followed by a several-day afebrile period and then a relapse into additional days of fever. In the Pel-Ebstein pattern, fever lasting 3–10 days is followed by afebrile periods of 3–10 days; this pattern can be classic for Hodgkin's disease and other lymphomas. In cyclic neutropenia, fevers occur every 21 days and accompany the neutropenia. There is no periodicity of fever in patients with familial Mediterranean fever. However, these patterns have limited or no diagnostic value compared with specific and rapid laboratory tests.

ANTICYTOKINE THERAPY TO REDUCE FEVER IN AUTOIMMUNE AND AUTOINFLAMMATORY DISEASES

Recurrent fever is documented at some point in most autoimmune diseases and nearly all autoinflammatory diseases. Although fever can be a manifestation of autoimmune diseases, recurrent fevers are characteristic of autoinflammatory diseases **(Table 15-1)**, including uncommon diseases such as adult and juvenile Still's disease, familial Mediterranean fever, and hyper-IgD syndrome but also common diseases such as idiopathic pericarditis and gout. In addition to recurrent fevers, neutrophilia and serosal inflammation characterize

autoinflammatory diseases. The fevers associated with these illnesses are dramatically reduced by blocking of IL-1 activity with anakinra or canakinumab. Anticytokines therefore reduce fever in autoimmune and autoinflammatory diseases. Although fevers in autoinflammatory diseases are mediated by IL-1β, patients also respond to antipyretics.

MECHANISMS OF ANTIPYRETIC AGENTS

The reduction of fever by lowering of the elevated hypothalamic set point is a direct function of reduction of the PGE_2 level in the thermoregulatory center. The synthesis of PGE_2 depends on the constitutively expressed enzyme cyclooxygenase. The substrate for cyclooxygenase is arachidonic acid released from the cell membrane, and this release is the rate-limiting step in the synthesis of PGE_2. Therefore, inhibitors of cyclooxygenase are potent antipyretics. The antipyretic potency of various drugs is directly correlated with the inhibition of brain cyclooxygenase. Acetaminophen is a poor cyclooxygenase inhibitor in peripheral tissue and lacks noteworthy anti-inflammatory activity; in the brain, however, acetaminophen is oxidized by the p450 cytochrome system, and the oxidized form inhibits cyclooxygenase activity. Moreover, in the brain, the inhibition of another enzyme, COX-3, by acetaminophen may account for the antipyretic effect of this agent. However, COX-3 is not found outside the CNS.

Oral aspirin and acetaminophen are equally effective in reducing fever in humans. NSAIDs such as ibuprofen and specific inhibitors of COX-2 also are excellent antipyretics. Chronic, high-dose therapy with antipyretics such as aspirin or any NSAID does not reduce normal core body temperature. Thus, PGE_2 appears to play no role in normal thermoregulation.

As effective antipyretics, glucocorticoids act at two levels. First, similar to the cyclooxygenase inhibitors, glucocorticoids reduce PGE_2 synthesis by inhibiting the activity of phospholipase A_2, which is needed to release arachidonic acid from the cell membrane. Second, glucocorticoids block the transcription of the mRNA for the pyrogenic cytokines. Limited experimental evidence indicates that ibuprofen and COX-2 inhibitors reduce IL-1-induced IL-6 production and may contribute to the antipyretic activity of NSAIDs.

REGIMENS FOR THE TREATMENT OF FEVER

The objectives in treating fever are first to reduce the elevated hypothalamic set point and second to facilitate heat loss. Reducing fever with antipyretics also reduces systemic symptoms of headache, myalgias, and arthralgias.

Oral aspirin and NSAIDs effectively reduce fever but can adversely affect platelets and the gastrointestinal tract. Therefore, acetaminophen is preferred as an antipyretic. In children, acetaminophen or oral ibuprofen must be used because aspirin increases the risk of Reye's syndrome. If the patient cannot take oral antipyretics, parenteral preparations of NSAIDs and rectal suppositories of various antipyretics can be used.

Treatment of fever in some patients is highly recommended. Fever increases the demand for oxygen (i.e., for every increase of 1°C over 37°C, there is a 13% increase in oxygen consumption) and can aggravate the condition of patients with preexisting impairment of cardiac, pulmonary, or CNS function. Children with a

history of febrile or nonfebrile seizure should be aggressively treated to reduce fever. However, it is unclear what triggers the febrile seizure, and there is no correlation between absolute temperature elevation and onset of a febrile seizure in susceptible children.

In hyperpyrexia, the use of cooling blankets facilitates the reduction of temperature; however, cooling blankets should not be used without oral antipyretics. In hyperpyretic patients with CNS disease or trauma (CNS bleeding), reducing core temperature mitigates the detrimental effects of high temperature on the brain.

For a discussion of treatment for hyperthermia, see Chap. 455.

TABLE 15-1

Autoinflammatory Diseases in Which Fever Is Characteristic

Adult and juvenile Still's disease
Cryopyrin-associated periodic syndromes (CAPS)
Familial Mediterranean fever
Hyper-IgD syndrome
Behçet's syndrome
Macrophage activation syndrome
Normocomplementemic urticarial vasculitis
Antisynthetase myositis
PAPA[a] syndrome
Blau syndrome
Gouty arthritis

[a]Pyogenic arthritis, pyoderma gangrenosum, and acne.

FURTHER READING

Dinarello CA et al: Treating inflammation by blocking interleukin-1 in a broad spectrum of diseases. Nature Rev 11:633, 2012.

Kullenberg T et al: Long-term safety profile of anakinra in patients with severe cryopyrin-associated periodic syndromes. Rheumatology 55:1499, 2016. [PubMed: 27143789]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 16: Fever and Rash

Elaine T. Kaye; Kenneth M. Kaye

INTRODUCTION

The acutely ill patient with fever and rash often presents a diagnostic challenge for physicians, yet the distinctive appearance of an eruption in concert with a clinical syndrome can facilitate a prompt diagnosis and the institution of life-saving therapy or critical infection-control interventions. **Representative images of many of the rashes discussed in this chapter are included in Chap. A1.**

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Fever and Rash

A thorough history of patients with fever and rash includes the following relevant information: immune status, medications taken within the previous month, specific travel history, immunization status, exposure to domestic pets and other animals, history of animal (including arthropod) bites, recent dietary exposures, existence of cardiac abnormalities, presence of prosthetic material, recent exposure to ill individuals, and sexual exposures. The history should also include the site of onset of the rash and its direction and rate of spread.

A thorough physical examination entails close attention to the rash, with an assessment and precise definition of its salient features. First, it is critical to determine what *type* of lesions make up the eruption. *Macules* are flat lesions defined by an area of changed color (i.e., a blanchable erythema). *Papules* are raised, solid lesions <5 mm in diameter; *plaques* are lesions >5 mm in diameter with a flat, plateau-like surface; and *nodules* are lesions >5 mm in diameter with a more rounded configuration. *Wheals* (urticaria, hives) are papules or plaques that are pale pink and may appear annular (ringlike) as they enlarge; classic (nonvasculitic) wheals are transient, lasting only 24 h in any defined area. *Vesicles* (<5 mm) and *bullae* (>5 mm) are circumscribed, elevated lesions containing fluid. *Pustules* are raised lesions containing purulent exudate; vesicular processes such as varicella or herpes simplex may evolve to pustules. *Nonpalpable purpura* is a flat lesion that is due to bleeding into the skin. If <3 mm in diameter, the purpuric lesions are termed *petechiae*; if >3 mm, they are termed *ecchymoses. Palpable purpura* is a raised lesion that is due to inflammation of the vessel wall (vasculitis) with subsequent hemorrhage. An *ulcer* is a defect in the skin extending at least into the upper layer of the dermis, and an *eschar* (tâche noire) is a necrotic lesion covered with a black crust.

Other pertinent features of rashes include their *configuration* (i.e., annular or target), the *arrangement* of their lesions, and their *distribution* (i.e., central or peripheral).

For further discussion, see Chaps. 52, 54, 117, and 124.

CLASSIFICATION OF RASH

This chapter reviews rashes that reflect systemic disease, but it does not include localized skin eruptions (i.e., cellulitis, impetigo) that may also be associated with fever (Chap. 124). The chapter is not intended to be all-inclusive, but it covers the most important and most common diseases associated with fever and rash. Rashes are classified herein on the basis of lesion morphology and distribution. For practical purposes, this classification system is based on the most typical disease presentations. However, morphology may vary as rashes evolve, and the presentation of diseases with rashes is subject to many variations (Chap. 54). For instance, the classic petechial rash of Rocky Mountain spotted fever (Chap. 182) may initially consist of blanchable erythematous macules distributed peripherally; at times, however, the rash associated with this disease may not be predominantly acral, or no rash may develop at all.

Diseases with fever and rash may be classified by type of eruption: centrally distributed maculopapular, peripheral, confluent desquamative erythematous, vesiculobullous, urticaria-like, nodular, purpuric, ulcerated, or with eschars. Diseases are listed by these categories in Table 16-1, and many are highlighted in the text. However, for a more detailed discussion of each disease associated with a rash, the reader is referred to the chapter dealing with that specific disease. **(Reference chapters are cited in the text and listed in Table 16-1.)**

TABLE 16-1

Diseases Associated with Fever and Rash

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Acute meningococcemia[a]	—	—	—	—	150
Drug reaction with eosinophilia and systemic symptoms (DRESS); also termed drug-induced hypersensitivity syndrome (DIHS)[b]	—	—	—	—	56
Rubeola (measles, first disease)	Paramyxovirus	Discrete lesions that become confluent as rash spreads from hairline downward, usually sparing palms and soles; lasts ≥3 days; Koplik's spots	Nonimmune individuals	Cough, conjunctivitis, coryza, severe prostration	200
Rubella (German measles, third disease)	Togavirus	Spreads from hairline downward, clearing as it spreads; Forschheimer spots	Nonimmune individuals	Adenopathy, arthritis	201

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Erythema infectiosum (fifth disease)	Human parvovirus B19	Bright-red "slapped-cheeks" appearance followed by lacy reticular rash that waxes and wanes over 3 weeks; rarely, papular-purpuric "gloves-and-socks" syndrome on hands and feet	Most common among children 3–12 years old; occurs in winter and spring	Mild fever; arthritis in adults; rash following resolution of fever	192
Exanthem subitum (roseola, sixth disease)	Human herpesvirus 6	Diffuse maculopapular eruption over trunk and neck; resolves within 2 days	Usually affects children <3 years old	Rash following resolution of fever; similar to Boston exanthem (echovirus 16); febrile seizures may occur	190
Primary HIV infection	HIV	Nonspecific diffuse macules and papules; less commonly, urticarial or vesicular oral or genital ulcers	Individuals recently infected with HIV	Pharyngitis, adenopathy, arthralgias	197

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Infectious mononucleosis	Epstein-Barr virus	Diffuse maculopapular eruption (5% of cases; 30–90% if ampicillin is given); urticaria, petechiae in some cases; periorbital edema (50%); palatal petechiae (25%)	Adolescents, young adults	Hepatosplenomegaly, pharyngitis, cervical lymphadenopathy, atypical lymphocytosis, heterophile antibody	189
Other viral exanthems	Echoviruses 2, 4, 9, 11, 16, 19, 25; coxsackieviruses A9, B1, B5; etc.	Wide range of skin findings that may mimic rubella or measles	Affect children more commonly than adults	Nonspecific viral syndromes	199
Exanthematous drug-induced eruption	Drugs (antibiotics, anticonvulsants, diuretics, etc.)	Intensely pruritic, bright-red macules and papules, symmetric on trunk and extremities; may become confluent	Occurs 2–3 days after exposure in previously sensitized individuals; otherwise, after 2–3 weeks (but can occur anytime, even shortly after drug is discontinued)	Variable findings: fever and eosinophilia	56

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Epidemic typhus	*Rickettsia prowazekii*	Maculopapular eruption appearing in axillae, spreading to trunk and later to extremities; usually spares face, palms, soles; evolves from blanchable macules to confluent eruption with petechiae; rash evanescent in recrudescent typhus (Brill-Zinsser disease)	Exposure to body lice; occurrence of recrudescent typhus as relapse after 30–50 years	Headache, myalgias; mortality rates 10–40% if untreated; milder clinical presentation in recrudescent form	182
Endemic (murine) typhus	*Rickettsia typhi*	Maculopapular eruption, usually sparing palms, soles	Exposure to rat or cat fleas	Headache, myalgias	182
Scrub typhus	*Orientia tsutsugamushi*	Diffuse macular rash starting on trunk; eschar at site of mite bite	Endemic in South Pacific, Australia, Asia; transmitted by mites	Headache, myalgias, regional adenopathy; mortality rates up to 30% if untreated	182
Rickettsial spotted fevers	*Rickettsia conorii* (boutonneuse fever), *Rickettsia australis* (North Queensland tick typhus), *Rickettsia sibirica* (Siberian tick typhus), and others	Eschar common at bite site; maculopapular (rarely, vesicular and petechial) eruption on proximal extremities, spreading to trunk and face	Exposure to ticks; *R. conorii* in Mediterranean region, India, Africa; *R. australis* in Australia; *R. sibirica* in Siberia, Mongolia	Headache, myalgias, regional adenopathy	182

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Human monocytotropic ehrlichiosis[c]	*Ehrlichia chaffeensis*	Maculopapular eruption (40% of cases), involves trunk and extremities; may be petechial	Tick-borne; most common in U.S. Southeast, southern Midwest, and mid-Atlantic regions	Headache, myalgias, leukopenia	182
Leptospirosis	*Leptospira interrogans* and other *Leptospira* species	Maculopapular eruption; conjunctivitis; scleral hemorrhage in some cases	Exposure to water contaminated with animal urine	Myalgias; aseptic meningitis; *fulminant form*: icterohemorrhagic fever (Weil's disease)	179
Lyme disease	*Borrelia burgdorferi* (sole cause in U.S.), *Borrelia afzelii, Borrelia garinii*	Papule expanding to erythematous annular lesion with central clearing (erythema migrans; average diameter, 15 cm), sometimes with concentric rings, sometimes with indurated or vesicular center; multiple secondary erythema migrans lesions in some cases	Bite of *Ixodes* tick vector	Headache, myalgias, chills, photophobia occurring acutely; CNS disease, myocardial disease, arthritis weeks to months later in some cases	181

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Southern tick-associated rash illness (STARI, Master's disease)	Unknown (possibly *Borrelia lonestari* or other *Borrelia* spirochetes)	Similar to erythema migrans of Lyme disease with several differences, including: multiple secondary lesions less likely; lesions tending to be smaller (average diameter, ~8 cm); central clearing more likely	Bite of tick vector *Amblyomma americanum* (Lone Star tick); often found in regions where Lyme disease is uncommon, including southern United States	Compared with Lyme disease: fewer constitutional symptoms, tick bite more likely to be recalled; other Lyme disease sequelae lacking	181
Typhoid fever	*Salmonella typhi*	Transient, blanchable erythematous macules and papules, 2–4 mm, usually on trunk (rose spots)	Ingestion of contaminated food or water (rare in U.S.)	Variable abdominal pain and diarrhea; headache, myalgias, hepatosplenomegaly	160

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Dengue fever[d]	Dengue virus (4 serotypes; flaviviruses)	Rash in 50% of cases; initially diffuse flushing; midway through illness, onset of maculopapular rash, which begins on trunk and spreads centrifugally to extremities and face; pruritus, hyperesthesia in some cases; after defervescence, petechiae on extremities may occur	Occurs in tropics and subtropics; transmitted by mosquito	Headache; musculoskeletal pain ("breakbone fever"); leukopenia; occasionally biphasic ("saddleback") fever	204
Rat-bite fever (sodoku)	*Spirillum minus*	Eschar at bite site; then blotchy violaceous or red-brown rash involving trunk and extremities	Rat bite; primarily found in Asia; rare in U.S.	Regional adenopathy; recurrent fevers if untreated	136
Relapsing fever	*Borrelia* species	Central rash at end of febrile episode; petechiae in some cases	Exposure to ticks or body lice	Recurrent fever, headache, myalgias, hepatosplenomegaly	180

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Erythema marginatum (rheumatic fever)	Group A *Streptococcus*	Erythematous annular papules and plaques occurring as polycyclic lesions in waves over trunk, proximal extremities; evolving and resolving within hours	Patients with rheumatic fever	Pharyngitis preceding polyarthritis, carditis, subcutaneous nodules, chorea	381
Systemic lupus erythematosus (SLE)	Autoimmune disease	Macular and papular erythema, often in sun-exposed areas; discoid lupus lesions (local atrophy, scale, pigmentary changes); periungual telangiectasis; malar rash; vasculitis sometimes causing urticaria, palpable purpura; oral erosions in some cases	Most common in young to middle-aged women; flares precipitated by sun exposure	Arthritis; cardiac, pulmonary, renal, hematologic, and vasculitic disease	352

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Still's disease	Autoimmune disease	Transient 2- to 5-mm erythematous papules appearing at height of fever on trunk, proximal extremities; lesions evanescent	Children and young adults	High spiking fever, polyarthritis, splenomegaly; erythrocyte sedimentation rate, >100 mm/h	—
African trypanosomiasis	*Trypanosoma brucei rhodesiense/gambiense*	Blotchy or annular erythematous macular and papular rash (trypanid), primarily on trunk; pruritus; chancre at site of tsetse fly bite may precede rash by several weeks	Tsetse fly bite in eastern (*T. brucei rhodesiense)* or western *(T. brucei gambiense)* Africa	Hemolymphatic disease followed by meningoencephalitis; Winterbottom's sign (posterior cervical lymphadenopathy) *(T. brucei gambiense)*	222
Arcanobacterial pharyngitis	*Arcanobacterium (Corynebacterium) haemolyticum*	Diffuse, erythematous, maculopapular eruption involving trunk and proximal extremities; may desquamate	Children and young adults	Exudative pharyngitis, lymphadenopathy	145

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
West Nile fever	West Nile virus	Maculopapular eruption involving the trunk, extremities, and head or neck; rash in 20–50% of cases	Mosquito bite; rarely, blood transfusion or transplanted organ	Headache, weakness, malaise, myalgia, neuroinvasive disease (encephalitis, meningitis, flaccid paralysis)	204
Zika virus infection	Zika virus	Pruritic macular and papular erythema; rash may begin on trunk and descend to lower body; conjunctival injection; palatal petechiae may occur	Mosquito bite; sexual transmission or blood transfusion less common	Arthralgia (especially of small joints), myalgia, lymphadenopathy, headache, low-grade fever; illness in pregnancy may cause severe birth defects, including microcephaly; neurologic complications, including Guillain-Barré, may occur	204
Peripheral Eruptions					
Chronic meningococcemia, disseminated gonococcal infection,[a] human parvovirus B19 infection[e]	—	—	—	—	150, 151, 192

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Rocky Mountain spotted fever	*Rickettsia rickettsii*	Rash beginning on wrists and ankles and spreading centripetally; appears on palms and soles later in disease; lesion evolution from blanchable macules to petechiae	Tick vector; widespread but more common in southeastern and southwest-central U.S.	Headache, myalgias, abdominal pain; mortality rates up to 40% if untreated	182
Secondary syphilis	*Treponema pallidum*	Coincident primary chancre in 10% of cases; copper-colored, scaly papular eruption, diffuse but prominent on palms and soles; rash never vesicular in adults; condyloma latum, mucous patches, and alopecia in some cases	Sexually transmitted	Fever, constitutional symptoms	177
Chikungunya fever	Chikungunya virus	Maculopapular eruption; typically occurs on trunk, but also occurs on extremities and face	*Aedes aegypti* and *A. albopictus* mosquito bites; tropical and subtropical regions	Severe polyarticular, migratory arthralgias, especially involving small joints (e.g., hands, wrists, ankles)	204

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Hand-foot-and-mouth disease	Coxsackievirus A16 and enterovirus 71 most common causes; coxsackievirus A6 associated with atypical syndrome	Tender vesicles, erosions in mouth; 0.25-cm papules on hands and feet with rim of erythema evolving into tender vesicles; shedding of nails can occur 1–2 months after acute illness; coxsackievirus A6 lesions extend to perioral area, extremities, trunk, buttocks, genitals, and areas affected by eczema	Summer and fall; primarily children <10 years old; multiple family members; coxsackievirus A6 infection also occurs in young adults	Transient fever; enterovirus 71 can be associated with brain stem encephalitis, flaccid paralysis resembling polio, or aseptic meningitis	199

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Erythema multiforme (EM)	Infection, drugs, idiopathic causes	Target lesions (central erythema surrounded by area of clearing and another rim of erythema) up to 2 cm; symmetric on knees, elbows, palms, soles; spreads centripetally; papular, sometimes vesicular; when extensive and involving mucous membranes, termed *EM major*	Herpes simplex virus or *Mycoplasma pneumoniae* infection; drug intake (i.e., sulfa, phenytoin, penicillin)	50% of patients <20 years old; fever more common in most severe form, EM major, which can be confused with Stevens-Johnson syndrome (but EM major lacks prominent skin sloughing)	—[f]
Rat-bite fever (Haverhill fever)	*Streptobacillus moniliformis*	Maculopapular eruption over palms, soles, and extremities; tends to be more severe at joints; eruption sometimes becoming generalized; may be purpuric; may desquamate	Rat bite, ingestion of contaminated food	Myalgias; arthritis (50%); fever recurrence in some cases	136

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Bacterial endocarditis	*Streptococcus*, *Staphylococcus*, etc.	*Subacute course* (e.g., viridans streptococci): Osler's nodes (tender pink nodules on finger or toe pads); petechiae on skin and mucosa; splinter hemorrhages. *Acute course* (e.g., *Staphylococcus aureus*): Janeway lesions (painless erythematous or hemorrhagic macules, usually on palms and soles)	Abnormal heart valve (e.g., viridans streptococci), intravenous drug use	New or changing heart murmur	123
Confluent Desquamative Erythemas					

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Scarlet fever (second disease)	Group A *Streptococcus* (pyrogenic exotoxins A, B, C)	Diffuse blanchable erythema beginning on face and spreading to trunk and extremities; circumoral pallor; "sandpaper" texture to skin; accentuation of linear erythema in skin folds (Pastia's lines); enanthem of white evolving into red "strawberry" tongue; desquamation in second week	Most common among children 2–10 years old; usually follows group A streptococcal pharyngitis	Fever, pharyngitis, headache	143
Kawasaki disease	Idiopathic causes	Rash similar to scarlet fever (scarlatiniform) or EM; fissuring of lips, strawberry tongue; conjunctivitis; edema of hands, feet; desquamation later in disease	Children <8 years old	Cervical adenopathy, pharyngitis, coronary artery vasculitis	54, 356

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Streptococcal toxic shock syndrome	Group A *Streptococcus* (associated with pyrogenic exotoxin A and/or B or certain M types)	When present, rash often scarlatiniform	May occur in setting of severe group A streptococcal infections (e.g., necrotizing fasciitis, bacteremia, pneumonia)	Multiorgan failure, hypotension; mortality rate 30%	143
Staphylococcal toxic shock syndrome	*S. aureus* (toxic shock syndrome toxin 1, enterotoxins B and others)	Diffuse erythema involving palms; pronounced erythema of mucosal surfaces; conjunctivitis; desquamation 7–10 days into illness	Colonization with toxin-producing *S. aureus*	Fever >39°C (>102°F), hypotension, multiorgan dysfunction	142
Staphylococcal scalded-skin syndrome	*S. aureus*, phage group II	Diffuse tender erythema, often with bullae and desquamation; Nikolsky's sign	Colonization with toxin-producing *S. aureus*; occurs in children <10 years old (termed *Ritter's disease* in neonates) or adults with renal dysfunction	Irritability; nasal or conjunctival secretions	142
Exfoliative erythroderma syndrome	Underlying psoriasis, eczema, drug eruption, mycosis fungoides	Diffuse erythema (often scaling) interspersed with lesions of underlying condition	Usually occurs in adults over age 50; more common among men	Fever, chills (i.e., difficulty with thermoregulation); lymphadenopathy	54, 56

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
DRESS [drug-induced hypersensitivity syndrome (DIHS)]	Aromatic anticonvulsants; other drugs, including sulfonamides, minocycline	Maculopapular eruption (mimicking exanthematous drug rash), sometimes progressing to exfoliative erythroderma; profound edema, especially facial; pustules may occur	Individuals genetically unable to detoxify arene oxides (anticonvulsant metabolites), patients with slow *N*-acetylating capacity (sulfonamides)	Lymphadenopathy, multiorgan failure (especially hepatic), eosinophilia, atypical lymphocytes; mimics sepsis	56
Stevens-Johnson syndrome (SJS), toxic epidermal necrolysis (TEN)	Drugs (80% of cases; often allopurinol, anticonvulsants, antibiotics), infection, idiopathic factors	Erythematous and purpuric macules, sometimes targetoid, or diffuse erythema progressing to bullae, with sloughing and necrosis of entire epidermis; Nikolsky's sign; involves mucosal surfaces; TEN (>30% epidermal necrosis) is maximal form; SJS involves <10% of epidermis; SJS/TEN overlap involves 10–30% of epidermis	Uncommon among children; more common among patients with HIV infection, systemic lupus erythematosus, certain HLA types, or slow acetylators	Dehydration, sepsis sometimes resulting from lack of normal skin integrity; mortality rates up to 30%	56
Vesiculobullous or Pustular Eruptions					

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Hand-foot-and-mouth syndrome[g]; staphylococcal scalded-skin syndrome; TEN[b];DRESS[b]	—	—	—	—	—[f]
Varicella (chickenpox)	VZV	Macules (2–3 mm) evolving into papules, then vesicles (sometimes umbilicated), on an erythematous base (“dewdrops on a rose petal”); pustules then forming and crusting; lesions appearing in crops; may involve scalp, mouth; intensely pruritic	Usually affects children; 10% of adults susceptible; most common in late winter and spring; incidence down by 90% in U.S. as a result of varicella vaccination	Malaise; generally mild disease in healthy children; more severe disease with complications in adults and immunocompromised children	188

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Pseudomonas "hot-tub" folliculitis	*Pseudomonas aeruginosa*	Pruritic erythematous follicular, papular, vesicular, or pustular lesions that may involve axillae, buttocks, abdomen, and especially areas occluded by bathing suits; can manifest as tender isolated nodules on palmar or plantar surfaces (the latter designated "*Pseudomonas* hot-foot syndrome")	Bathers in hot tubs or swimming pools; occurs in outbreaks	Earache, sore eyes and/or throat; fever may be absent; generally self-limited	159

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Variola (smallpox)	Variola major virus	Red macules on tongue and palate evolving to papules and vesicles; skin macules evolving to papules, then vesicles, then pustules over 1 week, with subsequent lesion crusting; lesions initially appearing on face and spreading centrifugally from trunk to extremities; differs from varicella in that (1) skin lesions in any given area are at same stage of development and (2) there is a prominent distribution of lesions on face and extremities (including palms, soles)	Nonimmune individuals exposed to smallpox	Prodrome of fever, headache, backache, myalgias; vomiting in 50% of cases	S2

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Primary herpes simplex virus (HSV) infection	HSV	Erythema rapidly followed by hallmark painful *grouped vesicles* that may evolve into pustules that ulcerate, especially on mucosal surfaces; lesions at site of inoculation: commonly gingivostomatitis for HSV-1 and genital lesions for HSV-2; recurrent disease milder (e.g., herpes labialis does not involve oral mucosa)	Primary infection most common among children and young adults for HSV-1 and among sexually active young adults for HSV-2; no fever in recurrent infection	Regional lymphadenopathy	187
Disseminated herpesvirus infection	Varicella-zoster virus (VZV) or HSV	Generalized vesicles that can evolve to pustules and ulcerations; individual lesions similar for VZV and HSV. *Zoster cutaneous dissemination*: >25 lesions extending outside involved dermatome. *HSV*: extensive, progressive mucocutaneous	Patients with immunosuppression, eczema; neonates	Visceral organ involvement (e.g., liver, lungs) in some cases; neonatal disease particularly severe	133, 187, 188

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
		lesions that may occur in absence of dissemination, sometimes disseminate in eczematous skin (eczema herpeticum); HSV visceral dissemination may occur with only localized mucocutaneous disease; in disseminated neonatal disease, skin lesions diagnostically helpful when present, but rash absent in a substantial minority of cases			
Rickettsialpox	*Rickettsia akari*	Eschar found at site of mite bite; generalized rash involving face, trunk, extremities; may involve palms and soles; <100 papules and plaques (2–10 mm); tops of lesions developing vesicles that may evolve into pustules	Seen in urban settings; transmitted by mouse mites	Headache, myalgias, regional adenopathy; mild disease	182

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Acute generalized exanthematous pustulosis	Drugs (mostly anticonvulsants or antimicrobials); also viral	Tiny sterile nonfollicular pustules on erythematous, edematous skin; begins on face and in body folds, then becomes generalized	Appears 2–21 days after start of drug therapy, depending on whether patient has been sensitized	Acute fever, pruritus, leukocytosis	56
Disseminated *Vibrio vulnificus* infection	*V. vulnificus*	Erythematous lesions evolving into hemorrhagic bullae and then into necrotic ulcers	Patients with cirrhosis, diabetes, renal failure; exposure by ingestion of contaminated saltwater, seafood	Hypotension; mortality rate 50%	163
Ecthyma gangrenosum	*P. aeruginosa*, other gram-negative rods, fungi	Indurated plaque evolving into hemorrhagic bulla or pustule that sloughs, resulting in eschar formation; erythematous halo; most common in axillary, groin, perianal regions	Usually affects neutropenic patients; occurs in up to 28% of individuals with *Pseudomonas* bacteremia	Clinical signs of sepsis	159
Urticaria-Like Eruptions					

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Urticarial vasculitis	Serum sickness, often due to infection (including hepatitis B viral, enteroviral, parasitic), drugs; connective tissue disease	Erythematous, edematous "urticaria-like" plaques, pruritic or burning; unlike urticaria: typical lesion duration >24 h (up to 5 days) and lack of complete lesion blanching with compression due to hemorrhage	Patients with serum sickness (including hepatitis B), connective tissue disease	Fever variable; arthralgias/arthritis	356[f]
Nodular Eruptions					
Disseminated infection	Fungal infections (e.g., candidiasis, histoplasmosis, cryptococcosis, sporotrichosis, coccidioidomycosis); mycobacteria	Subcutaneous nodules (up to 3 cm); fluctuance, draining common with mycobacteria; necrotic nodules (extremities, periorbital or nasal regions) common with *Aspergillus*, *Mucor*	Immunocompromised hosts (e.g., bone marrow transplant recipients, patients undergoing chemotherapy, HIV-infected patients)	Features vary with organism	__[f]

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Erythema nodosum (septal panniculitis)	Infections (e.g., streptococcal, fungal, mycobacterial, yersinial); drugs (e.g., sulfas, penicillins, oral contraceptives); sarcoidosis; idiopathic causes	Large, violaceous, nonulcerative, subcutaneous nodules; exquisitely tender; usually on lower legs but also on upper extremities	More common among girls and women 15–30 years old	Arthralgias (50%); features vary with associated condition	—[f]
Sweet syndrome (acute febrile neutrophilic dermatosis)	*Yersinia* infection; upper respiratory infection; inflammatory bowel disease; pregnancy; malignancy (usually hematologic); drugs (G-CSF)	Tender red or blue edematous nodules giving impression of vesiculation; usually on face, neck, upper extremities; when on lower extremities, may mimic erythema nodosum	More common among women and among persons 30–60 years old; 20% of cases associated with malignancy (men and women equally affected in this group)	Headache, arthralgias, leukocytosis	54
Bacillary angiomatosis	*Bartonella henselae, B. quintana*	Many forms, including erythematous, smooth vascular nodules; friable, exophytic lesions; erythematous plaques (may be dry, scaly); subcutaneous nodules (may be erythematous)	Immunosuppressed individuals, especially those with advanced HIV infection	Peliosis of liver and spleen in some cases; lesions sometimes involving multiple organs; bacteremia	167
Purpuric Eruptions					

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Rocky Mountain spotted fever, rat-bite fever, endocarditis[g]; epidemic typhus[e]; dengue fever[d,e]; human parvovirus B19 infection[e]	—	—	—	—	—[f]
Acute meningococcemia	*Neisseria meningitidis*	Initially pink maculopapular lesions evolving into petechiae; petechiae rapidly becoming numerous, sometimes enlarging and becoming vesicular; trunk, extremities most commonly involved; may appear on face, hands, feet; may include purpura fulminans (see below) reflecting DIC	Most common among children, individuals with asplenia or terminal complement component deficiency (C5–C8)	Hypotension, meningitis (sometimes preceded by upper respiratory infection)	150
Purpura fulminans	Severe DIC	Large ecchymoses with sharply irregular shapes evolving into hemorrhagic bullae and then into black necrotic lesions	Individuals with sepsis (e.g., involving *N. meningitidis*), malignancy, or massive trauma; asplenic patients at high risk for sepsis	Hypotension	150, 297

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Chronic meningococcemia	*N. meningitidis*	Variety of recurrent eruptions, including pink maculopapular; nodular (usually on lower extremities); petechial (sometimes developing vesicular centers); purpuric areas with pale blue-gray centers	Individuals with complement deficiencies	Fevers, sometimes intermittent; arthritis, myalgias, headache	150
Disseminated gonococcal infection	*Neisseria gonorrhoeae*	Papules (1–5 mm) evolving over 1–2 days into hemorrhagic pustules with gray necrotic centers; hemorrhagic bullae occurring rarely; lesions (usually <40) distributed peripherally near joints (more commonly on upper extremities)	Sexually active individuals (more often females), some with complement deficiency	Low-grade fever, tenosynovitis, arthritis	151

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Enteroviral petechial rash	Usually echovirus 9 or coxsackievirus A9	Disseminated petechial lesions (may also be maculopapular, vesicular, or urticarial)	Often occurs in outbreaks	Pharyngitis, headache; aseptic meningitis with echovirus 9	199
Viral hemorrhagic fever	Arboviruses (including dengue) and arenaviruses	Petechial rash	Residence in or travel to endemic areas, other virus exposure	Triad of fever, shock, hemorrhage from mucosa or gastrointestinal tract	204, 205
Thrombotic thrombocytopenic purpura/hemolytic-uremic syndrome	Idiopathic, bloody diarrhea caused by Shiga toxin–generating bacteria (e.g., *Escherichia coli* O157:H7), deficiency in ADAMTS13 (cleaves von Willebrand factor), drugs (e.g., quinine, chemotherapy, immunosuppression)	Petechiae	Individuals with *E. coli* O157:H7 gastroenteritis (especially children), cancer chemotherapy, HIV infection, autoimmune diseases, pregnant/postpartum women	Fever (not always present), microangiopathic hemolytic anemia, thrombocytopenia, renal dysfunction, neurologic dysfunction; coagulation studies normal	54, 96, 111, 156, 161
Cutaneous small-vessel vasculitis (leukocytoclastic vasculitis)	Infections (including group A streptococcal infection, hepatitis B or C), drugs, idiopathic factors	Palpable purpuric lesions appearing in crops on legs or other dependent areas; may become vesicular or ulcerative	Occurs in a wide spectrum of diseases, including connective tissue disease, cryoglobulinemia, malignancy, Henoch-Schönlein purpura (HSP); more common among children	Fever (not always present), malaise, arthralgias, myalgias; systemic vasculitis in some cases; renal, joint, and gastrointestinal involvement common in HSP	54
Eruptions with Ulcers and/or Eschars					

Disease	Etiology	Description	Group Affected/Epidemiologic Factors	Clinical Syndrome	Chapter
Centrally Distributed Maculopapular Eruptions					
Scrub typhus, rickettsial spotted fevers, rat-bite fever[e]; rickettsialpox, ecthyma gangrenosum[h]	—	—	—	—	—[f]
Tularemia	*Francisella tularensis*	Ulceroglandular form: erythematous, tender papule evolves into necrotic, tender ulcer with raised borders; in 35% of cases, eruptions (maculopapular, vesiculopapular, acneiform, or urticarial; erythema nodosum; or EM) may occur	Exposure to ticks, biting flies, infected animals	Fever, headache, lymphadenopathy	165
Anthrax	*Bacillus anthracis*	Pruritic papule enlarging and evolving into a 1- by 3-cm painless ulcer surrounded by vesicles and then developing a central eschar with edema; residual scar	Exposure to infected animals or animal products, other exposure to anthrax spores	Lymphadenopathy, headache	S2

[a]See "Purpuric Eruptions." [b]See "Confluent Desquamative Erythemas." [c]Rash is rare in human granulocytotropic ehrlichiosis or anaplasmosis (caused by *Anaplasma phagocytophila*; most common in the upper midwestern and northeastern United States). [d]See "Viral hemorrhagic fever" under "Purpuric Eruptions" for dengue hemorrhagic fever/dengue shock syndrome. [e]See "Centrally Distributed Maculopapular Eruptions." [f]See etiology-specific chapters. [g]See "Peripheral Eruptions." [h]See "Vesiculobullous or Pustular Eruptions."

Abbreviations: CNS, central nervous system; DIC, disseminated intravascular coagulation; G-CSF, granulocyte colony-stimulating factor; HLA, human leukocyte antigen.

CENTRALLY DISTRIBUTED MACULOPAPULAR ERUPTIONS

Centrally distributed rashes, in which lesions are primarily truncal, are the most common type of eruption. The rash of *rubeola* (measles) starts at the hairline 2–3 days into the illness and moves down the body, typically sparing the palms and soles **(Chap. 200)**. It begins as discrete erythematous lesions, which become confluent as the rash spreads. Koplik's spots (1- to 2-mm white or bluish lesions with an erythematous halo on the buccal mucosa) are pathognomonic for measles and are generally seen during the first 2 days of symptoms. They should not be confused with Fordyce's spots (ectopic sebaceous glands), which have no erythematous halos and are found in the mouth of healthy individuals. Koplik's spots may briefly overlap with the measles exanthem.

Rubella (German measles) also spreads from the hairline downward; unlike that of measles, however, the rash of rubella tends to clear from originally affected areas as it migrates, and it may be pruritic **(Chap. 201)**. Forchheimer spots (palatal petechiae) may develop but are nonspecific because they also develop in *infectious mononucleosis* **(Chap. 189)**, *scarlet fever* **(Chap. 143)**, and *Zika virus infection* **(Chap. 204)**. Postauricular and suboccipital adenopathy and arthritis are common among adults with rubella. Exposure of pregnant women to ill individuals should be avoided, as rubella causes severe congenital abnormalities. Numerous strains of *enteroviruses* **(Chap. 199)**, primarily echoviruses and coxsackieviruses, cause nonspecific syndromes of fever and eruptions that may mimic rubella or measles. Patients with *infectious mononucleosis* caused by Epstein-Barr virus **(Chap. 189)** or with *primary HIV infection* **(Chap. 197)** may exhibit pharyngitis, lymphadenopathy, and a nonspecific maculopapular exanthem.

The rash of *erythema infectiosum* (fifth disease), which is caused by human parvovirus B19, primarily affects children 3–12 years old; it develops after fever has resolved as a bright blanchable erythema on the cheeks ("slapped cheeks") with perioral pallor **(Chap. 192)**. A more diffuse rash (often pruritic) appears the next day on the trunk and extremities and then rapidly develops into a lacy reticular eruption that may wax and wane (especially with temperature change) over 3 weeks. Adults with fifth disease often have arthritis, and fetal hydrops can develop in association with this condition in pregnant women.

Exanthem subitum (roseola) is caused by human herpesvirus 6 and is most common among children <3 years of age **(Chap. 190)**. As in erythema infectiosum, the rash usually appears after fever has subsided. It consists of 2- to 3-mm rose-pink macules and papules that coalesce only rarely, occur initially on the trunk and sometimes on the extremities (sparing the face), and fade within 2 days.

Although drug reactions have many manifestations, including urticaria, exanthematous *drug-induced eruptions* **(Chap. 56)** are most common and are often difficult to distinguish from viral exanthems. Eruptions elicited by drugs are usually more intensely erythematous and pruritic than viral exanthems, but this distinction is not reliable. A history of new medications and an absence of prostration may help to distinguish a drug-related rash from an eruption of another etiology. Rashes may persist for up to 2 weeks after administration of the offending agent is discontinued. Certain populations are more prone

than others to drug rashes. Of HIV-infected patients, 50–60% develop a rash in response to sulfa drugs; 30–90% of patients with mononucleosis due to Epstein-Barr virus develop a rash when given ampicillin.

Rickettsial illnesses (Chap. 182) should be considered in the evaluation of individuals with centrally distributed maculopapular eruptions. The usual setting for *epidemic typhus* is a site of war or natural disaster in which people are exposed to body lice. Endemic typhus or *leptospirosis* (the latter caused by a spirochete) (Chap. 179) may be seen in urban environments where rodents proliferate. Outside the United States, other rickettsial diseases cause a spotted-fever syndrome and should be considered in residents of or travelers to endemic areas. Similarly, *typhoid fever*, a nonrickettsial disease caused by *Salmonella typhi* (Chap. 160), is usually acquired during travel outside the United States. *Dengue fever*, caused by a mosquito-transmitted flavivirus, occurs in tropical and subtropical regions of the world (Chap. 204).

Some centrally distributed maculopapular eruptions have distinctive features. Erythema migrans, the rash of *Lyme disease* (Chap. 181), typically manifests as single or multiple annular lesions. Untreated erythema migrans lesions usually fade within a month but may persist for more than a year. *Southern tick-associated rash illness* (STARI) (Chap. 181) has an erythema migrans–like rash, but is less severe than Lyme disease and often occurs in regions where Lyme is not endemic. Erythema marginatum, the rash of *acute rheumatic fever* (Chap. 352), has a distinctive pattern of enlarging and shifting transient annular lesions.

Collagen vascular diseases may cause fever and rash. Patients with *systemic lupus erythematosus* (Chap. 349) typically develop a sharply defined, erythematous eruption in a butterfly distribution on the cheeks (malar rash) as well as many other skin manifestations. *Still's disease* presents as an evanescent, salmon-colored rash on the trunk and proximal extremities that coincides with fever spikes.

Zika virus is a mosquito-transmitted flavivirus that is associated with severe birth defects (Chap. 204). Zika is rapidly spreading among tropical and subtropical regions of the world. The eruption of Zika virus infection is typically pruritic and often accompanied by conjunctival injection.

PERIPHERAL ERUPTIONS

These rashes are alike in that they are most prominent peripherally or begin in peripheral (acral) areas before spreading centripetally. Early diagnosis and therapy are critical in *Rocky Mountain spotted fever* (Chap. 182) because of its grave prognosis if untreated. Lesions evolve from macular to petechial, start on the wrists and ankles, spread centripetally, and appear on the palms and soles only later in the disease. The rash of *secondary syphilis* (Chap. 177), which may be generalized but is prominent on the palms and soles, should be considered in the differential diagnosis of pityriasis rosea, especially in sexually active patients. *Chikungunya fever* (Chap. 204), which is transmitted by mosquito bite in tropical and subtropical regions, is associated with a maculopapular eruption and severe polyarticular small-joint arthralgias. *Hand-foot-and-mouth disease* (Chap. 199), most commonly caused by coxsackievirus A16 or enterovirus 71, is distinguished by tender vesicles distributed on the hands and feet and in the mouth; coxsackievirus A6 causes an atypical syndrome with more extensive lesions. The classic target lesions of *erythema multiforme* appear symmetrically on the elbows, knees, palms, soles, and face. In severe cases, these lesions spread diffusely and involve mucosal surfaces. Lesions may develop on the hands and feet in *endocarditis* (Chap. 123).

CONFLUENT DESQUAMATIVE ERYTHEMAS

These eruptions consist of diffuse erythema frequently followed by desquamation. The eruptions caused by group A *Streptococcus* or *Staphylococcus aureus* are toxin-mediated. *Scarlet fever* (Chap. 143) usually follows pharyngitis; patients have a facial flush, a "strawberry" tongue, and accentuated petechiae in body folds (Pastia's lines). *Kawasaki disease*

(Chaps. 54 and 356) presents in the pediatric population as fissuring of the lips, a strawberry tongue, conjunctivitis, adenopathy, and sometimes cardiac abnormalities. *Streptococcal toxic shock syndrome* (Chap. 143) manifests with hypotension, multiorgan failure, and, often, a severe group A streptococcal infection (e.g., necrotizing fasciitis). *Staphylococcal toxic shock syndrome* (Chap. 142) also presents with hypotension and multiorgan failure, but usually only *S. aureus* colonization—not a severe *S. aureus* infection—is documented. *Staphylococcal scalded-skin syndrome* (Chap. 142) is seen primarily in children and in immunocompromised adults. Generalized erythema is often evident during the prodrome of fever and malaise; profound tenderness of the skin is distinctive. In the exfoliative stage, the skin can be induced to form bullae with light lateral pressure (Nikolsky's sign). In a mild form, a scarlatiniform eruption mimics scarlet fever, but the patient does not exhibit a strawberry tongue or circumoral pallor. In contrast to the staphylococcal scalded-skin syndrome, in which the cleavage plane is superficial in the epidermis, *toxic epidermal necrolysis* (Chap. 56), a maximal variant of *Stevens-Johnson syndrome*, involves sloughing of the entire epidermis, resulting in severe disease. *Exfoliative erythroderma syndrome* (Chaps. 54 and 56) is a serious reaction associated with systemic toxicity that is often due to eczema, psoriasis, a drug reaction, or mycosis fungoides. *Drug rash with eosinophilia and systemic symptoms (DRESS)*, often due to antiepileptic and antibiotic agents (Chap. 56), initially appears similar to an exanthematous drug reaction but may progress to exfoliative erythroderma; it is accompanied by multiorgan failure and has an associated mortality rate of ~10%.

VESICULOBULLOUS OR PUSTULAR ERUPTIONS

Varicella (Chap. 188) is highly contagious, often occurring in winter or spring, and is characterized by pruritic lesions that, within a given region of the body, are in different stages of development at any point in time. In immunocompromised hosts, varicella vesicles may lack the characteristic erythematous base or may appear hemorrhagic. Lesions of *Pseudomonas"hot-tub" folliculitis* (Chap. 159) are also pruritic and may appear similar to those of varicella. However, hot-tub folliculitis generally occurs in outbreaks after bathing in hot tubs or swimming pools, and lesions occur in regions occluded by bathing suits. Lesions of *variola (smallpox)* (Chap. S2) also appear similar to those of varicella but are all at the same stage of development in a given region of the body. Variola lesions are most prominent on the face and extremities, while varicella lesions are most prominent on the trunk. *Herpes simplex virus infection* (Chap. 187) is characterized by hallmark grouped vesicles on an erythematous base. Primary herpes infection is accompanied by fever and toxicity, while recurrent disease is milder. *Rickettsialpox* (Chap. 182) is often documented in urban settings and is characterized by vesicles followed by pustules. It can be distinguished from varicella by an eschar at the site of the mouse-mite bite and the papule/plaque base of each vesicle. *Acute generalized exanthematous pustulosis* should be considered in individuals who are acutely febrile and are taking new medications, especially anticonvulsant or antimicrobial agents (Chap. 56). Disseminated *Vibrio vulnificus* infection (Chap. 163) or *ecthyma gangrenosum* due to *Pseudomonas aeruginosa* (Chap. 159) should be considered in immunosuppressed individuals with sepsis and hemorrhagic bullae.

URTICARIA-LIKE ERUPTIONS

Individuals with classic urticaria ("hives") usually have a hypersensitivity reaction without associated fever. In the presence of fever, urticaria-like eruptions are most often due to *urticarial vasculitis* (Chap. 356). Unlike individual lesions of classic urticaria, which last up to 24 h, these lesions may last 3–5 days. Etiologies include serum sickness (often induced by drugs such as penicillins, sulfas, salicylates, or barbiturates), connective-tissue disease (e.g., systemic lupus erythematosus or Sjögren's syndrome), and infection (e.g., with hepatitis B virus, enteroviruses, or parasites). Malignancy, especially lymphoma, may be associated with fever and chronic urticaria (Chap. 54).

NODULAR ERUPTIONS

In immunocompromised hosts, nodular lesions often represent disseminated infection. Patients with disseminated *candidiasis* (often due to *Candida tropicalis*) may have a triad of fever, myalgias, and eruptive nodules (Chap. 211). Disseminated *cryptococcosis* lesions (Chap. 210) may resemble molluscum contagiosum (Chap. 191). Necrosis of nodules should raise the suspicion of *aspergillosis* (Chap. 212) or *mucormycosis* (Chap. 213). *Erythema nodosum* presents with exquisitely tender nodules on the lower extremities. *Sweet syndrome* (Chap. 54) should be considered in individuals with multiple nodules and plaques, often so edematous that they give the appearance of vesicles or bullae. Sweet syndrome may occur in individuals with infection, inflammatory bowel disease, or malignancy and can also be induced by drugs.

PURPURIC ERUPTIONS

Acute meningococcemia (Chap. 150) classically presents in children as a petechial eruption, but initial lesions may appear as blanchable macules or urticaria. Rocky Mountain spotted fever should be considered in the differential diagnosis of acute meningococcemia. *Echovirus 9 infection* (Chap. 199) may mimic acute meningococcemia; patients should be treated as if they have bacterial sepsis because prompt differentiation of these conditions may be impossible. Large ecchymotic areas of *purpura fulminans* (Chaps. 150 and 297) reflect severe underlying disseminated intravascular coagulation, which may be due to infectious or noninfectious causes. The lesions of *chronic meningococcemia* (Chap. 150) may have a variety of morphologies, including petechial. Purpuric nodules may develop on the legs and resemble erythema nodosum but lack its exquisite tenderness. Lesions of *disseminated gonococcemia* (Chap. 151) are distinctive, sparse, countable hemorrhagic pustules, usually located near joints. The lesions of chronic meningococcemia and those of gonococcemia may be indistinguishable in terms of appearance and distribution. *Viral hemorrhagic fever* (Chaps. 204 and 205) should be considered in patients with an appropriate travel history and a petechial rash. *Thrombotic thrombocytopenic purpura* (Chaps. 54, 96, and 111) and *hemolytic-uremic syndrome* (Chaps. 111, 156, and 161) are closely related and are noninfectious causes of fever and petechiae. *Cutaneous small-vessel vasculitis* (*leukocytoclastic vasculitis*) typically manifests as palpable purpura and has a wide variety of causes (Chap. 54).

ERUPTIONS WITH ULCERS OR ESCHARS

The presence of an ulcer or eschar in the setting of a more widespread eruption can provide an important diagnostic clue. For example, an eschar may suggest the diagnosis of *scrub typhus* or *rickettsialpox* (Chap. 182) in the appropriate setting. In other illnesses (e.g., anthrax) (Chap. S2), an ulcer or eschar may be the only skin manifestation.

FURTHER READING

Cherry JD: Cutaneous manifestations of systemic infections, in *Feigin and Cherry's Textbook of Pediatric Infectious Diseases,* 7th ed. JD Cherry et al (eds). Houston, Elsevier Saunders, 2014, pp 741–768.

Weber DJ et al: The acutely ill patient with fever and rash, in *Principles and Practice of Infectious Diseases,* vol 1, 8th ed. JI Bennett et al (eds). Philadelphia, Elsevier Saunders, 2015, pp 732–747.

Wolff K et al: *Fitzpatrick's Color Atlas and Synopsis of Clinical Dermatology,* 7th ed. New York, McGraw-Hill, 2013.

Wolff K et al (eds): *Fitzpatrick's Dermatology in General Medicine,* 8th ed. New York, McGraw-Hill, 2012.

Chapter 17: Fever of Unknown Origin

Chantal P. Bleeker-Rovers; Jos W. M. van der Meer

DEFINITION

Clinicians commonly refer to any febrile illness without an initially obvious etiology as *fever of unknown origin* (FUO). Most febrile illnesses either resolve before a diagnosis can be made or develop distinguishing characteristics that lead to a diagnosis. The term *FUO* should be reserved for prolonged febrile illnesses without an established etiology despite intensive evaluation and diagnostic testing. This chapter focuses on classic FUO in the adult patient.

FUO was originally defined by Petersdorf and Beeson in 1961 as an illness of >3 weeks' duration with fever of ≥38.3°C (≥101°F) on two occasions and an uncertain diagnosis despite 1 week of inpatient evaluation. Nowadays, most patients with FUO are hospitalized only if their clinical condition requires it, and not for diagnostic purposes alone; thus the in-hospital evaluation requirement has been eliminated from the definition. The definition of FUO has been further modified by the exclusion of immunocompromised patients, whose workup requires an entirely different diagnostic and therapeutic approach. For optimal comparison of patients with FUO in different geographic areas, it has been proposed that the quantitative criterion (diagnosis uncertain after 1 week of evaluation) be changed to a qualitative criterion that requires the performance of a specific list of investigations. Accordingly, FUO is now defined as follows:

1. Fever ≥38.3°C (≥101°F) on at least two occasions
2. Illness duration of ≥3 weeks
3. No known immunocompromised state
4. Diagnosis that remains uncertain after a thorough history-taking, physical examination, and the following obligatory investigations: determination of erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) level; platelet count; leukocyte count and differential; measurement of levels of hemoglobin, electrolytes, creatinine, total protein, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, lactate dehydrogenase, creatine kinase, ferritin, antinuclear antibodies, and rheumatoid factor; protein electrophoresis; urinalysis; blood cultures ($n = 3$); urine culture; chest x-ray; abdominal ultrasonography; and tuberculin skin test (TST) or interferon γ release assay (IGRA).

ETIOLOGY AND EPIDEMIOLOGY

The range of FUO etiologies has evolved over time as a result of changes in the spectrum of diseases causing FUO, the widespread use of antibiotics, and especially the availability of new diagnostic techniques. The proportion of cases caused by intraabdominal abscesses and tumors, for example, has decreased because of earlier detection by CT and ultrasound. In addition, infective endocarditis is a less frequent cause because blood culture and echocardiographic techniques have improved. Conversely, some diagnoses, such as acute HIV infection, were unknown four decades ago.

 Table 17-1 summarizes the findings of large studies on FUO conducted over the past 25 years. In general, infection accounts for about one-fifth of cases of FUO in Western countries; next in frequency are noninfectious inflammatory diseases (NIIDs, which include "collagen or rheumatic diseases," vasculitis syndromes, granulomatous disorders, and autoinflammatory syndromes) and neoplasms. In geographic areas outside the West, infections are a much more common cause of FUO (43% vs 17%), while the proportions of cases due to NIIDs and neoplasms are similar. Up to 50% of cases caused by infections in patients with FUO outside Western nations are due to tuberculosis, which is a less common cause in the United States and Western Europe. The number of FUO patients diagnosed with NIIDs probably will not decrease in the near future, as fever may precede more typical manifestations or serologic evidence of these diseases by months. Moreover, many NIIDs can be diagnosed only after prolonged observation and exclusion of other diseases.

TABLE 17-1

Etiology of Fever of Unknown Origin (FUO) Over the Past 25 Years: Findings from Large FUO Studies

First Author (Country, Year of Publication)	No. of Patients (Recruitment Period)	Percentage of Cases Due to Indicated Cause				
		Infections	Noninfectious Inflammatory Diseases	Neoplasms	Miscellaneous	Unknown
Western Countries						
De Kleijn et al.	167	26	24	13	8	30
(Netherlands, 1997)	(1992–1994)					
Vanderschueren et al.	185	11	18	10	8	53
(Belgium, 2003)	(1990–1999)					
Hot et al.	280	11	20	27	9	33
(France, 2005)	(1995–2005)					
Zenone et al.	144	23	26	10	15	26
(France, 2006)	(1999–2005)					
Bleeker-Rovers	73	16	22	7	4	51
(Netherlands, 2007)	(2003–2005)					
Mansueto et al.	91	32	12	14	10	32
(Italy, 2008)	(1991–2002)					
Vanderschueren et al.	114	15	22	13	10	40

First Author (Country, Year of Publication)	No. of Patients (Recruitment Period)	Percentage of Cases Due to Indicated Cause				
		Infections	Noninfectious Inflammatory Diseases	Neoplasms	Miscellaneous	Unknown
Western Countries						
(Belgium, 2009)	(2003–2007)					
Efstathiou et al.	112	30	33	11	5	21
(Greece, 2010)	(2001–2007)					
Pedersen et al.	52	19	33	8	0	40
(Denmark, 2012)	(2005–2010)					
Robine et al.	103	12	30	3	5	51
(France, 2014)	(2002–2012)					
Vanderschueren et al.	436	17	24	11	10	39
(Belgium, 2014)	(2000–2010)					
Total	**1757**	**19**	**24**	**12**	**8**	**38**
Other Geographic Locations						
Tabak et al.	117	34	29	19	4	14
(Turkey, 2003)	(1984–2001)					
Saltoglu et al.	87	59	18	14	2	7
(Turkey, 2004)	(1994–2002)					
Ergonul et al.	80	52	16	18	3	11

First Author (Country, Year of Publication)	No. of Patients (Recruitment Period)	Percentage of Cases Due to Indicated Cause				
		Infections	Noninfectious Inflammatory Diseases	Neoplasms	Miscellaneous	Unknown
Western Countries						
(Turkey, 2005)	(1993–1999)					
Brahim et al.	97	36	8	16	5	35
(Turkey, 2005)	(1990–2005)					
Chin et al.	94	57	7	9	9	18
(Taiwan, 2006)	(2001–2002)					
Colpan et al.	71	45	27	14	6	9
(Turkey, 2007)	(2001–2004)					
Hu et al.	142	36	32	13	5	14
(China, 2008)	(2002–2003)					
Kucukardali et al.	154	34	31	14	5	16
(Turkey, 2008)	(2003–2004)					
Ali-Eldin et al.	93	42	15	30	0	12
(Egypt, 2011)	(2009–2010)					
Bandyopadhya et al.	164	55	11	22	0	12
(India, 2011)	(2008–2009)					
Mete et al.	100	26	38	14	2	20

First Author (Country, Year of Publication)	No. of Patients (Recruitment Period)	Percentage of Cases Due to Indicated Cause				
		Infections	Noninfectious Inflammatory Diseases	Neoplasms	Miscellaneous	Unknown
Western Countries						
(Turkey, 2012)	(2001–2009)					
Ma et al.	397	49	18	16	7	10
(China, 2012)	(2000–2009)					
Ryuko et al.	174	41	27	7	6	19
(Japan, 2013)	(2004–2010)					
Mahmood et al.	205	49	20	13	2	17
(Pakistan, 2013)	(2006–2011)					
Alvi et al.	106	44	18	12	10	15
(Iran, 2013)	(2007–2011)					
Naito et al.	121	23	31	11	12	23
(Japan, 2013)	(2011)					
Yamanouchi et al.	256	28	18	10	15	29
(Japan, 2014)	(1994–2012)					
Moawad et al.	98	33	14	18	18	17
(Turkey, 2014)	(1995–2008)					
Yu et al.	107	30	17	18	14	22
(China, 2014)	(2010–2011)					

First Author (Country, Year of Publication)	No. of Patients (Recruitment Period)	Percentage of Cases Due to Indicated Cause				
		Infections	Noninfectious Inflammatory Diseases	Neoplasms	Miscellaneous	Unknown
Western Countries						
Mir et al.	91	44	12	12	4	27
(India, 2014)	(2010–2012)					
Kabapy et al.	979	79	17	1	1	2
(Egypt, 2015)	(2009–2010)					
Montasser et al.	217	66	7	7	12	8
(Egypt, 2015)	(unknown)					
Popovsa-Jovicic et al.	74	38	26	15	18	4
(Serbia, 2016)						
Total	**4024**	**43**	**20**	**14**	**7**	**16**

In the West, the proportion of patients who remain undiagnosed is higher than in non-Western populations and has been increasing over figures reported in studies before the 1990s. An important factor contributing to the seemingly high diagnostic failure rate is that a diagnosis is more often being established before 3 weeks have elapsed, given that patients with fever tend to seek medical advice earlier and that better diagnostic techniques, such as CT and MRI, are available; therefore, only the cases that are most difficult to diagnose continue to meet the criteria for FUO. Furthermore, most patients who have FUO without a diagnosis currently do well, and thus a less aggressive diagnostic approach may be used in clinically stable patients once diseases with immediate therapeutic or prognostic consequences have been ruled out to a reasonable extent. This factor may be especially relevant to patients with recurrent fever who are asymptomatic between febrile episodes. In patients with recurrent fever (defined as repeated episodes of fever interspersed with fever-free intervals of at least 2 weeks and apparent remission of the underlying disease), the chance of attaining an etiologic diagnosis is <50%.

DIFFERENTIAL DIAGNOSIS

The differential diagnosis for FUO is extensive. It is important to remember that FUO is far more often caused by an atypical presentation of a rather common disease than by a very rare disease. Table 17-2 presents an overview of possible causes of FUO. Atypical presentations of endocarditis, diverticulitis, vertebral osteomyelitis, and extrapulmonary tuberculosis are the more common infectious disease diagnoses. Q fever and Whipple's disease are quite rare but should always be kept in mind as a cause of FUO since the presenting symptoms can be nonspecific. Serologic testing for Q fever, which results from exposure to animals or animal products, should be performed when the patient lives in a rural area or has a history of heart valve disease, an aortic aneurysm, or a vascular prosthesis. In patients with unexplained symptoms localized to the central nervous system, gastrointestinal tract, or joints, polymerase chain reaction testing for *Tropheryma whipplei* should be performed. Travel to or (former) residence in tropical countries or the American Southwest should lead to consideration of infectious diseases such as malaria, leishmaniasis, histoplasmosis, or coccidioidomycosis. Fever with signs of endocarditis and negative blood culture results poses a special problem. Culture-negative endocarditis may be due to difficult-to-culture bacteria such as nutritionally variant bacteria, HACEK organisms (including *Haemophilus parainfluenzae*, *H. paraphrophilus*, *Aggregatibacter actinomycetemcomitans*, *A. aphrophilus*, *A. paraphrophilus*, *Cardiobacterium hominis*, *C. valvarum*, *Eikenella corrodens*, and *Kingella kingae*; discussed below), *Coxiella burnetii*, *T. whipplei*, and *Bartonella* species. Marantic endocarditis is a sterile thrombotic disease that occurs as a paraneoplastic phenomenon, especially with adenocarcinomas. Sterile endocarditis is also seen in the context of systemic lupus erythematosus and antiphospholipid syndrome.

TABLE 17-2

All Reported Causes of FUO[a]

Infections	
Bacterial, nonspecific	Abdominal abscess, adnexitis, apical granuloma, appendicitis, cholangitis, cholecystitis, diverticulitis, endocarditis, endometritis, epidural abscess, infected joint prosthesis, infected vascular catheter, infected vascular prosthesis, infectious arthritis, infective myonecrosis, intracranial abscess, liver abscess, lung abscess, malakoplakia, mastoiditis, mediastinitis, mycotic aneurysm, osteomyelitis, pelvic inflammatory disease, prostatitis, pyelonephritis, pylephlebitis, renal abscess, septic phlebitis, sinusitis, spondylodiscitis, xanthogranulomatous urinary tract infection
Bacterial, specific	Actinomycosis, atypical mycobacterial infection, bartonellosis, brucellosis, *Campylobacter* infection, *Chlamydia pneumoniae* infection, chronic meningococcemia, ehrlichiosis, gonococcemia, legionellosis, leptospirosis, listeriosis, louse-borne relapsing fever (*Borrelia recurrentis*), Lyme disease, melioidosis (*Pseudomonas pseudomallei*), *Mycoplasma* infection, nocardiosis, psittacosis, Q fever (*Coxiella burnetii),* rickettsiosis, *Spirillum minor* infection, *Streptobacillus moniliformis* infection, syphilis, tick-borne relapsing fever (*Borrelia duttonii*), tuberculosis, tularemia, typhoid fever and other salmonelloses, Whipple's disease (*Tropheryma whipplei*), yersiniosis
Fungal	Aspergillosis, blastomycosis, candidiasis, coccidioidomycosis, cryptococcosis, histoplasmosis, *Malassezia furfur* infection, paracoccidioidomycosis, *Pneumocystis jirovecii* pneumonia, sporotrichosis, zygomycosis
Parasitic	Amebiasis, babesiosis, echinococcosis, fascioliasis, malaria, schistosomiasis, strongyloidiasis, toxocariasis, toxoplasmosis, trichinellosis, trypanosomiasis, visceral leishmaniasis
Viral	Colorado tick fever, coxsackievirus infection, cytomegalovirus infection, dengue, Epstein-Barr virus infection, hantavirus infection, hepatitis (A, B, C, D, E), herpes simplex, HIV infection, human herpesvirus 6 infection, parvovirus infection, West Nile virus infection
Noninfectious Inflammatory Diseases	
Systemic rheumatic and autoimmune diseases	Ankylosing spondylitis, antiphospholipid syndrome, autoimmune hemolytic anemia, autoimmune hepatitis, Behçet's disease, cryoglobulinemia, dermatomyositis, Felty syndrome, gout, mixed connective-tissue disease, polymyositis, pseudogout, reactive

	arthritis, relapsing polychondritis, rheumatic fever, rheumatoid arthritis, Sjögren's syndrome, systemic lupus erythematosus, Vogt-Koyanagi-Harada syndrome
Vasculitis	Allergic vasculitis, eosinophilic granulomatosis with polyangiitis, giant cell vasculitis/polymyalgia rheumatica, granulomatosis with polyangiitis, hypersensitivity vasculitis, Kawasaki disease, polyarteritis nodosa, Takayasu arteritis, urticarial vasculitis
Granulomatous diseases	Idiopathic granulomatous hepatitis, sarcoidosis
Autoinflammatory syndromes	Adult-onset Still's disease, Blau syndrome, CAPS[b] (cryopyrin-associated periodic syndromes), Crohn's disease, DIRA (deficiency of the interleukin 1 receptor antagonist), familial Mediterranean fever, hemophagocytic syndrome, hyper-IgD syndrome (HIDS, also known as mevalonate kinase deficiency), juvenile idiopathic arthritis, PAPA syndrome (pyogenic sterile arthritis, pyoderma gangrenosum, and acne), PFAPA syndrome (periodic fever, aphthous stomatitis, pharyngitis, adenitis), recurrent idiopathic pericarditis, SAPHO (synovitis, acne, pustulosis, hyperostosis, osteomyelitis), Schnitzler syndrome, TRAPS (tumor necrosis factor receptor–associated periodic syndrome)
Neoplasms	
Hematologic malignancies	Amyloidosis, angioimmunoblastic lymphoma, Castleman's disease, Hodgkin's disease, hypereosinophilic syndrome, leukemia, lymphomatoid granulomatosis, malignant histiocytosis, multiple myeloma, myelodysplastic syndrome, myelofibrosis, non-Hodgkin's lymphoma, plasmacytoma, systemic mastocytosis, vaso-occlusive crisis in sickle cell disease
Solid tumors	Most solid tumors and metastases can cause fever. Those most commonly causing FUO are breast, colon, hepatocellular, lung, pancreatic, and renal cell carcinomas.
Benign tumors	Angiomyolipoma, cavernous hemangioma of the liver, craniopharyngioma, necrosis of dermoid tumor in Gardner's syndrome
Miscellaneous Causes	
	ADEM (acute disseminated encephalomyelitis), adrenal insufficiency, aneurysms, anomalous thoracic duct, aortic dissection, aortic-enteral fistula, aseptic meningitis (Mollaret's syndrome), atrial myxoma, brewer's yeast ingestion, Caroli disease, cholesterol emboli, cirrhosis, complex partial status epilepticus, cyclic neutropenia,

	drug fever, Erdheim-Chester disease, extrinsic allergic alveolitis, Fabry's disease, factitious disease, fire-eater's lung, fraudulent fever, Gaucher disease, Hamman-Rich syndrome (acute interstitial pneumonia), Hashimoto's encephalopathy, hematoma, hypersensitivity pneumonitis, hypertriglyceridemia, hypothalamic hypopituitarism, idiopathic normal-pressure hydrocephalus, inflammatory pseudotumor, Kikuchi's disease, linear IgA dermatosis, mesenteric fibromatosis, metal fume fever, milk protein allergy, myotonic dystrophy, nonbacterial osteitis, organic dust toxic syndrome, panniculitis, POEMS (polyneuropathy, organomegaly, endocrinopathy, monoclonal protein, skin changes), polymer fume fever, post–cardiac injury syndrome, primary biliary cirrhosis, primary hyperparathyroidism, pulmonary embolism, pyoderma gangrenosum, retroperitoneal fibrosis, Rosai-Dorfman disease, sclerosing mesenteritis, silicone embolization, subacute thyroiditis (de Quervain's), Sweet syndrome (acute febrile neutrophilic dermatosis), thrombosis, tubulointerstitial nephritis and uveitis syndrome (TINU), ulcerative colitis
Thermoregulatory Disorders	
Central	Brain tumor, cerebrovascular accident, encephalitis, hypothalamic dysfunction
Peripheral	Anhidrotic ectodermal dysplasia, exercise-induced hyperthermia, hyperthyroidism, pheochromocytoma

[a]This table includes all causes of FUO that have been described in the literature. [b]CAPS includes chronic infantile neurologic cutaneous and articular syndrome (CINCA, also known as neonatal-onset multisystem inflammatory disease, or NOMID), familial cold autoinflammatory syndrome (FCAS), and Muckle-Wells syndrome.

Of the NIIDs, large-vessel vasculitis, polymyalgia rheumatica, sarcoidosis, familial Mediterranean fever, and adult-onset Still's disease are rather common diagnoses in patients with FUO. The hereditary autoinflammatory syndromes are very rare and usually present in young patients. Schnitzler syndrome, which can present at any age, is uncommon but can often be diagnosed easily in a patient with FUO who presents with urticaria, bone pain, and monoclonal gammopathy.

Although most tumors can present with fever, malignant lymphoma is by far the most common diagnosis of FUO among the neoplasms. Sometimes the fever even precedes lymphadenopathy detectable by physical examination.

Apart from drug-induced fever and exercise-induced hyperthermia, none of the miscellaneous causes of fever is found very frequently in patients with FUO. Virtually all drugs can cause fever, even that commencing after long-term use. *Drug-induced fever*, including DRESS (drug reaction with eosinophilia and systemic symptoms; **Fig. A1-48**), is often accompanied by eosinophilia and also by lymphadenopathy, which can be extensive. More common causes of drug-induced fever are allopurinol, carbamazepine, lamotrigine,

phenytoin, sulfasalazine, furosemide, antimicrobial drugs (especially sulfonamides, minocycline, vancomycin, β-lactam antibiotics, and isoniazid), some cardiovascular drugs (e.g., quinidine), and some antiretroviral drugs (e.g., nevirapine). *Exercise-induced hyperthermia* (Chaps. 15 and 455) is characterized by an elevated body temperature that is associated with moderate to strenuous exercise lasting from half an hour up to several hours without an increase in CRP level or ESR; typically these patients sweat during the temperature elevation. *Factitious fever* (fever artificially induced by the patient—for example, by IV injection of contaminated water) should be considered in all patients but is more common among young women in health care professions. In *fraudulent fever*, the patient is normothermic but manipulates the thermometer. Simultaneous measurements at different body sites (rectum, ear, mouth) should rapidly identify this diagnosis. Another clue to fraudulent fever is a dissociation between pulse rate and temperature.

Previous studies of FUO have shown that a diagnosis is more likely in elderly patients than in younger age groups. In many cases, FUO in the elderly results from an atypical manifestation of a common disease, among which giant cell arteritis and polymyalgia rheumatica are most frequently involved. Tuberculosis is the most common infectious disease associated with FUO in elderly patients, occurring much more often than in younger patients. As many of these diseases are treatable, it is well worth pursuing the cause of fever in elderly patients.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Fever of Unknown Origin

FIRST-STAGE DIAGNOSTIC TESTS

Figure 17-1 shows a structured approach to patients presenting with FUO. The most important step in the diagnostic workup is the search for potentially diagnostic clues (PDCs) through complete and repeated history-taking and physical examination and the obligatory investigations listed above and in the figure. *PDCs* are defined as all localizing signs, symptoms, and abnormalities potentially pointing toward a diagnosis. Although PDCs are often misleading, only with their help can a concise list of probable diagnoses be made. The history should include information about the fever pattern (continuous or recurrent) and duration, previous medical history, present and recent drug use, family history, sexual history, country of origin, recent and remote travel, unusual environmental exposures associated with travel or hobbies, and animal contacts. A complete physical examination should be performed, with special attention to the eyes, lymph nodes, temporal arteries, liver, spleen, sites of previous surgery, entire skin surface, and mucous membranes. Before further diagnostic tests are initiated, antibiotic and glucocorticoid treatment, which can mask many diseases, should be stopped. For example, blood and other cultures are not reliable when samples are obtained during antibiotic treatment, and the size of enlarged lymph nodes usually decreases during glucocorticoid treatment, regardless of the cause of lymphadenopathy. Despite the high percentage of false-positive ultrasounds and the relatively low sensitivity of chest x-rays, the performance of these simple, low-cost diagnostic tests remains obligatory in all patients with FUO in order to separate cases that

are caused by easily diagnosed diseases from those that are not. Abdominal ultrasound is preferred to abdominal CT as an obligatory test because of relatively low cost, lack of radiation burden, and absence of side effects.

Only rarely do biochemical tests (beyond the obligatory tests needed to classify a patient's fever as FUO) lead directly to a definitive diagnosis in the absence of PDCs. The diagnostic yield of immunologic serology other than that included in the obligatory tests is relatively low. These tests more often yield false-positive rather than true-positive results and are of little use without PDCs pointing to specific immunologic disorders. Given the absence of specific symptoms in many patients and the relatively low cost of the test, investigation of cryoglobulins appears to be a valuable screening test in patients with FUO.

Multiple blood samples should be cultured in the laboratory long enough to ensure ample growth time for any fastidious organisms, such as HACEK organisms. It is critical to inform the laboratory of the intent to test for unusual organisms. Specialized media should be used when the history suggests uncommon microorganisms, such as *Histoplasma* or *Legionella*. Performing more than three blood cultures or more than one urine culture is useless in patients with FUO in the absence of PDCs (e.g., a high level of clinical suspicion of endocarditis). Repeating blood or urine cultures is useful only when previously cultured samples were collected during antibiotic treatment or within 1 week after its discontinuation. FUO with headache should prompt microbiologic examination of cerebrospinal fluid (CSF) for organisms including herpes simplex virus (especially type 2), *Cryptococcus neoformans*, and *Mycobacterium tuberculosis*. In central nervous system tuberculosis, the CSF typically has elevated protein and lowered glucose concentrations, with a mononuclear pleocytosis. CSF protein levels range from 100 to 500 mg/dL in most patients, the CSF glucose concentration is <45 mg/dL in 80% of cases, and the usual CSF cell count is between 100 and 500 cells/μL.

Microbiologic serology should not be included in the diagnostic workup of patients without PDCs for specific infections. A TST is included in the obligatory investigations, but it may yield false-negative results in patients with miliary tuberculosis, malnutrition, or immunosuppression. Although the IGRA is less influenced by prior vaccination with bacille Calmette-Guérin or by infection with nontuberculous mycobacteria, its sensitivity is similar to that of the TST; a negative TST or IGRA therefore does not exclude a diagnosis of tuberculosis. Miliary tuberculosis is especially difficult to diagnose. Granulomatous disease in liver or bone marrow biopsy samples, for example, should always lead to a (re)consideration of this diagnosis. If miliary tuberculosis is suspected, liver biopsy for acid-fast smear, culture, and polymerase chain reaction probably still has the highest diagnostic yield; however, biopsies of bone marrow, lymph nodes, or other involved organs also can be considered.

The diagnostic yield of echocardiography, sinus radiography, radiologic or endoscopic evaluation of the gastrointestinal tract, and bronchoscopy is very low in the absence of PDCs. Therefore, these tests should not be used as screening procedures.

After identification of all PDCs retrieved from the history, physical examination, and obligatory tests, a limited list of the most probable diagnoses should be made. Since most investigations are helpful only for patients who have PDCs for the diagnoses sought, further diagnostic procedures should be limited to specific

investigations aimed at confirming or excluding diseases on this list. In FUO, the diagnostic pointers are numerous and diverse but may be missed on initial examination, often being detected only by a very careful examination performed subsequently. In the absence of PDCs, the history and physical examination should therefore be repeated regularly. One of the first steps should be to rule out factitious or fraudulent fever, particularly in patients without signs of inflammation in laboratory tests. All medications, including nonprescription drugs and nutritional supplements, should be discontinued early in the evaluation to exclude drug fever. If fever persists beyond 72 h after discontinuation of the suspected drug, it is unlikely that this drug is the cause. In patients without PDCs or with only misleading PDCs, funduscopy by an ophthalmologist may be useful in the early stage of the diagnostic workup. When the first-stage diagnostic tests do not lead to a diagnosis, scintigraphy should be performed, especially when the ESR or the CRP level is elevated.

Recurrent Fever

In patients with recurrent fever, the diagnostic workup should consist of thorough history-taking, physical examination, and obligatory tests. The search for PDCs should be directed toward clues matching known recurrent syndromes **(Table 17-3)**. Patients should be asked to return during a febrile episode so that the history, physical examination, and laboratory tests can be repeated during a symptomatic phase. Further diagnostic tests, such as scintigraphic imaging (see below), should be performed only during a febrile episode because abnormalities may be absent between episodes. In patients with recurrent fever lasting >2 years, it is very unlikely that the fever is caused by infection or malignancy. Further diagnostic tests in that direction should be considered only when PDCs for infections, vasculitis syndromes, or malignancy are present or when the patient's clinical condition is deteriorating.

Scintigraphy

Scintigraphic imaging is a noninvasive method allowing delineation of foci in all parts of the body on the basis of functional changes in tissues. This procedure plays an important role in the diagnosis of patients with FUO in clinical practice. Conventional scintigraphic methods used in clinical practice are ^{67}Ga-citrate scintigraphy and ^{111}In- or ^{99m}Tc-labeled leukocyte scintigraphy. Focal infectious and inflammatory processes can also be detected by several radiologic techniques, such as CT, MRI, and ultrasound. However, because of the lack of substantial pathologic changes in the early phase, infectious and inflammatory foci cannot be detected at this time. Furthermore, distinguishing active infectious or inflammatory lesions from residual changes due to cured processes or surgery remains critical. Finally, CT and MRI routinely provide information on only part of the body, while scintigraphy readily allows whole-body imaging.

Fluorodeoxyglucose Positron Emission Tomography

^{18}F-Fluorodeoxyglucose (FDG) positron emission tomography (PET) combined with CT has become an established imaging procedure in FUO. FDG accumulates in tissues with a high rate of glycolysis, which occurs not only in malignant cells but also in activated leukocytes and thus permits the imaging of acute and chronic inflammatory processes. Normal uptake may obscure pathologic foci in the brain, heart, bowel, kidneys, and bladder. FDG uptake in the heart, which obscures endocarditis, may be prevented by consumption of a low-carbohydrate diet before the PET investigation. In patients with fever, bone marrow

uptake is frequently increased in a nonspecific way due to cytokine activation, which upregulates glucose transporters in bone marrow cells. Compared with conventional scintigraphy, FDG-PET/CT offers the advantages of higher resolution, greater sensitivity in chronic low-grade infections, and a high degree of accuracy in the central skeleton. Furthermore, vascular uptake of FDG is increased in patients with vasculitis (Fig. 17-2). The mechanisms responsible for FDG uptake do not allow differentiation among infection, sterile inflammation, and malignancy. However, since all of these disorders are causes of FUO, FDG-PET/CT can be used to guide additional diagnostic tests (e.g., targeted biopsies) that may yield the final diagnosis.

In recent years, many cohort studies and several meta-analyses have focused on the diagnostic yield of PET and PET/CT in FUO. Although these studies are highly variable in terms of the selection of patients, follow-up, and the selection of a gold-standard reference point, all meta-analyses report a high diagnostic yield for PET and PET/CT in the workup of FUO patients, with pooled sensitivity and specificity figures of ~85% and ~50%, respectively, and a total diagnostic yield of ~50% for PET/CT and ~40% for PET. In one study, FDG-PET was never helpful in diagnosing FUO in patients who had a normal CRP level and a normal ESR. In a meta-analysis on the performance, diagnostic yield, and management decision impact of nuclear imaging tests in patients with FUO, the diagnostic yield of gallium scintigraphy ranged from 21% to 54%, and, on average, the location of a source of fever was correctly localized in approximately one-third of patients. Moreover, in gallium scintigraphy, results do not become available for days, whereas FDG-PET/CT yields results within hours. In this meta-analysis, estimates of the diagnostic yield of labeled leukocyte scintigraphy ranged from 8% to 31%, and overall the cause of FUO was correctly identified on the basis of the scan results in only one-fifth of patients. Indirect comparisons of test performance suggested that FDG-PET/CT outperformed stand-alone FDG-PET, gallium scintigraphy, and leukocyte scintigraphy. Similarly, indirect comparisons of diagnostic yield suggested that FDG-PET/CT was more likely than alternative tests to correctly identify the cause of FUO.

Although scintigraphic techniques do not directly provide a definitive diagnosis, they often identify the anatomic location of a particular ongoing metabolic process and, with the help of other techniques such as biopsy and culture, facilitate timely diagnosis and treatment. Pathologic FDG uptake is quickly eradicated by treatment with glucocorticoids in many diseases, including vasculitis and lymphoma; therefore, glucocorticoid use should be stopped or postponed until after FDG-PET/CT is performed. Results reported in the literature and the advantages offered by FDG-PET/CT indicate that conventional scintigraphic techniques should be replaced by FDG-PET/CT in the investigation of patients with FUO at institutions where this technique is available. FDG-PET/CT is a relatively expensive procedure whose availability is still limited compared with that of CT and conventional scintigraphy. Nevertheless, FDG-PET/CT can be cost-effective in the FUO diagnostic workup if used at an early stage, helping to establish an early diagnosis, reducing days of hospitalization for diagnostic purposes, and obviating unnecessary and unhelpful tests.

LATER-STAGE DIAGNOSTIC TESTS

In some cases, more invasive tests are appropriate. Abnormalities found with scintigraphic techniques often need to be confirmed by pathology and/or culture of biopsy specimens. If lymphadenopathy is found, lymph node biopsy is necessary, even when the affected lymph nodes are hard to reach or when previous biopsies were inconclusive. In the case of skin lesions, skin biopsy should be undertaken. In one study, pulmonary

wedge excision, histologic examination of an excised tonsil, and biopsy of the peritoneum were performed in light of PDCs or abnormal FDG-PET results and yielded a diagnosis.

If no diagnosis is reached despite scintigraphic and PDC-driven histologic investigations or culture, second-stage screening diagnostic tests should be considered (Fig. 17-1). In three studies, the diagnostic yield of screening chest and abdominal CT in patients with FUO was ~20%. The specificity of chest CT was ~80%, but that of abdominal CT varied between 63% and 80%. Despite the relatively limited specificity of abdominal CT and the probably limited additional value of chest CT after normal FDG-PET/CT, chest and abdominal CT may be used as screening procedures at a later stage of the diagnostic protocol because of their noninvasive nature and high sensitivity. Bone marrow aspiration is seldom useful in the absence of PDCs for bone marrow disorders. With addition of FDG-PET/CT, which is highly sensitive in detecting lymphoma, carcinoma, and osteomyelitis, the value of bone marrow biopsy as a screening procedure is probably further reduced. Several studies have shown a high prevalence of giant cell arteritis among patients with FUO, with rates up to 17% among elderly patients. Giant cell arteritis often involves large arteries and in most cases can be diagnosed by FDG-PET/CT. However, temporal artery biopsy is still recommended for patients ≥55 years of age in a later stage of the diagnostic protocol: FDG-PET/CT will not be useful in vasculitis limited to the temporal arteries because of the small diameter of these vessels and the high levels of FDG uptake in the brain. In the past, liver biopsies have often been performed as a screening procedure in patients with FUO. In each of two recent studies, liver biopsy as part of the later stage of a screening diagnostic protocol was helpful in only one patient. Moreover, abnormal liver tests are not predictive of a diagnostic liver biopsy in FUO. Liver biopsy is an invasive procedure that carries the possibility of complications and even death. Therefore, it should not be used for screening purposes in patients with FUO except in those with PDCs for liver disease or miliary tuberculosis.

In patients with unexplained fever after all of the above procedures, the last steps in the diagnostic workup—with only a marginal diagnostic yield—come at an extraordinarily high cost in terms of both expense and discomfort for the patient. Repetition of a thorough history-taking and physical examination and review of laboratory results and imaging studies (including those from other hospitals) are recommended. Diagnostic delay often results from a failure to recognize PDCs in the available information. In these patients with persisting FUO, waiting for new PDCs to appear probably is better than ordering more screening investigations. Only when a patient's condition deteriorates without providing new PDCs should a further diagnostic workup be performed.

TREATMENT

TREATMENT

Fever of Unknown Origin

Empirical therapeutic trials with antibiotics, glucocorticoids, or antituberculous agents should be avoided in FUO except when a patient's condition is rapidly deteriorating after the aforementioned diagnostic tests have failed to provide a definite diagnosis.

ANTIBIOTICS AND ANTITUBERCULOUS THERAPY

Antibiotic or antituberculous therapy may irrevocably diminish the ability to culture fastidious bacteria or mycobacteria. However, hemodynamic instability or neutropenia is a good indication for empirical antibiotic therapy. If the TST or IGRA is positive or if granulomatous disease is present with anergy and sarcoidosis seems unlikely, a trial of therapy for tuberculosis should be started. Especially in miliary tuberculosis, it may be very difficult to obtain a rapid diagnosis. If the fever does not respond after 6 weeks of empirical antituberculous treatment, another diagnosis should be considered.

COLCHICINE, NONSTEROIDAL ANTI-INFLAMMATORY DRUGS, AND GLUCOCORTICOIDS

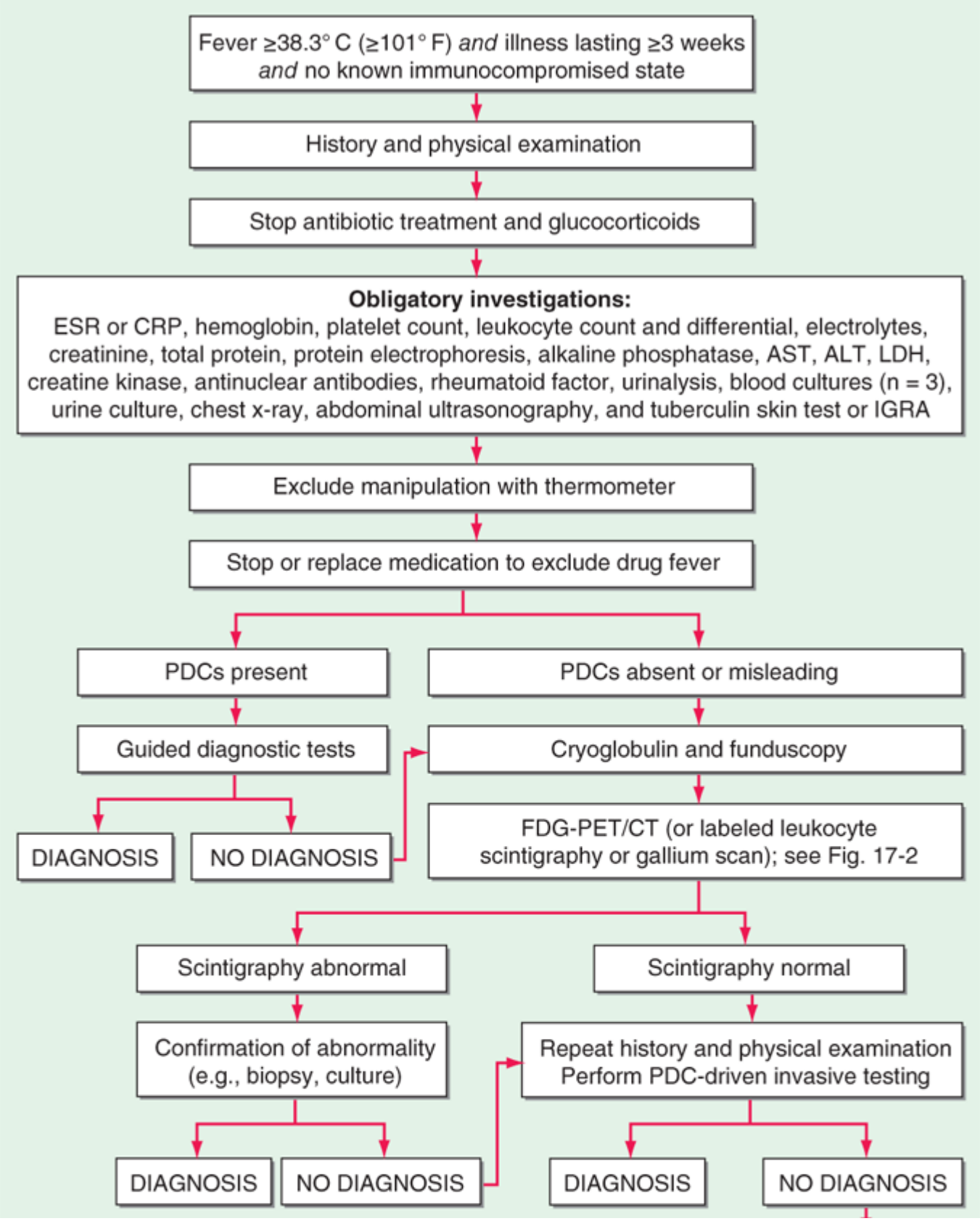
Colchicine is highly effective in preventing attacks of familial Mediterranean fever but is not always effective once an attack is well under way. When familial Mediterranean fever is suspected, the response to colchicine is not a completely reliable diagnostic tool in the acute phase, but with colchicine treatment most patients show remarkable improvements in the frequency and severity of subsequent febrile episodes within weeks to months. Therefore, colchicine may be tried in patients with features compatible with familial Mediterranean fever, especially when these patients originate from a high-prevalence region. If the fever persists and the source remains elusive after completion of the later-stage investigations, supportive treatment with nonsteroidal anti-inflammatory drugs (NSAIDs) can be helpful. The response of adult-onset Still's disease to NSAIDs is dramatic in some cases. The effects of glucocorticoids on giant cell arteritis and polymyalgia rheumatica are equally impressive. Early empirical trials with glucocorticoids, however, decrease the chances of reaching a diagnosis for which more specific and sometimes life-saving treatment might be more appropriate, such as malignant lymphoma. The ability of NSAIDs and glucocorticoids to mask fever while permitting the spread of infection or lymphoma dictates that their use should be avoided unless infectious diseases and malignant lymphoma have been largely ruled out and inflammatory disease is probable and is likely to be debilitating or threatening.

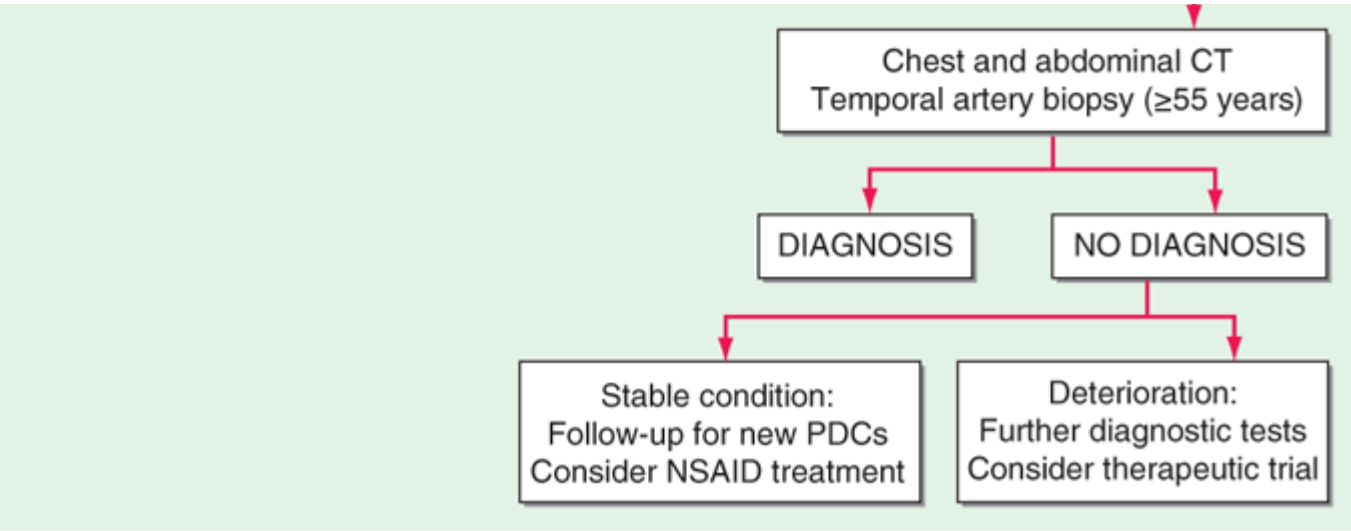
ANAKINRA

Interleukin (IL) 1 is a key cytokine in local and systemic inflammation and the febrile response. The availability of specific IL-1-targeting agents has revealed a pathologic role of IL-1-mediated inflammation in a growing list of diseases. Anakinra, a recombinant form of the naturally occurring IL-1 receptor antagonist (IL-1Ra), blocks the activity of both IL-1α and IL-1β. Anakinra is extremely effective in the treatment of many autoinflammatory syndromes, such as familial Mediterranean fever, cryopyrin-associated periodic syndrome, tumor necrosis factor receptor–associated periodic syndrome, mevalonate kinase deficiency (hyper IgD syndrome), and Schnitzler syndrome. There are many other chronic inflammatory disorders in which anti-IL-1 therapy is highly effective. A therapeutic trial with anakinra can be considered in patients whose FUO has not been diagnosed after later-stage diagnostic tests. Although most chronic inflammatory conditions without a known basis can be controlled with glucocorticoids, monotherapy with IL-1 blockade can provide improved control without the metabolic, immunologic, and gastrointestinal side effects of glucocorticoid administration.

FIGURE 17-1

Structured approach to patients with FUO. ALT, alanine aminotransferase; AST, aspartate aminotransferase; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; FDG-PET/CT, ^{18}F-fluorodeoxyglucose positron emission tomography combined with low-dose CT; IGRA, interferon γ release assay; LDH, lactate dehydrogenase; NSAID, nonsteroidal anti-inflammatory drug; PDCs, potentially diagnostic clues (all localizing signs, symptoms, and abnormalities potentially pointing toward a diagnosis).

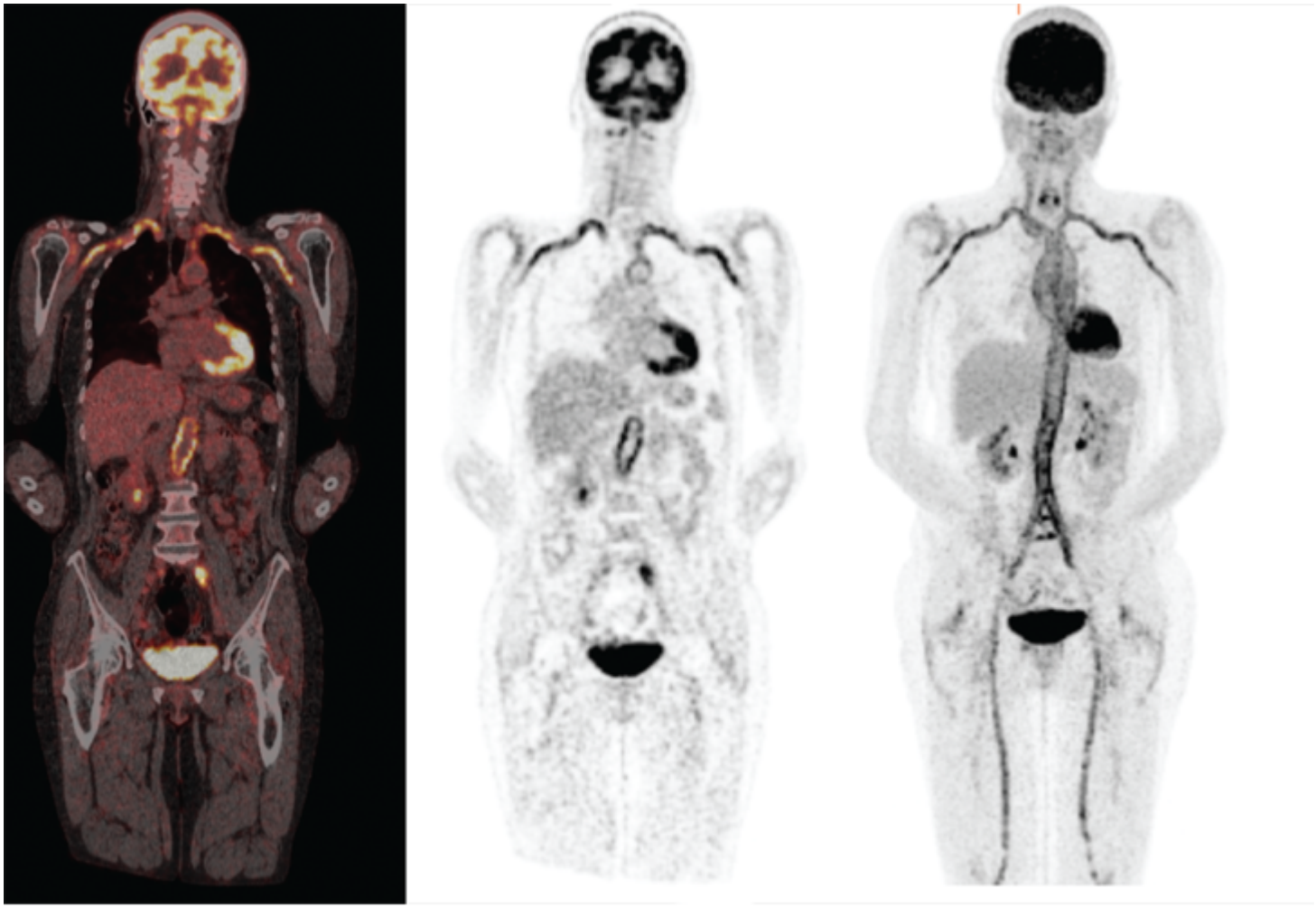




Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 17-2

FDG-PET/CT in a patient with FUO. This 72-year-old woman presented with a low-grade fever and severe fatigue of almost 3 months' duration. An extensive history was taken, but the patient had no specific complaints and had not traveled recently. Her previous history was unremarkable, and she did not use any drugs. Physical examination, including palpation of the temporal arteries, yielded completely normal results. Laboratory examination showed normocytic anemia, a C-reactive protein level of 43 mg/L, an erythrocyte sedimentation rate of 87 mm/h, and mild hypoalbuminemia. Results of the other obligatory tests were all normal. Since there were no potentially diagnostic clues, FDG-PET/CT was performed. This test showed increased FDG uptake in all major arteries (carotid, jugular, and subclavian arteries; thoracic and abdominal aorta; iliac, femoral, and popliteal arteries) and in the soft tissue around the shoulders, hips, and knees—findings compatible with large-vessel vasculitis and polymyalgia rheumatica. Within 1 week after the initiation of treatment with prednisone (60 mg once daily), the patient completely recovered. After 1 month, the prednisone dose was slowly tapered.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 17-3

All Reported Causes of Recurrent Fever[a]

Infections	
Bacterial, nonspecific	Apical granuloma, diverticulitis, prostatitis, recurrent bacteremia caused by colonic neoplasia or persistent focal infection, recurrent cellulitis, recurrent cholangitis or cholecystitis, recurrent pneumonia, recurrent sinusitis, recurrent urinary tract infection
Bacterial, specific	Bartonellosis, brucellosis, chronic gonococcemia, chronic meningococcemia, louse-borne relapsing fever (*Borrelia recurrentis*), melioidosis (*Pseudomonas pseudomallei*), Q fever (*Coxiella burnetii),* salmonellosis, *Spirillum minor* infection, *Streptobacillus moniliformis* infection, syphilis, tick-borne relapsing fever (*Borrelia duttonii*), tularemia, Whipple's disease (*Tropheryma whipplei*), yersiniosis
Fungal	Coccidioidomycosis, histoplasmosis, paracoccidioidomycosis
Parasitic	Babesiosis, malaria, toxoplasmosis, trypanosomiasis, visceral leishmaniasis
Viral	Cytomegalovirus infection, Epstein-Barr virus infection, herpes simplex
Noninfectious Inflammatory Diseases	
Systemic rheumatic and autoimmune diseases	Ankylosing spondylitis, antiphospholipid syndrome, autoimmune hemolytic anemia, autoimmune hepatitis, Behçet's disease, cryoglobulinemia, gout, polymyositis, pseudogout, reactive arthritis, relapsing polychondritis, systemic lupus erythematosus
Vasculitis	Churg-Strauss syndrome, giant cell vasculitis/polymyalgia rheumatica, hypersensitivity vasculitis, polyarteritis nodosa, urticarial vasculitis
Granulomatous diseases	Idiopathic granulomatous hepatitis, sarcoidosis

Infections	
Autoinflammatory syndromes	Adult-onset Still's disease, Blau syndrome, CANDLE (chronic atypical neutrophilic dermatitis with lipodystrophy and elevated temperature syndrome), CAPS[b] (cryopyrin-associated periodic syndrome), CRMO (chronic recurrent multifocal osteomyelitis), Crohn's disease, DIRA (deficiency of the interleukin 1 receptor antagonist), familial Mediterranean fever, hemophagocytic syndrome, hyper-IgD syndrome (HIDS, also known as mevalonate kinase deficiency), juvenile idiopathic arthritis, NLRC4-activating mutations, PAPA syndrome (pyogenic sterile arthritis, pyoderma gangrenosum, and acne), PFAPA syndrome (periodic fever, aphthous stomatitis, pharyngitis, adenitis), recurrent idiopathic pericarditis, SAPHO (synovitis, acne, pustulosis, hyperostosis, osteomyelitis), SAVI (stimulator of interferon genes [STING]–associated vasculopathy with onset in infancy), Schnitzler syndrome, TRAPS (tumor necrosis factor receptor–associated periodic syndrome)
Neoplasms	
	Angioimmunoblastic lymphoma, Castleman's disease, colon carcinoma, craniopharyngioma, Hodgkin's disease, malignant histiocytosis, mesothelioma, non-Hodgkin's lymphoma
Miscellaneous Causes	
	Adrenal insufficiency, aortic-enteral fistula, aseptic meningitis (Mollaret's syndrome), atrial myxoma, brewer's yeast ingestion, cholesterol emboli, cyclic neutropenia, drug fever, extrinsic allergic alveolitis, Fabry's disease, factitious disease, fraudulent fever, Gaucher disease, hypersensitivity pneumonitis, hypertriglyceridemia, hypothalamic hypopituitarism, inflammatory pseudotumor, metal fume fever, milk protein allergy, polymer fume fever, pulmonary embolism, sclerosing mesenteritis
Thermoregulatory Disorders	
Central	Hypothalamic dysfunction
Peripheral	Anhidrotic ectodermal dysplasia, exercise-induced hyperthermia, pheochromocytoma

[a]This table includes all causes of recurrent fever that have been described in the literature. [b]CAPS includes chronic infantile neurologic cutaneous and articular syndrome (CINCA, also known as neonatal-onset multisystem inflammatory disease, or NOMID), familial cold autoinflammatory syndrome (FCAS), and Muckle-Wells syndrome.

PROGNOSIS

FUO-related mortality rates have continuously declined over recent decades. The majority of fevers are caused by treatable diseases, and the risk of death related to FUO is, of course, dependent on the underlying disease. In a study by our group (Table 17-1), none of 37 FUO patients without a diagnosis died during a follow-up period of at least 6 months; 4 of 36 patients with a diagnosis died during follow-up as a result of infection ($n = 1$) or malignancy ($n = 3$). A large study on the prognosis of FUO (Vanderschueren et al, 2014; Table 17-1) included 436 patients and documented a mortality rate of 10%, of which 68% was related to the febrile illness—malignancy in most cases. In this study, only 4 of 168 patients in whom no diagnosis could be made died, all during their first admission. In two of these patients, diagnosis (lymphoma and pneumonia) was made during autopsy. Other studies have also shown that malignancy accounts for most FUO-related deaths. Non-Hodgkin's lymphoma carries a disproportionately high death toll. In nonmalignant FUO, fatality rates are very low. The good outcome in patients without a diagnosis confirms that potentially lethal occult diseases are very unusual and that empirical therapy with antibiotics, antituberculous agents, or glucocorticoids is rarely required in stable patients. In less affluent regions, infectious diseases are still a major cause of FUO, and outcomes may be different.

FURTHER READING

Bleeker-rovers cp et al: A prospective multicenter study on fever of unknown origin: The yield of a structured diagnostic protocol. Medicine (Baltimore) 86:26, 2007. [PubMed: 17220753]

Knockaert dc et al: Fever of unknown origin in adults: 40 years on. J Intern Med 253:263, 2003. [PubMed: 12603493]

Mulders-manders c et al: Fever of unknown origin. Clin Med 15:280, 2015.

Takeuchi m et al: Nuclear imaging for classical fever of unknown origin: Meta-analysis. J Nucl Med 57:1913, 2016. [PubMed: 27339873]

Vanderschueren s et al: Mortality in patients presenting with fever of unknown origin. Acta Clin Belg 69:12, 2014. [PubMed: 24635393]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 18: Syncope

Roy Freeman

INTRODUCTION

Syncope is a transient, self-limited loss of consciousness due to acute global impairment of cerebral blood flow. The onset is rapid, duration brief, and recovery spontaneous and complete. Other causes of transient loss of consciousness need to be distinguished from syncope; these include seizures, vertebrobasilar ischemia, hypoxemia, and hypoglycemia. A syncopal prodrome (*presyncope*) is common, although loss of consciousness may occur without any warning symptoms. Typical presyncopal symptoms include dizziness, lightheadedness or faintness, weakness, fatigue, and visual and auditory disturbances. The causes of syncope can be divided into three general categories: (1) neurally mediated syncope (also called *reflex or vasovagal syncope*), (2) orthostatic hypotension, and (3) cardiac syncope.

Neurally mediated syncope comprises a heterogeneous group of functional disorders that are characterized by a transient change in the reflexes responsible for maintaining cardiovascular homeostasis. Episodic vasodilation (or loss of vasoconstrictor tone) and bradycardia occur in varying combinations, resulting in temporary failure of blood pressure control. In contrast, in patients with orthostatic hypotension due to autonomic failure, these cardiovascular homeostatic reflexes are chronically impaired. Cardiac syncope may be due to arrhythmias or structural cardiac diseases that cause a decrease in cardiac output. The clinical features, underlying pathophysiologic mechanisms, therapeutic interventions, and prognoses differ markedly among these three causes.

EPIDEMIOLOGY AND NATURAL HISTORY

Syncope is a common presenting problem, accounting for ~3% of all emergency room visits and 1% of all hospital admissions. The annual cost for syncope-related hospitalization in the United States is ~$2.4 billion. Syncope has a lifetime cumulative incidence of up to 35% in the general population. The peak incidence in the young occurs between ages 10 and 30 years, with a median peak around 15 years. Neurally mediated syncope is the etiology in the vast majority of these cases. In elderly adults, there is a sharp rise in the incidence of syncope after 70 years.

In population-based studies, neurally mediated syncope is the most common cause of syncope. The incidence is slightly higher in females than males. In young subjects, there is often a family history in first-degree relatives. Cardiovascular disease due to structural disease or arrhythmias is the next most common cause in most series, particularly in emergency room settings and in older patients. Orthostatic hypotension also increases in prevalence with age because of the reduced baroreflex responsiveness, decreased cardiac compliance, and attenuation of the vestibulosympathetic reflex associated with aging. In the elderly, orthostatic hypotension is substantially more common in institutionalized (54–68%) than community-dwelling (6%) individuals, an observation most likely explained by the greater prevalence of predisposing neurologic disorders, physiologic impairment, and vasoactive medication use among institutionalized patients.

The prognosis after a single syncopal event for all age groups is generally benign. In particular, syncope of noncardiac and unexplained origin in younger individuals has an excellent prognosis; life expectancy is unaffected. By contrast, syncope due to a cardiac cause, either structural heart disease or primary arrhythmic disease, is associated with an increased risk of sudden cardiac death and mortality from other causes. Similarly, mortality rate is increased in individuals with syncope due to orthostatic hypotension related to age and the associated comorbid conditions (Table 18-1).

TABLE 18-1

High-Risk Features Indicating Hospitalization or Intensive Evaluation of Syncope

Chest pain suggesting coronary ischemia
Features of congestive heart failure
Moderate or severe valvular disease
Moderate or severe structural cardiac disease
Electrocardiographic features of ischemia
History of ventricular arrhythmias
Prolonged QT interval (>500 ms)
Repetitive sinoatrial block or sinus pauses
Persistent sinus bradycardia
Bi- or trifascicular block or intraventricular conduction delay with QRS duration ≥120 ms
Atrial fibrillation
Nonsustained ventricular tachycardia
Family history of sudden death
Preexcitation syndromes
Brugada pattern on ECG
Palpitations at time of syncope
Syncope at rest or during exercise

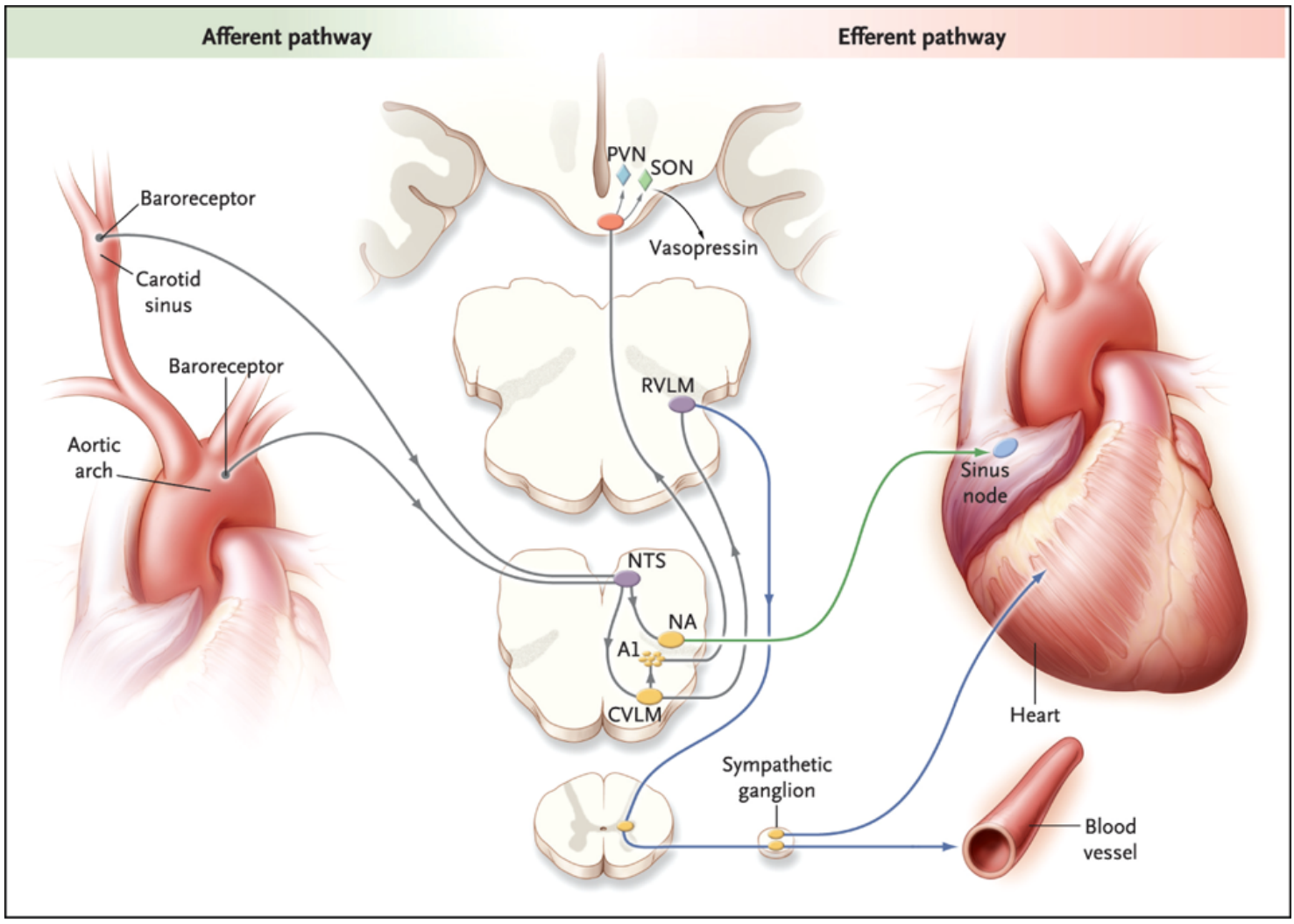
PATHOPHYSIOLOGY

The upright posture imposes a unique physiologic stress upon humans; most, although not all, syncopal episodes occur from a standing position. Standing results in pooling of 500–1000 mL of blood in the lower extremities and splanchnic circulation. There is a decrease in venous return to the heart and reduced ventricular filling that result in diminished cardiac output and blood pressure. These hemodynamic changes provoke a compensatory reflex response, initiated by the baroreceptors in the carotid sinus and aortic arch, resulting in increased sympathetic outflow and decreased vagal nerve activity **(Fig. 18-1)**. The reflex increases peripheral resistance, venous return to the heart, and cardiac output and thus limits the fall in blood pressure. If this response fails, as is the case chronically in orthostatic hypotension and transiently in neurally mediated syncope, cerebral hypoperfusion occurs.

FIGURE 18-1

The baroreflex. A decrease in arterial pressure unloads the baroreceptors—the terminals of afferent fibers of the glossopharyngeal and vagus nerves—that are situated in the carotid sinus and aortic arch. This leads to a reduction in the afferent impulses that are relayed from these mechanoreceptors through the glossopharyngeal and vagus nerves to the nucleus of the tractus solitarius (NTS) in the dorsomedial medulla. The reduced baroreceptor afferent activity produces a decrease in vagal nerve input to the sinus node that is mediated via connections of the NTS to the nucleus ambiguus (NA). There is an increase in sympathetic efferent activity that is mediated by the NTS projections to the caudal

ventrolateral medulla (CVLM) (an excitatory pathway) and from there to the rostral ventrolateral medulla (RVLM) (an inhibitory pathway). The activation of RVLM presympathetic neurons in response to hypotension is thus predominantly due to disinhibition. In response to a sustained fall in blood pressure, vasopressin release is mediated by projections from the A1 noradrenergic cell group in the ventrolateral medulla. This projection activates vasopressin-synthesizing neurons in the magnocellular portion of the paraventricular nucleus (PVN) and the supraoptic nucleus (SON) of the hypothalamus. *Blue* denotes sympathetic neurons, and *green* denotes parasympathetic neurons. *(From R Freeman: N Engl J Med 358:615, 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Syncope is a consequence of global cerebral hypoperfusion and thus represents a failure of cerebral blood flow autoregulatory mechanisms. Myogenic factors, local metabolites, and to a lesser extent autonomic neurovascular control are responsible for the autoregulation of cerebral blood flow (Chap. 301). The latency of the autoregulatory response is 5–10 s. Typically cerebral blood flow ranges from 50 to 60 mL/min per 100 g brain tissue and remains relatively constant over perfusion pressures ranging from 50 to 150 mmHg. Cessation of blood flow for 6–8 s will result in loss of consciousness, while impairment of consciousness ensues when blood flow decreases to 25 mL/min per 100 g brain tissue.

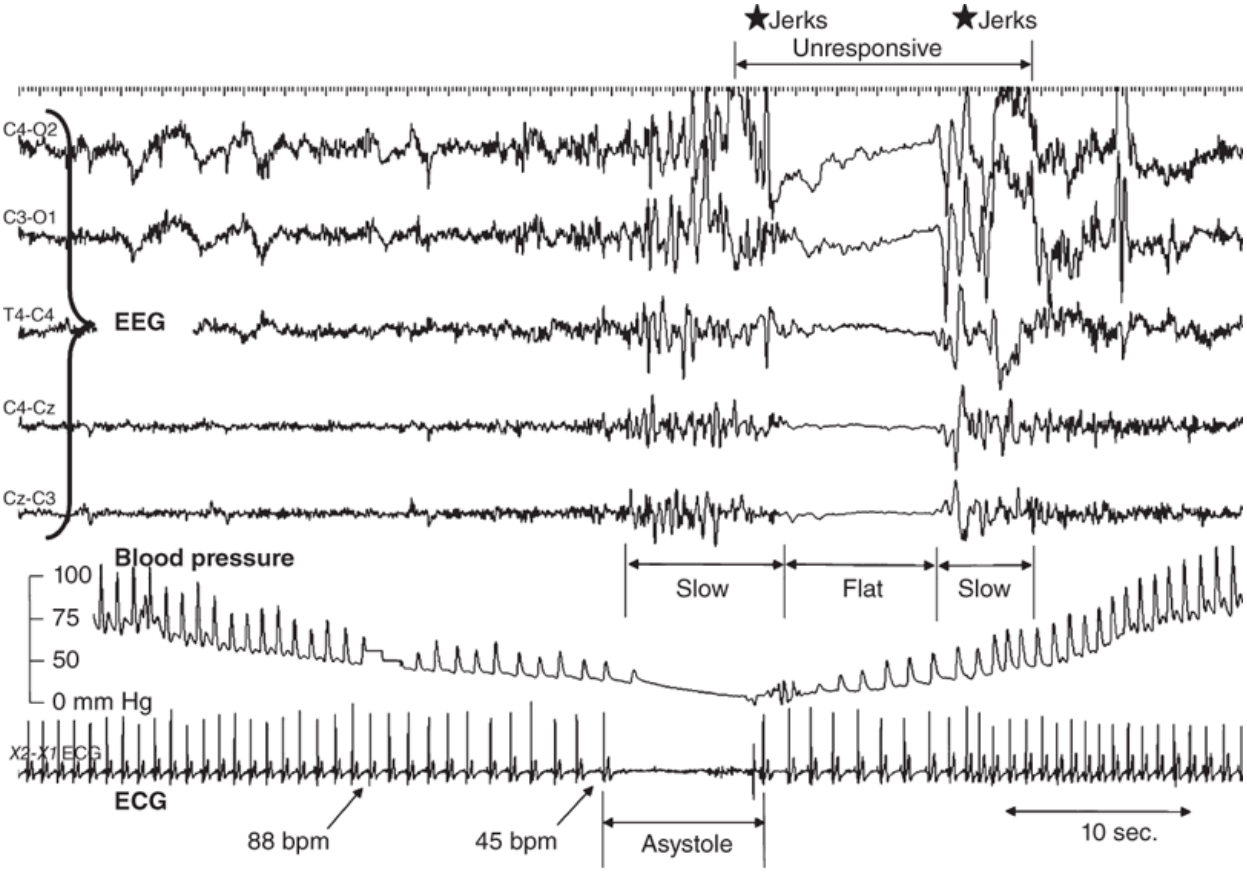
From the clinical standpoint, a fall in systemic systolic blood pressure to ~50 mmHg or lower will result in syncope. A decrease in cardiac output and/or systemic vascular resistance—the determinants of blood pressure—thus underlies the pathophysiology of syncope. Common causes of impaired cardiac output include decreased effective circulating blood volume; increased thoracic pressure; massive pulmonary embolus; cardiac brady- and tachyarrhythmias; valvular heart disease; and myocardial dysfunction. Systemic vascular resistance may be decreased by central and peripheral autonomic nervous system diseases, sympatholytic medications, and transiently during neurally mediated syncope. Increased cerebral vascular resistance, most frequently due to hypocarbia induced by hyperventilation, may also contribute to the pathophysiology of syncope.

Two patterns of electroencephalographic (EEG) changes occur in syncopal subjects. The first is a “slow-flat-slow” pattern (Fig. 18-2) in which normal background activity is replaced with high-amplitude slow delta waves. This is followed by sudden flattening of the EEG—a cessation or

attenuation of cortical activity—followed by the return of slow waves, and then normal activity. A second pattern, the “slow pattern,” is characterized by increasing and decreasing slow wave activity only. The EEG flattening that occurs in the slow-flat-slow pattern is a marker of more severe cerebral hypoperfusion. Despite the presence of myoclonic movements and other motor activity during some syncopal events, EEG seizure discharges are not detected.

FIGURE 18-2

The electroencephalogram (EEG) in vasovagal syncope. A 1-min segment of a tilt-table test with typical vasovagal syncope demonstrating the “slow-flat-slow” EEG pattern. Finger beat-to-beat blood pressure, electrocardiogram (ECG), and selected EEG channels are shown. EEG slowing starts when systolic blood pressure drops to ~50 mmHg; heart rate is then ~45 beats/min (bpm). Asystole occurred, lasting about 8 s. The EEG flattens for a similar period, but with a delay. A transient loss of consciousness, lasting 14 s, was observed. There were muscle jerks just before and just after the flat period of the EEG. *(Figure reproduced with permission from W Wieling et al: Brain 132:2630, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

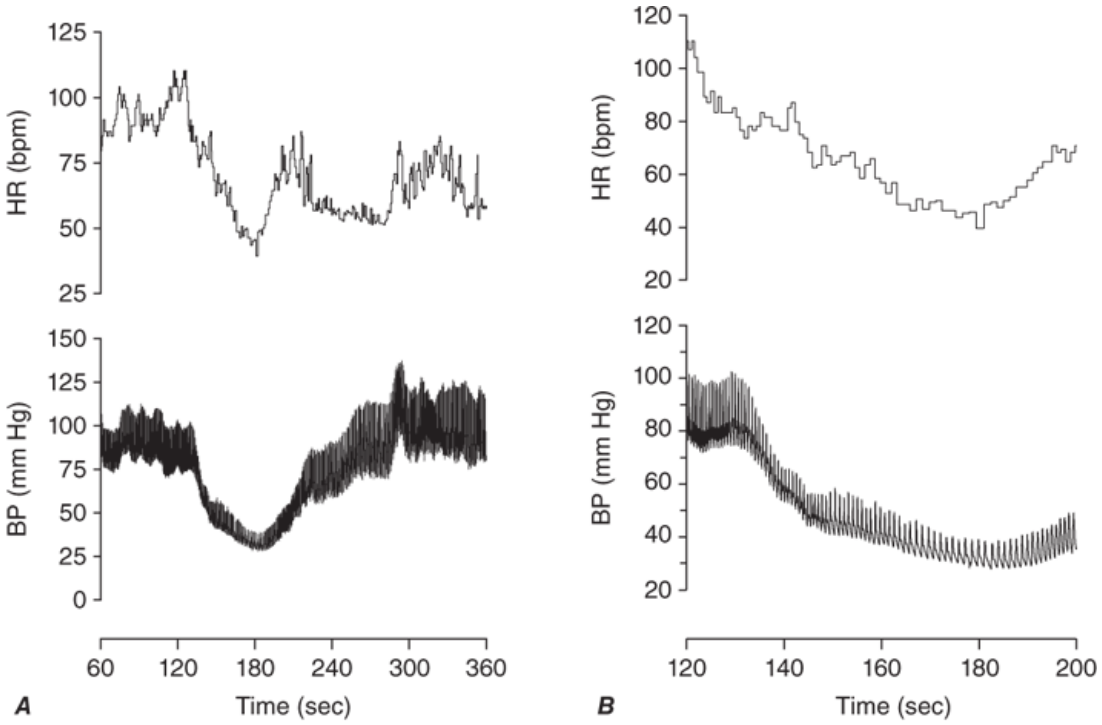
CLASSIFICATION

NEURALLY MEDIATED SYNCOPE

Neurally mediated (reflex; vasovagal) syncope is the final pathway of a complex central and peripheral nervous system reflex arc. There is a sudden, transient change in autonomic efferent activity with increased parasympathetic outflow, plus sympathoinhibition (the vasodepressor response), resulting in bradycardia, vasodilation, and/or reduced vasoconstrictor tone. The resulting fall in systemic blood pressure can then reduce cerebral blood flow to below the compensatory limits of autoregulation (Fig. 18-3). In order to elicit neurally mediated syncope, a functioning autonomic nervous system is necessary, in contrast to syncope resulting from autonomic failure (discussed below).

FIGURE 18-3

A. The paroxysmal hypotensive-bradycardic response that is characteristic of neurally mediated syncope. Noninvasive beat-to-beat blood pressure and heart rate are shown >5 min (from 60 to 360 s) of an upright tilt on a tilt table. ***B.*** The same tracing expanded to show 80 s of the episode (from 80 to 200 s). BP, blood pressure; bpm, beats per minute; HR, heart rate.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Multiple triggers of the afferent limb of the reflex arc can result in neurally mediated syncope. In some situations, these can be clearly defined, e.g., the carotid sinus, the gastrointestinal tract, or the bladder. Often, however, the trigger is less easily recognized and the cause is multifactorial. Under these circumstances, it is likely that different afferent pathways converge on the central autonomic network within the medulla that integrates the neural impulses and mediates the vasodepressor-bradycardic response.

Classification of Neurally Mediated Syncope

Neurally mediated syncope may be subdivided based on the afferent pathway and provocative trigger. Vasovagal syncope (the common faint) is provoked by intense emotion, pain, and/or orthostatic stress, whereas the situational reflex syncopes have specific localized stimuli that provoke the reflex vasodilation and bradycardia that leads to syncope. The underlying mechanisms have been identified and pathophysiology delineated for most of these situational reflex syncopes. The afferent trigger may originate in the pulmonary system, gastrointestinal system, urogenital system, heart, and carotid artery (Table 18-2). Hyperventilation leading to hypocarbia and cerebral vasoconstriction, and raised intrathoracic pressure that impairs venous return to the heart, play a central role in many of the situational reflex syncopes. The afferent pathway of the reflex arc differs among these disorders, but the efferent response via the vagus and sympathetic pathways is similar.

TABLE 18-2

Causes of Syncope

A. Neurally Mediated Syncope
Vasovagal syncope
Provoked fear, pain, anxiety, intense emotion, sight of blood, unpleasant sights and odors, orthostatic stress
Situational reflex syncope
Pulmonary
Cough syncope, wind instrument player's syncope, weightlifter's syncope, "mess trick"[a] and "fainting lark,"[b] sneeze syncope, airway instrumentation
Urogenital
Postmicturition syncope, urogenital tract instrumentation, prostatic massage
Gastrointestinal
Swallow syncope, glossopharyngeal neuralgia, esophageal stimulation, gastrointestinal tract instrumentation, rectal examination, defecation syncope
Cardiac
Bezold-Jarisch reflex, cardiac outflow obstruction
Carotid sinus
Carotid sinus sensitivity, carotid sinus massage
Ocular
Ocular pressure, ocular examination, ocular surgery
B. Orthostatic Hypotension
Primary autonomic failure due to idiopathic central and peripheral neurodegenerative diseases—the "synucleinopathies"
Lewy body diseases
Parkinson's disease
Lewy body dementia
Pure autonomic failure
Multiple system atrophy (Shy-Drager syndrome)
Secondary autonomic failure due to autonomic peripheral neuropathies
Diabetes

Hereditary amyloidosis (familial amyloid polyneuropathy)
Primary amyloidosis (AL amyloidosis; immunoglobulin light chain associated)
Hereditary sensory and autonomic neuropathies (HSAN) (especially type III—familial dysautonomia)
Idiopathic immune-mediated autonomic neuropathy
Autoimmune autonomic ganglionopathy
Sjögren's syndrome
Paraneoplastic autonomic neuropathy
HIV neuropathy
Postprandial hypotension
Iatrogenic (drug-induced)
Volume depletion
C. Cardiac Syncope
Arrhythmias
Sinus node dysfunction
Atrioventricular dysfunction
Supraventricular tachycardias
Ventricular tachycardias
Inherited channelopathies
Cardiac structural disease
Valvular disease
Myocardial ischemia
Obstructive and other cardiomyopathies
Atrial myxoma
Pericardial effusions and tamponade

[a]Hyperventilation for ~1 min, followed by sudden chest compression. [b]Hyperventilation (~20 breaths) in a squatting position, rapid rise to standing, then Valsalva.

Alternately, neurally mediated syncope may be subdivided based on the predominant efferent pathway. Vasodepressor syncope describes syncope predominantly due to efferent, sympathetic, vasoconstrictor failure; cardioinhibitory syncope describes syncope predominantly associated with bradycardia or asystole due to increased vagal outflow; and mixed syncope describes syncope in which there are both vagal and sympathetic reflex changes.

Features of Neurally Mediated Syncope

In addition to symptoms of orthostatic intolerance such as dizziness, lightheadedness, and fatigue, premonitory features of autonomic activation may be present in patients with neurally mediated syncope. These include diaphoresis, pallor, palpitations, nausea, hyperventilation, and yawning. During the syncopal event, proximal and distal myoclonus (typically arrhythmic and multifocal) may occur, raising the possibility of epilepsy. The eyes typically remain open and usually deviate upward. Pupils are usually dilated. Roving eye movements may occur. Grunting, moaning, snorting, and stertorous breathing may be present. Urinary incontinence may occur. Fecal incontinence is very rare. Postictal confusion is also rare, although visual and auditory hallucinations and near death and out-of-body experiences are sometimes reported.

Although some predisposing factors and provocative stimuli are well established (for example, motionless upright posture, warm ambient temperature, intravascular volume depletion, alcohol ingestion, hypoxemia, anemia, pain, the sight of blood, venipuncture, and intense emotion), the underlying basis for the widely different thresholds for syncope among individuals exposed to the same provocative stimulus is not known. A genetic basis for neurally mediated syncope may exist; several studies have reported an increased incidence of syncope in first-degree relatives of fainters, but no gene or genetic marker has been identified, and environmental, social, and cultural factors have not been excluded by these studies.

TREATMENT

TREATMENT

Neurally Mediated Syncope

Reassurance, avoidance of provocative stimuli, and plasma volume expansion with fluid and salt are the cornerstones of the management of neurally mediated syncope. Isometric counterpressure maneuvers of the limbs (leg crossing or handgrip and arm tensing) may raise blood pressure by increasing central blood volume and cardiac output. By maintaining pressure in the autoregulatory zone, these maneuvers avoid or delay the onset of syncope. Randomized controlled trials support this intervention.

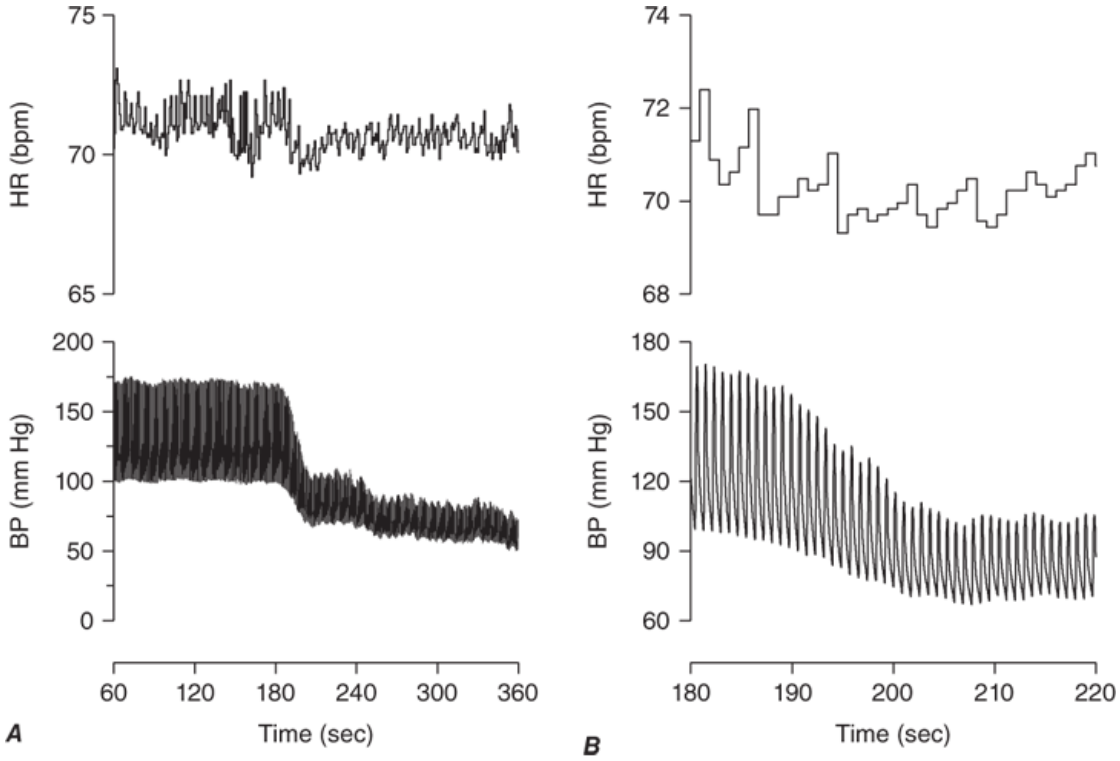
Fludrocortisone, vasoconstricting agents, and β-adrenoreceptor antagonists are widely used by experts to treat refractory patients, although there is no consistent evidence from randomized controlled trials for any pharmacotherapy to treat neurally mediated syncope. Because vasodilation is the dominant pathophysiologic syncopal mechanism in most patients, use of a cardiac pacemaker is rarely beneficial. Possible exceptions are older patients (>40 years) in whom syncope is associated with asystole or severe bradycardia and patients with prominent cardioinhibition due to carotid sinus syndrome. In these patients, dual-chamber pacing may be helpful although this continues to be an area of uncertainty.

ORTHOSTATIC HYPOTENSION

Orthostatic hypotension, defined as a reduction in systolic blood pressure of at least 20 mmHg or diastolic blood pressure of at least 10 mmHg within 3 min of standing or head-up tilt on a tilt table, is a manifestation of sympathetic vasoconstrictor (autonomic) failure **(Fig. 18-4)**. In many (but not all) cases, there is no compensatory increase in heart rate despite hypotension; with partial autonomic failure, heart rate may increase to some degree but is insufficient to maintain cardiac output. A variant of orthostatic hypotension is “delayed” orthostatic hypotension, which occurs beyond 3 min of standing; this may reflect a mild or early form of sympathetic adrenergic dysfunction. In some cases, orthostatic hypotension occurs within 15 s of standing (so-called “initial” orthostatic hypotension), a finding that may reflect a transient mismatch between cardiac output and peripheral vascular resistance and does not represent autonomic failure.

FIGURE 18-4

A. The gradual fall in blood pressure without a compensatory heart rate increase that is characteristic of orthostatic hypotension due to autonomic failure. Blood pressure and heart rate are shown >5 min (from 60 to 360 s) of an upright tilt on a tilt table. ***B.*** The same tracing expanded to show 40 s of the episode (from 180 to 220 s). BP, blood pressure; bpm, beats per minute; HR, heart rate.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Characteristic symptoms of orthostatic hypotension include light-headedness, dizziness, and presyncope (near-faintness) occurring in response to sudden postural change. However, symptoms may be absent or nonspecific, such as generalized weakness, fatigue, cognitive slowing, leg buckling, or headache. Visual blurring may occur, likely due to retinal or occipital lobe ischemia. Neck pain, typically in the suboccipital, posterior cervical, and shoulder region (the “coat-hanger headache”), most likely due to neck muscle ischemia, may be the only symptom. Patients may report orthostatic dyspnea (thought to reflect ventilation-perfusion mismatch due to inadequate perfusion of ventilated lung apices) or angina (attributed to impaired myocardial perfusion even with normal coronary arteries). Symptoms may be exacerbated by exertion, prolonged standing, increased ambient temperature, or meals. Syncope is usually preceded by warning symptoms, but may occur suddenly, suggesting the possibility of a seizure or cardiac cause.

Supine hypertension is common in patients with orthostatic hypotension due to autonomic failure, affecting >50% of patients in some series. Orthostatic hypotension may present after initiation of therapy for hypertension, and supine hypertension may follow treatment of orthostatic hypotension. However, in other cases, the association of the two conditions is unrelated to therapy; it may in part be explained by baroreflex dysfunction in the presence of residual sympathetic outflow, particularly in patients with central autonomic degeneration.

Causes of Neurogenic Orthostatic Hypotension

Causes of neurogenic orthostatic hypotension include central and peripheral autonomic nervous system dysfunction **(Chap. 432)**. Autonomic dysfunction of other organ systems (including the bladder, bowels, sexual organs, and sudomotor system) of varying severity frequently accompanies orthostatic hypotension in these disorders (Table 18-2).

The primary autonomic degenerative disorders are multiple system atrophy (Shy-Drager syndrome; **Chap. 432**), Parkinson’s disease **(Chap. 427)**, dementia with Lewy bodies **(Chap. 426)**, and pure autonomic failure **(Chap. 432)**. These are often grouped together as “synucleinopathies” due to the presence of α-synuclein, a small protein that aggregates predominantly in the cytoplasm of neurons in the Lewy body disorders (Parkinson’s disease, dementia with Lewy bodies, and pure autonomic failure) and in the glia in multiple system atrophy.

Peripheral autonomic dysfunction may also accompany small-fiber peripheral neuropathies such as those seen in diabetes mellitus, amyloid, immune-mediated neuropathies, hereditary sensory and autonomic neuropathies (HSAN; particularly HSAN type III, familial dysautonomia) **(Chaps. 438 and 439)**. Less frequently, orthostatic hypotension is associated with the peripheral neuropathies that accompany vitamin B_{12} deficiency, neurotoxic exposure, HIV and other infections, and porphyria.

Patients with autonomic failure and the elderly are susceptible to falls in blood pressure associated with meals. The magnitude of the blood pressure fall is exacerbated by large meals, meals high in carbohydrate, and alcohol intake. The mechanism of postprandial syncope is not fully elucidated.

Orthostatic hypotension is often iatrogenic. Drugs from several classes may lower peripheral resistance (e.g., α-adrenoreceptor antagonists used to treat hypertension and prostatic hypertrophy; antihypertensive agents of several classes; nitrates and other vasodilators; tricyclic agents and phenothiazines). Iatrogenic volume depletion due to diuresis and volume depletion due to medical causes (hemorrhage, vomiting, diarrhea, or decreased fluid intake) may also result in decreased effective circulatory volume, orthostatic hypotension, and syncope.

TREATMENT

TREATMENT

Orthostatic Hypotension

The first step is to remove reversible causes—usually vasoactive medications **(Table 432-6)**. Next, nonpharmacologic interventions should be introduced. These interventions include patient education regarding staged moves from supine to upright; warnings about the hypotensive effects of large meals; instructions about the isometric counterpressure maneuvers that increase intravascular pressure (see above); and raising the head of the bed to reduce supine hypertension. Intravascular volume should be expanded by increasing dietary fluid and salt. If these nonpharmacologic measures fail, pharmacologic intervention with fludrocortisone acetate and vasoconstricting agents such as midodrine, L-dihydroxyphenylserine, and pseudoephedrine should be introduced. Some patients with intractable symptoms require additional therapy with supplementary agents that include pyridostigmine, atomoxetine, yohimbine, desmopressin acetate (DDAVP), and erythropoietin **(Chap. 432)**.

CARDIAC SYNCOPE

Cardiac (or cardiovascular) syncope is caused by arrhythmias and structural heart disease. These may occur in combination because structural disease renders the heart more vulnerable to abnormal electrical activity.

Arrhythmias

Bradyarrhythmias that cause syncope include those due to severe sinus node dysfunction (e.g., sinus arrest or sinoatrial block) and atrioventricular (AV) block (e.g., Mobitz type II, high-grade, and complete AV block). The bradyarrhythmias due to sinus node dysfunction are often associated with an atrial tachyarrhythmia, a disorder known as the tachycardia-bradycardia syndrome. A prolonged pause following the termination of a tachycardic episode is a frequent cause of syncope in patients with the tachycardia-bradycardia syndrome. Medications of several classes may also cause bradyarrhythmias of sufficient severity to cause syncope. Syncope due to bradycardia or asystole is referred to as a Stokes-Adams attack.

Ventricular tachyarrhythmias frequently cause syncope. The likelihood of syncope with ventricular tachycardia is in part dependent on the ventricular rate; rates <200 beats/min are less likely to cause syncope. The compromised hemodynamic function during ventricular tachycardia is caused by ineffective ventricular contraction, reduced diastolic filling due to abbreviated filling periods, loss of AV synchrony, and concurrent myocardial ischemia.

Several disorders associated with cardiac electrophysiologic instability and arrhythmogenesis are due to mutations in ion channel subunit genes. These include the long QT syndrome, Brugada syndrome, and catecholaminergic polymorphic ventricular tachycardia. The long QT syndrome is a genetically heterogeneous disorder associated with prolonged cardiac repolarization and a predisposition to ventricular arrhythmias. Syncope and sudden death in patients with long QT syndrome result from a unique polymorphic ventricular tachycardia called *torsades des pointes* that degenerates into ventricular fibrillation. The long QT syndrome has been linked to genes encoding K^+ channel α-subunits, K^+ channel β-subunits, voltage-gated Na^+ channel, and a scaffolding protein, ankyrin B (ANK2). Brugada syndrome is characterized by idiopathic ventricular fibrillation in association with right ventricular electrocardiogram (ECG) abnormalities without structural heart disease. This disorder is also genetically heterogeneous, although it is most frequently linked to mutations in the Na^+ channel α-subunit, SCN5A. Catecholaminergic polymorphic tachycardia is an inherited, genetically heterogeneous disorder associated with exercise- or stress-induced ventricular arrhythmias, syncope, or sudden death. Acquired QT interval prolongation, most commonly due to drugs, may also result in ventricular arrhythmias and syncope. **These disorders are discussed in detail in Chap. 249.**

Structural Disease

Structural heart disease (e.g., valvular disease, myocardial ischemia, hypertrophic and other cardiomyopathies, cardiac masses such as atrial myxoma, and pericardial effusions) may lead to syncope by compromising cardiac output. Structural disease may also contribute to other pathophysiologic mechanisms of syncope. For example, cardiac structural disease may predispose to arrhythmogenesis; aggressive treatment of cardiac failure with diuretics and/or vasodilators may lead to orthostatic hypotension; and inappropriate reflex vasodilation may occur with structural disorders such as aortic stenosis and hypertrophic cardiomyopathy, possibly provoked by increased ventricular contractility.

TREATMENT

TREATMENT

Cardiac Syncope

Treatment of cardiac disease depends on the underlying disorder. Therapies for arrhythmias include cardiac pacing for sinus node disease and AV block, and ablation, antiarrhythmic drugs, and cardioverter-defibrillators for atrial and ventricular tachyarrhythmias. These disorders are best managed by physicians with specialized skills in this area.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Syncope

DIFFERENTIAL DIAGNOSIS

Syncope is easily diagnosed when the characteristic features are present; however, several disorders with transient real or apparent loss of consciousness may create diagnostic confusion.

Generalized and partial seizures may be confused with syncope; however, there are a number of differentiating features. Whereas tonic-clonic movements are the hallmark of a generalized seizure, myoclonic and other movements also may occur in up to 90% of syncopal episodes. Myoclonic jerks associated with syncope may be multifocal or generalized. They are typically arrhythmic and of short duration (<30 s). Mild flexor and extensor posturing also may occur. Partial or partial-complex seizures with secondary generalization are usually preceded by an aura, commonly an unpleasant smell; fear; anxiety; abdominal discomfort; or other visceral sensations. These phenomena should be differentiated from the premonitory features of syncope.

Autonomic manifestations of seizures (autonomic epilepsy) may provide a more difficult diagnostic challenge. Autonomic seizures have cardiovascular, gastrointestinal, pulmonary, urogenital, pupillary, and cutaneous manifestations that are similar to the premonitory features of syncope. Furthermore, the cardiovascular manifestations of autonomic epilepsy include clinically significant tachycardias and bradycardias that may be of sufficient magnitude to cause loss of consciousness. The presence of accompanying nonautonomic auras may help differentiate these episodes from syncope.

Loss of consciousness associated with a seizure usually lasts >5 min and is associated with prolonged postictal drowsiness and disorientation, whereas reorientation occurs almost immediately after a syncopal event. Muscle aches may occur after both syncope and seizures, although they tend to last longer and be more severe following a seizure. Seizures, unlike syncope, are rarely provoked by emotions or pain. Incontinence of urine may occur with both seizures and syncope; however, fecal incontinence occurs very rarely with syncope.

Hypoglycemia may cause transient loss of consciousness, typically in individuals with type 1 or type 2 diabetes treated with insulin. The clinical features associated with impending or actual hypoglycemia include tremor, palpitations, anxiety, diaphoresis, hunger, and paresthesias. These symptoms are due to autonomic activation to counter the falling blood glucose. Hunger, in particular, is not a typical premonitory feature of syncope. Hypoglycemia also impairs neuronal function, leading to fatigue, weakness, dizziness, and cognitive and behavioral symptoms. Diagnostic difficulties may occur in individuals in strict glycemic control; repeated hypoglycemia impairs the counterregulatory response and leads to a loss of the characteristic warning symptoms that are the hallmark of hypoglycemia.

Patients with cataplexy experience an abrupt partial or complete loss of muscular tone triggered by strong emotions, typically anger or laughter. Unlike syncope, consciousness is maintained throughout the attacks, which typically last between 30 s and 2 min. There are no premonitory symptoms. Cataplexy occurs in 60–75% of patients with narcolepsy.

The clinical interview and interrogation of eyewitnesses usually allow differentiation of syncope from falls due to vestibular dysfunction, cerebellar disease, extrapyramidal system dysfunction, and other gait disorders. A diagnosis of syncope can be particularly challenging in

patients with dementia who experience repeated falls and are unable to provide a clear history of the episodes. If the fall is accompanied by head trauma, a postconcussive syndrome, amnesia for the precipitating events, and/or the presence of loss of consciousness may also contribute to diagnostic difficulty.

Apparent loss of consciousness can be a manifestation of psychiatric disorders such as generalized anxiety, panic disorders, major depression, and somatization disorder. These possibilities should be considered in individuals who faint frequently without prodromal symptoms. Such patients are rarely injured despite numerous falls. There are no clinically significant hemodynamic changes concurrent with these episodes. In contrast, transient loss of consciousness due to vasovagal syncope precipitated by fear, stress, anxiety, and emotional distress is accompanied by hypotension, bradycardia, or both.

INITIAL EVALUATION

The goals of the initial evaluation are to determine whether the transient loss of consciousness was due to syncope; to identify the cause; and to assess risk for future episodes and serious harm (Table 18-1). The initial evaluation should include a detailed history, thorough questioning of eyewitnesses, and a complete physical and neurologic examination. Blood pressure and heart rate should be measured in the supine position and after 3 min of standing to determine whether orthostatic hypotension is present. An ECG should be performed if there is suspicion of syncope due to an arrhythmia or underlying cardiac disease. Relevant electrocardiographic abnormalities include bradyarrhythmias or tachyarrhythmias, AV block, ischemia, old myocardial infarction, long QT syndrome, and bundle branch block. This initial assessment will lead to the identification of a cause of syncope in ~50% of patients and also allows stratification of patients at risk for cardiac mortality.

Laboratory Tests

Baseline laboratory blood tests are rarely helpful in identifying the cause of syncope. Blood tests should be performed when specific disorders, e.g., myocardial infarction, anemia, and secondary autonomic failure, are suspected (Table 18-2).

Autonomic Nervous System Testing (Chap. 432)

Autonomic testing, including tilt-table testing, can be performed in specialized centers. Autonomic testing is helpful to uncover objective evidence of autonomic failure and also to demonstrate a predisposition to neurally mediated syncope. Autonomic testing includes assessments of parasympathetic autonomic nervous system function (e.g., heart rate variability to deep respiration and a Valsalva maneuver), sympathetic cholinergic function (e.g., thermoregulatory sweat response and quantitative sudomotor axon reflex test), and sympathetic adrenergic function (e.g., blood pressure response to a Valsalva maneuver and a tilt-table test with beat-to-beat blood pressure measurement). The hemodynamic abnormalities demonstrated on the tilt-table test (Figs. 18-3 and 18-4) may be useful in distinguishing orthostatic hypotension due to autonomic failure from the hypotensive bradycardic response of neurally mediated syncope. Similarly, the tilt-table test may help identify patients with syncope due to immediate or delayed orthostatic hypotension.

Carotid sinus massage should be considered in patients with symptoms suggestive of carotid sinus syncope and in patients >50 years with recurrent syncope of unknown etiology. This test should only be carried out under continuous ECG and blood pressure monitoring and should be avoided in patients with carotid bruits, plaques, or stenosis.

Cardiac Evaluation

ECG monitoring is indicated for patients with a high pretest probability of arrhythmia causing syncope. Patients should be monitored in hospital if the likelihood of a life-threatening arrhythmia is high, e.g., patients with severe structural or coronary artery disease, nonsustained ventricular tachycardia, trifascicular heart block, prolonged QT interval, Brugada syndrome ECG pattern, or family history of sudden cardiac death (Table 18-1). Outpatient Holter monitoring is recommended for patients who experience frequent syncopal episodes (one or more per week), whereas loop recorders, which continually record and erase cardiac rhythm, are indicated for patients with suspected arrhythmias with low risk of sudden cardiac death. Loop recorders may be external (recommended for evaluation of episodes that occur at a frequency of >1 per month) or implantable (if syncope occurs less frequently).

Echocardiography should be performed in patients with a history of cardiac disease or if abnormalities are found on physical examination or the ECG. Echocardiographic diagnoses that may be responsible for syncope include aortic stenosis, hypertrophic cardiomyopathy, cardiac tumors, aortic dissection, and pericardial tamponade. Echocardiography also has a role in risk stratification based on the left ventricular ejection fraction.

Treadmill exercise testing with ECG and blood pressure monitoring should be performed in patients who have experienced syncope during or shortly after exercise. Treadmill testing may help identify exercise-induced arrhythmias (e.g., tachycardia-related AV block) and exercise-induced exaggerated vasodilation.

Electrophysiologic studies are indicated in patients with structural heart disease and ECG abnormalities in whom noninvasive investigations have failed to yield a diagnosis. Electrophysiologic studies have low sensitivity and specificity and should only be performed when a high pretest probability exists. Currently, this test is rarely performed to evaluate patients with syncope.

Psychiatric Evaluation

Screening for psychiatric disorders may be appropriate in patients with recurrent unexplained syncope episodes. Tilt-table testing, with demonstration of symptoms in the absence of hemodynamic change, may be useful in reproducing syncope in patients with suspected psychogenic syncope.

FURTHER READING

Al-Khatib SM et al: Risk stratification for arrhythmic events in patients with asymptomatic pre-excitation: A systematic review for the 2015 ACC/AHA/HRS guideline for the management of adult patients with supraventricular tachycardia: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. J Am Coll Cardiol 67:1624, 2016. [PubMed: 26409260]

Freeman R et al: Consensus statement on the definition of orthostatic hypotension, neurally mediated syncope and the postural tachycardia syndrome. AutonNeurosci 161:46, 2011.

Gibbons CH et al: The recommendations of a consensus panel for the screening, diagnosis, and treatment of neurogenic orthostatic hypotension and associated supine hypertension. J Neurol 264:1567, 2017. [PubMed: 28050656]

Sheldon RS, Raj SR: Pacing and vasovagal syncope: Back to our physiologic roots. Clin Auton Res 27:213, 2017. [PubMed: 28710604]

Varosy PD et al: Pacing as a treatment for reflex-mediated (vasovagal, situational, or carotid sinus hypersensitivity) syncope: A systematic review for the 2017 ACC/AHA/HRS guideline for the evaluation and management of patients with syncope: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. J Am Coll Cardiol 70:664, 2017. [PubMed: 28286220]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 19: Dizziness and Vertigo

Mark F. Walker; Robert B. Daroff

FIGURE 19-1

INTRODUCTION

Dizziness is an imprecise symptom used to describe a variety of common sensations that include vertigo, light-headedness, faintness, and imbalance. *Vertigo* refers to a sense of spinning or other motion that may be physiological, occurring during or after a sustained head rotation, or pathological, due to vestibular dysfunction. The term *light-headedness* is classically applied to presyncopal sensations resulting from brain hypoperfusion but as used by patients has little specificity, as it may also refer to other symptoms such as disequilibrium and imbalance. A challenge to diagnosis is that patients often have difficulty distinguishing among these various symptoms, and the words they choose do not reliably indicate the underlying etiology.

There are many causes of dizziness. Vestibular dizziness (vertigo or imbalance) may be due to peripheral disorders that affect the labyrinths or vestibular nerves, or it may result from disruption of central vestibular pathways. It may be paroxysmal or due to a fixed unilateral or bilateral vestibular deficit. Acute unilateral lesions cause vertigo due to a sudden imbalance in vestibular inputs from the two labyrinths. Bilateral lesions cause imbalance and instability of vision when the head moves (*oscillopsia*) due to loss of normal vestibular reflexes.

Presyncopal dizziness occurs when cardiac dysrhythmia, orthostatic hypotension, medication effects, or another cause leads to brain hypoperfusion. Such presyncopal sensations vary in duration; they may increase in severity until loss of consciousness occurs, or they may resolve before loss of consciousness if the cerebral ischemia is corrected. Faintness and syncope, which are discussed in detail in Chap. 18, should always be considered when one is evaluating patients with brief episodes of dizziness or dizziness that occurs with upright posture. Other causes of dizziness include non-vestibular imbalance and gait disorders (e.g., loss of proprioception from sensory neuropathy, parkinsonism), and anxiety.

When evaluating patients with dizziness, questions to consider include the following: (1) Is it dangerous (e.g., arrhythmia, transient ischemic attack/stroke)? (2) Is it vestibular? (3) If vestibular, is it peripheral or central? A careful history and examination often provide sufficient information to answer these questions and determine whether additional studies or referral to a specialist is necessary.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Dizziness

HISTORY

When a patient presents with dizziness, the first step is to delineate more precisely the nature of the symptom. In the case of vestibular disorders, the physical symptoms depend on whether the lesion is unilateral or bilateral, and whether it is acute or chronic. Vertigo, an illusion of self or environmental motion, implies asymmetry of vestibular inputs from the two labyrinths or in their central pathway that is usually acute. Symmetric bilateral vestibular hypofunction causes imbalance but no vertigo. Because of the ambiguity in patients' descriptions of their symptoms, diagnosis based simply on symptom characteristics is typically unreliable. Thus, the history should focus closely on other features, including whether this is the first attack, the duration of this and any prior episodes, provoking factors, and accompanying symptoms.

Dizziness can be divided into episodes that last for seconds, minutes, hours, or days. Common causes of brief dizziness (seconds) include benign paroxysmal positional vertigo (BPPV) and orthostatic hypotension, both of which typically are provoked by changes in head and body position. Attacks of vestibular migraine and Ménière's disease often last hours. When episodes are of intermediate duration (minutes), transient ischemic attacks of the posterior circulation should be considered, although migraine and a number of other causes are also possible.

Symptoms that accompany vertigo may be helpful in distinguishing peripheral vestibular lesions from central causes. Unilateral hearing loss and other aural symptoms (ear pain, pressure, fullness) typically point to a peripheral cause. Because the auditory pathways quickly become bilateral upon entering the brainstem, central lesions are unlikely to cause unilateral hearing loss, unless the lesion lies near the root entry zone of the auditory nerve. Symptoms such as double vision, numbness, and limb ataxia suggest a brainstem or cerebellar lesion.

EXAMINATION

Because dizziness and imbalance can be a manifestation of a variety of neurologic disorders, the neurologic examination is important in the evaluation of these patients. Particular focus should be given to assessment of eye movements, vestibular function, and hearing. The range of eye movements and whether they are equal in each eye should be observed. Peripheral eye movement disorders (e.g., cranial neuropathies, eye muscle weakness) are usually disconjugate (different in the two eyes). One should check pursuit (the ability to follow a smoothly moving target) and saccades (the ability to look back and forth accurately between two targets). Poor pursuit or inaccurate (dysmetric) saccades usually indicate central pathology, often involving the cerebellum. Alignment of the two eyes can be checked with a cover test: while the patient is looking at a target, alternately cover the eyes and observe for corrective saccades. A vertical misalignment may indicate a brainstem or cerebellar lesion. Finally, one should look for spontaneous nystagmus, an involuntary back-and-forth movement of the eyes. Nystagmus is most often of the jerk type, in which a slow drift (slow phase) in one direction alternates with a rapid saccadic movement (quick phase or fast phase) in the opposite direction that resets the position of the eyes in the orbits. Except in the case of acute vestibulopathy (e.g., vestibular neuritis), if primary position nystagmus is easily seen in the light, it is probably due to a central cause. Two forms of nystagmus that are characteristic of lesions of the cerebellar pathways are vertical nystagmus with downward fast phases (downbeat nystagmus) and horizontal nystagmus that changes direction with gaze (gaze-evoked nystagmus). By contrast, peripheral lesions typically cause unidirectional horizontal nystagmus. Use of Frenzel eyeglasses (self-illuminated goggles with convex lenses that blur the patient's vision but allow the examiner to see the eyes greatly magnified) can aid in the detection of peripheral vestibular nystagmus, because they reduce the patient's ability to use visual fixation to suppress nystagmus. **Table 19-1** outlines key findings that help distinguish peripheral from central causes of vertigo.

The most useful bedside test of peripheral vestibular function is the head impulse test, in which the vestibuloocular reflex (VOR) is assessed with small-amplitude (~20 degrees) rapid head rotations. While the patient fixates on a target, the head is rotated to the left or right. If the VOR is deficient, the rotation is followed by a catch-up saccade in the opposite direction (e.g., a leftward saccade after a rightward rotation). The head impulse test can identify both unilateral (catch-up saccades after rotations toward the weak side) and bilateral vestibular hypofunction (catch-up saccades after rotations in both directions).

All patients with episodic dizziness, especially if provoked by positional change, should be tested with the Dix-Hallpike maneuver. The patient begins in a sitting position with the head turned 45 degrees; holding the back of the head, the examiner then lowers the patient into a supine position with the head extended backward by about 20 degrees while watching the eyes. Posterior canal BPPV can be diagnosed confidently if transient upbeating-torsional nystagmus is seen. If no nystagmus is observed after 15–20 s, the patient is raised to the sitting position, and the procedure is repeated with the head turned to the other side. Again, Frenzel goggles may improve the sensitivity of the test.

Dynamic visual acuity is a functional test that can be useful in assessing vestibular function. Visual acuity is measured with the head still and when the head is rotated back and forth by the examiner (about 1–2 Hz). A drop in visual acuity during head motion of more than one line on a near card or Snellen chart is abnormal and indicates vestibular dysfunction.

ANCILLARY TESTING

The choice of ancillary tests should be guided by the history and examination findings. Audiometry should be performed whenever a vestibular disorder is suspected. Unilateral sensorineural hearing loss supports a peripheral disorder (e.g., vestibular schwannoma). Predominantly low-frequency hearing loss is characteristic of Ménière's disease. Electronystagmography or videonystagmography includes recordings of spontaneous nystagmus (if present) and measurement of positional nystagmus. Caloric testing assesses the responses of the two horizontal semicircular canals. The test battery often includes recording of saccades and pursuit to assess central ocular motor function. Neuroimaging is important if a central vestibular disorder is suspected. In addition, patients with unexplained unilateral hearing loss or vestibular hypofunction should undergo magnetic resonance imaging (MRI) of the internal auditory canals, including administration of gadolinium, to rule out a schwannoma.

TABLE 19-1

Features of Peripheral and Central Vertigo

- Nystagmus from an acute peripheral lesion is unidirectional, with fast phases beating away from the ear with the lesion. Nystagmus that changes direction with gaze is due to a central lesion.
- Transient mixed vertical-torsional nystagmus occurs in benign paroxysmal positional vertigo (BPPV), but pure vertical or pure torsional nystagmus is a central sign.
- Nystagmus from a peripheral lesion may be inhibited by visual fixation, whereas central nystagmus is not suppressed.
- Absence of a head impulse sign in a patient with acute prolonged vertigo should suggest a central cause.
- Unilateral hearing loss suggests peripheral vertigo. Findings such as diplopia, dysarthria, and limb ataxia suggest a central disorder.

DIFFERENTIAL DIAGNOSIS AND TREATMENT

Treatment of vestibular symptoms should be driven by the underlying diagnosis. Simply treating dizziness with vestibular suppressant medications is often not helpful and may make the symptoms worse and prolong recovery. The diagnostic and specific treatment approaches for the most commonly encountered vestibular disorders are discussed below.

ACUTE PROLONGED VERTIGO (VESTIBULAR NEURITIS)

An acute unilateral vestibular lesion causes constant vertigo, nausea, vomiting, oscillopsia (motion of the visual scene), and imbalance. These symptoms are due to a sudden asymmetry of inputs from the two labyrinths or in their central connections, simulating a continuous rotation of the head. Unlike BPPV, continuous vertigo persists even when the head remains still.

When a patient presents with an acute vestibular syndrome, the most important question is whether the lesion is central (e.g., a cerebellar or brainstem infarct or hemorrhage), which may be life-threatening, or peripheral, affecting the vestibular nerve or labyrinth (vestibular neuritis). Attention should be given to any symptoms or signs that point to central dysfunction (diplopia, weakness or numbness, dysarthria). The pattern of spontaneous nystagmus, if present, may be helpful (Table 19-1). If the head impulse test is normal, an acute peripheral vestibular lesion is unlikely. A central lesion cannot always be excluded with certainty based on symptoms and examination alone; thus, older patients with vascular risk factors who present with an acute vestibular syndrome should be evaluated for the possibility of stroke even when there are no specific findings that indicate a central lesion.

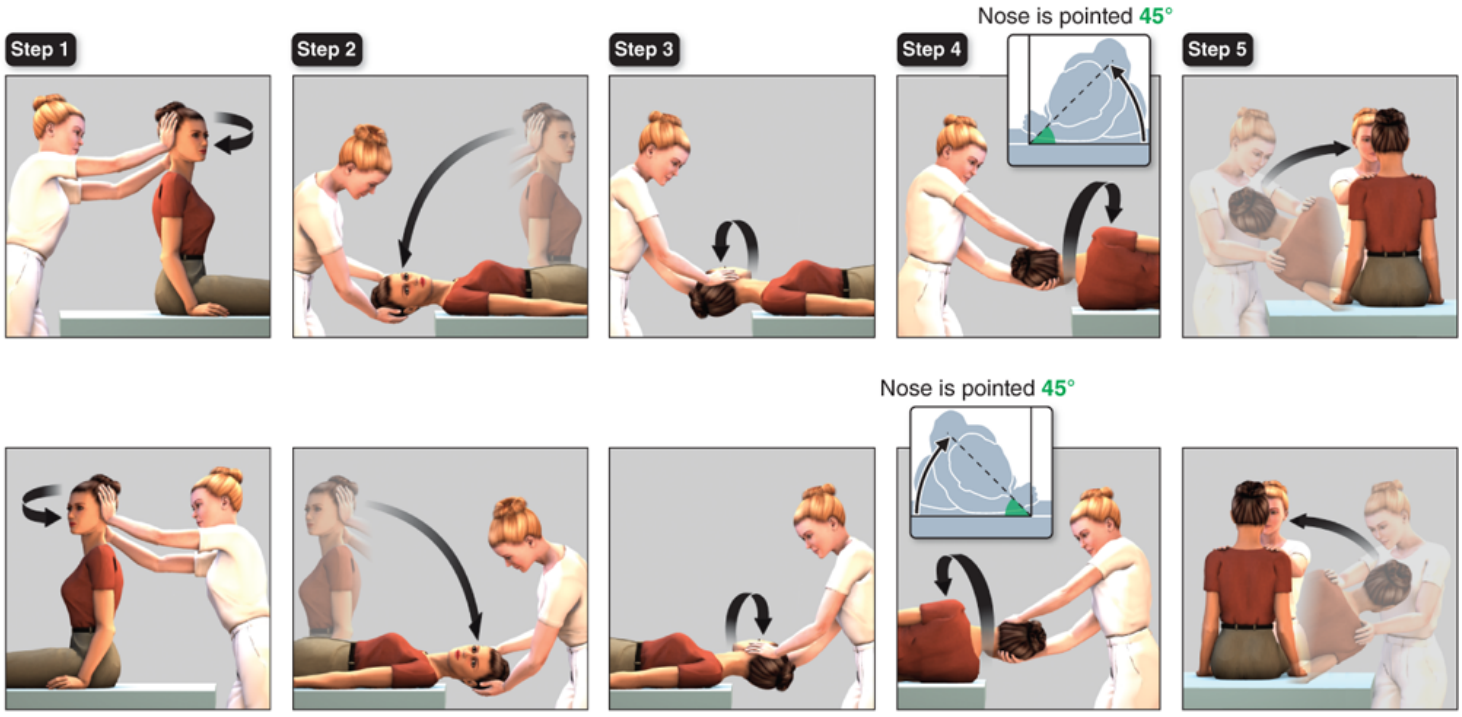
Most patients with vestibular neuritis recover spontaneously, but glucocorticoids can improve outcome if administered within 3 days of symptom onset. Antiviral medications are of no proven benefit and are not typically given unless there is evidence to suggest herpes zoster oticus (Ramsay Hunt syndrome). Vestibular suppressant medications may reduce acute symptoms but should be avoided after the first several days because they may impede central compensation and recovery. Patients should be encouraged to resume a normal level of activity as soon as possible, and directed vestibular rehabilitation therapy may accelerate improvement.

BENIGN PAROXYSMAL POSITIONAL VERTIGO

BPPV is a common cause of recurrent vertigo. Episodes are brief (<1 min and typically 15–20 s) and are always provoked by changes in head position relative to gravity, such as lying down, rolling over in bed, rising from a supine position, and extending the head to look upward. The attacks are caused by free-floating otoconia (calcium carbonate crystals) that have been dislodged from the utricular macula and have moved into one of the semicircular canals, usually the posterior canal. When head position changes, gravity causes the otoconia to move within the canal, producing vertigo and nystagmus. With posterior canal BPPV, the nystagmus beats upward and torsionally (the upper poles of the eyes beat toward the affected lower ear). Less commonly, the otoconia enter the horizontal canal, resulting in a horizontal nystagmus when the patient is lying with either ear down. Superior (also called anterior) canal involvement is rare. BPPV is treated with repositioning maneuvers that use gravity to remove the otoconia from the semicircular canal. For posterior canal BPPV, the Epley maneuver **(Fig. 19-1)** is the most commonly used procedure. For more refractory cases of BPPV, patients can be taught a variant of this maneuver that they can perform alone at home. A demonstration of the Epley maneuver is available online (*http://www.dizziness-and-balance.com/disorders/bppv/bppv.html*).

FIGURE 19-1

Modified Epley maneuver for treatment of benign paroxysmal positional vertigo of the right (*top panels*) and left (*bottom panels*) posterior semicircular canals. **Step 1.** With the patient seated, turn the head 45 degrees toward the affected ear. **Step 2.** Keeping the head turned, lower the patient to the head-hanging position and hold for at least 30 s and until nystagmus disappears. **Step 3.** Without lifting the head, turn it 90 degrees toward the other side. Hold for another 30 s. **Step 4.** Rotate the patient onto her side while turning the head another 90 degrees, so that the nose is pointed down 45 degrees. Hold again for 30 s. **Step 5.** Have the patient sit up on the side of the table. After a brief rest, the maneuver should be repeated to confirm successful treatment. *(Figure adapted from http://www.dizziness-and-balance.com/disorders/bppv/movies/Epley-480x640.avi.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

VESTIBULAR MIGRAINE

Vestibular migraine is a very common yet underdiagnosed cause of episodic vertigo. Vertigo sometimes precedes a typical migraine headache but more often occurs without headache or with only a mild headache. Some patients who have had frequent migraine headaches in the past present later in life with vestibular migraine as the predominant problem. In vestibular migraine, the duration of vertigo may be from minutes to hours, and some migraineurs also experience more prolonged periods of disequilibrium (lasting days to weeks). Motion sensitivity and sensitivity to visual motion (e.g., movies) are common. Even in the absence of headache, other migraine features may be present, such as photophobia, phonophobia, or a visual aura. Although data from controlled studies are generally lacking, vestibular migraine typically is treated with medications that are used for prophylaxis of migraine headaches **(Chap. 422)**. Antiemetics may be helpful to relieve symptoms at the time of an attack.

MÉNIÈRE'S DISEASE

Attacks of Ménière's disease consist of vertigo and hearing loss, as well as pain, pressure, and/or fullness in the affected ear. The low-frequency hearing loss and aural symptoms are key features that distinguish Ménière's disease from other peripheral vestibulopathies and from vestibular migraine. Audiometry at the time of an attack shows a characteristic asymmetric low-frequency hearing loss; hearing commonly improves between attacks, although permanent hearing loss may eventually occur. Ménière's disease is thought to be due to excess fluid (endolymph) in the inner ear; hence the term *endolymphatic hydrops*. Patients suspected of having Ménière's disease should be referred to an otolaryngologist for further evaluation. Diuretics and sodium restriction are typically the initial treatments. If attacks persist, injections of glucocorticoids or gentamicin into the middle ear may be considered. Non-ablative surgical options include decompression and shunting of the endolymphatic sac. Full ablative procedures (vestibular nerve section, labyrinthectomy) are seldom required.

VESTIBULAR SCHWANNOMA

Vestibular schwannomas (sometimes termed *acoustic neuromas*) and other tumors at the cerebellopontine angle cause slowly progressive unilateral sensorineural hearing loss and vestibular hypofunction. These patients typically do not have vertigo, because the gradual vestibular deficit is compensated centrally as it develops. The diagnosis often is not made until there is sufficient hearing loss to be noticed. The vestibular examination will show a deficient response to the head impulse test when the head is rotated toward the affected side, but nystagmus will not be prominent. As noted above, patients with unexplained unilateral sensorineural hearing loss or vestibular hypofunction require MRI of the internal auditory canals to look for a schwannoma.

BILATERAL VESTIBULAR HYPOFUNCTION

Patients with bilateral loss of vestibular function also typically do not have vertigo, because vestibular function is lost on both sides simultaneously, and there is no asymmetry of vestibular input. Symptoms include loss of balance, particularly in the dark, where vestibular input is most critical, and oscillopsia during head movement, such as while walking or riding in a car. Bilateral vestibular hypofunction may be (1) idiopathic and progressive, (2) part of a neurodegenerative disorder, or (3) iatrogenic, due to medication ototoxicity (most commonly gentamicin or other aminoglycoside antibiotics). Other causes include bilateral vestibular schwannomas (neurofibromatosis type 2), autoimmune disease, superficial siderosis, and meningeal-based infection or tumor. It also may occur in patients with peripheral polyneuropathy; in these patients, both vestibular loss and impaired proprioception may contribute to poor balance. Finally, unilateral processes such as vestibular neuritis and Ménière's disease may involve both ears sequentially, resulting in bilateral vestibulopathy.

Examination findings include diminished *dynamic visual acuity* (see above) due to loss of stable vision when the head is moving, abnormal head impulse responses in both directions, and a Romberg sign. Responses to caloric testing are reduced. Patients with bilateral vestibular hypofunction should be referred for vestibular rehabilitation therapy. Vestibular suppressant medications should not be used, as they will increase the imbalance. Evaluation by a neurologist is important not only to confirm the diagnosis but also to consider any other associated neurologic abnormalities that may clarify the etiology.

CENTRAL VESTIBULAR DISORDERS

Central lesions causing vertigo typically involve vestibular pathways in the brainstem and/or cerebellum. They may be due to discrete lesions, such as from ischemic or hemorrhagic stroke **(Chaps. 419, 420, 421)** demyelination **(Chap. 436)**, or tumors **(Chap. 86)**, or they may be due to neurodegenerative conditions that include the vestibulocerebellum **(Chaps. 423, 424, 425, 426)**. Subacute cerebellar degeneration may be due to immune, including paraneoplastic, processes **(Chaps. 90 and 431)**. Table 19-1 outlines important features of the history and examination that help to identify central vestibular disorders. Acute central vertigo is a medical emergency, due to the possibility of life-threatening stroke or hemorrhage. All patients with suspected central vestibular disorders should undergo brain MRI, and the patient should be referred for full neurologic evaluation.

PSYCHOSOMATIC AND FUNCTIONAL DIZZINESS

Psychological factors play an important role in chronic dizziness. First, dizziness may be a somatic manifestation of a psychiatric condition such as major depression, anxiety, or panic disorder **(Chap. 443)**. Second, patients may develop anxiety and autonomic symptoms as a consequence or comorbidity of an independent vestibular disorder. One particular form of this has been termed variously *phobic postural vertigo*, *psychophysiologic vertigo*, or *chronic subjective dizziness*, but is now referred to as *persistent postural-perceptual dizziness (PPPD)*. These patients have a chronic feeling (3 months or longer) of fluctuating dizziness and disequilibrium that is present at rest but worse while standing. There is an increased sensitivity to self-motion and visual motion (e.g., watching movies), and a particular intensification of symptoms when moving through complex visual environments such as supermarkets. Although there may be a past history of an acute vestibular disorder (e.g., vestibular neuritis), the neurootologic examination and vestibular testing are normal or indicative of a compensated vestibular deficit, indicating that the ongoing subjective dizziness cannot be explained by a primary vestibular pathology. Anxiety disorders are particularly common in patients with chronic dizziness; when present, they contribute substantially to the morbidity. Treatment approaches for PPPD include pharmacological therapy with selective serotonin reuptake inhibitors (SSRIs), cognitive-behavioral psychotherapy, and vestibular rehabilitation. Vestibular suppressant medications generally should be avoided.

TREATMENT

TREATMENT

Vertigo

Table 19-2 provides a list of commonly used medications for suppression of vertigo. As noted, these medications should be reserved for short-term control of active vertigo, such as during the first few days of acute vestibular neuritis, or for acute attacks of Ménière's disease. They are less helpful for chronic dizziness and, as previously stated, may hinder central compensation. An exception is that benzodiazepines may attenuate psychosomatic dizziness and the associated anxiety, although SSRIs are generally preferable in such patients.

Vestibular rehabilitation therapy promotes central adaptation processes that compensate for vestibular loss and also may help habituate motion sensitivity and other symptoms of psychosomatic dizziness. The general approach is to use a graded series of exercises that progressively challenge gaze stabilization and balance.

TABLE 19-2

Treatment of Vertigo

Agent[a]	Dose[b]
Antihistamines	
Meclizine	25–50 mg 3 times daily
Dimenhydrinate	50 mg 1–2 times daily
Promethazine	25 mg 2–3 times daily (also can be given rectally and IM)
Benzodiazepines	
Diazepam	2.5 mg 1–3 times daily
Clonazepam	0.25 mg 1–3 times daily
Anticholinergic	
Scopolamine transdermal[c]	Patch
Physical therapy	
Repositioning maneuvers[d]	
Vestibular rehabilitation	
Other	
Diuretics and/or low-sodium (1000 mg/d) diet[e]	
Antimigrainous drugs[f]	
Methylprednisolone[g]	100 mg daily days 1–3; 80 mg daily days 4–6; 60 mg daily days 7–9; 40 mg daily days 10–12; 20 mg daily days 13–15; 10 mg daily days 16–18, 20, 22
Selective serotonin reuptake inhibitors[h]	

[a]All listed drugs are approved by the U.S. Food and Drug Administration, but most are not approved for the treatment of vertigo. [b]Usual oral (unless otherwise stated) starting dose in adults; a higher maintenance dose can be reached by a gradual increase. [c]For motion sickness only. [d]For benign paroxysmal positional vertigo. [e]For Ménière's disease. [f]For vestibular migraine. [g]For acute vestibular neuritis (started within 3 days of onset). [h]For persistent postural-perceptual vertigo and anxiety.

FURTHER READING

Dieterich M, Staab JP: Functional dizziness: From phobic postural vertigo and chronic subjective dizziness to persistent postural-perceptual dizziness. Curr Opin Neurol 30:107, 2017. [PubMed: 28002123]

Kim JS, Zee DS: Benign paroxysmal positional vertigo. N Engl J Med 370:1138, 2014. [PubMed: 24645946]

von brevern M, Lempert T: Vestibular Migraine. Handb Clin Neurol 137:301, 2016. [PubMed: 27638080]

Chapter 20: Fatigue

Jeffrey M. Gelfand; Vanja C. Douglas

INTRODUCTION

Fatigue is one of the most common symptoms in clinical medicine. It is a prominent manifestation of a number of systemic, neurologic, and psychiatric syndromes, although a precise cause will not be identified in a substantial minority of patients. Fatigue refers to the subjective human experience of physical and mental weariness, sluggishness, low energy, and exhaustion. In the context of clinical medicine, fatigue is most typically and practically defined as difficulty initiating or maintaining voluntary mental or physical activity. Nearly everyone who has ever been ill with a self-limited infection has experienced this near-universal symptom, and fatigue is usually brought to medical attention only when it is either of unclear cause, fails to remit, or the severity is out of proportion with what would be expected for the associated trigger.

Fatigue should be distinguished from *muscle weakness*, a reduction of neuromuscular power (Chap. 21); most patients complaining of fatigue are not truly weak when direct muscle power is tested. Fatigue is also distinct from *somnolence*, which refers to sleepiness in the context of disturbed sleep-wake physiology (Chap. 27), and from *dyspnea on exertion*, although patients may use the word fatigue to describe those symptoms. The task facing clinicians when a patient presents with fatigue is to identify the underlying cause and to develop a therapeutic alliance, the goal of which is to spare patients expensive and fruitless diagnostic workups and steer them toward effective therapy.

EPIDEMIOLOGY AND GLOBAL CONSIDERATIONS

Variability in the definitions of fatigue and the survey instruments used in different studies makes it difficult to arrive at precise figures about the global burden of fatigue. The point prevalence of fatigue was 6.7% and the lifetime prevalence was 25% in a large National Institute of Mental Health survey of the U.S. general population. In primary care clinics in Europe and the United States, between 10 and 25% of patients surveyed endorsed symptoms of prolonged (present for >1 month) or chronic (present for >6 months) fatigue, but in only a minority was fatigue the primary reason for seeking medical attention. In a community survey of women in India, 12% reported chronic fatigue. By contrast, the prevalence of chronic fatigue syndrome, as defined by the U.S. Centers for Disease Control and Prevention, is low (Chap. 442).

DIFFERENTIAL DIAGNOSIS

Psychiatric Disease

Fatigue is a common somatic manifestation of many major psychiatric syndromes, including depression, anxiety, and somatoform disorders. Psychiatric symptoms are reported in more than three-quarters of patients with unexplained chronic fatigue. Even in patients with systemic or neurologic syndromes in which fatigue is independently recognized as a manifestation of disease, comorbid psychiatric symptoms or disease may still be an important source of interaction.

Neurologic Disease

Patients complaining of fatigue often say they feel weak, but upon careful examination objective muscle weakness is rarely discernible. If found, muscle weakness must then be localized to the central nervous system, peripheral nervous system, neuromuscular junction, or muscle and appropriate follow-up studies obtained **(Chap. 21)**. *Fatigability* of muscle power is a cardinal manifestation of some neuromuscular disorders such as myasthenia gravis and is distinguished from *fatigue* by finding clinically apparent diminution of the amount of force that a muscle generates upon repeated contraction **(Chap. 440)**. Fatigue is one of the most common and bothersome symptoms reported in multiple sclerosis (MS) **(Chap. 436)**, affecting nearly 90% of patients; fatigue in MS can persist between MS attacks and does not necessarily correlate with magnetic resonance imaging (MRI) disease activity. Fatigue is also increasingly identified as a troublesome feature of many neurodegenerative diseases, including Parkinson's disease, central dysautonomias, and amyotrophic lateral sclerosis. Fatigue after stroke is a well-described but poorly understood entity with a widely varying prevalence. Episodic fatigue can be a premonitory symptom of migraine. Fatigue is also a frequent result of traumatic brain injury, often occurring in association with depression and sleep disorders.

Sleep Disorders

Obstructive sleep apnea is an important cause of excessive daytime sleepiness in association with fatigue and should be investigated using overnight polysomnography, particularly in those with prominent snoring, obesity, or other predictors of obstructive sleep apnea **(Chap. 291)**. Whether the cumulative sleep deprivation that is common in modern society contributes to clinically apparent fatigue is not known **(Chap. 27)**.

Endocrine Disorders

Fatigue, sometimes in association with true muscle weakness, can be a heralding symptom of hypothyroidism, particularly in the context of hair loss, dry skin, cold intolerance, constipation, and weight gain. Fatigue associated with heat intolerance, sweating, and palpitations is typical of hyperthyroidism. Adrenal insufficiency can also manifest with unexplained fatigue as a primary or prominent symptom, often with anorexia, weight loss, nausea, myalgias, and arthralgias; hyponatremia, hyperkalemia, and hyperpigmentation may be present at time of diagnosis. Mild hypercalcemia can cause fatigue, which may be relatively vague, whereas severe hypercalcemia can lead to lethargy, stupor, and coma. Both hypoglycemia and hyperglycemia can cause lethargy, often in association with confusion; diabetes mellitus, and in

particular type 1 diabetes, is also associated with fatigue independent of glucose levels. Fatigue may also accompany Cushing's disease, hypoaldosteronism, and hypogonadism. Low vitamin D status has also been associated with fatigue.

Liver and Kidney Disease

Both chronic liver failure and chronic kidney disease can cause fatigue. Over 80% of hemodialysis patients complain of fatigue, which makes it one of the most common symptoms reported by patients in chronic kidney disease.

Obesity

Obesity is associated with fatigue and sleepiness independent of the presence of obstructive sleep apnea. Obese patients undergoing bariatric surgery experience improvement in daytime sleepiness sooner than would be expected if the improvement were solely the result of weight loss and resolution of sleep apnea. A number of other factors common in obese patients are likely contributors as well, including physical inactivity, diabetes, and depression.

Physical Inactivity

Physical inactivity is associated with fatigue, and increasing physical activity can improve fatigue in some patients.

Malnutrition

Although fatigue can be a presenting feature of malnutrition, nutritional status may also be an important comorbidity and contributor to fatigue in other chronic illnesses, including cancer-associated fatigue.

Infection

Both acute and chronic infections commonly lead to fatigue as part of the broader infectious syndrome. Evaluation for undiagnosed infection as the cause of unexplained fatigue, and particularly prolonged or chronic fatigue, should be guided by the history, physical examination, and infectious risk factors, with particular attention to risk for tuberculosis, HIV, chronic hepatitis, and endocarditis. Infectious mononucleosis may cause prolonged fatigue that persists for weeks to months following the acute illness, but infection with the Epstein-Barr virus is only very rarely the cause of unexplained chronic fatigue.

Drugs

Many medications, drugs, drug withdrawal, and chronic alcohol use can all lead to fatigue. Medications that are more likely to be causative include antidepressants, antipsychotics, anxiolytics, opiates, antispasticity agents, antiseizure agents, and beta blockers.

Cardiovascular and Pulmonary

Fatigue is one of the most taxing symptoms reported by patients with congestive heart failure and chronic obstructive pulmonary disease and negatively affects quality of life.

Malignancy

Fatigue, particularly in association with unexplained weight loss, can be a sign of occult malignancy, but cancer is rarely identified in patients with unexplained chronic fatigue in the absence of other telltale signs or symptoms. Cancer-related fatigue is experienced by 40% of patients at the time of diagnosis and by >80% at some time in the disease course.

Hematologic

Chronic or progressive anemia may present with fatigue, sometimes in association with exertional tachycardia and breathlessness. Anemia may also contribute to fatigue in chronic illness. Low serum ferritin in the absence of anemia may also cause fatigue that is reversible with iron replacement.

Systemic Inflammatory/Rheumatologic Disorders

Fatigue is a prominent complaint in many chronic inflammatory disorders, including systemic lupus erythematosus, polymyalgia rheumatica, rheumatoid arthritis, inflammatory bowel disease, antineutrophil cytoplasmic antibody (ANCA)—associated vasculitis, sarcoidosis, and Sjögren's syndrome, but is not usually an isolated symptom. Fatigue is also associated with primary immunodeficiency diseases.

Pregnancy

Fatigue is very commonly reported by women during all stages of pregnancy and postpartum.

Disorders of Unclear Cause

Chronic fatigue syndrome (Chap. 442) and fibromyalgia (Chap. 366) incorporate chronic fatigue as part of the syndromic definition when present in association with a number of other inclusion and exclusion criteria, as discussed in the respective chapters. Chronic multisymptom illness, also known as Gulf-War syndrome, is another symptom complex with prominent fatigue; it is most commonly, although not exclusively, observed in veterans of the 1991 Gulf war conflict (Chap. S6). Idiopathic chronic fatigue is used to describe the syndrome of unexplained chronic fatigue in the absence of enough additional clinical features to meet the diagnostic criteria for chronic fatigue syndrome.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Fatigue

A detailed history focusing on the quality, pattern, time-course, associated symptoms, and alleviating factors of fatigue is critical to define the syndrome and help direct further evaluation and treatment. It is important to determine if fatigue is the appropriate designation, whether symptoms are acute or chronic, and if the impairment is primarily mental, physical, or a combination of the two. The review of systems should attempt to distinguish fatigue from excessive sleepiness, dyspnea on exertion, exercise intolerance, and muscle weakness. The presence of fever, chills, night sweats, or weight loss should raise suspicion for an occult infection or malignancy. A careful review of prescription, over-the-counter, herbal, and recreational drug and alcohol use is required. Circumstances surrounding the onset of symptoms and potential triggers should be investigated. The social history is important, with attention paid to life stressors, workhours, the social support network, and domestic affairs including a screen for intimate partner violence. Sleep habits and sleep hygiene should be questioned. The impact of fatigue on daily functioning is important to understand the patient's experience and gauge recovery and the success of treatment.

The physical examination of patients with fatigue is guided by the history and differential diagnosis. A detailed mental status examination should be performed with particular attention to symptoms of depression and anxiety. A formal neurologic examination is required to determine whether objective muscle weakness is present. This is usually a straightforward exercise, although occasionally patients with fatigue have difficulty sustaining effort against resistance and sometimes report that generating full power requires substantial mental effort. On confrontational testing, full power can be generated for only a brief period before the patient suddenly gives way to the examiner. This type of weakness is often referred to as *breakaway weakness* and may or may not be associated with pain. This is contrasted with weakness due to lesions in the motor tracts or lower motor unit, in which the patient's resistance can be overcome in a smooth and steady fashion and full power can never be generated. Occasionally, a patient may demonstrate fatigable weakness, in which power is full when first tested but becomes weak upon repeat evaluation without interval rest. Fatigable weakness, which usually indicates a problem of neuromuscular transmission, never has the sudden breakaway quality that one occasionally observes in patients with fatigue. If the presence or absence of muscle weakness cannot be determined with the physical examination, electromyography with nerve conductions studies can be a helpful ancillary test.

The general physical examination should screen for signs of cardiopulmonary disease, malignancy, lymphadenopathy, organomegaly, infection, liver failure, kidney disease, malnutrition, endocrine abnormalities, and connective tissue disease. In patients with associated widespread musculoskeletal pain, assessment of tender points may help to reveal fibromyalgia. Although the diagnostic yield of the general physical examination may be relatively low in the context of evaluation of unexplained chronic fatigue, elucidating the cause of only 2% of cases in one prospective analysis, the yield of a detailed neuropsychiatric and mental status evaluation is likely to be much higher, revealing a potential explanation for fatigue in up to 75–80% of patients in some series. Furthermore, a complete physical examination demonstrates a serious and systematic approach to the patient's complaint and helps build trust and a therapeutic alliance.

Laboratory testing is likely to identify the cause of chronic fatigue in only about 5% of cases. Beyond a few standard screening tests, laboratory evaluation should be guided by the history and physical examination;

extensive testing is more likely to lead to false-positive results that require explanation and unnecessary follow-up investigation, and should be avoided in lieu of frequent clinical follow-up. A reasonable approach to screening includes a complete blood count with differential (to screen for anemia, infection, and malignancy), electrolytes (including sodium, potassium, and calcium), glucose, renal function, liver function, and thyroid function. Testing for HIV and adrenal function can also be considered. Published guidelines for chronic fatigue syndrome also recommend an erythrocyte sedimentation rate (ESR) as part of the evaluation for mimics, but unless the value is very high such nonspecific testing in the absence of other features is unlikely to clarify the situation. Routine screening with an antinuclear antibody (ANA) test is also unlikely to be informative in isolation and is frequently positive at low titers in otherwise healthy adults. Additional unfocused studies, such as whole-body imaging scans, are usually not indicated; in addition to their inconvenience, potential risk, and cost, they often reveal unrelated incidental findings that can prolong the workup unnecessarily.

TREATMENT

TREATMENT

Fatigue

The first priority of treatment is to address the underlying disorder or disorders that account for fatigue, because this can be curative in select contexts and palliative in others. Unfortunately, in many chronic illnesses fatigue may be refractory to traditional disease-modifying therapies, but it is nevertheless important in such cases to evaluate for other potential contributors, because the cause may be multifactorial. Antidepressant treatment (Chap. 444) may be helpful for treatment of chronic fatigue when symptoms of depression are present and may be most effective as part of a multimodal approach. However, antidepressants can also cause fatigue and should be discontinued if they are not clearly effective. Cognitive-behavioral therapy has also been demonstrated to be helpful in the context of chronic fatigue syndrome as well as cancer-associated fatigue. Both cognitive behavioral therapy and graded exercise therapy, in which physical exercise, most typically walking, is gradually increased with attention to target heart rates to avoid overexertion, were shown to modestly improve walking times and self-reported fatigue measures when compared to standard medical care in patients in the United Kingdom with chronic fatigue. These benefits were maintained after a median follow-up of 2.5 years. Psychostimulants such as amphetamines, modafinil, and armodafinil can help increase alertness and concentration and reduce excessive daytime sleepiness in certain clinical contexts, which may in turn help with symptoms of fatigue in a minority of patients, but they have generally proven to be unhelpful in randomized trials for treating fatigue in posttraumatic brain injury, Parkinson's disease, cancer, and MS. In patients with low vitamin D status, vitamin D replacement may lead to improvement in fatigue.

Development of more effective therapy for fatigue is hampered by limited knowledge of the biologic basis of this symptom, including how fatigue is detected and registered in the nervous system. Proinflammatory cytokines, such as interleukin 1α and 1β, and tumor necrosis factor α, might mediate fatigue in some patients. Preliminary data suggests that biological therapies that inhibit IL-1 or other cytokines can help to

ameliorate fatigue in some patients with inflammatory conditions in addition to, or as part of, their disease modifying effect; thus, cytokine antagonists represent one possible future approach.

PROGNOSIS

Acute fatigue significant enough to require medical evaluation is more likely to lead to an identifiable medical, neurologic, or psychiatric cause than unexplained chronic fatigue. Evaluation of unexplained chronic fatigue most commonly leads to diagnosis of a psychiatric condition or remains unexplained. Identification of a previously undiagnosed serious or life-threatening culprit etiology is rare on longitudinal follow-up in patients with unexplained chronic fatigue. Complete resolution of unexplained chronic fatigue is uncommon, at least over the short term, but multidisciplinary treatment approaches can lead to symptomatic improvements that can substantially improve quality of life.

FURTHER READING

David A et al: Tired, weak, or in need of rest: Fatigue among general practice attenders. BMJ 301:1199, 1990. [PubMed: 2261560]

Kroenke K et al: Chronic fatigue in primary care. Prevalence, patient characteristics, and outcome. JAMA 260:929–934, 1988. [PubMed: 3398197]

Roerink ME et al: Interleukin-1 as a mediator of fatigue in disease: A narrative review. J Neuroinflammation 14:16, 2017. [PubMed: 28109186]

Sharpe M et al: Rehabilitative treatments for chronic fatigue syndrome: Long-term follow-up from the PACE trial. Lancet Psychiatry 2:1067, 2015. [PubMed: 26521770]

White PD et al: Comparison of adaptive pacing therapy, cognitive behaviour therapy, graded exercise therapy, and specialist medical care for chronic fatigue syndrome (PACE): A randomised trial. Lancet 377:823, 2011. [PubMed: 21334061]

Chapter 21: Neurologic Causes of Weakness and Paralysis

Michael J. Aminoff

INTRODUCTION

Normal motor function involves integrated muscle activity that is modulated by the activity of the cerebral cortex, basal ganglia, cerebellum, red nucleus, brainstem reticular formation, lateral vestibular nucleus, and spinal cord. Motor system dysfunction leads to weakness or paralysis, discussed in this chapter, or to ataxia (Chap. 431) or abnormal movements (Chap. 428). *Weakness* is a reduction in the power that can be exerted by one or more muscles. It must be distinguished from increased *fatigability* (i.e., the inability to sustain the performance of an activity that should be normal for a person of the same age, sex, and size), limitation in function due to pain or articular stiffness, or impaired motor activity because severe *proprioceptive sensory loss* prevents adequate feedback information about the direction and power of movements. It is also distinct from *bradykinesia* (in which increased time is required for full power to be exerted) and *apraxia,* a disorder of planning and initiating a skilled or learned movement unrelated to a significant motor or sensory deficit (Chap. 26).

Paralysis or the suffix "-plegia" indicates weakness so severe that a muscle cannot be contracted at all, whereas *paresis* refers to less severe weakness. The prefix "hemi-" refers to one-half of the body, "para-" to both legs, and "quadri-" to all four limbs.

The *distribution* of weakness helps to localize the underlying lesion. Weakness from involvement of upper motor neurons occurs particularly in the extensors and abductors of the upper limb and the flexors of the lower limb. Lower motor neuron weakness depends on whether involvement is at the level of the anterior horn cells, nerve root, limb plexus, or peripheral nerve—only muscles supplied by the affected structure are weak. Myopathic weakness is generally most marked in proximal muscles. Weakness from impaired neuromuscular transmission has no specific pattern of involvement.

Weakness often is accompanied by other neurologic abnormalities that help indicate the site of the responsible lesion (Table 21-1).

TABLE 21-1

Signs That Distinguish the Origin of Weakness

Sign	Upper Motor Neuron	Lower Motor Neuron	Myopathic	Psychogenic
Atrophy	None	Severe	Mild	None
Fasciculations	None	Common	None	None
Tone	Spastic	Decreased	Normal/decreased	Variable/paratonia
Distribution of weakness	Pyramidal/regional	Distal/segmental	Proximal	Variable/inconsistent with daily activities
Muscle stretch reflexes	Hyperactive	Hypoactive/absent	Normal/hypoactive	Normal
Babinski sign	Present	Absent	Absent	Absent

Tone is the resistance of a muscle to passive stretch. Increased tone may be of several types. *Spasticity* is the increase in tone associated with disease of upper motor neurons. It is velocity-dependent, has a sudden release after reaching a maximum (the "clasp-knife" phenomenon), and predominantly affects the antigravity muscles (i.e., upper-limb flexors and lower-limb extensors). *Rigidity* is hypertonia that is present throughout the range of motion (a "lead pipe" or "plastic" stiffness) and affects flexors and extensors equally; it sometimes has a cogwheel quality that is enhanced by voluntary movement of the contralateral limb (reinforcement). Rigidity occurs with certain extrapyramidal disorders, such as Parkinson's disease. *Paratonia* (or *gegenhalten*) is increased tone that varies irregularly in a manner seemingly related to the degree of relaxation, is present throughout the range of motion, and affects flexors and extensors equally; it usually results from disease of the frontal lobes. Weakness with *decreased tone* (*flaccidity*) or normal tone occurs with disorders of *motor units.* A motor unit consists of a single lower motor neuron and all the muscle fibers that it innervates.

Muscle bulk generally is not affected by upper motor neuron lesions, although mild disuse atrophy eventually may occur. By contrast, atrophy is often conspicuous when a lower motor neuron lesion is responsible for weakness and also may occur with advanced muscle disease.

Muscle stretch (tendon) reflexes are usually increased with upper motor neuron lesions, but may be decreased or absent for a variable period immediately after onset of an acute lesion. Hyperreflexia is usually —but not invariably—accompanied by loss of *cutaneous reflexes* (such as superficial abdominals; **Chap. 415**)

and, in particular, by an extensor plantar (Babinski) response. The muscle stretch reflexes are depressed with lower motor neuron lesions directly involving specific reflex arcs. They generally are preserved in patients with myopathic weakness except in advanced stages, when they sometimes are attenuated. In disorders of the neuromuscular junction, reflex responses may be affected by preceding voluntary activity of affected muscles; such activity may lead to enhancement of initially depressed reflexes in Lambert-Eaton myasthenic syndrome and, conversely, to depression of initially normal reflexes in myasthenia gravis (Chap. 440).

The distinction of *neuropathic* (lower motor neuron) from *myopathic* weakness is sometimes difficult clinically, although distal weakness is likely to be neuropathic, and symmetric proximal weakness myopathic. *Fasciculations* (visible or palpable twitch within a muscle due to the spontaneous discharge of a motor unit) and early atrophy indicate that weakness is neuropathic.

PATHOGENESIS

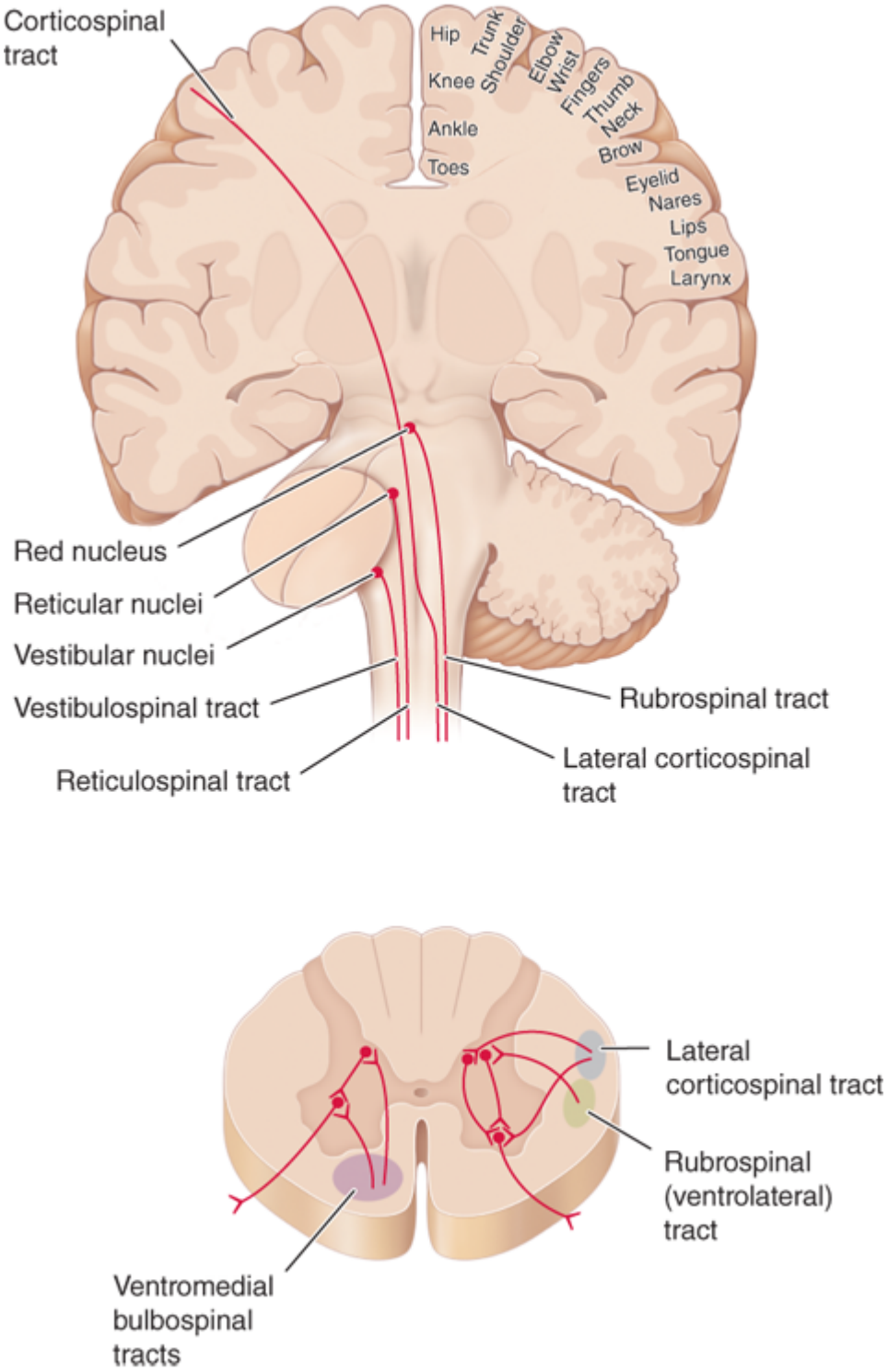
Upper Motor Neuron Weakness

Lesions of the upper motor neurons or their descending axons to the spinal cord (Fig. 21-1) produce weakness through decreased activation of lower motor neurons. In general, distal muscle groups are affected more severely than proximal ones, and axial movements are spared unless the lesion is severe and bilateral. Spasticity is typical but may not be present acutely. Rapid repetitive movements are slowed and coarse, but normal rhythmicity is maintained. With corticobulbar involvement, weakness occurs in the lower face and tongue; extraocular, upper facial, pharyngeal, and jaw muscles are typically spared. Bilateral corticobulbar lesions produce a *pseudobulbar palsy*: dysarthria, dysphagia, dysphonia, and emotional lability accompany bilateral facial weakness and a brisk jaw jerk. Electromyogram (EMG) (Chap. 438) shows that with weakness of the upper motor neuron type, motor units have a diminished maximal discharge frequency.

FIGURE 21-1

The corticospinal and bulbospinal upper motor neuron pathways. Upper motor neurons have their cell bodies in layer V of the primary motor cortex (the precentral gyrus, or Brodmann's area 4) and in the premotor and supplemental motor cortex (area 6). The upper motor neurons in the primary motor cortex are somatotopically organized (right side of figure). Axons of the upper motor neurons descend through the subcortical white matter and the posterior limb of the internal capsule. Axons of the *pyramidal* or *corticospinal system* descend through the brainstem in the cerebral peduncle of the midbrain, the basis pontis, and the medullary pyramids. At the cervicomedullary junction, most corticospinal axons decussate into the contralateral corticospinal tract of the lateral spinal cord, but 10–30% remain ipsilateral in the anterior spinal cord. Corticospinal neurons synapse on premotor interneurons, but some—especially in the cervical enlargement and those connecting with motor neurons to distal limb muscles—make direct monosynaptic connections with lower motor neurons. They innervate most densely the lower motor neurons of hand muscles and are involved in the execution of learned, fine movements. Corticobulbar neurons are similar to corticospinal neurons but innervate brainstem motor nuclei. *Bulbospinal upper motor neurons* influence strength and tone but are not part of the pyramidal system. The descending *ventromedial bulbospinal pathways* originate in the tectum of the midbrain (tectospinal pathway), the vestibular nuclei

(vestibulospinal pathway), and the reticular formation (reticulospinal pathway). These pathways influence axial and proximal muscles and are involved in the maintenance of posture and integrated movements of the limbs and trunk. The descending *ventrolateral bulbospinal pathways*, which originate predominantly in the red nucleus (rubrospinal pathway), facilitate distal limb muscles. The bulbospinal system sometimes is referred to as the *extrapyramidal upper motor neuron system*. In all figures, nerve cell bodies and axon terminals are shown, respectively, as closed circles and forks.



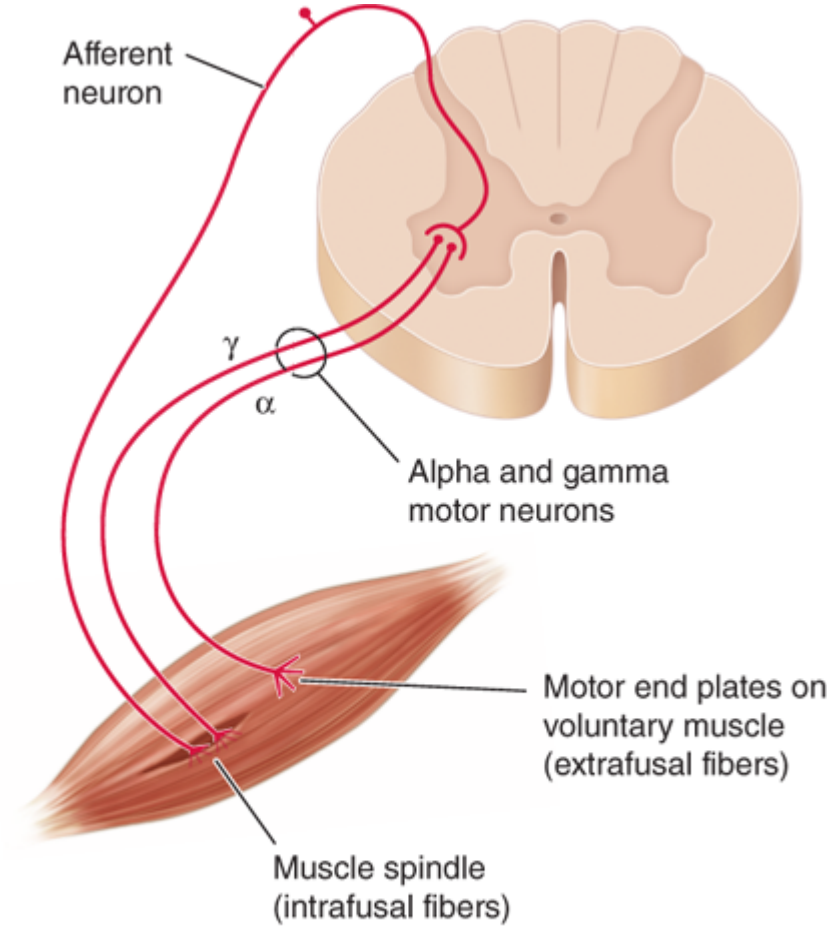
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Lower Motor Neuron Weakness

This pattern results from disorders of lower motor neurons in the brainstem motor nuclei and the anterior horn of the spinal cord or from dysfunction of the axons of these neurons as they pass to skeletal muscle **(Fig. 21-2)**. Weakness is due to a decrease in the number of muscle fibers that can be activated through a loss of α motor neurons or disruption of their connections to muscle. Loss of γ motor neurons does not cause weakness but decreases tension on the muscle spindles, which decreases muscle tone and attenuates the stretch reflexes. An absent stretch reflex suggests involvement of spindle afferent fibers.

FIGURE 21-2

Lower motor neurons are divided into α and γ types. The larger α motor neurons are more numerous and innervate the extrafusal muscle fibers of the motor unit. Loss of α motor neurons or disruption of their axons produces lower motor neuron weakness. The smaller, less numerous γ motor neurons innervate the intrafusal muscle fibers of the muscle spindle and contribute to normal tone and stretch reflexes. The α motor neuron receives direct excitatory input from corticomotoneurons and primary muscle spindle afferents. The α and γ motor neurons also receive excitatory input from other descending upper motor neuron pathways, segmental sensory inputs, and interneurons. The α motor neurons receive direct inhibition from Renshaw cell interneurons, and other interneurons indirectly inhibit the α and γ motor neurons. A muscle stretch (tendon) reflex requires the function of all the illustrated structures. A tap on a tendon stretches muscle spindles (which are tonically activated by γ motor neurons) and activates the primary spindle afferent neurons. These neurons stimulate the α motor neurons in the spinal cord, producing a brief muscle contraction, which is the familiar tendon reflex.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

When a motor unit becomes diseased, especially in anterior horn cell diseases, it may discharge spontaneously, producing *fasciculations*. When α motor neurons or their axons degenerate, the denervated muscle fibers also may discharge spontaneously. These single muscle fiber discharges, or *fibrillation potentials,* cannot be seen but can be recorded with EMG. Weakness leads to delayed or reduced recruitment of motor units, with fewer than normal activated at a particular discharge frequency.

Neuromuscular Junction Weakness

Disorders of the neuromuscular junctions produce weakness of variable degree and distribution. The number of muscle fibers that are activated varies over time, depending on the state of rest of the neuromuscular junctions. Strength is influenced by preceding activity of the affected muscle. In myasthenia gravis, for example, sustained or repeated contractions of affected muscle decline in strength despite continuing effort (Chap. 440). Thus, fatigable weakness is suggestive of disorders of the neuromuscular junction, which cause functional loss of muscle fibers due to failure of their activation.

Myopathic Weakness

Myopathic weakness is produced by a decrease in the number or contractile force of muscle fibers activated within motor units. With muscular dystrophies, inflammatory myopathies, or myopathies with muscle fiber

necrosis, the number of muscle fibers is reduced within many motor units. On EMG, the size of each motor unit action potential is decreased, and motor units must be recruited more rapidly than normal to produce the desired power. Some myopathies produce weakness through loss of contractile force of muscle fibers or through relatively selective involvement of type II (fast) fibers. These myopathies may not affect the size of individual motor unit action potentials and are detected by a discrepancy between the electrical activity and force of a muscle.

Psychogenic Weakness

Weakness may occur without a recognizable organic basis. It tends to be variable, inconsistent, and with a pattern of distribution that cannot be explained on a neuroanatomic basis. On formal testing, antagonists may contract when the patient is supposedly activating the agonist muscle. The severity of weakness is out of keeping with the patient's daily activities.

DISTRIBUTION OF WEAKNESS

Hemiparesis

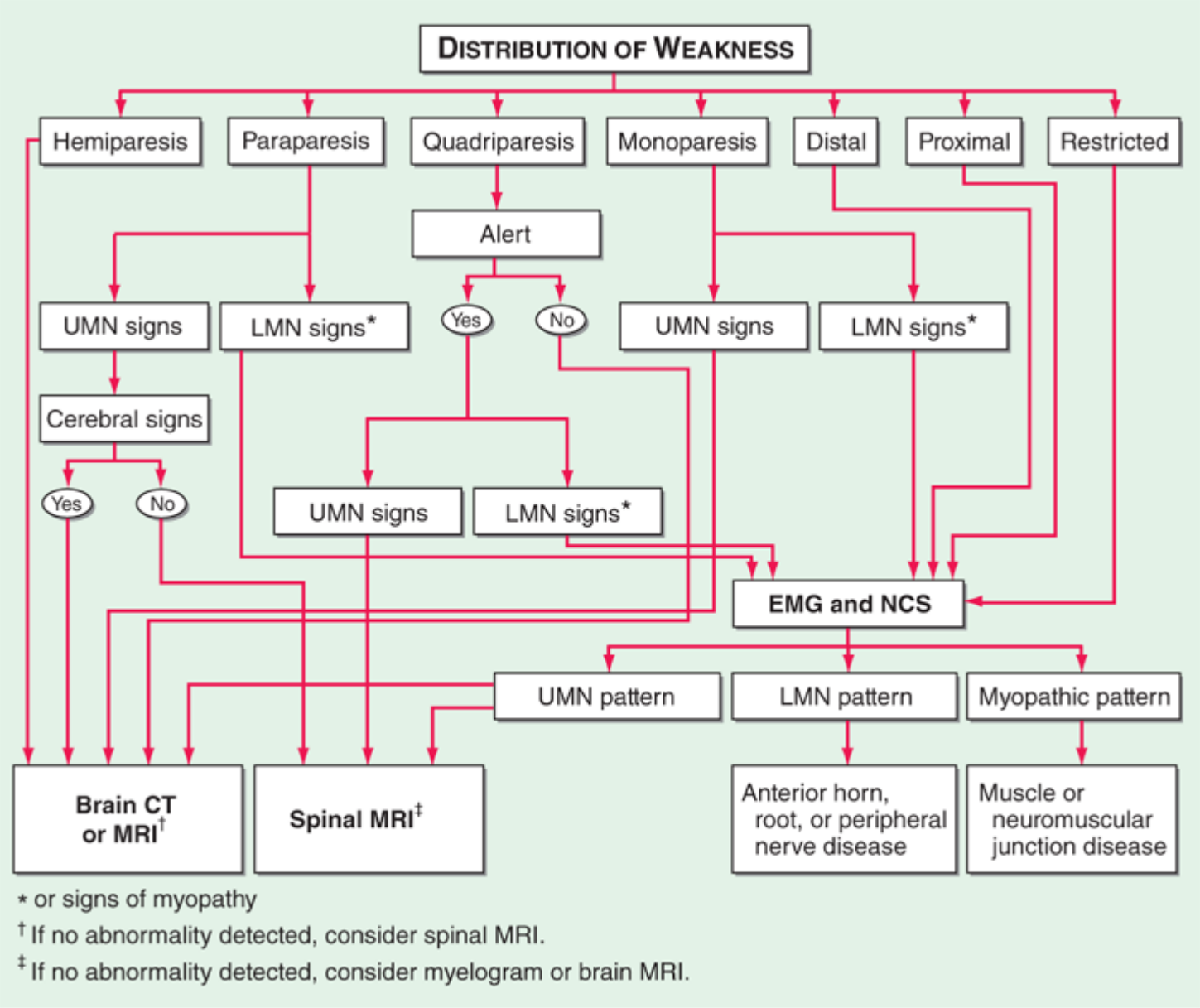
Hemiparesis results from an upper motor neuron lesion above the midcervical spinal cord; most such lesions are above the foramen magnum. The presence of other neurologic deficits helps localize the lesion. Thus, language disorders, for example, point to a cortical lesion. Homonymous visual field defects reflect either a cortical or a subcortical hemispheric lesion. A "pure motor" hemiparesis of the face, arm, and leg often is due to a small, discrete lesion in the posterior limb of the internal capsule, cerebral peduncle in the midbrain, or upper pons. Some brainstem lesions produce "crossed paralyses," consisting of ipsilateral cranial nerve signs and contralateral hemiparesis (Chap. 419). The absence of cranial nerve signs or facial weakness suggests that a hemiparesis is due to a lesion in the high cervical spinal cord, especially if associated with the Brown-Séquard syndrome (Chap. 434).

Acute or episodic hemiparesis usually results from focal structural lesions, particularly rapidly expanding lesions, or an inflammatory process. *Subacute hemiparesis* that evolves over days or weeks may relate to subdural hematoma, infectious or inflammatory disorders (e.g., cerebral abscess, fungal granuloma or meningitis, parasitic infection, multiple sclerosis, sarcoidosis), or primary or metastatic neoplasms. AIDS may present with subacute hemiparesis due to toxoplasmosis or primary central nervous system (CNS) lymphoma. *Chronic hemiparesis* that evolves over months usually is due to a neoplasm or vascular malformation, a chronic subdural hematoma, or a degenerative disease.

Investigation of hemiparesis (Fig. 21-3) of acute origin starts with a computed tomography (CT) scan of the brain and laboratory studies. If the CT is normal, or in subacute or chronic cases of hemiparesis, magnetic resonance imaging (MRI) of the brain and/or cervical spine (including the foramen magnum) is performed, depending on the clinical accompaniments.

FIGURE 21-3

An algorithm for the initial workup of a patient with weakness. CT, computed tomography; EMG, electromyography; LMN, lower motor neuron; MRI, magnetic resonance imaging; NCS, nerve conduction studies; UMN, upper motor neuron.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Paraparesis

Acute paraparesis is caused most commonly by an intraspinal lesion, but its spinal origin may not be recognized initially if the legs are flaccid and areflexic. Usually, however, there is sensory loss in the legs with an upper level on the trunk, a dissociated sensory loss suggestive of a central cord syndrome **(Chap. 434)**, or hyperreflexia in the legs with normal reflexes in the arms. Imaging the spinal cord (Fig. 21-3) may reveal compressive lesions, infarction (proprioception usually is spared), arteriovenous fistulas or other vascular anomalies, or transverse myelitis **(Chap. 434)**.

Diseases of the cerebral hemispheres that produce acute paraparesis include anterior cerebral artery ischemia (shoulder shrug also is affected), superior sagittal sinus or cortical venous thrombosis, and acute hydrocephalus.

Paraparesis may result from a cauda equina syndrome, for example, after trauma to the low back, a midline disk herniation, or an intraspinal tumor. The sphincters are commonly affected, whereas hip flexion often is spared, as is sensation over the anterolateral thighs. Rarely, paraparesis is caused by a rapidly evolving anterior horn cell disease (such as poliovirus or West Nile virus infection), peripheral neuropathy (such as Guillain-Barré syndrome; **Chap. 439**), or myopathy **(Chap. 441)**.

Subacute or chronic spastic paraparesis is caused by upper motor neuron disease. When associated with lower-limb sensory loss and sphincter involvement, a chronic spinal cord disorder should be considered **(Chap. 434)**. If hemispheric signs are present, a parasagittal meningioma or chronic hydrocephalus is likely. The absence of spasticity in a long-standing paraparesis suggests a lower motor neuron or myopathic etiology.

Investigations typically begin with spinal MRI, but when upper motor neuron signs are associated with drowsiness, confusion, seizures, or other hemispheric signs, brain MRI should also be performed, sometimes as the initial investigation. Electrophysiologic studies are diagnostically helpful when clinical findings suggest an underlying neuromuscular disorder.

Quadriparesis or Generalized Weakness

Generalized weakness may be due to disorders of the CNS or the motor unit. Although the terms often are used interchangeably, *quadriparesis* is commonly used when an upper motor neuron cause is suspected, and *generalized weakness* is used when a disease of the motor units is likely. Weakness from CNS disorders usually is associated with changes in consciousness or cognition and accompanied by spasticity, hyperreflexia, and sensory disturbances. Most neuromuscular causes of generalized weakness are associated with normal mental function, hypotonia, and hypoactive muscle stretch reflexes. The major causes of intermittent weakness are listed in **Table 21-2**. A patient with generalized fatigability without objective weakness may have the chronic fatigue syndrome **(Chap. 442)**.

TABLE 21-2

Causes of Episodic Generalized Weakness

1. Electrolyte disturbances, e.g., hypokalemia, hyperkalemia, hypercalcemia, hypernatremia, hyponatremia, hypophosphatemia, hypermagnesemia
2. Muscle disorders
    a. Channelopathies (periodic paralyses)
    b. Metabolic defects of muscle (impaired carbohydrate or fatty acid utilization; abnormal mitochondrial function)
3. Neuromuscular junction disorders
    a. Myasthenia gravis
    b. Lambert-Eaton myasthenic syndrome
4. Central nervous system disorders
    a. Transient ischemic attacks of the brainstem
    b. Transient global cerebral ischemia
    c. Multiple sclerosis
5. Lack of voluntary effort
    a. Anxiety
    b. Pain or discomfort
    c. Somatization disorder

ACUTE QUADRIPARESIS

Quadriparesis with onset over minutes may result from disorders of upper motor neurons (such as from anoxia, hypotension, brainstem or cervical cord ischemia, trauma, and systemic metabolic abnormalities) or muscle (electrolyte disturbances, certain inborn errors of muscle energy metabolism, toxins, and periodic paralyses). Onset over hours to weeks may, in addition to these disorders, be due to lower motor neuron disorders such as Guillain-Barré syndrome **(Chap. 439)**.

In obtunded patients, evaluation begins with a CT scan of the brain. If upper motor neuron signs are present but the patient is alert, the initial test is usually an MRI of the cervical cord. If weakness is lower motor neuron, myopathic, or uncertain in origin, the clinical approach begins with blood studies to determine the level of muscle enzymes and electrolytes and with EMG and nerve conduction studies.

SUBACUTE OR CHRONIC QUADRIPARESIS

Quadriparesis due to upper motor neuron disease may develop over weeks to years from chronic myelopathies, multiple sclerosis, brain or spinal tumors, chronic subdural hematomas, and various

metabolic, toxic, and infectious disorders. It may also result from lower motor neuron disease, a chronic neuropathy (in which weakness is often most profound distally), or myopathic weakness (typically proximal).

When *quadriparesis* develops acutely in obtunded patients, evaluation begins with a CT scan of the brain. If upper motor neuron signs have developed acutely but the patient is alert, the initial test is usually an MRI of the cervical cord. When onset has been gradual, disorders of the cerebral hemispheres, brainstem, and cervical spinal cord can usually be distinguished clinically, and imaging is directed first at the clinically suspected site of pathology. If weakness is lower motor neuron, myopathic, or uncertain in origin, laboratory studies to determine the levels of muscle enzymes and electrolytes, and EMG and nerve conduction studies help to localize the pathologic process.

Monoparesis

Monoparesis usually is due to lower motor neuron disease, with or without associated sensory involvement. Upper motor neuron weakness occasionally presents as a monoparesis of distal and nonantigravity muscles. Myopathic weakness rarely is limited to one limb.

ACUTE MONOPARESIS

If weakness is predominantly distal and of upper motor neuron type and is not associated with sensory impairment or pain, focal cortical ischemia is likely **(Chap. 420)**; diagnostic possibilities are similar to those for acute hemiparesis. Sensory loss and pain usually accompany acute lower motor neuron weakness; the weakness commonly localizes to a single nerve root or peripheral nerve, but occasionally reflects plexus involvement. If lower motor neuron weakness is likely, evaluation begins with EMG and nerve conduction studies.

SUBACUTE OR CHRONIC MONOPARESIS

Weakness and atrophy that develop over weeks or months are usually of lower motor neuron origin. When associated with sensory symptoms, a peripheral cause (nerve, root, or plexus) is likely; otherwise, anterior horn cell disease should be considered. In either case, an electrodiagnostic study is indicated. If weakness is of the upper motor neuron type, a discrete cortical (precentral gyrus) or cord lesion may be responsible, and appropriate imaging is performed.

Distal Weakness

Involvement of two or more limbs distally suggests lower motor neuron or peripheral nerve disease. Acute distal lower-limb weakness results occasionally from an acute toxic polyneuropathy or cauda equina syndrome. Distal symmetric weakness usually develops over weeks, months, or years and, when associated with numbness, is due to peripheral neuropathy **(Chap. 438)**. Anterior horn cell disease may begin distally but is typically asymmetric and without accompanying numbness **(Chap. 429)**. Rarely, myopathies present with distal weakness **(Chap. 441)**. Electrodiagnostic studies help localize the disorder (Fig. 21-3).

Proximal Weakness

Myopathy often produces symmetric weakness of the pelvic or shoulder girdle muscles **(Chap. 441)**. Diseases of the neuromuscular junction, such as myasthenia gravis **(Chap. 440)**, may present with symmetric proximal weakness often associated with ptosis, diplopia, or bulbar weakness and fluctuating in severity during the day. In anterior horn cell disease, proximal weakness is usually asymmetric, but it may be symmetric if familial. Numbness does not occur with any of these diseases. The evaluation usually begins with determination of the serum creatine kinase level and electrophysiologic studies.

Weakness in a Restricted Distribution

Weakness may not fit any of these patterns, being limited, for example, to the extraocular, hemifacial, bulbar, or respiratory muscles. If it is unilateral, restricted weakness usually is due to lower motor neuron or peripheral nerve disease, such as in a facial palsy. Weakness of part of a limb is commonly due to a peripheral nerve lesion such as an entrapment neuropathy. Relatively symmetric weakness of extraocular or bulbar muscles frequently is due to a myopathy **(Chap. 441)** or neuromuscular junction disorder **(Chap. 440)**. Bilateral facial palsy with areflexia suggests Guillain-Barré syndrome **(Chap. 439)**. Worsening of relatively symmetric weakness with fatigue is characteristic of neuromuscular junction disorders. Asymmetric bulbar weakness usually is due to motor neuron disease. Weakness limited to respiratory muscles is uncommon and usually is due to motor neuron disease, myasthenia gravis, or polymyositis/dermatomyositis **(Chap. 358)**.

FURTHER READING

Brazis P, Masdeu JC, Biller J: *Localization in Clinical Neurology*, 7th ed. Philadelphia, Lippincott William & Wilkins, 2016.

Campbell WW: *DeJong's The Neurological Examination*, 7th ed. Philadelphia, Lippincott William & Wilkins, 2012.

Guarantors of Brain: *Aids to the Examination of the Peripheral Nervous System*, 4th ed. Edinburgh, Saunders, 2000.

Chapter 22: Numbness, Tingling, and Sensory Loss

Michael J. Aminoff

INTRODUCTION

Normal somatic sensation reflects a continuous monitoring process, little of which reaches consciousness under ordinary conditions. By contrast, disordered sensation, particularly when experienced as painful, is alarming and dominates the patient's attention. Physicians should be able to recognize abnormal sensations by how they are described, know their type and likely site of origin, and understand their implications. **Pain is considered separately in Chap. 10.**

POSITIVE AND NEGATIVE SYMPTOMS

Abnormal sensory symptoms can be divided into two categories: positive and negative. The prototypical positive symptom is tingling (pins and needles); other positive sensory phenomena include itch and altered sensations that are described as pricking, bandlike, lightning-like shooting feelings (lancinations), aching, knifelike, twisting, drawing, pulling, tightening, burning, searing, electrical, or raw feelings. Such symptoms are often painful.

Positive phenomena usually result from trains of impulses generated at sites of lowered threshold or heightened excitability along a peripheral or central sensory pathway. The nature and severity of the abnormal sensation depend on the number, rate, timing, and distribution of ectopic impulses and the type and function of nervous tissue in which they arise. Because positive phenomena represent excessive activity in sensory pathways, they are not necessarily associated with a sensory deficit (loss) on examination.

Negative phenomena represent loss of sensory function and are characterized by diminished or absent feeling that often is experienced as numbness and by abnormal findings on sensory examination. In disorders affecting peripheral sensation, at least one-half the afferent axons innervating a particular site are probably lost or functionless before a sensory deficit can be demonstrated by clinical examination. If the rate of loss is slow, however, lack of cutaneous feeling may be unnoticed by the patient and difficult to demonstrate on examination, even though few sensory fibers are functioning; if it is rapid, both positive and negative phenomena are usually conspicuous. Subclinical degrees of sensory dysfunction may be revealed by sensory nerve conduction studies or somatosensory-evoked potentials.

Whereas sensory symptoms may be either positive or negative, sensory signs on examination are always a measure of negative phenomena.

TERMINOLOGY

Paresthesias and dysesthesias are general terms used to denote positive sensory symptoms. The term *paresthesias* typically refers to tingling or pins-and-needles sensations but may include a wide variety of other abnormal sensations, except pain; it sometimes implies that the abnormal sensations are perceived spontaneously. The more general term *dysesthesias* denotes all types of abnormal sensations, including painful ones, regardless of whether a stimulus is evident.

Another set of terms refers to sensory abnormalities found on examination. *Hypesthesia* or *hypoesthesia* refers to a reduction of cutaneous sensation to a specific type of testing such as pressure, light touch, and warm or cold stimuli; *anesthesia,* to a complete absence of skin sensation to the same stimuli plus pinprick; and *hypalgesia* or *analgesia,* to reduced or absent pain perception (nociception). *Hyperesthesia* means pain or increased sensitivity in response to touch. Similarly, *allodynia* describes the situation in which a nonpainful stimulus, once perceived, is experienced as painful, even excruciating. An example is elicitation of a painful sensation by application of a vibrating tuning fork. *Hyperalgesia* denotes severe pain in response to a mildly noxious stimulus, and *hyperpathia*, a broad term, encompasses all the phenomena described by hyperesthesia, allodynia, and hyperalgesia. With hyperpathia, the threshold for a sensory stimulus is increased and perception is delayed, but once felt, it is unduly painful.

Disorders of deep sensation arising from muscle spindles, tendons, and joints affect proprioception (position sense). Manifestations include imbalance (particularly with eyes closed or in the dark), clumsiness of precision movements, and unsteadiness of gait, which are referred to collectively as *sensory ataxia*. Other findings on examination usually, but not invariably, include reduced or absent joint position and vibratory sensibility and absent deep tendon reflexes in the affected limbs. The Romberg sign is positive, which means that the patient sways markedly or topples when asked to stand with feet close together and eyes closed. In severe states of deafferentation involving deep sensation, the patient cannot walk or stand unaided or even sit unsupported. Continuous involuntary movements (*pseudoathetosis*) of the outstretched hands and fingers occur, particularly with eyes closed.

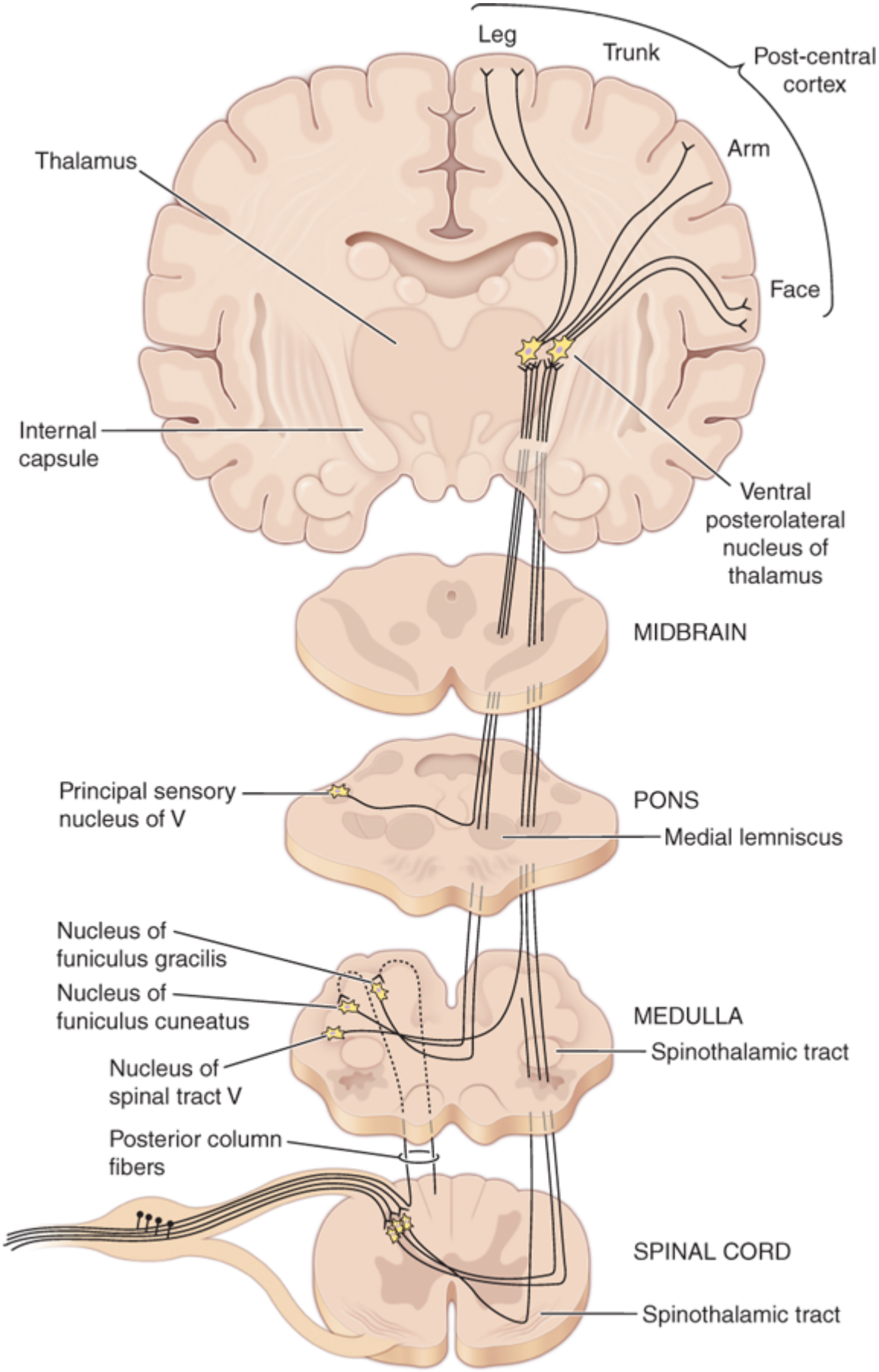
ANATOMY OF SENSATION

Cutaneous receptors are classified by the type of stimulus that optimally excites them. They consist of naked nerve endings (nociceptors, which respond to tissue-damaging stimuli, and thermoreceptors, which respond to noninjurious thermal stimuli) and encapsulated terminals (several types of mechanoreceptor, activated by physical deformation of the skin). Each type of receptor has its own set of sensitivities to specific stimuli, size and distinctness of receptive fields, and adaptational qualities.

Afferent fibers in peripheral nerve trunks traverse the dorsal roots and enter the dorsal horn of the spinal cord **(Fig. 22-1)**. From there, the polysynaptic projections of the smaller fibers (unmyelinated and small myelinated), which subserve mainly nociception, itch, temperature sensibility, and touch, cross and ascend in the opposite anterior and lateral columns of the spinal cord, through the brainstem, to the ventral posterolateral (VPL) nucleus of the thalamus and ultimately project to the postcentral gyrus of the parietal cortex and other cortical areas **(Chap. 10)**. This is the *spinothalamic pathway* or *anterolateral system*. The larger fibers, which subserve tactile and position sense and kinesthesia, project rostrally in the posterior and posterolateral columns on the same side of the spinal cord and make their first synapse in the gracile or cuneate nucleus of the lower medulla. Axons of second-order neurons decussate and ascend in the medial lemniscus located medially in the medulla and in the tegmentum of the pons and midbrain and synapse in the VPL nucleus; third-order neurons project to parietal cortex as well as to other cortical areas. This large-fiber system is referred to as the *posterior column–medial lemniscal pathway* (lemniscal, for short). Although the fiber types and functions that make up the spinothalamic and lemniscal systems are relatively well known, many other fibers, particularly those associated with touch, pressure, and position sense, ascend in a diffusely distributed pattern both ipsilaterally and contralaterally in the anterolateral quadrants of the spinal cord. This explains why a complete lesion of the posterior columns of the spinal cord may be associated with little sensory deficit on examination.

FIGURE 22-1

The main somatosensory pathways. The spinothalamic tract (pain, thermal sense) and the posterior column–lemniscal system (touch, pressure, joint position) are shown. Offshoots from the ascending anterolateral fasciculus (spinothalamic tract) to nuclei in the medulla, pons, and mesencephalon and nuclear terminations of the tract are indicated. *(From AH Ropper, MA Samuels: Adams and Victor's Principles of Neurology, 9th ed. New York, McGraw-Hill, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Nerve conduction studies and nerve biopsy are important means of investigating the peripheral nervous system, but they do not evaluate the function or structure of cutaneous receptors and free nerve endings or of unmyelinated or thinly myelinated nerve fibers in the nerve trunks. Skin biopsy can be used to evaluate these structures in the dermis and epidermis.

CLINICAL EXAMINATION OF SENSATION

The main components of the sensory examination are tests of primary sensation (pain, touch, vibration, joint position, and thermal sensation) (Table 22-1). The examiner must depend on patient responses, and this complicates interpretation. Further, examination may be limited in some patients. In a stuporous patient, for example, sensory examination is reduced to observing the briskness of withdrawal in response to a pinch or another noxious stimulus. Comparison of responses on the two sides of the body is essential. In an alert but uncooperative patient, it may not be possible to examine cutaneous sensation, but some idea of proprioceptive function may be gained by noting the patient's best performance of movements requiring balance and precision.

TABLE 22-1

Testing Primary Sensation

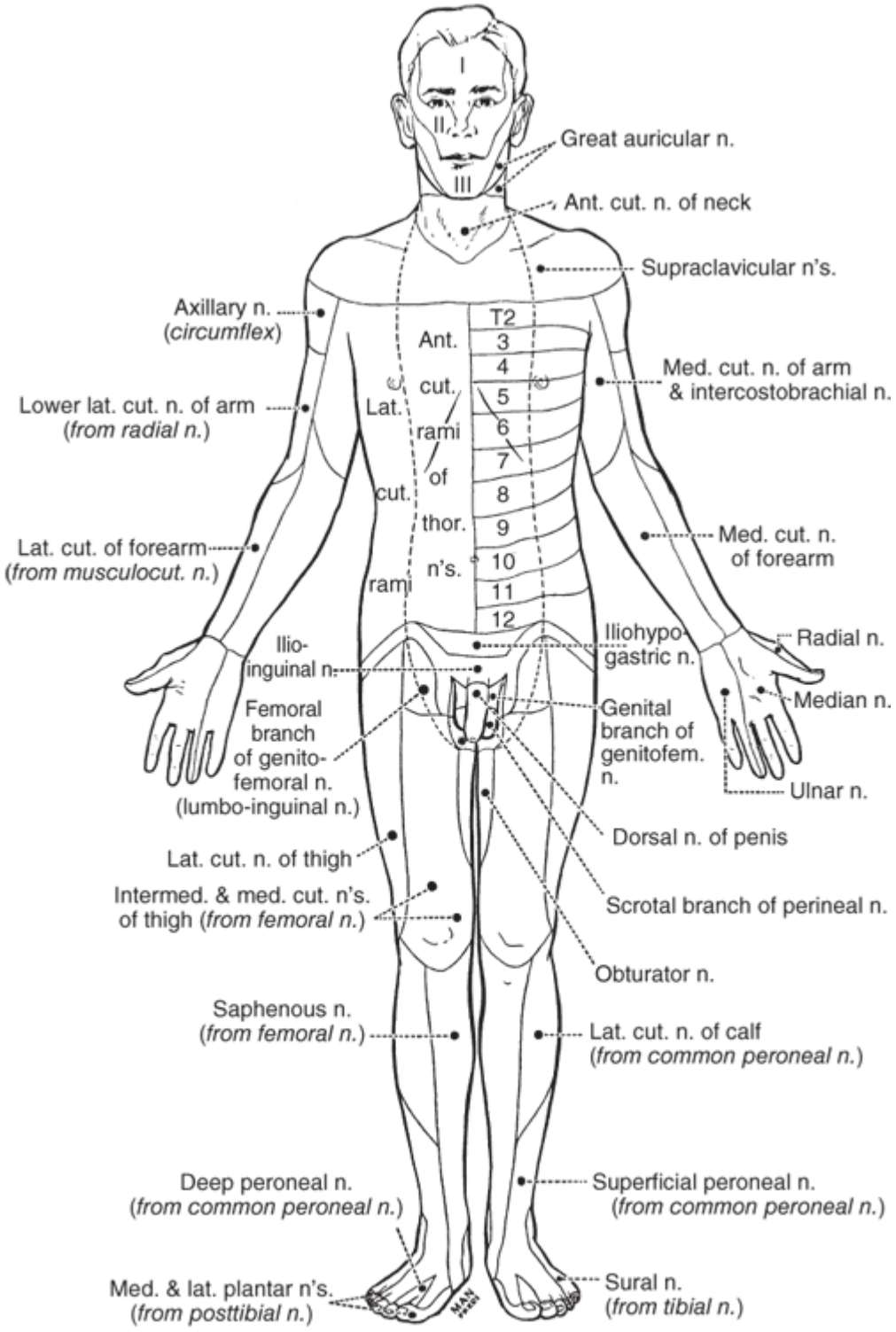
Sense	Test Device	Endings Activated	Fiber Size Mediating	Central Pathway
Pain	Pinprick	Cutaneous nociceptors	Small	SpTh, also D
Temperature, heat	Warm metal object	Cutaneous thermoreceptors for hot	Small	SpTh
Temperature, cold	Cold metal object	Cutaneous thermoreceptors for cold	Small	SpTh
Touch	Cotton wisp, fine brush	Cutaneous mechanoreceptors, also naked endings	Large and small	Lem, also D and SpTh
Vibration	Tuning fork, 128 Hz	Mechanoreceptors, especially pacinian corpuscles	Large	Lem, also D
Joint position	Passive movement of specific joints	Joint capsule and tendon endings, muscle spindles	Large	Lem, also D

Abbreviations: D, diffuse ascending projections in ipsilateral and contralateral anterolateral columns; Lem, posterior column and lemniscal projection, ipsilateral; SpTh, spinothalamic projection, contralateral.

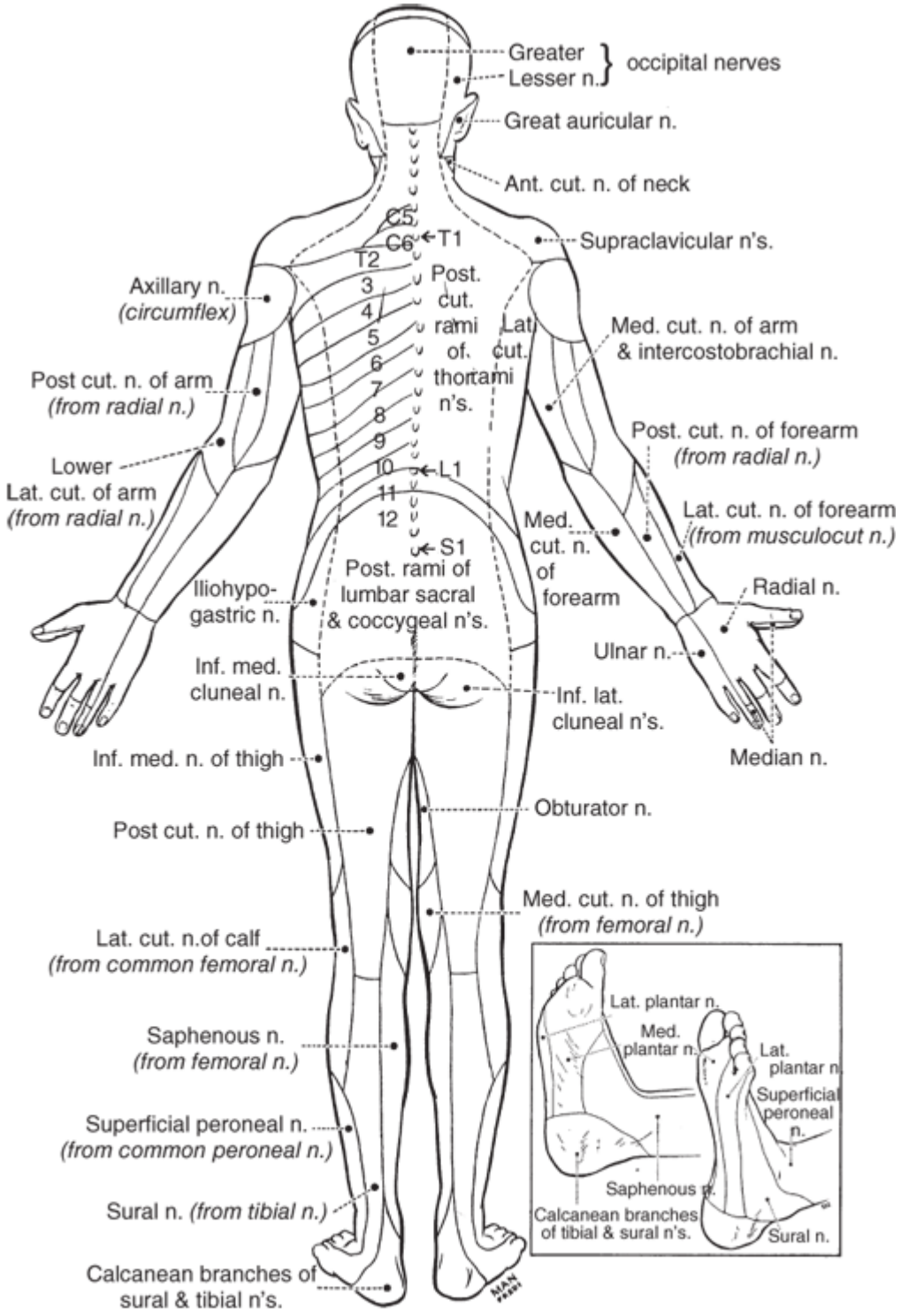
In patients with sensory complaints, testing should begin in the center of the affected region and proceed radially until sensation is perceived as normal. The distribution of any abnormality is defined and compared to root and peripheral nerve territories **(Figs. 22-2 and 22-3)**. Some patients present with sensory symptoms that do not fit an anatomic localization and are accompanied by either no abnormalities or gross inconsistencies on examination. The examiner should consider whether the sensory symptoms are a disguised request for help with psychologic or situational problems. Sensory examination of a patient who has no neurologic complaints can be brief and consist of pinprick, touch, and vibration testing in the hands and feet plus evaluation of stance and gait, including the Romberg maneuver **(Chap. V6)**. Evaluation of stance and gait also tests the integrity of motor and cerebellar systems.

FIGURE 22-2

The cutaneous fields of peripheral nerves. *(Reproduced by permission from W Haymaker, B Woodhall: Peripheral Nerve Injuries, 2nd ed. Philadelphia, Saunders, 1953.)*



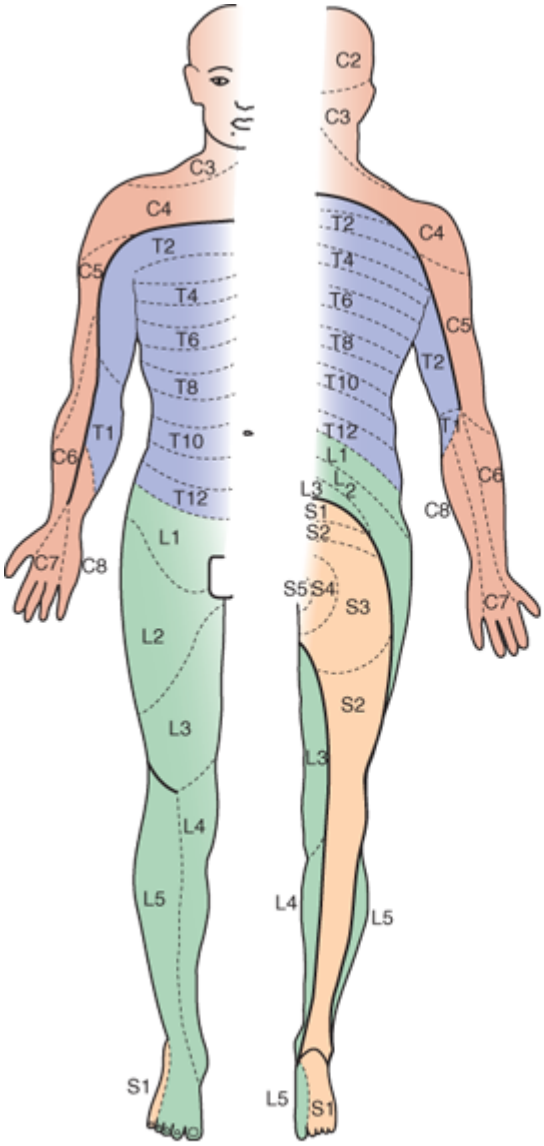
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 22-3

Distribution of the sensory spinal roots on the surface of the body (dermatomes). *(From D Sinclair: Mechanisms of Cutaneous Sensation. Oxford, UK, Oxford University Press, 1981; with permission from Dr. David Sinclair.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Primary Sensation

The sense of pain usually is tested with a clean pin, which is then discarded. The patient is asked to close the eyes and focus on the pricking or unpleasant quality of the stimulus, not just the pressure or touch sensation elicited. Areas of hypalgesia should be mapped by proceeding radially from the most hypalgesic site. Temperature sensation to both hot and cold is best tested with small containers filled with water of the desired temperature. An alternative way to test cold sensation is to touch a metal object, such as a tuning fork at room temperature, to the skin. For testing warm temperatures, the tuning fork or another metal object may be held under warm water of the desired temperature and then used. The appreciation of both cold and warmth should be tested because different receptors respond to each. Touch usually is tested with a wisp of cotton or a fine camel hair brush, minimizing pressure on the skin. In general, it is better to avoid testing touch on hairy skin because of the profusion of the sensory endings that surround each hair follicle. The patient is tested with the eyes closed and should indicate as soon as the stimulus is perceived, indicating its location.

Joint position testing is a measure of proprioception. With the patient's eyes closed, joint position is tested in the distal interphalangeal joint of the great toe and fingers. The digit is held by its sides, distal to the joint being tested, and moved passively while more proximal joints are stabilized—the patient indicates the change in position or direction of movement. If errors are made, more proximal joints are tested. A test of proximal joint position sense, primarily at the shoulder, is performed by asking the patient to bring the two index fingers together with arms extended and eyes closed. Normal individuals can do this accurately, with errors of 1 cm or less.

The sense of vibration is tested with an oscillating tuning fork that vibrates at 128 Hz. Vibration is tested over bony points, beginning distally; in the feet, it is tested over the dorsal surface of the distal phalanx of the big toes and at the malleoli of the ankles, and in the hands, it is tested dorsally at the distal phalanx of the fingers. If abnormalities are found, more proximal sites should be examined. Vibratory thresholds at the same site in the patient and the examiner may be compared for control purposes.

Quantitative Sensory Testing

Effective sensory testing devices are commercially available. Quantitative sensory testing is particularly useful for serial evaluation of cutaneous sensation in clinical trials. Threshold testing for touch and vibratory and thermal sensation is the most widely used application.

Cortical Sensation

The most commonly used tests of cortical function are two-point discrimination, touch localization, and bilateral simultaneous stimulation and tests for graphesthesia and stereognosis. Abnormalities of these sensory tests, in the presence of normal primary sensation in an alert cooperative patient, signify a lesion of the parietal cortex or thalamocortical projections. If primary sensation is altered, these cortical discriminative functions usually will be abnormal also. Comparisons should always be made between analogous sites on the two sides of the body because the deficit with a specific parietal lesion is likely to be unilateral.

Two-point discrimination is tested with special calipers, the points of which may be set from 2 mm to several centimeters apart and then applied simultaneously to the test site. On the fingertips, a normal individual can distinguish about a 3-mm separation of points.

Touch localization is performed by light pressure for an instant with the examiner's fingertip or a wisp of cotton wool; the patient, whose eyes are closed, is required to identify the site of touch. *Bilateral simultaneous stimulation* at analogous sites (e.g., the dorsum of both hands) can be carried out to determine whether the perception of touch is extinguished consistently on one side (*extinction* or *neglect*). *Graphesthesia* refers to the capacity to recognize, with eyes closed, letters or numbers drawn by the examiner's fingertip on the palm of the hand. Once again, interside comparison is of prime importance. Inability to recognize numbers or letters is termed *agraphesthesia*.

Stereognosis refers to the ability to identify common objects by palpation, recognizing their shape, texture, and size. Common standard objects such as keys, paper clips, and coins are best used. Patients with normal stereognosis should be able to distinguish a dime from a penny and a nickel from a quarter without looking. Patients should feel the object with only one hand at a time. If they are unable to identify it in one hand, it should be placed in the other for comparison. Individuals who are unable to identify common objects and coins in one hand but can do so in the other are said to have *astereognosis* of the abnormal hand.

LOCALIZATION OF SENSORY ABNORMALITIES

Sensory symptoms and signs can result from lesions at many different levels of the nervous system from the parietal cortex to the peripheral sensory receptor. Noting their distribution and nature is the most important way to localize their source. Their extent, configuration, symmetry, quality, and severity are the key observations.

Dysesthesias without sensory findings by examination may be difficult to interpret. To illustrate, tingling dysesthesias in an acral distribution (hands and feet) can be systemic in origin, for example, secondary to hyperventilation, or induced by a medication such as acetazolamide. Distal dysesthesias can also be an early event in an evolving polyneuropathy or may herald a myelopathy, such as from vitamin B_{12} deficiency.

Sometimes, distal dysesthesias have no definable basis. In contrast, dysesthesias that correspond in distribution to that of a particular peripheral nerve structure denote a lesion at that site. For instance, dysesthesias restricted to the fifth digit and the adjacent one-half of the fourth finger on one hand reliably point to disorder of the ulnar nerve, most commonly at the elbow.

Nerve and Root

In focal nerve trunk lesions, sensory abnormalities are readily mapped and generally have discrete boundaries (Figs. 22-2 and 22-3). Root (“radicular”) lesions frequently are accompanied by deep, aching pain along the course of the related nerve trunk. With compression of a fifth lumbar (L5) or first sacral (S1) root, as from a ruptured intervertebral disk, sciatica (radicular pain relating to the sciatic nerve trunk) is a common manifestation **(Chap. 14)**. With a lesion affecting a single root, sensory deficits may be minimal or absent because adjacent root territories overlap extensively.

Isolated mononeuropathies may cause symptoms beyond the territory supplied by the affected nerve, but abnormalities on examination typically are confined to appropriate anatomic boundaries. In multiple mononeuropathies, symptoms and signs occur in discrete territories supplied by different individual nerves and—as more nerves are affected—may simulate a polyneuropathy if deficits become confluent. With polyneuropathies, sensory deficits are generally graded, distal, and symmetric in distribution **(Chap. 438)**. Dysesthesias, followed by numbness, begin in the toes and ascend symmetrically. When dysesthesias reach the knees, they usually also have appeared in the fingertips. The process is nerve length–dependent, and the deficit is often described as “stocking-glove” in type. Involvement of both hands and feet also occurs with lesions of the upper cervical cord or the brainstem, but an upper level of the sensory disturbance may then

be found on the trunk and other evidence of a central lesion may be present, such as sphincter involvement or signs of an upper motor neuron lesion (Chap. 21). Although most polyneuropathies are pansensory and affect all modalities of sensation, selective sensory dysfunction according to nerve fiber size may occur. Small-fiber polyneuropathies are characterized by burning, painful dysesthesias with reduced pinprick and thermal sensation but with sparing of proprioception, motor function, and deep tendon reflexes. Touch is involved variably; when it is spared, the sensory pattern is referred to as exhibiting *sensory dissociation*. Sensory dissociation may occur also with spinal cord lesions. Large-fiber polyneuropathies are characterized by vibration and position sense deficits, imbalance, absent tendon reflexes, and variable motor dysfunction but preservation of most cutaneous sensation. Dysesthesias, if present at all, tend to be tingling or bandlike in quality.

Sensory neuronopathy (or ganglionopathy) is characterized by widespread but asymmetric sensory loss occurring in a non-length-dependent manner so that it may occur proximally or distally and in the arms, legs, or both. Pain and numbness progress to sensory ataxia and impairment of all sensory modalities with time. This condition is usually paraneoplastic or idiopathic in origin (Chaps. 90 and 438) or related to an autoimmune disease, particularly Sjögren's syndrome.

Spinal Cord

(See also Chap. 434) If the spinal cord is transected, all sensation is lost below the level of transection. Bladder and bowel function also are lost, as is motor function. Lateral hemisection of the spinal cord produces the Brown-Séquard syndrome, with absent pain and temperature sensation contralaterally and loss of proprioceptive sensation and power ipsilaterally below the lesion (see Figs. 22-1 and 434-1); ipsilateral pain or hyperesthesia may also occur.

Numbness or paresthesias in both feet may arise from a spinal cord lesion; this is especially likely when the upper level of the sensory loss extends to the trunk. When all extremities are affected, the lesion is probably in the cervical region or brainstem unless a peripheral neuropathy is responsible. The presence of upper motor neuron signs (Chap. 21) supports a central lesion; a hyperesthetic band on the trunk may suggest the level of involvement.

A dissociated sensory loss can reflect spinothalamic tract involvement in the spinal cord, especially if the deficit is unilateral and has an upper level on the torso. Bilateral spinothalamic tract involvement occurs with lesions affecting the center of the spinal cord, such as in syringomyelia. There is a dissociated sensory loss with impairment of pinprick and temperature appreciation but relative preservation of light touch, position sense, and vibration appreciation.

Dysfunction of the posterior columns in the spinal cord or of the posterior root entry zone may lead to a bandlike sensation around the trunk or a feeling of tight pressure in one or more limbs. Flexion of the neck sometimes leads to an electric shock–like sensation that radiates down the back and into the legs (Lhermitte's sign) in patients with a cervical lesion affecting the posterior columns, such as from multiple sclerosis, cervical spondylosis, or recent irradiation to the cervical region.

Brainstem

Crossed patterns of sensory disturbance, in which one side of the face and the opposite side of the body are affected, localize to the lateral medulla. Here a small lesion may damage both the ipsilateral descending trigeminal tract and the ascending spinothalamic fibers subserving the opposite arm, leg, and hemitorso (see “Lateral medullary syndrome” in **Fig. 419-7**). A lesion in the tegmentum of the pons and midbrain, where the lemniscal and spinothalamic tracts merge, causes pansensory loss contralaterally.

Thalamus

Hemisensory disturbance with tingling numbness from head to foot is often thalamic in origin but also can arise from the anterior parietal region. If abrupt in onset, the lesion is likely to be due to a small stroke (lacunar infarction), particularly if localized to the thalamus. Occasionally, with lesions affecting the VPL nucleus or adjacent white matter, a syndrome of thalamic pain, also called *Déjerine-Roussy syndrome*, may ensue. The persistent, unrelenting unilateral pain often is described in dramatic terms.

Cortex

With lesions of the parietal lobe involving either the cortex or the subjacent white matter, the most prominent symptoms are contralateral hemineglect, hemi-inattention, and a tendency not to use the affected hand and arm. On cortical sensory testing (e.g., two-point discrimination, graphesthesia), abnormalities are often found but primary sensation is usually intact. Anterior parietal infarction may present as a pseudothalamic syndrome with contralateral loss of primary sensation from head to toe. Dysesthesias or a sense of numbness and, rarely, a painful state may also occur.

Focal Sensory Seizures

These seizures generally are due to lesions in the area of the postcentral or precentral gyrus. The principal symptom of focal sensory seizures is tingling, but additional, more complex sensations may occur, such as a rushing feeling, a sense of warmth, or a sense of movement without detectable motion. Symptoms typically are unilateral; commonly begin in the arm or hand, face, or foot; and often spread in a manner that reflects the cortical representation of different bodily parts, as in a Jacksonian march. Their duration is variable; seizures may be transient, lasting only for seconds, or persist for an hour or more. Focal motor features may supervene, often becoming generalized with loss of consciousness and tonic-clonic jerking.

Psychogenic Symptoms

Sensory symptoms may have a psychogenic basis. Such symptoms may be generalized or have an anatomic boundary that is difficult to explain neurologically, for example, circumferentially at the groin or shoulder or around a specific joint. Pain is common, but the nature and intensity of any sensory disturbances are variable. The diagnosis should not be one of exclusion but based on suggestive findings that are otherwise difficult to explain, such as midline splitting of impaired vibration, pinprick, or light touch appreciation;

variability or poor reproducibility of sensory deficits; or normal performance of tasks requiring sensory input that is seemingly abnormal on formal testing, such as good performance with eyes closed of the finger-to-nose test despite an apparent loss of position sense in the upper limb. The side with abnormal sensation may be confused when the limbs are placed in an unusual position, such as crossed behind the back. Sensory complaints should not be regarded as psychogenic simply because they are unusual.

FURTHER READING

Brazis P, Masdeu JC, Biller J: *Localization in Clinical Neurology*, 7th ed. Philadelphia, Lippincott William & Wilkins, 2016.

Campbell WW: *DeJong's The Neurological Examination*, 7th ed. Philadelphia, Lippincott William & Wilkins, 2012.

Chapter 23: Gait Disorders, Imbalance, and Falls

Jessica M. Baker; Lewis R. Sudarsky

PREVALENCE, MORBIDITY, AND MORTALITY

Gait and balance problems are common in the elderly and contribute to the risk of falls and injury. Gait disorders have been described in 15% of individuals aged >65. By age 80 one person in four will use a mechanical aid to assist with ambulation. Among those aged ≥85, the prevalence of gait abnormality approaches 40%. In epidemiologic studies, gait disorders are consistently identified as a major risk factor for falls and injury.

ANATOMY AND PHYSIOLOGY

An upright bipedal gait depends on the successful integration of postural control and locomotion. These functions are widely distributed in the central nervous system. The biomechanics of bipedal walking are complex, and the performance is easily compromised by a neurologic deficit at any level. Command and control centers in the brainstem, cerebellum, and forebrain modify the action of spinal pattern generators to promote stepping. While a form of “fictive locomotion” can be elicited from quadrupedal animals after spinal transection, this capacity is limited in primates. Step generation in primates is dependent on locomotor centers in the pontine tegmentum, midbrain, and subthalamic region. Locomotor synergies are executed through the reticular formation and descending pathways in the ventromedial spinal cord. Cerebral control provides a goal and purpose for walking and is involved in avoidance of obstacles and adaptation of locomotor programs to context and terrain.

Postural control requires the maintenance of the center of mass over the base of support through the gait cycle. Unconscious postural adjustments maintain standing balance: long latency responses are measurable in the leg muscles, beginning 110 milliseconds after a perturbation. Forward motion of the center of mass provides propulsive force for stepping, but failure to maintain the center of mass within stability limits results in falls. The anatomic substrate for dynamic balance has not been well defined, but the vestibular nucleus and midline cerebellum contribute to balance control in animals. Patients with damage to these structures have impaired balance while standing and walking.

Standing balance depends on good-quality sensory information about the position of the body center with respect to the environment, support surface, and gravitational forces. Sensory information for postural control is primarily generated by the visual system, the vestibular system, and proprioceptive receptors in the

muscle spindles and joints. A healthy redundancy of sensory afferent information is generally available, but loss of two of the three pathways is sufficient to compromise standing balance. Balance disorders in older individuals sometimes result from multiple insults in the peripheral sensory systems (e.g., visual loss, vestibular deficit, peripheral neuropathy) that critically degrade the quality of afferent information needed for balance stability.

Older patients with cognitive impairment appear to be particularly prone to falls and injury. There is a growing body of literature on the use of attentional resources to manage gait and balance. Walking is generally considered to be unconscious and automatic, but the ability to walk while attending to a cognitive task (*dual-task walking*) may be compromised in the elderly. Older patients with deficits in executive function may have particular difficulty in managing the attentional resources needed for dynamic balance when distracted.

DISORDERS OF GAIT

Disorders of gait may be attributed to neurological and non-neurological causes, though significant overlap often exists. The *antalgic gait* results from avoidance of pain associated with weight-bearing and is commonly seen in osteoarthritis. Asymmetry is a common feature of gait disorders due to contractures and other orthopedic deformities. Impaired vision rounds out the list of common non-neurological causes of gait disorders.

Neurologic gait disorders are disabling and equally important to address. The heterogeneity of gait disorders observed in clinical practice reflects the large network of neural systems involved in the task. Walking is vulnerable to neurologic disease at every level. Gait disorders have been classified descriptively on the basis of abnormal physiology and biomechanics. One problem with this approach is that many failing gaits look fundamentally similar. This overlap reflects common patterns of adaptation to threatened balance stability and declining performance. *The gait disorder observed clinically must be viewed as the product of a neurologic deficit and a functional adaptation*. Unique features of the failing gait are often overwhelmed by the adaptive response. Some common patterns of abnormal gait are summarized next. Gait disorders can also be classified by etiology **(Table 23-1)**.

TABLE 23-1

Etiology of Gait Disorders

Etiology	No. of Cases	Percent
Sensory deficits	22	18.3
Myelopathy	20	16.7
Multiple infarcts	18	15.0
Parkinsonism	14	11.7
Cerebellar degeneration	8	6.7
Hydrocephalus	8	6.7
Toxic/metabolic causes	3	2.5
Psychogenic causes	4	3.3
Other	6	5.0
Unknown causes	17	14.2
Total	120	100

Source: Reproduced with permission from J Masdeu et al: *Gait Disorders of Aging*. Lippincott Raven, 1997.

CAUTIOUS GAIT

The term *cautious gait* is used to describe the patient who walks with an abbreviated stride, widened base and lowered center of mass, as if walking on a slippery surface. This disorder is both common and nonspecific. It is, in essence, an adaptation to a perceived postural threat. There may be an associated fear of falling. This disorder can be observed in more than one-third of older patients with gait impairment. Physical therapy often improves walking to the degree that follow-up observation may reveal a more specific underlying disorder.

STIFF-LEGGED GAIT

Spastic gait is characterized by stiffness in the legs, an imbalance of muscle tone, and a tendency to circumduct and scuff the feet. The disorder reflects compromise of corticospinal command and overactivity of spinal reflexes. The patient may walk on the toes. In extreme instances, the legs cross due to increased tone in the adductors ("scissoring" gait). Upper motor neuron signs are present on physical examination. The disorder may be cerebral or spinal in origin.

Myelopathy from cervical spondylosis is a common cause of spastic or spastic-ataxic gait in the elderly. Demyelinating disease and trauma are the leading causes of myelopathy in younger patients. In chronic progressive myelopathy of unknown cause, a workup with laboratory and imaging tests may establish a diagnosis. A structural lesion, such as a tumor or a spinal vascular malformation, should be excluded with appropriate testing. **Spinal cord disorders are discussed in detail in Chap. 434.**

With cerebral spasticity, asymmetry is common, the upper extremities are usually involved, and dysarthria is often an associated feature. Common causes include vascular disease (stroke), multiple sclerosis, motor neuron disease, and perinatal nervous system injury (cerebral palsy).

Other stiff-legged gaits include dystonia **(Chap. 428)** and stiff-person syndrome **(Chap. 90)**. Dystonia is a disorder characterized by sustained muscle contractions resulting in repetitive twisting movements and abnormal posture. It often has a genetic basis. Dystonic spasms can produce plantar flexion and inversion of the feet, sometimes with torsion of the trunk. In autoimmune stiff-person syndrome, exaggerated lordosis of the lumbar spine and overactivation of antagonist muscles restrict trunk and lower-limb movement and result in a wooden or fixed posture.

PARKINSONISM, FREEZING GAIT, AND OTHER MOVEMENT DISORDERSS

Parkinson's disease **(Chap. 427)** is common, affecting 1% of the population >55 years of age. The stooped posture and shuffling gait are characteristic and distinctive features. Patients sometimes accelerate (festinate) with walking, display retropulsion, or exhibit a tendency to turn en bloc. The step-to-step variability of the parkinsonian gait also contributes to fall risk. Dopamine replacement improves step length, arm swing, turning speed, and gait initiation. There is increasing evidence that deficits in cholinergic circuits in the pedunculopontine nucleus and cortex contribute to the gait disorder of Parkinson's disease. Cholinesterase inhibitors such as donepezil and rivastigmine have been shown in early studies to significantly decrease gait variability, instability, and fall frequency, even in the absence of cognitive impairment, perhaps through improvement in attention.

Freezing is defined as a brief, episodic absence of forward progression of the feet, despite the intention to walk. Freezing may be triggered by approaching a narrow doorway or crowd, may be overcome by visual cueing, and contributes to fall risk. Gait freezing is present in approximately one-quarter of Parkinson's patients within 5 years of onset, and its frequency increases further over time. In treated patients, an end-of-dose gait freezing is a common problem that may improve with more frequent administration of dopaminergic drugs, or with use of monoamine oxidase type B inhibitors such as rasagiline or selegiline **(Chap. 427)**.

Freezing of gait is also common in other neurodegenerative disorders associated with parkinsonism, including progressive supranuclear palsy (PSP), multiple-system atrophy, and corticobasal degeneration. Patients with these disorders frequently present with axial stiffness, postural instability, and a shuffling, freezing gait while lacking the characteristic pill-rolling tremor of Parkinson's disease. The gait of PSP is typically more erect compared with the stooped posture of typical Parkinson's disease, and falls within the first year also suggest the possibility of, PSP.

Hyperkinetic movement disorders also produce characteristic and recognizable disturbances in gait. In Huntington's disease **(Chap. 428)**, the unpredictable occurrence of choreic movements gives the gait a dancing quality. Tardive dyskinesia is the cause of many odd, stereotypic gait disorders seen in patients chronically exposed to antipsychotics and other drugs that block the D_2 dopamine receptor. *Orthostatic tremor* is a high frequency, low amplitude tremor predominantly involving the lower extremities. Patients often report shakiness or unsteadiness on standing, and improvement with sitting or walking. Falls are common. The tremor is often only appreciable by palpating the legs while standing.

FRONTAL GAIT DISORDER

Frontal gait disorder, also known as higher level gait disorder, is common in the elderly and has a variety of causes. The term is used to describe a shuffling, freezing gait with imbalance, and other signs of higher cerebral dysfunction. Typical features include a wide base of support, a short stride, shuffling along the floor, and difficulty with starts and turns. Many patients exhibit a difficulty with gait initiation that is descriptively characterized as the "slipping clutch" syndrome or gait ignition failure. The term *lower-body parkinsonism* is also used to describe such patients. Strength is generally preserved, and patients are able to make stepping movements when not standing and maintaining their balance at the same time. This disorder is best considered a higher-level motor control disorder, as opposed to an apraxia **(Chap. 26)**, though the term *gait apraxia* persists in the literature.

The most common cause of frontal gait disorder is vascular disease, particularly subcortical small-vessel disease in the deep frontal white matter and centrum ovale. Over three-quarters of patients with subcortical vascular dementia demonstrate gait abnormalities; decreased arm swing and a stooped posture are particularly prevalent features. The clinical syndrome also includes dysarthria, pseudobulbar affect (emotional disinhibition), increased tone, and hyperreflexia in the lower limbs.

Normal pressure (communicating) hydrocephalus (NPH) in adults also presents with a similar gait disorder. Other features of the diagnostic triad (mental changes, incontinence) may be absent in a substantial number of patients. MRI demonstrates ventricular enlargement, an enlarged flow void about the aqueduct, periventricular white-matter change, and high-convexity tightness (disproportionate widening of the sylvian fissures versus the cortical sulci). A lumbar puncture or dynamic test is necessary to confirm a diagnosis of NPH. Neurodegenerative dementias and mass lesions of the frontal lobes cause a similar clinical picture and can be differentiated from vascular disease and hydrocephalus by neuroimaging.

CEREBELLAR GAIT ATAXIA

Disorders of the cerebellum have a dramatic impact on gait and balance. Cerebellar gait ataxia is characterized by a wide base of support, lateral instability of the trunk, erratic foot placement, and decompensation of balance when attempting to walk on a narrow base. Difficulty maintaining balance when turning is often an early feature. Patients are unable to walk tandem heel to toe and display truncal sway in narrow-based or tandem stance. They show considerable variation in their tendency to fall in daily life.

Causes of cerebellar ataxia in older patients include stroke, trauma, tumor, and neurodegenerative disease such as multiple-system atrophy **(Chap. 432)** and various forms of hereditary cerebellar degeneration **(Chap. 431)**. A short expansion at the site of the fragile X mutation (*fragile X pre-mutation*) has been associated with gait ataxia in older men. Alcohol causes an acute and chronic cerebellar ataxia. In patients with ataxia due to cerebellar degeneration, MRI demonstrates the extent and topography of cerebellar atrophy.

SENSORY ATAXIA

As reviewed earlier in this chapter, balance depends on high-quality afferent information from the visual and the vestibular systems and proprioception. When this information is lost or degraded, balance during locomotion is impaired and instability results. The sensory ataxia of tabetic neurosyphilis is a classic example. The contemporary equivalent is the patient with neuropathy affecting large fibers. Vitamin B_{12} deficiency is a treatable cause of large-fiber sensory loss in the spinal cord and peripheral nervous system. Joint position and vibration sense are diminished in the lower limbs. The stance in such patients is destabilized by eye closure; they often look down at their feet when walking and do poorly in the dark. **Table 23-2** compares sensory ataxia with cerebellar ataxia and frontal gait disorder.

TABLE 23-2

Features of Cerebellar Ataxia, Sensory Ataxia, and Frontal Gait Disorders

Feature	Cerebellar Ataxia	Sensory Ataxia	Frontal Gait
Base of support	Wide-based	Narrow base, looks down	Wide-based
Velocity	Variable	Slow	Very slow
Stride	Irregular, lurching	Regular with path deviation	Short, shuffling
Romberg test	+/–	Unsteady, falls	+/–
Heel → shin	Abnormal	+/–	Normal
Initiation	Normal	Normal	Hesitant
Turns	Unsteady	+/–	Hesitant, multistep
Postural instability	+	+++	++++ Poor postural synergies rising from a chair
Falls	Late event	Frequent	Frequent

NEUROMUSCULAR DISEASE

Patients with neuromuscular disease often have an abnormal gait, occasionally as a presenting feature. With distal weakness (peripheral neuropathy), the step height is increased to compensate for foot drop, and the sole of the foot may slap on the floor during weight acceptance, termed the *steppage gait*. Patients with myopathy or muscular dystrophy more typically exhibit proximal weakness. Weakness of the hip girdle may result in some degree of excess pelvic sway during locomotion. The stooped posture of lumbar spinal stenosis ameliorates pain from the compression of the cauda equina occurring with a more upright posture while walking, and may mimic early parkinsonism.

TOXIC AND METABOLIC DISORDERS

Chronic toxicity from medications and metabolic disturbances can impair motor function and gait. Examination may reveal mental status changes, asterixis or myoclonus. Static equilibrium is disturbed, and such patients are easily thrown off balance. Disequilibrium is particularly evident in patients with chronic renal disease and those with hepatic failure, in whom asterixis may impair postural support. Sedative drugs, especially neuroleptics and long-acting benzodiazepines, affect postural control and increase the risk for falls. These disorders are especially important to recognize because they are often treatable.

FUNCTIONAL GAIT DISORDER

Functional disorders (formerly “psychogenic”) are common in neurologic practice, and the presentation often involves gait. The hallmark of a functional gait disorder is an internal inconsistency of deficits that may be incompatible with a neurological deficit. For example, odd gyrations of posture with wastage of muscular energy (astasia–abasia) appear superficially unsteady, yet in reality require significant postural control. Falls are rare, and there are often discrepancies between examination findings and the patient’s functional status. Extreme slow motion, an inappropriately overcautious gait, and dramatic fluctuations over time may improve with distraction, keeping in mind that numerous organic neurological diseases are also paroxysmal in nature. Preceding stress or trauma is variably present, and its absence no longer precludes the diagnosis of a functional neurological disorder. Functional gait disorders are among the most dramatic encountered, and should be differentiated from the slowness and psychomotor retardation seen in certain patients with major depression.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Slowly Progressive Disorder of Gait

When reviewing the history, it is helpful to inquire about the onset and progression of disability. Initial awareness of an unsteady gait often follows a fall. Stepwise evolution or sudden progression suggests vascular disease. Gait disorder may be associated with urinary urgency and incontinence, particularly in patients with cervical spine disease or hydrocephalus. It is always important to review the use of alcohol and medications that affect gait and balance. Information on localization derived from the neurologic examination can be helpful in narrowing the list of possible diagnoses.

Gait observation provides an immediate sense of the patient’s degree of disability. Arthritic and antalgic gaits are recognized by observation, though neurologic and orthopedic problems may coexist. Characteristic patterns of abnormality are sometimes seen, though, as stated previously, failing gaits often look fundamentally similar. Cadence (steps per minute), velocity, and stride length can be recorded by timing a patient over a fixed distance. Watching the patient rise from a chair provides a good functional assessment of balance.

Brain imaging studies may be informative in patients with an undiagnosed disorder of gait. MRI is sensitive for cerebral lesions of vascular or demyelinating disease and is a good screening test for occult

hydrocephalus. Patients with recurrent falls are at risk for subdural hematoma. As mentioned earlier, many elderly patients with gait and balance difficulty have white matter abnormalities in the periventricular region and centrum semiovale. While these lesions may be an incidental finding, a substantial burden of white matter disease will ultimately impact cerebral control of locomotion.

DISORDERS OF BALANCE

DEFINITION, ETIOLOGY, AND MANIFESTATIONS

Balance is the ability to maintain equilibrium—a dynamic state in which one's center of mass is controlled with respect to the lower extremities, gravity and the support surface despite external perturbations. The reflexes required to maintain upright posture require input from cerebellar, vestibular, and somatosensory systems; the premotor cortex, corticospinal and reticulospinal tracts mediate output to axial and proximal limb muscles. These responses are physiologically complex, and the anatomic representation they entail is not well understood. Failure can occur at any level and presents as difficulty maintaining posture while standing and walking.

The history and physical examination may differentiate underlying causes of imbalance. Patients with *cerebellar* ataxia do not generally complain of dizziness, though balance is visibly impaired. Neurologic examination reveals a variety of cerebellar signs. Postural compensation may prevent falls early on, but falls are inevitable with disease progression. The progression of neurodegenerative ataxia is often measured by the number of years to loss of stable ambulation.

Vestibular disorders (Chap. 19) have symptoms and signs that fall into three categories: (1) vertigo (the subjective inappropriate perception or illusion of movement); (2) nystagmus (involuntary eye movements); and (3) impaired standing balance. Not every patient has all manifestations. Patients with vestibular deficits related to ototoxic drugs may lack vertigo or obvious nystagmus, but their balance is impaired on standing and walking, and they cannot navigate in the dark. Laboratory testing is available to investigate vestibular deficits.

Somatosensory deficits also produce imbalance and falls. There is often a subjective sense of insecure balance and fear of falling. Postural control is compromised by eye closure (*Romberg's sign*); these patients also have difficulty navigating in the dark. A dramatic example is provided by the patient with autoimmune subacute sensory neuropathy, which is sometimes a paraneoplastic disorder (Chap. 90). Compensatory strategies enable such patients to walk in the virtual absence of proprioception, but the task requires active visual monitoring.

Patients with *higher-level disorders of equilibrium* have difficulty maintaining balance in daily life and may present with falls. Their awareness of balance impairment may be reduced. Patients taking sedating medications are in this category.

FALLS

Falls are common in the elderly; over one-third of people aged >65 who are living in the community fall each year. This number is even higher in nursing homes and hospitals. Elderly people are not only at higher risk for falls, but are more likely to suffer serious complications due to medical comorbidities such as osteoporosis. Hip fractures result in hospitalization, can lead to nursing home admission, and are associated with an increased mortality risk in the subsequent year. Falls may result in brain or spinal injury, the history of which may be difficult for the patient to provide. The proportion of spinal cord injuries due to falls in individuals aged >65 years has doubled in the last decade, perhaps due to increasing activity in this age group. Some falls result in a prolonged time lying on the ground; fractures and CNS injury are a particular concern in this context.

For each person who is physically disabled, there are others whose functional independence is limited by anxiety and fear of falling. Nearly one in five elderly individuals voluntarily restricts his or her activity because of fear of falling. With loss of ambulation, the quality of life diminishes, and rates of morbidity and mortality increase.

RISK FACTORS FOR FALLS

Risk factors for falls may be *intrinsic* (e.g., gait and balance disorders) or *extrinsic* (e.g., polypharmacy, and environmental factors); some risk factors are modifiable. The presence of multiple risk factors is associated with a substantially increased risk of falls. **(Table 23-3)** summarizes a meta-analysis of studies establishing the principal risk factors for falls. Polypharmacy (use of four or more prescription medications) has also been identified as an important risk factor.

TABLE 23-3

Meta-Analysis of Risk Factors for Falls in Older Persons

Risk Factor	Mean RR (OR)	Range
Muscle weakness	4.4	1.5–10.3
History of falls	3.0	1.7–7.0
Gait deficit	2.9	1.3–5.6
Balance deficit	2.9	1.6–5.4
Use assistive device	2.6	1.2–4.6
Visual deficit	2.5	1.6–3.5
Arthritis	2.4	1.9–2.9
Impaired ADL	2.3	1.5–3.1
Depression	2.2	1.7–2.5
Cognitive impairment	1.8	1.0–2.3
Age >80 years	1.7	1.1–2.5

Abbreviations: ADL, activity of daily living; OR, odds ratio from retrospective studies; RR, relative risk from prospective studies.

Source: Reproduced with permission from Guideline for the Prevention of Falls in Older Persons. J Am Geriatr Soc 49:664, 2001.

ASSESSMENT OF THE PATIENT WITH FALLS

The most productive approach is to identify the high-risk patient prospectively, before there is a serious injury. All community-dwelling adults should be asked about falls at least annually. The Timed Up and Go ("TUG") test involves timing a patient as they stand up from a chair, walk 10 ft, turn, then sit down. Patients with a history of falls, or those requiring >12 s to complete the TUG test, are high risk for falls and should undergo further assessment.

History

The history surrounding a fall is often problematic or incomplete, and the underlying mechanism or cause may be difficult to establish in retrospect. Patients should be queried about any provoking factors (including head turn, standing) or prodromal symptoms, such as dizziness, vertigo, pre-syncopal symptoms or focal weakness. A history of baseline mobility and medical comorbidities should be elicited. Patients at particular risk include those with mental status changes or dementia. Medications should be reviewed, with particular attention to neuroleptics, benzodiazepines, anti-depressants, anti-arrhythmics, and diuretics, all of which are associated with an increased risk of falls. It is equally important to distinguish *mechanical falls* (those caused by tripping or slipping) due to purely extrinsic or environmental factors from those in which a modifiable intrinsic factor contributes. *Recurrent falls* may indicate an underlying gait or balance disorder. Falls associated with loss of consciousness (syncope, seizure) may require appropriate cardiac or neurological evaluation and intervention **(Chaps. 18 and 418)**, though a patient's report of change in consciousness may be unreliable.

Physical Examination

Examination of the patient with falls should include a basic cardiac examination, including orthostatic blood pressure if indicated by history, and observation of any orthopedic abnormalities. Mental status is easily assessed while obtaining a history from the patient; the remainder of the neurological examination should include visual acuity, strength and sensation in the lower extremities, muscle tone, and cerebellar function, with particular attention to gait and balance as described earlier in this chapter.

Fall Patterns

The description of a fall event may provide further clues to the underlying etiology. While there is no standard nosology of falls, some common clinical patterns may emerge and provide a clue.

DROP ATTACKS AND COLLAPSING FALLS

Drop attacks and collapsing falls are associated with a sudden loss of postural tone. Patients may report that their legs just "gave out" underneath them, or that they "collapsed in a heap." Syncope or orthostatic hypotension may be a factor in some such falls. Neurological causes are relatively rare, but include atonic seizures, myoclonus and intermittent obstruction of the foramen of Monro by a colloid cyst of the third ventricle causing acute obstructive hydrocephalus. An emotional trigger suggests cataplexy. While collapsing falls are more common among older patients with vascular risk factors, drop attacks should not be confused with vertebrobasilar ischemic attacks.

TOPPLING FALLS

Some patients maintain tone in antigravity muscles but fall over like a tree trunk, as if postural defenses had disengaged. Causes include cerebellar pathology and lesions of the vestibular system. There may be a consistent direction to such falls. Toppling falls are an early feature of PSP, and a late feature of Parkinson's disease, once postural instability has developed. Thalamic lesions causing truncal instability (*thalamic astasia*) may also contribute to this type of fall.

FALLS DUE TO GAIT FREEZING

Freezing of gait is seen in Parkinson's disease and related disorders. The feet stick to the floor and the center of mass keeps moving, resulting in a disequilibrium from which the patient has difficulty recovering, resulting in a forward fall. Similarly, patients with Parkinson's disease and festinating gait may find their feet unable to keep up and may thus fall forward.

FALLS RELATED TO SENSORY LOSS

Patients with somatosensory, visual, or vestibular deficits are prone to falls. These patients have particular difficulty dealing with poor illumination or walking on uneven ground. They often report subjective imbalance, apprehension, and fear of falling. These patients may be especially responsive to a rehabilitation-based intervention.

FALLS RELATED TO WEAKNESS

Patients who lack strength in antigravity muscles have difficulty rising from a chair or maintaining their balance after a perturbation. These patients are often unable to get up after a fall and may have to remain on the floor for a prolonged period until help arrives. If due to deconditioning, this is often treatable. Resistance strength training can increase muscle mass and leg strength, even for people in their eighties and nineties.

TREATMENT

TREATMENT

Interventions to Reduce the Risk of Falls and Injury

Efforts should be made to define the etiology of the gait disorder and the mechanism underlying the falls by a given patient. Orthostatic changes in blood pressure and pulse should be recorded. Rising from a chair and walking should be evaluated for safety. Specific treatment may be possible once a diagnosis is established. Therapeutic intervention is often recommended for older patients at substantial risk for falls, even if no neurologic disease is identified. A home visit to look for environmental hazards can be helpful. A variety of modifications may be recommended to improve safety, including improved lighting and the installation of grab bars and nonslip surfaces.

Rehabilitative interventions aim to improve muscle strength and balance stability and to make the patient more resistant to injury. High-intensity resistance strength training with weights and machines is useful to improve muscle mass, even in frail older patients. Improvements realized in posture and gait should translate to reduced risk of falls and injury. Sensory balance training is another approach to improving balance stability. Measurable gains can be made in a few weeks of training, and benefits can be maintained over 6 months by a 10- to 20-min home exercise program. This strategy is particularly successful in patients with vestibular and somatosensory balance disorders. A Tai Chi exercise program has been demonstrated to reduce the risk of falls and injury in patients with Parkinson's disease.

FURTHER READING

American Geriatrics Society, British Geriatrics Society, American Academy of Orthopedic Surgeons Panel on Falls Prevention: Guideline for the Prevention of Falls in Older Persons. J Am Geriatr Soc 49:664, 2001.

[PubMed: 11380764]

Nutt JG: Classification of Gait and Balance Disorders. Adv Neurol 87:135, 2001. [PubMed: 11347216]

Pirker W, Katzenschlager R: Gait disorders in adults and the elderly. Wien Klin Wochenschr 129:81, 2017. [PubMed: 27770207]

Chapter 24: Confusion and Delirium

S. Andrew Josephson; Bruce L. Miller

INTRODUCTION

Confusion, a mental and behavioral state of reduced comprehension, coherence, and capacity to reason, is one of the most common problems encountered in medicine, accounting for a large number of emergency department visits, hospital admissions, and inpatient consultations. *Delirium*, a term used to describe an acute confusional state, remains a major cause of morbidity and mortality, costing billions of dollars yearly in health care costs in the United States alone. Despite increased efforts targeting awareness of this condition, delirium often goes unrecognized in the face of evidence that it is usually the cognitive manifestation of serious underlying medical or neurologic illness.

CLINICAL FEATURES OF DELIRIUM

A multitude of terms are used to describe patients with delirium, including *encephalopathy, acute brain failure, acute confusional state*, and *postoperative* or *intensive care unit (ICU) psychosis*. Delirium has many clinical manifestations, but it is defined as a relatively acute decline in cognition that fluctuates over hours or days. The hallmark of delirium is a deficit of attention, although all cognitive domains—including memory, executive function, visuospatial tasks, and language—are variably involved. Associated symptoms that may be present in some cases include altered sleep-wake cycles, perceptual disturbances such as hallucinations or delusions, affect changes, and autonomic findings that include heart rate and blood pressure instability.

Delirium is a clinical diagnosis that is made only at the bedside. Two subtypes have been described—hyperactive and hypoactive—based on differential psychomotor features. The cognitive syndrome associated with severe alcohol withdrawal (i.e., “delirium tremens”) remains the classic example of the hyperactive subtype, featuring prominent hallucinations, agitation, and hyperarousal, often accompanied by life-threatening autonomic instability. In striking contrast is the hypoactive subtype, exemplified by benzodiazepine intoxication, in which patients are withdrawn and quiet, with prominent apathy and psychomotor slowing.

This dichotomy between subtypes of delirium is a useful construct, but patients often fall somewhere along a spectrum between the hyperactive and hypoactive extremes, sometimes fluctuating from one to the other. Therefore, clinicians must recognize this broad range of presentations of delirium to identify all patients with this potentially reversible cognitive disturbance. Hyperactive patients are often easily recognized by their

characteristic severe agitation, tremor, hallucinations, and autonomic instability. Patients who are quietly hypoactive are more often overlooked on the medical wards and in the ICU.

The reversibility of delirium is emphasized because many etiologies, such as infection and medication effects, can be treated easily. The long-term cognitive consequences of delirium remain largely unknown. Some episodes of delirium continue for weeks, months, or even years. The persistence of delirium in some patients and its high recurrence rate may be due to inadequate initial treatment of the underlying etiology. In other instances, delirium appears to cause permanent neuronal damage and cognitive decline; therefore prevention strategies are important to implement. Even if an episode of delirium completely resolves, there may be lingering effects of the disorder; a patient's recall of events after delirium varies widely, ranging from complete amnesia to repeated re-experiencing of the frightening period of confusion, similar to what is seen in patients with posttraumatic stress disorder.

RISK FACTORS

An effective primary prevention strategy for delirium begins with identification of high-risk patients, including those preparing for elective surgery or being admitted to the hospital. Multiple validated scoring systems have been developed as a screen for asymptomatic patients, many of which emphasize well-established risk factors for delirium.

The two most consistently identified risk factors are older age and baseline cognitive dysfunction. Individuals who are aged >65 or exhibit low scores on standardized tests of cognition develop delirium upon hospitalization at a rate approaching 50%. Whether age and baseline cognitive dysfunction are truly independent risk factors is uncertain. Other predisposing factors include sensory deprivation, such as preexisting hearing and visual impairment, as well as indices for poor overall health, including baseline immobility, malnutrition, and underlying medical or neurologic illness.

In-hospital risks for delirium include the use of bladder catheterization, physical restraints, sleep and sensory deprivation, and the addition of three or more new medications. Avoiding such risks remains a key component of delirium prevention as well as treatment. Surgical and anesthetic risk factors for the development of postoperative delirium include procedures such as those involving cardiopulmonary bypass, inadequate or excessive treatment of pain in the immediate postoperative period, and perhaps specific agents such as inhalational anesthetics.

The relationship between delirium and dementia (Chap. 25) is complicated by significant overlap between the two conditions, and it is not always simple to distinguish between them. Dementia and preexisting cognitive dysfunction serve as major risk factors for delirium, and at least two-thirds of cases of delirium occur in patients with coexisting underlying dementia. A form of dementia with parkinsonism, *dementia with Lewy bodies*, is characterized by a fluctuating course, prominent visual hallucinations, parkinsonism, and an attentional deficit that clinically resembles hyperactive delirium; patients with this condition are particularly vulnerable to delirium. Delirium in the elderly often reflects an insult to a brain that is vulnerable due to an underlying neurodegenerative condition. Therefore, the development of delirium sometimes heralds the

onset of a previously unrecognized brain disorder, and after the acute delirious episode has cleared, careful screening for an underlying condition should occur in the outpatient setting.

EPIDEMIOLOGY

Delirium is common, but its reported incidence has varied widely with the criteria used to define this disorder. Estimates of delirium in hospitalized patients range from 10 to >50%, with higher rates reported for elderly patients and patients undergoing hip surgery. Older patients in the ICU have especially high rates of delirium that approach 75%. The condition is not recognized in up to one-third of delirious inpatients, and the diagnosis is especially problematic in the ICU environment, where cognitive dysfunction is often difficult to appreciate in the setting of serious systemic illness and sedation. Delirium in the ICU should be viewed as an important manifestation of organ dysfunction not unlike liver, kidney, or heart failure. Outside the acute hospital setting, delirium occurs in nearly one-quarter of patients in nursing homes and in 50–80% of those at the end of life. These estimates emphasize the remarkably high frequency of this cognitive syndrome in older patients, a population that continues to grow.

An episode of delirium was previously viewed as a transient condition that carried a benign prognosis. It is now recognized as a disorder with substantial morbidity and mortality, and that often represents the first manifestation of a serious underlying illness. Estimates of in-hospital mortality rates among delirious patients range from 25 to 33%, similar to mortality rates due to sepsis. Patients with an in-hospital episode of delirium have a fivefold higher mortality rate in the months after their illness compared with age-matched nondelirious hospitalized patients. Delirious hospitalized patients also have a longer length of stay, are more likely to be discharged to a nursing home, and are more likely to experience subsequent episodes of delirium and cognitive decline; as a result, this condition has an enormous economic cost.

PATHOGENESIS

The pathogenesis and anatomy of delirium are incompletely understood. The attentional deficit that serves as the neuropsychological hallmark of delirium has a diffuse localization within the brainstem, thalamus, prefrontal cortex, and parietal lobes. Rarely, focal lesions such as ischemic strokes have led to delirium in otherwise healthy persons; right parietal and medial dorsal thalamic lesions have been reported most commonly, pointing to the importance of these areas in delirium pathogenesis. In most cases, however, delirium results from widespread disturbances in cortical and subcortical regions of the brain. Electroencephalogram (EEG) usually reveals symmetric slowing, a nonspecific finding that supports diffuse cerebral dysfunction.

Multiple neurotransmitter abnormalities, proinflammatory factors, and specific genes likely play a role in the pathogenesis of delirium. Deficiency of acetylcholine may play a key role, and medications with anticholinergic properties can commonly precipitate delirium. As noted above, patients with preexisting dementia are particularly susceptible to episodes of delirium. Alzheimer’s disease, dementia with Lewy bodies, and Parkinson’s disease dementia are all associated with cholinergic deficiency due to degeneration

of acetylcholine-producing neurons in the basal forebrain. In addition, other neurotransmitters are also likely to be involved in this diffuse cerebral disorder. For example, increases in dopamine can lead to delirium, and patients with Parkinson's disease treated with dopaminergic medications can develop a delirium-like state that features visual hallucinations, fluctuations, and confusion.

Not all individuals exposed to the same insult will develop signs of delirium. A low dose of an anticholinergic medication may have no cognitive effects on a healthy young adult but produce a florid delirium in an elderly person with known underlying dementia, although even healthy young persons develop delirium with very high doses of anticholinergic medications. This concept of delirium developing as the result of an insult in predisposed individuals is currently the most widely accepted pathogenic construct. Therefore, if a previously healthy individual with no known history of cognitive illness develops delirium in the setting of a relatively minor insult such as elective surgery or hospitalization, an unrecognized underlying neurologic illness such as a neurodegenerative disease, multiple previous strokes, or another diffuse cerebral cause should be considered. In this context, delirium can be viewed as a "stress test for the brain" whereby exposure to known inciting factors such as systemic infection and offending drugs can unmask a decreased cerebral reserve and herald a serious underlying and potentially treatable illness.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Delirium

Because the diagnosis of delirium is clinical and is made at the bedside, a careful history and physical examination are necessary in evaluating patients with possible confusional states. Screening tools can aid physicians and nurses in identifying patients with delirium, including the Confusion Assessment Method (CAM); the Nursing Delirium Screening Scale (NuDESC); the Organic Brain Syndrome Scale; the Delirium Rating Scale; and, in the ICU, the ICU version of the CAM and the Delirium Detection Score. Using the well-validated CAM, a diagnosis of delirium is made if there is (1) an acute onset and fluctuating course and (2) inattention accompanied by either (3) disorganized thinking or (4) an altered level of consciousness (Table 24-1). These scales may not identify the full spectrum of patients with delirium, and all patients who are acutely confused should be presumed delirious regardless of their presentation due to the wide variety of possible clinical features. A course that fluctuates over hours or days and may worsen at night (termed *sundowning*) is typical but not essential for the diagnosis. Observation will usually reveal an altered level of consciousness or a deficit of attention. Other features that are sometimes present include alteration of sleep-wake cycles, thought disturbances such as hallucinations or delusions, autonomic instability, and changes in affect.

HISTORY

It may be difficult to elicit an accurate history in delirious patients who have altered levels of consciousness or impaired attention. Information from a collateral source such as a spouse or another family member is

therefore invaluable. The three most important pieces of history are the patient's baseline cognitive function, the time course of the present illness, and current medications.

Premorbid cognitive function can be assessed through the collateral source or, if needed, via a review of outpatient records. Delirium by definition represents a change that is relatively acute and usually developing over hours to days, from a cognitive baseline. An acute confusional state is nearly impossible to diagnose without some knowledge of baseline cognitive function. Without this information, many patients with dementia or longstanding depression may be mistaken as delirious during a single initial evaluation. Patients with a more hypoactive, apathetic presentation with psychomotor slowing may be identified as being different from baseline only through conversations with family members. A number of validated instruments have been shown to diagnose cognitive dysfunction accurately using a collateral source, including the modified Blessed Dementia Rating Scale and the Clinical Dementia Rating (CDR). Baseline cognitive impairment is common in patients with delirium. Even when no such history of cognitive impairment is elicited, there should still be a high suspicion for a previously unrecognized underlying neurologic disorder.

Establishing the time course of cognitive change is important not only to make a diagnosis of delirium but also to correlate the onset of the illness with potentially treatable etiologies such as recent medication changes or symptoms of systemic infection.

Medications remain a common cause of delirium, especially compounds with anticholinergic or sedative properties. It is estimated that nearly one-third of all cases of delirium are secondary to medications, especially in the elderly. Medication histories should include all prescription as well as over-the-counter and herbal substances taken by the patient and any recent changes in dosing or formulation, including substitution of generics for brand-name medications.

Other important elements of the history include screening for symptoms of organ failure or systemic infection, which often contributes to delirium in the elderly. A history of illicit drug use, alcoholism, or toxin exposure is common in younger delirious patients. Finally, asking the patient and collateral source about other symptoms that may accompany delirium, such as depression, may help identify potential therapeutic targets.

PHYSICAL EXAMINATION

The general physical examination in a delirious patient should include careful screening for signs of infection such as fever, tachypnea, pulmonary consolidation, heart murmur, and meningismus. The patient's fluid status should be assessed; both dehydration and fluid overload with resultant hypoxemia have been associated with delirium, and each is usually easily rectified. The appearance of the skin can be helpful, showing jaundice in hepatic encephalopathy, cyanosis in hypoxemia, or needle tracks in patients using intravenous drugs.

The neurologic examination requires a careful assessment of mental status. Patients with delirium often present with a fluctuating course; therefore, the diagnosis can be missed when one relies on a single time

point of evaluation. For patients who worsen in the evening (sundowning), assessment only during morning rounds may be falsely reassuring.

An altered level of consciousness ranging from hyperarousal to lethargy to coma is present in most patients with delirium and can be assessed easily at the bedside. In a patient with a relatively normal level of consciousness, a screen for an attentional deficit is in order, because this deficit is the classic neuropsychological hallmark of delirium. Attention can be assessed while taking a history from the patient. Tangential speech, a fragmentary flow of ideas, or inability to follow complex commands often signifies an attentional problem. There are formal neuropsychological tests to assess attention, but a simple bedside test of digit span forward is quick and fairly sensitive. In this task, patients are asked to repeat successively longer random strings of digits beginning with two digits in a row, said to the patient at one per second intervals. Healthy adults can repeat a string of five to seven digits before faltering; a digit span of four or less usually indicates an attentional deficit unless hearing or language barriers are present, and many patients with delirium have digit spans of three or fewer digits.

More formal neuropsychological testing can be helpful in assessing a delirious patient, but it is usually too cumbersome and time-consuming in the inpatient setting. A Mini-Mental State Examination (MMSE) provides information regarding orientation, language, and visuospatial skills **(Chap. 25)**; however, performance of many tasks on the MMSE, including the spelling of “world” backward and serial subtraction of digits, will be impaired by delirious patients’ attentional deficits, rendering the test unreliable.

The remainder of the screening neurologic examination should focus on identifying new focal neurologic deficits. Focal strokes or mass lesions in isolation are rarely the cause of delirium, but patients with underlying extensive cerebrovascular disease or neurodegenerative conditions may not be able to cognitively tolerate even relatively small new insults. Patients should be screened for other signs of neurodegenerative conditions such as parkinsonism, which is seen not only in idiopathic Parkinson’s disease but also in other dementing conditions including Alzheimer’s disease, dementia with Lewy bodies, and progressive supranuclear palsy. The presence of multifocal myoclonus or asterixis on the motor examination is nonspecific but usually indicates a metabolic or toxic etiology of the delirium.

ETIOLOGY

Some etiologies can be easily discerned through a careful history and physical examination, whereas others require confirmation with laboratory studies, imaging, or other ancillary tests. A large, diverse group of insults can lead to delirium, and the cause in many patients is multifactorial. Common etiologies are listed in **Table 24-2**.

Prescribed, over-the-counter, and herbal medications all can precipitate delirium. Drugs with anticholinergic properties, narcotics, and benzodiazepines are particularly common offenders, but nearly any compound can lead to cognitive dysfunction in a predisposed patient. Whereas an elderly patient with baseline dementia may become delirious upon exposure to a relatively low dose of a medication, in less susceptible individuals delirium occurs only with very high doses of the same medication. This observation emphasizes

the importance of correlating the timing of recent medication changes, including dose and formulation, with the onset of cognitive dysfunction.

In younger patients, illicit drugs and toxins are common causes of delirium. In addition to more classic drugs of abuse, the recent rise in availability of “bath salts,” synthetic cannabis, methylenedioxymethamphetamine (MDMA, ecstasy), γ-hydroxybutyrate (GHB), and the phencyclidine (PCP)-like agent ketamine has led to an increase in delirious young persons presenting to acute care settings **(Chap. 447)**. Many common prescription drugs such as oral narcotics and benzodiazepines are often abused and readily available on the street. Alcohol abuse leading to high serum levels causes confusion, but more commonly, it is withdrawal from alcohol that leads to a hyperactive delirium **(Chap. 445)**. Alcohol and benzodiazepine withdrawal should be considered in all cases of delirium because even patients who drink only a few servings of alcohol every day can experience relatively severe withdrawal symptoms upon hospitalization.

Metabolic abnormalities such as electrolyte disturbances of sodium, calcium, magnesium, or glucose can cause delirium, and mild derangements can lead to substantial cognitive disturbances in susceptible individuals. Other common metabolic etiologies include liver and renal failure, hypercarbia and hypoxemia, vitamin deficiencies of thiamine and B_{12}, autoimmune disorders including central nervous system (CNS) vasculitis, and endocrinopathies such as thyroid and adrenal disorders.

Systemic infections often cause delirium, especially in the elderly. A common scenario involves the development of an acute cognitive decline in the setting of a urinary tract infection in a patient with baseline dementia. Pneumonia, skin infections such as cellulitis, and frank sepsis also lead to delirium. This so-called septic encephalopathy, often seen in the ICU, is probably due to the release of proinflammatory cytokines and their diffuse cerebral effects. CNS infections such as meningitis, encephalitis, and abscess are less common etiologies of delirium as are cases of autoimmune or paraneoplastic encephalitis; however, in light of the high morbidity and mortality rates associated with these conditions when they are not treated, clinicians must always maintain a high index of suspicion.

In some susceptible individuals, exposure to the unfamiliar environment of a hospital itself can lead to delirium. This etiology usually occurs as part of a multifactorial delirium and should be considered a diagnosis of exclusion after all other causes have been thoroughly investigated. Many primary prevention and treatment strategies for delirium involve relatively simple methods to address the aspects of the inpatient setting that are most confusing.

Cerebrovascular etiologies of delirium are usually due to global hypoperfusion in the setting of systemic hypotension from heart failure, septic shock, dehydration, or anemia. Focal strokes in the right parietal lobe and medial dorsal thalamus rarely can lead to a delirious state. A more common scenario involves a new focal stroke or hemorrhage causing confusion in a patient who has decreased cerebral reserve. In these individuals, it is sometimes difficult to distinguish between cognitive dysfunction resulting from the new neurovascular insult itself and delirium due to the infectious, metabolic, and pharmacologic complications that can accompany hospitalization after stroke.

Because a fluctuating course often is seen in delirium, intermittent seizures may be overlooked when one is considering potential etiologies. Both nonconvulsive status epilepticus and recurrent focal or generalized seizures followed by postictal confusion can cause delirium; EEG remains essential for this diagnosis and should be considered whenever the etiology of delirium remains unclear following initial workup. Seizure activity spreading from an electrical focus in a mass or infarct can explain global cognitive dysfunction caused by relatively small lesions.

It is extremely common for patients to experience delirium at the end of life in palliative care settings. This condition, sometimes described as *terminal restlessness*, must be identified and treated aggressively because it is an important cause of patient and family discomfort at the end of life. It should be remembered that these patients also may be suffering from more common etiologies of delirium such as systemic infection.

LABORATORY AND DIAGNOSTIC EVALUATION

A cost-effective approach allows the history and physical examination to guide further tests. No single algorithm will fit all delirious patients due to the staggering number of potential etiologies, but one stepwise approach is detailed in Table 24-3. If a clear precipitant such as an offending medication is identified, further testing may not be required. If, however, no likely etiology is uncovered with initial evaluation, an aggressive search for an underlying cause should be initiated.

Basic screening labs, including a complete blood count, electrolyte panel, and tests of liver and renal function, should be obtained in all patients with delirium. In elderly patients, screening for systemic infection, including chest radiography, urinalysis and culture, and possibly blood cultures, is important. In younger individuals, serum and urine drug and toxicology screening may be appropriate earlier in the workup. Additional laboratory tests addressing other autoimmune, endocrinologic, metabolic, and infectious etiologies should be reserved for patients in whom the diagnosis remains unclear after initial testing.

Multiple studies have demonstrated that brain imaging in patients with delirium is often unhelpful. If, however, the initial workup is unrevealing, most clinicians quickly move toward imaging of the brain to exclude structural causes. A noncontrast computed tomography (CT) scan can identify large masses and hemorrhages but is otherwise unlikely to help determine an etiology of delirium. The ability of magnetic resonance imaging (MRI) to identify most acute ischemic strokes as well as to provide neuroanatomic detail that gives clues to possible infectious, inflammatory, neurodegenerative, and neoplastic conditions makes it the test of choice. Because MRI techniques are limited by availability, speed of imaging, patient's cooperation, and contraindications, many clinicians begin with CT scanning and proceed to MRI if the etiology of delirium remains elusive.

Lumbar puncture (LP) must be obtained immediately after neuroimaging for all patients in whom CNS infection is suspected. Spinal fluid examination can also be useful in identifying inflammatory and neoplastic conditions. As a result, LP should be considered in any delirious patient with a negative workup. EEG remains invaluable if seizures are considered or if there is no cause readily identified.

TABLE 24-1

The Confusion Assessment Method (CAM) Diagnostic Algorithm[a]

The diagnosis of delirium requires the presence of features 1 and 2 **and** *either* feature 3 or 4.
Feature 1. Acute Onset and Fluctuating Course
This feature is satisfied by positive responses to the following questions: Is there evidence of an acute change in mental status from the patient's baseline? Did the (abnormal) behavior fluctuate during the day, that is, tend to come and go, or did it increase and decrease in severity?
Feature 2. Inattention
This feature is satisfied by a positive response to the following question: Did the patient have difficulty focusing attention, for example, being easily distractible, or have difficulty keeping track of what was being said?
Feature 3. Disorganized Thinking
This feature is satisfied by a positive response to the following question: Was the patient's thinking disorganized or incoherent, such as rambling or irrelevant conversation, unclear or illogical flow of ideas, or unpredictable switching from subject to subject?
Feature 4. Altered Level of Consciousness
This feature is satisfied by any answer other than "alert" to the following question: Overall, how would you rate the patient's level of consciousness: alert (normal), vigilant (hyperalert), lethargic (drowsy, easily aroused), stupor (difficult to arouse), or coma (unarousable)?

[a]Information is usually obtained from a reliable reporter, such as a family member, caregiver, or nurse.

Source: Modified from SK Inouye et al: Clarifying confusion: The Confusion Assessment Method. A new method for detection of delirium. Ann Intern Med 113:941, 1990.

TABLE 24-2

Common Etiologies of Delirium

Toxins
Prescription medications: especially those with anticholinergic properties, narcotics, and benzodiazepines
Drugs of abuse: alcohol intoxication and alcohol withdrawal, opiates, ecstasy, LSD, GHB, PCP, ketamine, cocaine, “bath salts,” marijuana and its synthetic forms
Poisons: inhalants, carbon monoxide, ethylene glycol, pesticides
Metabolic Conditions
Electrolyte disturbances: hypoglycemia, hyperglycemia, hyponatremia, hypernatremia, hypercalcemia, hypocalcemia, hypomagnesemia
Hypothermia and hyperthermia
Pulmonary failure: hypoxemia and hypercarbia
Liver failure/hepatic encephalopathy
Renal failure/uremia
Cardiac failure
Vitamin deficiencies: B_{12}, thiamine, folate, niacin
Dehydration and malnutrition
Anemia
Infections
Systemic infections: urinary tract infections, pneumonia, skin and soft tissue infections, sepsis
CNS infections: meningitis, encephalitis, brain abscess
Endocrine Conditions
Hyperthyroidism, hypothyroidism

Hyperparathyroidism
Adrenal insufficiency
Cerebrovascular Disorders
Global hypoperfusion states
Hypertensive encephalopathy
Focal ischemic strokes and hemorrhages (rare): especially nondominant parietal and thalamic lesions
Autoimmune Disorders
CNS vasculitis
Cerebral lupus
Neurologic paraneoplastic and autoimmune encephalitis
Seizure-Related Disorders
Nonconvulsive status epilepticus
Intermittent seizures with prolonged postictal states
Neoplastic Disorders
Diffuse metastases to the brain
Gliomatosis cerebri
Carcinomatous meningitis
CNS lymphoma
Hospitalization
Terminal end-of-life delirium

Abbreviations: CNS, central nervous system; GHB, γ-hydroxybutyrate; LSD, lysergic acid diethylamide; PCP, phencyclidine.

TABLE 24-3

Stepwise Evaluation of a Patient with Delirium

Initial Evaluation
History with special attention to medications (including over-the-counter and herbals)
General physical examination and neurologic examination
Complete blood count
Electrolyte panel including calcium, magnesium, phosphorus
Liver function tests, including albumin
Renal function tests
First-tier Further Evaluation Guided by Initial Evaluation
Systemic infection screen
Urinalysis and culture
Chest radiograph
Blood cultures
Electrocardiogram
Arterial blood gas
Serum and/or urine toxicology screen (perform earlier in young persons)
Brain imaging with MRI with diffusion and gadolinium (preferred) or CT
Suspected CNS infection or other inflammatory disorder: lumbar puncture after brain imaging
Suspected seizure-related etiology: electroencephalogram (EEG) (if high suspicion, should be performed immediately)
Second-tier Further Evaluation

Vitamin levels: B_{12}, folate, thiamine
Endocrinologic laboratories: thyroid-stimulating hormone (TSH) and free T_4; cortisol
Serum ammonia
Sedimentation rate
Autoimmune serologies: antinuclear antibodies (ANA), complement levels; p-ANCA, c-ANCA, consider paraneoplastic/autoimmune encephalitis serologies
Infectious serologies: rapid plasmin reagin (RPR); fungal and viral serologies if high suspicion; HIV antibody
Lumbar puncture (if not already performed)
Brain MRI with and without gadolinium (if not already performed)

Abbreviations: c-ANCA, cytoplasmic antineutrophil cytoplasmic antibody; CNS, central nervous system; CT, computed tomography; MRI, magnetic resonance imaging; p-ANCA, perinuclear antineutrophil cytoplasmic antibody.

TREATMENT

TREATMENT

Delirium

Management of delirium begins with treatment of the underlying inciting factor (e.g., patients with systemic infections should be given appropriate antibiotics, and underlying electrolyte disturbances judiciously corrected). These treatments often lead to prompt resolution of delirium. Blindly targeting the symptoms of delirium pharmacologically only serves to prolong the time patients remain in the confused state and may mask important diagnostic information.

Relatively simple methods of supportive care can be highly effective. Reorientation by the nursing staff and family combined with visible clocks, calendars, and outside-facing windows can reduce confusion. Sensory isolation should be prevented by providing glasses and hearing aids to patients who need them. Sundowning can be addressed to a large extent through vigilance to appropriate sleep-wake cycles. During the day, a well-lit room should be accompanied by activities or exercises to prevent napping. At night, a quiet, dark environment with limited interruptions by staff can assure proper rest. These sleep-wake cycle interventions are especially important in the ICU setting as the usual constant 24-h activity commonly provokes delirium. Attempting to mimic the home environment as much as possible also has been shown to help treat and even prevent delirium. Visits from friends and family throughout the day minimize the anxiety associated with the

constant flow of new faces of staff and physicians. Allowing hospitalized patients to have access to home bedding, clothing, and nightstand objects makes the hospital environment less foreign and therefore less confusing. Simple standard nursing practices such as maintaining proper nutrition and volume status as well as managing pain, incontinence and skin breakdown also help alleviate discomfort and resulting confusion.

In some instances, patients pose a threat to their own safety or to the safety of staff members, and acute management is required. Bed alarms and personal sitters are more effective and much less disorienting than physical restraints. Chemical restraints should be avoided, but when necessary, very-low-dose typical or atypical antipsychotic medications administered on an as-needed basis can be used; however, there is little evidence that these medications are effective in delirium, and therefore they should be reserved for patients who display severe agitation and significant potential to harm themselves or staff. The recent association of antipsychotic use in the elderly with increased mortality rates underscores the importance of using these medications judiciously and only as a last resort. Benzodiazepines often worsen confusion through their sedative properties. Although many clinicians still use benzodiazepines to treat acute confusion, their use should be limited to cases in which delirium is caused by alcohol or benzodiazepine withdrawal.

PREVENTION

In light of the high morbidity associated with delirium and the tremendously increased health care costs that accompany it, development of an effective strategy to prevent delirium in hospitalized patients is extremely important. Successful identification of high-risk patients is the first step, followed by initiation of appropriate interventions. Increasingly, hospitals are using nursing or physician-administered tools to screen for high-risk individuals, triggering simple standardized protocols used to manage risk factors for delirium, including sleep-wake cycle reversal, immobility, visual impairment, hearing impairment, sleep deprivation, and dehydration. No specific medications have been definitively shown to be effective for delirium prevention, including trials of cholinesterase inhibitors and antipsychotic agents. Melatonin and its agonist ramelteon have shown some promising results in small preliminary trials. Recent studies in the ICU have focused both on identifying sedatives, such as dexmedetomidine, that are less likely to lead to delirium in critically ill patients and on developing protocols for daily awakenings in which infusions of sedative medications are interrupted and the patient is reorientated by the staff. All hospitals and health care systems should work toward decreasing the incidence of delirium and promptly recognizing and treating the disorder when it occurs.

FURTHER READING

Constantin JM et al: Efficacy and safety of sedation with dexmedetomidine in critical care patients: A meta-analysis of randomized controlled trials. Anaesth Crit Care Pain Med 35:7, 2016. [PubMed: 26700947]

Hatta K et al: Preventive effects of ramelteon on delirium: A randomized placebo-controlled trial. JAMA Psychiatry 71:397, 2014. [PubMed: 24554232]

Neufeld KJ et al: Antipsychotic medication for prevention and treatment of delirium in hospitalized adults: A systematic review and meta-analysis. J Am Geriatr Soc 64:705, 2016. [PubMed: 27004732]

Chapter 25: Dementia

William W. Seeley; Bruce L. Miller

INTRODUCTION

Dementia, a syndrome with many causes, affects >5 million people in the United States and results in a total annual health care cost in excess of $250 billion. Dementia is defined as an acquired deterioration in cognitive abilities that impairs the successful performance of activities of daily living. Episodic memory, the ability to recall events specific in time and place, is the cognitive function most commonly lost; 10% of persons aged >70 years and 20–40% of individuals aged >85 years have clinically identifiable memory loss. In addition to memory, dementia may erode other mental faculties, including language, visuospatial, praxis, calculation, judgment, and problem-solving abilities. Neuropsychiatric and social deficits also arise in many dementia syndromes, manifesting as depression, apathy, anxiety, hallucinations, delusions, agitation, insomnia, sleep disturbances, compulsions, or disinhibition. The clinical course may be slowly progressive, as in *Alzheimer's disease* (*AD*); static, as in anoxic encephalopathy; or may fluctuate from day to day or minute to minute, as in *dementia with Lewy bodies* (*DLB*). Most patients with AD, the most prevalent form of dementia, begin with episodic memory impairment, although in other dementias, such as *frontotemporal dementia* (*FTD*), memory loss is not typically a presenting feature. **Focal cerebral disorders are discussed in Chap. 26 and illustrated in a video library in Chap. V2; detailed discussions of AD can be found in Chap. 423; FTD and related disorders in Chap. 424; vascular dementia in Chap. 425; DLB in Chap. 426; Huntington's disease (HD) in Chap. 428; and prion diseases in Chap. 430.**

FUNCTIONAL ANATOMY OF THE DEMENTIAS

Dementia syndromes result from the disruption of specific large-scale neuronal networks; the location and severity of synaptic and neuronal loss combine to produce the clinical features **(Chap. 26)**. Behavior, mood, and attention are modulated by ascending noradrenergic, serotonergic, and dopaminergic pathways, whereas cholinergic signaling is critical for attention and memory functions. The dementias differ in the relative neurotransmitter deficit profiles; accordingly, accurate diagnosis guides effective pharmacologic therapy.

AD begins in the entorhinal region of the medial temporal lobe, spreads to the hippocampus, and then moves to lateral and posterior temporal and parietal neocortex, eventually causing a more widespread degeneration. *Vascular dementia* is associated with focal damage in a variable patchwork of cortical and subcortical regions or white matter tracts that disconnect nodes within distributed networks. In keeping with

its anatomy, AD typically presents with episodic memory loss accompanied later by aphasia, executive dysfunction, or navigational problems. In contrast, dementias that begin in frontal or subcortical regions, such as FTD or HD, are less likely to begin with memory problems and more likely to present with difficulties with judgment, mood, executive control, movement, and behavior.

Lesions of frontal-striatal[1] pathways produce specific and predictable effects on behavior. The dorsolateral prefrontal cortex has connections with a central band of the caudate nucleus. Lesions of either the caudate or dorsolateral prefrontal cortex, or their connecting white matter pathways, may result in executive dysfunction, manifesting as poor organization and planning, decreased cognitive flexibility, and impaired working memory. The lateral orbital frontal cortex connects with the ventromedial caudate, and lesions of this system cause impulsiveness, distractibility, and disinhibition. The anterior cingulate cortex and adjacent medial prefrontal cortex project to the nucleus accumbens, and interruption of this system produces apathy, poverty of speech, emotional blunting, or even akinetic mutism. All corticostriatal systems also include topographically organized projections through the globus pallidus and thalamus, and damage to these nodes can likewise reproduce the clinical syndrome associated with the corresponding cortical or striatal injuries.

THE CAUSES OF DEMENTIA

The single strongest risk factor for dementia is increasing age. The prevalence of disabling memory loss increases with each decade for those aged >50 and is usually associated with the microscopic changes of AD at autopsy. Yet some centenarians have intact memory function and no evidence of clinically significant dementia. Whether dementia is an inevitable consequence of normal human aging remains controversial.

The many causes of dementia are listed in **Table 25-1**. The frequency of each condition depends on the age group under study, access of the group to medical care, country of origin, and perhaps racial or ethnic background. AD is the most common cause of dementia in Western countries, accounting for more than half of all patients. Vascular disease is considered the second most frequent cause for dementia and is particularly common in elderly patients or populations with limited access to medical care, where vascular risk factors are undertreated. Often, vascular brain injury is mixed with neurodegenerative disorders, making it difficult, even for the neuropathologist, to estimate the contribution of cerebrovascular disease to the cognitive disorder in an individual patient. Dementias associated with Parkinson's disease (PD) are common and may develop years after onset of a parkinsonian disorder, as seen with PD-related dementia (PDD), or can occur concurrently with or preceding the motor syndrome, as in DLB. A mixed pathology is common, especially in very old individuals. In patients aged <65, FTD rivals AD as the most common cause of dementia. Chronic intoxications, including those resulting from alcohol and prescription drugs, are an important and often treatable cause of dementia. Other disorders listed in Table 25-1 are uncommon but important because many are reversible. The classification of dementing illnesses into reversible and irreversible disorders is a useful approach to differential diagnosis. When effective treatments for the neurodegenerative conditions emerge, this dichotomy will become obsolete.

TABLE 25-1

Differential Diagnosis of Dementia

Most Common Causes of Dementia	
Alzheimer's disease Vascular dementia Multi-infarct Diffuse white matter disease (Binswanger's)	Alcoholism[a] PDD/LBD spectrum Drug/medication intoxication[a]
Less Common Causes of Dementia	
Vitamin deficiencies Thiamine (B_1): Wernicke's encephalopathy[a] B_{12} (subacute combined degeneration)[a] Nicotinic acid (pellagra)[a] Endocrine and other organ failure Hypothyroidism[a] Adrenal insufficiency and Cushing's syndrome[a] Hypo- and hyperparathyroidism[a] Renal failure[a] Liver failure[a] Pulmonary failure[a] Chronic infections HIV Neurosyphilis[a] Papovavirus (JC virus) (progressive multifocal leukoencephalopathy) Tuberculosis, fungal, and protozoal[a] Whipple's disease[a] Head trauma and diffuse brain damage Chronic traumatic encephalopathy	Toxic disorders Drug, medication, and narcotic poisoning[a] Heavy metal intoxication[a] Organic toxins Psychiatric Depression (pseudodementia)[a] Schizophrenia[a] Conversion disorder[a] Degenerative disorders Huntington's disease Multisystem atrophy Hereditary ataxias (some forms) Frontotemporal lobar degeneration spectrum Multiple sclerosis Adult Down's syndrome with Alzheimer's disease ALS-parkinsonism-dementia complex of Guam Prion (Creutzfeldt-Jakob and Gerstmann-Sträussler-Scheinker diseases) Miscellaneous Sarcoidosis[a] Vasculitis[a] CADASIL, etc. Acute intermittent porphyria[a] Recurrent nonconvulsive seizures[a] Additional conditions in children or adolescents Pantothenate kinase–associated neurodegeneration

Chronic subdural hematoma[a] Postanoxia Postencephalitis Normal-pressure hydrocephalus[a] Intracranial hypotension Neoplastic Primary brain tumor[a] Metastatic brain tumor[a] Paraneoplastic/autoimmune limbic encephalitis[a]	Subacute sclerosing panencephalitis Metabolic disorders (e.g., Wilson's and Leigh's diseases, leukodystrophies, lipid storage diseases, mitochondrial mutations)

[a]Potentially reversible dementia.

Abbreviations: ALS, amyotrophic lateral sclerosis; CADASIL, cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy; LBD, Lewy body disease; PDD, Parkinson's disease dementia.

In a study of 1000 persons attending a memory disorders clinic, 19% had a potentially reversible cause of the cognitive impairment and 23% had a potentially reversible concomitant condition that may have contributed to the patient's impairment. The three most common potentially reversible diagnoses were depression, normal pressure hydrocephalus (NPH), and alcohol dependence; medication side effects are also common and should be considered in every patient (Table 25-1).

Subtle cumulative decline in episodic memory is a common part of aging. This frustrating experience, often the source of jokes and humor, is often referred to as *benign forgetfulness of the elderly*. *Benign* means that it is not so progressive or serious that it impairs reasonably successful and productive daily functioning, although the distinction between benign and more significant memory loss can be difficult to make. At age 85, the average person is able to learn and recall approximately one-half of the items (e.g., words on a list) that he or she could at age 18. A measurable cognitive problem that does not seriously disrupt daily activities is often referred to as *mild cognitive impairment* (MCI). Factors that predict progression from MCI to an AD dementia include a prominent memory deficit, family history of dementia, presence of an apolipoprotein ε4 (Apo ε4) allele, small hippocampal volumes, an AD-like signature of cortical atrophy, low cerebrospinal fluid Aβ, and elevated tau or evidence of brain amyloid deposition on positron emission tomography (PET) imaging.

The major degenerative dementias include AD, DLB, FTD and related disorders, HD, and prion diseases, including Creutzfeldt-Jakob disease (CJD). These disorders are all associated with the abnormal aggregation of a specific protein: $A\beta_{42}$ and tau in AD; α-synuclein in DLB; tau, TAR DNA-binding protein of 43 kDa (TDP-43), or *fu*sed in *s*arcoma (FUS) in FTD; huntingtin in HD; and misfolded prion protein (PrP^{Sc}) in CJD **(Table 25-2)**.

TABLE 25-2

The Molecular Basis for Degenerative Dementia

Dementia	Molecular Basis	Causal Genes (Chromosome)	Susceptibility Genes	Pathologic Findings
AD	Aβ/tau	*APP* (21), *PS-1* (14), *PS-2* (1) (<2% carry these mutations, most often in *PS-1*)	*Apo ε4* (19)	Amyloid plaques, neurofibrillary tangles, and neuropil threads
FTD	Tau	*MAPT* exon and intron mutations (17) (about 10% of familial cases)	H1 *MAPT* haplotype	Tau neuronal and glial inclusions varying in morphology and distribution
	TDP-43	*GRN* (10% of familial cases), *C9ORF72* (20–30% of familial cases), rare *VCP*, very rare *TARDBP, TBK1, TIA1*		TDP-43 neuronal and glial inclusions varying in morphology and distribution
	FUS	Very rare *FUS*		FUS neuronal and glial inclusions varying in morphology and distribution
DLB	α-Synuclein	Very rare *SNCA* (4)	Unknown	α-Synuclein neuronal inclusions (Lewy bodies)
CJD	PrP^{SC}	*PRNP* (20) (up to 15% of patients carry these dominant mutations)	Codon 129 homozygosity for methionine or valine	PrP^{SC} deposition, panlaminar spongiosis

Abbreviations: AD, Alzheimer's disease; CJD, Creutzfeldt-Jakob disease; DLB, dementia with Lewy bodies; FTD, frontotemporal dementia.

[1]The striatum comprises the caudate/putamen/nucleus accumbens.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Dementias

Three major issues should be kept at the forefront: (1) What is the best fit for a clinical diagnosis? (2) What component of the dementia syndrome is treatable or reversible? (3) Can the physician help to alleviate the burden on caregivers? A broad overview of the approach to dementia is shown in **Table 25-3**. The major degenerative dementias can usually be distinguished by the initial symptoms; neuropsychological, neuropsychiatric, and neurologic findings; and neuroimaging features **(Table 25-4)**.

HISTORY

The history should concentrate on the onset, duration, and tempo of progression. An acute or subacute onset of confusion may be due to delirium **(Chap. 24)** and should trigger the search for intoxication, infection, or metabolic derangement. An elderly person with slowly progressive memory loss over several years is likely to suffer from AD. Nearly 75% of patients with AD begin with memory symptoms, but other early symptoms include difficulty with managing money, driving, shopping, following instructions, finding words, or navigating. Personality change, disinhibition, and weight gain or compulsive eating suggest FTD, not AD. FTD is also suggested by prominent apathy, compulsivity, loss of empathy for others, or progressive loss of speech fluency or single-word comprehension and by a relative sparing of memory and visuospatial abilities. The diagnosis of DLB is suggested by early visual hallucinations; parkinsonism; proneness to delirium or sensitivity to psychoactive medications; rapid eye movement (REM) behavior disorder (RBD; the loss of skeletal muscle paralysis during dreaming); or Capgras syndrome, the delusion that a familiar person has been replaced by an impostor.

A history of stroke with irregular stepwise progression suggests vascular dementia. Vascular dementia is also commonly seen in the setting of hypertension, atrial fibrillation, peripheral vascular disease, and diabetes. In patients suffering from cerebrovascular disease, it can be difficult to determine whether the dementia is due to AD, vascular disease, or a mixture of the two because many of the risk factors for vascular dementia, including diabetes, high cholesterol, elevated homocysteine, and low exercise, are also putative risk factors for AD. Moreover, many patients with a major vascular contribution to their dementia lack a history of stepwise decline. Rapid progression with motor rigidity and myoclonus suggests CJD **(Chap. 430)**. Seizures may indicate strokes or neoplasm but also occur in AD, particularly early-age-of-onset AD. Gait disturbance is common in vascular dementia, PD/DLB, or NPH. A history of high-risk sexual behaviors or intravenous drug use should trigger a search for central nervous system (CNS) infection, especially HIV or syphilis. A history of recurrent head trauma could indicate chronic subdural hematoma, chronic traumatic encephalopathy (a progressive dementia best characterized in contact sport athletes such as boxers and American football players), intracranial hypotension, or NPH. Subacute onset of severe amnesia and psychosis with mesial temporal T2/fluid-attenuated inversion recovery (FLAIR) hyperintensities on magnetic resonance imaging (MRI) should raise concern for paraneoplastic limbic encephalitis, especially in a long-term smoker or other patients at risk for cancer. Related autoimmune conditions, such as voltage-gated potassium channel (VGKC)- or N-methyl-D-aspartate (NMDA)-receptor antibody-mediated encephalopathy, can present with a similar tempo and imaging signature with or without characteristic motor manifestations such as myokymia (anti-

VGKC) and faciobrachial dystonic seizures (anti-NMDA) **(Chap. 90)**. Alcohol abuse creates risk for malnutrition and thiamine deficiency. Veganism, bowel irradiation, an autoimmune diathesis, a remote history of gastric surgery, and chronic antihistamine therapy for dyspepsia or gastroesophageal reflux predispose to B_{12} deficiency. Certain occupations, such as working in a battery or chemical factory, might indicate heavy metal intoxication. Careful review of medication intake, especially for sedatives and analgesics, may raise the issue of chronic drug intoxication. An autosomal dominant family history is found in HD and in familial forms of AD, FTD, DLB, or prion disorders. A history of mood disorders, the recent death of a loved one, or depressive signs, such as insomnia or weight loss, raise the possibility of depression-related cognitive impairments.

PHYSICAL AND NEUROLOGIC EXAMINATION

A thorough general and neurologic examination is essential to document dementia, to look for other signs of nervous system involvement, and to search for clues suggesting a systemic disease that might be responsible for the cognitive disorder. Typical AD spares motor systems until later in the course. In contrast, FTD patients often develop axial rigidity, supranuclear gaze palsy, or a motor neuron disease reminiscent of amyotrophic lateral sclerosis (ALS). In DLB, the initial symptoms may include the new onset of a parkinsonian syndrome (resting tremor, cogwheel rigidity, bradykinesia, festinating gait), but DLB often starts with visual hallucinations or dementia. Symptoms referable to the lower brainstem (RBD, gastrointestinal or autonomic problems) may arise years or even decades before parkinsonism or dementia. Corticobasal syndrome (CBS) features asymmetric akinesia and rigidity, dystonia, myoclonus, alien limb phenomena, pyramidal signs, and prefrontal deficits such as nonfluent aphasia with or without motor speech impairment, executive dysfunction, apraxia, or a behavioral disorder. Progressive supranuclear palsy (PSP) is associated with unexplained falls, axial rigidity, dysphagia, and vertical gaze deficits. CJD is suggested by the presence of diffuse rigidity, an akinetic-mute state, and prominent, often startle-sensitive myoclonus.

Hemiparesis or other focal neurologic deficits suggest vascular dementia or brain tumor. Dementia with a myelopathy and peripheral neuropathy suggests vitamin B_{12} deficiency. Peripheral neuropathy could also indicate another vitamin deficiency, heavy metal intoxication, thyroid dysfunction, Lyme disease, or vasculitis. Dry, cool skin, hair loss, and bradycardia suggest hypothyroidism. Fluctuating confusion associated with repetitive stereotyped movements may indicate ongoing limbic, temporal, or frontal seizures. In the elderly, hearing impairment or visual loss may produce confusion and disorientation misinterpreted as dementia. Profound bilateral sensorineural hearing loss in a younger patient with short stature or myopathy, however, should raise concern for a mitochondrial disorder.

COGNITIVE AND NEUROPSYCHIATRIC EXAMINATION

Brief screening tools such as the Mini-Mental State Examination (MMSE), the Montreal Cognitive Assessment (MOCA), and Cognistat can be used to capture dementia and follow progression. None of these tests is highly sensitive to early-stage dementia or discriminates between dementia syndromes. The MMSE is a 30-point test of cognitive function, with each correct answer being scored as 1 point. It includes tests in the areas of: orientation (e.g., identify season/date/month/year/floor/hospital/town/state/country); registration (e.g., name and restate 3 objects); recall (e.g., remember the same three objects 5 min later); and language (e.g., name pencil and watch; repeat “no ifs ands or buts”; follow a 3-step command; obey a written command;

and write a sentence and copy a design). In most patients with MCI and some with clinically apparent AD, bedside screening tests may be normal, and a more challenging and comprehensive set of neuropsychological tests will be required. When the etiology for the dementia syndrome remains in doubt, a specially tailored evaluation should be performed that includes tasks of working and episodic memory, executive function, language, and visuospatial and perceptual abilities. In AD, the early deficits involve episodic memory, category generation ("name as many animals as you can in 1 minute"), and visuoconstructive ability. Usually deficits in verbal or visual episodic memory are the first neuropsychological abnormalities detected, and tasks that require the patient to recall a long list of words or a series of pictures after a predetermined delay will demonstrate deficits in most patients. In FTD, the earliest deficits on cognitive testing involve executive control or language (speech or naming) function, but some patients lack either finding despite profound social-emotional deficits. PDD or DLB patients have more severe deficits in visuospatial function but do better on episodic memory tasks than patients with AD. Patients with vascular dementia often demonstrate a mixture of executive control and visuospatial deficits, with prominent psychomotor slowing. In delirium, the most prominent deficits involve attention, working memory, and executive function, making the assessment of other cognitive domains challenging and often uninformative.

A functional assessment should also be performed to help the physician determine the day-to-day impact of the disorder on the patient's memory, community affairs, hobbies, judgment, dressing, and eating. Knowledge of the patient's functional abilities will help the clinician and the family to organize a therapeutic approach.

Neuropsychiatric assessment is important for diagnosis, prognosis, and treatment. In the early stages of AD, mild depressive features, social withdrawal, and irritability or anxiety are the most prominent psychiatric changes, but patients often maintain core social graces into the middle or late stages, when delusions, agitation, and sleep disturbance may emerge. In FTD, dramatic personality change with apathy, overeating, compulsions, disinhibition, euphoria, and loss of empathy are early and common. DLB is associated with visual hallucinations, delusions related to person or place identity, RBD, and excessive daytime sleepiness. Dramatic fluctuations occur not only in cognition but also in arousal. Vascular dementia can present with psychiatric symptoms such as depression, anxiety, delusions, disinhibition, or apathy.

LABORATORY TESTS

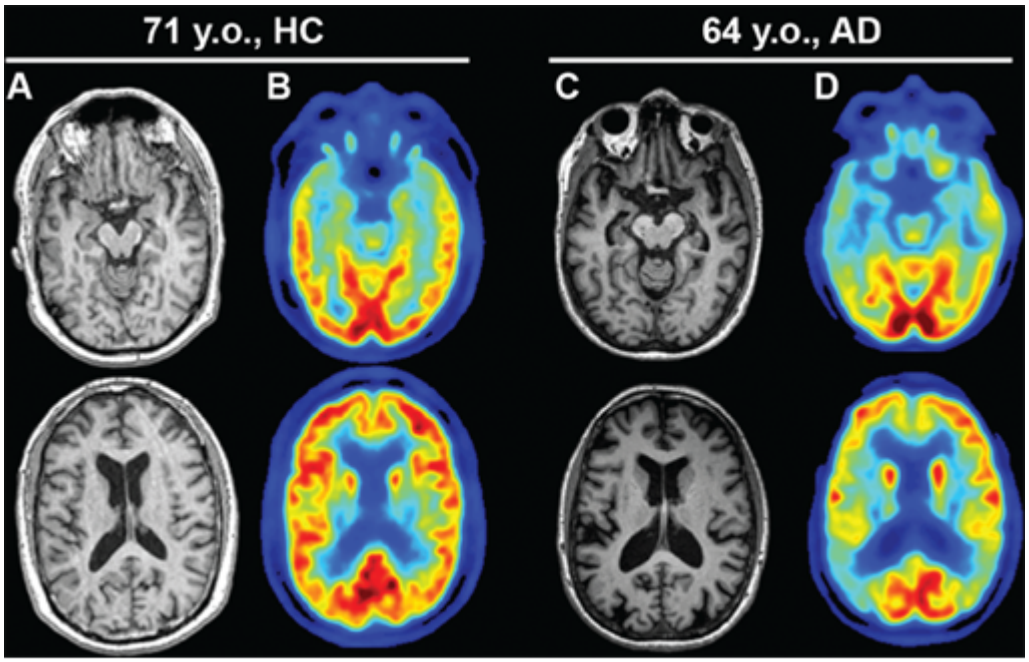
The choice of laboratory tests in the evaluation of dementia is complex and should be tailored to the individual patient. The physician must take measures to avoid missing a reversible or treatable cause, yet no single treatable etiology is common; thus, a screen must use multiple tests, each of which has a low yield. Cost/benefit ratios are difficult to assess, and many laboratory screening algorithms for dementia discourage multiple tests. Nevertheless, even a test with only a 1–2% positive rate is worth undertaking if the alternative is missing a treatable cause of dementia. Table 25-3 lists most screening tests for dementia. The American Academy of Neurology recommends the routine measurement of a complete blood count, electrolytes, renal and thyroid function, a vitamin B_{12} level, and a neuroimaging study (computed tomography [CT] or MRI).

Neuroimaging studies, especially MRI, help to rule out primary and metastatic neoplasms, locate areas of infarction or inflammation, detect subdural hematomas, and suggest NPH or diffuse white matter disease. They also help to establish a regional pattern of atrophy. Support for the diagnosis of AD includes hippocampal atrophy in addition to posterior-predominant cortical atrophy **(Fig. 25-1)**. Focal frontal, insular, and/or anterior temporal atrophy suggests FTD **(Chap. 424)**. DLB often features less prominent atrophy, with greater involvement of amygdala than hippocampus. In CJD, magnetic resonance (MR) diffusion-weighted imaging reveals restricted diffusion within the cortical ribbon and/or basal ganglia in most patients. Extensive multifocal white matter abnormalities suggest a vascular etiology **(Fig. 25-2)**. Communicating hydrocephalus with vertex effacement (crowding of dorsal convexity gyri/sulci), gaping Sylvian fissures despite minimal cortical atrophy, and additional features shown in **Fig. 25-3** suggest NPH. Single-photon emission computed tomography (SPECT) and PET scanning show temporal-parietal hypoperfusion or hypometabolism in AD and frontotemporal deficits in FTD, but these changes often reflect atrophy and can therefore be detected with MRI alone in many patients. Recently, amyloid imaging has shown promise for the diagnosis of AD, and Pittsburgh Compound-B (PiB) (not available outside of research settings) and ^{18}F-AV-45 (florbetapir; approved by the U.S. Food and Drug Administration in 2013) are reliable radioligands for detecting brain amyloid associated with amyloid angiopathy or neuritic plaques of AD **(Fig. 25-4)**. Because these abnormalities can be seen in cognitively normal older persons (~25% of individuals at age 65), however, amyloid imaging may also detect preclinical or incidental AD in patients lacking an AD-like dementia syndrome. Currently, the main clinical value of amyloid imaging is to exclude AD as the likely cause of dementia in patients who have negative scans. Once disease-modifying therapies become available, use of these biomarkers may help to identify treatment candidates before irreversible brain injury has occurred. In the meantime, the prognostic value of detecting brain amyloid in an asymptomatic elder remains a topic of vigorous investigation. Similarly, MRI perfusion and structural/functional connectivity methods are being explored as potential treatment-monitoring strategies.

Lumbar puncture need not be done routinely in the evaluation of dementia, but it is indicated when CNS infection or inflammation are credible diagnostic possibilities. Cerebrospinal fluid (CSF) levels of $A\beta_{42}$ and tau proteins show differing patterns with the various dementias, and the presence of low $A\beta_{42}$ and mildly elevated CSF tau is highly suggestive of AD. The routine use of lumbar puncture in the diagnosis of dementia is debated, but the sensitivity and specificity of AD diagnostic measures are not yet high enough to warrant routine use. Formal psychometric testing helps to document the severity of cognitive disturbance, suggest psychogenic causes, and provide a more formal method for following the disease course. Electroencephalogram (EEG) is not routinely used but can help to suggest CJD (repetitive bursts of diffuse high-amplitude sharp waves, or “periodic complexes”) or an underlying nonconvulsive seizure disorder (epileptiform discharges). Brain biopsy (including meninges) is not advised except to diagnose vasculitis, potentially treatable neoplasms, or unusual infections when the diagnosis is uncertain. Systemic disorders with CNS manifestations, such as sarcoidosis, can usually be confirmed through biopsy of lymph node or solid organ rather than brain. MR angiography should be considered when cerebral vasculitis or cerebral venous thrombosis is a possible cause of the dementia.

FIGURE 25-1

Alzheimer's disease (AD). Axial T1-weighted magnetic resonance images of a healthy 71-year-old (***A***) and a 64-year-old with AD (***C***). Note the reduction in medial temporal lobe volume in the patient with AD. Fluorodeoxyglucose positron emission tomography scans of the same individuals (***B*** and ***D***) demonstrate reduced glucose metabolism in the posterior temporoparietal regions bilaterally in AD, a typical finding in this condition. HC, healthy control. *(Images courtesy of Gil Rabinovici, University of California, San Francisco and William Jagust, University of California, Berkeley.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 25-2

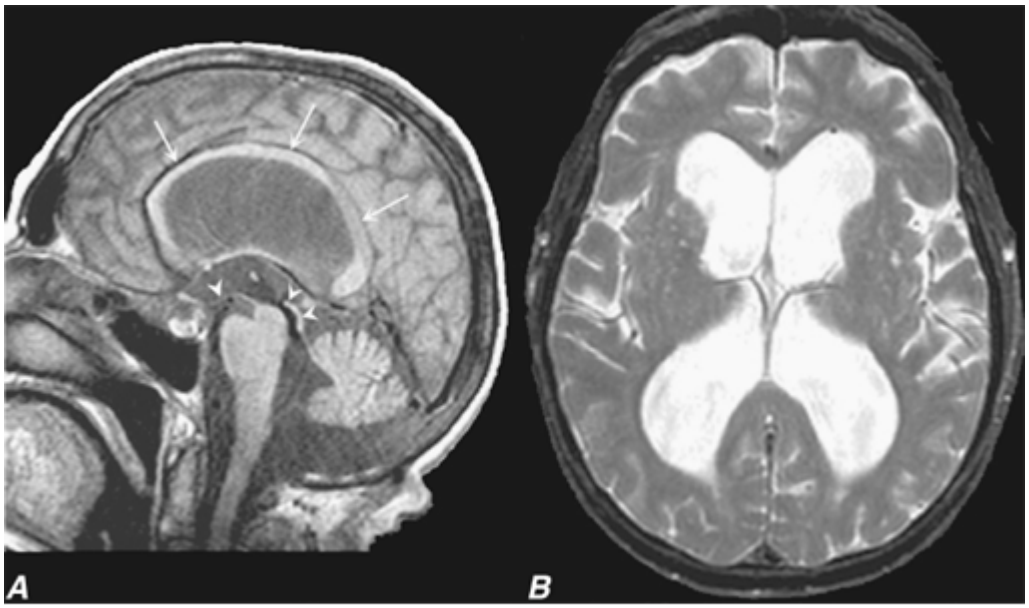
Diffuse white matter disease. Axial fluid-attenuated inversion recovery (FLAIR) magnetic resonance image through the lateral ventricles reveals multiple areas of hyperintensity (*arrows*) involving the periventricular white matter as well as the corona radiata and striatum. Although seen in some individuals with normal cognition, this appearance is more pronounced in patients with dementia of a vascular etiology.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 25-3

Normal-pressure hydrocephalus. ***A.*** Sagittal T1-weighted magnetic resonance image (MRI) demonstrates dilation of the lateral ventricle and stretching of the corpus callosum (*arrows*), depression of the floor of the third ventricle (*single arrowhead*), and enlargement of the aqueduct (*double arrowheads*). Note the diffuse dilation of the lateral, third, and fourth ventricles with a patent aqueduct, typical of communicating hydrocephalus. ***B.*** Axial T2-weighted MRIs demonstrate dilation of the lateral ventricles. This patient underwent successful ventriculoperitoneal shunting.

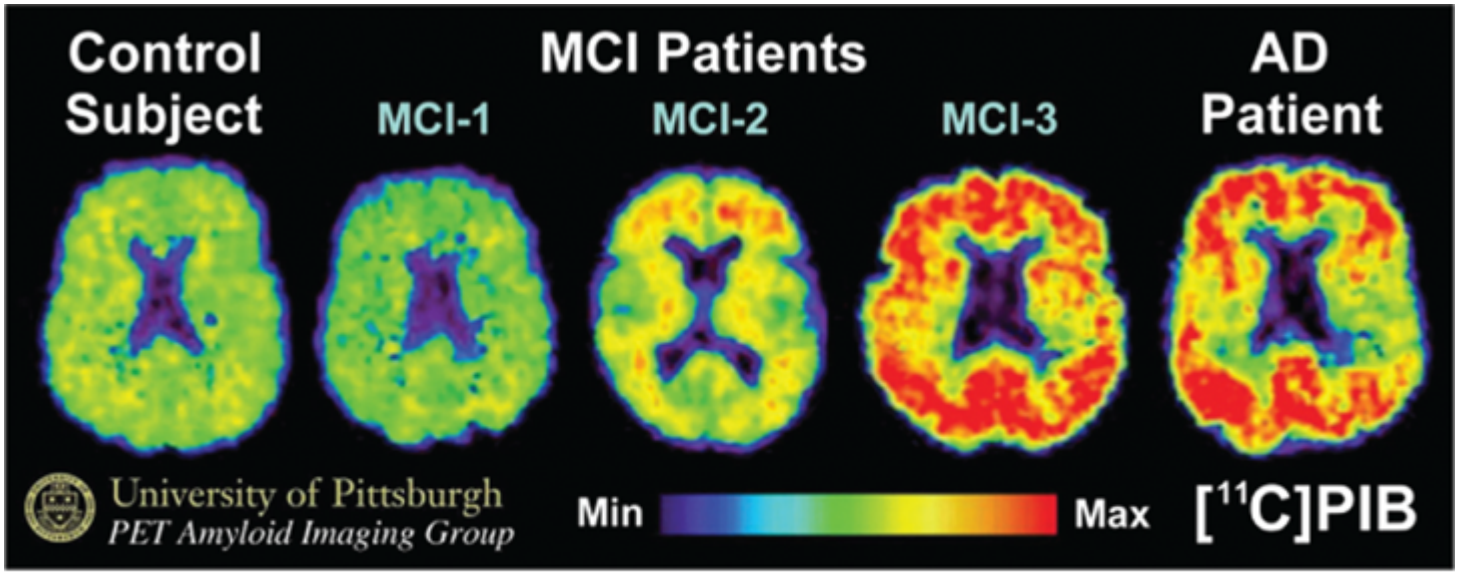


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 25-4

Positron emission tomography (PET) images obtained with the amyloid-imaging agent Pittsburgh Compound-B ([^{11}C]PIB) in a normal control (*left*); three different patients with mild cognitive impairment

(MCI; *center*); and a patient with mild Alzheimer's disease (AD; *right*). Some MCI patients have control-like levels of amyloid, some have AD-like levels of amyloid, and some have intermediate levels. *(Images courtesy of William Klunk and Chester Mathis, University of Pittsburgh.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 25-3

Evaluation of the Patient with Dementia

Routine Evaluation	**Optional Focused Tests**	**Occasionally Helpful Tests**
History Physical examination Laboratory tests Thyroid function (TSH) Vitamin B_{12} Complete blood count Electrolytes CT/MRI	Psychometric testing Chest x-ray Lumbar puncture Liver function Renal function Urine toxin screen HIV Apolipoprotein E RPR or VDRL	EEG Parathyroid function Adrenal function Urine heavy metals RBC sedimentation rate Angiogram Brain biopsy SPECT PET Lab screen for autoantibodies
Diagnostic Categories		
Reversible Causes	**Irreversible/Degenerative Dementias**	**Psychiatric Disorders**
Examples Hypothyroidism Thiamine deficiency Vitamin B_{12} deficiency Normal-pressure hydrocephalus Subdural hematoma Chronic infection Brain tumor Drug intoxication Autoimmune encephalopathy	Examples Alzheimer's Frontotemporal dementia Huntington's Dementia with Lewy bodies Vascular Leukoencephalopathies Parkinson's	Depression Schizophrenia Conversion reaction
Associated Treatable Conditions		
	Depression Seizures Insomnia	Agitation Caregiver "burnout" Drug side effects

Abbreviations: CT, computed tomography; EEG, electroencephalogram; MRI, magnetic resonance imaging; PET, positron emission tomography; RBC, red blood cell; RPR, rapid plasma reagin (test); SPECT, single-photon emission

computed tomography; TSH, thyroid-stimulating hormone; VDRL, Venereal Disease Research Laboratory (test for syphilis).

TABLE 25-4

Clinical Differentiation of the Major Dementias

Disease	First Symptom	Mental Status	Neuropsychiatry	Neurology	Imaging
AD	Memory loss	Episodic memory loss	Irritability, anxiety, depression	Initially normal	Entorhinal cortex and hippocampal atrophy
FTD	Apathy; poor judgment/insight, speech/language; hyperorality	Frontal/executive and/or language; spares drawing	Apathy, disinhibition, overeating, compulsivity	May have vertical gaze palsy, axial rigidity, dystonia, alien hand, or MND	Frontal, insular, and/or temporal atrophy; usually spares posterior parietal lobe
DLB	Visual hallucinations, REM sleep behavior disorder, delirium, Capgras syndrome, parkinsonism	Drawing and frontal/executive; spares memory; delirium-prone	Visual hallucinations, depression, sleep disorder, delusions	Parkinsonism	Posterior parietal atrophy; hippocampi larger than in AD
CJD	Dementia, mood, anxiety, movement disorders	Variable, frontal/executive, focal cortical, memory	Depression, anxiety, psychosis in some	Myoclonus, rigidity, parkinsonism	Cortical ribboning and basal ganglia or thalamus hyperintensity on diffusion/FLAIR MRI

Disease	First Symptom	Mental Status	Neuropsychiatry	Neurology	Imaging
Vascular	Often but not always sudden; variable; apathy, falls, focal weakness	Frontal/executive, cognitive slowing; can spare memory	Apathy, delusions, anxiety	Usually motor slowing, spasticity; can be normal	Cortical and/or subcortical infarctions, confluent white matter disease

Abbreviations: AD, Alzheimer's disease; CBD, cortical basal degeneration; CJD, Creutzfeldt-Jakob disease; DLB, dementia with Lewy bodies; FLAIR, fluid-attenuated inversion recovery; FTD, frontotemporal dementia; MND, motor neuron disease; MRI, magnetic resonance imaging; PSP, progressive supranuclear palsy; REM, rapid eye movement.

GLOBAL CONSIDERATIONS

Vascular dementia (Chap. 425) is more common in Asian countries, due to the higher prevalence of intracranial atherosclerosis. Rates of vascular dementia are also on the rise in developing countries as vascular risk factors such as hypertension, hypercholesterolemia, and diabetes mellitus become more widespread. CNS infections, particularly with HIV (and associated opportunistic infections), syphilis, and tuberculosis, likewise represent major contributors to dementia in the developing world. Isolated populations have also contributed to our understanding of neurodegenerative dementia. Kuru, the cannibalism-associated rapidly progressive dementia seen in tribal New Guinea, played a role in the discovery of human prion disease. Amyotrophic lateral sclerosis-parkinsonism-dementia complex of Guam (or, Lytico-Bodig disease) is a poly-proteinopathy, often with tau, TDP-43, and alpha-synuclein aggregation. The root cause of the disease remains uncertain, but its incidence has declined sharply over the past 60 years.

TREATMENT

TREATMENT

Dementia

The major goals of dementia management are to treat reversible causes and to provide comfort and support to the patient and caregivers. Treatment of underlying causes includes thyroid replacement for hypothyroidism; vitamin therapy for thiamine or B_{12} deficiency or for elevated serum homocysteine; antimicrobials for opportunistic infections or antiretrovirals for HIV; ventricular shunting for NPH; or appropriate surgical, radiation, and/or chemotherapeutic treatment for CNS neoplasms. Removal of cognition-impairing drugs or medications is frequently useful. If the patient's cognitive complaints stem from a psychiatric disorder, vigorous treatment of this condition should seek to eliminate the cognitive complaint

or confirm that it persists despite adequate resolution of the mood or anxiety symptoms. Patients with degenerative diseases may also be depressed or anxious, and those aspects of their condition often respond to therapy. Antidepressants, such as selective serotonin reuptake inhibitors (SSRIs) or serotonin-norepinephrine reuptake inhibitors (SNRIs) **(Chap. 443)**, which feature anxiolytic properties but few cognitive side effects, provide the mainstay of treatment when necessary. Anticonvulsants are used to control seizures. Levetiracetam may be particularly useful, but there have as yet been no randomized trials for treatment of AD-associated seizures.

Agitation, hallucinations, delusions, and confusion are difficult to treat. These behavioral problems represent major causes for nursing home placement and institutionalization. Before treating these behaviors with medications, the clinician should aggressively seek out modifiable environmental or metabolic factors. Hunger, lack of exercise, toothache, constipation, urinary tract or respiratory infection, electrolyte imbalance, and drug toxicity all represent easily correctable causes that can be remedied without psychoactive drugs. Drugs such as phenothiazines and benzodiazepines may ameliorate the behavior problems but have untoward side effects such as sedation, rigidity, dyskinesia, and occasionally paradoxical disinhibition (benzodiazepines). Despite their unfavorable side effect profile, second-generation antipsychotics such as quetiapine (starting dose, 12.5–25 mg daily) can be used for patients with agitation, aggression, and psychosis, although the risk profile for these compounds is significant. When patients do not respond to treatment, it is usually a mistake to advance to higher doses or to use anticholinergic drugs or sedatives (such as barbiturates or benzodiazepines). It is important to recognize and treat depression; treatment can begin with a low dose of an SSRI (e.g., escitalopram, starting dose 5 mg daily, target dose 5–10 mg daily) while monitoring for efficacy and toxicity. Sometimes apathy, visual hallucinations, depression, and other psychiatric symptoms respond to the cholinesterase inhibitors, especially in DLB, obviating the need for other more toxic therapies.

Cholinesterase inhibitors are being used to treat AD (donepezil, rivastigmine, galantamine) and PDD (rivastigmine). Recent work has focused on developing antibodies against $A\beta_{42}$ as a treatment for AD.

Although the initial randomized controlled trials failed, there was some evidence for efficacy in the mildest patient groups. Therefore, researchers have begun to focus on patients with very mild disease and asymptomatic individuals at risk for AD, such as those who carry autosomal dominantly inherited genetic mutations or healthy elders with CSF or amyloid imaging biomarker evidence supporting presymptomatic AD. Memantine proves useful when treating some patients with moderate to severe AD; its major benefit relates to decreasing caregiver burden, most likely by decreasing resistance to dressing and grooming support. In moderate to severe AD, the combination of memantine and a cholinesterase inhibitor delayed nursing home placement in several studies, although other studies have not supported the efficacy of adding memantine to the regimen.

A proactive strategy has been shown to reduce the occurrence of delirium in hospitalized patients. This strategy includes frequent orientation, cognitive activities, sleep-enhancement measures, vision and hearing aids, and correction of dehydration.

Nondrug behavior therapy has an important place in dementia management. The primary goals are to make the patient's life comfortable, uncomplicated, and safe. Preparing lists, schedules, calendars, and labels can be helpful in the early stages. It is also useful to stress familiar routines, walks, and simple physical exercises. For many demented patients, memory for events is worse than their ability to carry out routine activities, and they may still be able to take part in activities such as walking, bowling, dancing, singing, bingo, and golf. Demented patients often object to losing control over familiar tasks such as driving, cooking, and handling finances. Attempts to help or take over may be greeted with complaints, depression, or anger. Hostile responses on the part of the caregiver are counterproductive and sometimes even harmful. Reassurance, distraction, and calm positive statements are more productive in this setting. Eventually, tasks such as finances and driving must be assumed by others, and the patient will conform and adjust. Safety is an important issue that includes not only driving but controlling the kitchen, bathroom, and sleeping area environments, as well as stairways. These areas need to be monitored, supervised, and made as safe as possible. A move to a retirement complex, assisted-living center, or nursing home can initially increase confusion and agitation. Repeated reassurance, reorientation, and careful introduction to the new personnel will help to smooth the process. Providing activities that are known to be enjoyable to the patient can be of considerable benefit.

The clinician must pay special attention to frustration and depression among family members and caregivers. Caregiver guilt and burnout are common. Family members often feel overwhelmed and helpless and may vent their frustrations on the patient, each other, and health care providers. Caregivers should be encouraged to take advantage of day-care facilities and respite services. Education and counseling about dementia are important. Local and national support groups, such as the Alzheimer's Association (*www.alz.org*), can provide considerable help.

FURTHER READING

Barton C et al: Non-pharmacological management of behavioral symptoms in frontotemporal and other dementias. Curr Neurol Neurosci Rep 16:14, 2016. [PubMed: 26750129]

Griem J et al: Psychologic/functional forms of memory disorder. Handb Clin Neurol 139:407, 2017.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 26: Aphasia, Memory Loss, Hemispatial Neglect, Frontal Syndromes, and Other Cerebral Disorders

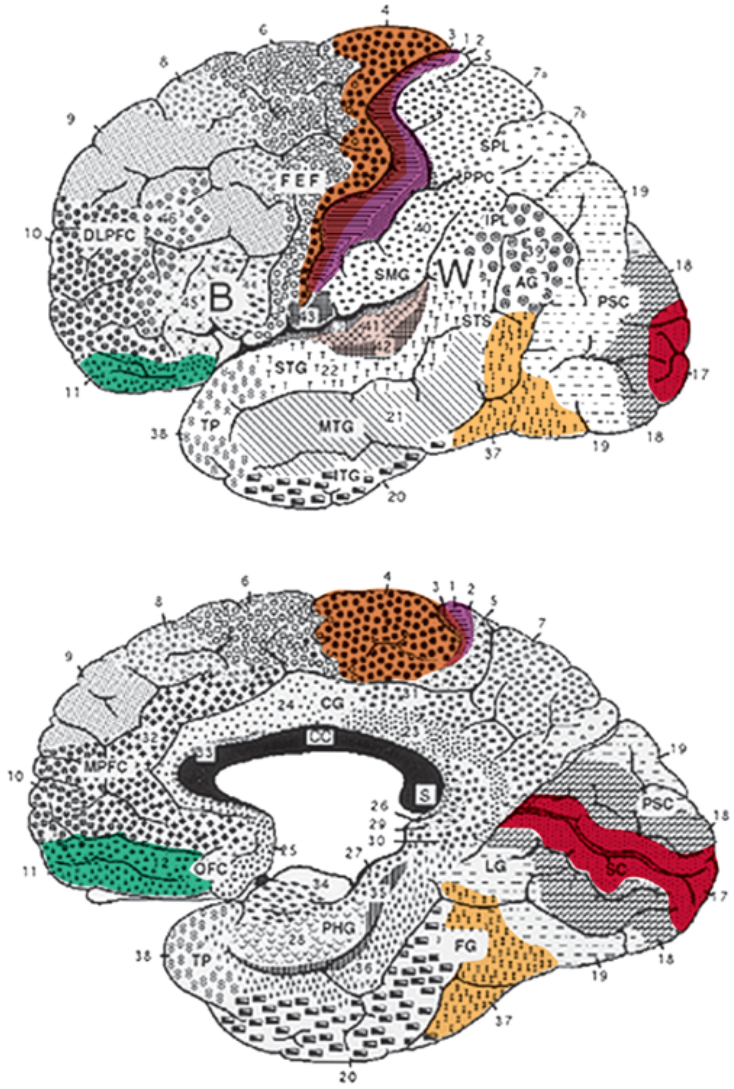
M.-Marsel Mesulam

INTRODUCTION

The cerebral cortex of the human brain contains ~20 billion neurons spread over an area of 2.5 m^2. The primary sensory and motor areas constitute 10% of the cerebral cortex. The rest is subsumed by modality-selective, heteromodal, paralimbic, and limbic areas collectively known as the *association cortex* (Fig. 26-1). The association cortex mediates the integrative processes that subserve cognition, emotion, and comportment. A systematic testing of these mental functions is necessary for the effective clinical assessment of the association cortex and its diseases. According to current thinking, there are no centers for "hearing words," "perceiving space," or "storing memories." Cognitive and behavioral functions (domains) are coordinated by intersecting *large-scale neural networks* that contain interconnected cortical and subcortical components. Five anatomically defined *large-scale networks* are most relevant to clinical practice: (1) a left-dominant perisylvian network for language, (2) a right-dominant parietofrontal network for spatial orientation, (3) an occipitotemporal network for face and object recognition, (4) a limbic network for explicit episodic memory, and (5) a prefrontal network for the executive control of cognition and comportment. Investigations based on functional imaging have also identified a *default mode network*, which becomes activated when the person is not engaged in a specific task requiring attention to external events. The clinical consequences of damage to this network are not yet fully defined.

FIGURE 26-1

Lateral (*top*) and medial (*bottom*) views of the cerebral hemispheres. The numbers refer to the Brodmann cyto-architectonic designations. Area 17 corresponds to the primary visual cortex, 41–42 to the primary auditory cortex, 1–3 to the primary somatosensory cortex, and 4 to the primary motor cortex. The rest of the cerebral cortex contains association areas. AG, angular gyrus; B, Broca's area; CC, corpus callosum; CG, cingulate gyrus; DLPFC, dorsolateral prefrontal cortex; FEF, frontal eye fields (premotor cortex); FG, fusiform gyrus; IPL, inferior parietal lobule; ITG, inferior temporal gyrus; LG, lingual gyrus; MPFC, medial prefrontal cortex; MTG, middle temporal gyrus; OFC, orbitofrontal cortex; PHG, parahippocampal gyrus; PPC, posterior parietal cortex; PSC, peristriate cortex; SC, striate cortex; SMG, supramarginal gyrus; SPL, superior parietal lobule; STG, superior temporal gyrus; STS, superior temporal sulcus; TP, temporopolar cortex; W, Wernicke's area.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE LEFT PERISYLVIAN NETWORK FOR LANGUAGE AND APHASIAS

The production and comprehension of words and sentences is dependent on the integrity of a distributed network located along the perisylvian region of the language-dominant (usually left) hemisphere. One hub, situated in the inferior frontal gyrus, is known as *Broca's area*. Damage to this region impairs fluency of verbal output and the grammatical structure of sentences. The location of a second hub, critical for language comprehension, is less clearly settled. Accounts of patients with focal cerebrovascular lesions identified *Wernicke's area*, located at the parietotemporal junction, as a critical hub for word and sentence comprehension. Occlusive or embolic strokes involving this area interfere with the ability to understand spoken or written language as well as the ability to express thoughts through meaningful words and statements. However, investigations of patients with the neurodegenerative syndrome of primary progressive aphasia (PPA) have shown that sentence comprehension is a widely distributed faculty jointly subserved by Broca's and Wernicke's areas, and that the areas critical for word comprehension are more closely associated with the anterior temporal lobe rather than Wernicke's area. All components of the language network are interconnected with each other and with surrounding parts of the frontal, parietal, and temporal lobes. Damage to this network gives rise to language impairments known as aphasia. Aphasia should be diagnosed only when there are deficits in the formal aspects of language, such as word finding, word choice, comprehension, spelling, or grammar. Dysarthria, apraxia of speech and mutism do not by themselves lead to a diagnosis of aphasia. In ~90% of right-handers and 60% of left-handers, aphasia occurs only after lesions of the left hemisphere.

CLINICAL EXAMINATION

The clinical examination of language should include the assessment of naming, spontaneous speech, comprehension, repetition, reading, and writing. A deficit of naming (*anomia*) is the single most common finding in aphasic patients. When asked to name a common object, the patient may fail to come up with the appropriate word, may provide a circumlocutious description of the object ("the thing for writing"), or may come up with the wrong word (*paraphasia*). If the patient offers an incorrect but related word ("pen" for "pencil"), the naming error is known as a *semantic paraphasia*; if the word

approximates the correct answer but is phonetically inaccurate ("plentil" for "pencil"), it is known as a *phonemic paraphasia*. In most anomias, the patient cannot retrieve the appropriate name when shown an object but can point to the appropriate object when the name is provided by the examiner. This is known as a one-way (or retrieval-based) naming deficit. A two-way (comprehension-based or semantic) naming deficit exists if the patient can neither provide nor recognize the correct name. *Spontaneous speech* is described as "fluent" if it maintains appropriate output volume, phrase length, and melody or as "nonfluent" if it is sparse and halting and average utterance length is below four words. The examiner also should note the integrity of *grammar* as manifested by word order (syntax), tenses, suffixes, prefixes, plurals, and possessives. *Comprehension* can be tested by assessing the patient's ability to follow conversation, asking yes-no questions ("Can a dog fly?" "Does it snow in summer?"), asking the patient to point to appropriate objects ("Where is the source of illumination in this room?"), or asking for verbal definitions of single words. *Repetition* is assessed by asking the patient to repeat single words, short sentences, or strings of words such as "No ifs, ands, or buts." The testing of repetition with tongue twisters such as "hippopotamus" and "Irish constabulary" provides a better assessment of dysarthria and palilalia than of aphasia. It is important to make sure that the number of words does not exceed the patient's attention span. Otherwise, the failure of repetition becomes a reflection of the narrowed attention span (auditory working memory) rather than an indication of an aphasic deficit caused by dysfunction of a hypothetical *phonological loop* in the language network. *Reading* should be assessed for deficits in reading aloud as well as comprehension. *Alexia* describes an inability to either read aloud or comprehend written words and sentences; *agraphia* (or dysgraphia) is used to describe an acquired deficit in spelling.

Aphasias can arise acutely in cerebrovascular accidents (CVAs) or gradually in neurodegenerative diseases. In CVAs damage encompasses cerebral cortex as well as deep white matter pathways interconnecting otherwise unaffected cortical areas. The syndromes listed in **Table 26-1** are most applicable to this group, where gray matter and white matter at the lesion site are abruptly and jointly destroyed. Progressive neurodegenerative diseases can have cellular, laminar, and regional specificity for the cerebral cortex, giving rise to a different set of aphasias that will be described separately.

TABLE 26-1

Clinical Features of Aphasias and Related Conditions Commonly Seen in Cerebrovascular Accidents

	Comprehension	Repetition of Spoken Language	Naming	Fluency
Wernicke's	Impaired	Impaired	Impaired	Preserved or increased
Broca's	Preserved (except grammar)	Impaired	Impaired	Decreased
Global	Impaired	Impaired	Impaired	Decreased
Conduction	Preserved	Impaired	Impaired	Preserved
Nonfluent (anterior) transcortical	Preserved	Preserved	Impaired	Impaired
Fluent (posterior) transcortical	Impaired	Preserved	Impaired	Preserved
Isolation	Impaired	Echolalia	Impaired	No purposeful speech
Anomic	Preserved	Preserved	Impaired	Preserved except for word-finding pauses
Pure word deafness	Impaired only for spoken language	Impaired	Preserved	Preserved
Pure alexia	Impaired only for reading	Preserved	Preserved	Preserved

Wernicke's Aphasia

Comprehension is impaired for spoken and written words and sentences. Language output is fluent but is highly paraphasic and circumlocutious. Paraphasic errors may lead to strings of neologisms, which lead to "jargon aphasia." Speech contains few substantive nouns. The output is therefore voluminous but uninformative. For example, a patient attempts to describe how his wife accidentally threw away something important, perhaps his dentures: "We don't need it anymore, she says. And with it when that was downstairs was my teeth-tick … a … den … dentith … my dentist. And they

happened to be in that bag … see? …Where my two … two little pieces of dentist that I use … that I … all gone. If she throws the whole thing away … visit some friends of hers and she can't throw them away."

Gestures and pantomime do not improve communication. The patient may not realize that his or her language is incomprehensible and may appear angry and impatient when the examiner fails to decipher the meaning of a severely paraphasic statement. In some patients this type of aphasia can be associated with severe agitation and paranoia. The ability to follow commands aimed at axial musculature may be preserved. The dissociation between the failure to understand simple questions ("What is your name?") in a patient who rapidly closes his or her eyes, sits up, or rolls over when asked to do so is characteristic of Wernicke's aphasia and helps differentiate it from deafness, psychiatric disease, or malingering. Patients with Wernicke's aphasia cannot express their thoughts in meaning-appropriate words and cannot decode the meaning of words in any modality of input. This aphasia therefore has expressive as well as receptive components. Repetition, naming, reading, and writing also are impaired.

The lesion site most commonly associated with Wernicke's aphasia caused by CVAs is the posterior portion of the language network. An embolus to the inferior division of the middle cerebral artery (MCA), to the posterior temporal or angular branches in particular, is the most common etiology **(Chap. 419)**. Intracerebral hemorrhage, head trauma, and neoplasm are other causes of Wernicke's aphasia. A coexisting right hemianopia or superior quadrantanopia is common, and mild right nasolabial flattening may be found, but otherwise, the examination is often unrevealing. The paraphasic, neologistic speech in an agitated patient with an otherwise unremarkable neurologic examination may lead to the suspicion of a primary psychiatric disorder such as schizophrenia or mania, but the other components characteristic of acquired aphasia and the absence of prior psychiatric disease usually settle the issue. Prognosis for recovery of language function is guarded.

Broca's Aphasia

Speech is nonfluent, labored, interrupted by many word-finding pauses, and usually dysarthric. It is impoverished in function words but enriched in meaning-appropriate nouns. Abnormal word order and the inappropriate deployment of *bound morphemes* (word endings used to denote tenses, possessives, or plurals) lead to a characteristic agrammatism. Speech is telegraphic and pithy but quite informative. In the following passage, a patient with Broca's aphasia describes his medical history: "I see … the dotor, dotor sent me … Bosson. Go to hospital. Dotor … kept me beside. Two, tee days, doctor send me home."

Output may be reduced to a grunt or single word ("yes" or "no"), which is emitted with different intonations in an attempt to express approval or disapproval. In addition to fluency, naming and repetition are impaired. Comprehension of spoken language is intact except for syntactically difficult sentences with a passive voice structure or embedded clauses, indicating that Broca's aphasia is not just an "expressive" or "motor" disorder and that it also may involve a comprehension deficit in decoding syntax. Patients with Broca's aphasia can be tearful, easily frustrated, and profoundly depressed. Insight into their condition is preserved, in contrast to Wernicke's aphasia. Even when spontaneous speech is severely dysarthric, the patient may be able to display a relatively normal articulation of words when singing. This dissociation has been used to develop specific therapeutic approaches (melodic intonation therapy) for Broca's aphasia. Additional neurologic deficits include right facial weakness, hemiparesis or hemiplegia, and a buccofacial apraxia characterized by an inability to carry out motor commands involving oropharyngeal and facial musculature (e.g., patients are unable to demonstrate how to blow out a match or suck through a straw). The cause is most often infarction of Broca's area (the inferior frontal convolution; "B" in Fig. 26-1) and surrounding anterior perisylvian and insular cortex due to occlusion of the superior division of the MCA **(Chap. 419)**. Mass lesions, including tumor, intracerebral hemorrhage, and abscess, also may be responsible. When the cause of Broca's aphasia is stroke, recovery of language function generally peaks within 2–6 months, after which time further progress is limited. Speech therapy is more successful than in Wernicke's aphasia.

Conduction Aphasia

Speech output is fluent but contains many phonemic paraphasias, comprehension of spoken language is intact, and repetition is severely impaired. Naming elicits phonemic paraphasias, and spelling is impaired. Reading aloud is impaired, but reading comprehension is preserved. The responsible lesion, usually a CVA in the temporoparietal or dorsal perisylvian region, interferes with the function of the phonological loop interconnecting Broca's area with Wernicke's area. Occasionally, a transient Wernicke's aphasia may rapidly resolve into a conduction aphasia. The paraphasic output in conduction aphasia interferes with the ability to express meaning, but this deficit is not nearly as severe as the one displayed by patients with Wernicke's aphasia. Associated neurologic signs in conduction aphasia vary according to the primary lesion site.

Transcortical Aphasias: Fluent and Nonfluent

Clinical features of *fluent (posterior) transcortical aphasia* are similar to those of Wernicke's aphasia, but repetition is intact. The lesion site disconnects the intact core of the language network from other temporoparietal association areas. Associated neurologic findings may include hemianopia. Cerebrovascular lesions (e.g., infarctions in the posterior watershed zone) and neoplasms that involve the temporoparietal cortex posterior to Wernicke's area are common causes. The features of *nonfluent (anterior) transcortical aphasia* are similar to those of Broca's aphasia, but repetition is intact and agrammatism is less pronounced. The neurologic examination may be otherwise intact, but a right hemiparesis also can exist. The lesion site

disconnects the intact language network from prefrontal areas of the brain and usually involves the anterior watershed zone between anterior and MCA territories or the supplementary motor cortex in the territory of the anterior cerebral artery.

Global and Isolation Aphasias

Global aphasia represents the combined dysfunction of Broca's and Wernicke's areas and usually results from strokes that involve the entire MCA distribution in the left hemisphere. Speech output is nonfluent, and comprehension of language is severely impaired. Related signs include right hemiplegia, hemisensory loss, and homonymous hemianopia. *Isolation aphasia* represents a combination of the two transcortical aphasias. Comprehension is severely impaired, and there is no purposeful speech output. The patient may parrot fragments of heard conversations (*echolalia*), indicating that the neural mechanisms for repetition are at least partially intact. This condition represents the pathologic function of the language network when it is isolated from other regions of the brain. Broca's and Wernicke's areas tend to be spared, but there is damage to the surrounding frontal, parietal, and temporal cortex. Lesions are patchy and can be associated with anoxia, carbon monoxide poisoning, or complete watershed zone infarctions.

Anomic Aphasia

This form of aphasia may be considered the "minimal dysfunction" syndrome of the language network. Articulation, comprehension, and repetition are intact, but confrontation naming, word finding, and spelling are impaired. Word-finding pauses are uncommon, so language output is fluent but paraphasic, circumlocutious, and uninformative. The lesion sites can be anywhere within the left hemisphere language network, including the middle and inferior temporal gyri. *Anomic aphasia is the single most common language disturbance seen in head trauma, metabolic encephalopathy, and Alzheimer's disease.*

Pure Word Deafness

The most common causes are either bilateral or left-sided MCA strokes affecting the superior temporal gyrus. The net effect of the underlying lesion is to interrupt the flow of information from the auditory association cortex to the language network. Patients have no difficulty understanding written language and can express themselves well in spoken or written language. They have no difficulty interpreting and reacting to environmental sounds if the primary auditory cortex and auditory association areas of the right hemisphere are spared. Because auditory information cannot be conveyed to the language network, however, it cannot be decoded into neural word representations, and the patient reacts to speech as if it were in an alien tongue that cannot be deciphered. Patients cannot repeat spoken language but have no difficulty naming objects. In time, patients with pure word deafness teach themselves lipreading and may appear to have improved. There may be no additional neurologic findings, but agitated paranoid reactions are common in the acute stages. Cerebrovascular lesions are the most common cause.

Pure Alexia Without Agraphia

This is the visual equivalent of pure word deafness. The lesions (usually a combination of damage to the left occipital cortex and to a posterior sector of the corpus callosum—the splenium) interrupt the flow of visual input into the language network. There is usually a right hemianopia, but the core language network remains unaffected. The patient can understand and produce spoken language, name objects in the left visual hemifield, repeat, and write. However, the patient acts as if illiterate when asked to read even the simplest sentence because the visual information from the written words (presented to the intact left visual hemifield) cannot reach the language network. Objects in the left hemifield may be named accurately because they activate nonvisual associations in the right hemisphere, which in turn can access the language network through transcallosal pathways anterior to the splenium. Patients with this syndrome also may lose the ability to name colors, although they can match colors. This is known as a *color anomia*. The most common etiology of pure alexia is a vascular lesion in the territory of the posterior cerebral artery or an infiltrating neoplasm in the left occipital cortex that involves the optic radiations as well as the crossing fibers of the splenium. Because the posterior cerebral artery also supplies medial temporal components of the limbic system, a patient with pure alexia also may experience an amnesia, but this is usually transient because the limbic lesion is unilateral.

Apraxia and Aphemia

Apraxia designates a complex motor deficit that cannot be attributed to pyramidal, extrapyramidal, cerebellar, or sensory dysfunction and that does not arise from the patient's failure to understand the nature of the task. *Apraxia of speech* is used to designate articulatory abnormalities in the duration, fluidity, and stress of syllables that make up words. It can arise with CVAs in the posterior part of Broca's area or in the course of frontotemporal lobar degeneration (FTLD) with tauopathy. *Aphemia* is a severe form of acute speech apraxia that presents with severely impaired fluency (often mutism). Recovery is the rule and involves an intermediate stage of hoarse whispering. Writing, reading, and comprehension are intact, and so this is not a true aphasic syndrome. CVAs in parts of Broca's area or subcortical lesions that undercut its connections with other parts of the brain may be present. Occasionally, the lesion site is on the medial aspects of the frontal lobes and may involve the supplementary motor cortex of the left hemisphere. *Ideomotor apraxia* is diagnosed when commands to perform a specific motor act ("cough," "blow out a match") or pantomime the use of a common tool (a comb, hammer, straw, or toothbrush) in the absence of the real object cannot be followed. The patient's ability to comprehend the command is

ascertained by demonstrating multiple movements and establishing that the correct one can be recognized. Some patients with this type of apraxia can imitate the appropriate movement when it is demonstrated by the examiner and show no impairment when handed the real object, indicating that the sensorimotor mechanisms necessary for the movement are intact. Some forms of ideomotor apraxia represent a disconnection of the language network from pyramidal motor systems so that commands to execute complex movements are understood but cannot be conveyed to the appropriate motor areas. *Buccofacial apraxia* involves apraxic deficits in movements of the face and mouth. Ideomotor *limb apraxia* encompasses apraxic deficits in movements of the arms and legs. Ideomotor apraxia almost always is caused by lesions in the left hemisphere and is commonly associated with aphasic syndromes, especially Broca's aphasia and conduction aphasia. Because the handling of real objects is not impaired, ideomotor apraxia by itself causes no major limitation of daily living activities. Patients with lesions of the anterior corpus callosum can display ideomotor apraxia confined to the left side of the body, a sign known as *sympathetic dyspraxia*. A severe form of sympathetic dyspraxia, known as the *alien hand* syndrome, is characterized by additional features of motor disinhibition on the left hand. *Ideational apraxia* refers to a deficit in the sequencing of goal-directed movements in patients who have no difficulty executing the individual components of the sequence. For example, when the patient is asked to pick up a pen and write, the sequence of uncapping the pen, placing the cap at the opposite end, turning the point toward the writing surface, and writing may be disrupted, and the patient may be seen trying to write with the wrong end of the pen or even with the removed cap. These motor sequencing problems usually are seen in the context of confusional states and dementias rather than focal lesions associated with aphasic conditions. *Limb-kinetic apraxia* involves clumsiness in the use of tools or objects that cannot be attributed to sensory, pyramidal, extrapyramidal, or cerebellar dysfunction. This condition can emerge in the context of focal premotor cortex lesions or *corticobasal degeneration* and can interfere with the use of tools and utensils.

Gerstmann's Syndrome

The combination of *acalculia* (impairment of simple arithmetic), *dysgraphia* (impaired writing), *finger anomia* (an inability to name individual fingers such as the index and thumb), and *right-left confusion* (an inability to tell whether a hand, foot, or arm of the patient or examiner is on the right or left side of the body) is known as Gerstmann's syndrome. In making this diagnosis, it is important to establish that the finger and left-right naming deficits are not part of a more generalized anomia and that the patient is not otherwise aphasic. When Gerstmann's syndrome arises acutely and in isolation, it is commonly associated with damage to the inferior parietal lobule (especially the angular gyrus) in the left hemisphere.

Pragmatics and Prosody

Pragmatics refers to aspects of language that communicate attitude, affect, and the figurative rather than literal aspects of a message (e.g., "green thumb" does not refer to the actual color of the finger). One component of pragmatics, *prosody,* refers to variations of melodic stress and intonation that influence attitude and the inferential aspect of verbal messages. For example, the two statements "He *is* clever." and "He is *clever*?" contain an identical word choice and syntax but convey vastly different messages because of differences in the intonation with which the statements are uttered. Damage to right hemisphere regions corresponding to Broca's area impairs the ability to introduce meaning-appropriate prosody into spoken language. The patient produces grammatically correct language with accurate word choice, but the statements are uttered in a monotone that interferes with the ability to convey the intended stress and effect. Patients with this type of *aprosodia* give the mistaken impression of being depressed or indifferent. Other aspects of pragmatics, especially the ability to infer the figurative aspect of a message, become impaired by damage to the right hemisphere or frontal lobes.

Subcortical Aphasia

Damage to subcortical components of the language network (e.g., the striatum and thalamus of the left hemisphere) also can lead to aphasia. The resulting syndromes contain combinations of deficits in the various aspects of language but rarely fit the specific patterns described in Table 26-1. In a patient with a CVA, an anomic aphasia accompanied by dysarthria or a fluent aphasia with hemiparesis should raise the suspicion of a subcortical lesion site.

CLINICAL PRESENTATION AND DIAGNOSIS OF PPA

Aphasias caused by CVAs start suddenly and display maximal deficits at the onset. These are the "classic" aphasias described above. Aphasias caused by neurodegenerative diseases have an insidious onset and relentless progression. The neuropathology can be selective not only for gray matter but also for specific layers and cell types. The clinico-anatomic patterns are therefore different from those described in Table 26-1.

Several neurodegenerative syndromes, such as typical Alzheimer-type (amnestic; **Chap. 423**) and frontotemporal (behavioral; **Chap. 424**) dementias, can also include language impairments as the disease progresses. In these cases, the aphasia is an ancillary component of the overall syndrome. A diagnosis of PPA is justified only if the language disorder (i.e., aphasia) arises in relative isolation, becomes the primary concern that brings the patient to medical attention, and remains the most salient deficit for 1–2 years. PPA can be caused by either FTLD or Alzheimer's disease (AD) pathology. Rarely, an identical syndrome can be caused by Creutzfeldt-Jacob disease (CJD) but with a more rapid progression **(Chap. 430)**.

LANGUAGE IN PPA

The impairments of language in PPA have slightly different patterns from those seen in CVA-caused aphasias. For example, the full syndrome of Wernicke's aphasia is almost never seen in PPA, confirming the view that sentence comprehension and word comprehension are controlled by different regions of the language network. Three major subtypes of PPA can be recognized.

Agrammatic PPA

The *agrammatic variant* is characterized by consistently low fluency and impaired grammar but intact word comprehension. It most closely resembles Broca's aphasia or anterior transcortical aphasia but usually lacks the right hemiparesis or dysarthria and may have more profound impairments of grammar. Peak sites of neuronal loss (gray matter atrophy) include the left inferior frontal gyrus where Broca's area is located. The neuropathology is usually a FTLD with tauopathy but can also be an atypical form of AD pathology.

Semantic PPA

The *semantic variant* is characterized by preserved fluency and syntax but poor single-word comprehension and profound two-way naming impairments. This kind of aphasia is not seen with CVAs. It differs from Wernicke's aphasia or posterior transcortical aphasia because speech is usually informative and repetition is intact. Comprehension of sentences is relatively preserved if the meaning is not too dependent on words that fail to be understood allowing the patient to surmise the gist of the conversation through contextual cues. Such patients may appear unimpaired in the course of casual small talk but become puzzled upon encountering an undecipherable word such as "pumpkin" or "umbrella." Peak atrophy sites are located in the left anterior temporal lobe, indicating that this part of the brain plays a critical role in the comprehension of words, especially words that denote concrete objects. This is a part of the brain that was not included within the classic language network, probably because it is not a common site for focal CVAs. The neuropathology is frequently an FTLD with abnormal precipitates of the 43-kDa transactive response DNA-binding protein TDP-43 of type C.

Logopenic PPA

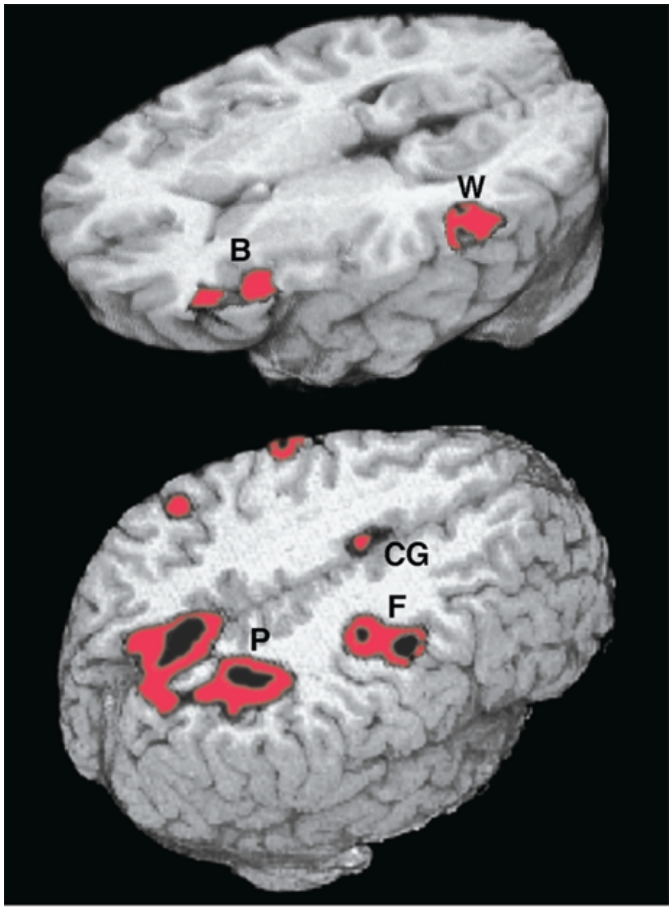
The *logopenic variant* is characterized by preserved syntax and comprehension but frequent and severe word-finding pauses, anomia, circumlocutions, and simplifications during spontaneous speech. Repetition is usually impaired. Peak atrophy sites are located in the temporoparietal junction and posterior temporal lobe, partially overlapping with traditional location of Wernicke's area. However, the comprehension impairment of *Wernicke's aphasia* is absent probably because the underlying deep white matter, frequently damaged by CVAs, remains relatively intact in PPA. The repetition impairment suggests that parts of Wernicke's area are critical for phonological loop functionality. In contrast to Broca's aphasia or agrammatic PPA, the interruption of fluency is variable so that speech may appear entirely normal if the patient is allowed to engage in small talk. Logopenic PPA resembles the anomic aphasia of Table 26-1 but usually has longer and more frequent word-finding pauses. When repetition is impaired the aphasia resembles the *conduction aphasia* in Table 26-1. Of all PPA subtypes, this is the one most commonly associated with the pathology of AD, but FTLD can also be the cause. In addition to these three major subtypes, there is also a *mixed* type of PPA where grammar, fluency and word comprehension are jointly impaired. This is most like the global aphasia of Table 26-1. Rarely, PPA can present with patterns reminiscent of *pure word deafness* or *Gerstmann's syndrome*.

THE PARIETOFRONTAL NETWORK FOR SPATIAL ORIENTATION

Adaptive spatial orientation is subserved by a large-scale network containing three major cortical components. The *cingulate cortex* provides access to a motivational mapping of the extrapersonal space, the *posterior parietal cortex* to a sensorimotor representation of salient extrapersonal events, and the *frontal eye fields* to motor strategies for attentional behaviors **(Fig. 26-2)**. Subcortical components of this network include the striatum and the thalamus. Damage to this network can undermine the distribution of attention within the extrapersonal space, giving rise to hemispatial neglect, simultanagnosia and object finding failures. The integration of egocentric (self-centered) with allocentric (object-centered) coordinates can also be disrupted, giving rise to impairments in route finding, the ability to avoid obstacles, and the ability to dress.

FIGURE 26-2

Functional magnetic resonance imaging of language and spatial attention in neurologically intact subjects. The red and black areas show regions of task-related significant activation. (*Top*) The subjects were asked to determine if two words were synonymous. This language task led to the simultaneous activation of the two components of the language network, Broca's area (B) and Wernicke's area (W). The activations are exclusively in the left hemisphere. (*Bottom*) The subjects were asked to shift spatial attention to a peripheral target. This task led to the simultaneous activation of the three epicenters of the attentional network: the posterior parietal cortex (P), the frontal eye fields (F), and the cingulate gyrus (CG). The activations are predominantly in the right hemisphere. *(Courtesy of Darren Gitelman, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HEMISPATIAL NEGLECT

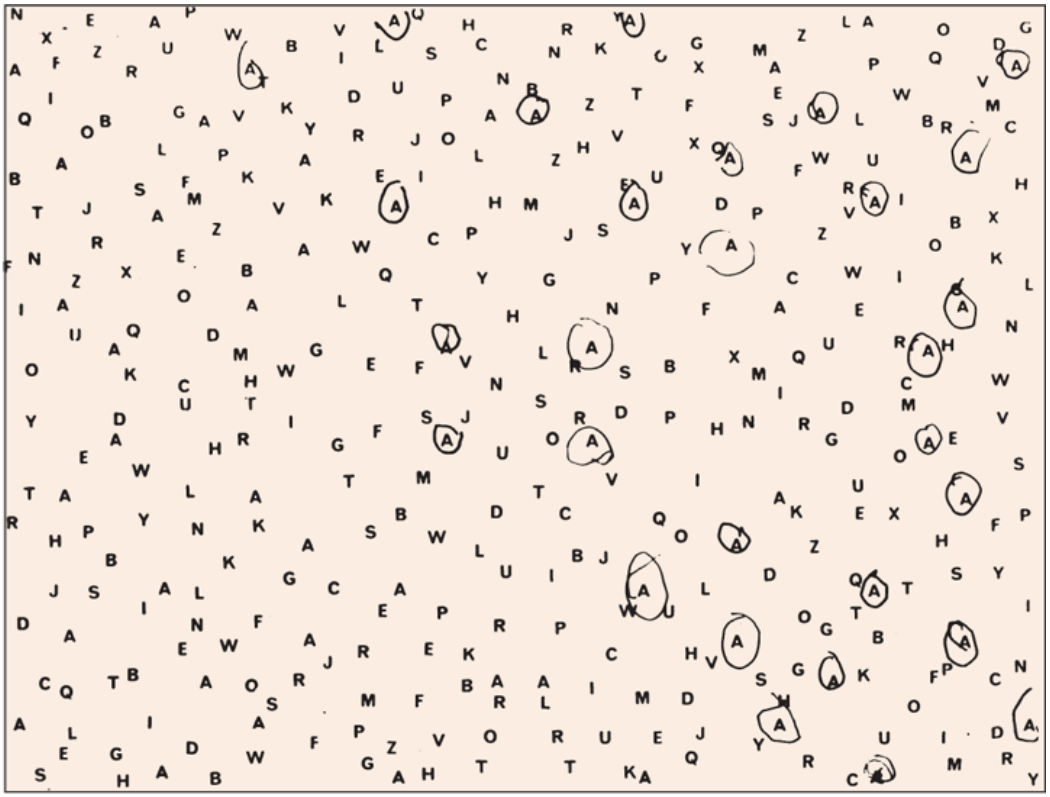
Contralesional hemispatial neglect represents one outcome of damage to the cortical or subcortical components of this network. *The traditional view that hemispatial neglect always denotes a parietal lobe lesion is inaccurate*. According to one model of spatial cognition, the right hemisphere directs attention within the *entire* extrapersonal space, whereas the left hemisphere directs attention mostly within the contralateral right hemispace. Consequently, left hemisphere lesions do not give rise to much contralesional neglect because the global attentional mechanisms of the right hemisphere can compensate for the loss of the *contralaterally* directed attentional functions of the left hemisphere. Right hemisphere lesions, however, give rise to severe contralesional left hemispatial neglect because the unaffected left hemisphere does not contain ipsilateral attentional mechanisms. This model is consistent with clinical experience, which shows that contralesional neglect is more common, more severe, and longer lasting after damage to the right hemisphere than after damage to the left hemisphere. Severe neglect for the right hemispace is rare, even in left-handers with left hemisphere lesions.

Clinical Examination

Patients with severe neglect may fail to dress, shave, or groom the left side of the body; fail to eat food placed on the left side of the tray; and fail to read the left half of sentences. When asked to copy a simple line drawing, the patient fails to copy detail on the left, and when the patient is asked to write, there is a tendency to leave an unusually wide margin on the left. Two bedside tests that are useful in assessing neglect are *simultaneous bilateral stimulation* and *visual target cancellation*. In the former, the examiner provides either unilateral or simultaneous bilateral stimulation in the visual, auditory, and tactile modalities. After right hemisphere injury, patients who have no difficulty detecting unilateral stimuli on either side experience the bilaterally presented stimulus as coming only from the right. This phenomenon is known as *extinction* and is a manifestation of the sensory-representational aspect of hemispatial neglect. In the target detection task, targets (e.g., A's) are interspersed with foils (e.g., other letters of the alphabet) on a 21.5- to 28.0-cm (8.5–11 in.) sheet of paper, and the patient is asked to circle all the targets. A failure to detect targets on the left is a manifestation of the exploratory (motor) deficit in hemispatial neglect (Fig. 26-3A). Hemianopia is not by itself sufficient to cause the target detection failure because the patient is free to turn the head and eyes to the left. Target detection failures therefore reflect a distortion of spatial attention, not just of sensory input. Some patients with neglect also may deny the existence of hemiparesis and may even deny ownership of the paralyzed limb, a condition known as *anosognosia*.

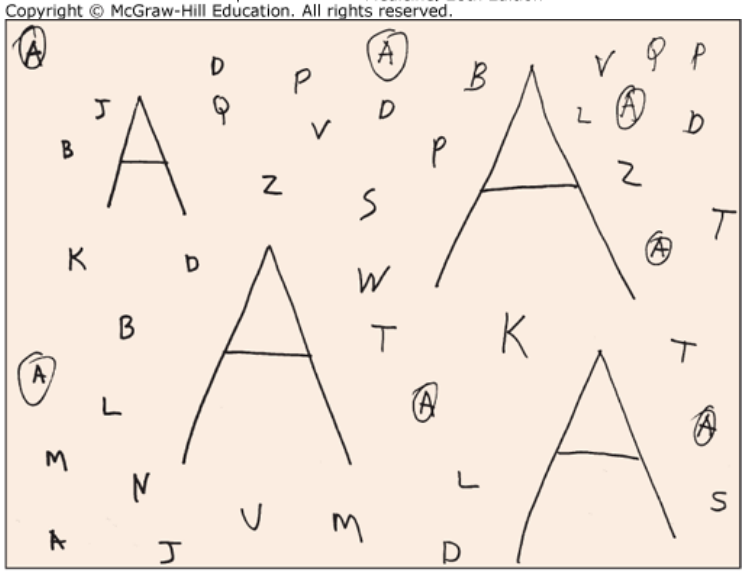
FIGURE 26-3

A. A 47-year-old man with a large frontoparietal lesion in the right hemisphere was asked to circle all the 'A's. Only targets on the right are circled. This is a manifestation of left hemispatial neglect. ***B.*** A 70-year-old woman with a 2-year history of degenerative dementia was able to circle most of the small targets but ignored the larger ones. This is a manifestation of simultanagnosia.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

BÁLINT'S SYNDROME, SIMULTANAGNOSIA, DRESSING APRAXIA, CONSTRUCTION APRAXIA, AND ROUTE FINDING IMPAIRMENTS

Bilateral involvement of the network for spatial attention, especially its parietal components, leads to a state of severe spatial disorientation known as *Bálint's syndrome*. Bálint's syndrome involves deficits in the orderly visuomotor scanning of the environment (*oculomotor apraxia*), accurate manual reaching toward visual targets (*optic ataxia*), and the ability to integrate visual information in the center of gaze with more peripheral information (*simultanagnosia*). A patient with simultanagnosia "misses the forest for the trees." For example, a patient who is shown a table lamp and asked to

name the object may look at its circular base and call it an ashtray. Some patients with simultanagnosia report that objects they look at may vanish suddenly, probably indicating an inability to compute the oculomotor return to the original point of gaze after brief saccadic displacements. Movement and distracting stimuli greatly exacerbate the difficulties of visual perception. Simultanagnosia can occur without the other two components of Bálint's syndrome especially in association with Alzheimer's disease.

A modification of the letter cancellation task described above can be used for the bedside diagnosis of simultanagnosia. In this modification, some of the targets (e.g., A's) are made to be much larger than the others (7.5–10 cm vs 2.5 cm [3–4 in. vs 1 in.] in height), and all targets are embedded among foils. Patients with simultanagnosia display a counterintuitive but characteristic tendency to miss the larger targets **(Fig. 26-3B)**. This occurs because the information needed for the identification of the larger targets cannot be confined to the immediate line of gaze and requires the integration of visual information across multiple fixation points. The greater difficulty in the detection of the larger targets also indicates that poor acuity is not responsible for the impairment of visual function and that the problem is central rather than peripheral. The test shown in Fig. 26-3B is not by itself sufficient to diagnose simultanagnosia because some patients with a frontal network syndrome may omit the strange looking large letters, perhaps because they lack the mental flexibility needed to realize that the two types of targets are symbolically identical despite being superficially different.

Bilateral parietal lesions can impair the integration of egocentric with allocentric spatial coordinates. One manifestation is *dressing apraxia*. A patient with this condition is unable to align the body axis with the axis of the garment and can be seen struggling as he or she holds a coat from its bottom or extends his or her arm into a fold of the garment rather than into its sleeve. Lesions that involve the posterior parietal cortex also lead to severe difficulties in copying simple line drawings. This is known as a *construction apraxia* and is much more severe if the lesion is in the right hemisphere. In some patients with right hemisphere lesions, the drawing difficulties are confined to the left side of the figure and represent a manifestation of hemispatial neglect; in others, there is a more universal deficit in reproducing contours and three-dimensional perspective. Impairments of route finding can be included in this group of disorders, which reflect an inability to orient the self with respect to external objects and landmarks.

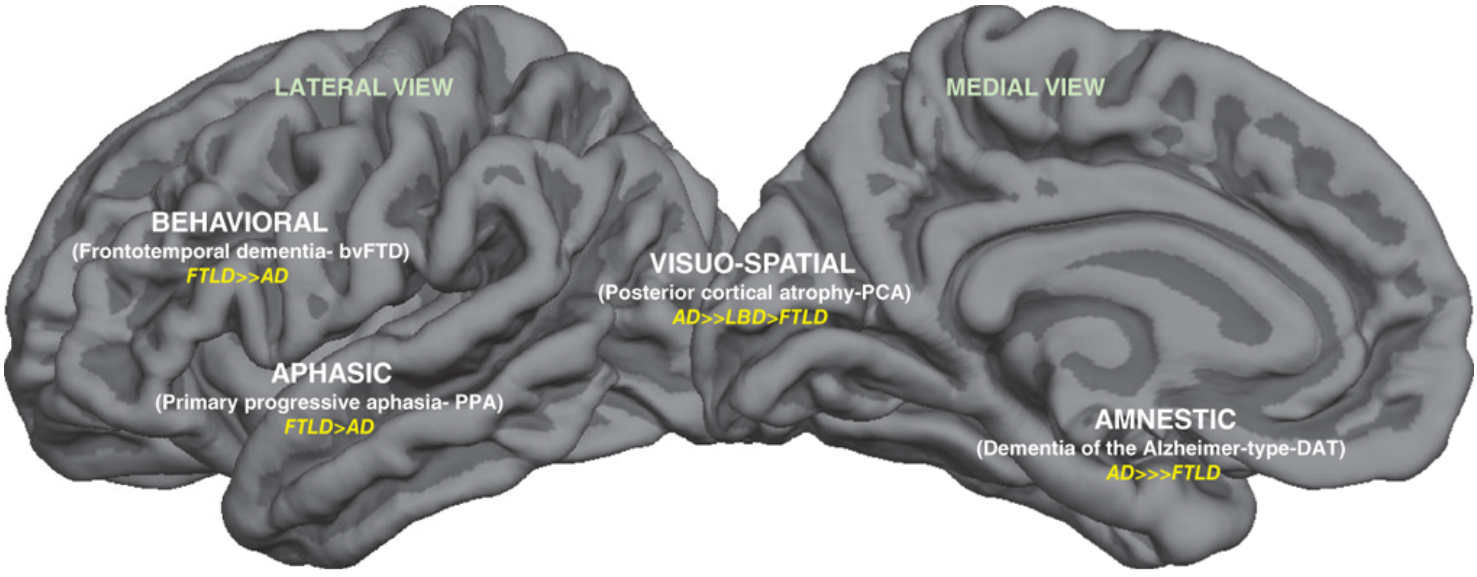
Causes of Spatial Disorientation and the Posterior Cortical Atrophy Syndrome

Cerebrovascular lesions and neoplasms in the right hemisphere are common causes of hemispatial neglect. Depending on the site of the lesion, a patient with neglect also may have hemiparesis, hemihypesthesia, and hemianopia on the left, but these are not invariant findings. The majority of these patients display considerable improvement of hemispatial neglect, usually within the first several weeks. Bálint's syndrome, dressing apraxia, and route finding impairments are more likely to result from bilateral dorsal parietal lesions; common settings for acute onset include watershed infarction between the middle and posterior cerebral artery territories, hypoglycemia, and sagittal sinus thrombosis.

A progressive form of spatial disorientation, known as the *posterior cortical atrophy* (PCA) syndrome, most commonly represents a variant of AD with unusual concentrations of neurofibrillary degeneration in the parieto-occipital cortex and the superior colliculus **(Fig. 26-4)**. Lewy body disease (LBD), CJD, and FTLD (corticobasal degeneration type) are other possible causes. The patient displays progressive hemispatial neglect, Bálint's syndrome, and route finding impairments, usually accompanied by dressing and construction apraxia.

FIGURE 26-4

Four focal dementia syndromes and their most likely neuropathologic correlates. AD, Alzheimer's disease; bvFTD, behavioral variant frontotemporal dementia; DAT, amnestic dementia of the Alzheimer-type; FTLD, frontotemporal lobar degeneration (tau or TDP-43 type); LBD, Lewy body disease; PCA, posterior cortical atrophy syndrome; PPA, primary progressive aphasia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE OCCIPITOTEMPORAL NETWORK FOR FACE AND OBJECT RECOGNITION

A patient with *prosopagnosia* cannot recognize familiar faces, including, sometimes, the reflection of his or her own face in the mirror. This is not a perceptual deficit because prosopagnosic patients easily can tell whether two faces are identical. Furthermore, a prosopagnosic patient who cannot recognize a familiar face by visual inspection alone can use auditory cues to reach appropriate recognition if allowed to listen to the person's voice. The deficit in prosopagnosia is therefore modality-specific and reflects the existence of a lesion that prevents the activation of otherwise intact multimodal associative templates by relevant visual input. Prosopagnosic patients characteristically have no difficulty with the generic identification of a face as a face or a car as a car, but may not recognize the identity of an individual face or the make of an individual car. This reflects a visual recognition deficit for proprietary features that characterize individual members of an object class. When recognition problems become more generalized and extend to the generic identification of common objects, the condition is known as *visual object agnosia*. A patient with anomia cannot name the object but can describe its use. In contrast, a patient with visual agnosia is unable either to name a visually presented object or to describe its use. Face and object recognition disorders also can result from the simultanagnosia of Bálint's syndrome, in which case they are known as *apperceptive* agnosias as opposed to the *associative* agnosias that result from inferior temporal lobe lesions.

CAUSES AND RELATION TO SEMANTIC DEMENTIA

The characteristic lesions in prosopagnosia and visual object agnosia of acute onset consist of bilateral infarctions in the territory of the posterior cerebral arteries that involve the fusiform gyrus. Associated deficits can include visual field defects (especially superior quadrantanopias) and a centrally based color blindness known as achromatopsia. Rarely, the responsible lesion is unilateral. In such cases, prosopagnosia is associated with lesions in the right hemisphere, and object agnosia with lesions in the left. Degenerative diseases of anterior and inferior temporal cortex can cause progressive associative prosopagnosia and object agnosia. The combination of progressive associative agnosia and a fluent aphasia with word comprehension impairment is known as *semantic dementia*. Patients with semantic dementia fail to recognize faces and objects and cannot understand the meaning of words denoting objects. This needs to be differentiated from the semantic type of PPA where there is severe impairment in understanding words that denote objects and in naming faces and objects but a relative preservation of face and object recognition. The anterior temporal lobe atrophy is usually bilateral in semantic dementia whereas it tends to affect mostly the left hemisphere in semantic PPA. Acute onset of the semantic dementia syndrome can be associated with herpes simplex encephalitis.

LIMBIC NETWORK FOR EXPLICIT MEMORY AND AMNESIA

Limbic and paralimbic areas (such as the hippocampus, amygdala, and entorhinal cortex), the anterior and medial nuclei of the thalamus, the medial and basal parts of the striatum, and the hypothalamus collectively constitute a distributed network known as the *limbic system*. The behavioral affiliations of this network include the coordination of emotion, motivation, autonomic tone, and endocrine function. An additional area of specialization for the limbic network and the one that is of most relevance to clinical practice is that of declarative (explicit) memory for recent episodes and experiences. A disturbance in this function is known as an *amnestic state*. In the absence of deficits in motivation, attention, language, or visuospatial function, the clinical diagnosis of a persistent global amnestic state is always associated with bilateral damage to the limbic network, usually within the hippocampo-entorhinal complex or the thalamus. Damage to the limbic network does not necessarily destroy memories but interferes with their conscious recall in coherent form. The individual fragments of information remain preserved despite the limbic lesions and can sustain what is known as *implicit memory*. For example, patients with amnestic states can acquire new motor or perceptual skills even though they may have no conscious knowledge of the experiences that led to the acquisition of these skills.

The memory disturbance in the amnestic state is multimodal and includes retrograde and anterograde components. The *retrograde amnesia* involves an inability to recall experiences that occurred before the onset of the amnestic state. Relatively recent events are more vulnerable to retrograde amnesia than are more remote and more extensively consolidated events. A patient who comes to the emergency room complaining that he cannot remember his or her identity but can remember the events of the previous day almost certainly does not have a neurologic cause of memory disturbance. The second and most important component of the amnestic state is the *anterograde amnesia*, which indicates an inability to store, retain, and recall new knowledge. Patients with amnestic states cannot remember what they ate a few hours ago or the details of an important event they may have experienced in the recent past. In the acute stages, there also may be a tendency to fill in memory gaps with inaccurate, fabricated, and often implausible information. This is known as *confabulation*. Patients with the amnestic syndrome forget that they forget and tend to deny the existence of a memory problem when questioned. Confabulation is more common in cases where the underlying lesion also interferes with parts of the frontal network, as in the case of the Wernicke-Korsakoff syndrome or traumatic head injury.

CLINICAL EXAMINATION

A patient with an amnestic state is almost always disoriented, especially to time, and has little knowledge of current news. The anterograde component of an amnestic state can be tested with a list of four to five words read aloud by the examiner up to five times or until the patient can immediately repeat the entire list without an intervening delay. The next phase of the recall occurs after a period of 5–10 min during which the patient is engaged in other tasks. Amnestic patients fail this phase of the task and may even forget that they were given a list of words to remember. Accurate recognition of the

words by multiple choice in a patient who cannot recall them indicates a less severe memory disturbance that affects mostly the retrieval stage of memory. The retrograde component of an amnesia can be assessed with questions related to autobiographical or historic events. The anterograde component of amnestic states is usually much more prominent than the retrograde component. In rare instances, occasionally associated with temporal lobe epilepsy or herpes simplex encephalitis, the retrograde component may dominate. Confusional states caused by toxic-metabolic encephalopathies and some types of frontal lobe damage lead to secondary memory impairments, especially at the stages of encoding and retrieval, even in the absence of limbic lesions. This sort of memory impairment can be differentiated from the amnestic state by the presence of additional impairments in the attention-related tasks described below in the section on the frontal lobes.

CAUSES, INCLUDING ALZHEIMER'S DISEASE

Neurologic diseases that give rise to an amnestic state include tumors (of the sphenoid wing, posterior corpus callosum, thalamus, or medial temporal lobe), infarctions (in the territories of the anterior or posterior cerebral arteries), head trauma, herpes simplex encephalitis, Wernicke-Korsakoff encephalopathy, autoimmune limbic encephalitis, and degenerative dementias such as AD and Pick's disease. The one common denominator of all these diseases is the presence of bilateral lesions within one or more components in the limbic network. Occasionally, unilateral left-sided hippocampal lesions can give rise to an amnestic state, but the memory disorder tends to be transient. Depending on the nature and distribution of the underlying neurologic disease, the patient also may have visual field deficits, eye movement limitations, or cerebellar findings.

The most common cause of progressive memory impairments in the elderly is AD. This is why a predominantly amnestic dementia is also known as a dementia of the Alzheimer-type (DAT). A prodromal stage of DAT, when daily living activities are generally preserved, is known as amnestic mild cognitive impairment (MCI). The predilection of the entorhinal cortex and hippocampus for early neurofibrillary degeneration by typical AD pathology is responsible for the initially selective impairment of episodic memory. In time, additional impairments in language, attention, and visuospatial skills emerge as the neurofibrillary degeneration spreads to additional neocortical areas. Less frequently, amnestic dementias can also be caused by FTLD.

Transient global amnesia is a distinctive syndrome usually seen in late middle age. Patients become acutely disoriented and repeatedly ask who they are, where they are, and what they are doing. The spell is characterized by anterograde amnesia (inability to retain new information) and a retrograde amnesia for relatively recent events that occurred before the onset. The syndrome usually resolves within 24–48 h and is followed by the filling in of the period affected by the retrograde amnesia, although there is persistent loss of memory for the events that occurred during the ictus. Recurrences are noted in ~20% of patients. Migraine, temporal lobe seizures, and perfusion abnormalities in the posterior cerebral territory have been postulated as causes of transient global amnesia. The absence of associated neurologic findings occasionally may lead to the incorrect diagnosis of a psychiatric disorder.

THE PREFRONTAL NETWORK FOR EXECUTIVE FUNCTION AND BEHAVIOR

The frontal lobes can be subdivided into motor-premotor, dorsolateral prefrontal, medial prefrontal, and orbitofrontal components. The terms *frontal lobe syndrome* and *prefrontal cortex* refer only to the last three of these four components. These are the parts of the cerebral cortex that show the greatest phylogenetic expansion in primates, especially in humans. The dorsolateral prefrontal, medial prefrontal, and orbitofrontal areas, along with the subcortical structures with which they are interconnected (i.e., the head of the caudate and the dorsomedial nucleus of the thalamus), collectively make up a large-scale network that coordinates exceedingly complex aspects of human cognition and behavior. The term *salience network* has been introduced to designate parts of the frontal network and their interactions with adjacent paralimbic cortices of the insula and cingulate gyrus. Impairments of social conduct and empathy seen in neurodegenerative frontal dementias are attributed to pathology of the salience network.

The prefrontal network plays an important role in behaviors that require multitasking and the integration of thought with emotion. Cognitive operations impaired by prefrontal cortex lesions often are referred to as "executive functions." The most common clinical manifestations of damage to the prefrontal network take the form of two relatively distinct syndromes. In the *frontal abulic syndrome*, the patient shows a loss of initiative, creativity, and curiosity and displays a pervasive emotional blandness, apathy, and lack of empathy. In the *frontal disinhibition syndrome*, the patient becomes socially disinhibited and shows severe impairments of judgment, insight, foresight, and the ability to mind rules of conduct. The dissociation between intact intellectual function and a total lack of even rudimentary common sense is striking. Despite the preservation of all essential memory functions, the patient cannot learn from experience and continues to display inappropriate behaviors without appearing to feel emotional pain, guilt, or regret when those behaviors repeatedly lead to disastrous consequences. The impairments may emerge only in real-life situations when behavior is under minimal external control and may not be apparent within the structured environment of the medical office. Testing judgment by asking patients what they would do if they detected a fire in a theater or found a stamped and addressed envelope on the road is not very informative because patients who answer these questions wisely in the office may still act very foolishly in real-life settings. The physician must therefore be prepared to make a diagnosis of frontal lobe disease based on historic information alone even when the mental state is quite intact in the office examination.

CLINICAL EXAMINATION

The emergence of developmentally primitive reflexes, also known as frontal release signs, such as grasping (elicited by stroking the palm) and sucking (elicited by stroking the lips) are seen primarily in patients with large structural lesions that extend into the premotor components of the frontal lobes or

in the context of metabolic encephalopathies. The vast majority of patients with prefrontal lesions and frontal lobe behavioral syndromes do not display these reflexes. Damage to the frontal lobe disrupts a variety of attention-related functions, including working memory (the transient online holding and manipulation of information), concentration span, the effortful scanning and retrieval of stored information, the inhibition of immediate but inappropriate responses, and mental flexibility. Digit span (which should be seven forward and five reverse) is decreased, reflecting poor working memory; the recitation of the months of the year in reverse order (which should take <15 s) is slowed as another indication of poor working memory; and the fluency in producing words starting with the letter a, f, or s that can be generated in 1 min (normally ≥12 per letter) is diminished even in nonaphasic patients, indicating an impairment in the ability to search and retrieve information from long-term stores. In "go–no go" tasks (where the instruction is to raise the finger upon hearing one tap but keep it still upon hearing two taps), the patient shows a characteristic inability to inhibit the response to the "no go" stimulus. Mental flexibility (tested by the ability to shift from one criterion to another in sorting or matching tasks) is impoverished; distractibility by irrelevant stimuli is increased; and there is a pronounced tendency for impersistence and perseveration. The ability for abstracting similarities and interpreting proverbs is also undermined.

The attentional deficits disrupt the orderly registration and retrieval of new information and lead to *secondary* deficits of explicit memory. The distinction of the underlying neural mechanisms is illustrated by the observation that severely amnestic patients who cannot remember events that occurred a few minutes ago may have intact if not superior working memory capacity as shown in tests of digit span. The use of the term "memory" to designate two completely different mental faculties is confusing. Working memory depends on the on-line holding of information for brief periods of time whereas explicit memory depends on the off-line storage and subsequent retrieval of the information.

CAUSES: TRAUMA, NEOPLASM, AND FRONTOTEMPORAL DEMENTIA

The abulic syndrome tends to be associated with damage in dorsolateral or dorsomedial prefrontal cortex, and the disinhibition syndrome with damage in orbitofrontal or ventromedial cortex. These syndromes tend to arise almost exclusively after bilateral lesions. Unilateral lesions confined to the prefrontal cortex may remain silent until the pathology spreads to the other side; this explains why thromboembolic CVA is an unusual cause of the frontal lobe syndrome. When behavioral syndromes of the frontal network arise in conjunction with asymmetric disease, the lesion tends to be predominantly on the right side of the brain. Common settings for frontal lobe syndromes include head trauma, ruptured aneurysms, hydrocephalus, tumors (including metastases, glioblastoma, and falx or olfactory groove meningiomas), and focal degenerative diseases, especially FTLD. The most prominent neurodegenerative frontal syndrome is known as the behavioral variant of frontotemporal dementia (bvFTD). In many patients with bvFTD the atrophy extends into the anterior temporal lobes. Occasionally, atrophy prodominantly in the right anterior temporal lobe presents with the bvFTD syndrome. The behavioral changes in these patients can range from apathy to shoplifting, compulsive gambling, sexual indiscretions, remarkable lack of common sense, new ritualistic behaviors, and alterations in dietary preferences, usually leading to increased taste for sweets or rigid attachment to specific food items. In many patients with AD, neurofibrillary degeneration eventually spreads to prefrontal cortex and gives rise to components of the frontal lobe syndrome, but almost always on a background of severe memory impairment. Rarely, the bvFTD syndrome can arise in isolation in the context of an atypical form of AD pathology.

Lesions in the caudate nucleus or in the dorsomedial nucleus of the thalamus (subcortical components of the prefrontal network) also can produce a frontal lobe syndrome affecting mostly executive functions. This is one reason why the changes in mental state associated with degenerative basal ganglia diseases such as Parkinson's disease and Huntington's disease display components of the frontal lobe syndrome. Bilateral multifocal lesions of the cerebral hemispheres, none of which are individually large enough to cause specific cognitive deficits such as aphasia and neglect, can collectively interfere with the connectivity and therefore integrating (executive) function of the prefrontal cortex. A frontal lobe syndrome, usually of the abulic form, is therefore the single most common behavioral profile associated with a variety of bilateral multifocal brain diseases, including metabolic encephalopathy, multiple sclerosis, and vitamin B_{12} deficiency, among others. Many patients with the clinical diagnosis of a frontal lobe syndrome tend to have lesions that do not involve prefrontal cortex but involve either the subcortical components of the prefrontal network or its connections with other parts of the brain. To avoid making a diagnosis of "frontal lobe syndrome" in a patient with no evidence of frontal cortex disease, it is advisable to use the diagnostic term *frontal network syndrome*, with the understanding that the responsible lesions can lie anywhere within this distributed network. A patient with frontal lobe disease raises potential dilemmas in differential diagnosis: the abulia and blandness may be misinterpreted as depression, and the disinhibition as idiopathic mania or acting out. Appropriate intervention may be delayed while a treatable tumor keeps expanding.

CARING FOR PATIENTS WITH DEFICITS OF HIGHER CEREBRAL FUNCTION

Spontaneous improvement of cognitive deficits following stroke or trauma is common. It is most rapid in the first few weeks but may continue for up to 2 years, especially in young individuals with single brain lesions. Some of the initial deficits in such cases appear to arise from remote dysfunction (diaschisis) in brain regions that are interconnected with the site of initial injury. Improvement in these patients may reflect, at least in part, a normalization of the remote dysfunction. Other mechanisms may involve functional reorganization in surviving neurons adjacent to the injury or the compensatory use of homologous structures, e.g., the right superior temporal gyrus with recovery from Wernicke's aphasia. In contrast, neurodegenerative diseases show a progression of impairment but at rates that vary greatly from patient to patient.

Pharmacologic and Non-pharmacologic Interventions

Some of the deficits described in this chapter are so complex that they may bewilder not only the patient and family but also the physician. The care of patients with such deficits requires a careful evaluation of the history, cognitive test results and diagnostic procedures. Each piece of information needs to be interpreted cautiously and placed in context. A complaint of “poor memory,” for example, may reflect an anomia; poor scores on a learning task may reflect a weakness of attention rather than explicit memory; a report of depression or indifference may reflect impaired prosody rather than a change in mood or empathy; jocularity may arise from poor insight rather than good mood. Although there are few well-controlled studies, several non-pharmacologic interventions have been used to treat higher cortical deficits. These include speech therapy for aphasias, behavioral modification for comportmental disorders, and cognitive training for visuospatial disorientation and amnestic syndromes. More practical interventions, usually delivered through occupational therapy, aim to improve daily living activities through assistive devices and modifications of the home environment. Determining driving competence is challenging, especially in the early stages of dementing diseases. An on-the-road driving test and reports from family members may help time decisions related to this very important activity. In neurodegenerative conditions such as PPA, transcranial magnetic (or direct current) stimulation has had mixed success in eliciting symptomatic improvement. The goal is to activate remaining neurons at sites of atrophy or in unaffected regions of the contralateral hemisphere. Depression and sleep disorders can intensify the cognitive disorders and should be treated with appropriate modalities. If neuroleptics become absolutely necessary for the control of agitation, atypical neuroleptics are preferable because of their lower extrapyramidal side effects. Treatment with neuroleptics in elderly patients with dementia requires weighing the potential benefits against the potentially serious side effects. This is especially relevant to the case of patients with Lewy body dementia, who can be unusually sensitive to side effects.

As in all other branches of medicine, a crucial step in patient care is to identify the underlying cause of the impairment. This is easily done in cases of CVA, head trauma or encephalitis but becomes particularly challenging in the dementias because the same progressive clinical syndrome can be caused by one of several neuropathologic entities. The advent of imaging, blood, and CSF biomarkers now makes it possible to address this question with reasonable success and to make specific diagnoses of AD, LBD, CJD, FTLD. A specific etiological diagnosis allows the physician to recommend medications or clinical trials that are the most appropriate for the underlying disease process. A clinical assessment that identifies the principal domain of behavioral and cognitive impairment followed by the judicious use of biomarker information to surmise the nature of the underlying disease allows a personalized approach to patients with higher cognitive impairment.

FURTHER READING

Mesulam M-M: Behavioral neuroanatomy: Large-scale networks, association cortex, frontal syndromes, the limbic system and hemispheric specialization, in *Principles of Behavioral and Cognitive Neurology*, M-M Mesulam (ed). New York, Oxford University Press, 2000, pp 1–120.

Mesulam M-M et al: Case 1-2017: A 70-year-old woman with gradually progressive loss of language. N Engl J Med 376:158, 2017. [PubMed: 28076711]

Miller BL, Boeve BF (eds): *The Behavioral Neurology of Dementia*, 2nd ed. Cambridge University Press, 2017.

Teichmann M et al: Direct current stimulation over the anterior temporal areas boosts semantic processing in primary progressive aphasia. Ann Neurol 80:693, 2016. [PubMed: 27553723]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 27: Sleep Disorders

Thomas E. Scammell; Clifford B. Saper; Charles A. Czeisler

INTRODUCTION

Disturbed sleep is one of the most common health complaints that physicians encounter. More than one-half of adults in the United States experience at least intermittent sleep disturbance, and only 30% of adult Americans report consistently obtaining a sufficient amount of sleep. The National Academy of Medicine has estimated that 50–70 million Americans suffer from a chronic disorder of sleep and wakefulness, which can adversely affect daytime functioning as well as physical and mental health. A high prevalence of sleep disorders across all cultures is also now increasingly recognized, and these problems are expected to further increase in the years ahead as the global population ages. Over the last 20 years, the field of sleep medicine has emerged as a distinct specialty in response to the impact of sleep disorders and sleep deficiency on overall health. Nonetheless, over 80% of patients with sleep disorders remain undiagnosed and untreated—costing the U.S. economy over $400 billion annually in increased health care costs, lost productivity, accidents and injuries, and leading to the development of workplace-based sleep health education and sleep disorders screening programs designed to address this unmet medical need.

PHYSIOLOGY OF SLEEP AND WAKEFULNESS

Adults need at least 7 h of sleep per night to promote optimal health, although the timing, duration, and internal structure of sleep vary among individuals. In the United States, adults tend to have one consolidated sleep episode each night, although in some cultures sleep may be divided into a mid-afternoon nap and a shortened night sleep. This pattern changes considerably over the life span, as infants and young children sleep considerably more than older people.

The stages of human sleep are defined on the basis of characteristic patterns in the electroencephalogram (EEG), the electrooculogram (EOG—a measure of eye-movement activity), and the surface electromyogram (EMG) measured on the chin, neck, and legs. The continuous recording of these electrophysiologic parameters to define sleep and wakefulness is termed *polysomnography*.

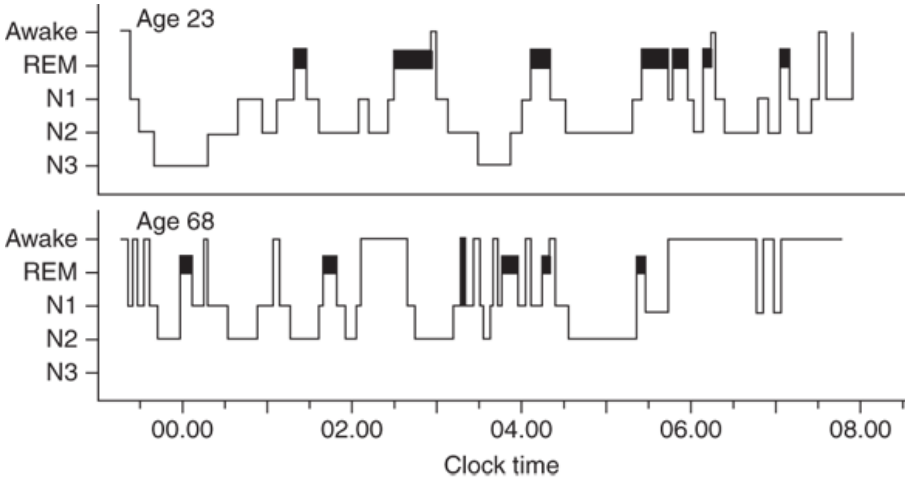
Polysomnographic profiles define two basic states of sleep: (1) rapid eye movement (REM) sleep and (2) non–rapid eye movement (NREM) sleep. NREM sleep is further subdivided into three stages: N1, N2, and N3, characterized by increasing arousal threshold and slowing of the cortical EEG. REM sleep is characterized by a low-amplitude, mixed-frequency EEG similar to that of NREM stage N1 sleep, and the EOG shows REMs which tend to occur in flurries or bursts. EMG activity is absent in nearly all skeletal muscles except those involved in respiration, reflecting the brainstem-mediated muscle paralysis that is characteristic of REM sleep.

ORGANIZATION OF HUMAN SLEEP

Normal nocturnal sleep in adults displays a consistent organization from night to night **(Fig. 27-1)**. After sleep onset, sleep usually progresses through NREM stages N1–N3 sleep within 45–60 min. NREM stage N3 sleep (also known as slow-wave sleep) predominates in the first third of the night and comprises 15–25% of total nocturnal sleep time in young adults. Sleep deprivation increases the rapidity of sleep onset and both the intensity and amount of slow-wave sleep.

FIGURE 27-1

Wake-sleep architecture. Alternating stages of wakefulness, the three stages of non–rapid eye movement sleep (N1–N3), and rapid eye movement (REM) sleep (*solid bars*) occur over the course of the night for representative young and older adult men. Characteristic features of sleep in older people include reduction of N3 slow-wave sleep, frequent spontaneous awakenings, early sleep onset, and early morning awakening.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The first REM sleep episode usually occurs in the second hour of sleep. NREM and REM sleep alternate through the night with an average period of 90–110 min (the "ultradian" sleep cycle). Overall, in a healthy young adult, REM sleep constitutes 20–25% of total sleep, and NREM stages N1 and N2 constitute 50–60%.

Age has a profound impact on sleep state organization (Fig. 27-1). N3 sleep is most intense and prominent during childhood, decreasing with puberty and across the second and third decades of life. N3 sleep declines during adulthood to the point where it may be completely absent in older adults. The remaining NREM sleep becomes more fragmented, with many more frequent awakenings from NREM sleep. It is the increased frequency of awakenings, rather than a decreased ability to fall back asleep, that accounts for the increased wakefulness during the sleep episode in older people. While REM sleep may account for 50% of total sleep time in infancy, the percentage falls off sharply over the first postnatal year as a mature REM-NREM cycle develops; thereafter, REM sleep occupies about 25% of total sleep time.

Sleep deprivation degrades cognitive performance, particularly on tests that require continual vigilance. Paradoxically, older people are less vulnerable to the neurobehavioral performance impairment induced by acute sleep deprivation than young adults, maintaining their reaction time and sustaining vigilance with fewer lapses of attention. However, it is more difficult for older adults to obtain recovery sleep after staying awake all night, as the ability to sleep during the daytime declines with age.

After sleep deprivation, NREM sleep is generally recovered first, followed by REM sleep. However, because REM sleep tends to be most prominent in the second half of the night, sleep truncation (e.g., by an alarm clock) results in selective REM sleep deprivation. This may increase REM sleep pressure to the point where the first REM sleep may occur much earlier in the nightly sleep episode. Because several disorders (see below) also cause sleep fragmentation, it is important that the patient have sufficient sleep opportunity (at least 8 h per night) for several nights prior to a diagnostic polysomnogram.

There is growing evidence that inadequate sleep in humans is associated with glucose intolerance that may contribute to the development of diabetes, obesity, and the metabolic syndrome, plus impaired immune responses, accelerated atherosclerosis, and increased risk of cardiac disease, cognitive impairment, Alzheimer's disease, and stroke. For these reasons, the National Academy of Medicine declared sleep deficiency and sleep disorders "an unmet public health problem."

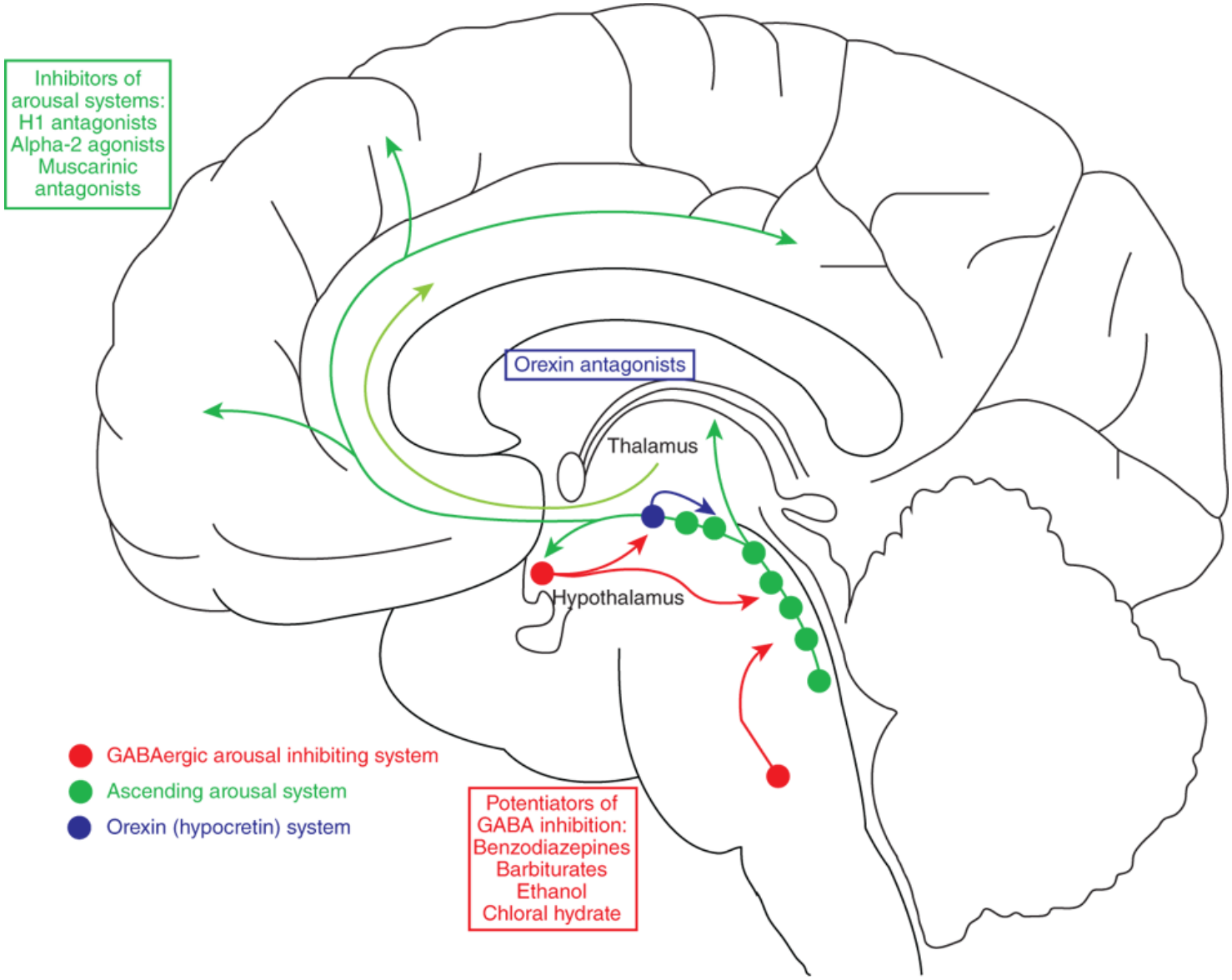
WAKE AND SLEEP ARE REGULATED BY BRAIN CIRCUITS

Two principal neural systems govern the expression of sleep and wakefulness. The ascending arousal system, illustrated in green in **Fig. 27-2**, consists of clusters of nerve cells extending from the upper pons to the hypothalamus and basal forebrain that activate the cerebral cortex, thalamus (which is necessary to relay sensory information to the cortex), and other forebrain regions. The ascending arousal neurons use monoamines (norepinephrine, dopamine, serotonin, and histamine), glutamate, or acetylcholine as neurotransmitters to activate their target neurons. Some basal forebrain neurons use GABA to disinhibit cortical inhibitory interneurons, thus promoting arousal. Additional wake-promoting neurons in the hypothalamus use the peptide neurotransmitter orexin (also known as hypocretin, shown in blue) to reinforce activity in the other arousal cell groups.

FIGURE 27-2

Relationship of drugs for insomnia with wake-sleep systems. The arousal system in the brain (*green*) includes monoaminergic, glutamatergic, and cholinergic neurons in the brainstem that activate neurons in the hypothalamus, thalamus, basal forebrain, and cerebral cortex. Orexin

neurons (*blue*) in the hypothalamus, which are lost in narcolepsy, reinforce and stabilize arousal by activating other components of the arousal system. The sleep-promoting system (*red*) consists of GABAergic neurons in the preoptic area and brainstem that inhibit the components of the arousal system, thus allowing sleep to occur. Drugs used to treat insomnia include those that block the effects of arousal system neurotransmitters (*green and blue*) and those that enhance the effects of γ-aminobutyric acid (GABA) produced by the sleep system (*red*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Damage to the arousal system at the level of the rostral pons and lower midbrain causes coma, indicating that the ascending arousal influence from this level is critical in maintaining wakefulness. Injury to the hypothalamic branch of the arousal system causes profound sleepiness, but usually not coma. Specific loss of the orexin neurons produces the sleep disorder narcolepsy (see below). Damage to the thalamus causes loss of the content of wakefulness, but wake-sleep cycles are largely preserved.

The arousal system is turned off during sleep by inhibitory inputs from cell groups in the sleep-promoting system, shown in Fig. 27-2 in red. These neurons in the preoptic area and pons use γ-aminobutyric acid (GABA) to inhibit the arousal system. Additional neurons in the lateral hypothalamus containing the peptide melanin-concentrating hormone promote REM sleep. Many sleep-promoting neurons are themselves inhibited by inputs from the arousal system. This mutual inhibition between the arousal- and sleep-promoting systems forms a neural circuit akin to what electrical engineers call a “flip-flop switch.” A switch of this type tends to promote rapid transitions between the on (wake) and off (sleep) states, while avoiding intermediate states. The relatively rapid transitions between waking and sleeping states, as seen in the EEG of humans and animals, is consistent with this model.

Neurons in the ventrolateral preoptic nucleus, one of the key sleep-promoting sites, are lost during normal human aging, correlating with reduced ability to maintain sleep (sleep fragmentation). The ventrolateral preoptic neurons are also injured in Alzheimer's disease, which may in part account for the poor sleep quality in those patients.

Transitions between NREM and REM sleep appear to be governed by a similar switch in the brainstem. GABAergic REM-Off neurons have been identified in the lower midbrain that inhibit REM-On neurons in the upper pons. The REM-On group contains both GABAergic neurons that inhibit the REM-Off group (thus satisfying the conditions for a REM sleep flip-flop switch) as well as glutamatergic neurons that project widely in the central nervous system (CNS) to cause the key phenomena associated with REM sleep. REM-On neurons that project to the medulla and spinal cord activate inhibitory (GABA and glycine-containing) interneurons, which in turn hyperpolarize the motor neurons, producing the paralysis of REM sleep. REM-On neurons that project to the forebrain may be important in producing dreams.

The REM sleep switch receives cholinergic input, which favors transitions to REM sleep, and monoaminergic (norepinephrine and serotonin) input that prevents REM sleep. As a result, drugs that increase monoamine tone (e.g., serotonin or norepinephrine reuptake inhibitors) tend to reduce the amount of REM sleep. Damage to the neurons that promote REM sleep paralysis can produce REM sleep behavior disorder, a condition in which patients act out their dreams (see below).

SLEEP-WAKE CYCLES ARE DRIVEN BY HOMEOSTATIC, ALLOSTATIC, AND CIRCADIAN INPUTS

The gradual increase in sleep drive with prolonged wakefulness, followed by deeper slow-wave sleep and prolonged sleep episodes, demonstrates that there is a *homeostatic* mechanism that regulates sleep. The neurochemistry of sleep homeostasis is only partially understood, but with prolonged wakefulness, adenosine levels rise in parts of the brain. Adenosine may act through A1 receptors to directly inhibit many arousal-promoting brain regions. In addition, adenosine promotes sleep through A2a receptors; blockade of these receptors by caffeine is one of the chief ways in which people fight sleepiness. Other humoral factors, such as prostaglandin D_2, have also been implicated in this process. Both adenosine and prostaglandin D_2 activate the sleep-promoting neurons in the ventrolateral preoptic nucleus.

Allostasis is the physiologic response to a challenge such as physical danger or psychological threat that cannot be managed by homeostatic mechanisms. These stress responses can severely impact the need for and ability to sleep. For example, insomnia is very common in patients with anxiety and other psychiatric disorders. Stress-induced insomnia is even more common, affecting most people at some time in their lives. Positron emission tomography (PET) studies in patients with chronic insomnia show hyperactivation of components of the ascending arousal system, as well as their targets in the limbic system in the forebrain (e.g., cingulate cortex and amygdala). The limbic areas are not only targets for the arousal system, but they also send excitatory outputs back to the arousal system, which contributes to a vicious cycle of anxiety about insomnia that makes it more difficult to sleep. Approaches to treating insomnia may employ drugs that either inhibit the output of the ascending arousal system (green and blue in Fig. 27-2) or potentiate the output of the sleep-promoting system (red in Fig. 27-2). However, behavioral approaches (cognitive behavioral therapy [CBT] and sleep hygiene) that may reduce forebrain limbic activity at bedtime are often the best long term treatment.

Sleep is also regulated by a strong *circadian* timing signal, driven by the suprachiasmatic nuclei (SCN) of the hypothalamus, as described below. The SCN sends outputs to key sites in the hypothalamus, which impose 24-h rhythms on a wide range of behaviors and body systems, including the wake-sleep cycle.

PHYSIOLOGY OF CIRCADIAN RHYTHMICITY

The wake-sleep cycle is the most evident of many 24-h rhythms in humans. Prominent daily variations also occur in endocrine, thermoregulatory, cardiac, pulmonary, renal, immune, gastrointestinal, and neurobehavioral functions. At the molecular level, endogenous circadian rhythmicity is driven by self-sustaining transcriptional/translational feedback loops. In evaluating daily rhythms in humans, it is important to distinguish between diurnal components passively evoked by periodic environmental or behavioral changes (e.g., the increase in blood pressure and heart rate that occurs upon assumption of the upright posture) and circadian rhythms actively driven by an endogenous oscillatory process (e.g., the circadian variations in adrenal cortisol and pineal melatonin secretion that persist across a variety of environmental and behavioral conditions).

While it is now recognized that most cells in the body have circadian clocks that regulate diverse physiologic processes, most of these disparate clocks when placed in isolation in a tissue explant are unable to maintain the long-term synchronization with each other that is required to produce useful 24-h rhythms aligned with the external light-dark cycle. The neurons in the SCN are interconnected with one another in such a way as to produce a near-24-h synchronous rhythm of neural activity even in prolonged slice culture. They also receive

visual input to synchronize them with the external world and have outputs to transmit that signal to the rest of the body. Bilateral destruction of the SCN results in a loss of most endogenous circadian rhythms including wake-sleep behavior and rhythms in endocrine and metabolic systems. The genetically determined period of this endogenous neural oscillator, which averages ~24.15 h in humans, is normally synchronized to the 24-h period of the environmental light-dark cycle through direct input from intrinsically photosensitive ganglion cells in the retina to the SCN. Humans are exquisitely sensitive to the resetting effects of light, particularly the shorter wavelengths (~460–500 nm) in the blue part of the visible spectrum. Small differences in circadian period contribute to variations in diurnal preference. For example, young adults typically have long intrinsic circadian periods and consequently go to bed late and rise late, whereas others have short periods and go to bed and rise earlier. Changes in homeostatic sleep regulation may underlie age-related changes in sleep-wake timing.

The timing and internal architecture of sleep are directly coupled to the output of the endogenous circadian pacemaker. Paradoxically, the endogenous circadian rhythm for wake propensity peaks just before the habitual bedtime, whereas that of sleep propensity peaks near the habitual wake time. These rhythms are thus timed to oppose the rise of sleep tendency throughout the usual waking day and the decline of sleep propensity during the habitual sleep episode, respectively. Misalignment of the endogenous circadian pacemaker with the desired wake-sleep cycle can, therefore, induce insomnia, decrease alertness, and impair performance, posing health problems for night-shift workers and airline travelers.

BEHAVIORAL AND PHYSIOLOGIC CORRELATES OF SLEEP STATES AND STAGES

Polysomnographic staging of sleep correlates with behavioral changes during specific states and stages. During the transitional state (stage N1) between wakefulness and deeper sleep, individuals may respond to faint auditory or visual signals. Formation of short-term memories is inhibited at the onset of NREM stage N1 sleep, which may explain why individuals aroused from that transitional sleep stage frequently lack situational awareness. After sleep deprivation, such transitions may intrude upon behavioral wakefulness notwithstanding attempts to remain continuously awake (see "Shift-Work Disorder," below).

Subjects woken from REM sleep recall vivid dream imagery >80% of the time, especially later in the night. Less vivid imagery may also be reported after NREM sleep interruptions. Certain disorders may occur during specific sleep stages and are described below under "Parasomnias." These include sleepwalking, night terrors, and enuresis (bed wetting), which occur most commonly in children during deep (N3) NREM sleep, and REM sleep behavior disorder, which occurs mainly among older men who fail to maintain full paralysis during REM sleep, and often call out, thrash around, or even act out fragments of dreams.

All major physiologic systems are influenced by sleep. Blood pressure and heart rate decrease during NREM sleep, particularly during N3 sleep. During REM sleep, bursts of eye movements are associated with large variations in both blood pressure and heart rate mediated by the autonomic nervous system. Cardiac dysrhythmias may occur selectively during REM sleep. Respiratory function also changes. In comparison to relaxed wakefulness, respiratory rate becomes slower but more regular during NREM sleep (especially N3 sleep) and becomes irregular during bursts of eye movements in REM sleep. Decreases in minute ventilation during NREM sleep are out of proportion to the decrease in metabolic rate, resulting in a slightly higher P_{CO_2}.

Within the brain itself, neurotransmission is supported by ion gradients across the cell membranes of neurons and astrocytes. These ion flows are accompanied by increases in intracellular volume, so that during wake, there is very little extracellular space in the brain. During sleep, intracellular volume is reduced, resulting in increased extracellular space, which has higher calcium and lower potassium concentrations, supporting hyperpolarization and reduced firing of neurons. This expansion of the extracellular space during sleep increases diffusion of substances that accumulate extracellularly, like β-amyloid peptide, enhancing their clearance from the brain via cerebrospinal fluid flow. Recent evidence suggests that lack of adequate sleep may contribute to extracellular accumulation of β-amyloid peptide, a key step in the pathogenesis of Alzheimer's disease.

Endocrine function also varies with sleep. N3 sleep is associated with secretion of growth hormone in men, while sleep in general is associated with augmented secretion of prolactin in both men and women. Sleep has a complex effect on the secretion of luteinizing hormone (LH): during puberty, sleep is associated with increased LH secretion, whereas sleep in postpubertal women inhibits LH secretion in the early follicular phase of the menstrual cycle. Sleep onset (and probably N3 sleep) is associated with inhibition of thyroid-stimulating hormone and of the adrenocorticotropic hormone–cortisol axis, an effect that is superimposed on the prominent circadian rhythms in the two systems.

The pineal hormone melatonin is secreted predominantly at night in both day- and night-active species, reflecting the direct modulation of pineal activity by the SCN via the sympathetic nervous system which innervates the pineal gland. Melatonin secretion does not require sleep,

but melatonin secretion is inhibited by ambient light, an effect mediated by the neural connection from the retina to the pineal gland via the SCN. Sleep efficiency is highest when sleep coincides with endogenous melatonin secretion. When endogenous melatonin levels are low, such as during the biological day or at the desired bedtime in patients with delayed sleep-wake phase disorder (DSWPD), administration of exogenous melatonin can hasten sleep onset and increase sleep efficiency, but it does not increase sleep efficiency if administered when endogenous melatonin levels are elevated. This may explain why melatonin is often ineffective in the treatment of patients with primary insomnia. On the other hand, patients with sympathetic denervation of the pineal gland, such as occurs in cervical spinal cord injury or in patients with Parkinson's disease, often have low melatonin levels, and administration of melatonin (3 mg 30 min before bedtime) may help them sleep.

Sleep is accompanied by alterations of thermoregulatory function. NREM sleep is associated with an increase in the firing of warm-responsive neurons in the preoptic area and a fall in body temperature; conversely, skin warming without increasing core body temperature has been found to increase NREM sleep. REM sleep is associated with reduced thermoregulatory responsiveness.

DISORDERS OF SLEEP AND WAKEFULNESS

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Sleep Disorders

Patients may seek help from a physician because of: (1) sleepiness or tiredness during the day; (2) difficulty initiating or maintaining sleep at night (insomnia); or (3) unusual behaviors during sleep itself (parasomnias).

Obtaining a careful history is essential. In particular, the duration, severity, and consistency of the symptoms are important, along with the patient's estimate of the consequences of the sleep disorder on waking function. Information from a bed partner or family member is often helpful because some patients may be unaware of symptoms such as heavy snoring or may underreport symptoms such as falling asleep at work or while driving. Physicians should inquire about when the patient typically goes to bed, when they fall asleep and wake up, whether they awaken during sleep, whether they feel rested in the morning, and whether they nap during the day. Depending on the primary complaint, it may be useful to ask about snoring, witnessed apneas, restless sensations in the legs, movements during sleep, depression, anxiety, and behaviors around the sleep episode. The physical examination may provide evidence of a small airway, large tonsils, or a neurologic or medical disorder that contributes to the main complaint.

It is important to remember that, rarely, seizures may occur exclusively during sleep, mimicking a primary sleep disorder; such sleep-related seizures typically occur during episodes of NREM sleep and may take the form of generalized tonic-clonic movements (sometimes with urinary incontinence or tongue biting) or stereotyped movements in partial complex epilepsy **(Chap. 418)**.

It is often helpful for the patient to complete a daily sleep log for 1–2 weeks to define the timing and amounts of sleep. When relevant, the log can also include information on levels of alertness, work times, and drug and alcohol use, including caffeine and hypnotics.

Polysomnography is necessary for the diagnosis of several disorders such as sleep apnea, narcolepsy, and periodic limb movement disorder (PLMD). A conventional polysomnogram performed in a clinical sleep laboratory allows measurement of sleep stages, respiratory effort and airflow, oxygen saturation, limb movements, heart rhythm, and additional parameters. A home sleep test usually focuses on just respiratory measures and is helpful in patients with a moderate to high likelihood of having obstructive sleep apnea. The multiple sleep latency test (MSLT) is used to measure a patient's propensity to sleep during the day and can provide crucial evidence for diagnosing narcolepsy and some other causes of sleepiness. The maintenance of wakefulness test is used to measure a patient's ability to sustain wakefulness during the daytime and can provide important evidence for evaluating the efficacy of therapies for improving sleepiness in conditions such as narcolepsy and obstructive sleep apnea.

EVALUATION OF DAYTIME SLEEPINESS

Up to 25% of the adult population has persistent daytime sleepiness that impairs an individual's ability to perform optimally in school, at work, while driving, and in other conditions that require alertness. Sleepy students often have trouble staying alert and performing well in school, and sleepy adults struggle to stay awake and focused on their work. More than half of Americans have fallen asleep while driving. An estimated 1.2 million motor vehicle crashes per year are due to drowsy drivers, causing about 20% of all serious crash injuries and deaths. One

needn't fall asleep to have an accident, as the inattention and slowed responses of drowsy drivers are a major contributor. Twenty-four hours of continuous wakefulness impairs reaction time as much as a blood alcohol concentration of 0.10 g/dL (which is legally drunk in all 50 states).

Identifying and quantifying sleepiness can be challenging. First, patients may describe themselves as "sleepy," "fatigued," or "tired," and the meanings of these words may differ between patients. For clinical purposes, it is best to use the term "sleepiness" to describe a propensity to fall asleep; whereas "fatigue" is best used to describe a feeling of low physical or mental energy but without a tendency to actually sleep. Sleepiness is usually most evident when the patient is sedentary, whereas fatigue may interfere with more active pursuits. Sleepiness generally occurs with disorders that reduce the quality or quantity of sleep or that interfere with the neural mechanisms of arousal, whereas fatigue is more common in inflammatory disorders such as cancer, multiple sclerosis **(Chap. 436)**, fibromyalgia **(Chap. 366)**, chronic fatigue syndrome **(Chap. 442)**, or endocrine deficiencies such as hypothyroidism **(Chap. 376)** or Addison's disease **(Chap. 379)**. Second, sleepiness can affect judgment in a manner analogous to ethanol, such that patients may have limited insight into the condition and the extent of their functional impairment. Finally, patients may be reluctant to admit that sleepiness is a problem because they may have become unfamiliar with feeling fully alert and because sleepiness is sometimes viewed pejoratively as reflecting poor motivation or bad sleep habits.

Table 27-1 outlines the diagnostic and therapeutic approach to the patient with a complaint of excessive daytime sleepiness.

TABLE 27-1

Evaluation of the Patient with Excessive Daytime Sleepiness

Findings on History and Physical Examination	Diagnostic Evaluation	Diagnosis	Therapy
Difficulty waking in the morning, rebound sleep on weekends and vacations with improvement in sleepiness	Sleep log	Insufficient sleep	Sleep education and behavioral modification to increase amount of sleep
Obesity, snoring, hypertension	Polysomnogram or home sleep test	Obstructive sleep apnea **(Chap. 291)**	Continuous positive airway pressure; upper airway surgery (e.g., uvulopalatopharyngoplasty); dental appliance; weight loss
Cataplexy, hypnogogic hallucinations, sleep paralysis	Polysomnogram and multiple sleep latency test	Narcolepsy	Stimulants (e.g., modafinil, methylphenidate); REM sleep-suppressing antidepressants (e.g., venlafaxine); sodium oxybate
Restless legs, kicking movements during sleep	Assessment for predisposing medical conditions (e.g., iron deficiency or renal failure)	Restless legs syndrome with or without periodic limb movements	Treatment of predisposing condition; dopamine agonists (e.g., pramipexole, ropinirole); gabapentin; opiates
Sedating medications, stimulant withdrawal, head trauma, systemic inflammation, Parkinson's disease and other neurodegenerative disorders, hypothyroidism, encephalopathy	Thorough medical history and examination including detailed neurologic examination	Sleepiness due to a drug or medical condition	Change medications, treat underlying condition, consider stimulants

To determine the extent and impact of sleepiness on daytime function, it is helpful to ask patients about the occurrence of sleep episodes during normal waking hours, both intentional and unintentional. Specific areas to be addressed include the occurrence of inadvertent sleep episodes while driving or in other safety-related settings, sleepiness while at work or school (and the relationship of sleepiness to work and school performance), and the effect of sleepiness on social and family life. Standardized questionnaires such as the Epworth Sleepiness Scale are often used clinically to measure sleepiness.

Eliciting a history of daytime sleepiness is usually adequate, but objective quantification is sometimes necessary. The MSLT measures a patient's propensity to sleep under quiet conditions. An overnight polysomnogram should precede the MSLT to establish that the patient has had an adequate amount of good-quality nighttime sleep. The MSLT consists of five 20-min nap opportunities every 2 h across the day. The patient is instructed to try to fall asleep, and the major endpoints are the average latency to sleep and the occurrence of REM sleep during the naps. An average sleep latency across the naps of <8 min is considered objective evidence of excessive daytime sleepiness. REM sleep normally occurs only during the nighttime sleep episode, and the occurrence of REM sleep in two or more of the MSLT naps provides support for the diagnosis of narcolepsy.

For the safety of the individual and the general public, physicians have a responsibility to help manage issues around driving in patients with sleepiness. Legal reporting requirements vary from state to state, but at a minimum, physicians should inform sleepy patients about their increased risk of having an accident and advise such patients not to drive a motor vehicle until the sleepiness has been treated effectively. This discussion is especially important for commercial drivers, and it should be documented in the patient's medical record.

INSUFFICIENT SLEEP

Insufficient sleep is probably the most common cause of excessive daytime sleepiness. The average adult needs 7.5–8 h of sleep, but on weeknights, the average U.S. adult gets only 6.75 h of sleep. Only 30% of the U.S. adult population reports consistently obtaining sufficient sleep. Insufficient sleep is especially common among shift workers, individuals working multiple jobs, and people in lower socioeconomic groups. Most teenagers need ≥9 h of sleep, but many fail to get enough sleep because of circadian phase delay, plus social pressures to stay up late coupled with early school start times. Late evening light exposure, television viewing, video-gaming, social media, texting, and smartphone use often delay bedtimes despite the fixed, early wake times required for work or school. As is typical with any disorder that causes sleepiness, individuals with chronically insufficient sleep may feel inattentive, irritable, unmotivated, and depressed, and have difficulty with school, work, and driving. Individuals differ in their optimal amount of sleep, and it can be helpful to ask how much sleep the patient obtains on a quiet vacation when he or she can sleep without restrictions. Some patients may think that a short amount of sleep is normal or advantageous, and they may not appreciate their biological need for more sleep, especially if coffee and other stimulants mask the sleepiness. A 2-week sleep log documenting the timing of sleep and daily level of alertness is diagnostically useful and provides helpful feedback for the patient. Extending sleep to the optimal amount on a regular basis can resolve the sleepiness and other symptoms. As with any lifestyle change, extending sleep requires commitment and adjustments, but the improvements in daytime alertness make this change worthwhile.

SLEEP APNEA SYNDROMES

Respiratory dysfunction during sleep is a common, serious cause of excessive daytime sleepiness as well as of disturbed nocturnal sleep. At least 24% of middle-aged men and 9% of middle-aged women in the United States have a reduction or cessation of breathing dozens or more times each night during sleep, with 9% of men and 4% of women doing so more than a hundred times per night. These episodes may be due to an occlusion of the airway (*obstructive sleep apnea*), absence of respiratory effort (*central sleep apnea*), or a combination of these factors. Failure to recognize and treat these conditions appropriately may lead to impairment of daytime alertness, increased risk of sleep-related motor vehicle crashes, depression, hypertension, myocardial infarction, diabetes, stroke, and increased mortality. Sleep apnea is particularly prevalent in overweight men and in the elderly, yet it is estimated to go undiagnosed in most affected individuals. This is unfortunate because several effective treatments are available. **Readers are referred to Chap. 291 for a comprehensive review of the diagnosis and treatment of patients with sleep apnea.**

NARCOLEPSY

Narcolepsy is characterized by difficulty sustaining wakefulness, poor regulation of REM sleep, and disturbed nocturnal sleep. All patients with narcolepsy have excessive daytime sleepiness. This sleepiness is usually moderate to severe, and in contrast to patients with disrupted sleep (e.g., sleep apnea), people with narcolepsy usually feel well rested upon awakening and then feel tired throughout much of the day. In addition, they often experience symptoms related to an intrusion of REM sleep characteristics. REM sleep is characterized by dreaming and muscle paralysis, and people with narcolepsy can have: (1) sudden muscle weakness without a loss of consciousness, which is usually triggered by strong emotions (cataplexy; **Video 27-1**); (2) dream-like hallucinations at sleep onset (hypnagogic hallucinations) or upon awakening (hypnopompic hallucinations); and (3) muscle paralysis upon awakening (sleep paralysis). With severe cataplexy, an individual may be laughing at a joke and then suddenly collapse to the ground, immobile but awake for 1–2 min. With milder episodes, patients may have partial weakness of the face or neck. Narcolepsy is one of the more common causes of chronic sleepiness and affects about 1 in 2000 people in the United States. Narcolepsy typically begins between age 10 and 20; once established, the disease persists for life.

Video 027-1: Classic Cataplexy

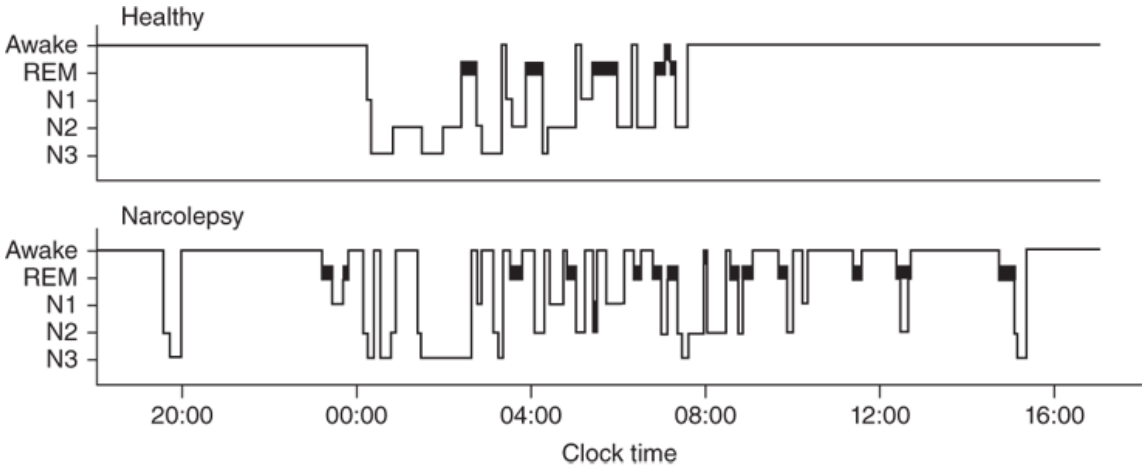
A typical episode of severe cataplexy. The patient is joking and then falls to the ground with an abrupt loss of muscle tone. The electromyogram recordings (four lower traces on the right) show reductions in muscle activity during the period of paralysis. The electroencephalogram (top two traces) shows wakefulness throughout the episode. *(Video courtesy of Giuseppe Plazzi, University of Bologna.)*

Image not available.
Play Video

Narcolepsy is caused by loss of the hypothalamic neurons that produce the orexin neuropeptides (also known as hypocretins). Research in mice and dogs first demonstrated that a loss of orexin signaling due to null mutations of either the orexin neuropeptides or one of the orexin receptors causes sleepiness and cataplexy nearly identical to that seen in people with narcolepsy. Although genetic mutations rarely cause human narcolepsy, researchers soon discovered that patients with narcolepsy with cataplexy (now called type 1 narcolepsy) have very low or undetectable levels of orexins in their cerebrospinal fluid, and autopsy studies showed a nearly complete loss of the orexin-producing neurons in the hypothalamus. The orexins normally promote long episodes of wakefulness and suppress REM sleep, and thus, loss of orexin signaling results in frequent intrusions of sleep during the usual waking episode, with REM sleep and fragments of REM sleep at any time of day **(Fig. 27-3)**. Patients with narcolepsy but no cataplexy (type 2 narcolepsy) usually have normal orexin levels and may have other yet uncharacterized causes of their excessive daytime sleepiness.

FIGURE 27-3

Polysomnographic recordings of a healthy individual and a patient with narcolepsy. The healthy individual has a long period or NREM sleep before entering REM sleep, but the individual with narcolepsy enters rapid eye movement (REM) sleep quickly at night and has moderately fragmented sleep. During the day, the healthy subject stays awake from 8:00 AM until midnight, but the patient with narcolepsy dozes off frequently, with many daytime naps that include REM sleep.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Extensive evidence suggests that an autoimmune process likely causes this selective loss of the orexin-producing neurons. Certain human leukocyte antigens (HLAs) can increase the risk of autoimmune disorders **(Chap. 343)**, and narcolepsy has the strongest known HLA association. HLA DQB1*06:02 is found in >90% of people with type 1 narcolepsy, whereas it occurs in only 12–25% of the general population. Researchers now hypothesize that in people with DQB1*06:02, an immune response against influenza, *Streptococcus*, or other infections may also damage the orexin-producing neurons through a process of molecular mimicry. This mechanism may account for the eight- to twelvefold increase in new cases of narcolepsy among children in Europe who received a particular brand of H1N1 influenza A vaccine (Pandemrix). Traumatic brain injury can also damage orexin-containing neurons, inducing type 2 narcolepsy.

On rare occasions, narcolepsy can occur with neurologic disorders such as tumors or strokes that directly damage the orexin-producing neurons in the hypothalamus or their projections.

Diagnosis

Narcolepsy is most commonly diagnosed by the history of chronic sleepiness plus cataplexy or other symptoms. Many disorders can cause feelings of weakness, but with true cataplexy, patients will describe definite functional weakness (e.g., slurred speech, dropping a cup, slumping into a chair) that has consistent emotional triggers such as heartfelt mirth when laughing at a great joke, happy surprise at unexpectedly seeing a friend, or intense anger. Cataplexy occurs in about half of all narcolepsy patients and is diagnostically very helpful

because it occurs in almost no other disorder. In contrast, occasional hypnagogic hallucinations and sleep paralysis occur in about 20% of the general population, and these symptoms are not as diagnostically specific.

When narcolepsy is suspected, the diagnosis should be firmly established with a polysomnogram followed the next day by an MSLT. The polysomnogram helps rule out other possible causes of sleepiness such as sleep apnea and establishes that the patient was not sleep deprived the night before, and the MSLT provides essential, objective evidence of sleepiness plus REM sleep dysregulation. Across the five naps of the MSLT, most patients with narcolepsy will fall asleep in <8 min on average, and they will have episodes of REM sleep in at least two of the naps. Abnormal regulation of REM sleep is also manifested by the appearance of REM sleep within 15 min of sleep onset at night, which is rare in healthy individuals sleeping at their habitual bedtime. Stimulants should be stopped 1 week before the MSLT and antidepressants should be stopped 3 weeks prior, because these medications can affect the MSLT. In addition, patients should be encouraged to obtain a fully adequate amount of sleep each night for the week prior to the test to eliminate any effects of insufficient sleep.

TREATMENT

TREATMENT

Narcolepsy

The treatment of narcolepsy is symptomatic. Most patients with narcolepsy feel more alert after sleep, and they should be encouraged to get adequate sleep each night and to take a 15- to 20-min nap in the afternoon. This nap may be sufficient for some patients with mild narcolepsy, but most also require treatment with wake-promoting medications. Modafinil is used quite often because it has fewer side effects than amphetamines and a relatively long half-life; for most patients, 200–400 mg each morning is very effective. Methylphenidate (10–20 mg bid) or dextroamphetamine (10 mg bid) are often effective, but sympathomimetic side effects, anxiety, and the potential for abuse can be concerns. These medications are available in slow-release formulations, extending their duration of action and allowing easier dosing. Sodium oxybate (gamma hydroxybutyrate) is given twice each night and is often very valuable in improving alertness, but it can produce excessive sedation, nausea, and confusion.

Cataplexy is usually much improved with antidepressants that increase noradrenergic or serotonergic tone because these neurotransmitters strongly suppress REM sleep and cataplexy. Venlafaxine (37.5–150 mg each morning) and fluoxetine (10–40 mg each morning) are often quite effective. The tricyclic antidepressants, such as protriptyline (10–40 mg/d) or clomipramine (25–50 mg/d) are potent suppressors of cataplexy, but their anticholinergic effects, including sedation and dry mouth, make them less attractive.[1] Sodium oxybate, given at bedtime and 3–4 h later, is also very helpful in reducing cataplexy.

[1]No antidepressant has been approved by the U.S. Food and Drug Administration (FDA) for treating narcolepsy.

EVALUATION OF INSOMNIA

Insomnia is the complaint of poor sleep and usually presents as difficulty initiating or maintaining sleep. People with insomnia are dissatisfied with their sleep and feel that it impairs their ability to function well in work, school, and social situations. Affected individuals often experience fatigue, decreased mood, irritability, malaise, and cognitive impairment.

Chronic insomnia, lasting >3 months, occurs in about 10% of adults and is more common in women, older adults, people of lower socioeconomic status, and individuals with medical, psychiatric, and substance abuse disorders. Acute or short-term insomnia affects over 30% of adults and is often precipitated by stressful life events such as a major illness or loss, change of occupation, medications, and substance abuse. If the acute insomnia triggers maladaptive behaviors such as increased nocturnal light exposure, frequently checking the clock, or attempting to sleep more by napping, it can lead to chronic insomnia.

Most insomnia begins in adulthood, but many patients may be predisposed and report easily disturbed sleep predating the insomnia, suggesting that their sleep is lighter than usual. Clinical studies and animal models indicate that insomnia is associated with activation during sleep of brain areas normally active only during wakefulness. The polysomnogram is rarely used in the evaluation of insomnia, as it typically confirms the patient's subjective report of long latency to sleep and numerous awakenings but usually adds little new information. Many patients with insomnia have increased fast (beta) activity in the EEG during sleep; this fast activity is normally present only during wakefulness, which may explain why some patients report feeling awake for much of the night. The MSLT is rarely used in the evaluation of insomnia because, despite their feelings of low energy, most people with insomnia do not easily fall asleep during the day, and on the MSLT, their average sleep latencies are usually longer than normal.

Many factors can contribute to insomnia, and obtaining a careful history is essential so one can select therapies targeting the underlying factors. The assessment should focus on identifying predisposing, precipitating, and perpetuating factors.

Psychophysiologic Factors

Many patients with insomnia have negative expectations and conditioned arousal that interfere with sleep. These individuals may worry about their insomnia during the day and have increasing anxiety as bedtime approaches if they anticipate a poor night of sleep. While attempting to sleep, they may frequently check the clock, which only heightens anxiety and frustration. They may find it easier to sleep in a new environment rather than their bedroom, as it lacks the negative associations.

Inadequate Sleep Hygiene

Patients with insomnia sometimes develop counterproductive behaviors that contribute to their insomnia. These can include daytime napping that reduces sleep drive at night; an irregular sleep-wake schedule that disrupts their circadian rhythms; use of wake-promoting substances (e.g., caffeine, tobacco) too close to bedtime; engaging in alerting or stressful activities close to bedtime (e.g., arguing with a partner, work-related emailing and texting while in bed, sleeping with a smartphone or tablet at the bedside); and routinely using the bedroom for activities other than sleep or sex (e.g., TV, work), so the bedroom becomes associated with arousing or stressful feelings.

Psychiatric Conditions

About 80% of patients with psychiatric disorders have sleep complaints, and about half of all chronic insomnia occurs in association with a psychiatric disorder. Depression is classically associated with early morning awakening, but it can also interfere with the onset and maintenance of sleep. Mania and hypomania can disrupt sleep and often are associated with substantial reductions in the total amount of sleep. Anxiety disorders can lead to racing thoughts and rumination that interfere with sleep and can be very problematic if the patient's mind becomes active midway through the night. Panic attacks can arise from sleep and need to be distinguished from other parasomnias. Insomnia is common in schizophrenia and other psychoses, often resulting in fragmented sleep, less deep NREM sleep, and sometimes reversal of the day-night sleep pattern.

Medications and Drugs of Abuse

A wide variety of psychoactive drugs can interfere with sleep. Caffeine, which has a half-life of 6–9 h, can disrupt sleep for up to 8–14 h, depending on the dose, variations in metabolism, and an individual's caffeine sensitivity. Insomnia can also result from use of prescription medications too close to bedtime (e.g., antidepressants, stimulants, glucocorticoids, theophylline). Conversely, withdrawal of sedating medications such as alcohol, narcotics, or benzodiazepines can cause insomnia. Alcohol taken just before bed can shorten sleep latency, but it often produces rebound insomnia 2–3 h later as it wears off. This same problem with sleep maintenance can occur with short-acting benzodiazepines such as alprazolam.

Medical Conditions

A large number of medical conditions disrupt sleep. Pain from rheumatologic disorders or a painful neuropathy commonly disrupts sleep. Some patients may sleep poorly because of respiratory conditions such as asthma, chronic obstructive pulmonary disease, cystic fibrosis, congestive heart failure, or restrictive lung disease, and some of these disorders are worse at night in bed due to circadian variations in airway resistance and postural changes that can result in nocturnal dyspnea. Many women experience poor sleep with the hormonal changes of menopause. Gastroesophageal reflux is also a common cause of difficulty sleeping.

Neurologic Disorders

Dementia (Chap. 25) is often associated with poor sleep, probably due to a variety of factors, including napping during the day, altered circadian rhythms, and perhaps a weakened output of the brain's sleep-promoting mechanisms. In fact, insomnia and nighttime wandering are some of the most common causes for institutionalization of patients with dementia, because they place a larger burden on caregivers. Conversely, in cognitively intact elderly men, fragmented sleep and poor sleep quality are associated with subsequent cognitive decline. Patients with Parkinson's disease may sleep poorly due to rigidity, dementia, and other factors. Fatal familial insomnia is a very rare neurodegenerative condition caused by mutations in the prion protein gene, and although insomnia is a common early symptom, most patients present with other obvious neurologic signs such dementia, myoclonus, dysarthria, or autonomic dysfunction.

TREATMENT

TREATMENT

Insomnia

Treatment of insomnia improves quality of life and can promote long-term health. With improved sleep, patients often report less daytime fatigue, improved cognition, and more energy. Treating the insomnia can also improve the comorbid disease. For example, management of insomnia at the time of diagnosis of major depression often improves the response to antidepressants and reduces the risk of relapse. Sleep loss can heighten the perception of pain, so a similar approach is warranted in acute and chronic pain management.

The treatment plan should target all putative contributing factors: establish good sleep hygiene, treat medical disorders, use behavioral therapies for anxiety and negative conditioning, and use pharmacotherapy and/or psychotherapy for psychiatric disorders. Behavioral therapies should be the first-line treatment, followed by judicious use of sleep-promoting medications if needed.

TREATMENT OF MEDICAL AND PSYCHIATRIC DISEASE

If the history suggests that a medical or psychiatric disease contributes to the insomnia, then it should be addressed by, for example, treating the pain, improving breathing, and switching or adjusting the timing of medications.

IMPROVE SLEEP HYGIENE

Attention should be paid to improving sleep hygiene and avoiding counterproductive, arousing behaviors before bedtime. Patients should establish a regular bedtime and wake time, even on weekends, to help synchronize their circadian rhythms and sleep patterns. The amount of time allocated for sleep should not be more than their actual total amount of sleep. In the 30 min before bedtime, patients should establish a relaxing “wind-down” routine that can include a warm bath, listening to music, meditation, or other relaxation techniques. The bedroom should be off-limits to computers, televisions, radios, smartphones, videogames, and tablets. Once in bed, patients should try to avoid thinking about anything stressful or arousing such as problems with relationships or work. If they cannot fall asleep within 20 min, it often helps to get out of bed and read or listen to relaxing music in dim light as a form of distraction from any anxiety, but artificial light, including light from a television, cell phone, or computer, should be avoided, because light itself suppresses melatonin secretion and is arousing.

Table 27-2 outlines some of the key aspects of good sleep hygiene to improve insomnia.

COGNITIVE BEHAVIORAL THERAPY

CBT uses a combination of the techniques above plus additional methods to improve insomnia. A trained therapist may use cognitive psychology techniques to reduce excessive worrying about sleep and to reframe faulty beliefs about the insomnia and its daytime consequences. The therapist may also teach the patient relaxation techniques, such as progressive muscle relaxation or meditation, to reduce autonomic arousal, intrusive thoughts, and anxiety.

MEDICATIONS FOR INSOMNIA

If insomnia persists after treatment of these contributing factors, pharmacotherapy is often used on a nightly or intermittent basis. A variety of sedatives can improve sleep.

Antihistamines, such as diphenhydramine, are the primary active ingredient in most over-the-counter sleep aids. These may be of benefit when used intermittently, but can produce tolerance and anticholinergic side effects such as dry mouth and constipation, which limit their use, particularly in the elderly.

Benzodiazepine receptor agonists (BzRAs) are an effective and well-tolerated class of medications for insomnia. BzRAs bind to the $GABA_A$ receptor and potentiate the postsynaptic response to GABA. $GABA_A$ receptors are found throughout the brain, and BzRAs may globally reduce neural activity and may enhance the activity of specific sleep-promoting GABAergic pathways. Classic BzRAs include lorazepam, triazolam, and clonazepam, whereas newer agents such as zolpidem and zaleplon have more selective affinity for the α_1 subunit of the $GABA_A$ receptor.

Specific BzRAs are often chosen based on the desired duration of action. The most commonly prescribed agents in this family are zaleplon (5–20 mg), with a half-life of 1–2 h; zolpidem (5–10 mg) and triazolam (0.125–0.25 mg), with half-lives of 2–4 h; eszopiclone (1–3 mg), with a half-life of 5–8 h; and temazepam (15–30 mg), with a half-life of 8–20 h. Generally, side effects are minimal when the dose is kept low and the serum concentration is minimized during the waking hours (by using the shortest-acting effective agent). For chronic insomnia, intermittent use is recommended, unless the consequences of untreated insomnia outweigh concerns regarding chronic use.

The heterocyclic *antidepressants* (trazodone, amitriptyline,[2] and doxepin) are the most commonly prescribed alternatives to BzRAs due to their lack of abuse potential and lower cost. Trazodone (25–100 mg) is used more commonly than the tricyclic antidepressants, because it has a much shorter half-life (5–9 h) and less anticholinergic activity.

The orexin receptor antagonist suvorexant (10–20 mg) can also improve insomnia by blocking the wake-promoting effects of the orexin neuropeptides. It has a long half-life and can produce morning sedation, and as it reduces orexin signaling, it can rarely produce hypnagogic hallucinations and sleep paralysis (see narcolepsy section above).

Medications for insomnia are now among the most commonly prescribed medications, but they should be used cautiously. All sedatives increase the risk of injurious falls and confusion in the elderly, and therefore if needed, these medications should be used at the lowest effective dose. Morning sedation can interfere with driving and judgment, and when selecting a medication, one should consider the duration of action. Benzodiazepines carry a risk of addiction and abuse, especially in patients with a history of alcohol or sedative abuse. In patients with depression, all sedatives can worsen the depression. Like alcohol, some sleep-promoting medications can worsen sleep apnea. Sedatives can also produce complex behaviors during sleep, such as sleep walking and sleep eating, although this seems more likely at higher doses.

[2]Trazodone and amitriptyline have not been approved by the FDA for treating insomnia.

TABLE 27-2

Methods to Improve Sleep Hygiene in Insomnia Patients

Helpful Behaviors	Behaviors to Avoid
Use the bed only for sleep and sex • If you cannot sleep within 20 min, get out of bed and read or do other relaxing activities in dim light before returning to bed	Avoid behaviors that interfere with sleep physiology, including: • Napping, especially after 3:00 PM • Attempting to sleep too early • Caffeine after lunchtime
Make quality sleep a priority • Go to bed and get up at the same time each day • Ensure a restful environment (comfortable bed, bedroom quiet and dark)	In the 2–3 h before bedtime, avoid: • Heavy eating • Smoking or alcohol • Vigorous exercise
Develop a consistent bedtime routine. For example: • Prepare for sleep with 20–30 min of relaxation (e.g., soft music, meditation, yoga, pleasant reading) • Take a warm bath	When trying to fall asleep, avoid: • Solving problems • Thinking about life issues • Reviewing events of the day

RESTLESS LEGS SYNDROME

Patients with restless legs syndrome (RLS) report an irresistible urge to move the legs. Many patients report a creepy-crawly or unpleasant deep ache within the thighs or calves, and those with more severe RLS may have discomfort in the arms as well. For most patients with RLS, these dysesthesias and restlessness are much worse in the evening and first half of the night. The symptoms appear with inactivity and can make sitting still in an airplane or when watching a movie a miserable experience. The sensations are temporarily relieved by movement, stretching, or massage. This nocturnal discomfort usually interferes with sleep, and patients may report daytime sleepiness as a consequence. RLS is very common, affecting 5–10% of adults and is more common in women and older adults.

A variety of factors can cause RLS. Iron deficiency is the most common treatable cause, and iron replacement should be considered if the ferritin level is <75 ng/mL. RLS can also occur with peripheral neuropathies and uremia and can be worsened by pregnancy, caffeine, alcohol, antidepressants, lithium, neuroleptics, and antihistamines. Genetic factors contribute to RLS, and polymorphisms in a variety of genes

(*BTBD9, MEIS1, MAP2K5/LBXCOR*, and *PTPRD*) have been linked to RLS, although as yet, the mechanism through which they cause RLS remains unknown. Roughly one-third of patients (particularly those with an early age of onset) have multiple affected family members.

RLS is treated by addressing the underlying cause such as iron deficiency if present. Otherwise, treatment is symptomatic, and dopamine agonists or alpha-2-delta calcium channel ligands are used most frequently. Agonists of dopamine $D_{2/3}$ receptors such as pramipexole (0.25–0.5 mg q7PM) or ropinirole (0.5–4 mg q7PM) are usually quite effective, but about 25% of patients taking dopamine agonists develop augmentation, a worsening of RLS such that symptoms begin earlier in the day and can spread to other body regions. Other possible side effects of dopamine agonists include nausea, morning sedation, and increases in rewarding behavior such as gambling and sex. Alpha-2-delta calcium channel ligands such as gabapentin (300–600 mg q7PM) and pregabalin (150–450 mg q7PM) can also be quite effective; these do not cause augmentation and they can be especially helpful in patients with concomitant pain, neuropathy or anxiety. Opioids and benzodiazepines may also be of therapeutic value. Most patients with restless legs also experience PLMD, although the reverse is not the case.

PERIODIC LIMB MOVEMENT DISORDER

PLMD involves rhythmic twitches of the legs that disrupt sleep. The movements resemble a triple flexion reflex with extensions of the great toe and dorsiflexion of the foot for 0.5–5.0 s, which recur every 20–40 s during NREM sleep, in episodes lasting from minutes to hours. PLMD is diagnosed by a polysomnogram that includes recordings of the anterior tibialis and sometimes other muscles. The EEG shows that the movements of PLMD frequently cause brief arousals that disrupt sleep and can cause insomnia and daytime sleepiness. PLMD can be caused by the same factors that cause RLS (see above), and the frequency of leg movements improves with the same medications as used for RLS, including dopamine agonists. Recent genetic studies identified polymorphisms associated with both RLS and PLMD, suggesting that they may have a common pathophysiology.

PARASOMNIAS

Parasomnias are abnormal behaviors or experiences that arise from or occur during sleep. A variety of parasomnias can occur during NREM sleep, from brief confusional arousals to sleepwalking and night terrors. The presenting complaint is usually related to the behavior itself, but the parasomnias can disturb sleep continuity or lead to mild impairments in daytime alertness. Two main parasomnias occur in REM sleep: REM sleep behavior disorder (RBD) and nightmares.

Sleepwalking (Somnambulism)

Patients affected by this disorder carry out automatic motor activities that range from simple to complex. Individuals may walk, urinate inappropriately, eat, exit the house, or drive a car with minimal awareness. It may be difficult to arouse the patient to wakefulness, and occasional individuals may respond to attempted awakening with agitation or violence. In general it is safest to lead the patient back to bed, at which point he or she will often fall back asleep. Sleepwalking arises from NREM stage N3 sleep, usually in the first few hours of the night, and the EEG initially shows the slow cortical activity of deep NREM sleep even when the patient is moving about. Sleepwalking is most common in children and adolescents, when deep NREM sleep is most abundant. About 15% of children have occasional sleepwalking, and it persists in about 1% of adults. Episodes are usually isolated but may be recurrent in 1–6% of patients. The cause is unknown, although it has a familial basis in roughly one-third of cases. Sleepwalking can be worsened by insufficient sleep, which subsequently causes an increase in deep NREM sleep; alcohol; and stress. These should be addressed if present. Small studies have shown some efficacy of antidepressants and benzodiazepines; relaxation techniques and hypnosis can also be helpful. Patients and their families should improve home safety (e.g., replace glass doors, remove low tables to avoid tripping) to minimize the chance of injury if sleepwalking occurs.

Sleep Terrors

This disorder occurs primarily in young children during the first few hours of sleep during NREM stage N3 sleep. The child often sits up during sleep and screams, exhibiting autonomic arousal with sweating, tachycardia, large pupils, and hyperventilation. The individual may be difficult to arouse and rarely recalls the episode on awakening in the morning. Treatment usually consists of reassuring the parents that the condition is self-limited and benign, and like sleepwalking, it may improve by avoiding insufficient sleep.

Sleep Enuresis

Bedwetting, like sleepwalking and night terrors, is another parasomnia that occurs during sleep in the young. Before age 5 or 6 years, nocturnal enuresis should be considered a normal feature of development. The condition usually improves spontaneously by puberty, persists in 1–3% of adolescents, and is rare in adulthood. Treatment consists of bladder training exercises and behavioral therapy.

Symptomatic pharmacotherapy is usually accomplished in adults with desmopressin (0.2 mg qhs), oxybutynin chloride (5 mg qhs), or imipramine (10–25 mg qhs). Important causes of nocturnal enuresis in patients who were previously continent for 6–12 months include urinary tract infections or malformations, cauda equina lesions, emotional disturbances, epilepsy, sleep apnea, and certain medications.

Sleep Bruxism

Bruxism is an involuntary, forceful grinding of teeth during sleep that affects 10–20% of the population. The patient is usually unaware of the problem. The typical age of onset is 17–20 years, and spontaneous remission usually occurs by age 40. In many cases, the diagnosis is made during dental examination, damage is minor, and no treatment is indicated. In more severe cases, treatment with a mouth guard is necessary to prevent tooth injury. Stress management, benzodiazepines, and biofeedback can be useful when bruxism is a manifestation of psychological stress.

REM Sleep Behavior Disorder (RBD)

RBD **(Video 27-2)** is distinct from other parasomnias in that it occurs during REM sleep. The patient or the bed partner usually reports agitated or violent behavior during sleep, and upon awakening, the patient can often report a dream that matches the accompanying movements. During normal REM sleep, nearly all non-respiratory skeletal muscles are paralyzed, but in patients with RBD, dramatic limb movements such as punching or kicking lasting seconds to minutes occur during REM sleep, and it is not uncommon for the patient or the bed partner to be injured.

Video 027-2: REM Sleep Behavior Disorder

Typical aggressive movements in rapid eye movement (REM) sleep behavior disorder. *(Video courtesy of Dr. Carlos Schenck, University of Minnesota Medical School.)*

Image not available.
Play Video

The prevalence of RBD increases with age, afflicting about 2% of adults aged >70, and is about twice as common in men. Most already have or will develop a neurodegenerative disorder. Within 12 years of disease onset, half of RBD patients develop a synucleinopathy such as Parkinson's disease **(Chap. 427)** or dementia with Lewy bodies **(Chap. 426)**, or occasionally multiple system atrophy **(Chap. 432)**, and over 90% develop a synucleinopathy by 25 years. RBD can occur in patients taking antidepressants, and in some, these medications may unmask this early indicator of neurodegeneration. Synucleinopathies probably cause neuronal loss in brainstem regions that regulate muscle paralysis during REM sleep, and loss of these neurons permits movements to break through during REM sleep. RBD also occurs in about 30% of patients with narcolepsy, but the underlying cause is probably different, as they seem to be at no increased risk of a neurodegenerative disorder.

Many patients with RBD have sustained improvement with clonazepam (0.5–2.0 mg qhs).[3] Melatonin at doses up to 9 mg nightly may also prevent attacks.

[3]No medications have been approved by the FDA for the treatment of RBD.

CIRCADIAN RHYTHM SLEEP DISORDERS

A subset of patients presenting with either insomnia or hypersomnia may have a disorder of sleep *timing* rather than sleep *generation*. Disorders of sleep timing can be either organic (i.e., due to an abnormality of circadian pacemaker[s]) or environmental/behavioral (i.e., due to a disruption of environmental synchronizers). Effective therapies aim to entrain the circadian rhythm of sleep propensity to an appropriate phase.

Delayed Sleep-Wake Phase Disorder

DSWPD is characterized by: (1) reported sleep onset and wake times persistently later than desired; (2) actual sleep times at nearly the same clock hours daily; and (3) if conducted at the habitual delayed sleep time, essentially normal sleep on polysomnography (except for delayed sleep onset). Patients with DSWPD exhibit an abnormally delayed endogenous circadian phase, which can be assessed by measuring the onset of secretion of melatonin in either the blood or saliva; this is best done in a dimly lit environment as light suppresses melatonin secretion. Dim-light melatonin onset (DLMO) in DSWPD patients occurs later in the evening than normal, which is about 8:00–9:00 P.M. (i.e., about 1–2 h before habitual bedtime). Patients tend to be young adults. The delayed circadian phase could be due to: (1) an abnormally long, genetically determined intrinsic period of the endogenous circadian pacemaker; (2) reduced phase-advancing capacity of the pacemaker; (3) slower rate of buildup of homeostatic sleep drive during wakefulness; or (4) an irregular prior sleep-wake schedule, characterized by frequent

nights when the patient chooses to remain awake while exposed to artificial light well past midnight (for personal, social, school, or work reasons). In most cases, it is difficult to distinguish among these factors, as patients with either a behaviorally induced or biologically driven circadian phase delay may both exhibit a similar circadian phase delay in DLMO, and both factors make it difficult to fall asleep at the desired hour. Late onset of dim-light melatonin secretion can help distinguish DSWD from other forms of sleep-onset insomnia. DSWPD is a chronic condition that can persist for years and may not respond to attempts to reestablish normal bedtime hours. Treatment methods involving phototherapy with blue-enriched light during the morning hours and/or melatonin administration in the evening hours show promise in these patients, although the relapse rate is high.

Advanced Sleep-Wake Phase Disorder

Advanced sleep-wake phase disorder (ASWPD) is the converse of DSWPD. Most commonly, this syndrome occurs in older people, 15% of whom report that they cannot sleep past 5:00 A.M., with twice that number complaining that they wake up too early at least several times per week. Patients with ASWPD are sleepy during the evening hours, even in social settings. Sleep-wake timing in ASWPD patients can interfere with a normal social life. Patients with this circadian rhythm sleep disorder can be distinguished from those who have early wakening due to insomnia because ASWPD patients show early onset of dim-light melatonin secretion.

In addition to age-related ASWPD, an early-onset familial variant of this condition has also been reported. In two families in which ASWPD was inherited in an autosomal dominant pattern, the syndrome was due to missense mutations in a circadian clock component (in the casein kinase binding domain of *PER2* in one family, and in casein kinase I delta in the other) that shortens the circadian period. Patients with ASWPD may benefit from bright light and/or blue enriched phototherapy during the evening hours to reset the circadian pacemaker to a later hour.

Non-24-h Sleep-Wake Rhythm Disorder

Non-24-h sleep-wake rhythm disorder (N24SWRD) most commonly occurs when the primary synchronizing input (i.e., the light-dark cycle) from the environment to the circadian pacemaker is lost (as occurs in many blind people with no light perception), and the maximal phase-advancing capacity of the circadian pacemaker in response to non-photic cues cannot accommodate the difference between the 24-h geophysical day and the intrinsic period of the patient's circadian pacemaker, resulting in loss of entrainment to the 24-h day. The sleep of most blind patients with N24SWRD is restricted to the nighttime hours due to social or occupational demands. Despite this regular sleep-wake schedule, affected patients with N24SWRD are nonetheless unable to maintain a stable phase relationship between the output of the non-entrained circadian pacemaker and the 24-h day. Therefore, most blind patients present with intermittent bouts of insomnia. When the blind patient's endogenous circadian rhythms are out of phase with the local environment, nighttime insomnia coexists with excessive daytime sleepiness. Conversely, when the endogenous circadian rhythms of those same patients are in phase with the local environment, symptoms remit. The interval between symptomatic phases may last several weeks to several months in blind patients with N24SWRD, depending on the period of the underlying nonentrained rhythm and the 24-h day. Nightly low-dose (0.5 mg) melatonin administration may improve sleep and, in some cases, induce synchronization of the circadian pacemaker. In sighted patients, N24SWRD is usually caused by self-selected exposure to artificial light that inadvertently entrains the circadian pacemaker to a >24-h schedule, and these individuals present with an incremental pattern of successive delays in sleep timing, progressing in and out of phase with local time—a clinical presentation that is seldom seen in blind patients with N24SWRD.

Shift-Work Disorder

More than 7 million workers in the United States regularly work at night, either on a permanent or rotating schedule. Many more begin the commute to work or school between 4:00 A.M. and 7:00 A.M., requiring them to commute and then work during a time of day that they would otherwise be asleep. In addition, each week, millions of "day" workers and students elect to remain awake at night or awaken very early in the morning to work or study to meet work or school deadlines, drive long distances, compete in sporting events, or participate in recreational activities. Such schedules can result in both sleep loss and misalignment of circadian rhythms with respect to the sleep-wake cycle.

The circadian timing system usually fails to adapt successfully to the inverted schedules required by overnight work or the phase advance required by early morning (4:00 A.M. to 7:00 A.M.) start times. This leads to a misalignment between the desired work-rest schedule and the output of the pacemaker and to disturbed daytime sleep in most such individuals. Excessive work hours (per day or per week), insufficient time off between consecutive days of work or school, and frequent travel across time zones may be contributing factors. Sleep deficiency, increased length of time awake prior to work, and misalignment of circadian phase produce decreased alertness and performance, increased reaction time, and increased risk of performance lapses, thereby resulting in greater safety hazards among night workers and other sleep-deprived individuals. Sleep disturbance nearly doubles the risk of a fatal work accident. In addition, long-term night shift workers have higher

rates of breast, colorectal, and prostate cancer and of cardiac, gastrointestinal, metabolic, and reproductive disorders. The World Health Organization has added night-shift work to its list of probable carcinogens.

Sleep onset begins in local brain regions before gradually sweeping over the entire brain as sensory thresholds rise and consciousness is lost. A sleepy individual struggling to remain awake may attempt to continue performing routine and familiar motor tasks during the transition state between wakefulness and stage N1 sleep, while unable to adequately process sensory input from the environment. Such sleep-related attentional failures typically last only seconds but are known on occasion to persist for longer durations. Motor vehicle operators who fail to heed the warning signs of sleepiness are especially vulnerable to sleep-related accidents, as sleep processes can slow reaction times, induce automatic behavior, and intrude involuntarily upon the waking brain, causing catastrophic consequences—including 6400 fatalities and 50,000 debilitating injuries in the United States annually. For this reason, an expert consensus panel has concluded that individuals who have slept <2 h in the prior 24 h are unfit to drive a motor vehicle. There is a significant increase in the risk of sleep-related, fatal-to-the-driver highway crashes in the early morning and late afternoon hours, coincident with bimodal peaks in the daily rhythm of sleep tendency.

Physicians who work prolonged shifts, especially intermittent overnight shifts, constitute another group of workers at greater risk for accidents and other adverse consequences of lack of sleep and misalignment of the circadian rhythm. Recurrent scheduling of resident physicians to work shifts of ≥24 consecutive hours impairs psychomotor performance to a degree that is comparable to alcohol intoxication, doubles the risk of attentional failures among intensive care unit resident physicians working at night, and significantly increases the risk of serious medical errors in intensive care units, including a fivefold increase in the risk of serious diagnostic mistakes. Some 20% of hospital resident physicians report making a fatigue-related mistake that injured a patient, and 5% admit making a fatigue-related mistake that resulted in the death of a patient. Moreover, working for >24 consecutive hours increases the risk of percutaneous injuries and more than doubles the risk of motor vehicle crashes during the commute home. For these reasons, in 2008, the National Academy of Medicine concluded that the practice of scheduling resident physicians to work for >16 consecutive hours without sleep is hazardous for both resident physicians and their patients.

From 5 to 15% of individuals scheduled to work at night or in the early morning hours have much greater-than-average difficulties remaining awake during night work and sleeping during the day; these individuals are diagnosed with chronic and severe shift-work disorder (SWD). Patients with this disorder have a level of excessive sleepiness during work at night or in the early morning and insomnia during day sleep that the physician judges to be clinically significant; the condition is associated with an increased risk of sleep-related accidents and with some of the illnesses associated with night-shift work. Patients with chronic and severe SWD are profoundly sleepy at work. In fact, their sleep latencies during night work average just 2 min, comparable to mean daytime sleep latency durations of patients with narcolepsy or severe sleep apnea.

TREATMENT

TREATMENT

Shift-Work Disorder

Caffeine is frequently used by night workers to promote wakefulness. However, it cannot forestall sleep indefinitely, and it does not shield users from sleep-related performance lapses. Postural changes, exercise, and strategic placement of nap opportunities can sometimes temporarily reduce the risk of fatigue-related performance lapses. Properly timed exposure to blue-enriched light or bright white light can directly enhance alertness and facilitate more rapid adaptation to night-shift work.

Modafinil (200 mg) or armodafinil (150 mg) 30–60 min before the start of an 8-h overnight shift is an effective treatment for the excessive sleepiness during night work in patients with SWD. Although treatment with modafinil or armodafinil significantly improves performance and reduces sleep propensity and the risk of lapses of attention during night work, affected patients remain excessively sleepy.

Fatigue risk management programs for night shift workers should promote education about sleep, increase awareness of the hazards associated with sleep deficiency and night work, and screen for common sleep disorders. Work schedules should be designed to minimize: (1) exposure to night work; (2) the frequency of shift rotations; (3) the number of consecutive night shifts; and (4) the duration of night shifts.

Jet Lag Disorder

Each year, >60 million people fly from one time zone to another, often resulting in excessive daytime sleepiness, sleep-onset insomnia, and frequent arousals from sleep, particularly in the latter half of the night. The syndrome is transient, typically lasting 2–14 d depending on the number of time zones crossed, the direction of travel, and the traveler's age and phase-shifting capacity. Travelers who spend more time

outdoors at their destination reportedly adapt more quickly than those who remain in hotel or seminar rooms, presumably due to brighter (outdoor) light exposure. Avoidance of antecedent sleep loss or napping on the afternoon prior to overnight travel can reduce the difficulties associated with extended wakefulness. Laboratory studies suggest that low doses of melatonin can enhance sleep efficiency, but only if taken when endogenous melatonin concentrations are low (i.e., during the biologic daytime).

In addition to jet lag associated with travel across time zones, many patients report a behavioral pattern that has been termed *social jet lag*, in which bedtimes and wake times on weekends or days off occur 4–8 h later than during the week. Such recurrent displacement of the timing of the sleep-wake cycle is common in adolescents and young adults and is associated with delayed circadian phase, sleep-onset insomnia, excessive daytime sleepiness, poorer academic performance, and increased risk of both obesity and depressive symptoms.

MEDICAL IMPLICATIONS OF CIRCADIAN RHYTHMICITY

Prominent circadian variations have been reported in the incidence of acute myocardial infarction, sudden cardiac death, and stroke, the leading causes of death in the United States. Platelet aggregability is increased in the early morning hours, coincident with the peak incidence of these cardiovascular events. Recurrent circadian disruption combined with chronic sleep deficiency, such as occurs during night-shift work, is associated with increased plasma glucose concentrations after a meal due to inadequate pancreatic insulin secretion. Night shift workers with elevated fasting glucose have an increased risk of progressing to diabetes. Blood pressure of night workers with sleep apnea is higher than that of day workers. A better understanding of the possible role of circadian rhythmicity in the acute destabilization of a chronic condition such as atherosclerotic disease could improve the understanding of its pathophysiology.

Diagnostic and therapeutic procedures may also be affected by the time of day at which data are collected. Examples include blood pressure, body temperature, the dexamethasone suppression test, and plasma cortisol levels. The timing of chemotherapy administration has been reported to have an effect on the outcome of treatment. In addition, both the toxicity and effectiveness of drugs can vary with time of day. For example, more than a fivefold difference has been observed in mortality rates following administration of toxic agents to experimental animals at different times of day. Anesthetic agents are particularly sensitive to time-of-day effects. Finally, the physician must be aware of the public health risks associated with the ever-increasing demands made by the 24/7 schedules in our round-the-clock society.

ACKNOWLEDGMENT

John W. Winkelman, MD, PhD and Gary S. Richardson, MD contributed to this chapter in prior editions, and some material from their work has been retained here.

FURTHER READING

Ding F et al: Changes in the composition of brain interstitial ions control the sleep-wake cycle. Science 352:550, 2016. [PubMed: 27126038]

Ju YE et al: Sleep and Alzheimer disease pathology—A bidirectional relationship. Nat Rev Neurol 10:115, 2014. [PubMed: 24366271]

Lee ML et al: High risk of near-crash driving events following night-shift work. Proc Natl Acad Sci USA 113:176, 2016. [PubMed: 26699470]

Lim AS et al: Sleep is related to neuron numbers in the ventrolateral preoptic/intermediate nucleus in older adults with and without Alzheimer's disease. Brain 137:2847, 2014. [PubMed: 25142380]

Liu Y et al: Prevalence of healthy sleep duration among adults—United States, 2014. MMWR Morb Mortal Wkly Rep 65:137, 2016. [PubMed: 26890214]

Riemann D et al: The neurobiology, investigation, and treatment of chronic insomnia. Lancet Neurol 14:547, 2015. [PubMed: 25895933]

Scammell TE: Narcolepsy. N Engl J Med 373:2654, 2015. [PubMed: 26716917]

Scammell TE et al: Neural circuitry of wakefulness and sleep. Neuron 93:747, 2017. [PubMed: 28231463]

Stothard ER et al: Circadian entrainment to the natural light-dark cycle across seasons and the weekend. Curr Biol 27:508, 2017. [PubMed: 28162893]

Xie L et al: Sleep drives metabolite clearance from the adult brain. Science 342:373, 2013. [PubMed: 24136970]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 28: Disorders of the Eye

Jonathan C. Horton

THE HUMAN VISUAL SYSTEM

The visual system provides a supremely efficient means for the rapid assimilation of information from the environment to aid in the guidance of behavior. The act of seeing begins with the capture of images focused by the cornea and lens on a light-sensitive membrane in the back of the eye called the *retina*. The retina is actually part of the brain, banished to the periphery to serve as a transducer for the conversion of patterns of light energy into neuronal signals. Light is absorbed by pigment in two types of photoreceptors: rods and cones. In the human retina there are 100 million rods and 5 million cones. The rods operate in dim (scotopic) illumination. The cones function under daylight (photopic) conditions. The cone system is specialized for color perception and high spatial resolution. The majority of cones are within the macula, the portion of the retina that serves the central 10° of vision. In the middle of the macula a small pit termed the *fovea*, packed exclusively with cones, provides the best visual acuity.

Photoreceptors hyperpolarize in response to light, activating bipolar, amacrine, and horizontal cells in the inner nuclear layer. After processing of photoreceptor responses by this complex retinal circuit, the flow of sensory information ultimately converges on a final common pathway: the ganglion cells. These cells translate the visual image impinging on the retina into a continuously varying barrage of action potentials that propagates along the primary optic pathway to visual centers within the brain. There are a million ganglion cells in each retina and hence a million fibers in each optic nerve.

Ganglion cell axons sweep along the inner surface of the retina in the nerve fiber layer, exit the eye at the optic disc, and travel through the optic nerve, optic chiasm, and optic tract to reach targets in the brain. The majority of fibers synapse on cells in the lateral geniculate body, a thalamic relay station. Cells in the lateral geniculate body project in turn to the primary visual cortex. This afferent retinogeniculocortical sensory pathway provides the neural substrate for visual perception. Although the lateral geniculate body is the main target of the retina, separate classes of ganglion cells project to other subcortical visual nuclei involved in different functions. Ganglion cells that mediate pupillary constriction and circadian rhythms are light sensitive owing to a novel visual pigment, melanopsin. Pupil responses are mediated by input to the pretectal olivary nuclei in the midbrain. The pretectal nuclei send their output to the Edinger-Westphal nuclei, which in turn provide parasympathetic innervation to the iris sphincter via an interneuron in the ciliary ganglion. Circadian rhythms are timed by a retinal projection to the suprachiasmatic nucleus. Visual orientation and eye movements are served by retinal input to the superior colliculus. Gaze stabilization and optokinetic reflexes are governed by a group of small retinal targets known collectively as the *brainstem accessory optic system*.

The eyes must be rotated constantly within their orbits to place and maintain targets of visual interest on the fovea. This activity, called *foveation*, or looking, is governed by an elaborate efferent motor system. Each eye is moved by six extraocular muscles that are supplied by cranial nerves from the oculomotor (III), trochlear (IV), and abducens (VI) nuclei. Activity in these ocular motor nuclei is coordinated by pontine and midbrain mechanisms for smooth pursuit, saccades, and gaze stabilization during head and body movements. Large regions of the frontal and parietooccipital cortex control these brainstem eye movement centers by providing descending supranuclear input.

CLINICAL ASSESSMENT OF VISUAL FUNCTION

REFRACTIVE STATE

In approaching a patient with reduced vision, the first step is to decide whether refractive error is responsible. In *emmetropia*, parallel rays from infinity are focused perfectly on the retina. Sadly, this condition is enjoyed by only a minority of the population. In *myopia*, the globe is too long, and light rays come to a focal point in front of the retina. Near objects can be seen clearly, but distant objects require a diverging lens in front of the eye. In *hyperopia*, the globe is too short, and hence a converging lens is used to supplement the refractive

power of the eye. In *astigmatism*, the corneal surface is not perfectly spherical, necessitating a cylindrical corrective lens. Most patients elect to wear eyeglasses or contact lenses to neutralize refractive error. An alternative is to permanently alter the refractive properties of the cornea by performing laser in situ keratomileusis (LASIK) or photorefractive keratectomy (PRK).

With the onset of middle age, *presbyopia* develops as the lens within the eye becomes unable to increase its refractive power to accommodate on near objects. To compensate for presbyopia an emmetropic patient must use reading glasses. A patient already wearing glasses for distance correction usually switches to bifocals. The only exception is a myopic patient, who may achieve clear vision at near simply by removing glasses containing the distance prescription.

Refractive errors usually develop slowly and remain stable after adolescence, except in unusual circumstances. For example, the acute onset of diabetes mellitus can produce sudden myopia because of lens edema induced by hyperglycemia. Testing vision through a pinhole aperture is a useful way to screen quickly for refractive error. If visual acuity is better through a pinhole than it is with the unaided eye, the patient needs refraction to obtain best corrected visual acuity.

VISUAL ACUITY

The Snellen chart is used to test acuity at a distance of 6 m (20 ft). For convenience, a scale version of the Snellen chart called the Rosenbaum card is held at 36 cm (14 in.) from the patient **(Fig. 28-1)**. All subjects should be able to read the 6/6 m (20/20 ft) line with each eye using their refractive correction, if any. Patients who need reading glasses because of presbyopia must wear them for accurate testing with the Rosenbaum card. If 6/6 (20/20) acuity is not present in each eye, the deficiency in vision must be explained. If it is worse than 6/240 (20/800), acuity should be recorded in terms of counting fingers, hand motions, light perception, or no light perception. Legal blindness is defined by the Internal Revenue Service as a best corrected acuity of 6/60 (20/200) or less in the better eye or a binocular visual field subtending 20° or less. Loss of vision in one eye only does not constitute legal blindness. For driving the laws vary by state, but most require a corrected acuity of 6/12 (20/40) in at least one eye for unrestricted privileges. Patients who develop a homonymous hemianopia should not drive.

FIGURE 28-1

The Rosenbaum card is a miniature, scale version of the Snellen chart for testing visual acuity at near. When the visual acuity is recorded, the Snellen distance equivalent should bear a notation indicating that vision was tested at near, not at 6 m (20 ft), or else the Jaeger number system should be used to report the acuity.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PUPILS

The pupils should be tested individually in dim light with the patient fixating on a distant target. There is no need to check the near response if the pupils respond briskly to light, because isolated loss of constriction (miosis) to accommodation does not occur. For this reason, the ubiquitous abbreviation PERRLA (pupils equal, round, and reactive to light and accommodation) implies a wasted effort with the last step. However, it is important to test the near response if the light response is poor or absent. Light-near dissociation occurs with neurosyphilis (Argyll Robertson pupil), with lesions of the dorsal midbrain (*Parinaud's syndrome*), and after aberrant regeneration (oculomotor nerve palsy, Adie's tonic pupil).

An eye with no light perception has no pupillary response to direct light stimulation. If the retina or optic nerve is only partially injured, the direct pupillary response will be weaker than the consensual pupillary response evoked by shining a light into the healthy fellow eye. A *relative afferent pupillary defect* (Marcus Gunn pupil) is elicited with the swinging flashlight test (Fig. 28-2). It is an extremely useful sign in retrobulbar optic neuritis and other optic nerve diseases, in which it may be the sole objective evidence for disease. In bilateral optic neuropathy, no afferent pupil defect is present if the optic nerves are affected equally.

FIGURE 28-2

Demonstration of a relative afferent pupil defect (Marcus Gunn pupil) in the left eye, done with the patient fixating on a distant target. ***A.*** With dim background lighting, the pupils are equal and relatively large. ***B.*** Shining a flashlight into the right eye evokes equal, strong constriction of both pupils. ***C.*** Swinging the flashlight over to the damaged left eye causes dilation of both pupils, although they remain smaller than in ***A***. Swinging the flashlight back over to the healthy right eye would result in symmetric constriction back to the appearance shown in ***B***. Note that the pupils always remain equal; the damage to the left retina/optic nerve is revealed by weaker bilateral pupil constriction to a flashlight in the left eye compared with the right eye. *(From P Levatin: Arch Ophthalmol 62:768, 1959. Copyright © 1959 American Medical Association. All rights reserved.)*



A



B



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Subtle inequality in pupil size, up to 0.5 mm, is a fairly common finding in normal persons. The diagnosis of essential or physiologic anisocoria is secure as long as the relative pupil asymmetry remains constant as ambient lighting varies. Anisocoria that increases in dim light indicates a sympathetic paresis of the iris dilator muscle. The triad of miosis with ipsilateral ptosis and anhidrosis constitutes *Horner's syndrome*, although anhidrosis is an inconstant feature. Brainstem stroke, carotid dissection, and neoplasm impinging on the sympathetic chain occasionally are identified as the cause of Horner's syndrome, but most cases are idiopathic.

Anisocoria that increases in bright light suggests a parasympathetic palsy. The first concern is an oculomotor nerve paresis. This possibility is excluded if the eye movements are full and the patient has no ptosis or diplopia. Acute pupillary dilation (mydriasis) can result from damage to the ciliary ganglion in the orbit. Common mechanisms are infection (herpes zoster, influenza), trauma (blunt, penetrating, surgical), and ischemia (diabetes, temporal arteritis). After denervation of the iris sphincter the pupil does not respond well to light, but the response to near is often relatively intact. When the near stimulus is removed, the pupil redilates very slowly compared with the normal pupil, hence the term *tonic pupil*. In *Adie's syndrome* a tonic pupil is present, sometimes in conjunction with weak or absent tendon reflexes in the lower extremities. This benign disorder, which occurs predominantly in healthy young women, is assumed to represent a mild dysautonomia. Tonic pupils are also associated with multiple system atrophy, segmental hypohidrosis, diabetes, and amyloidosis. Occasionally, a tonic pupil is discovered incidentally in an otherwise completely normal, asymptomatic individual. The diagnosis is confirmed by placing a drop of dilute (0.125%) pilocarpine into each eye. Denervation hypersensitivity produces pupillary constriction in a tonic pupil, whereas the normal pupil shows no response. Pharmacologic dilatation from accidental or deliberate instillation of anticholinergic (atropine, scopolamine) drops can produce pupillary mydriasis. Gardener's pupil refers to mydriasis induced by exposure to tropane alkaloids, contained in plants such as deadly nightshade, jimsonweed, or angel's trumpet. When an anticholinergic agent is responsible for pupil dilation, 1% pilocarpine causes no constriction.

Both pupils are affected equally by systemic medications. They are small with narcotic use (morphine, oxycodone) and large with anticholinergics (scopolamine). Parasympathetic agents (pilocarpine) used to treat glaucoma produce miosis. In any patient with an unexplained pupillary abnormality, a slit-lamp examination is helpful to exclude surgical trauma to the iris, an occult foreign body, perforating injury, intraocular inflammation, adhesions (synechia), angle-closure glaucoma, and iris sphincter rupture from blunt trauma.

EYE MOVEMENTS AND ALIGNMENT

Eye movements are tested by asking the patient, with both eyes open, to pursue a small target such as a pen tip into the cardinal fields of gaze. Normal ocular versions are smooth, symmetric, full, and maintained in all directions without nystagmus. Saccades, or quick refixation eye movements, are assessed by having the patient look back and forth between two stationary targets. The eyes should move rapidly and accurately in a single jump to their target. Ocular alignment can be judged by holding a penlight directly in front of the patient at about 1 m. If the eyes are straight, the corneal light reflex will be centered in the middle of each pupil. To test eye alignment more precisely, the cover test is useful. The patient is instructed to look at a small fixation target in the distance. One eye is occluded with a paddle or hand, while the other eye is observed. If the viewing eye shifts position to take up fixation on the target, it was misaligned. If it remains motionless, the first eye is uncovered and the test is repeated on the second eye. If neither eye moves the eyes are aligned orthotropically. If the eyes are orthotropic in primary gaze but the patient complains of diplopia, the cover test should be performed with the head tilted or turned in whatever direction elicits diplopia. With practice, the examiner can detect an ocular deviation (heterotropia) as small as 1–2° with the cover test. In a patient with vertical diplopia, a small deviation can be difficult to detect and easy to dismiss. The magnitude of the deviation can be measured by placing a prism in front of the misaligned eye to determine the power required to neutralize the fixation shift evoked by covering the other eye. Temporary press-on plastic Fresnel prisms, prism eyeglasses, or eye muscle surgery can be used to restore binocular alignment.

STEREOPSIS

Stereoacuity is determined by presenting targets with retinal disparity separately to each eye by using polarized images. The most popular office tests measure a range of thresholds from 800 to 40 s of arc. Normal stereoacuity is 40 s of arc. If a patient achieves this level of stereoacuity, one is assured that the eyes are aligned orthotropically and that vision is intact in each eye. Random dot stereograms have no monocular depth cues and provide an excellent screening test for strabismus.

COLOR VISION

The retina contains three classes of cones, with visual pigments of differing peak spectral sensitivity: red (560 nm), green (530 nm), and blue (430 nm). The red and green cone pigments are encoded on the X chromosome, and the blue cone pigment on chromosome 7. Mutations of the blue cone pigment are exceedingly rare. Mutations of the red and green pigments cause congenital X-linked color blindness in 8% of males. Affected individuals are not truly color blind; rather, they differ from normal subjects in the way they perceive color and how they combine primary monochromatic lights to match a particular color. Anomalous trichromats have three cone types, but a mutation in one cone pigment (usually red or green) causes a shift in peak spectral sensitivity, altering the proportion of primary colors required to achieve a color match. Dichromats have only two cone types and therefore will accept a color match based on only two

primary colors. Anomalous trichromats and dichromats have 6/6 (20/20) visual acuity, but their hue discrimination is impaired. Ishihara color plates can be used to detect red-green color blindness. The test plates contain a hidden number that is visible only to subjects with color confusion from red-green color blindness. Because color blindness is almost exclusively X-linked, it is worth screening only male children.

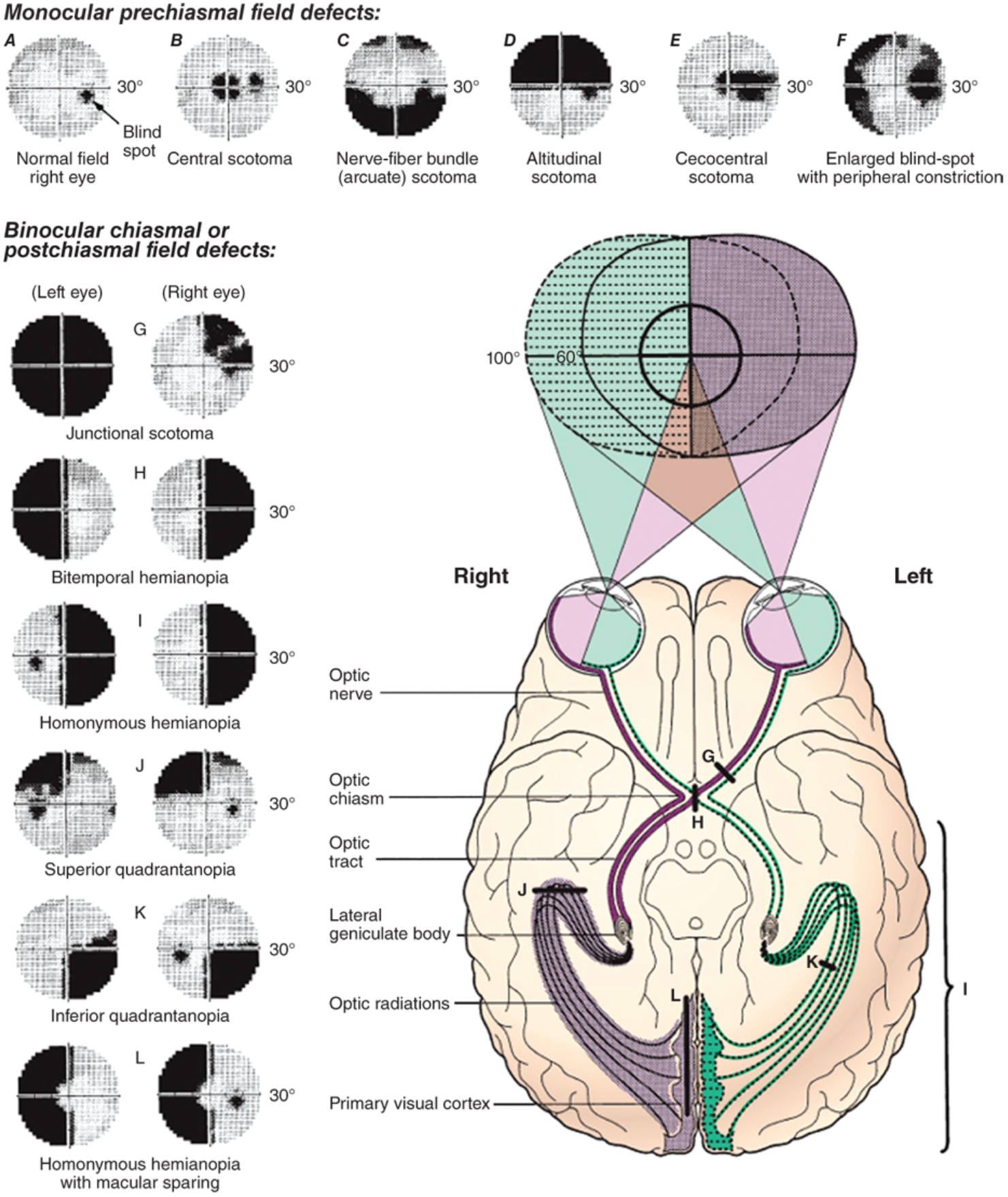
The Ishihara plates often are used to detect acquired defects in color vision, although they are intended as a screening test for congenital color blindness. Acquired defects in color vision frequently result from disease of the macula or optic nerve. For example, patients with a history of optic neuritis often complain of color desaturation long after their visual acuity has returned to normal. Color blindness also can result from bilateral strokes involving the ventral portion of the occipital lobe (cerebral achromatopsia). Such patients can perceive only shades of gray and also may have difficulty recognizing faces (prosopagnosia). Infarcts of the dominant occipital lobe sometimes give rise to color anomia. Affected patients can discriminate colors but cannot name them.

VISUAL FIELDS

Vision can be impaired by damage to the visual system anywhere from the eyes to the occipital lobes. One can localize the site of the lesion with considerable accuracy by mapping the visual field deficit by finger confrontation and then correlating it with the topographic anatomy of the visual pathway **(Fig. 28-3)**. Quantitative visual field mapping is performed by computer-driven perimeters that present a target of variable intensity at fixed positions in the visual field **(Fig. 28-3A)**. By generating an automated printout of light thresholds, these static perimeters provide a sensitive means of detecting scotomas in the visual field. They are exceedingly useful for serial assessment of visual function in chronic diseases such as glaucoma and pseudotumor cerebri.

FIGURE 28-3

Ventral view of the brain, correlating patterns of visual field loss with the sites of lesions in the visual pathway. The visual fields overlap partially, creating 120° of central binocular field flanked by a 40° monocular crescent on either side. The visual field maps in this figure were done with a computer-driven perimeter (Humphrey Instruments, Carl Zeiss, Inc.). It plots the retinal sensitivity to light in the central 30° by using a gray scale format. Areas of visual field loss are shown in black. The examples of common monocular, prechiasmal field defects are all shown for the right eye. By convention, the visual fields are always recorded with the left eye's field on the left and the right eye's field on the right, just as the patient sees the world.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The crux of visual field analysis is to decide whether a lesion is before, at, or behind the optic chiasm. If a scotoma is confined to one eye, it must be due to a lesion anterior to the chiasm, involving either the optic nerve or the retina. Retinal lesions produce scotomas that

correspond optically to their location in the fundus. For example, a superior-nasal retinal detachment results in an inferior-temporal field cut. Damage to the macula causes a central scotoma (Fig. 28-3B).

Optic nerve disease produces characteristic patterns of visual field loss. Glaucoma selectively destroys axons that enter the superotemporal or inferotemporal poles of the optic disc, resulting in arcuate scotomas shaped like a Turkish scimitar, which emanate from the blind spot and curve around fixation to end flat against the horizontal meridian (Fig. 28-3C). This type of field defect mirrors the arrangement of the nerve fiber layer in the temporal retina. Arcuate or nerve fiber layer scotomas also result from optic neuritis, ischemic optic neuropathy, optic disc drusen, and branch retinal artery or vein occlusion.

Damage to the entire upper or lower pole of the optic disc causes an altitudinal field cut that follows the horizontal meridian (Fig. 28-3D). This pattern of visual field loss is typical of ischemic optic neuropathy but also results from retinal vascular occlusion, advanced glaucoma, and optic neuritis.

About half the fibers in the optic nerve originate from ganglion cells serving the macula. Damage to papillomacular fibers causes a cecocentral scotoma that encompasses the blind spot and macula (Fig. 28-3E). If the damage is irreversible, pallor eventually appears in the temporal portion of the optic disc. Temporal pallor from a cecocentral scotoma may develop in optic neuritis, nutritional optic neuropathy, toxic optic neuropathy, Leber's hereditary optic neuropathy, Kjer's dominant optic atrophy, and compressive optic neuropathy. It is worth mentioning that the temporal side of the optic disc is slightly paler than the nasal side in most normal individuals. Therefore, it sometimes can be difficult to decide whether the temporal pallor visible on fundus examination represents a pathologic change. Pallor of the nasal rim of the optic disc is a less equivocal sign of optic atrophy.

At the optic chiasm, fibers from nasal ganglion cells decussate into the contralateral optic tract. Crossed fibers are damaged more by compression than are uncrossed fibers. As a result, mass lesions of the sellar region cause a temporal hemianopia in each eye. Tumors anterior to the optic chiasm, such as meningiomas of the tuberculum sella, produce a junctional scotoma characterized by an optic neuropathy in one eye and a superior-temporal field cut in the other eye (Fig. 28-3G). More symmetric compression of the optic chiasm by a pituitary adenoma (see Fig. 373-1), meningioma, craniopharyngioma, glioma, or aneurysm results in a bitemporal hemianopia (Fig. 28-3H). The insidious development of a bitemporal hemianopia often goes unnoticed by the patient and will escape detection by the physician unless each eye is tested separately.

It is difficult to localize a postchiasmal lesion accurately, because injury anywhere in the optic tract, lateral geniculate body, optic radiations, or visual cortex can produce a homonymous hemianopia (i.e., a temporal hemifield defect in the contralateral eye and a matching nasal hemifield defect in the ipsilateral eye) (Fig. 28-3I). A unilateral postchiasmal lesion leaves the visual acuity in each eye unaffected, although the patient may read the letters on only the left or right half of the eye chart. Lesions of the optic radiations tend to cause poorly matched or incongruous field defects in each eye. Damage to the optic radiations in the temporal lobe (Meyer's loop) produces a superior quadrantic homonymous hemianopia (Fig. 28-3J), whereas injury to the optic radiations in the parietal lobe results in an inferior quadrantic homonymous hemianopia (Fig. 28-3K). Lesions of the primary visual cortex give rise to dense, congruous hemianopic field defects. Occlusion of the posterior cerebral artery supplying the occipital lobe is a common cause of total homonymous hemianopia. Some patients with hemianopia after occipital stroke have macular sparing, because the macular representation at the tip of the occipital lobe is supplied by collaterals from the middle cerebral artery (Fig. 28-3L). Destruction of both occipital lobes produces cortical blindness. This condition can be distinguished from bilateral prechiasmal visual loss by noting that the pupil responses and optic fundi remain normal.

Partial recovery of homonymous hemianopia has been reported through computer-based rehabilitation therapy. During daily training sessions, patients fixate a central target while visual stimuli are presented within the blind region. The premise of vision restoration programs is that extra stimulation can promote recovery of partially damaged tissue located at the fringe of a cortical lesion. When fixation is controlled rigorously, however, no real improvement of the visual fields can be demonstrated. No effective treatment has been devised for homonymous hemianopia caused by loss of visual cortex.

DISORDERS

RED OR PAINFUL EYE

Corneal Abrasions

Corneal abrasions are seen best by placing a drop of fluorescein in the eye and looking with the slit lamp, using a cobalt-blue light. A penlight with a blue filter will suffice if a slit lamp is not available. Damage to the corneal epithelium is revealed by yellow fluorescence of the exposed basement membrane underlying the epithelium. It is important to check for foreign bodies. To search the conjunctival fornices, the lower lid should be pulled down and the upper lid everted. A foreign body can be removed with a moistened cotton-tipped applicator after a drop of a topical anesthetic such as proparacaine has been placed in the eye. Alternatively, it may be possible to flush the foreign body from the eye by irrigating copiously with saline or artificial tears. If the corneal epithelium has been abraded, antibiotic ointment and a patch should be applied to the eye. A drop of an intermediate-acting cycloplegic such as cyclopentolate hydrochloride 1% helps reduce pain by relaxing the ciliary body. The eye should be reexamined the next day. Minor abrasions may not require patching, antibiotics, or cycloplegia.

Subconjunctival Hemorrhage

This results from rupture of small vessels bridging the potential space between the episclera and the conjunctiva. Blood dissecting into this space can produce a spectacular red eye, but vision is not affected and the hemorrhage resolves without treatment. Subconjunctival hemorrhage is usually spontaneous but can result from blunt trauma, eye rubbing, or vigorous coughing. Occasionally it is a clue to an underlying bleeding disorder.

Pinguecula

Pinguecula is a small, raised conjunctival nodule, usually at the nasal limbus. In adults such lesions are extremely common and have little significance unless they become inflamed (pingueculitis). They are more apt to occur in workers with frequent outdoor exposure. A *pterygium* resembles a pinguecula but has crossed the limbus to encroach on the corneal surface. Removal is justified when symptoms of irritation or blurring develop, but recurrence is a common problem.

Blepharitis

This refers to inflammation of the eyelids. The most common form occurs in association with acne rosacea or seborrheic dermatitis. The eyelid margins usually are colonized heavily by staphylococci. Upon close inspection, they appear greasy, ulcerated, and crusted with scaling debris that clings to the lashes. Treatment consists of strict eyelid hygiene, using warm compresses and eyelash scrubs with baby shampoo. An external *hordeolum* (sty) is caused by staphylococcal infection of the superficial accessory glands of Zeis or Moll located in the eyelid margins. An internal hordeolum occurs after suppurative infection of the oil-secreting meibomian glands within the tarsal plate of the eyelid. Topical antibiotics such as bacitracin/polymyxin B ophthalmic ointment can be applied. Systemic antibiotics, usually tetracyclines or azithromycin, sometimes are necessary for treatment of meibomian gland inflammation (meibomitis) or chronic, severe blepharitis. A *chalazion* is a painless, chronic granulomatous inflammation of a meibomian gland that produces a pealike nodule within the eyelid. It can be incised and drained, but injection with glucocorticoids is equally effective. Basal cell, squamous cell, or meibomian gland carcinoma should be suspected with any nonhealing ulcerative lesion of the eyelids.

Dacryocystitis

An inflammation of the lacrimal drainage system, dacryocystitis can produce epiphora (tearing) and ocular injection. Gentle pressure over the lacrimal sac evokes pain and reflux of mucus or pus from the tear puncta. Dacryocystitis usually occurs after obstruction of the lacrimal system. It is treated with topical and systemic antibiotics, followed by probing, silicone stent intubation, or surgery to reestablish patency. *Entropion* (inversion of the eyelid) or *ectropion* (sagging or eversion of the eyelid) can also lead to epiphora and ocular irritation.

Conjunctivitis

Conjunctivitis is the most common cause of a red, irritated eye. Pain is minimal, and visual acuity is reduced only slightly. The most common viral etiology is adenovirus infection. It causes a watery discharge, a mild foreign-body sensation, and photophobia. Bacterial infection tends to produce a more mucopurulent exudate. Mild cases of infectious conjunctivitis usually are treated empirically with broad-spectrum topical ocular antibiotics such as sulfacetamide 10%, polymyxin-bacitracin, or a trimethoprim-polymyxin combination. Smears and cultures usually are reserved for severe, resistant, or recurrent cases of conjunctivitis. To prevent contagion, patients should be admonished to wash their hands frequently, not to touch their eyes, and to avoid direct contact with others.

Allergic Conjunctivitis

This condition is extremely common and often is mistaken for infectious conjunctivitis. Itching, redness, and epiphora are typical. The palpebral conjunctiva may become hypertropic with giant excrescences called cobblestone papillae. Irritation from contact lenses or any chronic foreign body also can induce formation of cobblestone papillae. *Atopic conjunctivitis* occurs in subjects with atopic dermatitis or asthma. Symptoms caused by allergic conjunctivitis can be alleviated with cold compresses, topical vasoconstrictors, antihistamines (olopatadine), and mast cell stabilizers (cromolyn). Topical glucocorticoid solutions provide dramatic relief of immune-mediated forms of conjunctivitis, but their long-term use is ill advised because of the complications of glaucoma, cataract, and secondary infection. Topical nonsteroidal anti-inflammatory drugs (ketorolac) are better alternatives.

Keratoconjunctivitis Sicca

Also known as dry eye, this produces a burning foreign-body sensation, injection, and photophobia. In mild cases the eye appears surprisingly normal, but tear production measured by wetting of a filter paper (Schirmer strip) is deficient. A variety of systemic drugs, including antihistaminic, anticholinergic, and psychotropic medications, result in dry eye by reducing lacrimal secretion. Disorders that involve the lacrimal gland directly, such as sarcoidosis and Sjögren's syndrome, also cause dry eye. Patients may develop dry eye after radiation therapy if the treatment field includes the orbits. Problems with ocular drying are also common after lesions affecting cranial nerve V or VII. Corneal anesthesia is particularly dangerous, because the absence of a normal blink reflex exposes the cornea to injury without pain to warn the patient. Dry eye is managed by frequent and liberal application of artificial tears and ocular lubricants. In severe cases the tear puncta can be plugged or cauterized to reduce lacrimal outflow.

Keratitis

Keratitis is a threat to vision because of the risk of corneal clouding, scarring, and perforation. Worldwide, the two leading causes of blindness from keratitis are trachoma from chlamydial infection and vitamin A deficiency related to malnutrition. In the United States, contact lenses play a major role in corneal infection and ulceration. They should not be worn by anyone with an active eye infection. In evaluating the cornea, it is important to differentiate between a superficial infection (*keratoconjunctivitis*) and a deeper, more serious ulcerative process. The latter is accompanied by greater visual loss, pain, photophobia, redness, and discharge. Slit-lamp examination shows disruption of the corneal epithelium, a cloudy infiltrate or abscess in the stroma, and an inflammatory cellular reaction in the anterior chamber. In severe cases, pus settles at the bottom of the anterior chamber, giving rise to a hypopyon. Immediate empirical antibiotic therapy should be initiated after corneal scrapings are obtained for Gram's stain, Giemsa stain, and cultures. Fortified topical antibiotics are most effective, supplemented with subconjunctival antibiotics as required. A fungal etiology should always be considered in a patient with keratitis. Fungal infection is common in warm humid climates, especially after penetration of the cornea by plant or vegetable material. Acanthamoeba keratitis is associated with improper disinfection of contact lenses.

Herpes Simplex

The *herpesviruses* are a major cause of blindness from keratitis. Most adults in the United States have serum antibodies to herpes simplex, indicating prior viral infection **(Chap. 187)**. Primary ocular infection generally is caused by herpes simplex type 1 rather than type 2. It manifests as a unilateral follicular blepharoconjunctivitis that is easily confused with adenoviral conjunctivitis, unless telltale vesicles are present on the eyelids or conjunctiva. A dendritic pattern of corneal epithelial ulceration revealed by fluorescein staining is pathognomonic for herpes infection but is seen in only a minority of primary infections. Recurrent ocular infection arises from reactivation of the latent herpesvirus. Viral eruption in the corneal epithelium may result in the characteristic herpes dendrite. Involvement of the corneal stroma produces edema, vascularization, and iridocyclitis. Herpes keratitis is treated with cycloplegia, and either a topical antiviral (trifluridine, ganciclovir) or an oral antiviral (acyclovir, ganciclovir) agent. Topical glucocorticoids are effective in mitigating corneal scarring but are generally reserved for cases involving stromal damage, because of the danger of corneal melting and perforation. Topical glucocorticoids also carry the risk of prolonging infection and inducing glaucoma.

Herpes Zoster

Herpes zoster from reactivation of latent varicella (chickenpox) virus causes a dermatomal pattern of painful vesicular dermatitis **(Chap. 188)**. Ocular symptoms can occur after zoster eruption in any branch of the trigeminal nerve but are particularly common when vesicles form on the nose, reflecting nasociliary (V1) nerve involvement (Hutchinson's sign). Herpes zoster ophthalmicus produces corneal dendrites, which can be difficult to distinguish from those seen in herpes simplex. Stromal keratitis, anterior uveitis, raised intraocular pressure, ocular motor nerve palsies, acute retinal necrosis, and postherpetic scarring and neuralgia are other common sequelae. Herpes

zoster ophthalmicus is treated with antiviral agents and cycloplegics. In severe cases, glucocorticoids may be added to prevent permanent visual loss from corneal scarring.

Episcleritis

This is an inflammation of the episclera, a thin layer of connective tissue between the conjunctiva and the sclera. Episcleritis resembles conjunctivitis, but it is a more localized process and discharge is absent. Most cases of episcleritis are idiopathic, but some occur in the setting of an autoimmune disease. *Scleritis* refers to a deeper, more severe inflammatory process that frequently is associated with a connective tissue disease such as rheumatoid arthritis, lupus erythematosus, polyarteritis nodosa, granulomatosis with polyangiitis, or relapsing polychondritis. The inflammation and thickening of the sclera can be diffuse or nodular. In anterior forms of scleritis, the globe assumes a violet hue and the patient complains of severe ocular tenderness and pain. With posterior scleritis, the pain and redness may be less marked, but there is often proptosis, choroidal effusion, reduced motility, and visual loss. Episcleritis and scleritis should be treated with NSAIDs. If these agents fail, topical or even systemic glucocorticoid therapy may be necessary, especially if an underlying autoimmune process is active.

Uveitis

Involving the anterior structures of the eye, uveitis also is called *iritis* or *iridocyclitis*. The diagnosis requires slit-lamp examination to identify inflammatory cells floating in the aqueous humor or deposited on the corneal endothelium (keratic precipitates). Anterior uveitis develops in sarcoidosis, ankylosing spondylitis, juvenile rheumatoid arthritis, inflammatory bowel disease, psoriasis, reactive arthritis, and Behçet's disease. It also is associated with herpes infections, syphilis, Lyme disease, onchocerciasis, tuberculosis, and leprosy. Although anterior uveitis can occur in conjunction with many diseases, no cause is found to explain the majority of cases. For this reason, laboratory evaluation usually is reserved for patients with recurrent or severe anterior uveitis. Treatment is aimed at reducing inflammation and scarring by judicious use of topical glucocorticoids. Dilatation of the pupil reduces pain and prevents the formation of synechiae.

Posterior Uveitis

This is diagnosed by observing inflammation of the vitreous, retina, or choroid on fundus examination. It is more likely than anterior uveitis to be associated with an identifiable systemic disease. Some patients have panuveitis, or inflammation of both the anterior and posterior segments of the eye. Posterior uveitis is a manifestation of autoimmune diseases such as sarcoidosis, Behçet's disease, Vogt-Koyanagi-Harada syndrome, and inflammatory bowel disease. It also accompanies diseases such as toxoplasmosis, onchocerciasis, cysticercosis, coccidioidomycosis, toxocariasis, and histoplasmosis; infections caused by organisms such as *Candida*, *Pneumocystis carinii*, *Cryptococcus*, *Aspergillus*, herpes, and cytomegalovirus **(see Fig. 190-1)**; and other diseases, such as syphilis, Lyme disease, tuberculosis, cat-scratch disease, Whipple's disease, and brucellosis. In multiple sclerosis, chronic inflammatory changes can develop in the extreme periphery of the retina (pars planitis or intermediate uveitis). Glucocorticoids have been the mainstay of treatment for noninfectious uveitis. Monoclonal antibodies which target proinflammatory cytokines, such as the tumor necrosis factor alpha (TNF-α) inhibitor adalimumab, are effective at preventing vision loss in chronic uveitis.

Acute Angle-Closure Glaucoma

This is an unusual but frequently misdiagnosed cause of a red, painful eye. Asian populations have a particularly high risk of angle-closure glaucoma. Susceptible eyes have a shallow anterior chamber because the eye has either a short axial length (hyperopia) or a lens enlarged by the gradual development of cataract. When the pupil becomes mid-dilated, the peripheral iris blocks aqueous outflow via the anterior chamber angle and the intraocular pressure rises abruptly, producing pain, injection, corneal edema, obscurations, and blurred vision. In some patients, ocular symptoms are overshadowed by nausea, vomiting, or headache, prompting a fruitless workup for abdominal or neurologic disease. The diagnosis is made by measuring the intraocular pressure during an acute attack or by performing gonioscopy, a procedure that allows one to observe a narrow chamber angle with a mirrored contact lens. Acute angle closure is treated with acetazolamide (PO or IV), topical beta blockers, prostaglandin analogues, α_2-adrenergic agonists, and pilocarpine to induce miosis. If these measures fail, a laser can be used to create a hole in the peripheral iris to relieve pupillary block. Many physicians are reluctant to dilate patients routinely for fundus examination because they fear precipitating an angle-closure glaucoma. The risk is actually remote and more than outweighed by the potential benefit to patients of discovering a hidden fundus lesion visible only through a fully dilated

pupil. Moreover, a single attack of angle closure after pharmacologic dilatation rarely causes any permanent damage to the eye and serves as an inadvertent provocative test to identify patients with narrow angles who would benefit from prophylactic laser iridectomy.

Endophthalmitis

This results from bacterial, viral, fungal, or parasitic infection of the internal structures of the eye. It usually is acquired by hematogenous seeding from a remote site. Chronically ill, diabetic, or immunosuppressed patients, especially those with a history of indwelling IV catheters or positive blood cultures, are at greatest risk for endogenous endophthalmitis. Although most patients have ocular pain and injection, visual loss is sometimes the only symptom. Septic emboli from a diseased heart valve or a dental abscess that lodge in the retinal circulation can give rise to endophthalmitis. White-centered retinal hemorrhages known as Roth's spots **(Fig. 28-4)** are considered pathognomonic for subacute bacterial endocarditis, but they also appear in leukemia, diabetes, and many other conditions. Endophthalmitis also occurs as a complication of ocular surgery, especially glaucoma filtering, occasionally months or even years after the operation. An occult penetrating foreign body or unrecognized trauma to the globe should be considered in any patient with unexplained intraocular infection or inflammation.

FIGURE 28-4

Roth's spot, cotton-wool spot, and retinal hemorrhages in a 48-year-old liver transplant patient with candidemia from immunosuppression.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TRANSIENT OR SUDDEN VISUAL LOSS

Amaurosis Fugax

This term refers to a transient ischemic attack of the retina **(Chap. 420)**. Because neural tissue has a high rate of metabolism, interruption of blood flow to the retina for more than a few seconds results in *transient monocular blindness*, a term used interchangeably with amaurosis fugax. Patients describe a rapid fading of vision like a curtain descending, sometimes affecting only a portion of the visual field. Amaurosis fugax usually results from an embolus that becomes stuck within a retinal arteriole **(Fig. 28-5)**. If the embolus breaks up or passes, flow is restored and vision returns quickly to normal without permanent damage. With prolonged interruption of blood flow, the inner retina suffers infarction. Ophthalmoscopy reveals zones of whitened, edematous retina following the distribution of branch retinal arterioles. Complete occlusion of the central retinal artery produces arrest of blood flow and a milky retina with a cherry-red fovea **(Fig. 28-6)**. Emboli are composed of cholesterol (Hollenhorst plaque), calcium, or platelet-fibrin debris. The most common source is an atherosclerotic plaque in the carotid artery or aorta, although emboli also can arise from the heart, especially in patients with diseased valves, atrial fibrillation, or wall motion abnormalities.

FIGURE 28-5

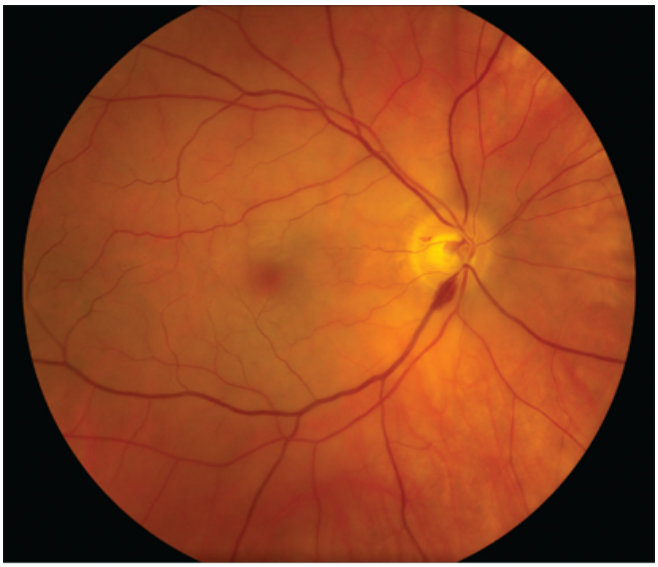
Hollenhorst plaque lodged at the bifurcation of a retinal arteriole proves that a patient is shedding emboli from the carotid artery, great vessels, or heart.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 28-6

Central retinal artery occlusion in a 78-year-old man reducing acuity to counting fingers in the right eye. Note the splinter hemorrhage on the optic disc and the slightly milky appearance to the macula with a cherry-red fovea.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

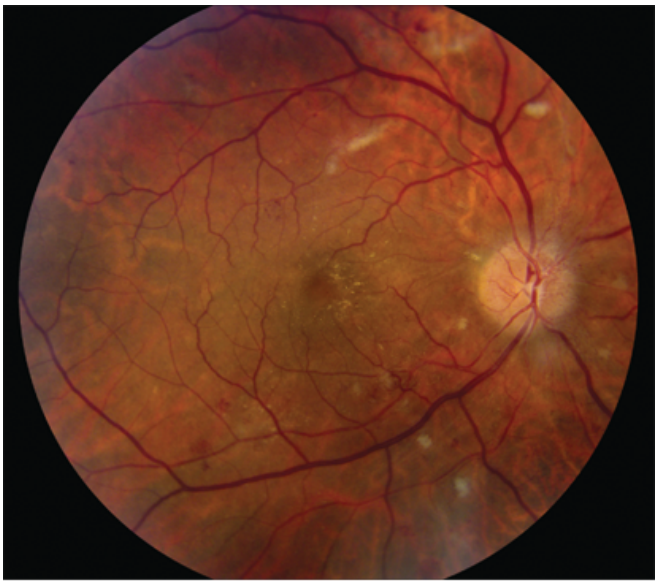
In rare instances, amaurosis fugax results from low central retinal artery perfusion pressure in a patient with a critical stenosis of the ipsilateral carotid artery and poor collateral flow via the circle of Willis. In this situation, amaurosis fugax develops when there is a dip in systemic blood pressure or a slight worsening of the carotid stenosis. Sometimes there is contralateral motor or sensory loss, indicating concomitant hemispheric cerebral ischemia.

Retinal arterial occlusion also occurs rarely in association with retinal migraine, lupus erythematosus, anticardiolipin antibodies, anticoagulant deficiency states (protein S, protein C, and antithrombin deficiency), Susac's syndrome, pregnancy, IV drug abuse, blood dyscrasias, dysproteinemias, and temporal arteritis.

Marked *systemic hypertension* causes sclerosis of retinal arterioles, splinter hemorrhages, focal infarcts of the nerve fiber layer (cotton-wool spots), and leakage of lipid and fluid (hard exudate) into the macula (Fig. 28-7). In hypertensive crisis, sudden visual loss can result from vasospasm of retinal arterioles and retinal ischemia. In addition, acute hypertension may produce visual loss from ischemic swelling of the optic disc. Patients with acute hypertensive retinopathy should be treated by lowering the blood pressure. However, the blood pressure should not be reduced precipitously, because there is a danger of optic disc infarction from sudden hypoperfusion.

FIGURE 28-7

Hypertensive retinopathy with blurred optic disc, scattered hemorrhages, cotton-wool spots (nerve fiber layer infarcts), and foveal exudate in a 62-year-old man with chronic renal failure and a systolic blood pressure of 220.

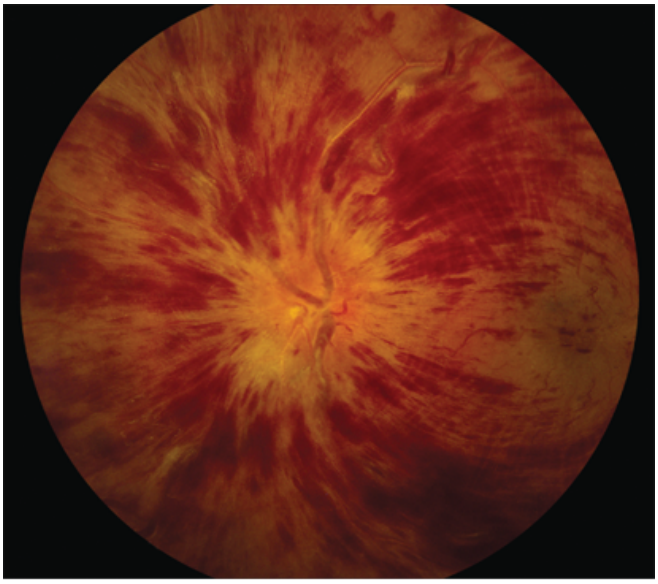


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Impending *branch* or *central retinal vein occlusion* can produce prolonged visual obscurations that resemble those described by patients with amaurosis fugax. The veins appear engorged and phlebitic, with numerous retinal hemorrhages (Fig. 28-8). In some patients, venous blood flow recovers spontaneously, whereas others evolve a frank obstruction with extensive retinal bleeding ("blood and thunder" appearance), infarction, and visual loss. Venous occlusion of the retina is often idiopathic, but hypertension, diabetes, and glaucoma are prominent risk factors. Polycythemia, thrombocythemia, or other factors leading to an underlying hypercoagulable state should be corrected; aspirin treatment may be beneficial.

FIGURE 28-8

Central retinal vein occlusion can produce massive retinal hemorrhage ("blood and thunder"), ischemia, and vision loss.



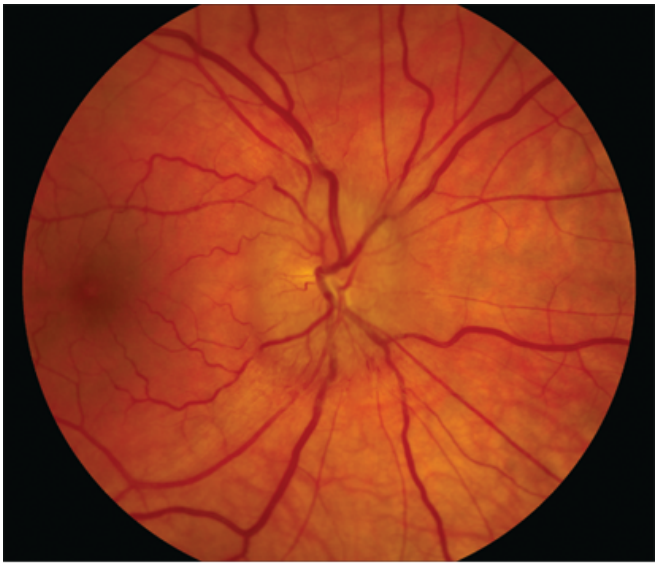
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Anterior Ischemic Optic Neuropathy (AION)

This is caused by insufficient blood flow through the posterior ciliary arteries that supply the optic disc. It produces painless monocular visual loss that is sudden in onset, followed sometimes by stuttering progression. The optic disc is edematous and usually bordered by nerve fiber layer splinter hemorrhages (Fig. 28-9). AION is divided into two forms: arteritic and nonarteritic. The nonarteritic form is most common. No specific cause is known, although diabetes, renal failure, and hypertension are common risk factors. Case reports have linked erectile dysfunction drugs to AION, but a causal association is doubtful. Evidence is strong that a crowded disc architecture and small optic cup predispose to the development of nonarteritic AION. In patients with a "disc-at-risk," the advent of AION in one eye increases the likelihood of the same event occurring in the other eye. No treatment is available for nonarteritic AION; glucocorticoids should not be prescribed.

FIGURE 28-9

Anterior ischemic optic neuropathy from temporal arteritis in a 64-year-old woman with acute disc swelling, splinter hemorrhages, visual loss, and an erythrocyte sedimentation rate of 60 mm/h.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

About 5% of patients, especially Caucasian females aged >60, develop the arteritic form of AION in conjunction with giant-cell (temporal) arteritis (Chap. 356). It is urgent to recognize arteritic AION so that high doses of glucocorticoids can be instituted immediately to prevent

blindness in the second eye. Tocilizumab is an effective alternative to glucocorticoids for sustained suppression of symptoms of giant cell arteritis. Symptoms of polymyalgia rheumatica may be present; the sedimentation rate and C-reactive protein level are usually elevated. In a patient with visual loss from suspected arteritic AION, temporal artery biopsy is mandatory to confirm the diagnosis. Administer glucocorticoids immediately, without waiting for the biopsy to be completed. The biopsy should be obtained as soon as practical, because prolonged glucocorticoid treatment can hide inflammatory changes. It is important to harvest an arterial segment at least 3 cm long and to examine a sufficient number of tissue sections. The histological features of granulomatous inflammation are often quite subtle in temporal artery specimens. If the biopsy is declared negative by an experienced pathologist, the diagnosis of arteritic AION is highly unlikely and glucocorticoids should usually be discontinued.

Posterior Ischemic Optic Neuropathy

This is an uncommon cause of acute visual loss, induced by the combination of severe anemia and hypotension. Cases have been reported after major blood loss during surgery (especially in patients undergoing cardiac or lumbar spine operations), shock, gastrointestinal bleeding, and renal dialysis. The fundus usually appears normal, although optic disc swelling develops if the process extends anteriorly far enough to reach the globe. Vision can be salvaged in some patients by immediate blood transfusion and reversal of hypotension.

Optic Neuritis

This is a common inflammatory disease of the optic nerve. In the Optic Neuritis Treatment Trial (ONTT), the mean age of patients was 32 years, 77% were female, 92% had ocular pain (especially with eye movements), and 35% had optic disc swelling. In most patients, the demyelinating event was retrobulbar and the ocular fundus appeared normal on initial examination (Fig. 28-10), although optic disc pallor slowly developed over subsequent months.

FIGURE 28-10

Retrobulbar optic neuritis is characterized by a normal fundus examination initially, hence the rubric “the doctor sees nothing, and the patient sees nothing.” Optic atrophy develops after severe or repeated attacks.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Virtually all patients experience a gradual recovery of vision after a single episode of optic neuritis, even without treatment. This rule is so reliable that failure of vision to improve after a first attack of optic neuritis casts doubt on the original diagnosis. Treatment with high-dose IV methylprednisolone (250 mg every 6 h for 3 days) followed by oral prednisone (1 mg/kg per day for 11 days) makes no difference in ultimate acuity 6 months after the attack, but the recovery of visual function occurs more rapidly. Therefore, when visual loss is severe (worse than 20/100), IV followed by PO glucocorticoids are often recommended.

For some patients, optic neuritis remains an isolated event. However, the ONTT showed that the 15-year cumulative probability of developing clinically definite multiple sclerosis after optic neuritis is 50%. A brain magnetic resonance (MR) scan is advisable in every

patient with a first attack of optic neuritis. If two or more plaques are present on initial imaging, treatment should be considered to prevent the development of additional demyelinating lesions (Chap. 436).

A particularly severe form of optic neuritis occurs in neuromyelitis optica (NMO); it is typically longitudinally extensive, and may be bilateral or associated with myelitis. NMO can occur as a primary disorder, in the setting of systemic autoimmune disease or rarely as a paraneoplastic condition. Detection of circulating antibodies directed against aquaporin-4 is diagnostic. Treatment for acute episodes consists of glucocorticoids and, in resistant cases, plasma exchange. **Neuromyelitis optica is discussed in detail in Chap. 437.**

LEBER'S HEREDITARY OPTIC NEUROPATHY

This disease usually affects young men, causing gradual, painless, severe central visual loss in one eye, followed weeks to years later by the same process in the other eye. Acutely, the optic disc appears mildly plethoric with surface capillary telangiectasias but no vascular leakage on fluorescein angiography. Eventually optic atrophy ensues. Leber's optic neuropathy is caused by a point mutation at codon 11778 in the mitochondrial gene encoding nicotinamide adenine dinucleotide dehydrogenase (NADH) subunit 4. Additional mutations responsible for the disease have been identified, most in mitochondrial genes that encode proteins involved in electron transport. Mitochondrial mutations that cause Leber's neuropathy are inherited from the mother by all her children, but for unknown reasons, daughters are rarely affected. Early stage clinical trials of gene therapy for this condition are in progress.

Toxic Optic Neuropathy

This can result in acute visual loss with bilateral optic disc swelling and cecocentral scotomas. Cases have been reported from exposure to ethambutol, methyl alcohol (moonshine), ethylene glycol (antifreeze), or carbon monoxide. In toxic optic neuropathy, visual loss also can develop gradually and produce optic atrophy (Fig. 28-11) without a phase of acute optic disc edema. Many agents have been implicated in toxic optic neuropathy, but evidence supporting the association is often weak. The following is a partial list of potential offending drugs or toxins: disulfiram, ethchlorvynol, chloramphenicol, amiodarone, monoclonal anti-CD3 antibody, ciprofloxacin, digitalis, streptomycin, lead, arsenic, thallium, D-penicillamine, isoniazid, emetine, and sulfonamides. Metallosis (chromium, cobalt, nickel) from hip implant failure is a rare cause of toxic optic neuropathy. Deficiency states induced by starvation, malabsorption, or alcoholism can lead to insidious visual loss. Thiamine, vitamin B_{12}, and folate levels should be checked in any patient with unexplained bilateral central scotomas and optic pallor.

FIGURE 28-11

Optic atrophy is not a specific diagnosis but refers to the combination of optic disc pallor, arteriolar narrowing, and nerve fiber layer destruction produced by a host of eye diseases, especially optic neuropathies.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Papilledema

This connotes bilateral optic disc swelling from raised intracranial pressure **(Fig. 28-12)**. Headache is a common but not invariable accompaniment. All other forms of optic disc swelling (e.g., from optic neuritis or ischemic optic neuropathy) should be called "optic disc edema." This convention is arbitrary but serves to avoid confusion. Often it is difficult to differentiate papilledema from other forms of optic disc edema by fundus examination alone. Transient visual obscurations are a classic symptom of papilledema. They can occur in only one eye or simultaneously in both eyes. They usually last seconds but can persist longer. Obscurations follow abrupt shifts in posture or happen spontaneously. When obscurations are prolonged or spontaneous, the papilledema is more threatening. Visual acuity is not affected by papilledema unless the papilledema is severe, long-standing, or accompanied by macular edema and hemorrhage. Visual field testing shows enlarged blind spots and peripheral constriction (Fig. 28-3F). With unremitting papilledema, peripheral visual field loss progresses in an insidious fashion while the optic nerve develops atrophy. In this setting, reduction of optic disc swelling is an ominous sign of a dying nerve rather than an encouraging indication of resolving papilledema.

FIGURE 28-12

Papilledema means optic disc edema from raised intracranial pressure. This young woman developed acute papilledema, with hemorrhages and cotton-wool spots, as a rare side effect of treatment with tetracycline for acne.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Evaluation of papilledema requires neuroimaging to exclude an intracranial lesion. MR angiography is appropriate in selected cases to search for a dural venous sinus occlusion or an arteriovenous shunt. If neuroradiologic studies are negative, the subarachnoid opening pressure should be measured in the lateral decubitus position by lumbar puncture. Inaccurate pressure readings are a common pitfall. An elevated pressure, with normal cerebrospinal fluid, points by exclusion to the diagnosis of *pseudotumor cerebri* (idiopathic intracranial hypertension). Almost all patients are female, and most are obese. Treatment with a carbonic anhydrase inhibitor such as acetazolamide lowers intracranial pressure by reducing the production of cerebrospinal fluid and improves the visual fields. Weight reduction is vital: bariatric surgery should be considered in patients who cannot lose weight by diet control. If vision loss is severe or progressive, a shunt should be performed without delay to prevent blindness. Optic nerve sheath fenestration is less efficacious, and does not address other neurological symptoms. Occasionally, fulminant papilledema produces rapid onset of blindness. In such patients, emergency surgery should be performed to install a shunt without delay.

Optic Disc Drusen

These are refractile deposits within the substance of the optic nerve head **(Fig. 28-13)**. They are unrelated to drusen of the retina, which occur in age-related macular degeneration. Optic disc drusen are most common in people of northern European descent. Their diagnosis is obvious when they are visible as glittering particles on the surface of the optic disc. However, in many patients they are hidden beneath the surface, producing pseudopapilledema. It is important to recognize optic disc drusen to avoid an unnecessary evaluation for papilledema. When optic disc drusen are buried, B-ultrasound is the most sensitive way to detect them. They appear hyperechoic because they contain calcium. They are also visible on computed tomography (CT) or optical coherence tomography (OCT), a technique for acquiring cross-section images of the retina. In most patients, optic disc drusen are an incidental, innocuous finding, but they can produce

visual obscurations. On perimetry they give rise to enlarged blind spots and arcuate scotomas from damage to the optic disc. With increasing age, drusen tend to become more exposed on the disc surface as optic atrophy develops. Hemorrhage, choroidal neovascular membrane, and AION are more likely to occur in patients with optic disc drusen. No treatment is available.

FIGURE 28-13

Optic disc drusen are calcified, mulberry-like deposits of unknown etiology within the optic disc, giving rise to "pseudopapilledema."



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Vitreous Degeneration

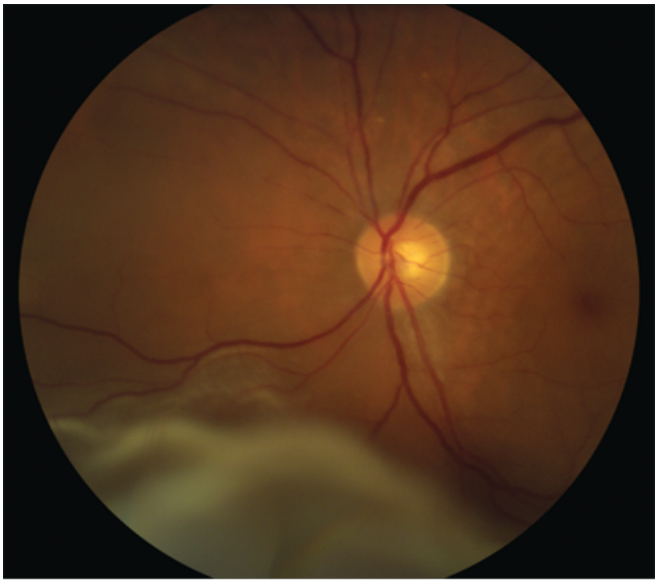
This occurs in all individuals with advancing age, leading to visual symptoms. Opacities develop in the vitreous, casting annoying shadows on the retina. As the eye moves, these distracting "floaters" move synchronously, with a slight lag caused by inertia of the vitreous gel. Vitreous traction on the retina causes mechanical stimulation, resulting in perception of flashing lights. This photopsia is brief and is confined to one eye, in contrast to the bilateral, prolonged scintillations of cortical migraine. Contraction of the vitreous can result in sudden separation from the retina, heralded by an alarming shower of floaters and photopsia. This process, known as *vitreous detachment*, is a common involutional event in the elderly. It is not harmful unless it damages the retina. A careful examination of the dilated fundus is important in any patient complaining of floaters or photopsia to search for peripheral tears or holes. If such a lesion is found, laser application can forestall a retinal detachment. Occasionally a tear ruptures a retinal blood vessel, causing vitreous hemorrhage and sudden loss of vision. On attempted ophthalmoscopy the fundus is hidden by a dark haze of blood. Ultrasound is required to examine the interior of the eye for a retinal tear or detachment. If the hemorrhage does not resolve spontaneously, the vitreous can be removed surgically. Vitreous hemorrhage also results from the fragile neovascular vessels that proliferate on the surface of the retina in diabetes, sickle cell anemia, and other ischemic ocular diseases.

Retinal Detachment

This produces symptoms of floaters, flashing lights, and a scotoma in the peripheral visual field corresponding to the detachment **(Fig. 28-14)**. If the detachment includes the fovea, there is an afferent pupil defect and the visual acuity is reduced. In most eyes, retinal detachment starts with a hole, flap, or tear in the peripheral retina (rhegmatogenous retinal detachment). Patients with peripheral retinal thinning (lattice degeneration) are particularly vulnerable to this process. Once a break has developed in the retina, liquefied vitreous is free to enter the subretinal space, separating the retina from the pigment epithelium. The combination of vitreous traction on the retinal surface and passage of fluid behind the retina leads inexorably to detachment. Patients with a history of myopia, trauma, or prior cataract extraction are at greatest risk for retinal detachment. The diagnosis is confirmed by ophthalmoscopic examination of the dilated eye.

FIGURE 28-14

Retinal detachment appears as an elevated sheet of retinal tissue with folds. In this patient, the fovea was spared, so acuity was normal, but an inferior detachment produced a superior scotoma.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Classic Migraine

(See also Chap. 422) This usually occurs with a visual aura lasting about 20 min. In a typical attack, a small central disturbance in the field of vision marches toward the periphery, leaving a transient scotoma in its wake. The expanding border of migraine scotoma has a scintillating, dancing, or zigzag edge, resembling the bastions of a fortified city, hence the term *fortification spectra*. Patients' descriptions of fortification spectra vary widely and can be confused with amaurosis fugax. Migraine patterns usually last longer and are perceived in both eyes, whereas amaurosis fugax is briefer and occurs in only one eye. Migraine phenomena also remain visible in the dark or with the eyes closed. Generally they are confined to either the right or the left visual hemifield, but sometimes both fields are involved simultaneously. Patients often have a long history of stereotypic attacks. After the visual symptoms recede, headache develops in most patients.

Transient Ischemic Attacks

Vertebrobasilar insufficiency may result in acute homonymous visual symptoms. Many patients mistakenly describe symptoms in the left or right eye when in fact the symptoms are occurring in the left or right hemifield of both eyes. Interruption of blood supply to the visual cortex causes a sudden fogging or graying of vision, occasionally with flashing lights or other positive phenomena that mimic migraine. Cortical ischemic attacks are briefer in duration than migraine, occur in older patients, and are not followed by headache. There may be associated signs of brainstem ischemia, such as diplopia, vertigo, numbness, weakness, and dysarthria.

Stroke

Stroke occurs when interruption of blood supply from the posterior cerebral artery to the visual cortex is prolonged. The only finding on examination is a homonymous visual field defect that stops abruptly at the vertical meridian. Occipital lobe stroke usually is due to thrombotic occlusion of the vertebrobasilar system, embolus, or dissection. Lobar hemorrhage, tumor, abscess, and arteriovenous malformation are other common causes of hemianopic cortical visual loss.

Factitious (Functional, Nonorganic) Visual Loss

This is claimed by hysterics or malingerers. The latter account for the vast majority, seeking sympathy, special treatment, or financial gain by feigning loss of sight. The diagnosis is suspected when the history is atypical, physical findings are lacking or contradictory, inconsistencies emerge on testing, and a secondary motive can be identified. In our litigious society, the fraudulent pursuit of recompense has spawned an epidemic of factitious visual loss.

CHRONIC VISUAL LOSS

Cataract

Cataract is a clouding of the lens sufficient to reduce vision. Most cataracts develop slowly as a result of aging, leading to gradual impairment of vision. The formation of cataract occurs more rapidly in patients with a history of uveitis, diabetes mellitus, ocular trauma or vitrectomy. Cataracts are acquired in a variety of genetic diseases, such as myotonic dystrophy, neurofibromatosis type 2, and galactosemia. Radiation therapy and glucocorticoid treatment can induce cataract as a side effect. The cataracts associated with radiation or glucocorticoids have a typical posterior subcapsular location. Cataract can be detected by noting an impaired red reflex when viewing light reflected from the fundus with an ophthalmoscope or by examining the dilated eye with the slit lamp.

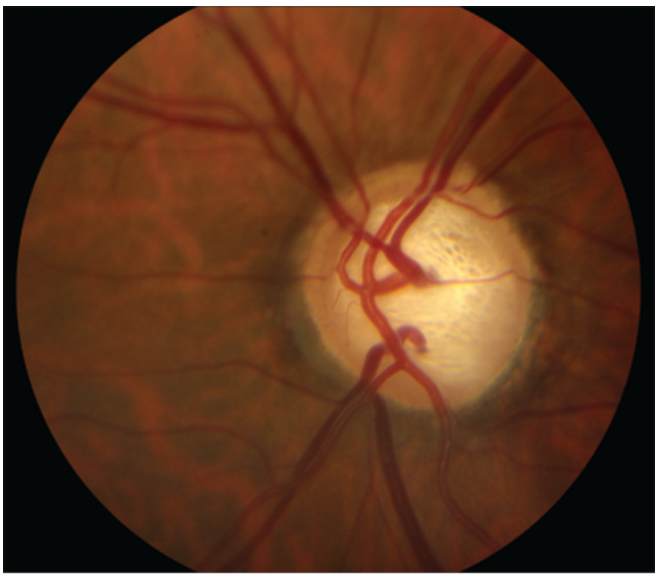
The only treatment for cataract is surgical extraction of the opacified lens. Millions of cataract operations are performed each year around the globe. The operation generally is done under local anesthesia on an outpatient basis. A plastic or silicone intraocular lens is placed within the empty lens capsule in the posterior chamber, substituting for the natural lens and leading to rapid recovery of sight. More than 95% of patients who undergo cataract extraction can expect an improvement in vision. In some patients, the lens capsule remaining in the eye after cataract extraction eventually turns cloudy, causing secondary loss of vision. A small opening, called a posterior capsulotomy, is made in the lens capsule with a laser to restore clarity.

Glaucoma

Glaucoma is a slowly progressive, insidious optic neuropathy that usually is associated with chronic elevation of intraocular pressure. After cataract, it is the most common cause of blindness in the world. It is especially prevalent in people of African descent. The mechanism by which raised intraocular pressure injures the optic nerve is not understood. Axons entering the inferotemporal and superotemporal aspects of the optic disc are damaged first, producing typical nerve fiber bundle or arcuate scotomas on perimetric testing. As fibers are destroyed, the neural rim of the optic disc shrinks and the physiologic cup within the optic disc enlarges **(Fig. 28-15)**. This process is referred to as pathologic "cupping." The cup-to-disc diameter is expressed as a fraction (e.g., 0.2). The cup-to-disc ratio ranges widely in normal individuals, making it difficult to diagnose glaucoma reliably simply by observing an unusually large or deep optic cup. Careful documentation of serial examinations is helpful. In a patient with physiologic cupping the large cup remains stable, whereas in a patient with glaucoma it expands relentlessly over the years. Observation of progressive cupping and detection of an arcuate scotoma or a nasal step on computerized visual field testing is sufficient to establish the diagnosis of glaucoma. OCT reveals corresponding loss of fibers along the arcuate pathways in the nerve fiber layer.

FIGURE 28-15

Glaucoma results in "cupping" as the neural rim is destroyed and the central cup becomes enlarged and excavated. The cup-to-disc ratio is about 0.8 in this patient.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The preponderance of patients with glaucoma have open anterior chamber angles. In most affected individuals the intraocular pressure is elevated. The cause of elevated intraocular pressure is unknown, but it is associated with gene mutations in the heritable forms.

Surprisingly, a third of patients with open-angle glaucoma have an intraocular pressure within the normal range of 10–20 mmHg. For this so-called normal or low-tension form of glaucoma, high myopia is a risk factor.

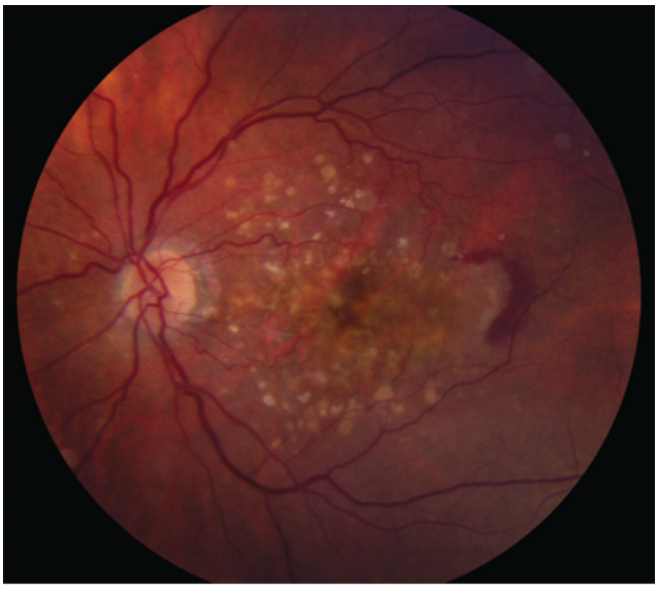
Chronic angle-closure glaucoma and chronic open-angle glaucoma are usually asymptomatic. Only acute angle-closure glaucoma causes a red or painful eye, from abrupt elevation of intraocular pressure. In all forms of glaucoma, foveal acuity is spared until end-stage disease is reached. For these reasons, severe and irreversible damage can occur before either the patient or the physician recognizes the diagnosis. Screening of patients for glaucoma by noting the cup-to-disc ratio on ophthalmoscopy and by measuring intraocular pressure is vital. Glaucoma is treated with topical adrenergic agonists, cholinergic agonists, beta blockers, prostaglandin analogues, and carbonic anhydrase inhibitors. Occasionally, systemic absorption of beta blocker from eyedrops can be sufficient to cause side effects of bradycardia, hypotension, heart block, bronchospasm, or depression. Laser treatment of the trabecular meshwork in the anterior chamber angle improves aqueous outflow from the eye. If medical or laser treatments fail to halt optic nerve damage from glaucoma, a filter must be constructed surgically (trabeculectomy) or a drainage device placed to release aqueous from the eye in a controlled fashion.

Macular Degeneration

This is a major cause of gradual, painless, bilateral central visual loss in the elderly. It occurs in a nonexudative (dry) form and an exudative (wet) form. Inflammation may be important in both forms of macular degeneration; susceptibility is associated with variants in the gene for complement factor H, an inhibitor of the alternative complement pathway. The nonexudative process begins with the accumulation of extracellular deposits called drusen underneath the retinal pigment epithelium. On ophthalmoscopy, they are pleomorphic but generally appear as small discrete yellow lesions clustered in the macula **(Fig. 28-16)**. With time they become larger, more numerous, and confluent. The retinal pigment epithelium becomes focally detached and atrophic, causing visual loss by interfering with photoreceptor function. Treatment with vitamins C and E, beta-carotene, and zinc may retard dry macular degeneration.

FIGURE 28-16

Age-related macular degeneration consisting of scattered yellow drusen in the macula (dry form) and a crescent of fresh hemorrhage temporal to the fovea from a subretinal neovascular membrane (wet form).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Exudative macular degeneration, which develops in only a minority of patients, occurs when neovascular vessels from the choroid grow through defects in Bruch's membrane and proliferate underneath the retinal pigment epithelium or the retina. Leakage from these vessels produces elevation of the retina, with distortion (metamorphopsia) and blurring of vision. Although the onset of these symptoms is usually gradual, bleeding from a subretinal choroidal neovascular membrane sometimes causes acute visual loss. Neovascular membranes can be difficult to see on fundus examination because they are located beneath the retina. Fluorescein angiography and OCT are extremely useful for their detection. Major or repeated hemorrhage under the retina from neovascular membranes results in fibrosis, development of a round (disciform) macular scar, and permanent loss of central vision.

A major therapeutic advance has occurred with the discovery that exudative macular degeneration can be treated with intraocular injection of antagonists to vascular endothelial growth factor. Bevacizumab, ranibizumab, or aflibercept is administered by direct injection into the vitreous cavity, beginning on a monthly basis. These antibodies cause the regression of neovascular membranes by blocking the action of vascular endothelial growth factor, thereby improving visual acuity.

Central Serous Chorioretinopathy

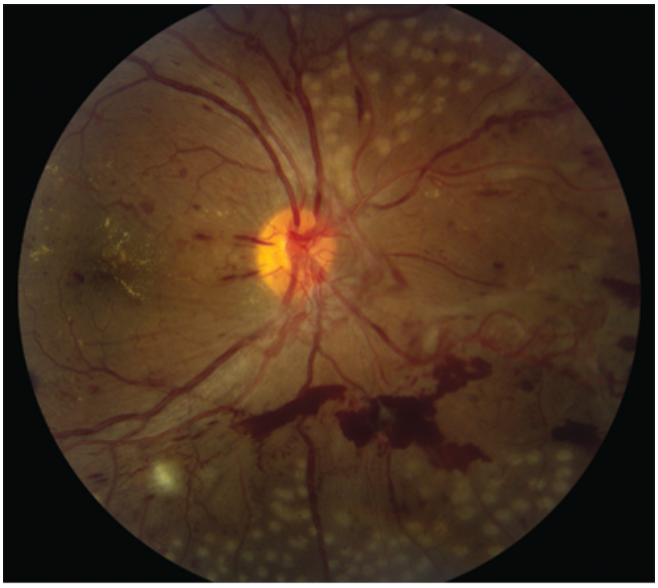
This primarily affects males between the ages of 20 and 50 years. Leakage of serous fluid from the choroid causes small, localized detachment of the retinal pigment epithelium and the neurosensory retina. These detachments produce acute or chronic symptoms of metamorphopsia and blurred vision when the macula is involved. They are difficult to visualize with a direct ophthalmoscope because the detached retina is transparent and only slightly elevated. OCT shows fluid beneath the retina, and fluorescein angiography shows dye streaming into the subretinal space. The cause of central serous chorioretinopathy is unknown. Symptoms may resolve spontaneously if the retina reattaches, but recurrent detachment is common. Laser photocoagulation has benefited some patients with this condition.

Diabetic Retinopathy

A rare disease until 1921, when the discovery of insulin resulted in a dramatic improvement in life expectancy for patients with diabetes mellitus, diabetic retinopathy is now a leading cause of blindness in the United States. The retinopathy takes years to develop but eventually appears in nearly all cases. Regular surveillance of the dilated fundus is crucial for any patient with diabetes. In advanced diabetic retinopathy, the proliferation of neovascular vessels leads to blindness from vitreous hemorrhage, retinal detachment, and glaucoma **(Fig. 28-17)**. These complications can be avoided in most patients by administration of panretinal laser photocoagulation at the appropriate point in the evolution of the disease. Anti-vascular endothelial growth factor antibody treatment is equally effective, but intraocular injections must be given repeatedly. **For further discussion of the manifestations and management of diabetic retinopathy, see Chaps. 396, 397, 398.**

FIGURE 28-17

Proliferative diabetic retinopathy in a 25-year-old man with an 18-year history of diabetes, showing neovascular vessels emanating from the optic disc, retinal and vitreous hemorrhage, cotton-wool spots, and macular exudate. Round spots in the periphery represent recently applied panretinal photocoagulation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

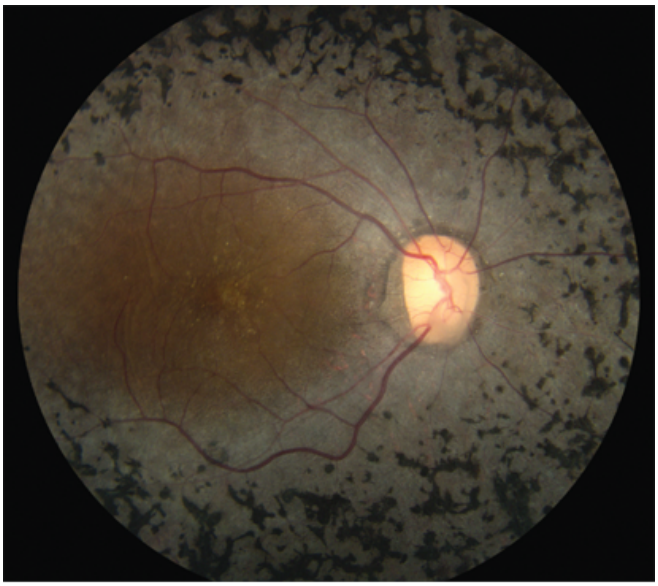
Retinitis Pigmentosa

This is a general term for a disparate group of rod-cone dystrophies characterized by progressive night blindness, visual field constriction with a ring scotoma, loss of acuity, and an abnormal electroretinogram (ERG). It occurs sporadically or in an autosomal recessive, dominant, or X-linked pattern. Irregular black deposits of clumped pigment in the peripheral retina, called *bone spicules* because of their

vague resemblance to the spicules of cancellous bone, give the disease its name (Fig. 28-18). The name is actually a misnomer because retinitis pigmentosa is not an inflammatory process. Most cases are due to a mutation in the gene for rhodopsin, the rod photopigment, or in the gene for peripherin, a glycoprotein located in photoreceptor outer segments. Vitamin A (15,000 IU/d) slightly retards the deterioration of the ERG in patients with retinitis pigmentosa but has no beneficial effect on visual acuity or fields.

FIGURE 28-18

Retinitis pigmentosa with black clumps of pigment known as "bone spicules." The patient had peripheral visual field loss with sparing of central (macular) vision.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Leber's congenital amaurosis, a rare cone dystrophy, has been treated by replacement of the missing RPE65 protein through gene therapy, resulting in slight improvement in visual function. Some forms of retinitis pigmentosa occur in association with rare, hereditary systemic diseases (olivopontocerebellar degeneration, Bassen-Kornzweig disease, Kearns-Sayre syndrome, Refsum's disease). Chronic treatment with chloroquine, hydroxychloroquine, and phenothiazines (especially thioridazine) can produce visual loss from a toxic retinopathy that resembles retinitis pigmentosa. Patients receiving long-term treatment with hydroxychloroquine require regular eye examinations to monitor for potential development of a bull's eye maculopathy.

Epiretinal Membrane

This is a fibrocellular tissue that grows across the inner surface of the retina, causing metamorphopsia and reduced visual acuity from distortion of the macula. A crinkled, cellophane-like membrane is visible on the retinal examination. Epiretinal membrane is most common in patients aged >50 years and is usually unilateral. Most cases are idiopathic, but some occur as a result of hypertensive retinopathy, diabetes, retinal detachment, or trauma. When visual acuity is reduced to the level of about 6/24 (20/80), vitrectomy and surgical peeling of the membrane to relieve macular puckering are recommended. Contraction of an epiretinal membrane sometimes gives rise to a *macular hole*. Most macular holes, however, are caused by local vitreous traction within the fovea. Vitrectomy can improve acuity in selected cases.

Melanoma and Other Tumors

Melanoma is the most common primary tumor of the eye (Fig. 28-19). It causes photopsia, an enlarging scotoma, and loss of vision. A small melanoma is often difficult to differentiate from a benign choroidal nevus. Serial examinations are required to document a malignant pattern of growth. Treatment of melanoma is controversial. Options include enucleation, local resection, and irradiation. *Metastatic tumors* to the eye outnumber primary tumors. Breast and lung carcinomas have a special propensity to spread to the choroid or iris. Leukemia and lymphoma also commonly invade ocular tissues. Sometimes their only sign on eye examination is cellular debris in the vitreous, which can masquerade as a chronic posterior uveitis.

FIGURE 28-19

Melanoma of the choroid, appearing as an elevated dark mass in the inferior fundus, with overlying hemorrhage. The black line denotes the plane of the optical coherence tomography scan (*below*) showing the subretinal tumor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In a patient with vision loss, CT or MR scanning should be considered if the cause remains unknown after careful review of the history, visual fields, and thorough examination of the eye. Optic nerve sheath meningioma is a common retrobulbar tumor. It produces the classic triad of opto-ciliary shunt vessels, optic atrophy, and progressive visual loss. Optic disc swelling and proptosis are also frequent signs. Optic nerve glioma in young patients is usually a pilocytic astrocytoma and has a good prognosis for preservation of vision, especially in neurofibromatosis type 1 (Chap. 118). In adults, optic nerve glioma is rare and highly malignant. Chiasmal tumors (pituitary adenoma, meningioma, craniopharyngioma) produce visual loss with few objective findings except for optic disc pallor. Loss of the temporal visual field in each eye is typically described, but in fact, patients complain of vision loss in just one eye. A high degree of vigilance is necessary to avoid missing chiasmal tumors. Although symptoms progress gradually, in rare instances the sudden expansion of a pituitary adenoma from infarction and bleeding (*pituitary apoplexy*) causes acute retrobulbar visual loss, with headache, nausea, and ocular motor nerve palsies.

PROPTOSIS

When the globes appear asymmetric, the clinician must first decide which eye is abnormal. Is one eye recessed within the orbit (*enophthalmos*), or is the other eye protuberant (*exophthalmos*, or *proptosis*)? A small globe or a Horner's syndrome can give the appearance of enophthalmos. True enophthalmos occurs commonly after trauma, from atrophy of retrobulbar fat, or from fracture of the orbital floor. The position of the eyes within the orbits is measured by using a Hertel exophthalmometer, a handheld instrument that records the position of the anterior corneal surface relative to the lateral orbital rim. If this instrument is not available, relative eye position can be judged by bending the patient's head forward and looking down upon the orbits. A proptosis of only 2 mm in one eye is detectable from this perspective. The development of proptosis implies a space-occupying lesion in the orbit and usually warrants CT or MR imaging.

Graves' Ophthalmopathy

This is the leading cause of proptosis in adults (Chap. 375). The proptosis is often asymmetric and can even appear to be unilateral. Orbital inflammation and engorgement of the extraocular muscles, particularly the medial rectus and the inferior rectus, account for the protrusion of the globe. Corneal exposure, lid retraction, lid lag on downgaze, conjunctival injection, restriction of gaze, diplopia, and visual loss from optic nerve compression are cardinal symptoms. Graves' eye disease is a clinical diagnosis, but laboratory testing can be useful. The serum level of thyroid-stimulating immunoglobulins is often elevated. Orbital imaging usually reveals enlarged extraocular eye

muscles, but not always. Graves' ophthalmopathy can be treated with oral prednisone (60 mg/d) for 1 month, followed by a taper over several months. Worsening of symptoms upon glucocorticoid withdrawal is common. Topical lubricants, taping the eyelids closed at night, moisture chambers, and eyelid surgery are helpful to limit exposure of ocular tissues. Radiation therapy is not effective. Orbital decompression should be performed for severe, symptomatic exophthalmos or if visual function is reduced by optic nerve compression. In patients with diplopia, prisms or eye muscle surgery can be used to restore ocular alignment in primary gaze.

Orbital Pseudotumor

This is an idiopathic, inflammatory orbital syndrome that is distinguished from Graves' ophthalmopathy by the prominent complaint of pain. Other symptoms include diplopia, ptosis, proptosis, and orbital congestion. Evaluation for sarcoidosis, granulomatosis with polyangiitis, and other types of orbital vasculitis or collagen-vascular disease is negative. Imaging often shows swollen eye muscles (orbital myositis) with enlarged tendons. By contrast, in Graves' ophthalmopathy, the tendons of the eye muscles usually are spared. The Tolosa-Hunt syndrome (Chap. 433) may be regarded as an extension of orbital pseudotumor through the superior orbital fissure into the cavernous sinus. The diagnosis of orbital pseudotumor is difficult. Biopsy of the orbit frequently yields nonspecific evidence of fat infiltration by lymphocytes, plasma cells, and eosinophils. A dramatic response to a therapeutic trial of systemic glucocorticoids indirectly provides the best confirmation of the diagnosis.

Orbital Cellulitis

This causes pain, lid erythema, proptosis, conjunctival chemosis, restricted motility, decreased acuity, afferent pupillary defect, fever, and leukocytosis. It often arises from the paranasal sinuses, especially by contiguous spread of infection from the ethmoid sinus through the lamina papyracea of the medial orbit. A history of recent upper respiratory tract infection, chronic sinusitis, thick mucus secretions, or dental disease is significant in any patient with suspected orbital cellulitis. Blood cultures should be obtained, but they are usually negative. Most patients respond to empirical therapy with broad-spectrum IV antibiotics. Occasionally, orbital cellulitis follows an overwhelming course, with massive proptosis, blindness, septic cavernous sinus thrombosis, and meningitis. To avert this disaster, orbital cellulitis should be managed aggressively in the early stages, with immediate imaging of the orbits and antibiotic therapy that includes coverage of methicillin-resistant *Staphylococcus aureus* (MRSA). Prompt surgical drainage of an orbital abscess or paranasal sinusitis is indicated if optic nerve function deteriorates despite antibiotics.

Tumors

Tumors of the orbit cause painless, progressive proptosis. The most common primary tumors are cavernous hemangioma, lymphangioma, neurofibroma, schwannoma, dermoid cyst, adenoid cystic carcinoma, optic nerve glioma, optic nerve meningioma, and benign mixed tumor of the lacrimal gland. Metastatic tumor to the orbit occurs frequently in breast carcinoma, lung carcinoma, and lymphoma. Diagnosis by fine-needle aspiration followed by urgent radiation therapy sometimes can preserve vision.

Carotid Cavernous Fistulas

With anterior drainage through the orbit, these fistulas produce proptosis, diplopia, glaucoma, and corkscrew, arterialized conjunctival vessels. Direct fistulas usually result from trauma. They are easily diagnosed because of the prominent signs produced by high-flow, high-pressure shunting. Indirect fistulas, or dural arteriovenous malformations, are more likely to occur spontaneously, especially in older women. The signs are more subtle, and the diagnosis frequently is missed. The combination of slight proptosis, diplopia, enlarged muscles, and an injected eye often is mistaken for thyroid ophthalmopathy. A bruit heard upon auscultation of the head or reported by the patient is a valuable diagnostic clue. Imaging shows an enlarged superior ophthalmic vein in the orbits. Carotid cavernous shunts can be eliminated by intravascular embolization.

PTOSIS

Blepharoptosis

This is an abnormal drooping of the eyelid. Unilateral or bilateral ptosis can be congenital, from dysgenesis of the levator palpebrae superioris, or from abnormal insertion of its aponeurosis into the eyelid. Acquired ptosis can develop so gradually that the patient is unaware of the problem. Inspection of old photographs is helpful in dating the onset. A history of prior trauma, eye surgery, contact lens use, diplopia, systemic symptoms (e.g., dysphagia or peripheral muscle weakness), or a family history of ptosis should be sought.

Fluctuating ptosis that worsens late in the day is typical of myasthenia gravis. Ptosis evaluation should focus on evidence for proptosis, eyelid masses or deformities, inflammation, pupil inequality, or limitation of motility. The width of the palpebral fissures is measured in primary gaze to determine the degree of ptosis. The ptosis will be underestimated if the patient compensates by lifting the brow with the frontalis muscle.

Mechanical Ptosis

This occurs in many elderly patients from stretching and redundancy of eyelid skin and subcutaneous fat (dermatochalasis). The extra weight of these sagging tissues causes the lid to droop. Enlargement or deformation of the eyelid from infection, tumor, trauma, or inflammation also results in ptosis on a purely mechanical basis.

Aponeurotic Ptosis

This is an acquired dehiscence or stretching of the aponeurotic tendon, which connects the levator muscle to the tarsal plate of the eyelid. It occurs commonly in older patients, presumably from loss of connective tissue elasticity. Aponeurotic ptosis is also a common sequela of eyelid swelling from infection or blunt trauma to the orbit, cataract surgery, or contact lens use.

Myogenic Ptosis

The causes of *myogenic ptosis* include myasthenia gravis **(Chap. 440)** and a number of rare myopathies that manifest with ptosis. The term *chronic progressive external ophthalmoplegia* refers to a spectrum of systemic diseases caused by mutations of mitochondrial DNA. As the name implies, the most prominent findings are symmetric, slowly progressive ptosis and limitation of eye movements. In general, diplopia is a late symptom because all eye movements are reduced equally. In the *Kearns-Sayre* variant, retinal pigmentary changes and abnormalities of cardiac conduction develop. Peripheral muscle biopsy shows characteristic “ragged-red fibers.” *Oculopharyngeal dystrophy* is a distinct autosomal dominant disease with onset in middle age, characterized by ptosis, limited eye movements, and trouble swallowing. *Myotonic dystrophy*, another autosomal dominant disorder, causes ptosis, ophthalmoparesis, cataract, and pigmentary retinopathy. Patients have muscle wasting, myotonia, frontal balding, and cardiac abnormalities.

Neurogenic Ptosis

This results from a lesion affecting the innervation to either of the two muscles that open the eyelid: Müller’s muscle or the levator palpebrae superioris. Examination of the pupil helps distinguish between these two possibilities. In Horner’s syndrome, the eye with ptosis has a smaller pupil and the eye movements are full. In an oculomotor nerve palsy, the eye with the ptosis has a larger or a normal pupil. If the pupil is normal but there is limitation of adduction, elevation, and depression, a pupil-sparing oculomotor nerve palsy is likely (see next section). Rarely, a lesion affecting the small, central subnucleus of the oculomotor complex will cause bilateral ptosis with normal eye movements and pupils.

DOUBLE VISION (DIPLOPIA)

The first point to clarify is whether diplopia persists in either eye after the opposite eye is covered. If it does, the diagnosis is monocular diplopia. The cause is usually intrinsic to the eye and therefore has no dire implications for the patient. Corneal aberrations (e.g., keratoconus, pterygium), uncorrected refractive error, cataract, or foveal traction may give rise to monocular diplopia. Occasionally it is a symptom of malingering or psychiatric disease. Diplopia alleviated by covering one eye is binocular diplopia and is caused by disruption of ocular alignment. Inquiry should be made into the nature of the double vision (purely side-by-side versus partial vertical displacement of images), mode of onset, duration, intermittency, diurnal variation, and associated neurologic or systemic symptoms. If the patient has diplopia while being examined, motility testing should reveal a deficiency corresponding to the patient’s symptoms. However, subtle limitation of ocular excursions is often difficult to detect. For example, a patient with a slight left abducens nerve paresis may appear to have full eye movements despite a complaint of horizontal diplopia upon looking to the left. In this situation, the cover test provides a more sensitive method for demonstrating the ocular misalignment. It should be conducted in primary gaze and then with the head turned and tilted in each direction. In the above example, a cover test with the head turned to the right will maximize the fixation shift evoked by the cover test.

Occasionally, a cover test performed in an asymptomatic patient during a routine examination will reveal an ocular deviation. If the eye movements are full and the ocular misalignment is equal in all directions of gaze (comitant deviation), the diagnosis is strabismus. In this

condition, which affects about 1% of the population, fusion is disrupted in infancy or early childhood. To avoid diplopia, retinal input from the nonfixating eye may be partially suppressed. In some children, this leads to impaired vision (amblyopia, or “lazy” eye) in the deviated eye.

Binocular diplopia results from a wide range of processes: infectious, neoplastic, metabolic, degenerative, inflammatory, and vascular. One must decide whether the diplopia is neurogenic in origin or is due to restriction of globe rotation by local disease in the orbit. Orbital pseudotumor, myositis, infection, tumor, thyroid disease, and muscle entrapment (e.g., from a blowout fracture) cause restrictive diplopia. The diagnosis of restriction is usually made by recognizing other associated signs and symptoms of local orbital disease. Dedicated, high-resolution orbital imaging is helpful when the cause of diplopia is not evident.

Myasthenia Gravis

(See also Chap. 440) This is a major cause of painless diplopia. The diplopia is often intermittent, variable, and not confined to any single ocular motor nerve distribution. The pupils are always normal. Serial measurements of a variable, fatigable ptosis, often accompanied by diplopia, are helpful to establish the diagnosis. Many patients have a purely ocular form of the disease, with no evidence of systemic muscular weakness. The diagnosis can be confirmed by an IV edrophonium injection, which produces a transient reversal of eyelid or eye muscle weakness. Blood tests for antibodies against the acetylcholine receptor or the MuSK protein are frequently negative in the purely ocular form of myasthenia gravis. *Botulism* from food or wound poisoning can mimic ocular myasthenia.

If restrictive orbital disease and myasthenia gravis are excluded, a lesion of a cranial nerve supplying innervation to the extraocular muscles is the most likely cause of binocular diplopia.

Oculomotor Nerve

The third cranial nerve innervates the medial, inferior, and superior recti; inferior oblique; levator palpebrae superioris; and the iris sphincter. Total palsy of the oculomotor nerve causes ptosis, a dilated pupil, and leaves the eye “down and out” because of the unopposed action of the lateral rectus and superior oblique. This combination of findings is obvious. More challenging is the diagnosis of early or partial oculomotor nerve palsy. In this setting any combination of ptosis, pupil dilation, and weakness of the eye muscles supplied by the oculomotor nerve may be encountered. Frequent serial examinations during the rapidly evolving phase of the palsy help ensure that the diagnosis is not missed. The advent of an oculomotor nerve palsy with a pupil involvement, especially when accompanied by pain, suggests a compressive lesion, such as a tumor or circle of Willis aneurysm. Urgent neuroimaging should be obtained, along with a CT or MR angiogram. With improvement in the resolution of these non-invasive techniques, catheter angiography is rarely necessary to exclude an aneurysm.

A lesion of the oculomotor nucleus in the rostral midbrain produces signs that differ from those caused by a lesion of the nerve itself. There is bilateral ptosis because the levator muscle is innervated by a single central subnucleus. There is also weakness of the contralateral superior rectus, because it is supplied by the oculomotor nucleus on the other side. Occasionally both superior recti are weak. Isolated nuclear oculomotor palsy is rare. Usually neurologic examination reveals additional signs that suggest brainstem damage from infarction, hemorrhage, tumor, or infection.

Injury to structures surrounding fascicles of the oculomotor nerve descending through the midbrain has given rise to a number of classic eponymic designations. In *Nothnagel’s syndrome*, injury to the superior cerebellar peduncle causes ipsilateral oculomotor palsy and contralateral cerebellar ataxia. In *Benedikt’s syndrome*, injury to the red nucleus results in ipsilateral oculomotor palsy and contralateral tremor, chorea, and athetosis. *Claude’s syndrome* incorporates features of both of these syndromes, by injury to both the red nucleus and the superior cerebellar peduncle. Finally, in *Weber’s syndrome*, injury to the cerebral peduncle causes ipsilateral oculomotor palsy with contralateral hemiparesis.

In the subarachnoid space the oculomotor nerve is vulnerable to aneurysm, meningitis, tumor, infarction, and compression. In cerebral herniation, the nerve becomes trapped between the edge of the tentorium and the uncus of the temporal lobe. Oculomotor palsy also can result from midbrain torsion and hemorrhage during herniation. In the cavernous sinus, oculomotor palsy arises from carotid aneurysm, carotid cavernous fistula, cavernous sinus thrombosis, tumor (pituitary adenoma, meningioma, metastasis), herpes zoster infection, and the Tolosa-Hunt syndrome.

The etiology of an isolated, pupil-sparing oculomotor palsy often remains an enigma even after neuroimaging and extensive laboratory testing. Most cases are thought to result from microvascular infarction of the nerve somewhere along its course from the brainstem to the orbit. Usually the patient complains of pain. Diabetes, hypertension, and vascular disease are major risk factors. Spontaneous recovery over a period of months is the rule. If this fails to occur or if new findings develop, the diagnosis of microvascular oculomotor nerve palsy should be reconsidered. Aberrant regeneration is common when the oculomotor nerve is injured by trauma or compression (tumor, aneurysm). Miswiring of sprouting fibers to the levator muscle and the rectus muscles results in elevation of the eyelid upon downgaze or adduction. The pupil also constricts upon attempted adduction, elevation, or depression of the globe. Aberrant regeneration is not seen after oculomotor palsy from microvascular infarct and hence vitiates that diagnosis.

Trochlear Nerve

The fourth cranial nerve originates in the midbrain, just caudal to the oculomotor nerve complex. Fibers exit the brainstem dorsally and cross to innervate the contralateral superior oblique. The principal actions of this muscle are to depress and intort the globe. A palsy therefore results in hypertropia and excyclotorsion. The cyclotorsion seldom is noticed by patients. Instead, they complain of vertical diplopia, especially upon reading or looking down. The vertical diplopia also is exacerbated by tilting the head toward the side with the muscle palsy and alleviated by tilting it away. This “head tilt test” is a cardinal diagnostic feature.

Isolated trochlear nerve palsy results from all the causes listed above for the oculomotor nerve except aneurysm. The trochlear nerve is particularly apt to suffer injury after closed head trauma. The free edge of the tentorium is thought to impinge on the nerve during a concussive blow. Most isolated trochlear nerve palsies are idiopathic and hence are diagnosed by exclusion as “microvascular.” Spontaneous improvement occurs over a period of months in most patients. A base-down prism (conveniently applied to the patient’s glasses as a stick-on Fresnel lens) may serve as a temporary measure to alleviate diplopia. If the palsy does not resolve, the eyes can be realigned by weakening the inferior oblique muscle.

Abducens Nerve

The sixth cranial nerve innervates the lateral rectus muscle. A palsy produces horizontal diplopia, worse on gaze to the side of the lesion. A nuclear lesion has different consequences, because the abducens nucleus contains interneurons that project via the medial longitudinal fasciculus to the medial rectus subnucleus of the contralateral oculomotor complex. Therefore, an abducens nuclear lesion produces a complete lateral gaze palsy from weakness of both the ipsilateral lateral rectus and the contralateral medial rectus. *Foville’s syndrome* after dorsal pontine injury includes lateral gaze palsy, ipsilateral facial palsy, and contralateral hemiparesis incurred by damage to descending corticospinal fibers. *Millard-Gubler syndrome* from ventral pontine injury is similar except for the eye findings. There is lateral rectus weakness only, instead of gaze palsy, because the abducens fascicle is injured rather than the nucleus. Infarct, tumor, hemorrhage, vascular malformation, and multiple sclerosis are the most common etiologies of brainstem abducens palsy.

After leaving the ventral pons, the abducens nerve runs forward along the clivus to pierce the dura at the petrous apex, where it enters the cavernous sinus. Along its subarachnoid course it is susceptible to meningitis, tumor (meningioma, chordoma, carcinomatous meningitis), subarachnoid hemorrhage, trauma, and compression by aneurysm or dolichoectatic vessels. At the petrous apex, mastoiditis can produce deafness, pain, and ipsilateral abducens palsy (*Gradenigo’s syndrome*). In the cavernous sinus, the nerve can be affected by carotid aneurysm, carotid cavernous fistula, tumor (pituitary adenoma, meningioma, nasopharyngeal carcinoma), herpes infection, and Tolosa-Hunt syndrome.

Unilateral or bilateral abducens palsy is a classic sign of raised intracranial pressure. The diagnosis can be confirmed if papilledema is observed on fundus examination. The mechanism is still debated but probably is related to rostral-caudal displacement of the brainstem. The same phenomenon accounts for abducens palsy from Chiari malformation or low intracranial pressure (e.g., after lumbar puncture, spinal anesthesia, or spontaneous dural cerebrospinal fluid leak).

Treatment of abducens palsy is aimed at prompt correction of the underlying cause. However, the cause remains obscure in many instances despite diligent evaluation. As was mentioned above for isolated trochlear or oculomotor palsy, most cases are assumed to represent microvascular infarcts because they often occur in the setting of diabetes or other vascular risk factors. Some cases may develop as a postinfectious mononeuritis (e.g., after a viral flu). Patching one eye, occluding one eyeglass lens with tape, or applying a temporary prism will provide relief of diplopia until the palsy resolves. If recovery is incomplete, eye muscle surgery nearly always can realign the eyes, at least in primary position. A patient with an abducens palsy that fails to improve should be reevaluated for an occult etiology (e.g.,

chordoma, carcinomatous meningitis, carotid cavernous fistula, myasthenia gravis). Skull base tumors are easily missed even on contrast-enhanced neuroimaging studies.

Multiple Ocular Motor Nerve Palsies

These should not be attributed to spontaneous microvascular events affecting more than one cranial nerve at a time. This remarkable coincidence does occur, especially in diabetic patients, but the diagnosis is made only in retrospect after all other diagnostic alternatives have been exhausted. Neuroimaging should focus on the cavernous sinus, superior orbital fissure, and orbital apex, where all three ocular motor nerves are in close proximity. In a diabetic or immunocompromised host, fungal infection (*Aspergillus*, Mucorales, *Cryptococcus*) is a common cause of multiple nerve palsies. In a patient with systemic malignancy, carcinomatous meningitis is a likely diagnosis. Cytologic examination may be negative despite repeated sampling of the cerebrospinal fluid. The cancer-associated Lambert-Eaton myasthenic syndrome also can produce ophthalmoplegia. Giant cell (temporal) arteritis occasionally manifests as diplopia from ischemic palsies of extraocular muscles. Fisher's syndrome, an ocular variant of Guillain-Barré, produces ophthalmoplegia with areflexia and ataxia. Often the ataxia is mild, and the reflexes are normal. Antiganglioside antibodies (GQ1b) can be detected in about 50% of cases.

Supranuclear Disorders of Gaze

These are often mistaken for multiple ocular motor nerve palsies. For example, Wernicke's encephalopathy can produce nystagmus and a partial deficit of horizontal and vertical gaze that mimics a combined abducens and oculomotor nerve palsy. The disorder occurs in patients who are malnourished, alcoholic, or following bariatric surgery, and can be reversed by thiamine. Infarct, hemorrhage, tumor, multiple sclerosis, encephalitis, vasculitis, and Whipple's disease are other important causes of supranuclear gaze palsy. Disorders of vertical gaze, especially downward saccades, are an early feature of progressive supranuclear palsy. Smooth pursuit is affected later in the course of the disease. Parkinson's disease, Huntington's disease, and olivopontocerebellar degeneration also can affect vertical gaze.

The *frontal eye field* of the cerebral cortex is involved in generation of saccades to the contralateral side. After hemispheric stroke, the eyes usually deviate toward the lesioned side because of the unopposed action of the frontal eye field in the normal hemisphere. With time, this deficit resolves. Seizures generally have the opposite effect: the eyes deviate conjugately away from the irritative focus. *Parietal lesions* disrupt smooth pursuit of targets moving toward the side of the lesion. Bilateral parietal lesions produce *Bálint's syndrome*, which is characterized by impaired eye-hand coordination (optic ataxia), difficulty initiating voluntary eye movements (ocular apraxia), and visuospatial disorientation (simultanagnosia).

Horizontal Gaze

Descending cortical inputs mediating horizontal gaze ultimately converge at the level of the pons. Neurons in the paramedian pontine reticular formation are responsible for controlling conjugate gaze toward the same side. They project directly to the ipsilateral abducens nucleus. A lesion of either the paramedian pontine reticular formation or the abducens nucleus causes an ipsilateral conjugate gaze palsy. Lesions at either locus produce nearly identical clinical syndromes, with the following exception: vestibular stimulation (oculocephalic maneuver or caloric irrigation) will succeed in driving the eyes conjugately to the side in a patient with a lesion of the paramedian pontine reticular formation but not in a patient with a lesion of the abducens nucleus.

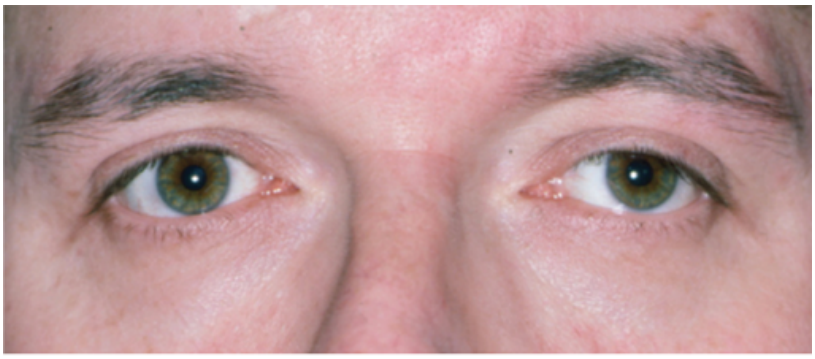
INTERNUCLEAR OPHTHALMOPLEGIA

This results from damage to the medial longitudinal fasciculus ascending from the abducens nucleus in the pons to the oculomotor nucleus in the midbrain (hence, "internuclear"). Damage to fibers carrying the conjugate signal from abducens interneurons to the contralateral medial rectus motoneurons results in a failure of adduction on attempted lateral gaze. For example, a patient with a left internuclear ophthalmoplegia (INO) will have slowed or absent adducting movements of the left eye **(Fig. 28-20)**. A patient with bilateral injury to the medial longitudinal fasciculus will have bilateral INO. Multiple sclerosis is the most common cause, although tumor, stroke, trauma, or any brainstem process may be responsible. *One-and-a-half syndrome* is due to a lesion of the medial longitudinal fasciculus combined with a lesion of either the abducens nucleus or the paramedian pontine reticular formation on the same side. The patient's only horizontal eye movement is abduction of the eye on the other side.

FIGURE 28-20

Left internuclear ophthalmoplegia (INO). ***A.*** In primary position of gaze, the eyes appear normal. ***B.*** Horizontal gaze to the left is intact. ***C.*** On attempted horizontal gaze to the right, the left eye fails to adduct. In mildly affected patients, the eye may adduct partially or more

slowly than normal. Nystagmus is usually present in the abducted eye. ***D.*** T2-weighted axial magnetic resonance image through the pons showing a demyelinating plaque in the left medial longitudinal fasciculus (*arrow*).



A

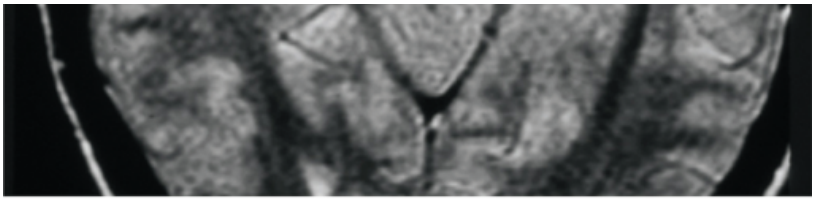


B



C





D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Vertical Gaze

This is controlled at the level of the midbrain. The neuronal circuits affected in disorders of vertical gaze are not fully elucidated, but lesions of the rostral interstitial nucleus of the medial longitudinal fasciculus and the interstitial nucleus of Cajal cause supranuclear paresis of upgaze, downgaze, or all vertical eye movements. Distal basilar artery ischemia is the most common etiology. *Skew deviation* refers to a vertical misalignment of the eyes, usually constant in all positions of gaze. The finding has poor localizing value because skew deviation has been reported after lesions in widespread regions of the brainstem and cerebellum.

PARINAUD'S SYNDROME

Also known as dorsal midbrain syndrome, this is a distinct supranuclear vertical gaze disorder caused by damage to the posterior commissure. It is a classic sign of hydrocephalus from aqueductal stenosis. Pineal region or midbrain tumors, cysticercosis, and stroke also cause Parinaud's syndrome. Features include loss of upgaze (and sometimes downgaze), convergence-retraction nystagmus on attempted upgaze, downward ocular deviation ("setting sun" sign), lid retraction (Collier's sign), skew deviation, pseudoabducens palsy, and light-near dissociation of the pupils.

Nystagmus

This is a rhythmic oscillation of the eyes, occurring physiologically from vestibular and optokinetic stimulation or pathologically in a wide variety of diseases (Chap. 19). Abnormalities of the eyes or optic nerves, present at birth or acquired in childhood, can produce a complex, searching nystagmus with irregular pendular (sinusoidal) and jerk features. Examples are albinism, Leber's congenital amaurosis, and bilateral cataract. This nystagmus is commonly referred to as *congenital sensory nystagmus*. This is a poor term because even in children with congenital lesions, the nystagmus does not appear until weeks after birth. *Congenital motor nystagmus*, which looks similar to congenital sensory nystagmus, develops in the absence of any abnormality of the sensory visual system. Visual acuity also is reduced in congenital motor nystagmus, probably by the nystagmus itself, but seldom below a level of 20/200.

JERK NYSTAGMUS

This is characterized by a slow drift off the target, followed by a fast corrective saccade. By convention, the nystagmus is named after the quick phase. Jerk nystagmus can be downbeat, upbeat, horizontal (left or right), and torsional. The pattern of nystagmus may vary with gaze position. Some patients will be oblivious to their nystagmus. Others will complain of blurred vision or a subjective to-and-fro movement of the environment (oscillopsia) corresponding to the nystagmus. Fine nystagmus may be difficult to see on gross examination of the eyes. Observation of nystagmoid movements of the optic disc on ophthalmoscopy is a sensitive way to detect subtle nystagmus.

GAZE-EVOKED NYSTAGMUS

This is the most common form of jerk nystagmus. When the eyes are held eccentrically in the orbits, they have a natural tendency to drift back to primary position. The subject compensates by making a corrective saccade to maintain the deviated eye position. Many normal patients have mild gaze-evoked nystagmus. Exaggerated gaze-evoked nystagmus can be induced by drugs (sedatives, anticonvulsants, alcohol); muscle paresis; myasthenia gravis; demyelinating disease; and cerebellopontine angle, brainstem, and cerebellar lesions.

VESTIBULAR NYSTAGMUS

Vestibular nystagmus results from dysfunction of the labyrinth (Ménière's disease), vestibular nerve, or vestibular nucleus in the brainstem. Peripheral vestibular nystagmus often occurs in discrete attacks, with symptoms of nausea and vertigo. There may be associated tinnitus and hearing loss. Sudden shifts in head position may provoke or exacerbate symptoms.

DOWNBEAT NYSTAGMUS

Downbeat nystagmus results from lesions near the craniocervical junction (Chiari malformation, basilar invagination). It also has been reported in brainstem or cerebellar stroke, lithium or anticonvulsant intoxication, alcoholism, and multiple sclerosis. *Upbeat nystagmus* is

associated with damage to the pontine tegmentum from stroke, demyelination, or tumor.

Opsoclonus

This rare, dramatic disorder of eye movements consists of bursts of consecutive saccades (saccadomania). When the saccades are confined to the horizontal plane, the term *ocular flutter* is preferred. It can result from viral encephalitis, trauma, or a paraneoplastic effect of neuroblastoma, breast carcinoma, and other malignancies. It has also been reported as a benign, transient phenomenon in otherwise healthy patients.

FURTHER READING

Bainbridge JW et al: Long-term effect of gene therapy on Leber's congenital amaurosis. N Engl J Med 372:1887, 2015. [PubMed: 25938638]

Buttgerei TF et al: Polymyalgia rheumatica and giant cell arteritis. JAMA 315:2442, 2016. [PubMed: 27299619]

Campochiaro PA et al: Anti-vascular endothelial growth factor agents in the treatment of retinal disease. Ophthalmology 123:S78, 2016. [PubMed: 27664289]

Gross JG et al: Panretinal photocoagulation vs intravitreous ranibizumab for proliferative diabetic retinopathy. JAMA 314:2137, 2015 [PubMed: 26565927]

Jaffe GJ et al: Adalimumbab in patients with active noninfectious uveitis. N Engl J Med 375:932, 2016 [PubMed: 27602665]

Pearson RA et al: Donor and host photoreceptors engage in material transfer following transplantation of post-mitotic photoreceptor precursors. Nat Commun 7:13029, 2016. [PubMed: 27701378]

Stone JH et al: Trial of tocilizumab in giant-cell arteritis. N Engl J Med 377:317, 2017.

Wall M et al: Effect of acetazolamide on visual function in patients with idiopathic intracranial hypertension and mild visual loss: The idiopathic intracranial hypertension treatment trial. JAMA 311:1641, 2014. [PubMed: 24756514]

Williams PA et al: Vitamin B3 modulates mitochrondrial vulnerability and prevents glaucoma in aged mice. Science 355:756, 2017. [PubMed: 28209901]

Yanoff M, Duker J: *Ophthalmology*, 4th ed. Atlanta, Georgia, Saunders, 2014.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 29: Disorders of Smell and Taste

Richard L. Doty; Steven M. Bromley

INTRODUCTION

All environmental chemicals necessary for life enter the body by the nose and mouth. The senses of smell (olfaction) and taste (gustation) monitor such chemicals, determine the flavor and palatability of foods and beverages, and warn of dangerous environmental conditions, including fire, air pollution, leaking natural gas, and bacteria-laden foodstuffs. These senses contribute significantly to quality of life and, when dysfunctional, can have untoward physical and psychological consequences. Indeed, a recent longitudinal study of 1162 non-demented elderly persons found, even after controlling for confounders, that those with the lowest baseline olfactory test scores had a 45% mortality rate over a 4-year period, compared to an 18% mortality rate for those with the highest olfactory test scores. A basic understanding of these senses in health and disease is critical for the physician, because thousands of patients present to doctors' offices each year with complaints of chemosensory dysfunction. Among the more important recent developments in neurology is the discovery that decreased smell function is among the first signs, if not the first sign, of such neurodegenerative diseases as Parkinson's disease (PD) and Alzheimer's disease (AD), signifying their "presymptomatic" phase.

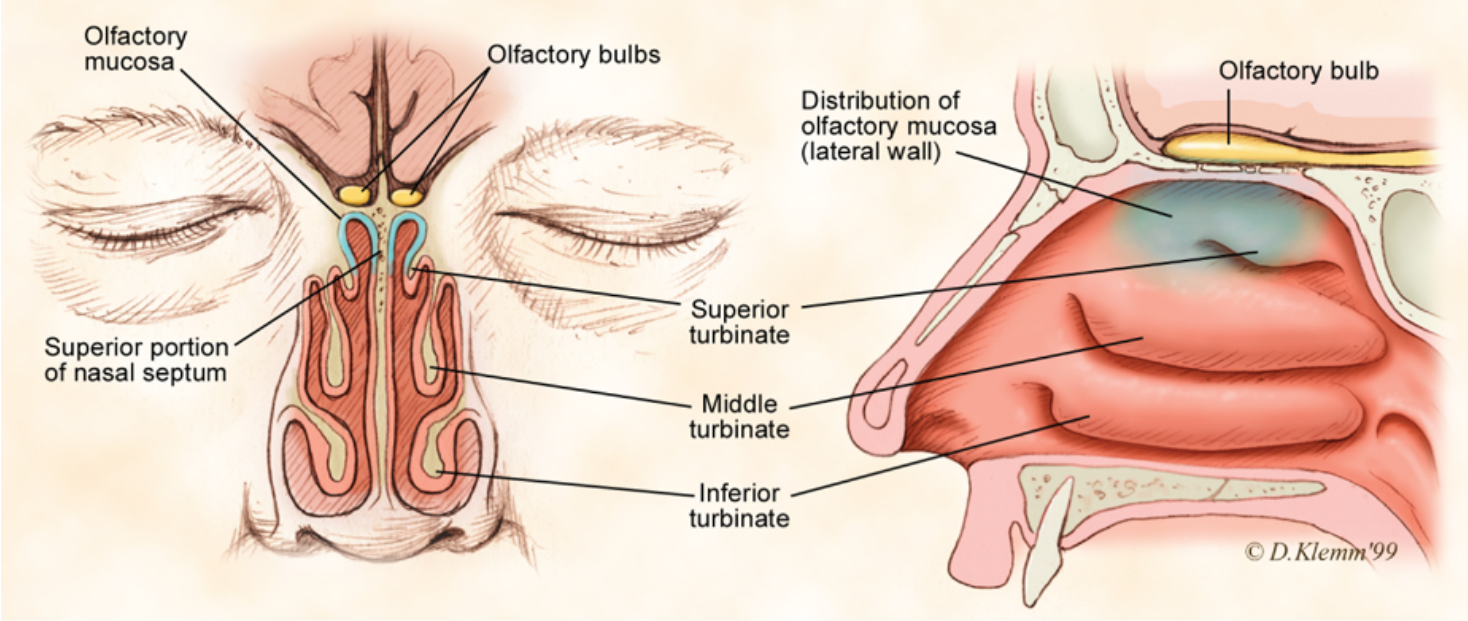
ANATOMY AND PHYSIOLOGY

Olfactory System

Odorous chemicals enter the front of nose during inhalation and active sniffing, as well as the back of the nose (nasopharynx) during deglutition. After reaching the highest recesses of the nasal cavity, they dissolve in the olfactory mucus and diffuse or are actively transported by specialized proteins to receptors located on the cilia of olfactory receptor cells. The cilia, dendrites, cell bodies, and proximal axonal segments of these bipolar cells are located within a unique neuroepithelium covering the cribriform plate, the superior nasal septum, superior turbinate, and sectors of the middle turbinate **(Fig. 29-1)**. Nearly 400 types of G-protein-coupled odor receptors (GPCRs) are expressed on the cilia of the receptor cells, with only one type of GPCR receptor being expressed on a given cell. Other receptors, including trace amine-associated receptors and members of the non-GPCR membrane-spanning 4-domain family, subfamily A (MS4A) protein family, are also present on some receptor cells. Such a plethora of receptor cell types does not exist in any other sensory system. Importantly, when damaged, the receptor cells can be replaced by stem cells near the basement membrane, although such replacement is often incomplete.

FIGURE 29-1

Anatomy of the nose, showing the distribution of olfactory receptors in the roof of the nasal cavity. *(Copyright David Klemm, Faculty and Curriculum Support [FACS], Georgetown University Medical Center; used with permission.)*

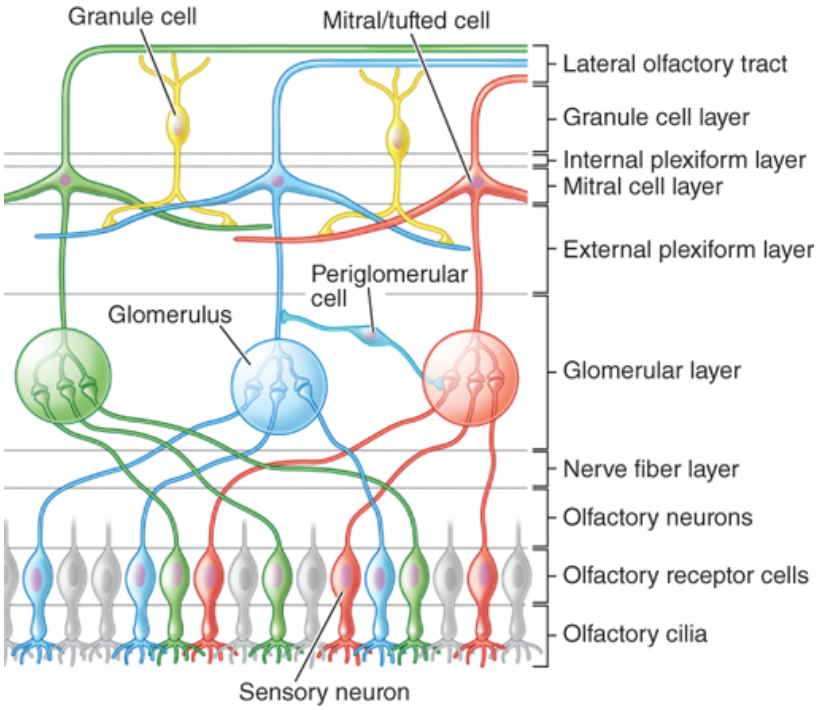


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

After coalescing into bundles surrounded by glia-like ensheathing cells (termed fila), the receptor cell axons pass through the cribriform plate to the olfactory bulbs, where they synapse with dendrites of other cell types within the glomeruli **(Fig. 29-2)**. These spherical structures, which make up a distinct layer of the olfactory bulb, are a site of convergence of information, because many more fibers enter than leave them. Receptor cells that express the same type of receptor project to the same glomeruli, effectively making each glomerulus a functional unit. The major projection neurons of the olfactory system—the mitral and tufted cells—send primary dendrites into the glomeruli, connecting not only with the incoming receptor cell axons, but with dendrites of periglomerular cells. The activity of the mitral/tufted cells is modulated by the periglomerular cells, secondary dendrites from other mitral/tufted cells, and granule cells, the most numerous cells of the bulb. The latter cells, which are largely GABAergic, receive inputs from central brain structures and modulate the output of the mitral/tufted cells. Interestingly, like the olfactory receptor cells, some cells within the bulb undergo replacement. Thus, neuroblasts formed within the anterior subventricular zone of the brain migrate along the rostral migratory stream, ultimately becoming granule and periglomerular cells.

FIGURE 29-2

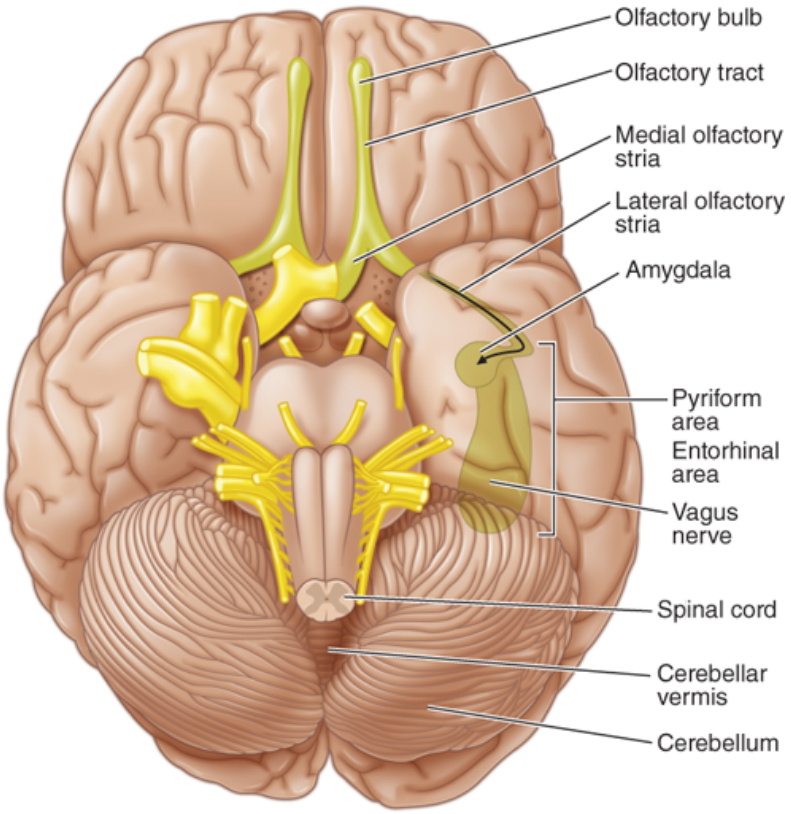
Schematic of the layers and wiring of the olfactory bulb. Each receptor type (red, green, blue) projects to a common glomerulus. The neural activity within each glomerulus is modulated by periglomerular cells. The activity of the primary projection cells, the mitral and tufted cells, is modulated by granule cells, periglomerular cells, and secondary dendrites from adjacent mitral and tufted cells. *(From www.med.yale.edu/neurosurg/treloar/index.html.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The axons of the mitral and tufted cells synapse within secondary olfactory structures which largely comprise the primary olfactory cortex (POC) **(Fig. 29-3)**. The POC is defined as those cortical structures that receive direct projections from the olfactory bulb, most notably the piriform and entorhinal cortices. Although olfaction is unique in that its initial afferent projections bypass the thalamus, persons with damage to the thalamus can exhibit olfactory deficits, particularly ones of odor identification. Such deficits likely reflect the involvement of thalamic connections between the POC and the orbitofrontal cortex (OFC), where odor identification largely occurs. The close anatomic ties between the olfactory system and the amygdala, hippocampus, and hypothalamus help to explain the intimate associations between odor perception and cognitive functions such as memory, motivation, arousal, autonomic activity, digestion, and sex.

FIGURE 29-3
Anatomy of the base of the brain showing the primary olfactory cortex.



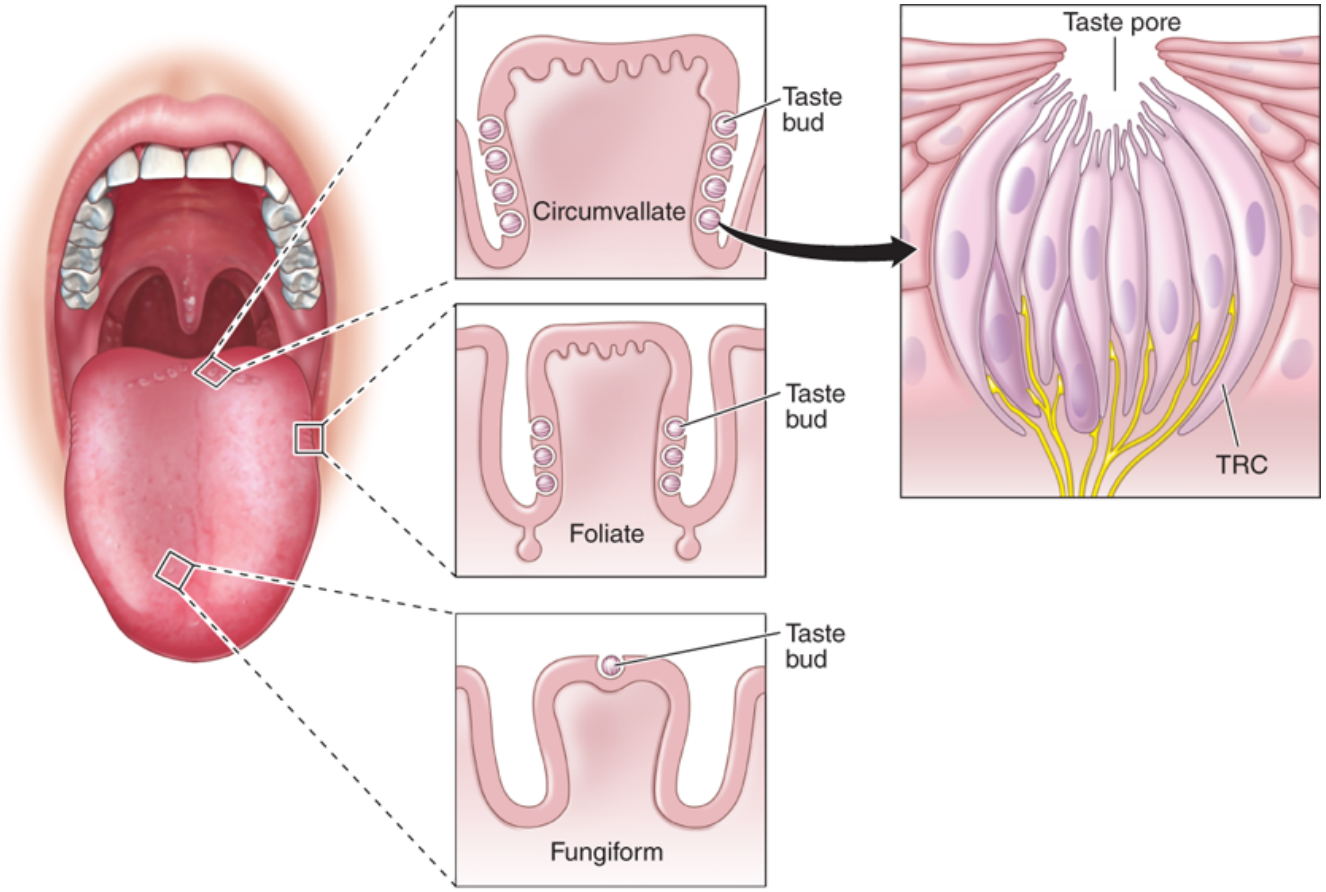
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Taste System

Tastants are sensed by specialized receptor cells present within taste buds—small grapefruit-like segmented structures located on the lateral margins and dorsum of the tongue, roof of the mouth, pharynx, larynx, and superior esophagus (Fig. 29-4). Lingual taste buds are embedded in well-defined protuberances, termed fungiform, foliate, and circumvallate papillae. After dissolving in a liquid, tastants enter the opening of the taste bud—the taste pore—and bind to receptors on microvilli, small extensions of receptor cells within each taste bud. Such binding changes the electrical potential across the taste cell, resulting in neurotransmitter release onto the first-order taste neurons. Although humans have ~7500 taste buds, not all harbor taste-sensitive cells; some contain only one class of receptor (e.g., cells responsive only to sugars), whereas others contain cells sensitive to more than one class. The number of taste receptor cells per taste bud ranges from zero to well over 100. A small family of three G-protein-coupled receptors (GPCRs), namely T1R1, T1R2, and T1R3, mediate sweet and umami taste sensations. Bitter sensations, on the other hand, depend on T2R receptors, a family of ~30 GPCRs expressed on cells different from those that express the sweet and umami receptors. T2Rs sense a wide range of bitter substances but do not distinguish among them. Sour tastants are sensed by the PKD2L1 receptor, a member of the transient receptor potential protein (TRP) family. Perception of salty sensations, such as induced by sodium chloride, arises from the entry of Na^+ ions into the cells via specialized membrane channels, such as the amiloride-sensitive Na^+ channel.

FIGURE 29-4

Schematic of the taste bud and its opening (pore), as well as the location of buds on the three major types of papillae: Fungiform (anterior), foliate (lateral), and circumvallate (posterior).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

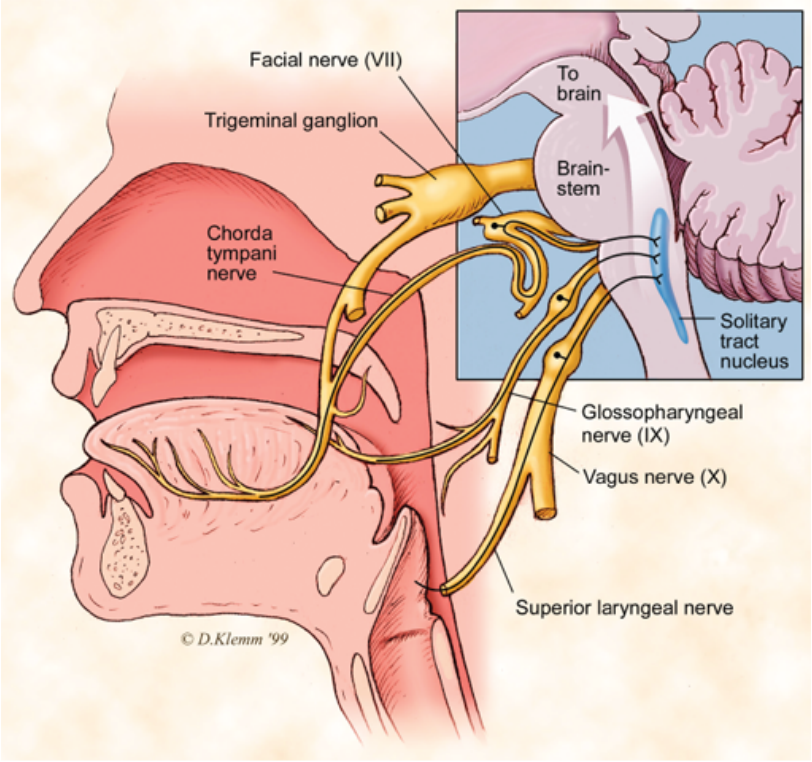
It is now well established that both bitter and sweet taste-related receptors are also present elsewhere in the body, most notably in the alimentary and respiratory tracts. This important discovery generalizes the concept of taste-related chemoreception to areas of the body beyond the mouth and throat, with α-gustducin, the taste-specific G-protein α-subunit, expressed in so-called brush cells found specifically within the human trachea, lung, pancreas, and gallbladder. These brush cells are rich in nitric oxide (NO) synthase, known to defend against xenobiotic organisms, protect the mucosa from acid-induced lesions, and, in the case of the gastrointestinal tract, stimulate vagal and splanchnic afferent neurons. NO further acts on nearby cells, including enteroendocrine cells, absorptive or secretory epithelial cells, mucosal blood vessels, and cells of the immune system. Members of the T2R family of bitter receptors and the sweet receptors of the T1R family have been identified within the gastrointestinal tract and in enteroendocrine cell lines. In some cases, these receptors are important for metabolism, with the T1R3 receptors and gustducin playing decisive roles in the sensing and transport of dietary sugars from the intestinal lumen into absorptive enterocytes via a sodium-dependent glucose transporter and in regulation of hormone release from gut enteroendocrine cells. In other cases, these receptors may be important for airway protection, with a number of T2R bitter receptors in the motile cilia of the human airway that responded to bitter compounds by increasing their beat frequency. One specific T2R38 taste receptor is expressed in human upper respiratory epithelia and responds to acyl-monoserine lactone quorum-sensing molecules secreted by *Pseudomonas aeruginosa* and other gram-negative bacteria. Differences in T2R38 functionality, as related to TAS2R38 genotype, correlate with susceptibility to upper respiratory infections in humans.

Taste information is sent to the brain via three cranial nerves (CNs): CN VII (the *facial nerve*, which involves the intermediate nerve with its branches, the greater petrosal and chorda tympani nerves), CN IX (the *glossopharyngeal nerve*), and CN X (the *vagus nerve*) **(Fig. 29-5)**. CN VII innervates the anterior tongue and all of the soft palate, CN IX innervates the posterior tongue, and CN X innervates the laryngeal surface of the epiglottis, larynx, and proximal portion of the esophagus. The mandibular branch of CN V (V_3) conveys somatosensory information (e.g., touch, burning, cooling, irritation) to the brain. Although not technically a gustatory nerve, CN V shares primary nerve routes with many of the gustatory nerve fibers and adds temperature, texture, pungency, and spiciness to the taste experience. The chorda tympani nerve is famous for taking a recurrent course through the facial canal in the petrosal portion of the temporal bone, passing through the middle ear, and then exiting the skull via the petrotympanic fissure, where it joins the lingual nerve (a division of CN

V) near the tongue. This nerve also carries parasympathetic fibers to the submandibular and sublingual glands, whereas the greater petrosal nerve supplies the palatine glands, thereby influencing saliva production.

FIGURE 29-5

Schematic of the cranial nerves (CNs) that mediate taste function, including the chorda tympani nerve (CN VII), the glossopharyngeal nerve (CN IX), and the vagus nerve (CN X). *(Copyright David Klemm, Faculty and Curriculum Support [FACS], Georgetown University Medical Center; used with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The axons of the projection cells, which synapse with taste buds, enter the rostral portion of the nucleus of the solitary tract (NTS) within the medulla of the brainstem (Fig. 29-5). From the NTS, neurons then project to a division of the ventroposteromedial thalamic nucleus (VPM) via the medial lemniscus. From here, projections are made to the rostral part of the frontal operculum and adjoining insula, a brain region considered the *primary taste cortex* (PTC). Projections from the PTC then go to the *secondary taste cortex*, namely the caudolateral OFC. This brain region is involved in the conscious recognition of taste qualities. Moreover, because it contains cells that are activated by several sensory modalities, it is likely a center for establishing "flavor."

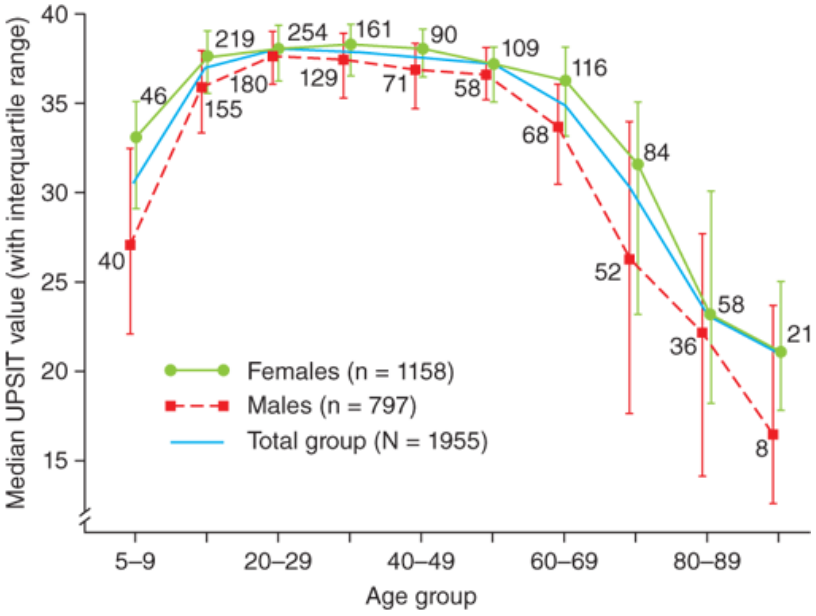
DISORDERS OF OLFACTION

The ability to smell is influenced, in everyday life, by such factors as age, gender, general health, nutrition, smoking, and reproductive state. Women typically outperform men on tests of olfactory function and retain normal smell function to a later age than do men.

Estimates of the prevalence of olfactory dysfunction in the general population vary; a recent cross-sectional analysis from the National Health and Nutrition Examination Survey (NHANES 2013–2014) found an overall prevalence of 13.5%. However, it is apparent that significant decrements in the ability to smell are present in >50% of the population between 65 and 80 years of age and in 75% of those aged ≥80 years **(Fig. 29-6)**. Such presbyosmia helps to explain why many elderly report that food has little flavor, a problem that can result in nutritional disturbances. This also helps to explain why a disproportionate number of elderly die in accidental gas poisonings. A relatively complete listing of conditions and disorders that have been associated with olfactory dysfunction is presented in **Table 29-1**.

FIGURE 29-6

Scores on the University of Pennsylvania Smell Identification Test (UPSIT) as a function of subject age and sex. Numbers by each data point indicate sample sizes. Note that women identify odorants better than men at all ages. *(From RL Doty et al: Science 226:1421, 1984. Copyright © 1984 American Association for the Advancement of Science.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 29-1

Disorders and Conditions Associated with Compromised Olfactory Function, as Measured by Olfactory Testing

22q11 deletion syndrome	Korsakoff's psychosis
AIDS/HIV infection	Laryngopharyngeal reflux disease
Adenoid hypertrophy	Legionnaires' disease
Adrenal cortical insufficiency	Leprosy
Age	Liver disease
Alcoholism	Lubag disease
Allergies	Medications
Alzheimer's disease	Migraine
Amyotrophic lateral sclerosis (ALS)	Multiple sclerosis
Anorexia nervosa	Multi-infarct dementia
Asperger's syndrome	Myasthenia gravis
Ataxias	Narcolepsy with cataplexy
Attention deficit/hyperactivity disorder	Neoplasms, cranial/nasal
Behcet's disease	Nutritional deficiencies
Bardet-Biedl syndrome	Obstructive pulmonary disease
Chagas' disease	Obesity
Chemical exposure	Obsessive compulsive disorder
Chronic obstructive pulmonary disease	Orthostatic tremor
Congenital	Panic disorder
Cushing's syndrome	Parkinson's disease (PD)
Cystic fibrosis	Pick's disease
Degenerative ataxias	Posttraumatic stress disorder
Depression	Pregnancy
Diabetes	Pseudohypoparathyroidism
Down's syndrome	Psychopathy
Epilepsy	Radiation (therapeutic, cranial)
Facial paralysis	REM behavior disorder
Fibromyalgia	Refsum's disease
Frontotemporal lobe degeneration	Renal failure/end-stage kidney disease
Gonadal dysgenesis (Turner's syndrome)	Restless leg syndrome
Granulomatosis with Polyangiitis (Wegener's)	Rhinosinusitis/polyposis
Guamanian ALS/PD/dementia syndrome	Schizophrenia
Head trauma	Seasonal affective disorder
Herpes simplex encephalitis	Sjögren's syndrome
Hypothyroidism	Stroke
Huntington's disease	Systemic sclerosis
Iatrogenesis	Tobacco smoking
Idiopathic inflammatory myopathies	Toxic chemical exposure
Kallmann's syndrome	Upper respiratory infections
	Usher syndrome
	Vitamin B_{12} deficiency

Aside from aging, the three most common identifiable causes of long-lasting or permanent smell loss seen in the clinic are, in order of frequency, severe upper respiratory infections, head trauma, and chronic rhinosinusitis. The physiologic basis for most head trauma–related losses is the shearing and subsequent scarring of the olfactory fila as they pass from the nasal cavity into the brain cavity. The cribriform plate does not have to be fractured or show pathology for smell loss to be present. Severity of trauma, as indexed by a poor

Glasgow Coma Scale score on presentation and the length of posttraumatic amnesia, is associated with higher risk of olfactory impairment. Less than 10% of posttraumatic anosmic patients will recover age-related normal function over time. This increases to nearly 25% of those with less-than-total loss. Upper respiratory infections, such as those associated with the common cold, influenza, pneumonia, or HIV, can directly and permanently harm the olfactory epithelium by decreasing receptor cell number, damaging cilia on remaining receptor cells, and inducing the replacement of sensory epithelium with respiratory epithelium. The smell loss associated with chronic rhinosinusitis is related to disease severity, with most loss occurring in cases where rhinosinusitis and polyposis are both present. Although systemic glucocorticoid therapy can usually induce short-term functional improvement, it does not, on average, return smell test scores to normal, implying that chronic permanent neural loss is present and/or that short-term administration of systemic glucocorticoids does not completely mitigate the inflammation. It is well established that microinflammation in an otherwise seemingly normal epithelium can influence smell function.

A number of neurodegenerative diseases are accompanied by olfactory impairment, including PD, AD, Huntington's disease, parkinsonism-dementia complex of Guam, dementia with Lewy bodies (DLB), multiple system atrophy, corticobasal degeneration, frontotemporal dementia, and Down's syndrome; smell loss can also occur in idiopathic rapid eye movement (REM) behavioral sleep disorder (iRBD), as well as in multiple sclerosis (MS) related to lesions within olfaction-related structures. Olfactory impairment in PD often predates the clinical diagnosis by a number of years. In staged cases, studies of the sequence of formation of abnormal α-synuclein aggregates and Lewy bodies suggest that the olfactory bulbs may be, along with the dorsomotor nucleus of the vagus, the first site of neural damage in PD. In postmortem studies of patients with very mild "presymptomatic" signs of AD, poorer smell function has been associated with higher levels of AD-related pathology. Smell loss is more marked in patients with early clinical manifestations of DLB than in those with mild AD. Interestingly, smell loss is minimal or nonexistent in progressive supranuclear palsy and 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)-induced parkinsonism. The relative contributions of disease-specific pathology or differential damage to forebrain neuromodulator/neurotransmitter systems in explaining different degrees of olfactory dysfunction among the various neurodegenerative diseases are presently unknown.

The smell loss seen in iRBD is of the same magnitude as that found in PD. This is of particular interest because patients with iRBD frequently develop PD and hyposmia. REM behavior disorder is not only seen in its idiopathic form, but can also be associated with narcolepsy **(Chap. 27)**. A study of narcoleptic patients with and without REM behavior disorder demonstrated that narcolepsy, independent of REM behavior disorder, was associated with impairments in olfactory function. Loss of hypothalamic neurons expressing orexin (also known as hypocretin) neuropeptides is believed to be responsible for narcolepsy and cataplexy. Orexin-containing neurons project throughout the entire olfactory system (from the olfactory epithelium to the olfactory cortex), and damage to these projections may be one underlying mechanism for impaired olfactory performance in narcoleptic patients. Administration of intranasal orexin A (hypocretin-1) improved olfactory function, supporting the notion that mild olfactory impairment is not only a primary feature of narcolepsy with cataplexy, but that orexin deficiency may be directly responsible for the loss of smell in this condition.

DISORDERS OF TASTE

The majority of patients who present with taste dysfunction exhibit olfactory, not taste, loss. This is because most flavors attributed to taste actually depend on retronasal stimulation of the olfactory receptors during deglutition. As noted earlier, taste buds only mediate basic tastes such as sweet, sour, bitter, salty, and umami. Significant impairment of whole-mouth gustatory function is rare outside of generalized metabolic disturbances or systemic use of some medications, because taste bud regeneration occurs and peripheral damage alone would require the involvement of multiple CN pathways. Taste function can be influenced by age, diet, smoking behavior, use of medications, and other subject-related factors including (1) the release of foul-tasting materials from the oral cavity from oral medical conditions (e.g., gingivitis, purulent sialadenitis) or appliances; (2) transport problems of tastants to the taste buds (e.g., drying, infections, or inflammatory conditions of the orolingual mucosa), (3) damage to the taste buds themselves (e.g., local trauma, invasive carcinomas), (4) damage to the neural pathways innervating the taste buds (e.g., middle ear infections), (5) damage to central structures (e.g., multiple sclerosis, tumor, epilepsy, stroke), and (6) systemic disturbances of metabolism (e.g., diabetes, thyroid disease, medications). Unlike CN VII, CN IX is relatively protected along its path, although iatrogenic interventions such as tonsillectomy, bronchoscopy, laryngoscopy, endotracheal intubation, and radiation therapy can result in selective injury. CN VII damage commonly results from mastoidectomy, tympanoplasty, and stapedectomy, in some cases inducing persistent metallic sensations. Bell's palsy **(Chap. 433)** is one of the most common causes of CN VII injury that results in taste disturbance. On rare occasions, migraine **(Chap. 422)** is associated with a gustatory prodrome or aura, and in some cases tastants can trigger a migraine attack. Interestingly, dysgeusia occurs in some cases of *burning mouth syndrome* (also termed *glossodynia* or *glossalgia*), as does dry mouth and thirst. Burning mouth syndrome

is likely associated with dysfunction of the trigeminal nerve (CN V). Some of the etiologies suggested for this poorly understood syndrome are amenable to treatment, including (1) nutritional deficiencies (e.g., iron, folic acid, B vitamins, zinc), (2) diabetes mellitus (possibly predisposing to oral candidiasis), (3) denture allergy, (4) mechanical irritation from dentures or oral devices, (5) repetitive movements of the mouth (e.g., tongue thrusting, teeth grinding, jaw clenching), (6) tongue ischemia as a result of temporal arteritis, (7) periodontal disease, (8) reflux esophagitis, and (9) geographic tongue.

Although both taste and smell can be adversely influenced by drugs, taste alterations are more common. Indeed, over 250 medications have been reported to alter the ability to taste. Major offenders include antineoplastic agents, antirheumatic drugs, antibiotics, and blood pressure medications. Terbinafine, a commonly used antifungal, has been linked to taste disturbance lasting up to 3 years. In a recent controlled trial, nearly two-thirds of individuals taking eszopiclone (Lunesta) experienced a bitter dysgeusia that was stronger in women, systematically related to the time since drug administration, and positively correlated with both blood and saliva levels of the drug. Intranasal use of nasal gels and sprays containing zinc, which are common over-the-counter prophylactics for upper respiratory viral infections, has been implicated in loss of smell function. Whether their efficacy in preventing such infections, which are the most common cause of anosmia and hyposmia, outweighs their potential detriment to smell function requires study. Dysgeusia occurs commonly in the context of drugs used to treat or minimize symptoms of cancer, with a weighted prevalence from 56 to 76% depending on the type of cancer treatment. Attempts to prevent taste problems from such drugs using prophylactic zinc sulfate or amifostine have proven to be minimally beneficial. Although antiepileptic medications are occasionally used to treat smell or taste disturbances, the use of topiramate has been reported to result in a reversible loss of an ability to detect and recognize tastes and odors during treatment.

As with olfaction, a number of systemic disorders can affect taste. These include, but are not limited to, chronic renal failure, end-stage liver disease, vitamin and mineral deficiencies, diabetes mellitus, and hypothyroidism. In diabetes, there appears to be a progressive loss of taste beginning with glucose and then extending to other sweeteners, salty stimuli, and then all stimuli. Psychiatric conditions can be associated with chemosensory alterations (e.g., depression, schizophrenia, bulimia). A recent review of tactile, gustatory, and olfactory hallucinations demonstrated that no one type of hallucinatory experience is pathognomonic to any given diagnosis.

Pregnancy is a unique condition with regard to taste function. There appears to be an increase in dislike and intensity of bitter tastes during the first trimester that may help to ensure that pregnant women avoid poisons during a critical phase of fetal development. Similarly, a relative increase in the preference for salt and bitter in the second and third trimesters may support the ingestion of much needed electrolytes to expand fluid volume and support a varied diet.

CLINICAL EVALUATION

In most cases, a careful clinical history will establish the probable etiology of a chemosensory problem, including questions about its nature, onset, duration, and pattern of fluctuations. *Sudden loss* suggests the possibility of head trauma, ischemia, infection, or a psychiatric condition. *Gradual loss* can reflect the development of a progressive obstructive lesion, although gradual loss can also follow head trauma. *Intermittent loss* suggests the likelihood of an inflammatory process. The patient should be asked about potential precipitating events, such as cold or flu infections prior to symptom onset, because these often go underappreciated. Information regarding head trauma, smoking habits, drug and alcohol abuse (e.g., intranasal cocaine, chronic alcoholism), exposures to pesticides and other toxic agents, and medical interventions is also informative. A determination of all the medications that the patient was taking before and at the time of symptom onset is important, because many can cause chemosensory disturbances. Comorbid medical conditions associated with smell impairment, such as renal failure, liver disease, hypothyroidism, diabetes, or dementia, should be assessed. Delayed puberty in association with anosmia (with or without midline craniofacial abnormalities, deafness, and renal anomalies) suggests the possibility of Kallmann's syndrome. Recollection of epistaxis, discharge (clear, purulent, or bloody), nasal obstruction, allergies, and somatic symptoms, including headache or irritation, may have localizing value. Questions related to memory, parkinsonian symptoms, and seizure activity (e.g., automatisms, blackouts, auras, déjà vu) should be posed. Pending litigation and the possibility of malingering should be considered. Modern forced-choice olfactory tests can detect malingering from improbable responses.

Neurologic and otorhinolaryngologic (ORL) examinations, along with appropriate brain and nasosinus imaging, aid in the evaluation of patients with olfactory or gustatory complaints. The neural evaluation should focus on CN function, with particular attention to possible skull base and intracranial lesions. Visual acuity, field, and optic disc examinations aid in the detection of intracranial mass lesions that produce raised intracranial pressure (papilledema) and optic atrophy. Foster Kennedy syndrome refers to raised intracranial pressure plus a compressive optic neuropathy; typical causes are olfactory groove meningiomas or other frontal lobe tumors. The ORL examination

should thoroughly assess the intranasal architecture and mucosal surfaces. Polyps, masses, and adhesions of the turbinates to the septum may compromise the flow of air to the olfactory receptors, because less than a fifth of the inspired air traverses the olfactory cleft in the unobstructed state. Blood tests may be helpful to identify such conditions as diabetes, infection, heavy metal exposure, nutritional deficiency (e.g., vitamin B_6 or B_{12}), allergy, and thyroid, liver, and kidney disease.

As with other sensory disorders, quantitative sensory testing is advised. Self-reports of patients can be misleading, and a number of patients who complain of chemosensory dysfunction have normal function for their age and gender. Quantitative smell and taste testing provides objective information for worker's compensation and other legal claims, as well as a way to accurately assess the effects of treatment interventions. A number of standardized olfactory and taste tests are commercially available. The most widely used of these tests, the 40-item University of Pennsylvania Smell Identification Test (UPSIT), uses norms based on nearly 4000 normal subjects. A determination is made of both absolute dysfunction (i.e., mild loss, moderate loss, severe loss, total loss, probable malingering) and relative dysfunction (percentile rank for age and gender). Although electrophysiologic testing is available at some smell and taste centers (e.g., odor event-related potentials), they require complex stimulus presentation and recording equipment and rarely provide additional diagnostic information. With the exception of electrogustometers, commercially available taste tests have only recently become available. Most use filter paper strips impregnated with tastants, so no stimulus preparation is required.

TREATMENT AND MANAGEMENT

Given the various mechanisms by which olfactory and gustatory disturbance can occur, management of patients tends to be condition-specific. For example, patients with hypothyroidism, diabetes, or infections often benefit from specific treatments to correct the underlying disease process that is adversely influencing chemoreception. For most patients who present primarily with obstructive/transport loss affecting the nasal and paranasal regions (e.g., allergic rhinitis, polyposis, intranasal neoplasms, nasal deviations), medical and/or surgical intervention is often beneficial. Antifungal and antibiotic treatments may reverse taste problems secondary to candidiasis or other oral infections. Chlorhexidine mouthwash mitigates some salty or bitter dysgeusias, conceivably as a result of its strong positive charge. Excessive dryness of the oral mucosa is a problem with many medications and conditions, and artificial saliva (e.g., Xerolube) or oral pilocarpine treatments may prove beneficial. Other methods to improve salivary flow include the use of mints, lozenges, or sugarless gum. Flavor enhancers may make food more palatable (e.g., monosodium glutamate), but caution is advised to avoid overusing ingredients containing sodium or sugar, particularly in circumstances when a patient also has underlying hypertension or diabetes. Medications that induce distortions of taste can often be discontinued and replaced with other types of medications or modes of therapy. As mentioned earlier, pharmacologic agents result in taste disturbances much more frequently than smell disturbances. It is important to note, however, that many drug-related effects are long lasting and not reversed by short-term drug discontinuance.

A recent study of endoscopic sinus surgery in patients with chronic rhinosinusitis and hyposmia revealed that patients with severe olfactory dysfunction prior to the surgery had a more dramatic and sustained improvement over time compared to patients with more mild olfactory dysfunction prior to intervention. In the case of intranasal and sinus-related inflammatory conditions, such as seen with allergy, viruses, and traumas, the use of intranasal or systemic glucocorticoids may also be helpful. One common approach is to use a tapering course of oral prednisone. Topical intranasal administration of glucocorticoids was found to be less effective in general than systemic administration, however the effects of different nasal administration techniques were not analyzed; for example, intranasal glucocorticoids are more effective if administered in the Moffett's position (head in the inverted position such as over the edge of the bed with the bridge of the nose perpendicular to the floor). After head trauma, an initial trial of glucocorticoids may help to reduce local edema and the potential deleterious deposition of scar tissue around olfactory fila at the level of the cribriform plate.

Treatments are limited for patients with chemosensory loss or primary injury to neural pathways. Nonetheless, spontaneous recovery can occur. In a follow-up study of 542 patients presenting to our center with smell loss from a variety of causes, modest improvement occurred over an average time period of 4 years in about half of the participants. However, only 11% of the anosmic and 23% of the hyposmic patients regained normal age-related function. Interestingly, the amount of dysfunction present at the time of presentation, not etiology, was the best predictor of prognosis. Other predictors were age and the duration of dysfunction prior to initial testing.

Several studies have reported that patients with hyposmia may benefit from repeated smelling of odors over the course of weeks or months. The usual paradigm is to smell odors such as eucalyptol, citronella, eugenol, and phyenyl ethyl alcohol before going to bed and immediately upon awakening each day. The rationale for such an approach comes from animal studies demonstrating that prolonged exposure to odorants can induce increased neural activity within the olfactory bulb. There is also limited evidence that α-lipoic acid (400

mg/d), an essential cofactor for many enzyme complexes with possible antioxidant effects, may be beneficial in mitigating smell loss following viral infection of the upper respiratory tract. However, double-blind studies are needed to confirm this observation. α-lipoic acid has also been suggested to be useful in some cases of hypogeusia and burning mouth syndrome.

The use of zinc and vitamin A in treating olfactory disturbances is controversial, and there does not appear to be much benefit beyond replenishing established deficiencies. However, zinc has been shown to improve taste function secondary to hepatic deficiencies, and retinoids (bioactive vitamin A derivatives) are known to play an essential role in the survival of olfactory neurons. One protocol in which zinc was infused with chemotherapy treatments suggested a possible protective effect against developing taste impairment. Diseases of the alimentary tract can not only influence chemoreceptive function, but also occasionally influence vitamin B_{12} absorption. This can result in a relative deficiency of vitamin B_{12}, theoretically contributing to olfactory nerve disturbance. Vitamin B_2 (riboflavin) and magnesium supplements are reported in the alternative literature to aid in the management of migraine that, in turn, may be associated with smell dysfunction. Because vitamin D deficiency is a cofactor of chemotherapy-induced mucocutaneous toxicity and dysgeusia, adding vitamin D_3, 1000–2000 units per day, may benefit some patients with smell and taste complaints during or following chemotherapy.

A number of medications have reportedly been used with success in ameliorating olfactory symptoms, although strong scientific evidence for efficacy is generally lacking. A report that theophylline improved smell function was uncontrolled and failed to account for the fact that some meaningful improvement occurs without treatment; indeed, the percentage of responders was about the same (~50%) as that noted by others to show spontaneous improvement over a similar time period. Antiepileptics and some antidepressants (e.g., amitriptyline) have been used to treat dysosmias and smell distortions, particularly following head trauma. Ironically, amitriptyline is also frequently on the list of medications that can ultimately distort smell and taste function, possibly from its anticholinergic effects. One study suggested that the centrally acting acetylcholinesterase inhibitor donepezil in AD resulted in improvements on smell identification measures that correlated with overall clinician-based impressions of change in dementia severity scores.

Alternative therapies, such as acupuncture, meditation, cognitive-behavioral therapy, and yoga, can help patients manage uncomfortable experiences associated with chemosensory disturbance and oral pain syndromes and to cope with the psychosocial stressors surrounding the impairment. Additionally, modification of diet and eating habits is also important. By accentuating the other sensory experiences of a meal, such as food texture, aroma, temperature, and color, one can optimize the overall eating experience for a patient. In some cases, a flavor enhancer like monosodium glutamate (MSG) can be added to foods to increase palatability and encourage intake.

Proper oral and nasal hygiene and routine dental care are extremely important ways for patients to protect themselves from disorders of the mouth and nose that can ultimately result in chemosensory disturbance. Patients should be warned not to overcompensate for their taste loss by adding excessive amounts of sugar or salt. Smoking cessation and the discontinuance of oral tobacco use are essential in the management of any patient with smell and/or taste disturbance and should be repeatedly emphasized.

A major and often overlooked element of therapy comes from chemosensory testing itself. Confirmation or lack of conformation of loss is beneficial to patients who come to believe, in light of unsupportive family members and medical providers, that they may be “crazy.” In cases where the loss is minor, patients can be informed of the likelihood of a more positive prognosis. Importantly, quantitative testing places the patient’s problem into overall perspective. Thus, it is often therapeutic for an older person to know that, while his or her smell function is not what it used to be, it still falls above the average of his or her peer group. Without testing, many such patients are simply told that they are getting old and nothing can be done for them, leading in some cases to depression and decreased self-esteem.

FURTHER READING

Devanand DP et al: Olfactory identification deficits are associated with increased mortality in a multiethnic urban community. Ann Neurol 78:401, 2015. [PubMed: 26031760]

Doty RL: Olfaction in Parkinson’s disease and related disorders. Neurobiol Dis 46:527, 2012. [PubMed: 22192366]

Doty RL: Neurotoxic exposure and alterations in olfaction and gustation. Handbook Clin Neurol 131:299, 2015.

Doty RL (ed): *Handbook of Olfaction and Gustation*, 3rd ed. Hoboken, Wiley-Liss, 2015.

Doty RL et al: Influences of hormone replacement therapy on olfactory and cognitive function in the menopause. Neurobiol Aging 36:2053, 2015. [PubMed: 25850354]

Doty RL et al: Taste function in early stage treated and untreated Parkinson's disease. J Neurol 262:547, 2015. [PubMed: 25480568]

Kohli P et al: The association between olfaction and depression: A systematic review. Chem Senses 41:479, 2016. [PubMed: 27170667]

Liu G et al: Prevalence and risk factors of taste and smell impairment in a nationwide sample of the US population: A cross-sectional study. BMJ Open 6:e013246, 2016. [PubMed: 28157672]

London B et al: Predictors of prognosis in patients with olfactory disturbance. Ann Neurol 63:159, 2008. [PubMed: 18058814]

Pekala K et al: Efficacy of olfactory training in patients with olfactory loss: A systematic review and meta-analysis. Int Forum Allergy Rhinol 6:299, 2016. [PubMed: 26624966]

Perricone C et al: Smell and autoimmunity: A comprehensive review. Clin Rev Allergy Immunol 45:87, 2013. [PubMed: 23233263]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 30: Disorders of Hearing

Anil K. Lalwani

INTRODUCTION

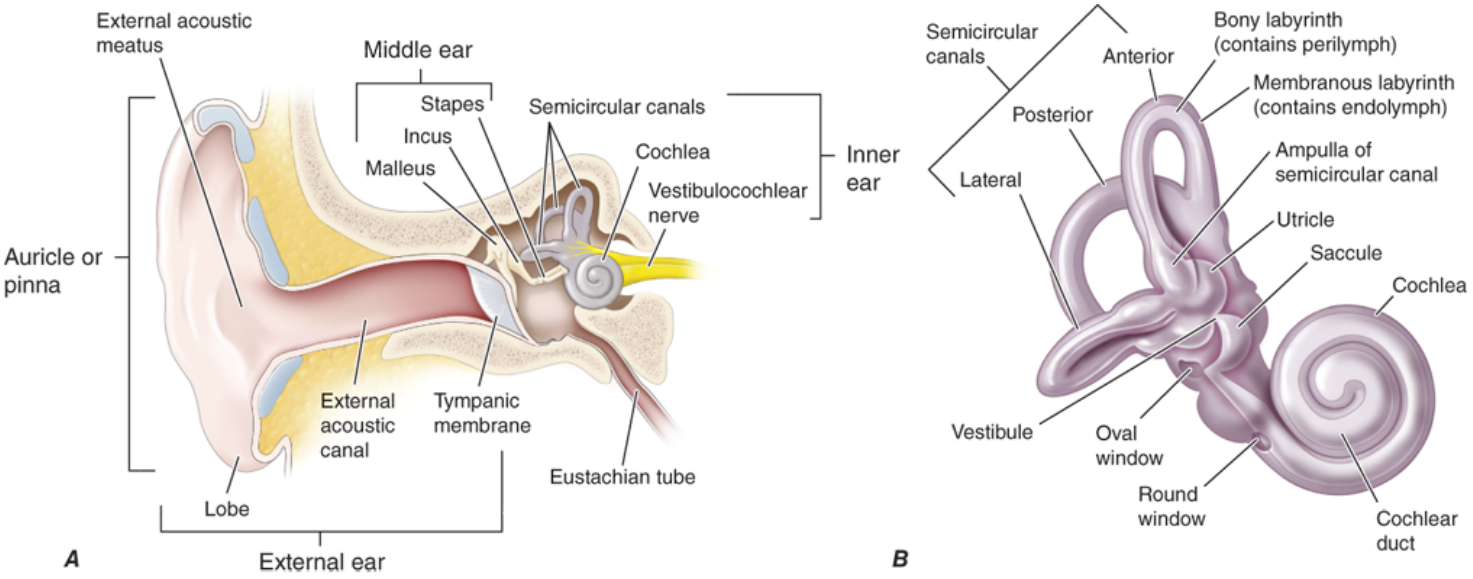
Hearing loss can present at any age and is one of the most common sensory disorders in humans. Nearly 10% of the adult population has some hearing loss, and one-third of individuals age >65 years have a hearing loss of sufficient magnitude to require a hearing aid.

PHYSIOLOGY OF HEARING

The function of the external and middle ear is to amplify sound to facilitate conversion of the mechanical energy of the sound wave into an electrical signal by the inner ear hair cells, a process called mechanotransduction **(Fig. 30-1)**. Sound waves enter the external auditory canal and set the tympanic membrane (eardrum) in motion, which in turn moves the malleus, incus, and stapes of the middle ear. Movement of the footplate of the stapes causes pressure changes in the fluid-filled inner ear, eliciting a traveling wave in the basilar membrane of the cochlea. The tympanic membrane and the ossicular chain in the middle ear serve as an impedance-matching mechanism, improving the efficiency of energy transfer from air to the fluid-filled inner ear. In its absence, nearly 99.9% of the acoustical energy would be reflected and thus not heard. Instead, the ear drum and the ossicles boost the sound energy nearly 200-fold by the time it reaches the inner ear.

FIGURE 30-1

Ear anatomy. ***A.*** Drawing of modified coronal section through external ear and temporal bone, with structures of the middle and inner ear demonstrated. ***B.*** High-resolution view of inner ear.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Within the cochlea of the inner ear, there are two types of hair cells that aid in hearing: inner and outer. The inner and outer hair cells of the organ of Corti have different innervation patterns, but both are mechanoreceptors; they detect the mechanical energy of the acoustical signal and aid its conversion to an electrical signal that travels by the auditory nerve. The afferent innervation relates principally to the inner hair cells while the efferent innervation relates principally to the outer hair cells. The outer hair cells outnumber the inner hair cells by nearly 6:1 (20,000 vs 3500). The motility of the outer hair cells alters the micromechanics of the inner hair cells, creating a cochlear amplifier, which explains the exquisite sensitivity and frequency selectivity of the cochlea.

Stereocilia of the hair cells of the organ of Corti, which rests on the basilar membrane, are in contact with the tectorial membrane and are deformed by the traveling wave. The deformation stretches tiny filamentous connections (tip links) between stereocilia, leading to opening of ion channels, influx of potassium, and hair cell depolarization and subsequent neurotransmission. A point of maximal displacement of the basilar membrane is determined by

the frequency of the stimulating tone. High-frequency tones cause maximal displacement of the basilar membrane near the base of the cochlea, whereas for low-frequency sounds, the point of maximal displacement is toward the apex of the cochlea.

Beginning in the cochlea, the frequency specificity is maintained at each point of the central auditory pathway: dorsal and ventral cochlear nuclei, trapezoid body, superior olivary complex, lateral lemniscus, inferior colliculus, medial geniculate body, and auditory cortex. At low frequencies, individual auditory nerve fibers can respond more or less synchronously with the stimulating tone. At higher frequencies, phase-locking occurs so that neurons alternate in response to particular phases of the cycle of the sound wave. Intensity is encoded by the amount of neural activity in individual neurons, the number of neurons that are active, and the specific neurons that are activated.

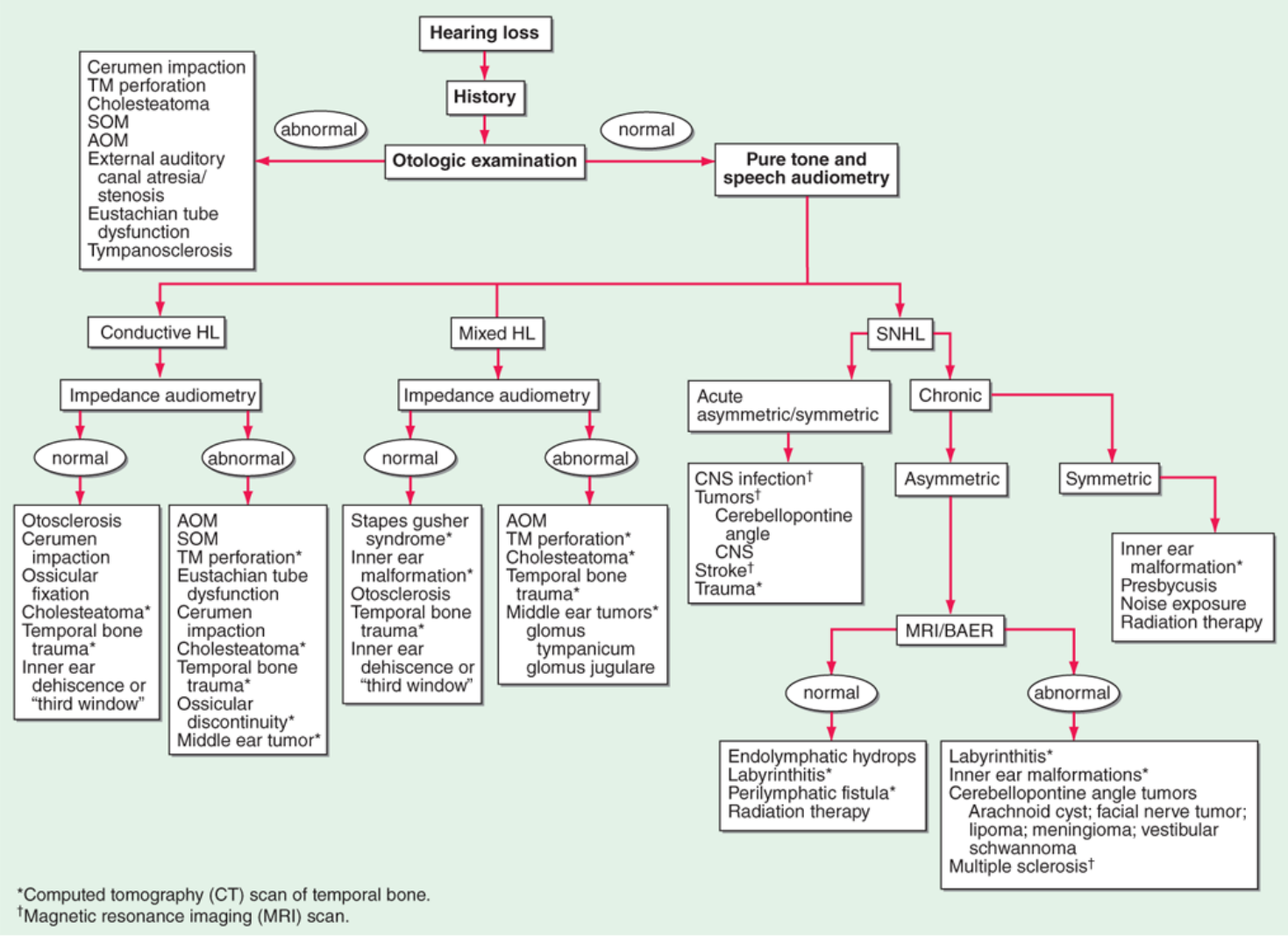
There is evidence that the right and left ears as well as the central nervous system may process speech asymmetrically. Generally, a sound is processed symmetrically from the peripheral to the central auditory system. However, a “right ear advantage” exists for dichotic listening tasks, in which subjects are asked to report on competing sounds presented to each ear. In most individuals, a perceptual right ear advantage for consonant-vowel syllables, stop consonants, and words also exists. Similarly, whereas central auditory processing for sounds is symmetric with minimal lateral specialization for the most part, speech processing is lateralized. There is specialization of the left auditory cortex for speech recognition and production, and of the right hemisphere for emotional and tonal aspects of speech. Left hemisphere dominance for speech is found in 95–98% of right-handed persons and 70–80% of left-handed persons.

DISORDERS OF THE SENSE OF HEARING

Hearing loss can result from disorders of the auricle, external auditory canal, middle ear, inner ear, or central auditory pathways **(Fig. 30-2)**. In general, lesions in the auricle, external auditory canal, or middle ear that impede the transmission of sound from the external environment to the inner ear cause conductive hearing loss, whereas lesions that impair mechanotransduction in the inner ear or transmission of the electrical signal along the eighth nerve to the brain cause sensorineural hearing loss.

FIGURE 30-2

An algorithm for the approach to hearing loss. AOM, acute otitis media; BAER, brainstem auditory-evoked response; CNS, central nervous system; HL, hearing loss; SNHL, sensorineural hearing loss; SOM, serous otitis media; TM, tympanic membrane.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Conductive Hearing Loss

The external ear, the external auditory canal, and the middle ear apparatus are designed to collect and amplify sound and efficiently transfer the mechanical energy of the sound wave to the fluid-filled cochlea. Factors that obstruct the transmission of sound or dampen the acoustical energy result in conductive hearing loss. Conductive hearing loss can occur from obstruction of the external auditory canal by cerumen, debris, and foreign bodies; swelling of the lining of the canal; atresia or neoplasms of the canal; perforations of the tympanic membrane; disruption of the ossicular chain, as occurs with necrosis of the long process of the incus in trauma or infection; otosclerosis; or fluid, scarring, or neoplasms in the middle ear. Rarely, inner ear malformations or pathologies, such as superior semicircular canal dehiscence, lateral semicircular canal dysplasia, incomplete partition of the inner ear, and large vestibular aqueduct, are also associated with conductive hearing loss.

Eustachian tube dysfunction is extremely common in adults and may predispose to acute otitis media (AOM) or serous otitis media (SOM). Trauma, AOM, and chronic otitis media are the usual factors responsible for tympanic membrane perforation. While small perforations often heal spontaneously, larger defects usually require surgical intervention. Tympanoplasty is highly effective (>90%) in the repair of tympanic membrane perforations. Otoscopy is usually sufficient to diagnose AOM, SOM, chronic otitis media, cerumen impaction, tympanic membrane perforation, and eustachian tube dysfunction; tympanometry can be useful to confirm the clinical suspicion of these conditions.

Cholesteatoma, a benign tumor composed of stratified squamous epithelium in the middle ear or mastoid, occurs frequently in adults. This is a slowly growing lesion that destroys bone and normal ear tissue. Theories of pathogenesis include traumatic immigration and invasion of squamous epithelium through a retraction pocket of the tympanic membrane, implantation of squamous epithelia in the middle ear through a perforation or surgery, and metaplasia following chronic infection and irritation. A chronically draining ear that fails to respond to appropriate antibiotic therapy should raise suspicion of a cholesteatoma. On examination, there is often a perforation of the tympanic membrane filled with cheesy white squamous debris. The presence of an aural polyp obscuring the tympanic membrane is also highly suggestive of an underlying cholesteatoma. Conductive hearing

loss secondary to ossicular erosion is common. Bony destruction visualized on computerized tomography (CT) of the temporal bone is highly suggestive of cholesteatoma. Surgery is required to remove this destructive process and reconstruct the ossicles.

Conductive hearing loss with a normal ear canal and intact tympanic membrane suggests either ossicular pathology or the presence of "third window" in the inner ear (see below). Fixation of the stapes from *otosclerosis* is a common cause of low-frequency conductive hearing loss. It occurs equally in men and women and is inherited as an autosomal dominant trait with incomplete penetrance; in some cases, it may be a manifestation of osteogenesis imperfecta. Hearing impairment usually presents between the late teens and the forties. In women, the otosclerotic process is accelerated during pregnancy, and the hearing loss is often first noticeable at this time. A hearing aid or a simple outpatient surgical procedure (stapedectomy) can provide excellent auditory rehabilitation. Extension of otosclerosis beyond the stapes footplate to involve the cochlea (cochlear otosclerosis) can lead to mixed or sensorineural hearing loss. Fluoride therapy to prevent hearing loss from cochlear otosclerosis is of uncertain value.

Disorders that lead to the formation of a pathologic "third window" in the inner ear can be associated with conductive hearing loss. There are normally two major openings, or windows, that connect the inner ear with the middle ear and serve as conduits for transmission of sound; these are, respectively, the oval and round windows. A third window is formed where the normally hard otic bone surrounding the inner ear is eroded; dissipation of the acoustic energy at the third window is responsible for the "inner ear conductive hearing loss." The superior semicircular canal dehiscence syndrome resulting from erosion of the otic bone over the superior circular canal can present with conductive hearing loss that mimics otosclerosis. A common symptom is vertigo evoked by loud sounds (Tullio phenomenon), by Valsalva maneuvers that change middle ear pressure, or by applying positive pressure on the tragus (the cartilage anterior to the external opening of the ear canal). Patients with this syndrome also complain of fullness of the ear, pulsatile tinnitus, and being able to hear the movement of their eyes and neck. A large jugular bulb or jugular bulb diverticulum can create a "third window" by eroding into the vestibular aqueduct or posterior semicircular canal; the symptoms are similar to those of the superior semicircular canal dehiscence syndrome. Low activation threshold on the vestibular-evoked myogenic potential test (VEMP test, see below) and inner ear erosion on CT are diagnostic. Recalcitrant vertigo and dizziness may respond to surgical repair of the dehiscence.

Sensorineural Hearing Loss

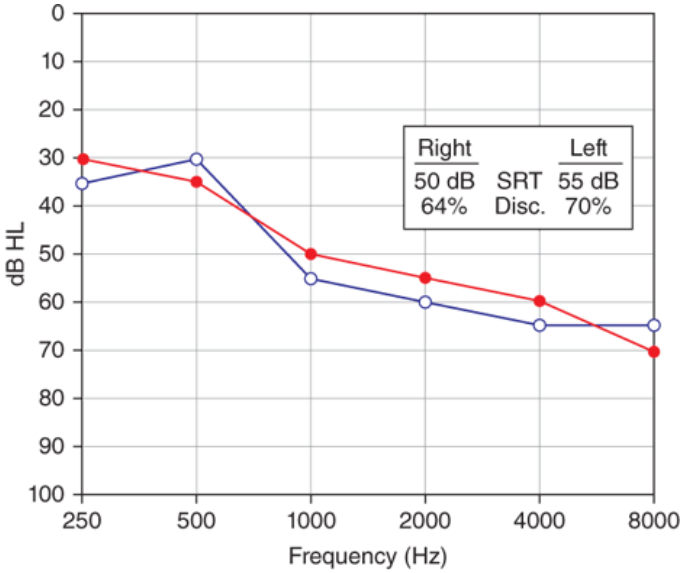
Sensorineural hearing loss results from either damage to the mechanotransduction apparatus of the cochlea or disruption of the electrical conduction pathway from the inner ear to the brain. Thus, injury to hair cells, supporting cells, auditory neurons, or the central auditory pathway can cause sensorineural hearing loss. Damage to the hair cells of the organ of Corti may be caused by intense noise, viral infections, ototoxic drugs (e.g., salicylates, quinine and its synthetic analogues, aminoglycoside antibiotics, loop diuretics such as furosemide and ethacrynic acid, and cancer chemotherapeutic agents such as cisplatin), fractures of the temporal bone, meningitis, cochlear otosclerosis (see above), Ménière's disease, and aging. Congenital malformations of the inner ear may be the cause of hearing loss in some adults. Genetic predisposition alone or in concert with environmental exposures may also be responsible (see below).

Exposure to loud noise, either a short burst or over a more prolonged period of time, can lead to noise-induced hearing loss. Acute exposure to noise can lead to either temporary or permanent threshold shifts, depending on the intensity and duration of sound, due to hair cell injury and/or death. Typically, with permanent hearing loss there is a "noise notch" with elevated hearing thresholds at 3000–4000 Hz. More recently, loud noise exposure has also been associated with "hidden hearing loss"—hidden, because routine audiometry shows the pure tone hearing to be normal. Patients usually complains of not being able to hear clearly and are more bothered by the presence of background noise. In contrast to hair cell loss, hidden hearing loss is thought to be due to loss of auditory synapses on hair cells following noise exposure. In an increasingly noisy world, avoiding acoustic trauma with ear plugs or earmuffs is highly recommended to prevent noise-induced or hidden hearing loss.

Presbycusis (age-associated hearing loss) is the most common cause of sensorineural hearing loss in adults. It is estimated to affect over half of the adults aged >75 in the United States, a population that is expected to double in size over the next 40 years. In the early stages, it is characterized by symmetric, gentle to sharply sloping, high-frequency hearing loss **(Fig. 30-3)**. With progression, the hearing loss involves all frequencies. More importantly, the hearing impairment is associated with significant loss in clarity. There is a loss of discrimination for phonemes, recruitment (abnormal growth of loudness), and particular difficulty in understanding speech in noisy environments such as at restaurants and social events. Poor hearing is also associated with an increased incidence of cognitive impairment and rate of cognitive decline. In the elderly, left untreated, hearing loss leads to diminished quality of life, and has been shown to increase overall morbidity and mortality through falls and accidents. Hearing aids are helpful in enhancing the signal-to-noise ratio by amplifying sounds that are close to the listener. Hearing aid use has been shown to reduce cognitive decline. Although hearing aids are able to amplify sounds, they cannot restore the clarity of hearing. Thus, amplification with hearing aids may provide only limited rehabilitation once the word recognition score deteriorates below 50%. Cochlear implants are the treatment of choice when hearing aids prove inadequate, even when hearing loss is incomplete (see below).

FIGURE 30-3

Presbyacusis or age-related hearing loss. The audiogram shows a moderate to severe downsloping sensorineural hearing loss typical of presbyacusis. The loss of high-frequency hearing is associated with a decreased speech discrimination score; consequently, patients complain of lack of clarity of hearing, especially in a noisy background. HL, hearing threshold level; SRT, speech reception threshold.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ménière's disease is characterized by episodic vertigo, fluctuating sensorineural hearing loss, tinnitus, and aural fullness. Tinnitus and/or deafness may be absent during the initial attacks of vertigo, but it invariably appears as the disease progresses and increases in severity during acute attacks. The annual incidence of Ménière's disease is 0.5–7.5 per 1000; onset is most frequently in the fifth decade of life but may also occur in young adults or the elderly. Histologically, there is distention of the endolymphatic system (endolymphatic hydrops) leading to degeneration of vestibular and cochlear hair cells. This may result from endolymphatic sac dysfunction secondary to infection, trauma, autoimmune disease, inflammatory causes, or tumor; an idiopathic etiology constitutes the largest category and is most accurately referred to as Ménière's disease. Although any pattern of hearing loss can be observed, typically, low-frequency, unilateral sensorineural hearing impairment is present. An abnormal VEMP test may be helpful in detecting Ménière's disease in a clinically unaffected contralateral ear. Magnetic resonance imaging (MRI) should be obtained to exclude retrocochlear pathology such as a cerebellopontine angle tumor or demyelinating disorder. Therapy is directed toward the control of vertigo. A 2-g/d low-salt diet is the mainstay of treatment for control of rotatory vertigo. Diuretics, a short course of oral glucocorticoids, intratympanic glucocorticoids, or intratympanic gentamicin may also be useful adjuncts in recalcitrant cases. Surgical therapy of vertigo is reserved for unresponsive cases and includes endolymphatic sac decompression, labyrinthectomy, and vestibular nerve section. Both labyrinthectomy and vestibular nerve section abolish rotatory vertigo in >90% of cases. Unfortunately, there is no effective therapy for hearing loss, tinnitus, or aural fullness from Ménière's disease.

Sensorineural hearing loss may also result from any neoplastic, vascular, demyelinating, infectious, or degenerative disease or trauma affecting the central auditory pathways. Characteristically, a reduction in clarity of hearing and speech comprehension is much greater than the loss of the ability to hear pure tone. Auditory testing is consistent with an auditory neuropathy; normal otoacoustic emissions (OAEs) and an abnormal auditory brainstem response (ABR) is typical (see below). Hearing loss can accompany hereditary sensorimotor neuropathies and inherited disorders of myelin. Tumors of the cerebellopontine angle such as vestibular schwannoma and meningioma **(Chap. 86)** usually present with asymmetric sensorineural hearing loss with greater deterioration of speech understanding than pure tone hearing. Multiple sclerosis **(Chap. 436)** may present with acute unilateral or bilateral hearing loss; typically, pure tone testing remains relatively stable while speech understanding fluctuates. Isolated labyrinthine infarction can present with acute hearing loss and vertigo due to a cerebrovascular accident involving the posterior circulation, usually the anterior inferior cerebellar artery; it may also be the heralding sign of impending catastrophic basilar artery infarction **(Chap. 419)**. HIV **(Chap. 197)**, which can produce both peripheral and central auditory system pathology, is another consideration in the evaluation of sensorineural hearing impairment.

A finding of conductive and sensorineural hearing loss in combination is termed *mixed hearing loss*. Mixed hearing losses can result from pathology of both the middle and inner ear, as can occur in otosclerosis involving the ossicles and the cochlea, head trauma, chronic otitis media, cholesteatoma, middle ear tumors, and some inner ear malformations.

Trauma resulting in temporal bone fractures may be associated with conductive, sensorineural, or mixed hearing loss. If the fracture spares the inner ear, there may simply be conductive hearing loss due to rupture of the tympanic membrane or disruption of the ossicular chain. These abnormalities can be surgically corrected. Profound hearing loss and severe vertigo are associated with temporal bone fractures involving the inner ear. A perilymphatic fistula associated with leakage of inner ear fluid into the middle ear can occur and may require surgical repair. An associated facial nerve injury is not uncommon. CT is best suited to assess fracture of the traumatized temporal bone, evaluate the ear canal, and determine the integrity of the ossicular chain and involvement of the inner ear. Cerebrospinal fluid leaks that accompany temporal bone fractures are usually self-limited; the value of prophylactic antibiotics is uncertain.

Tinnitus is defined as the perception of a sound when there is no sound in the environment. It can have a buzzing, roaring, or ringing quality and may be pulsatile (synchronous with the heartbeat). Tinnitus is often associated with either a conductive or sensorineural hearing loss. The pathophysiology of

tinnitus is not well understood. The cause of the tinnitus can usually be determined by finding the cause of the associated hearing loss. Tinnitus may be the first symptom of a serious condition such as a vestibular schwannoma. Pulsatile tinnitus requires evaluation of the vascular system of the head to exclude vascular tumors such as glomus jugulare tumors, aneurysms, dural arteriovenous fistulas, and stenotic arterial lesions; it may also occur with SOM, superior semicircular dehiscence, and inner ear dehiscence. It is most commonly associated with some abnormality of the jugular bulb such as a large jugular bulb or jugular bulb diverticulum.

GENETIC CAUSES OF HEARING LOSS

 More than half of childhood hearing impairment is thought to be hereditary; hereditary hearing impairment (HHI) can also manifest later in life. HHI may be classified as either nonsyndromic, when hearing loss is the only clinical abnormality, or syndromic, when hearing loss is associated with anomalies in other organ systems. Nearly two-thirds of HHIs are nonsyndromic. Between 70 and 80% of nonsyndromic HHI is inherited in an autosomal recessive manner and designated DFNB; another 15–20% is autosomal dominant (DFNA). Less than 5% is X-linked (DFNX) or maternally inherited via the mitochondria.

More than 150 loci harboring genes for nonsyndromic HHI have been mapped, with recessive loci outnumbering dominant ones; numerous genes have now been identified **(Table 30-1)**. The hearing genes fall into the categories of structural proteins (*MYH9, MYO7A, MYO15, TECTA, DIAPH1*), transcription factors (*POU3F4, POU4F3*), ion channels (*KCNQ4, SLC26A4*), and gap junction proteins (*GJB2, GJB3, GJB6*). Several of these genes, including *GJB2, TECTA*, and *TMC1*, cause both autosomal dominant and recessive forms of nonsyndromic HHI. In general, the hearing loss associated with dominant genes has its onset in adolescence or adulthood, varies in severity, and progresses with age, whereas the hearing loss associated with recessive inheritance is congenital and profound. Connexin 26, product of the *GJB2* gene, is particularly important because it is responsible for nearly 20% of all cases of childhood deafness; half of genetic deafness in children is *GJB2*-related. Two frameshift mutations, 35delG and 167delT, account for >50% of the cases; however, screening for these two mutations alone is insufficient, and sequencing of the entire gene is required to fully capture *GJB2*-related recessive deafness. The 167delT mutation is highly prevalent in Ashkenazi Jews; ~1 in 1765 individuals in this population are homozygous and affected. *GJB2* hearing loss can also vary among the members of the same family, suggesting that other genes or factors influence the auditory phenotype. A single mutation in *GJB2* in combination with a single mutation in *GJB6* (connexin 30) can also lead to hearing loss and is an example of digenic inheritance of hearing loss.

TABLE 30-1

Hereditary Hearing Impairment Genes

Designation	Gene	Function
Autosomal Dominant		
	CRYM	Thyroid hormone–binding protein
DFNA1	*DIAPH1*	Cytoskeletal protein
DFNA2A	*KCNQ4*	Potassium channel
DFNA2B	*GJB3 (Cx31)*	Gap junction
DFNA3A	*GJB2 (Cx26)*	Gap junction
DFNA3B	*GJB6 (Cx30)*	Gap junction
DFNA4	*MYH14*	Class II nonmuscle myosin
	CEACAM16	Cell adhesion molecule
DFNA5	*DFNA5*	Unknown
DFNA6/14/38	*WFS1*	Transmembrane protein
DFNA8/12	*TECTA*	Tectorial membrane protein
DFNA9	*COCH*	Unknown
DFNA10	*EYA4*	Developmental gene
DFNA11	*MYO7A*	Cytoskeletal protein
DFNA13	*COL11A2*	Cytoskeletal protein
DFNA15	*POU4F3*	Transcription factor
DFNA17	*MYH9*	Cytoskeletal protein
DFNA20/26	*ACTG1*	Cytoskeletal protein
DFNA22	*MYO6*	Unconventional myosin
DFNA23	*SIX1*	Developmental gene
DFNA25	*SLC17A8*	Vesicular glutamate transporter
DFNA28	*GRHL2*	Transcription factor
DFNA36	*TMC1*	Transmembrane protein
DFNA41	*P2RX2*	Purinergic receptor
DFNA44	*CCDC50*	Effector of epidermal growth factor–mediated signaling
DFNA50	*MIRN96*	MicroRNA

Designation	Gene	Function
DFNA51	*TJP2*	Tight junction protein
DFNA56	*TNC*	Extracellular matrix protein
DFNA64	*SMAC/DIABLO*	Mitochondrial proapoptotic protein
DFNA65	*TBC1D24*	ARF6-interacting protein
DFNA66	*CD164*	Sialomucin
DFNA67	*OSBPL2*	Intracellular lipid receptor
DFNA68	*HOMER2*	Stereociliary scaffolding protein
DFNA69	*KITLG*	Ligand for KIT receptor
DFNA70	*MCM2*	Initiation and elongation during DNA replication
DFNA71	*DMXL2*	Regulator of Notch signaling
Autosomal Recessive		
DFNB1A	*GJB2 (CX26)*	Gap junction
DFNB1B	*GJB6 (CX30)*	Gap junction
DFNB2	*MYO7A*	Cytoskeletal protein
DFNB3	*MYO15*	Cytoskeletal protein
DFNB4	*PDS (SLC26A4)*	Chloride/iodide transporter
DFNB6	*TMIE*	Transmembrane protein
DFNB7/B11	*TMC1*	Transmembrane protein
DFNB9	*OTOF*	Trafficking of membrane vesicles
DFNB8/10	*TMPRSS3*	Transmembrane serine protease
DFNB12	*CDH23*	Intercellular adherence protein
DFNB15/72/95	*GIPC3*	PDZ domain–containing protein
DFNB16	*STRC*	Stereocilia protein
DFNB18	*USH1C*	Unknown
DFNB18B	*OTOG*	Tectorial membrane protein
DFNB21	*TECTA*	Tectorial membrane protein
DFNB22	*OTOA*	Gel attachment to nonsensory cell
DFNB23	*PCDH15*	Morphogenesis and cohesion
DFNB24	*RDX*	Cytoskeletal protein

Designation	Gene	Function
DFNB25	*GRXCR1*	Reversible *S*-glutathionylation of proteins
DFNB28	*TRIOBP*	Cytoskeletal-organizing protein
DFNB29	*CLDN14*	Tight junctions
DFNB30	*MYO3A*	Hybrid motor-signaling myosin
DFNB31	*WHRN*	PDZ domain–containing protein
DFNB35	*ESRRB*	Estrogen-related receptor beta protein
DFNB36	*ESPN*	Ca-insensitive actin-bundling protein
DFNB37	*MYO6*	Unconventional myosin
DFNB39	*HFG*	Hepatocyte growth factor
DFNB42	*ILDR1*	Ig-like domain–containing receptor
DFNB44	*ADCY1*	Adenylate cyclase
DFNB48	*CIB2*	Calcium and integrin binding protein
DFNB49	*BDP1*	Subunit of RNA polymerase
DFNB49	*MARVELD2*	Tight junction protein
DFNB53	*COL11A2*	Collagen protein
DFNB59	*PJVK*	Zn-binding protein
DFNB60	*SLC22A4*	Prestin, motor protein of cochlear outer hair cell
DFNB61	*SLC26A5*	Motor protein
DFNB63	*LRTOMT/COMT2*	Putative methyltransferase
DFNB66	*DCDC2*	Ciliary protein
DFNB66/67	*LHFPL5*	Tetraspan protein
DFNB68	*S1PR2*	Tetraspan membrane protein of hair cell stereocilia
DFNB70	*PNPT1*	Mitochondrial–RNA–import protein
DFNB73	*BSND*	Beta subunit of chloride channel
DFNB74	*MSRB3*	Methionine sulfoxide reductase
DFNB76	*SYNE4*	Part of *LINC* tethering complex
DFNB77	*LOXHD1*	Stereociliary protein
DFNB79	*TPRN*	Unknown
DFNB82	*GPSM2*	G protein signaling modulator

Designation	Gene	Function
DFNB84	*PTPRQ*	Type III receptor-like protein-tyrosine phosphatase family
DFNB84	*OTOGL*	Otogelin-like protein
DFNB86	*TBC1D24*	GTPase-activating protein
DFNB88	*ELMOD3*	GTPase-activating protein
DFNB89	*KARS*	Lysyl-tRNA synthetase
DFNB91	*SERPINB6*	Protease inhibitor
DFNB93	*CABP2*	Calcium-binding protein
DFNA97	*MET*	Oncogene/hepatocyte growth factor receptor
DFNB98	*TSPEAR*	Epilepsy-associated repeats containing protein
DFNB99	*TMEM132E*	Transmembrane protein
DFNB101	*GRXCR2*	Maintaining stereocilia bundles
DFNB102	*EPS8*	Epidermal growth factor receptor
DFNB103	*CLIC5*	Chloride ion transport
DFNB105	*CDC14A*	Protein phosphatase involved in hair cell ciliogenesis
	FAM65B	Membrane-associated protein in stereocilia
	EPS8L2	Actin remodeling in response to *EGF* stimulation
	ROR1	Receptor tyrosine kinase-like orphan receptor

In addition to *GJB2*, several other nonsyndromic genes are associated with hearing loss that progresses with age. The contribution of genetics to presbycusis is also becoming better understood. Sensitivity to aminoglycoside ototoxicity can be maternally transmitted through a mitochondrial mutation. Susceptibility to noise-induced hearing loss may also be genetically determined.

There are >400 syndromic forms of hearing loss. These include Usher's syndrome (retinitis pigmentosa and hearing loss), Waardenburg's syndrome (pigmentary abnormality and hearing loss), Pendred's syndrome (thyroid organification defect and hearing loss), Alport's syndrome (renal disease and hearing loss), Jervell and Lange-Nielsen syndrome (prolonged QT interval and hearing loss), neurofibromatosis type 2 (bilateral acoustic schwannoma), and mitochondrial disorders (mitochondrial encephalopathy, lactic acidosis, and stroke-like episodes [MELAS]; myoclonic epilepsy and ragged red fibers [MERRF]; and progressive external ophthalmoplegia [PEO]) **(Table 30-2)**.

TABLE 30-2

Syndromic Hereditary Hearing Impairment Genes

Syndrome	Gene	Function
Alport's syndrome	*COL4A3-5*	Cytoskeletal protein
BOR syndrome	*EYA1*	Developmental gene
	SIX5	Developmental gene
	SIX1	Developmental gene
Jervell and Lange-Nielsen syndrome	*KCNQ1*	Delayed rectifier K^+ channel
	KCNE1	Delayed rectifier K^+ channel
Norrie's disease	*NDP*	Cell–cell interactions
Pendred's syndrome	*SLC26A4*	Chloride/iodide transporter
	FOXI1	Transcriptional activator of *SLC26A4*
	KCNJ10	Inwardly rectifying K^+ channel
Treacher Collins syndrome	*TCOF1*	Nucleolar-cytoplasmic transport
	POLR1D	Subunit of RNA polymerases I and III
	POLR1C	Subunit of RNA polymerases I and III
Usher's syndrome	*MYO7A*	Cytoskeletal protein
	USH1C	Unknown
	CDH23	Intercellular adherence protein
	PCDH15	Cell adhesion molecule
	SANS	Harmonin-associated protein
	CIB2	Calcium- and integrin-binding protein
	USH2A	Cell adhesion molecule
	VLGR1	G protein–coupled receptor
	WHRN	PDZ domain–containing protein
	CLRN1	Cellular synapse protein
	HARS	Histidyl-tRNA synthetase
	PDZD7	PDZ domain–containing protein
WS type I, III	*PAX3*	Transcription factor
WS type II	*MITF*	Transcription factor

Syndrome	Gene	Function
	SNAI2	Transcription factor
WS type IV	*EDNRB*	Endothelin B receptor
	EDN3	Endothelin B receptor ligand
	SOX10	Transcription factor

Abbreviations: BOR, branchio-oto-renal syndrome; WS, Waardenburg's syndrome.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Disorders of the Sense of Hearing

The goal in the evaluation of a patient with auditory complaints is to determine (1) the nature of the hearing impairment (conductive vs sensorineural vs mixed), (2) the severity of the impairment (mild, moderate, severe, or profound), (3) the anatomy of the impairment (external ear, middle ear, inner ear, or central auditory pathway), and (4) the etiology. The presence of signs and symptoms associated with hearing loss should be ascertained **(Table 30-3)**. The history should elicit characteristics of the hearing loss, including the duration of deafness, unilateral versus bilateral involvement, nature of onset (sudden vs insidious), and rate of progression (rapid vs slow). Symptoms of tinnitus, vertigo, imbalance, aural fullness, otorrhea, headache, facial nerve dysfunction, and head and neck paresthesias should be noted. Information regarding head trauma, exposure to ototoxins, occupational or recreational noise exposure, and family history of hearing impairment may also be important. A sudden onset of unilateral hearing loss, with or without tinnitus, may represent a viral infection of the inner ear, vestibular schwannoma, or a stroke. Patients with unilateral hearing loss (sensory or conductive) usually complain of reduced hearing, poor sound localization, and difficulty hearing clearly in the presence of background noise. Gradual progression of a hearing deficit is common with otosclerosis, noise-induced hearing loss, vestibular schwannoma, or Ménière's disease. Small vestibular schwannomas typically present with asymmetric hearing impairment, tinnitus, and imbalance (rarely vertigo); cranial neuropathy, in particular of the trigeminal or facial nerve, may accompany larger tumors. In addition to hearing loss, Ménière's disease may be associated with episodic vertigo, tinnitus, and aural fullness. Hearing loss with otorrhea is most likely due to chronic otitis media or cholesteatoma.

Examination should include the auricle, external ear canal, and tympanic membrane. In the elderly, the external ear canal is often dry and fragile; it is preferable to clean cerumen with wall-mounted suction or cerumen loops and to avoid irrigation. In examining the eardrum, the topography of the tympanic membrane is more important than the presence or absence of the light reflex. In addition to the pars tensa (the lower two-thirds of the tympanic membrane), the pars flaccida (upper one-third of the tympanic membrane) above the short process of the malleus should also be examined for retraction pockets that may be evidence of chronic eustachian tube dysfunction or cholesteatoma. Insufflation of the ear canal is necessary to assess tympanic membrane mobility and compliance. Careful inspection of the nose, nasopharynx, and upper respiratory tract is important. Unilateral serous effusion or unexplained otalgia should prompt a fiberoptic examination of the nasopharynx and larynx to exclude neoplasms. Cranial nerves should be evaluated with special attention to facial and trigeminal nerves, which are commonly affected with tumors involving the cerebellopontine angle.

The Rinne and Weber tuning fork tests, with a 512-Hz tuning fork, are used to screen for hearing loss, differentiate conductive from sensorineural hearing losses, and confirm the findings of audiologic evaluation. The Rinne test compares the ability to hear by air conduction with the ability to hear by bone conduction. The tines of a vibrating tuning fork are held near the opening of the external auditory canal, and then the stem is placed on the mastoid process; for direct contact, it may be placed on teeth or dentures. The patient is asked to indicate whether the tone is louder by air conduction or bone conduction. Normally, and in the presence of sensorineural hearing loss, a tone is heard louder by air conduction than by bone conduction; however, with conductive hearing loss of ≥30 dB (see "Audiologic Assessment," below), the bone-conduction stimulus is perceived as louder than the air-conduction stimulus. For the Weber test, the stem of a vibrating tuning fork is placed on the head in the midline and the patient is asked whether the tone is heard in both ears or better in one ear than in the other. With a unilateral conductive hearing loss, the tone is perceived in the affected ear. With a unilateral sensorineural hearing loss, the tone is perceived in the unaffected ear. A 5-dB difference in hearing between the two ears is required for lateralization.

TABLE 30-3

Signs and Symptoms Suggestive of Hearing Loss

Saying “huh” a great deal
Reduced clarity of hearing
Difficulty understanding conversations in background noise
Family complaining of hearing loss
Tinnitus
Turning the volume up on radio or television
Sensitivity to noises
Fullness in the ear
Avoiding social settings

LABORATORY ASSESSMENT OF HEARING

Audiologic Assessment

The minimum audiologic assessment for hearing loss should include the measurement of pure tone air-conduction and bone-conduction thresholds, speech reception threshold, word recognition score, tympanometry, acoustic reflexes, and acoustic-reflex decay. This test battery provides a screening evaluation of the entire auditory system and allows one to determine whether further differentiation of a sensory (cochlear) from a neural (retrocochlear) hearing loss is indicated.

Pure tone audiometry assesses hearing acuity for pure tones. The test is administered by an audiologist and is performed in a sound-attenuated chamber. The pure tone stimulus is delivered with an audiometer, an electronic device that allows the presentation of specific frequencies (generally between 250 and 8000 Hz) at specific intensities. Air- and bone-conduction thresholds are established for each ear. Air-conduction thresholds are determined by presenting the stimulus in air with the use of headphones. Bone-conduction thresholds are determined by placing the stem of a vibrating tuning fork or an oscillator of an audiometer in contact with the head. In the presence of a hearing loss, broad-spectrum noise is presented to the nontest ear for *masking* purposes so that responses are based on perception from the ear under test.

The responses are measured in decibels (dBs). An *audiogram* is a plot of intensity in dBs of hearing threshold versus frequency. A dB is equal to 20 times the logarithm of the ratio of the sound pressure required to achieve threshold in the patient to the sound pressure required to achieve threshold in a normal-hearing person. Therefore, a change of 6 dB represents doubling of sound pressure, and a change of 20 dB represents a tenfold change in sound pressure. Loudness, which depends on the frequency, intensity, and duration of a sound, doubles with approximately each 10-dB increase in sound pressure level. Pitch, on the other hand, does not directly correlate with frequency. The perception of pitch changes slowly in the low and high frequencies. In the middle tones, which are important for human speech, pitch varies more rapidly with changes in frequency.

Pure tone audiometry establishes the presence and severity of hearing impairment, unilateral versus bilateral involvement, and the type of hearing loss. Conductive hearing losses with a large mass component, as is often seen in middle ear effusions, produce elevation of thresholds that predominate in the higher frequencies. Conductive hearing losses with a large stiffness component, as in fixation of the footplate of the stapes in early otosclerosis, produce threshold elevations in the lower frequencies. Often, the conductive hearing loss involves all frequencies, suggesting involvement of both stiffness and mass. In general, sensorineural hearing losses such as presbycusis affect higher frequencies more than lower frequencies (Fig. 30-3). An exception is Ménière’s disease, which is characteristically associated with low-frequency sensorineural hearing loss (though any frequency can be affected). Noise-induced hearing loss has an unusual pattern of hearing impairment in which the loss at 4000 Hz is greater than at higher frequencies. Vestibular schwannomas characteristically affect the higher frequencies, but any pattern of hearing loss can be observed.

Speech recognition requires greater synchronous neural firing than is necessary for appreciation of pure tones. *Speech audiometry* tests the clarity with which one hears. The *speech reception threshold* (SRT) is defined as the intensity at which speech is recognized as a meaningful symbol and is obtained by presenting two-syllable words with an equal accent on each syllable. The intensity at which the patient can repeat 50% of the words correctly is the SRT. Once the SRT is determined, discrimination or word recognition ability is tested by presenting one-syllable words at 25–40 dB above the SRT. The

words are phonetically balanced in that the phonemes (speech sounds) occur in the list of words at the same frequency that they occur in ordinary conversational English. An individual with normal hearing or conductive hearing loss can repeat 88–100% of the phonetically balanced words correctly. Patients with a sensorineural hearing loss have variable loss of discrimination. As a general rule, neural lesions produce greater deficits in discrimination than do cochlear lesions. For example, in a patient with mild asymmetric sensorineural hearing loss, a clue to the diagnosis of vestibular schwannoma is the presence of greater than expected deterioration in discrimination ability. Deterioration in discrimination ability at higher intensities above the SRT also suggests a lesion in the eighth nerve or central auditory pathways.

Tympanometry measures the impedance of the middle ear to sound and is useful in diagnosis of middle ear effusions. A *tympanogram* is the graphic representation of change in impedance or compliance as the pressure in the ear canal is changed. Normally, the middle ear is most compliant at atmospheric pressure, and the compliance decreases as the pressure is increased or decreased (type A); this pattern is seen with normal hearing or in the presence of sensorineural hearing loss. Compliance that does not change with change in pressure suggests middle ear effusion (type B). With a negative pressure in the middle ear, as with eustachian tube obstruction, the point of maximal compliance occurs with negative pressure in the ear canal (type C). A tympanogram in which no point of maximal compliance can be obtained is most commonly seen with discontinuity of the ossicular chain (type A_d). A reduction in the maximal compliance peak can be seen in otosclerosis (type A_s).

During tympanometry, an intense tone elicits contraction of the stapedius muscle. The change in compliance of the middle ear with contraction of the stapedius muscle can be detected. The presence or absence of this *acoustic reflex* is important in determining the etiology of hearing loss as well as in the anatomic localization of facial nerve paralysis. The acoustic reflex can help differentiate between conductive hearing loss due to otosclerosis and that caused by an inner ear “third window”: it is absent in otosclerosis and present in inner ear conductive hearing loss. Normal or elevated acoustic reflex thresholds in an individual with sensorineural hearing impairment suggest a cochlear hearing loss. An absent acoustic reflex in the setting of sensorineural hearing loss is not helpful in localizing the site of lesion. Assessment of *acoustic reflex decay* helps differentiate sensory from neural hearing losses. In neural hearing loss, such as with vestibular schwannoma, the reflex adapts or decays with time.

OAEs generated by outer hair cells only can be measured with microphones inserted into the external auditory canal. The emissions may be spontaneous or evoked with sound stimulation. The presence of OAEs indicates that the outer hair cells of the organ of Corti are intact and can be used to assess auditory thresholds and to distinguish sensory from neural hearing losses.

Evoked Responses

Electrocochleography measures the earliest evoked potentials generated in the cochlea and the auditory nerve. Receptor potentials recorded include the cochlear microphonic, generated by the outer hair cells of the organ of Corti, and the summating potential, generated by the inner hair cells in response to sound. The whole nerve action potential representing the composite firing of the first-order neurons can also be recorded during electrocochleography. Clinically, the test is useful in the diagnosis of Ménière’s disease, in which an elevation of the ratio of summating potential to action potential is seen.

Brainstem auditory-evoked responses (BAERs), also known as (ABRs), are useful in differentiating the site of sensorineural hearing loss. In response to sound, five distinct electrical potentials arising from different stations along the peripheral and central auditory pathway (eighth nerve, cochlear nucleus, superior olivary complex, lateral lemniscus, and inferior colliculus) can be identified using computer averaging from scalp surface electrodes. BAERs are valuable in situations in which patients cannot or will not give reliable voluntary thresholds. They are also used to assess the integrity of the auditory nerve and brainstem in various clinical situations, including intraoperative monitoring, and in determination of brain death.

The *VEMP test* investigates otolith and vestibular nerve function by presenting a high-level acoustic stimuli and evoking a short-latency electromyographic potential; cVEMP (or cervical VEMP) and oVEMP (or ocular VEMP) have been described. The cVEMP elicits a vestibulocollic reflex whose afferent limb arises from acoustically sensitive cells in the saccule, with signals conducted via the inferior vestibular nerve. cVEMP is a biphasic, short-latency response recorded from the tonically contracted sternocleidomastoid muscle in response to loud auditory clicks or tones. cVEMPs may be diminished or absent in patients with early and late Ménière’s disease, vestibular neuritis, benign paroxysmal positional vertigo, and vestibular schwannoma. On the other hand, the threshold for VEMPs may be lower in cases of superior canal dehiscence, other inner ear dehiscence, and perilymphatic fistula. The oVEMP, in contrast, is a response involving the utricle primarily and superior vestibular nerve. The oVEMP excitatory response is recorded from the extraocular muscle. The oVEMP is abnormal in superior vestibular neuritis.

Imaging Studies

The choice of radiologic tests is largely determined by whether the goal is to evaluate the bony anatomy of the external, middle, and inner ear or to image the auditory nerve and brain. Axial and coronal CT of the temporal bone with fine 0.3-mm cuts is ideal for determining the caliber of the external auditory canal, integrity of the ossicular chain, and presence of middle ear or mastoid disease; it can also detect inner ear malformations. CT is also ideal for the detection of bone erosion with chronic otitis media and cholesteatoma. Pöschl reformatting in the plane of the superior semicircular canal is required for the identification of dehiscence or absence of bone over the superior semicircular canal. MRI is superior to CT for imaging of retrocochlear pathology such as vestibular schwannoma, meningioma, other lesions of the cerebellopontine angle, demyelinating lesions of the

brainstem, and brain tumors. Both CT and MRI are equally capable of identifying inner ear malformations and assessing cochlear patency for preoperative evaluation of patients for cochlear implantation.

TREATMENT

TREATMENT

Disorders of the Sense of Hearing

In general, conductive hearing losses are amenable to surgical correction, whereas sensorineural hearing losses are usually managed medically. Atresia of the ear canal can be surgically repaired, often with significant improvement in hearing. Alternatively, the conductive hearing loss associated with atresia can be addressed with a bone-anchored hearing aid (BAHA). Tympanic membrane perforations due to chronic otitis media or trauma can be repaired with an outpatient tympanoplasty. Likewise, conductive hearing loss associated with otosclerosis can be treated by stapedectomy, which is successful in >95% of cases. Tympanostomy tubes allow the prompt return of normal hearing in individuals with middle ear effusions. Hearing aids are effective and well tolerated in patients with conductive hearing losses.

Patients with mild, moderate, and severe sensorineural hearing losses are regularly rehabilitated with hearing aids of varying configuration and strength. Hearing aids have been improved to provide greater fidelity and have been miniaturized. The current generation of hearing aids can be placed entirely within the ear canal, thus reducing any stigma associated with their use. In general, the more severe the hearing impairment, the larger the hearing aid required for auditory rehabilitation. Digital hearing aids lend themselves to individual programming, and multiple and directional microphones at the ear level may be helpful in noisy surroundings. Because all hearing aids amplify noise as well as speech, the only absolute solution to the problem of noise is to place the microphone closer to the speaker than the noise source. This arrangement is not possible with a self-contained, cosmetically acceptable device. A significant limitation of rehabilitation with a hearing aid is that although it is able to enhance detection of sound with amplification, it cannot restore clarity of hearing that is lost with presbycusis.

The cost of a single hearing aid (~$2300 US) is a significant obstacle for many hearing-impaired individuals and usually bilateral amplification is recommended. To reduce cost and spur innovation, efforts are underway to create a new category for “basic” hearing aids that could be sold over-the-counter, similar to some eyeglasses and contact lenses. By reducing the cost of hearing aids to consumers, promoting innovation, and increasing competition, this new class of devices could fundamentally change the way hearing rehabilitation is delivered.

Patients with unilateral deafness have difficulty with sound localization and reduced clarity of hearing in background noise. They may benefit from a contralateral routing of signal (CROS) hearing aid in which a microphone is placed on the hearing-impaired side, and the sound is transmitted to the receiver placed on the contralateral ear. The same result may be obtained with a BAHA, in which a hearing aid clamps to a screw integrated into the skull on the hearing-impaired side. Like the CROS hearing aid, the BAHA transfers the acoustic signal to the contralateral hearing ear, but it does so by vibrating the skull. Patients with profound deafness on one side and some hearing loss in the better ear are candidates for a BICROS hearing aid; it differs from the CROS hearing aid in that the patient wears a hearing aid, and not simply a receiver, in the better ear. Unfortunately, while CROS and BAHA devices provide benefit, they do not restore hearing in the deaf ear. Only cochlear implants can restore hearing (see below). Increasingly, cochlear implants are being investigated for the treatment of patients with single-sided deafness; early reports show great promise in not only restoring hearing and reducing tinnitus, but also improving sound localization and performance in background noise.

In many situations, including lectures and the theater, hearing-impaired persons benefit from assistive devices that are based on the principle of having the speaker closer to the microphone than any source of noise. Assistive devices include infrared and frequency-modulated (FM) transmission as well as an electromagnetic loop around the room for transmission to the individual’s hearing aid. Hearing aids with telecoils can also be used with properly equipped telephones in the same way.

In the event that the hearing aid provides inadequate rehabilitation, cochlear implants may be appropriate (Fig. 30-4). Criteria for implantation include severe to profound hearing loss with open-set sentence cognition of ≤40% under best-aided conditions. Worldwide, >600,000 hearing-impaired individuals have received cochlear implants. Cochlear implants are neural prostheses that convert sound energy to electrical energy and can be used to stimulate the auditory division of the eighth nerve directly. In most cases of profound hearing impairment, the auditory hair cells are lost but the ganglionic cells of the auditory division of the eighth nerve are preserved. Cochlear implants consist of electrodes that are inserted into the cochlea through the round window, speech processors that extract acoustical elements of speech for conversion to electrical currents, and a means of transmitting the electrical energy through the skin. Patients with implants experience sound that helps with speech reading, allows open-set word recognition, and helps in modulating the person’s own voice. Usually, within the first 3–6 months after implantation, adult patients can understand speech without visual cues. With the current generation of multichannel cochlear implants, nearly 75% of patients are able to converse on the telephone. Bilateral cochlear implantations are increasingly being performed, especially in children; these patients perform better in background noise, have better sound localization, and are less fatigued by the “work” compared to monaural hearing.

The first hybrid cochlear implant for the treatment of high-frequency hearing loss has now been approved by the U.S. Food and Drug Administration. Patients with presbyacusis typically have normal low-frequency hearing, while suffering from high-frequency hearing loss associated with loss of clarity

that cannot always be adequately rehabilitated with a hearing aid. However, these patients are not candidates for conventional cochlear implants because they have too much residual hearing. The hybrid implant has been specifically designed for this patient population; it has a shorter electrode than a conventional cochlear implant and can be introduced into the cochlea atraumatically, thus preserving low-frequency hearing. Individuals with a hybrid implant use their own natural low-frequency “acoustic” hearing and rely on the implant for providing “electrical” high-frequency hearing. Patients who have received the hybrid implant perform better on speech discrimination tests in both quiet and noisy backgrounds.

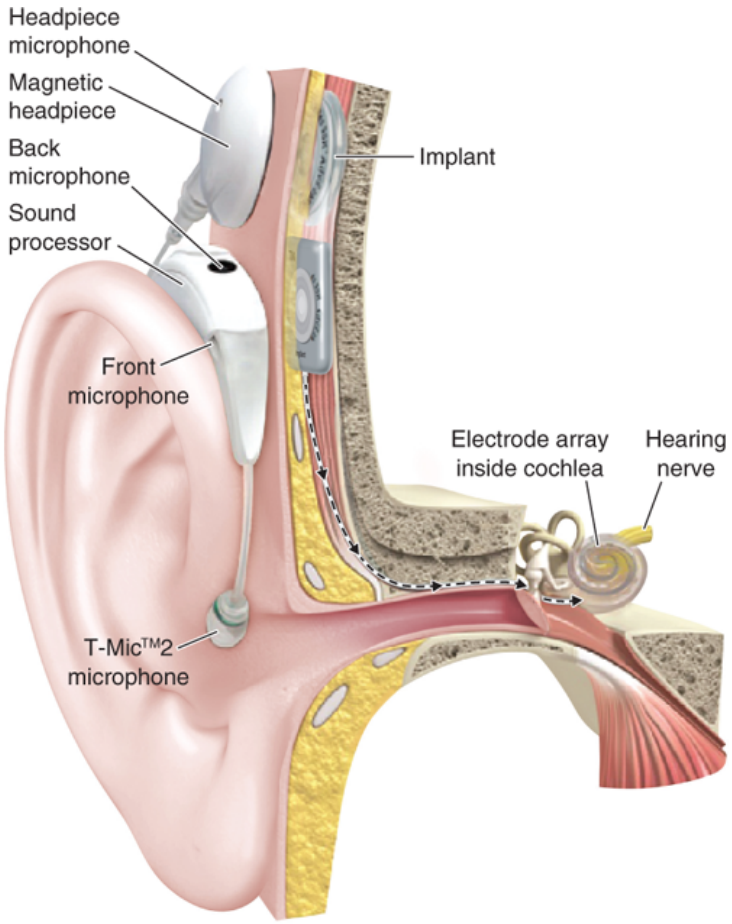
For individuals who have had both eighth nerves destroyed by trauma or bilateral vestibular schwannomas (e.g., neurofibromatosis type 2), brainstem auditory implants placed near the cochlear nucleus may provide auditory rehabilitation. Currently, brainstem implants provide sound awareness but unfortunately speech understanding remains elusive.

Tinnitus often accompanies hearing loss. As for background noise, tinnitus can degrade speech comprehension in individuals with hearing impairment. Patients with tinnitus should be advised to minimize caffeine ingestion, avoid high dosage of nonsteroidal anti-inflammatory drugs (NSAIDs), and reduce stress. Therapy for tinnitus is usually directed toward minimizing the appreciation of tinnitus. Relief of the tinnitus may be obtained by masking it with background music. Hearing aids are also helpful in tinnitus suppression, as are tinnitus maskers, devices that present a sound to the affected ear that is more pleasant to listen to than the tinnitus. The use of a tinnitus masker is often followed by several hours of inhibition of the tinnitus. Antidepressants have also been shown to be beneficial in helping patients cope with tinnitus.

Hard-of-hearing individuals often benefit from a reduction in unnecessary noise in the environment (e.g., radio or television) to enhance the signal-to-noise ratio. Speech comprehension is aided by lip reading; therefore, the impaired listener should be seated so that the face of the speaker is well illuminated and easily seen. Although speech should be in a loud, clear voice, one should be aware that in sensorineural hearing losses in general and in hard-of-hearing elderly in particular, recruitment (abnormal perception of loud sounds) may be troublesome. Above all, optimal communication cannot take place without both parties giving it their full and undivided attention.

FIGURE 30-4

A cochlear implant is composed of an external microphone and speech processor worn on the ear and a receiver implanted underneath the temporalis muscle. The internal receiver is attached to an electrode that is placed surgically in the cochlea.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PREVENTION

Conductive hearing losses may be prevented by prompt antibiotic therapy of adequate duration for AOM and by ventilation of the middle ear with tympanostomy tubes in middle ear effusions lasting ≥12 weeks. Loss of vestibular function and deafness due to aminoglycoside antibiotics can largely be prevented by careful monitoring of serum peak and trough levels.

Some 10 million Americans have noise-induced hearing loss, and 20 million are exposed to hazardous noise in their employment. Noise-induced hearing loss can be prevented by avoidance of exposure to loud noise or by regular use of ear plugs or fluid-filled ear muffs to attenuate intense sound. Table 30-4 lists loudness levels for a variety of environmental sounds. High-risk activities for noise-induced hearing loss include use of electrical equipment for wood and metal working and target practice or hunting with small firearms. All internal-combustion and electric engines, including snow and leaf blowers, snowmobiles, outboard motors, and chainsaws, require protection of the user with hearing protectors. Virtually all noise-induced hearing loss is preventable through education, which should begin before the teenage years. Programs for conservation of hearing in the workplace are required by the Occupational Safety and Health Administration (OSHA) whenever the exposure over an 8-h period averages 85 dB. OSHA mandates that workers in such noisy environments have hearing monitoring and protection programs that include a preemployment screen, an annual audiologic assessment, and the mandatory use of hearing protectors. Exposure to loud sounds above 85 dB in the work environment is restricted by OSHA, with halving of allowed exposure time for each increment of 5 dB above this threshold; for example, exposure to 90 dB is permitted for 8 h; 95 dB for 4 h, and 100 dB for 2 h (Table 30-5).

TABLE 30-4

Decibel (Loudness) Level of Common Environmental Noise

Source	Decibel (dB)
Weakest sound heard	0
Whisper	30
Normal conversation	55–65
City traffic inside car	85
OSHA Monitoring Requirement Begins	**90**
Jackhammer	95
Subway train at 200 ft	95
Power mower	107
Power saw	110
Painful Sound	**125**
Jet engine at 100 feet	140
12-gauge shotgun blast	165
Loudest sound that can occur	194

Abbreviation: OSHA, Occupational Safety and Health Administration.

TABLE 30-5

OSHA Daily Permissible Noise Level Exposure

Sound Level (dB)	Duration per Day (h)
90	8
92	6
95	4
97	3
100	2
102	1.5
105	1
110	0.5
115	≤0.25

Note: Exposure to impulsive or impact noise should not exceed 140-dB peak sound pressure level.

Source: From *https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=standards&p_id=9735*.

FURTHER READING

Espinosa-Sanchez JM, Lopez-Escamez JA: Menière's disease. Handb Clin Neurol 137:257, 2016. [PubMed: 27638077]

Moser T, Starr A: Auditory neuropathy—neural and synaptic mechanisms. Nat Rev Neurol 12:135, 2016. [PubMed: 26891769]

Patel M et al: Intratympanic methylprednisolone versus gentamicin in patients with unilateral Ménière's disease: A randomised, double-blind, comparative effectiveness trial. Lancet 388:2753, 2016. [PubMed: 27865535]

Tikka C et al: Interventions to prevent occupational noise-induced hearing loss. Cochrane Database Syst Rev 7:CD006396, 2017. [PubMed: 28685503]

Wilson BS et al: Global hearing health care: New findings and perspectives. Lancet 390:2503, 2017.

Chapter 31: Sore Throat, Earache, and Upper Respiratory Symptoms

Michael A. Rubin; Larry C. Ford; Ralph Gonzales

INTRODUCTION

Infections of the upper respiratory tract (URIs) have a tremendous impact on public health. They are among the most common reasons for visits to primary care providers, and although the illnesses are typically mild, their high incidence and transmission rates place them among the leading causes of time lost from work or school. Even though a minority (~25%) of cases are caused by bacteria, URIs are the leading diagnoses for which antibiotics are prescribed on an outpatient basis in the United States, often inappropriately. Antibiotics are more often misprescribed in adults than in pediatric populations. The enormous consumption of antibiotics for these illnesses has contributed to the rise in antibiotic resistance among common community-acquired pathogens such as *Streptococcus pneumoniae*—a trend that in itself has an enormous influence on public health and on the individual patient.

Although most URIs are caused by viruses, distinguishing patients with primary viral infection from those with primary bacterial infection is difficult. Signs and symptoms of bacterial and viral URIs are typically indistinguishable. Until consistent, inexpensive, and rapid testing becomes available and is used widely, acute infections will be diagnosed largely on clinical grounds. The judicious use and potential for misuse of antibiotics in this setting pose ongoing challenges.

NONSPECIFIC INFECTIONS OF THE UPPER RESPIRATORY TRACT

Nonspecific URIs are a broadly defined group of disorders that collectively constitute the leading cause of ambulatory care visits in the United States. By definition, nonspecific URIs have no prominent localizing features. They are identified by a variety of descriptive names, including *acute infective rhinitis*, *acute rhinopharyngitis/nasopharyngitis*, *acute coryza*, and *acute nasal catarrh*, as well as by the inclusive label *common cold*.

ETIOLOGY

The large assortment of URI classifications reflects the wide variety of causative infectious agents and the varied manifestations of common pathogens. Nearly all nonspecific URIs are caused by viruses spanning multiple virus families and many antigenic types. For instance, there are at least 100 immunotypes of

rhinovirus (Chap. 194), the most common cause of URI (~30–40% of cases); other causes include influenza virus (three immunotypes; Chap. 195) as well as parainfluenza virus (four immunotypes), coronavirus (at least three immunotypes), and adenovirus (47 immunotypes) (Chap. 194). Respiratory syncytial virus (RSV), a well-established pathogen in pediatric populations, is also a recognized cause of significant disease in elderly and immunocompromised individuals. A host of additional viruses, including some viruses not typically associated with URIs (e.g., enteroviruses, rubella virus, and varicella-zoster virus), account for a small percentage of cases in adults each year. Although new diagnostic modalities (e.g., nasopharyngeal swab for polymerase chain reaction) can assign a viral etiology, there are few specific treatment options, and no pathogen is identified in a substantial proportion of cases. A specific diagnostic workup beyond a clinical diagnosis is generally unnecessary in an otherwise healthy adult.

CLINICAL MANIFESTATIONS

The signs and symptoms of nonspecific URI are similar to those of other URIs but lack a pronounced localization to one particular anatomic location, such as the sinuses, pharynx, or lower airway. Nonspecific URI commonly presents as an acute, mild, and self-limited catarrhal syndrome with a median duration of ~1 week (range, 2–10 days). Signs and symptoms are diverse and frequently variable across patients, even when caused by the same virus. The principal signs and symptoms of nonspecific URI include rhinorrhea (with or without purulence), nasal congestion, cough, and sore throat. Other manifestations, such as fever, malaise, sneezing, lymphadenopathy, and hoarseness, are more variable, with fever more common among infants and young children. This varying presentation may reflect differences in host response as well as in infecting organisms; myalgias and fatigue, for example, sometimes are seen with influenza and parainfluenza infections, whereas conjunctivitis may suggest infection with adenovirus or enterovirus. Cough secondary to upper respiratory inflammation after such an illness frequently lasts 2–3 weeks and can be misinterpreted as an indication of a process that necessitates antibiotic therapy. Findings on physical examination are frequently nonspecific and unimpressive. Between 0.5 and 2% of colds are complicated by secondary bacterial infections (e.g., rhinosinusitis, otitis media, and pneumonia), particularly in higher-risk populations such as infants, elderly persons, and chronically ill or immunosuppressed individuals. Secondary bacterial infections usually are associated with a prolonged course of illness, increased severity of illness, and localization of signs and symptoms, often as a rebound after initial clinical improvement (the “double-dip” sign). Purulent secretions from the nares or throat often are misinterpreted as an indication of bacterial sinusitis or pharyngitis. These secretions, however, can be seen in nonspecific URI and, in the absence of other clinical features, are poor predictors of bacterial infection.

TREATMENT

TREATMENT

Nonspecific Upper Respiratory Infections

Antibiotics have no role in the treatment of uncomplicated nonspecific URI, and their misuse facilitates the emergence of antimicrobial resistance; in healthy volunteers, a single course of a commonly prescribed

antibiotic like azithromycin can result in macrolide resistance in oral streptococci many months later. In the absence of clinical evidence of bacterial infection, treatment remains entirely symptom based, with use of decongestants and nonsteroidal anti-inflammatory drugs. Clinical trials of zinc, vitamin C, echinacea, and other alternative remedies have revealed no consistent benefit in the treatment of nonspecific URI.

INFECTIONS OF THE SINUS

Rhinosinusitis refers to an inflammatory condition involving the nasal sinuses. Although most cases of sinusitis involve more than one sinus, the maxillary sinus is most commonly involved; next, in order of frequency, are the ethmoid, frontal, and sphenoid sinuses. Each sinus is lined with a respiratory epithelium that produces mucus, which is transported out by ciliary action through the sinus ostium and into the nasal cavity. Normally, mucus does not accumulate in the sinuses, which remain mostly sterile despite their adjacency to the bacterium-filled nasal passages. When the sinus ostia are obstructed or when ciliary clearance is impaired or absent, the secretions can be retained, producing the typical signs and symptoms of sinusitis. As these secretions accumulate with obstruction, they become more susceptible to infection with a variety of pathogens, including viruses, bacteria, and, rarely, fungi. Sinusitis affects a tremendous proportion of the population, accounts for millions of visits to primary care physicians each year, and is the fifth leading diagnosis for which antibiotics are prescribed. It typically is classified by duration of illness (acute vs. chronic); by etiology (infectious vs. noninfectious); and, when infectious, by the offending pathogen type (viral, bacterial, or fungal).

ACUTE RHINOSINUSITIS

Acute rhinosinusitis—defined as sinusitis of <4 weeks' duration—constitutes the vast majority of sinusitis cases. Most cases are diagnosed in the ambulatory care setting and occur primarily as a consequence of a preceding viral URI. Differentiating acute bacterial from viral sinusitis on clinical grounds is difficult. Therefore, it is perhaps not surprising that antibiotics are prescribed frequently (in 85–98% of all cases) for this condition.

Etiology

The ostial obstruction in rhinosinusitis can arise from both infectious and noninfectious causes. Noninfectious etiologies include allergic rhinitis (with either mucosal edema or polyp obstruction), barotrauma (e.g., from deep-sea diving or air travel), and exposure to chemical irritants. Obstruction can also occur with nasal and sinus tumors (e.g., squamous cell carcinoma) or granulomatous diseases (e.g., granulomatosis with polyangiitis, rhinoscleroma), and conditions leading to altered mucus content (e.g., cystic fibrosis) can cause sinusitis through impaired mucus clearance. In intensive care units, nasotracheal intubation and nasogastric tubes are major risk factors for nosocomial sinusitis.

Viral rhinosinusitis is far more common than bacterial sinusitis, although relatively few studies have sampled sinus aspirates for the presence of different viruses. In the studies that have done so, the viruses most commonly isolated—both alone and with bacteria—have been rhinovirus, parainfluenza virus, and influenza

virus. Bacterial causes of sinusitis have been better described. Among community-acquired cases, *S. pneumoniae* and nontypable *Haemophilus influenzae* are the most common pathogens, accounting for 50–60% of cases. *Moraxella catarrhalis* causes disease in a significant percentage (20%) of children but a lesser percentage of adults. Other streptococcal species and *Staphylococcus aureus* cause only a small percentage of cases, although there is increasing concern about methicillin-resistant *S. aureus* (MRSA) as an emerging cause. It is difficult to assess whether a cultured bacterium represents a true infecting organism, an insufficiently deep sample (which would not be expected to be sterile), or—especially in the case of previous sinus surgeries—a colonizing organism. Anaerobes occasionally are found in association with infections of the roots of premolar teeth that spread to the adjacent maxillary sinuses. The role of atypical organisms like *Chlamydia pneumoniae* and *Mycoplasma pneumoniae* in the pathogenesis of acute sinusitis is unclear. Nosocomial cases commonly are associated with bacteria prevalent in the hospital environment, including *S. aureus*, *Pseudomonas aeruginosa*, *Serratia marcescens*, *Klebsiella pneumoniae*, and *Enterobacter* species. Often, these infections are polymicrobial and can involve organisms that are highly resistant to numerous antibiotics. Fungi also are established causes of sinusitis, although most acute cases affect immunocompromised patients and represent invasive, life-threatening infections. The best-known example is rhinocerebral mucormycosis caused by fungi of the order Mucorales, which includes *Rhizopus*, *Rhizomucor*, *Mucor*, *Lichtheimia* (formerly *Mycocladus*, formerly *Absidia*), and *Cunninghamella* (Chap. 213). These infections classically occur in diabetic patients with ketoacidosis but can also develop in transplant recipients, patients with hematologic malignancies, and patients receiving chronic glucocorticoid or deferoxamine therapy. Other hyaline molds, such as *Aspergillus* and *Fusarium* species, also are occasional causes of this disease.

Clinical Manifestations

Most cases of acute sinusitis present after or in conjunction with a viral URI, and it can be difficult to discriminate the clinical features of one from the other, with timing becoming important in diagnosis (see below). A large proportion of patients with colds have sinus inflammation, although true bacterial sinusitis complicates only 0.2–2% of these viral infections. Common presenting symptoms of sinusitis include nasal drainage and congestion, facial pain or pressure, and headache. Thick, purulent or discolored nasal discharge is often thought to indicate bacterial sinusitis but also occurs early in viral infections such as the common cold and is not specific to bacterial infection. Other nonspecific manifestations include cough, sneezing, and fever. Tooth pain, most often involving the upper molars, as well as halitosis are occasionally associated with bacterial sinusitis.

In acute sinusitis, sinus pain or pressure often localizes to the involved sinus (particularly the maxillary sinus) and can be worse when the patient bends over or is supine. Although rare, manifestations of advanced sphenoid or ethmoid sinus infection can be profound, including severe frontal or retroorbital pain radiating to the occiput, thrombosis of the cavernous sinus, and signs of orbital cellulitis. Acute focal sinusitis is uncommon but should be considered with severe symptoms involving the maxillary sinus and fever, regardless of illness duration. This condition is typically associated with red, hot, and swollen sinuses that are extremely tender to palpation; is of staphylococcal etiology; and requires emergent debridement and

initial IV administration of antibiotics. Similarly, patients with advanced frontal sinusitis can present with a condition known as *Pott's puffy tumor*, with soft-tissue swelling and pitting edema over the frontal bone from a communicating subperiosteal abscess. Life-threatening complications of sinusitis are rare but include meningitis, epidural abscess, and cerebral abscess.

Patients with acute fungal rhinosinusitis (such as mucormycosis; Chap. 213) often present with symptoms related to pressure effects, particularly when the infection has spread to the orbits and cavernous sinus. Signs such as orbital swelling and cellulitis, proptosis, ptosis, and decreased extraocular movement are common, as is retro- or periorbital pain. Nasopharyngeal ulcerations, epistaxis, and headaches are also common, and involvement of cranial nerves V and VII has been described in more advanced cases. Bony erosion may be evident on examination or endoscopy. Often the patient does not appear to be seriously ill despite the rapidly progressive nature of these infections.

Patients with acute nosocomial sinusitis are often critically ill and thus do not manifest the typical clinical features of sinus disease. This diagnosis should be suspected, however, when hospitalized patients with appropriate risk factors (e.g., nasotracheal intubation) develop fever without another apparent cause.

Diagnosis

Distinguishing viral from bacterial rhinosinusitis in the ambulatory setting is usually difficult because of the relatively low sensitivity and specificity of the common clinical features. One clinical feature that has been used to help guide diagnostic and therapeutic decision-making is illness duration. Because acute bacterial sinusitis is uncommon in patients whose symptoms have lasted <10 days, expert panels now recommend reserving this diagnosis for patients with "persistent" symptoms (i.e., symptoms lasting >10 days in adults or >10–14 days in children) accompanied by the three cardinal signs of purulent nasal discharge, nasal obstruction, and facial pain (Table 31-1). The fact that, even among patients who meet these criteria, only 40–50% have true bacterial sinusitis prompts some authorities to favor 14 days of symptoms before considering treatment. The use of CT or sinus radiography is not recommended for acute disease, particularly early in the course of illness (i.e., at <10 days) in light of the high prevalence of similar findings among patients with acute viral rhinosinusitis. In the evaluation of persistent, recurrent, or chronic sinusitis, CT of the sinuses becomes the radiographic study of choice.

TABLE 31-1

Guidelines for the Diagnosis and Treatment of Acute Bacterial Sinusitis in Adults

Diagnostic Criteria	Treatment Recommendations[a]
Moderate symptoms (e.g., nasal purulence/congestion or cough) for >10 d ***or*** Severe symptoms of any duration, including unilateral/focal facial swelling or tooth pain	*Initial therapy:* Amoxicillin/clavulanate, 500/125 mg PO tid or 875/125 mg PO bid[b] *Penicillin allergy:* Doxycycline, 100 mg PO bid; ***or*** An antipneumococcal fluoroquinolone (e.g., moxifloxacin, 400 mg/d PO daily)[c] *Exposure to antibiotics within 30 d or >30% prevalence of penicillin-resistant Streptococcus pneumoniae:* Amoxicillin/clavulanate (extended release), 2000/125 mg PO bid; ***or*** Doxycycline, 100 mg PO bid; ***or*** An antipneumococcal fluoroquinolone (e.g., moxifloxacin, 400 mg PO daily)[c] *Recent treatment failure:* Amoxicillin/clavulanate (extended release), 2000 mg PO bid; ***or*** An antipneumococcal fluoroquinolone (e.g., moxifloxacin, 400 mg PO daily)[c]

[a]The duration of therapy is 5–7 days if symptoms improve within the first few days of treatment but can be up to 7–10 days, with appropriate follow-up. Severe disease may warrant IV antibiotics and consideration of hospital admission.

[b]In areas where the prevalence of antibiotic resistance is low, amoxicillin can be considered as initial therapy in patients without recent antibiotic exposure. [c]Fluoroquinolones carry a risk of tendinitis and neuropathy and should be used only if other options are not reasonable, with consideration of risks and benefits.

The clinical history and/or setting often can identify cases of acute anaerobic bacterial sinusitis, acute fungal sinusitis, or sinusitis from noninfectious causes (e.g., allergic rhinosinusitis). In the case of an immunocompromised patient with acute fungal sinus infection, immediate examination by an otolaryngologist is required. In addition to cultures, biopsy specimens from involved areas should be examined by a pathologist for evidence of fungal hyphal elements and tissue invasion. Cases of suspected acute nosocomial sinusitis should be confirmed by sinus CT. Because therapy should target the offending organism, a sinus aspirate for culture and susceptibility testing should be obtained, whenever possible,

before the initiation of antimicrobial therapy. As the ability to isolate the sometimes-myriad components of the sinus microbiome is augmented by molecular techniques, the hope is for an even more tailored treatment regimen.

TREATMENT

TREATMENT

Acute Rhinosinusitis

Most patients with a clinical diagnosis of acute rhinosinusitis improve without antibiotic therapy. The preferred initial approach in patients with mild to moderate symptoms of short duration is therapy aimed at symptom relief and facilitation of sinus drainage, such as with oral and topical decongestants, nasal saline lavage, and—at least in patients with a history of chronic sinusitis or allergies—nasal glucocorticoids. Newer studies have cast doubt on the role of antibiotics and nasal glucocorticoids in acute rhinosinusitis. In one notable double-blind, randomized, placebo-controlled trial, neither antibiotics nor topical glucocorticoids had a significant impact on cure in the study population of patients, the majority of whom had had symptoms for <7 days. Similarly, another high-profile randomized trial comparing antibiotics to placebo in patients with acute rhinosinusitis demonstrated no significant improvement in symptoms by the third day of therapy. Still, antibiotic therapy can be considered for adult patients whose condition does not improve after 10–14 days, and patients with more severe symptoms (regardless of duration) should be treated with antibiotics (Table 31-1). However, watchful waiting remains a viable option in many cases.

Empirical antibiotic therapy for community-acquired sinusitis in adults should consist of the narrowest-spectrum agent active against the most common bacterial pathogens, including *S. pneumoniae* and *H. influenzae*—e.g., amoxicillin/clavulanate (with the decision guided by local rates of β-lactamase-producing *H. influenzae*). No clinical trials support the use of broader-spectrum agents for routine cases of bacterial sinusitis, even in the current era of drug-resistant *S. pneumoniae*. For those patients who do not respond to initial antimicrobial therapy, sinus aspiration and/or lavage by an otolaryngologist should be considered. Antibiotic prophylaxis to prevent episodes of recurrent acute bacterial sinusitis is not recommended.

Surgical intervention and IV antibiotic administration usually are reserved for patients with severe disease or those with intracranial complications such as abscess and orbital involvement. Immunocompromised patients with acute invasive fungal sinusitis usually require extensive surgical debridement and treatment with IV antifungal agents active against fungal hyphal forms, such as amphotericin B. Specific therapy should be individualized according to the fungal species and its susceptibilities as well as the individual patient's characteristics.

Treatment of nosocomial sinusitis should begin with broad-spectrum antibiotics to cover common and often resistant pathogens such as *S. aureus* and gram-negative bacilli. Therapy then should be tailored to the results of culture and susceptibility testing of sinus aspirates.

CHRONIC SINUSITIS

Chronic sinusitis is characterized by symptoms of sinus inflammation lasting >12 weeks. This illness is most commonly associated with either bacteria or fungi, and clinical cure in most cases is very difficult. Many patients have undergone treatment with repeated courses of antibacterial agents and multiple sinus surgeries, increasing their risk of colonization with antibiotic-resistant pathogens and of surgical complications. These patients often have high rates of morbidity, sometimes over many years.

In *chronic bacterial sinusitis*, infection is thought to be due to the impairment of mucociliary clearance from repeated infections rather than to persistent bacterial infection. The pathogenesis of this condition, however, is poorly understood. The role of biofilms in such chronic infections continues to be explored, including the contribution that low-virulence pathogens may play in this complex, interacting milieu. Although certain conditions (e.g., cystic fibrosis) can predispose patients to chronic bacterial sinusitis, most patients with chronic rhinosinusitis do not have obvious underlying conditions that result in the obstruction of sinus drainage, the impairment of ciliary action, or immune dysfunction. Patients experience constant nasal congestion and sinus pressure, with intermittent periods of greater severity, which may persist for years. CT can be helpful in determining the extent of disease, detecting an underlying anatomic defect or obstructing process (e.g., a polyp), and assessing the response to therapy. Management should involve an otolaryngologist to conduct endoscopic examinations and obtain tissue samples for histologic examination and culture. An endoscopy-derived culture not only has a higher yield but also allows direct visualization for abnormal anatomy.

Chronic fungal sinusitis is a disease of immunocompetent hosts and is usually noninvasive, although slowly progressive invasive disease is sometimes seen. Noninvasive disease, which typically is associated with hyaline molds such as *Aspergillus* species and dematiaceous molds such as *Curvularia* or *Bipolaris* species, can present as a number of different scenarios. In mild, indolent disease, which usually occurs in the setting of repeated failures of antibacterial therapy, only nonspecific mucosal changes may be seen on sinus CT. Although there is some controversy on this point, endoscopic surgery is usually curative in these cases, with no need for antifungal therapy. Another form of disease presents as long-standing, often unilateral symptoms and opacification of a single sinus on imaging studies as a result of a mycetoma (fungus ball) within the sinus. Treatment for this condition also is surgical, although systemic antifungal therapy may be warranted in the rare case in which bony erosion occurs. A third form of disease, known as *allergic fungal sinusitis*, is seen in patients with a history of nasal polyposis and asthma, who often have had multiple sinus surgeries. Patients with this condition produce a thick, eosinophil-laden mucus with the consistency of peanut butter that contains sparse fungal hyphae on histologic examination. These patients often present with pansinusitis.

TREATMENT

TREATMENT

Chronic Sinusitis

Treatment of chronic bacterial sinusitis can be challenging and consists primarily of repeated culture-guided courses of antibiotics, sometimes for 3–4 weeks or longer at a time; administration of intranasal glucocorticoids; and mechanical irrigation of the sinus with sterile saline solution. When this management approach fails, sinus surgery may be indicated and sometimes provides significant, albeit short-term, alleviation. Treatment of chronic fungal sinusitis consists of surgical removal of impacted mucus. Recurrence, unfortunately, is common.

INFECTIONS OF THE EAR AND MASTOID

Infections of the ear and associated structures can involve both the middle and the external ear, including the skin, cartilage, periosteum, ear canal, and tympanic and mastoid cavities. Both viruses and bacteria are known causes of these infections, some of which result in significant morbidity if not treated appropriately.

INFECTIONS OF EXTERNAL EAR STRUCTURES

Infections involving the structures of the external ear are often difficult to differentiate from noninfectious inflammatory conditions with similar clinical manifestations. Clinicians should consider inflammatory disorders as possible causes of external ear irritation, particularly in the absence of local or regional adenopathy. Aside from the more salient causes of inflammation, such as trauma, insect bite, and overexposure to sunlight or extreme cold, the differential diagnosis should include less common conditions such as autoimmune disorders (e.g., lupus or relapsing polychondritis) and vasculitides (e.g., granulomatosis with polyangiitis).

Auricular Cellulitis

Auricular cellulitis is an infection of the skin overlying the external ear and typically follows minor local trauma. It presents as the typical signs and symptoms of cellulitis, with tenderness, erythema, swelling, and warmth of the external ear (particularly the lobule) but without apparent involvement of the ear canal or inner structures. Treatment consists of warm compresses and oral antibiotics such as cephalexin or dicloxacillin that are active against typical skin and soft-tissue pathogens (specifically, *S. aureus* and streptococci). IV antibiotics such as a first-generation cephalosporin (e.g., cefazolin) or a penicillinase-resistant penicillin (e.g., nafcillin) occasionally are needed for more severe cases, with consideration of MRSA if either risk factors or failure of therapy point to this organism.

Perichondritis

Perichondritis, an infection of the perichondrium of the auricular cartilage, typically follows local trauma (e.g., piercings, burns, or lacerations). Occasionally, when the infection spreads down to the cartilage of the pinna itself, patients may develop chondritis. The infection may closely resemble auricular cellulitis, with erythema, swelling, and extreme tenderness of the pinna, although the lobule is less often involved in perichondritis. The most common pathogens are *P. aeruginosa* and *S. aureus*, although other gram-negative and gram-positive organisms occasionally are involved. Treatment consists of systemic antibiotics active

against both *P. aeruginosa* and *S. aureus*. An antipseudomonal penicillin (e.g., piperacillin) or a combination of a penicillinase-resistant penicillin and an antipseudomonal quinolone (e.g., nafcillin plus ciprofloxacin) is typically used. Incision and drainage may be helpful for culture and for resolution of infection, which often takes weeks. When perichondritis fails to respond to adequate antimicrobial therapy, clinicians should consider a noninfectious inflammatory etiology such as relapsing polychondritis.

Otitis Externa

The term *otitis externa* refers to a collection of diseases involving primarily the auditory meatus. Otitis externa usually results from a combination of heat and retained moisture, with desquamation and maceration of the epithelium of the outer ear canal. The disease exists in several forms: localized, diffuse, chronic, and invasive. All forms are predominantly bacterial in origin, with *P. aeruginosa* and *S. aureus* the most common pathogens.

Acute localized otitis externa (*furunculosis*) can develop in the outer third of the ear canal, where skin overlies cartilage and hair follicles are numerous. As in furunculosis elsewhere on the body, *S. aureus* is the usual pathogen, and treatment typically consists of an oral antistaphylococcal penicillin (e.g., dicloxacillin or cephalexin), with incision and drainage in cases of abscess formation.

Acute diffuse otitis externa is also known as *swimmer's ear*, although it can develop in patients who have not recently been swimming. Heat, humidity, and the loss of protective cerumen lead to excessive moisture and elevation of the pH in the ear canal, which in turn lead to skin maceration and irritation. Infection may then follow; the predominant pathogen is *P. aeruginosa*, although other bacteria—and rarely yeasts—have been recovered from patients with this condition. The illness often starts with itching and progresses to severe pain, which is usually elicited by manipulation of the pinna or tragus. The onset of pain is generally accompanied by the development of an erythematous, swollen ear canal, often with scant white, clumpy discharge. Treatment consists of cleansing the canal to remove debris and enhance the activity of topical therapeutic agents—usually hypertonic saline or mixtures of alcohol and acetic acid. Inflammation can also be decreased by adding glucocorticoids to the treatment regimen or by using Burow's solution (aluminum acetate in water). Antibiotics are most effective when given topically. Otic mixtures provide adequate pathogen coverage; these preparations usually combine neomycin with polymyxin, with or without glucocorticoids. Systemic antimicrobial agents typically are reserved for severe disease or infections in immunocompromised hosts.

Chronic otitis externa is caused primarily by repeated local irritation, most commonly arising from persistent drainage from a chronic middle-ear infection. Other causes of repeated irritation, such as insertion of cotton swabs or other foreign objects into the ear canal, can lead to this condition, as can rare chronic infections such as syphilis, tuberculosis, and leprosy. Chronic otitis externa typically presents as erythematous, scaling dermatitis in which the predominant symptom is pruritus rather than pain; this condition must be differentiated from several others that produce a similar clinical picture, such as atopic dermatitis, seborrheic dermatitis, psoriasis, and dermatomycosis. Therapy consists of identifying and treating or removing the offending process, although successful resolution is frequently difficult.

Invasive otitis externa, also known as *malignant* or *necrotizing* otitis externa, is an aggressive and potentially life-threatening disease that occurs predominantly in elderly diabetic patients and other immunocompromised persons. The disease begins in the external canal as a soft-tissue infection that progresses slowly over weeks to months and often is difficult to distinguish from a severe case of chronic otitis externa because of the presence of purulent otorrhea and an erythematous swollen ear and external canal. Severe, deep-seated otalgia, frequently out of proportion to findings on examination, is often noted and can help differentiate invasive from chronic otitis externa. The characteristic finding on examination is granulation tissue in the posteroinferior wall of the external canal, near the junction of bone and cartilage. If left unchecked, the infection can migrate to the base of the skull (resulting in skull-base osteomyelitis) and onward to the meninges and brain, with a high mortality rate. Cranial nerve involvement is seen occasionally, with the facial nerve usually affected first and most often. Thrombosis of the sigmoid sinus can occur if the infection extends to the area. CT, which can reveal osseous erosion of the temporal bone and skull base, can be used to help determine the extent of disease, as can gallium and technetium-99 scintigraphy studies. *P. aeruginosa* is by far the most common offender, although *S. aureus*, *Staphylococcus epidermidis*, *Aspergillus*, *Actinomyces*, and some gram-negative bacteria also have been associated with this disease. In all cases, the external ear canal should be cleansed and a biopsy specimen of the granulation tissue within the canal (or of deeper tissues) obtained for culture of the offending organism. IV antibiotic therapy should be given for a prolonged course (6–8 weeks) and directed specifically toward the recovered pathogen. For *P. aeruginosa*, the regimen typically includes an antipseudomonal penicillin or cephalosporin (e.g., piperacillin or cefepime), sometimes with an aminoglycoside or a fluoroquinolone, the latter of which can even be administered orally given its excellent bioavailability. In addition, antibiotic drops containing an agent active against *Pseudomonas* (e.g., ciprofloxacin) are usually prescribed and are combined with glucocorticoids to reduce inflammation. Cases of invasive *Pseudomonas* otitis externa recognized in the early stages can sometimes be treated with oral and otic fluoroquinolones alone, albeit with close follow-up. Extensive surgical debridement, once an important component of the treatment approach, is now rarely indicated.

In *necrotizing otitis externa*, recurrence is documented up to 20% of the time. Aggressive glycemic control in diabetics is important not only for effective treatment but also for prevention of recurrence. The role of hyperbaric oxygen has not been clearly established.

INFECTIONS OF MIDDLE-EAR STRUCTURES

Otitis media is an inflammatory condition of the middle ear that results from dysfunction of the eustachian tube in association with a number of illnesses, including URIs and chronic rhinosinusitis. The inflammatory response in these conditions leads to the development of a sterile transudate within the middle-ear and mastoid cavities. Infection may occur if bacteria or viruses from the nasopharynx contaminate this fluid, producing an acute (or sometimes chronic) illness.

Acute Otitis Media

Acute otitis media results when pathogens from the nasopharynx are introduced into the inflammatory fluid collected in the middle ear (e.g., by nose blowing during a URI). Pathogenic proliferation in this space leads to

the development of the typical signs and symptoms of acute middle-ear infection. The diagnosis of acute otitis media requires the demonstration of fluid in the middle ear (with tympanic membrane [TM] immobility) and the accompanying signs or symptoms of local or systemic illness (Table 31-2).

TABLE 31-2

Guidelines for the Diagnosis and Treatment of Acute Otitis Media

Illness Severity	Diagnostic Criteria	Treatment Recommendations
Mild to moderate	>2 yrs *or* 6 mo to 2 yrs without middle-ear effusion	*Observation alone* (deferring antibiotic therapy for 48–72 h and limiting management to symptom relief)

Illness Severity	Diagnostic Criteria	Treatment Recommendations
	<6 mo; ***or*** 6 mo to 2 yrs with middle-ear effusion (fluid in the middle ear, evidenced by decreased TM mobility, air/fluid level behind TM, bulging TM, purulent otorrhea) *and* acute onset of signs and symptoms of middle-ear inflammation, including fever, otalgia, decreased hearing, tinnitus, vertigo, erythematous TM; ***or*** >2 yrs with bilateral disease, TM perforation, high fever, immunocompromise, emesis	*Initial therapy*[a]: Amoxicillin, 80–90 mg/kg qd (up to 2 g) PO in divided doses (bid or tid); ***or*** Cefdinir, 14 mg/kg qd PO in 1 dose or divided doses (bid); ***or*** Cefuroxime, 30 mg/kg qd PO in divided doses (bid); ***or*** Azithromycin, 10 mg/kg qd PO on day 1 followed by 5 mg/kg qd PO for 4 d *Exposure to antibiotics within 30 d or recent treatment failure*[a,b]: Amoxicillin, 90 mg/kg qd (up to 2 g) PO in divided doses (bid), plus clavulanate, 6.4 mg/kg qd PO in divided doses (bid); ***or*** Ceftriaxone, 50 mg/kg IV/IM qd for 3 d; ***or*** Clindamycin, 30–40 mg/kg qd PO in divided doses (tid)

Illness Severity	Diagnostic Criteria	Treatment Recommendations
Severe	As above, with temperature ≥39.0°C (≥102°F); ***or*** Moderate to severe otalgia	*Initial therapy*[a]: Amoxicillin, 90 mg/kg qd (up to 2 g) PO in divided doses (bid), plus clavulanate, 6.4 mg/kg qd PO in divided doses (bid); ***or*** Ceftriaxone, 50 mg/kg IV/IM qd for 3 d *Exposure to antibiotics within 30 d or recent treatment failure*[a,b]: Ceftriaxone, 50 mg/kg IV/IM qd for 3 d; ***or*** Clindamycin, 30–40 mg/kg qd PO in divided doses (tid); ***or*** Consider tympanocentesis with culture

[a]Duration (unless otherwise specified): 10 days for patients <6 years old and patients with severe disease; 5–7 days (with consideration of observation only in previously healthy individuals with mild disease) for patients ≥6 years old.

[b]Failure to improve and/or clinical worsening after 48–72 h of observation or treatment.

Abbreviation: TM, tympanic membrane.

Source: American Academy of Pediatrics Subcommittee on Management of Acute Otitis Media, 2004.

ETIOLOGY

Acute otitis media typically follows a viral URI. The causative viruses (most commonly RSV, influenza virus, rhinovirus, and enterovirus) can themselves cause subsequent acute otitis media; more often, they predispose the patient to bacterial otitis media. Studies using tympanocentesis have consistently found *S. pneumoniae* to be the most important bacterial cause, isolated in up to 35% of cases. *H. influenzae* (nontypable strains) and *M. catarrhalis* also are common bacterial causes of acute otitis media, and concern is increasing with MRSA as an emerging etiologic agent. Viruses, such as those mentioned above, have been recovered either alone or with bacteria in 17–40% of cases.

CLINICAL MANIFESTATIONS

Fluid in the middle ear is typically demonstrated or confirmed with pneumatic otoscopy. In the absence of fluid, the TM moves visibly with the application of positive and negative pressure, but this movement is dampened when fluid is present. With bacterial infection, the TM can also be erythematous, bulging, or retracted and occasionally can perforate spontaneously. The signs and symptoms accompanying infection can be local or systemic, including otalgia, otorrhea, diminished hearing, and fever. Erythema of the TM is often evident but is nonspecific as it frequently is seen in association with inflammation of the upper respiratory mucosa. Other signs and symptoms occasionally reported include vertigo, nystagmus, and tinnitus.

TREATMENT

TREATMENT

Acute Otitis Media

There has been considerable debate on the usefulness of antibiotics for the treatment of acute otitis media. A higher proportion of treated than untreated patients are free of illness 3–5 days after diagnosis. The difficulty of predicting which patients will benefit from antibiotic therapy has led to different approaches. In the Netherlands, for instance, physicians typically manage acute otitis media with initial observation, administering anti-inflammatory agents for aggressive pain management and reserving antibiotics for high-risk patients, patients with complicated disease, or patients whose condition does not improve after 48–72 h. In contrast, many experts in the United States continue to recommend antibiotic therapy for children <6 months old in light of the higher frequency of secondary complications in this young and functionally immunocompromised population. However, observation without antimicrobial therapy is now the recommended option in the United States for acute otitis media in children >2 years of age and for mild to moderate disease without middle-ear effusion in children 6 months to 2 years of age. Treatment is typically indicated for patients <6 months old; for children 6 months to 2 years old who have middle-ear effusion and signs/symptoms of middle-ear inflammation; for all patients >2 years old who have bilateral disease, TM perforation, immunocompromise, or emesis; and for any patient who has severe symptoms, including a fever ≥39°C or moderate to severe otalgia (Table 31-2).

Because most studies of the etiologic agents of acute otitis media consistently document similar pathogen profiles, therapy is generally empirical except in those few cases in which tympanocentesis is warranted—e.g., cases refractory to therapy and cases in patients who are severely ill or immunodeficient. Despite resistance to penicillin and amoxicillin in roughly one-quarter of *S. pneumoniae* isolates, one-third of *H. influenzae* isolates, and nearly all *M. catarrhalis* isolates, outcome studies continue to find that amoxicillin is as successful as any other agent, and it remains the drug of first choice in recommendations from multiple sources (Table 31-2). Therapy for uncomplicated acute otitis media typically is administered for 5–7 days to patients aged ≥6 years; longer courses (e.g., 10 days) should be reserved for immunocompromised patients or patients with severe disease, in whom short-course therapy may be inadequate.

A switch in regimen is recommended if there is no clinical improvement by the third day of therapy, given the possibility of infection with a β-lactamase-producing strain of *H. influenzae* or *M. catarrhalis* or with a strain

of penicillin-resistant *S. pneumoniae*. Decongestants and antihistamines are frequently used as adjunctive agents to reduce congestion and relieve obstruction of the eustachian tube, but clinical trials have yielded no significant evidence of benefit with either class of agents.

Recurrent Acute Otitis Media

Recurrent acute otitis media (more than three episodes within 6 months or four episodes within 12 months) generally is due to relapse or reinfection, although data indicate that the majority of early recurrences are new infections. In general, the same pathogens responsible for acute otitis media cause recurrent disease; even so, the recommended treatment consists of antibiotics active against β-lactamase-producing organisms. Antibiotic prophylaxis (e.g., with amoxicillin) can reduce recurrences in patients with recurrent acute otitis media by an average of one episode per year, which benefit is small compared with the high likelihood of colonization with antibiotic-resistant pathogens. Other approaches, including placement of tympanostomy tubes, adenoidectomy, and tonsillectomy plus adenoidectomy, are of questionable overall value in light of the relatively small benefit compared with the potential for complications.

Serous Otitis Media

In serous otitis media (otitis media with effusion), fluid is present in the middle ear for an extended period in the absence of signs and symptoms of infection. In general, acute effusions are self-limited; most resolve in 2–4 weeks. In some cases, however (in particular after an episode of acute otitis media), effusions can persist for months. These chronic effusions are often associated with significant hearing loss in the affected ear. The great majority of cases of otitis media with effusion resolve spontaneously within 3 months without antibiotic therapy. Antibiotic therapy or myringotomy with insertion of tympanostomy tubes typically is reserved for patients in whom bilateral effusion (1) has persisted for at least 3 months and (2) is associated with significant bilateral hearing loss. With this conservative approach and the application of strict diagnostic criteria for acute otitis media and otitis media with effusion, an estimated 6–8 million courses of antibiotics could be avoided each year in the United States.

Chronic Otitis Media

Chronic suppurative otitis media is characterized by persistent or recurrent purulent otorrhea in the setting of TM perforation. Usually, there is also some degree of conductive hearing loss. This condition can be categorized as active or inactive. Inactive disease is characterized by a central perforation of the TM, which allows drainage of purulent fluid from the middle ear. When the perforation is more peripheral, squamous epithelium from the auditory canal may invade the middle ear through the perforation, forming a mass of keratinaceous debris (*cholesteatoma*) at the site of invasion. This mass can enlarge and has the potential to erode bone and promote further infection, which can lead to meningitis, brain abscess, or paralysis of cranial nerve VII. Treatment of chronic active otitis media is surgical; mastoidectomy, myringoplasty, and tympanoplasty can be performed as outpatient surgical procedures, with an overall success rate of ~80%. Chronic inactive otitis media is more difficult to cure, usually requiring repeated courses of topical antibiotic

drops during periods of drainage. Systemic antibiotics may offer better cure rates, but their role in the treatment of this condition remains unclear.

Mastoiditis

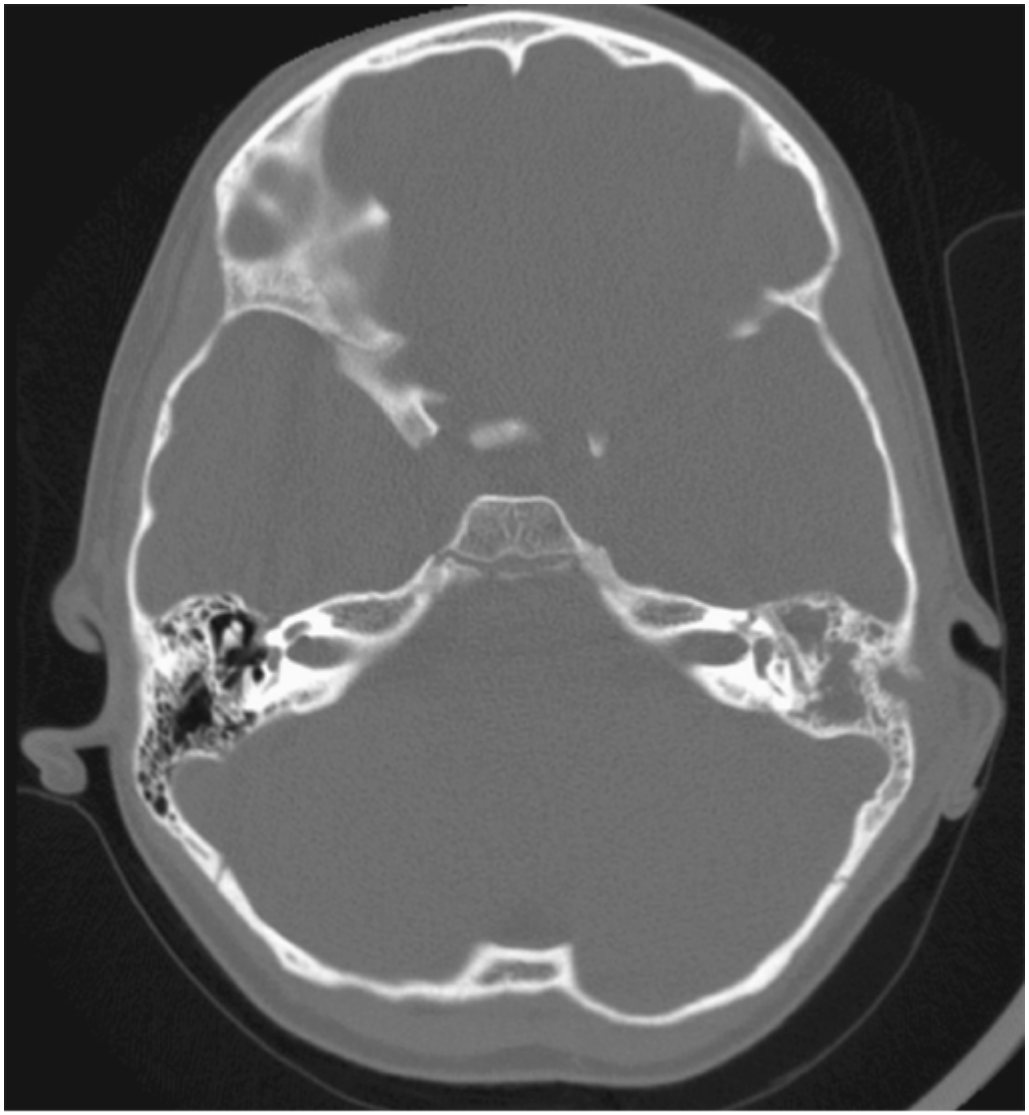
Acute mastoiditis was relatively common among children before the introduction of antibiotics. Because the mastoid air cells connect with the middle ear, the process of fluid collection and infection is usually the same in the mastoid as in the middle ear. Early and frequent treatment of acute otitis media is most likely the reason that the incidence of acute mastoiditis has declined to only 1.2–2.0 cases per 100,000 person-years in countries with high prescribing rates for acute otitis media.

In countries such as the Netherlands, where antibiotics are used sparingly for acute otitis media, the incidence rate of acute mastoiditis is roughly twice that in countries like the United States. However, neighboring Denmark has a rate of acute mastoiditis similar to that in the Netherlands but an antibiotic-prescribing rate for acute otitis media more similar to that in the United States.

In typical acute mastoiditis, purulent exudate collects in the mastoid air cells **(Fig. 31-1)**, producing pressure that may result in erosion of the surrounding bone and formation of abscess-like cavities that are usually evident on CT. Patients typically present with pain, erythema, and swelling of the mastoid process along with displacement of the pinna, usually in conjunction with the typical signs and symptoms of acute middle-ear infection. Rarely, patients can develop severe complications if the infection tracks under the periosteum of the temporal bone to cause a subperiosteal abscess, erodes through the mastoid tip to cause a deep neck abscess, or extends posteriorly to cause septic thrombosis of the lateral sinus.

FIGURE 31-1

Acute mastoiditis. Axial CT image shows an acute fluid collection within the mastoid air cells on the left.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Purulent fluid should be cultured whenever possible to help guide antimicrobial therapy. Initial empirical therapy usually is directed against the typical organisms associated with acute otitis media, such as *S. pneumoniae*, *H. influenzae*, and *M. catarrhalis*. Patients with more severe or prolonged courses of illness should be treated for infection with *S. aureus* and gram-negative bacilli (including *Pseudomonas*). Broad-spectrum empirical therapy should be narrowed once culture results become available. Most patients can be treated conservatively with IV antibiotics; surgery (cortical mastoidectomy) is reserved for complicated cases and those in which conservative treatment has failed.

INFECTIONS OF THE PHARYNX AND ORAL CAVITY

Oropharyngeal infections range from mild, self-limited viral illnesses to serious, life-threatening bacterial infections. The most common presenting symptom is sore throat—one of the most common reasons for ambulatory care visits by both adults and children. Although sore throat is a symptom in many noninfectious illnesses as well, the overwhelming majority of patients with a new sore throat have acute pharyngitis of viral or bacterial etiology.

ACUTE PHARYNGITIS

Millions of visits to primary care providers each year are for sore throat; the majority of cases of acute pharyngitis are caused by typical respiratory viruses. The most important source of concern is infection with group A β-hemolytic *Streptococcus* (*S. pyogenes*), which is associated with acute glomerulonephritis and acute rheumatic fever. The risk of rheumatic fever can be reduced by timely penicillin therapy.

Etiology

A wide variety of organisms cause acute pharyngitis. The relative importance of the different pathogens can only be estimated, since a significant proportion of cases (~30%) have no identified cause. Together, respiratory viruses are the most common identifiable cause of acute pharyngitis, with rhinoviruses and coronaviruses accounting for large proportions of cases (~20% and at least 5%, respectively). Influenza virus, parainfluenza virus, and adenovirus also account for a measurable share of cases, with the former two more seasonal and the latter as part of the more clinically severe syndrome of pharyngoconjunctival fever. Other important but less common viral causes include herpes simplex virus (HSV) types 1 and 2, coxsackievirus A, cytomegalovirus (CMV), and Epstein-Barr virus (EBV). Acute HIV infection can present as acute pharyngitis and should always be considered in at-risk populations.

Acute bacterial pharyngitis is typically caused by *S. pyogenes*, which accounts for ~5–15% of all cases of acute pharyngitis in adults; rates vary with the season and with utilization of the health care system. Group A streptococcal pharyngitis is primarily a disease of children aged 5–15 years; it is uncommon among children <3 years old, as is rheumatic fever. Streptococci of groups C and G account for a minority of cases, although these serogroups are nonrheumatogenic. *Fusobacterium necrophorum* has been increasingly recognized as a cause of pharyngitis in adolescents and young adults and, when sought, is isolated nearly as often as group A streptococci. This organism is important because of the rare but life-threatening *Lemierre disease*, which is generally associated with *F. necrophorum* and is usually preceded by pharyngitis (see "Oral Infections," below). The remaining bacterial causes of acute pharyngitis are seen infrequently (<1% of cases each) but should be considered in appropriate exposure groups because of the severity of illness if left untreated; these etiologic agents include *Neisseria gonorrhoeae*, *Corynebacterium diphtheriae*, *Corynebacterium ulcerans*, *Yersinia enterocolitica*, and *Treponema pallidum* (in secondary syphilis). Anaerobic bacteria also can cause acute pharyngitis (*Vincent angina*) and can contribute to more serious polymicrobial infections, such as peritonsillar or retropharyngeal abscesses (see below). Atypical organisms such as *M. pneumoniae* and *C. pneumoniae* have been recovered from patients with acute pharyngitis; whether these agents are commensals or causes of acute infection is debatable.

Clinical Manifestations

Although the signs and symptoms accompanying acute pharyngitis are not reliable predictors of the etiologic agent, the clinical presentation occasionally suggests one etiology over another. Acute pharyngitis due to respiratory viruses such as rhinovirus or coronavirus usually is not severe and typically is associated with a constellation of coryzal symptoms better characterized as nonspecific URI. Findings on physical examination

are uncommon; fever is rare, and tender cervical adenopathy and pharyngeal exudates are not seen. In contrast, acute pharyngitis from influenza virus can be severe and is much more likely to be associated with fever as well as with myalgias, headache, and cough. The presentation of pharyngoconjunctival fever due to adenovirus infection is similar. Since pharyngeal exudate may be present on examination, this condition can be difficult to differentiate from streptococcal pharyngitis. However, adenoviral pharyngitis is distinguished by the presence of conjunctivitis in one-third to one-half of patients. Acute pharyngitis from primary HSV infection can also mimic streptococcal pharyngitis in some cases, with pharyngeal inflammation and exudate, but the presence of vesicles and shallow ulcers on the palate can help differentiate the two diseases. This HSV syndrome is distinct from pharyngitis caused by coxsackievirus (*herpangina*), which is associated with small vesicles that develop on the soft palate and uvula and then rupture to form shallow white ulcers. Acute pharyngitis coupled with fever, fatigue, generalized lymphadenopathy, and (on occasion) splenomegaly is characteristic of infectious mononucleosis due to EBV or CMV. Acute primary infection with HIV is frequently associated with fever and acute pharyngitis as well as with myalgias, arthralgias, malaise, and occasionally a nonpruritic maculopapular rash, which may be followed by lymphadenopathy and mucosal ulcerations without exudate.

The clinical features of acute pharyngitis caused by streptococci of groups A, C, and G are similar, ranging from a relatively mild illness without many accompanying symptoms to clinically severe cases with profound pharyngeal pain, fever, chills, and abdominal pain. A hyperemic pharyngeal membrane with tonsillar hypertrophy and exudate is usually seen, along with tender anterior cervical adenopathy. Coryzal manifestations, including cough, are typically absent; when present, they suggest a viral etiology. Strains of *S. pyogenes* that generate erythrogenic toxin can also produce scarlet fever characterized by an erythematous rash and strawberry tongue. The other types of acute bacterial pharyngitis (e.g., gonococcal, diphtherial, and yersinial) often present as exudative pharyngitis with or without other clinical features. Their etiologies are often suggested only by the clinical history.

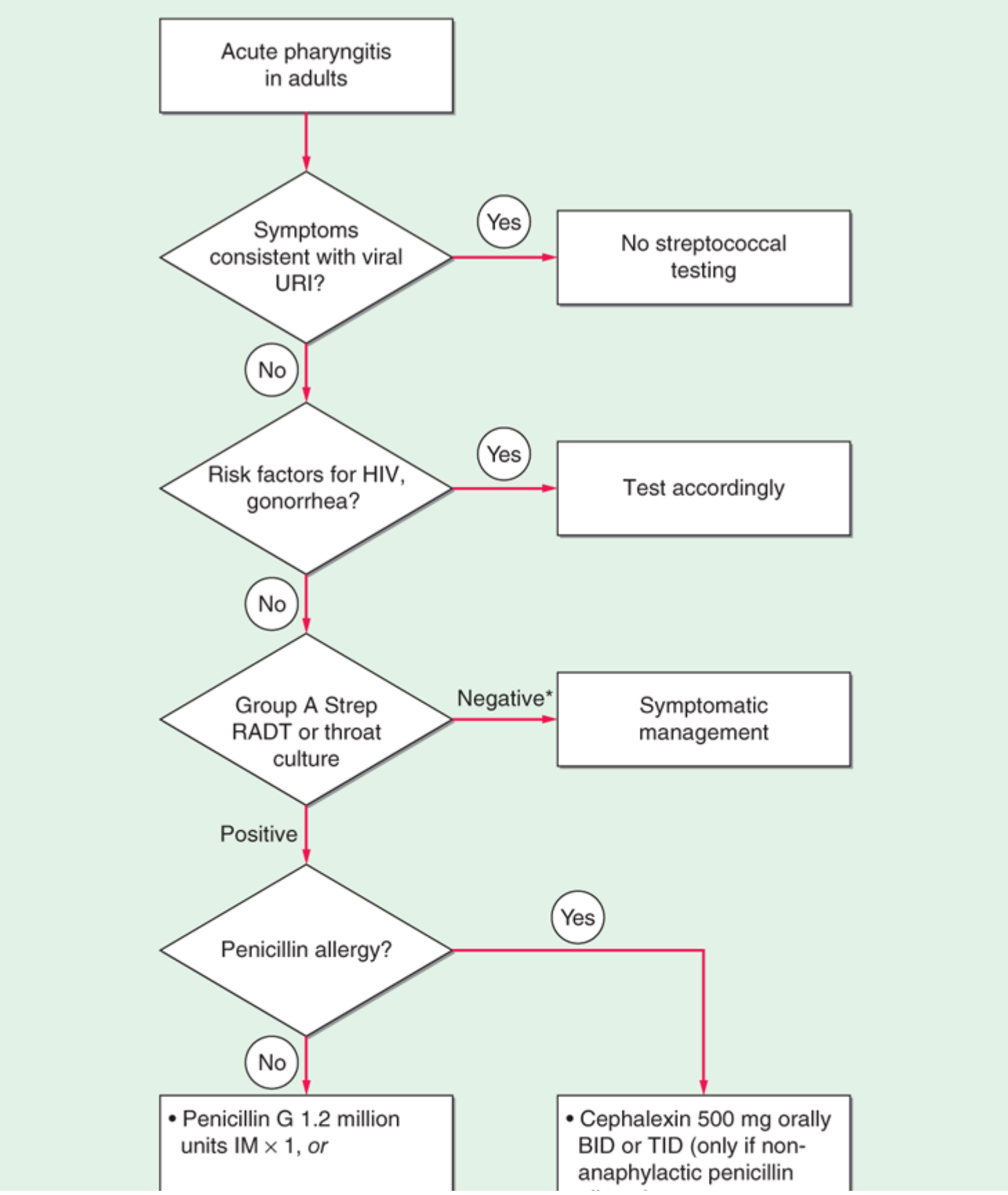
Diagnosis

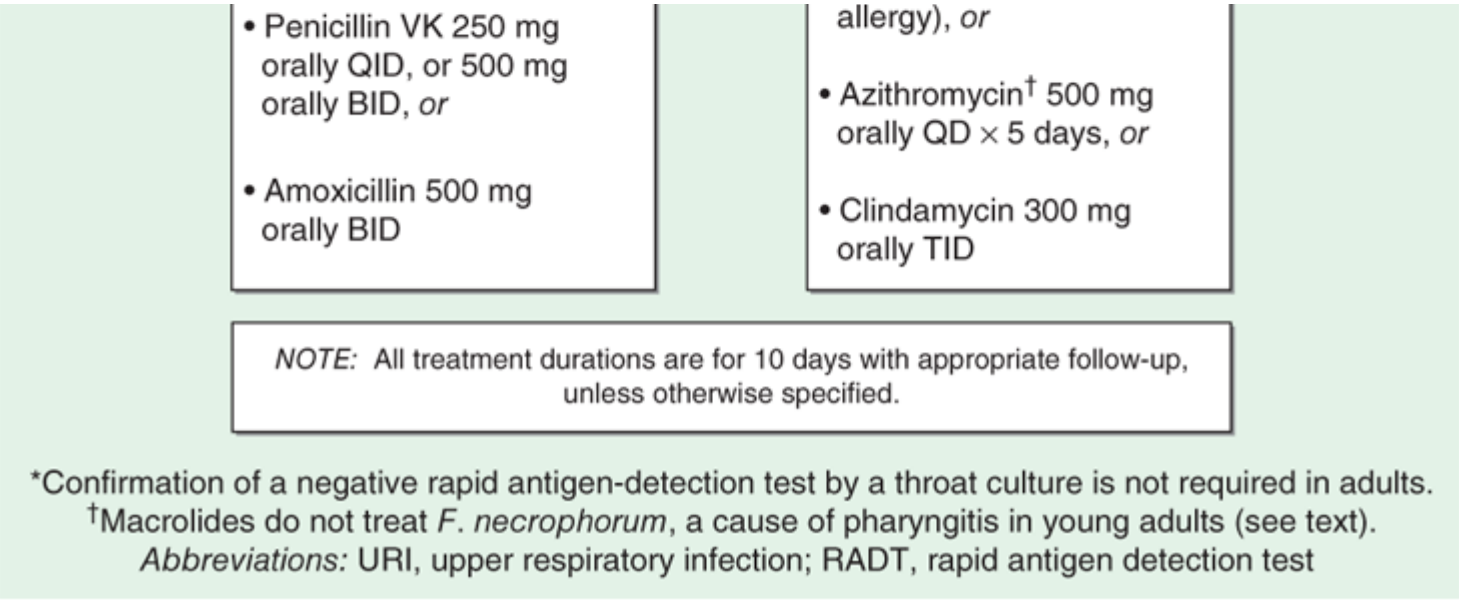
The primary goal of diagnostic testing is to separate acute streptococcal pharyngitis from pharyngitis of other etiologies (particularly viral) so that antibiotics can be prescribed more efficiently for patients in whom they may be beneficial. The most appropriate standard for the diagnosis of streptococcal pharyngitis, however, has not been established definitively. Throat swab culture is generally regarded as the most appropriate but cannot distinguish between infection and colonization and requires 24–48 h to yield results that vary with technique and culture conditions. Rapid antigen-detection tests offer good specificity (>90%) but lower sensitivity when implemented in routine practice. Sensitivity has also been shown to vary across the clinical spectrum of disease (65–90%). Several clinical prediction systems **(Fig. 31-2)** can increase the sensitivity of rapid antigen-detection tests to >90% in controlled settings. Since the sensitivities achieved in routine clinical practice are often lower, several medical and professional societies continue to recommend that all negative rapid antigen-detection tests in children be confirmed by a throat culture to limit transmission and complications of illness caused by group A streptococci. The Centers for Disease Control and Prevention, the Infectious Diseases Society of America, and the American Academy of Family Physicians

do not recommend backup culture when adults have negative results from a highly sensitive rapid antigen-detection test, however, because of the lower prevalence and smaller benefit in this age group.

FIGURE 31-2

Algorithm for the diagnosis and treatment of acute pharyngitis.





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Cultures and rapid diagnostic tests for other causes of acute pharyngitis, such as influenza virus, adenovirus, HSV, EBV, CMV, and *M. pneumoniae*, are available in many locations and can be used when these pathogens are suspected. The diagnosis of acute EBV infection depends primarily on the detection of antibodies to the virus with a heterophile agglutination assay (monospot slide test) or enzyme-linked immunosorbent assay. Testing for HIV, ideally through a combination antigen/antibody method, should be performed when acute primary HIV infection is suspected. If other bacterial causes are suspected (particularly *N. gonorrhoeae*, *C. diphtheriae*, or *Y. enterocolitica*), specific cultures should be requested since these organisms may be missed on routine throat swab culture.

TREATMENT

TREATMENT

Pharyngitis

Antibiotic treatment of pharyngitis due to *S. pyogenes* confers numerous benefits, including a decrease in the risk of rheumatic fever—the primary focus of treatment. The magnitude of this benefit is fairly small, since rheumatic fever is now a rare disease, even among untreated patients. Nevertheless, when therapy is started within 48 h of illness onset, symptom duration is decreased modestly. An additional benefit of therapy is the potential to reduce the transmission of streptococcal pharyngitis, particularly in areas of overcrowding or close contact. Antibiotic therapy for acute pharyngitis is therefore recommended in cases in which *S. pyogenes* is confirmed as the etiologic agent by rapid antigen-detection test or throat swab culture. Otherwise, antibiotics should be given in routine cases only when another bacterial cause has been identified. Effective therapy for streptococcal pharyngitis consists of either a single dose of IM benzathine penicillin or a full 10-day course of oral penicillin (Fig. 31-2).

Azithromycin can be used in place of penicillin, although its potential utility is waning and its use in some parts of the world (particularly Europe) is prohibited as a result of resistance among *S. pyogenes* strains. Broader-spectrum (and often more expensive) antibiotics also are active against streptococci but offer no greater efficacy than the agents mentioned above. Testing for cure is unnecessary and may reveal only chronic colonization. There is no evidence to support antibiotic treatment of group C or G streptococcal pharyngitis or pharyngitis in which mycoplasmas or chlamydiae have been recovered. Cultures can be of benefit because *F. necrophorum*, an increasingly common cause of bacterial pharyngitis in young adults, is not covered by macrolide therapy. Long-term penicillin prophylaxis (benzathine penicillin G, 1.2 million units IM every 3–4 weeks; or penicillin VK, 250 mg PO twice daily) is indicated for patients at risk of recurrent rheumatic fever in order to prevent what could be catastrophic sequelae of recurrent streptococcal pharyngitis.

Antibiotic shortages, sometimes the result of manufacturing difficulties or delays, natural disasters, and regulatory or other issues, can preclude the use of the optimal antibiotic. These shortages can be regional, national, or international. Communication with pharmacists and the use of antibiotic stewardship teams can help mitigate the effects of shortages, yield recommendations for the use of alternative agents, and prevent delays in treatment that can affect patients' access to antibiotics.

Treatment of viral pharyngitis is entirely symptom-based except in infection with influenza virus or HSV. For influenza, the armamentarium includes the adamantanes amantadine and rimantadine and the neuraminidase inhibitors oseltamivir and zanamivir. Administration of all these agents needs to be started within 48 h of symptom onset to reduce illness duration meaningfully. Among these agents, only oseltamivir and zanamivir are active against both influenza A and influenza B and therefore can be used when local patterns of infection and antiviral resistance are unknown. Oropharyngeal HSV infection sometimes responds to treatment with antiviral agents such as acyclovir, although these drugs are often reserved for immunosuppressed patients.

Complications

Although rheumatic fever is the best-known complication of acute streptococcal pharyngitis, the risk of its following acute infection remains quite low. Other complications include acute glomerulonephritis and numerous suppurative conditions, such as peritonsillar abscess (*quinsy*), otitis media, mastoiditis, sinusitis, bacteremia, and pneumonia—all of which occur at low rates. Although antibiotic treatment of acute streptococcal pharyngitis can prevent the development of rheumatic fever, there is no evidence that it can prevent acute glomerulonephritis. Some evidence supports antibiotic use to prevent the suppurative complications of streptococcal pharyngitis, particularly peritonsillar abscess, which can also involve oral anaerobes such as *Fusobacterium*. Abscesses usually are accompanied by severe pharyngeal pain, dysphagia, fever, and dehydration; in addition, medial displacement of the tonsil and lateral displacement of the uvula are often evident on examination. Although early use of IV antibiotics (e.g., clindamycin, penicillin G with metronidazole) may eliminate the need for surgical drainage in some cases, treatment typically involves needle aspiration or incision and drainage.

ORAL INFECTIONS

Aside from periodontal diseases such as gingivitis, infections of the oral cavity most commonly involve HSV or *Candida* species. In addition to causing painful cold sores on the lips, HSV can infect the tongue and buccal mucosa, causing the formation of irritating vesicles. Although topical antiviral agents (e.g., acyclovir and penciclovir) can be used externally for cold sores, with possible benefit, oral or IV acyclovir is often needed for primary infections, extensive oral infections, and infections in immunocompromised patients. Oropharyngeal candidiasis (*thrush*) is caused by a variety of *Candida* species, most often *C. albicans*. Thrush occurs predominantly in neonates, immunocompromised patients (especially those with AIDS), and recipients of prolonged antibiotic or glucocorticoid therapy. In addition to sore throat, patients often report a burning tongue or abnormal taste, and physical examination reveals friable white or gray plaques on the gingiva, tongue, and oral mucosa, often with underlying erythema. Treatment, which usually consists of a topical antifungal (nystatin or clotrimazole) or oral fluconazole, is typically successful. In the uncommon cases of fluconazole-refractory thrush that are seen in some patients with HIV/AIDS or in patients with resistant organisms that can sometimes complicate the treatment of recurrent oral candidiasis, other therapeutic options include oral voriconazole, an IV echinocandin (caspofungin, micafungin, or anidulafungin), or amphotericin B deoxycholate, if needed. In these cases, therapy based on culture and susceptibility test results is ideal.

Vincent angina, also known as *acute necrotizing ulcerative gingivitis* or *trench mouth*, is a unique and dramatic form of gingivitis characterized by painful, inflamed gingiva with ulcerations of the interdental papillae that bleed easily. Since oral anaerobes are the cause, patients typically have halitosis and frequently present with fever, malaise, and lymphadenopathy. Treatment consists of debridement and oral administration of penicillin plus metronidazole, with clindamycin or doxycycline alone as an alternative.

Ludwig angina is a rapidly progressive, potentially fulminant form of cellulitis that involves the bilateral sublingual and submandibular spaces and that typically originates from an infected or recently extracted tooth, most commonly a lower second or third molar. Improved dental care has reduced the incidence of this disorder substantially. Infection in these areas leads to dysphagia, odynophagia, and “woody” edema in the sublingual region, forcing the tongue up and back with the potential for airway obstruction. Fever, dysarthria, and drooling also may occur, and patients may speak in a “hot potato” voice. Intubation or tracheostomy may be necessary to secure the airway, as asphyxiation is the most common cause of death. Patients should be admitted to the hospital and closely monitored during treatment with IV antibiotics directed against streptococci and oral anaerobes. Recommended agents include ampicillin/sulbactam, clindamycin, or high-dose penicillin plus metronidazole.

Septic thrombophlebitis of the internal jugular vein (*Lemierre disease*) is a rare anaerobic oropharyngeal infection caused predominantly by *F. necrophorum*. The illness typically starts as a sore throat (most commonly in adolescents and young adults), which may present as exudative tonsillitis or peritonsillar abscess. Infection of the deep pharyngeal tissue allows organisms to drain into the lateral pharyngeal space, which contains the carotid artery and internal jugular vein. Septic thrombophlebitis of the internal jugular vein can result, with associated pain, dysphagia, and unilateral neck swelling and stiffness. Sepsis usually

occurs 3–10 days after the onset of sore throat and is often coupled with metastatic infection to the lung and other distant sites, with pulmonary abscess or empyema. Occasionally, the infection can extend along the carotid sheath and into the posterior mediastinum, resulting in mediastinitis, or it can erode into the carotid artery, with the early sign of repeated small bleeds into the mouth. The mortality rate from these invasive infections can be as high as 50%. Treatment consists of IV antibiotics (clindamycin or ampicillin/sulbactam) and surgical drainage of any purulent collections. The concomitant use of anticoagulants to prevent embolization remains controversial and is not typically advised; both the risks and the benefits of their use must be carefully considered.

INFECTIONS OF THE LARYNX AND EPIGLOTTIS

LARYNGITIS

Laryngitis is defined as any inflammatory process involving the larynx and can be caused by a variety of infectious and noninfectious processes. The vast majority of laryngitis cases seen in clinical practice in developed countries are acute. Acute laryngitis is a common syndrome caused predominantly by the same viruses responsible for many other URIs. In fact, most cases of acute laryngitis occur in the setting of a viral URI.

Etiology

Nearly all major respiratory viruses have been implicated in acute viral laryngitis, including rhinovirus, influenza virus, parainfluenza virus, adenovirus, coxsackievirus, coronavirus, and RSV. Acute laryngitis can also be associated with acute bacterial respiratory infections such as those caused by group A *Streptococcus* or *C. diphtheriae* (although diphtheria has been virtually eliminated in the United States). Another bacterial pathogen thought to play a role (albeit unclear) in the pathogenesis of acute laryngitis is *M. catarrhalis*, which has been recovered from nasopharyngeal cultures in a significant percentage of cases.

Chronic laryngitis of infectious etiology is much less common in developed than in developing countries. Laryngitis due to *Mycobacterium tuberculosis* is often difficult to distinguish from laryngeal cancer, in part because of the frequent absence of signs, symptoms, and radiographic findings typical of pulmonary disease. *Histoplasma* and *Blastomyces* may cause laryngitis, often as a complication of systemic infection. *Candida* species can cause laryngitis as well, often in association with thrush or esophagitis and particularly in immunosuppressed patients. Rare cases of chronic laryngitis are due to *Coccidioides* and *Cryptococcus*.

Clinical Manifestations

Laryngitis is characterized by hoarseness and also can be associated with reduced vocal pitch or aphonia. As acute laryngitis is caused primarily by respiratory viruses, these symptoms usually occur in association with other symptoms and signs of URI, including rhinorrhea, nasal congestion, cough, and sore throat. Direct laryngoscopy often reveals diffuse laryngeal erythema and edema, along with vascular engorgement of the

vocal folds. In addition, chronic disease (e.g., tuberculous laryngitis) often includes mucosal nodules and ulcerations visible on laryngoscopy; these lesions are sometimes mistaken for laryngeal cancer.

TREATMENT

TREATMENT

Laryngitis

Acute laryngitis is usually treated with humidification and voice rest alone. Antibiotics are not recommended except when group A *Streptococcus* is cultured, in which case penicillin is the drug of choice. The choice of therapy for chronic laryngitis depends on the pathogen, whose identification usually requires biopsy with culture. Patients with laryngeal tuberculosis are highly contagious because of the large number of organisms that are easily aerosolized. These patients should be managed in the same way as patients with active pulmonary disease.

CROUP

The term *croup* actually denotes a group of diseases collectively referred to as “croup syndrome,” all of which are acute and predominantly viral respiratory illnesses characterized by marked swelling of the subglottic region of the larynx. Croup primarily affects children <6 years old. For a detailed discussion of this entity, the reader should consult a textbook of pediatric medicine.

EPIGLOTTITIS

Acute epiglottitis (supraglottitis) is an acute, rapidly progressive form of cellulitis of the epiglottis and adjacent structures that can result in complete—and potentially fatal—airway obstruction in both children and adults. Before the widespread use of *H. influenzae* type b (Hib) vaccine, this entity was much more common among children, with a peak incidence at ~3.5 years of age. In some countries, mass vaccination against Hib has reduced the annual incidence of acute epiglottitis in children by >90%; in contrast, the annual incidence in adults has changed little since the introduction of Hib vaccine. Because of the danger of airway obstruction, acute epiglottitis constitutes a medical emergency, particularly in children, and prompt diagnosis and airway protection are of the utmost importance.

Etiology

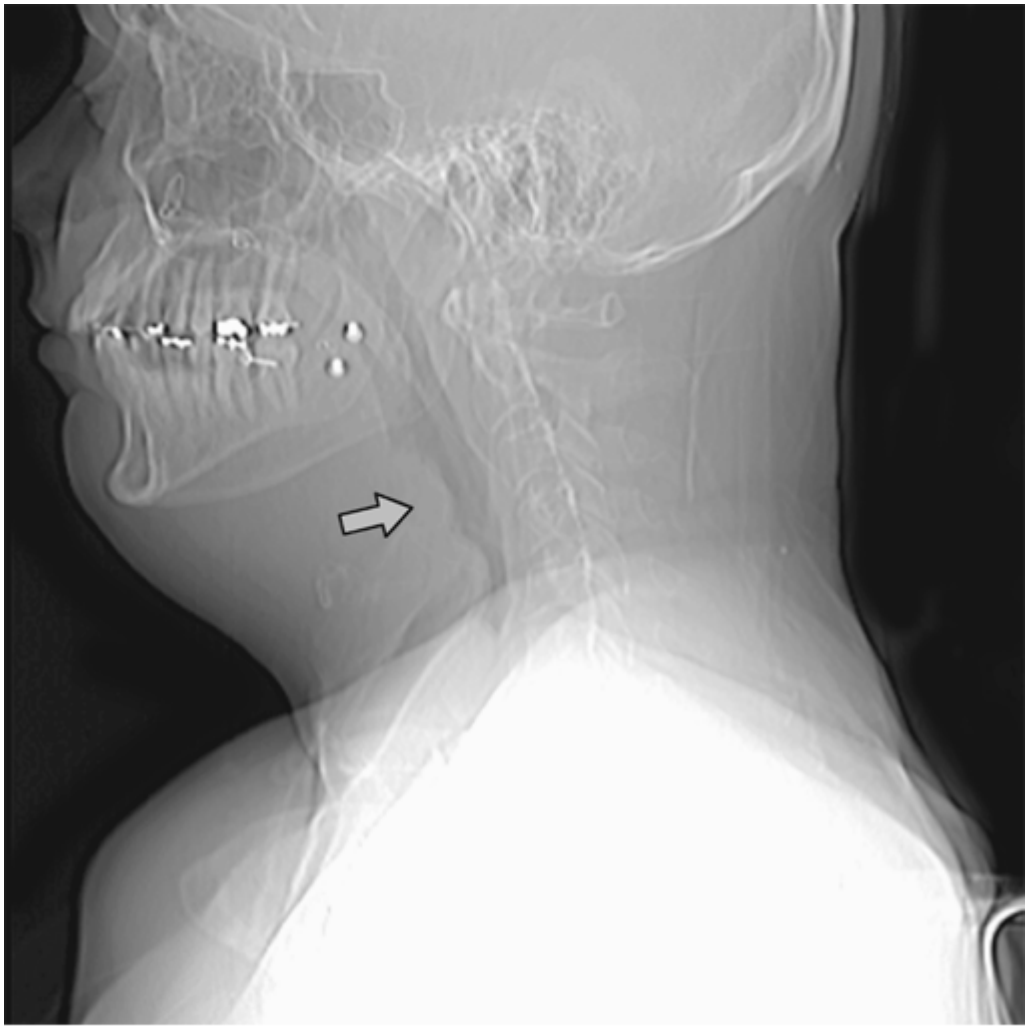
After the introduction of the Hib vaccine in the mid-1980s, disease incidence among children in the United States declined dramatically. Nevertheless, lack of vaccination or vaccine failure has meant that many pediatric cases seen today are still due to Hib. In adults and (more recently) in children, a variety of other bacterial pathogens have been associated with epiglottitis, the most common being group A *Streptococcus*. Other pathogens—seen less frequently— include *S. pneumoniae*, *Haemophilus parainfluenzae*, and *S. aureus* (including MRSA). Viruses have not been established as causes of acute epiglottitis.

Clinical Manifestations and Diagnosis

Epiglottitis typically presents more acutely in young children than in adolescents or adults. On presentation, most children have had symptoms for <24 h, including high fever, severe sore throat, tachycardia, systemic toxicity, and (in many cases) drooling while sitting forward. Symptoms and signs of respiratory obstruction also may be present and may progress rapidly. The somewhat milder illness in adolescents and adults often follows 1–2 days of severe sore throat and is commonly accompanied by dyspnea, drooling, and stridor. Physical examination of patients with acute epiglottitis may reveal moderate or severe respiratory distress, with inspiratory stridor and retractions of the chest wall. These findings *diminish* as the disease progresses and the patient tires. Conversely, oropharyngeal examination reveals infection that is much less severe than would be predicted from the symptoms—a finding that should alert the clinician to a cause of symptoms and obstruction that lies beyond the tonsils. The diagnosis often is made on clinical grounds, although direct fiberoptic laryngoscopy is frequently performed in a controlled environment (e.g., an operating room) to visualize and culture the typical edematous “cherry-red” epiglottis and facilitate placement of an endotracheal tube. Direct visualization in an examination room (i.e., with a tongue blade and indirect laryngoscopy) is not recommended because of the risk of immediate laryngospasm and complete airway obstruction. Lateral neck radiographs and laboratory tests can assist in the diagnosis but may delay the critical securing of the airway and cause the patient to be moved or repositioned more than is necessary, thereby increasing the risk of further airway compromise. Neck radiographs typically reveal an enlarged edematous epiglottis (the “thumbprint sign,” Fig. 31-3), usually with a dilated hypopharynx and normal subglottic structures. Laboratory tests characteristically document mild to moderate leukocytosis with a predominance of neutrophils. Blood cultures are positive in a significant proportion of cases.

FIGURE 31-3

Acute epiglottitis. In this lateral soft-tissue radiograph of the neck, the arrow indicates the enlarged edematous epiglottis (the “thumbprint sign”).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Epiglottitis

Security of the airway is always of primary concern in acute epiglottitis, even if the diagnosis is only suspected. Mere observation for signs of impending airway obstruction is not routinely recommended, particularly in children. Many adults have been managed with observation only since the illness is perceived to be milder in this age group, but some data suggest that this approach may be risky and probably should be reserved only for adult patients who have yet to develop dyspnea or stridor. Once the airway has been secured and specimens of blood and epiglottis tissue have been obtained for culture, treatment with IV antibiotics should be given to cover the most likely organisms, particularly *H. influenzae*. Because rates of ampicillin resistance in this organism have risen significantly in recent years, therapy with a β-lactam/β-lactamase inhibitor combination or a third-generation cephalosporin is recommended. Typically, ampicillin/sulbactam, cefotaxime, or ceftriaxone is given, with clindamycin and trimethoprim-sulfamethoxazole reserved for patients allergic to β-lactams. Antibiotic therapy should be continued for 7–10

days and should be tailored to the organism recovered in culture. If the household contacts of a patient with *H. influenzae* epiglottitis include an unvaccinated child aged <4 years, all members of the household (including the patient) should receive prophylactic rifampin for 4 days to eradicate carriage of *H. influenzae*.

INFECTIONS OF DEEP NECK STRUCTURES

Deep neck infections are usually extensions of infection from other primary sites, most often within the pharynx or oral cavity. Many of these infections are life-threatening but are difficult to detect at early stages, when they may be more easily managed. Three of the most clinically relevant spaces in the neck are the submandibular (and sublingual) space, the lateral pharyngeal (or parapharyngeal) space, and the retropharyngeal space. These spaces communicate with one another and with other important structures in the head, neck, and thorax, providing pathogens with easy access to areas that include the mediastinum, carotid sheath, skull base, and meninges. Once infection reaches these sensitive areas, mortality rates can be as high as 20–50%.

Infection of the submandibular and/or sublingual space typically originates from an infected or recently extracted lower tooth. The result is the severe, life-threatening infection referred to as Ludwig angina (see "Oral Infections," above). Infection of the lateral pharyngeal (or parapharyngeal) space is most often a complication of common infections of the oral cavity and upper respiratory tract, including tonsillitis, peritonsillar abscess, pharyngitis, mastoiditis, and periodontal infection. This space, situated deep in the lateral wall of the pharynx, contains a number of sensitive structures, including the carotid artery, internal jugular vein, cervical sympathetic chain, and portions of cranial nerves IX through XII; at its distal end, it opens into the posterior mediastinum. Involvement of this space with infection can therefore be rapidly fatal. Examination may reveal some tonsillar displacement, trismus, and neck rigidity, but swelling of the lateral pharyngeal wall can easily be missed. The diagnosis can be confirmed by CT. Treatment consists of airway management, operative drainage of fluid collections, and at least 10 days of IV therapy with an antibiotic active against streptococci and oral anaerobes (e.g., ampicillin/sulbactam). A particularly severe form of this infection involving the components of the carotid sheath (postanginal septicemia, Lemierre disease) is described above (see "Oral Infections"). Infection of the retropharyngeal space also can be extremely dangerous, as this space runs posterior to the pharynx from the skull base to the superior mediastinum. Infections in this space are more common among children <5 years old because of the presence of several small retropharyngeal lymph nodes that typically atrophy by age 4 years. Infection is usually a consequence of extension from another site of infection—most commonly, acute pharyngitis. Other sources include otitis media, tonsillitis, dental infections, Ludwig angina, and anterior extension of vertebral osteomyelitis. Retropharyngeal space infection also can follow penetrating trauma to the posterior pharynx (e.g., from an endoscopic procedure). Infections are commonly polymicrobial, involving a mixture of aerobes and anaerobes; group A β-hemolytic streptococci and *S. aureus* are the most common pathogens. *M. tuberculosis* was a common cause in the past but now is rarely involved in the United States.

Patients with retropharyngeal abscess typically present with sore throat, fever, dysphagia, and neck pain and are often drooling because of difficulty and pain with swallowing. Examination may reveal tender cervical

adenopathy, neck swelling, and diffuse erythema and edema of the posterior pharynx as well as a bulge in the posterior pharyngeal wall that may not be obvious on routine inspection. A soft-tissue mass is usually demonstrable by lateral neck radiography or CT. Because of the risk of airway obstruction, treatment begins with securing of the airway, which is followed by a combination of surgical drainage and IV antibiotic administration. Initial empirical therapy should cover streptococci, oral anaerobes, and *S. aureus*; ampicillin/sulbactam, clindamycin plus ceftriaxone, or meropenem is usually effective. Complications result primarily from extension to other areas (e.g., rupture into the posterior pharynx may lead to aspiration pneumonia and empyema). Extension may also occur to the lateral pharyngeal space and mediastinum, resulting in mediastinitis and pericarditis, or into nearby major blood vessels. All these events are associated with a high mortality rate.

FURTHER READING

Brook I: Microbiology of chronic rhinosinusitis. Eur j Clin Microbiol Infect Dis 35:1059, 2016.

Fletcher-Lartey S et al: Why do general practitioners prescribe antibiotics for upper respiratory tract infections to meet patient expectations: A mixed methods study. BMJ Open 6:e012244, 2016.

Jensen A et al: *Fusobacterium necrophorum* tonsillitis: An important cause of tonsillitis in adolescents and young adults. Clin Microbiol Infect 21:266.e1, 2015. [PubMed: 25636943]

Lee GC et al: Outpatient antibiotic prescribing in the United States: 2000 to 2010. BMC Med 12:96, 2014.

Chapter 32: Oral Manifestations of Disease

Samuel C. Durso

INTRODUCTION

As primary care physicians and consultants, internists are often asked to evaluate patients with disease of the oral soft tissues, teeth, and pharynx. Knowledge of the oral milieu and its unique structures is necessary to guide preventive services and recognize oral manifestations of local or systemic disease (Chap. A2). Furthermore, internists frequently collaborate with dentists in the care of patients who have a variety of medical conditions that affect oral health or who undergo dental procedures that increase their risk of medical complications.

DISEASES OF THE TEETH AND PERIODONTAL STRUCTURES

Tooth formation begins during the sixth week of embryonic life and continues through 17 years of age. Teeth start to develop in utero and continue to develop until after the tooth erupts. Normally, all 20 deciduous teeth have erupted by age 3 and have been shed by age 13. Permanent teeth, eventually totaling 32, begin to erupt by age 6 and have completely erupted by age 14, though third molars (“wisdom teeth”) may erupt later.

The erupted tooth consists of the visible *crown* covered with enamel and the root submerged below the gum line and covered with bonelike *cementum*. *Dentin*, a material that is denser than bone and exquisitely sensitive to pain, forms the majority of the tooth substance, surrounding a core of myxomatous *pulp* containing the vascular and nerve supply. The tooth is held firmly in the alveolar socket by the *periodontium*, supporting structures that consist of the gingivae, alveolar bone, cementum, and periodontal ligament. The periodontal ligament tenaciously binds the tooth’s cementum to the alveolar bone. Above this ligament is a collar of attached gingiva just below the crown. A few millimeters of unattached or free gingiva (1–3 mm) overlap the base of the crown, forming a shallow sulcus along the gum-tooth margin.

Dental Caries, Pulpal and Periapical Disease, and Complications

Dental caries usually begin asymptomatically as a destructive infectious process of the enamel. Bacteria—principally *Streptococcus mutans*—colonize the organic buffering biofilm (*plaque*) on the tooth surface. If not removed by brushing or by the natural cleansing and antibacterial action of saliva, bacterial acids can demineralize the enamel. Fissures and pits on the occlusal surfaces are the most frequent sites of early decay. Surfaces between the teeth, adjacent to tooth restorations and exposed roots, are also vulnerable, particularly as individuals age. Over time, dental caries extend to the underlying dentin, leading to cavitation

of the enamel. Without management, the caries will penetrate to the tooth pulp, producing *acute pulpitis*. At this stage, when the pulp infection is limited, the tooth may become sensitive to percussion and to hot or cold, and pain resolves immediately when the irritating stimulus is removed. Should the infection spread throughout the pulp, *irreversible pulpitis* occurs, leading to *pulp necrosis*. At this later stage, pain can be severe and has a sharp or throbbing visceral quality that may be worse when the patient lies down. Once pulp necrosis is complete, pain may be constant or intermittent, but cold sensitivity is lost.

Treatment of caries involves removal of the softened and infected hard tissue and restoration of the tooth structure with silver amalgam, glass ionomer, composite resin, or gold. Once irreversible pulpitis occurs, root canal therapy becomes necessary; removal of the contents of the pulp chamber and root canal is followed by thorough cleaning and filling with an inert material. Alternatively, the tooth may be extracted.

Pulpal infection leads to *periapical abscess* formation, which can produce pain on chewing. If the infection is mild and chronic, a *periapical granuloma* or eventually a *periapical cyst* forms, either of which produces radiolucency at the root apex. When unchecked, a periapical abscess can erode into the alveolar bone, producing osteomyelitis; penetrate and drain through the gingivae, producing a parulis (gumboil); or track along deep fascial planes, producing virulent cellulitis (Ludwig's angina) involving the submandibular space and floor of the mouth **(Chap. 172)**. Elderly patients, patients with diabetes mellitus, and patients taking glucocorticoids may experience little or no pain or fever as these complications develop.

Periodontal Disease

Periodontal disease and dental caries are the primary causes of tooth loss. Like dental caries, chronic infection of the gingiva and anchoring structures of the tooth begins with formation of bacterial plaque. The process begins at the gum line. Plaque and *calculus* (calcified plaque) are preventable by appropriate daily oral hygiene, including periodic professional cleaning. Left undisturbed, chronic inflammation can ensue and produce hyperemia of the free and attached gingivae (*gingivitis*), which then typically bleed with brushing. If this issue is ignored, severe *periodontitis* can develop, leading to deepening of the physiologic sulcus and destruction of the periodontal ligament. Gingival pockets develop around the teeth. As the periodontium (including the supporting bone) is destroyed, the teeth loosen. A role for chronic inflammation due to chronic periodontal disease in promoting coronary heart disease and stroke has been proposed. Epidemiologic studies have demonstrated a moderate but significant association between chronic periodontal inflammation and atherogenesis, though a causal role remains unproven.

Acute and aggressive forms of periodontal disease are less common than the chronic forms described above. However, if the host is stressed or exposed to a new pathogen, rapidly progressive and destructive disease of the periodontal tissue can occur. A virulent example is *acute necrotizing ulcerative gingivitis*. Stress and poor oral hygiene are risk factors. The presentation includes sudden gingival inflammation, ulceration, bleeding, interdental gingival necrosis, and fetid halitosis. *Localized juvenile periodontitis*, which is seen in adolescents, is particularly destructive and appears to be associated with impaired neutrophil chemotaxis. *AIDS-related periodontitis* resembles acute necrotizing ulcerative gingivitis in some patients and a more destructive form of adult chronic periodontitis in others. It may also produce a gangrene-like destructive

process of the oral soft tissues and bone that resembles *noma*, an infectious condition seen in severely malnourished children in developing nations.

Prevention of Tooth Decay and Periodontal Infection

Despite the reduced prevalences of dental caries and periodontal disease in the United States (due in large part to water fluoridation and improved dental care, respectively), both diseases constitute a major public health problem worldwide, particularly in certain groups. The internist should promote preventive dental care and hygiene as part of health maintenance. Populations at high risk for dental caries and periodontal disease include those with hyposalivation and/or xerostomia, diabetics, alcoholics, tobacco users, persons with Down syndrome, and those with gingival hyperplasia. Furthermore, patients lacking access to dental care (e.g., as a result of low socioeconomic status) and patients with a reduced ability to provide self-care (e.g., individuals with disabilities, nursing home residents, and persons with dementia or upper-extremity disability) suffer at a disproportionate rate. It is important to provide counseling regarding regular dental hygiene and professional cleaning, use of fluoride-containing toothpaste, professional fluoride treatments, and (for patients with limited dexterity) use of electric toothbrushes and also to instruct persons caring for those who are not capable of self-care. Cost, fear of dental care, and differences in language and culture create barriers that prevent some people from seeking preventive dental services.

Developmental and Systemic Disease Affecting the Teeth and Periodontium

In addition to posing cosmetic issues, *malocclusion*, the most common developmental oral problem, can interfere with mastication unless corrected through orthodontic and surgical techniques. Impacted third molars are common and can become infected or erupt into an insufficient space. Acquired prognathism due to *acromegaly* may also lead to malocclusion, as may deformity of the maxilla and mandible due to *Paget's disease* of the bone. Delayed tooth eruption, a receding chin, and a protruding tongue are occasional features of *cretinism* and *hypopituitarism*. Congenital syphilis produces tapering, notched (*Hutchinson's*) incisors and finely nodular (*mulberry*) molar crowns. *Enamel hypoplasia* results in crown defects ranging from pits to deep fissures of primary or permanent teeth. Intrauterine infection (syphilis, rubella), vitamin deficiency (A, C, or D), disorders of calcium metabolism (malabsorption, vitamin D–resistant rickets, hypoparathyroidism), prematurity, high fever, and rare inherited defects (*amelogenesis imperfecta*) are all causes. Tetracycline, given in sufficiently high doses during the first 8 years of life, may produce enamel hypoplasia and discoloration. Exposure to endogenous pigments can discolor developing teeth; etiologies include *erythroblastosis fetalis* (green or bluish-black), congenital liver disease (green or yellow-brown), and porphyria (red or brown that fluoresces with ultraviolet light). *Mottled enamel* occurs if excessive fluoride is ingested during development. Worn enamel is seen with age, bruxism, or excessive acid exposure (e.g., chronic gastric reflux or bulimia). Celiac disease is associated with nonspecific enamel defects in children but not in adults.

Total or partial tooth loss resulting from periodontitis is seen with cyclic neutropenia, Papillon-Lefévre syndrome, Chédiak-Higashi syndrome, and leukemia. Rapid focal tooth loosening is most often due to infection, but rarer causes include Langerhans cell histiocytosis, Ewing's sarcoma, osteosarcoma, and

Burkitt's lymphoma. Early loss of primary teeth is a feature of *hypophosphatasia*, a rare congenital error of metabolism.

Pregnancy may produce gingivitis and localized *pyogenic granulomas*. Severe periodontal disease occurs in uncontrolled diabetes mellitus. *Gingival hyperplasia* may be caused by phenytoin, calcium channel blockers (e.g., nifedipine), and cyclosporine, though excellent daily oral care can prevent or reduce its occurrence. *Idiopathic familial gingival fibromatosis* and several syndrome-related disorders cause similar conditions. Discontinuation of the medication may reverse the drug-induced form, though surgery may be needed to control both of the latter entities. *Linear gingival erythema* is variably seen in patients with advanced HIV infection and probably represents immune deficiency and decreased neutrophil activity. Diffuse or focal gingival swelling may be a feature of early or late acute myelomonocytic leukemia as well as of other lymphoproliferative disorders. A rare but pathognomonic sign of granulomatosis with polyangiitis is a red-purplish, granular gingivitis (*strawberry gums*).

DISEASES OF THE ORAL MUCOSA

Infections

Most oral mucosal diseases involve microorganisms **(Table 32-1)**.

TABLE 32-1

Vesicular, Bullous, or Ulcerative Lesions of the Oral Mucosa

Condition	Usual Location	Clinical Features	Course
Viral Diseases			
Primary acute herpetic gingivostomatitis (HSV type 1; rarely type 2)	Lip and oral mucosa (buccal, gingival, lingual mucosa)	Labial vesicles that rupture and crust, and intraoral vesicles that quickly ulcerate; extremely painful; acute gingivitis, fever, malaise, foul odor, and cervical lymphadenopathy; occurs primarily in infants, children, and young adults	Heals spontaneously in 10–14 days; unless secondarily infected, lesions lasting >3 weeks are not due to primary HSV infection
Recurrent herpes labialis	Mucocutaneous junction of lip, perioral skin	Eruption of groups of vesicles that may coalesce, then rupture and crust; painful to pressure or spicy foods	Lasts ~1 week, but condition may be prolonged if secondarily infected; if severe, topical or oral antiviral treatment may reduce healing time
Recurrent intraoral herpes simplex	Palate and gingiva	Small vesicles on keratinized epithelium that rupture and coalesce; painful	Heals spontaneously in ~1 week; if severe, topical, or oral antiviral treatment may reduce healing time
Chickenpox (VZV)	Gingiva and oral mucosa	Skin lesions may be accompanied by small vesicles on oral mucosa that rupture to form shallow ulcers; may coalesce to form large bullous lesions that ulcerate; mucosa may have generalized erythema	Lesions heal spontaneously within 2 weeks

Condition	Usual Location	Clinical Features	Course
Herpes zoster (VZV reactivation)	Cheek, tongue, gingiva, or palate	Unilateral vesicular eruptions and ulceration in linear pattern following sensory distribution of trigeminal nerve or one of its branches	Gradual healing without scarring unless secondarily infected; postherpetic neuralgia is common; oral acyclovir, famciclovir, or valacyclovir reduces healing time and postherpetic neuralgia
Infectious mononucleosis (Epstein-Barr virus)	Oral mucosa	Fatigue, sore throat, malaise, fever, and cervical lymphadenopathy; numerous small ulcers usually appear several days before lymphadenopathy; gingival bleeding and multiple petechiae at junction of hard and soft palates	Oral lesions disappear during convalescence; no treatment is given, though glucocorticoids are indicated if tonsillar swelling compromises the airway
Herpangina (coxsackievirus A; also possibly coxsackievirus B and echovirus)	Oral mucosa, pharynx, tongue	Sudden onset of fever, sore throat, and oropharyngeal vesicles, usually in children <4 years old, during summer months; diffuse pharyngeal congestion and vesicles (1–2 mm), grayish-white surrounded by red areola; vesicles enlarge and ulcerate	Incubation period of 2–9 days; fever for 1–4 days; recovery uneventful
Hand-foot-and-mouth disease (most commonly coxsackievirus A16)	Oral mucosa, pharynx, palms, and soles	Fever, malaise, headache with oropharyngeal vesicles that become painful, shallow ulcers; highly infectious; usually affects children under age 10	Incubation period 2–18 days; lesions heal spontaneously in 2–4 weeks

Condition	Usual Location	Clinical Features	Course
Primary HIV infection	Gingiva, palate, and pharynx	Acute gingivitis and oropharyngeal ulceration, associated with febrile illness resembling mononucleosis and including lymphadenopathy	Followed by HIV seroconversion, asymptomatic HIV infection, and usually ultimately by HIV disease
Bacterial or Fungal Diseases			
Acute necrotizing ulcerative gingivitis (“trench mouth”)	Gingiva	Painful, bleeding gingiva characterized by necrosis and ulceration of gingival papillae and margins plus lymphadenopathy and foul breath	Debridement and diluted (1:3) peroxide lavage provide relief within 24 h; antibiotics in acutely ill patients; relapse may occur
Prenatal (congenital) syphilis	Palate, jaws, tongue, and teeth	Gummatous involvement of palate, jaws, and facial bones; Hutchinson’s incisors, mulberry molars, glossitis, mucous patches, and fissures at corner of mouth	Tooth deformities in permanent dentition irreversible
Primary syphilis (chancre)	Lesion appearing where organism enters body; may occur on lips, tongue, or tonsillar area	Small papule developing rapidly into a large, painless ulcer with indurated border; unilateral lymphadenopathy; chancre and lymph nodes containing spirochetes; serologic tests positive by third to fourth weeks	Healing of chancre in 1–2 months, followed by secondary syphilis in 6–8 weeks
Secondary syphilis	Oral mucosa frequently involved with mucous patches, which occur primarily on palate and also at commissures of mouth	Maculopapular lesions of oral mucosa, 5–10 mm in diameter with central ulceration covered by grayish membrane; eruptions occurring on various mucosal surfaces and skin, accompanied by fever, malaise, and sore throat	Lesions may persist from several weeks to a year

Condition	Usual Location	Clinical Features	Course
Tertiary syphilis	Palate and tongue	Gummatous infiltration of palate or tongue followed by ulceration and fibrosis; atrophy of tongue papillae produces characteristic bald tongue and glossitis	Gumma may destroy palate, causing complete perforation
Gonorrhea	Lesions may occur in mouth at site of inoculation or secondarily by hematogenous spread from a primary focus	Most pharyngeal infection is asymptomatic; may produce burning or itching sensation; oropharynx and tonsils may be ulcerated and erythematous; saliva viscous and fetid	More difficult to eradicate than urogenital infection, though pharyngitis usually resolves with appropriate antimicrobial treatment
Tuberculosis	Tongue, tonsillar area, soft palate	Painless, solitary, 1- to 5-cm, irregular ulcer covered with persistent exudate; ulcer has firm undermined border	Autoinoculation from pulmonary infection is usual; lesions resolve with appropriate antimicrobial therapy
Cervicofacial actinomycosis	Swellings in region of face, neck, and floor of mouth	Infection may be associated with extraction, jaw fracture, or eruption of molar tooth; in acute form, resembles acute pyogenic abscess, but contains yellow “sulfur granules” (gram-positive mycelia and their hyphae)	Typically, swelling is hard and grows painlessly; multiple abscesses with draining tracts develop; penicillin first choice; surgery usually necessary
Histoplasmosis	Any area of the mouth, particularly tongue, gingiva, or palate	Nodular, verrucous, or granulomatous lesions; ulcers are indurated and painful; usual source hematogenous or pulmonary, but may be primary	Systemic antifungal therapy necessary
Candidiasis[a]			
Dermatologic Diseases			

Condition	Usual Location	Clinical Features	Course
Mucous membrane pemphigoid	Typically produces marked gingival erythema and ulceration; other areas of oral cavity, esophagus, and vagina may be affected	Painful, grayish-white collapsed vesicles or bullae of full-thickness epithelium with peripheral erythematous zone; gingival lesions desquamate, leaving ulcerated area	Protracted course with remissions and exacerbations; involvement of different sites develops slowly; glucocorticoids may temporarily reduce symptoms but do not control disease
EM minor and EM major (Stevens-Johnson syndrome)	Primarily oral mucosa and skin of hands and feet	Intraoral ruptured bullae surrounded by inflammatory area; lips may show hemorrhagic crusts; “iris” or “target” lesion on skin is pathognomonic; patient may have severe signs of toxicity	Onset very rapid; usually idiopathic, but may be associated with trigger such as drug reaction; condition may last 3–6 weeks; mortality rate for untreated EM major is 5–15%
Pemphigus vulgaris	Oral mucosa and skin; sites of mechanical trauma (soft/hard palate, frenulum, lips, buccal mucosa)	Usually (>70%) presents with oral lesions; fragile, ruptured bullae and ulcerated oral areas; mostly in older adults	With repeated occurrence of bullae, toxicity may lead to cachexia, infection, and death within 2 years; often controllable with oral glucocorticoids
Lichen planus	Oral mucosa and skin	White striae in mouth; purplish nodules on skin at sites of friction; occasionally causes oral mucosal ulcers and erosive gingivitis	White striae alone usually asymptomatic; erosive lesions often difficult to treat, but may respond to glucocorticoids
Other Conditions			

Condition	Usual Location	Clinical Features	Course
Recurrent aphthous ulcers	Usually on nonkeratinized oral mucosa (buccal and labial mucosa, floor of mouth, soft palate, lateral and ventral tongue)	Single or clustered painful ulcers with surrounding erythematous border; lesions may be 1–2 mm in diameter in crops (herpetiform), 1–5 mm (minor), or 5–15 mm (major)	Lesions heal in 1–2 weeks but may recur monthly or several times a year; protective barrier with benzocaine and topical glucocorticoids relieve symptoms; systemic glucocorticoids may be needed in severe cases
Behçet's syndrome	Oral mucosa, eyes, genitalia, gut, and CNS	Multiple aphthous ulcers in mouth; inflammatory ocular changes, ulcerative lesions on genitalia; inflammatory bowel disease and CNS disease	Oral lesions often first manifestation; persist several weeks and heal without scarring
Traumatic ulcers	Anywhere on oral mucosa; dentures frequently responsible for ulcers in vestibule	Localized, discrete ulcerated lesions with red border; produced by accidental biting of mucosa, penetration by foreign object, or chronic irritation by dentures	Lesions usually heal in 7–10 days when irritant is removed, unless secondarily infected
Squamous cell carcinoma	Any area of mouth, most commonly on lower lip, lateral borders of tongue, and floor of mouth	Red, white, or red and white ulcer with elevated or indurated border; failure to heal; pain not prominent in early lesions	Invades and destroys underlying tissues; frequently metastasizes to regional lymph nodes

Condition	Usual Location	Clinical Features	Course
Acute myeloid leukemia (usually monocytic)	Gingiva	Gingival swelling and superficial ulceration followed by hyperplasia of gingiva with extensive necrosis and hemorrhage; deep ulcers may occur elsewhere on mucosa, complicated by secondary infection	Usually responds to systemic treatment of leukemia; occasionally requires local irradiation
Lymphoma	Gingiva, tongue, palate, and tonsillar area	Elevated, ulcerated area that may proliferate rapidly, giving appearance of traumatic inflammation	Fatal if untreated; may indicate underlying HIV infection
Chemical or thermal burns	Any area in mouth	White slough due to contact with corrosive agents (e.g., aspirin, hot cheese) applied locally; removal of slough leaves raw, painful surface	Lesion heals in several weeks if not secondarily infected

[a]See Table 32-3.

Abbreviations: CNS, central nervous system; EM, erythema multiforme; HSV, herpes simplex virus; VZV, varicella-zoster virus.

Pigmented Lesions

See **Table 32-2**.

TABLE 32-2

Pigmented Lesions of the Oral Mucosa

Condition	Usual Location	Clinical Features	Course
Oral melanotic macule	Any area of mouth	Discrete or diffuse, localized, brown to black macule	Remains indefinitely; no growth
Diffuse melanin pigmentation	Any area of mouth	Diffuse pale to dark-brown pigmentation; may be physiologic (“racial”) or due to smoking	Remains indefinitely
Nevi	Any area of mouth	Discrete, localized, brown to black pigmentation	Remains indefinitely
Malignant melanoma	Any area of mouth	Can be flat and diffuse, painless, brown to black; or can be raised and nodular	Expands and invades early; metastasis leads to death
Addison’s disease	Any area of mouth, but mostly buccal mucosa	Blotches or spots of bluish-black to dark-brown pigmentation occurring early in disease, accompanied by diffuse pigmentation of skin; other symptoms of adrenal insufficiency	Condition controlled by adrenal steroid replacement
Peutz-Jeghers syndrome	Any area of mouth	Dark-brown spots on lips, buccal mucosa, with characteristic distribution of pigment around lips, nose, and eyes and on hands; concomitant intestinal polyposis	Oral pigmented lesions remain indefinitely; gastrointestinal polyps may become malignant
Drug ingestion (neuroleptics, oral contraceptives, minocycline, zidovudine, quinine derivatives)	Any area of mouth	Brown, black, or gray areas of pigmentation	Gradually disappears following cessation of drug intake

Condition	Usual Location	Clinical Features	Course
Amalgam tattoo	Gingiva and alveolar mucosa	Small blue-black pigmented areas associated with embedded amalgam particles in soft tissues; may show up on radiographs as radiopaque particles in some cases	Remains indefinitely
Heavy metal pigmentation (bismuth, mercury, lead)	Gingival margin	Thin blue-black pigmented line along gingival margin; rarely seen except in children exposed to lead-based paint	Indicative of systemic absorption; no significance for oral health
Black hairy tongue	Dorsum of tongue	Elongation of filiform papillae of tongue, which become stained by coffee, tea, tobacco, or pigmented bacteria	Improves within 1–2 weeks with gentle brushing of tongue or (if due to bacterial overgrowth) discontinuation of antibiotic
Fordyce spots	Buccal and labial mucosa	Numerous small yellowish spots just beneath mucosal surface; no symptoms; due to hyperplasia of sebaceous glands	Benign; remains without apparent change
Kaposi’s sarcoma	Palate most common, but may occur at any other site	Red or blue plaques of variable size and shape; often enlarge, become nodular, and may ulcerate	Usually indicative of HIV infection or non-Hodgkin’s lymphoma; rarely fatal, but may require treatment for comfort or cosmesis
Mucous retention cysts	Buccal and labial mucosa	Bluish, clear fluid–filled cyst due to extravasated mucus from injured minor salivary gland	Benign; painless unless traumatized; may be removed surgically

Dermatologic Diseases

See Tables 32-1, 32-2, and **32-3** and **Chaps. 52,53,54,55,56,57**.

TABLE 32-3

White Lesions of Oral Mucosa

Condition	Usual Location	Clinical Features	Course
Lichen planus	Buccal mucosa, tongue, gingiva, and lips; skin	Striae, white plaques, red areas, ulcers in mouth; purplish papules on skin; may be asymptomatic, sore, or painful; lichenoid drug reactions may look similar	Protracted; responds to topical glucocorticoids
White sponge nevus	Oral mucosa, vagina, anal mucosa	Painless white thickening of epithelium; adolescence/early adulthood onset; familial	Benign and permanent
Smoker's leukoplakia and smokeless tobacco lesions	Any area of oral mucosa, sometimes related to location of habit	White patch that may become firm, rough, or red-fissured and ulcerated; may become sore and painful but is usually painless	May or may not resolve with cessation of habit; 2% of patients develop squamous cell carcinoma; early biopsy essential
Erythroplakia with or without white patches	Floor of mouth commonly affected in men; tongue and buccal mucosa in women	Velvety, reddish plaque; occasionally mixed with white patches or smooth red areas	High risk of squamous cell cancer; early biopsy essential

Condition	Usual Location	Clinical Features	Course
Candidiasis	Any area in mouth	*Pseudomembranous type* ("thrush"): creamy white curdlike patches that reveal a raw, bleeding surface when scraped; found in sick infants, debilitated elderly patients receiving high-dose glucocorticoids or broad-spectrum antibiotics, and patients with AIDS	Responds favorably to antifungal therapy and correction of predisposing causes where possible
		Erythematous type: flat, red, sometimes sore areas in same groups of patients	Course same as for pseudomembranous type
		Candidal leukoplakia: nonremovable white thickening of epithelium due to *Candida*	Responds to prolonged antifungal therapy
		Angular cheilitis: sore fissures at corner of mouth	Responds to topical antifungal therapy
Hairy leukoplakia	Usually on lateral tongue, rarely elsewhere on oral mucosa	White areas ranging from small and flat to extensive accentuation of vertical folds; found in HIV carriers (all risk groups for AIDS)	Due to Epstein-Barr virus; responds to high-dose acyclovir but recurs; rarely causes discomfort unless secondarily infected with *Candida*
Warts (human papillomavirus)	Anywhere on skin and oral mucosa	Single or multiple papillary lesions with thick, white, keratinized surfaces containing many pointed projections; cauliflower lesions covered with normal-colored mucosa or multiple pink or pale bumps (focal epithelial hyperplasia)	Lesions grow rapidly and spread; squamous cell carcinoma must be ruled out with biopsy; excision or laser therapy; may regress in HIV-infected patients receiving antiretroviral therapy

Diseases of the Tongue

See **Table 32-4**.

TABLE 32-4

Alterations of the Tongue

Type of Change	Clinical Features
Size or Morphology	
Macroglossia	Enlarged tongue that may be part of a syndrome found in developmental conditions such as Down syndrome, Simpson-Golabi-Behmel syndrome, or Beckwith-Wiedemann syndrome; may be due to tumor (hemangioma or lymphangioma), metabolic disease (e.g., primary amyloidosis), or endocrine disturbance (e.g., acromegaly or cretinism); may occur when all teeth are removed
Fissured ("scrotal") tongue	Dorsal surface and sides of tongue covered by painless shallow or deep fissures that may collect debris and become irritated
Median rhomboid glossitis	Congenital abnormality with ovoid, denuded area in median posterior portion of tongue; may be associated with candidiasis and may respond to antifungal treatment
Color	
"Geographic" tongue (benign migratory glossitis)	Asymptomatic inflammatory condition of tongue, with rapid loss and regrowth of filiform papillae leading to appearance of denuded red patches "wandering" across surface
Hairy tongue	Elongation of filiform papillae of medial dorsal surface area due to failure of keratin layer of papillae to desquamate normally; brownish-black coloration may be due to staining by tobacco, food, or chromogenic organisms
"Strawberry" and "raspberry" tongue	Appearance of tongue during scarlet fever due to hypertrophy of fungiform papillae as well as changes in filiform papillae

Type of Change	Clinical Features
“Bald” tongue	Atrophy may be associated with xerostomia, pernicious anemia, iron-deficiency anemia, pellagra, or syphilis; may be accompanied by painful burning sensation; may be an expression of erythematous candidiasis and respond to antifungal treatment

HIV Disease and AIDS

See Tables 32-1, 32-2, 32-3, and **32-5**; **Chap. 197;** and **Fig. 189-3**.

TABLE 32-5

Oral Lesions Associated with HIV Infection

Lesion Morphology	Etiologies
Papules, nodules, plaques	Candidiasis (hyperplastic and pseudomembranous)[a] Condyloma acuminatum (human papillomavirus infection) Squamous cell carcinoma (preinvasive and invasive) Non-Hodgkin's lymphoma[a] Hairy leukoplakia[a]
Ulcers	Recurrent aphthous ulcers[a] Angular cheilitis Squamous cell carcinoma Acute necrotizing ulcerative gingivitis[a] Necrotizing ulcerative periodontitis[a] Necrotizing ulcerative stomatitis Non-Hodgkin's lymphoma[a] Viral infection (herpes simplex, herpes zoster, cytomegalovirus infection) Infection caused by *Mycobacterium tuberculosis* or *Mycobacterium avium-intracellulare* Fungal infection (histoplasmosis, cryptococcosis, candidiasis, geotrichosis, aspergillosis) Bacterial infection (*Escherichia coli, Enterobacter cloacae, Klebsiella pneumoniae, Pseudomonas aeruginosa*) Drug reactions (single or multiple ulcers)
Pigmented lesions	Kaposi's sarcoma[a] Bacillary angiomatosis (skin and visceral lesions more common than oral) Zidovudine pigmentation (skin, nails, and occasionally oral mucosa) Addison's disease
Miscellaneous	Linear gingival erythema[a]

[a]Strongly associated with HIV infection.

Ulcers

Ulceration is the most common oral mucosal lesion. Although there are many causes, the host and the pattern of lesions, including the presence of organ system features, narrow the differential diagnosis (Table 32-1). Most acute ulcers are painful and self-limited. Recurrent aphthous ulcers and herpes simplex account for the majority. Persistent and deep aphthous ulcers can be idiopathic or can accompany HIV/AIDS. Aphthous lesions are often the presenting symptom in *Behçet's syndrome* **(Chap. 357)**. Similar-appearing, though less painful, lesions may occur in reactive arthritis, and aphthous ulcers are occasionally present during phases of *discoid* or *systemic lupus erythematosus* **(Chap. 353)**. Aphthous-like ulcers are seen in *Crohn's disease* **(Chap. 319)**, but, unlike the common aphthous variety, they may exhibit granulomatous inflammation on histologic examination. Recurrent aphthae are more prevalent in patients with *celiac disease* and have been reported to remit with elimination of gluten.

Of major concern are chronic, relatively painless ulcers and mixed red/white patches (erythroplakia and leukoplakia) of >2 weeks' duration. Squamous cell carcinoma and premalignant dysplasia should be considered early and a diagnostic biopsy performed. This awareness and this procedure are critically important because early-stage malignancy is vastly more treatable than late-stage disease. High-risk sites include the lower lip, floor of the mouth, ventral and lateral tongue, and soft palate–tonsillar pillar complex. Significant risk factors for oral cancer in Western countries include sun exposure (lower lip), tobacco and alcohol use, and human papillomavirus infection. In India and some other Asian countries, smokeless tobacco mixed with betel nut, slaked lime, and spices is a common cause of oral cancer. Rarer causes of chronic oral ulcer, such as tuberculosis, fungal infection, granulomatosis with polyangiitis, and midline granuloma may look identical to carcinoma. Making the correct diagnosis depends on recognizing other clinical features and performing a biopsy of the lesion. The syphilitic chancre is typically painless and therefore easily missed. Regional lymphadenopathy is invariably present. The syphilitic etiology is confirmed with appropriate bacterial and serologic tests.

Disorders of mucosal fragility often produce painful oral ulcers that fail to heal within 2 weeks. *Mucous membrane pemphigoid* and *pemphigus vulgaris* are the major acquired disorders. While their clinical features are often distinctive, a biopsy or immunohistochemical examination should be performed to diagnose these entities and to distinguish them from *lichen planus* and drug reactions.

Hematologic and Nutritional Disease

Internists are more likely to encounter patients with acquired, rather than congenital, bleeding disorders. Bleeding should stop 15 min after minor trauma and within an hour after tooth extraction if local pressure is applied. More prolonged bleeding, if not due to continued injury or rupture of a large vessel, should lead to investigation for a clotting abnormality. In addition to bleeding, petechiae and ecchymoses are prone to occur at the vibrating line between the soft and hard palates in patients with platelet dysfunction or thrombocytopenia.

All forms of leukemia, but particularly *acute myelomonocytic leukemia*, can produce gingival bleeding, ulcers, and gingival enlargement. Oral ulcers are a feature of agranulocytosis, and ulcers and mucositis are often severe complications of chemotherapy and radiation therapy for hematologic and other malignancies.

Plummer-Vinson syndrome (iron deficiency, angular stomatitis, glossitis, and dysphagia) raises the risk of oral squamous cell cancer and esophageal cancer at the postcricoidal tissue web. Atrophic papillae and a red, burning tongue may occur with pernicious anemia. Deficiencies in B-group vitamins produce many of these same symptoms as well as oral ulceration and cheilosis. Consequences of *scurvy* include swollen, bleeding gums; ulcers; and loosening of the teeth.

NONDENTAL CAUSES OF ORAL PAIN

Most, but not all, oral pain emanates from inflamed or injured tooth pulp or periodontal tissues. Nonodontogenic causes are often overlooked. In most instances, toothache is predictable and proportional to the stimulus applied, and an identifiable condition (e.g., caries, abscess) is found. Local anesthesia eliminates pain originating from dental or periodontal structures, but not referred pains. The most common nondental source of pain is myofascial pain referred from muscles of mastication, which become tender and ache with increased use. Many sufferers exhibit *bruxism* (grinding of the teeth) secondary to stress and anxiety. *Temporomandibular joint disorder* is closely related. It affects both sexes, with a higher prevalence among women. Features include pain, limited mandibular movement, and temporomandibular joint sounds. The etiologies are complex; malocclusion does not play the primary role once attributed to it. *Osteoarthritis* is a common cause of masticatory pain. Anti-inflammatory medication, jaw rest, soft foods, and heat provide relief. The temporomandibular joint is involved in 50% of patients with *rheumatoid arthritis*, and its involvement is usually a late feature of severe disease. Bilateral preauricular pain, particularly in the morning, limits range of motion.

Migrainous neuralgia may be localized to the mouth. Episodes of pain and remission without an identifiable cause and a lack of relief with local anesthesia are important clues. *Trigeminal neuralgia* (*tic douloureux*) can involve the entire branch or part of the mandibular or maxillary branch of the fifth cranial nerve and can produce pain in one or a few teeth. Pain may occur spontaneously or may be triggered by touching the lip or gingiva, brushing the teeth, or chewing. *Glossopharyngeal neuralgia* produces similar acute neuropathic symptoms in the distribution of the ninth cranial nerve. Swallowing, sneezing, coughing, or pressure on the tragus of the ear triggers pain that is felt in the base of the tongue, pharynx, and soft palate and may be referred to the temporomandibular joint. *Neuritis* involving the maxillary and mandibular divisions of the trigeminal nerve (e.g., maxillary sinusitis, neuroma, and leukemic infiltrate) is distinguished from ordinary toothache by the neuropathic quality of the pain. Occasionally, *phantom pain* follows tooth extraction. Pain and hyperalgesia behind the ear and on the side of the face in the day or so before facial weakness develops often constitute the earliest symptom of *Bell's palsy*. Likewise, similar symptoms may precede visible lesions of herpes zoster infecting the seventh nerve (*Ramsey-Hunt syndrome*) or trigeminal nerve. *Postherpetic neuralgia* may follow either condition. *Coronary ischemia* may produce pain exclusively in the face and jaw; as in typical angina pectoris, this pain is usually reproducible with increased myocardial demand. Aching in several upper molar or premolar teeth that is unrelieved by anesthetizing the teeth may point to *maxillary sinusitis*.

Giant cell arteritis is notorious for producing headache, but it may also produce facial pain or sore throat without headache. Jaw and tongue claudication with chewing or talking is relatively common. Tongue infarction is rare. Patients with subacute thyroiditis often experience pain referred to the face or jaw before the tenderness of the thyroid gland and transient hyperthyroidism are appreciated.

"Burning mouth syndrome" (*glossodynia*) occurs in the absence of an identifiable cause (e.g., vitamin B_{12} deficiency, iron deficiency, diabetes mellitus, low-grade *Candida* infection, food sensitivity, or subtle xerostomia) and predominantly affects postmenopausal women. The etiology may be neuropathic. Clonazepam, α-lipoic acid, and cognitive behavioral therapy have benefited some patients. Some cases associated with an angiotensin-converting enzyme inhibitor have remitted when treatment with the drug was discontinued.

DISEASES OF THE SALIVARY GLANDS

Saliva is essential to oral health. Its absence leads to dental caries, periodontal disease, and difficulties in wearing dental prostheses, masticating, and speaking. Its major components, water and mucin, serve as a cleansing solvent and lubricating fluid. In addition, saliva contains antimicrobial factors (e.g., lysozyme, lactoperoxidase, secretory IgA), epidermal growth factor, minerals, and buffering systems. The major salivary glands secrete intermittently in response to autonomic stimulation, which is high during a meal but low otherwise. Hundreds of minor glands in the lips and cheeks secrete mucus continuously throughout the day and night. Consequently, oral function becomes impaired when salivary function is reduced. The sensation of a dry mouth (*xerostomia*) is perceived when salivary flow is reduced by 50%. The most common etiology is medication, especially drugs with anticholinergic properties but also alpha and beta blockers, calcium channel blockers, and diuretics. Other causes include Sjögren's syndrome, chronic parotitis, salivary duct obstruction, diabetes mellitus, HIV/AIDS, and radiation therapy that includes the salivary glands in the field (e.g., for Hodgkin's lymphoma and for head and neck cancer). Management involves the elimination or limitation of drying medications, preventive dental care, and supplementation with oral liquid or salivary substitutes. Sugarless mints or chewing gum may stimulate salivary secretion if dysfunction is mild. When sufficient exocrine tissue remains, pilocarpine or cevimeline has been shown to increase secretions. Commercial saliva substitutes or gels relieve dryness. Fluoride supplementation is critical to prevent caries.

Sialolithiasis presents most often as painful swelling but in some instances as only swelling or only pain. Conservative therapy consists of local heat, massage, and hydration. Promotion of salivary secretion with mints or lemon drops may flush out small stones. Antibiotic treatment is necessary when bacterial infection in suspected. In adults, *acute bacterial parotitis* is typically unilateral and most commonly affects postoperative, dehydrated, and debilitated patients. *Staphylococcus aureus* (including methicillin-resistant strains) and anaerobic bacteria are the most common pathogens. Chronic bacterial *sialadenitis* results from lowered salivary secretion and recurrent bacterial infection. When suspected bacterial infection is not responsive to therapy, the differential diagnosis should be expanded to include benign and malignant neoplasms, lymphoproliferative disorders, Sjögren's syndrome, sarcoidosis, tuberculosis, lymphadenitis,

actinomycosis, and granulomatosis with polyangiitis. Bilateral nontender parotid enlargement occurs with diabetes mellitus, cirrhosis, bulimia, HIV/AIDS, and drugs (e.g., iodide, propylthiouracil).

Pleomorphic adenoma comprises two-thirds of all salivary neoplasms. The parotid is the principal salivary gland affected, and the tumor presents as a firm, slow-growing mass. Although this tumor is benign, its recurrence is common if resection is incomplete. Malignant tumors such as mucoepidermoid carcinoma, adenoid cystic carcinoma, and adenocarcinoma tend to grow relatively fast, depending upon grade. They may ulcerate and invade nerves, producing numbness and facial paralysis. Surgical resection is the primary treatment. Radiation therapy (particularly neutron-beam therapy) is used when surgery is not feasible and as post-resection for certain histologic types with a high risk of recurrence. Malignant salivary gland tumors have a 5-year survival rate of ~68%.

Dental Care for Medically Complex Patients

Routine dental care (e.g., uncomplicated extraction, scaling and cleaning, tooth restoration, and root canal) is remarkably safe. The most common concerns regarding care of dental patients with medical disease are excessive bleeding for patients taking anticoagulants, infection of the heart valves and prosthetic devices from hematogenous seeding by the oral flora, and cardiovascular complications resulting from vasopressors used with local anesthetics during dental treatment. Experience confirms that the risk of any of these complications is very low.

Patients undergoing tooth extraction or alveolar and gingival surgery rarely experience uncontrolled bleeding when warfarin anticoagulation is maintained within the therapeutic range currently recommended for prevention of venous thrombosis, atrial fibrillation, or mechanical heart valve. Embolic complications and death, however, have been reported during subtherapeutic anticoagulation. Therapeutic anticoagulation should be confirmed before and continued through the procedure. Likewise, low-dose aspirin (e.g., 81–325 mg) can safely be continued. For patients taking aspirin and another antiplatelet medication (e.g., clopidogrel), the decision to continue the second antiplatelet medication should be based on individual consideration of the risks of thrombosis and bleeding. The newer target-specific oral anticoagulants (dabigatran, apixaban, rivaroxaban, and edoxaban) are in increasingly common use. Simple extractions of 1–3 teeth, periodontal surgery, abscess drainage, and implant positioning do not typically require interruption of therapy. More extensive surgery may necessitate delaying or holding a dose of the anticoagulant or more elaborate measures to manage the risk of thrombosis and bleeding.

Patients at risk for bacterial endocarditis (Chap. 123) should maintain optimal oral hygiene, including flossing, and have regular professional cleanings. Currently, guidelines recommend that prophylactic antibiotics be restricted to those patients at high risk for bacterial endocarditis who undergo dental and oral procedures involving significant manipulation of gingival or periapical tissue or penetration of the oral mucosa. If unexpected bleeding occurs, antibiotics given within 2 h after the procedure provide effective prophylaxis.

Hematogenous bacterial seeding from oral infection can undoubtedly produce late prosthetic-joint infection and therefore requires removal of the infected tissue (e.g., drainage, extraction, root canal) and appropriate antibiotic therapy. However, evidence that late prosthetic-joint infection follows routine dental procedures is lacking. For this reason, antibiotic prophylaxis is generally not recommended before oral surgery or oral mucosal manipulation for patients who have undergone joint replacement surgery. Exceptions to this may be considered for patients who have experienced joint replacement complications.

Concern often arises regarding the use of vasoconstrictors to treat patients with hypertension and heart disease. Vasoconstrictors enhance the depth and duration of local anesthesia, thus reducing the anesthetic dose and potential toxicity. If intravascular injection is avoided, 2% lidocaine with 1:100,000 epinephrine (limited to a total of 0.036 mg of epinephrine) can be used safely in patients with controlled hypertension and stable coronary heart disease, arrhythmia, or congestive heart failure. Precautions should be taken with patients taking tricyclic antidepressants and nonselective beta blockers because these drugs may potentiate the effect of epinephrine.

Elective dental treatments should be postponed for at least 1 month and preferably for 6 months after myocardial infarction, after which the risk of reinfarction is low provided the patient is medically stable (e.g., stable rhythm, stable angina, and no heart failure). Patients who have suffered a stroke should have elective dental care deferred for 9 months. In both situations, effective stress reduction requires good pain control, including the use of the minimal amount of vasoconstrictor necessary to provide good hemostasis and local anesthesia.

Bisphosphonate therapy is associated with *osteonecrosis* of the jaw. However, the risk with oral bisphosphonate therapy is very low. Most patients affected have received high-dose aminobisphosphonate therapy for multiple myeloma or metastatic breast cancer and have undergone tooth extraction or dental surgery. Intraoral lesions, of which two-thirds are painful, appear as exposed yellow-white hard bone involving the mandible or maxilla. Screening tests for determining risk of osteonecrosis are unreliable. Patients slated for aminobisphosphonate therapy should receive preventive dental care that reduces the risk of infection and the need for future dentoalveolar surgery.

Halitosis

Halitosis typically emanates from the oral cavity or nasal passages. Volatile sulfur compounds resulting from bacterial decay of food and cellular debris account for the malodor. Periodontal disease, caries, acute forms of gingivitis, poorly fitting dentures, oral abscess, and tongue coating are common causes. Treatment includes correcting poor hygiene, treating infection, and tongue brushing. Hyposalivation can produce and exacerbate halitosis. Pockets of decay in the tonsillar crypts, esophageal diverticulum, esophageal stasis (e.g., achalasia, stricture), sinusitis, and lung abscess account for some instances. A few systemic diseases produce distinctive odors: renal failure (ammoniacal), hepatic (fishy), and ketoacidosis (fruity). *Helicobacter pylori* gastritis can also produce ammoniacal breath. If a patient presents because of concern about halitosis but no odor is detectable, then pseudohalitosis or halitophobia must be considered.

While tooth loss and dental disease are not normal consequences of aging, a complex array of structural and functional changes that occur with age can affect oral health. Subtle changes in tooth structure (e.g., diminished pulp space and volume, sclerosis of dentinal tubules, and altered proportions of nerve and vascular pulp content) result in the elimination or diminution of pain sensitivity and a reduction in the reparative capacity of the teeth. In addition, age-associated fatty replacement of salivary acini may reduce physiologic reserve, thus increasing the risk of hyposalivation. In healthy older adults, there is minimal, if any, reduction in salivary flow.

Poor oral hygiene often results when general health fails or when patients lose manual dexterity and upper-extremity flexibility. This situation is particularly common among frail older adults and nursing home residents and must be emphasized because regular oral cleaning and dental care reduce the incidence of pneumonia and oral disease as well as the mortality risk in this population. Other risks for dental decay include limited lifetime fluoride exposure. Without assiduous care, decay can become quite advanced yet remain asymptomatic. Consequently, much of a tooth—or the entire tooth—can be destroyed before the patient is aware of the process.

Periodontal disease, a leading cause of tooth loss, is indicated by loss of alveolar bone height. More than 90% of the U.S. population has some degree of periodontal disease by age 50. Healthy adults who have not had significant alveolar bone loss by the sixth decade of life do not typically experience significant worsening with advancing age.

With the passing of those born in the first half of the twentieth century, complete edentulousness in the United States is becoming increasingly restricted to impoverished populations. When it is present, speech, mastication, and facial contours are dramatically affected. Edentulousness may also exacerbate obstructive sleep apnea, particularly in asymptomatic individuals who wear dentures. Dentures can improve verbal articulation and restore diminished facial contours. Mastication can also be restored; however, patients expecting dentures to facilitate oral intake are often disappointed. Accommodation to dentures requires a period of adjustment. Pain can result from friction or traumatic lesions produced by loose dentures. Poor fit and poor oral hygiene may permit the development of candidiasis. This fungal infection may be either asymptomatic or painful and is suggested by erythematous smooth or granular tissue conforming to an area covered by the appliance. Individuals with dentures and no natural teeth need regular (annual) professional oral examinations.

FURTHER READING

Durso SC: Interaction with other health team members in caring for elderly patients. Dent Clin North Am 49:377, 2005. [PubMed: 15755411]

Elad S et al: Novel anticoagulants: General overview and practical considerations for dental practitioners. Oral Dis 22:23, 2016. [PubMed: 26386350]

Sollecito TP et al: The use of prophylactic antibiotics prior to dental procedures in patients with prosthetic joints: Evidence-based guidelines for dental practitioners—a report of the American Dental Association Council on Scientific Affairs. J Am Dent Assoc 146:11, 2015. [PubMed: 25569493]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 33: Dyspnea

Rebecca M. Baron

DYSPNEA

Definition: The American Thoracic Society consensus statement defines *dyspnea* as a "subjective experience of breathing discomfort that consists of qualitatively distinct sensations that vary in intensity. The experience derives from interactions among multiple physiological, psychological, social, and environmental factors and may induce secondary physiological and behavioral responses." Dyspnea, a symptom, can be perceived only by the person experiencing it and, therefore, must be self-reported. In contrast, signs of increased work of breathing, such as tachypnea, accessory muscle use, and intercostal retraction, can be measured and reported by clinicians.

Epidemiology: Dyspnea is a common, and it has been reported that up to one half of inpatients and one quarter of ambulatory patients experience dyspnea, with a prevalence of 9–13% in the community that increases to as high as 37% for adults aged ≥70 years. Dyspnea is a frequent cause for emergency room visits, accounting for as many as 3–4 million visits per year. Furthermore, it is increasingly appreciated that the degree of dyspnea may better predict outcomes in chronic obstructive pulmonary disease (COPD) than does the forced expiratory volume in 1 s (FEV1), and formal measures of dyspnea have been incorporated into the Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2017 COPD severity assessment guidelines. Dyspnea may also predict outcomes in other chronic heart and lung diseases as well. Dyspnea can arise from a diverse array of pulmonary, cardiac, and neurologic underlying causes, and elucidation of particular symptoms may point toward a specific etiology and/or mechanism driving dyspnea (although additional diagnostic testing is often required as will be further discussed below).

MECHANISMS UNDERLYING DYSPNEA

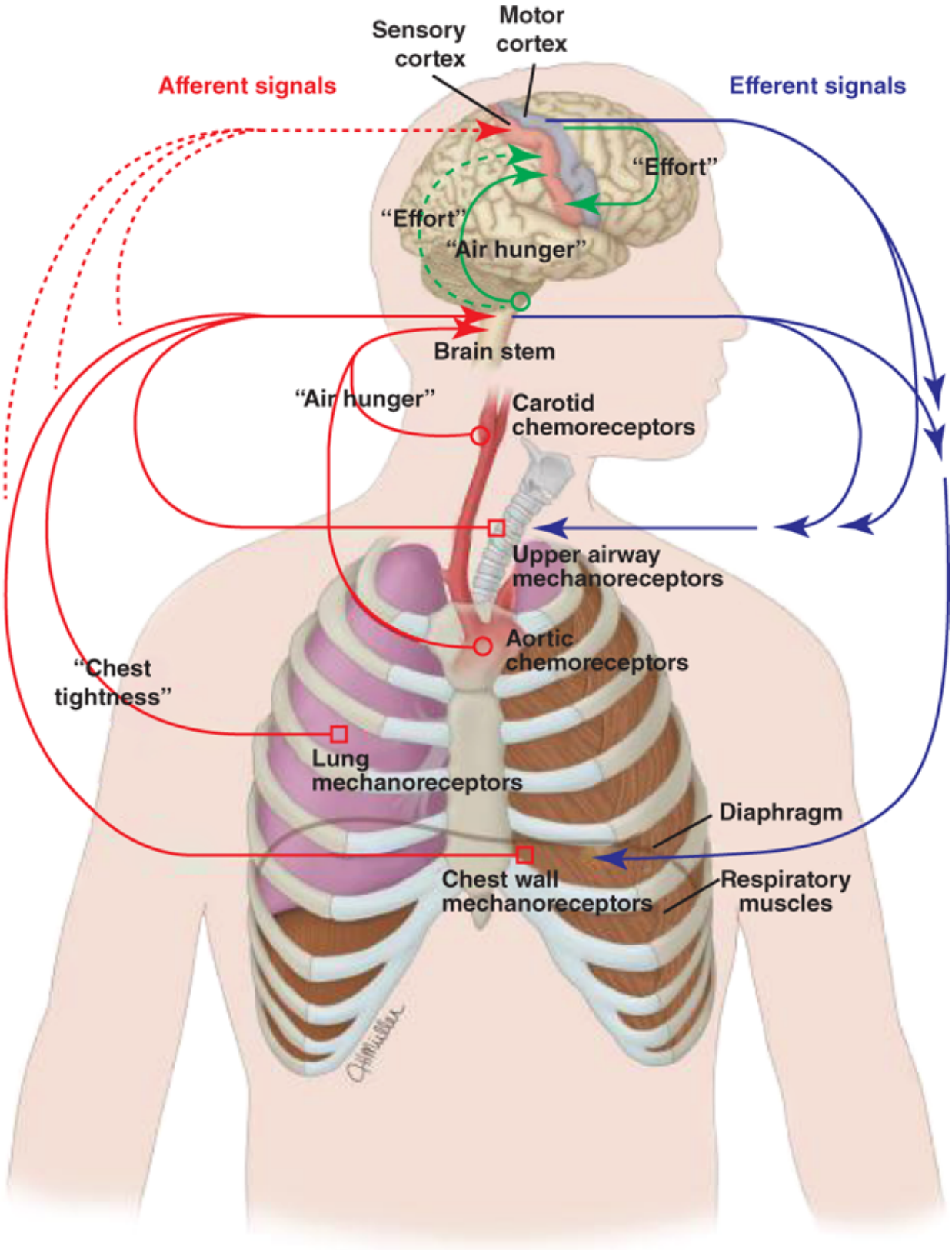
The mechanisms underlying dyspnea are complex, as it can arise from different contributory respiratory sensations. While a large body of research has increased our understanding of mechanisms underlying particular respiratory sensations such as "chest tightness" or "air hunger" it is likely that a given disease state might produce the sensation of dyspnea via more than one underlying mechanism. Dyspnea can arise from a variety of pathways, including generation of *afferent* signals from the respiratory system to the central nervous system (CNS), *efferent* signals from the CNS to the respiratory muscles, and particularly when there is a mismatch in the integrative signaling between these two pathways, termed "efferent-reafferent mismatch" (Fig. 33-1).

FIGURE 33-1

Signalling pathways underlying dyspnea. Dyspnea arises from a range of sensory inputs, many of which lead to distinct descriptive phrases used by patients (shown in quotes in the figure). The sensation of respiratory effort

likely arises from signals transmitted from the motor cortex to the sensory cortex (*green arrow*) when outgoing motor commands are sent to the ventilatory muscles (*efferent signals, blue arrow*). Motor output from the brain stem (*blue arrow*) may also be accompanied by signals transmitted to the sensory cortex and contribute to the sensation of effort (*dotted green arrow*). The sensation of air hunger probably derives from a combination of stimuli that increase the drive to breathe such as hypoxemia or hypercapnia (mediated by signals from chemoreceptors in the carotid body and aortic arch, indicated by afferent signals in red), acute hypercapnia or acidemia (mediated by signals from the peripheral and central chemoreceptors, indicated by afferent signals in red), airway and interstitial inflammation (mediated by pulmonary afferents, indicated by afferent signals in red), and pulmonary vascular receptors. Dyspnea arises in part from a perceived mismatch between the outgoing efferent messages to the ventilatory muscles and incoming afferent signals from the lungs and chest wall. Chest tightness, often associated with bronchospasm, is largely mediated by stimulation of vagal-irritant receptors. Afferent signals (*red arrows*) from airway, lung, and chest wall mechanoreceptors most likely pass through the brain stem before being transmitted to sensory cortex, although it is also possible that some afferent information bypasses the brain stem and goes directly to sensory cortex (*dotted arrow*).

Red arrows and text: afferent signals; Blue arrows and text: efferent signals; Green arrows: signals within the central nervous system; Dotted lines: hypothetical pathways; Hollow Red Circles: chemoreceptors; Hollow Red Squares: mechanoreceptors. *(Adapted from UpToDate 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Afferent signals trigger the CNS (brainstem and/or cortex) and include primarily: (a) peripheral chemoreceptors in the carotid body and aortic arch and central chemoreceptors in the medulla that are activated by hypoxemia, hypercapnia, or academia, and might produce a sense of "air hunger"; and (b) mechanoreceptors in the upper airways, lungs (including stretch receptors, irritant receptors, and J receptors), and chest wall (including muscle

spindles as stretch receptors and tendon organs that monitor force generation) that are activated in the setting of an increased work load from a disease state producing an increase in airway resistance that may be associated with symptoms of chest tightness (e.g., asthma or COPD) or decreased lung or chest wall compliance (e.g., pulmonary fibrosis). Other afferent signals that trigger dyspnea within the respiratory system can arise from pulmonary vascular receptor responses to changes in pulmonary artery pressure and skeletal muscle (termed metaboreceptors) that are believed to sense changes in the biochemical environment.

Efferent signals are sent from the CNS (motor cortex and brainstem) to the respiratory muscles, and are also transmitted by corollary discharge to the sensory cortex that are believed to underlie sensations of respiratory effort (or "work of breathing") and perhaps contribute to sensations of "air hunger," especially in response to an increased ventilatory load in a disease state such as COPD. In addition, fear or anxiety may heighten the sense of dyspnea through exacerbating the underlying physiologic disturbance in response to an increased respiratory rate or disordered breathing pattern.

ASSESSING DYSPNEA

While it is well appreciated that dyspnea is a difficult quality to reliably measure due to multiple relevant possible domains that can be measured (e.g., sensory-perceptual experience, affective distress, and symptom impact or burden), and there exist no uniformly agreed upon tools for dyspnea assessment, consensus opinion is that dyspnea should be formally assessed in a context most relevant and beneficial for patient management; furthermore, that the specific domains being measured are adequately described. There are a number of emerging tools that have been developed for formal dyspnea assessment. As an example, the GOLD 2017 criteria advocate use of a dyspnea assessment tool such as the Modified Medical Research Council Dyspnea Scale (MMRC, Table 33-1) to assess symptom/impact burden in COPD.

TABLE 33-1

An Example of a Clinical Method for Rating Dyspnea: The Modified Medical Research Council Dyspnea Scale[a]

Grade of Dyspnea	Description
0	Not troubled by breathlessness, except with strenuous exercise
1	Shortness of breath walking on level ground or with walking up a slight hill
2	Walks slower than people of similar age on level ground due to breathlessness, or has to stop to rest when walking at own pace on level ground
3	Stops to rest after walking 100 m or after walking a few minutes on level ground
4	Too breathless to leave the house, or breathless with activities of daily living (e.g., dressing/undressing)

[a]Which has been incorporated into the GOLD 2017 guidelines as a possible tool for rating dyspnea in COPD.

Source: Modified from DA Mahler, CK Wells: Evaluation of clinical methods for rating dyspnea. Chest 93:580, 1988.

DIFFERENTIAL DIAGNOSIS

This chapter focuses largely on chronic dyspnea, which is defined as symptoms lasting longer than 1 month and can arise from a broad array of different underlying conditions, most commonly attributable to pulmonary or cardiac conditions that account for as many as 85% of the underlying causes of dyspnea. However, as many as one-third of patients may have multifactorial reasons underlying dyspnea. Examples of a wide array of conditions that underlie dyspnea with possible mechanisms underlying the presenting symptoms are described in Table 33-2.

TABLE 33-2

Differential Diagnosis of Disease Processes Underlying Dyspnea

System	Type of Process	Example of Disease Process	Possible Presenting Dyspnea Symptoms	Possible Physical Findings	Possible Mechanisms Underlying Dyspnea	Initial Diagnostic Studies (and possible findings)
Pulmonary	Airways disease	Asthma, COPD	Chest tightness, tachypnea, increased WOB, air hunger, inability to get a deep breath	Wheezing, accessory muscle use, exertional hypoxemia (especially with COPD)	Increased WOB, hypoxemia, hypercapnia, stimulation of pulmonary receptors	Peak flow (reduced); Spirometry (OVD); CXR (hyper-inflation; loss of lung parenchyma in COPD)
	Parenchymal disease	Interstitial lung disease[a]	Air hunger, inability to get a deep breath	Dry end-inspiratory crackles, clubbing, exertional hypoxemia	Increased WOB, increased respiratory drive, hypoxemia, hypercapnia, stimulation of pulmonary receptors	Spirometry and lung volumes (RVD); CXR and chest CT (interstitial lung disease)
	Chest wall disease	Kyphoscoliosis, Neuromuscular (NM) weakness	Increased WOB, inability to get a deep breath	Decreased diaphragm excursion; atelectasis	Increased WOB; stimulation ofpulmonary receptors (if atelectasis is present)	Spirometry and lung volumes (RVD); MIP and MEPs (reduced in NM weakness)

System	Type of Process	Example of Disease Process	Possible Presenting Dyspnea Symptoms	Possible Physical Findings	Possible Mechanisms Underlying Dyspnea	Initial Diagnostic Studies (and possible findings)
Pulmonary and cardiac	Pulmonary vasculature	Pulmonary Hypertension	Tachypnea	Elevated R heart pressures, exertional hypoxemia	Increased respiratory drive, hypoxemia, stimulation of vascular receptors	Diffusion capacity (reduced); ECG; ECHO (to evaluate PA pressures)[b]
Cardiac	Left heart failure _____ Pericardial disease	Coronary artery disease, cardio-myopathy[c] _____ Restrictive pericarditis; Cardiac tamponade	Chest tightness, air hunger	Elevated L heart pressures; wet crackles on lung examination; pulsus paradoxus (pericardial disease)	Increased WOB and drive, hypoxemia, stimulation of vascular and pulmonary receptors^	Consider BNP testing in the acute setting; ECG, ECHO, may need stress testing and/or LHC
Other	Variable	Anemia Deconditioning Psychological	Exertional breathlessness Poor fitness Anxiety	Variable	Metabo-receptors (anemia, poor fitness); chemoreceptors (anaerobic metabolism from poor fitness); some subjects may have increased sensitivity to hypercapnia	Hematocrit for anemia; exclude other causes

[a]Differential diagnosis of interstitial lung disease includes idiopathic pulmonary fibrosis, collagen vascular disease, drug or occupation-induced pneumonitis, lymphangitic spread of malignancy; processes that are more alveolar rather than

interstitial in nature can also less commonly contribute to parenchymal lung disease underlying chronic dyspnea and include entities such as hypersensitivity pneumonitis, bronchiolitis obliterans organizing pneumonia, etc.

[b]Would additionally consider these patients for CT angiography to evaluate for presence of thromboemboli, ventilation/perfusion scanning to evaluate for the presence of chronic thromboembolic disease, and right heart catheterization (RHC) to further evaluate pulmonary hypertension.

[c]Diastolic dysfunction in the setting of a stiff left ventricle is often seen and contributes significantly to insidious dyspnea that can be difficult to treat.

^May stimulate metaboreceptors if cardiac output is sufficiently reduced to a result in a lactic acidosis.

Abbreviations: BNP, brain natriuretic peptide; COPD, chronic obstructive pulmonary disease; CT, computed tomography; CT angio, CT angiography; CXR, chest x-ray; ECHO, echocardiogram; ECG, electrocardiogram; LHC, left heart catheterization; MIP/MEP, maximal inspiratory and maximal expiratory pressures (obtained in the PFT laboratory); OVD, obstructive ventilatory defect; RVD, restrictive ventilatory defect; WOB, work of breathing.

Respiratory system causes include diseases of the airways (e.g., asthma and COPD), diseases of the parenchyma (more commonly interstitial lung diseases are seen in the setting of chronic dyspnea, but alveolar filling processes, such as hypersensitivity pneumonitis or bronchiolitis obliterans organizing pneumonia [BOOP], can also present with similar symptoms), diseases affecting the chest wall (e.g., bony abnormalities such as kyphoscoliosis, or neuromuscular weakness conditions such as amyotrophic lateral sclerosis), and diseases affecting the pulmonary vasculature (e.g., pulmonary hypertension that can arise from a variety of underlying causes, or chronic thromboemolic disease). Diseases affecting the cardiovascular system that can present with dyspnea include processes affecting left heart function, such as coronary artery disease and cardiomyopathy, as well as disease processes affecting the pericardium, including restrictive pericarditis and cardiac tamponade. Other conditions underyling dyspnea that might not directly emanate from the pulmonary or cardiovascular systems include anemia (thereby potentially affecting oxygen-carrying capacity), deconditioning, and psychological processes such as anxiety. Distinguishing between the myriad of underlying processes that might present with dyspnea can be challenging. A graded approach that begins with a history and physical examination, followed by selected laboratory testing that might then advance to additional diagnostics and potentially subspecialty referral may help elucidate the underlying cause of dyspnea. However, a substantial proportion of patients may have persistent dyspnea despite treatment for an underlying process, or may not have a specific underlying process identified that is driving the dyspnea.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Dyspnea

OVERALL

For patients with a known prior pulmonary, cardiac, or neuromuscular condition and worsening dyspnea, the initial focus of the evaluation will usually address determining whether the known condition has progressed or whether a new process has developed that is causing dyspnea (See Fig. 33-2). For patients without a prior known

potential cause of dyspnea, the initial evaluation will focus on determining an underlying etiology. Determining the underlying cause, if possible, is extremely important, as the treatment may vary dramatically based upon the predisposing condition. An initial history and physical examination remain fundamental to the evaluation followed by initial diagnostic testing as indicated that might prompt subspecialty referral (e.g., pulmonary, cardiology, neurology, sleep, and/or specialized dyspnea clinic) if the cause of dyspnea remains elusive (Fig. 33-2). As many as two-thirds of patients will require diagnostic testing beyond the initial clinical presentation.

HISTORY

The patient should be asked to describe in his/her own words what the discomfort feels like as well as the effect of position, infections, and environmental stimuli on the dyspnea, as descriptors may be helpful in pointing toward an etiology. For example, symptoms of chest tightness might suggest the possibility of bronchoconstriction, and the sensation of inability to take a deep breath may correlate with dynamic hyperinflation from COPD. Orthopnea is a common indicator of congestive heart failure (CHF), mechanical impairment of the diaphragm associated with obesity, or asthma triggered by esophageal reflux. Nocturnal dyspnea suggests CHF or asthma. Acute, intermittent episodes of dyspnea are more likely to reflect episodes of myocardial ischemia, bronchospasm, or pulmonary embolism, while chronic persistent dyspnea is more typical of COPD, interstitial lung disease, and chronic thromboembolic disease. Information on risk factors for drug-induced or occupational lung disease and for coronary artery disease should be elicited. Left atrial myxoma or hepatopulmonary syndrome should be considered when the patient complains of *platypnea*—i.e., dyspnea in the upright position with relief in the supine position.

PHYSICAL EXAMINATION

Initial vital signs might be helpful in pointing toward an underlying etiology in the context of the remainder of the evaluation. For example, the presence of fever might point toward an underlying infectious or inflammatory process; the presence of hypertension in the setting of a heart failure might point toward diastolic dysfunction; the presence of tachycardia might be associated with many different underlying processes including fever, cardiac dysfunction, and deconditioning; and the presence of resting hypoxemia suggests processes involving hypercapnia, ventilation-perfusion mismatch, shunt, or impairment in diffusion capacity might be involved. An exertional oxygen saturation should also be obtained as described below. The physical examination should begin during the interview of the patient. Inability of the patient to speak in full sentences before stopping to get a deep breath suggests a condition that leads to stimulation of the controller or impairment of the ventilatory pump with reduced vital capacity. Evidence of increased work of breathing (supraclavicular retractions; use of accessory muscles of ventilation; and the tripod position, characterized by sitting with the hands braced on the knees) is indicative of increased airway resistance or stiffness of the lungs and the chest wall. When measuring the vital signs, the physician should accurately assess the respiratory rate and measure the pulsus paradoxus **(Chap. 265)**; if the systolic pressure decreases by >10 mmHg, the presence of COPD, acute asthma, or pericardial disease should be considered. During the general examination, signs of anemia (pale conjunctivae), cyanosis, and cirrhosis (spider angiomata, gynecomastia) should be sought. Examination of the chest should focus on symmetry of movement; percussion (dullness is indicative of pleural effusion; hyperresonance is a sign of emphysema); and auscultation (wheezes, rhonchi, prolonged expiratory phase, and diminished breath sounds are clues to disorders of the airways; rales suggest interstitial edema or fibrosis). The cardiac examination should focus on signs of elevated right heart pressures (jugular venous distention, edema, accentuated pulmonic

component to the second heart sound); left ventricular dysfunction (S3 and S4 gallops); and valvular disease (murmurs). When examining the abdomen with the patient in the supine position, the physician should note whether there is paradoxical movement of the abdomen as well as the presence of increased respiratory distress in the supine position: inward motion during inspiration is a sign of diaphragmatic weakness, and rounding of the abdomen during exhalation is suggestive of pulmonary edema. Clubbing of the digits may be an indication of interstitial pulmonary fibrosis or bronchiectasis, and joint swelling or deformation as well as changes consistent with Raynaud's disease may be indicative of a collagen-vascular process that can be associated with pulmonary disease.

Patients should be asked to walk under observation with oximetry in order to reproduce the symptoms. The patient should be examined during and at the end of exercise for new findings that were not present at rest (e.g., presence of wheezing), and for changes in oxygen saturation.

CHEST IMAGING

After the history elicitation and the physical examination, a chest radiograph should be obtained if the diagnosis remains elusive. The lung volumes should be assessed: hyperinflation is consistent with obstructive lung disease, whereas low lung volumes suggest interstitial edema or fibrosis, diaphragmatic dysfunction, or impaired chest wall motion. The pulmonary parenchyma should be examined for evidence of interstitial disease, infiltrates, and emphysema. Prominent pulmonary vasculature in the upper zones indicates pulmonary venous hypertension, while enlarged central pulmonary arteries may suggest pulmonary arterial hypertension. An enlarged cardiac silhouette can point toward dilated cardiomyopathy or valvular disease. Bilateral pleural effusions are typical of CHF and some forms of collagen-vascular disease. Unilateral effusions raise the specter of carcinoma and pulmonary embolism but may also occur in heart failure or in the case of a parapneumonic effusion. CT of the chest is generally reserved for further evaluation of the lung parenchyma (interstitial lung disease) and possible pulmonary embolism if there remains diagnostic uncertainty.

LABORATORY STUDIES

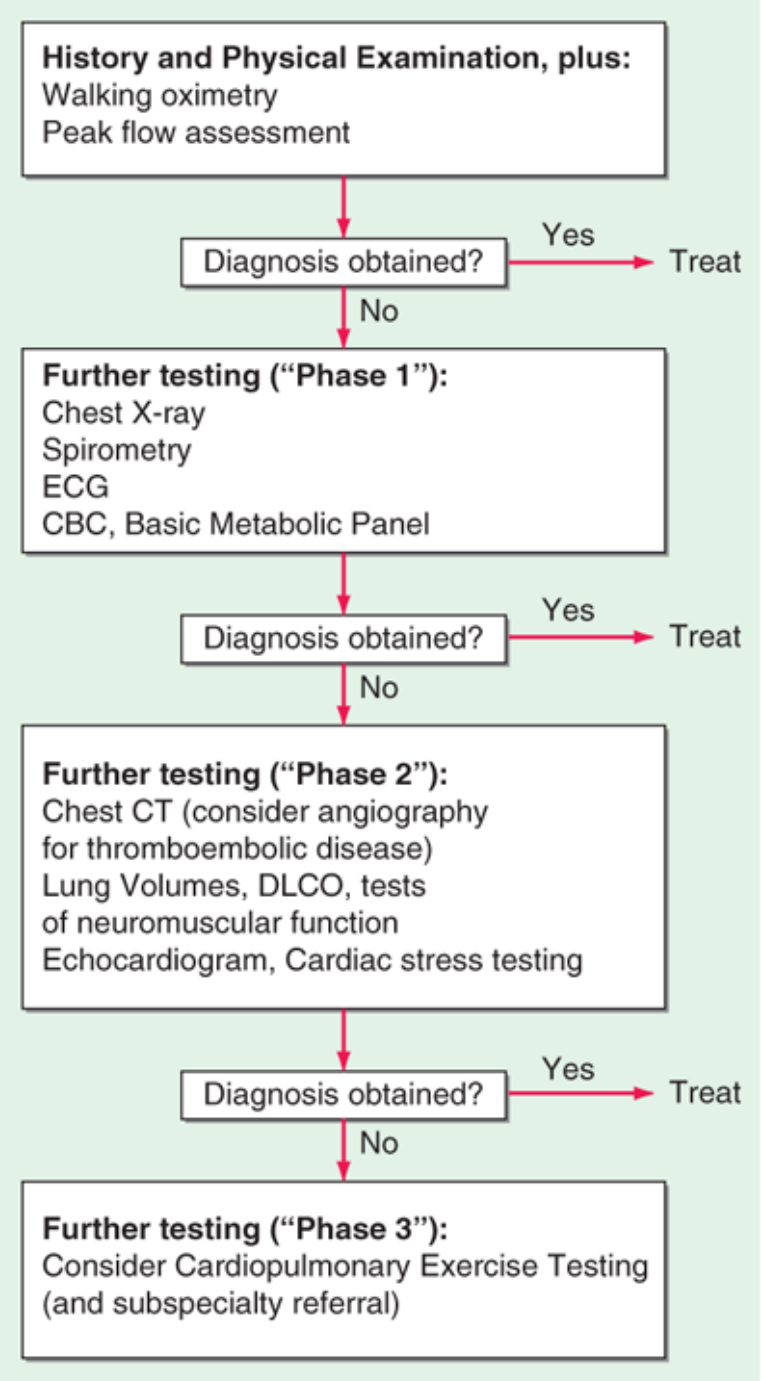
Initial laboratory testing should include a hematocrit to exclude occult anemia as an underlying cause of reduced oxygen-carrying capacity contributing to dyspnea, and a basic metabolic panel may be helpful to exclude a significant underlying metabolic acidosis (and conversely, an elevated bicarbonate might point toward the possibility of carbon dioxide retention that might be seen in chronic respiratory failure—in such a setting, an arterial blood gas may provide useful additional information). Additional laboratory studies should include electrocardiography to seek evidence of ventricular hypertrophy and prior myocardial infarction and spirometry that can be diagnostic of the presence of an obstructive ventilatory defect, and suggest the possibility of a restrictive ventilatory defect (that then might prompt additional pulmonary function laboratory testing, including lung volumes, diffusion capacity, and possible tests of neuromuscular function). Echocardiography is indicated when systolic dysfunction, pulmonary hypertension, or valvular heart disease is suspected. Bronchoprovocation testing and/or home peak-flow monitoring may be useful in patients with intermittent symptoms suggestive of asthma who have a normal physical examination and spirometry; up to one-third of patients with the clinical diagnosis of asthma do not have reactive airways disease when formally tested. Measurement of brain natriuretic peptide levels in serum is increasingly used to assess for CHF in patients presenting with acute dyspnea but may be elevated in the presence of right ventricular strain as well.

DISTINGUISHING CARDIOVASCULAR FROM RESPIRATORY SYSTEM DYSPNEA

If a patient has evidence of both pulmonary and cardiac disease that is either not responsive to treatment, or it remains unclear what factors are primarily driving dyspnea, a cardiopulmonary exercise test (CPET) can be carried out to determine which system is responsible for the exercise limitation. CPET includes incremental symptom-limited exercise (cycling or treadmill) with measurements of ventilation and pulmonary gas exchange, and in some cases includes non-invasive and invasive measures of pulmonary vascular pressures and cardiac output. If, at peak exercise, the patient achieves predicted maximal ventilation, demonstrates an increase in dead space or hypoxemia, or develops bronchospasm, the respiratory system may be the cause of the problem. Alternatively, if the heart rate is >85% of the predicted maximum, if the anaerobic threshold occurs early, if the blood pressure becomes excessively high or decreases during exercise, if the O_2 pulse (O_2 consumption/heart rate, an indicator of stroke volume) falls, or if there are ischemic changes on the electrocardiogram, an abnormality of the cardiovascular system is likely the explanation for the breathing discomfort. Additionally, a CPET may also help point toward a peripheral extraction deficit, or metabolic/neuromuscular disease as potential underlying processes driving dyspnea.

FIGURE 33-2

Possible algorithm for the evaluation of the patient with dyspnea. As described in the text, the approach should begin with a detailed history and physical examination, followed by progressive testing and ultimately more invasive testing and subspecialty referral as is indicated to determine the underlying cause of dyspnea. *(Adapted from NG Karnani et al: Am Fam Physician 71:1529, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Dyspnea

The first goal is to correct the underlying condition(s) driving dyspnea and address potentially reversible causes with appropriate treatment for the particular condition. Multiple different interventions may be necessary, given that dyspnea often arises from multifactorial causes. If relief of dyspnea with treatment of the underlying condition(s) is not fully possible, an effort is made to lessen the intensity of the symptom and its effect on the

patient's quality of life. Despite an increased understanding of the mechanisms underlying dyspnea, there has been limited progress in treatment strategies for dyspnea. Supplemental O_2 should be administered if the resting O_2 saturation is ≤88% or if the patient's saturation drops to these levels with activity or sleep. In particular, for patients with COPD, supplemental oxygen for those with hypoxemia has been shown to improve mortality, and pulmonary rehabilitation programs have demonstrated positive effects on dyspnea, exercise capacity, and rates of hospitalization. Opioids have been shown to reduce symptoms of dyspnea, largely through reducing air hunger, thus, likely suppressing respiratory drive and influencing cortical activity. However, opioids should be considered for each patient individually based upon the risk-benefit profile as regards the effects of respiratory depression. Studies of anxiolytics for dyspnea have not demonstrated consistent benefit. Additional approaches are under study for dyspnea, including inhaled furosemide that might alter afferent sensory information.

ACKNOWLEDGMENT

With prior contributions from Richard M. Schwartzstein.

FURTHER READING

Banzett RB et al: Multidimensional dyspnea profile: An instrument for clinical and laboratory research. Eur Respir J 45:1681, 2015. [PubMed: 25792641]

Laviolette L, Laveneziana P on behalf of the ERS Research Seminar Faculty: Dyspnoea: A multidimensional and multidisciplinary approach. Eur Respir J 43:1750, 2014. [PubMed: 24525437]

Parshall MB et al: An Official American Thoracic Society Statement: Update on the mechanisms, assessment, and management of dyspnea. Am J Respir Crit Care Med 185:435, 2012. [PubMed: 22336677]

Wahls SA: Causes and evaluation of chronic dyspnea. Am Fam Physician 86:173, 2012. [PubMed: 22962929]

Chapter 34: Cough

Christopher H. Fanta

FIGURE 34-1

COUGH

Cough performs an essential protective function for human airways and lungs. Without an effective cough reflex, we are at risk for retained airway secretions and aspirated material predisposing to infection, atelectasis, and respiratory compromise. At the other extreme, excessive coughing can be exhausting; can be complicated by emesis, syncope, muscular pain, or rib fractures; can aggravate low back pain, abdominal or inguinal hernias, and urinary incontinence; and can be a major impediment to social interactions. Cough is often a clue to the presence of respiratory disease. In many instances, cough is an expected and accepted manifestation of disease, as in acute respiratory tract infection. However, persistent cough in the absence of other respiratory symptoms commonly causes patients to seek medical attention.

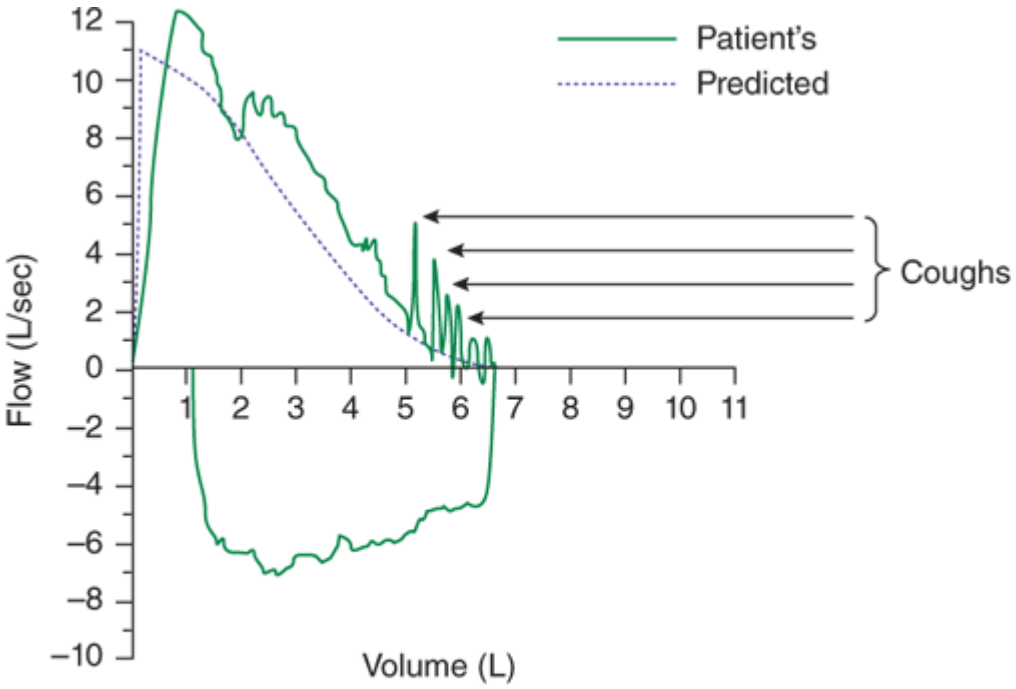
COUGH MECHANISM

Spontaneous cough is triggered by stimulation of sensory nerve endings that are thought to be primarily rapidly adapting receptors and C fibers. Both chemical (e.g., capsaicin) and mechanical (e.g., particulates in air pollution) stimuli may initiate the cough reflex. A cationic ion channel—the transient receptor potential vanilloid 1 (TRPV1)—found on rapidly adapting receptors and C fibers is the receptor for capsaicin, and its expression is increased in patients with chronic cough. Afferent nerve endings richly innervate the pharynx, larynx, and airways to the level of the terminal bronchioles and extend into the lung parenchyma. They may also be located in the external auditory meatus (the auricular branch of the vagus nerve, or Arnold's nerve) and in the esophagus. Sensory signals travel via the vagus and superior laryngeal nerves to a region of the brainstem in the nucleus tractus solitarius vaguely identified as the "cough center." The cough reflex involves a highly orchestrated series of involuntary muscular actions, with the potential for input from cortical pathways as well. The vocal cords adduct, leading to transient upper-airway occlusion. Expiratory muscles contract, generating positive intrathoracic pressures as high as 300 mmHg. With sudden release of the laryngeal contraction, rapid expiratory flows are generated, exceeding the normal "envelope" of maximal expiratory flow seen on the flow-volume curve (Fig. 34-1). Bronchial smooth-muscle contraction together with dynamic compression of airways narrows airway lumens and maximizes the velocity of exhalation. The kinetic energy available to dislodge mucus from the inside of airway walls is directly proportional to the square of the velocity of expiratory airflow. A deep breath preceding a cough optimizes the function of the

expiratory muscles; a series of repetitive coughs at successively lower lung volumes sweeps the point of maximal expiratory velocity progressively further into the lung periphery.

FIGURE 34-1

Flow-volume curve shows spikes of high expiratory flow achieved with cough.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IMPAIRED COUGH

Weak or ineffective cough compromises the ability to clear lower respiratory tract secretions, predisposing to more serious infections and their sequelae. Weakness or paralysis of the expiratory (abdominal and intercostal) muscles and pain in the chest wall or abdomen are foremost on the list of causes of impaired cough (Table 34-1). Cough strength is generally assessed qualitatively; peak expiratory flow or maximal expiratory pressure at the mouth can be used as a surrogate marker for cough strength. A variety of assistive devices and techniques have been developed to improve cough strength, running the gamut from simple (splinting of the abdominal muscles with a tightly held pillow to reduce postoperative pain while coughing) to complex (a mechanical cough-assist device supplied via face mask or tracheal tube that applies a cycle of positive pressure followed rapidly by negative pressure). Cough may fail to clear secretions despite a preserved ability to generate normal expiratory velocities; such failure may be due to either abnormal airway secretions (e.g., bronchiectasis due to cystic fibrosis) or structural abnormalities of the airways (e.g., tracheomalacia with excessive expiratory collapse of the trachea during cough).

TABLE 34-1

Causes of Impaired Cough

Decreased respiratory muscle strength
Chest wall or abdominal pain
Chest wall deformity (e.g., severe kyphoscoliosis)
Impaired glottic closure or tracheostomy
Tracheobronchomalacia
Abnormal airway secretions
Central respiratory depression (e.g., anesthesia, sedation, or coma)

SYMPTOMATIC COUGH

Cough may occur in the context of other respiratory symptoms that together point to a diagnosis; for example, cough accompanied by wheezing, shortness of breath, and chest tightness after exposure to a cat or other sources of allergens suggests asthma. At times, however, cough is the dominant or sole symptom of disease, and it may be of sufficient duration and severity that relief is sought. The duration of cough is a clue to its etiology, at least retrospectively. Acute cough (<3 weeks) is most commonly due to a respiratory tract infection, aspiration, or inhalation of noxious chemicals or smoke. Subacute cough (3–8 weeks in duration) is a common residuum of tracheobronchitis, as in pertussis or “postviral tussive syndrome.” Chronic cough (>8 weeks) may be caused by a wide variety of cardiopulmonary diseases, including those of inflammatory, infectious, neoplastic, and cardiovascular etiologies. When initial assessment with chest examination and radiography is normal, cough-variant asthma, gastroesophageal reflux, nasopharyngeal drainage, and medications (angiotensin-converting enzyme [ACE] inhibitors) are the most common identifiable causes of chronic cough. In a long-time cigarette smoker, an early-morning, productive cough suggests chronic bronchitis. A dry, irritative cough that lingers for >2 months following one or more respiratory tract infections (“post-bronchitic cough”) is a very common cause of chronic cough, especially in the winter months.

ASSESSMENT OF CHRONIC COUGH

Except for our ability to detect the sound of excess airway secretions, details as to the resonance of the cough, its time of occurrence during the day, and the pattern of coughing (e.g., occurring in paroxysms) infrequently provide useful etiologic clues. Regardless of cause, cough often worsens upon first lying down at

night, with talking, or with the hyperpnea of exercise; it frequently improves with sleep. An exception may involve the cough that occurs only with certain allergic exposures or exercise in cold air, as in asthma. Useful historical questions include what circumstances surrounded the onset of cough, what makes the cough better or worse, and does the cough produce sputum.

The physical examination seeks clues suggesting the presence of cardiopulmonary disease, including findings such as wheezing or crackles on chest examination. Examination of the auditory canals and tympanic membranes (for irritation of the latter resulting in stimulation of Arnold's nerve), the nasal passageways (for rhinitis or polyps), and the nails (for clubbing) may also provide etiologic clues. Because cough can be a manifestation of a systemic disease such as sarcoidosis or vasculitis, a thorough general examination is likewise important.

In virtually all instances, evaluation of chronic cough merits a chest radiograph. The list of diseases that can cause persistent cough without other symptoms and without detectable abnormalities on physical examination is long. It includes serious illnesses such as sarcoidosis or Hodgkin's disease in young adults, lung cancer in older patients, and (worldwide) pulmonary tuberculosis. An abnormal chest film prompts an evaluation aimed at explaining the radiographic abnormality. In a patient with chronic productive cough, examination of expectorated sputum is warranted, because determining the cause of mucus hypersecretion is critically important. Purulent-appearing sputum should be sent for routine bacterial culture and, in certain circumstances, mycobacterial culture as well. Cytologic examination of mucoid sputum may be useful to assess for malignancy and oropharyngeal aspiration and to distinguish neutrophilic from eosinophilic bronchitis. Expectoration of blood—whether streaks of blood, blood mixed with airway secretions, or pure blood—deserves a special approach to assessment and management.

CHRONIC COUGH WITH A NORMAL CHEST RADIOGRAPH

It is commonly held that (alone or in combination) the use of an ACE inhibitor; postnasal drainage; gastroesophageal reflux; and asthma account for >90% of cases of chronic cough with a normal or noncontributory chest radiograph. However, clinical experience does not support this contention, and strict adherence to this concept discourages the search for alternative explanations by both clinicians and researchers. In recent years, the concept of a distinct "cough hypersensitivity syndrome" has emerged, emphasizing the putative role of sensitized sensory nerve endings and afferent neural pathways in causing chronic refractory cough, akin to chronic neuropathic pain. It presents with a dry or minimally productive cough and a tickle or sensitivity in the throat, made worse with talking, laughing, or exertion. It is more common in women than men and can last for years. Specific diagnostic criteria are lacking; the diagnosis is suspected when alternative etiologies are excluded by diagnostic testing or failed therapeutic trials. It is uncertain whether persistent daily coughing elicits an inflammatory response and is thereby self-perpetuating.

ACE inhibitor–induced cough occurs in 5–30% of patients taking these agents and is not dose-dependent. ACE metabolizes bradykinin and other tachykinins, such as substance P. The mechanism of ACE inhibitor–associated cough may involve sensitization of sensory nerve endings due to accumulation of bradykinin. Any

patient with chronic unexplained cough who is taking an ACE inhibitor should have a trial period off the medication, regardless of the timing of the onset of cough relative to the initiation of ACE inhibitor therapy. In most instances, a safe alternative is available; angiotensin-receptor blockers do not cause cough. Failure to observe a decrease in cough after 1 month off medication argues strongly against this etiology. Postnasal drainage of any etiology can cause cough as a response to stimulation of sensory receptors of the cough-reflex pathway in the hypopharynx or aspiration of draining secretions into the trachea. Clues suggesting this etiology include postnasal drip, frequent throat clearing, and sneezing and rhinorrhea. On speculum examination of the nose, excess mucoid or purulent secretions, inflamed and edematous nasal mucosa, and/or polyps may be seen; in addition, secretions or a cobblestoned appearance of the mucosa along the posterior pharyngeal wall may be noted. Unfortunately, there is no means by which to quantitate postnasal drainage. In many instances, this diagnosis must rely on subjective information provided by the patient. This assessment must also be counterbalanced by the fact that many people who have chronic postnasal drainage do not experience cough.

Linking gastroesophageal reflux to chronic cough poses similar challenges. It is thought that reflux of gastric contents into the lower esophagus may trigger cough via reflex pathways initiated in the esophageal mucosa. Reflux to the level of the pharynx (laryngopharyngeal reflux), with consequent aspiration of gastric contents, causes a chemical bronchitis and possibly pneumonitis that can elicit cough for days afterward, but it is a rare finding among persons with chronic cough. Retrosternal burning after meals or on recumbency, frequent eructation, hoarseness, and throat pain may be indicative of gastroesophageal reflux. Nevertheless, reflux may also elicit minimal or no symptoms. Glottic inflammation detected on laryngoscopy may be a manifestation of recurrent reflux to the level of the throat, but it is a nonspecific finding. Quantification of the frequency and level of reflux requires a somewhat invasive procedure to measure esophageal pH (either nasopharyngeal placement of a catheter with a pH probe into the esophagus for 24 h or endoscopic placement of a radiotransmitter capsule into the esophagus) and, with newer techniques, non-acid reflux. The precise interpretation of test results that permits an etiologic linking of reflux events and cough remains debated. Again, assigning the cause of cough to gastroesophageal reflux must be weighed against the observation that many people with symptomatic reflux do not experience chronic cough.

Cough alone as a manifestation of asthma is common among children but not among adults. Cough due to asthma in the absence of wheezing, shortness of breath, and chest tightness is referred to as “cough-variant asthma.” A history suggestive of cough-variant asthmaties the onset of cough to exposure to typical triggers for asthma and the resolution of cough to discontinuation of exposure. Objective testing can establish the diagnosis of asthma (airflow obstruction on spirometry that varies over time or reverses in response to a bronchodilator) or exclude it with certainty (a negative response to a bronchoprovocation challenge—e.g., with methacholine). In a patient capable of taking reliable measurements, home expiratory peak flow monitoring can be a cost-effective method to support or discount a diagnosis of asthma.

Chronic eosinophilic bronchitis causes chronic cough with a normal chest radiograph. This condition is characterized by sputum eosinophilia in excess of 3% without airflow obstruction or bronchial hyperresponsiveness and is successfully treated with inhaled glucocorticoids.

Treatment of chronic cough in a patient with a normal chest radiograph is often empirical and is targeted at the most likely cause(s) of cough as determined by history, physical examination, and possibly pulmonary-function testing. Therapy for postnasal drainage depends on the presumed etiology (infection, allergy, or vasomotor rhinitis) and may include systemic antihistamines; decongestants; antibiotics; nasal saline irrigation; and nasal pump sprays with glucocorticoids, antihistamines, or anticholinergics. Antacids, histamine type 2 (H2) receptor antagonists, and proton-pump inhibitors are used to neutralize or decrease the production of gastric acid in gastroesophageal reflux disease; dietary changes, elevation of the head and torso during sleep, and medications to improve gastric emptying are additional therapeutic measures. Cough-variant asthma typically responds well to inhaled glucocorticoids and intermittent use of inhaled β-agonist bronchodilators.

Patients who fail to respond to treatment targeting the common causes of chronic cough or who have had these causes excluded by appropriate diagnostic testing should undergo chest CT. Diseases causing cough that may be missed on chest x-ray include tumors, early interstitial lung disease, bronchiectasis, and atypical mycobacterial pulmonary infection. On the other hand, patients with chronic cough who have normal findings on chest examination, lung function testing, oxygenation assessment, and chest CT can be reassured as to the absence of serious pulmonary pathology.

GLOBAL CONSIDERATIONS


Regular exposure to air pollution can cause chronic cough and throat clearing, as well as lower respiratory tract disease. Smoke from cooking and heating fuels in poorly ventilated homes; toxic exposures in work settings lacking implementation of occupational safety standards; and ambient chemicals and particulates in highly polluted outdoor air are all forms of air pollution causing cough. Limited therapeutic options are available; treatment focuses on improving environmental air quality (e.g., use of a stove chimney in the home), removal from the exposure, and use of an appropriate face mask.

SYMPTOM-BASED TREATMENT OF COUGH

Empiric treatment of chronic idiopathic cough with inhaled corticosteroids, inhaled anticholinergic bronchodilators, and macrolide antibiotics has been tried without consistent success. Currently available cough suppressants are only modestly effective. Most potent are narcotic cough suppressants, such as codeine or hydrocodone, which are thought to act in the “cough center” in the brainstem. The tendency of narcotic cough suppressants to cause drowsiness and constipation and their potential for addictive dependence limit their appeal for long-term use. *Dextromethorphan* is an over-the-counter, centrally acting cough suppressant with fewer side effects and less efficacy than the narcotic cough suppressants. Dextromethorphan is thought to have a different site of action than narcotic cough suppressants and can be used in combination with them if necessary. *Benzonatate* is thought to inhibit neural activity of sensory nerves in the cough-reflex pathway. It is generally free of side effects; however, its effectiveness in suppressing cough is variable and unpredictable. Attempts to treat cough hypersensitivity syndrome have focused on inhibition of neural pathways. Small case series and randomized clinical trials have indicated

benefit from off-label use of gabapentin, pregabalin, or amitriptyline. Recent studies suggest a role for behavioral modification using specialized speech therapy techniques, but widespread application of this modality is currently not practical. Novel cough suppressants without the limitations of currently available agents are greatly needed. Approaches that are being explored include the development of neurokinin receptor antagonists, TRPV1 ion channel antagonists, and novel opioid and opioid-like receptor agonists.

FURTHER READING

Brightling CE et al: Eosinophilic bronchitis as an important cause of chronic cough. Am J Respir Crit Care Med 160:406, 1999. [PubMed: 10430705]

Gibson PG, Vertigan AE: Management of chronic refractory cough. BMJ 351:h5590, 2015. [PubMed: 26666537]

Kahrilas PJ et al: Chronic cough due to gastroesophageal reflux in adults: CHEST Guideline and Expert Panel Report. Chest 150:1341, 2016. [PubMed: 27614002]

Ramsay LE et al: Double-blind comparison of losartan, lisinopril and hydrochlorothiazide in hypertensive patients with previous angiotensin converting enzyme inhibitor-associated cough. J Hypertens Suppl 13:S73, 1995. [PubMed: 18800460]

Ryan NM et al: Gabapentin for refractory chronic cough: a randomized, double-blind, placebo-controlled trial. Lancet 380:1583, 2012. [PubMed: 22951084]

Smith JA, Woodcock A: Chronic cough. N Engl J Med 375:1544, 2016. [PubMed: 27797316]

Chapter 35: Hemoptysis

Anna K. Brady; Patricia A. Kritek

FIGURE 35-1

INTRODUCTION

Hemoptysis is the expectoration of blood from the respiratory tract. The first step in evaluation is to ascertain whether the bleeding is coming from the respiratory tree or instead originating from the nasal cavities (i.e., epistaxis) or the gastrointestinal tract (i.e., hematemesis) as the therapies for these etiologies will be significantly different. Once established as hemoptysis, the exact nature of the expectoration is important as the term can be applied to blood-tinged phlegm, the pink frothy sputum of pulmonary edema, or frank blood. Next steps include identifying the source and etiology of bleeding.

ANATOMY AND PHYSIOLOGY OF HEMOPTYSIS

Hemoptysis can arise from anywhere in the respiratory tract; from the glottis to the alveolus. Most commonly, bleeding arises from the bronchi or medium sized airways, but a thorough evaluation of the entire respiratory tree is often necessary.

A unique feature of the lung that predisposes to hemoptysis of varied severity is its dual blood supply—the pulmonary and bronchial circulations. The former is a low-pressure system that is essential to gas exchange at the alveolar level; in contrast, the bronchial arteries originate from the aorta and are under systemic pressure. The bronchial arteries supply the airways and have the ability to neovascularize tumors, dilate airways of bronchiectasis, and cavitary lesions. Most hemoptysis is due to vessels in the bronchial circulation and is, therefore, under systemic pressure, making it more challenging to arrest the bleeding.

ETIOLOGY

Hemoptysis commonly results from infection, malignancy, or vascular disease; however, the differential for bleeding from the respiratory tree is varied and broad.

Infections

Most blood-tinged sputum and small-volume hemoptysis is due to viral bronchitis. Patients with chronic bronchitis are at risk for bacterial superinfection with organisms such as *Streptococcus pneumoniae*,

Haemophilus influenzae, or *Moraxella catarrhalis*, increasing airway inflammation and potential for bleeding. Similarly, patients with bronchiectasis are prone to hemoptysis with exacerbations of disease. Due to recurrent bacterial infection, bronchiectatic airways are dilated, inflamed, and highly vascular, supplied by the bronchial circulation. In several case series, bronchiectasis is the leading cause of massive hemoptysis and subsequent death.

Tuberculosis had long been the most common cause of hemoptysis worldwide, but it is now surpassed in industrialized countries by bronchitis and bronchiectasis. In patients with tuberculosis, development of cavitary disease is frequently the source of bleeding but rarer complications such as the erosion of a pulmonary artery aneurysm into a preexisting cavity (i.e., Rasmussen's aneurysm) can also be the source.

Other infectious agents such as endemic fungi, *Nocardia,* and non-tuberculous mycobacteria can present as cavitary lung disease complicated by hemoptysis. In addition, *Aspergillus* species can develop into mycetomas within preexisting cavities, with neovascularization to these inflamed spaces leading to bleeding. Pulmonary abscesses and necrotizing pneumonia can cause bleeding by devitalizing lung parenchyma. Common responsible organisms include *Staphylococcus aureus, Klebsiella pneumoniae*, and oral anaerobes.

 Paragonimiasis can mimic tuberculosis and is another significant cause of hemoptysis seen globally; it is common in Southeast Asia and China, although cases have been reported in North America from raw crayfish ingestion. It should be considered as a cause of hemoptysis in recent immigrants from endemic areas.

Vascular

Hemoptysis commonly results from pulmonary edema due to elevated left ventricular end-diastolic pressure. While the classic description of the sputum expectorated in pulmonary edema is "pink and frothy," a spectrum of hemoptysis including frank blood can be seen.

A pulmonary embolism with parenchymal infarction can present with hemoptysis, although most pulmonary emboli do not cause hemoptysis and will present with other signs and symptoms. An ectatic vessel in an airway or a pulmonary arteriovenous malformation can be a source of bleeding. While rare, rupture of an aortobronchial fistula can result in massive bleeding and sudden death; these fistulae arise in the setting of aortic pathology such as aneurysm or pseudoaneurysm and can cause small bleeding episodes that herald massive hemoptysis.

Diffuse alveolar hemorrhage (DAH), despite causing significant bleeding into the lung parenchyma, uncommonly results in hemoptysis. A range of insults cause DAH, including immune-mediated capillaritis from diseases such as systemic lupus erythematosus, toxicity from cocaine and other inhalants, and stem cell transplantation. The so-called "pulmonary-renal" syndromes, including granulomatosis with polyangiitis and anti-glomerular basement membrane disease, may lead to both hemoptysis and hematuria (though one manifestation may be present without the other). DAH more commonly presents with diffuse ground glass opacities on imaging and anemia, so the absence of hemoptysis should not exclude the diagnosis.

Malignancy

Bronchogenic carcinoma of any histology is a common cause of hemoptysis (both massive and non-massive) in modern published series. Hemoptysis often indicates airway involvement of the tumor and can be a presenting symptom of carcinoid tumors, vascular lesions that frequently arise in the proximal airways. Small cell and squamous cell carcinomas are frequently central in nature and more likely to erode into major pulmonary vessels, resulting in massive hemoptysis. Pulmonary metastases from distant tumors (e.g., melanoma, sarcoma, adenocarcinomas of the breast and colon) can also cause bleeding. Kaposi's sarcoma, seen in advanced acquired immunodeficiency syndrome, is very vascular and can develop anywhere along the respiratory tract, from the bronchi to the oral cavity.

Mechanical and Other Causes

In addition to infection, vascular disease, and malignancy, other insults to the pulmonary system can cause hemoptysis. Pulmonary endometriosis causes cyclical bleeding known as catamenial hemoptysis. Foreign body aspiration can lead to airway irritation and bleeding. Diagnostic and therapeutic procedures are also potential offenders: pulmonary vein stenosis can result from left atrial procedures, such as pulmonary vein isolation, and pulmonary artery catheters can lead to rupture of the pulmonary artery if the distal balloon is kept inflated. Finally, in the setting of thrombocytopenia, coagulopathy, anticoagulation, or antiplatelet therapy, even minor insults can cause hemoptysis.

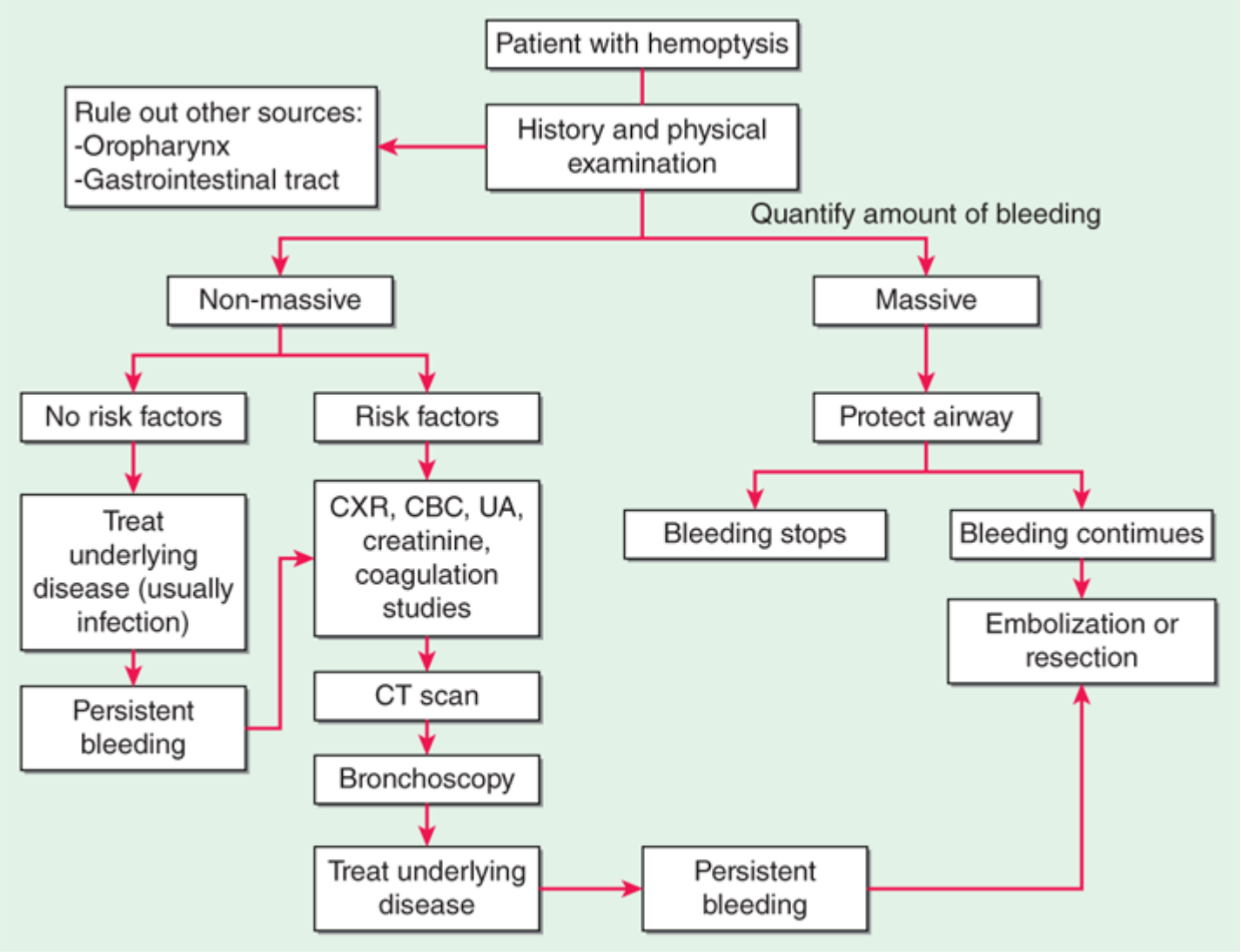
EVALUATION AND MANAGEMENT

History

The first step in evaluating hemoptysis is to determine the amount or severity of bleeding. A patient's description of the sputum (e.g., flecks of blood, pink-tinged, or frank blood or clot) is helpful if you cannot examine it. An approach to management of hemoptysis is outlined in **Fig. 35-1**.

FIGURE 35-1

Approach to the management of hemoptysis. CBC, complete blood count; CT, computed tomography; CXR, chest x-ray; UA, urinalysis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

It is crucial to determine whether the amount of blood expectorated is *massive*; while there is no agreed-upon volume, blood loss of 400 mL in 24 hours or 100–150 mL expectorated at one time are considered *massive hemoptysis*. These numbers derive from the volume of the tracheobronchial tree (generally 100–200 mL). This determination is clinically important as patients rarely die of exsanguination and, instead, are at risk of death due to asphyxiation from blood filling the airways and airspaces. Most patients cannot describe the volume of their hemoptysis in mL, so using referents like cups (one U.S. cup is 236 mL) can be helpful. Fortunately, massive hemoptysis only accounts for 5–15% of cases of hemoptysis.

Careful history may point to the cause of hemoptysis. Fever, chills, or antecedent cough may suggest infection. A history of smoking or unintentional weight loss makes malignancy more likely. Patients should be asked about inhalational exposures. A thorough medical history with careful attention to chronic pulmonary disease should be obtained, and the clinician should determine risk factors for malignancy and bronchiectatic lung disease (e.g., cystic fibrosis, sarcoidosis).

Physical Examination

Reviewing the vital signs is an important first step. The presence of hypoxemia, tachypnea, and tachycardia should raise concern. Clinicians should examine the nasal and oral cavities; observe the patient's breathing pattern, with careful attention to any respiratory distress; and auscultate the lungs. Clubbing can suggest underlying lung disease such as lung cancer or cystic fibrosis. Signs of bleeding diathesis (e.g., skin or mucosal ecchymoses and petechiae) or teleangectasias may suggest other predispositions to hemoptysis.

Diagnostic Studies

Initial studies should include measurement of a complete blood count to assess for infection, anemia, or thrombocytopenia, coagulation parameters, measurement of electrolytes and renal function, as well as urinalysis to exclude pulmonary-renal disease.

In patients with small, non-massive hemoptysis, outpatient evaluation can be pursued. All patients with hemoptysis need chest imaging. A chest radiograph is usually obtained first, though it frequently does not localize bleeding and can appear normal. In patients without risk factors for malignancy and with a normal chest radiograph, treating for bronchitis and ensuring close follow-up is a reasonable strategy, with further diagnostic workup if bleeding persists.

In contrast, patients with risk factors for malignancy (i.e., age >40 or a smoking history) should undergo additional testing. First, chest computed tomography (CT) should be obtained to better identify masses, bronchiectasis, and parenchymal lesions. Following CT, a flexible bronchoscopy should be performed to exclude bronchogenic carcinoma unless imaging reveals a lesion that can be sampled without bronchoscopy. Small case series show that patients with hemoptysis and unrevealing bronchoscopies have good outcomes.

Interventions

When the amount of hemoptysis is massive, there are three simultaneous goals: first, protect the non-bleeding lung; second, locate the site of bleeding; and third, control the bleeding.

Protecting the airway and non-bleeding lung is paramount in the management of massive hemoptysis, since asphyxiation can happen quickly. If the side of bleeding is known, the patient should be positioned with the bleeding side down, to use gravitational advantage to keep blood out of the non-bleeding lung. Endotracheal intubation should be avoided unless truly necessary, since suctioning through an endotracheal tube is a less effective means of removing blood and clot than the cough reflex. If intubation is required, take steps to protect the non-bleeding lung either by selective intubation of one lung (i.e., the non-bleeding lung) or insertion of a double-lumen endotracheal tube.

Locating the bleeding site is sometimes obvious, but frequently it can be difficult to determine the source of hemoptysis. A chest radiograph, if it shows new opacities, can be helpful in localizing the side or site of bleeding, though this test is not adequate by itself. CT angiography helps by localizing active extravasation. Flexible bronchoscopy may be useful to identify the side of bleeding (although it has only a 50% chance of locating the site). Experts do not agree on the timing of bronchoscopy, though in some cases—cystic fibrosis,

for instance—bronchoscopy is *not* recommended because it may delay definitive management. Finally, proceeding directly to angiography is also a reasonable strategy given that it has both diagnostic and therapeutic capabilities.

Controlling the bleeding during an episode of massive hemoptysis can be accomplished in one of three ways: from the airway lumen, from the involved blood vessel, or by surgical resection of both airway and vessel involved. Bronchoscopic measures are generally only temporizing: a flexible bronchoscope can be used to suction clot and insert a balloon catheter that occludes the involved airway. Rigid bronchoscopy, done by an interventional pulmonologist or thoracic surgeon, may allow therapeutic interventions of bleeding airway lesions such as photocoagulation and cautery. Because most massive hemoptysis arises from the bronchial circulation, bronchial artery embolization is the procedure of choice for control of massive hemoptysis. It is not without risk—embolization of the anterior spinal artery is a known complication—but is generally successful in the short term, with >80% success rate at controlling bleeding immediately, though bleeding can recur if the underlying disease (e.g., a mycetoma) is not treated. Surgical resection has a high mortality rate (up to 15–40%) and should not be pursued unless initial measures have failed and bleeding is ongoing. Ideal candidates for surgery have localized disease but otherwise normal lung parenchyma.

FURTHER READING

Adelman M et al: Cryptogenic hemoptysis: Clinical features, bronchoscopic findings, and natural history in 67 patients. Ann Int Med 102:829, 1985. [PubMed: 3994194]

Flume PA et al: CF pulmonary guidelines. Pulmonary complications: Hemoptysis and pneumothorax. AJRCCM 182:298, 2010.

Hirshberg B et al: Hemoptysis: Etiology, evaluation, and outcome in a tertiary care hospital. Chest 112:440, 1997. [PubMed: 9266882]

Johnson JL: Manifestations of hemoptysis: How to manage minor, moderate, and massive bleeding. Postgrad Med 112:4:101, 2002. [PubMed: 12400152]

Lordan JL et al: The pulmonary physician in critical care: Illustrative case 7. Assessment and management of massive hemoptysis. Thorax 58:814, 2003. [PubMed: 12947147]

Sopko DR, Smith TP: Bronchial artery embolization for massive hemoptysis. Semin Intervent Radiol 28:48, 2011. [PubMed: 22379276]

Chapter 36: Hypoxia and Cyanosis

Joseph Loscalzo

HYPOXIA

The fundamental purpose of the cardiorespiratory system is to deliver O_2 and nutrients to cells and to remove CO_2 and other metabolic products from them. Proper maintenance of this function depends not only on intact cardiovascular and respiratory systems, but also on an adequate number of red blood cells and hemoglobin, and a supply of inspired gas containing adequate O_2.

RESPONSES TO HYPOXIA

Decreased O_2 availability to cells results in an inhibition of oxidative phosphorylation and increased anaerobic glycolysis. This switch from aerobic to anaerobic metabolism, the Pasteur effect, reduces the rate of adenosine 5′-triphosphate (ATP) production. In severe hypoxia, when ATP production is inadequate to meet the energy requirements of ionic and osmotic equilibrium, cell membrane depolarization leads to uncontrolled Ca^{2+} influx and activation of Ca^{2+}-dependent phospholipases and proteases. These events, in turn, cause cell swelling, activation of apoptotic pathways, and, ultimately, cell death.

The adaptations to hypoxia are mediated, in part, by the upregulation of genes encoding a variety of proteins, including glycolytic enzymes, such as phosphoglycerate kinase and phosphofructokinase, as well as the glucose transporters Glut-1 and Glut-2; and by growth factors, such as vascular endothelial growth factor (VEGF) and erythropoietin, which enhance erythrocyte production. The hypoxia-induced increase in expression of these key proteins is governed by the hypoxia-sensitive transcription factor, hypoxia-inducible factor-1 (HIF-1).

During hypoxia, systemic arterioles dilate, at least in part, by opening of K_{ATP} channels in vascular smooth-muscle cells due to the hypoxia-induced reduction in ATP concentration. By contrast, in pulmonary vascular smooth-muscle cells, inhibition of K^+ channels causes depolarization which, in turn, activates voltage-gated Ca^{2+} channels raising the cytosolic [Ca^{2+}] and causing smooth-muscle cell contraction. Hypoxia-induced pulmonary arterial constriction shunts blood away from poorly ventilated portions toward better ventilated portions of the lung; however, it also increases pulmonary vascular resistance and right ventricular afterload.

Effects on the Central Nervous System

Changes in the central nervous system (CNS), particularly the higher centers, are especially important consequences of hypoxia. Acute hypoxia causes impaired judgment, motor incoordination, and a clinical picture resembling acute alcohol intoxication. High-altitude illness is characterized by headache secondary to cerebral vasodilation, gastrointestinal symptoms, dizziness, insomnia, fatigue, or somnolence. Pulmonary arterial and sometimes venous constriction causes capillary leakage and high-altitude pulmonary edema (HAPE) **(Chap. 33)**, which intensifies hypoxia, further promoting vasoconstriction. Rarely, high-altitude cerebral edema (HACE) develops, which is manifest by severe headache and papilledema and can cause coma. As hypoxia becomes more severe, the regulatory centers of the brainstem are affected, and death usually results from respiratory failure.

Effects on the Cardiovascular System

Acute hypoxia stimulates the chemoreceptor reflex arc to induce venoconstriction and systemic arterial vasodilation. These acute changes are accompanied by transiently increased myocardial contractility, which is followed by depressed myocardial contractility with prolonged hypoxia.

CAUSES OF HYPOXIA

Respiratory Hypoxia

When hypoxia occurs from respiratory failure, Pa_{O_2} declines, and when respiratory failure is persistent, the hemoglobin-oxygen (Hb-O_2) dissociation curve **(see Fig. 94-2)** is displaced to the right, with greater quantities of O_2 released at any level of tissue P_{O_2}. Arterial hypoxemia, that is, a reduction of O_2 saturation of arterial blood (Sa_{O_2}), and consequent cyanosis are likely to be more marked when such depression of Pa_{O_2} results from pulmonary disease than when the depression occurs as the result of a decline in the fraction of oxygen in inspired air (F_{IO_2}). In this latter situation, Pa_{CO_2} falls secondary to anoxia-induced hyperventilation and the Hb-O_2 dissociation curve is displaced to the left, limiting the decline in Sa_{O_2} at any level of Pa_{O_2}.

The most common cause of respiratory hypoxia is *ventilation-perfusion mismatch* resulting from perfusion of poorly ventilated alveoli. Respiratory hypoxemia may also be caused by *hypoventilation*, in which case it is associated with an elevation of Pa_{CO_2} **(Chap. 279)**. These two forms of respiratory hypoxia are usually correctable by inspiring 100% O_2 for several minutes. A third cause of respiratory hypoxia is shunting of blood across the lung from the pulmonary arterial to the venous bed (*intrapulmonary right-to-left shunting*) by perfusion of nonventilated portions of the lung, as in pulmonary atelectasis or through pulmonary arteriovenous connections. The low Pa_{O_2} in this situation is only partially corrected by an F_{IO_2} of 100%.

Hypoxia Secondary to High Altitude

As one ascends rapidly to 3000 m (~10,000 ft), the reduction of the O_2 content of inspired air (F_{IO_2}) leads to a decrease in alveolar P_{O_2} to ~60 mmHg, and a condition termed *high-altitude illness* develops (see above). At

higher altitudes, arterial saturation declines rapidly and symptoms become more serious; and at 5000 m, unacclimated individuals usually cease to be able to function normally owing to the changes in CNS function described above.

Hypoxia Secondary to Right-to-Left Extrapulmonary Shunting

From a physiologic viewpoint, this cause of hypoxia resembles intrapulmonary right-to-left shunting but is caused by congenital cardiac malformations, such as tetralogy of Fallot, transposition of the great arteries, and Eisenmenger's syndrome **(Chap. 264)**. As in pulmonary right-to-left shunting, the Pa_{O_2} cannot be restored to normal with inspiration of 100% O_2.

Anemic Hypoxia

A reduction in hemoglobin concentration of the blood is accompanied by a corresponding decline in the O_2-carrying capacity of the blood. Although the Pa_{O_2} is normal in anemic hypoxia, the absolute quantity of O_2 transported per unit volume of blood is diminished. As the anemic blood passes through the capillaries and the usual quantity of O_2 is removed from it, the P_{O_2} and saturation in the venous blood decline to a greater extent than normal.

Carbon Monoxide (CO) Intoxication

(See also Chap. S11) Hemoglobin that binds with CO (carboxy-hemoglobin, COHb) is unavailable for O_2 transport. In addition, the presence of COHb shifts the Hb-O_2 dissociation curve to the left **(see Fig. 94-2)** so that O_2 is unloaded only at lower tensions, further contributing to tissue hypoxia.

Circulatory Hypoxia

As in anemic hypoxia, the Pa_{O_2} is usually normal, but venous and tissue P_{O_2} values are reduced as a consequence of reduced tissue perfusion and greater tissue O_2 extraction. This pathophysiology leads to an increased arterial-mixed venous O_2 difference (a-v-O_2 difference), or gradient. Generalized circulatory hypoxia occurs in heart failure **(Chap. 252)** and in most forms of shock **(Chap. 296)**.

Specific Organ Hypoxia

Localized circulatory hypoxia may occur as a result of decreased perfusion secondary to arterial obstruction, as in localized atherosclerosis in any vascular bed, or as a consequence of vasoconstriction, as observed in Raynaud's phenomenon **(Chap. 275)**. Localized hypoxia may also result from venous obstruction and the resultant expansion of interstitial fluid causing arteriolar compression and, thereby, reduction of arterial inflow. Edema, which increases the distance through which O_2 must diffuse before it reaches cells, can also cause localized hypoxia. In an attempt to maintain adequate perfusion to more vital organs in patients with

reduced cardiac output secondary to heart failure or hypovolemic shock, vasoconstriction may reduce perfusion in the limbs and skin, causing hypoxia of these regions.

Increased O_2 Requirements

If the O_2 consumption of tissues is elevated without a corresponding increase in perfusion, tissue hypoxia ensues and the Po_2 in venous blood declines. Ordinarily, the clinical picture of patients with hypoxia due to an elevated metabolic rate, as in fever or thyrotoxicosis, is quite different from that in other types of hypoxia: the skin is warm and flushed owing to increased cutaneous blood flow that dissipates the excessive heat produced, and cyanosis is usually absent.

Exercise is a classic example of increased tissue O_2 requirements. These increased demands are normally met by several mechanisms operating simultaneously: (1) increase in the cardiac output and ventilation and, thus, O_2 delivery to the tissues; (2) a preferential shift in blood flow to the exercising muscles by changing vascular resistances in the circulatory beds of exercising tissues, directly and/or reflexly; (3) an increase in O_2 extraction from the delivered blood and a widening of the arteriovenous O_2 difference; and (4) a reduction in the pH of the tissues and capillary blood, shifting the Hb-O_2 curve to the right **(see Fig. 94-2)**, and unloading more O_2 from hemoglobin. If the capacity of these mechanisms is exceeded, then hypoxia, especially of the exercising muscles, will result.

Improper Oxygen Utilization

Cyanide **(Chap. 450)** and several other similarly acting poisons cause cellular hypoxia. The tissues are unable to use O_2, and, as a consequence, the venous blood tends to have a high O_2 tension. This condition has been termed *histotoxic hypoxia*.

ADAPTATION TO HYPOXIA

An important component of the respiratory response to hypoxia originates in special chemosensitive cells in the carotid and aortic bodies and in the respiratory center in the brainstem. The stimulation of these cells by hypoxia increases ventilation, with a loss of CO_2, and can lead to respiratory alkalosis. When combined with the metabolic acidosis resulting from the production of lactic acid, the serum bicarbonate level declines **(Chap. 51)**.

With the reduction of Pao_2, cerebrovascular resistance decreases and cerebral blood flow increases in an attempt to maintain O_2 delivery to the brain. However, when the reduction of Pao_2 is accompanied by hyperventilation and a reduction of $Paco_2$, cerebrovascular resistance rises, cerebral blood flow falls, and tissue hypoxia intensifies.

The diffuse, systemic vasodilation that occurs in generalized hypoxia increases the cardiac output. In patients with underlying heart disease, the requirements of peripheral tissues for an increase of cardiac output with hypoxia may precipitate congestive heart failure. In patients with ischemic heart disease, a reduced Pa_{O_2} may intensify myocardial ischemia and further impair left ventricular function.

One of the important compensatory mechanisms for chronic hypoxia is an increase in the hemoglobin concentration and in the number of red blood cells in the circulating blood, that is, the development of polycythemia secondary to erythropoietin production **(Chap. 99)**. In persons with chronic hypoxemia secondary to prolonged residence at a high altitude (>13,000 ft, 4200 m), a condition termed *chronic mountain sickness* develops. This disorder is characterized by a blunted respiratory drive, reduced ventilation, erythrocytosis, cyanosis, weakness, right ventricular enlargement secondary to pulmonary hypertension, and even stupor.

CYANOSIS

Cyanosis refers to a bluish color of the skin and mucous membranes resulting from an increased quantity of reduced hemoglobin (i.e., deoxygenated hemoglobin) or of hemoglobin derivatives (e.g., methemoglobin or sulfhemoglobin) in the small blood vessels of those tissues. It is usually most marked in the lips, nail beds, ears, and malar eminences. Cyanosis, especially if developed recently, is more commonly detected by a family member than the patient. The florid skin characteristic of polycythemia vera **(Chap. 99)** must be distinguished from the true cyanosis discussed here. A cherry-colored flush, rather than cyanosis, is caused by COHb **(Chap. 450)**.

The degree of cyanosis is modified by the color of the cutaneous pigment and the thickness of the skin, as well as by the state of the cutaneous capillaries. The accurate clinical detection of the presence and degree of cyanosis is difficult, as proved by oximetric studies. In some instances, central cyanosis can be detected reliably when the Sa_{O_2} has fallen to 85%; in others, particularly in dark-skinned persons, it may not be detected until it has declined to 75%. In the latter case, examination of the mucous membranes in the oral cavity and the conjunctivae rather than examination of the skin is more helpful in the detection of cyanosis.

The increase in the quantity of reduced hemoglobin in the mucocutaneous vessels that produces cyanosis may be brought about either by an increase in the quantity of venous blood as a result of dilation of the venules (including precapillary venules) or by a reduction in the Sa_{O_2} in the capillary blood. In general, cyanosis becomes apparent when the concentration of reduced hemoglobin in capillary blood exceeds 40 g/L (4 g/dL).

It is the *absolute*, rather than the *relative*, quantity of reduced hemoglobin that is important in producing cyanosis. Thus, in a patient with severe anemia, the *relative* quantity of reduced hemoglobin in the venous blood may be very large when considered in relation to the total quantity of hemoglobin in the blood. However, since the concentration of the latter is markedly reduced, the *absolute* quantity of reduced hemoglobin may still be low, and, therefore, patients with severe anemia and even *marked* arterial

desaturation may not display cyanosis. Conversely, the higher the total hemoglobin content, the greater the tendency toward cyanosis; thus, patients with marked polycythemia tend to be cyanotic at higher levels of Sa_{O_2} than patients with normal hematocrit values. Likewise, local passive congestion, which causes an increase in the total quantity of reduced hemoglobin in the vessels in a given area, may cause cyanosis. Cyanosis is also observed when nonfunctional hemoglobin, such as methemoglobin (consequential or acquired) or sulfhemoglobin (Chap. 94), is present in blood.

Cyanosis may be subdivided into central and peripheral types. In *central* cyanosis, the Sa_{O_2} is reduced or an abnormal hemoglobin derivative is present, and the mucous membranes and skin are both affected. *Peripheral* cyanosis is due to a slowing of blood flow and abnormally great extraction of O_2 from normally saturated arterial blood; it results from vasoconstriction and diminished peripheral blood flow, such as occurs in cold exposure, shock, congestive failure, and peripheral vascular disease. Often in these conditions, the mucous membranes of the oral cavity or those beneath the tongue may be spared. Clinical differentiation between central and peripheral cyanosis may not always be straightforward, and in conditions such as cardiogenic shock with pulmonary edema, there may be a mixture of both types.

DIFFERENTIAL DIAGNOSIS

Central Cyanosis

(Table 36-1) Decreased Sa_{O_2} results from a marked reduction in the Pa_{O_2}. This reduction may be brought about by a decline in the F_{IO_2} without sufficient compensatory alveolar hyperventilation to maintain alveolar P_{O_2}. Cyanosis usually becomes manifest in an ascent to an altitude of 4000 m (13,000 ft).

TABLE 36-1

Causes of Cyanosis

Central Cyanosis
Decreased arterial oxygen saturation
Decreased atmospheric pressure—high altitude
Impaired pulmonary function
Alveolar hypoventilation
Inhomogeneity in pulmonary ventilation and perfusion (perfusion of hypoventilated alveoli)
Impaired oxygen diffusion
Anatomic shunts
Certain types of congenital heart disease
Pulmonary arteriovenous fistulas
Multiple small intrapulmonary shunts
Hemoglobin with low affinity for oxygen
Hemoglobin abnormalities
Methemoglobinemia—hereditary, acquired
Sulfhemoglobinemia—acquired
Carboxyhemoglobinemia (not true cyanosis)
Peripheral Cyanosis
Reduced cardiac output
Cold exposure
Redistribution of blood flow from extremities

Arterial obstruction
Venous obstruction

Seriously *impaired pulmonary function*, through perfusion of unventilated or poorly ventilated areas of the lung or alveolar hypoventilation, is a common cause of central cyanosis **(Chap. 279)**. This condition may occur acutely, as in extensive pneumonia or pulmonary edema, or chronically, with chronic pulmonary diseases (e.g., emphysema). In the latter situation, secondary polycythemia is generally present and clubbing of the fingers (see below) may occur. Another cause of reduced Sao_2 is *shunting of systemic venous blood into the arterial circuit*. Certain forms of congenital heart disease are associated with cyanosis on this basis (see above and **Chap. 264**).

Pulmonary arteriovenous fistulae may be congenital or acquired, solitary or multiple, microscopic or massive. The severity of cyanosis produced by these fistulae depends on their size and number. They occur with some frequency in hereditary hemorrhagic telangiectasia. Sao_2 reduction and cyanosis may also occur in some patients with cirrhosis, presumably as a consequence of pulmonary arteriovenous fistulae or portal vein–pulmonary vein anastomoses.

In patients with cardiac or pulmonary right-to-left shunts, the presence and severity of cyanosis depend on the size of the shunt relative to the systemic flow and on the Hb-O_2 saturation of the venous blood. With increased extraction of O_2 from the blood by the exercising muscles, the venous blood returning to the right side of the heart is more unsaturated than at rest, and shunting of this blood intensifies the cyanosis. Secondary polycythemia occurs frequently in patients in this setting and contributes to the cyanosis.

Cyanosis can be caused by small quantities of circulating methemoglobin (Hb Fe^{3+}) and by even smaller quantities of sulfhemoglobin **(Chap. 94)**; both of these hemoglobin derivatives impair oxygen delivery to the tissues. Although they are uncommon causes of cyanosis, these abnormal hemoglobin species should be sought by spectroscopy when cyanosis is not readily explained by malfunction of the circulatory or respiratory systems. Generally, digital clubbing does not occur with them.

Peripheral Cyanosis

Probably the most common cause of peripheral cyanosis is the normal vasoconstriction resulting from exposure to cold air or water. When cardiac output is reduced, cutaneous vasoconstriction occurs as a compensatory mechanism so that blood is diverted from the skin to more vital areas such as the CNS and heart, and cyanosis of the extremities may result even though the arterial blood is normally saturated.

Arterial obstruction to an extremity, as with an embolus, or arteriolar constriction, as in cold-induced vasospasm (Raynaud's phenomenon) **(Chap. 275)**, generally results in pallor and coldness, and there may be

associated cyanosis. Venous obstruction, as in thrombophlebitis or deep venous thrombosis, dilates the subpapillary venous plexuses and thereby intensifies cyanosis.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Cyanosis

Certain features are important in arriving at the cause of cyanosis:

1. It is important to ascertain the time of onset of cyanosis. Cyanosis present since birth or infancy is usually due to congenital heart disease.
2. Central and peripheral cyanosis must be differentiated. Evidence of disorders of the respiratory or cardiovascular systems is helpful. Massage or gentle warming of a cyanotic extremity will increase peripheral blood flow and abolish peripheral, but not central, cyanosis.
3. The presence or absence of clubbing of the digits (see below) should be ascertained. The combination of cyanosis and clubbing is frequent in patients with congenital heart disease and right-to-left shunting and is seen occasionally in patients with pulmonary disease, such as lung abscess or pulmonary arteriovenous fistulae. In contrast, peripheral cyanosis or acutely developing central cyanosis is *not* associated with clubbed digits.
4. Pa_{O_2} and Sa_{O_2} should be determined, and, in patients with cyanosis in whom the mechanism is obscure, spectroscopic examination of the blood should be performed to look for abnormal types of hemoglobin (critical in the differential diagnosis of cyanosis).

CLUBBING

The selective bulbous enlargement of the distal segments of the fingers and toes due to proliferation of connective tissue, particularly on the dorsal surface, is termed *clubbing*; there is also increased sponginess of the soft tissue at the base of the clubbed nail. Clubbing may be hereditary, idiopathic, or acquired and associated with a variety of disorders, including cyanotic congenital heart disease (see above), infective endocarditis, and a variety of pulmonary conditions (among them primary and metastatic lung cancer, bronchiectasis, asbestosis, sarcoidosis, lung abscess, cystic fibrosis, tuberculosis, and mesothelioma), as well as with some gastrointestinal diseases (including inflammatory bowel disease and hepatic cirrhosis). In some instances, it is occupational, for example, in jackhammer operators.

Clubbing in patients with primary and metastatic lung cancer, mesothelioma, bronchiectasis, or hepatic cirrhosis may be associated with *hypertrophic osteoarthropathy*. In this condition, the subperiosteal formation of new bone in the distal diaphyses of the long bones of the extremities causes pain and symmetric arthritis-like changes in the shoulders, knees, ankles, wrists, and elbows. The diagnosis of

hypertrophic osteoarthropathy may be confirmed by bone radiograph or magnetic resonance imaging (MRI). Although the mechanism of clubbing is unclear, it appears to be secondary to humoral substances that cause dilation of the vessels of the distal digits as well as growth factors released from platelet precursors in the digital circulation. In certain circumstances, clubbing is reversible, such as following lung transplantation for cystic fibrosis.

FURTHER READING

Callemeyn J et al: Clubbing and hypertrophic osteoarthropathy: Insights into diagnosis, pathophysiology, and clinical significance. Acta Clin Belg 22:1, 2016.

MacIntyre NR: Tissue hypoxia: Implications for the respiratory clinician. Respir Care 59:1590, 2014. [PubMed: 25161296]

Chapter 37: Edema

Eugene Braunwald; Joseph Loscalzo

FIGURE 37-1

PLASMA AND INTERSTITIAL FLUID EXCHANGE

About two-thirds of total body water is intracellular and one-third is extracellular. Approximately one-fourth of the latter is in the plasma and the remainder comprises the interstitial fluid. Edema represents an excess of interstitial fluid that has become evident clinically.

There is constant interchange of fluid between the two compartments of the extracellular fluid. The hydrostatic pressure within the capillaries and the colloid oncotic pressure in the interstitial fluid promote the movement of water and diffusible solutes from plasma to the interstitium. This movement is most prominent at the arterial origin of the capillary and falls progressively with the decline in intracapillary pressure and the rise in oncotic pressure toward the venular end. Fluid is returned from the interstitial space into the vascular system largely through the lymphatic system. These interchanges of fluids are normally balanced so that the volumes of the intravascular and interstitial compartments remain constant. However, a net movement of fluid from the intravascular to the interstitial spaces takes place and may be responsible for the development of edema under the following conditions: (1) an increase in intracapillary hydrostatic pressure; (2) inadequate lymphatic drainage; (3) reductions in the oncotic pressure in the plasma; (4) damage to the capillary endothelial barrier; and (5) increases in the oncotic pressure in the interstitial space.

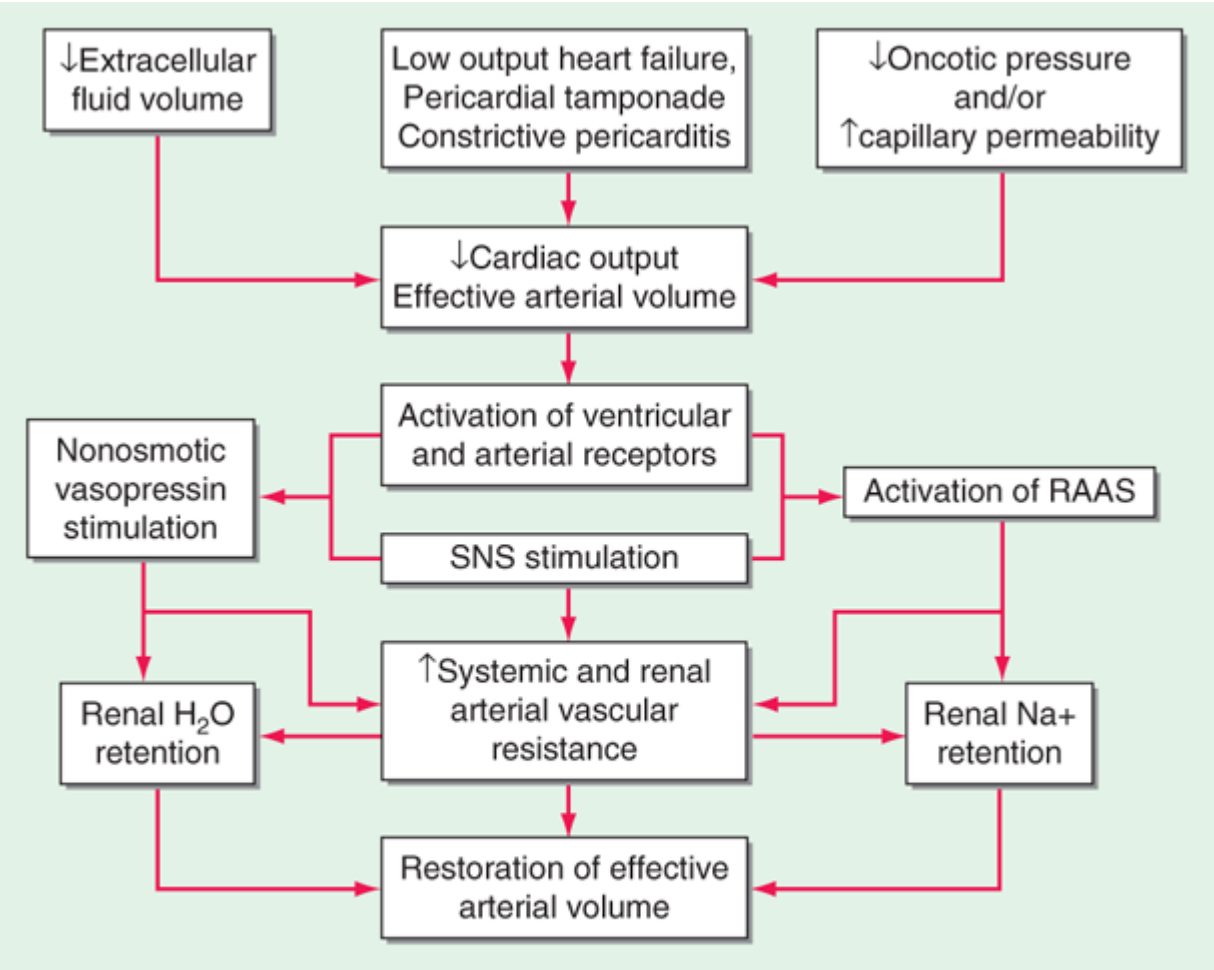
REDUCTION OF EFFECTIVE ARTERIAL VOLUME

In many forms of edema, the effective arterial blood volume, a parameter that represents the filling of the arterial tree and that effectively perfuses the tissues, is reduced. Underfilling of the arterial tree may be caused by a reduction of cardiac output and/or systemic vascular resistance, by the pooling of blood in the splanchnic veins (as in cirrhosis), and by hypoalbuminemia **(Fig. 37-1A)**. As a consequence of this underfilling, a series of physiologic responses designed to restore the effective arterial volume to normal are set into motion. A key element of these responses is the renal retention of sodium and, therefore, water, thereby restoring effective arterial volume, but sometimes also leading to the development or intensification of edema.

FIGURE 37-1

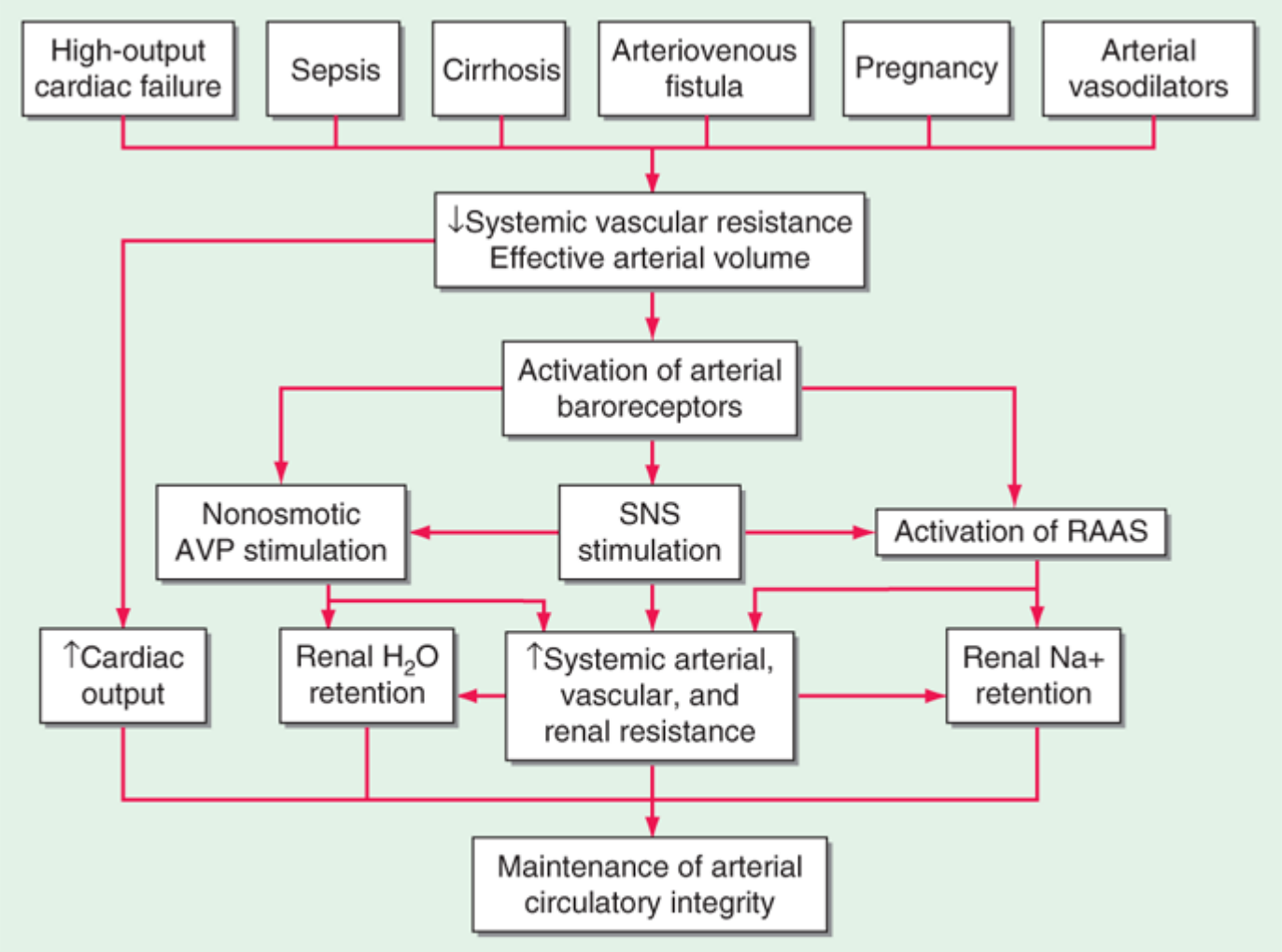
Clinical conditions in which a decrease in cardiac output (*A*) and systemic vascular resistance (*B*) cause arterial underfilling with resulting neurohumoral activation and renal sodium and water retention. In

addition to activating the neurohumoral axis, adrenergic stimulation causes renal vasoconstriction and enhances sodium and fluid transport by the proximal tubule epithelium. RAAS, renin-angiotensin aldosterone system; SNS, sympathetic nervous system. *(Modified from RW Schrier: Ann Intern Med 113:155, 1990.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

RENAL FACTORS AND THE RENIN-ANGIOTENSIN-ALDOSTERONE SYSTEM

The diminished renal blood flow characteristic of states in which the effective arterial blood volume is reduced is translated by the renal juxtaglomerular cells (specialized myoepithelial cells surrounding the afferent arteriole) into a signal for increased renin release. Renin is an enzyme with a molecular mass of about 40,000 Da that acts on its substrate, angiotensinogen, an α_2-globulin synthesized by the liver, to release angiotensin I, a decapeptide, which in turn is converted to angiotensin II (AII), an octapeptide. AII has generalized vasoconstrictor properties, particularly on the renal efferent arterioles. This action reduces the hydrostatic pressure in the peritubular capillaries, whereas the increased filtration fraction raises the colloid osmotic pressure in these vessels, thereby enhancing salt and water reabsorption in the proximal tubule as well as in the ascending limb of the loop of Henle.

The renin-angiotensin-aldosterone system (RAAS) operates as both a hormonal and paracrine system. Its activation causes sodium and water retention and thereby contributes to edema formation. Blockade of the conversion of angiotensin I to AII and blockade of the AII receptors enhance sodium and water excretion and

reduce many forms of edema. AII that enters the systemic circulation stimulates the production of aldosterone by the zona glomerulosa of the adrenal cortex. Aldosterone in turn enhances sodium reabsorption (and potassium excretion) by the collecting tubule, further favoring edema formation. Blockade of the action of aldosterone by spironolactone or eplerenone (aldosterone antagonists) or by amiloride (a blocker of epithelial sodium channels) often induces a moderate diuresis in edematous states.

ARGININE VASOPRESSIN

(See also Chap. 374) The secretion of arginine vasopressin (AVP) by the posterior pituitary gland occurs in response to increased intracellular osmolar concentration; by stimulating V_2 receptors, AVP increases the reabsorption of free water in the distal tubules and collecting ducts of the kidneys, thereby increasing total-body water. Circulating AVP is elevated in many patients with heart failure secondary to a nonosmotic stimulus associated with decreased effective arterial volume and reduced compliance of the left atrium. Such patients fail to show the normal reduction of AVP with a reduction of osmolality, contributing to edema formation and hyponatremia.

ENDOTHELIN-1

This potent peptide vasoconstrictor is released by endothelial cells. Its concentration in the plasma is elevated in patients with severe heart failure and contributes to renal vasoconstriction, sodium retention, and edema.

NATRIURETIC PEPTIDES

Atrial distention causes release into the circulation of atrial natriuretic peptide (ANP), a polypeptide. A high-molecular-weight precursor of ANP is stored in secretory granules within atrial myocytes. A closely related natriuretic peptide (pre-prohormone brain natriuretic peptide) is stored primarily in ventricular myocytes and is released when ventricular diastolic pressure rises. Released ANP and BNP (which is derived from its precursor) bind to the natriuretic receptor-A, which causes: (1) excretion of sodium and water by augmenting glomerular filtration rate, inhibiting sodium reabsorption in the proximal tubule, and inhibiting release of renin and aldosterone; and (2) dilation of arterioles and venules by antagonizing the vasoconstrictor actions of AII, AVP, and sympathetic stimulation. Thus, elevated levels of natriuretic peptides have the capacity to oppose sodium retention in hypervolemic and edematous states.

Although circulating levels of ANP and BNP are elevated in heart failure and in cirrhosis with ascites, these natriuretic peptides are not sufficiently potent to prevent edema formation. Indeed, in edematous states, resistance to the actions of natriuretic peptides may be increased, further reducing their effectiveness.

Further discussion of the control of sodium and water balance is found in **Chap. S1**.

CLINICAL CAUSES OF EDEMA

A weight gain of several kilograms usually precedes overt manifestations of generalized edema. *Anasarca* refers to gross, generalized edema. *Ascites* (Chap. 46) and *hydrothorax* refer to accumulation of excess fluid in the peritoneal and pleural cavities, respectively, and are considered special forms of edema.

Edema is recognized by the persistence of an indentation of the skin after pressure known as “pitting” edema. In its more subtle form, edema may be detected by noting that after the stethoscope is removed from the chest wall, the rim of the bell leaves an indentation on the skin of the chest for a few minutes. Edema may be present when the ring on a finger fits more snugly than in the past or when a patient complains of difficulty putting on shoes, particularly in the evening. Edema may also be recognized by puffiness of the face, which is most readily apparent in the periorbital areas.

GENERALIZED EDEMA

The differences among the major causes of generalized edema are shown in Table 37-1. Cardiac, renal, hepatic, or nutritional disorders are responsible for a large majority of patients with generalized edema. Consequently, the differential diagnosis of generalized edema should be directed toward identifying or excluding these several conditions.

TABLE 37-1

Principal Causes of Generalized Edema: History, Physical Examination, and Laboratory Findings

Organ System	History	Physical Examination	Laboratory Findings
Cardiac	Dyspnea with exertion prominent—often associated with orthopnea—or paroxysmal nocturnal dyspnea	Elevated jugular venous pressure, ventricular (S_3) gallop; occasionally with displaced or dyskinetic apical pulse; peripheral cyanosis, cool extremities, small pulse pressure when severe	Elevated urea nitrogen-to-creatinine ratio common; serum sodium often diminished; elevated natriuretic peptides
Hepatic	Dyspnea uncommon, except if associated with significant degree of ascites; most often a history of ethanol abuse	Frequently associated with ascites; jugular venous pressure normal or low; blood pressure lower than in renal or cardiac disease; one or more additional signs of chronic liver disease (jaundice, palmar erythema, Dupuytren's contracture, spider angiomata, male gynecomastia; asterixis and other signs of encephalopathy) may be present	If severe, reductions in serum albumin, cholesterol, other hepatic proteins (transferrin, fibrinogen); liver enzymes elevated, depending on the cause and acuity of liver injury; tendency toward hypokalemia, respiratory alkalosis; macrocytosis from folate deficiency
Renal (CRF)	Usually chronic: may be associated with uremic signs and symptoms, including decreased appetite, altered (metallic or fishy) taste, altered sleep pattern, difficulty concentrating, restless legs, or myoclonus; dyspnea can be present, but generally less prominent than in heart failure	Elevated blood pressure; hypertensive retinopathy; nitrogenous fetor; pericardial friction rub in advanced cases with uremia	Elevation of serum creatinine and cystatin C; albuminuria; hyperkalemia, metabolic acidosis, hyperphosphatemia, hypocalcemia, anemia (usually normocytic)

Organ System	History	Physical Examination	Laboratory Findings
Renal (NS)	Childhood diabetes mellitus; plasma cell dyscrasias	Periorbital edema; hypertension	Proteinuria (≥3.5 g/d); hypoalbuminemia; hypercholesterolemia; microscopic hematuria

Abbreviations: CRF, chronic renal failure; NS, nephrotic syndrome.

Source: Modified from GM Chertow: Approach to the patient with edema, in *Primary Cardiology*, 2nd ed, E Braunwald, L Goldman (eds). Philadelphia, Saunders, 2003, pp 117–128.

Heart Failure

(See also Chap. 252) In heart failure, the impaired systolic emptying of the ventricle(s) and/or the impairment of ventricular relaxation promotes an accumulation of blood in the venous circulation at the expense of the effective arterial volume. In addition, the activation of the sympathetic nervous system and the RAAS (see above) acts in concert to cause renal vasoconstriction and reduction of glomerular filtration and salt and water retention. Sodium and water retention continue, and the increment in blood volume accumulates in the venous circulation, raising venous and intracapillary pressure resulting in edema (Fig. 37-1).

The presence of overt cardiac disease, as manifested by cardiac enlargement and/or ventricular hypertrophy, together with clinical evidence of cardiac failure, such as dyspnea, basilar rales, venous distention, and hepatomegaly, usually indicates that edema results from heart failure. Noninvasive tests such as electrocardiography, echocardiography, and measurements of BNP (or NTproBNP) are helpful in establishing the diagnosis of heart disease. The edema of heart failure typically occurs in the dependent portions of the body.

Edema of Renal Disease

(See also Chap. 308) The edema that occurs during the acute phase of glomerulonephritis is characteristically associated with hematuria, proteinuria, and hypertension. In most instances, the edema results from primary retention of sodium and water by the kidneys owing to renal dysfunction. This state differs from most forms of heart failure in that it is characterized by a normal (or sometimes even increased) cardiac output. Patients with *chronic* renal failure may also develop edema due to primary renal retention of sodium and water.

Nephrotic Syndrome and Other Hypoalbuminemic States

The primary alteration in the nephrotic syndrome is a diminished colloid oncotic pressure due to losses of large quantities (≥3.5 g/d) of protein into the urine, and hypoalbuminemia (<3.0 g/dL). As a result of the reduced colloid osmotic pressure, the sodium and water that are retained cannot be confined within the vascular compartment, and total and effective arterial blood volumes decline. This process initiates the edema-forming sequence of events described above, including activation of the RAAS. The nephrotic syndrome may occur during the course of a variety of kidney diseases, including glomerulonephritis, diabetic glomerulosclerosis, and hypersensitivity reactions. The edema is diffuse, symmetric, and most prominent in the dependent areas; periorbital edema is most prominent in the morning.

Hepatic Cirrhosis

(See also Chap. 337) This condition is characterized in part by hepatic venous outflow obstruction, which in turn expands the splanchnic blood volume, and hepatic lymph formation. Intrahepatic hypertension acts as a stimulus for renal sodium retention and causes a reduction of effective arterial blood volume. These alterations are frequently complicated by hypoalbuminemia secondary to reduced hepatic synthesis of albumin, as well as peripheral arterial vasodilation. These effects reduce the effective arterial blood volume, leading to activation of the sodium- and water-retaining mechanisms described above **(Fig. 37-1B)**. The concentration of circulating aldosterone often is elevated by the failure of the liver to metabolize this hormone. Initially, the excess interstitial fluid is localized preferentially proximal (upstream) to the congested portal venous system, causing ascites **(Chap. 46)**. In later stages, particularly when there is severe hypoalbuminemia, peripheral edema may develop. A sizable accumulation of ascitic fluid may increase intraabdominal pressure and impede venous return from the lower extremities and contribute to the accumulation of the edema.

Drug-Induced Edema

A large number of widely used drugs can cause edema **(Table 37-2)**. Mechanisms include renal vasoconstriction (NSAIDs and cyclosporine), arteriolar dilation (vasodilators), augmented renal sodium reabsorption (steroid hormones), and capillary damage.

TABLE 37-2

Drugs Associated with Edema Formation

Nonsteroidal anti-inflammatory drugs
Antihypertensive agents
Direct arterial/arteriolar vasodilators
Hydralazine
Clonidine
Methyldopa
Guanethidine
Minoxidil
Calcium channel antagonists
α-Adrenergic antagonists
Thiazolidinediones
Steroid hormones
Glucocorticoids
Anabolic steroids
Estrogens
Progestins
Cyclosporine
Growth hormone
Immunotherapies
Interleukin 2

OKT3 monoclonal antibody

Source: Modified from GM Chertow: Approach to the patient with edema, in *Primary Cardiology*, 2nd ed, E Braunwald, L Goldman (eds). Philadelphia, Saunders, 2003, pp 117–128.

Edema of Nutritional Origin

A diet grossly deficient in calories and particularly in protein over a prolonged period may produce hypoproteinemia and edema. The latter may be intensified by the development of beriberi heart disease, which also is of nutritional origin, in which multiple peripheral arteriovenous fistulae result in reduced effective systemic perfusion and effective arterial blood volume, thereby enhancing edema formation **(Chap. 326)** (Fig. 37-1B). Edema develops or becomes intensified when famished subjects are first provided with an adequate diet. The ingestion of more food may increase the quantity of sodium ingested, which is then retained along with water. So-called refeeding edema also may be linked to increased release of insulin, which directly increases tubular sodium reabsorption. In addition to hypoalbuminemia, hypokalemia and caloric deficits may be involved in the edema of starvation.

LOCALIZED EDEMA

In thrombophlebitis, varicose veins, and in primary venous valve failure, the hydrostatic pressure in the capillary bed upstream (proximal) of the obstruction increases so that an abnormal quantity of fluid is transferred from the vascular to the interstitial space, which may give rise to localized edema. The latter may also occur in lymphatic obstruction caused by chronic lymphangitis, resection of regional lymph nodes, filariasis, and genetic (frequently called primary) lymphedema. The latter is particularly intractable because restriction of lymphatic flow results in both an increase in intracapillary pressure and increased protein concentration in the interstitial fluid, which act in concert to aggravate fluid retention.

Other Causes of Edema

These causes include hypothyroidism (myxedema) due to deposition of hyaluronic acid, and hyperthyroidism (pretibial myxedema secondary to Graves' disease), in which edema is typically nonpitting and, in Graves' disease, exogenous hyperadrenocortism; pregnancy; and administration of estrogens and vasodilators, particularly dihydropyridines such as nifedipine.

DISTRIBUTION OF EDEMA

The distribution of edema is an important guide to its cause. Edema associated with heart failure tends to be more extensive in the legs and to be accentuated in the evening, a feature also determined largely by posture. When patients with heart failure are confined to bed, edema may be most prominent in the presacral region.

Edema resulting from hypoproteinemia, as occurs in the nephrotic syndrome, characteristically is generalized, but it is especially evident in the very soft tissues of the eyelids and face and tends to be most pronounced in the morning owing to the recumbent posture assumed during the night. Less common causes of facial edema include trichinosis, allergic reactions, and myxedema. Edema limited to one leg or to one or both arms is usually the result of venous and/or lymphatic obstruction. Unilateral paralysis reduces lymphatic and venous drainage on the affected side and may also be responsible for unilateral edema. In patients with obstruction of the superior vena cava, edema is confined to the face, neck, and upper extremities in which the venous pressure is elevated compared with that in the lower extremities.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Edema

An important first question is whether the edema is localized or generalized. If it is localized, the local phenomena that may be responsible should be identified. If the edema is generalized, one should determine if there is serious hypoalbuminemia, e.g., serum albumin <3.0 g/dl. If so, the history, physical examination, urinalysis, and other laboratory data will help evaluate the question of cirrhosis, severe malnutrition, or the nephrotic syndrome as the underlying disorder. If hypoalbuminemia is not present, it should be determined if there is evidence of heart failure severe enough to promote generalized edema. Finally, it should be ascertained as to whether or not the patient has an adequate urine output or if there is significant oliguria or anuria. **These abnormalities are discussed in Chaps. 48, 304, and 305.**

FURTHER READING

Clark AL, Cleland JG: Causes and treatment of oedema in patients with heart failure. Nature Rev Cardiol 10:156, 2013.

Damman K et al: Congestion in chronic systolic heart failure is related to renal dysfunction and increased mortality. Eur J Heart Fail 12:974, 2010. [PubMed: 20685688]

Ferrell RE et al: *GJC2* missense mutations cause human lymphedema. Am J Hum Genet 86:943, 2010. [PubMed: 20537300]

Frison S et al: Omitting edema measurement: How much acute malnutrition are we missing? Am J Clin Nutr 102:1176, 2015. [PubMed: 26377162]

Levick JR, Michel CC: Microvascular fluid exchange and the revised Starling principle. Cardiovascular Res 87:198, 2010.

Mortimer PS, Rockson SG: New developments in clinical aspects of lymphatic disease. J Clin Invest 124:915, 2014. [PubMed: 24590276]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 38: Approach to the Patient with a Heart Murmur

Patrick T. O'Gara; Joseph Loscalzo

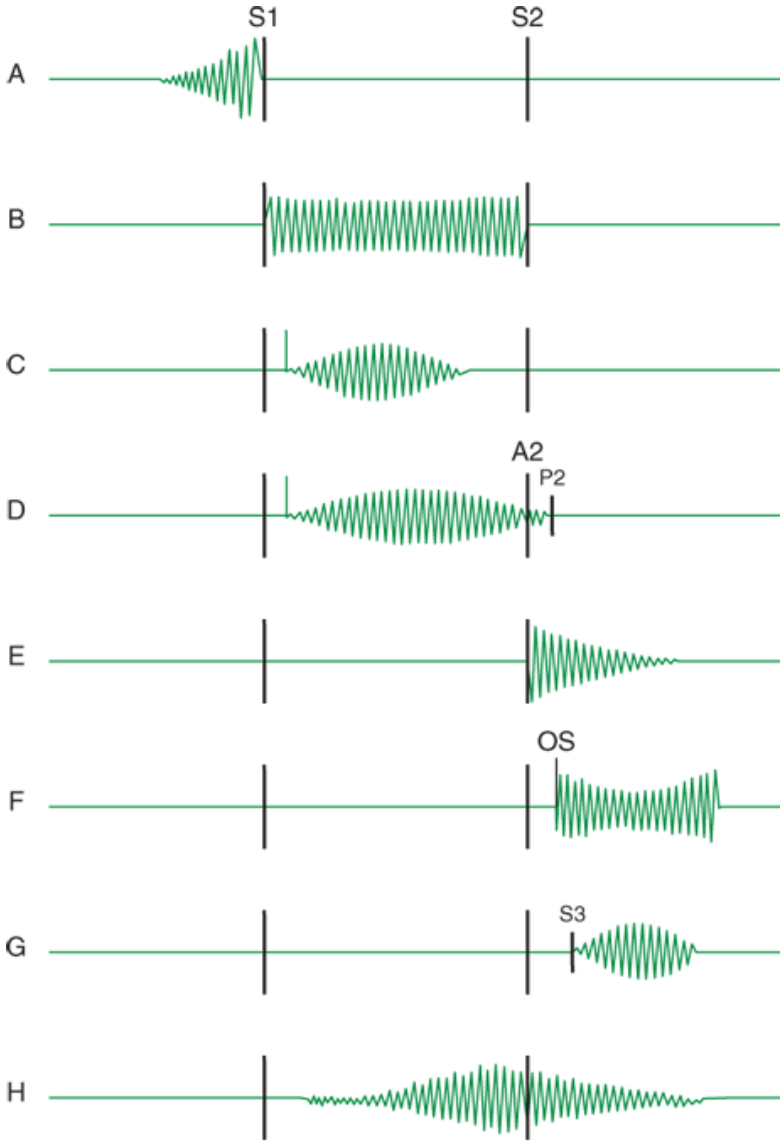
INTRODUCTION

The differential diagnosis of a heart murmur begins with a careful assessment of its major attributes and response to bedside maneuvers. The history, clinical context, and associated physical examination findings provide additional clues to help establish the significance of a heart murmur. Accurate bedside identification of a heart murmur can inform decisions regarding the indications for noninvasive testing and the need for referral to a cardiovascular specialist. Preliminary discussions can be held with the patient regarding antibiotic or rheumatic fever prophylaxis, the need to restrict various forms of physical activity, and the potential role for family screening.

Heart murmurs are caused by audible vibrations that are due to increased turbulence from accelerated blood flow through normal or abnormal orifices; flow through a narrowed or irregular orifice into a dilated vessel or chamber; or backward flow through an incompetent valve, ventricular septal defect, or patent ductus arteriosus. They traditionally are defined by their timing within the cardiac cycle (Fig. 38-1). *Systolic murmurs* begin with or after the first heart sound (S_1) and terminate at or before the component (A_2 or P_2) of the second heart sound (S_2) that corresponds to their site of origin (left or right, respectively). *Diastolic murmurs* begin with or after the associated component of S_2 and end at or before the subsequent S_1. *Continuous murmurs* are not confined to either phase of the cardiac cycle but instead begin in early systole and proceed through S_2 into all or part of diastole. The accurate timing of heart murmurs is the first step in their identification. The distinction between S_1 and S_2, and therefore systole and diastole is usually a straightforward process but can be difficult in the setting of a tachyarrhythmia, in which case the heart sounds can be distinguished by simultaneous palpation of the carotid upstroke, which should closely follow S_1.

FIGURE 38-1

Diagram depicting principal heart murmurs. ***A.*** Presystolic murmur of mitral or tricuspid stenosis. ***B.*** Holosystolic (pansystolic) murmur of mitral or tricuspid regurgitation or of ventricular septal defect. ***C.*** Aortic ejection murmur beginning with an ejection click and fading before the second heart sound. ***D.*** Systolic murmur in pulmonic stenosis spilling through the aortic second sound, pulmonic valve closure being delayed. ***E.*** Aortic or pulmonary diastolic murmur. ***F.*** Long diastolic murmur of mitral stenosis after the opening snap (OS). ***G.*** Short mid-diastolic inflow murmur after a third heart sound. ***H.*** Continuous murmur of patent ductus arteriosus. *(Adapted from P Wood: Diseases of the Heart and Circulation, London, Eyre & Spottiswood, 1968. Permission granted courtesy of Antony and Julie Wood.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Duration and Character

The duration of a heart murmur depends on the length of time over which a pressure difference exists between two cardiac chambers, the left ventricle and the aorta, the right ventricle and the pulmonary artery, or the great vessels. The magnitude and variability of this pressure difference, coupled with the geometry and compliance of the involved chambers or vessels, dictate the velocity of flow; the degree of turbulence; and the resulting frequency, configuration, and intensity of the murmur. The diastolic murmur of chronic aortic regurgitation (AR) is a blowing, high-frequency event, whereas the murmur of mitral stenosis (MS), indicative of the left atrial–left ventricular diastolic pressure gradient, is a low-frequency event, heard as a rumbling sound with the bell of the stethoscope. The frequency components of a heart murmur may vary at different sites of auscultation. The coarse systolic murmur of aortic stenosis (AS) may sound higher pitched and more acoustically pure at the apex, a phenomenon eponymously referred to as the *Gallavardin effect*. Some murmurs may have a distinct or unusual quality, such as the "honking" sound appreciated in some patients with mitral regurgitation (MR) due to mitral valve prolapse (MVP).

The configuration of a heart murmur may be described as crescendo, decrescendo, crescendo-decrescendo, or plateau. The decrescendo configuration of the murmur of chronic AR **(Fig. 38-1E)** can be understood in terms of the progressive decline in the diastolic pressure gradient between the aorta and the left ventricle. The crescendo-decrescendo configuration of the murmur of AS reflects the changes in the systolic pressure gradient between the left ventricle and the aorta as ejection occurs, whereas the

plateau configuration of the murmur of chronic MR (Fig. 38-1B) is consistent with the large and nearly constant pressure difference between the left ventricle and the left atrium.

Intensity

The intensity of a heart murmur is graded on a scale of 1–6 (or I–VI). A grade 1 murmur is very soft and is heard only with great effort. A grade 2 murmur is easily heard but not particularly loud. A grade 3 murmur is loud but is not accompanied by a palpable thrill over the site of maximal intensity. A grade 4 murmur is very loud and accompanied by a thrill. A grade 5 murmur is loud enough to be heard with only the edge of the stethoscope touching the chest, whereas a grade 6 murmur is loud enough to be heard with the stethoscope slightly off the chest. Murmurs of grade 3 or greater intensity usually signify important structural heart disease and indicate high blood flow velocity at the site of murmur production. Small ventricular septal defects (VSDs), for example, are accompanied by loud, usually grade 4 or greater, systolic murmurs as blood is ejected at high velocity from the left ventricle to the right ventricle. Low-velocity events, such as left-to-right shunting across an atrial septal defect (ASD), are usually silent. The intensity of a heart murmur may be diminished by any process that increases the distance between the intracardiac source and the stethoscope on the chest wall, such as obesity, obstructive lung disease, or a large pericardial effusion. The intensity of a murmur also may be misleadingly soft when cardiac output is reduced significantly or when the pressure gradient between the involved cardiac structures is low.

Location and Radiation

Recognition of the location and radiation of the murmur help facilitate its accurate identification (Fig. 38-2). Adventitious sounds, such as a systolic click or diastolic snap, or abnormalities of S_1 or S_2 may provide additional clues. Careful attention to the characteristics of the murmur and other heart sounds during the respiratory cycle and the performance of simple bedside maneuvers complete the auscultatory examination. These features, along with recommendations for further testing, are discussed below in the context of specific systolic, diastolic, and continuous heart murmurs (Table 38-1).

TABLE 38-1

Principal Causes of Heart Murmurs

Systolic Murmurs
Early systolic
Mitral
Acute MR
VSD
Muscular
Nonrestrictive with pulmonary hypertension
Tricuspid
TR with normal pulmonary artery pressure
Midsystolic
Aortic
Obstructive
Supravalvular–supravalvular AS, coarctation of the aorta
Valvular–AS and aortic sclerosis
Subvalvular–discrete, tunnel or HOCM
Increased flow, hyperkinetic states, AR, complete heart block
Dilation of ascending aorta, atheroma, aortitis
Pulmonary
Obstructive
Supravalvular–pulmonary artery stenosis
Valvular–pulmonic valve stenosis
Subvalvular–infundibular stenosis (dynamic)
Increased flow, hyperkinetic states, left-to-right shunt (e.g., ASD)
Dilation of pulmonary artery

Late systolic
Mitral
MVP, acute myocardial ischemia
Tricuspid
TVP
Holosystolic
Atrioventricular valve regurgitation (MR, TR)
Left-to-right shunt at ventricular level (VSD)
Early Diastolic Murmurs
AR
Valvular: congenital (bicuspid valve), rheumatic deformity, endocarditis, prolapse, trauma, post-valvulotomy
Dilation of valve ring: aorta dissection, annuloaortic ectasia, cystic medial degeneration, hypertension, ankylosing spondylitis
Widening of commissures: syphilis
Pulmonic regurgitation
Valvular: post-valvulotomy, endocarditis, rheumatic fever, carcinoid
Dilation of valve ring: pulmonary hypertension; Marfan syndrome
Congenital: isolated or associated with tetralogy of Fallot, VSD, pulmonic stenosis
Mid-Diastolic Murmurs
Mitral
MS
Carey-Coombs murmur (mid-diastolic apical murmur in acute rheumatic fever)
Increased flow across nonstenotic mitral valve (e.g., MR, VSD, PDA, high-output states, and complete heart block)
Tricuspid
Tricuspid stenosis
Increased flow across nonstenotic tricuspid valve (e.g., TR, ASD, and anomalous pulmonary venous return)
Left and right atrial tumors (myxoma)

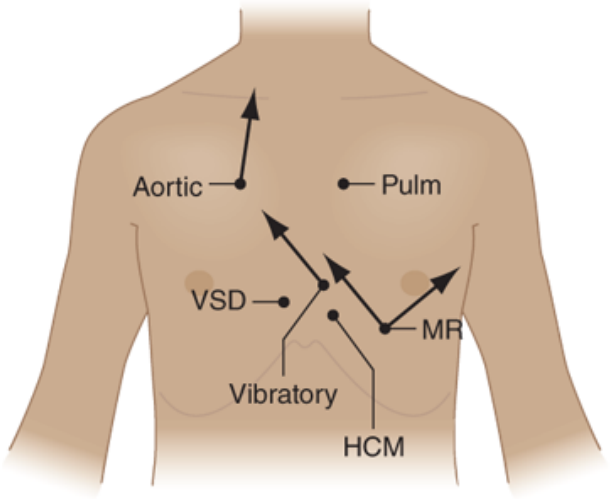
Severe AR (Austin Flint murmur)	
Continuous Murmurs	
Patent ductus arteriosus	Proximal coronary artery stenosis
Coronary AV fistula	Mammary souffle of pregnancy
Ruptured sinus of Valsalva aneurysm	Pulmonary artery branch stenosis
Aortic septal defect	Bronchial collateral circulation
Cervical venous hum	Small (restrictive) ASD with MS
Anomalous left coronary artery	Intercostal AV fistula

Abbreviations: AR, aortic regurgitation; AS, aortic stenosis; ASD, atrial septal defect; AV, arteriovenous; HOCM, hypertrophic obstructive cardiomyopathy; MR, mitral regurgitation; MS, mitral stenosis; MVP, mitral valve prolapse; PDA, patent ductus arteriosus; TR, tricuspid regurgitation; TVP, tricuspid valve prolapse; VSD, ventricular septal defect.

Source: E Braunwald, JK Perloff, in D Zipes et al (eds): *Braunwald's Heart Disease*, 7th ed. Philadelphia, Elsevier, 2005; PJ Norton, RA O'Rourke, in E Braunwald, L Goldman (eds): *Primary Cardiology*, 2nd ed. Philadelphia, Elsevier, 2003.

FIGURE 38-2

Maximal intensity and radiation of six isolated systolic murmurs. Aortic, aortic stenosis; HOCM, hypertrophic obstructive cardiomyopathy; MR, mitral regurgitation; Pulm, pulmonary stenosis; VSD, ventricular septal defect. *(From JB Barlow: Perspectives on the Mitral Valve. Philadelphia, FA Davis, 1987, p 140.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SYSTOLIC HEART MURMURS

Early Systolic Murmurs

Early systolic murmurs begin with S_1 and extend for a variable period, ending well before S_2. Their causes are relatively few in number. *Acute*, *severe MR* into a normal-sized, relatively noncompliant left atrium results in an early, decrescendo systolic

murmur best heard at or just medial to the apical impulse. These characteristics reflect the progressive attenuation of the pressure gradient between the left ventricle and the left atrium during systole owing to the rapid rise in left atrial pressure caused by the sudden volume load into an unprepared, noncompliant chamber and contrast sharply with the auscultatory features of chronic MR. Clinical settings in which acute, severe MR occur include (1) papillary muscle rupture complicating acute myocardial infarction (MI) **(Chap. 269)**, (2) rupture of chordae tendineae in the setting of myxomatous mitral valve disease (MVP, **Chap. 260**), (3) infective endocarditis **(Chap. 123)**, and (4) blunt chest wall trauma.

Acute, severe MR from papillary muscle rupture usually accompanies an inferior, posterior, or lateral MI, and occurs 2–7 days after presentation. It often is signaled by chest pain, hypotension, and pulmonary edema, but a murmur may be absent in up to 50% of cases. The posteromedial papillary muscle is involved 6 to 10 times more frequently than the anterolateral papillary muscle. The murmur is to be distinguished from that associated with post-MI ventricular septal rupture, which is accompanied by a systolic thrill at the left sternal border in nearly all patients and is holosystolic in duration. A new heart murmur after an MI is an indication for transthoracic echocardiography (TTE) **(Chap. 236)**, which allows bedside delineation of its etiology and pathophysiologic significance. The distinction between acute MR and ventricular septal rupture also can be achieved with right-sided heart catheterization, sequential determination of oxygen saturations, and analysis of the pressure waveforms (tall *v* wave in the pulmonary artery wedge pressure in MR). Post-MI mechanical complications of this nature mandate aggressive medical stabilization and prompt referral for surgical repair.

Spontaneous chordal rupture can complicate the course of myxomatous mitral valve disease (MVP) and result in new-onset or “acute on chronic” severe MR. MVP may occur as an isolated phenomenon, or the lesion may be part of a more generalized connective tissue disorder as seen, for example, in patients with Marfan syndrome. Acute, severe MR as a consequence of infective endocarditis results from destruction of leaflet tissue, chordal rupture, or both. Blunt chest wall trauma is usually self-evident but may be disarmingly trivial; it can result in papillary muscle contusion and rupture, chordal detachment, or leaflet avulsion. TTE is indicated in all cases of suspected acute, severe MR to define its mechanism and severity, delineate left ventricular size and systolic function, and provide an assessment of suitability for primary valve repair.

A congenital, small muscular VSD **(Chap. 264)** may be associated with an early systolic murmur. The defect closes progressively during septal contraction, and thus the murmur is confined to early systole. It is localized to the left sternal border (Fig. 38-2) and is usually of grade 4 or 5 intensity. Signs of pulmonary hypertension or left ventricular volume overload are absent. Anatomically large and uncorrected VSDs, which usually involve the membranous portion of the septum, may lead to pulmonary hypertension. The murmur associated with the left-to-right shunt, which earlier may have been holosystolic, becomes limited to the first portion of systole as the elevated pulmonary vascular resistance leads to an abrupt rise in right ventricular pressure and an attenuation of the interventricular pressure gradient during the remainder of the cardiac cycle. In such instances, signs of pulmonary hypertension (right ventricular lift, loud and single or closely split S_2) may predominate. The murmur is best heard along the left sternal border but is softer. Suspicion of a VSD is an indication for TTE.

Tricuspid regurgitation (TR) with normal pulmonary artery pressures, as may occur with infective endocarditis, may produce an early systolic murmur. The murmur is soft (grade 1 or 2), is best heard at the lower left sternal border and may increase in intensity with inspiration (Carvallo’s sign). Regurgitant *c-v* waves may be visible in the jugular venous pulse. TR in this setting is not associated with signs of right heart failure.

Midsystolic Murmurs

Midsystolic murmurs begin at a short interval after S_1, end before S_2 **(Fig. 38-1C)** and are usually crescendo-decrescendo in configuration. AS is the most common cause of a midsystolic murmur in an adult. The murmur of AS is usually loudest to the right of the sternum in the second intercostal space (aortic area, Fig. 38-2) and radiates into the carotids. Transmission of the midsystolic murmur to the apex, where it becomes higher-pitched, is common (Gallavardin effect; see above).

Differentiation of this apical systolic murmur from MR can be difficult. The murmur of AS will increase in intensity or become louder, in the beat after a premature beat, whereas the murmur of MR will have constant intensity from beat to beat. The intensity

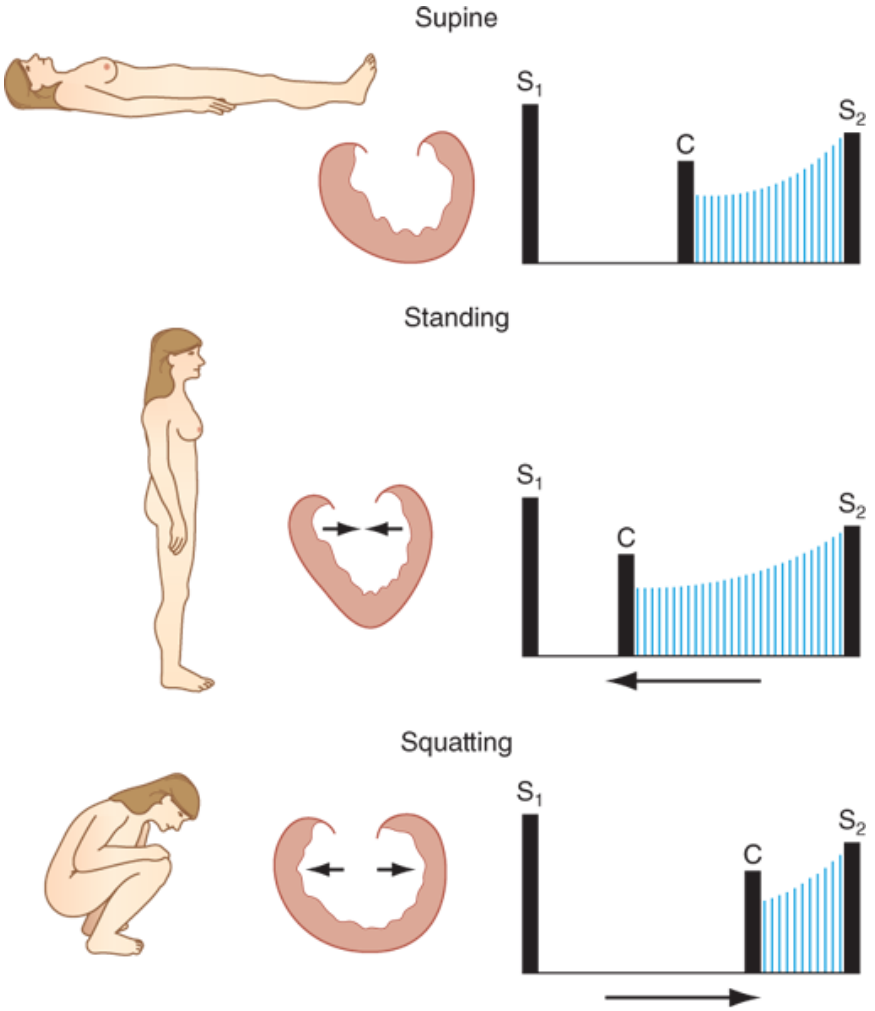
of the AS murmur also varies directly with the cardiac output. With a normal cardiac output, a systolic thrill and a grade 4 or higher murmur suggest severe AS. The murmur is softer in the setting of heart failure and low cardiac output. Other auscultatory findings of severe AS include a soft or absent A_2, paradoxical splitting of S_2, an apical S_4, and a late-peaking systolic murmur. In children, adolescents, and young adults with congenital valvular AS, an early ejection sound (click) is usually audible, more often along the left sternal border than at the base. Its presence signifies a flexible, noncalcified bicuspid valve (or one of its variants) and localizes the left ventricular outflow obstruction to the valvular (rather than sub- or supravalvular) level.

Assessment of the volume and rate of rise of the carotid pulse can provide additional information. A small and delayed upstroke (*parvus et tardus*) is consistent with severe AS. The carotid pulse examination is less discriminatory, however, in older patients with stiffened arteries. The electrocardiogram (ECG) shows signs of left ventricular hypertrophy (LVH) as the severity of the stenosis increases. TTE is indicated to assess the anatomic features of the aortic valve, the severity of the stenosis, left ventricular size, wall thickness and function, and the size and contour of the aortic root and proximal ascending aorta.

The obstructive form of hypertrophic cardiomyopathy (HOCM) is associated with a midsystolic murmur that is usually loudest along the left sternal border or between the left lower sternal border and the apex (**Chap. 254**, Fig. 38-2). The murmur is produced by both dynamic left ventricular outflow tract obstruction and MR, and thus, its configuration is a hybrid between ejection and regurgitant phenomena. The intensity of the murmur may vary from beat to beat and after provocative maneuvers but usually does not exceed grade 3. The murmur classically will increase in intensity with maneuvers that result in increasing degrees of outflow tract obstruction, such as a reduction in preload or afterload (Valsalva, standing, vasodilators), or with an augmentation of contractility (inotropic stimulation). Maneuvers that increase preload (squatting, passive leg raising, volume administration) or afterload (squatting, vasopressors) or that reduce contractility (β-adrenoreceptor blockers) decrease the intensity of the murmur. In rare patients, there may be reversed splitting of S_2. A sustained left ventricular apical impulse and an S_4 may be appreciated. In contrast to AS, the carotid upstroke is rapid and of normal volume. Rarely, it is bisferiens or bifid in contour **(see Fig. 234-2D)** due to midsystolic closure of the aortic valve. LVH is present on the ECG, and the diagnosis is confirmed by TTE. Although the systolic murmur associated with MVP behaves similarly to that due to HOCM in response to the Valsalva maneuver and to standing/squatting **(Fig. 38-3)**, these two lesions can be distinguished on the basis of their associated findings, such as the presence of LVH in HOCM or a nonejection click in MVP.

FIGURE 38-3

A midsystolic nonejection sound (C) occurs in mitral valve prolapse and is followed by a late systolic murmur that crescendos to the second heart sound (S_2). Standing decreases venous return; the heart becomes smaller; C moves closer to the first heart sound (S_1), and the mitral regurgitant murmur has an earlier onset. With prompt squatting, venous return and afterload increase; the heart becomes larger; C moves toward S_2; and the duration of the murmur shortens. The systolic murmur of hypertrophic obstructive cardiomyopathy behaves similarly. *(From JA Shaver, JJ Leonard, DF Leon: Examination of the Heart, Part IV, Auscultation of the Heart. Dallas, American Heart Association, 1990, p 13. Copyright, American Heart Association.)*

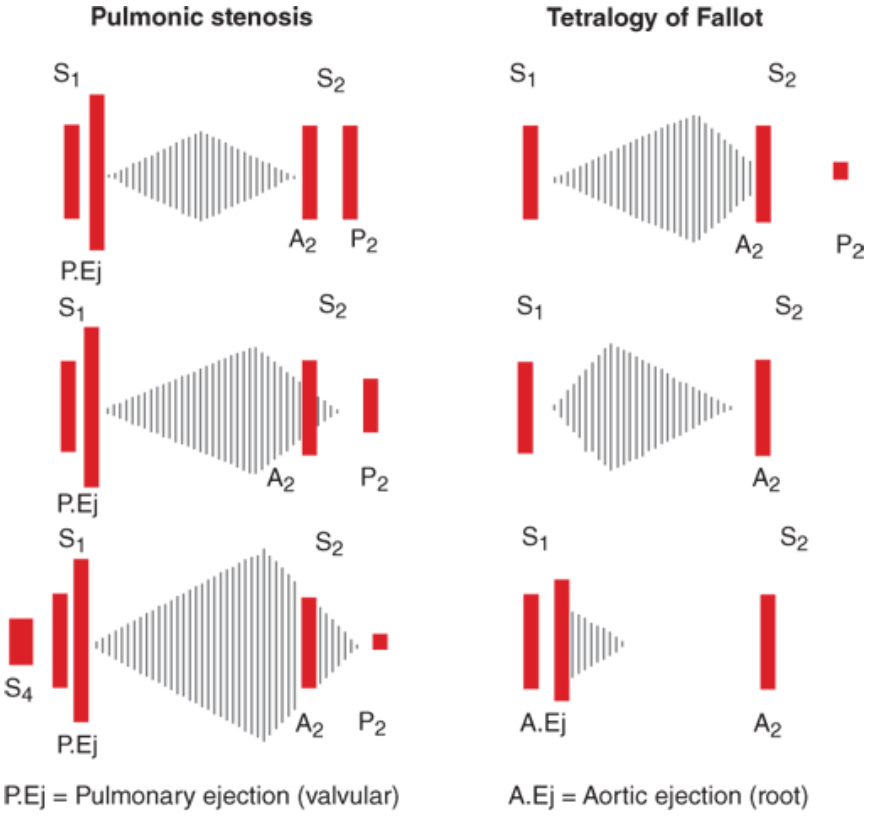


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The midsystolic, crescendo-decrescendo murmur of congenital pulmonic stenosis (PS, Chap. 264) is best appreciated in the second and third left intercostal spaces (pulmonic area) (Figs. 38-2 and 38-4). The duration of the murmur lengthens and the intensity of P_2 diminishes with increasing degrees of valvular stenosis (Fig. 38-1D). An early ejection sound, the intensity of which decreases with inspiration, is heard in younger patients. A parasternal lift and ECG evidence of right ventricular hypertrophy indicate severe pressure overload. If obtained, the chest x-ray may show poststenotic dilation of the main pulmonary artery. TTE is recommended for complete characterization.

FIGURE 38-4

Left. In valvular pulmonic stenosis with intact ventricular septum, right ventricular systolic ejection becomes progressively longer, with increasing obstruction to flow. As a result, the murmur becomes longer and louder, enveloping the aortic component of the second heart sound (A_2). The pulmonic component (P_2) occurs later, and splitting becomes wider but more difficult to hear because A_2 is lost in the murmur and P_2 becomes progressively fainter and lower pitched. As the pulmonic gradient increases, the isometric contraction phase shortens until the pulmonic valve ejection sound fuses with the first heart sound (S_1). In severe pulmonic stenosis with concentric hypertrophy and decreasing right ventricular compliance, a fourth heart sound appears. ***Right.*** In tetralogy of Fallot with increasing obstruction at the pulmonic infundibular area, an increasing amount of right ventricular blood is shunted across the silent ventricular septal defect and flow across the obstructed outflow tract decreases. Therefore, with increasing obstruction the murmur becomes shorter, earlier, and fainter. P_2 is absent in severe tetralogy of Fallot. A large aortic root receives almost all cardiac output from both ventricular chambers, and the aorta dilates and is accompanied by a root ejection sound that does not vary with respiration. *(From JA Shaver, JJ Leonard, DF Leon: Examination of the Heart, Part IV, Auscultation of the Heart. Dallas, American Heart Association, 1990, p 45. Copyright, American Heart Association.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Significant left-to-right intracardiac shunting due to an ASD **(Chap. 264)** leads to an increase in pulmonary blood flow and a grades 2–3 midsystolic murmur at the middle to upper left sternal border attributed to increased flow rates across the pulmonic valve with fixed splitting of S_2. Ostium secundum ASDs are the most common cause of these shunts in adults. Features suggestive of a primum ASD include the coexistence of MR due to a cleft anterior mitral valve leaflet and left axis deviation of the QRS complex on the ECG. With sinus venosus ASDs, the left-to-right shunt is usually not large enough to result in a systolic murmur, although the ECG may show abnormalities of sinus node function. A grade 2 or 3 midsystolic murmur may also be heard best at the upper left sternal border in patients with idiopathic dilation of the pulmonary artery; a pulmonary ejection sound is also present in these patients. TTE is indicated to evaluate a grade 2 or 3 midsystolic murmur when there are other signs of cardiac disease.

An isolated grade 1 or 2 midsystolic murmur, heard in the absence of symptoms or signs of heart disease, is most often a benign finding for which no further evaluation, including TTE, is necessary. The most common example of a murmur of this type in an older adult patient is the crescendo-decrescendo murmur of aortic valve sclerosis, heard at the second right interspace (Fig. 38-2). Aortic sclerosis is defined as focal thickening and calcification of the aortic valve to a degree that does not interfere with leaflet opening. The carotid upstrokes are normal, and electrocardiographic LVH is not present. A grade 1 or 2 midsystolic murmur often can be heard at the left sternal border with pregnancy, hyperthyroidism, or anemia, physiologic states that are associated with accelerated blood flow. *Still's murmur* refers to a benign grade 2, vibratory or musical midsystolic murmur at the mid or lower left sternal border in normal children and adolescents, best heard in the supine position (Fig. 38-2).

Late Systolic Murmurs

A late systolic murmur that is best heard at the left ventricular apex is usually due to MVP **(Chap. 260)**. Often, this murmur is introduced by one or more nonejection clicks. The radiation of the murmur can help identify the specific mitral leaflet involved in the process of prolapse or flail. The term *flail* refers to the movement made by an unsupported portion of the leaflet (usually the tip) after loss of its chordal attachment(s). With posterior leaflet prolapse or flail, the resultant jet of MR is directed anteriorly and medially, as a result of which the murmur radiates to the base of the heart and masquerades as AS. Anterior leaflet prolapse or flail results in a posteriorly directed MR jet that radiates to the axilla or left infrascapular region. Leaflet flail is associated with a

murmur of grade 3 or 4 intensity that can be heard throughout the precordium in thin-chested patients. The presence of an S_3 or a short, rumbling mid-diastolic murmur due to enhanced flow signifies severe MR.

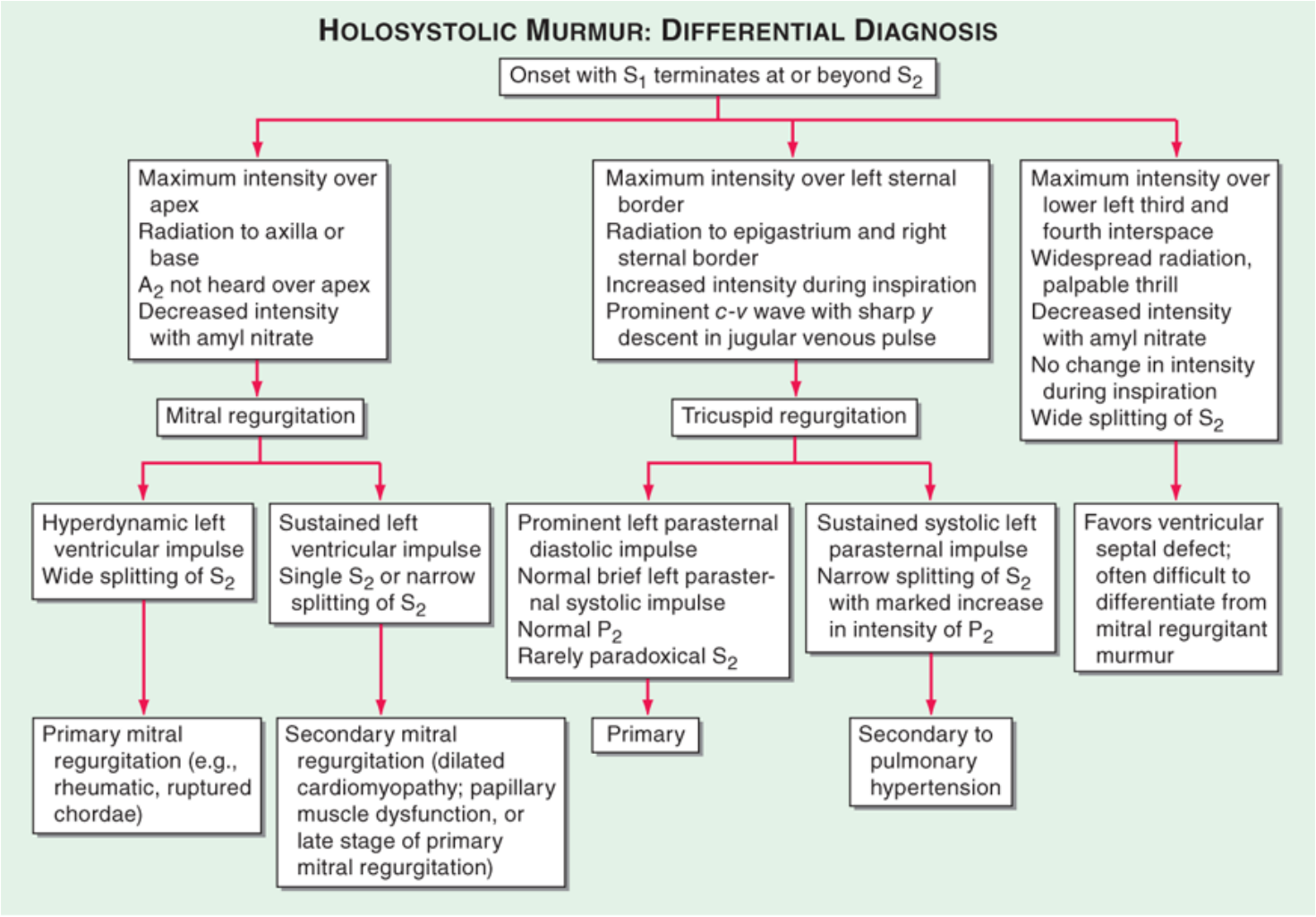
Bedside maneuvers that decrease left ventricular preload, such as standing, will cause the click and murmur of MVP to move closer to the first heart sound, as leaflet prolapse occurs earlier in systole. Standing also causes the murmur to become louder and longer. With squatting, left ventricular preload and afterload are increased abruptly, leading to an increase in left ventricular volume, and the click and murmur move away from the first heart sound as leaflet prolapse is delayed; the murmur becomes softer and shorter in duration (Fig. 38–3). As noted above, these responses to standing and squatting are directionally similar to those observed in patients with HOCM.

A late, apical systolic murmur indicative of MR may be heard transiently in the setting of acute myocardial ischemia; it is due to apical tethering and malcoaptation of the leaflets in response to structural and functional changes of the ventricle and mitral annulus. The intensity of the murmur varies as a function of left ventricular afterload and will increase in the setting of hypertension. TTE is recommended for assessment of late systolic murmurs.

Holosystolic Murmurs

(Figs. 38-1B and 38-5) Holosystolic murmurs begin with S_1 and continue through systole to S_2. They are usually indicative of chronic mitral or tricuspid valve regurgitation or a VSD and warrant TTE for further characterization. The holosystolic murmur of chronic MR is best heard at the left ventricular apex and radiates to the axilla (Fig. 38-2); it is usually high-pitched and plateau in configuration because of the wide difference between left ventricular and left atrial pressure throughout systole. In contrast to acute MR, left atrial compliance is normal or even increased in chronic MR. As a result, there is only a small increase in left atrial pressure for any increase in regurgitant volume.

FIGURE 38-5
Differential diagnosis of a holosystolic murmur.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Several conditions are associated with chronic MR and an apical holosystolic murmur, including rheumatic scarring of the leaflets, mitral annular calcification, postinfarction left ventricular remodeling, and severe left ventricular chamber enlargement. The circumference of the mitral annulus increases as the left ventricle enlarges and leads to failure of leaflet coaptation with central MR in patients with dilated cardiomyopathy **(Chap. 254)**. The severity of the MR is worsened by any contribution from apical displacement of the papillary muscles and leaflet tethering (remodeling). Because the mitral annulus is contiguous with the left atrial endocardium, gradual enlargement of the left atrium from chronic MR will result in further stretching of the annulus and more MR; thus, "MR begets MR." Chronic severe MR results in enlargement and leftward displacement of the left ventricular apex beat and, in some patients, a diastolic filling complex, as described previously **(Fig. 38-1G)**.

The holosystolic murmur of chronic TR is generally softer than that of MR, is loudest at the left lower sternal border, and usually increases in intensity with inspiration (Carvallo's sign). Associated signs include *c-v* waves in the jugular venous pulse, an enlarged and pulsatile liver, ascites, and peripheral edema. The abnormal jugular venous waveforms are the predominant finding and seen very often in the absence of an audible murmur despite Doppler echocardiographic verification of TR. Causes of *primary* TR include myxomatous disease (prolapse), endocarditis, rheumatic disease, radiation, carcinoid, Ebstein's anomaly, and chordal detachment as a complication of right ventricular endomyocardial biopsy. TR is much more commonly a passive process that results secondarily from annular enlargement due to right ventricular dilation in the face of volume or pressure overload or adverse right ventricular remodeling.

The holosystolic murmur of a VSD is loudest at the mid- to lower-left sternal border (Fig. 38-2) and radiates widely. A thrill is present at the site of maximal intensity in the majority of patients. There is no change in the intensity of the murmur with inspiration. The intensity of the murmur varies as a function of the anatomic size of the defect. Small, restrictive VSDs, as

exemplified by the *maladie de Roger*, create a very loud murmur due to the significant and sustained systolic pressure gradient between the left and right ventricles. With large defects, the ventricular pressures tend to equalize, shunt flow is balanced, and a murmur is not appreciated. The distinction between post-MI ventricular septal rupture and MR has been reviewed previously.

DIASTOLIC HEART MURMURS

Early Diastolic Murmurs

(Fig. 38-1E) Chronic AR results in a high-pitched, blowing, decrescendo, early- to mid-diastolic murmur that begins after the aortic component of S_2 (A_2), and is best heard at the second right interspace. The murmur may be soft and difficult to hear unless auscultation is performed with the patient leaning forward at end expiration. This maneuver brings the aortic root closer to the anterior chest wall. Radiation of the murmur may provide a clue to the cause of the AR. With primary valve disease, such as that due to congenital bicuspid disease, prolapse, or endocarditis, the diastolic murmur tends to radiate along the left sternal border, where it is often louder than appreciated in the second right interspace. When AR is caused by aortic root disease, the diastolic murmur may radiate along the right sternal border. Diseases of the aortic root cause dilation or distortion of the aortic annulus and failure of leaflet coaptation. Causes include Marfan syndrome with aneurysm formation, annuloaortic ectasia, ankylosing spondylitis, and aortic dissection.

Chronic, severe AR also may produce a lower-pitched mid to late, grade 1 or 2 diastolic murmur at the apex (Austin Flint murmur), which is thought to reflect turbulence at the mitral inflow area from the admixture of regurgitant (aortic) and forward (mitral) blood flow. This lower-pitched, apical diastolic murmur can be distinguished from that due to MS by the absence of an opening snap and the response of the murmur to a vasodilator challenge. Lowering afterload with an agent such as amyl nitrite will decrease the duration and magnitude of the aortic–left ventricular diastolic pressure gradient, and thus the Austin Flint murmur of severe AR will become shorter and softer. The intensity of the diastolic murmur of MS (Fig. 38-6) may either remain constant or increase with afterload reduction because of the reflex increase in cardiac output and mitral valve flow.

Although AS and AR may coexist, a grade 2 or 3 crescendo-decrescendo midsystolic murmur frequently is heard at the base of the heart in patients with isolated, severe AR and is due to an increased volume and rate of systolic flow. Accurate bedside identification of coexistent AS can be difficult unless the carotid pulse examination is abnormal or the midsystolic murmur is of grade 4 or greater intensity. In the absence of heart failure, chronic severe AR is accompanied by several peripheral signs of significant diastolic runoff, including a wide pulse pressure, a “water-hammer” carotid upstroke (Corrigan’s pulse), and Quincke’s pulsations of the nail beds. The diastolic murmur of *acute*, *severe AR* is notably shorter in duration and lower pitched than the murmur of chronic AR. It can be very difficult to appreciate in the presence of a rapid heart rate. These attributes reflect the abrupt rate of rise of diastolic pressure within the unprepared and noncompliant left ventricle and the correspondingly rapid decline in the aortic–left ventricular diastolic pressure gradient. Left ventricular diastolic pressure may increase sufficiently to result in premature closure of the mitral valve and a soft first heart sound. Peripheral signs of significant diastolic runoff are not present.

Pulmonic regurgitation (PR) results in a decrescendo, early to mid-diastolic murmur (*Graham Steell murmur*) that begins after the pulmonic component of S_2 (P_2), is best heard at the second left interspace, and radiates along the left sternal border. The intensity of the murmur may increase with inspiration. PR is most commonly due to dilation of the valve annulus from chronic elevation of the pulmonary artery pressure. Signs of pulmonary hypertension, including a right ventricular lift and a loud, single or narrowly split S_2, are present. These features also help distinguish PR from AR as the cause of a decrescendo diastolic murmur heard along the left sternal border. PR in the absence of pulmonary hypertension can occur with endocarditis or a congenitally deformed valve. It is usually present after repair of tetralogy of Fallot in childhood. When pulmonary hypertension is not present, the diastolic murmur is softer and lower pitched than the classic Graham Steell murmur, and the severity of the PR can be difficult to appreciate.

TTE is indicated for the further evaluation of a patient with an early to mid-diastolic murmur. Longitudinal assessment of lesion severity, ventricular size, and of systolic function helps guide a potential decision for surgical management. TTE also can provide

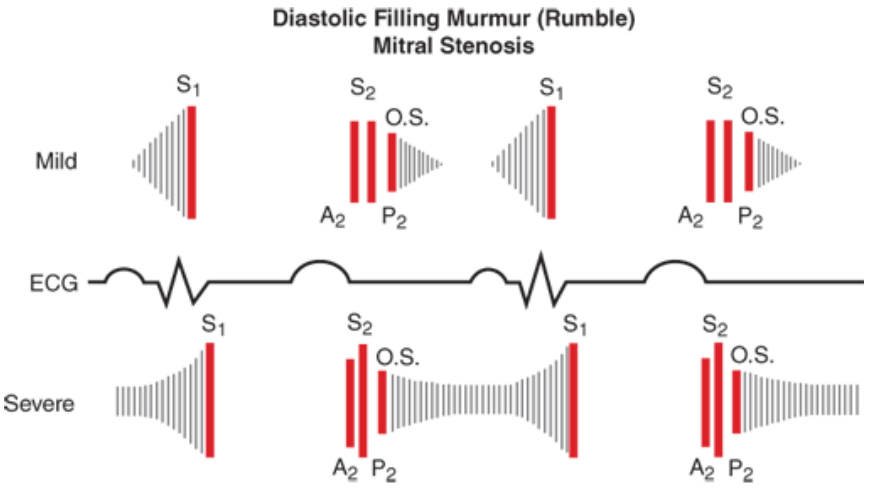
anatomic information regarding the root and proximal ascending aorta, although computed tomographic or magnetic resonance angiography may be indicated for more precise characterization **(Chap. 236)**.

Mid-Diastolic Murmurs

(**Figs. 38-1F** and 38-1G) Mid-diastolic murmurs result from obstruction and/or augmented flow at the level of the mitral or tricuspid valve. Rheumatic fever is the most common cause of MS (Fig. 38-6). In younger patients with pliable valves, S_1 is loud and the murmur begins after an opening snap, which is a high-pitched sound that occurs shortly after S_2. The interval between the pulmonic component of the second heart sound (P_2) and the opening snap is inversely related to the magnitude of the left atrial–left ventricular pressure gradient. The murmur of MS is low-pitched and thus is best heard with the bell of the stethoscope. It is loudest at the left ventricular apex and often is appreciated only when the patient is turned in the left lateral decubitus position. It is usually of grade 1 or 2 intensity but may be absent when the cardiac output is severely reduced despite significant obstruction. The intensity of the murmur increases during maneuvers that increase cardiac output and mitral valve flow, such as exercise. The duration of the murmur reflects the length of time over which left atrial pressure exceeds left ventricular diastolic pressure. An increase in the intensity of the murmur just before S_1, a phenomenon known as *presystolic accentuation* (Figs. 38-1A and 38-6), occurs in patients in sinus rhythm and is due to a late increase in transmitral flow with atrial contraction. Presystolic accentuation does not occur in patients with atrial fibrillation.

FIGURE 38-6

Diastolic filling murmur (rumble) in mitral stenosis. In mild mitral stenosis, the diastolic gradient across the valve is limited to the phases of rapid ventricular filling in early diastole and presystole. The rumble may occur during either or both periods. As the stenotic process becomes severe, a large pressure gradient exists across the valve during the entire diastolic filling period, and the rumble persists throughout diastole. As the left atrial pressure becomes greater, the interval between A_2 (or P_2) and the opening snap (O.S.) shortens. In severe mitral stenosis, secondary pulmonary hypertension develops and results in a loud P_2 and the splitting interval usually narrows. ECG, electrocardiogram. *(From JA Shaver, JJ Leonard, DF Leon: Examination of the Heart, Part IV, Auscultation of the Heart. Dallas, American Heart Association, 1990, p 55. Copyright, American Heart Association.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The mid-diastolic murmur associated with tricuspid stenosis is best heard at the lower left sternal border and increases in intensity with inspiration. A prolonged *y* descent may be visible in the jugular venous waveform. This murmur is very difficult to hear and often is obscured by left-sided acoustical events.

There are several other causes of mid-diastolic murmurs. Large left atrial myxomas may prolapse across the mitral valve and cause variable degrees of obstruction to left ventricular inflow **(Chap. 266)**. The murmur associated with an atrial myxoma may change in duration and intensity with changes in body position. An opening snap is not present, and there is no presystolic

accentuation. Augmented mitral diastolic flow can occur with isolated severe MR or with a large left-to-right shunt at the ventricular or great vessel level and produce a soft, rapid filling sound (S_3) followed by a short, low-pitched mid-diastolic apical murmur (Fig. 38-1G). The Austin Flint murmur of severe, chronic AR has already been described.

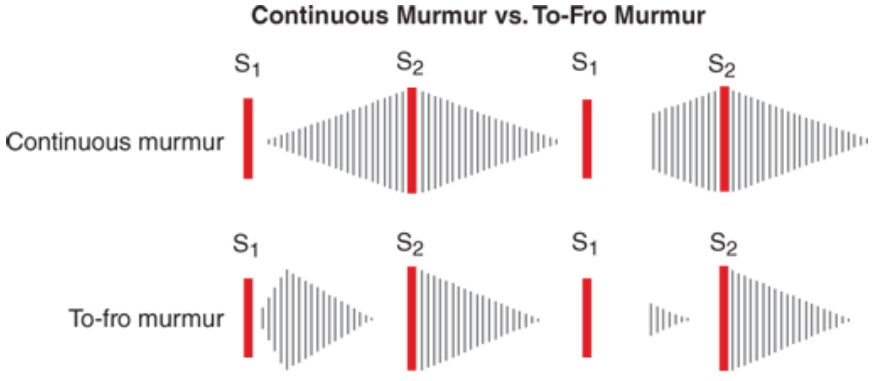
A short, mid-diastolic murmur is rarely heard during an episode of acute rheumatic fever (Carey-Coombs murmur) and probably is due to flow through an edematous mitral valve. An opening snap is not present in the acute phase, and the murmur dissipates with resolution of the acute attack. Complete heart block with dyssynchronous atrial and ventricular activation may be associated with intermittent mid- to late diastolic murmurs if atrial contraction occurs when the mitral valve is partially closed. Mid-diastolic murmurs indicative of increased tricuspid valve flow can occur with severe, isolated TR and with large ASDs and significant left-to-right shunting. Other signs of an ASD are present **(Chap. 264)**, including fixed splitting of S_2 and a midsystolic murmur at the mid- to upper left sternal border. TTE is indicated for evaluation of a patient with a mid- to late diastolic murmur. Findings specific to the diseases discussed above will help guide management.

CONTINUOUS MURMURS

(Figs. 38-1H and 38-7) Continuous murmurs begin in systole, peak near the second heart sound, and continue into all or part of diastole. Their presence throughout the cardiac cycle implies a pressure gradient between two chambers or vessels during both systole and diastole. The continuous murmur associated with a patent ductus arteriosus is best heard at the upper left sternal border. Large, uncorrected shunts may lead to pulmonary hypertension, attenuation or obliteration of the diastolic component of the murmur, reversal of shunt flow, and differential cyanosis of the lower extremities. A ruptured sinus of Valsalva aneurysm creates a continuous murmur of abrupt onset at the upper right sternal border. Rupture typically occurs into a right heart chamber, and the murmur is indicative of a continuous pressure difference between the aorta and either the right ventricle or the right atrium. A continuous murmur also may be audible along the left sternal border with a coronary arteriovenous fistula and at the site of an arteriovenous fistula used for hemodialysis access. Enhanced flow through enlarged intercostal collateral arteries in patients with aortic coarctation may produce a continuous murmur along the course of one or more ribs. A cervical bruit with both systolic and diastolic components (a to-fro murmur, Fig. 38-7) usually indicates a high-grade carotid artery stenosis.

FIGURE 38-7

Comparison of the continuous murmur and the to-fro murmur. During abnormal communication between high-pressure and low-pressure systems, a large pressure gradient exists throughout the cardiac cycle, producing a continuous murmur. A classic example is patent ductus arteriosus. At times, this type of murmur can be confused with a to-fro murmur, which is a combination of systolic ejection murmur and a murmur of semilunar valve incompetence. A classic example of a to-fro murmur is aortic stenosis and regurgitation. A continuous murmur crescendos to near the second heart sound (S_2), whereas a to-fro murmur has two components. The midsystolic ejection component decrescendos and disappears as it approaches S_2. *(From JA Shaver, JJ Leonard, DF Leon: Examination of the Heart, Part IV, Auscultation of the Heart. Dallas, American Heart Association, 1990, p 55. Copyright, American Heart Association.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Not all continuous murmurs are pathologic. A continuous venous hum can be heard in healthy children and young adults, especially during pregnancy; it is best appreciated in the right supraclavicular fossa and can be obliterated by pressure over the right internal jugular vein or by having the patient turn his or her head toward the examiner. The continuous mammary souffle of pregnancy is created by enhanced arterial flow through engorged breasts and usually appears during the late third trimester or early puerperium. The murmur is louder in systole. Firm pressure with the diaphragm of the stethoscope can eliminate the diastolic portion of the murmur.

DYNAMIC AUSCULTATION

(Table 38-2; see Table 234-1) Careful attention to the behavior of heart murmurs during simple maneuvers that alter cardiac hemodynamics can provide important clues to their cause and significance.

TABLE 38-2

Dynamic Auscultation: Bedside Maneuvers That Can Be Used to Change the Intensity of Cardiac Murmurs (See Text)

1. Respiration
2. Isometric exercise (handgrip)
3. Transient arterial occlusion
4. Pharmacologic manipulation of preload and/or afterload
5. Valsalva maneuver
6. Rapid standing/squatting
7. Passive leg raising
8. Post-premature beat

Respiration

Auscultation should be performed during quiet respiration or with a modest increase in inspiratory effort, as more forceful movement of the chest tends to obscure the heart sounds. Left-sided murmurs may be best heard at end expiration, when lung volumes are minimized and the heart and great vessels are brought closer to the chest wall. This phenomenon is characteristic of the murmur of AR. Murmurs of right-sided origin, such as tricuspid or pulmonic regurgitation, increase in intensity during inspiration. The intensity of left-sided murmurs either remains constant or decreases with inspiration.

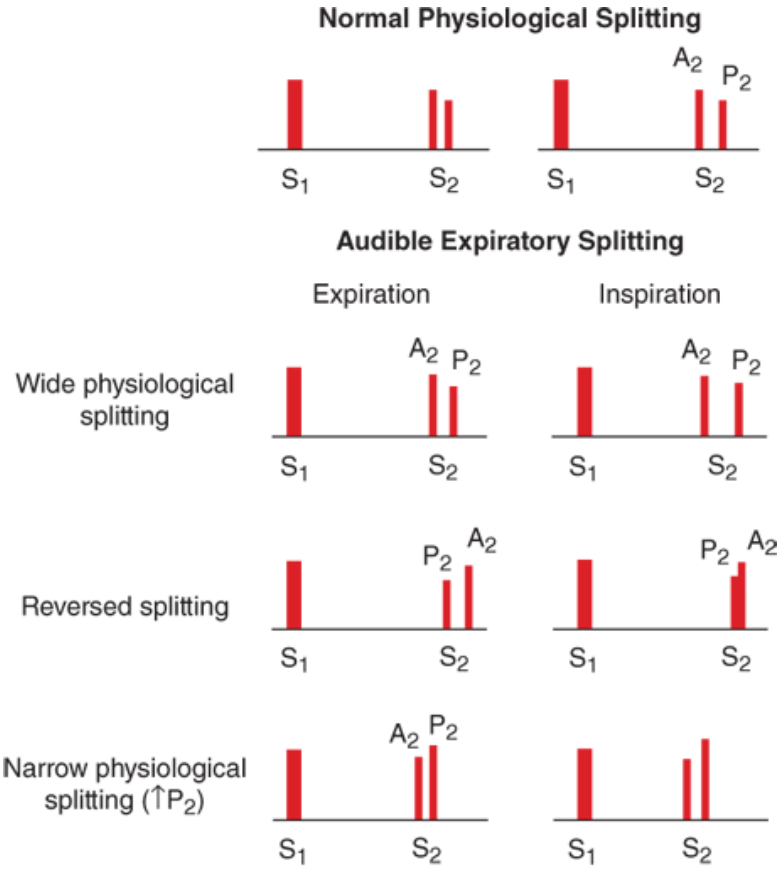
Bedside assessment also should evaluate the behavior of S_2 with respiration and the dynamic relationship between the aortic and pulmonic components **(Fig. 38-8)**. Reversed splitting can be a feature of severe AS, HOCM, left bundle branch block, right ventricular pacing, or acute myocardial ischemia. Fixed splitting of S_2 in the presence of a grade 2 or 3 midsystolic murmur at the mid- or upper left sternal border indicates an ASD. Physiologic but wide splitting during the respiratory cycle implies either premature aortic valve closure, as can occur with severe MR, or delayed pulmonic valve closure due to PS or right bundle branch block.

FIGURE 38-8

Top. Normal physiologic splitting. During expiration, the aortic (A_2) and pulmonic (P_2) components of the second heart sound are separated by <30 ms and are appreciated as a single sound. During inspiration, the splitting interval widens, and A_2 and P_2 are clearly separated into two distinct sounds. ***Bottom.*** Audible expiratory splitting. Wide physiologic splitting is caused by a delay in P_2 (as, for example, with right bundle branch block) or by early closure of the aortic valve (A_2, as for example with severe mitral regurgitation). Reversed splitting is caused by a delay in A_2, resulting in paradoxical movement; i.e., with inspiration P_2 moves toward A_2, and the splitting interval narrows. Narrow physiologic splitting occurs in pulmonary hypertension, and both A_2 and P_2

are heard during expiration at a narrow splitting interval because of the increased intensity and high-frequency composition of P_2.

(From JA Shaver, JJ Leonard, DF Leon: Examination of the Heart, Part IV, Auscultation of the Heart. Dallas, American Heart Association, 1990, p 17. Copyright, American Heart Association.)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Alterations of Systemic Vascular Resistance

Murmurs can change characteristics after maneuvers that alter systemic vascular resistance and left ventricular afterload. The systolic murmurs of MR and VSD become louder during sustained handgrip, simultaneous inflation of blood pressure cuffs on both upper extremities to pressures 20–40 mmHg above systolic pressure for 20 s, or infusion of a vasopressor agent. The murmurs associated with AS or HOCM will become softer or remain unchanged with these maneuvers. The diastolic murmur of AR becomes louder in response to interventions that raise systemic vascular resistance.

Opposite changes in systolic and diastolic murmurs may occur with the use of pharmacologic agents that lower systemic vascular resistance. Inhaled amyl nitrite is now rarely used for this purpose but can help distinguish the murmur of AS or HOCM from that of either MR or VSD, if necessary. The former two murmurs increase in intensity, whereas the latter two become softer after exposure to amyl nitrite. As noted previously, the Austin Flint murmur of severe AR becomes softer, but the mid-diastolic rumble of MS becomes louder, in response to the abrupt lowering of systemic vascular resistance with amyl nitrite.

Changes in Venous Return

The Valsalva maneuver results in an increase in intrathoracic pressure, followed by a decrease in venous return, ventricular filling, and cardiac output. The majority of murmurs decrease in intensity during the strain phase of the maneuver. Two notable exceptions are the murmurs associated with MVP and obstructive HOCM, both of which become louder during the Valsalva maneuver. The murmur of MVP may also become longer as leaflet prolapse occurs earlier in systole at smaller ventricular

volumes. These murmurs behave in a similar and parallel fashion with standing. Both the click and the murmur of MVP move closer in timing to S_1 on rapid standing from a squatting position (Fig. 38-3). The increase in the intensity of the murmur of HOCM is predicated on the augmentation of the dynamic left ventricular outflow tract gradient that occurs with reduced ventricular filling. Squatting results in abrupt increases in both venous return (preload) and left ventricular afterload that increase ventricular volume, changes that predictably cause a decrease in the intensity and duration of the murmurs associated with MVP and HOCM; the click and murmur of MVP move away from S_1 with squatting. Passive leg raising can be used to increase venous return in patients who are unable to squat and stand. This maneuver may lead to a decrease in the intensity of the murmur associated with HOCM but has less effect in patients with MVP.

Post-premature Ventricular Contraction

A change in the intensity of a systolic murmur in the first beat after a premature beat, or in the beat after a long cycle length in patients with atrial fibrillation, can help distinguish AS from MR, particularly in an older patient in whom the murmur of AS is well transmitted to the apex. Systolic murmurs due to left ventricular outflow obstruction, including that due to AS, increase in intensity in the beat after a premature beat because of the combined effects of enhanced left ventricular filling and post-extrasystolic potentiation of contractile function. Forward flow accelerates, causing an increase in the gradient and a louder murmur. The intensity of the murmur of MR does not change in the post-premature beat as there is relatively little further increase in mitral valve flow or change in the left ventricular–left atrial gradient.

THE CLINICAL CONTEXT

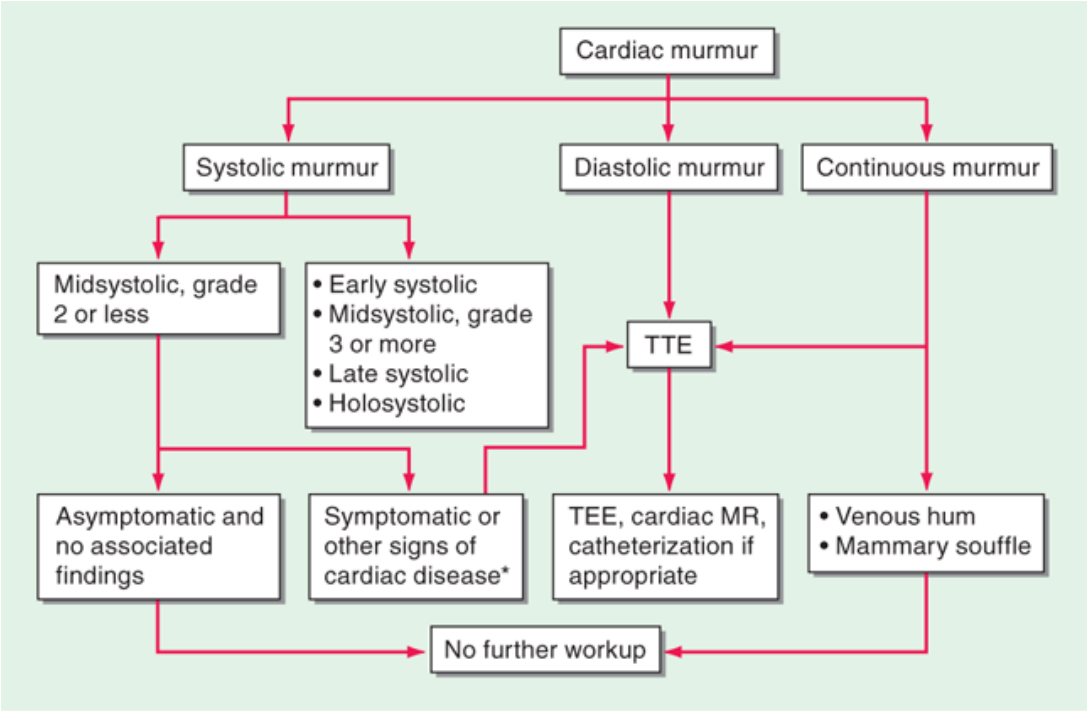
Additional clues to the etiology and importance of a heart murmur can be gleaned from the history and other physical examination findings. Symptoms suggestive of cardiovascular, neurologic, or pulmonary disease help focus the differential diagnosis, as do findings relevant to the jugular venous pressure and waveforms, the arterial pulses, other heart sounds, the lungs, the abdomen, the skin, and the extremities. In many instances, laboratory studies, an ECG, and/or a chest x-ray may have been obtained earlier and may contain valuable information. A patient with suspected infective endocarditis, for example, may have a murmur in the setting of fever, chills, anorexia, fatigue, dyspnea, splenomegaly, petechiae, and positive blood cultures. A new systolic murmur in a patient with a marked fall in blood pressure after a recent MI suggests myocardial rupture. By contrast, an isolated grade 1 or 2 midsystolic murmur at the left sternal border in a healthy, active, and asymptomatic young adult is most likely a benign finding for which no further evaluation is indicated. The context in which the murmur is appreciated often dictates the need for further testing and the pace of the evaluation.

ECHOCARDIOGRAPHY

(Fig. 38–9; Chaps. 234 and 236) Echocardiography with color flow and spectral Doppler is a valuable tool for the assessment of cardiac murmurs. Information regarding valve structure and function, chamber size, wall thickness, ventricular function, estimated pulmonary artery pressures, intracardiac shunt flow, pulmonary and hepatic vein flow, and aortic flow can be ascertained readily. It is important to note that Doppler signals of trace or mild valvular regurgitation of no clinical consequence can be detected with structurally normal tricuspid, pulmonic, and mitral valves. Such signals are not likely to generate enough turbulence to create an audible murmur.

FIGURE 38-9

Strategy for evaluating heart murmurs. *If an electrocardiogram or chest x-ray has been obtained and is abnormal, echocardiography is indicated. TTE, transthoracic echocardiography; TEE, transesophageal echocardiography; MR, magnetic resonance. *(Adapted from RO Bonow et al: J Am Coll Cardiol 32:1486, 1998.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Echocardiography is indicated for the evaluation of patients with early, late, or holosystolic murmurs and patients with grade 3 or louder midsystolic murmurs. Patients with grade 1 or 2 midsystolic murmurs but other symptoms or signs of cardiovascular disease, including those from ECG or chest x-ray, should also undergo echocardiography. Echocardiography is also indicated for the evaluation of any patient with a diastolic murmur and for patients with continuous murmurs not due to a venous hum or mammary souffle. Echocardiography should be considered when there is a clinical need to verify normal cardiac structure and function in a patient whose symptoms and signs are probably noncardiac in origin. The performance of serial echocardiography to follow the course of asymptomatic individuals with valvular heart disease is a central feature of their longitudinal assessment, and it provides valuable information that may have an impact on decisions regarding the timing of surgery. Routine echocardiography is *not* recommended for asymptomatic patients with a grade 1 or 2 midsystolic murmur without other signs of heart disease. For this category of patients, referral to a cardiovascular specialist should be considered if there is doubt about the significance of the murmur after the initial examination.

The selective use of echocardiography outlined above has not been subjected to rigorous analysis of its cost-effectiveness. For some clinicians, handheld or miniaturized cardiac ultrasound devices have replaced the stethoscope. Although several reports attest to the improved sensitivity of such devices for the detection of valvular heart disease (e.g., rheumatic heart disease in susceptible populations), accuracy is highly operator-dependent, and incremental cost considerations and outcomes have not been addressed adequately for most patient scenarios. The use of electronic or digital stethoscopes with spectral display capabilities has also been proposed as a method to improve the characterization of heart murmurs and the mentored teaching of cardiac auscultation.

OTHER CARDIAC TESTING

(**Chap. 236**, Fig. 38-9) In relatively few patients, clinical assessment and TTE do not adequately characterize the origin and significance of a heart murmur. Transesophageal echocardiography (TEE) can be considered for further evaluation, especially when the TTE windows are limited by body size, chest configuration, or intrathoracic pathology. TEE offers enhanced sensitivity for the detection of a wide range of structural cardiac disorders. Electrocardiographically gated cardiac magnetic resonance (CMR) imaging, although limited in its ability to display valvular morphology, can provide quantitative information regarding valvular function, stenosis severity, regurgitant fraction, regurgitant volume, shunt flow, chamber and great vessel size, ventricular function, and myocardial perfusion. CMR has largely supplanted the need for cardiac catheterization and invasive hemodynamic

assessment when there is a discrepancy between the clinical and echocardiographic findings. Invasive coronary angiography is performed routinely in most adult patients before valve surgery, especially when there is a suspicion of coronary artery disease predicated on symptoms, risk factors, and/or age. The use of computed tomography coronary angiography (CTCA) to exclude coronary artery disease in selected patients with a low pretest probability of disease before valve surgery has gained wider acceptance.

INTEGRATED APPROACH

The accurate identification of a heart murmur begins with a systematic approach to cardiac auscultation. Characterization of its major attributes, as reviewed above, allows the examiner to construct a preliminary differential diagnosis, which is then refined by integration of information available from the history, associated cardiac findings, the general physical examination, and the clinical context. The need for and urgency of further testing follow sequentially. Correlation of the findings on auscultation with the noninvasive data provides an educational feedback loop and an opportunity for improving physical examination skills. Cost constraints mandate that noninvasive imaging be justified on the basis of its incremental contribution to diagnosis, treatment, and outcome. Cardiac auscultation using a stethoscope remains a time-honored tradition in medicine, the benefits of which extend beyond accurate recognition of heart sounds. Selective augmentation with, rather than wholesale replacement by, handheld ultrasound and newer technologies may improve diagnostic accuracy and better guide therapeutic decisions.

FURTHER READING

Edelman ER, Weber BN: Tenuous tether. N Engl J Med 373:2199, 2015. [PubMed: 26630141]

Fang LC, O'Gara PT: The history and physical examination. An evidence-based approach, in *Braunwald's Heart Disease. A Textbook of Cardiovascular Medicine*, 10th ed, DL Mann et al (eds). Philadelphia, Elsevier/Saunders, 2015, pp 95–113.

Fuster V: The stethoscope's prognosis. Very much alive and very necessary. J Am Coll Cardiol 67:1118, 2016. [PubMed: 26780238]

Kimura BJ et al: Cardiac limited ultrasound examination techniques to augment the bedside cardiac physical examination. J Ultrasound Med 34:1683, 2015. [PubMed: 26269293]

Lai LS et al: Computerized automatic diagnosis of innocent and pathologic murmurs in pediatrics: A pilot study. Congen Heart Dis 11:386, 2016.

Nishimura R et al: 2014 AHA/ACC guideline for the management of patients with valvular heart disease. J Am Coll Cardiol 63:2438, 2014. [PubMed: 24603192]

Shrestha NR et al: Prevalence of subclinical rheumatic heart disease in Eastern Nepal: A school-based cross-sectional study. JAMA Cardiol 1:89, 2016. [PubMed: 27437661]

Stokke TM et al: Brief group training of medical students in focused cardiac ultrasound may improve diagnostic accuracy of physical examination. J Am Soc Echocardiogr 27:1238, 2014. [PubMed: 25216765]

Chapter 39: Palpitations

Joseph Loscalzo

INTRODUCTION

Palpitations are extremely common among patients who present to their internists and can best be defined as a “thumping,” “pounding,” or “fluttering” sensation in the chest. This sensation can be either intermittent or sustained and either regular or irregular. Most patients interpret palpitations as an unusual awareness of the heartbeat and become especially concerned when they sense that they have had “skipped” or “missing” heartbeats. Palpitations are often noted when the patient is quietly resting, during which time other stimuli are minimal. Palpitations that are positional generally reflect a structural process within (e.g., atrial myxoma) or adjacent to (e.g., mediastinal mass) the heart.

Palpitations are brought about by cardiac (43%), psychiatric (31%), miscellaneous (10%), and unknown (16%) causes, according to one large series. Among the cardiovascular causes are premature atrial and ventricular contractions, supraventricular and ventricular arrhythmias, mitral valve prolapse (with or without associated arrhythmias), aortic insufficiency, atrial myxoma, myocarditis, and pulmonary embolism. Intermittent palpitations are commonly caused by premature atrial or ventricular contractions: the post-extrasystolic beat is sensed by the patient owing to the increase in ventricular end-diastolic dimension following the pause in the cardiac cycle and the increased strength of contraction (post-extrasystolic potentiation) of that beat. Regular, sustained palpitations can be caused by regular supraventricular and ventricular tachycardias. Irregular, sustained palpitations can be caused by atrial fibrillation. It is important to note that most arrhythmias are not associated with palpitations. In those that are, it is often useful either to ask the patient to “tap out” the rhythm of the palpitations or to take his/her pulse during palpitations. In general, hyperdynamic cardiovascular states caused by catecholaminergic stimulation from exercise, stress, or pheochromocytoma can lead to palpitations. Palpitations are common among athletes, especially older endurance athletes. In addition, the enlarged ventricle of aortic regurgitation and accompanying hyperdynamic precordium frequently lead to the sensation of palpitations. Other factors that enhance the strength of myocardial contraction, including tobacco, caffeine, aminophylline, atropine, thyroxine, cocaine, and amphetamines, can cause palpitations.

Psychiatric causes of palpitations include panic attacks or disorders, anxiety states, and somatization, alone or in combination. Patients with psychiatric causes for palpitations more commonly report a longer duration of the sensation (>15 min) and other accompanying symptoms than do patients with other causes. Among

the miscellaneous causes of palpitations are thyrotoxicosis, drugs (see above) and ethanol, spontaneous skeletal muscle contractions of the chest wall, pheochromocytoma, and systemic mastocytosis.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Palpitations

The principal goal in assessing patients with palpitations is to determine whether the symptom is caused by a life-threatening arrhythmia. Patients with preexisting coronary artery disease (CAD) or risk factors for CAD are at greatest risk for ventricular arrhythmias **(Chap. 241)** as a cause for palpitations. In addition, the association of palpitations with other symptoms suggesting hemodynamic compromise, including syncope or lightheadedness, supports this diagnosis. Palpitations caused by sustained tachyarrhythmias in patients with CAD can be accompanied by angina pectoris or dyspnea, and, in patients with ventricular dysfunction (systolic or diastolic), aortic stenosis, hypertrophic cardiomyopathy, or mitral stenosis (with or without CAD), can be accompanied by dyspnea from increased left atrial and pulmonary venous pressure.

Key features of the physical examination that will help confirm or refute the presence of an arrhythmia as a cause for palpitations (as well as its adverse hemodynamic consequences) include measurement of the vital signs, assessment of the jugular venous pressure and pulse, and auscultation of the chest and precordium. A resting electrocardiogram can be used to document the arrhythmia. If exertion is known to induce the arrhythmia and accompanying palpitations, exercise electrocardiography can be used to make the diagnosis. If the arrhythmia is sufficiently infrequent, other methods must be used, including continuous electrocardiographic (Holter) monitoring; telephonic monitoring, through which the patient can transmit an electrocardiographic tracing during a sensed episode; loop recordings (external or implantable), which can capture the electrocardiographic event for later review; and mobile cardiac outpatient telemetry. Data suggest that Holter monitoring is of limited clinical utility, while the implantable loop recorder and mobile cardiac outpatient telemetry are safe and possibly more cost-effective in the assessment of patients with (infrequent) recurrent, unexplained palpitations.

Most patients with palpitations do not have serious arrhythmias or underlying structural heart disease. If sufficiently troubling to the patient, occasional benign atrial or ventricular premature contractions can often be managed with beta-blocker therapy. Palpitations incited by alcohol, tobacco, or illicit drugs need to be managed by abstention, while those caused by pharmacologic agents should be addressed by considering alternative therapies when appropriate or possible. Psychiatric causes of palpitations may benefit from cognitive therapy or pharmacotherapy. The physician should note that palpitations are at the very least bothersome and, on occasion, frightening to the patient. Once serious causes for the symptom have been excluded, the patient should be reassured that the palpitations will not adversely affect prognosis.

FURTHER READING

Crossland S, Berkin L: Problem based review: The patient with palpitations. Acute Med 11:169, 2012.

Jamshed N, Dubin J, Eldagah Z: Emergency management of palpitations in the elderly: Epidemiology, diagnostic approaches, and therapeutic options. Clin Geriatr Med 29:205, 2013. [PubMed: 23177608]

Sedaghat-Yazdi F, Koenig PR: The teenager with palpitations. Pediatr Clin North Am 61:63, 2014. [PubMed: 24267458]

Weber BE, Kapoor WN: Evaluation and outcomes of patients with palpitations. Am J Med 100:138, 1996. [PubMed: 8629647]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 40: Dysphagia

Ikuo Hirano; Peter J. Kahrilas

INTRODUCTION

Dysphagia—difficulty with swallowing—refers to problems with the transit of food or liquid from the mouth to the hypopharynx or through the esophagus. Severe dysphagia can compromise nutrition, cause aspiration, and reduce quality of life. Additional terminology pertaining to swallowing dysfunction is as follows. *Aphagia* (inability to swallow) typically denotes complete esophageal obstruction, most commonly encountered in the acute setting of a food bolus or foreign body impaction. *Odynophagia* refers to painful swallowing, typically resulting from mucosal ulceration within the oropharynx or esophagus. It commonly is accompanied by dysphagia, but the converse is not true. *Globus pharyngeus* is a foreign body sensation localized in the neck that does not interfere with swallowing and sometimes is relieved by swallowing. *Transfer dysphagia* frequently results in nasal regurgitation and pulmonary aspiration during swallowing and is characteristic of oropharyngeal dysphagia. *Phagophobia* (fear of swallowing) and *refusal to swallow* may be psychogenic or related to anticipatory anxiety about food bolus obstruction, odynophagia, or aspiration.

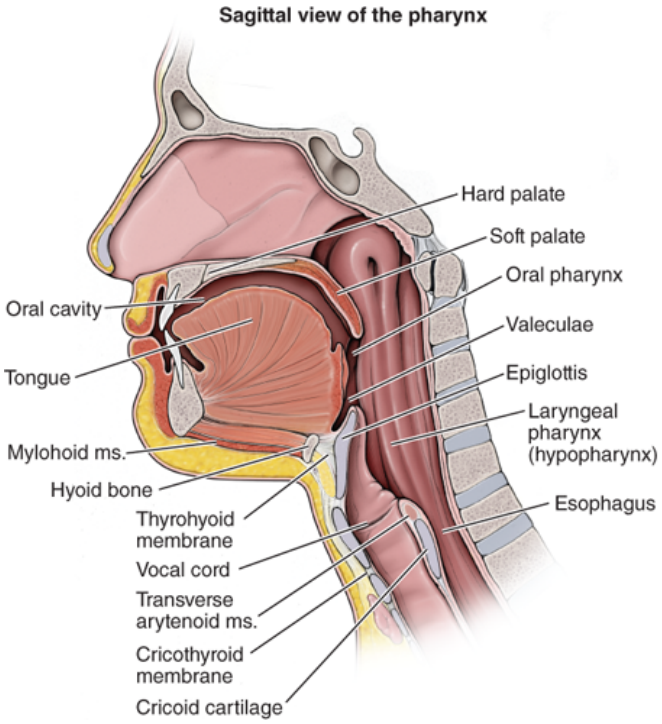
PHYSIOLOGY OF SWALLOWING

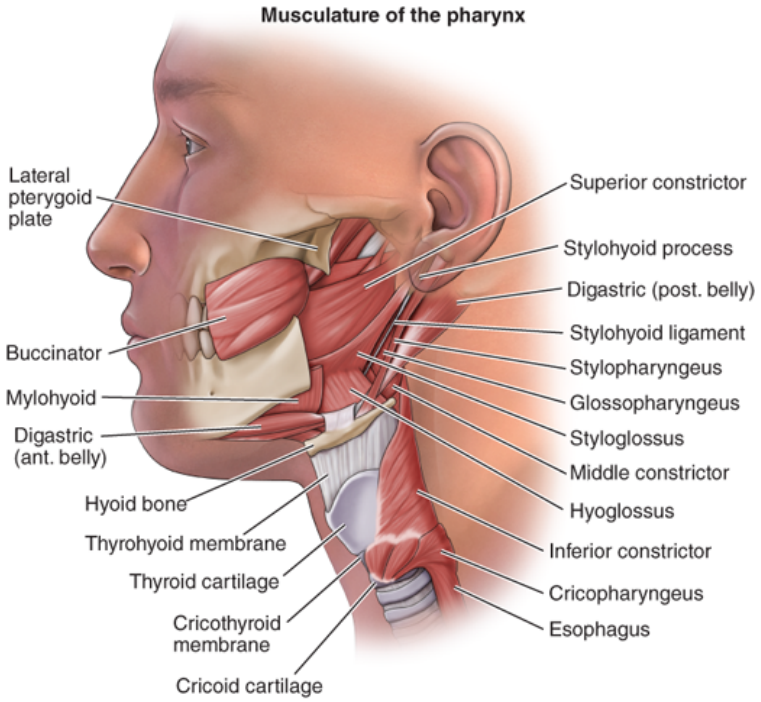
Swallowing begins with a voluntary (oral) phase that includes preparation during which food is masticated and mixed with saliva. This is followed by a transfer phase during which the bolus is pushed into the pharynx by the tongue. Bolus entry into the hypopharynx initiates the pharyngeal swallow response, which is centrally mediated and involves a complex series of actions, the net result of which is to propel food through the pharynx into the esophagus while preventing its entry into the airway. To accomplish this, the larynx is elevated and pulled forward, actions that also facilitate upper esophageal sphincter (UES) opening. Tongue pulsion then propels the bolus through the UES, followed by a peristaltic contraction that clears residue from the pharynx and through the esophagus. The lower esophageal sphincter (LES) relaxes as the food enters the esophagus and remains relaxed until the peristaltic contraction has delivered the bolus into the stomach. Peristaltic contractions elicited in response to a swallow are called *primary peristalsis* and involve sequenced inhibition followed by contraction of the musculature along the entire length of the esophagus. The inhibition that precedes the peristaltic contraction is called *deglutitive inhibition*. Local distention of the esophagus anywhere along its length, as may occur with gastroesophageal reflux, activates *secondary peristalsis* that begins at the point of distention and proceeds distally. Tertiary esophageal contractions are nonperistaltic, disordered esophageal contractions that may be observed to occur spontaneously during fluoroscopic observation.

The musculature of the oral cavity, pharynx, UES, and cervical esophagus is striated and directly innervated by lower motor neurons carried in cranial nerves (Fig. 40-1). Oral cavity muscles are innervated by the fifth (trigeminal) and seventh (facial) cranial nerves; the tongue, by the twelfth (hypoglossal) cranial nerve. Pharyngeal muscles are innervated by the ninth (glossopharyngeal) and tenth (vagus) cranial nerves.

FIGURE 40-1

Sagittal and diagrammatic views of the musculature involved in enacting oropharyngeal swallowing. Note the dominance of the tongue in the sagittal view and the intimate relationship between the entrance to the larynx (airway) and the esophagus. In the resting configuration illustrated, the esophageal inlet is closed. This is transiently reconfigured such that the esophageal inlet is open and the laryngeal inlet closed during swallowing. *(Adapted from PJ Kahrilas, in DW Gelfand and JE Richter [eds]: Dysphagia: Diagnosis and Treatment. New York: Igaku-Shoin Medical Publishers, 1989, pp. 11–28.)*





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Physiologically, the UES consists of the cricopharyngeus muscle, the adjacent inferior pharyngeal constrictor, and the proximal portion of the cervical esophagus. UES innervation is derived from the vagus nerve, whereas the innervation to the musculature acting on the UES to facilitate its opening during swallowing comes from the fifth, seventh, and twelfth cranial nerves. The UES remains closed at rest owing to both its inherent elastic properties and neurogenically mediated contraction of the cricopharyngeus muscle. UES opening during swallowing involves both cessation of vagal excitation to the cricopharyngeus and simultaneous contraction of the suprahyoid and geniohyoid muscles that pull open the UES in conjunction with the upward and forward displacement of the larynx.

The neuromuscular apparatus for peristalsis is distinct in proximal and distal parts of the esophagus. The cervical esophagus, like the pharyngeal musculature, consists of striated muscle and is directly innervated by lower motor neurons of the vagus nerve. Peristalsis in the proximal esophagus is governed by the sequential activation of the vagal motor neurons in the nucleus ambiguus. In contrast, the distal esophagus and LES are composed of smooth muscle and are controlled by excitatory and inhibitory neurons within the esophageal myenteric plexus. Medullary preganglionic neurons from the dorsal motor nucleus of the vagus trigger peristalsis via these ganglionic neurons during primary peristalsis. Neurotransmitters of the excitatory ganglionic neurons are acetylcholine and substance P; those of the inhibitory neurons are vasoactive intestinal peptide and nitric oxide. Peristalsis results from the patterned activation of inhibitory followed by excitatory ganglionic neurons, with progressive dominance of the inhibitory neurons distally. Similarly, LES relaxation occurs with the onset of deglutitive inhibition and persists until the peristaltic sequence is complete. At rest, the LES is contracted because of excitatory ganglionic stimulation and its intrinsic myogenic tone, a property that distinguishes it from the adjacent esophagus. The function of the LES is supplemented by the surrounding muscle of the right diaphragmatic crus, which acts as an external sphincter during inspiration, cough, or abdominal straining.

PATHOPHYSIOLOGY OF DYSPHAGIA

Dysphagia can be subclassified both by location and by the circumstances in which it occurs. With respect to location, distinct considerations apply to oral, pharyngeal, or esophageal dysphagia. Normal transport of an ingested bolus depends on the consistency and size of the bolus, the caliber of the lumen, the integrity of peristaltic contraction, and deglutitive inhibition of both the UES and the LES. Dysphagia caused by an oversized bolus or a narrow lumen is called *structural dysphagia*, whereas dysphagia due to abnormalities of peristalsis or impaired sphincter relaxation after swallowing is called *propulsive* or *motor dysphagia*. More than one mechanism may be operative in a patient with dysphagia. Scleroderma commonly presents with absent peristalsis as well as a weakened LES that predisposes patients to peptic stricture formation. Likewise, radiation therapy for head and neck cancer may compound the functional deficits in the oropharyngeal swallow attributable to the tumor and cause cervical esophageal stenosis. It is worth noting that in addition to bolus transit, symptom reporting of dysphagia is dependent upon intact sensory innervation and central nervous system perception.

Oral and Pharyngeal (Oropharyngeal) Dysphagia

Oral-phase dysphagia is associated with poor bolus formation and control so that food has prolonged retention within the oral cavity and may seep out of the mouth. Drooling and difficulty in initiating swallowing are other characteristic signs. Poor bolus control also may lead to premature spillage of food into the hypopharynx with resultant aspiration into the trachea or regurgitation into the nasal cavity. Pharyngeal-phase dysphagia is associated with retention of food in the pharynx due to poor tongue or pharyngeal propulsion or obstruction at the UES. Signs and symptoms of concomitant hoarseness or cranial nerve dysfunction may be associated with oropharyngeal dysphagia.

Oropharyngeal dysphagia may be due to neurologic, muscular, structural, iatrogenic, infectious, and metabolic causes. Iatrogenic, neurologic, and structural pathologies are most common. Iatrogenic causes include surgery and radiation, often in the setting of head and neck cancer. Neurogenic dysphagia resulting from cerebrovascular accidents, Parkinson's disease, and amyotrophic lateral sclerosis is a major source of morbidity related to aspiration and malnutrition. Medullary nuclei directly innervate the oropharynx. Lateralization of pharyngeal dysphagia implies either a structural pharyngeal lesion or a neurologic process that selectively targeted the ipsilateral brainstem nuclei or cranial nerve. Advances in functional brain imaging have elucidated an important role of the cerebral cortex in swallow function and dysphagia. Asymmetry in the cortical representation of the pharynx provides an explanation for the dysphagia that occurs as a consequence of unilateral cortical cerebrovascular accidents.

Oropharyngeal structural lesions causing dysphagia include Zenker's diverticulum, cricopharyngeal bar, and neoplasia. Zenker's diverticulum typically is encountered in elderly patients. In addition to dysphagia, patients may present with regurgitation of particulate food debris, aspiration, and halitosis. The pathogenesis is related to stenosis of the cricopharyngeus that causes diminished opening of the UES and results in increased hypopharyngeal pressure during swallowing with development of a pulsion diverticulum immediately above the cricopharyngeus in a region of potential weakness known as Killian's dehiscence. A cricopharyngeal bar, appearing as a prominent indentation behind the lower third of the cricoid cartilage, is related to Zenker's diverticulum in that it involves limited distensibility of the cricopharyngeus and can lead to the formation of a Zenker's diverticulum. However, a cricopharyngeal bar is a common radiographic finding, and most patients with transient cricopharyngeal bars are asymptomatic, making it important to rule out alternative etiologies of dysphagia before treatment. Furthermore, cricopharyngeal bars may be secondary to other neuromuscular disorders that impair opening of the UES.

Since the pharyngeal phase of swallowing occurs in less than a second, rapid-sequence fluoroscopy is necessary to evaluate for functional abnormalities. Adequate fluoroscopic examination requires that the patient be conscious and cooperative. The study incorporates recordings of swallow sequences during ingestion of food and liquids of varying consistencies. The pharynx is examined to detect bolus retention, regurgitation into the nose, or aspiration into the trachea. Timing and integrity of pharyngeal contraction and opening of the UES with a swallow are analyzed to assess both aspiration risk and the potential for swallow therapy. Structural abnormalities of the oropharynx, especially those which may require biopsies, also should be assessed by direct laryngoscopic examination.

Esophageal Dysphagia

The adult esophagus measures 18–26 cm in length and is anatomically divided into the cervical esophagus, extending from the pharyngoesophageal junction to the suprasternal notch, and the thoracic esophagus, which continues to the diaphragmatic hiatus. When distended, the esophageal lumen has internal dimensions of about 2 cm in the anteroposterior plane and 3 cm in the lateral plane. Solid food dysphagia becomes common when the lumen is narrowed to <13 mm, but also can occur with larger diameters in the setting of poorly masticated food or motor dysfunction. Circumferential lesions are more likely to cause dysphagia than are lesions that involve only a partial circumference of the esophageal wall. The most common structural causes of dysphagia are Schatzki's rings, eosinophilic esophagitis, and peptic strictures. Dysphagia also occurs in the setting of gastroesophageal reflux disease without a stricture, perhaps on the basis of altered esophageal sensation, reduced esophageal mural distensibility, or motor dysfunction.

Propulsive disorders leading to esophageal dysphagia result from abnormalities of peristalsis and/or deglutitive inhibition, potentially affecting the cervical or thoracic esophagus. Since striated muscle pathology usually involves both the oropharynx and the cervical esophagus, the clinical manifestations usually are dominated by oropharyngeal dysphagia. Diseases affecting smooth muscle involve both the thoracic esophagus and the LES. A dominant manifestation of this, absent peristalsis, refers to either the complete absence of swallow-induced contraction (absent contractility) or the presence of nonperistaltic, disordered contractions. Absent peristalsis and failure of deglutitive LES relaxation are the defining features of achalasia. In diffuse esophageal spasm (DES), LES function is normal, with the disordered motility restricted to the esophageal body. Absent contractility combined with severe weakness of the LES is a nonspecific pattern commonly found in patients with scleroderma.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Dysphagia

Figure 40-2 shows an algorithm for the approach to a patient with dysphagia.

HISTORY

The patient history is extremely valuable in making a presumptive diagnosis or at least substantially restricting the differential diagnoses in most patients. Key elements of the history are the localization of dysphagia, the circumstances in which dysphagia is experienced, other symptoms associated with dysphagia, and progression. Dysphagia that localizes to the suprasternal notch may indicate either an oropharyngeal or an esophageal etiology as distal dysphagia is referred proximally about 30% of the time. Dysphagia that localizes to the chest is esophageal in origin. Nasal regurgitation and tracheobronchial aspiration manifest by coughing with swallowing are hallmarks of oropharyngeal dysphagia. Severe cough with swallowing may also be a sign of a tracheoesophageal fistula. The presence of hoarseness may be another important diagnostic clue. When hoarseness precedes dysphagia, the primary lesion is usually laryngeal; hoarseness that occurs after the development of dysphagia may result from compromise of the recurrent laryngeal nerve by a malignancy. The type of food causing dysphagia is a crucial detail. Intermittent dysphagia that occurs only with solid food implies structural dysphagia, whereas constant dysphagia with both liquids and solids strongly suggests a motor abnormality. Two caveats to this pattern are that despite having a motor abnormality, patients with scleroderma generally develop mild dysphagia for solids only and, somewhat paradoxically, that patients with oropharyngeal dysphagia often have greater difficulty managing liquids than solids. Dysphagia that is progressive over the course of weeks to months raises concern for neoplasia. Episodic dysphagia to solids that is unchanged or slowly progressive over years indicates a benign disease process such as a Schatzki ring or eosinophilic esophagitis. Food impaction with a prolonged inability to pass an ingested bolus even with ingestion of liquid is typical of a structural dysphagia. Chest pain frequently accompanies dysphagia whether it is related to motor disorders, structural disorders, or reflux disease. A prolonged history of heartburn preceding the onset of dysphagia is suggestive of peptic stricture and, infrequently, esophageal adenocarcinoma. A history of prolonged nasogastric intubation, esophageal or head and neck surgery, ingestion of caustic agents or pills, previous radiation or chemotherapy, or associated mucocutaneous diseases may help isolate the cause of dysphagia. With accompanying odynophagia, which usually is indicative of ulceration, infectious or pill-induced esophagitis should be suspected. In patients with AIDS or other immunocompromised states, esophagitis due to opportunistic infections such as *Candida*, herpes simplex virus, or cytomegalovirus and to tumors such as Kaposi's sarcoma and lymphoma should be considered. A strong history of atopy increases concerns for eosinophilic esophagitis, especially in younger Caucasian male patients.

PHYSICAL EXAMINATION

Physical examination is important in the evaluation of oral and pharyngeal dysphagia because dysphagia is usually only one of many manifestations of a more global disease process. Signs of bulbar or pseudobulbar palsy, including dysarthria, dysphonia, ptosis, tongue atrophy, and hyperactive jaw jerk, in addition to evidence of generalized neuromuscular disease, should be elicited. The neck should be examined for thyromegaly. A careful inspection of the mouth and pharynx should disclose lesions that may interfere with passage of food. Missing dentition can interfere with mastication and exacerbate an existing cause of dysphagia. Physical examination is less helpful in the evaluation of esophageal dysphagia as most relevant pathology is restricted to the esophagus. The notable exception is skin disease. Changes in the skin may suggest a diagnosis of scleroderma or mucocutaneous diseases such as pemphigoid, lichen planus, and epidermolysis bullosa, all of which can involve the esophagus.

DIAGNOSTIC PROCEDURES

Although most instances of dysphagia are attributable to benign disease processes, dysphagia is also a cardinal symptom of several malignancies, making it an important symptom to evaluate. Cancer may result in dysphagia due to intraluminal obstruction (esophageal or proximal gastric cancer, metastatic deposits), extrinsic compression (lymphoma, lung cancer), or paraneoplastic syndromes. Even when not attributable to malignancy, dysphagia is usually a manifestation of an identifiable and treatable disease entity, making its evaluation beneficial to the patient and gratifying to the practitioner. The specific diagnostic algorithm to pursue is guided by the details of the history (Fig. 40-2). If oral or pharyngeal dysphagia is suspected, a fluoroscopic swallow study, usually done by a swallow therapist, is the procedure of choice. Otolaryngoscopic and neurologic evaluation also can be important, depending on the circumstances. For suspected esophageal dysphagia, upper endoscopy is the single most useful test. Endoscopy allows better visualization of mucosal lesions than does barium radiography and also allows one to obtain mucosal biopsies. Endoscopic or histologic abnormalities are evident in the leading causes of esophageal dysphagia: Schatzki's ring, gastroesophageal reflux disease, and eosinophilic esophagitis. Furthermore, therapeutic intervention with esophageal dilation can be done as part of the procedure if it is deemed necessary. The emergence of eosinophilic esophagitis as a leading cause of dysphagia in both children and adults has led to the recommendation that esophageal mucosal biopsies be obtained routinely in the evaluation of unexplained dysphagia even if characteristic, endoscopically identified esophageal mucosal features are absent. For cases of suspected esophageal motility disorders, endoscopy is still the appropriate initial evaluation as neoplastic and inflammatory conditions can secondarily produce patterns of either achalasia or esophageal spasm. Esophageal manometry is done if dysphagia is not adequately explained by endoscopy or to confirm the diagnosis of a suspected esophageal motor disorder. Barium radiography can provide useful adjunctive information in cases of subtle or complex esophageal strictures, prior esophageal surgery, esophageal diverticula, or paraesophageal herniation. In specific cases, computed tomography (CT) examination and endoscopic ultrasonography may be useful.

TREATMENT

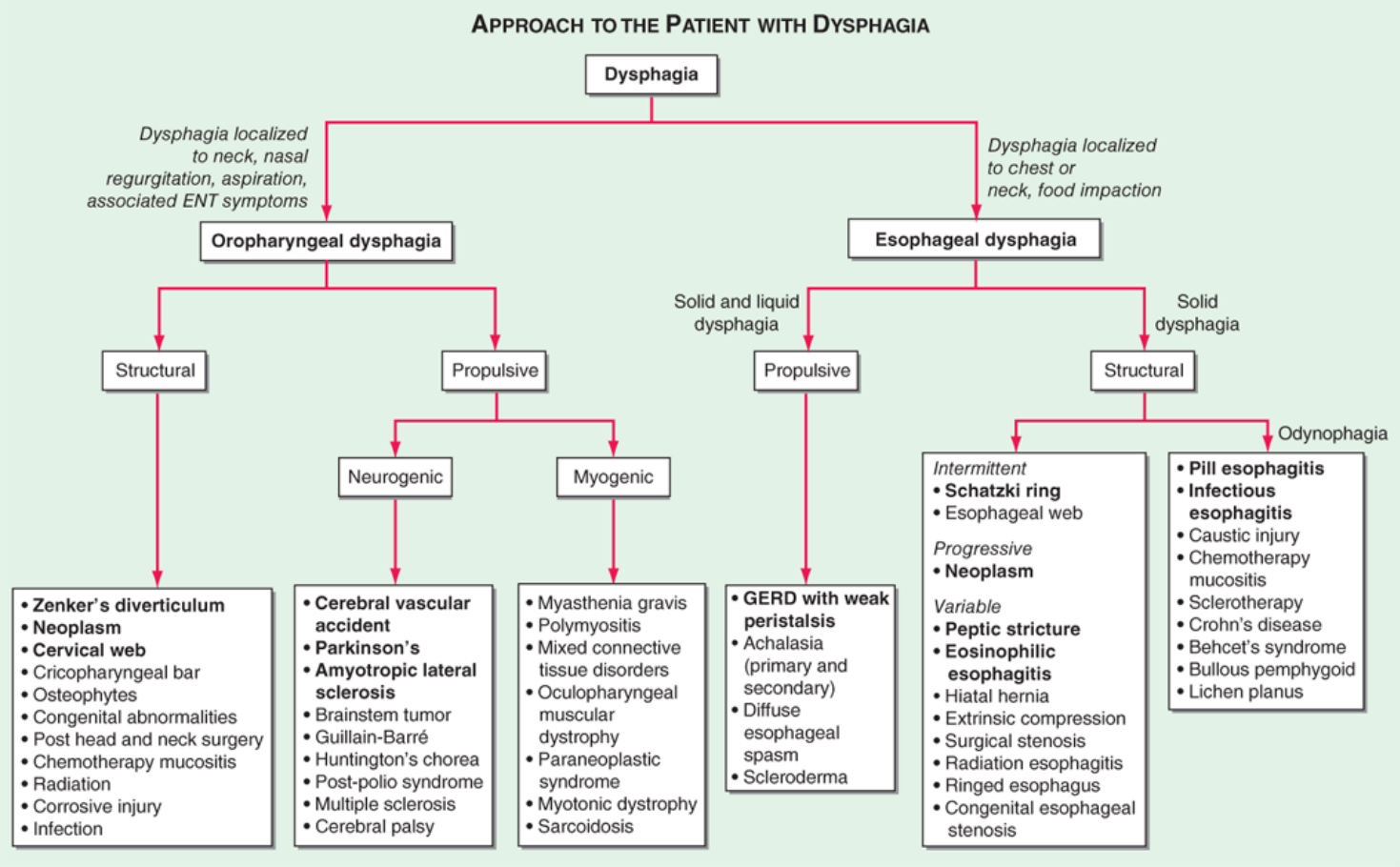
Treatment of dysphagia depends on both the locus and the specific etiology. Oropharyngeal dysphagia most commonly results from functional deficits caused by neurologic disorders. In such circumstances, the treatment focuses on utilizing postures or maneuvers devised to reduce pharyngeal residue and enhance airway protection learned under the direction of a trained swallow therapist. Aspiration risk may be reduced by altering the consistency of ingested food and liquid. Dysphagia resulting from a cerebrovascular accident usually, but not always, spontaneously improves within the first few weeks after the event. More severe and persistent cases may require gastrostomy and enteral feeding. Patients with myasthenia gravis **(Chap. 440)** and polymyositis **(Chap. 358)** may respond to medical treatment of the primary neuromuscular disease. Surgical intervention with cricopharyngeal

myotomy is usually not helpful, with the exception of specific disorders such as the idiopathic cricopharyngeal bar, Zenker's diverticulum, and oculopharyngeal muscular dystrophy. Chronic neurologic disorders such as Parkinson's disease and amyotrophic lateral sclerosis may manifest with severe oropharyngeal dysphagia. Feeding by a nasogastric tube or an endoscopically placed gastrostomy tube may be considered for nutritional support; however, these maneuvers do not provide protection against aspiration of salivary secretions or refluxed gastric contents.

Treatment of esophageal dysphagia is covered in detail in **Chap. 316**. The majority of causes of esophageal dysphagia are effectively managed by means of esophageal dilatation using bougie or balloon dilators. Cancer and achalasia are often managed surgically, although endoscopic techniques are available for both palliation and primary therapy, respectively. Infectious etiologies respond to antimicrobial medications or treatment of the underlying immunosuppressive state. Finally, eosinophilic esophagitis has emerged as an important cause of dysphagia that is amenable to treatment by elimination of dietary allergens or administration of swallowed, topically acting glucocorticoids.

FIGURE 40-2

Approach to the patient with dysphagia. Etiologies in bold print are the most common. ENT, ear, nose, and throat; GERD, gastroesophageal reflux disease.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Cook IJ: Oropharyngeal dysphagia. Gastroenterol Clin North Am 38:411, 2009. [PubMed: 19699405]

Hirano I: Esophagus: Anatomy and structural anomalies, in *Yamada Atlas of Gastroenterology*, 6th ed. Wiley-Blackwell Publishing Co. 2016, pp 42–59.

Kahrilas PJ et al: The Chicago Classification of esophageal motility disorders, v3.0. Neurogastroenterol Motil 27:160, 2015. [PubMed: 25469569]

Pandolfino JP, Kahrilas PJ: Esophageal neuromuscular function and motility disorders, in *Sleisenger and Fordtran's Gastrointestinal and Liver Disease*, 10th ed, Feldman M, Friedman LS, Brandt LJ (eds). Philadelphia, Elsevier, 2016, pp 701–732.

Shaker R et al (eds): *Principles of Deglutition: A Multidisciplinary Text for Swallowing and Its Disorders*. New York, Springer, 2013.

Chapter 41: Nausea, Vomiting, and Indigestion

William L. Hasler

INTRODUCTION

Nausea is the subjective feeling of a need to vomit. *Vomiting* (emesis) is the oral expulsion of gastrointestinal contents due to gut and thoracoabdominal wall contractions. Vomiting is contrasted with *regurgitation*, the effortless passage of gastric contents into the mouth. *Rumination* is the repeated regurgitation of food residue, which may be rechewed and reswallowed. In contrast to emesis, these phenomena exhibit volitional control. *Indigestion* is a term encompassing a range of complaints including nausea, vomiting, heartburn, regurgitation, and dyspepsia (symptoms thought to originate in the gastroduodenal region). Some individuals with dyspepsia experience postprandial fullness, early satiety (an inability to complete a meal due to premature fullness), bloating, eructation (belching), and anorexia. Others report predominantly epigastric burning or pain.

NAUSEA AND VOMITING

MECHANISMS

Vomiting is coordinated by the brainstem and is effected by responses in the gut, pharynx, and somatic musculature. Mechanisms underlying nausea are poorly understood but likely involve the cerebral cortex, as nausea requires conscious perception. This is supported by functional brain imaging studies showing activation of cerebral cortical regions during nausea.

Coordination of Emesis

Brainstem nuclei—including the nucleus tractus solitarius; dorsal vagal and phrenic nuclei; medullary nuclei regulating respiration; and nuclei that control pharyngeal, facial, and tongue movements—coordinate initiation of emesis involving neurokinin NK_1, serotonin 5-HT_3, and vasopressin pathways.

Somatic and visceral muscles respond stereotypically during emesis. Inspiratory thoracic and abdominal wall muscles contract, producing high intrathoracic and intraabdominal pressures that evacuate the stomach. The gastric cardia herniates above the diaphragm, and the larynx moves upward to propel the vomitus. Distally migrating gut contractions are normally regulated by an electrical phenomenon, the slow wave, which cycles at 3 cycles/min in the stomach and 11 cycles/min in the duodenum. During emesis, the

slow wave is abolished and replaced by orally propagating spikes that evoke retrograde contractions that assist in expulsion of gut contents.

Activators of Emesis

Emetic stimuli act at several sites. Emesis evoked by unpleasant thoughts or smells originates in the brain, whereas cranial nerves mediate vomiting after gag reflex activation. Motion sickness and inner ear disorders act on labyrinthine pathways. Gastric irritants and cytotoxic agents like cisplatin stimulate gastroduodenal vagal afferent nerves. Nongastric afferents are activated by bowel obstruction and mesenteric ischemia. The area postrema, in the medulla, responds to bloodborne stimuli (emetogenic drugs, bacterial toxins, uremia, hypoxia, ketoacidosis) and is termed the *chemoreceptor trigger zone*.

Neurotransmitters mediating vomiting are selective for different sites. Labyrinthine disorders stimulate vestibular muscarinic M_1 and histaminergic H_1 receptors. Vagal afferent stimuli activate 5-HT_3 receptors. The area postrema is served by nerves acting on 5-HT_3, M_1, H_1, and dopamine D_2 subtypes. Central NK_1 receptors mediate both nausea and vomiting. Cannabinoid CB_1 pathways may participate in the cerebral cortex and brainstem. Optimal pharmacologic therapy of vomiting requires understanding these pathways.

DIFFERENTIAL DIAGNOSIS

Nausea and vomiting are caused by conditions within and outside the gut, by drugs, and by circulating toxins **(Table 41-1)**. Unexplained causes of chronic nausea and vomiting are relatively rare, being reported by 2–3% of the population.

TABLE 41-1

Causes of Nausea and Vomiting

Intraperitoneal	Extraperitoneal	Medications/Metabolic Disorders
Obstructing disorders Pyloric obstruction Small-bowel obstruction Colonic obstruction Superior mesenteric artery syndrome Enteric infections Viral Bacterial Inflammatory diseases Cholecystitis Pancreatitis Appendicitis Hepatitis Altered sensorimotor function Gastroparesis Intestinal pseudoobstruction Gastroesophageal reflux Chronic nausea vomiting syndrome Cyclic vomiting syndrome Cannabinoid hyperemesis syndrome Rumination syndrome Biliary colic Abdominal irradiation	Cardiopulmonary disease Cardiomyopathy Myocardial infarction Labyrinthine disease Motion sickness Labyrinthitis Malignancy Intracerebral disorders Malignancy Hemorrhage Abscess Hydrocephalus Psychiatric illness Anorexia and bulimia nervosa Depression Postoperative vomiting	Drugs Cancer chemotherapy Antibiotics Cardiac antiarrhythmics Digoxin Oral hypoglycemics Oral contraceptives Antidepressants Restless legs/Parkinson's therapies Smoking cessation agents Endocrine/metabolic disease Pregnancy Uremia Ketoacidosis Thyroid and parathyroid disease Adrenal insufficiency Toxins Liver failure Ethanol

Intraperitoneal Disorders

Visceral obstruction and inflammation of hollow and solid viscera may elicit vomiting. Gastric obstruction results from ulcers and malignancy. Small-bowel and colon blockage occur because of adhesions, benign or malignant tumors, volvulus, intussusception, or inflammatory diseases like Crohn's disease. The superior mesenteric artery syndrome, occurring after weight loss or prolonged bed rest, results when the duodenum

is compressed by the overlying superior mesenteric artery. Abdominal irradiation impairs intestinal motor function and induces strictures. Biliary colic causes nausea by acting on local afferent nerves. Vomiting with pancreatitis, cholecystitis, and appendicitis result from visceral irritation and induction of ileus. Enteric infections with viruses like norovirus or rotavirus or bacteria like *Staphylococcus aureus* and *Bacillus cereus* cause vomiting, especially in children. Opportunistic infections like cytomegalovirus or herpes simplex virus induce emesis in immunocompromised individuals.

Gut sensorimotor dysfunction often causes nausea and vomiting. *Gastroparesis* presents with symptoms of gastric retention with evidence of delayed gastric emptying and occurs after vagotomy or with pancreatic carcinoma, mesenteric vascular insufficiency, or organic diseases like diabetes, scleroderma, and amyloidosis. Idiopathic gastroparesis is the most common etiology. It occurs in the absence of systemic illness and follows a viral illness in ~15–20% of cases, suggesting an infectious trigger. *Intestinal pseudoobstruction* is characterized by disrupted intestinal and colonic motor activity with retention of food residue and secretions; bacterial overgrowth; nutrient malabsorption; and symptoms of nausea, vomiting, bloating, pain, and altered defecation. Intestinal pseudoobstruction may be idiopathic, inherited as a familial visceral myopathy or neuropathy, result from systemic disease like scleroderma or an infiltrative process like amyloidosis, or occur as a paraneoplastic consequence of malignancy (e.g., small-cell lung carcinoma). Patients with gastroesophageal reflux report nausea and vomiting, as do some with irritable bowel syndrome (IBS) or chronic constipation.

Other functional gastroduodenal disorders without organic abnormalities have been characterized. *Chronic nausea vomiting syndrome* is defined as bothersome nausea at least one day and/or one or more vomiting episodes weekly in the absence of an eating disorder or psychiatric disease. *Cyclic vomiting syndrome* causes 3–14% of cases of unexplained nausea and vomiting and presents with periodic discrete episodes of relentless vomiting in children and adults and shows an association with migraine headaches, suggesting that some cases may be migraine variants. Some adult cases have been associated with rapid gastric emptying. A related condition, *cannabinoid hyperemesis syndrome*, presents with cyclical vomiting with intervening well periods in individuals (mostly men) who use large quantities of cannabis over many years and resolves with its discontinuation. Pathologic behaviors such as taking prolonged hot baths or showers are associated with the syndrome. *Rumination syndrome*, characterized by repetitive regurgitation of recently ingested food, is often misdiagnosed as refractory vomiting.

Extraperitoneal Disorders

Myocardial infarction and congestive heart failure may cause nausea and vomiting. Postoperative emesis occurs after 25% of surgeries, most commonly abdominal and orthopedic surgery. Increased intracranial pressure from tumors, bleeding, abscess, or blockage of cerebrospinal fluid outflow produces vomiting with or without nausea. Patients with psychiatric illnesses including anorexia nervosa, bulimia nervosa, anxiety, and depression often report significant nausea that may be associated with delayed gastric emptying.

Medications and Metabolic Disorders

Drugs evoke vomiting by action on the stomach (analgesics, erythromycin) or area postrema (opiates, anti-parkinsonian drugs). Other emetogenic agents include antibiotics, cardiac antiarrhythmics, antihypertensives, oral hypoglycemics, antidepressants (selective serotonin and serotonin norepinephrine reuptake inhibitors), smoking cessation drugs (varenicline, nicotine), and contraceptives. Cancer chemotherapy causes vomiting that is acute (within hours of administration), delayed (after 1 or more days), or anticipatory. Acute emesis from highly emetogenic agents (e.g., cisplatin) is mediated by 5-HT_3 pathways. Delayed emesis is less dependent on 5-HT_3 pathways with greater mediation by NK_1 mechanisms. Anticipatory nausea may respond to anxiolytic therapy rather than antiemetics.

Metabolic disorders elicit nausea and vomiting. Pregnancy is the most prevalent endocrinologic cause, and nausea affects 70% of women in the first trimester. Hyperemesis gravidarum is a severe form of nausea of pregnancy that produces significant dehydration and electrolyte disturbances. Uremia, ketoacidosis, adrenal insufficiency, and parathyroid and thyroid disease are other metabolic etiologies.

Circulating toxins evoke emesis via effects on the area postrema. Endogenous toxins are generated in fulminant liver failure, whereas exogenous enterotoxins may be produced by enteric bacterial infection. Ethanol intoxication is a common toxic etiology of nausea and vomiting.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Nausea and Vomiting

HISTORY AND PHYSICAL EXAMINATION

The history helps define the etiology of nausea and vomiting. Drugs, toxins, and infections often cause acute symptoms, whereas established illnesses evoke chronic complaints. Gastroparesis and pyloric obstruction elicit vomiting within an hour of eating. Emesis from intestinal blockage occurs later. Vomiting occurring minutes after meal consumption prompts consideration of rumination syndrome. With severe gastric emptying delays, the vomitus may contain food residue ingested days before. Hematemesis raises suspicion of an ulcer, malignancy, or Mallory-Weiss tear. Feculent emesis is noted with distal intestinal or colonic obstruction. Bilious vomiting excludes gastric obstruction, whereas emesis of undigested food is consistent with a Zenker's diverticulum or achalasia. Vomiting can relieve abdominal pain from a bowel obstruction, but has no effect in pancreatitis or cholecystitis. Profound weight loss raises concern about malignancy or obstruction. Fevers suggest inflammation. An intracranial source is considered if there are headaches or visual field changes. Vertigo or tinnitus indicates labyrinthine disease.

The physical examination complements the history. Orthostatic hypotension and reduced skin turgor indicate intravascular fluid loss. Pulmonary abnormalities raise concern for aspiration of vomitus. Bowel sounds may be absent with ileus. High-pitched rushes suggest bowel obstruction, whereas a succussion splash upon abrupt lateral movement of the patient is found with gastroparesis or pyloric obstruction. Tenderness or involuntary guarding raises suspicion of inflammation. Fecal blood suggests mucosal injury

from ulcer, ischemia, or tumor. Neurologic disease presents with papilledema, visual field loss, or focal neural abnormalities. Neoplasm is suggested by palpable masses or adenopathy.

DIAGNOSTIC TESTING

For intractable symptoms or an elusive diagnosis, selected screening tests can direct clinical care. Electrolyte replacement is indicated for hypokalemia or metabolic alkalosis. Iron-deficiency anemia mandates a search for mucosal injury. Pancreaticobiliary disease is indicated by abnormal pancreatic or liver biochemistries. Endocrinologic, rheumatologic, or paraneoplastic etiologies are suggested by hormone or serologic abnormalities. If obstruction is suspected, supine and upright abdominal radiographs may show intestinal air-fluid levels with reduced colonic air. Ileus is characterized by diffusely dilated air-filled bowel loops.

Anatomic studies may be indicated if initial testing is nondiagnostic. Upper endoscopy detects ulcers, malignancy, and retained food residue in gastroparesis. Small-bowel barium radiography or computed tomography (CT) diagnoses partial bowel obstruction. Colonoscopy or contrast enema radiography detects colonic obstruction. Ultrasound or CT defines intraperitoneal inflammation; CT and magnetic resonance imaging (MRI) enterography provide define inflammation in Crohn's disease. Brain CT or MRI can delineate intracranial disease. Mesenteric angiography, CT, or MRI is useful for suspected ischemia.

Gastrointestinal motility testing may detect an underlying motor disorder. Gastroparesis commonly is diagnosed by gastric scintigraphy, by which measures emptying of a radiolabeled meal. A non-radioactive ^{13}C-labelled gastric emptying breath test was FDA-approved in 2015 and may be a cost-effective alternative to scintigtraphy. Intestinal pseudoobstruction is suggested by abnormal barium transit and luminal dilation on small-bowel contrast radiography. Wireless motility capsule methods measure transit in the stomach, small bowel, and colon by detecting pH changes between regions and also can diagnose gastroparesis and small bowel dysmotility. Small-intestinal manometry can confirm the diagnosis of pseudoobstruction and characterize the motor abnormality as neuropathic or myopathic based on contractile patterns. Manometry can obviate the need for surgical intestinal biopsy to detect smooth muscle or neuronal degeneration. Combined ambulatory esophageal pH/impedance testing and high-resolution manometry facilitates diagnosis of rumination syndrome.

TREATMENT

TREATMENT

Nausea and Vomiting

GENERAL PRINCIPLES

Therapy of vomiting is tailored to correcting remediable abnormalities if possible. Hospitalization is considered for severe dehydration, especially if oral fluid replenishment cannot be sustained. Once oral intake is tolerated, nutrients are restarted with low-fat liquids, because lipids delay gastric emptying. A low residue, small particle diet has shown efficacy in gastroparesis in a controlled study. Controlling blood

glucose in poorly controlled diabetics can reduce hospitalizations in gastroparesis and may improve nausea and vomiting.

ANTIEMETIC MEDICATIONS

The most commonly used antiemetic agents act on central nervous system sites **(Table 41-2)**. Antihistamines like dimenhydrinate and meclizine and anticholinergics like scopolamine act on labyrinthine pathways to treat motion sickness and labyrinthine disorders. D_2 antagonists treat emesis evoked by area postrema stimuli and are used for medication, toxic, and metabolic etiologies. Dopamine antagonists cross the blood-brain barrier and cause anxiety, movement disorders, and hyperprolactinemic effects (galactorrhea, sexual dysfunction).

Other classes exhibit antiemetic properties. 5-HT_3 antagonists like ondansetron and granisetron prevent postoperative vomiting, radiation therapy–induced symptoms, and cancer chemotherapy–induced emesis, but also are used for other causes of emesis. NK_1 antagonists like aprepitant are approved for chemotherapy-induced vomiting and also reduce gastroparesis symptoms. Tricyclic antidepressants reduce symptoms in some patients with functional causes of vomiting, but did not show benefits in a controlled trial in gastroparesis. Other antidepressants such as mirtazapine and olanzapine and the pain-modulating agent gabapentin also may exhibit antiemetic effects.

GASTROINTESTINAL MOTOR STIMULANTS

Drugs that stimulate gastric emptying are used for gastroparesis (Table 41-2). Metoclopramide, a combined 5-HT_4 agonist and D_2 antagonist, is effective in gastroparesis, but antidopaminergic side effects, including dystonias and mood disturbances, limit use in ~25% of cases. Erythromycin increases gastroduodenal motility by action on receptors for motilin, an endogenous fasting motor stimulant. Intravenous erythromycin is useful for inpatients with refractory gastroparesis. Utility of oral forms is limited by development of tolerance. Domperidone, a D_2 antagonist not available in the United States, exhibits prokinetic and antiemetic effects but does not cross into most brain regions; thus, dystonic reactions are rare. Domperidone can induce hyperprolactinemic side effects via effects on pituitary regions served by a porous blood-brain barrier. Prucalopride, a 5-HT_4 agonist available in Canada and Europe, has shown efficacy in a preliminary gastroparesis trial.

Refractory motility disorders pose challenges. Intestinal pseudoobstruction may respond to the somatostatin analogue octreotide, which induces propagative small-intestinal motor complexes. Acetylcholinesterase inhibitors like pyridostigmine may benefit some patients with small-bowel dysmotility. Pyloric botulinum toxin injections are reported in uncontrolled studies to reduce gastroparesis symptoms, but small controlled trials observe benefits no greater than sham treatments. Surgical pyloroplasty and peroral endoscopic myotomy (POEM) of the pylorus has improved symptoms in case series. Placing a feeding jejunostomy reduces hospitalizations and improves overall health in some patients with refractory gastroparesis. Postvagotomy gastroparesis may improve with near-total gastric resection; similar operations are being tried for other gastroparesis etiologies. Implanted gastric electrical stimulators may reduce

symptoms, enhance nutrition, improve quality of life, and decrease health care expenditures in medication-refractory gastroparesis, but small controlled trials do not report convincing benefits.

SAFETY CONSIDERATIONS

Safety concerns have been raised about selected antiemetics. Centrally acting antidopaminergics, especially metoclopramide, can cause irreversible movement disorders like tardive dyskinesia, particularly in older patients. This complication should be explained and documented in the medical record. Domperidone, erythromycin, tricyclic antidepressants, and 5-HT_3 antagonists can induce dangerous cardiac arrhythmias, especially in those with QTc interval prolongation on electrocardiography (ECG). Surveillance ECG testing has been advocated for some of these agents.

SELECTED CLINICAL SETTINGS

Some cancer chemotherapies are intensely emetogenic **(Chap. 69)**. Combining a 5-HT_3 antagonist, an NK_1 antagonist, and a glucocorticoid can control both acute and delayed vomiting after highly emetogenic chemotherapy. Unlike other drugs in the same class, the 5-HT_3 antagonist palonosetron can prevent delayed chemotherapy-induced vomiting. Benzodiazepines like lorazepam reduce anticipatory nausea and vomiting. Miscellaneous therapies with benefit in chemotherapy-induced emesis include cannabinoids, olanzapine, and alternative therapies like ginger. Most antiemetic regimens produce greater reductions in vomiting than nausea.

Clinicians should exercise caution in managing pregnant patients with nausea. Studies of the teratogenic effects of antiemetic agents provide conflicting results. Few controlled trials have been performed in nausea of pregnancy. Antihistamines like meclizine and doxylamine, antidopaminergics like prochlorperazine, and antiserotonergics like ondansetron demonstrate limited efficacy. Some obstetricians offer alternative therapies including pyridoxine, acupressure, or ginger.

Managing cyclic vomiting syndrome is challenging. Prophylaxis with tricyclic antidepressants, cyproheptadine, or β-adrenoceptor antagonists can reduce the severity and frequency of attacks. Intravenous 5-HT_3 antagonists combined with the sedating effects of a benzodiazepine like lorazepam are a mainstay for treating acute flares. Small studies report benefits with antimigraine agents, including the 5-HT_1 agonist sumatriptan, and selected anticonvulsants like topiramate, zonisamide, and levetiracetam.

TABLE 41-2

Treatment of Nausea and Vomiting

Treatment	Mechanism	Examples	Clinical Indications
Antiemetic agents	Antihistaminergic	Dimenhydrinate, meclizine	Motion sickness, inner ear disease
	Anticholinergic	Scopolamine	Motion sickness, inner ear disease
	Antidopaminergic	Prochlorperazine, thiethylperazine	Medication-, toxin-, or metabolic-induced emesis
	5-HT_3 antagonist	Ondansetron, granisetron	Chemotherapy- and radiation-induced emesis, postoperative emesis
	NK_1 antagonist	Aprepitant	Chemotherapy-induced nausea and vomiting
	Tricyclic antidepressant	Amitriptyline, nortriptyline	Chronic nausea vomiting syndrome, cyclic vomiting syndrome, ?gastroparesis
	Other antidepressant	Mirtazapine, olanzapine	?Chronic nausea vomiting syndrome, ? gastroparesis
Prokinetic agents	5-HT_4 agonist and antidopaminergic	Metoclopramide	Gastroparesis
	Motilin agonist	Erythromycin	Gastroparesis, ?intestinal pseudoobstruction
	Peripheral antidopaminergic	Domperidone	Gastroparesis
	Somatostatin analogue	Octreotide	Intestinal pseudoobstruction
	Acetylcholinesterase inhibitor	Pyridostigmine	?Small-intestinal dysmotility/pseudoobstruction

Treatment	Mechanism	Examples	Clinical Indications
Special settings	Benzodiazepines	Lorazepam	Anticipatory nausea and vomiting with chemotherapy
	Glucocorticoids	Methylprednisolone, dexamethasone	Chemotherapy-induced emesis
	Cannabinoids	Tetrahydrocannabinol	Chemotherapy-induced emesis

Note: ?, indication is uncertain.

INDIGESTION

MECHANISMS

The most common causes of indigestion are gastroesophageal reflux and functional dyspepsia. Other cases are a consequence of organic illness.

Gastroesophageal Reflux

Gastroesophageal reflux results from many physiologic defects. Reduced lower esophageal sphincter (LES) tone contributes to reflux in scleroderma and pregnancy and may be a factor in some patients without systemic illness. Others exhibit frequent transient LES relaxations (TLESRs) that permit bathing of the esophagus by acid or nonacidic fluid. Reductions in esophageal body motility or salivary secretion prolong fluid exposure. Increased intragastric pressure promotes gastroesophageal reflux in obese patients. The role of hiatal hernias is controversial—most reflux patients have hiatal hernias, but most with hiatal hernias do not report excess heartburn.

Gastric Motor Dysfunction

Disturbed gastric motility may contribute to gastroesophageal reflux in up to one-third of cases. Delayed gastric emptying is also found in ~30% of functional dyspeptics, while rapid gastric emptying affects 5%. The relation of these defects to symptom induction is uncertain; studies show poor correlation between symptom severity and degrees of motor dysfunction. Impaired gastric fundus relaxation after eating (i.e., accommodation) may underlie selected dyspeptic symptoms like bloating, nausea, and early satiety in ~40% of patients and may predispose to TLESRs and acid reflux.

Visceral Afferent Hypersensitivity

Disturbed gastric sensation is another pathogenic factor in functional dyspepsia. Approximately 35% of dyspeptic patients note discomfort with fundic distention to lower pressures than in healthy controls. Others with dyspepsia exhibit hypersensitivity to chemical stimulation with capsaicin or with acid or lipid perfusion of the duodenum. Some individuals with functional heartburn without increased acid or nonacid reflux may have heightened perception of normal esophageal acidity.

Other Factors

Helicobacter pylori has a clear etiologic role in peptic ulcer disease, but ulcers cause a minority of dyspepsia cases. *H. pylori* is a minor factor in the genesis of functional dyspepsia. Anxiety and depression may play contributing roles in some functional dyspepsia cases. Functional MRI studies show increased activation of several brain regions, emphasizing contributions from central nervous system pathways. Inflammatory factors like duodenal eosinophilia (and possibly increased duodenal mast cells) may contribute to early satiety and pain in functional dyspepsia. Up to 20% of functional dyspepsia patients report symptom onset after a viral illness, suggestive of an infectious cause. Analgesics cause dyspepsia, whereas nitrates, calcium channel blockers, theophylline, and progesterone promote gastroesophageal reflux. Other stimuli that induce reflux include ethanol, tobacco, and caffeine via LES relaxation. Genetic factors predispose to development of reflux and dyspepsia.

DIFFERENTIAL DIAGNOSIS

Gastroesophageal Reflux Disease

Gastroesophageal reflux disease (GERD) is prevalent. Heartburn or regurgitation are reported weekly by 18–28%. Most cases of heartburn result from excess acid reflux, but reflux of nonacidic fluid produces similar symptoms. Alkaline reflux esophagitis produces GERD-like symptoms most often in patients who have had surgery for peptic ulcer disease. Ten percent of patients with heartburn exhibit no increase in acid or nonacidic esophageal reflux (functional heartburn).

Functional Dyspepsia

Nearly 25% of the populace has dyspepsia at least six times yearly, but only 10–20% present to clinicians. Functional dyspepsia, the cause of symptoms in >70% of dyspeptic patients, is defined as bothersome postprandial fullness, early satiety, or epigastric pain or burning with symptom onset at least 6 months before diagnosis in the absence of organic cause. Functional dyspepsia is subdivided into postprandial distress syndrome, characterized by meal-induced fullness and early satiety, and epigastric pain syndrome, which presents with epigastric pain or burning which may or may not be meal-related. Most cases follow a benign course, but some with *H. pylori* infection or on nonsteroidal anti-inflammatory drugs (NSAIDs) develop ulcers.

Ulcer Disease

In most GERD patients, there is no injury to the esophagus. However, 5% develop esophageal ulcers, and some form strictures. Symptoms cannot distinguish nonerosive from erosive or ulcerative esophagitis. A minority of cases of dyspepsia stem from gastric or duodenal ulcers. The most common causes of ulcers are *H. pylori* infection and NSAID use. Other rare causes of gastroduodenal ulcers include Crohn's disease **(Chap. 319)** and Zollinger-Ellison syndrome **(Chap. 317)**, resulting from gastrin overproduction by an endocrine tumor.

Malignancy

Dyspeptic patients often seek care because of fear of cancer, but few cases result from malignancy. Esophageal squamous cell carcinoma occurs most often with long-standing tobacco or ethanol intake. Other risks include prior caustic ingestion, achalasia, and the hereditary disorder tylosis. Esophageal adenocarcinoma usually complicates prolonged acid reflux. Eight to 20% of GERD patients exhibit esophageal intestinal metaplasia, termed *Barrett's metaplasia*, which predisposes to esophageal adenocarcinoma **(Chap. 76)**. Gastric malignancies include adenocarcinoma, which is prevalent in certain Asian societies, and lymphoma.

Other Causes

Opportunistic fungal or viral esophageal infections may produce heartburn but more often cause odynophagia. Other causes of esophageal inflammation include eosinophilic esophagitis and pill esophagitis. Biliary colic is in the differential diagnosis of unexplained upper abdominal pain, but most patients with biliary colic report discrete acute episodes of right upper quadrant or epigastric pain rather than the chronic burning or fullness of dyspepsia. Twenty percent of gastroparesis patients report a predominance of pain rather than nausea and vomiting. Intestinal lactase deficiency as a cause of gas, bloating, and discomfort occurs in 15–25% of whites of northern European descent but is more common in blacks and Asians. Intolerance of other carbohydrates (e.g., fructose, sorbitol) produces similar symptoms. Small-intestinal bacterial overgrowth may cause dyspepsia, often associated with bowel dysfunction, distention, and malabsorption. Celiac disease, pancreatic disease (chronic pancreatitis, malignancy), hepatocellular carcinoma, Ménétrier's disease, infiltrative diseases (sarcoidosis, eosinophilic gastroenteritis), mesenteric ischemia, thyroid and parathyroid disease, and abdominal wall strain cause dyspepsia. Gluten sensitivity in the absence of celiac disease can elicit unexplained upper abdominal symptoms. Extraperitoneal etiologies of indigestion include congestive heart failure and tuberculosis.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Indigestion

HISTORY AND PHYSICAL EXAMINATION

Management of indigestion requires a thorough interview. GERD classically produces heartburn, a substernal warmth that moves toward the neck. Heartburn often is exacerbated by meals and may awaken the patient. Associated symptoms include regurgitation of acid or nonacidic fluid and water brash, the reflex release of salty salivary secretions into the mouth. Atypical symptoms include pharyngitis, asthma, cough, bronchitis, hoarseness, and chest pain that mimics angina. Some patients with acid reflux on esophageal pH testing do not report heartburn, but note abdominal pain or other symptoms.

Dyspeptic patients typically report symptoms referable to the upper abdomen that may be meal-related, as with postprandial distress syndrome, or possibly independent of food ingestion in epigastric pain syndrome. Functional dyspepsia overlaps with other disorders including GERD, IBS, and idiopathic gastroparesis.

The physical examination with GERD and functional dyspepsia usually is normal. In atypical GERD, pharyngeal erythema and wheezing may be noted. Recurrent acid regurgitation may cause poor dentition. Dyspeptics may exhibit epigastric tenderness or distention.

Discriminating functional from organic causes of indigestion mandates excluding certain historic and examination features. Odynophagia suggests esophageal infection. Dysphagia is concerning for a benign or malignant esophageal blockage. Other alarm features include unexplained weight loss, recurrent vomiting, occult or gross bleeding, jaundice, palpable mass or adenopathy, and a family history of gastrointestinal neoplasm.

DIAGNOSTIC TESTING

Because indigestion is prevalent and most cases result from GERD or functional dyspepsia, a general principle is to perform only limited and directed diagnostic testing in selected individuals.

Once alarm factors are excluded **(Table 41-3)**, patients with typical GERD do not need further evaluation and are treated empirically. Upper endoscopy is indicated to exclude mucosal injury in cases with atypical symptoms or alarm factors. For heartburn >5 years in duration, especially in patients >50 years old, endoscopy is advocated to screen for Barrett's metaplasia. Endoscopy is not needed in low risk patients who exhibit a therapeutic response to acid suppressants. Ambulatory esophageal pH testing using a catheter method or a wireless capsule endoscopically attached to the esophageal wall is considered for drug-refractory symptoms and atypical symptoms like unexplained chest pain. High-resolution esophageal manometry is ordered when surgical treatment of GERD is considered. A low LES pressure predicts failure of drug therapy and provides a rationale to proceed to surgery. Poor esophageal body peristalsis raises concern about postoperative dysphagia and directs the choice of surgical technique. Nonacidic reflux may be detected by combined esophageal impedance-pH testing in medication-unresponsive patients.

Upper endoscopy is recommended as the initial test in patients with unexplained dyspepsia who are >55 years old or who have alarm factors because of the purported elevated risks of malignancy and ulcer in these groups. However, findings of endoscopy performed for uninvestigated dyspepsia include erosive esophagitis in 13%, peptic ulcer in 8%, and gastric or esophageal malignancy in only 0.3%. Management of patients <55 years old without alarm factors depends on the local prevalence of *H. pylori* infection. In regions with low *H.*

pylori prevalence (<10%), a 4-week trial of an acid-suppressing medication such as a proton pump inhibitor (PPI) is recommended. If this fails, a “test and treat” approach is most commonly applied. *H. pylori* status is determined with urea breath testing or stool antigen measurement. Those who are *H. pylori* positive are given therapy to eradicate the infection. If symptoms resolve on either regimen, no further intervention is required. For patients in areas with high *H. pylori* prevalence (>10%), an initial test and treat approach is advocated, with a subsequent trial of an acid-suppressing regimen offered for those in whom *H. pylori* treatment fails or for those who are negative for the infection. In each of these patient subsets, upper endoscopy is reserved for those whose symptoms fail to respond to therapy.

Further testing is indicated in some settings. If bleeding is noted, a blood count can exclude anemia. Thyroid chemistries or calcium levels screen for metabolic disease, whereas specific serologies may suggest celiac disease. Pancreatic and liver chemistries are obtained for possible pancreaticobiliary causes which are further investigated with ultrasound, CT, or MRI. Gastric emptying testing is considered to exclude gastroparesis for dyspeptic symptoms that resemble postprandial distress when drug therapy fails and in some GERD patients, especially if surgical intervention is an option. Breath testing after carbohydrate ingestion detects lactase deficiency, intolerance to other carbohydrates, or small-intestinal bacterial overgrowth.

TABLE 41-3

Alarm Symptoms in Gastroesophageal Reflux Disease

Odynophagia or dysphagia
Unexplained weight loss
Recurrent vomiting
Occult or gross gastrointestinal bleeding
Jaundice
Palpable mass or adenopathy
Family history of gastroesophageal malignancy

TREATMENT

TREATMENT

Indigestion

GENERAL PRINCIPLES

For mild indigestion, reassurance that a careful evaluation revealed no serious organic disease may be the only intervention needed. Drugs that cause gastroesophageal reflux or dyspepsia should be stopped, if possible. Patients with GERD should limit ethanol, caffeine, chocolate, and tobacco use due to their effects on the LES. Other measures in GERD include ingesting a low-fat diet, avoiding snacks before bedtime, and elevating the head of the bed. Patients with functional dyspepsia also may be advised to reduce intake of fat, spicy foods, caffeine, and alcohol.

Specific therapies for organic disease should be offered when possible. Surgery is appropriate for biliary colic. Diet changes are indicated for lactase deficiency or celiac disease. Peptic ulcers may be cured by specific medical regimens. However, because most indigestion is caused by GERD or functional dyspepsia, medications that reduce gastric acid, modulate motility, or blunt gastric sensitivity are used.

ACID-SUPPRESSING OR NEUTRALIZING MEDICATIONS

Drugs that reduce or neutralize gastric acid are often prescribed for GERD. Histamine H_2 antagonists like cimetidine, ranitidine, famotidine, and nizatidine are useful in mild to moderate GERD. For severe symptoms or for many cases of erosive or ulcerative esophagitis, PPIs like omeprazole, lansoprazole, rabeprazole, pantoprazole, esomeprazole, or dexlansoprazole are needed. These drugs inhibit gastric H^+, K^+-ATPase and are more potent than H_2 antagonists. Up to one-third of GERD patients do not respond to standard PPI doses; one-third of these patients have nonacidic reflux, whereas 10% have persistent acid-related disease. Heartburn typically responds better to PPI therapy than regurgitation or atypical GERD symptoms. Some individuals respond to doubling of the PPI dose or adding an H_2 antagonist at bedtime. Infrequent complications of long-term PPI therapy include diarrhea (from *Clostridium difficile* infection or microscopic colitis), small-intestinal bacterial overgrowth, nutrient deficiency (vitamin B_{12}, iron, calcium), hypomagnesemia, bone demineralization, interstitial nephritis, and impaired medication absorption (e.g., clopidogrel). Many patients started on a PPI can be stepped down to an H_2 antagonist or switched to an on-demand schedule.

Acid-suppressing drugs are also effective in selected patients with functional dyspepsia. A meta-analysis of 10 controlled trials calculated a risk ratio of 0.87, with a 95% confidence interval of 0.80–0.96, favoring PPI therapy over placebo. H_2 antagonists also reportedly improve symptoms in functional dyspepsia; however, findings of trials of this drug class likely are influenced by inclusion of large numbers of GERD patients.

Antacids are useful for short-term control of mild GERD but have less benefit in severe cases unless given at high doses that cause side effects (diarrhea and constipation with magnesium- and aluminum-containing agents, respectively). Alginic acid combined with antacids forms a floating barrier to reflux in patients with upright symptoms. Sucralfate, a salt of aluminum hydroxide and sucrose octasulfate that buffers acid and binds pepsin and bile salts, shows efficacy in GERD similar to H_2 antagonists.

HELICOBACTER PYLORI ERADICATION

H. pylori eradication is definitively indicated only for peptic ulcer and mucosa-associated lymphoid tissue gastric lymphoma. The benefits of eradication therapy in functional dyspepsia are limited but are statistically significant. A systematic review of 25 controlled trials calculated a pooled risk ratio of 1.24, with a 95% confidence interval of 1.12–1.37, favoring *H. pylori* eradication over placebo. Most drug combinations **(Chaps. 158 and 317)** include 10–14 days of a PPI or bismuth subsalicylate with two antibiotics. *H. pylori* infection is associated with reduced prevalence of GERD, especially in the elderly. However, eradication of the infection does not worsen GERD symptoms. No consensus recommendations regarding *H. pylori* eradication in GERD patients have been offered.

AGENTS THAT MODIFY GASTROINTESTINAL MOTOR ACTIVITY

Prokinetics like metoclopramide, erythromycin, and domperidone have limited utility in GERD. The γ-aminobutyric acid B (GABA-B) agonist baclofen reduces esophageal exposure to acid and nonacidic fluids by reducing TLESRs by 40%; this drug is proposed as adjunctive therapy for refractory acid and nonacid reflux. Several studies have promoted the efficacy of motor-stimulating drugs in functional dyspepsia with 33% relative risk reductions, but publication bias and small sample sizes raise questions about reported benefits of these agents. Some clinicians suggest that patients with the postprandial distress subtype may respond preferentially to prokinetic drugs. The 5-HT_{1A} agonists buspirone and tandospirone may improve some functional dyspepsia symptoms by enhancing meal-induced gastric accommodation. Acotiamide promotes gastric emptying and augments accommodation by enhancing acetylcholine release via muscarinic receptor antagonism and acetylcholinesterase inhibition. This agent is approved for functional dyspepsia in Japan.

ANTIDEPRESSANTS

Some patients with refractory functional heartburn may respond to antidepressants in tricyclic and selective serotonin reuptake inhibitor (SSRI) classes, although studies are limited. Their mechanism of action may involve blunting of visceral pain processing in the brain. In a recent controlled trial in functional dyspepsia, the tricyclic drug amitriptyline produced symptom reductions while the SSRI escitalopram had no benefit in a 3-way comparison with placebo. In another controlled trial in functional dyspepsia, the antidepressant mirtazapine produced superior symptom reductions versus placebo.

OTHER OPTIONS

Antireflux surgery (fundoplication) to increase LES pressure may be offered to GERD patients who are young and require lifelong therapy, have typical heartburn and regurgitation, are responsive to PPIs, and show acid reflux on pH monitoring. Surgery also is effective for some cases of nonacidic reflux. Individuals who respond less well to fundoplication include those with atypical symptoms or who have esophageal body motor disturbances. Dysphagia, gas-bloat syndrome, and gastroparesis are long-term complications of fundoplication; ~60% develop recurrent GERD symptoms over time. Studies assessing the utility and safety of gastroesophageal junction endoscopic therapies (radiofrequency therapy, transoral fundoplication, endoscopic stapling, antireflux mucosectomy) and laparoscopic magnetic sphincter augmentation to enhance gastroesophageal barrier function in GERD are ongoing.

Gas and bloating can be troubling symptoms in some patients with indigestion that are difficult to treat. Dietary exclusion of gas-producing foods such as legumes and use of simethicone or activated charcoal provide benefits in some cases. Low FODMAP (fermentable oligosaccharide, disaccharide, monosaccharide, and polyol) diets and therapies to modify gut flora (nonabsorbable antibiotics, probiotics) reduce gaseous symptoms in some IBS patients. The utility of low-FODMAP diets, antibiotics, and probiotics in functional dyspepsia is unproven. Herbal remedies such as STW 5 (Iberogast, a mixture of nine herbal agents) are useful in some dyspeptic patients. Psychological treatments (e.g., behavioral therapy, psychotherapy, hypnotherapy) may be offered for refractory functional dyspepsia, but no convincing data confirm their efficacy.

FURTHER READING

Hasler WL: Newest drugs for unexplained nausea and vomiting. Curr Treat Options Gastroenterol 14:371, 2016. [PubMed: 27726068]

Patti MG: An evidence-based approach to the treatment of gastroesophageal reflux disease. JAMA Surg 151:73, 2016. [PubMed: 26629969]

Scarpellini E et al: Management of refractory typical GERD symptoms. Nat Rev Gastroenterol Hepatol 13:281, 2016. [PubMed: 27075264]

Stanghellini V et al: Gastroduodenal disorders. Gastroenterology 150:1380, 2016. [PubMed: 27147122]

Talley NJ, Ford AC: Functional dyspepsia. N Engl J Med 373:1853, 2015. [PubMed: 26535514]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 42: Diarrhea and Constipation

Michael Camilleri; Joseph A. Murray

INTRODUCTION

Diarrhea and constipation are exceedingly common and, together, exact an enormous toll in terms of mortality, morbidity, social inconvenience, loss of work productivity, and consumption of medical resources. Worldwide, >1 billion individuals suffer one or more episodes of acute diarrhea each year. Among the 100 million persons affected annually by acute diarrhea in the United States, nearly half must restrict activities, 10% consult physicians, ~250,000 require hospitalization, and ~5000 die (primarily the elderly). The annual economic burden to society may exceed $20 billion. Acute infectious diarrhea remains one of the most common causes of mortality in developing countries, particularly among impoverished infants, accounting for 1.8 million deaths per year. Recurrent, acute diarrhea in children in tropical countries results in environmental enteropathy with long-term impacts on physical and intellectual development.

Constipation, by contrast, is rarely associated with mortality and is exceedingly common in developed countries, leading to frequent self-medication and, in a third of those, to medical consultation. Population statistics on chronic diarrhea and constipation are more uncertain, perhaps due to variable definitions and reporting, but the frequency of these conditions is also high. U.S. population surveys put prevalence rates for chronic diarrhea at 2–7% and for chronic constipation at 12–19%, with women being affected twice as often as men. Diarrhea and constipation are among the most common patient complaints presenting to internists and primary care physicians, and they account for nearly 50% of referrals to gastroenterologists.

Although diarrhea and constipation may present as mere nuisance symptoms at one extreme, they can be severe or life threatening at the other. Even mild symptoms may signal a serious underlying gastrointestinal (GI) lesion, such as colorectal cancer, or systemic disorder, such as thyroid disease. Given the heterogeneous causes and potential severity of these common complaints, it is imperative for clinicians to appreciate the pathophysiology, etiologic classification, diagnostic strategies, and principles of management of diarrhea and constipation, so that rational and cost-effective care can be delivered.

NORMAL PHYSIOLOGY

While the primary function of the small intestine is the digestion and assimilation of nutrients from food, the small intestine and colon together perform important functions that regulate the secretion and absorption of water and electrolytes, the storage and subsequent transport of intraluminal contents aborally, and the salvage of some nutrients that are not absorbed in the small intestine after bacterial metabolism of carbohydrate allows salvage of short-chain fatty acids. The main motor functions are summarized in Table 42-1. Alterations in fluid and electrolyte handling contribute significantly to diarrhea. Alterations in motor and sensory functions of the colon result in highly prevalent syndromes such as irritable bowel syndrome (IBS), chronic diarrhea, and chronic constipation.

TABLE 42-1

Normal Gastrointestinal Motility: Functions at Different Anatomic Levels

Stomach and Small Bowel
Synchronized MMC in fasting
Accommodation, trituration, mixing, transit
Stomach ~3 h
Small bowel ~3 h
Ileal reservoir empties boluses
Colon: Irregular Mixing, Fermentation, Absorption, Transit
Ascending, transverse: reservoirs
Descending: conduit
Sigmoid/rectum: volitional reservoir

Abbreviation: MMC, migrating motor complex.

NEURAL CONTROL

The small intestine and colon have intrinsic and extrinsic innervation. The *intrinsic innervation*, also called the enteric nervous system, comprises myenteric, submucosal, and mucosal neuronal layers. The function of these layers is modulated by interneurons through the actions of neurotransmitter amines or peptides, including acetylcholine, vasoactive intestinal peptide (VIP), opioids, norepinephrine, serotonin, adenosine triphosphate (ATP), and nitric oxide (NO). The myenteric plexus regulates smooth-muscle function through intermediary pacemaker-like cells called the interstitial cells of Cajal, and the submucosal plexus affects secretion, absorption, and mucosal blood flow. The enteric nervous system receives input from the extrinsic nerves, but it is capable of independent control of these functions.

The *extrinsic innervations* of the small intestine and colon are part of the autonomic nervous system and also modulate motor and secretory functions. The parasympathetic nerves convey visceral sensory pathways from and excitatory pathways to the small intestine and colon. Parasympathetic fibers via the vagus nerve reach the small intestine and proximal colon along the branches of the superior mesenteric artery. The distal colon is supplied by sacral parasympathetic nerves (S_{2-4}) via the pelvic plexus; these fibers course through the wall of the colon as ascending intracolonic fibers as far as, and in some instances including, the proximal colon. The chief excitatory neurotransmitters controlling motor function are acetylcholine and the tachykinins, such as substance P. The sympathetic nerve supply modulates motor functions and reaches the small intestine and colon alongside their arterial vessels. Sympathetic input to the gut is generally excitatory to sphincters and inhibitory to non-sphincteric muscle. Visceral afferents convey

sensation from the gut to the central nervous system (CNS). Some afferent fibers synapse in the prevertebral ganglia and reflexly modulate intestinal motility, blood flow, and secretion.

INTESTINAL FLUID ABSORPTION AND SECRETION

On an average day, 9 L of fluid enter the GI tract, ~1 L of residual fluid reaches the colon, and the stool excretion of fluid constitutes about 0.2 L/d. The colon has a large capacitance and functional reserve and may recover up to four times its usual volume of 0.8 L/d, provided the rate of flow permits reabsorption to occur. Thus, the colon can partially compensate for excess fluid delivery to the colon that may result from intestinal absorptive or secretory disorders.

In the small intestine and colon, sodium absorption is predominantly electrogenic (i.e., it can be measured as an ionic current across the membrane because there is not an equivalent loss of a cation from the cell), and uptake takes place at the apical membrane; it is compensated for by the export functions of the basolateral sodium pump. There are several active transport proteins at the apical membrane, especially in the small intestine, whereby sodium ion entry is coupled to monosaccharides (e.g., glucose through the transporter SGLT1, or fructose through GLUT-5). Glucose then exits the basal membrane through a specific transport protein, GLUT-5, creating a glucose concentration and osmotic gradient between the lumen and the intercellular space, drawing water and electrolytes passively from the lumen. A variety of neural and nonneural mediators regulate colonic fluid and electrolyte balance, including cholinergic, adrenergic, and serotonergic mediators. Angiotensin and aldosterone also influence colonic absorption, reflecting the common embryologic development of the distal colonic epithelium and the renal tubules.

SMALL-INTESTINAL MOTILITY

During the fasting period, the motility of the small intestine is characterized by a cyclical event called the migrating motor complex (MMC), which serves to clear nondigestible residue from the small intestine (the intestinal “housekeeper”). This organized, propagated series of contractions lasts, on average, 4 min, occurs every 60–90 min, and usually involves the entire small intestine. After food ingestion, the small intestine produces irregular, mixing contractions of relatively low amplitude, except in the distal ileum where more powerful contractions occur intermittently and empty the ileum by bolus transfers.

ILEOCOLONIC STORAGE AND SALVAGE

The distal ileum acts as a reservoir, emptying intermittently by bolus movements. This action allows time for salvage of fluids, electrolytes, and nutrients. Segmentation by haustra compartmentalizes the colon and facilitates mixing, retention of residue, and formation of solid stools. There is increased appreciation of the intimate interaction between the colonic function and the luminal ecology. The resident microorganisms, predominantly anaerobic bacteria, in the colon are necessary for the digestion of unabsorbed carbohydrates that reach the colon even in health, thereby providing a vital source of nutrients to the mucosa. Normal intestinal flora also keeps pathogens at bay by a variety of mechanisms including a crucial role in the development and maintenance of a potent but well-regulated immune response capacity to pathogens and tolerance to normal ingesta. In health, the ascending and transverse regions of colon function as reservoirs (average transit time, 15 h), and the descending colon acts as a conduit (average transit time, 3 h). The colon is efficient at conserving sodium and water, a function that is particularly important in sodium-depleted patients in whom the small intestine alone is unable to maintain sodium balance. Diarrhea or constipation may result from alteration in the reservoir function of the proximal colon or the propulsive function of the left colon. Constipation may also result from disturbances of the rectal or sigmoid reservoir, typically as a result of dysfunction of the pelvic floor, the anal sphincters, the coordination of defecation, or dehydration.

COLONIC MOTILITY AND TONE

The small-intestinal MMC only rarely continues into the colon. However, short duration or phasic contractions mix colonic contents and high-amplitude (>75 mmHg) propagated contractions (HAPCs) are sometimes associated with mass movements through the colon and normally occur approximately five times per day, usually on awakening in the morning and postprandially. Increased frequency of HAPCs may result in diarrhea or urgency. The predominant phasic contractions in the colon are irregular and nonpropagated and serve a “mixing” function.

Colonic tone refers to the background contractility upon which phasic contractile activity (typically contractions lasting <15 s) is superimposed. It is an important cofactor in the colon’s capacitance (volume accommodation) and sensation.

COLONIC MOTILITY AFTER MEAL INGESTION

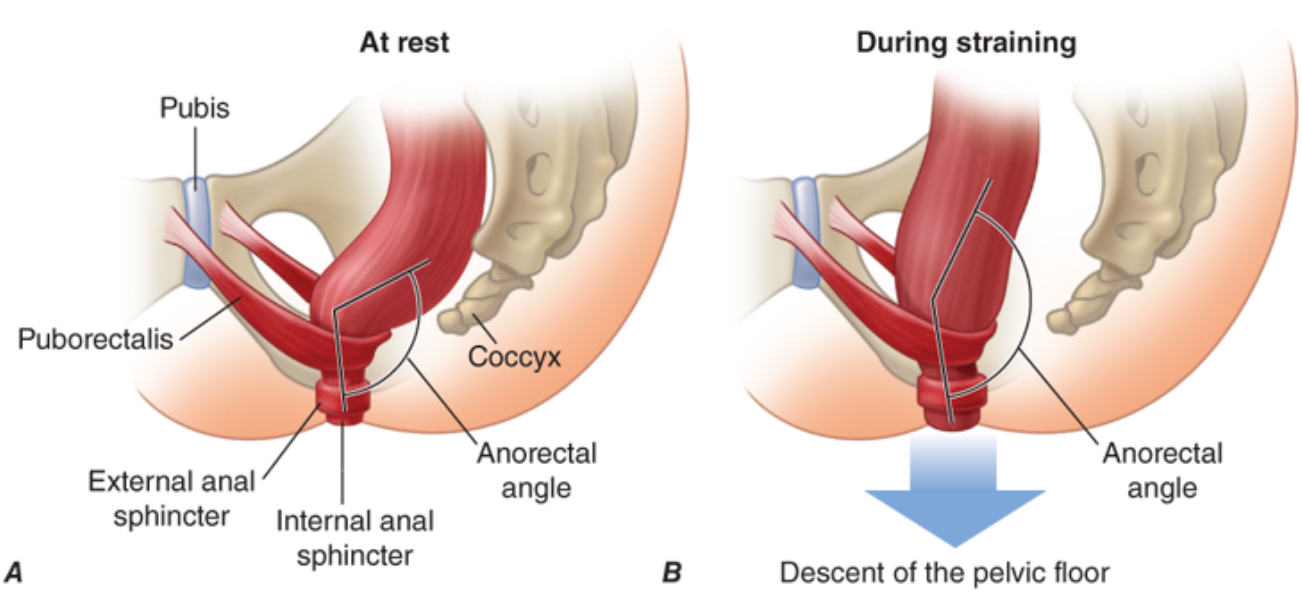
After meal ingestion, colonic phasic and tonic contractility increases for a period of ~2 h. The initial phase (~10 min) is mediated by the vagus nerve in response to mechanical distention of the stomach. The subsequent response of the colon requires caloric stimulation (e.g., intake of at least 500 kcal) and is mediated, at least in part, by hormones (e.g., gastrin and serotonin).

DEFECATION

Tonic contraction of the puborectalis muscle, which forms a sling around the rectoanal junction, is important to maintain continence; during defecation, sacral parasympathetic nerves relax this muscle, facilitating the straightening of the rectoanal angle **(Fig. 42-1)**. Distention of the rectum results in transient relaxation of the internal anal sphincter via intrinsic and reflex sympathetic innervation. As sigmoid and rectal contractions, as well as straining (Valsalva maneuver), which increases intraabdominal pressure, increase the pressure within the rectum, the rectosigmoid angle opens by >15°. Voluntary relaxation of the external anal sphincter (striated muscle innervated by the pudendal nerve) in response to the sensation produced by distention permits the evacuation of feces. Defecation can also be delayed voluntarily by contraction of the external anal sphincter.

FIGURE 42-1

Sagittal view of the anorectum (*A*) at rest and (*B*) during straining to defecate. Continence is maintained by normal rectal sensation and tonic contraction of the internal anal sphincter and the puborectalis muscle, which wraps around the anorectum, maintaining an anorectal angle between 80° and 110°. During defecation, the pelvic floor muscles (including the puborectalis) relax, allowing the anorectal angle to straighten by at least 15°, and the perineum descends by 1–3.5 cm. The external anal sphincter also relaxes and reduces pressure on the anal canal. *(Reproduced with permission from A Lembo, M Camilleri: N Engl J Med 349:1360, 2003.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIARRHEA

DEFINITION

Diarrhea is loosely defined as passage of abnormally liquid or unformed stools at an increased frequency. For adults on a typical Western diet, stool weight >200 g/d can generally be considered diarrheal. Diarrhea may be further defined as *acute* if <2 weeks, *persistent* if 2–4 weeks, and *chronic* if >4 weeks in duration.

Two common conditions, usually associated with the passage of stool totaling <200 g/d, must be distinguished from diarrhea, because diagnostic and therapeutic algorithms differ. *Pseudodiarrhea*, or the frequent passage of small volumes of stool, is often associated with rectal urgency, tenesmus, or a feeling of incomplete evacuation, and accompanies IBS or proctitis. *Fecal incontinence* is the involuntary discharge of rectal contents and is most often caused by neuromuscular disorders or structural anorectal problems. Diarrhea and urgency, especially if severe, may aggravate or cause incontinence. Pseudodiarrhea and fecal incontinence occur at prevalence rates comparable to or higher than that of chronic diarrhea and should always be considered in patients complaining of "diarrhea." Overflow diarrhea may occur in nursing home patients due to fecal impaction that is readily detectable by rectal examination. A careful history and physical examination generally allow these conditions to be discriminated from true diarrhea.

ACUTE DIARRHEA

More than 90% of cases of acute diarrhea are caused by infectious agents; these cases are often accompanied by vomiting, fever, and abdominal pain. The remaining 10% or so are caused by medications, toxic ingestions, ischemia, food indiscretions, and other conditions.

Infectious Agents

Most infectious diarrheas are acquired by fecal-oral transmission or, more commonly, via ingestion of food or water contaminated with pathogens from human or animal feces. In the immunocompetent person, the resident fecal microflora, containing >500 taxonomically distinct species, are rarely the source of diarrhea and may actually play a role in

suppressing the growth of ingested pathogens. Disturbances of flora by antibiotics can lead to diarrhea by reducing the digestive function or by allowing the overgrowth of pathogens, such as *Clostridium difficile* (Chap. 129). Acute infection or injury occurs when the ingested agent overwhelms or bypasses the host's mucosal immune and nonimmune (gastric acid, digestive enzymes, mucus secretion, peristalsis, and suppressive resident flora) defenses. Established clinical associations with specific enteropathogens may offer diagnostic clues.

In the United States, five high-risk groups are recognized:

1. *Travelers*. Nearly 40% of tourists to endemic regions of Latin America, Africa, and Asia develop so-called traveler's diarrhea, most commonly due to enterotoxigenic or enteroaggregative *Escherichia coli* as well as to *Campylobacter*, *Shigella*, *Aeromonas*, norovirus, *Coronavirus*, and *Salmonella*. Visitors to Russia (especially St. Petersburg) may have increased risk of *Giardia*-associated diarrhea; visitors to Nepal may acquire *Cyclospora*. Campers, backpackers, and swimmers in wilderness areas may become infected with *Giardia*. Cruise ships may be affected by outbreaks of gastroenteritis caused by agents such as norovirus.

2. *Consumers of certain foods*. Diarrhea closely following food consumption at a picnic, banquet, or restaurant may suggest infection with *Salmonella*, *Campylobacter*, or *Shigella* from chicken; enterohemorrhagic *E. coli* (O157:H7) from undercooked hamburger; *Bacillus cereus* from fried rice or other reheated food; *Staphylococcus aureus* or *Salmonella* from mayonnaise or creams; *Salmonella* from eggs; *Listeria* from fresh or frozen uncooked foods or soft cheeses; and *Vibrio* species, *Salmonella*, or acute hepatitis A from seafood, especially if raw. State departments of public health issue communications regarding food-related illnesses, which may have originated domestically or been imported, but ultimately cause epidemics in the United States (e.g., the *Cyclospora* epidemic of 2013 in midwestern states that resulted from bagged salads).

3. *Immunodeficient persons*. Individuals at risk for diarrhea include those with either primary immunodeficiency (e.g., IgA deficiency, common variable hypogammaglobulinemia, chronic granulomatous disease) or the much more common secondary immunodeficiency states (e.g., AIDS, senescence, pharmacologic suppression). Common enteric pathogens often cause a more severe and protracted diarrheal illness, and, particularly in persons with AIDS, opportunistic infections, such as by *Mycobacterium* species, certain viruses (cytomegalovirus, adenovirus, and herpes simplex), and protozoa (*Cryptosporidium*, *Isospora belli*, Microsporida, and *Blastocystis hominis*) may also play a role (Chap. 197). In patients with AIDS, agents transmitted venereally per rectum or by extension from vaginal infection (e.g., *Neisseria gonorrhoeae*, *Treponema pallidum*, *Chlamydia*) may contribute to proctocolitis. Symptoms suggesting anorectal disease, particularly pain, may result from constipation occurring coincidentally in a person with immunodeficiency. Persons with hemochromatosis are especially prone to invasive, even fatal, enteric infections with *Vibrio* species and *Yersinia* infections and should avoid raw fish.

4. *Daycare attendees and their family members*. Infections with *Shigella*, *Giardia*, *Cryptosporidium*, rotavirus, and other agents are very common and should be considered.

5. *Institutionalized persons*. Infectious diarrhea is one of the most frequent categories of nosocomial infections in many hospitals and long-term care facilities; the causes are a variety of microorganisms but most commonly *C. difficile*. *C. difficile* can affect those with no history of antibiotic use and may be acquired in the community.

The pathophysiology underlying acute diarrhea by infectious agents produces specific clinical features that may also be helpful in diagnosis (Table 42-2). Profuse, watery diarrhea secondary to small-bowel hypersecretion occurs with ingestion of preformed bacterial toxins, enterotoxin-producing bacteria, and enteroadherent pathogens. Diarrhea associated with marked vomiting and minimal or no fever may occur abruptly within a few hours after ingestion of the former two types; vomiting is usually less, abdominal cramping or bloating is greater, and fever is higher with the latter. Cytotoxin-producing

and invasive microorganisms all cause high fever and abdominal pain. Invasive bacteria and *Entamoeba histolytica* often cause bloody diarrhea (referred to as *dysentery*). *Yersinia* invades the terminal ileal and proximal colon mucosa and may cause especially severe abdominal pain with tenderness mimicking acute appendicitis.

TABLE 42-2

Association Between Pathobiology of Causative Agents and Clinical Features in Acute Infectious Diarrhea

Pathobiology/Agents	Incubation Period	Vomiting	Abdominal Pain	Fever	Diarrhea
Toxin producers					
Preformed toxin					
Bacillus cereus, *Staphylococcus aureus*, *Clostridium perfringens*	1–8 h 8–24 h	3–4+	1–2+	0–1+	3–4+, watery
Enterotoxin					
Vibrio cholerae, enterotoxigenic *Escherichia coli*, *Klebsiella pneumoniae*, *Aeromonas* species	8–72 h	2–4+	1–2+	0–1+	3–4+, watery
Enteroadherent					
Enteropathogenic and enteroadherent *E. coli*, *Giardia* organisms, cryptosporidiosis, helminths	1–8 d	0–1+	1–3+	0–2+	1–2+, watery, mushy
Cytotoxin producers					
Clostridium difficile	1–3 d	0–1+	3–4+	1–2+	1–3+, usually watery, occasionally bloody
Hemorrhagic *E. coli*	12–72 h	0–1+	3–4+	1–2+	1–3+, initially watery, quickly bloody
Invasive organisms					
Minimal inflammation					
Rotavirus and norovirus	1–3 d	1–3+	2–3+	3–4+	1–3+, watery
Variable inflammation					

Pathobiology/Agents	Incubation Period	Vomiting	Abdominal Pain	Fever	Diarrhea
Salmonella, *Campylobacter*, and *Aeromonas* species, *Vibrio parahaemolyticus*, *Yersinia*	12 h–11 d	0–3+	2–4+	3–4+	1–4+, watery or bloody
Severe inflammation					
Shigella species, enteroinvasive *E. coli*, *Entamoeba histolytica*	12 h–8 d	0–1+	3–4+	3–4+	1–2+, bloody

Source: Adapted from DW Powell, in T Yamada (ed): *Textbook of Gastroenterology and Hepatology*, 4th ed. Philadelphia, Lippincott Williams & Wilkins, 2003.

Finally, infectious diarrhea may be associated with systemic manifestations. Reactive arthritis (formerly known as Reiter's syndrome), arthritis, urethritis, and conjunctivitis may accompany or follow infections by *Salmonella*, *Campylobacter*, *Shigella*, and *Yersinia*. Yersiniosis may also lead to an autoimmune-type thyroiditis, pericarditis, and glomerulonephritis. Both enterohemorrhagic *E. coli* (O157:H7) and *Shigella* can lead to the *hemolytic-uremic syndrome* with an attendant high mortality rate. The syndrome of postinfectious IBS has now been recognized as a complication of infectious diarrhea. Similarly, acute gastroenteritis may precede the diagnosis of celiac disease or Crohn's disease. Acute diarrhea can also be a major symptom of several systemic infections including *viral hepatitis*, *listeriosis*, *legionellosis*, and *toxic shock syndrome*.

Other Causes

Side effects from medications are probably the most common noninfectious causes of acute diarrhea, and etiology may be suggested by a temporal association between use and symptom onset. Although innumerable medications may produce diarrhea, some of the more frequently incriminated include antibiotics, cardiac antidysrhythmics, antihypertensives, nonsteroidal anti-inflammatory drugs (NSAIDs), certain antidepressants, chemotherapeutic agents, bronchodilators, antacids, and laxatives. Occlusive or nonocclusive ischemic colitis typically occurs in persons aged >50 years; often presents as acute lower abdominal pain preceding watery, then bloody diarrhea; and generally results in acute inflammatory changes in the sigmoid or left colon while sparing the rectum. Acute diarrhea may accompany colonic diverticulitis and graft-versus-host disease. Acute diarrhea, often associated with systemic compromise, can follow ingestion of toxins including organophosphate insecticides, amanita and other mushrooms, arsenic, and preformed toxins in seafood such as ciguatera (from algae that the fish eat) and scombroid (an excess of histamine due to inadequate refrigeration). Acute anaphylaxis to food ingestion can have a similar presentation. Conditions causing chronic diarrhea can also be confused with acute diarrhea early in their course. This confusion may occur with inflammatory bowel disease (IBD) and some of the other inflammatory chronic diarrheas that may have an abrupt rather than insidious onset and exhibit features that mimic infection.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Acute Diarrhea

The decision to evaluate acute diarrhea depends on its severity and duration and on various host factors **(Fig. 42-2)**. Most episodes of acute diarrhea are mild and self-limited and do not justify the cost and potential morbidity rate of diagnostic or pharmacologic interventions. Indications for evaluation include profuse diarrhea with dehydration, grossly bloody stools, fever ≥38.5°C (≥101°F), duration >48 h without improvement, recent antibiotic use, new community outbreaks, associated severe abdominal pain in patients aged >50 years, and elderly (≥70 years) or immunocompromised patients. In some cases of moderately severe febrile diarrhea associated with fecal leukocytes (or increased fecal levels of the leukocyte proteins, such as calprotectin) or with gross blood, a diagnostic evaluation might be avoided in favor of an empirical antibiotic trial (see below).

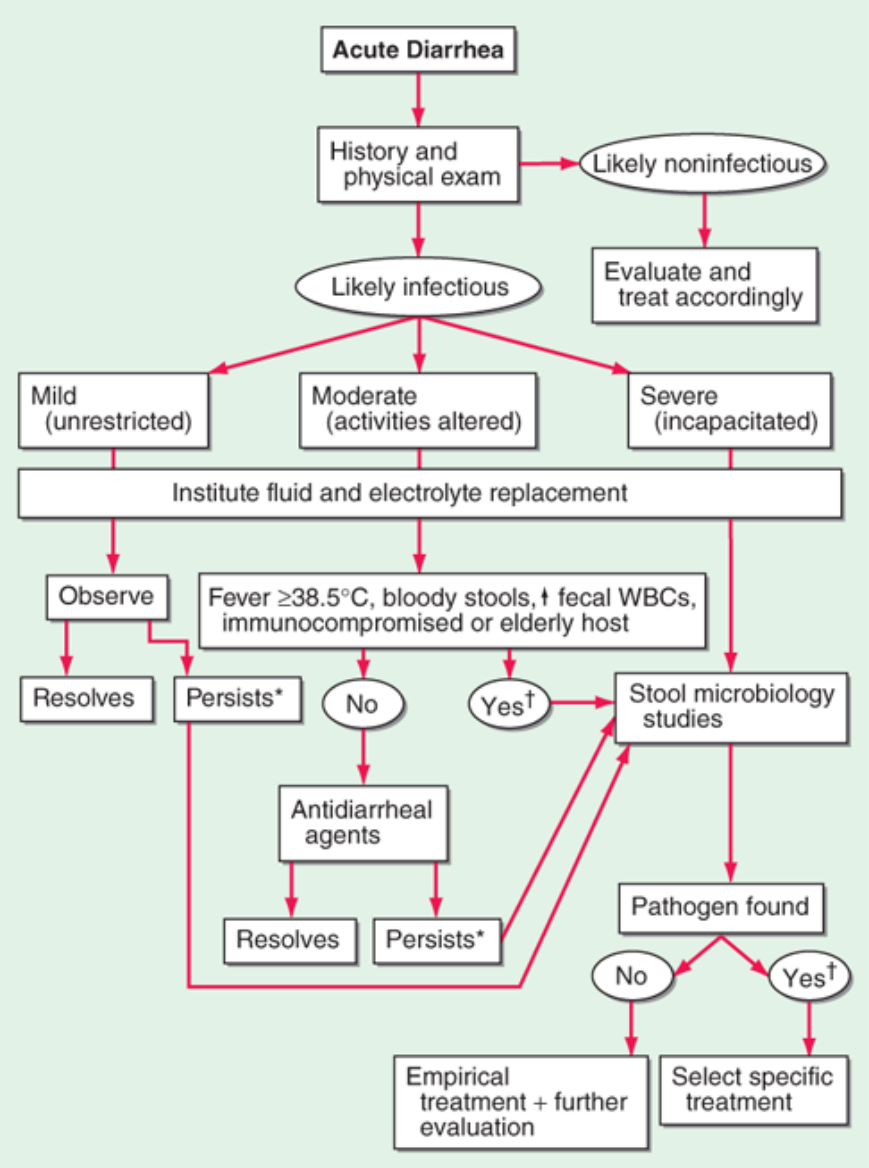
The cornerstone of diagnosis in those suspected of severe acute infectious diarrhea is microbiologic analysis of the stool. Workup includes cultures for bacterial and viral pathogens; direct inspection for ova and parasites; and immunoassays for certain bacterial toxins (*C. difficile*), viral antigens (rotavirus), and protozoal antigens (*Giardia*, *E. histolytica*). The aforementioned clinical and epidemiologic associations may assist in focusing the evaluation. If a particular pathogen or set of possible pathogens is so implicated, either the whole panel of routine studies may not be necessary or, in some instances, special cultures may be appropriate as for enterohemorrhagic and other types of *E. coli*, *Vibrio* species, and *Yersinia*. Molecular diagnosis of pathogens in stool can be made by identification of unique DNA sequences, and evolving microarray technologies have led to more rapid, sensitive, specific, and cost-effective diagnosis.

Persistent diarrhea is commonly due to *Giardia* **(Chap. 218)**, but additional causative organisms that should be considered include *C. difficile* (especially if antibiotics had been administered), *E. histolytica*, *Cryptosporidium*, *Campylobacter*, and others. If stool studies are unrevealing, flexible sigmoidoscopy with biopsies and upper endoscopy with duodenal aspirates and biopsies may be indicated. Brainerd diarrhea is an increasingly recognized entity characterized by an abrupt-onset diarrhea that persists for at least 4 weeks, but may last 1–3 years, and is thought to be of infectious origin. It may be associated with subtle inflammation of the distal small intestine or proximal colon.

Structural examination by sigmoidoscopy, colonoscopy, or abdominal computed tomography (CT) scanning (or other imaging approaches) may be appropriate in patients with uncharacterized persistent diarrhea to exclude IBD or as an initial approach in patients with suspected noninfectious acute diarrhea such as might be caused by ischemic colitis, diverticulitis, or partial bowel obstruction.

FIGURE 42-2

Algorithm for the management of acute diarrhea. Consider empirical treatment before evaluation with (*) metronidazole and with (†) quinolone. WBCs, white blood cells.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Acute Diarrhea

Fluid and electrolyte replacement are of central importance to all forms of acute diarrhea. Fluid replacement alone may suffice for mild cases. Oral sugar-electrolyte solutions (iso-osmolar sport drinks or designed formulations) should be instituted promptly with severe diarrhea to limit dehydration, which is the major cause of death. Profoundly dehydrated patients, especially infants and the elderly, require IV rehydration.

In moderately severe nonfebrile and nonbloody diarrhea, antimotility and antisecretory agents such as loperamide can be useful adjuncts to control symptoms. Such agents should be avoided with febrile dysentery, which may be exacerbated or prolonged by them. Bismuth subsalicylate may reduce symptoms of vomiting and diarrhea but should not be used to treat immunocompromised patients or those with renal impairment because of the risk of bismuth encephalopathy.

Judicious use of antibiotics is appropriate in selected instances of acute diarrhea and may reduce its severity and duration (Fig. 42-2). Many physicians treat moderately to severely ill patients with febrile dysentery empirically without diagnostic evaluation using a quinolone, such as ciprofloxacin (500 mg bid for 3–5 d). Empirical treatment can also be considered for suspected giardiasis with metronidazole (250 mg qid for 7 d). Selection of antibiotics and dosage regimens are otherwise dictated by specific pathogens, geographic patterns of resistance, and conditions found **(Chaps. 128, 156, and 160, 161, 162, 163, 164, 165, 166)**. Because of resistance to first-line treatments, newer agents such as nitazoxanide may be required for *Giardia* and *Cryptosporidium* infections. Antibiotic coverage is indicated, whether or not a causative organism is discovered, in patients who are immunocompromised, have mechanical heart valves or recent vascular grafts, or are elderly. Bismuth subsalicylate may reduce the frequency of traveler's diarrhea. Antibiotic prophylaxis is only indicated for certain patients traveling to high-risk countries in whom the likelihood or seriousness of acquired diarrhea would be especially high, including those with immunocompromise, IBD, hemochromatosis, or gastric achlorhydria. Use of ciprofloxacin, azithromycin, or rifaximin may reduce bacterial diarrhea in such travelers by 90%, though rifaximin is not suitable for invasive disease but rather as treatment for uncomplicated traveler's diarrhea. There is little role for endoscopic evaluation in most circumstances except in immunocompromised patients. Finally, physicians should be vigilant to identify if an outbreak of diarrheal illness is occurring and to alert the public health authorities promptly. This may reduce the ultimate size of the affected population.

CHRONIC DIARRHEA

Diarrhea lasting >4 weeks warrants evaluation to exclude serious underlying pathology. In contrast to acute diarrhea, most of the causes of chronic diarrhea are noninfectious. The classification of chronic diarrhea by pathophysiologic mechanism facilitates a rational approach to management, although many diseases cause diarrhea by more than one mechanism **(Table 42-3)**.

TABLE 42-3

Major Causes of Chronic Diarrhea According to Predominant Pathophysiologic Mechanism

Secretory Causes
Exogenous stimulant laxatives
Chronic ethanol ingestion
Other drugs and toxins
Endogenous laxatives (dihydroxy bile acids)
Idiopathic secretory diarrhea or bile acid diarrhea
Certain bacterial infections
Bowel resection, disease, or fistula (↓ absorption)
Partial bowel obstruction or fecal impaction
Hormone-producing tumors (carcinoid, VIPoma, medullary cancer of thyroid, mastocytosis, gastrinoma, colorectal villous adenoma)
Addison's disease
Congenital electrolyte absorption defects
Osmotic Causes
Osmotic laxatives (Mg^{2+}, PO_4^{-3}, SO_4^{-2})
Lactase and other disaccharide deficiencies
Nonabsorbable carbohydrates (sorbitol, lactulose, polyethylene glycol)
Gluten and FODMAP intolerance
Steatorrheal Causes
Intraluminal maldigestion (pancreatic exocrine insufficiency, bacterial overgrowth, bariatric surgery, liver disease)
Mucosal malabsorption (celiac sprue, Whipple's disease, infections, abetalipoproteinemia, ischemia, drug-induced enteropathy)
Postmucosal obstruction (1° or 2° lymphatic obstruction)

Inflammatory Causes
Idiopathic inflammatory bowel disease (Crohn's, chronic ulcerative colitis)
Lymphocytic and collagenous colitis
Immune-related mucosal disease (1° or 2° immunodeficiencies, food allergy, eosinophilic gastroenteritis, graft-versus-host disease)
Infections (invasive bacteria, viruses, and parasites, Brainerd diarrhea)
Radiation injury
Gastrointestinal malignancies
Dysmotile Causes
Irritable bowel syndrome (including postinfectious IBS)
Visceral neuromyopathies
Hyperthyroidism
Drugs (prokinetic agents)
Postvagotomy
Factitial Causes
Munchausen
Eating disorders
Iatrogenic Causes
Cholecystectomy
Ileal resection
Bariatric surgery
Vagotomy, fundoplication

Abbreviation: FODMAP, fermentable oligosaccharides, disaccharides, monosaccharides, and polyols.

Secretory Causes

Secretory diarrheas are due to derangements in fluid and electrolyte transport across the enterocolonic mucosa. They are characterized clinically by watery, large-volume fecal outputs that are typically painless and persist with fasting. Because there is no malabsorbed solute, stool osmolality is accounted for by normal endogenous electrolytes with no fecal osmotic gap.

MEDICATIONS

Side effects from regular ingestion of drugs and toxins are the most common secretory causes of chronic diarrhea. Hundreds of prescription and over-the-counter medications (see earlier section, "Acute Diarrhea, Other Causes") may produce diarrhea. Surreptitious or habitual use of stimulant laxatives (e.g., senna, cascara, bisacodyl, ricinoleic acid [castor oil]) must also be considered. Chronic ethanol consumption may cause a secretory-type diarrhea due to enterocyte injury with impaired sodium and water absorption as well as rapid transit and other alterations. Inadvertent ingestion of certain environmental toxins (e.g., arsenic) may lead to chronic rather than acute forms of diarrhea. Certain bacterial infections may occasionally persist and be associated with a secretory-type diarrhea. The oral angiotensin-receptor blocker, olmesartan, is associated with diarrhea due to sprue-like enteropathy.

BOWEL RESECTION, MUCOSAL DISEASE, OR ENTEROCOLIC FISTULA

These conditions may result in a secretory-type diarrhea because of inadequate surface for reabsorption of secreted fluids and electrolytes. Unlike other secretory diarrheas, this subset of conditions tends to worsen with eating. With disease (e.g., Crohn's ileitis) or resection of <100 cm of terminal ileum, dihydroxy bile acids may escape absorption and stimulate colonic secretion (cholerheic diarrhea). This mechanism may contribute to so-called *idiopathic secretory diarrhea or bile acid diarrhea (BAD)*, in which bile acids are functionally malabsorbed from a normal-appearing terminal ileum. This *idiopathic bile acid malabsorption (BAM)* may account for an average of 40% of unexplained chronic diarrhea. Reduced negative feedback regulation of bile acid synthesis in hepatocytes by fibroblast growth factor 19 (FGF-19) produced by ileal enterocytes results in a degree of bile-acid synthesis that exceeds the normal capacity for ileal reabsorption, producing BAD. An alternative cause of BAD is a genetic variation in the receptor proteins (β-klotho and fibroblast growth factor 4) on the hepatocyte that normally mediate the effect of FGF-19. Dysfunction of these proteins prevents FGF-19 inhibition of hepatocyte bile acid synthesis. Another mechanism is based on genetic variation in the bile acid receptor (TGR5) in the colon, resulting in accelerated colonic transit.

Partial bowel obstruction, ostomy stricture, or fecal impaction may paradoxically lead to increased fecal output due to fluid hypersecretion.

HORMONES

Although uncommon, the classic examples of secretory diarrhea are those mediated by hormones. *Metastatic gastrointestinal carcinoid tumors* or, rarely, *primary bronchial carcinoids* may produce watery diarrhea alone or as part of the carcinoid syndrome that comprises episodic flushing, wheezing, dyspnea, and right-sided valvular heart disease. Diarrhea is due to the release into the circulation of potent intestinal secretagogues including serotonin, histamine, prostaglandins, and various kinins. Pellagra-like skin lesions may rarely occur as the result of serotonin overproduction with niacin depletion. *Gastrinoma*, one of the most common neuroendocrine tumors, most typically presents with refractory peptic ulcers, but diarrhea occurs in up to one-third of cases and may be the only clinical manifestation in 10%. While other secretagogues released with gastrin may play a role, the diarrhea most often results from fat maldigestion owing to pancreatic enzyme inactivation by low intraduodenal pH. The watery diarrhea hypokalemia achlorhydria syndrome, also called *pancreatic cholera*, is due to a non-β cell pancreatic adenoma, referred to as a *VIPoma*, that secretes VIP and a host of other peptide hormones including pancreatic polypeptide, secretin, gastrin, gastrin-inhibitory polypeptide (also called glucose-dependent insulinotropic peptide), neurotensin, calcitonin, and prostaglandins. The secretory diarrhea is often massive with stool volumes >3 L/d; daily volumes as high as 20 L have been reported. Life-threatening dehydration; neuromuscular dysfunction from associated hypokalemia, hypomagnesemia, or hypercalcemia; flushing; and hyperglycemia may accompany a VIPoma. *Medullary carcinoma of the thyroid* may present with watery

diarrhea caused by calcitonin, other secretory peptides, or prostaglandins. Prominent diarrhea is often associated with metastatic disease and poor prognosis. *Systemic mastocytosis*, which may be associated with the skin lesion urticaria pigmentosa, may cause diarrhea that is either secretory and mediated by histamine or inflammatory due to intestinal infiltration by mast cells. Large *colorectal villous adenomas* may rarely be associated with a secretory diarrhea that may cause hypokalemia, can be inhibited by NSAIDs, and are apparently mediated by prostaglandins.

CONGENITAL DEFECTS IN ION ABSORPTION

Rarely, defects in specific carriers associated with ion absorption cause watery diarrhea from birth. These disorders include defective Cl^-/HCO_3^- exchange (*congenital chloridorrhea*) with alkalosis (which results from a mutated *DRA* [down-regulated in adenoma] gene) and defective Na^+/H^+ exchange (*congenital sodium diarrhea*), which results from a mutation in the *NHE3* (sodium-hydrogen exchanger) gene and results in acidosis.

Some hormone deficiencies may be associated with watery diarrhea, such as occurs with adrenocortical insufficiency (Addison's disease) that may be accompanied by skin hyperpigmentation.

Osmotic Causes

Osmotic diarrhea occurs when ingested, poorly absorbable, osmotically active solutes draw enough fluid into the lumen to exceed the reabsorptive capacity of the colon. Fecal water output increases in proportion to such a solute load. Osmotic diarrhea characteristically ceases with fasting or with discontinuation of the causative agent.

OSMOTIC LAXATIVES

Ingestion of magnesium-containing antacids, health supplements, or laxatives may induce osmotic diarrhea typified by a stool osmotic gap (>50 mosmol/L): serum osmolarity (typically 290 mosmol/kg) − (2 × [fecal sodium + potassium concentration]). Measurement of fecal osmolarity is no longer recommended because, even when measured immediately after evacuation, it may be erroneous because carbohydrates are metabolized by colonic bacteria, causing an increase in osmolarity.

CARBOHYDRATE MALABSORPTION

Carbohydrate malabsorption due to acquired or congenital defects in brush-border disaccharidases and other enzymes leads to osmotic diarrhea with a low pH. One of the most common causes of chronic diarrhea in adults is *lactase deficiency*, which affects three-fourths of nonwhites worldwide and 5–30% of persons in the United States; the total lactose load at any one time influences the symptoms experienced. Most patients learn to avoid milk products without requiring treatment with enzyme supplements. Some sugars, such as sorbitol, lactulose, or fructose, are frequently malabsorbed, and diarrhea ensues with ingestion of medications, gum, or candies sweetened with these poorly or incompletely absorbed sugars.

WHEAT AND FODMAP INTOLERANCE

Chronic diarrhea, bloating, and abdominal pain are recognized as symptoms of non-celiac gluten intolerance (which is associated with impaired intestinal or colonic barrier function) and intolerance of fermentable oligosaccharides, disaccharides, monosaccharides, and polyols (FODMAPs). The latter's effects represent the interaction between the GI microbiome and the nutrients.

Steatorrheal Causes

Fat malabsorption may lead to greasy, foul-smelling, difficult-to-flush diarrhea often associated with weight loss and nutritional deficiencies due to concomitant malabsorption of amino acids and vitamins. Increased fecal output is caused by the osmotic effects of fatty acids, especially after bacterial hydroxylation, and, to a lesser extent, by the neutral fat. Quantitatively, steatorrhea is defined as stool fat exceeding the normal 7 g/d; rapid-transit diarrhea may result in fecal fat

up to 14 g/d; daily fecal fat averages 15–25 g with small-intestinal diseases and is often >32 g with pancreatic exocrine insufficiency. Intraluminal maldigestion, mucosal malabsorption, or lymphatic obstruction may produce steatorrhea.

INTRALUMINAL MALDIGESTION

This condition most commonly results from pancreatic exocrine insufficiency, which occurs when >90% of pancreatic secretory function is lost. *Chronic pancreatitis*, usually a sequel of ethanol abuse, most frequently causes pancreatic insufficiency. Other causes include *cystic fibrosis*, *pancreatic duct obstruction*, and, rarely, *somatostatinoma*. Bacterial overgrowth in the small intestine may deconjugate bile acids and alter micelle formation, impairing fat digestion; it occurs with stasis from a blind-loop, small-bowel diverticulum or dysmotility and is especially likely in the elderly. Finally, cirrhosis or biliary obstruction may lead to mild steatorrhea due to deficient intraluminal bile acid concentration.

MUCOSAL MALABSORPTION

Mucosal malabsorption occurs from a variety of enteropathies, but it most commonly occurs from *celiac disease*. This gluten-sensitive enteropathy affects all ages and is characterized by villous atrophy and crypt hyperplasia in the proximal small bowel and can present with fatty diarrhea associated with multiple nutritional deficiencies of varying severity. Celiac disease is much more frequent than previously thought; it affects ~1% of the population, frequently presents without steatorrhea, can mimic IBS, and has many other GI and extraintestinal manifestations. *Tropical sprue* may produce a similar histologic and clinical syndrome but occurs in residents of or travelers to tropical climates; abrupt onset and response to antibiotics suggest an infectious etiology. *Whipple's disease*, due to the bacillus *Tropheryma whipplei* and histiocytic infiltration of the small-bowel mucosa, is a less common cause of steatorrhea that most typically occurs in young or middle-aged men; it is frequently associated with arthralgias, fever, lymphadenopathy, and extreme fatigue, and it may affect the CNS and endocardium. A similar clinical and histologic picture results from *Mycobacterium avium-intracellulare* infection in patients with AIDS. *Abetalipoproteinemia* is a rare defect of chylomicron formation and fat malabsorption in children, associated with acanthocytic erythrocytes, ataxia, and retinitis pigmentosa. Several other conditions may cause mucosal malabsorption including infections, especially with protozoa such as *Giardia*, numerous medications (e.g., olmesartan, mycophenolate mofetil, colchicine, cholestyramine, neomycin), amyloidosis, and chronic ischemia.

POSTMUCOSAL LYMPHATIC OBSTRUCTION

The pathophysiology of this condition, which is due to the rare *congenital intestinal lymphangiectasia* or to *acquired lymphatic obstruction* secondary to trauma, tumor, cardiac disease or infection, leads to the unique constellation of fat malabsorption with enteric losses of protein (often causing edema) and lymphocytopenia. Carbohydrate and amino acid absorption are preserved.

Inflammatory Causes

Inflammatory diarrheas are generally accompanied by pain, fever, bleeding, or other manifestations of inflammation. The mechanism of diarrhea may not only be exudation but, depending on lesion site, may include fat malabsorption, disrupted fluid/electrolyte absorption, and hypersecretion or hypermotility from release of cytokines and other inflammatory mediators. The unifying feature on stool analysis is the presence of leukocytes or leukocyte-derived proteins such as calprotectin. With severe inflammation, exudative protein loss can lead to anasarca (generalized edema). Any middle-aged or older person with chronic inflammatory-type diarrhea, especially with blood, should be carefully evaluated to exclude a colorectal tumor.

IDIOPATHIC INFLAMMATORY BOWEL DISEASE

The illnesses in this category, which include *Crohn's disease* and *chronic ulcerative colitis*, are among the most common organic causes of chronic diarrhea in adults and range in severity from mild to fulminant and life-threatening. They may be associated with uveitis, polyarthralgias, cholestatic liver disease (primary sclerosing cholangitis), and skin lesions (erythema nodosum, pyoderma gangrenosum). *Microscopic colitis*, including both lymphocytic and *collagenous colitis*, is

an increasingly recognized cause of chronic watery diarrhea, especially in middle-aged women and those on NSAIDs, statins, proton pump inhibitors (PPIs), and selective serotonin reuptake inhibitors (SSRIs); biopsy of a normal-appearing colon is required for histologic diagnosis. It may coexist with symptoms suggesting IBS or with celiac sprue or drug-induced enteropathy. It typically responds well to anti-inflammatory drugs (e.g., bismuth), the opioid agonist loperamide, or to budesonide.

PRIMARY OR SECONDARY FORMS OF IMMUNODEFICIENCY

Immunodeficiency may lead to prolonged infectious diarrhea. With selective IgA deficiency or common variable *hypogammaglobulinemia*, diarrhea is particularly prevalent and often the result of giardiasis, bacterial overgrowth, or sprue.

EOSINOPHILIC GASTROENTERITIS

Eosinophil infiltration of the mucosa, muscularis, or serosa at any level of the GI tract may cause diarrhea, pain, vomiting, or ascites. Affected patients often have an atopic history, Charcot-Leyden crystals due to extruded eosinophil contents may be seen on microscopic inspection of stool, and peripheral eosinophilia is present in 50–75% of patients. While hypersensitivity to certain foods occurs in adults, true food allergy causing chronic diarrhea is rare.

OTHER CAUSES

Chronic inflammatory diarrhea may be caused by *radiation enterocolitis*, *chronic graft-versus-host disease*, *Behçet's syndrome*, and *Cronkhite-Canada syndrome*, among others.

Dysmotility Causes

Rapid transit may accompany many diarrheas as a secondary or contributing phenomenon, but primary dysmotility is an unusual etiology of true diarrhea. Stool features often suggest a secretory diarrhea, but mild steatorrhea of up to 14 g of fat per day can be produced by maldigestion from rapid transit alone. *Hyperthyroidism*, *carcinoid syndrome*, and certain drugs (e.g., prostaglandins, prokinetic agents) may produce hypermotility with resultant diarrhea. Primary visceral neuromyopathies or idiopathic acquired intestinal pseudoobstruction may lead to stasis with secondary bacterial overgrowth causing diarrhea. *Diabetic diarrhea*, often accompanied by peripheral and generalized autonomic neuropathies, may occur in part because of intestinal dysmotility.

The exceedingly common IBS (10% point prevalence, 1–2% per year incidence) is characterized by disturbed intestinal and colonic motor and sensory responses to various stimuli. Symptoms of stool frequency typically cease at night, alternate with periods of constipation, are accompanied by abdominal pain relieved with defecation, and rarely result in weight loss.

Factitial Causes

Factitial diarrhea accounts for up to 15% of unexplained diarrheas referred to tertiary care centers. Either as a form of *Munchausen syndrome* (deception or self-injury for secondary gain) or *eating disorders*, some patients covertly self-administer laxatives alone or in combination with other medications (e.g., diuretics) or surreptitiously add water or urine to stool sent for analysis. Such patients are typically women, often with histories of psychiatric illness, and disproportionately from careers in health care. Hypotension and hypokalemia are common co-presenting features. The evaluation of such patients may be difficult: contamination of the stool with water or urine is suggested by very low or high stool osmolarity, respectively. Such patients often deny this possibility when confronted, but they do benefit from psychiatric counseling when they acknowledge their behavior.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Chronic Diarrhea

The laboratory tools available to evaluate the very common problem of chronic diarrhea are extensive, and many are costly and invasive. As such, the diagnostic evaluation must be rationally directed by a careful history, including medications, and physical examination **(Fig. 42-3)**. When this strategy is unrevealing, simple triage tests are often warranted to direct the choice of more complex investigations (Fig. 42-3). The history, physical examination **(Table 42-4)**, and routine blood studies should attempt to characterize the mechanism of diarrhea, identify diagnostically helpful associations, and assess the patient's fluid/electrolyte and nutritional status. Patients should be questioned about the onset, duration, pattern, aggravating (especially diet) and relieving factors, and stool characteristics of their diarrhea. The presence or absence of fecal incontinence, fever, weight loss, pain, certain exposures (travel, medications, contacts with diarrhea), and common extraintestinal manifestations (skin changes, arthralgias, oral aphthous ulcers) should be noted. A family history of inflammatory bowel disease (IBD) or sprue may indicate those possibilities. Physical findings may offer clues such as a thyroid mass, wheezing, heart murmurs, edema, hepatomegaly, abdominal masses, lymphadenopathy, mucocutaneous abnormalities, perianal fistulas, or anal sphincter laxity. Peripheral blood leukocytosis, elevated sedimentation rate, or C-reactive protein suggests inflammation; anemia reflects blood loss or nutritional deficiencies; or eosinophilia may occur with parasitoses, neoplasia, collagen-vascular disease, allergy, or eosinophilic gastroenteritis. Blood chemistries may demonstrate electrolyte, hepatic, or other metabolic disturbances. Measuring IgA tissue transglutaminase antibodies may help detect celiac disease. Bile acid diarrhea is confirmed by a scintigraphic radiolabeled bile acid retention test; however, this is not available in many countries. Alternative approaches are a screening blood test (serum C4 or FGF-19), measurement of fecal bile acids, or a therapeutic trial with a bile acid sequestrant (e.g., cholestyramine or colesevelam).

A therapeutic trial is often appropriate, definitive, and highly cost-effective when a specific diagnosis is suggested on the initial physician encounter. For example, chronic watery diarrhea, which ceases with fasting in an otherwise healthy young adult, may justify a trial of a lactose-restricted diet; bloating and diarrhea persisting since a mountain backpacking trip may warrant a trial of metronidazole for likely giardiasis; and postprandial diarrhea persisting following resection of terminal ileum might be due to bile acid malabsorption and be treated with cholestyramine or colesevelam before further evaluation. Persistent symptoms require additional investigation.

Certain diagnoses may be suggested on the initial encounter (e.g., idiopathic IBD); however, additional focused evaluations may be necessary to confirm the diagnosis and characterize the severity or extent of disease so that treatment can be best guided. Patients suspected of having IBS should be initially evaluated with flexible sigmoidoscopy with colorectal biopsies to exclude IBD, or particularly microscopic colitis, which is clinically indistinguishable from IBS with diarrhea or functional diarrhea; those with normal findings might be reassured and, as indicated, treated empirically with antispasmodics, antidiarrheals, or antidepressants (e.g., tricyclic agents). Any patient who presents with chronic diarrhea and hematochezia should be evaluated with stool microbiologic studies and colonoscopy.

In an estimated two-thirds of cases, the cause for chronic diarrhea remains unclear after the initial encounter, and further testing is required. Quantitative stool collection and analyses can yield important objective data that may establish a diagnosis or characterize the type of diarrhea as a triage for focused additional studies (Fig. 42-3). If stool weight is >200 g/d, additional stool analyses should be performed that might include electrolyte concentration, pH, occult blood testing, leukocyte inspection (or leukocyte protein assay), fat quantitation, and laxative screens.

For secretory diarrheas (watery, normal osmotic gap), possible medication-related side effects or surreptitious laxative use should be reconsidered. Microbiologic studies should be done including fecal bacterial cultures (including media for *Aeromonas* and *Plesiomonas*), inspection for ova and parasites, and *Giardia* antigen assay (the most sensitive test for

giardiasis). Small-bowel bacterial overgrowth can be excluded by intestinal aspirates with quantitative cultures or with glucose or lactulose breath tests involving measurement of breath hydrogen, methane, or other metabolite. However, interpretation of these breath tests may be confounded by disturbances of intestinal transit. Upper endoscopy and colonoscopy with biopsies and small-bowel x-rays (formerly barium, but increasingly CT with enterography or magnetic resonance with enteroclysis) are helpful to rule out structural or occult inflammatory disease. When suggested by history or other findings, screens for peptide hormones should be pursued (e.g., serum gastrin, VIP, calcitonin, and thyroid hormone/thyroid-stimulating hormone, urinary 5-hydroxyindolacetic acid, histamine).

Further evaluation of osmotic diarrhea should include tests for lactose intolerance and magnesium ingestion, the two most common causes. Low fecal pH suggests carbohydrate malabsorption; lactose malabsorption can be confirmed by lactose breath testing or by a therapeutic trial with lactose exclusion and observation of the effect of lactose challenge (e.g., a liter of milk). Lactase determination on small-bowel biopsy is not generally available. If fecal magnesium or laxative levels are elevated, inadvertent or surreptitious ingestion should be considered and psychiatric help should be sought.

For those with proven fatty diarrhea, endoscopy with small-bowel biopsy (including aspiration for *Giardia* and quantitative cultures) should be performed; if this procedure is unrevealing, a small-bowel radiograph is often an appropriate next step. If small-bowel studies are negative or if pancreatic disease is suspected, pancreatic exocrine insufficiency should be excluded with direct tests, such as the secretin-cholecystokinin stimulation test or a variation that could be performed endoscopically. In general, indirect tests such as assay of fecal elastase or chymotrypsin activity or a bentiromide test have fallen out of favor because of low sensitivity and specificity.

Chronic inflammatory-type diarrheas should be suspected by the presence of blood or leukocytes in the stool. Such findings warrant stool cultures; inspection for ova and parasites; *C. difficile* toxin assay; colonoscopy with biopsies; and, if indicated, small-bowel contrast studies.

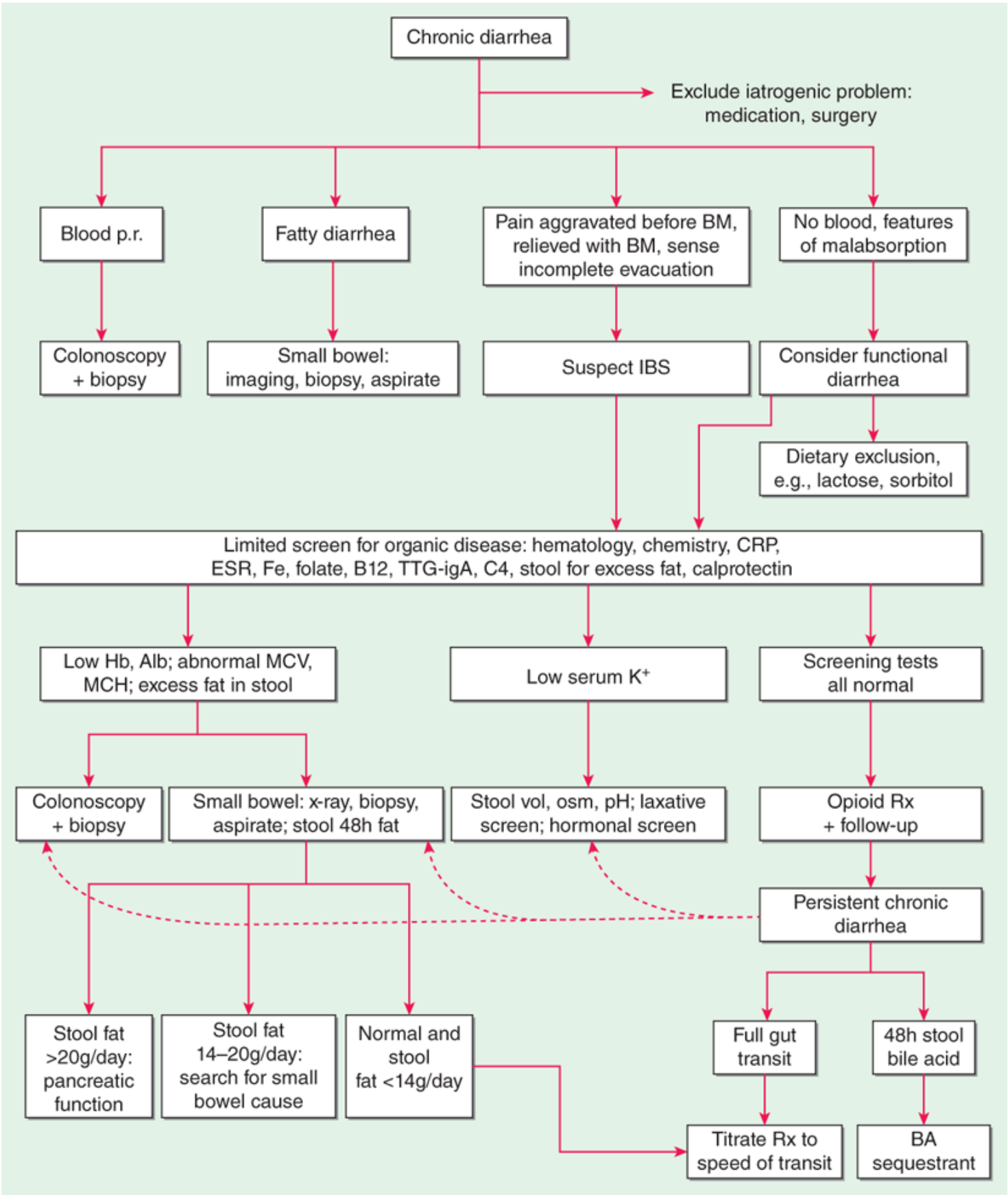
TABLE 42-4

Physical Examination in Patients with Chronic Diarrhea

1. Are there general features to suggest malabsorption or inflammatory bowel disease (IBD) such as anemia, dermatitis herpetiformis, edema, or clubbing?
2. Are there features to suggest underlying autonomic neuropathy or collagen-vascular disease in the pupils, orthostasis, skin, hands, or joints?
3. Is there an abdominal mass or tenderness?
4. Are there any abnormalities of rectal mucosa, rectal defects, or altered anal sphincter functions?
5. Are there any mucocutaneous manifestations of systemic disease such as dermatitis herpetiformis (celiac disease), erythema nodosum (ulcerative colitis), flushing (carcinoid), or oral ulcers for IBD or celiac disease?

FIGURE 42-3

Algorithm for management of chronic diarrhea. Patients undergo an initial evaluation based on different symptom presentations, leading to selection of patients for imaging, biopsy analysis, and limited screens for organic diseases. Alb, albumin; BA, bile acid; BM, bowel movement; C4, 7 α-hydroxy-4-cholesten-3-one; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; Hb, hemoglobin; Hx, history; IBS, irritable bowel syndrome; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; osm, osmolality; p.r., per rectum; Rx, treatment. *(Reprinted from M Camilleri, JH Sellin, KE Barrett: Pathophysiology, evaluation, and management of chronic watery diarrhea. Gastroenterology 152:515, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Chronic Diarrhea

Treatment of chronic diarrhea depends on the specific etiology and may be curative, suppressive, or empirical. If the cause can be eradicated, treatment is curative as with resection of a colorectal cancer, antibiotic administration for Whipple's disease or tropical sprue, or discontinuation of a drug. For many chronic conditions, diarrhea can be controlled by suppression of the underlying mechanism. Examples include elimination of dietary lactose for lactase deficiency or gluten for celiac sprue, use of glucocorticoids or other anti-inflammatory agents for idiopathic IBDs, bile acid sequestrants for bile acid malabsorption, PPIs for the gastric hypersecretion of gastrinomas, somatostatin analogues such as octreotide for malignant carcinoid syndrome, prostaglandin inhibitors such as indomethacin for medullary carcinoma of the thyroid, and pancreatic enzyme replacement for pancreatic insufficiency. When the specific cause or mechanism of chronic diarrhea evades diagnosis, empirical therapy may be beneficial. Mild opiates, such as diphenoxylate or loperamide, are often helpful in mild or moderate watery diarrhea. For those with more severe diarrhea, codeine or tincture of opium may be beneficial. Such antimotility agents should be avoided with severe IBD, because toxic megacolon may be precipitated. Clonidine, an α_2-adrenergic agonist, may allow control of diabetic diarrhea, although the medication may be poorly tolerated because it causes postural hypotension. The 5-HT_3 receptor antagonists (e.g., alosetron, ondansetron) may relieve diarrhea and urgency in patients with IBS diarrhea. Other medications approved for the treatment of diarrhea associated with IBS are the nonabsorbed antibiotic, rifaximin, and the mixed μ-opioid receptor (OR) and κ-OR agonist and δ-OR antagonist, eluxadoline. The latter may induce sphincter of Oddi spasm and subsequent acute pancreatitis, usually in patients with prior cholecystectomy. For all patients with chronic diarrhea, fluid and electrolyte repletion is an important component of management (see "Acute Diarrhea," earlier). Replacement of fat-soluble vitamins may also be necessary in patients with chronic steatorrhea.

CONSTIPATION

DEFINITION

Constipation is a common complaint in clinical practice and usually refers to persistent, difficult, infrequent, or seemingly incomplete defecation. Because of the wide range of normal bowel habits, constipation is difficult to define precisely. Most persons have at least three bowel movements per week; however, low stool frequency alone is not the sole criterion for the diagnosis of constipation. Many constipated patients have a normal frequency of defecation but complain of excessive straining, hard stools, lower abdominal fullness, or a sense of incomplete evacuation. The individual patient's symptoms must be analyzed in detail to ascertain what is meant by "constipation" or "difficulty" with defecation.

Stool form and consistency are well correlated with the time elapsed from the preceding defecation. Hard, pellety stools occur with slow transit, whereas loose, watery stools are associated with rapid transit. Both small pellety or very large stools are more difficult to expel than normal stools.

The perception of hard stools or excessive straining is more difficult to assess objectively, and the need for enemas or digital disimpaction is a clinically useful way to corroborate the patient's perceptions of difficult defecation.

Psychosocial or cultural factors may also be important. A person whose parents attached great importance to daily defecation will become greatly concerned when he or she misses a daily bowel movement; some children withhold stool to gain attention or because of fear of pain from anal irritation; and some adults habitually ignore or delay the call to have a bowel movement.

CAUSES

Pathophysiologically, chronic constipation generally results from inadequate fiber or fluid intake or from disordered colonic transit or anorectal function. These result from neurogastroenterologic disturbance, certain drugs, advancing age, or in association with a large number of systemic diseases that affect the GI tract (Table 42-5). Constipation of recent onset may be a symptom of significant organic disease such as tumor, anorectal irritation, or stricture. In *idiopathic constipation*, a subset of patients exhibits delayed emptying of the ascending and transverse colon with prolongation of transit (often in the proximal colon) and a reduced frequency of propulsive HAPCs. *Outlet obstruction to defecation* (also called *evacuation disorders*) accounts for about a quarter of cases presenting with constipation in tertiary care and may cause delayed colonic transit, which is usually corrected by biofeedback retraining of the disordered defecation. Constipation of any cause may be exacerbated by hospitalization or chronic illnesses that lead to physical or mental impairment and result in inactivity or physical immobility.

TABLE 42-5

Causes of Constipation in Adults

Types of Constipation and Causes	Examples
Recent Onset	
Colonic obstruction	Neoplasm; stricture: ischemic, diverticular, inflammatory
Anal sphincter spasm	Anal fissure, painful hemorrhoids
Medications	
Chronic	
Irritable bowel syndrome	Constipation-predominant, alternating
Medications	Ca^{2+} blockers, antidepressants
Colonic pseudoobstruction	Slow-transit constipation, megacolon (rare Hirschsprung's, Chagas' diseases)
Disorders of rectal evacuation	Pelvic floor dysfunction; anismus; descending perineum syndrome; rectal mucosal prolapse; rectocele
Endocrinopathies	Hypothyroidism, hypercalcemia, pregnancy
Psychiatric disorders	Depression, eating disorders, drugs
Neurologic disease	Parkinsonism, multiple sclerosis, spinal cord injury
Generalized muscle disease	Progressive systemic sclerosis

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Constipation

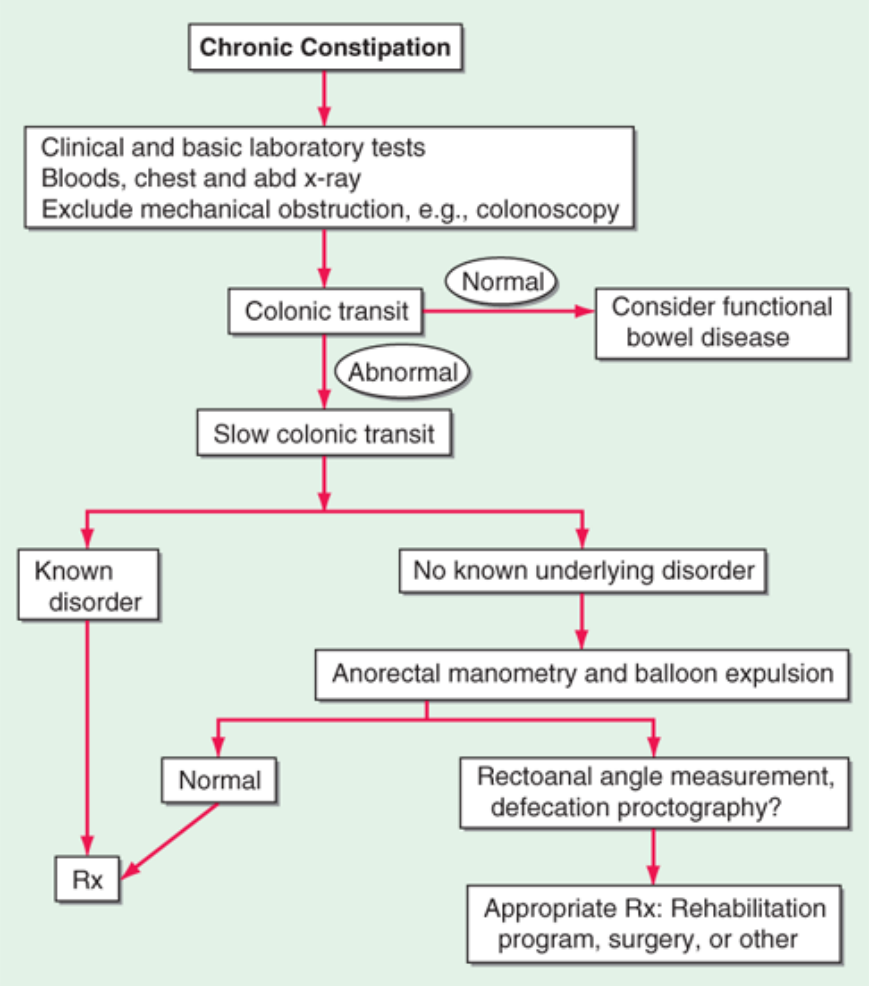
A careful history should explore the patient's symptoms and confirm whether he or she is indeed constipated based on frequency (e.g., fewer than three bowel movements per week), consistency (lumpy/hard), excessive straining, prolonged defecation time, or need to support the perineum or digitate the anorectum to facilitate stool evacuation. In the vast majority of cases (probably >90%), there is no underlying cause (e.g., cancer, depression, or hypothyroidism), and constipation responds to ample hydration, exercise, and supplementation of dietary fiber (15–25 g/d). A good diet and medication history and attention to psychosocial issues are key. Physical examination and, particularly, a rectal examination should exclude fecal impaction and most of the important diseases that present with constipation and possibly indicate features suggesting an evacuation disorder (e.g., high anal sphincter tone, failure of perineal descent, or paradoxical puborectalis contraction during straining to simulate stool evacuation).

The presence of weight loss, rectal bleeding, or anemia with constipation mandates either flexible sigmoidoscopy plus barium enema or colonoscopy alone, particularly in patients aged >40 years, to exclude structural diseases such as cancer or strictures. Colonoscopy alone is most cost-effective in this setting because it provides an opportunity to biopsy mucosal lesions, perform polypectomy, or dilate strictures. Barium enema has advantages over colonoscopy in the patient with isolated constipation because it is less costly and identifies colonic dilation and all significant mucosal lesions or strictures that are likely to present with constipation. Melanosis coli, or pigmentation of the colon mucosa, indicates the use of anthraquinone laxatives such as cascara or senna; however, this is usually apparent from a careful history. An unexpected disorder such as megacolon or cathartic colon may also be detected by colonic radiographs. Measurement of serum calcium, potassium, and thyroid-stimulating hormone levels will identify rare patients with metabolic disorders.

Patients with more troublesome constipation may not respond to fiber alone and may be helped by a bowel-training regimen, which involves taking an osmotic laxative (e.g., magnesium salts, lactulose, sorbitol, polyethylene glycol) and evacuating with enema or suppository (e.g., glycerin or bisacodyl) as needed. After breakfast, a distraction-free 15–20 min on the toilet without straining is encouraged. Excessive straining may lead to development of hemorrhoids and, if there is weakness of the pelvic floor or injury to the pudendal nerve, may result in obstructed defecation from descending perineum syndrome several years later. Those few who do not benefit from the simple measures delineated above or require long-term treatment or fail to respond to potent laxatives should undergo further investigation **(Fig. 42-4)**. Novel agents that induce secretion (e.g., lubiprostone, a chloride channel activator, or linaclotide, a guanylate cyclase C agonist that activates chloride secretion) are also available.

FIGURE 42-4

Algorithm for the management of constipation. abd, abdominal.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

INVESTIGATION OF SEVERE CONSTIPATION

A small minority (probably <5%) of patients have severe or "intractable" constipation; about 25% have evacuation disorders. These are the patients most likely to require evaluation by gastroenterologists or in referral centers. Further observation of the patient may occasionally reveal a previously unrecognized cause, such as an evacuation disorder, laxative abuse, malingering, or psychological disorder. In these patients, evaluations of the physiologic function of the colon and pelvic floor and of psychological status aid in the rational choice of treatment. Even among these highly selected patients with severe constipation, a cause can be identified in only about one-third of tertiary referral patients, with the others being diagnosed with normal transit constipation.

Measurement of Colonic Transit

Radiopaque marker transit tests are easy, repeatable, generally safe, inexpensive, reliable, and highly applicable in evaluating constipated patients in clinical practice. Several validated methods are very simple. For example, radiopaque markers are ingested; an abdominal flat film taken 5 days later should indicate passage of 80% of the markers out of the colon without the use of laxatives or enemas. This test does not provide useful information about the transit profile of the stomach and small bowel. An alternative approach involves ingestion of 24 radiopaque markers on 3 successive days and an abdominal radiograph on the fourth day. The number of markers counted in the radiograph is an estimate of the colonic transit in hours. The collection of gas in the rectum between the level of the ischial spines and the lower border of the sacroiliac joints may suggest the presence of a rectal evacuation disorder as the cause of constipation.

Radioscintigraphy with a delayed-release capsule containing radiolabeled particles has been used to noninvasively characterize normal, accelerated, or delayed colonic function over 24–48 h with low radiation exposure. This approach simultaneously assesses gastric, small bowel (which may be important in ~20% of patients with delayed colonic transit because they reflect a more generalized GI motility disorder), and colonic transit. The disadvantages are the greater cost and the need for specific materials prepared in a nuclear medicine laboratory.

Anorectal and Pelvic Floor Tests

Pelvic floor dysfunction is suggested by the inability to evacuate the rectum, a feeling of persistent rectal fullness, rectal pain, the need to extract stool from the rectum digitally, application of pressure on the posterior wall of the vagina, support of the perineum during straining, and excessive straining. These significant symptoms should be contrasted with the simple sense of incomplete rectal evacuation, which is common in IBS.

Formal psychological evaluation may identify eating disorders, “control issues,” depression, or posttraumatic stress disorders that may respond to cognitive or other intervention and may be important in restoring quality of life to patients who might present with chronic constipation.

A simple clinical test in the office to document a nonrelaxing puborectalis muscle is to have the patient strain to expel the index finger during a digital rectal examination. Motion of the puborectalis posteriorly during straining indicates proper coordination of the pelvic floor muscles. Motion anteriorly with paradoxical contraction or limited perineal descent (<1.5 cm) during simulated evacuation indicates pelvic floor dysfunction.

Measurement of perineal descent is relatively easy to gauge clinically by placing the patient in the left decubitus position and watching the perineum to detect inadequate descent (<1.5 cm, a sign of pelvic floor dysfunction) or perineal ballooning during straining relative to bony landmarks (>4 cm, suggesting excessive perineal descent).

A useful overall test of evacuation is the balloon expulsion test. A balloon-tipped urinary catheter is placed and inflated with 50 mL of water. Normally, a patient can expel it while seated on a toilet or in the left lateral decubitus position. In the lateral position, the weight needed to facilitate expulsion of the balloon is determined; normally, expulsion occurs with <200 g added or unaided within 1 minute.

Anorectal manometry, when used in the evaluation of patients with severe constipation, may find an excessively high resting (>80 mmHg) or squeeze anal sphincter tone, suggesting anismus (anal sphincter spasm). This test also identifies rare syndromes, such as adult Hirschsprung’s disease, by the absence of the rectoanal inhibitory reflex.

Defecography (a dynamic barium enema including lateral views obtained during barium expulsion or a magnetic resonance defecogram) reveals “soft abnormalities” in many patients; the most relevant findings are the measured changes in rectoanal angle, anatomic defects of the rectum such as internal mucosal prolapse, and enteroceles or rectoceles. Surgically remediable conditions are identified in only a few patients. These include severe, whole-thickness intussusception with complete outlet obstruction due to funnel-shaped plugging at the anal canal or an extremely large rectocele that fills preferentially during attempts at defecation instead of expulsion of the barium through the anus. In summary, defecography requires an interested and experienced radiologist, and abnormalities are not pathognomonic for pelvic floor dysfunction. The most common cause of outlet obstruction is failure of the puborectalis muscle to relax; this is not identified by barium defecography but can be demonstrated by magnetic resonance defecography, which provides more information about the structure and function of the pelvic floor, distal colorectum, and anal sphincters.

Neurologic testing (electromyography) is more helpful in the evaluation of patients with incontinence than of those with symptoms suggesting obstructed defecation. The absence of neurologic signs in the lower extremities suggests that any

documented denervation of the puborectalis results from pelvic (e.g., obstetric) injury or from stretching of the pudendal nerve by chronic, long-standing straining. Constipation is common among patients with spinal cord injuries, neurologic diseases such as Parkinson's disease, multiple sclerosis, and diabetic neuropathy.

Spinal-evoked responses during electrical rectal stimulation or stimulation of external anal sphincter contraction by applying magnetic stimulation over the lumbosacral cord identify patients with limited sacral neuropathies with sufficient residual nerve conduction to attempt biofeedback training.

In summary, a balloon expulsion test is an important screening test for anorectal dysfunction. Rarely, an anatomic evaluation of the rectum or anal sphincters and an assessment of pelvic floor relaxation are the tools for evaluating patients in whom obstructed defecation is suspected and is associated with symptoms of rectal mucosal prolapse, pressure of the posterior wall of the vagina to facilitate defecation (suggestive of anterior rectocele), or prior pelvic surgery that may be complicated by enterocele.

TREATMENT

TREATMENT

Constipation

After the cause of constipation is characterized, a treatment decision can be made. Slow-transit constipation requires aggressive medical or surgical treatment; anismus or pelvic floor dysfunction usually responds to biofeedback management (Fig. 42-4). The remaining ~60% of patients with constipation has normal colonic transit and can be treated symptomatically. Patients with spinal cord injuries or other neurologic disorders require a dedicated bowel regimen that often includes rectal stimulation, enema therapy, and carefully timed laxative therapy.

Patients with constipation are treated with bulk, osmotic, prokinetic, secretory, and stimulant laxatives including fiber, psyllium, milk of magnesia, lactulose, polyethylene glycol (colonic lavage solution), lubiprostone, linaclotide, and bisacodyl, or, in some countries, prucalopride, a 5-HT_4 agonist. If a 3- to 6-month trial of medical therapy fails, unassociated with obstructed defecation, the patients should be considered for laparoscopic colectomy with ileorectostomy; however, this should not be undertaken if there is continued evidence of an evacuation disorder or a generalized GI dysmotility. Referral to a specialized center for further tests of colonic motor function is warranted. The decision to resort to surgery is facilitated in the presence of megacolon and megarectum. The complications after surgery include small-bowel obstruction (11%) and fecal soiling, particularly at night during the first postoperative year. Frequency of defecation is 3–8 per day during the first year, dropping to 1–3 per day from the second year after surgery.

Patients who have a combined (evacuation and transit/motility) disorder should first pursue pelvic floor retraining (biofeedback and muscle relaxation), psychological counseling, and dietetic advice. If symptoms are intractable despite biofeedback and optimized medical therapy, colectomy and ileorectostomy could be considered as long as the evacuation disorder is resolved and optimized medical therapy is unsuccessful. In patients with pelvic floor dysfunction alone, biofeedback training has a 70–80% success rate, measured by the acquisition of comfortable stool habits. Attempts to manage pelvic floor dysfunction with operations (internal anal sphincter or puborectalis muscle division) or injections with botulinum toxin have achieved only mediocre success and have been largely abandoned.

FURTHER READING

Assi R et al: Sexually transmitted infections of the anus and rectum. World J Gastroenterol 20:15262, 2014. [PubMed: 25386074]

Bharucha AE, Rao SS: An update on anorectal disorders for gastroenterologists. Gastroenterology 146:37, 2014. [PubMed: 24211860]

Bharucha AE, Pemberton JH, Locke GR 3rd: American Gastroenterological Association technical review on constipation. Gastroenterology 144:218, 2013. [PubMed: 23261065]

Boeckxstaens G et al: Fundamentals of neurogastroenterology: Physiology/motility—sensation. Gastroenterology pii: S0016-5085(16)00221–3, 2016. doi: 10.1053/j.gastro.2016.02.030. [Epub ahead of print]

Camilleri M, Sellin JH, Barrett KE: Pathophysiology, evaluation, and management of chronic watery diarrhea. Gastroenterology 152:515, 2017. [PubMed: 27773805]

Lembo A, Camilleri M: Chronic constipation. N Engl J Med 349:1360, 2003. [PubMed: 14523145]

Riddle MS, DuPont HL, Connor BA: ACG Clinical Guideline: Diagnosis, treatment, and prevention of acute diarrheal infections in adults. Am J Gastroenterol 111:602, 2016. [PubMed: 27068718]

Rubio-Tapia A et al: American College of Gastroenterology. ACG clinical guidelines: Diagnosis and management of celiac disease. Am J Gastroenterol 108:656, 2013. [PubMed: 23609613]

Schiller LR, Pardi DS, Sellin JH: Chronic diarrhea: Diagnosis and management. Clin Gastroenterol Hepatol 15:182, 2017. [PubMed: 27496381]

Uzzan M et al: Gastrointestinal disorders associated with common variable immune deficiency (CVID) and chronic granulomatous disease (CGD). Curr Gastroenterol Rep 18:17, 2016. [PubMed: 26951230]

Chapter 43: Unintentional Weight Loss

J. Larry Jameson

INTRODUCTION

Unintentional weight loss (UWL) is frequently insidious and can have important implications, often serving as a harbinger of serious underlying disease. Clinically important weight loss is defined as the loss of 10 pounds (4.5 kg) or >5% of one's body weight over a period of 6–12 months. UWL is encountered in up to 8% of all adult outpatients and 27% of frail persons aged ≥65 years. There is no identifiable cause in up to one-quarter of patients despite extensive investigation. Conversely, up to half of people who claim to have lost weight have no documented evidence of weight loss. People with no known cause of weight loss generally have a better prognosis than do those with known causes, particularly when the source is neoplastic. Weight loss in older persons is associated with a variety of deleterious effects, including falls and fractures, pressure ulcers, impaired immune function, and decreased functional status. Not surprisingly, significant weight loss is associated with increased mortality, which can range from 9% to as high as 38% within 1–2.5 years in the absence of clinical awareness and attention.

PHYSIOLOGY OF WEIGHT REGULATION WITH AGING

(See also Chaps. 463 and 394) Among healthy aging people, total body weight peaks in the sixth decade of life and generally remains stable until the ninth decade, after which it gradually falls. In contrast, lean body mass (fat-free mass) begins to decline at a rate of 0.3 kg per year in the third decade, and the rate of decline increases further beginning at age 60 in men and age 65 in women. These changes in lean body mass largely reflect the age-dependent decline in growth hormone secretion and, consequently, circulating levels of insulin-like growth factor type I (IGF-I) that occur with normal aging. Loss of sex steroids, at menopause in women and more gradually with aging in men, also contributes to these changes in body composition. In the healthy elderly, an increase in fat tissue balances the loss in lean body mass until very old age, when loss of both fat and skeletal muscle occurs. Age-dependent changes also occur at the cellular level. Telomeres shorten, and body cell mass—the fat-free portion of cells—declines steadily with aging.

Between ages 20 and 80, mean energy intake is reduced by up to 1200 kcal/d in men and 800 kcal/d in women. Decreased hunger is a reflection of reduced physical activity and loss of lean body mass, producing lower demand for calories and food intake. Several important age-associated physiologic changes also predispose elderly persons to weight loss, such as declining chemosensory function (smell and taste), reduced efficiency of chewing, slowed gastric emptying, and alterations in the neuroendocrine axis,

including changes in levels of leptin, cholecystokinin, neuropeptide Y, and other hormones and peptides. These changes are associated with early satiety and a decline in both appetite and the hedonistic appreciation of food. Collectively, they contribute to the “anorexia of aging.” As noted below, these physiologic changes with aging may be accompanied by social isolation and/or poverty, further contributing to undernutrition.

CAUSES OF UNINTENTIONAL WEIGHT LOSS

Most causes of UWL belong to one of four categories: (1) malignant neoplasms, (2) chronic inflammatory or infectious diseases, (3) metabolic disorders (e.g., hyperthyroidism and diabetes), or (4) psychiatric disorders (Table 43-1). Not infrequently, more than one of these causes can be responsible for UWL. In most series, UWL is caused by malignant disease in a quarter of patients and by organic disease in one-third, with the remainder due to psychiatric disease, medications, or uncertain causes.

TABLE 43-1

Causes of Involuntary Weight Loss

Cancer	**Medications**
Colon	Sedatives
Hepatobiliary	Antibiotics
Hematologic	Nonsteroidal anti-inflammatory drugs
Lung	Serotonin reuptake inhibitors
Breast	Metformin
Genitourinary	Levodopa
Ovarian	Angiotensin-converting enzyme inhibitors
Prostate	Other drugs
Gastrointestinal disorders	**Disorders of the mouth and teeth**
Malabsorption	Caries
Peptic ulcer	Dysgeusia
Inflammatory bowel disease	**Age-related factors**
Pancreatitis	Physiologic changes
Obstruction/constipation	Visual impairment
Pernicious anemia	Decreased taste and smell
Endocrine and metabolic	Functional disabilities
Hyperthyroidism	**Neurologic**
Diabetes mellitus	Stroke
Pheochromocytoma	Parkinson’s disease
Adrenal insufficiency	Neuromuscular disorders
Cardiac disorders	Dementia
Chronic ischemia	**Social**
Chronic congestive heart failure	Isolation
Respiratory disorders	Economic hardship
Emphysema	**Psychiatric and behavioral**
Chronic obstructive pulmonary disease	Depression
Renal insufficiency	Anxiety
Rheumatologic disease	Paranoia
Infections	Bereavement
HIV	Alcoholism
Tuberculosis	Eating disorders
Parasitic infection	Increased activity or exercise
Subacute bacterial endocarditis	**Idiopathic**

The most common malignant causes of UWL are gastrointestinal, hepatobiliary, hematologic, lung, breast, genitourinary, ovarian, and prostate. Half of all patients with cancer lose some body weight; one-third lose more than 5% of their original body weight, and up to 20% of all cancer deaths are caused directly by cachexia (through immobility and/or cardiac/respiratory failure). The greatest incidence of weight loss is seen among patients with solid tumors. Malignancy that reveals itself through significant weight loss usually has a very poor prognosis.

In addition to malignancies, gastrointestinal causes are among the most prominent causes of UWL. Peptic ulcer disease, inflammatory bowel disease, dysmotility syndromes, chronic pancreatitis, celiac disease, constipation, and atrophic gastritis are some of the more common entities. Oral and dental problems are easily overlooked and may manifest with halitosis, poor oral hygiene, xerostomia, inability to chew, reduced masticatory force, nonocclusion, temporomandibular joint syndrome, edentulousness, and pain due to caries or abscesses.

Tuberculosis, fungal diseases, parasites, subacute bacterial endocarditis, and HIV are well-documented causes of UWL. Cardiovascular and pulmonary diseases cause UWL through increased metabolic demand and decreased appetite and caloric intake. Repeated surgeries may lead to weight loss because of reduced caloric intake and increased metabolic demands resulting from a systemic inflammatory response. Uremia produces nausea, anorexia, and vomiting. Connective tissue diseases may increase metabolic demand and disrupt nutritional balance. As the incidence of diabetes mellitus increases with aging, the associated glucosuria can contribute to weight loss. Hyperthyroidism in the elderly may have less prominent sympathomimetic features and may present as “apathetic hyperthyroidism” or T_3 toxicosis (Chap. 375).

Neurologic injuries such as stroke, quadriplegia, and multiple sclerosis may lead to visceral and autonomic dysfunction that can impair caloric intake. Dysphagia from these neurologic insults is a common mechanism. Functional disability that compromises activities of daily living (ADLs) is a common cause of undernutrition in the elderly. Visual impairment from ophthalmic or central nervous system disorders such as a tremor can limit the ability of people to prepare and eat meals. UWL may be one of the earliest manifestations of Alzheimer’s dementia.

Isolation and depression are significant causes of UWL that may manifest as an inability to care for oneself, including nutritional needs. A cytokine-mediated inflammatory metabolic cascade can be both a cause of and a manifestation of depression. Bereavement can be a cause of UWL and, when present, is often more pronounced in men. More intense forms of mental illness such as paranoid disorders may lead to delusions about food and cause weight loss. Alcoholism can be a significant source of weight loss and malnutrition.

Elderly persons living in poverty may have to choose whether to purchase food or use the money for other expenses, including medications. Institutionalization is an independent risk factor, as up to 30–50% of nursing home patients have inadequate food intake.

Medications can cause anorexia, nausea, vomiting, gastrointestinal distress, diarrhea, dry mouth, and changes in taste. This is particularly an issue in the elderly, many of whom take five or more medications.

ASSESSMENT

The four major manifestations of UWL are (1) anorexia (loss of appetite), (2) sarcopenia (loss of muscle mass), (3) cachexia (a syndrome that combines weight loss, loss of muscle and adipose tissue, anorexia, and weakness), and (4) dehydration. The current obesity epidemic adds complexity, as excess adipose tissue can mask the development of sarcopenia and delay awareness of the development of cachexia. If it is not possible to measure weight directly, a change in clothing size, corroboration of weight loss by a relative or friend, and a numeric estimate of weight loss provided by the patient are suggestive of true weight loss.

Initial assessment includes a comprehensive history and physical, a complete blood count, tests of liver enzyme levels, C-reactive protein, erythrocyte sedimentation rate, renal function studies, thyroid function tests, chest radiography, and an abdominal ultrasound **(Table 43-2)**. Age, sex, and risk factor–specific cancer screening tests, such as mammography and colonoscopy, should be performed **(Chap. 66)**. Patients at risk should have HIV testing. All elderly patients with weight loss should undergo screening for dementia and depression by using instruments such as the Mini-Mental State Examination and the Geriatric Depression Scale, respectively **(Chap. 464)**. The Mini Nutritional Assessment (*www.mna-elderly.com*) and the Nutrition Screening Initiative (*http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/pmc/articles/PMC1694757/*) are also available for the nutritional assessment of elderly patients. Almost all patients with a malignancy and >90% of those with other organic diseases have at least one laboratory abnormality. In patients presenting with substantial UWL, major organic and malignant diseases are unlikely when a baseline evaluation is completely normal. Careful follow-up rather than undirected testing is advised since the prognosis of weight loss of undetermined cause is generally favorable.

TABLE 43-2

Assessment and Testing for Involuntary Weight Loss

Indications	**Laboratory**
5% weight loss in 30 d	Complete blood count
10% weight loss in 180 d	Comprehensive electrolyte and metabolic panel, including liver and renal function tests
Body mass index <21	Thyroid function tests
25% of food left uneaten after 7 d	Erythrocyte sedimentation rate
Change in fit of clothing	C-reactive protein
Change in appetite, smell, or taste	Ferritin
Abdominal pain, nausea, vomiting, diarrhea, constipation, dysphagia	HIV testing, if indicated
Assessment	**Radiology**
Complete physical examination, including dental evaluation	Chest x-ray Abdominal ultrasound
Medication review	
Recommended cancer screening	
Mini-Mental State Examination[a]	
Mini-Nutritional Assessment[a]	
Nutrition Screening Initiative[a]	
Simplified Nutritional Assessment Questionnaire[a]	
Observation of eating[a]	

Indications	Laboratory
Activities of daily living[a]	
Instrumental activities of daily living[a]	

[a]May be more specific to assess weight loss in the elderly.

TREATMENT

TREATMENT

Unintentional Weight Loss

The first priority in managing weight loss is to identify and treat the underlying causes. Treatment of underlying metabolic, psychiatric, infectious, or other systemic disorders may be sufficient to restore weight and functional status gradually. Medications that cause nausea or anorexia should be withdrawn or changed, if possible. For those with unexplained UWL, oral nutritional supplements such as high-energy drinks sometimes reverse weight loss. Advising patients to consume supplements between meals rather than with a meal may help minimize appetite suppression and facilitate increased overall intake. Orexigenic, anabolic, and anticytokine agents are under investigation. In selected patients, the antidepressant mirtazapine results in a significant increase in body weight, body fat mass, and leptin concentration. Patients with wasting conditions who can comply with an appropriate exercise program gain muscle protein mass, strength, and endurance and may be more capable of performing ADLs.

ACKNOWLEDGMENT

The author is grateful to Russell G. Robertson, MD for contributions to this chapter in prior editions.

FURTHER READING

Alibhai SM et al: An approach to the management of unintentional weight loss in elderly people. CMAJ 172:773, 2005. [PubMed: 15767612]

Compston JE et al: Increase in fracture risk following unintentional weight loss in postmenopausal women: The global longitudinal study of osteoporosis in women. J Bone Miner Res 31:1466, 2016. [PubMed: 26861139]

Gaddey HL, Holder K: Unintentional weight loss in older adults. Am Fam Physician 89:718, 2014. [PubMed: 24784334]

McMinn J et al: Investigation and management of unintentional weight loss in older adults. BMJ 342:d1732, 2011. [PubMed: 21447571]

Miller SL, Wolfe RR: The danger of weight loss in the elderly. J Nutr Health and Aging 12:487, 2008.

Vanderschueren S et al: The diagnostic spectrum of unintentional weight loss. Eur J Intern Med 16:160, 2005. [PubMed: 15967329]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 44: Gastrointestinal Bleeding

Loren Laine

INTRODUCTION

Gastrointestinal bleeding (GIB) is the most common gastrointestinal condition leading to hospitalization in the United States, accounting for over 507,000 admissions and $4.85 billion in direct costs annually. Upper GIB (UGIB) incidence has decreased in recent decades, primarily due to decreases in GIB from ulcers. The ratio of UGIB to lower GIB (LGIB) among GIB admissions from U.S. emergency rooms is ~1.3. The case fatality of patients hospitalized with GIB has also decreased and is <3% in the United States. Patients generally die from decompensation of other underlying illnesses rather than exsanguination.

GIB presents as either overt or occult bleeding. *Overt GIB* is manifested by *hematemesis*, vomitus of red blood or "coffee-grounds" material; *melena*, black, tarry stool; and/or *hematochezia*, passage of red or maroon blood from the rectum. In the absence of overt bleeding, *occult GIB* may present with *symptoms of blood loss or anemia* such as lightheadedness, syncope, angina, or dyspnea; or with iron-deficiency anemia or a positive fecal occult blood test on routine testing. GIB is also categorized by the site of bleeding as UGIB (esophagus, stomach, duodenum), LGIB (colonic), small intestinal, or obscure GIB (if the source is unclear).

SOURCES OF GASTROINTESTINAL BLEEDING

Upper Gastrointestinal Sources of Bleeding

PEPTIC ULCERS

Peptic ulcers are the most common cause of UGIB, accounting for ~50% of UGIB hospitalizations. Features of an ulcer at endoscopy provide important prognostic information that guides subsequent management decisions as outlined in **Figs. 315-3** and **315-4**. Approximately 20% of patients with bleeding ulcers have the highest risk findings of active bleeding or a nonbleeding visible vessel: one-third of such patients have further bleeding that requires urgent surgery if they are treated conservatively. These patients benefit from endoscopic therapy with bipolar electrocoagulation, heater probe, injection therapy (e.g., absolute alcohol, 1:10,000 epinephrine), and/or clips with reductions in bleeding, hospital stay, mortality, and costs. In contrast, patients with clean-based ulcers have rates of serious recurrent bleeding approaching zero. If stable with no other reason for hospitalization, such patients may be discharged home after endoscopy.

Randomized controlled trials document that high-dose, constant-infusion IV proton pump inhibitor (PPI) (80-mg bolus and 8-mg/h infusion), designed to sustain intragastric pH >6 and enhance clot stability, decreases further bleeding and mortality in patients with high-risk ulcers (active bleeding, nonbleeding visible vessel, adherent clot) when given after endoscopic therapy. Recent meta-analysis of randomized trials documents that high-dose intermittent PPIs are non-inferior to constant-infusion PPI therapy and thus may be substituted in this population. Patients with lower-risk findings (flat pigmented spot or clean base) do not require endoscopic therapy and receive standard doses of oral PPI.

Approximately 10–50% of patients with bleeding ulcers will rebleed within the next year if no preventive strategies are employed. Prevention of recurrent bleeding focuses on the three main factors in ulcer pathogenesis, *Helicobacter pylori*, nonsteroidal anti-inflammatory drugs (NSAIDs), and acid. Eradication of *H. pylori* in patients with bleeding ulcers decreases rebleeding rates to <5%. If a bleeding ulcer develops in a patient taking NSAIDs, the NSAIDs should be discontinued. If NSAIDs must be given, a cyclooxygenase (COX)-2 selective NSAID plus a PPI is recommended, based on results of a randomized trial. Patients with established cardiovascular disease who develop bleeding ulcers while taking low-dose aspirin for secondary prevention should restart aspirin as soon as possible after their bleeding episode (1–7 days). A randomized trial showed that failure to restart aspirin was associated with no significant difference in rebleeding (5% vs 10%) at 30 days but a significant increase in mortality (9% vs 1%) compared with immediate reinstitution of aspirin. In contrast, aspirin probably should be discontinued in most patients taking aspirin for primary prevention of cardiovascular events who develop UGIB. Patients with bleeding ulcers unrelated to *H. pylori* or NSAIDs should remain on PPI therapy indefinitely given a 42% incidence of rebleeding at 7 years without protective therapy. **Peptic ulcers are discussed in Chap. 317.**

MALLORY-WEISS TEARS

Mallory-Weiss tears account for ~2–10% of UGIB hospitalizations. The classic history is vomiting, retching, or coughing preceding hematemesis, especially in an alcoholic patient. Bleeding from these tears, which are usually on the gastric side of the gastroesophageal junction, stops spontaneously in 80–90% of patients and recurs in only 0–10%. Endoscopic therapy is indicated for actively bleeding Mallory-Weiss tears. **Mallory-Weiss tears are discussed in Chap. 316.**

ESOPHAGEAL VARICES

The proportion of UGIB hospitalizations due to varices varies widely, from ~2–40%, depending on the population. Patients with variceal hemorrhage have poorer outcomes than patients with other sources of UGIB. Urgent endoscopy within 12 h is recommended in cirrhotics with UGIB, and if esophageal varices are present, endoscopic ligation is performed and an IV vasoactive medication (octreotide, somatostatin, vapreotide, terlipressin) is given for 2–5 days. Combination of endoscopic and medical therapy is superior to either therapy alone in decreasing rebleeding. Over the long term, treatment with nonselective beta blockers plus endoscopic ligation is recommended because the combination is more effective than either alone in reduction of recurrent esophageal variceal bleeding. Transjugular intrahepatic portosystemic shunt (TIPS) is recommended in patients who have persistent or recurrent bleeding despite endoscopic and medical therapy. TIPS should also be considered in the first 1–2 days of hospitalization for acute variceal bleeding in patients with advanced liver disease (e.g., Child-Pugh class C with Child-Pugh score 10–13), because randomized trials show significant decreases in rebleeding and mortality compared with standard endoscopic and medical therapy.

Portal hypertension is also responsible for bleeding from gastric varices, varices in the small and large intestine, and portal hypertensive gastropathy and enterocolopathy. Bleeding gastric varices due to cirrhosis are treated with endoscopic injection of tissue adhesive (e.g., *n*-butyl cyanoacrylate), if available; if not, TIPS is performed.

EROSIVE DISEASE

Erosions are endoscopically visualized breaks which are confined to the mucosa and do not cause major bleeding due to the absence of arteries and veins in the mucosa. Erosions in the esophagus, stomach, or duodenum commonly cause mild UGIB, with erosive gastritis and duodenitis accounting for perhaps ~10–15% and erosive esophagitis (primarily due to gastroesophageal reflux disease) ~1–10% of UGIB hospitalizations. The most important cause of gastric and duodenal erosions is NSAID use: ~50% of patients who chronically ingest NSAIDs may have gastric erosions. Other potential causes of gastric erosions include alcohol intake, *H. pylori* infection, and stress-related mucosal injury.

Stress-related gastric mucosal injury occurs only in extremely sick patients, such as those who have experienced serious trauma, major surgery, burns covering more than one-third of the body surface area, major intracranial disease, or severe medical illness (i.e., ventilator dependence, coagulopathy). Severe bleeding should not develop unless ulceration occurs. The mortality rate in these patients is high because of their serious underlying illnesses.

The incidence of bleeding from stress-related gastric mucosal injury has decreased dramatically in recent years, most likely due to better care of critically ill patients. Pharmacologic prophylaxis for bleeding may be considered in the high-risk patients mentioned above. Meta-analyses of randomized trials indicate that PPIs are more effective than H_2-receptor antagonists in reduction of overt and clinically important UGIB without differences in mortality or nosocomial pneumonia.

OTHER CAUSES

Less common causes of UGIB include neoplasms, vascular ectasias (including hereditary hemorrhagic telangiectasias [Osler-Weber-Rendu] and gastric antral vascular ectasia ["watermelon stomach"]), Dieulafoy's lesion (in which an aberrant vessel in the mucosa bleeds from a pinpoint mucosal defect), prolapse gastropathy (prolapse of proximal stomach into esophagus with retching, especially in alcoholics), aortoenteric fistulas, and hemobilia or hemosuccus pancreaticus (bleeding from the bile duct or pancreatic duct).

Small-Intestinal Sources of Bleeding

Patients without a source of GIB identified on upper endoscopy and colonoscopy were previously labeled as having obscure GIB. With the advent of improved diagnostic modalities, ~75% of GIB previously labeled obscure is now estimated to originate in the small intestine beyond the extent of a standard upper endoscopic exam. Small-intestinal GIB may account for up to ~5–10% of GIB cases. The most common causes in adults >40 years are vascular ectasias, neoplasm (e.g., GI stromal tumor, carcinoid, adenocarcinoma, lymphoma, metastases), and NSAID-induced erosions and ulcers. Meckel's diverticulum is the most common cause of significant small-intestinal GIB in children, decreasing in frequency as a cause of bleeding with age. Other causes in patients <40 years include Crohn's disease, polyposis syndromes, or neoplasm. Less common causes of small-intestinal GIB include infection, ischemia, vasculitis, small-bowel varices, diverticula, intussusception, Dieulafoy's lesions, aortoenteric fistulas, and duplication cysts.

Small-intestinal vascular ectasias are treated with endoscopic therapy if possible based on observational studies suggesting initial efficacy. However, rebleeding is common: 45% over a mean follow-up of 26 months in a recent systematic review. Estrogen/progesterone compounds are not recommended because a multicenter double-blind trial found no benefit in prevention of recurrent bleeding. Octreotide is used, based on positive results from case series but no randomized trials. A randomized trial reported significant benefit of thalidomide and awaits further confirmation. Other isolated lesions, such as tumors, generally require surgical resection.

Colonic Sources of Bleeding

Hemorrhoids are probably the most common cause of LGIB; anal fissures also cause minor bleeding and pain. If these local anal processes, which rarely require hospitalization, are excluded, the most common cause of LGIB in adults is diverticulosis, followed by vascular ectasias (especially in the proximal colon of patients >70 years), neoplasms (primarily adenocarcinoma), colitis (ischemic, infectious, Crohn's or ulcerative colitis, NSAID-induced colitis or ulcers), postpolypectomy bleeding, and radiation proctopathy. Rarer causes include solitary rectal ulcer syndrome, trauma, varices (most commonly rectal),

lymphoid nodular hyperplasia, vasculitis, and aortocolic fistulas. In children and adolescents, the most common colonic causes of significant GIB are inflammatory bowel disease and juvenile polyps.

Diverticular bleeding is abrupt in onset, usually painless, sometimes massive, and often from the right colon; chronic or occult bleeding is not characteristic. Colonic diverticula stop bleeding spontaneously in ~80–90% of patients and, on long-term follow-up, rebleed in ~15–40% of patients. Case series suggest endoscopic therapy may decrease recurrent bleeding in the uncommon case when colonoscopy identifies the specific bleeding diverticulum. When diverticular bleeding is found at angiography, transcatheter arterial embolization by superselective technique stops bleeding in a majority of patients. Segmental surgical resection is recommended for persistent or refractory diverticular bleeding.

Bleeding from colonic vascular ectasias may be overt or occult; it tends to be chronic and only occasionally is hemodynamically significant. Endoscopic hemostatic therapy may be used in the treatment of vascular ectasias, as well as discrete bleeding ulcers and postpolypectomy bleeding. Transcatheter arterial embolization also may be attempted for persistent bleeding from vascular ectasias and other discrete lesions. Surgical therapy is generally required for major persistent or recurrent bleeding from colonic sources that cannot be treated medically, endoscopically, or angiographically. Patients with Heyde's syndrome (bleeding vascular ectasias and aortic stenosis) appear to benefit from aortic valve replacement.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Gastrointestinal Bleeding

INITIAL ASSESSMENT

Measurement of the heart rate and blood pressure is the best way to initially assess a patient with GIB. Clinically significant bleeding leads to postural changes in heart rate or blood pressure, tachycardia, and, finally, recumbent hypotension. In contrast, hemoglobin does not fall immediately with acute GIB, due to proportionate reductions in plasma and red cell volumes (people bleed whole blood). Thus, hemoglobin may be normal or only minimally decreased at the initial presentation of a severe bleeding episode. As extravascular fluid enters the vascular space to restore volume, the hemoglobin falls, but this process may take up to 72 h. Transfusion is recommended when the hemoglobin drops below 7 g/dL, based on a large randomized trial showing this restrictive transfusion strategy decreases rebleeding and death in acute UGIB compared with a transfusion threshold of 9 g/dL. Patients with slow, chronic GIB may have very low hemoglobin values despite normal blood pressure and heart rate. With the development of iron-deficiency anemia, the mean corpuscular volume is low and red blood cell distribution width is increased.

DIFFERENTIATION OF UGIB FROM LGIB

Hematemesis indicates an UGIB source. Melena indicates blood has been present in the GI tract for ≥14 h, and as long as 3–5 days. The more proximal the bleeding site, the more likely melena will occur. Hematochezia usually represents a lower GI source of bleeding, although an upper GI lesion may bleed so briskly that blood transits the bowel before melena develops. When hematochezia is the presenting symptom of UGIB, it is associated with hemodynamic instability and dropping hemoglobin. Bleeding lesions of the small bowel may present as melena or hematochezia. Other clues to UGIB include hyperactive bowel sounds and an elevated blood urea nitrogen (due to volume depletion and blood proteins absorbed in the small intestine).

A nonbloody nasogastric aspirate may be seen in ~15% of patients with UGIB who present with clinically serious hematochezia. A bile-stained appearance does not exclude UGIB because reports of bile in the aspirate are incorrect in ~50% of cases. Testing of aspirates that are not grossly bloody for occult blood is not useful.

EVALUATION AND MANAGEMENT OF UGIB

Baseline characteristics predictive of rebleeding and death include hemodynamic compromise (tachycardia or hypotension), increasing age, and comorbidities **(Fig. 44-1)**. Risk assessment tools may be used to identify patients with very low risk. Discharge from the emergency room with outpatient management has been suggested for patients with a Glasgow-Blatchford score (possible range 0–23, **Table 44-1**) of 0–1 or 0–2 among patients <70 years because when hospitalized <1% of such patients require intervention and <0.5% die.

PPI infusion may be considered at presentation: it decreases high-risk ulcer stigmata (e.g., active bleeding) and need for endoscopic therapy but does not improve clinical outcomes such as further bleeding, surgery, or death. The promotility agent erythromycin, 250 mg intravenously ~30 min before endoscopy, also may be considered to improve visualization at endoscopy: it provides a small but significant increase in diagnostic yield and decrease in red cell transfusions. Cirrhotic patients presenting with UGIB should be given an antibiotic (quinolone or ceftriaxone) and IV vasoactive medication upon presentation, even before endoscopy. Antibiotics decrease bacterial infections, rebleeding, and mortality, and vasoactive medications may improve control of bleeding in the first 12 h after presentation.

Upper endoscopy should be performed within 24 h in most patients with UGIB. Patients at higher risk (e.g., hemodynamic instability, cirrhosis) may benefit from more urgent endoscopy within 12 h. Early endoscopy is also beneficial in low-risk patients for management decisions (e.g., discharge). Patients with major bleeding and high-risk endoscopic findings (e.g., varices, ulcers with active bleeding or a visible vessel) benefit from endoscopic hemostatic therapy, whereas patients with low-risk lesions (e.g., clean-based ulcers, erosions, nonbleeding Mallory-Weiss tears) who have stable vital signs and hemoglobin and no other medical problems may be discharged home.

EVALUATION AND MANAGEMENT OF LGIB

Patients with hematochezia and hemodynamic instability should have upper endoscopy to rule out an upper GI source before evaluation of the lower GI tract (Fig. 44-2).

Colonoscopy after an oral lavage solution is the procedure of choice in most patients admitted with LGIB unless bleeding is too massive, in which case angiography is recommended. Computed tomography (CT) angiography is often suggested prior to angiography to document evidence and location of active bleeding. Sigmoidoscopy is used primarily in patients <40 years old with minor bleeding. In patients with no source identified on colonoscopy, imaging studies may be employed. ^{99m}Tc-labeled red cell scan allows repeated imaging for up to 24 h and may identify the general location of bleeding. However, radionuclide scans should be interpreted with caution because results, especially from later images, are highly variable. Multidetector CT angiography is likely superior to nuclear scintigraphy and increasingly used in its place. In active LGIB, angiography can detect the site of bleeding (extravasation of contrast into the gut) and permits treatment with embolization.

EVALUATION AND MANAGEMENT OF SMALL-INTESTINAL OR OBSCURE GIB

In patients with massive bleeding suspected to be from the small intestine, current guidelines suggest angiography as the initial test, with CT angiography or ^{99m}Tc-labeled red cell scan prior to angiography if the patient's clinical status permits. For others, repeat upper and lower endoscopy may be considered as the initial evaluation because second-look procedures identify a source in up to ~25% of upper endoscopies and colonoscopies; a push enteroscopy, usually performed with a pediatric colonoscope to inspect the entire duodenum and proximal jejunum, may be substituted for a repeat standard upper endoscopy. If second-look procedures are negative, evaluation of the entire small intestine is performed, usually with video capsule endoscopy. A systematic review of comparative studies showed the yield of "clinically significant findings" greater with capsule than push enteroscopy (56% vs 26%) or small bowel barium radiography (42% vs 6%). However, capsule endoscopy does not allow full visualization of the small intestine, tissue sampling, or application of therapy.

CT enterography may be used initially instead of video capsule in patients with possible small bowel narrowing (e.g., stricture, prior surgery or radiation, Crohn's disease) and may follow a negative video capsule for suspected small-intestinal GIB, given its higher sensitivity for small-intestinal masses.

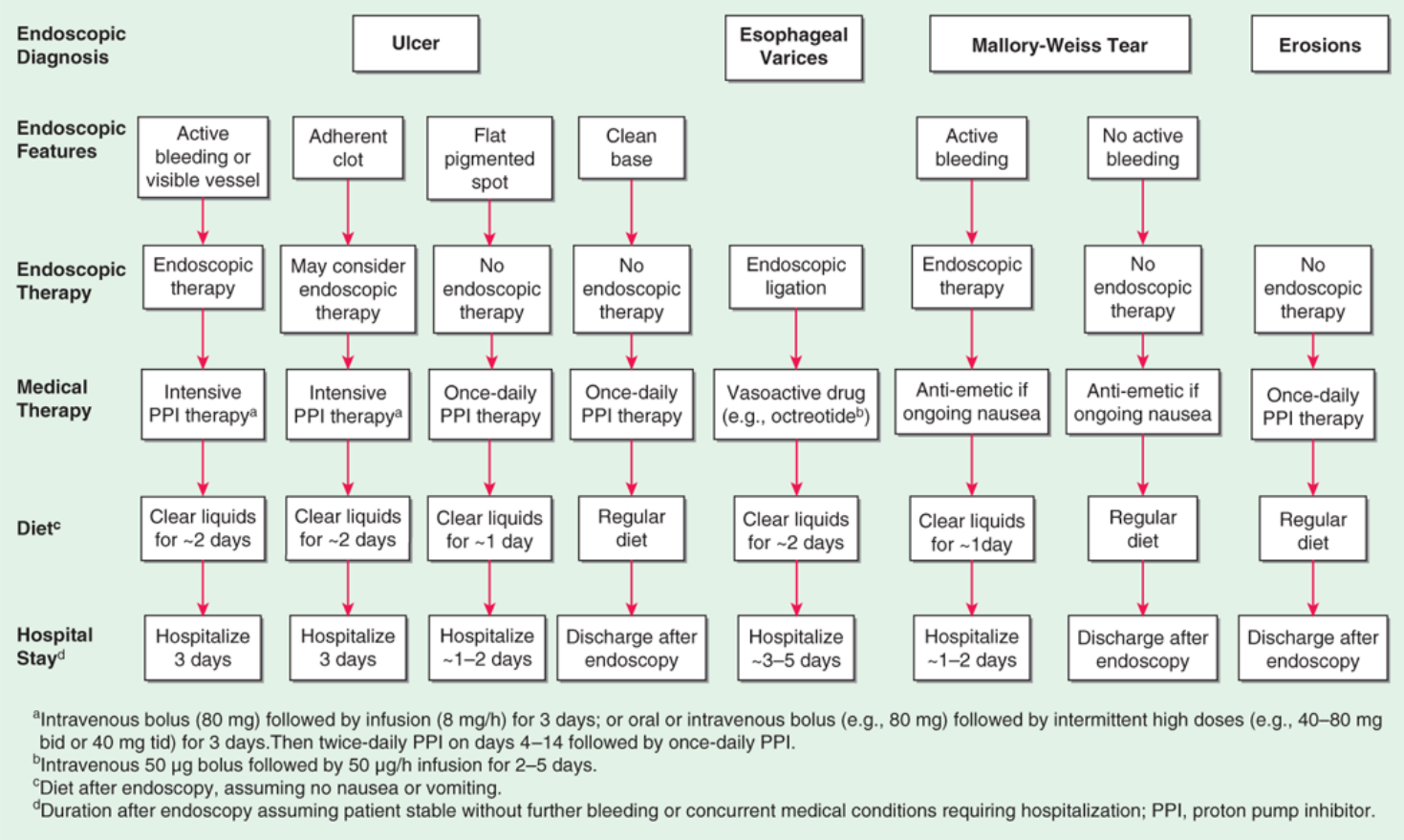
If capsule endoscopy is positive, management is dictated by the finding. If capsule endoscopy is negative, current recommendations suggest patients may either be observed or, if their clinical course mandates (e.g., need for transfusions), undergo further testing. "Deep" enteroscopy (double-balloon, single-balloon, or spiral enteroscopy) is commonly the next test undertaken for clinically important GIB documented or suspected to be from the small intestine because it allows the endoscopist to examine, obtain specimens from, and provide therapy to much or all of the small intestine. Other imaging techniques sometimes used in evaluation of obscure GIB include ^{99m}Tc-labeled red blood cell scintigraphy, CT angiography, angiography, and ^{99m}Tc-pertechnetate scintigraphy for Meckel's diverticulum (especially in young patients). If all tests are unrevealing, intraoperative endoscopy is indicated in patients with severe recurrent or persistent bleeding requiring repeated transfusions.

POSITIVE FECAL OCCULT BLOOD TEST

Fecal occult blood testing is recommended only for colorectal cancer screening, beginning at age 50 in average-risk adults. A positive test necessitates colonoscopy. If evaluation of the colon is negative, further workup is not recommended unless iron-deficiency anemia or GI symptoms are present.

FIGURE 44-1

Suggested algorithm for patients with acute upper gastrointestinal bleeding based on endoscopic findings.



[a]Intravenous bolus (80 mg) followed by infusion (8 mg/h) for 3 days; or oral or intravenous bolus (e.g., 80 mg) followed by intermittent high doses (e.g., 40–80 mg bid or 40 mg tid) for 3 days.Then twice-daily PPI on days 4–14 followed by once-daily PPI.
[b]Intravenous 50 µg bolus followed by 50 µg/h infusion for 2–5 days.
[c]Diet after endoscopy, assuming no nausea or vomiting.
[d]Duration after endoscopy assuming patient stable without further bleeding or concurrent medical conditions requiring hospitalization; PPI, proton pump inhibitor.

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

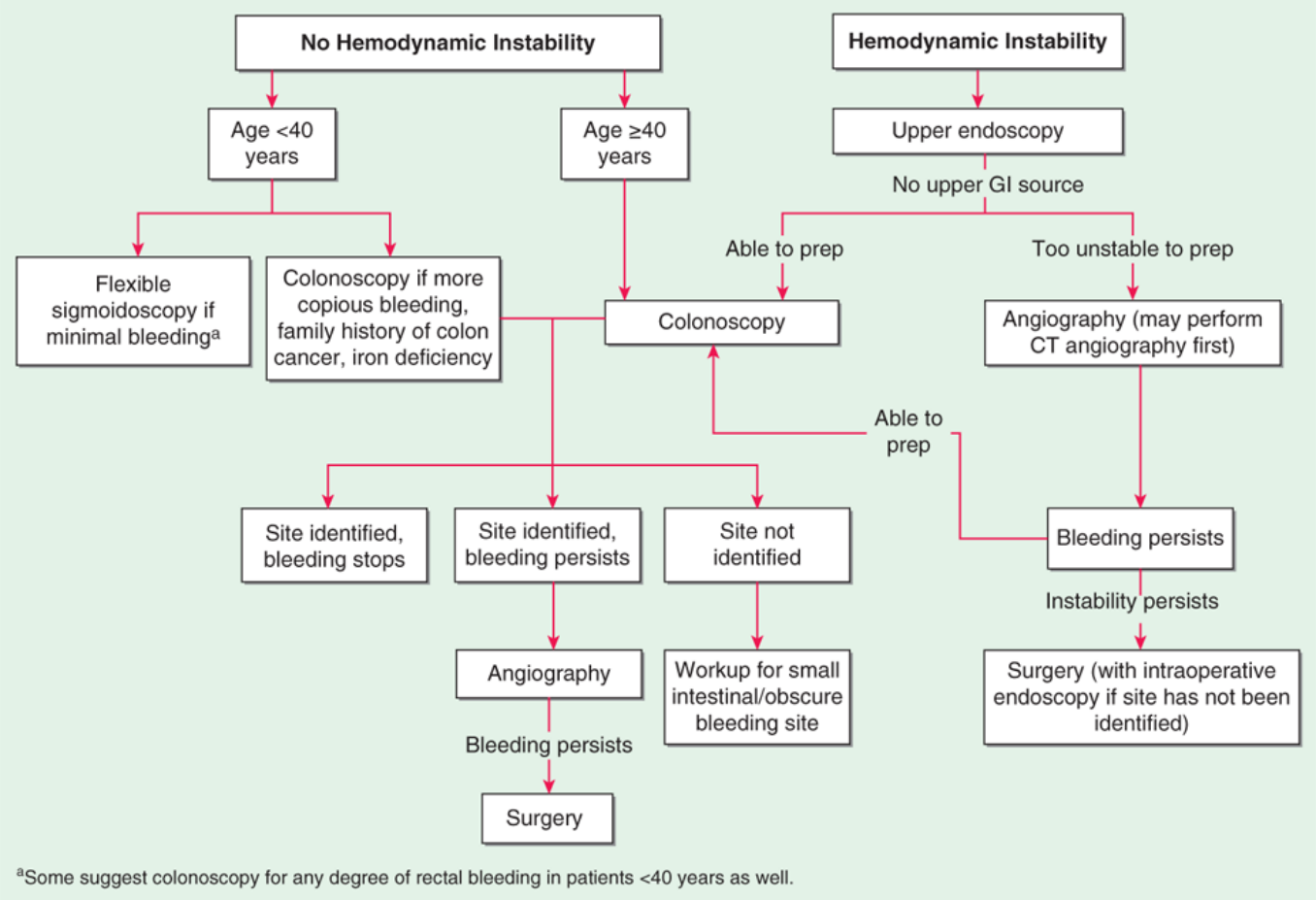
TABLE 44-1

Glasgow-Blatchford Score

Admission Marker	Score
Blood urea nitrogen (mg/dL)	
18.2 to <22.4	2
22.4 to <28.0	3
28.0 to <70.0	4
≥70.0	6
Hemoglobin (g/dL)	
12.0 to <13.0 (men); 10.0 to <12.0 (women)	1
10.0 to <12.0 (men)	3
<10.0	6
Systolic blood pressure (mmHg)	
100–109	1
90–99	2
<90	3
Heart rate (beats per minute)	
≥100	1
Other markers	
Melena	1
Syncope	2
Hepatic disease	2
Cardiac failure	2

FIGURE 44-2

Suggested algorithm for patients with acute lower gastrointestinal bleeding.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

De Franchis R: Expanding consensus in portal hypertension. J Hepatol 63:743, 2015. [PubMed: 26047908]

Garcia-Tsao G et al: Portal hypertensive bleeding in cirrhosis: Risk stratification, diagnosis, and management: 2016 practice guidance by the American Association for the Study of Liver Diseases. Hepatology 65:310, 2017. [PubMed: 27786365]

Gerson LB et al: ACG clinical guideline: Diagnosis and management of small bowel bleeding. Am J Gastroenterol 110:1265, 2015. [PubMed: 26303132]

Gralnek IM et al: Diagnosis and management of upper gastrointestinal hemorrhage: European Society of Gastrointestinal Endoscopy (ESGE) guideline. Endoscopy 47:1, 2015.

Laine L: Upper gastrointestinal bleeding due to a peptic ulcer. N Engl J Med 374:2367, 2016.

Laine L, Jensen DM: ACG Practice Guidelines: Management of patients with ulcer bleeding. Am J Gastroenterol 107:345, 2012. [PubMed: 22310222]

Strate LL, Gralnek KM: ACG clinical guideline: Management of patients with acute lower gastrointestinal bleeding. Am J Gastroenterol 111:459, 2016. [PubMed: 26925883]

Sung JJ et al: Continuation of low-dose aspirin therapy in peptic ulcer bleeding: A randomized trial. Ann Intern Med 152:1, 2010. [PubMed: 19949136]

Villaneuva C et al: Transfusion strategies for acute upper gastrointestinal bleeding. N Engl J Med 368:11, 2013. [PubMed: 23281973]

Chapter 45: Jaundice

Savio John; Daniel S. Pratt

FIGURE 45-1

INTRODUCTION

Jaundice is a yellowish discoloration of body tissues resulting from the deposition of bilirubin. Tissue deposition of bilirubin occurs only in the presence of serum hyperbilirubinemia and is a sign of either liver disease or, less often, a hemolytic disorder or disorder of bilirubin metabolism. The degree of serum bilirubin elevation can be estimated by physical examination. Slight increases in serum bilirubin level are best detected by examining the sclerae for icterus. Sclerae have a particular affinity for bilirubin due to their high elastin content, and the presence of scleral icterus indicates a serum bilirubin level of at least 51 μmol/L (3 mg/dL). The ability to detect scleral icterus is made more difficult if the examining room has fluorescent lighting. If the examiner suspects scleral icterus, a second site to examine is underneath the tongue. As serum bilirubin levels rise, the skin will eventually become yellow in light-skinned patients and even green if the process is long-standing; the green color is produced by oxidation of bilirubin to biliverdin.

The differential diagnosis for yellowing of the skin is limited. In addition to jaundice, it includes carotenoderma, the use of the drug quinacrine, and excessive exposure to phenols. Carotenoderma is the yellow color imparted to the skin of healthy individuals who ingest excessive amounts of vegetables and fruits that contain carotene, such as carrots, leafy vegetables, squash, peaches, and oranges. In jaundice the yellow coloration of the skin is uniformly distributed over the body, whereas in carotenoderma the pigment is concentrated on the palms, soles, forehead, and nasolabial folds. Carotenoderma can be distinguished from jaundice by the sparing of the sclerae. Quinacrine causes a yellow discoloration of the skin in 4–37% of patients treated with it.

Another sensitive indicator of increased serum bilirubin is darkening of the urine, which is due to the renal excretion of conjugated bilirubin. Patients often describe their urine as tea- or cola-colored. Bilirubinuria indicates an elevation of the direct serum bilirubin fraction and, therefore, the presence of liver or biliary disease.

Serum bilirubin levels increase when an imbalance exists between bilirubin production and clearance. A logical evaluation of the patient who is jaundiced requires an understanding of bilirubin production and metabolism.

PRODUCTION AND METABOLISM OF BILIRUBIN

(See Chap. 331) Bilirubin, a tetrapyrrole pigment, is a breakdown product of heme (ferroprotoporphyrin IX). About 80–85% of the 4 mg/kg body weight of bilirubin produced each day is derived from the breakdown of hemoglobin in senescent red blood cells. The remainder comes from prematurely destroyed erythroid cells in bone marrow and from the turnover of hemoproteins such as myoglobin and cytochromes found in tissues throughout the body.

The formation of bilirubin occurs in reticuloendothelial cells, primarily in the spleen and liver. The first reaction, catalyzed by the microsomal enzyme heme oxygenase, oxidatively cleaves the α bridge of the porphyrin group and opens the heme ring. The end products of this reaction are biliverdin, carbon monoxide, and iron. The second reaction, catalyzed by the cytosolic enzyme biliverdin reductase, reduces the central methylene bridge of biliverdin and converts it to bilirubin. Bilirubin formed in the reticuloendothelial cells is virtually insoluble in water due to tight internal hydrogen bonding between the water-soluble moieties of bilirubin—that is, the bonding of the propionic acid carboxyl groups of one dipyrrolic half of the molecule with the imino and lactam groups of the opposite half. This configuration blocks solvent access to the polar residues of bilirubin and places the hydrophobic residues on the outside. To be transported in blood, bilirubin must be solubilized. Solubilization is accomplished by the reversible, noncovalent binding of bilirubin to albumin. Unconjugated bilirubin bound to albumin is transported to the liver. There, the bilirubin—but not the albumin—is taken up by hepatocytes via a process that at least partly involves carrier-mediated membrane transport. No specific bilirubin transporter has yet been identified (**Chap. 331**, **Fig. 331-1**).

After entering the hepatocyte, unconjugated bilirubin is bound in the cytosol to a number of proteins including proteins in the glutathione-S-transferase superfamily. These proteins serve both to reduce efflux of bilirubin back into the serum and to present the bilirubin for conjugation. In the endoplasmic reticulum, bilirubin is made aqueous soluble by conjugation to glucuronic acid, a process that disrupts the hydrophobic internal hydrogen bonds and yields bilirubin monoglucuronide and diglucuronide. The conjugation of glucuronic acid to bilirubin is catalyzed by bilirubin uridine diphosphate-glucuronosyl transferase (UDPGT). The now-hydrophilic bilirubin conjugates diffuse from the endoplasmic reticulum to the canalicular membrane, where bilirubin monoglucuronide and diglucuronide are actively transported into canalicular bile by an energy-dependent mechanism involving the multidrug resistance–associated protein 2 (MRP2). A portion of bilirubin glucuronides is transported into the sinusoids and portal circulation by MRP3 and is subjected to reuptake into the hepatocyte by the sinusoidal organic anion transport protein 1B1 (OATP1B1) and OATP1B3. The conjugated bilirubin excreted into bile drains into the duodenum and passes unchanged through the proximal small bowel. Conjugated bilirubin is not reabsorbed by the intestinal mucosa due to its hydrophilicity and increased molecular size. When the conjugated bilirubin reaches the distal ileum and colon, it is hydrolyzed to unconjugated bilirubin by bacterial β-glucuronidases. The unconjugated bilirubin is reduced by normal gut bacteria to form a group of colorless tetrapyrroles called *urobilinogens* and other products, the nature and relative amounts of which depend on the bacterial flora. About 80–90% of these products are excreted in feces, either unchanged or oxidized to orange derivatives called *urobilins*. The

remaining 10–20% of the urobilinogens undergo enterohepatic cycling. A small fraction (usually <3 mg/dL) escapes hepatic uptake, filters across the renal glomerulus, and is excreted in urine. Increased urinary excretion of urobilinogen can be due to increased bilirubin production, increased hepatic reabsorption of urobilinogen from the colon, or decreased hepatic clearance of urobilinogen.

MEASUREMENT OF SERUM BILIRUBIN

The terms *direct* and *indirect* bilirubin—that is, conjugated and unconjugated bilirubin, respectively—are based on the original van den Bergh reaction. This assay, or a variation of it, is still used in most clinical chemistry laboratories to determine the serum bilirubin level. In this assay, bilirubin is exposed to diazotized sulfanilic acid and splits into two relatively stable dipyrrylmethene azopigments that absorb maximally at 540 nm, allowing photometric analysis. The direct fraction is that which reacts with diazotized sulfanilic acid in the absence of an accelerator substance such as alcohol. The direct fraction provides an approximation of the conjugated bilirubin level in serum. The *total* serum bilirubin is the amount that reacts after the addition of alcohol. The indirect fraction is the difference between the total and the direct bilirubin levels and provides an estimate of the unconjugated bilirubin in serum. Unconjugated bilirubin also reacts with diazo reagents, albeit slowly, even when the accelerator is absent. Thus the calculated indirect bilirubin may underestimate the true amount of unconjugated bilirubin in circulation.

With the van den Bergh method, the normal serum bilirubin concentration usually is between 17 and 26 µmol/L (1 and 1.5 mg/dL). Total serum bilirubin concentrations are between 3.4 and 15.4 µmol/L (0.2 and 0.9 mg/dL) in 95% of a normal population. Unconjugated hyperbilirubinemia is present when the direct fraction is <15% of the total serum bilirubin. The presence of even limited amounts of true conjugated bilirubin in serum suggests significant hepatobiliary pathology. As conjugated hyperbilirubinemia is always associated with bilirubinuria (except in the presence of delta bilirubin in prolonged cholestasis when jaundice is overt), detection of bilirubin in urine via dipstick test is extremely helpful to confirm the presence of conjugated hyperbilirubinemia in a patient with mildly elevated direct fraction.

Several new techniques, although less convenient to perform, have added considerably to our understanding of bilirubin metabolism. First, studies using these methods demonstrate that, in normal persons or those with Gilbert's syndrome, almost 100% of the serum bilirubin is unconjugated; <3% is monoconjugated bilirubin. Second, in jaundiced patients with hepatobiliary disease, the total serum bilirubin concentration measured by these new, more accurate methods is lower than the values found with diazo methods. This finding suggests that there are diazo-positive compounds distinct from bilirubin in the serum of patients with hepatobiliary disease. Third, these studies indicate that, in jaundiced patients with hepatobiliary disease, monoglucuronides of bilirubin predominate over diglucuronides. Fourth, part of the direct-reacting bilirubin fraction includes conjugated bilirubin that is covalently linked to albumin. This albumin-linked fraction of conjugated bilirubin (*delta fraction, delta bilirubin*, or *biliprotein*) represents an important fraction of total serum bilirubin in patients with cholestasis and hepatobiliary disorders. The delta bilirubin is formed in serum when hepatic excretion of bilirubin glucuronides is impaired and the glucuronides accumulate in serum. By virtue of its tight binding to albumin, the clearance rate of delta

bilirubin from serum approximates the half-life of albumin (12–14 days) rather than the short half-life of bilirubin (about 4 h).

The prolonged half-life of albumin-bound conjugated bilirubin accounts for two previously unexplained enigmas in jaundiced patients with liver disease: (1) that some patients with conjugated hyperbilirubinemia do not exhibit bilirubinuria during the recovery phase of their disease because the delta bilirubin, although conjugated, is covalently bound to albumin and therefore not filtered by the renal glomeruli, and (2) that the elevated serum bilirubin level declines more slowly than expected in some patients who otherwise appear to be recovering satisfactorily. Late in the recovery phase of hepatobiliary disorders, all the conjugated bilirubin may be in the albumin-linked form.

MEASUREMENT OF URINE BILIRUBIN

Unconjugated bilirubin is always bound to albumin in the serum, is not filtered by the kidney, and is not found in the urine. Conjugated bilirubin is filtered at the glomerulus, and the majority is reabsorbed by the proximal tubules; a small fraction is excreted in the urine. Any bilirubin found in the urine is conjugated bilirubin. The presence of bilirubinuria on urine dipstick test (Ictotest) indicates an elevation of the conjugated bilirubin fraction that cannot be excreted from the liver, and implies the presence of hepatobiliary disease. A false-negative result is possible in patients with prolonged cholestasis due to the predominance of delta bilirubin, which is covalently bound to albumin and therefore not filtered by the renal glomeruli.

APPROACH TO THE PATIENT

Jaundice

The goal of this chapter is not to provide an encyclopedic review of all of the conditions that can cause jaundice. Rather, the chapter is intended to offer a framework that helps a physician to evaluate the patient with jaundice in a logical way **(Fig. 45-1)**.

Simply stated, the initial step is to perform appropriate blood tests in order to determine whether the patient has an isolated elevation of serum bilirubin. If so, is the bilirubin elevation due to an increased unconjugated or conjugated fraction? If the hyperbilirubinemia is accompanied by other liver test abnormalities, is the disorder hepatocellular or cholestatic? If cholestatic, is it intra- or extrahepatic? All of these questions can be answered with a thoughtful history, physical examination, and interpretation of laboratory and radiologic tests and procedures.

The bilirubin present in serum represents a balance between input from the production of bilirubin and hepatic/biliary removal of the pigment. Hyperbilirubinemia may result from (1) overproduction of bilirubin; (2) impaired uptake, conjugation, or excretion of bilirubin; or (3) regurgitation of unconjugated or conjugated bilirubin from damaged hepatocytes or bile ducts. An increase in unconjugated bilirubin in serum results from overproduction, impaired uptake, or conjugation of bilirubin. An increase in conjugated bilirubin is due to decreased excretion into the bile ductules or backward leakage of the pigment. The initial steps in

evaluating the patient with jaundice are to determine (1) whether the hyperbilirubinemia is predominantly conjugated or unconjugated in nature and (2) whether other biochemical liver tests are abnormal. The thoughtful interpretation of limited data permits a rational evaluation of the patient (Fig. 45-1). The following discussion will focus solely on the evaluation of the adult patient with jaundice.

ISOLATED ELEVATION OF SERUM BILIRUBIN

Unconjugated Hyperbilirubinemia

The differential diagnosis of isolated unconjugated hyperbilirubinemia is limited **(Table 45-1)**. The critical determination is whether the patient is suffering from a hemolytic process resulting in an overproduction of bilirubin (hemolytic disorders and ineffective erythropoiesis) or from impaired hepatic uptake/conjugation of bilirubin (drug effect or genetic disorders).

Hemolytic disorders that cause excessive heme production may be either inherited or acquired. Inherited disorders include spherocytosis, sickle cell anemia, thalassemia, and deficiency of red cell enzymes such as pyruvate kinase and glucose-6-phosphate dehydrogenase. In these conditions, the serum bilirubin level rarely exceeds 86 µmol/L (5 mg/dL). Higher levels may occur when there is coexistent renal or hepatocellular dysfunction or in acute hemolysis, such as a sickle cell crisis. In evaluating jaundice in patients with chronic hemolysis, it is important to remember the high incidence of pigmented (calcium bilirubinate) gallstones found in these patients, which increases the likelihood of choledocholithiasis as an alternative explanation for hyperbilirubinemia.

Acquired hemolytic disorders include microangiopathic hemolytic anemia (e.g., hemolytic-uremic syndrome), paroxysmal nocturnal hemoglobinuria, spur cell anemia, immune hemolysis, and parasitic infections (e.g., malaria and babesiosis). Ineffective erythropoiesis occurs in cobalamin, folate, and iron deficiencies. Resorption of hematomas and massive blood transfusions both can result in increased hemoglobin release and overproduction of bilirubin.

In the absence of hemolysis, the physician should consider a problem with the hepatic uptake or conjugation of bilirubin. Certain drugs, including rifampin and probenecid, may cause unconjugated hyperbilirubinemia by diminishing hepatic uptake of bilirubin. Impaired bilirubin conjugation occurs in three genetic conditions: Crigler-Najjar syndrome types I and II and Gilbert's syndrome. *Crigler-Najjar type I* is an exceptionally rare condition found in neonates and characterized by severe jaundice (bilirubin >342 µmol/L [>20 mg/dL]) and neurologic impairment due to kernicterus, frequently leading to death in infancy or childhood. These patients have a complete absence of bilirubin UDPGT activity; are totally unable to conjugate bilirubin; and hence cannot excrete it.

Crigler-Najjar type II is somewhat more common than type I. Patients live into adulthood with serum bilirubin levels of 103–428 µmol/L (6–25 mg/dL). In these patients, mutations in the bilirubin *UDPGT* gene cause the reduction—typically ≤10%—of the enzyme's activity. Bilirubin UDPGT activity can be induced by the administration of phenobarbital, which can reduce serum bilirubin levels in these patients. Despite marked jaundice, these patients usually survive into adulthood, although they may be susceptible to kernicterus under the stress of concurrent illness or surgery.

Gilbert's syndrome is also marked by the impaired conjugation of bilirubin due to reduced bilirubin UDPGT activity (typically 10–35% of normal). Patients with Gilbert's syndrome have mild unconjugated hyperbilirubinemia, with serum levels almost always <103 μmol/L (6 mg/dL). The serum levels may fluctuate, and jaundice is often identified only during periods of stress, concurrent illness, alcohol use, or fasting. Unlike both Crigler-Najjar syndromes, Gilbert's syndrome is very common. The reported incidence is 3–7% of the population, with males predominating over females by a ratio of 1.5–7:1.

Conjugated Hyperbilirubinemia

Elevated conjugated hyperbilirubinemia is found in two rare inherited conditions: *Dubin-Johnson syndrome* and *Rotor syndrome* (Table 45-1). Patients with either condition present with asymptomatic jaundice. The defect in Dubin-Johnson syndrome is the presence of mutations in the gene for MRP2. These patients have altered excretion of bilirubin into the bile ducts. Rotor syndrome may represent a deficiency of the major hepatic drug reuptake transporters OATP1B1 and OATP1B3. Differentiating between these syndromes is possible but is clinically unnecessary due to their benign nature.

ELEVATION OF SERUM BILIRUBIN WITH OTHER LIVER TEST ABNORMALITIES

The remainder of this chapter will focus on the evaluation of patients with conjugated hyperbilirubinemia in the setting of other liver test abnormalities. This group of patients can be divided into those with a primary hepatocellular process and those with intra- or extrahepatic cholestasis. This distinction, which is based on the history and physical examination as well as the pattern of liver test abnormalities, guides the clinician's evaluation (Fig. 45-1).

History

A complete medical history is perhaps the single most important part of the evaluation of the patient with unexplained jaundice. Important considerations include the use of or exposure to any chemical or medication, whether physician-prescribed, over-the-counter, complementary, or alternative medicines (e.g., herbal and vitamin preparations) or other drugs such as anabolic steroids. The patient should be carefully questioned about possible parenteral exposures, including transfusions, intravenous and intranasal drug use, tattooing, and sexual activity. Other important points include recent travel history; exposure to people with jaundice; exposure to possibly contaminated foods; occupational exposure to hepatotoxins; alcohol consumption; the duration of jaundice; and the presence of any accompanying signs and symptoms, such as arthralgias, myalgias, rash, anorexia, weight loss, abdominal pain, fever, pruritus, and changes in the urine and stool. While none of the latter manifestations is specific for any one condition, any of them can suggest a particular diagnosis. A history of arthralgias and myalgias predating jaundice suggests hepatitis, either viral or drug-related. Jaundice associated with the sudden onset of severe right-upper-quadrant pain and shaking chills suggests choledocholithiasis and ascending cholangitis.

Physical Examination

The general assessment should include evaluation of the patient's nutritional status. Temporal and proximal muscle wasting suggests long-standing disease such as pancreatic cancer or cirrhosis. Stigmata of chronic liver disease, including spider nevi, palmar erythema, gynecomastia, caput medusae, Dupuytren's

contractures, parotid gland enlargement, and testicular atrophy, are commonly seen in advanced alcoholic (Laennec's) cirrhosis and occasionally in other types of cirrhosis. An enlarged left supraclavicular node (Virchow's node) or a periumbilical nodule (Sister Mary Joseph's nodule) suggests an abdominal malignancy. Jugular venous distention, a sign of right-sided heart failure, suggests hepatic congestion. Right pleural effusion even in the absence of clinically apparent ascites may be seen in advanced cirrhosis.

The abdominal examination should focus on the size and consistency of the liver, on whether the spleen is palpable and hence enlarged, and on whether ascites is present. Patients with cirrhosis may have an enlarged left lobe of the liver, which is felt below the xiphoid, and an enlarged spleen. A grossly enlarged nodular liver or an obvious abdominal mass suggests malignancy. An enlarged tender liver could signify viral or alcoholic hepatitis; an infiltrative process such as amyloidosis; or, less often, an acutely congested liver secondary to right-sided heart failure. Severe right-upper-quadrant tenderness with respiratory arrest on inspiration (Murphy's sign) suggests cholecystitis. Ascites in the presence of jaundice suggests either cirrhosis or malignancy with peritoneal spread.

Laboratory Tests

A battery of tests are helpful in the initial evaluation of a patient with unexplained jaundice. These include total and direct serum bilirubin measurement with fractionation; determination of serum aminotransferase, alkaline phosphatase, and albumin concentrations; and prothrombin time tests. Enzyme tests (alanine aminotransferase [ALT], aspartate aminotransferase [AST], and alkaline phosphatase [ALP]) are helpful in differentiating between a hepatocellular process and a cholestatic process (**Table 330-1;** Fig. 45-1)—a critical step in determining what additional workup is indicated. Patients with a hepatocellular process generally have a rise in the aminotransferases that is disproportionate to that in ALP, whereas patients with a cholestatic process have a rise in ALP that is disproportionate to that of the aminotransferases. The serum bilirubin can be prominently elevated in both hepatocellular and cholestatic conditions and therefore is not necessarily helpful in differentiating between the two.

In addition to enzyme tests, all jaundiced patients should have additional blood tests—specifically, an albumin level and a prothrombin time—to assess liver function. A low albumin level suggests a chronic process such as cirrhosis or cancer. A normal albumin level is suggestive of a more acute process such as viral hepatitis or choledocholithiasis. An elevated prothrombin time indicates either vitamin K deficiency due to prolonged jaundice and malabsorption of vitamin K or significant hepatocellular dysfunction. The failure of the prothrombin time to correct with parenteral administration of vitamin K indicates severe hepatocellular injury.

The results of the bilirubin, enzyme, albumin, and prothrombin time tests will usually indicate whether a jaundiced patient has a hepatocellular or a cholestatic disease and offer some indication of the duration and severity of the disease. The causes and evaluations of hepatocellular and cholestatic diseases are quite different.

Hepatocellular Conditions

Hepatocellular diseases that can cause jaundice include viral hepatitis, drug or environmental toxicity, alcohol, and end-stage cirrhosis from any cause **(Table 45-2)**. Wilson's disease occurs primarily in young adults. Autoimmune hepatitis is typically seen in young to middle-aged women, but may affect men and women of any age. Alcoholic hepatitis can be differentiated from viral and toxin-related hepatitis by the pattern of the aminotransferases: patients with alcoholic hepatitis typically have an AST-to-ALT ratio of at least 2:1, and the AST level rarely exceeds 300 U/L. Patients with acute viral hepatitis and toxin-related injury severe enough to produce jaundice typically have aminotransferase levels >500 U/L, with the ALT greater than or equal to the AST. While ALT and AST values <8 times normal may be seen in either hepatocellular or cholestatic liver disease, values 25 times normal or higher are seen primarily in acute hepatocellular diseases. Patients with jaundice from cirrhosis can have normal or only slightly elevated aminotransferase levels.

When the clinician determines that a patient has a hepatocellular disease, appropriate testing for acute viral hepatitis includes a hepatitis A IgM antibody assay, a hepatitis B surface antigen and core IgM antibody assay, a hepatitis C viral RNA test, and, depending on the circumstances, a hepatitis E IgM antibody assay. Because it can take many weeks for hepatitis C antibody to become detectable, its assay is an unreliable test if acute hepatitis C is suspected. Studies for hepatitis D and E viruses, Epstein-Barr virus (EBV), and cytomegalovirus (CMV) may also be indicated. Ceruloplasmin is the initial screening test for Wilson's disease. Testing for autoimmune hepatitis usually includes an antinuclear antibody assay and measurement of specific immunoglobulins.

Drug-induced hepatocellular injury can be classified as either predictable or unpredictable. Predictable drug reactions are dose-dependent and affect all patients who ingest a toxic dose of the drug in question. The classic example is acetaminophen hepatotoxicity. Unpredictable or idiosyncratic drug reactions are not dose-dependent and occur in a minority of patients. A great number of drugs can cause idiosyncratic hepatic injury. Environmental toxins are also an important cause of hepatocellular injury. Examples include industrial chemicals such as vinyl chloride, herbal preparations containing pyrrolizidine alkaloids (Jamaica bush tea) or Kava, and the mushrooms *Amanita phalloides* and *A. verna*, which contain highly hepatotoxic amatoxins.

Cholestatic Conditions

When the pattern of the liver tests suggests a cholestatic disorder, the next step is to determine whether it is intra- or extrahepatic cholestasis (Fig. 45-1). Distinguishing intrahepatic from extrahepatic cholestasis may be difficult. History, physical examination, and laboratory tests often are not helpful. The next appropriate test is an ultrasound. The ultrasound is inexpensive, does not expose the patient to ionizing radiation, and can detect dilation of the intra- and extrahepatic biliary tree with a high degree of sensitivity and specificity. The absence of biliary dilation suggests intrahepatic cholestasis, while its presence indicates extrahepatic cholestasis. False-negative results occur in patients with partial obstruction of the common bile duct or in patients with cirrhosis or primary sclerosing cholangitis (PSC), in which scarring prevents the intrahepatic ducts from dilating.

Although ultrasonography may indicate extrahepatic cholestasis, it rarely identifies the site or cause of obstruction. The distal common bile duct is a particularly difficult area to visualize by ultrasound because of

overlying bowel gas. Appropriate next tests include CT, magnetic resonance cholangiopancreatography (MRCP), endoscopic retrograde cholangiopancreatography (ERCP), percutaneous transhepatic cholangiography (PTC), and endoscopic ultrasound (EUS). CT scanning and MRCP are better than ultrasonography for assessing the head of the pancreas and for identifying choledocholithiasis in the distal common bile duct, particularly when the ducts are not dilated. ERCP is the "gold standard" for identifying choledocholithiasis. Beyond its diagnostic capabilities, ERCP allows therapeutic interventions, including the removal of common bile duct stones and the placement of stents. PTC can provide the same information as ERCP and it also allows for intervention in patients in whom ERCP is unsuccessful due to proximal biliary obstruction or altered gastrointestinal anatomy. MRCP has replaced ERCP as the initial diagnostic test in cases where the need for intervention is thought to be small. EUS displays sensitivity and specificity comparable to that of MRCP in the detection of bile duct obstruction. EUS also allows biopsy of suspected malignant lesions, but is invasive and requires sedation.

In patients with apparent *intrahepatic cholestasis*, the diagnosis is often made by serologic testing in combination with percutaneous liver biopsy. The list of possible causes of intrahepatic cholestasis is long and varied **(Table 45-3)**. A number of conditions that typically cause a hepatocellular pattern of injury can also present as a cholestatic variant. Both hepatitis B and C viruses can cause cholestatic hepatitis (fibrosing cholestatic hepatitis). This disease variant has been reported in patients who have undergone solid organ transplantation. Hepatitis A and E, alcoholic hepatitis, and EBV or CMV infections may also present as cholestatic liver disease.

Drugs may cause intrahepatic cholestasis that is usually reversible after discontinuation of the offending agent, although it may take many months for cholestasis to resolve. Drugs most commonly associated with cholestasis are the anabolic and contraceptive steroids. Cholestatic hepatitis has been reported with chlorpromazine, imipramine, tolbutamide, sulindac, cimetidine, and erythromycin estolate. It also occurs in patients taking trimethoprim; sulfamethoxazole; and penicillin-based antibiotics such as ampicillin, dicloxacillin, and clavulanic acid. Rarely, cholestasis may be chronic and associated with progressive fibrosis despite early discontinuation of the offending drug. Chronic cholestasis has been associated with chlorpromazine and prochlorperazine.

Primary biliary cholangitis is an autoimmune disease predominantly affecting middle-aged women and characterized by progressive destruction of interlobular bile ducts. The diagnosis is made by the detection of antimitochondrial antibody, which is found in 95% of patients. *Primary sclerosing cholangitis* is characterized by the destruction and fibrosis of larger bile ducts. The diagnosis of PSC is made with cholangiography (either MRCP or ERCP), which demonstrates the pathognomonic segmental strictures. Approximately 75% of patients with PSC have inflammatory bowel disease.

The *vanishing bile duct syndrome* and *adult bile ductopenia* are rare conditions in which a decreased number of bile ducts are seen in liver biopsy specimens. The histologic picture is similar to that in primary biliary cholangitis. This picture is seen in patients who develop chronic rejection after liver transplantation and in those who develop graft-versus-host disease after bone marrow transplantation. Vanishing bile duct

syndrome also occurs in rare cases of sarcoidosis, in patients taking certain drugs (including chlorpromazine), and idiopathically.

There are also familial forms of intrahepatic cholestasis. The familial intrahepatic cholestatic syndromes include *progressive familial intrahepatic cholestasis* (PFIC) *types 1–3* and *benign recurrent intrahepatic cholestasis* (BRIC) types 1 and 2. BRIC is characterized by episodic attacks of pruritus, cholestasis, and jaundice beginning at any age, which can be debilitating but does not lead to chronic liver disease. Serum bile acids are elevated during episodes, but serum γ-glutamyltransferase (γ-GT) activity is normal. PFIC disorders begin at childhood and are progressive in nature. All three types of PFIC are associated with progressive cholestasis, elevated levels of serum bile acids, similar phenotypes but different genetic mutations. Only type 3 PFIC is associated with high levels of γ-GT. *Cholestasis of pregnancy* occurs in the second and third trimesters and resolves after delivery. Its cause is unknown, but the condition is probably inherited, and cholestasis can be triggered by estrogen administration.

Other causes of intrahepatic cholestasis include total parenteral nutrition (TPN); nonhepatobiliary sepsis; benign postoperative cholestasis; and a paraneoplastic syndrome associated with a number of different malignancies, including Hodgkin's disease, medullary thyroid cancer, renal cell cancer, renal sarcoma, T cell lymphoma, prostate cancer, and several gastrointestinal malignancies. The term *Stauffer's syndrome* has been used for intrahepatic cholestasis specifically associated with renal cell cancer. In patients developing cholestasis in the intensive care unit, the major considerations should be sepsis, ischemic hepatitis ("shock liver"), and TPN jaundice. Jaundice occurring after bone marrow transplantation is most likely due to veno-occlusive disease or graft-versus-host disease. In addition to hemolysis, sickle cell disease may cause intrahepatic and extrahepatic cholestasis. Jaundice is a late finding in heart failure caused by hepatic congestion and hepatocellular hypoxia. Ischemic hepatitis is a distinct entity of acute hypoperfusion characterized by an acute and dramatic elevation in the serum aminotransferases followed by a gradual peak in serum bilirubin.

Jaundice with associated liver dysfunction can be seen in severe cases of *Plasmodium falciparum* malaria. The jaundice in these cases is due to a combination of indirect hyperbilirubinemia from hemolysis and both cholestatic and hepatocellular jaundice. Weil's disease, a severe presentation of leptospirosis, is marked by jaundice with renal failure, fever, headache, and muscle pain.

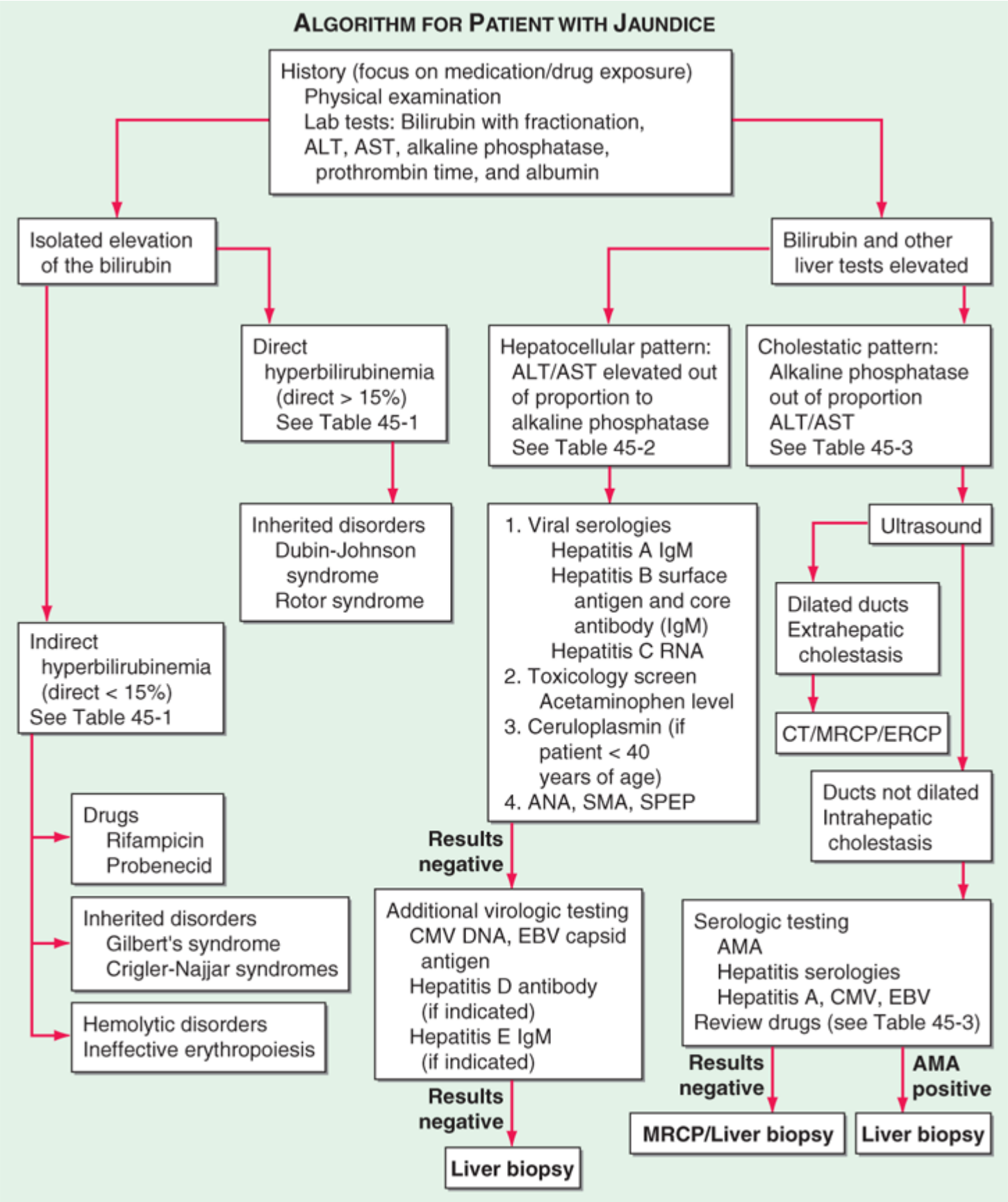
Causes of *extrahepatic cholestasis* can be split into malignant and benign (Table 45-3). Malignant causes include pancreatic, gallbladder, and ampullary cancers as well as cholangiocarcinoma. This last malignancy is most commonly associated with PSC and is exceptionally difficult to diagnose because its appearance is often identical to that of PSC. Pancreatic and gallbladder tumors as well as cholangiocarcinoma are rarely resectable and have poor prognoses. Ampullary carcinoma has the highest surgical cure rate of all the tumors that present as painless jaundice. Hilar lymphadenopathy due to metastases from other cancers may cause obstruction of the extrahepatic biliary tree.

Choledocholithiasis is the most common cause of extrahepatic cholestasis. The clinical presentation can range from mild right-upper-quadrant discomfort with only minimal elevations of enzyme test values to

ascending cholangitis with jaundice, sepsis, and circulatory collapse. PSC may occur with clinically important strictures limited to the extrahepatic biliary tree. IgG4-associated cholangitis is marked by stricturing of the biliary tree. It is critical that the clinician differentiate this condition from PSC as it is responsive to glucocorticoid therapy. In rare instances, chronic pancreatitis causes strictures of the distal common bile duct, where it passes through the head of the pancreas. AIDS cholangiopathy is a condition that is usually due to infection of the bile duct epithelium with CMV or cryptosporidia and has a cholangiographic appearance similar to that of PSC. The affected patients usually present with greatly elevated serum alkaline phosphatase levels (mean, 800 IU/L), but the bilirubin level is often near normal. These patients do not typically present with jaundice.

FIGURE 45-1

Evaluation of the patient with jaundice. ALT, alanine aminotransferase; AMA, antimitochondrial antibody; ANA, antinuclear antibody; AST, aspartate aminotransferase; CMV, cytomegalovirus; EBV, Epstein-Barr virus; LKM, liver-kidney microsomal antibody; MRCP, magnetic resonance cholangiopancreatography; SMA, smooth-muscle antibody; SPEP, serum protein electrophoresis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 45-1

Causes of Isolated Hyperbilirubinemia

I. Indirect hyperbilirubinemia
   A. Hemolytic disorders
   B. Ineffective erythropoiesis
   C. Increased bilirubin production
      1. Massive blood transfusion
      2. Resorption of hematoma
   D. Drugs
      1. Rifampin
      2. Probenecid
      3. Ribavirin
      4. Protease inhibitors (Atazanavir, Indinavir)
   E. Inherited conditions
      1. Crigler-Najjar types I and II
      2. Gilbert's syndrome

II. Direct hyperbilirubinemia (inherited conditions)
   A. Dubin-Johnson syndrome
   B. Rotor syndrome

TABLE 45-2

Hepatocellular Conditions That May Produce Jaundice

Viral hepatitis
Hepatitis A, B, C, D, and E
Epstein-Barr virus
Cytomegalovirus
Herpes simplex virus
Alcoholic hepatitis
Chronic liver disease and cirrhosis
Drug toxicity
Predictable, dose-dependent (e.g., acetaminophen)
Unpredictable, idiosyncratic (e.g., isoniazid)
Environmental toxins
Vinyl chloride
Jamaica bush tea—pyrrolizidine alkaloids
Kava Kava
Wild mushrooms—*Amanita phalloides, A. verna*
Wilson’s disease
Autoimmune hepatitis

TABLE 45-3

Cholestatic Conditions That May Produce Jaundice

I. Intrahepatic
- A. Viral hepatitis
  - 1. Fibrosing cholestatic hepatitis—hepatitis B and C
  - 2. Hepatitis A, Epstein-Barr virus infection, cytomegalovirus infection
- B. Alcoholic hepatitis
- C. Drug toxicity
  - 1. Pure cholestasis—anabolic and contraceptive steroids
  - 2. Cholestatic hepatitis—chlorpromazine, erythromycin estolate
  - 3. Chronic cholestasis—chlorpromazine and prochlorperazine
- D. Primary biliary cholangitis
- E. Primary sclerosing cholangitis
- F. Vanishing bile duct syndrome
  - 1. Chronic rejection of liver transplants
  - 2. Sarcoidosis
  - 3. Drugs
- G. Congestive hepatopathy and ischemic hepatitis
- H. Inherited conditions
  - 1. Progressive familial intrahepatic cholestasis
  - 2. Benign recurrent intrahepatic cholestasis
- I. Cholestasis of pregnancy
- J. Total parenteral nutrition
- K. Nonhepatobiliary sepsis
- L. Benign postoperative cholestasis
- M. Paraneoplastic syndrome
- N. Veno-occlusive disease
- O. Graft-versus-host disease
- P. Infiltrative disease
  - 1. Tuberculosis
  - 2. Lymphoma
  - 3. Amyloidosis
- Q. Infections
  - 1. Malaria
  - 2. Leptospirosis

II. Extrahepatic
- A. Malignant
  - 1. Cholangiocarcinoma
  - 2. Pancreatic cancer

3. Gallbladder cancer
4. Ampullary cancer
5. Malignant involvement of the porta hepatis lymph nodes

B. Benign

1. Choledocholithiasis
2. Postoperative biliary strictures
3. Primary sclerosing cholangitis
4. Chronic pancreatitis
5. AIDS cholangiopathy
6. Mirizzi's syndrome
7. Parasitic disease (ascariasis)

GLOBAL CONSIDERATIONS

 While extrahepatic biliary obstruction and drugs are common causes of new-onset jaundice in developed countries, infections remain the leading cause in developing countries. Liver involvement and jaundice are observed with numerous infections, particularly malaria, babesiosis, severe leptospirosis, infections due to *Mycobacterium tuberculosis* and the *Mycobacterium avium* complex, typhoid fever, infection with hepatitis viruses A–E, EBV, CMV, Ebola virus, late phases of yellow fever, dengue hemorrhagic fever, schistosomiasis, fascioliasis, clonorchiasis, opisthorchiasis, ascariasis, echinococcosis, hepatosplenic candidiasis, disseminated histoplasmosis, cryptococcosis, coccidioimycosis, ehrlichiosis, chronic Q fever, yersiniosis, brucellosis, syphilis, and leprosy. Bacterial infections that do not necessarily involve the liver and bile ducts may also lead to jaundice, as in cholestasis of sepsis. The presence of fever or abdominal pain suggests concurrent infection, sepsis, or complications from gallstones. The development of encephalopathy and coagulopathy in a jaundiced patient with no preexisting liver disease signifies acute liver failure, which warrants urgent liver transplant evaluation.

ACKNOWLEDGMENT

This chapter is a revised version of chapters that have appeared in prior editions of Harrison's in which Marshall M. Kaplan was a co-author together with Daniel Pratt.

FURTHER READING

Erlinger S, Arias IM, Dhumeaux D: Inherited disorders of bilirubin transport and conjugation: New insights into molecular mechanisms and consequences. Gastroenterology 146:1625, 2014. [PubMed: 24704527]

Wolkoff AW et al: Bilirubin metabolism and jaundice, in *Schiff's Diseases of the Liver*, 11th ed, Schiff ER et al (eds). Oxford, UK, John Wiley & Sons, Ltd, 2012, pp 120–150.

Chapter 46: Abdominal Swelling and Ascites

Kathleen E. Corey; Lawrence S. Friedman

ABDOMINAL SWELLING

Abdominal swelling is a manifestation of numerous diseases. Patients may complain of bloating or abdominal fullness and may note increasing abdominal girth on the basis of increased clothing or belt size. Abdominal discomfort is often reported, but pain is less frequent. When abdominal pain does accompany swelling, it is frequently the result of an intraabdominal infection, peritonitis, or pancreatitis. Patients with abdominal distention from *ascites* (fluid in the abdomen) may report the new onset of an inguinal or umbilical hernia. Dyspnea may result from pressure against the diaphragm and the inability to expand the lungs fully.

CAUSES

The causes of abdominal swelling can be remembered conveniently as the *six Fs*: flatus, fat, fluid, fetus, feces, or a “fatal growth” (often a neoplasm).

Flatus

Abdominal swelling may be the result of increased intestinal gas. The normal small intestine contains ~200 mL of gas made up of nitrogen, oxygen, carbon dioxide, hydrogen, and methane. Nitrogen and oxygen are consumed (swallowed), whereas carbon dioxide, hydrogen, and methane are produced intraluminally by bacterial fermentation. Increased intestinal gas can occur in a number of conditions. *Aerophagia*, the swallowing of air, can result in increased amounts of oxygen and nitrogen in the small intestine and lead to abdominal swelling. Aerophagia typically results from gulping food; chewing gum; smoking; or as a response to anxiety, which can lead to repetitive belching. In some cases, increased intestinal gas is the consequence of bacterial metabolism of excess fermentable substances such as lactose and other oligosaccharides, which can lead to production of hydrogen, carbon dioxide, or methane. In many cases, the precise cause of abdominal distention cannot be determined. In some persons, particularly those with irritable bowel syndrome and bloating, the subjective sense of abdominal pressure is attributable to impaired intestinal transit of gas rather than increased gas volume. Abdominal distention—an objective increase in girth—is the result of a lack of coordination between diaphragmatic contraction and anterior abdominal wall relaxation, a response in some cases to an increase in intraabdominal volume loads. Occasionally, increased lumbar lordosis accounts for apparent abdominal distention.

Fat

Weight gain with an increase in abdominal fat can result in an increase in abdominal girth and can be perceived as abdominal swelling. Abdominal fat may be caused by an imbalance between caloric intake and energy expenditure associated with a poor diet and sedentary lifestyle; it also can be a manifestation of certain diseases, such as Cushing's syndrome. Excess abdominal fat has been associated with an increased risk of insulin resistance and cardiovascular disease.

Fluid

The accumulation of fluid within the abdominal cavity (ascites) often results in abdominal distention and is discussed in detail below.

Fetus

Pregnancy results in increased abdominal girth. Typically, an increase in abdominal size is first noted at 12–14 weeks of gestation, when the uterus moves from the pelvis into the abdomen. Abdominal distention may be seen before this point as a result of fluid retention and relaxation of the abdominal muscles.

Feces

In the setting of severe constipation or intestinal obstruction, increased stool in the colon leads to increased abdominal girth. These conditions are often accompanied by abdominal discomfort or pain, nausea, and vomiting and can be diagnosed by imaging studies.

Fatal Growth

An abdominal mass can result in abdominal swelling. Neoplasms, abscesses, or cysts can grow to sizes that lead to increased abdominal girth. Enlargement of the intraabdominal organs, specifically the liver (hepatomegaly) or spleen (splenomegaly), or an abdominal aortic aneurysm can result in abdominal distention. Bladder distention also may result in abdominal swelling.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Abdominal Swelling

HISTORY

Determining the etiology of abdominal swelling begins with history-taking and a physical examination. Patients should be questioned regarding symptoms suggestive of malignancy, including weight loss, night sweats, and anorexia. Inability to pass stool or flatus together with nausea or vomiting suggests bowel obstruction, severe constipation, or an ileus (lack of peristalsis). Increased eructation and flatus may point toward aerophagia or increased intestinal production of gas. Patients should be questioned about risk

factors for or symptoms of chronic liver disease, including excessive alcohol use and jaundice, which suggest ascites. Patients should also be asked about symptoms of other medical conditions, including heart failure and tuberculosis, which may cause ascites.

PHYSICAL EXAMINATION

Physical examination should include an assessment for signs of systemic disease. The presence of lymphadenopathy, especially supraclavicular lymphadenopathy (*Virchow's node*), suggests metastatic abdominal malignancy. Care should be taken during the cardiac examination to evaluate for elevation of jugular venous pressure (JVP); *Kussmaul's sign* (elevation of the JVP during inspiration); a pericardial knock, which may be seen in heart failure or constrictive pericarditis; or a murmur of tricuspid regurgitation. Spider angiomas, palmar erythema, dilated superficial veins around the umbilicus (*caput medusae*), and gynecomastia suggest chronic liver disease.

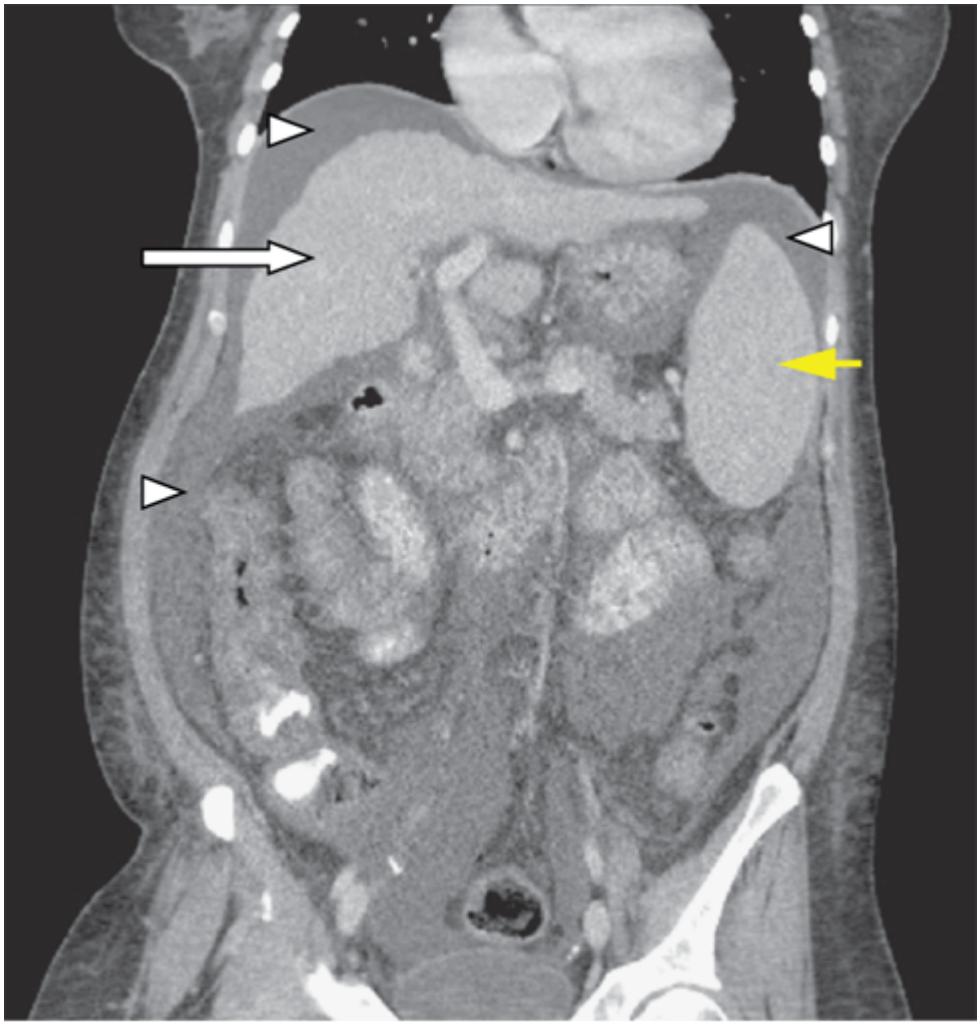
The abdominal examination should begin with inspection for the presence of uneven distention or an obvious mass. Auscultation should follow. The absence of bowel sounds or the presence of high-pitched localized bowel sounds points toward an ileus or intestinal obstruction. An umbilical venous hum may suggest the presence of portal hypertension, and a harsh bruit over the liver is heard rarely in patients with hepatocellular carcinoma or alcoholic hepatitis. Abdominal swelling caused by intestinal gas can be differentiated from swelling caused by fluid or a solid mass by percussion; an abdomen filled with gas is tympanic, whereas an abdomen containing a mass or fluid is dull to percussion. The absence of abdominal dullness, however, does not exclude ascites, because a minimum of 1500 mL of ascitic fluid is required for detection on physical examination. Finally, the abdomen should be palpated to assess for tenderness, a mass, enlargement of the spleen or liver, or presence of a nodular liver suggesting cirrhosis or tumor. Light palpation of the liver may detect pulsations suggesting retrograde vascular flow from the heart in patients with right-sided heart failure, particularly tricuspid regurgitation.

IMAGING AND LABORATORY EVALUATION

Abdominal x-rays can be used to detect dilated loops of bowel suggesting intestinal obstruction or ileus. Abdominal ultrasonography can detect as little as 100 mL of ascitic fluid, hepatosplenomegaly, a nodular liver, or a mass. Ultrasonography is often inadequate to detect retroperitoneal lymphadenopathy or a pancreatic lesion because of overlying bowel gas. If malignancy or pancreatic disease is suspected, CT can be performed. CT may also detect changes associated with advanced cirrhosis and portal hypertension **(Fig. 46-1)**.

FIGURE 46-1

CT of a patient with a cirrhotic, nodular liver (*white arrow*), splenomegaly (*yellow arrow*), and ascites (*arrowheads*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Laboratory evaluation should include liver biochemical testing, serum albumin level measurement, and prothrombin time determination (international normalized ratio) to assess hepatic function as well as a complete blood count to evaluate for the presence of cytopenias that may result from portal hypertension or of leukocytosis, anemia, and thrombocytosis that may result from systemic infection. Serum amylase and lipase levels should be checked to evaluate the patient for acute pancreatitis. Urinary protein quantitation is indicated when nephrotic syndrome, which may cause ascites, is suspected.

In selected cases, the hepatic venous pressure gradient (pressure across the liver between the portal and hepatic veins) can be measured via cannulation of the hepatic vein to confirm that ascites is caused by cirrhosis (Chap. 337). In some cases, a liver biopsy may be necessary to confirm cirrhosis.

ASCITES

PATHOGENESIS IN THE PRESENCE OF CIRRHOSIS

Ascites in patients with cirrhosis is the result of portal hypertension and renal salt and water retention. Similar mechanisms contribute to ascites formation in heart failure. Portal hypertension signifies elevation of

the pressure within the portal vein. According to Ohm's law, pressure is the product of resistance and flow. Increased hepatic resistance occurs by several mechanisms. First, the development of hepatic fibrosis, which defines cirrhosis, disrupts the normal architecture of the hepatic sinusoids and impedes normal blood flow through the liver. Second, activation of hepatic stellate cells, which mediate fibrogenesis, leads to smooth-muscle contraction and fibrosis. Finally, cirrhosis is associated with a decrease in endothelial nitric oxide synthetase (eNOS) production, which results in decreased nitric oxide production and increased intrahepatic vasoconstriction.

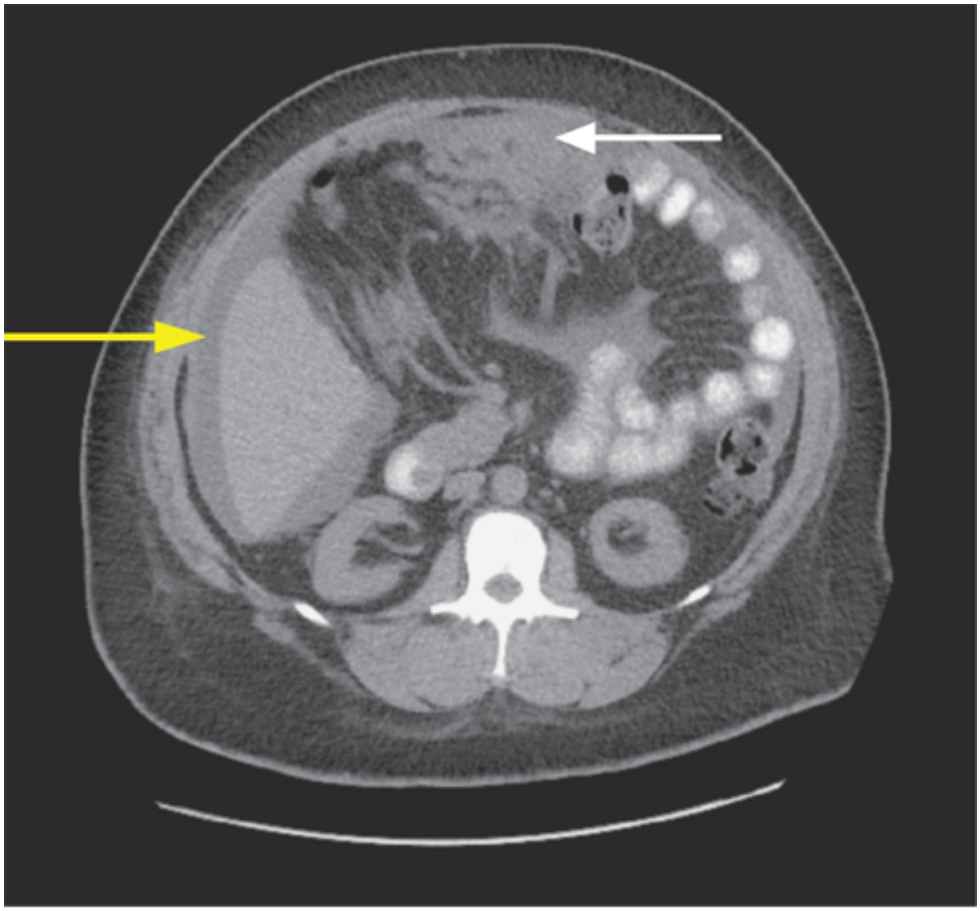
The development of cirrhosis is also associated with increased systemic circulating levels of nitric oxide (contrary to the decrease seen intrahepatically) as well as increased levels of vascular endothelial growth factor and tumor necrosis factor that result in splanchnic arterial vasodilation. Vasodilation of the splanchnic circulation results in pooling of blood and a decrease in the effective circulating volume, which is perceived by the kidneys as hypovolemia. Compensatory vasoconstriction via release of antidiuretic hormone ensues; the consequences are free water retention and activation of the sympathetic nervous system and the renin angiotensin aldosterone system, which lead in turn to renal sodium and water retention.

PATHOGENESIS IN THE ABSENCE OF CIRRHOSIS

Ascites in the absence of cirrhosis generally results from peritoneal carcinomatosis, peritoneal infection, or pancreatic disease. Peritoneal carcinomatosis can result from primary peritoneal malignancies such as mesothelioma or sarcoma, abdominal malignancies such as gastric or colonic adenocarcinoma, or metastatic disease from breast or lung carcinoma or melanoma **(Fig. 46-2)**. The tumor cells lining the peritoneum produce a protein-rich fluid that contributes to the development of ascites. Fluid from the extracellular space is drawn into the peritoneum, further contributing to the development of ascites. Tuberculous peritonitis causes ascites via a similar mechanism; tubercles deposited on the peritoneum exude a proteinaceous fluid. Pancreatic ascites results from leakage of pancreatic enzymes into the peritoneum.

FIGURE 46-2

CT of a patient with peritoneal carcinomatosis (*white arrow*) and ascites (*yellow arrow*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CAUSES

Cirrhosis accounts for 84% of cases of ascites. Cardiac ascites, peritoneal carcinomatosis, and "mixed" ascites resulting from cirrhosis and a second disease account for 10–15% of cases. Less common causes of ascites include massive hepatic metastasis, infection (tuberculosis, *Chlamydia* infection), pancreatitis, and renal disease (nephrotic syndrome). Rare causes of ascites include hypothyroidism and familial Mediterranean fever.

EVALUATION

Once the presence of ascites has been confirmed, the etiology of the ascites is best determined by *paracentesis*, a bedside procedure in which a needle or small catheter is passed transcutaneously to extract ascitic fluid from the peritoneum. The lower quadrants are the most frequent sites for paracentesis. The left lower quadrant is preferred because of the greater depth of ascites and the thinner abdominal wall. Paracentesis is a safe procedure even in patients with coagulopathy; complications, including abdominal wall hematomas, hypotension, hepatorenal syndrome, and infection, are infrequent.

Once ascitic fluid has been extracted, its gross appearance should be examined. Turbid fluid can result from the presence of infection or tumor cells. White, milky fluid indicates a triglyceride level >200 mg/dL (and

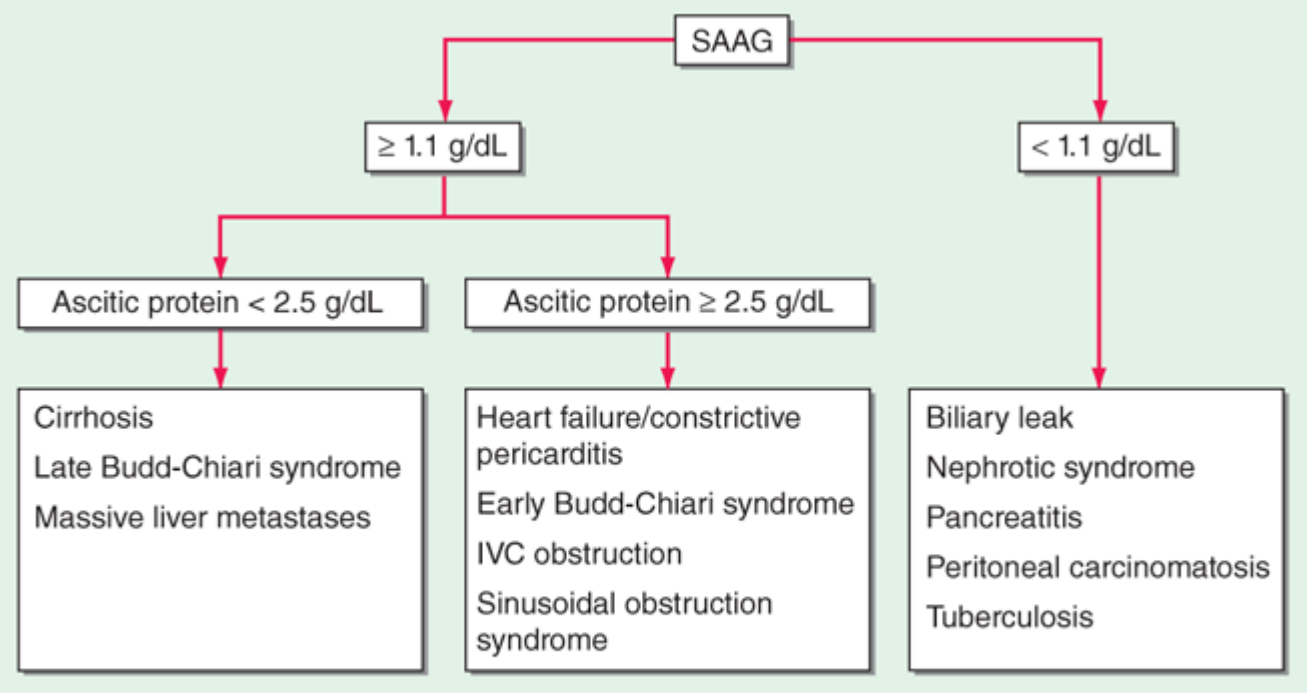
often >1000 mg/dL), which is the hallmark of *chylous ascites*. Chylous ascites results from lymphatic disruption that may occur with trauma, cirrhosis, tumor, tuberculosis, or certain congenital abnormalities. Dark brown fluid can reflect a high bilirubin concentration and indicates biliary tract perforation. Black fluid may indicate the presence of pancreatic necrosis or metastatic melanoma.

The ascitic fluid should be sent for measurement of albumin and total protein levels, cell and differential counts, and, if infection is suspected, Gram's stain and culture, with inoculation into blood culture bottles at the patient's bedside to maximize the yield. A serum albumin level should be measured simultaneously to permit calculation of the *serum-ascites albumin gradient* (SAAG).

The SAAG is useful for distinguishing ascites caused by portal hypertension from nonportal hypertensive ascites **(Fig. 46-3)**. The SAAG reflects the pressure within the hepatic sinusoids and correlates with the hepatic venous pressure gradient. The SAAG is calculated by subtracting the ascitic albumin concentration from the serum albumin level and does not change with diuresis. A SAAG ≥1.1 g/dL reflects the presence of portal hypertension and indicates that the ascites is due to increased pressure in the hepatic sinusoids. According to Starling's law, a high SAAG reflects the oncotic pressure that counterbalances the portal pressure. Possible causes include cirrhosis, cardiac ascites, hepatic vein thrombosis (Budd-Chiari syndrome), sinusoidal obstruction syndrome (veno-occlusive disease), or massive liver metastases. A SAAG <1.1 g/dL indicates that the ascites is not related to portal hypertension, as in tuberculous peritonitis, peritoneal carcinomatosis, or pancreatic ascites.

FIGURE 46-3

Algorithm for the diagnosis of ascites according to the serum-ascites albumin gradient (SAAG). IVC, inferior vena cava.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

For high-SAAG (≥1.1) ascites, the ascitic protein level can provide further clues to the etiology (Fig. 46-3). An ascitic protein level of ≥2.5 g/dL indicates that the hepatic sinusoids are normal and are allowing passage of protein into the ascites, as occurs in cardiac ascites, early Budd-Chiari syndrome, or sinusoidal obstruction syndrome. An ascitic protein level <2.5 g/dL indicates that the hepatic sinusoids have been damaged and scarred and no longer allow passage of protein, as occurs with cirrhosis, late Budd-Chiari syndrome, or massive liver metastases. Pro-brain-type natriuretic peptide (BNP) is a natriuretic hormone released by the heart as a result of increased volume and ventricular wall stretch. High levels of BNP in serum occur in heart failure and may be useful in identifying heart failure as the cause of high-SAAG ascites.

Further tests are indicated only in specific clinical circumstances. When secondary peritonitis resulting from a perforated hollow viscus is suspected, ascitic glucose and lactate dehydrogenase (LDH) levels can be measured. In contrast to "spontaneous" bacterial peritonitis, which may complicate cirrhotic ascites (see "Complications," below), secondary peritonitis is suggested by an ascitic glucose level <50 mg/dL, an ascitic LDH level higher than the serum LDH level, and the detection of multiple pathogens on ascitic fluid culture. When pancreatic ascites is suspected, the ascitic amylase level should be measured and is typically >1000 mg/dL. Cytology can be useful in the diagnosis of peritoneal carcinomatosis. At least 50 mL of fluid should be obtained and sent for immediate processing. Tuberculous peritonitis is typically associated with ascitic fluid lymphocytosis but can be difficult to diagnose by paracentesis. A smear for acid-fast bacilli has a diagnostic sensitivity of only 0 to 3%; a culture increases the sensitivity to 35–50%. In patients without cirrhosis, an elevated ascitic adenosine deaminase level has a sensitivity of >90% when a cut-off value of 30–45 U/L is used. When the cause of ascites remains uncertain, laparotomy or laparoscopy with peritoneal biopsies for histology and culture remains the gold standard.

TREATMENT

TREATMENT

Ascites

The initial treatment for cirrhotic ascites is restriction of sodium intake to 2 g/d. When sodium restriction alone is inadequate to control ascites, oral diuretics—typically the combination of spironolactone and furosemide—are used. Spironolactone is an aldosterone antagonist that inhibits sodium resorption in the distal convoluted tubule of the kidney. Use of spironolactone may be limited by hyponatremia, hyperkalemia, and painful gynecomastia. If the gynecomastia is distressing, amiloride (5–40 mg/d) may be substituted for spironolactone. Furosemide is a loop diuretic that is generally combined with spironolactone in a ratio of 40:100; maximal daily doses of spironolactone and furosemide are 400 mg and 160 mg, respectively. Fluid intake may be restricted in patients with hyponatremia.

Refractory cirrhotic ascites is defined by the persistence of ascites despite sodium restriction and maximal (or maximally tolerated) diuretic use. Pharmacologic therapy for refractory ascites includes the addition of midodrine, an α_1-adrenergic agonist, or clonidine, an α_2-adrenergic agonist, to diuretic therapy. These agents act as vasoconstrictors, counteracting splanchnic vasodilation. Midodrine alone or in combination

with clonidine improves systemic hemodynamics and control of ascites over that obtained with diuretics alone. Although β-adrenergic blocking agents (beta blockers) are often prescribed to prevent variceal hemorrhage in patients with cirrhosis, the use of beta blockers in patients with refractory ascites may be associated with decreased survival rates.

When medical therapy alone is insufficient, refractory ascites can be managed by repeated large-volume paracentesis (LVP) or a transjugular intrahepatic peritoneal shunt (TIPS)—a radiologically placed portosystemic shunt that decompresses the hepatic sinusoids. Intravenous infusion of albumin accompanying LVP decreases the risk of "post-paracentesis circulatory dysfunction" and death. Patients undergoing LVP should receive IV albumin infusions of 6–8 g/L of ascitic fluid removed. TIPS placement is superior to LVP in reducing the reaccumulation of ascites but is associated with an increased frequency of hepatic encephalopathy, with no difference in mortality rates.

Malignant ascites does not respond to sodium restriction or diuretics. Patients must undergo serial LVPs, transcutaneous drainage catheter placement, or, rarely, creation of a peritoneovenous shunt (a shunt from the abdominal cavity to the vena cava).

Ascites caused by tuberculous peritonitis is treated with standard antituberculosis therapy. Noncirrhotic ascites of other causes is treated by correction of the precipitating condition.

COMPLICATIONS

Spontaneous bacterial peritonitis (SBP; **Chap. 127**) is a common and potentially lethal complication of cirrhotic ascites. Occasionally, SBP also complicates ascites caused by nephrotic syndrome, heart failure, acute hepatitis, and acute liver failure but is rare in malignant ascites. Patients with SBP generally note an increase in abdominal girth; however, abdominal tenderness is found in only 40% of patients, and rebound tenderness is uncommon. Patients may present with fever, nausea, vomiting, or the new onset of or exacerbation of preexisting hepatic encephalopathy.

In hospitalized patients with ascites, paracentesis within 12 hours of admission reduces mortality because of early detection of SBP. SBP is defined by a polymorphonuclear neutrophil (PMN) count of ≥250/μL in the ascitic fluid. Cultures of ascitic fluid typically reveal one bacterial pathogen. The presence of multiple pathogens in the setting of an elevated ascitic PMN count suggests *secondary peritonitis* from a ruptured viscus or abscess (**Chap. 127**). The presence of multiple pathogens without an elevated PMN count suggests bowel perforation from the paracentesis needle. SBP is generally the result of enteric bacteria that have translocated across an edematous bowel wall. The most common pathogens are gram-negative rods, including *Escherichia coli* and *Klebsiella*, as well as streptococci and enterococci.

Treatment of SBP with an antibiotic such as IV cefotaxime is effective against gram-negative and gram-positive aerobes. A 5-day course of treatment is sufficient if the patient improves clinically. Nosocomial or health care–acquired SBP is frequently caused by multidrug-resistant bacteria, and initial antibiotic therapy should be guided by the local bacterial epidemiology.

Cirrhotic patients with a history of SBP, an ascitic fluid total protein concentration <1 g/dL, or active gastrointestinal bleeding should receive prophylactic antibiotics to prevent SBP; oral daily norfloxacin is commonly used. Diuresis increases the activity of ascitic fluid protein opsonins and may decrease the risk of SBP.

Hepatic hydrothorax occurs when ascites, often caused by cirrhosis, migrates via fenestrae in the diaphragm into the pleural space. This condition can result in shortness of breath, hypoxia, and infection. Treatment is similar to that for cirrhotic ascites and includes sodium restriction, diuretics, and, if needed, thoracentesis or TIPS placement. Chest tube placement should be avoided.

FURTHER READING

Becker G et al: Malignant ascites: Systematic review and guideline for treatment. Eur J Cancer 42:589, 2006. [PubMed: 16434188]

Bernardi M et al: Albumin infusion in patients undergoing large-volume paracentesis: A meta-analysis of randomized trials. Hepatology 55:1172, 2012. [PubMed: 22095893]

Farias AQ et al: Serum B-type natriuretic peptide in the initial workup of patients with new onset ascites: A diagnostic accuracy study. Hepatology 59:1043, 2014. [PubMed: 23907731]

Fernandez J et al: Prevalence and risk factors of infections by multiresistant bacteria in cirrhosis: A prospective study. Hepatology 55:1551, 2012. [PubMed: 22183941]

Ge PS, Runyon BA: Role of plasma BNP in patients with ascites: Advantages and pitfalls. Hepatology 59: 751, 2014. [PubMed: 23959954]

Orman ES et al: Paracentesis is associated with reduced mortality in patients hospitalized with cirrhosis and ascites. Clin Gastroenterol Hepatol 12:496, 2014. [PubMed: 23978348]

Runyon BA: Introduction to the revised American Association for the Study of Liver Diseases Practice Guideline management of adult patients with ascites due to cirrhosis 2012. Hepatology 57:165, 2013.

Runyon BA et al: The serum-ascites albumin gradient is superior to the exudate-transudate concept in the differential diagnosis of ascites. Ann Intern Med 117:215, 1992. [PubMed: 1616215]

Sort P et al: Effect of intravenous albumin on renal impairment and mortality in patients with cirrhosis and spontaneous bacterial peritonitis. N Engl J Med 341:403, 1999. [PubMed: 10432325]

Williams JW Jr, Simel DL: The rational clinical examination. Does this patient have ascites? How to divine fluid in the abdomen. JAMA 267:2645, 1992. [PubMed: 1573754]

Chapter 47: Dysuria, Bladder Pain, and the Interstitial Cystitis/Bladder Pain Syndrome

John W. Warren

INTRODUCTION

Dysuria and bladder pain are two symptoms that commonly call attention to the lower urinary tract.

DYSURIA

Dysuria, or pain that occurs during urination, is commonly perceived as burning or stinging in the urethra and is a symptom of several syndromes. The presence or absence of *other* symptoms is often helpful in distinguishing among these conditions. Some of these syndromes differ between men and women.

Women

Approximately 50% of women experience dysuria at some time in their lives; ~20% report having had dysuria within the past year. Most dysuria syndromes in women can be categorized into two broad groups: bacterial cystitis and lower genital tract infections.

Bacterial cystitis is usually caused by *Escherichia coli*; a few other gram-negative rods and *Staphylococcus saprophyticus* also can be responsible. Bacterial cystitis is acute in onset and manifests not only as dysuria but also as urinary frequency, urinary urgency, suprapubic pain, and/or hematuria.

The lower genital tract infections include vaginitis, urethritis, and ulcerative lesions; many of these infections are caused by sexually transmitted organisms and should be considered particularly in young women who have new or multiple sexual partners or whose partners do not use condoms. The onset of dysuria associated with these syndromes is more gradual than in bacterial cystitis and is thought (but not proven) to result from the flow of urine over damaged epithelium. Frequency, urgency, suprapubic pain, and hematuria are reported less frequently than in bacterial cystitis. Vaginitis, caused by *Candida albicans* or *Trichomonas vaginalis*, presents as vaginal discharge or irritation. Urethritis is a consequence of infection by *Chlamydia trachomatis* or *Neisseria gonorrhoeae*. Ulcerative genital lesions may be caused by herpes simplex virus and several other specific organisms.

Among women presenting with dysuria, the probability of bacterial cystitis is ~50%. This figure rises to >90% if four criteria are met: dysuria and frequency without vaginal discharge or irritation. Present standards suggest that women meeting these four criteria, if they are otherwise healthy, are not pregnant, and have an apparently normal urinary tract, can be diagnosed with uncomplicated bacterial cystitis and treated empirically with appropriate antibiotics. Other women with dysuria should be further evaluated by urine dipstick, urine culture, and a pelvic examination.

Men

Dysuria is less common among men. The syndromes presenting as dysuria are similar to those in women but with some important distinctions.

In the majority of men with dysuria, frequency, urgency, and/or suprapubic, penile, and/or perineal pain, the prostate is involved, either as the source of infection or as an obstruction to urine flow. Bacterial prostatitis is usually caused by *E. coli* or another gram-negative rod, with one of two presentations. *Acute bacterial prostatitis* presents with fever and chills; prostate examination should be gentle or not performed at all, as massage may result in a wave of bacteremia. *Chronic bacterial prostatitis* presents as recurrent episodes of bacterial cystitis; prostate examination with massage demonstrates prostatic bacteria and leukocytes. *Benign prostatic hyperplasia* (BPH) can obstruct urine flow, with consequent symptoms of weak stream, hesitancy, and dribbling. If a bacterial infection develops behind the obstructing prostate, dysuria and other symptoms of cystitis will occur. Men whose symptoms are consistent with bacterial cystitis should be evaluated with urinalysis and urine culture.

Several sexually transmitted infections can manifest as dysuria. Urethritis (usually without urinary frequency) presents as a urethral discharge and can be caused by *C. trachomatis*, *N. gonorrhoeae*, *Mycoplasma genitalium*, *Ureaplasma urealyticum*, or *T. vaginalis*. Herpes simplex, chancroid, and other ulcerous lesions may present as dysuria, again without urinary frequency.

For further discussion, see Chaps. 130 and 131.

Either Women or Men

Other causes of dysuria may be found in patients of either sex. Some cases are acute and include lower urinary tract stones, trauma, and urethral exposure to topical chemicals. Others may be relatively chronic and attributable to lower urinary tract cancers, certain medications, Behçet's syndrome, reactive arthritis, a poorly understood entity known as *chronic urethral syndrome*, or interstitial cystitis/bladder pain syndrome (see below).

BLADDER PAIN

Studies indicate that patients perceive pain as coming from the urinary bladder if it is suprapubic in location, alters with bladder filling or emptying, and/or is associated with urinary symptoms such as urgency and

frequency. Bladder pain occurring acutely (i.e., over hours or a day or two) is helpful in distinguishing bacterial cystitis from urethritis, vaginitis, and other genital infections. Chronic or recurrent bladder pain may accompany lower urinary tract stones; bladder, uterine, cervical, vaginal, urethral, or prostate cancer; urethral diverticulum; cystitis induced by radiation or certain medications; tuberculous cystitis; bladder neck obstruction; neurogenic bladder; urogenital prolapse; or BPH. In the absence of these conditions, the diagnosis of interstitial cystitis/bladder pain syndrome (IC/BPS) should be considered.

INTERSTITIAL CYSTITIS/BLADDER PAIN SYNDROME

Most clinicians with outpatient practices see undiagnosed cases of IC/BPS. This chronic condition is characterized by pain perceived to be from the urinary bladder, urinary urgency and frequency, and nocturia. As currently diagnosed, the majority of cases occur in women. Symptoms wax and wane for months or years or possibly even for the rest of the patient's life. The spectrum of symptom intensity is broad. The pain can be excruciating, urgency can be distressing, frequency can be up to 60 times per 24 h, and nocturia can cause sleep deprivation. These symptoms can disrupt daily activities, work schedules, and personal relationships; patients with IC/BPS report less life satisfaction than do those with end-stage renal disease.

IC/BPS is not a new disease, having first been described in the late nineteenth century in a patient with the symptoms described above and a single ulcer visible on cystoscopy (now called a *Hunner lesion* after the urologist who first reported it). Over the ensuing decades, it became clear that many patients with similar symptoms had no ulcer. It is now appreciated that ≤10% of patients with IC/BPS have a Hunner lesion. The definition of IC/BPS, its diagnostic features, and even its name continue to evolve. The American Urological Association has defined IC/BPS as "an unpleasant sensation (pain, pressure, discomfort) perceived to be related to the urinary bladder, associated with lower urinary tract symptoms of more than 6 weeks' duration, in the absence of infection or other identifiable causes."

Many patients with IC/BPS also have other syndromes, such as fibromyalgia, chronic fatigue syndrome, and irritable bowel syndrome. These syndromes collectively are known as *functional somatic syndromes* (FSSs): chronic conditions in which pain and fatigue are prominent features but laboratory tests and histologic findings are normal. Like IC/BPS, the FSSs often are associated with depression and anxiety. The majority of FSSs affect more women than men, and more than one FSS can affect a single patient. Because of its similar features and comorbidity, IC/BPS sometimes is considered an FSS.

Epidemiology

Contemporary population studies of IC/BPS in the United States indicate a prevalence of 3–6% among women and 2–4% among men. For decades, it was thought that IC/BPS occurred mostly in women. These prevalence findings, however, have generated research aimed at determining the proportion of men who have symptoms usually diagnosed as chronic prostatitis (now known as *chronic prostatitis/chronic pelvic pain syndrome*) but who actually have IC/BPS.

Among women, the average age at onset of IC/BPS symptoms is the early forties, but the range is from childhood through the early sixties. Risk factors (antecedent features that distinguish cases from controls) primarily have been FSSs. Indeed, the odds of IC/BPS increase with the number of such syndromes present. Surgery was long thought to be a risk factor for IC/BPS, but analyses adjusting for FSSs refuted that association. About one-third of patients appear to have bacterial cystitis at the onset of IC/BPS.

The natural history of IC/BPS is not known. Although studies from urology and urogynecology practices have been interpreted as showing that IC/BPS lasts for the lifetime of the patient, population studies suggest that some individuals with IC/BPS do not consult specialists and may not seek medical care at all, and most prevalence studies do not show an upward trend with age—a pattern that would be expected with incident cases throughout adulthood followed by lifetime persistence of a nonfatal disease. It may be reasonable to conclude that patients in a urology practice represent those with the most severe and recalcitrant IC/BPS.

Pathology

For the ≤10% of IC/BPS patients who have a Hunner lesion, the term *interstitial cystitis* may indeed describe the histopathologic picture. Most of these patients have substantive inflammation, mast cells, and granulation tissue. However, in the 90% of patients without such lesions, the bladder mucosa and interstitium are relatively normal, with scant inflammation.

Etiology

Numerous hypotheses about the pathogenesis of IC/BPS have been put forward. It is not surprising that most early theories focused on the bladder. For instance, IC/BPS has been investigated as a chronic bladder infection. Sophisticated technologies have not identified a causative organism in urine or in bladder tissue; however, the patients studied by these methods had IC/BPS of long duration, and the results do not preclude the possibility that infection may trigger the syndrome or may be a feature of early IC/BPS. Other inflammatory factors, including a role for mast cells, have been postulated, but (as noted above) the 90% of patients who do not have a Hunner ulcer have little bladder inflammation and do not have a prominence of mast cells in urine or in bladder tissue. Autoimmunity has been considered, but autoantibodies are low in titer, nonspecific, and thought to be a result rather than a cause of IC/BPS. Increased permeability of the bladder mucosa due to defective epithelium or glycosaminoglycan (the bladder's mucous coating) has been studied frequently, but the findings have been inconclusive.

Investigations of causes outside the bladder have been prompted by the presence of comorbid FSSs. Many patients with FSSs have abnormal pain sensitivity as evidenced by (1) low pain thresholds in body areas unrelated to the diagnosed syndrome, (2) dysfunctional descending neurologic control of tactile signals, and (3) enhanced brain responses to touch in functional neuroimaging studies. Moreover, in patients with IC/BPS, body surfaces remote from the bladder are more sensitive to pain than is the case in individuals without IC/BPS. All these findings are consistent with upregulation of sensory processing in the brain. Indeed, a prevailing theory is that these concomitantly occurring syndromes have in common an abnormality of brain

processing of sensory input. However, antecedence is a critical criterion for causality, and no study has demonstrated that abnormal pain sensitivity precedes either IC/BPS or the FSSs.

Clinical Presentation

In some patients, IC/BPS has a gradual onset and/or the cardinal symptoms of pain, urgency, frequency, and nocturia appear sequentially in no consistent order. Other patients can identify the exact date of onset of IC/BPS symptoms. More than half of the latter patients describe dysuria beginning on that date. Only a minority of IC/BPS patients who obtain medical care soon after symptom onset have uropathogenic bacteria or leukocytes in the urine. These patients—and many others with new-onset IC/BPS—are treated with antibiotics for presumptive bacterial cystitis or, if male, chronic bacterial prostatitis. Persistent or recurring symptoms without bacteriuria eventually prompt a differential diagnosis, and IC/BPS is considered. Traditionally, the diagnosis of IC/BPS has been delayed for years, but recent interest in the disease appears to have shortened this interval.

Two-thirds of women with IC/BPS report two or more sites of pain. The most common site (involved in 80% of women) and generally the one with the most severe pain is the suprapubic area. About 35% of female patients have pain in the urethra, 25% in other parts of the vulva, and 30% in non-urogenital areas, mostly the low back and also the anterior or posterior thighs or the buttocks. The pain of IC/BPS is most commonly described as aching, pressing, throbbing, tender, and/or piercing. What may distinguish IC/BPS from other pelvic pain is that, in 95% of patients, bladder filling exacerbates the pain and/or bladder emptying relieves it. Almost as many patients report a puzzling pattern in which certain dietary substances worsen the pain of IC/BPS. Smaller majorities report that their IC/BPS pain is worsened by menstruation, stress, tight clothing, exercise, and riding in a car as well as during or after vaginal intercourse.

The urethral and vulvar pains of IC/BPS merit special mention. In addition to the descriptive adjectives for IC/BPS mentioned above, these pains commonly are described as burning, stinging, and sharp and as being worsened by touch, tampons, and vaginal intercourse. Patients report that urethral pain increases during urination and generally lessens afterward. These characteristics have commonly resulted in the diagnosis of the urethral pain of IC/BPS as chronic urethral syndrome and of the vulvar pain as vulvodynia.

In many patients with IC/BPS, there is a link between pain and urinary urgency; that is, two-thirds of patients describe the urge to urinate as a desire to relieve their bladder pain. Only 20% report that the urge stems from a desire to prevent incontinence; indeed, very few patients with IC/BPS are incontinent. As mentioned above, urinary frequency can be severe, with ~85% of patients voiding more than 10 times per 24 h and some as often as 60 times. Voiding continues through the night, and nocturia is common, frequent, and often associated with sleep deprivation.

Beyond these common symptoms of IC/BPS, additional urinary and other symptoms may be present. Among the urinary symptoms are difficulty in starting urine flow, perceptions of difficulty in emptying the bladder, and bladder spasms. Among the non-urinary symptoms are the manifestations of comorbid FSSs as well as

symptoms that do not constitute recognized syndromes, such as numbness, muscle spasms, dizziness, ringing in the ears, and blurred vision.

The pain, urgency, and frequency of IC/BPS can be debilitating. Proximity to a bathroom is a continual focus, and patients report difficulties in the workplace, leisure activities, travel, and simply leaving home. Familial and sexual relationships can be strained.

Diagnosis

Traditionally, IC/BPS has been considered a rare condition that is diagnosed by urologists at cystoscopy. However, this disorder is much more common than once was thought; it is now being considered earlier in its course and is being diagnosed and managed more often by primary care clinicians. Results of physical examination, urinalysis, and urologic procedures are insensitive and/or nonspecific. Thus, diagnosis is based on the presence of appropriate symptoms and the exclusion of diseases with a similar presentation.

Three categories of disorders can be considered in the differential diagnosis of IC/BPS. The first comprises diseases that manifest as bladder pain or urinary symptoms. Among the latter diseases is *overactive bladder*, a chronic condition of women and men that presents as urgency and frequency and can be distinguished from IC/BPS by the patient's history: pain is not a feature of overactive bladder, and its urgency arises from the need to avoid incontinence. Endometriosis is a special case: it can be asymptomatic or can cause pelvic pain, dysmenorrhea, and dyspareunia—i.e., types of pain that mimic IC/BPS. Endometrial implants on the bladder (although uncommon) can cause urinary symptoms, and the resulting syndrome can mimic IC/BPS. Even if endometriosis is identified, it is difficult in the absence of bladder implants to determine whether it is causative of or incidental to the symptoms of IC/BPS in a specific woman.

The second category of disorders encompasses the FSSs that can accompany IC/BPS. IC/BPS can be misdiagnosed as gynecologic chronic pelvic pain, irritable bowel syndrome, or fibromyalgia. The correct diagnosis may be entertained only when either changes of pain with altered bladder volume or urinary symptoms become more prominent.

The third category involves syndromes that IC/BPS mimics by way of its referred pain, such as vulvodynia and chronic urethral syndrome. Therefore, IC/BPS should be considered in the differential diagnosis of persistent or recurrent "urinary tract infection" (UTI) with sterile urine cultures; "overactive bladder" with pain; chronic pelvic pain, endometriosis, vulvodynia, or FSSs with urinary symptoms; and "chronic prostatitis." Important clues to the diagnosis of IC/BPS are changes of pain with bladder volume or with certain foods or drinks.

Cystoscopy under anesthesia formerly was thought to be necessary for the diagnosis of IC/BPS because of its capacity to reveal a Hunner lesion or—in the 90% of patients without an ulcer—petechial hemorrhages after bladder distention. However, because Hunner lesions are uncommon in IC/BPS and petechiae are nonspecific, cystoscopy is no longer necessary for diagnosis. Accordingly, the indications for urologic referral have evolved toward the need to rule out other diseases or to administer more advanced treatment.

A typical patient presents to the primary clinician after days, weeks, or months of pain, urgency, frequency, and/or nocturia. The presence of urinary nitrites, leukocytes, or uropathogenic bacteria should prompt treatment for UTI in women and for chronic bacterial prostatitis in men. Persistence or recurrence of symptoms in the absence of bacteriuria should prompt a pelvic examination for women, an assay for serum prostate-specific antigen for men, and urine cytology and inclusion of IC/BPS in the differential diagnosis for both sexes.

In the diagnosis of IC/BPS, inquiries about pain, pressure, and discomfort are useful; IC/BPS should be considered if any of these sensations are noted in one or more anterior or posterior sites between the umbilicus and the upper thighs. Nondirective questions about the effect of bladder volume changes include “As your next urination approaches, does this pain get better, get worse, or stay the same?” and “After you urinate, does this pain get better, get worse, or stay the same?” Establishing that the pain is exacerbated by the consumption of certain foods and drinks not only supports the diagnosis of IC/BPS but also serves as the basis for one of the first steps in managing this syndrome. A nondirective way to ask about urgency is to describe it to the patient as a compelling urge to urinate that is difficult to postpone; follow-up questions can determine whether this urge is intended to relieve pain or prevent incontinence. To assess severity and provide quantitative baseline measures, pain and urgency should be estimated by the patient on a scale of 0–10, with 0 being none and 10 the worst imaginable. Frequency per 24-h period should be determined and nocturia assessed as the number of times per night the patient is awakened by the need to urinate.

About half of patients with IC/BPS have intermittent or persistent microscopic hematuria; this manifestation and the need to exclude bladder stones or cancer require urologic or urogynecologic referral. Initiation of therapy for IC/BPS does not hamper subsequent urologic evaluation.

TREATMENT

TREATMENT

Interstitial Cystitis/Bladder Pain Syndrome

The goal of therapy is to relieve the symptoms of IC/BPS; the challenge lies in the fact that no treatment is uniformly successful. However, most patients eventually obtain relief, generally with a multifaceted approach. The American Urological Association’s guidelines for management of IC/BPS are an excellent resource. The correct strategy is to begin with conservative therapies and proceed to riskier measures only if necessary and under the supervision of a urologist or urogynecologist. Conservative tactics include education, stress reduction, dietary changes, medications, pelvic-floor physical therapy, and treatment of associated FSSs.

Months or even years may have passed since the onset of symptoms, and the patient’s life may have been disrupted continually, with repeated medical visits provoking frustration and dismay in both the patient and the physician. In this circumstance, simply giving a name to the syndrome is beneficial. The physician should discuss the disease, the diagnostic and therapeutic strategies, and the prognosis with the patient and with the spouse and/or other pertinent family members, who may need to be made aware that although IC/BPS

has no visible manifestations, the patient is undergoing substantial pain and suffering. This information is particularly important for sexual partners, as exacerbation of pain during and after intercourse is a common feature of IC/BPS. Because stress can worsen IC/BPS symptoms, stress reduction and active measures such as yoga or meditation exercises may be suggested. The Interstitial Cystitis Association (*www.ichelp.com*) and the Interstitial Cystitis Network (*www.ic-network.com*) can be useful in this educational process.

Over time, many patients identify particular foods and drinks that exacerbate their symptoms. Common among these are chilies, chocolate, citrus fruits, tomatoes, alcohol, caffeinated drinks, and carbonated beverages; full lists of common trigger foods are available at the websites cited above. In constructing a benign diet, some patients find it useful to exclude all possible offenders and add items back into the diet one at a time to identify those that worsen their symptoms. Patients also should experiment with fluid volumes; some find relief with less fluid, others with more.

The pelvic floor is often tender in IC/BPS patients. Two randomized controlled trials showed that weekly physical therapy directed at relaxation of the pelvic muscles yielded significantly more relief than a similar schedule of general body massage. This intervention can be initiated under the direction of a knowledgeable physical therapist who recognizes that the objective is to relax the pelvic floor, not to strengthen it.

Among oral medications, nonsteroidal anti-inflammatory drugs are commonly used but are controversial and often unsuccessful. Two randomized controlled trials showed that amitriptyline can diminish IC/BPS symptoms if an adequate dose (≥50 mg per night) can be given. This drug is used not for its antidepressant activity but because of its proven effects on neuropathic pain; however, it is not approved by the U.S. Food and Drug Administration for treatment of IC/BPS. An initial dose of 10 mg at bedtime is increased weekly up to 75 mg (or less if a lower dose adequately relieves symptoms). Side effects can be expected and include dry mouth, weight gain, sedation, and constipation. If this regimen does not control symptoms adequately, pentosan polysulfate, a semisynthetic polysaccharide, can be added at a dose of 100 mg three times a day. Its theoretical effect is to replenish a possibly defective glycosaminoglycan layer over the bladder mucosa; randomized controlled trials suggest only a modest benefit over placebo. Adverse reactions are uncommon and include gastrointestinal symptoms, headache, and alopecia. Pentosan polysulfate has weak anticoagulant effects and probably should be avoided by patients with coagulation abnormalities.

Anecdotal reports suggest that successful therapy for one FSS is accompanied by diminished symptoms of other FSSs. As has been noted here, IC/BPS often is associated with one or several FSSs. Thus, it seems reasonable to hope that, to the extent that accompanying FSSs are treated successfully, the symptoms of IC/BPS will be relieved as well.

If several months of these therapies in combination do not relieve symptoms adequately, the patient should be referred to a urologist or urogynecologist who has access to additional modalities. Cystoscopy under anesthesia allows distention of the bladder with water, a procedure that provides ~40% of patients with several months of relief and can be repeated. For those few patients with a Hunner lesion, fulguration may offer relief. Solutions containing lidocaine, hyaluronic acid, or dimethyl sulfoxide can be instilled into the bladder, or botulinum toxin can be injected into the bladder wall. Physicians experienced in the care of

IC/BPS patients have used anticonvulsants, narcotics, and cyclosporine as components of therapy. Pain specialists can be of assistance. Sacral neuromodulation can be tested with a temporary percutaneous electrode and, if effective, can be administered with an implanted device. In a very small number of patients with recalcitrant symptoms, surgeries, including cystoplasty, partial or total cystectomy, and urinary diversion, may provide relief.

FURTHER READING

Fitzgerald MP et al: Randomized multicenter clinical trial of myofascial physical therapy in women with interstitial cystitis/painful bladder syndrome and pelvic floor tenderness. J Urol 187:2113, 2012. [PubMed: 22503015]

Hanno PM et al: AUA guideline for the diagnosis and treatment of interstitial cystitis/bladder pain syndrome. J Urol 185:2162, 2011. [PubMed: 21497847]

Hanno PM et al: Diagnosis and treatment of interstitial cystitis/bladder pain syndrome: AUA guideline amendment. J Urol 193:1545, 2015. [PubMed: 25623737]

Shorter B et al: Effect of comestibles on symptoms of interstitial cystitis. J Urol 178:145, 2007. [PubMed: 17499305]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 48: Azotemia and Urinary Abnormalities

David B. Mount

INTRODUCTION

Normal kidney functions occur through numerous cellular processes to maintain body homeostasis. Disturbances in any of these functions can lead to abnormalities that may be detrimental to survival. Clinical manifestations of these disorders depend on the pathophysiology of renal injury and often are identified as a complex of symptoms, abnormal physical findings, and laboratory changes that constitute specific syndromes. These renal syndromes **(Table 48-1)** may arise from systemic illness or as primary renal disease. Nephrologic syndromes usually consist of several elements that reflect the underlying pathologic processes, typically including one or more of the following: (1) reduction in glomerular filtration rate (GFR), (2) abnormalities of urine sediment (red blood cells [RBCs], white blood cells [WBCs], casts, and crystals), (3) abnormal excretion of serum proteins (proteinuria), (4) disturbances in urine volume (oliguria, anuria, polyuria), (5) presence of hypertension and/or expanded total body fluid volume (edema), (6) electrolyte abnormalities, and (7) in some syndromes, fever/pain. The specific combination of these findings should permit identification of one of the major nephrologic syndromes (Table 48-1) and allow differential diagnoses to be narrowed so that the appropriate diagnostic and therapeutic course can be determined. All these syndromes and their associated diseases are discussed in more detail in subsequent chapters. This chapter focuses on several aspects of renal abnormalities that are critically important for distinguishing among those processes: (1) reduction in GFR, (2) alterations of the urinary sediment and/or protein excretion, and (3) abnormalities of urinary volume.

TABLE 48-1

Initial Clinical and Laboratory Database for Defining Major Syndromes in Nephrology

Syndrome	Important Clues to Diagnosis	Common Findings	Chap(s). Discussing Disease-Causing Syndrome
Acute or rapidly progressive renal failure	Anuria	Hypertension, hematuria	304, 308, 310, 313
	Oliguria	Proteinuria, pyuria	
	Documented recent decline in GFR	Casts, edema	
Acute nephritis	Hematuria, RBC casts	Proteinuria	308
	Azotemia, reduced GFR, oliguria	Pyuria	
	Edema, hypertension	Circulatory congestion	
Chronic renal failure	Azotemia for >3 months	Proteinuria, casts	305
	Symptoms or signs of uremia, (late manifestation), casts	Hypocalcemia, hyperphosphatemia, hyperparathyroidism	
	Symptoms or signs of renal osteodystrophy	Polyuria, nocturia	
	Kidneys reduced in size bilaterally	Edema, hypertension	
	Broad casts in urinary sediment	Hyperkalemia, metabolic acidosis	
Nephrotic syndrome	Proteinuria, with >3.5 g/24 h per 1.73 m^2	Casts	308
	Hypoalbuminemia	Lipiduria	
	Edema	Hypercoagulable state	

Syndrome	Important Clues to Diagnosis	Common Findings	Chap(s). Discussing Disease-Causing Syndrome
	Hyperlipidemia		
Asymptomatic urinary abnormalities	Hematuria		308
	Proteinuria (below nephrotic range)		
	Sterile pyuria, casts		
Urinary tract infection/pyelonephritis	Bacteriuria, with $>10^5$ cfu/mL	Hematuria	130
	Other infectious agent documented in urine	Mild azotemia and reduced GFR	
	Pyuria, leukocyte casts	Mild proteinuria	
	Frequency, urgency	Fever	
	Bladder tenderness, flank tenderness		
Renal tubular defects	Electrolyte disorders	Hematuria	309, 310
	Polyuria, nocturia	“Tubular” proteinuria (<1 g/24 h)	
	Renal calcification	Enuresis	
	Large kidneys	Electrolyte and/or acid-base abnormalities	
	Renal transport defects	Other electrolyte issues, e.g. hypomagnesemia	
Hypertension	Systolic/diastolic hypertension	Proteinuria	271, 311

Syndrome	Important Clues to Diagnosis	Common Findings	Chap(s). Discussing Disease-Causing Syndrome
		Casts	
		Azotemia	
Nephrolithiasis	Previous history of stone passage or removal	Hematuria	312
	Previous history of stone seen by x-ray	Pyuria	
	Renal colic	Frequency, urgency	
Urinary tract obstruction	Azotemia, oliguria, anuria	Hematuria	313
	Polyuria, nocturia, urinary retention	Pyuria	
	Slowing of urinary stream	Enuresis, dysuria	
	Large prostate, large kidneys		
	Flank tenderness, full bladder after voiding		

Abbreviations: cfu, colony-forming units; GFR; glomerular filtration rate; RBC, red blood cell.

AZOTEMIA

ASSESSMENT OF GFR

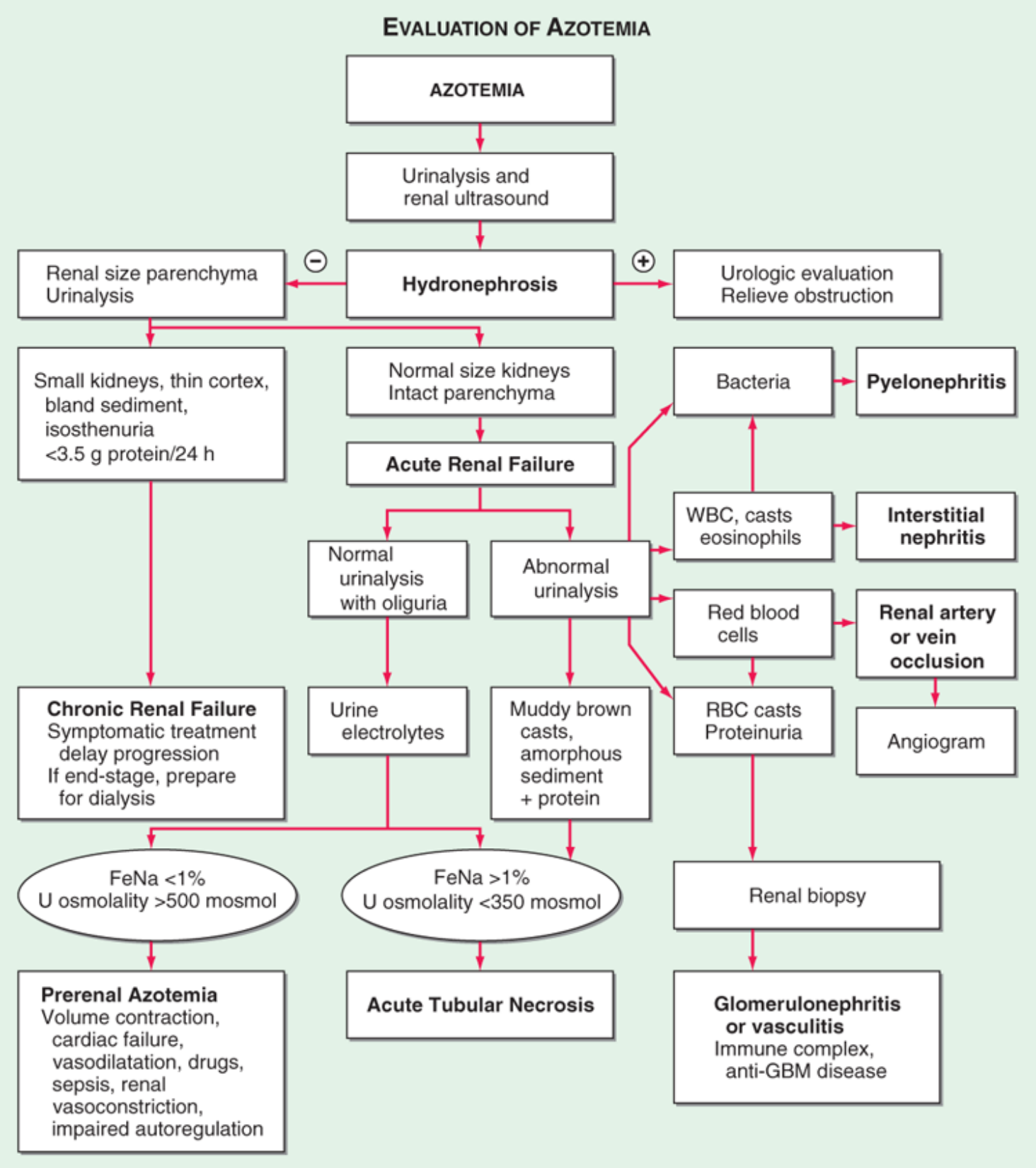
Monitoring the GFR is important in both hospital and outpatient settings, and several different methodologies are available. GFR is the primary metric for kidney "function," and its direct measurement involves administration of a radioactive isotope (such as inulin or iothalamate) that is filtered at the glomerulus into the urinary space but is neither reabsorbed nor secreted throughout the tubule. GFR—i.e., the clearance of inulin or iothalamate in milliliters per minute—is calculated from the rate of appearance of the isotope in the urine over several hours. In most clinical circumstances, direct GFR measurement is not feasible, and the plasma creatinine

level is used as a surrogate to estimate GFR. Plasma creatinine (P_{Cr}) is the most widely used marker for GFR, which is related directly to urine creatinine (U_{Cr}) excretion and inversely to P_{Cr}. On the basis of this relationship (with some important caveats, as discussed below), GFR will fall in roughly inverse proportion to the rise in P_{Cr}. Failure to account for GFR reductions in drug dosing can lead to significant morbidity and death from drug toxicities (e.g., digoxin, imipenem). In the outpatient setting, P_{Cr} serves as an estimate for GFR (although much less accurate; see below). In patients with chronic progressive renal disease, there is an approximately linear relationship between $1/P_{Cr}$ (*y* axis) and time (*x* axis). The slope of that line will remain constant for an individual; when values deviate, an investigation for a superimposed acute process (e.g., volume depletion, drug reaction) should be initiated. Signs and symptoms of uremia, the clinical symptom complex associated with renal failure, develop at significantly different levels of P_{Cr}, depending on the patient (size, age, and sex), underlying renal disease, existence of concurrent diseases, and true GFR. Generally, patients do not develop symptomatic uremia until renal insufficiency is severe (GFR <15 mL/min).

A significantly reduced GFR (either acute or chronic) is usually reflected in a rise in P_{Cr}, leading to retention of nitrogenous waste products (defined as azotemia) such as urea. Azotemia may result from reduced renal perfusion, intrinsic renal disease, or postrenal processes (ureteral obstruction; see below and **Fig. 48-1**). Precise determination of GFR is problematic, as both commonly measured indices (urea and creatinine) have characteristics that affect their accuracy as markers of clearance. Urea clearance may underestimate GFR significantly because of urea reabsorption by the tubule. In contrast, creatinine is derived from muscle metabolism of creatine, and its generation varies little from day to day.

FIGURE 48-1

Approach to the patient with azotemia. FeNa, fractional excretion of sodium; GBM, glomerular basement membrane; RBC, red blood cell; WBC, white blood cell.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Creatinine clearance (CrCl), an approximation of GFR, is measured from plasma and urinary creatinine excretion rates for a defined period (usually 24 h) and is expressed in milliliters per minute: $CrCl = (U_{vol} \times U_{Cr})/(P_{Cr} \times T_{min})$.

The “adequacy” or “completeness” of the urinary collection is estimated by the urinary volume and creatinine

content; creatinine is produced from muscle and excreted at a relatively constant rate. For a 20- to 50-year-old man, creatinine excretion should be 18.5–25.0 mg/kg body weight; for a woman of the same age, it should be 16.5–22.4 mg/kg body weight. For example, an 80-kg man should excrete between ~1500 and 2000 mg of creatinine in an "adequate" collection. Creatinine is useful for estimating GFR because it is a small, freely filtered solute that is not reabsorbed by the tubules. P_{Cr} levels can increase acutely from dietary ingestion of cooked meat, however, and creatinine can be secreted into the proximal tubule through an organic cation pathway (especially in advanced progressive chronic kidney disease), leading to overestimation of GFR. When a timed collection for CrCl is not available, decisions about drug dosing must be based on P_{Cr} alone. Two formulas are used widely to estimate kidney function from P_{Cr}: (1) Cockcroft-Gault and (2) four-variable MDRD (Modification of Diet in Renal Disease).

Cockcroft-Gault: CrCl (mL/min) = (140 – age (years) × weight (kg) × [0.85 if female])/(72 × P_{Cr} (mg/dL).

MDRD: eGFR (mL/min per 1.73 m^2) = 186.3 × P_{Cr} ($e^{-1.154}$) × age ($e^{-0.203}$) × (0.742 if female) × (1.21 if black).

Numerous websites are available to assist with these calculations (*www.kidney.org/professionals/kdoqi/gfr_calculator.cfm*). A newer CKD-EPI eGFR, which was developed by pooling several cohorts with and without kidney disease who had data on directly measured GFR, appears to be more accurate:

CKD-EPI: eGFR = 141 × min (P_{Cr}/k, 1)^a × max (P_{Cr}/k, 1)$^{-1.209}$ × 0.993^{Age} × 1.018 [if female] × 1.159 [if black],

where P_{Cr} is plasma creatinine, *k* is 0.7 for females and 0.9 for males, *a* is –0.329 for females and –0.411 for males, *min* indicates the minimum of P_{Cr}/k or 1, and *max* indicates the maximum of P_{Cr}/k or 1 (*http://www.qxmd.com/renal/Calculate-CKD-EPI-GFR.php*).

There are limitations to all creatinine-based estimates of GFR. Each equation, along with 24-h urine collection for measurement of creatinine clearance, is based on the assumption that the patient is in *steady state*, without daily increases or decreases in P_{Cr} as a result of rapidly changing GFR. The MDRD equation is better correlated with true GFR when the GFR is <60 mL/min per 1.73 m^2. The gradual loss of muscle from chronic illness, chronic use of glucocorticoids, or malnutrition can mask significant changes in GFR with small or imperceptible changes in P_{Cr}. Cystatin C, a member of the cystatin superfamily of cysteine protease inhibitors, is produced at a relatively constant rate from all nucleated cells. Serum cystatin C has been proposed to be a more sensitive marker of early GFR decline than is P_{Cr}; however, like serum creatinine, cystatin C is influenced by the patient's age, race, and sex and also is associated with diabetes, smoking, and markers of inflammation.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Azotemia

Once GFR reduction has been established, the physician must decide if it represents acute or chronic renal injury. The clinical situation, history, and laboratory data often make this an easy distinction. However, the laboratory abnormalities characteristic of chronic renal failure, including anemia, hypocalcemia, and hyperphosphatemia, are also often present in patients presenting with acute renal failure. Radiographic evidence of renal osteodystrophy **(Chap. 305)** can be seen only in chronic renal failure but is a very late finding, typically in patients with end-stage renal disease (ESRD) maintained on dialysis. The urinalysis and renal ultrasound can facilitate distinguishing acute from chronic renal failure. An approach to the evaluation of azotemic patients is shown in Fig. 48-1. Patients with advanced chronic renal insufficiency often have some proteinuria, nonconcentrated urine (isosthenuria; isosmotic with plasma), and small kidneys on ultrasound, characterized by increased echogenicity and cortical thinning. Treatment should be directed toward slowing the progression of renal disease and providing symptomatic relief for edema, acidosis, anemia, and hyperphosphatemia, as discussed in **Chap. 305**. Acute renal failure **(Chap. 304)** can result from processes that affect and blood flow and glomerular perfusion (prerenal azotemia), intrinsic renal diseases (affecting small vessels, glomeruli, or tubules), or postrenal processes (obstruction of urine flow in ureters, bladder, or urethra) **(Chap. 313)**.

PRERENAL FAILURE

Decreased renal perfusion accounts for 40–80% of cases of acute renal failure and, if appropriately treated, is readily reversible. The etiologies of prerenal azotemia include any cause of decreased circulating blood volume (gastrointestinal hemorrhage, burns, diarrhea, diuretics), volume sequestration (pancreatitis, peritonitis, rhabdomyolysis), or decreased effective arterial volume (cardiogenic shock, sepsis). Renal and glomerular perfusion also can be affected by reductions in cardiac output from peripheral vasodilation (sepsis, drugs) or profound renal vasoconstriction (severe heart failure, hepatorenal syndrome, agents such as nonsteroidal anti-inflammatory drugs [NSAIDs]). True or “effective” arterial hypovolemia leads to a fall in mean arterial pressure, which in turn triggers a series of neural and humoral responses, including activation of the sympathetic nervous and renin-angiotensin-aldosterone systems and vasopressin (AVP) release. GFR is maintained by prostaglandin-mediated dilatation of afferent arterioles and angiotensin II–mediated constriction of efferent arterioles. Once the mean arterial pressure falls below 80 mmHg, GFR declines steeply.

Blockade of prostaglandin production by NSAIDs can result in severe vasoconstriction and acute renal failure. Blocking angiotensin action with angiotensin-converting enzyme (ACE) inhibitors or angiotensin receptor blockers (ARBs) decreases efferent arteriolar tone and in turn decreases glomerular capillary perfusion pressure. Patients taking NSAIDs and/or ACE inhibitors/ARBs are most susceptible to hemodynamically mediated acute renal failure when blood volume or arterial perfusion pressure is reduced for any reason; under these circumstances, preservation of GFR is dependent on afferent vasodilation due to prostaglandins and efferent vasoconstriction due to angiotensin-II. Patients with bilateral renal artery stenosis (or stenosis in a solitary kidney) can also be dependent on efferent arteriolar vasoconstriction for maintenance of glomerular filtration pressure and are particularly susceptible to a precipitous decline in GFR when given ACE inhibitors or ARBs.

Prolonged renal hypoperfusion may lead to acute tubular necrosis (ATN), an intrinsic renal disease that is discussed below. The urinalysis and urinary electrolyte measurements can be useful in distinguishing prerenal azotemia from ATN **(Table 48-2)**. The urine Na and osmolality of patients with prerenal azotemia can be predicted from the stimulatory actions of norepinephrine, angiotensin II, AVP, aldosterone, and low tubule fluid flow rate. In prerenal conditions, the tubules are intact, leading to a concentrated urine (>500 mosmol), avid Na retention (urine Na concentration, <20 mmol/L; fractional excretion of Na, <1%), and U_{Cr}/P_{Cr} >40 (Table 48-2).

The FE_{Na} is typically >1% in ATN, but may be <1% in patients with milder, nonoliguric ATN (e.g., from rhabdomyolysis) and in pts with underlying "prerenal" disorders, such as congestive heart failure (CHF) or cirrhosis or hepatorenal syndrome. The prerenal urine sediment is usually normal or has hyaline and granular casts, whereas the sediment of ATN usually is filled with cellular debris, tubular epithelial casts, and dark (muddy brown) granular casts. The measurement of urinary biomarkers associated with tubular injury is a promising technique to detect subclinical ATN and/or help further diagnose the exact cause of acute renal failure.

POSTRENAL AZOTEMIA

Urinary tract obstruction accounts for <5% of cases of acute renal failure but is usually reversible and must be ruled out early in the evaluation (Fig. 48-1). Since a single kidney is capable of adequate clearance, complete obstructive acute renal failure requires obstruction at the urethra or bladder outlet, bilateral ureteral obstruction, or unilateral obstruction in a patient with a single functioning kidney. Obstruction is usually diagnosed by the presence of ureteral and renal pelvic dilation on renal ultrasound. However, early in the course of obstruction or if the ureters are unable to dilate (e.g., encasement by pelvic or periureteral tumors or by retroperitoneal fibrosis), the ultrasound examination may be negative. Other imaging, such as a furosemide renogram (MAG3 nuclear medicine study), may be required to better define the presence or absence of obstructive uropathy. **The specific urologic conditions that cause obstruction are discussed in Chap. 313.**

INTRINSIC RENAL DISEASE

When prerenal and postrenal azotemia have been excluded as etiologies of renal failure, an intrinsic parenchymal renal disease is present. Intrinsic renal disease can arise from processes involving large renal vessels, intrarenal microvasculature and glomeruli, or the tubulointerstitium. Ischemic and toxic ATN account for ~90% of cases of acute intrinsic renal failure. As outlined in Fig. 48-1, the clinical setting and urinalysis are helpful in separating the possible etiologies. Prerenal azotemia and ATN are part of a spectrum of renal hypoperfusion; evidence of structural tubule injury is present in ATN, whereas prompt reversibility occurs with prerenal azotemia upon restoration of adequate renal perfusion. Thus, ATN often can be distinguished from prerenal azotemia by urinalysis and urine electrolyte composition (Table 48-2 and Fig. 48-1). Ischemic ATN is observed most frequently in patients who have undergone major surgery, trauma, severe hypovolemia, overwhelming sepsis, or extensive burns. Nephrotoxic ATN complicates the administration of many common medications, usually by inducing a combination of intrarenal vasoconstriction, direct tubule toxicity, and/or tubular obstruction. The kidney is vulnerable to toxic injury by virtue of its rich blood supply (25% of cardiac output) and its ability to concentrate and metabolize toxins. A diligent search for hypotension and nephrotoxins usually uncovers the specific etiology of ATN. Discontinuation of nephrotoxins and stabilization of blood

pressure often suffice without the need for dialysis, with ongoing regeneration of tubular cells. **An extensive list of potential drugs and toxins implicated in ATN is found in Chap. 304.**

Processes involving the tubules and interstitium can lead to acute kidney injury (AKI), a subtype of acute renal failure. These processes include drug-induced interstitial nephritis (especially by antibiotics, NSAIDs, and diuretics), severe infections (both bacterial and viral), systemic diseases (e.g., systemic lupus erythematosus), and systemic disorders (e.g., sarcoidosis, Sjögren's syndrome, lymphoma, or leukemia). A list of drugs associated with allergic interstitial nephritis is found in **Chap. 310**. Urinalysis usually shows mild to moderate proteinuria, hematuria, and pyuria (~75% of cases) and occasionally WBC casts. The finding of RBC casts in interstitial nephritis has been reported but should prompt a search for glomerular diseases (Fig. 48-1). Occasionally, renal biopsy will be needed to distinguish among these possibilities. The classic sediment finding in allergic interstitial nephritis is a predominance (>10%) of urinary eosinophils with Wright's or Hansel's stain; however, urinary eosinophils can be increased in several other causes of AKI, such that measurement of urine eosinophils has no diagnostic utility in renal disease.

Occlusion of large renal vessels, including arteries and veins, is an uncommon cause of acute renal failure. A significant reduction in GFR by this mechanism suggests bilateral processes or, in a patient with a single functioning kidney, a unilateral process. In patients with preexisting renal artery stenosis, a substantial renal collateral circulation can develop over time and sustain renal perfusion—typically not enough to sustain glomerular filtration—in the event of total renal artery occlusion. Renal arteries can be occluded with atheroemboli, thromboemboli, in situ thrombosis, aortic dissection, or vasculitis. Atheroembolic renal failure can occur spontaneously but most often is associated with recent aortic instrumentation. The emboli are cholesterol-rich and lodge in medium and small renal arteries, with a consequent eosinophil-rich inflammatory reaction. Patients with atheroembolic acute renal failure often have a normal urinalysis, but the urine may contain eosinophils and casts. The diagnosis can be confirmed by renal biopsy, but this procedure is often unnecessary when other stigmata of atheroemboli are present (livedo reticularis, distal peripheral infarcts, eosinophilia). Renal artery thrombosis may lead to mild proteinuria and hematuria, whereas renal vein thrombosis typically occurs in the context of heavy proteinuria and hematuria. **These vascular complications often require angiography for confirmation and are discussed in Chap. 311.**

Diseases of the glomeruli (glomerulonephritis and vasculitis) and the renal microvasculature (hemolytic-uremic syndromes, thrombotic thrombocytopenic purpura, and malignant hypertension) usually present with various combinations of glomerular injury: proteinuria, hematuria, reduced GFR, and alterations of sodium excretion that lead to hypertension, edema, and circulatory congestion (acute nephritic syndrome). These findings may occur as primary renal diseases or as renal manifestations of systemic diseases. The clinical setting and other laboratory data help distinguish primary renal diseases from systemic diseases. The finding of RBC casts in the urine is an indication for early renal biopsy (Fig. 48-1), as the pathologic pattern has important implications for diagnosis, prognosis, and treatment. Hematuria without RBC casts can also be an indication of glomerular disease, since RBC casts are highly specific but very insensitive for glomerulonephritis. The specificity of urine microscopy can be enhanced by examining urine with a phase contrast microscope capable of detecting dysmorphic red cells ("acanthocytes") that are associated with glomerular disease. This evaluation is

summarized in Fig. 48-2. **A detailed discussion of glomerulonephritis and diseases of the microvasculature is found in Chap. 310.**

OLIGURIA AND ANURIA

Oliguria refers to a 24-h urine output <400 mL, and *anuria* is the complete absence of urine formation (<100 mL). Anuria can be caused by complete bilateral urinary tract obstruction; a vascular catastrophe (dissection or arterial occlusion); renal vein thrombosis; acute cast nephropathy in myeloma; renal cortical necrosis; severe ATN; combined therapy with nonsteroidal anti-inflammatory drugs, ACE inhibitors, and/or ARBs; and hypovolemic, cardiogenic, or septic shock. Oliguria is never normal, since at least 400 mL of maximally concentrated urine must be produced to excrete the obligate daily osmolar load. *Nonoliguria* refers to urine output >400 mL/d in patients with acute or chronic azotemia. With nonoliguric ATN, disturbances of potassium and hydrogen balance are less severe than in oliguric patients, and recovery to normal renal function is usually more rapid.

TABLE 48-2

Laboratory Findings in Acute Renal Failure

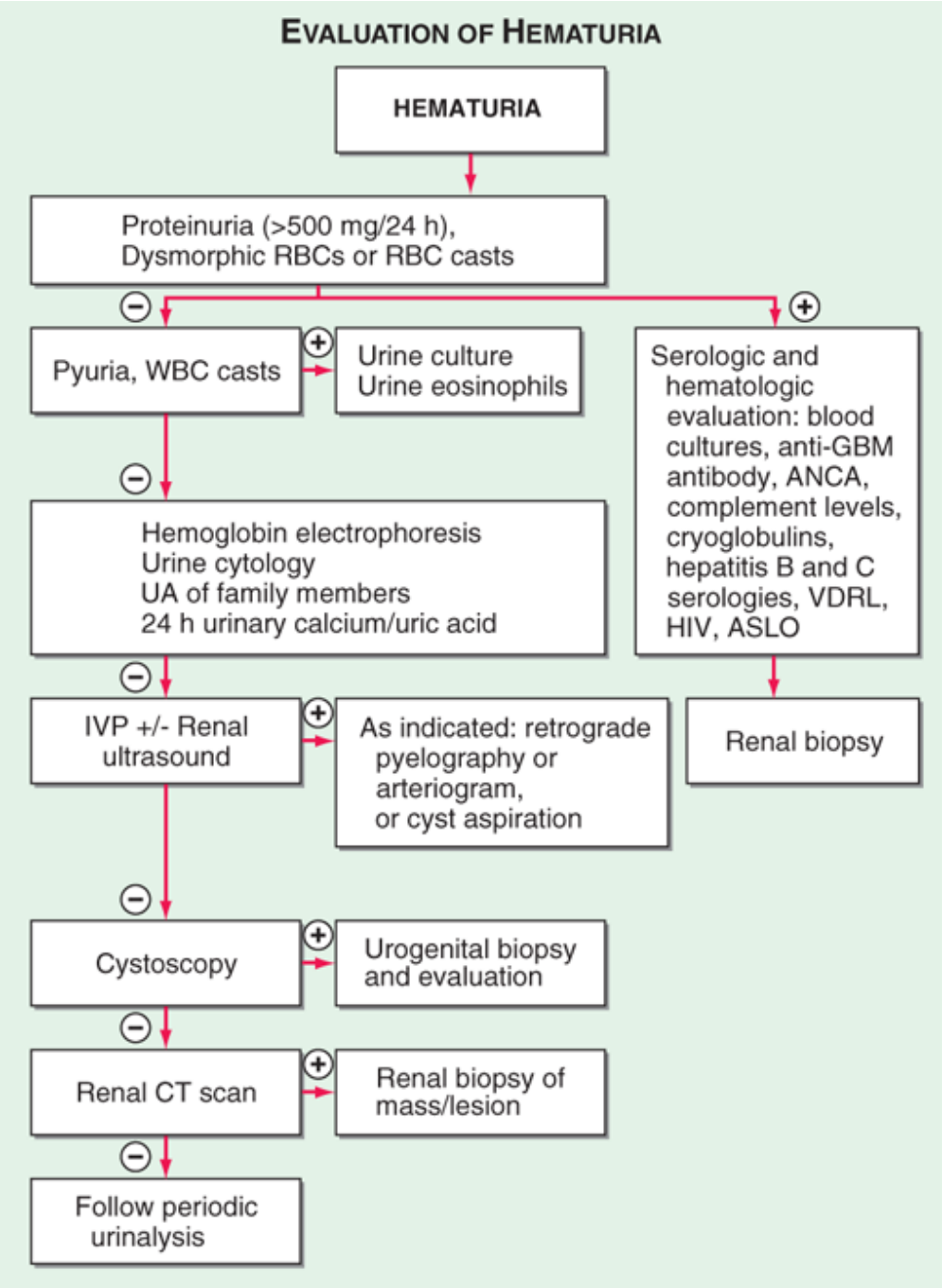
INDEX	PRERENAL AZOTEMIA	OLIGURIC ACUTE RENAL FAILURE
BUN/P_{Cr} ratio	>20:1	10–15:1
Urine sodium U_{Na}, meq/L	<20	>40
Urine osmolality, mosmol/L H_2O	>500	<350
Fractional excretion of sodium[a]	<1%	>2%
Urine/plasma creatinine U_{Cr}/P_{Cr}	>40	<20
Urinalysis (casts)	None or hyaline/granular	Muddy brown

$$^{a}FE_{Na} = \frac{U_{Na} \times P_{Cr} \times 100}{P_{Na} \times U_{Cr}}$$

Abbreviations: BUN, blood urea nitrogen; P_{Cr}, plasma creatinine concentration; P_{Na}, plasma sodium concentration; U_{Cr}, urine creatinine concentration; U_{Na}, urine sodium concentration.

FIGURE 48-2

Approach to the patient with hematuria. ANCA, antineutrophil cytoplasmic antibody; ASLO, antistreptolysin O; CT, computed tomography; GBM, glomerular basement membrane; IVP, intravenous pyelography; RBC, red blood cell; UA, urinalysis; VDRL, Venereal Disease Research Laboratory; WBC, white blood cell.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ABNORMALITIES OF THE URINE

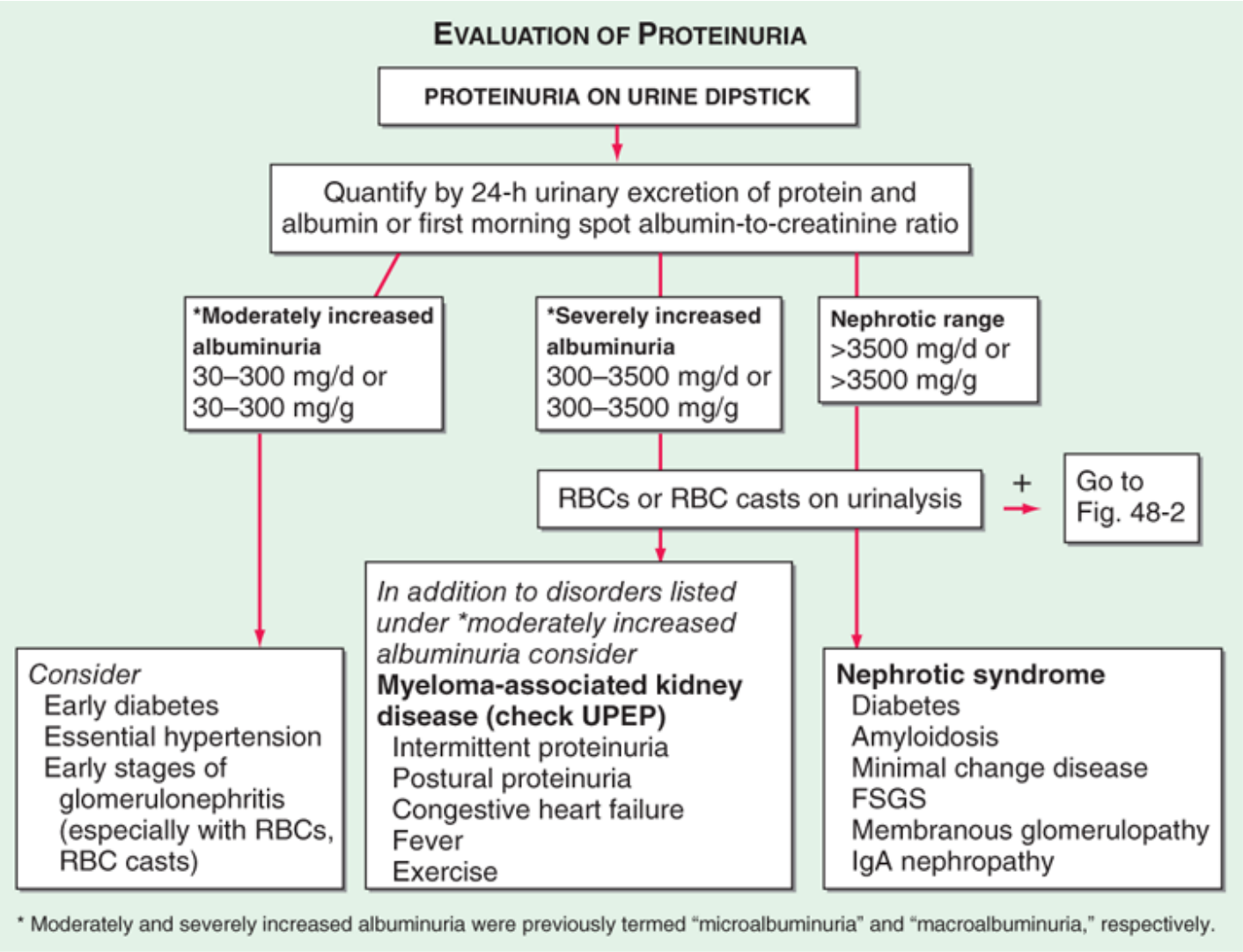
PROTEINURIA

The evaluation of proteinuria is shown schematically in Fig. 48-3 and typically is initiated after detection of proteinuria by dipstick examination. The dipstick measurement detects only albumin and gives false-positive results at pH >7.0 or when the urine is very concentrated or contaminated with blood. Because the dipstick relies on urinary albumin concentration, a very dilute urine may obscure significant proteinuria on dipstick examination. Quantification of urinary albumin on a spot urine sample (ideally from a first morning void) by

measurement of an albumin-to-creatinine ratio (ACR) is helpful in approximating a 24-h albumin excretion rate (AER), where ACR (mg/g) ≈AER (mg/24 h). Furthermore, proteinuria that is not predominantly due to albumin will be missed by dipstick screening. This information is particularly important for the detection of Bence-Jones proteins in the urine of patients with multiple myeloma. Tests to measure total urine protein concentration accurately rely on precipitation with sulfosalicylic or trichloracetic acid (Fig. 48-3). As with albuminuria, the ratio of protein to creatinine in a random, "spot" urine can also provide a rough estimate of protein excretion; for example, a protein/creatinine ratio of 3.0 correlates to ~3.0 g of proteinuria per day. Formal assessment of urinary protein excretion requires a 24-h urine protein collection (see "Measurement of GFR," above).

FIGURE 48-3

Approach to the patient with proteinuria. Investigation of proteinuria is often initiated by a positive dipstick on routine urinalysis. Conventional dipsticks detect predominantly albumin and provide a semiquantitative assessment (trace, 1+, 2+, or 3+), which is influenced by urinary concentration as reflected by urine specific gravity (minimum, <1.005; maximum, 1.030). However, more exact determination of proteinuria should employ a spot morning protein/creatinine ratio (mg/g) or a 24-h urine collection (mg/24 h). FSGS, focal segmental glomerulosclerosis; RBC, red blood cell; UPEP, urine protein electrophoresis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The magnitude of proteinuria and its composition in the urine depend on the mechanism of renal injury that leads to protein losses. Both charge and size selectivity normally prevent virtually all plasma albumin, globulins, and other high-molecular-weight proteins from crossing the glomerular wall; however, if this barrier is disrupted, plasma proteins may leak into the urine (glomerular proteinuria; Fig. 48-3). Smaller proteins (<20 kDa) are freely filtered but are readily reabsorbed by the proximal tubule. Typically, healthy individuals excrete <150 mg/d of total protein and <30 mg/d of albumin. However, even at albuminuria levels <30 mg/d, risk for progression to overt nephropathy or subsequent cardiovascular disease is increased. The remainder of the protein in the urine is secreted by the tubules (Tamm-Horsfall, IgA, and urokinase) or represents small amounts of filtered β_2-microglobulin, apoproteins, enzymes, and peptide hormones. Another mechanism of proteinuria entails excessive production of an abnormal protein that exceeds the capacity of the tubule for reabsorption. This situation most commonly occurs with plasma cell dyscrasias, such as multiple myeloma, amyloidosis, and lymphomas, that are associated with monoclonal production of immunoglobulin light chains.

The normal glomerular endothelial cell forms a barrier composed of pores of ~100 nm that retain blood cells but offer little impediment to passage of most proteins. The glomerular basement membrane traps most large proteins (>100 kDa), and the foot processes of epithelial cells (podocytes) cover the urinary side of the glomerular basement membrane and produce a series of narrow channels (slit diaphragms) to allow molecular passage of small solutes and water but not proteins. Some glomerular diseases, such as minimal change disease, cause fusion of glomerular epithelial cell foot processes, resulting in predominantly “selective” (Fig. 48-3) loss of albumin. Other glomerular diseases can present with disruption of the basement membrane and slit diaphragms (e.g., by immune complex deposition), resulting in losses of albumin and other plasma proteins. The fusion of foot processes causes increased pressure across the capillary basement membrane, resulting in areas with larger pore sizes (and more severe “nonselective” proteinuria (Fig. 48-3).

When the total daily urinary excretion of protein is >3.5 g, hypoalbuminemia, hyperlipidemia, and edema (nephrotic syndrome; Fig. 48-3) are often present as well. However, total daily urinary protein excretion >3.5 g can occur without the other features of the nephrotic syndrome in a variety of other renal diseases, including diabetes (Fig. 48-3). Plasma cell dyscrasias (multiple myeloma) can be associated with large amounts of excreted light chains in the urine, which may not be detected by dipstick. The light chains are filtered by the glomerulus and overwhelm the reabsorptive capacity of the proximal tubule. Renal failure from these disorders occurs through a variety of mechanisms, including but not limited to proximal tubule injury, tubule obstruction (cast nephropathy), amyloid deposition, and light chain deposition **(Chap. 310)**. The specific renal lesion is dictated by the sequence and structural characteristics of the monoclonal light chain; however, not all excreted light chains are nephrotoxic.

Hypoalbuminemia in nephrotic syndrome occurs through excessive urinary losses and increased proximal tubule catabolism of filtered albumin. Edema results from renal sodium retention and reduced plasma oncotic pressure, which favors fluid movement from capillaries to interstitium. To compensate for the perceived decrease in effective intravascular volume, activation of the renin-angiotensin system, stimulation of AVP, and activation of the sympathetic nervous system take place, promoting continued renal salt and water reabsorption and progressive edema. Filtered proteases, normally retained by the glomerular filtration barrier, can also directly activate sodium reabsorption by the epithelial Na channels in principal cells (ENaC) in

nephrotic syndrome. Despite these changes, hypertension is uncommon in primary kidney diseases resulting in the nephrotic syndrome (Fig. 48-3 and **Chap. 308**). The urinary loss of regulatory proteins and changes in hepatic synthesis contribute to the other manifestations of the nephrotic syndrome. A hypercoagulable state may arise from urinary losses of antithrombin III, reduced serum levels of proteins S and C, hyperfibrinogenemia, and enhanced platelet aggregation. Hypercholesterolemia may be severe and results from increased hepatic lipoprotein synthesis. Loss of immunoglobulins contributes to an increased risk of infection. Many diseases (some listed in Fig. 48-3) and drugs can cause the nephrotic syndrome; a complete list is found in **Chap. 308**.

HEMATURIA, PYURIA, AND CASTS

Isolated hematuria without proteinuria, other cells, or casts is often indicative of bleeding from the urinary tract. Hematuria is defined as two to five RBCs per high-power field (HPF) and can be detected by dipstick. A false-positive dipstick for hematuria (where no RBCs are seen on urine microscopy) may occur when myoglobinuria is present, often in the setting of rhabdomyolysis. Common causes of isolated hematuria include stones, neoplasms, tuberculosis, trauma, and prostatitis. Gross hematuria with blood clots usually is not an intrinsic renal process; rather, it suggests a postrenal source in the urinary collecting system. Evaluation of patients presenting with microscopic hematuria is outlined in Fig. 48-2. A single urinalysis with hematuria is common and can result from menstruation, viral illness, allergy, exercise, or mild trauma. Persistent or significant hematuria (>3 RBCs/HPF on three urinalyses, a single urinalysis with >100 RBCs, or gross hematuria) is associated with significant renal or urologic lesions in 9.1% of cases. The level of suspicion for urogenital neoplasms in patients with isolated painless hematuria and nondysmorphic RBCs increases with age. Neoplasms are rare in the pediatric population, and isolated hematuria is more likely to be “idiopathic” or associated with a congenital anomaly. Hematuria with pyuria and bacteriuria is typical of infection and should be treated with antibiotics after appropriate cultures. Acute cystitis or urethritis in women can cause gross hematuria. Hypercalciuria and hyperuricosuria are also risk factors for unexplained isolated hematuria in both children and adults. In some of these patients (50–60%), reducing calcium and uric acid excretion through dietary interventions can eliminate the microscopic hematuria.

Isolated microscopic hematuria can be a manifestation of glomerular diseases. The RBCs of glomerular origin are often dysmorphic when examined by phase-contrast microscopy. Irregular shapes of RBCs may also result from pH and osmolarity changes produced along the distal nephron. Observer variability in detecting dysmorphic RBCs is common. The most common etiologies of isolated glomerular hematuria are IgA nephropathy, hereditary nephritis, and thin basement membrane disease. IgA nephropathy and hereditary nephritis can lead to episodic gross hematuria. A family history of renal failure is often present in hereditary nephritis, and patients with thin basement membrane disease often have family members with microscopic hematuria. A renal biopsy is needed for the definitive diagnosis of these disorders, which are discussed in more detail in **Chap. 308**. Hematuria with dysmorphic RBCs, RBC casts, and protein excretion >500 mg/d is virtually diagnostic of glomerulonephritis. RBC casts form as RBCs that enter the tubule fluid and become trapped in a cylindrical mold of gelled Tamm-Horsfall protein. Even in the absence of azotemia, these patients should undergo serologic evaluation and renal biopsy as outlined in Fig. 48-2.

Isolated pyuria is unusual since inflammatory reactions in the kidney or collecting system also are associated with hematuria. The presence of bacteria suggests infection, and WBC casts with bacteria are indicative of pyelonephritis. WBCs and/or WBC casts also may be seen in acute glomerulonephritis as well as in tubulointerstitial processes such as interstitial nephritis and transplant rejection.

Casts can be seen in chronic renal diseases. Degenerated cellular casts called *waxy casts* or *broad casts* (arising in the dilated tubules that have undergone compensatory hypertrophy in response to reduced renal mass) may be seen in the urine.

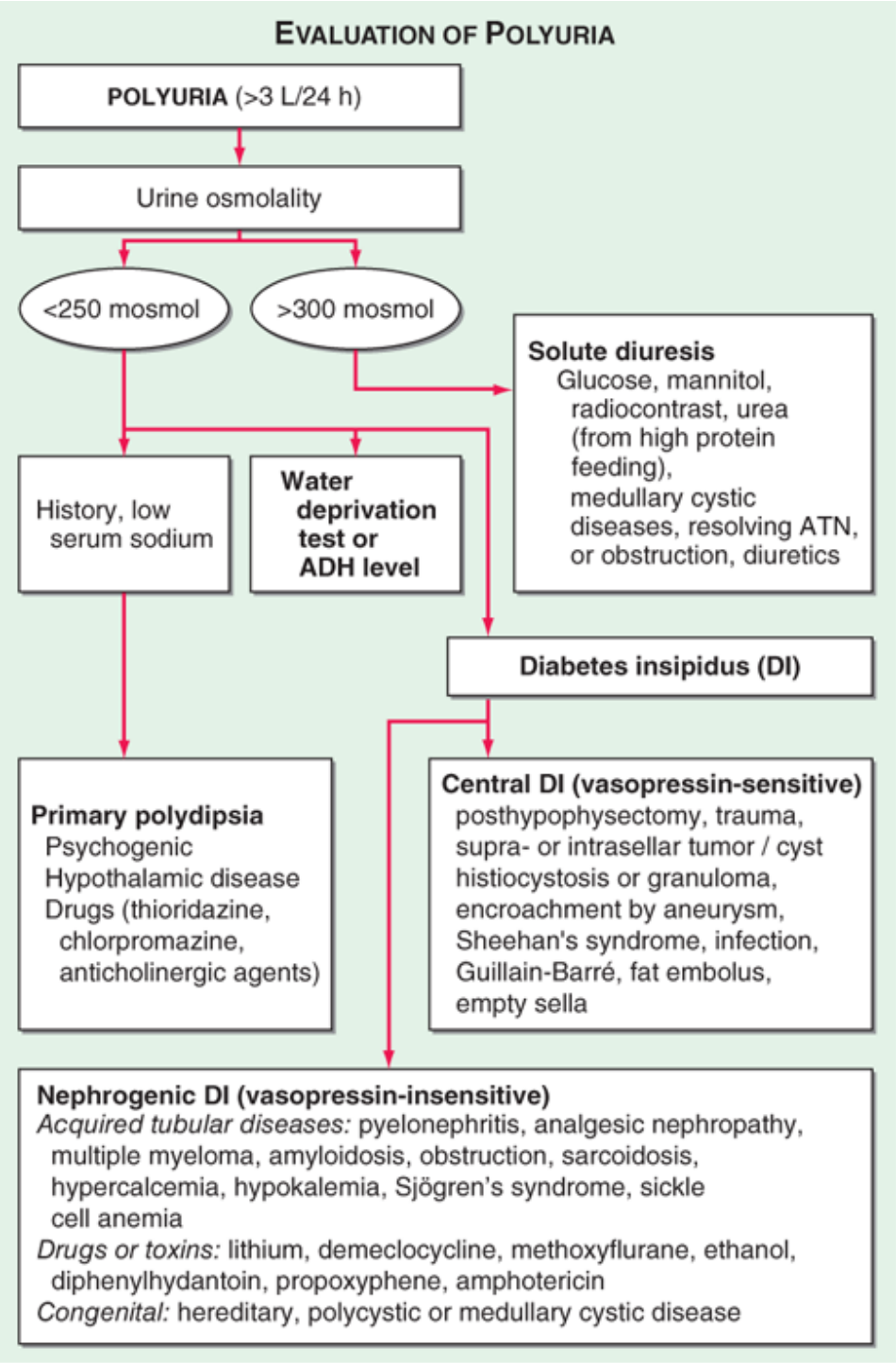
ABNORMALITIES OF URINE VOLUME

POLYURIA

By history, it is often difficult for patients to distinguish urinary frequency (often of small volumes) from true polyuria (>3 L/d), and a quantification of volume by 24-h urine collection may be needed **(Fig. 48-4)**. Polyuria results from two potential mechanisms: (1) excretion of nonabsorbable solutes (such as glucose) or (2) excretion of water (usually from a defect in AVP production or renal responsiveness). To distinguish a solute diuresis from a water diuresis and to determine whether the diuresis is appropriate for the clinical circumstances, urine osmolality is measured. The average person excretes between 600 and 800 mosmol of solutes per day, primarily as urea and electrolytes. If the urine output is >3 L/d and the urine is dilute (<250 mosmol/L), total osmolar excretion is normal and a water diuresis is present. This circumstance could arise from polydipsia, inadequate secretion of vasopressin (*central diabetes insipidus*), or failure of renal tubules to respond to vasopressin (*nephrogenic diabetes insipidus*). If the urine volume is >3 L/d and urine osmolality is >300 mosmol/L, a solute diuresis is clearly present and a search for the responsible solute(s) is mandatory.

FIGURE 48-4

Approach to the patient with polyuria. AVP, antidiuretic hormone; ATN, acute tubular necrosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Excessive filtration of a poorly reabsorbed solute such as glucose or mannitol can depress reabsorption of NaCl and water in the proximal tubule and lead to enhanced excretion in the urine. Poorly controlled diabetes mellitus with glucosuria is the most common cause of a solute diuresis, leading to volume depletion and serum hypertonicity. Since the urine sodium concentration is less than that of blood, more water than sodium is lost, causing hypernatremia and hypertonicity. Common iatrogenic solute diuresis occurs in association with mannitol administration, radiocontrast media, and high-protein feedings (enteral or parenteral), leading to increased urea production and excretion. Less commonly, excessive sodium loss may result from cystic renal diseases or Bartter's syndrome or may develop during a tubulointerstitial process (such as resolving ATN). In

these so-called salt-wasting disorders, the tubule damage results in direct impairment of sodium reabsorption and indirectly reduces the responsiveness of the tubule to aldosterone. Usually, the sodium losses are mild, and the obligatory urine output is <2 L/d; resolving ATN and postobstructive diuresis are exceptions and may be associated with significant natriuresis and polyuria.

Formation of large volumes of dilute urine is usually due to polydipsic states or diabetes insipidus. Primary polydipsia can result from habit, psychiatric disorders, neurological lesions, or medications. During deliberate polydipsia, extracellular fluid volume is normal or expanded and plasma AVP levels are reduced because serum osmolality tends to be near the lower limits of normal. Urine osmolality is also maximally dilute at 50 mosmol/L.

Central diabetes insipidus may be idiopathic in origin or secondary to a variety of conditions, including hypophysectomy, trauma, neoplastic, inflammatory, vascular, or infectious hypothalamic diseases. Idiopathic central diabetes insipidus is associated with selective destruction of the vasopressin-secreting neurons in the supraoptic and paraventricular nuclei and can either be inherited as an autosomal dominant trait or occur spontaneously. Nephrogenic diabetes insipidus can occur in a variety of clinical situations, as summarized in Fig. 48-4.

A plasma AVP level is recommended as the best method for distinguishing between central and nephrogenic diabetes insipidus. Alternatively, a water deprivation test plus exogenous vasopressin may distinguish primary polydipsia from central and nephrogenic diabetes insipidus. **For a detailed discussion, see Chap. 374.**

ACKNOWLEDGMENT

This chapter was adapted and updated from the prior version written by Julie Lin and Bradley Denker.

FURTHER READING

Emmett M et al: Approach to the patient with kidney disease, in *Brenner and Rector's The Kidney*, 10th ed, K Skorecki et al (eds). Philadelphia, W.B. Saunders & Company, 2016, pp 754–779.

Köhler H et al: Acanthocyturia—A characteristic marker for glomerular bleeding. Kidney Int 40:115, 1991. [PubMed: 1921146]

Levey AS et al: Glomerular filtration rate and albuminuria for detection and staging of acute and chronic kidney disease in adults: A systematic review. JAMA 313:837, 2015. [PubMed: 25710660]

Perazella MA: The urine sediment as a biomarker of kidney disease. Am J Kidney Dis 66:748, 2015. [PubMed: 25943719]

Sharfuddin AA et al: Acute kidney injury, in *Brenner and Rector's The Kidney*, 10th ed, K Skorecki et al (eds). Philadelphia, W.B. Saunders & Company, 2016, pp 958–1011.

Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 49: Fluid and Electrolyte Disturbances

David B. Mount

SODIUM AND WATER

COMPOSITION OF BODY FLUIDS

Water is the most abundant constituent in the body, comprising ~50% of body weight in women and 60% in men. Total-body water is distributed in two major compartments: 55–75% is intracellular (intracellular fluid [ICF]), and 25–45% is extracellular (extracellular fluid [ECF]). The ECF is further subdivided into intravascular (plasma water) and extravascular (interstitial) spaces in a ratio of 1:3. Fluid movement between the intravascular and interstitial spaces occurs across the capillary wall and is determined by Starling forces, i.e., capillary hydraulic pressure and colloid osmotic pressure. The transcapillary hydraulic pressure gradient exceeds the corresponding oncotic pressure gradient, thereby favoring the movement of plasma ultrafiltrate into the extravascular space. The return of fluid into the intravascular compartment occurs via lymphatic flow.

The solute or particle concentration of a fluid is known as its osmolality, expressed as milliosmoles per kilogram of water (mOsm/kg). Water easily diffuses across most cell membranes to achieve osmotic equilibrium (ECF osmolality = ICF osmolality). Notably, the extracellular and intracellular solute compositions differ considerably owing to the activity of various transporters, channels, and ATP-driven membrane pumps. The major ECF particles are Na^+ and its accompanying anions Cl^- and HCO_3^-, whereas K^+ and organic phosphate esters (ATP, creatine phosphate, and phospholipids) are the predominant ICF osmoles. Solutes that are restricted to the ECF or the ICF determine the “tonicity” or effective osmolality of that compartment. Certain solutes, particularly urea, do not contribute to water shifts across most membranes and are thus known as *ineffective osmoles*.

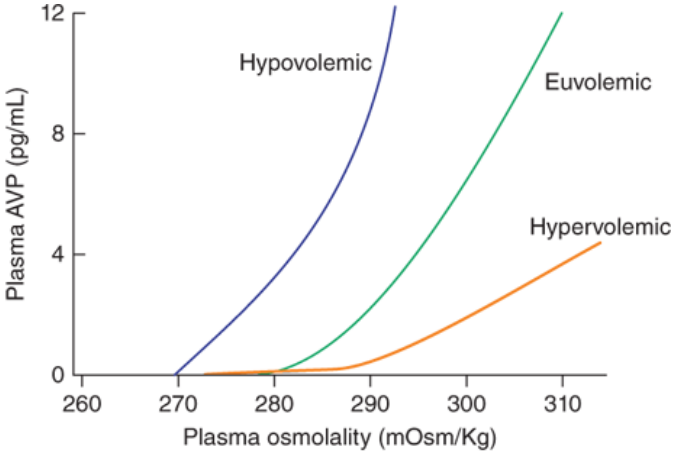
Water Balance

Vasopressin secretion, water ingestion, and renal water transport collaborate to maintain human body fluid osmolality between 280 and 295 mOsm/kg. Vasopressin (AVP) is synthesized in magnocellular neurons within the hypothalamus; the distal axons of these neurons project to the posterior pituitary or neurohypophysis, from which AVP is released into the circulation. A network of central “osmoreceptor” neurons, which includes the AVP-expressing magnocellular neurons themselves, sense circulating osmolality via nonselective, stretch-activated cation channels. These osmoreceptor neurons are activated or inhibited by modest increases and decreases in circulating osmolality, respectively; activation leads to AVP release and thirst.

AVP secretion is stimulated as systemic osmolality increases above a threshold level of ~285 mOsm/kg, above which there is a linear relationship between osmolality and circulating AVP **(Fig. 49-1)**. Thirst and thus water ingestion are also activated at ~285 mOsm/kg, beyond which there is an equivalent linear increase in the perceived intensity of thirst as a function of circulating osmolality. Changes in blood volume and blood pressure are also direct stimuli for AVP release and thirst, albeit with a less sensitive response profile. Of perhaps greater clinical relevance to the pathophysiology of water homeostasis, ECF volume strongly modulates the relationship between circulating osmolality and AVP release, such that hypovolemia reduces the osmotic threshold and increases the slope of the response curve to osmolality; *hypervolemia* has an opposite effect, increasing the osmotic threshold and reducing the slope of the response curve (Fig. 49-1). Notably, AVP has a half-life in the circulation of only 10–20 min; thus, changes in ECF volume and/or circulating osmolality can rapidly affect water homeostasis. In addition to volume status, a number of other “nonosmotic” stimuli have potent activating effects on osmosensitive neurons and AVP release, including nausea, intracerebral angiotensin II, serotonin, and multiple drugs.

FIGURE 49-1

Circulating levels of vasopressin (AVP) in response to changes in osmolality. Plasma AVP becomes detectable in euvolemic, healthy individuals at a threshold of ~285 mOsm/kg, above which there is a linear relationship between osmolality and circulating AVP. The vasopressin response to osmolality is modulated strongly by volume status. The osmotic threshold is thus slightly lower in hypovolemia, with a steeper response curve; hypervolemia reduces the sensitivity of circulating AVP levels to osmolality.

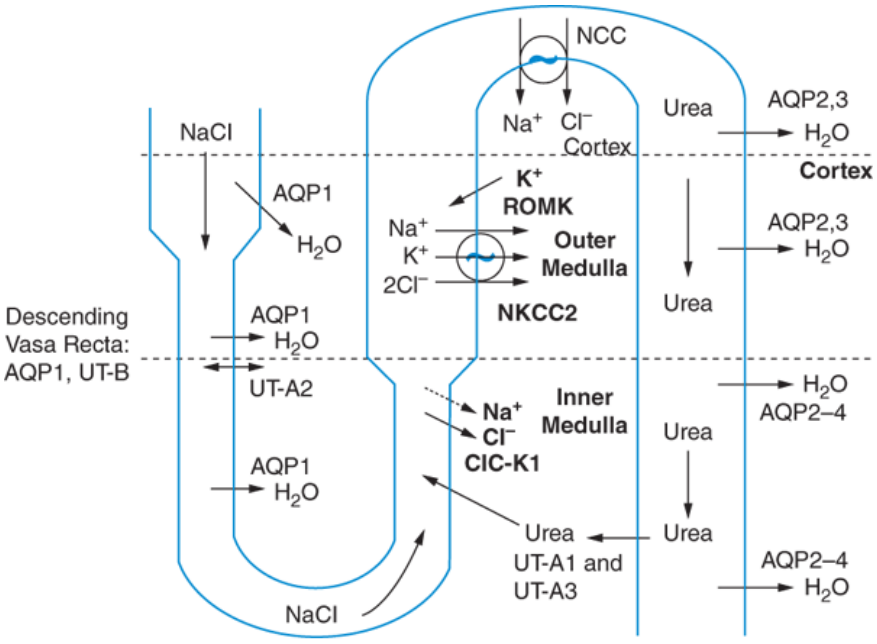


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The excretion or retention of electrolyte-free water by the kidney is modulated by circulating AVP. AVP acts on renal, V_2-type receptors in the thick ascending limb of Henle and principal cells of the collecting duct (CD), increasing intracellular levels of cyclic AMP and activating protein kinase A (PKA)–dependent phosphorylation of multiple transport proteins. The AVP- and PKA-dependent activation of Na^+-Cl^- and K^+ transport by the thick ascending limb of the loop of Henle (TALH) is a key participant in the countercurrent mechanism **(Fig. 49-2)**. The countercurrent mechanism ultimately increases the interstitial osmolality in the inner medulla of the kidney, driving water absorption across the renal CD. However, water, salt, and solute transport by both proximal and distal nephron segments participates in the renal concentrating mechanism (Fig. 49-2). Water transport across apical and basolateral aquaporin-1 water channels in the descending thin limb of the loop of Henle is thus involved, as is passive absorption of Na^+-Cl^- by the thin ascending limb, via apical and basolateral CLC-K1 chloride channels and paracellular Na^+ transport. Renal urea transport in turn plays important roles in the generation of the medullary osmotic gradient and the ability to excrete solute-free water under conditions of both high and low protein intake (Fig. 49-2).

FIGURE 49-2

The renal concentrating mechanism. Water, salt, and solute transport by both proximal and distal nephron segments participates in the renal concentrating mechanism (see text for details). Diagram showing the location of the major transport proteins involved; a loop of Henle is depicted on the *left,* collecting duct on the *right*. AQP, aquaporin; CLC-K1, chloride channel; NKCC2, Na-K-2Cl cotransporter; ROMK, renal outer medullary K^+ channel; UT, urea transporter. *(Used with permission from JM Sands: Molecular approaches to urea transporters. J Am Soc Nephrol 13:2795, 2002.)*



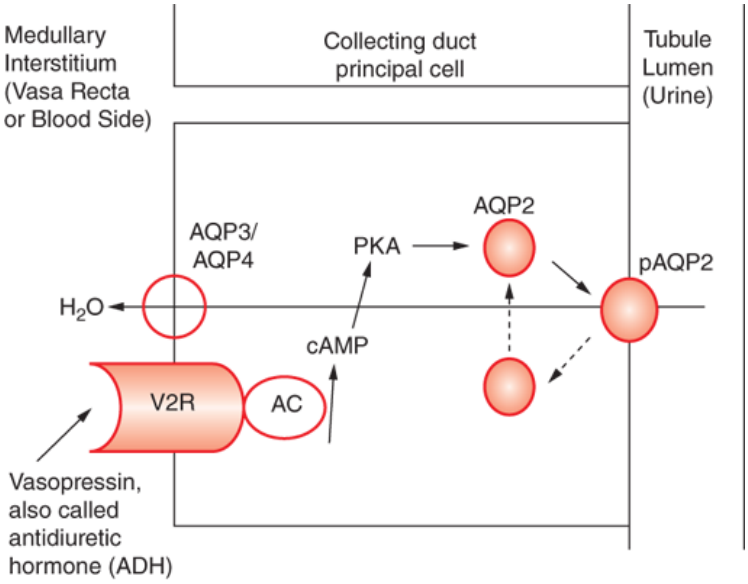
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

AVP-induced, PKA-dependent phosphorylation of the aquaporin-2 water channel in principal cells stimulates the insertion of active water channels into the lumen of the CD, resulting in transepithelial water absorption down the medullary osmotic gradient **(Fig. 49-3)**. Under "antidiuretic" conditions, with increased circulating AVP, the kidney reabsorbs water filtered by the glomerulus, equilibrating the osmolality across the CD epithelium to excrete a

hypertonic, “concentrated” urine (osmolality of up to 1200 mOsm/kg). In the absence of circulating AVP, insertion of aquaporin-2 channels and water absorption across the CD is essentially abolished, resulting in secretion of a hypotonic, dilute urine (osmolality as low as 30–50 mOsm/kg). Abnormalities in this “final common pathway” are involved in most disorders of water homeostasis, e.g., a reduced or absent insertion of active aquaporin-2 water channels into the membrane of principal cells in diabetes insipidus (DI).

FIGURE 49-3

Vasopressin and the regulation of water permeability in the renal collecting duct. Vasopressin binds to the type 2 vasopressin receptor (V2R) on the basolateral membrane of principal cells, activates adenylyl cyclase (AC), increases intracellular cyclic adenosine monophosphatase (cAMP), and stimulates protein kinase A (PKA) activity. Cytoplasmic vesicles carrying aquaporin-2 (AQP) water channel proteins are inserted into the luminal membrane in response to vasopressin, thereby increasing the water permeability of this membrane. When vasopressin stimulation ends, water channels are retrieved by an endocytic process and water permeability returns to its low basal rate. The AQP3 and AQP4 water channels are expressed on the basolateral membrane and complete the transcellular pathway for water reabsorption. pAQP2, phosphorylated aquaporin-2. *(From JM Sands, DG Bichet: Nephrogenic diabetes insipidus. Ann Intern Med 144:186, 2006, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Maintenance of Arterial Circulatory Integrity

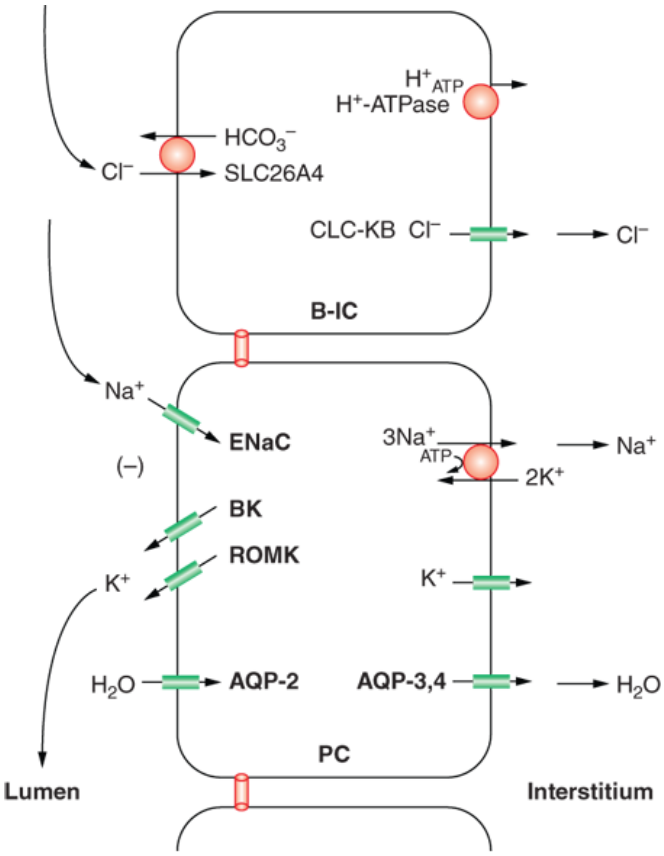
Sodium is actively pumped out of cells by the Na^+/K^+-ATPase membrane pump. In consequence, 85–90% of body Na^+ is extracellular, and the ECF volume (ECFV) is a function of total-body Na^+ content. Arterial perfusion and circulatory integrity are, in turn, determined by renal Na^+ retention or excretion, in addition to the modulation of systemic arterial resistance. Within the kidney, Na^+ is filtered by the glomeruli and then sequentially reabsorbed by the renal tubules. The Na^+ cation is typically reabsorbed with the chloride anion (Cl^-), and, thus, chloride homeostasis also affects the ECFV. On a quantitative level, at a glomerular filtration rate (GFR) of 180 L/d and serum Na^+ of ~140 m*M*, the kidney filters some 25,200 mmol/d of Na^+. This is equivalent to ~1.5 kg of salt, which would occupy roughly 10 times the extracellular space; 99.6% of filtered Na^+-Cl^- must be reabsorbed to excrete 100 m*M* per day. Minute changes in renal Na^+-Cl^- excretion will thus have significant effects on the ECFV, leading to edema syndromes or hypovolemia.

Approximately two-thirds of filtered Na^+-Cl^- is reabsorbed by the renal proximal tubule, via both paracellular and transcellular mechanisms. The TALH subsequently reabsorbs another 25–30% of filtered Na^+-Cl^- via the apical, furosemide-sensitive Na^+-K^+-$2Cl^-$ cotransporter. The adjacent aldosterone-sensitive distal nephron, comprising the distal convoluted tubule (DCT), connecting tubule (CNT), and CD, accomplishes the “fine-tuning” of renal Na^+-Cl^- excretion. The thiazide-sensitive apical Na^+-Cl^- cotransporter (NCC) reabsorbs 5–10% of filtered Na^+-Cl^- in the DCT. Principal cells in the CNT and CD reabsorb Na^+ via electrogenic, amiloride-sensitive epithelial Na^+ channels (ENaC); Cl^- ions are primarily reabsorbed by adjacent intercalated cells, via apical Cl^- exchange (Cl^--OH^- and Cl^--HCO_3^- exchange, mediated by the SLC26A4 anion exchanger) **(Fig. 49-4)**.

FIGURE 49-4

Sodium, water, and potassium transport in principal cells (PC) and adjacent β-intercalated cells (B-IC). The absorption of Na^+ via the amiloride-sensitive epithelial sodium channel (ENaC) generates a lumen-negative potential difference, which drives K^+ excretion through the apical secretory K^+ channel ROMK (renal outer medullary K^+ channel) and/or the flow-dependent BK channel. Transepithelial Cl^- transport occurs in adjacent β-intercalated cells,

via apical Cl^--HCO_3^- and Cl^--OH^- exchange (SLC26A4 anion exchanger, also known as pendrin) basolateral CLC chloride channels. Water is absorbed down the osmotic gradient by principal cells, through the apical aquaporin-2 (AQP-2) and basolateral aquaporin-3 and aquaporin-4 (Fig. 49-3).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Renal tubular reabsorption of filtered Na^+-Cl^- is regulated by multiple circulating and paracrine hormones, in addition to the activity of renal nerves. Angiotensin II activates proximal Na^+-Cl^- reabsorption, as do adrenergic receptors under the influence of renal sympathetic innervation; locally generated dopamine, in contrast, has a *natriuretic* effect. Aldosterone primarily activates Na^+-Cl^- reabsorption within the aldosterone-sensitive distal nephron. In particular, aldosterone activates the ENaC channel in principal cells, inducing Na^+ absorption and promoting K^+ excretion (Fig. 49-4).

Circulatory integrity is critical for the perfusion and function of vital organs. "Underfilling" of the arterial circulation is sensed by ventricular and vascular pressure receptors, resulting in a neurohumoral activation (increased sympathetic tone, activation of the renin-angiotensin-aldosterone axis, and increased circulating AVP) that synergistically increases renal Na^+-Cl^- reabsorption, vascular resistance, and renal water reabsorption. This occurs in the context of decreased cardiac output, as occurs in hypovolemic states, low-output cardiac failure, decreased oncotic pressure, and/or increased capillary permeability. Alternatively, excessive arterial vasodilation results in *relative* arterial underfilling, leading to neurohumoral activation in the defense of tissue perfusion. These physiologic responses play important roles in many of the disorders discussed in this chapter. In particular, it is important to appreciate that AVP functions in the defense of circulatory integrity, inducing vasoconstriction, increasing sympathetic nervous system tone, increasing renal retention of both water and Na^+-Cl^-, and modulating the arterial baroreceptor reflex. Most of these responses involve activation of systemic V_{1A} AVP receptors, but concomitant activation of V_2 receptors in the kidney can result in renal water retention and hyponatremia.

HYPOVOLEMIA

Etiology

True volume depletion, or hypovolemia, generally refers to a state of combined salt and water loss, leading to contraction of the ECFV. The loss of salt and water may be renal or nonrenal in origin.

RENAL CAUSES

Excessive urinary Na^+-Cl^- and water loss is a feature of several conditions. A high filtered load of endogenous solutes, such as glucose and urea, can impair tubular reabsorption of Na^+-Cl^- and water, leading to an osmotic diuresis. Exogenous mannitol, often used to decrease intracerebral pressure, is filtered by glomeruli but not reabsorbed by the proximal tubule, thus causing an osmotic diuresis. Pharmacologic diuretics selectively impair Na^+-Cl^-

reabsorption at specific sites along the nephron, leading to increased urinary Na^+-Cl^- excretion. Other drugs can induce natriuresis as a side effect. For example, acetazolamide can inhibit proximal tubular Na^+-Cl^- absorption via its inhibition of carbonic anhydrase; other drugs, such as the antibiotics trimethoprim (TMP) and pentamidine, inhibit distal tubular Na^+ reabsorption through the amiloride-sensitive ENaC channel, leading to urinary Na^+-Cl^- loss. Hereditary defects in renal transport proteins are also associated with reduced reabsorption of filtered Na^+-Cl^- and/or water. Alternatively, mineralocorticoid deficiency, mineralocorticoid resistance, or inhibition of the mineralocorticoid receptor (MLR) can reduce Na^+-Cl^- reabsorption by the aldosterone-sensitive distal nephron. Finally, tubulointerstitial injury, as occurs in interstitial nephritis, acute tubular injury, or obstructive uropathy, can reduce distal tubular Na^+-Cl^- and/or water absorption.

Excessive excretion of free water, i.e., water without electrolytes, can also lead to hypovolemia. However, the effect on ECFV is usually less marked, given that two-thirds of the water volume is lost from the ICF. Excessive renal water excretion occurs in the setting of decreased circulating AVP or renal resistance to AVP (central and nephrogenic DI, respectively).

EXTRARENAL CAUSES

Nonrenal causes of hypovolemia include fluid loss from the gastrointestinal tract, skin, and respiratory system. Accumulations of fluid within specific tissue compartments, typically the interstitium, peritoneum, or gastrointestinal tract, can also cause hypovolemia.

Approximately 9 L of fluid enter the gastrointestinal tract daily, 2 L by ingestion and 7 L by secretion; almost 98% of this volume is absorbed, such that daily fecal fluid loss is only 100–200 mL. Impaired gastrointestinal reabsorption or enhanced secretion of fluid can cause hypovolemia. Because gastric secretions have a low pH (high H^+ concentration), whereas biliary, pancreatic, and intestinal secretions are alkaline (high HCO_3^- concentration), vomiting and diarrhea are often accompanied by metabolic alkalosis and acidosis, respectively.

Evaporation of water from the skin and respiratory tract (so-called “insensible losses”) constitutes the major route for loss of solute-free water, which is typically 500–650 mL/d in healthy adults. This evaporative loss can increase during febrile illness or prolonged heat exposure. Hyperventilation can also increase insensible losses via the respiratory tract, particularly in ventilated patients; the humidity of inspired air is another determining factor. In addition, increased exertion and/or ambient temperature will increase insensible losses via sweat, which is hypotonic to plasma. Profuse sweating without adequate repletion of water and Na^+-Cl^- can thus lead to both hypovolemia and hypertonicity. Alternatively, replacement of these insensible losses with a surfeit of free water, without adequate replacement of electrolytes, may lead to hypovolemic hyponatremia.

Excessive fluid accumulation in interstitial and/or peritoneal spaces can also cause intravascular hypovolemia. Increases in vascular permeability and/or a reduction in oncotic pressure (hypoalbuminemia) alter Starling forces, resulting in excessive “third spacing” of the ECFV. This occurs in sepsis syndrome, burns, pancreatitis, nutritional hypoalbuminemia, and peritonitis. Alternatively, distributive hypovolemia can occur due to accumulation of fluid within specific compartments, for example within the bowel lumen in gastrointestinal obstruction or ileus. Hypovolemia can also occur after extracorporeal hemorrhage or after significant hemorrhage into an expandable space, for example, the retroperitoneum.

Diagnostic Evaluation

A careful history will usually determine the etiologic cause of hypovolemia. Symptoms of hypovolemia are nonspecific and include fatigue, weakness, thirst, and postural dizziness; more severe symptoms and signs include oliguria, cyanosis, abdominal and chest pain, and confusion or obtundation. Associated electrolyte disorders may cause additional symptoms, for example, muscle weakness in patients with hypokalemia. On examination, diminished skin turgor and dry oral mucous membranes are less than ideal markers of a decreased ECFV in adult patients; more reliable signs of hypovolemia include a decreased jugular venous pressure (JVP), orthostatic tachycardia (an increase of >15–20 beats/min upon standing), and orthostatic hypotension (a >10–20 mmHg drop in blood pressure on standing). More severe fluid loss leads to hypovolemic shock, with hypotension, tachycardia, peripheral vasoconstriction, and peripheral hypoperfusion; these patients may exhibit peripheral cyanosis, cold extremities, oliguria, and altered mental status.

Routine chemistries may reveal an increase in blood urea nitrogen (BUN) and creatinine, reflective of a decrease in GFR. Creatinine is the more dependable measure of GFR, because BUN levels may be influenced by an increase in tubular reabsorption (“prerenal azotemia”), an increase in urea generation in catabolic states, hyperalimentation, or gastrointestinal bleeding, and/or a decreased urea generation in decreased protein intake. In hypovolemic shock, liver function tests and cardiac biomarkers may show evidence of hepatic and cardiac ischemia, respectively. Routine chemistries and/or blood gases may reveal evidence of acid-base disorders. For example, bicarbonate loss due to diarrheal illness is a very common cause of metabolic acidosis; alternatively, patients with severe hypovolemic shock may develop lactic acidosis with an elevated anion gap.

The neurohumoral response to hypovolemia stimulates an increase in renal tubular Na^+ and water reabsorption. Therefore, the urine Na^+ concentration is typically <20 m*M* in nonrenal causes of hypovolemia, with a urine osmolality of >450 mOsm/kg. The reduction in both GFR and distal tubular Na^+ delivery may cause a defect in renal potassium excretion, with an increase in plasma K^+ concentration. Of note, patients with hypovolemia and a hypochloremic alkalosis due to vomiting, diarrhea, or diuretics will typically have a urine Na^+ concentration >20 m*M* and urine pH of >7.0, due to the

increase in filtered HCO_3^-; the urine Cl^- concentration in this setting is a more accurate indicator of volume status, with a level <25 m*M* suggestive of hypovolemia. The urine Na^+ concentration is often >20 m*M* in patients with *renal* causes of hypovolemia, such as acute tubular necrosis; similarly, patients with DI will have an inappropriately dilute urine.

TREATMENT

TREATMENT

Hypovolemia

The therapeutic goals in hypovolemia are to restore normovolemia and replace ongoing fluid losses. Mild hypovolemia can usually be treated with oral hydration and resumption of a normal maintenance diet. More severe hypovolemia requires intravenous hydration, tailoring the choice of solution to the underlying pathophysiology. Isotonic, "normal" saline (0.9% NaCl, 154 m*M* Na^+) is the most appropriate resuscitation fluid for normonatremic or hyponatremic patients with severe hypovolemia; colloid solutions such as intravenous albumin are not demonstrably superior for this purpose. Hypernatremic patients should receive a hypotonic solution, 5% dextrose if there has only been water loss (as in DI), or hypotonic saline (1/2 or 1/4 normal saline) if there has been water and Na^+-Cl^- loss; changes in free water administration should be made if necessary, based on frequent measuring of serum chemistries. Patients with bicarbonate loss and metabolic acidosis, as occur frequently in diarrhea, should receive intravenous bicarbonate, either an isotonic solution (150 meq of Na^+-HCO_3^- in 5% dextrose) or a more hypotonic bicarbonate solution in dextrose or dilute saline. Patients with severe hemorrhage or anemia should receive red cell transfusions, without increasing the hematocrit beyond 35%.

SODIUM DISORDERS

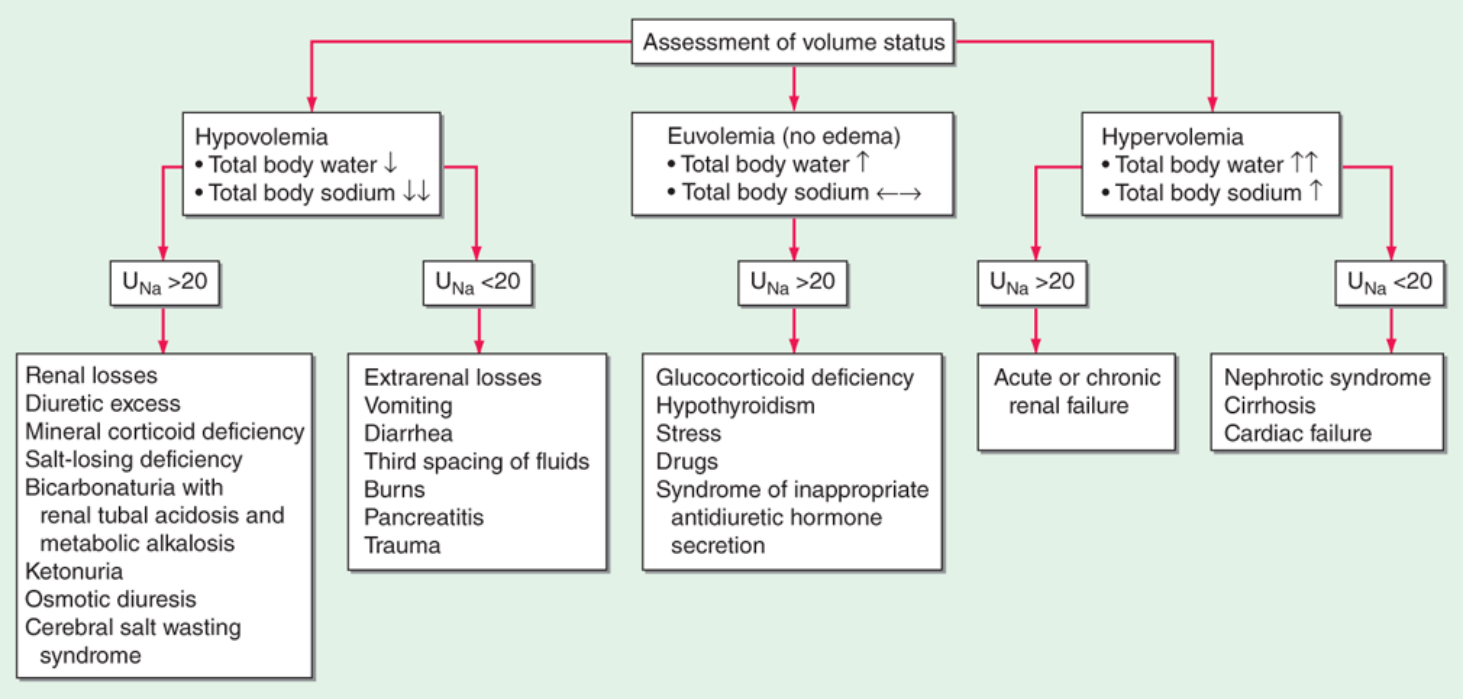
Disorders of serum Na^+ concentration are caused by abnormalities in water homeostasis, leading to changes in the relative ratio of Na^+ to body water. Water intake and circulating AVP constitute the two key effectors in the defense of serum osmolality; defects in one or both of these two defense mechanisms cause most cases of hyponatremia and hypernatremia. In contrast, abnormalities in sodium homeostasis *per se* lead to a deficit or surplus of whole-body Na^+-Cl^- content, a key determinant of the ECFV and circulatory integrity. Notably, volume status also modulates the release of AVP by the posterior pituitary, such that hypovolemia is associated with higher circulating levels of the hormone at each level of serum osmolality. Similarly, in "hypervolemic" causes of arterial underfilling, e.g., heart failure and cirrhosis, the associated neurohumoral activation encompasses an increase in circulating AVP, leading to water retention and hyponatremia. Therefore, a key concept in sodium disorders is that the absolute plasma Na^+ concentration tells one nothing about the volume status of a given patient, which furthermore must be taken into account in the diagnostic and therapeutic approach.

HYPONATREMIA

Hyponatremia, which is defined as a plasma Na^+ concentration <135 m*M*, is a very common disorder, occurring in up to 22% of hospitalized patients. This disorder is almost always the result of an increase in circulating AVP and/or increased renal sensitivity to AVP, combined with an intake of free water; a notable exception is hyponatremia due to low solute intake (see below). The underlying pathophysiology for the exaggerated or "inappropriate" AVP response differs in patients with hyponatremia as a function of their ECFV. Hyponatremia is thus subdivided diagnostically into three groups, depending on clinical history and volume status, i.e., "hypovolemic," "euvolemic," and "hypervolemic" **(Fig. 49-5)**.

FIGURE 49-5

The diagnostic approach to hyponatremia. *(From S Kumar, T Berl: Diseases of water metabolism, in Atlas of Diseases of the Kidney, RW Schrier [ed]. Philadelphia, Current Medicine, Inc, 1999; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Hypovolemic Hyponatremia

Hypovolemia causes a marked neurohumoral activation, increasing circulating levels of AVP. The increase in circulating AVP helps preserve blood pressure via vascular and baroreceptor V_{1A} receptors and increases water reabsorption via renal V_2 receptors; activation of V_2 receptors can lead to hyponatremia in the setting of increased free water intake. Nonrenal causes of hypovolemic hyponatremia include GI loss (e.g., vomiting, diarrhea, tube drainage) and insensible loss (sweating, burns) of Na^+-Cl^- and water, in the absence of adequate oral replacement; urine Na^+ concentration is typically <20 m*M*. Notably, these patients may be clinically classified as euvolemic, with only the reduced urinary Na^+ concentration to indicate the cause of their hyponatremia. Indeed, a urine Na^+ concentration <20 m*M*, in the absence of a cause of hypervolemic hyponatremia, predicts a rapid increase in plasma Na^+ concentration in response to intravenous normal saline; saline therapy thus induces a water diuresis in this setting, as circulating AVP levels plummet.

The *renal* causes of hypovolemic hyponatremia share an inappropriate loss of Na^+-Cl^- in the urine, leading to volume depletion and an increase in circulating AVP; urine Na^+ concentration is typically >20 m*M* (Fig. 49-5). A deficiency in circulating aldosterone and/or its renal effects can lead to hyponatremia in primary adrenal insufficiency and other causes of hypoaldosteronism; hyperkalemia and hyponatremia in a hypotensive and/or hypovolemic patient with high urine Na^+ concentration (much greater than 20 m*M*) should strongly suggest this diagnosis. Salt-losing nephropathies may lead to hyponatremia when sodium intake is reduced, due to impaired renal tubular function; typical causes include reflux nephropathy, interstitial nephropathies, postobstructive uropathy, medullary cystic disease, and the recovery phase of acute tubular necrosis. Thiazide diuretics cause hyponatremia via a number of mechanisms, including polydipsia and diuretic-induced volume depletion. Notably, thiazides do not inhibit the renal concentrating mechanism, such that circulating AVP retains a full effect on renal water retention. In contrast, loop diuretics, which are less frequently associated with hyponatremia, inhibit Na^+-Cl^- and K^+ absorption by the TALH, blunting the countercurrent mechanism and reducing the ability to concentrate the urine. Increased excretion of an osmotically active nonreabsorbable or poorly reabsorbable solute can also lead to volume depletion and hyponatremia; important causes include glycosuria, ketonuria (e.g., in starvation or in diabetic or alcoholic ketoacidosis), and bicarbonaturia (e.g., in renal tubular acidosis or metabolic alkalosis, where the associated bicarbonaturia leads to loss of Na^+).

Finally, the syndrome of "cerebral salt wasting" is a rare cause of hypovolemic hyponatremia, encompassing hyponatremia with clinical hypovolemia and inappropriate natriuresis in association with intracranial disease; associated disorders include subarachnoid hemorrhage, traumatic brain injury, craniotomy, encephalitis, and meningitis. Distinction from the more common syndrome of inappropriate antidiuresis (SIAD) is critical because cerebral salt wasting will typically respond to aggressive Na^+-Cl^- repletion.

Hypervolemic Hyponatremia

Patients with hypervolemic hyponatremia develop an increase in total-body Na^+-Cl^- that is accompanied by a proportionately *greater* increase in total-body water, leading to a reduced plasma Na^+ concentration. As in hypovolemic hyponatremia, the causative disorders can be separated by the effect on urine Na^+ concentration, with acute or chronic renal failure uniquely associated with an increase in urine Na^+ concentration (Fig. 49-5). The pathophysiology of hyponatremia in the sodium-avid edematous disorders (congestive heart failure [CHF], cirrhosis, and nephrotic syndrome) is similar to that in hypovolemic hyponatremia, except that arterial filling and circulatory integrity is decreased due to the specific etiologic factors (e.g., cardiac dysfunction in CHF, peripheral vasodilation in cirrhosis). Urine Na^+ concentration is typically very low, i.e., <10 m*M*, even after hydration with normal saline; this Na^+-avid state may be obscured by diuretic therapy. The degree of hyponatremia provides an indirect index of the associated neurohumoral activation and is an important prognostic indicator in hypervolemic hyponatremia.

Euvolemic Hyponatremia

Euvolemic hyponatremia can occur in moderate to severe hypothyroidism, with correction after achieving a euthyroid state. Severe hyponatremia can also be a consequence of secondary adrenal insufficiency due to pituitary disease; whereas the deficit in circulating aldosterone in primary adrenal insufficiency causes *hypovolemic* hyponatremia, the predominant glucocorticoid deficiency in secondary adrenal failure is associated with *euvolemic* hyponatremia. Glucocorticoids exert a negative feedback on AVP release by the posterior pituitary such that hydrocortisone replacement in these patients can rapidly normalize the AVP response to osmolality, reducing circulating AVP.

The SIAD is the most frequent cause of euvolemic hyponatremia **(Table 49-1)**. The generation of hyponatremia in SIAD requires an intake of free water, with persistent intake at serum osmolalities that are lower than the usual threshold for thirst; as one would expect, the osmotic threshold and osmotic response curves for the sensation of thirst are shifted downward in patients with SIAD. Four distinct patterns of AVP secretion have been recognized in patients with SIAD, independent for the most part of the underlying cause. Unregulated, erratic AVP secretion is seen in about a third of patients, with no obvious correlation between serum osmolality and circulating AVP levels. Other patients fail to suppress AVP secretion at lower serum osmolalities, with a normal response curve to hyperosmolar conditions; others have a “reset osmostat,” with a lower threshold osmolality and a left-shifted osmotic response curve. Finally, the fourth subset of patients have essentially no detectable circulating AVP, suggesting either a gain in function in renal water reabsorption or a circulating antidiuretic substance that is distinct from AVP. Gain-in-function mutations of a single specific residue in the V_2 AVP receptor have been described in some of these patients, leading to constitutive activation of the receptor in the absence of AVP and “nephrogenic” SIAD.

TABLE 49-1

Causes of the Syndrome of Inappropriate Antidiuresis (SIAD)

Malignant Diseases	Pulmonary Disorders	Disorders of the Central Nervous System	Drugs	Other Causes
Carcinoma Lung Small cell Mesothelioma Oropharynx Gastrointestinal tract Stomach Duodenum Pancreas Genitourinary tract Ureter Bladder Prostate Endometrium Endocrine thymoma Lymphomas Sarcomas Ewing's sarcoma	Infections Bacterial pneumonia Viral pneumonia Pulmonary abscess Tuberculosis Aspergillosis Asthma Cystic fibrosis Respiratory failure associated with positive-pressure breathing	Infection Encephalitis Meningitis Brain abscess Rocky Mountain spotted fever AIDS Bleeding and masses Subdural hematoma Subarachnoid hemorrhage Cerebrovascular accident Brain tumors Head trauma Hydrocephalus Cavernous sinus thrombosis Other Multiple sclerosis Guillain-Barré syndrome Shy-Drager syndrome Delirium tremens Acute intermittent porphyria	Drugs that stimulate release of AVP or enhance its action Chlorpropamide SSRIs Tricyclic antidepressants Clofibrate Carbamazepine Vincristine Nicotine Narcotics Antipsychotic drugs Ifosfamide Cyclophosphamide Nonsteroidal anti-inflammatory drugs MDMA ("ecstasy") AVP analogues Desmopressin Oxytocin Vasopressin	Hereditary (gain-of-function mutations in the vasopressin V_2 receptor) Idiopathic Transient Endurance exercise General anesthesia Nausea Pain Stress

Abbreviations: AVP, vasopressin; MDMA; 3,4-methylenedioxymethamphetamine; SSRI, selective serotonin reuptake inhibitor.

Source: From DH Ellison, T Berl: Syndrome of inappropriate antidiuresis. N Engl J Med 356:2064, 2007.

Strictly speaking, patients with SIAD are not euvolemic but are subclinically volume-expanded, due to AVP-induced water and Na^+-Cl^- retention; "AVP escape" mechanisms invoked by sustained increases in AVP serve to limit distal renal tubular transport, preserving a modestly hypervolemic steady state. Serum uric acid is often low (<4 mg/dL) in patients with SIAD, consistent with suppressed proximal tubular transport in the setting of increased distal tubular Na^+-Cl^- and water transport; in contrast, patients with hypovolemic hyponatremia will often be hyperuricemic, due to a shared activation of proximal tubular Na^+-Cl^- and urate transport.

Common causes of SIAD include pulmonary disease (e.g., pneumonia, tuberculosis, pleural effusion) and central nervous system (CNS) diseases (e.g., tumor, subarachnoid hemorrhage, meningitis). SIAD also occurs with malignancies, most commonly with small-cell lung carcinoma (75% of malignancy-associated SIAD); ~10% of patients with this tumor will have a plasma Na^+ concentration of <130 m*M* at presentation. SIAD is also a frequent complication of certain drugs, most commonly the selective serotonin reuptake inhibitors (SSRIs). Other drugs can potentiate the renal effect of AVP, without exerting direct effects on circulating AVP levels (Table 49-1).

Low Solute Intake and Hyponatremia

Hyponatremia can occasionally occur in patients with a very low intake of dietary solutes. Classically, this occurs in alcoholics whose sole nutrient is beer, hence the diagnostic label of *beer potomania*; beer is very low in protein and salt content, containing only 1–2 m*M* of Na^+. The syndrome has also been described in nonalcoholic patients with highly restricted solute intake due to nutrient-restricted diets, e.g., extreme vegetarian diets. Patients with hyponatremia due to low solute intake typically present with a very low urine osmolality (<100–200 mOsm/kg) with a urine Na^+ concentration that is <10–20 m*M*. The fundamental abnormality is the inadequate dietary intake of solutes; the reduced urinary solute excretion limits water excretion such that hyponatremia ensues after relatively modest polydipsia. AVP levels have not been reported in patients with beer potomania but are expected to be suppressed or rapidly suppressible with saline hydration; this fits with the overly rapid correction in plasma Na^+ concentration that can be seen with saline hydration. Resumption of a normal diet and/or saline hydration will also correct the causative deficit in urinary solute excretion, such that patients with beer potomania typically correct their plasma Na^+ concentration promptly after admission to the hospital.

Clinical Features of Hyponatremia

Hyponatremia induces generalized cellular swelling, a consequence of water movement down the osmotic gradient from the hypotonic ECF to the ICF. The symptoms of hyponatremia are primarily neurologic, reflecting the development of cerebral edema within a rigid skull. The initial CNS response to acute hyponatremia is an increase in interstitial pressure, leading to shunting of ECF and solutes from the interstitial space into the cerebrospinal fluid and then on into the systemic circulation. This is accompanied by an efflux of the major intracellular ions, Na^+, K^+, and Cl^-, from brain cells. Acute hyponatremic encephalopathy ensues when these volume regulatory mechanisms are overwhelmed by a rapid decrease in tonicity, resulting in acute cerebral edema. Early symptoms can include nausea, headache, and vomiting. However, severe complications can rapidly evolve, including seizure activity, brainstem herniation, coma, and death. A key complication of acute hyponatremia is normocapneic or hypercapneic respiratory failure; the associated hypoxia may amplify the neurologic injury. Normocapneic respiratory failure in this setting is typically due to noncardiogenic, "neurogenic" pulmonary edema, with a normal pulmonary capillary wedge pressure.

Acute symptomatic hyponatremia is a medical emergency, occurring in a number of specific settings **(Table 49-2)**. Women, particularly before menopause, are much more likely than men to develop encephalopathy and severe neurologic sequelae. Acute hyponatremia often has an iatrogenic component, e.g., when hypotonic intravenous fluids are given to postoperative patients with an increase in circulating AVP. Exercise-associated hyponatremia, an important clinical issue at marathons and other endurance events, has similarly been linked to both a "nonosmotic" increase in circulating AVP and excessive free water intake. The recreational drugs Molly and ecstasy, which share an active ingredient (MDMA, 3,4-methylenedioxymethamphetamine), cause a rapid and potent induction of both thirst and AVP, leading to severe acute hyponatremia.

TABLE 49-2

Causes of Acute Hyponatremia

Iatrogenic
Postoperative: premenopausal women
Hypotonic fluids with cause of ↑ vasopressin
Glycine irrigation: TURP, uterine surgery
Colonoscopy preparation
Recent institution of thiazides
Polydipsia
MDMA ("ecstasy," "Molly") ingestion
Exercise induced
Multifactorial, e.g., thiazide and polydipsia

Abbreviations: MDMA, 3,4-methylenedioxymethamphetamine; TURP, transurethral resection of the prostate.

Persistent, chronic hyponatremia results in an efflux of organic osmolytes (creatine, betaine, glutamate, myoinositol, and taurine) from brain cells; this response reduces intracellular osmolality and the osmotic gradient favoring water entry. This reduction in intracellular osmolytes is largely complete within 48 h, the time period that clinically defines chronic hyponatremia; this temporal definition has considerable relevance for the treatment of hyponatremia (see below). The cellular response to chronic hyponatremia does not fully protect patients from symptoms, which can include vomiting, nausea, confusion, and seizures, usually at plasma Na^+ concentration <125 m*M*. Even patients who are judged "asymptomatic" can manifest subtle gait and cognitive defects that reverse with correction of hyponatremia; notably, chronic "asymptomatic" hyponatremia increases the risk of falls. Chronic hyponatremia also increases the risk of bony fractures owing to the associated neurologic dysfunction and to a hyponatremia-associated reduction in bone density. Therefore, every attempt should be made to safely correct the plasma Na^+ concentration in patients with chronic hyponatremia, even in the absence of overt symptoms (see the section on treatment of hyponatremia below).

The management of chronic hyponatremia is complicated significantly by the asymmetry of the cellular response to correction of plasma Na^+ concentration. Specifically, the *reaccumulation* of organic osmolytes by brain cells is attenuated and delayed as osmolality increases after correction of hyponatremia, sometimes resulting in degenerative loss of oligodendrocytes and an osmotic demyelination syndrome (ODS). Overly rapid correction of hyponatremia (>8–10 m*M* in 24 h or 18 m*M* in 48 h) is also associated with a disruption in integrity of the blood-brain barrier, allowing the entry of immune mediators that may contribute to demyelination. The lesions of ODS classically affect the pons, a neuroanatomic structure wherein the delay in the reaccumulation of osmotic osmolytes is particularly pronounced; clinically, patients with central pontine myelinolysis can present 1 or more days after overcorrection of hyponatremia with paraparesis or quadriparesis, dysphagia, dysarthria, diplopia, a "locked-in syndrome," and/or loss of consciousness. Other regions of the brain can also be involved in ODS, most commonly in association with lesions of the pons but occasionally in isolation; in order of frequency, the lesions of extrapontine myelinolysis can occur in the cerebellum, lateral geniculate body, thalamus, putamen, and cerebral cortex or subcortex. Clinical presentation of ODS can, therefore, vary as a function of the extent and localization of extrapontine myelinolysis, with the reported development of ataxia, mutism, parkinsonism, dystonia, and catatonia. Relowering of plasma Na^+ concentration after overly rapid correction can prevent or attenuate ODS (see the section on treatment of hyponatremia below). However, even appropriately slow correction can be associated with ODS, particularly in patients with additional risk factors; these include alcoholism, malnutrition, hypokalemia, and liver transplantation.

Diagnostic Evaluation of Hyponatremia

Clinical assessment of hyponatremic patients should focus on the underlying cause; a detailed drug history is particularly crucial (Table 49-1). A careful clinical assessment of volume status is obligatory for the classical diagnostic approach to hyponatremia (Fig. 49-5). Hyponatremia is frequently multifactorial, particularly when severe; clinical evaluation should consider *all* the possible causes for excessive circulating AVP, including volume status, drugs, and the presence of nausea and/or pain. Radiologic imaging may also be appropriate to assess whether patients have a pulmonary or CNS cause for hyponatremia. A screening chest x-ray may fail to detect a small-cell carcinoma of the lung; computed tomography (CT) scanning of the thorax should be considered in patients at high risk for this tumor (e.g., patients with a smoking history).

Laboratory investigation should include a measurement of serum osmolality to exclude pseudohyponatremia, which is defined as the coexistence of hyponatremia with a normal or increased plasma tonicity. Most clinical laboratories measure plasma Na^+ concentration by testing diluted samples with automated ion-sensitive electrodes, correcting for this dilution by assuming that plasma is 93% water. This correction factor can be inaccurate in patients with pseudohyponatremia due to extreme hyperlipidemia and/or hyperproteinemia, in whom serum lipid or protein makes up a greater percentage of plasma volume. The measured osmolality should also be converted to the effective osmolality (tonicity) by subtracting the measured concentration of urea (divided by 2.8, if in mg/dL); patients with hyponatremia have an effective osmolality of <275 mOsm/kg.

Elevated BUN and creatinine in routine chemistries can also indicate renal dysfunction as a potential cause of hyponatremia, whereas hyperkalemia may suggest adrenal insufficiency or hypoaldosteronism. Serum glucose should also be measured; plasma Na^+ concentration falls by ~1.6–2.4 m*M* for every 100-mg/dL increase in glucose, due to glucose-induced water efflux from cells; this "true" hyponatremia resolves after correction of hyperglycemia. Measurement of serum uric acid should also be performed; whereas patients with SIAD-type physiology will typically be hypouricemic (serum uric acid <4 mg/dL), volume-depleted patients will often be hyperuricemic. In the appropriate clinical setting, thyroid, adrenal, and pituitary function should also be tested; hypothyroidism and secondary adrenal failure due to pituitary insufficiency are important causes of euvolemic hyponatremia, whereas primary adrenal failure causes hypovolemic hyponatremia. A cosyntropin stimulation test is necessary to assess for primary adrenal insufficiency.

Urine electrolytes and osmolality are crucial tests in the initial evaluation of hyponatremia. A urine Na^+ concentration <20–30 m*M* is consistent with hypovolemic hyponatremia, in the clinical absence of a hypervolemic, Na^+-avid syndrome such as CHF (Fig. 49-5). In contrast, patients with SIAD will typically excrete urine with an Na^+ concentration that is >30 m*M*. However, there can be substantial overlap in urine Na^+ concentration values in patients with SIAD and hypovolemic hyponatremia, particularly in the elderly; the ultimate "gold standard" for the diagnosis of hypovolemic hyponatremia is the demonstration that plasma Na^+ concentration corrects after hydration with normal saline. Patients with thiazide-associated hyponatremia may also present with higher than expected urine Na^+ concentration and other findings suggestive of SIAD; one should defer making a diagnosis of SIAD in these patients until 1–2 weeks after discontinuing the thiazide. A urine osmolality <100 mOsm/kg is suggestive of polydipsia; urine osmolality >400 mOsm/kg indicates that AVP excess is playing a more dominant role, whereas intermediate values are more consistent with multifactorial pathophysiology (e.g., AVP excess with a significant component of polydipsia). Patients with hyponatremia due to decreased solute intake (beer potomania) typically have urine Na^+ concentration <20 m*M* and urine osmolality in the range of <100 to the low 200s. Finally, the measurement of urine K^+ concentration is required to calculate the urine-to-plasma electrolyte ratio, which is useful to predict the response to fluid restriction (see the section on treatment of hyponatremia below).

TREATMENT

TREATMENT

Hyponatremia

Three major considerations guide the therapy of hyponatremia. First, the presence and/or severity of symptoms determine the urgency and goals of therapy. Patients with acute hyponatremia (Table 49-2) present with symptoms that can range from headache, nausea, and/or vomiting, to seizures, obtundation, and central herniation; patients with chronic hyponatremia, present for >48 h, are less likely to have severe symptoms. Second, patients with chronic hyponatremia are at risk for ODS if plasma Na^+ concentration is corrected by >8–10 m*M* within the first 24 h and/or by >18 m*M* within the first 48 h. Third, the response to interventions such as hypertonic saline, isotonic saline, or AVP antagonists can be highly unpredictable, such that frequent monitoring of plasma Na^+ concentration during corrective therapy is imperative.

Once the urgency in correcting the plasma Na^+ concentration has been established and appropriate therapy instituted, the focus should be on treatment or withdrawal of the underlying cause. Patients with euvolemic hyponatremia due to SIAD, hypothyroidism, or secondary adrenal failure will respond to successful treatment of the underlying cause, with an increase in plasma Na^+ concentration. However, not all causes of SIAD are immediately reversible, necessitating pharmacologic therapy to increase the plasma Na^+ concentration (see below). Hypovolemic hyponatremia will respond to intravenous hydration with isotonic normal saline, with a rapid reduction in circulating AVP and a brisk water diuresis; it may be necessary to reduce the rate of correction if the history suggests that hyponatremia has been chronic, i.e., present for more than 48 h (see below). Hypervolemic hyponatremia due to CHF will often respond to improved therapy of the underlying cardiomyopathy, e.g., following the institution or intensification of angiotensin-converting enzyme (ACE) inhibition. Finally, patients with hyponatremia due to beer potomania and low solute intake will respond very rapidly to intravenous saline and the resumption of a normal diet. Notably, patients with beer potomania have a very high risk of developing ODS, due to the associated hypokalemia, alcoholism, malnutrition, and high risk of overcorrecting the plasma Na^+ concentration.

Water deprivation has long been a cornerstone of the therapy of chronic hyponatremia. However, patients who are excreting minimal electrolyte-free water will require aggressive fluid restriction; this can be very difficult for patients with SIAD to tolerate, given that their thirst is also inappropriately stimulated. The urine-to-plasma electrolyte ratio (urinary $[Na^+] + [K^+]$/plasma $[Na^+]$) can be exploited as a quick indicator of electrolyte-free water excretion **(Table 49-3)**; patients with a ratio of >1 should be more aggressively restricted (<500 mL/d), those with a ratio of ~1 should be restricted to 500–700 mL/d, and those with a ratio <1 should be restricted to <1 L/d. In hypokalemic patients, potassium replacement will serve to increase plasma Na^+ concentration, given that the plasma Na^+ concentration is a functional of both exchangeable Na^+ and exchangeable K^+ divided by total-body water; a corollary is that aggressive repletion of K^+ has the potential to overcorrect the plasma Na^+ concentration even in the absence of hypertonic saline. Plasma Na^+ concentration will also tend to respond to an increase in dietary solute intake, which increases the ability to excrete free water; this can be accomplished with oral salt tablets and with newly available, palatable preparations of oral urea.

Patients in whom therapy with fluid restriction, potassium replacement, and/or increased solute intake fails may merit pharmacologic therapy to increase their plasma Na^+ concentration. Many patients with SIAD respond to combined therapy with oral furosemide, 20 mg twice a day (higher doses may be necessary in renal insufficiency), and oral salt tablets; furosemide serves to inhibit the renal countercurrent mechanism and blunt urinary concentrating ability, whereas the salt tablets counteract diuretic-associated natriuresis. Demeclocycline is a potent inhibitor of principal cells and can be used in patients whose Na levels do not increase in response to furosemide and salt tablets. However, this agent can be associated with a reduction in GFR, due to excessive natriuresis and/or direct renal toxicity; it should be avoided in cirrhotic patients in particular, who are at higher risk of nephrotoxicity due to drug accumulation. If available, palatable preparations of oral urea can also be used to manage SIAD; the increase in solute excretion with oral urea ingestion increases free water excretion, thus reducing the plasma Na^+.

AVP antagonists (vaptans) are highly effective in SIAD and in hypervolemic hyponatremia due to heart failure or cirrhosis, reliably increasing plasma Na^+ concentration due to their “aquaretic” effects (augmentation of free water clearance). Most of these agents specifically antagonize the V_2 AVP receptor; tolvaptan is currently the only oral V_2 antagonist to be approved by the U.S. Food and Drug Administration. Conivaptan, the only available intravenous vaptan, is a mixed V_{1A}/V_2 antagonist, with a modest risk of hypotension due to V_{1A} receptor inhibition. Therapy with vaptans must be initiated in a hospital setting, with a liberalization of fluid restriction (>2 L/d) and close monitoring of plasma Na^+ concentration. Although approved for the management of all but hypovolemic hyponatremia and acute hyponatremia, the clinical indications are limited. Oral tolvaptan is perhaps most appropriate for the management of significant and persistent SIAD (e.g., in small-cell lung carcinoma) that has not responded to water restriction and/or oral furosemide and salt tablets. Abnormalities in liver function tests have been reported with chronic tolvaptan therapy; hence, the use of this agent should be restricted to <1–2 months.

Treatment of acute symptomatic hyponatremia should include hypertonic 3% saline (513 m*M*) to acutely increase plasma Na^+ concentration by 1–2 m*M*/h to a total of 4–6 m*M*; this modest increase is typically sufficient to alleviate severe acute symptoms, after which corrective guidelines for chronic hyponatremia are appropriate (see below). A number of equations have been developed to estimate the required rate of hypertonic saline, which has an Na^+-Cl^- concentration of 513 m*M*. The traditional approach is to calculate an Na^+ deficit, where the Na^+ deficit = 0.6 × body weight × (target plasma Na^+ concentration – starting plasma Na^+ concentration), followed by a calculation of the required rate. Regardless of the method used to determine the

rate of administration, the increase in plasma Na^+ concentration can be highly unpredictable during treatment with hypertonic saline, due to rapid changes in the underlying physiology; plasma Na^+ concentration should be monitored every 2–4 h during treatment, with appropriate changes in therapy based on the observed rate of change. The administration of supplemental oxygen and ventilatory support is also critical in acute hyponatremia, in the event that patients develop acute pulmonary edema or hypercapneic respiratory failure. Intravenous loop diuretics will help treat acute pulmonary edema and will also increase free water excretion, by interfering with the renal countercurrent multiplication system. AVP antagonists do *not* have an approved role in the management of acute hyponatremia.

The rate of correction should be comparatively slow in *chronic* hyponatremia (<8–10 m*M* in the first 24 h and <18 m*M* in the first 48 h), so as to avoid ODS; lower target rates are appropriate in patients at particular risk for ODS, such as alcoholics or hypokalemic patients. Overcorrection of the plasma Na^+ concentration can occur when AVP levels rapidly normalize, for example following the treatment of patients with chronic hypovolemic hyponatremia with intravenous saline or following glucocorticoid replacement of patients with hypopituitarism and secondary adrenal failure. Approximately 10% of patients treated with vaptans will overcorrect; the risk is increased if water intake is not liberalized. In the event that the plasma Na^+ concentration overcorrects following therapy, be it with hypertonic saline, isotonic saline, or a vaptan, hyponatremia can be safely reinduced or stabilized by the administration of the AVP *agonist* desmopressin acetate (DDAVP) and/or the administration of free water, typically intravenous D_5W; the goal is to prevent or reverse the development of ODS. Alternatively, the treatment of patients with marked hyponatremia can be initiated with the twice-daily administration of DDAVP to maintain constant AVP bioactivity, combined with the administration of hypertonic saline to slowly correct the serum sodium in a more controlled fashion, thus reducing upfront the risk of overcorrection.

TABLE 49-3

Management of Hypernatremia

Water Deficit
1. Estimate total-body water (TBW): 50% of body weight in women and 60% in men 2. Calculate free-water deficit: $[(Na^+ - 140)/140] \times TBW$ 3. Administer deficit over 48–72 h, without decrease in plasma Na^+ concentration by >10 m*M*/24 h
Ongoing Water Losses
4. Calculate free-water clearance, C_eH_2O: $C_eH_2O = V \times \left(1 - \frac{U_{Na} + U_k}{P_{Na}}\right)$ where V is urinary volume, U_{Na} is urinary $[Na^+]$, U_K is urinary $[K^+]$, and P_{Na} is plasma $[Na^+]$
Insensible Losses
5. ~10 mL/kg per day: less if ventilated, more if febrile
Total
6. Add components to determine water deficit and ongoing water loss; correct the water deficit over 48–72 h and replace daily water loss. Avoid correction of plasma $[Na^+]$ by >10 m*M*/d.

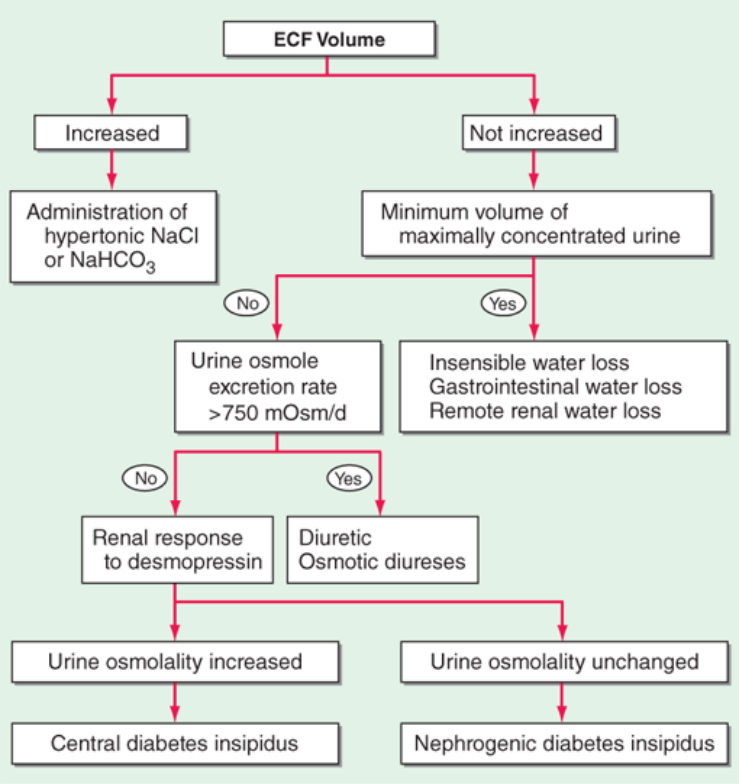
HYPERNATREMIA

Etiology

Hypernatremia is defined as an increase in the plasma Na^+ concentration to >145 m*M*. Considerably less common than hyponatremia, hypernatremia is nonetheless associated with mortality rates of as high as 40–60%, mostly due to the severity of the associated underlying disease processes. Hypernatremia is usually the result of a combined water and electrolyte deficit, with losses of H_2O in excess of Na^+. Less frequently, the ingestion or iatrogenic administration of excess Na^+ can be causative, for example after IV administration of excessive hypertonic Na^+-Cl^- or Na^+-HCO_3^- **(Fig. 49-6)**.

FIGURE 49-6

The diagnostic approach to hypernatremia. ECF, extracellular fluid.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Elderly individuals with reduced thirst and/or diminished access to fluids are at the highest risk of developing hypernatremia. Patients with hypernatremia may rarely have a central defect in hypothalamic osmoreceptor function, with a mixture of both decreased thirst and reduced AVP secretion. Causes of this adipsic DI include primary or metastatic tumor, occlusion or ligation of the anterior communicating artery, trauma, hydrocephalus, and inflammation.

Hypernatremia can develop following the loss of water via both renal and nonrenal routes. Insensible losses of water may increase in the setting of fever, exercise, heat exposure, severe burns, or mechanical ventilation. Diarrhea is, in turn, the most common gastrointestinal cause of hypernatremia. Notably, osmotic diarrhea and viral gastroenteritides typically generate stools with Na^+ and K^+ <100 m*M*, thus leading to water loss and hypernatremia; in contrast, secretory diarrhea typically results in isotonic stool and thus hypovolemia with or without hypovolemic hyponatremia.

Common causes of renal water loss include osmotic diuresis secondary to hyperglycemia, excess urea, postobstructive diuresis, or mannitol; these disorders share an increase in urinary solute excretion and urinary osmolality (see "Diagnostic Approach," below). Hypernatremia due to a water diuresis occurs in central or nephrogenic DI (NDI).

NDI is characterized by renal resistance to AVP, which can be partial or complete (see "Diagnostic Approach," below). Genetic causes include loss-of-function mutations in the X-linked V_2 receptor; mutations in the AVP-responsive aquaporin-2 water channel can cause autosomal recessive and autosomal dominant NDI, whereas recessive deficiency of the aquaporin-1 water channel causes a more modest concentrating defect (Fig. 49-2). Hypercalcemia can also cause polyuria and NDI; calcium signals directly through the calcium-sensing receptor to downregulate Na^+, K^+, and Cl^- transport by the TALH and water transport in principal cells, thus reducing renal concentrating ability in hypercalcemia. Another common acquired cause of NDI is hypokalemia, which inhibits the renal response to AVP and downregulates aquaporin-2 expression. Several drugs can cause acquired NDI, in particular lithium, ifosfamide, and several antiviral agents. Lithium causes NDI by multiple mechanisms, including direct inhibition of renal glycogen synthase kinase-3 (GSK3), a kinase thought to be the pharmacologic target of lithium in bipolar disease; GSK3 is required for the response of principal cells to AVP. The entry of lithium through the amiloride-sensitive Na^+ channel ENaC (Fig. 49-4) is required for the effect of the drug on principal cells, such that combined therapy within lithium and amiloride can mitigate lithium-associated NDI. However, lithium causes chronic tubulointerstitial scarring and chronic kidney disease after prolonged therapy, such that patients may have a persistent NDI long after stopping the drug, with a reduced therapeutic benefit from amiloride.

Finally, gestational DI is a rare complication of late-term pregnancy wherein increased activity of a circulating placental protease with "vasopressinase" activity leads to reduced circulating AVP and polyuria, often accompanied by hypernatremia. DDAVP is an effective therapy for this syndrome, given its resistance to the vasopressinase enzyme.

Clinical Features

Hypernatremia increases osmolality of the ECF, generating an osmotic gradient between the ECF and ICF, an efflux of intracellular water, and cellular shrinkage. As in hyponatremia, the symptoms of hypernatremia are predominantly neurologic. Altered mental status is the most frequent manifestation, ranging from mild confusion and lethargy to deep coma. The sudden shrinkage of brain cells in acute hypernatremia may lead to parenchymal or subarachnoid hemorrhages and/or subdural hematomas; however, these vascular complications are primarily encountered in pediatric and neonatal patients. Osmotic damage to muscle membranes can also lead to hypernatremic rhabdomyolysis. Brain cells accommodate to a chronic increase in ECF osmolality (>48 h) by activating membrane transporters that mediate influx and intracellular accumulation of organic osmolytes (creatine, betaine, glutamate, myoinositol, and taurine); this results in an increase in ICF water and normalization of brain parenchymal volume. In consequence, patients with *chronic* hypernatremia are less likely to develop severe neurologic compromise. However, the cellular response to chronic hypernatremia predisposes these patients to the development of cerebral edema and seizures during overly rapid hydration (overcorrection of plasma Na^+ concentration by >10 m*M*/d).

Diagnostic Approach

The history should focus on the presence or absence of thirst, polyuria, and/or an extrarenal source for water loss, such as diarrhea. The physical examination should include a detailed neurologic exam and an assessment of the ECFV; patients with a particularly large water deficit and/or a combined deficit in electrolytes and water may be hypovolemic, with reduced JVP and orthostasis. Accurate documentation of daily fluid intake and daily urine output is also critical for the diagnosis and management of hypernatremia.

Laboratory investigation should include a measurement of serum and urine osmolality, in addition to urine electrolytes. The appropriate response to hypernatremia and a serum osmolality >295 mOsm/kg is an increase in circulating AVP and the excretion of low volumes (<500 mL/d) of maximally concentrated urine, i.e., urine with osmolality >800 mOsm/kg; should this be the case, then an extrarenal source of water loss is primarily responsible for the generation of hypernatremia. Many patients with hypernatremia are polyuric; should an osmotic diuresis be responsible, with excessive excretion of Na^+-Cl^-, glucose, and/or urea, then daily solute excretion will be >750–1000 mOsm/d (>15 mOsm/kg body water per day) (Fig. 49-6). More commonly, patients with hypernatremia and polyuria will have a predominant water diuresis, with excessive excretion of hypotonic, dilute urine.

Adequate differentiation between nephrogenic and central causes of DI requires the measurement of the response in urinary osmolality to DDAVP, combined with measurement of circulating AVP in the setting of hypertonicity. By definition, patients with baseline hypernatremia are hypertonic, with an adequate stimulus for AVP by the posterior pituitary. Therefore, in contrast to polyuric patients with a normal or reduced baseline plasma Na^+ concentration and osmolality, a water deprivation test **(Chap. 48)** is unnecessary in hypernatremia; indeed, water deprivation is absolutely contraindicated in this setting, given the risk for worsening the hypernatremia. Patients with NDI will fail to respond to DDAVP, with a urine osmolality that increases by <50% or <150 mOsm/kg from baseline, in combination with a normal or high circulating AVP level; patients with central DI will respond to DDAVP, with a reduced circulating AVP. Patients may exhibit a partial response to DDAVP, with a >50% rise in urine osmolality that nonetheless fails to reach 800 mOsm/kg; the level of circulating AVP will help differentiate the underlying cause, i.e., NDI versus central DI. In pregnant patients, AVP assays should be drawn in tubes containing the protease inhibitor 1,10-phenanthroline, to prevent in vitro degradation of AVP by placental vasopressinase.

For patients with hypernatremia due to renal loss of water, it is critical to quantify *ongoing* daily losses using the calculated electrolyte-free water clearance, in addition to calculation of the baseline water deficit (the relevant formulas are discussed in Table 49-3). This requires daily measurement of urine electrolytes, combined with accurate measurement of daily urine volume.

TREATMENT

TREATMENT

Hypernatremia

The underlying cause of hypernatremia should be withdrawn or corrected, be it drugs, hyperglycemia, hypercalcemia, hypokalemia, or diarrhea. The approach to the correction of hypernatremia is outlined in Table 49-3. It is imperative to correct hypernatremia slowly to avoid cerebral edema, typically replacing the calculated free water deficit over 48 h. Notably, the plasma Na^+ concentration should be corrected by no >10 m*M*/d, which may take longer than 48 h in patients with severe hypernatremia (>160 m*M*). A rare exception is patients with acute hypernatremia (<48 h) due to sodium loading, who can safely be corrected rapidly at a rate of 1 m*M*/h.

Water should ideally be administered by mouth or by nasogastric tube, as the most direct way to provide free water, i.e., water without electrolytes. Alternatively, patients can receive free water in dextrose-containing IV solutions, such as 5% dextrose (D_5W); blood glucose should be monitored in case hyperglycemia occurs. Depending on the history, blood pressure, or clinical volume status, it may be appropriate to initially treat with hypotonic saline solutions (1/4 or 1/2 normal saline); normal saline is usually inappropriate in the absence of very severe hypernatremia, where normal saline is proportionally more hypotonic relative to plasma, or frank hypotension. Calculation of urinary electrolyte-free water clearance (Table 49-3) is required to estimate daily, ongoing loss of free water in patients with NDI or central DI, which should be replenished daily.

Additional therapy may be feasible in specific cases. Patients with central DI should respond to the administration of intravenous, intranasal, or oral DDAVP. Patients with NDI due to lithium may reduce their polyuria with amiloride (2.5–10 mg/d), which decreases entry of lithium into principal cells by inhibiting ENaC (see above); in practice, however, most patients with lithium-associated DI are able to compensate for their polyuria by simply increasing their daily water intake. Thiazides may reduce polyuria due to NDI, ostensibly by inducing hypovolemia and increasing proximal tubular water reabsorption. Occasionally, nonsteroidal anti-inflammatory drugs (NSAIDs) have been used to treat polyuria associated with NDI, reducing the negative effect of intrarenal prostaglandins on urinary concentrating mechanisms; however, this assumes the risks of NSAID-associated gastric and/or renal toxicity. Furthermore, it must be emphasized that thiazides, amiloride, and NSAIDs are only appropriate for *chronic* management of polyuria from NDI and have *no* role in the acute management of associated hypernatremia, where the focus is on replacing free water deficits and ongoing free water loss.

POTASSIUM DISORDERS

Homeostatic mechanisms maintain plasma K^+ concentration between 3.5 and 5.0 m*M*, despite marked variation in dietary K^+ intake. In a healthy individual at steady state, the entire daily intake of potassium is excreted, ~90% in the urine and 10% in the stool; thus, the kidney plays a dominant role in potassium homeostasis. However, >98% of total-body potassium is intracellular, chiefly in muscle; buffering of extracellular K^+ by this large intracellular pool plays a crucial role in the regulation of plasma K^+ concentration. Changes in the exchange and distribution of intra- and extracellular K^+ can thus lead to marked hypo- or hyperkalemia. A corollary is that massive necrosis and the attendant release of tissue K^+ can cause severe hyperkalemia, particularly in the setting of acute kidney injury and reduced excretion of K^+.

Changes in whole-body K^+ content are primarily mediated by the kidney, which *reabsorbs* filtered K^+ in hypokalemic, K^+-deficient states and *secretes* K^+ in hyperkalemic, K^+-replete states. Although K^+ is transported along the entire nephron, it is the principal cells of the connecting segment (CNT) and cortical CD that play a dominant role in renal K^+ secretion, whereas alpha-intercalated cells of the outer medullary CD function in renal tubular reabsorption of filtered K^+ in K^+-deficient states. In principal cells, apical Na^+ entry via the amiloride-sensitive ENaC generates a lumen-negative potential difference, which drives passive K^+ exit through apical K^+ channels (Fig. 49-4). Two major K^+ channels mediate distal tubular K^+ secretion: the secretory K^+ channel ROMK (renal outer medullary K^+ channel; also known as Kir1.1 or KcnJ1) and the flow-sensitive “big potassium” (BK) or maxi-K K^+ channel. ROMK is thought to mediate the bulk of constitutive K^+ secretion, whereas increases in distal flow rate and/or genetic absence of ROMK activate K^+ secretion via the BK channel.

An appreciation of the relationship between ENaC-dependent Na^+ entry and distal K^+ secretion (Fig. 49-4) is required for the bedside interpretation of potassium disorders. For example, decreased distal delivery of Na^+, as occurs in hypovolemic, prerenal states, tends to blunt the ability to excrete K^+, leading to hyperkalemia; on the other hand, an *increase* in distal delivery of Na^+ and distal flow rate, as occurs after treatment with thiazide and loop diuretics, can enhance K^+ secretion and lead to hypokalemia. Hyperkalemia is also a predictable consequence of drugs that directly inhibit ENaC, due to the role of this Na^+ channel in generating a lumen-negative potential difference. Aldosterone in turn has a major influence on potassium excretion, increasing the activity of ENaC channels and thus amplifying the driving force for K^+ secretion across the luminal membrane of principal cells. Abnormalities in the renin-angiotensin-aldosterone system can thus cause both hypokalemia and hyperkalemia. Notably, however, potassium excess and potassium restriction have opposing, aldosterone-independent effects on the density and activity of apical K^+ channels in the distal nephron, i.e., factors other than aldosterone modulate the renal capacity to secrete K^+. In addition, potassium restriction and hypokalemia activates aldosterone-independent distal *reabsorption* of filtered K^+, activating apical H^+/K^+-ATPase activity in intercalated cells within the outer medullary CD. Reflective perhaps of this physiology, changes in plasma K^+ concentration are not universal in disorders associated with changes in aldosterone activity.

HYPOKALEMIA

Hypokalemia, defined as a plasma K^+ concentration of <3.5 m*M*, occurs in up to 20% of hospitalized patients. Hypokalemia is associated with a tenfold increase in in-hospital mortality, due to adverse effects on cardiac rhythm, blood pressure, and cardiovascular morbidity. Mechanistically, hypokalemia can be caused by redistribution of K^+ between tissues and the ECF or by renal and nonrenal loss of K^+ **(Table 49-4)**. Systemic hypomagnesemia can also cause treatment-resistant hypokalemia, due to a combination of reduced cellular uptake of K^+ and exaggerated renal secretion. Spurious hypokalemia or “pseudohypokalemia” can occasionally result from in vitro cellular uptake of K^+ after venipuncture, for example, due to profound leukocytosis in acute leukemia.

TABLE 49-4

Causes of Hypokalemia

I. Decreased intake
- A. Starvation
- B. Clay ingestion

II. Redistribution into cells
- A. Acid-base
  1. Metabolic alkalosis
- B. Hormonal
  1. Insulin
  2. Increased β_2-adrenergic sympathetic activity: post–myocardial infarction, head injury
  3. β_2-Adrenergic agonists—bronchodilators, tocolytics
  4. α-Adrenergic antagonists
  5. Thyrotoxic periodic paralysis
  6. Downstream stimulation of Na^+/K^+-ATPase: theophylline, caffeine
- C. Anabolic state
  1. Vitamin B_{12} or folic acid administration (red blood cell production)
  2. Granulocyte-macrophage colony-stimulating factor (white blood cell production)
  3. Total parenteral nutrition
- D. Other
  1. Pseudohypokalemia
  2. Hypothermia
  3. Familial hypokalemic periodic paralysis
  4. Barium toxicity: systemic inhibition of "leak" K^+ channels

III. Increased loss
- A. Nonrenal
  1. Gastrointestinal loss (diarrhea)
  2. Integumentary loss (sweat)
- B. Renal
  1. Increased distal flow and distal Na+ delivery: diuretics, osmotic diuresis, salt-wasting nephropathies
  2. Increased secretion of potassium
     a. Mineralocorticoid excess: primary hyperaldosteronism (aldosterone-producing adenomas, primary or unilateral adrenal hyperplasia, idiopathic hyperaldosteronism due to bilateral adrenal hyperplasia, and adrenal carcinoma), genetic hyperaldosteronism (familial hyperaldosteronism types I/II/III, congenital adrenal hyperplasias), secondary hyperaldosteronism (malignant hypertension, renin-secreting tumors, renal artery stenosis, hypovolemia), Cushing's syndrome, Bartter's syndrome, Gitelman's syndrome
     b. Apparent mineralocorticoid excess: genetic deficiency of 11β-dehydrogenase-2 (syndrome of apparent mineralocorticoid excess), inhibition of 11β-dehydrogenase-2 (glycyrrhetinic/glycyrrhizinic acid and/or carbenoxolone; licorice, food products, drugs), Liddle's syndrome (genetic activation of epithelial Na^+ channels)
     c. Distal delivery of nonreabsorbed anions: vomiting, nasogastric suction, proximal renal tubular acidosis, diabetic ketoacidosis, glue-sniffing (toluene abuse), penicillin derivatives (penicillin, nafcillin, dicloxacillin, ticarcillin, oxacillin, and carbenicillin)
  3. Magnesium deficiency

Redistribution and Hypokalemia

Insulin, β_2-adrenergic activity, thyroid hormone, and alkalosis promote Na^+/K^+-ATPase-mediated cellular uptake of K^+, leading to hypokalemia.

Inhibition of the passive *efflux* of K^+ can also cause hypokalemia, albeit rarely; this typically occurs in the setting of systemic inhibition of K^+ channels by toxic barium ions. Exogenous insulin can cause iatrogenic hypokalemia, particularly during the management of K^+-deficient states such as diabetic ketoacidosis. Alternatively, the stimulation of *endogenous* insulin can provoke hypokalemia, hypomagnesemia, and/or hypophosphatemia in malnourished patients given a carbohydrate load. Alterations in the activity of the endogenous sympathetic nervous system can cause hypokalemia in several settings, including alcohol withdrawal, hyperthyroidism, acute myocardial infarction, and severe head injury. β_2 agonists, including both

bronchodilators and tocolytics (ritodrine), are powerful activators of cellular K^+ uptake; "hidden" sympathomimetics, such as pseudoephedrine and ephedrine in cough syrup or dieting agents, may also cause unexpected hypokalemia. Finally, xanthine-dependent activation of cAMP-dependent signaling, downstream of the β_2 receptor, can lead to hypokalemia, usually in the setting of overdose (theophylline) or marked overingestion (dietary caffeine).

Redistributive hypokalemia can also occur in the setting of hyperthyroidism, with periodic attacks of hypokalemic paralysis (thyrotoxic periodic paralysis [TPP]). Similar episodes of hypokalemic weakness in the absence of thyroid abnormalities occur in *familial* hypokalemic periodic paralysis, usually caused by missense mutations of voltage sensor domains within the α_1 subunit of L-type calcium channels or the skeletal Na^+ channel; these mutations generate an abnormal gating pore current activated by hyperpolarization. TPP develops more frequently in patients of Asian or Hispanic origin; this shared predisposition has been linked to genetic variation in Kir2.6, a muscle-specific, thyroid hormone–responsive K^+ channel. Patients with TPP typically present with weakness of the extremities and limb girdles, with paralytic episodes that occur most frequently between 1 and 6 A.M. Signs and symptoms of hyperthyroidism are not invariably present. Hypokalemia is usually profound and almost invariably accompanied by hypophosphatemia and hypomagnesemia. The hypokalemia in TPP is also attributed to both direct and indirect activation of the Na^+/K^+-ATPase, resulting in increased uptake of K^+ by muscle and other tissues. Increases in β-adrenergic activity play an important role in that high-dose propranolol (3 mg/kg) rapidly reverses the associated hypokalemia, hypophosphatemia, and paralysis.

Nonrenal Loss of Potassium

The loss of K^+ in sweat is typically low, except under extremes of physical exertion. Direct gastric losses of K^+ due to vomiting or nasogastric suctioning are also minimal; however, the ensuing hypochloremic alkalosis results in persistent kaliuresis due to secondary hyperaldosteronism and bicarbonaturia, i.e., a *renal* loss of K^+. Diarrhea is a globally important cause of hypokalemia, given the worldwide prevalence of infectious diarrheal disease. Noninfectious gastrointestinal processes such as celiac disease, ileostomy, villous adenomas, inflammatory bowel disease, colonic pseudo-obstruction (Ogilvie's syndrome), VIPomas, and chronic laxative abuse can also cause significant hypokalemia; an exaggerated intestinal secretion of potassium by upregulated colonic BK channels has been directly implicated in the pathogenesis of hypokalemia in many of these disorders.

Renal Loss of Potassium

Drugs can increase renal K^+ excretion by a variety of different mechanisms. Diuretics are a particularly common cause, due to associated increases in distal tubular Na^+ delivery and distal tubular flow rate, in addition to secondary hyperaldosteronism. Thiazides have a greater effect on plasma K^+ concentration than loop diuretics, despite their lesser natriuretic effect. The diuretic effect of thiazides is largely due to inhibition of the Na^+-Cl^- cotransporter NCC in DCT cells. This leads to a direct increase in the delivery of luminal Na^+ to the principal cells immediately downstream in the CNT and cortical CD, which augments Na^+ entry via ENaC, increases the lumen-negative potential difference, and amplifies K^+ secretion. The higher propensity of thiazides to cause hypokalemia may also be secondary to thiazide-associated hypocalciuria, versus the *hypercalciuria* seen with loop diuretics; the increases in downstream luminal calcium in response to loop diuretics inhibit ENaC in principal cells, thus reducing the lumen-negative potential difference and attenuating distal K^+ excretion. High doses of penicillin-related antibiotics (nafcillin, dicloxacillin, ticarcillin, oxacillin, and carbenicillin) can increase obligatory K^+ excretion by acting as nonreabsorbable anions in the distal nephron. Finally, several renal tubular toxins cause renal K^+ and magnesium wasting, leading to hypokalemia and hypomagnesemia; these drugs include aminoglycosides, amphotericin, foscarnet, cisplatin, and ifosfamide (see also "Magnesium Deficiency and Hypokalemia," below).

Aldosterone activates the ENaC channel in principal cells via multiple synergistic mechanisms, thus increasing the driving force for K^+ excretion. In consequence, increases in aldosterone bioactivity and/or gains in function of aldosterone-dependent signaling pathways are associated with hypokalemia. Increases in circulating aldosterone (hyperaldosteronism) may be primary or secondary. Increased levels of circulating renin in secondary forms of hyperaldosteronism lead to increased angiotensin II and thus aldosterone; renal artery stenosis is perhaps the most frequent cause (Table 49-4). Primary hyperaldosteronism may be genetic or acquired. Hypertension and hypokalemia, due to increases in circulating 11-deoxycorticosterone, occur in patients with congenital adrenal hyperplasia caused by defects in either steroid 11β-hydroxylase or steroid 17α-hydroxylase; deficient 11β-hydroxylase results in associated virilization and other signs of androgen excess, whereas reduced sex steroids in 17α-hydroxylase deficiency lead to hypogonadism.

The major forms of *isolated* primary genetic hyperaldosteronism are familial hyperaldosteronism type I (FH-I, also known as glucocorticoid-remediable hyperaldosteronism [GRA]) and familial hyperaldosteronism types II and III (FH-II and FH-III), in which aldosterone production is not repressible by exogenous glucocorticoids. FH-I is caused by a chimeric gene duplication between the homologous 11β-hydroxylase (*CYP11B1*) and aldosterone synthase (*CYP11B2*) genes, fusing the adrenocorticotropic hormone (ACTH)–responsive 11β-hydroxylase promoter to the coding region of aldosterone synthase; this chimeric gene is under the control of ACTH and thus repressible by glucocorticoids. FH-III is caused by mutations in the *KCNJ5* gene, which encodes the G-protein-activated inward rectifier K^+ channel 4 (GIRK4); these mutations lead to the acquisition of sodium permeability in the

mutant GIRK4 channels, causing an exaggerated membrane depolarization in adrenal glomerulosa cells and the activation of voltage-gated calcium channels. The resulting calcium influx is sufficient to produce aldosterone secretion and cell proliferation, leading to adrenal adenomas and hyperaldosteronism.

Acquired causes of primary hyperaldosteronism include aldosterone-producing adenomas (APAs), primary or unilateral adrenal hyperplasia (PAH), idiopathic hyperaldosteronism (IHA) due to bilateral adrenal hyperplasia, and adrenal carcinoma; APA and IHA account for close to 60 and 40%, respectively, of diagnosed hyperaldosteronism. Acquired somatic mutations in *KCNJ5* or less frequently in the *ATP1A1* (an Na^+/K^+ ATPase α subunit) and *ATP2B3* (a Ca^{2+} ATPase) genes can be detected in APAs; as in FH-III (see above), the exaggerated depolarization of adrenal glomerulosa cells caused by these mutations is implicated in the excessive adrenal proliferation and the exaggerated release of aldosterone.

Random testing of plasma renin activity (PRA) and aldosterone is a helpful screening tool in hypokalemic and/or hypertensive patients, with an aldosterone:PRA ratio of >50 suggestive of primary hyperaldosteronism. Hypokalemia and multiple antihypertensive drugs may alter the aldosterone:PRA ratio by suppressing aldosterone or increasing PRA, leading to a ratio of <50 in patients who do in fact have primary hyperaldosteronism; therefore, the clinical context should always be considered when interpreting these results.

The glucocorticoid cortisol has equal affinity for the MLR to that of aldosterone, with resultant “mineralocorticoid-like” activity. However, cells in the aldosterone-sensitive distal nephron are protected from this “illicit” activation by the enzyme 11β-hydroxysteroid dehydrogenase-2 (11βHSD-2), which converts cortisol to cortisone; cortisone has minimal affinity for the MLR. Recessive loss-of-function mutations in the *11βHSD-2* gene are thus associated with cortisol-dependent activation of the MLR and the syndrome of apparent mineralocorticoid excess (SAME), encompassing hypertension, hypokalemia, hypercalciuria, and metabolic alkalosis, with suppressed PRA and suppressed aldosterone. A similar syndrome is caused by biochemical inhibition of 11βHSD-2 by glycyrrhetinic/glycyrrhizinic acid and/or carbenoxolone. Glycyrrhizinic acid is a natural sweetener found in licorice root, typically encountered in licorice and its many guises or as a flavoring agent in tobacco and food products.

Finally, hypokalemia may also occur with systemic increases in glucocorticoids. In Cushing’s syndrome caused by increases in pituitary ACTH **(Chap. 379)**, the incidence of hypokalemia is only 10%, whereas it is 60–100% in patients with ectopic secretion of ACTH, despite a similar incidence of hypertension. Indirect evidence suggests that the activity of renal 11βHSD-2 is reduced in patients with ectopic ACTH compared with Cushing’s syndrome, resulting in SAME.

Finally, defects in multiple renal tubular transport pathways are associated with hypokalemia. For example, loss-of-function mutations in subunits of the acidifying H^+-ATPase in alpha-intercalated cells cause hypokalemic distal renal tubular acidosis, as do many acquired disorders of the distal nephron. Liddle’s syndrome is caused by autosomal dominant gain-in-function mutations of ENaC subunits. Disease-associated mutations either activate the channel directly or abrogate aldosterone-inhibited retrieval of ENaC subunits from the plasma membrane; the end result is increased expression of activated ENaC channels at the plasma membrane of principal cells. Patients with Liddle’s syndrome classically manifest severe hypertension with hypokalemia, unresponsive to spironolactone yet sensitive to amiloride. Hypertension and hypokalemia are, however, variable aspects of the Liddle’s phenotype; more consistent features include a blunted aldosterone response to ACTH and reduced urinary aldosterone excretion.

Loss of the transport functions of the TALH and DCT nephron segments causes hereditary hypokalemic alkalosis, Bartter’s syndrome (BS) and Gitelman’s syndrome (GS), respectively. Patients with classic BS typically suffer from polyuria and polydipsia, due to the reduction in renal concentrating ability. They may have an increase in urinary calcium excretion, and 20% are hypomagnesemic. Other features include marked activation of the renin-angiotensin-aldosterone axis. Patients with antenatal BS suffer from a severe systemic disorder characterized by marked electrolyte wasting, polyhydramnios, and hypercalciuria with nephrocalcinosis; renal prostaglandin synthesis and excretion are significantly increased, accounting for much of the systemic symptoms. There are five disease genes for BS, all of them functioning in some aspect of regulated Na^+, K^+, and Cl^- transport by the TALH. In contrast, GS is genetically homogeneous, caused almost exclusively by loss-of-function mutations in the thiazide-sensitive Na^+-Cl^- cotransporter of the DCT. Patients with GS are uniformly hypomagnesemic and exhibit marked hypocalciuria, rather than the hypercalciuria typically seen in BS; urinary calcium excretion is thus a critical diagnostic test in GS. GS is a milder phenotype than BS; however, patients with GS may suffer from chondrocalcinosis, an abnormal deposition of calcium pyrophosphate dihydrate (CPPD) in joint cartilage **(Chap. 309)**.

Magnesium Deficiency and Hypokalemia

Magnesium depletion has inhibitory effects on muscle Na^+/K^+-ATPase activity, reducing influx into muscle cells and causing a secondary kaliuresis. In addition, magnesium depletion causes exaggerated K^+ secretion by the distal nephron; this effect is attributed to a reduction in the magnesium-dependent, intracellular block of K^+ efflux through the secretory K^+ channel of principal cells (ROMK; Fig. 49-4). In consequence, hypomagnesemic patients are clinically refractory to K^+ replacement in the absence of Mg^{2+} repletion. Notably, magnesium deficiency is also a common concomitant of hypokalemia because many disorders of the distal nephron may cause both potassium and magnesium wasting **(Chap. 309)**.

Clinical Features

Hypokalemia has prominent effects on cardiac, skeletal, and intestinal muscle cells. In particular, hypokalemia is a major risk factor for both ventricular and atrial arrhythmias. Hypokalemia predisposes to digoxin toxicity by a number of mechanisms, including reduced competition between K^+ and digoxin for shared binding sites on cardiac Na^+/K^+-ATPase subunits. Electrocardiographic changes in hypokalemia include broad flat T waves, ST depression, and QT prolongation; these are most marked when serum K^+ is <2.7 mmol/L. Hypokalemia can thus be an important precipitant of arrhythmia in patients with additional genetic or acquired causes of QT prolongation. Hypokalemia also results in hyperpolarization of skeletal muscle, thus impairing the capacity to depolarize and contract; weakness and even paralysis may ensue. It also causes a skeletal myopathy and predisposes to rhabdomyolysis. Finally, the paralytic effects of hypokalemia on intestinal smooth muscle may cause intestinal ileus.

The functional effects of hypokalemia on the kidney can include Na^+-Cl^- and HCO_3^- retention, polyuria, phosphaturia, hypocitraturia, and an activation of renal ammoniagenesis. Bicarbonate retention and other acid-base effects of hypokalemia can contribute to the generation of metabolic alkalosis. Hypokalemic polyuria is due to a combination of central polydipsia and an AVP-resistant renal concentrating defect. Structural changes in the kidney due to hypokalemia include a relatively specific vacuolizing injury to proximal tubular cells, interstitial nephritis, and renal cysts. Hypokalemia also predisposes to acute kidney injury and can lead to end-stage renal disease (ESRD) in patients with long-standing hypokalemia due to eating disorders and/or laxative abuse.

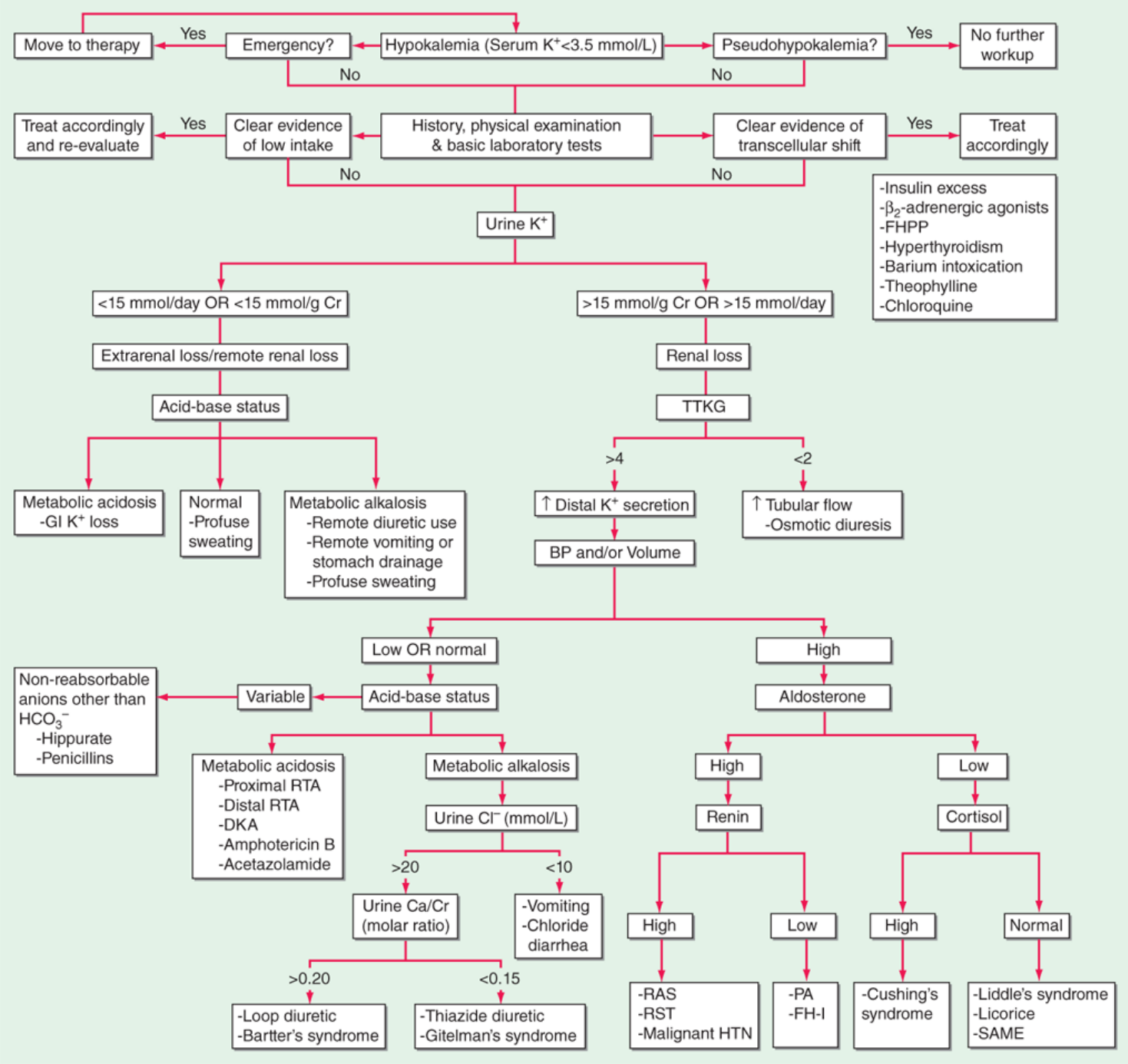
Hypokalemia and/or reduced dietary K^+ are implicated in the pathophysiology and progression of hypertension, heart failure, and stroke. For example, short-term K^+ restriction in healthy humans and patients with essential hypertension induces Na^+-Cl^- retention and hypertension. Correction of hypokalemia is particularly important in hypertensive patients treated with diuretics, in whom blood pressure improves with potassium supplementation and the establishment of normokalemia.

Diagnostic Approach

The cause of hypokalemia is usually evident from history, physical examination, and/or basic laboratory tests. The history should focus on medications (e.g., laxatives, diuretics, antibiotics), diet and dietary habits (e.g., licorice), and/or symptoms that suggest a particular cause (e.g., periodic weakness, diarrhea). The physical examination should pay particular attention to blood pressure, volume status, and signs suggestive of specific hypokalemic disorders, e.g., hyperthyroidism and Cushing's syndrome. Initial laboratory evaluation should include electrolytes, BUN, creatinine, serum osmolality, Mg^{2+}, Ca^{2+}, a complete blood count, and urinary pH, osmolality, creatinine, and electrolytes **(Fig. 49-7)**. The presence of a non–anion gap acidosis suggests a distal, hypokalemic renal tubular acidosis or diarrhea; calculation of the urinary anion gap can help differentiate these two diagnoses. Renal K^+ excretion can be assessed with a 24-h urine collection; a 24-h K^+ excretion of <15 mmol is indicative of an extrarenal cause of hypokalemia (Fig. 49-7). If only a random, spot urine sample is available, serum and urine osmolality can be used to calculate the transtubular K^+ gradient (TTKG), which should be <3 in the presence of hypokalemia (see also "Hyperkalemia"). Alternatively, a urinary K^+-to-creatinine ratio of >13 mmol/g creatinine (>1.5 mmol/mmol creatinine) is compatible with excessive renal K^+ excretion. Urine Cl^- is usually decreased in patients with hypokalemia from a nonreabsorbable anion, such as antibiotics or HCO_3^-. The most common causes of chronic hypokalemic alkalosis are surreptitious vomiting, diuretic abuse, and GS; these can be distinguished by the pattern of urinary electrolytes. Hypokalemic patients with vomiting due to bulimia will thus typically have a urinary Cl^- <10 mmol/L; urine Na^+, K^+, and Cl^- are persistently elevated in GS, due to loss of function in the thiazide-sensitive Na^+-Cl^- cotransporter, but less elevated in diuretic abuse and with greater variability. Urine diuretic screens for loop diuretics and thiazides may be necessary to further exclude diuretic abuse.

FIGURE 49-7

The diagnostic approach to hypokalemia. See text for details. AME, apparent mineralocorticoid excess; BP, blood pressure; CCD, cortical collecting duct; DKA, diabetic ketoacidosis; FH-I, familial hyperaldosteronism type I; FHPP, familial hypokalemic periodic paralysis; GI, gastrointestinal; GRA, glucocorticoid remediable aldosteronism; HTN, hypertension; PA, primary aldosteronism; RAS, renal artery stenosis; RST, renin-secreting tumor; RTA, renal tubular acidosis; SAME, syndrome of apparent mineralocorticoid excess; TTKG, transtubular potassium gradient. *(Used with permission from DB Mount, K Zandi-Nejad K: Disorders of potassium balance, in Brenner and Rector's The Kidney, 8th ed, BM Brenner [ed]. Philadelphia, W.B. Saunders & Company, 2008, pp 547–587.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Other tests, such as urinary Ca^{2+}, thyroid function tests, and/or PRA and aldosterone levels, may also be appropriate in specific cases. A plasma aldosterone:PRA ratio of >50, due to suppression of circulating renin and an elevation of circulating aldosterone, is suggestive of hyperaldosteronism. Patients with hyperaldosteronism or apparent mineralocorticoid excess may require further testing, for example adrenal vein sampling (Chap. 379) or the clinically available testing for specific genetic causes (e.g., FH-I, SAME, Liddle's syndrome). Patients with primary aldosteronism should thus be tested for the chimeric FH-I/GRA gene (see above) if they are younger than 20 years of age or have a family history of primary aldosteronism or stroke at a young age (<40 years). Preliminary differentiation of Liddle's syndrome due to mutant ENaC channels from SAME due to mutant 11βHSD-2 (see above), both of which cause hypokalemia and hypertension with aldosterone suppression, can be made on a clinical basis and then confirmed by genetic analysis; patients with Liddle's syndrome should respond to amiloride (ENaC inhibition) but not spironolactone, whereas patients with SAME will respond to spironolactone.

TREATMENT

TREATMENT

Hypokalemia

The goals of therapy in hypokalemia are to prevent life-threatening and/or serious chronic consequences, to replace the associated K^+ deficit, and to correct the underlying cause and/or mitigate future hypokalemia. The urgency of therapy depends on the severity of hypokalemia, associated clinical factors (e.g., cardiac disease, digoxin therapy), and the rate of decline in serum K^+. Patients with a prolonged QT interval and/or other risk factors for arrhythmia should be monitored by continuous cardiac telemetry during repletion. Urgent but cautious K^+ replacement should be considered in patients with severe redistributive hypokalemia (plasma K^+ concentration <2.5 m*M*) and/or when serious complications ensue; however, this approach has a risk of rebound hyperkalemia following acute resolution of the underlying cause. When excessive activity of the sympathetic nervous system is thought to play a dominant role in redistributive hypokalemia, as in TPP, theophylline overdose, and acute head injury, high-dose propranolol (3 mg/kg) should be considered; this nonspecific b-adrenergic blocker will correct hypokalemia without the risk of rebound hyperkalemia.

Oral replacement with K^+-Cl^- is the mainstay of therapy in hypokalemia. Potassium phosphate, oral or IV, may be appropriate in patients with combined hypokalemia and hypophosphatemia. Potassium bicarbonate or potassium citrate should be considered in patients with concomitant metabolic acidosis. Notably, hypomagnesemic patients are refractory to K^+ replacement alone, such that concomitant Mg^{2+} deficiency should *always* be corrected with oral or intravenous repletion. The deficit of K^+ and the rate of correction should be estimated as accurately as possible; renal function, medications, and comorbid conditions such as diabetes should also be considered, so as to gauge the risk of overcorrection. In the absence of abnormal K^+ redistribution, the total deficit correlates with serum K^+, such that serum K^+ drops by ~0.27 m*M* for every 100-mmol reduction in total-body stores; loss of 400–800 mmol of total-body K^+ results in a reduction in serum K^+ by ~2.0 m*M*. Notably, given the delay in redistributing potassium into intracellular compartments, this deficit must be replaced gradually over 24–48 h, with frequent monitoring of plasma K^+ concentration to avoid transient overrepletion and transient hyperkalemia.

The use of intravenous administration should be limited to patients unable to use the enteral route or in the setting of severe complications (e.g., paralysis, arrhythmia). Intravenous K^+-Cl^- should always be administered in saline solutions, rather than dextrose, because the dextrose-induced increase in insulin can acutely exacerbate hypokalemia. The peripheral intravenous dose is usually 20–40 mmol of K^+-Cl^- per liter; higher concentrations can cause localized pain from chemical phlebitis, irritation, and sclerosis. If hypokalemia is severe (<2.5 mmol/L) and/or critically symptomatic, intravenous K^+-Cl^- can be administered through a central vein with cardiac monitoring in an intensive care setting, at rates of 10–20 mmol/h; higher rates should be reserved for acutely life-threatening complications. The absolute amount of administered K^+ should be restricted (e.g., 20 mmol in 100 mL of saline solution) to prevent inadvertent infusion of a large dose. Femoral veins are preferable, because infusion through internal jugular or subclavian central lines can acutely increase the local concentration of K^+ and affect cardiac conduction.

Strategies to minimize K^+ losses should also be considered. These measures may include minimizing the dose of non-K^+-sparing diuretics, restricting Na^+ intake, and using clinically appropriate combinations of non-K^+-sparing and K^+-sparing medications (e.g., loop diuretics with ACE inhibitors).

HYPERKALEMIA

Hyperkalemia is defined as a plasma potassium level of 5.5 m*M*, occurring in up to 10% of hospitalized patients; severe hyperkalemia (>6.0 m*M*) occurs in ~1%, with a significantly increased risk of mortality. Although redistribution and reduced tissue uptake can acutely cause hyperkalemia, a decrease in renal K^+ excretion is the most frequent underlying cause **(Table 49-5)**. Excessive intake of K^+ is a rare cause, given the adaptive capacity to increase renal secretion; however, dietary intake can have a major effect in susceptible patients, e.g., diabetics with hyporeninemic hypoaldosteronism and chronic kidney disease. Drugs that impact on the renin-angiotensin-aldosterone axis are also a major cause of hyperkalemia.

TABLE 49-5

Causes of Hyperkalemia

I. Pseudohyperkalemia
- A. Cellular efflux; thrombocytosis, erythrocytosis, leukocytosis, in vitro hemolysis
- B. Hereditary defects in red cell membrane transport

II. Intra- to extracellular shift
- A. Acidosis
- B. Hyperosmolality; radiocontrast, hypertonic dextrose, mannitol
- C. β_2-Adrenergic antagonists (noncardioselective agents)
- D. Digoxin and related glycosides (yellow oleander, foxglove, bufadienolide)
- E. Hyperkalemic periodic paralysis
- F. Lysine, arginine, and ε-aminocaproic acid (structurally similar, positively charged)
- G. Succinylcholine; thermal trauma, neuromuscular injury, disuse atrophy, mucositis, or prolonged immobilization
- H. Rapid tumor lysis

III. Inadequate excretion
- A. Inhibition of the renin-angiotensin-aldosterone axis; ↑ risk of hyperkalemia when used in combination
  1. Angiotensin-converting enzyme (ACE) inhibitors
  2. Renin inhibitors; aliskiren (in combination with ACE inhibitors or angiotensin receptor blockers [ARBs])
  3. Angiotensin receptor blockers (ARBs)
  4. Blockade of the mineralocorticoid receptor: spironolactone, eplerenone, drospirenone
  5. Blockade of the epithelial sodium channel (ENaC): amiloride, triamterene, trimethoprim, pentamidine, nafamostat
- B. Decreased distal delivery
  1. Congestive heart failure
  2. Volume depletion
- C. Hyporeninemic hypoaldosteronism
  1. Tubulointerstitial diseases: systemic lupus erythematosus (SLE), sickle cell anemia, obstructive uropathy
  2. Diabetes, diabetic nephropathy
  3. Drugs: nonsteroidal anti-inflammatory drugs (NSAIDs), cyclooxygenase 2 (COX2) inhibitors, β-blockers, cyclosporine, tacrolimus
  4. Chronic kidney disease, advanced age
  5. Pseudohypoaldosteronism type II: defects in WNK1 or WNK4 kinases, Kelch-like 3 (KLHL3), or Cullin 3 (CUL3)
- D. Renal resistance to mineralocorticoid
  1. Tubulointerstitial diseases: SLE, amyloidosis, sickle cell anemia, obstructive uropathy, post–acute tubular necrosis
  2. Hereditary: pseudohypoaldosteronism type I; defects in the mineralocorticoid receptor *or* the epithelial sodium channel (ENaC)
- E. Advanced renal insufficiency
  1. Chronic kidney disease
  2. End-stage renal disease
  3. Acute oliguric kidney injury
- F. Primary adrenal insufficiency
  1. Autoimmune: Addison's disease, polyglandular endocrinopathy
  2. Infectious: HIV, cytomegalovirus, tuberculosis, disseminated fungal infection
  3. Infiltrative: amyloidosis, malignancy, metastatic cancer
  4. Drug-associated: heparin, low-molecular-weight heparin
  5. Hereditary: adrenal hypoplasia congenita, congenital lipoid adrenal hyperplasia, aldosterone synthase deficiency
  6. Adrenal hemorrhage or infarction, including in antiphospholipid syndrome

Pseudohyperkalemia

Hyperkalemia should be distinguished from factitious hyperkalemia or "pseudohyperkalemia," an artifactual increase in serum K^+ due to the release of K^+ during or after venipuncture. Pseudohyperkalemia can occur in the setting of excessive muscle activity during venipuncture (e.g., fist clenching), a marked increase in cellular elements (thrombocytosis, leukocytosis, and/or erythrocytosis) with in vitro efflux of K^+, and acute anxiety during venipuncture with respiratory alkalosis and redistributive hyperkalemia. Cooling of blood following venipuncture is another cause, due to reduced cellular uptake; the converse is the increased uptake of K^+ by cells at high ambient temperatures, leading to normal values for hyperkalemic patients

and/or to spurious hypokalemia in normokalemic patients. Finally, there are multiple genetic subtypes of hereditary pseudohyperkalemia, caused by increases in the passive K^+ permeability of erythrocytes. For example, causative mutations have been described in the red cell anion exchanger (AE1, encoded by the *SLC4A1* gene), leading to reduced red cell anion transport, hemolytic anemia, the acquisition of a novel AE1-mediated K^+ leak, and pseudohyperkalemia.

Redistribution and Hyperkalemia

Several different mechanisms can induce an efflux of intracellular K^+ and hyperkalemia. Acidemia is associated with cellular uptake of H^+ and an associated efflux of K^+; it is thought that this effective K^+-H^+ exchange serves to help maintain extracellular pH. Notably, this effect of acidosis is limited to non–anion gap causes of metabolic acidosis and, to a lesser extent, respiratory causes of acidosis; hyperkalemia due to an acidosis-induced shift of potassium from the cells into the ECF does *not* occur in the anion gap acidoses lactic acidosis and ketoacidosis. Hyperkalemia due to hypertonic mannitol, hypertonic saline, and intravenous immune globulin is generally attributed to a "solvent drag" effect, as water moves out of cells along the osmotic gradient. Diabetics are also prone to osmotic hyperkalemia in response to intravenous hypertonic glucose, when given without adequate insulin. Cationic amino acids, specifically lysine, arginine, and the structurally related drug epsilon-aminocaproic acid, cause efflux of K^+ and hyperkalemia, through an effective cation-K^+ exchange of unknown identity and mechanism. Digoxin inhibits Na^+/K^+-ATPase and impairs the uptake of K^+ by skeletal muscle, such that digoxin overdose predictably results in hyperkalemia. Structurally related glycosides are found in specific plants (e.g., yellow oleander, foxglove) and in the cane toad, *Bufo marinus* (bufadienolide); ingestion of these substances and extracts thereof can also cause hyperkalemia. Finally, fluoride ions also inhibit Na^+/K^+-ATPase, such that fluoride poisoning is typically associated with hyperkalemia.

Succinylcholine depolarizes muscle cells, causing an efflux of K^+ through acetylcholine receptors (AChRs). The use of this agent is contraindicated in patients who have sustained thermal trauma, neuromuscular injury, disuse atrophy, mucositis, or prolonged immobilization. These disorders share a marked increase and redistribution of AChRs at the plasma membrane of muscle cells; depolarization of these upregulated AChRs by succinylcholine leads to an exaggerated efflux of K^+ through the receptor-associated cation channels, resulting in acute hyperkalemia.

Hyperkalemia Caused by Excess Intake or Tissue Necrosis

Increased intake of even small amounts of K^+ may provoke severe hyperkalemia in patients with predisposing factors; hence, an assessment of dietary intake is crucial. Foods rich in potassium include tomatoes, bananas, and citrus fruits; occult sources of K^+, particularly K^+-containing salt substitutes, may also contribute significantly. Iatrogenic causes include simple overreplacement with K^+-Cl^- or the administration of a potassium-containing medication (e.g., K^+-penicillin) to a susceptible patient. Red cell transfusion is a well-described cause of hyperkalemia, typically in the setting of massive transfusions. Finally, severe tissue necrosis, as in acute tumor lysis syndrome and rhabdomyolysis, will predictably cause hyperkalemia from the release of intracellular K^+.

Hypoaldosteronism and Hyperkalemia

Aldosterone release from the adrenal gland may be reduced by hyporeninemic hypoaldosteronism, medications, primary hypoaldosteronism, or isolated deficiency of ACTH (secondary hypoaldosteronism). Primary hypoaldosteronism may be genetic or acquired **(Chap. 379)** but is commonly caused by autoimmunity, either in Addison's disease or in the context of a polyglandular endocrinopathy. HIV has surpassed tuberculosis as the most important infectious cause of adrenal insufficiency. The adrenal involvement in HIV disease is usually subclinical; however, adrenal insufficiency may be precipitated by stress, drugs such as ketoconazole that inhibit steroidogenesis, or the acute withdrawal of steroid agents such as megestrol.

Hyporeninemic hypoaldosteronism is a very common predisposing factor in several overlapping subsets of hyperkalemic patients: diabetics, the elderly, and patients with renal insufficiency. Classically, patients should have suppressed PRA and aldosterone; ~50% have an associated acidosis, with a reduced renal excretion of NH_4^+, a positive urinary anion gap, and urine pH <5.5. Most patients are volume expanded, with secondary increases in circulating atrial natriuretic peptide (ANP) that inhibit both renal renin release and adrenal aldosterone release.

Renal Disease and Hyperkalemia

Chronic kidney disease and end-stage kidney disease are very common causes of hyperkalemia, due to the associated deficit or absence of functioning nephrons. Hyperkalemia is more common in oliguric acute kidney injury; distal tubular flow rate and Na^+ delivery are less limiting factors in nonoliguric patients. Hyperkalemia out of proportion to GFR can also be seen in the context of tubulointerstitial disease that affects the distal nephron, such as amyloidosis, sickle cell anemia, interstitial nephritis, and obstructive uropathy.

Hereditary renal causes of hyperkalemia have overlapping clinical features with hypoaldosteronism, hence the diagnostic label *pseudohypoaldosteronism* (PHA). PHA type I (PHA-I) has both an autosomal recessive and an autosomal dominant form. The autosomal dominant form is due to loss-of-function mutations in the MLR; the recessive form is caused by various combinations of mutations in the three subunits of ENaC,

resulting in impaired Na^+ channel activity in principal cells and other tissues. Patients with recessive PHA-I suffer from lifelong salt wasting, hypotension, and hyperkalemia, whereas the phenotype of autosomal dominant PHA-I due to MLR dysfunction improves in adulthood. PHA type II (PHA-II; also known as *hereditary hypertension with hyperkalemia*) is in every respect the mirror image of GS caused by loss of function in NCC, the thiazide-sensitive Na^+-Cl^- cotransporter (see above); the clinical phenotype includes hypertension, hyperkalemia, hyperchloremic metabolic acidosis, suppressed PRA and aldosterone, hypercalciuria, and reduced bone density. PHA-II thus behaves like a gain of function in NCC, and treatment with thiazides results in resolution of the entire clinical phenotype. However, the NCC gene is not directly involved in PHA-II, which is caused by mutations in the WNK1 and WNK4 serine-threonine kinases or the upstream Kelch-like 3 (KLHL3) and Cullin 3 (CUL3) proteins, two components of an E3 ubiquitin ligase complex that regulates these kinases; these proteins collectively regulate NCC activity, with PHA-II-associated activation of the transporter.

Medication-Associated Hyperkalemia

Most medications associated with hyperkalemia cause inhibition of some component of the renin-angiotensin-aldosterone axis. ACE inhibitors, angiotensin receptor blockers, renin inhibitors, and MLRs are predictable and common causes of hyperkalemia, particularly when prescribed in combination. The oral contraceptive agent Yasmin-28 contains the progestin drospirenone, which inhibits the MLR and can cause hyperkalemia in susceptible patients. Cyclosporine, tacrolimus, NSAIDs, and cyclooxygenase 2 (COX2) inhibitors cause hyperkalemia by multiple mechanisms, but share the ability to cause hyporeninemic hypoaldosteronism. Notably, most drugs that affect the renin-angiotensin-aldosterone axis also block the local adrenal response to hyperkalemia, thus attenuating the *direct* stimulation of aldosterone release by increased plasma K^+ concentration.

Inhibition of apical ENaC activity in the distal nephron by amiloride and other K^+-sparing diuretics results in hyperkalemia, often with a voltage-dependent hyperchloremic acidosis and/or hypovolemic hyponatremia. Amiloride is structurally similar to the antibiotics TMP and pentamidine, which also block ENaC; risk factors for TMP-associated hyperkalemia include the administered dose, renal insufficiency, and hyporeninemic hypoaldosteronism. Indirect inhibition of ENaC at the plasma membrane is also a cause of drug-associated hyperkalemia; nafamostat, a protease inhibitor used in some countries for anticoagulation and for the management of pancreatitis, inhibits aldosterone-induced renal proteases that activate ENaC by proteolytic cleavage.

Clinical Features

Hyperkalemia is a medical emergency due to its effects on the heart. Cardiac arrhythmias associated with hyperkalemia include sinus bradycardia, sinus arrest, slow idioventricular rhythms, ventricular tachycardia, ventricular fibrillation, and asystole. Mild increases in extracellular K^+ affect the repolarization phase of the cardiac action potential, resulting in changes in T-wave morphology; further increase in plasma K^+ concentration depresses intracardiac conduction, with progressive prolongation of the PR and QRS intervals. Severe hyperkalemia results in loss of the P wave and a progressive widening of the QRS complex; development of a sine-wave sinoventricular rhythm suggests impending ventricular fibrillation or asystole. Hyperkalemia can also cause a type I Brugada pattern in the electrocardiogram (ECG), with a pseudo–right bundle branch block and persistent coved ST segment elevation in at least two precordial leads. This hyperkalemic Brugada's sign occurs in critically ill patients with severe hyperkalemia and can be differentiated from genetic Brugada's syndrome by an absence of P waves, marked QRS widening, and an abnormal QRS axis. Classically, the electrocardiographic manifestations in hyperkalemia progress from tall peaked T waves (5.5–6.5 m*M*), to a loss of P waves (6.5–7.5 m*M*) to a widened QRS complex (7.0–8.0 m*M*), and, ultimately, a to a sine wave pattern (>8.0 m*M*). However, these changes are notoriously insensitive, particularly in patients with chronic kidney disease or ESRD.

Hyperkalemia from a variety of causes can also present with ascending paralysis, denoted *secondary hyperkalemic paralysis* to differentiate it from familial hyperkalemic periodic paralysis (HYPP). The presentation may include diaphragmatic paralysis and respiratory failure. Patients with familial HYPP develop myopathic weakness during hyperkalemia induced by increased K^+ intake or rest after heavy exercise. Depolarization of skeletal muscle by hyperkalemia unmasks an inactivation defect in skeletal Na^+ channel; autosomal dominant mutations in the *SCN4A* gene encoding this channel are the predominant cause.

Within the kidney, hyperkalemia has negative effects on the ability to excrete an acid load, such that hyperkalemia per se can contribute to metabolic acidosis. This defect appears to be due in part to competition between K^+ and NH_4^+ for reabsorption by the TALH and subsequent countercurrent multiplication, ultimately reducing the medullary gradient for NH_3/NH_4 excretion by the distal nephron. Regardless of the underlying mechanism, restoration of normokalemia can, in many instances, correct hyperkalemic metabolic acidosis.

Diagnostic Approach

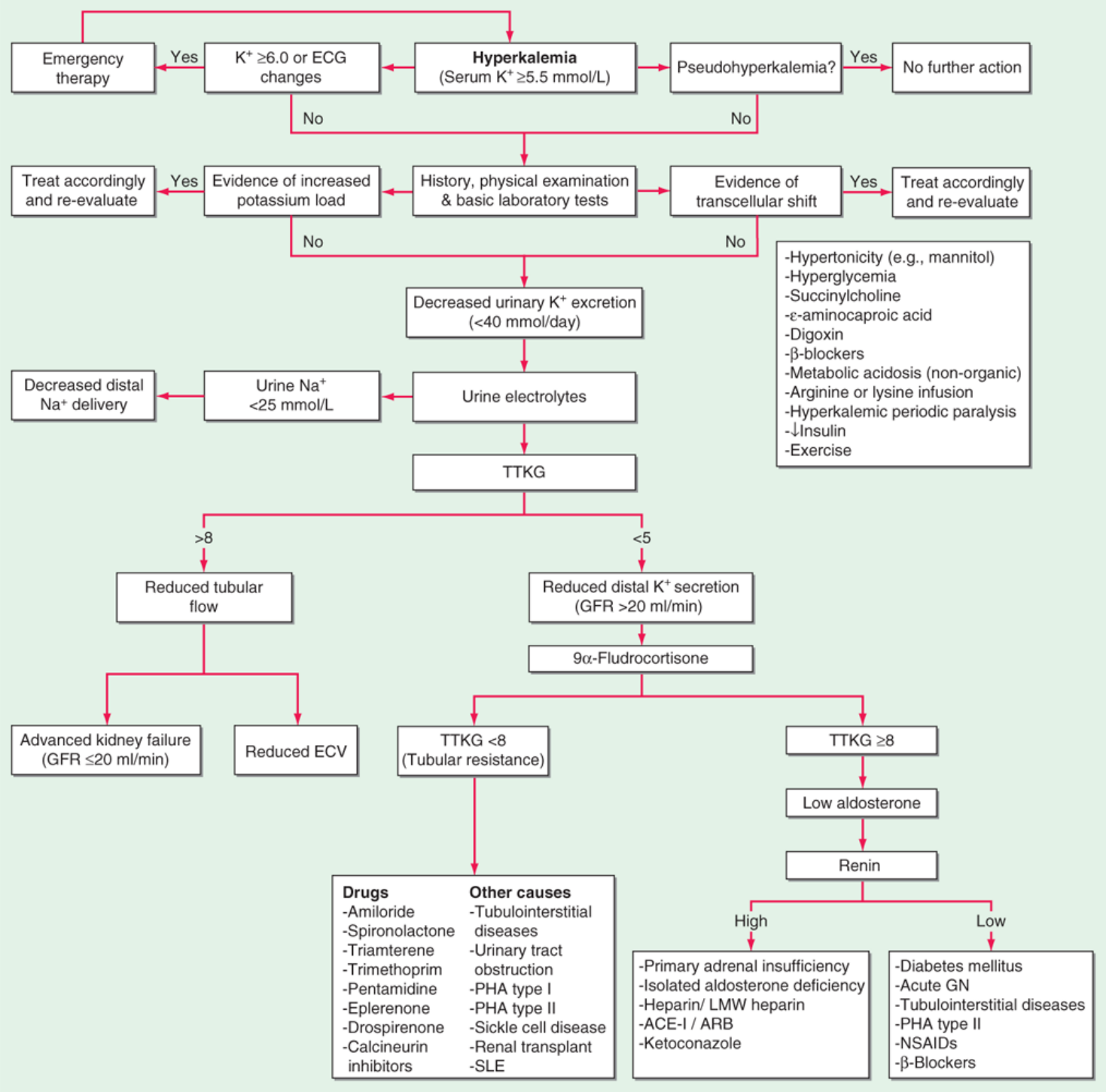
The first priority in the management of hyperkalemia is to assess the need for emergency treatment, followed by a comprehensive workup to determine the cause **(Fig. 49-8)**. History and physical examination should focus on medications, diet and dietary supplements, risk factors for kidney failure, reduction in urine output, blood pressure, and volume status. Initial laboratory tests should include electrolytes, BUN, creatinine, serum osmolality, Mg^{2+} and Ca^{2+}, a complete blood count, and urinary pH, osmolality, creatinine, and electrolytes. A urine Na^+ concentration of <20 m*M* indicates that

distal Na^+ delivery is a limiting factor in K^+ excretion; volume repletion with 0.9% saline or treatment with furosemide may be effective in reducing plasma K^+ concentration. Serum and urine osmolality are required for calculation of the transtubular K^+ gradient (TTKG) (Fig. 49-8). The expected values of the TTKG are largely based on historical data, and are <3 in the presence of hypokalemia and >7–8 in the presence of hyperkalemia.

$$\text{TTKG} = \frac{[K^+]_{\text{urine}} \times \text{Osm}_{\text{serum}}}{[K^+]_{\text{serum}} \times \text{Osm}_{\text{urine}}}$$

FIGURE 49-8

The diagnostic approach to hyperkalemia. See text for details. ACE-I, angiotensin-converting enzyme inhibitor; ARB, angiotensin II receptor blocker; CCD, cortical collecting duct; ECG, electrocardiogram; ECV, effective circulatory volume; GFR, glomerular filtration rate; GN, glomerulonephritis; HIV, human immunodeficiency virus; LMW heparin, low-molecular-weight heparin; NSAIDs, nonsteroidal anti-inflammatory drugs; PHA, pseudohypoaldosteronism; SLE, systemic lupus erythematosus; TTKG, transtubular potassium gradient. *(Used with permission from DB Mount, K Zandi-Nejad K: Disorders of potassium balance, in Brenner and Rector's The Kidney, 8th ed, BM Brenner [ed]. Philadelphia, W.B. Saunders & Company, 2008, pp 547–587.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Hyperkalemia

Electrocardiographic manifestations of hyperkalemia should be considered a medical emergency and treated urgently. However, patients with significant hyperkalemia (plasma K^+ concentration ≥6.5 m*M*) in the absence of ECG changes should also be aggressively managed, given the limitations of ECG changes as a predictor of cardiac toxicity. Urgent management of hyperkalemia includes admission to the hospital, continuous cardiac monitoring, and immediate treatment. The treatment of hyperkalemia is divided into three stages:

1. *Immediate antagonism of the cardiac effects of hyperkalemia.* Intravenous calcium serves to protect the heart, whereas other measures are taken to correct hyperkalemia. Calcium raises the action potential threshold and reduces excitability, without changing the resting membrane potential. By restoring the difference between resting and threshold potentials, calcium reverses the depolarization blockade due to hyperkalemia. The recommended dose is 10 mL of 10% calcium gluconate (3–4 mL of calcium chloride), infused intravenously over 2–3 min with cardiac monitoring. The effect of the infusion starts in 1–3 min and lasts 30–60 min; the dose should be repeated if there is no change in ECG findings or if they recur after initial improvement. Hypercalcemia potentiates the cardiac toxicity of digoxin; hence, intravenous calcium should be used with extreme caution in patients taking this medication; if judged necessary, 10 mL of 10% calcium gluconate can be added to 100 mL of 5% dextrose in water and infused over 20–30 min to avoid acute hypercalcemia.

2. *Rapid reduction in plasma K^+ concentration by redistribution into cells.* Insulin lowers plasma K^+ concentration by shifting K^+ into cells. The recommended dose is 10 units of intravenous regular insulin followed immediately by 50 mL of 50% dextrose ($D_{50}W$, 25 g of glucose total); the effect begins in 10–20 min, peaks at 30–60 min, and lasts for 4–6 h. Bolus $D_{50}W$ without insulin is *never* appropriate, given the risk of acutely worsening hyperkalemia due to the osmotic effect of hypertonic glucose. Hypoglycemia is common with insulin plus glucose; hence, this should be followed by an infusion of 10% dextrose at 50–75 mL/h, with close monitoring of plasma glucose concentration. In hyperkalemic patients with glucose concentrations of ≥200–250 mg/dL, insulin should be administered *without* glucose, again with close monitoring of glucose concentrations.

   β_2-agonists, most commonly albuterol, are effective but underused agents for the acute management of hyperkalemia. Albuterol and insulin with glucose have an additive effect on plasma K^+ concentration; however, ~20% of patients with ESRD are resistant to the effect of β_2-agonists; hence, these drugs should not be used without insulin. The recommended dose for inhaled albuterol is 10–20 mg of nebulized albuterol in 4 mL of normal saline, inhaled over 10 min; the effect starts at about 30 min, reaches its peak at about 90 min, and lasts for 2–6 h. Hyperglycemia is a side effect, along with tachycardia. β_2-Agonists should be used with caution in hyperkalemic patients with known cardiac disease.

   Intravenous bicarbonate has no role in the acute treatment of hyperkalemia, but may slowly attenuate hyperkalemia with sustained administration over several hours. It should not be given repeatedly as a hypertonic intravenous bolus of undiluted ampules, given the risk of associated hypernatremia, but should instead be infused in an isotonic or hypotonic fluid (e.g., 150 mEqu in 1 L of D_5W). In patients with metabolic acidosis, a delayed drop in plasma K^+ concentration can be seen after 4–6 h of isotonic bicarbonate infusion.

3. *Removal of potassium.* This is typically accomplished using cation exchange resins, diuretics, and/or dialysis. The cation exchange resin sodium polystyrene sulfonate (SPS) exchanges Na^+ for K^+ in the gastrointestinal tract and increases the fecal excretion of K^+; alternative calcium-based resins, when available, may be more appropriate in patients with an increased ECFV. The recommended dose of SPS is 15–30 g of powder, almost always given in a premade suspension with 33% sorbitol. The effect of SPS on plasma K^+ concentration is slow; the full effect may take up to 24 h and usually requires repeated doses every 4–6 h. Intestinal necrosis, typically of the colon or ileum, is a rare but usually fatal complication of SPS. Intestinal necrosis is more common in patients administered SPS via enema and/or in patients with reduced intestinal motility (e.g., in the postoperative state or after treatment with opioids). The coadministration of SPS with sorbitol appears to increase the risk of intestinal necrosis; however, this complication can also occur with SPS alone. The low but real risk of intestinal necrosis with SPS, which can sometimes be the only available or appropriate therapy for the removal of potassium, must be weighed against the delayed onset of efficacy. Whenever possible, alternative therapies for the acute management of hyperkalemia (i.e., aggressive redistributive therapy, isotonic bicarbonate infusion, diuretics, and/or hemodialysis) should be used instead of SPS.

   Novel intestinal potassium binders have recently become available for the management of hyperkalemia. These agents appear to lack the intestinal toxicity of SPS. Patiromer is a nonabsorbed polymer provided as a powder for suspension, which binds K^+ in exchange for Ca^{2+}. In healthy adults, patiromer causes a decrease in urinary potassium, magnesium, and sodium excretion, suggesting the binding of the polymer to these cations in the intestine; notably, a side-effect of the medication is hypomagnesemia. ZS-9 is an inorganic, nonabsorbable crystalline compound that exchanges both Na^+ and H^+ ions in exchange for K^+ and NH_4^+ in the intestine. These agents promise to revolutionize the management of both chronic and acute hyperkalemia. In particular, the availability of safe, well-tolerated potassium binders is expected to allow for greater intensity of RAAS inhibition in both renal and cardiac disease.

   Therapy with intravenous saline may be beneficial in hypovolemic patients with oliguria and decreased distal delivery of Na^+, with the associated reductions in renal K^+ excretion. Loop and thiazide diuretics can be used to reduce plasma K^+ concentration in volume-replete or hypervolemic patients with sufficient renal function for a diuretic response; this may need to be combined with intravenous saline or isotonic bicarbonate to achieve or maintain euvolemia.

   Hemodialysis is the most effective and reliable method to reduce plasma K^+ concentration; peritoneal dialysis is considerably less effective. Patients with acute kidney injury require temporary, urgent venous access for hemodialysis, with the attendant risks; in contrast, patients with ESRD or advanced chronic kidney disease may have a preexisting venous access. The amount of K^+ removed during hemodialysis depends on the relative

distribution of K^+ between ICF and ECF (potentially affected by prior therapy for hyperkalemia), the type and surface area of the dialyzer used, dialysate and blood flow rates, dialysate flow rate, dialysis duration, and the plasma-to-dialysate K^+ gradient.

FURTHER READING

Choi M et al: K^+ channel mutations in adrenal aldosterone-producing adenomas and hereditary hypertension. Science 331:768, 2011. [PubMed: 21311022]

Mount DB, Zandi-Nejad K: Disorders of potassium balance, in *Brenner and Rector's The Kidney*, 10th ed, K Skorecki et al (eds). Philadelphia, W.B. Saunders & Company, 2016, pp 559–600.

Packham DK et al: Sodium zirconium cyclosilicate in hyperkalemia. N Engl J Med 372:222, 2015. [PubMed: 25415807]

Perianayagam A et al: DDAVP is effective in preventing and reversing inadvertent overcorrection of hyponatremia. Clin J Am Soc Nephrol 3:331, 2008. [PubMed: 18235152]

Schrier RW: Decreased effective blood volume in edematous disorders: what does this mean? J Am Soc Nephrol 18:2028, 2007. [PubMed: 17568020]

Sood L et al: Hypertonic saline and desmopressin: a simple strategy for safe correction of severe hyponatremia. Am J Kidney Dis 61:571, 2013. [PubMed: 23266328]

Soupart A et al: Efficacy and tolerance of urea compared with vaptans for long-term treatment of patients with SIADH. Clin J Am Soc Nephrol 7:742, 2012. [PubMed: 22403276]

Weir MR et al: Patiromer in patients with kidney disease and hyperkalemia receiving RAAS inhibitors. N Engl J Med 372:211, 2015. [PubMed: 25415805]

Chapter 50: Hypercalcemia and Hypocalcemia

Sundeep Khosla

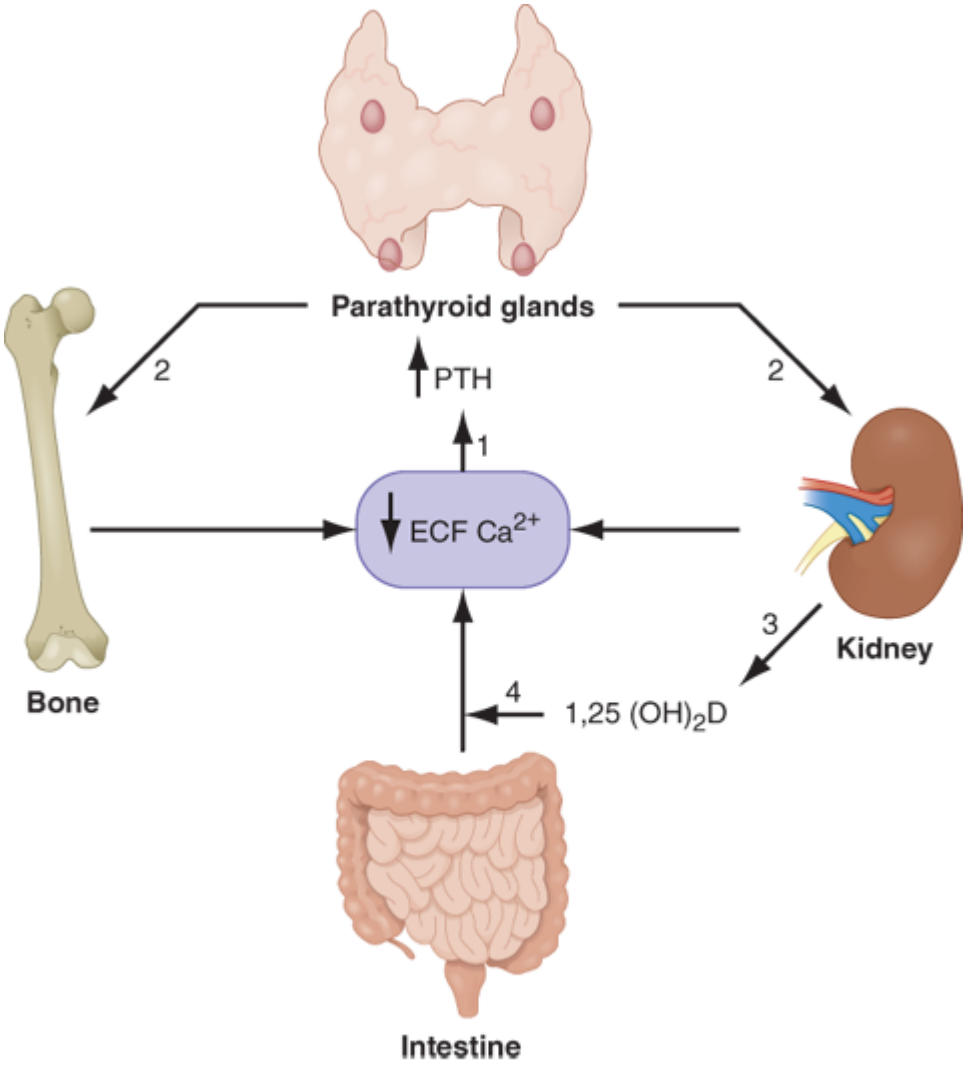
FIGURE 50-1

INTRODUCTION

The calcium ion plays a critical role in normal cellular function and signaling, regulating diverse physiologic processes such as neuromuscular signaling, cardiac contractility, hormone secretion, and blood coagulation. Thus, extracellular calcium concentrations are maintained within an exquisitely narrow range through a series of feedback mechanisms that involve parathyroid hormone (PTH) and the active vitamin D metabolite 1,25-dihydroxyvitamin D [1,25$(OH)_2$D]. These feedback mechanisms are orchestrated by integrating signals between the parathyroid glands, kidney, intestine, and bone **(Fig. 50-1; Chap. 402)**. Disorders of serum calcium concentration are relatively common and often serve as a harbinger of underlying disease. This chapter provides a brief summary of the approach to patients with altered serum calcium levels. **See Chap. 403 for a detailed discussion of this topic.**

FIGURE 50-1

Feedback mechanisms maintaining extracellular calcium concentrations within a narrow, physiologic range (8.9–10.1 mg/dL [2.2–2.5 m*M*]). A decrease in extracellular (ECF) calcium (Ca^{2+}) triggers an increase in parathyroid hormone (PTH) secretion (1) via the calcium sensor receptor on parathyroid cells. PTH, in turn, results in increased tubular reabsorption of calcium by the kidney (2) and resorption of calcium from bone (2) and also stimulates renal 1,25$(OH)_2$D production (3). 1,25$(OH)_2$D, in turn, acts principally on the intestine to increase calcium absorption (4). Collectively, these homeostatic mechanisms serve to restore serum calcium levels to normal.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HYPERCALCEMIA

ETIOLOGY

The causes of hypercalcemia can be understood and classified based on derangements in the normal feedback mechanisms that regulate serum calcium (Table 50-1). Excess PTH production, which is not appropriately suppressed by increased serum calcium concentrations, occurs in primary neoplastic disorders of the parathyroid glands (parathyroid adenomas; hyperplasia; or, rarely, carcinoma) that are associated with increased parathyroid cell mass and impaired feedback inhibition by calcium. Inappropriate PTH secretion for the ambient level of serum calcium also occurs in familial hypocalciuric hypercalcemia (FHH), which is an autosomal dominant syndrome most commonly involving inactivating mutations in the calcium sensor receptor (*CaSR*; FHH type 1), with rare families having mutations in the $G\alpha_{11}$ protein (*GNA11*; FHH type 2) or the adaptor-related protein complex 2, σ-2 subunit (*AP2S1;* FHH type 3); all of these mutations impair extracellular calcium sensing by the parathyroid glands and the kidneys, leading to inappropriate PTH secretion and increased renal tubular calcium reabsorption. Although PTH secretion by tumors is extremely

rare, many solid tumors produce PTH-related peptide (PTHrP), which shares homology with PTH in the first 13 amino acids and binds the PTH receptor, thus mimicking effects of PTH on bone and the kidney. In PTHrP-mediated hypercalcemia of malignancy, PTH levels are suppressed by the high serum calcium levels. Hypercalcemia associated with granulomatous disease (e.g., sarcoidosis) or lymphomas is caused by enhanced conversion of 25(OH)D to the potent $1,25(OH)_2D$. In these disorders, $1,25(OH)_2D$ enhances intestinal calcium absorption, resulting in hypercalcemia and suppressed PTH. Disorders that directly increase calcium mobilization from bone, such as hyperthyroidism or osteolytic metastases, also lead to hypercalcemia with suppressed PTH secretion as does exogenous calcium overload, as in milk-alkali syndrome, or total parenteral nutrition with excessive calcium supplementation.

TABLE 50-1

Causes of Hypercalcemia

Excessive PTH production
Primary hyperparathyroidism (adenoma, hyperplasia, rarely carcinoma)
Tertiary hyperparathyroidism (long-term stimulation of PTH secretion in renal insufficiency)
Ectopic PTH secretion (very rare)
FHH
Alterations in CaSR function (lithium therapy)
Hypercalcemia of malignancy
Overproduction of PTHrP (many solid tumors)
Lytic skeletal metastases (breast, myeloma)
Excessive 1,25$(OH)_2$D production
Granulomatous diseases (sarcoidosis, tuberculosis, silicosis)
Lymphomas
Vitamin D intoxication
Primary increase in bone resorption
Hyperthyroidism
Immobilization
Excessive calcium intake
Milk-alkali syndrome
Total parenteral nutrition
Other causes

Endocrine disorders (adrenal insufficiency, pheochromocytoma, VIPoma)
Medications (thiazides, vitamin A, antiestrogens)

Abbreviations: CaSR, calcium sensor receptor; FHH, familial hypocalciuric hypercalcemia; PTH, parathyroid hormone; PTHrP, PTH-related peptide.

CLINICAL MANIFESTATIONS

Mild hypercalcemia (up to 11–11.5 mg/dL) is usually asymptomatic and recognized only on routine calcium measurements. Some patients may complain of vague neuropsychiatric symptoms, including trouble concentrating, personality changes, or depression. Other presenting symptoms may include peptic ulcer disease or nephrolithiasis, and fracture risk may be increased. More severe hypercalcemia (>12–13 mg/dL), particularly if it develops acutely, may result in lethargy, stupor, or coma, as well as gastrointestinal symptoms (nausea, anorexia, constipation, or pancreatitis). Hypercalcemia decreases renal concentrating ability, which may cause polyuria and polydipsia. With long-standing hyperparathyroidism, patients may present with bone pain or pathologic fractures. Finally, hypercalcemia can result in significant electrocardiographic changes, including bradycardia, AV block, and short QT interval; changes in serum calcium can be monitored by following the QT interval.

DIAGNOSTIC APPROACH

The first step in the diagnostic evaluation of hyper- or hypocalcemia is to ensure that the alteration in serum calcium levels is not due to abnormal albumin concentrations. About 50% of total calcium is ionized, and the rest is bound principally to albumin. Although direct measurements of ionized calcium are possible, they are easily influenced by collection methods and other artifacts; thus, it is generally preferable to measure total calcium and albumin to “correct” the serum calcium. When serum albumin concentrations are reduced, a corrected calcium concentration is calculated by adding 0.2 m*M* (0.8 mg/dL) to the total calcium level for every decrement in serum albumin of 1.0 g/dL below the reference value of 4.1 g/dL for albumin, and, conversely, for elevations in serum albumin.

A detailed history may provide important clues regarding the etiology of the hypercalcemia (Table 50-1). Chronic hypercalcemia is most commonly caused by primary hyperparathyroidism, as opposed to the second most common etiology of hypercalcemia, an underlying malignancy. The history should include medication use, previous neck surgery, and systemic symptoms suggestive of sarcoidosis or lymphoma.

Once true hypercalcemia is established, the second most important laboratory test in the diagnostic evaluation is a PTH level using a two-site assay for the intact hormone. Increases in PTH are often accompanied by hypophosphatemia. In addition, serum creatinine should be measured to assess renal function; hypercalcemia may impair renal function, and renal clearance of PTH may be altered depending on

the fragments detected by the assay. If the PTH level is increased (or “inappropriately normal”) in the setting of elevated calcium and low phosphorus, the diagnosis is almost always primary hyperparathyroidism. Because individuals with FHH may also present with mildly elevated PTH levels and hypercalcemia, this diagnosis should be considered and excluded because parathyroid surgery is ineffective in this condition. A calcium/creatinine clearance ratio (calculated as urine calcium/serum calcium divided by urine creatinine/serum creatinine) of <0.01 is suggestive of FHH, particularly when there is a family history of mild, asymptomatic hypercalcemia. In addition, sequence analysis of the *CASR* gene is now commonly performed for the definitive diagnosis of FHH, although as noted above, in rare families FHH may be caused by mutations in the *GNA11* or *AP2S1* genes. Ectopic PTH secretion is extremely rare.

A suppressed PTH level in the face of hypercalcemia is consistent with non-parathyroid-mediated hypercalcemia, most often due to underlying malignancy. Although a tumor that causes hypercalcemia is generally overt, a PTHrP level may be needed to establish the diagnosis of hypercalcemia of malignancy. Serum 1,25$(OH)_2$D levels are increased in granulomatous disorders, and clinical evaluation in combination with laboratory testing will generally provide a diagnosis for the various disorders listed in Table 50-1.

TREATMENT

TREATMENT

Hypercalcemia

Mild, asymptomatic hypercalcemia does not require immediate therapy, and management should be dictated by the underlying diagnosis. By contrast, significant, symptomatic hypercalcemia usually requires therapeutic intervention independent of the etiology of hypercalcemia. Initial therapy of significant hypercalcemia begins with volume expansion because hypercalcemia invariably leads to dehydration; 4–6 L of intravenous saline may be required over the first 24 h, keeping in mind that underlying comorbidities (e.g., congestive heart failure) may require the use of loop diuretics to enhance sodium and calcium excretion. However, loop diuretics should not be initiated until the volume status has been restored to normal. If there is increased calcium mobilization from bone (as in malignancy or severe hyperparathyroidism), drugs that inhibit bone resorption should be considered. Zoledronic acid (e.g., 4 mg intravenously over ~30 min), pamidronate (e.g., 60–90 mg intravenously over 2–4 h), and ibandronate (2 mg intravenously over 2 h) are bisphosphonates that are commonly used for the treatment of hypercalcemia of malignancy in adults. Onset of action is within 1–3 days, with normalization of serum calcium levels occurring in 60–90% of patients. Bisphosphonate infusions may need to be repeated if hypercalcemia relapses. An alternative to the bisphosphonates is gallium nitrate (200 mg/m^2 intravenously daily for 5 days), which is also effective, but has potential nephrotoxicity. More recently, the potent inhibitor of bone resorption, denosumab (120 mg sc on days 1, 8, 15, and 29, and then every 4 weeks), has also been shown to be effective in treating hypercalcemia refractory to bisphosphonates. In rare instances, dialysis may be necessary. Finally, although intravenous phosphate chelates calcium and decreases serum calcium levels, this therapy can be toxic because calcium-phosphate complexes may deposit in tissues and cause extensive organ damage.

In patients with $1,25(OH)_2D$-mediated hypercalcemia, glucocorticoids are the preferred therapy, as they decrease $1,25(OH)_2D$ production. Intravenous hydrocortisone (100–300 mg daily) or oral prednisone (40–60 mg daily) for 3–7 days is used most often. Other drugs, such as ketoconazole, chloroquine, and hydroxychloroquine, may also decrease $1,25(OH)_2D$ production and are used occasionally.

HYPOCALCEMIA

ETIOLOGY

The causes of hypocalcemia can be differentiated according to whether serum PTH levels are low (hypoparathyroidism) or high (secondary hyperparathyroidism). Although there are many potential causes of hypocalcemia, impaired PTH production and impaired vitamin D production are the most common etiologies **(Table 50-2) (Chap. 403)**. Because PTH is the main defense against hypocalcemia, disorders associated with deficient PTH production or secretion may be associated with profound, life-threatening hypocalcemia. In adults, hypoparathyroidism most commonly results from inadvertent damage to all four glands during thyroid or parathyroid gland surgery. Hypoparathyroidism is a cardinal feature of autoimmune endocrinopathies **(Chap. 381)**; rarely, it may be associated with infiltrative diseases such as sarcoidosis. Impaired PTH secretion may be secondary to magnesium deficiency or to activating mutations in the CaSR or in the G proteins that mediate CaSR signaling (autosomal dominant hypocalcemia), which suppress PTH, leading to effects that are opposite to those that occur in FHH.

TABLE 50-2

Causes of Hypocalcemia

Low Parathyroid Hormone Levels (Hypoparathyroidism)
Parathyroid agenesis
Isolated
DiGeorge's syndrome
Parathyroid destruction
Surgical
Radiation
Infiltration by metastases or systemic diseases
Autoimmune
Reduced parathyroid function
Hypomagnesemia
Autosomal dominant hypocalcemia
High Parathyroid Hormone Levels (Secondary Hyperparathyroidism)
Vitamin D deficiency or impaired $1,25(OH)_2D$ production/action
Nutritional vitamin D deficiency (poor intake or absorption)
Renal insufficiency with impaired $1,25(OH)_2D$ production
Vitamin D resistance, including receptor defects
Parathyroid hormone resistance syndromes
PTH receptor mutations

Low Parathyroid Hormone Levels (Hypoparathyroidism)
Pseudohypoparathyroidism (G protein mutations)
Drugs
Calcium chelators
Inhibitors of bone resorption (bisphosphonates, plicamycin)
Altered vitamin D metabolism (phenytoin, ketoconazole)
Miscellaneous causes
Acute pancreatitis
Acute rhabdomyolysis
Hungry bone syndrome after parathyroidectomy
Osteoblastic metastases with marked stimulation of bone formation (prostate cancer)

Abbreviations: CaSR, calcium sensor receptor; PTH, parathyroid hormone.

Vitamin D deficiency, impaired $1,25(OH)_2D$ production (primarily secondary to renal insufficiency), or vitamin D resistance also cause hypocalcemia. However, the degree of hypocalcemia in these disorders is generally not as severe as that seen with hypoparathyroidism because the parathyroids are capable of mounting a compensatory increase in PTH secretion. Hypocalcemia may also occur in conditions associated with severe tissue injury such as burns, rhabdomyolysis, tumor lysis, or pancreatitis. The cause of hypocalcemia in these settings may include a combination of low albumin, hyperphosphatemia, tissue deposition of calcium, and impaired PTH secretion.

CLINICAL MANIFESTATIONS

Patients with hypocalcemia may be asymptomatic if the decreases in serum calcium are relatively mild and chronic, or they may present with life-threatening complications. Moderate to severe hypocalcemia is associated with paresthesias, usually of the fingers, toes, and circumoral regions, and is caused by increased neuromuscular irritability. On physical examination, a Chvostek's sign (twitching of the circumoral muscles in response to gentle tapping of the facial nerve just anterior to the ear) may be elicited, although it is also present in ~10% of normal individuals. Carpal spasm may be induced by inflation of a blood pressure cuff to

20 mmHg above the patient's systolic blood pressure for 3 min (Trousseau's sign). Severe hypocalcemia can induce seizures, carpopedal spasm, bronchospasm, laryngospasm, and prolongation of the QT interval.

DIAGNOSTIC APPROACH

In addition to measuring serum calcium, it is useful to determine albumin, phosphorus, and magnesium levels. As for the evaluation of hypercalcemia, determining the PTH level is central to the evaluation of hypocalcemia. A suppressed (or "inappropriately low") PTH level in the setting of hypocalcemia establishes absent or reduced PTH secretion (hypoparathyroidism) as the cause of the hypocalcemia. Further history will often elicit the underlying cause (i.e., parathyroid agenesis vs. destruction). By contrast, an elevated PTH level (secondary hyperparathyroidism) should direct attention to the vitamin D axis as the cause of the hypocalcemia. Nutritional vitamin D deficiency is best assessed by obtaining serum 25-hydroxyvitamin D levels, which reflect vitamin D stores. In the setting of renal insufficiency or suspected vitamin D resistance, serum 1,25$(OH)_2$D levels are informative.

TREATMENT

TREATMENT

Hypocalcemia

The approach to treatment depends on the severity of the hypocalcemia, the rapidity with which it develops, and the accompanying complications (e.g., seizures, laryngospasm). Acute, symptomatic hypocalcemia is initially managed with calcium gluconate, 10 mL 10% wt/vol (90 mg or 2.2 mmol) intravenously, diluted in 50 mL of 5% dextrose or 0.9% sodium chloride, given intravenously over 5 min. Continuing hypocalcemia often requires a constant intravenous infusion (typically 10 ampules of calcium gluconate or 900 mg of calcium in 1 L of 5% dextrose or 0.9% sodium chloride administered over 24 h). Accompanying hypomagnesemia, if present, should be treated with appropriate magnesium supplementation.

Chronic hypocalcemia due to hypoparathyroidism is treated with calcium supplements (1000–1500 mg/d elemental calcium in divided doses) and either vitamin D_2 or D_3 (25,000–100,000 U daily) or calcitriol [1,25$(OH)_2$D, 0.25–2 μg/d]. Other vitamin D metabolites (dihydrotachysterol, alfacalcidiol) are now used less frequently. Importantly, PTH (1-84) (Natpara) has recently been approved by the FDA for the treatment of refractory hypoparathyroidism, representing an important advance in treatment of these patients. Vitamin D deficiency is best treated using vitamin D supplementation, with the dose depending on the severity of the deficit and the underlying cause. Thus, nutritional vitamin D deficiency generally responds to relatively low doses of vitamin D (50,000 U, 2–3 times per week for several months), whereas vitamin D deficiency due to malabsorption may require much higher doses (100,000 U/d or more). The treatment goal is to bring serum calcium into the low normal range and to avoid hypercalciuria, which may lead to nephrolithiasis.

GLOBAL CONSIDERATIONS

 In countries with more limited access to health care or screening laboratory testing of serum calcium levels, primary hyperparathyroidism often presents in its severe form with skeletal complications (osteitis fibrosa cystica) in contrast to the asymptomatic form that is common in developed countries. In addition, vitamin D deficiency is paradoxically common in some countries despite extensive sunlight (e.g., India) due to avoidance of sun exposure and poor dietary vitamin D intake.

FURTHER READING

Eastell R et al: Diagnosis of asymptomatic primary hyperparathyroidism: Proceedings of the 4th International Workshop. J Clin Endocrinol Metab 99:3570, 2014. [PubMed: 25162666]

Kim ES, Keating GM: Recombinant human parathyroid hormone (1-84): A review in hypoparathyroidism. Drugs 75:1293, 2015. [PubMed: 26177893]

Mayr B et al: Genetics in endocrinology: Gain and loss of function mutations of the calcium-sensing receptor and associated proteins: Current treatment concepts. Eur J Endocrinol 174:R189, 2016. [PubMed: 26646938]

Minisola S et al: The diagnosis and management of hypercalcemia. BMJ 350:h2723, 2015. [PubMed: 26037642]

Thakker RV: The calcium-sensing receptor: And its involvement in parathyroid pathology. Ann Endocrinol 76:81, 2015.

Chapter 51: Acidosis and Alkalosis

Thomas D. DuBose, Jr.

FIGURE 51-1

NORMAL ACID-BASE HOMEOSTASIS

Systemic arterial pH is maintained between 7.35 and 7.45 by extracellular and intracellular chemical buffering together with respiratory and renal regulatory mechanisms. The control of arterial CO_2 tension ($Paco_2$) by the central nervous system (CNS) and respiratory system and the control of plasma bicarbonate by the kidneys stabilize the arterial pH by excretion or retention of acid or alkali. The metabolic and respiratory components that regulate systemic pH are described by the Henderson-Hasselbalch equation:

$$pH = 6.1 + \log \frac{HCO_3^-}{PaCO_2 \times 0.03001}$$

Under most circumstances, CO_2 production and excretion are matched, and the usual steady-state $Paco_2$ is maintained at 40 mmHg. Underexcretion of CO_2 produces hypercapnia, and overexcretion causes hypocapnia. Nevertheless, production and excretion are again matched at a new steady-state $Paco_2$. Therefore, the $Paco_2$ is regulated primarily by neural respiratory factors and is not subject to regulation by the rate of CO_2 production. Hypercapnia is usually the result of hypoventilation rather than of increased CO_2 production. Increases or decreases in $Paco_2$ represent derangements of neural respiratory control or are due to compensatory changes in response to a primary alteration in the plasma [HCO_3^-].

DIAGNOSIS OF GENERAL TYPES OF DISTURBANCES

The most common clinical disturbances are simple acid-base disorders, that is, metabolic acidosis or alkalosis or respiratory acidosis or alkalosis.

SIMPLE ACID-BASE DISORDERS

Primary respiratory disturbances (primary changes in $Paco_2$) invoke compensatory metabolic responses (secondary changes in $[HCO_3^-]$), and primary metabolic disturbances elicit predictable compensatory respiratory responses (secondary changes in $Paco_2$). Physiologic compensation can be predicted from the relationships displayed in **Table 51-1**. In general, with one exception, compensatory responses return the pH toward, but not to, the normal value. Chronic respiratory alkalosis when prolonged is an exception to this rule and may return the pH to a normal value. Metabolic acidosis due to an increase in endogenous acid production (e.g., ketoacidosis) lowers extracellular fluid $[HCO_3^-]$ and decreases extracellular pH. This stimulates the medullary chemoreceptors to increase ventilation and to return the ratio of $[HCO_3^-]$ to $Paco_2$, and thus pH, toward, but not to, normal. The degree of respiratory compensation expected in a metabolic acidosis can be predicted from the relationship: $Paco_2 = (1.5 \times [HCO_3^-]) + 8 \pm 2$. Thus, a patient with metabolic acidosis and $[HCO_3^-]$ of 12 mmol/L would be expected to have a $Paco_2$ of ~26 mmHg. Values for $Paco_2$ <24 or >28 mmHg define a mixed disturbance (metabolic acidosis and respiratory alkalosis or metabolic acidosis and respiratory acidosis, respectively). Compensatory responses for primary metabolic disorders move the $Paco_2$ in the same direction as the change in $[HCO_3^-]$, whereas, conversely, compensation for primary respiratory disorders moves the $[HCO_3^-]$ in the same direction as the primary change in $Paco_2$ (Table 51-1). Therefore, changes in $Paco_2$ and $[HCO_3^-]$ in **opposite directions** (i.e., $Paco_2$ or $[HCO_3^-]$ is increased, whereas the other value is decreased) indicate a **mixed acid-base disturbance**. Another way to judge the appropriateness of the response in $[HCO_3^-]$ or $Paco_2$ is to use an acid-base nomogram **(Fig. 51-1)**. While the shaded areas of the nomogram show the 95% confidence limits for physiologic compensation in simple disturbances, finding acid-base values within the shaded area does not necessarily rule out a mixed disturbance. Imposition of one disorder over another may result in values lying within the area of a third. Thus, the nomogram, while convenient, is not a substitute for the equations in Table 51-1.

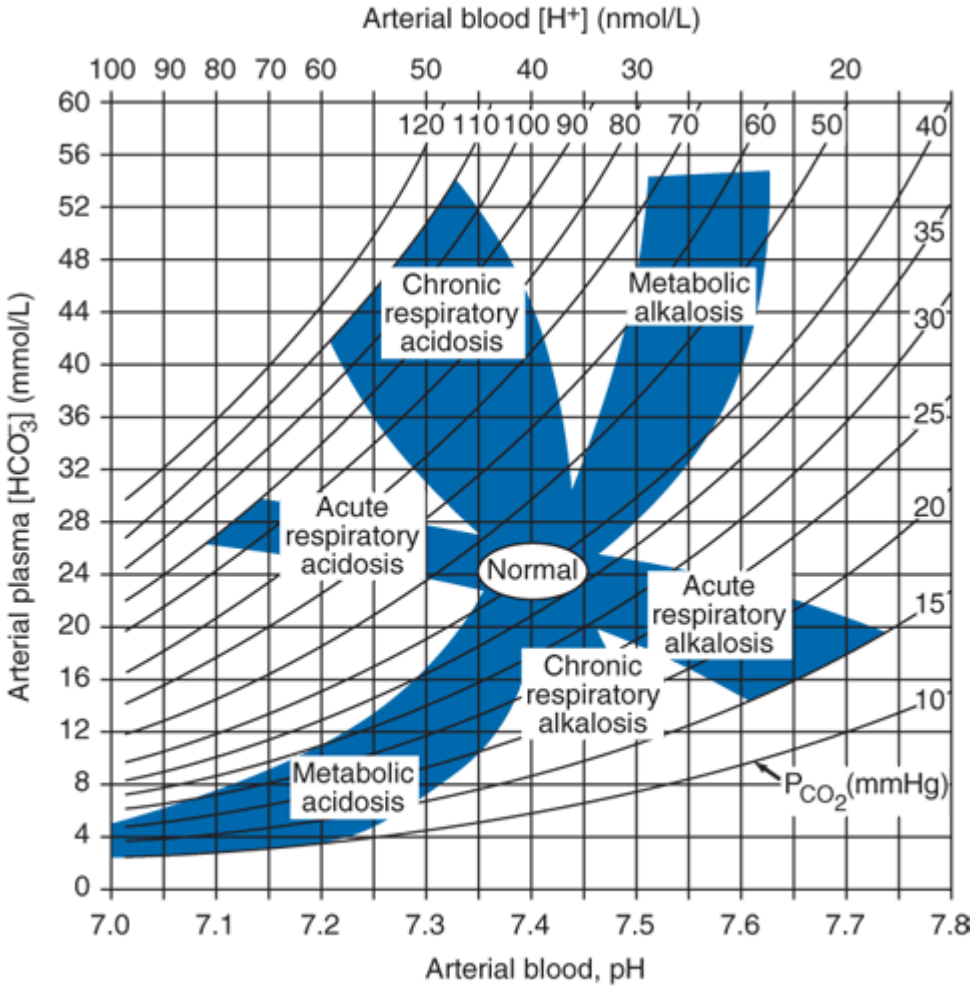
TABLE 51-1

Prediction of Compensatory Responses to Simple Acid-Base Disturbances and Pattern of Changes

Disorder	Prediction of Compensation	Range of Values		
		pH	HCO_3^-	$Paco_2$
Metabolic acidosis	$Paco_2 = (1.5 \times HCO_3^-) + 8 \pm 2$ *or* $Paco_2$ will ↓ 1.25 mmHg per mmol/L ↓ in $[HCO_3^-]$ *or* $Paco_2 = [HCO_3^-] + 15$	Low	Low	Low
Metabolic alkalosis	$Paco_2$ will ↑ 0.75 mmHg per mmol/L ↑ in $[HCO_3^-]$ *or* $Paco_2$ will ↑ 6 mmHg per 10 mmol/L ↑ in $[HCO_3^-]$ *or* $Paco_2 = [HCO_3^-] + 15$	High	High	High
Respiratory alkalosis		High	Low	Low
Acute	$[HCO_3^-]$ will ↓ 0.2 mmol/L per mmHg ↓ in $Paco_2$			
Chronic	$[HCO_3^-]$ will ↓ 0.4 mmol/L per mmHg ↓ in $Paco_2$			
Respiratory acidosis		Low	High	High
Acute	$[HCO_3^-]$ will ↑ 0.1 mmol/L per mmHg ↑ in $Paco_2$			
Chronic	$[HCO_3^-]$ will ↑ 0.4 mmol/L per mmHg ↑ in $Paco_2$			

FIGURE 51-1

Acid-base nomogram. Shown are the 90% confidence limits (range of values) of the normal respiratory and metabolic compensations for primary acid-base disturbances. *(From TD DuBose Jr: Acid-Base Disorders, in Brenner and Rector's The Kidney, 10th ed, K Skorecki, GM Chertow, PA Marsden, MW Taal, and Alan SL Yu [eds]. Philadelphia, Saunders, 2016, p. 522; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MIXED ACID-BASE DISORDERS

Mixed acid-base disorders—defined as independently coexisting disorders, not merely compensatory responses—are often seen in patients in critical care units and can lead to dangerous extremes of pH (Table 51-2). The diagnosis of mixed acid-base disorders requires consideration of the anion gap (AG), and requires the presence of or correction to a normal serum albumin of 4.5 g/dL. A patient with diabetic ketoacidosis (metabolic acidosis) may develop an independent respiratory problem (e.g., pneumonia) leading to a superimposed respiratory acidosis or alkalosis. Patients with underlying pulmonary disease (e.g., chronic obstructive pulmonary disease) may not respond to metabolic acidosis with an appropriate ventilatory response because of insufficient respiratory reserve. Such imposition of respiratory acidosis on metabolic acidosis can lead to severe acidemia. When metabolic acidosis and metabolic alkalosis coexist in the same patient, the pH may be in the normal range. In this circumstance, it is the presence of an elevated AG (see below) that denotes the presence of a metabolic acidosis. Assuming a normal value for the AG of 10 mmol/L, an incongruity in the ΔAG (prevailing minus normal AG) and the ΔHCO_3^- (normal value of 25 mmol/L minus abnormal HCO_3^- in the patient) indicates the presence of a mixed high-gap acidosis—metabolic alkalosis (see example below). A diabetic patient with ketoacidosis may have renal dysfunction resulting in

simultaneous metabolic acidosis. Patients who have ingested an overdose of drug combinations such as sedatives and salicylates may have mixed disturbances as a result of the acid-base response to the individual drugs (metabolic acidosis mixed with respiratory acidosis or respiratory alkalosis, respectively). Triple acid-base disturbances are more complex. For example, patients with metabolic acidosis due to alcoholic ketoacidosis may develop metabolic alkalosis due to vomiting and superimposed respiratory alkalosis due to the hyperventilation of hepatic dysfunction or alcohol withdrawal.

TABLE 51-2

Examples of Mixed Acid-Base Disorders

Mixed Metabolic and Respiratory
Metabolic acidosis—respiratory alkalosis
Key: High- or normal-AG metabolic acidosis; prevailing $Paco_2$ *below* predicted value (Table 51-1)
Example: Na^+, 140; K^+, 4.0; Cl^-, 106; HCO_3^-, 14; AG, 20; $Paco_2$, 24; pH, 7.39 (lactic acidosis, sepsis in ICU)
Metabolic acidosis—respiratory acidosis
Key: High- or normal-AG metabolic acidosis; prevailing $Paco_2$ *above* predicted value (Table 51-1)
Example: Na^+, 140; K^+, 4.0; Cl^-, 102; HCO_3^-, 18; AG, 20; $Paco_2$, 38; pH, 7.30 (severe pneumonia, pulmonary edema)
Metabolic alkalosis—respiratory alkalosis
Key: $Paco_2$ does not increase as predicted; pH higher than expected
Example: Na^+, 140; K^+, 4.0; Cl^-, 91; HCO_3^-, 33; AG, 16; $Paco_2$, 38; pH, 7.55 (liver disease and diuretics)
Metabolic alkalosis—respiratory acidosis
Key: $Paco_2$ higher than predicted; pH normal
Example: Na^+, 140; K^+, 3.5; Cl^-, 88; HCO_3^-, 42; AG, 10; $Paco_2$, 67; pH, 7.42 (COPD on diuretics)
Mixed Metabolic Disorders
Metabolic acidosis—metabolic alkalosis
Key: Only detectable with high-AG acidosis; $\Delta AG >> \Delta HCO_3^-$
Example: Na^+, 140; K^+, 3.0; Cl^-, 95; HCO_3^-, 25; AG, 20; $Paco_2$, 40; pH, 7.42 (uremia with vomiting)
Metabolic acidosis—metabolic acidosis

Key: Mixed high-AG—normal-AG acidosis; ΔHCO_3^- accounted for by combined change in ΔAG and ΔCl^-
Example: Na^+, 135; K^+, 3.0; Cl^-, 110; HCO_3^-, 10; AG, 15; $Paco_2$, 25; pH, 7.20 (diarrhea and lactic acidosis, toluene toxicity, treatment of diabetic ketoacidosis)

Abbreviations: AG, anion gap; COPD, chronic obstructive pulmonary disease; ICU, intensive care unit.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Acid-Base Disorders

A stepwise approach to the diagnosis of acid-base disorders follows **(Table 51-3)**. Blood for electrolytes and arterial blood gases should be drawn simultaneously prior to therapy. An increase in $[HCO_3^-]$ occurs with either metabolic alkalosis or respiratory acidosis. Conversely, a decrease in $[HCO_3^-]$ occurs with either metabolic acidosis or respiratory alkalosis. In the determination of arterial blood gases by the clinical laboratory, both pH and $Paco_2$ are measured, and the $[HCO_3^-]$ is calculated from the Henderson-Hasselbalch equation. This calculated value should be compared with the measured $[HCO_3^-]$ (total CO_2) on the electrolyte panel. These two values should agree within 2 mmol/L. If they do not, the values may not have been drawn simultaneously, or a laboratory error may be present. After verifying the blood acid-base values, the precise acid-base disorder can then be identified.

CALCULATE THE ANION GAP

All evaluations of acid-base disorders should include a simple calculation of the AG. The AG is calculated as follows: $AG = Na^+ - (Cl^- + HCO_3^-)$. In the United States, the value for plasma $[K^+]$ is typically omitted from the calculation of the AG. The "normal" value for the AG reported by clinical laboratories has declined with improved methodology for measuring plasma electrolytes, and ranges from 6 to 12 mmol/L, with an average of ~10 mmol/L. The clinician is encouraged to be aware of the normal value for the AG in their clinical chemistry laboratory. The unmeasured anions normally present in plasma include anionic proteins (e.g., albumin), phosphate, sulfate, and organic anions. When acid anions, such as acetoacetate and lactate, accumulate in extracellular fluid, the AG increases, causing a **high-AG acidosis**. An increase in the AG is most often due to an increase in unmeasured anions and, less commonly, may be due to a decrease in unmeasured cations (calcium, magnesium, potassium). In addition, the AG may increase with an increase in anionic albumin. A decrease in the AG can be due to (1) an increase in unmeasured cations; (2) the addition to the blood of abnormal cations, such as lithium (lithium intoxication) or cationic immunoglobulins (plasma cell dyscrasias); (3) a reduction in the plasma anion albumin concentration (nephrotic syndrome, liver disease or malabsorption); or (4) hyperviscosity and severe hyperlipidemia, which can lead to an

underestimation of sodium and chloride concentrations. Because the normal AG of 10 mmol/L assumes that the serum albumin is normal, if hypoalbuminemia is present, the value for the AG must be corrected. For example, for each g/dL of serum albumin below the normal value (4.5 g/dL), 2.5 mmol/L should be added to the reported (uncorrected) AG. Thus, in a patient with a serum albumin of 2.5 g/dL (2 g/dL below the normal value), and an uncorrected AG of 15, the corrected AG is calculated by adding 5 mmol/L (2.5 × 2 = 5; 5 + 15 = corrected AG of 20 mmol/L). The clinical disorders that cause a high-AG acidosis are displayed in Table 51-3.

A high AG is usually due to accumulation of non–chloride-containing acids that contain inorganic (phosphate, sulfate), organic (ketoacids, lactate, uremic organic anions), exogenous (salicylate or ingested toxins with organic acid production), or unidentified anions. The high AG is significant clinically even if the $[HCO_3^-]$ or pH is normal. Simultaneous metabolic acidosis of the high-AG variety plus either chronic respiratory acidosis or metabolic alkalosis represents such a situation in which $[HCO_3^-]$ may be normal or even high (Table 51-3). In cases of high-AG metabolic acidosis it is valuable to compare the decline in $[HCO_3^-]$ (ΔHCO_3^-: 25 – patient's $[HCO_3^-]$) with the increase in the AG (ΔAG: patient's AG – 10).

Similarly, normal values for $[HCO_3^-]$, $Paco_2$, and pH do not ensure the absence of an acid-base disturbance. For instance, an alcoholic who has been vomiting may develop a metabolic alkalosis with a pH of 7.55, $Paco_2$ of 47 mmHg, $[HCO_3^-]$ of 40 mmol/L, $[Na^+]$ of 135, $[Cl^-]$ of 80, and $[K^+]$ of 2.8. If such a patient were then to develop a superimposed alcoholic ketoacidosis with a β-hydroxybutyrate concentration of 15 mmol/L, arterial pH would fall to 7.40, the $[HCO_3^-]$ to 25 mmol/L, and the $Paco_2$ to 40 mmHg. Although these blood gases are normal, the AG is elevated at 30 mmol/L, indicating a mixed metabolic alkalosis and metabolic acidosis. A mixture of high-gap acidosis and metabolic alkalosis is recognized easily by comparing the differences (Δ values) in the normal to prevailing patient values. In this example, the ΔHCO_3^- is 0 (25 − 25 mmol/L), but the ΔAG is 20 (30 – 10 mmol/L). Therefore, 20 mmol/L is unaccounted for in the Δ/Δ value (ΔAG to ΔHCO_3^-).

TABLE 51-3

Steps in Acid-Base Diagnosis

1. Obtain arterial blood gas (ABG) and electrolytes simultaneously.
2. Compare [HCO_3^-] on ABG and electrolytes to verify accuracy.
3. Calculate anion gap (AG), but correct to a normal albumin concentration of 4.5 g/dL.
4. Know four causes of high-AG acidosis (ketoacidosis, lactic acid acidosis, renal failure, and toxins).
5. Know two causes of hyperchloremic or nongap acidosis (bicarbonate loss from gastrointestinal tract, renal tubular acidosis).
6. Estimate compensatory response (Table 51-1).
7. Compare ΔAG and ΔHCO_3^-.
8. Compare change in [Cl^-] with change in [Na^+].

METABOLIC ACIDOSIS

Metabolic acidosis can occur because of an increase in endogenous acid production (such as lactate and ketoacids), loss of bicarbonate (as in diarrhea), or accumulation of endogenous acids because of inappropriately low excretion of net acid by the kidney (as in chronic kidney disease [CKD]). Metabolic acidosis has profound effects on the respiratory, cardiac, and nervous systems. The fall in blood pH is accompanied by a characteristic increase in ventilation, especially the tidal volume (Kussmaul respiration). Intrinsic cardiac contractility may be depressed, but inotropic function can be normal because of catecholamine release. Both peripheral arterial vasodilation and central venoconstriction can be present; the decrease in central and pulmonary vascular compliance predisposes to pulmonary edema with even minimal volume overload. CNS function is depressed, with headache, lethargy, stupor, and, in some cases, even coma. Glucose intolerance may also occur.

There are two major categories of clinical metabolic acidosis: high-AG and non-AG acidosis (Table 51-3 and **Table 51-4**). The presence of metabolic acidosis, a normal AG, and hyperchloremia denotes the presence of a normal AG metabolic acidosis.

TABLE 51-4

Causes of High-Anion Gap Metabolic Acidosis

Lactic acidosis	Toxins
Ketoacidosis	Ethylene glycol
Diabetic	Methanol
Alcoholic	Salicylates
Starvation	Propylene glycol
	Pyroglutamic acid (5-oxoproline)
	Renal failure (acute and chronic)

TREATMENT

TREATMENT

Metabolic Acidosis

Treatment of metabolic acidosis with alkali should be reserved for severe acidemia except when the patient has no "potential HCO_3^-" in plasma. The potential [HCO_3^-] can be estimated from the increment (Δ) in the AG (ΔAG = patient's AG – 10), only if the acid anion that has accumulated in plasma is metabolizable (i.e., β-hydroxybutyrate, acetoacetate, and lactate). Conversely non-metabolizable anions that may accumulate in advanced stage CKD or after toxin ingestion are not metabolizable and do not represent "potential" HCO_3^-.

With acute CKD improvement in kidney function to replenish the [HCO_3^-] deficit is a slow and often unpredictable process. Consequently, patients with a normal AG acidosis (hyperchloremic acidosis) or an AG attributable to a non-metabolizable anion due to advanced kidney failure should receive alkali therapy, either PO ($NaHCO_3$ or Shohl's solution) or IV ($NaHCO_3$), in an amount necessary to slowly increase the plasma [HCO_3^-] to a target value of 22 mmol/L. Nevertheless, overcorrection should be avoided.

Controversy exists in regard to the use of alkali in patients with a pure AG acidosis owing to accumulation of a metabolizable organic acid anion (ketoacidosis or lactic acidosis). In general, severe acidemia (pH <7.10) in an adult patient (especially the elderly and patients with severe heart disease) warrants the IV administration of 50 meq of $NaHCO_3$ diluted in 300 mL of sterile water over 30–45 min, during the initial 1–2 h of therapy.

Provision of such modest quantities of alkali in this situation seems to provide an added measure of safety. Administration of alkali requires careful monitoring of plasma electrolytes, especially the plasma [K^+], during the course of therapy. A reasonable initial goal is to increase the [HCO_3^-] to 10–12 mmol/L and the pH to ~7.20, but clearly not to increase these values to normal. Estimation of the "bicarbonate deficit" by calculation of the volume of distribution of bicarbonate is often taught but is unnecessary and may result in administration of excessive amounts of alkali.

HIGH-ANION GAP ACIDOSES

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

There are four principal causes of a high-AG acidosis: (1) lactic acidosis, (2) ketoacidosis, (3) ingested toxins, and (4) acute and chronic renal failure (Table 51-4). Initial screening to differentiate the high-AG acidoses should include (1) a probe of the history for evidence of drug and toxin ingestion and measurement of arterial blood gas to detect coexistent respiratory alkalosis (salicylates); (2) determination of whether diabetes mellitus is present (diabetic ketoacidosis); (3) a search for evidence of alcoholism or increased levels of β-hydroxybutyrate (alcoholic ketoacidosis); (4) observation for clinical signs of uremia and determination of the blood urea nitrogen (BUN) and creatinine (uremic acidosis); (5) inspection of the urine for oxalate crystals (ethylene glycol); and (6) recognition of the numerous clinical settings in which lactate levels may be increased (hypotension, shock, cardiac failure, leukemia, cancer, and drug or toxin ingestion).

Lactic Acidosis

An increase in plasma l-lactate may be secondary to poor tissue perfusion (type A)—circulatory insufficiency (shock, cardiac failure), severe anemia, mitochondrial enzyme defects, and inhibitors (carbon monoxide, cyanide)—or to aerobic disorders (type B)—malignancies, nucleoside analogue reverse transcriptase inhibitors in HIV, diabetes mellitus, renal or hepatic failure, thiamine deficiency, severe infections (cholera, malaria), seizures, or drugs/toxins (biguanides, ethanol, and the toxic alcohols: ethylene glycol (EG), methanol, or propylene glycol). Unrecognized bowel ischemia or infarction in a patient with severe atherosclerosis or cardiac decompensation receiving vasopressors is a common cause of lactic acidosis in elderly patients. Pyroglutamic acidemia may occur in critically ill patients receiving acetaminophen, which causes depletion of glutathione and accumulation of 5-oxyprolene. D-Lactic acid acidosis, which may be associated with jejunoileal bypass, short bowel syndrome, or intestinal obstruction, is due to formation of D-lactate by gut bacteria.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

L-Lactic Acid Acidosis

The underlying condition that disrupts lactate metabolism should be corrected preemptively, if possible; tissue perfusion must be restored when inadequate, but vasoconstrictors should be avoided, if possible, because they may worsen tissue perfusion. Alkali therapy is generally advocated for acute, severe acidemia (pH <7.00) to improve cardiovascular function. However, $NaHCO_3$ therapy may paradoxically depress cardiac performance and exacerbate acidosis by enhancing lactate production (HCO_3^- stimulates phosphofructokinase). While the use of alkali in moderate lactic acidosis is controversial, it is generally agreed that attempts to return the pH or [HCO_3^-] to normal by administration of exogenous $NaHCO_3$ are deleterious. A reasonable approach is to infuse sufficient $NaHCO_3$ to raise the arterial pH to no more than 7.2 or the [HCO_3^-] to no more than 12, over 30–40 min.

$NaHCO_3$ therapy can cause fluid overload and hypertension because the amount required can be massive when accumulation of lactic acid is relentless. Fluid administration is poorly tolerated, especially in the oliguric patient, when central venoconstriction coexists. When the underlying cause of the lactic acidosis can be remedied, blood lactate will be converted to HCO_3^- and may result in an overshoot alkalosis if excess $NaHCO_3$ has been administered excessively.

Ketoacidosis

DIABETIC KETOACIDOSIS (DKA)

This condition is caused by increased fatty acid metabolism and the accumulation of ketoacids (acetoacetate and β-hydroxybutyrate). DKA usually occurs in insulin-dependent diabetes mellitus in association with cessation of insulin or an intercurrent illness such as an infection, gastroenteritis, pancreatitis, or myocardial infarction, which increases insulin requirements temporarily and acutely. The accumulation of ketoacids accounts for the increment in the AG and is accompanied most often by hyperglycemia (glucose >17 mmol/L [300 mg/dL]). The relationship between the ΔAG and ΔHCO_3^- is usually 1:1 in DKA. It should be noted that because insulin prevents production of ketones, bicarbonate therapy is rarely needed except with extreme acidemia (pH < 7.10), and then in only limited amounts. Patients with DKA are typically volume depleted and require fluid resuscitation with isotonic saline. Volume overexpansion with IV isotonic fluid administration is not uncommon, however, and contributes to the development of a hyperchloremic acidosis during treatment of DKA. The mainstay for treatment of this condition is IV regular insulin and is described in **Chap. 396** in more detail.

ALCOHOLIC KETOACIDOSIS (AKA)

Chronic alcoholics can develop ketoacidosis when alcohol consumption is abruptly curtailed and nutrition is poor. AKA is usually associated with binge drinking, vomiting, abdominal pain, starvation, and volume depletion. The glucose concentration is variable, and acidosis may be severe because of elevated ketones, predominantly β-hydroxybutyrate. Hypoperfusion may enhance lactic acid production, chronic respiratory alkalosis may accompany liver disease, and metabolic alkalosis can result from vomiting (refer to the relationship between ΔAG and ΔHCO_3^-). Thus, mixed acid-base disorders are common in AKA. As the

circulation is restored by administration of isotonic saline, the preferential accumulation of β-hydroxybutyrate is then shifted to acetoacetate. This explains the common clinical observation of an increasingly positive nitroprusside reaction (ketones) as the patient improves. The nitroprusside ketone reaction (Acetest) can detect acetoacetic acid but not β-hydroxybutyrate, so that the degree of ketosis and ketonuria can not only change with therapy, but can be underestimated initially. Patients with AKA usually present with relatively normal renal function, as opposed to DKA, where renal function is often compromised because of volume depletion (osmotic diuresis) or diabetic nephropathy. The AKA patient with normal renal function may excrete relatively large quantities of ketoacids in the urine and, therefore, may have a relatively normal AG and a discrepancy in the $\Delta AG/\Delta HCO_3^-$ relationship.

TREATMENT

TREATMENT

Alcoholic Ketoacidosis

Extracellular fluid deficits almost always accompany AKA and should be repleted by IV administration of saline and glucose (5% dextrose in 0.9% NaCl). Hypophosphatemia, hypokalemia, and hypomagnesemia may coexist and should be monitored carefully and corrected when indicated. Hypophosphatemia typically emerges 12–24 h after admission, may be exacerbated by glucose infusion, and, if severe, may induce rhabdomyolysis or even respiratory arrest. Upper gastrointestinal hemorrhage, pancreatitis, and pneumonia may accompany this disorder.

Drug- and Toxin-Induced Acidosis

Salicylates

Salicylate intoxication in adults usually causes respiratory alkalosis or a mixture of high-AG metabolic acidosis and respiratory alkalosis (See also Chap. 449). Only a portion of the AG is due to salicylates. Lactic acid production is also often increased.

TREATMENT

TREATMENT

Salicylate-Induced Acidosis

Vigorous gastric lavage with isotonic saline (not $NaHCO_3$) should be initiated immediately. All patients should receive at least one round of activated charcoal per nasogastric tube (1 g/kg up to 50 g). In the acidotic patient, to facilitate removal of salicylate, IV $NaHCO_3$ is administered in amounts adequate to alkalinize the urine and to maintain urine output (urine pH >7.5), because raising the urine pH from 6.5 to 7.5 increases salicylate clearance fivefold. Patients with coexisting respiratory alkalosis should also receive $NaHCO_3$, but with caution to avoid excessive alkalemia. Acetazolamide may be administered in the face of alkalemia, when an alkaline diuresis cannot be achieved, or to ameliorate volume overload associated with $NaHCO_3$ administration, but this drug can cause systemic metabolic acidosis if the excreted HCO_3^- is not replaced, a circumstance that can markedly reduce salicylate clearance.

Hypokalemia should be anticipated with vigorous bicarbonate therapy and should be treated promptly and aggressively. Glucose-containing fluids should be administered because of the danger of hypoglycemia. Excessive insensible fluid losses may cause severe volume depletion and hypernatremia. If renal failure prevents rapid clearance of salicylate, hemodialysis can be performed against a bicarbonate–containing dialysate.

ALCOHOLS

Under most physiologic conditions, sodium, urea, and glucose generate the osmotic pressure of blood. Plasma osmolality is calculated according to the following expression: $P_{osm} = 2Na^+ + Glu + BUN$ (all in mmol/L), or, using conventional laboratory values in which glucose and BUN are expressed in milligrams per deciliter: $P_{osm} = 2Na^+ + Glu/18 + BUN/2.8$. The calculated and determined osmolality should agree within 10–15 mmol/kg H_2O. When the measured osmolality exceeds the calculated osmolality by >10–15 mmol/kg H_2O, one of two circumstances prevails. Either the serum sodium is spuriously low, as with hyperlipidemia or hyperproteinemia (pseudohyponatremia), or osmolytes other than sodium salts, glucose, or urea have accumulated in plasma. Examples of such osmolytes include mannitol, radiocontrast media, ethanol, isopropyl alcohol, EG, propylene glycol, methanol, and acetone. In this situation, the difference between the calculated osmolality and the measured osmolality (*osmolar gap*) is proportional to the concentration of the unmeasured solute. With an appropriate clinical history and index of suspicion, identification of an osmolar gap is helpful in identifying the presence of toxic alcohol-associated AG acidosis. Three alcohols may cause fatal intoxications: EG, methanol, and isopropyl alcohol. All cause an elevated osmolal gap, but only the first two cause a high-AG acidosis. Isopropyl alcohol ingestion does not typically elevate the AG unless extreme overdose causes hypotension and lactic acid acidosis.

ETHYLENE GLYCOL

Ingestion of EG (commonly used in antifreeze) leads to a metabolic acidosis and severe damage to the CNS, heart, lungs, and kidneys (See also Chap. 449). The combination of a high AG and high osmolar gap is highly suspicious for EG or methanol intoxication. The increased AG and osmolar gap in EG intoxication are attributable to EG and its metabolites, oxalic acid, glycolic acid, and other organic acids. Lactic acid production increases secondary to inhibition of the tricarboxylic acid cycle and altered intracellular redox state. In addition to the presence of elevated osmolar and AGs, the diagnosis is further enabled by recognition of oxalate crystals in the urine. Use of a Wood's lamp to visualize the fluorescent additive to commercial antifreeze in the urine of patients with EG ingestion, has been reported, but is not reliable. The combination of a high AG and high osmolar gap in a patient suspected of EG ingestion should be taken as evidence of EG toxicity. Treatment should not be delayed while awaiting measurement of EG levels in this setting.

TREATMENT

TREATMENT

Ethylene Glycol-Induced Acidosis

This includes the prompt institution of a saline or osmotic diuresis, thiamine and pyridoxine supplements, fomepizole, and usually, hemodialysis. The IV administration of the alcohol dehydrogenase inhibitor

fomepizole (4-methylpyrazole; 15 mg/kg as a loading dose) is the agent of choice and offers the advantages of a predictable decline in EG levels without excessive obtundation as seen during ethyl alcohol infusion. If used, ethanol IV should be infused to achieve a blood level of 22 mmol/L (100 mg/dL). Both fomepizole and ethanol reduce toxicity because they compete with EG for metabolism by alcohol dehydrogenase. Hemodialysis is indicated when the arterial pH is <7.3 or the osmolar gap exceeds 20 mOsm/kg.

METHANOL

The ingestion of methanol (wood alcohol) causes metabolic acidosis, and its metabolites formaldehyde and formic acid cause severe optic nerve and CNS damage (See also Chap. 449). Lactic acid, ketoacids, and other unidentified organic acids may contribute to the acidosis. Due to its low molecular mass (32 Da), an osmolar gap is usually present.

TREATMENT

TREATMENT

Methanol-Induced Acidosis

This is similar to that for EG intoxication, including general supportive measures, fomepizole, and hemodialysis (as above).

PROPYLENE GLYCOL

Propylene glycol is the vehicle used in IV administration of diazepam, lorazepam, phenobarbital, nitroglycerine, etomidate, enoximone, and phenytoin. Propylene glycol is generally safe for limited use in these IV preparations, but toxicity has been reported, most often in the setting of the intensive care unit in patients receiving frequent or continuous therapy. This form of high-gap acidosis should be considered in patients with unexplained high-gap acidosis, hyperosmolality, and clinical deterioration, especially in the setting of treatment for alcohol withdrawal. Propylene glycol, like EG and methanol, is metabolized by alcohol dehydrogenase. With intoxication by propylene glycol, the first response is to stop the offending infusion. Additionally, fomepizole should also be administered in acidotic patients.

ISOPROPYL ALCOHOL

Ingested isopropanol is absorbed rapidly and may be fatal when as little as 150 mL of rubbing alcohol, solvent, or deicer is consumed. A plasma level >400 mg/dL is life-threatening. Isopropyl alcohol is metabolized by alcohol dehydrogenase to acetone. The characteristic features differ significantly from EG and methanol intoxication in that the parent compound, not the metabolites, causes toxicity, and a high AG acidosis is ***not*** present because acetone is rapidly excreted. Both isopropyl alcohol and acetone increase the osmolar gap, and hypoglycemia is common. Alternative diagnoses should be considered if the patient does not improve significantly within a few hours. Patients with hemodynamic instability with plasma levels above 400 mg/dL should be considered for hemodialysis.

TREATMENT

TREATMENT

Isopropyl Alcohol Toxicity

Isopropanol alcohol toxicity is treated by supportive therapy, IV fluids, pressors, ventilatory support if needed, and occasionally hemodialysis for prolonged coma, hemodynamic instability, or levels >400 mg/dL.

PYROGLUTAMIC ACID

Acetaminophen-induced high-AG metabolic acidosis is uncommon but is being recognized more often in either patients with acetaminophen overdose or malnourished or critically ill patients receiving acetaminophen in typical dosage. 5-Oxoproline accumulation after acetaminophen should be suspected in the setting of an unexplained high-AG acidosis without elevation of the osmolar gap in patients receiving acetaminophen. The first step in treatment is to immediately discontinue the drug. Additionally, sodium bicarbonate IV should be given. Although *N*-acetylcysteine has been suggested, it is not known if it hastens the metabolism of 5-oxoproline by increasing intracellular glutathione concentrations in this setting.

Chronic Kidney Disease

The hyperchloremic acidosis of moderate CKD (Stage 3) is eventually converted to the high-AG acidosis of advanced renal failure (Stages 4 and 5 CKD) (See also Chap. 305). Poor filtration and reabsorption of organic anions contribute to the pathogenesis. As renal disease progresses, the number of functioning nephrons eventually becomes insufficient to keep pace with net acid production. Uremic acidosis in advanced CKD is characterized, therefore, by a reduced rate of NH_4^+ production and excretion. Alkaline salts from bone buffer the acid retained in chronic kidney disease. Despite significant retention of acid (up to 20 mmol/d), the serum [HCO_3^-] does not typically decrease further, indicating participation of buffers outside the extracellular compartment. Therefore, the trade-off in untreated chronic metabolic acidosis of CKD stages 3 and 4 is significant loss of bone mass due to reduction in bone calcium carbonate. Chronic acidosis also increases urinary calcium excretion, proportional to cumulative acid retention, and contributes significantly to muscle wasting.

TREATMENT

TREATMENT

Metabolic Acidosis of Chronic Kidney Disease

Because of the association of metabolic acidosis in advanced CKD with muscle catabolism, bone disease and more rapid progression of CKD, both the “uremic acidosis” of ESRD and the non-AG metabolic acidosis of stages 3 and 4 CKD require oral alkali replacement to maintain the [HCO_3^-] to approximately the normal value (25 mmol/L). This can be accomplished with relatively modest amounts of alkali (1.0–1.5 mmol/kg body weight per day). Either $NaHCO_3$ tablets (650-mg tablets contain 7.8 meq) or sodium citrate (Shohl’s solution) is effective.

NON–ANION GAP METABOLIC ACIDOSES

Alkali can be lost from the gastrointestinal tract as a result of diarrhea or from the kidneys due to renal tubular abnormalities (e.g., renal tubular acidosis [RTA]). In these disorders **(Table 51-5)**, reciprocal changes in $[Cl^-]$ and $[HCO_3^-]$ result in a normal AG. In pure non–AG acidosis, therefore, the increase in $[Cl^-]$ above the normal value approximates the decrease in $[HCO_3^-]$. The absence of such a relationship suggests a mixed disturbance.

TABLE 51-5

Causes of Non–Anion Gap Acidosis

I. Gastrointestinal bicarbonate loss
- A. Diarrhea
- B. External pancreatic or small-bowel drainage
- C. Ureterosigmoidostomy, jejunal loop, ileal loop
- D. Drugs
  1. Calcium chloride (acidifying agent)
  2. Magnesium sulfate (diarrhea)
  3. Cholestyramine (bile acid diarrhea)

II. Renal acidosis
- A. Hypokalemia
  1. Proximal RTA (type 2)

     Drug-induced: acetazolamide, topiramate
  2. Distal (classic) RTA (type 1)

     Drug-induced: amphotericin B, ifosfamide
- B. Hyperkalemia
  1. Generalized distal nephron dysfunction (type 4 RTA)
     a. Mineralocorticoid deficiency
     b. Mineralocorticoid resistance (PHA I, autosomal dominant)
     c. Voltage defect (PHA I, autosomal recessive, and PHA II)
     d. Tubulointerstitial disease
- C. Normokalemia
  1. Chronic progressive kidney disease

III. Drug-induced hyperkalemia (with renal insufficiency)
- A. Potassium-sparing diuretics (amiloride, triamterene, spironolactone, eplerenone)
- B. Trimethoprim
- C. Pentamidine
- D. ACE-Is and ARBs
- E. Nonsteroidal anti-inflammatory drugs
- F. Calcineurin inhibitors
- G. Heparin in critically ill patients

IV. Other
- A. Acid loads (ammonium chloride, hyperalimentation)
- B. Loss of potential bicarbonate: ketosis with ketone excretion
- C. Expansion acidosis (rapid saline administration)
- D. Hippurate
- E. Cation exchange resins

Abbreviations: ACE-I, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; PHA, pseudohypoaldosteronism; RTA, renal tubular acidosis.

Stool contains a higher concentration of HCO_3^- and decomposed HCO_3^- than plasma so that metabolic acidosis develops in diarrhea. Instead of an acid urine pH (as anticipated with systemic acidosis), urine pH is usually >6 because metabolic acidosis and hypokalemia increase renal synthesis and excretion of NH_4^+, thus providing a urinary buffer that increases urine pH. Metabolic acidosis due to gastrointestinal losses with a high urine pH can be differentiated from RTA because urinary NH_4^+ excretion is typically low in RTA and high with diarrhea. Urinary NH_4^+ levels can be estimated by calculating the urine AG (UAG): UAG = $[Na^+ + K^+]_u - [Cl^-]_u$. When $[Cl^-]_u > [Na^+ + K^+]_u$, the UAG is negative by definition. This indicates that the urine ammonium level is appropriately increased, suggesting an extrarenal cause of the acidosis. Conversely, when the UAG is positive, the urine ammonium level is low, suggesting a renal cause of the acidosis.

Proximal RTA (type 2 RTA) **(Chap. 309)** is most often due to generalized proximal tubular dysfunction manifested by glycosuria, generalized aminoaciduria, and phosphaturia (Fanconi syndrome). When the plasma $[HCO_3^-]$ is low the urine pH is acid (pH <5.5), but exceeds 5.5 with alkali therapy. The fractional excretion of $[HCO_3^-]$ may exceed 10–15% when the serum HCO_3^- is >20 mmol/L. Because HCO_3^- is not reabsorbed normally in the proximal tubule, therapy with $NaHCO_3$ will enhance delivery of HCO_3^- to the distal nephron and enhance renal potassium secretion, thereby, causing hypokalemia.

The typical findings in acquired or inherited forms of **classic distal RTA** (type 1 RTA) include hypokalemia, a non-AG metabolic acidosis, low urinary NH_4^+ excretion (positive UAG, low urine $[NH_4^+]$), and inappropriately high urine pH (pH > 5.5). Most patients have hypocitraturia and hypercalciuria, so nephrolithiasis, nephrocalcinosis, and bone disease are common. In **generalized distal RTA** (type 4 RTA), hyperkalemia is disproportionate to the reduction in glomerular filtration rate (GFR) because of coexisting dysfunction of potassium and acid secretion. Urinary ammonium excretion is invariably depressed, and kidney function may be compromised, for example, due to diabetic nephropathy, obstructive uropathy, or chronic tubulointerstitial disease.

Hyporeninemic hypoaldosteronism typically causes non-AG metabolic acidosis, most commonly in older adults with diabetes mellitus or tubulointerstitial disease and CKD. Patients usually have mild to moderate CKD (GFR, 20–50 mL/min) and acidosis, with elevation in serum $[K^+]$ (5.2–6.0 mmol/L), concurrent hypertension, and congestive heart failure. Both the metabolic acidosis and the hyperkalemia are out of proportion to impairment in GFR. Nonsteroidal anti-inflammatory drugs, trimethoprim, pentamidine, angiotensin-converting enzyme (ACE) inhibitors, and aldosterone receptor blockers (ARBs), can also increase the risk for a hyperkalemia and a non-AG metabolic acidosis in patients with CKD (Table 51-5).

TREATMENT

TREATMENT

Non–Anion Gap Metabolic Acidoses

For non-renal causes of non-AG acidosis due to gastrointestinal losses of bicarbonate, $NaHCO_3$ may be administered intravenously or orally, as determined by the severity of both the acidosis and the accompanying volume depletion. Proximal RTA is the most challenging of the RTAs to treat if the goal is to restore the serum [HCO_3^-] to normal, because administration of oral alkali increases urinary excretion of potassium. In patients with proximal RTA (type 1), potassium administration is typically required. An oral solution of a combination of sodium and potassium citrate (citric acid 334 mg, sodium citrate 500 mg, and potassium citrate 550 mg per 5 mL) may be prescribed for this purpose and is available commercially as Virtrate-3. The syrup preparation is not recommended for chronic administration. In classical distal RTA (type 2), potassium should be administered in the acutely acidotic patient with hypokalemia. For chronic therapy, most patients respond to replacement with either sodium citrate (Shohl's solution) or $NaHCO_3$ tablets (650-mg tablets contain 7.8 meq) with the goal of correcting the serum [HCO_3^-] to normal. These patients typically respond to chronic alkali therapy readily and the benefits of adequate alkali therapy include a decrease in the frequency of nephrolithiasis, improvement in bone density, resumption of normal growth patterns in children, and preservation of kidney function in both adults and children. For type 4 RTA, attention must be paid to the dual goals of correction of the metabolic acidosis, using the same approach as for cDRTA, but in addition, effort toward correcting the plasma [K^+] is necessary. This latter goal deserves emphasis because restoration of normokalemia increases urinary net acid excretion and in that way can greatly improve the metabolic acidosis. Chronic administration of oral sodium polystyrene solfonate (15 g of power prepared as an oral solution, and without sorbitol, once daily 2–3 times per week) is sometimes used. Additionally, the diet should be low in potassium-containing foods, all potassium-retaining medications should be discontinued, and a loop diuretic may be administered. The recent release of a new a non-absorbed, calcium-potassium cation exchange polymer, patiromer, may prove to be very useful for type 4 RTA patients with significant hyperkalemia. However, patiromer has not yet been investigated in this population of patients. Finally, patients with demonstrated adrenal insufficiency should also receive fludocortisone, but the dose varies with the cause of the hormone deficiency, and should be assiduously avoided in patients with hyporeninemic-hypoaldosteronism.

METABOLIC ALKALOSIS

Metabolic alkalosis is established by an elevated arterial pH, an increase in the serum [HCO_3^-], and an increase in $Paco_2$ as a result of compensatory alveolar hypoventilation (Table 51-1). It is often accompanied by hypochloremia and hypokalemia. The arterial pH establishes the diagnosis, because it is increased in

metabolic alkalosis and decreased in respiratory acidosis. Metabolic alkalosis frequently occurs as a mixed acid base disorder in association with either respiratory acidosis, respiratory alkalosis, or metabolic acidosis.

PATHOGENESIS

Metabolic alkalosis occurs as a result of net gain of [HCO_3^-] or loss of nonvolatile acid (usually HCl by vomiting) from the extracellular fluid. When vomiting causes loss of HCl from the stomach, HCO_3^- secretion cannot be initiated in the small bowel and thus HCO_3 is added to the extracellular fluid. Thus, vomiting or nasogastric (NG) suction is an example of the *generation stage*, in which the loss of acid typically causes alkalosis. Upon cessation of vomiting, the *maintenance stage*, typically ensues because secondary factors prevent the kidneys from compensating by excreting HCO_3^-.

Maintenance of metabolic alkalosis, therefore, represents a failure of the kidneys to eliminate excess HCO_3^- from the extracellular compartment. The kidneys will retain, rather than excrete, the excess alkali and maintain the alkalosis if (1) volume deficiency, chloride deficiency, and K^+ deficiency exist in combination with a reduced GFR; or (2) hypokalemia exists because of autonomous hyperaldosteronism. In the first example, alkalosis is corrected by administration of NaCl and KCl, whereas, in the latter, it may be necessary to repair the alkalosis by pharmacologic or surgical intervention, not with saline administration.

DIFFERENTIAL DIAGNOSIS

To establish the cause of metabolic alkalosis **(Table 51-6)**, it is necessary to assess the status of the extracellular fluid volume (ECFV), the recumbent and upright blood pressure (to determine if orthostasis is present), the serum [K^+], and in some circumstances, an assessment of the renin-aldosterone system. For example, the presence of chronic hypertension and chronic hypokalemia in an alkalotic patient suggests either mineralocorticoid excess or that the hypertensive patient is receiving diuretics. Low plasma renin activity and normal values for both the urine [Na^+] and [Cl^-], in a patient who is not taking diuretics, suggest primary mineralocorticoid excess. The combination of hypokalemia and alkalosis in a normotensive, nonedematous patient can be due to Bartter's or Gitelman's syndrome, magnesium deficiency, vomiting, exogenous alkali, or diuretic ingestion. Measurement of urine electrolytes (especially the urine [Cl^-]) and screening of the urine for diuretics is recommended. If the urine is alkaline, with an elevated $[Na^+]_u$ and $[K^+]_u$ but low $[Cl^-]_u$, the diagnosis is usually either vomiting (overt or surreptitious) or alkali ingestion. If the urine is relatively acid and has low concentrations of Na^+, K^+, and Cl^-, the most likely possibilities are prior vomiting, the posthypercapnic state, or prior diuretic ingestion. If, on the other hand, neither the urine sodium, potassium, nor chloride concentrations are depressed, magnesium deficiency, Bartter's or Gitelman's syndrome, or current diuretic ingestion should be considered. Bartter's syndrome is distinguished from Gitelman's syndrome because of hypocalciuria in the latter disorder.

TABLE 51-6

Causes of Metabolic Alkalosis

I. Exogenous HCO_3^- loads
   A. Acute alkali administration
   B. Milk-alkali syndrome

II. Effective ECFV contraction, normotension, K^+ deficiency, and secondary hyperreninemic hyperaldosteronism
   A. Gastrointestinal origin
      1. Vomiting
      2. Gastric aspiration
      3. Congenital chloridorrhea
      4. Gastrocystoplasty
      5. Villous adenoma
   B. Renal origin
      1. Diuretics
      2. Posthypercapnic state
      3. Hypercalcemia/hypoparathyroidism
      4. Recovery from lactic acidosis or ketoacidosis
      5. Nonreabsorbable anions including penicillin, carbenicillin
      6. Mg^{2+} deficiency
      7. K^+ depletion
      8. Bartter's syndrome (loss of function mutations of transporters and ion channels in TALH)
      9. Gitelman's syndrome (loss of function mutation of Na^+-Cl^- cotransporter in DCT)

III. ECFV expansion, hypertension, K^+ deficiency, and mineralocorticoid excess
   A. High renin
      1. Renal artery stenosis
      2. Accelerated hypertension
      3. Renin-secreting tumor
      4. Estrogen therapy
   B. Low renin
      1. Primary aldosteronism
         a. Adenoma
         b. Hyperplasia
         c. Carcinoma
      2. Adrenal enzyme defects
         a. 11β-Hydroxylase deficiency
         b. 17α-Hydroxylase deficiency
      3. Cushing's syndrome or disease

4. Other
    a. Licorice
    b. Carbenoxolone
    c. Chewer's tobacco

IV. Gain-of-function mutation of sodium channel in DCT with ECFV expansion, hypertension, K^+ deficiency, and hyporeninemic-hypoaldosteronism
    A. Liddle's syndrome

Abbreviations: DCT, distal convoluted tubule; ECFV, extracellular fluid volume; TALH, thick ascending limb of Henle's loop.

Alkali Administration

Chronic administration of alkali to individuals with normal renal function rarely causes alkalosis. However, in patients with coexistent hemodynamic disturbances associated with effective ECF volume depletion, alkalosis can develop because the normal capacity to excrete HCO_3^- is diminished or there may be enhanced reabsorption of HCO_3^-. Such patients include those who receive $NaHCO_3$ (PO or IV), citrate loads (transfusions of whole blood, or therapeutic apheresis), or antacids plus cation-exchange resins (aluminum hydroxide and sodium polystyrene sulfonate). Nursing home patients receiving enteral tube feedings (an often overlooked source of alkali loads) have a higher incidence of metabolic alkalosis than nursing home patients receiving regular diets.

METABOLIC ALKALOSIS ASSOCIATED WITH ECFV CONTRACTION, K^+ DEPLETION, AND SECONDARY HYPERRENINEMIC HYPERALDOSTERONISM

Gastrointestinal Origin

Gastrointestinal loss of H^+ from vomiting or gastric aspiration causes simultaneous addition of HCO_3^- into the extracellular fluid. During active vomiting, the filtered load of bicarbonate reaching the kidneys is acutely increased and will exceed the reabsorptive capacity of the proximal tubule for HCO_3^- absorption. Subsequently, enhanced delivery of HCO_3 to the distal nephron will cause excretion of alkaline urine that is high in potassium. When vomiting ceases, the persistence of volume, potassium, and chloride depletion triggers maintenance of the alkalosis because these conditions promote HCO_3^- reabsorption. Correction of the contracted ECFV with NaCl and repair of K^+ deficits with KCl corrects the acid-base disorder by restoring the ability of the kidney to excrete the excess bicarbonate.

Renal Origin

DIURETICS

Diuretics such as thiazides and loop diuretics (furosemide, bumetanide, torsemide) increase excretion of salt and acutely diminish the ECFV without altering the total body bicarbonate content (See also Chap. 252). The serum [HCO_3^-] increases because the reduced ECFV "contracts" around the [HCO_3^-] in the plasma (contraction alkalosis). The chronic administration of diuretics tends to generate an alkalosis by increasing distal salt delivery, so that both K^+ and H^+ secretion are stimulated. The alkalosis is maintained by persistence of the contraction of the ECFV, secondary hyperaldosteronism, K^+ deficiency, and the direct effect of the diuretic (as long as diuretic administration continues). Discontinuing the diuretic and providing isotonic saline to correct the ECFV deficit will repair the alkalosis.

SOLUTE LOSING DISORDERS: BARTTER'S SYNDROME AND GITELMAN'S SYNDROME

See Chap. 309.

NONREABSORBABLE ANIONS AND MAGNESIUM DEFICIENCY

Administration of large quantities of the penicillin derivatives carbenicillin or tricarcillin cause their nonreabsorbable anions to appear in the urine. This increases the transepithelial potential difference in the collecting tubule, and thereby enhances H^+ and K^+ secretion. Mg^{2+} deficiency, may occur with chronic administration of thiazide diuretics, alcoholism, and malnutrition, and in Gitelman's syndrome potentiates the development of hypokalemic alkalosis by enhancing distal acidification through stimulation of renin and hence aldosterone secretion.

POTASSIUM DEPLETION

Chronic K^+ depletion may cause metabolic alkalosis by increasing urinary acid excretion. The renal generation of NH_4^+ (ammoniagenesis) is upregulated directly by hypokalemia. Chronic K^+ deficiency also upregulates the renal H^+, K^+-ATPase to increase K^+ absorption at the expense of enhanced H^+ secretion. Alkalosis associated with severe K^+ depletion is resistant to salt administration, but repair of the K^+ deficiency corrects the alkalosis. Potassium depletion often occurs concomitant with magnesium deficiency in alcoholics with malnutrition.

AFTER TREATMENT OF LACTIC ACIDOSIS OR KETOACIDOSIS

When an underlying stimulus for the generation of lactic acid or ketoacid is corrected by treatment of the underlying disorder, such as correction shock or severe volume depletion by volume restoration, or with insulin therapy, respectively, the lactate or ketones are metabolized to yield an equivalent amount of HCO_3^-. Exogenous sources of HCO_3^- will be additive with that amount generated by organic anion metabolism to create a surfeit of HCO_3^-. Acidosis-induced contraction of the ECFV and K^+ deficiency act in concert to sustain the alkalosis.

POSTHYPERCAPNIA

Prolonged CO_2 retention with chronic respiratory acidosis enhances renal HCO_3^- absorption and the generation of new HCO_3^- (increased net acid excretion). Metabolic alkalosis results from the effect of the

persistently elevated [HCO_3^-] when the elevated $Paco_2$ is abruptly returned toward normal.

METABOLIC ALKALOSIS ASSOCIATED WITH ECFV EXPANSION, HYPERTENSION, AND HYPERALDOSTERONISM

Increased aldosterone levels may be the result of autonomous primary adrenal overproduction or of secondary aldosterone release due to renal overproduction of renin. Mineralocorticoid excess increases net acid excretion and may result in metabolic alkalosis, which is typically exacerbated by associated K^+ deficiency. Salt retention is due to upregulation of the epithelial Na^+ channels in the collecting tubule to aldosterone, as a result of the associated ECFV expansion, causes hypertension. The kaliuresis persists because of mineralocorticoid excess and distal Na^+ absorption causing enhanced K^+ excretion, continued K^+ depletion with polydipsia, inability to concentrate the urine, and polyuria.

Liddle's syndrome (Chap. 309) results from an inherited gain of function mutation of genes that regulate the collecting duct Na^+ channel (ENaC). Liddle's is a rare monogenic form of hypertension due to volume expansion manifested as hypokalemic alkalosis and normal aldosterone levels.

Symptoms

With metabolic alkalosis, changes in CNS and peripheral nervous system function are similar to those of hypocalcemia (Chap. 402); symptoms include mental confusion; obtundation; and a predisposition to seizures, paresthesia, muscular cramping, tetany, aggravation of arrhythmias, and hypoxemia in chronic obstructive pulmonary disease. Related electrolyte abnormalities include hypokalemia and hypophosphatemia.

TREATMENT

TREATMENT

Metabolic Alkalosis

This is primarily directed at correcting the underlying stimulus for HCO_3^- generation. If primary aldosteronism or Cushing's syndrome is present, correction of the underlying cause, when successful, will reverse the hypokalemia and alkalosis. [H^+] loss by the stomach or kidneys can be mitigated by the use of proton pump inhibitors or the discontinuation of diuretics. The second aspect of treatment is to remove the factors that sustain the inappropriate increase in HCO_3^- reabsorption, such as ECFV contraction or K^+ deficiency. K^+ deficits should always be repaired. Isotonic saline is recommended to reverse the alkalosis when ECFV contraction is present. If associated conditions preclude infusion of saline, renal HCO_3^- loss can be accelerated by administration of acetazolamide, a carbonic anhydrase inhibitor (125–250 mg IV), which is usually effective in patients with adequate renal function but can worsen K^+ losses. Dilute hydrochloric acid (0.1 *N* HCl) has been advocated historically in extreme cases, but can cause hemolysis, and must be delivered

slowly in a central vein. This preparation is not available generally and must be mixed by the pharmacist. Because serious errors or harm may occur, its use is not recommended.

RESPIRATORY ACIDOSIS

Respiratory acidosis can be due to severe pulmonary disease, respiratory muscle fatigue, or abnormalities in ventilatory control and is recognized by an increase in $Paco_2$ and decrease in pH **(Table 51-7)**. In acute respiratory acidosis, there is a compensatory elevation (due to cellular buffering mechanisms) in HCO_3^-, which increases 1 mmol/L for every 10-mmHg increase in $Paco_2$. In chronic respiratory acidosis (>24 h), renal adaptation increases the $[HCO_3^-]$ by 4 mmol/L for every 10-mmHg increase in $Paco_2$. The serum HCO_3^- usually does not increase above 38 mmol/L.

TABLE 51-7

Respiratory Acid-Base Disorders

I. Alkalosis
- A. Central nervous system stimulation
  1. Pain
  2. Anxiety, psychosis
  3. Fever
  4. Cerebrovascular accident
  5. Meningitis, encephalitis
  6. Tumor
  7. Trauma
- B. Hypoxemia or tissue hypoxia
  1. High altitude
  2. Pneumonia, pulmonary edema
  3. Aspiration
  4. Severe anemia
- C. Drugs or hormones
  1. Pregnancy, progesterone
  2. Salicylates
  3. Cardiac failure
- D. Stimulation of chest receptors
  1. Hemothorax
  2. Flail chest
  3. Cardiac failure
  4. Pulmonary embolism
- E. Miscellaneous
  1. Septicemia
  2. Hepatic failure
  3. Mechanical hyperventilation
  4. Heat exposure
  5. Recovery from metabolic acidosis

II. Acidosis
- A. Central
  1. Drugs (anesthetics, morphine, sedatives)
  2. Stroke
  3. Infection
- B. Airway
  1. Obstruction
  2. Asthma

C. Parenchyma
   1. Emphysema
   2. Pneumoconiosis
   3. Bronchitis
   4. Adult respiratory distress syndrome
   5. Barotrauma

D. Neuromuscular
   1. Poliomyelitis
   2. Kyphoscoliosis
   3. Myasthenia
   4. Muscular dystrophies

E. Miscellaneous
   1. Obesity
   2. Hypoventilation
   3. Permissive hypercapnia

The clinical features vary according to the severity and duration of the respiratory acidosis, the underlying disease, and whether there is accompanying hypoxemia. A rapid increase in $Paco_2$ may cause anxiety, dyspnea, confusion, psychosis, and hallucinations and may progress to coma. Lesser degrees of dysfunction in chronic hypercapnia include sleep disturbances; loss of memory; daytime somnolence; personality changes; impairment of coordination; and motor disturbances such as tremor, myoclonic jerks, and asterixis. Headaches and other signs that mimic raised intracranial pressure, such as papilledema, abnormal reflexes, and focal muscle weakness, are due to vasoconstriction secondary to loss of the vasodilator effects of CO_2.

Depression of the respiratory center by a variety of drugs, injury, or disease can produce respiratory acidosis. This may occur acutely with general anesthetics, sedatives, and head trauma or chronically with sedatives, alcohol, intracranial tumors, and the syndromes of sleep-disordered breathing including the primary alveolar and obesity-hypoventilation syndromes **(Chaps. 290 and 291)**. Abnormalities or disease in the motor neurons, neuromuscular junction, and skeletal muscle can cause hypoventilation via respiratory muscle fatigue. Mechanical ventilation, when not properly adjusted and supervised, may result in respiratory acidosis, particularly if CO_2 production suddenly rises (because of fever, agitation, sepsis, or overfeeding) or alveolar ventilation falls because of worsening pulmonary function. High levels of positive end-expiratory pressure in the presence of reduced cardiac output may cause hypercapnia as a result of large increases in alveolar dead space **(Chap. 279)**. Permissive hypercapnia may be used to minimize intrinsic positive end-expiratory pressure in acute lung injury/acute respiratory distress syndrome and severe obstructive lung disease. The respiratory acidosis associated with permissive hypercapnia may require administration of $NaHCO_3$ to increase the arterial pH to ~7.15–7.20, but correction of the acidemia to a normal arterial pH is deleterious.

Acute hypercapnia follows sudden occlusion of the upper airway or generalized bronchospasm as in severe asthma, anaphylaxis, inhalational burn, or toxin injury. Chronic hypercapnia and respiratory acidosis occur in end-stage obstructive lung disease. Restrictive disorders involving both the chest wall and the lungs can cause respiratory acidosis because the high metabolic cost of respiration causes ventilatory muscle fatigue. Advanced stages of intrapulmonary and extrapulmonary restrictive defects present as chronic respiratory acidosis.

The diagnosis of respiratory acidosis requires the measurement of $Paco_2$ and arterial pH. A detailed history and physical examination often indicate the cause. Pulmonary function studies **(Chap. 279)**, including spirometry, diffusion capacity for carbon monoxide, lung volumes, and arterial $Paco_2$ and O_2 saturation, usually make it possible to determine if respiratory acidosis is secondary to lung disease. The workup for nonpulmonary causes should include a detailed drug history, measurement of hematocrit, and assessment of upper airway, chest wall, pleura, and neuromuscular function.

TREATMENT

TREATMENT

Respiratory Acidosis

The management of respiratory acidosis depends on its severity and rate of onset. Acute respiratory acidosis can be life-threatening, and measures to reverse the underlying cause should be undertaken simultaneously with restoration of adequate alveolar ventilation. This may necessitate tracheal intubation and assisted mechanical ventilation. Oxygen administration should be titrated carefully in patients with severe obstructive pulmonary disease and chronic CO_2 retention who are breathing spontaneously **(Chap. 286)**. When oxygen is used injudiciously, these patients may experience progression of the respiratory acidosis causing severe acidemia. Aggressive and rapid correction of hypercapnia should be avoided, because the falling $Paco_2$ may provoke the same complications noted with acute respiratory alkalosis (i.e., cardiac arrhythmias, reduced cerebral perfusion, and seizures). The $Paco_2$ should be lowered gradually in chronic respiratory acidosis, aiming to restore the $Paco_2$ to baseline levels and to provide sufficient Cl^- and K^+ to enhance the renal excretion of HCO_3^-.

Chronic respiratory acidosis is frequently difficult to correct, but measures aimed at improving lung function **(Chap. 286)** should be the primary focus of treatment.

RESPIRATORY ALKALOSIS

Alveolar hyperventilation decreases $Paco_2$ and increases the $HCO_3^-/Paco_2$ ratio, thus increasing pH (Table 51-7). Nonbicarbonate cellular buffers respond by consuming HCO_3^-. Hypocapnia develops when a

sufficiently strong ventilatory stimulus causes CO_2 output in the lungs to exceed its metabolic production by tissues. Plasma pH and [HCO_3^-] appear to vary proportionately with $Paco_2$ over a range from 40–15 mmHg. The relationship between arterial [H^+] concentration and $Paco_2$ is ~0.7 mmol/L per mmHg (or 0.01 pH unit/mmHg), and that for plasma [HCO_3^-] is 0.2 mmol/L per mmHg. Hypocapnia sustained for >2–6 h is further compensated by a decrease in renal ammonium and titratable acid excretion and a reduction in filtered HCO_3^- reabsorption. Full renal adaptation to respiratory alkalosis may take several days and requires normal volume status and renal function. The kidneys appear to respond directly to the lowered $Paco_2$ rather than to alkalosis per se. In chronic respiratory alkalosis a 1-mmHg decrease in $Paco_2$ causes a 0.4- to 0.5-mmol/L drop in [HCO_3^-] and a 0.3-mmol/L decrease (or 0.003 increase in pH) in [H^+].

The effects of respiratory alkalosis vary according to duration and severity but are primarily those of the underlying disease. Reduced cerebral blood flow as a consequence of a rapid decline in $Paco_2$ may cause dizziness, mental confusion, and seizures, even in the absence of hypoxemia. The cardiovascular effects of acute hypocapnia in the conscious human are generally minimal, but in the anesthetized or mechanically ventilated patient, cardiac output and blood pressure may fall because of the depressant effects of anesthesia and positive-pressure ventilation on heart rate, systemic resistance, and venous return. Cardiac arrhythmias may occur in patients with heart disease as a result of changes in oxygen unloading by blood from a left shift in the hemoglobin-oxygen dissociation curve (Bohr effect). Acute respiratory alkalosis causes intracellular shifts of Na^+, K^+, and PO_4^{2-} and reduces free [Ca^{2+}] by increasing the protein-bound fraction. Hypocapnia-induced hypokalemia is usually minor.

Chronic respiratory alkalosis is the most common acid-base disturbance in critically ill patients and, when severe, portends a poor prognosis. Many cardiopulmonary disorders manifest respiratory alkalosis in their early to intermediate stages, and the finding of normocapnia and hypoxemia in a patient with hyperventilation may herald the onset of rapid respiratory failure and should prompt an assessment to determine if the patient is becoming fatigued. Respiratory alkalosis is common during mechanical ventilation.

The hyperventilation syndrome may be disabling. Paresthesia; circumoral numbness; chest wall tightness or pain; dizziness; inability to take an adequate breath; and, rarely, tetany may be sufficiently stressful to perpetuate the disorder. Arterial blood-gas analysis demonstrates an acute or chronic respiratory alkalosis, often with hypocapnia in the range of 15–30 mmHg and no hypoxemia. CNS diseases or injury can produce several patterns of hyperventilation and sustained $Paco_2$ levels of 20–30 mmHg. Hyperthyroidism, high caloric loads, and exercise raise the basal metabolic rate, but ventilation usually rises in proportion so that arterial blood gases are unchanged and respiratory alkalosis does not develop. Salicylates are the most common cause of drug-induced respiratory alkalosis as a result of direct stimulation of the medullary chemoreceptor **(Chap. 449)**. The methylxanthines, theophylline, and aminophylline stimulate ventilation and

increase the ventilatory response to CO_2. Progesterone increases ventilation and lowers arterial $Paco_2$ by as much as 5–10 mmHg. Therefore, chronic respiratory alkalosis is a common feature of pregnancy. Respiratory alkalosis is also prominent in liver failure, and the severity correlates with the degree of hepatic insufficiency. Respiratory alkalosis is often an early finding in gram-negative septicemia, before fever, hypoxemia, or hypotension develops.

The diagnosis of respiratory alkalosis depends on measurement of arterial pH and $Paco_2$. The plasma $[K^+]$ is often reduced and the $[Cl^-]$ increased. In the acute phase, respiratory alkalosis is not associated with increased renal HCO_3^- excretion, but within hours net acid excretion is reduced. In general, the HCO_3^- concentration falls by 2.0 mmol/L for each 10-mmHg decrease in $Paco_2$. If the hypocapnia persists for >3–5 days, chronic respiratory alkalosis is present, and the decline in $Paco_2$ reduces the serum $[HCO_3^-]$ by 4–5 mmol/L for each 10-mmHg decrease in $Paco_2$. It is unusual to observe a plasma HCO_3^- <12 mmol/L as a result of a pure respiratory alkalosis. Moreover, the compensatory reduction in the plasma HCO_3^- concentration is so effective in chronic respiratory alkalosis that the pH does not decline significantly from the normal value. In this regard, chronic respiratory alkalosis is the only acid-base disorder that may return the pH to the normal value.

When a diagnosis of respiratory alkalosis is made, its cause should be investigated. The diagnosis of hyperventilation syndrome is made by exclusion. In difficult cases, it may be important to rule out other conditions such as pulmonary embolism, coronary artery disease, and hyperthyroidism.

TREATMENT

TREATMENT

Respiratory Alkalosis

The management of respiratory alkalosis is directed toward alleviation of the underlying disorder. If respiratory alkalosis complicates ventilator management, changes in dead space, tidal volume, and frequency can minimize the hypocapnia. Patients with the hyperventilation syndrome may benefit from reassurance, rebreathing from a paper bag during symptomatic attacks, and attention to underlying psychological stress. Antidepressants and sedatives are not recommended. β-adrenergic blockers may ameliorate peripheral manifestations of the hyperadrenergic state.

FURTHER READING

Berend K, et al: Physiological approach to assessment of acid-base disturbances. N Engl J Med 371:1434, 2014. [PubMed: 25295502]

DuBose TD: Disorders of acid-base balance. In *Brenner and Rector's The Kidney,* 10th ed. Skorecki K, et al. (eds). Philadelphia, Elsevier, 2016, pp. 511–558.

DuBose TD: Etiologic causes of metabolic acidosis I: The high anion gap acidosis, In *Metabolic Acidosis*. Wesson DE (ed). New York, Springer, 2016, pp. 17–26.

DuBose TD: Etiologic causes of metabolic acidosis II: The normal anion gap acidosis, In *Metabolic Acidosis*. Wesson DE (ed). New York, Springer, 2016, pp. 27–38.

Kurtz I, et al: Acid-base analysis: A critique of the Stewart and bicarbonate-centered approaches. Am J Physiol Renal Physiol 294: F1009, 2008. [PubMed: 18184741]

Palmer BF, Clegg DJ: Electrolyte and acid–base disturbances in patients with diabetes mellitus. N Engl J Med 373:548, 2015. [PubMed: 26244308]

Chapter 52: Approach to the Patient with a Skin Disorder

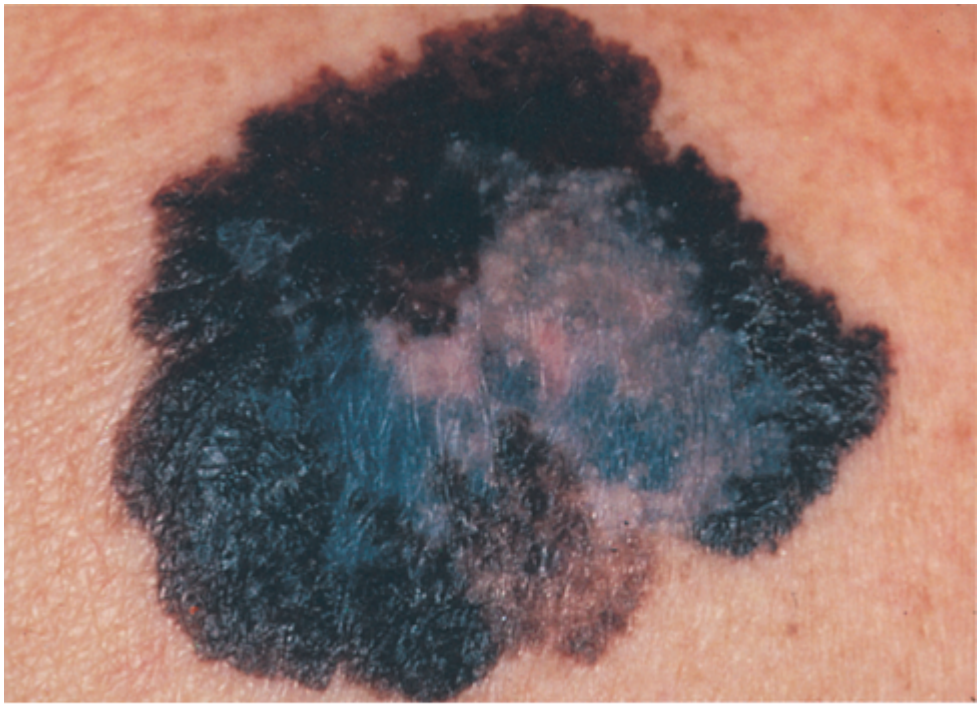
Kim B. Yancey; Thomas J. Lawley

INTRODUCTION

The challenge of examining the skin lies in distinguishing normal from abnormal findings, distinguishing significant findings from trivial ones, and integrating pertinent signs and symptoms into an appropriate differential diagnosis. The fact that the largest organ in the body is visible is both an advantage and a disadvantage to those who examine it. It is advantageous because no special instrumentation is necessary and because the skin can be biopsied with little morbidity. However, the casual observer can be misled by a variety of stimuli and overlook important, subtle signs of skin or systemic disease. For instance, the sometimes minor differences in color and shape that distinguish a melanoma (Fig. 52-1) from a benign nevomelanocytic nevus (Fig. 52-2) can be difficult to recognize. A variety of descriptive terms have been developed that characterize cutaneous lesions (Tables 52-1, 52-2, and 52-3; Fig. 52-3), thereby aiding in their interpretation and in the formulation of a differential diagnosis (Table 52-4). For example, the finding of scaling papules, which are present in psoriasis or atopic dermatitis, places the patient in a different diagnostic category than would hemorrhagic papules, which may indicate vasculitis or sepsis (Figs. 52-4 and 52-5, respectively). It is also important to differentiate primary from secondary skin lesions. If the examiner focuses on linear erosions overlying an area of erythema and scaling, he or she may incorrectly assume that the erosion is the primary lesion and that the redness and scale are secondary, whereas the correct interpretation would be that the patient has a pruritic eczematous dermatitis with erosions caused by scratching.

FIGURE 52-1

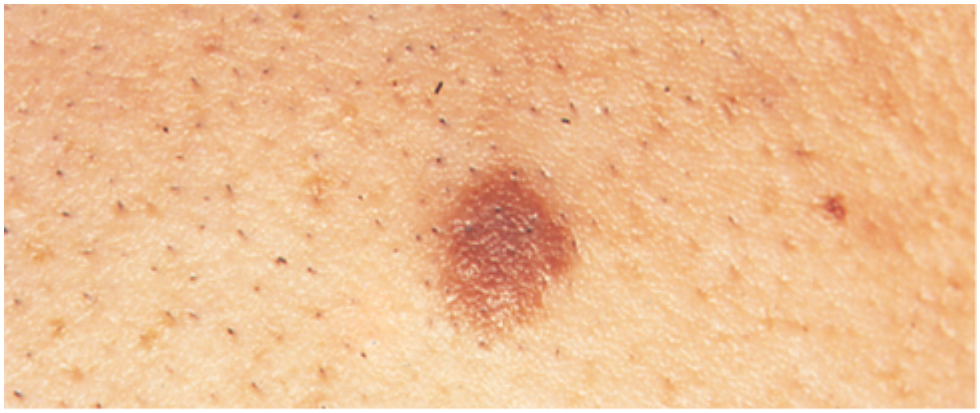
Superficial spreading melanoma. This is the most common type of melanoma. Such lesions usually demonstrate asymmetry, border irregularity, color variegation (black, blue, brown, pink, and white), a diameter >6 mm, and a history of change (e.g., an increase in size or development of associated symptoms such as pruritus or pain).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 52-2

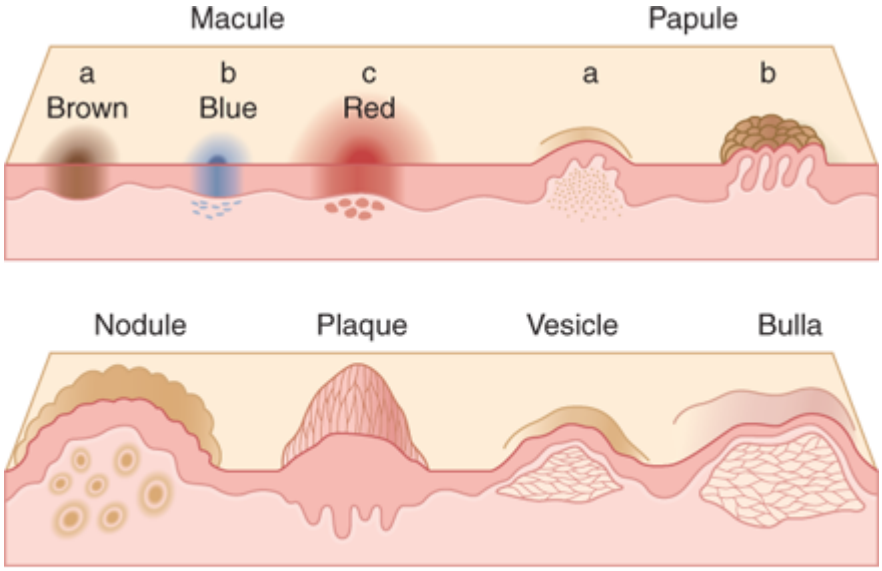
Nevomelanocytic nevus. Nevi are benign proliferations of nevomelanocytes characterized by regularly shaped hyperpigmented macules or papules of a uniform color.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 52-3

A schematic representation of several common primary skin lesions (see Table 52-1).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 52-4

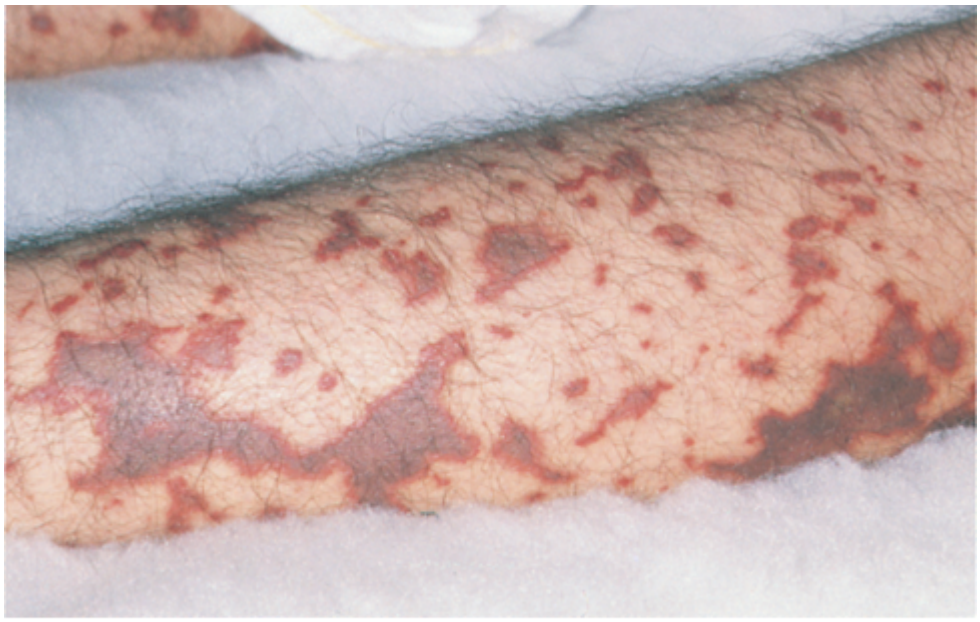
Necrotizing vasculitis. Palpable purpuric papules on the lower legs are seen in this patient with cutaneous small-vessel vasculitis. *(Courtesy of Robert Swerlick, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 52-5

Meningococcemia. An example of fulminant meningococcemia with extensive angular purpuric patches. *(Courtesy of Stephen E. Gellis, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 52-1

Description of Primary Skin Lesions

Macule: A flat, colored lesion, <2 cm in diameter, not raised above the surface of the surrounding skin. A “freckle,” or ephelid, is a prototypical pigmented macule.
Patch: A large (>2 cm) flat lesion with a color different from the surrounding skin. This differs from a macule only in size.
Papule: A small, solid lesion, <0.5 cm in diameter, raised above the surface of the surrounding skin and thus palpable (e.g., a closed comedone, or whitehead, in acne).
Nodule: A larger (0.5 to 5.0 cm), firm lesion raised above the surface of the surrounding skin. This differs from a papule only in size (e.g., a large dermal nevomelanocytic nevus).
Tumor: A solid, raised growth >5 cm in diameter.
Plaque: A large (>1 cm), flat-topped, raised lesion; edges may either be distinct (e.g., in psoriasis) or gradually blend with surrounding skin (e.g., in eczematous dermatitis).
Vesicle: A small, fluid-filled lesion, <0.5 cm in diameter, raised above the plane of surrounding skin. Fluid is often visible, and the lesions are translucent (e.g., vesicles in allergic contact dermatitis caused by *Toxicodendron* [poison ivy]).
Pustule: A vesicle filled with leukocytes. Note: The presence of pustules does not necessarily signify the existence of an infection.
Bulla: A fluid-filled, raised, often translucent lesion >0.5 cm in diameter.
Wheal: A raised, erythematous, edematous papule or plaque, usually representing short-lived vasodilation and vasopermeability.
Telangiectasia: A dilated, superficial blood vessel.

TABLE 52-2

Description of Secondary Skin Lesions

Lichenification: A distinctive thickening of the skin that is characterized by accentuated skin-fold markings.
Scale: Excessive accumulation of stratum corneum.
Crust: Dried exudate of body fluids that may be either yellow (i.e., serous crust) or red (i.e., hemorrhagic crust).
Erosion: Loss of epidermis without an associated loss of dermis.
Ulcer: Loss of epidermis and at least a portion of the underlying dermis.
Excoriation: Linear, angular erosions that may be covered by crust and are caused by scratching.
Atrophy: An acquired loss of substance. In the skin, this may appear as a depression with intact epidermis (i.e., loss of dermal or subcutaneous tissue) or as sites of shiny, delicate, wrinkled lesions (i.e., epidermal atrophy).
Scar: A change in the skin secondary to trauma or inflammation. Sites may be erythematous, hypopigmented, or hyperpigmented depending on their age or character. Sites on hair-bearing areas may be characterized by destruction of hair follicles.

TABLE 52-3

Common Dermatologic Terms

Alopecia: Hair loss, partial or complete.
Annular: Ring-shaped.
Cyst: A soft, raised, encapsulated lesion filled with semisolid or liquid contents.
Herpetiform: In a grouped configuration.
Lichenoid eruption: Violaceous to purple, polygonal lesions that resemble those seen in lichen planus.
Milia: Small, firm, white papules filled with keratin.
Morbilliform rash: Generalized, small erythematous macules and/or papules that resemble lesions seen in measles.
Nummular: Coin-shaped.
Poikiloderma: Skin that displays variegated pigmentation, atrophy, and telangiectases.
Polycyclic lesions: A configuration of skin lesions formed from coalescing rings or incomplete rings.
Pruritus: A sensation that elicits the desire to scratch. Pruritus is often the predominant symptom of inflammatory skin diseases (e.g., atopic dermatitis, allergic contact dermatitis); it is also commonly associated with xerosis and aged skin. Systemic conditions that can be associated with pruritus include chronic renal disease, cholestasis, pregnancy, malignancy, thyroid disease, polycythemia vera, and delusions of parasitosis.

TABLE 52-4

Selected Common Dermatologic Conditions

Diagnosis	Common Distribution	Usual Morphology	Diagnosis	Common Distribution	Usual Morphology
Acne vulgaris	Face, upper back, chest	Open and closed comedones, erythematous papules, pustules, cysts	Seborrheic keratosis	Trunk, face, extremities	Brown plaques with adherent, greasy scale; “stuck on” appearance
Rosacea	Blush area of cheeks, nose, forehead, chin	Erythema, telangiectases, papules, pustules	Folliculitis Impetigo	Any hair-bearing area Anywhere	Follicular pustules Papules, vesicles, pustules, often with honey-colored crusts
Seborrheic dermatitis	Scalp, eyebrows, perinasal areas	Erythema with greasy yellow-brown scale	Herpes simplex	Lips, genitalia	Grouped vesicles progressing to crusted erosions
Atopic dermatitis	Antecubital and popliteal fossae; may be widespread	Patches and plaques of erythema, scaling, and lichenification; pruritus	Herpes zoster	Dermatomal, usually trunk but may be anywhere	Vesicles limited to a dermatome (often painful)

Diagnosis	Common Distribution	Usual Morphology	Diagnosis	Common Distribution	Usual Morphology
Stasis dermatitis	Ankles, lower legs over medial malleoli	Patches of erythema and scaling on background of hyperpigmentation associated with signs of venous insufficiency	Varicella	Face, trunk, relative sparing of extremities	Lesions arise in crops and quickly progress from erythematous macules, to papules, to vesicles, to pustules, to crusted sites.
Dyshidrotic eczema	Palms, soles, sides of fingers and toes	Deep vesicles	Pityriasis rosea	Trunk (Christmas tree pattern); herald patch followed by multiple smaller lesions	Symmetric erythematous papules and plaques with a collarette of scale
Allergic contact dermatitis	Anywhere	Localized erythema, vesicles, scale, and pruritus (e.g., fingers, earlobes—nickel; dorsal aspect of foot—shoe; exposed surfaces—poison ivy)	Tinea versicolor	Chest, back, abdomen, proximal extremities	Scaly hyper- or hypopigmented macules
Psoriasis	Elbows, knees, scalp, lower back, fingernails (may be generalized)	Papules and plaques covered with silvery scale; nails have pits	Candidiasis	Groin, beneath breasts, vagina, oral cavity	Erythematous macerated areas with satellite pustules; white, friable patches on mucous membranes

Diagnosis	Common Distribution	Usual Morphology	Diagnosis	Common Distribution	Usual Morphology
Lichen planus	Wrists, ankles, mouth (may be widespread)	Violaceous flat-topped papules and plaques	Dermatophytosis	Feet, groin, beard, or scalp	Varies with site (e.g., tinea corporis—scaly annular plaque)
Keratosis pilaris	Extensor surfaces of arms and thighs, buttocks	Keratotic follicular papules with surrounding erythema	Scabies	Groin, axillae, between fingers and toes, beneath breasts	Excoriated papules, burrows, pruritus
Melasma	Forehead, cheeks, temples, upper lip	Tan to brown patches	Insect bites	Anywhere	Erythematous papules with central puncta
Vitiligo	Periorificial, trunk, extensor surfaces of extremities, flexor wrists, axillae	Chalk-white macules	Cherry angioma Keloid Dermatofibroma	Trunk	Red, blood-filled papules
				Anywhere (site of previous injury)	Firm tumor, pink, purple, or brown
				Anywhere	Firm red to brown nodule that shows dimpling of overlying skin with lateral compression

Diagnosis	Common Distribution	Usual Morphology	Diagnosis	Common Distribution	Usual Morphology
Actinic keratosis	Sun-exposed areas	Skin-colored or red-brown macule or papule with dry, rough, adherent scale	Acrochordons (skin tags)	Groin, axilla, neck	Fleshy papules
Basal cell carcinoma	Face	Papule with pearly, telangiectatic border on sun-damaged skin	Urticaria	Anywhere	Wheals, sometimes with surrounding flare; pruritus
Squamous cell carcinoma	Face, especially lower lip, ears	Indurated and possibly hyperkeratotic lesions often showing ulceration and/or crusting	Transient acantholytic dermatosis	Trunk, especially anterior chest	Erythematous papules
			Xerosis	Extensor extremities, especially legs	Dry, erythematous, scaling patches; pruritus

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Skin Disorder

In examining the skin it is usually advisable to assess the patient before taking an extensive history. This approach ensures that the entire cutaneous surface will be evaluated, and objective findings can be integrated with relevant historical data. Four basic features of a skin lesion must be noted and considered during a physical examination: the *distribution* of the eruption, the *types* of primary and secondary lesions, the *shape* of individual lesions, and the *arrangement* of the lesions. An ideal skin examination includes evaluation of the skin, hair, and nails as well as the mucous membranes of the mouth, eyes, nose, nasopharynx, and anogenital region. In the initial examination, it is important that the patient be disrobed as completely as possible to minimize chances of missing important individual skin lesions and permit accurate assessment of the distribution of the eruption. The patient should first be viewed from a distance of about 1.5–2 m (4–6 ft) so that the general character of the skin and the distribution of lesions can be evaluated.

Indeed, the distribution of lesions often correlates highly with diagnosis **(Fig. 52-6)**. For example, a hospitalized patient with a generalized erythematous exanthem is more likely to have a drug eruption than is a patient with a similar rash limited to the sun-exposed portions of the face. Once the distribution of the lesions has been established, the nature of the primary lesion must be determined. Thus, when lesions are distributed on elbows, knees, and scalp, the most likely possibility based solely on distribution is psoriasis or dermatitis herpetiformis **(Figs. 52-7 and 52-8, respectively)**. The primary lesion in psoriasis is a scaly papule that soon forms erythematous plaques covered with a white scale, whereas that of dermatitis herpetiformis is an urticarial papule that quickly becomes a small vesicle. In this manner, identification of the primary lesion directs the examiner toward the proper diagnosis. Secondary changes in skin can also be quite helpful. For example, scale represents excessive epidermis, while crust is the result of a discontinuous epithelial cell layer. Palpation of skin lesions can yield insight into the character of an eruption. Thus, red papules on the lower extremities that blanch with pressure can be a manifestation of many different diseases, but hemorrhagic red papules that do not blanch with pressure indicate palpable purpura characteristic of necrotizing vasculitis (Fig. 52-4).

The shape of lesions is also an important feature. Flat, round, erythematous papules and plaques are common in many cutaneous diseases. However, target-shaped lesions that consist in part of erythematous plaques are specific for erythema multiforme **(Fig. 52-9)**. Likewise, the arrangement of individual lesions is important. Erythematous papules and vesicles can occur in many conditions, but their arrangement in a specific linear array suggests an external etiology such as allergic contact dermatitis **(Fig. 52-10)** or primary irritant dermatitis. In contrast, lesions with a generalized arrangement are common and suggest a systemic etiology.

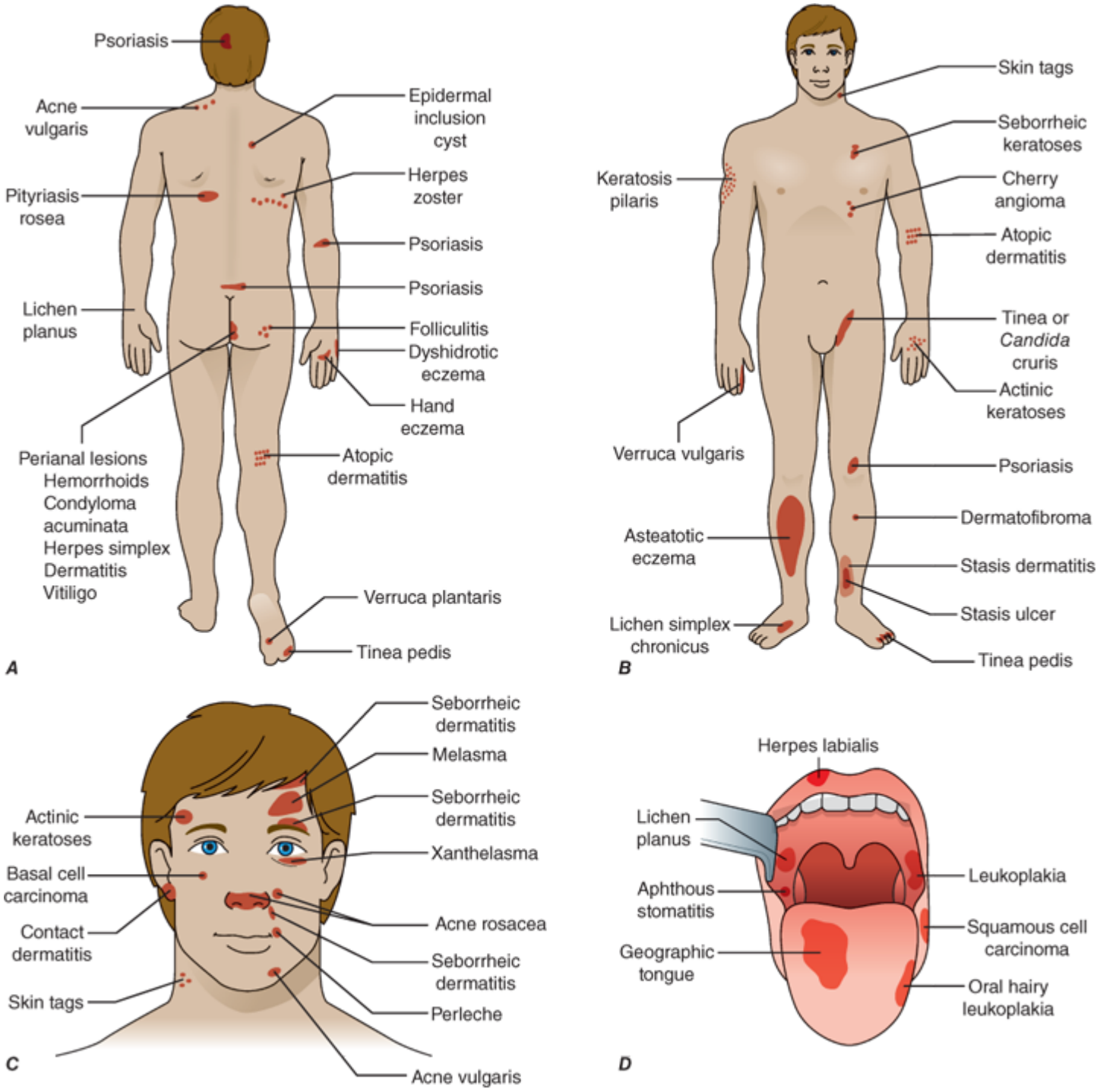
As in other branches of medicine, a complete history should be obtained to emphasize the following features:

1. Evolution of lesions
    a. Site of onset
    b. Manner in which the eruption progressed or spread
    c. Duration
    d. Periods of resolution or improvement in chronic eruptions
2. Symptoms associated with the eruption
    a. Itching, burning, pain, numbness
    b. What, if anything, has relieved symptoms
    c. Time of day when symptoms are most severe

3. Current or recent medications (prescribed as well as over-the-counter)
4. Associated systemic symptoms (e.g., malaise, fever, arthralgias)
5. Ongoing or previous illnesses
6. History of allergies
7. Presence of photosensitivity
8. Review of systems
9. Family history (particularly relevant for patients with melanoma, atopy, psoriasis, or acne)
10. Social, sexual, or travel history

FIGURE 52-6

Distribution of some common dermatologic diseases and lesions.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 52-7

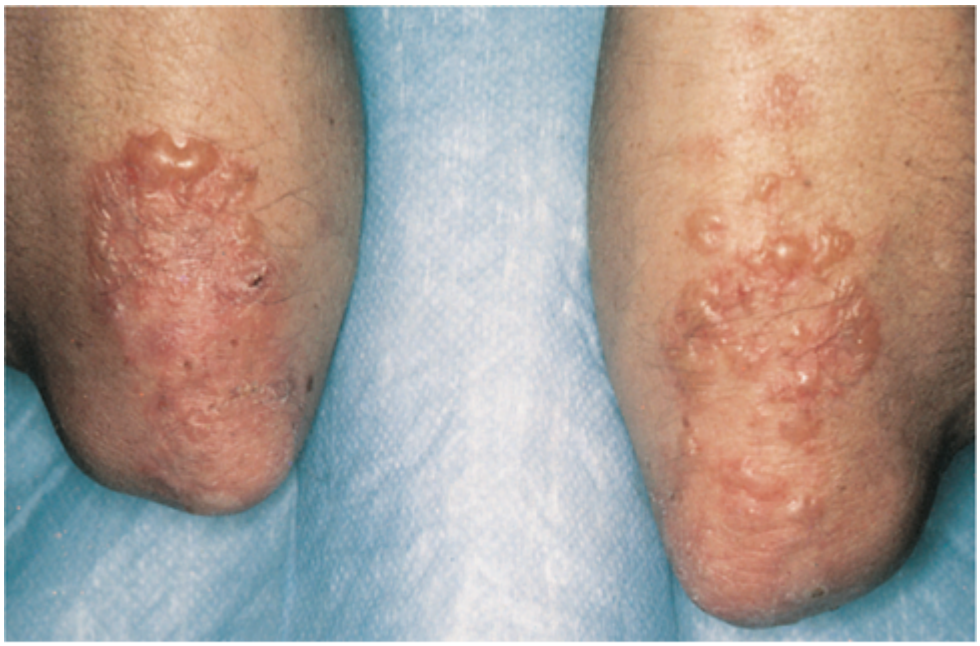
Psoriasis. This papulosquamous skin disease is characterized by small and large erythematous papules and plaques with overlying adherent silvery scale.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 52-8

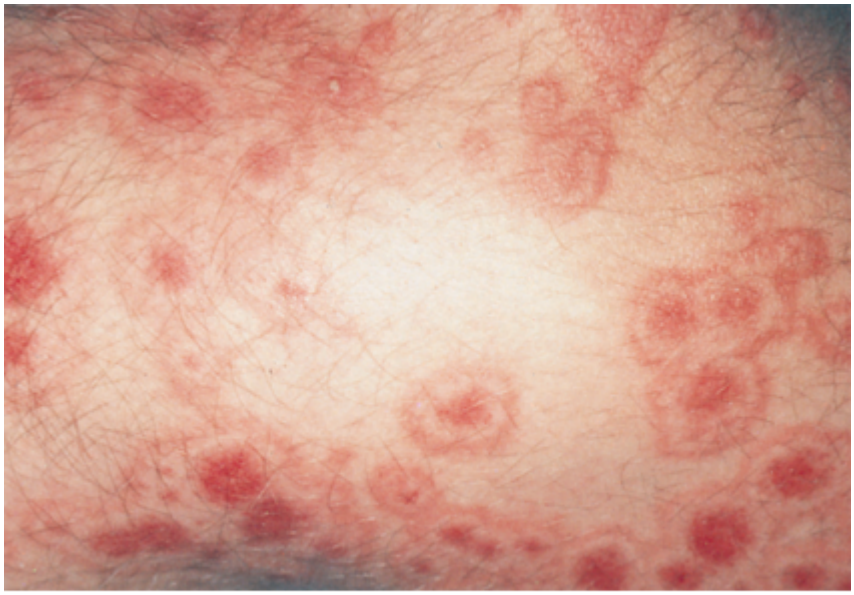
Dermatitis herpetiformis. This disorder typically displays pruritic, grouped papulovesicles on elbows, knees, buttocks, and posterior scalp. Vesicles are often excoriated due to associated pruritus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 52-9

Erythema multiforme. This eruption is characterized by multiple erythematous plaques with a target or iris morphology. It usually represents a hypersensitivity reaction to drugs (e.g., sulfonamides) or infections (e.g., HSV). *(Courtesy of the Yale Resident's Slide Collection; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 52-10

Allergic contact dermatitis (ACD). ***A.*** An example of ACD in its acute phase, with sharply demarcated, weeping, eczematous plaques in a perioral distribution. ***B.*** ACD in its chronic phase, with an erythematous, lichenified, weeping plaque on skin chronically exposed to nickel in a metal snap. *(B, Courtesy of Robert Swerlick, MD; with permission.)*





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSTIC TECHNIQUES

Many skin diseases can be diagnosed on the basis of gross clinical appearance, but sometimes relatively simple diagnostic procedures can yield valuable information. In most instances, they can be performed at the bedside with a minimum of equipment.

Skin Biopsy

A skin biopsy is a straightforward minor surgical procedure; however, it is important to biopsy a lesion that is most likely to yield diagnostic findings. This decision may require expertise in skin diseases and knowledge of superficial anatomic structures in selected areas of the body. In this procedure, a small area of skin is anesthetized with 1% lidocaine with or without epinephrine. The skin lesion in question can be excised or saucerized with a scalpel or removed by punch biopsy. In the latter technique, a punch is pressed against the

surface of the skin and rotated with downward pressure until it penetrates to the subcutaneous tissue. The circular biopsy is then lifted with forceps, and the bottom is cut with iris scissors. Biopsy sites may or may not need suture closure, depending on size and location.

KOH Preparation

A potassium hydroxide (KOH) preparation is performed on scaling skin lesions where a fungal infection is suspected. The edge of such a lesion is scraped gently with a no. 15 scalpel blade. The removed scale is collected on a glass microscope slide and then treated with 1 or 2 drops of a solution of 10–20% KOH. KOH dissolves keratin and allows easier visualization of fungal elements. Brief heating of the slide accelerates dissolution of keratin. When the preparation is viewed under the microscope, the refractile hyphae are seen more easily when the light intensity is reduced and the condenser is lowered. This technique can be used to identify hyphae in dermatophyte infections, pseudohyphae and budding yeasts in *Candida* infections, and "spaghetti and meatballs" yeast forms in tinea versicolor. The same sampling technique can be used to obtain scale for culture of selected pathogenic organisms.

Tzanck Smear

A Tzanck smear is a cytologic technique most often used in the diagnosis of herpesvirus infections (herpes simplex virus [HSV] or varicella zoster virus [VZV]) **(see Figs. 188-1 and 188-3)**. An early vesicle, not a pustule or crusted lesion, is unroofed, and the base of the lesion is scraped gently with a scalpel blade. The material is placed on a glass slide, air-dried, and stained with Giemsa or Wright's stain. Multinucleated epithelial giant cells suggest the presence of HSV or VZV; culture, immunofluorescence microscopy, or genetic testing must be performed to identify the specific virus.

Diascopy

Diascopy is designed to assess whether a skin lesion will blanch with pressure as, for example, in determining whether a red lesion is hemorrhagic or simply blood-filled. Urticaria **(Fig. 52-11)** will blanch with pressure, whereas a purpuric lesion caused by necrotizing vasculitis (Fig. 52-4) will not. Diascopy is performed by pressing a microscope slide or magnifying lens against a lesion and noting the amount of blanching that occurs. Granulomas often have an opaque to transparent, brown-pink "apple jelly" appearance on diascopy.

FIGURE 52-11

Urticaria. Discrete and confluent, edematous, erythematous papules and plaques are characteristic of this whealing eruption.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Wood's Light

A Wood's lamp generates 360-nm ultraviolet ("black") light that can be used to aid the evaluation of certain skin disorders. For example, a Wood's lamp will cause erythrasma (a superficial, intertriginous infection caused by *Corynebacterium minutissimum*) to show a characteristic coral pink color, and wounds colonized by *Pseudomonas* will appear pale blue. Tinea capitis caused by certain dermatophytes (e.g., *Microsporum canis* or *M. audouinii*) exhibits a yellow fluorescence. Pigmented lesions of the epidermis such as freckles are accentuated, while dermal pigment such as postinflammatory hyperpigmentation fades under a Wood's light. Vitiligo **(Fig. 52-12)** appears totally white under a Wood's lamp, and previously unsuspected areas of involvement often become apparent. A Wood's lamp may also aid in the demonstration of tinea versicolor, sites of depigmentation within and/or surrounding melanomas, and in recognition of ash leaf spots in patients with tuberous sclerosis.

FIGURE 52-12

Vitiligo. Characteristic lesions display an acral distribution and striking depigmentation as a result of loss of melanocytes.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Patch Tests

Patch testing is designed to document sensitivity to a specific antigen. In this procedure, a battery of suspected allergens is applied to the patient's back under occlusive dressings and allowed to remain in contact with the skin for 48 h. The dressings are removed, and the area is examined for evidence of delayed hypersensitivity reactions (e.g., erythema, edema, or papulovesicles). This test is best performed by physicians with special expertise in patch testing and is often helpful in the evaluation of patients with chronic dermatitis.

FURTHER READING

Bolognia JL et al (eds): *Dermatology*, 4th ed. Philadelphia, Elsevier, 2018.

Goldsmith LA et al (eds): *Fitzpatrick's Dermatology in General Medicine*, 8th ed. New York, McGraw-Hill, 2012.

James WD: *Andrews' Diseases of the Skin: Clinical Dermatology*, 12th ed. Philadelphia, Elsevier, 2016.

Chapter 53: Eczema, Psoriasis, Cutaneous Infections, Acne, and Other Common Skin Disorders

Leslie P. Lawley; Calvin O. McCall; Thomas J. Lawley

ECZEMA AND DERMATITIS

Eczema is a type of dermatitis, and these terms are often used synonymously (e.g., atopic eczema or atopic dermatitis [AD]). Eczema is a reaction pattern that presents with variable clinical findings and the common histologic finding of *spongiosis* (intercellular edema of the epidermis). Eczema is the final common expression for a number of disorders, including those discussed in the following sections. Primary lesions may include erythematous macules, papules, and vesicles, which can coalesce to form patches and plaques. In severe eczema, secondary lesions from infection or excoriation, marked by weeping and crusting, may predominate. In chronic eczematous conditions, *lichenification* (cutaneous hypertrophy and accentuation of normal skin markings) may alter the characteristic appearance of eczema.

ATOPIC DERMATITIS

AD is the cutaneous expression of the atopic state, characterized by a family history of asthma, allergic rhinitis, or eczema. The prevalence of AD is increasing worldwide. Some of its features are shown in Table 53-1.

TABLE 53-1

Clinical Features of Atopic Dermatitis

1. Pruritus and scratching
2. Course marked by exacerbations and remissions
3. Lesions typical of eczematous dermatitis
4. Personal or family history of atopy (asthma, allergic rhinitis, food allergies, or eczema)
5. Clinical course lasting >6 weeks
6. Lichenification of skin
7. Presence of dry skin



The etiology of AD is only partially defined, but there is a clear genetic predisposition. When both parents are affected by AD, >80% of their children manifest the disease. When only one parent is affected, the prevalence drops to slightly >50%. A characteristic defect in AD that contributes to the pathophysiology is an impaired epidermal barrier. In many patients, a mutation in the gene encoding filaggrin, a structural protein in the stratum corneum, is responsible. Patients with AD may display a variety of immunoregulatory abnormalities, including increased IgE synthesis; increased serum IgE levels; and impaired, delayed-type hypersensitivity reactions.

The clinical presentation often varies with age. Half of patients with AD present within the first year of life, and 80% present by 5 years of age. About 80% ultimately coexpress allergic rhinitis or asthma. The infantile pattern is characterized by weeping inflammatory patches and crusted plaques on the face, neck, and extensor surfaces. The childhood and adolescent pattern is typified by dermatitis of flexural skin, particularly in the antecubital and popliteal fossae **(Fig. 53-1)**. AD may resolve spontaneously, but approximately 40% of all individuals affected as children will have dermatitis in adult life. The distribution of lesions in adults may be similar to those seen in childhood; however, adults frequently have localized disease manifesting as lichen simplex chronicus or hand eczema (see below). In patients with localized disease, AD may be suspected because of a typical personal or family history or the presence of cutaneous stigmata of AD such as perioral pallor, an extra fold of skin beneath the lower eyelid (Dennie-Morgan folds), increased palmar skin markings, and an increased incidence of cutaneous infections, particularly with *Staphylococcus aureus*. Regardless of other manifestations, pruritus is a prominent characteristic of AD in all age groups and is exacerbated by dry skin. Many of the cutaneous findings in affected patients, such as lichenification, are secondary to rubbing and scratching.

FIGURE 53-1

Atopic dermatitis. Hyperpigmentation, lichenification, and scaling in the antecubital fossae are seen in this patient with atopic dermatitis. *(Courtesy of Robert Swerlick, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Atopic Dermatitis

Therapy for AD should include avoidance of cutaneous irritants, adequate moisturizing through the application of emollients, judicious use of topical anti-inflammatory agents, and prompt treatment of secondary infection. Patients should be instructed to bathe no more often than daily, using warm or cool water, and to use only mild bath soap. Immediately after bathing, while the skin is still moist, a topical anti-inflammatory agent in a cream or ointment base should be applied to areas of dermatitis, and all other skin areas should be lubricated with a moisturizer. Approximately 30 g of a topical agent is required to cover the entire body surface of an average adult.

Low- to mid-potency topical glucocorticoids are employed in most treatment regimens for AD. Skin atrophy and the potential for systemic absorption are constant concerns, especially with more potent agents. Low-potency topical glucocorticoids or nonglucocorticoid anti-inflammatory agents should be selected for use on the face and in intertriginous areas to minimize the risk of skin atrophy. Two nonglucocorticoid anti-inflammatory agents are available: tacrolimus ointment and pimecrolimus cream. These agents are macrolide immunosuppressants that are approved by the U.S. Food and Drug Administration (FDA) for topical use in AD. Reports of broader effectiveness appear in the literature. These agents do not cause skin atrophy, nor do they suppress the hypothalamic-pituitary-adrenal axis. However, concerns have emerged regarding the potential for lymphomas in patients treated with these agents. Thus, caution should be exercised when these agents are considered. Currently, they are also more costly than topical

glucocorticoids. Barrier-repair products that attempt to restore the impaired epidermal barrier are also nonglucocorticoid agents and are gaining popularity in the treatment of AD.

Secondary infection of eczematous skin may lead to exacerbation of AD. Crusted and weeping skin lesions may be infected with *S. aureus*. When secondary infection is suspected, eczematous lesions should be cultured and patients treated with systemic antibiotics active against *S. aureus*. The initial use of penicillinase-resistant penicillins or cephalosporins is preferable. Dicloxacillin or cephalexin (250 mg qid for 7–10 days) is generally adequate for adults; however, antibiotic selection must be directed by culture results and clinical response. More than 50% of *S. aureus* isolates are now methicillin resistant in some communities. Current recommendations for the treatment of infection with these community-acquired methicillin-resistant *S. aureus* (CA-MRSA) strains in adults include trimethoprim-sulfamethoxazole (one double-strength tablet bid), minocycline (100 mg bid), doxycycline (100 mg bid), or clindamycin (300–450 mg qid). Duration of therapy should be 7–10 days. Inducible resistance may limit clindamycin's usefulness. Such resistance can be detected by the double-disk diffusion test, which should be ordered if the isolate is erythromycin resistant and clindamycin sensitive. As an adjunct, antibacterial washes or dilute sodium hypochlorite baths (0.005% bleach) and intermittent nasal mupirocin may be useful.

Control of pruritus is essential for treatment, because AD often represents "an itch that rashes." Antihistamines are most often used to control pruritus. Diphenhydramine (25 mg every 4–6 h), hydroxyzine (10–25 mg every 6 h), or doxepin (10–25 mg at bedtime) are useful primarily due to their sedating action. Higher doses of these agents may be required, but sedation can become bothersome. Patients need to be counseled about driving or operating heavy equipment after taking these medications. When used at bedtime, sedating antihistamines may improve the patient's sleep. Although they are effective in urticaria, nonsedating antihistamines and selective H_2 blockers are of little use in controlling the pruritus of AD.

Treatment with systemic glucocorticoids should be limited to severe exacerbations unresponsive to topical therapy. In the patient with chronic AD, therapy with systemic glucocorticoids will generally clear the skin only briefly, and cessation of the systemic therapy will invariably be accompanied by a return, if not a worsening, of the dermatitis. Patients who do not respond to conventional therapies should be considered for patch testing to rule out allergic contact dermatitis (ACD). The role of dietary allergens in AD is controversial, and there is little evidence that they play any role outside of infancy, during which a small percentage of patients with AD may be affected by food allergens.

LICHEN SIMPLEX CHRONICUS

Lichen simplex chronicus may represent the end stage of a variety of pruritic and eczematous disorders, including AD. It consists of a circumscribed plaque or plaques of lichenified skin due to chronic scratching or rubbing. Common areas involved include the posterior nuchal region, dorsum of the feet, and ankles. Treatment of lichen simplex chronicus centers on breaking the cycle of chronic itching and scratching. High-potency topical glucocorticoids are helpful in most cases, but, in recalcitrant cases, application of topical glucocorticoids under occlusion or intralesional injection of glucocorticoids may be required.

CONTACT DERMATITIS

Contact dermatitis is an inflammatory skin process caused by an exogenous agent or agents that directly or indirectly injure the skin. In *irritant* contact dermatitis (ICD), this injury is caused by an inherent characteristic of a compound—for example, a concentrated acid or base. Agents that cause *allergic* contact dermatitis (ACD) induce an antigen-specific immune response (e.g., poison ivy dermatitis). The clinical lesions of contact dermatitis may be acute (wet and edematous) or chronic (dry, thickened, and scaly), depending on the persistence of the insult **(see Chap. 52, Fig. 52-10)**.

Irritant Contact Dermatitis

ICD is generally well demarcated and often localized to areas of thin skin (eyelids, intertriginous areas) or areas where the irritant was occluded. Lesions may range from minimal skin erythema to areas of marked edema, vesicles, and ulcers. Prior exposure to the offending agent is not necessary, and the reaction develops in minutes to a few hours. Chronic low-grade irritant dermatitis is the most common type of ICD, and the most common area of involvement is the hands (see below). The most common irritants encountered are chronic wet work, soaps, and detergents. Treatment should be directed toward the avoidance of irritants and the use of protective gloves or clothing.

Allergic Contact Dermatitis

ACD is a manifestation of delayed-type hypersensitivity mediated by memory T lymphocytes in the skin. Prior exposure to the offending agent is necessary to develop the hypersensitivity reaction, which may take as little as 12 h or as much as 72 h to develop. The most common cause of ACD is exposure to plants, especially to members of the family Anacardiaceae, including the genus *Toxicodendron*. Poison ivy, poison oak, and poison sumac are members of this genus and cause an allergic reaction marked by erythema, vesiculation, and severe pruritus. The eruption is often linear or angular, corresponding to areas where plants have touched the skin. The sensitizing antigen common to these plants is urushiol, an oleoresin containing the active ingredient pentadecylcatechol. The oleoresin may adhere to skin, clothing, tools, and pets, and contaminated articles may cause dermatitis even after prolonged storage. Blister fluid does not contain urushiol and is not capable of inducing skin eruption in exposed subjects.

TREATMENT

TREATMENT

Contact Dermatitis

If contact dermatitis is suspected and an offending agent is identified and removed, the eruption will resolve. Usually, treatment with high-potency topical glucocorticoids is enough to relieve symptoms while the dermatitis runs its course. For those patients who require systemic therapy, daily oral prednisone—beginning at 1 mg/kg, but usually ≤60 mg/d—is sufficient. The dose should be tapered over 2–3 weeks, and each daily dose should be taken in the morning with food.

Identification of a contact allergen can be a difficult and time-consuming task. Allergic contact dermatitis should be suspected in patients with dermatitis unresponsive to conventional therapy or with an unusual and patterned distribution. Patients should be questioned carefully regarding occupational exposures and topical medications. Common sensitizers include preservatives in topical preparations, nickel sulfate, potassium dichromate, thimerosal, neomycin sulfate, fragrances, formaldehyde, and rubber-curing agents. Patch testing is helpful in identifying these agents but should not be attempted when patients have widespread active dermatitis or are taking systemic glucocorticoids.

HAND ECZEMA

Hand eczema is a very common, chronic skin disorder in which both exogenous and endogenous factors play important roles. It may be associated with other cutaneous disorders such as AD, and contact with various agents may be involved. Hand eczema represents a large proportion of cases of occupation-associated skin disease. Chronic, excessive exposure to water and detergents, harsh chemicals, or allergens may initiate or aggravate this disorder. It may present with dryness and cracking of the skin of the hands as well as with variable amounts of erythema and edema. Often, the dermatitis will begin under rings, where water and irritants are trapped. *Dyshidrotic* eczema, a variant of hand eczema, presents with multiple, intensely pruritic, small papules and vesicles on the thenar and hypothenar eminences and the sides of the fingers (Fig. 53-2). Lesions tend to occur in crops that slowly form crusts and then heal.

FIGURE 53-2

Dyshidrotic eczema. This example is characterized by deep-seated vesicles and scaling on palms and lateral fingers, and the disease is often associated with an atopic diathesis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The evaluation of a patient with hand eczema should include an assessment of potential occupation-associated exposures. The history should be directed to identifying possible irritant or allergen exposures.

TREATMENT

TREATMENT

Hand Eczema

Therapy for hand eczema is directed toward avoidance of irritants, identification of possible contact allergens, treatment of coexistent infection, and application of topical glucocorticoids. Whenever possible, the hands should be protected by gloves, preferably vinyl. The use of rubber gloves (latex) to protect dermatitic skin is sometimes associated with the development of hypersensitivity reactions to components of the gloves, which could be a type I hypersensitivity reaction to the latex manifested by the development of hives, itching, angioedema, and possibly anaphylaxis within minutes to hours of exposure or a type IV hypersensitivity reaction to rubber accelerators with worsening of eczematous eruptions days after exposure. Patients can be treated with cool moist compresses followed by application of a mid- to high-potency topical glucocorticoid in a cream or ointment base. As in AD, treatment of secondary infection is essential for good control. In addition, patients with hand eczema should be examined for dermatophyte infection by potassium hydroxide (KOH) preparation and culture (see below).

NUMMULAR ECZEMA

Nummular eczema is characterized by circular or oval “coinlike” lesions, beginning as small edematous papules that become crusted and scaly. The etiology of nummular eczema is unknown, but dry skin is a contributing factor. Common locations are the trunk or the extensor surfaces of the extremities, particularly on the pretibial areas or dorsum of the hands. Nummular eczema occurs more frequently in men and is most common in middle age. The treatment of nummular eczema is similar to that for AD.

ASTEATOTIC ECZEMA

Asteatotic eczema, also known as *xerotic eczema* or “winter itch,” is a mildly inflammatory dermatitis that develops in areas of extremely dry skin, especially during the dry winter months. Clinically, there may be considerable overlap with nummular eczema. This form of eczema accounts for a large number of physician visits because of the associated pruritus. Fine cracks and scale, with or without erythema, characteristically develop in areas of dry skin, especially on the anterior surfaces of the lower extremities in elderly patients. Asteatotic eczema responds well to topical moisturizers and the avoidance of cutaneous irritants. Overbathing and the use of harsh soaps exacerbate asteatotic eczema.

STASIS DERMATITIS AND STASIS ULCERATION

Stasis dermatitis develops on the lower extremities secondary to venous incompetence and chronic edema. Patients may give a history of deep venous thrombosis and may have evidence of vein removal or varicose veins. Early findings in stasis dermatitis consist of mild erythema and scaling associated with pruritus. The typical initial site of involvement is the medial aspect of the ankle, often over a distended vein **(Fig. 53-3)**.

FIGURE 53-3

Stasis dermatitis. An example of stasis dermatitis showing erythematous, scaly, and oozing patches over the lower leg. Several stasis ulcers are also seen in this patient.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Stasis dermatitis may become acutely inflamed, with crusting and exudate. In this state, it is easily confused with cellulitis. Of note, symmetrical and bilateral involvement is more likely stasis dermatitis whereas unilateral involvement may represent cellulitis. Chronic stasis dermatitis is often associated with dermal fibrosis that is recognized clinically as brawny edema of the skin. As the disorder progresses, the dermatitis becomes progressively pigmented due to chronic erythrocyte extravasation leading to cutaneous hemosiderin deposition. Stasis dermatitis may be complicated by secondary infection and contact dermatitis. Severe stasis dermatitis may precede the development of stasis ulcers.

TREATMENT

TREATMENT

Stasis Dermatitis and Stasis Ulceration

Patients with stasis dermatitis and stasis ulceration benefit greatly from leg elevation and the routine use of compression stockings with a gradient of at least 30–40 mmHg. Stockings providing less compression, such as antiembolism hose, are poor substitutes. Use of emollients and/or mid-potency topical glucocorticoids and avoidance of irritants are also helpful in treating stasis dermatitis. Protection of the legs from injury, including scratching, and control of chronic edema are essential to prevent ulcers. Diuretics may be required to adequately control chronic edema.

Stasis ulcers are difficult to treat, and resolution is slow. It is extremely important to elevate the affected limb as much as possible. The ulcer should be kept clear of necrotic material by gentle debridement and covered with a semipermeable dressing and a compression dressing or compression stocking. Glucocorticoids should not be applied to ulcers, because they may retard healing; however, they may be applied to the surrounding

skin to control itching, scratching, and additional trauma. Secondarily infected lesions should be treated with appropriate oral antibiotics, but it should be noted that all ulcers will become colonized with bacteria, and the purpose of antibiotic therapy should not be to clear all bacterial growth. Care must be taken to exclude treatable causes of leg ulcers (hypercoagulation, vasculitis) before beginning the chronic management outlined above.

SEBORRHEIC DERMATITIS

Seborrheic dermatitis is a common, chronic disorder characterized by greasy scales overlying erythematous patches or plaques. Induration and scale are generally less prominent than in psoriasis, but clinical overlap exists between these diseases ("sebopsoriasis"). The most common location is in the scalp, where it may be recognized as severe dandruff. On the face, seborrheic dermatitis affects the eyebrows, eyelids, glabella, and nasolabial folds **(Fig. 53-4)**. Scaling of the external auditory canal is common in seborrheic dermatitis. In addition, the postauricular areas often become macerated and tender. Seborrheic dermatitis may also develop in the central chest, axilla, groin, submammary folds, and gluteal cleft. Rarely, it may cause widespread generalized dermatitis. Pruritus is variable.

FIGURE 53-4

Seborrheic dermatitis. Central facial erythema with overlying greasy, yellowish scale is seen in this patient. *(Courtesy of Jean Bolognia, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Seborrheic dermatitis may be evident within the first few weeks of life, and within this context it typically occurs in the scalp ("cradle cap"), face, or groin. It is rarely seen in children beyond infancy but becomes

evident again during adolescent and adult life. Although it is frequently seen in patients with Parkinson's disease, in those who have had cerebrovascular accidents, and in those with HIV infection, the overwhelming majority of individuals with seborrheic dermatitis have no underlying disorder.

TREATMENT

TREATMENT

Seborrheic Dermatitis

Treatment with low-potency topical glucocorticoids in conjunction with a topical antifungal agent, such as ketoconazole cream or ciclopirox cream, is often effective. The scalp and beard areas may benefit from antidandruff shampoos, which should be left in place 3–5 min before rinsing. High-potency topical glucocorticoid solutions (betamethasone or clobetasol) are effective for control of severe scalp involvement. High-potency glucocorticoids should not be used on the face because this treatment is often associated with steroid-induced rosacea or atrophy.

PAPULOSQUAMOUS DISORDERS

(TABLE 53-2)

TABLE 53-2

Papulosquamous Disorders

	Clinical Features	Other Notable Features	Histologic Features
Psoriasis	Sharply demarcated, erythematous plaques with mica-like scale; predominantly on elbows, knees, and scalp; atypical forms may localize to intertriginous areas; eruptive forms may be associated with infection	May be aggravated by certain drugs, infection; severe forms seen in association with HIV	Acanthosis, vascular proliferation
Lichen planus	Purple polygonal papules marked by severe pruritus; lacy white markings, especially associated with mucous membrane lesions	Certain drugs may induce: thiazides, antimalarial drugs	Interface dermatitis
Pityriasis rosea	Rash often preceded by herald patch; oval to round plaques with trailing scale; most often affects trunk; eruption lines up in skinfolds giving a “fir tree–like” appearance; generally spares palms and soles	Variable pruritus; self-limited, resolving in 2–8 weeks; may be imitated by secondary syphilis	Pathologic features often nonspecific
Dermatophytosis	Polymorphous appearance depending on dermatophyte, body site, and host response; sharply defined to ill-demarcated scaly plaques with or without inflammation; may be associated with hair loss	KOH preparation may show branching hyphae; culture helpful	Hyphae and neutrophils in stratum corneum

Abbreviations: HIV, human immunodeficiency virus; KOH, potassium hydroxide.

PSORIASIS

Psoriasis is one of the most common dermatologic diseases, affecting up to 2% of the world’s population. It is an immune-mediated disease clinically characterized by erythematous, sharply demarcated papules and rounded plaques covered by silvery micaceous scale. The skin lesions of psoriasis are variably pruritic. Traumatized areas often develop lesions of psoriasis (the *Koebner* or isomorphic phenomenon). In addition, other external factors may exacerbate psoriasis, including infections, stress, and medications (lithium, beta blockers, and antimalarial drugs).

The most common variety of psoriasis is called *plaque-type*. Patients with plaque-type psoriasis have stable, slowly enlarging plaques, which remain basically unchanged for long periods of time. The most commonly involved areas are the elbows, knees, gluteal cleft, and scalp. Involvement tends to be symmetric. Plaque psoriasis generally develops slowly and runs an indolent course. It rarely remits spontaneously. *Inverse psoriasis* affects the intertriginous regions, including the axilla, groin, submammary region, and navel; it also tends to affect the scalp, palms, and soles. The individual lesions are sharply demarcated plaques **(see Chap. 52, Fig. 52-7)**, but they may be moist and without scale due to their locations.

Guttate psoriasis (eruptive psoriasis) is most common in children and young adults. It develops acutely in individuals without psoriasis or in those with chronic plaque psoriasis. Patients present with many small erythematous, scaling papules, frequently after upper respiratory tract infection with β-hemolytic streptococci. The differential diagnosis should include pityriasis rosea and secondary syphilis.

In *pustular psoriasis*, patients may have disease localized to the palms and soles, or the disease may be generalized. Regardless of the extent of disease, the skin is erythematous, with pustules and variable scale. Localized to the palms and soles, it is easily confused with eczema. When it is generalized, episodes are characterized by fever (39°–40°C [102.2°–104.0°F]) lasting several days, an accompanying generalized eruption of sterile pustules, and a background of intense erythema; patients may become erythrodermic. Episodes of fever and pustules are recurrent. Local irritants, pregnancy, medications, infections, and systemic glucocorticoid withdrawal can precipitate this form of psoriasis. Oral retinoids are the treatment of choice in nonpregnant patients.

Fingernail involvement, appearing as punctate pitting, onycholysis, nail thickening, or subungual hyperkeratosis, may be a clue to the diagnosis of psoriasis when the clinical presentation is not classic.

According to the National Psoriasis Foundation, up to 30% of patients with psoriasis have psoriatic arthritis (PsA). It develops most commonly between the ages of 30 and 50 years. There are five subtypes of PsA: symmetric PsA, asymmetric PsA, distal PsA, spondylitis, and arthritis mutilans. Approximately 50% of PsA is classified as symmetric, which may resemble rheumatoid arthritis. Asymmetric arthritis comprises about 35% of cases. It can involve any joint and may present as “sausage digits.” Distal PsA is the classic form; however, it occurs in only about 5% of patients with PsA. It can involve fingers and toes; fingernails and toenails are often dystrophic, including nail pitting. Spondylitis also occurs in ~5% of patients with PsA. Arthritis mutilans is severe and deforming, and affects primarily the small joints of the hands and feet. It accounts for fewer than 5% of PsA cases.

An increased risk of metabolic syndrome, including increased morbidity and mortality from cardiovascular events, has been demonstrated in psoriasis patients. Appropriate screening tests should be performed. The etiology of psoriasis is still poorly understood, but there is clearly a genetic component to the disease. In various studies, 30–50% of patients with psoriasis report a positive family history. Psoriatic lesions contain infiltrates of activated T cells that are thought to elaborate cytokines responsible for keratinocyte hyperproliferation, which results in the characteristic clinical findings. Agents inhibiting T cell activation,

clonal expansion, or release of proinflammatory cytokines are often effective for the treatment of severe psoriasis (see below).

TREATMENT

TREATMENT

Psoriasis

Treatment of psoriasis depends on the type, location, and extent of disease. All patients should be instructed to avoid excess drying or irritation of their skin and to maintain adequate cutaneous hydration. Most cases of localized, plaque-type psoriasis can be managed with mid-potency topical glucocorticoids, although their long-term use is often accompanied by loss of effectiveness (tachyphylaxis) and atrophy of the skin. A topical vitamin D analogue (calcipotriene) and a retinoid (tazarotene) are also efficacious in the treatment of limited psoriasis and have largely replaced other topical agents such as coal tar, salicylic acid, and anthralin.

Ultraviolet (UV) light, natural or artificial, is an effective therapy for many patients with widespread psoriasis. Ultraviolet B (UVB), narrowband UVB, and ultraviolet A (UVA) light with either oral or topical psoralens (PUVA) are used clinically. UV light's immunosuppressive properties are thought to be responsible for its therapeutic activity in psoriasis. It is also mutagenic, potentially leading to an increased incidence of nonmelanoma and melanoma skin cancer. UV-light therapy is contraindicated in patients receiving cyclosporine and should be used with great care in all immunocompromised patients due to the increased risk of skin cancer.

Various systemic agents can be used for severe, widespread psoriatic disease **(Table 53-3)**. Oral glucocorticoids should not be used for the treatment of psoriasis due to the potential for development of life-threatening pustular psoriasis when therapy is discontinued. Methotrexate is an effective agent, especially in patients with PsA. The synthetic retinoid acitretin is useful, especially when immunosuppression must be avoided; however, teratogenicity limits its use. Apremilast is a new oral agent that inhibits phosphodiesterase type 4. It is approved for both psoriasis and PsA. It must be used cautiously in the presence of renal failure or depression.

The evidence implicating psoriasis as a T cell–mediated disorder has directed therapeutic efforts to immunoregulation. Cyclosporine and other immunosuppressive agents can be very effective in the treatment of psoriasis, and much attention is currently directed toward the development of biologic agents with more selective immunosuppressive properties and better safety profiles **(Table 53-4)**. Experience with some of these biologic agents is limited, and information regarding combination therapy and adverse events continues to emerge. These biologic agents appear to be quite efficacious in treatment of psoriasis and are well tolerated; however, caution with certain patient comorbidities must be exercised. Use of tumor necrosis factor-α (TNF-α) inhibitors may worsen congestive heart failure (CHF), and they should be used with caution in patients at risk for or known to have CHF. Further, none of the immunosuppressive agents used in the treatment of psoriasis should be initiated if the patient has a severe infection (including TB, HIV, hepatitis B or C); patients on such therapy should be routinely screened for tuberculosis. There have been reports of progressive multifocal leukoencephalopathy and lupus erythematosus in association with treatment with

the TNF-α inhibitors. Malignancies, including a risk or history of certain malignancies, may limit the use of these systemic agents. In general, immunosuppressive agents have also been linked to an increase risk of skin cancer and patients receiving these agents should be monitored for the development of skin cancer.

TABLE 53-3

FDA-Approved Systemic Therapy for Psoriasis

		Administration		
Agent	**Medication Class**	**Route**	**Frequency**	**Adverse Events (Selected)**
Methotrexate	Antimetabolite	Oral	Weekly[a]	Hepatotoxicity, pulmonary toxicity, pancytopenia, potential for increased malignancies, ulcerative stomatitis, nausea, diarrhea, teratogenicity
Acitretin	Retinoid	Oral	Daily	Teratogenicity, hepatotoxicity, hyperostosis, hyperlipidemia/pancreatitis, depression, ophthalmologic effects, pseudotumor cerebri
Cyclosporine	Calcineurin inhibitor	Oral	Twice daily	Renal dysfunction, hypertension, hyperkalemia, hyperuricemia, hypomagnesemia, hyperlipidemia, increased risk of malignancies
Apremilast	Phosphodiesterase type 4 inhibitor	Oral	Twice daily[b]	Hypersensitivity reaction, depression, nausea, diarrhea, vomiting, dyspepsia, weight loss, headache, fatigue

Abbreviation: FDA, Food and Drug Administration.

[a]Initial test dose is required.

[b]Initial dose escalation is required.

TABLE 53-4

FDA-Approved Biologics for Psoriasis or Psoriatic Arthritis

		Administration			
Agent	**Mechanism of Action**	**Indication**	**Route**	**Frequency**	**Warnings, Selected**
Etanercept	Anti-TNF-α	Ps, PsA	SC	Once or twice weekly[a]	Serious infections, hepatotoxicity, CHF, hematologic events, hypersensitivity reactions, neurologic events, potential for increased malignancies
Adalimumab	Anti-TNF-α	Ps, PsA	SC	Every other week[a]	Serious infections, hepatotoxicity, CHF, hematologic events, hypersensitivity reactions, neurologic events, potential for increased malignancies
Infliximab	Anti-TNF-α	Ps, PsA	IV	Every 8 weeks[a]	Serious infections, hepatotoxicity, CHF, hematologic events, hypersensitivity reactions, neurologic events, potential for increased malignancies
Golimumab	Anti-TNF-α	PsA	SC	Every 4 or 8 weeks	Serious infections, hepatotoxicity, CHF, hypersensitivity reactions, neurologic events, potential for increased malignancies
Ustekinumab	Anti-IL-12 and anti-IL-23	Ps, PsA	SC	Every 12 weeks[a]	Serious infections, neurologic events, potential for increased malignancies
Certolizumab pegol	Anti-TNF-α	PsA	SC	Every 2 or 4 weeks[a]	Serious infections, CHF, hematologic events, hypersensitivity reactions, neurologic events, potential for increased malignancies, hepatotoxicity

		Administration			
Agent	**Mechanism of Action**	**Indication**	**Route**	**Frequency**	**Warnings, Selected**
Secukinumab	Anti-IL-17	Ps, PsA	SC	Every 4 weeks[a]	Serious infections, hypersensitivity reaction, inflammatory bowel disease
Ixekizumab	Anti-IL-17	Ps	SC	Every 4 weeks[a]	Serious infections, hypersensitivity reaction, inflammatory bowel disease

[a]Initial dose modifications required.

Abbreviations: CHF, congestive heart failure; IL, interleukin; IV, intravenous; Ps, psoriasis; PsA, psoriatic arthritis; SC, subcutaneous; TNF-α, tumor necrosis factor-α.

LICHEN PLANUS

Lichen planus (LP) is a papulosquamous disorder that may affect the skin, scalp, nails, and mucous membranes. The primary cutaneous lesions are pruritic, polygonal, flat-topped, violaceous papules. Close examination of the surface of these papules often reveals a network of gray lines (*Wickham's striae*). The skin lesions may occur anywhere but have a predilection for the wrists, shins, lower back, and genitalia **(Fig. 53-5)**. Involvement of the scalp (*lichen planopilaris*) may lead to scarring alopecia, and nail involvement may lead to permanent deformity or loss of fingernails and toenails. LP commonly involves mucous membranes, particularly the buccal mucosa, where it can present on a spectrum ranging from a mild, white, reticulate eruption of the mucosa to a severe, erosive stomatitis. Erosive stomatitis may persist for years and may be linked to an increased risk of oral squamous cell carcinoma. Cutaneous eruptions clinically resembling LP have been observed after administration of numerous drugs, including thiazide diuretics, gold, antimalarial agents, penicillamine, and phenothiazines, and in patients with skin lesions of chronic graft-versus-host disease. In addition, LP may be associated with hepatitis C infection. The course of LP is variable, but most patients have spontaneous remissions 6 months to 2 years after the onset of disease. Topical glucocorticoids are the mainstay of therapy.

FIGURE 53-5

Lichen planus. An example of lichen planus showing multiple flat-topped, violaceous papules and plaques. Nail dystrophy, as seen in this patient's thumbnail, may also be a feature. *(Courtesy of Robert Swerlick, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PITYRIASIS ROSEA

Pityriasis rosea (PR) is a papulosquamous eruption of unknown etiology occurring more commonly in the spring and fall. Its first manifestation is the development of a 2- to 6-cm annular lesion (the herald patch). This is followed in a few days to a few weeks by the appearance of many smaller annular or papular lesions with a predilection to occur on the trunk **(Fig. 53-6)**. The lesions are generally oval, with their long axis parallel to the skinfold lines. Individual lesions may range in color from red to brown and have a trailing scale. PR shares many clinical features with the eruption of secondary syphilis, but palm and sole lesions are extremely rare in PR and common in secondary syphilis. The eruption tends to be moderately pruritic and lasts 3–8 weeks. Treatment is directed at alleviating pruritus and consists of oral antihistamines; mid-potency topical glucocorticoids; and, in some cases, UVB phototherapy.

FIGURE 53-6

Pityriasis rosea. In this patient with pityriasis rosea, multiple round to oval erythematous patches with fine central scale are distributed along the skin tension lines on the trunk.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CUTANEOUS INFECTIONS

(TABLE 53-5)

TABLE 53-5

Common Skin Infections

	Clinical Features	Etiologic Agent	Treatment
Impetigo	Honey-colored crusted papules, plaques, or bullae	Group A *Streptococcus* and *Staphylococcus aureus*	Systemic or topical antistaphylococcal and antistreptococcal antibiotics
Dermatophytosis	Inflammatory or noninflammatory annular scaly plaques; may involve hair loss; groin involvement spares scrotum; hyphae on KOH preparation	*Trichophyton, Epidermophyton*, or *Microsporum* spp.	Topical azoles, systemic griseofulvin, terbinafine, or azoles
Candidiasis	Inflammatory papules and plaques with satellite pustules, frequently in intertriginous areas; may involve scrotum; pseudohyphae on KOH preparation	*Candida albicans* and other *Candida* spp.	Topical nystatin or azoles; systemic azoles for resistant disease
Tinea versicolor	Hyper- or hypopigmented scaly patches on trunk; characteristic mixture of hyphae and spores ("spaghetti and meatballs") on KOH preparation	*Malassezia furfur*	Topical selenium sulfide lotion or azoles

Abbreviation: KOH, potassium hydroxide.

IMPETIGO, ECTHYMA, AND FURUNCULOSIS

Impetigo is a common superficial bacterial infection of skin caused most often by *S. aureus* **(Chap. 142)** and in some cases by group A β-hemolytic streptococci **(Chap. 143)**. The primary lesion is a superficial pustule that ruptures and forms a characteristic yellow-brown honey-colored crust **(see Chap. 143, Fig. 143-3)**. Lesions may occur on normal skin (primary infection) or in areas already affected by another skin disease (secondary infection). Lesions caused by staphylococci may be tense, clear bullae, and this less common form of the disease is called *bullous impetigo*. Blisters are caused by the production of exfoliative toxin by *S. aureus* phage type II. This is the same toxin responsible for staphylococcal scalded-skin syndrome, often

resulting in dramatic loss of the superficial epidermis due to blistering. The latter syndrome is much more common in children than in adults; however, it should be considered along with toxic epidermal necrolysis and severe drug eruptions in patients with widespread blistering of the skin. *Ecthyma* is a deep nonbullous variant of impetigo that causes punched-out ulcerative lesions. It is more often caused by a primary or secondary infection with *Streptococcus pyogenes*. Ecthyma is a deeper infection than typical impetigo and resolves with scars. Treatment of both ecthyma and impetigo involves gentle debridement of adherent crusts, which is facilitated by the use of soaks and topical antibiotics in conjunction with appropriate oral antibiotics.

Furunculosis is also caused by *S. aureus*, and this disorder has gained prominence in the last decade because of CA-MRSA. A furuncle, or boil, is a painful, erythematous nodule that can occur on any cutaneous surface. The lesions may be solitary but are most often multiple. Patients frequently believe they have been bitten by spiders or insects. Family members or close contacts may also be affected. Furuncles can rupture and drain spontaneously or may need incision and drainage, which may be adequate therapy for small solitary furuncles without cellulitis or systemic symptoms. Whenever possible, lesional material should be sent for culture. Current recommendations for methicillin-sensitive infections are β-lactam antibiotics. Therapy for CA-MRSA is discussed previously (see "Atopic Dermatitis"). Warm compresses and nasal mupirocin are helpful therapeutic additions. Severe infections may require IV antibiotics.

ERYSIPELAS AND CELLULITIS

See Chap. 124.

DERMATOPHYTOSIS

Dermatophytes are fungi that infect skin, hair, and nails and include members of the genera *Trichophyton*, *Microsporum*, and *Epidermophyton* **(Chap. 214)**. *Tinea corporis*, or infection of the relatively hairless skin of the body (glabrous skin), may have a variable appearance depending on the extent of the associated inflammatory reaction. Typical infections consist of erythematous, scaly plaques, with an annular appearance that accounts for the common name "ringworm." Deep inflammatory nodules or granulomas occur in some infections, most often those inappropriately treated with mid- to high-potency topical glucocorticoids. Involvement of the groin (*tinea cruris*) is more common in males than in females. It presents as a scaling, erythematous eruption sparing the scrotum. Infection of the foot (*tinea pedis*) is the most common dermatophyte infection and is often chronic; it is characterized by variable erythema, edema, scaling, pruritus, and occasionally vesiculation. The infection may be widespread or localized but generally involves the web space between the fourth and fifth toes. Infection of the nails (*tinea unguium* or *onychomycosis*) occurs in many patients with tinea pedis and is characterized by opacified, thickened nails and subungual debris. The distal-lateral variant is most common. Proximal subungual onychomycosis may be a marker for HIV infection or other immunocompromised states. Dermatophyte infection of the scalp (*tinea capitis*) continues to be common, particularly affecting inner-city children but also affecting adults. The predominant organism is *Trichophyton tonsurans*, which can produce a relatively noninflammatory

infection with mild scale and hair loss that is diffuse or localized. *T. tonsurans* and *Microsporum canis* can also cause a markedly inflammatory dermatosis with edema and nodules. This latter presentation is a *kerion*.

The diagnosis of tinea can be made from skin scrapings, nail scrapings, or hair by culture or direct microscopic examination with KOH. Nail clippings may be sent for histologic examination with periodic acid–Schiff (PAS) stain.

TREATMENT

TREATMENT

Dermatophytosis

Both topical and systemic therapies may be used in dermatophyte infections. Treatment depends on the site involved and the type of infection. Topical therapy is generally effective for uncomplicated tinea corporis, tinea cruris, and limited tinea pedis. Topical agents are not effective as monotherapy for tinea capitis or onychomycosis (see below), and nystatin is not active against dermatophytes. Topicals are generally applied twice daily, and treatment should continue for 1 week beyond clinical resolution of the infection. Tinea pedis often requires longer treatment courses and frequently relapses. Oral antifungal agents may be required for recalcitrant tinea pedis or tinea corporis.

For dermatophyte infections involving the hair and nails and for other infections unresponsive to topical therapy, oral antifungal agents are often used. Markedly inflammatory tinea capitis may result in scarring and hair loss, and a systemic antifungal agent plus systemic or topical glucocorticoids may be helpful in preventing these sequelae. A fungal etiology should be confirmed by direct microscopic examination or by culture before oral antifungal agents are prescribed for any infection. All of the oral agents may cause hepatotoxicity. They should not be used in women who are pregnant or breast-feeding.

Griseofulvin is approved in the United States for dermatophyte infections involving the skin, hair, or nails. Common side effects of griseofulvin include gastrointestinal distress, headache, and urticaria.

Two newer oral antifungal agents, itraconazole and terbinafine, are sometimes prescribed “off-label” for superficial fungal infections. Oral itraconazole is approved for onychomycosis. Itraconazole has the potential for serious interactions with other drugs requiring the P450 enzyme system for metabolism. Itraconazole should not be administered to patients with evidence of ventricular dysfunction or patients with known CHF.

Terbinafine is also approved for onychomycosis, and the granule version is approved for treatment of tinea capitis. Terbinafine has fewer interactions with other drugs than itraconazole; however, caution should be used with patients who are on multiple medications. The risk/benefit ratio should be considered when an asymptomatic toenail infection is treated with systemic agents.

The FDA has limited the use of a third oral agent due to potential hepatotoxicity and published the following: “Nizoral [ketoconazole] oral tablets should not be a first-line treatment for any fungal infection.” The topical form of ketoconazole is not affected by this action.

TINEA (PITYRIASIS) VERSICOLOR

Tinea versicolor is caused by a nondermatophytic, dimorphic fungus, *Malassezia furfur*, a normal inhabitant of the skin. The expression of infection is promoted by heat and humidity. The typical lesions consist of oval scaly macules, papules, and patches concentrated on the chest, shoulders, and back but only rarely on the face or distal extremities. On dark skin the lesions often appear as hypopigmented areas, whereas on light skin they are slightly erythematous or hyperpigmented. A KOH preparation from scaling lesions will demonstrate a confluence of short hyphae and round spores ("spaghetti and meatballs"). Lotions or shampoos containing sulfur, salicylic acid, or selenium sulfide are the treatments of choice and will clear the infection if used daily for 1–2 weeks and then weekly thereafter. These preparations are irritating if left on the skin for >10 min; thus, they should be washed off completely. Treatment with some oral antifungal agents is also effective, but they do not provide lasting results and are not FDA approved for this indication.

CANDIDIASIS

Candidiasis is a fungal infection caused by a related group of yeasts whose manifestations may be localized to the skin and mucous membranes or, rarely, may be systemic and life-threatening **(Chap. 211)**. The causative organism is usually *Candida albicans*. These organisms are normal saprophytic inhabitants of the gastrointestinal tract but may overgrow due to broad-spectrum antibiotic therapy, diabetes mellitus, or immunosuppression and cause disease. Candidiasis is a very common infection in HIV-infected individuals **(Chap. 197)**. The oral cavity is commonly involved. Lesions may occur on the tongue or buccal mucosa (*thrush*) and appear as white plaques. Fissured, macerated lesions at the corners of the mouth (*perlèche*) are often seen in individuals with poorly fitting dentures and may also be associated with candidal infection. In addition, candidal infections have an affinity for sites that are chronically wet and macerated, including the skin around nails (onycholysis and paronychia), and in intertriginous areas. Intertriginous lesions are characteristically edematous, erythematous, and scaly, with scattered "satellite pustules." In males, there is often involvement of the penis and scrotum as well as the inner aspect of the thighs. In contrast to dermatophyte infections, candidal infections are frequently painful and accompanied by a marked inflammatory response. Diagnosis of candidal infection is based on the clinical pattern and demonstration of yeast on KOH preparation or culture.

TREATMENT

TREATMENT

Candidiasis

Treatment involves removal of any predisposing factors such as antibiotic therapy or chronic wetness and the use of appropriate topical or systemic antifungal agents. Effective topicals include nystatin or azoles (miconazole, clotrimazole, econazole, or ketoconazole). The associated inflammatory response accompanying candidal infection on glabrous skin can be treated with a mild glucocorticoid lotion or cream (2.5% hydrocortisone). Systemic therapy is usually reserved for immunosuppressed patients or individuals

with chronic or recurrent disease who fail to respond to appropriate topical therapy. Oral fluconazole is most commonly prescribed for cutaneous candidiasis. Oral nystatin is effective only for candidiasis of the gastrointestinal tract.

WARTS

Warts are cutaneous neoplasms caused by papillomaviruses. More than 100 different human papillomaviruses (HPVs) have been described. A typical wart, *verruca vulgaris*, is sessile, dome-shaped, and usually about a centimeter in diameter. Its surface is hyperkeratotic, consisting of many small filamentous projections. HPV also causes typical plantar warts, flat warts (*verruca plana*), and filiform warts. Plantar warts are endophytic and are covered by thick keratin. Paring of the wart will generally reveal a central core of keratinized debris and punctate bleeding points. Filiform warts are most commonly seen on the face, neck, and skinfolds, and present as papillomatous lesions on a narrow base. Flat warts are only slightly elevated and have a velvety, nonverrucous surface. They have a propensity for the face, arms, and legs, and are often spread by shaving.

Genital warts begin as small papillomas that may grow to form large, fungating lesions. In women, they may involve the labia, perineum, or perianal skin. In addition, the mucosa of the vagina, urethra, and anus can be involved as well as the cervical epithelium. In men, the lesions often occur initially in the coronal sulcus but may be seen on the shaft of the penis, the scrotum, or the perianal skin or in the urethra.

Appreciable evidence has accumulated indicating that HPV plays a role in the development of neoplasia of the uterine cervix and anogenital skin **(Chap. 85)**. HPV types 16 and 18 have been most intensely studied and are the major risk factors for intraepithelial neoplasia and squamous cell carcinoma of the cervix, anus, vulva, and penis. The risk is higher among patients immunosuppressed after solid organ transplantation and among those infected with HIV. Recent evidence also implicates other HPV types. Histologic examination of biopsied samples from affected sites may reveal changes associated with typical warts and/or features typical of intraepidermal carcinoma (Bowen's disease). Squamous cell carcinomas associated with HPV infections have also been observed in extragenital skin **(Chap. 72)**, most commonly in patients immunosuppressed after organ transplantation. Patients on long-term immunosuppression should be monitored for the development of squamous cell carcinoma and other cutaneous malignancies.

TREATMENT

TREATMENT

Warts

Treatment of warts, other than anogenital warts, should be tempered by the observation that a majority of warts in normal individuals resolve spontaneously within 1–2 years. There are many modalities available to treat warts, but no single therapy is universally effective. Factors that influence the choice of therapy include the location of the wart, the extent of disease, the age and immunologic status of the patient, and the patient's desire for therapy. Perhaps the most useful and convenient method for treating warts in almost any

location is cryotherapy with liquid nitrogen. Equally effective for nongenital warts, but requiring much more patient compliance, is the use of keratolytic agents such as salicylic acid plasters or solutions. For genital warts, in-office application of a podophyllin solution is moderately effective but may be associated with marked local reactions. Prescription preparations of dilute, purified podophyllin are available for home use. Topical imiquimod, a potent inducer of local cytokine release, has been approved for treatment of genital warts. A new topical compound composed of green tea extracts (sinecatechins) is also available. Conventional and laser surgical procedures may be required for recalcitrant warts. Recurrence of warts appears to be common with all these modalities. A highly effective vaccine for selected types of HPV has been approved by the FDA, and its use is reported to reduce the incidence of anogenital and cervical carcinoma.

HERPES SIMPLEX

See Chap. 187.

HERPES ZOSTER

See Chap. 188.

ACNE

ACNE VULGARIS

Acne vulgaris is a self-limited disorder primarily of teenagers and young adults, although perhaps 10–20% of adults may continue to experience some form of the disorder. The permissive factor for the expression of the disease in adolescence is the increase in sebum production by sebaceous glands after puberty. Small cysts, called *comedones*, form in hair follicles due to blockage of the follicular orifice by retention of keratinous material and sebum. The activity of bacteria (*Propionibacterium acnes*) within the comedones releases free fatty acids from sebum, causes inflammation within the cyst, and results in rupture of the cyst wall. An inflammatory foreign-body reaction develops as result of extrusion of oily and keratinous debris from the cyst.

The clinical hallmark of acne vulgaris is the comedone, which may be closed (*whitehead*) or open (*blackhead*). Closed comedones appear as 1- to 2-mm pebbly white papules, which are accentuated when the skin is stretched. They are the precursors of inflammatory lesions of acne vulgaris. The contents of closed comedones are not easily expressed. Open comedones, which rarely result in inflammatory acne lesions, have a large dilated follicular orifice and are filled with easily expressible oxidized, darkened, oily debris. Comedones are usually accompanied by inflammatory lesions: papules, pustules, or nodules.

The earliest lesions seen in adolescence are generally mildly inflamed or noninflammatory comedones on the forehead. Subsequently, more typical inflammatory lesions develop on the cheeks, nose, and chin **(Fig. 53-7)**. The most common location for acne is the face, but involvement of the chest and back is common.

Most disease remains mild and does not lead to scarring. A small number of patients develop large inflammatory cysts and nodules, which may drain and result in significant scarring. Regardless of the severity, acne may affect a patient's quality of life. With adequate treatment, this effect may be transient. In the case of severe, scarring acne, the effects can be permanent and profound. Early therapeutic intervention in severe acne is essential.

FIGURE 53-7

Acne vulgaris. An example of acne vulgaris with inflammatory papules, pustules, and comedones. *(Courtesy of Kalman Watsky, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Exogenous and endogenous factors can alter the expression of acne vulgaris. Friction and trauma (from headbands or chin straps of athletic helmets), application of comedogenic topical agents (cosmetics or hair preparations), or chronic topical exposure to certain industrial compounds may elicit or aggravate acne. Glucocorticoids, topical or systemic, may also elicit acne. Other systemic medications such as oral contraceptive pills, lithium, isoniazid, androgenic steroids, halogens, phenytoin, and phenobarbital may produce acneiform eruptions or aggravate preexisting acne. Genetic factors and polycystic ovary disease may also play a role.

TREATMENT

TREATMENT

Acne Vulgaris

Treatment of acne vulgaris is directed toward elimination of comedones by normalizing follicular keratinization and decreasing sebaceous gland activity, the population of *P. acnes*, and inflammation. Minimal to moderate pauci-inflammatory disease may respond adequately to local therapy alone. Although areas affected with acne should be kept clean, overly vigorous scrubbing may aggravate acne due to

mechanical rupture of comedones. Topical agents such as retinoic acid, benzoyl peroxide, or salicylic acid may alter the pattern of epidermal desquamation, preventing the formation of comedones and aiding in the resolution of preexisting cysts. Topical antibacterial agents (such as azelaic acid, erythromycin, clindamycin, or dapsone) are also useful adjuncts to therapy. Benzoyl peroxide products should be used in combination with topical antibiotics (erythromycin and clindamycin) to prevent development of bacterial resistance.

Patients with moderate to severe acne with a prominent inflammatory component will benefit from the addition of systemic therapy, such as tetracycline in doses of 250–500 mg bid or doxycycline in doses of 100 mg bid. Minocycline is also useful. Such antibiotics appear to have anti-inflammatory effects independent of their antibacterial effects. If the patient is not showing appropriate response within 3 months, changes in the plan should be considered. Female patients who do not respond to oral antibiotics may benefit from hormonal therapy. Several oral contraceptives are now approved by the FDA for use in the treatment of acne vulgaris.

Patients with severe nodulocystic acne unresponsive to the therapies discussed above may benefit from treatment with the synthetic retinoid isotretinoin. Its dose is based on the patient's weight, and it is given once daily for 5 months. Results are excellent in appropriately selected patients. Its use is highly regulated due to its potential for severe adverse events, primarily teratogenicity and depression. In addition, patients receiving this medication develop extremely dry skin and cheilitis and must be followed for development of hypertriglyceridemia.

At present, prescribers must enroll in a program designed to prevent pregnancy and adverse events while patients are taking isotretinoin. These measures are imposed to ensure that all prescribers are familiar with the risks of isotretinoin, that all female patients have two negative pregnancy tests prior to initiation of therapy and a negative pregnancy test prior to each refill, and that all patients have been warned about the risks associated with isotretinoin.

ACNE ROSACEA

Acne rosacea, commonly referred to simply as *rosacea*, is an inflammatory disorder predominantly affecting the central face. Persons most often affected are Caucasians of northern European background, but rosacea also occurs in patients with dark skin. Rosacea is seen almost exclusively in adults, only rarely affecting patients <30 years old. Rosacea is more common in women, but those most severely affected are men. It is characterized by the presence of erythema, telangiectases, and superficial pustules **(Fig. 53-8)** but is not associated with the presence of comedones. Rosacea rarely involves the chest or back.

FIGURE 53-8

Acne rosacea. Prominent facial erythema, telangiectasia, scattered papules, and small pustules are seen in this patient with acne rosacea. *(Courtesy of Robert Swerlick, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

There is a relationship between the tendency for facial flushing and the subsequent development of acne rosacea. Often, individuals with rosacea initially demonstrate a pronounced flushing reaction. This may be in response to heat, emotional stimuli, alcohol, hot drinks, or spicy foods. As the disease progresses, the flush persists longer and longer and may eventually become permanent. Papules, pustules, and telangiectases can become superimposed on the persistent flush. Rosacea of very long standing may lead to connective tissue overgrowth, particularly of the nose (*rhinophyma*). Rosacea may also be complicated by various inflammatory disorders of the eye, including keratitis, blepharitis, iritis, and recurrent chalazion. These ocular problems are potentially sight-threatening and warrant ophthalmologic evaluation.

TREATMENT

TREATMENT

Acne Rosacea

Acne rosacea can be treated topically or systemically. Mild disease often responds to topical metronidazole, sodium sulfacetamide, azelaic acid, topical ivermectin, or topical brimonidine. More severe disease requires oral tetracyclines: tetracycline, 250–500 mg bid; doxycycline, 100 mg bid; or minocycline, 50–100 mg bid. Residual telangiectasia may respond to laser therapy. Topical glucocorticoids, especially potent agents, should be avoided because chronic use of these preparations may elicit rosacea. Application of topical agents to the skin is not effective treatment for ocular disease.

SKIN DISEASES AND SMALLPOX VACCINATION

Although smallpox vaccinations were discontinued several decades ago for the general population, they are still required for certain military personnel and first responders. In the absence of a bioterrorism attack and a real or potential exposure to smallpox, such vaccination is contraindicated in persons with a history of skin

diseases such as AD, eczema, and psoriasis, who have a higher incidence of adverse events associated with smallpox vaccination. In the case of such exposure, the risk of smallpox infection outweighs that of adverse events from the vaccine (Chap. S2).

FURTHER READING

Bolognia JL, Jorizzo JL, Schaffer JV (eds): *Dermatology*, 3rd ed. Philadelphia, Saunders, 2012.

Goldsmith LA et al (eds): *Fitzpatrick's Dermatology in General Medicine*, 8th ed. New York, McGraw-Hill, 2012.

James WD, Berger TG, Elston DM (eds): *Andrew's Diseases of the Skin Clinical Dermatology*, 12th ed. Philadelphia, Elsevier, 2016.

Wolff K, Johnson RA, Saavedra AP (eds): *Fitzpatrick's Color Atlas and Synopsis of Clinical Dermatology*, 7th ed. New York, McGraw-Hill, 2013.

Chapter 54: Skin Manifestations of Internal Disease

Jean L. Bolognia; Irwin M. Braverman

INTRODUCTION

It is a generally accepted concept in medicine that the skin can develop signs of internal disease. Therefore, in textbooks of medicine, one finds a chapter describing in detail the major systemic disorders that can be identified by cutaneous signs. The underlying assumption of such a chapter is that the clinician has been able to identify the specific disorder in the patient and needs only to read about it in the textbook. In reality, concise differential diagnoses and the identification of these disorders are actually difficult for the nondermatologist because he or she is not well-versed in the recognition of cutaneous lesions or their spectrum of presentations. Therefore, this chapter covers this particular topic of cutaneous medicine not by simply focusing on individual diseases, but by describing the various presenting clinical signs and symptoms that point to specific disorders. Concise differential diagnoses will be generated in which the significant diseases will be distinguished from the more common cutaneous disorders that have minimal or no significance with regard to associated internal disease. The latter disorders are reviewed in table form and always need to be excluded when considering the former. For a detailed description of individual diseases, the reader should consult a dermatologic text.

PAPULOSQUAMOUS SKIN LESIONS

(Table 54-1) When an eruption is characterized by elevated lesions, either papules (<1 cm) or plaques (>1 cm), in association with scale, it is referred to as *papulosquamous*. The most common papulosquamous diseases—*tinea*, *psoriasis*, *pityriasis rosea*, and *lichen planus*—are primary cutaneous disorders (Chap. 53). When psoriatic lesions are accompanied by arthritis, the possibility of psoriatic arthritis or reactive arthritis should be considered. A history of oral ulcers, conjunctivitis, uveitis, and/or urethritis points to the latter diagnosis. Lithium, beta blockers, HIV or streptococcal infections, and a rapid taper of systemic glucocorticoids are known to exacerbate psoriasis; despite being used to treat psoriasis, TNF-α inhibitors can also induce psoriatic lesions. Comorbidities in patients with psoriasis include cardiovascular disease and metabolic syndrome.

TABLE 54-1

Selected Causes of Papulosquamous Skin Lesions

1. Primary cutaneous disorders
    a. Tinea[a]—widespread disease may be sign of immunosuppression
    b. Psoriasis[a]—widespread or resistant disease may be sign of HIV infection
    c. Pityriasis rosea[a]
    d. Lichen planus[a]
    e. Parapsoriasis, small plaque and large plaque
    f. Bowen's disease (squamous cell carcinoma in situ)[b]
2. Drugs
3. Systemic diseases
    a. Lupus erythematosus, primarily subacute or chronic (discoid) lesions[c]
    b. Cutaneous T cell lymphoma, in particular, mycosis fungoides[d]
    c. Secondary syphilis
    d. Reactive arthritis
    e. Sarcoidosis[e]—with scale less common than without scale

[a]Discussed in detail in **Chap. 53**; cardiovascular disease and the metabolic syndrome are comorbidities in psoriasis; primarily in Europe, hepatitis C virus is associated with oral lichen planus. [b]Associated with chronic sun exposure more often than exposure to arsenic; usually one or a few lesions. [c]See also Red Lesions in "Papulonodular Skin Lesions." [d]Also cutaneous lesions of HTLV-1-associated adult T cell leukemia/lymphoma. [e]See also Red-Brown Lesions in "Papulonodular Skin Lesions."

Abbreviation: HIV, human immunodeficiency virus.

Whenever the diagnosis of pityriasis rosea or lichen planus is made, it is important to review the patient's medications because the eruption may resolve by simply discontinuing the offending agent. Pityriasis rosea–like drug eruptions are seen most commonly with beta blockers, angiotensin-converting enzyme (ACE) inhibitors, and metronidazole, whereas the drugs that can produce a lichenoid eruption include thiazides, antimalarials, quinidine, beta blockers, TNF-α inhibitors, anti-PD-1/PD-L1 Ab, and ACE inhibitors. In some populations, there is a higher prevalence of hepatitis C viral infection in patients with oral lichen planus. Lichen planus–like lesions are also observed in chronic graft-versus-host disease.

In its early stages, the mycosis fungoides (MF) form of *cutaneous T cell lymphoma* (CTCL) may be confused with eczema or psoriasis, but it often fails to respond to appropriate therapy for those inflammatory diseases. MF can develop within lesions of large-plaque parapsoriasis and is suggested by an increase in the thickness of the lesions. The diagnosis of MF is established by skin biopsy in which collections of atypical T

lymphocytes are found in the epidermis and dermis. As the disease progresses, cutaneous tumors and lymph node involvement may appear.

In *secondary syphilis*, there are scattered red-brown papules with thin scale. The eruption often involves the palms and soles and can resemble pityriasis rosea. Associated findings are helpful in making the diagnosis and include annular plaques on the face, nonscarring alopecia, condyloma lata (broad-based and moist), and mucous patches as well as lymphadenopathy, malaise, fever, headache, and myalgias. The interval between the primary chancre and the secondary stage is usually 4–8 weeks, and spontaneous resolution without appropriate therapy occurs.

ERYTHRODERMA

(Table 54-2) *Erythroderma* is the term used when the majority of the skin surface is erythematous (red in color). There may be associated scale, erosions, or pustules as well as shedding of the hair and nails. Potential systemic manifestations include fever, chills, hypothermia, reactive lymphadenopathy, peripheral edema, hypoalbuminemia, and high-output cardiac failure. The major etiologies of erythroderma are (1) cutaneous diseases such as psoriasis and dermatitis (Table 54-3); (2) drugs; (3) systemic diseases, most commonly CTCL; and (4) idiopathic. In the first three groups, the location and description of the initial lesions, prior to the development of the erythroderma, aid in the diagnosis. For example, a history of red scaly plaques on the elbows and knees would point to psoriasis. It is also important to examine the skin carefully for a migration of the erythema and associated secondary changes such as pustules or erosions. Migratory waves of erythema studded with superficial pustules are seen in *pustular psoriasis*.

TABLE 54-2

Causes of Erythroderma

1. Primary cutaneous disorders
    a. Psoriasis[a]
    b. Dermatitis (atopic > contact >> stasis [with autosensitization] or seborrheic [primarily infants])[a]
    c. Pityriasis rubra pilaris
2. Drugs
3. Systemic diseases
    a. Cutaneous T cell lymphoma (Sézary syndrome, erythrodermic mycosis fungoides)
    b. Other lymphomas
4. Idiopathic (usually older men)

[a]Discussed in detail in Chap. 53.

TABLE 54-3

Erythroderma (Primary Cutaneous Disorders)

	Initial Lesions	Location of Initial Lesions	Other Findings	Diagnostic Aids	Treatment
Psoriasis[a]	Pink-red, silvery scale, sharply demarcated	Elbows, knees, scalp, presacral area, intergluteal fold	Nail dystrophy, arthritis, pustules, SAPHO syndrome[b]	Skin biopsy	Topical glucocorticoids, vitamin D; UV-B (narrowband) > PUVA; oral retinoid; MTX, cyclosporine, anti-TNF agents, apremilast, anti-IL-12/23 Ab, anti-IL-17A or -IL-17 receptor Ab
Dermatitis[a]					

	Initial Lesions	Location of Initial Lesions	Other Findings	Diagnostic Aids	Treatment
Atopic	Acute: Erythema, fine scale, crust, indistinct borders, excoriations Chronic: Lichenification (increased skin markings), excoriations	Antecubital and popliteal fossae, neck, hands, eyelids	Pruritus Personal and/or family history of atopy, including asthma, allergic rhinitis or conjunctivitis, and atopic dermatitis Exclude secondary infection with *Staphylococcus aureus* or HSV Exclude superimposed irritant or allergic contact dermatitis	Skin biopsy	Topical glucocorticoids, tacrolimus, pimecrolimus, tar, and antipruritics; oral antihistamines; open wet dressings; UV-B ± UV-A > PUVA; oral/IM glucocorticoids (short-term); MTX; mycophenolate mofetil; azathioprine; cyclosporine; anti-IL-4/13 Ab Topical or oral antibiotics

	Initial Lesions	Location of Initial Lesions	Other Findings	Diagnostic Aids	Treatment
Contact	Local: Erythema, crusting, vesicles, and bullae Systemic: Erythema, fine scale, crust	Depends on offending agent Generalized vs major intertriginous zones (especially groin)	Irritant—onset often within hours Allergic—delayed-type hypersensitivity; lag time of 48 h with re-challenge Patient has history of allergic contact dermatitis to topical agent and then receives systemic medication that is structurally related, e.g., formaldehyde (skin), aspartame (oral)	Patch testing; repeat open application test Patch testing	Remove irritant or allergen; topical glucocorticoids; oral antihistamines; oral/IM glucocorticoids (short-term) Same as local
Seborrheic (rare in adults)	Pink-red to pink-orange, greasy scale	Scalp, nasolabial folds, eyebrows, intertriginous zones	Flares with stress, HIV infection Associated with Parkinson’s disease	Skin biopsy	Topical glucocorticoids and imidazoles

	Initial Lesions	Location of Initial Lesions	Other Findings	Diagnostic Aids	Treatment
Stasis (with autosensitization)	Erythema, crusting, excoriations	Lower extremities	Pruritus, lower extremity edema, varicosities, hemosiderin deposits, lipodermatosclerosis History of venous ulcers, thrombophlebitis, and/or cellulitis Exclude cellulitis Exclude superimposed contact dermatitis, e.g., topical neomycin	Skin biopsy	Topical glucocorticoids; open wet dressings; leg elevation; pressure stockings; pressure wraps if associated ulcers
Pityriasis rubra pilaris	Orange-red (salmon-colored), perifollicular papules	Generalized, but characteristic "skip" areas of normal skin	Wax-like palmoplantar keratoderma Exclude cutaneous T cell lymphoma	Skin biopsy	Isotretinoin or acitretin; MTX; perhaps anti-IL-12/23 Ab, anti-TNF agents, anti-IL-17 Ab

[a]Discussed in detail in Chap. 53. [b]SAPHO syndrome occurs more commonly in patients with palmoplantar pustulosis than in those with erythrodermic psoriasis.

Abbreviations: Ab, antibody; HSV, herpes simplex virus; IL, interleukin; IM, intramuscular; MTX, methotrexate; PUVA, psoralens + ultraviolet A irradiation; SAPHO, *s*ynovitis, *a*cne, *p*ustulosis, *h*yperostosis, and *o*steitis (a subtype is chronic recurrent multifocal osteomyelitis); TNF, tumor necrosis factor; UV-A, ultraviolet A irradiation; UV-B, ultraviolet B irradiation.

Drug-induced erythroderma (exfoliative dermatitis) may begin as an exanthematous (morbilliform) eruption (Chap. 56) or may arise as diffuse erythema. A number of drugs can produce an erythroderma, including penicillins, sulfonamides, carbamazepine, phenytoin, and allopurinol. Fever and peripheral eosinophilia often accompany the eruption, and there may also be facial swelling, hepatitis, myocarditis, thyroiditis, and

allergic interstitial nephritis; this constellation is frequently referred to as *drug reaction with eosinophilia and systemic symptoms* (DRESS) or *drug-induced hypersensitivity reaction* (DIHS). In addition, these reactions, especially to aromatic anticonvulsants, can lead to a pseudolymphoma syndrome (with adenopathy and circulating atypical lymphocytes), while reactions to allopurinol may be accompanied by gastrointestinal bleeding.

The most common malignancy that is associated with erythroderma is CTCL; in some series, up to 25% of the cases of erythroderma were due to CTCL. The patient may progress from isolated plaques and tumors, but more commonly, the erythroderma is present throughout the course of the disease (Sézary syndrome). In Sézary syndrome, there are circulating clonal atypical T lymphocytes, pruritus, and lymphadenopathy. In cases of erythroderma where there is no apparent cause (idiopathic), longitudinal evaluation is mandatory to monitor for the possible development of CTCL. There have been isolated case reports of erythroderma secondary to some solid tumors—lung, liver, prostate, thyroid, and colon—but it is primarily during a late stage of the disease.

ALOPECIA

(Table 54-4) The two major forms of alopecia are scarring and nonscarring. *Scarring alopecia* is associated with fibrosis, inflammation, and loss of hair follicles. A smooth scalp with a decreased number of follicular openings is usually observed clinically, but in some patients, the changes are seen only in biopsy specimens from affected areas. In *nonscarring alopecia*, the hair shafts are absent or miniaturized, but the hair follicles are preserved, explaining the reversible nature of nonscarring alopecia.

TABLE 54-4

Causes of Alopecia

I. Nonscarring alopecia
  A. Primary cutaneous disorders
    1. Androgenetic alopecia
    2. Telogen effluvium
    3. Alopecia areata
    4. Tinea capitis
    5. Traumatic alopecia[a]
    6. Psoriasiform alopecia, including TNF-α inhibitor-induced
  B. Drugs
  C. Systemic diseases
    1. Systemic lupus erythematosus
    2. Secondary syphilis
    3. Hypothyroidism
    4. Hyperthyroidism
    5. Hypopituitarism
    6. Deficiencies of protein, biotin, zinc, and perhaps iron
II. Scarring alopecia
  A. Primary cutaneous disorders
    1. Cutaneous lupus (chronic discoid lesions)[b]
    2. Lichen planus, including frontal fibrosing alopecia
    3. Central centrifugal cicatricial alopecia
    4. Folliculitis decalvans
    5. Linear morphea (linear scleroderma)[c]
  B. Systemic diseases
    1. Discoid lesions in the setting of systemic lupus erythematosus[b]
    2. Sarcoidosis
    3. Cutaneous metastases

[a]Most patients with trichotillomania or early stages of traction alopecia and some patients with pressure-induced alopecia.

[b]While the majority of patients with discoid lesions have only cutaneous disease, these lesions do represent one of the 11 American College of Rheumatology criteria (1982) for systemic lupus erythematosus.

[c]Can involve underlying muscles and osseous structures and rarely in linear morphea of the frontal scalp (*en coup de sabre*), there is involvement of the meninges and brain.

The most common causes of nonscarring alopecia include *androgenetic alopecia*, *telogen effluvium*, *alopecia areata*, *tinea capitis*, and the early phase of *traumatic alopecia* (Table 54-5). In women with androgenetic alopecia, an elevation in circulating levels of androgens may be seen as a result of ovarian or adrenal gland dysfunction or neoplasm. When there are signs of virilization, such as a deepened voice and enlarged clitoris, the possibility of an ovarian or adrenal gland tumor should be considered.

TABLE 54-5

Nonscarring Alopecia (Primary Cutaneous Disorders)

	Clinical Characteristics	Pathogenesis	Treatment
Telogen effluvium	Diffuse shedding of normal hairs Follows major stress (high fever, severe infection) or change in hormone levels (postpartum) Reversible without treatment	Stress causes more of the asynchronous growth cycles of individual hairs to become synchronous; therefore, larger numbers of growing (anagen) hairs simultaneously enter the dying (telogen) phase	Observation; discontinue any drugs that have alopecia as a side effect; must exclude underlying metabolic causes, e.g., hypothyroidism, hyperthyroidism
Androgenetic alopecia (male pattern; female pattern)	Miniaturization of hairs along the midline of the scalp Recession of the anterior scalp line in men and some women	Increased sensitivity of affected hairs to the effects of androgens Increased levels of circulating androgens (ovarian or adrenal source in women)	If no evidence of hyperandrogenemia, then topical minoxidil; finasteride[a]; spironolactone (women); hair transplant
Alopecia areata	Well-circumscribed, circular areas of hair loss, 2–5 cm in diameter In extensive cases, coalescence of lesions and/or involvement of other hair-bearing surfaces of the body Pitting or sandpapered appearance of the nails	The germinative zones of the hair follicles are surrounded by T lymphocytes Occasional associated diseases: hyperthyroidism, hypothyroidism, vitiligo, Down syndrome	Topical anthralin or tazarotene; intralesional glucocorticoids; topical contact sensitizers; JAK inhibitors

	Clinical Characteristics	Pathogenesis	Treatment
Tinea capitis	Varies from scaling with minimal hair loss to discrete patches with "black dots" (sites of broken infected hairs) to boggy plaque with pustules (kerion)[b]	Invasion of hairs by dermatophytes, most commonly *Trichophyton tonsurans*	Oral griseofulvin or terbinafine plus 2.5% selenium sulfide or ketoconazole shampoo; examine family members
Traumatic alopecia[c]	Broken hairs, often of varying lengths Irregular outline in trichotillomania and traction alopecia	Traction with curlers, rubber bands, tight braiding Exposure to heat or chemicals (e.g., hair straighteners) Mechanical pulling (trichotillomania)	Discontinuation of offending hair style or chemical treatments; diagnosis of trichotillomania may require observation of shaved hairs (for growth) or biopsy, possibly followed by psychotherapy

[a]To date, Food and Drug Administration–approved for men. [b]Scarring alopecia can occur at sites of kerions. [c]May also be scarring, especially late-stage traction alopecia.

Exposure to various drugs can also cause diffuse hair loss, usually by inducing a telogen effluvium. An exception is the anagen effluvium observed with antimitotic agents such as daunorubicin. Alopecia is a side effect of the following drugs: warfarin, heparin, propylthiouracil, carbimazole, isotretinoin, acitretin, lithium, beta blockers, interferons, colchicine, and amphetamines. Fortunately, spontaneous regrowth usually follows discontinuation of the offending agent.

Less commonly, nonscarring alopecia is associated with *lupus erythematosus* and *secondary syphilis.* In systemic lupus there are two forms of alopecia—one is scarring secondary to discoid lesions (see below), and the other is nonscarring. The latter form coincides with flares of systemic disease and may involve the entire scalp or just the frontal scalp, with the appearance of multiple short hairs ("lupus hairs") as a sign of initial regrowth. Scattered, poorly circumscribed patches of alopecia with a "moth-eaten" appearance are a manifestation of the secondary stage of syphilis. Diffuse thinning of the hair is also associated with hypothyroidism and hyperthyroidism (Table 54-4).

Scarring alopecia is more frequently the result of a primary cutaneous disorder such as *lichen planus*, *chronic cutaneous (discoid) lupus*, *central centrifugal cicatricial alopecia*, *folliculitis decalvans*, or *linear scleroderma* (*morphea*) than it is a sign of systemic disease. Although the scarring lesions of *discoid lupus*

can be seen in patients with systemic lupus, in the majority of patients, the disease process is limited to the skin. Less common causes of scarring alopecia include *sarcoidosis* (see "Papulonodular Skin Lesions," below) and *cutaneous metastases*.

In the early phases of discoid lupus, lichen planus, and folliculitis decalvans, there are circumscribed areas of alopecia. Fibrosis and subsequent loss of hair follicles are observed primarily in the center of these alopecic patches, whereas the inflammatory process is most prominent at the periphery. The areas of active inflammation in discoid lupus are erythematous with scale, whereas the areas of previous inflammation are often hypopigmented with a rim of hyperpigmentation. In lichen planus, perifollicular macules at the periphery are usually violet-colored. A complete examination of the skin and oral mucosa combined with a biopsy and direct immunofluorescence microscopy of inflamed skin will aid in distinguishing these two entities. The peripheral active lesions in folliculitis decalvans are follicular pustules; these patients can develop a reactive arthritis.

FIGURATE SKIN LESIONS

(Table 54-6) In *figurate eruptions*, the lesions form rings and arcs that are usually erythematous but can be skin-colored to brown. Most commonly, they are due to primary cutaneous diseases such as *tinea*, *urticaria*, *granuloma annulare*, and *erythema annulare centrifugum* **(Chaps. 53 and 55)**. An underlying systemic illness is found in a second, less common group of migratory annular erythemas. It includes *erythema migrans*, *erythema gyratum repens*, *erythema marginatum*, and *necrolytic migratory erythema*.

TABLE 54-6

Causes of Figurate Skin Lesions

I. Primary cutaneous disorders
  A. Tinea
  B. Urticaria (primary in ≥90% of patients)
  C. Granuloma annulare
  D. Erythema annulare centrifugum
  E. Psoriasis, annular pustular psoriasis
  F. Interstitial granulomatous drug reaction
II. Systemic diseases
  A. Migratory
    1. Erythema migrans (CDC case definition is ≥5 cm in diameter)
    2. Urticaria (≤10% of patients)
    3. Erythema gyratum repens
    4. Erythema marginatum
    5. Pustular psoriasis (generalized and annular forms)
    6. Necrolytic migratory erythema (glucagonoma syndrome)[a]
  B. Nonmigratory
    1. Sarcoidosis
    2. Subacute cutaneous lupus erythematosus, LE tumidus
    3. Annular erythema of Sjögren's syndrome
    4. Secondary syphilis (especially the face)
    5. Cutaneous T cell lymphoma (especially mycosis fungoides)
    6. Interstitial granulomatous dermatitis[b]

[a]Migratory erythema with erosions; favors lower extremities and girdle area. [b]Underlying diseases include rheumatoid arthritis, LE, and granulomatosis with polyangiitis.

Abbreviations: CDC, Centers for Disease Control and Prevention; LE, lupus erythematosus.

In erythema gyratum repens, one sees numerous mobile concentric arcs and wavefronts that resemble the grain in wood. A search for an underlying malignancy is mandatory in a patient with this eruption. Erythema migrans is the cutaneous manifestation of Lyme disease, which is caused by the spirochete *Borrelia burgdorferi*. In the initial stage (3–30 days after tick bite), a single annular lesion is usually seen, which can expand to ≥10 cm in diameter. Within several days, up to half of the patients develop multiple smaller erythematous lesions at sites distant from the bite. Associated symptoms include fever, headache, photophobia, myalgias, arthralgias, and malar rash. Erythema marginatum is seen in patients with

rheumatic fever, primarily on the trunk. Lesions are pink-red in color, flat to minimally elevated, and transient.

There are additional cutaneous diseases that present as annular eruptions but lack an obvious migratory component. Examples include *CTCL*, *subacute cutaneous lupus*, *secondary syphilis*, and *sarcoidosis* (see "Papulonodular Skin Lesions," below).

ACNE

(Table 54-7) In addition to *acne vulgaris* and *acne rosacea*, the two major forms of acne **(Chap. 53)**, there are drugs and systemic diseases that can lead to acneiform eruptions.

TABLE 54-7

Causes of Acneiform Eruptions

I. Primary cutaneous disorders
  A. Acne vulgaris
  B. Acne rosacea
II. Drugs, e.g., anabolic steroids, glucocorticoids, lithium, EGFR inhibitors, MEK inhibitors iodides
III. Systemic diseases
  A. Increased androgen production
    1. Adrenal origin, e.g., Cushing's disease, 21-hydroxylase deficiency
    2. Ovarian origin, e.g., polycystic ovary syndrome, ovarian hyperthecosis
  B. Cryptococcosis, disseminated
  C. Dimorphic fungal infections
  D. Behçet's disease

Abbreviation: EGFR, epidermal growth factor receptor; MEK, MAP (mitogen activated protein) kinase.

Patients with the *carcinoid syndrome* have episodes of flushing of the head, neck, and sometimes the trunk. Resultant skin changes of the face, in particular telangiectasias, may mimic the clinical appearance of erythematotelangiectatic acne rosacea.

PUSTULAR LESIONS

Acneiform eruptions (see "Acne," above) and *folliculitis* represent the most common pustular dermatoses. An important consideration in the evaluation of follicular pustules is a determination of the associated pathogen, for example, normal flora (culture-negative), *Staphylococcus aureus*, *Pseudomonas aeruginosa* ("hot tub" folliculitis), *Malassezia*, dermatophytes (Majocchi's granuloma), and *Demodex* spp. Noninfectious forms of folliculitis include HIV- or immunosuppression-associated eosinophilic folliculitis and folliculitis

secondary to drugs such as glucocorticoids, lithium, and epidermal growth factor receptor (EGFR) or MEK inhibitors. Administration of high-dose systemic glucocorticoids can result in a widespread eruption of follicular pustules on the trunk, characterized by lesions in the same stage of development. With regard to underlying systemic diseases, nonfollicular-based pustules are a characteristic component of pustular psoriasis (sterile) and can be seen in septic emboli of bacterial or fungal origin (see "Purpura," below). In patients with acute generalized exanthematous pustulosis (AGEP) due primarily to medications (e.g., cephalosporins), there are large areas of erythema studded with multiple sterile pustules in addition to neutrophilia.

TELANGIECTASIAS

(Table 54-8) To distinguish the various types of telangiectasias, it is important to examine the shape and configuration of the dilated blood vessels. *Linear telangiectasias* are seen on the face of patients with *actinically damaged skin* and *acne rosacea,* and they are found on the legs of patients with *venous hypertension* and first appear on the legs in *generalized essential telangiectasia*. Patients with an unusual form of *mastocytosis* (telangiectasia macularis eruptiva perstans) and the *carcinoid syndrome* (see "Acne," above) also have linear telangiectasias. Lastly, linear telangiectasias are found in areas of cutaneous inflammation. For example, longstanding lesions of discoid lupus frequently have telangiectasias within them.

TABLE 54-8

Causes of Telangiectasias

I. Primary cutaneous disorders
  A. Linear/branching
    1. Acne rosacea (face)
    2. Actinically damaged skin (face, neck, V of chest)
    3. Venous hypertension (legs)
    4. Generalized essential telangiectasia
    5. Cutaneous collagenous vasculopathy
    6. Within basal cell carcinomas or cutaneous lymphoma
  B. Poikiloderma
    1. Ionizing radiation[a]
  C. Spider angioma
    1. Idiopathic
    2. Pregnancy
II. Systemic diseases
  A. Linear/branching
    1. Carcinoid (head, neck, upper trunk)
    2. Ataxia-telangiectasia (bulbar conjunctivae, head and neck)
    3. Mastocytosis (within lesions)
  B. Poikiloderma
    1. Dermatomyositis, lupus erythematosus
    2. Mycosis fungoides, patch stage
    3. Genodermatoses, e.g., xeroderma pigmentosum, Kindler syndrome
  C. Mat
    1. Systemic sclerosis (scleroderma)
  D. Cuticular/periungual
    1. Lupus erythematosus
    2. Systemic sclerosis (scleroderma)
    3. Dermatomyositis
    4. Hereditary hemorrhagic telangiectasia
  E. Papular
    1. Hereditary hemorrhagic telangiectasia
  F. Spider angioma
    1. Cirrhosis

[a]Becoming less common.

Poikiloderma is a term used to describe a patch of skin with: (1) reticulated hypo- and hyperpigmentation, (2) wrinkling secondary to epidermal atrophy, and (3) telangiectasias. Poikiloderma does not imply a single disease entity—although it is becoming less common, it is seen in skin damaged by *ionizing radiation* as well as in patients with autoimmune connective tissue diseases, primarily *dermatomyositis* (DM), and rare genodermatoses (e.g., Kindler syndrome).

In *systemic sclerosis (scleroderma)* the dilated blood vessels have a unique configuration and are known as *mat telangiectasias*. The lesions are broad macules that usually measure 2–7 mm in diameter but occasionally are larger. Mats have a polygonal or oval shape, and their erythematous color may appear uniform, but, upon closer inspection, the erythema is the result of delicate telangiectasias. The most common locations for mat telangiectasias are the face, oral mucosa, and hands—peripheral sites that are prone to intermittent ischemia. The limited form of systemic sclerosis, often referred to as the CREST (*c*alcinosis cutis, *R*aynaud's phenomenon, *e*sophageal dysmotility, *s*clerodactyly, and *t*elangiectasia) variant (Chap. 353), is associated with a chronic course and anticentromere antibodies. Mat telangiectasias are an important clue to the diagnosis of this variant as well as the diffuse form of systemic sclerosis because they may be the only cutaneous finding.

Cuticular telangiectasias are pathognomonic signs of the three major autoimmune connective tissue diseases: *lupus erythematosus*, *systemic sclerosis*, and *DM*. They are easily visualized by the naked eye and occur in at least two-thirds of these patients. In both DM and lupus, there is associated nailfold erythema, and in DM, the erythema is often accompanied by "ragged" cuticles and fingertip tenderness. Under 10× magnification, the blood vessels in the nailfolds of lupus patients are tortuous and resemble "glomeruli," whereas in systemic sclerosis and DM, there is a loss of capillary loops and those that remain are markedly dilated.

In *hereditary hemorrhagic telangiectasia* (Osler-Rendu-Weber disease), the lesions usually appear during adolescence (mucosal) and adulthood (cutaneous) and are most commonly seen on the mucous membranes (nasal, orolabial), face, and distal extremities, including under the nails. They represent arteriovenous (AV) malformations of the dermal microvasculature, are dark red in color, and are usually slightly elevated. When the skin is stretched over an individual lesion, an eccentric punctum with radiating legs is seen. Although the degree of systemic involvement varies in this autosomal dominant disease (due primarily to mutations in either the endoglin or activin receptor–like kinase gene), the major symptoms are recurrent epistaxis and gastrointestinal bleeding. The fact that these mucosal telangiectasias are actually AV communications helps to explain their tendency to bleed.

HYPOPIGMENTATION

(Table 54-9) Disorders of hypopigmentation are often classified as either diffuse or localized. The classic example of *diffuse hypopigmentation* is *oculocutaneous albinism* (OCA). The most common forms are due to mutations in the tyrosinase gene (type I) or the *P* gene (type II); patients with type IA OCA have a total lack of enzyme activity. At birth, different forms of OCA can appear similar—white hair, gray-blue eyes, and pink-

white skin. However, the patients with no tyrosinase activity maintain this phenotype, whereas those with decreased activity will acquire some pigmentation of the eyes, hair, and skin as they age. The degree of pigment formation is also a function of racial background, and the pigmentary dilution is more readily apparent when patients are compared to their first-degree relatives. The ocular findings in OCA correlate with the degree of hypopigmentation and include decreased visual acuity, nystagmus, photophobia, strabismus, and a lack of normal binocular vision.

TABLE 54-9

Causes of Hypopigmentation

I. Primary cutaneous disorders
   A. Diffuse
      1. Generalized vitiligo[a]
   B. Localized
      1. Idiopathic guttate hypomelanosis
      2. Postinflammatory
      3. Pityriasis (tinea) versicolor
      4. Vitiligo[a]
      5. Chemical- or drug-induced leukoderma, e.g., topical imiquimod, oral imatinib
      6. Nevus depigmentosus
      7. Piebaldism[a]

II. Systemic diseases
   A. Diffuse
      1. Oculocutaneous albinism[b]
      2. Hermansky-Pudlak syndrome[b,c]
      3. Chédiak-Higashi syndrome[b,d]
      4. Phenylketonuria
   B. Localized
      1. Systemic sclerosis (scleroderma)
      2. Melanoma-associated leukoderma, spontaneous or immunotherapy-induced
      3. Vogt-Koyanagi-Harada syndrome
      4. Onchocerciasis
      5. Sarcoidosis
      6. Cutaneous T cell lymphoma (especially mycosis fungoides)
      7. Tuberculoid and indeterminate leprosy
      8. Linear nevoid hypopigmentation (hypomelanosis of Ito)[e]
      9. Incontinentia pigmenti (stage IV)
      10. Tuberous sclerosis
      11. Waardenburg syndrome and Shah-Waardenburg syndrome

[a]Absence of melanocytes in areas of leukoderma. [b]Normal number of melanocytes. [c]Platelet storage defect and restrictive lung disease secondary to deposits of ceroid-like material or immunodeficiency; due to mutations in β or δ subunit of adaptor-related protein complex 3 as well as subunits of *b*iogenesis of *l*ysosome-related *o*rganelles *c*omplex

(BLOC)-1, 2, and 3. [d]Giant lysosomal granules and recurrent infections. [e]Minority of patients in a nonreferral setting have systemic abnormalities (musculoskeletal, central nervous system, ocular).

The differential diagnosis of *localized hypomelanosis* includes the following primary cutaneous disorders: *idiopathic guttate hypomelanosis*, *postinflammatory hypopigmentation*, *pityriasis* (*tinea*) *versicolor*, *vitiligo*, *chemical- or drug-induced leukoderma*, *nevus depigmentosus* (see below), and *piebaldism* **(Table 54-10)**. In this group of diseases, the areas of involvement are macules or patches with a decrease or absence of pigmentation. Patients with vitiligo also have an increased incidence of several autoimmune disorders, including Hashimoto's thyroiditis, Graves' disease, pernicious anemia, Addison's disease, uveitis, alopecia areata, chronic mucocutaneous candidiasis, and the autoimmune polyendocrine syndromes (types I and II). Diseases of the thyroid gland are the most frequently associated disorders, occurring in up to 30% of patients with vitiligo. Circulating autoantibodies are often found, and the most common ones are antithyroglobulin, antimicrosomal, and antithyroid-stimulating hormone receptor antibodies.

TABLE 54-10

Hypopigmentation (Primary Cutaneous Disorders, Localized)

	Clinical Characteristics	Wood's Lamp Examination (UV-A; Peak = 365 nm)	Skin Biopsy Specimen	Pathogenesis	Treatment
Idiopathic guttate hypomelanosis	Common; acquired; usually 2–4 mm in diameter Shins and extensor forearms	Less enhancement than vitiligo	Abrupt decrease in epidermal melanin content	Possible somatic mutations as a reflection of aging or UV exposure	None
Postinflammatory hypopigmentation	Can develop within active lesions, as in subacute cutaneous lupus, or after the lesion fades, as in atopic dermatitis	Depends on particular disease Usually less enhancement than in vitiligo	Type of inflammatory infiltrate depends on specific disease	Block in transfer of melanin from melanocytes to keratinocytes could be secondary to edema or decrease in contact time Destruction of melanocytes if inflammatory cells attack basal layer of epidermis	Treat underlying inflammatory disease

	Clinical Characteristics	Wood's Lamp Examination (UV-A; Peak = 365 nm)	Skin Biopsy Specimen	Pathogenesis	Treatment
Pityriasis (tinea) versicolor	Common disorder Upper trunk and neck (shawl-like distribution), groin Young adults Macules have fine white scale when scratched	Golden fluorescence	Hyphal forms and budding yeast in stratum corneum	Invasion of stratum corneum by the yeast *Malassezia* Yeast is lipophilic and produces C_9 and C_{11} dicarboxylic acids, which *in vitro* inhibit tyrosinase	Selenium sulfide 2.5% shampoo; topical imidazoles; oral triazoles
Vitiligo	Acquired; progressive Symmetric areas of complete pigment loss Periorificial—around mouth, nose, eyes, nipples, umbilicus, anus Other areas—flexor wrists, extensor distal extremities Segmental form is less common—unilateral, dermatomal-like	More apparent Chalk-white	Absence of melanocytes in well-developed lesions Mild inflammation	Autoimmune phenomenon that results in destruction of melanocytes—primarily cellular (circulating skin-homing autoreactive T cells)	Topical glucocorticoids; topical calcineurin inhibitors; UV-B (narrowband); PUVA; JAK inhibitors transplants, if stable; depigmentation (topical MBEH), if widespread and treatment-resistant

	Clinical Characteristics	Wood's Lamp Examination (UV-A; Peak = 365 nm)	Skin Biopsy Specimen	Pathogenesis	Treatment
Chemical- or drug-induced leukoderma	Similar appearance to vitiligo Often begins on hands when associated with chemical exposure Satellite lesions in areas not exposed to chemicals	More apparent Chalk-white	Decreased number or absence of melanocytes	Exposure to chemicals that selectively destroy melanocytes, in particular phenols and catechols (germicides; rubber products) or ingestion of drugs such as imatinib Release of cellular antigens and activation of circulating lymphocytes may explain satellite phenomenon Possible inhibition of KIT receptor	Avoid exposure to offending agent, then treat as vitiligo Drug-induced variant may undergo repigmentation when medication is discontinued

	Clinical Characteristics	Wood's Lamp Examination (UV-A; Peak = 365 nm)	Skin Biopsy Specimen	Pathogenesis	Treatment
Piebaldism	Autosomal dominant Congenital, stable White forelock Areas of amelanosis contain normally pigmented and hyperpigmented macules of various sizes Symmetric involvement of central forehead, ventral trunk, and mid regions of upper and lower extremities	Enhancement of leukoderma and hyperpigmented macules	Amelanotic areas—few to no melanocytes	Defect in migration of melanoblasts from neural crest to involved skin or failure of melanoblasts to survive or differentiate in these areas Mutations within the *KIT* protooncogene that encodes the tyrosine kinase receptor for stem cell growth factor (kit ligand)	None; occasionally transplants

Abbreviations: MBEH, monobenzylether of hydroquinone; UV-B, ultraviolet B irradiation; PUVA, psoralens + ultraviolet A irradiation.

There are four systemic diseases that should be considered in a patient with skin findings suggestive of vitiligo—*Vogt-Koyanagi-Harada syndrome*, *systemic sclerosis*, *onchocerciasis*, and *melanoma-associated leukoderma*. A history of aseptic meningitis, nontraumatic uveitis, tinnitus, hearing loss, and/or dysacousia points to the diagnosis of the Vogt-Koyanagi-Harada syndrome. In these patients, the face and scalp are the most common locations of pigment loss. The vitiligo-like leukoderma seen in patients with systemic sclerosis has a clinical resemblance to idiopathic vitiligo that has begun to repigment as a result of treatment; that is, perifollicular macules of normal pigmentation are seen within areas of depigmentation. The basis of this leukoderma is unknown; there is no evidence of inflammation in areas of involvement, but it can resolve if the underlying connective tissue disease becomes inactive. In contrast to idiopathic vitiligo, melanoma-

associated leukoderma often begins on the trunk, and its appearance, if spontaneous, should prompt a search for metastatic disease. It is also seen in patients undergoing immunotherapy for melanoma, including ipilimumab, with cytotoxic T lymphocytes presumably recognizing cell surface antigens common to melanoma cells and melanocytes, and is associated with a greater likelihood of a clinical response.

There are two systemic disorders (neurocristopathies) that may have the cutaneous findings of piebaldism (Table 54-9). They are *Shah-Waardenburg syndrome* and *Waardenburg syndrome*. A possible explanation for both disorders is an abnormal embryonic migration or survival of two neural crest–derived elements, one of them being melanocytes and the other myenteric ganglion cells (leading to Hirschsprung disease in Shah-Waardenburg syndrome) or auditory nerve cells (Waardenburg syndrome). The latter syndrome is characterized by congenital sensorineural hearing loss, dystopia canthorum (lateral displacement of the inner canthi but normal interpupillary distance), heterochromic irises, and a broad nasal root, in addition to the piebaldism. The facial dysmorphism can be explained by the neural crest origin of the connective tissues of the head and neck. Patients with Waardenburg syndrome have been shown to have mutations in four genes, including *PAX-3* and *MITF*, all of which encode transcription factors, whereas patients with Hirschsprung disease plus white spotting have mutations in one of three genes—endothelin 3, endothelin B receptor, and *SOX-10*.

In *tuberous sclerosis*, the earliest cutaneous sign is macular hypomelanosis, referred to as an ash leaf spot. These lesions are often present at birth and are usually multiple; however, detection may require Wood's lamp examination, especially in fair-skinned individuals. The pigment within them is reduced, but not absent. The average size is 1–3 cm, and the common shapes are polygonal and lance-ovate. Examination of the patient for additional cutaneous signs such as multiple angiofibromas of the face (adenoma sebaceum), ungual and intraoral fibromas, fibrous cephalic plaques, and connective tissue nevi (shagreen patches) is recommended. It is important to remember that an ash leaf spot on the scalp will result in a circumscribed patch of lightly pigmented hair. Internal manifestations include seizures, intellectual disability, central nervous system (CNS) and retinal hamartomas, pulmonary lymphangioleiomyomatosis (women), renal angiomyolipomas, and cardiac rhabdomyomas. The latter can be detected in up to 60% of children (<18 years) with tuberous sclerosis by echocardiography.

Nevus depigmentosus is a stable, well-circumscribed hypomelanosis that is present at birth. There is usually a single oval or rectangular lesion, but when there are multiple lesions, the possibility of tuberous sclerosis needs to be considered. In *linear nevoid hypopigmentation*, a term that is replacing hypomelanosis of Ito and segmental or systematized nevus depigmentosus, streaks and swirls of hypopigmentation are observed. Up to one-third of patients in a tertiary care setting had associated abnormalities involving the musculoskeletal system (asymmetry), the CNS (seizures and intellectual disability), and the eyes (strabismus and hypertelorism). Chromosomal mosaicism has been detected in these patients, lending support to the hypothesis that the cutaneous pattern is the result of the migration of two clones of primordial melanocytes, each with a different pigment potential.

Localized areas of decreased pigmentation are commonly seen as a result of cutaneous inflammation (Table 54-10) and have been observed in the skin overlying active lesions of sarcoidosis (see "Papulonodular Skin

Lesions," below) as well as in CTCL. Cutaneous infections also present as disorders of hypopigmentation, and in *tuberculoid leprosy*, there are a few asymmetric patches of hypomelanosis that have associated anesthesia, anhidrosis, and alopecia. Biopsy specimens of the palpable border show dermal granulomas that contain rare, if any, *Mycobacterium leprae* organisms.

HYPERPIGMENTATION

(Table 54-11) Disorders of hyperpigmentation are also divided into two major groups—localized and diffuse. The localized forms are due to an epidermal alteration, a proliferation of melanocytes, or an increase in pigment production. Both seborrheic keratoses and acanthosis nigricans belong to the first group. *Seborrheic keratoses* are common lesions, but in one rare clinical setting, they are a sign of systemic disease, and that setting is the sudden appearance of multiple lesions, often with an inflammatory base and in association with acrochordons (skin tags) and acanthosis nigricans. This is termed the *sign of Leser-Trélat* and alerts the clinician to search for an internal malignancy. *Acanthosis nigricans* can also be a reflection of an internal malignancy, most commonly of the gastrointestinal tract, and it appears as velvety hyperpigmentation, primarily in flexural areas. However, in the majority of patients, acanthosis nigricans is associated with obesity and insulin resistance, although it may be a reflection of an endocrinopathy such as acromegaly, Cushing's syndrome, polycystic ovary syndrome, or insulin-resistant diabetes mellitus (type A, type B, and lipodystrophic forms).

TABLE 54-11

Causes of Hyperpigmentation

I. Primary cutaneous disorders
  A. Localized
    1. Epidermal alteration
      a. Seborrheic keratosis
      b. Pigmented actinic keratosis
    2. Proliferation of melanocytes
      a. Lentigo
      b. Melanocytic nevus (mole)
      c. Melanoma
    3. Increased pigment production
      a. Ephelide (freckle)
      b. Café au lait macule
      c. Postinflammatory hyperpigmentation
      d. Melasma
    4. Dermal pigmentation
      a. Fixed drug eruption
  B. Localized and diffuse
    1. Drugs (e.g., minocycline, hydroxychloroquine, bleomycin)

II. Systemic diseases
  A. Localized
    1. Epidermal alteration
      a. Seborrheic keratoses (sign of Leser-Trélat)
      b. Acanthosis nigricans (insulin resistance, other endocrine disorders, paraneoplastic)
    2. Proliferation of melanocytes
      a. Lentigines (Peutz-Jeghers and LEOPARD/Noonan with multiple lentigines syndromes; xeroderma pigmentosum)
      b. Melanocytic nevi (Carney complex [LAMB and NAME syndromes])[a]
    3. Increased pigment production
      a. Café au lait macules (neurofibromatosis, McCune-Albright syndrome[b])
      b. Urticaria pigmentosa[c]
    4. Dermal pigmentation
      a. Incontinentia pigmenti (stage III)
      b. Dyskeratosis congenita
  B. Diffuse
    1. Endocrinopathies
      a. Addison’s disease

b. Nelson syndrome
c. Ectopic ACTH syndrome
d. Hyperthyroidism

2. Metabolic
   a. Porphyria cutanea tarda
   b. Hemochromatosis
   c. Vitamin B_{12}, folate deficiency
   d. Pellagra
   e. Malabsorption, including Whipple's disease
3. Melanosis secondary to metastatic melanoma
4. Autoimmune
   a. Biliary cirrhosis
   b. Systemic sclerosis (scleroderma)
   c. POEMS syndrome
   d. Eosinophilia-myalgia syndrome[d]
5. Drugs (e.g. cyclophosphamide) and metals (e.g. silver)

[a]Also lentigines. [b]Polyostotic fibrous dysplasia. [c]See also "Papulonodular Skin Lesions." [d]Late 1980s.

Abbreviations: LAMB, *l*entigines, *a*trial *m*yxomas, *m*ucocutaneous myxomas, and *b*lue nevi; LEOPARD, *l*entigines, *E*CG abnormalities, *o*cular hypertelorism, *p*ulmonary stenosis and subaortic valvular stenosis, *a*bnormal genitalia, *r*etardation of growth, and *d*eafness (sensorineural); NAME, *n*evi, *a*trial myxoma, *m*yxoid neurofibroma, and *e*phelides (freckles); POEMS, *p*olyneuropathy, *o*rganomegaly, *e*ndocrinopathies, *M*-protein, and *s*kin changes.

A proliferation of melanocytes results in the following pigmented lesions: *lentigo*, *melanocytic nevus*, and *melanoma* (Chap. 72). In an adult, the majority of lentigines are related to sun exposure, which explains their distribution. However, in the Peutz-Jeghers and LEOPARD (*l*entigines; *E*CG abnormalities, primarily conduction defects; *o*cular hypertelorism; *p*ulmonary stenosis and subaortic valvular stenosis; *a*bnormal genitalia [cryptorchidism, hypospadias]; *r*etardation of growth; and *d*eafness [sensorineural]) syndromes, lentigines do serve as a clue to systemic disease. In *LEOPARD/Noonan with multiple lentigines syndrome*, hundreds of lentigines develop during childhood and are scattered over the entire surface of the body. The lentigines in patients with *Peutz-Jeghers syndrome* are located primarily around the nose and mouth, on the hands and feet, and within the oral cavity. While the pigmented macules on the face may fade with age, the oral lesions persist. However, similar intraoral lesions are also seen in Addison's disease, in Laugier-Hunziker syndrome (no internal manifestations), and as a normal finding in darkly pigmented individuals. Patients with this autosomal dominant syndrome (due to mutations in a novel serine threonine kinase gene) have multiple benign polyps of the gastrointestinal tract, testicular or ovarian tumors, and an increased risk of developing gastrointestinal (primarily colon) and pancreatic cancers.

In the *Carney complex,* numerous lentigines are also seen, but they are in association with cardiac myxomas. This autosomal dominant disorder is also known as the *LAMB* (*l*entigines, *a*trial myxomas, *m*ucocutaneous myxomas, and *b*lue nevi) *syndrome* or *NAME* (*n*evi, *a*trial myxoma, *m*yxoid neurofibroma, and *e*phelides [freckles]) *syndrome*. These patients can also have evidence of endocrine overactivity in the form of Cushing's syndrome (pigmented nodular adrenocortical disease) and acromegaly.

The third type of localized hyperpigmentation is due to a local increase in pigment production, and it includes *ephelides* and *café au lait macules* (CALMs). While a single CALM can be seen in up to 10% of the normal population, the presence of multiple or large-sized CALMs raises the possibility of an associated genodermatosis, for example, neurofibromatosis (NF) or McCune-Albright syndrome. *CALMs* are flat, uniformly brown in color (usually two shades darker than uninvolved skin), and can vary in size from 0.5 to 12+ cm. More than 90% of adult patients with *type I NF* will have six or more CALMs measuring ≥1.5 cm in diameter. Additional findings are discussed in the section on neurofibromas (see "Papulonodular Skin Lesions," below). In comparison with NF, the CALMs in patients with *McCune-Albright syndrome* (polyostotic fibrous dysplasia with precocious puberty in females due to mosaicism for an activating mutation in a G protein [$G_s\alpha$] gene) are usually larger, are more irregular in outline, and tend to respect the midline.

In incontinentia pigmenti, dyskeratosis congenita, and bleomycin pigmentation, the areas of localized hyperpigmentation form a pattern—swirled in the first, reticulated in the second, and flagellate in the third. In *dyskeratosis congenita*, atrophic reticulated hyperpigmentation is seen on the neck, trunk, and thighs and is accompanied by nail dystrophy, pancytopenia, and leukoplakia of the oral and anal mucosae. The latter often develops into squamous cell carcinoma. In addition to the flagellate pigmentation (linear streaks) on the trunk, patients receiving bleomycin often have hyperpigmentation overlying the elbows, knees, and small joints of the hand.

Localized hyperpigmentation is seen as a side effect of several other *systemic medications*, including those that produce fixed drug reactions (nonsteroidal anti-inflammatory drugs [NSAIDs], sulfonamides, barbiturates, and tetracyclines) and those that can complex with melanin or iron (antimalarials and minocycline). Fixed drug eruptions recur in the exact same location as circular areas of erythema that can become bullous and then resolve as brown macules. The eruption usually appears within hours of re-administration of the offending agent, and common locations include the genitalia, distal extremities, and perioral region. Chloroquine and hydroxychloroquine produce gray-brown to blue-black discoloration of the shins, hard palate, and face, while blue macules (often misdiagnosed as bruises) can be seen on the lower extremities and in sites of inflammation with prolonged minocycline administration. Estrogen in oral contraceptives can induce melasma—symmetric brown patches on the face, especially the cheeks, upper lip, and forehead. Similar changes are seen in pregnancy and in patients receiving phenytoin.

In the diffuse forms of hyperpigmentation, the darkening of the skin may be of equal intensity over the entire body or may be accentuated in sun-exposed areas. The causes of diffuse hyperpigmentation can be divided into four major groups—endocrine, metabolic, autoimmune, and drugs. The endocrinopathies that frequently have associated hyperpigmentation include *Addison's disease*, *Nelson's syndrome*, and *ectopic*

ACTH syndrome. In these diseases, the increased pigmentation is diffuse but is accentuated in sun-exposed areas, as well as in the palmar creases, sites of friction, and scars. An overproduction of the pituitary hormones α-MSH (melanocyte-stimulating hormone) and ACTH can lead to an increase in melanocyte activity. These peptides are products of the proopiomelanocortin gene and exhibit homology, for example, α-MSH and ACTH share 13 amino acids. A minority of patients with Cushing's disease or hyperthyroidism have generalized hyperpigmentation.

The metabolic causes of hyperpigmentation include *porphyria cutanea tarda* (PCT), *hemochromatosis*, *vitamin B_{12} deficiency*, *folic acid deficiency*, *pellagra*, and *malabsorption*, including *Whipple's disease*. In patients with PCT (see "Vesicles/Bullae," below), the skin darkening is seen in sun-exposed areas and is a reflection of the photoreactive properties of porphyrins. The increased level of iron in the skin of patients with type 1 hemochromatosis stimulates melanin pigment production and leads to the classic bronze color. Patients with pellagra have a brown discoloration of the skin, especially in sun-exposed areas, as a result of nicotinic acid (niacin) deficiency. In the areas of increased pigmentation, there is a thin, varnish-like scale. These changes are also seen in patients who are vitamin B_6 deficient, have functioning carcinoid tumors (increased consumption of niacin), or take isoniazid. Approximately 50% of the patients with Whipple's disease have an associated generalized hyperpigmentation in association with diarrhea, weight loss, arthritis, and lymphadenopathy. A diffuse, slate-blue to gray-brown color is seen in patients with *melanosis secondary to metastatic melanoma*. The color reflects widespread deposition of melanin within the dermis as a result of the high concentration of circulating melanin precursors.

Of the autoimmune diseases associated with diffuse hyperpigmentation, *biliary cirrhosis* and *systemic sclerosis* are the most common, and occasionally, both disorders are seen in the same patient. The skin is dark brown in color, especially in sun-exposed areas. In biliary cirrhosis, the hyperpigmentation is accompanied by pruritus, jaundice, and xanthomas, whereas in systemic sclerosis, it is accompanied by sclerosis of the extremities, face, and, less commonly, the trunk. Additional clues to the diagnosis of systemic sclerosis are mat and cuticular telangiectasias, calcinosis cutis, Raynaud's phenomenon, and distal ulcerations (see "Telangiectasias," above). The differential diagnosis of cutaneous sclerosis with hyperpigmentation includes POEMS (*p*olyneuropathy; *o*rganomegaly [liver, spleen, lymph nodes]; *e*ndocrinopathies [impotence, gynecomastia]; *M*-protein; and *s*kin changes) syndrome. The skin changes include hyperpigmentation, induration, hypertrichosis, angiomas, clubbing, and facial lipoatrophy.

Diffuse hyperpigmentation that is due to drugs or metals can result from one of several mechanisms—induction of melanin pigment formation, complexing of the drug or its metabolites to melanin, and deposits of the drug in the dermis. Busulfan, cyclophosphamide, 5-fluorouracil, and inorganic arsenic induce pigment production. Complexes containing melanin or iron plus the drug or its metabolites are seen in patients receiving minocycline, and a diffuse, brown-gray, muddy appearance within sun-exposed areas may develop, in addition to pigmentation of the mucous membranes, teeth, nails, bones, and thyroid. Administration of amiodarone can result in both a phototoxic eruption (exaggerated sunburn) and/or a slate-gray to violaceous discoloration of sun-exposed skin. Biopsy specimens of the latter show yellow-brown granules in dermal macrophages, which represent intralysosomal accumulations of lipids, amiodarone, and its metabolites.

Actual deposits of a particular drug or metal in the skin are seen with silver (argyria), where the skin appears blue-gray in color; gold (chrysiasis), where the skin has a brown to blue-gray color; and clofazimine, where the skin appears reddish brown. The associated pigmentation is accentuated in sun-exposed areas, and discoloration of the eye is seen with gold (sclerae) and clofazimine (conjunctivae).

VESICLES/BULLAE

(Table 54-12) Depending on their size, cutaneous blisters are referred to as *vesicles* (<1 cm) or *bullae* (>1 cm). The primary autoimmune blistering disorders include *pemphigus vulgaris*, *pemphigus foliaceus*, *paraneoplastic pemphigus*, *bullous pemphigoid*, *gestational pemphigoid*, *cicatricial pemphigoid*, *epidermolysis bullosa acquisita*, *linear IgA bullous dermatosis* (*LABD*), and *dermatitis herpetiformis* (Chap. 55).

TABLE 54-12

Causes of Vesicles/Bullae

I. Primary mucocutaneous diseases
  A. Primary blistering diseases (autoimmune)
    1. Pemphigus, foliaceus and vulgaris[a]
    2. Bullous pemphigoid[b]
    3. Gestational pemphigoid[b]
    4. Cicatricial pemphigoid[b]
    5. Dermatitis herpetiformis[b,c]
    6. Linear IgA bullous dermatosis[b]
    7. Epidermolysis bullosa acquisita[b,d]
  B. Secondary blistering diseases
    1. Contact dermatitis[a,b]
    2. Erythema multiforme[e]
    3. Stevens-Johnson syndrome[e]
    4. Toxic epidermal necrolysis[e]
  C. Infections
    1. Varicella-zoster virus[a,f]
    2. Herpes simplex virus[a,f]
    3. Enteroviruses, e.g., hand-foot-and-mouth disease[f]
    4. Staphylococcal scalded-skin syndrome[a,g]
    5. Bullous impetigo[a]
II. Systemic diseases
  A. Autoimmune
    1. Paraneoplastic pemphigus[a]
  B. Infections
    1. Cutaneous emboli[b]
  C. Metabolic
    1. Diabetic bullae[a,b]
    2. Porphyria cutanea tarda[b]
    3. Porphyria variegata[b]
    4. Pseudoporphyria[b]
    5. Bullous dermatosis of hemodialysis[b]
  D. Ischemia
    1. Coma bullae

E. Secondary blistering diseases

1. Toxic epidermal necrolysis[e] (respiratory and GI tracts can be involved)

[a]Intraepidermal. [b]Subepidermal. [c]Associated with gluten enteropathy. [d]Associated with inflammatory bowel disease.

[e]Degeneration of cells within the basal layer of the epidermis can give impression split is subepidermal. [f]Also systemic.

[g]In adults, associated with renal failure and immunocompromised state.

Vesicles and bullae are also seen in *contact dermatitis*, both allergic and irritant forms **(Chap. 53)**. When there is a linear arrangement of vesicular lesions, an exogenous cause or herpes zoster should be suspected. Bullous disease secondary to the ingestion of drugs can take one of several forms, including phototoxic eruptions, isolated bullae, Stevens-Johnson syndrome (SJS), and toxic epidermal necrolysis (TEN) **(Chap. 56)**. Clinically, phototoxic eruptions resemble an exaggerated sunburn with diffuse erythema and bullae in sun-exposed areas. The most commonly associated drugs are doxycycline, quinolones, thiazides, NSAIDs, voriconazole, and psoralens. The development of a phototoxic eruption is dependent on the doses of both the drug and ultraviolet (UV)-A irradiation.

Toxic epidermal necrolysis is characterized by bullae that arise on widespread areas of tender erythema and then slough. This results in large areas of denuded skin. The associated morbidity, such as sepsis, and mortality rates are relatively high and are a function of the extent of epidermal necrosis. In addition, these patients may also have involvement of the mucous membranes and respiratory and intestinal tracts. Drugs are the primary cause of TEN, and the most common offenders are aromatic anticonvulsants (phenytoin, barbiturates, carbamazepine), sulfonamides, aminopenicillins, allopurinol, and NSAIDs. Severe acute graft-versus-host disease (grade 4), vancomycin-induced LABD, and flares of lupus can also resemble TEN.

In *erythema multiforme* (EM), the primary lesions are pink-red macules and edematous papules, the centers of which may become vesicular. In contrast to a morbilliform exanthem, the clue to the diagnosis of EM, and especially SJS, is the development of a “dusky” violet color in the center of the lesions. Target lesions are also characteristic of EM and arise as a result of active centers and borders in combination with centrifugal spread. However, target lesions need not be present to make the diagnosis of EM.

EM has been subdivided into two major groups: (1) EM minor due to herpes simplex virus (HSV); and (2) EM major due to HSV, *Mycoplasma pneumonia*, or, occasionally, drugs. Involvement of the mucous membranes (ocular, nasal, oral, and genital) is seen more commonly in the latter form. Hemorrhagic crusts of the lips are characteristic of EM major and SJS as well as herpes simplex, pemphigus vulgaris, and paraneoplastic pemphigus. Fever, malaise, myalgias, sore throat, and cough may precede or accompany the eruption. The lesions of EM usually resolve over 2–4 weeks but may be recurrent, especially when due to HSV. In addition to HSV (in which lesions usually appear 7–12 days after the viral eruption), EM can also follow vaccinations, radiation therapy, and exposure to environmental toxins, including the oleoresin in poison ivy.

Induction of SJS is most often due to drugs, especially sulfonamides, phenytoin, barbiturates, lamotrigine, aminopenicillins, nonnucleoside reverse transcriptase inhibitors (e.g., nevirapine), and carbamazepine. Widespread dusky macules and significant mucosal involvement are characteristic of SJS, and the cutaneous lesions may or may not develop epidermal detachment. If the latter occurs, by definition, it is limited to <10% of the body surface area (BSA). Greater involvement leads to the diagnosis of SJS/TEN overlap (10–30% BSA) or TEN (>30% BSA).

In addition to primary blistering disorders and hypersensitivity reactions, bacterial and viral infections can lead to vesicles and bullae. The most common infectious agents are HSV **(Chap. 187)**, varicella-zoster virus **(Chap. 188)**, and *S. aureus* **(Chap. 142)**.

Staphylococcal scalded-skin syndrome (SSSS) and *bullous impetigo* are two blistering disorders associated with staphylococcal (phage group II) infection. In SSSS, the initial findings are redness and tenderness of the central face, neck, trunk, and intertriginous zones. This is followed by short-lived flaccid bullae and a slough or exfoliation of the superficial epidermis. Crusted areas then develop, characteristically around the mouth in a radial pattern. SSSS is distinguished from TEN by the following features: younger age group (primarily infants), more superficial site of blister formation, no oral lesions, shorter course, lower morbidity and mortality rates, and an association with staphylococcal exfoliative toxin (“exfoliatin”), not drugs. A rapid diagnosis of SSSS versus TEN can be made by a frozen section of the blister roof or exfoliative cytology of the blister contents. In SSSS, the site of staphylococcal infection is usually extracutaneous (conjunctivitis, rhinorrhea, otitis media, pharyngitis, tonsillitis), and the cutaneous lesions are sterile, whereas in bullous impetigo, the skin lesions are the site of infection. Impetigo is more localized than SSSS and usually presents with honey-colored crusts. Occasionally, superficial purulent blisters also form. *Cutaneous emboli* from gram-negative infections may present as isolated bullae, but the base of the lesion is purpuric or necrotic, and it may develop into an ulcer (see “Purpura,” below).

Several metabolic disorders are associated with blister formation, including diabetes mellitus, renal failure, and porphyria. Local hypoxemia secondary to decreased cutaneous blood flow can also produce blisters, which explains the presence of bullae over pressure points in comatose patients (coma bullae). In *diabetes mellitus*, tense bullae with clear sterile viscous fluid arise on normal skin. The lesions can be as large as 6 cm in diameter and are located on the distal extremities. There are several types of porphyria, but the most common form with cutaneous findings is *porphyria cutanea tarda* (PCT). In sun-exposed areas (primarily the hands), the skin is very fragile, with trauma leading to erosions mixed with tense vesicles. These lesions then heal with scarring and formation of milia; the latter are firm, 1- to 2-mm white or yellow papules that represent epidermoid inclusion cysts. Associated findings can include hypertrichosis of the lateral malar region (men) or face (women) and, in sun-exposed areas, hyperpigmentation and firm sclerotic plaques. An elevated level of urinary uroporphyrins confirms the diagnosis and is due to a decrease in uroporphyrinogen decarboxylase activity. PCT can be exacerbated by alcohol, hemochromatosis and other forms of iron overload, chlorinated hydrocarbons, hepatitis C virus and HIV infections, and hepatomas.

The differential diagnosis of PCT includes (1) *porphyria variegata*—the skin signs of PCT plus the systemic findings of acute intermittent porphyria; it has a diagnostic plasma porphyrin fluorescence emission at 626

nm; (2) *drug-induced pseudoporphyria*—the clinical and histologic findings are similar to PCT, but porphyrins are normal; etiologic agents include naproxen and other NSAIDs, furosemide, tetracycline, and voriconazole; (3) *bullous dermatosis of hemodialysis*—the same appearance as PCT, but porphyrins are usually normal or occasionally borderline elevated; patients have chronic renal failure and are on hemodialysis; (4) *PCT associated with hepatomas and hemodialysis;* and (5) *epidermolysis bullosa acquisita* (Chap. 55).

EXANTHEMS

(Table 54-13) Exanthems are characterized by an acute generalized eruption. The most common presentation is erythematous macules and papules (morbilliform) and less often confluent blanching erythema (scarlatiniform). *Morbilliform* eruptions are usually due to either drugs or viral infections. For example, up to 5% of patients receiving penicillins, sulfonamides, phenytoin, or nevirapine will develop a maculopapular eruption. Accompanying signs may include pruritus, fever, eosinophilia, and transient lymphadenopathy. Similar maculopapular eruptions are seen in the classic childhood viral exanthems, including (1) *rubeola* (measles)—a prodrome of coryza, cough, and conjunctivitis followed by Koplik's spots on the buccal mucosa; the eruption begins behind the ears, at the hairline, and on the forehead and then spreads down the body, often becoming confluent; (2) *rubella*—the eruption begins on the forehead and face and then spreads down the body; it resolves in the same order and is associated with retroauricular and suboccipital lymphadenopathy; and (3) *erythema infectiosum* (fifth disease)—erythema of the cheeks is followed by a reticulated pattern on the extremities; it is secondary to a parvovirus B19 infection, and an associated arthritis is seen in adults.

TABLE 54-13

Causes of Exanthems

I. Morbilliform
- A. Drugs
- B. Viral
  1. Rubeola (measles)
  2. Rubella
  3. Erythema infectiosum (erythema of cheeks; reticulated on extremities)
  4. Epstein-Barr virus, echovirus, coxsackievirus, CMV, adenovirus, HHV-6/HHV-7[a], dengue virus, Zika virus, Chikungunya, and West Nile virus infections
  5. HIV seroconversion exanthem (plus mucosal ulcerations)
- C. Bacterial
  1. Typhoid fever
  2. Early secondary syphilis
  3. Early *Rickettsia* infections
  4. Early meningococcemia
  5. Ehrlichiosis
- D. Acute graft-versus-host disease
- E. Kawasaki disease

II. Scarlatiniform
- A. Scarlet fever
- B. Toxic shock syndrome
- C. Kawasaki disease
- D. Early staphylococcal scalded-skin syndrome

[a]Primary infection in infants and reactivation in the setting of immunosuppression.

Abbreviations: CMV, cytomegalovirus; HHV, human herpesvirus; HIV, human immunodeficiency virus.

Both measles and rubella can occur in unvaccinated adults, and an atypical form of measles is seen in adults immunized with either killed measles vaccine or killed vaccine followed in time by live vaccine. In contrast to classic measles, the eruption of atypical measles begins on the palms, soles, wrists, and ankles, and the lesions may become purpuric. The patient with atypical measles can have pulmonary involvement and be quite ill. Rubelliform and roseoliform eruptions are also associated with *Epstein-Barr virus* (5–15% of patients), *echovirus*, *coxsackievirus*, *cytomegalovirus*, *adenovirus*, *dengue virus*, *Zika virus*, and *West Nile virus* infections. Detection of specific IgM antibodies or fourfold elevations in IgG antibodies often allows the proper diagnosis, but polymerase chain reaction (PCR) is gradually replacing serologic assays. Occasionally,

a maculopapular drug eruption is a reflection of an underlying viral infection. For example, ~95% of the patients with infectious mononucleosis who are given ampicillin will develop a rash.

Of note, early in the course of infections with *Rickettsia* and meningococcus, prior to the development of petechiae and purpura, the lesions may be erythematous macules and papules. This is also the case in chickenpox prior to the development of vesicles. Maculopapular eruptions are associated with early *HIV infection*, early secondary *syphilis*, *typhoid fever*, and *acute graft-versus-host disease*. In the last, lesions frequently begin on the dorsal hands and forearms; the macular rose spots of typhoid fever involve primarily the anterior trunk.

The prototypic *scarlatiniform* eruption is seen in *scarlet fever* and is due to an erythrogenic toxin produced by bacteriophage-containing group A β-hemolytic streptococci, most commonly in the setting of pharyngitis. This eruption is characterized by diffuse erythema, which begins on the neck and upper trunk, and red follicular puncta. Additional findings include a white strawberry tongue (white coating with red papillae) followed by a red strawberry tongue (red tongue with red papillae); petechiae of the palate; a facial flush with circumoral pallor; linear petechiae in the antecubital fossae; and desquamation of the involved skin, palms, and soles 5–20 days after onset of the eruption. A similar desquamation of the palms and soles is seen in toxic shock syndrome (TSS), in Kawasaki disease, and after severe febrile illnesses. Certain strains of staphylococci also produce an erythrotoxin that leads to the same clinical findings as in streptococcal scarlet fever, except that the anti-streptolysin O or DNase B titers are not elevated.

In *toxic shock syndrome*, staphylococcal (phage group I) infections produce an exotoxin (TSST-1) that causes the fever and rash as well as enterotoxins. Initially, the majority of cases were reported in menstruating women who were using tampons. However, other sites of infection, including wounds and nasal packing, can lead to TSS. The diagnosis of TSS is based on clinical criteria **(Chap. 142)**, and three of these involve mucocutaneous sites (diffuse erythema of the skin, desquamation of the palms and soles 1–2 weeks after onset of illness, and involvement of the mucous membranes). The latter is characterized as hyperemia of the vagina, oropharynx, or conjunctivae. Similar systemic findings have been described in *streptococcal toxic shock syndrome* **(Chap. 143)**, and although an exanthem is seen less often than in TSS due to a staphylococcal infection, the underlying infection is often in the soft tissue (e.g., cellulitis).

The cutaneous eruption in *Kawasaki disease* **(Chap. 356)** is polymorphous, but the two most common forms are morbilliform and scarlatiniform. Additional mucocutaneous findings include bilateral conjunctival injection; erythema and edema of the hands and feet followed by desquamation; and diffuse erythema of the oropharynx, red strawberry tongue, and dry fissured lips. This clinical picture can resemble TSS and scarlet fever, but clues to the diagnosis of Kawasaki disease are cervical lymphadenopathy, cheilitis, and thrombocytosis. The most serious associated systemic finding in this disease is coronary aneurysms secondary to arteritis. Scarlatiniform eruptions are also seen in the early phase of SSSS (see "Vesicles/Bullae," above), in young adults with *Arcanobacterium haemolyticum* infection, and as reactions to drugs.

URTICARIA

(Table 54-14) *Urticaria* (hives) are transient lesions that are composed of a central wheal surrounded by an erythematous halo or flare. Individual lesions are round, oval, or figurate and are often pruritic. Acute and chronic urticarias have a wide variety of allergic etiologies and reflect edema in the dermis. Urticarial lesions can also be seen in patients with mastocytosis (urticaria pigmentosa), hypo- or hyperthyroidism, Schnitzler's syndrome, and systemic-onset juvenile idiopathic arthritis (Still's disease). In both juvenile- and adult-onset Still's disease, the lesions coincide with the fever spike, are transient, and are due to dermal infiltrates of neutrophils.

TABLE 54-14

Causes of Urticaria and Angioedema

- I. Primary cutaneous disorders
  - A. Acute and chronic urticaria[a]
  - B. Physical urticaria
    - 1. Dermographism
    - 2. Solar urticaria[b]
    - 3. Cold urticaria[b]
    - 4. Cholinergic urticaria[b]
  - C. Angioedema (hereditary and acquired)[b,c]
- II. Systemic diseases
  - A. Urticarial vasculitis
  - B. Hepatitis B or C viral infection
  - C. Serum sickness
  - D. Angioedema (hereditary and acquired)

[a]A small minority develop anaphylaxis. [b]Also systemic. [c]Acquired angioedema can be idiopathic, associated with a lymphoproliferative disorder, or due to a drug, e.g., angiotensin-converting enzyme (ACE) inhibitors.

The common *physical urticarias* include dermographism, solar urticaria, cold urticaria, and cholinergic urticaria. Patients with *dermo-graphism* exhibit linear wheals following minor pressure or scratching of the skin. It is a common disorder, affecting ~5% of the population. *Solar urticaria* characteristically occurs within minutes of sun exposure and is a skin sign of one systemic disease—erythropoietic protoporphyria. In addition to the urticaria, these patients have subtle pitted scarring of the nose and hands. *Cold urticaria* is precipitated by exposure to the cold, and therefore exposed areas are usually affected. In occasional patients, the disease is associated with abnormal circulating proteins—more commonly cryoglobulins and less commonly cryofibrinogens. Additional systemic symptoms include wheezing and syncope, thus explaining the need for these patients to avoid swimming in cold water. Autosomal dominantly inherited cold urticaria is associated with dysfunction of cryopyrin. *Cholinergic urticaria* is precipitated by heat, exercise, or

emotion and is characterized by small wheals with relatively large flares. It is occasionally associated with wheezing.

Whereas urticarias are the result of dermal edema, subcutaneous edema leads to the clinical picture of *angioedema*. Sites of involvement include the eyelids, lips, tongue, larynx, and gastrointestinal tract as well as the subcutaneous tissue. Angioedema occurs alone or in combination with urticaria, including urticarial vasculitis and the physical urticarias. Both acquired and hereditary (autosomal dominant) forms of angioedema occur (Chap. 347), and in the latter, urticaria is rarely, if ever, seen.

Urticarial vasculitis is an immune complex disease that may be confused with simple urticaria. In contrast to simple urticaria, individual lesions tend to last longer than 24 h and usually develop central petechiae that can be observed even after the urticarial phase has resolved. The patient may also complain of burning rather than pruritus. On biopsy, there is a leukocytoclastic vasculitis of the small dermal blood vessels. Although urticarial vasculitis may be idiopathic in origin, it can be a reflection of an underlying systemic illness such as lupus erythematosus, Sjögren's syndrome, or hereditary complement deficiency. There is a spectrum of urticarial vasculitis that ranges from purely cutaneous to multisystem involvement. The most common systemic signs and symptoms are arthralgias and/or arthritis, nephritis, and crampy abdominal pain, with asthma and chronic obstructive lung disease seen less often. Hypocomplementemia occurs in one- to two-thirds of patients, even in the idiopathic cases. Urticarial vasculitis can also be seen in patients with *hepatitis B* and *hepatitis C* infections, *serum sickness*, and *serum sickness–like illnesses* (e.g., due to cefaclor, minocycline).

PAPULONODULAR SKIN LESIONS

(Table 54-15) In the *papulonodular diseases*, the lesions are elevated above the surface of the skin and may coalesce to form larger plaques. The location, consistency, and color of the lesions are the keys to their diagnosis; this section is organized on the basis of color.

TABLE 54-15

Papulonodular Skin Lesions According to Color Groups

I. White
- A. Calcinosis cutis
- B. Osteoma cutis (also skin-colored or blue)

II. Skin-colored
- A. Rheumatoid nodules
- B. Neurofibromas (von Recklinghausen's disease [NF1])
- C. Angiofibromas (tuberous sclerosis, MEN syndrome, type 1)
- D. Neuromas (MEN syndrome, type 2b)
- E. Adnexal tumors
  1. Basal cell carcinomas (basal cell nevus syndrome)
  2. Tricholemmomas (Cowden disease)
- F. Osteomas (arise in skull and jaw in Gardner syndrome)
- G. Primary cutaneous disorders
  1. Epidermal inclusion cysts[a]
  2. Lipomas

III. Pink/translucent[b]
- A. Amyloidosis, primary systemic
- B. Papular mucinosis/scleromyxedema
- C. Multicentric reticulohistiocytosis

IV. Yellow
- A. Xanthomas
- B. Tophi
- C. Necrobiosis lipoidica
- D. Pseudoxanthoma elasticum
- E. Sebaceous adenomas (Muir-Torre syndrome)

V. Red[b]
- A. Papules
  1. Angiokeratomas (Fabry disease)
  2. Bacillary angiomatosis (primarily in AIDS)
- B. Papules/plaques
  1. Cutaneous lupus
  2. Lymphoma cutis
  3. Leukemia cutis
  4. Sweet syndrome
- C. Nodules
  1. Panniculitis

2. Medium-sized vessel vasculitis (e.g., cutaneous polyarteritis nodosa)

D. Primary cutaneous disorders

1. Arthropod bites
2. Cherry hemangiomas
3. Infections, e.g., streptococcal cellulitis, sporotrichosis
4. Polymorphous light eruption
5. Cutaneous lymphoid hyperplasia (lymphocytoma cutis, pseudolymphoma)

VI. Red-brown[b]

A. Sarcoidosis
B. Urticaria pigmentosa
C. Erythema elevatum diutinum (chronic leukocytoclastic vasculitis)
D. Lupus vulgaris

VII. Blue[b]

A. Venous malformations (e.g., blue rubber bleb syndrome)
B. Primary cutaneous disorders

1. Venous lake
2. Blue nevus

VIII. Violaceous

A. Lupus pernio (sarcoidosis)
B. Lymphoma cutis
C. Cutaneous lupus

IX. Purple

A. Kaposi's sarcoma
B. Angiosarcoma
C. Palpable purpura (see Table 54-16)

X. Brown-black[c]

XI. Any color

A. Metastases

[a]If multiple with childhood onset, consider Gardner syndrome. [b]May have darker hue in more darkly pigmented individuals. [c]See also "Hyperpigmentation."

Abbreviation: MEN, multiple endocrine neoplasia.

WHITE LESIONS

In *calcinosis cutis*, there are firm white to white-yellow papules with an irregular surface. When the contents are expressed, a chalky white material is seen. *Dystrophic calcification* is seen at sites of previous

inflammation or damage to the skin. It develops in acne scars as well as on the distal extremities of patients with systemic sclerosis and in the subcutaneous tissue and intermuscular fascial planes in DM. The latter is more extensive and is more commonly seen in children. An elevated calcium phosphate product, most commonly due to secondary hyperparathyroidism in the setting of renal failure, can lead to nodules of *metastatic calcinosis cutis*, which tend to be subcutaneous and periarticular. These patients can also develop calcification of muscular arteries and subsequent ischemic necrosis (calciphylaxis). *Osteoma cutis*, in the form of small papules, most commonly occurs on the face of individuals with a history of acne vulgaris, whereas plate-like lesions occur in rare genetic syndromes.

SKIN-COLORED LESIONS

There are several types of skin-colored lesions, including epidermoid inclusion cysts, lipomas, rheumatoid nodules, neurofibromas, angiofibromas, neuromas, and adnexal tumors such as tricholemmomas. Both *epidermoid inclusion cysts* and *lipomas* are very common mobile subcutaneous nodules—the former are rubbery and drain cheeselike material (sebum and keratin) if incised. Lipomas are firm and somewhat lobulated on palpation. When extensive facial epidermoid inclusion cysts develop during childhood or there is a family history of such lesions, the patient should be examined for other signs of Gardner syndrome, including osteomas and desmoid tumors. *Rheumatoid nodules* are firm 0.5- to 4-cm nodules that favor the extensor aspect of joints, especially the elbows. They are seen in ~20% of patients with rheumatoid arthritis and 6% of patients with Still's disease. Biopsies of the nodules show palisading granulomas. Similar lesions that are smaller and shorter-lived are seen in rheumatic fever.

Neurofibromas (benign Schwann cell tumors) are soft papules or nodules that exhibit the "button-hole" sign; that is, they invaginate into the skin with pressure in a manner similar to a hernia. Single lesions are seen in normal individuals, but multiple neurofibromas, usually in combination with six or more CALMs measuring >1.5 cm (see "Hyperpigmentation," above), axillary freckling, and multiple Lisch nodules, are seen in von Recklinghausen's disease (NF type I) **(Chap. 86)**. In some patients, the neurofibromas are localized and unilateral due to somatic mosaicism.

Angiofibromas are firm pink to skin-colored papules that measure from 3 mm to 1.5 cm in diameter. When multiple lesions are located on the central cheeks (adenoma sebaceum), the patient has tuberous sclerosis or multiple endocrine neoplasia (MEN) syndrome, type 1. The former is an autosomal disorder due to mutations in two different genes, and the associated findings are discussed in the section on ash leaf spots as well as in **Chap. 86**.

Neuromas (benign proliferations of nerve fibers) are also firm, skin-colored papules. They are more commonly found at sites of amputations and in rudimentary polydacyly. However, when there are multiple neuromas on the eyelids, lips, distal tongue, and/or oral mucosa, the patient should be investigated for other signs of MEN syndrome, type 2b. Associated findings include marfanoid habitus, protuberant lips, intestinal ganglioneuromas, and medullary thyroid carcinoma (>75% of patients; **Chap. 381**).

Adnexal tumors are derived from pluripotent cells of the epidermis that can differentiate toward hair, sebaceous, apocrine or eccrine glands, or remain undifferentiated. *Basal cell carcinomas* (BCCs) are examples of adnexal tumors that have little or no evidence of differentiation. Clinically, they are translucent papules with rolled borders, telangiectasias, and central erosion. BCCs commonly arise in sun-damaged skin of the head and neck as well as the upper trunk. When a patient has multiple BCCs, especially prior to age 30, the possibility of the basal cell nevus syndrome should be raised. It is inherited as an autosomal dominant trait and is associated with jaw cysts, palmar and plantar pits, frontal bossing, medulloblastomas, and calcification of the falx cerebri and diaphragma sellae. *Tricholemmomas* are also skin-colored adnexal tumors but differentiate toward hair follicles and can have a wartlike appearance. The presence of multiple tricholemmomas on the face and cobblestoning of the oral mucosa points to the diagnosis of Cowden disease (multiple hamartoma syndrome) due to mutations in the phosphatase and tensin homolog (*PTEN*) gene. Internal organ involvement (in decreasing order of frequency) includes fibrocystic disease and carcinoma of the breast, adenomas and carcinomas of the thyroid, and gastrointestinal polyposis. Keratoses of the palms, soles, and dorsal aspect of the hands are also seen.

PINK LESIONS

The cutaneous lesions associated with primary systemic *amyloidosis* are often pink to pink-orange in color and translucent. Common locations are the face, especially the periorbital and perioral regions, and flexural areas. On biopsy, homogeneous deposits of amyloid are seen in the dermis and in the walls of blood vessels; the latter lead to an increase in vessel wall fragility. As a result, petechiae and purpura develop in clinically normal skin as well as in lesional skin following minor trauma, hence the term *pinch purpura*. Amyloid deposits are also seen in the striated muscle of the tongue and result in macroglossia.

Even though specific mucocutaneous lesions are present in only ~30% of the patients with primary systemic (AL) amyloidosis, the diagnosis can be made via histologic examination of abdominal subcutaneous fat, in conjunction with a serum free light chain assay. By special staining, amyloid deposits are seen around blood vessels or individual fat cells in 40–50% of patients. There are also three forms of amyloidosis that are limited to the skin and that should not be construed as cutaneous lesions of systemic amyloidosis. They are macular amyloidosis (upper back), lichen amyloidosis (usually lower extremities), and nodular amyloidosis. In macular and lichen amyloidosis, the deposits are composed of altered epidermal keratin. Early-onset macular and lichen amyloidosis have been associated with MEN syndrome, type 2a.

Patients with *multicentric reticulohistiocytosis* also have pink-colored papules and nodules on the face and mucous membranes as well as on the extensor surface of the hands and forearms. They have a polyarthritis that can mimic rheumatoid arthritis clinically. On histologic examination, the papules have characteristic giant cells that are not seen in biopsies of rheumatoid nodules. Pink to skin-colored papules that are firm, 2–5 mm in diameter, and often in a linear arrangement are seen in patients with *papular mucinosis*. This disease is also referred to as *scleromyxedema*. The latter name comes from the induration of the face and extremities that may accompany the papular eruption. Biopsy specimens of the papules show localized

mucin deposition, and serum protein electrophoresis plus immunofixation electrophoresis demonstrates a monoclonal spike of IgG, usually with a λ light chain.

YELLOW LESIONS

Several systemic disorders are characterized by yellow-colored cutaneous papules or plaques—hyperlipidemia (xanthomas), gout (tophi), diabetes (necrobiosis lipoidica), pseudoxanthoma elasticum, and Muir-Torre syndrome (sebaceous tumors). Eruptive xanthomas are the most common form of *xanthomas* and are associated with hypertriglyceridemia (primarily hyperlipoproteinemia types I, IV, and V). Crops of yellow papules with erythematous halos occur primarily on the extensor surfaces of the extremities and the buttocks, and they spontaneously involute with a fall in serum triglycerides. Types II and III result in one or more of the following types of xanthoma: xanthelasma, tendon xanthomas, and plane xanthomas. Xanthelasma are found on the eyelids, whereas tendon xanthomas are frequently associated with the Achilles and extensor finger tendons; plane xanthomas are flat and favor the palmar creases and flexural folds. Tuberous xanthomas are frequently associated with hypercholesterolemia; however, they are also seen in patients with hypertriglyceridemia and are found most frequently over the large joints or hand. Biopsy specimens of xanthomas show collections of lipid-containing macrophages (foam cells).

Patients with several disorders, including biliary cirrhosis, can have a secondary form of hyperlipidemia with associated tuberous and plane xanthomas. However, patients with plasma cell dyscrasias have *normolipemic plane xanthomas*. This latter form of xanthoma may be ≥12 cm in diameter and is most frequently seen on the neck, upper trunk, and flexural folds. It is important to note that the most common setting for eruptive xanthomas is uncontrolled diabetes mellitus. The least specific sign for hyperlipidemia is xanthelasma, because at least 50% of the patients with this finding have normal lipid profiles.

In *tophaceous gout*, there are deposits of monosodium urate in the skin around the joints, particularly those of the hands and feet. Additional sites of *tophi* formation include the helix of the ear and the olecranon and prepatellar bursae. The lesions are firm, yellow to yellow-white in color, and occasionally discharge a chalky material. Their size varies from 1 mm to 7 cm, and the diagnosis can be established by polarized light microscopy of the aspirated contents of a tophus. Lesions of *necrobiosis lipoidica* are found primarily on the shins (90%), and patients can have diabetes mellitus or develop it subsequently. Characteristic findings include a central yellow color, atrophy (transparency), telangiectasias, and a red to red-brown border. Ulcerations can also develop within the plaques. Biopsy specimens show necrobiosis of collagen and granulomatous inflammation.

In *pseudoxanthoma elasticum* (PXE), due to mutations in the gene *ABCC6*, there is an abnormal deposition of calcium on the elastic fibers of the skin, eye, and blood vessels. In the skin, the flexural areas such as the neck, axillae, antecubital fossae, and inguinal area are the primary sites of involvement. Yellow papules coalesce to form reticulated plaques that have an appearance similar to that of plucked chicken skin. In severely affected skin, hanging, redundant folds develop. Biopsy specimens of involved skin show swollen and irregularly clumped elastic fibers with deposits of calcium. In the eye, the calcium deposits in Bruch's membrane lead to angioid streaks and choroiditis; in the arteries of the heart, kidney, gastrointestinal tract,

and extremities, the deposits lead to angina, hypertension, gastrointestinal bleeding, and claudication, respectively.

Adnexal tumors that have differentiated toward sebaceous glands include sebaceous adenoma, sebaceous carcinoma, and sebaceous hyperplasia. Except for sebaceous hyperplasia, which is commonly seen on the face, these tumors are fairly rare. Patients with Muir-Torre syndrome have one or more *sebaceous adenoma(s)*, and they can also have sebaceous carcinomas and sebaceous hyperplasia as well as keratoacanthomas. The internal manifestations of Muir-Torre syndrome include *multiple* carcinomas of the gastrointestinal tract (primarily colon) as well as cancers of the genitourinary tract.

RED LESIONS

Cutaneous lesions that are red in color have a wide variety of etiologies; in an attempt to simplify their identification, they will be subdivided into papules, papules/plaques, and subcutaneous nodules. Common red papules include *arthropod bites* and *cherry hemangiomas*; the latter are small, bright-red, dome-shaped papules that represent a benign proliferation of capillaries. In patients with AIDS **(Chap. 197)**, the development of multiple red hemangioma-like lesions points to bacillary angiomatosis, and biopsy specimens show clusters of bacilli that stain positively with the Warthin-Starry stain; the pathogens have been identified as *Bartonella henselae* and *Bartonella quintana*. Disseminated visceral disease is seen primarily in immunocompromised hosts but can occur in immunocompetent individuals.

Multiple *angiokeratomas* are seen in Fabry disease, an X-linked recessive lysosomal storage disease that is due to a deficiency of α-galactosidase A. The lesions are red to red-blue in color and can be quite small in size (1–3 mm), with the most common location being the lower trunk. Associated findings include chronic renal disease, peripheral neuropathy, and corneal opacities (cornea verticillata). Electron photomicrographs of angiokeratomas and clinically normal skin demonstrate lamellar lipid deposits in fibroblasts, pericytes, and endothelial cells that are diagnostic of this disease. Widespread acute eruptions of erythematous papules are discussed in the section on exanthems.

There are several infectious diseases that present as erythematous papules or nodules in a lymphocutaneous or sporotrichoid pattern, that is, in a linear arrangement along the lymphatic channels. The two most common etiologies are *Sporothrix schenckii* (sporotrichosis) and the atypical mycobacterium *Mycobacterium marinum*. The organisms are introduced as a result of trauma, and a primary inoculation site is often seen in addition to the lymphatic nodules. Additional causes include *Nocardia*, *Leishmania*, and other atypical mycobacteria and dimorphic fungi; culture or PCR of lesional tissue will aid in the diagnosis.

The diseases that are characterized by erythematous plaques with scale are reviewed in the papulosquamous section, and the various forms of dermatitis are discussed in the section on erythroderma. Additional disorders in the differential diagnosis of red papules/plaques include *cellulitis*, *polymorphous light eruption* (PMLE), *cutaneous lymphoid hyperplasia* (lymphocytoma cutis), *cutaneous lupus*, *lymphoma cutis*, and *leukemia cutis*. The first three diseases represent primary cutaneous disorders, although cellulitis may be accompanied by a bacteremia. PMLE is characterized by erythematous papules and plaques in a

primarily sun-exposed distribution—dorsum of the hand, extensor forearm, and upper trunk. Lesions follow exposure to UV-B and/or UV-A, and in higher latitudes, PMLE is most severe in the late spring and early summer. A process referred to as "hardening" occurs with continued UV exposure, and the eruption fades, but in temperate climates, it recurs the next spring. PMLE must be differentiated from cutaneous lupus, and this is accomplished by observation of the natural history, histologic examination, and sometimes direct immunofluorescence of the lesions. Cutaneous lymphoid hyperplasia (pseudolymphoma) is a *benign* polyclonal proliferation of lymphocytes within the skin that presents as infiltrated pink-red to red-purple papules and plaques; it must be distinguished from lymphoma cutis.

Several types of red plaques are seen in patients with systemic *lupus*, including (1) erythematous urticarial plaques across the cheeks and nose in the classic butterfly rash; (2) erythematous discoid lesions with fine or "carpet-tack" scale, telangiectasias, central hypopigmentation, peripheral hyperpigmentation, follicular plugging, and atrophy located on the scalp, face, external ears, arms, and upper trunk; and (3) psoriasiform or annular lesions of subacute cutaneous lupus with hypopigmented centers located primarily on the extensor arms and upper trunk. Additional mucocutaneous findings include (1) a violaceous flush on the face and V of the neck; (2) photosensitivity; (3) urticarial vasculitis (see "Urticaria," above); (4) lupus panniculitis (see below); (5) diffuse alopecia; (6) alopecia secondary to discoid lesions; (7) cuticular telangiectasias and erythema; (8) EM- or TEN-like lesions that may become bullous; (9) oral or nasal ulcers; (10) livedo reticularis; and (11) distal ulcerations secondary to Raynaud's phenomenon, vasculitis, or livedoid vasculopathy. Patients with only discoid lesions usually have the form of lupus that is limited to the skin. However, up to 10–15% of these patients eventually develop systemic lupus. Direct immunofluorescence of involved skin, in particular discoid lesions, shows deposits of IgG or IgM and C3 in a granular distribution along the dermal-epidermal junction.

In *lymphoma cutis*, there is a clonal proliferation of malignant lymphocytes within the skin, and the clinical appearance resembles that of cutaneous lymphoid hyperplasia—infiltrated pink-red to red-purple papules and plaques. Lymphoma cutis can occur anywhere on the surface of the skin, whereas the sites of predilection for lymphocytomas include the malar ridge, tip of the nose, and earlobes. Patients with non-Hodgkin's lymphomas have specific cutaneous lesions more often than those with Hodgkin's disease, and, occasionally, the skin nodules precede the development of extracutaneous non-Hodgkin's lymphoma or represent the only site of involvement (e.g., primary cutaneous B cell lymphoma). Arcuate lesions are sometimes seen in lymphoma and lymphocytoma cutis as well as in CTCL. *Adult T cell leukemia/lymphoma* that develops in association with HTLV-1 infection is characterized by cutaneous plaques, hypercalcemia, and circulating CD25+ lymphocytes. *Leukemia cutis* has the same appearance as lymphoma cutis, and specific lesions are seen more commonly in monocytic leukemias than in lymphocytic or granulocytic leukemias. Cutaneous chloromas (granulocytic sarcomas) may precede the appearance of circulating blasts in acute myelogenous leukemia and, as such, represent a form of aleukemic leukemia cutis.

Sweet syndrome is characterized by pink-red to red-brown edematous plaques that are frequently painful and occur primarily on the head, neck, and upper extremities. The patients also have fever, neutrophilia, and a dense dermal infiltrate of neutrophils in the lesions. In ~10% of the patients, there is an associated

malignancy, most commonly acute myelogenous leukemia. Sweet syndrome has also been reported with inflammatory bowel disease, systemic lupus erythematosus, and solid tumors (primarily of the genitourinary tract) as well as drugs (e.g., all-*trans*-retinoic acid, granulocyte colony-stimulating factor [G-CSF]). The differential diagnosis includes neutrophilic eccrine hidradenitis; bullous forms of pyoderma gangrenosum; and, occasionally, cellulitis. Extracutaneous sites of involvement include joints, muscles, eyes, kidneys (proteinuria, occasionally glomerulonephritis), and lungs (neutrophilic infiltrates). The idiopathic form of Sweet syndrome is seen more often in women, following a respiratory tract infection.

Common causes of erythematous subcutaneous nodules include inflamed epidermoid inclusion cysts, acne cysts, and furuncles. *Panniculitis*, an inflammation of the fat, also presents as subcutaneous nodules and is frequently a sign of systemic disease. There are several forms of panniculitis, including erythema nodosum, erythema induratum/nodular vasculitis, lupus panniculitis, lipodermatosclerosis, α_1-antitrypsin deficiency, factitial, and fat necrosis secondary to pancreatic disease. Except for erythema nodosum, these lesions may break down and ulcerate or heal with a scar. The shin is the most common location for the nodules of erythema nodosum, whereas the calf is the most common location for lesions of erythema induratum. In erythema nodosum, the nodules are initially red but then develop a blue color as they resolve. Patients with erythema nodosum but no underlying systemic illness can still have fever, malaise, leukocytosis, arthralgias, and/or arthritis. However, the possibility of an underlying illness should be excluded, and the most common associations are streptococcal infections, upper respiratory viral infections, sarcoidosis, and inflammatory bowel disease, in addition to drugs (oral contraceptives, sulfonamides, penicillins, bromides, iodides, BRAF inhibitors). Less common associations include bacterial gastroenteritis (*Yersinia*, *Salmonella*) and coccidioidomycosis followed by tuberculosis, histoplasmosis, brucellosis, and infections with *Chlamydophila pneumoniae*, *Chlamydia trachomatis, Mycoplasma pneumoniae,* or hepatitis B virus.

Erythema induratum and nodular vasculitis have overlapping features clinically and histologically, and whether they represent two separate entities or the ends of a single disease spectrum is a point of debate; in general, the latter is usually idiopathic and the former is associated with the presence of *Mycobacterium tuberculosis* DNA by PCR within skin lesions. The lesions of lupus panniculitis are found primarily on the cheeks, upper arms, and buttocks (sites of abundant fat) and are seen in both the cutaneous and systemic forms of lupus. The overlying skin may be normal, erythematous, or have the changes of discoid lupus. The subcutaneous fat necrosis that is associated with pancreatic disease is presumably secondary to circulating lipases and is seen in patients with pancreatic carcinoma as well as in patients with acute and chronic pancreatitis. In this disorder, there may be an associated arthritis, fever, and inflammation of visceral fat. Histologic examination of deep incisional biopsy specimens will aid in the diagnosis of the particular type of panniculitis.

Subcutaneous erythematous nodules are also seen in cutaneous polyarteritis nodosa and as a manifestation of *systemic vasculitis* when there is involvement of medium-sized vessels, for example, systemic polyarteritis nodosa, eosinophilic granulomatosis with polyangiitis, or granulomatosis with polyangiitis (Chap. 356). Cutaneous polyarteritis nodosa presents with painful subcutaneous nodules and ulcers within a red-purple, netlike pattern of livedo reticularis. The latter is due to slowed blood flow through the superficial horizontal

venous plexus. The majority of lesions are found on the lower extremities, and while arthralgias and myalgias may accompany cutaneous polyarteritis nodosa, there is no evidence of systemic involvement. In both the cutaneous and systemic forms of vasculitis, skin biopsy specimens of the associated nodules will show the changes characteristic of a necrotizing vasculitis and/or granulomatous inflammation.

RED-BROWN LESIONS

The cutaneous lesions in *sarcoidosis* (Chap. 360) are classically red to red-brown in color, and with diascopy (pressure with a glass slide), a yellow-brown residual color is observed that is secondary to the granulomatous infiltrate. The waxy papules and plaques may be found anywhere on the skin, but the face is the most common location. Usually there are no surface changes, but occasionally the lesions will have scale. Biopsy specimens of the papules show "naked" granulomas in the dermis, that is, granulomas surrounded by a minimal number of lymphocytes. Other cutaneous findings in sarcoidosis include annular lesions with an atrophic or scaly center, papules within scars, hypopigmented papules and patches, alopecia, acquired ichthyosis, erythema nodosum, and lupus pernio (see below).

The differential diagnosis of sarcoidosis includes foreign-body granulomas produced by chemicals such as beryllium and zirconium, late secondary syphilis, and *lupus vulgaris*. Lupus vulgaris is a form of cutaneous tuberculosis that is seen in previously infected and sensitized individuals. There is often underlying active tuberculosis elsewhere, usually in the lungs or lymph nodes. Lesions occur primarily in the head and neck region and are red-brown plaques with a yellow-brown color on diascopy. Secondary scarring can develop within the central portion of the plaques. Cultures or PCR analysis of the lesions should be performed, along with an interferon γ release assay of peripheral blood, because it is rare for the acid-fast stain to show bacilli within the dermal granulomas.

A generalized distribution of red-brown macules and papules is seen in the form of mastocytosis known as *urticaria pigmentosa* (Chap. 347). Each lesion represents a collection of mast cells in the dermis, with hyperpigmentation of the overlying epidermis. Stimuli such as rubbing cause these mast cells to degranulate, and this leads to the formation of localized urticaria (Darier's sign). Additional symptoms can result from mast cell degranulation and include headache, flushing, diarrhea, and pruritus. Mast cells also infiltrate various organs such as the liver, spleen, and gastrointestinal tract, and accumulations of mast cells in the bones may produce either osteosclerotic or osteolytic lesions on radiographs. In the majority of these patients, however, the internal involvement remains indolent. A subtype of chronic cutaneous small-vessel vasculitis, *erythema elevatum diutinum* (EED), also presents with papules that are red-brown in color. The papules coalesce into plaques on the extensor surfaces of knees, elbows, and the small joints of the hand. Flares of EED have been associated with streptococcal infections.

BLUE LESIONS

Lesions that are blue in color are the result of vascular ectasias, hyperplasias and tumors or melanin pigment within the dermis. *Venous lakes* (ectasias) are compressible dark-blue lesions that are found commonly in the head and neck region. *Venous malformations* are also compressible blue papulonodules and plaques

that can occur anywhere on the body, including the oral mucosa. When there are multiple papulonodules rather than a single congenital lesion, the patient may have the blue rubber bleb syndrome or Maffucci's syndrome. Patients with the blue rubber bleb syndrome also have vascular anomalies of the gastrointestinal tract that may bleed, whereas patients with Maffucci's syndrome have associated osteochondromas. *Blue nevi* (moles) are seen when there are collections of pigment-producing nevus cells in the dermis. These benign papular lesions are dome-shaped and occur most commonly on the dorsum of the hand or foot or in the head and neck region.

VIOLACEOUS LESIONS

Violaceous papules and plaques are seen in *lupus pernio*, *lymphoma cutis*, and *cutaneous lupus*. Lupus pernio is a particular type of sarcoidosis that involves the tip and alar rim of the nose as well as the earlobes, with lesions that are violaceous in color rather than red-brown. This form of sarcoidosis is associated with involvement of the upper respiratory tract. The plaques of lymphoma cutis and cutaneous lupus may be red or violaceous in color and were discussed above.

PURPLE LESIONS

Purple-colored papules and plaques are seen in vascular tumors, such as *Kaposi's sarcoma* (Chap. 197) and *angiosarcoma*, and when there is extravasation of red blood cells into the skin in association with inflammation, as in *palpable purpura* (see "Purpura," below). Patients with congenital or acquired AV fistulas and venous hypertension can develop purple papules on the lower extremities that can resemble Kaposi's sarcoma clinically and histologically; this condition is referred to as pseudo-Kaposi's sarcoma (acral angiodermatitis). Angiosarcoma is found most commonly on the scalp and face of elderly patients or within areas of chronic lymphedema and presents as purple papules and plaques. In the head and neck region, the tumor often extends beyond the clinically defined borders and may be accompanied by facial edema.

BROWN AND BLACK LESIONS

Brown- and black-colored papules are reviewed in "Hyperpigmentation," above.

CUTANEOUS METASTASES

These are discussed last because they can have a wide range of colors. Most commonly, they present as either firm, skin-colored subcutaneous nodules or firm, red to red-brown papulonodules while metastatic melanoma can be pink, blue, or black in color. Cutaneous metastases develop from hematogenous or lymphatic spread and are most often due to the following primary carcinomas: in men, melanoma, oropharynx, lung, and colon; and in women, breast, melanoma, and ovary. These metastatic lesions may be the initial presentation of the carcinoma, especially when the primary site is the lung.

PURPURA

(Table 54-16) *Purpura* are seen when there is an extravasation of red blood cells into the dermis and, as a result, the lesions do not blanch with pressure. This is in contrast to those erythematous or violet-colored lesions that are due to localized vasodilatation—they do blanch with pressure. Purpura (≥3 mm) and petechiae (≤2 mm) are divided into two major groups: palpable and nonpalpable. The most frequent causes of *nonpalpable* petechiae and purpura are primary cutaneous disorders such as *trauma*, *solar* (*actinic*) *purpura*, and *capillaritis*. Less common causes are *steroid purpura* and *livedoid vasculopathy* (see "Ulcers," below). Solar purpura are seen primarily on the extensor forearms, whereas steroid purpura secondary to potent topical glucocorticoids or endogenous or exogenous Cushing's syndrome can be more widespread. In both cases, there is alteration of the supporting connective tissue that surrounds the dermal blood vessels. In contrast, the petechiae that result from capillaritis are found primarily on the lower extremities. In capillaritis, there is an extravasation of erythrocytes as a result of perivascular lymphocytic inflammation. The petechiae are bright red, 1–2 mm in size, and scattered within yellow-brown patches. The yellow-brown color is caused by hemosiderin deposits within the dermis.

TABLE 54-16

Causes of Purpura

I. Primary cutaneous disorders
  A. Nonpalpable
    1. Trauma
    2. Solar (actinic, senile) purpura
    3. Steroid purpura
    4. Capillaritis
    5. Livedoid vasculopathy in the setting of venous hypertension[a]
II. Drugs (e.g. anti-platelet agents, anti-coagulants)
III. Systemic diseases
  A. Nonpalpable
    1. Clotting disturbances
      a. Thrombocytopenia (including ITP)
      b. Abnormal platelet function
      c. Clotting factor defects
    2. Vascular fragility
      a. Amyloidosis (within normal-appearing skin)
      b. Ehlers-Danlos syndrome
      c. Scurvy
    3. Thrombi
      a. Disseminated intravascular coagulation
      b. Warfarin (Coumadin®)-induced necrosis
      c. Heparin-induced thrombocytopenia and thrombosis
      d. Antiphospholipid antibody syndrome
      e. Monoclonal cryoglobulinemia
      f. Vasculopathy induced by levamisole-adulterated cocaine
      g. Thrombotic thrombocytopenic purpura
      h. Thrombocytosis
      i. Homozygous protein C or protein S deficiency
    4. Emboli
      a. Cholesterol
      b. Fat
    5. Possible immune complex
      a. Gardner-Diamond syndrome (autoerythrocyte sensitivity)
      b. Waldenström's hypergammaglobulinemic purpura
  B. Palpable
    1. Vasculitis

a. Cutaneous small-vessel vasculitis, including in the setting of systemic vasculitides

2. Emboli[b]
   a. Acute meningococcemia
   b. Disseminated gonococcal infection
   c. Rocky Mountain spotted fever
   d. Ecthyma gangrenosum

[a]Also associated with underlying disorders that lead to hypercoagulability/thrombophilia, e.g., factor V Leiden, protein C dysfunction/deficiency. [b]Bacterial (including rickettsial), fungal, or parasitic.

Abbreviation: ITP, idiopathic thrombocytopenic purpura.

Systemic causes of nonpalpable purpura fall into several categories, and those secondary to clotting disturbances and vascular fragility will be discussed first. The former group includes *thrombocytopenia* (Chap. 111), *abnormal platelet function* as is seen in uremia, and *clotting factor defects*. The initial site of presentation for thrombocytopenia-induced petechiae is the distal lower extremity. Capillary fragility leads to nonpalpable purpura in patients with systemic *amyloidosis* (see "Papulonodular Skin Lesions," above), disorders of collagen production such as *Ehlers-Danlos syndrome*, and *scurvy*. In scurvy, there are flattened corkscrew hairs with surrounding hemorrhage on the lower extremities, in addition to gingivitis. Vitamin C is a cofactor for lysyl hydroxylase, an enzyme involved in the posttranslational modification of procollagen that is necessary for cross-link formation.

In contrast to the previous group of disorders, the noninflammatory purpura seen in the following group of diseases are associated with thrombi formation within vessels and have a retiform configuration. It is important to note that these thrombi are demonstrable in skin biopsy specimens. This group of disorders includes disseminated intravascular coagulation (DIC), monoclonal cryoglobulinemia, thrombocytosis, thrombotic thrombocytopenic purpura, antiphospholipid antibody syndrome, and reactions to warfarin and heparin (heparin-induced thrombocytopenia and thrombosis). DIC is triggered by several types of infection (gram-negative, gram-positive, viral, and rickettsial) as well as by tissue injury and neoplasms. Widespread purpura and hemorrhagic infarcts of the distal extremities are seen. Similar lesions are found in purpura fulminans, which is a form of DIC associated with fever and hypotension that occurs more commonly in children following an infectious illness such as varicella, scarlet fever, or an upper respiratory tract infection. In both disorders, hemorrhagic bullae can develop in involved skin.

Monoclonal cryoglobulinemia is associated with plasma cell dyscrasias, chronic lymphocytic leukemia, and lymphoma. Purpura, primarily of the lower extremities, and hemorrhagic infarcts of the fingers, toes, and ears are seen in these patients. Exacerbations of disease activity can follow cold exposure or an increase in serum viscosity. Biopsy specimens show precipitates of the cryoglobulin within dermal vessels. Similar deposits have been found in the lung, brain, and renal glomeruli. Patients with *thrombotic thrombocytopenic purpura* can also have hemorrhagic infarcts as a result of intravascular thromboses.

Additional signs include microangiopathic hemolytic anemia and fluctuating neurologic abnormalities, especially headaches and confusion.

Administration of *warfarin* can result in painful areas of erythema that become purpuric and then necrotic with an adherent black eschar; the condition is referred to as warfarin-induced necrosis. This reaction is seen more often in women and in areas with abundant subcutaneous fat—breasts, abdomen, buttocks, thighs, and calves. The erythema and purpura develop between the third and tenth day of therapy, most likely as a result of a transient imbalance in the levels of anticoagulant and procoagulant vitamin K–dependent factors. Continued therapy does not exacerbate preexisting lesions, and patients with an inherited or acquired deficiency of protein C are at increased risk for this particular reaction as well as for purpura fulminans and calciphylaxis.

Purpura secondary to *cholesterol emboli* are usually seen on the lower extremities of patients with atherosclerotic vascular disease. They often follow anticoagulant therapy or an invasive vascular procedure such as an arteriogram but also occur spontaneously from disintegration of atheromatous plaques. Associated findings include livedo reticularis, gangrene, cyanosis, and ischemic ulcerations. Multiple step sections of the biopsy specimen may be necessary to demonstrate the cholesterol clefts within the vessels. Petechiae are also an important sign of *fat embolism* and occur primarily on the upper body 2–3 days after a major injury. By using special fixatives, the emboli can be demonstrated in biopsy specimens of the petechiae. Emboli of tumor or thrombus are seen in patients with atrial myxomas and marantic endocarditis.

In the *Gardner-Diamond syndrome* (autoerythrocyte sensitivity), female patients develop large ecchymoses within areas of painful, warm erythema. Intradermal injections of autologous erythrocytes or phosphatidyl serine derived from the red cell membrane can reproduce the lesions in some patients; however, there are instances where a reaction is seen at an injection site of the forearm but not in the midback region. The latter has led some observers to view Gardner-Diamond syndrome as a cutaneous manifestation of severe emotional stress. More recently, the possibility of platelet dysfunction (as assessed via aggregation studies) has been raised. *Waldenström's hypergammaglobulinemic purpura* is a chronic disorder characterized by recurrent crops of petechiae and larger purpuric macules on the lower extremities. There are circulating complexes of IgG–anti-IgG molecules, and exacerbations are associated with prolonged standing or walking.

Palpable purpura are further subdivided into vasculitic and embolic. In the group of vasculitic disorders, cutaneous small-vessel vasculitis, also known as *leukocytoclastic vasculitis* (LCV), is the one most commonly associated with palpable purpura (Chap. 356). Underlying etiologies include drugs (e.g., antibiotics), infections (e.g., hepatitis C virus), and autoimmune connective tissue diseases (e.g., rheumatoid arthritis, Sjögren's syndrome, lupus). *Henoch-Schönlein purpura* (HSP) is a subtype of acute LCV that is seen more commonly in children and adolescents following an upper respiratory infection. The majority of lesions are found on the lower extremities and buttocks. Systemic manifestations include fever, arthralgias (primarily of the knees and ankles), abdominal pain, gastrointestinal bleeding, and nephritis. Direct immunofluorescence examination shows deposits of IgA within dermal blood vessel walls. Renal disease is of particular concern in adults with HSP.

Several types of infectious emboli can give rise to palpable purpura. These embolic lesions are usually *irregular* in outline as opposed to the lesions of LCV, which are *circular* in outline. The irregular outline is indicative of a cutaneous infarct, and the size corresponds to the area of skin that received its blood supply from that particular arteriole or artery. The palpable purpura in LCV are circular because the erythrocytes simply diffuse out evenly from the postcapillary venules as a result of inflammation. Infectious emboli are most commonly due to gram-negative cocci (meningococcus, gonococcus), gram-negative rods (Enterobacteriaceae), and gram-positive cocci (*Staphylococcus*). Additional causes include *Rickettsia* and, in immunocompromised patients, *Aspergillus* and other opportunistic fungi.

The embolic lesions in *acute meningococcemia* are found primarily on the trunk, lower extremities, and sites of pressure, and a gunmetal-gray color often develops within them. Their size varies from a few millimeters to several centimeters, and the organisms can be cultured from the lesions. Associated findings include a preceding upper respiratory tract infection; fever; meningitis; DIC; and, in some patients, a deficiency of the terminal components of complement. In *disseminated gonococcal infection* (arthritis-dermatitis syndrome), a small number of inflammatory papules and vesicopustules, often with central purpura or hemorrhagic necrosis, are found on the distal extremities. Additional symptoms include arthralgias, tenosynovitis, and fever. To establish the diagnosis, a Gram stain of these lesions should be performed. *Rocky Mountain spotted fever* is a tick-borne disease that is caused by *Rickettsia rickettsii*. A several-day history of fever, chills, severe headache, and photophobia precedes the onset of the cutaneous eruption. The initial lesions are erythematous macules and papules on the wrists, ankles, palms, and soles. With time, the lesions spread centripetally and become purpuric.

Lesions of *ecthyma gangrenosum* begin as edematous, erythematous papules or plaques and then develop central purpura and necrosis. Bullae formation also occurs in these lesions, and they are frequently found in the girdle region. The organism that is classically associated with ecthyma gangrenosum is *Pseudomonas aeruginosa*, but other gram-negative rods such as *Klebsiella*, *Escherichia coli*, and *Serratia* can produce similar lesions. In immunocompromised hosts, the list of potential pathogens is expanded to include *Candida* and other opportunistic fungi (e.g., *Aspergillus*, *Fusarium)*.

ULCERS

The approach to the patient with a cutaneous ulcer is outlined in **Table 54-17**. Peripheral vascular diseases of the extremities are reviewed in **Chap. 275**, as is Raynaud's phenomenon.

TABLE 54-17

Causes of Mucocutaneous Ulcers

I. Primary cutaneous disorders
   A. Peripheral vascular disease **(Chap. 275)**
      1. Venous
      2. Arterial[a]
   B. Livedoid vasculopathy in the setting of venous hypertension[b]
   C. Squamous cell carcinoma (e.g., within scars), basal cell carcinomas
   D. Infections, e.g., ecthyma caused by *Streptococcus* **(Chap. 143)**
   E. Physical, e.g., trauma, pressure
   F. Drugs, e.g., hydroxyurea

II. Systemic diseases
   A. Lower legs
      1. Small-vessel and medium-vessel vasculitis[c]
      2. Hemoglobinopathies **(Chap. 94)**
      3. Cryoglobulinemia,[c] cryofibrinogenemia
      4. Cholesterol emboli[a,c]
      5. Necrobiosis lipoidica[d]
      6. Antiphospholipid syndrome **(Chap. 112)**
      7. Neuropathic[e] **(Chap. 396)**
      8. Panniculitis
      9. Kaposi's sarcoma, acral angiodermatitis
      10. Diffuse dermal angiomatosis
   B. Hands and feet
      1. Raynaud's phenomenon **(Chap. 275)**
      2. Buerger disease
   C. Generalized
      1. Pyoderma gangrenosum, but most commonly legs
      2. Calciphylaxis **(Chap. 403)**
      3. Infections, e.g., dimorphic fungi, leishmaniasis
      4. Lymphoma
   D. Face, especially perioral, and anogenital
      1. Chronic herpes simplex[f]

III. Mucosal
   A. Behçet's syndrome **(Chap. 357)**
   B. Erythema multiforme major, Stevens-Johnson syndrome, TEN
   C. Primary blistering disorders **(Chap. 55)**

D. Lupus erythematosus, lichen planus
E. Inflammatory bowel disease
F. Acute HIV infection
G. Reactive arthritis

[a]Underlying atherosclerosis. [b]Also associated with underlying disorders that lead to hypercoagulability/thrombophilia, e.g., factor V Leiden, protein C dysfunction/deficiency, antiphospholipid antibodies. [c]Reviewed in section on Purpura. [d]Reviewed in section on Papulonodular Skin Lesions. [e]Favors plantar surface of the foot. [f]Sign of immunosuppression.

Abbreviations: HIV, human immunodeficiency virus; TEN, toxic epidermal necrolysis.

Livedoid vasculopathy (livedoid vasculitis; atrophie blanche) represents a combination of a vasculopathy plus intravascular thrombosis. Purpuric lesions and livedo reticularis are found in association with *painful* ulcerations of the lower extremities. These ulcers are often slow to heal, but when they do, irregularly shaped white scars form. The majority of cases are secondary to venous hypertension, but possible underlying illnesses include disorders of hypercoagulability, for example, antiphospholipid syndrome, factor V Leiden (Chaps. 113 and 350).

In *pyoderma gangrenosum*, the border of untreated active ulcers has a characteristic appearance consisting of an undermined necrotic violaceous edge and a peripheral erythematous halo. The ulcers often begin as pustules that then expand rather rapidly to a size as large as 20 cm. Although these lesions are most commonly found on the lower extremities, they can arise anywhere on the surface of the body, including at sites of trauma (pathergy). An estimated 30–50% of cases are idiopathic, and the most common associated disorders are ulcerative colitis and Crohn's disease. Less commonly, pyoderma gangrenosum is associated with seropositive rheumatoid arthritis, acute and chronic myelogenous leukemia, hairy cell leukemia, myelofibrosis, or a monoclonal gammopathy, usually IgA. Because the histology of pyoderma gangrenosum may be nonspecific (dermal infiltrate of neutrophils when in untreated state), the diagnosis requires clinicopathologic correlation, in particular, the exclusion of similar-appearing ulcers such as necrotizing vasculitis, Meleney's ulcer (synergistic infection at a site of trauma or surgery), dimorphic fungi, cutaneous amebiasis, spider bites, and factitial. In the myeloproliferative disorders, the ulcers may be more superficial with a pustulobullous border, and these lesions provide a connection between classic pyoderma gangrenosum and acute febrile neutrophilic dermatosis (Sweet syndrome).

FEVER AND RASH

The major considerations in a patient with a fever and a rash are inflammatory diseases versus infectious diseases. In the hospital setting, the most common scenario is a patient who has a drug rash plus a fever secondary to an underlying infection. However, it should be emphasized that a drug reaction can lead to both a cutaneous eruption and a fever ("drug fever"), especially in the setting of DRESS, AGEP, or serum sickness–like reaction. Additional inflammatory diseases that are often associated with a fever include

pustular psoriasis, erythroderma, and Sweet syndrome. Lyme disease, secondary syphilis, and viral and bacterial exanthems (see "Exanthems," above) are examples of infectious diseases that produce a rash and a fever. Lastly, it is important to determine whether or not the cutaneous lesions represent septic emboli (see "Purpura," above). Such lesions usually have evidence of ischemia in the form of purpura, necrosis, or impending necrosis (gunmetal-gray color). In the patient with thrombocytopenia, however, purpura can be seen in inflammatory reactions such as morbilliform drug eruptions and infectious lesions.

FURTHER READING

Bolognia JL, Schaffer JV, Cerroni L (eds): *Dermatology*, 4th ed. Philadelphia, Elsevier, 2018.

Callen JP et al (eds): *Dermatological Signs of Systemic Disease*, 5th ed. Edinburgh, Elsevier, 2017.

Rigopoulos D, Larios G, Katsambas A: Skin signs of systemic diseases. Clin Dermatol 29:531, 2011. [PubMed: 21855729]

Taylor SC et al (eds): *Taylor and Kelly's Dermatology for Skin of Color*, 2nd ed. New York, McGraw-Hill, 2016.

Thiers BH, Sahn RE, Callen JP: Cutaneous manifestations of internal malignancy. CA: Cancer J Clin 59:73, 2009. [PubMed: 19258446]

Chapter 55: Immunologically Mediated Skin Diseases

Kim B. Yancey; Thomas J. Lawley

INTRODUCTION

A number of immunologically mediated skin diseases and immunologically mediated systemic disorders with cutaneous manifestations are now recognized as distinct entities with consistent clinical, histologic, and immunopathologic findings. Clinically, these disorders are characterized by morbidity (pain, pruritus, disfigurement) and, in some instances, result in death (largely due to loss of epidermal barrier function and/or secondary infection). The major features of the more common immunologically mediated skin diseases are summarized in this chapter **(Table 55-1)**, as are autoimmune systemic disorders with cutaneous manifestations.

TABLE 55-1

Immunologically Mediated Blistering Diseases

Disease	Clinical Manifestations	Histology	Immunopathology	Autoantigens[a]
Pemphigus vulgaris	Flaccid blisters, denuded skin, oromucosal lesions	Acantholytic blister formed in suprabasal layer of epidermis	Cell surface deposits of IgG on keratinocytes	Dsg3 (plus Dsg1 in patients with skin involvement)
Pemphigus foliaceus	Crusts and shallow erosions on scalp, central face, upper chest, and back	Acantholytic blister formed in superficial layer of epidermis	Cell surface deposits of IgG on keratinocytes	Dsg1
Paraneoplastic pemphigus	Painful stomatitis with papulosquamous or lichenoid eruptions that may progress to blisters	Acantholysis, keratinocyte necrosis, and vacuolar interface dermatitis	Cell surface deposits of IgG and C3 on keratinocytes and (variably) similar immunoreactants in epidermal BMZ	Plakin protein family members and desmosomal cadherins (see text for details)
Bullous pemphigoid	Large tense blisters on flexor surfaces and trunk	Subepidermal blister with eosinophil-rich infiltrate	Linear band of IgG and/or C3 in epidermal BMZ	BPAG1, BPAG2
Pemphigoid gestationis	Pruritic, urticarial plaques rimmed by vesicles and bullae on the trunk and extremities	Teardrop-shaped, subepidermal blisters in dermal papillae; eosinophil-rich infiltrate	Linear band of C3 in epidermal BMZ	BPAG2 (plus BPAG1 in some patients)

Disease	Clinical Manifestations	Histology	Immunopathology	Autoantigens[a]
Dermatitis herpetiformis	Extremely pruritic small papules and vesicles on elbows, knees, buttocks, and posterior neck	Subepidermal blister with neutrophils in dermal papillae	Granular deposits of IgA in dermal papillae	Epidermal transglutaminase
Linear IgA disease	Pruritic small papules on extensor surfaces; occasionally larger, arciform blisters	Subepidermal blister with neutrophil-rich infiltrate	Linear band of IgA in epidermal BMZ	BPAG2 (see text for specific details)
Epidermolysis bullosa acquisita	Blisters, erosions, scars, and milia on sites exposed to trauma; widespread, inflammatory, tense blisters may be seen initially	Subepidermal blister that may or may not include a leukocytic infiltrate	Linear band of IgG and/or C3 in epidermal BMZ	Type VII collagen
Mucous membrane pemphigoid	Erosive and/or blistering lesions of mucous membranes and possibly the skin; scarring of some sites	Subepidermal blister that may or may not include a leukocytic infiltrate	Linear band of IgG, IgA, and/or C3 in epidermal BMZ	BPAG2, laminin-332, or others

[a]Autoantigens bound by these patients' autoantibodies are defined as follows: Dsg1, desmoglein 1; Dsg3, desmoglein 3; BPAG1, bullous pemphigoid antigen 1; BPAG2, bullous pemphigoid antigen 2.

Abbreviation: BMZ, basement membrane zone.

AUTOIMMUNE CUTANEOUS DISEASES

PEMPHIGUS VULGARIS

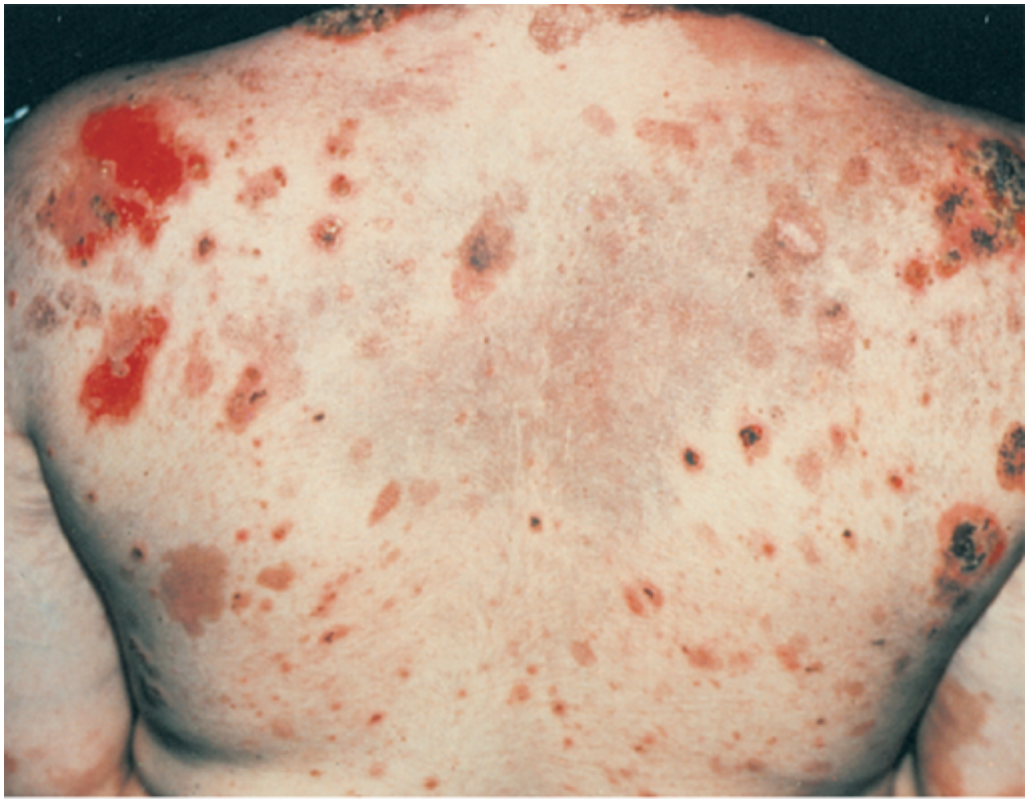
Pemphigus refers to a group of autoantibody-mediated intraepidermal blistering diseases characterized by loss of cohesion between epidermal cells (a process termed *acantholysis*). Manual pressure to the skin of

these patients may elicit the separation of the epidermis (*Nikolsky's sign*). This finding, while characteristic of pemphigus, is not specific to this group of disorders and is also seen in toxic epidermal necrolysis, Stevens-Johnson syndrome, and a few other skin diseases.

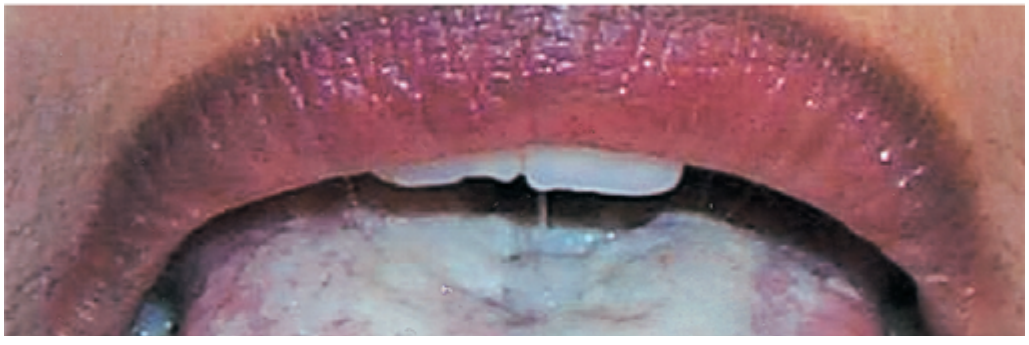
Pemphigus vulgaris (PV) is a mucocutaneous blistering disease that predominantly occurs in patients >40 years of age. PV typically begins on mucosal surfaces and often progresses to involve the skin. This disease is characterized by fragile, flaccid blisters that rupture to produce extensive denudation of mucous membranes and skin (Fig. 55-1). The mouth, scalp, face, neck, axilla, groin, and trunk are typically involved. PV may be associated with severe skin pain; some patients experience pruritus as well. Lesions usually heal without scarring except at sites complicated by secondary infection or mechanically induced dermal wounds. Postinflammatory hyperpigmentation is usually present for some time at sites of healed lesions.

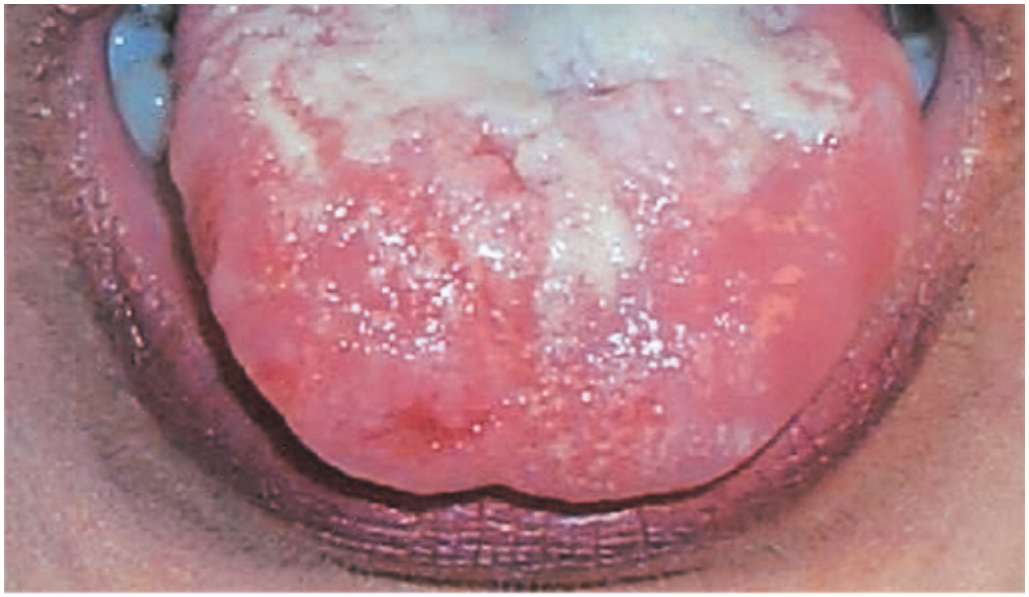
FIGURE 55-1

Pemphigus vulgaris. ***A.*** Flaccid bullae are easily ruptured, resulting in multiple erosions and crusted plaques. ***B.*** Involvement of the oral mucosa, which is almost invariable, may present with erosions on the gingiva, buccal mucosa, palate, posterior pharynx, or tongue. *(B, Courtesy of Robert Swerlick, MD; with permission.)*



A





B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Biopsies of early lesions demonstrate intraepidermal vesicle formation secondary to loss of cohesion between epidermal cells (i.e., acantholytic blisters). Blister cavities contain acantholytic epidermal cells, which appear as round homogeneous cells containing hyperchromatic nuclei. Basal keratinocytes remain attached to the epidermal basement membrane; hence, blister formation takes place within the suprabasal portion of the epidermis. Lesional skin may contain focal collections of intraepidermal eosinophils within blister cavities; dermal alterations are slight, often limited to an eosinophil-predominant leukocytic infiltrate. Direct immunofluorescence microscopy of lesional or intact patient skin shows deposits of IgG on the surface of keratinocytes; deposits of complement components are typically found in lesional but not in uninvolved skin. Deposits of IgG on keratinocytes are derived from circulating autoantibodies to cell-surface autoantigens. Such circulating autoantibodies can be demonstrated in 80–90% of PV patients by indirect immunofluorescence microscopy; monkey esophagus is the optimal substrate for these studies. Patients with PV have IgG autoantibodies to *desmogleins* (Dsgs), transmembrane desmosomal glycoproteins that belong to the cadherin family of calcium-dependent adhesion molecules. Such autoantibodies can be precisely quantitated by enzyme-linked immunosorbent assay (ELISA). Patients with early PV (i.e., mucosal disease) have IgG autoantibodies to Dsg3; patients with advanced PV (i.e., mucocutaneous disease) have IgG autoantibodies to both Dsg3 and Dsg1. Experimental studies have shown that autoantibodies from patients with PV are pathogenic (i.e., responsible for blister formation) and that their titer correlates with disease activity. Recent studies have shown that the anti-Dsg autoantibody profile in these patients' sera as well as the tissue distribution of Dsg3 and Dsg1 determine the site of blister formation in patients with PV. Coexpression of Dsg3 and Dsg1 by epidermal cells protects against pathogenic IgG antibodies to either of these cadherins but not against pathogenic autoantibodies to both.

PV can be life-threatening. Prior to the availability of glucocorticoids, mortality rates ranged from 60% to 90%; the current figure is ~5%. Common causes of morbidity and death are infection and complications of treatment. Bad prognostic factors include advanced age, widespread involvement, and the requirement for

high doses of glucocorticoids (with or without other immunosuppressive agents) for control of disease. The course of PV in individual patients is variable and difficult to predict. Some patients experience remission, while others may require long-term treatment or succumb to complications of their disease or its treatment. The mainstay of treatment is systemic glucocorticoids. Patients with moderate to severe PV are usually started on prednisone at 1 mg/kg per day. If new lesions continue to appear after 1–2 weeks of treatment, the dose may need to be increased and/or prednisone may need to be combined with other immunosuppressive agents such as azathioprine (2–2.5 mg/kg per day), mycophenolate mofetil (20–35 mg/kg per day), rituximab (375 mg/m^2 per week × 4, or 1000 mg on days 1 and 15), or cyclophosphamide (1–2 mg/kg per day). Patients with severe, treatment-resistant disease may derive benefit from plasmapheresis (six high-volume exchanges [i.e., 2–3 L per exchange] over ~2 weeks) and/or IV immunoglobulin (IVIg) (2 g/kg over 3–5 days every 6–8 weeks). It is important to bring severe or progressive disease under control quickly in order to lessen the severity and/or duration of this disorder. Increasingly, rituximab and daily glucocorticoids are used early in PV patients to avert the development of advanced and/or treatment-resistant disease.

PEMPHIGUS FOLIACEUS

Pemphigus foliaceus (PF) is distinguished from PV by several features. In PF, acantholytic blisters are located high within the epidermis, usually just beneath the stratum corneum. Hence, PF is a more superficial blistering disease than PV. The distribution of lesions in the two disorders is much the same, except that in PF mucous membranes are almost always spared. Patients with PF rarely have intact blisters but rather exhibit shallow erosions associated with erythema, scale, and crust formation. Mild cases of PF can resemble severe seborrheic dermatitis; severe PF may cause extensive exfoliation. Sun exposure (ultraviolet irradiation) may be an aggravating factor.

PF has immunopathologic features in common with PV. Specifically, direct immunofluorescence microscopy of perilesional skin demonstrates IgG on the surface of keratinocytes. Similarly, patients with PF have circulating IgG autoantibodies directed against the surface of keratinocytes. In PF, autoantibodies are directed against Dsg1, a 160-kDa desmosomal cadherin. These autoantibodies can be quantitated by ELISA. As noted for PV, the autoantibody profile in patients with PF (i.e., anti-Dsg1 IgG) and the tissue distribution of this autoantigen (i.e., expression in oral mucosa that is compensated by coexpression of Dsg3) are thought to account for the distribution of lesions in this disease.

Endemic forms of PF are found in south-central rural Brazil, where the disease is known as *fogo salvagem* (FS), as well as in selected sites in Latin America and Tunisia. Endemic PF, like other forms of this disease, is mediated by IgG autoantibodies to Dsg1. Clusters of FS overlap with those of leishmaniasis, a disease transmitted by bites of the sand fly *Lutzomyia longipalis*. Recent studies have shown that sand-fly salivary antigens (specifically, the LJM11 salivary protein) are recognized by IgG autoantibodies from FS patients (as well as by monoclonal antibodies to Dsg1 derived from these patients). Moreover, mice immunized with LJM11 produce antibodies to Dsg1. Thus, these findings suggest that insect bites may deliver salivary antigens that initiate a cross-reactive humoral immune response, which may lead to FS in genetically susceptible individuals.

Although pemphigus has been associated with several autoimmune diseases, its association with thymoma and/or myasthenia gravis is particularly notable. To date, >30 cases of thymoma and/or myasthenia gravis have been reported in association with pemphigus, usually with PF. Patients may also develop pemphigus as a consequence of drug exposure; drug-induced pemphigus usually resembles PF rather than PV. Drugs containing a thiol group in their chemical structure (e.g., penicillamine, captopril, enalapril) are most commonly associated with drug-induced pemphigus. Nonthiol drugs linked to pemphigus include penicillins, cephalosporins, and piroxicam. Some cases of drug-induced pemphigus are durable and require treatment with systemic glucocorticoids and/or immunosuppressive agents.

PF is generally a less severe disease than PV and usually carries a better prognosis. Localized disease can sometimes be treated with topical or intralesional glucocorticoids; more active cases can usually be controlled with systemic glucocorticoids either alone or in combination with other immunosuppressive agents. Patients with severe, treatment-resistant disease may require more aggressive interventions, as described above for patients with PV.

PARANEOPLASTIC PEMPHIGUS

Paraneoplastic pemphigus (PNP) is an autoimmune acantholytic mucocutaneous disease associated with an occult or confirmed neoplasm. Patients with PNP typically have painful stomatitis in association with papulosquamous and/or lichenoid eruptions that often progress to blisters. Palm and sole involvement are common in these patients and raise the possibility that prior reports of neoplasia-associated erythema multiforme actually may have represented unrecognized cases of PNP. Biopsies of lesional skin from these patients show varying combinations of acantholysis, keratinocyte necrosis, and vacuolar-interface dermatitis. Direct immunofluorescence microscopy of a patient's skin shows deposits of IgG and complement on the surface of keratinocytes and (variably) similar immunoreactants in the epidermal basement membrane zone. Patients with PNP have IgG autoantibodies to cytoplasmic proteins that are members of the plakin family (e.g., desmoplakins I and II, bullous pemphigoid antigen [BPAG]1, envoplakin, periplakin, and plectin) and to cell-surface proteins that are members of the cadherin family (e.g., Dsg1 and Dsg3). Passive transfer studies have shown that autoantibodies from patients with PNP are pathogenic in animal models.

The predominant neoplasms associated with PNP are non-Hodgkin's lymphoma, chronic lymphocytic leukemia, thymoma, spindle cell tumors, Waldenström's macroglobulinemia, and Castleman's disease; the last-mentioned neoplasm is particularly common among children with PNP. Rare cases of seronegative PNP have been reported in patients with B cell malignancies previously treated with rituximab. In addition to severe skin lesions, many patients with PNP develop life-threatening bronchiolitis obliterans. PNP is generally resistant to conventional therapies (i.e., those used to treat PV); rarely, a patient's disease may ameliorate or even remit following ablation or removal of underlying neoplasms.

BULLOUS PEMPHIGOID

Bullous pemphigoid (BP) is a polymorphic autoimmune subepidermal blistering disease usually seen in the elderly. Initial lesions may consist of urticarial plaques; most patients eventually display tense blisters on either normal-appearing or erythematous skin **(Fig. 55-2)**. The lesions are usually distributed over the lower abdomen, groin, and flexor surface of the extremities; oral mucosal lesions are found in some patients. Pruritus may be nonexistent or severe. As lesions evolve, tense blisters tend to rupture and be replaced by erosions with or without surmounting crust. Nontraumatized blisters heal without scarring. The major histocompatibility complex class II allele HLA-DQβ1*0301 is prevalent in patients with BP. Despite isolated reports, several studies have shown that patients with BP do not have a higher incidence of malignancy than appropriately age- and gender-matched controls.

FIGURE 55-2

Bullous pemphigoid with tense vesicles and bullae on erythematous, urticarial bases. *(Courtesy of the Yale Resident's Slide Collection; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Biopsies of early lesional skin demonstrate subepidermal blisters and histologic features that roughly correlate with the clinical character of the particular lesion under study. Lesions on normal-appearing skin generally contain a sparse perivascular leukocytic infiltrate with some eosinophils; conversely, biopsies of inflammatory lesions typically show an eosinophil-rich infiltrate at sites of vesicle formation and in perivascular areas. In addition to eosinophils, cell-rich lesions also contain mononuclear cells and

neutrophils. It is not possible to distinguish BP from other subepidermal blistering diseases by routine histologic studies alone.

Direct immunofluorescence microscopy of normal-appearing perilesional skin from patients with BP shows linear deposits of IgG and/or C3 in the epidermal basement membrane. The sera of ~70% of these patients contain circulating IgG autoantibodies that bind the epidermal basement membrane of normal human skin in indirect immunofluorescence microscopy. IgG from an even higher percentage of patients reacts with the epidermal side of 1 *M* NaCl split skin (an alternative immunofluorescence microscopy test substrate used to distinguish circulating IgG autoantibodies to the basement membrane in patients with BP from those in patients with similar, yet different, subepidermal blistering diseases; see below). In BP, circulating autoantibodies recognize 230- and 180-kDa hemidesmosome-associated proteins in basal keratinocytes (i.e., BPAG1 and BPAG2, respectively). Autoantibodies to BPAG2 are thought to deposit in situ, activate complement, produce dermal mast-cell degranulation, and generate granulocyte-rich infiltrates that cause tissue damage and blister formation.

BP may persist for months to years, with exacerbations or remissions. Extensive involvement may result in widespread erosions and compromise cutaneous integrity; elderly and/or debilitated patients may die. The mainstay of treatment is systemic glucocorticoids. Local or minimal disease can sometimes be controlled with topical glucocorticoids alone; more extensive lesions generally respond to systemic glucocorticoids either alone or in combination with other immunosuppressive agents. Patients usually respond to prednisone (0.75–1 mg/kg per day). In some instances, azathioprine (2–2.5 mg/kg per day), mycophenolate mofetil (20–35 mg/kg per day), or rituximab (375 mg/m^2 per week x 4, or 1000 mg on days 1 and 15) are necessary adjuncts.

PEMPHIGOID GESTATIONIS

Pemphigoid gestationis (PG), also known as *herpes gestationis*, is a rare, nonviral, subepidermal blistering disease of pregnancy and the puerperium. PG may begin during any trimester of pregnancy or present shortly after delivery. Lesions are usually distributed over the abdomen, trunk, and extremities; mucous membrane lesions are rare. Skin lesions in these patients may be quite polymorphic and consist of erythematous urticarial papules and plaques, vesiculopapules, and/or frank bullae. Lesions are almost always extremely pruritic. Severe exacerbations of PG frequently follow delivery, typically within 24–48 h. PG tends to recur in subsequent pregnancies, often beginning earlier during such gestations. Brief flare-ups of disease may occur with resumption of menses and may develop in patients later exposed to oral contraceptives. Occasionally, infants of affected mothers have transient skin lesions.

Biopsies of early lesional skin show teardrop-shaped subepidermal vesicles forming in dermal papillae in association with an eosinophil-rich leukocytic infiltrate. Differentiation of PG from other subepidermal bullous diseases by light microscopy is difficult. However, direct immunofluorescence microscopy of perilesional skin from PG patients reveals the immunopathologic hallmark of this disorder: linear deposits of C3 in the epidermal basement membrane. These deposits develop as a consequence of complement activation produced by low-titer IgG anti–basement membrane autoantibodies directed against BPAG2, the

same hemidesmosome-associated protein that is targeted by autoantibodies in patients with BP—a subepidermal bullous disease that resembles PG clinically, histologically, and immunopathologically.

The goals of therapy in patients with PG are to prevent the development of new lesions, relieve intense pruritus, and care for erosions at sites of blister formation. Many patients require treatment with moderate doses of daily glucocorticoids (i.e., 20–40 mg of prednisone) at some point in their course. Mild cases (or brief flare-ups) may be controlled by vigorous use of potent topical glucocorticoids. Infants born of mothers with PG appear to be at increased risk of being born slightly premature or "small for dates." Current evidence suggests that there is no difference in the incidence of uncomplicated live births between PG patients treated with systemic glucocorticoids and those managed more conservatively. If systemic glucocorticoids are administered, newborns are at risk for development of reversible adrenal insufficiency.

DERMATITIS HERPETIFORMIS

Dermatitis herpetiformis (DH) is an intensely pruritic, papulovesicular skin disease characterized by lesions symmetrically distributed over extensor surfaces (i.e., elbows, knees, buttocks, back, scalp, and posterior neck) **(see Fig. 52-8)**. Primary lesions in this disorder consist of papules, papulovesicles, or urticarial plaques. Because pruritus is prominent, patients may present with excoriations and crusted papules but no observable primary lesions. Patients sometimes report that their pruritus has a distinctive burning or stinging component; the onset of such local symptoms reliably heralds the development of distinct clinical lesions 12–24 h later. Almost all DH patients have associated, usually subclinical, gluten-sensitive enteropathy **(Chap. 318)**, and >90% express the HLA-B8/DRw3 and HLA-DQw2 haplotypes. DH may present at any age, including in childhood; onset in the second to fourth decades is most common. The disease is typically chronic.

Biopsy of early lesional skin reveals neutrophil-rich infiltrates within dermal papillae. Neutrophils, fibrin, edema, and microvesicle formation at these sites are characteristic of early disease. Older lesions may demonstrate nonspecific features of a subepidermal bulla or an excoriated papule. Because the clinical and histologic features of this disease can be variable and resemble those of other subepidermal blistering disorders, the diagnosis is confirmed by direct immunofluorescence microscopy of normal-appearing perilesional skin. Such studies demonstrate granular deposits of IgA (with or without complement components) in the papillary dermis and along the epidermal basement membrane zone. IgA deposits in the skin are unaffected by control of disease with medication; however, these immunoreactants diminish in intensity or disappear in patients maintained for long periods on a strict gluten-free diet (see below). Patients with DH have granular deposits of IgA in their epidermal basement membrane zone and should be distinguished from individuals with linear IgA deposits at this site (see below).

Although most DH patients do not report overt gastrointestinal symptoms or have laboratory evidence of malabsorption, biopsies of the small bowel usually reveal blunting of intestinal villi and a lymphocytic infiltrate in the lamina propria. As is true for patients with celiac disease, this gastrointestinal abnormality can be reversed by a gluten-free diet. Moreover, if maintained, this diet alone may control the skin disease and eventuate in clearance of IgA deposits from these patients' epidermal basement membrane zones.

Subsequent gluten exposure in such patients alters the morphology of their small bowel, elicits a flare-up of their skin disease, and is associated with the reappearance of IgA in their epidermal basement membrane zones. As in patients with celiac disease, dietary gluten sensitivity in patients with DH is associated with IgA anti-endomysial autoantibodies that target tissue transglutaminase. Studies indicate that patients with DH also have high-avidity IgA autoantibodies to epidermal transglutaminase and that the latter is co-localized with granular deposits of IgA in the papillary dermis of DH patients. Patients with DH also have an increased incidence of thyroid abnormalities, achlorhydria, atrophic gastritis, and autoantibodies to gastric parietal cells. These associations likely relate to the high frequency of the HLA-B8/DRw3 haplotype in these patients, since this marker is commonly linked to autoimmune disorders. The mainstay of treatment of DH is dapsone, a sulfone. Patients respond rapidly (24–48 h) to dapsone (50–200 mg/d), but require careful pretreatment evaluation and close follow-up to ensure that complications are avoided or controlled. All patients taking dapsone at >100 mg/d will have some hemolysis and methemoglobinemia, which are expected pharmacologic side effects of this agent. Gluten restriction can control DH and lessen dapsone requirements; this diet must rigidly exclude gluten to be of maximal benefit. Many months of dietary restriction may be necessary before a beneficial result is achieved. Good dietary counseling by a trained dietitian is essential.

LINEAR IgA DISEASE

Linear IgA disease, once considered a variant form of DH, is actually a separate and distinct entity. Clinically, patients with linear IgA disease may resemble individuals with DH, BP, or other subepidermal blistering diseases. Lesions typically consist of papulovesicles, bullae, and/or urticarial plaques that develop predominantly on central or flexural sites. Oral mucosal involvement occurs in some patients. Severe pruritus resembles that is seen in patients with DH. Patients with linear IgA disease do not have an increased frequency of the HLA-B8/DRw3 haplotype or an associated enteropathy and therefore are not candidates for treatment with a gluten-free diet.

Histologic alterations in early lesions may be virtually indistinguishable from those in DH. However, direct immunofluorescence microscopy of normal-appearing perilesional skin reveals a linear band of IgA (and often C3) in the epidermal basement membrane zone. Most patients with linear IgA disease have circulating IgA anti-basement membrane autoantibodies directed against neoepitopes in the proteolytically processed extracellular domain of BPAG2. These patients generally respond to treatment with dapsone (50–200 mg/d).

EPIDERMOLYSIS BULLOSA ACQUISITA

Epidermolysis bullosa acquisita (EBA) is a rare, noninherited, polymorphic, chronic, subepidermal blistering disease. **(The inherited form is discussed in Chap. 406.)** Patients with classic or noninflammatory EBA have blisters on noninflamed skin, atrophic scars, milia, nail dystrophy, and oral lesions. Because lesions generally occur at sites exposed to minor trauma, classic EBA is considered a mechanobullous disease. Other patients with EBA have widespread inflammatory scarring and bullous lesions that resemble severe BP. Inflammatory EBA may evolve into the classic, noninflammatory form of this disease. Rarely, patients present with lesions that predominate on mucous membranes. The HLA-DR2 haplotype is found with increased frequency in EBA

patients. Studies suggest that EBA is sometimes associated with inflammatory bowel disease (especially Crohn's disease).

The histology of lesional skin varies with the character of the lesion being studied. Noninflammatory bullae are subepidermal, feature a sparse leukocytic infiltrate, and resemble the lesions in patients with porphyria cutanea tarda. Inflammatory lesions consist of neutrophil-rich subepidermal blisters. EBA patients have continuous deposits of IgG (and frequently C3) in a linear pattern within the epidermal basement membrane zone. Ultrastructurally, these immunoreactants are found in the sublamina densa region in association with anchoring fibrils. Approximately 50% of EBA patients have demonstrable circulating IgG anti-basement membrane autoantibodies directed against type VII collagen—the collagen species that makes up anchoring fibrils. Such IgG autoantibodies bind the dermal side of 1 *M* NaCl split skin (in contrast to IgG autoantibodies in patients with BP). Studies have shown that passive transfer of experimental or patient IgG against type VII collagen can produce lesions in mice that clinically, histologically, and immunopathologically resemble those in patients with EBA.

Treatment of EBA is generally unsatisfactory. Some patients with inflammatory EBA may respond to systemic glucocorticoids, either alone or in combination with immunosuppressive agents. Other patients (especially those with neutrophil-rich inflammatory lesions) may respond to dapsone. The chronic, noninflammatory form of EBA is largely resistant to treatment, although some patients may respond to cyclosporine, azathioprine, IVIg, or rituximab.

MUCOUS MEMBRANE PEMPHIGOID

Mucous membrane pemphigoid (MMP) is a rare, acquired, subepithelial immunobullous disease characterized by erosive lesions of mucous membranes and skin that result in scarring of at least some sites of involvement. Common sites include the oral mucosa (especially the gingiva) and conjunctiva; other sites that may be affected include the nasopharyngeal, laryngeal, esophageal, and anogenital mucosa. Skin lesions (present in about one-third of patients) tend to predominate on the scalp, face, and upper trunk and generally consist of a few scattered erosions or tense blisters on an erythematous or urticarial base. MMP is typically a chronic and progressive disorder. Serious complications may arise as a consequence of ocular, laryngeal, esophageal, or anogenital lesions. Erosive conjunctivitis may result in shortened fornices, symblepharon, ankyloblepharon, entropion, corneal opacities, and (in severe cases) blindness. Similarly, erosive lesions of the larynx may cause hoarseness, pain, and tissue loss that, if unrecognized and untreated, may eventuate in complete destruction of the airway. Esophageal lesions may result in stenosis and/or strictures that could place patients at risk for aspiration. Strictures may also complicate anogenital involvement.

Biopsies of lesional tissue generally show subepithelial vesiculo-bullae and a mononuclear leukocytic infiltrate. Neutrophils and eosinophils may be seen in biopsies of early lesions; older lesions may demonstrate a scant leukocytic infiltrate and fibrosis. Direct immunofluorescence microscopy of perilesional tissue typically reveals deposits of IgG, IgA, and/or C3 in the epidermal basement membrane. Because many patients with MMP exhibit no evidence of circulating anti-basement membrane autoantibodies, testing of

perilesional skin is important diagnostically. Although MMP was once thought to be a single nosologic entity, it is now largely regarded as a disease phenotype that may develop as a consequence of an autoimmune reaction to a variety of molecules in the epidermal basement membrane (e.g., BPAG2, laminin-332, type VII collagen, $\alpha_6\beta_4$ integrin) and other antigens yet to be completely defined. Studies suggest that MMP patients with autoantibodies to laminin-332 have an increased relative risk for cancer. Treatment of MMP is largely dependent upon the sites of involvement. Due to potentially severe complications, patients with ocular, laryngeal, esophageal, and/or anogenital involvement require aggressive systemic treatment with dapsone, prednisone, or the latter in combination with another immunosuppressive agent (e.g., azathioprine, mycophenolate mofetil, cyclophosphamide, or rituximab), or IVIg. Less threatening forms of the disease may be managed with topical or intralesional glucocorticoids.

AUTOIMMUNE SYSTEMIC DISEASES WITH PROMINENT CUTANEOUS FEATURES

DERMATOMYOSITIS

The cutaneous manifestations of dermatomyositis **(Chap. 358)** are often distinctive but at times may resemble those of systemic lupus erythematosus (SLE) **(Chap. 349)**, scleroderma **(Chap. 353)**, or other overlapping connective tissue diseases **(Chap. 353)**. The extent and severity of cutaneous disease may or may not correlate with the extent and severity of the myositis. The cutaneous manifestations of dermatomyositis are similar, whether the disease appears in children or in the elderly, except that calcification of subcutaneous tissue is a common late sequela in childhood dermatomyositis.

The cutaneous signs of dermatomyositis may precede or follow the development of myositis by weeks to years. Cases lacking muscle involvement (i.e., *dermatomyositis sine myositis* or *amyopathic dermatomyositis*) have also been reported. The most common manifestation is a purple-red discoloration of the upper eyelids, sometimes associated with scaling ("heliotrope" erythema; **Fig. 55-3**) and periorbital edema. Erythema on the cheeks and nose in a "butterfly" distribution may resemble the malar eruption of SLE. Erythematous or violaceous scaling patches are common on the upper anterior chest, posterior neck, scalp, and the extensor surfaces of the arms, legs, and hands. Erythema and scaling may be particularly prominent over the elbows, knees, and dorsal interphalangeal joints. Approximately one-third of patients have violaceous, flat-topped papules over the dorsal interphalangeal joints that are pathognomonic of dermatomyositis (Gottron's papules) **(Fig. 55-4)**. Thin violaceous papules and plaques on the elbows and knees of patients with dermatomyositis are referred to as *Gottron's sign* (Fig. 55-4). These lesions can be contrasted with the erythema and scaling on the dorsum of the fingers that spares the skin over the interphalangeal joints of some SLE patients. Periungual telangiectases and edema may be prominent in patients with dermatomyositis. Lacy or reticulated erythema may be associated with fine scaling on the extensor and lateral surfaces of the thighs and upper arms. Other patients, particularly those with long-standing disease, develop areas of hypopigmentation, hyperpigmentation, mild atrophy, and telangiectasia known as *poikiloderma*. Poikiloderma is rare in both SLE and scleroderma and thus can serve as a clinical

sign that distinguishes dermatomyositis from these two diseases. Cutaneous changes may be similar in dermatomyositis and various overlap syndromes where thickening and binding down of the skin of the hands (*sclerodactyly*) as well as Raynaud's phenomenon can be seen. However, the presence of severe muscle disease, Gottron's papules, heliotrope erythema, and poikiloderma serve to distinguish patients with dermatomyositis. Skin biopsy of the erythematous, scaling lesions of dermatomyositis may reveal only mild nonspecific inflammation, but sometimes may show changes indistinguishable from those found in cutaneous lupus erythematosus (LE), including epidermal atrophy, hydropic degeneration of basal keratinocytes, and dermal changes consisting of edema of the upper dermis, interstitial mucin deposition, and a mild mononuclear cell infiltrate. Direct immunofluorescence microscopy of lesional skin is usually negative, although granular deposits of immunoglobulin(s) and complement in the epidermal basement membrane zone have been described in some patients. Treatment should be directed at the systemic disease. Topical glucocorticoids are sometimes useful; patients should avoid exposure to ultraviolet irradiation and aggressively use photoprotective measures, including broad-spectrum sunscreens.

FIGURE 55-3

Dermatomyositis. Periorbital violaceous erythema characterizes the classic heliotrope rash. *(Courtesy of James Krell, MD; with permission.*)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 55-4

Gottron's papules. Dermatomyositis often involves the hands as erythematous flat-topped papules over the knuckles. Periungual telangiectases are also evident.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

LUPUS ERYTHEMATOSUS

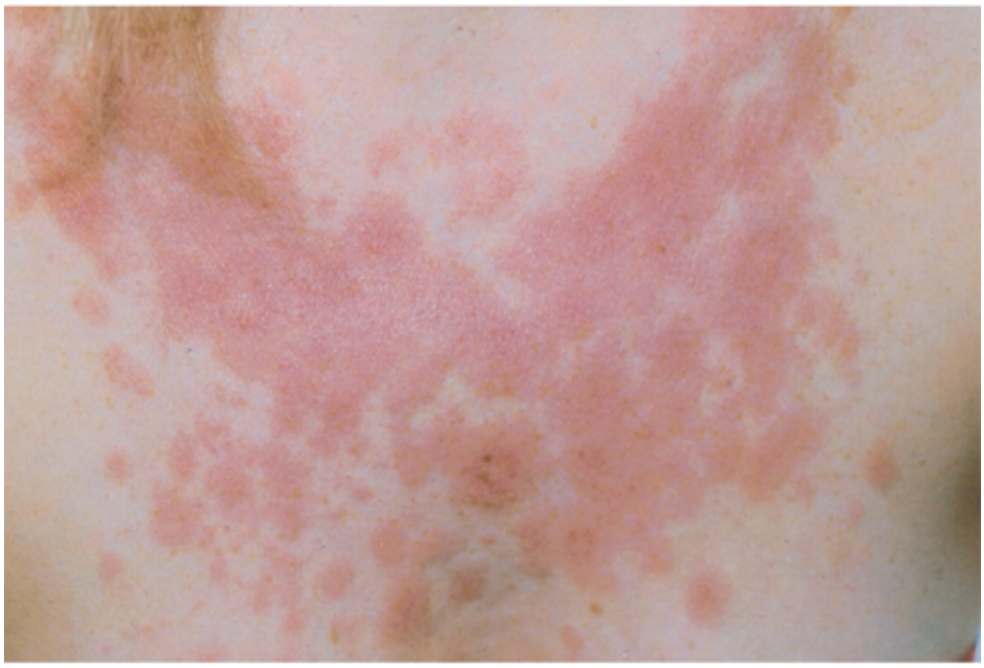
The cutaneous manifestations of LE **(Chap. 349)** can be divided into acute, subacute, and chronic or discoid types. *Acute cutaneous LE* is characterized by erythema of the nose and malar eminences in a "butterfly" distribution **(Fig. 55-5A)**. The erythema is often sudden in onset, accompanied by edema and fine scale, and correlated with systemic involvement. Patients may have widespread involvement of the face as well as erythema and scaling of the extensor surfaces of the extremities and upper chest **(Fig. 55-5B)**. These acute lesions, while sometimes evanescent, usually last for days and are often associated with exacerbations of systemic disease. Skin biopsy of acute lesions typically shows hydropic degeneration of basal keratinocytes, dermal edema, and (in some cases) a sparse infiltrate of mononuclear cells in the upper dermis as well as dermal mucin. Direct immunofluorescence microscopy of lesional skin frequently reveals deposits of immunoglobulin(s) and complement in the epidermal basement membrane zone. Treatment is aimed at control of systemic disease. Photoprotection is very important in this as well as in other forms of LE.

FIGURE 55-5

Acute cutaneous lupus erythematosus (LE). ***A.*** Acute cutaneous LE on the face, showing prominent, scaly, malar erythema. Involvement of other sun-exposed sites is also common. ***B.*** Acute cutaneous LE on the upper chest, demonstrating brightly erythematous and slightly edematous papules and plaques. *(B, Courtesy of Robert Swerlick, MD; with permission.)*



A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Subacute cutaneous lupus erythematosus (SCLE) is characterized by a widespread photosensitive, nonscarring eruption. In most patients, renal and central nervous system involvement is mild or absent. SCLE may present as a papulosquamous eruption that resembles psoriasis or as annular polycyclic lesions. In the papulosquamous form, discrete erythematous papules arise on the back, chest, shoulders, extensor surfaces of the arms, and dorsum of the hands; lesions are uncommon on the central face and the flexor

surfaces of the arms as well as below the waist. These slightly scaling papules tend to merge into large plaques, some with a reticulate appearance. The annular form involves the same areas and presents with erythematous papules that evolve into oval, circular, or polycyclic lesions. The lesions of SCLE are more widespread but have less tendency for scarring than lesions of discoid LE. In many patients with SCLE, drugs (e.g., hydrochlorothiazide, calcium channel blockers, proton pump inhibitors) may induce or exacerbate disease. Skin biopsy typically reveals epidermal changes that include atrophy, hydropic degeneration of basal keratinocytes, and apoptosis accompanied by an infiltrate of mononuclear cells in the upper dermis. Direct immunofluorescence microscopy of lesional skin reveals deposits of immunoglobulin(s) in the epidermal basement membrane zone in about one-half of these cases. A particulate pattern of IgG deposition throughout the epidermis has been associated with SCLE. Most SCLE patients have anti-Ro autoantibodies. Local therapy alone is usually unsuccessful. Most patients require treatment with aminoquinoline antimalarial drugs. Low-dose therapy with oral glucocorticoids is sometimes necessary. Photoprotective measures against both ultraviolet B and ultraviolet A wavelengths are very important.

Discoid lupus erythematosus (DLE, also called *chronic cutaneous LE*) is characterized by discrete lesions, most often found on the face, scalp, and/or external ears. The lesions are erythematous papules or plaques with a thick, adherent scale that occludes hair follicles (follicular plugging). When the scale is removed, its underside shows small excrescences that correlate with the openings of hair follicles (so-called “carpet tacking”), a finding relatively specific for DLE. Long-standing lesions develop central atrophy, scarring, and hypopigmentation but frequently have erythematous, sometimes raised borders **(Fig. 55-6)**. These lesions persist for years and tend to expand slowly. Up to 15% of patients with DLE eventually meet the American College of Rheumatology criteria for SLE. Typical discoid lesions are frequently seen in patients with SLE. Biopsy of DLE lesions shows hyperkeratosis, follicular plugging, atrophy of the epidermis, hydropic degeneration of basal keratinocytes, thickening of the epidermal basement membrane zone, and a mononuclear cell infiltrate adjacent to epidermal, adnexal, and microvascular basement membranes. Direct immunofluorescence microscopy demonstrates immunoglobulin(s) and complement deposits at the basement membrane zone in ~90% of cases. Treatment is focused on control of local cutaneous disease and consists mainly of photoprotection and topical or intralesional glucocorticoids. If local therapy is ineffective, use of aminoquinoline antimalarial agents may be indicated.

FIGURE 55-6

Discoid (chronic cutaneous) lupus erythematosus (LE). Violaceous, hyperpigmented, atrophic plaques, follicular plugging, and scarring are typical features of chronic cutaneous LE.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SCLERODERMA AND MORPHEA

The skin changes of scleroderma (Chap. 353) usually begin on the fingers, hands, toes, feet, and face, with episodes of recurrent nonpitting edema. Sclerosis of the skin commences distally on the fingers (sclerodactyly) and spreads proximally, usually accompanied by resorption of bone of the fingertips, which may have punched out ulcers, stellate scars, or areas of hemorrhage (Fig. 55-7). The fingers may actually shrink and become sausage-shaped, and, because the fingernails are usually unaffected, they may curve over the end of the fingertips. Periungual telangiectases are usually present, but periungual erythema is rare. In advanced cases, the extremities show contractures and calcinosis cutis. Facial involvement includes a smooth, unwrinkled brow, taut skin over the nose, shrinkage of tissue around the mouth, and perioral radial furrowing (Fig. 55-8). Matlike telangiectases are often present, particularly on the face and hands. Involved skin feels indurated, smooth, and bound to underlying structures; hyper- and hypopigmentation are common as well. *Raynaud's phenomenon* (i.e., cold-induced blanching, cyanosis, and reactive hyperemia) is documented in almost all patients and can precede development of scleroderma by many years. *Linear scleroderma* is a limited form of disease that presents in a linear, bandlike distribution and tends to involve deep as well as superficial layers of skin. The combination of calcinosis cutis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectases has been termed as the *CREST syndrome*. Anti-centromere autoantibodies have been reported in a very high percentage of patients with CREST syndrome but in only a small minority of patients with scleroderma. Skin biopsy reveals thickening of the dermis, homogenization of collagen bundles, atrophic pilosebaceous and eccrine glands, and a sparse mononuclear

cell infiltrate in the dermis and subcutaneous fat. Direct immunofluorescence microscopy of lesional skin is usually negative.

FIGURE 55-7

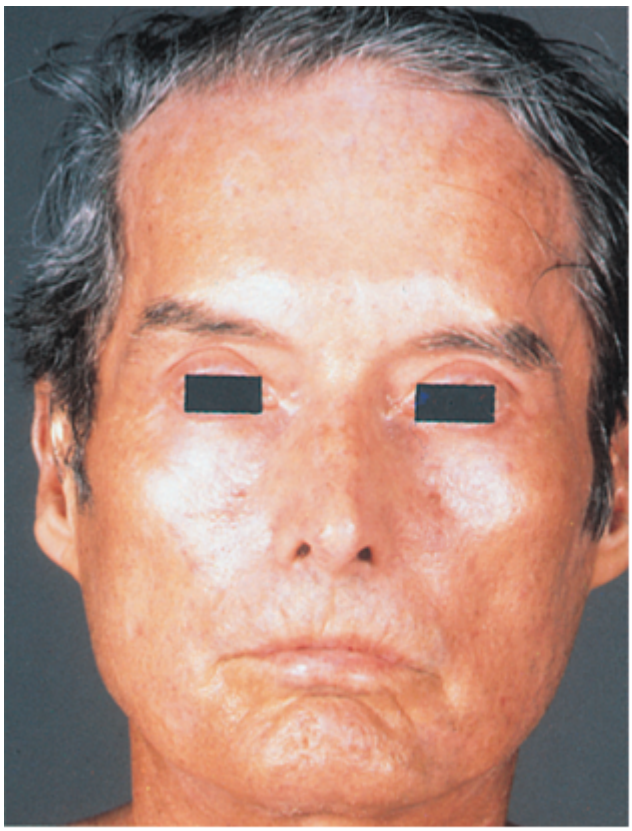
Scleroderma showing acral sclerosis and focal digital ulcers.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 55-8

Scleroderma often eventuates in development of an expressionless, masklike facies.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Morphea is characterized by localized thickening and sclerosis of skin; it dominates on the trunk. This disorder may affect children or adults. Morphea begins as erythematous or flesh-colored plaques that become sclerotic, develop central hypopigmentation, and have an erythematous border. In most cases, patients have one or a few lesions, and the disease is termed *localized morphea*. In some patients, widespread cutaneous lesions may occur without systemic involvement (*generalized morphea*). Many adults with generalized morphea have concomitant rheumatic or other autoimmune disorders. Skin biopsy of morphea is generally indistinguishable from that of scleroderma. Scleroderma and morphea are usually quite resistant to therapy. For this reason, physical therapy to prevent joint contractures and to maintain function is employed and is often helpful. Treatment options for early, rapidly progressive disease include phototherapy (UVA1 [ultraviolet A1 irradiation] or PUVA [psoralens + ultraviolet A irradiation]) or methotrexate (15–20 mg/week) alone or in combination with daily glucocorticoids.

Diffuse fasciitis with eosinophilia is a clinical entity that can sometimes be confused with scleroderma. There is usually a sudden onset of swelling, induration, and erythema of the extremities, frequently following significant physical exertion. The proximal portions of the extremities (upper arms, forearms, thighs, calves) are more often involved than are the hands and feet. While the skin is indurated, it usually displays a woody, dimpled, or "pseudocellulite" appearance rather than being bound down as in scleroderma; contractures may occur early secondary to fascial involvement. The latter may also cause muscle groups to be separated and veins to appear depressed (i.e., the "groove sign"). These skin findings are accompanied by peripheral-blood eosinophilia, increased erythrocyte sedimentation rate, and sometimes hypergammaglobulinemia. Deep biopsy of affected areas of skin reveals inflammation and thickening of the deep fascia overlying

muscle. An inflammatory infiltrate composed of eosinophils and mononuclear cells is usually found. Patients with eosinophilic fasciitis appear to be at increased risk for developing bone marrow failure or other hematologic abnormalities. While the ultimate course of eosinophilic fasciitis is uncertain, many patients respond favorably to treatment with prednisone in doses of 40–60 mg/d.

The *eosinophilia-myalgia syndrome*, a disorder with epidemic numbers of cases reported in 1989 and linked to ingestion of L-tryptophan manufactured by a single company in Japan, is a multisystem disorder characterized by debilitating myalgias and absolute eosinophilia in association with varying combinations of arthralgias, pulmonary symptoms, and peripheral edema. In a later phase (3–6 months after initial symptoms), these patients often develop localized sclerodermatous skin changes, weight loss, and/or neuropathy **(Chap. 353)**. The precise cause of this syndrome, which may resemble other sclerotic skin conditions, is unknown. However, the implicated lots of L-tryptophan contained the contaminant 1,1-ethylidene bis[tryptophan]. This contaminant may be pathogenic or may be a marker for another substance that provokes the disorder.

FURTHER READING

Bolognia JL et al (eds): *Dermatology*, 4th ed. Philadelphia, Elsevier, 2018.

Goldsmith LA et al (eds): *Fitzpatrick's Dermatology in General Medicine*, 8th ed. New York, McGraw-Hill, 2012.

Hammers CM, Stanley JR: Mechanisms of disease: Pemphigus and bullous pemphigoid. Annu Rev Pathol 11:175, 2016. [PubMed: 26907530]

Schmidt E, Zillikens D: Pemphigoid diseases. Lancet 381:320, 2013. [PubMed: 23237497]

Chapter 56: Cutaneous Drug Reactions

Robert Micheletti; Misha Rosenbach; Bruce U. Wintroub; Kanade Shinkai

INTRODUCTION

Cutaneous reactions are among the most frequent adverse reactions to drugs. Most are benign, but a few can be life threatening. Prompt recognition of severe reactions, drug withdrawal, and appropriate therapeutic interventions can minimize toxicity. This chapter focuses on adverse cutaneous reactions to systemic medications; it covers their incidence, patterns, and pathogenesis, and provides some practical guidelines on treatment, assessment of causality, and future use of drugs.

USE OF PRESCRIPTION DRUGS IN THE UNITED STATES

In the United States, more than 3 billion prescriptions for >60,000 drug products, which include >2000 different active agents, are dispensed annually. Hospital inpatients alone annually receive about 120 million courses of drug therapy, and half of adult Americans receive prescription drugs on a regular outpatient basis. Adverse effects of a prescription medication may result in 4.5 million urgent or emergency care visits each year in the United States. Many patients use over-the-counter medicines that may cause adverse cutaneous reactions.

INCIDENCE OF CUTANEOUS REACTIONS

Several large cohort studies established that acute cutaneous reactions to drugs affect about 3% of hospitalized patients. Reactions usually occur a few days to 4 weeks after initiation of therapy.

Many drugs of common use are associated with a 1–2% rate of rashes during premarketing clinical trials. The risk is often higher when medications are used in general, unselected populations. The rate may reach 3–7% for amoxicillin, sulfamethoxazole, many anticonvulsants, and anti-HIV agents.

In addition to acute eruptions, a variety of skin diseases can be induced or exacerbated by prolonged use of drugs (e.g., pruritus, pigmentation, nail or hair disorders, psoriasis, bullous pemphigoid, photosensitivity, and even cutaneous neoplasms). These drug reactions are not frequent, but neither their incidence nor their impact on public health has been evaluated.

In a series of 48,005 inpatients over a 20-year period, morbilliform rash (91%) and urticaria (6%) were the most frequent skin reactions. Severe reactions are too rare to be detected in such cohorts. Although rare,

severe cutaneous reactions to drugs have an important impact on health because of significant sequelae, including mortality. Adverse drug rashes are responsible for hospitalization, increase the duration of hospital stay, and can be life threatening. Some populations are at increased risk of drug reactions, including elderly patients, patients with autoimmune disease, hematopoietic stem cell transplant recipients, and those with acute Epstein-Barr virus (EBV) or human immunodeficiency virus (HIV) infection. The pathophysiology underlying this association is unknown but may be related to immunocompromise or immune dysregulation. Individuals with advanced HIV disease (e.g., CD4 T lymphocyte count <200 cells/μL) have a 40- to 50-fold increased risk of adverse reactions to sulfamethoxazole **(Chap. 197)** and increased risk of severe hypersensitivity reactions.

PATHOGENESIS OF DRUG REACTIONS

Adverse cutaneous responses to drugs can arise as a result of immunologic or nonimmunologic mechanisms.

NONIMMUNOLOGIC DRUG REACTIONS

Examples of nonimmunologic drug reactions are pigmentary changes due to dermal accumulation of medications or their metabolites, alteration of hair follicles by antimetabolites and signaling inhibitors, and lipodystrophy associated with metabolic effects of anti-HIV medications. These side effects are predictable and sometimes can be prevented.

IMMUNOLOGIC DRUG REACTIONS

Evidence suggests an immunologic basis for most acute drug eruptions. Drug reactions may result from immediate release of preformed mediators (e.g., urticaria, anaphylaxis), antibody-mediated reactions, immune complex deposition, and antigen-specific responses. Drug-specific T cell clones can be derived from the blood or from skin lesions of patients with a variety of drug allergies, strongly suggesting that these T cells mediate drug allergy in an antigen-specific manner. Specific clones are generated by medications that are frequently a cause of drug eruptions: penicillin G, amoxicillin, cephalosporins, sulfamethoxazole, phenobarbital, carbamazepine, and lamotrigine. Both CD4 and CD8 clones have been obtained; however, their specific roles in drug allergy have not been elucidated. Drug presentation to T cells is major histocompatibility complex (MHC)-restricted and likely involves drug-peptide complex recognition by specific T cell receptors (TCRs).

Once a drug has induced an immune response, the final phenotype of the reaction is determined by the nature of effectors: cytotoxic (CD8+) T cells in blistering and certain hypersensitivity reactions, chemokines for reactions mediated by neutrophils or eosinophils, and B cell collaboration for production of specific antibodies for urticarial reactions. Immunologic reactions have recently been classified into further subtypes that provide a useful framework for designating adverse drug reactions based on involvement of specific immune pathways **(Table 56-1)**.

TABLE 56-1

Classification of Adverse Drug Reactions Based on Immune Pathway

Type	Key Pathway	Key Immune Mediators	Adverse Drug Reaction Type
Type I	IgE	IgE	Urticaria, angioedema, anaphylaxis
Type II	IgG-mediated cytotoxicity	IgG	Drug-induced hemolysis, thrombocytopenia (e.g., penicillin)
Type III	Immune complex	IgG + antigen	Vasculitis, serum sickness, drug-induced lupus
Type IVa	T lymphocyte–mediated macrophage inflammation	IFN-γ, TNF-α T_H1 cells	Tuberculin skin test, contact dermatitis
Type IVb	T lymphocyte–mediated eosinophil inflammation	IL-4, IL-5, IL-13 T_H2 cells Eosinophils	DIHS Morbilliform eruption
Type IVc	T lymphocyte–mediated cytotoxic T lymphocyte inflammation	Cytotoxic T lymphocytes Granzyme Perforin Granulysin (SJS/TEN] only)	SJS/TEN Morbilliform eruption
Type IVd	T lymphocyte–mediated neutrophil inflammation	CXCL8, IL-17, GM-CSF Neutrophils	AGEP

Abbreviations: AGEP, acute generalized exanthematous pustulosis; DIHS, drug-induced hypersensitivity syndrome; GM-CSF, granulocyte-macrophage colony-stimulating factor; IFN, interferon; IL, interleukin; SJS, Stevens-Johnson syndrome; TEN, toxic epidermal necrolysis; TNF, tumor necrosis factor.

Immediate Reactions

Immediate reactions depend on the release of mediators of inflammation by tissue mast cells or circulating basophils. These mediators include histamine, leukotrienes, prostaglandins, bradykinins, platelet-activating factor, enzymes, and proteoglycans. Drugs can trigger mediator release either directly ("anaphylactoid" reaction) or through IgE-specific antibodies. These reactions usually manifest in the skin and gastrointestinal, respiratory, and cardiovascular systems **(Chap. 346)**. Primary symptoms and signs include pruritus, urticaria, nausea, vomiting, abdominal cramps, bronchospasm, laryngeal edema, and, occasionally, anaphylactic shock with hypotension and death. They occur within minutes of drug exposure. Nonsteroidal anti-inflammatory drugs (NSAIDs), including aspirin, and radiocontrast media are frequent causes of direct mast cell degranulation or anaphylactoid reactions, which can occur on first exposure. Penicillins and muscle relaxants used in general anesthesia are the most frequent causes of IgE-dependent reactions to drugs, which require prior sensitization. Release of mediators is triggered when polyvalent drug protein conjugates cross-link IgE molecules fixed to sensitized cells. Certain routes of administration favor different clinical patterns (e.g., gastrointestinal effects from oral route, circulatory effects from intravenous route).

Immune Complex–Dependent Reactions

Serum sickness is produced by tissue deposition of circulating immune complexes with consumption of complement. It is characterized by fever, arthritis, nephritis, neuritis, edema, and an urticarial, papular, or purpuric rash **(Chap. 356)**. First described following administration of nonhuman sera, it currently occurs in the setting of monoclonal antibodies and similar medications. In classic serum sickness, symptoms develop 6 or more days after drug exposure, the latent period representing the time needed to synthesize antibody. Vasculitis, a relatively rare complication of drugs, may also be a result of immune complex deposition **(Chap. 356)**. Cephalosporin and other medications, including monoclonal antibodies such as infliximab, rituximab, and omalizumab, may be associated with clinically similar "serum sickness–like" reactions. The mechanism of this reaction is unknown but is unrelated to immune complex formation and complement activation.

Delayed Hypersensitivity

While not completely understood, delayed hypersensitivity directed by drug-specific T cells is an important mechanism underlying the most common drug eruptions, that is, morbilliform eruptions, and also rare and severe forms such as drug-induced hypersensitivity syndrome (DIHS) (also known as drug rash with eosinophilia and systemic symptoms [DRESS]), acute generalized exanthematous pustulosis (AGEP), Stevens-Johnson syndrome (SJS), and toxic epidermal necrolysis (TEN) (Table 56-1). Drug-specific T cells have been detected in these types of drug eruptions. In TEN, skin lesions contain T lymphocytes reactive to autologous lymphocytes and keratinocytes in a drug-specific, human leukocyte antigen (HLA)-restricted, and perforin/granzyme-mediated pathway.

The mechanism(s) by which medications result in T cell activation is unknown. Two hypotheses prevail: first, that the antigens driving these reactions may be the native drug itself or components of the drug covalently complexed with endogenous proteins, presented in association with HLA molecules to T cells through the classic antigen presentation pathway or, alternatively, through direct interaction of the drug/metabolite with the TCR or peptide-loaded HLA (e.g., the pharmacologic interaction of drugs with immune receptors, or p-i

hypothesis). Recent x-ray crystallography data characterizing binding between specific HLA molecules to particular drugs known to cause hypersensitivity reactions demonstrate unique alterations to the MHC peptide-binding groove, suggesting a molecular basis for T cell activation in the development of hypersensitivity reactions.

GENETIC FACTORS AND CUTANEOUS DRUG REACTIONS



Genetic determinants may predispose individuals to severe drug reactions by affecting either drug metabolism or immune responses to drugs. Polymorphisms in cytochrome P450 enzymes, drug acetylation, methylation (such as thiopurine methyltransferase activity and azathioprine), and other forms of metabolism (such as glucose-6-phosphate dehydrogenase and dapsone) may increase susceptibility to drug toxicity or underdosing, highlighting a role for differential pharmacokinetic or pharmacodynamic effects. The value of routine screening of P450 enzymes has not been determined, though its cost-effectiveness in certain populations (e.g., patients with seizure disorder) has been suggested.

Associations between drug hypersensitivities and HLA haplotypes suggest a key role for immune mechanisms. Hypersensitivity to the anti-HIV medication abacavir is strongly associated with HLA-B*57:01 (**Chap. 197**). In Taiwan, within a homogeneous Han Chinese population, a 100% association was observed between SJS/TEN (but not DIHS) related to carbamazepine and HLA-B*15:02. In the same population, another 100% association was found between HLA-B*58:01 and SJS, TEN, or DIHS related to allopurinol. These associations are drug and phenotype specific; that is, HLA-specific T cell stimulation by medications leads to distinct reactions. However, the strong associations found in Taiwan have not been observed in other countries with more heterogeneous populations.

GLOBAL CONSIDERATIONS



Recognition of HLA associations with drug hypersensitivity has resulted in recommendations to screen high-risk populations. Genetic screening for HLA-B*57:01 to prevent abacavir hypersensitivity, which carries a 100% negative predictive value when patch test confirmed and 55% positive predictive value generalizable across races, is becoming the clinical standard of care worldwide (number needed to treat = 13). The U.S. Food and Drug Administration has recommended HLA-B*15:02 screening of Asian individuals prior to a new prescription of carbamazepine. The American College of Rheumatology has recommended HLA-B*58:01 screening of Han Chinese patients prescribed allopurinol. To date, screening for a single HLA (but not multiple HLA haplotypes) in specific populations has been determined to be cost-effective.

Several investigators have proposed that specific HLA haplotypes associated with drug hypersensitivity indeed play a pathogenic role; stimulation of carbamazepine-specific cytotoxic T lymphocytes (CTLs) in the context of HLA-B*15:02 results in production of a putative mediator of keratinocyte necrosis in TEN. Other studies have identified CTLs reactive to carbamazepine that use highly restricted V-alpha and V-beta TCR

repertoires in patients with carbamazepine hypersensitivity that are not found in carbamazepine-tolerant individuals. Genetic testing for specific HLA haplotypes and functional screening for TCR repertoire to identify patients at risk is becoming more widely available and heralds the era of personalized medicine and pharmacogenomics.

CLINICAL PRESENTATION OF CUTANEOUS DRUG REACTIONS

NONIMMUNE CUTANEOUS REACTIONS

Exacerbation or Induction of Dermatologic Diseases

A variety of drugs can exacerbate preexisting diseases or induce—or unmask—a disease that may or may not disappear after withdrawal of the inducing medication. For example, NSAIDs, lithium, beta blockers, tumor necrosis factor (TNF) antagonists, interferon (IFN) α, and angiotensin-converting enzyme (ACE) inhibitors can exacerbate plaque psoriasis, whereas antimalarials and withdrawal of systemic glucocorticoids can worsen pustular psoriasis. The situation of TNF-α inhibitors is unusual, as this class of medications is used to treat psoriasis; however, they may induce psoriasis (especially palmoplantar) in patients being treated for other conditions. Acne may be induced by glucocorticoids, androgens, lithium, and antidepressants. Follicular papular or pustular eruptions of the face and trunk resembling acne frequently occur with epidermal growth factor receptor (EGFR) antagonists. The severity of the eruption correlates with a better anticancer effect. This rash is typically responsive to and prevented by tetracycline antibiotics.

Several medications induce or exacerbate autoimmune disease. Interleukin (IL) 2, IFN-α, and anti-TNF-α are associated with new-onset systemic lupus erythematosus (SLE). Drug-induced lupus is classically marked by antinuclear and antihistone antibodies and, in some cases, anti-double-stranded DNA (D-penicillamine, anti-TNF-α) or perinuclear anti-neutrophil cytoplasmic antibody (p-ANCA) (minocycline) antibodies. Subacute lupus erythematosus (SCLE) can be induced by a growing list of drugs, including thiazide diuretics, TNF-inhibitors, terbinafine, and minocycline. IFN and TNF-inhibitors can induce granulomatous disease and sarcoidosis. Autoimmune blistering diseases may be drug induced as well: pemphigus by D-penicillamine and ACE inhibitors, bullous pemphigoid by furosemide and PD-1 inhibitors, and linear IgA bullous dermatosis by vancomycin. Other medications may cause highly specific cutaneous reactions. Gadolinium contrast has been associated with nephrogenic systemic fibrosis, a condition of sclerosing skin with rare internal organ involvement; advanced renal compromise may be an important risk factor. Granulocyte colony-stimulating factor, azacitidine, all-trans retinoic acid, and the *FLT3*-inhibitor class of drugs may induce neutrophilic dermatoses. In this setting, the hypothesis that a drug may be responsible should always be considered, even after the treatment is complete. In addition, reactions may develop in cases of long-term medication therapy due to small changes in dosing or host metabolism. Resolution of the cutaneous reaction may be delayed upon discontinuation of the medication.

Photosensitivity Eruptions

Photosensitivity eruptions are usually most marked in sun-exposed areas, but they may extend to sun-protected areas. The mechanism is almost always phototoxicity. Phototoxic reactions resemble sunburn and can occur with first exposure to a drug. Blistering may occur in drug-related pseudoporphyria, most commonly with NSAIDs. The severity of the reaction depends on the tissue level of the drug, its efficiency as a photosensitizer, and the extent of exposure to the activating wavelengths of ultraviolet (UV) light **(Chap. 57)**.

Common orally administered photosensitizing drugs include fluoroquinolones, tetracycline antibiotics, and trimethoprim/sulfamethoxazole. Other drugs less frequently implicated are chlorpromazine, thiazides, NSAIDs, and BRAF inhibitors. Voriconazole may result in severe photosensitivity, accelerated photoaging, and cutaneous carcinogenesis.

Because UV-A and visible light, which trigger these reactions, are not easily absorbed by nonopaque sunscreens and are transmitted through window glass, photosensitivity reactions may be difficult to block. Photosensitivity reactions abate with removal of either the drug or UV radiation, use of sunscreens that block UV-A light, and treatment of the reaction as one would a sunburn. Rarely, individuals develop persistent reactivity to light, necessitating long-term avoidance of sun exposure. Some chemotherapeutic agents, such as methotrexate, can induce a UV-recall reaction characterized by an erythematous, slightly scaly eruption at sites of prior severe sun exposure.

Pigmentation Changes

Drugs, either systemic or topical, may cause a variety of pigmentary changes in the skin by triggering melanocyte production of melanin (as in the case of oral contraceptives causing melasma) or due to deposition of drug or drug metabolites. Long-term minocycline and amiodarone may cause blue-gray pigmentation. Phenothiazine, gold, and bismuth result in gray-brown pigmentation of sun-exposed areas. Numerous cancer chemotherapeutic agents may be associated with characteristic patterns of pigmentation (e.g., bleomycin, busulfan, daunorubicin, cyclophosphamide, hydroxyurea, fluorouracil, and methotrexate). Clofazimine causes a drug-induced lipofuscinosis with characteristic red-brown coloration. Hyperpigmentation of the face, mucous membranes, and pretibial and subungual areas occurs with antimalarials. Quinacrine causes generalized yellow discoloration. Pigmentation changes may also occur in mucous membranes (busulfan, bismuth), conjunctiva (chlorpromazine, thioridazine, imipramine, clomipramine), nails (zidovudine, doxorubicin, cyclophosphamide, bleomycin, fluorouracil, hydroxyurea), hair, and teeth (tetracyclines).

Warfarin Necrosis of Skin

This rare reaction (0.01–0.1%) usually occurs between the third and tenth days of therapy with warfarin, usually in women. Common sites are breasts, thighs, and buttocks **(Fig. 56-1)**. Lesions are sharply demarcated, erythematous, or purpuric, and may progress to form large, hemorrhagic bullae with necrosis and eschar formation.

FIGURE 56-1

Warfarin necrosis involving the breasts.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Warfarin anticoagulation in protein C or S deficiency causes an additional reduction in already low circulating levels of endogenous anticoagulants, permitting hypercoagulability and thrombosis in the cutaneous microvasculature, with consequent areas of necrosis. Heparin-induced necrosis may have clinically similar features but is probably due to heparin-induced platelet aggregation with subsequent occlusion of blood vessels; it can affect areas adjacent to the injection site or more distant sites if infused.

Warfarin-induced cutaneous necrosis is treated with vitamin K, heparin, surgical debridement, and intensive wound care. Treatment with protein C concentrates may also be helpful. Newer anticoagulants such as dabigatran etexilate may avoid warfarin necrosis in high-risk patients.

Drug-Induced Hair Disorders • DRUG-INDUCED HAIR LOSS

Medications may affect hair follicles at two different phases of their growth cycle: anagen (growth) or telogen (resting). *Anagen effluvium* occurs within days of drug administration, especially with antimetabolite or other chemotherapeutic drugs. In contrast, in *telogen effluvium*, the delay is 2–4 months following initiation of a new medication. Both present as diffuse, nonscarring alopecia most often reversible after discontinuation of the responsible agent.

A considerable number of drugs have been associated with hair loss. These include antineoplastic agents (alkylating agents, bleomycin, vinca alkaloids, platinum compounds), anticonvulsants (carbamazepine, valproate), beta blockers, antidepressants, antithyroid drugs, IFNs, oral contraceptives, and cholesterol-lowering agents.

DRUG-INDUCED HAIR GROWTH

Medications may also cause hair growth. Hirsutism is an excessive growth of terminal hair with masculine hair growth pattern in a female, most often on the face and trunk, due to androgenic stimulation of hormone-sensitive hair follicles (anabolic steroids, oral contraceptives, testosterone, corticotropin). Hypertrichosis is a

distinct pattern of hair growth, not in a masculine pattern, typically located on the forehead and temporal regions of the face. Drugs responsible for hypertrichosis include anti-inflammatory drugs, glucocorticoids, vasodilators (diazoxide, minoxidil), diuretics (acetazolamide), anticonvulsants (phenytoin), immunosuppressive agents (cyclosporine A), psoralens, and zidovudine.

Changes in hair color or structure are uncommon adverse effects from medications. Hair discoloration may occur with chloroquine, IFN-α, chemotherapeutic agents, and tyrosine kinase inhibitors. Changes in hair structure have been observed in patients given EGFR inhibitors, BRAF inhibitors, tyrosine kinase inhibitors, and acitretin.

Drug-Induced Nail Disorders

Drug-related nail disorders usually involve all 20 nails and need months to resolve after withdrawal of the medication. The pathogenesis is most often toxic. Drug-induced nail changes include Beau's line (transverse depression of the nail plate), onycholysis (detachment of the distal part of the nail plate), onychomadesis (detachment of the proximal part of the nail plate), pigmentation, and paronychia (inflammation of periungual skin).

ONYCHOLYSIS

Onycholysis occurs with tetracyclines, fluoroquinolones, retinoids, NSAIDs, and others, including many chemotherapeutic agents, and may be triggered by exposure to sunlight.

ONYCHOMADESIS

Onychomadesis is caused by temporary arrest of nail matrix mitotic activity. Common drugs reported to induce onychomadesis include carbamazepine, lithium, retinoids, and chemotherapeutic agents.

PARONYCHIA

Paronychia and multiple pyogenic granuloma with progressive and painful periungual abscess of fingers and toes are side effects of systemic retinoids, lamivudine, indinavir, and anti-EGFR monoclonal antibodies.

NAIL DISCOLORATION

Some drugs—including anthracyclines, taxanes, fluorouracil, psoralens, and zidovudine—may induce nail bed hyperpigmentation through melanocyte stimulation. It appears to be reversible and dose dependent.

Toxic Erythema of Chemotherapy and Other Chemotherapy Reactions

Because many agents used in cancer chemotherapy inhibit cell division, rapidly proliferating elements of the skin, including hair, mucous membranes, and appendages, are sensitive to their effects. A broad spectrum of chemotherapy-related skin toxicities have been reported, including neutrophilic eccrine hidradenitis, sterile cellulitis, exfoliative dermatitis, and flexural erythema; recent nomenclature classifies these under the unifying diagnosis of toxic erythema of chemotherapy (TEC) (**Fig. 56-2**). Acral erythema is marked by dysesthesia and an erythematous, edematous eruption of the palms and soles. Common causes include cytarabine, doxorubicin, methotrexate, hydroxyurea, fluorouracil, and capecitabine.

FIGURE 56-2

Toxic erythema of chemotherapy.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

The recent introduction of many new monoclonal antibody and small molecular signaling inhibitors for the treatment of cancer has been accompanied by numerous reports of skin and hair toxicity; only the most common of these are mentioned here. EGFR antagonists induce follicular eruptions and nail toxicity after a mean interval of 10 days in a majority of patients. Xerosis, eczematous eruptions, acneiform eruptions, and pruritus are common. Erlotinib is associated with marked hair textural changes. Sorafenib, a tyrosine kinase inhibitor, may result in follicular eruptions and focal bullous eruptions at palmoplantar, flexural sites or areas of frictional pressure. BRAF inhibitors are associated with photosensitivity, palmoplantar hyperkeratosis, hair curling, dyskeratotic (Grover's-like) rash, hyperkeratotic benign cutaneous neoplasms, and keratoacanthoma-like squamous cell carcinomas. Rash, pruritus, and vitiliginous depigmentation have been reported in association with ipilimumab (anti-CTLA4) treatment. Up to 50% of patients experience immune-mediated skin eruptions, including granulomatous reactions, dermatomyositis, panniculitis, and vasculitis.

IMMUNE CUTANEOUS REACTIONS: COMMON

Maculopapular Eruptions

Morbilliform or maculopapular eruptions (Fig. 56-3) are the most common of all drug-induced reactions, often start on the trunk or intertriginous areas, and consist of blanching erythematous macules and papules that are symmetric and confluent. Nonblanching, dusky, or bright-red macules should raise concern for a more severe reaction. Involvement of mucous membranes is rare and should prompt consideration of SJS. Facial involvement in morbilliform eruptions is also uncommon, and the presence of extensive facial lesions with facial edema suggests DIHS. Diagnosis of morbilliform eruptions is rarely assisted by laboratory testing. Skin biopsy often shows nonspecific inflammatory changes.

FIGURE 56-3

Morbilliform drug eruption.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Morbilliform eruptions may be associated with moderate to severe pruritus and fever. A viral exanthem is another differential diagnostic consideration, especially in children, and graft-versus-host disease is also a consideration in the proper clinical setting. Absence of enanthems; absence of ear, nose, throat, and upper respiratory tract symptoms; and polymorphism of the skin lesions support a drug rather than a viral eruption. Common offenders include aminopenicillins, cephalosporins, antibacterial sulfonamides, allopurinol, and antiepileptic drugs. Beta blockers, calcium channel blockers, and ACE inhibitors are rarely the culprit; however, any drug can cause a morbilliform exanthem. Certain medications carry very high rates of morbilliform eruption, including nevirapine and lamotrigine, even in the absence of DIHS reactions. Lamotrigine morbilliform rash is associated with higher starting doses, rapid dose escalation, concomitant use of valproate (which increases lamotrigine levels and half-life), and use in children.

Maculopapular reactions usually develop within 1 week of initiation of therapy and last less than 2 weeks. Occasionally, these eruptions resolve despite continued use of the responsible drug. Because the eruption may also worsen, the suspect drug should be discontinued unless it is essential. It is important to note that

the rash may continue to progress for a few days up to 1 week following medication discontinuation. Oral antihistamines and emollients may help relieve pruritus. Short courses of potent topical glucocorticoids can reduce inflammation and symptoms. Systemic glucocorticoid treatment is rarely indicated.

Pruritus

Pruritus is associated with almost all drug eruptions and, in some cases, may represent the only symptom of the adverse cutaneous reaction. It may be alleviated by antihistamines such as hydroxyzine or diphenhydramine. Pruritus stemming from specific medications may require distinct treatment, such as selective opiate antagonists for opiate-related pruritus.

Urticaria/Angioedema/Anaphylaxis

Urticaria, the second most frequent type of cutaneous reaction to drugs, is characterized by pruritic, red wheals of varying size rarely lasting more than 24 hours. It has been observed in association with nearly all drugs, most frequently ACE inhibitors, aspirin, NSAIDs, penicillin, and blood products. However, medications account for no more than 10–20% of acute urticaria cases. Deep edema within dermal and subcutaneous tissues is known as angioedema and may involve respiratory and gastrointestinal mucous membranes. Urticaria and angioedema may be part of a life-threatening anaphylactic reaction.

Drug-induced urticaria may be caused by three mechanisms: an IgE-dependent mechanism, circulating immune complexes (serum sickness), and nonimmunologic activation of effector pathways. IgE-dependent urticarial reactions usually occur within 36 hours of drug exposure but can occur within minutes. Immune complex–induced urticaria associated with serum sickness–like reactions usually occur 6–12 days after first exposure. In this syndrome, the urticarial eruption (typically polycyclic plaques over distal joints) may be accompanied by fever, hematuria, arthralgias, hepatic dysfunction, and neurologic symptoms. Certain drugs, such as NSAIDs, ACE inhibitors, angiotensin II antagonists, radiographic dye, and opiates, may induce urticarial reactions, angioedema, and anaphylaxis in the absence of drug-specific antibodies through direct mast-cell degranulation.

Radiocontrast agents are a common cause of urticaria and, in rare cases, can cause anaphylaxis. High-osmolality radiocontrast media are about five times more likely to induce urticaria (1%) or anaphylaxis than are newer low-osmolality media. About one-third of those with mild reactions to previous exposure react on reexposure. Pretreatment with prednisone and diphenhydramine reduces reaction rates.

The treatment of urticaria or angioedema depends on the severity of the reaction. In severe cases with respiratory or cardiovascular compromise, epinephrine and intravenous glucocorticoids are the mainstay of therapy. For patients with urticaria without symptoms of angioedema or anaphylaxis, drug withdrawal and oral antihistamines are usually sufficient. Future drug avoidance is recommended; rechallenge, especially in individuals with severe reactions, should only occur in an intensive care setting.

Anaphylactoid Reactions

Vancomycin is associated with red man syndrome, a histamine-related anaphylactoid reaction characterized by flushing, diffuse maculopapular eruption, and hypotension. In rare cases, cardiac arrest may be associated with rapid IV infusion of the medication.

Irritant/Allergic Contact Dermatitis

Patients using topical medications may develop an irritant or allergic contact dermatitis to the medication itself or to a preservative or other component of the formulation. Reactions to neomycin sulfate, bacitracin, and polymyxin B are common. Contact dermatitis may be seen to adhesive tapes, leading to irritation or blisters around ports and IV sites **(Fig. 56-4)**. Harsh disinfectant skin cleansers may lead to localized irritant dermatitis.

FIGURE 56-4

Allergic contact dermatitis (bullous) due to adhesive tape.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Fixed Drug Eruptions

These less common reactions are characterized by one or more sharply demarcated, dull red to brown lesions, sometimes with central dusky violaceous erythema and central bulla **(Fig. 56-5)**. Hyperpigmentation often results after resolution of the acute inflammation. With rechallenge, the process recurs in the same (fixed) location but may spread to new areas as well. Lesions often involve the lips, hands, legs, face, genitalia, and oral mucosa, and cause a burning sensation. Most patients have multiple lesions. Fixed drug

eruptions have been associated with pseudoephedrine (frequently a nonpigmenting reaction), phenolphthalein (in laxatives), sulfonamides, tetracyclines, NSAIDs, barbiturates, and others.

FIGURE 56-5

Fixed drug eruption.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

IMMUNE CUTANEOUS REACTIONS: RARE AND SEVERE

Drug-Induced Hypersensitivity Syndrome

DIHS is a systemic drug reaction also known as DRESS (drug reaction with eosinophilia and systemic symptoms) syndrome; since eosinophilia is not always present, the term *DIHS* is now preferred. Clinically, DIHS presents with a prodrome of fever and flu-like symptoms for several days, followed by the appearance of a diffuse morbilliform eruption usually involving the face **(Fig. 56-6)**. Facial swelling and hand/foot swelling are often present. Systemic manifestations include lymphadenopathy, fever, and leukocytosis (often with eosinophilia or atypical lymphocytosis), as well as hepatitis, nephritis, pneumonitis, myositis, and gastroenteritis, in descending order. Distinct patterns of timing of onset and organ involvement may exist; for example allopurinol classically induces DIHS with renal involvement, cardiac and lung involvements are more common with minocycline, gastrointestinal involvement is almost exclusively seen with abacavir, and some medications typically lack eosinophilia (abacavir, dapsone, lamotrigine). The cutaneous reaction usually begins 2–8 weeks after the drug is started and persists after drug cessation. Signs and symptoms may continue for several weeks, especially those associated with hepatitis. The eruption recurs with rechallenge, and cross-reactions among aromatic anticonvulsants, including phenytoin, carbamazepine, and

phenobarbital, are common. Other drugs causing DIHS include antibacterial sulfonamides and other antibiotics. Hypersensitivity to reactive drug metabolites, hydroxylamine for sulfamethoxazole and arene oxide for aromatic anticonvulsants, may be involved in the pathogenesis of DIHS. Reactivation of herpes viruses, in particular human herpesviruses 6 and 7, EBV, and cytomegalovirus (CMV), has been frequently reported in this syndrome, although the causal role of viral infection has been debated. Recent research suggests that inciting drugs may reactivate quiescent herpes viruses, resulting in expansion of viral-specific CD8+ T lymphocytes and subsequent end-organ damage. Viral reactivation may be associated with a worse clinical prognosis. Mortality rates as high as 10% have been reported, with most fatalities resulting from liver failure. Systemic glucocorticoids (1.5–2 mg/kg/d prednisone equivalent) should be started and tapered slowly over 8–12 weeks, during which time clinical symptoms and labs (including complete blood count with differential, basic metabolic panel, and liver function tests) should be followed carefully. A steroid-sparing agent such as mycophenolate mofetil may be indicated in cases of rapid recurrence upon steroid taper. In all cases, immediate withdrawal of the suspected culprit drug is required. Given the severe long-term complications of myocarditis, patients should undergo cardiac evaluation in cases of severe DIHS or if heart involvement is suspected due to hypotension or arrhythmia. Patients should be closely monitored for resolution of organ dysfunction and for development of late-onset autoimmune thyroiditis and diabetes (up to 6 months).

FIGURE 56-6

Drug-induced hypersensitivity syndrome/drug rash with eosinophilia and systemic symptoms (DIHS/DRESS).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis

SJS and *TEN* are characterized by blisters and mucosal/epidermal detachment resulting from full-thickness epidermal necrosis in the absence of substantial dermal inflammation. The term *Stevens-Johnson syndrome* (SJS) describes cases in which the total body surface area of blistering and eventual detachment is <10% **(Fig. 56-7)**. The term *Stevens-Johnson syndrome/toxic epidermal necrolysis (SJS/TEN) overlap* is used to describe cases with 10–30% epidermal detachment **(Fig. 56-8)**, and *TEN* is used to describe cases with >30% detachment **(Figs. 56-9 and 56-10)**.

FIGURE 56-7

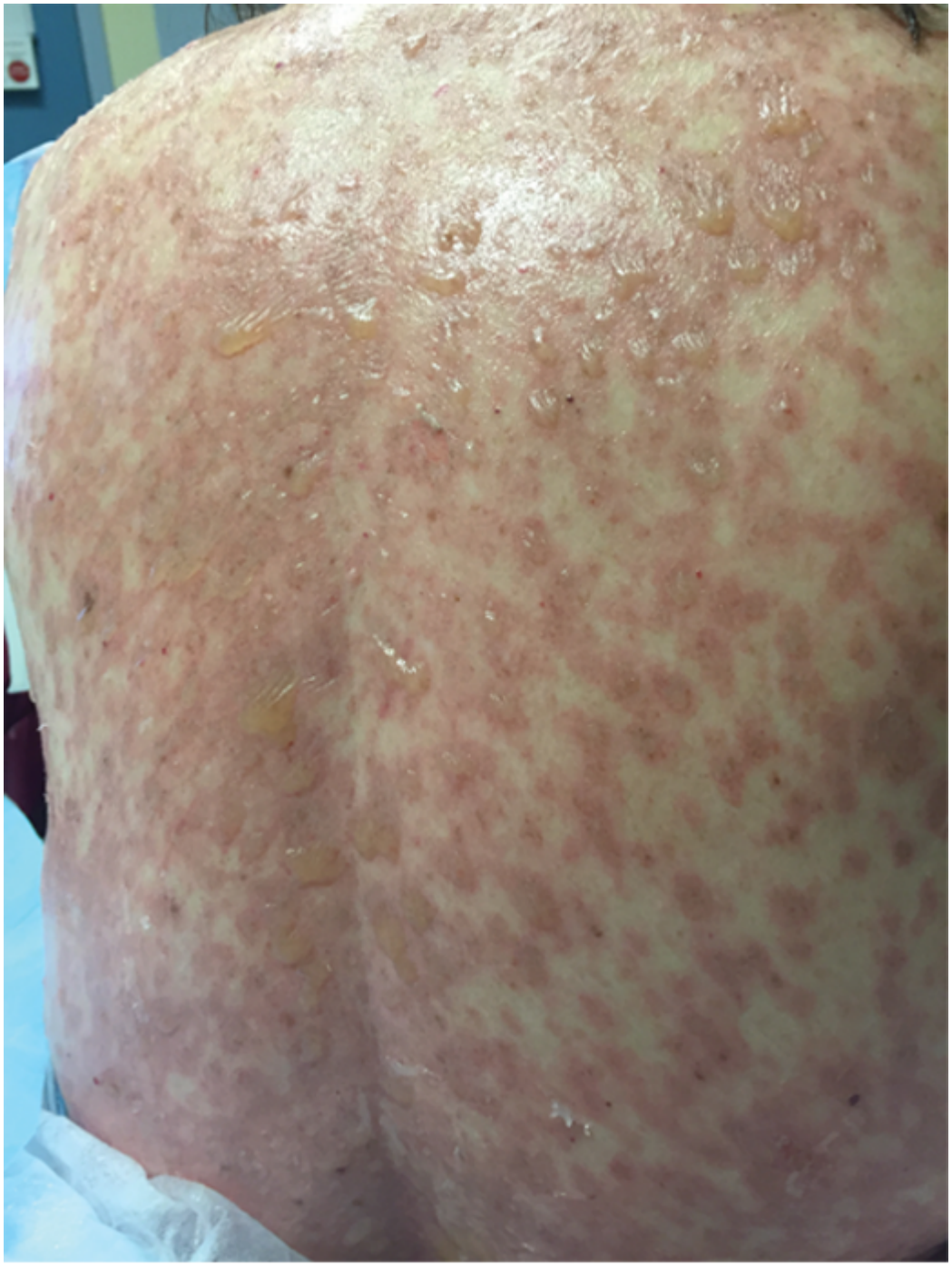
Stevens-Johnson syndrome (SJS).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 56-8

SJS-TEN overlap.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 56-9

Toxic epidermal necrolysis, hand.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 56-10

Toxic epidermal necrolysis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Other blistering eruptions with concomitant mucositis may be confused with SJS/TEN. Erythema multiforme (EM) associated with herpes simplex virus is characterized by painful mucosal erosions and target lesions, typically with an acral distribution and limited skin detachment. *Mycoplasma* infection in children causes a clinically distinct presentation with prominent mucositis and limited cutaneous involvement. The name *Mycoplasma*-induced rash and mucositis has been proposed to help differentiate this clinical entity, which some believe may be the syndrome originally described by Stevens and Johnson.

Patients with SJS/TEN initially present with fever >39°C (102.2°F); sore throat; conjunctivitis; and acute onset of painful dusky, atypical, target-like lesions **(Fig. 56-11)**. Intestinal and upper respiratory tract involvement are associated with a poor prognosis, as are older age and greater extent of epidermal detachment. At least 10% of those with SJS and 30% of those with TEN die from the disease. Drugs that most commonly cause SJS/TEN are sulfonamides, allopurinol, antiepileptics (e.g., lamotrigine, phenytoin, carbamazepine), oxicam NSAIDs, β-lactam and other antibiotics, and nevirapine. Frozen-section skin biopsy may aid in rapid diagnosis. At this time, there is no consensus on the most effective treatment for SJS/TEN. The best outcomes stem from early diagnosis, immediate discontinuation of the suspected drug, and meticulous

supportive therapy in an intensive care or burn unit. Issues such as fluid management, atraumatic wound care, infection prevention and treatment, and ophthalmologic and respiratory support are critical. Systemic glucocorticoid therapy (prednisone 1–2 mg/kg) may be useful early in disease evolution; however, long-term or late systemic glucocorticoid use has been associated with increased mortality. After initial enthusiasm for the use of intravenous immunoglobulin (IVIG) in the treatment of SJS/TEN, more recent data question whether it is beneficial. There are emerging data to support treatment with cyclosporine and etanercept. Randomized studies to evaluate potential therapies are lacking and difficult to perform.

FIGURE 56-11

Target-like lesion in SJS.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Pustular Eruptions

AGEP is a rare reaction pattern affecting 3–5 people per million per year. It is thought to be secondary to medication exposure in >90% of cases (**Fig. 56-12**). Patients typically present with diffuse erythema or erythroderma, as well as high spiking fevers, and leukocytosis. One to two days later, innumerable pinpoint pustules develop overlying the erythema. The pustules are most pronounced in body fold areas; however, they may become generalized and, when coalescent, can lead to superficial erosion. In such cases, differentiating the eruption from SJS in its initial stages may be difficult; in AGEP, any erosions tend to be more superficial, and prominent mucosal involvement is lacking. Skin biopsy shows collections of neutrophils and sparse necrotic keratinocytes in the upper part of the epidermis, unlike the full-thickness

epidermal necrosis that characterizes SJS. Before the pustules appear, AGEP may also mimic DIHS due to the prominent fever and erythroderma.

FIGURE 56-12

Acute generalized exanthematous pustulosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

The principal differential diagnosis for AGEP is acute pustular psoriasis, which has an identical clinical and histologic appearance. Many patients with AGEP have a personal or family history of psoriasis. AGEP classically begins within 24–48 hours of drug exposure, though it may occur as much as 1–2 weeks later. β-Lactam antibiotics, calcium channel blockers, macrolide antibiotics, and other inciting agents (including radiocontrast and dialysates) have been reported. Patch testing with the responsible drug often results in a localized pustular eruption.

Overlap Hypersensitivity Syndromes

An important concept in the clinical approach to severe drug eruptions is the presence of overlap syndromes, most notably DIHS with TEN-like features, DIHS with pustular eruption (AGEP-like), and AGEP with TEN-like features. In several case series of AGEP, 50% of cases had TEN-like or DRESS-like features, and 20% of cases had mucosal involvement resembling SJS/TEN. In one study, up to 20% of all severe drug eruptions had overlap features, suggesting that AGEP, DIHS, and SJS/TEN represent a clinical spectrum with some common pathophysiologic mechanisms. Designation of a single diagnosis based on cutaneous and extracutaneous involvement may not always be possible in cases of hypersensitivity; in such instances, treatment should be geared toward addressing the dominant clinical features. The timing of rash onset with respect to drug

administration, which is usually much more delayed in DIHS, and the presence of systemic manifestations such as hepatitis are helpful clues to that diagnosis.

Vasculitis

Cutaneous small-vessel vasculitis (CSVV) typically presents with purpuric papules and macules involving the lower extremities and other dependent areas (**Fig. 56-13**) (**Chap. 356**). Pustular and hemorrhagic vesicles as well as rounded ulcers also occur. Importantly, vasculitis may involve other organs, including the kidneys, joints, gastrointestinal tract, and lungs, necessitating a thorough clinical evaluation for systemic involvement. Drugs are implicated as a cause of roughly 15% of all cases of small vessel vasculitis. Antibiotics, particularly β-lactams, are commonly implicated; however, almost any drug can cause vasculitis. Vasculitis may also be idiopathic or due to underlying infection, connective tissue disease, or (rarely) malignancy.

FIGURE 56-13

Cutaneous small-vessel vasculitis (CSVV, leukocytoclastic vasculitis).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Rare but important types of drug-induced vasculitis include drug-induced ANCA vasculitis. Such patients commonly present with cutaneous manifestations but can develop the full range of symptoms associated with ANCA vasculitis, including crescentic glomerulonephritis and alveolar hemorrhage. Propylthiouracil, methimazole, and hydralazine are common culprits. Drug-induced polyarteritis nodosa has been associated with long-term exposure to minocycline. The presence of perivascular eosinophils on skin biopsy can be a clue to possible drug etiology.

MANAGEMENT OF THE PATIENT WITH SUSPECTED DRUG ERUPTION

There are four main questions to answer regarding a suspected drug eruption:

1. Is the observed rash caused by a medication?
2. Is the reaction severe or evolving?
3. Which drug or drugs are suspected, and should they be withdrawn?
4. What recommendation can be made for future medication use?

EARLY DIAGNOSIS OF SEVERE ERUPTIONS

Rapid recognition of potentially serious or life-threatening reactions is paramount. In this regard, a suspected drug eruption is best defined initially by what it is not (e.g., SJS/TEN, DIHS). **Table 56-2** lists clinical and laboratory features that, if present, suggest the presence of a severe reaction. **Table 56-3** lists the most important of these reactions, along with their key features and commonly associated medications. Any concern for a serious reaction should prompt immediate consultation with a dermatologist and/or referral of the patient to a specialized center.

TABLE 56-2

Clinical and Laboratory Findings Suggestive of Severe Cutaneous Adverse Drug Reaction

Cutaneous
Generalized erythema Facial edema Skin pain Palpable purpura Dusky or target-like lesions Skin necrosis Blisters or epidermal detachment Positive Nikolsky sign Mucous membrane erosions Swelling of lips or tongue
General
High fever Enlarged lymph nodes Arthralgias or arthritis Shortness of breath, hoarseness, wheezing, hypotension
Laboratory Results
Eosinophil count >1000/μL Lymphocytosis with atypical lymphocytes Abnormal liver or kidney function tests

Source: Adapted from JC Roujeau, RS Stern: Severe adverse cutaneous reactions to drugs. N Engl J Med 331:1272, 1994.

TABLE 56-3

Clinical Features of Severe Cutaneous Drug Reactions

DIAGNOSIS	MUCOSAL LESIONS	TYPICAL SKIN LESIONS	FREQUENT SIGNS AND SYMPTOMS	MOST COMMON CULPRIT DRUGS
Stevens-Johnson syndrome (SJS)	Erosions usually at two or more sites	Small blisters form from dusky macules or atypical targets; rare areas of confluence; detachment ≤10% body surface area	Most cases involve fever	Sulfonamides, anticonvulsants, allopurinol, nonsteroidal anti-inflammatory drugs (NSAIDs)
Toxic epidermal necrolysis (TEN)[a]	Erosions usually at two or more sites	Individual lesions like those seen in SJS; confluent dusky erythema; large sheets of necrotic epidermis; total detachment of >30% body surface area	Nearly all cases involve fever, “acute skin failure,” leukopenia	Same as for SJS
Drug-induced hypersensitivity syndrome/drug rash with eosinophilia and systemic symptoms (DIHS/DRESS)	Mucositis reported in as many as 30%	Diffuse, deep red morbilliform eruption with facial involvement; facial and acral swelling	Fever, lymphadenopathy, hepatitis, nephritis, myocarditis, eosinophilia, atypical lymphocytosis	Anticonvulsants, sulfonamides, allopurinol, minocycline
Acute generalized exanthematous pustulosis (AGEP)	Oral erosions in perhaps 20%	Innumerable pinpoint pustules overlying a diffuse erythematous eruption; may develop superficial erosions	High fever, leukocytosis (neutrophilia), hypocalcemia	β-Lactam antibiotics, calcium channel blockers, macrolide antibiotics

DIAGNOSIS	MUCOSAL LESIONS	TYPICAL SKIN LESIONS	FREQUENT SIGNS AND SYMPTOMS	MOST COMMON CULPRIT DRUGS
Serum sickness or serum sickness-like reaction	Absent	Urticarial serpiginous or polycyclic rash; purpuric eruption along the sides of the feet and hands is characteristic	Fever, arthralgias	Antithymocyte globulin, cephalosporins, monoclonal antibodies
Anticoagulant-induced necrosis	Infrequent	Purpura and necrosis, especially of central, fatty areas	Pain in affected areas	Warfarin, heparin
Angioedema	Often involved	Urticaria or swelling of the central face, other areas	Respiratory distress, cardiovascular collapse	Angiotensin-converting enzyme (ACE) inhibitors, NSAIDs, contrast dye

[a]Overlap of SJS and TEN have features of both, and attachment of 10–30% of body surface area may occur.

Source: Adapted from JC Roujeau, RS Stern: Severe adverse cutaneous reactions to drugs. N Engl J Med 331:1272, 1994.

CONFIRMATION OF DRUG REACTION

The probability of drug etiology varies with the pattern of the reaction. Only fixed drug eruptions are always drug-induced. Morbilliform eruptions are usually viral in children and drug-induced in adults. Among severe reactions, drugs account for 10–20% of anaphylaxis and vasculitis and between 70% and 90% of AGEP, DIHS, SJS, and TEN. Skin biopsy helps characterize the reaction but does not indicate drug causality. Blood counts and liver and renal function tests are important for evaluating organ involvement. The association of mild elevation of liver enzymes and high eosinophil count is frequent but not specific for a drug reaction. Blood tests that could identify an alternative cause, serologic tests (to rule out drug-induced lupus), and serology or polymerase chain reaction for infections may be of great importance to determine a cause.

WHAT DRUG(S) TO SUSPECT AND WITHDRAW

Most cases of drug eruptions occur during the first course of treatment with a new medication. A notable exception is IgE-mediated urticaria and anaphylaxis that need presensitization and develop a few minutes to a few hours after rechallenge. Characteristic timing of onset following drug administration is as follows: 4–14

days for morbilliform eruption, 2–4 days for AGEP, 5–28 days for SJS/TEN, and 14–48 days for DIHS. A drug chart, compiling information of all current and past medications/supplements and the timing of administration relative to the rash, is a key diagnostic tool for identifying the inciting drug. Medications introduced for the first time in the relevant time frame are prime suspects. Two other important elements to suspect causality at this stage are (1) previous experience with the drug in the population and (2) alternative etiologic candidates.

The decision to continue or discontinue any medication depends on the severity of the reaction, the severity of the primary disease undergoing treatment, the degree of suspicion of causality, and the feasibility of finding an alternative safer treatment. In any potentially fatal drug reaction, elimination of all possible suspect drugs or unnecessary medications should be immediately attempted. Some rashes may resolve when “treating through” a benign drug-related eruption. The decision to treat through an eruption should, however, remain the exception and withdrawal of every suspect drug the general rule. On the other hand, drugs that are not suspected and are important for the patient (e.g., antihypertensive agents) generally should not be quickly withdrawn. This approach may permit judicious use of these agents in the future.

RECOMMENDATION FOR FUTURE USE OF DRUGS

The aims are to (1) prevent the recurrence of the drug eruption and (2) avoid compromising future treatment by inaccurately excluding otherwise useful medications.

A thorough assessment of drug causality is based on timing of the reaction, evaluation of other possible causes, and effect of drug withdrawal or continuation. The RegiSCAR group has proposed the Algorithm of Drug Causality for Epidermal Necrolysis (ALDEN) to rank likelihood of drug causality in SJS/TEN; validation of this and other instruments, such as the Naranjo adverse drug reaction probability scale, is limited. Medication(s) with a “definite” or “probable” causality should be contraindicated, a warning card or medical alert tag (e.g., wristband) should be given to the patient, and the drugs should be listed in the patient’s medical chart as allergies.

CROSS-SENSITIVITY

Because of possible cross-sensitivity among chemically related drugs, many physicians recommend avoidance of not only the medication that induced the reaction but also all drugs of the same pharmacologic class.

There are two types of cross-sensitivity. Reactions that depend on a pharmacologic interaction may occur with all drugs that target the same pathway, whether the drugs are structurally similar or not. This is the case with angioedema caused by NSAIDs and ACE inhibitors. In this situation, the risk of recurrence varies from drug to drug in a particular class; however, avoidance of all drugs in the class is usually recommended. Immune recognition of structurally related drugs is the second mechanism by which cross-sensitivity occurs. A classic example is hypersensitivity to aromatic antiepileptics (barbiturates, phenytoin, carbamazepine) with up to 50% reaction to a second drug in patients who reacted to one. For other drugs, in vitro and in vivo

data have suggested that cross-reactivity exists only between compounds with very similar chemical structures. Sulfamethoxazole-specific lymphocytes may be activated by other antibacterial sulfonamides but not diuretics, antidiabetic drugs, or anti-COX2 NSAIDs with a sulfonamide group. Approximately 10% of patients with penicillin allergies will also develop allergic reactions to cephalosporin class antibiotics.

Recent data suggest that although the risk of developing a drug eruption to another drug is increased in persons with a prior reaction, “cross-sensitivity” is probably not the explanation. As an example, those with a history of an allergic-like reaction to penicillin are at greater risk of developing a reaction to antibacterial sulfonamides than to cephalosporins.

These data suggest that the list of drugs to avoid after a drug reaction should be limited to the causative one(s) and to a few very similar medications.

Because of growing evidence that some severe cutaneous reactions to drugs are associated with HLA genes, it is recommended that first-degree family members of patients with severe cutaneous reactions also should avoid causative agents. This may be most relevant for sulfonamides and antiepileptic medications.

ROLE OF TESTING FOR CAUSALITY AND DRUG RECHALLENGE

The usefulness of laboratory tests, skin-prick, or patch testing to determine causality is debated. Many in vitro immunologic assays have been developed for research purposes; however, the predictive value of these tests has not been validated in large series of affected patients. In some cases, diagnostic rechallenge may be appropriate, even for drugs with high rates of adverse reactions.

Skin-prick testing has clinical value in limited settings. In patients with a history suggesting immediate IgE-mediated reactions to penicillin, skin-prick testing with penicillins or cephalosporins has proven useful for identifying patients at risk of anaphylactic reactions to these agents. Negative skin tests do not totally rule out IgE-mediated reactivity; however, the risk of anaphylaxis in response to penicillin administration in patients with negative skin tests is about 1%. In contrast, two-thirds of patients with a positive skin test experience an allergic response upon rechallenge. The skin tests themselves carry a small risk of anaphylaxis.

For patients with delayed-type hypersensitivity, the clinical utility of skin tests remains questionable. At least one of a combination of several tests (prick, patch, and intradermal) is positive in 50–70% of patients with a reaction “definitely” attributed to a single medication. This low sensitivity corresponds to the observation that readministration of drugs with negative skin testing results in eruptions in 17% of cases.

Desensitization can be considered in those with a history of reaction to a medication that must be used again. Efficacy of such procedures has been demonstrated in cases of immediate reaction to penicillin and positive skin tests, anaphylactic reactions to platinum chemotherapy, and delayed reactions to sulfonamides in patients with AIDS. Desensitization is often successful in HIV-infected patients with morbilliform eruptions to sulfonamides but is not recommended in HIV-infected patients who developed erythroderma or a bullous reaction in response to prior sulfonamide exposure. Various protocols are available, including oral and

parenteral approaches. Oral desensitization appears to have a lower risk of serious anaphylactic reaction. Desensitization carries the risk of anaphylaxis regardless of how it is performed and should be performed in monitored clinical settings such as an intensive care unit. After desensitization, many patients experience non-life-threatening reactions during therapy with the culprit drug.

REPORTING

Any severe reaction to drugs should be reported to a regulatory agency or to pharmaceutical companies. Because severe reactions are too rare to be detected in premarketing clinical trials, spontaneous reports are of critical importance for early detection of unexpected life-threatening events. To be useful, the report should contain enough details to permit ascertainment of severity and drug causality.

ACKNOWLEDGMENT

We acknowledge the contribution of Drs. Jean-Claude Roujeau and Robert S. Stern to this chapter in previous editions.

FURTHER READING

Belum VR: Characterisation and management of dermatologic adverse events to agents targeting the PD-1 receptor. Eur J Cancer 60:12, 2016. [PubMed: 27043866]

Cornejo-Garcia JA et al: The genetics of drug hypersensitivity reactions. J Investig Allergol Clin Immunol 26:222, 2016. [PubMed: 27375032]

Creamer D et al: U.K. guidelines for the management of Stevens-Johnson syndrome/toxic epidermal necrolysis in adults 2016. Br J Dermatol 174:1194, 2016. [PubMed: 27317286]

Harp JL et al: Severe cutaneous adverse reactions: impact of immunology, genetics, and pharmacology. Semin Cutan Med Surg 33:17, 2014. [PubMed: 25037255]

Ko TM et al: Use of HLA-B*5801 genotyping to prevent allopurinol induced severe cutaneous adverse reactions in Taiwan: National prospective cohort study. BMJ 351:h4848, 2015. [PubMed: 26399967]

Lacouture ME et al: Ipilimumab in patients with cancer and the management of dermatologic adverse events. J Am Acad Dermatol 71:161, 2014. [PubMed: 24767731]

Mayorga C et al: In vitro tests for drug hypersensitivity reactions: An ENDA/EAACI Drug Allergy Interest Group position paper. Allergy 71:1103, 2016. [PubMed: 26991315]

Oussalah A et al: Genetic variants associated with drug-induced immediate hypersensitivity reactions: A PRISMA-compliant systematic review. Allergy 71:443, 2016. [PubMed: 26678823]

Petrelli F et al: Antibiotic prophylaxis for skin toxicity induced by antiepidermal growth factor receptor agents: A systematic review and meta-analysis. Br J Dermatol, 2016 ePub ahead of print. Accessed September 28, 2016.

Sassolas B et al: ALDEN, an algorithm for assessment of drug causality in Stevens-Johnson syndrome and toxic epidermal necrolysis: Comparison with case-control analysis. Clin Pharmacol Ther 88:60, 2010. [PubMed: 20375998]

White KD et al: Evolving models of the immunopathogenesis of T cell-mediated drug allergy: The role of host, pathogens, and drug response. J Allergy Clin Immunol 136:219, 2015. [PubMed: 26254049]

Wolverton SE: Practice gaps: Drug reactions. Dermatol Clin 34:311, 2016. [PubMed: 27363888]

Chapter 57: Photosensitivity and Other Reactions to Light

Alexander G. Marneros; David R. Bickers

FIGURE 57-1

SOLAR RADIATION

Sunlight is the most visible and obvious source of comfort in the environment. The sun provides the beneficial effects of warmth and vitamin D synthesis. However, acute and chronic sun exposure also has pathologic consequences. Cutaneous exposure to sunlight is a major cause of human skin cancer and can have immunosuppressive effects as well.

The sun's energy reaching the earth's surface is limited to components of the ultraviolet (UV) spectrum, the visible spectrum, and portions of the infrared spectrum. The cutoff at the short end of the UV spectrum at ~290 nm is due primarily to stratospheric ozone—formed by highly energetic ionizing radiation—that prevents penetration to the earth's surface of the shorter, more energetic, potentially more harmful wavelengths of solar radiation. Indeed, concern about destruction of the ozone layer by chlorofluorocarbons released into the atmosphere has led to international agreements to reduce production of those chemicals.

Measurements of solar flux showed a 20-fold regional variation in the amount of energy at 300 nm that reaches the earth's surface. This variability relates to seasonal effects, the path that sunlight traverses through ozone and air, the altitude (a 4% increase for each 300 m of elevation), the latitude (increasing intensity with decreasing latitude), and the amount of cloud cover, fog, and pollution.

The major components of the photobiologic action spectrum that are capable of affecting human skin include the UV and visible wavelengths between 290 and 700 nm. In addition, the wavelengths beyond 700 nm in the infrared spectrum primarily emit heat and in certain circumstances may exacerbate the pathologic effects of energy in the UV and visible spectra.

The UV spectrum reaching the Earth represents <10% of total incident solar energy and is arbitrarily divided into two major segments, UV-B and UV-A, which constitute the wavelengths from 290 to 400 nm. UV-B consists of wavelengths between 290 and 320 nm. This portion of the photobiologic action spectrum is the most efficient in producing redness or erythema in human skin and thus is sometimes known as the "sunburn spectrum." UV-A includes wavelengths between 320 and 400 nm and is ~1000-fold less efficient in producing skin redness than is UV-B.

The wavelengths between 400 and 700 nm are visible to the human eye. The photon energy in the visible spectrum is not capable of damaging human skin in the absence of a photosensitizing chemical. Without the absorption of energy by a molecule, there can be no photosensitivity. Thus, the *absorption spectrum* of a molecule is defined as the range of wavelengths it absorbs, whereas the *action spectrum* for an effect of incident radiation is defined as the range of wavelengths that evoke the response.

Photosensitivity occurs when a photon-absorbing chemical (*chromophore*) present in the skin absorbs incident energy, becomes excited, and transfers the absorbed energy to various structures or to molecular oxygen.

UV RADIATION (UVR) AND SKIN STRUCTURE AND FUNCTION

Human skin consists of two major compartments: the outer epidermis, which is a stratified squamous epithelium, and the underlying dermis, which is rich in matrix proteins such as collagens and elastin. Both compartments are susceptible to damage from sun exposure. The epidermis and the dermis contain several chromophores capable of absorbing incident solar energy, including nucleic acids, proteins, and lipids. The outermost epidermal layer, the stratum corneum, is a major absorber of UV-B, and <10% of incident UV-B wavelengths penetrate through the epidermis to the dermis. Approximately 3% of radiation below 300 nm, 20% of radiation below 360 nm, and 33% of short visible radiation reach the basal cell layer in untanned human skin. UV-A readily penetrates to the dermis and is capable of altering structural and matrix proteins that contribute to photoaging of chronically sun-exposed skin, particularly in individuals of light complexion. Thus, longer wavelengths can penetrate more deeply into the skin.

Molecular Targets for UVR-Induced Skin Effects

Epidermal DNA—predominantly in keratinocytes and in Langerhans cells, which are dendritic antigen-presenting cells—absorbs UV-B and undergoes structural changes between adjacent pyrimidine bases (thymine or cytosine), including the formation of cyclobutane dimers and 6,4-photoproducts. These structural changes are potentially mutagenic and are found in most basal cell and squamous cell carcinomas (BCCs and SCCs, respectively). They can be repaired by cellular mechanisms that result in their recognition and excision and the restoration of normal base sequences. The efficient repair of these structural aberrations is crucial, since individuals with defective DNA repair are at high risk for the development of cutaneous cancer. For example, patients with xeroderma pigmentosum, an autosomal recessive disorder, have a variably deficient repair of UV-induced photoproducts. The skin of these patients often shows the dry, leathery appearance of prematurely photoaged skin, and these patients have an increased frequency of skin cancer already in the first two decades of life. Studies in transgenic mice have verified the importance of functional genes that regulate these repair pathways in preventing the development of UV-induced skin cancer. DNA damage to Langerhans cells may also contribute to the known immunosuppressive effects of UV-B (see "Photoimmunology," later).

In addition to DNA, molecular oxygen is a target for incident solar UVR, leading to the generation of reactive oxygen species (ROS). These ROS can damage skin components through oxidative damage to DNA, oxidation

of polyunsaturated fatty acids in lipids (lipid peroxidation), oxidation of amino acids in proteins, or they can lead to oxidative deactivation of specific enzymes. UVR can also promote increased cross-linking and degradation of dermal matrix proteins and accumulation of abnormal dermal elastin leading to photoaging changes known as *solar elastosis*.

Cutaneous Optics and Chromophores

Chromophores are endogenous or exogenous chemical components that can absorb physical energy. Endogenous chromophores are of two types: (1) normal components of skin, including nucleic acids, proteins, lipids, and 7-dehydrocholesterol (the precursor of vitamin D); and (2) components that are synthesized elsewhere in the body and that circulate in the bloodstream and diffuse into the skin, such as porphyrins. Normally, only trace amounts of porphyrins are present in the skin, but, in selected diseases known as the *porphyrias* **(Chap. 409)**, porphyrins are released into the circulation in increased amounts from the bone marrow and the liver and are transported to the skin, where they absorb incident energy both in the Soret band (~400 nm; short visible) and, to a lesser extent, in the red portion of the visible spectrum (580–660 nm). This energy absorption results in the generation of ROS that can mediate structural damage to the skin, manifested as erythema, edema, urticaria, or blister formation. It is of interest that photoexcited porphyrins are currently used in the treatment of BCCs and SCCs and their precursor lesions, actinic keratoses. Known as *photodynamic therapy* (PDT), this modality generates ROS in the skin, leading to cell death. Topical photosensitizers used in PDT are the porphyrin precursors 5-aminolevulinic acid and methyl aminolevulinate, which are converted to porphyrins in the skin. It is believed that PDT targets tumor cells for destruction more selectively than it targets adjacent nonneoplastic cells. The efficacy of such therapy requires appropriate timing of the application of methyl aminolevulinate or 5-aminolevulinic acid to the affected skin followed by exposure to artificial sources of visible light. High-intensity blue light has been used successfully for the treatment of thin actinic keratoses. Red light has a longer wavelength, penetrates more deeply into the skin, and is more beneficial in the treatment of superficial BCCs.

Acute Effects of Sun Exposure

The acute effects of skin exposure to sunlight include sunburn and vitamin D synthesis.

SUNBURN

This painful skin condition is an acute inflammatory response of the skin, predominantly to UV-B. Generally, an individual's ability to tolerate sunlight is inversely proportional to that individual's degree of melanin pigmentation. Melanin, a complex polymer of tyrosine derivatives, is synthesized in specialized epidermal dendritic cells known as *melanocytes* and is packaged into *melanosomes* that are transferred via dendritic processes into *keratinocytes*, thereby providing photoprotection (dissipating the vast majority of absorbed UVR in the skin) and simultaneously darkening the skin. Sun-induced melanogenesis is a consequence of increased tyrosinase activity in melanocytes. Central to the suntan response is the melanocortin-1 receptor (*MC1R*), and mutations in this gene contribute to the wide variation in human skin and hair color; individuals with red hair and fair skin typically have low MC1R activity. In the skin there are two main types of melanin: eumelanin (providing brown and black pigmentation associated with high MC1R activity) and pheomelanin

(providing red pigmentation associated with low MC1R activity). Pheomelanin is a cysteine-containing red polymer of benzothiazine units and has much weaker shielding capacity against UVR compared to eumelanin. This may explain why individuals with a higher proportion of pheomelanin (red hair/fair skin appearance) have an increased risk of melanoma formation. In addition, pheomelanin may also promote melanoma formation through induction of oxidative damage by amplifying UV-A-induced ROS but also through UVR-independent mechanisms.

Genetic studies have revealed additional genes that influence skin color variation in humans, such as the gene for tyrosinase (*TYR*) and the genes *APBA2[OCA2]*, *SLC45A2*, and *SLC24A5*. The human *MC1R* gene encodes a G protein–coupled receptor that binds α-melanocyte-stimulating hormone (α-MSH), which is secreted in the skin mainly by keratinocytes in response to UVR. The UV-induced expression of this hormone is controlled by the tumor suppressor p53, and absence of functional p53 attenuates the tanning response. Activation of the melanocortin receptor leads to increased intracellular cyclic adenosine 5′-monophosphate (cAMP) and protein kinase A activation, resulting in an increased transcription of the microphthalmia-associated transcription factor (MITF), which stimulates melanogenesis. Since the precursor of α-MSH, proopiomelanocortin produced by keratinocytes, is also the precursor of β-endorphins, UVR may result in not only increased pigmentation but also in increased β-endorphin production in the skin, an effect that has been hypothesized to promote sun-seeking behaviors and even mediate addiction to tanning.

The Fitzpatrick classification of human skin phototypes is based on the efficiency of the epidermal-melanin unit, which usually can be ascertained by asking an individual two questions: (1) Do you burn after sun exposure? (2) Do you tan after sun exposure? The answers to these questions permit division of the population into six skin types, varying from type I (always burn, never tan) to type VI (never burn, always tan) **(Table 57-1)**.

TABLE 57-1

Skin Type and Sunburn Sensitivity (Fitzpatrick Classification)

Type	Description
I	Always burn, never tan
II	Always burn, sometimes tan
III	Sometimes burn, sometimes tan
IV	Sometimes burn, always tan
V	Never burn, sometimes tan
VI	Never burn, always tan

Sunburn erythema is due to vasodilation of dermal blood vessels. There is a lag time (usually 4–12 h) between skin exposure to sunlight and the development of visible redness. The action spectrum for sunburn erythema includes UV-B and UV-A, although UV-B is much more efficient than UV-A in evoking the response. However, UV-A may contribute to sunburn erythema at midday, when much more UV-A than UV-B is present in the solar spectrum. The erythema that accompanies the inflammatory response induced by UVR results from the orchestrated release of cytokines along with growth factors and the generation of ROS. Furthermore, UV-induced activation of nuclear factor κB–dependent gene transcription can augment release of several proinflammatory cytokines and vasoactive mediators. These cytokines and mediators accumulate locally in sunburned skin, providing chemotactic factors that attract neutrophils, macrophages, and T lymphocytes, which promote the inflammatory response. UVR also stimulates infiltration of inflammatory cells through induced expression of adhesion molecules such as E-selectin and intercellular adhesion molecule 1 on endothelial cells and keratinocytes. UVR also has been shown to activate phospholipase A_2, resulting in increases in eicosanoids such as prostaglandin E_2, which is known to be a potent inducer of sunburn erythema. The role of eicosanoids in this reaction has been verified by studies showing that nonsteroidal anti-inflammatory drugs (NSAIDs) can reduce sunburn erythema.

Epidermal changes in sunburn include the induction of “sunburn cells,” which are keratinocytes undergoing p53-dependent apoptosis as a defense, with elimination of cells that harbor UV-B-induced structural DNA damage.

VITAMIN D SYNTHESIS AND PHOTOCHEMISTRY

Cutaneous exposure to UV-B causes photolysis of epidermal 7-dehydrocholesterol, converting it to pre–vitamin D_3, which then undergoes temperature-dependent isomerization to form the stable hormone

vitamin D_3. This compound diffuses to the dermal vasculature and circulates to the liver and kidney, where it is converted to the dihydroxylated functional hormone 1,25-dihydroxyvitamin D_3. Vitamin D metabolites from the circulation and those produced in the skin itself can augment epidermal differentiation signaling and inhibit keratinocyte proliferation. These effects are exploited therapeutically in psoriasis with the topical application of synthetic vitamin D analogues. In addition, vitamin D is increasingly thought to have beneficial effects in several other inflammatory conditions, and some evidence suggests that—besides its classic physiologic effects on calcium metabolism and bone homeostasis—it is associated with a reduced risk of various internal malignancies. There is controversy regarding the risk-to-benefit ratio of sun exposure for vitamin D homeostasis. At present, it is important to emphasize that no clear-cut evidence suggests that the use of sunscreens substantially diminishes vitamin D levels. Since aging also substantially decreases the ability of human skin to photocatalytically produce vitamin D_3, the widespread use of sunscreens that filter out UV-B has led to concerns that the elderly might be unduly susceptible to vitamin D deficiency. However, the amount of sunlight needed to produce sufficient vitamin D is small and does not justify the risks of skin cancer and other types of photodamage linked to increased sun exposure or tanning behavior. Nutritional supplementation of vitamin D is a preferable strategy for patients with vitamin D deficiency.

Chronic Effects of Sun Exposure: Nonmalignant

The clinical features of photoaging (*dermatoheliosis*) consist of wrinkling, blotchiness, and telangiectasia, as well as a roughened, irregular, “weather-beaten” leathery appearance.

UVR is important in the pathogenesis of photoaging in human skin, and ROS are likely involved. The dermis and its connective tissue matrix are major targets for sun-associated chronic damage that manifests as solar elastosis, a massive increase in thickened irregular masses of abnormal-appearing elastic fibers. Collagen fibers are also abnormally clumped in the deeper dermis of sun-damaged skin. The chromophores, the action spectra, and the specific biochemical events orchestrating these changes are only partially understood, although more deeply penetrating UV-A seems to be primarily involved. Chronologically aged sun-protected skin and photoaged skin share important molecular features, including connective tissue damage and elevated levels of matrix metalloproteinases (MMPs). MMPs are enzymes involved in the degradation of the extracellular matrix. UV-A induces expression of some MMPs, including MMP-1 and MMP-3, leading to increased collagen breakdown. In addition, UV-A reduces type I procollagen messenger RNA (mRNA) expression. Thus, chronic UVR alters the structure and function of dermal collagen both by inhibiting its synthesis and enhancing its breakdown. On the basis of these observations, it is not surprising that high-dose UV-A phototherapy may have beneficial effects in some patients with localized fibrotic diseases of the skin, such as localized scleroderma.

Chronic Effects of Sun Exposure: Malignant

One of the major known consequences of chronic excessive skin exposure to sunlight is nonmelanoma skin cancer (NMSC). The two most common types of NMSC are BCC and SCC **(Chap. 72)**. A model for skin cancer induction involves three major steps: initiation, promotion, and progression. Exposure of human skin to

sunlight results in *initiation*, a step by which structural (mutagenic) changes in DNA evoke an irreversible alteration in the target cell (keratinocyte) that begins the tumorigenic process. Exposure to a tumor initiator such as UV-B is believed to be a necessary but not a sufficient step in the malignant process, since initiated skin cells not exposed to tumor promoters generally do not develop into tumors. The second stage in tumor development is *promotion*, a multistep process by which chronic exposure to sunlight evokes further changes that culminate in the clonal expansion of initiated cells and cause the development of premalignant growths known as *actinic keratoses*, which may progress to form SCCs. As a result of extensive studies, it seems clear that UV-B is a *complete carcinogen*, meaning that it can act as both a tumor initiator and a tumor promoter. The third and final step in the malignant process is *malignant conversion* of benign precursors into malignant lesions, a process thought to require additional genetic alterations.

On a molecular level, skin carcinogenesis results from the accumulation of gene mutations that cause inactivation of tumor suppressors, activation of oncogenes, or reactivation of cellular signaling pathways that normally are expressed only during embryologic epidermal development. Interestingly, a large number of UV-induced oncogenic driver mutations that are present in SCCs can already be found in aged sun-exposed normal skin, leading to a growth advantage and innumerable precancerous clones carrying cancer-causing mutations. These mutations occur particularly often in genes that affect proliferation of epidermal stem cells (e.g., NOTCH receptor genes). The pattern of oncogenic gene mutations in aged sun-exposed skin shows considerable overlap with the mutations identified in SCCs, while there is little overlap with the mutations identified in BCCs or melanomas. For example, ~20% of normal aged sun-exposed skin cells and ~60% of SCCs carry driver mutations in *NOTCH1*. Additionally, the accumulation of mutations in the tumor-suppressor gene *p53* can also promote skin carcinogenesis. Indeed, the majority of both human and murine UV-induced skin cancers have characteristic UVR-induced *p53* mutations (C → T and CC → TT transitions). Studies in mice have shown that sunscreens can substantially reduce the frequency of these signature mutations in *p53* and inhibit the induction of tumors. The comparison of UVR-induced gene mutations between aged sun-exposed normal skin and SCCs supports the hypothesis of a progressive accumulation of additional oncogenic mutations that eventually lead to the transition from precancerous cell clones to SCCs. It has been estimated that SCCs harbor ~10 times more oncogenic driver mutations per cell than cells in aged sun-exposed normal skin. Furthermore, while aged sun-exposed skin and SCCs carry similar UVR-induced mutations in *p53* or NOTCH receptors, oncogenic mutations in other genes (e.g., *CDKN2A*) were mainly found in SCCs and not in aged sun-exposed skin, which are thus likely to play a critical role in malignant progression.

Compared to SCCs, BCCs carry a distinct mutational profile in specific genes that are critical for their formation. BCCs harbor inactivating mutations particularly in the tumor-suppressor gene *patched* or activating mutations in the oncogene *smoothened*, which results in the constitutive activation of the sonic hedgehog signaling pathway and increased cell proliferation. New evidence links alterations in the Wnt/β-catenin signaling pathway, which is known to be critical for hair follicle development, to skin cancer as well. Thus, interactions between this pathway and the hedgehog signaling pathway appear to be involved in both skin carcinogenesis and embryologic development of the skin and hair follicles.

Clonal analysis in mouse models of BCC revealed that tumor cells arise from stem cells of the interfollicular epidermis and the upper infundibulum of the hair follicle. These BCC-initiating cells are reprogrammed to resemble embryonic hair follicle progenitors, whose tumor-initiating ability depends on activation of the Wnt/β-catenin signaling pathway.

SCC initiation occurs both in the interfollicular epidermis and in the hair follicle bulge stem cell populations. In mouse models, the combination of mutant K-Ras and p53 is sufficient to induce invasive SCCs from these cell populations.

The transcription factor Myc is important for stem cell maintenance in the skin, and oncogenic activation of Myc has been implicated in the development of BCCs and SCCs. Thus, NMSC involves mutations and alterations in multiple genes and pathways that occur as a result of their chronic accumulation driven by exposure to environmental factors such as solar UVR.

Epidemiologic studies have linked excessive sun exposure to an increased risk of NMSCs and melanoma of the skin; the evidence is far more direct for NMSCs (BCCs and SCCs) than for melanoma. Approximately 80% of NMSCs develop on sun-exposed body areas, including the face, neck, and hands. Major risk factors include male sex, childhood sun exposures, older age, fair skin, and residence at latitudes relatively close to the equator. Individuals with darker-pigmented skin have a lower risk of skin cancer than do fair-skinned individuals. More than 2 million individuals in the United States develop NMSC annually, and the lifetime risk that a fair-skinned individual will develop such a neoplasm is estimated at ~15%. The incidence of NMSC in the population is increasing at a rate of 2–3% per year.

The relationship of sun exposure to melanoma development is less direct, but strong evidence supports an association. Clear-cut risk factors include a positive family or personal history of melanoma and multiple dysplastic nevi. Melanomas can occur during adolescence; the implication is that the latent period for tumor growth is shorter than that for NMSC. For reasons that are only partially understood, melanomas are among the most rapidly increasing human malignancies (Chap. 72). One potential explanation is the widespread use of indoor tanning. It is estimated that 30 million people tan indoors in the United States annually, including >2 million adolescents. Furthermore, epidemiologic studies suggest that life in a sunny climate from birth or early childhood may increase the risk of melanoma development. In general, risk does not correlate with cumulative sun exposure but may be related to the duration and extent of exposure in childhood.

However, in contrast to NMSCs, melanoma frequently develops in non–sun-exposed skin, and oncogenic mutations in melanoma may also not be UVR-signature mutations. These observations suggest that UVR-independent factors may contribute to melanomagenesis, which is consistent with findings in mouse models showing that pheomelanin can promote melanoma formation through UVR-independent mechanisms.

Importantly, mutations in BRAF and NRAS that lead to activation of a growth-promoting signaling cascade are frequently found in melanoma (but not in SCCs or BCCs), which has led to the development of specific inhibitors of this pathway for the treatment of BRAF-mutant melanoma. However, a high mutational load in melanoma may not be equated with a more unfavorable prognosis. Tumor-specific missense mutations in

melanomas can result in neoantigens that facilitate an immune response to the tumor cell. A novel therapeutic approach for melanoma, termed immune checkpoint blockade, targets inhibitors of T cell activation (such as CTLA-4 or PD-1) that in a subset of patients has resulted in a durable and potent immune destruction of melanoma cells, resulting in prolonged survival of patients with metastatic melanoma. It has recently been shown that a high mutational load in melanomas correlated indeed with improved therapeutic outcome to immune checkpoint blockade, consistent with the hypothesis that acquired missense mutations in the tumor cells lead to neoantigens that increase the vulnerability of these melanoma cells to attack by activated T cells.

GLOBAL CONSIDERATIONS



The frequency of skin cancer shows strong geographic variation, depending on the skin phototype of the majority of the population in these geographic areas, but also depending on the intensity of UVR. For example, both melanoma and NMSCs are particularly common in Australia.

Photoimmunology

Exposure to solar radiation causes both local immunosuppression (inhibition of immune responses to antigens applied at the irradiated site) and systemic immunosuppression (inhibition of immune responses to antigens applied at remote, unirradiated sites). For example, human skin exposure to modest doses of UV-B can deplete the epidermal antigen-presenting cells known as Langerhans cells, thereby reducing the degree of allergic sensitization to application of the potent contact allergen dinitrochlorobenzene at the irradiated skin site.

An example of the systemic immunosuppressive effects of higher doses of UVR is the diminished immunologic response to antigens introduced either epicutaneously or intracutaneously at sites distant from the irradiated site. Various immunomodulatory factors and immune cells have been implicated in UVR-induced systemic immunosuppression, including tumor necrosis factor α, interleukin 4, interleukin 10, *cis*-urocanic acid, and eicosanoids. Experimental evidence suggests that prostaglandin E_2 signaling through prostaglandin E receptor subtype 4 mediates UVR-induced systemic immunosuppression by elevating the number of regulatory T cells, and this effect can be inhibited with NSAIDs.

The major chromophores in the upper epidermis that are known to initiate UV-mediated immunosuppression include DNA, *trans*-urocanic acid, and membrane components. The action spectrum for UV-induced immunosuppression closely mimics the absorption spectrum of DNA. Pyrimidine dimers in Langerhans cells may inhibit antigen presentation. The absorption spectrum of epidermal urocanic acid closely mimics the action spectrum for UV-B-induced immunosuppression. Urocanic acid is a metabolic product of the essential amino acid histidine and accumulates in the upper epidermis through breakdown of the histidine-rich protein filaggrin due to the absence of its catabolizing enzyme in keratinocytes. Urocanic acid is synthesized as a *trans*-isomer, and UV-induced *trans-cis* isomerization of urocanic acid in the stratum corneum drives immunosuppression. *Cis*-urocanic acid may exert its immunosuppressive effects through a variety of mechanisms, including inhibition of antigen presentation by Langerhans cells.

One important consequence of chronic sun exposure and associated immunosuppression is an enhanced risk of skin cancer. In part, UV-B activates regulatory T cells that suppress antitumor immune responses via interleukin 10 expression, whereas in the absence of high UV-B exposure, epidermal Langerhans cells present tumor-associated antigens and induce protective immunity, thereby inhibiting skin tumorigenesis. UV-induced DNA damage is a major molecular trigger of this immunosuppressive effect.

Perhaps the most graphic demonstration of the role of immunosuppression in enhancing the risk of NMSC comes from studies of organ transplant recipients who require lifelong immunosuppressive/antirejection drug regimens. More than 50% of organ transplant recipients develop BCCs and SCCs, and these cancers are the most common types of malignancies arising in these patients. Rates of BCC and SCC increase with the duration and degree of immunosuppression. These patients ideally should be screened prior to organ transplantation, be monitored closely thereafter, and adhere to rigorous photoprotection measures, including the use of sunscreens and protective clothing as well as sun avoidance. Notably, immunosuppressive drugs that target the mTOR pathway, such as sirolimus and everolimus, may reduce the risk of NMSC in organ transplant recipients compared to that associated with the use of calcineurin inhibitors (cyclosporine and tacrolimus). The latter may contribute to NMSC formation not only through their immunosuppressive effects but also through suppression of p53-dependent cancer cell senescence pathways independent of host immunity.

PHOTOSENSITIVITY DISEASES

The diagnosis of photosensitivity requires elicitation of a careful history to define the duration of signs and symptoms, the length of time between exposure to sunlight and the development of subjective symptoms, and visible changes in the skin. The age of onset can also be a helpful diagnostic clue. For example, the acute photosensitivity of erythropoietic protoporphyria (EPP) almost always begins in infancy or early childhood, whereas the chronic photosensitivity of porphyria cutanea tarda (PCT) typically begins in the fourth and fifth decades of life. A patient's history of exposure to topical and systemic drugs and chemicals may provide important diagnostic clues. Many classes of drugs can cause photosensitivity on the basis of either phototoxicity or photoallergy. Fragrances such as musk ambrette that were previously present in numerous cosmetic products are also potent photosensitizers.

Examination of the skin may offer important clues. Anatomic areas that are naturally protected from direct sunlight, such as the hairy scalp, the upper eyelids, the retroauricular areas, and the infranasal and submental regions, may be spared, whereas exposed areas show characteristic features of the pathologic process. These anatomic localization patterns are often helpful, but not infallible, in making the diagnosis. For example, airborne contact sensitizers that are blown onto the skin may produce dermatitis that can be difficult to distinguish from photosensitivity despite the fact that such material may trigger skin reactivity in areas shielded from direct sunlight.

Many dermatologic conditions may be caused or aggravated by sunlight **(Table 57-2)**. The role of light in evoking these responses may be dependent on genetic abnormalities ranging from well-described defects in

DNA repair that occur in xeroderma pigmentosum to the inherited abnormalities in heme synthesis that characterize the porphyrias.

TABLE 57-2

Classification of Photosensitivity Diseases

Type	Disease
Genetic	Erythropoietic porphyria
	Erythropoietic protoporphyria
	Porphyria cutanea tarda—familial
	Variegate porphyria
	Hepatoerythropoietic porphyria
	Albinism
	Xeroderma pigmentosum
	Rothmund-Thomson syndrome
	Bloom syndrome
	Cockayne syndrome
	Kindler syndrome
	Phenylketonuria
Metabolic	Porphyria cutanea tarda—sporadic
	Hartnup disease
	Kwashiorkor
	Pellagra
	Carcinoid syndrome
Phototoxic	
Internal	Drugs

Type	Disease
External	Drugs, plants, food
Photoallergic	
Immediate	Solar urticaria
Delayed	Drug photoallergy
	Persistent light reaction/chronic actinic dermatitis
Neoplastic and degenerative	Photoaging
	Actinic keratosis
	Melanoma and nonmelanoma skin cancer
Idiopathic	Polymorphous light eruption
	Hydroa aestivale
	Actinic prurigo
Photoaggravated	Lupus erythematosus
	Systemic
	Subacute cutaneous
	Discoid
	Dermatomyositis
	Herpes simplex
	Lichen planus actinicus
	Acne vulgaris (aestivale)

Polymorphous Light Eruption

The most common type of photosensitivity disease is *polymorphous light eruption* (PMLE). Many affected individuals never seek medical attention because the condition is often transient, becoming manifest in the spring with initial sun exposure but then subsiding spontaneously with continuing exposure, a phenomenon known as “hardening.” The major manifestations of PMLE include (often intensely) pruritic erythematous papules that may coalesce into plaques in a patchy distribution on exposed areas of the trunk and forearms. The face is usually less seriously involved. Whereas the morphologic skin findings remain similar for each patient with subsequent recurrences, significant interindividual variations in skin findings are characteristic (hence the term “polymorphous”).

A skin biopsy and phototest procedures in which skin is exposed to multiple erythemal doses of UV-A and UV-B may aid in the diagnosis. The action spectrum for PMLE is usually within these portions of the solar spectrum.

Whereas the treatment of an acute flare of PMLE may require topical or systemic glucocorticoids, approaches to preventing PMLE are important and include the use of high-SPF broad-spectrum sunscreens as well as the induction of “hardening” by the cautious administration of artificial UV-B (broad-band or narrow-band) and/or UV-A radiation or the use of psoralen plus UV-A (PUVA) photochemotherapy for ~4 weeks before initial sun exposure. Such prophylactic phototherapy or photochemotherapy at the beginning of spring may prevent the occurrence of PMLE throughout the summer.

Phototoxicity and Photoallergy

These photosensitivity disorders are related to the topical or systemic administration of drugs and other chemicals that can act as chromophores. Both reactions require the absorption of energy by a drug or chemical with consequent production of an excited-state photosensitizer that can transfer its absorbed energy to a bystander molecule or to molecular oxygen, thereby generating tissue-destructive chemical species, including ROS.

Phototoxicity is a nonimmunologic reaction that can be caused by a broad range of drugs and chemicals, a few of which are listed in **Table 57-3**. The usual clinical manifestations include erythema resembling a sunburn reaction that quickly desquamates, or “peels,” within several days. In addition, edema, vesicles, and bullae may occur.

TABLE 57-3

Drugs That May Cause a Phototoxic Reaction

Drug	Topical	Systemic
Amiodarone		+
Dacarbazine		+
Fluoroquinolones		+
5-Fluorouracil	+	+
Furosemide		+
Nalidixic acid		+
Phenothiazines		+
Psoralens	+	+
Retinoids	+/−	+
Sulfonamides		+
Sulfonylureas		+
Tetracyclines		+
Thiazides		+
Vinblastine		+

Photoallergy is much less common and is distinct in that it is an immunopathologic process. The excited-state photosensitizer may create highly unstable haptenic free radicals that bind covalently to macromolecules to form a functional antigen capable of evoking a delayed-type hypersensitivity response. Some drugs and chemicals that can produce photoallergy are listed in **Table 57-4**. The clinical manifestations typically differ from those of phototoxicity in that an intensely pruritic eczematous dermatitis tends to predominate and evolves into lichenified, thickened, “leathery” changes in sun-exposed areas. A small subset (perhaps 5–10%) of patients with photoallergy may develop a persistent exquisite hypersensitivity to

light even when the offending drug or chemical is identified and eliminated, a condition known as *persistent light reaction*.

TABLE 57-4

Drugs That May Cause a Photoallergic Reaction

Drug	Topical	Systemic
6-Methylcoumarin	+	
Aminobenzoic acid and esters	+	
Bithionol	+	
Chlorpromazine		+
Diclofenac		+
Fluoroquinolones		+
Halogenated salicylanilides	+	
Hypericin (St. John's wort)	+	+
Musk ambrette	+	
Piroxicam		+
Promethazine		+
Sulfonamides		+
Sulfonylureas		+

A very uncommon type of persistent photosensitivity is known as *chronic actinic dermatitis*. The affected patients are typically elderly men with a long history of preexisting allergic contact dermatitis or photosensitivity. These individuals are usually exquisitely sensitive to UV-B, UV-A, and visible wavelengths.

Phototoxicity and photoallergy often can be diagnostically confirmed by phototest procedures. In patients with suspected phototoxicity, determining the minimal erythemal dose (MED) while the patient is exposed to a suspected agent and then repeating the MED after discontinuation of the agent may provide a clue to the

causative drug or chemical. Photopatch testing can be performed to confirm the diagnosis of photoallergy. In this simple variant of ordinary patch testing, a series of known photoallergens is applied to the skin in duplicate, and one set is irradiated with a suberythemal dose of UV-A. The development of eczematous changes at sites exposed to sensitizer and light is a positive result. The characteristic abnormality in patients with persistent light reaction is a diminished threshold to erythema evoked by UV-B. Patients with chronic actinic dermatitis usually manifest a broad spectrum of UV hyperresponsiveness and require meticulous photoprotection, including avoidance of sun exposure, use of high-SPF (>30) sunscreens, and, in severe cases, systemic immunosuppression, such as with azathioprine.

The management of drug photosensitivity involves first and foremost the elimination of exposure to the chemical agents responsible for the reaction and the minimization of sun exposure. The acute symptoms of phototoxicity may be ameliorated by cool moist compresses, topical glucocorticoids, and systemically administered NSAIDs. In severely affected individuals, a tapered course of systemic glucocorticoids may be useful. Judicious use of analgesics may be necessary.

Photoallergic reactions require a similar management approach. Furthermore, patients with persistent light reaction and chronic actinic dermatitis must be meticulously protected against light exposure. In selected patients to whom chronic systemic high-dose glucocorticoids pose unacceptable risks, it may be necessary to employ an immunosuppressive drug such as azathioprine, cyclophosphamide, cyclosporine, or mycophenolate mofetil.

Porphyria

The porphyrias **(Chap. 409)** are a group of diseases that have in common inherited or acquired derangements in the synthesis of heme. Heme is an iron-chelated tetrapyrrole or porphyrin, and the nonmetal chelated porphyrins are potent photosensitizers that absorb light intensely in both the short (400–410 nm) and the long (580–650 nm) portions of the visible spectrum.

Heme cannot be reutilized and must be synthesized continuously. The two body compartments with the largest capacity for its production are the bone marrow and the liver. Accordingly, the porphyrias originate in one or the other of these organs, with an end result of excessive endogenous production of potent photosensitizing porphyrins. The porphyrins circulate in the bloodstream and diffuse into the skin, where they absorb solar energy, become photoexcited, generate ROS, and evoke cutaneous photosensitivity. The mechanism of porphyrin photosensitization is known to be photodynamic, or oxygen-dependent, and is mediated by ROS such as singlet oxygen and superoxide anions.

The group of cutaneous porphyrias can be classified as either causing (1) chronic blistering photosensitivity or (2) acute nonblistering photosensitivity. Chronic cutaneous porphyrias include porphyria cutanea tarda (PCT), congenital erythropoietic porphyria (CEP), hepatoerythropoietic porphyria (HEP), hereditary coproporphyria (HCP), and variegate porphyria (VP). CEP, HEP, and PCT manifest only with cutaneous symptoms, while HCP and VP have acute neurovisceral symptoms in addition to the skin photosensitivity.

Acute cutaneous nonblistering porphyrias include EPP and X-linked protoporphyria (XLP). Representative examples of chronic and acute cutaneous porphyrias are discussed below.

Porphyria cutanea tarda (PCT) is the most common type of porphyria and is associated with decreased activity of the heme pathway enzyme uroporphyrinogen decarboxylase (UROD) to <20% of normal. Increased iron and various acquired factors (e.g., alcohol consumption, estrogens, smoking, hepatitis C or HIV infection) can reduce UROD activity. There are two basic types of PCT: (1) the sporadic or acquired type, generally seen in individuals ingesting ethanol or receiving estrogens; and (2) the inherited type, in which there is autosomal dominant transmission of deficient enzyme activity (resulting in heterozygosity for UROD with a reduction to 50% of UROD enzymatic activity and thus predisposing the individual to PCT). Both forms are associated with increased hepatic iron stores.

In both types of PCT, the predominant feature is chronic photosensitivity characterized by increased fragility of sun-exposed skin, particularly areas subject to repeated trauma such as the dorsa of the hands, the forearms, the face, and the ears. The predominant skin lesions are vesicles and bullae that rupture, producing moist erosions (often with a hemorrhagic base) that heal slowly, with crusting and purplish discoloration of the affected skin. Hypertrichosis, mottled pigmentary change, and scleroderma-like induration are associated features. The diagnosis can be confirmed biochemically by measurement of urinary porphyrin excretion, plasma porphyrin assay, and assay of erythrocyte and/or hepatic UROD. Multiple mutations of the *UROD* gene have been identified in human populations. Some patients with PCT have associated mutations in the *HFE* gene, which is linked to hemochromatosis and leads to increased iron absorption by reducing hepcidin expression; these mutations could contribute to the iron overload precipitating PCT, although iron status as measured by serum ferritin, iron levels, and transferrin saturation is no different from that in PCT patients without *HFE* mutations.

Treatment of PCT consists of repeated phlebotomies to diminish the excessive hepatic iron stores and/or intermittent (twice weekly) low doses of orally administered hydroxychloroquine. This treatment is highly effective for PCT but not suited for treatment of other porphyrias. Long-term remission of the disease can often be achieved if the patient eliminates exposure to porphyrinogenic agents such as ethanol or estrogens and avoids sun exposure.

Erythropoietic protoporphyria (EPP) is an acute nonblistering cutaneous porphyria, originates in the bone marrow, and is due to genetic mutations that in most cases decrease the activity of the mitochondrial enzyme ferrochelatase. The major clinical features include acute photosensitivity characterized by painful burning and stinging of exposed skin that often develops during or just after sun exposure. There may be associated skin swelling and, after repeated episodes, a waxlike scarring.

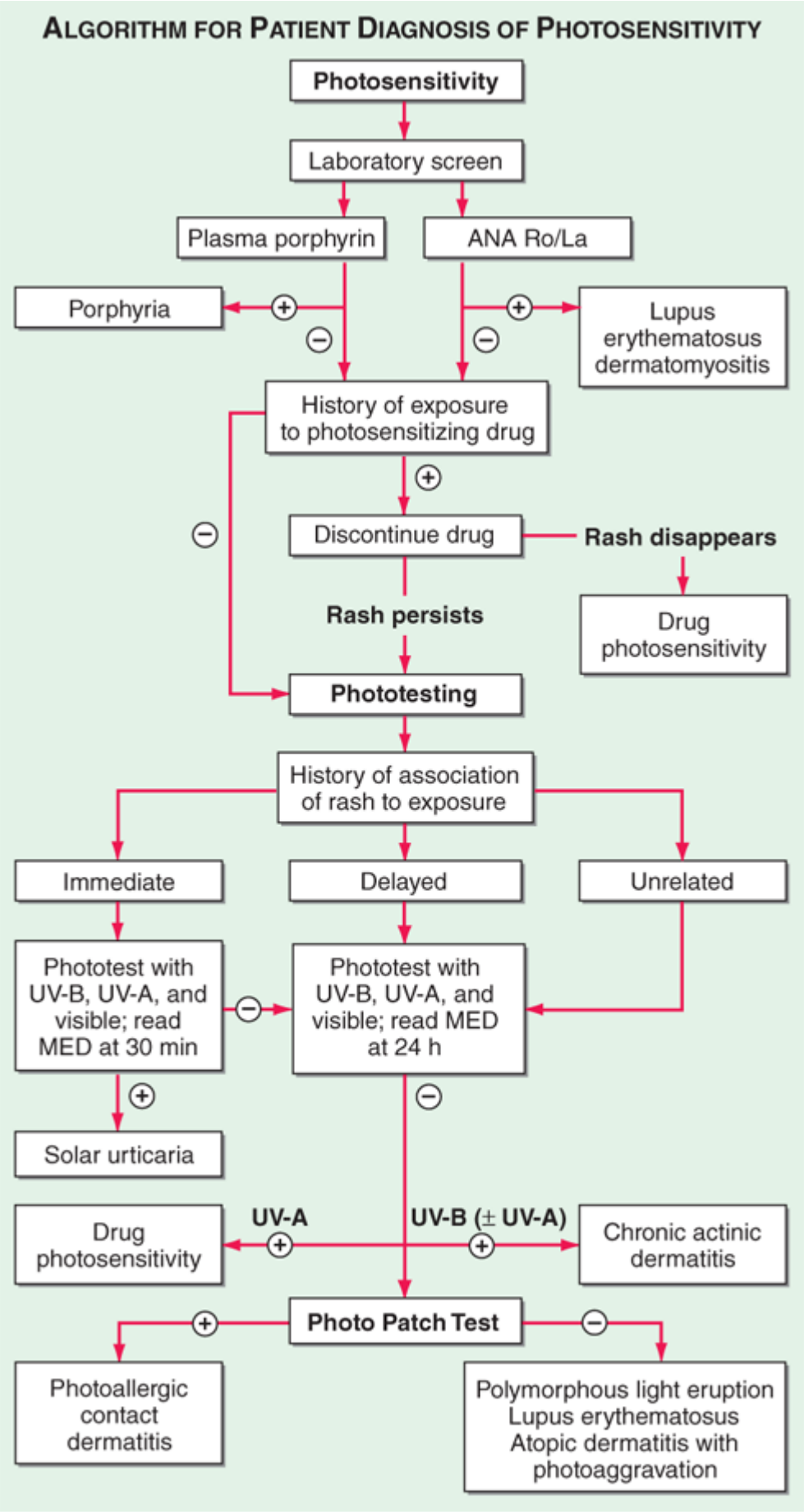
The diagnosis is confirmed by demonstration of elevated levels of free erythrocyte protoporphyrin. Detection of increased plasma protoporphyrin helps distinguish EPP from lead poisoning and iron-deficiency anemia, in both of which erythrocyte protoporphyrin levels are elevated in the absence of cutaneous photosensitivity and elevated plasma protoporphyrin levels.

Rigorous sunlight protection is essential in the management of EPP. Therapies that may increase sunlight tolerance in patients with EPP may be helpful as well, such as oral administration of β-carotene, which is an effective scavenger of free radicals. Notably, a recent clinical trial showed that a synthetic peptide analogue of α-MSH, afamelanotide, increased skin pigmentation through melanogenesis and thereby enhanced tolerance to sunlight in patients with EPP. Patients treated with afamelanotide tolerated sun exposure without pain for longer periods of time and had an improved quality of life as compared to untreated patients. Interestingly, initial studies suggest that afamelanotide may also be beneficial when combined with NB-UVB in the treatment of patients with vitiligo (in patients with skin phototypes IV–VI).

An algorithm for managing patients with photosensitivity is presented in **Fig. 57-1**.

FIGURE 57-1

Algorithm for the diagnosis of a patient with photosensitivity. ANA, antinuclear antibody; MED, minimal erythemal dose; UV-A and UV-B, ultraviolet spectrum segments including wavelengths of 320–400 nm and 290–320 nm, respectively.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PHOTOPROTECTION

Since photosensitivity of the skin results from exposure to sunlight, it follows that absolute avoidance of sunlight will eliminate these disorders. However, contemporary lifestyles make this approach impractical for most individuals. Thus, better approaches to photoprotection have been sought.

Natural photoprotection is provided by structural proteins in the epidermis, particularly keratins and melanin. The amount of melanin and its distribution in cells are genetically regulated, and individuals of darker complexion (skin types IV–VI) are at decreased risk for the development of acute sunburn and cutaneous malignancy.

Other forms of photoprotection include clothing and sunscreens. Clothing constructed of tightly woven sun-protective fabrics, irrespective of color, affords substantial protection. Wide-brimmed hats, long sleeves, and trousers all reduce direct exposure. Sunscreens are now considered over-the-counter drugs, and a monograph from the U.S. Food and Drug Administration (FDA) has recognized category I ingredients as safe and effective. Those ingredients are listed in Table 57-5. Sunscreens are rated for their photoprotective effect by their sun protection factor (SPF). The SPF is simply a ratio of the time required to produce sunburn erythema with and without sunscreen application. The SPF of most sunscreens reflects protection from UV-B but not from UV-A. The FDA monograph stipulates that sunscreens must be rated on a scale ranging from minimal (SPF ≥2 and <12) to moderate (SPF ≥12 and <30) to high (SPF ≥30, labeled as 30+).

TABLE 57-5

FDA Category I Monographed Sunscreen Ingredients

Ingredients	Maximum Concentration, %
p-Aminobenzoic acid (PABA)	15
Avobenzone	3
Cinoxate	3
Dioxybenzone (benzophenone-8)	3
Ecamsule	15
Homosalate	15
Methyl anthranilate	5
Octocrylene	10
Octyl methoxycinnamate	7.5
Octyl salicylate	5
Oxybenzone (benzophenone-3)	6
Padimate O (octyl dimethyl PABA)	8
Phenylbenzimidazole sulfonic acid	4
Sulisobenzone (benzophenone-4)	10
Titanium dioxide	25
Trolamine salicylate	12
Zinc oxide	25

Abbreviation: FDA, U.S. Food and Drug Administration.

Broad-spectrum sunscreens contain both UV-B-absorbing and UV-A-absorbing chemicals, the latter including avobenzone and ecamsule (terephthalylidene dicamphor sulfonic acid). These chemicals absorb UVR and transfer the absorbed energy to surrounding cells. In contrast, physical UV blockers (zinc oxide and titanium dioxide) scatter or reflect UVR.

In addition to light absorption, a critical determinant of the sustained photoprotective effect of sunscreens is their water resistance. The FDA monograph has defined strict testing criteria for sunscreens that claim to possess a high degree of water resistance. Some degree of photoprotection can be achieved by limiting the time of sun exposure during the day. Since a large part of an individual's total lifetime sun exposure may occur by age 18, it is important to educate parents and young children about the hazards of sunlight. Eliminating exposure at midday will substantially reduce lifetime UVR exposure.

PHOTOTHERAPY AND PHOTOCHEMOTHERAPY

UVR can be used therapeutically. The administration of UV-B alone or in combination with topically applied agents can induce remissions of many dermatologic diseases, including psoriasis and atopic dermatitis. In particular, narrow-band UV-B treatments (with fluorescent bulbs emitting radiation at ~311 nm) have enhanced efficacy over that obtained with broad-band UV-B in the treatment of psoriasis.

Photochemotherapy in which topically applied or systemically administered psoralens are combined with UV-A (PUVA) is effective in treating psoriasis and the early stages of cutaneous T cell lymphoma and vitiligo. Psoralens are tricyclic furocoumarins that, when intercalated into DNA and exposed to UV-A, form adducts with pyrimidine bases and eventually form DNA cross-links. These structural changes are thought to decrease DNA synthesis and to be related to the amelioration of psoriasis. Why PUVA photochemotherapy is effective in cutaneous T cell lymphoma is only partially understood, but it has been shown to induce apoptosis of atypical T lymphocyte populations in the skin. Consequently, direct treatment of circulating atypical lymphocytes by extracorporeal photochemotherapy (photopheresis) has been used in Sézary syndrome as well as in other severe systemic diseases with circulating atypical lymphocytes, such as graft-versus-host disease.

In addition to its effects on DNA, PUVA photochemotherapy stimulates epidermal thickening and melanin synthesis; the latter property, together with its anti-inflammatory effects, provides the rationale for use of PUVA in the depigmenting disease vitiligo. Oral 8-methoxypsoralen and UV-A appear to be most effective in this regard, but as many as 100 treatments extending over 12–18 months may be required for satisfactory repigmentation.

Not surprisingly, the major side effects of long-term UV-B phototherapy and PUVA photochemotherapy mimic those seen in individuals with chronic sun exposure. Despite these risks, the therapeutic index of these modalities continues to be excellent. It is important to choose the most appropriate phototherapeutic approach for a specific dermatologic disease. For example, narrow-band UV-B has been reported in several studies to be as effective as PUVA photochemotherapy in the treatment of psoriasis but to pose a lower risk of skin cancer development than PUVA.

FURTHER READING

Fell GL et al: Skin beta-endorphin mediates addiction to UV light. Cell 157:1527, 2014. [PubMed: 24949966]

Jansen R et al: Photoprotection: Part II. Sunscreen: development, efficacy, and controversies. J Am Acad Dermatol 69:867, 2013. [PubMed: 24238180]

Martincorena I et al: Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348:880, 2015. [PubMed: 25999502]

Sanchez-Danes A et al: Defining the clonal dynamics leading to mouse skin tumour initiation. Nature 536:298, 2016. [PubMed: 27459053]

Van Allen EM et al: Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science 350:207, 2015. [PubMed: 26359337]

Chapter 58: Interpreting Peripheral Blood Smears

Dan L. Longo

INTRODUCTION

Some of the relevant findings in peripheral blood, enlarged lymph nodes, and bone marrow are illustrated in this chapter. Systematic histologic examination of the bone marrow and lymph nodes is beyond the scope of a general medicine textbook. However, every internist should know how to examine a peripheral blood smear.

The examination of a peripheral blood smear is one of the most informative exercises a physician can perform. Although advances in automated technology have made the examination of a peripheral blood smear by a physician seem less important, the technology is not a completely satisfactory replacement for a blood smear interpretation by a trained medical professional who also knows the patient's clinical history, family history, social history, and physical findings. It is useful to ask the laboratory to generate a Wright's-stained peripheral blood smear and examine it.

The best place to examine blood cell morphology is the feathered edge of the blood smear where red cells lie in a single layer, side by side, just barely touching one another but not overlapping. The author's approach is to look at the smallest cellular elements, the platelets, first and work his way up in size to red cells and then white cells.

Using an oil immersion lens that magnifies the cells 100-fold, one counts the platelets in five to six fields, averages the number per field, and multiplies by 20,000 to get a rough estimate of the platelet count. The platelets are usually 1–2 μm in diameter and have a blue granulated appearance. There is usually 1 platelet for every 20 or so red cells. Of course, the automated counter is much more accurate, but gross disparities between the automated and manual counts should be assessed. Large platelets may be a sign of rapid platelet turnover, as young platelets are often larger than old ones; alternatively, certain rare inherited syndromes can produce large platelets. If the platelet count is low, the absence of large (young) platelets may be an indicator of marrow production problems. Platelet clumping visible on the smear can be associated with falsely low automated platelet counts. Clumping may be caused by the anticoagulant into which the blood is drawn. Similarly, neutrophil fragmentation can be a source of falsely elevated automated platelet counts. The absence of platelet granules may be an artifact of the handling of the blood or may indicate marrow disease or a rare congenital anomaly, gray platelet syndrome. Elevated platelet counts usually signify a myeloproliferative disorder or a reaction to systemic inflammation.

Next one examines the red blood cells. One can gauge their size by comparing the red cell to the nucleus of a small lymphocyte. Both are normally about 8-µm wide. Red cells that are smaller than the small lymphocyte nucleus may be microcytic; those larger than the small lymphocyte nucleus may be macrocytic. Macrocytic cells also tend to be more oval than spherical in shape and are sometimes called macroovalocytes. The automated mean corpuscular volume (MCV) can assist in making a classification. However, some patients may have both iron and vitamin B_{12} deficiency, which will produce an MCV in the normal range but wide variation in red cell size. When the red cells vary greatly in size, *anisocytosis* is said to be present. When the red cells vary greatly in shape, *poikilocytosis* is said to be present. The electronic cell counter provides an independent assessment of variability in red cell size. It measures the range of red cell volumes and reports the results as "red cell distribution width" (RDW). This value is calculated from the MCV; thus, cell width is not being measured but cell volume is. The term is derived from the curve displaying the frequency of cells at each volume, also called the distribution. The width of red cell volume distribution curve is what determines the RDW. The RDW is calculated as follows: RDW = (standard deviation of MCV ÷ mean MCV) × 100. In the presence of morphologic anisocytosis, RDW (normally 11–14%) increases to 15–18%. The RDW is useful in at least two clinical settings. In patients with microcytic anemia, the differential diagnosis is generally between iron deficiency and thalassemia. In thalassemia, the small red cells are generally of uniform size with a normal small RDW. In iron deficiency, the size variability and the RDW are large. In addition, a large RDW can suggest a dimorphic anemia when a chronic atrophic gastritis can produce both vitamin B_{12} malabsorption to produce macrocytic anemia and blood loss to produce iron deficiency. In such settings, RDW is also large. An elevated RDW also has been reported as a risk factor for all-cause mortality in population-based studies, a finding that is unexplained currently.

After red cell size is assessed, one examines the hemoglobin content of the cells. They are either normal in color (*normochromic*) or pale in color (*hypochromic*). They are never "hyperchromic." If more than the normal amount of hemoglobin is made, the cells get larger—they do not become darker. In addition to hemoglobin content, the red cells are examined for inclusions. Red cell inclusions are the following:

1. *Basophilic stippling*—diffuse fine or coarse blue dots in the red cell usually representing RNA residue—especially common in lead poisoning

2. *Howell-Jolly bodies*—dense blue circular inclusions that represent nuclear remnants—their presence implies defective splenic function

3. *Nuclei*—red cells may be released or pushed out of the marrow prematurely before nuclear extrusion—often implies a myelophthisic process or a vigorous narrow response to anemia, usually hemolytic anemia

4. *Parasites*—red cell parasites include malaria and babesia (Chap. A6)

5. *Polychromatophilia*—the red cell cytoplasm has a bluish hue, reflecting the persistence of ribosomes still actively making hemoglobin in a young red cell

Vital stains are necessary to see precipitated hemoglobin called *Heinz bodies*.

Red cells can take on a variety of different shapes. All abnormally shaped red cells are *poikilocytes*. Small red cells without the central pallor are *spherocytes*; they can be seen in hereditary spherocytosis, hemolytic anemias of other causes, and clostridial sepsis. *Dacrocytes* are teardrop-shaped cells that can be seen in hemolytic anemias, severe iron deficiency, thalassemias, myelofibrosis, and myelodysplastic syndromes. *Schistocytes* are helmet-shaped cells that reflect microangiopathic hemolytic anemia or fragmentation on an artificial heart valve. *Echinocytes* are spiculated red cells with the spikes evenly spaced; they can represent an artifact of abnormal drying of the blood smear or reflect changes in stored blood. They also can be seen in renal failure and malnutrition and are often reversible. *Acanthocytes* are spiculated red cells with the spikes irregularly distributed. This process tends to be irreversible and reflects underlying renal disease, abetalipoproteinemia, or splenectomy. *Elliptocytes* are elliptical-shaped red cells that can reflect an inherited defect in the red cell membrane, but they also are seen in iron deficiency, myelodysplastic syndromes, megaloblastic anemia, and thalassemias. *Stomatocytes* are red cells in which the area of central pallor takes on the morphology of a slit instead of the usual round shape. Stomatocytes can indicate an inherited red cell membrane defect and also can be seen in alcoholism. *Target cells* have an area of central pallor that contains a dense center, or bull's-eye. These cells are seen classically in thalassemia, but they are also present in iron deficiency, cholestatic liver disease, and some hemoglobinopathies. They also can be generated artifactually by improper slide making.

One last feature of the red cells to assess before moving to the white blood cells is the distribution of the red cells on the smear. In most individuals, the cells lie side by side in a single layer. Some patients have red cell clumping (called *agglutination*) in which the red cells pile upon one another; it is seen in certain paraproteinemias and autoimmune hemolytic anemias. Another abnormal distribution involves red cells lying in single cell rows on top of one another like stacks of coins. This is called *rouleaux formation* and reflects abnormal serum protein levels.

Finally, one examines the white blood cells. Three types of granulocytes are usually present: neutrophils, eosinophils, and basophils, in decreasing frequency. Neutrophils are generally the most abundant white cell. They are round, are 10–14 μm wide, and contain a lobulated nucleus with two to five lobes connected by a thin chromatin thread. Bands are immature neutrophils that have not completed nuclear condensation and have a U-shaped nucleus. Bands reflect a left shift in neutrophil maturation in an effort to make more cells more rapidly. Neutrophils can provide clues to a variety of conditions. Vacuolated neutrophils may be a sign of bacterial sepsis. The presence of 1- to 2-μm blue cytoplasmic inclusions, called *Döhle bodies*, can reflect infections, burns, or other inflammatory states. If the neutrophil granules are larger than normal and stain a darker blue, "toxic granulations" are said to be present, and they also suggest a systemic inflammation. The presence of neutrophils with more than five nuclear lobes suggests megaloblastic anemia. Large misshapen granules may reflect the inherited Chédiak-Higashi syndrome.

Eosinophils are slightly larger than neutrophils, have bilobed nuclei, and contain large red granules. Diseases of eosinophils are associated with too many of them rather than any morphologic or qualitative change. They normally total less than one-thirtieth the number of neutrophils. Basophils are even more rare than

eosinophils in the blood. They have large dark blue granules and may be increased as part of chronic myeloid leukemia.

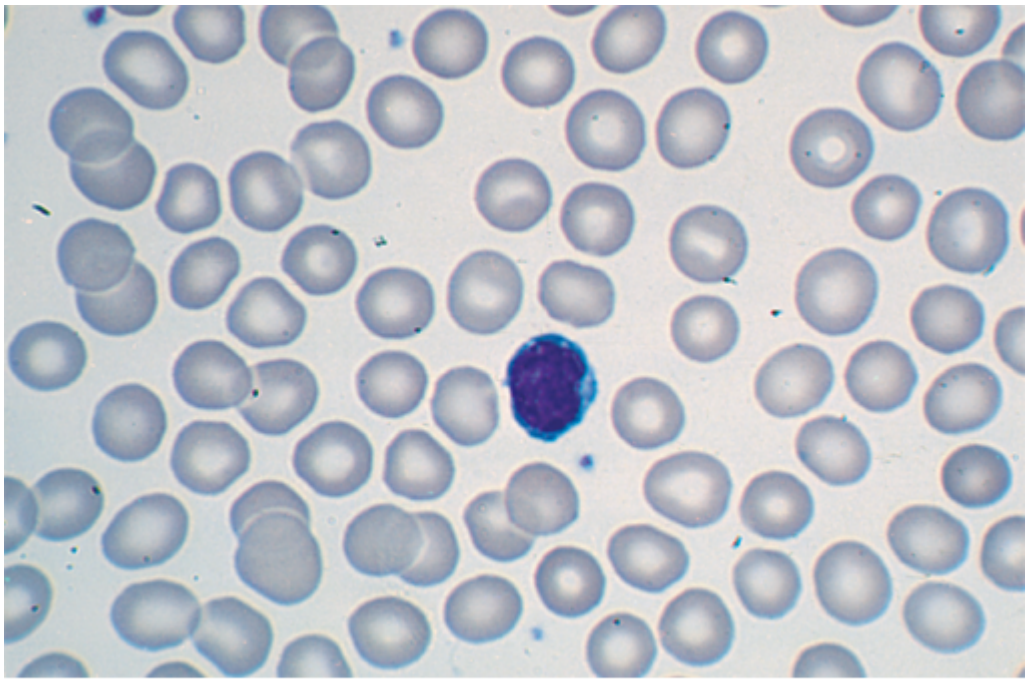
Lymphocytes can be present in several morphologic forms. Most common in healthy individuals are small lymphocytes with a small dark nucleus and scarce cytoplasm. In the presence of viral infections, more of the lymphocytes are larger, about the size of neutrophils, with abundant cytoplasm and a less condensed nuclear chromatin. These cells are called *reactive lymphocytes*. About 1% of lymphocytes are larger and contain blue granules in a light blue cytoplasm; they are called *large granular lymphocytes*. In chronic lymphoid leukemia, the small lymphocytes are increased in number, and many of them are ruptured in making the blood smear, leaving a smudge of nuclear material without a surrounding cytoplasm or cell membrane; they are called *smudge cells* and are rare in the absence of chronic lymphoid leukemia.

Monocytes are the largest white blood cells, ranging from 15 to 22 μm in diameter. The nucleus can take on a variety of shapes but usually appears to be folded; the cytoplasm is gray.

Abnormal cells may appear in the blood. Most often the abnormal cells originate from neoplasms of bone marrow–derived cells, including lymphoid cells, myeloid cells, and occasionally red cells. More rarely, other types of tumors can get access to the bloodstream, and rare epithelial malignant cells may be identified. The chances of seeing such abnormal cells are increased by examining blood smears made from buffy coats, the layer of cells that is visible on top of sedimenting red cells when blood is left in the test tube for an hour. Smears made from finger sticks may include rare endothelial cells.

FIGURE 58-1

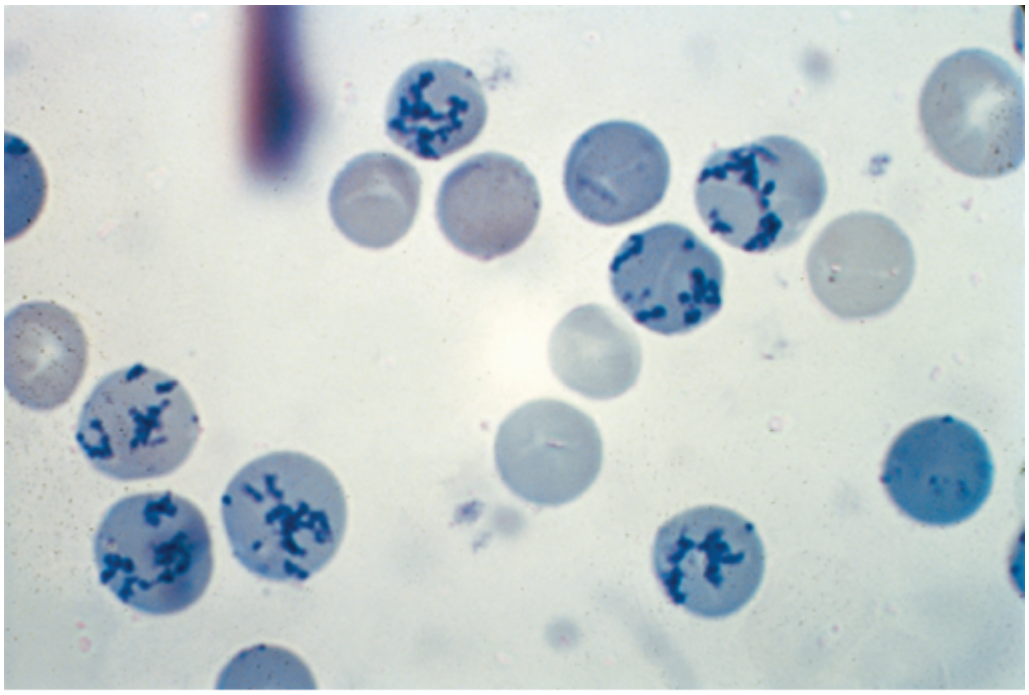
Normal peripheral blood smear. Small lymphocyte in center of field. Note that the diameter of the red blood cell is similar to the diameter of the small lymphocyte nucleus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-2

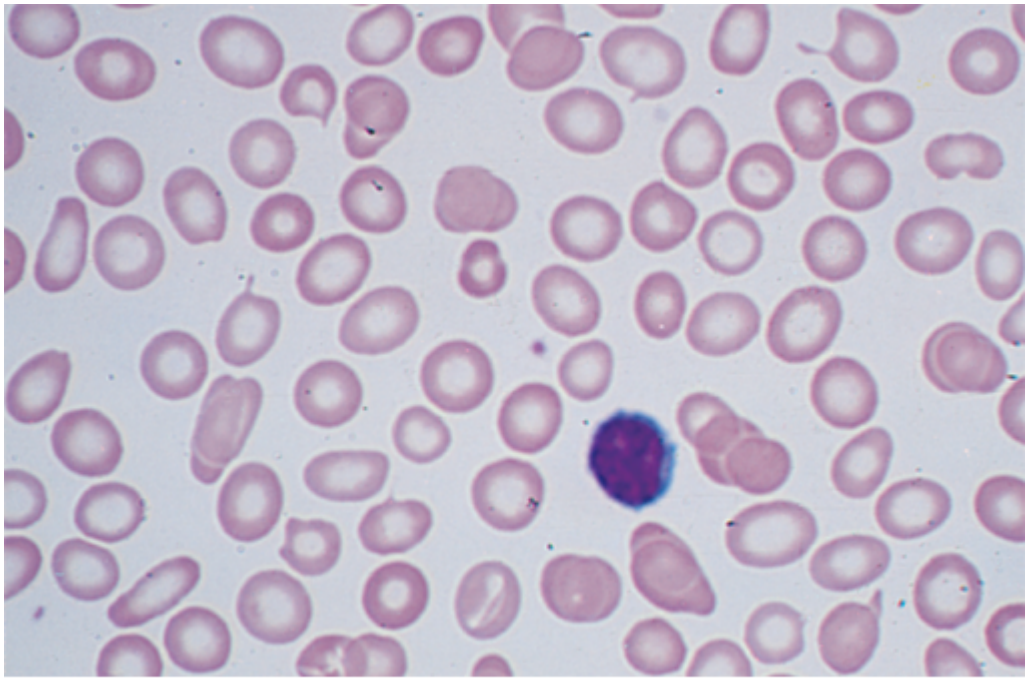
Reticulocyte count preparation. This new methylene blue–stained blood smear shows large numbers of heavily stained reticulocytes (the cells containing the dark blue–staining RNA precipitates).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-3

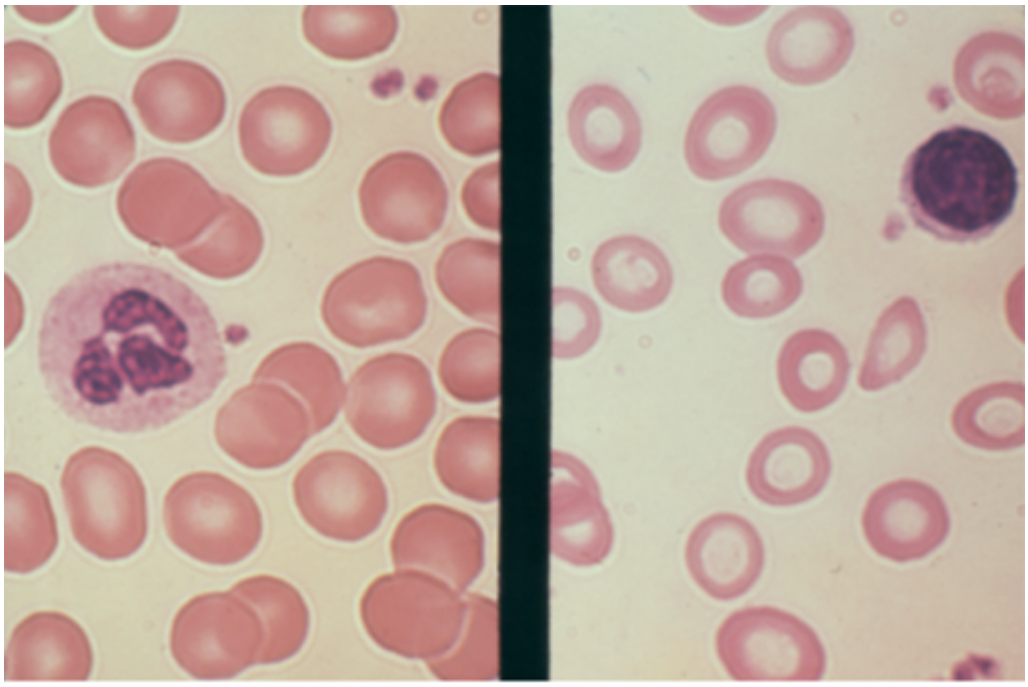
Hypochromic microcytic anemia of iron deficiency. Small lymphocyte in field helps assess the red blood cell size.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-4

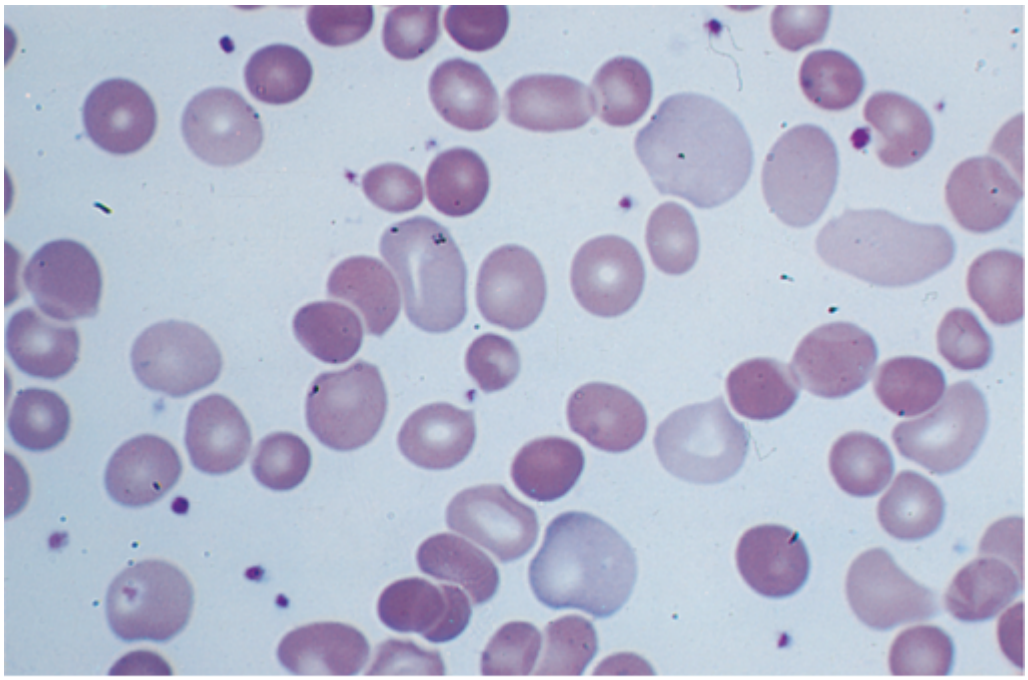
Iron deficiency anemia next to normal red blood cells. Microcytes (*right panel*) are smaller than normal red blood cells (cell diameter <7 μm) and may or may not be poorly hemoglobinized (hypochromic).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-5

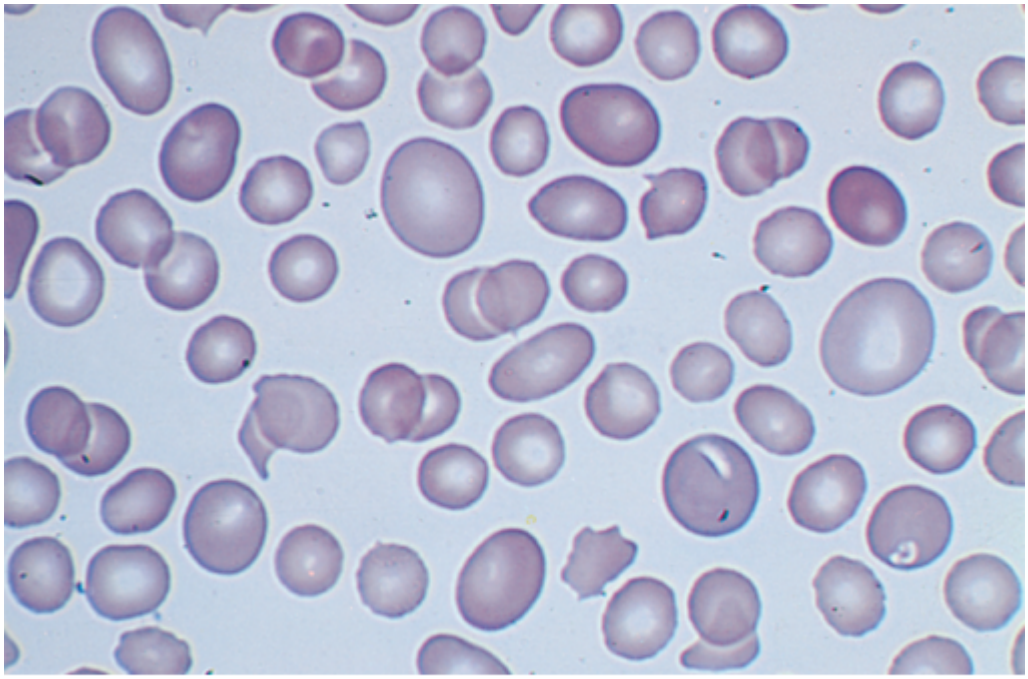
Polychromatophilia. Note large red cells with light purple coloring.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-6

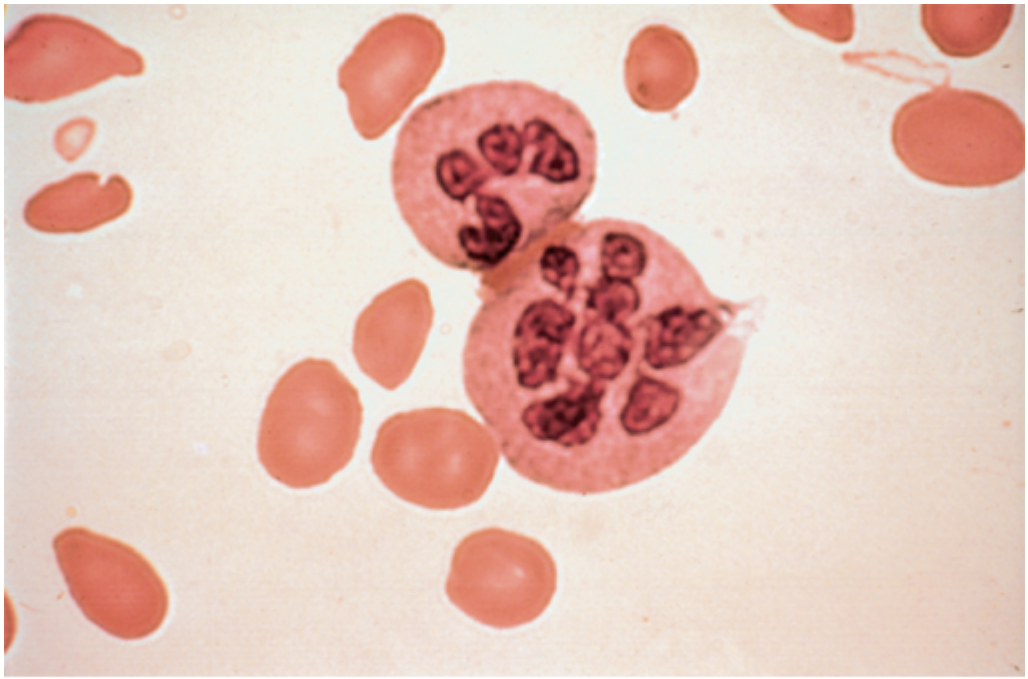
Macrocytosis. These cells are both larger than normal (mean corpuscular volume >100) and somewhat oval in shape. Some morphologists call these cells macroovalocytes.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-7

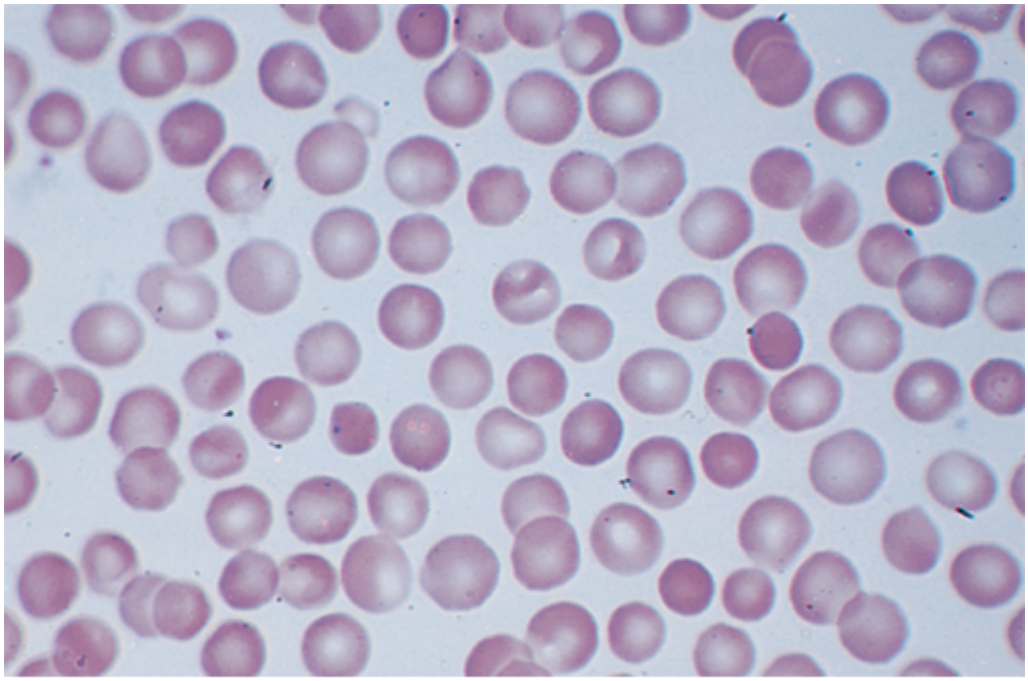
Hypersegmented neutrophils. Hypersegmented neutrophils (multilobed polymorphonuclear leukocytes) are larger than normal neutrophils with five or more segmented nuclear lobes. They are commonly seen with folic acid or vitamin B_{12} deficiency.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-8

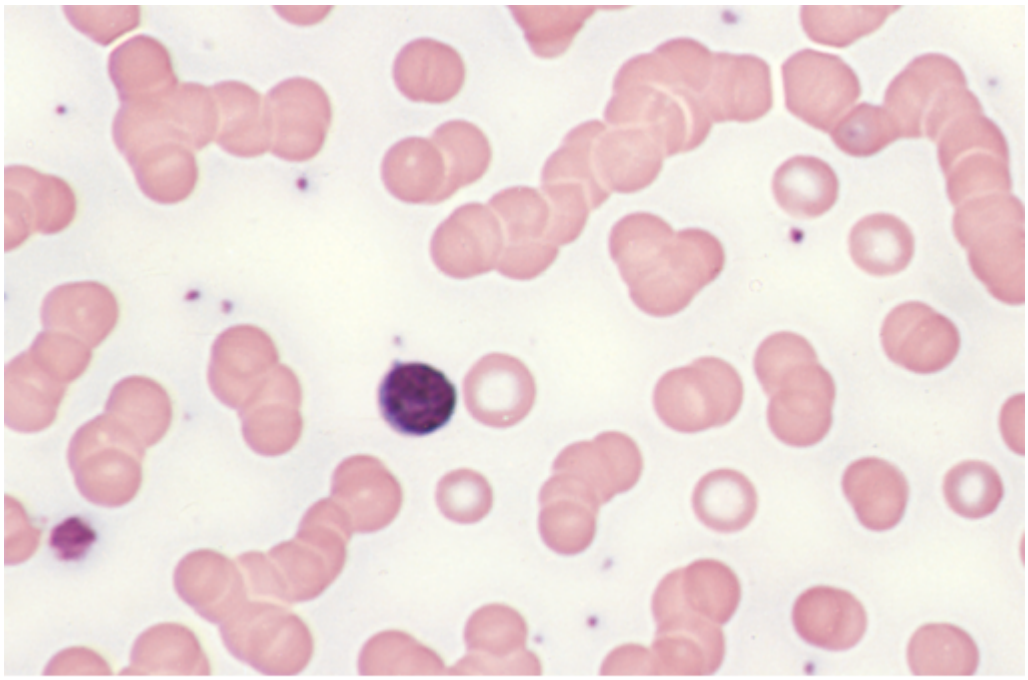
Spherocytosis. Note small hyperchromatic cells without the usual clear area in the center.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-9

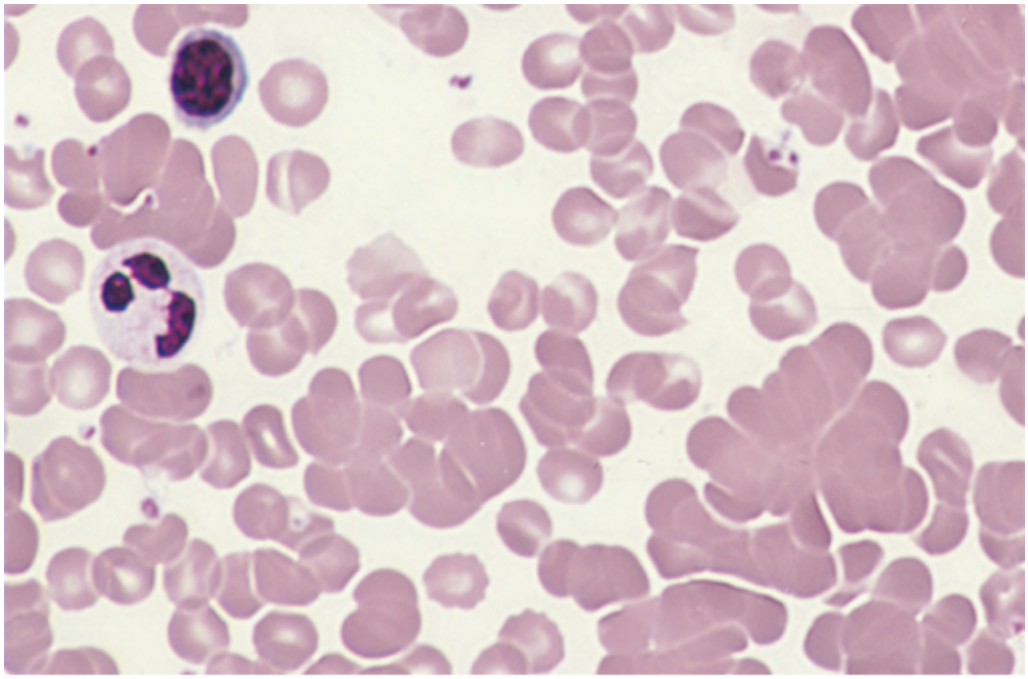
Rouleaux formation. Small lymphocyte in center of field. These red cells align themselves in stacks and are related to increased serum protein levels.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-10

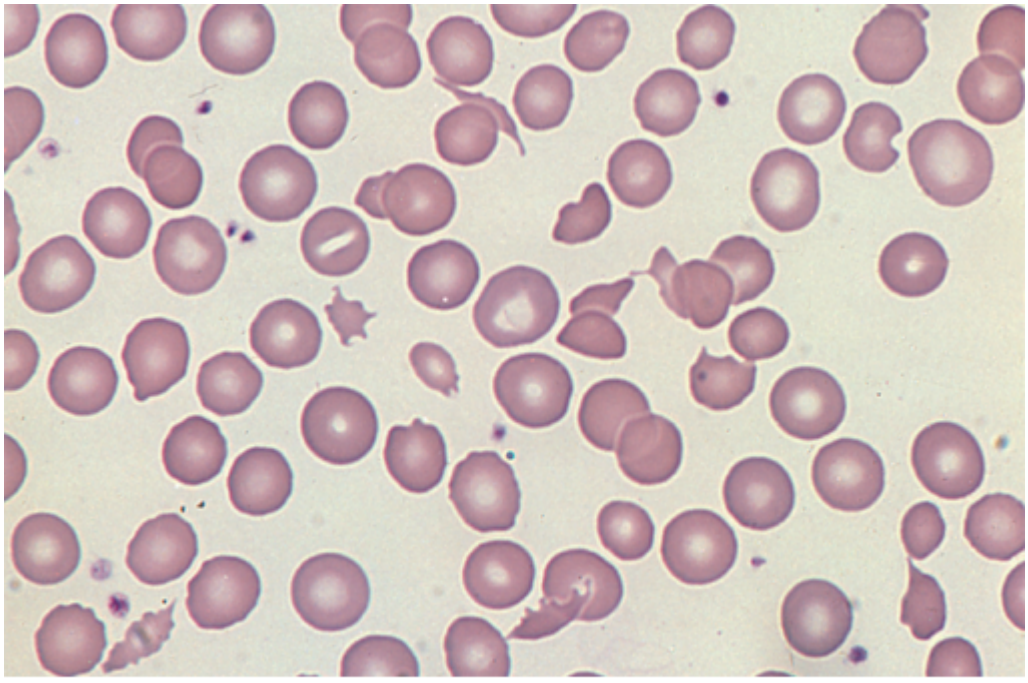
Red cell agglutination. Small lymphocyte and segmented neutrophil in upper left center. Note irregular collections of aggregated red cells.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-11

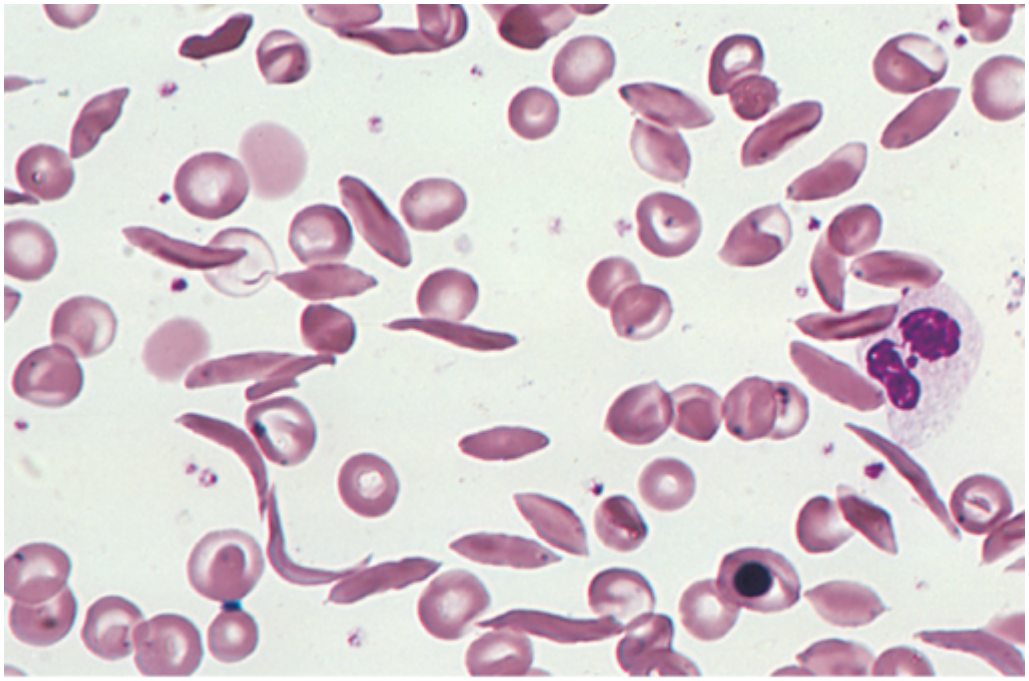
Fragmented red cells. Heart valve hemolysis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-12

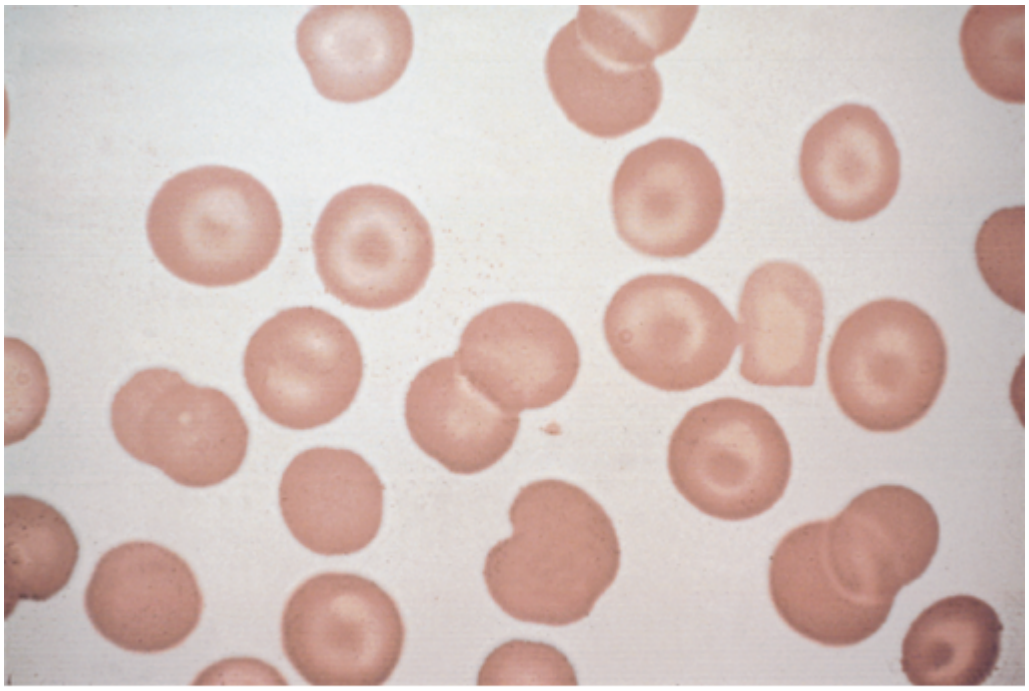
Sickle cells. Homozygous sickle cell disease. A nucleated red cell and neutrophil are also in the field.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-13

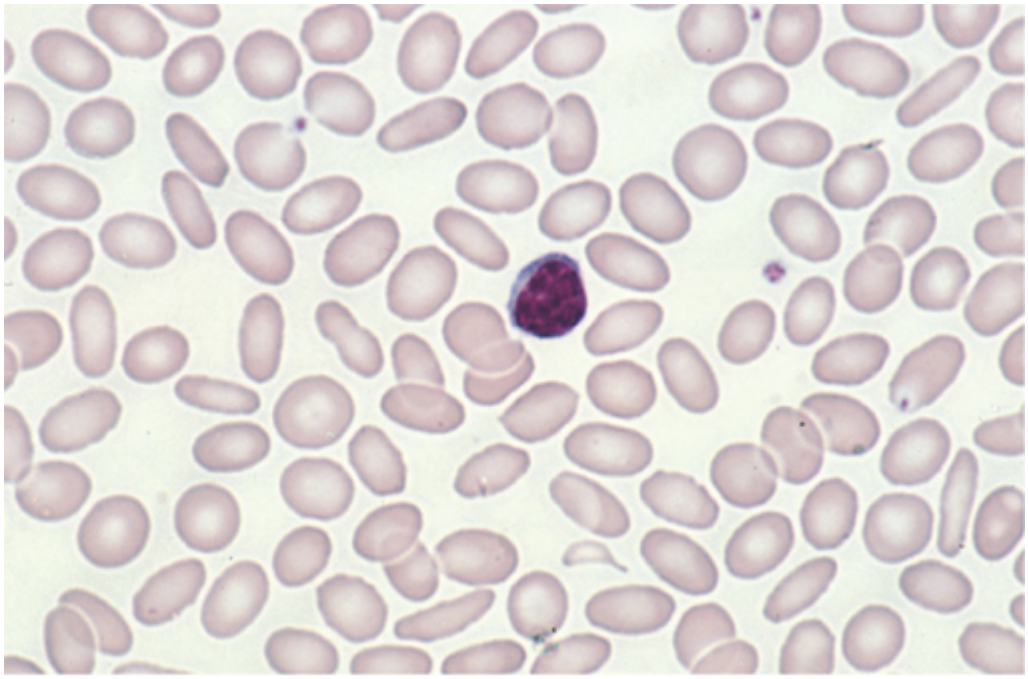
Target cells. Target cells are recognized by the bull's-eye appearance of the cell. Small numbers of target cells are seen with liver disease and thalassemia. Larger numbers are typical of hemoglobin C disease.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-14

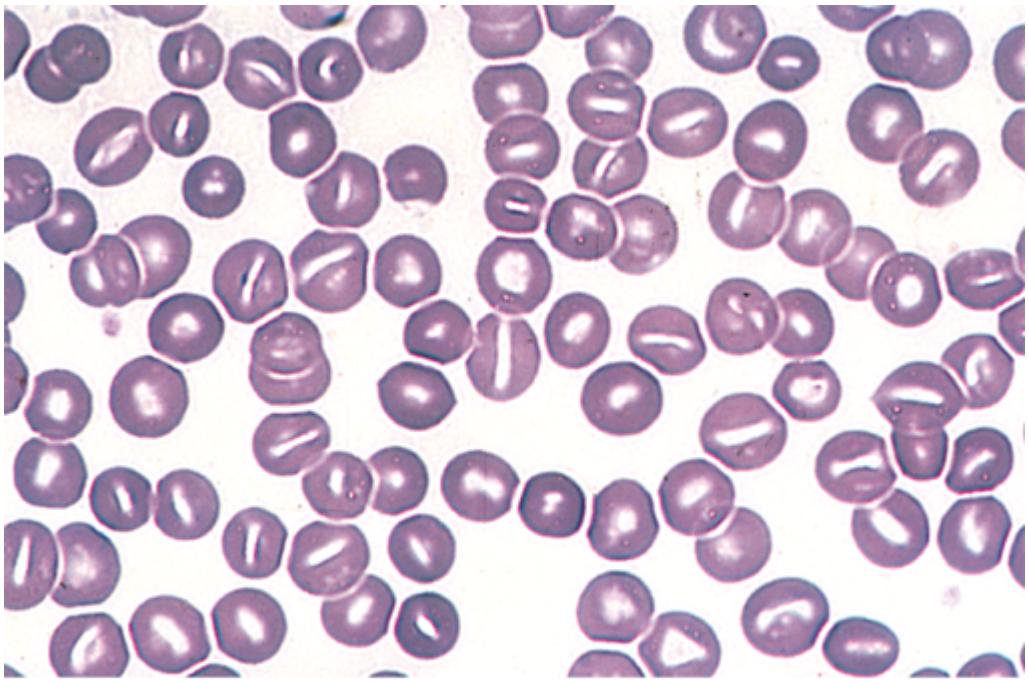
Elliptocytosis. Small lymphocyte in center of field. Elliptical shape of red cells related to weakened membrane structure, usually due to mutations in spectrin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-15

Stomatocytosis. Red cells characterized by a wide transverse slit or stoma. This often is seen as an artifact in a dehydrated blood smear. These cells can be seen in hemolytic anemias and in conditions in which the red cell is overhydrated or dehydrated.

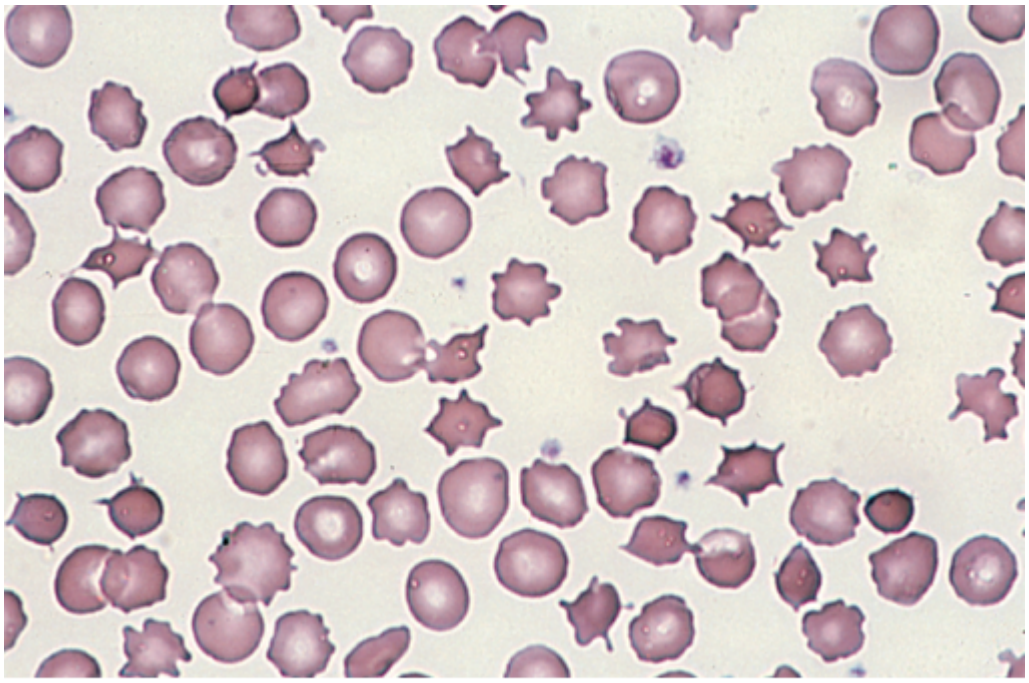


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-16

Acanthocytosis. Spiculated red cells are of two types: *acanthocytes* are contracted dense cells with irregular membrane projections that vary in length and width; *echinocytes* have small, uniform, and evenly spaced membrane projections. Acanthocytes are present in severe liver disease, in patients with

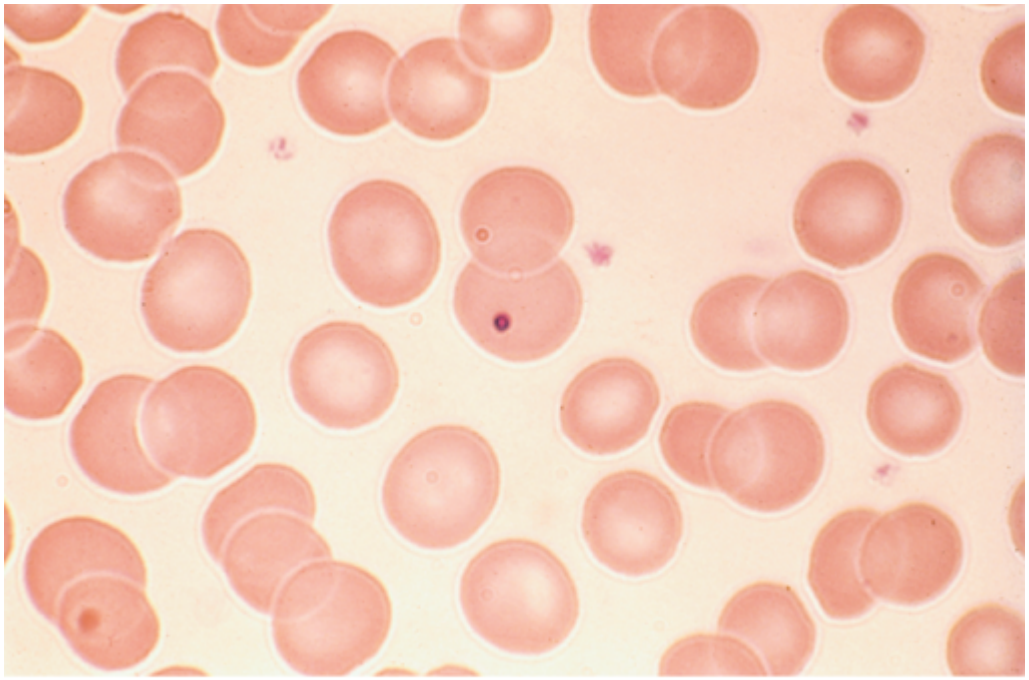
abetalipoproteinemia, and in rare patients with McLeod blood group. Echinocytes are found in patients with severe uremia, in glycolytic red cell enzyme defects, and in microangiopathic hemolytic anemia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-17

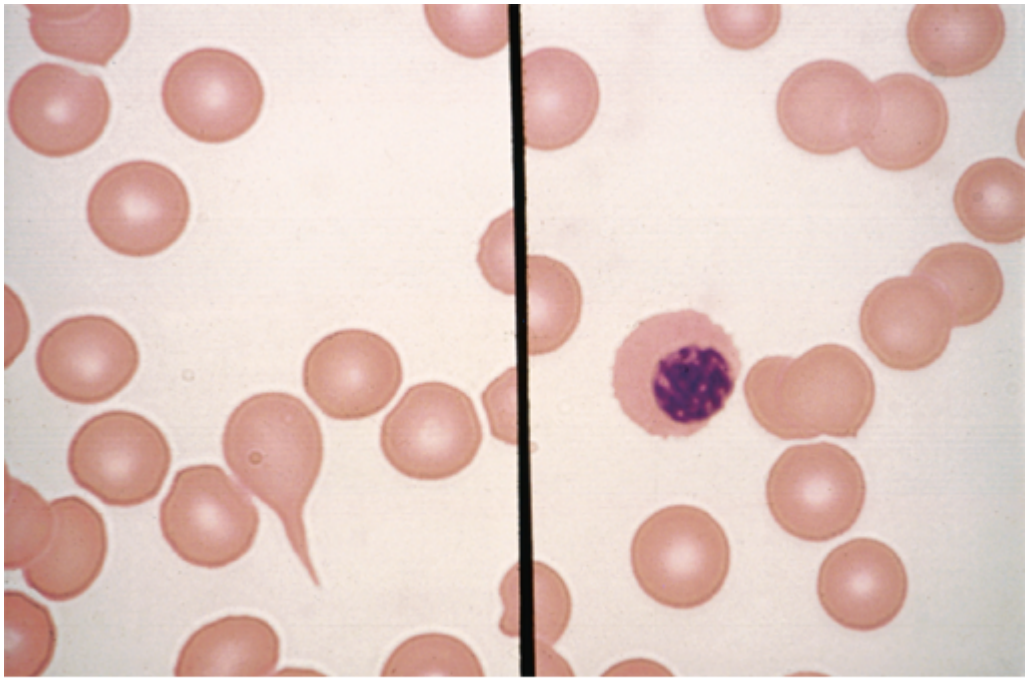
Howell-Jolly bodies. Howell-Jolly bodies are tiny nuclear remnants that normally are removed by the spleen. They appear in the blood after splenectomy (defect in removal) and with maturation/dysplastic disorders (excess production).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-18

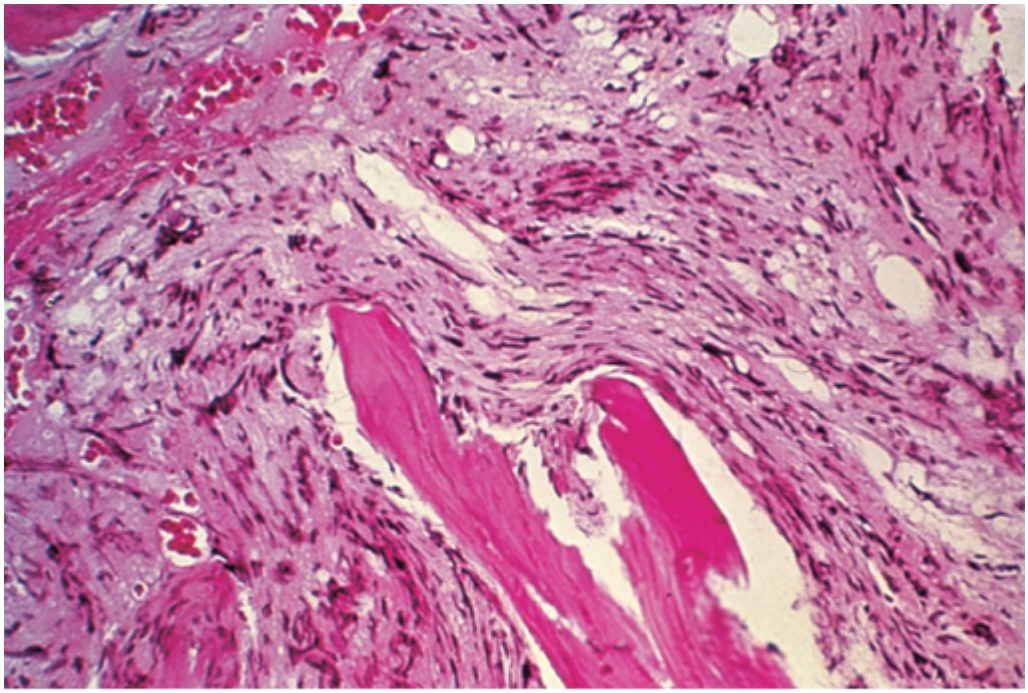
Teardrop cells and nucleated red blood cells characteristic of myelofibrosis. A teardrop-shaped red blood cell (*left panel*) and a nucleated red blood cell (*right panel*) as typically seen with myelofibrosis and extramedullary hematopoiesis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-19

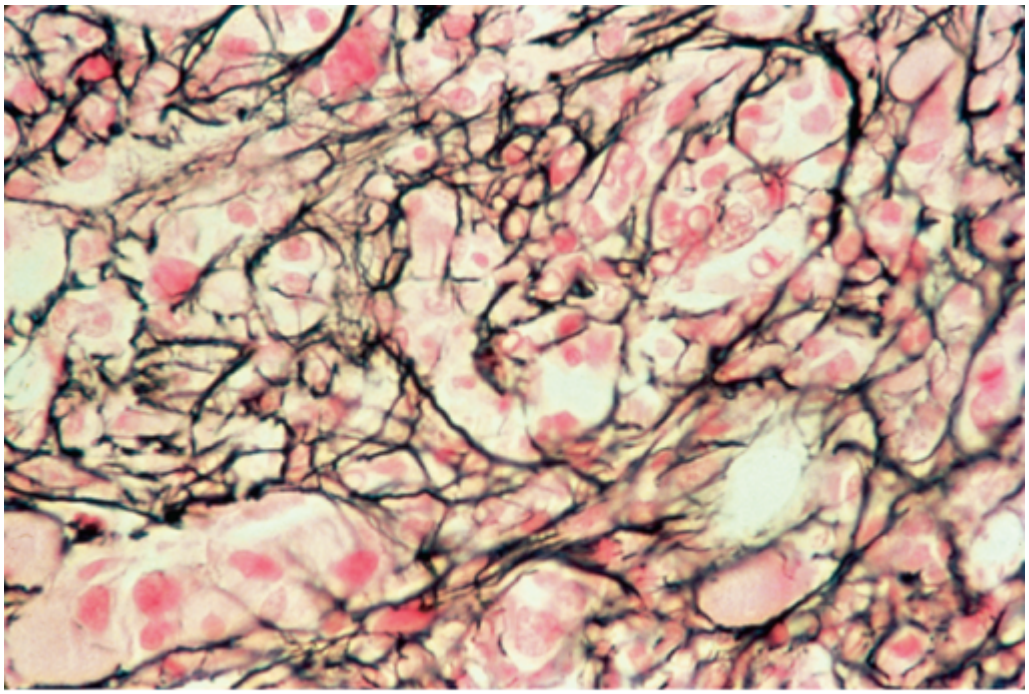
Myelofibrosis of the bone marrow. Total replacement of marrow precursors and fat cells by a dense infiltrate of reticulin fibers and collagen (H&E stain).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-20

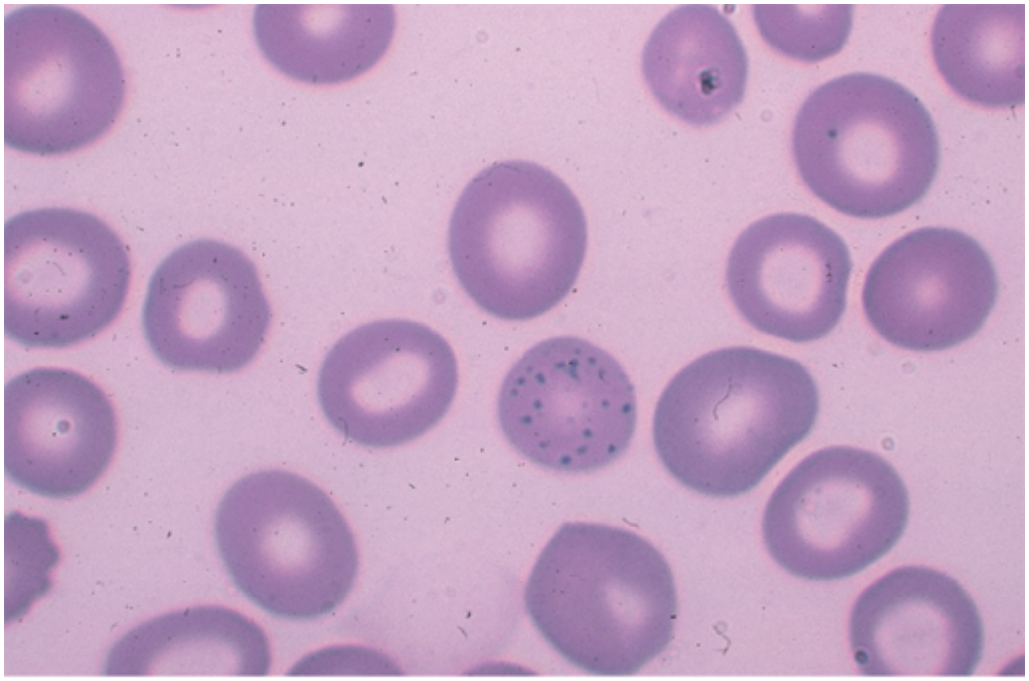
Reticulin stain of marrow myelofibrosis. Silver stain of a myelofibrotic marrow showing an increase in reticulin fibers (black-staining threads).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-21

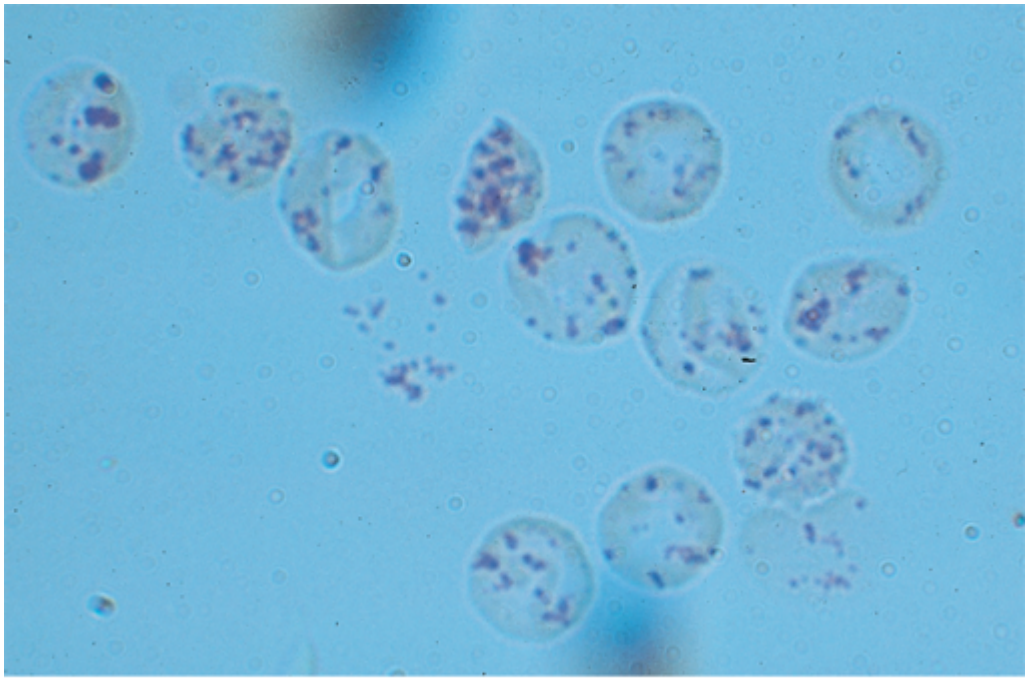
Stippled red cell in lead poisoning. Mild hypochromia. Coarsely stippled red cell.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-22

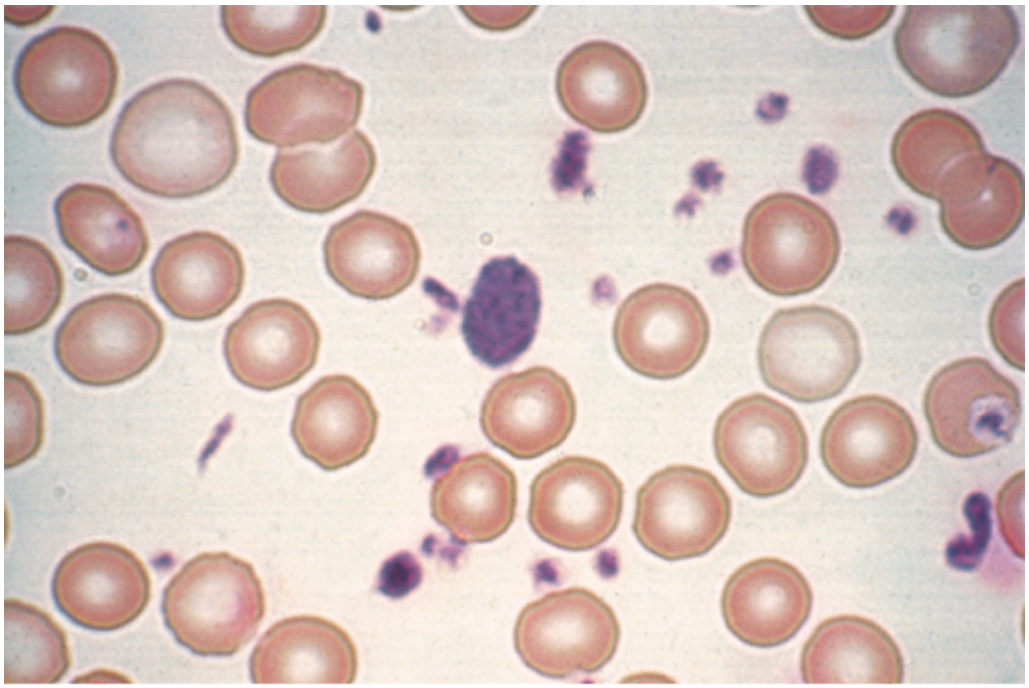
Heinz bodies. Blood mixed with hypotonic solution of crystal violet. The stained material is precipitates of denatured hemoglobin within cells.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-23

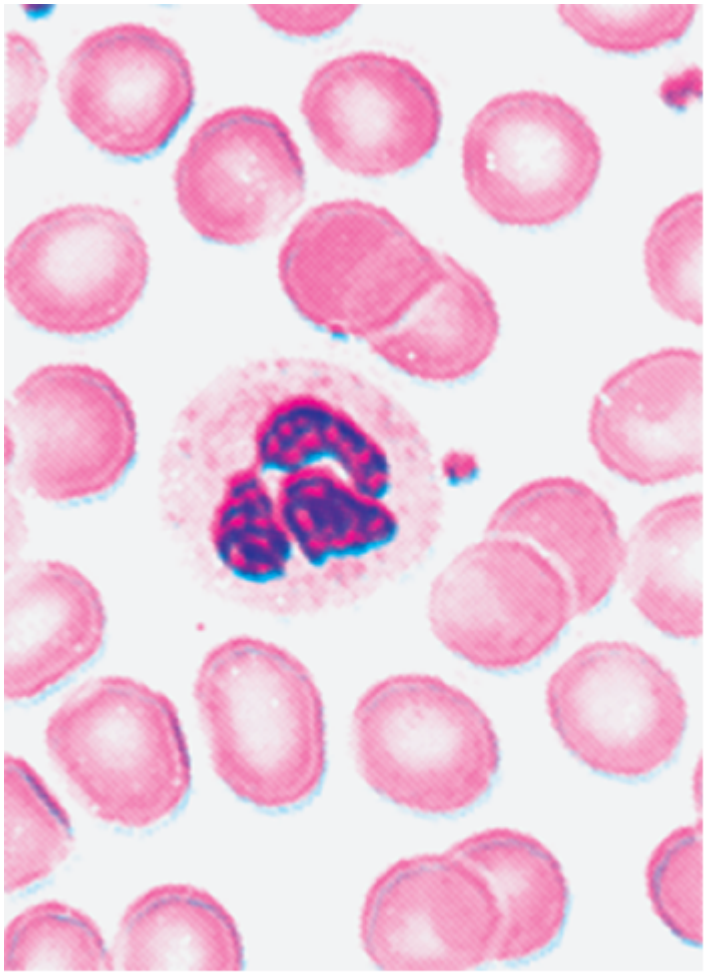
Giant platelets. Giant platelets, together with a marked increase in the platelet count, are seen in myeloproliferative disorders, especially primary thrombocythemia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-24

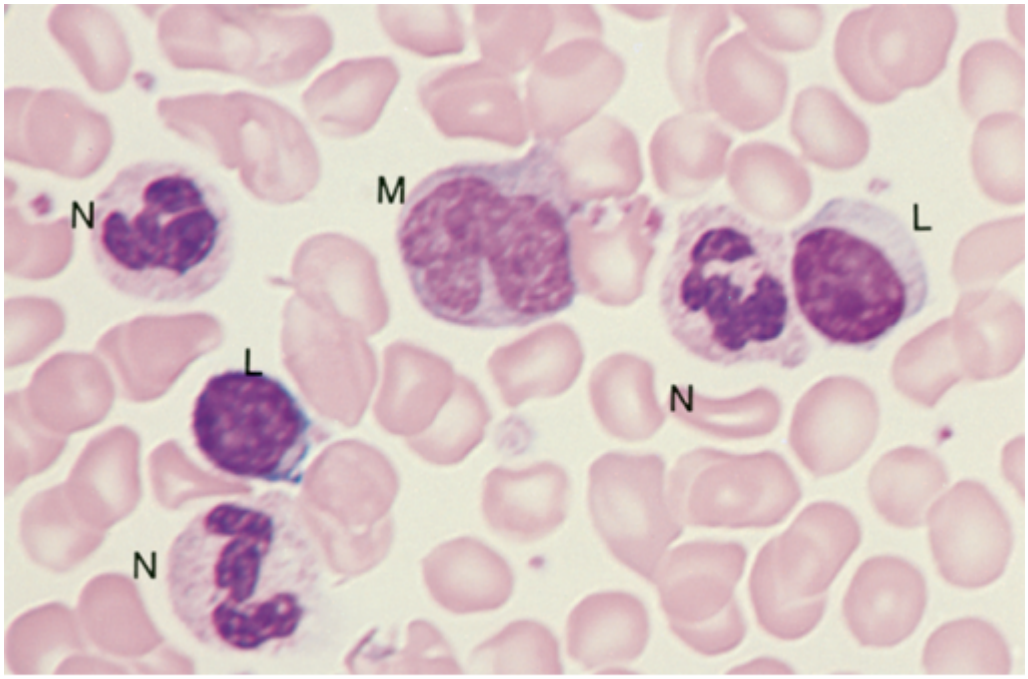
Normal granulocytes. The normal granulocyte has a segmented nucleus with heavy, clumped chromatin; fine neutrophilic granules are dispersed throughout the cytoplasm.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-25

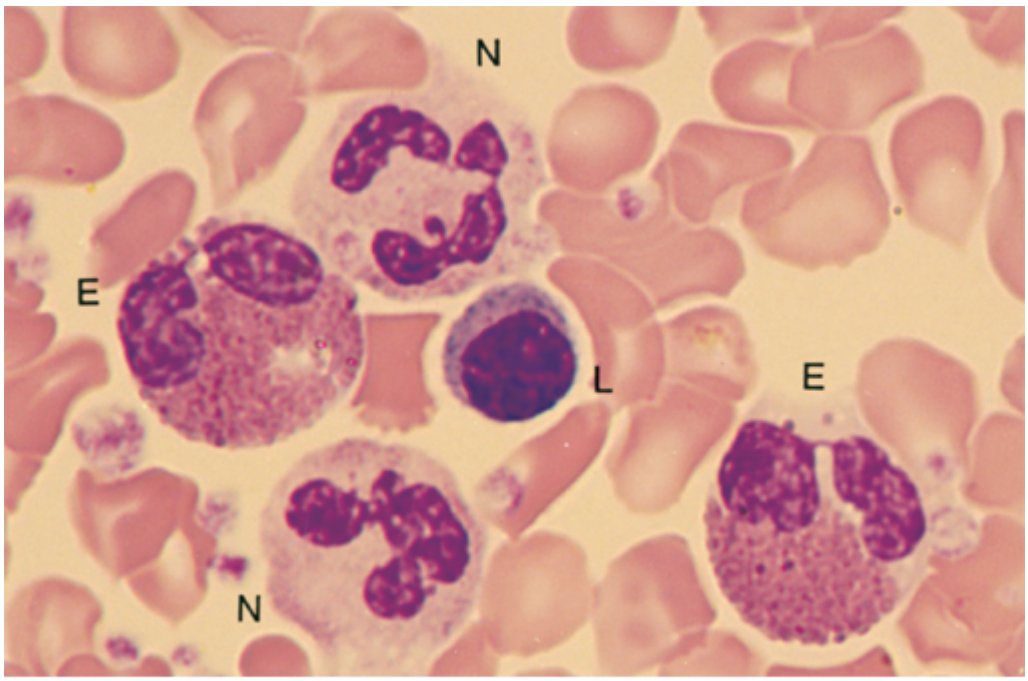
Normal monocytes. The film was prepared from the buffy coat of the blood from a normal donor. L, lymphocyte; M, monocyte; N, neutrophil.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-26

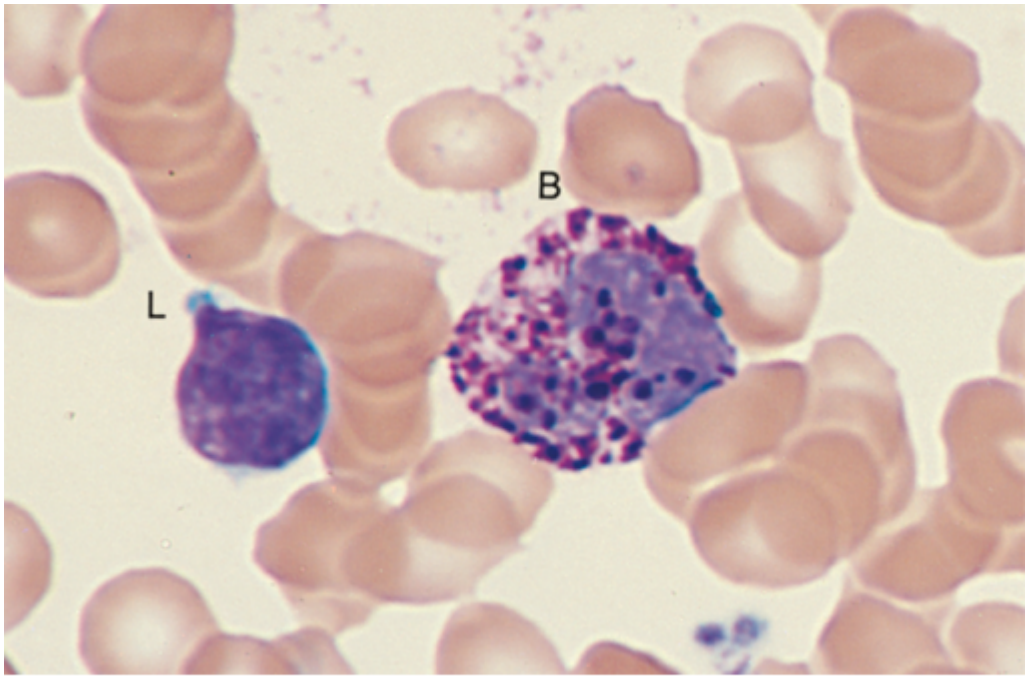
Normal eosinophils. The film was prepared from the buffy coat of the blood from a normal donor. E, eosinophil; L, lymphocyte; N, neutrophil.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-27

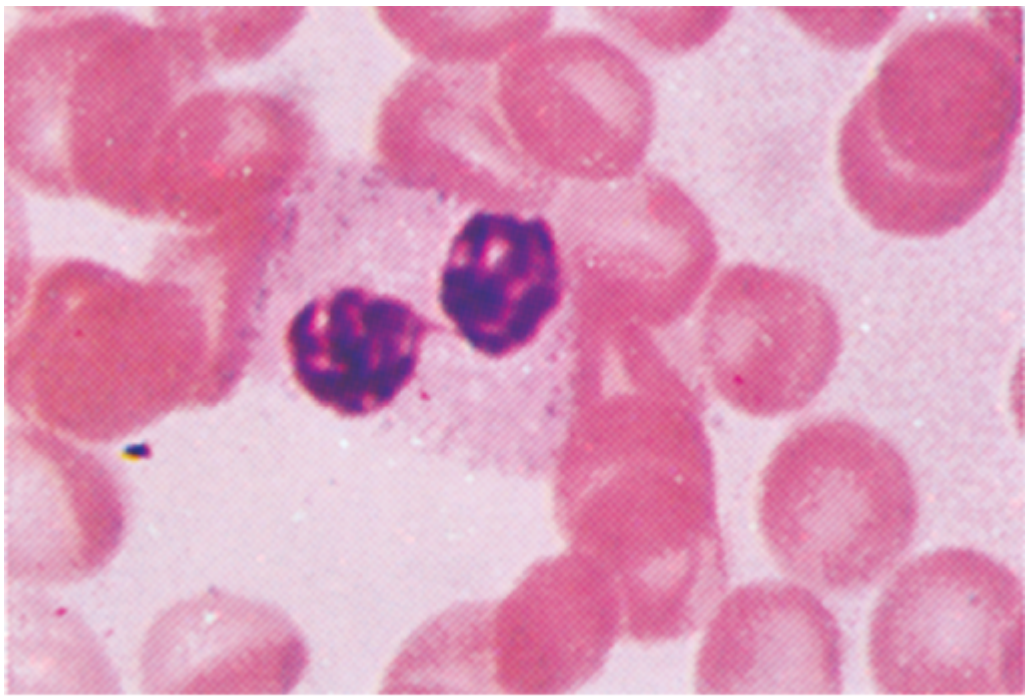
Normal basophil. The film was prepared from the buffy coat of the blood from a normal donor. B, basophil; L, lymphocyte.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-28

Pelger-Hüet anomaly. In this benign disorder, the majority of granulocytes are bilobed. The nucleus frequently has a spectacle-like, or "pince-nez," configuration.

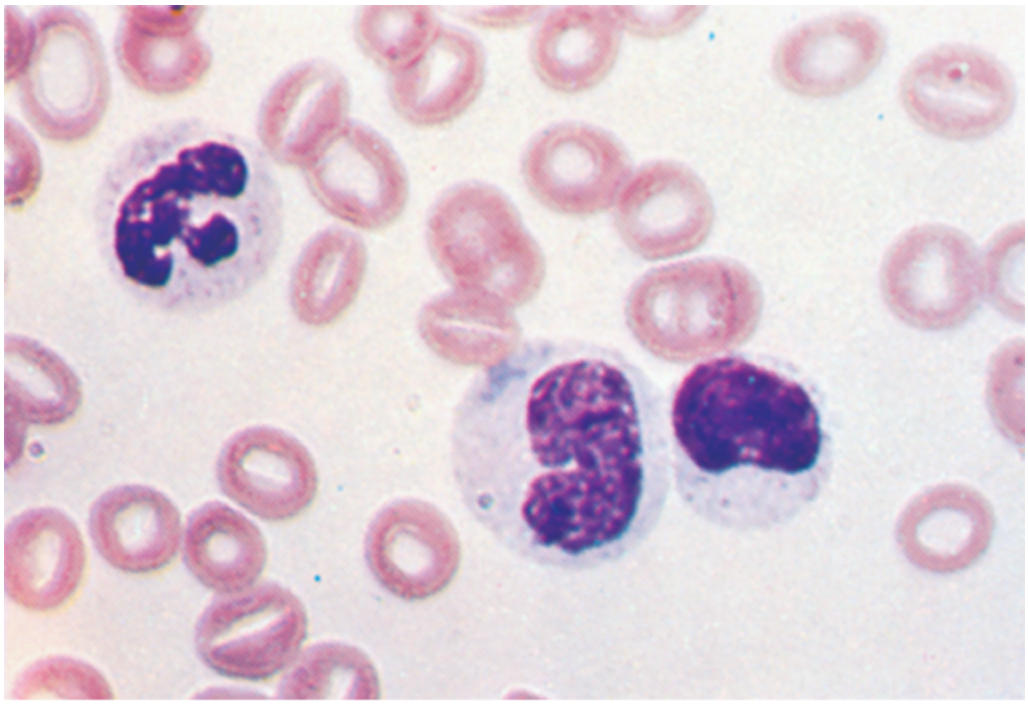


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-29

Döhle body. Neutrophil band with Döhle body. The neutrophil with a sausage-shaped nucleus in the center of the field is a band form. Döhle bodies are discrete, blue-staining nongranular areas found in the periphery

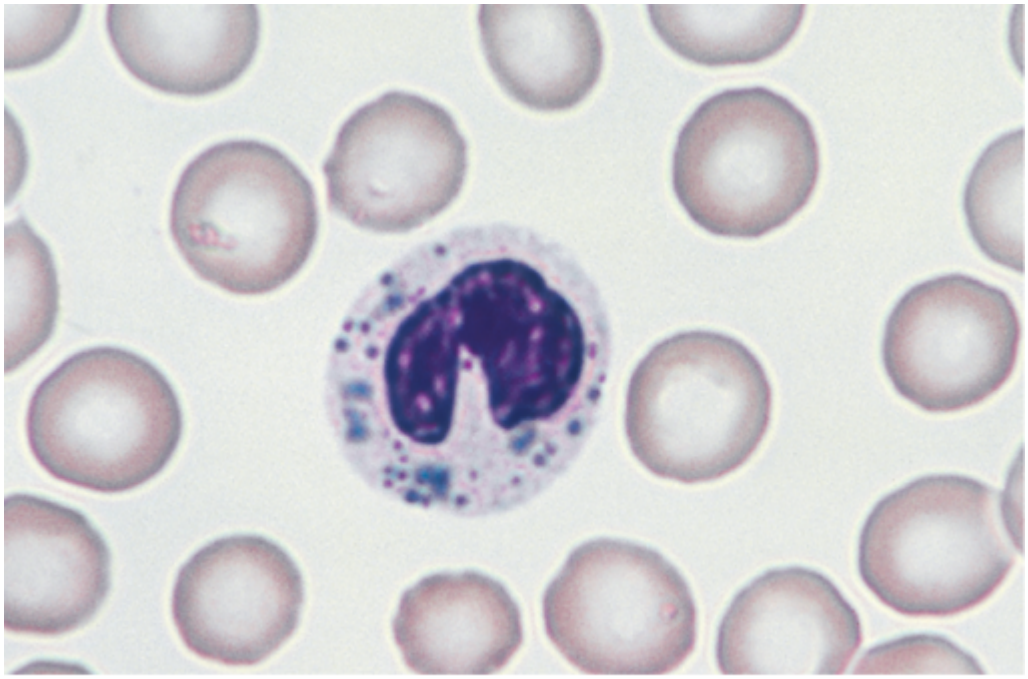
of the cytoplasm of the neutrophil in infections and other toxic states. They represent aggregates of rough endoplasmic reticulum.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 58-30

Chédiak-Higashi disease. Note giant granules in neutrophil.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ACKNOWLEDGMENT

Figures in this chapter were borrowed from Williams Hematology, 7th edition, M Lichtman et al (eds). New York, McGraw-Hill, 2005; Hematology in General Practice, 4th edition, RS Hillman, KA Ault, New York, McGraw-Hill, 2005.

Chapter 59: Anemia and Polycythemia

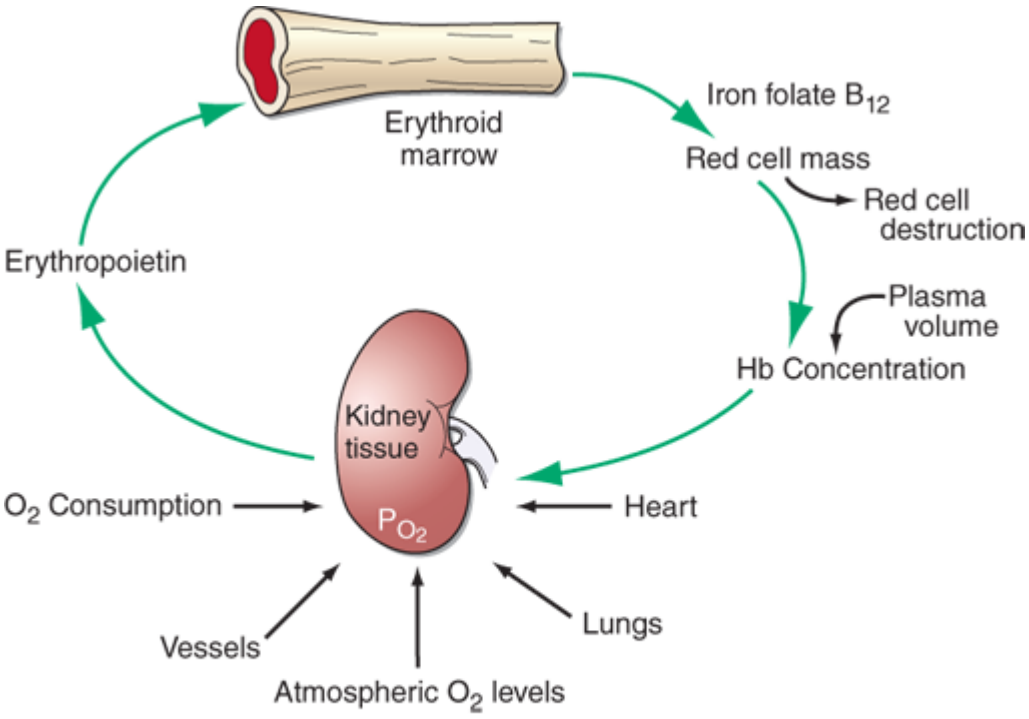
John W. Adamson; Dan L. Longo

HEMATOPOIESIS AND THE PHYSIOLOGIC BASIS OF RED CELL PRODUCTION

Hematopoiesis is the process by which the formed elements of blood are produced. The process is regulated through a series of steps beginning with the hematopoietic stem cell. Stem cells are capable of producing red cells, all classes of granulocytes, monocytes, platelets, and the cells of the immune system. The precise molecular mechanism by which the stem cell becomes committed to a given lineage is not fully defined. However, experiments in mice suggest that erythroid cells come from a common erythroid/megakaryocyte progenitor that does not develop in the absence of expression of the GATA-1 and FOG-1 (friend of GATA-1) transcription factors **(Chap. 92)**. Following lineage commitment, hematopoietic progenitor and precursor cells come increasingly under the regulatory influence of growth factors and hormones. For red cell production, erythropoietin (EPO) is the primary regulatory hormone. EPO is required for the maintenance of committed erythroid progenitor cells that, in the absence of the hormone, undergo programmed cell death (*apoptosis*). The regulated process of red cell production is *erythropoiesis*, and its key elements are illustrated in **Fig. 59-1**.

FIGURE 59-1

The physiologic regulation of red cell production by tissue oxygen tension. Hb, hemoglobin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In the bone marrow, the first morphologically recognizable erythroid precursor is the pronormoblast. This cell can undergo four to five cell divisions, which result in the production of 16–32 mature red cells. With increased EPO production, or the administration of EPO as a drug, early progenitor cell numbers are amplified and, in turn, give rise to increased numbers of erythrocytes. The regulation of EPO production itself is linked to tissue oxygenation.

In mammals, O_2 is transported to tissues bound to the hemoglobin contained within circulating red cells. The mature red cell is 8 μm in diameter, anucleate, discoid in shape, and extremely pliable in order to traverse the microcirculation successfully; its membrane integrity is maintained by the intracellular generation of ATP. Normal red cell production results in the daily replacement of 0.8–1% of all circulating red cells in the body, since the average red cell lives 100–120 days. The organ responsible for red cell production is called the *erythron*. The erythron is a dynamic organ made up of a rapidly proliferating pool of marrow erythroid precursor cells and a large mass of mature circulating red blood cells. The size of the red cell mass reflects the balance of red cell production and destruction. The physiologic basis of red cell production and destruction provides an understanding of the mechanisms that can lead to anemia.

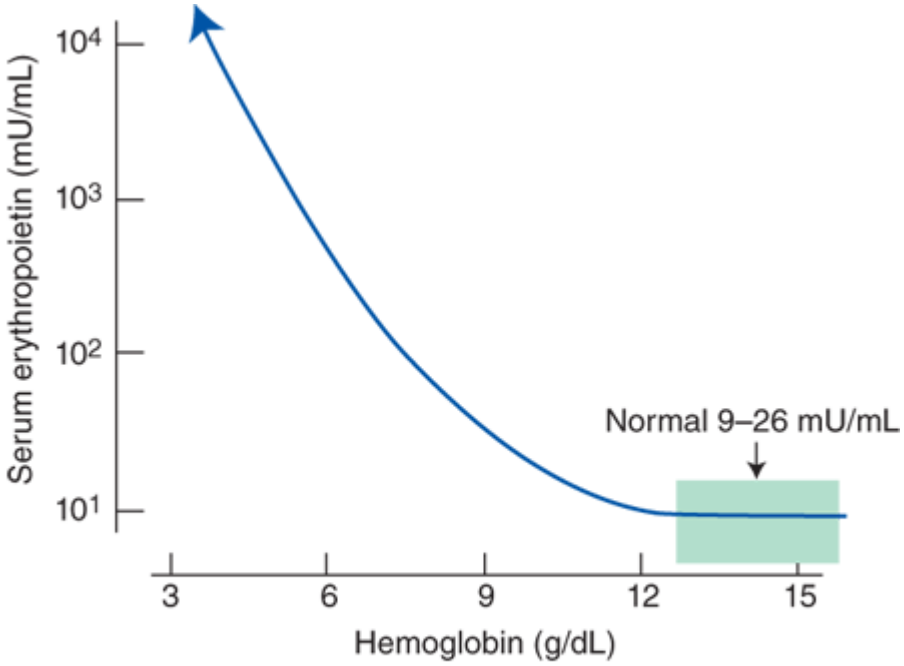
The physiologic regulator of red cell production, the glycoprotein hormone EPO, is produced and released by peritubular capillary lining cells within the kidney. These cells are highly specialized epithelial-like cells. A small amount of EPO is produced by hepatocytes. The fundamental stimulus for EPO production is the availability of O_2 for tissue metabolic needs. Key to EPO gene regulation is hypoxia-inducible factor (HIF)-1α. In the presence of O_2, HIF-1α is hydroxylated at a key proline, allowing HIF-1α to be ubiquitinated and degraded via the proteasome pathway. If O_2 becomes limiting, this critical hydroxylation step does not occur,

allowing HIF-1α to partner with other proteins, translocate to the nucleus, and upregulate the expression of the EPO gene, among others.

Impaired O_2 delivery to the kidney can result from a decreased red cell mass (*anemia*), impaired O_2 loading of the hemoglobin molecule or a high O_2 affinity mutant hemoglobin (*hypoxemia*), or, rarely, impaired blood flow to the kidney (renal artery stenosis). EPO governs the day-to-day production of red cells, and ambient levels of the hormone can be measured in the plasma by sensitive immunoassays—the normal level being 10–25 U/L. When the hemoglobin concentration falls below 100–120 g/L (10–12 g/dL), plasma EPO levels increase in proportion to the severity of the anemia **(Fig. 59-2)**. In circulation, EPO has a half-clearance time of 6–9 h. EPO acts by binding to specific receptors on the surface of marrow erythroid precursors, inducing them to proliferate and to mature. With EPO stimulation, red cell production can increase four- to fivefold within a 1- to 2-week period, but only in the presence of adequate nutrients, especially iron. The functional capacity of the erythron, therefore, requires normal renal production of EPO, a functioning erythroid marrow, and an adequate supply of substrates for hemoglobin synthesis. A defect in any of these key components can lead to anemia. Generally, anemia is recognized in the laboratory when a patient's hemoglobin level or hematocrit is reduced below an expected value (the normal range). The likelihood and severity of anemia are defined based on the deviation of the patient's hemoglobin/hematocrit from values expected for age- and sex-matched normal subjects. The hemoglobin concentration in adults has a Gaussian distribution. The mean hematocrit value for adult males is 47% (standard deviation, ±7%) and that for adult females is 42% (±5%). Any single hematocrit or hemoglobin value carries with it a likelihood of associated anemia. Thus, a hematocrit of <39% in an adult male or <35% in an adult female has only about a 25% chance of being normal. Hematocrit levels are less useful than hemoglobin levels in assessing anemia because they are calculated rather than measured directly. Suspected low hemoglobin or hematocrit values are more easily interpreted if previous values for the same patient are known for comparison. The World Health Organization (WHO) defines anemia as a hemoglobin level <130 g/L (13 g/dL) in men and <120 g/L (12 g/dL) in women.

FIGURE 59-2

Erythropoietin (EPO) levels in response to anemia. When the hemoglobin level falls to 120 g/L (12 g/dL), plasma EPO levels increase logarithmically. In the presence of chronic kidney disease or chronic inflammation, EPO levels are typically lower than expected for the degree of anemia. As individuals age, the level of EPO needed to sustain normal hemoglobin levels appears to increase. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The critical elements of erythropoiesis—EPO production, iron availability, the proliferative capacity of the bone marrow, and effective maturation of red cell precursors—are used for the initial classification of anemia (see below).

ANEMIA

CLINICAL PRESENTATION OF ANEMIA

Signs and Symptoms

Anemia is most often recognized by abnormal screening laboratory tests. Patients less commonly present with advanced anemia and its attendant signs and symptoms. Acute anemia is due to blood loss or hemolysis. If blood loss is mild, enhanced O_2 delivery is achieved through changes in the O_2–hemoglobin dissociation curve mediated by a decreased pH or increased CO_2 (*Bohr effect*). With acute blood loss, hypovolemia dominates the clinical picture, and the hematocrit and hemoglobin levels do not reflect the volume of blood lost. Signs of vascular instability appear with acute losses of 10–15% of the total blood volume. In such patients, the issue is not anemia but hypotension and decreased organ perfusion. When >30% of the blood volume is lost suddenly, patients are unable to compensate with the usual mechanisms of vascular contraction and changes in regional blood flow. The patient prefers to remain supine and will show postural hypotension and tachycardia. If the volume of blood lost is >40% (i.e., >2 L in the average-sized adult), signs of hypovolemic shock including confusion, dyspnea, diaphoresis, hypotension, and tachycardia appear (Chap. 97). Such patients have significant deficits in vital organ perfusion and require immediate volume replacement.

With acute hemolysis, the signs and symptoms depend on the mechanism that leads to red cell destruction. Intravascular hemolysis with release of free hemoglobin may be associated with acute back pain, free hemoglobin in the plasma and urine, and renal failure. Symptoms associated with more chronic or progressive anemia depend on the age of the patient and the adequacy of blood supply to critical organs. Symptoms associated with moderate anemia include fatigue, loss of stamina, breathlessness, and tachycardia (particularly with physical exertion). However, because of the intrinsic compensatory mechanisms that govern the O_2–hemoglobin dissociation curve, the gradual onset of anemia—particularly in young patients—may not be associated with signs or symptoms until the anemia is severe (hemoglobin <70–80 g/L [7–8 g/dL]). When anemia develops over a period of days or weeks, the total blood volume is normal to slightly increased, and changes in cardiac output and regional blood flow help compensate for the overall loss in O_2-carrying capacity. Changes in the position of the O_2–hemoglobin dissociation curve account for some of the compensatory response to anemia. With chronic anemia, intracellular levels of 2,3-bisphosphoglycerate rise, shifting the dissociation curve to the right and facilitating O_2 unloading. This compensatory mechanism can only maintain normal tissue O_2 delivery in the face of a 20–30 g/L (2–3 g/dL) deficit in hemoglobin concentration. Finally, further protection of O_2 delivery to vital organs is achieved by the shunting of blood away from organs that are relatively rich in blood supply, particularly the kidney, gut, and skin.

Certain disorders are commonly associated with anemia. Chronic inflammatory states (e.g., infection, rheumatoid arthritis, cancer) are associated with mild to moderate anemia, whereas lymphoproliferative disorders, such as chronic lymphocytic leukemia and certain other B cell neoplasms, may be associated with autoimmune hemolysis.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Anemia

The evaluation of the patient with anemia requires a careful history and physical examination. Nutritional history related to drugs or alcohol intake and family history of anemia should always be assessed. Certain geographic backgrounds and ethnic origins are associated with an increased likelihood of an inherited disorder of the hemoglobin molecule or intermediary metabolism. Glucose-6-phosphate dehydrogenase (G6PD) deficiency and certain hemoglobinopathies are seen more commonly in those of Middle Eastern or African origin, including African Americans who have a high frequency of G6PD deficiency. Other information that may be useful includes exposure to certain toxic agents or drugs and symptoms related to other disorders commonly associated with anemia. These include symptoms and signs such as bleeding, fatigue, malaise, fever, weight loss, night sweats, and other systemic symptoms. Clues to the mechanisms of anemia may be provided on physical examination by findings of infection, blood in the stool, lymphadenopathy, splenomegaly, or petechiae. Splenomegaly and lymphadenopathy suggest an underlying

lymphoproliferative disease, whereas petechiae suggest platelet dysfunction. Past laboratory measurements are helpful to determine a time of onset.

In the anemic patient, physical examination may demonstrate a forceful heartbeat, strong peripheral pulses, and a systolic “flow” murmur. The skin and mucous membranes may be pale if the hemoglobin is <80–100 g/L (8–10 g/dL). This part of the physical examination should focus on areas where vessels are close to the surface such as the mucous membranes, nail beds, and palmar creases. If the palmar creases are lighter in color than the surrounding skin when the hand is hyperextended, the hemoglobin level is usually <80 g/L (8 g/dL).

LABORATORY EVALUATION

Table 59-1 lists the tests used in the initial workup of anemia. A routine complete blood count (CBC) is required as part of the evaluation and includes the hemoglobin, hematocrit, and red cell indices: the mean cell volume (MCV) in femtoliters, mean cell hemoglobin (MCH) in picograms per cell, and mean concentration of hemoglobin per volume of red cells (MCHC) in grams per liter (non-SI: grams per deciliter). The MCH is the least useful of the indices; it tends to track with the MCV. The red cell indices are calculated as shown in Table 59-2, and the normal variations in the hemoglobin and hematocrit with age are shown in Table 59-3. A number of physiologic factors affect the CBC, including age, sex, pregnancy, smoking, and altitude. High-normal hemoglobin values may be seen in men and women who live at altitude or smoke heavily. Hemoglobin elevations due to smoking reflect normal compensation due to the displacement of O_2 by CO in hemoglobin binding. Other important information is provided by the reticulocyte count and measurements of iron supply including *serum iron*, *total iron-binding capacity* (TIBC; an indirect measure of serum transferrin), and *serum ferritin*. Marked alterations in the red cell indices usually reflect disorders of maturation or iron deficiency. A careful evaluation of the peripheral blood smear is important, and clinical laboratories often provide a description of both the red and white cells, a white cell differential count, and the platelet count. In patients with severe anemia and abnormalities in red blood cell morphology and/or low reticulocyte counts, a bone marrow aspirate or biopsy can assist in the diagnosis. Other tests of value in the diagnosis of specific anemias are discussed in chapters on specific disease states.

The components of the CBC also help in the classification of anemia. *Microcytosis* is reflected by a lower than normal MCV (<80), whereas high values (>100) reflect *macrocytosis*. The MCHC reflect defects in hemoglobin synthesis (*hypochromia*). Automated cell counters describe the red cell volume distribution width (RDW). The MCV (representing the peak of the distribution curve) is insensitive to the appearance of small populations of macrocytes or microcytes. An experienced laboratory technician will be able to identify minor populations of large or small cells or hypochromic cells before the red cell indices change.

Peripheral Blood Smear

The peripheral blood smear provides important information about defects in red cell production (Chap. 58). As a complement to the red cell indices, the blood smear also reveals variations in cell size (*anisocytosis*) and shape (*poikilocytosis*). The degree of anisocytosis usually correlates with increases in the RDW or the range of cell sizes. Poikilocytosis suggests a defect in the maturation of red cell precursors in the bone marrow or

fragmentation of circulating red cells. The blood smear may also reveal *polychromasia*—red cells that are slightly larger than normal and grayish blue in color on the Wright-Giemsa stain. These cells are reticulocytes that have been prematurely released from the bone marrow, and their color represents residual amounts of ribosomal RNA. These cells appear in circulation in response to EPO stimulation or to architectural damage of the bone marrow (fibrosis, infiltration of the marrow by malignant cells, etc.) that results in their disordered release from the marrow. The appearance of nucleated red cells, Howell-Jolly bodies, target cells, sickle cells, and others may provide clues to specific disorders **(Figs. 59-3, 59-4, 59-5, 59-6, 59-7, 59-8, 59-9, 59-10, 59-11)**.

Reticulocyte Count

An accurate reticulocyte count is key to the initial classification of anemia. Reticulocytes are red cells that have been recently released from the bone marrow. They are identified by staining with a supravital dye that precipitates the ribosomal RNA **(Fig. 59-12)**. These precipitates appear as blue or black punctate spots and can be counted manually or, currently, by fluorescent emission of dyes that bind to RNA. This residual RNA is metabolized over the first 24–36 h of the reticulocyte's life span in circulation. Normally, the reticulocyte count ranges from 1 to 2% and reflects the daily replacement of 0.8–1.0% of the circulating red cell population. A corrected reticulocyte percentage or the absolute number of reticulocytes provides a reliable measure of effective red cell production.

In the initial classification of anemia, the patient's reticulocyte count is compared with the expected reticulocyte response. In general, if the EPO and erythroid marrow responses to moderate anemia [hemoglobin <100 g/L (10 g/dL)] are intact, the red cell production rate increases to two to three times normal within 10 days following the onset of anemia. In the face of established anemia, a reticulocyte response less than two to three times normal indicates an inadequate marrow response.

To use the reticulocyte count to estimate marrow response, two corrections are necessary. The first correction adjusts the reticulocyte count based on the reduced number of circulating red cells. With anemia, the percentage of reticulocytes may be increased while the absolute number is unchanged. To correct for this effect, the reticulocyte percentage is multiplied by the ratio of the patient's hemoglobin or hematocrit to the expected hemoglobin/hematocrit for the age and sex of the patient **(Table 59-4)**. This provides an estimate of the reticulocyte count corrected for anemia. To convert the corrected reticulocyte count to an index of marrow production, a further correction is required, depending on whether some of the reticulocytes in circulation have been released from the marrow prematurely. For this second correction, the peripheral blood smear is examined to see if there are polychromatophilic macrocytes present.

These cells, representing prematurely released reticulocytes, are referred to as "shift" cells, and the relationship between the degree of shift and the necessary shift correction factor is shown in **Fig. 59-13**. The correction is necessary because these prematurely released cells survive as reticulocytes in circulation for >1 day, thereby providing a falsely high estimate of daily red cell production. If polychromasia is increased, the reticulocyte count, already corrected for anemia, should be corrected again by 2 to account for the prolonged reticulocyte maturation time. The second correction factor varies from 1 to 3 depending on the severity of anemia. In general, a correction of 2 is simply used. An appropriate correction is shown in Table 59-4. If

polychromatophilic cells are not seen on the blood smear, the second correction is not indicated. The now doubly corrected reticulocyte count is the *reticulocyte production index*, and it provides an estimate of marrow production relative to normal. In many hospital laboratories, the reticulocyte count is reported not only as a percentage but also in absolute numbers. If so, no correction for dilution is required. A summary of the appropriate marrow response to varying degrees of anemia is shown in **Table 59-5**.

Premature release of reticulocytes is normally due to increased EPO stimulation. However, if the integrity of the bone marrow release process is lost through tumor infiltration, fibrosis, or other disorders, the appearance of nucleated red cells or polychromatophilic macrocytes should still invoke the second reticulocyte correction. The shift correction should always be applied to a patient with anemia and a very high reticulocyte count to provide a true index of effective red cell production. Patients with severe chronic hemolytic anemia may increase red cell production as much as six- to sevenfold. This measure alone confirms the fact that the patient has an appropriate EPO response, a normally functioning bone marrow, and sufficient iron available to meet the demands for new red cell formation. If the reticulocyte production index is <2 in the face of established anemia, a defect in erythroid marrow proliferation or maturation must be present.

Tests of Iron Supply and Storage

The laboratory measurements that reflect the availability of iron for hemoglobin synthesis include the serum iron, the TIBC, and the percent transferrin saturation. The percent transferrin saturation is derived by dividing the serum iron level (× 100) by the TIBC. The normal serum iron ranges from 9 to 27 µmol/L (50–150 µg/dL), whereas the normal TIBC is 54–64 µmol/L (300–360 µg/dL); the normal transferrin saturation ranges from 25 to 50%. A diurnal variation in the serum iron leads to a variation in the percent transferrin saturation. The serum ferritin is used to evaluate total body iron stores. Adult males have serum ferritin levels that average ~100 µg/L, corresponding to iron stores of ~1 g. Adult females have lower serum ferritin levels averaging 30 µg/L, reflecting lower iron stores (~300 mg). A serum ferritin level of 10–15 µg/L indicates depletion of body iron stores. However, ferritin is also an acute-phase reactant and, in the presence of acute or chronic inflammation, may rise several-fold above baseline levels. As a rule, a serum ferritin >200 µg/L means there is at least some iron in tissue stores.

Bone Marrow Examination

A bone marrow aspirate and smear or a needle biopsy can be useful in the evaluation of some patients with anemia. In patients with hypoproliferative anemia and normal iron status, a bone marrow is indicated. Marrow examination can diagnose primary marrow disorders such as myelofibrosis, a red cell maturation defect, or an infiltrative disease (**Figs. 59-14**, **59-15**, **59-16**). The increase or decrease of one cell lineage (myeloid vs erythroid) compared to another is obtained by a differential count of nucleated cells in a bone marrow smear (the myeloid/erythroid [M/E] ratio). A patient with a hypoproliferative anemia (see below) and a reticulocyte production index <2 will demonstrate an M/E ratio of 2 or 3:1. In contrast, patients with hemolytic disease and a production index >3 will have an M/E ratio of at least 1:1. Maturation disorders are identified from the discrepancy between the M/E ratio and the reticulocyte production index (see below). Either the marrow smear or biopsy can be stained for the presence of iron stores or iron in developing red

cells. The storage iron is in the form of ferritin or *hemosiderin*. On carefully prepared bone marrow smears, small ferritin granules can normally be seen under oil immersion in 20–40% of developing erythroblasts. Such cells are called *sideroblasts*.

OTHER LABORATORY MEASUREMENTS

Additional laboratory tests may be of value in confirming specific diagnoses. **For details of these tests and how they are applied in individual disorders, see Chaps. 93, 94, 95, 96, 97.**

TABLE 59-1

Laboratory Tests in Anemia Diagnosis

I. Complete blood count (CBC)
- A. Red blood cell count
  1. Hemoglobin
  2. Hematocrit
  3. Reticulocyte count
- B. Red blood cell indices
  1. Mean cell volume (MCV)
  2. Mean cell hemoglobin (MCH)
  3. Mean cell hemoglobin concentration (MCHC)
  4. Red cell distribution width (RDW)
- C. White blood cell count
  1. Cell differential
  2. Nuclear segmentation of neutrophils
- D. Platelet count
- E. Cell morphology
  1. Cell size
  2. Hemoglobin content
  3. Anisocytosis
  4. Poikilocytosis
  5. Polychromasia

II. Iron supply studies
- A. Serum iron
- B. Total iron-binding capacity
- C. Serum ferritin

III. Marrow examination
- A. Aspirate
  1. M/E ratio[a]
  2. Cell morphology
  3. Iron stain
- B. Biopsy
  1. Cellularity
  2. Morphology

[a]M/E ratio, ratio of myeloid to erythroid precursors.

TABLE 59-2

Red Blood Cell Indices

Index	Normal Value
Mean cell volume (MCV) = (hematocrit × 10)/(red cell count × 10^6)	90 ± 8 fL
Mean cell hemoglobin (MCH) = (hemoglobin × 10)/(red cell count × 10^6)	30 ± 3 pg
Mean cell hemoglobin concentration = (hemoglobin × 10)/hematocrit, or MCH/MCV	33 ± 2%

TABLE 59-3

Changes in Normal Hemoglobin/Hematocrit Values with Age, Sex, and Pregnancy

Age/Sex	Hemoglobin, g/dL	Hematocrit, %
At birth	17	52
Childhood	12	36
Adolescence	13	40
Adult man	16 (±2)	47 (±6)
Adult woman (menstruating)	13 (±2)	40 (±6)
Adult woman (postmenopausal)	14 (±2)	42 (±6)
During pregnancy	12 (±2)	37 (±6)

Source: From RS Hillman et al: *Hematology in Clinical Practice*, 5th ed. New York, McGraw-Hill, 2010.

TABLE 59-4

Calculation of Reticulocyte Production Index

Correction #1 for Anemia:
This correction produces the corrected reticulocyte count.
In a person whose reticulocyte count is 9%, hemoglobin 7.5 g/dL, and hematocrit 23%, the absolute reticulocyte count = 9 × (7.5/15) [or × (23/45)] = 4.5%
Note. This correction is not done if the reticulocyte count is reported in absolute numbers (e.g., 50,000/μL of blood)
Correction #2 for Longer Life of Prematurely Released Reticulocytes in the Blood:
This correction produces the reticulocyte production index.
In a person whose reticulocyte count is 9%, hemoglobin 7.5 gm/dL, and hematocrit 23%, the reticulocyte production index $= 9 \times \frac{(7.5/15)(\text{hemoglobin correction})}{2(\text{maturation time correction})} = 2.25$

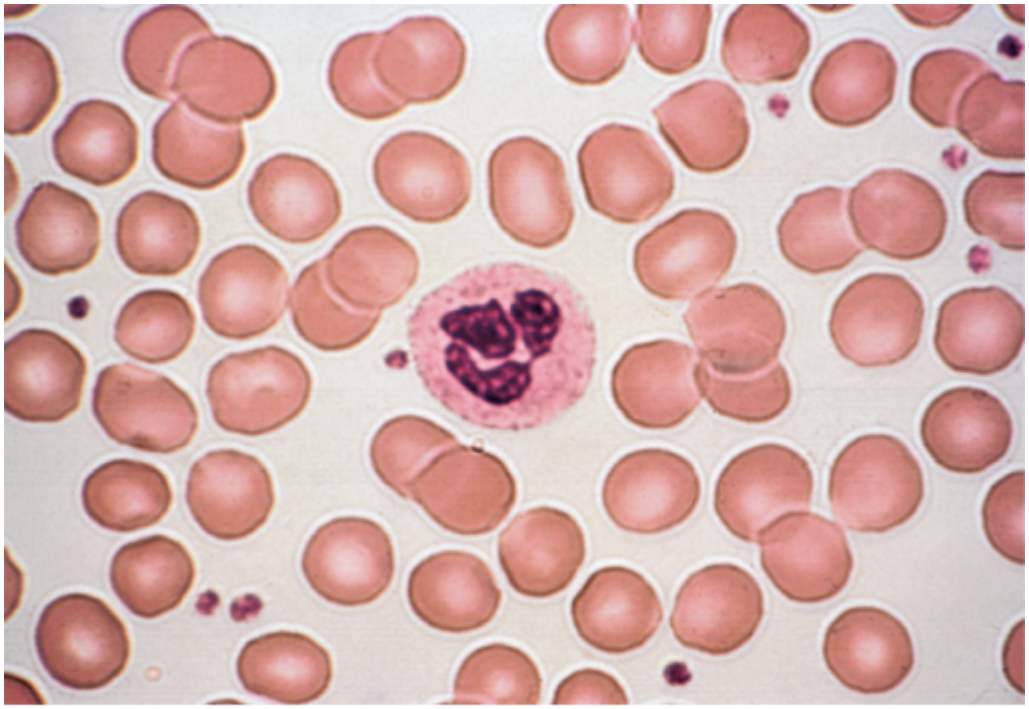
TABLE 59-5

Normal Marrow Response to Anemia

Hemoglobin	Production Index	Reticulocyte Count
15 g/dL	1	50,000/μL
11 g/dL	2.0–2.5	100–150,000/μL
8 g/dL	3.0–4.0	300–400,000/μL

FIGURE 59-3

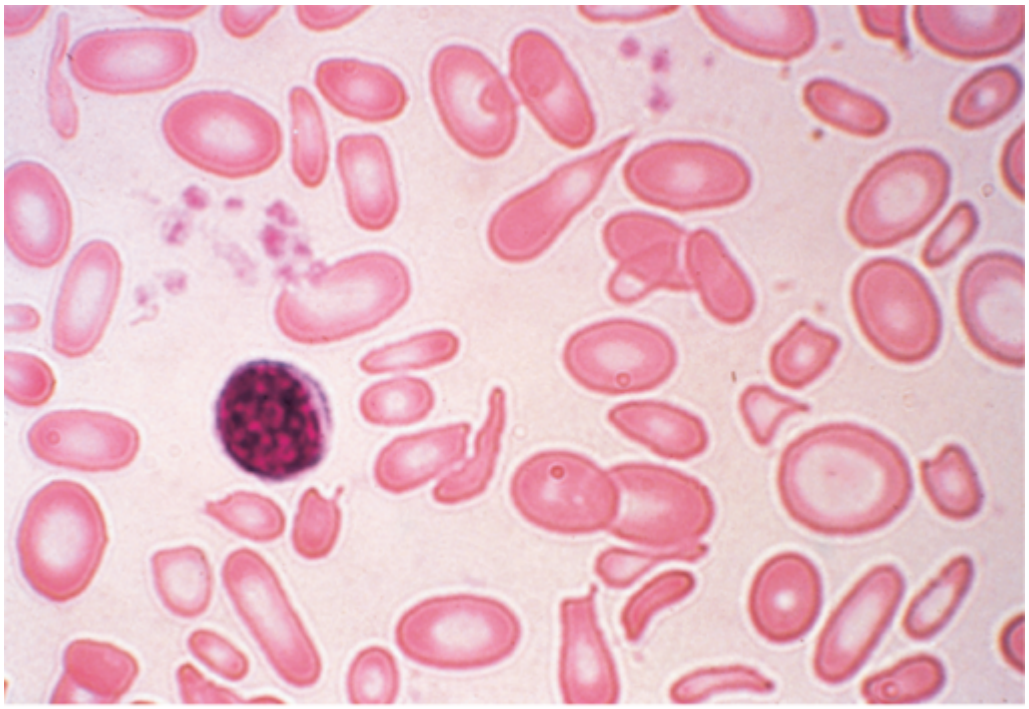
Normal blood smear (Wright stain). High-power field showing normal red cells, a neutrophil, and a few platelets. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-4

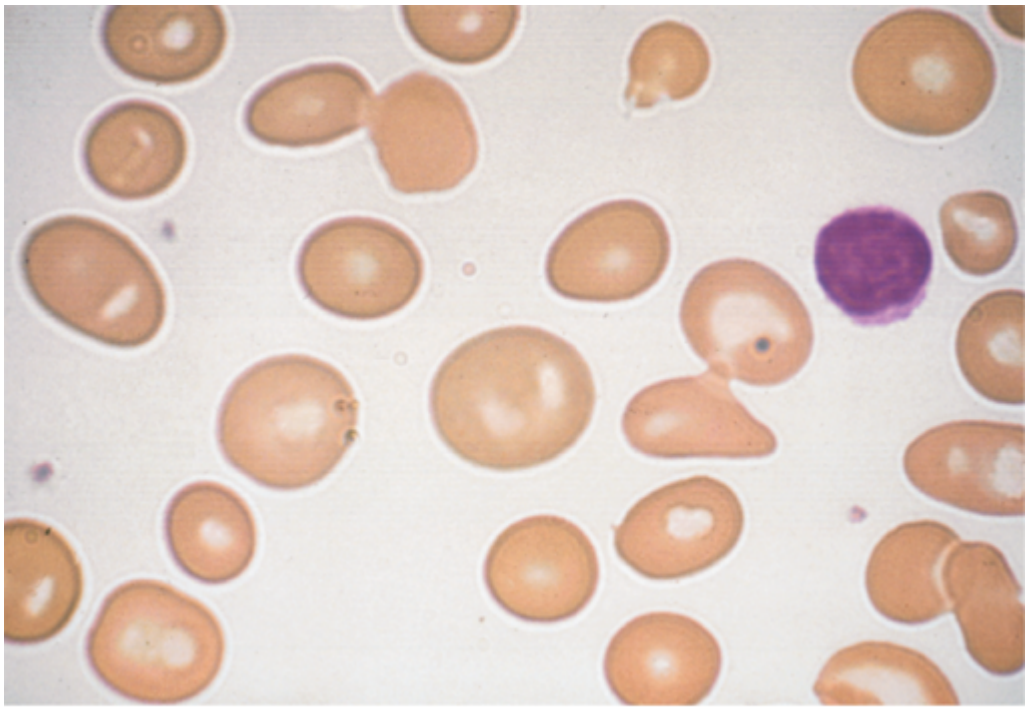
Severe iron-deficiency anemia. Microcytic and hypochromic red cells smaller than the nucleus of a lymphocyte associated with marked variation in size (anisocytosis) and shape (poikilocytosis). *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-5

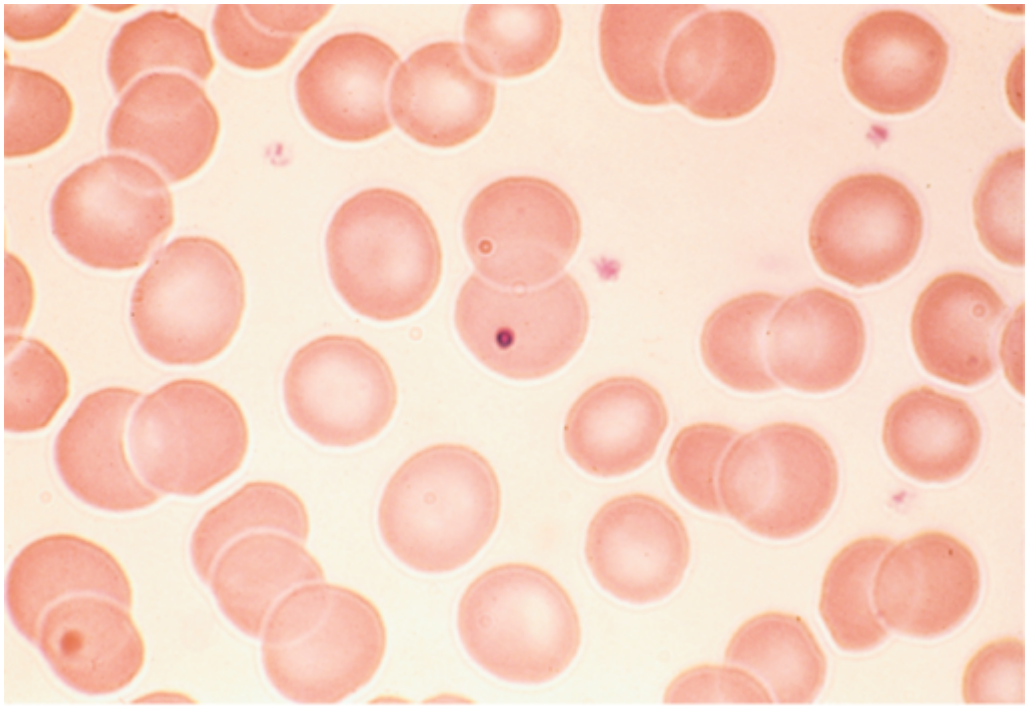
Macrocytosis. Red cells are larger than a small lymphocyte and well hemoglobinized. Often macrocytes are oval shaped (macro-ovalocytes).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-6

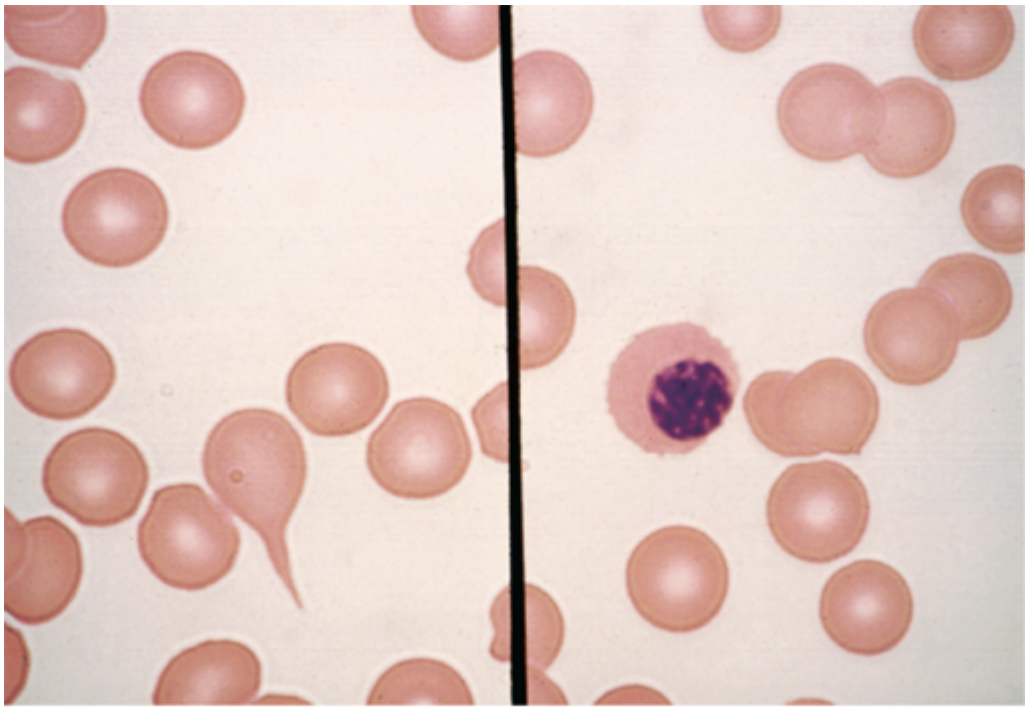
Howell-Jolly bodies. In the absence of a functional spleen, nuclear remnants are not culled from the red cells and remain as small homogeneously staining blue inclusions on Wright stain. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-7

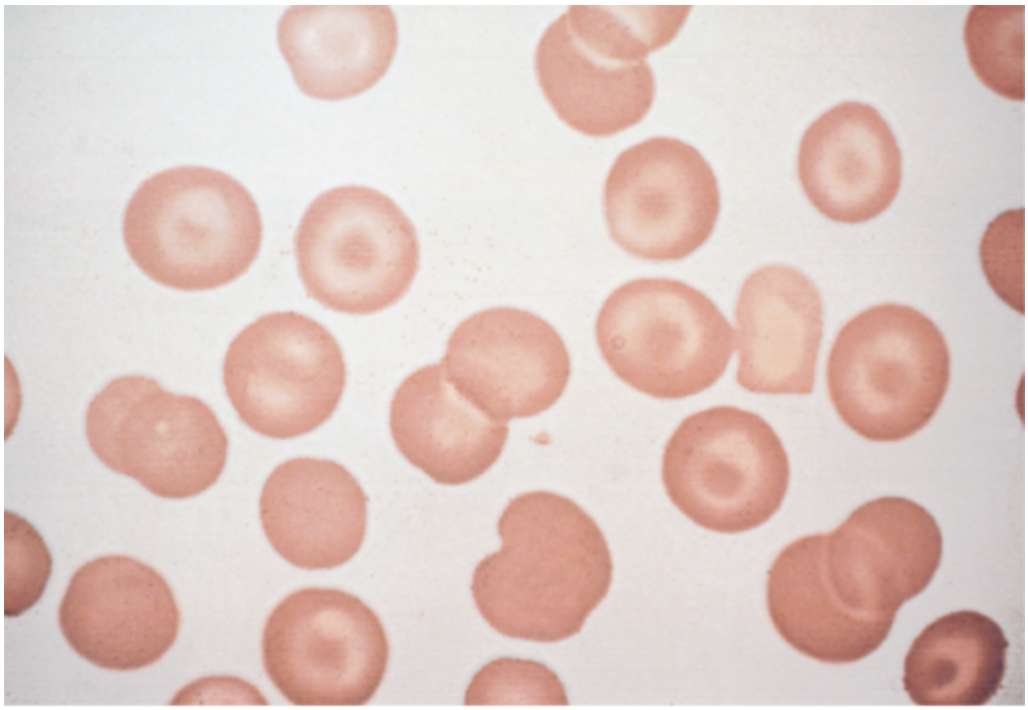
Red cell changes in myelofibrosis. The left panel shows a teardrop-shaped cell. The right panel shows a nucleated red cell. These forms can be seen in myelofibrosis. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-8

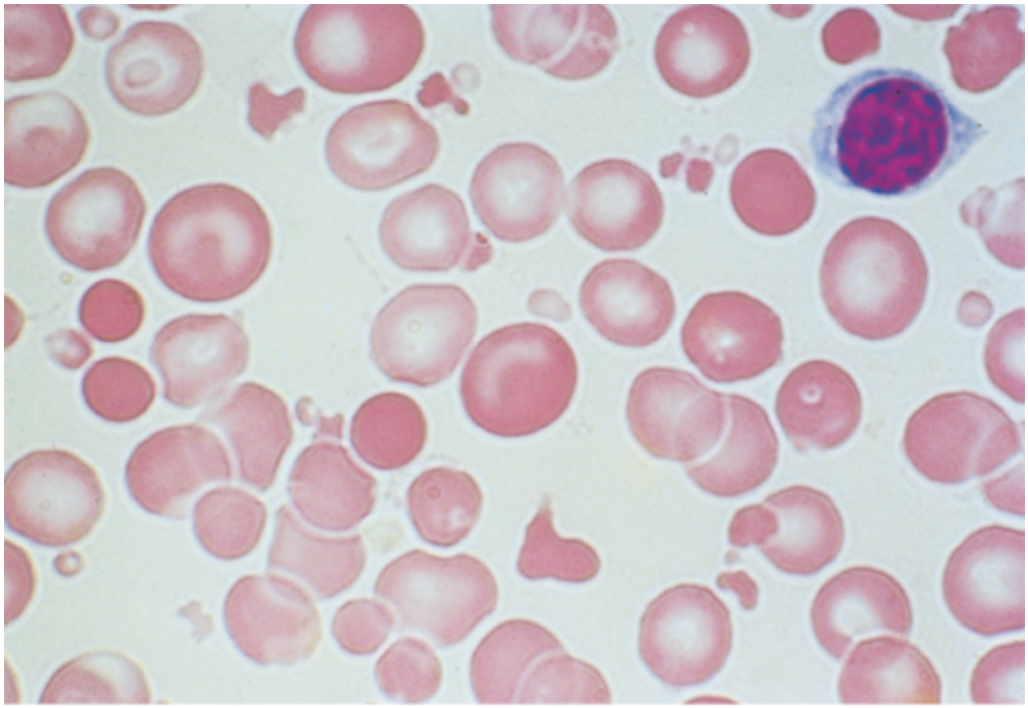
Target cells. Target cells have a bull's-eye appearance and are seen in thalassemia and in liver disease. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-9

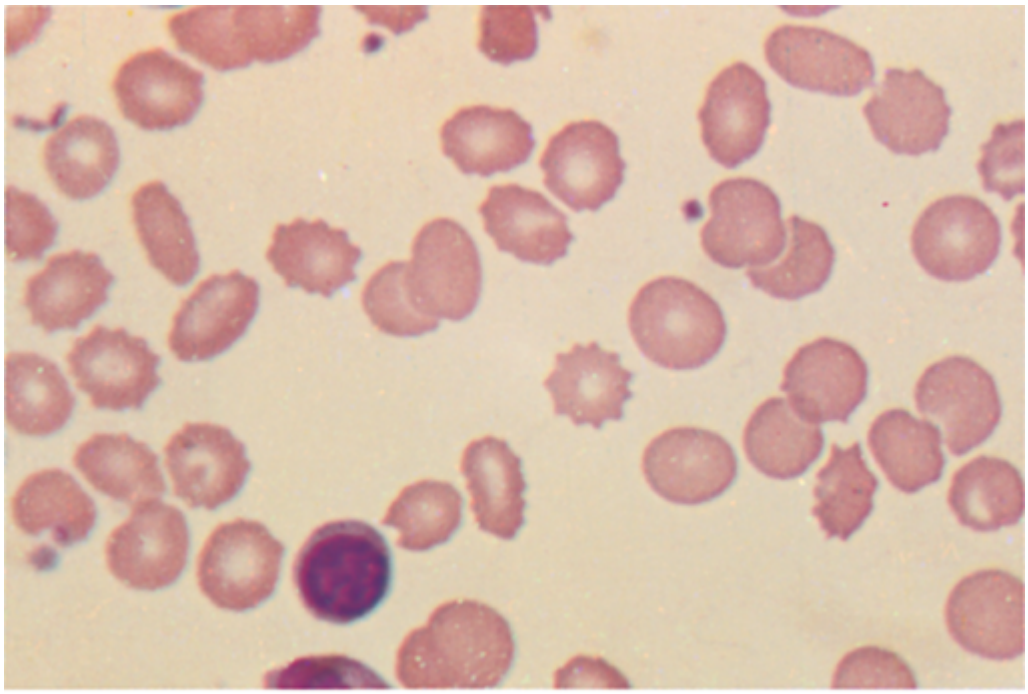
Red cell fragmentation. Red cells may become fragmented in the presence of foreign bodies in the circulation, such as mechanical heart valves, or in the setting of thermal injury. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-10

Uremia. The red cells in uremia may acquire numerous regularly spaced, small, spiny projections. Such cells, called burr cells or echinocytes, are readily distinguishable from irregularly spiculated acanthocytes shown in Fig. 59-11.

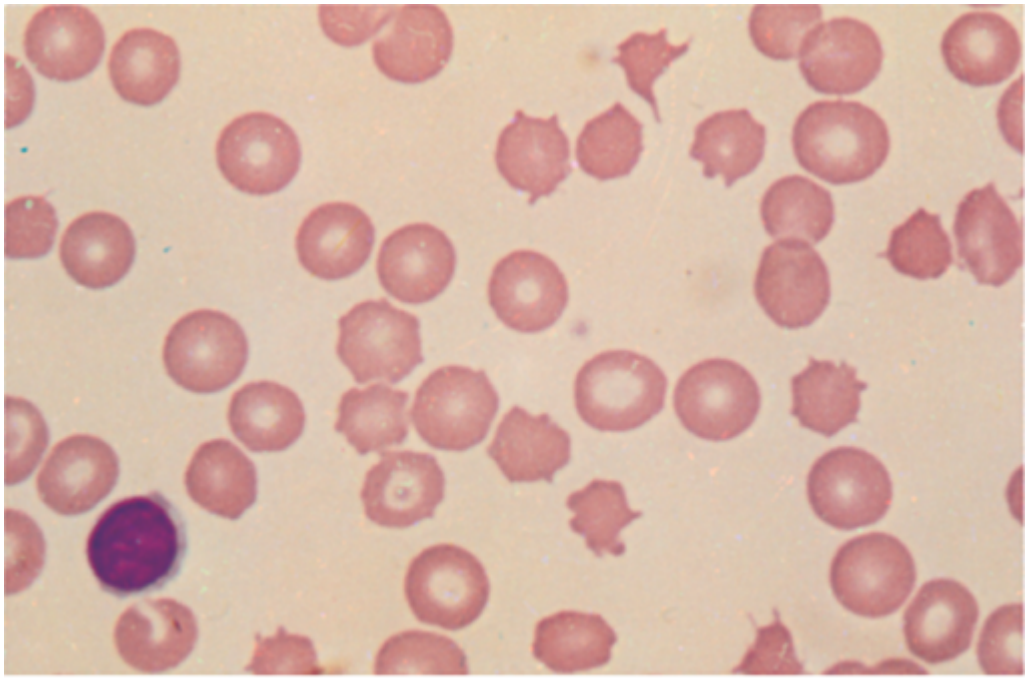


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-11

Spur cells. Spur cells are recognized as distorted red cells containing several irregularly distributed thorn-like projections. Cells with this morphologic abnormality are also called acanthocytes. *(From RS Hillman et al:*

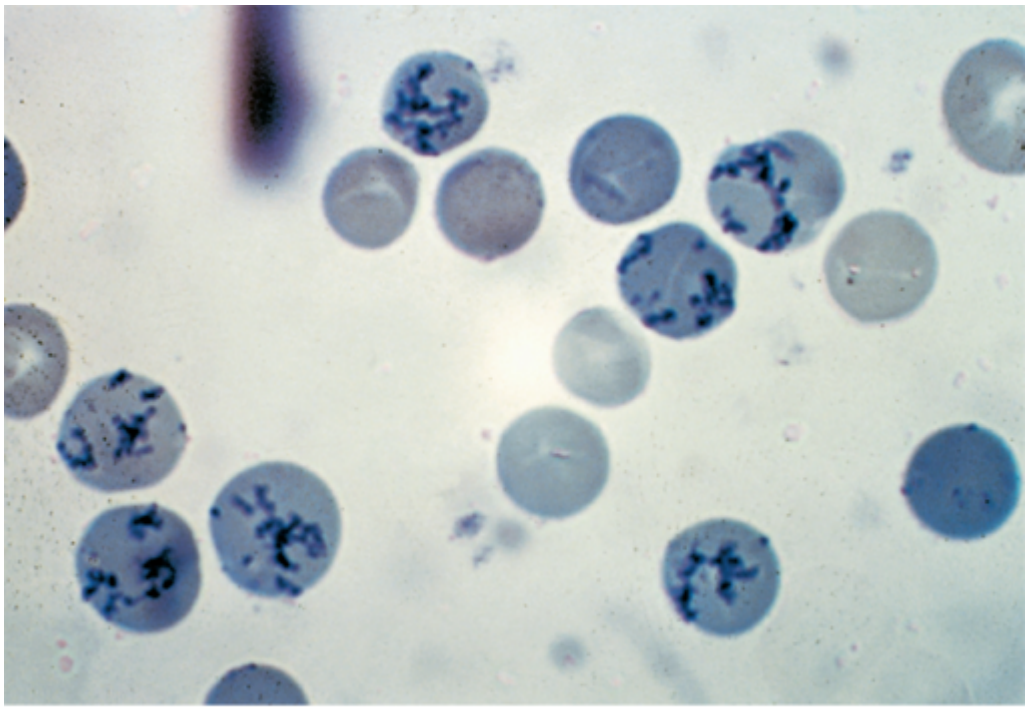
Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-12

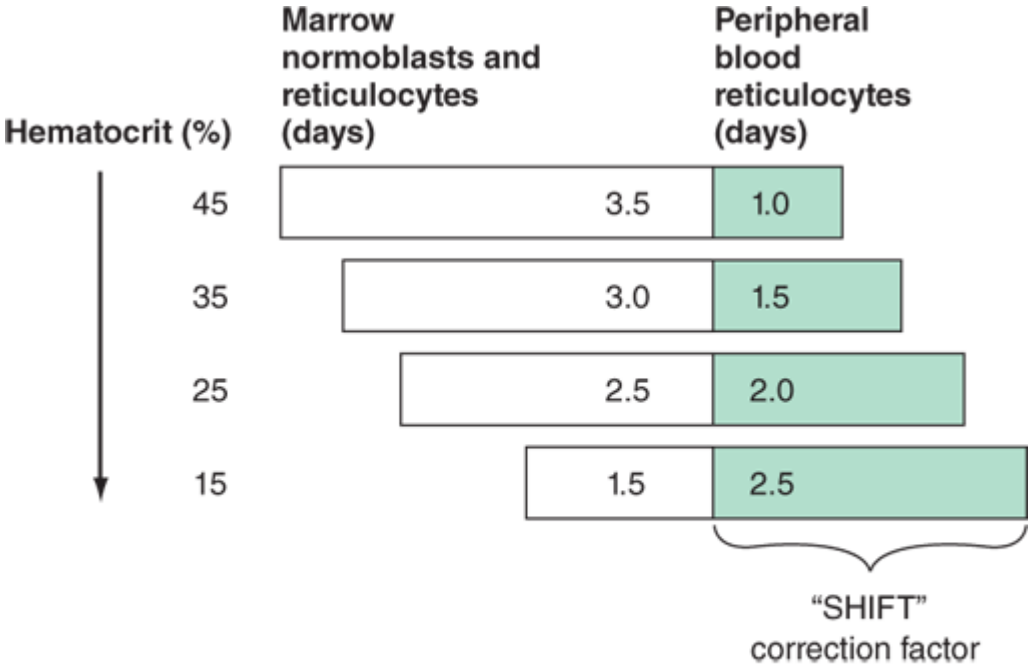
Reticulocytes. Methylene blue stain demonstrates residual RNA in newly made red cells. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-13

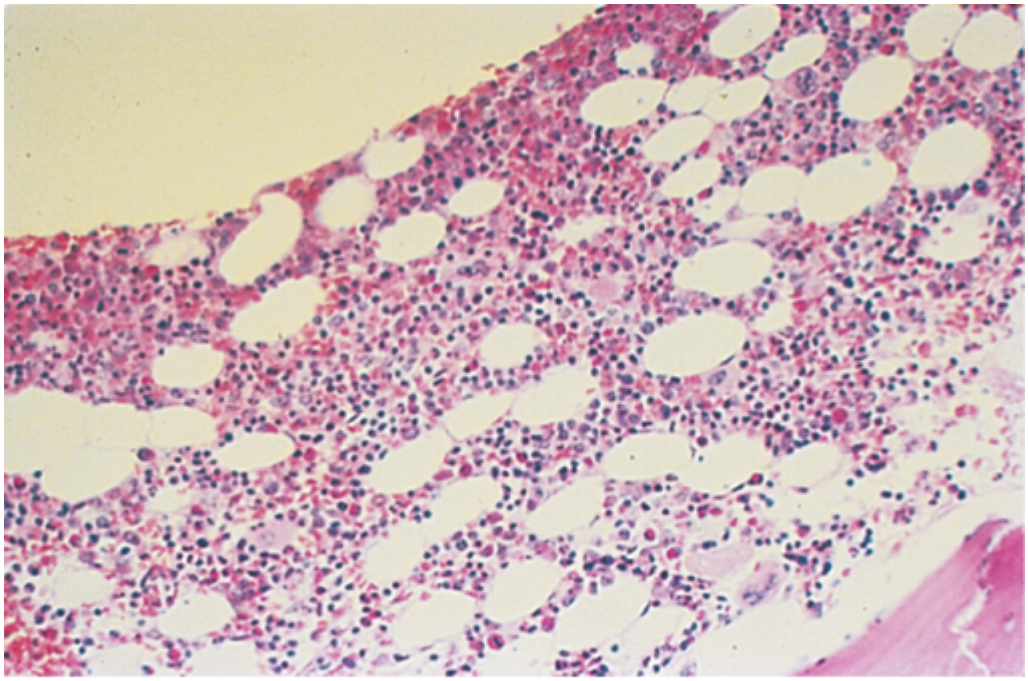
Correction of the reticulocyte count. To use the reticulocyte count as an indicator of effective red cell production, the reticulocyte number must be corrected based on the level of anemia and the circulating life span of the reticulocytes. Erythroid cells take ~4.5 days to mature. At a normal hemoglobin, reticulocytes are released to the circulation with ~1 day left as reticulocytes. However, with different levels of anemia, reticulocytes (and even earlier erythroid cells) may be released from the marrow prematurely. Most patients come to clinical attention with hematocrits in the mid-20s, and thus a correction factor of 2 is commonly used because the observed reticulocytes will live for 2 days in the circulation before losing their RNA.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-14

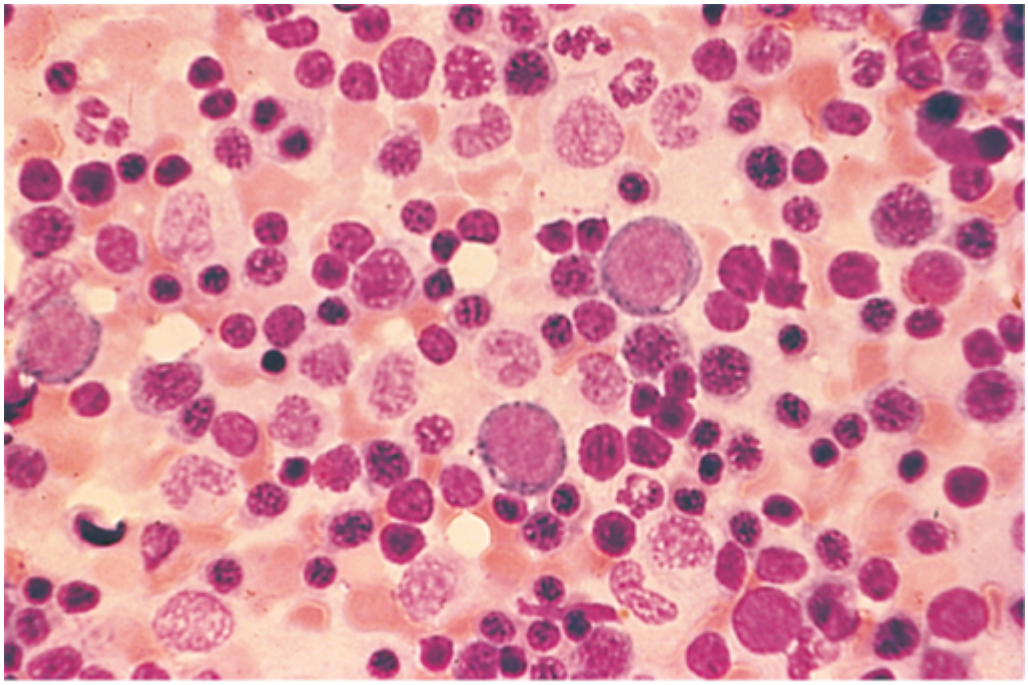
Normal bone marrow. This is a low-power view of a section of a normal bone marrow biopsy stained with hematoxylin and eosin (H&E). Note that the nucleated cellular elements account for ~40–50% and the fat (clear areas) accounts for ~50–60% of the area. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-15

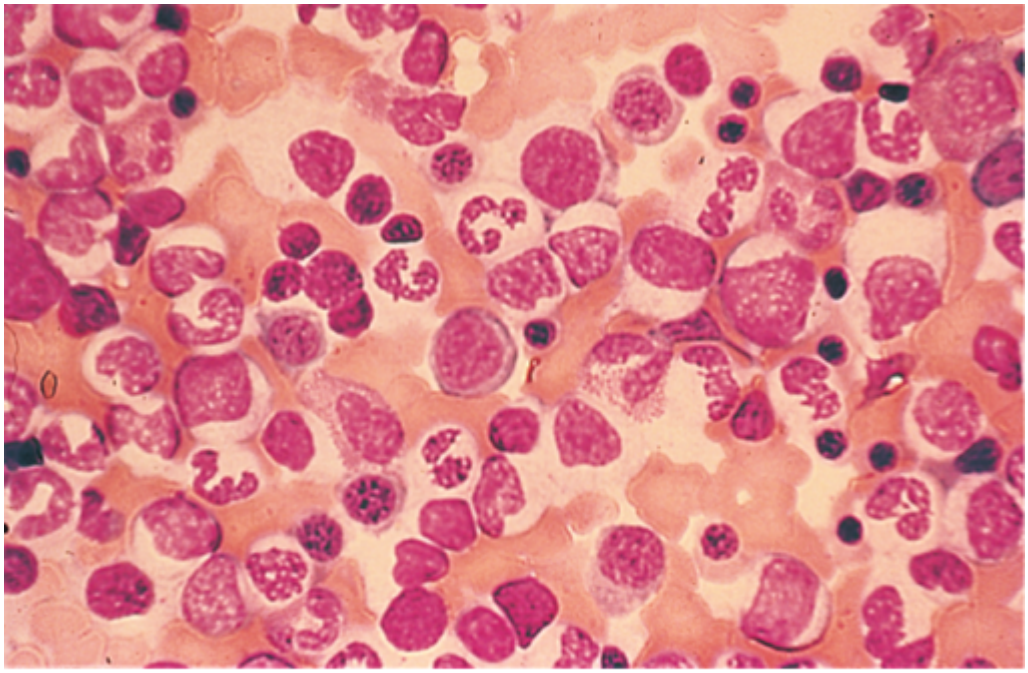
Erythroid hyperplasia. This marrow shows an increase in the fraction of cells in the erythroid lineage as might be seen when a normal marrow compensates for acute blood loss or hemolysis. The myeloid/erythroid (M/E) ratio is about 1:1. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 59-16

Myeloid hyperplasia. This marrow shows an increase in the fraction of cells in the myeloid or granulocytic lineage as might be seen in a normal marrow responding to infection. The myeloid/erythroid (M/E) ratio is >3:1. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

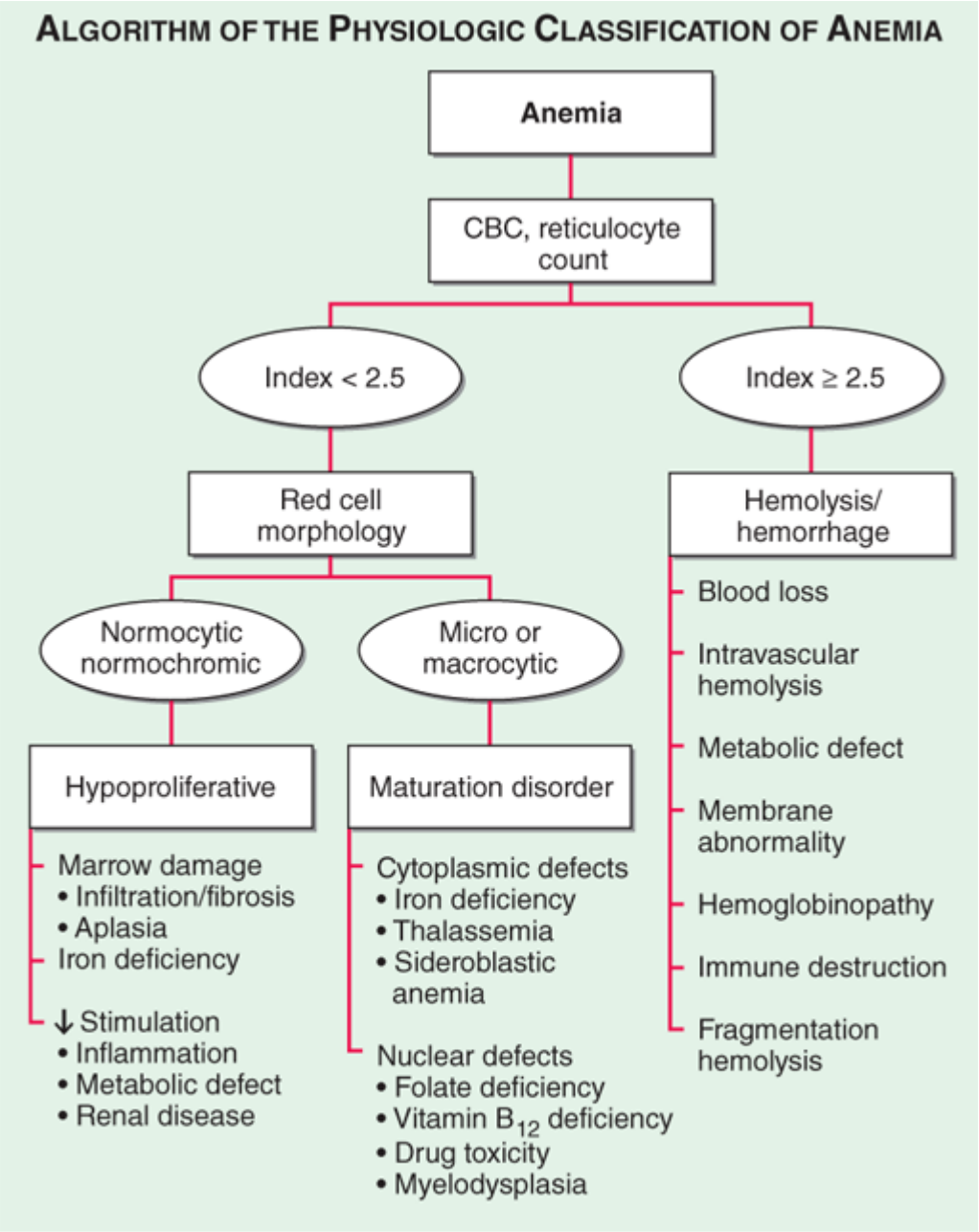
DEFINITION AND CLASSIFICATION OF ANEMIA

Initial Classification of Anemia

The functional classification of anemia has three major categories. These are (1) marrow production defects (*hypoproliferation*), (2) red cell maturation defects (*ineffective erythropoiesis*), and (3) decreased red cell survival (*blood loss/hemolysis*). The classification is shown in **Fig. 59-17**. A hypoproliferative anemia is typically seen with a low reticulocyte production index together with little or no change in red cell morphology (a normocytic, normochromic anemia) **(Chap. 93)**. Maturation disorders typically have a slight to moderately elevated reticulocyte production index that is accompanied by either macrocytic **(Chap. 95)** or microcytic **(Chaps. 93, 94)** red cell indices. Increased red blood cell destruction secondary to hemolysis results in an increase in the reticulocyte production index to at least three times normal **(Chap. 96)**, provided sufficient iron is available. Hemorrhagic anemia does not typically result in production indices of more than 2.0–2.5 times normal because of the limitations placed on expansion of the erythroid marrow by iron availability **(Chap. 97)**.

FIGURE 59-17

The physiologic classification of anemia. CBC, complete blood count.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In the first branch point of the classification of anemia, a reticulocyte production index >2.5 indicates that hemolysis is most likely. A reticulocyte production index <2 indicates either a hypoproliferative anemia or maturation disorder. The latter two possibilities can often be distinguished by the red cell indices, by examination of the peripheral blood smear, or by a marrow examination. If the red cell indices are normal, the anemia is almost certainly hypoproliferative in nature. Maturation disorders are characterized by ineffective red cell production and a low reticulocyte production index. Bizarre red cell shapes—macrocytes or hypochromic microcytes—are seen on the peripheral blood smear. With a hypoproliferative anemia, no erythroid hyperplasia is noted in the marrow, whereas patients with ineffective red cell production have erythroid hyperplasia and an M/E ratio <1:1.

Hypoproliferative Anemias

At least 75% of all cases of anemia are hypoproliferative in nature. A hypoproliferative anemia reflects absolute or relative marrow failure in which the erythroid marrow has not proliferated appropriately for the degree of anemia. The majority of hypoproliferative anemias are due to mild to moderate iron deficiency or inflammation. A hypoproliferative anemia can result from marrow damage, iron deficiency, or inadequate EPO stimulation. The last may reflect impaired renal function, suppression of EPO production by inflammatory cytokines such as interleukin 1, or reduced tissue needs for O_2 from metabolic disease such as hypothyroidism. Only occasionally is the marrow unable to produce red cells at a normal rate, and this is most prevalent in patients with renal failure. With diabetes mellitus or myeloma, the EPO deficiency may be more marked than would be predicted by the degree of renal insufficiency. In general, hypoproliferative anemias are characterized by normocytic, normochromic red cells, although microcytic, hypochromic cells may be observed with mild iron deficiency or long-standing chronic inflammatory disease. The key laboratory tests in distinguishing between the various forms of hypoproliferative anemia include the serum iron and iron-binding capacity, evaluation of renal and thyroid function, a marrow biopsy or aspirate to detect marrow damage or infiltrative disease, and serum ferritin to assess iron stores. An iron stain of the marrow will determine the pattern of iron distribution. Patients with the anemia of acute or chronic inflammation show a distinctive pattern of serum iron (low), TIBC (normal or low), percent transferrin saturation (low), and serum ferritin (normal or high). These changes in iron values are brought about by hepcidin, the iron regulatory hormone that is produced by the liver and is increased in inflammation **(Chap. 93)**. A distinct pattern of results is noted in mild to moderate iron deficiency (low serum iron, high TIBC, low percent transferrin saturation, low serum ferritin) **(Chap. 93)**. Marrow damage by drugs, infiltrative disease such as leukemia or lymphoma, or marrow aplasia is diagnosed from the peripheral blood and bone marrow morphology. With infiltrative disease or fibrosis, a marrow biopsy is required.

Maturation Disorders

The presence of anemia with an inappropriately low reticulocyte production index, macro- or microcytosis on smear, and abnormal red cell indices suggests a maturation disorder. Maturation disorders are divided into two categories: nuclear maturation defects, associated with macrocytosis, and cytoplasmic maturation defects, associated with microcytosis and hypochromia usually from defects in hemoglobin synthesis. The inappropriately low reticulocyte production index is a reflection of the ineffective erythropoiesis that results from the destruction within the marrow of developing erythroblasts. Bone marrow examination shows erythroid hyperplasia.

Nuclear maturation defects result from vitamin B_{12} or folic acid deficiency, drug damage, or myelodysplasia. Drugs that interfere with cellular DNA synthesis, such as methotrexate or alkylating agents, can produce a nuclear maturation defect. Alcohol, alone, is also capable of producing macrocytosis and a variable degree of anemia, but this is usually associated with folic acid deficiency. Measurements of folic acid and vitamin B_{12} are critical not only in identifying the specific vitamin deficiency but also because they reflect different pathogenetic mechanisms **(Chap. 95)**.

Cytoplasmic maturation defects result from severe iron deficiency or abnormalities in globin or heme synthesis. Iron deficiency occupies an unusual position in the classification of anemia. If the iron-deficiency anemia is mild to moderate, erythroid marrow proliferation is blunted and the anemia is classified as hypoproliferative. However, if the anemia is severe and prolonged, the erythroid marrow will become hyperplastic despite the inadequate iron supply, and the anemia will be classified as ineffective erythropoiesis with a cytoplasmic maturation defect. In either case, an inappropriately low reticulocyte production index, microcytosis, and a classic pattern of iron values make the diagnosis clear and easily distinguish iron deficiency from other cytoplasmic maturation defects such as the thalassemias. Defects in heme synthesis, in contrast to globin synthesis, are less common and may be acquired or inherited **(Chap. 409)**. Acquired abnormalities are usually associated with myelodysplasia, may lead to either a macro- or microcytic anemia, and are frequently associated with mitochondrial iron loading. In these cases, iron is taken up by the mitochondria of the developing erythroid cell but not incorporated into heme. The iron-encrusted mitochondria surround the nucleus of the erythroid cell, forming a ring. Based on the distinctive finding of so-called ringed sideroblasts on the marrow iron stain, patients are diagnosed as having a sideroblastic anemia—almost always reflecting myelodysplasia. Again, studies of iron parameters are helpful in the differential diagnosis of these patients.

Blood Loss/Hemolytic Anemia

In contrast to anemias associated with an inappropriately low reticulocyte production index, hemolysis is associated with red cell production indices ≥2.5 times normal. The stimulated erythropoiesis is reflected in the blood smear by the appearance of increased numbers of polychromatophilic macrocytes. A marrow examination is rarely indicated if the reticulocyte production index is increased appropriately. The red cell indices are typically normocytic or slightly macrocytic, reflecting the increased number of reticulocytes. Acute blood loss is not associated with an increased reticulocyte production index because of the time required to increase EPO production and, subsequently, marrow proliferation **(Chap. 97)**. Subacute blood loss may be associated with modest reticulocytosis. Anemia from chronic blood loss presents more often as iron deficiency than with the picture of increased red cell production.

The evaluation of blood loss anemia is usually not difficult. Most problems arise when a patient presents with an increased red cell production index from an episode of acute blood loss that went unrecognized. The cause of the anemia and increased red cell production may not be obvious. The confirmation of a recovering state may require observations over a period of 2–3 weeks, during which the hemoglobin concentration will rise and the reticulocyte production index fall **(Chap. 97)**.

Hemolytic disease, while dramatic, is among the least common forms of anemia. The ability to sustain a high reticulocyte production index reflects the ability of the erythroid marrow to compensate for hemolysis and, in the case of extravascular hemolysis, the efficient recycling of iron from the destroyed red cells to support red cell production. With intravascular hemolysis, such as paroxysmal nocturnal hemoglobinuria, the loss of iron may limit the marrow response. The level of response depends on the severity of the anemia and the nature of the underlying disease process.

Hemoglobinopathies, such as sickle cell disease and the thalassemias, present a mixed picture. The reticulocyte index may be high but is inappropriately low for the degree of marrow erythroid hyperplasia (Chap. 94).

Hemolytic anemias present in different ways. Some appear suddenly as an acute, self-limited episode of intravascular or extravascular hemolysis, a presentation pattern often seen in patients with autoimmune hemolysis or with inherited defects of the Embden-Meyerhof pathway or the glutathione reductase pathway. Patients with inherited disorders of the hemoglobin molecule or red cell membrane generally have a lifelong clinical history typical of the disease process. Those with chronic hemolytic disease, such as hereditary spherocytosis, may actually present not with anemia but with a complication stemming from the prolonged increase in red cell destruction such as symptomatic bilirubin gallstones or splenomegaly. Patients with chronic hemolysis are also susceptible to aplastic crises if an infectious process interrupts red cell production.

The differential diagnosis of an acute or chronic hemolytic event requires the careful integration of family history, the pattern of clinical presentation, and—whether the disease is congenital or acquired—careful examination of the peripheral blood smear. Precise diagnosis may require more specialized laboratory tests, such as hemoglobin electrophoresis or a screen for red cell enzymes. Acquired defects in red cell survival are often immunologically mediated and require a direct or indirect antiglobulin test or a cold agglutinin titer to detect the presence of hemolytic antibodies or complement-mediated red cell destruction (Chap. 96).

TREATMENT

TREATMENT

Anemia

An overriding principle is to initiate treatment of mild to moderate anemia only when a specific diagnosis is made. Rarely, in the acute setting, anemia may be so severe that red cell transfusions are required before a specific diagnosis is available. Whether the anemia is of acute or gradual onset, the selection of the appropriate treatment is determined by the documented cause(s) of the anemia. Often, the cause of the anemia is multifactorial. For example, a patient with severe rheumatoid arthritis who has been taking anti-inflammatory drugs may have a hypoproliferative anemia associated with chronic inflammation as well as chronic blood loss associated with intermittent gastrointestinal bleeding. In every circumstance, it is important to evaluate the patient's iron status fully before and during the treatment of any anemia. **Transfusion is discussed in Chap. 109; iron therapy is discussed in Chap. 93; treatment of megaloblastic anemia is discussed in Chap. 95; treatment of other entities is discussed in their respective chapters (sickle cell anemia, Chap. 94; hemolytic anemias, Chap. 96; aplastic anemia and myelodysplasia, Chap. 98).**

Therapeutic options for the treatment of anemias have expanded dramatically during the past 30 years. Blood component therapy is available and safe. Recombinant EPO as an adjunct to anemia management has transformed the lives of patients with chronic renal failure on dialysis and reduced transfusion needs of anemic cancer patients receiving chemotherapy. Eventually, patients with inherited disorders of globin

synthesis or mutations in the globin gene, such as sickle cell disease, may benefit from the successful introduction of targeted genetic therapy (Chap. 458).

POLYCYTHEMIA

Polycythemia is defined as an increase in the hemoglobin above normal. This increase may be real or only apparent because of a decrease in plasma volume (spurious or relative polycythemia). The term *erythrocytosis* may be used interchangeably with polycythemia, but some draw a distinction between them: erythrocytosis implies documentation of increased red cell mass, whereas polycythemia refers to any increase in red cells. Often patients with polycythemia are detected through an incidental finding of elevated hemoglobin or hematocrit levels. Concern that the hemoglobin level may be abnormally high is usually triggered at 170 g/L (17 g/dL) for men and 150 g/L (15 g/dL) for women. Hematocrit levels >50% in men or >45% in women may be abnormal. Hematocrits >60% in men and >55% in women are almost invariably associated with an increased red cell mass. Given that the machine that quantitates red cell parameters actually measures hemoglobin concentrations and calculates hematocrits, hemoglobin levels may be a better index.

Features of the clinical history that are useful in the differential diagnosis include smoking history; current living at high altitude; or a history of diuretic use, congenital heart disease, sleep apnea, or chronic lung disease.

Patients with polycythemia may be asymptomatic or experience symptoms related to the increased red cell mass or the underlying disease process that leads to the increased red cell mass. The dominant symptoms from an increased red cell mass are related to hyperviscosity and thrombosis (both venous and arterial), because the blood viscosity increases logarithmically at hematocrits >55%. Manifestations include neurologic symptoms such as vertigo, tinnitus, headache, and visual disturbances. Hypertension is often present. Patients with *polycythemia vera* may have aquagenic pruritus, symptoms related to hepatosplenomegaly, easy bruising, epistaxis, or bleeding from the gastrointestinal tract. Peptic ulcer disease is common. Such patients also may present with digital ischemia, Budd–Chiari syndrome, hepatic or splenic/mesenteric vein thrombosis. Patients with hypoxemia may develop cyanosis on minimal exertion or have headache, impaired mental acuity, and fatigue.

The physical examination usually reveals a ruddy complexion. Splenomegaly favors polycythemia vera as the diagnosis (Chap. 99). The presence of cyanosis or evidence of a right-to-left shunt suggests congenital heart disease presenting in the adult, particularly tetralogy of Fallot or Eisenmenger's syndrome (Chap. 264). Increased blood viscosity raises pulmonary artery pressure; hypoxemia can lead to increased pulmonary vascular resistance. Together, these factors can produce cor pulmonale.

Polycythemia can be spurious (related to a decrease in plasma volume; Gaisbock's syndrome), primary, or secondary in origin. The secondary causes are all mediated by EPO: either a physiologically adapted appropriate level based on tissue hypoxia (lung disease, high altitude, CO poisoning, high-affinity hemoglobinopathy) or an abnormal overproduction (renal cysts, renal artery stenosis, tumors with ectopic

EPO production). A rare familial form of polycythemia is associated with normal EPO levels but hyperresponsive EPO receptors due to mutations.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

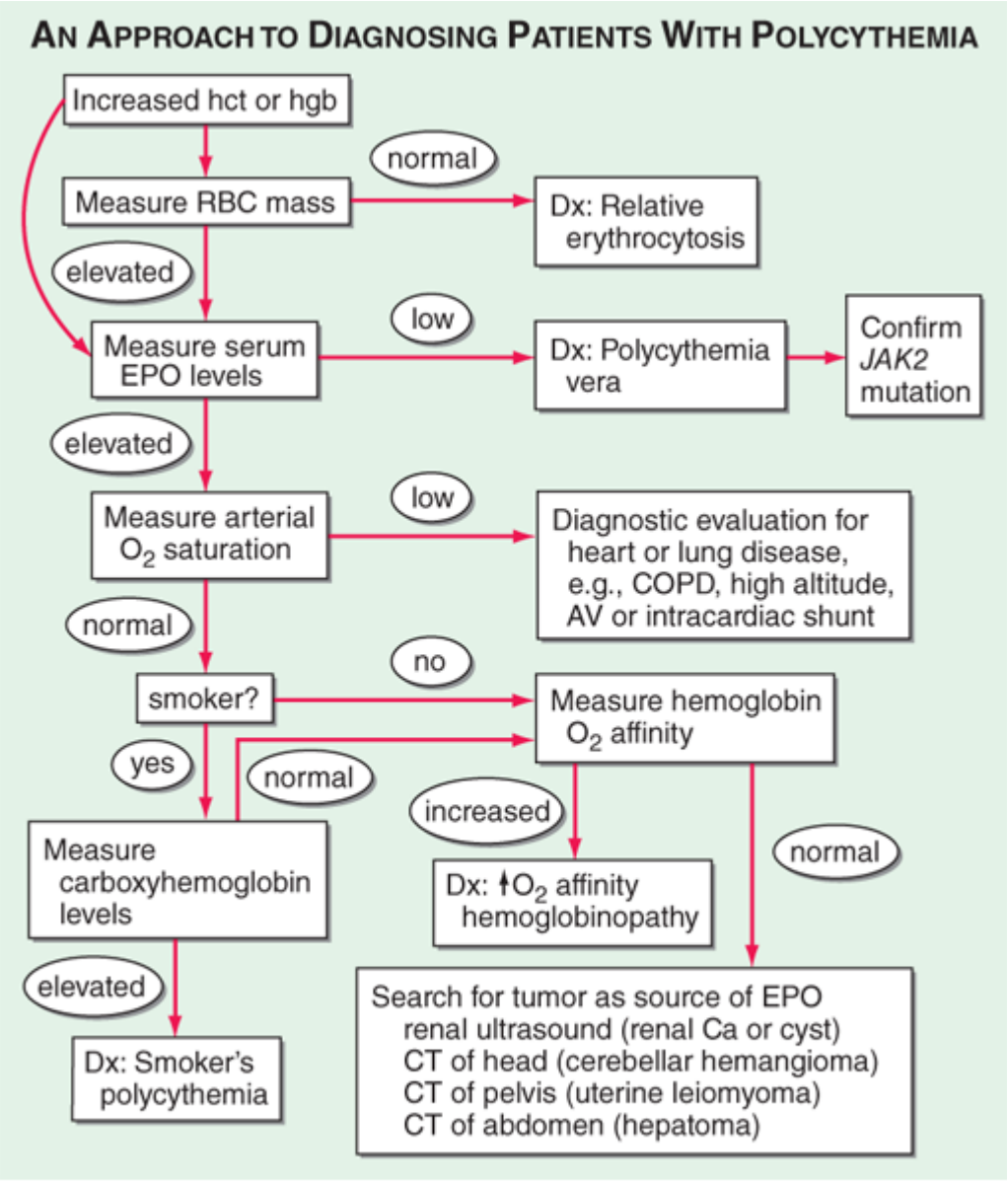
Polycythemia

As shown in **Fig. 59-18**, the first step is to document the presence of an increased red cell mass using the principle of isotope dilution by administering ^{51}Cr-labeled autologous red blood cells to the patient and sampling blood radioactivity over a 2-h period. If the red cell mass is normal (<36 mL/kg in men, <32 mL/kg in women), the patient has spurious or relative polycythemia. If the red cell mass is increased (>36 mL/kg in men, >32 mL/kg in women), serum EPO levels should be measured. If EPO levels are low or unmeasurable, the patient most likely has polycythemia vera. A mutation in *JAK2* (Val617Phe), a key member of the cytokine intracellular signaling pathway, can be found in 90–95% of patients with polycythemia vera. Many of those without this particular *JAK2* mutation have mutations in exon 12. As a practical matter, few centers assess red cell mass in the setting of an increased hemoglobin level. The alternative workup is to measure EPO levels, check for *JAK2* mutation(s), and perform an abdominal ultrasound to assess spleen size. Tests that support the diagnosis of polycythemia vera include elevated white blood cell count, increased absolute basophil count, and thrombocytosis.

If serum EPO levels are elevated, one needs to distinguish whether the elevation is a physiologic response to hypoxia or related to autonomous EPO production. Patients with low arterial O_2 saturation (<92%) should be further evaluated for the presence of heart or lung disease, if they are not living at high altitude. Patients with normal O_2 saturation who are smokers may have elevated EPO levels because of CO displacement of O_2. If carboxyhemoglobin (COHb) levels are high, the diagnosis is “smoker’s polycythemia.” Such patients should be urged to stop smoking. Those who cannot stop smoking require phlebotomy to control their polycythemia. Patients with normal O_2 saturation who do not smoke either have an abnormal hemoglobin that does not deliver O_2 to the tissues (evaluated by finding elevated O_2–hemoglobin affinity) or have a source of EPO production that is not responding to the normal feedback inhibition. Further workup is dictated by the differential diagnosis of EPO-producing neoplasms. Hepatoma, uterine leiomyoma, and renal cancer or cysts are all detectable with abdominopelvic computed tomography scans. Cerebellar hemangiomas may produce EPO, but they present with localizing neurologic signs and symptoms rather than polycythemia-related symptoms.

FIGURE 59-18

An approach to the differential diagnosis of patients with an elevated hemoglobin (possible polycythemia). AV, atrioventricular; COPD, chronic obstructive pulmonary disease; CT, computed tomography; EPO, erythropoietin; hct, hematocrit; hgb, hemoglobin; IVP, intravenous pyelogram; RBC, red blood cell.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Hillman RS et al: *Hematology in Clinical Practice*, 5th ed. New York, McGraw-Hill, 2010.

McMullin MF et al: Guidelines for the diagnosis, investigation and management of polycythaemia/erythrocytosis. Br J Haematol 130:174, 2005. [PubMed: 16029446]

Sankaran VG, Weiss MJ: Anemia: Progress in molecular mechanisms and therapies. Nat Med 21:221, 2015. [PubMed: 25742458]

Chapter 60: Disorders of Granulocytes and Monocytes

Steven M. Holland; John I. Gallin

INTRODUCTION

Leukocytes, the major cells comprising inflammatory and immune responses, include neutrophils, T and B lymphocytes, natural killer (NK) cells, monocytes, eosinophils, and basophils. These cells have specific functions, such as antibody production by B lymphocytes or destruction of bacteria by neutrophils, but in no single infectious disease is the exact role of the cell types completely established. Thus, whereas neutrophils are classically thought to be critical to host defense against bacteria, they may also play important roles in defense against viral infections.

The blood delivers leukocytes to the various tissues from the bone marrow, where they are produced. Normal blood leukocyte counts are 4.3–10.8 × 10^9/L, with neutrophils representing 45–74% of the cells, bands 0–4%, lymphocytes 16–45%, monocytes 4–10%, eosinophils 0–7%, and basophils 0–2%. Variation among individuals and among different ethnic groups can be substantial, with lower leukocyte numbers for certain African-American ethnic groups. The various leukocytes are derived from a common stem cell in the bone marrow. Three-fourths of the nucleated cells of bone marrow are committed to the production of leukocytes. Leukocyte maturation in the marrow is under the regulatory control of a number of different factors, known as colony-stimulating factors (CSFs) and interleukins (ILs). Because an alteration in the number and type of leukocytes is often associated with disease processes, total white blood cell (WBC) count (cells per μL) and differential counts are informative. This chapter focuses on neutrophils, monocytes, and eosinophils. **Lymphocytes and basophils are discussed in Chaps. 342 and 346, respectively.**

NEUTROPHILS

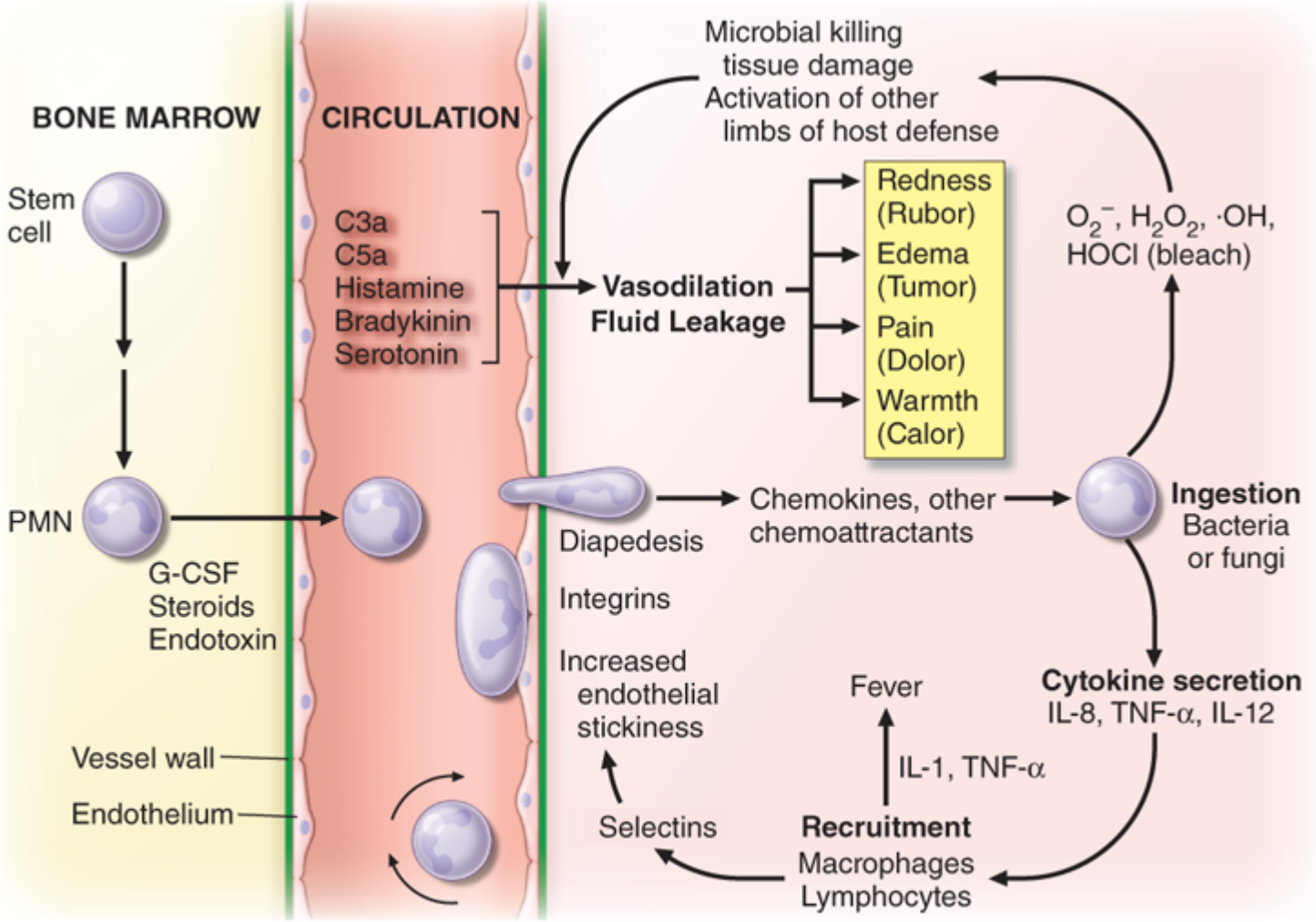
MATURATION

Important events in neutrophil life are summarized in **Fig. 60-1**. In normal humans, neutrophils are produced only in the bone marrow. The minimum number of stem cells necessary to support hematopoiesis is estimated to be 400–500 at any one time. Human blood monocytes, tissue macrophages, and stromal cells produce CSFs, hormones required for the growth of monocytes and neutrophils in the bone marrow. The hematopoietic system not only produces enough neutrophils (~1.3 × 10^{11} cells per 80-kg person per day) to carry out physiologic functions but also has a large reserve stored in the marrow, which can be mobilized in response to inflammation or infection. An increase in the number of blood neutrophils is called *neutrophilia*,

and the presence of immature cells is termed a *shift to the left*. A decrease in the number of blood neutrophils is called *neutropenia*.

FIGURE 60-1

Schematic events in neutrophil production, recruitment, and inflammation. The four cardinal signs of inflammation (rubor, tumor, calor, dolor) are indicated, as are the interactions of neutrophils with other cells and cytokines. G-CSF, granulocyte colony-stimulating factor; IL, interleukin; PMN, polymorphonuclear leukocyte; TNF-α, tumor necrosis factor α.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Neutrophils and monocytes evolve from pluripotent stem cells under the influence of cytokines and CSFs (Fig. 60-2). The proliferation phase through the metamyelocyte takes about 1 week, while the maturation phase from metamyelocyte to mature neutrophil takes another week. The myeloblast is the first recognizable precursor cell and is followed by the *promyelocyte*. The promyelocyte evolves when the classic lysosomal granules, called the *primary*, or *azurophil, granules* are produced. The primary granules contain hydrolases, elastase, myeloperoxidase, cathepsin G, cationic proteins, and bactericidal/permeability-increasing protein, which is important for killing gram-negative bacteria. Azurophil granules also contain *defensins*, a family of cysteine-rich polypeptides with broad antimicrobial activity against bacteria, fungi and certain enveloped viruses. The promyelocyte divides to produce the *myelocyte*, a cell responsible for the

synthesis of the *specific*, or *secondary, granules*, which contain unique (specific) constituents such as lactoferrin, vitamin B_{12}–binding protein, membrane components of the reduced nicotinamide-adenine dinucleotide phosphate (NADPH) oxidase required for hydrogen peroxide production, histaminase, and receptors for certain chemoattractants and adherence-promoting factors (CR3) as well as receptors for the basement membrane component, laminin. The secondary granules do not contain acid hydrolases and therefore are not classic lysosomes. Packaging of secondary granule contents during myelopoiesis is controlled by CCAAT/enhancer binding protein-ε. Secondary granule contents are readily released extracellularly, and their mobilization is important in modulating inflammation. During the final stages of maturation, no cell division occurs, and the cell passes through the metamyelocyte stage and then to the band neutrophil with a sausage-shaped nucleus **(Fig. 60-3)**. As the band cell matures, the nucleus assumes a lobulated configuration. The nucleus of neutrophils normally contains up to four segments **(Fig. 60-4)**. Excessive segmentation (>5 nuclear lobes) may be a manifestation of folate or vitamin B_{12} deficiency or the congenital neutropenia syndrome of warts, hypogammaglobulinemia, infections, and myelokathexis (WHIM) described below. The Pelger-Hüet anomaly **(Fig. 60-5)**, an infrequent dominant benign inherited trait, results in neutrophils with distinctive bilobed nuclei that must be distinguished from band forms. Acquired bilobed nuclei, pseudo Pelger-Hüet anomaly, can occur with acute infections or in myelodysplastic syndromes. The physiologic role of the normal multilobed nucleus of neutrophils is unknown, but it may allow great deformation of neutrophils during migration into tissues at sites of inflammation.

FIGURE 60-2

Stages of neutrophil development shown schematically. Granulocyte colony-stimulating factor (G-CSF) and granulocyte-macrophage colony-stimulating factor (GM-CSF) are critical to this process. Identifying cellular characteristics and specific cell-surface markers are listed for each maturational stage.

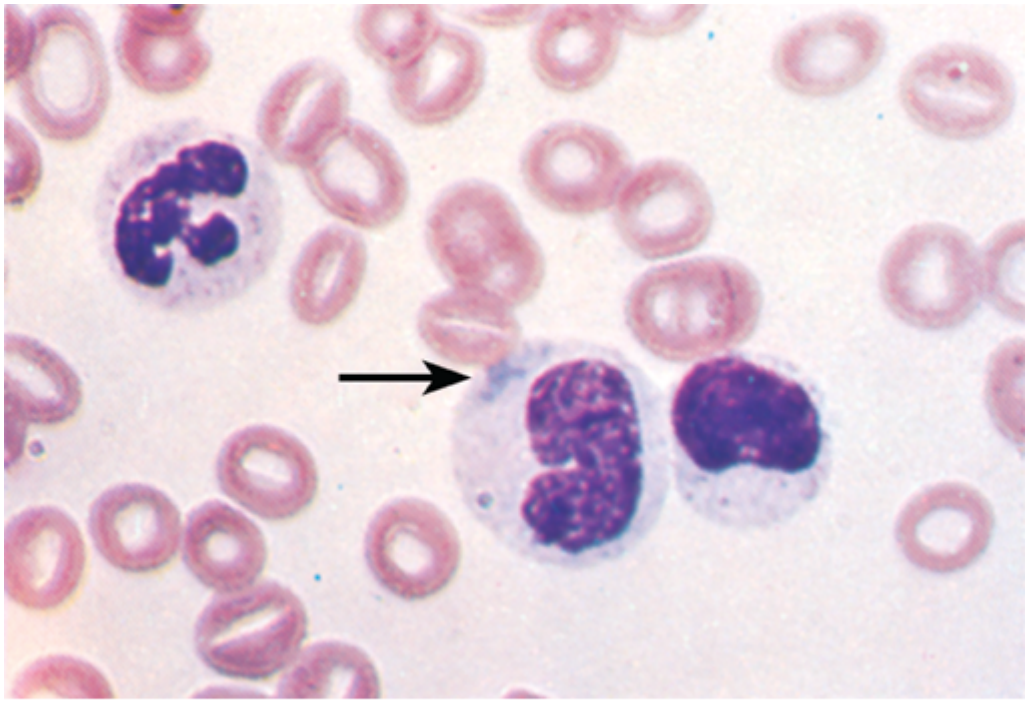
Cell	Stage	Surface Markers[a]	Characteristics
	MYELOBLAST	CD33, CD13, CD15	Prominent nucleoli
	PROMYELOCYTE	CD33, CD13, CD15	Large cell Primary granules appear
	MYELOCYTE	CD33, CD13, CD15, CD14, CD11b	Secondary granules appear
	METAMYELOCYTE	CD33, CD13, CD15, CD14, CD11b	Kidney bean–shaped nucleus
	BAND FORM	CD33, CD13, CD15, CD14, CD11b CD10, CD16	Condensed, band–shaped nucleus
	NEUTROPHIL	CD33, CD13, CD15, CD14, CD11b CD10, CD16	Condensed, multilobed nucleus

[a]CD = Cluster Determinant; ● Nucleolus; ● Primary granule; • Secondary granule.

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 60-3

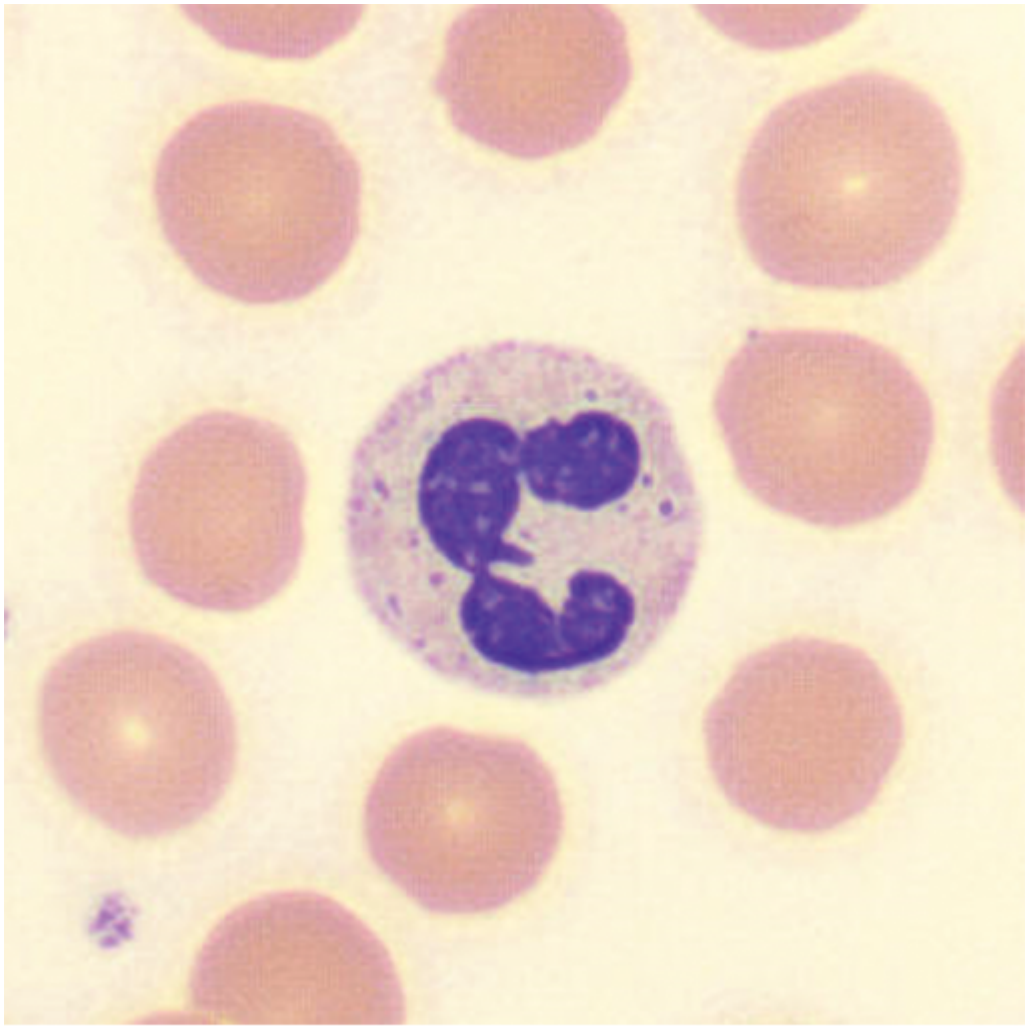
Neutrophil band with Döhle body. The neutrophil with a sausage-shaped nucleus in the center of the field is a band form. Döhle bodies are discrete, blue-staining, nongranular areas found in the periphery of the cytoplasm of the neutrophil in infections and other toxic states. They represent aggregates of rough endoplasmic reticulum.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

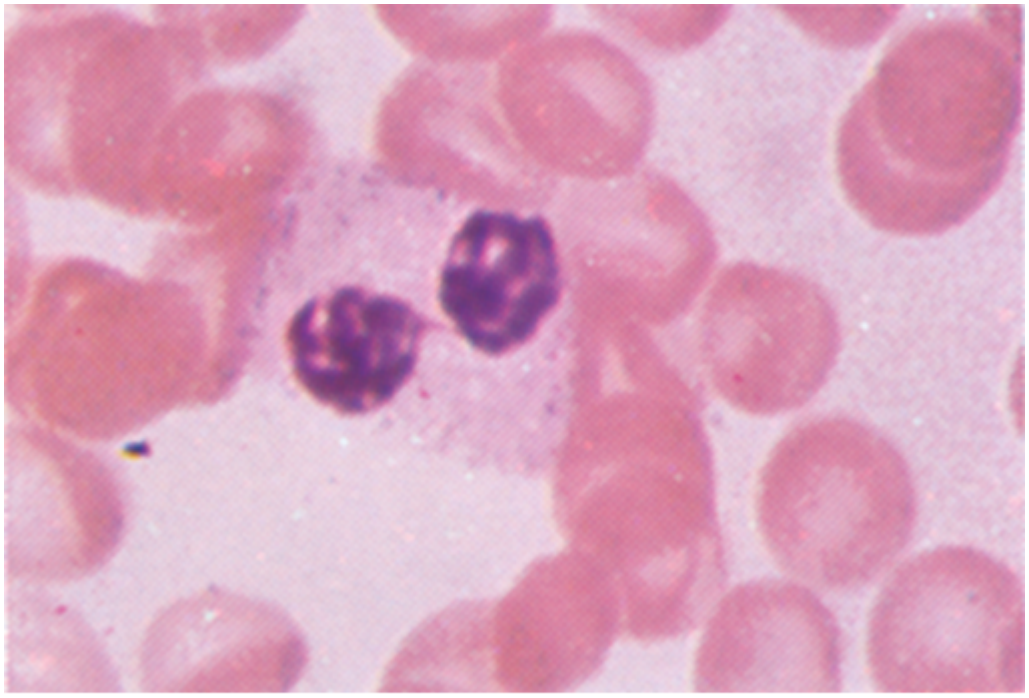
FIGURE 60-4

Normal granulocyte. The normal granulocyte has a segmented nucleus with heavy, clumped chromatin; fine neutrophilic granules are dispersed throughout the cytoplasm.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.
FIGURE 60-5

Pelger-Hüet anomaly. In this benign disorder, the majority of granulocytes are bilobed. The nucleus frequently has a spectacle-like, or "pince-nez," configuration.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

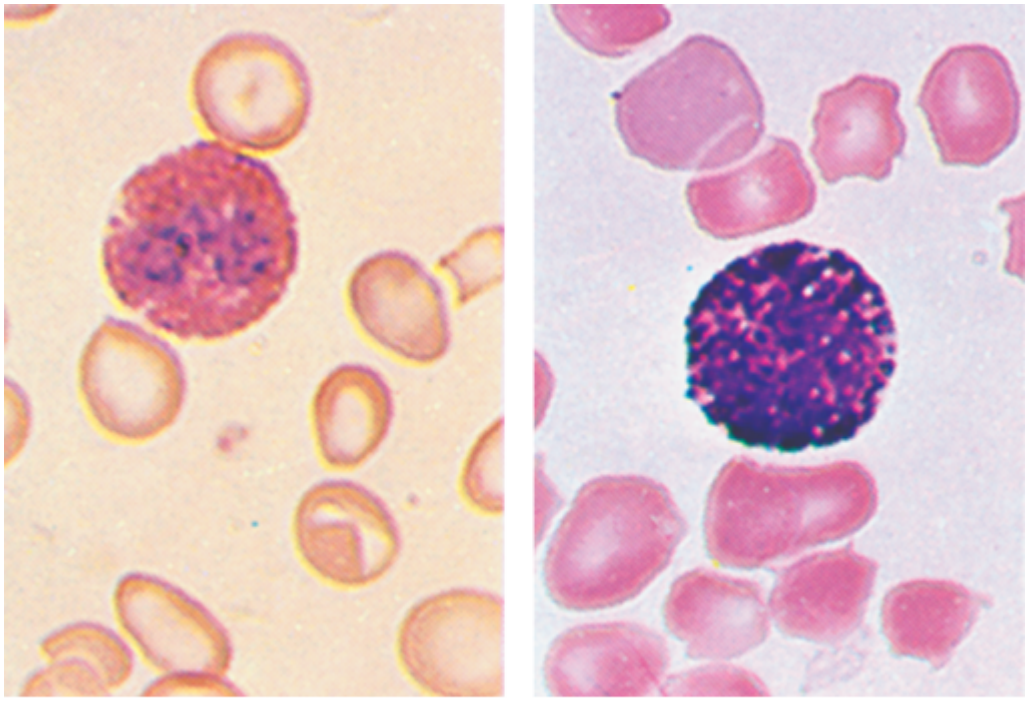
In severe acute bacterial infection, prominent neutrophil cytoplasmic granules, called *toxic granulations*, are occasionally seen. Toxic granulations are immature or abnormally staining azurophil granules. Cytoplasmic inclusions, also called *Döhle bodies* (Fig. 60-3), can be seen during infection and are fragments of ribosome-rich endoplasmic reticulum. Large neutrophil vacuoles are often present in acute bacterial infection and probably represent pinocytosed (internalized) membrane.

Neutrophils are heterogeneous in function. Monoclonal antibodies have been developed that recognize only a subset of mature neutrophils. The meaning of neutrophil heterogeneity is not known.

The morphology of eosinophils and basophils is shown in **Fig. 60-6**.

FIGURE 60-6

Normal eosinophil (*left*) and basophil (*right*). The eosinophil contains large, bright orange granules and usually a bilobed nucleus. The basophil contains large purple-black granules that fill the cell and obscure the nucleus.



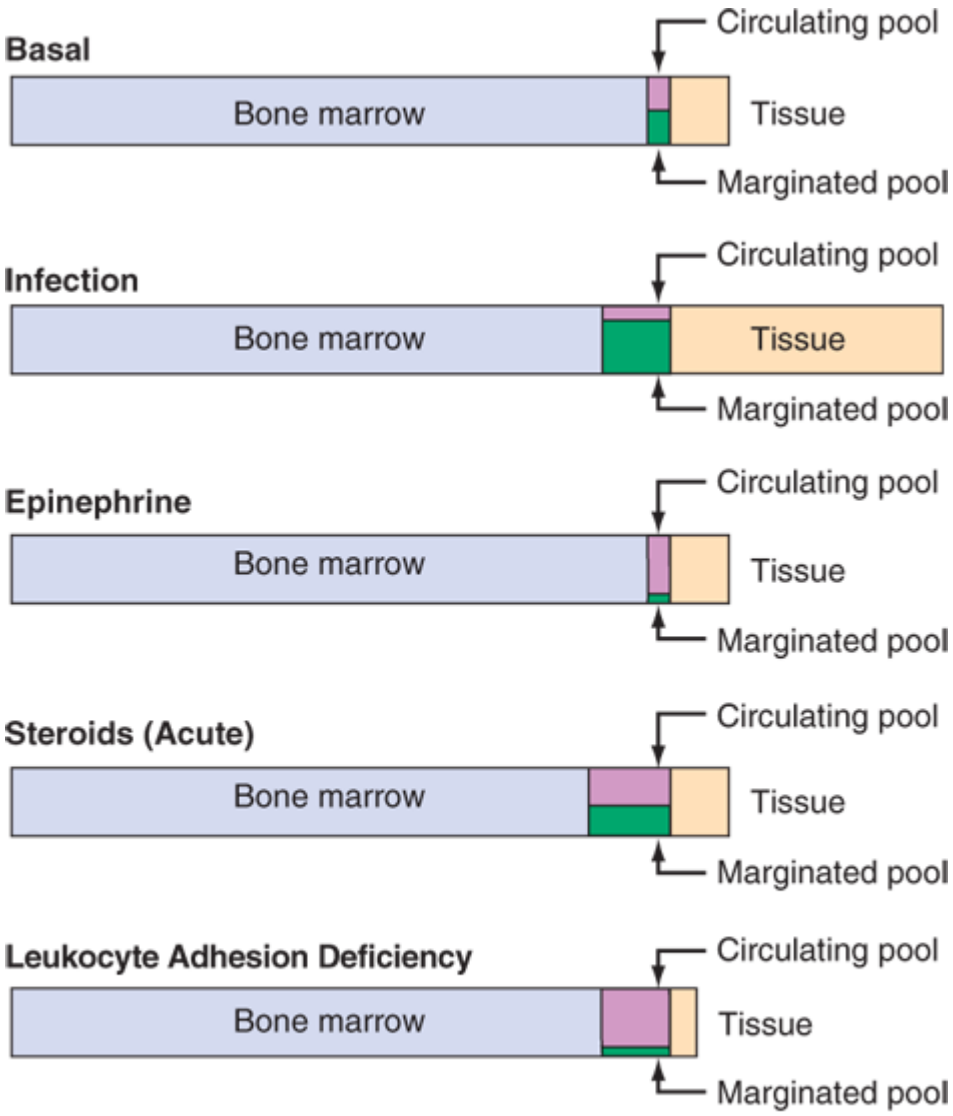
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MARROW RELEASE AND CIRCULATING COMPARTMENTS

Specific signals, including IL-1, tumor necrosis factor α (TNF-α), the CSFs, complement fragments, and chemokines, mobilize leukocytes from the bone marrow and deliver them to the blood in an unstimulated state. Under normal conditions, ~90% of the neutrophil pool is in the bone marrow, 2–3% in the circulation, and the remainder in the tissues **(Fig. 60-7)**.

FIGURE 60-7

Schematic neutrophil distribution and kinetics between the different anatomic and functional pools.



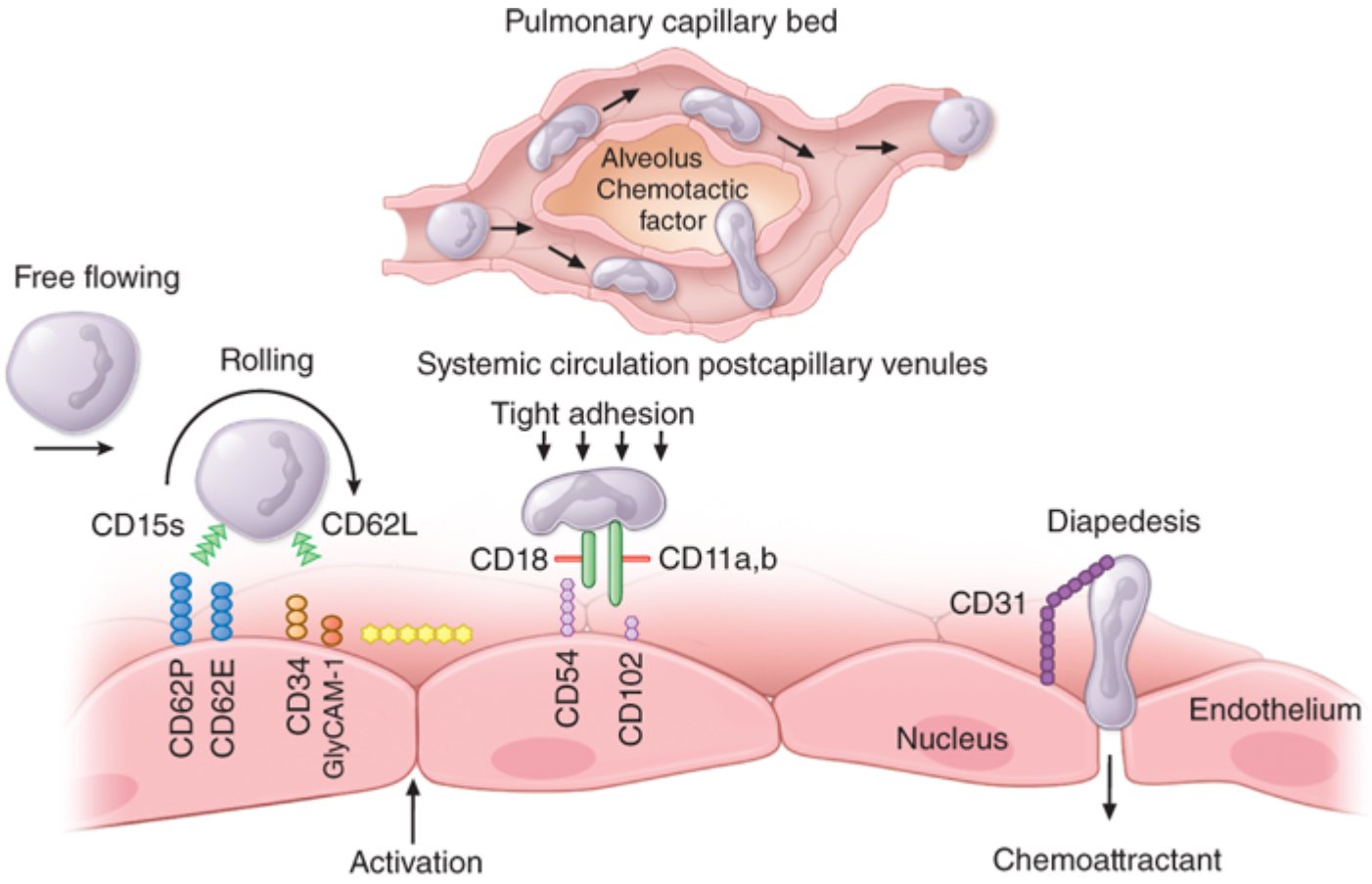
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The circulating pool exists in two dynamic compartments: one freely flowing and one marginated. The freely flowing pool is about one-half the neutrophils in the basal state and is composed of those cells that are in the blood and not in contact with the endothelium. Marginated leukocytes are those that are in close physical contact with the endothelium **(Fig. 60-8)**. In the pulmonary circulation, where an extensive capillary bed (~1000 capillaries per alveolus) exists, margination occurs because the capillaries are about the same size as a mature neutrophil. Therefore, neutrophil fluidity and deformability are necessary to make the transit through the pulmonary bed. Increased neutrophil rigidity and decreased deformability lead to augmented neutrophil trapping and margination in the lung. In contrast, in the systemic postcapillary venules, margination is mediated by the interaction of specific cell-surface molecules called *selectins*. Selectins are glycoproteins expressed on neutrophils and endothelial cells, among others, that cause a low-affinity interaction, resulting in "rolling" of the neutrophil along the endothelial surface. On neutrophils, the molecule L-selectin (cluster determinant [CD] 62L) binds to glycosylated proteins on endothelial cells (e.g., glycosylation-dependent cell adhesion molecule [GlyCAM1] and CD34). Glycoproteins on neutrophils, most importantly sialyl-Lewis^x (SLe^x, CD15s), are targets for binding of selectins expressed on endothelial cells (E-selectin [CD62E] and P-selectin [CD62P]) and other leukocytes. In response to chemotactic stimuli from

injured tissues (e.g., complement product C5a, leukotriene B_4, IL-8) or bacterial products (e.g., *N*-formylmethionylleucylphenylalanine [f-met-leu-phe]), neutrophil adhesiveness increases through mobilization of intracellular adhesion proteins stored in specific granules to the cell surface, and the cells "stick" to the endothelium through *integrins*. The integrins are leukocyte glycoproteins that exist as complexes of a common CD18 β chain with CD11a (LFA-1), CD11b (called Mac-1, CR3, or the C3bi receptor), and CD11c (called p150,95 or CR4). CD11a/CD18 and CD11b/CD18 bind to specific endothelial receptors (intercellular adhesion molecules [ICAM] 1 and 2).

FIGURE 60-8

Neutrophil travel through the pulmonary capillaries is dependent on neutrophil deformability. Neutrophil rigidity (e.g., caused by C5a) enhances pulmonary trapping and response to pulmonary pathogens in a way that is not so dependent on cell-surface receptors. Intraalveolar chemotactic factors, such as those caused by certain bacteria (e.g., *Streptococcus pneumoniae*), lead to diapedesis of neutrophils from the pulmonary capillaries into the alveolar space. Neutrophil interaction with the endothelium of the systemic postcapillary venules is dependent on molecules of attachment. The neutrophil "rolls" along the endothelium using selectins: neutrophil CD15s (sialyl-Lewis^X) binds to CD62E (E-selectin) and CD62P (P-selectin) on endothelial cells; CD62L (L-selectin) on neutrophils binds to CD34 and other molecules (e.g., GlyCAM-1) expressed on endothelium. Chemokines or other activation factors stimulate integrin-mediated "tight adhesion": CD11a/CD18 (LFA-1) and CD11b/CD18 (Mac-1, CR3) bind to CD54 (ICAM-1) and CD102 (ICAM-2) on the endothelium. Diapedesis occurs between endothelial cells: CD31 (PECAM-1) expressed by the emigrating neutrophil interacts with CD31 expressed at the endothelial cell-cell junction. CD, cluster determinant; GlyCAM, glycosylation-dependent cell adhesion molecule; ICAM, intercellular adhesion molecule; PECAM, platelet/endothelial cell adhesion molecule.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

On cell stimulation, L-selectin is shed from neutrophils, and E-selectin increases in the blood, presumably because it is shed from endothelial cells; receptors for chemoattractants and opsonins are mobilized; and the phagocytes orient toward the chemoattractant source in the extravascular space, increase their motile activity (chemokinesis), and migrate directionally (chemotaxis) into tissues. The process of migration into tissues is called *diapedesis* and involves the crawling of neutrophils between postcapillary endothelial cells that open junctions between adjacent cells to permit leukocyte passage. Diapedesis involves platelet/endothelial cell adhesion molecule (PECAM) 1 (CD31), which is expressed on both the emigrating leukocyte and the endothelial cells. The endothelial responses (increased blood flow from increased vasodilation and permeability) are mediated by anaphylatoxins (e.g., C3a and C5a) as well as vasodilators such as histamine, bradykinin, serotonin, nitric oxide, vascular endothelial growth factor (VEGF), and prostaglandins E and I. Cytokines regulate some of these processes (e.g., TNF-α induction of VEGF, interferon [IFN] γ inhibition of prostaglandin E).

In the healthy adult, most neutrophils leave the body by migration through the mucous membrane of the gastrointestinal tract. Normally, neutrophils spend a short time in the circulation (half-life, 6–7 h). Senescent neutrophils are cleared from the circulation by macrophages in the lung and spleen. Once in the tissues, neutrophils release enzymes, such as collagenase and elastase, which may help establish abscess cavities. Neutrophils ingest pathogenic materials that have been opsonized by IgG and C3b. Fibronectin and the tetrapeptide tuftsin also facilitate phagocytosis.

With phagocytosis comes a burst of oxygen consumption and activation of the hexose-monophosphate shunt. A membrane-associated NADPH oxidase, consisting of membrane and cytosolic components, is assembled and catalyzes the univalent reduction of oxygen to superoxide anion, which is then converted by superoxide dismutase to hydrogen peroxide and other toxic oxygen products (e.g., hydroxyl radical). Hydrogen peroxide + chloride + neutrophil myeloperoxidase generate hypochlorous acid (bleach), hypochlorite, and chlorine. These products oxidize and halogenate microorganisms and tumor cells and, when uncontrolled, can damage host tissue. Strongly cationic proteins, defensins, elastase, cathepsins, and probably nitric oxide also participate in microbial killing. Lactoferrin chelates iron, an important growth factor for microorganisms, especially fungi. Other enzymes, such as lysozyme and acid proteases, help digest microbial debris. After 1–4 days in tissues, neutrophils die. The apoptosis of neutrophils is also cytokine-regulated; granulocyte colony-stimulating factor (G-CSF) and IFN-γ prolong their life span. Under certain conditions, such as in delayed-type hypersensitivity, monocyte accumulation occurs within 6–12 h of initiation of inflammation. Neutrophils, monocytes, microorganisms in various states of digestion, and altered local tissue cells make up the inflammatory exudate, pus. Myeloperoxidase confers the characteristic green color to pus and may participate in turning off the inflammatory process by inactivating chemoattractants and immobilizing phagocytic cells.

Neutrophils respond to certain cytokines (IFN-γ, granulocyte-macrophage colony-stimulating factor [GM-CSF], IL-8) and produce cytokines and chemotactic signals (TNF-α, IL-8, macrophage inflammatory protein [MIP] 1) that modulate the inflammatory response. In the presence of fibrinogen, f-met-leu-phe or leukotriene B_4 induce IL-8 production by neutrophils, providing autocrine amplification of inflammation.

Chemokines (*chemo*attractant cyto*kines*) are small proteins produced by many different cell types, including endothelial cells, fibroblasts, epithelial cells, neutrophils, and monocytes, that regulate neutrophil, monocyte, eosinophil, and lymphocyte recruitment and activation. Chemokines transduce their signals through heterotrimeric G protein–linked receptors that have seven cell membrane–spanning domains, the same type of cell-surface receptor that mediates the response to the classic chemoattractants f-met-leu-phe and C5a. Four major groups of chemokines are recognized based on the cysteine structure near the N terminus: C, CC, CXC, and CXXXC. The CXC cytokines such as IL-8 mainly attract neutrophils; CC chemokines such as MIP-1 attract lymphocytes, monocytes, eosinophils, and basophils; the C chemokine lymphotactin is T-cell tropic; the CXXXC chemokine fractalkine attracts neutrophils, monocytes, and T cells. These molecules and their receptors not only regulate the trafficking and activation of inflammatory cells, but specific chemokine receptors also serve as co-receptors for HIV infection **(Chap. 197)** and have a role in other viral infections such as West Nile infection and atherogenesis.

NEUTROPHIL ABNORMALITIES

Defects in the neutrophil life cycle can lead to dysfunction and compromised host defenses. Inflammation is often depressed, and the clinical result is often recurrent, severe bacterial and fungal infections. Aphthous ulcers of mucous membranes (gray ulcers without pus) and gingivitis and periodontal disease suggest a phagocytic cell disorder. Patients with congenital phagocyte defects can have infections within the first few days of life. Skin, ear, upper and lower respiratory tract, and bone infections are common. Sepsis and

meningitis are rare. In some disorders, the frequency of infection is variable, and patients can go for months or even years without major infection. Aggressive management of these congenital diseases, including hematopoietic stem cell transplantation and gene therapy, has extended the life span of patients well into adulthood.

Neutropenia

The consequences of absent neutrophils are dramatic. Susceptibility to infectious diseases increases sharply when neutrophil counts fall to <1000 cells/μL. When the absolute neutrophil count (ANC; band forms and mature neutrophils combined) falls to <500 cells/μL, control of endogenous microbial flora (e.g., mouth, gut) is impaired; when the ANC is <200/μL, the local inflammatory process is absent. Neutropenia can be due to depressed production, increased peripheral destruction, or excessive peripheral pooling. A falling neutrophil count or a significant decrease in the number of neutrophils below steady-state levels, together with a failure to increase neutrophil counts in the setting of infection or other challenge, requires investigation. Acute neutropenia, such as that caused by cancer chemotherapy, is more likely to be associated with increased risk of infection than neutropenia of long duration (months to years) that reverses in response to infection or carefully controlled administration of endotoxin (see "Laboratory Diagnosis and Management," below).

Some causes of inherited and acquired neutropenia are listed in **Table 60-1**. The most common neutropenias are iatrogenic, resulting from the use of cytotoxic or immunosuppressive therapies for malignancy or control of autoimmune disorders. These drugs cause neutropenia because they result in decreased production of rapidly growing progenitor (stem) cells of the marrow. Certain antibiotics such as chloramphenicol, trimethoprim-sulfamethoxazole, flucytosine, vidarabine, and the antiretroviral drug zidovudine may cause neutropenia by inhibiting proliferation of myeloid precursors. Azathioprine and 6-mercaptopurine are metabolized by the enzyme thiopurine methyltransferase (TMPT), hypofunctional polymorphisms that are found in 11% of whites and can lead to accumulation of 6-thioguanine and profound marrow toxicity. The marrow suppression is generally dose-related and dependent on continued administration of the drug. Cessation of the offending agent and recombinant human G-CSF usually reverse these forms of neutropenia.

TABLE 60-1

Causes of Neutropenia

Decreased Production
Drug-induced—alkylating agents (nitrogen mustard, busulfan, chlorambucil, cyclophosphamide); antimetabolites (methotrexate, 6-mercaptopurine, 5-flucytosine); noncytotoxic agents (antibiotics [chloramphenicol, penicillins, sulfonamides], phenothiazines, tranquilizers [meprobamate], anticonvulsants [carbamazepine], antipsychotics [clozapine], certain diuretics, anti-inflammatory agents, antithyroid drugs, many others)
Hematologic diseases—idiopathic, cyclic neutropenia, Chédiak-Higashi syndrome, aplastic anemia, infantile genetic disorders (see text)
Tumor invasion, myelofibrosis
Nutritional deficiency—vitamin B_{12}, folate (especially alcoholics)
Infection—tuberculosis, typhoid fever, brucellosis, tularemia, measles, infectious mononucleosis, malaria, viral hepatitis, leishmaniasis, AIDS
Peripheral Destruction
Antineutrophil antibodies and/or splenic or lung trapping
Autoimmune disorders—Felty's syndrome, rheumatoid arthritis, lupus erythematosus
Drugs as haptens—aminopyrine, α-methyldopa, phenylbutazone, mercurial diuretics, some phenothiazines
Granulomatosis with polyangiitis (Wegener's)
Peripheral Pooling (Transient Neutropenia)
Overwhelming bacterial infection (acute endotoxemia)
Hemodialysis
Cardiopulmonary bypass

Another important mechanism for iatrogenic neutropenia is the effect of drugs that serve as immune haptens and sensitize neutrophils or neutrophil precursors to immune-mediated peripheral destruction. This

form of drug-induced neutropenia can be seen within 7 days of exposure to the drug; with previous drug exposure, resulting in preexisting antibodies, neutropenia may occur a few hours after administration of the drug. Although any drug can cause this form of neutropenia, the most frequent causes are commonly used antibiotics, such as sulfa-containing compounds, penicillins, and cephalosporins. Fever and eosinophilia may also be associated with drug reactions, but often these signs are not present. Drug-induced neutropenia can be severe, but discontinuation of the sensitizing drug is sufficient for recovery, which is usually seen within 5–7 days and is complete by 10 days. Readministration of the sensitizing drug should be avoided, because abrupt neutropenia will often result. For this reason, diagnostic challenge should be avoided.

Autoimmune neutropenias caused by circulating antineutrophil antibodies are another form of acquired neutropenia that results in increased destruction of neutrophils. Acquired neutropenia may also be seen with viral infections, including acute infection with HIV. Acquired neutropenia may be cyclic in nature, occurring at intervals of several weeks. Acquired cyclic or stable neutropenia may be associated with an expansion of large granular lymphocytes (LGLs), which may be T cells, NK cells, or NK-like cells. Patients with large granular lymphocytosis may have moderate blood and bone marrow lymphocytosis, neutropenia, polyclonal hypergammaglobulinemia, splenomegaly, rheumatoid arthritis, and absence of lymphadenopathy. Such patients may have a chronic and relatively stable course. Recurrent bacterial infections are frequent. Benign and malignant forms of this syndrome occur. In some patients, a spontaneous regression has occurred even after 11 years, suggesting an immunoregulatory defect as the basis for at least one form of the disorder. Glucocorticoids, cyclosporine, and methotrexate are commonly used to manage these cytopenias.

Hereditary Neutropenias

Hereditary neutropenias are rare and may manifest in early childhood as a profound constant neutropenia or agranulocytosis. Congenital forms of neutropenia include Kostmann's syndrome (neutrophil count <100/μL), which is often fatal and due to mutations in the antiapoptosis gene *HAX-1*; severe chronic neutropenia (neutrophil count of 300–1500/μL) due to mutations in neutrophil elastase (*ELANE*); hereditary cyclic neutropenia, or, more appropriately, cyclic hematopoiesis, also due to mutations in neutrophil elastase (*ELANE*); the cartilage-hair hypoplasia syndrome due to mutations in the mitochondrial RNA-processing endoribonuclease *RMRP*; Shwachman-Diamond syndrome associated with pancreatic insufficiency due to mutations in the Shwachman-Bodian-Diamond syndrome gene *SBDS*; the WHIM (*w*arts, *h*ypogammaglobulinemia, *i*nfections, *m*yelokathexis [retention of WBCs in the marrow]) syndrome, characterized by neutrophil hypersegmentation and bone marrow myeloid arrest due to mutations in the chemokine receptor *CXCR4*; and neutropenias associated with other immune defects, such as X-linked agammaglobulinemia, Wiskott-Aldrich syndrome, and CD40 ligand deficiency. Mutations in the G-CSF receptor can develop in severe congenital neutropenia and are linked to leukemia. Absence of both myeloid and lymphoid cells is seen in reticular dysgenesis, due to mutations in the nuclear genome-encoded mitochondrial enzyme adenylate kinase-2 (*AK2*).

Maternal factors can be associated with neutropenia in the newborn. Transplacental transfer of IgG directed against antigens on fetal neutrophils can result in peripheral destruction. Drugs (e.g., thiazides) ingested

during pregnancy can cause neutropenia in the newborn by either depressed production or peripheral destruction.

In Felty's syndrome—the triad of rheumatoid arthritis, splenomegaly, and neutropenia **(Chap. 351)**—spleen-produced antibodies can shorten neutrophil life span, while large granular lymphocytes can attack marrow neutrophil precursors. Splenectomy may increase the neutrophil count in Felty's syndrome and lower serum neutrophil-binding IgG. Some Felty's syndrome patients also have neutropenia associated with an increased number of LGLs. Splenomegaly with peripheral trapping and destruction of neutrophils is also seen in lysosomal storage diseases and in portal hypertension.

Neutrophilia

Neutrophilia results from increased neutrophil production, increased marrow release, or defective margination **(Table 60-2)**. The most important acute cause of neutrophilia is infection. Neutrophilia from acute infection represents both increased production and increased marrow release. Increased production is also associated with chronic inflammation and certain myeloproliferative diseases. Increased marrow release and mobilization of the marginated leukocyte pool are induced by glucocorticoids. Release of epinephrine, as with vigorous exercise, excitement, or stress, will demarginate neutrophils in the spleen and lungs and double the neutrophil count in minutes. Cigarette smoking can elevate neutrophil counts above the normal range. Leukocytosis with cell counts of 10,000–25,000/μL occurs in response to infection and other forms of acute inflammation and results from both release of the marginated pool and mobilization of marrow reserves. Persistent neutrophilia with cell counts of ≥30,000–50,000/μL is called a *leukemoid reaction*, a term often used to distinguish this degree of neutrophilia from leukemia. In a leukemoid reaction, the circulating neutrophils are usually mature and not clonally derived.

TABLE 60-2

Causes of Neutrophilia

Increased Production
Idiopathic
Drug-induced—glucocorticoids, G-CSF
Infection—bacterial, fungal, sometimes viral
Inflammation—thermal injury, tissue necrosis, myocardial and pulmonary infarction, hypersensitivity states, collagen vascular diseases
Myeloproliferative diseases—myelocytic leukemia, myeloid metaplasia, polycythemia vera
Increased Marrow Release
Glucocorticoids
Acute infection (endotoxin)
Inflammation—thermal injury
Decreased or Defective Margination
Drugs—epinephrine, glucocorticoids, nonsteroidal anti-inflammatory agents
Stress, excitement, vigorous exercise
Leukocyte adhesion deficiency type 1 (CD18); leukocyte adhesion deficiency type 2 (selectin ligand, CD15s); leukocyte adhesion deficiency type 3 (FERMT3)
Miscellaneous
Metabolic disorders—ketoacidosis, acute renal failure, eclampsia, acute poisoning
Drugs—lithium
Other—metastatic carcinoma, acute hemorrhage or hemolysis

Abbreviation: G-CSF, granulocyte colony-stimulating factor.

Abnormal Neutrophil Function

Inherited and acquired abnormalities of phagocyte function are listed in **Table 60-3**. The resulting diseases are best considered in terms of the functional defects of adherence, chemotaxis, and microbicidal activity. The distinguishing features of the important inherited disorders of phagocyte function are shown in **Table 60-4**.

TABLE 60-3

Types of Granulocyte and Monocyte Disorders

Cause of Indicated Dysfunction			
Function	**Drug-Induced**	**Acquired**	**Inherited**
Adherence-aggregation	Aspirin, colchicine, alcohol, glucocorticoids, ibuprofen, piroxicam	Neonatal state, hemodialysis	Leukocyte adhesion deficiency types 1, 2, and 3
Deformability		Leukemia, neonatal state, diabetes mellitus, immature neutrophils	
Chemokinesis-chemotaxis	Glucocorticoids (high dose), auranofin, colchicine (weak effect), phenylbutazone, naproxen, indomethacin, interleukin 2	Thermal injury, malignancy, malnutrition, periodontal disease, neonatal state, systemic lupus erythematosus, rheumatoid arthritis, diabetes mellitus, sepsis, influenza virus infection, herpes simplex virus infection, acrodermatitis enteropathica, AIDS	Chédiak-Higashi syndrome, neutrophil-specific granule deficiency, hyper IgE–recurrent infection (Job's) syndrome (in some patients), Down's syndrome, α-mannosidase deficiency, leukocyte adhesion deficiencies, Wiskott-Aldrich syndrome
Microbicidal activity	Colchicine, cyclophosphamide, glucocorticoids (high dose), TNF-α-blocking antibodies	Leukemia, aplastic anemia, certain neutropenias, tuftsin deficiency, thermal injury, sepsis, neonatal state, diabetes mellitus, malnutrition, AIDS	Chédiak-Higashi syndrome, neutrophil-specific granule deficiency, chronic granulomatous disease, defects in IFNγ/IL-12 axis

Abbreviations: IFNγ, interferon γ; IL, interleukin; TNF-α, tumor necrosis factor alpha.

TABLE 60-4

Inherited Disorders of Phagocyte Function: Differential Features

Clinical Manifestations	Cellular or Molecular Defects	Diagnosis
Chronic Granulomatous Diseases (70% X-Linked, 30% Autosomal Recessive)		
Severe infections of skin, ears, lungs, liver, and bone with catalase-positive microorganisms such as *Staphylococcus aureus, Burkholderia cepacia* complex*, Aspergillus* spp., *Chromobacterium violaceum*; often hard to culture organism; excessive inflammation with granulomas, frequent lymph node suppuration; granulomas can obstruct GI or GU tracts; gingivitis, aphthous ulcers, seborrheic dermatitis	No respiratory burst due to the lack of one of five NADPH oxidase subunits in neutrophils, monocytes, and eosinophils	DHR or NBT test; no superoxide and H_2O_2 production by neutrophils; immunoblot for NADPH oxidase components; genetic detection
Chédiak-Higashi Syndrome (Autosomal Recessive)		
Recurrent pyogenic infections, especially with *S. aureus*; many patients get lymphoma-like illness during adolescence; periodontal disease; partial oculocutaneous albinism, nystagmus, progressive peripheral neuropathy, mental retardation in some patients	Reduced chemotaxis and phagolysosome fusion, increased respiratory burst activity, defective egress from marrow, abnormal skin window; defect in *CHS1*	Giant primary granules in neutrophils and other granule-bearing cells (Wright's stain); genetic detection
Specific Granule Deficiency (Autosomal Recessive and Dominant)		
Recurrent infections of skin, ears, and sinopulmonary tract; delayed wound healing; decreased inflammation; bleeding diathesis	Abnormal chemotaxis, impaired respiratory burst and bacterial killing, failure to upregulate chemotactic and adhesion receptors with stimulation, defect in transcription of granule proteins; defect in *CEBPE*	Lack of secondary (specific) granules in neutrophils (Wright's stain), no neutrophil-specific granule contents (i.e., lactoferrin), no defensins, platelet α granule abnormality; genetic detection

Clinical Manifestations	Cellular or Molecular Defects	Diagnosis
Myeloperoxidase Deficiency (Autosomal Recessive)		
Clinically normal except in patients with underlying disease such as diabetes mellitus; then candidiasis or other fungal infections	No myeloperoxidase due to pre- and posttranslational defects in myeloperoxidase deficiency	No peroxidase in neutrophils; genetic detection
Leukocyte Adhesion Deficiency		
Type 1: Delayed separation of umbilical cord, sustained neutrophilia, recurrent infections of skin and mucosa, gingivitis, periodontal disease	Impaired phagocyte adherence, aggregation, spreading, chemotaxis, phagocytosis of C3bi-coated particles; defective production of CD18 subunit common to leukocyte integrins	Reduced phagocyte surface expression of the CD18-containing integrins with monoclonal antibodies against LFA-1 (CD18/CD11a), Mac-1 or CR3 (CD18/CD11b), p150,95 (CD18/CD11c); genetic detection
Type 2: Mental retardation, short stature, Bombay (hh) blood phenotype, recurrent infections, neutrophilia	Impaired phagocyte rolling along endothelium; due to defects in fucose transporter	Reduced phagocyte surface expression of Sialyl-Lewis^x, with monoclonal antibodies against CD15s; genetic detection
Type 3: Petechial hemorrhage, recurrent infections	Impaired signaling for integrin activation resulting in impaired adhesion due to mutation in *FERMT3*	Reduced signaling for adhesion through integrins; genetic detection
Phagocyte Activation Defects (X-Linked and Autosomal Recessive)		

Clinical Manifestations	Cellular or Molecular Defects	Diagnosis
NEMO deficiency: mild hypohidrotic ectodermal dysplasia; broad-based immune defect: pyogenic and encapsulated bacteria, viruses, *Pneumocystis*, mycobacteria; X-linked	Impaired phagocyte activation by IL-1, IL-18, TLR, CD40L, TNF-α leading to problems with inflammation and antibody production	Poor in vitro response to endotoxin; impaired NF-κB activation; genetic detection
IRAK4 and MyD88 deficiency: susceptibility to pyogenic bacteria such as staphylococci, streptococci, clostridia; resistant to *Candida*; autosomal recessive	Impaired phagocyte activation by endotoxin through TLR and other pathways; TNF-α signaling preserved	Poor in vitro response to endotoxin; lack of NF-κB activation by endotoxin; genetic detection
Hyper IgE–Recurrent Infection Syndrome (Autosomal Dominant) (Job's Syndrome)		
Eczematoid or pruritic dermatitis, "cold" skin abscesses, recurrent pneumonias with *S. aureus* with bronchopleural fistulae and cyst formation, mild eosinophilia, mucocutaneous candidiasis, characteristic facies, restrictive lung disease, scoliosis, delayed primary dental deciduation	Reduced chemotaxis in some patients, reduced memory T and B cells; mutation in *STAT3*	Somatic and immune features involving lungs, skeleton, and immune system; serum IgE >2000 IU/mL; genetic testing
DOCK8 deficiency (autosomal recessive), severe eczema, atopic dermatitis, cutaneous abscesses, HSV, HPV, and molluscum infections, severe allergies, cancer	Impaired T-cell proliferation to mitogens; mutation in *DOCK8*	Severe allergies, viral infections, high IgE, eosinophilia, low IgM, progressive lymphopenia, genetic detection
Mycobacteria Susceptibility (Autosomal Dominant and Recessive Forms)		

Clinical Manifestations	Cellular or Molecular Defects	Diagnosis
Severe extrapulmonary or disseminated infections with bacille Calmette-Guérin (BCG), nontuberculous mycobacteria, salmonella, histoplasmosis, coccidioidomycosis, poor granuloma formation	Inability to kill intracellular organisms due to low IFN-γ production or response; mutations in IFN-γ receptors, IL-12 receptors, IL-12 p40, *STAT1*, *NEMO*, *ISG15*, *GATA2*	Abnormally low or very high levels of IFN-γ receptor 1; functional assays of cytokine production and response; genetic detection
GATA2 Deficiency (Autosomal Dominant)		
Persistent or disseminated warts, disseminated mycobacterial disease, low monocytes, NK cells, B cells; hypoplastic myelodysplasia, leukemia, cytogenetic abnormalities, pulmonary alveolar proteinosis	Impaired macrophage activity, cytopenias; mutations in *GATA2*	Profound circulating monocytopenia, NK and B-cell cytopenias; genetic detection

Abbreviations: C/EBPε, CCAAT/enhancer binding protein-ε; DHR, dihydrorhodamine (oxidation test); DOCK8, dedicator of cytokinesis 8; GI, gastrointestinal; GU, genitourinary; HPV, human papilloma virus; HSV, herpes simplex virus; IFN, interferon; IL, interleukin; IRAK4, IL-1 receptor–associated kinase 4; LFA-1, leukocyte function–associated antigen 1; MyD88, myeloid differentiation primary response gene 88; NADPH, nicotinamide–adenine dinueleotide phosphate; NBT, nitroblue tetrazolium (dye test); NEMO, NF-κB essential modulator; NF-κB, nuclear factor-κB; NK, natural killer; STAT1–3, signal transducer and activator of transcription 1–3; TLR, Toll-like receptor; TNF, tumor necrosis factor.

DISORDERS OF ADHESION

Three main types of leukocyte adhesion deficiency (LAD) have been described. All are autosomal recessive and result in the inability of neutrophils to exit the circulation to sites of infection, leading to leukocytosis and increased susceptibility to infection (Fig. 60-8). Patients with LAD 1 have mutations in *CD18*, the common component of the integrins LFA-1, Mac-1, and p150,95, leading to a defect in tight adhesion between neutrophils and the endothelium. The heterodimer formed by CD18/CD11b (Mac-1) is also the receptor for the complement-derived opsonin C3bi (CR3). The *CD18* gene is located on distal chromosome 21q. The severity of the defect determines the severity of clinical disease. Complete lack of expression of the leukocyte integrins results in a severe phenotype in which inflammatory stimuli do not increase the expression of leukocyte integrins on neutrophils or activated T and B cells. Neutrophils (and monocytes) from patients with LAD 1 adhere poorly to endothelial cells and protein-coated surfaces and exhibit defective spreading, aggregation, and chemotaxis. The inability of neutrophils to exit the vasculature to the tissue deprives the

tissue macrophage of its expected neutrophil ingestion, leading to macrophage production of IL-23, which induces T- cell production of IL-17, a potent proinflammatory cytokine. These processes conspire to drive inflammation in LAD1. Patients with LAD 1 have recurrent bacterial infections involving the skin, oral and genital mucosa, and respiratory and intestinal tracts; persistent leukocytosis (resting neutrophil counts of 15,000–20,000/μL) because cells do not marginate; and, in severe cases, a history of delayed separation of the umbilical stump. Infections, especially of the skin, may become necrotic with progressively enlarging borders, slow healing, and development of dysplastic scars. The most common bacteria are *Staphylococcus aureus* and enteric gram-negative bacteria. LAD 2 is caused by an abnormality of fucosylation of SLe^X (CD15s), the ligand on neutrophils that interacts with selectins on endothelial cells and is responsible for neutrophil rolling along the endothelium. Infection susceptibility in LAD 2 appears to be less severe than in LAD 1. LAD 2 is also known as *congenital disorder of glycosylation IIc* (CDGIIc) due to mutation in a GDP-fucose transporter (*SLC35C1*). LAD 3 is characterized by infection susceptibility, leukocytosis, and petechial hemorrhage due to impaired integrin activation caused by mutations in the gene *FERMT3*.

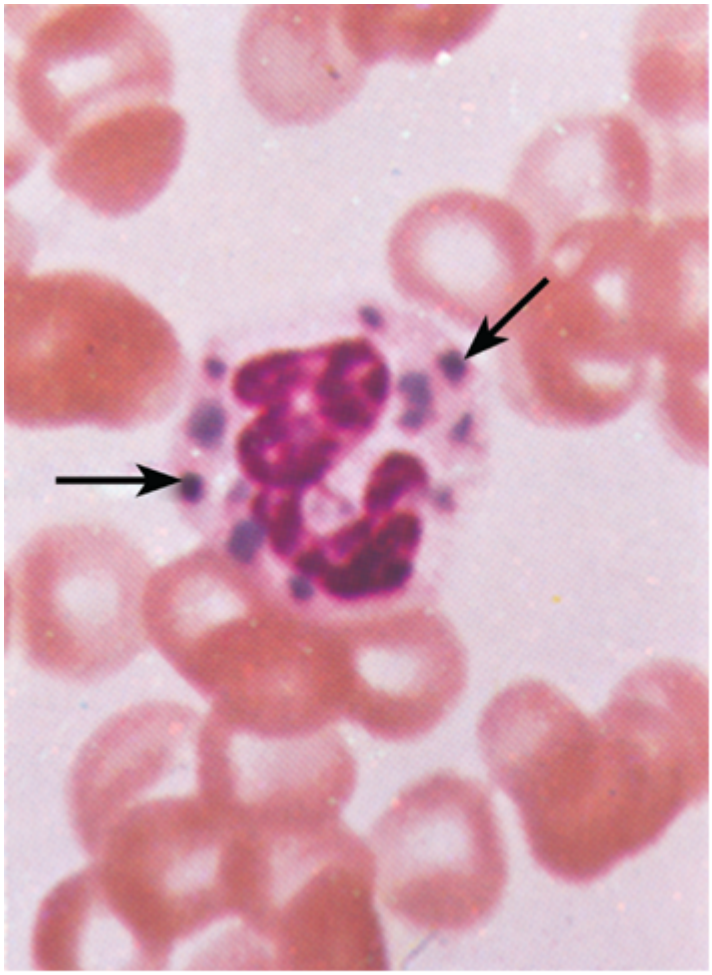
DISORDERS OF NEUTROPHIL GRANULES

The most common neutrophil defect is myeloperoxidase deficiency, a primary granule defect inherited as an autosomal recessive trait; the incidence is ~1 in 2000 persons. Isolated myeloperoxidase deficiency is not associated with clinically compromised defenses, presumably because other defense systems such as hydrogen peroxide generation are amplified. Microbicidal activity of neutrophils is delayed but not absent. Myeloperoxidase deficiency may make other acquired host defense defects more serious, and patients with myeloperoxidase deficiency and diabetes are more susceptible to *Candida* infections. An acquired form of myeloperoxidase deficiency occurs in myelomonocytic leukemia and acute myeloid leukemia.

Chédiak-Higashi syndrome (CHS) is a rare disease with autosomal recessive inheritance due to defects in the lysosomal transport protein LYST, encoded by the gene *CHS1* at 1q42. This protein is required for normal packaging and disbursement of granules. Neutrophils (and all cells containing lysosomes) from patients with CHS characteristically have large granules **(Fig. 60-9)**, making it a systemic disease. Patients with CHS have nystagmus, partial oculocutaneous albinism, and an increased number of infections resulting from many bacterial agents. Some CHS patients develop an “accelerated phase” in childhood with a hemophagocytic syndrome and an aggressive lymphoma requiring bone marrow transplantation. CHS neutrophils and monocytes have impaired chemotaxis and abnormal rates of microbial killing due to slow rates of fusion of the lysosomal granules with phagosomes. NK cell function is also impaired. CHS patients may develop a severe disabling peripheral neuropathy in adulthood.

FIGURE 60-9

Chédiak-Higashi syndrome. The granulocytes contain huge cytoplasmic granules formed from aggregation and fusion of azurophilic and specific granules. Large abnormal granules are found in other granule-containing cells throughout the body.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Specific granule deficiency is a rare autosomal recessive disease in which the production of secondary granules and their contents, as well as the primary granule component defensins, is defective. The defect in killing leads to severe bacterial infections. One type of specific granule deficiency is due to a mutation in the CCAAT/enhancer binding protein-ε, a regulator of expression of granule components. A dominant mutation in *C/EBP-ε* has also been described.

CHRONIC GRANULOMATOUS DISEASE

Chronic granulomatous disease (CGD) is a group of disorders of granulocyte and monocyte oxidative metabolism. Although CGD is rare, with an incidence of ~1 in 200,000 individuals, it is an important model of defective neutrophil oxidative metabolism. In about two-thirds of patients, CGD is inherited as an X-linked recessive trait; the remainder patients inherit the disease in an autosomal recessive pattern. Mutations in the genes for the five proteins that assemble at the plasma membrane account for all patients with CGD. Two proteins (a 91-kDa protein, abnormal in X-linked CGD, and a 22-kDa protein, absent in one form of autosomal recessive CGD) form the heterodimer cytochrome b-558 in the plasma membrane. Three other proteins (40, 47, and 67 kDa, abnormal in the other autosomal recessive forms of CGD) are cytoplasmic in origin and interact with the cytochrome after cell activation to form the NADPH oxidase, required for hydrogen peroxide production. Leukocytes from patients with CGD have severely diminished hydrogen peroxide production. The genes involved in each of the defects have been cloned and sequenced and the chromosome locations

identified. Patients with CGD characteristically have increased numbers of infections due to catalase-positive microorganisms (organisms that destroy their own hydrogen peroxide) such as *S. aureus*, *Burkholderia cepacia* complex, and *Aspergillus* species. When patients with CGD become infected, they often have extensive inflammatory reactions, and lymph node suppuration is common despite the administration of appropriate antibiotics. Aphthous ulcers and chronic inflammation of the nares are often present. Granulomas are frequent and can obstruct the gastrointestinal or genitourinary tracts. The excessive inflammation is due to failure to downregulate inflammation, reflecting a failure to inhibit the synthesis of, degradation of, or response to ILs or chemoattractants, leading to persistent myeloid reaction. Impaired killing of intracellular microorganisms by macrophages may lead to persistent cell-mediated immune activation and granuloma formation. Autoimmune complications such as immune thrombocytopenic purpura and juvenile rheumatoid arthritis are also increased in CGD. In addition, for unexplained reasons, discoid lupus is more common in X-linked carriers. Late complications, including nodular regenerative hyperplasia and portal hypertension, are increasingly recognized in older patients with CGD.

DISORDERS OF PHAGOCYTE ACTIVATION

Phagocytes depend on cell-surface stimulation to induce signals that evoke multiple levels of the inflammatory response, including cytokine synthesis, chemotaxis, and antigen presentation. Mutations affecting the major pathway that signals through NF-κB have been noted in patients with a variety of infection susceptibility syndromes. If the defects are at a very late stage of signal transduction, in the protein critical for NF-κB activation known as the NF-κB essential modulator (NEMO), then affected males develop ectodermal dysplasia and severe immune deficiency with susceptibility to bacteria, fungi, mycobacteria, and viruses. If the defects in NF-κB activation are closer to the cell-surface receptors, in the proteins transducing Toll-like receptor signals, IL-1 receptor–associated kinase 4 (IRAK4), and myeloid differentiation primary response gene 88 (MyD88), then children have a marked susceptibility to pyogenic infections early in life but develop resistance to infection later.

MONONUCLEAR PHAGOCYTES

The mononuclear phagocyte system is composed of monoblasts, promonocytes, and monocytes, in addition to the structurally diverse tissue macrophages that make up what was previously referred to as the reticuloendothelial system. Macrophages are long-lived phagocytic cells capable of many of the functions of neutrophils. They are also secretory cells that participate in many immunologic and inflammatory processes distinct from neutrophils. Monocytes leave the circulation by diapedesis more slowly than neutrophils and have a half-life in the blood of 12–24 h.

After blood monocytes arrive in the tissues, they differentiate into macrophages (“big eaters”) with specialized functions suited for specific anatomic locations. Macrophages are particularly abundant in capillary walls of the lung, spleen, liver, and bone marrow, where they function to remove microorganisms and other noxious elements from the blood. Alveolar macrophages, liver Kupffer cells, splenic macrophages, peritoneal macrophages, bone marrow macrophages, lymphatic macrophages, brain microglial cells, and dendritic macrophages all have specialized functions. Macrophage-secreted products include lysozyme,

neutral proteases, acid hydrolases, arginase, complement components, enzyme inhibitors (plasmin, α_2-macroglobulin), binding proteins (transferrin, fibronectin, transcobalamin II), nucleosides, and cytokines (TNF-α; IL-1, 8, 12, 18). IL-1 **(Chaps. 15 and 342)** has many functions, including initiating fever in the hypothalamus, mobilizing leukocytes from the bone marrow, and activating lymphocytes and neutrophils. TNF-α is a pyrogen that duplicates many of the actions of IL-1 and plays an important role in the pathogenesis of gram-negative shock **(Chap. 297)**. TNF-α stimulates production of hydrogen peroxide and related toxic oxygen species by macrophages and neutrophils. In addition, TNF-α induces catabolic changes that contribute to the profound wasting (cachexia) associated with many chronic diseases.

Other macrophage-secreted products include reactive oxygen and nitrogen metabolites, bioactive lipids (arachidonic acid metabolites and platelet-activating factors), chemokines, CSFs, and factors stimulating fibroblast and vessel proliferation. Macrophages help regulate the replication of lymphocytes and participate in the killing of tumors, viruses, and certain bacteria (*Mycobacterium tuberculosis* and *Listeria monocytogenes*). Macrophages are key effector cells in the elimination of intracellular microorganisms. Their ability to fuse to form giant cells that coalesce into granulomas in response to some inflammatory stimuli is important in the elimination of intracellular microbes and is under the control of IFN-γ. Nitric oxide induced by IFN-γ is an important effector against intracellular parasites, including tuberculosis and *Leishmania*.

Macrophages play an important role in the immune response **(Chap. 342)**. They process and present antigen to lymphocytes and secrete cytokines that modulate and direct lymphocyte development and function. Macrophages participate in autoimmune phenomena by removing immune complexes and other substances from the circulation. Polymorphisms in macrophage receptors for immunoglobulin (FcγRII) determine susceptibility to some infections and autoimmune diseases. In wound healing, they dispose of senescent cells, and they contribute to atheroma development. Macrophage elastase mediates development of emphysema from cigarette smoking.

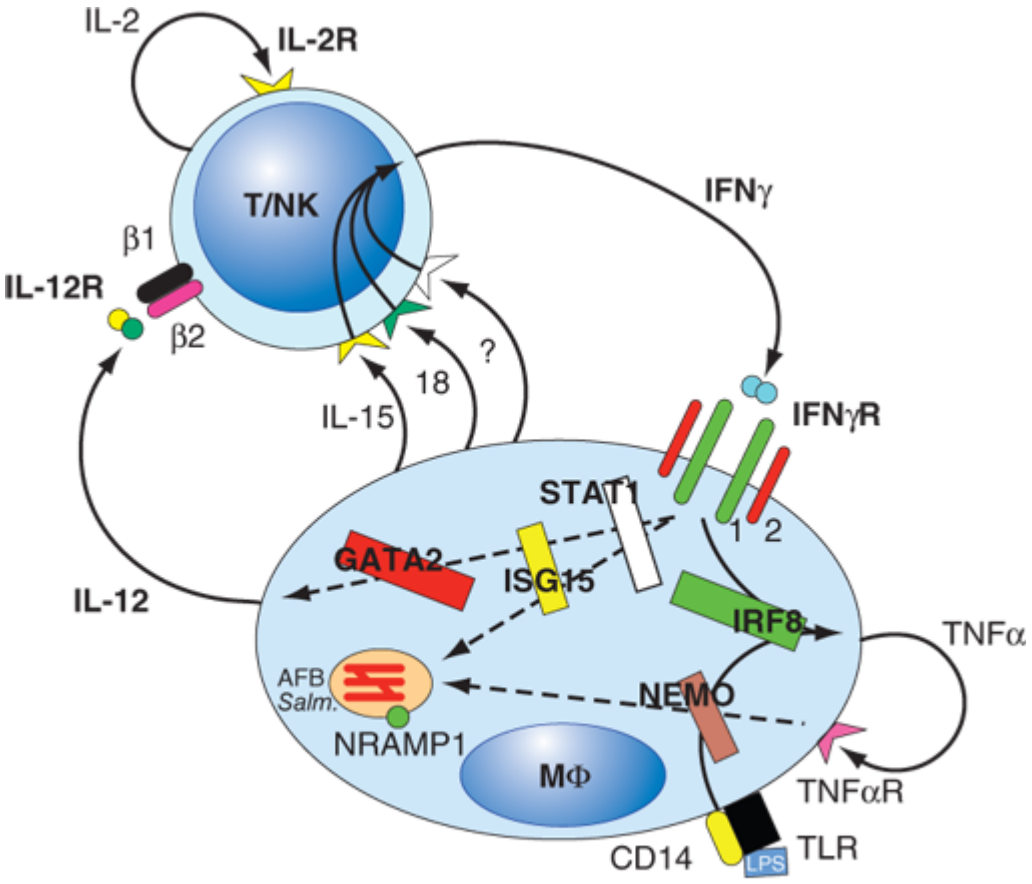
DISORDERS OF THE MONONUCLEAR PHAGOCYTE SYSTEM

Many disorders of neutrophils extend to mononuclear phagocytes. Monocytosis is associated with tuberculosis, brucellosis, subacute bacterial endocarditis, Rocky Mountain spotted fever, malaria, and visceral leishmaniasis (kala azar). Monocytosis also occurs with malignancies, leukemias, myeloproliferative syndromes, hemolytic anemias, chronic idiopathic neutropenias, and granulomatous diseases such as sarcoidosis, regional enteritis, and some collagen vascular diseases. Patients with LAD, hyperimmunoglobulin E–recurrent infection (Job's) syndrome, CHS, and CGD all have defects in the mononuclear phagocyte system.

Monocyte cytokine production or response is impaired in some patients with disseminated nontuberculous mycobacterial infection who are not infected with HIV. Genetic defects in the pathways regulated by IFN-γ and IL-12 lead to impaired killing of intracellular bacteria, mycobacteria, salmonellae, and certain viruses **(Fig. 60-10)**.

FIGURE 60-10

Lymphocyte-macrophage interactions underlying resistance to mycobacteria and other intracellular pathogens such as *Salmonella, Histoplasma,* and *Coccidioides*. Mycobacteria (and others) infect macrophages, leading to the production of IL-12, which activates T or NK cells through its receptor, leading to production of IL-2 and IFN-γ. IFN-γ acts through its receptor on macrophages to upregulate TNF-γ and IL-12 and kill intracellular pathogens. Other critical interacting molecules include signal transducer and activator of transcription 1 (STAT1), interferon regulatory factor 8 (IRF8), GATA2, and ISG15. Mutant forms of the cytokines and receptors shown in *bold type* have been found in severe cases of nontuberculous mycobacterial infection, salmonellosis and other intracellular pathogens. AFB, acid-fast bacilli; IFN, interferon; IL, interleukin; NEMO, nuclear factor-κB essential modulator; NK, natural killer; TLR, Toll-like receptor; TNF, tumor necrosis factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Certain viral infections impair mononuclear phagocyte function. For example, influenza virus infection causes abnormal monocyte chemotaxis. Mononuclear phagocytes can be infected by HIV using CCR5, the chemokine receptor that acts as a co-receptor with CD4 for HIV. T lymphocytes produce IFN-γ, which induces FcR expression and phagocytosis and stimulates hydrogen peroxide production by mononuclear phagocytes and neutrophils. In certain diseases, such as AIDS, IFN-γ production may be deficient, whereas in other diseases, such as T-cell lymphomas, excessive release of IFN-γ may be associated with erythrophagocytosis by splenic macrophages.

Autoinflammatory diseases are characterized by abnormal cytokine regulation, leading to excess inflammation in the absence of infection. These diseases can mimic infectious or immunodeficient syndromes. Gain-of-function mutations in the TNF-α receptor cause TNF-α receptor–associated periodic syndrome (TRAPS), which is characterized by recurrent fever in the absence of infection, due to persistent stimulation of the TNF-α receptor (Chap. 362). Diseases with abnormal IL-1 regulation leading to fever include familial Mediterranean fever due to mutations in *PYRIN*. Mutations in *cold-induced autoinflammatory syndrome 1* (*CIAS1*) lead to neonatal-onset multisystem autoinflammatory disease, familial cold urticaria, and Muckle-Wells syndrome. The syndrome of *p*yoderma gangrenosum, *a*cne, and sterile *p*yogenic *a*rthritis (PAPA syndrome) is caused by mutations in *PSTPIP1*. In contrast to these syndromes of overexpression of proinflammatory cytokines, blockade of TNF-α by the antagonists infliximab, adalimumab, certolizumab, golimumab, or etanercept has been associated with severe infections due to tuberculosis, nontuberculous mycobacteria, and fungi (Chap. 362).

Monocytopenia occurs with acute infections, with stress, and after treatment with glucocorticoids. Drugs that suppress neutrophil production in the bone marrow can cause monocytopenia. Persistent severe circulating monocytopenia is seen in GATA2 deficiency, even though macrophages are found at the sites of inflammation. Monocytopenia also occurs in aplastic anemia, hairy cell leukemia, acute myeloid leukemia, and as a direct result of myelotoxic drugs.

EOSINOPHILS

Eosinophils and neutrophils share similar morphology, many lysosomal constituents, phagocytic capacity, and oxidative metabolism. Eosinophils express a specific chemoattractant receptor and respond to a specific chemokine, eotaxin, but little is known about their required role. Eosinophils are much longer lived than neutrophils, and unlike neutrophils, tissue eosinophils can recirculate. During most infections, eosinophils appear unimportant. However, in invasive helminthic infections, such as hookworm, schistosomiasis, strongyloidiasis, toxocariasis, trichinosis, filariasis, echinococcosis, and cysticercosis, the eosinophil plays a central role in host defense. Eosinophils are associated with bronchial asthma, cutaneous allergic reactions, and other hypersensitivity states.

The distinctive feature of the red-staining (Wright's stain) eosinophil granule is its crystalline core consisting of an arginine-rich protein (major basic protein) with histaminase activity, important in host defense against parasites. Eosinophil granules also contain a unique eosinophil peroxidase that catalyzes the oxidation of many substances by hydrogen peroxide and may facilitate killing of microorganisms.

Eosinophil peroxidase, in the presence of hydrogen peroxide and halide, initiates mast cell secretion in vitro and thereby promotes inflammation. Eosinophils contain cationic proteins, some of which bind to heparin and reduce its anticoagulant activity. Eosinophil-derived neurotoxin and eosinophil cationic protein are ribonucleases that can kill respiratory syncytial virus. Eosinophil cytoplasm contains Charcot-Leyden crystal protein, a hexagonal bipyramidal crystal first observed in a patient with leukemia and then in sputum of

patients with asthma; this protein is lysophospholipase and may function to detoxify certain lysophospholipids.

Several factors enhance the eosinophil's function in host defense. T cell–derived factors enhance the ability of eosinophils to kill parasites. Mast cell–derived eosinophil chemotactic factor of anaphylaxis (ECFa) increases the number of eosinophil complement receptors and enhances eosinophil killing of parasites. Eosinophil CSFs (e.g., IL-5) produced by macrophages increase eosinophil production in the bone marrow and activate eosinophils to kill parasites.

EOSINOPHILIA

Eosinophilia is the presence of >500 eosinophils per µL of blood and is common in many settings besides parasite infection. Significant tissue eosinophilia can occur without an elevated blood count. A common cause of eosinophilia is allergic reaction to drugs (iodides, aspirin, sulfonamides, nitrofurantoin, penicillins, and cephalosporins). Allergies such as hay fever, asthma, eczema, serum sickness, allergic vasculitis, and pemphigus are associated with eosinophilia. Eosinophilia also occurs in collagen vascular diseases (e.g., rheumatoid arthritis, eosinophilic fasciitis, allergic angiitis, and periarteritis nodosa) and malignancies (e.g., Hodgkin's disease; mycosis fungoides; chronic myeloid leukemia; and cancer of the lung, stomach, pancreas, ovary, or uterus), as well as in STAT3 deficient Job's syndrome, DOCK8 deficiency (see below), and CGD. Eosinophilia is commonly present in helminthic infections. IL-5 is the dominant eosinophil growth factor. Therapeutic administration of the cytokines IL-2 or GM-CSF frequently leads to transient eosinophilia. The most dramatic hypereosinophilic syndromes are Loeffler's syndrome, tropical pulmonary eosinophilia, Loeffler's endocarditis, eosinophilic leukemia, and idiopathic hypereosinophilic syndrome (50,000–100,000/µL). IL-5 is the dominant eosinophil growth factor and can be specifically inhibited with the monoclonal antibody mepolizumab.

The idiopathic hypereosinophilic syndrome represents a heterogeneous group of disorders with the common feature of prolonged eosinophilia of unknown cause and organ system dysfunction, including the heart, central nervous system, kidneys, lungs, gastrointestinal tract, and skin. The bone marrow is involved in all affected individuals, but the most severe complications involve the heart and central nervous system. Clinical manifestations and organ dysfunction are highly variable. Eosinophils are found in the involved tissues and likely cause tissue damage by local deposition of toxic eosinophil proteins such as eosinophil cationic protein and major basic protein. In the heart, the pathologic changes lead to thrombosis, endocardial fibrosis, and restrictive endomyocardiopathy. The damage to tissues in other organ systems is similar. Some cases are due to mutations involving the platelet-derived growth factor receptor, and these are extremely sensitive to the tyrosine kinase inhibitor imatinib. Glucocorticoids, hydroxyurea, and IFN-α each have been used successfully, as have therapeutic antibodies against IL-5. Cardiovascular complications are managed aggressively.

The *eosinophilia-myalgia syndrome* is a multisystem disease, with prominent cutaneous, hematologic, and visceral manifestations, that frequently evolves into a chronic course and can occasionally be fatal. The syndrome is characterized by eosinophilia (eosinophil count >1000/µL) and generalized disabling myalgias

without other recognized causes. Eosinophilic fasciitis, pneumonitis, and myocarditis; neuropathy culminating in respiratory failure; and encephalopathy may occur. The disease is caused by ingesting contaminants in L-tryptophan–containing products. Eosinophils, lymphocytes, macrophages, and fibroblasts accumulate in the affected tissues, but their role in pathogenesis is unclear. Activation of eosinophils and fibroblasts and the deposition of eosinophil-derived toxic proteins in affected tissues may contribute. IL-5 and transforming growth factor β have been implicated as potential mediators. Treatment is withdrawal of products containing L-tryptophan and the administration of glucocorticoids. Most patients recover fully, remain stable, or show slow recovery, but the disease can be fatal in up to 5% of patients.

Eosinophilic neoplasms are discussed in Chap. 106.

EOSINOPENIA

Eosinopenia occurs with stress, such as acute bacterial infection, and after treatment with glucocorticoids. The mechanism of eosinopenia of acute bacterial infection is unknown but is independent of endogenous glucocorticoids, because it occurs in animals after total adrenalectomy. There is no known adverse effect of eosinopenia.

HYPERIMMUNOGLOBULIN E–RECURRENT INFECTION SYNDROME

The hyperimmunoglobulin E–recurrent infection syndrome, or Job's syndrome, is a rare multisystem disease in which the immune and somatic systems are affected, including neutrophils, monocytes, T cells, B cells, and osteoclasts. Autosomal *dominant* inhibitory mutations in signal transducer and activator of transcription 3 (STAT3) lead to inhibition of normal STAT signaling with broad and profound effects. Patients have characteristic facies with broad nose, kyphoscoliosis, and eczema. The primary teeth erupt normally but do not deciduate, often requiring extraction. Patients develop recurrent sinopulmonary and cutaneous infections that tend to be much less inflamed than appropriate for the degree of infection and have been referred to as "cold abscesses." Characteristically, pneumonias cavitate, leading to pneumatoceles. Coronary artery aneurysms are common, as are cerebral demyelinated plaques that accumulate with age. Importantly, IL-17–producing T cells, which are thought responsible for protection against extracellular and mucosal infections, are profoundly reduced in Job's syndrome. Despite very high IgE levels, these patients have only mildly elevated levels of allergy. An important syndrome with clinical overlap with the dominant negative STAT3 deficiency is due to autosomal recessive defects in dedicator of cytokinesis 8 (DOCK8). In DOCK8 deficiency, IgE elevation is joined to severe allergy, viral susceptibility, and increased rates of cancer. Autosomal dominant *gain-of-function* mutations in STAT3 lead to a disease characterized by onset in childhood of lymphadenopathy, autoimmune cytopenias, multiorgan autoimmunity, infections, and interstitial lung disease.

LABORATORY DIAGNOSIS AND MANAGEMENT

Initial studies of WBC and differential and often a bone marrow examination may be followed by assessment of bone marrow reserves (steroid challenge test), marginated circulating pool of cells (epinephrine challenge test), and marginating ability (endotoxin challenge test) (Fig. 60-7). In vivo assessment of inflammation is possible with a Rebuck skin window test or an in vivo skin blister assay, which measures the ability of leukocytes and inflammatory mediators to accumulate locally in the skin. In vitro tests of phagocyte aggregation, adherence, chemotaxis, phagocytosis, degranulation, and microbicidal activity (for *S. aureus*) may help pinpoint cellular or humoral lesions. Deficiencies of oxidative metabolism are detected with either the nitroblue tetrazolium (NBT) dye test or the dihydrorhodamine (DHR) oxidation test. These tests are based on the ability of products of oxidative metabolism to alter the oxidation states of reporter molecules so that they can be detected microscopically (NBT) or by flow cytometry (DHR). Qualitative studies of superoxide and hydrogen peroxide production may further define neutrophil oxidative function.

Patients with leukopenias or leukocyte dysfunction often have delayed inflammatory responses. Therefore, clinical manifestations may be minimal despite overwhelming infection, and unusual infections must always be suspected. Early signs of infection demand prompt, aggressive culturing for microorganisms, use of antibiotics, and drainage of abscesses. Prolonged courses of antibiotics are often required. In patients with CGD, prophylactic antibiotics (trimethoprim-sulfamethoxazole) and antifungals (itraconazole) markedly diminish the frequency of life-threatening infections. Glucocorticoids may relieve gastrointestinal or genitourinary tract obstruction by granulomas in patients with CGD. Although TNF-α-blocking agents may markedly relieve inflammatory bowel symptoms, extreme caution must be exercised in their use in CGD inflammatory bowel disease, because it profoundly increases these patients' already heightened susceptibility to infection. Recombinant human IFN-γ, which nonspecifically stimulates phagocytic cell function, reduces the frequency of infections in patients with CGD by 70% and reduces the severity of infection. This effect of IFN-γ in CGD is additive to the effect of prophylactic antibiotics. The recommended dose is 50 $\mu g/m^2$ subcutaneously three times weekly. IFN-γ has also been used successfully in the treatment of leprosy, nontuberculous mycobacteria, and visceral leishmaniasis.

Rigorous oral hygiene reduces but does not eliminate the discomfort of gingivitis, periodontal disease, and aphthous ulcers; chlorhexidine mouthwash and tooth brushing with a hydrogen peroxide–sodium bicarbonate paste help many patients. Oral antifungal agents (fluconazole, itraconazole, voriconazole, posaconazole) have reduced mucocutaneous candidiasis in patients with Job's syndrome. Androgens, glucocorticoids, lithium, and immunosuppressive therapy have been used to restore myelopoiesis in patients with neutropenia due to impaired production. Recombinant G-CSF is useful in the management of certain forms of neutropenia due to depressed neutrophil production, including those related to cancer chemotherapy. Patients with chronic neutropenia with evidence of a good bone marrow reserve need not receive prophylactic antibiotics. Patients with chronic or cyclic neutrophil counts <500/μL may benefit from prophylactic antibiotics and G-CSF during periods of neutropenia. Oral trimethoprim-sulfamethoxazole (160/800 mg) twice daily can prevent infection. Increased numbers of fungal infections are not seen in patients with CGD on this regimen. Oral quinolones such as levofloxacin and ciprofloxacin are alternatives.

In the setting of cytotoxic chemotherapy with severe, persistent lymphocyte dysfunction, trimethoprim-sulfamethoxazole prevents *Pneumocystis jiroveci* pneumonia. These patients, and patients with phagocytic cell dysfunction, should avoid heavy exposure to airborne soil, dust, or decaying matter (mulch, manure), which are often rich in *Nocardia* and the spores of *Aspergillus* and other fungi. Restriction of activities or social contact has no proven role in reducing risk of infection for phagocyte defects.

Although aggressive medical care for many patients with phagocytic disorders can allow them to go for years without a life-threatening infection, there may still be delayed effects of prolonged antimicrobials and other inflammatory complications. Cure of most congenital phagocyte defects is possible by bone marrow transplantation, and rates of success are improving **(Chap. 110)**. The identification of specific gene defects in patients with LAD 1, CGD, and other immunodeficiencies has led to gene therapy trials in a number of genetic white cell disorders.

FURTHER READING

Casanova JL: Severe infectious diseases of childhood as monogenic inborn errors of immunity. Proc Natl Acad Sci USA 112:E7128, 2015. [PubMed: 26621750]

Kolaczkowska E, Kubes P: Neutrophil recruitment and function in health and inflammation. Nat Rev Immunol 13:159, 2013. [PubMed: 23435331]

Leiding JW et al (eds): GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993–2017. 2012 August 9 [updated 2016 February 11].

Lionakis MS et al: Immunity against fungi. JCI Insight 2: pii: 93156, 2017.

Moutsopoulos NM et al: Interleukin-12 and interleukin-23 blockade in leukocyte adhesion deficiency type 1. N Engl J Med 376:1141, 2017. [PubMed: 28328326]

Soehnlein O et al: Neutrophils as protagonists and targets in chronic inflammation. Nat Rev Immunol 17:248, 2017. [PubMed: 28287106]

Williams KW et al: Eosinophilia associated with disorders of immune deficiency or immune dysregulation. Immunol Allergy Clin North Am 35:523, 2015. [PubMed: 26209898]

Wu UI, Holland SM: Host susceptibility to non-tuberculous mycobacterial infections. Lancet Infect Dis 15:968, 2015. [PubMed: 26049967]

Chapter 61: Bleeding and Thrombosis

Barbara A. Konkle

INTRODUCTION

The human hemostatic system provides a natural balance between procoagulant and anticoagulant forces. The procoagulant forces include platelet adhesion and aggregation and fibrin clot formation; anticoagulant forces include the natural inhibitors of coagulation and fibrinolysis. Under normal circumstances, hemostasis is regulated to promote blood flow; however, it is also prepared to clot blood rapidly to arrest blood flow and prevent exsanguination. After bleeding is successfully halted, the system remodels the damaged vessel to restore normal blood flow. The major components of the hemostatic system, which function in concert, are (1) platelets and other formed elements of blood, such as monocytes and red cells; (2) plasma proteins (the coagulation and fibrinolytic factors and inhibitors); and (3) the vessel wall.

STEPS OF NORMAL HEMOSTASIS

PLATELET PLUG FORMATION

On vascular injury, platelets adhere to the site of injury, usually the denuded vascular intimal surface. Platelet adhesion is mediated primarily by Von Willebrand factor (VWF), a large multimeric protein present in both plasma and the extracellular matrix of the subendothelial vessel wall, which serves as the primary “molecular glue,” providing sufficient strength to withstand the high levels of shear stress that would tend to detach them with the flow of blood. Platelet adhesion is also facilitated by direct binding to subendothelial collagen through specific platelet membrane collagen receptors.

Platelet adhesion results in subsequent platelet activation and aggregation. This process is enhanced and amplified by humoral mediators in plasma (e.g., epinephrine, thrombin); mediators released from activated platelets (e.g., adenosine diphosphate, serotonin); and vessel wall extracellular matrix constituents that come in contact with adherent platelets (e.g., collagen, VWF). Activated platelets undergo the release reaction, during which they secrete contents that further promote aggregation and inhibit the naturally anticoagulant endothelial cell factors. During platelet aggregation (platelet-platelet interaction), additional platelets are recruited from the circulation to the site of vascular injury, leading to the formation of an occlusive platelet thrombus. The platelet plug is anchored and stabilized by the developing fibrin mesh.

The platelet glycoprotein (Gp) IIb/IIIa ($\alpha_{IIb}\beta_3$) complex is the most abundant receptor on the platelet surface.

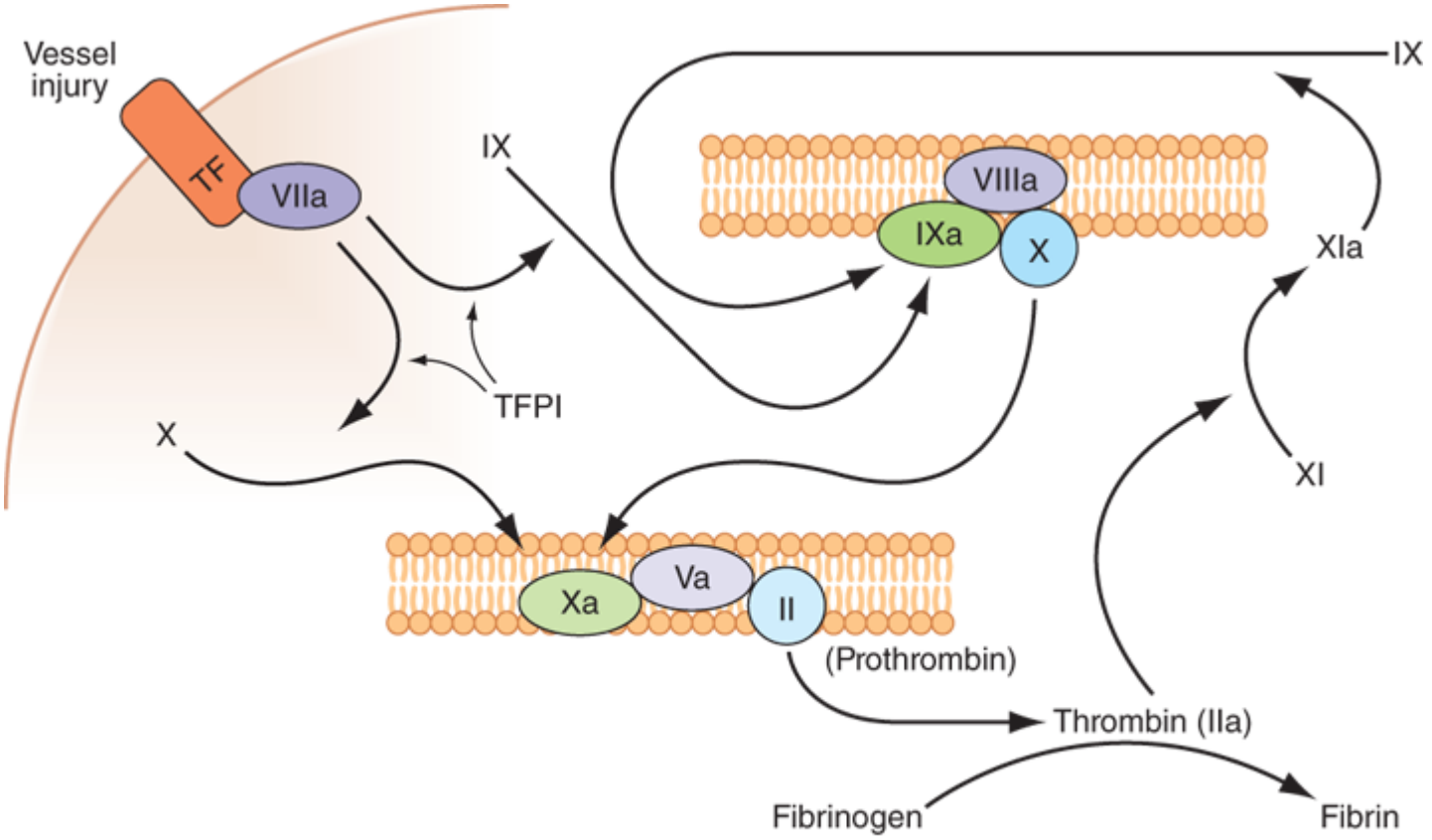
Platelet activation converts the normally inactive Gp IIb/IIIa receptor into an active receptor, enabling binding to fibrinogen and VWF. Because the surface of each platelet has about 50,000 Gp IIb/IIIa–binding sites, numerous activated platelets recruited to the site of vascular injury can rapidly form an occlusive aggregate by means of a dense network of intercellular fibrinogen bridges. Because this receptor is the key mediator of platelet aggregation, it has become an effective target for antiplatelet therapy.

FIBRIN CLOT FORMATION

Plasma coagulation proteins (*clotting factors*) normally circulate in plasma in their inactive forms. The sequence of coagulation protein reactions that culminate in the formation of fibrin was originally described as a *waterfall* or a *cascade*. Two pathways of blood coagulation have been described in the past: the so-called extrinsic, or tissue factor, pathway and the so-called intrinsic, or contact activation, pathway. We now know that coagulation is normally initiated through tissue factor (TF) exposure and activation through the classic *extrinsic pathway* but with critically important amplification through elements of the classic *intrinsic pathway*, as illustrated in **Fig. 61-1**. These reactions take place on phospholipid surfaces, usually the activated platelet surface. Coagulation testing in the laboratory can reflect other influences due to the artificial nature of the in vitro systems used (see below).

FIGURE 61-1

Coagulation is initiated by tissue factor (TF) exposure, which, with factor (F) VIIa, activates FIX and FX, which in turn, with FVIII and FV as cofactors, respectively, results in thrombin formation and subsequent conversion of fibrinogen to fibrin. Thrombin activates FXI, FVIII, and FV, amplifying the coagulation signal. Once the TF/FVIIa/FXa complex is formed, tissue factor pathway inhibitor (TFPI) inhibits the TF/FVIIa pathway, making coagulation dependent on the amplification loop through FIX/FVIII. Coagulation requires calcium (not shown) and takes place on phospholipid surfaces, usually the activated platelet membrane.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

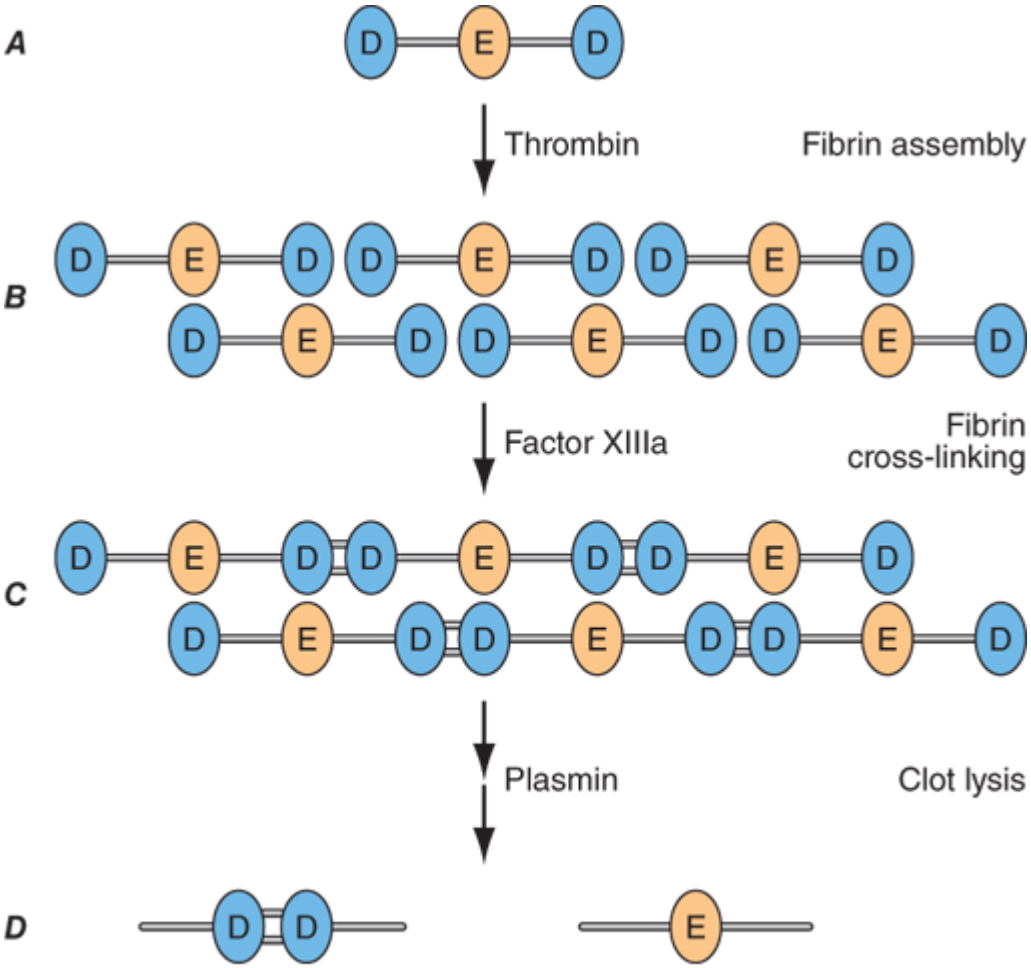
The immediate trigger for coagulation is vascular damage that exposes blood to TF that is constitutively expressed on the surfaces of subendothelial cellular components of the vessel wall, such as smooth muscle cells and fibroblasts. TF is also present in circulating microparticles, presumably shed from cells including monocytes and platelets. TF binds the serine protease factor VIIa; the complex activates factor X to factor Xa. Alternatively, the complex can indirectly activate factor X by initially converting factor IX to factor IXa, which then activates factor X. The participation of factor XI in hemostasis is not primarily dependent on its activation by factor XIIa but rather on its positive feedback activation by thrombin. Thus, factor XIa functions in the propagation and amplification, rather than in the initiation, of the coagulation cascade.

Factor Xa can be formed through the actions of either the TF/factor VIIa complex or factor IXa (with factor VIIIa as a cofactor) and converts prothrombin to thrombin, the pivotal protease of the coagulation system. The essential cofactor for this reaction is factor Va. Like the homologous factor VIIIa, factor Va is produced by thrombin-induced limited proteolysis of factor V. Thrombin is a multifunctional enzyme that converts soluble plasma fibrinogen to an insoluble fibrin matrix. Fibrin polymerization involves an orderly process of intermolecular associations **(Fig. 61-2)**. Thrombin also activates factor XIII (fibrin-stabilizing factor) to factor XIIIa, which covalently cross-links and thereby stabilizes the fibrin clot.

FIGURE 61-2

Fibrin formation and dissolution. (***A***) Fibrinogen is a trinodular structure consisting of two D domains and one E domain. Thrombin activation results in an ordered lateral assembly of protofibrils (***B***) with noncovalent associations. Factor XIIIa cross-links the D domains on adjacent molecules (***C***). Fibrin and fibrinogen (not

shown) lysis by plasmin occurs at discrete sites and results in intermediary fibrin(ogen) degradation products (not shown). D-Dimers are the product of complete lysis of fibrin (***D***), maintaining the cross-linked D domains.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The assembly of the clotting factors on activated cell membrane surfaces greatly accelerates their reaction rates and also serves to localize blood clotting to sites of vascular injury. The critical cell membrane components, acidic phospholipids, are not normally exposed on resting cell membrane surfaces. However, when platelets, monocytes, and endothelial cells are activated by vascular injury or inflammatory stimuli, the procoagulant head groups of the membrane anionic phospholipids become translocated to the surfaces of these cells or released as part of microparticles, making them available to support and promote the plasma coagulation reactions.

ANTITHROMBOTIC MECHANISMS

Several physiologic antithrombotic mechanisms act in concert to prevent clotting under normal circumstances. These mechanisms operate to preserve blood fluidity and to limit blood clotting to specific focal sites of vascular injury. Endothelial cells have many antithrombotic effects. They produce prostacyclin, nitric oxide, and ectoADPase/CD39, which act to inhibit platelet binding, secretion, and aggregation. Endothelial cells produce anticoagulant factors including heparan proteoglycans, antithrombin, TF pathway

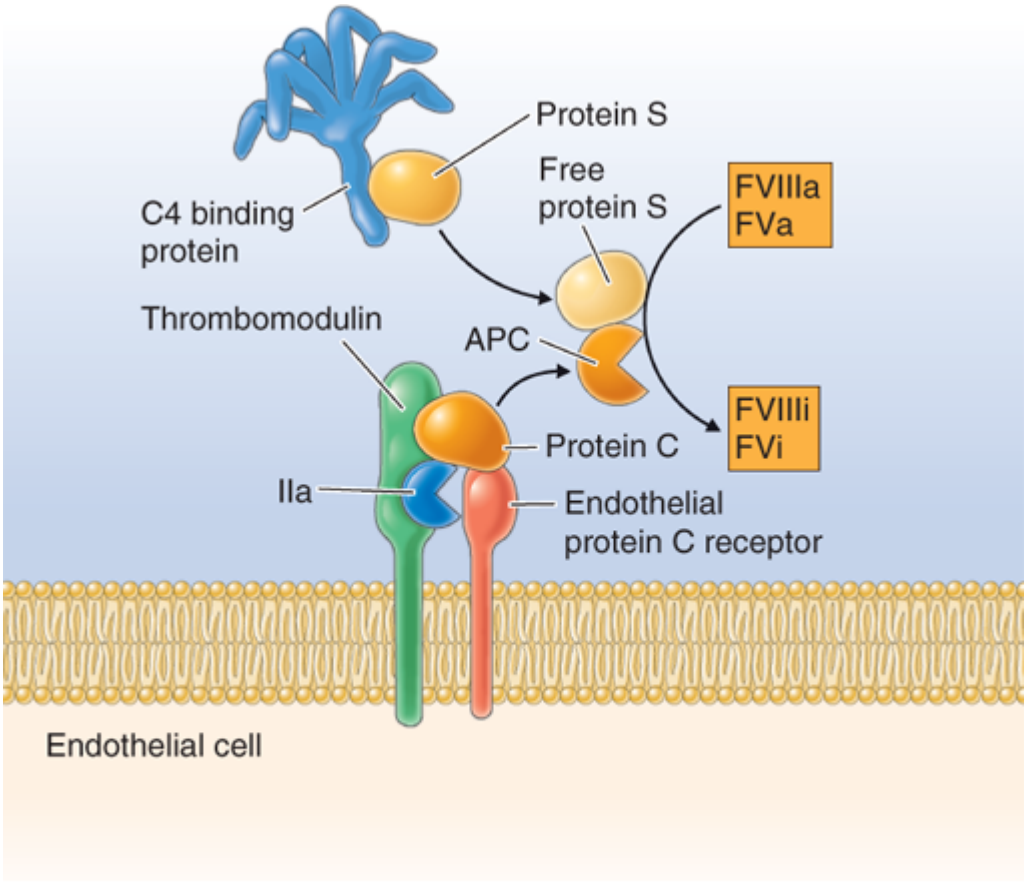
inhibitor, and thrombomodulin. They also activate fibrinolytic mechanisms through the production of tissue plasminogen activator 1, urokinase, plasminogen activator inhibitor, and annexin-2.

Antithrombin is the major plasma protease inhibitor of thrombin and the other clotting factors in coagulation. Antithrombin neutralizes thrombin and other activated coagulation factors by forming a complex between the active site of the enzyme and the reactive center of antithrombin. The rate of formation of these inactivating complexes increases by a factor of several thousand in the presence of heparin. Antithrombin inactivation of thrombin and other activated clotting factors occurs physiologically on vascular surfaces, where glycosoaminoglycans, including heparan sulfates, are present to catalyze these reactions. Inherited quantitative or qualitative deficiencies of antithrombin lead to a lifelong predisposition to venous thromboembolism.

Protein C is a plasma glycoprotein that becomes an anticoagulant when it is activated by thrombin. The thrombin-induced activation of protein C occurs physiologically on thrombomodulin, a transmembrane proteoglycan-binding site for thrombin on endothelial cell surfaces. The binding of protein C to its receptor on endothelial cells places it in proximity to the thrombin-thrombomodulin complex, thereby enhancing its activation efficiency. **(See Fig. 61-3.)** Activated protein C acts as an anticoagulant by cleaving and inactivating activated factors V and VIII. This reaction is accelerated by a cofactor, protein S, which, like protein C, is a glycoprotein that undergoes vitamin K–dependent posttranslational modification. Quantitative or qualitative deficiencies of protein C or protein S, or resistance to the action of activated protein C by a specific mutation at its target cleavage site in factor Va (factor V Leiden), lead to hypercoagulable states.

FIGURE 61-3

The activated protein C pathway in regulation of thrombosis. Thrombin generation results in protein C activation through interaction with thrombomodulin and protein C bound to the endothelial protein C receptor (EPCR). Activated protein C (APC) with free protein S converts activated factors (F) VIII and V to inactivate forms, thus in turn decreasing thrombin generation. APC, activated protein C; C4BP, C4 binding protein; EC, endothelial cell; EPCR, endothelial protein C receptor; F, factor; IIa, thrombin; PC, protein C; PS, protein S; TM, thrombomodulin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Tissue factor pathway inhibitor (TFPI) is a plasma protease inhibitor that regulates the TF-induced extrinsic pathway of coagulation. TFPI inhibits the TF/factor VIIa/factor Xa complex, essentially turning off the TF/factor VIIa initiation of coagulation, which then becomes dependent on the "amplification loop" via factor XI and factor VIII activation by thrombin. TFPI is bound to lipoprotein and can also be released by heparin from endothelial cells, where it is bound to glycosaminoglycans, and from platelets. The heparin-mediated release of TFPI may play a role in the anticoagulant effects of unfractionated and low-molecular-weight heparins (LMWH).

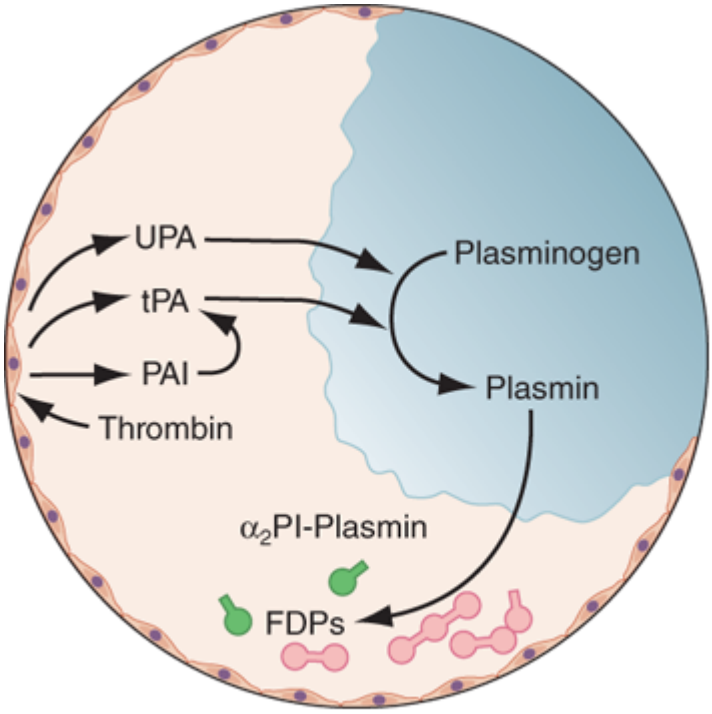
THE FIBRINOLYTIC SYSTEM

Any thrombin that escapes the inhibitory effects of the physiologic anticoagulant systems is available to convert fibrinogen to fibrin. In response, the endogenous fibrinolytic system is then activated to dispose of intravascular fibrin and thereby maintain or reestablish the patency of the circulation. Just as thrombin is the key protease enzyme of the coagulation system, plasmin is the major protease enzyme of the fibrinolytic system, acting to digest fibrin to fibrin degradation products. The general scheme of fibrinolysis and its control is shown in **Fig. 61-4**.

FIGURE 61-4

A schematic diagram of the fibrinolytic system. Tissue plasminogen activator (tPA) is released from endothelial cells, binds the fibrin clot, and activates plasminogen to plasmin. Excess fibrin is degraded by

plasmin to distinct degradation products (FDPs). Any free plasmin is complexed with α_2-antiplasmin (α_2Pl). PAI, plasminogen activator inhibitor; UPA, urokinase-type plasminogen activator.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The plasminogen activators, tissue type plasminogen activator (tPA) and the urokinase-type plasminogen activator (uPA), cleave the Arg560-Val561 bond of plasminogen to generate the active enzyme plasmin. The lysine-binding sites of plasmin (and plasminogen) permit it to bind to fibrin, so that physiologic fibrinolysis is "fibrin specific." Both plasminogen (through its lysine-binding sites) and tPA possess specific affinity for fibrin and thereby bind selectively to clots. The assembly of a ternary complex, consisting of fibrin, plasminogen, and tPA, promotes the localized interaction between plasminogen and tPA and greatly accelerates the rate of plasminogen activation to plasmin. Moreover, partial degradation of fibrin by plasmin exposes new plasminogen and tPA-binding sites in carboxy-terminus lysine residues of fibrin fragments to enhance these reactions further. This creates a highly efficient mechanism to generate plasmin focally on the fibrin clot, which then becomes plasmin's substrate for digestion to fibrin degradation products.

Plasmin cleaves fibrin at distinct sites of the fibrin molecule, leading to the generation of characteristic fibrin fragments during the process of fibrinolysis (Fig. 61-2). The sites of plasmin cleavage of fibrin are the same as those in fibrinogen. However, when plasmin acts on covalently cross-linked fibrin, D-dimers are released; hence, D-dimers can be measured in plasma as a relatively specific test of fibrin (rather than fibrinogen) degradation. D-Dimer assays can be used as sensitive markers of blood clot formation and have been validated for clinical use to exclude the diagnosis of deep-venous thrombosis (DVT) and pulmonary embolism in selected populations. In addition, D-dimer measurement can be used to stratify patients, particularly women, for risk of recurrent venous thromboembolism (VTE) when measured 1 month after

discontinuation of anticoagulation given for treatment of an initial idiopathic event. D-Dimer levels increase with age. Whether a higher cut-off should be used in the elderly is controversial.

Physiologic regulation of fibrinolysis occurs primarily at three levels: (1) plasminogen activator inhibitors (PAIs), specifically PAI-1 and PAI-2, inhibit the physiologic plasminogen activators; (2) the thrombin-activatable fibrinolysis inhibitor (TAFI) limits fibrinolysis; and (3) α_2-antiplasmin inhibits plasmin. PAI-1 is the primary inhibitor of tPA and uPA in plasma. TAFI cleaves the N-terminal lysine residues of fibrin, which aid in localization of plasmin activity. α_2-Antiplasmin is the main inhibitor of plasmin in human plasma, inactivating any nonfibrin clot-associated plasmin.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Bleeding and Thrombosis

CLINICAL PRESENTATION

Disorders of hemostasis may be either inherited or acquired. A detailed personal and family history is key in determining the chronicity of symptoms and the likelihood of the disorder being inherited, as well as providing clues to underlying conditions that have contributed to the bleeding or thrombotic state. In addition, the history can give clues as to the etiology by determining (1) the bleeding (mucosal and/or joint) or thrombosis (arterial and/or venous) site and (2) whether an underlying bleeding or clotting tendency was enhanced by another medical condition or the introduction of medications or dietary supplements.

History of Bleeding

A history of bleeding is the most important predictor of bleeding risk. In evaluating a patient for a bleeding disorder, a history of at-risk situations, including the response to past surgeries, should be assessed. Does the patient have a history of spontaneous or trauma/surgery-induced bleeding? Spontaneous hemarthroses are a hallmark of moderate and severe factor VIII and IX deficiency and, in rare circumstances, of other clotting factor deficiencies. Mucosal bleeding symptoms are more suggestive of underlying platelet disorders or Von Willebrand disease (VWD), termed *disorders of primary hemostasis or platelet plug formation.* Disorders affecting primary hemostasis are shown in **Table 61-1**.

A bleeding score has been validated as a tool to predict patients more likely to have type 1 VWD (International Society on Thrombosis and Haemostasis Bleeding Assessment Tool [*www.isth.org/resource/resmgr/ssc/isth-ssc_bleeding_assessment.pdf*]). This is the most useful tool in excluding the diagnosis of a bleeding disorder, and thus avoiding unnecessary testing. One study found that a low bleeding score (≤3) and a normal activated partial thromboplastin time (aPTT) had 99.6% negative predictive value for the diagnosis of VWD. Bleeding symptoms that appear to be more common in patients with bleeding disorders include prolonged bleeding with surgery, dental procedures and extractions, and/or trauma, heavy menstrual bleeding (HMB), or postpartum hemorrhage (PPH), and large bruises (often described with lumps).

Easy bruising and HMB are common complaints in patients with and without bleeding disorders. Easy bruising can also be a sign of medical conditions in which there is no identifiable coagulopathy; instead, the conditions are caused by an abnormality of blood vessels or their supporting tissues. In Ehlers-Danlos syndrome, there may be posttraumatic bleeding and a history of joint hyperextensibility. Cushing's syndrome, chronic steroid use, and aging result in changes in skin and subcutaneous tissue, and subcutaneous bleeding occurs in response to minor trauma. The latter has been termed *senile purpura*.

Epistaxis is a common symptom, particularly in children and in dry climates, and may not reflect an underlying bleeding disorder. However, it is the most common symptom in hereditary hemorrhagic telangiectasia and in boys with VWD. Clues that epistaxis is a symptom of an underlying bleeding disorder include lack of seasonal variation and bleeding that requires medical evaluation or treatment, including cauterization. Bleeding with eruption of primary teeth is seen in children with more severe bleeding disorders, such as moderate and severe hemophilia. It is uncommon in children with mild bleeding disorders. Patients with disorders of primary hemostasis (platelet adhesion) may have increased bleeding after dental cleanings and other procedures that involve gum manipulation.

Heavy menstrual bleeding is defined quantitatively as a loss of >80 mL of blood per cycle, based on the quantity of blood loss required to produce iron-deficiency anemia. A complaint of heavy menses is subjective and has a poor correlation with excessive blood loss. Predictors of HMB include bleeding resulting in iron-deficiency anemia or a need for blood transfusion, passage of clots >1 in. in diameter, and changing a pad or tampon more than hourly. HMB is a common symptom in women with underlying bleeding disorders and is reported in the majority of women with VWD, women with factor XI deficiency, and symptomatic carriers of hemophilia. Women with underlying bleeding disorders are more likely to have other bleeding symptoms, including bleeding after dental extractions, postoperative bleeding, and postpartum bleeding, and are much more likely to have HMB beginning at menarche than women with HMB due to other causes.

PPH is a common symptom in women with underlying bleeding disorders. In women with type 1 VWD and symptomatic carriers of hemophilia A in whom levels of VWF and factor VIII usually normalize during pregnancy, PPH may be delayed. Women with a history of PPH may have a higher risk of recurrence with subsequent pregnancies. Rupture of ovarian cysts with intraabdominal hemorrhage has also been reported in women with underlying bleeding disorders.

Tonsillectomy is a major hemostatic challenge, because intact hemostatic mechanisms are essential to prevent excessive bleeding from the tonsillar bed. Bleeding may occur early after surgery or after ~7 days postoperatively, with loss of the eschar at the operative site. Similar delayed bleeding is seen after colonic polyp resection. Gastrointestinal (GI) bleeding and hematuria are usually due to underlying pathology, and procedures to identify and treat the bleeding site should be undertaken, even in patients with known bleeding disorders. VWD, particularly types 2 and 3, has been associated with angiodysplasia of the bowel and GI bleeding.

Hemarthroses and spontaneous muscle hematomas are characteristic of moderate or severe congenital factor VIII or IX deficiency. They can also be seen in moderate and severe deficiencies of fibrinogen,

prothrombin, and factors V, VII, and X. Spontaneous hemarthroses occur rarely in other bleeding disorders except for severe VWD, with associated factor VIII levels <5%. Muscle and soft tissue bleeds are also common in acquired factor VIII deficiency. Bleeding into a joint results in severe pain and swelling, as well as loss of function, but is rarely associated with discoloration from bruising around the joint. Life-threatening sites of bleeding include bleeding into the oropharynx, where bleeding can obstruct the airway, into the central nervous system, and into the retroperitoneum. Central nervous system bleeding is the major cause of bleeding-related deaths in patients with severe congenital factor deficiencies.

Prohemorrhagic Effects of Medications and Dietary Supplements

Aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDs) that inhibit cyclooxygenase 1 impair primary hemostasis and may exacerbate bleeding from another cause or even unmask a previously occult mild bleeding disorder such as VWD. All NSAIDs, however, can precipitate GI bleeding, which may be more severe in patients with underlying bleeding disorders. This aspirin effect lasts for the life of the platelet, though in individuals with typical platelet turnover, the functional defect reverts to near-normal within 2–3 days after the last dose. The effect of other NSAIDs is shorter, as the inhibitor effect is reversed when the drug is removed. Inhibitors of the ADP $P2Y_{12}$ receptor (clopidogrel, prasugrel, and ticagrelor) inhibit ADP-mediated platelet aggregation and, like NSAIDs, can precipitate or exacerbate bleeding symptoms. The risk of bleeding with these drugs is higher than with NSAIDs.

Many herbal supplements can impair hemostatic function **(Table 61-2)**. Some are more convincingly associated with a bleeding risk than others. Fish oil or concentrated omega-3 fatty acid supplements impair platelet function. They alter platelet biochemistry to produce more PGI_3, a more potent platelet inhibitor than prostacyclin (PGI_2), and more thromboxane A_3, a less potent platelet activator than thromboxane A_2. In fact, diets naturally rich in omega-3 fatty acids can result in a prolonged bleeding time and abnormal platelet aggregation studies, but the actual associated bleeding risk is unclear. Vitamin E appears to inhibit protein kinase C–mediated platelet aggregation and nitric oxide production. In patients with unexplained bruising or bleeding, it is prudent to review any new medications or supplements and discontinue those that may be associated with bleeding.

Underlying Systemic Diseases that Cause or Exacerbate a Bleeding Tendency

Acquired bleeding disorders are commonly secondary to, or associated with, systemic disease. The clinical evaluation of a patient with a bleeding tendency must therefore include a thorough assessment for evidence of underlying disease. Bruising or mucosal bleeding may be the presenting complaint in liver disease, severe renal impairment, hypothyroidism, paraproteinemias or amyloidosis, and conditions causing bone marrow failure. All coagulation factors are synthesized in the liver, and hepatic failure results in combined factor deficiencies. This is often compounded by thrombocytopenia associated with liver failure and portal hypertension. Coagulation factors II, VII, IX, and X and proteins C, S, and Z are dependent on vitamin K for posttranslational modification. Although vitamin K is required in both procoagulant and anticoagulant processes, the phenotype of vitamin K deficiency or the warfarin effect on coagulation is bleeding.

The normal blood platelet count is 150,000–450,000/μL. Thrombocytopenia results from decreased production, increased destruction, and/or sequestration. Although the bleeding risk varies somewhat by the reason for the thrombocytopenia, bleeding rarely occurs in isolated thrombocytopenia at counts >50,000/μL and usually not until <10,000–20,000/μL. Coexisting coagulopathies, as is seen in liver failure or disseminated coagulation; infection; platelet-inhibitory drugs; and underlying medical conditions can all increase the risk of bleeding in the thrombocytopenic patient. Most procedures can be performed in patients with a platelet count of 50,000/μL. The level needed for major surgery will depend on the type of surgery and the patient's underlying medical state, although a count of ~80,000/μL is likely sufficient.

HISTORY OF THROMBOSIS

The risk of thrombosis, like that of bleeding, is influenced by both genetic and environmental influences. The major risk factor for arterial thrombosis is atherosclerosis, whereas for venous thrombosis, the risk factors are immobility, surgery, underlying medical conditions such as malignancy, medications such as hormonal therapy, obesity, and genetic predispositions. Factors that increase risks for venous and for both venous and arterial thromboses are shown in **Table 61-3**.

The most important point in a history related to venous thrombosis is determining whether the thrombotic event was idiopathic (meaning there was no clear precipitating factor) or was a precipitated event. In patients without underlying malignancy, having an idiopathic event is the strongest predictor of recurrence of VTE. In patients who have a vague history of thrombosis, a history of being treated with warfarin suggests a past DVT. Age is an important risk factor for venous thrombosis—the risk of DVT increases per decade, with an approximate incidence of 1/100,000 per year in early childhood to 1/200 per year among octogenarians. Family history is helpful in determining if there is a genetic predisposition and how strong that predisposition appears to be. A genetic thrombophilia that confers a relatively small increased risk, such as being a heterozygote for the prothrombin G20210A or factor V Leiden mutation, is a minor determinant of risk in an elderly individual undergoing a high-risk surgical procedure. As illustrated in **Fig. 61-5**, a thrombotic event usually has more than one contributing factor. Predisposing factors must be carefully assessed to determine the risk of recurrent thrombosis and, with consideration of the patient's bleeding risk, determine the length of anticoagulation. Testing for inherited thrombophilias in adults should be limited to instances where results would change clinical care.

LABORATORY EVALUATION

Careful history taking and clinical examination are essential components in the assessment of bleeding and thrombotic risk. The use of laboratory tests of coagulation complement, but cannot substitute for, clinical assessment. No test exists that provides a global assessment of hemostasis. The bleeding time has been used to assess bleeding risk; however, it does not predict bleeding risk with surgery and it is not recommended for this indication. The PFA-100, an instrument that measures platelet-dependent coagulation under flow conditions, is more sensitive and specific for VWD than the bleeding time; however, it is not sensitive enough to rule out mild bleeding disorders. PFA-100 closure times are prolonged in patients with some, but not all, inherited platelet disorders. Also, its utility in predicting bleeding risk has not been determined.

Thromboelastography can be useful in guiding intraoperative transfusion but is not broadly applicable for the diagnosis of disorders of hemostasis and thrombosis.

For routine preoperative and pre-procedure testing, an abnormal prothrombin time (PT) may detect liver disease or vitamin K deficiency that had not been previously appreciated. Studies have not confirmed the usefulness of an aPTT in preoperative evaluations in patients with a negative bleeding history. The primary use of coagulation testing should be to confirm the presence and type of bleeding disorder in a patient with a suspicious clinical history.

Because of the nature of coagulation assays, proper sample acquisition and handling is critical to obtaining valid results. In patients with abnormal coagulation assays who have no bleeding history, repeat studies with attention to these factors frequently result in normal values. Most coagulation assays are performed in sodium citrate anticoagulated plasma that is recalcified for the assay. Because the anticoagulant is in liquid solution and needs to be added to blood in proportion to the plasma volume, incorrectly filled or inadequately mixed blood collection tubes will give erroneous results. Vacutainer tubes should be filled to >90% of the recommended fill, which is usually denoted by a line on the tube. An elevated hematocrit (>55%) can result in a false value due to a decreased plasma-to-anticoagulant ratio.

Screening Assays

The most commonly used screening tests are the PT, aPTT, and platelet count. The PT assesses the factors I (fibrinogen), II (prothrombin), V, VII, and X **(Fig. 61-6)**. The PT measures the time for clot formation of the citrated plasma after recalcification and addition of thromboplastin, a mixture of TF and phospholipids. The sensitivity of the assay varies by the source of thromboplastin. The relationship between defects in secondary hemostasis (fibrin formation) and coagulation test abnormalities is shown in **Table 61-4**. To adjust for this variability, the overall sensitivity of different thromboplastins to reduction of the vitamin K–dependent clotting factors II, VII, IX, and X in anticoagulation patients is expressed as the International Sensitivity Index (ISI). The international normalized ratio (INR) is determined based on the formula: INR = $(PT_{patient}/PT_{normal\ mean})^{ISI}$.

The INR was developed to assess stable anticoagulation due to reduction of vitamin K–dependent coagulation factors; it is commonly used in the evaluation of patients with liver disease. Although it does allow comparison between laboratories, reagent sensitivity as used to determine the ISI is not the same in liver disease as with warfarin anticoagulation. In addition, progressive liver failure is associated with variable changes in coagulation factors; the degree of prolongation of either the PT or the INR only roughly predicts the bleeding risk. Thrombin generation has been shown to be normal in many patients with mild to moderate liver dysfunction. Because the PT only measures one aspect of hemostasis affected by liver dysfunction, we likely overestimate the bleeding risk of a mildly elevated INR in this setting. PT reagents have variable sensitivity to the direct Xa inhibitors and the PT is usually normal in patients on apixaban.

The aPTT assesses the intrinsic and common coagulation pathways; factors XI, IX, VIII, X, V, and II; fibrinogen; prekallikrein; high-molecular-weight kininogen; and factor XII (Fig. 61-6). The aPTT reagent contains phospholipids derived from either animal or vegetable sources that function as a platelet substitute in the

coagulation pathways and includes an activator of the intrinsic coagulation system, such as nonparticulate ellagic acid or the particulate activators kaolin, celite, or micronized silica.

The phospholipid composition of aPTT reagents varies, which influences the sensitivity of individual reagents to clotting factor deficiencies and to inhibitors such as heparin and lupus anticoagulants. Thus, aPTT results will vary from one laboratory to another, and the normal range in the laboratory where the testing occurs should be used in the interpretation. Local laboratories can relate their aPTT values to the therapeutic heparin anticoagulation by correlating aPTT values with direct measurements of heparin activity (anti-Xa or protamine titration assays) in samples from heparinized patients, although correlation between these assays is often poor. The aPTT reagent will vary in sensitivity to individual factor deficiencies and usually becomes prolonged with individual factor deficiencies of 30–50%.

Mixing Studies

Mixing studies are used to evaluate a prolonged aPTT or, less commonly PT, to distinguish between a factor deficiency and an inhibitor. In this assay, normal plasma and patient plasma are mixed in a 1:1 ratio, and the aPTT or PT is determined immediately and after incubation at 37°C for varying times, typically 30, 60, and/or 120 min. With isolated factor deficiencies, the aPTT will correct with mixing and stay corrected with incubation. With aPTT prolongation due to a lupus anticoagulant, the mixing and incubation will show no correction. In acquired neutralizing factor antibodies, notably an acquired factor VIII inhibitor, the initial assay may or may not correct immediately after mixing but will prolong or remain prolonged with incubation at 37°C. Failure to correct with mixing can also be due to the presence of other inhibitors or interfering substances such as heparin, fibrin split products, and paraproteins.

Specific Factor Assays

Decisions to proceed with specific clotting factor assays will be influenced by the clinical situation and the results of coagulation screening tests. Precise diagnosis and effective management of inherited and acquired coagulation deficiencies necessitate quantitation of the relevant factors. When bleeding is severe, specific assays are urgently required to guide appropriate therapy. Individual factor assays are usually performed as modifications of the mixing study, where the patient's plasma is mixed with plasma deficient in the factor being studied. This will correct all factor deficiencies to >50%, thus making prolongation of clot formation due to a factor deficiency dependent on the factor missing from the added plasma.

Testing for Antiphospholipid Antibodies

Antibodies to phospholipids (cardiolipin) or phospholipid-binding proteins (β_2-microglobulin and others) are detected by enzyme-linked immunosorbent assay (ELISA). When these antibodies interfere with phospholipid-dependent coagulation tests, they are termed *lupus anticoagulants*. The aPTT has variability sensitivity to lupus anticoagulants, depending in part on the aPTT reagents used. An assay using a sensitive reagent has been termed an *LA-PTT*. The dilute Russell viper venom test (dRVVT) and the tissue thromboplastin inhibition (TTI) test are modifications of standard tests with the phospholipid reagent decreased, thus increasing the sensitivity to antibodies that interfere with the phospholipid component. The tests, however, are not specific for lupus anticoagulants, because factor deficiencies or other inhibitors will

also result in prolongation. Documentation of a lupus anticoagulant requires not only prolongation of a phospholipid-dependent coagulation test but also lack of correction when mixed with normal plasma and correction with the addition of activated platelet membranes or certain phospholipids (e.g., hexagonal phase).

Other Coagulation Tests

The thrombin time and the reptilase time measure fibrinogen conversion to fibrin and are prolonged when the fibrinogen level is low (usually <80–100 mg/dL) or qualitatively abnormal, as seen in inherited or acquired dysfibrinogenemias, or when fibrin/fibrinogen degradation products interfere. The thrombin time, but not the reptilase time, is prolonged in the presence of heparin. The thrombin time is markedly prolonged in the presence of the direct thrombin inhibitor, dabigatran; a dilute thrombin time can be used to assess drug activity. Measurement of anti–factor Xa plasma inhibitory activity is a test frequently used to assess LMWH levels, as a direct measurement of unfractionated heparin (UFH) activity, or to assess activity of the direct Xa inhibitors rivaroxaban, apixaban, and edoxaban. Drug in the patient sample inhibits the enzymatic conversion of an Xa-specific chromogenic substrate to colored product by factor Xa. Standard curves are created using multiple concentrations of the specific drug and are used to calculate the concentration of anti-Xa activity in the patient plasma.

Laboratory Testing for Thrombophilia

Laboratory assays to detect thrombophilic states include molecular diagnostics and immunologic and functional assays. These assays vary in their sensitivity and specificity for the condition being tested. Furthermore, acute thrombosis, acute illnesses, inflammatory conditions, pregnancy, and medications affect levels of many coagulation factors and their inhibitors. Antithrombin is decreased by heparin and in the setting of acute thrombosis. Protein C and S levels may be increased in the setting of acute thrombosis and are decreased by warfarin. Antiphospholipid antibodies are frequently transiently positive in acute illness. Testing for genetic thrombophilias should, in general, only be performed when there is a strong family history of thrombosis and results would affect clinical decision-making.

Because thrombophilia evaluations are usually performed to assess the need to extend anticoagulation, testing, if indicated, should be performed in a steady state, remote from the acute event. In most instances, warfarin anticoagulation can be stopped after the initial 3–6 months of treatment, and testing can be performed at least 3 weeks later. As a sensitive marker of coagulation activation, the quantitative D-dimer assay, drawn 4 weeks after stopping anticoagulation, can be used to stratify risk of recurrent thrombosis in patients, particularly women, who have an idiopathic event.

Measures of Platelet Function

The bleeding time has been used to assess bleeding risk; however, it has not been found to predict bleeding risk with surgery, and it is not recommended for use for this indication. The PFA-100 and similar instruments that measure platelet-dependent coagulation under flow conditions are generally more sensitive and specific for platelet disorders and VWD than the bleeding time; however, data are insufficient to support their use to predict bleeding risk or monitor response to therapy, and they will be normal in some patients with platelet

disorders or mild VWD. When they are used in the evaluation of a patient with bleeding symptoms, abnormal results, as with the bleeding time, require specific testing, such as VWF assays and/or platelet aggregation studies. Because all of these “screening” assays may miss patients with mild bleeding disorders, further studies are needed to define their role in hemostasis testing.

For classic platelet aggregometry, various agonists are added to the patient’s platelet-rich plasma or whole blood and platelet aggregation is measured. Tests of platelet secretion in response to agonists can also be measured. These tests are affected by many factors, including numerous medications, and the association between minor defects in aggregation or secretion in these assays and bleeding risk is not clearly established.

TABLE 61-1

Primary Hemostatic (Platelet Plug) Disorders

Defects of Platelet Adhesion
Von Willebrand disease
Bernard-Soulier syndrome (absence or dysfunction of platelet Gp Ib-IX-V)
Defects of Platelet Aggregation
Glanzmann's thrombasthenia (absence or dysfunction of platelet glycoprotein [Gp] IIb/IIIa)
Afibrinogenemia
Defects of Platelet Secretion
Decreased cyclooxygenase activity
Drug-induced (aspirin, nonsteroidal anti-inflammatory agents, thienopyridines)
Inherited
Granule storage pool defects
Inherited
Acquired
Nonspecific inherited secretory defects
Nonspecific drug effects
Uremia
Platelet coating (e.g., paraprotein, penicillin)
Defect of Platelet Coagulant Activity
Scott's syndrome

TABLE 61-2

Herbal Supplements Associated with Increased Bleeding

Herbs with Potential Antiplatelet Activity
Ginkgo (*Ginkgo biloba L.*)
Garlic (*Allium sativum*)
Bilberry (*Vaccinium myrtillus*)
Ginger (*Gingiber officinale*)
Dong quai (*Angelica sinensis*)
Feverfew (*Tanacetum parthenium*)
Asian ginseng (*Panax ginseng*)
American ginseng (*Panax quinquefolius*)
Siberian ginseng/eleuthero (*Eleutherococcus senticosus*)
Turmeric (*Circuma longa*)
Meadowsweet (*Filipendula ulmaria*)
Willow (*Salix* spp.)
Coumarin-Containing Herbs
Motherwort (*Leonurus cardiaca*)
Chamomile (*Matricaria recutita, Chamaemelum mobile*)
Horse chestnut (*Aesculus hippocastanum*)
Red clover (*Trifolium pratense*)
Fenugreek (*Trigonella foenum-graecum*)

TABLE 61-3

Risk Factors for Thrombosis

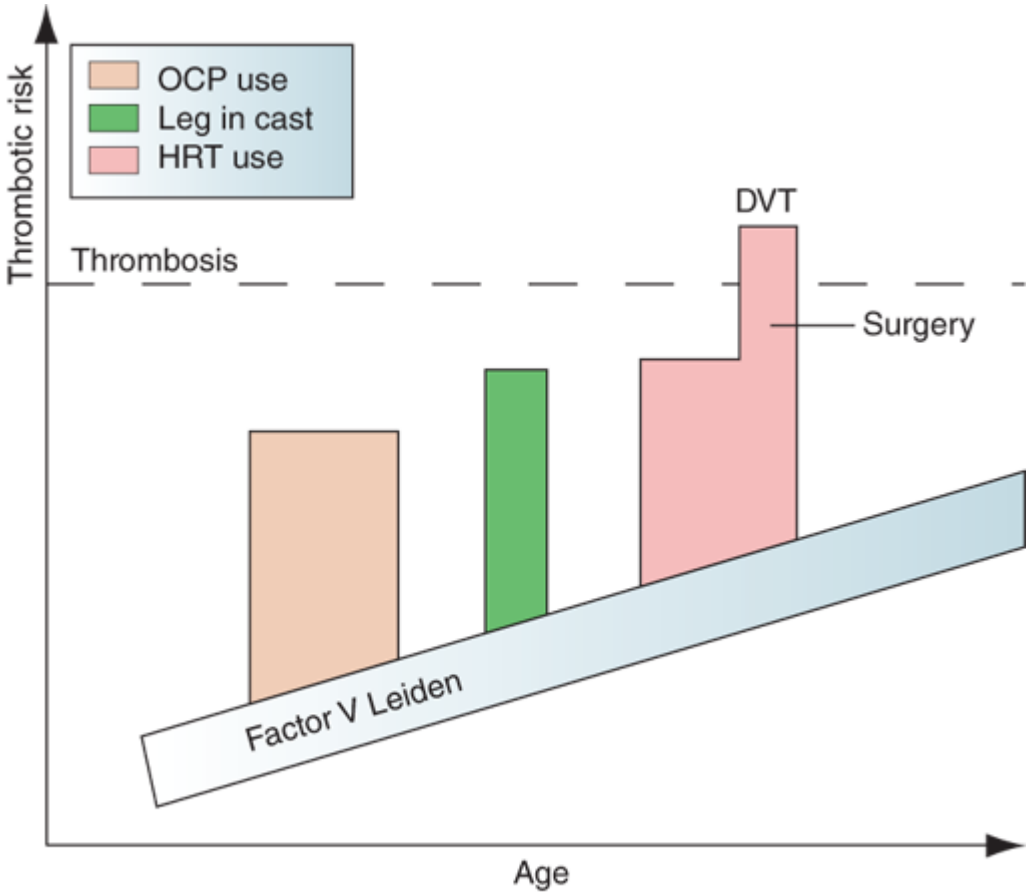
Venous	Venous and Arterial
Inherited	**Inherited**
Factor V Leiden	Homocystinuria
Prothrombin G20210A	Dysfibrinogenemia
Antithrombin deficiency	**Acquired**
Protein C deficiency	Malignancy
Protein S deficiency	Antiphospholipid antibody syndrome
Elevated factor VIII	Hormonal therapy
Acquired	Polycythemia vera
Age	Essential thrombocythemia
Previous thrombosis	Paroxysmal nocturnal hemoglobinuria
Immobilization	Thrombotic thrombocytopenic purpura
Major surgery	Heparin-induced thrombocytopenia
Pregnancy and puerperium	Disseminated intravascular coagulation
Hospitalization	**Unknown**[a]
Obesity	Elevated factor II, IX, XI
Infection	Elevated TAFI levels
APC resistance, nongenetic	Low levels of TFPI
Smoking	

[a]Unknown whether risk is inherited or acquired.

Abbreviations: APC, activated protein C; TAFI, thrombin-activatable fibrinolysis inhibitor; TFPI, tissue factor pathway inhibitor.

FIGURE 61-5

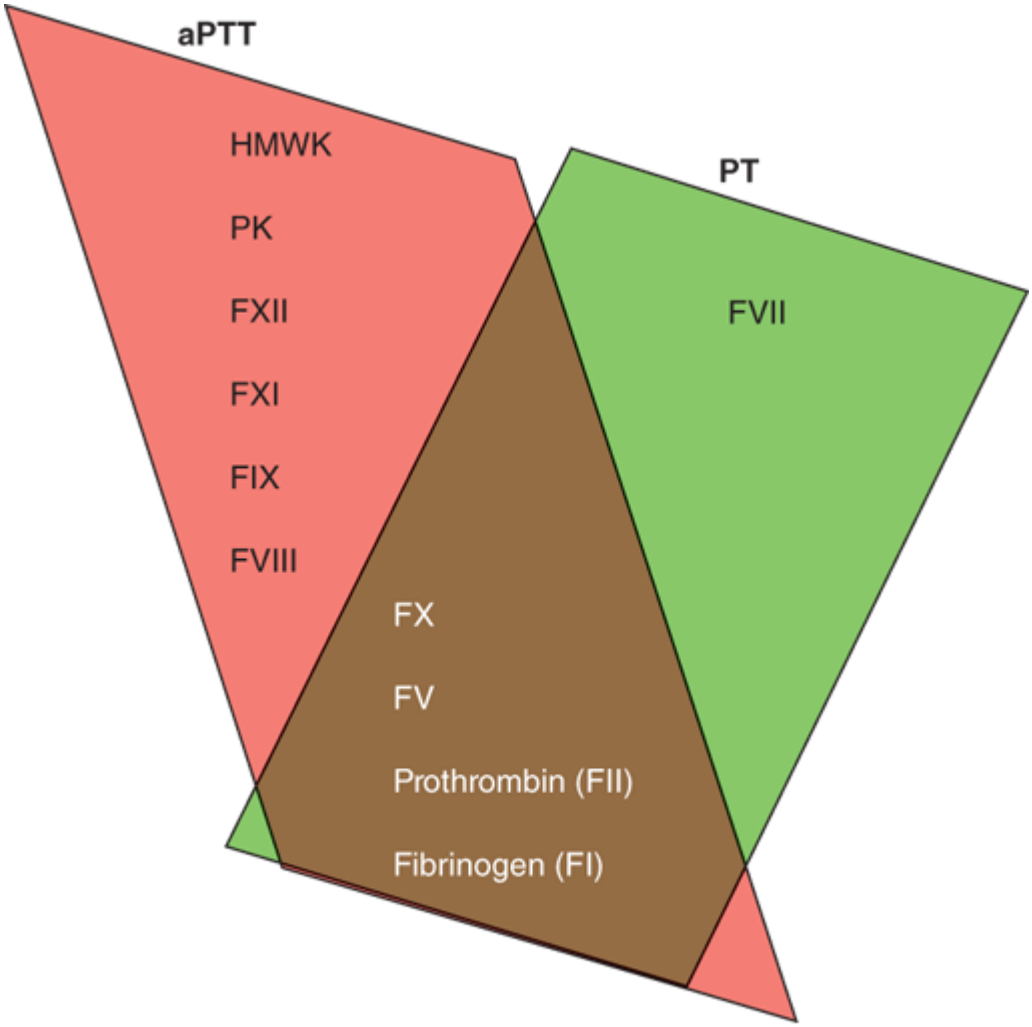
Thrombotic risk over time. Shown schematically is an individual's thrombotic risk over time. An underlying factor V Leiden mutation provides a "theoretically" constant increased risk. The thrombotic risk increases with age and, intermittently, with oral contraceptive (OCP) or hormone replacement therapy (HRT) use; other events may increase the risk further. At some point, the cumulative risk may increase to the threshold for thrombosis and result in deep-venous thrombosis (DVT). Note: The magnitude and duration of risk portrayed in the figure are meant for example only and may not precisely reflect the relative risk determined by clinical study. *(From BA Konkle, A Schafer, in DP Zipes et al [eds]: Braunwald's Heart Disease, 7th ed. Philadelphia, Saunders, 2005; modified with permission from FR Rosendaal: Venous thrombosis: A multicausal disease. Lancet 353:1167, 1999.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 61–6

Coagulation factor activity tested in the activated partial thromboplastin time (aPTT) in red and prothrombin time (PT) in green, or both. F, factor; HMWK, high-molecular-weight kininogen; PK, prekallikrein.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 61-4

Hemostatic Disorders and Coagulation Test Abnormalities

Prolonged Activated Partial Thromboplastin Time (aPTT)
No clinical bleeding—↓ factor XII, high-molecular-weight kininogen, prekallikrein
Variable, but usually mild, bleeding—↓ factor XI, mild ↓ factor VIII and factor IX
Frequent, severe bleeding—severe deficiencies of factors VIII and IX
Heparin and direct thrombin inhibitors
Prolonged Prothrombin Time (PT)
Factor VII deficiency
Vitamin K deficiency—early
Warfarin anticoagulation
Direct Xa inhibitors (rivaroxaban, edoxaban, apixaban– note PT may be normal)
Prolonged aPTT and PT
Factors II, V, X, or fibrinogen deficiency
Vitamin K deficiency—late
Direct thrombin inhibitors
Prolonged Thrombin Time
Heparin or heparin-like inhibitors
Direct thrombin inhibitors (e.g., dabigatran, argatroban, bivalirudin)
Mild or no bleeding—dysfibrinogenemia
Frequent, severe bleeding—afibrinogenemia
Prolonged PT and/or aPTT Not Corrected with Mixing with Normal Plasma

Bleeding—specific factor inhibitor
No symptoms, or clotting and/or pregnancy loss—lupus anticoagulant
Disseminated intravascular coagulation
Heparin or direct thrombin inhibitor
Abnormal Clot Solubility
Factor XIII deficiency
Inhibitors or defective cross-linking
Rapid Clot Lysis
Deficiency of α_2-antiplasmin or plasminogen activator inhibitor 1
Treatment with fibrinolytic therapy

FURTHER READING

Giannakopoulos B, Krilis SA: The pathogenesis of the antiphospholipid syndrome. N Engl J Med 368:11, 2013. [PubMed: 23281973]

Hicks LK et al: The ASH choosing wisely® campaign: Five hematologic tests and treatments to question. Blood 1222:3879, 2013.

Konkle BA: Direct oral anticoagulants: Monitoring anticoagulant effect, in *Direct Oral Anticoagulants in Clinical Practice*, Connors JM, ed., Hematol Oncol Clin North Am 30:995, 2016.

Mackie I et al: Guidelines on the laboratory aspect of assays used in haemostasis and thrombosis. Int Jnl Lab Hem 35:1, 2013.

Middeldorp S: Evidence-based approach to thrombophilia testing. J Thromb Haemost 31:275, 2011.

Rydz N, James PD: The evolution and value of bleeding assessment tools. J Thromb Haemost 10:2223, 2012. [PubMed: 22974079]

Wagenman BL et al: The laboratory approach to inherited and acquired coagulation factor deficiencies. Clini Lab Med 29:229, 2009.

Yau JW et al: Endothelial cell control of thrombosis. BMC Cardiovasc Disord 15:130, 2015. [PubMed: 26481314]

Chapter 62: Enlargement of Lymph Nodes and Spleen

Dan L. Longo

FIGURE 62-1

INTRODUCTION

This chapter is intended to serve as a guide to the evaluation of patients who present with enlargement of the lymph nodes (*lymphadenopathy*) or the spleen (*splenomegaly*). Lymphadenopathy is a rather common clinical finding in primary care settings, whereas palpable splenomegaly is less so.

LYMPHADENOPATHY

Lymphadenopathy may be an incidental finding in patients being examined for various reasons, or it may be a presenting sign or symptom of the patient's illness. The physician must eventually decide whether the lymphadenopathy is a normal finding or one that requires further study, up to and including biopsy. Soft, flat, submandibular nodes (<1 cm) are often palpable in healthy children and young adults; healthy adults may have palpable inguinal nodes of up to 2 cm, which are considered normal. Further evaluation of these normal nodes is not warranted. In contrast, if the physician believes the node(s) to be abnormal, then pursuit of a more precise diagnosis is needed.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Lymphadenopathy

Lymphadenopathy may be a primary or secondary manifestation of numerous disorders, as shown in **Table 62-1**. Many of these disorders are infrequent causes of lymphadenopathy. In primary care practice, more than two-thirds of patients with lymphadenopathy have nonspecific causes or upper respiratory illnesses (viral or bacterial) and <1% have a malignancy. In one study, 84% of patients referred for evaluation of lymphadenopathy had a "benign" diagnosis. The remaining 16% had a malignancy (lymphoma or metastatic adenocarcinoma). Of the patients with benign lymphadenopathy, 63% had a nonspecific or reactive etiology (no causative agent found), and the remainder had a specific cause demonstrated, most commonly infectious mononucleosis, toxoplasmosis, or tuberculosis. Thus, the vast majority of patients with lymphadenopathy will have a nonspecific etiology requiring few diagnostic tests.

CLINICAL ASSESSMENT

The physician will be aided in the pursuit of an explanation for the lymphadenopathy by a careful medical history, physical examination, selected laboratory tests, and perhaps an excisional lymph node biopsy.

The *medical history* should reveal the setting in which lymphadenopathy is occurring. Symptoms such as sore throat, cough, fever, night sweats, fatigue, weight loss, or pain in the nodes should be sought. The patient's age, sex, occupation, exposure to pets, sexual behavior, and use of drugs such as diphenylhydantoin are other important historic points. For example, children and young adults usually have benign (i.e., nonmalignant) disorders that account for the observed lymphadenopathy such as viral or bacterial upper respiratory infections; infectious mononucleosis; toxoplasmosis; and, in some countries, tuberculosis. In contrast, after age 50, the incidence of malignant disorders increases and that of benign disorders decreases.

The *physical examination* can provide useful clues such as the extent of lymphadenopathy (localized or generalized), size of nodes, texture, presence or absence of nodal tenderness, signs of inflammation over the node, skin lesions, and splenomegaly. A thorough ear, nose, and throat (ENT) examination is indicated in adult patients with cervical adenopathy and a history of tobacco use. Localized or regional adenopathy implies involvement of a single anatomic area. Generalized adenopathy has been defined as involvement of three or more noncontiguous lymph node areas. Many of the causes of lymphadenopathy (Table 62-1) can produce localized *or* generalized adenopathy, so this distinction is of limited utility in the differential diagnosis. Nevertheless, generalized lymphadenopathy is frequently associated with nonmalignant disorders such as infectious mononucleosis (Epstein-Barr virus [EBV] or cytomegalovirus [CMV]), toxoplasmosis, AIDS, other viral infections, systemic lupus erythematosus (SLE), and mixed connective tissue disease. Acute and chronic lymphocytic leukemias and malignant lymphomas also produce generalized adenopathy in adults.

The site of localized or regional adenopathy may provide a useful clue about the cause. Occipital adenopathy often reflects an infection of the scalp, and preauricular adenopathy accompanies conjunctival infections and cat-scratch disease. The most frequent site of regional adenopathy is the neck, and most of the causes are benign—upper respiratory infections, oral and dental lesions, infectious mononucleosis, or other viral illnesses. The chief malignant causes include metastatic cancer from head and neck, breast, lung, and thyroid primaries. Enlargement of supraclavicular and scalene nodes is always abnormal. Because these nodes drain regions of the lung and retroperitoneal space, they can reflect lymphomas, other cancers, or infectious processes arising in these areas. Virchow's node is an enlarged left supraclavicular node infiltrated with metastatic cancer from a gastrointestinal primary. Metastases to supraclavicular nodes also occur from lung, breast, testis, or ovarian cancers. Tuberculosis, sarcoidosis, and toxoplasmosis are nonneoplastic causes of supraclavicular adenopathy. Axillary adenopathy is usually due to injuries or localized infections of the ipsilateral upper extremity. Malignant causes include melanoma or lymphoma and, in women, breast cancer. Inguinal lymphadenopathy is usually secondary to infections or trauma of the lower extremities and may accompany sexually transmitted diseases such as lymphogranuloma venereum, primary syphilis,

genital herpes, or chancroid. These nodes may also be involved by lymphomas and metastatic cancer from primary lesions of the rectum, genitalia, or lower extremities (melanoma).

The size and texture of the lymph node(s) and the presence of pain are useful parameters in evaluating a patient with lymphadenopathy. Nodes <1.0 cm^2 in area (1.0 cm × 1.0 cm or less) are almost always secondary to benign, nonspecific reactive causes. In one retrospective analysis of younger patients (9–25 years) who had a lymph node biopsy, a maximum diameter of >2 cm served as one discriminant for predicting that the biopsy would reveal malignant or granulomatous disease. Another study showed that a lymph node size of 2.25 cm^2 (1.5 cm × 1.5 cm) was the best size limit for distinguishing malignant or granulomatous lymphadenopathy from other causes of lymphadenopathy. Patients with node(s) ≤1.0 cm^2 should be observed after excluding infectious mononucleosis and/or toxoplasmosis unless there are symptoms and signs of an underlying systemic illness.

The texture of lymph nodes may be described as soft, firm, rubbery, hard, discrete, matted, tender, movable, or fixed. Tenderness is found when the capsule is stretched during rapid enlargement, usually secondary to an inflammatory process. Some malignant diseases such as acute leukemia may produce rapid enlargement and pain in the nodes. Nodes involved by lymphoma tend to be large, discrete, symmetric, rubbery, firm, mobile, and nontender. Nodes containing metastatic cancer are often hard, nontender, and nonmovable because of fixation to surrounding tissues. The coexistence of splenomegaly in the patient with lymphadenopathy implies a systemic illness such as infectious mononucleosis, lymphoma, acute or chronic leukemia, SLE, sarcoidosis, toxoplasmosis, cat-scratch disease, or other less common hematologic disorders. The patient's story should provide helpful clues about the underlying systemic illness.

Nonsuperficial presentations (thoracic or abdominal) of adenopathy are usually detected as the result of a symptom-directed diagnostic workup. Thoracic adenopathy may be detected by routine chest radiography or during the workup for superficial adenopathy. It may also be found because the patient complains of a cough or wheezing from airway compression; hoarseness from recurrent laryngeal nerve involvement; dysphagia from esophageal compression; or swelling of the neck, face, or arms secondary to compression of the superior vena cava or subclavian vein. The differential diagnosis of mediastinal and hilar adenopathy includes primary lung disorders and systemic illnesses that characteristically involve mediastinal or hilar nodes. In the young, mediastinal adenopathy is associated with infectious mononucleosis and sarcoidosis. In endemic regions, histoplasmosis can cause unilateral paratracheal lymph node involvement that mimics lymphoma. Tuberculosis can also cause unilateral adenopathy. In older patients, the differential diagnosis includes primary lung cancer (especially among smokers), lymphomas, metastatic carcinoma (usually lung), tuberculosis, fungal infection, and sarcoidosis.

Enlarged intraabdominal or retroperitoneal nodes are usually malignant. Although tuberculosis may present as mesenteric lymphadenitis, these masses usually contain lymphomas or, in young men, germ cell tumors.

LABORATORY INVESTIGATION

The laboratory investigation of patients with lymphadenopathy must be tailored to elucidate the etiology suspected from the patient's history and physical findings. One study from a family practice clinic evaluated 249 younger patients with "enlarged lymph nodes, not infected" or "lymphadenitis." No laboratory studies were obtained in 51%. When studies were performed, the most common were a complete blood count (CBC) (33%), throat culture (16%), chest x-ray (12%), or monospot test (10%). Only eight patients (3%) had a node biopsy, and half of those were normal or reactive. The CBC can provide useful data for the diagnosis of acute or chronic leukemias, EBV or CMV mononucleosis, lymphoma with a leukemic component, pyogenic infections, or immune cytopenias in illnesses such as SLE. Serologic studies may demonstrate antibodies specific to components of EBV, CMV, HIV, and other viruses; *Toxoplasma gondii*; *Brucella*; etc. If SLE is suspected, antinuclear and anti-DNA antibody studies are warranted.

The chest x-ray is usually negative, but the presence of a pulmonary infiltrate or mediastinal lymphadenopathy would suggest tuberculosis, histoplasmosis, sarcoidosis, lymphoma, primary lung cancer, or metastatic cancer and demands further investigation.

A variety of imaging techniques (CT, MRI, ultrasound, color Doppler ultrasonography) have been employed to differentiate benign from malignant lymph nodes, especially in patients with head and neck cancer. CT and MRI are comparably accurate (65–90%) in the diagnosis of metastases to cervical lymph nodes. Ultrasonography has been used to determine the long (L) axis, short (S) axis, and a ratio of long to short axis in cervical nodes. An L/S ratio of <2.0 has a sensitivity and a specificity of 95% for distinguishing benign and malignant nodes in patients with head and neck cancer. This ratio has greater specificity and sensitivity than palpation or measurement of either the long or the short axis alone.

The indications for lymph node biopsy are imprecise, yet it is a valuable diagnostic tool. The decision to biopsy may be made early in a patient's evaluation or delayed for up to two weeks. Prompt biopsy should occur if the patient's history and physical findings suggest a malignancy; examples include a solitary, hard, nontender cervical node in an older patient who is a chronic user of tobacco; supraclavicular adenopathy; and solitary or generalized adenopathy that is firm, movable, and suggestive of lymphoma. If a primary head and neck cancer is suspected as the basis of a solitary, hard cervical node, then a careful ENT examination should be performed. Any mucosal lesion that is suspicious for a primary neoplastic process should be biopsied first. If no mucosal lesion is detected, an excisional biopsy of the largest node should be performed. Fine-needle aspiration should not be performed as the first diagnostic procedure. Most diagnoses require more tissue than such aspiration can provide, and it often delays a definitive diagnosis. Fine-needle aspiration should be reserved for thyroid nodules and for confirmation of relapse in patients whose primary diagnosis is known. If the primary physician is uncertain about whether to proceed to biopsy, consultation with a hematologist or medical oncologist should be helpful. In primary care practices, <5% of lymphadenopathy patients will require a biopsy. That percentage will be considerably larger in referral practices, i.e., hematology, oncology, or ENT.

Two groups have reported algorithms that they claim will identify more precisely those lymphadenopathy patients who should have a biopsy. Both reports were retrospective analyses in referral practices. The first study involved patients 9–25 years of age who had a node biopsy performed. Three variables were identified

that predicted those young patients with peripheral lymphadenopathy who should undergo biopsy; lymph node size >2 cm in diameter and abnormal chest x-ray had positive predictive values, whereas recent ENT symptoms had negative predictive values. The second study evaluated 220 lymphadenopathy patients in a hematology unit and identified five variables (lymph node size, location [supraclavicular or nonsupraclavicular], age [>40 years or <40 years], texture [nonhard or hard], and tenderness) that were used in a mathematical model to identify those patients requiring a biopsy. Positive predictive value was found for age >40 years, supraclavicular location, node size >2.25 cm^2, hard texture, and lack of pain or tenderness. Negative predictive value was evident for age <40 years, node size <1.0 cm^2, nonhard texture, and tender or painful nodes. Ninety-one percent of those who required biopsy were correctly classified by this model. Because both of these studies were retrospective analyses and one was limited to young patients, it is not known how useful these models would be if applied prospectively in a primary care setting.

Most lymphadenopathy patients do not require a biopsy, and at least half require no laboratory studies. If the patient's history and physical findings point to a benign cause for lymphadenopathy, careful follow-up at a 2- to 4-week interval can be employed. The patient should be instructed to return for reevaluation if there is an increase in the size of the nodels). Antibiotics are not indicated for lymphadenopathy unless strong evidence of a bacterial infection is present. Glucocorticoids should not be used to treat lymphadenopathy because their lympholytic effect obscures some diagnoses (lymphoma, leukemia, Castleman's disease) and they contribute to delayed healing or activation of underlying infections. An exception to this statement is the life-threatening pharyngeal obstruction by enlarged lymphoid tissue in Waldeyer's ring that is occasionally seen in infectious mononucleosis.

TABLE 62-1

Diseases Associated with Lymphadenopathy

1. Infectious diseases
    a. Viral—infectious mononucleosis syndromes (EBV, CMV), infectious hepatitis, herpes simplex, herpesvirus-6, varicella-zoster virus, rubella, measles, adenovirus, HIV, epidemic keratoconjunctivitis, vaccinia, herpesvirus-8
    b. Bacterial—streptococci, staphylococci, cat-scratch disease, brucellosis, tularemia, plague, chancroid, melioidosis, glanders, tuberculosis, atypical mycobacterial infection, primary and secondary syphilis, diphtheria, leprosy, bartonella
    c. Fungal—histoplasmosis, coccidioidomycosis, paracoccidioidomycosis
    d. Chlamydial—lymphogranuloma venereum, trachoma
    e. Parasitic—toxoplasmosis, leishmaniasis, trypanosomiasis, filariasis
    f. Rickettsial—scrub typhus, rickettsialpox, Q fever
2. Immunologic diseases
    a. Rheumatoid arthritis
    b. Juvenile rheumatoid arthritis
    c. Mixed connective tissue disease
    d. Systemic lupus erythematosus
    e. Dermatomyositis
    f. Sjögren's syndrome
    g. Serum sickness
    h. Drug hypersensitivity—diphenylhydantoin, hydralazine, allopurinol, primidone, gold, carbamazepine, etc.
    i. Angioimmunoblastic lymphadenopathy
    j. Primary biliary cirrhosis
    k. Graft-vs.-host disease
    l. Silicone-associated
    m. Autoimmune lymphoproliferative syndrome
    n. IgG4-related disease
    o. Immune reconstitution inflammatory syndrome (IRIS)
3. Malignant diseases
    a. Hematologic—Hodgkin's disease, non-Hodgkin's lymphomas, acute or chronic lymphocytic leukemia, hairy cell leukemia, malignant histiocytosis, amyloidosis
    b. Metastatic—from numerous primary sites
4. Lipid storage diseases—Gaucher's, Niemann-Pick, Fabry, Tangier
5. Endocrine diseases—hyperthyroidism
6. Other disorders
    a. Castleman's disease (giant lymph node hyperplasia)
    b. Sarcoidosis

c. Dermatopathic lymphadenitis
d. Lymphomatoid granulomatosis
e. Histiocytic necrotizing lymphadenitis (Kikuchi's disease)
f. Sinus histiocytosis with massive lymphadenopathy (Rosai-Dorfman disease)
g. Mucocutaneous lymph node syndrome (Kawasaki's disease)
h. Histiocytosis X
i. Familial Mediterranean fever
j. Severe hypertriglyceridemia
k. Vascular transformation of sinuses
l. Inflammatory pseudotumor of lymph node
m. Congestive heart failure

Abbreviations: CMV, cytomegalovirus; EBV, Epstein-Barr virus.

SPLENOMEGALY

STRUCTURE AND FUNCTION OF THE SPLEEN

The spleen is a reticuloendothelial organ that has its embryologic origin in the dorsal mesogastrium at about five weeks' gestation. It arises in a series of hillocks, migrates to its normal adult location in the left upper quadrant (LUQ), and is attached to the stomach via the gastrolienal ligament and to the kidney via the lienorenal ligament. When the hillocks fail to unify into a single tissue mass, accessory spleens may develop in around 20% of persons. The function of the spleen has been elusive. Galen believed it was the source of "black bile" or melancholia, and the word *hypochondria* (literally, beneath the ribs) and the idiom "to vent one's spleen" attest to the beliefs that the spleen had an important influence on the psyche and emotions. In humans, its normal physiologic roles seem to be the following:

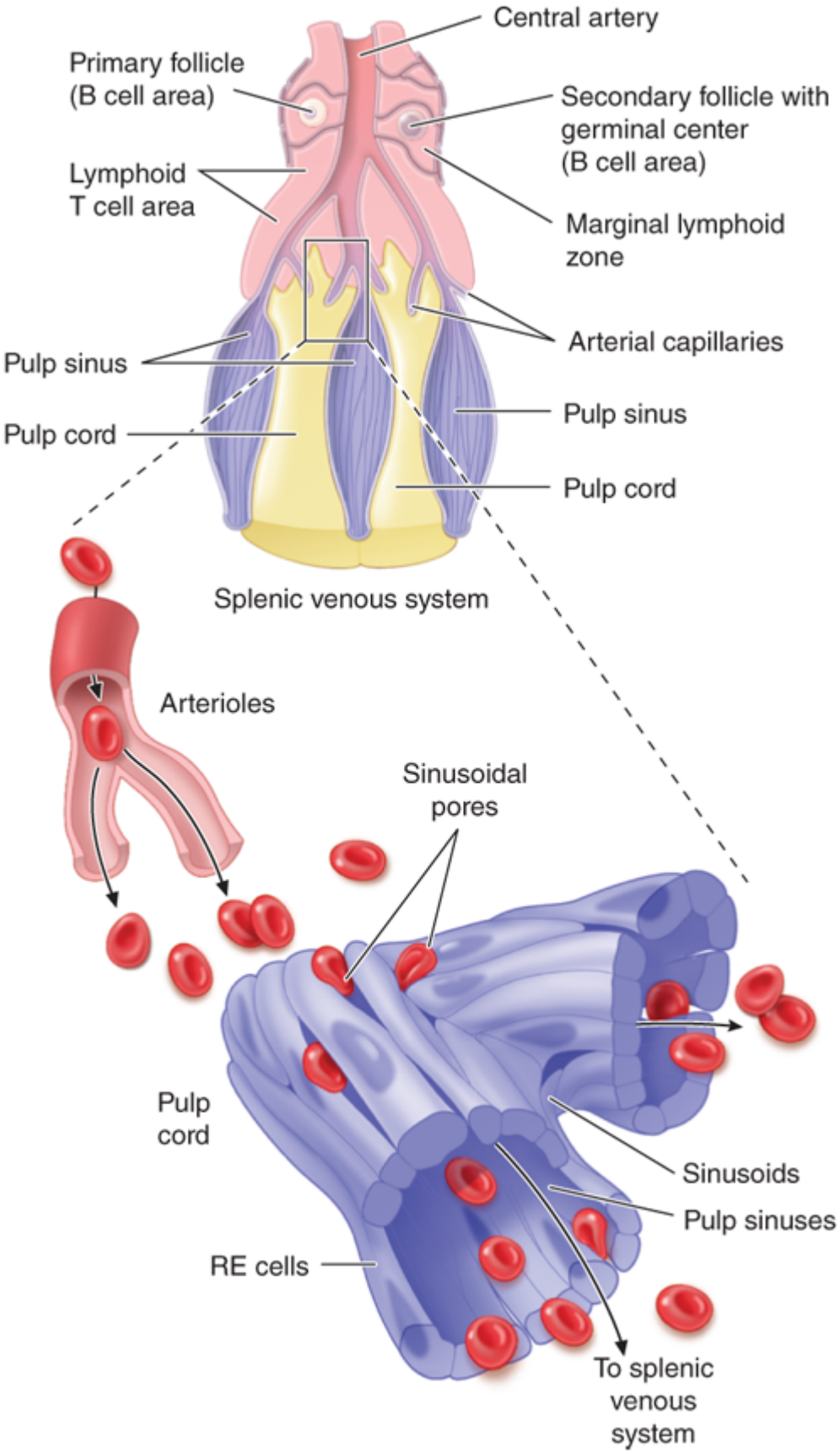
1. Maintenance of quality control over erythrocytes in the red pulp by removal of senescent and defective red blood cells. The spleen accomplishes this function through a unique organization of its parenchyma and vasculature **(Fig. 62-1)**.

2. Synthesis of antibodies in the white pulp.

3. The removal of antibody-coated bacteria and antibody-coated blood cells from the circulation.

FIGURE 62-1
Schematic spleen structure. The spleen comprises many units of red and white pulp centered around small branches of the splenic artery, called *central arteries*. White pulp is lymphoid in nature and contains B-cell follicles, a marginal zone around the follicles, and T-cell–rich areas sheathing arterioles. The red pulp areas include pulp sinuses and pulp cords. The cords are dead ends. In order to regain access to the circulation, red

blood cells must traverse tiny openings in the sinusoidal lining. Stiff, damaged, or old red cells cannot enter the sinuses. RE, reticuloendthelial. *(Bottom portion of figure from RS Hillman, KA Ault: Hematology in Clinical Practice, 4th ed. New York, McGraw-Hill, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

An increase in these normal functions may result in splenomegaly.

The spleen is composed of *red pulp* and *white pulp*, which are Malpighi's terms for the red blood–filled sinuses and reticuloendothelial cell–lined cords and the white lymphoid follicles arrayed within the red pulp matrix. The spleen is in the portal circulation. The reason for this is unknown but may relate to the fact that lower blood pressure allows less rapid flow and minimizes damage to normal erythrocytes. Blood flows into the spleen at a rate of about 150 mL/min through the splenic artery, which ultimately ramifies into central arterioles. Some blood goes from the arterioles to capillaries and then to splenic veins and out of the spleen, but the majority of blood from central arterioles flows into the macrophage-lined sinuses and cords. The blood entering the sinuses reenters the circulation through the splenic venules, but the blood entering the cords is subjected to an inspection of sorts. To return to the circulation, the blood cells in the cords must squeeze through slits in the cord lining to enter the sinuses that lead to the venules. Old and damaged erythrocytes are less deformable and are retained in the cords, where they are destroyed and their components recycled. Red cell–inclusion bodies such as parasites **(Chaps. 219, 220, and A6)**, nuclear residua (Howell-Jolly bodies, **see Fig. 59-6**), or denatured hemoglobin (Heinz bodies) are pinched off in the process of passing through the slits, a process called *pitting*. The culling of dead and damaged cells and the pitting of cells with inclusions appear to occur without significant delay because the blood transit time through the spleen is only slightly slower than in other organs.

The spleen is also capable of assisting the host in adapting to its hostile environment. It has at least three adaptive functions: (1) clearance of bacteria and particulates from the blood, (2) the generation of immune responses to certain pathogens, and (3) the generation of cellular components of the blood under circumstances in which the marrow is unable to meet the needs (i.e., extramedullary hematopoiesis). The latter adaptation is a recapitulation of the blood-forming function the spleen plays during gestation. In some animals, the spleen also serves a role in the vascular adaptation to stress because it stores red blood cells (often hemoconcentrated to higher hematocrits than normal) under normal circumstances and contracts under the influence of β-adrenergic stimulation to provide the animal with an autotransfusion and improved oxygen-carrying capacity. However, the normal human spleen does not sequester or store red blood cells and does not contract in response to sympathetic stimuli. The normal human spleen contains approximately one-third of the total body platelets and a significant number of marginated neutrophils. These sequestered cells are available when needed to respond to bleeding or infection.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Splenomegaly

CLINICAL ASSESSMENT

The most common *symptoms* produced by diseases involving the spleen are pain and a heavy sensation in the LUQ. Massive splenomegaly may cause early satiety. Pain may result from acute swelling of the spleen with stretching of the capsule, infarction, or inflammation of the capsule. For many years, it was believed that splenic infarction was clinically silent, which, at times, is true. However, Soma Weiss, in his classic 1942 report of the self-observations by a Harvard medical student on the clinical course of subacute bacterial

endocarditis, documented that severe LUQ and pleuritic chest pain may accompany thromboembolic occlusion of splenic blood flow. Vascular occlusion, with infarction and pain, is commonly seen in children with sickle cell crises. Rupture of the spleen, from either trauma or infiltrative disease that breaks the capsule, may result in intraperitoneal bleeding, shock, and death. The rupture itself may be painless.

A palpable spleen is the major *physical sign* produced by diseases affecting the spleen and suggests enlargement of the organ. The normal spleen weighs <250 g, decreases in size with age, normally lies entirely within the rib cage, has a maximum cephalocaudad diameter of 13 cm by ultrasonography or maximum length of 12 cm and/or width of 7 cm by radionuclide scan, and is usually not palpable. However, a palpable spleen was found in 3% of 2200 asymptomatic, male, freshman college students. Follow-up at 3 years revealed that 30% of those students still had a palpable spleen without any increase in disease prevalence. Ten-year follow-up found no evidence for lymphoid malignancies. Furthermore, in some tropical countries (e.g., New Guinea), the incidence of splenomegaly may reach 60%. Thus, the presence of a palpable spleen does not always equate with presence of disease. Even when disease is present, splenomegaly may not reflect the primary disease but rather a reaction to it. For example, in patients with Hodgkin's disease, only two-thirds of the palpable spleens show involvement by the cancer.

Physical examination of the spleen uses primarily the techniques of palpation and percussion. Inspection may reveal fullness in the LUQ that descends on inspiration, a finding associated with a massively enlarged spleen. Auscultation may reveal a venous hum or friction rub.

Palpation can be accomplished by bimanual palpation, ballotment, and palpation from above (Middleton maneuver). For bimanual palpation, which is at least as reliable as the other techniques, the patient is supine with flexed knees. The examiner's left hand is placed on the lower rib cage and pulls the skin toward the costal margin, allowing the fingertips of the right hand to feel the tip of the spleen as it descends while the patient inspires slowly, smoothly, and deeply. Palpation is begun with the right hand in the left lower quadrant with gradual movement toward the left costal margin, thereby identifying the lower edge of a massively enlarged spleen. When the spleen tip is felt, the finding is recorded as centimeters below the left costal margin at some arbitrary point, i.e., 10–15 cm, from the midpoint of the umbilicus or the xiphisternal junction. This allows other examiners to compare findings or the initial examiner to determine changes in size over time. Bimanual palpation in the right lateral decubitus position adds nothing to the supine examination.

Percussion for splenic dullness is accomplished with any of three techniques described by Nixon, Castell, or Barkun:

1. *Nixon's method*: The patient is placed on the right side so that the spleen lies above the colon and stomach. Percussion begins at the lower level of pulmonary resonance in the posterior axillary line and proceeds diagonally along a perpendicular line toward the lower midanterior costal margin. The upper border of dullness is normally 6–8 cm above the costal margin. Dullness >8 cm in an adult is presumed to indicate splenic enlargement.

2. *Castell's method*: With the patient supine, percussion in the lowest intercostal space in the anterior axillary line (8th or 9th) produces a resonant note if the spleen is normal in size. This is true during expiration or full inspiration. A dull percussion note on full inspiration suggests splenomegaly.

3. *Percussion of Traube's semilunar space*: The borders of Traube's space are the sixth rib superiorly, the left midaxillary line laterally, and the left costal margin inferiorly. The patient is supine with the left arm slightly abducted. During normal breathing, this space is percussed from medial to lateral margins, yielding a normal resonant sound. A dull percussion note suggests splenomegaly.

Studies comparing methods of percussion and palpation with a standard of ultrasonography or scintigraphy have revealed sensitivity of 56–71% for palpation and 59–82% for percussion. Reproducibility among examiners is better for palpation than percussion. Both techniques are less reliable in obese patients or patients who have just eaten. Thus, the physical examination techniques of palpation and percussion are imprecise at best. It has been suggested that the examiner perform percussion first and, if positive, proceed to palpation; if the spleen is palpable, then one can be reasonably confident that splenomegaly exists. However, not all LUQ masses are enlarged spleens; gastric or colon tumors and pancreatic or renal cysts or tumors can mimic splenomegaly.

The presence of an enlarged spleen can be more precisely determined, if necessary, by liver-spleen radionuclide scan, CT, MRI, or ultrasonography. The latter technique is the current procedure of choice for routine assessment of spleen size (normal = a maximum cephalocaudad diameter of 13 cm) because it has high sensitivity and specificity and is safe, noninvasive, quick, mobile, and less costly. Nuclear medicine scans are accurate, sensitive, and reliable but are costly, require greater time to generate data, and use immobile equipment. They have the advantage of demonstrating accessory splenic tissue. CT and MRI provide accurate determination of spleen size, but the equipment is immobile and the procedures are expensive. MRI appears to offer no advantage over CT. Changes in spleen structure such as mass lesions, infarcts, inhomogeneous infiltrates, and cysts are more readily assessed by CT, MRI, or ultrasonography. None of these techniques is very reliable in the detection of patchy infiltration (e.g., Hodgkin's disease).

DIFFERENTIAL DIAGNOSIS

Many of the diseases associated with splenomegaly are listed in **Table 62-2**. They are grouped according to the presumed basic mechanisms responsible for organ enlargement:

1. Hyperplasia or hypertrophy related to a particular splenic function such as reticuloendothelial hyperplasia (work hypertrophy) in diseases such as hereditary spherocytosis or thalassemia syndromes that require removal of large numbers of defective red blood cells; immune hyperplasia in response to systemic infection (infectious mononucleosis, subacute bacterial endocarditis) or to immunologic diseases (immune thrombocytopenia, SLE, Felty's syndrome).

2. Passive congestion due to decreased blood flow from the spleen in conditions that produce portal hypertension (cirrhosis, Budd-Chiari syndrome, congestive heart failure).

3. Infiltrative diseases of the spleen (lymphomas, metastatic cancer, amyloidosis, Gaucher's disease, myeloproliferative disorders with extramedullary hematopoiesis).

The differential diagnostic possibilities are much fewer when the spleen is "massively enlarged," palpable >8 cm below the left costal margin or its drained weight is ≥1000 g **(Table 62-3)**. The vast majority of such patients will have non-Hodgkin's lymphoma, chronic lymphocytic leukemia, hairy cell leukemia, chronic myeloid leukemia, myelofibrosis with myeloid metaplasia, or polycythemia vera.

LABORATORY ASSESSMENT

The major laboratory abnormalities accompanying splenomegaly are determined by the underlying systemic illness. Erythrocyte counts may be normal, decreased (thalassemia major syndromes, SLE, cirrhosis with portal hypertension), or increased (polycythemia vera). Granulocyte counts may be normal, decreased (Felty's syndrome, congestive splenomegaly, leukemias), or increased (infections or inflammatory disease, myeloproliferative disorders). Similarly, the platelet count may be normal, decreased when there is enhanced sequestration or destruction of platelets in an enlarged spleen (congestive splenomegaly, Gaucher's disease, immune thrombocytopenia), or increased in the myeloproliferative disorders such as polycythemia vera.

The CBC may reveal cytopenia of one or more blood cell types, which should suggest *hypersplenism*. This condition is characterized by splenomegaly, cytopenia(s), normal or hyperplastic bone marrow, and a response to splenectomy. The latter characteristic is less precise because reversal of cytopenia, particularly granulocytopenia, is sometimes not sustained after splenectomy. The cytopenias result from increased destruction of the cellular elements secondary to reduced flow of blood through enlarged and congested cords (congestive splenomegaly) or to immune-mediated mechanisms. In hypersplenism, various cell types usually have normal morphology on the peripheral blood smear, although the red cells may be spherocytic due to loss of surface area during their longer transit through the enlarged spleen. The increased marrow production of red cells should be reflected as an increased reticulocyte production index, although the value may be less than expected due to increased sequestration of reticulocytes in the spleen.

The need for additional laboratory studies is dictated by the differential diagnosis of the underlying illness of which splenomegaly is a manifestation.

TABLE 62-2

Diseases Associated with Splenomegaly Grouped by Pathogenic Mechanism

Enlargement Due to Increased Demand for Splenic Function	
Reticuloendothelial system hyperplasia (for removal of defective erythrocytes)	Malaria
Spherocytosis	Leishmaniasis
Early sickle cell anemia	Trypanosomiasis
Ovalocytosis	Ehrlichiosis
Thalassemia major	Disordered immunoregulation
Hemoglobinopathies	Rheumatoid arthritis (Felty's syndrome)
Paroxysmal nocturnal hemoglobinuria	Systemic lupus erythematosus
Pernicious anemia	Collagen vascular diseases
Immune hyperplasia	Serum sickness
Response to infection (viral, bacterial, fungal, parasitic)	Immune hemolytic anemias
Infectious mononucleosis	Immune thrombocytopenias
AIDS	Immune neutropenias
Viral hepatitis	Drug reactions
Cytomegalovirus	Angioimmunoblastic lymphadenopathy
Subacute bacterial endocarditis	Sarcoidosis
Bacterial septicemia	Thyrotoxicosis (benign lymphoid hypertrophy)
Congenital syphilis	Interleukin 2 therapy
Splenic abscess	Extramedullary hematopoiesis

Tuberculosis	Myelofibrosis
Histoplasmosis	Marrow damage by toxins, radiation, strontium
	Marrow infiltration by tumors, leukemias, Gaucher's disease
Enlargement Due to Abnormal Splenic or Portal Blood Flow	
Cirrhosis	Splenic artery aneurysm
Hepatic vein obstruction	Hepatic schistosomiasis
Portal vein obstruction, intrahepatic or extrahepatic	Congestive heart failure
Cavernous transformation of the portal vein	Hepatic echinococcosis
Splenic vein obstruction	Portal hypertension (any cause including the above): "Banti's disease"
Infiltration of the Spleen	
Intracellular or extracellular depositions	Hodgkin's disease
Amyloidosis	Myeloproliferative syndromes (e.g., polycythemia vera, essential thrombocytosis)
Gaucher's disease	Angiosarcomas
Niemann-Pick disease	Metastatic tumors (melanoma is most common)
Tangier disease	Eosinophilic granuloma
Hurler's syndrome and other mucopolysaccharidoses	Histiocytosis X
Hyperlipidemias	Hamartomas
Benign and malignant cellular infiltrations	Hemangiomas, fibromas, lymphangiomas
Leukemias (acute, chronic, lymphoid, myeloid,	Splenic cysts

monocytic)	
Lymphomas	
Unknown Etiology	
Idiopathic splenomegaly	Iron-deficiency anemia
Berylliosis	

TABLE 62-3

Diseases Associated with Massive Splenomegaly*

Chronic myeloid leukemia	Gaucher's disease
Lymphomas	Chronic lymphocytic leukemia
Hairy cell leukemia	Sarcoidosis
Myelofibrosis with myeloid metaplasia	Autoimmune hemolytic anemia
Polycythemia vera	Diffuse splenic hemangiomatosis

*The spleen extends >8 cm below left costal margin and/or weighs >1000 g.

SPLENECTOMY

Splenectomy is infrequently performed for diagnostic purposes, especially in the absence of clinical illness or other diagnostic tests that suggest underlying disease. More often, splenectomy is performed for symptom control in patients with massive splenomegaly, for disease control in patients with traumatic splenic rupture, or for correction of cytopenias in patients with hypersplenism or immune-mediated destruction of one or more cellular blood elements. Splenectomy is necessary for staging of patients with Hodgkin's disease only in those with clinical stage I or II disease in whom radiation therapy alone is contemplated as the treatment. Noninvasive staging of the spleen in Hodgkin's disease is not a sufficiently reliable basis for treatment decisions because one-third of normal-sized spleens will be involved with Hodgkin's disease and one-third of enlarged spleens will be tumor-free. The widespread use of systemic therapy to test all stages of Hodgkin's disease has made staging laparotomy with splenectomy unnecessary. Although splenectomy in chronic myeloid leukemia (CML) does not affect the natural history of disease, removal of the massive spleen usually

makes patients significantly more comfortable and simplifies their management by significantly reducing transfusion requirements. The improvements in therapy of CML have reduced the need for splenectomy for symptom control. Splenectomy is an effective secondary or tertiary treatment for two chronic B cell leukemias, hairy cell leukemia and prolymphocytic leukemia, and for the very rare splenic mantle cell or marginal zone lymphoma. Splenectomy in these diseases may be associated with significant tumor regression in bone marrow and other sites of disease. Similar regressions of systemic disease have been noted after splenic irradiation in some types of lymphoid tumors, especially chronic lymphocytic leukemia and prolymphocytic leukemia. This has been termed the *abscopal effect*. Such systemic tumor responses to local therapy directed at the spleen suggest that some hormone or growth factor produced by the spleen may affect tumor cell proliferation, but this conjecture is not yet substantiated. A common therapeutic indication for splenectomy is traumatic or iatrogenic splenic rupture. In a fraction of patients with splenic rupture, peritoneal seeding of splenic fragments can lead to *splenosis*—the presence of multiple rests of spleen tissue not connected to the portal circulation. This ectopic spleen tissue may cause pain or gastrointestinal obstruction, as in endometriosis. A large number of hematologic, immunologic, and congestive causes of splenomegaly can lead to destruction of one or more cellular blood elements. In the majority of such cases, splenectomy can correct the cytopenias, particularly anemia and thrombocytopenia. In a large series of patients seen in two tertiary care centers, the indication for splenectomy was diagnostic in 10% of patients, therapeutic in 44%, staging for Hodgkin's disease in 20%, and incidental to another procedure in 26%. Perhaps the only contraindication to splenectomy is the presence of marrow failure, in which the enlarged spleen is the only source of hematopoietic tissue.

Often the splenectomy is done by laparoscopy, which is associated with shorter hospital stays and faster recovery than the open procedure; however, concern has emerged that the laparoscopic approach is associated with a higher risk of postoperative portal venous system thrombosis and Budd-Chiari syndrome.

The absence of the spleen has minimal long-term effects on the hematologic profile. In the immediate postsplenectomy period, leukocytosis (up to 25,000/μL) and thrombocytosis (up to 1×10^6/μL) may develop, but within 2–3 weeks, blood cell counts and survival of each cell lineage are usually normal. The chronic manifestations of splenectomy are marked variation in size and shape of erythrocytes (anisocytosis, poikilocytosis) and the presence of Howell-Jolly bodies (nuclear remnants), Heinz bodies (denatured hemoglobin), basophilic stippling, and an occasional nucleated erythrocyte in the peripheral blood. When such erythrocyte abnormalities appear in a patient whose spleen has not been removed, one should suspect splenic infiltration by tumor that has interfered with its normal culling and pitting function.

The most serious consequence of splenectomy is increased susceptibility to bacterial infections, particularly those with capsules such as *Streptococcus pneumoniae*, *Haemophilus influenzae*, and some gram-negative enteric organisms. Patients aged <20 years are particularly susceptible to overwhelming sepsis with *S. pneumoniae*, and the overall actuarial risk of sepsis in patients who have had their spleens removed is about 7% in 10 years. The case-fatality rate for pneumococcal sepsis in splenectomized patients is 50–80%. About 25% of patients without spleens will develop a serious infection at some time in their life. The frequency is highest within the first three years after splenectomy. About 15% of the infections are polymicrobial, and

lung, skin, and blood are the most common sites. No increased risk of viral infection has been noted in patients who have no spleen. The susceptibility to bacterial infections relates to the inability to remove opsonized bacteria from the bloodstream and a defect in making antibodies to T cell–independent antigens such as the polysaccharide components of bacterial capsules. Pneumococcal vaccine should be administered to all patients 2 weeks before elective splenectomy. The Advisory Committee on Immunization Practices recommends that these patients receive repeat vaccination 5 years post-splenectomy. Efficacy has not been proven for this group, and the recommendation discounts the possibility that administration of the vaccine may actually lower the titer of specific pneumococcal antibodies. A more effective pneumococcal conjugate vaccine that involves T cells in the response is now available (Prevenar, 7-valent). The vaccine to *Neisseria meningitidis* should also be given to patients in whom elective splenectomy is planned. Although efficacy data for *Haemophilus influenzae* type b vaccine are not available for older children or adults, it may be given to patients who have had a splenectomy.

Splenectomized patients should be educated to consider any unexplained fever as a medical emergency. Prompt medical attention with evaluation and treatment of suspected bacteremia may be lifesaving. Routine chemoprophylaxis with oral penicillin can result in the emergence of drug-resistant strains and is not recommended.

In addition to an increased susceptibility to bacterial infections, splenectomized patients are also more susceptible to the parasitic disease babesiosis. The splenectomized patient should avoid areas where the parasite *Babesia* is endemic (e.g., Cape Cod, MA).

Surgical removal of the spleen is an obvious cause of hyposplenism. Patients with sickle cell disease often suffer from autosplenectomy as a result of splenic destruction by the numerous infarcts associated with sickle cell crises during childhood. Indeed, the presence of a palpable spleen in a patient with sickle cell disease after age 5 suggests a coexisting hemoglobinopathy, e.g., thalassemia or hemoglobin C. In addition, patients who receive splenic irradiation for a neoplastic or autoimmune disease are also functionally hyposplenic. The term *hyposplenism* is preferred to *asplenism* in referring to the physiologic consequences of splenectomy because asplenia is a rare, specific, and fatal congenital abnormality in which there is a failure of the left side of the coelomic cavity (which includes the splenic anlagen) to develop normally. Infants with asplenia have no spleens, but that is the least of their problems. The right side of the developing embryo is duplicated on the left so there is liver where the spleen should be, there are two right lungs, and the heart comprises two right atria and two right ventricles.

ACKNOWLEDGMENT

Patrick H. Henry, MD, friend and mentor now deceased, contributed significantly to the chapter in past editions and much of his work remains in this chapter.

FURTHER READING

Barkun AN et al: The bedside assessment of splenic enlargement. Am J Med 91:512, 1991. [PubMed: 1951414]

Facchetti F: Tumors of the spleen. Int J Surg Pathol 18:136S, 2010. [PubMed: 20484279]

Girard E et al: Management of splenic and pancreatic trauma. J Visc Surg 153(suppl 4):45, 2016. [PubMed: 27402320]

Graves SA et al: Does this patient have splenomegaly? JAMA 270:2218, 1993. [PubMed: 8411607]

Kim DK et al: Advisory committee on immunization practices reocommended immunization schedule for adults aged 19 years or older—United States, 2017. MMWR 66:136, 2017. [PubMed: 28182599]

Kraus MD et al: The spleen as a diagnostic specimen: A review of ten years' experience at two tertiary care institutions. Cancer 91:2001, 2001. [PubMed: 11391578]

McIntyre OR, Ebaugh FG Jr: Palpable spleens: Ten year follow-up. Ann Intern Med 90:130, 1979. [PubMed: 420452]

Pangalis GA et al: Clinical approach to lymphadenopathy. Semin Oncol 20:570, 1993. [PubMed: 8296196]

Williamson HA Jr: Lymphadenopathy in a family practice: A descriptive study of 240 cases. J Fam Pract 20:449, 1985. [PubMed: 3989485]

Chapter 63: Principles of Clinical Pharmacology

Dan M. Roden

INTRODUCTION

Drugs are the cornerstone of modern therapeutics. Nevertheless, it is well recognized among healthcare providers and the lay community that the outcome of drug therapy varies widely among individuals. While this variability has been perceived as an unpredictable, and therefore inevitable, accompaniment of drug therapy, this is not the case. The goal of this chapter is to describe the principles of clinical pharmacology that can be used for the safe and optimal use of available and new drugs.

Drugs interact with specific target molecules to produce their beneficial and adverse effects. The chain of events between administration of a drug and production of these effects in the body can be divided into two components, both of which contribute to variability in drug actions. The first component comprises the processes that determine drug delivery to, and removal from, molecular targets. The resulting description of the relationship between drug concentration and time is termed *pharmacokinetics*. The second component of variability in drug action comprises the processes that determine variability in drug actions despite equivalent drug delivery to effector drug sites. This description of the relationship between drug concentration and effect is termed *pharmacodynamics*. As discussed further below, pharmacodynamic variability can arise as a result of variability in function of the target molecule itself or of variability in the broad biologic context in which the drug-target interaction occurs to achieve drug effects.

Two important goals of clinical pharmacology are (1) to provide a description of conditions under which drug actions vary among human subjects; and (2) to determine mechanisms underlying this variability, with the goal of improving therapy with available drugs as well as pointing to mechanisms whose targeting by new drugs may be effective in the treatment of human disease. The drug development process is briefly described at the end of this chapter.

The first steps in the discipline of clinical pharmacology were empirical descriptions of the influence of disease on drug actions and of individuals or families with unusual sensitivities to adverse drug effects. These important descriptive findings are now being replaced by an understanding of the molecular mechanisms underlying variability in drug actions. Importantly, it is often the personal interaction of the patient with the physician or other health care provider that first identifies unusual variability in drug actions; maintained alertness to unusual drug responses continues to be a key component of improving drug safety.

One useful unifying framework is to consider that the effects of disease, drug coadministration, or familial factors in modulating drug action reflect variability in expression or function of specific genes whose products determine pharmacokinetics and pharmacodynamics. This idea forms the basis for pharmacogenomic science; a few examples are cited in this chapter, and further details are addressed in Chap. 64.

GLOBAL CONSIDERATIONS


It is true across all cultures and diseases that factors such as compliance, genetic variants affecting pharmacokinetics, or pharmacodynamics (which themselves vary by ancestry), and drug interactions contribute to drug responses. Cost issues or cultural factors may determine the likelihood that specific drugs, drug combinations, or over-the-counter (OTC) remedies are prescribed. The broad principles of clinical pharmacology enunciated here can be used to analyze the mechanisms underlying successful or unsuccessful therapy with any drug.

INDICATIONS FOR DRUG THERAPY: RISK VERSUS BENEFIT

It is self-evident that the benefits of drug therapy should outweigh the risks. Benefits fall into two broad categories: those designed to alleviate a symptom and those designed to prolong useful life. An increasing emphasis on the principles of evidence-based medicine and techniques such as large clinical trials and meta-analyses has defined benefits of drug therapy in broad patient populations. However, establishing the balance between risk and benefit is not always simple. An increasing body of evidence supports the idea, with which practitioners are very familiar, that individual patients may display responses that are not expected from large population studies and often have comorbidities that typically exclude them from large clinical trials. In addition, therapies that provide symptomatic benefits but shorten life may be entertained in patients with serious and highly symptomatic diseases such as heart failure or cancer. These considerations illustrate the continuing, highly personal nature of the relationship between the prescriber and the patient.

Adverse Effects

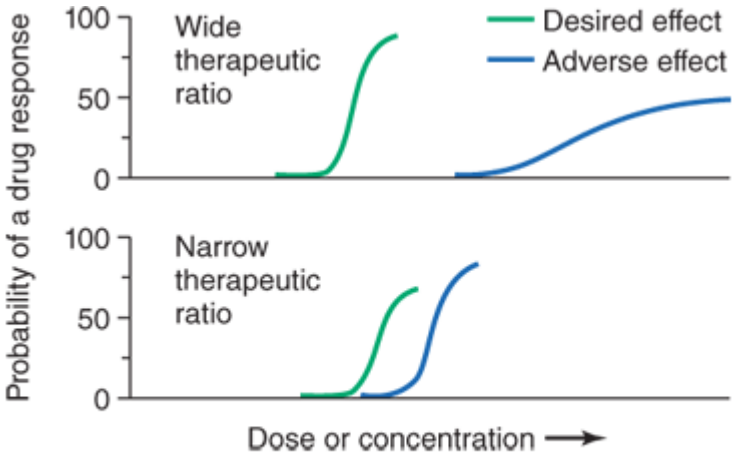
Some adverse effects are so common and so readily associated with drug therapy that they are identified very early during clinical use of a drug. By contrast, serious adverse drug reactions may be sufficiently uncommon that they escape detection for many years after a drug begins to be widely used. The issue of how to identify rare but serious adverse effects (that can profoundly affect the benefit-risk perception in an individual patient) has not been satisfactorily resolved. Potential approaches range from an increased understanding of the molecular and genetic basis of variability in drug actions to expanded post-marketing surveillance mechanisms. None of these have been completely effective, so practitioners must be continuously vigilant to the possibility that unusual symptoms may be related to specific drugs, or combinations of drugs, that their patients receive.

Therapeutic Index

Beneficial and adverse reactions to drug therapy can be described by a series of dose-response relations (Fig. 63-1). Well-tolerated drugs demonstrate a wide margin, termed the *therapeutic ratio*, *therapeutic index*, or *therapeutic window*, between the doses required to produce a therapeutic effect and those producing toxicity. In cases where there is a similar relationship between plasma drug concentration and effects, monitoring plasma concentrations can be a highly effective aid in managing drug therapy by enabling concentrations to be maintained above the minimum required to produce an effect and below the concentration range likely to produce toxicity. Such monitoring has been widely used to guide therapy with specific agents, such as certain antiarrhythmics, anticonvulsants, and antibiotics. Many of the principles in clinical pharmacology and examples outlined below, which can be applied broadly to therapeutics, have been developed in these arenas.

FIGURE 63-1

The concept of a therapeutic ratio. Each panel illustrates the relationship between increasing dose and cumulative probability of a desired or adverse drug effect. ***Top.*** A drug with a wide therapeutic ratio, that is, a wide separation of the two curves. ***Bottom.*** A drug with a narrow therapeutic ratio; here, the likelihood of adverse effects at therapeutic doses is increased because the curves are not well separated. Further, a steep dose-response curve for adverse effects is especially undesirable, as it implies that even small dosage increments may sharply increase the likelihood of toxicity. When there is a definable relationship between drug concentration (usually measured in plasma) and desirable and adverse effect curves, concentration may be substituted on the abscissa. Note that not all patients necessarily demonstrate a therapeutic response (or adverse effect) at any dose, and that some effects (notably some adverse effects) may occur in a dose-independent fashion.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PRINCIPLES OF PHARMACOKINETICS

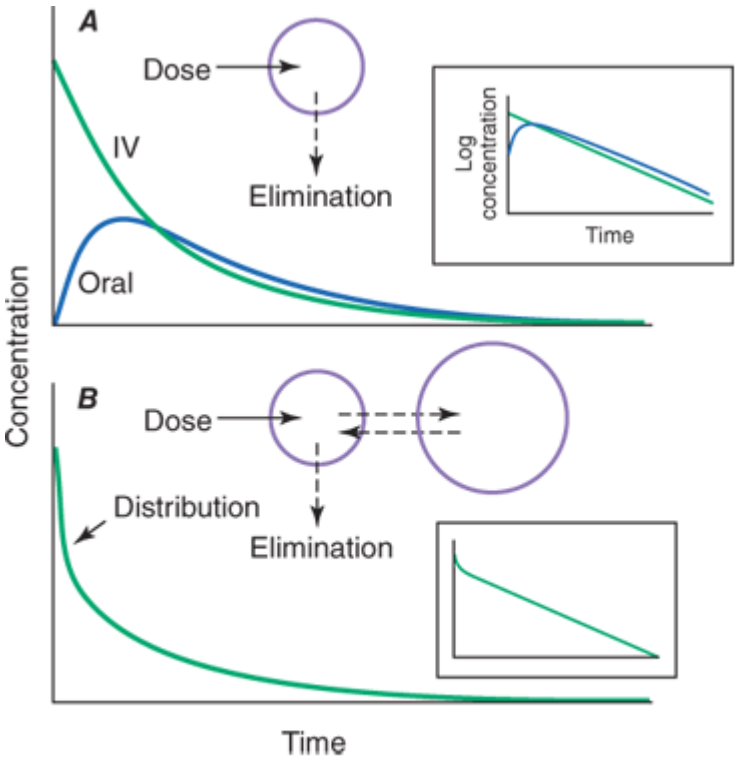
The processes of absorption, distribution, metabolism, and excretion—collectively termed *drug disposition*—determine the concentration of drug delivered to target effector molecules.

ABSORPTION AND BIOAVAILABILITY

When a drug is administered orally, subcutaneously, intramuscularly, rectally, sublingually, or directly into desired sites of action, the amount of drug actually entering the systemic circulation may be less than with the intravenous route (Fig. 63-2A). The fraction of drug available to the systemic circulation by other routes is termed *bioavailability*. Bioavailability may be <100% for two main reasons: (1) absorption is reduced, or (2) the drug undergoes metabolism or elimination prior to entering the systemic circulation. Occasionally, the administered drug formulation is inconsistent or has degraded with time; for example, the anticoagulant dabigatran degrades rapidly (over weeks) once exposed to air, so the amount administered may be less than prescribed.

FIGURE 63-2

Idealized time-plasma concentration curves after a single dose of drug. ***A.*** The time course of drug concentration after an instantaneous IV bolus or an oral dose in the one-compartment model shown. The area under the time-concentration curve is clearly less with the oral drug than the IV, indicating incomplete bioavailability. Note that despite this incomplete bioavailability, concentration after the oral dose can be higher than after the IV dose at some time points. The inset shows that the decline of concentrations over time is linear on a log-linear plot, characteristic of first-order elimination, and that oral and IV drugs have the same elimination (parallel) time course. ***B.*** The decline of central compartment concentration when drug is distributed both to and from a peripheral compartment and eliminated from the central compartment. The rapid initial decline of concentration reflects not drug elimination but distribution.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

When a drug is administered by a non-intravenous route, the peak concentration occurs later and is lower than after the same dose given by rapid intravenous injection, reflecting absorption from the site of administration **(Fig. 63-2)**. The extent of absorption may be reduced because a drug is incompletely released from its dosage form, undergoes destruction at its site of administration, or has physicochemical properties

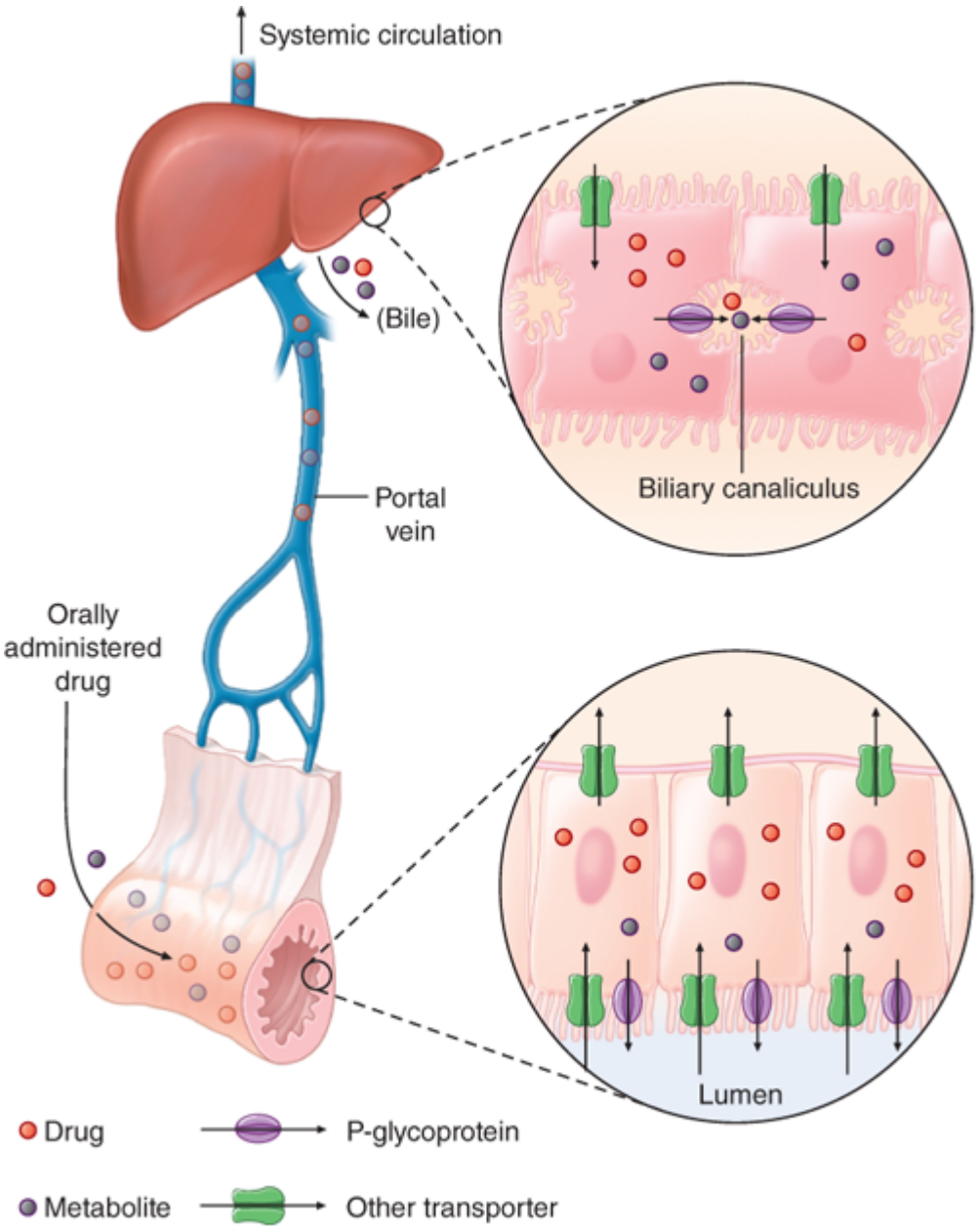
such as insolubility that prevent complete absorption from its site of administration. Slow absorption rates are deliberately designed into “slow-release” or “sustained-release” drug formulations in order to minimize variation in plasma concentrations during the interval between doses.

“First-Pass” Effect

When a drug is administered orally, it must traverse the intestinal epithelium, the portal venous system, and the liver prior to entering the systemic circulation **(Fig. 63-3)**. Once a drug enters the enterocyte, it may undergo metabolism, be transported into the portal vein, or be excreted back into the intestinal lumen. Both excretion into the intestinal lumen and metabolism decrease systemic bioavailability. Once a drug passes this enterocyte barrier, it may also be taken up into the hepatocyte, where bioavailability can be further limited by metabolism or excretion into the bile. This elimination in intestine and liver, which reduces the amount of drug delivered to the systemic circulation, is termed *presystemic elimination*, *presystemic extraction*, or *first-pass elimination*.

FIGURE 63-3

Mechanism of presystemic clearance. After drug enters the enterocyte, it can undergo metabolism, excretion into the intestinal lumen, or transport into the portal vein. Similarly, the hepatocyte may accomplish metabolism and biliary excretion prior to the entry of drug and metabolites to the systemic circulation. *(Adapted by permission from DM Roden, in DP Zipes, J Jalife [eds]: Cardiac Electrophysiology: From Cell to Bedside, 4th ed. Philadelphia, Saunders, 2003. Copyright 2003 with permission from Elsevier.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DRUG TRANSPORT

Drug movement across the membrane of any cell, including enterocytes and hepatocytes, is a combination of passive diffusion and active transport, mediated by specific drug uptake and efflux molecules. One widely studied drug transport molecule is the drug efflux pump P-glycoprotein, the product of the *ABCB1* (or *MDR1*) gene. P-glycoprotein is expressed on the apical aspect of the enterocyte and on the canalicular aspect of the hepatocyte (Fig. 63-3). In both locations, it serves as an efflux pump, limiting availability of drug to the systemic circulation. P-glycoprotein–mediated drug efflux from cerebral capillaries limits drug brain penetration and is an important component of the blood-brain barrier. Other transporters mediate uptake into cells of drugs and endogenous substrates such as vitamins or nutrients.

DRUG METABOLISM

Drug metabolism generates compounds that are usually more polar and, hence, more readily excreted than parent drug. Metabolism takes place predominantly in the liver but can occur at other sites such as kidney, intestinal epithelium, lung, and plasma. “Phase I” metabolism involves chemical modification, most often oxidation accomplished by members of the cytochrome P450 (CYP) monooxygenase superfamily. CYPs and other molecules that are especially important for drug metabolism are presented in **Table 63-1**, and each drug may be a substrate for one or more of these enzymes. “Phase II” metabolism involves conjugation of specific endogenous compounds to drugs or their metabolites. The enzymes that accomplish phase II reactions include glucuronyl-, acetyl-, sulfo-, and methyltransferases. Drug metabolites may exert important pharmacologic activity, as discussed further below.

TABLE 63-1

Molecular Pathways Mediating Drug Disposition

MOLECULE	SUBSTRATES[a]	INHIBITORS[a]
CYP3A	Calcium channel blockers	Amiodarone
	Antiarrhythmics (lidocaine, quinidine, mexiletine)	Ketoconazole, itraconazole
	HMG-CoA reductase inhibitors ("statins"; see text)	Erythromycin, clarithromycin
	Cyclosporine, tacrolimus	Ritonavir
	Indinavir, saquinavir, ritonavir	
CYP2D6[b]	Timolol, metoprolol, carvedilol Propafenone, flecainide Tricyclic antidepressants Fluoxetine, paroxetine	Quinidine (even at ultra-low doses) Tricyclic antidepressants Fluoxetine, paroxetine
CYP2C9[b]	Warfarin	Amiodarone
	Phenytoin	Fluconazole
	Glipizide	Phenytoin
	Losartan	
CYP2C19[b]	Omeprazole	Omeprazole
	Mephenytoin	
	Clopidogrel	
CYP2B6[b]	Efavirenz	
Thiopurine S-methyltransferase[b]	6-Mercaptopurine, azathioprine	

MOLECULE	SUBSTRATES[a]	INHIBITORS[a]
N-acetyltransferase[b]	Isoniazid	
	Procainamide	
	Hydralazine	
	Some sulfonamides	
UGT1A1[b]	Irinotecan	
Pseudocholinesterase[b]	Succinylcholine	
P-glycoprotein	Digoxin	Quinidine
	HIV protease inhibitors	Amiodarone
	Many CYP3A substrates	Verapamil
		Cyclosporine
		Itraconazole
		Erythromycin
SLCO1B1[b]	Simvastatin and some other statins	

[a]Inhibitors affect the molecular pathway, and thus may affect substrate. [b]Clinically important genetic variants described; **see Chap. 64**.

Note: A listing of CYP substrates, inhibitors, and inducers is maintained at *http://medicine.iupui.edu/clinpharm/ddis/main-table*.

Clinical Implications of Altered Bioavailability

Some drugs undergo near-complete presystemic metabolism and thus cannot be administered orally. Nitroglycerin cannot be used orally because it is completely extracted prior to reaching the systemic circulation. The drug is, therefore, used by the sublingual, transdermal, or intravascular routes, which bypass presystemic metabolism.

Some drugs with very extensive presystemic metabolism can still be administered by the oral route, using much higher doses than those required intravenously. Thus, a typical intravenous dose of verapamil is 1–5 mg, compared to a usual single oral dose of 40–120 mg. Administration of low-dose aspirin can result in exposure of cyclooxygenase in platelets in the portal vein to the drug, but systemic sparing because of first-pass aspirin deacylation in the liver. This is an example of presystemic metabolism being exploited to therapeutic advantage.

HALF-LIFE

Most pharmacokinetic processes, such as elimination, are first-order; that is, the rate of the process depends on the amount of drug present. Elimination can occasionally be zero-order (fixed amount eliminated per unit time), and this can be clinically important (see “Principles of Dose Selection”). In the simplest pharmacokinetic model (Fig. 63-2A), a drug bolus (D) is administered instantaneously to a central compartment, from which drug elimination occurs as a first-order process. Occasionally, central and other compartments correspond to physiologic spaces (e.g., plasma volume), whereas in other cases they are simply mathematical functions used to describe drug disposition. The first-order nature of drug elimination leads directly to the relationship describing drug concentration (C) at any time (*t*) following the bolus:

$$C = \frac{D}{V_c} \bullet e^{(-0.69t/t_{1/2})}$$

where V_c is the volume of the compartment into which drug is delivered and $t_{1/2}$ is elimination half-life. As a consequence of this relationship, a plot of the logarithm of concentration versus time is a straight line (Fig. 63-2A, inset). *Half-life* is the time required for 50% of a first-order process to be completed. Thus, 50% of drug elimination is achieved after one drug-elimination half-life, 75% after two, 87.5% after three, etc. In practice, first-order processes such as elimination are near-complete after four–five half-lives.

In some cases, drug is removed from the central compartment not only by elimination but also by distribution into peripheral compartments. In this case, the plot of plasma concentration versus time after a bolus may demonstrate two (or more) exponential components (Fig. 63-2B). In general, the initial rapid drop in drug concentration represents not elimination but drug distribution into and out of peripheral tissues (also first-order processes), while the slower component represents drug elimination; the initial precipitous decline is usually evident with administration by intravenous but not by other routes. Drug concentrations at peripheral sites are determined by a balance between drug distribution to and redistribution from those sites, as well as by elimination. Once distribution is near-complete (four–five distribution half-lives), plasma and tissue concentrations decline in parallel.

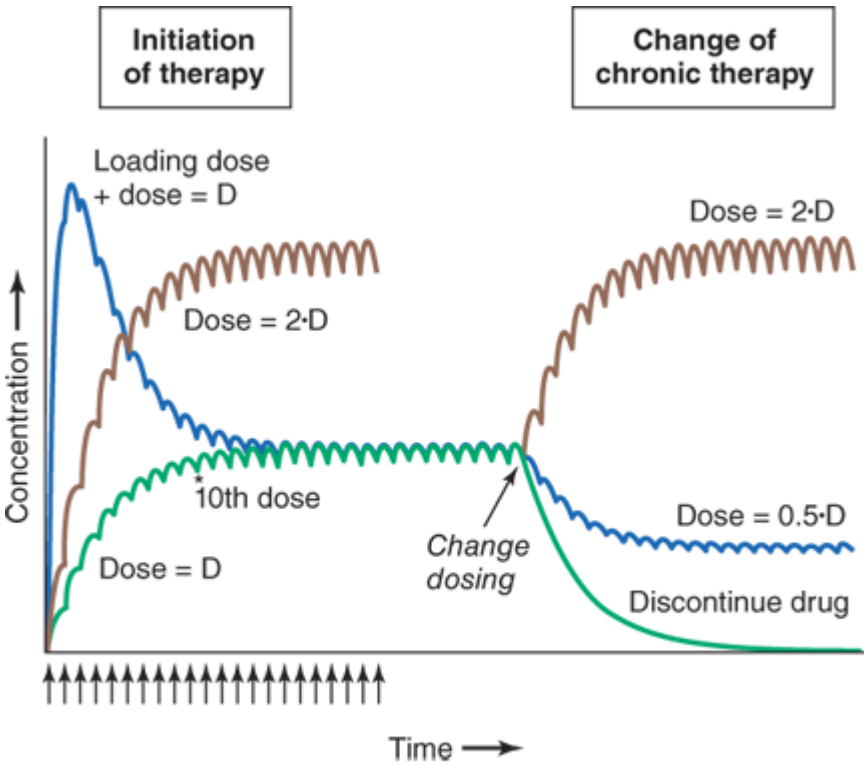
Clinical Implications of Half-Life Measurements

The elimination half-life not only determines the time required for drug concentrations to fall to near-immeasurable levels after a single bolus, it is also the sole determinant of the time required for steady-state plasma concentrations to be achieved after any change in drug dosing **(Fig. 63-4)**. This applies to the

initiation of chronic drug therapy (whether by multiple oral doses or by continuous intravenous infusion), a change in chronic drug dose or dosing interval, or discontinuation of drug.

FIGURE 63-4

Drug accumulation to steady state. In this simulation, drug was administered (arrows) at intervals = 50% of the elimination half-life. Steady state is achieved during initiation of therapy after ~5 elimination half-lives, or 10 doses. A loading dose did not alter the eventual steady state achieved. A doubling of the dose resulted in a doubling of the steady state but the same time course of accumulation. Once steady state is achieved, a change in dose (increase, decrease, or drug discontinuation) results in a new steady state in ~5 elimination half-lives. *(Adapted by permission from DM Roden, in DP Zipes, J Jalife [eds]: Cardiac Electrophysiology: From Cell to Bedside, 4th ed. Philadelphia, Saunders, 2003. Copyright 2003 with permission from Elsevier.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Steady state describes the situation during chronic drug administration when the amount of drug administered per unit time equals drug eliminated per unit time. With a continuous intravenous infusion, plasma concentrations at steady state are stable, while with chronic oral drug administration, plasma concentrations vary during the dosing interval but the time-concentration profile between dosing intervals is stable (Fig. 63-4).

DRUG DISTRIBUTION

In a typical 70-kg human, plasma volume is ~3 L, blood volume is ~5.5 L, and extracellular water outside the vasculature is ~20 L. The volume of distribution of drugs extensively bound to plasma proteins but not to tissue components approaches plasma volume; warfarin is an example. By contrast, for drugs highly bound to tissues, the volume of distribution can be far greater than any physiologic space. For example, the volume

of distribution of digoxin and tricyclic antidepressants is hundreds of liters, obviously exceeding total-body volume. Such drugs are not readily removed by dialysis, an important consideration in overdose.

Clinical Implications of Drug Distribution

In some cases, pharmacologic effects require drug distribution to peripheral sites. In this instance, the time course of drug delivery to and removal from these sites determines the time course of drug effects; anesthetic uptake into the central nervous system (CNS) is an example.

LOADING DOSES

For some drugs, the indication may be so urgent that administration of "loading" dosages is required to achieve rapid elevations of drug concentration and therapeutic effects earlier than with chronic maintenance therapy (Fig. 63-4). Nevertheless, the time required for true steady state to be achieved is still determined only by the elimination half-life.

RATE OF INTRAVENOUS DRUG ADMINISTRATION

Although the simulations in Fig. 63-2 use a single intravenous bolus, this is usually inappropriate in practice because side effects related to transiently very high concentrations can result. Rather, drugs are more usually administered orally or as a slower intravenous infusion. Some drugs are so predictably lethal when infused too rapidly that special precautions should be taken to prevent accidental boluses. For example, solutions of potassium for intravenous administration >20 mEq/L should be avoided in all but the most exceptional and carefully monitored circumstances. This minimizes the possibility of cardiac arrest due to accidental increases in infusion rates of more concentrated solutions.

Transiently high drug concentrations after rapid intravenous administration can occasionally be used to advantage. The use of midazolam for intravenous sedation, for example, depends upon its rapid uptake by the brain during the distribution phase to produce sedation quickly, with subsequent egress from the brain during the redistribution of the drug as equilibrium is achieved.

Similarly, adenosine must be administered as a rapid bolus in the treatment of reentrant supraventricular tachycardias **(Chap. 241)** to prevent elimination by very rapid ($t_{1/2}$ of seconds) uptake into erythrocytes and endothelial cells before the drug can reach its clinical site of action, the atrioventricular node.

Clinical Implications of Altered Protein Binding

Many drugs circulate in the plasma partly bound to plasma proteins. Since only unbound (free) drug can distribute to sites of pharmacologic action, drug response is related to the free rather than the total circulating plasma drug concentration. In chronic kidney or liver disease, protein binding may be decreased and thus drug actions increased. In some situations (myocardial infarction, infection, surgery), acute phase reactants transiently increase binding of some drugs and thus decrease efficacy. These changes assume the greatest clinical importance for drugs that are highly protein-bound since even a small change in protein binding can result in large changes in free drug; for example, a decrease in binding from 99 to 98% doubles

the free drug concentration from 1 to 2%. For some drugs (e.g., phenytoin), monitoring free rather than total drug concentrations can be useful.

DRUG ELIMINATION

Drug elimination reduces the amount of drug in the body over time. An important approach to quantifying this reduction is to consider that drug concentrations at the beginning and end of a time period are unchanged and that a specific volume of the body has been "cleared" of the drug during that time period. This defines clearance as volume/time. Clearance includes both drug metabolism and excretion.

Clinical Implications of Altered Clearance

While elimination half-life determines the time required to achieve steady-state plasma concentration (C_{ss}), the *magnitude* of that steady state is determined by clearance (Cl) and dose alone. For a drug administered as an intravenous infusion, this relationship is:

C_{ss} = dosing rate/Cl or dosing rate = $Cl \bullet C_{ss}$

When drug is administered orally, the average plasma concentration within a dosing interval ($C_{avg,ss}$) replaces C_{ss}, and the dosage (dose per unit time) must be increased if bioavailability (F) is <100%:

Dose/time = $Cl \bullet C_{avg,ss}/F$

Genetic variants, drug interactions, or diseases that reduce the activity of drug-metabolizing enzymes or excretory mechanisms lead to decreased clearance and, hence, a requirement for downward dose adjustment to avoid toxicity. Conversely, some drug interactions and genetic variants increase the function of drug elimination pathways, and hence, increased drug dosage is necessary to maintain a therapeutic effect.

ACTIVE DRUG METABOLITES

Metabolites may produce effects similar to, overlapping with, or distinct from those of the parent drug. Accumulation of the major metabolite of procainamide, *N*-acetylprocainamide (NAPA), likely accounts for marked QT prolongation and torsades des pointes ventricular tachycardia **(Chap. 247)** during therapy with procainamide. Neurotoxicity during therapy with the opioid analgesic meperidine is likely due to accumulation of normeperidine, especially in renal disease.

Prodrugs are inactive compounds that require metabolism to generate active metabolites that mediate the drug effects. Examples include many angiotensin-converting enzyme (ACE) inhibitors, the angiotensin receptor blocker losartan, the antineoplastic irinotecan, the anti-estrogen tamoxifen, the analgesic codeine

(whose active metabolite morphine probably underlies the opioid effect during codeine administration), and the antiplatelet drug clopidogrel. Drug metabolism has also been implicated in bioactivation of procarcinogens and in generation of reactive metabolites that mediate certain adverse drug effects (e.g., acetaminophen hepatotoxicity, discussed below).

THE CONCEPT OF HIGH-RISK PHARMACOKINETICS

When plasma concentrations of active drug depend exclusively on a single metabolic pathway, any condition that inhibits that pathway (be it disease-related, genetic, or due to a drug interaction) can lead to dramatic changes in drug concentrations and marked variability in drug action. Two mechanisms can generate highly variable drug concentrations and effects through such "high-risk pharmacokinetics." *First*, variability in bioactivation of a prodrug can lead to striking variability in drug action; examples include decreased CYP2D6 activity, which prevents analgesia by codeine, and decreased CYP2C19 activity, which reduces the antiplatelet effects of clopidogrel. The s*econd* setting is drug elimination that relies on a single pathway. In this case, inhibition of the elimination pathway by genetic variants or by administration of inhibiting drugs leads to marked elevation of drug concentration and, for drugs with a narrow therapeutic window, an increased likelihood of dose-related toxicity. The active S-enantiomer of the anticoagulant warfarin is eliminated by CYP2C9, and co-administration of amiodarone or phenytoin, CYP2C9 inhibitors, may therefore increase the risk of bleeding unless the dose is decreased. When drugs undergo elimination by multiple-drug metabolizing or excretory pathways, absence of one pathway (due to a genetic variant or drug interaction) is much less likely to have a large impact on drug concentrations or drug actions.

PRINCIPLES OF PHARMACODYNAMICS

The Onset of Drug Action

For drugs used in the urgent treatment of acute symptoms, little or no delay is anticipated (or desired) between the drug-target interaction and the development of a clinical effect. Examples of such acute situations include vascular thrombosis, shock, or status epilepticus.

For many conditions, however, the indication for therapy is less urgent, and a delay between the interaction of a drug with its pharmacologic target(s) and a clinical effect is clinically acceptable. Common pharmacokinetic mechanisms that can contribute to such a delay include slow elimination (resulting in slow accumulation to steady state), uptake into peripheral compartments, or accumulation of active metabolites. A common pharmacodynamic explanation for such a delay is that the clinical effect develops as a downstream consequence of the initial molecular effect the drug produces. Thus, administration of a proton pump inhibitor or an H_2-receptor blocker produces an immediate increase in gastric pH but ulcer healing that is delayed. Cancer chemotherapy similarly produces delayed therapeutic effects.

Drug Effects May Be Disease Specific

A drug may produce no action or a different spectrum of actions in unaffected individuals compared to patients with underlying disease. Further, concomitant disease can complicate interpretation of response to drug therapy, especially adverse effects. For example, high doses of anticonvulsants such as phenytoin may cause neurologic symptoms, which may be confused with the underlying neurologic disease. Similarly, increasing dyspnea in a patient with chronic lung disease receiving amiodarone therapy could be due to the drug, underlying disease, or an intercurrent cardiopulmonary problem. As a result, alternate antiarrhythmic therapies are preferable in patients with chronic lung disease.

While drugs interact with specific molecular receptors, drug effects may vary over time, even if stable drug and metabolite concentrations are maintained. The drug-receptor interaction occurs in a complex biologic milieu that can vary to modulate the drug effect. For example, ion channel blockade by drugs, an important anticonvulsant and antiarrhythmic effect, is often modulated by membrane potential, itself a function of factors such as extracellular potassium or local ischemia. Receptors may be up- or down-regulated by disease or by the drug itself. For example, β-adrenergic blockers upregulate β-receptor density during chronic therapy. While this effect does not usually result in resistance to the therapeutic effect of the drugs, it may produce severe agonist-mediated effects (such as hypertension or tachycardia) if the blocking drug is abruptly withdrawn.

PRINCIPLES OF DOSE SELECTION

The desired goal of therapy with any drug is to maximize the likelihood of a beneficial effect while minimizing the risk of adverse effects. Previous experience with the drug, in controlled clinical trials or in post-marketing use, defines the relationships between dose or plasma concentration and these dual effects (Fig. 63-1) and has important implications for initiation of drug therapy:

1. *The target drug effect should be defined when drug treatment is started*. With some drugs, the desired effect may be difficult to measure objectively, or the onset of efficacy can be delayed for weeks or months; drugs used in the treatment of cancer and psychiatric disease are examples. Sometimes a drug is used to treat a symptom, such as pain or palpitations, and here it is the patient who will report whether the selected dose is effective. In yet other settings, such as anticoagulation or hypertension, the desired response can be repeatedly and objectively assessed by simple clinical or laboratory tests.

2. *The nature of anticipated toxicity often dictates the starting dose*. If side effects are minor, it may be acceptable to start chronic therapy at a dose highly likely to achieve efficacy and down-titrate if side effects occur. However, this approach is rarely, if ever, justified if the anticipated toxicity is serious or life-threatening; in this circumstance, it is more appropriate to initiate therapy with the lowest dose that may produce a desired effect. In cancer chemotherapy, it is common practice to use maximum-tolerated doses.

3. *The above considerations do not apply if these relationships between dose and effects cannot be defined*. This is especially relevant to some adverse drug effects (discussed further below) whose development is not readily related to drug dose.

4. *If a drug dose does not achieve its desired effect, a dosage increase is justified only if toxicity is absent and the likelihood of serious toxicity is small.*

Failure of Efficacy

Assuming the diagnosis is correct and the correct drug is prescribed, explanations for failure of efficacy include drug interactions, noncompliance, or unexpectedly low drug concentration due to administration of expired or degraded drug. These are situations in which measurement of plasma drug concentrations, if available, can be especially useful. Noncompliance is an especially frequent problem in the long-term treatment of diseases such as hypertension and epilepsy, occurring in ≥25% of patients in therapeutic environments in which no special effort is made to involve patients in the responsibility for their own health. Multidrug regimens with multiple doses per day are especially prone to noncompliance.

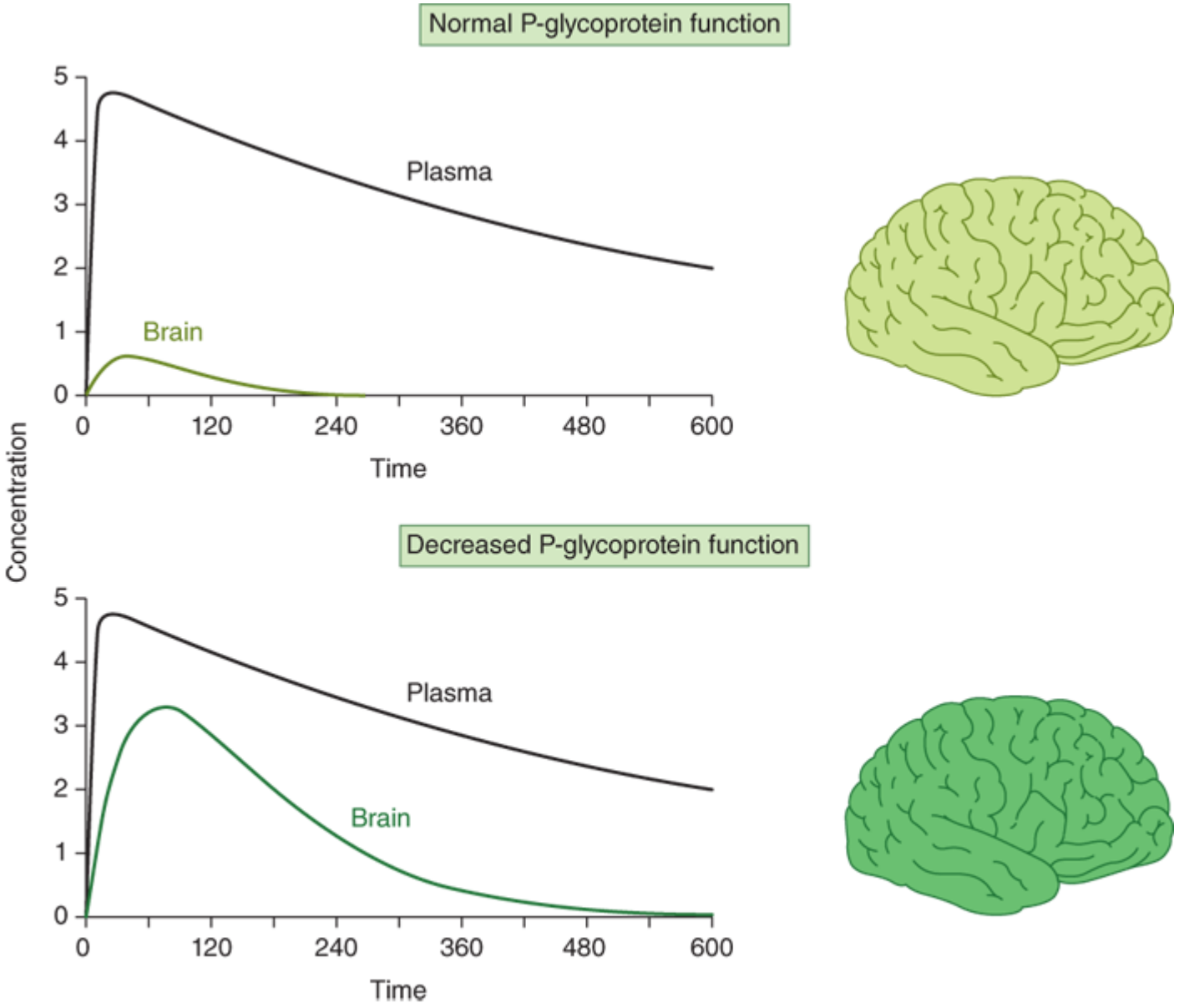
Monitoring response to therapy, by physiologic measures or by plasma concentration measurements, requires an understanding of the relationships between plasma concentration and anticipated effects. For example, measurement of QT interval is used during treatment with sotalol or dofetilide to avoid marked QT prolongation that can herald serious arrhythmias. In this setting, evaluating the electrocardiogram at the time of anticipated peak plasma concentration and effect (e.g., 1–2 h post-dose at steady state) is most appropriate. Maintained high vancomycin levels carry a risk of nephrotoxicity, so dosages should be adjusted on the basis of plasma concentrations measured at trough (pre-dose). Similarly, for dose adjustment of other drugs (e.g., anticonvulsants), concentration should be measured at its lowest during the dosing interval, just prior to a dose at steady state (Fig. 63-4), to ensure a maintained therapeutic effect.

Concentration of Drugs in Plasma as a Guide to Therapy

Factors such as interactions with other drugs, disease-induced alterations in elimination and distribution, and genetic variation in drug disposition combine to yield a wide range of plasma levels in patients given the same dose. Hence, if a predictable relationship can be established between plasma drug concentration and beneficial or adverse drug effect, measurement of plasma levels can provide a valuable tool to guide selection of an optimal dose, especially when there is a narrow range between the plasma levels yielding therapeutic and adverse effects. Monitoring is commonly used with certain types of drugs including many anticonvulsants, antirejection agents, antiarrhythmics, and antibiotics. By contrast, if no such relationship can be established (e.g., if drug access to important sites of action outside plasma is highly variable), monitoring plasma concentration may not provide an accurate guide to therapy **(Fig. 63-5)**.

FIGURE 63-5

The efflux pump P-glycoprotein excludes drugs from the endothelium of capillaries in the brain and so constitutes a key element of the blood-brain barrier. Thus, reduced P-glycoprotein function (e.g., due to drug interactions) increases penetration of substrate drugs into the brain, even when plasma concentrations are unchanged.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The common situation of first-order elimination implies that average, maximum, and minimum steady-state concentrations are related linearly to the dosing rate. Accordingly, the maintenance dose may be adjusted on the basis of the ratio between the desired and measured concentrations *at steady state*; for example, if a doubling of the steady-state plasma concentration is desired, the dose should be doubled. This does not apply to drugs eliminated by zero-order kinetics (fixed amount per unit time), where small dosage increases will produce disproportionate increases in plasma concentration; examples include phenytoin and theophylline.

An increase in dosage is usually best achieved by changing the drug dose but not the dosing interval (e.g., by giving 200 mg every 8 h instead of 100 mg every 8 h). However, this approach is acceptable only if the resulting maximum concentration is not toxic and the trough value does not fall below the minimum effective concentration for an undesirable period of time. Alternatively, the steady state may be changed by altering the frequency of intermittent dosing but not the size of each dose. In this case, the magnitude of the

fluctuations around the average steady-state level will change—the shorter the dosing interval, the smaller the difference between peak and trough levels.

EFFECTS OF DISEASE ON DRUG CONCENTRATION AND RESPONSE

RENAL DISEASE

Renal excretion of parent drug and metabolites is generally accomplished by glomerular filtration and by specific drug transporters. If a drug or its metabolites are primarily excreted through the kidneys and increased drug levels are associated with adverse effects (an example of “high-risk pharmacokinetics” described above), drug dosages must be reduced in patients with renal dysfunction to avoid toxicity. The antiarrhythmics dofetilide and sotalol undergo predominant renal excretion and carry a risk of QT prolongation and arrhythmias if doses are not reduced in renal disease. In end-stage renal disease, sotalol has been given as 40 mg after dialysis (every second day), compared to the usual daily dose, 80–120 mg every 12 h. At approved doses, the anticoagulant edoxaban appears to be somewhat more effective in subjects with mild renal dysfunction, possibly reflecting higher drug levels. The narcotic analgesic meperidine undergoes extensive hepatic metabolism, so that renal failure has little effect on its plasma concentration. However, its metabolite, normeperidine, does undergo renal excretion, accumulates in renal failure, and probably accounts for the signs of CNS excitation, such as irritability, twitching, and seizures, that appear when multiple doses of meperidine are administered to patients with renal disease. Protein binding of some drugs (e.g., phenytoin) may be altered in uremia, so measuring free drug concentration may be desirable.

In non-end-stage renal disease, changes in renal drug clearance are generally proportional to those in creatinine clearance, which may be measured directly or estimated from the serum creatinine. This estimate, coupled with the knowledge of how much drug is normally excreted renally versus non-renally, allows an estimate of the dose adjustment required. In practice, most decisions involving dosing adjustment in patients with renal failure use published recommended adjustments in dosage or dosing interval based on the severity of renal dysfunction indicated by creatinine clearance. Any such modification of dose is a first approximation and should be followed by plasma concentration data (if available) and clinical observation to further optimize therapy for the individual patient.

LIVER DISEASE

Standard tests of liver function are not useful in adjusting doses in diseases like hepatitis or cirrhosis. First-pass metabolism may decrease, leading to increased oral bioavailability as a consequence of disrupted hepatocyte function, altered liver architecture, and portacaval shunts. The oral bioavailability for high first-pass drugs such as morphine, meperidine, midazolam, and nifedipine is almost doubled in patients with cirrhosis, compared to those with normal liver function. Therefore, the size of the oral dose of such drugs should be reduced in this setting.

HEART FAILURE AND SHOCK

Under conditions of decreased tissue perfusion, the cardiac output is redistributed to preserve blood flow to the heart and brain at the expense of other tissues (Chap. 252). As a result, drugs may be distributed into a smaller volume of distribution, higher drug concentrations will be present in the plasma, and the tissues that are best perfused (the brain and heart) will be exposed to these higher concentrations, resulting in increased CNS or cardiac effects. As well, decreased perfusion of the kidney and liver may impair drug clearance. Another consequence of severe heart failure is decreased gut perfusion, which may reduce drug absorption and, thus, lead to reduced or absent effects of orally administered therapies.

DRUG USE IN THE ELDERLY

In the elderly, multiple pathologies and medications used to treat them result in more drug interactions and adverse effects. Aging also results in changes in organ function, especially of the organs involved in drug disposition. Initial doses should be less than the usual adult dosage and should be increased slowly. The number of medications, and doses per day, should be kept as low as possible.

Even in the absence of kidney disease, renal clearance may be reduced by 35–50% in elderly patients. Dosages should be adjusted on the basis of creatinine clearance. Aging also results in a decrease in the size of, and blood flow to, the liver and possibly in the activity of hepatic drug-metabolizing enzymes; accordingly, the hepatic clearance of some drugs is impaired in the elderly. As with liver disease, these changes are not readily predicted.

Elderly patients may display altered drug sensitivity. Examples include increased analgesic effects of opioids, increased sedation from benzodiazepines and other CNS depressants, and increased risk of bleeding while receiving anticoagulant therapy, even when clotting parameters are well controlled. Exaggerated responses to cardiovascular drugs are also common because of the impaired responsiveness of normal homeostatic mechanisms. Conversely, the elderly display decreased sensitivity to β-adrenergic receptor blockers.

Adverse drug reactions are especially common in the elderly because of altered pharmacokinetics and pharmacodynamics, the frequent use of multidrug regimens, and concomitant disease. For example, use of long half-life benzodiazepines is linked to the occurrence of hip fractures in elderly patients, perhaps reflecting both a risk of falls from these drugs (due to increased sedation) and the increased incidence of osteoporosis in elderly patients. In population surveys of the noninstitutionalized elderly, as many as 10% had at least one adverse drug reaction in the previous year.

DRUG USE IN CHILDREN

While most drugs used to treat disease in children are the same as those in adults, there are few studies that provide solid data to guide dosing. Drug metabolism pathways mature at different rates after birth, and disease mechanisms may be different in children. In practice, doses are adjusted for size (weight or body surface area) as a first approximation unless age-specific data are available.

INTERACTIONS BETWEEN DRUGS

Drug interactions can complicate therapy by increasing or decreasing the action of a drug; interactions may be based on changes in drug disposition or in drug response in the absence of changes in drug levels. *Interactions must be considered in the differential diagnosis of any unusual response occurring during drug therapy*. Prescribers should recognize that patients often come to them with a legacy of drugs acquired during previous medical experiences, often with multiple physicians who may not be aware of all the patient's medications. A meticulous drug history should include examination of the patient's medications and, if necessary, calls to the pharmacist to identify prescriptions. It should also address the use of agents not often volunteered during questioning, such as OTC drugs, health food supplements, and topical agents such as eye drops. Lists of interactions are available from a number of electronic sources. While it is unrealistic to expect the practicing physician to memorize these, certain drugs consistently run the risk of generating interactions, often by inhibiting or inducing specific drug elimination pathways. Examples are presented below and in Table 63-2. Accordingly, when these drugs are started or stopped, prescribers must be especially alert to the possibility of interactions.

TABLE 63-2

Drugs with a High Risk of Generating Pharmacokinetic Interactions

DRUG	MECHANISM	EXAMPLES
Antacids	Reduced absorption	Antacids/tetracyclines
Bile acid sequestrants		Cholestyramine/digoxin
Proton pump inhibitors	Altered gastric pH	Ketoconazole absorption decreased
H_2-receptor blockers		
Rifampin	Induction of CYPs and/or P-glycoprotein	Decreased concentration and effects of
Carbamazepine		warfarin
Barbiturates		quinidine
Phenytoin		cyclosporine
St. John's wort		losartan
Glutethimide		oral contraceptives
Nevirapine (CYP3A; CYP2B6)		methadone, dabigatran
Tricyclic antidepressants	Inhibitors of CYP2D6	Increased effect of many β blockers
Fluoxetine		Decreased codeine effect; possible decreased tamoxifen effect
Quinidine		

DRUG	MECHANISM	EXAMPLES
Cimetidine	Inhibitor of multiple CYPs	Increased concentration and effects of warfarin theophylline phenytoin
Ketoconazole, itraconazole	Inhibitor of CYP3A	Increased concentration and toxicity of some HMG-CoA reductase inhibitors, colchicine
Erythromycin, clarithromycin		Cyclosporine, cisapride, terfenadine (now withdrawn)
Calcium channel blockers		Increased concentration and effects of indinavir (with ritonavir)
Ritonavir		Decreased clearance and dose requirement for cyclosporine (with calcium channel blockers)
Allopurinol	Xanthine oxidase inhibitor	Azathioprine and 6-mercaptopurine toxicity
Amiodarone	Inhibitor of many CYPs and of P-glycoprotein	Decreased clearance (risk of toxicity) for warfarin digoxin quinidine
Gemfibrozil (and other fibrates)	CYP3A inhibition	Rhabdomyolysis when co-prescribed with some HMG-CoA reductase inhibitors
Quinidine	P-glycoprotein inhibition	Risk of toxicity with P-glycoprotein substrates (e.g., digoxin, dabigatran)
Amiodarone		
Verapamil		
Cyclosporine		
Itraconazole		
Erythromycin		

DRUG	MECHANISM	EXAMPLES
Phenylbutazone	Inhibition of renal tubular transport	Increased risk of methotrexate toxicity with salicylates
Probenecid		
Salicylates		

PHARMACOKINETIC INTERACTIONS CAUSING DECREASED DRUG EFFECTS

Gastrointestinal absorption can be reduced if a drug interaction results in drug binding in the gut, as with aluminum-containing antacids, kaolin-pectin suspensions, or bile acid sequestrants. Drugs such as histamine H_2-receptor antagonists or proton pump inhibitors that alter gastric pH may decrease the solubility and hence absorption of weak bases such as ketoconazole.

Expression of some genes responsible for drug elimination, notably *CYP3A* and *ABCB1*, can be markedly increased by inducing drugs, such as rifampin, carbamazepine, phenytoin, St. John's wort, and glutethimide, and by smoking, exposure to chlorinated insecticides, and chronic alcohol ingestion. Administration of inducing agents lowers plasma levels, and thus effects, over 2–3 weeks as gene expression is increased. If a drug dose is stabilized in the presence of an inducer that is subsequently stopped, major toxicity can occur as clearance returns to preinduction levels and drug concentrations rise. Individuals vary in the extent to which drug metabolism can be induced, likely through genetic mechanisms.

Interactions that inhibit the bioactivation of prodrugs will decrease drug effects (Table 63-1).

Interactions that decrease drug delivery to intracellular sites of action can decrease drug effects: tricyclic antidepressants can blunt the antihypertensive effect of clonidine by decreasing its uptake into adrenergic neurons. Reduced CNS penetration of multiple human immunodeficiency virus (HIV) protease inhibitors (with the attendant risk of facilitating viral replication in a sanctuary site) appears attributable to P-glycoprotein-mediated exclusion of the drug from the CNS; indeed, inhibition of P-glycoprotein has been proposed as a therapeutic approach to enhance drug entry to the CNS (Fig. 63-5).

PHARMACOKINETIC INTERACTIONS CAUSING INCREASED DRUG EFFECTS

The most common mechanism here is inhibition of drug elimination. In contrast to induction, new protein synthesis is not involved, and the effect develops as drug and any metabolites accumulate (a function of their elimination half-lives). Since shared substrates of a single enzyme can compete for access to the active site of the protein, many CYP substrates are also inhibitors. However, some drugs are especially potent as inhibitors (and occasionally may not even be substrates) of specific drug elimination pathways, and so it is in the use of these agents that clinicians must be most alert to the potential for interactions (Table 63-2).

Commonly implicated interacting drugs of this type include amiodarone, cimetidine, erythromycin and some other macrolide antibiotics (clarithromycin but not azithromycin), ketoconazole and other azole antifungals, the antiretroviral agent ritonavir, and high concentrations of grapefruit juice. The consequences of such interactions will depend on the drug whose elimination is being inhibited (see “The Concept of High-Risk Pharmacokinetics,” above). Examples include CYP3A inhibitors increasing the risk of cyclosporine toxicity or of rhabdomyolysis with some 3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA) reductase inhibitors (lovastatin, simvastatin, atorvastatin, but not pravastatin), and P-glycoprotein inhibitors increasing the risk of toxicity with digoxin therapy or of bleeding with the thrombin inhibitor dabigatran.

These interactions can occasionally be exploited to therapeutic benefit. The antiviral ritonavir is a very potent CYP3A4 inhibitor that has been added to anti-HIV regimens, not because of its antiviral effects but because it decreases clearance, and hence increases efficacy, of other anti-HIV agents. Similarly, calcium channel blockers have been deliberately coadministered with cyclosporine to reduce its clearance and thus its maintenance dosage and cost.

Phenytoin, an inducer of many systems, including CYP3A, inhibits CYP2C9 and thus can reduce the bioactivation of losartan, with potential loss of antihypertensive effect, or the elimination of S-warfarin, with attendant increased bleeding risk.

Grapefruit (but not orange) juice inhibits CYP3A, especially at high doses; patients receiving drugs where even modest CYP3A inhibition may increase the risk of adverse effects (e.g., cyclosporine, some HMG-CoA reductase inhibitors) should therefore avoid grapefruit juice.

CYP2D6 is markedly inhibited by quinidine, a number of neuroleptic drugs (chlorpromazine and haloperidol), and the selective serotonin reuptake inhibitors (SSRIs) fluoxetine and paroxetine. The clinical consequences of fluoxetine’s interaction with CYP2D6 substrates may not be apparent for weeks after the drug is started, because of its very long half-life and slow generation of a CYP2D6-inhibiting metabolite.

Azathioprine is metabolized to 6-mercaptopurine, which is then metabolized by thiopurine methyltransferase and by xanthine oxidase. When allopurinol, an inhibitor of xanthine oxidase, is administered with standard doses of azathioprine or 6-mercaptopurine, life-threatening toxicity (bone marrow suppression) can result.

A number of drugs are secreted by the renal tubular transport systems for organic anions. Inhibition of these systems can cause excessive drug accumulation. Salicylate, for example, reduces the renal clearance of methotrexate, an interaction that may lead to methotrexate toxicity. Renal tubular secretion contributes substantially to the elimination of penicillin, which can be inhibited (to increase its therapeutic effect) by probenecid. Similarly, inhibition of tubular cation transport by cimetidine decreases the renal clearance of dofetilide.

DRUG INTERACTIONS NOT MEDIATED BY CHANGES IN DRUG DISPOSITION

Drugs may act on separate components of a common process to generate effects greater than either has alone. While antithrombotic therapy with combinations of antiplatelet agents (glycoprotein IIb/IIIa inhibitors, aspirin, clopidogrel) and anticoagulants (e.g., warfarin, heparins, dabigatran, apixaban, rivaraxaban, edoxaban) is often used in the treatment of vascular disease, such combinations do carry an increased risk of bleeding.

Nonsteroidal anti-inflammatory drugs (NSAIDs) cause gastric ulcers, and in patients treated with oral anticoagulants, the risk of upper gastrointestinal bleeding is increased almost threefold by concomitant use of an NSAID.

Indomethacin, piroxicam, and probably other NSAIDs antagonize the antihypertensive effects of β-adrenergic receptor blockers, diuretics, ACE inhibitors, and other drugs. The resulting elevation in blood pressure ranges from trivial to severe. This effect is not seen with aspirin and sulindac but has been found with the cyclooxygenase 2 (COX-2) inhibitor celecoxib.

Torsades de pointes ventricular tachycardia during administration of QT-prolonging antiarrhythmics (quinidine, sotalol, dofetilide) occurs much more frequently in patients receiving diuretics, probably reflecting hypokalemia. Low potassium not only prolongs the QT interval in the absence of drug but also potentiates drug block of ion channels that results in QT prolongation. Also, some diuretics have direct electrophysiologic actions that prolong QT.

The administration of supplemental potassium leads to more frequent and more severe hyperkalemia when potassium elimination is reduced by concurrent treatment with ACE inhibitors, spironolactone, eplerenone, amiloride, or triamterene.

The pharmacologic effects of sildenafil result from inhibition of the phosphodiesterase type 5 isoform that inactivates cyclic guanosine monophosphate (GMP) in the vasculature. Nitroglycerin and related nitrates used to treat angina produce vasodilation by elevating cyclic GMP. Thus, coadministration of these nitrates with sildenafil can cause profound hypotension, which can be catastrophic in patients with coronary disease.

Sometimes, combining drugs can increase overall efficacy and/or reduce drug-specific toxicity. Such therapeutically useful interactions are described in chapters dealing with specific disease entities.

ADVERSE DRUG REACTIONS

The beneficial effects of drugs are coupled with the inescapable risk of untoward effects. The morbidity and mortality from these adverse effects often present diagnostic problems because they can involve every organ and system of the body and may be mistaken for signs of underlying disease. As well, some surveys have suggested that drug therapy for a range of chronic conditions such as psychiatric disease or hypertension does not achieve its desired goal in up to half of treated patients; thus, the most common “adverse” drug effect may be failure of efficacy.

Adverse reactions can be classified in two broad groups. Type A reactions result from exaggeration of an intended pharmacologic action of the drug, such as increased bleeding with anticoagulants or bone marrow suppression with some antineoplastics. Type B reactions result from toxic effects unrelated to the intended pharmacologic actions. The latter effects are often unanticipated (especially with new drugs) and frequently severe and may result from recognized (often immunologic) as well as previously undescribed mechanisms.

Drugs may increase the frequency of an event that is common in a general population, and this may be especially difficult to recognize; an excellent example is the increase in myocardial infarctions with the COX-2 inhibitor rofecoxib. Drugs can also cause rare and serious adverse effects, such as hematologic abnormalities, arrhythmias, severe skin reactions, or hepatic or renal dysfunction. Prior to regulatory approval and marketing (see below), new drugs are tested in relatively few patients who tend to be less sick and to have fewer concomitant diseases than those patients who subsequently receive the drug therapeutically. Because of the relatively small number of patients studied in clinical trials and the selected nature of these patients, rare adverse effects are generally not detected prior to a drug's approval; indeed, if they are detected, the new drugs are generally not approved. Therefore, physicians need to be cautious in the prescription of new drugs and alert for the appearance of previously unrecognized adverse events.

Elucidating mechanisms underlying adverse drug effects can assist development of safer compounds or allow a patient subset at especially high risk to be excluded from drug exposure. National adverse reaction reporting systems, such as those operated by the FDA (suspected adverse reactions can be reported online at *http://www.fda.gov/safety/medwatch/default.htm*) and the Committee on Safety of Medicines in Great Britain, can prove useful. The publication or reporting of a newly recognized adverse reaction can in a short time stimulate many similar such reports of reactions that previously had gone unrecognized.

Occasionally, "adverse" effects may be exploited to develop an entirely new indication for a drug. Unwanted hair growth during minoxidil treatment of severely hypertensive patients led to development of the drug for hair growth. Sildenafil was initially developed as an antianginal, but its effects to alleviate erectile dysfunction not only led to a new drug indication but also to increased understanding of the role of type 5 phosphodiesterase in erectile tissue. These examples further reinforce the concept that prescribers must remain vigilant to the possibility that unusual symptoms may reflect unappreciated drug effects.

Some 25–50% of patients make errors in self-administration of prescribed medicines, and these errors can be responsible for adverse drug effects. Similarly, patients commit errors in taking OTC drugs by not reading or following the directions on the containers. Health care providers must recognize that providing directions with prescriptions does not always guarantee compliance.

In hospitals, drugs are administered in a controlled setting, and patient compliance is, in general, ensured. Errors may occur nevertheless—the wrong drug or dose may be given or the drug may be given to the wrong patient—and improved drug distribution and administration systems should help with this problem.

SCOPE OF THE PROBLEM

One estimate in the United Kingdom was that 6.5% of all hospital admissions are due to adverse drug reactions, and that 2.3% of these patients (0.15%) died as a result. The most common culprit drugs were aspirin, other NSAIDs, diuretics, warfarin, ACE inhibitors, antidepressants, opiates, digoxin, steroids, and clopidogrel. One study in the late 1990s suggested that adverse drug reactions were responsible for >100,000 in-hospital deaths in the United States, making them the 4th to 6th commonest cause of in-hospital death. Another study 10 years later showed no change in this trend.

In hospital, patients receive, on average, 10 different drugs during each hospitalization. The sicker the patient, the more drugs are given, and there is a corresponding increase in the likelihood of adverse drug reactions. When <6 different drugs are given to hospitalized patients, the probability of an adverse reaction is ~5%, but if >15 drugs are given, the probability is >40%. Serious adverse reactions are also well-recognized with "herbal" remedies and OTC compounds; examples include kava-associated hepatotoxicity, L-tryptophan-associated eosinophilia-myalgia, and phenylpropanolamine-associated stroke, each of which has caused fatalities.

TOXICITY UNRELATED TO A DRUG'S PRIMARY PHARMACOLOGIC ACTIVITY

Drugs or more commonly reactive metabolites generated by CYPs can covalently bind to tissue macromolecules (such as proteins or DNA) to cause tissue toxicity. Because of the reactive nature of these metabolites, covalent binding often occurs close to the site of production, typically the liver.

Acetaminophen

The most common cause of drug-induced hepatotoxicity is acetaminophen overdosage (Chap. 333). Normally, reactive metabolites are detoxified by combining with hepatic glutathione. When glutathione becomes depleted, the metabolites bind instead to hepatic protein, with resultant hepatocyte damage. The hepatic necrosis produced by the ingestion of acetaminophen can be prevented or attenuated by the administration of substances such as *N*-acetylcysteine that reduce the binding of electrophilic metabolites to hepatic proteins. The risk of acetaminophen-related hepatic necrosis is increased in patients receiving drugs such as phenobarbital or phenytoin, which increase the rate of drug metabolism, or ethanol, which exhausts glutathione stores. Such toxicity has even occurred with therapeutic dosages, so patients at risk through these mechanisms should be warned.

Immunologic Reactions

Most pharmacologic agents are haptens, small molecules with low molecular weights (<2000) that are therefore poor immunogens. Generation of an immune response to a drug therefore often requires in vivo activation and covalent linkage to protein, carbohydrate, or nucleic acid.

Drug stimulation of antibody production may mediate tissue injury by several mechanisms. The antibody may attack the drug when the drug is covalently attached to a cell and thereby destroy the cell. This occurs in penicillin-induced hemolytic anemia. Antibody-drug-antigen complexes may be passively adsorbed by a

bystander cell, which is then destroyed by activation of complement; this occurs in quinine- and quinidine-induced thrombocytopenia. Heparin-induced thrombocytopenia arises when antibodies against complexes of platelet factor 4 peptide and heparin generate immune complexes that activate platelets; thus, the thrombocytopenia is accompanied by "paradoxical" thrombosis and is treated with thrombin inhibitors. Drugs or their reactive metabolites may alter a host tissue, rendering it antigenic and eliciting autoantibodies. For example, hydralazine and procainamide (or their reactive metabolites) can chemically alter nuclear material, stimulating the formation of antinuclear antibodies and occasionally causing lupus erythematosus. Drug-induced pure red cell aplasia (Chap. 98) is due to an immune-based drug reaction.

Serum sickness (Chap. 345) results from the deposition of circulating drug-antibody complexes on endothelial surfaces. Complement activation occurs, chemotactic factors are generated locally, and an inflammatory response develops at the site of complex entrapment. Arthralgias, urticaria, lymphadenopathy, glomerulonephritis, or cerebritis may result. Foreign proteins (vaccines, streptokinase, therapeutic antibodies) and antibiotics are common causes. Many drugs, particularly antimicrobial agents, ACE inhibitors, and aspirin, can elicit anaphylaxis with production of IgE, which binds to mast cell membranes. Contact with a drug antigen initiates a series of biochemical events in the mast cell and results in the release of mediators that can produce the characteristic urticaria, wheezing, flushing, rhinorrhea, and (occasionally) hypotension.

Drugs may also elicit cell-mediated immune responses. One serious reaction is Steven-Johnson Syndrome/Toxic Epidermal Necrolysis (SJS/TEN), which can result in death due to T-cell-mediated massive skin sloughing. As described in Chap. 64, specific genetic variants appear necessary but not sufficient to elicit SJS/TEN. The mechanism is thought to be T cell activation by hapten-"self-peptide" interactions or direct binding of drug to HLA or T cell receptors.

DIAGNOSIS AND TREATMENT OF ADVERSE DRUG REACTIONS

The manifestations of drug-induced diseases frequently resemble those of other diseases, and a given set of manifestations may be produced by different and dissimilar drugs. Recognition of the role of a drug or drugs in an illness depends on appreciation of the possible adverse reactions to drugs in any disease, on identification of the temporal relationship between drug administration and development of the illness, and on familiarity with the common manifestations of the drugs.

A suspected adverse drug reaction developing after introduction of a new drug naturally implicates that drug; however, it is also important to remember that a drug interaction may be responsible. Thus, for example, a patient on a chronic stable warfarin dose may develop a bleeding complication after introduction of amiodarone; this does not reflect a direct reaction to amiodarone but rather its effect to inhibit warfarin metabolism. Many associations between particular drugs and specific reactions have been described, but there is always a "first time" for a novel association, and any drug should be suspected of causing an adverse effect if the clinical setting is appropriate.

Illness related to a drug's intended pharmacologic action is often more easily recognized than illness attributable to immune or other mechanisms.

For example, side effects such as cardiac arrhythmias in patients receiving digitalis, hypoglycemia in patients given insulin, or bleeding in patients receiving anticoagulants are more readily related to a specific drug than are symptoms such rash, which may be caused by many drugs or by other factors. Drug fever often escapes initial diagnosis because fever is such a common manifestation of disease.

Electronic listings of adverse drug reactions can be useful. However, exhaustive compilations often provide little sense of perspective in terms of frequency and seriousness, which can vary considerably among patients.

Eliciting a drug history from each patient is important for diagnosis. Attention must be directed to OTC drugs and herbal preparations as well as to prescription drugs. Each type can be responsible for adverse drug effects, and adverse interactions may occur between OTC drugs and prescribed drugs. Loss of efficacy of oral contraceptives or cyclosporine with concurrent use of St. John's wort (a P-glycoprotein inducer) is an example. In addition, it is common for patients to be cared for by several physicians, and duplicative, additive, antagonistic, or synergistic drug combinations may therefore be administered if the physicians are not aware of the patients' drug histories. Every physician should determine what drugs a patient has been taking, for the previous month or two ideally, before prescribing any medications. Medications stopped for inefficacy or adverse effects should be documented to avoid pointless and potentially dangerous reexposure. A frequently overlooked source of additional drug exposure is topical therapy; for example, a patient complaining of bronchospasm may not mention that an ophthalmic beta blocker is being used unless specifically asked. A history of previous adverse drug effects in patients is common. Since these patients have shown a predisposition to drug-induced illnesses, such a history should dictate added caution in prescribing new drugs.

Laboratory studies may include demonstration of serum antibody in some persons with drug allergies involving cellular blood elements, as in agranulocytosis, hemolytic anemia, and thrombocytopenia. For example, both quinine and quinidine can produce platelet agglutination in vitro in the presence of complement and the serum from a patient who has developed thrombocytopenia following use of this drug. Biochemical abnormalities such as G6PD deficiency, serum pseudocholinesterase level, or genotyping may also be useful in diagnosis, especially after an adverse effect has occurred in the patient or a family member (**see Chap. 64**).

Once an adverse reaction is suspected, discontinuation of the suspected drug followed by disappearance of the reaction is presumptive evidence of a drug-induced illness. Confirming evidence may be sought by cautiously reintroducing the drug and seeing if the reaction reappears. However, that should be done only if confirmation would be useful in the future management of the patient and if the attempt would not entail undue risk. With concentration-dependent adverse reactions, lowering the dosage may cause the reaction to disappear, and raising it may cause the reaction to reappear. When the reaction is thought to be immunologic, however, readministration of the drug may be hazardous, since anaphylaxis may develop.

If the patient is receiving many drugs when an adverse reaction is suspected, the drugs likeliest to be responsible can usually be identified; this should include both potential culprit agents as well as drugs that alter their elimination. All drugs may be discontinued at once or, if this is not practical, discontinued one at a time, starting with the ones most suspect, and the patient observed for signs of improvement. The time needed for a concentration-dependent adverse effect to disappear depends on the time required for the concentration to fall below the range associated with the adverse effect; that, in turn, depends on the initial blood level and on the rate of elimination or metabolism of the drug. Adverse effects of drugs with long half-lives or those not directly related to serum concentration may take a considerable time to disappear.

THE DRUG DEVELOPMENT PROCESS

Drug therapy is an ancient feature of human culture. The first treatments were plant extracts discovered empirically to be effective for indications like fever, pain, or breathlessness. This symptom-based empiric approach to drug development was supplanted in the twentieth century by identification of compounds targeting more fundamental biologic processes, such as bacterial growth or elevated blood pressure. The term “magic bullet,” coined by Paul Ehrlich to describe the search for effective compounds for syphilis, captures the essence of the hope that understanding basic biologic processes will lead to highly effective new therapies.

A common starting point for the development of many widely used modern therapies has been basic biologic discovery that implicates potential target molecules: examples of such target molecules include HMG-CoA reductase, a key step in cholesterol biosynthesis, or the *BRAF* V600E mutation that appears to drive the development of some malignant melanomas and other tumors. The development of compounds targeting these molecules has not only revolutionized treatment for diseases such as hypercholesterolemia or malignant melanoma, but has also revealed new biologic features of disease. Thus, for example, initial spectacular successes with vemurafenib (which targets *BRAF* V600E) were followed by near-universal tumor relapse, strongly suggesting that inhibition of this pathway alone would be insufficient for tumor control. This reasoning, in turn, supports a view that many complex diseases will not lend themselves to cure by targeting a single magic bullet, but rather single drugs or combinations that attack multiple pathways whose perturbation results in disease. The use of combination therapy in settings such as hypertension, tuberculosis, HIV infection, and many cancers highlights the potential for such a “systems biology” view of drug therapy.

A common approach in contemporary drug development is to start with a high-throughput screening procedure to identify “lead” chemical(s) modulating the activity of a potential drug target. The next step is application of increasingly sophisticated medicinal chemistry-based modification of the “lead” to develop compounds with specificity for the chosen target, lack of “off-target” effects, and pharmacokinetic properties suitable for human use (e.g., consistent bioavailability, long elimination half-life, and no high-risk pharmacokinetic features). Drug evaluation in human subjects then proceeds from initial safety and tolerance (phase 1), dose finding (phase 2), and efficacy (phase 3). This is a very expensive process and the vast majority of lead compounds fail at some point. Thus, new approaches to identify likely successes and

failures early are needed. One idea, described further in **Chap. 64**, is to use genomic and other high throughput profiling approaches not only to identify new drug targets but also to identify disease subsets for which drugs approved for other indications might be "repurposed" thereby avoiding the costly development process.

SUMMARY

Modern clinical pharmacology aims to replace empiricism in the use of drugs with therapy based on in-depth understanding of factors that determine an individual's response to drug treatment. Molecular pharmacology, pharmacokinetics, genetics, clinical trials, and the educated prescriber all contribute to this process. No drug response should ever be termed *idiosyncratic*; all responses have a mechanism whose understanding will help guide further therapy with that drug or successors. This rapidly expanding understanding of variability in drug actions makes the process of prescribing drugs increasingly daunting for the practitioner. However, fundamental principles should guide this process:

The benefits of drug therapy, however defined, should always outweigh the risk.

The smallest dosage necessary to produce the desired effect should be used.

The number of medications and doses per day should be minimized.

Although the literature is rapidly expanding, accessing it is becoming easier; electronic tools to search databases of literature and unbiased opinion will become increasingly commonplace.

Genetics play a role in determining variability in drug response and may become a part of clinical practice.

Electronic medical record and pharmacy systems will increasingly incorporate prescribing advice, such as indicated medications not used; unindicated medications being prescribed; and potential dosing errors, drug interactions, or genetically determined drug responses.

Prescribers should be particularly wary when adding or stopping specific drugs that are especially liable to provoke interactions and adverse reactions.

Prescribers should use only a limited number of drugs, with which they are thoroughly familiar.

FURTHER READING

Landrigan CP et al: Temporal trends in rates of patient harm resulting from medical care. N Engl J Med 363:2124, 2010. [PubMed: 21105794]

Lazarou J et al: Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. JAMA 279:1200, 1998. [PubMed: 9555760]

MacRae CA et al: The future of cardiovascular therapeutics. Circulation 133:2610, 2016. [PubMed: 27324356]

Pirmohamed M et al: Adverse drug reactions as cause of admission to hospital: Prospective analysis of 18 820 patients. Br Med J 329:15, 2004.

Wheatley LM et al: Report from the National Institute of Allergy and Infectious Diseases workshop on drug allergy. J Allergy Clin Immunol 136:262, 2015. [PubMed: 26254053]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 64: Pharmacogenomics

Dan M. Roden

FIGURE 64-1

INTRODUCTION

The previous chapter discussed mechanisms underlying variability in drug action, highlighting pharmacokinetic and pharmacodynamic pathways to beneficial and adverse drug events. Work in the past several decades has defined how genetic variation can play a prominent role in modulating these pathways. Initial studies described unusual drug responses due to single genetic variants in individual subjects, defining the field of pharmacogenetics. A more recent view extends this idea to multiple genetic variants across populations, and the term “pharmacogenomics” is often used. Understanding the role of genetic variation in drug response could improve the use of current drugs, avoid drug use in those at increased risk for adverse drug reactions (ADRs), guide development of new drugs, and even be used as a lens through which to understand mechanisms of diseases themselves. This chapter will outline the principles of pharmacogenomics, the evidence as currently available that genetic factors play a role in variable drug actions, and outline areas of controversy and future work.

PRINCIPLES OF GENETIC VARIATION AND DRUG RESPONSE

A goal of traditional Mendelian genetics is to identify DNA variants associated with a distinct phenotype in multiple related family members **(Chap. 457)** (See Also Chaps. 456 and 457). However, it is unusual for a drug response phenotype to be accurately measured in more than one family member, let alone across a kindred. Some clinical studies have examined drug disposition traits (such as urinary drug excretion after a fixed test dose) in twins, and have in some instances shown greater concordance in monozygotic compared to dizygotic pairs, supporting a genetic contribution to the trait under study. However, in general, non-family-based approaches are generally used to identify and validate DNA variants contributing to variable drug actions.

Types of Genetic Variants Influencing Drug Response

The commonest type of genetic variant is a single nucleotide polymorphism (SNP), and nonsynonymous SNPs (i.e., those that alter primary amino acid sequence encoded by a gene) are a common cause of variant function in genes regulating drug responses, often termed *pharmacogenes* **(Table 64-1)**. Small insertions and deletions can similarly alter protein function, or lead to functionally important splice variation. Examples of synonymous coding region variants altering pharmacogene function have also been described; the postulated mechanism is an alteration in the rate of RNA translation, and hence in folding of the nascent protein. Variation in pharmacogene

promoters has been described, and copy number variation (gene deletion or multiple copies of the same gene) is also well described.

TABLE 64-1

Examples of Genetic Variation and Ancestry

Structural Variant	Example		Functional Effect	Minor Allele Frequency (%)[a]		
	Common Name	dbSNP		European	African	East Asian
Single nucleotide polymorphism (SNP) (or single nucleotide variant, SNV)	CYP2C9*2	rs1799853	R144C: Reduction of function	12.7	2.4	b
	CYP2C9*3	rs1057910	I359L: Loss of function	6.9	1.3	3.4
	CYP2C9*8	rs7900194	R150H: Reduction of function	b	5.6	b
	CYP2C19*2	rs4244285	Splicing defect: Loss of function	14.8	18.1	31.0
	CYP2C19*3	rs4986893	Premature stop: Loss of function	b	b	6.7
	CYP2C19*17	rs12248560	Gain of function	45	45	<5
	CYP2D6*4[c]	rs3892097	Splicing defect: Loss of function	23.1	11.9	0.4
	CYP2D6*10[c]	Multiple SNPs define CYP2D6*10 (reduction of function allele):				
		rs1065852	P34S	24.9	15.1	59.1
		rs1135840	S486T			
	CYP3A5*3	rs776746	Splicing defect: Loss of function	90	33	85
	VKORC1*2	rs9923231	Promoter variant associated with decreased warfarin dose	39	11	91

Structural Variant	Example		Functional Effect	Minor Allele Frequency (%)[a]		
	Common Name	dbSNP		European	African	East Asian
	VKORC1	rs61742245	D36Y: Reduction of function, associated with increased warfarin dose	5% in East Africa, Middle East, Oceania; rare elsewhere		
	ABCB1	rs1045642	Synonymous variant; may affect mRNA stability and protein folding	47.2	79.8	62.5
Insertion/deletion	UGT1A1*28		Reduction of function promoter variant (7 TA repeats versus 6 repeats in reference allele); homozygotes have Gilbert's syndrome	31.6	39.1	14.8
Multiple variants constituting specific haplotypes	HLA-B*15:01		Predispose to immunologically mediated adverse drug reactions	b	b	5
	HLA-B*57:01			6.8	1.0	1.6
Gene deletion	CYP2D6*5		Loss of function	2.7	6	5.6
Gene duplication	CYP2D6*1xN	Duplication of normal allele	Ultra-rapid metabolizer phenotype	0.8	1.5	0.3
				Up to 3% in North Africa and the Middle East		
	CYP2D6*4xN	Duplication of loss of function allele	Extensive or poor metabolizer phenotype, depending on the opposite allele	0.3	1.4	b

Note: Allele frequencies from http://exac.broadinstitute.org/ and https://cpicpgx.org/.

[a]Includes heterozygotes and homozygotes. [b]Allele frequency <0.05%. [c]*CYP2D6* is highly polymorphic and multiple SNPs may be required to define a specific variant. For example, rs1065852 is present in both *4 and *10 variants. See http://www.cypalleles.ki.se.

Table 64-1 lists examples of individual types of genomic variation and the impact they can have on function of pharmacogenes. Multiple genotyping approaches may be needed to detect important variants; for example, SNP assays may fail to detect large gene duplications, and highly polymorphic regions (such as human leukocyte antigens, HLA-B) are currently best evaluated by sequencing.

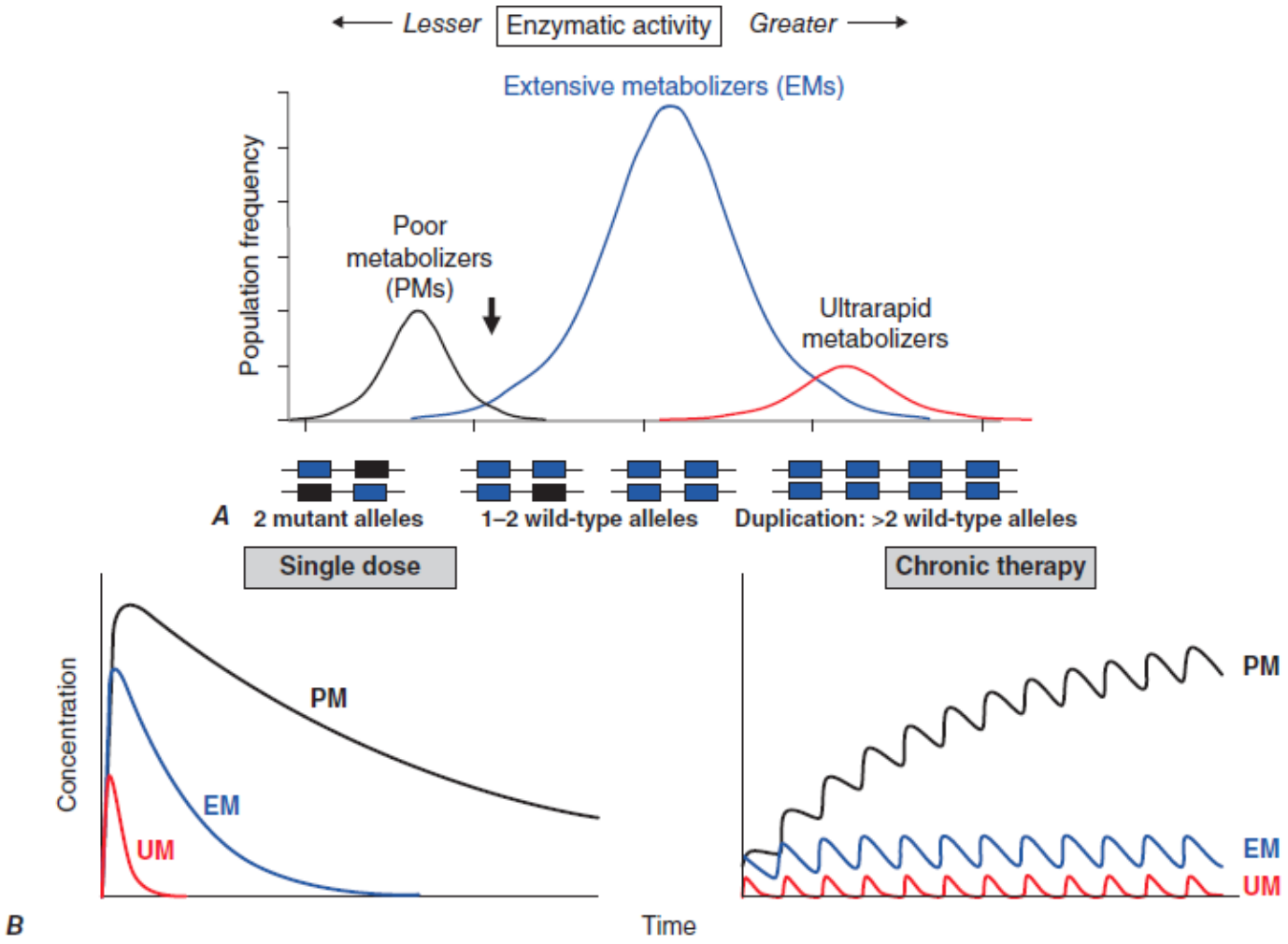
Table 64-1 highlights the fact that the frequency of important variation across pharmacogenes can vary strikingly by ancestry, with the result that certain ethnic groups may be at unusually high risk of displaying variant response to specific drugs.

Candidate Gene Approaches

Most studies to date have used an understanding of the molecular mechanisms modulating drug action to identify candidate genes in which variants could explain variable drug responses. One very common scenario is that variable drug actions can be attributed to variability in plasma drug concentrations. When plasma drug concentrations vary widely (e.g., more than an order of magnitude), especially if their distribution is non-unimodal as in **Fig. 64-1**, variants in single genes controlling drug concentrations often contribute. In this case, the most obvious candidate genes are those responsible for drug metabolism and elimination. Other candidate genes are those encoding the target molecules with which drugs interact to produce their effects or molecules modulating that response, including those involved in disease pathogenesis.

FIGURE 64-1

A. Distribution of CYP2D6 metabolic activity across a population. The heavy arrow indicates an antimode, separating poor metabolizer subjects (PMs, black), with two loss-of-function CYP2D6 alleles (black), indicated by the intron-exon structures below the chart. Individuals with one or two functional alleles are grouped together as extensive metabolizers (EMs, blue). Also shown are ultra-rapid metabolizers (UMs, red), with 2–12 functional copies of the gene, displaying the greatest enzyme activity. *(Adapted from M-L Dahl et al: J Pharmacol Exp Ther 274:516, 1995.)* ***B.*** These simulations show the predicted effects of CYP2D6 genotype on disposition of a substrate drug. With a single dose (left), there is an inverse “gene-dose” relationship between the number of active alleles and the areas under the time-concentration curves (smallest in UM subjects; highest in PM subjects); this indicates that clearance is greatest in UM subjects. In addition, elimination half-life is longest in PM subjects. The right panel shows that these single dose differences are exaggerated during chronic therapy: steady-state concentration is much higher in PM subjects (decreased clearance), as is the time required to achieve steady state (longer elimination half-life).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Genome-Wide Association Studies

The field has also had some success with "unbiased" approaches such as genome-wide association (GWA) (Chap. 456), particularly in identifying single variants associated with high risk for certain forms of drug toxicity. GWA studies have identified variants in the HLA-B locus that are associated with high risk for severe skin rashes during treatment with the anticonvulsant carbamazepine and hepatotoxicity with flucloxacillin, an antibiotic never marketed in the United States. A GWA study of simvastatin-associated myopathy identified a single noncoding SNP in *SLCO1B1*, encoding OATP1B1, a drug transporter known to modulate simvastatin uptake into the liver, which accounts for 60% of myopathy risk. GWA approaches have also implicated interferon variants in antileukemic responses and in response to therapy in hepatitis C. African-American subjects are known to have higher dose requirements to achieve stable anticoagulation with warfarin, due in part to variation in *CYP2C9* and *VKORC1*, discussed below. In addition, a GWA study identified novel SNPs near *CYP2C9* that contribute to this effect in African Americans.

GENETIC VARIANTS AFFECTING PHARMACOKINETICS

Clinically important genetic variants have been described in multiple molecular pathways of drug disposition (**Table 64-2**). A distinct multimodal distribution of drug disposition (as shown in Fig. 64-1) argues for a predominant effect of variants in a single gene in the metabolism of that substrate. Individuals with two alleles (variants) encoding for nonfunctional protein make up one group, often termed *poor metabolizers* (PM phenotype). For most genes, many variants can produce such a loss of function, and assessing whether they are on the same or different alleles (i.e., the *diplotype*) can complicate the use of genotyping in clinical practice. Furthermore, some variants produce only partial loss of function, and the presence of more than one variant may be required to define a specific allele. Individuals with one functional allele, or multiple reduction of function alleles, make up a second (*intermediate metabolizers*) and may or may not be distinguishable from those with two functional alleles (normal metabolizers, often termed *extensive metabolizers*, EMs). *Ultra-rapid metabolizers* (UMs) with especially high enzymatic activity (occasionally due to gene duplication; Table 64-1 and Fig. 64-1) have also been described for some traits. Many drugs in widespread use can inhibit specific drug disposition pathways (**see Chap. 63, Table 63-1**), and so EM individuals receiving such inhibitors can respond like PM patients (*phenocopying*). Polymorphisms in genes encoding drug uptake or drug efflux transporters may be other contributors to variability in drug delivery to target sites and, hence, in drug effects.

TABLE 64-2

Genetic Variants and Drug Responses

Gene	Drugs	Effect of Genetic Variants[a]
Variants in Drug Metabolism Pathways		
CYP2C9	losartan	Decreased bioactivation and effects (PMs)
	warfarin	Decreased dose requirements; possible increased bleeding risk (PMs)
	phenytoin	Decreased dose requirement (PMs)
CYP2C19	omeprazole, voriconazole	Decreased effect in EMs
	celecoxib	Exaggerated effect in PMs
	clopidogrel	Decreased effect in PMs and IMs Consider alternate drug in PMs and alternate drug or dose increase in IMs Possible increased bleeding risk in carriers of gain of function variants
	citalopram, escitalopram	Choose alternate drug in UMs; reduce dose in PMs
CYP2D6	codeine, tamoxifen	Decreased bioactivation and drug effects in PMs
	codeine	Respiratory depression in UMs
	tricyclic antidepressants[b]	Increased adverse effects in PMs: Consider dose decrease Decreased therapeutic effects in UMs: Consider alternate drug
	metoprolol, carvedilol, timolol, propafenone	Increased beta blockade in PMs
	Fluvoxamine	Reduce dose or chose alternate drug in PMs
CYP3A5	tacrolimus, vincristine	Decreased drug concentrations and effect (CYP3A5*3 carriers)

Gene	Drugs	Effect of Genetic Variants[a]
Dihydropyrimidine dehydrogenase (*DPYD*)	capecitabine, 5-fluorouracil, tegafur	Possible severe toxicity (PMs)
NAT2	rifampin, isoniazid, pyrazinamide, hydralazine, procainamide	Increased risk of toxicity in PMs
Thiopurine S-methyltransferase (*TPMT*)	azathioprine, 6-mercaptopurine, thioguanine	PMs: Increased risk of bone marrow aplasia EMs: Possible decreased drug action at usual dosages
Uridine diphosphate glucuronosyltransferase (*UGT1A1*)	irinotecan	PM homozygotes: Increased risk of severe adverse effects (diarrhea, bone marrow aplasia)
	atazanavir	High risk of hyperbilirubinemia during treatment; can result in drug discontinuation
Pseudocholinesterase (*BCHE*)	succinylcholine and other muscle relaxants	Prolonged paralysis (autosomal recessive). Diagnosis established by genotyping or by measuring serum cholinesterase activity.
Variants in Other Genes		
Glucose 6-phosphate dehydrogenase (*G6PD*)	rasburicase, primaquine, chloroquine	Increased risk of hemolytic anemia in G6PD-deficient subjects
HLA-B*15:02	carbamazepine	Carriers (1 or 2 alleles) at increased risk of SJS/TEN (mainly Asian subjects)
HLA-B*31:01	carbamazepine	Carriers (1 or 2 alleles) at increased risk of SJS/TEN and milder skin toxicities (Caucasian and Asian subjects)
HLA-B*15:02	phenytoin	Carriers (1 or 2 alleles) at increased risk of SJS/TEN
HLA-B*57:01	abacavir	Carriers (1 or 2 alleles) at increased risk of SJS/TEN
HLA-B*58:01	allopurinol	Carriers (1 or 2 alleles) at increased risk of SJS/TEN
IFNL3 (IL28B)	interferon	Variable response in hepatitis C therapy

Gene	Drugs	Effect of Genetic Variants[a]
SLCO1B1	simvastatin	Encodes a drug uptake transporter; variant non-synonymous single nucleotide polymorphism increases myopathy risk especially at higher dosages
VKORC1	warfarin	Decreased dose requirements with variant promoter haplotype Increased dose requirement in individuals with non-synonymous loss of function variants
ITPA	ribavirin	Variants modulate risk for hemolytic anemia
RYR1	general anesthetics	Variants predispose to malignant hyperthermia
CFTR	ivacaftor, lumacaftor	Targeted therapies for cystic fibrosis indicated only in certain genotypes
Variants in Other Genomes (Infectious Agents, Tumors)		
Chemokine C-C motif receptor (CCR5)	maraviroc	Drug effective only in HIV strains with CCR5 detectible
C-KIT	imatinib	In gastrointestinal stromal tumors, drug indicated only with c-kit–positive cases
ALK (anaplastic lymphoma kinase)	Crizotinib	Indicated in patients with non-small cell lung cancer and ALK mutations
Her2/neu overexpression	trastuzumab, lapatinib	Drugs indicated only with tumor overexpression
K-ras mutation	panitumumab, cetuximab	Lack of efficacy with *KRAS* mutation
Philadelphia chromosome	dasatinib, nilotinib, imatinib	Decreased efficacy in Philadelphia chromosome–negative chronic myelogenous leukemia

[a]Drug effect in homozygotes unless otherwise specified. [b]Many tricyclic antidepressants and selective serotonin uptake inhibitors are metabolized by either CYP2D6, CYP2C19, or both, and some metabolites have pharmacologic activity. See https://www.pharmgkb.org/view/dosing-guidelines.do.

Note: EM, extensive metabolizer (normal enzymatic activity); IM, intermediate metabolizer (heterozygote for loss of function allele); PM, poor metabolizer (homozygote for reduced or loss of function allele); UM, ultra-rapid metabolizer (enzymatic activity much greater than normal, e.g., with gene duplication, Fig. 64-1). SJS/TEN: Stevens-Johnson Syndrome/Toxic Epidermal Necrolysis.

Further data at:

U.S. Food and Drug Administration:
http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm

Pharmacogenetics Research Network/Knowledge Base: *http://www.pharmgkb.org*

The Clinical Pharmacogenomics Implementation Consortium*: https://www.pharmgkb.org/page/cpic*

CYP3A

Members of the CYP3A family (*CYP3A4*, *CYP3A5*) metabolize the greatest number of drugs in therapeutic use. CYP3A4 activity is highly variable (up to an order of magnitude) among individuals, but non-synonymous coding region polymorphisms (those that change the encoded amino acid) are rare. Thus, the underlying mechanism likely reflects genetic variation in regulatory regions.

Most subjects of European or Asian origin carry a polymorphism that disrupts splicing in the closely related *CYP3A5* gene. As a result, these individuals display reduced CYP3A5 activity whereas CYP3A5 activity tends to be greater in subjects of African origin. Decreased efficacy of the antirejection agent tacrolimus in subjects of African origin has been attributed to more rapid CYP3A5-mediated elimination and a lower risk of vincristine-associated neuropathy has been reported in CYP3A5 “expressers.”

CYP2D6

CYP2D6 is second to CYP3A4 in the number of commonly used drugs that it metabolizes. CYP2D6 activity is polymorphically distributed, with 5–10% of European- and African-derived populations (but very few Asians) displaying the PM phenotype (Fig. 64-1). Dozens of loss-of-function variants in *CYP2D6* have been described; the PM phenotype arises in individuals with two such alleles. In addition, ultra-rapid metabolizers with multiple functional copies of *CYP2D6* have been identified especially in East Africa, the Middle East, and Oceania. PMs have slower elimination rates and lower clearance of substrate drugs; as a consequence (Fig. 64-1B), steady state concentrations are higher and the time taken to achieve steady state is longer than in EMs (**see Chap. 63**). Conversely, UMs display very low steady state parent drug concentrations and an abbreviated time to steady state.

Codeine is biotransformed by CYP2D6 to the potent active metabolite morphine, so its effects are blunted in PMs and exaggerated in UMs. Deaths due to respiratory depression in children given codeine after tonsillectomy have been attributed to the UM trait, and the U.S. Food and Drug Administration (FDA) has revised the package insert to include a prominent “black box” warning against its use in this setting. In the case of drugs with beta-blocking properties metabolized by CYP2D6, greater signs of beta blockade (e.g., bronchospasm, bradycardia) are seen in PM subjects than in EMs. This can be seen not only with orally administered beta blockers such as metoprolol and carvedilol, but also with ophthalmic timolol and with the sodium channel–blocking antiarrhythmic propafenone,

a CYP2D6 substrate with beta-blocking properties. Ultra-rapid metabolizers may require very high dosages of nortriptyline and other tricyclic antidepressants to achieve a therapeutic effect. Tamoxifen undergoes CYP2D6-mediated biotransformation to an active metabolite, so its efficacy may be in part related to this polymorphism. In addition, the widespread use of selective serotonin reuptake inhibitors (SSRIs) to treat tamoxifen-related hot flashes may also alter the drug's effects because many SSRIs, notably fluoxetine and paroxetine, are also CYP2D6 inhibitors.

CYP2C19

The PM phenotype for CYP2C19 is common (20%) among Asians and rarer (2–3%) in other populations. The impact of polymorphic CYP2C19-mediated metabolism has been demonstrated with the proton pump inhibitor omeprazole, where ulcer cure rates with "standard" dosages were much lower in EM patients (29%) than in PMs (100%). Thus, understanding the importance of this polymorphism would have been important in developing the drug, and knowing a patient's *CYP2C19* genotype should improve therapy. CYP2C19 is responsible for bioactivation of the antiplatelet drug clopidogrel, and several large retrospective studies have documented decreased efficacy (e.g., increased myocardial infarction after placement of coronary stents or increased stroke or transient ischemic attacks) among subjects with one or two reduction of function alleles. In addition, some studies suggest that omeprazole and possibly other proton pump inhibitors phenocopy this effect by inhibiting CYP2C19.

CYP2C9

There are common variants in *CYP2C9* that encode proteins with reduction or loss of catalytic function. These variant alleles are associated with increased rates of neurologic complications with phenytoin, hypoglycemia with glipizide, and reduced warfarin dose required to maintain stable anticoagulation. Rare patients homozygous for loss of function alleles may require very low warfarin dosages. Up to 50% of the variability in steady-state warfarin dose requirement is attributable to polymorphisms in *CYP2C9* and in the promoter of *VKORC1*, which encodes the warfarin target with lesser contributions by genes controlling vitamin K metabolism such as *CYP4F2*. The angiotensin-receptor blocker losartan is a prodrug that is bioactivated by CYP2C9; as a result, PMs and those receiving inhibitor drugs may display little response to therapy.

DPYD

Individuals homozygous for loss of function alleles in dihydropyrimidine dehydrogenase, encoded by DPYD, are at high risk for severe toxicity when exposed to the substrate anticancer drug 5-Fluorouracil (5-FU), as well as to capecitabine and tegafur, which are metabolized to 5-FU. Dose reductions have been recommended in intermediate metabolizers.

Transferase Variants

Thiopurine S-methyltransferase (TPMT) bioinactivates the antileukemic drug 6-mercaptopurine (6-MP) and 6-MP is itself an active metabolite of the immunosuppressive azathioprine. Homozygotes for alleles encoding inactive TPMT (1/300 individuals) predictably exhibit severe and potentially fatal pancytopenia on standard doses of

azathioprine or 6-MP. On the other hand, homozygotes for fully functional alleles may display less anti-inflammatory or antileukemic effect with standard doses of the drugs.

N-acetylation is catalyzed by hepatic *N*-acetyl transferase (NAT), which represents the activity of two genes, *NAT1* and *NAT2*. Both enzymes transfer an acetyl group from acetyl coenzyme A to the drug; polymorphisms in *NAT2* are thought to underlie individual differences in the rate at which drugs are acetylated and thus define "rapid acetylators" and "slow acetylators." Slow acetylators make up ~50% of European and African populations but are less common among East Asians. Slow acetylators have an increased incidence of the drug-induced lupus syndrome during procainamide and hydralazine therapy and of hepatitis with isoniazid.

Individuals homozygous for a common promoter polymorphism that reduces transcription of uridine diphosphate glucuronosyltransferase (*UGT1A1*) have benign hyperbilirubinemia (Gilbert's syndrome; **Chap. 330**). This variant has also been associated with diarrhea and increased bone marrow depression with the antineoplastic prodrug irinotecan, whose active metabolite is normally detoxified by UGT1A1-mediated glucuronidation. The antiretroviral atazanavir is a UGT1A1 inhibitor, and individuals with the Gilbert's variant develop higher bilirubin levels during treatment. While this is benign, the hyperbilirubinemia can complicate clinical care because it may raise the question of whether coexistent hepatic injury is present.

Transporter Variants

The risk for myotoxicity with simvastatin and possibly other statins appears increased with variants in *SLCO1B1*. Variants in *ABCB1*, encoding the drug efflux transporter P-glycoprotein, may increase digoxin toxicity. Variants in the uptake transporters *MATE1* and *MATE2* have been reported to modulate metformin's glucose-lowering activity.

GENETIC VARIANTS AFFECTING PHARMACODYNAMICS

A variant in the *VKORC1* promoter, especially common in Asian subjects (Table 64-1), reduces transcriptional activity and warfarin dose requirement. Multiple polymorphisms identified in the β_2-adrenergic receptor appear to be linked to specific phenotypes in asthma and congestive heart failure, diseases in which β_2-receptor function might be expected to determine prognosis. Polymorphisms in the β_2-receptor gene have also been associated with response to inhaled β_2-receptor agonists, while those in the β_1-adrenergic receptor gene have been associated with variability in heart rate slowing and blood pressure lowering. In addition, in heart failure, the arginine allele of the common β_1-adrenergic receptor gene polymorphism R389G has been associated with decreased mortality and decreased incidence of atrial fibrillation during treatment with the investigational beta blocker bucindolol.

Drugs may also interact with genetic pathways of disease to elicit or exacerbate symptoms of the underlying conditions. In the porphyrias, CYP inducers are thought to increase the activity of enzymes proximal to the deficient enzyme, exacerbating or triggering attacks **(Chap. 409)**. Deficiency of glucose-6-phosphate dehydrogenase (G6PD), most often in individuals of African, Mediterranean, or South Asian descent, increases the risk of hemolytic anemia in response to the antimalarial primaquine **(Chap. 96)** and the uric acid-lowering agent rasburicase, which does not cause hemolysis in patients with normal amounts of the enzyme. Patients with

mutations in *RYR1* encoding the skeletal muscle intracellular release calcium (also termed type 1 ryanodine receptor) are asymptomatic until exposed to certain general anesthetics, which can trigger the rare syndrome of malignant hyperthermia. Certain antiarrhythmics and other drugs can produce marked QT prolongation and torsades de pointes **(Chap. 241)**, and in some patients, this adverse effect represents unmasking of previously subclinical congenital long QT syndrome.

Immunologically Mediated Drug Reactions

The Stevens-Johnson syndrome (SJS) and toxic epidermal necrolysis (TEN) are potentially fatal skin reactions now increasingly recognized to be linked to specific HLA alleles (see Table 64-2). Some cases of hepatotoxicity have also been linked to variants in this region. The frequency of risk alleles often varies by ancestry (Table 64-1). The HLA risk alleles appear to be necessary but not sufficient to elicit these reactions. For example, HLA-B*57:01 is a risk allele for abacavir-related SJS/TEN and flucloxacillin-related hepatotoxicity. However, while 55% of abacavir-exposed subjects will develop reaction, only 1/10,000 subjects exposed to flucloxacillin develop hepatotoxicity. Thus, a third factor, the nature of which has not yet been established, seems necessary.

Tumor and Infectious Agent Genomes

The actions of drugs used to treat infectious or neoplastic disease may be modulated by variants in these nonhuman germline genomes. Genotyping tumors is a rapidly evolving approach to target therapies to underlying mechanisms and to avoid potentially toxic therapy in patients who would derive no benefit **(Chap. 67)**. Trastuzumab, which potentiates anthracycline-related cardiotoxicity, is ineffective in breast cancers that do not express the herceptin receptor. Imatinib targets a specific tyrosine kinase, BCR-Abl1, that is generated by the translocation that creates the Philadelphia chromosome typical of chronic myelogenous leukemia (CML). BCR-Abl1 is not only active but may be central to the pathogenesis of CML; use of imatinib and other BCR-Abl1 inhibitors has resulted in remarkable efficacy not only in CML but also in other BCR-Abl1-positive tumors such as gastrointestinal stromal tumors (**see Chap. 67**). Similarly, the anti-epidermal growth factor receptor (EGFR) antibodies cetuximab and panitumumab appear especially effective in colon cancers in which K-ras, a G protein in the EGFR pathway, is not mutated. Vemurafenib does not inhibit wild-type *BRAF* but is active against the V600E mutant form of the kinase. Crizotinib is highly effective in non-small cell lung cancers harboring anaplastic lymphoma kinase (ALK) mutations.

INCORPORATING PHARMACOGENETIC INFORMATION INTO CLINICAL PRACTICE

The discovery of common variant alleles with relatively large effects on drug response raises the prospect that these variants could be used to guide therapy. Desired outcomes could be better ways of choosing likely effective drugs and dosages, or avoiding drugs that are likely to produce severe adverse drug events or be ineffective in individual subjects. Indeed, the FDA now incorporates pharmacogenetic data into package inserts meant to guide prescribing. A decision to adopt pharmacogenetically guided dosing for a given drug depends on multiple factors. The most important are the magnitude and clinical importance of the genetic effect and the strength of evidence linking genetic variation to variable drug effects (e.g., anecdote versus post-hoc analysis of clinical trial data versus randomized clinical trial, RCT). The evidence can be strengthened if statistical arguments from clinical trial

data are complemented by an understanding of underlying physiologic mechanisms. Cost versus expected benefit may also be a factor.

Reactive versus Preemptive Approaches

Two approaches to pharmacogenetic implementation have been put in place at both “early adopter” institutions and are currently being evaluated. In the first, variant-specific assays are ordered at the time of drug prescription and delivered rapidly (often within an hour or two) and the results then used to guide therapy with that specific drug. The alternative to this “reactive” approach is a “preemptive” approach in which pharmacogenetic testing for large numbers of potential variants across many drugs is undertaken prior to prescription of any such drug. The data are then available in electronic health record (EHR) systems and coupled to real time clinical decision support (CDS). When a drug whose effects are known to be influenced by pharmacogenetic variants is prescribed, the EHR system looks up whether variants likely to affect response are present; if so, CDS will alert healthcare providers that an alternate drug or a different dose may be required.

Challenges

There are multiple challenges in putting in place either system. Assay validity and reproducibility have been issues in the past, but are less likely now. National consortia are now being put in place to develop standards for pharmacogenetic CDS. While common variants in genes such as those listed in Table 64-1 have been clearly associated with variable drug responses, the effect of rare variants, now readily discoverable by large scale sequencing, is unknown. The extent to which a dose adjustment might be recommended may vary depending on whether zero, one, or two variant alleles are present, and whether such variants are reduction of function, loss-of-function, or gain of function. The Clinical Pharmacogenetics Implementation Consortium (CPIC) has developed and published guidelines for multiple drug-gene pairs focusing on the question of what might be an appropriate drug dose adjustment given the availability of genetic data. CPIC does not, however, address the question of when or how such genetic testing should be undertaken.

Developing Evidence that Pharmacogenetic Testing Alters Drug Outcomes

A major issue is whether pharmacogenetic testing affects important drug response outcomes. When the evidence is compelling, alternate therapies are not available, and there are clear recommendations for dosage adjustment in subjects with variants, there is a strong argument for deploying genetic testing as a guide to prescribing; HLA-B*57:01 testing for abacavir is an example described below. In other situations, the arguments are less compelling: the magnitude of the genetic effect may be smaller, the consequences may be less serious, alternate therapies may be available, or the drug effect may be amenable to monitoring by other approaches.

One school argues that the physiology and pharmacology are known, and that RCTs are, therefore, unnecessary (and conceivably unethical). The analogy is sometimes drawn to well-recognized dose adjustment of renally excreted drugs in the presence of renal dysfunction. RCTs have not been conducted and the idea of such dose adjustment is well accepted in the medical community and recommended in FDA-approved drug labels. Others have argued that the effect of genetic variants is generally modest and variability in drug actions has many non-genetic sources, so genetic testing might provide marginal benefit at best.

Efforts to demonstrate the value of pharmacogenetic testing have met with mixed results. An RCT clearly showed that HLA-B*57:01 testing eliminates SJS/TEN due to abacavir. Similarly, regulatory authorities in some countries in Southeast Asia mandated HLA-B*15:02 testing prior to initiation of carbamazepine; however, in this case, an unfortunate outcome was that while the use of carbamazepine dropped, it was often substituted by phenytoin (another drug associated with SJS/TEN), so the incidence of the severe ADR was unchanged.

RCTs evaluating the effect of using pharmacogenetically guided therapy to optimize warfarin treatment have shown either no effect or a modest benefit of incorporating genetic information into prescribing the drug. These RCTs focused on time in therapeutic range in the first 4–12 weeks of treatment, and were not powered to examine outcomes such as recurrent thrombosis or bleeding. Retrospective analyses of bleeding cases vs non-bleeding controls in EHRs and administrative databases have suggested a role for CYP2C9*3 or the variants in V433M variant in CYP4F2 in mediating this risk.

While large retrospective analyses indicate that CYP2C19 loss of function variants decrease clopidogrel efficacy, RCTs are difficult to design: many argue that it is unethical to randomize individuals known to be homozygous for loss of function alleles, since administering clopidogrel is then tantamount to administering placebo. However, trials examining outcomes in only heterozygotes might require very large numbers of subjects.

New effective alternate therapies to warfarin and clopidogrel that appear to lack important pharmacogenetic variants have emerged. One approach to therapy, therefore, is to use pharmacogenetic testing to identify subjects in whom variants are absent and therefore standard doses of the conventional inexpensive drugs are likely to be effective and reserve alternate more expensive therapies for subjects likely to have variant responses to warfarin or clopidogrel. As price drops and as experience grows with newer agents, it is likely that clopidogrel and warfarin will be largely supplanted.

GENETICS AND DRUG DEVELOPMENT

Genetic tools are now being increasingly used to identify or validate new drug targets. Initial studies this field suggest that a new drug development program is more likely to succeed if evidence from human genetics supports the role of a possible drug target in disease pathogenesis and suggests that the risk of toxicity due to high risk pharmacokinetics or other mechanisms is small.

Finding Protective Alleles Can Identify Drug Targets

One example of using genetics to identify a new drug target started with the discovery that very rare gain of function variants in *PCSK9* are a rare cause of familial hypercholesterolemia. Subsequently, population studies showed that carriers of loss of function SNPs (2.5% of African Americans) had decreased low-density lipoprotein, decreased incidence of coronary artery disease, and no deleterious consequences in other organ systems. These data triggered the development of PCSK9 antagonists which were marketed less than 10 years after the initial population studies. Other targets implicated by similar population genetic studies include SLC30A8 for the prevention of type 2 diabetes and APOC3 for hypertriglyceridemia. In the latter examples, the identification of an apparently protective effect of rare loss-of-function alleles required very large datasets (>100,000) coupling DNA to longitudinal clinical information; long-term epidemiologic studies like the Framingham Heart Study or EHR systems are now being harnessed to address this opportunity.

Cancer

In cancer, tumor sequencing has identified new targets for drug development, often constitutively active kinases. A problem in this area has been the rapid emergence of drug resistance, often after extraordinary initial responses. For example, 40% of melanomas appear to be driven by the V600E mutant form of BRAF, and the specific inhibitor vemurafenib can produce clinically spectacular remission. However, durable responses are rare, and it is now apparent that combination therapy, often with inhibitors of the MEK pathway, can provide improved therapy. Another approach that is rapidly gaining wide use in cancer are drugs that reverse immune system inhibition (Chap. 69). In some patients the release of this "break" can provide durable remissions, whereas in others, severe adverse events, including colitis, pneumonitis, and myocarditis, have been reported. Understanding the mechanisms underlying variability to these therapies is a major emerging challenge in the field.

Using Multiple Data Types

The development of methods to understand associations across multiple large datasets is another approach that is being explored in drug development. For example, a GWA of risk of rheumatoid arthritis identified multiple risk loci and many encode proteins that are known targets for intervention in the disease. Interestingly, others encode proteins that are targets for drugs used in other conditions, such as certain cancers, raising the question of whether such drugs could be "repurposed" for rheumatoid arthritis. An extension of this approach is the broader issue of systems pharmacology, in which multiple sources of data are used to identify potential molecules or pathways that would be amenable to treatment, by new drugs or by existing agents, using analysis of genomic, transcriptomic, proteomic, and other large datasets. Similar approaches are being developed to predict toxicity expected from targeting specific genes or disease pathways.

SUMMARY

The science of pharmacogenomics has evolved from isolated examples of rare adverse drug actions to a more comprehensive view of the role of genetic variation in mediating the effects of most drugs. Current principles include:

Genetic variants with an important effect on drug actions can be common and their frequencies often vary by ancestry.

One common mechanism is modulation of drug concentrations.

No practitioner can be expected to remember all variants important for all drugs. Electronic data systems can now be accessed to describe this information. Ultimately, this information will be used by linking individual pharmacogenetic data to smart electronic health record systems.

Incorporating genetic approaches into drug development projects hold the promise of more rapid development of targeted, safe, and effective therapies.

FURTHER READING

Mallal S et al: HLA-B*5701 screening for hypersensitivity to abacavir. N Engl J Med 358:568, 2008. [PubMed: 18256392]

Nelson MR et al: The support of human genetic evidence for approved drug indications. Nat Genet 47:856, 2015. [PubMed: 26121088]

Relling MV, Evans WE: Pharmacogenomics in the clinic. Nature 526:343, 2015. [PubMed: 26469045]

Wang L et al: Genomics and drug response. N Engl J Med 364:1144, 2011. [PubMed: 21428770]

Weeke P, Roden DM: Applied pharmacogenomics in cardiovascular medicine. Annu Rev Med 65:81, 2014. [PubMed: 24111889]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 65: Approach to the Patient with Cancer

Dan L. Longo

INTRODUCTION

The application of current treatment techniques (surgery, radiation therapy, chemotherapy, and biologic therapy) results in the cure of nearly two of three patients diagnosed with cancer. Nevertheless, patients experience the diagnosis of cancer as one of the most traumatic and revolutionary events that has ever happened to them. Independent of prognosis, the diagnosis brings with it a change in a person's self-image and in his or her role in the home and workplace. The prognosis of a person who has just been found to have pancreatic cancer is the same as the prognosis of the person with aortic stenosis who develops the first symptoms of congestive heart failure (median survival, ~8 months). However, the patient with heart disease may remain functional and maintain a self-image as a fully intact person with just a malfunctioning part, a diseased organ ("a bum ticker"). By contrast, the patient with pancreatic cancer has a completely altered self-image and is viewed differently by family and anyone who knows the diagnosis. He or she is being attacked and invaded by a disease that could be anywhere in the body. Every ache or pain takes on desperate significance. Cancer is an exception to the coordinated interaction among cells and organs. In general, the cells of a multicellular organism are programmed for collaboration. Many diseases occur because the specialized cells fail to perform their assigned task. Cancer takes this malfunction one step further. Not only is there a failure of the cancer cell to maintain its specialized function, but it also strikes out on its own; the cancer cell competes to survive using natural mutability and natural selection to seek advantage over normal cells in a recapitulation of evolution. One consequence of the traitorous behavior of cancer cells is that the patient feels betrayed by his or her body. The cancer patient feels that he or she, and not just a body part, is diseased.

THE MAGNITUDE OF THE PROBLEM

No nationwide cancer registry exists; therefore, the incidence of cancer is estimated on the basis of the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) database, which tabulates cancer incidence and death figures from 13 sites, accounting for about 10% of the U.S. population, and from population data from the U.S. Census Bureau. In 2017, 1.688 million new cases of invasive cancer (836,150 men and 852,630 women) were diagnosed, and 600,920 persons (318,420 men and 282,500 women) died from cancer. The percent distribution of new cancer cases and cancer deaths by site for men and women is shown in Table 65-1. Cancer incidence has been declining by about 2% each year since 1992. Cancer is the cause of one in four deaths in the United States.

TABLE 65-1

Distribution of Cancer Incidence and Deaths for 2017

Male			Female		
Sites	**%**	**Number**	**Sites**	**%**	**Number**
Cancer Incidence					
Prostate	19	161,360	Breast	30	252,710
Lung	14	116,990	Lung	12	105,510
Colorectal	9	71,420	Colorectal	8	64,010
Bladder	7	60,490	Endometrial	6	61,380
Melanoma	6	52,170	Thyroid	5	42,470
Kidney	5	40,610	Melanoma	4	34,940
Lymphoma	5	40,080	Lymphoma	4	32,160
Leukemia	4	36,290	Leukemia	3	25,840
Oral Cavity	4	35,720	Pancreas	3	25,700
Liver	3	29,200	Kidney	3	23,380
All others	23	191,820	All others	22	184,530
All sites	100	836,150	All sites	100	852,630
Cancer Deaths					
Lung	27	84,590	Lung	25	71,280
Colorectal	9	27,150	Breast	14	40,610
Prostate	8	26,730	Colorectal	9	23,110
Pancreas	7	22,300	Pancreas	7	20,790
Liver	6	19,610	Ovary	5	14,080
Leukemia	4	14,300	Endometrial	4	10,920
Esophagus	4	12,720	Leukemia	4	10,200
Bladder	4	12,240	Liver	3	9310
Lymphoma	4	11,450	Lymphoma	3	8690
CNS	3	9620	CNS	3	7080

Male			Female		
Sites	**%**	**Number**	**Sites**	**%**	**Number**
All others	24	77,710	All others	24	66,880
All sites	100	318,420	All sites	100	282,500

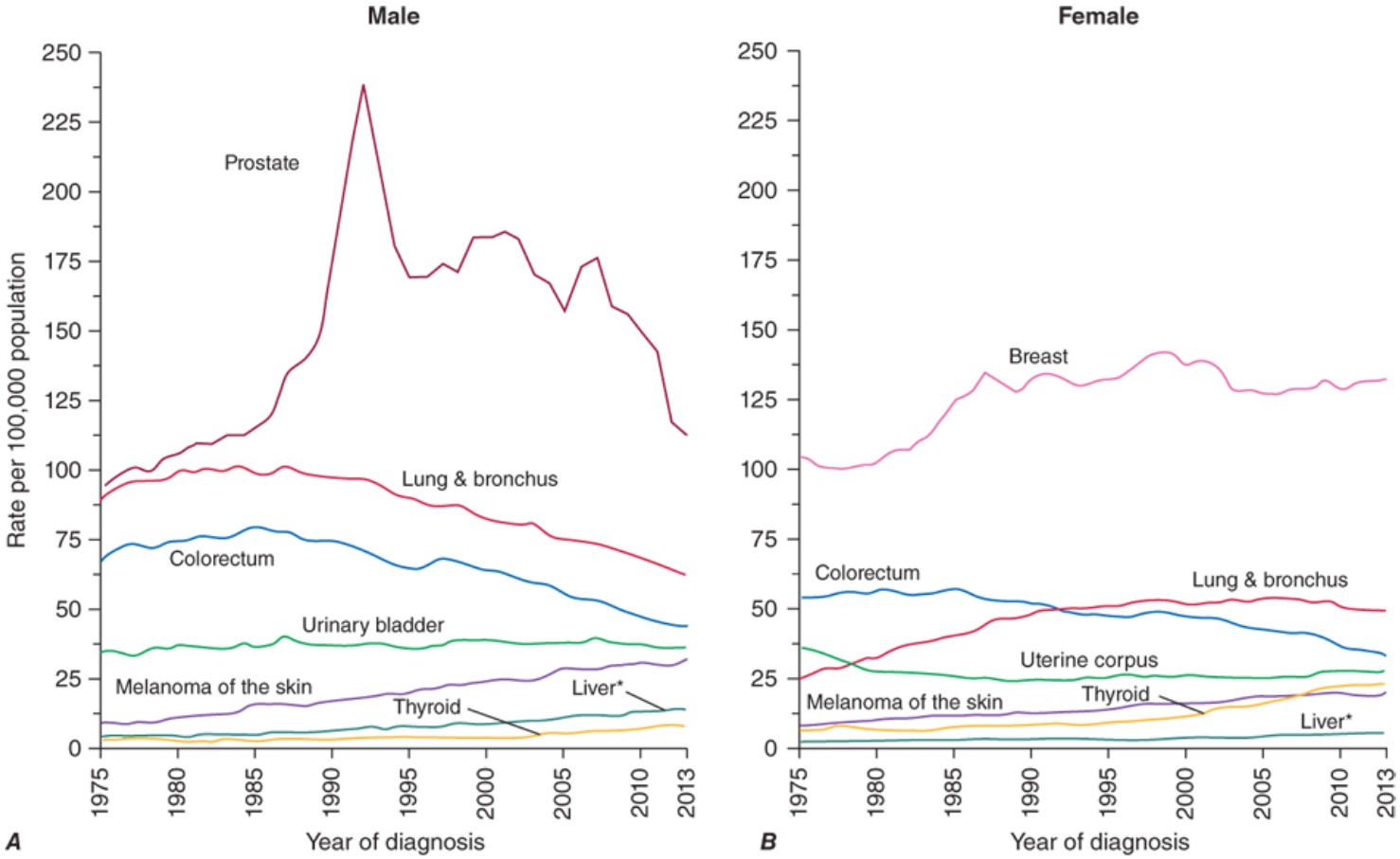
Source: From RL Siegel et al: Cancer statistics, 2017. CA Cancer J Clin 67:7, 2017.

The most significant risk factor for cancer overall is age; two-thirds of all cases were in those aged >65 years. Cancer incidence increases as the third, fourth, or fifth power of age in different sites. For the interval between birth and age 49 years, 1 in 29 men and 1 in 19 women will develop cancer; for the interval between ages 50 and 59 years, 1 in 15 men and 1 in 17 women will develop cancer; for the interval between ages 60 and 69 years, 1 in 6 men and 1 in 10 women will develop cancer; and for people aged ≥70, 1 in 3 men and 1 in 4 women will develop cancer. Overall, men have a 44% risk of developing cancer at some time during their lives; women have a 38% lifetime risk.

Cancer is the second leading cause of death behind heart disease. Deaths from heart disease have declined 45% in the United States since 1950 and continue to decline. Cancer has overtaken heart disease as the number one cause of death in persons aged <85 years. Incidence trends over time are shown in Fig. 65-1. After a 70-year period of increase, cancer deaths began to decline in 1990–1991 (Fig. 65-2). Between 1990 and 2010, cancer deaths decreased by 21% among men and 12.3% among women. The magnitude of the decline is illustrated in Fig. 65-3. The five leading causes of cancer deaths are shown for various populations in Table 65-2. The 5-year survival for white patients was 39% in 1960–1963 and 69% in 2003–2009. Cancers are more often deadly in blacks; the 5-year survival was 61% for the 2003–2009 interval; however, the racial differences are narrowing over time. Incidence and mortality vary among racial and ethnic groups (Table 65-3). The basis for these differences is unclear.

FIGURE 65-1

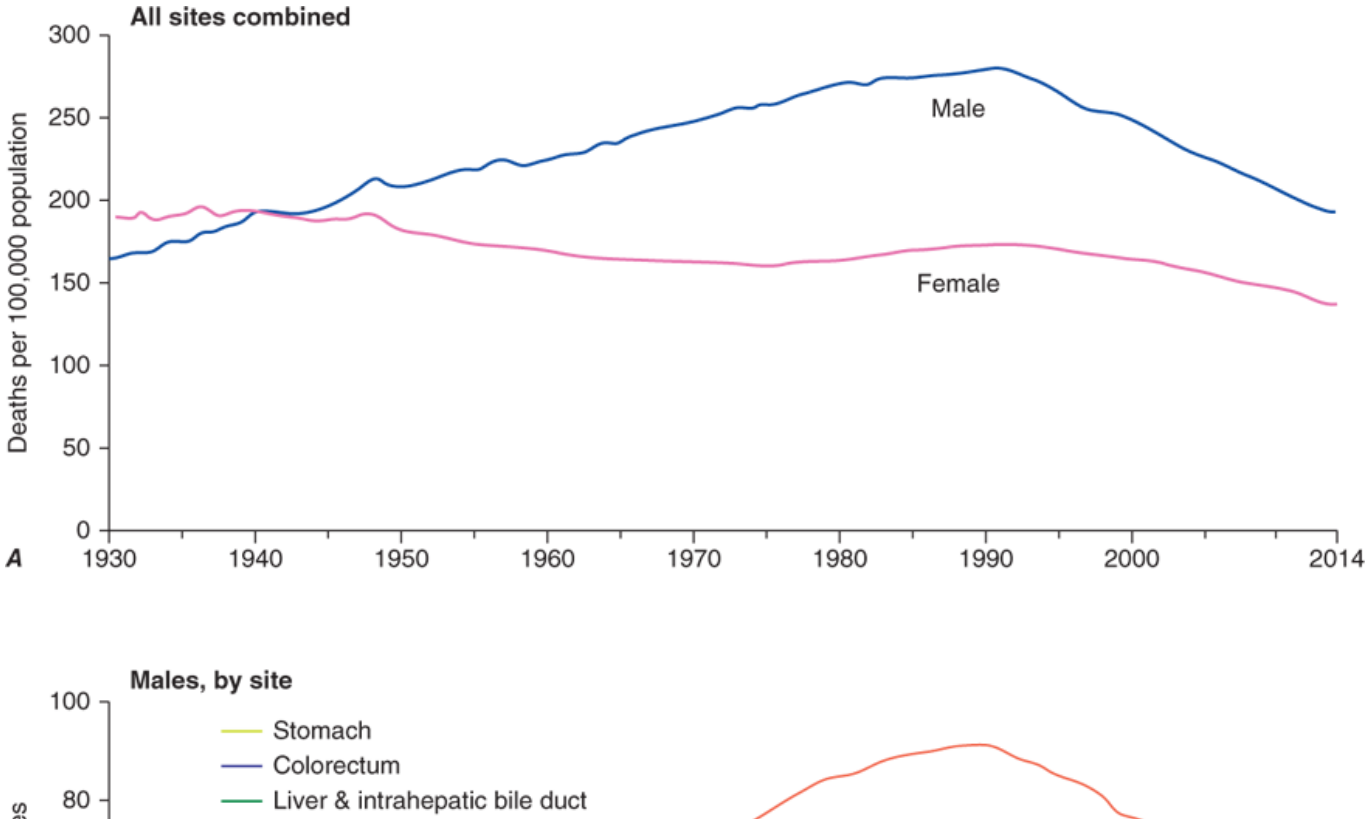
Incidence rates for particular types of cancer over the last 38 years in men (***A***) and women (***B***). *(From RL Siegel et al: CA Cancer J Clin 67:7, 2017.)*

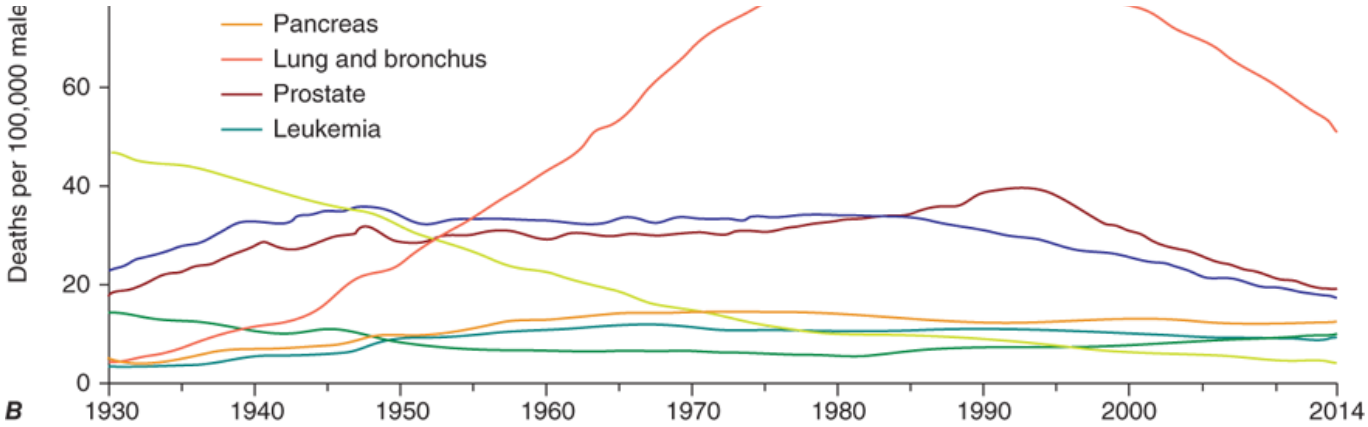


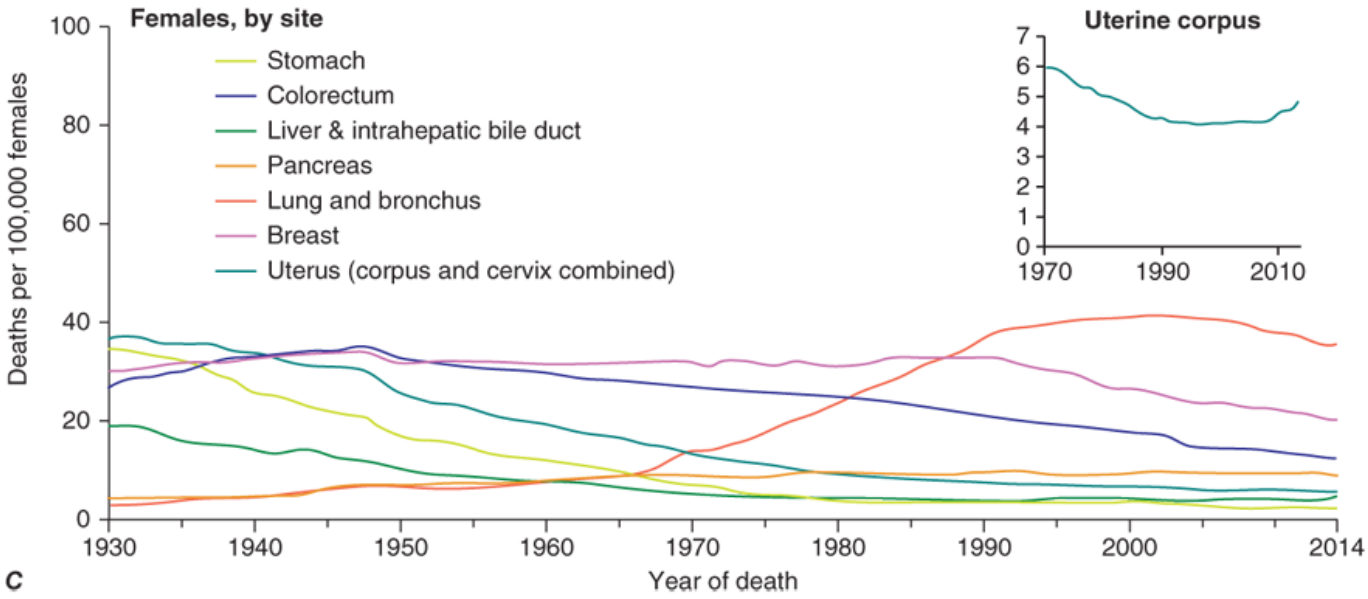
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 65-2

Eighty-five-year trend in cancer death rates for (***A***) women and (***B***) men by site in the United States, 1930–2014. Rates are per 100,000 age-adjusted to the 2000 U.S. standard population. All sites combined (***A***), individual sites in men (***B***) and individual sites in women (***C***) are shown. *(From RL Siegel et al: CA Cancer J Clin 67:7, 2017.)*



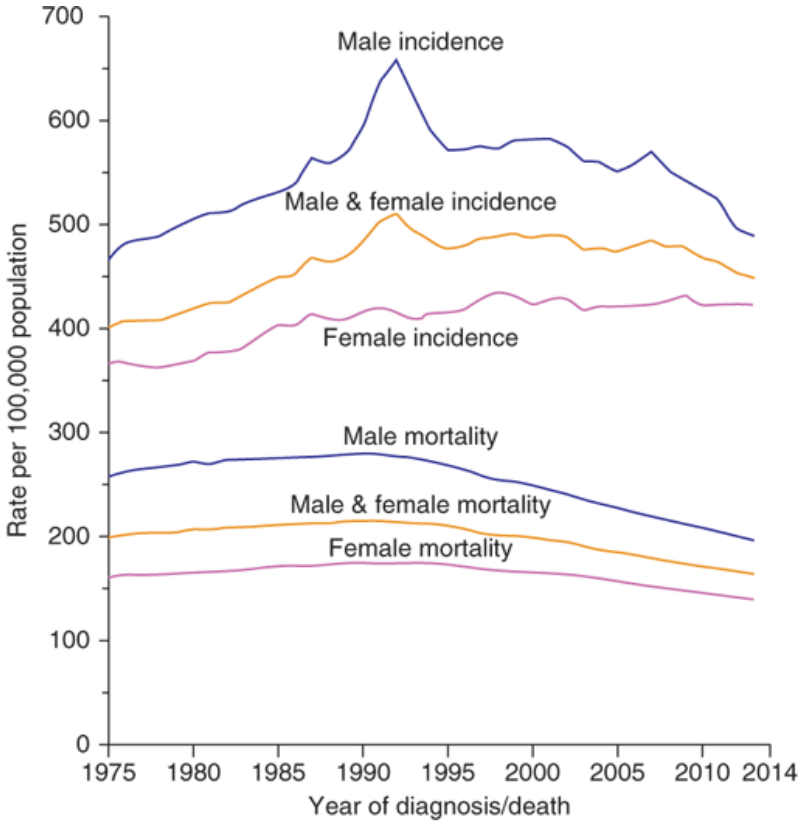




Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 65-3

Trends in cancer incidence and death rates for men and women from 1975 to 2014. *(From RL Siegel et al: CA Cancer J Clin 67:7, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 65-2

The Five Leading Primary Tumor Sites for Patients Dying of Cancer Based on Age and Sex in 2017

			Age, years				
Rank	**Sex**	**All Ages**	**Under 20**	**20–39**	**40–59**	**60–79**	**>80**
1	M	Lung	CNS	CNS	Lung	Lung	Lung
	F	Lung	CNS	Breast	Lung	Lung	Lung
2	M	Prostate	Leukemia	Leukemia	Colorectal	Colorectal	Prostate
	F	Breast	Leukemia	Cervix	Breast	Breast	Breast
3	M	Colorectal	Bone sarcoma	Colorectal	Liver	Prostate	Colorectal
	F	Colorectal	Bone sarcoma	Colorectal	Colorectal	Colorectal	Colorectal
4	M	Pancreas	Soft tissue sarcoma	Lymphoma	Pancreas	Pancreas	Bladder
	F	Pancreas	Soft tissue sarcoma	Leukemial	Ovary	Pancreas	Pancreas
5	M	Liver	Lymphoma	Lung	Esophagus	Liver	Pancreas
	F	Ovary	Lymphoma	CNS	Pancreas	Ovary	Leukemia

Abbreviations: CNS, central nervous system; F, female; M, male.

Source: From RL Siegel et al: Cancer statistics, 2017. CA Cancer J Clin 67:7, 2017.

TABLE 65-3

Cancer Incidence and Mortality in Racial and Ethnic Groups, United States, 2009–2013

Site	Sex	White	Black	Asian/Pacific Islander	American Indian[a]	Hispanic
Incidence per 100,000 Population						
All	M	519.3	577.3	310.2	426.7	498.1
	F	436.0	408.5	287.1	387.3	329.6
Breast		128.3	125.1	89.3	98.1	91.7
Colorectal	M	46.1	58.3	37.8	51.4	42.8
	F	35.2	42.7	27.8	41.2	29.8
Kidney	M	21.9	24.4	10.8	29.9	20.7
	F	11.3	13.0	4.8	17.6	11.9
Liver	M	9.7	16.9	20.4	18.5	19.4
	F	3.3	5.0	7.6	8.9	7.5
Lung	M	77.7	90.8	46.6	71.3	42.2
	F	58.2	51.0	28.3	56.2	25.6
Prostate		114.8	198.4	63.5	85.1	104.9
Cervix		7.0	9.8	6.1	9.7	9.9
Deaths per 100,000 Population						
All	M	204.0	253.4	122.7	183.6	142.5
	F	145.5	165.9	88.8	129.1	97.7
Breast		21.1	30.0	11.3	14.1	14.4
Colorectal	M	17.3	25.9	12.4	19.5	15.0
	F	12.3	16.9	8.8	14.0	9.2
Kidney	M	5.8	5.7	2.7	8.9	4.9
	F	2.5	2.5	1.1	4.2	2.3
Liver	M	8.0	13.3	14.3	14.9	13.1
	F	3.3	4.6	6.1	6.8	5.8
Lung	M	58.3	69.8	31.7	46.2	27.3

Site	Sex	White	Black	Asian/Pacific Islander	American Indian[a]	Hispanic
	F	39.8	35.5	18.0	30.8	13.4
Prostate		20.0	42.8	8.8	19.4	16.5
Cervix		2.3	3.9	1.7	2.8	2.6

[a]Based on Indian Health Service delivery areas.

Abbreviations: F, female; M, male.

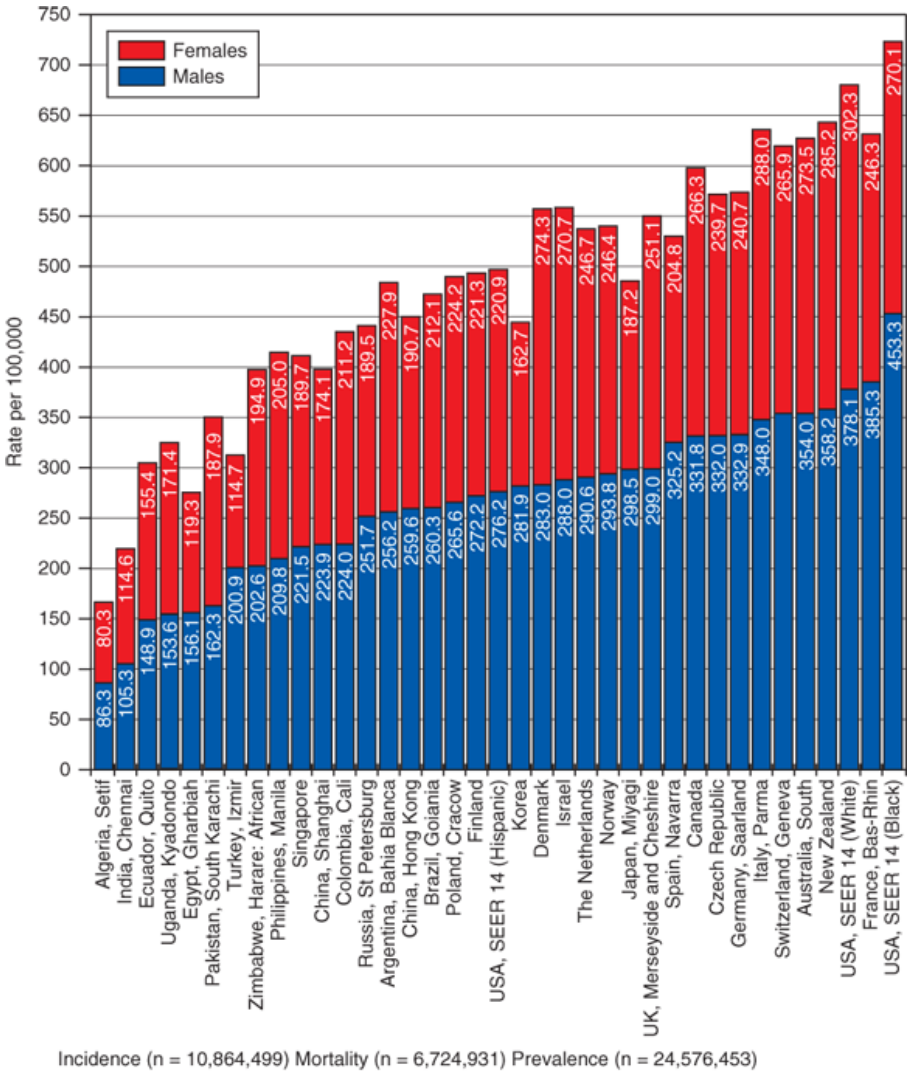
Source: From RL Siegel R et al: Cancer statistics, 2017. CA Cancer J Clin 67:7, 2017.

CANCER AROUND THE WORLD

In 2008, 12.7 million new cancer cases and 7.6 million cancer deaths were estimated worldwide, according to estimates of GLOBOCAN 2008, developed by the International Agency for Research on Cancer (IARC). When broken down by region of the world, ~45% of cases were in Asia, 26% in Europe, 14.5% in North America, 7.1% in Central/South America, 6% in Africa, and 1% in Australia/New Zealand **(Fig. 65-4)**. Lung cancer is the most common cancer and the most common cause of cancer death in the world. Its incidence is highly variable, affecting only 2 per 100,000 African women but as many as 61 per 100,000 North American men. Breast cancer is the second most common cancer worldwide; however, it ranks fifth as a cause of death behind lung, stomach, liver, and colorectal cancer. Among the eight most common forms of cancer, lung (2-fold), breast (3-fold), prostate (2.5-fold), and colorectal (3-fold) cancers are more common in more developed countries than in less developed countries. By contrast, liver (twofold), cervical (twofold), and esophageal (two- to threefold) cancers are more common in less developed countries. Stomach cancer incidence is similar in more and less developed countries but is much more common in Asia than North America or Africa. The most common cancers in Africa are cervical, breast, and liver cancers. It has been estimated that nine modifiable risk factors are responsible for more than one-third of cancers worldwide. These include smoking, alcohol consumption, obesity, physical inactivity, low fruit and vegetable consumption, unsafe sex, air pollution, indoor smoke from household fuels, and contaminated injections.

FIGURE 65-4

Worldwide overall annual cancer incidence, mortality, and 5-year prevalence for the period of 1993–2001. *(Adapted from A Jemal et al: Cancer Epidemiol Biomarkers Prev 19:1893, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATIENT MANAGEMENT

Important information is obtained from every portion of the routine history and physical examination. The duration of symptoms may reveal the chronicity of disease. The past medical history may alert the physician to the presence of underlying diseases that may affect the choice of therapy or the side effects of treatment. The social history may reveal occupational exposure to carcinogens or habits, such as smoking or alcohol consumption, that may influence the course of disease and its treatment. The family history may suggest an underlying familial cancer predisposition and point out the need to begin surveillance or other preventive therapy for unaffected siblings of the patient. The review of systems may suggest early symptoms of metastatic disease or a paraneoplastic syndrome.

DIAGNOSIS

The diagnosis of cancer relies most heavily on invasive tissue biopsy and should never be made without obtaining tissue; no noninvasive diagnostic test is sufficient to define a disease process as cancer. Although in rare clinical settings (e.g., thyroid nodules), fine-needle aspiration is an acceptable diagnostic procedure, the diagnosis generally depends on obtaining adequate tissue to permit careful evaluation of the histology of the tumor, its grade, and its invasiveness and to yield further molecular diagnostic information, such as the expression of cell-surface markers or intracellular proteins that typify a particular cancer, or the presence of a molecular marker, such as the t(8;14) translocation of Burkitt's lymphoma. Increasing evidence links the expression of certain genes with the prognosis and response to therapy (Chaps. 67 and 68).

Occasionally, a patient will present with a metastatic disease process that is defined as cancer on biopsy but has no apparent primary site of disease. Efforts should be made to define the primary site based on age, sex, sites of involvement, histology and tumor markers, and personal

and family history. Particular attention should be focused on ruling out the most treatable causes (Chap. 88).

Once the diagnosis of cancer is made, the management of the patient is best undertaken as a multidisciplinary collaboration among the primary care physician, medical oncologists, surgical oncologists, radiation oncologists, oncology nurse specialists, pharmacists, social workers, rehabilitation medicine specialists, and a number of other consulting professionals working closely with each other and with the patient and family.

DEFINING THE EXTENT OF DISEASE AND THE PROGNOSIS

The first priority in patient management after the diagnosis of cancer is established and shared with the patient is to determine the extent of disease. The curability of a tumor usually is inversely proportional to the tumor burden. Ideally, the tumor will be diagnosed before symptoms develop or as a consequence of screening efforts (Chap. 66). A very high proportion of such patients can be cured. However, most patients with cancer present with symptoms related to the cancer, caused either by mass effects of the tumor or by alterations associated with the production of cytokines or hormones by the tumor.

For most cancers, the extent of disease is evaluated by a variety of noninvasive and invasive diagnostic tests and procedures. This process is called *staging*. There are two types. *Clinical staging* is based on physical examination, radiographs, isotopic scans, computed tomography (CT) scans, and other imaging procedures; *pathologic staging* takes into account information obtained during a surgical procedure, which might include intraoperative palpation, resection of regional lymph nodes and/or tissue adjacent to the tumor, and inspection and biopsy of organs commonly involved in disease spread. Pathologic staging includes histologic examination of all tissues removed during the surgical procedure. Surgical procedures performed may include a simple lymph node biopsy or more extensive procedures such as thoracotomy, mediastinoscopy, or laparotomy. Surgical staging may occur in a separate procedure or may be done at the time of definitive surgical resection of the primary tumor.

Knowledge of the predilection of particular tumors for spreading to adjacent or distant organs helps direct the staging evaluation.

Information obtained from staging is used to define the extent of disease as localized, as exhibiting spread outside of the organ of origin to regional but not distant sites, or as metastatic to distant sites. The most widely used system of staging is the tumor, node, metastasis (TNM) system codified by the International Union Against Cancer and the American Joint Committee on Cancer. The TNM classification is an anatomically based system that categorizes the tumor on the basis of the size of the primary tumor lesion (T1–4, where a higher number indicates a tumor of larger size), the presence of nodal involvement (usually N0 and N1 for the absence and presence, respectively, of involved nodes, although some tumors have more elaborate systems of nodal grading), and the presence of metastatic disease (M0 and M1 for the absence and presence, respectively, of metastases). The various permutations of T, N, and M scores (sometimes including tumor histologic grade [G]) are then broken into stages, usually designated by the roman numerals I through IV. Tumor burden increases and curability decreases with increasing stage. Other anatomic staging systems are used for some tumors, e.g., the Dukes classification for colorectal cancers, the International Federation of Gynecologists and Obstetricians classification for gynecologic cancers, and the Ann Arbor classification for Hodgkin's disease.

Certain tumors cannot be grouped on the basis of anatomic considerations. For example, hematopoietic tumors such as leukemia, myeloma, and lymphoma are often disseminated at presentation and do not spread like solid tumors. For these tumors, other prognostic factors have been identified (Chaps. 101,102,103,104,105,106,107).

In addition to tumor burden, a second major determinant of treatment outcome is the physiologic reserve of the patient. Patients who are bedridden before developing cancer are likely to fare worse, stage for stage, than fully active patients. Physiologic reserve is a determinant of how a patient is likely to cope with the physiologic stresses imposed by the cancer and its treatment. This factor is difficult to assess directly. Instead, surrogate markers for physiologic reserve are used, such as the patient's age or Karnofsky performance status (Table 65-4) or Eastern Cooperative Oncology Group (ECOG) performance status (Table 65-5). Older patients and those with a Karnofsky performance status <70 or ECOG performance status ≥3 have a poor prognosis unless the poor performance is a reversible consequence of the tumor.

TABLE 65-4

Karnofsky Performance Index

Performance Status	Functional Capability of the Patient
100	Normal; no complaints; no evidence of disease
90	Able to carry on normal activity; minor signs or symptoms of disease
80	Normal activity with effort; some signs or symptoms of disease
70	Cares for self; unable to carry on normal activity or do active work
60	Requires occasional assistance but is able to care for most needs
50	Requires considerable assistance and frequent medical care
40	Disabled; requires special care and assistance
30	Severely disabled; hospitalization is indicated, although death is not imminent
20	Very sick; hospitalization is necessary; active supportive treatment is necessary
10	Moribund, fatal processes progressing rapidly
0	Dead

TABLE 65-5

The Eastern Cooperative Oncology Group (ECOG) Performance Scale

ECOG Grade 0: Fully active, able to carry on all predisease performance without restriction
ECOG Grade 1: Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light housework, office work
ECOG Grade 2: Ambulatory and capable of all self-care but unable to carry out any work activities. Up and about >50% of waking hours
ECOG Grade 3: Capable of only limited self-care, confined to bed or chair >50% of waking hours
ECOG Grade 4: Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair
ECOG Grade 5: Dead

Source: From MM Oken et al: Am J Clin Oncol 5:649, 1982.

Increasingly, biologic features of the tumor are being related to prognosis. The expression of particular oncogenes, drug-resistance genes, apoptosis-related genes, and genes involved in metastasis is being found to influence response to therapy and prognosis. The presence of selected cytogenetic abnormalities may influence survival. Tumors with higher growth fractions, as assessed by expression of proliferation-related markers such as proliferating cell nuclear antigen, behave more aggressively than tumors with lower growth fractions. Information obtained from studying the tumor itself will increasingly be used to influence treatment decisions. Host genes involved in drug metabolism can influence the safety and efficacy of particular treatments.

Enormous heterogeneity has been noted by studying tumors; we have learned that morphology is not capable of discerning certain distinct subsets of patients whose tumors have different sets of abnormalities. Tumors that look the same by light microscopy can be very different. Similarly, tumors that look quite different from one another histologically can share genetic lesions that predict responses to treatments. Furthermore, tumor cells vary enormously within a single patient even though the cells share a common origin.

MAKING A TREATMENT PLAN

From information on the extent of disease and the prognosis and in conjunction with the patient's wishes, it is determined whether the treatment approach should be curative or palliative in intent. Cooperation among the various professionals involved in cancer treatment is of the utmost importance in treatment planning. For some cancers, chemotherapy or chemotherapy plus radiation therapy delivered before the use of definitive surgical treatment (so-called neoadjuvant therapy) may improve the outcome, as seems to be the case for locally advanced breast cancer and head and neck cancers. In certain settings in which combined-modality therapy is intended, coordination among the medical oncologist, radiation oncologist, and surgeon is crucial to achieving optimal results. Sometimes the chemotherapy and radiation therapy need to be delivered sequentially, and other times concurrently. Surgical procedures may precede or follow other treatment approaches. It is best for the treatment plan either to follow a standard protocol precisely or else to be part of an ongoing clinical research protocol evaluating new treatments. Ad hoc modifications of standard protocols are likely to compromise treatment results.

The choice of treatment approaches was formerly dominated by the local culture in both the university and the practice settings. However, it is now possible to gain access electronically to standard treatment protocols and to every approved clinical research study in North America through a personal computer interface with the Internet.[1]

The skilled physician also has much to offer the patient for whom curative therapy is no longer an option. Often a combination of guilt and frustration over the inability to cure the patient and the pressure of a busy schedule greatly limit the time a physician spends with a patient who is receiving only palliative care. Resist these forces. In addition to the medicines administered to alleviate symptoms (see below), it is important to remember the comfort that is provided by holding the patient's hand, continuing regular examinations, and taking time to talk.

[1]The National Cancer Institute maintains a database called PDQ (Physician Data Query) that is accessible on the Internet under the name CancerNet at *www.cancer.gov/cancertopics/pdq/cancerdatabase*. Information can be obtained through a facsimile machine using CancerFax by dialing 301-402-5874. Patient information is also provided by the National Cancer Institute in at least three formats: on the Internet via CancerNet at *www.cancer.gov*, through the CancerFax number listed above, or by calling 1-800-4-CANCER. The quality control for the information provided through these services is rigorous.

MANAGEMENT OF DISEASE AND TREATMENT COMPLICATIONS

Because cancer therapies are toxic **(Chap. 69)**, patient management involves addressing complications of both the disease and its treatment as well as the complex psychosocial problems associated with cancer. In the short term during a course of curative therapy, the patient's functional status may decline. Treatment-induced toxicity is less acceptable if the goal of therapy is palliation. The most common side effects of treatment are nausea and vomiting (see below), febrile neutropenia **(Chap. 70)**, and myelosuppression **(Chap. 69)**. Tools are now available to minimize the acute toxicity of cancer treatment.

New symptoms developing in the course of cancer treatment should always be assumed to be reversible until proven otherwise. The fatalistic attribution of anorexia, weight loss, and jaundice to recurrent or progressive tumor could result in a patient dying from a reversible intercurrent cholecystitis. Intestinal obstruction may be due to reversible adhesions rather than progressive tumor. Systemic infections, sometimes with unusual pathogens, may be a consequence of the immunosuppression associated with cancer therapy. Some drugs used to treat cancer or its complications (e.g., nausea) may produce central nervous system symptoms that look like metastatic disease or may mimic paraneoplastic syndromes such as the syndrome of inappropriate antidiuretic hormone. A definitive diagnosis should be pursued and may even require a repeat biopsy.

A critical component of cancer management is assessing the response to treatment. In addition to a careful physical examination in which all sites of disease are physically measured and recorded in a flow chart by date, response assessment usually requires periodic repeating of imaging tests that were abnormal at the time of staging. If imaging tests have become normal, repeat biopsy of previously involved tissue is performed to document complete response by pathologic criteria. Biopsies are not usually required if there is macroscopic residual disease. A *complete response* is defined as disappearance of all evidence of disease, and a *partial response* as >50% reduction in the sum of the products of the perpendicular diameters of all measurable lesions. The determination of partial response may also be based on a 30%

decrease in the sums of the longest diameters of lesions (Response Evaluation Criteria in Solid Tumors [RECIST]). *Progressive disease* is defined as the appearance of any new lesion or an increase of >25% in the sum of the products of the perpendicular diameters of all measurable lesions (or an increase of 20% in the sums of the longest diameters by RECIST). Tumor shrinkage or growth that does not meet any of these criteria is considered *stable disease*. Some sites of involvement (e.g., bone) or patterns of involvement (e.g., lymphangitic lung or diffuse pulmonary infiltrates) are considered unmeasurable. No response is complete without biopsy documentation of their resolution, but partial responses may exclude their assessment unless clear objective progression has occurred.

For some hematologic neoplasms, flow cytometric and genetic assays may determine the presence of residual tumor cells that escape microscopic detection. In general, these techniques can reliably detect as few as 1 tumor cell among 10,000 cells. If such tests do not detect tumor cells, the patient is said to have minimal residual disease negativity, a finding generally associated with more durable remissions. Accumulating data are defining interventions in patients with minimal residual disease positivity that can extend remission duration and survival.

Tumor markers may be useful in patient management in certain tumors. Response to therapy may be difficult to gauge with certainty. However, some tumors produce or elicit the production of markers that can be measured in the serum or urine, and in a particular patient, rising and falling levels of the marker are usually associated with increasing or decreasing tumor burden, respectively. Some clinically useful tumor markers are shown in **Table 65-6**. Tumor markers are not in themselves specific enough to permit a diagnosis of malignancy to be made, but once a malignancy has been diagnosed and shown to be associated with elevated levels of a tumor marker, the marker can be used to assess response to treatment.

TABLE 65-6

Tumor Markers

Tumor Markers	Cancer	Nonneoplastic Conditions
Hormones		
Human chorionic gonadotropin	Gestational trophoblastic disease, gonadal germ cell tumor	Pregnancy
Calcitonin	Medullary cancer of the thyroid	
Catecholamines	Pheochromocytoma	
Oncofetal Antigens		
α Fetoprotein	Hepatocellular carcinoma, gonadal germ cell tumor	Cirrhosis, hepatitis
Carcinoembryonic antigen	Adenocarcinomas of the colon, pancreas, lung, breast, ovary	Pancreatitis, hepatitis, inflammatory bowel disease, smoking
Enzymes		
Prostatic acid phosphatase	Prostate cancer	Prostatitis, prostatic hypertrophy
Neuron-specific enolase	Small-cell cancer of the lung, neuroblastoma	
Lactate dehydrogenase	Lymphoma, Ewing's sarcoma	Hepatitis, hemolytic anemia, many others
Tumor-Associated Proteins		
Prostate-specific antigen	Prostate cancer	Prostatitis, prostatic hypertrophy
Monoclonal immunoglobulin	Myeloma	Infection, MGUS
CA-125	Ovarian cancer, some lymphomas	Menstruation, peritonitis, pregnancy
CA 19-9	Colon, pancreatic, breast cancer	Pancreatitis, ulcerative colitis
CD30	Hodgkin's disease, anaplastic large-cell lymphoma	—
CD25	Hairy cell leukemia, adult T-cell leukemia/lymphoma	—

Abbreviation: MGUS, monoclonal gammopathy of uncertain significance.

The recognition and treatment of depression are important components of management. The incidence of depression in cancer patients is ~25% overall and may be greater in patients with greater debility. This diagnosis is likely in a patient with a depressed mood (dysphoria) and/or a loss of interest in pleasure (anhedonia) for at least 2 weeks. In addition, three or more of the following symptoms are usually present: appetite change, sleep problems, psychomotor retardation or agitation, fatigue, feelings of guilt or worthlessness, inability to concentrate, and suicidal ideation. Patients with these symptoms should receive therapy. Medical therapy with a serotonin reuptake inhibitor such as fluoxetine (10–20 mg/d), sertraline (50–150 mg/d), or paroxetine (10–20 mg/d) or a tricyclic antidepressant such as amitriptyline (50–100 mg/d) or desipramine (75–150 mg/d) should be tried, allowing 4–6 weeks for response. Effective therapy should be continued at least 6

months after resolution of symptoms. If therapy is unsuccessful, other classes of antidepressants may be used. In addition to medication, psychosocial interventions such as support groups, psychotherapy, and guided imagery may be of benefit.

Many patients opt for unproven or unsound approaches to treatment when it appears that conventional medicine is unlikely to be curative. Those seeking such alternatives are often well educated and may be early in the course of their disease. Unsound approaches are usually hawked on the basis of unsubstantiated anecdotes and not only cannot help the patient but may be harmful. Physicians should strive to keep communications open and nonjudgmental, so that patients are more likely to discuss with the physician what they are actually doing. The appearance of unexpected toxicity may be an indication that a supplemental therapy is being taken.[2]

[2]Information about unsound methods may be obtained from the National Council Against Health Fraud, Box 1276, Loma Linda, CA 92354, or from the Center for Medical Consumers and Health Care Information, 237 Thompson Street, New York, NY 10012.

LONG-TERM FOLLOW-UP/LATE COMPLICATIONS

At the completion of treatment, sites originally involved with tumor are reassessed, usually by radiography or imaging techniques, and any persistent abnormality is biopsied. If disease persists, the multidisciplinary team discusses a new salvage treatment plan. If the patient has been rendered disease-free by the original treatment, the patient is followed regularly for disease recurrence. The optimal guidelines for follow-up care are not known. For many years, a routine practice has been to follow the patient monthly for 6–12 months, then every other month for a year, every 3 months for a year, every 4 months for a year, every 6 months for a year, and then annually. At each visit, a battery of laboratory and radiographic and imaging tests were obtained on the assumption that it is best to detect recurrent disease before it becomes symptomatic. However, where follow-up procedures have been examined, this assumption has been found to be untrue. Studies of breast cancer, melanoma, lung cancer, colon cancer, and lymphoma have all failed to support the notion that asymptomatic relapses are more readily cured by salvage therapy than symptomatic relapses. In view of the enormous cost of a full battery of diagnostic tests and their manifest lack of impact on survival, new guidelines are emerging for less frequent follow-up visits, during which the history and physical examination are the major investigations performed.

As time passes, the likelihood of recurrence of the primary cancer diminishes. For many types of cancer, survival for 5 years without recurrence is tantamount to cure. However, important medical problems can occur in patients treated for cancer and must be examined (Chap. 91). Some problems emerge as a consequence of the disease and some as a consequence of the treatment. An understanding of these disease- and treatment-related problems may help in their detection and management.

Despite these concerns, most patients who are cured of cancer return to normal lives.

SUPPORTIVE CARE

In many ways, the success of cancer therapy depends on the success of the supportive care. Failure to control the symptoms of cancer and its treatment may lead patients to abandon curative therapy. Of equal importance, supportive care is a major determinant of quality of life. Even when life cannot be prolonged, the physician must strive to preserve its quality. Quality-of-life measurements have become common endpoints of clinical research studies. Furthermore, palliative care has been shown to be cost-effective when approached in an organized fashion. A credo for oncology could be to cure sometimes, to extend life often, and to comfort always.

Pain

Pain occurs with variable frequency in the cancer patient: 25–50% of patients present with pain at diagnosis, 33% have pain associated with treatment, and 75% have pain with progressive disease. The pain may have several causes. In ~70% of cases, pain is caused by the tumor itself—by invasion of bone, nerves, blood vessels, or mucous membranes or obstruction of a hollow viscus or duct. In ~20% of cases, pain is related to a surgical or invasive medical procedure, to radiation injury (mucositis, enteritis, or plexus, or spinal cord injury), or to chemotherapy injury (mucositis, peripheral neuropathy, phlebitis, steroid-induced aseptic necrosis of the femoral head). In 10% of cases, pain is unrelated to cancer or its treatment.

Assessment of pain requires the methodical investigation of the history of the pain, its location, character, temporal features, provocative and palliative factors, and intensity (Chap. 10); a review of the oncologic history and past medical history as well as personal and social history; and a thorough physical examination. The patient should be given a 10-division visual analogue scale on which to indicate the severity of the pain. The clinical condition is often dynamic, making it necessary to reassess the patient frequently. Pain therapy should not be withheld while the cause of pain is being sought.

A variety of tools are available with which to address cancer pain. About 85% of patients will have pain relief from pharmacologic intervention. However, other modalities, including antitumor therapy (such as surgical relief of obstruction, radiation therapy, and strontium-89 or samarium-153 treatment for bone pain), neurostimulatory techniques, regional analgesia, or neuroablative procedures, are effective in an additional 12% or so. Thus, very few patients will have inadequate pain relief if appropriate measures are taken. **A specific approach to pain relief is detailed in Chap. 9.**

Nausea

Emesis in the cancer patient is usually caused by chemotherapy **(Chap. 69)**. Its severity can be predicted from the drugs used to treat the cancer. Three forms of emesis are recognized on the basis of their timing with regard to the noxious insult. *Acute emesis*, the most common variety, occurs within 24 h of treatment. *Delayed emesis* occurs 1–7 days after treatment; it is rare, but, when present, usually follows cisplatin administration. *Anticipatory emesis* occurs before administration of chemotherapy and represents a conditioned response to visual and olfactory stimuli previously associated with chemotherapy delivery.

Acute emesis is the best understood form. Stimuli that activate signals in the chemoreceptor trigger zone in the medulla, the cerebral cortex, and peripherally in the intestinal tract lead to stimulation of the vomiting center in the medulla, the motor center responsible for coordinating the secretory and muscle contraction activity that leads to emesis. Diverse receptor types participate in the process, including dopamine, serotonin, histamine, opioid, and acetylcholine receptors. The serotonin receptor antagonists ondansetron and granisetron are effective drugs against highly emetogenic agents, as are neurokinin receptor antagonists like aprepitant and fosaprepitant **(see Chap. 69)**.

As with the analgesia ladder, emesis therapy should be tailored to the situation. For mildly and moderately emetogenic agents, prochlorperazine, 5–10 mg PO or 25 mg PR, is effective. Its efficacy may be enhanced by administering the drug before the chemotherapy is delivered. Dexamethasone, 10–20 mg IV, is also effective and may enhance the efficacy of prochlorperazine. For highly emetogenic agents such as cisplatin, mechlorethamine, dacarbazine, and streptozocin, combinations of agents work best and administration should begin 6–24 h before treatment. Ondansetron, 8 mg PO every 6 h the day before therapy and IV on the day of therapy, plus dexamethasone, 20 mg IV before treatment, is an effective regimen. Addition of oral aprepitant (a substance P/neurokinin 1 receptor antagonist) to this regimen (125 mg on day 1, 80 mg on days 2 and 3) further decreases the risk of both acute and delayed vomiting. Like pain, emesis is easier to prevent than to alleviate.

Delayed emesis may be related to bowel inflammation from the therapy and can be controlled with oral dexamethasone and oral metoclopramide, a dopamine receptor antagonist that also blocks serotonin receptors at high dosages. The best strategy for preventing anticipatory emesis is to control emesis in the early cycles of therapy to prevent the conditioning from taking place. If this is unsuccessful, prophylactic antiemetics the day before treatment may help. Experimental studies are evaluating behavior modification.

Effusions

Fluid may accumulate abnormally in the pleural cavity, pericardium, or peritoneum. Asymptomatic malignant effusions may not require treatment. Symptomatic effusions occurring in tumors responsive to systemic therapy usually do not require local treatment but respond to the treatment for the underlying tumor. Symptomatic effusions occurring in tumors unresponsive to systemic therapy may require local treatment in patients with a life expectancy of at least 6 months.

Pleural effusions due to tumors may or may not contain malignant cells. Lung cancer, breast cancer, and lymphomas account for ~75% of malignant pleural effusions. Their exudative nature is usually gauged by an effusion/serum protein ratio of ≥0.5 or an effusion/serum lactate dehydrogenase ratio of ≥0.6. When the condition is symptomatic, thoracentesis is usually performed first. In most cases, symptomatic improvement occurs for <1 month. Chest tube drainage is required if symptoms recur within 2 weeks. Fluid is aspirated until the flow rate is <100 mL in 24 h. Then either 60 units of bleomycin or 1 g of doxycycline is infused into the chest tube in 50 mL of 5% dextrose in water; the tube is clamped; the patient is rotated on four sides, spending 15 min in each position; and, after 1–2 h, the tube is again attached to suction for another 24 h. The tube is then disconnected from suction and allowed to drain by gravity. If <100 mL drains over the next 24 h, the chest tube is pulled, and a radiograph is taken 24 h later. If the chest tube continues to drain fluid at an unacceptably high rate, sclerosis can be repeated. Bleomycin may be somewhat more effective than doxycycline but is very expensive. Doxycycline is usually the drug of first choice. If neither doxycycline nor bleomycin is effective, talc can be used.

Symptomatic pericardial effusions are usually treated by creating a pericardial window or by stripping the pericardium. If the patient's condition does not permit a surgical procedure, sclerosis can be attempted with doxycycline and/or bleomycin.

Malignant ascites is usually treated with repeated paracentesis of small volumes of fluid. If the underlying malignancy is unresponsive to systemic therapy, peritoneovenous shunts may be inserted. Despite the fear of disseminating tumor cells into the circulation, widespread metastases are an unusual complication. The major complications are occlusion, leakage, and fluid overload. Patients with severe liver disease may develop disseminated intravascular coagulation.

Nutrition

Cancer and its treatment may lead to a decrease in nutrient intake of sufficient magnitude to cause weight loss and alteration of intermediary metabolism. The prevalence of this problem is difficult to estimate because of variations in the definition of cancer cachexia, but most patients with advanced cancer experience weight loss and decreased appetite. A variety of both tumor-derived factors (e.g., bombesin, adrenocorticotropic hormone) and host-derived factors (e.g., tumor necrosis factor, interleukins 1 and 6, growth hormone) contribute to the altered metabolism, and a vicious cycle is established in which protein catabolism, glucose intolerance, and lipolysis cannot be reversed by the provision of calories.

It remains controversial how to assess nutritional status and when and how to intervene. Efforts to make the assessment objective have included the use of a prognostic nutritional index based on albumin levels, triceps skinfold thickness, transferrin levels, and delayed-type hypersensitivity skin testing. However, a simpler approach has been to define the threshold for nutritional intervention as <10% unexplained body weight loss, serum transferrin level <1500 mg/L (150 mg/dL), and serum albumin <34 g/L (3.4 g/dL).

The decision is important, because it appears that cancer therapy is substantially more toxic and less effective in the face of malnutrition. Nevertheless, it remains unclear whether nutritional intervention can alter the natural history. Unless some pathology is affecting the absorptive function of the gastrointestinal tract, enteral nutrition provided orally or by tube feeding is preferred over parenteral supplementation. However, the risks associated with the tube may outweigh the benefits. Megestrol acetate, a progestational agent, has been advocated as a pharmacologic intervention to improve nutritional status. Research in this area may provide more tools in the future as cytokine-mediated mechanisms are further elucidated.

Psychosocial Support

The psychosocial needs of patients vary with their situation. Patients undergoing treatment experience fear, anxiety, and depression. Self-image is often seriously compromised by deforming surgery and loss of hair. Women who receive cosmetic advice that enables them to look better also feel better. Loss of control over how one spends time can contribute to the sense of vulnerability. Juggling the demands of work and family with the demands of treatment may create enormous stresses. Sexual dysfunction is highly prevalent and needs to be discussed openly with the patient. An empathetic health care team is sensitive to the individual patient's needs and permits negotiation where such flexibility will not adversely affect the course of treatment.

Cancer survivors have other sets of difficulties. Patients may have fears associated with the termination of a treatment they associate with their continued survival. Adjustments are required to physical losses and handicaps, real and perceived. Patients may be preoccupied with minor physical problems. They perceive a decline in their job mobility and view themselves as less desirable workers. They may be victims of job and/or insurance discrimination. Patients may experience difficulty reentering their normal past life. They may feel guilty for having survived and may carry a sense of vulnerability to colds and other illnesses. Perhaps the most pervasive and threatening concern is the ever-present fear of relapse (the Damocles syndrome).

Patients in whom therapy has been unsuccessful have other problems related to the end of life.

Death and Dying

The most common causes of death in patients with cancer are infection (leading to circulatory failure), respiratory failure, hepatic failure, and renal failure. Intestinal blockage may lead to inanition and starvation. Central nervous system disease may lead to seizures, coma, and central hypoventilation. About 70% of patients develop dyspnea preterminally. However, many months usually pass between the diagnosis of cancer and the occurrence of these complications, and during this period, the patient is severely affected by the possibility of death. The path of unsuccessful cancer treatment usually occurs in three phases. First, there is optimism at the hope of cure; when the tumor recurs, there is the acknowledgment of an incurable disease, and the goal of palliative therapy is embraced in the hope of being able to live with disease; finally, at the disclosure of imminent death, another adjustment in outlook takes place. The patient imagines the worst in preparation for the end of life and may go through stages of adjustment to the diagnosis. These stages include denial, isolation, anger, bargaining, depression, acceptance, and hope. Of course, patients do not all progress through all the stages or proceed through them in the same order or at the

same rate. Nevertheless, developing an understanding of how the patient has been affected by the diagnosis and is coping with it is an important goal of patient management.

It is best to speak frankly with the patient and the family regarding the likely course of disease. These discussions can be difficult for the physician as well as for the patient and family. The critical features of the interaction are to reassure the patient and family that everything that can be done to provide comfort will be done. They will not be abandoned. Many patients prefer to be cared for in their homes or in a hospice setting rather than a hospital. The American College of Physicians has published a book called *Home Care Guide for Cancer: How to Care for Family and Friends at Home* that teaches an approach to successful problem-solving in home care. With appropriate planning, it should be possible to provide the patient with the necessary medical care as well as the psychological and spiritual support that will prevent the isolation and depersonalization that can attend in-hospital death.

The care of dying patients may take a toll on the physician. A “burnout” syndrome has been described that is characterized by fatigue, disengagement from patients and colleagues, and a loss of self-fulfillment. Efforts at stress reduction, maintenance of a balanced life, and setting realistic goals may combat this disorder.

End-of-Life Decisions

Unfortunately, a smooth transition in treatment goals from curative to palliative may not be possible in all cases because of the occurrence of serious treatment-related complications or rapid disease progression. Vigorous and invasive medical support for a reversible disease or treatment complication is assumed to be justified. However, if the reversibility of the condition is in doubt, the patient’s wishes determine the level of medical care. These wishes should be elicited before the terminal phase of illness and reviewed periodically. Information about advance directives can be obtained from the American Association of Retired Persons, 601 E Street, NW, Washington, DC 20049, 202-434-2277, or Choice in Dying, 250 West 57th Street, New York, NY 10107, 212-366-5540. Some states allow physicians to assist patients who choose to end their lives. This subject is challenging from an ethical and a medical point of view. Discussions of end-of-life decisions should be candid and involve clear informed consent, waiting periods, second opinions, and documentation. **A full discussion of end-of-life management is in Chap. 9**.

FURTHER READING

Brandt JM et al: Chronic and refractory pain: A systematic review of pharmacologic management in oncology. Clin J Oncol Nurs 21:31, 2017. [PubMed: 28524906]

Kelley AS, Morrison RS: Palliative care for the seriously ill. N Engl J Med 373:747, 2015. [PubMed: 26287850]

Navari RM, Aapro M: Antiemetic prophylaxis for chemotherapy-induced nausea and vomiting. N Engl J Med 374:1356, 2016. [PubMed: 27050207]

Siegel RL et al: Cancer statistics, 2017. CA Cancer J Clin 67:7, 2017. [PubMed: 28055103]

Chapter 66: Prevention and Early Detection of Cancer

Jennifer M. Croswell; Otis W. Brawley; Barnett S. Kramer

INTRODUCTION

Improved understanding of carcinogenesis has allowed cancer prevention and early detection to expand beyond the identification and avoidance of carcinogens. Specific interventions to reduce cancer mortality by preventing cancer in those at risk, and effective screening for early detection of cancer, are the goals.

Carcinogenesis is a process that usually extends over years, a continuum of discrete tissue and cellular changes over time resulting in aberrant physiologic processes. Prevention concerns the identification and manipulation of the biologic, environmental, social, and genetic factors in the causal pathway of cancer.

EDUCATION AND HEALTHFUL HABITS

Public education on the avoidance of identified risk factors for cancer and encouraging healthy habits contributes to cancer prevention. The clinician is a powerful messenger in this process. The patient-provider encounter provides an opportunity to teach patients about the hazards of smoking, features of a healthy lifestyle, and use of proven cancer screening methods.

SMOKING CESSATION

Tobacco smoking is a strong, modifiable risk factor for cardiovascular disease, pulmonary disease, and cancer. Smokers have an ~1 in 3 lifetime risk of dying prematurely from a tobacco-related cancer, cardiovascular, or pulmonary disease. Tobacco use causes more deaths from cardiovascular disease than from cancer. Lung cancer and cancers of the larynx, oropharynx, esophagus, kidney, bladder, colon, pancreas, and stomach are all tobacco-related.

The number of cigarettes smoked per day and the level of inhalation of cigarette smoke are correlated with risk of lung cancer mortality. Light- and low-tar cigarettes are not safer, because smokers tend to inhale them more frequently and deeply.

Those who stop smoking have a 30–50% lower 10-year lung cancer mortality rate compared to those who continue smoking, despite the fact that some carcinogen-induced gene mutations persist for years after smoking cessation. Smoking cessation and avoidance would save more lives than any other public health activity.

The risk of tobacco smoke is not limited to the smoker. Environmental tobacco smoke, known as secondhand or passive smoke, causes lung cancer and other cardiopulmonary diseases in nonsmokers.

Tobacco use prevention is a pediatric issue. More than 80% of adult American smokers began smoking before the age of 18 years. Approximately 13% of Americans in grades 9 through 12 reported using two or more tobacco products in the past month. Electronic cigarettes have been advanced as a tool to achieve smoking cessation in adult smokers, but there is concern that they serve as a "gateway" to cigarette uptake in adolescents and are increasing in use. Counseling of adolescents and young adults is critical to prevent smoking. A clinician's simple advice can be of benefit. Providers should query patients on tobacco use and offer smokers assistance in quitting.

Current approaches to smoking cessation recognize smoking as an addiction **(Chap. 448)**. The smoker who is quitting goes through identifiable stages including: contemplation of quitting, an action phase in which the smoker quits, and a maintenance phase. Smokers who quit completely are more likely to be successful than those who gradually reduce the number of cigarettes smoked or change to lower-tar or lower-nicotine cigarettes. More than 90% of the Americans who have successfully quit smoking did so on their own, without participation in an organized cessation program, but cessation programs are helpful for some. The Community Intervention Trial for Smoking Cessation (COMMIT) was a 4-year program showing that light smokers (<25 cigarettes per day) were more likely to benefit from simple cessation messages and cessation programs than those who did not receive an intervention. Quit rates were 30.6% in the intervention group and 27.5% in the control group. The COMMIT interventions were unsuccessful in heavy smokers (>25 cigarettes per day). Heavy smokers may need an intensive broad-based cessation program that includes counseling, behavioral strategies, and pharmacologic adjuncts, such as nicotine replacement (gum, patches, sprays, lozenges, and inhalers), bupropion, and/or varenicline.

The health risks of cigars are similar to those of cigarettes. Smoking one or two cigars daily doubles the risk for oral and esophageal cancers; smoking three or four cigars daily increases the risk of oral cancers more than eightfold and esophageal cancer fourfold. The risks of occasional use are unknown.

Smokeless tobacco also represents a substantial health risk. Chewing tobacco is a carcinogen linked to dental caries, gingivitis, oral leukoplakia, and oral cancer. The systemic effects of smokeless tobacco (including snuff) may increase risks for other cancers. Esophageal cancer is linked to carcinogens in tobacco dissolved in saliva and swallowed. The net effects of e-cigarettes on health are poorly studied.

PHYSICAL ACTIVITY

Physical activity is associated with a decreased risk of colon and breast cancer. A variety of mechanisms have been proposed. However, such studies are prone to confounding factors such as recall bias, association of exercise with other health-related practices, and effects of preclinical cancers on exercise habits (reverse causality).

DIET MODIFICATION

International epidemiologic studies suggest that diets high in fat are associated with increased risk for cancers of the breast, colon, prostate, and endometrium. These cancers have their highest incidence and mortalities in Western cultures, where fat composes an average of one-third of the total calories consumed.

Despite correlations, dietary fat has not been proven to cause cancer. Case-control and cohort epidemiologic studies give conflicting results. In addition, diet is a highly complex exposure to many nutrients and chemicals. Low-fat diets are associated with many dietary changes beyond simple subtraction of fat. Other lifestyle changes are also associated with adherence to a low-fat diet.

In observational studies, dietary fiber is associated with a reduced risk of colonic polyps and invasive cancer of the colon. However, cancer-protective effects of increasing fiber and lowering dietary fat have not been proven in the context of a prospective clinical trial. The putative protective mechanisms are complex and speculative. Fiber binds oxidized bile acids and generates soluble fiber products, such as butyrate, that may have differentiating properties. Fiber does not increase bowel transit times. Two large prospective cohort studies of >100,000 health professionals showed no association between fruit and vegetable intake and risk of cancer.

The Polyp Prevention Trial randomly assigned 2000 elderly persons, who had polyps removed, to a low-fat, high-fiber diet versus routine diet for 4 years. No differences were noted in polyp formation.

The U.S. National Institutes of Health Women's Health Initiative, launched in 1994, was a long-term clinical trial enrolling >100,000 women age 45–69 years. It placed women into 22 intervention groups. Participants received calcium/vitamin D supplementation; hormone replacement therapy; and counseling to increase exercise, eat a low-fat diet with increased consumption of fruits, vegetables, and fiber, and cease smoking. The study showed that although dietary fat intake was lower in the diet intervention group, invasive breast cancers were not reduced over an 8-year follow-up period compared to the control group. No reduction was seen in the incidence of colorectal cancer in the dietary intervention arm. The difference in dietary fat averaged ~10% between the two groups. Evidence does not currently establish the anticarcinogenic value of vitamin, mineral, or nutritional supplements in amounts greater than those provided by a balanced diet.

ENERGY BALANCE

Risk of certain cancers appears to increase modestly (relative risks generally in the 1.0–2.0 range) as body mass index (BMI) increases beyond 25 kg/m^2. A cohort study of >5 million adults included in the U.K. Clinical Practice Research Datalink (a primary care database) found that each 5 kg/m^2 increase in BMI was linearly associated with cancers of the uterus, gallbladder, kidney, cervix, thyroid, and leukemia. Positive associations were also noted between BMI and colon, liver, ovarian, and postmenopausal breast cancers, but these associations were not linear and the effect varied by individual characteristics. High BMI appears to have an inverse association with prostate and premenopausal breast cancer.

SUN AVOIDANCE

Nonmelanoma skin cancers (basal cell and squamous cell) are induced by cumulative exposure to ultraviolet (UV) radiation. Intermittent acute sun exposure and sun damage have been linked to melanoma, but the evidence is inconsistent. Sunburns, especially in childhood and adolescence, may be associated with an increased risk of melanoma in adulthood. Reduction of sun exposure through use of protective clothing and changing patterns of outdoor activities can reduce skin cancer risk. Sunscreens decrease the risk of actinic keratoses, the precursor to squamous cell skin cancer, but melanoma risk may not be reduced. Sunscreens prevent burning, but they may encourage more prolonged exposure to the sun and may not filter out wavelengths of energy that cause melanoma.

Appearance-focused behavioral interventions in young women can decrease indoor tanning use and other UV exposures and may be more effective than messages about long-term cancer risks. Self-examination for skin pigment characteristics associated with skin cancer, such as freckling, may be useful in identifying people at high risk. Those who recognize themselves as being at risk tend to be more compliant with sun-avoidance recommendations. Risk factors for melanoma include a propensity to sunburn, a large number of benign melanocytic nevi, and atypical nevi.

CANCER CHEMOPREVENTION

Chemoprevention involves the use of specific natural or synthetic chemical agents to reverse, suppress, or prevent carcinogenesis before the development of invasive malignancy.

Cancer develops through an accumulation of tissue abnormalities associated with genetic and epigenetic changes, and growth regulatory pathways that are potential points of intervention to prevent cancer. The initial changes are termed *initiation*. The alteration can be inherited or acquired through the action of physical, infectious, or chemical carcinogens. Like most human diseases, cancer arises from an interaction between genetics and environmental exposures **(Table 66-1)**. Influences that cause the initiated cell and its surrounding tissue microenvironment to progress through the carcinogenic process and change phenotypically are termed *promoters*. Promoters include hormones such as androgens, linked to prostate cancer, and estrogen, linked to breast and endometrial cancer. The distinction between an initiator and promoter is indistinct; some components of cigarette smoke are “complete carcinogens,” acting as both initiators and promoters. Cancer can be prevented or controlled through interference with the factors that cause cancer initiation, promotion, or progression. Compounds of interest in chemoprevention often have antimutagenic, hormone modulation, anti-inflammatory, antiproliferative, or proapoptotic activity (or a combination).

TABLE 66-1

Suspected Carcinogens

Carcinogens[a]	Associated Cancer or Neoplasm
Alkylating agents	Acute myeloid leukemia, bladder cancer
Androgens	Prostate cancer
Aromatic amines (dyes)	Bladder cancer
Arsenic	Cancer of the lung, skin
Asbestos	Cancer of the lung, pleura, peritoneum
Benzene	Acute myelocytic leukemia
Chromium	Lung cancer
Diethylstilbestrol (prenatal)	Vaginal cancer (clear cell)
Epstein-Barr virus	Burkitt's lymphoma, nasal T-cell lymphoma
Estrogens	Cancer of the endometrium, liver, breast
Ethyl alcohol	Cancer of the breast, liver, esophagus, head and neck
Helicobacter pylori	Gastric cancer, gastric MALT lymphoma
Hepatitis B or C virus	Liver cancer
Human immunodeficiency virus	Non-Hodgkin's lymphoma, Kaposi's sarcoma, squamous cell carcinomas (especially of the urogenital tract)
Human papilloma virus	Cancers of the cervix, anus, oropharynx
Human T-cell lymphotropic virus type 1 (HTLV-1)	Adult T-cell leukemia/lymphoma
Immunosuppressive agents (azathioprine, cyclosporine, glucocorticoids)	Non-Hodgkin's lymphoma

Carcinogens[a]	Associated Cancer or Neoplasm
Ionizing radiation (therapeutic or diagnostic)	Breast, bladder, thyroid, soft tissue, bone, hematopoietic, and many more
Nitrogen mustard gas	Cancer of the lung, head and neck, nasal sinuses
Nickel dust	Cancer of the lung, nasal sinuses
Diesel exhaust	Lung cancer (miners)
Phenacetin	Cancer of the renal pelvis and bladder
Polycyclic hydrocarbons	Cancer of the lung, skin (especially squamous cell carcinoma of scrotal skin)
Radon gas	Lung cancer
Schistosomiasis	Bladder cancer (squamous cell)
Sunlight (ultraviolet)	Skin cancer (squamous cell and melanoma)
Tobacco (including smokeless)	Cancer of the upper aerodigestive tract, bladder
Vinyl chloride	Liver cancer (angiosarcoma)

[a]Agents that are thought to act as cancer initiators and/or promoters.

CHEMOPREVENTION OF CANCERS OF THE UPPER AERODIGESTIVE TRACT

Smoking causes diffuse epithelial injury in the oral cavity, neck, esophagus, and lung. Patients cured of squamous cell cancers of the lung, esophagus, oral cavity, and neck are at risk (as high as 5% per year) of developing second cancers of the upper aerodigestive tract. Cessation of cigarette smoking does not markedly decrease the cured cancer patient's risk of second malignancy, even though it does lower the cancer risk in those who have never developed a malignancy. Smoking cessation may halt the early stages of the carcinogenic process (such as metaplasia), but it may have no effect on late stages of carcinogenesis. This "field carcinogenesis" hypothesis for upper aerodigestive tract cancer has made "cured" patients an important population for chemoprevention of second malignancies.

Persistent oral human papilloma virus (HPV) infection, particularly HPV-16, increases the risk for cancers of the oropharynx. This association exists even in the absence of other risk factors such as smoking or alcohol use (although the magnitude of increased risk appears greater than additive when HPV infection and smoking are both present). Oral HPV infection is believed to be largely sexually acquired. Although the evidence is not definitive, the introduction of the HPV vaccine may eventually reduce oropharyngeal cancer rates.

Oral leukoplakia, a premalignant lesion commonly found in smokers, has been used as an intermediate marker of chemopreventive activity in smaller shorter-duration, randomized, placebo-controlled trials. Response was associated with upregulation of retinoic acid receptor-β (RAR-β). Therapy with high, relatively toxic doses of isotretinoin (13-*cis*-retinoic acid) causes regression of oral leukoplakia. However, the lesions recur when the therapy is withdrawn, suggesting the need for long-term administration. More tolerable doses of isotretinoin have not shown benefit in the prevention of head and neck cancer. Isotretinoin did not prevent second malignancies in patients cured of early-stage non-small cell lung cancer; mortality rates were actually increased in current smokers.

Several large-scale trials have assessed agents in the chemoprevention of lung cancer in patients at high risk. In the α-tocopherol/β-carotene (ATBC) Lung Cancer Prevention Trial, participants were male smokers, age 50–69 years at entry. Participants had smoked an average of one pack of cigarettes per day for 35.9 years. Participants received α-tocopherol, β-carotene, and/or placebo in a randomized, two-by-two factorial design. After median follow-up of 6.1 years, lung cancer incidence and mortality were statistically significantly increased in those receiving β-carotene. α-Tocopherol had no effect on lung cancer mortality, with no apparent interaction between the two drugs. Patients receiving α-tocopherol had a higher incidence of hemorrhagic stroke.

The β-Carotene and Retinol Efficacy Trial (CARET) involved 17,000 American smokers and workers with asbestos exposure. Entrants were randomly assigned to one of four arms and received β-carotene, retinol, and/or placebo in a two-by-two factorial design. This trial also demonstrated harm from β-carotene: a lung cancer rate of 5 per 1000 subjects per year for those taking placebo versus 6 per 1000 subjects per year for those taking β-carotene.

The ATBC and CARET results demonstrate the importance of testing chemoprevention hypotheses thoroughly before widespread implementation because the results contradict a number of observational studies. The Physicians' Health Trial showed no change in the risk of lung cancer for those taking β-carotene; however, fewer of its participants were smokers than those in the ATBC and CARET studies.

CHEMOPREVENTION OF COLON CANCER

Many colon cancer prevention trials are based on the premise that most colorectal cancers develop from adenomatous polyps. These trials use adenoma recurrence or disappearance as a surrogate endpoint (not yet validated) for colon cancer prevention. Early clinical trial results suggest that nonsteroidal anti-inflammatory drugs (NSAIDs), such as piroxicam, sulindac, and aspirin, may prevent adenoma formation or

cause regression of adenomatous polyps. The mechanism of action of NSAIDs is unknown, but they are presumed to work through the cyclooxygenase pathway. Although two randomized controlled trials (the Physicians' Health Study and the Women's Health Study) did not show an effect of aspirin on colon cancer or adenoma incidence in persons with no previous history of colonic lesions after 10 years of therapy, these trials did show an approximately 18% relative risk reduction for colonic adenoma incidence in persons with a previous history of adenomas after 1 year. A meta-analysis of four randomized controlled trials (albeit primarily designed to examine aspirin's effects on cardiovascular events) found that aspirin at doses of at least 75 mg/d resulted in a 33% relative reduction in colorectal cancer incidence after 20 years, with no clear increase in efficacy at higher doses. Based on a systematic review of evidence from randomized trials for primary prevention of cardiovascular disease, the U.S. Preventive Services Task Force concluded that the balance of benefits and harms favored initiating low-dose aspirin for colorectal cancer prevention in adults age 50–59 if they have a 10% or greater 10-year risk of cardiovascular disease. Cyclooxygenase-2 (COX-2) inhibitors have also been considered for colorectal cancer and polyp prevention. Trials with COX-2 inhibitors were initiated, but an increased risk of cardiovascular events in those taking the COX-2 inhibitors was noted, suggesting that these agents are not suitable for chemoprevention in the general population.

Epidemiologic studies suggest that diets high in calcium lower colon cancer risk. Calcium binds bile and fatty acids, which cause proliferation of colonic epithelium. It is hypothesized that calcium reduces intraluminal exposure to these compounds. The randomized controlled Calcium Polyp Prevention Study found that calcium supplementation decreased the absolute risk of adenomatous polyp recurrence by 7% at 4 years; extended observational follow-up demonstrated a 12% absolute risk reduction 5 years after cessation of treatment. However, in the Women's Health Initiative, combined use of calcium carbonate and vitamin D twice daily did not reduce the incidence of invasive colorectal cancer compared with placebo after 7 years.

The Women's Health Initiative demonstrated that postmenopausal women taking estrogen plus progestin have a 44% lower relative risk of colorectal cancer compared to women taking placebo. Of >16,600 women randomized and followed for a median of 5.6 years, 43 invasive colorectal cancers occurred in the hormone group and 72 in the placebo group. The positive effect on colon cancer is mitigated by the modest increase in cardiovascular and breast cancer risks associated with combined estrogen plus progestin therapy.

Most case-control and cohort studies have not confirmed early reports of an association between regular statin use and a reduced risk of colorectal cancer. No randomized controlled trials have addressed this hypothesis. A meta-analysis of statin use showed no protective effect of statins on overall cancer incidence or death.

CHEMOPREVENTION OF BREAST CANCER

Tamoxifen is an antiestrogen with partial estrogen agonistic activity in some tissues, such as endometrium and bone. One of its actions is to upregulate transforming growth factor β, which decreases breast cell proliferation. In a randomized placebo-controlled prevention trial involving >13,000 pre- and postmenopausal women at high risk, tamoxifen decreased the risk of developing breast cancer by 49% (from 43.4 to 22 per 1000 women) after a median follow-up of nearly 6 years. Tamoxifen also reduced bone

fractures; a small increase in risk of endometrial cancer, stroke, pulmonary emboli, and deep vein thrombosis was noted. The International Breast Cancer Intervention Study (IBIS-I) and the Italian Randomized Tamoxifen Prevention Trial also demonstrated a reduction in breast cancer incidence with tamoxifen use. A trial comparing tamoxifen with another selective estrogen receptor modulator, raloxifene, performed in postmenopausal women showed that raloxifene is comparable to tamoxifen in cancer prevention, but without the risk of endometrial cancer. Raloxifene was associated with more invasive breast cancers and a trend toward more noninvasive breast cancers, but fewer thromboembolic events than tamoxifen; the drugs are similar in risks of other cancers, fractures, ischemic heart disease, and stroke. Both tamoxifen and raloxifene (the latter for postmenopausal women only) have been approved by the U.S. Food and Drug Administration (FDA) for reduction of breast cancer in women at high risk for the disease (1.66% risk at 5 years based on the Gail risk model: *http://www.cancer.gov/bcrisktool/*).

Because the aromatase inhibitors are even more effective than tamoxifen in adjuvant breast cancer therapy, it has been hypothesized that they would be more effective in breast cancer prevention. A randomized, placebo-controlled trial of exemestane reported a 65% relative reduction (from 5.5 to 1.9 per 1000 women) in the incidence of invasive breast cancer in women at elevated risk after a median follow-up of about 3 years. Common adverse effects included arthralgias, hot flashes, fatigue, and insomnia. No trial has directly compared aromatase inhibitors with selective estrogen receptor modulators for breast cancer chemoprevention.

CHEMOPREVENTION OF PROSTATE CANCER

Finasteride and dutasteride are 5-α-reductase inhibitors. They inhibit conversion of testosterone to dihydrotestosterone (DHT), a potent stimulator of prostate cell proliferation. The Prostate Cancer Prevention Trial (PCPT) randomly assigned men age 55 years or older at average risk of prostate cancer to finasteride or placebo. All men in the trial were being regularly screened with prostate-specific antigen (PSA) levels and digital rectal examination. After 7 years of therapy, the incidence of prostate cancer was 18.4% in the finasteride arm, compared with 24.4% in the placebo arm, a statistically significant difference. However, the finasteride group had more patients with tumors of Gleason score 7 and higher compared with the placebo arm (6.4 vs 5.1%). Long-term (10–15 years) follow-up did not reveal any statistically significant differences in overall mortality between all men in the finasteride and placebo arms or in men diagnosed with prostate cancer, but the power to detect a difference was limited.

Dutasteride has also been evaluated as a preventive agent for prostate cancer. The Reduction by Dutasteride of Prostate Cancer Events (REDUCE) trial was a randomized double-blind trial in which ~8200 men with an elevated PSA (2.5–10 ng/mL for men age 50–60 years and 3–10 ng/mL for men age 60 years or older) and negative prostate biopsy on enrollment received daily 0.5 mg of dutasteride or placebo. The trial found a statistically significant 23% relative risk reduction in the incidence of biopsy-detected prostate cancer in the dutasteride arm at 4 years of treatment (659 cases vs 858 cases, respectively). Overall, across years 1 through 4, there was no difference between the arms in the number of tumors with a Gleason score of 7 to 10;

however, during years 3 and 4, there was a statistically significant difference in tumors with Gleason score of 8 to 10 in the dutasteride arm (12 tumors vs 1 tumor, respectively).

The clinical importance of the apparent increased incidence of higher-grade tumors in the 5-α-reductase inhibitor arms of these trials is controversial. It may represent an increased sensitivity of PSA and digital rectal exam for high-grade tumors in men receiving these agents. The FDA has analyzed both trials, and it determined that the use of a 5-α-reductase inhibitor for prostate cancer chemoprevention would result in one additional high-grade (Gleason score 8 to 10) prostate cancer for every three to four lower-grade (Gleason score <6) tumors averted. Although it acknowledged that detection bias may have accounted for the finding, a causative role for 5-α-reductase inhibitors could not be conclusively dismissed. These agents are therefore not FDA-approved for prostate cancer prevention.

Because all men in both the PCPT and REDUCE trials were being screened and because screening approximately doubles the rate of prostate cancer, it is not known if finasteride or dutasteride decreases the risk of prostate cancer in men who are not being screened or simply reduces the risk of non-life threatening cancers detectable by screening.

Several favorable laboratory and observational studies led to the formal evaluation of selenium and α-tocopherol (vitamin E) as potential prostate cancer preventives. The Selenium and Vitamin E Cancer Prevention Trial (SELECT) assigned 35,533 men to receive 200 µg/d selenium, 400 IU/d α-tocopherol, selenium plus vitamin E, or placebo. After a median follow-up of 7 years, a trend toward an increased risk of developing prostate cancer was observed for those men taking vitamin E alone as compared to the placebo arm (hazard ratio 1.17; 95% confidence interval, 1.004–1.36).

VACCINES AND CANCER PREVENTION

A number of infectious agents cause cancer. Hepatitis B and C are linked to liver cancer; some HPV strains are linked to cervical, anal, and head and neck cancer; and *Helicobacter pylori* is associated with gastric adenocarcinoma and gastric lymphoma. Vaccines to protect against these agents may therefore reduce the risk of their associated cancers.

The hepatitis B vaccine is effective in preventing hepatitis and hepatomas due to chronic hepatitis B infection.

A nonavalent vaccine (covering HPV strains 6, 11, 16, 18, 31, 33, 45, 52, and 58) is available for use in the United States. HPV types 6 and 11 cause genital papillomas. The remaining HPV types cause cervical and anal cancer; reduction in HPV types 16 and 18 alone could prevent >70% of cervical cancers worldwide. For individuals not previously infected with these HPV strains, the vaccine demonstrates high efficacy in preventing persistent strain-specific HPV infections. Studies also confirm the vaccine's ability to prevent preneoplastic lesions (cervical or anal intraepithelial neoplasia [CIN/AIN] I, II, and III). The durability of the immune response beyond 8–10 years is not currently known. The vaccine does not appear to impact preexisting infections and the efficacy appears to be lower for populations that had previously been exposed

to vaccine-specific HPV strains. A two-dose schedule is currently recommended in the United States for females and males age 9–14 years; teens and young adults who start the series between 15 and 26 years are recommended to receive three doses of the vaccine.

SURGICAL PREVENTION OF CANCER

Some organs in some individuals are at such high risk of developing cancer that surgical removal of the organ at risk may be considered. Women with severe cervical dysplasia are treated with laser or loop electrosurgical excision or conization and occasionally even hysterectomy. Colectomy is used to prevent colon cancer in patients with familial polyposis or ulcerative colitis.

Prophylactic bilateral mastectomy may be chosen for breast cancer prevention among women with genetic predisposition to breast cancer. In a prospective series of 139 women with *BRCA1* and *BRCA2* mutations, 76 chose to undergo prophylactic mastectomy and 63 chose close surveillance. At 3 years, no cases of breast cancer had been diagnosed in those opting for surgery, but eight patients in the surveillance group had developed breast cancer. A larger (n = 639) retrospective cohort study reported that three patients developed breast cancer after prophylactic mastectomy compared with an expected incidence of 30–53 cases: a 90–94% reduction in breast cancer risk. Postmastectomy breast cancer–related deaths were reduced by 81–94% for high-risk women compared with sister controls and by 100% for moderate-risk women when compared with expected rates.

Prophylactic salpingo-oophorectomy may also be employed for the prevention of ovarian and breast cancers among high-risk women. A prospective cohort study evaluating the outcomes of *BRCA* mutation carriers demonstrated a statistically significant association between prophylactic salpingo-oophorectomy and a reduced incidence of ovarian or primary peritoneal cancer (36% relative risk reduction, or a 4.5% absolute difference). Studies of prophylactic oophorectomy for prevention of breast cancer in women with genetic mutations have shown relative risk reductions of approximately 50%; the risk reduction may be greatest for women having the procedure at younger (i.e., <50 years) ages. The observation that most high-grade serous “ovarian cancers” actually arise in the fallopian tube fimbria raises the possibility that this lethal subtype may be prevented by ovary-sparing salpingectomy.

All of the evidence concerning the use of prophylactic mastectomy and salpingo-oophorectomy for prevention of breast and ovarian cancer in high-risk women has been observational in nature; such studies are prone to a variety of biases, including case selection bias, family relationships between patients and controls, and inadequate information about hormone use. Thus, they may give an overestimate of the magnitude of benefit.

CANCER SCREENING

Screening is a means of early detection in asymptomatic individuals, with the goal of decreasing morbidity and mortality. While screening can potentially reduce disease-specific deaths and has been shown to do so in cervical, colon, lung, and breast cancer, it is also subject to a number of biases that can suggest a benefit

when actually there is none. Biases can even mask net harm. Early detection does not in itself confer benefit. Cause-specific mortality, rather than survival after diagnosis, is the preferred endpoint (see below).

Because screening is done on asymptomatic, healthy persons, it should offer substantial likelihood of benefit that outweighs harm. Screening tests and their appropriate use should be carefully evaluated before their use is widely encouraged in screening programs.

A large and increasing number of genetic mutations and nucleotide polymorphisms have been associated with an increased risk of cancer. Testing for these genetic mutations could in theory define a high-risk population. However, most of the identified mutations have very low penetrance and individually provide limited predictive accuracy. The ability to predict the development of a particular cancer may someday present therapeutic options as well as ethical dilemmas. It may eventually allow for early intervention to prevent a cancer or limit its severity. People at high risk may be ideal candidates for chemoprevention and screening; however, efficacy of these interventions in the high-risk population should be investigated. Currently, persons at high risk for a particular cancer can engage in intensive screening. While this course is clinically reasonable, it is not known if it reduces mortality in these populations.

The Accuracy of Screening

A screening test's accuracy or ability to discriminate disease is described by four indices: sensitivity, specificity, positive predictive value, and negative predictive value **(Table 66-2)**. *Sensitivity*, also called the true-positive rate, is the proportion of persons with the disease who test positive in the screen (i.e., the ability of the test to detect disease when it is present). *Specificity*, or 1 minus the false-positive rate, is the proportion of persons who do not have the disease that test negative in the screening test (i.e., the ability of a test to correctly indicate that the disease is not present). The *positive predictive value* is the proportion of persons who test positive that actually have the disease. Similarly, *negative predictive value* is the proportion testing negative that do not have the disease. The sensitivity and specificity of a test are independent of the underlying prevalence (or risk) of the disease in the population screened, but the predictive values depend strongly on the prevalence of the disease.

TABLE 66-2

Assessment of the Value of a Diagnostic Test[a]

	Condition Present	**Condition Absent**
Positive test	*a*	*b*
Negative test	*c*	*d*
a = true positive		
b = false positive		
c = false negative		
d = true negative		
Sensitivity	The proportion of persons with the condition who test positive: $a/(a+c)$	
Specificity	The proportion of persons without the condition who test negative: $d/(b+d)$	
Positive predictive value (PPV)	The proportion of persons with a positive test who have the condition: $a/(a+b)$	
Negative predictive value	The proportion of persons with a negative test who do not have the condition: $d/(c+d)$	
Prevalence, sensitivity, and specificity determine PPV $PPV = \frac{\text{prevalence} \times \text{sensitivity}}{(\text{prevalence} \times \text{sensitivity}) + (1 - \text{prevalence})(1 - \text{specificity})}$		

[a]For diseases of low prevalence, such as cancer, poor specificity has a dramatic adverse effect on PPV such that only a small fraction of positive tests are true positives.

Screening is most beneficial, efficient, and economical when the target disease is common in the population being screened. Specificity is at least as important to the ultimate feasibility and success of a screening test as sensitivity.

Potential Biases of Screening Tests

Common biases of screening are lead time, length-biased sampling, and selection. These biases can make a screening test seem beneficial when actually it is not (or even causes net harm). Whether beneficial or not, screening can create the false impression of an epidemic by increasing the number of cancers diagnosed. It can also produce a shift in the *proportion* of patients diagnosed at an early stage (even without a reduction in absolute incidence of late-stage disease) and inflate survival statistics without reducing mortality (i.e., the number of deaths from a given cancer relative to the number of those at risk for the cancer). In such a case, the *apparent* duration of survival (measured from date of diagnosis) increases without lives being saved or life expectancy changed.

Lead-time bias occurs whether or not a test influences the natural history of the disease; the patient is merely diagnosed at an earlier date. Survival *appears* increased even if life is not prolonged. The screening test only prolongs the time the subject is aware of the disease and spends as a patient.

Length-biased sampling occurs because screening tests generally can more easily detect slow-growing, less aggressive cancers than fast-growing cancers. Cancers diagnosed due to the onset of symptoms between scheduled screenings are on average more aggressive, and treatment outcomes are not as favorable. An extreme form of length bias sampling is termed *overdiagnosis*, the detection of “pseudo disease.” The reservoir of some undetected slow-growing tumors is large. Many of these tumors fulfill the histologic criteria of cancer but will never become clinically significant or cause death during the patient’s remaining lifespan. This problem is compounded by the fact that the most common cancers appear most frequently at ages when competing causes of death are more frequent.

Selection bias occurs because the population most likely to seek screening often differs from the general population to which the screening test might be applied. In general, volunteers for studies are more health conscious and likely to have a better prognosis or lower mortality rate, irrespective of the screening result. This is termed the *healthy volunteer effect*.

Potential Drawbacks of Screening

Risks associated with screening include harm caused by the screening intervention itself, harm due to the further investigation of persons with positive tests (both true and false positives), and harm from the treatment of persons with a true-positive result, whether or not life is extended by treatment (e.g., even if a screening test reduces relative cause-specific mortality by 20–30%, 70–80% of those diagnosed still go on to die of the target cancer). The diagnosis and treatment of cancers that would never have caused medical problems can lead to the harm of unnecessary treatment and give patients the anxiety of a cancer diagnosis. The psychosocial impact of cancer screening can be substantial when applied to the entire population.

Assessment of Screening Tests

Good clinical trial design can offset some biases of screening and demonstrate the relative risks and benefits of a screening test. A randomized controlled screening trial with cause-specific mortality as the endpoint provides the strongest support for a screening intervention. Overall mortality should also be reported to

detect an adverse effect of screening and treatment on other disease outcomes (e.g., cardiovascular disease). In a randomized trial, two like populations are randomly established. One is given the usual standard of care (which may be no screening at all) and the other receives the screening intervention being assessed. Efficacy for the population studied is established when the group receiving the screening test has a better cause-specific mortality rate than the control group. Studies showing a reduction in the incidence of advanced-stage disease, improved survival, or a stage shift are weaker (and possibly misleading) evidence of benefit. These latter criteria are early indicators but not sufficient to establish the value of a screening test.

Although a randomized, controlled screening trial provides the strongest evidence to support a screening test, it is not perfect. Unless the trial is population-based, it does not remove the question of generalizability to the target population. Screening trials generally involve thousands of persons and last for years. Less definitive study designs are therefore often used to estimate the effectiveness of screening practices. However, every nonrandomized study design is subject to strong confounders. In descending order of strength, evidence may also be derived from the findings of internally controlled trials using intervention allocation methods other than randomization (e.g., allocation by birth date, date of clinic visit); the findings of analytic observational studies; or the results of multiple time series studies with or without the intervention.

Screening for Specific Cancers

Screening for cervical, colon, and breast cancer has the potential to be beneficial for certain age groups. Depending on age and smoking history, lung cancer screening can also be beneficial in specific settings. Special surveillance of those at high risk for a specific cancer because of a family history or a genetic risk factor may be prudent, but few studies have assessed the effect on mortality. A number of organizations have considered whether or not to endorse routine use of certain screening tests. Because criteria have varied, they have arrived at different recommendations. The American Cancer Society (ACS) and the U.S. Preventive Services Task Force (USPSTF) publish screening guidelines **(Table 66-3)**; the American Academy of Family Practitioners (AAFP) often follow/endorse the USPSTF recommendations; and the American College of Physicians (ACP) develops recommendations based on structured reviews of other organizations' guidelines.

TABLE 66-3

Screening Recommendations for Asymptomatic Subjects Not Known to Be at Increased Risk for the Target Condition[a]

Cancer Type	Test or Procedure	USPSTF	ACS
Breast	Self-examination	"D"[b] (Not in current recommendations; from 2009)	Women, all ages: No specific recommendation
	Clinical examination	Women ≥40 years: "I" (as a stand-alone without mammography) (Not in current recommendations; from 2009)	Women, all ages: Do not recommend
	Mammography	Women 40–49 years: The decision to start screening mammography in women prior to age 50 years should be an individual one. Women who place a higher value on the potential benefit than the potential harms may choose to begin biennial screening between the ages of 40 and 49 years. ("C")	Women 40–44 years: Provide the opportunity to begin annual screening Women 45–54 years: Screen annually Women ≥55 years: Transition to biennial screening or have the opportunity to continue annual screening Women ≥40 should continue screening mammography as long as their overall health is good and they have a life expectancy of 10 years or longer
		Women 50–74 years: Every 2 years ("B")	
		Women ≥75 years: "I"	
	Magnetic resonance imaging (MRI)	"I" (Not in current recommendations; from 2009)	Women with >20% lifetime risk of breast cancer: Screen with MRI plus mammography annually
			Women with 15–20% lifetime risk of breast cancer: Discuss option of MRI plus mammography annually
			Women with <15% lifetime risk of breast cancer: Do not screen annually with MRI

Cancer Type	Test or Procedure	USPSTF	ACS
	Tomosynthesis	Women, all ages: “I”	No specific recommendation
Cervical	Pap test (cytology)	Women 21–65 years: Screen every 3 years (“A”)	Women 21–29 years: Screen every 3 years
		Women <21 years: “D” Women >65 years, with adequate, normal prior Pap screenings: “D”	Women 30–65 years: Acceptable approach to screen with cytology every 3 years (see HPV test below) Women <21 years: No screening Women >65 years: No screening following adequate negative prior screening
		Women after total hysterectomy for noncancerous causes: “D”	Women after total hysterectomy for noncancerous causes: Do not screen
	HPV test	Women 30–65 years: Screen in combination with cytology every 5 years if woman desires to lengthen the screening interval (see Pap test above) (“A”) Women <30 years: “D” Women >65 years, with adequate, normal prior Pap screenings: “D” Women after total hysterectomy for noncancerous causes: “D”	Women 30–65 years: Preferred approach to screen with HPV and cytology co-testing every 5 years (see Pap test above) Women <30 years: Do not use HPV testing Women >65 years: No screening following adequate negative prior screening Women after total hysterectomy for noncancerous causes: Do not screen

Cancer Type	Test or Procedure	USPSTF	ACS
Colorectal	Sigmoidoscopy	Adults, 50–75 years: “A” Screen for colorectal cancer; the risks and benefits of the different screening methods vary Adults, 76 to 85 years: “C” The decision to screen should be an individual one, taking into account the patient’s overall health and prior screening history Every 5 years; modeling suggests improved benefit if performed every 10 years in combination with annual FIT	Adults ≥50 years: Screen every 5 years
	Fecal occult blood testing (FOBT)	Every year	Adults ≥50 years: Screen every year
	Colonoscopy	Every 10 years	Adults ≥50 years: Screen every 10 years
	Fecal DNA testing	Every 1 or 3 years	Adults ≥50 years: Screen, but interval uncertain
	Fecal immuno-chemical testing (FIT)	Every year	Adults ≥50 years: Screen every year
	CT colonography	Every 5 years	Adults ≥50 years: Screen every 5 years

Cancer Type	Test or Procedure	USPSTF	ACS
Lung	Low-dose computed tomography (CT) scan	Adults 55–80 years, with a ≥30 pack-year smoking history, still smoking or have quit within past 15 years: “B” Discontinue once a person has not smoked for 15 years or develops a health problem that substantially limits life expectancy or the ability to have curative lung surgery	Men and women, 55–74 years, with ≥30 pack-year smoking history, still smoking or have quit within past 15 years: Discuss benefits, limitations, and potential harms of screening; only perform screening in facilities with the right type of CT scanner and with high expertise/specialists
Ovarian	CA-125 Transvaginal ultrasound	Women, all ages: “D” Women, all ages: “D”	There is no sufficiently accurate test proven effective in the early detection of ovarian cancer. For women at high risk of ovarian cancer and/or who have unexplained, persistent symptoms, the combination of CA-125 and transvaginal ultrasound with pelvic exam may be offered.
Prostate	Prostate-specific antigen (PSA)	Men, all ages: “D”	Starting at age 50, men should talk to a doctor about the pros and cons of testing so they can decide if testing is the right choice for them. If African American or have a father or brother who had prostate cancer before age 65, men should have this talk starting at age 45. How often they are tested will depend on their PSA level
	Digital rectal examination (DRE)	No individual recommendation	As for PSA; if men decide to be tested, they should have the PSA blood test with or without a rectal exam
Skin	Complete skin examination by clinician or patient	Adults, all ages: “I”	Self-examination monthly; clinical exam as part of routine cancer-related checkup

[a]Summary of the screening procedures recommended for the general population by the USPSTF and the ACS. These recommendations refer to asymptomatic persons who are not known to have risk factors, other than age or gender, for the targeted condition. [b]USPSTF lettered recommendations are defined as follows: "A": The USPSTF recommends the service, because there is high certainty that the net benefit is substantial; "B": The USPSTF recommends the service, because there is high certainty that the net benefit is moderate or moderate certainty that the net benefit is moderate to substantial; "C": The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences; there is at least moderate certainty that the net benefit is small; "D": The USPSTF recommends against the service because there is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits; "I": The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service.

Abbreviations: ACS, American Cancer Society; USPSTF, U.S. Preventive Services Task Force.

BREAST CANCER

Breast self-examination, clinical breast examination by a caregiver, mammography, and magnetic resonance imaging (MRI) have all been variably advocated as useful screening tools.

A number of trials have suggested that annual or biennial screening with mammography or mammography plus clinical breast examination in normal-risk women older than age 50 years decreases breast cancer mortality. Each trial has been criticized for design flaws. In most trials, breast cancer-related mortality rates were decreased by 15–30%. Experts disagree on whether average-risk women age 40–49 years should receive regular screening (Table 66-3). The U.K. Age Trial, the only randomized trial of breast cancer screening to specifically evaluate the impact of mammography in women age 40–49 years, found no statistically significant difference in breast cancer mortality for screened women versus controls after about 11 years of follow-up (relative risk 0.83; 95% confidence interval 0.66–1.04); however, <70% of women received screening in the intervention arm, potentially diluting the observed effect. A meta-analysis of nine large randomized trials showed an 8% relative reduction in mortality (relative risk 0.92; 95% confidence interval 0.75–1.02) from mammography screening for women age 39–49 years after 11–20 years of follow-up. This is equivalent to 3 breast cancer deaths prevented per 10,000 women >10 years (although the result is not statistically significant). At the same time, nearly half of women age 40–49 years screened annually will have false-positive mammograms necessitating further evaluation, often including biopsy. Estimates of overdiagnosis range from 10 to 40% of diagnosed invasive cancers. In the United States, widespread screening over the last several decades has not been accompanied by a reduction in incidence of metastatic breast cancer despite a large increase in early-stage disease, suggesting a substantial amount of overdiagnosis at the population level.

Digital breast tomosynthesis is an emerging method of breast cancer screening that reconstructs multiple x-ray images of the breast into superimposed "three-dimensional" slices. Although some evidence is available concerning the test characteristics of this modality, there are currently no data on its effects on health outcomes such as breast cancer-related morbidity, mortality, or overdiagnosis rates.

No study of breast self-examination has shown it to decrease mortality. A randomized controlled trial of approximately 266,000 women in China demonstrated no difference in breast cancer mortality between a group that received intensive breast self-exam instruction and reinforcement/reminders and controls at 10 years of follow-up. However, more benign breast lesions were discovered and more breast biopsies were performed in the self-examination arm.

Genetic screening for *BRCA1* and *BRCA2* mutations and other markers of breast cancer risk has identified a group of women at high risk for breast cancer. Unfortunately, when to begin and the optimal frequency of screening have not been defined. Mammography is less sensitive at detecting breast cancers in women carrying *BRCA1* and *BRCA2* mutations, possibly because such cancers occur in younger women, in whom mammography is known to be less sensitive. MRI screening may be more sensitive than mammography in women at high risk due to genetic predisposition or in women with very dense breast tissue, but specificity may be lower. An increase in overdiagnosis may accompany the higher sensitivity. The impact of MRI on breast cancer mortality with or without concomitant use of mammography has not been evaluated in a randomized controlled trial.

CERVICAL CANCER

Screening with Papanicolaou (Pap) smears decreases cervical cancer mortality. The cervical cancer mortality rate has fallen substantially since the widespread use of the Pap smear. With the onset of sexual activity comes the risk of sexual transmission of HPV, the fundamental etiologic factor for cervical cancer. Screening guidelines recommend regular Pap testing for all women who have reached the age of 21 (before this age, even in individuals that have begun sexual activity, screening may cause more harm than benefit). The recommended interval for Pap screening is 3 years. Screening more frequently adds little benefit but leads to important harms, including unnecessary procedures and overtreatment of transient lesions. Beginning at age 30, guidelines also offer the alternative of combined Pap smear and HPV testing for women. The screening interval for women who test normal using this approach may be lengthened to 5 years.

An upper age limit at which screening ceases to be effective is not known, but women age 65 years with no abnormal results in the previous 10 years may choose to stop screening. Screening should be discontinued in women who have undergone a hysterectomy with cervical excision for noncancerous reasons.

Although the efficacy of the Pap smear in reducing cervical cancer mortality has never been directly confirmed in a randomized, controlled setting, a clustered randomized trial in India evaluated the impact of one-time cervical visual inspection and immediate colposcopy, biopsy, and/or cryotherapy (where indicated) versus counseling on cervical cancer deaths in women age 30–59 years. After 7 years of follow-up, the age-standardized rate of death due to cervical cancer was 39.6 per 100,000 person-years in the intervention group versus 56.7 per 100,000 person-years in controls.

COLORECTAL CANCER

Fecal occult blood testing (FOBT), digital rectal examination (DRE), rigid and flexible sigmoidoscopy, colonoscopy, and computed tomography (CT) colonography have been considered for colorectal cancer screening. A meta-analysis of five randomized controlled trials demonstrated a 22% relative reduction in colorectal cancer mortality after 2 to 9 rounds of biennial FOBT at 30 years of follow-up; annual screening

was shown to result in a greater colorectal cancer mortality reduction in a single trial (a 32% relative reduction). The sensitivity for FOBT is increased if specimens are rehydrated before testing, but at the cost of lower specificity. The false-positive rate for rehydrated FOBT is high; 1–5% of persons tested have a positive test. Only 2–10% of those with occult blood in the stool have cancer. The high false-positive rate of FOBT substantially increases the number of colonoscopies performed.

Fecal immunochemical tests (FIT) have higher sensitivity for colorectal cancer than nonrehydrated FOBT tests. Multi-targeted stool DNA testing is an emerging screening modality that combines FIT with testing for altered DNA biomarkers in cells that are shed into the stool. Although limited evidence demonstrates that it has a higher single-test sensitivity for colorectal cancer than fecal immunochemical testing alone, its specificity is much lower, resulting in a higher number of false-positive tests and follow-up colonoscopies. There are no studies evaluating its effects on colorectal cancer incidence, morbidity, or mortality.

A blood test for the methylated *SEPT9* gene associated with colorectal cancer is available. However, its sensitivity is low, no longitudinal data have been collected on its performance or efficacy, and it is not recommended as a first-line screening test.

Two meta-analyses of five randomized controlled trials of sigmoidoscopy (i.e., the NORCCAP, SCORE, PLCO, Telemark, and U.K. trials) found an 18% relative reduction in colorectal cancer incidence and a 28% relative reduction in colorectal cancer mortality. Participant ages ranged from 50 to 74 years, with follow-up ranging from 6 to 13 years. Diagnosis of adenomatous polyps by sigmoidoscopy should lead to evaluation of the entire colon with colonoscopy. The most efficient interval for screening sigmoidoscopy is unknown, but an interval of 5 years is often recommended. Case-control studies suggest that intervals of up to 15 years may confer benefit; the randomized U.K. trial demonstrated benefit with one-time screening.

One-time colonoscopy detects ~25% more advanced lesions (polyps >10 mm, villous adenomas, adenomatous polyps with high-grade dysplasia, invasive cancer) than one-time FOBT with sigmoidoscopy; comparative *programmatic* performance of the two modalities over time is not known. Perforation rates are about 4/10,000 for colonoscopy and 1/10,000 for sigmoidoscopy. Debate continues on whether colonoscopy is too expensive and invasive and whether sufficient provider capacity exists to be recommended as the preferred screening tool in standard-risk populations. Some observational studies suggest that efficacy of colonoscopy to decrease colorectal cancer mortality is primarily limited to the left side of the colon.

CT colonography, if done at expert centers, appears to have a sensitivity for polyps ≥6 mm comparable to colonoscopy. However, the rate of extracolonic findings of abnormalities of uncertain significance that must nevertheless be worked up is high (~5–37%); the long-term cumulative radiation risk of repeated colonography screenings is also a concern.

LUNG CANCER

Chest x-ray and sputum cytology have been evaluated in several randomized lung cancer screening trials. The most recent and largest (n = 154,901) of these, a component of the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial, found that, compared with usual care, annual chest x-ray did not reduce the risk of dying from lung cancer (relative risk 0.99; 95% confidence interval 0.87–1.22) after 13 years.

Low-dose CT has also been evaluated in several randomized trials. The largest and longest of these, the National Lung Screening Trial (NLST), was a randomized controlled trial of screening for lung cancer in ~53,000 persons age 55–74 years with a 30+ pack-year smoking history. It demonstrated a statistically significant relative reduction of about 15–20% in lung cancer mortality in the CT arm compared to the chest x-ray arm (or about 3 fewer deaths per 1000 people screened with CT). However, the harms include the potential radiation risks associated with multiple scans, the discovery of incidental findings of unclear significance, and a high rate of false-positive test results. Both incidental findings and false-positive tests can lead to invasive diagnostic procedures associated with anxiety, expense, and complications (e.g., pneumo- or hemothorax after lung biopsy). The NLST was performed at experienced screening centers, and the balance of benefits and harms may differ in the community setting at less experienced centers.

OVARIAN CANCER

Adnexal palpation, transvaginal ultrasound (TVUS), and serum CA-125 assay have been considered for ovarian cancer screening. A large randomized controlled trial has shown that an annual screening program of TVUS and CA-125 in average-risk women does not reduce deaths from ovarian cancer (relative risk 1.21; 95% confidence interval 0.99–1.48). Adnexal palpation was dropped early in the study because it did not detect any ovarian cancers that were not detected by either TVUS or CA-125. A second large randomized trial that used a two-stage screening approach incorporating a risk of ovarian cancer algorithm which determined whether additional testing with CA-125 or TVUS was required. At 14 years of follow-up, there was no statistically significant reduction in ovarian cancer deaths. The risks and costs associated with the high number of false-positive results are impediments to routine use of these modalities for screening. In the PLCO trial, 10% of participants had a false-positive result from TVUS or CA-125, and one-third of these women underwent a major surgical procedure; the ratio of surgeries to screen-detected ovarian cancer was approximately 20:1. In September 2016, the FDA issued a safety communication recommending against using any screening test, including the risk of ovarian cancer algorithm, for ovarian cancer.

PROSTATE CANCER

The most common prostate cancer screening modalities are digital rectal exam (DRE) and serum PSA assay. An emphasis on PSA screening has caused prostate cancer to become the most common nonskin cancer diagnosed in American males. This disease is prone to lead-time bias, length bias, and overdiagnosis, and substantial debate continues among experts as to whether screening should be offered unless the patient specifically asks to be screened. Virtually all organizations stress the importance of informing men about the uncertainty regarding screening efficacy and the associated harms. Prostate cancer screening clearly detects many asymptomatic cancers, but the ability to distinguish tumors that are lethal but still curable from those that pose little or no threat to health is limited, and randomized trials indicate that the effect of PSA screening on prostate cancer mortality across a population is, at best, small. Men older than age 50 years have a high prevalence of indolent, clinically insignificant prostate cancers (about 30–50% of men, increasing further as men age).

Two major randomized controlled trials of the impact of PSA screening on prostate cancer mortality have been published. The PLCO Cancer Screening Trial was a multicenter U.S. trial that randomized almost 77,000 men age 55–74 years to receive either annual PSA testing for 6 years or usual care. At 13 years of follow-up, no

statistically significant difference in the number of prostate cancer deaths were noted between the arms (rate ratio 1.09; 95% confidence interval 0.87–1.36). More than half of men in the control arm received at least one PSA test during the trial, which may have potentially diluted a small effect.

The European Randomized Study of Screening for Prostate Cancer (ERSPC) was a multinational study that randomized ~182,000 men between age 50 and 74 years (with a predefined "core" screening group of men age 55–69 years) to receive PSA testing or no screening. Recruitment and randomization procedures, as well as actual frequency of PSA testing, varied by country. After a median follow-up of 13 years, a 21% relative reduction in the risk of prostate cancer death in the screened arm was noted in the "core" screening group. The trial found that 781 (95% CI 490–1,929) men would need to be invited to screening, and 27 (95% CI 17–66) cases of prostate cancer detected, to avert 1 death from prostate cancer. Of the seven countries included in the mortality analysis, two demonstrated statistically significant reductions in prostate cancer deaths, whereas five did not. There was also an imbalance in treatment between the two study arms, with a higher proportion of men with clinically localized cancer receiving radical prostatectomy in the screening arm and receiving it at experienced referral centers.

Screening must be linked to effective therapy in order to have any benefit. In a trial conducted in the United States after the initiation of widespread PSA testing, random assignment to radical prostatectomy compared with "watchful waiting" did not result in a statistically significant decrease in prostate cancer deaths (absolute risk reduction 2.7%; 95% confidence interval—1.3 to 6.2%). Likewise, in a randomized trial conducted in the U.K. comparing monitoring (no curative treatment) to radical prostatectomy and to radiotherapy in men diagnosed in a screening program, prostate-cancer specific survival was very good (about 99%), and nearly identical, in all three study arms at a median of 10 years follow-up. Treatments for low-stage prostate cancer, such as surgery and radiation therapy, can cause substantial morbidity, including impotence and urinary incontinence.

SKIN CANCER

Visual examination of all skin surfaces by the patient or by a health care provider is used in screening for basal and squamous cell cancers and melanoma. No prospective randomized study has been performed to look for a mortality decrease. Unfortunately, screening is associated with a substantial rate of overdiagnosis.

FURTHER READING

Carter JL, Coletti RJ, Harris RP: Quantifying and monitoring overdiagnosis in cancer screening: A systematic review of methods. BMJ 350:g7773, 2015. [PubMed: 25569206]

Chubak J et al: Aspirin for the prevention of cancer incidence and mortality: Systematic evidence review for the U.S. Preventive Services Task Force. Ann Intern Med 164:814, 2016. [PubMed: 27064482]

Future II Study Group: Quadrivalent vaccine against human papillomavirus to prevent high-grade cervical lesions. N Engl J Med 356(19):1915 2007. [PubMed: 17494925]

Hamdy FC et al: 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. N Engl J Med 375:1415, 2016. [PubMed: 27626136]

Humphrey LL et al: Screening for lung cancer with low-dose computed tomography: A systematic review to update the U.S. Preventive Services Task Force recommendation. Ann Intern Med 159:411, 2013. [PubMed: 23897166]

Ilic D et al: Screening for prostate cancer. Cochrane Database of Systematic Reviews 0.1002/14651858.CD004720.pub3, 2013.

Kramer BS, Croswell JM: Cancer screening: The clash of science and intuition. Annu Rev Med 60:125, 2009. [PubMed: 18803476]

Lin JS et al: Screening for colorectal cancer: Updated evidence report and systematic review for the U.S. Preventive Services Task Force. JAMA 315:2576, 2016. [PubMed: 27305422]

Pace LE, Keating NL: A systematic assessment of benefits and risks to guide breast cancer screening decisions. JAMA 311:1327, 2014. [PubMed: 24691608]

Peirson L et al: Screening for cervical cancer: A systematic review and meta-analysis. Syst Rev 2:35, 2013. [PubMed: 23706117]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 67: Cancer Genetics

Fred Bunz; Bert Vogelstein

CANCER IS A GENETIC DISEASE

Cancer arises through a series of somatic alterations in DNA that result in unrestrained cellular proliferation. Most of these alterations involve subtle sequence changes in DNA (i.e., mutations). The somatic mutations may originate as a consequence of random replication errors or exposure to carcinogens (e.g., radiation) and can be exacerbated by faulty DNA repair processes. While most cancers arise sporadically, clustering of cancers occurs in families that carry a germline mutation in a cancer gene.

HISTORICAL PERSPECTIVE

The idea that cancer progression is driven by sequential somatic mutations in specific genes has only gained general acceptance in the past 30 years. Before the advent of the microscope, cancer was believed to be composed of aggregates of mucus or other noncellular matter. By the middle of the nineteenth century, it became clear that tumors were masses of cells and that these cells arose from the normal cells of the tissue from which the cancer originated. The molecular basis for the uncontrolled proliferation of cancer cells was to remain a mystery for another century. During that time, a number of theories for the origin of cancer were postulated. The great biochemist Otto Warburg proposed the combustion theory of cancer, which stipulated that cancer was due to abnormal oxygen metabolism. Others believed that all cancers were caused by viruses, and that cancer was in fact a contagious disease.

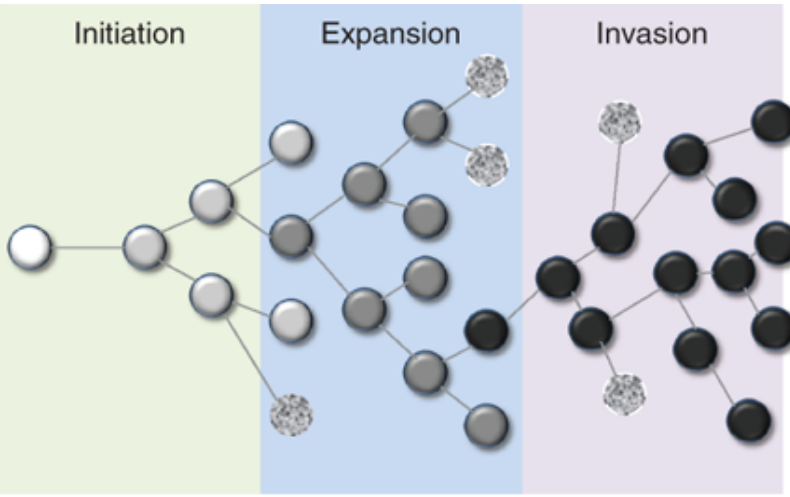
In the end, observations of cancer occurring in chimney sweeps, studies of x-rays, and the overwhelming data demonstrating cigarette smoke as a causative agent in lung cancer, together with Ames's work on chemical mutagenesis, were consistent with the idea that cancer originated through changes in DNA. However, it was not until the somatic mutations responsible for cancer were identified at the molecular level that the genetic basis of cancer was definitively established. Although the viral theory of cancer did not prove to be generally accurate (with the exception of human papillomaviruses, which can cause cervical and other cancers), the study of retroviruses led to the discovery of the first human *oncogenes* in the late 1970s. Oncogenes are one of the two major classes of cancer driver genes. The study of families with genetic predisposition to cancer was instrumental to the discovery of the other major class of cancer driver genes, called *tumor-suppressor genes*. Current technologies permit the sequence analysis of entire cancer genomes, and provide a comprehensive view of the genetic changes that cause tumors to arise and become malignant. The field that studies the various types of mutations, as well as the consequences of these mutations in tumor cells, is now known as *cancer genetics*.

THE CLONAL ORIGIN AND MULTISTEP NATURE OF CANCER

Nearly all cancers originate from a single cell; this clonal origin is a critical discriminating feature between neoplasia and hyperplasia. Multiple cumulative mutational events are invariably required for the progression of a tumor from normal to fully malignant phenotype. The process can be seen as Darwinian microevolution in which, at each successive step, the mutated cells gain a growth advantage resulting in the expansion of a neoplastic clone **(Fig. 67-1)**. Based on observations of cancer frequency increases during aging, the epidemiologists Armitage and Doll and Nordling independently proposed that cancer is a result of three discrete cellular changes. Remarkably, this early model has been validated by extensive sequencing of cancer genomes. These studies revealed that just three causal mutations are required for the development of several of the most common cancers. Overall, it is currently believed that most common solid tumors require a minimum of three mutated cancer driver genes (either oncogenes or tumor suppressor genes) for their development. One or two mutations is sufficient for benign tumorigenesis, but not for the invasive capacity that distinguishes cancers from benign tumors. Less common tumors, such as liquid tumors (leukemias or lymphomas), sarcomas, and childhood tumors, require two driver gene alterations for malignancy. Note that a cancer driver gene is best defined as one containing a mutation that increases the selective growth advantage of the cell containing it. Normally, cell birth and cell death are in perfect equilibrium; every time a cell is born, another in the same lineage dies. Cancer driver gene mutations alter this equilibrium, so that more cells are born than die. The imbalance is often slight, so that the difference between cell birth and cell death is <1%. This explains why tumorigenesis—the journey from a normal cell to a typical malignant, solid tumor—often takes decades.

FIGURE 67-1

Multistep clonal development of malignancy. In this diagram a series of three cumulative mutations, each with a modest growth advantage acting alone, eventually results in a malignant tumor. Note that not all such alterations result in progression. The actual number of cumulative mutations necessary to transform from the normal to the malignant state has been estimated to be three for several of the most common types of cancer. *(After P Nowell: Science 194:23, 1976, with permission.)*

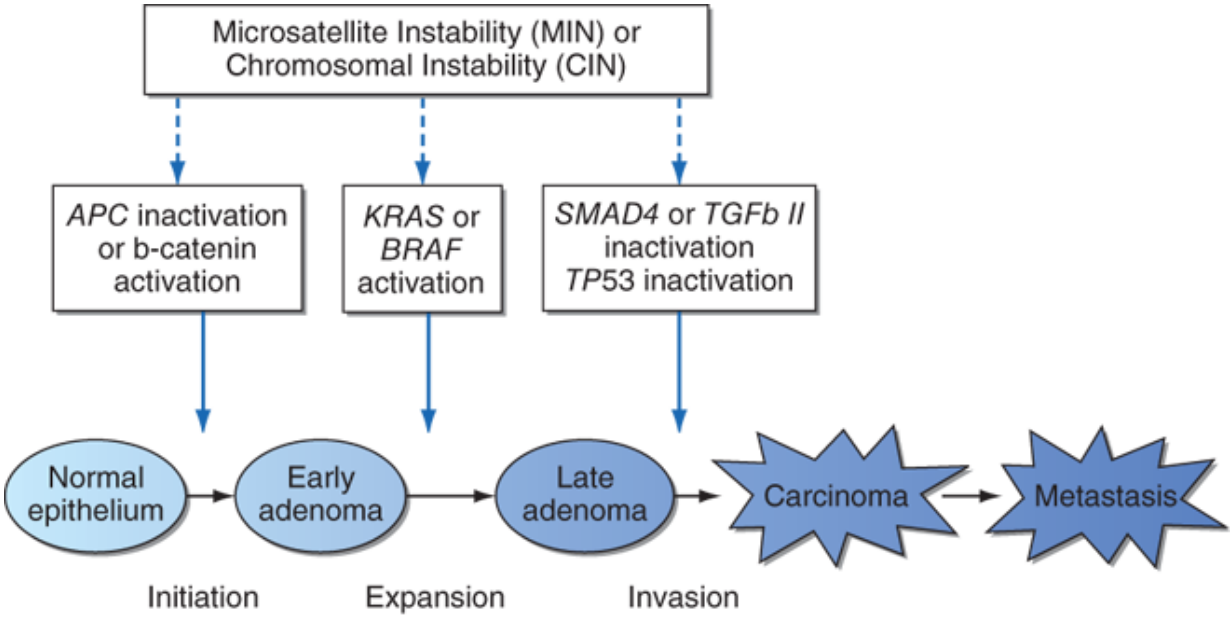


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

We now know the precise nature of the genetic alterations responsible for nearly all malignancies and are beginning to understand how these alterations promote the distinct stages of tumor growth. The prototypical example is colon cancer, in which analyses of genomes from the entire spectrum of neoplastic growths—from normal colon epithelium through adenoma to carcinoma—have identified mutations that are highly characteristic of each type of lesion **(Fig. 67-2)**.

FIGURE 67-2

Progressive somatic mutational steps in the development of colon carcinoma. The accumulation of alterations in a number of different genes results in the progression from normal epithelium through adenoma to full-blown carcinoma. Genetic instability (microsatellite or chromosomal) accelerates the progression by increasing the likelihood of mutation at each step. Patients with familial polyposis are already one step into this pathway, because they inherit a germline alteration of the *APC* gene. TGF, transforming growth factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TWO TYPES OF CANCER GENES: ONCOGENES AND TUMOR-SUPPRESSOR GENES

As briefly mentioned above, there are two major types of cancer genes. The first type comprises genes that positively influence growth and are known as *tumor-suppressor genes*. Both oncogenes and tumor-suppressor genes exert their effects on tumor growth through their ability to determine cell fates, influence cell survival and contribute to genome maintenance. The underlying molecular mechanisms can be extremely complex. While tightly regulated in normal cells, oncogenes acquire mutations that typically relieve this control and lead to increased activity of the gene products. This activating mutational event occurs in a single allele and acts in a dominant fashion. In contrast, the normal function of tumor-suppressor genes is usually to restrain cell growth, and this function is lost in cancer. Because of the diploid nature of mammalian cells, both alleles must be inactivated for a cell to completely lose the function of a tumor-suppressor gene. Thus, it requires two genetic events to inactivate a tumor-suppressor gene mutation, while only one genetic event is required to activate an oncogene.

A subset of tumor-suppressor genes controls the ability of the cell to maintain the integrity of its genome. Cells with a deficiency in these genes acquire an increased number of mutations throughout their genomes, including those in oncogenes and tumor-suppressor genes. This “mutator” phenotype was first hypothesized by Loeb to explain how the multiple rare mutational events required for tumorigenesis can occur in the lifetime of an individual. A mutator phenotype underlies several forms of cancer, such as those associated with deficiencies in DNA mismatch repair. The great majority of cancers do not harbor repair deficiencies, and their rate of mutation is similar to that observed in

normal cells. Many of these cancers, however, appear to harbor a different kind of genetic instability, affecting the loss or gains of whole chromosomes or large parts thereof (as explained in more detail below).

ONCOGENES IN HUMAN CANCER

Work by Peyton Rous in the early 1900s revealed that a chicken sarcoma could be transmitted from animal to animal in cell-free extracts, suggesting that cancer could be induced by an agent acting positively to promote tumor formation. The agent responsible for the transmission of the cancer was a retrovirus (Rous sarcoma virus, RSV) and the oncogene responsible was identified 75 years later as *V-SRC*. Other oncogenes were also discovered through their presence in the genomes of retroviruses that are capable of causing cancers in chickens, mice, and rats. The non-mutated cellular homologues of these viral genes are called proto-oncogenes and are often targets of mutation or aberrant regulation in human cancer. Whereas many oncogenes were discovered on the basis of their presence in retroviruses, other oncogenes, particularly those involved in translocations characteristic of particular leukemias and lymphomas, were identified through genomic approaches. Investigators cloned the sequences surrounding the chromosomal translocations observed cytogenetically and identified the genes activated at the breakpoints (see below). Some of these were oncogenes previously found in retroviruses (like *ABL*, involved in chronic myeloid leukemia [CML]), whereas others were new (like *BCL2*, involved in B-cell lymphoma). In the normal cellular environment, proto-oncogenes have crucial roles in cell proliferation and differentiation. Table 67-1 is a partial list of oncogenes known to be involved in human cancer.

TABLE 67-1

Oncogenes Commonly Altered in Human Cancers

Oncogene	Function	Alteration in Cancer	Neoplasm
AKT1	Serine/threonine kinase	Point mutation	Skin
BRAF	Serine/threonine kinase	Point mutation	Melanoma, thyroid, colorectal
CCND1	Cell cycle progression	Amplification	Esophageal, head and neck
CTNNB1	Signal transduction	Point mutation	Colon, liver, uterine, melanoma
EGFR	Signal transduction	Point mutation	Lung
FLT3	Signal transduction	Point mutation	AML
IDH1	Chromatin modification	Point mutation	Glioma
MDM2	Inhibitor of p53	Amplification	Sarcoma, glioma
MDM4	Inhibitor of p53	Amplification	Breast
MYC	Transcription factor	Amplification	Prostate, ovarian, breast, liver, pancreatic
MYCL1	Transcription factor	Amplification	Ovarian, bladder
MYCN	Transcription factor	Amplification	Neuroblastoma
PIK3CA	Phosphoinositol-3-kinase	Point Mutation	Multiple cancers
KRAS	GTPase	Point mutation	Pancreatic, colorectal, lung
NRAS	GTPase	Point mutation	Melanoma

Abbreviation: AML, acute myeloid leukemia.

The normal growth and differentiation of cells is controlled by growth factors that bind to receptors on the surface of the cell. The signals generated by the membrane receptors are transmitted inside the cells through signaling cascades involving kinases, G proteins, and other regulatory proteins. Ultimately, these signals affect the activity of transcription factors in the nucleus, which regulate the expression of genes crucial in cell proliferation, cell differentiation, and cell death. Oncogene products have been found to function at critical steps in these pathways (Chap. 68). Inappropriate activation of these pathways can lead to tumorigenesis.

MECHANISMS OF ONCOGENE ACTIVATION

POINT MUTATION

Point mutation (alternatively known as single nucleotide substitution) is a common mechanism of oncogene activation. For example, mutations in *KRAS* are present in > 95% of pancreatic cancers and 40% of colon cancers but are less common in other cancer types, although they can occur at significant frequencies in leukemia, lung, and thyroid cancers. Remarkably—and in contrast to the diversity of mutations found in tumor-suppressor genes—most of the activated *K*RAS alleles contain point mutations in codons 12, 13, or 61. These mutations reduce RAS GTPase activity, leading to constitutive activation of the mutant RAS protein. The restricted pattern of mutations observed in oncogenes compared to that of tumor-suppressor genes reflects the fact that gain-of-function mutations must occur at specific sites, while a broad variety of mutations can lead to loss of activity. Indeed, inactivation of a gene can in theory be accomplished through the introduction of a stop codon anywhere in the coding sequence, whereas activations require precise substitutions at residues that can somehow lead to an increase in the activity of the encoded protein under particular circumstances within the cell.

DNA AMPLIFICATION

The second mechanism for activation of oncogenes is DNA sequence amplification, leading to overexpression of the gene product. This increase in DNA copy number may cause cytologically recognizable chromosome alterations referred to as *homogeneous staining regions* (HSRs) if integrated within chromosomes, or *double minutes* (dmins) if extrachromosomal. The recognition of DNA amplification is accomplished through various DNA sequence-based methods for copy number analysis. With both microarray and sequencing technologies, the entire genome can be surveyed for gains and losses of DNA sequences, thus pinpointing chromosomal regions likely to contain genes important in the development or progression of cancer.

Numerous genes have been reported to be amplified in cancer. Several of these genes, including *NMYC* and *LMYC*, were identified through their presence within the amplified DNA sequences of a tumor and had homology to known oncogenes. Because the region amplified often includes hundreds of thousands of base pairs, multiple oncogenes may be amplified in a single amplicon in some cancers (particularly in sarcomas). Indeed, *MDM2*, *GLI*, *CDK4*, and *TPSPAN31* at chromosomal location 12q13-15 have been shown to be co-amplified in several types of sarcomas and other tumors. Amplification of a cellular gene is often a predictor of poor prognosis; for example, *ERBB2*/*HER2* and *NMYC* are often amplified in aggressive breast cancers and neuroblastoma, respectively.

CHROMOSOMAL REARRANGEMENT

Chromosomal alterations provide important clues to the genetic changes in cancer. The chromosomal alterations in human solid tumors such as carcinomas are heterogeneous and complex and occur as a result of the frequent chromosomal instability observed in these tumors (see below). In contrast, the chromosome alterations in myeloid and lymphoid tumors are often simple translocations, that is, reciprocal transfers of chromosome arms from one chromosome to another. The breakpoints of recurring chromosome abnormalities usually occur at the site of cellular oncogenes. Table 67-2 lists representative examples of recurring chromosome alterations in malignancy and the associated gene(s) rearranged or deregulated by the chromosomal rearrangement. Translocations are often observed in liquid tumors in general and are particularly common in lymphoid tumors, probably because these cell types have the capability to rearrange their DNA to generate antigen receptors. Indeed, antigen receptor genes are commonly

involved in the translocations, implying that an imperfect regulation of receptor gene rearrangement may be involved in their pathogenesis. In addition to transcription factors and signal transduction molecules, translocation may result in the overexpression of cell cycle regulatory proteins or proteins such as cyclins and of proteins that regulate cell death. Recurrent translocations have more recently been identified in solid tumors such as prostate cancers. Fusions between *TMPRSS2* and *ERG*, which are normally located in tandem on chromosome 21, contribute to about one-third of all prostate cancers and correlate with more aggressive disease.

TABLE 67-2

Representative Oncogenes at Chromosomal Translocations

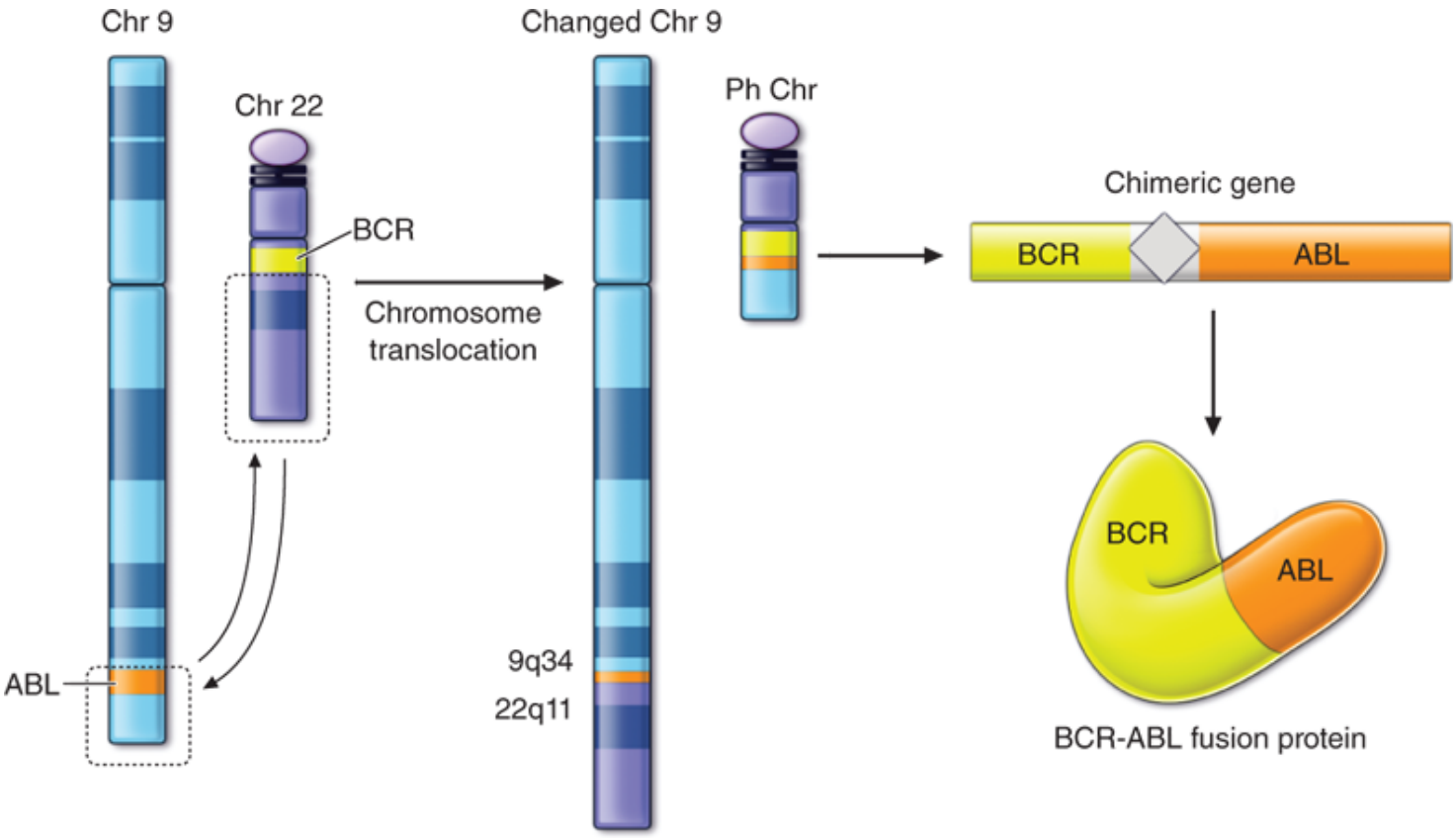
Gene (Chromosome)	Translocation	Malignancy
BCR-ABL	(9;22)(q34;q11)	Chronic myeloid leukemia
BCL1 (11q13.3)–*IgH* (14q32)	(11;14)(q13;q32)	Mantle cell lymphoma
BCL2 (18q21.3)–*IgH* (14q32)	(14;18)(q32;q21)	Follicular lymphoma
FLI-EWSR1	(11;22)(q24;q12)	Ewing's sarcoma
LCK – TCRB	(1;7)(p34;q35)	T-cell acute lymphocytic leukemia
PAX3– FOXO1	(2;13)(q35;q14)	Rhabdomyosarcoma
PAX8-PPARG	(2;3)(q13;p25)	Thyroid
IL21R-BCL6	(3;16)(q27;p11)	Non-Hodgkin's lymphoma
TAL1– TCTA	(1;3)(p34;p21)	Acute T cell leukemia
TMPRSS2-ERG	Rearrangement on Chr21q22	Prostate

The first reproducible chromosome abnormality detected in human malignancy was the Philadelphia chromosome detected in CML. This cytogenetic abnormality is generated by reciprocal translocation involving the *ABL* oncogene on chromosome 9, encoding a tyrosine kinase, being placed in proximity to the breakpoint cluster region (*BCR*) gene on chromosome 22. **Figure 67-3** illustrates the generation of the translocation and its protein product. The consequence of expression of the *BCR-ABL* gene product is the activation of signal transduction pathways leading to cell growth independent of normal external signals. Imatinib (marketed as Gleevec), a drug that specifically blocks the activity of Abl tyrosine kinase, has shown remarkable efficacy with little toxicity in patients with CML. The successful targeting of *BCR-ABL* by imatinib is the paradigm for molecularly targeted anti-cancer therapies.

FIGURE 67-3

Specific translocation seen in chronic myeloid leukemia (CML). The Philadelphia chromosome (Ph) is derived from a reciprocal translocation between chromosomes 9 and 22 with the breakpoint joining the sequences of the *ABL*

oncogene with the *BCR* gene. The fusion of these DNA sequences allows the generation of an entirely novel fusion protein with modified function.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CHROMOSOMAL INSTABILITY IN SOLID TUMORS

Solid tumors generally contain an abnormal number of chromosomes, a state known as aneuploidy; chromosomes from aneuploid tumors exhibit structural alterations such as translocations, deletions, and amplifications. These abnormalities reflect an underlying defect in cancer cells known as chromosomal instability. While aneuploidy is a striking cellular phenotype, chromosomal instability is manifest as only a small increase in the tendency of cells to gain, lose, or rearrange chromosomes during any given cell cycle. This intrinsically low rate of chromosome aberration implies that cancer cells become aneuploid only after many generations of clonal expansion. The molecular basis of aneuploidy remains incompletely understood. It is widely believed that defects in checkpoints, the quality-control mechanisms that halt the cell cycle if chromosomes are damaged or misaligned, contribute to chromosomal instability. This hypothesis emerged from experimental observations that the tumor suppressor p53 controls checkpoints that regulate the initiation of DNA replication and the onset of mitosis. These processes are therefore defective in many cancer cells. The mitotic spindle checkpoint, which ensures proper chromosome attachment to the mitotic spindle before allowing the sister chromatids to separate, is also altered in some cancers, irrespective of p53 status. The precise relationship between checkpoint deficiency and chromosomal instability remains unclear, but it is believed that even a subtle perturbation of the highly orchestrated process of cell division can impact the ability of a cell to faithfully replicate and segregate its complement of chromosomes. From a

therapeutic standpoint, the checkpoint defects that are prevalent in cancers have been proposed as vulnerabilities that may be exploited by novel agents and combinatorial strategies.

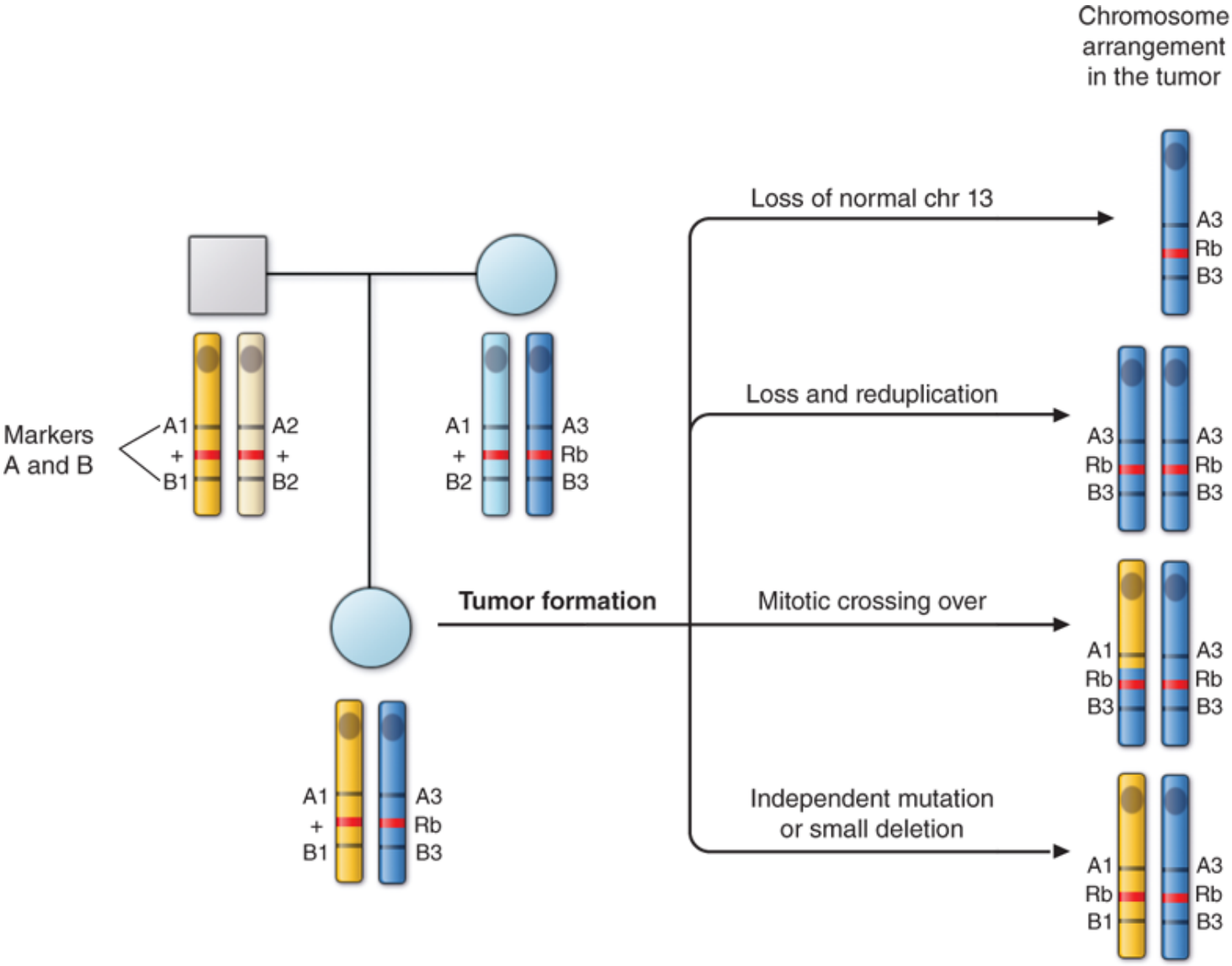
In contrast to the genome-wide cytogenetic changes that are typical indications of an underlying chromosomal instability, more focal patterns of chromosomal rearrangement have been recurrently detected in several cancer types. A curious phenomenon known as *chromothripsis* causes dozens of distinct breakpoints that are localized on one or several chromosomes. These striking structural alterations are thought to reflect a single event in which a chromosome is fragmented and then imprecisely reassembled. While the exact process that underlies chromothripsis remains obscure, and its effects on driver genes is not yet clear, a transient period of extreme instability stands in contrast to the gradual loss, gain and rearrangement of chromosomes that is typically observed in serially cultured cancer cells.

TUMOR-SUPPRESSOR GENE INACTIVATION IN CANCER

The first indication of the functional existence of tumor-suppressor genes came from experiments showing that fusion of mouse cancer cells with normal mouse fibroblasts led to a nonmalignant phenotype in the fused cells. The normal role of tumor-suppressor genes is to restrain cell growth, and the function of these genes is inactivated in cancer. The three major types of somatic lesions observed in tumor-suppressor genes during tumor development are *point mutations*, small insertions and/or deletions known as *indels*, and *large deletions*. Point mutations or indels in the coding region of tumor-suppressor genes will frequently lead to truncated protein products or allele-specific loss of RNA expression by the process of *nonsense-mediated decay*. Unlike the highly recurrent point mutations that are found in critical positions of activated oncogenes, known as mutational *hotspots*, the point mutations that cause tumor-suppressor gene inactivation tend to be distributed throughout the open reading frame. Large deletions lead to the loss of a functional product and sometimes encompass the entire gene or even the entire chromosome arm, leading to loss of heterozygosity (LOH) in the tumor DNA compared to the corresponding normal tissue DNA **(Fig. 67-4)**. LOH in tumor DNA often indicates the presence of a tumor-suppressor gene at a particular chromosomal location, and LOH studies have been useful in the positional cloning of many tumor-suppressor genes. The rate of LOH is increased in the presence of chromosomal instability, a relationship that would account for the high prevalence of aneuploidy in late-stage cancers.

FIGURE 67-4

Diagram of possible mechanisms for tumor formation in an individual with hereditary (familial) retinoblastoma. On the left is shown the pedigree of an affected individual who has inherited the abnormal (Rb) allele from her affected mother. The normal allele is shown as a (+). The four chromosomes of her two parents are drawn to indicate their origin. Flanking the retinoblastoma locus are genetic markers (A and B) also analyzed in this family. Markers A3 and B3 are on the chromosome carrying the retinoblastoma disease gene. Tumor formation results when the normal allele, which this patient inherited from her father, is inactivated. On the right are shown four possible ways in which this could occur. In each case, the resulting chromosome 13 arrangement is shown. Note that in the first three situations, the normal allele (B1) has been lost in the tumor tissue, which is referred to as loss of heterozygosity (LOH) at this locus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Gene silencing, an epigenetic change that leads to the loss of gene expression, occurs in conjunction with hypermethylation of the promoter and histone deacetylation, and is another mechanism of tumor-suppressor gene inactivation. An *epigenetic modification* refers to a covalent modification of chromatin, heritable by cell progeny that may involve DNA but does not involve a change in the DNA sequence. The inactivation of the second X chromosome in female cells is an example of an epigenetic silencing that prevents gene expression from the inactivated chromosome. Genomic regions of hypermethylated and hypomethylated DNA can be detected by specialized techniques, and a subset of these regional modifications has consequences on the cell's behavior.

FAMILIAL CANCER SYNDROMES

A small fraction of cancers occurs in patients with a genetic predisposition. Based on studies of inherited and sporadic forms of retinoblastoma, Knudson and others formulated a hypothesis that explains the differences between sporadic and inherited forms of the same tumor type. In inherited forms of cancer, called *cancer predisposition syndromes*, one allele of a particular tumor suppressor gene is inherited in mutant form. This germline mutation is not sufficient to initiate a tumor, however; the other allele, inherited from the unaffected parent, must

become somatically mutated in a normal stem cell for tumorigenesis to be initiated. In sporadic (non-inherited) forms of the same disease, all cells in the body start out with two normal copies of the tumor suppressor gene. A single cell must then sequentially acquire mutations in both alleles of the tumor suppressor gene to initiate a tumor. Thus bi-allelic mutations of the same tumor suppressor gene are required for both inherited and non-inherited forms of the disease; the only difference is that individuals with the inherited form have a "head-start": they already have one allele mutated, from conception, and only need one additional mutation to initiate the process (Fig. 67-4). This distinction explains why those with inherited forms of the disease develop more cancers, at an earlier age, than the general population. It also explains why, even though every cell in an individual with a cancer predisposition syndrome has a mutant gene, only a relatively small number of tumors arise during his/her lifetime. The reason is that the vast majority of cells within such individuals are functionally normal because one of the two alleles of the tumor suppressor gene is normal. Mutations are uncommon events, and only the rare cells that develop a mutation in the remaining normal allele will exhibit uncontrolled proliferation. The same principle applies to virtually all types of cancer predisposition syndromes, though the particular genes differ. For example, inherited mutations in *RB1*, *WT1*, *VHL*, *APC*, and *BRCA1* lead to predispositions to retinoblastomas, Wilms' tumors, renal cell carcinomas, colorectal carcinomas, and breast carcinomas, respectively **(Table 67-3)**. Also note that the biallelic inactivation of any of these genes is not sufficient to develop cancer; it requires other, additional somatic mutations for the initiating cells to evolve to malignancy, as noted above.

Roughly 100 familial cancer syndromes have been reported; the great majority are very rare. Most of these syndromes exhibit an autosomal dominant pattern of inheritance, although some of those associated with DNA repair abnormalities (xeroderma pigmentosum, Fanconi's anemia, ataxia telangiectasia) are inherited in an autosomal recessive fashion. **Table 67-3** shows a number of cancer predisposition syndromes and the responsible genes.

TABLE 67-3

Cancer Predisposition Syndromes and Associated Genes

Syndrome	Gene	Chromosome	Inheritance	Tumors
Ataxia telangiectasia	*ATM*	11q22-q23	AR	Breast
Autoimmune lymphoproliferative syndrome	*FAS* *FASL*	10q24 1q23	AD	Lymphomas
Bloom's syndrome	*BLM*	15q26.1	AR	Various
Cowden's syndrome	*PTEN*	10q23	AD	Breast, thyroid
Familial adenomatous polyposis	*APC* *MUTYH*	5q21 1p34.1	AD AR	Colorectal (early onset)
Familial melanoma	*CDKN2A*	9p21	AD	Melanoma, pancreatic
Familial Wilms' tumor	*WT1*	11p13	AD	Kidney (pediatric)
Hereditary breast/ovarian cancer	*BRCA1* *BRCA2*	17q21 13q12.3	AD	Breast, ovarian, prostate
Hereditary diffuse gastric cancer	*CDH1*	16q22	AD	Stomach
Hereditary multiple exostoses	*EXT1* *EXT2*	8q24 11p11-12	AD	Exostoses, chondrosarcoma
Hereditary retinoblastoma	*RB1*	13q14.2	AD	Retinoblastoma, osteosarcoma
Hereditary nonpolyposis colon cancer (HNPCC)	*MSH2* *MLH1* *MSH6* *PMS2*	2p16 3p21.3 2p16 7p22	AD	Colon, endometrial, ovarian, stomach, small bowel, ureter carcinoma
Hereditary papillary renal carcinoma	*MET*	7q31	AD	Papillary kidney
Juvenile polyposis syndrome	*SMAD4* *BMPR1A*	18q21	AD	Gastrointestinal, pancreatic
Li-Fraumeni syndrome	*TP53*	17p13.1	AD	Sarcoma, breast

Syndrome	Gene	Chromosome	Inheritance	Tumors
Multiple endocrine neoplasia type 1	*MEN1*	11q13	AD	Parathyroid, endocrine, pancreas, and pituitary
Multiple endocrine neoplasia type 2a	*RET*	10q11.2	AD	Medullary thyroid carcinoma, pheochromocytoma
Neurofibromatosis type 1	*NF1*	17q11.2	AD	Neurofibroma, neurofibrosarcoma, brain
Neurofibromatosis type 2	*NF2*	22q12.2	AD	Vestibular schwannoma, meningioma, spine
Nevoid basal cell carcinoma syndrome (Gorlin's syndrome)	*PTCH1*	9q22.3	AD	Basal cell carcinoma, medulloblastoma, jaw cysts
Tuberous sclerosis	*TSC1* *TSC2*	9q34 16p13.3	AD	Angiofibroma, renal angiomyolipoma
von Hippel–Lindau disease	*VHL*	3p25-26	AD	Kidney, cerebellum, pheochromocytoma

Abbreviations: AD, autosomal dominant; AR, autosomal recessive.

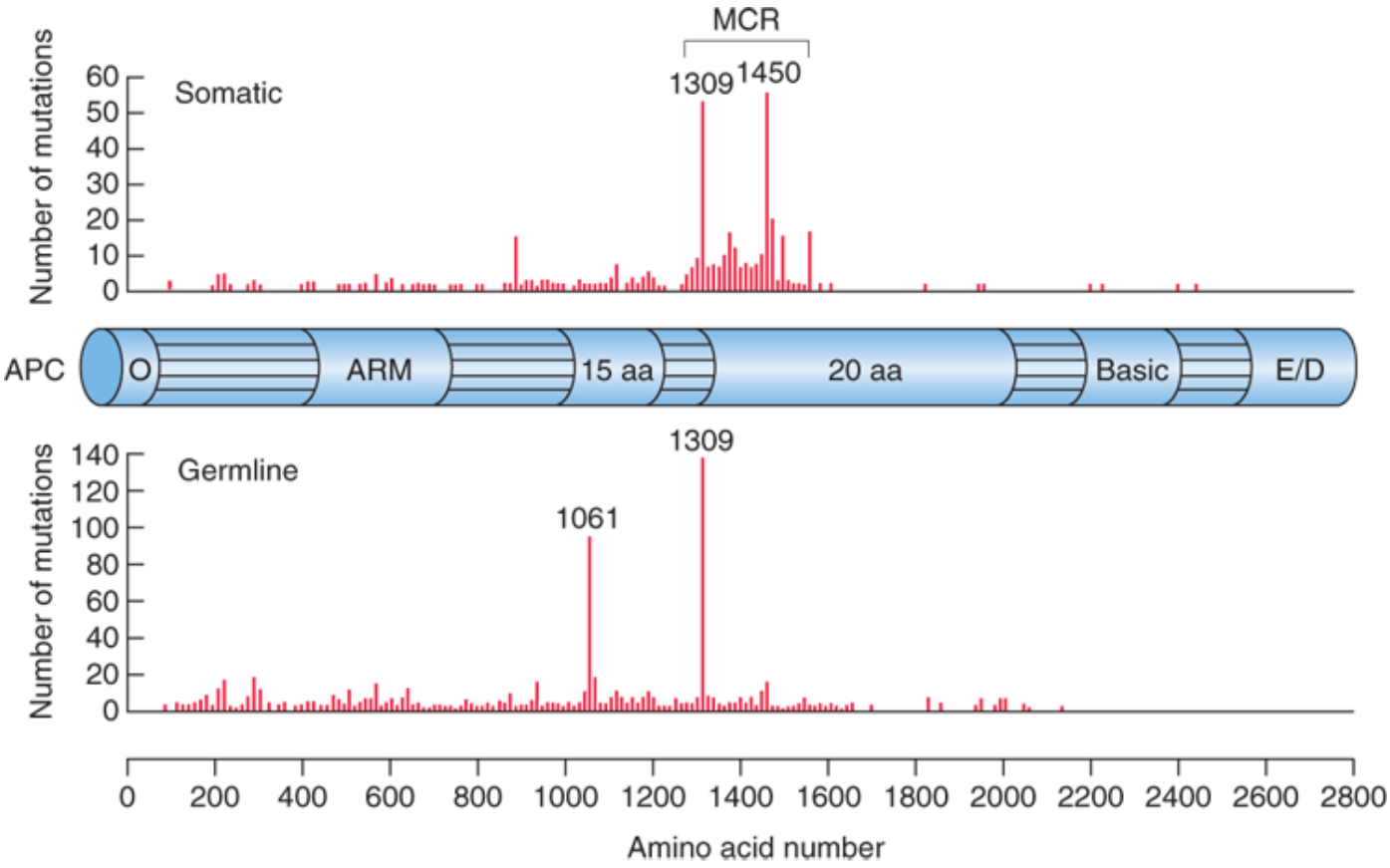
The next section examines inherited colon cancer predispositions in detail because several lessons of general importance have been derived from the study of these syndromes.

Familial adenomatous polyposis (FAP) is a dominantly inherited colon cancer syndrome caused by germline mutations in the adenomatous polyposis coli (*APC*) tumor-suppressor gene on chromosome 5. Affected individuals develop hundreds to thousands of adenomas in the colon. In each of these adenomas, the *APC* allele inherited from the affected parent has been inactivated by virtue of a somatic mutation (Fig. 67-2). This inactivation usually occurs through a gross chromosomal event resulting in loss of all or a large part of the long arm of chromosome 5, where *APC* resides. In other cases, the remaining allele is inactivated by a subtle intragenic mutation of *APC*, which as a single base substitution resulting in a nonsense codon. Gross chromosomal losses occur more commonly than point mutations in normal cells, explaining why these are the predominant mechanism underlying the inactivation of the normal allele of *APC*. The same is true for other cancer predisposition syndromes caused by other inherited tumor suppressor gene mutations; gross chromosomal events are generally responsible for inactivation of the tumor suppressor gene allele inherited from the non-affected parent. Several thousand adenomas form in FAP patients, and a small subset of the billions of cells within these adenomas will acquire a second mutation, leading to tumor progression, that is, a larger adenoma. A third mutation in such a larger adenoma may convert it to a carcinoma. If untreated (by colectomy), at least one of the adenomas will progress to cancer by the time patients are in their mid-40s. *APC* can be considered to be a gatekeeper for colon tumorigenesis in that the absence of mutation of this gatekeeper (or a gene acting within the same pathway), a colorectal tumor simply cannot be initiated. **Figure 67-5**

shows the germline and somatic mutations found in the *APC* gene. A negative regulator of a signaling pathway that determines cell fate during development, the APC protein provides differentiation and apoptotic cues to colonic epithelial cells as they migrate up the crypts. Defects in this process can lead to abnormal accumulation of cells that would otherwise differentiate and eventually undergo apoptosis.

FIGURE 67-5

Germline and somatic mutations in the tumor-suppressor gene adenomatous polyposis coli (*APC*). *APC* encodes a 2843-amino-acid protein with six major domains: an oligomerization region (O), armadillo repeats (ARM), 15-amino-acid repeats (15 aa), 20-amino-acid repeats (20 aa), a basic region, and a domain involved in binding EB1 and the *Drosophila* discs large homologue (E/D). Shown are 650 somatic and 826 germline mutations representative of the mutations that occur within the *APC* gene (from the *APC* database at *www.umd.be/APC*). All known pathogenic mutations of *APC* result in the truncation of the APC protein. Germline mutations are found to be relatively evenly distributed up to codon 1600 except for two mutation hotspots surrounding amino acids 1061 and 1309, which together account for one-third of the mutations found in familial adenomatous polyposis (FAP) families.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In contrast to patients with FAP, patients with hereditary nonpolyposis colon cancer (HNPCC, or Lynch's syndrome) do not develop polyposis, but instead develop only one or a small number of adenomas that rapidly progress to cancer. HNPCC is due to inherited mutations in one of four DNA mismatch repair genes (Table 67-3) that are components of a repair system responsible for correcting errors in newly replicated DNA. Germline mutations in *MSH2* and *MLH1* account for more than 90% of HNPCC cases, and mutations in *MSH6* and *PMS2* account for the remainder. When a somatic mutation inactivates the remaining wild-type allele of a mismatch repair gene, the cell develops a hypermutable phenotype characterized by profound genomic instability that is most readily apparent in short repeated sequences called *microsatellites* and is sometimes called microsatellite instability (MSI). The high rate of

mutation in such cells impacts all genes, including oncogenes and tumor suppressor genes, and thereby accelerates the activation of the former and the inactivation of the latter (Fig. 67-2). HNPCC can be considered a disease of tumor progression; once tumors are initiated (by an inactivating mutation of *APC* or by some other gene in the APC pathway), tumors rapidly progress because of the accelerated mutation rate. Progression from a tiny adenoma to carcinoma takes only a few years in HNPCC patients instead of the two or three decades this progression takes in patients with FAP (or in patients with sporadic colorectal tumors). Approximately half of HNPCC patients develop colorectal cancers by the time they are in their mid-40s—similar to that of FAP patients. This coincidence in age of onset emphasizes that both tumor initiation (abnormal in FAP patients) and tumor progression (abnormal in HNPCC patients) are the two pillars of cancer development and are equally important for cancer development.

Another general principle is apparent from the comparison between FAP and HNPCC patients. The tumors in FAP patients, like those in patients without hereditary predisposition to cancers, are chromosomal instability. MSI and chromosomal instability appear to be mutually exclusive in colon cancers, suggesting that they represent alternative mechanisms for the generation of genomic instability (Fig. 67-2). Other cancer types rarely exhibit MSI. Chromosomal instability is far more prevalent than MSI among all cancer types, perhaps explaining why nearly all cancers are aneuploid.

Although most autosomal dominant inherited cancer syndromes are due to mutations in tumor-suppressor genes (Table 67-3), there are a few interesting exceptions. Multiple endocrine neoplasia type 2, a dominant disorder characterized by pituitary adenomas, medullary carcinoma of the thyroid, and (in some pedigrees) pheochromocytoma, is due to gain-of-function mutations in the proto-oncogene *RET* on chromosome 10. Similarly, gain-of-function mutations in the tyrosine kinase domain of the *MET* oncogene lead to hereditary papillary renal carcinoma. Interestingly, loss-of-function mutations in the *RET* gene cause a completely different disease, Hirschsprung's disease (aganglionic megacolon **[Chaps. 321 and 381]**).

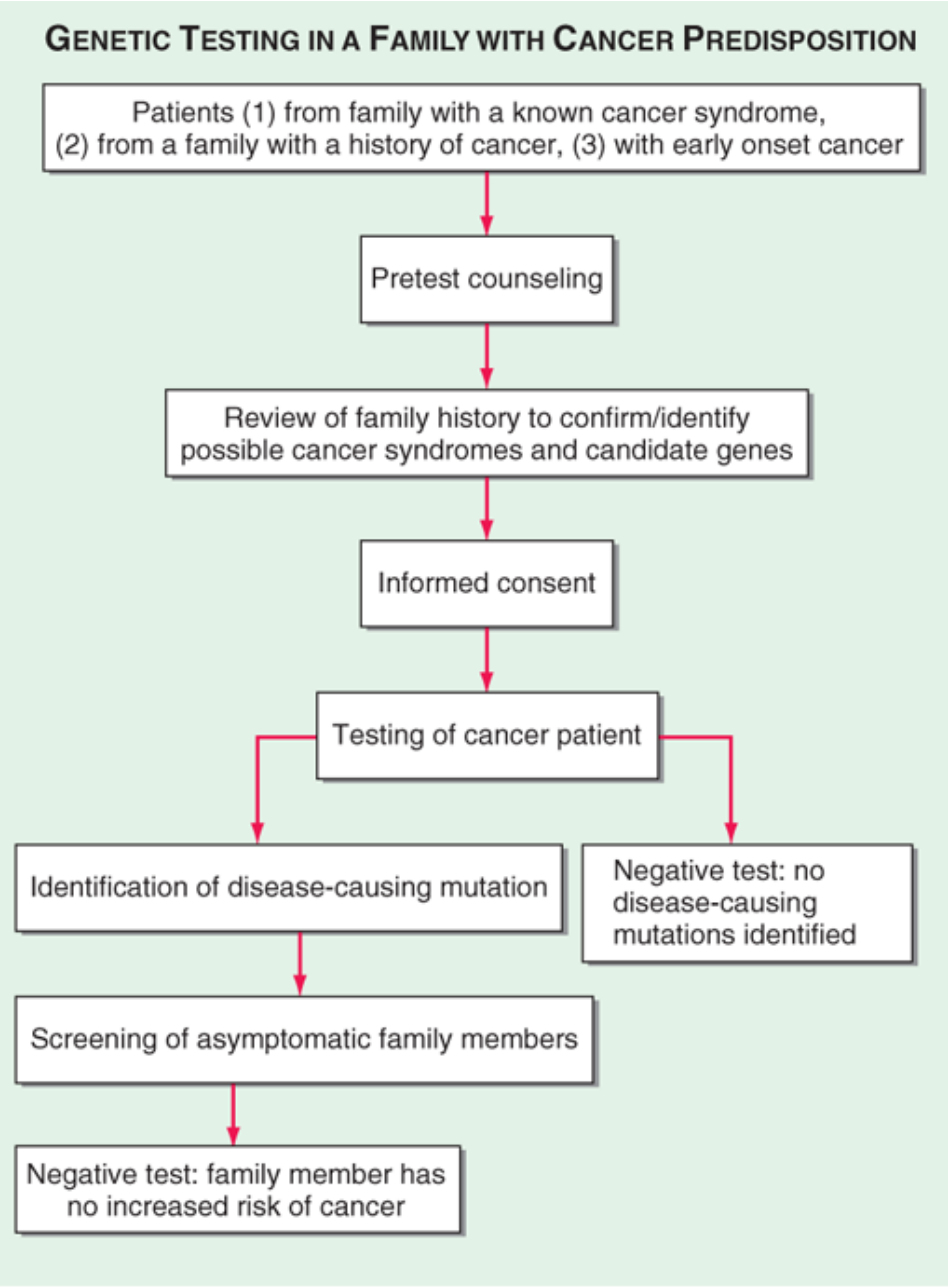
Although the heritable forms of cancer have taught us much about the mechanisms of growth control, most forms of cancer do not follow simple Mendelian patterns of inheritance. The majority of human cancers arise in a sporadic fashion, solely as a result of somatic mutation, and in the absence of any mutations in cancer-predisposing genes in their germlines.

GENETIC TESTING FOR FAMILIAL CANCER

The discovery of cancer susceptibility genes raises the possibility of DNA testing to predict the risk of cancer in individuals of affected families. An algorithm for cancer risk assessment and decision making in high-risk families using genetic testing is shown in **Fig. 67-6**. Once a mutation is discovered in a family, subsequent testing of asymptomatic family members can be crucial in patient management. A negative gene test in these individuals can prevent years of anxiety in the knowledge that their cancer risk is no higher than that of the general population. On the other hand, a positive test may lead to alteration of clinical management, such as increased frequency of cancer screening and, when feasible and appropriate, prophylactic surgery. Potential negative consequences of a positive test result include psychological distress (anxiety, depression) and discrimination, although the Genetic Information Nondiscrimination Act (GINA) makes it illegal for predictive genetic information to be used to discriminate in health insurance or employment. Testing should therefore not be conducted without counseling before and after disclosure of the test result.

FIGURE 67-6

Algorithm for genetic testing in a family with cancer predisposition. The key step is the identification of a disease-mutation in a cancer patient, which is an indication for the testing of asymptomatic family members. Asymptomatic family members who test positive may require increased screening or surgery, whereas those who test negative are at no greater risk for cancer than the general population. It should be emphasized that no molecular assay used for this sort of testing is 100% sensitive; negative results must be interpreted with this caveat in mind.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Recent technological developments have made it feasible to obtain high-quality sequence of all of the protein-coding DNA sequences, and even of the entire genome, in any given individual. The redundant nature of modern DNA sequencing provides an extremely high level of sensitivity, such that mutations and polymorphisms will inevitably be identified in every subject. In patients lacking a clear family history, the significance of these DNA sequence findings will not be apparent. Even mutations in tumor suppressor genes are difficult to interpret unless there is an obvious functional implication, such as the truncation of the open reading frame, or that particular mutation has previously

been associated with cancer. Such germline mutations are very rare in the general population. Vastly more common are *variants of unknown significance* (*VUS*). VUS that are found during genetic testing cannot be used to evaluate the relative risk of cancer, but may nonetheless cause anxiety because they represent a deviation from the reference allele that is established as "normal." Because of the low yield of informative mutations that modify cancer risk and the frequent identification of VUS, it is generally not appropriate to use DNA sequencing to assess cancer risk in individuals unless the family history is suggestive of a germline mutation. Conversely, testing may be appropriate in some subpopulations with a known increased risk, even without a defined family history. For example, two mutations in the breast cancer susceptibility gene *BRCA1*, 185delAG and 5382insC, exhibit a sufficiently high frequency in the Ashkenazi Jewish population that genetic testing based on ethnicity alone may be warranted.

It is important that genetic test results be communicated to families by trained genetic counselors, especially for high-risk high-penetrance conditions such as the hereditary breast and ovarian cancer syndrome (*BRCA1/BRCA2*). To ensure that the families clearly understand its advantages and disadvantages and the impact it may have on disease management and psyche, genetic testing should never be done before counseling. Significant expertise is needed to communicate the results of genetic testing to families.

VIRUSES IN HUMAN CANCER

Several human malignancies are associated with viruses. Examples include Burkitt's lymphoma (Epstein-Barr virus; **Chap. 189**), hepatocellular carcinoma (hepatitis viruses), cervical cancer (human papillomavirus [HPV]; **Chap. 193**), and T cell leukemia (retroviruses; **Chap. 196**). There are several types of HPV, including the high-risk types 16 and 18 that are strongly associated with the development of cervical, vulvar, vaginal, penile, anal, and oropharyngeal cancer. The mechanisms of action of all these viruses involve inactivation of tumor suppressor genes. For example, HPV proteins E6 and E7 bind to and inactivate cellular tumor suppressors p53 and pRB, respectively. This is the reason that HPV is such a potent initiator of cancer: infection with a virus is tantamount to having two of the three mutant driver genes required for cancer, that is, one viral oncogene inactivates p53 and the other inactivates Rb. Though these two inactivated gene products are not sufficient for tumorigenesis, only one additional mutant gene is required to develop a malignancy.

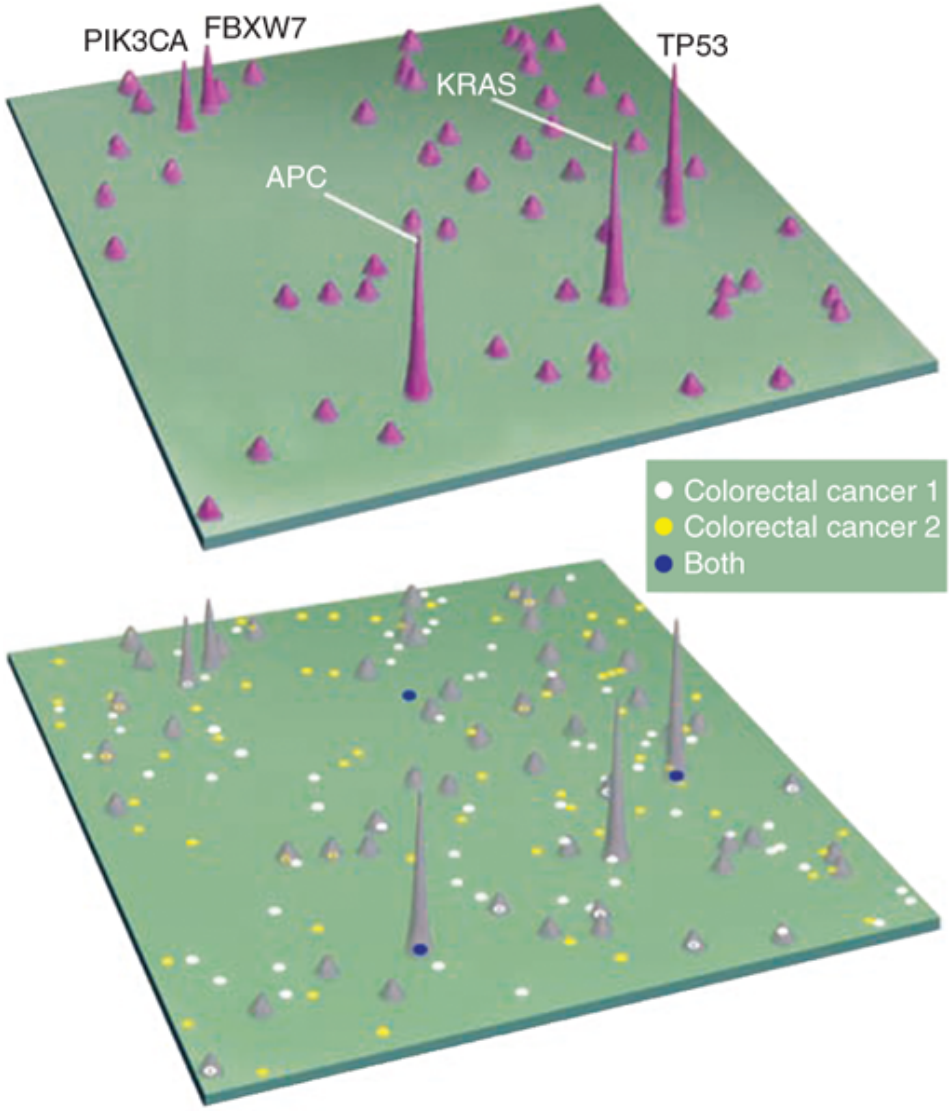
CANCER GENOMES

The advent of relatively inexpensive technologies for rapid and high-throughput DNA sequencing has facilitated the comprehensive analysis of numerous genomes from many types of tumors. This unprecedented view into the genetic nature of cancer has provided remarkable insights. Most cancers do not arise in the context of a mutator phenotype, and accordingly the number of mutations in even the most advanced cancers is relatively modest. Common solid tumors harbor 30–70 subtle mutations that are non-synonymous (i.e., result in an amino acid change in the encoded protein). Liquid tumors such as lymphomas and leukemias, as well as pediatric tumors, typically have fewer than 20 mutations. The vast majority of the mutations detected in tumors are not functionally significant, they simply arose by chance in a single cell that gave rise to an expanding clone. Such mutations, which provide no selective advantage to the cell in which they occur, are known as *passenger* mutations. As noted above, only a small number of the mutations confer a selective growth advantage and thereby promote tumorigenesis. These functional mutations are known as *driver* mutations, and the genes in which they occur are called driver genes.

The frequency and distribution of driver mutations within a single tumor type can be represented as a topographical landscape **(Fig. 67-7)**. The picture that emerges from these studies reveals that most genes that are mutated in tumors are actually mutated at relatively low frequencies, as would be expected of passenger genes, whereas a small number of genes (the driver genes) are mutated in a large proportion of tumors. There are a total of ~200 driver genes that are responsible for the development of all solid tumors, representing only ~1% of the total number of human protein-encoding genes. The majority of the mutations in these driver genes provide a direct selective growth advantage by altering the signaling pathways that mediate cell survival or the determination of cell fate. The remaining driver gene mutations indirectly provide a selective growth advantage by accelerating the mutation rate of proto-oncogenes and tumor suppressor genes. The functions of all these driver genes can be organized into a dozen signaling pathways, as shown in **Table 67-4**.

FIGURE 67-7

The mutational topography of colorectal cancer. The two-dimensional landscape represents the positions of the individual genes along the chromosomes. The height of each peak represents the mutation frequency at that locus. The top map is a representation of many sequenced colorectal cancers. The taller peaks represent the genes that are commonly mutated in colon cancer, while the smaller hills indicate the genes that are mutated at lower frequency. On the lower map, the mutations of two individual tumors are indicated. Note that there is little overlap between the mutated genes of the two colorectal tumors shown. These differences represent Type I heterogeneity, as noted in the test, which is the foundation for personalized medicine in cancer. *(From LD Wood et al: Science 318:1108, 2007, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 67-4

Signaling Pathways Altered in Cancer

Process	Pathway	Representative Driver Genes
Cell survival	Cell cycle regulation/apoptosis	*RB1, BCL2*
	RAS	*KRAS, BRAF*
	PIK3CA	*PTEN, PIK3CA*
	JAK/STAT	*JAK2, FLT3*
	MAPK	*MAP3K. ERK*
	TGF-β	*BMPR1A, SMAD4*
Cell fate	Notch	*NOTCH1, FBWX7*
	Hedgehog	*PTCH1, SMO*
	WNT/APC	*APC, CTNNB1*
	Chromatin modification	*DNMT1, IDH1*
	Transcriptional regulation	*AR, KLF4*
Genome maintenance	DNA damage signaling and repair	*ATM BRCA1*

TUMOR HETEROGENEITY

The mutant cells that compose a single tumor are not genetically identical. Rather, cells obtained from different sites on a tumor will harbor common mutations as well as mutations that are unique to each sample. Genetic heterogeneity results from the ongoing acquisition of mutations during tumor growth. Each time a genome is replicated, there is a small but quantifiable probability that a mutation will spontaneously arise as a result of a replication error and be passed on to the cellular progeny. This is true in normal cells or in tumor cells. Any randomly chosen cell from the skin of one individual will harbor hundreds of genetic alterations that distinguish it from a different randomly chosen skin cell, and the same is true for all organs of self-renewing tissues. Tumors are actually *less* genetically heterogeneous than normal cells; any two randomly chosen cells from a tumor of an individual will have fewer differences than any two randomly chosen cells from that individual's normal tissues. The reason for this decrease in heterogeneity is clonal expansion, the fundamental feature of tumorigenesis. Every time a clonal expansion occurs, a genetic bottleneck wipes out heterogeneity among the cells that didn't expand; these unexpanded cells either die or form only a minute proportion of the total cells in the expanding tumor.

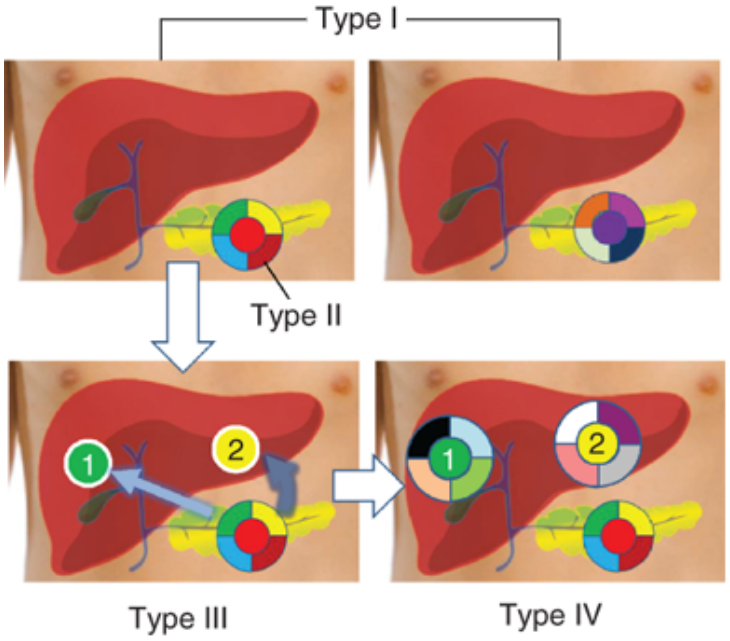
The mutations that vary between cells of a given tumor are invariably passenger mutations that arose since the last evolutionary bottleneck, that is, those mutations that arose during the expansion of the founder cell that gave rise to the final clonal expansion. In contrast, the passenger mutations that were present in the founder cell will be uniformly present in every cell in the tumor. In that respect, these passenger mutations that are not heterogeneously distributed, that is., those that are present in every cancer cell, are like the driver gene mutations, which are also present in virtually all cancer cells. The total number of mutations and their distribution within tumor cells therefore represents a complex interplay between the age of the patient (the older the patient, the more passenger mutations will have accumulated in the founding cell of the first clonal expansion) and the evolutionary history of the cancer (its age and number of clonal expansions it experienced).

Tumor heterogeneity has been recognized for decades at the cytogenetic, biochemical, and histopathologic levels. However, it is only recently, with the advent of a deep understanding of cancer genetics that genetic heterogeneity can be interpreted in a medically relevant fashion. The first important point to recognize about tumor heterogeneity is that it is only the variation in driver gene alterations that is important; the cellular distribution of passenger gene mutations is completely irrelevant. In this discussion of heterogeneity, we can expand the definition of “driver genes” to include those that provide a selective growth advantage in the face of therapy in addition to those that provide a selective growth advantage during tumor evolution, prior to treatment.

Type I heterogeneity refers to that among tumors of the same type from different patients **(Fig. 67-8)**. Though adenocarcinomas of the lung generally harbor mutations in three or more driver genes, the genes differ among the patients and the precise mutations within the same gene can vary considerably. Type I heterogeneity is the basis for precision medicine, where the goal is to treat patients with drugs that target the proteins encoded by genetic alterations within their specific tumors. Type 2 heterogeneity refers to the genetic heterogeneity among different cells from the same primary tumor. Tumors continue to evolve as they grow, and different cells of the same cancer, in its original site (e.g., the colon), may acquire another driver gene mutations that are not shared among the other cells of the tumor. Such a mutation can result in a small clonal expansion that may or may not be important biologically. In cases in which the primary tumor can be surgically excised, such mutations are unimportant unless they give rise to Type III heterogeneity (described below). The reason they are important is because all primary tumor cells, whether homogeneous or not, are removed by the surgical procedure. In primary tumors that cannot be completely excised (such as most advanced brain tumors and many pancreatic ductal adenocarcinomas), heterogeneity is biomedically important because it can give rise to drug resistance, analogously to that described for Type IV heterogeneity (see below). Type III heterogeneity refers to the genetic differences among the founder cells of the metastatic lesions from the same patient. For example, a patient with melanoma may have 100 different metastases distributed throughout various organs. Only if a mutant *BRAF* is present in every founder cell of every metastasis, then the patient has a chance at a complete response to a *BRAF* inhibitor. There have been several recent detailed studies of the metastases from various tumor types. Fortunately, these studies suggest there is very little, if any, Type III heterogeneity among driver genes, a necessary prerequisite for the successful implementation of future targeted therapies. Finally, Type IV heterogeneity refers to that among cells of individual metastatic lesions. As the founder cell of each metastasis expands to become detectable, it acquires mutations, a small number of which can act as “drivers” if the patient is exposed to therapeutics. This type of heterogeneity is of major clinical importance, as it has been shown to be responsible for the development of resistance in virtually all targeted therapies. The development of such resistance is a fait accompli based simply on known mutation rates and known genetic resistance mechanisms. The only way to circumvent acquired resistance is to treat metastatic tumors earlier (i.e., in adjuvant setting, before much tumor expansion has occurred) or to treat with combinations of drugs for which cross-resistance is genetically impossible.

FIGURE 67-8

The four types of tumor heterogeneity. Tumor heterogeneity is the inevitable result of cell proliferation, as new mutations are introduced during clonal expansion. This concept is illustrated by a primary tumor in the pancreas and two metastatic tumors in the liver. The tumors of the founding populations are shown in the middle of each circle, while the distinct subclones are shown around the periphery. Type I: the heterogeneity of tumors that occur among different patients. Type II: the heterogeneity among the cells of a primary tumor, also known as intratumoral heterogeneity. Type III: the heterogeneity among the founding cells of distinct metastatic lesions (marked as 1 and 2) that arise in the same patient, also known as intermetastatic heterogeneity. Type IV: heterogeneity among the cells of each metastasis that develops as each tumor grows, also known as intrametastatic heterogeneity.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PERSONALIZED CANCER DETECTION AND TREATMENT

High-throughput DNA sequencing has led to an unprecedented understanding of cancer at the molecular level. A comprehensive mutation profile provides a molecular history of a given tumor and insights into how it arose. Because tumor cells and tumor DNA are shed into the blood and other bodily fluids, common driver mutations can be used as highly specific biomarkers for early detection. For diagnosed tumors, tumor-specific mutations can be used to estimate tumor burden, to assess treatment responses and to detect recurrence.

In some cases, information regarding specific genes and pathways that are altered provides patients and physicians with options for personalized therapy. This general approach is sometimes referred to as *precision medicine*. Because tumor behavior is highly variable, even within a tumor type, personalized information-based medicine can supplement and perhaps eventually supplant histology-based tumor assessment, especially in the case of tumors that are resistant to conventional therapeutic approaches. Conversely, molecular nosology has revealed similarities in tumors of diverse histotype. The success of the precision medicine approach in any given patient depends on the presence of tumor-associated genetic alterations that are actionable (i.e., can be targeted with a specific drug). Examples of currently actionable changes include mutations in *BRAF* (targeted by the drug vemurafenib) and *RET* (targeted by sunitinib and sorafenib), and *ALK* rearrangements (targeted by crizotinib). At present, the proportion of

tumors that can be treated with such precision medicine approaches is small, but future therapeutic development will hopefully change this situation. The development of new targeted agents is at present hindered by the fact that such agents can only target activated oncogenes, while the great majority of genetic alterations in common solid tumors are those that inactivate tumor suppressor genes. Because all drugs, whether for use in oncology or any other purpose, can only inhibit protein actions, drugs cannot be used to directly target the proteins encoded by inactivated tumor suppressor genes; these proteins are already inactive. More information about the pathways through which tumor suppressor genes act may provide a way around this obstacle. For example, when a tumor suppressor gene is inactivated, some downstream component of the pathway is likely to be activated, thereby presenting a realistic target. An example of this is provided by PARP-1 inhibitors, which have been successfully used to treat patients whose tumors have inactivating mutations of genes involved in DNA repair processes, such as *BRCA1*. Patterns of global gene expression can be used to help unravel such pathways and are already being used to predict drug sensitivities and provide prognostic information in addition to that provided by DNA sequence analysis. Evaluation of proteomic and metabolomics patterns may also prove useful.

THE FUTURE

A revolution in cancer genetics has occurred in the past 30 years. Most types of cancer are now understood at the DNA sequence level and this accomplishment has led us to an increasingly refined understanding of tumorigenesis. Cancer gene mutations have proven to be reliable biomarkers for cancer detection and monitoring as well as for informing therapeutics through precision medicine approaches. Gene-based tests are already standard of care for certain tumor types, such as melanoma, colorectal and pancreatic cancers, and the utility of these tests will undoubtedly be expanding greatly in the coming years as new therapies and ways of predicting responses to therapies are developed. While effective treatment of advanced cancers remains difficult, it is expected that breakthroughs in these areas will continue to emerge and be applicable to an ever-increasing number of cancers. Moreover, with the hoped-for advances in diagnostics, particularly in the earlier detection of cancers, the new and old therapies for cancer can be expected to have a much greater impact on reducing cancer deaths.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the past contributions of Pat J. Morin, Jeff Trent, and Francis Collins to earlier versions of this chapter.

FURTHER READING

Bunz F: *Principles of Cancer Genetics*, 2nd ed. Dordrecht, Springer, 2016.

Simon R, Roychowdhury S: Implementing personalized cancer genomics in clinical trials. Nat Rev Drug Disc 12:358, 2013.

Vogelstein B et al: Cancer genome landscapes. Science 339:1546, 2013. [PubMed: 23539594]

Vogelstein B, Kinzler KW: The path to cancer—three strikes and you're out. N Engl J Med 373:1895, 2015. [PubMed: 26559569]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 68: Cancer Cell Biology

Jeffrey W. Clark; Dan L. Longo

CANCER CELL BIOLOGY

Cancers are characterized by unregulated cell division, avoidance of cell death, tissue invasion, and the ability to metastasize. A neoplasm is *benign* when it grows in an unregulated fashion without tissue invasion. The presence of unregulated growth and tissue invasion is characteristic of *malignant* neoplasms. Cancers are named based on their origin: those derived from epithelial tissue are called *carcinomas*, those derived from mesenchymal tissues are *sarcomas*, and those derived from hematopoietic tissue are *leukemias*, *lymphomas,* and *plasma cell dyscrasias* (including *multiple myeloma*).

Cancers nearly always arise as a consequence of genetic alterations, the vast majority of which begin in a single cell and therefore are monoclonal in origin. However, because a wide variety of genetic and epigenetic changes can occur in different cells within malignant tumors over time, most cancers are characterized by marked heterogeneity in the populations of cells. This heterogeneity significantly complicates the treatment of most cancers because it is likely that there are subsets of cells that will be resistant to therapy and will therefore survive and proliferate even if the majority of cells are killed.

A few cancers appear to, at least initially, be primarily driven by an alteration in a dominant gene that produces uncontrolled cell proliferation. Examples include chronic myeloid leukemia (*abl*), about half of melanomas (*braf*), Burkitt's lymphoma (c-*myc*), and subsets of lung adenocarcinomas (*egfr, alk, ros1, met, and ret*). The genes that can promote cell growth when altered are often called *oncogenes.* They were first identified as critical elements of viruses that cause animal tumors; it was subsequently found that the viral genes had normal counterparts with important functions in the cell and had been captured and mutated by viruses as they passed from host to host.

However, the vast majority of human cancers are characterized by a multiple step process involving many genetic abnormalities, each of which contributes to the loss of control of cell proliferation and differentiation and the acquisition of capabilities, such as tissue invasion, the ability to metastasize, and angiogenesis (development of new blood vessels required for tumor growth). These properties are not found in the normal adult cell from which the tumor is derived. Indeed, normal cells have a large number of safeguards against DNA damage (including multiple DNA repair and extensive DNA damage response mechanisms), uncontrolled proliferation, and invasion. Many cancers go through recognizable steps of progressively more abnormal phenotypes: hyperplasia, to adenoma, to dysplasia, to carcinoma in situ, to invasive cancer with the ability to metastasize **(Table 68-1)**. For most cancers, these changes occur over a prolonged period of time, usually many years.

TABLE 68-1

Phenotypic Characteristics of Malignant Cells

Deregulated cell proliferation: Loss of function of negative growth regulators (tumor suppressor genes, i.e., *Rb*, *p53*), and increased action of positive growth regulators (oncogenes, i.e., *Ras, Myc*). Leads to aberrant cell cycle control and includes loss of normal checkpoint responses.
Failure to differentiate: Arrest at a stage before terminal differentiation. May retain stem cell properties. (Frequently observed in leukemias due to transcriptional repression of developmental programs by the gene products of chromosomal translocations.)
Loss of normal apoptosis pathways: Inactivation of p53, increases in Bcl-2 (anti-apoptic) family members. This defect enhances the survival of cells with oncogenic mutations and genetic instability and allows clonal expansion and diversification within the tumor without activation of physiologic cell death pathways.
Genetic instability: Defects in DNA repair pathways leading to either single or oligo-nucleotide mutations (as in microsatellite instability, MIN) or more commonly chromosomal instability (CIN) leading to aneuploidy (abnormal number of chromosomes in a cell). Caused by loss of function of a number of proteins including p53, BRCA1/2, mismatch repair genes, DNA repair enzymes, and the spindle checkpoint. Leads to accumulation of a variety of mutations in different cells within the tumor and heterogeneity.
Loss of replicative senescence: Normal cells stop dividing in vitro after 25–50 population doublings. Arrest is mediated by the Rb, $p16^{INK4a}$, and p53 pathways. While most cells remain arrested, genetic and epigenetic changes in a subset of cells allows further replication leading to telomere loss, with crisis leading to death of many cells. Cells that survive often harbor gross chromosomal abnormalities and the ability to continue to proliferate. These cells express telomerase which maintains telomeres and is important for ongoing growth of these cells. Relevance to human in vivo cancer remains uncertain. Many human cancers express telomerase.
Non-responsiveness to external growth-inhibiting signals: Cancer cells have lost responsiveness to signals normally present to stop proliferating when they have overgrown the niche normally occupied by the organ from which they are derived. Our understanding about this mechanism of growth regulation remains limited.
Increased angiogenesis: Due to increased gene expression of proangiogenic factors (VEGF, FGF, IL-8, ANGIOPOEITIN) by tumor or stromal cells, or loss of negative regulators (endostatin, tumstatin, thrombospondin).
Invasion: Cell mobility and ability to move through extracellular matrix and into other tissues or organs. Loss of cell-cell contacts (gap junctions, cadherins) and increased production of matrix metalloproteinases (MMPs). Can take the form of epithelial-to-mesenchymal transition (EMT), with anchored epithelial cells becoming more like motile fibroblasts.
Metastasis: Spread of tumor cells to lymph nodes or distant tissue sites. Limited by the ability of tumor cells to migrate out of initial site and to survive in a foreign environment, including evading the immune system (see below).
Evasion of the immune system: Downregulation of MHC class I and II molecules; induction of T-cell tolerance; inhibition of normal dendritic cell and/or T-cell function; antigenic loss variants and clonal heterogeneity; increase in regulatory T cells.
Shift in cell metabolism: Complex changes including alterations due to tumor stress such as hypoxia, energy generation shifts from oxidative phosphorylation to aerobic glycolysis, generate building blocks for malignant cell production and proliferation.

Abbreviations: FGF, fibroblast growth factor; IL, interleukin; MHC, major histocompatibility complex; VEGF, vascular endothelial growth factor.

In most organs, only primitive undifferentiated cells are capable of proliferating and the cells lose the capacity to proliferate as they differentiate and acquire functional capability. The expansion of the primitive cells (stem cells) is linked to some functional need in the host through receptors that receive signals from the local environment or through hormonal and other influences delivered by the vascular supply. In the absence of such signals, the cells are at rest. The signals that keep the primitive cells at rest remain incompletely understood. These signals must be environmental, based on the observations that a regenerating liver stops growing when it has replaced the portion that has been surgically removed post partial hepatectomy and regenerating bone marrow stops growing when the peripheral blood counts return to normal. Cancer

cells clearly have lost responsiveness to such controls and do not recognize when they have overgrown the niche normally occupied by the organ from which they are derived. A better understanding of the mechanisms of growth regulation is evolving.

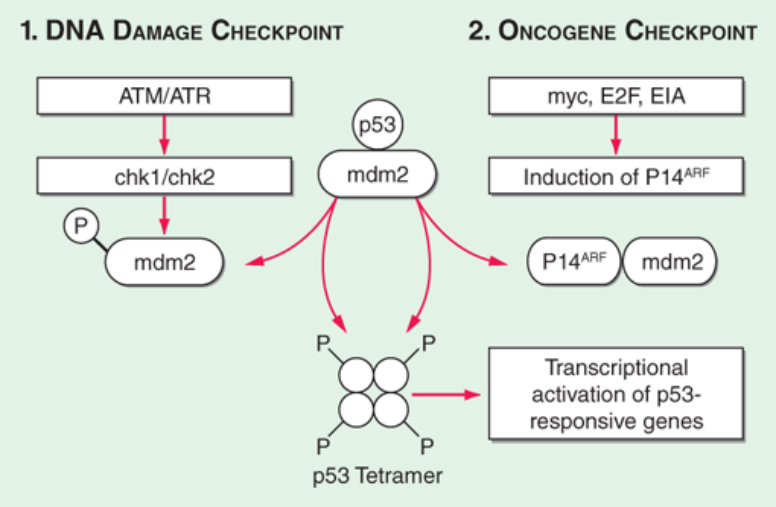
CELL CYCLE CHECKPOINTS

The cell division cycle consists of four phases—G1 (growth and preparation for DNA synthesis), S (DNA synthesis), G2 (preparation to divide), and M (mitosis, cell division). Cells can also exit the cell cycle and be quiescent (G0). Progression of a cell through the cell cycle is tightly regulated at a number of checkpoints (especially at the G1/S boundary, the G2/M boundary, and during M [spindle checkpoint]) by an array of genes that are targeted by specific genetic alterations in cancer. Critical proteins in these control processes that are frequently mutated or otherwise inactivated in cancers are called tumor-suppressor genes. Examples include p53 and Rb (discussed below). In the first phase, G_1, preparations are made to replicate the genetic material. The cell stops before entering the DNA synthesis phase, or S phase, to take inventory. Are we ready to replicate our DNA? Is the DNA repair machinery in place to fix any mutations that are detected? Are the DNA replicating enzymes available? Is there an adequate supply of nucleotides? Is there sufficient energy to proceed? The main brake on the process is the retinoblastoma protein, Rb. When the cell determines that it is prepared to move ahead, sequential activation of cyclin-dependent kinases (CDKs) results in the inactivation of the brake, Rb, by phosphorylation. Phosphorylated Rb releases the S phase–regulating transcription factor, E2F/DP1, and genes required for S phase progression are expressed. If the cell determines that it is unready to move ahead with DNA replication, a number of inhibitors are capable of blocking the action of the CDKs, including $p21^{Cip2/Waf1}$, $p16^{Ink4a}$, and $p27^{Kip1}$. *Nearly every cancer has one or more defects in the G_1 checkpoint that permit progression to S phase despite abnormalities in DNA repair machinery or other deficiencies that would affect normal DNA synthesis*.

At the end of the G2 phase and prior to the M phase, after the cell has exactly duplicated its DNA content, a second inventory is taken at the G2 checkpoint. Have all of the chromosomes been fully duplicated? Were all segments of DNA copied only once? Has all damaged DNA been repaired? Do we have the right number of chromosomes and the right amount of DNA? If so, the cell proceeds to G_2, in which the cell prepares for division by synthesizing mitotic spindle and other proteins needed to produce two daughter cells. When DNA damage is detected, the p53 pathway is normally activated. Called the guardian of the genome, p53 is a transcription factor that is normally present in the cell in very low levels. Its level is generally regulated through its rapid turnover. Normally, p53 is bound to mdm2, an ubiquitin ligase that both inhibits p53 transcriptional activation and also targets p53 for degradation in the proteasome. When damage is sensed, the ATM (ataxia-telangiectasia mutated) pathway is activated; ATM phosphorylates mdm2, which no longer binds to p53, and p53 then stops cell cycle progression, directs the synthesis of repair enzymes, or if the damage is too great, initiates apoptosis (programmed cell death) of the cell to prevent the propagation of a damaged cell (Fig. 68-1).

FIGURE 68-1

Induction of p53 by the DNA damage and oncogene checkpoints. In response to noxious stimuli, p53 and mdm2 are phosphorylated by the ataxia-telangiectasia mutated (ATM) and related ATR serine/threonine kinases, as well as the immediate downstream checkpoint kinases, Chk1 and Chk2. This causes dissociation of p53 from mdm2, leading to increased p53 protein levels and transcription of genes leading to cell cycle arrest ($p21^{Cip1/Waf1}$) or apoptosis (e.g., the proapoptotic Bcl-2 family members Noxa and Puma). Inducers of p53 include hypoxemia, DNA damage (caused by ultraviolet radiation, gamma irradiation, or chemotherapy), ribonucleotide depletion, and telomere shortening. A second mechanism of p53 induction is activated by oncogenes such as *Myc*, which promote aberrant G_1/S transition. This pathway is regulated by a second product of the Ink4a locus, $p14^{ARF}$ (p19 in mice), which is encoded by an *a*lternative *r*eading *f*rame (ARF) of the same stretch of DNA that codes for $p16^{Ink4a}$. Levels of ARF are upregulated by *Myc* and E2F, and ARF binds to mdm2 and rescues p53 from its inhibitory effect. This *oncogene checkpoint* leads to the death or senescence (an irreversible arrest in G_1 of the cell cycle) of renegade cells that attempt to enter S phase without appropriate physiologic signals. Senescent cells have been identified in patients whose premalignant lesions harbor activated oncogenes, for instance, dysplastic nevi that encode an activated form of BRAF (see below), demonstrating that induction of senescence is a protective mechanism that operates in humans to prevent the outgrowth of neoplastic cells.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A second method of activating p53 involves the induction of $p14^{ARF}$ by hyperproliferative signals from oncogenes. $p14^{ARF}$ competes with p53 for binding to mdm2, allowing p53 to escape the effects of mdm2 and accumulate in the cell. Then p53 stops cell cycle progression by activating CDK inhibitors such as p21 and/or initiating the apoptosis pathway. Not surprisingly given its critical role in controlling cell cycle progression, mutations in the gene for p53 on chromosome 17p are among the most frequent mutations in human cancers, although percentages vary between different cancers. Most commonly these mutations are acquired in the malignant tissue in one allele and the second allele is inactivated (such as by deletion), leaving the cell unprotected from DNA-damaging agents or activated oncogenes. Some environmental exposures produce signature mutations in p53; for example, aflatoxin exposure leads to mutation of arginine to serine at codon 249 and leads to hepatocellular carcinoma. In rare instances, p53 mutations are in the germline (Li-Fraumeni syndrome) and produce a familial cancer syndrome. The absence of p53 leads to chromosome instability and the accumulation of DNA damage including the acquisition of properties that give the abnormal cell a proliferative and survival advantage. *Like Rb dysfunction, most cancers have mutations that disable the p53 pathway*. Indeed, the importance of p53 and Rb in the development of cancer is underscored by the neoplastic transformation mechanism of human papillomavirus. This virus has two main oncogenes, E6 and E7. E6 acts to increase the rapid turnover of p53, and E7 acts to inhibit Rb function; inhibition of these two targets is required for transformation of epithelial cells.

Another cell cycle checkpoint exists when the cell is undergoing division (M phase), the spindle checkpoint which acts to ensure that there is proper attachment of chromosomes to the mitotic spindle before progression through the cell cycle can occur. If the spindle apparatus does not properly align the chromosomes for division, if the chromosome number is abnormal (i.e., greater or less than 4*n*), or if the centromeres are not properly paired with their duplicated partners, then the cell initiates a cell death pathway to prevent the production of aneuploid progeny (having an altered number of chromosomes). Abnormalities in the spindle checkpoint facilitate the development of aneuploidy which is frequently found in cancers. In some tumors, aneuploidy is a predominant genetic feature. In others, a defect in the cells' ability to repair errors in the DNA, such as due to mutations in genes coding for the proteins critical for mismatched DNA repair, is the primary genetic lesion. Mismatch repair is usually detected by finding alterations in repeat sequences of DNA (called microsatellites), or microsatellite instability, in malignant cells. In general, tumors either have defects in chromosome number or defective DNA repair pathways such as microsatellite instability, but not both. Defects that lead to cancer include abnormal cell cycle checkpoints, inadequate DNA repair, and failure to preserve genome integrity leading to DNA damage. These defects and the stress of the resultant increased DNA damage make cancer cells more vulnerable to additional DNA damage which can be exploited by chemotherapy, radiation therapy, and immunotherapy which are the major systemic therapeutic approaches effective against cancer.

Efforts are also under way to therapeutically restore the defects in cell cycle regulation that characterize cancer, although this remains a challenging problem because it is much more difficult to restore normal biologic function than to inhibit abnormal function of proteins driving cell proliferation, such as occurs with oncogenes. Newer approaches to gene editing (e.g., Clustered Regularly Interspaced Short Palindromic Repeats [CRISPR]) should make this more feasible.

CANCER AS AN ORGAN THAT IGNORES ITS NICHE

The fundamental cellular defects that create a malignant neoplasm act at the cellular level and some of these are cell autonomous. However, that is not the entire story. Cancers consist of both malignant cells as well as other cells in the cancer microenvironment and behave as organs

that have lost their specialized function and stopped responding to signals that would limit their growth in tightly regulated normal tissue homeostasis. Human cancers usually become clinically detectable when a primary mass is at least 1 cm in diameter—such a mass consists of about 10^9 cells. More commonly patients present with tumors that are at least 10^{10} cells. A lethal tumor burden is about 10^{12}–10^{13} cells, although there is significant variability depending on the type and location of the cancer. If all malignant cells were dividing at the time of diagnosis, patients would reach a lethal tumor burden in a very short time. However, human tumors grow by Gompertzian kinetics—this means that not every daughter cell produced by a cell division is actively dividing. In addition, the overall growth rate of a tumor depends on differences between growth rates of different cells within the tumor and rate of cell loss. The growth fraction of a tumor declines with time, largely due to factors in the microenvironment. The growth fraction of the first malignant cell is 100%, and by the time a patient presents for medical care, the growth fraction is estimated to be <10%, although the fraction varies between different types of cancers and even different cancers of the same type in different individuals. This fraction is similar to the growth fraction of normal bone marrow and normal intestinal epithelium, the most highly proliferative normal tissues in the human body, a fact that may explain the dose-limiting toxicities of agents that target dividing cells.

The implication of these data is that the tumor is slowing its own growth over time. How does it do this? The tumor cells have multiple genetic lesions that tend to promote proliferation, yet by the time the tumor is clinically detectable, its capacity for proliferation has declined. Better understanding of how a tumor slows its own growth would provide important clues for better cancer treatment. A number of factors, including those in the tumor microenvironment, are known to contribute to the failure of tumor cells to proliferate in vivo. Some cells are hypoxemic and have inadequate supply of nutrients and energy. Some have sustained too much genetic damage to complete the cell cycle but have lost the capacity to undergo apoptosis and therefore survive but do not proliferate. However, an important subset is not actively dividing but retains the capacity to divide and can start dividing again under certain conditions such as when the tumor mass is reduced by treatments leading to improved conditions in the tumor microenvironment favorable for cell proliferation. Just as the bone marrow increases its rate of proliferation in response to bone marrow–damaging agents, the tumor also seems to sense when tumor cell numbers have been reduced and can respond by increasing growth rate. However, the critical difference is that the marrow stops growing when it has reached its production goals whereas tumors do not.

Additional tumor cell vulnerabilities are likely to be detected when we learn more about how normal cells respond to “stop” signals from their environment, and why and how tumor cells fail to heed such signals.

IS IN VITRO SENESCENCE RELEVANT TO CARCINOGENESIS?

When normal cells are placed in culture in vitro, most are not capable of sustained growth. Fibroblasts are an exception to this rule. When they are cultured, fibroblasts may divide 30–50 times and then they undergo what has been termed a “crisis” during which the majority of cells stop dividing (usually due to an increase in p21 expression, a CDK inhibitor), many die, and a small fraction emerge that have acquired genetic and epigenetic changes that permit their uncontrolled growth. The cessation of growth of normal cells in culture has been termed “senescence” and whether this phenomenon is relevant to any physiologic event in vivo is still an area of investigation, including identifying biomarkers of senescence in vivo.

Among the cellular changes during in vitro propagation is telomere shortening. DNA polymerase is unable to replicate the tips of chromosomes, resulting in the loss of DNA at the specialized ends of chromosomes (called *telomeres*) with each replication cycle. At birth, human telomeres are 15- to 20-kb pairs long and are composed of tandem repeats of a six-nucleotide sequence (TTAGGG) that associates with specialized telomere-binding proteins to form a T-loop structure that protects the ends of chromosomes from being mistakenly recognized as damaged. The loss of telomeric repeats with each cell division cycle causes gradual telomere shortening, leading to growth arrest (called *senescence*) when one or more critically short telomeres trigger a p53-regulated DNA-damage checkpoint response. Cells can bypass this growth arrest if pRb and p53 are nonfunctional, but cell death usually ensues when the unprotected ends of chromosomes lead to chromosome fusions or other catastrophic DNA rearrangements. *The ability to bypass telomere-based growth limitations is thought to be a critical step in the evolution of most malignancies*. This occurs by the reactivation of telomerase expression in cancer cells. Telomerase is an enzyme that adds TTAGGG repeats onto the 3′ ends of chromosomes. It contains a catalytic subunit with reverse transcriptase activity (hTERT) and an RNA component that provides the template for telomere extension. Most normal somatic cells do not express sufficient telomerase to prevent telomere attrition with each cell division. Exceptions include stem cells (such as those found in hematopoietic tissues, gut and skin epithelium, and germ cells) that require extensive cell division to maintain tissue homeostasis. More than 90% of human cancers express high levels of telomerase that prevent telomere shortening to critical levels and allow indefinite cell proliferation. In vitro experiments indicate that inhibition of telomerase activity leads to tumor cell apoptosis. Major efforts are underway to develop methods to inhibit telomerase activity in cancer cells. For example, the protein component of telomerase (hTERT) may act as one of the most widely expressed tumor-associated antigens and can be targeted by vaccine

approaches. However, a caveat to targeting telomerase for anticancer treatment is an inadequate understanding of how important its presence is in certain normal cells to maintaining the normal physiologic state.

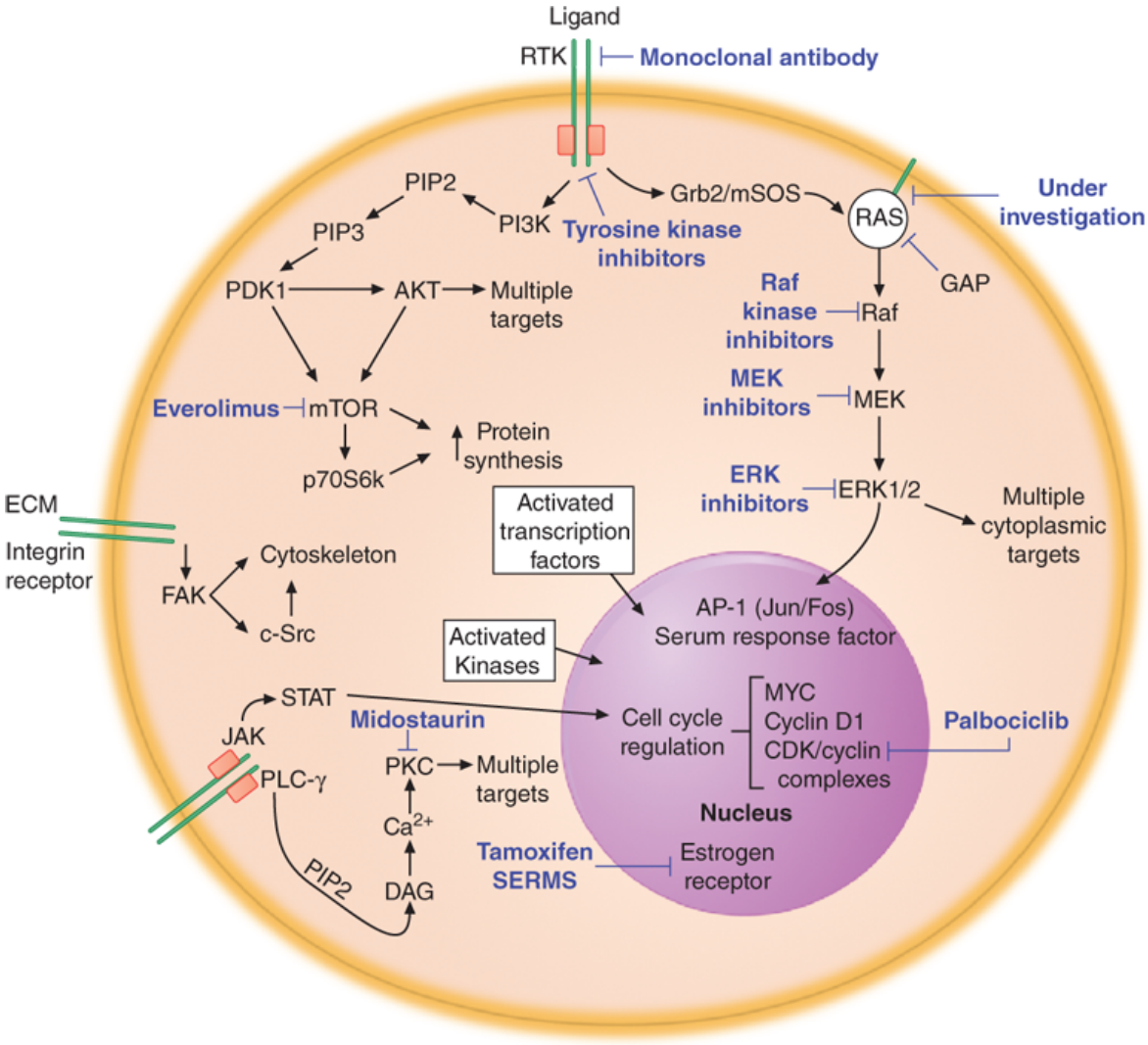
Although most of the functions of telomerase relate to cell division, it also has several other effects including interfering with the differentiated functions of at least certain stem cells. However, the impact on differentiated function of normal non-stem cells is less clear. The picture is further complicated by the fact that rare genetic defects in the telomerase enzyme seem to cause pulmonary fibrosis, aplastic anemia, or dyskeratosis congenita (characterized by abnormalities in skin, nails, and oral mucasa with increased risk for certain malignancies) but not defects in nutrient absorption in the gut, a site that might be presumed to be highly sensitive to defective cell proliferation. Much remains to be learned about how telomere shortening and telomere maintenance are related to human illness in general and cancer in particular.

SIGNAL TRANSDUCTION PATHWAYS IN CANCER CELLS

Signals that affect cell behavior come from adjacent cells, the stroma in which the cells are located, hormonal signals that originate remotely, and from the cells themselves (autocrine signaling). These signals generally exert their influence on the receiving cell through activation of signal transduction pathways that have as their end result the induction of activated transcription factors that mediate a change in cell behavior or function or the acquisition of effector machinery to accomplish a new task. Although signal transduction pathways can lead to a wide variety of outcomes, many such pathways rely on cascades of signals that sequentially activate different proteins or glycoproteins and lipids or glycolipids, and the activation steps often involve the addition or removal of one or more phosphate groups on a downstream target. Other chemical changes can result from signal transduction pathways, but phosphorylation and dephosphorylation play a major role. The proteins that add phosphate groups to proteins are called kinases. There are two major distinct classes of kinases; one class acts on tyrosine residues and the other acts on serine/threonine residues. The tyrosine kinases often play critical roles in signal transduction pathways; they may be receptor tyrosine kinases (RTKs) or they may be linked to other cell-surface receptors through associated docking proteins and transmit the signal into the cell **(Fig. 68-2)**.

FIGURE 68-2

Therapeutic targeting of signal transduction pathways in cancer cells. Three major signal transduction pathways are activated by receptor tyrosine kinases (RTK). *1.* The protooncogene Ras is activated by the Grb2/mSOS guanine nucleotide exchange factor, which induces an association with Raf and activation of downstream kinases (MEK and ERK1/2). *2.* Activated PI3K phosphorylates the membrane lipid PIP_2 to generate PIP_3, which acts as a membrane-docking site for a number of cellular proteins including the serine/threonine kinases PDK1 and Akt. PDK1 has numerous cellular targets, including Akt and mTOR. Akt phosphorylates target proteins that promote resistance to apoptosis and enhance cell cycle progression, while mTOR and its target p70S6K upregulate protein synthesis to potentiate cell growth. *3.* Activation of PLC-γ leads the formation of diacylglycerol (DAG) and increased intracellular calcium, with activation of multiple isoforms of PKC and other enzymes regulated by the calcium/calmodulin system. Other important signaling pathways involve non-RTKs that are activated by cytokine or integrin receptors. Janus kinases (JAK) phosphorylate STAT (signal transducer and activator of transcription) transcription factors, which translocate to the nucleus and activate target genes. Integrin receptors mediate cellular interactions with the extracellular matrix (ECM), inducing activation of FAK (focal adhesion kinase) and c-Src, which activate multiple downstream pathways, including modulation of the cell cytoskeleton. Many activated kinases and transcription factors migrate into the nucleus, where they regulate gene transcription, thus completing the path from extracellular signals, such as growth factors, to a change in cell phenotype, such as induction of differentiation or cell proliferation. The nuclear targets of these processes include transcription factors (e.g., Myc, AP-1, and serum response factor) and the cell cycle machinery (CDKs and cyclins). Inhibitors of many of these pathways have been developed for the treatment of human cancers. Examples of inhibitors that are either approved or are currently being evaluated in clinical trials are shown in purple type.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Normally, tyrosine kinase activity is short-lived and reversed by protein tyrosine phosphatases (PTPs). However, in many human cancers, tyrosine kinases or components of their downstream pathways are activated by mutation, gene amplification, or chromosomal translocations. Because these pathways regulate proliferation, survival, migration, and angiogenesis, they have been identified as important targets for cancer therapeutics.

Inhibition of kinase activity is effective in the treatment of a number of neoplasms. Lung cancers with mutations in the epidermal growth factor receptor are highly responsive to erlotinib and gefitinib **(Table 68-2)**. Lung cancers with activation of anaplastic lymphoma kinase (ALK) or ROS1 by translocations respond to crizotinib, an ALK and ROS1 inhibitor and additional ALK inhibitors including ceritinib and alectinib are available for treating lung cancers with a number of additional inhibitors currently in trials. BRAF inhibitors are highly effective in melanomas and thyroid cancers in which BRAF is mutated. Targeting a protein (MEK) downstream of BRAF also has activity against BRAF mutant melanomas and combined inhibition of BRAF and MEK is more effective than either alone. Janus kinase (JAK) inhibitors are active in myeloproliferative syndromes in which JAK2 activation is a pathogenetic event. Imatinib (which targets a number of tyrosine kinases) is an effective agent in tumors that have translocations of the c-Abl and BCR gene (such as chronic myeloid leukemia), mutant c-Kit (gastrointestinal stromal cell tumors), or mutant platelet-derived growth factor receptor (PDGFR a; gastrointestinal stromal tumors); second-generation inhibitors of BCR-Abl, dasatinib, and nilotinib are even more effective and the third generation agent bosutinib has activity in some patients who have progressed on other inhibitors, while the third generation ponatinib has activity against the T315I mutation, which is resistant to the other agents. Although almost all tyrosine kinase inhibitors are not entirely selective for one protein, certain inhibitors have significant activity against a broad number of proteins. These include sorafenib, regorafenib, cabozantinib, sunitinib, and lenvatinib. These have shown antitumor activity in various malignancies, including renal cell cancer (RCC) (sorafenib, sunitinib, cabozantinib, lenvatinib), hepatocellular carcinoma (sorafenib, regorafenib, lenvatinib), gastrointestinal stromal tumor (GIST) (sunitinib, regorafenib), thyroid cancer (sorafenib, cabozantinib, lenvatinib), colorectal cancer (regorafenib), and pancreatic neuroendocrine tumors (sunitinib). Inhibitors of the mammalian target of rapamycin (mTOR) are active in RCC, pancreatic neuroendocrine tumors, and breast cancer. The list of active agents and treatment indications is growing rapidly (Table 68-2). These

agents have ushered in a new era of personalized therapy. It is becoming more common for resected tumors to be assessed for specific molecular changes that predict response and to have clinical decision-making guided by those results. This is now an important component of standard therapy for metastatic lung, gastroesophageal, melanoma, breast, and colorectal cancers as well as in adjuvant therapy for breast cancer.

TABLE 68-2

Some FDA-Approved Molecularly Targeted Agents for the Treatment of Cancer

Drug	Molecular Target	Disease	Mechanism of Action
All-*trans* retinoic acid	PML-RARα oncogene	Acute promyelocytic leukemia M3 AML; t(15;17)	Inhibits transcriptional repression by PML-RARα
Imatinib	Bcr-Abl, c-Abl, c-Kit, PDGFR-α/β	Chronic myeloid leukemia; GIST	Blocks ATP binding to tyrosine kinase active site
Dasatinib, Nilotinib, Ponatinib, Bosutinib	Bcr-Abl (primarily)	Chronic myeloid leukemia	Blocks ATP binding to tyrosine kinase active site
Sunitinib	c-Kit, VEGFR-2, PDGFR-β, Flt-3	GIST; RCC; PNET	Inhibits activated c-Kit and PDGFR in GIST; inhibits VEGFR in RCC and probably in PNET
Sorafenib	RAF, VEGFR-2, PDGFR-α/β, Flt-3, c-Kit	RCC; hepatocellular carcinoma, differentiated thyroid cancer, desmoid	Targets VEGFR pathways in RCC and HCC. Possible activity against BRAF in thyroid cancer
Regorafenib	VEGFR1-3,TIE-2, FGFR1,KIT,RET, PDGFR	Colorectal cancer; GIST; HCC	Competitive inhibitor ATP binding site of tyrosine kinase domain multiple kinases including VEGFR
Axitinib	VEGFR 1-3	RCC	Competitive inhibitor ATP binding site of tyrosine kinase domain VEGF receptors
Erlotinib	EGFR	NSCLC; pancreatic cancer	Competitive inhibitor of the ATP-binding site of the EGFR
Afitinib	EGFR (and other HER family)	NSCLC	Irreversible inhibitor of ATP-binding site of HER family members
Osimertinib	EGFR(T790M)	NSCLC	Inhibits EGFR mutations including T790M mutant NSCLC
Lapatinib	HER2/neu	Breast Cancer	Competitive inhibitor of the ATP binding site of HER2
Crizotinib, Ceritinib, Alectinib	ALK, ROS1	NSCLC	Inhibitor of ALK and ROS1 tyrosine kinase
Palbociclib, Ribociclib, Abemaciclib	CDK4/6	Breast	Inhibitor of CDK4/6
Bortezomib, Carfilzomib, Ixazomib	Proteasome	Multiple myeloma	Inhibits proteolytic degradation of multiple cellular proteins
Vemurafenib, Dabrafenib	BRAF	Melanoma	Inhibitor of serine-threonine kinase domain of V600E mutant of BRAF

Drug	Molecular Target	Disease	Mechanism of Action
Trametinib, Cobimetinib	MEK	Melanoma	Inhibitor of serine-threonine kinase domain of MEK
Cabozantinib	RET, MET, VEGFR	MTC, RCC	Competitive inhibitor ATP binding site of tyrosine kinase domain multiple kinases, including VEGFR2 and RET
Vandetanib	RET, VEGFR, EGFR	MTC	Competitive inhibitor ATP-binding site of tyrosine kinase domain multiple kinases, including RET
Temsiroliumus	mTOR	RCC	Competitive inhibitor of mTOR serine-threonine kinase
Everolimus	mTOR	RCC; PNET	Binds to immunophilin FK binding protein-12 which forms complex that inhibits mTOR kinase
Vorinostat, Romidepsin, Belinostat	HDAC	CTCL/PTL	HDAC inhibitor, epigenetic modulation
Panobinostat	HDAC	MM	HDAC inhibitor, epigenetic modulation
Ruxolitinib	JAK-1, 2	Myelofibrosis	Competitive inhibitor of tyrosine kinase
Vismodegib	Hedgehog pathway	Basel cell cancer (skin)	Inhibits smoothened in Hedgehog Pathway
Lenvatinib	Multi-Kinase inhibitor (VEGFR, FGFR, PGFR-a, others)	RCC, Thyroid cancer	Competitive inhibitor ATP-binding site of tyrosine kinase domain multiple kinases
Olaparib, rucaparib	PARP	BRCA mutant ovarian (both) and breast (olaparib) cancers	Inhibits PARP and DNA repair
Venetoclax	BCL-2	CLL (with 17p deletion)	Inhibits BCL-2 and enhances apoptosis
Ibrutinib	Bruton Tyrosine Kinase (BTK)	CLL, MCL, MZL, SLL, WM	Inhibitor of BTK
Idealisib	PI3K-delta	CLL, SLL, FL	Inhibits PI3k-delta,preventing proliferation and inducing apoptosis
Monoclonal Antibodies Alone			
Trastuzumab	HER2/neu (ERBB2)	Breast cancer	Binds HER2 on tumor cell surface and induces receptor internalization
Pertuzumab	HER2/neu (ERBB2)	Breast cancer	Binds HER2 on tumor cell surface at distinct site from Trastuzumab and prevents binding to other receptors
Cetuximab	EGFR	Colon cancer, squamous cell carcinoma of the head and neck	Binds extracellular domain of EGFR and blocks binding of EGF and TGF-α; induces receptor internalization. Potentiates the efficacy of chemotherapy and radiotherapy

Drug	Molecular Target	Disease	Mechanism of Action
Panitumumab	EGFR	Colon cancer	Similar to Cetuximab but fully humanized rather than chimeric
Necitumumab	EGFR	Squamous NSCLC	Binds EGFR
Rituximab	CD20	B-cell lymphomas and leukemias that express CD20	Multiple potential mechanisms, including direct induction of tumor cell apoptosis and immune mechanisms
Alemtuzumab	CD52	Chronic lymphocytic leukemia and CD52-expressing lymphoid tumors	Immune mechanisms
Bevacizumab	VEGF	Colorectal, lung cancers, RCC, glioblastoma	Inhibits angiogenesis by high-affinity binding to VEGF
Ziv-Aflibercept	VEGFA, VEGFB, PLGF	Colorectal cancers	Inhibits angiogenesis by high-affinity binding to VEGFA, B and PLGF
Ramucirumab	VEGFR	gastric, colorectal, lung cancers	Inhibits angiogenesis by binding to VEGFR
Ipilumumab	CTLA-4	Melanoma	Blocks CTLA-4 preventing interaction with CD80/86 and T-cell inhibition
Nivolumab Pembrolizumab	PD-1	Melanoma, Head and Neck Cancer, NSCLC, Hodgkins Disease, Urothelial cancer, RCC, HCC, gastric cancer, MSI high cancers	Blocks PD-1 preventing interaction with PDL-1 and T-cell inhibition
Atezolizumab, Durvalumab	PDL1	NSCLC, Urothelial cancer	Blocks PDL1 preventing interaction with PD-1 and T-cell inhibition
Denosumab	Rank ligand	Breast, Prostate	Inhibits Rank ligand, primary signal for bone removal
Dinutuximab	Glycolipid GD2	Neuroblastoma (pediatric)	Immune mediated attack on GD2 expressing cells
Daratumumab	CD38	MM	Binds to CD38 on MM cells causing apoptosis by antibody dependent or compliment mediated cytotoxicity
Elotuzumab	SLAMF7	MM	Activating NK cells to kill MM cells
Olarutumab	PDGFRa	Soft Tissue Sarcomas	Blocks PDGFRa activity
Blinotumumab	CD19 and CD3	PH-relapsed precursor Bcell ALL	Binds CD19 on ALL cells and CD3 on T cells; Immune attack on CD19 expressing cells
Antibody-Chemotherapy Conjugates			
Brentuximab vedotin	CD30	HD, Anaplastic Lymphoma	Delivery of chemotherapeutic agent (MMAE) to CD30 expressing tumor cells
Ado-Trastuzumab emtansine	HER2	Breast Cancer	Delivery of chemotherapeutic agent emtansine to HER2 expressing breast cancer cells
CAR-T Cells			

Drug	Molecular Target	Disease	Mechanism of Action
Kymria, Yescarta	CD19	ALL (Kymria), DLBCL (Yescarta)	Targeted T-cells to protein on surface of malignant cells

Abbreviations: ALL, acute lymphocytic leukemia; AML, acute myeloid leukemia; DLBCL, diffuse large B-cell lymphoma; EGFR, epidermal growth factor receptor; Flt-3, fms-like tyrosine kinase-3; GIST, gastrointestinal stromal tumor; HDAC-histone deacetylases; MCL, mantle cell lymphoma; MSI, microsatellite instability; MZL, mantle zone lymphoma; NSCLC, nonsmall cell lung cancer; PARP, poly ADP ribose, polymerase; PDGFR, platelet-derived growth factor receptor; PLGF, placenta growth factor; PML-RARα, promyelocytic leukemia-retinoic acid receptor-alpha; PNET, pancreatic neuroendocrine tumors; RCC, renal cell cancer; t(15;17), translocation between chromosomes 15 and 17; SLL, small lymphocytic lymphoma; TGF-α, transforming growth factor-alpha; VEGFR, vascular endothelial growth factor receptor; WM, Waldenstroms macroglobulinemia.

However, none of these therapies has yet been curative by themselves for any malignancy, although prolonged periods of disease control lasting many years frequently occur in CML, including a >80% survival rate at 10 years. The reasons for the failure to cure are not completely defined, although resistance to the treatment ultimately develops in most patients. In some tumors, resistance to kinase inhibitors is related to an acquired mutation in the target kinase that inhibits drug binding. Many of these kinase inhibitors act as competitive inhibitors of the ATP-binding pocket. ATP is the phosphate donor in these phosphorylation reactions. For example, mutation in the critical BCR-ABL kinase in the ATP-binding pocket (such as the threonine to isoleucine change at codon 315 [T315I]) can prevent imatinib binding. Other resistance mechanisms include alterations in other signal transduction pathways to bypass the inhibited pathway. As resistance mechanisms become better defined, rational strategies to overcome resistance will emerge. In addition, many kinase inhibitors are less specific for an oncogenic target than was hoped, and toxicities related to off-target inhibition of kinases limits the use of the agent at a dose that would optimally inhibit the cancer-relevant kinase.

Targeted agents can also be used to deliver highly toxic compounds. An important component of the technology for developing effective conjugates is the design of the linker between the two which needs to be stable. Examples of currently approved antibody drug conjugates include brentuximab vedotin, which links the microtubule toxin monomethyl auristatin E (MMAE) to an antibody targeting the cell surface antigen CD30, which is expressed on a number of malignant cells but especially in Hodgkin's lymphoma and anaplastic lymphoma. The linker in this case is cleavable which allows diffusion of the drug out of the cell after delivery. A second approved conjugate is ado-trastuzumab emtansine which links the microtubule formation inhibitor mertansine and the monoclonal antibody trastuzumab targeted against HER2 on breast cancer cells. In this case the linker is non-cleavable, thus trapping the chemotherapeutic agent within the cells. There are theoretical pluses and minuses to having either cleavable or non-cleavable linkers and it is likely that both will be used in future developments of antibody-drug conjugates.

Another strategy to enhance the antitumor effects of targeted agents is to use them in rational combinations with each other as well as with chemotherapy or immunotherapy agents that kill cells in ways distinct from agents targeting specific mutant or overexpressed proteins. Combinations of trastuzumab (a monoclonal antibody that targets the HER2 receptor [member of the EGFR family]) with chemotherapy have significant activity against breast and stomach cancers that have high levels of expression of the HER2 protein. The activity of trastuzumab and chemotherapy can be enhanced further by combinations with another targeted monoclonal antibody (pertuzumab) which prevents dimerization of the HER2 receptor with other HER family members including HER3.

Although targeted therapies have not yet resulted in cures when used alone, their use in the adjuvant setting and when combined with other effective treatments has substantially increased the fraction of patients cured. For example, the addition of rituximab, an anti-CD20 antibody, to combination chemotherapy in patients with diffuse large B cell lymphoma improves cure rates by ~15%. The addition of trastuzumab, antibody to HER2, to combination chemotherapy in the adjuvant treatment of HER2-positive breast cancer significantly improves overall survival.

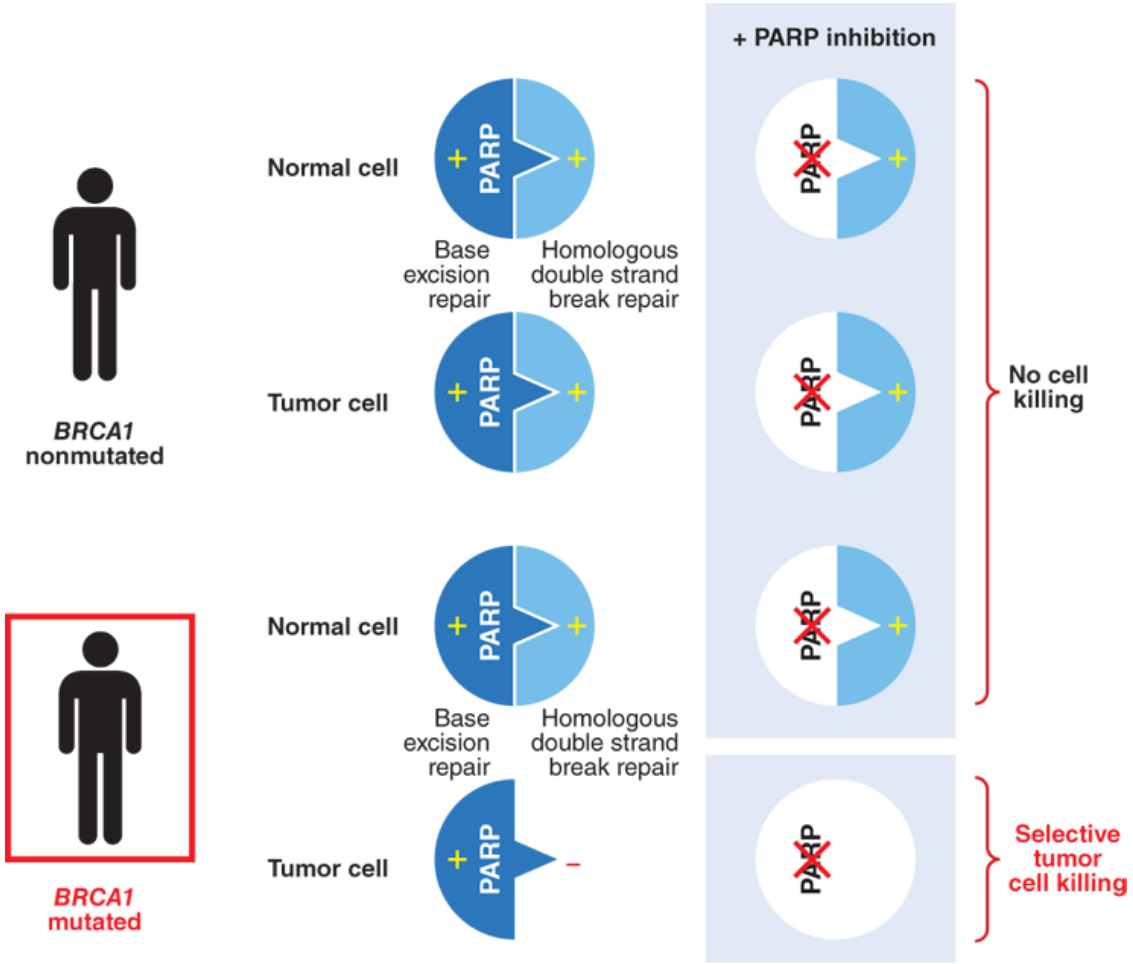
A major effort is underway to develop targeted therapies for mutations in the *ras* family of genes which are the most common mutations in oncogenes in cancers (especially *kras*) but have proved to be very difficult targets for a number of reasons related to the structure of RAS proteins as well as mechanisms of activation and inactivation. Targeted therapies against a subset of proteins downstream of RAS in the signaling pathway (including BRAF and mitogen-activated protein [MAP] kinase) have proven to have significant antitumor activity against V600E BRAF mutant melanoma, with improved efficacy when they are used in combination. However, similar activity is not seen against *ras* mutant tumors. Additional targeted therapies against other proteins downstream of RAS (including ERK, or combinations of MAP kinase inhibitors and immunotherapy) are currently being studied, both individually and in combination. However, at this time, there is no established effective approach to inhibiting RAS mutant tumors. Inhibitors of phospholipid signaling pathways such as the phosphatidylinositol-3-kinase (PI3K) and phospholipase C-gamma pathways, which are involved in a large number of cellular processes that are important in cancer development and

progression, are being evaluated. The targeting of a variety of other pathways that are activated in malignant cells, such as the MET pathway, hedgehog pathway, and various angiogenesis pathways are also being explored.

One of the strategies for new drug development is to take advantage of so-called oncogene addiction. This situation (Fig. 68-3) is created when a tumor cell develops an activating mutation in an oncogene that becomes a dominant pathway for survival and growth with reduced contributions from other pathways, even when there may be abnormalities in those pathways. This dependency on a single pathway creates a cell that is vulnerable to inhibitors of that oncogene pathway. For example, cells harboring mutations in *BRAF* are very sensitive to MEK inhibitors that inhibit downstream signaling in the BRAF pathway.

FIGURE 68-3

Synthetic lethality. Genes are said to have a synthetic lethal relationship when mutation of either gene alone is tolerated by the cell, but mutation of both genes leads to lethality, as originally noted by Bridges and later named by Dobzhansky. Thus, mutant *gene a* and *gene b* have a synthetic lethal relationship, implying that the loss of one gene makes the cell dependent on the function of the other gene. In cancer cells, loss of function of a DNA repair gene like *BRCA1*, which repairs double-strand breaks, makes the cell dependent on base excision repair mediated in part by *PARP.* If the PARP gene product is inhibited, the cell attempts to repair the break using the error-prone non-homologous end-joining method, which results in tumor cell death. High-throughput screens can now be performed using isogenic cell line pairs in which one cell line has a defined defect in a DNA repair pathway. Compounds can be identified that selectively kill the mutant cell line; targets of these compounds have a synthetic lethal relationship to the repair pathway, and are potentially important targets for future therapeutics.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Proteins critical for transcription of other proteins essential for malignant cell survival or proliferation provide another potential target for treating cancers. The transcription factor NF-κB is a heterodimer composed of p65 and p50 subunits that associate with an inhibitor, IκB, in the cell cytoplasm. In response to growth factor or cytokine signaling, a multi-subunit kinase called IKK (IκB-kinase) phosphorylates IκB and directs its degradation by the ubiquitin/proteasome system. NF-κB, free of its inhibitor, translocates to the nucleus and activates target genes, many of which promote the survival of tumor cells. One of the mechanisms by which novel drugs called *proteasome inhibitors* are thought to produce an anticancer effect is by blocking the proteolysis of IκB, thereby preventing NF-κB activation. For reasons that have not been fully elucidated, this has a differential toxicity affect on tumor, as compared to normal, cells. Although this mechanism appears to be an important aspect of the

antitumor effects of proteasome inhibitors, there are other effects involving the inhibition of the degradation of multiple cellular proteins important in malignant cell survival or proliferation. Proteasome inhibitors (bortezomib, carfilzomib, ixazomib) have activity in patients with multiple myeloma, including partial and complete remissions. Inhibitors of IKK are also in development, with the hope of more selectively blocking the degradation of IκB, thus "locking" NF-κB in an inhibitory complex and rendering the cancer cell more susceptible to apoptosis-inducing agents. Many other transcription factors are activated by phosphorylation, which can be prevented by tyrosine- or serine/threonine kinase inhibitors, a number of which are currently in clinical trials.

Estrogen receptors (ERs) and androgen receptors (ARs), members of the steroid hormone family of nuclear receptors, are targets of inhibition by drugs used to treat breast and prostate cancers, respectively. Selective estrogen receptor modulators (SERMs) have been developed as a treatment approach for ER positive breast cancer. Tamoxifen, a partial agonist and antagonist of ER function, is frequently used in breast cancer and can mediate tumor regression in metastatic breast cancer and can prevent disease recurrence in the adjuvant setting. Tamoxifen binds to the ER and modulates its transcriptional activity, inhibiting activity in the breast but promoting activity in bone but unfortunately also in uterine epithelium leading to a small increased risk of uterine cancer. Attempts have been made to develop SERMs that would have anti-estrogeneic effects in both breast and uterus while maintaining protective effects on bone. However, none of these to date has been an improvement over tamoxifen. Aromatase inhibitors, which block the conversion of androgens to estrogens in breast and subcutaneous fat tissues, have demonstrated improved clinical efficacy compared with tamoxifen in postmenopausal women and are often used as first-line therapy in post menopausal patients with ER-positive disease. They are occasionally utilized in premenopausal patients with ER positive disease in combination with ovarian suppression approaches such as leutinizing hormone receptor (LHRH) agonists. A number of approaches have been developed for blocking androgen stimulation of prostate cancer, including decreasing production by the testicles (e.g., orchiectomy, LHRH agonists or antagonists), directly blocking actions of androgen (a number of agents have been developed to do this), or blocking production by inhibiting the enzyme CYP17 which is central in production of androgens from cholesterol **(Chap. 75)**.

CANCER-SPECIFIC GENETIC CHANGES AND SYNTHETIC LETHALITY

The concepts of oncogene addiction and synthetic lethality have spurred new drug development targeting oncogene- and tumor-suppressor pathways. As discussed earlier in this chapter and outlined in Fig. 68-3, cancer cells can become dependent upon signaling pathways containing activated oncogenes; this can effect proliferation (i.e., mutated KRAS, BRAF, overexpressed Myc, or activated tyrosine kinases). Additional genetic changes in malignant cells or unique features of tumors including defects in DNA repair (e.g., loss of *BRCA1* or *BRCA2* gene function), modifications in cell cycle control (e.g., changes in protein levels or mutations in cyclins and cyclin dependent kinases), enhanced survival mechanisms (overexpression of Bcl-2 or NF-κB), altered cell metabolism (such as occurs when mutant KRAS enhances glucose uptake and aerobic glycolysis), tumor-stromal interactions, and angiogenesis (e.g., production of vascular endothelial growth factor [VEGF] in response to HIF-2α in RCC) can also be successfully exploited to relatively specifically target cancers. However, resistance to inhibition of specific oncogenic pathways almost always eventually develops. In addition, targeting defects in tumor- suppressor genes has been much more difficult, both because the target of mutation is often deleted and because it is much more difficult to restore normal function than to inhibit abnormal function of a protein.

Synthetic lethality occurs when loss of function in either of two or more genes individually has limited effects on cell survival but loss of function in both (or more) genes leads to cell death. In the case of oncogene addicted pathways, identifying genes that have a synthetic lethal relationship with the activated pathway may allow enhanced cell killing and decreased resistance by targeting those genes or their proteins. In the case of mutant tumor-suppressor genes, identifying genes that have a synthetic lethal relationship to those mutated pathways may allow targeting by inhibiting proteins required uniquely by those cells for survival or proliferation (Fig. 68-3). This is a much more tractable approach than attempting to repair normal function of the mutant suppressor gene itself. Examples of synthetic lethality with potential clinical impact have been identified. For instance, cells with mutations in the *BRCA1* or *BRCA2* tumor-suppressor genes (e.g., a subset of breast and ovarian cancers) are unable to repair DNA damage by homologous recombination. Poly ADP ribose, polymerase (PARP) are a family of proteins important for single-strand break (SSB) DNA repair. PARP inhibition results in selective killing of cancer cells which have lost *BRCA1* or *BRCA2* function. Trials have shown effectiveness of PARP inhibition in patients with BRCA mutant ovarian and breast cancers. Both olaparib (ovarian, breast) and rucoparib (ovarian) have been approved for this indication and others are in trials. The concept of synthetic lethality provides a framework for genetic screens to identify other synthetic lethal combinations involving known tumor-suppressor genes, and development of novel therapeutic agents to target dependent pathways. Other unique aspects of malignant tumors, including those outlined elsewhere in the chapter, may also be vulnerable to synthetic lethal interactions.

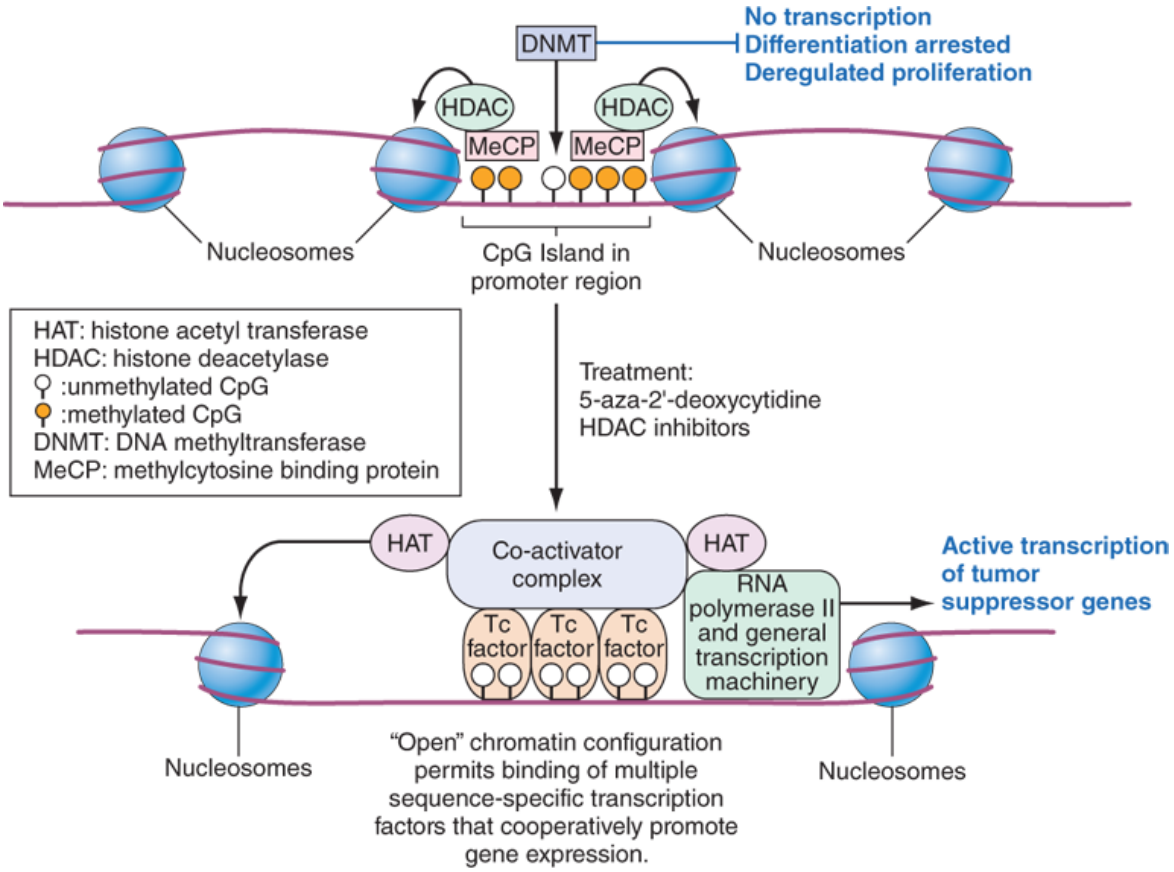
EPIGENETIC INFLUENCES ON CANCER GENE TRANSCRIPTION

Chromatin structure regulates the hierarchical order of sequential gene transcription that governs differentiation and tissue homeostasis. Disruption of chromatin remodeling (the process of modifying chromatin structure to control exposure of specific genes to transcriptional proteins, thereby controlling the expression of those genes) leads to aberrant gene expression that can significantly alter the biology of cells including inducing proliferation or migration of cells. *Epigenetics* is defined as changes that alter the pattern of gene expression that persist across at least one cell division, but are not caused by changes in the DNA code.

Epigenetic changes include alterations of chromatin structure mediated by methylation of cytosine residues of DNA (primarily in context of CpG dinucleotides in somatic cells), modification of histones by altering acetylation or methylation, or changes in higher-order chromosome structure **(Fig. 68-4)**. Appropriate control of DNA methylation is essential for normal cell function and development and both methylation and hypomethylation of histones occurs in cancers. Hypermethylation of DNA promoter regions is a common mechanism by which tumor-suppressor loci are epigenetically silenced in cancer cells. Thus one allele of a tumor suppressor gene may be inactivated by mutation or deletion (as occurs in loss of heterozygosity), while expression of the other allele is epigenetically silenced, usually by methylation leading to loss of gene function. Aberrant hypomethylation is also frequently found in a number of cancers consistent with the dysregulated pattern of gene transcription that is a hallmark of cancer cells with some genes being inappropriately turned off while others are inappropriately turned on.

FIGURE 68-4

Epigenetic regulation of gene expression in cancer cells. Tumor-suppressor genes are often epigenetically silenced in cancer cells. In the upper portion, a CpG island within the promoter and enhancer regions of the gene has been methylated, resulting in the recruitment of methyl-cytosine binding proteins (MeCP) and complexes with histone deacetylase (HDAC) activity. Chromatin is in a condensed, nonpermissive conformation that inhibits transcription. Clinical trials are under way utilizing the combination of demethylating agents such as 5-aza-2′-deoxycytidine plus HDAC inhibitors, which together confer an open, permissive chromatin structure (*lower portion*). Transcription factors bind to specific DNA sequences in promoter regions and, through protein-protein interactions, recruit coactivator complexes containing histone acetyl transferase (HAT) activity. This enhances transcription initiation by RNA polymerase II and associated general transcription factors. The expression of the tumor-suppressor gene commences, with phenotypic changes that may include growth arrest, differentiation, or apoptosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Acetylation of the amino terminus of the core histones H3 and H4 induces an open chromatin conformation that promotes transcription initiation. Histone acetylases are components of coactivator complexes recruited to promoter/enhancer regions by sequence-specific transcription factors during the activation of genes (Fig. 68-4). Histone deacetylases (HDACs; multiple HDACs are encoded in the human genome)

are recruited to genes by transcriptional repressors and prevent the initiation of gene transcription. Methylated cytosine residues in promoter regions become associated with methyl cytosine–binding proteins that recruit protein complexes with HDAC activity. The balance between permissive and inhibitory chromatin structure is therefore largely determined by the activity of transcription factors in modulating the "histone code" and the methylation status of the genetic regulatory elements of genes.

The pattern of gene transcription is aberrant in all human cancers, and in many cases, epigenetic events are responsible. Epigenetic events play a critical role in carcinogenesis (e.g., long-lasting changes in methylation induced by smoking) and are found in pre-malignant lesions. Unlike genetic events that alter DNA primary structure (e.g., deletions), epigenetic changes are potentially reversible and appear amenable to therapeutic intervention. In certain human cancers, including a subset of pancreatic cancers and multiple myeloma, the $p16^{Ink4a}$ promoter is inactivated by methylation, thus permitting the unchecked activity of CDK4/cyclin D and rendering pRb nonfunctional. In sporadic forms of renal, breast, and colon cancer, the von Hippel–Lindau (*VHL*), breast cancer 1 (*BRCA1*), and serine/threonine kinase 11 (*STK11*) genes, respectively, can be epigenetically silenced. Other targeted genes include the $p15^{Ink4b}$ CDK inhibitor, glutathione-S-transferase (which detoxifies reactive oxygen species [ROS]), and the E-cadherin molecule (important for junction formation between epithelial cells). Epigenetic silencing can affect genes involved in DNA repair, thus predisposing to further genetic damage. Examples include MLH1 (mut L homologue in sporadic colon cancers that have microsatellite instability) and MSH2 in a subset of hereditary non-polyposis colon cancer patients who have a mutation in the 3' end of epithelial cell adhesion molecule (EPCAM). These are critical genes involved in repair of mismatched bases that occur during DNA synthesis and their silencing can lead to mutations in the DNA.

Human leukemias often have chromosomal translocations that code for novel fusion proteins with activities that alter chromatin structure by interacting with HDACs or histone acetyl transferases (HATs). For example, the promyelocytic leukemia–retinoic acid receptorα (PML-RARα) fusion protein, generated by the t(15;17) translocation observed in most cases of acute promyelocytic leukemia (APL), binds to promoters containing retinoic acid response elements and recruits HDACs to these promoters, effectively inhibiting gene expression. This arrests differentiation at the promyelocyte stage and promotes tumor cell proliferation and survival. Treatment with pharmacologic doses of all-*trans* retinoic acid (ATRA), the ligand for RARα, results in the release of HDAC activity and the recruitment of coactivators, which overcome the differentiation block. This induced differentiation of APL cells has improved treatment of these patients but also has led to a novel treatment toxicity when newly differentiated tumor cells infiltrate the lungs. ATRA represents a treatment paradigm for the reversal of epigenetic changes in cancer. Other leukemia-associated fusion proteins, such as Tel-acute myeloid leukemia (AML1), AML1-eight-twenty-one (ETO), and the MLL fusion proteins seen in AML and acute lymphocytic leukemia, also lead to repression through the HDAC complex. Therefore, efforts are ongoing to determine the structural basis for interactions between translocation fusion proteins and chromatin-remodeling proteins and to use this information to rationally design small molecules that will disrupt specific protein-protein associations, although this has proven to be technically difficult. Several drugs that block the enzymatic activity of HDACs are approved for cancer treatment and others are being tested. HDAC inhibitors have demonstrated sufficient antitumor activity against cutaneous T-cell lymphoma (vorinostat, romidepsin), peripheral T-cell lymphoma (romidepsin, belinostat), and multiple myeloma (panobinostat) to be approved by the FDA.

HDAC inhibitors (HDACi) have also demonstrated antitumor activity in clinical studies against some solid tumors and additional studies are ongoing. HDACi may target cancer cells via a number of mechanisms including both epigenetic modulation via histone acetylation as well as effects on other proteins which are acetylated. Some of HDACi's pleiotropic effects include: enhancement of apoptosis by upregulation of a number of proteins that enhance apoptosis including death receptors (DR4/5, FAS, and their ligands) and downregulation of proteins that inhibit apoptosis (e.g., X-linked inhibitor of apoptosis (XIAP); upregulation of proteins that inhibit cell cycle progression (e.g., p21Cip1/Waf1); inhibition of DNA repair and generation of ROS leading to increased DNA damage; and disruption of the chaperone protein HSP90.

Efforts are also under way to modulate other epigenetic processes such as reversing the hypermethylation of CpG islands that characterizes many malignancies. Drugs that induce DNA demethylation, such as 5-aza-2-deoxycytidine, can lead to reexpression of silenced genes in cancer cells with restoration of function, and 5-aza-2-deoxycytidine is approved for use in myelodysplastic syndrome (MDS). However, 5-aza-2-deoxycytidine has limited aqueous solubility and is myelosuppressive limiting its usefulness. Other inhibitors of DNA methyltransferases are in development. In ongoing clinical trials, inhibitors of DNA methylation are being combined with HDAC inhibitors, with the idea that reversing coexisting epigenetic changes will reverse the deregulated patterns of gene transcription in cancer cells. Epigenetic gene regulation can also occur via microRNAs or long non-coding RNAs (lncRNA). MicroRNAs are short (average 22 nucleotides in length) RNA molecules that silence gene expression after transcription by binding and inhibiting the translation or promoting the degradation of mRNA transcripts. It is estimated that >1000 microRNAs are encoded in the human genome. Each tissue has a distinctive repertoire of microRNA expression and this pattern is altered in specific ways in cancers. Specific correlations between microRNA expression and tumor biology and clinical behavior are just now emerging. Therapies targeting microRNAs are not currently at hand but represent an ongoing area of treatment development. LncRNAs are longer than 200 nucleotides and comprise the largest group of noncoding RNAs. Some of them have been shown to play important roles in gene regulation. The

potential for altering these RNAs for therapeutic benefit is an area of active investigation, although much more needs to be learned before this will be feasible.

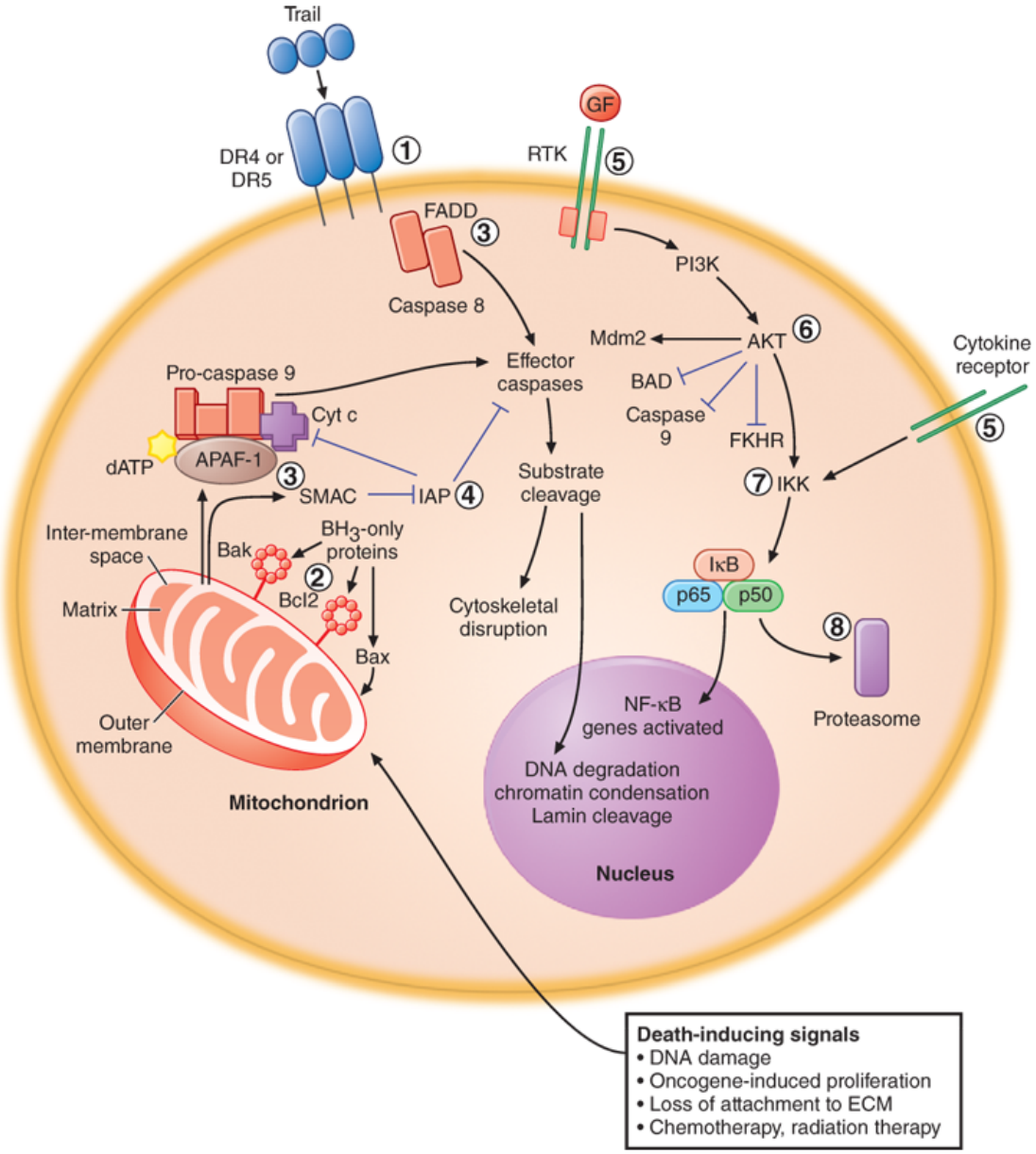
APOPTOSIS AND OTHER MECHANISMS OF CELL DEATH

Tissue homeostasis requires a balance between the death of aged, terminally differentiated cells or severely damaged cells and their renewal by proliferation of committed progenitors. Genetic damage to growth-regulating genes of stem cells could lead to catastrophic results for the host as a whole. Thus, genetic events causing activation of oncogenes or loss of tumor suppressors, which would be predicted to lead to unregulated cell proliferation unless corrected, usually activate signal transduction pathways that block aberrant cell proliferation. These pathways can lead to a form of programmed cell death (*apoptosis*) or irreversible growth arrest (*senescence*). Much as a panoply of intra- and extracellular signals impinge upon the core cell cycle machinery to regulate cell division, so too these signals are transmitted to a core enzymatic machinery that regulates cell death and survival.

Apoptosis is a tightly regulated process induced by two main pathways **(Fig. 68-5)**. The extrinsic pathway of apoptosis is activated by cross-linking members of the tumor necrosis factor (TNF) receptor superfamily, such as CD95 (Fas) and death receptors DR4 and DR5, by their ligands, Fas ligand or TRAIL (TNF-related apoptosis-inducing ligand), respectively. This induces the association of FADD (Fas-associated death domain) and procaspase-8 to death domain motifs of the receptors. Caspase-8 is activated and then cleaves and activates effector caspases-3 and -7, which then target cellular constituents (including caspase-activated DNAse, cytoskeletal proteins, and a number of regulatory proteins), inducing the morphologic appearance characteristic of apoptosis, which pathologists term “karyorrhexis.” The intrinsic pathway of apoptosis is initiated by the release of cytochrome *c* and SMAC (second mitochondrial activator of caspases) from the mitochondrial intermembrane space in response to a variety of noxious stimuli, including DNA damage, loss of adherence to the extracellular matrix (ECM), oncogene-induced proliferation, and growth factor deprivation. Upon release into the cytoplasm, cytochrome *c* associates with dATP, procaspase-9, and the adaptor protein APAF-1, leading to the sequential activation of caspase-9 and effector caspases. SMAC binds to and blocks the function of inhibitor of apoptosis proteins (IAP), negative regulators of caspase activation.

FIGURE 68-5

Therapeutic strategies to overcome aberrant survival pathways in cancer cells. *1.* The extrinsic pathway of apoptosis can be selectively induced in cancer cells by TRAIL (the ligand for death receptors 4 and 5) or by agonistic monoclonal antibodies. *2.* Inhibition of antiapoptotic Bcl-2 family members with antisense oligonucleotides or inhibitors of the BH_3-binding pocket will promote formation of Bak- or Bax-induced pores in the mitochondrial outer membrane. *3.* Epigenetic silencing of APAF-1, caspase-8, and other proteins can be overcome using demethylating agents and inhibitors of histone deacetylases. *4.* Inhibitor of apoptosis proteins (IAP) blocks activation of caspases; small-molecule inhibitors of IAP function (mimicking SMAC action) should lower the threshold for apoptosis. *5.* Signal transduction pathways originating with activation of receptor tyrosine kinase receptors (RTKs) or cytokine receptors promote survival of cancer cells by a number of mechanisms. Inhibiting receptor function with monoclonal antibodies, such as trastuzumab or cetuximab, or inhibiting kinase activity with small-molecule inhibitors can block the pathway. *6.* The Akt kinase phosphorylates many regulators of apoptosis to promote cell survival; inhibitors of Akt may render tumor cells more sensitive to apoptosis-inducing signals; however, the possibility of toxicity to normal cells may limit the therapeutic value of these agents. *7* and *8.* Activation of the transcription factor NF-κB (composed of p65 and p50 subunits) occurs when its inhibitor, IκB, is phosphorylated by IκB-kinase (IKK), with subsequent degradation of IκB by the proteasome. Inhibition of IKK activity should selectively block the activation of NF-κB target genes, many of which promote cell survival. Inhibitors of proteasome function are FDA-approved and may work in part by preventing destruction of IκB, thus blocking NF-κB nuclear localization. NF-κB is unlikely to be the only target for proteasome inhibitors.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The release of apoptosis-inducing proteins from the mitochondria is regulated by pro- and antiapoptotic members of the Bcl-2 family. Antiapoptotic members (e.g., Bcl-2, Bcl-XL, and Mcl-1) associate with the mitochondrial outer membrane via their carboxyl termini, exposing to the cytoplasm a hydrophobic binding pocket composed of Bcl-2 homology (BH) domains 1, 2, and 3 that is crucial for their activity. Perturbations of normal physiologic processes in specific cellular compartments lead to the activation of BH3-only proapoptotic family members (such as Bad, Bim, Bid, Puma, Noxa, and others) that can alter the conformation of the outer-membrane proteins Bax and Bak, which then oligomerize to form pores in the mitochondrial outer membrane resulting in cytochrome *c* release. If proteins comprised only by BH3 domains are sequestered by Bcl-2, Bcl-XL, or Mcl-1, pores do not form and apoptosis-inducing proteins are not released from the mitochondria. The ratio of levels of antiapoptotic Bcl-2 family members and the levels of proapoptotic BH3-only proteins at the mitochondrial membrane determines the activation state of the intrinsic pathway. The mitochondrion must therefore be recognized not only as an organelle with vital roles in intermediary metabolism and oxidative phosphorylation but also as a central regulatory structure of the apoptotic process.

The evolution of tumor cells to a more malignant phenotype requires the acquisition of genetic changes that subvert apoptosis pathways and promote cancer cell survival and resistance to anticancer therapies. However, cancer cells may be more vulnerable than normal cells to therapeutic interventions that target the apoptosis pathways that cancer cells depend upon. For instance, overexpression of Bcl-2 as a result of the t(14;18) translocation contributes to follicular lymphoma and it is highly expressed in many lymphoid malignancies including chronic lymphocytic leukemia (CLL). Upregulation of Bcl-2 expression is also observed in other cancers including prostate, breast, and lung cancers and melanoma. Targeting of antiapoptotic Bcl-2 family members has been accomplished by the identification of several low-molecular-weight compounds that bind to the hydrophobic pockets of either Bcl-2 or Bcl-XL and block their ability to associate with death-inducing BH3-only

proteins. These compounds inhibit the antiapoptotic activities of Bcl-2 and Bcl-XL at nanomolar concentrations. An oral BH3 mimetic inhibitor of BCL-2, venetoclax, is approved for use in patients with refractory CLL with 17p deletion.

Preclinical studies targeting death receptors DR4 and -5 have demonstrated that recombinant, soluble, human TRAIL or humanized monoclonal antibodies with agonist activity against DR4 or -5 can induce apoptosis of tumor cells while sparing normal cells. The mechanisms for this selectivity may include expression of decoy receptors or elevated levels of intracellular inhibitors (such as FLIP, which competes with caspase-8 for FADD) by normal cells but not tumor cells. Synergy has been shown between TRAIL-induced apoptosis and chemotherapeutic agents in some preclinical studies. However, clinical studies have not yet shown significant activity of approaches targeting the TRAIL pathway.

Many of the signal transduction pathways perturbed in cancer promote tumor cell survival (Fig. 68-5). These include activation of the PI3K/Akt pathway, increased levels of the NF-κB transcription factor, and epigenetic silencing of genes such as *APAF-1(apoptosis protease activating factor-1 involved in activating caspase-9 and essential for apoptosis)* and *caspase-8*. Each of these pathways is a target for therapeutic agents that, in addition to affecting cancer cell proliferation or gene expression, may render cancer cells more susceptible to apoptosis, thus promoting synergy when combined with other chemotherapeutic agents.

Some tumor cells resist drug-induced apoptosis indirectly by eliminating the noxious stimulus-inducing apoptosis through expression of one or more members of the ABC (ATP-binding casette proteins) family of ATP-dependent efflux pumps that mediate the multidrug-resistance (MDR) phenotype. The prototype member of this family, P-glycoprotein (PGP), spans the plasma membrane 12 times and has two ATP-binding sites. Hydrophobic drugs (e.g., anthracyclines and vinca alkaloids) are recognized by PGP as they enter the cell and are pumped out. Numerous clinical studies have failed to demonstrate that drug resistance can be overcome using inhibitors of PGP. However, ABC transporters have different substrate specificities, and inhibition of a single family member may not be sufficient to overcome the MDR phenotype. Efforts to reverse PGP-mediated drug resistance continue.

Cells, including cancer cells, can also undergo other mechanisms of cell death including *autophagy* (degradation of proteins and organelles by lysosomal proteases) and *necrosis* (digestion of cellular components and rupturing of the cell membrane). Necrosis usually occurs in response to external forces resulting in release of cellular components, which leads to inflammation and damage to surrounding tissues. Although necrosis was thought to be unprogrammed, evidence now suggests that at least some aspects may also be programmed. The exact role of necrosis in cancer cell death in various settings is still being determined. In addition to its role in cell death, autophagy can also serve as a homeostatic mechanism to promote survival for the cell by recycling cellular components to provide necessary energy. The mechanisms that control the balance between enhancing survival versus leading to cell death are still not fully understood. Autophagy appears to play conflicting roles in the development and survival of cancer. Early in the carcinogenic process it can act as a tumor suppressor by preventing the cell from accumulating abnormal proteins and organelles. However, in established tumors, it may serve as a mechanism of survival for cancer cells when they are stressed by damage such as from chemotherapy. Preclinical studies have indicated that inhibition of this process can enhance the sensitivity of cancer cells to chemotherapy and ongoing trials are evaluating inhibitors of autophagy in combination with chemotherapy. Better understanding of the factors that control the survival-promoting versus death-inducing aspects of autophagy is required in order to know how to best manipulate it for therapeutic benefit.

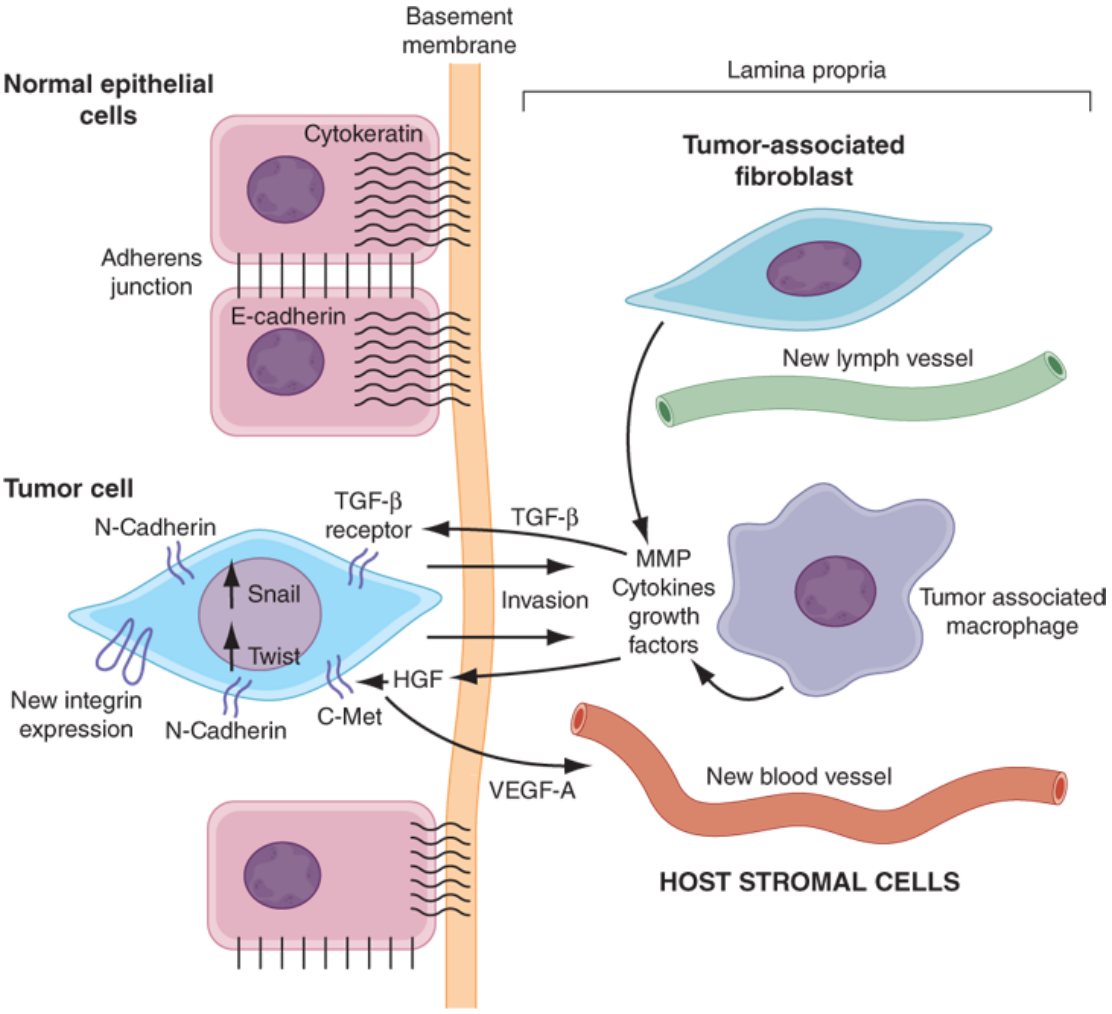
METASTASIS

The metastatic process accounts for the vast majority of deaths from solid tumors and therefore an understanding of this process is critical for improvements in survival from cancer. The biology of metastasis is complex and requires multiple steps. The initial step involves cell migration and invasion through the ECM. The three major features of tissue invasion are cell adhesion to the basement membrane, local proteolysis of the membrane, and movement of the cell through the rent in the membrane and the ECM. Cells that lose contact with the ECM normally undergo programmed cell death (anoikis-apoptosis induced by the loss of contact) and this process has to be suppressed in cells that metastasize. Another process important for many, but not necessarily all, metastasizing epithelial cancer cells is epithelial-mesenchymal transition (EMT). This is a process by which cells lose their epithelial properties and gain mesenchymal properties. This normally occurs during the developmental process in embryos, allowing cells to migrate to their appropriate destinations in the embryo. It also occurs in wound healing, tissue regeneration, and fibrotic reactions, but in all of these processes, cells stop proliferating when the process is complete. Malignant cells that metastasize often undergo EMT as an important step in that process but retain the capacity for unregulated proliferation. However, there is evidence that not all metastasizing cancer cells require EMT, and the exact role of EMT in different metastasizing cancer cells continues to be elucidated. Malignant cells that gain access to the circulation must then repeat those steps at a remote site, find a hospitable niche in a foreign tissue, avoid detection and elimination by host defenses including the immune system, and induce the growth of new blood vessels. Some metastatic cells occur as oligoclonal clusters which appear to be more potent in establishing metastasis than single cells, perhaps through differential and cooperative effects in evading host defenses. The rate-limiting step for metastasis is the ability for tumor cells to survive and expand in the novel microenvironment of the metastatic site, and multiple host-tumor interactions determine the ultimate outcome (Fig. 68-6).

Few drugs have been developed to attempt to directly target the process of metastasis, in part because the specifics of the critical steps in the process that would be potentially good targets for drugs are still being identified. However, a number of potential targets are known. HER2 can enhance the metastatic potential of breast cancer cells and as discussed above, the monoclonal antibody trastuzumab which targets HER2, improves survival in the adjuvant setting for HER2+ breast cancer patients. A number of other potential targets that increase metastatic potential of cells in preclinical studies include: HIF-1 and 2, transcription factors induced by hypoxia within tumors; growth factors (e.g., cMET and VEGFR); oncogenes (e.g., SRC); adhesion molecules (e.g., focal adhesion kinase, FAK); ECM proteins (e.g., matrix metaloproteinases 1 and 2); and inflammatory molecules (e.g., COX-2).

FIGURE 68-6

Oncogene signaling pathways are activated during tumor progression and promote metastatic potential. This figure shows a cancer cell that has undergone epithelial to mesenchymal transition (EMT) under the influence of several environmental signals. Critical components include activated transforming growth factor beta (TGF-β) and the hepatocyte growth factor (HGF)/c-Met pathways, as well as changes in the expression of adhesion molecules that mediate cell-cell and cell–extracellular matrix interactions. Important changes in gene expression are mediated by the Snail and Twist family of transcriptional repressors (whose expression is induced by the oncogenic pathways), leading to reduced expression of E-cadherin, a key component of adherens junctions between epithelial cells. This, in conjunction with upregulation of N-cadherin, a change in the pattern of expression of integrins (which mediate cell–extracellular matrix associations that are important for cell motility), and a switch in intermediate filament expression from cytokeratin to vimentin, results in the phenotypic change from adherent highly organized epithelial cells to motile and invasive cells with a fibroblast or mesenchymal morphology. EMT is thought to be an important step leading to metastasis in some human cancers. Host stromal cells, including tumor-associated fibroblasts and macrophages, play an important role in modulating tumor cell behavior through secretion of growth factors and proangiogenic cytokines, and matrix metalloproteinases that degrade the basement membrane. VEGF-A, -C, and -D are produced by tumor cells and stromal cells in response to hypoxemia or oncogenic signals, and induce production of new blood vessels and lymphatic channels through which tumor cells metastasize to lymph nodes or tissues.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The metastatic phenotype is likely restricted to a small fraction of tumor cells (Fig. 68-6). A number of genetic and epigenetic changes are required for tumor cells to be able to metastasize, including activation of metastatic-promoting genes and inhibition of genes that suppress the metastatic ability. Given the role of microRNAs in controlling gene expression (see epigenetic section) including those critical to the metastatic process, efforts are under way to modulate these to try to inhibit metastasis. Cells with metastatic capability frequently express chemokine receptors that are likely important in the metastatic process. A number of candidate metastasis-suppressor genes have been identified, including genes coding for proteins that enhance apoptosis, suppress cell division, are involved in the interactions of cells with each other or the ECM, or suppress cell migration. The loss of function of these genes enhances metastasis. Gene expression profiling is being used to study the metastatic process and other properties of tumor cells that may predict susceptibilities.

An example of the ability of malignant cells to survive and grow in a novel microenvironment is bone metastases. Bone metastases can be extremely painful, cause fractures of weight-bearing bones, can lead to hypercalcemia, and are a major cause of morbidity for cancer patients. Osteoclasts and their monocyte-derived precursors express the surface receptor RANK (receptor activator of NF-κB), which is required for terminal differentiation and activation of osteoclasts. Osteoblasts and other stromal cells express RANK ligand (RANKL), as both a membrane-bound and soluble cytokine. Osteoprotegerin (OPG), a soluble receptor for RANKL produced by stromal cells, acts as a decoy receptor to inhibit RANK activation. The relative balance of RANKL and OPG determines the activation state of RANK on osteoclasts. Many tumors increase osteoclast activity by secretion of substances such as parathyroid hormone (PTH), PTH-related peptide, interleukin (IL)-1, or Mip1 that perturb the homeostatic balance of bone remodeling by increasing RANK signaling. One example is multiple myeloma, where tumor cell–stromal cell interactions activate osteoclasts and inhibit osteoblasts, leading to the development of multiple lytic bone lesions. Inhibition of RANK ligand by an antibody (denosumab) can prevent further bone destruction. Bisphosphonates are also effective inhibitors of osteoclast function that are used in the treatment of cancer patients with bone metastases.

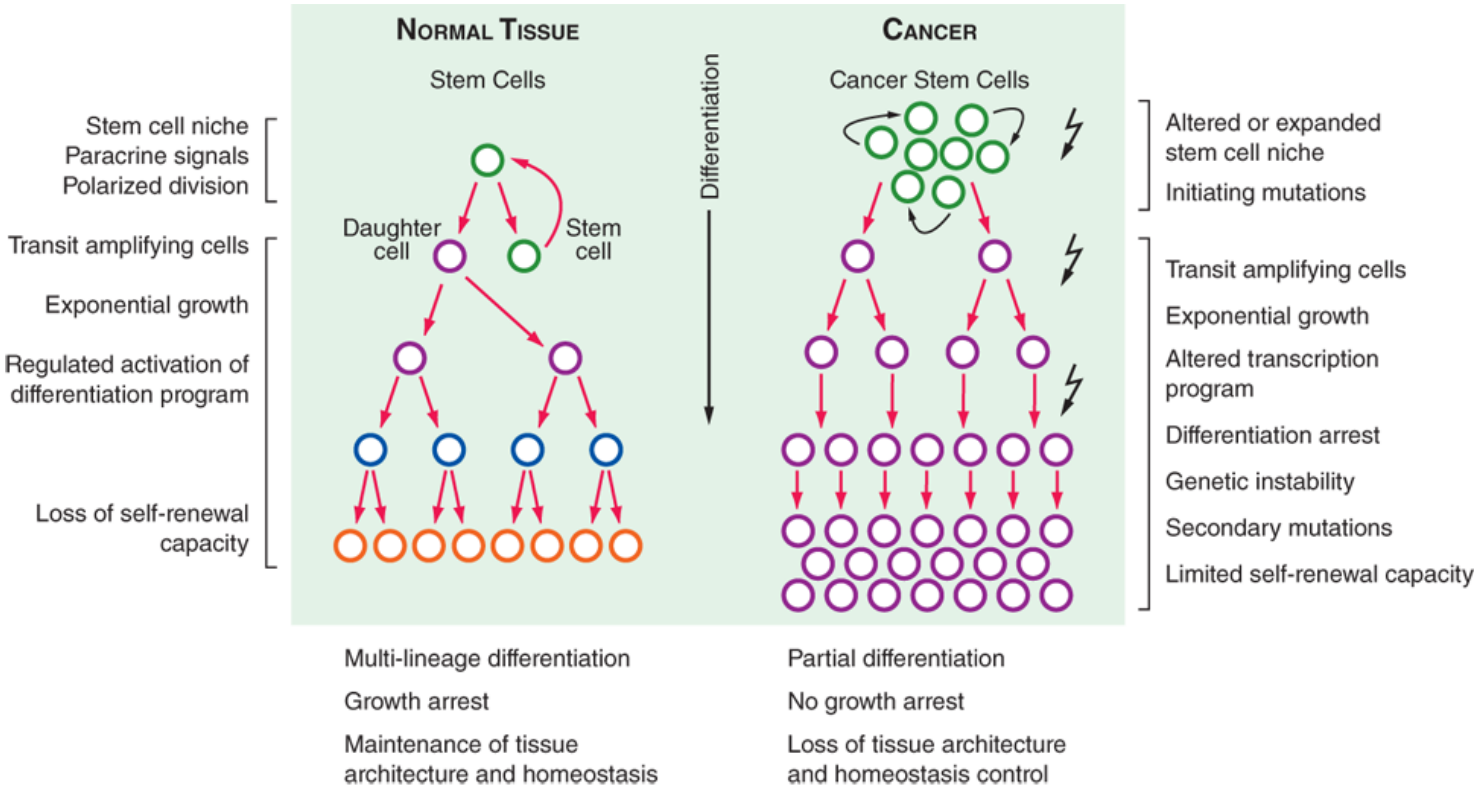
CANCER STEM CELLS

Normal tissues have stem cells capable of self-renewal and repairing damaged tissue whereas the majority of cells in normal tissues do not have this capacity. Similarly, only a small proportion of the cells within a tumor are capable of initiating colonies in vitro or forming tumors at high efficiency when injected into immunocompromised NOD/SCID mice. For example, acute and chronic myeloid leukemias (AML and CML) have a small population of cells (estimated to be <1%) that have properties of stem cells, such as unlimited self-renewal and the capacity to cause leukemia when serially transplanted in mice. These cells have an undifferentiated phenotype (Thy1–CD34+CD38– and do not express other differentiation markers) and resemble normal stem cells in many ways, but are no longer under homeostatic control **(Fig. 68-7)**. Solid tumors may also contain a population of stem cells. It is not yet known how often cancers may originate within a stem cell population. Cancer stem cells, like their normal counterparts, have unlimited proliferative capacity and paradoxically traverse the cell cycle at a slow rate; cancer growth occurs largely due to expansion of the stem cell pool, the unregulated proliferation of an amplifying population, and failure of apoptosis pathways (Fig. 68-7). Slow cell cycle progression and high levels of expression of antiapoptotic Bcl-2 family members and drug efflux pumps of the MDR family render cancer stem cells less vulnerable to cancer chemotherapy or radiation therapy. Implicit in the cancer stem cell hypothesis is the idea that failure to cure most human cancers is due to the fact that current therapeutic agents do not kill the stem cells. Identification and isolation of cancer stem cells will allow determination of the aberrant signaling pathways that distinguish these cells from normal tissue stem cells. These would serve as potential therapeutic targets. Evidence that cells with stem cell properties can arise from other epithelial cells within the cancer by processes such as epithelial mesenchymal transition also implies that it is essential to treat all of the cancer cells, and not just those with current stem cell-like properties, in order to eliminate the self -renewing cancer cell population. The exact nature of cancer stem cells remains an area of investigation. One of the unanswered questions is the exact origin of cancer stem cells for the different cancers.

FIGURE 68-7

Cancer stem cells play a critical role in the initiation, progression, and resistance to therapy of malignant neoplasms. In normal tissues (*left*), homeostasis is maintained by asymmetric division of stem cells leading to one progeny cell that will differentiate and one cell that will maintain the stem cell pool. This occurs within highly specific niches unique to each tissue, such as in close apposition to osteoblasts in bone marrow, or at the base of crypts in the colon. Here, paracrine signals from stromal cells, such as sonic hedgehog or Notch-ligands, as well as upregulation of β-catenin and telomerase, help to maintain stem cell features of unlimited self-renewal while preventing differentiation or cell death. This occurs in part through upregulation of the transcriptional repressor Bmi-1 and inhibition of the $p16^{Ink4a}$/Arf and p53 pathways. Daughter cells leave the stem cells niche and enter a proliferative phase (referred to as *transit-amplifying*) for a specified number of cell divisions, during which time a developmental program is activated, eventually giving rise to fully differentiated cells that have lost proliferative potential. Cell renewal equals cell death, and homeostasis is maintained. In this hierarchical system, only stem cells are long-lived. The hypothesis is that cancers harbor stem cells that make up a small fraction (i.e., 0.001–1%) of all cancer cells. These cells share several features with normal stem cells, including an undifferentiated phenotype, unlimited self-renewal potential, a capacity for some degree of differentiation; however, due to initiating mutations (mutations are indicated by lightning bolts), they are no longer regulated by environmental cues. The cancer stem cell pool

is expanded, and rapidly proliferating progeny, through additional mutations, may attain stem cell properties, although most of this population is thought to have a limited proliferative capacity. Differentiation programs are dysfunctional due to reprogramming of the pattern of gene transcription by oncogenic signaling pathways. Within the cancer transit-amplifying population, genomic instability generates aneuploidy and clonal heterogeneity as cells attain a fully malignant phenotype with metastatic potential. The cancer stem cell hypothesis has led to the idea that current cancer therapies may be effective at killing the bulk of tumor cells but do not kill tumor stem cells, leading to a regrowth of tumors that is manifested as tumor recurrence or disease progression. Research is in progress to identify unique molecular features of cancer stem cells that can lead to their direct targeting by novel therapeutic agents.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PLASTICITY AND RESISTANCE

Cancer cells, and especially stem cells, have the capacity for significant plasticity allowing them to alter multiple aspects of cell biology in response to external factors (e.g., chemotherapy, radiation therapy, inflammation, immune response). In addition, heterogeneity between the different clones of cells within the tumor population and their interactions with each other and the tumor microenvironment provides the tumor with the capacity for significant plasticity in dealing with both internal and external stresses. Thus, a major problem in cancer therapy is that malignancies have a wide spectrum of mechanisms for both initial and adaptive resistance to treatments. These include inhibiting drug delivery to the cancer cells, blocking drug uptake and retention, increasing drug metabolism, altering levels of target proteins making them less sensitive to drugs, acquiring mutations in target proteins making them no longer sensitive to the drug, modifying metabolism and cell signaling pathways, using alternate signaling pathways, adjusting the cell replication process including mechanisms by which the cell deals with DNA damage, inhibiting apoptosis, and evading the immune system. Thus, most metastatic cancers (except those curable with chemotherapy such as germ cell tumors) eventually become resistant to the therapy being utilized. Overcoming resistance is a major area of research.

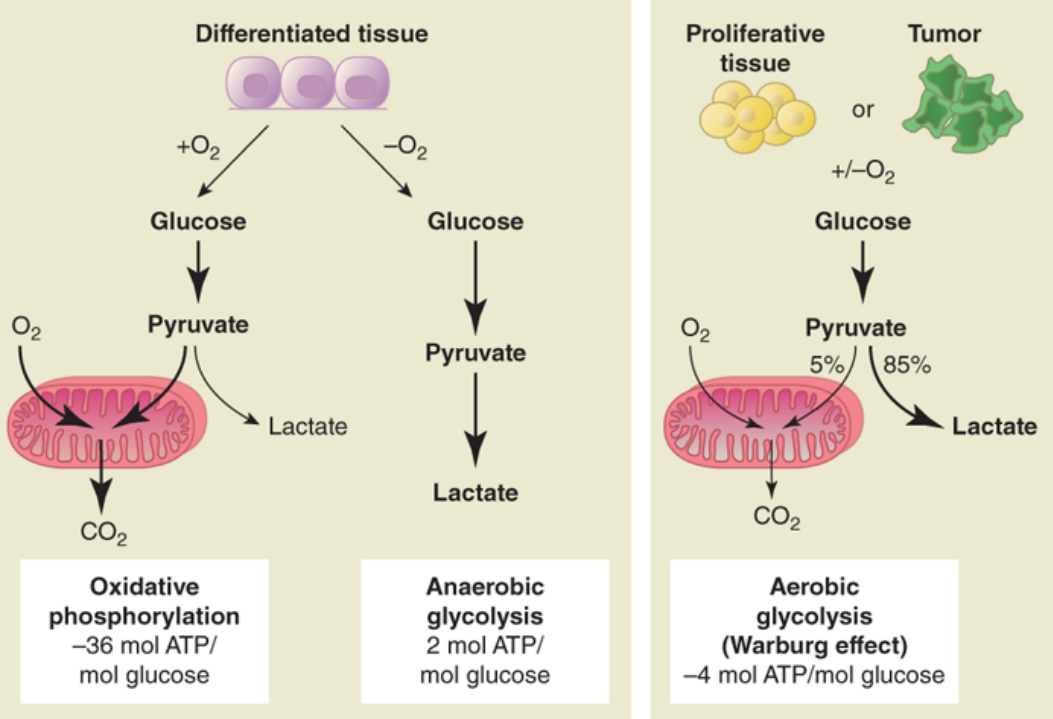
CANCER METABOLISM

One of the distinguishing characteristics of cancer cells is that they have altered metabolism as compared with normal cells in supporting survival, their high rates of proliferation, and ability to metastasize. Complicating studies evaluating metabolic differences between normal and malignant cells is that there is heterogeneity in metabolism between different cells within a cancer. Malignant cells must focus a significant fraction of their energy resources into synthesis of proteins and other molecules (building blocks required for the production of new cells) while still maintaining sufficient ATP production to survive and grow. Although normal proliferating cells also have similar needs, there are differences in how cancer cells metabolize glucose and a number of other compounds including the amino acid glutamine as compared to normal cells in

part because of genetic and epigenetic changes within cancer cells but also likely due to differences in the environments of cancer and normal cells. Many cancer cells utilize aerobic glycolysis (the Warburg effect) **(Fig. 68-8)** to metabolize glucose leading to increased lactic acid production whereas normal cells utilize oxidative phosphorylation in mitochondria under aerobic conditions, a much more efficient process for generating ATP for energy utilization but one that does not produce the same level of building blocks needed for new cells. One consequence is increased glucose uptake by cancer cells, a fact utilized in fluorodeoxyglucose (FDG)-positron emission tomography (PET) scanning to detect tumors. A number of proteins in cancer cells, including cMYC, HIF1, RAS, p53, pRB, and AKT are all involved in modulating glycolytic processes and controlling the Warburg effect. Although these pathways remain difficult to target therapeutically, both the pI3kinase pathway with signaling through mTOR and the AMP-activated kinase (AMPK) pathway that inhibits mTORC1 (a protein complex that includes mTOR) are important in controlling the glycolytic process and thus provide potential targets for inhibiting this process. An inhibitor of MTOR is approved for use against RCC (temsirolimus) and another inhibitor (everolimus) has activity against breast, neuroendocrine, and RCC. Other MTOR inhibitors are in trials and modulators of AMPK are being investigated. The inefficient utilization of glucose by malignant cells also leads to a need for alternative metabolic pathways for other compounds as well, one of which is glutamine. Similar to glucose, this provides both a source for structural molecules as well as energy production. Similarly to glucose, glutamine is also inefficiently utilized by cancer cells. Cancer cells can also take up nutrients released by surrounding cells and tissues increasing the complexity of successfully therapeutically inhibiting metabolism in cancer.

FIGURE 68-8

Warburg vs oxidative phosphorylation. In most normal tissues, the vast majority of cells are differentiated and dedicated to a particular function within the organ in which they reside. The metabolic needs are mainly for energy and not for building blocks for new cells. In these tissues, ATP is generated by oxidative phosphorylation that efficiently generates about 36 molecules of ATP for each molecule of glucose metabolized. By contrast, proliferative tumor tissues, especially in the setting of hypoxia, a typical condition within tumors, use aerobic glycolysis to generate energy for cell survival and generation of building blocks for new cells.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Mutations in genes involved in the metabolic process occur in a number of cancers. Among the most frequently found to date are mutations in isocitrate dehydrogenases 1 and 2 (IDH1 and 2). These have been most commonly seen in gliomas, acute myeloid leukemias (AML), and intra-hepatic cholangiocarcinomas. These mutations lead to the production of an oncometabolite (2-hydroxyglutarate, 2HG) instead of the normal product α-ketoglutarate. Although the exact mechanisms of oncogenesis by 2HG are still being elucidated, α-ketoglutarate is a key cofactor for a number of dioxygenases involved in controlling DNA methylation. 2HG can act as a competitive inhibitor for α-ketoglutarate leading to alterations in methylation status (primarily hypermethylation) of genes (leading to epigenetic changes) that can have profound effects on a number of cellular processes including differentiation. Inhibitors of mutants IDH1 and IDH2 are being developed. To date, they have had some activity against IDH mutant AML but less activity against glioblastomas or cholangiocarcinomas.

Much needs to be learned about the specific differences in metabolism between cancer cells and normal cells; however, even with the currently limited state of knowledge, modulators of metabolism are being tested clinically. The first of these is the anti-diabetic agent metformin, both alone and in combination with chemotherapeutic agents. Metformin inhibits gluconeogenesis and may have direct effects on tumor cells by activating the 5'-adenosine monophosphate-activated kinase (AMPK), a serine/threonine protein kinase which is downstream of the LKB1 tumor suppressor, and thus inhibiting mammalian target of rapamycin complex 1 (mTORC1). This leads to decreased protein synthesis and proliferation. Studies to date have not yet established metformin to have a clear role as an anticancer agent. Additional approaches being evaluated include other modulators of glucose metabolism (e.g., piaglitazone) and inhibiting glutaminase (important for glutamine utilization).

TUMOR MICROENVIRONMENT, ANGIOGENESIS, AND IMMUNE EVASION

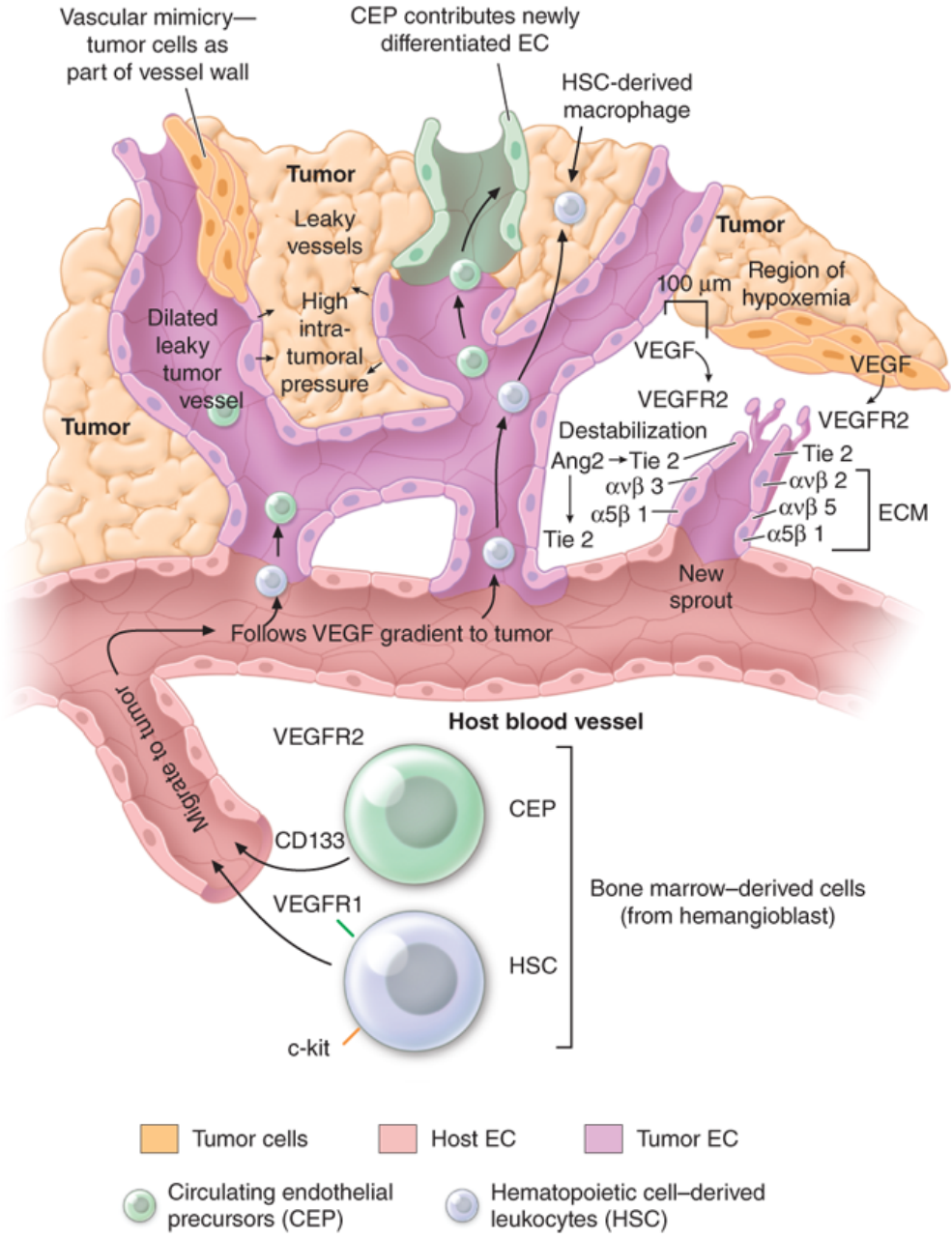
Tumors consist not only of malignant cells but also of a complex microenvironment including many other types of cells (including inflammatory cells), ECM, secreted factors (including growth factors), reactive oxygen and nitrogen species, mechanical factors, blood vessels, and lymphatics. This microenvironment is not static but rather is dynamic and continually evolving. Both the complexity and dynamic nature of the microenvironment enhance the difficulty of treating tumors. There are also a number of mechanisms by which the microenvironment can contribute to resistance to anti-cancer therapies.

One of the critical elements of tumor cell proliferation is delivery of oxygen, nutrients, and circulating factors important for growth and survival. The diffusion limit for oxygen in tissues is ~100–200 μm and thus a critical aspect in the growth of tumors is the development of new blood vessels, or angiogenesis. The growth of primary and metastatic tumors to larger than a few millimeters requires the recruitment of blood vessels and vascular endothelial cells (ECs) to support their metabolic requirements. Thus, a critical element in growth of primary tumors and formation of metastatic sites is the *angiogenic switch*: the ability of the tumor to promote the formation of new capillaries from preexisting host vessels. The angiogenic switch is a phase in tumor development when the dynamic balance of pro- and antiangiogenic factors is tipped in favor of vessel formation by the effects of the tumor on its immediate environment. Stimuli for tumor angiogenesis include hypoxemia, inflammation, and genetic lesions in oncogenes or tumor suppressors that alter tumor cell gene expression. Angiogenesis consists of several steps, including the stimulation of ECs by growth factors, degradation of the ECM by proteases, proliferation and migration of ECs into the tumor, and the eventual formation of new capillary tubes.

Tumor blood vessels are not normal; they have chaotic architecture and blood flow. Due to an imbalance of angiogenic regulators such as VEGF and angiopoietins (see below), tumor vessels are tortuous and dilated with an uneven diameter, excessive branching, and shunting. Tumor blood flow is variable, with areas of hypoxemia and acidosis leading to the selection of variants that are resistant to hypoxemia-induced apoptosis (often due to the loss of p53 expression). Tumor vessel walls have numerous openings, widened interendothelial junctions, and discontinuous or absent basement membrane; this contributes to the high vascular permeability of these vessels and, together with lack of functional intratumoral lymphatics, causes increased interstitial pressure within the tumor (which also interferes with the delivery of therapeutics to the tumor; **Figs. 68-9**, **68-10, and 68-11**). Tumor blood vessels lack perivascular cells such as pericytes and smooth-muscle cells that normally regulate flow in response to tissue metabolic needs.

FIGURE 68-9

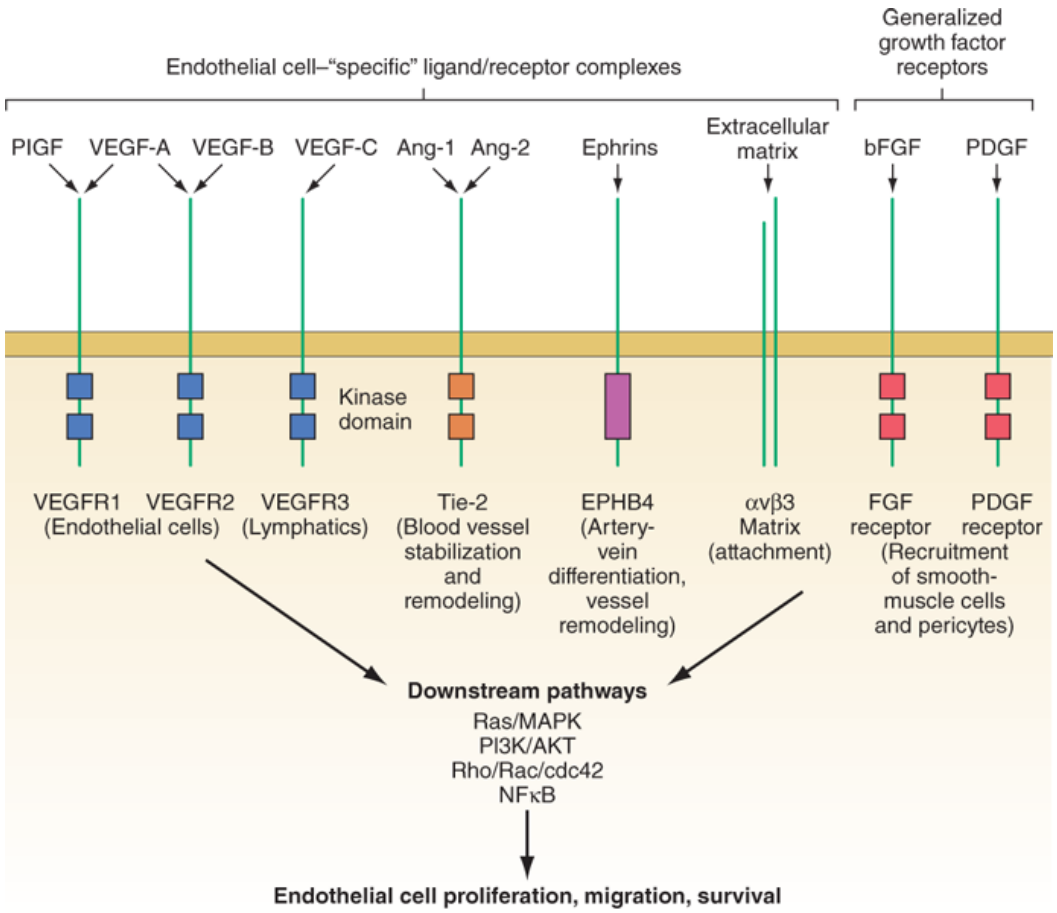
Tumor angiogenesis is a complex process involving many different cell types that must proliferate, migrate, invade, and differentiate in response to signals from the tumor microenvironment. Endothelial cells (ECs) sprout from host vessels in response to VEGF, bFGF, Ang2, and other proangiogenic stimuli. Sprouting is stimulated by VEGF/VEGFR2, Ang2/Tie-2, and integrin/extracellular matrix (ECM) interactions. Bone marrow–derived circulating endothelial precursors (CEPs) migrate to the tumor in response to VEGF and differentiate into ECs, while hematopoietic stem cells differentiate into leukocytes, including tumor-associated macrophages that secrete angiogenic growth factors and produce MMPs that remodel the ECM and release bound growth factors. Tumor cells themselves may directly form parts of vascular channels within tumors. The pattern of vessel formation is haphazard: vessels are tortuous, dilated, leaky, and branch in random ways. This leads to uneven blood flow within the tumor, with areas of acidosis and hypoxemia (which stimulate release of angiogenic factors) and high intratumoral pressures that inhibit delivery of therapeutic agents.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 68-10

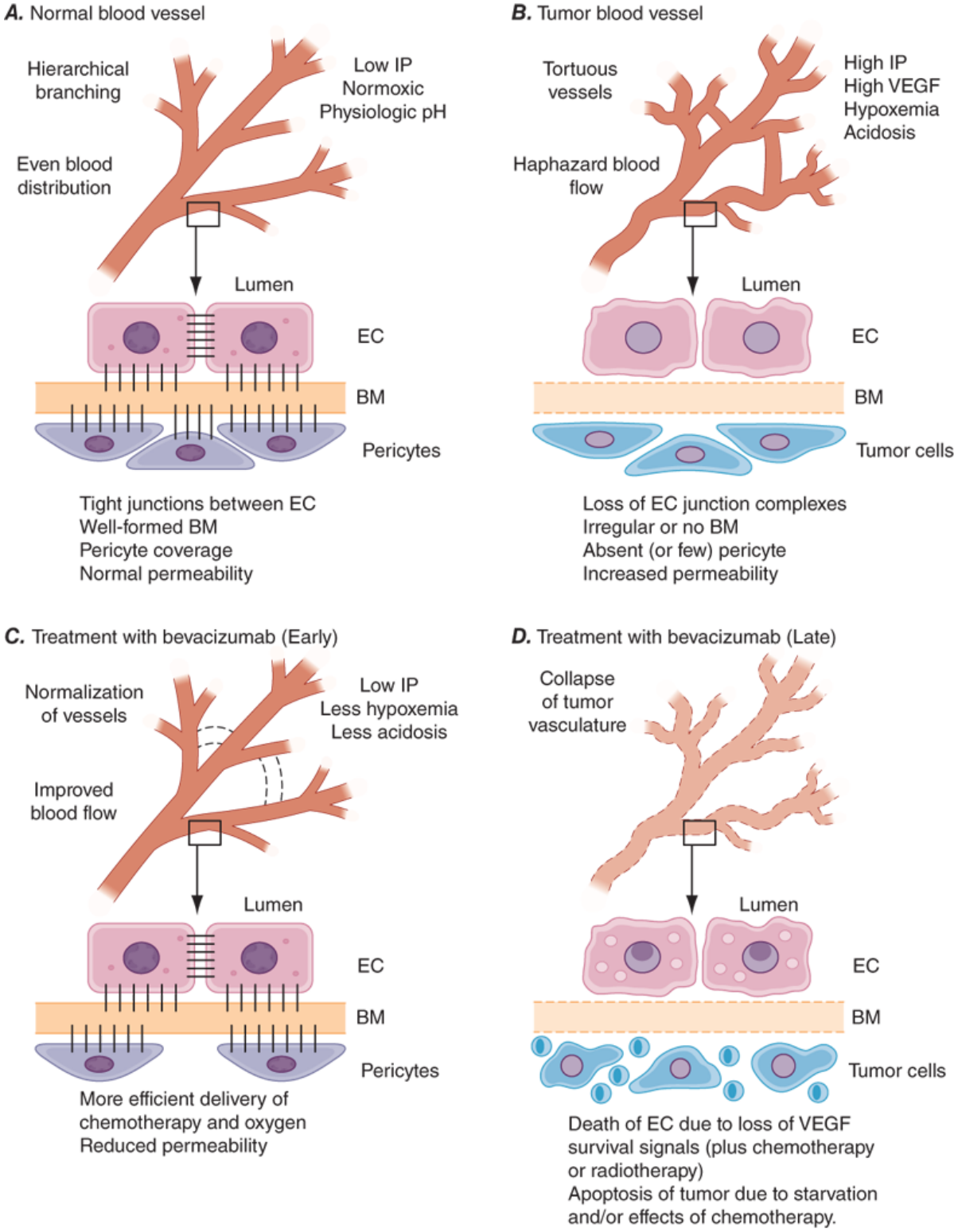
Critical molecular determinants of endothelial cell biology. Angiogenic endothelium expresses a number of receptors not found on resting endothelium. These include receptor tyrosine kinases (RTKs) and integrins that bind to the extracellular matrix and mediate endothelial cell (EC) adhesion, migration, and invasion. ECs also express RTK (i.e., the fibroblast growth factor [FGF] and platelet-derived growth factor [PDGF] receptors) that are found on many other cell types. Critical functions mediated by activated RTK include proliferation, migration, and enhanced survival of endothelial cells, as well as regulation of the recruitment of perivascular cells and bloodborne circulating endothelial precursors and hematopoietic stem cells to the tumor. Intracellular signaling via EC-specific RTK utilizes molecular pathways that may be targets for future antiangiogenic therapies.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 68-11

Normalization of tumor blood vessels due to inhibition of VEGF signaling. *A.* Blood vessels in normal tissues exhibit a regular hierarchical branching pattern that delivers blood to tissues in a spatially and temporally efficient manner to meet the metabolic needs of the tissue (top). At the microscopic level, tight junctions are maintained between endothelial cells (ECs), which are adherent to a thick and evenly distributed basement membrane (BM). Pericytes form a surrounding layer that provides trophic signals to the EC and helps maintain proper vessel tone. Vascular permeability is regulated, interstitial fluid pressure is low, and oxygen tension and pH are physiologic. *B.* Tumors have abnormal vessels with tortuous branching and dilated, irregular interconnecting branches, causing uneven blood flow with areas of hypoxemia and acidosis. This harsh environment selects genetic events that result in resistant tumor variants, such as the loss of p53. High levels of VEGF (secreted by tumor cells) disrupt gap junction communication, tight junctions, and adherens junctions between EC via src-mediated phosphorylation of proteins such as connexin 43, zonula occludens-1, VE-cadherin, and α/β-catenins. Tumor vessels have thin, irregular BM, and pericytes are sparse or absent. Together, these molecular abnormalities result in a vasculature that is permeable to serum macromolecules, leading to high tumor interstitial pressure, which can prevent the delivery of drugs to the tumor cells. This is made worse by the binding and activation of platelets at sites of exposed BM, with release of stored VEGF and microvessel clot formation, creating more abnormal blood flow and regions of hypoxemia. *C.* In experimental systems, treatment with bevacizumab or blocking antibodies to VEGFR2 leads to changes in the tumor vasculature that has been termed *vessel normalization*. During the first week of treatment, abnormal vessels are eliminated or pruned (dotted lines), leaving a more normal branching pattern. ECs partially regain features such as cell-cell junctions, adherence to a more normal BM, and pericyte coverage. These changes lead to a decrease in vascular permeability, reduced interstitial pressure, and a transient increase in blood flow within the tumor. Note that in murine models, this normalization period lasts only for ~5–6 days. *D.* After continued anti-VEGF/VEGFR therapy (which is often combined with chemo- or radiotherapy), ECs die, leading to tumor cell death (either due to direct effects of the chemotherapy or lack of blood flow).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Unlike normal blood vessels, the vascular lining of tumor vessels is not a homogeneous layer of ECs but often consists of a mosaic of ECs and tumor cells with upregulated genes seen in ECs and vessel formation that can occur in hypoxic conditions because of their plasticity; the concept of cancer cell–derived vascular channels, which may be lined by ECM secreted by the tumor cells, is referred to as *vascular mimicry*. During tumor angiogenesis, ECs are highly proliferative and express a number of plasma membrane proteins that are characteristic of activated endothelium, including growth factor receptors and adhesion molecules such as integrins.

MECHANISMS OF TUMOR VESSEL FORMATION

Tumors use a number of mechanisms to promote vascularization, subverting normal angiogenic processes for this purpose (Fig. 68-9). Primary or metastatic tumor cells sometimes arise in proximity to host blood vessels and grow around these vessels, parasitizing nutrients by co-opting the local blood supply. However, most tumor blood vessels arise by the process of *sprouting*, in which tumors secrete trophic angiogenic molecules, the most potent being VEGFs, that induce the proliferation and migration of host ECs into the tumor. Sprouting in normal and pathogenic angiogenesis is regulated by three families of transmembrane RTKs expressed on ECs and their ligands (VEGFs, angiopoietins, ephrins; Fig. 68-10), which are produced by tumor cells, inflammatory cells, or stromal cells in the tumor microenvironment.

When tumor cells arise in or metastasize to an avascular area, they grow to a size limited by hypoxemia and nutrient deprivation. Hypoxemia, a key regulator of tumor angiogenesis, causes the transcriptional induction of the genes encoding VEGF family members. VEGFs and their receptors are required for embryonic *vasculogenesis* (development of new blood vessels when none pre-exist) and normal (wound healing, corpus luteum formation) and pathologic angiogenesis (tumor angiogenesis, inflammatory conditions such as rheumatoid arthritis). VEGF-A is a heparin-binding glycoprotein with at least four isoforms (splice variants) that regulates blood vessel formation by binding to the RTKs VEGFR1 and VEGFR2, which are expressed on all ECs in addition to a subset of hematopoietic cells (Fig. 68-9). VEGFR2 regulates EC proliferation, migration, and survival, while VEGFR1 may act as an antagonist of R2 in ECs but is probably also important for angioblast differentiation during embryogenesis. Tumor vessels may be more dependent on VEGFR signaling for growth and survival than normal ECs. While VEGF signaling is a critical initiator of angiogenesis, this is a complex process regulated by additional signaling pathways (Fig. 68-10). The angiopoietin, Ang1, produced by stromal cells, binds to the EC RTK Tie-2 and promotes the interaction of ECs with the ECM and perivascular cells, such as pericytes and smooth-muscle cells, to form tight, nonleaky vessels. PDGF and basic fibroblast growth factor (bFGF) help to recruit these perivascular cells. Ang1 is required for maintaining the quiescence and stability of mature blood vessels and prevents the vascular permeability normally induced by VEGF and inflammatory cytokines.

For tumor cell–derived VEGF to initiate sprouting from host vessels, the stability conferred by the Ang1/Tie2 pathway must be perturbed; this occurs by the secretion of Ang2 by ECs that are undergoing active remodeling. Ang2 binds to Tie2 and is a competitive inhibitor of Ang1 action: under the influence of Ang2, preexisting blood vessels become more responsive to remodeling signals, with less adherence of ECs to stroma and associated perivascular cells and more responsiveness to VEGF. Therefore, Ang2 is required at early stages of tumor angiogenesis for destabilizing the vasculature by making host ECs more sensitive to angiogenic signals. In the presence of Ang2, there is no stabilization by the Ang1/Tie2 interaction, and tumor blood vessels are leaky, hemorrhagic, and have poor association of ECs with underlying stroma. Sprouting tumor ECs express high levels of the transmembrane protein ephrin-B2 and its receptor, the RTK EPH, whose signaling appears to work with the angiopoietins during vessel remodeling. During embryogenesis, EPH receptors are expressed on the endothelium of primordial venous vessels while the transmembrane ligand ephrin-B2 is expressed by cells of primordial arteries; the reciprocal expression may regulate differentiation and patterning of the vasculature.

A number of ubiquitously expressed host molecules play critical roles in normal and pathologic angiogenesis. Proangiogenic cytokines, chemokines, and growth factors secreted by stromal cells or inflammatory cells make important contributions to neovascularization, including bFGF, transforming growth factor-α (TGF-α), TNF-α, and IL-8. In contrast to normal endothelium, angiogenic endothelium overexpresses specific members of the integrin family of ECM-binding proteins that mediate EC adhesion, migration, and survival. Specifically, expression of integrins $\alpha_v\beta_3$, $\alpha_v\beta_5$, and $\alpha_5\beta_1$ mediates spreading and migration of ECs and is required for angiogenesis induced by VEGF and bFGF, which in turn can upregulate EC integrin expression. The $\alpha_v\beta_3$ integrin physically associates with VEGFR2 in the plasma membrane and promotes signal transduction from each receptor to promote EC proliferation (via focal adhesion kinase, src, PI3K, and other pathways) and survival (by inhibition of p53 and increasing the Bcl-2/Bax expression ratio). In addition, $\alpha_v\beta_3$ forms cell-surface complexes with matrix metalloproteinases (MMPs), zinc-requiring proteases that cleave ECM proteins, leading to enhanced EC migration and the release of heparin-binding growth factors, including VEGF and bFGF. EC adhesion molecules can be upregulated (i.e., by VEGF, TNF-α) or downregulated (by TGF-β); this, together with chaotic blood flow, explains poor leukocyte-endothelial interactions in tumor blood vessels and may help tumor cells avoid immune surveillance.

Lymphatic vessels also exist within tumors. Development of tumor lymphatics is associated with expression of VEGFR3 and its ligands VEGF-C and VEGF-D. The role of these vessels in tumor cell metastasis to regional lymph nodes remains to be determined. However, VEGF-C levels correlate significantly with metastasis to regional lymph nodes in lung, prostate, and colorectal cancers.

ANTIANGIOGENIC THERAPY

Angiogenesis inhibitors function by targeting the critical molecular pathways involved in EC proliferation, migration, and/or survival, many of which are highly expressed in the activated endothelium in tumors. Inhibition of growth factor and adhesion-dependent signaling pathways can induce EC apoptosis with concomitant inhibition of tumor growth. Different types of tumors can use distinct combinations of molecular

mechanisms to activate the angiogenic switch. Therefore, it is doubtful that a single antiangiogenic strategy will suffice for all human cancers; rather, a number of agents or combinations of agents will be needed, depending on distinct programs of angiogenesis used by different human cancers. Despite this, experimental data indicate that for some tumor types, blockade of a single growth factor (e.g., VEGF) may inhibit tumor-induced vascular growth.

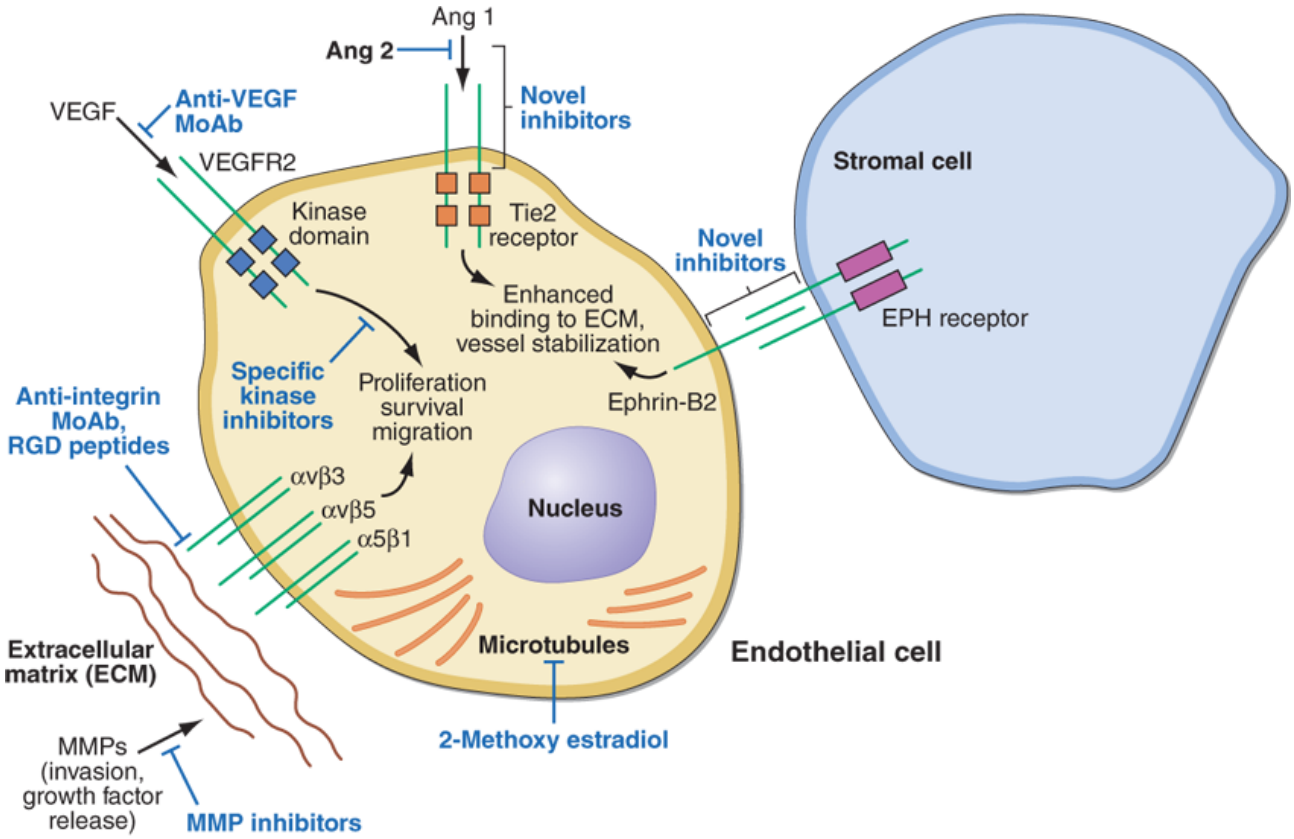
Bevacizumab, an antibody which binds VEGF, potentiates the effects of a number of different types of active chemotherapeutic regimens used to treat a variety of different tumor types including colon, lung, ovarian, and cervical cancers. It also has activity in combination with interferon against RCCs and alone for glioblastomas. Other protein inhibitors of the VEGF signaling pathway approved for anticancer therapy include ramucirumab (a monoclonal antibody directed against VEGFR2, approved for use against gastric/gastroesophageal, colon and lung cancers) and ziv-aflibercept (a recombinant protein inhibitor of VEGF, approved for colorectal cancer). Hypertension is the most common side effect of inhibitors of VEGF (or its receptors), but can be treated with antihypertensive agents and uncommonly requires discontinuation of therapy. Rare but serious potential risks include arterial thromboembolic events, including stroke and myocardial infarction, hemorrhage, bowel perforation, and inhibition of wound healing.

Several small-molecule inhibitors (SMI) that target VEGF RTK activity but are also inhibitory to other kinases have also been approved to treat certain cancers. Sunitinib (see above and Table 68-2) has activity directed against mutant c-Kit receptors (approved for GIST), but also targets VEGFR and PDGFR, and has antitumor activity against pancreatic neuroendocrine and metastatic renal cell carcinomas (RCC), presumably on the basis of its antiangiogenic activity. Similarly, sorafenib, originally developed as a Raf kinase inhibitor but with potent activity against VEGFR and PDGFR, has activity against RCC, differentiated thyroid and hepatocellular cancers as well as desmoid tumors. A closely related molecule to sorafenib, regorafenib, has activity against colorectal cancer, GIST, and hepatocellular cancer. Other inhibitors of the VEGF pathway approved for the treatment of various cancers include axitinib, pazopanib, lenvatinib, and cabozantinib.

The success in targeting tumor angiogenesis has led to enhanced enthusiasm for the development of drugs that target other aspects of the angiogenic process; some of these therapeutic approaches are outlined in **Fig. 68-12**. There is also evidence suggesting potential enhanced activity when anti-VEGF agents are used in combination with immunomodulators including immune check point inhibitors. However, it is not yet known whether this will produce a clinically meaningful enhancement of anti-tumor activity.

FIGURE 68-12

Knowledge of the molecular events governing tumor angiogenesis has led to a number of therapeutic strategies to block tumor blood vessel formation. The successful therapeutic targeting of VEGF and its receptors VEGFR is described in the text. Other endothelial cell–specific receptor tyrosine kinase pathways (e.g., angiopoietin/Tie2 and ephrin/EPH) are likely targets for the future. Ligation of the $\alpha_v\beta_3$ integrin is required for EC survival. Integrins are also required for EC migration and are important regulators of matrix metalloproteinase (MMP) activity, which modulates EC movement through the ECM as well as release of bound growth factors. Targeting of integrins includes development of blocking antibodies, small peptide inhibitors of integrin signaling, and arg-gly-asp–containing peptides that prevent integrin:ECM binding. Peptides derived from normal proteins by proteolytic cleavage, including endostatin and tumstatin, inhibit angiogenesis by mechanisms that include interfering with integrin function. Signal transduction pathways that are dysregulated in tumor cells indirectly regulate EC function. Inhibition of EGF-family receptors, whose signaling activity is upregulated in a number of human cancers (e.g., breast, colon, and lung cancers), results in downregulation of VEGF and IL-8, while increasing expression of the antiangiogenic protein thrombospondin-1. The Ras/MAPK, PI3K/Akt, and Src kinase pathways constitute important antitumor targets that also regulate the proliferation and survival of tumor-derived EC. The discovery that ECs from normal tissues express tissue-specific “vascular addressins” on their cell surface suggests that targeting specific EC subsets may be possible.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EVASION OF THE IMMUNE SYSTEM BY CANCERS

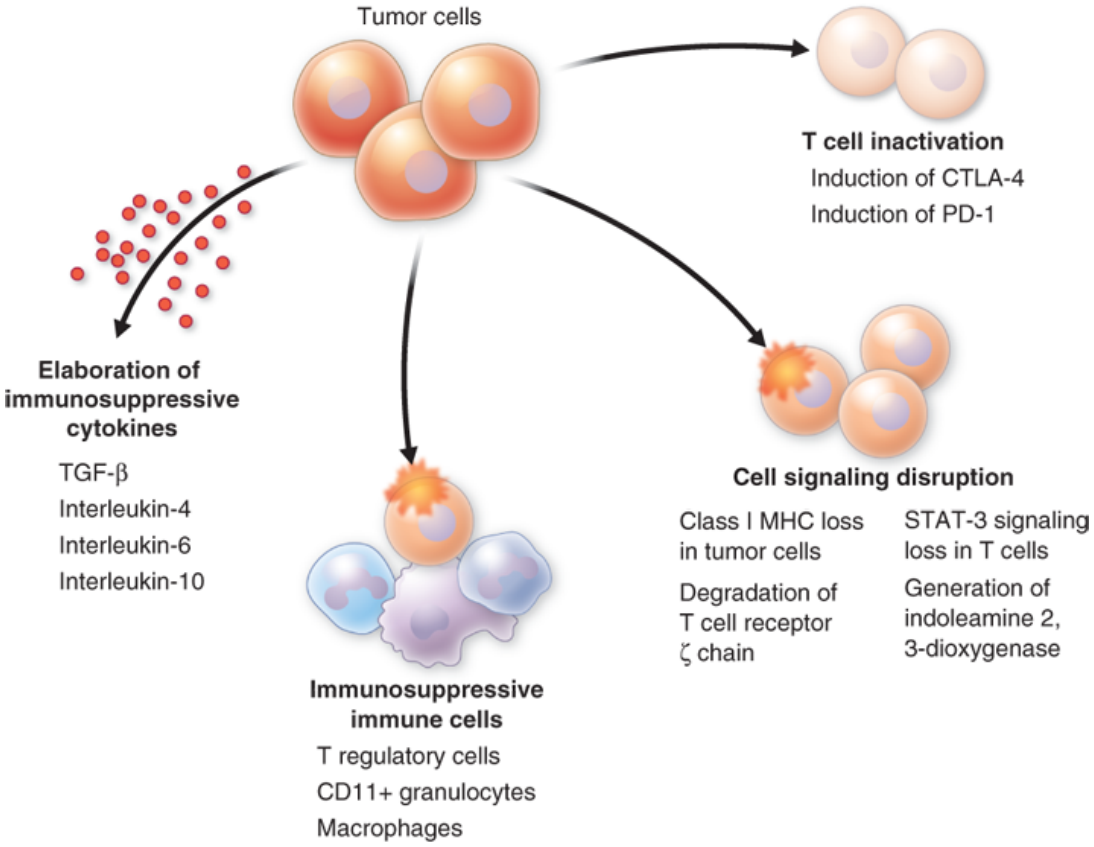
There is a complex interaction between tumors and the host from the initiation of the cancer until the establishment of a clinical cancer. Cancers have a number of mechanisms that allow them to evade detection and elimination by the immune system. These include downregulation of cell surface proteins involved in immune recognition (including MHC proteins and tumor-specific antigens), expression of other cell surface proteins that inhibit immune function (including members of the B7 family of proteins such as PD-L1), secretion of proteins and other molecules that are immunosuppressive, recruitment and expansion of immunosuppressive cells such as regulatory T cells, induction of T-cell tolerance, and down regulation of death receptors. Due to the marked heterogeneity of cells within a cancer, a variety of immune suppressive mechanisms are continuously occurring and changing. In addition, the inflammatory effects of some of the immune mediator cells in the tumor microenvironment (especially tissue-associated macrophages and myeloid-derived suppressor cells) can suppress T-cell responses to the tumor as well as stimulate inflammation that can enhance tumor growth. Immunotherapy approaches to treat cancer aimed at activating the immune response against tumors using immunostimulatory molecules such as interferons, IL-2, and monoclonal antibodies have had some successes. A more direct approach to enhance the activity of T cells directed against specific tumors involves isolating T cells from patients and re-engineering the cells to express chimeric antigen receptors (CAR-T cells) that recognize antigens present on the cells of that individual's tumor. The most commonly studied approach to date has been to engineer the cells to express receptors targeting the CD19 antigen on ALL and DLBCL cells. These have been shown to have significant antitumor activity in the treatment of patients with ALL and DLBCL including durable remissions in patients refractory to standard therapies and are approved for these malignancies. However, there have also been significant issues with toxicity including cytokine release syndrome, organ toxicity felt to be due to inadvertent targeting of antigens present in the organ, and neurotoxicity. These patients often require aggressive supportive care by individuals experienced in the delivery of CAR-T cells. In addition, as is true for most anticancer therapies, mechanisms of resistance have developed, most commonly the outgrowth of tumor cells no longer expressing the antigen. Mechanisms for preventing the development of resistant cells are being explored.

Another approach that has shown particular clinical promise is the targeting of proteins or cells (such as regulatory T cells) involved in normal homeostatic control to prevent autoimmune damage to the host but which malignant cells and their stroma can also utilize to inhibit the immune response directed against them. The approach that is furthest along clinically has involved targeting CTLA-4, PD-1, and PDL-1, co-inhibitory molecules that are expressed on the surface of cancer cells, cells of the immune system, and/or stromal cells and are involved in inhibiting the immune response against cancer (Fig. 68-13). A monoclonal antibody directed against CTLA-4 is approved for the treatment of

melanoma and antibodies targeting PD-1 or PDL-1 are approved for use against melanoma, RCC, lung cancer, head and neck cancer, urothelial cancer, HCC, gastric cancer, MSI high cancers, and Hodgkin's lymphoma. There is evidence of activity against other cancers including gastroesophageal and hepatocellular cancers and they continue to be evaluated against other malignancies as well. Combination approaches targeting more than one protein or with other anticancer approaches (targeted agents, chemotherapy, radiation therapy) are also being explored and have shown promise in early studies. An important aspect of these approaches is balancing sufficient release of the negative control of the immune response to allow immune mediated attack on the tumors while not allowing too much release and inducing severe autoimmune effects (such as against lung, liver, skin, thyroid, pituitary gland, or the GI tract).

FIGURE 68-13

Tumor-host interactions that suppress the immune response to the tumor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SUMMARY

Although each of the biological aspects of cancers and examples of targeting them has been addressed individually, clearly there is complicated cross-talk between these that occurs in all cancers which needs to be understood to optimally treat different cancers. The explosion of information on tumor cell biology, metastasis, and tumor-host interactions (including angiogenesis, other tumor-stromal interactions, and immune evasion by tumors) has ushered in a new era of rational targeted therapy for cancer. Furthermore, it has become clear that specific molecular factors detected in individual tumors (specific gene mutations, gene-expression profiles, microRNA expression, overexpression of specific proteins) can be used to tailor therapy and maximize antitumor effects.

ACKNOWLEDGMENT

Robert G. Fenton contributed to this chapter in prior editions and important material from those prior chapters has been included here.

FURTHER READING

Boussiotis VA: Molecular and biochemical aspects of the PD-1 checkpoint pathway. N Engl J Med 375:1767, 2016. [PubMed: 27806234]

De Palma M, Biziato D, Petrova TV: Microenvironmental regulation of tumour angiogenesis. Nat Rev Cancer 17:457, 2017.

Du W, Elemento O: Cancer systems biology: Embracing complexity to develop better anticancer therapeutic strategies. Oncogen 34:3215, 2015.

Fleuren ED et al: The kinome “at large” in cancer. Nat Rev Cancer 16:83, 2016. [PubMed: 26822576]

He S, Sharpless NE: Senescence in health and disease. Cell 169:1000,11, 2017.

Lambert AW et al: Emerging biological principles of metastasis. Cell 168:670, 2017. [PubMed: 28187288]

Otto T, Sicinski P: Cell cycle proteins as promising targets in cancer therapy. Nat Rev Cancer 17:93, 2017. [PubMed: 28127048]

Tomasetti C et al: Stem cell divisions, somatic mutations, cancer etiology and cancer prevention. Science 355:1330, 2017. [PubMed: 28336671]

Vander Heiden MG, DeBerardinis RJ: Understainding the intersections between metabolism and cancer biology. Cell 168:657, 2017. [PubMed: 28187287]

Vogelstein B et al: Cancer genome landscapes. Science 339:1546, 2013. [PubMed: 23539594]

Chapter 69: Principles of Cancer Treatment

Edward A. Sausville; Dan L. Longo

CANCER PRESENTATION

Localized or systemic cancer is frequent in the differential diagnosis of a variety of common complaints. Although not all forms of cancer are curable at initial diagnosis, affording patients the greatest opportunity for cure or meaningful prolongation of life is greatly aided by diagnosing cancer early in its natural history, and defining treatments that prevent or retard its systemic spread. Indeed, certain forms of cancer, notably breast, colon, and possibly lung cancers in certain patients, can be prevented by screening appropriately selected asymptomatic patients; screening is arguably the earliest point in the spectrum of possible cancer-related interventions where cure is possible **(Table 69-1)**.

TABLE 69-1

Spectrum of Cancer-Related Interventions

Screening for cancer in an asymptomatic patient
Consideration of cancer in a differential diagnosis
Physical examination, imaging, or endoscopy to define a possible tumor
Diagnosis of cancer by biopsy or removal:
Routine histology
Specialized histology: immunohistochemistry
Molecular studies
Cytogenetic studies
Staging the cancer: Where has it spread?
Treatment
Localized
Systemic
Supportive care
During treatment: related to tumor effects on patient
During treatment to counteract side effects of treatment
Palliative and end of life
When useful treatments are not feasible or desired

DETECTION OF A CANCER

The term *cancer*, as used here, is synonymous with the term *tumor*, whose original derivation from Latin simply meant “swelling,” not otherwise specified. We now understand that swelling as a common physical manifestation of a tumor reflects increased interstitial fluid pressure and increased cellular and stromal mass per volume, compared to normal tissue. Leukemias are a special case of a cancer of the blood-forming tissues presenting in a disseminated form frequently without definable tumor masses. In addition to localized swelling, tumors present by altered function of the organ they afflict, such as dyspnea on exertion from the anemia caused by leukemia replacing normal hematopoietic cells, cough from lung cancers, jaundice from tumors disrupting the hepatobiliary tree, or seizures and neurologic signs from brain tumors. Hemorrhage is also a frequent presenting sign of tumors involving hollow viscera, but also may reflect decreases in the number of platelets or altered blood coagulation. Tumors may also present owing to the effects of substances they secrete called a “paraneoplastic” syndrome. Thus, although statistically the fraction of patients with cancer underlying a particular presenting sign or symptom may be low, the implications for a patient with cancer of missing an early-stage tumor call for vigilance; therefore, persistent signs or symptoms should be evaluated as possibly coming from an early-stage tumor.

Evidence of a tumor's existence can objectively be established by careful physical examination, detecting enlarged lymph nodes in lymphomas or a palpable mass in a breast or soft tissue site. A mass may also be detected or confirmed by an imaging modality, such as plain x-ray, computed tomography (CT) scan, ultrasound, positron emission tomography (PET) imaging, or nuclear magnetic resonance approaches. Another way of initially establishing the existence of a possible tumor is through direct visualization of an afflicted organ by endoscopy.

ESTABLISHING A CANCER DIAGNOSIS

Once the existence of a likely tumor is defined, unequivocally establishing the diagnosis is the next step in the intervention spectrum. This is usually accomplished by a biopsy procedure and the emergence after pathologic examination of an unequivocal statement that cancer is present, or a non-cancer diagnosis explains the abnormality. Due to tumor heterogeneity, pathologists are better able to make the diagnosis when they have more tissue to examine. In addition to light microscopic inspection of a tumor, sufficient tissue also allows definition of genetic abnormalities and protein expression patterns, such as hormone receptor expression in breast cancers, that may aid in the differential diagnosis or provide information about prognosis or likely response to treatment. Efforts to define "personalized" information from the biology of each patient's tumor and pertinent to each patient's treatment plan are becoming increasingly important in selecting treatment options. The general internist should make sure that a patient's cancer biopsy is appropriately referred from the surgical suite for important molecular studies that can advise the best treatment (Table 69-2).

TABLE 69-2

Diagnostic Biopsy: Standard of Care Molecular and Special Studies

Breast cancer: primary and suspected metastatic
Hormone receptors: estrogen, progesterone
HER2/neu oncoprotein
Lung cancer: primary and suspected metastatic
If nonsquamous non-small cell: epidermal growth factor receptor mutation; alk oncoprotein gene fusion; programmed cell death ligand-1
Colon cancer: suspected metastatic
Ki-ras mutation
Gastrointestinal stromal tumor
c-kit oncoprotein mutation
Melanoma
B-raf oncoprotein mutation
c-kit expression and mutation
Brain tumor gliomas
1p/19q co-deletion
Alkylguanine alkyltransferase promoter methylation
Leukemia (peripheral blood mononuclear cells and/or bone marrow)
Cytogenetics
Flow cytometry
Treatment-defining chromosomal translocations
Bcr-Abl fusion protein
t(15,17)
inversion 16
t(8,21)
Lymphoma
Immunohistochemistry for CD20, CD30, T cell markers
Treatment defining chromosomal translocations:
t(14,18)
t(8,14)

Coordination among the surgeon, pathologist, and primary care physician is essential to ensure that the amount of information learned from the biopsy material is maximized. These goals are best met by an *excisional biopsy* in which the entire tumor mass is removed with a small margin of normal tissue surrounding it. If an excisional biopsy cannot be performed, *incisional biopsy* is the procedure of second choice. A wedge of tissue is removed, and an effort is made to include the majority of the cross-sectional diameter of the tumor in the biopsy to minimize sampling error. Biopsy techniques that involve cutting into

tumor carry with them a risk of facilitating the spread of the tumor, and consideration with a surgeon of whether the biopsy might be the prelude to a curative surgery if certain diagnoses are established should inform the actual approach taken. *Core-needle biopsy* usually obtains considerably less tissue, but this procedure often provides enough information to plan a definitive surgical procedure. *Fine-needle aspiration* generally obtains only a suspension of cells from within a mass. This procedure is minimally invasive, and if positive for cancer, it may allow inception of systemic treatment when metastatic disease is evident, or it can provide a basis for planning a more meticulous and extensive surgical procedure. However, a negative fine-needle aspiration for a neoplastic diagnosis cannot be taken as definitive evidence that a tumor is absent or make a definitive diagnosis in someone not known to have a cancer.

CANCER STAGING

An essential component of correct patient management in many cancer types is defining the extent of disease, because this information critically informs whether localized treatments, “combined-modality” approaches, or systemic treatments should initially be considered. Radiographic and other imaging tests can be helpful in defining the clinical stage; however, pathologic staging requires defining the extent of involvement by documenting the histologic presence of tumor in tissue biopsies obtained through a surgical procedure. Axillary lymph node sampling in breast cancer and lymph node sampling at laparotomy for testicular, colon, and other intraabdominal cancers may provide crucial information for treatment planning and may determine the extent and nature of primary cancer treatment.

For tumors associated with a potential “primary site,” staging systems have evolved to define a “T” component related to the size of the tumor or its invasion into local structures, an “N” component related to the number and nature of lymph node groups adjacent to the tumor with evidence of tumor spread, and an “M” component, based on the presence of local or distant metastatic sites. The various “TNM” components are then aggregated to stages, usually stage I to III or IV, depending on the anatomic site. The numerical stages reflect similar long-term survival outcomes of the aggregated TNM groupings in a numeric stage after treatment tailored to the stage. In general, stage I tumors are T1 (reflecting small size), N0 or N1 (reflecting no or minimal node spread), and M0 (no metastases). Such early-stage tumors are amenable to curative approaches with local treatments. On the other hand, stage IV tumors usually have metastasized to distant sites or locally invaded viscera in a nonresectable way and are dealt with using techniques that have palliative intent, except for those diseases with exceptional sensitivity to systemic treatments such as chemotherapy or immunotherapy. Also, the TNM staging system is not useful in diseases such as leukemia, where bone marrow infiltration is never really localized, or central nervous system (CNS) tumors, where tumor histology and the extent of anatomically feasible resection are more important in driving prognosis.

CANCER TREATMENT

The goal of cancer treatment is first to eradicate the cancer. If this primary goal cannot be accomplished, the goal of cancer treatment shifts to palliation, the amelioration of symptoms, and preservation of quality of life while striving to extend life. The dictum *primum non nocere* may not always be the guiding principle of

cancer therapy. When cure of cancer is possible, cancer treatments may be considered despite the certainty of severe and perhaps life-threatening toxicities. Every cancer treatment has the potential to cause harm, and treatment may be given that produces toxicity with no benefit. The therapeutic index of many interventions may be quite narrow, with treatments given to the point of toxicity. Conversely, when the clinical goal is palliation, careful attention to minimizing the toxicity of treatments becomes a significant goal.

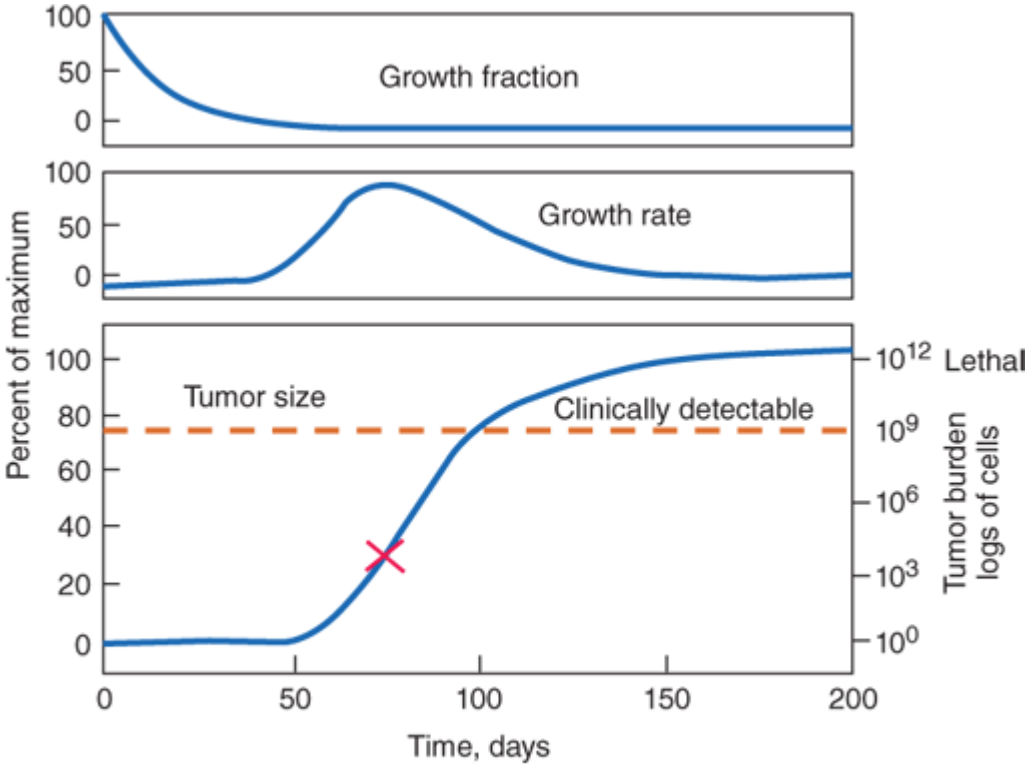
Cancer treatments are divided into two main types: local and systemic. Local treatments include surgery, radiation therapy (including photodynamic therapy), and ablative approaches, including radiofrequency and cryosurgical approaches. Systemic treatments include chemotherapy (including hormonal therapy and molecularly targeted therapy) and biologic therapy (including immunotherapy). The modalities are often used in combination, and agents in one category can act by several mechanisms. For example, cancer chemotherapy agents can induce differentiation, and antibodies (a form of immunotherapy) can be used to deliver radiation therapy. *Oncology*, the study of tumors including treatment approaches, is a multidisciplinary effort with surgical, radiation, and internal medicine–related areas of oncologic expertise. Treatments for patients with hematologic malignancies are often shared by hematologists and medical oncologists.

Normal organs and cancers share the property of having a population of cells actively progressing through the cell cycle with their division providing a basis for tumor growth, and a population of cells not in cycle; these include *cancer stem cells*, whose properties are being elucidated, as they may serve as a basis for giving rise to tumor initiating or repopulating cells. The stem cell fraction may define new targets for therapies that will retard their ability to reenter the cell cycle.

Tumors follow a Gompertzian growth curve **(Fig. 69-1)**, with the apparent growth fraction of a neoplasm being high with small tumor burdens and declining until, at the time of diagnosis, with a tumor burden of 1–5 $\times 10^9$ tumor cells, the growth fraction is usually 1–4% for many solid tumors. By this view, the most rapid growth rate occurs before the tumor is detectable. An alternative explanation for such growth properties may also emerge from the ability of tumors at metastatic sites to recruit circulating tumor cells from the primary tumor or other metastases. An additional key feature of a successful tumor is the ability to stimulate the development of a new supporting stroma through angiogenesis and production of proteases to allow invasion through basement membranes and normal tissue barriers **(Chap. 68)**.

FIGURE 69-1

Gompertzian tumor growth. The growth fraction of a tumor declines exponentially over time (*top*). The growth rate of a tumor peaks before it is clinically detectable (*middle*). Tumor size increases slowly, goes through an exponential phase, and slows again as the tumor reaches the size at which limitation of nutrients or autoregulatory or host regulatory influences can occur. The maximum growth rate occurs at 1/e, the point at which the tumor is about 37% of its maximum size (*marked with an X*). Tumor becomes detectable at a burden of about 10^9 (1 cm^3) cells and kills the patient at a tumor cell burden of about 10^{12} (1 kg). Efforts to treat the tumor and reduce its size can result in an increase in the growth fraction and an increase in growth rate.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

LOCALIZED CANCER TREATMENTS

SURGERY

Surgery is unquestionably the most effective means of treating cancer. Today at least 40% of cancer patients are cured by surgery. Unfortunately, a large fraction of patients with solid tumors (perhaps 60%) have metastatic disease that is not accessible for removal. Even when cancer is not curable by surgery alone, the removal of tumor can obtain important benefits, including local control of tumor, preservation of organ function, debulking that permits subsequent therapy to be more effective, and staging information on extent of involvement. Cancer surgery aiming for cure is usually planned to excise the tumor completely with an adequate margin of normal tissue (the margin varies with the tumor and the anatomy), touching the tumor as little as possible to prevent vascular and lymphatic spread, and minimizing operative risk. Such a resection is defined as an R0 resection. R1 and R2 resections, in contrast, are imprecisely defined pathologically as having microscopic or macroscopic, respectively, tumor at resection margins. Such outcomes may be necessitated by proximity of the tumor to vital structures or recognition only in the resected specimen of the extent of tumor involvement, and may be the basis for reoperation to obtain optimal margins if feasible. Extending the procedure to resect draining lymph nodes obtains prognostic information and may, in some anatomic locations, improve survival.

Increasingly, laparoscopic approaches are being used to address primary abdominal and pelvic tumors. Lymph node spread may be assessed using the sentinel node approach, in which the first draining lymph node a spreading tumor would encounter is defined by injecting a dye or radioisotope into the tumor site at

operation and then resecting the first node to turn blue or collect isotope. The sentinel node assessment is continuing to undergo clinical evaluation but appears to provide reliable information without the risks (lymphedema, lymphangiosarcoma) associated with resection of all the regional nodes. Advances in adjuvant chemotherapy (chemotherapy given systemically after removal of all local disease by operation and without evidence of active metastatic disease) and radiation therapy following surgery have permitted a substantial decrease in the extent of primary surgery necessary to obtain the best outcomes. Thus, lumpectomy with radiation therapy is as effective as modified radical mastectomy for breast cancer, and limb-sparing surgery followed or preceded by adjuvant radiation therapy and chemotherapy has replaced radical primary surgical procedures involving amputation and disarticulation for childhood rhabdomyosarcomas and osteosarcomas. More limited surgery is also being used to spare organ function, as in larynx and bladder cancer. In some settings (e.g., bulky testicular cancer or stage III breast cancer), surgery is not the first treatment modality used. After an initial diagnostic biopsy, chemotherapy and/or radiation therapy is delivered to reduce the size of the tumor and clinically control undetected metastatic disease. Such therapy is followed by a surgical procedure to remove residual masses; this is called *neoadjuvant therapy*. Because the sequence of treatment is critical to success and is different from the standard surgery-first approach, coordination among the surgical oncologist, radiation oncologist, and medical oncologist is crucial.

Surgery may be curative in a subset of patients with metastatic disease. Patients with lung metastases from osteosarcoma may be cured by resection of the lung lesions. In patients with colon cancer who have fewer than five liver metastases restricted to one lobe and no extrahepatic metastases, hepatic lobectomy may produce long-term disease-free survival in 25% of selected patients. Surgery can also be associated with systemic antitumor effects. In the setting of hormonally responsive tumors, oophorectomy and/or adrenalectomy may eliminate estrogen production, and orchiectomy may reduce androgen production, hormones that drive certain breast and all prostate cancers, respectively; both procedures can have useful effects on metastatic tumor growth. In selecting a surgeon or center for primary cancer treatment, consideration must be given to the volume of cancer surgeries undertaken by the site. Studies in a variety of cancers have shown that increased annual procedure volume appears to correlate with outcome. In addition, facilities with extensive support systems—e.g., for joint thoracic and abdominal surgical teams with cardiopulmonary bypass, if needed—may allow resection of certain tumors that would otherwise not be possible.

Surgery is used in a number of ways for palliative or supportive care of the cancer patient, not related to the goal of curing the cancer. These include insertion and care of central venous catheters, control of pleural and pericardial effusions and ascites, caval interruption for recurrent pulmonary emboli, stabilization of cancer-weakened weight-bearing bones, and control of hemorrhage, among others. Surgical bypass of gastrointestinal, urinary tract, or biliary tree obstruction can alleviate symptoms and prolong survival. Surgical procedures may provide relief of otherwise intractable pain or reverse neurologic dysfunction (cord decompression). Splenectomy may relieve symptoms and reverse hypersplenism. Intrathecal or intrahepatic therapy relies on surgical placement of appropriate infusion portals. Surgery may correct other treatment-related toxicities such as adhesions or strictures. Surgical procedures are also valuable in rehabilitative

efforts to restore health or function. Orthopedic procedures may be necessary to ensure proper ambulation. Breast reconstruction can make an enormous impact on the patient's perception of successful therapy. Plastic and reconstructive surgery can correct the effects of disfiguring primary treatment.

Surgery is also a tool valuable in the prevention of cancers in high-risk populations. Prophylactic mastectomy, colectomy, oophorectomy, and thyroidectomy are mainstays of prevention of genetic cancer syndromes. Resection of premalignant skin and uterine cervix lesions and colonic polyps prevents progression to frank malignancy.

RADIATION

Radiation Biology and Medicine

Therapeutic radiation is ionizing, causing breaks in DNA and generation of free radicals from cell water that may damage cell membranes, proteins, and organelles. Radiation damage is augmented by oxygen; hypoxic cells are more resistant. Augmentation of oxygen presence is one basis for radiation sensitization. X-rays and gamma rays are the forms of ionizing radiation most commonly used to treat cancer. They are both electromagnetic, nonparticulate waves that cause the ejection of an orbital electron when absorbed. This orbital electron ejection results in ionization. These waves behave biologically as packets of energy, called *photons*. Particulate ionizing radiation using protons has also become available. Most radiation-induced cell damage is due to the formation of hydroxyl radicals from tissue water:

$$\text{Ionizing radiation} + H_2O \rightarrow H_2O^+ + e^-$$
$$H_2O^+ + H_2O \rightarrow H_3O^+ + OH^\bullet$$
$$OH^\bullet \rightarrow \text{cell damage}$$

Radiation is quantitated based on the amount of radiation absorbed by the tumor in the patient; it is not based on the amount of radiation generated by the machine. The International System (SI) unit for radiation absorbed is the Gray (Gy): 1 Gy refers to 1 J/kg of tissue; 1 Gy equals 100 centigrays (cGy) of absorbed dose. A historically used unit appearing in the oncology literature, the *rad* (*r*adiation *a*bsorbed *d*ose), is defined as 100 ergs of energy absorbed per gram of tissue and is equivalent to 1 cGy. Radiation dosage is defined by the energy absorbed per mass of tissue. Radiation dose is measured by placing detectors at the body surface or based on radiating phantoms that resemble human form and substance, containing internal detectors. The features that make a particular cell more sensitive or more resistant to the biologic effects of radiation are not completely defined and critically involve DNA repair proteins that, in their physiologic role, protect against environmentally related DNA damage.

Localized Radiation Therapy

Radiation effect is influenced by three determinants: total absorbed dose, number of fractions, and time of treatment. A frequent error is to omit the number of fractions and the duration of treatment. Thus, a typical course of radiation therapy should be described as 4500 cGy delivered to a particular target (e.g., mediastinum) over 5 weeks in 180-cGy fractions. Most curative radiation treatment programs are delivered

once a day, 5 days a week, in 150- to 200-cGy fractions. Nondividing cells are more resistant than dividing cells, and this is one rationale for delivering radiation in repeated fractions, to ultimately expose a larger number of tumor cells that have entered the division cycle. In addition to these biologic parameters, physical parameters of the radiation are also crucial. The energy of the radiation determines its ability to penetrate tissue. Low-energy x-rays (150–400 kV) scatter when they strike the body, much like light diffuses when it strikes particles in the air. Such beams result in more damage to adjacent normal tissues and less radiation delivered to the tumor. Megavoltage radiation (>1 MeV) has very low lateral scatter; this produces a skin-sparing effect, more homogeneous distribution of the radiation energy, and greater deposit of the energy in the tumor, or *target volume*. The tissues that the beam passes through to get to the tumor are called the *transit volume*. The maximum dose in the target volume is often the cause of complications to tissues in the transit volume, and the minimum dose in the target volume influences the likelihood of tumor recurrence. Dose homogeneity in the target volume is the goal. Computational approaches and delivery of many beams to converge on a target lesion are the basis for “gamma knife” and related approaches to deliver high doses to small volumes of tumor, sparing normal tissue.

Therapeutic radiation is delivered in three ways: (1) *teletherapy*, with focused beams of radiation generated at a distance and aimed at the tumor within the patient; (2) *brachytherapy*, with encapsulated sources of radiation implanted directly into or adjacent to tumor tissues; and (3) *systemic therapy*, with radionuclides administered, for example, intravenously but targeted by some means to a tumor site. Teletherapy with x-ray or gamma-ray photons is the most commonly used form of radiation therapy. Particulate forms of radiation are also used in certain circumstances, such as the use of proton beams. The difference between photons and protons relates to the volume in which the greatest delivery of energy occurs. Typically, protons have a much narrower range of energy deposition, theoretically resulting in more precise delivery of radiation with improvement in the degree to which adjacent structures may be affected, in comparison to photons. Electron beams are a particulate form of radiation that, in contrast to photons and protons, have a very low tissue penetrance and are used to treat cutaneous tumors. Certain drugs used in cancer treatment may also act as radiation sensitizers. For example, compounds that incorporate into DNA and alter its stereochemistry (e.g., halogenated pyrimidines, cisplatin) augment radiation effects at local sites, as does hydroxyurea, another DNA synthesis inhibitor. These are important adjuncts to the local treatment of certain tumors, such as squamous head and neck, uterine cervix, and rectal cancers.

Toxicity of Radiation Therapy

Although radiation therapy is most often administered to a local region, systemic effects, including fatigue, anorexia, nausea, and vomiting, may develop that are related in part to the volume of tissue irradiated, dose fractionation, radiation fields, and individual susceptibility. Injured tissues release cytokines that act systemically to produce these effects. Bone is among the most radioresistant organs, with radiation effects being manifested mainly in children through premature fusion of the epiphyseal growth plate. By contrast, the male testis, female ovary, and bone marrow are the most sensitive organs. Any bone marrow in a radiation field will be eradicated by therapeutic irradiation. Organs with less need for cell renewal, such as heart, skeletal muscle, and nerves, are more resistant to radiation effects. In radiation-resistant organs, the

vascular endothelium is the most sensitive component. Organs with more self-renewal as a part of normal homeostasis, such as the hematopoietic system and mucosal lining of the intestinal tract, are more sensitive. Acute toxicities include mucositis, skin erythema (ulceration in severe cases), and bone marrow toxicity. Often these can be alleviated by interruption of treatment.

Chronic toxicities are more serious. Radiation of the head and neck region often produces thyroid failure. Cataracts and retinal damage can lead to blindness. Salivary glands stop making saliva, which leads to dental caries and poor dentition. Taste and smell can be affected. Mediastinal irradiation leads to a threefold increased risk of fatal myocardial infarction. Other late vascular effects include chronic constrictive pericarditis, lung fibrosis, viscus stricture, spinal cord transection, and radiation enteritis. A serious late toxicity is the development of second solid tumors in or adjacent to the radiation fields. Such tumors can develop in any organ or tissue and occur at a rate of ~1% per year beginning in the second decade after treatment. Some organs vary in susceptibility to radiation carcinogenesis. A woman who receives mantle field radiation therapy for Hodgkin's disease at age 25 years has a 30% risk of developing breast cancer by age 55 years. This is comparable in magnitude to genetic breast cancer syndromes. Women treated after age 30 years have little or no increased risk of breast cancer. No data suggest that a threshold dose of therapeutic radiation exists below which the incidence of second cancers is decreased. High rates of second tumors occur in people who receive as little as 1000 cGy.

OTHER LOCALIZED CANCER TREATMENTS

Endoscopy techniques may allow the placement of stents to unblock viscera by mechanical means, palliating, for example, gastrointestinal or biliary obstructions. Radiofrequency ablation (RFA) refers to the use of focused microwave radiation to induce thermal injury within a volume of tissue. RFA can be useful in the control of metastatic lesions, particularly in liver, that may threaten biliary drainage (as one example) and threaten quality and duration of useful life in patients with otherwise unresectable disease. Cryosurgery uses extreme cold to sterilize lesions in certain sites, such as prostate and kidney, when at a very early stage, eliminating the need for modalities with more side effects such as surgery or radiation.

Some chemicals (porphyrins, phthalocyanines) are preferentially taken up by cancer cells by mechanisms not fully defined. When light, usually delivered by a laser, is shone on cells containing these compounds, free radicals are generated and the cells die. Hematoporphyrins and light (phototherapy) are being used with increasing frequency to treat skin cancer; ovarian cancer; and cancers of the lung, colon, rectum, and esophagus. Palliation of recurrent locally advanced disease can sometimes be dramatic and last many months.

Infusion of chemotherapeutic or biologic agents or radiation-bearing delivery devices such as isotope-coated glass spheres into local sites through catheters inserted into specific vascular sites such as liver or an extremity have been used in an effort to control disease limited to that site; in selected cases, prolonged control of truly localized disease has been possible.

SYSTEMIC CANCER TREATMENTS

The concept that systemically administered agents may have a useful effect on cancers was historically derived from three sets of observations. Paul Ehrlich in the nineteenth century observed that different dyes reacted with different cell and tissue components. He hypothesized the existence of compounds that would be “magic bullets” that might bind to tumors, owing to the affinity of the agent for the tumor. A second observation was the toxic effects of certain mustard gas derivatives on the bone marrow during World War I, leading to the idea that smaller doses of these agents might be used to treat tumors of marrow-derived cells. Finally, the observation that certain tumors from hormone-responsive tissues, e.g., breast tumors, could shrink after oophorectomy led to the idea that endogenous substances promoting the growth of a tumor might be antagonized. Chemicals achieving each of the goals are actually or intellectually the forbearers of the currently used cancer chemotherapy agents.

Systemic cancer treatments are of four broad types. *Conventional “cytotoxic” chemotherapy agents* were historically derived by the empirical observation that these “small molecules” (generally with molecular mass <1500 Da) could cause major regression of experimental tumors growing in animals. These agents mainly target DNA structure or segregation of DNA as chromosomes in mitosis. *Targeted agents* refer to small molecules or “biologics” (generally macromolecules such as antibodies or cytokines) designed and developed to interact with a defined molecular target important in maintaining the malignant state or expressed by the tumor cells. As described in **Chap. 68**, successful tumors have activated biochemical pathways that lead to uncontrolled proliferation through the action of, e.g., oncogene products, loss of cell cycle inhibitors, or loss of cell death regulation, and have acquired the capacity to replicate chromosomes indefinitely, invade, metastasize, and evade the immune system. Targeted therapies seek to capitalize on the biology behind the aberrant cellular behavior as a basis for therapeutic effects. *Hormonal therapies* (the first form of targeted therapy) capitalize on the biochemical pathways underlying estrogen and androgen function and action as a therapeutic basis for approaching patients with tumors of breast, prostate, and uterus. *Biologic therapies* are often macromolecules that have a particular target (e.g., anti-growth factor receptor or cytokine antibodies) or may have the capacity to induce a host immune response to kill tumor cells.

CANCER CHEMOTHERAPY

Principles

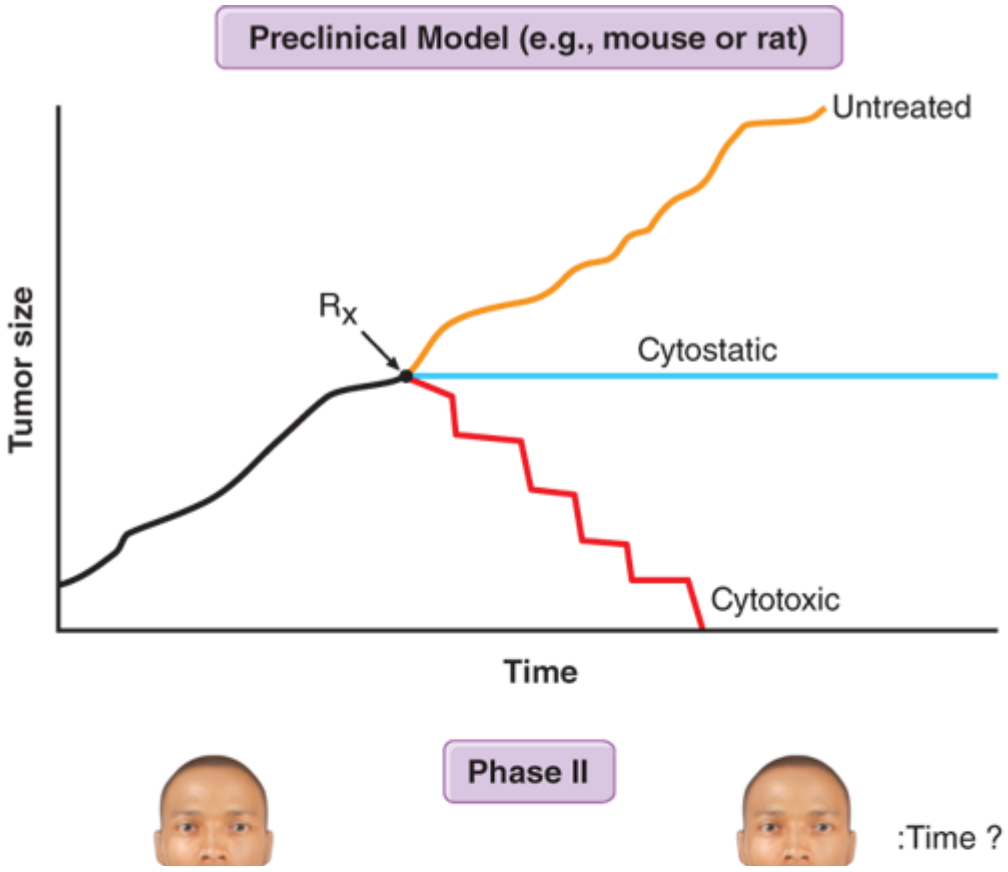
The usefulness of any drug is governed by the extent to which a given dose causes a therapeutic effect (in the case of anticancer agents, toxicity to tumor cells) as opposed to a toxic effect to the host. The *therapeutic index* is the degree of separation between toxic and therapeutic doses. Really useful drugs have large therapeutic indices, and this usually occurs when the drug target is expressed in the disease-causing compartment as opposed to the normal compartment. Currently used chemotherapeutic agents have the unfortunate property that their targets are present in both normal and tumor tissues. Therefore, they have relatively narrow therapeutic indices.

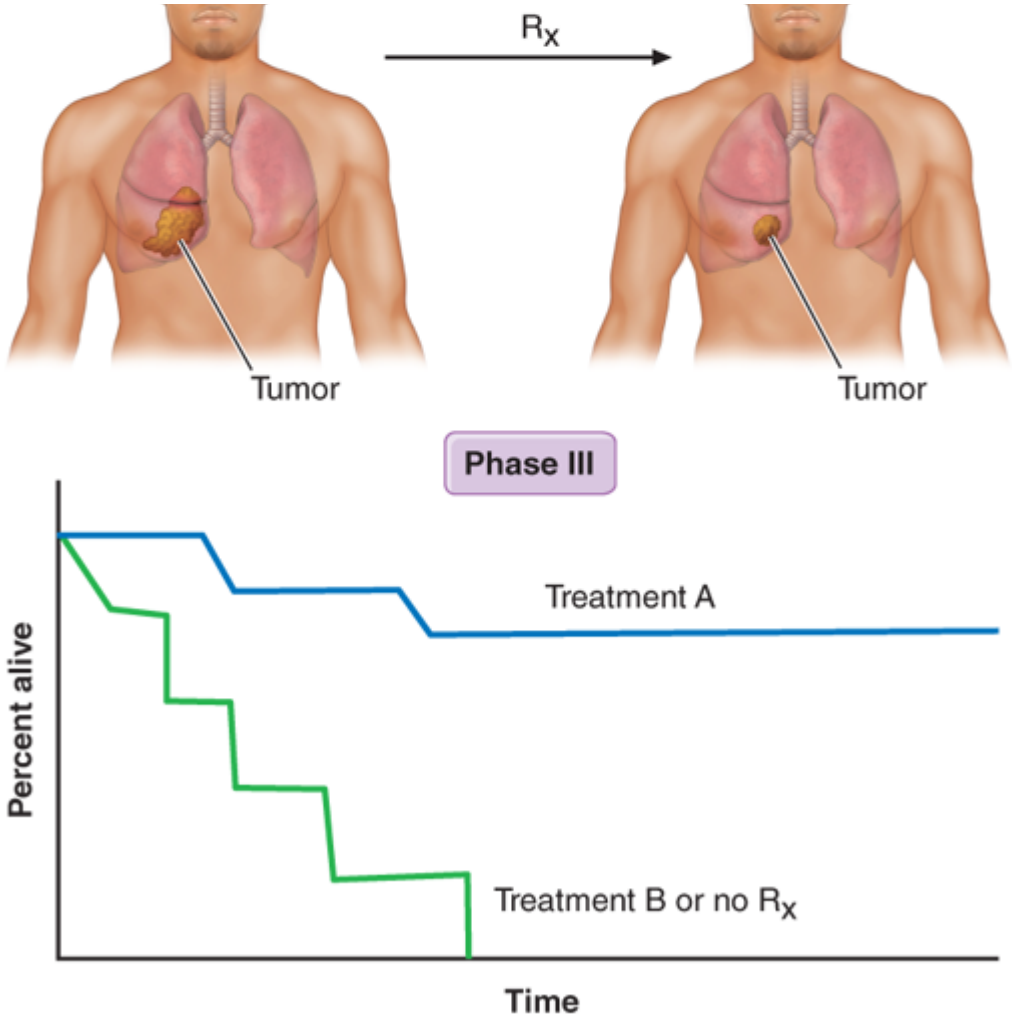
Figure 69-2 illustrates steps in cancer drug development. Following demonstration of antitumor activity in animal models, potentially useful anticancer agents are further evaluated to define an optimal schedule of

administration and arrive at a drug formulation designed for a given route of administration and schedule. Safety testing in two species on an analogous schedule of administration defines the starting dose for a phase 1 trial in humans, usually but not always in patients with cancer who have exhausted “standard” (already approved) treatments. The initial dose is usually one-sixth to one-tenth of the dose just causing easily reversible toxicity in the more sensitive animal species. Escalating doses of the drug are then given during the human phase 1 trial until reversible toxicity is observed. Dose-limiting toxicity (DLT) defines a dose that conveys greater toxicity than would be acceptable in routine practice, allowing definition of a lower maximum-tolerated dose (MTD). The occurrence of toxicity is, if possible, correlated with plasma drug concentrations. The MTD or a dose just lower than the MTD is usually the dose suitable for phase 2 trials, where a fixed dose is administered to a relatively homogeneous set of patients with a particular tumor type in an effort to define whether the drug causes regression of tumors. In a phase 3 trial, evidence of improved overall survival or improvement in the time to progression of disease on the part of the new drug is sought in comparison to an appropriate control population, which is usually receiving an acceptable “standard of care” approach. A favorable outcome of a phase 3 trial is the basis for application to a regulatory agency for approval of the new agent for commercial marketing as safe and possessing a measure of clinical effectiveness.

FIGURE 69-2

Steps in cancer drug discovery and development. Preclinical activity (*top*) in animal models of cancers may be used as evidence to support the entry of the drug candidate into phase 1 trials in humans to define a correct dose and observe any clinical antitumor effect that may occur. The drug may then be advanced to phase 2 trials directed against specific cancer types, with rigorous quantitation of antitumor effects (*middle*). Phase 3 trials then may reveal activity superior to standard or no treatment (*bottom*).





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Response, defined as tumor shrinkage, is the most immediate indicator of drug effect. To be clinically valuable, responses must translate into clinical benefit. This is conventionally established by a beneficial effect on overall survival, or at least an increased time to further progression of disease. Karnofsky was among the first to champion the evaluation of a chemotherapeutic agent's benefit by carefully quantitating its effect on tumor size and using these measurements to objectively decide the basis for further treatment of a particular patient or further clinical evaluation of a drug's potential. A partial response (PR) is defined conventionally as a decrease by at least 50% in a tumor's bidimensional area; a complete response (CR) connotes disappearance of all tumor; progression of disease signifies an increase in size of existing lesions by >25% from baseline or best response or development of new lesions; and stable disease fits into none of the above categories. Newer evaluation systems, such as Response Evaluation Criteria in Solid Tumors (RECIST), use unidimensional measurement, but the intent is similar in rigorously defining evidence for the activity of the agent in assessing its value to the patient. An active chemotherapy agent conventionally has PR rates of at least 20–25% with reversible non-life-threatening side effects, and it may then be suitable for study in phase 3 trials to assess efficacy in comparison to standard or no therapy. Active efforts are being made to quantitate effects of anticancer agents on quality of life. Cancer drug clinical trials conventionally use a toxicity grading scale where grade 1 toxicities do not require treatment, grade 2 toxicities may require

symptomatic treatment but are not life-threatening, grade 3 toxicities are potentially life-threatening if untreated, grade 4 toxicities are actually life-threatening, and grade 5 toxicities are those that result in the patient's death.

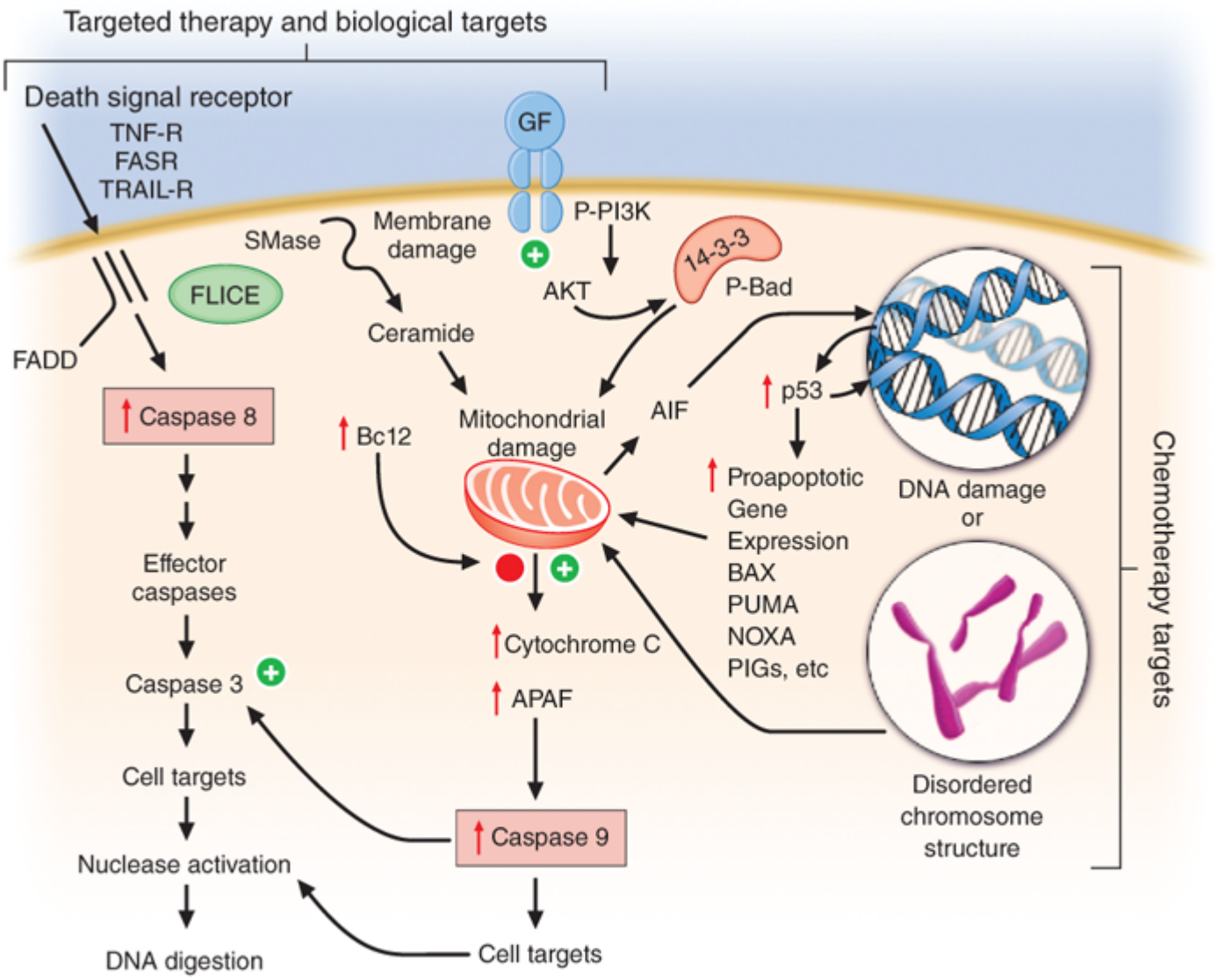
Development of targeted agents may proceed quite differently. While phase 1–3 trials are still conducted, molecular analysis of human tumors may allow the precise definition of target expression in a patient's tumor that is necessary for or relevant to the drug's action. This information might then allow selection of patients expressing the drug target for participation in all trial phases. These patients may then have a greater chance of developing a useful response to the drug by virtue of expressing the target in the tumor. Clinical trials may be designed to incorporate an assessment of the behavior of the target in relation to the drug (pharmacodynamic studies). Ideally, the plasma concentration that affects the drug target is known, so escalation to MTD may not be necessary. Rather, the correlation of host toxicity while achieving an "optimal biologic dose" becomes a more relevant endpoint for phase 1 and early phase 2 trials with targeted agents.

Useful cancer drug treatment strategies using conventional chemotherapy agents, targeted agents, hormonal treatments, or biologics have one of two valuable outcomes. They can induce cancer cell death, resulting in tumor shrinkage with corresponding improvement in patient survival, or increase the time until the disease progresses. Another potential outcome is to induce cancer cell *differentiation* or *dormancy* with loss of tumor cell replicative potential and reacquisition of phenotypic properties resembling normal cells. A general view of how cancer treatments work is that the interaction of a chemotherapeutic drug with its target induces a "cascade" of further signaling steps. These signals ultimately lead to cell death by triggering an "execution phase" where proteases, nucleases, and endogenous regulators of the cell death pathway are activated **(Fig. 69-3)**.

FIGURE 69-3

Integration of cell death responses. Cell death through apoptosis requires active participation of the cell. In response to interruption of growth factor (GF) or propagation of certain cytokine death signals (e.g., tumor necrosis factor receptor [TNF-R]), there is activation of "upstream" cysteine aspartyl proteases (caspases), which then directly digest cytoplasmic and nuclear proteins, resulting in activation of "downstream" caspases; these cause activation of nucleases, resulting in the characteristic DNA fragmentation that is a hallmark of apoptosis. Chemotherapy agents that create lesions in DNA or alter mitotic spindle function seem to activate aspects of this process by damage ultimately conveyed to the mitochondria, perhaps by activating the transcription of genes whose products can produce or modulate the toxicity of free radicals. In addition, membrane damage with activation of sphingomyelinases results in the production of ceramides that can have a direct action at mitochondria. The antiapoptotic protein bcl2 attenuates mitochondrial toxicity, while proapoptotic gene products such as bax antagonize the action of bcl2. Damaged mitochondria release cytochrome C and apoptosis-activating factor (APAF), which can directly activate caspase 9, resulting in propagation of a direct signal to other downstream caspases through protease activation. Apoptosis-inducing factor (AIF) is also released from the mitochondrion and then can translocate to the nucleus, bind to DNA, and generate free radicals to further damage DNA. An additional proapoptotic stimulus is the *bad* protein, which can heterodimerize with *bcl2* gene family members to antagonize apoptosis. Importantly, though, bad protein function can be retarded by its sequestration as phospho-*bad* through the 14-3-3

adapter proteins. The phosphorylation of bad is mediated by the action of the AKT kinase in a way that defines how growth factors that activate this kinase can retard apoptosis and promote cell survival.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Targeted agents differ from chemotherapy agents in that they do not indiscriminately cause macromolecular lesions but regulate the action of particular pathways. For example, the $p210^{bcr\text{-}abl}$ fusion protein tyrosine kinase drives chronic myeloid leukemia (CML), and HER2/neu stimulates the proliferation of certain breast cancers. The tumor has been described as "addicted" to the function of these molecules in the sense that without the pathway's continued action, the tumor cell cannot survive. In this way, targeted agents directed at $p210^{bcr\text{-}abl}$ or HER2/neu may alter the "threshold" tumors driven by these molecules may have for undergoing cell death without actually creating any molecular lesions such as direct DNA strand breakage or altered membrane function.

Chemotherapy agents may be used for the treatment of active, clinically apparent cancer. The goal of such treatment in some cases is cure of the cancer, that is, elimination of all clinical and pathologic evidence of cancer and return of the patient to an expected survival no different than the general population. **Table 69-3,**

A lists those tumors considered curable by conventionally available chemotherapeutic agents when used to address disseminated or metastatic cancers. If a tumor is localized to a single site, serious consideration of surgery or primary radiation therapy should be given, because these treatment modalities may be curative as local treatments. Chemotherapy may then be used after the failure of these modalities to eradicate a local tumor or as part of multimodality approaches to offer primary treatment to a clinically localized tumor. In this event, it can allow organ preservation when given with radiation, as in the larynx or other upper airway sites, or sensitize tumors to radiation when given, e.g., to patients concurrently receiving radiation for lung or cervix cancer (Table 69-3, B). Chemotherapy can be administered as an *adjuvant*, i.e., in addition to surgery or radiation (Table 69-3, C), even after all clinically apparent disease has been removed. This use of chemotherapy has curative potential in breast and colorectal neoplasms, as it attempts to eliminate clinically unapparent tumor that may have already disseminated. *Neoadjuvant* chemotherapy refers to administration of chemotherapy before any surgery or radiation to a local tumor in an effort to enhance the effect of the local treatment.

TABLE 69-3

Curability of Cancers with Chemotherapy

A. Advanced Cancers with Possible Cure

Acute lymphoid and acute myeloid leukemia (pediatric/adult)

Hodgkin's disease (pediatric/adult)

Lymphomas—certain types (pediatric/adult)

Germ cell neoplasms

- Embryonal carcinoma
- Teratocarcinoma
- Seminoma or dysgerminoma
- Choriocarcinoma

Gestational trophoblastic neoplasia

Pediatric neoplasms

- Wilms' tumor
- Embryonal rhabdomyosarcoma
- Ewing's sarcoma
- Peripheral neuroepithelioma
- Neuroblastoma

Small-cell lung carcinoma

Ovarian carcinoma

B. Advanced Cancers Possibly Cured by Chemotherapy and Radiation

Squamous carcinoma (head and neck)

Squamous carcinoma (anus)

Breast carcinoma

Carcinoma of the uterine cervix

Non-small-cell lung carcinoma (stage III)

Small-cell lung carcinoma

C. Cancers Possibly Cured with Chemotherapy as Adjuvant to Surgery

Breast carcinoma

Colorectal carcinoma[a]

Osteogenic sarcoma

Soft tissue sarcoma

D. Cancers Possibly Cured with "High-Dose" Chemotherapy with Stem Cell Support

Relapsed leukemias, lymphoid and myeloid

Relapsed lymphomas, Hodgkin's and non-Hodgkin's

Chronic myeloid leukemia

Multiple myeloma

E. Cancers Responsive with Useful Palliation, But Not Cure, by Chemotherapy

Bladder carcinoma

Chronic myeloid leukemia
Hairy cell leukemia
Chronic lymphocytic leukemia
Lymphoma—certain types
Multiple myeloma
Gastric carcinoma
Cervix carcinoma
Endometrial carcinoma
Soft tissue sarcoma
Head and neck cancer
Adrenocortical carcinoma
Islet cell neoplasms
Breast carcinoma
Colorectal carcinoma
Renal carcinoma

F. Tumors Poorly Responsive in Advanced Stages to Chemotherapy

Pancreatic carcinoma
Biliary tract neoplasms
Thyroid carcinoma
Carcinoma of the vulva
Non-small-cell lung carcinoma
Prostate carcinoma
Melanoma (subsets)
Hepatocellular carcinoma
Salivary gland cancer

[a]Rectum also receives radiation therapy.

Chemotherapy is routinely used in “conventional” dose regimens. In general, these doses produce reversible acute side effects, primarily consisting of transient myelosuppression with or without gastrointestinal toxicity (usually nausea), which are readily managed. “High-dose” chemotherapy regimens are predicated on the observation that the dose-response curve for many anticancer agents is rather steep, and increased dose can produce markedly increased therapeutic effect, although at the cost of potentially life-threatening complications that require intensive support, usually in the form of hematopoietic stem cell support from the patient (*autologous*) or from donors matched for histocompatibility loci (*allogeneic*), or pharmacologic “rescue” strategies to repair the effect of the high-dose chemotherapy on normal tissues. High-dose regimens have definite curative potential in defined clinical settings (Table 69-3, D).

If cure is not possible, chemotherapy may be undertaken with the goal of palliating some aspect of the tumor’s effect on the host. In this usage, value is perceived by the demonstration of improved symptom

relief, progression-free survival, or overall survival at a certain time from the inception of treatment in the treated population, compared to a relevant control population established as the result of a clinical research protocol as a basis for U.S. Food and Drug Administration (FDA) approval of a particular cancer treatment as safe and effective. Common tumors that may be meaningfully addressed by chemotherapy with palliative intent are listed in **Table 69-3, E**.

Usually, tumor-related symptoms manifest as pain, weight loss, or some local symptom related to the tumor's effect on normal structures. Patients treated with palliative intent should be aware of their diagnosis and the limitations of the proposed treatments, have access to supportive care, and have suitable "performance status," according to assessment algorithms such as the one developed by Karnofsky **(see Table 65-4)** or by the Eastern Cooperative Oncology Group (ECOG) **(see Table 65-5)**. ECOG performance status 0 (PS0) patients are without symptoms; PS1 patients are ambulatory but restricted in strenuous physical activity; PS2 patients are ambulatory but unable to work and are up and about 50% or more of the time; PS3 patients are capable of limited self-care and are up <50% of the time; and PS4 patients are totally confined to bed or chair and incapable of self-care. Only PS0, PS1, and PS2 patients are generally considered suitable for palliative (noncurative) treatment. If there is curative potential, even poor–performance status patients may be treated, but their prognosis is usually inferior to that of good–performance status patients treated with similar regimens.

An important perspective the primary care provider may bring to patients and their families facing incurable cancer is that, given the limited value of chemotherapeutic approaches at some point in the natural history of most metastatic cancers, *palliative care* or *hospice-based* approaches, with meticulous and ongoing attention to symptom relief and with family, psychological, and spiritual support, should receive prominent attention as a valuable therapeutic plan **(Chaps. 9 and 65)**. Optimizing the quality of life rather than attempting to extend it becomes a valued intervention. Patients facing the impending progression of disease in a life-threatening way frequently choose to undertake toxic treatments of little to no potential value, and support provided by the primary caregiver in accessing palliative and hospice-based options in contrast to receiving toxic and ineffective regimen can be critical in providing a basis for patients to make sensible choices.

Cytotoxic Chemotherapy Agents

Table 69-4 lists commonly used cytotoxic cancer chemotherapy agents and pertinent clinical aspects of their use, with particular reference to adverse effects that might be encountered by the generalist in the care of patients. The drugs listed may be usefully grouped into two general categories: those affecting DNA and those affecting microtubules.

TABLE 69-4

Cytotoxic Chemotherapy Agents

Drug	Toxicity	Interactions, Issues
Direct DNA-Interacting Agents		
Alkylator		
Cyclophosphamide	Marrow (relative platelet sparing) Cystitis Common alkylator[a] Cardiac (high dose)	Liver metabolism required to activate to phosphoramide mustard + acrolein Mesna protects against “high-dose” bladder damage
Melphalan	Marrow (delayed nadir) GI (high dose)	Decreased renal function delays clearance
Carmustine (BCNU)	Marrow (delayed nadir) GI, liver (high dose) Renal	
Lomustine (CCNU)	Marrow (delayed nadir)	
Ifosfamide	Myelosuppressive Bladder Neurologic Metabolic acidosis	Analogue of cyclophosphamide Must use mesna Greater activity vs testicular neoplasms and sarcomas
Procarbazine	Marrow Nausea Neurologic Common alkylator[a]	Liver and tissue metabolism required Disulfiram-like effect with ethanol Acts as MAOI HBP after tyrosinase-rich foods
Dacarbazine (DTIC)	Marrow Nausea Flulike	Metabolic activation

Drug	Toxicity	Interactions, Issues
Temozolomide	Nausea/vomiting Headache/fatigue Constipation	Infrequent myelosuppression
Cisplatin	Nausea Neuropathy Auditory Marrow platelets > WBCs Renal Mg^{2+}, Ca^{2+}	Maintain high urine flow; osmotic diuresis, monitor intake/output K^{+}, Mg^{2+} Emetogenic—prophylaxis needed Full dose if CrCl >60 mL/min and tolerate fluid push
Carboplatin	Marrow platelets > WBCs Nausea Renal (high dose)	Reduce dose according to CrCl: to AUC of 5–7 mg/mL per min [AUC = dose/(CrCl + 25)]
Oxaliplatin	Nausea Anemia	Acute reversible neurotoxicity; chronic sensory neurotoxicity cumulative with dose; reversible laryngopharyngeal spasm
Antitumor Antibiotics and Topoisomerase Poisons		
Bleomycin	Pulmonary Skin effects Raynaud's Hypersensitivity	Inactivate by bleomycin hydrolase (decreased in lung/skin) O_2 enhances pulmonary toxicity Cisplatin-induced decrease in CrCl may increase skin/lung toxicity Reduce dose if CrCl <60 mL/min
Dactinomycin	Marrow Nausea Mucositis Vesicant Alopecia	Radiation recall

Drug	Toxicity	Interactions, Issues
Etoposide (VP16-213)	Marrow (WBCs > platelet) Alopecia Hypotension Hypersensitivity (rapid IV) Nausea Mucositis (high dose)	Hepatic metabolism—renal 30% Reduce doses with renal failure Schedule-dependent (5-day schedule better than 1-day) Late leukemogenic Accentuate antimetabolite action
Topotecan	Marrow Mucositis Nausea Mild alopecia	Reduce dose with renal failure No liver toxicity
Irinotecan	Diarrhea: “early onset” with cramping, flushing, vomiting; “late onset” after several doses Marrow Alopecia Nausea Vomiting Pulmonary	Prodrug requires enzymatic clearance to active drug “SN 38” Early diarrhea due to acetylcholine release Late diarrhea, use “high-dose” loperamide (2 mg q2–4 h)
Doxorubicin and daunorubicin	Marrow Mucositis Alopecia Cardiovascular acute/chronic Vesicant	Heparin aggregate; coadministration increases clearance Acetaminophen, BCNU increase liver toxicity Radiation recall
Idarubicin	Marrow Cardiac (less than doxorubicin)	None established
Epirubicin	Marrow Cardiac	None established

Drug	Toxicity	Interactions, Issues
Mitoxantrone	Marrow Cardiac (less than doxorubicin) Vesicant (mild) Blue urine, sclerae, nails	Interacts with heparin Less alopecia, nausea than doxorubicin Radiation recall Less alopecia, nausea than doxorubicin
Indirectly DNA-Interacting Agents		
Antimetabolites		
6-Mercaptopurine (6-MP)	Marrow Liver Nausea	Variable bioavailability Metabolize by xanthine oxidase Decrease dose with allopurinol Increased toxicity with thiopurine methyltransferase deficiency
6-Thioguanine	Marrow Liver Nausea	Variable bioavailability Increased toxicity with thiopurine methyltransferase deficiency
2-Chlorodeoxyadenosine	Marrow Renal Fever	Notable use in hairy cell leukemia
Hydroxyurea	Marrow Nausea Mucositis Skin changes Rare renal, liver, lung, CNS	Decrease dose with renal failure Augments antimetabolite effect
Methotrexate	Marrow Liver/lung Renal tubular Mucositis	Toxicity lessened by “rescue” with leucovorin Excreted in urine Decrease dose in renal failure; NSAIDs increase renal toxicity
Pemetrexed	Anemia Neutropenia	Supplement folate/B_{12} Caution in renal failure

Drug	Toxicity	Interactions, Issues
Pralatrexate	Thrombocytopenia Myelosuppression Mucositis	Active in peripheral T cell lymphoma
5-Fluorouracil (5FU)	Marrow Mucositis Neurologic Skin changes	Toxicity enhanced by leucovorin by increasing "ternary complex" with thymidylate synthase; dihydropyrimidine dehydrogenase deficiency increases toxicity; metabolism in tissue
Capecitabine	Diarrhea Hand-foot syndrome	Prodrug of 5FU due to intratumoral metabolism
Cytosine arabinoside	Marrow Mucositis Neurologic (high dose) Conjunctivitis (high dose) Noncardiogenic pulmonary edema	Enhances activity of alkylating agents Metabolizes in tissues by deamination but renal excretion prominent at doses >500 mg; therefore, dose reduce in "high-dose" regimens in patients with decreased CrCl
Azacitidine Decitabine	Marrow Nausea Liver Neurologic Myalgia	Use limited to leukemia/myelodysplastic syndrome Altered methylation of DNA alters gene expression
Gemcitabine	Marrow Nausea Hepatic Fever/"flu syndrome"	
Fludarabine phosphate	Marrow Neurologic Lung	Dose reduction with renal failure Metabolized to F-ara converted to F-ara ATP in cells by deoxycytidine kinase

Drug	Toxicity	Interactions, Issues
Asparaginase	Decrease protein synthesis; indirect inhibition of DNA synthesis by decreased histone synthesis Clotting factors Glucose Albumin Hypersensitivity CNS Pancreatitis Hepatic	Blocks methotrexate action
Antimitotic Agents		
Vincristine	Vesicant Marrow Neurologic GI: ileus/constipation; bladder hypotoxicity; SIADH Cardiovascular	Hepatic clearance Dose reduction for bilirubin >1.5 mg/dL Prophylactic bowel regimen
Vinblastine	Vesicant Marrow Neurologic (less common but similar spectrum to other vincas) Hypertension Raynaud's	Hepatic clearance Dose reduction as with vincristine

Drug	Toxicity	Interactions, Issues
Vinorelbine	Vesicant Marrow Allergic/bronchospasm (immediate) Dyspnea/cough (subacute) Neurologic (less prominent but similar spectrum to other vincas)	Hepatic clearance
Paclitaxel	Hypersensitivity Marrow Mucositis Alopecia Sensory neuropathy CV conduction disturbance Nausea—infrequent	Premedicate with steroids, H_1 and H_2 blockers Hepatic clearance Dose reduction as with vincas
Docetaxel	Hypersensitivity Fluid retention syndrome Marrow Dermatologic Sensory neuropathy Nausea infrequent Some stomatitis	Premedicate with steroids, H_1 and H_2 blockers
Nab-paclitaxel (protein bound)	Neuropathy Anemia Neutropenia Thrombocytopenia	Caution in hepatic insufficiency
Ixabepilone	Myelosuppression Neuropathy	

[a]Common allkylator: alopecia, pulmonary, infertility, plus teratogenesis.

Abbreviations: ALL, acute lymphocytic leukemia; AUC, area under the curve; CHF, congestive heart failure; CNS, central nervous system; CrCl, creatinine clearance; CV, cardiovascular; GI, gastrointestinal; HBP, high blood pressure; MAOI, monoamine oxidase inhibitor; NSAID, nonsteroidal anti-inflammatory drug; SIADH, syndrome of inappropriate antidiuretic hormone secretion.

DNA-INTERACTIVE AGENTS

DNA replication occurs during the synthesis or S-phase of the cell cycle, with chromosome segregation of the replicated DNA occurring in the M, or mitosis, phase. The G_1 and G_2 "gap phases" precede S and M, respectively. Historically, chemotherapeutic agents have been divided into "phase-nonspecific" agents, which can act in any phase of the cell cycle, and "phase-specific" agents, which require the cell to be at a particular cell cycle phase to cause greatest effect. "Checkpoints" in the cell cycle exist where the drug-related damage may be assessed and either repaired or cell death initiated.

Alkylating agents as a class are cell cycle phase–nonspecific agents. They break down, either spontaneously or after normal organ or tumor cell metabolism, to reactive intermediates that covalently modify bases in DNA. This leads to cross-linkage of DNA strands or the appearance of breaks in DNA as a result of repair efforts. "Broken" or cross-linked DNA is intrinsically unable to complete normal replication or cell division; in addition, it is a potent activator of cell cycle checkpoints and further activates cell-signaling pathways that can precipitate apoptosis. Alkylating agents share similar toxicities: myelosuppression, alopecia, gonadal dysfunction, mucositis, and pulmonary fibrosis. They differ greatly in a spectrum of normal organ toxicities. They also share the capacity to cause "second" neoplasms, particularly leukemia, many years after use, particularly when used in low doses for protracted periods.

Cyclophosphamide is inactive unless metabolized by the liver to 4-hydroxy-cyclophosphamide, which decomposes into an alkylating species, as well as to chloroacetaldehyde and acrolein. The latter causes chemical cystitis; therefore, excellent hydration must be maintained while using cyclophosphamide. If severe, the cystitis may be attenuated or prevented altogether (if expected from the dose of cyclophosphamide to be used) by mesna (2-*m*ercapto*e*thane*s*ulfo*na*te). Liver disease impairs cyclophosphamide activation. Sporadic interstitial pneumonitis leading to pulmonary fibrosis can accompany the use of cyclophosphamide, and high doses used in conditioning regimens for bone marrow transplant can cause cardiac dysfunction. Ifosfamide is a cyclophosphamide analogue also activated in the liver, but more slowly, and it requires coadministration of mesna to prevent bladder injury. CNS effects, including somnolence, confusion, and psychosis, can follow ifosfamide use; the incidence appears related to low body surface area or decreased creatinine clearance.

Several alkylating agents are less commonly used. Bendamustine is a nitrogen mustard derivative with evidence of activity in chronic lymphocytic leukemia and certain lymphomas. Busulfan can cause profound myelosuppression, alopecia, and pulmonary toxicity but is relatively "lymphocyte sparing." Its routine use in treatment of CML has been curtailed in favor of imatinib (Gleevec) or dasatinib, but it is still used in transplant preparation regimens. Melphalan shows variable oral bioavailability and undergoes extensive binding to albumin and α_1-acidic glycoprotein. Mucositis appears more prominently; however, it has prominent activity in multiple myeloma.

Nitrosoureas break down to carbamylating species that not only cause a distinct pattern of DNA base pair–directed toxicity but also can covalently modify proteins. They share the feature of causing relatively delayed bone marrow toxicity, which can be cumulative and long-lasting. Procarbazine is metabolized in the liver and possibly in tumor cells to yield a variety of free radical and alkylating species. In addition to myelosuppression, it causes hypnotic and other CNS effects, including vivid nightmares. It can cause a disulfiram-like syndrome on ingestion of ethanol. Dacarbazine (DTIC) is activated in the liver to yield the highly reactive methyl diazonium cation. It causes only modest myelosuppression 21–25 days after a dose but causes prominent nausea on day 1. Temozolomide is structurally related to dacarbazine but was designed to be activated by nonenzymatic hydrolysis in tumors and is bioavailable orally. Brain tumors with alkylguanine alkyl transferase deficiency are selectively susceptible to temozolomide, which alkylates the O^6 position of guanine.

Cisplatin was discovered fortuitously by observing that bacteria present in electrolysis solutions with platinum electrodes could not divide. Only the *cis* diamine configuration is active as an antitumor agent. In the intracellular environment, a chloride is lost from each position, being replaced by a water molecule. The resulting positively charged species is an efficient bifunctional interactor with DNA, forming Pt-based cross-links. Cisplatin requires administration with adequate hydration, including forced diuresis with mannitol to prevent kidney damage; even with the use of hydration, gradual decrease in kidney function is common, along with noteworthy anemia. Hypomagnesemia frequently attends cisplatin use and can lead to hypocalcemia and tetany. Other common toxicities include neurotoxocity with stocking-and-glove sensorimotor neuropathy. Hearing loss occurs in 50% of patients treated with conventional doses. Cisplatin is intensely emetogenic, requiring prophylactic antiemetics. Myelosuppression is less evident than with other alkylating agents. Chronic vascular toxicity (Raynaud's phenomenon, coronary artery disease) is a more unusual toxicity. Carboplatin displays less nephro-, oto-, and neurotoxicity. However, myelosuppression is more frequent, and because the drug is exclusively cleared through the kidney, adjustment of dose for creatinine clearance must be accomplished through use of various dosing nomograms. Oxaliplatin is a platinum analogue with noteworthy activity in colon cancers refractory to other treatments. It is prominently neurotoxic.

ANTITUMOR ANTIBIOTICS AND TOPOISOMERASE POISONS

Antitumor antibiotics are substances produced by bacteria that in nature appear to provide a chemical defense against other hostile microorganisms. As a class, they bind to DNA directly and can frequently undergo electron transfer reactions to generate free radicals in close proximity to DNA, leading to DNA damage in the form of single-strand breaks or cross-links. Topoisomerase poisons include natural products or semisynthetic species derived ultimately from plants, and they modify enzymes that regulate the capacity of DNA to unwind to allow normal replication or transcription. These include topoisomerase I, which creates single-strand breaks that then rejoin following the passage of the other DNA strand through the break. Topoisomerase II creates double-strand breaks through which another segment of DNA duplex passes before rejoining. Owing to the role of topoisomerase I in the procession of the replication fork, topoisomerase I poisons cause lethality if the topoisomerase I–induced lesions are made in S-phase.

Doxorubicin can intercalate into DNA, thereby altering DNA structure, replication, and topoisomerase II function. It can also undergo reduction reactions by accepting electrons into its quinone ring system, with the capacity to undergo reoxidation to form reactive oxygen radicals after reoxidation. It causes predictable myelosuppression, alopecia, nausea, and mucositis. In addition, it causes acute cardiotoxicity in the form of atrial and ventricular dysrhythmias, but these are rarely of clinical significance. In contrast, cumulative doses >550 mg/m^2 are associated with a 10% incidence of chronic cardiomyopathy. The incidence of cardiomyopathy appears to be related to schedule (peak serum concentration), with low-dose, frequent treatment or continuous infusions better tolerated than intermittent higher-dose exposures. Cardiotoxicity has been related to iron-catalyzed oxidation and reduction of doxorubicin. Cardiotoxicity is related to peak plasma dose; thus, lower doses and continuous infusions are less likely to cause heart damage. Doxorubicin's cardiotoxicity is increased when given together with trastuzumab (Herceptin), the anti-HER2/neu antibody. Radiation recall or interaction with concomitantly administered radiation to cause local site complications is frequent. The drug is a powerful vesicant, with necrosis of tissue apparent 4–7 days after an extravasation; therefore, it should be administered into a rapidly flowing intravenous line. Dexrazoxane is an antidote to doxorubicin-induced extravasation. Doxorubicin is metabolized by the liver, so doses must be reduced by 50–75% in the presence of liver dysfunction. Daunorubicin is closely related to doxorubicin and was actually introduced first into leukemia treatment, where it remains part of curative regimens and has been shown preferable to doxorubicin owing to less mucositis and colonic damage. Idarubicin is also used in acute myeloid leukemia treatment and may be preferable to daunorubicin in activity. Encapsulation of daunorubicin into a liposomal formulation has attenuated cardiac toxicity and antitumor activity in Kaposi's sarcoma, other sarcomas, multiple myeloma, and ovarian cancer.

Bleomycin refers to a mixture of glycopeptides that have the unique feature of forming complexes with Fe^{2+} while also bound to DNA. It remains an important component of curative regimens for Hodgkin's disease and germ cell neoplasms. Oxidation of Fe^{2+} gives rise to superoxide and hydroxyl radicals. The drug causes little, if any, myelosuppression. The drug is cleared rapidly, but augmented skin and pulmonary toxicity in the presence of renal failure has led to the recommendation that doses be reduced by 50–75% in the face of a creatinine clearance <25 mL/min. Bleomycin is not a vesicant and can be administered intravenously, intramuscularly, or subcutaneously. Common side effects include fever and chills, facial flush, and Raynaud's phenomenon. The most feared complication of bleomycin treatment is pulmonary fibrosis, which increases in incidence at >300 cumulative units administered and is minimally responsive to treatment (e.g., glucocorticoids). The earliest indicator of an adverse effect is usually a decline in the carbon monoxide diffusing capacity (DL_{CO}) or coughing, although cessation of drug immediately upon documentation of a decrease in DL_{CO} may not prevent further decline in pulmonary function. Bleomycin is inactivated by a bleomycin hydrolase, whose concentration is diminished in skin and lung. Because bleomycin-dependent electron transport is dependent on O_2, bleomycin toxicity may become apparent after exposure to transient very high fraction of inspired oxygen (FI_{O2}). Thus, during surgical procedures, patients with prior exposure to bleomycin should be maintained on the lowest FI_{O2} consistent with maintaining adequate tissue oxygenation.

Mitoxantrone is a synthetic compound that was designed to recapitulate features of doxorubicin but with less cardiotoxicity. It is quantitatively less cardiotoxic (comparing the ratio of cardiotoxic to therapeutically effective doses), but is still associated with a 10% incidence of cardiotoxicity at cumulative doses of >150 mg/m^2. It also causes alopecia. Etoposide binds directly to topoisomerase II and DNA in a reversible ternary complex. It stabilizes the covalent intermediate in the enzyme's action where the enzyme is covalently linked to DNA. Prominent clinical effects include myelosuppression, nausea, and transient hypotension related to the speed of administration of the agent. Etoposide is a mild vesicant but is relatively free from other large-organ toxicities. Camptothecins target topoisomerase I. Topotecan is a camptothecin-derivative approved for use in gynecologic tumors and small-cell lung cancer. Toxicity is limited to myelosuppression and mucositis. CPT-11, or irinotecan, is a camptothecin with evidence of activity in colon carcinoma. In addition to myelosuppression, it causes a secretory diarrhea related to the toxicity of a metabolite called SN-38. Levels of SN-38 are particularly high in the setting of Gilbert's disease, characterized by defective glucuronyl transferase and indirect hyperbilirubinemia, a condition that affects about 10% of the white population in the United States. The diarrhea can be treated effectively with loperamide or octreotide.

ANTIMETABOLITES

A broad definition of antimetabolites would include compounds with structural similarity to precursors of purines or pyrimidines, or compounds that interfere with purine or pyrimidine synthesis. Some antimetabolites can cause DNA damage indirectly, through misincorporation into DNA, abnormal timing or progression through DNA synthesis, or altered function of pyrimidine and purine biosynthetic enzymes. They tend to convey greatest toxicity to cells in S-phase, and the degree of toxicity increases with duration of exposure. Common toxic manifestations include stomatitis, diarrhea, and myelosuppression. Second malignancies are not associated with their use.

Methotrexate inhibits dihydrofolate reductase, which regenerates reduced folates from the oxidized folates produced when thymidine monophosphate is formed from deoxyuridine monophosphate. Without reduced folates, cells die a "thymine-less" death. *N*5-Tetrahydrofolate or *N*5-formyltetrahydrofolate (leucovorin) can bypass this block and rescue cells from methotrexate, which is maintained in cells by polyglutamylation. The drug and other reduced folates are transported into cells by a membrane carrier, and high concentrations of drug can bypass this carrier and allow diffusion of drug directly into cells. These properties have suggested the design of "high-dose" methotrexate regimens with leucovorin rescue of normal marrow and mucosa as part of curative approaches to osteosarcoma in the adjuvant setting and hematopoietic neoplasms of children and adults. Methotrexate is cleared by the kidney via both glomerular filtration and tubular secretion, and toxicity is augmented by renal dysfunction and drugs such as salicylates, probenecid, and nonsteroidal anti-inflammatory agents that undergo tubular secretion. With normal renal function, 15 mg/m^2 leucovorin will rescue 10^{-8}–10^{-6} *M* methotrexate in 3–4 doses. However, with decreased creatinine clearance, doses of 50–100 mg/m^2 are continued until methotrexate levels are $<5 \times 10^{-8}$ *M*. In addition to bone marrow suppression and mucosal irritation, methotrexate can cause renal failure itself at high doses owing to crystallization in renal tubules; therefore, high-dose regimens require alkalinization of urine with increased flow by hydration. Methotrexate can be sequestered in third-space collections and diffuse back into the general circulation, causing prolonged myelosuppression. Less frequent adverse effects include

reversible increases in transaminases and hypersensitivity-like pulmonary syndrome. Chronic low-dose methotrexate can cause hepatic fibrosis. When administered to the intrathecal space, methotrexate can cause chemical arachnoiditis and CNS dysfunction.

Pemetrexed is a folate-directed antimetabolite. It inhibits the activity of several enzymes, including thymidylate synthetase (TS), dihydrofolate reductase, and glycinamide ribonucleotide formyltransferase, thereby affecting the synthesis of both purine and pyrimidine nucleic acid precursors. To avoid significant toxicity to the normal tissues, patients receiving pemetrexed should also receive low-dose folate and vitamin B_{12} supplementation. Pemetrexed has notable activity against certain lung cancers and, in combination with cisplatin, also against mesotheliomas. Pralatrexate is an antifolate approved for use in T-cell lymphoma that is very efficiently transported into cancer cells.

5-Fluorouracil (5FU) represents an early example of "rational" drug design in that it originated from the observation that tumor cells incorporate radiolabeled uracil more efficiently into DNA than normal cells, especially gut. 5FU is metabolized in cells to 5′FdUMP, which inhibits TS. In addition, misincorporation can lead to single-strand breaks, and RNA can aberrantly incorporate FUMP. 5FU is metabolized by dihydropyrimidine dehydrogenase, and deficiency of this enzyme can lead to excessive toxicity from 5FU. Oral bioavailability varies unreliably, but prodrugs such as capecitabine have been developed that allow at least equivalent activity to many parenteral 5FU-based approaches. Intravenous administration of 5FU leads to bone marrow suppression after short infusions but to stomatitis after prolonged infusions. Leucovorin augments the activity of 5FU by promoting formation of the ternary covalent complex of 5FU, the reduced folate, and TS. Less frequent toxicities include CNS dysfunction, with prominent cerebellar signs, and endothelial toxicity manifested by thrombosis, including pulmonary embolus and myocardial infarction.

Cytosine arabinoside (ara-C) is incorporated into DNA after formation of ara-CTP, resulting in S-phase–related toxicity. Continuous infusion schedules allow maximal efficiency, with uptake maximal at 5–7 μ*M*. Ara-C can be administered intrathecally. Adverse effects include nausea, diarrhea, stomatitis, chemical conjunctivitis, and cerebellar ataxia. Gemcitabine is a cytosine derivative that is similar to ara-C in that it is incorporated into DNA after anabolism to the triphosphate, rendering DNA susceptible to breakage and repair synthesis, which differs from that in ara-C in that gemcitabine-induced lesions are very inefficiently removed. In contrast to ara-C, gemcitabine appears to have useful activity in a variety of solid tumors, with limited nonmyelosuppressive toxicities.

6-Thioguanine and 6-mercaptopurine (6MP) are used in the treatment of acute lymphoid leukemia. Although administered orally, they display variable bioavailability. 6MP is metabolized by xanthine oxidase and therefore requires dose reduction when used with allopurinol. 6MP is also metabolized by thiopurine methyltransferase; genetic deficiency of thiopurine methyltransferase results in excessive toxicity.

Fludarabine phosphate is a prodrug of F-adenine arabinoside (F-ara-A), which in turn was designed to diminish the susceptibility of ara-A to adenosine deaminase. F-ara-A is incorporated into DNA and can cause delayed cytotoxicity even in cells with low growth fraction, including chronic lymphocytic leukemia and follicular B-cell lymphoma. CNS and peripheral nerve dysfunction and T-cell depletion leading to

opportunistic infections can occur in addition to myelosuppression. 2-Chlorodeoxyadenosine is a similar compound with activity in hairy cell leukemia. Hydroxyurea inhibits ribonucleotide reductase, resulting in S-phase block. It is orally bioavailable and useful for the acute management of myeloproliferative states.

Asparaginase is a bacterial enzyme that causes breakdown of extracellular asparagine required for protein synthesis in certain leukemic cells. This effectively stops tumor cell DNA synthesis, as DNA synthesis requires concurrent protein synthesis. The outcome of asparaginase action is therefore very similar to the result of the small-molecule antimetabolites. Because asparaginase is a foreign protein, hypersensitivity reactions are common, as are effects on organs such as pancreas and liver that normally require continuing protein synthesis. This may result in decreased insulin secretion with hyperglycemia, with or without hyperamylasemia and clotting function abnormalities. Close monitoring of clotting functions should accompany use of asparaginase. Paradoxically, owing to depletion of rapidly turning over anticoagulant factors, thromboses particularly affecting the CNS may also be seen with asparaginase.

MITOTIC SPINDLE INHIBITORS

Microtubules are cellular structures that form the mitotic spindle, and in interphase cells, they are responsible for the cellular "scaffolding" along which various motile and secretory processes occur. Microtubules are composed of repeating noncovalent multimers of a heterodimer of α and β isoform of the protein tubulin. Vincristine binds to the tubulin dimer with the result that microtubules are disaggregated. This results in the block of growing cells in M-phase; however, toxic effects in G_1 and S-phase are also evident, reflecting effects on normal cellular activities of microtubules. Vincristine is metabolized by the liver, and dose adjustment in the presence of hepatic dysfunction is required. It is a powerful vesicant, and infiltration can be treated by local heat and infiltration of hyaluronidase. At clinically used intravenous doses, neurotoxicity in the form of glove-and-stocking neuropathy is frequent. Acute neuropathic effects include jaw pain, paralytic ileus, urinary retention, and the syndrome of inappropriate antidiuretic hormone secretion. Myelosuppression is not seen. Vinblastine is similar to vincristine, except that it tends to be more myelotoxic, with more frequent thrombocytopenia and also mucositis and stomatitis. Vinorelbine is a vinca alkaloid that appears to have differences in resistance patterns in comparison to vincristine and vinblastine; it may be administered orally.

The taxanes include paclitaxel and docetaxel. These agents differ from the vinca alkaloids in that the taxanes stabilize microtubules against depolymerization. The "stabilized" microtubules function abnormally and are not able to undergo the normal dynamic changes of microtubule structure and function necessary for cell cycle completion. Taxanes are among the most broadly active antineoplastic agents for use in solid tumors, with evidence of activity in ovarian cancer, breast cancer, Kaposi's sarcoma, and lung tumors. They are administered intravenously, and paclitaxel requires use of a Cremophor-containing vehicle that can cause hypersensitivity reactions. Premedication with dexamethasone (8–16 mg orally or intravenously 12 and 6 h before treatment) and diphenhydramine (50 mg) and cimetidine (300 mg), both 30 min before treatment, decreases but does not eliminate the risk of hypersensitivity reactions to the paclitaxel vehicle. A protein-bound formulation of paclitaxel (called *nab-paclitaxel*) has at least equivalent antineoplastic activity and decreased risk of hypersensitivity reactions. Paclitaxel may also cause hypersensitivity reactions, myelosuppression, neurotoxicity in the form of glove-and-stocking numbness, and paresthesia. Docetaxel

causes comparable degrees of myelosuppression and neuropathy. Docetaxel uses a polysorbate 80 formulation that can cause fluid retention in addition to hypersensitivity reactions,; dexamethasone premedication with or without antihistamines is frequently used. Cabazitaxel is a taxane with somewhat better activity in prostate cancers than earlier generations of taxanes, perhaps due to superior delivery to sites of disease.

Epothilones represent a class microtubule-stabilizing agents that have been conscientiously optimized for activity in taxane-resistant tumors. Ixabepilone has clear evidence of activity in breast cancers resistant to taxanes and anthracyclines such as doxorubicin. It retains acceptable expected side effects, including myelosuppression, and can also cause peripheral sensory neuropathy. Eribulin is a microtubule-directed agent with activity in patients who have had progression of disease on taxanes. It alters dynamics of microtubule re-modeling in cells.

Targeted Chemotherapy

HORMONE RECEPTOR–DIRECTED THERAPY

Steroid hormone receptor–related molecules have emerged as prominent targets for small molecules useful in cancer treatment. When bound to their cognate ligands, these receptors can alter gene transcription and, in certain tissues, induce apoptosis. The pharmacologic effect is a mirror or parody of the normal effects of the agents acting on nontransformed normal tissues. While in some cases, such as breast cancer, demonstration of the target hormone receptor is necessary, in other cases such prostate cancer (androgen receptor) and lymphoid neoplasms (glucocorticoid receptor), the relevant receptor is always present in the tumor.

Glucocorticoids are generally given in “pulsed” high doses in leukemias and lymphomas, where they induce cell death in tumor cells. Cushing’s syndrome and inadvertent adrenal suppression on withdrawal from high-dose glucocorticoids can be significant complications, along with infections common in immunosuppressed patients, in particular *Pneumocystis* pneumonia, which classically appears a few days after completing a course of high-dose glucocorticoids.

Tamoxifen is a partial estrogen receptor antagonist; it has a tenfold greater antitumor activity in breast cancer patients whose tumors express estrogen receptors than in those who have low or no levels of expression. It might be considered the prototypic “molecularly targeted” agent. Owing to its agonistic activities in vascular and uterine tissue, side effects include a somewhat increased risk of cardiovascular complications, such as thromboembolic phenomena, and a small increased incidence of endometrial carcinoma, which appears after chronic use (usually >5 years). Progestational agents—including medroxyprogesterone acetate, androgens including fluoxymesterone (Halotestin), and, paradoxically, estrogens—have approximately the same degree of activity in primary hormonal treatment of breast cancers that have elevated expression of estrogen receptor protein. Estrogen itself is not used often owing to prominent cardiovascular and uterotropic activity.

Aromatase refers to a family of enzymes that catalyze the formation of estrogen in various tissues, including the ovary and peripheral adipose tissue and some tumor cells. Aromatase inhibitors are of two types, the

irreversible steroid analogues such as exemestane and the reversible inhibitors such as anastrozole or letrozole. Anastrozole is superior to tamoxifen in the adjuvant treatment of breast cancer in postmenopausal patients with estrogen receptor–positive tumors. Letrozole treatment affords benefit following tamoxifen treatment. Adverse effects of aromatase inhibitors may include an increased risk of osteoporosis.

Metastatic prostate cancer is treated by androgen deprivation. Orchiectomy causes responses in 80% of patients. In the event that orchiectomy is not accepted by the patient, testicular androgen suppression can also be effected by luteinizing hormone–releasing hormone (LHRH) agonists such as leuprolide and goserelin. These agents cause tonic stimulation of the LHRH receptor, with the loss of its normal pulsatile activation resulting in decreased output of LH by the anterior pituitary. Therefore, as primary hormonal manipulation in prostate cancer, one can choose orchiectomy or leuprolide, but not both. The addition of androgen receptor blockers, including flutamide or bicalutamide, is of uncertain additional benefit in extending overall response duration, although pretreatment with these agents before LHRH agonists is important to avoid a surge in testosterone after initial LH release. Enzalutamide also binds to the androgen receptor and antagonizes androgen action in a mechanistically distinct way. Somewhat analogous to inhibitors of aromatase, agents have been derived that inhibit testosterone and other androgen synthesis in the testis, adrenal gland, and prostate tissue. Abiraterone inhibits 17 α-hydroxylase/C17,20 lyase (CYP 17A1) and has been shown to be active in prostate cancer patients experiencing progression despite androgen blockade.

Tumors that respond to a primary hormonal manipulation may frequently respond to second and third hormonal manipulations. Thus, breast tumors that had previously responded to tamoxifen have, on relapse, notable response rates to withdrawal of tamoxifen itself or to subsequent addition of an aromatase inhibitor or progestin. Likewise, initial treatment of prostate cancers with leuprolide plus flutamide may be followed after disease progression by response to withdrawal of flutamide. These responses may result from the removal of antagonists from mutant steroid hormone receptors that have come to depend on the presence of the antagonist as a growth-promoting influence.

DIAGNOSTICALLY GUIDED TARGETED THERAPY

The basis for discovery of drugs of this type was the prior knowledge of oncogene directed pathways driving tumor growth. Figure 69-4 summarizes how FDA-approved targeted agents act. In the case of diagnostically guided targeted chemotherapy, prior demonstration of a specific target is necessary to guide the rational use of the agent, while in the case of targeted agents directed at oncogenic pathways, specific diagnosis of pathway activation is not yet necessary or in some cases feasible, although this is an area of ongoing clinical research. Table 69-5 lists currently approved targeted chemotherapy agents, with features of their use.

TABLE 69-5

Molecularly Targeted Agents

DRUG	TARGET	ADVERSE EVENTS	NOTES
Diagnostically Guided Protein Kinase Antagonists			
Imatinib	Bcr-Abl fusion protein (CML/ALL); c-kit mutants, PDGFR variants (GI stromal tumor; eosinophilic syndromes)	Nausea Periorbital edema Rare CHF QTc prolongation	Myelosuppression not frequent in solid tumor indications
Nilotinib	Bcr-Abl fusion protein (CML) and some imatinib-resistant variants	Interaction with CYP3A4-metabolized drugs CHF Hepatotoxicity Hypothyroidism	Chronic phase and in patients resistant to imatinib
Dasatinib	Bcr-Abl fusion protein (CML/ALL); wild-type and imatinib-resistant mutants	Myelosuppression (bleeding, infection) Pulmonary hypertension CHF Fluid retention QTc prolongation	Chronic phase and imatinib or nilotinib resistant
Bosutinib	Bcr-Abl fusion protein (CML); wild-type and imatinib-resistant mutants	Myelosuppression Hepatic QTc prolongation	Chronic phase and imatinib or nilotinib resistant
Ponatinib	T315I mutation of Bcr-Abl fusion protein (CML)	Clotting Hepatic CHF Pancreatitis Neuropathy Rash	
Gefitinib	First-line treatment of NSCLC with ATP site mutation of EGFR	Diarrhea	In the United States, only with prior

		Interstitial pneumonitis	documented benefit in second-line treatment of NSCLC
Erlotinib	First-line treatment of NSCLC with ATP site mutation of EGFR; second-line treatment of wild-type EGFR NSCLC	Rash Diarrhea Rare interstitial pneumonitis	1 h before, 2 h after meals
Afatinib	First-line treatment of NSCLC with ATP site mutation of EGFR	Diarrhea Cutaneous	Interacts with Pgp inhibitors
Crizotinib	EML4-Alk fusion protein	Interstitial pneumonitis Hepatic QTc prolongation Bradycardia	
Vemurafenib	*BRAF* V600E in melanoma	Nausea Rash Cutaneous Second cutaneous neoplasms	
Dabrafenib	*BRAF* V600E in melanoma	Cutaneous Second cutaneous neoplasms	
Trametinib	*BRAF* V600E in melanoma (both as single agent and in combination with dabrafenib)	Rash Diarrhea Lymphedema	In combination with dabrafenib, second neoplasms, hemorrhage, venous thrombosis, CHF, ocular, hyperglycemia
DRUG	**INDICATION**	**ADVERSE EVENTS**	**NOTES**
Diagnostically Guided Retinoid			
Tretinoin	APL t(15,17)	Teratogenic	APL differentiation

		Cutaneous	syndrome: pulmonary dysfunction/infiltrate, pleural/pericardial effusion, fever
Multikinase Inhibitors			
Sorafenib	Renal cell, hepatocellular, differentiated thyroid carcinoma	Diarrhea Hand-foot syndrome Other rash Hypertension CHF	Targets c-raf, VEGFR
Pazopanib	Renal cell carcinoma, soft tissue sarcoma	Fatigue Diarrhea/GI Hypertension Thromboses QTc	Target VEGFR, c-kit, PDGFR
Regorafenib	Second-line colorectal cancer; GI stromal tumor	Hypertension Hand-foot syndrome Thromboses Perforations	VEGFR/TIE2
Sunitinib	Renal cell carcinoma, pancreatic neuroendocrine tumor, GI stromal tumor	Fatigue Diarrhea Neutropenia	Target VEGFR
Vandetanib	Medullary thyroid cancer	Diarrhea Rash Hypertension Prolonged QTc Thromboses	Target VEGFR, ret, EGFR
Cabozantinib	Medullary thyroid cancer	Hypertension Wound healing Fistulas Osteonecrosis	Target VEGFR, c-met

		Proteinuria	
Axitinib Osimertinib	Renal cell carcoinoma, second line Non-small cell lung cancer, EGFR T790M mutation	Diarrhea/other GI Fatigue Hand-foot syndrome Interstitial lung disease QTc prolongation Cardiomyopathy	Target VEGFR, PDGFR, c-kit
Proteasome Inhibitors			
Bortezomib	Multiple myeloma, mantle cell lymphoma	Neuropathy Thrombocytopenia GI	
Carfilzomib	Multiple myeloma, second line	Infusion reaction CHF Thrombocytopenia Pulmonary Tumor lysis	
Histone Deacetylase Inhibitors			
Vorinostat	Cutaneous T-cell lymphoma, second line	Fatigue Diarrhea Thrombocytopenia Embolism	
Romidepsin	Cutaneous T-cell lymphoma, second line	Nausea Vomiting Cytopenias Cardiac conduction	
mTOR Inhibitors			
Temsirolimus	Renal cell carcinoma, second line or poor prognosis	Stomatitis Thrombocytopenia Nausea	

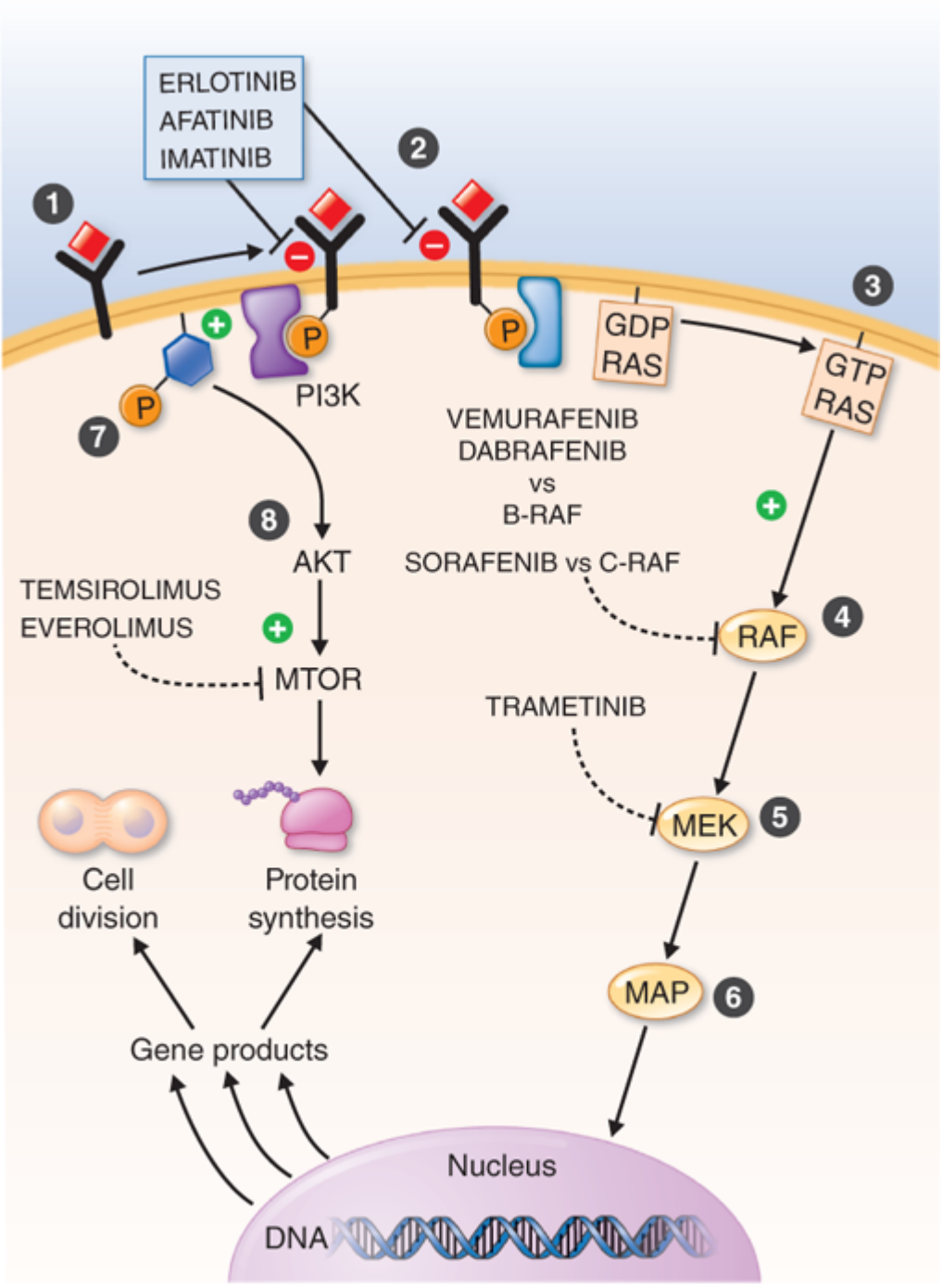
		Anorexia, fatigue Metabolic (glucose, lipid)	
Everolimus	Renal cell carcinoma, advanced; subependymal giant-cell astrocytoma; breast cancer, hormone receptor positive, resistant to antiestrogen; pancreatic neuroendocrine	Stomatitis Fatigue	
Miscellaneous			
Arsenic trioxide	APL	↑ QT_C	APL differentiation syndrome (see under tretinoin)
Vismodegib	Metastatic basal cell carcinoma	GI Hair loss Fatigue Muscle spasm Dysgeusia	Target smoothened receptor in hedgehog pathway

Abbreviations: APL, acute promyelocytic leukemia; ALL, acute lymphocytic leukemia; CHF, congestive heart failure; CML, chronic myeloid leukemia; EGFR, epidermal growth factor receptor; GI, gastrointestinal; mTOR, mammalian target of rapamycin kinase; NSCLC, non-small-cell lung cancer; PDGFR, platelet-derived growth factor receptor; Pgp, P-glycoprotein; VEGFR, vascular endothelial growth factor receptor.

FIGURE 69-4

Targeted chemotherapeutic agents act in most instances by interrupting cell growth factor-mediated signaling pathways. After a growth factor binds to cognate receptor (1), in many cases there is activation of tyrsosine kinase activity particularly after dimerization of the receptors (2). This leads to autophosphorylation of the receptor and docking of “adaptor” proteins. One important pathway activated occurs after exchange of GDP for GTP in the RAS family of protooncogene products (3). GTP-RAS activates the RAF proto-oncogene kinase (4), leading to a phosphorylation cascade of kinases (5, 6) that ultimately impart signals to regulators of gene function to produce transcripts which activate cell cycle progression and increase protein synthesis. In parallel, tyrosine phosphorylated receptors can activate the phosphatidylinositol-3-kinase to produce the phosphorylated lipid phosphatidyl-inositol-3-phosphate (7). This leads to the activation of the AKT kinase (8) which in turn stimulates the mammalian “Target of Rapamycin” kinase (mTOR), which directly increases the translation of key mRNAs for gene products regulating cell growth. Erlotinib and afatinib, are examples of Epidermal Growth Factor receptor tyrosine kinase inhibitors; imatinib can act on the nonreceptor tyrosine kinase bcr-abl or c-KIT membrane bound

tyrosine kinase. Vemurafenib and Dabrafenib act on the B isoform of RAF uniquely in melanoma, and c-RAF is inhibited by sorafenib. Trametinib acts on MEK. Temsirolimus and everolimus inhibit mTOR kinase to downregulate translation of oncogenic mRNAs.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In hematologic tumors, the prototypic agent of this type is imatinib, which targets the ATP binding site of the $p210^{bcr\text{-}abl}$ protein tyrosine kinase that is formed as the result of the chromosome 9;22 translocation producing the Philadelphia chromosome in CML. Imatinib is superior to interferon (IFN) plus chemotherapy in the initial treatment of the chronic phase of this disorder. It has lesser activity in the blast phase of CML, where the cells may have acquired additional mutations in $p210^{bcr\text{-}abl}$ itself or other genetic lesions. Its side

effects are relatively tolerable in most patients and include hepatic dysfunction, diarrhea, and fluid retention. Rarely, patients receiving imatinib have decreased cardiac function, which may persist after discontinuation of the drug. The quality of response to imatinib enters into the decision about when to refer patients with CML for consideration of transplant approaches. Nilotinib is a tyrosine protein kinase inhibitor with a similar spectrum of activity to imatinib, but with increased potency and perhaps better tolerance by certain patients. Dasatinib, another inhibitor of the $p210^{bcr\text{-}abl}$ oncoproteins, is active in certain mutant variants of $p210^{bcr\text{-}abl}$ that are refractory to imatinib and arise during therapy with imatinib or are present de novo. Dasatinib also has inhibitory action against kinases belonging to the src tyrosine protein kinase family; this activity may contribute to its effects in hematopoietic tumors and suggest a role in solid tumors where src kinases are active. The T315I mutant of $p210^{bcr\text{-}abl}$ is resistant to imatinib, nilotinib, bosutinib, and dasatinib; ponatinib has activity in patients with this $p210^{bcr\text{-}abl}$ variant, but ponatinib has noteworthy associated thromboembolic toxicity. Use of this class of targeted agents is thus critically guided not only by the presence of the $p210^{bcr\text{-}abl}$ tyrosine kinase, but also by the presence of different mutations in the ATP binding site.

All-*trans*-retinoic acid (ATRA) targets the PML-retinoic acid receptor (RAR) α fusion protein, which is the result of the chromosome 15;17 translocation pathogenic for most forms of APL. Administered orally, it causes differentiation of the neoplastic promyelocytes to mature granulocytes and attenuates the rate of hemorrhagic complications. Adverse effects include headache with or without pseudotumor cerebri and gastrointestinal and cutaneous toxicities.

In epithelial solid tumors, the small-molecule epidermal growth factor (EGF) antagonists act at the ATP binding site of the EGF receptor tyrosine kinase. In early clinical trials, gefitinib showed evidence of responses in a small fraction of patients with non-small-cell lung cancer (NSCLC). Side effects were generally acceptable, consisting mostly of acneiform rash (treated with glucocorticoid creams and clindamycin gel) and diarrhea. Subsequent analysis of responding patients revealed a high frequency of activating mutations in the EGF receptor. Patients with such activating mutations who initially responded to gefitinib but who then had progression of the disease then acquired additional mutations in the enzyme, analogous functionally to mutational variants responsible for imatinib resistance in CML. Erlotinib is another EGF receptor tyrosine kinase antagonist where the presence of EGF receptor tyrosine kinase mutations has recently been shown to be a basis for recommending erlotinib and afatinib for first-line treatment of advanced NSCLC. Osimertinib is uniquely active in lung cancers with the T790M mutation. Likewise, crizotinib targeting the *alk* protooncogene fusion protein has value in the initial treatment of *alk*-positive NSCLC. Lapatinib is a tyrosine kinase inhibitor with both EGF receptor and HER2/neu antagonist activity, which is important in the treatment of breast cancers expressing the HER2/neu oncoprotein.

In addition to the $p210^{bcr\text{-}abl}$ kinase, imatinib also has activity against the c-kit tyrosine kinase (the receptor for the *steel* growth factor, also called stem cell factor) and the platelet-derived growth factor receptor (PDGFR), both of which can be expressed in gastrointestinal stromal sarcoma (GIST). Imatinib has found clinical utility in GIST, a tumor previously notable for its refractoriness to chemotherapeutic approaches. Imatinib's degree of activity varies with the specific mutational variant of kit or PDGFR present in a particular patient's tumor.

The *BRAF* V600E mutation has been detected in a notable fraction of melanomas, thyroid tumors, and hairy cell leukemia, and preclinical models supported the concept that *BRAF* V600E drives oncogenic signaling in these tumors. Vemurafenib and dabrafenib, with selective capacity to inhibit the *BRAF* V600E serine kinase activity, were each shown to cause noteworthy responses in patients with *BRAF* V600E–mutated melanomas, although early relapse occurred in many patients treated with the drugs as single agents. Trametinib, acting downstream of *BRAF* V600E by directly inhibiting the MEK serine kinase by a non-ATP binding site mechanism, also displayed noteworthy responses in *BRAF* V600E–mutated melanomas, and the combination of trametinib and dabrafenib is even more active, by targeting the *BRAF* V600E–driven pathway at two points in the pathway leading to gene activation.

ONCOGENICALLY ACTIVATED PATHWAYS

Agents in this class also target specific regulatory molecules in promoting the viability of tumor cells, but they do not require the diagnostically verified presence of a particular target or target variant at this time.

"Multitargeted" kinase antagonists are small-molecule ATP site-directed antagonists that inhibit more than one protein kinase and have value in the treatment of several solid tumors. Drugs of this type with prominent activity against the vascular endothelial growth factor receptor (VEGFR) tyrosine kinase have activity in renal cell carcinoma. Sorafenib is a VEGFR antagonist with activity against the *raf* serine-threonine protein kinase, and regorafenib is a closely related drug with value in relapsed advanced colon cancer. Pazopanib also prominently targets VEGFR and has activity in renal carcinoma and soft tissue sarcomas. Sunitinib has anti-VEGFR, anti-PDGFR, and anti-c-kit activity. It causes prominent responses and stabilization of disease in renal cell cancers and GISTs. Side effects for agents with anti-VEGFR activity prominently include hypertension, proteinuria, and, more rarely, bleeding and clotting disorders and perforation of scarred gastrointestinal lesions. Also encountered are fatigue, diarrhea, and the hand-foot syndrome, with erythema and desquamation of the distal extremities, in some cases requiring dose modification, particularly with sorafenib.

Temsirolimus and everolimus are mammalian target of rapamycin (mTOR) inhibitors with activity in renal cancers. They produce stomatitis, fatigue, and some hyperlipidemia (10%), myelosuppression (10%), and rare lung toxicity. Everolimus is also useful in patients with hormone receptor–positive breast cancers displaying resistance to hormonal inhibition and in certain neuroendocrine and brain tumors, the latter arising in patients with sporadic or inherited mutations in the pathway activating mTOR. Cyclin dependent kinases (CDKs) are activated as the result of oncogene pathway activity. Palbociclib, a selective inhibititor of CDKs 4 and 6, has noteworthy activity in conjunction with the mTOR inhibitors in advanced breast cancers also expressing the estrogen receptor.

In hematologic neoplasms, bortezomib is an inhibitor of the proteasome, the multisubunit assembly of protease activities responsible for the selective degradation of proteins important in regulating activation of transcription factors, including nuclear factor-κB (NF-κB) and proteins regulating cell cycle progression. It has activity in multiple myeloma and certain lymphomas. Adverse effects include neuropathy, orthostatic hypotension with or without hyponatremia, and reversible thrombocytopenia. Carfilzomib is a proteasome inhibitor chemically unrelated to bortezomib without prominent neuropathy, but with evidence of a cytokine

release syndrome, which can be a cardiopulmonary stress. Other agents active in multiple myeloma and certain other hematologic neoplasms include the immunomodulatory agents related to thalidomide, including lenalidomide and pomalidomide. All these agents collectively inhibit aberrant angiogenesis in the bone marrow microenvironment, as well as influence stromal cell immune functions to alter the cytokine milieu supporting the growth of myeloma cells. Thalidomide, although clinically active, has prominent cytopenic, neuropathic, procoagulant, and CNS toxicities that have been somewhat attenuated in the other drugs of the class, although use of these agents frequently entails concomitant anticoagulant prophylaxis.

Ibrutinib and idelalisib are representative of novel classes of inhibitors directed at Bruton's tyrosine kinase and phosphatidyl inositide-3 kinase-δ, respectively, expressed in normal and neoplastic B cells. Initially approved for use in mantle cell lymphoma and chronic lymphocytic leukemia, respectively, they are potentially applicable to a number of B-cell neoplasms that depend on signals through the B-cell antigen receptor. Janus kinases likewise function downstream of a variety of cytokine receptors to amplify cytokine signals, and Janus kinase inhibitors including ruxolitinib have approved activity in myelofibrosis to ameliorate splenomegaly and systemic symptoms.

Vorinostat is an inhibitor of histone deacetylases, which are responsible for maintaining the proper orientation of histones on DNA, with resulting capacity for transcriptional readiness. Acetylated histones allow access of transcription factors to target genes and therefore increase expression of genes that are selectively repressed in tumors. The result can be differentiation with the emergence of a more normal cellular phenotype, or cell cycle arrest with expression of endogenous regulators of cell cycle progression. Vorinostat is approved for clinical use in cutaneous T-cell lymphoma, with dramatic skin clearing and very few side effects. Romidepsin is a distinct molecular class of histone deacetylase inhibitor also active in cutaneous T-cell lymphoma. Panobinostat has activity in multiple myeloma. DNA methyltransferase inhibitors, including 5-aza-cytidine and 2′-deoxy-5-azacytidine (decitabine), can also increase transcription of genes "silenced" during the pathogenesis of a tumor by causing demethylation of the methylated cytosines that are acquired as an "epigenetic" (i.e., after the DNA is replicated) modification of DNA. These drugs were originally considered antimetabolites but have clinical value in myelodysplastic syndromes and certain leukemias when administered at low doses.

Additional toxicities with several therapies affecting ongogene-activated pathways include poorly predicted hepatic and cardiac toxicities (imatinib, dasatinib, sorafenib, pazopanib) or cardiac conduction deficits including prolonged QT interval (pazopanib), and atrial fibrillation (ibrutinib). The occurrence of new cardiac or liver abnormalities in a patient receiving treatment with a protein kinase antagonist should lead to a consideration of the risk versus benefit and the possible relation of the agent to the new adverse event. The existence of prior cardiac dysfunction is a relative contraindication to the use of certain targeted therapies (e.g., trastuzumab), although each patient's needs should be individualized.

CANCER BIOLOGIC THERAPY

Principles

The goal of biologic therapy is to manipulate the host–tumor interaction in favor of the host, potentially at an optimum biologic dose that might be different than the MTD. As a class, biologic therapies may be distinguished from molecularly targeted agents in that many biologic therapies require an active response (e.g., reexpression of silenced genes or antigen expression) on the part of the tumor cell or on the part of the host (e.g., immunologic effects) to allow therapeutic effect. This may be contrasted with the more narrowly defined antiproliferative or apoptotic response that is the ultimate goal of molecularly targeted agents discussed above. However, there is much commonality in the strategies to evaluate and use molecularly targeted and biologic therapies.

Antibody-Mediated Therapeutic Approaches

In general, antibodies are not very effective at killing cancer cells. Because the tumor seems to influence the host toward making antibodies rather than generating cellular immunity, it is inferred that antibodies are easier for the tumor to fend off. Many patients can be shown to have serum antibodies directed at their tumors, but these do not appear to influence disease progression. However, the ability to grow very large quantities of high-affinity antibody directed at a tumor has led to the application of antibodies in the treatment of cancer. In this approach, antibodies are derived where the antigen-combining regions are grafted onto human immunoglobulin gene products (chimerized or humanized) or derived de novo from mice bearing human immunoglobulin gene loci. Three general strategies have emerged using antibodies. *Tumor-regulatory antibodies* target tumor cells directly or indirectly to modulate intracellular functions or attract immune or stromal cells. *Immunoregulatory antibodies* target antigens expressed on the tumor cells or host immune cells to modulate primarily the host's immune responsiveness to the tumor. Finally, *antibody conjugates* can be made with the antibody linked to drugs, toxins, or radioisotopes to target these "warheads" for delivery to the tumor. Table 69-6 lists features of currently used or promising antibodies for cancer treatment.

TABLE 69-6

Antibodies Used in Cancer Treatment

Drug	Target	Indications and Features of Use
Tumor Regulatory Antibodies		
Rituximab	CD20	B cell neoplasms (also emerging role in autoimmune disease); chimeric antibody
		with frequent mouse-derived sequences; frequent infusion reactions, particularly on initial doses; reactivation of infections, particularly hepatitis; progressive multifocal leukoencephalopathy; tumor lysis syndrome
Ofatumumab	CD20	active in CLL; fully human antibody with distinct binding site compared to rituximab; decreased intensity infusion reactions
Trastuzumab	HER2/neu	Active in breast cancer and GI cancers expressing HER2/neu; cardiotoxicity, particularly in setting of prior anthracyclines, requires monitoring; infusion reactions
Pertuzumab	HER2/neu	Breast cancer; targets distinct binding site from trastuzumab, inhibiting dimerization of HER2 family members; infusion reactions; cardiac toxicity
Cetuximab	EGFR	Colorectal cancers with wild-type Ki-ras oncoprotein; head and neck cancers with radiation; rash, diarrhea, infusion reactions
Panitumumab	EGFR	Colorectal cancers with wild-type Ki-ras oncoprotein; fully humanized; decreased infusion reactions; different IgG subtype than cetuximab
Bevacizumab Daratumumab Elotuzumab Olaratumab	VEGF CD38 CD319 PDGF-R	Metastatic colorectal cancer and non-small-cell lung cancer (nonsquamous) with chemotherapy; renal cancer and glioblastoma as single agents; prominent HBP, proteinuria, GI perforations, hemorrhage, thrombosis (venous and arterial) Multiple myeloma Multiple myeloma, with revlimid and dexamethasone Soft tissue sarcoma, in conjunction with doxorubicin
Immunoregulatory Antibodies		

Drug	Target	Indications and Features of Use
Alemtuzumab	CD52	CLL, T-cell lymphomas; activates complement after binding to cell surface; infusion reactions, hypersensitivity, tumor lysis, activation of infections, cytopenias
Ipilimumab	CTLA4	Melanoma; inhibits the negative proliferative signal to T cells acting through CTLA4, resulting in prominent T cell activation; side effects include immune-mediated toxicity to liver, skin, pituitary, gut, which if severe calls for steroids, which inhibit antineoplastic effect
Pembrolizumab	PD-1	Non-small cell lung cancer as a first- or second-line treatment if PDL1(+) and no actionable mutations;and as a second-line treatment for head and neck squamous cell carcinoma, after platinum-based chemotherapy; can cause immune-related colitis, hepatitis, hypophysitis, nephritis, and altered thyroid function; also consider steroids for treatment of severe adverse events
Nivolumab	PD-1	Metastatic melanoma in combination with ipilimumab if B-RAF mutation negative; melanoma following treatment with ipilimumab and after a BRAF inhibitor if relevant; second-line treatment for squamous non-small cell lung cancer, renal cancer and for relapsed and refractory Hodgkin's Disease; side effects similar to pembrolizumab
Atezolizumab	PD-L1	Locally advanced or metastatic urothelial carcinoma treatment after failure of chemo- or radiotherapy; metastatic non-small cell lung cancer (NSCLC) whose disease progressed during or following platinum-containing chemotherapy, without actionable mutations

Abbreviations: CLL, chronic lymphocytic leukemia; EGFR, epidermal growth factor receptor; GI, gastrointestinal; HBP, high blood pressure; VEGF, vascular endothelial growth factor.

TUMOR-REGULATORY ANTIBODIES

Humanized antibodies against the CD20 molecule expressed on B-cell lymphomas (rituximab and ofatumumab) are exemplary of antibodies that affect both signaling events driving lymphomagenesis as well as activating immune responses against B-cell neoplasms. They are used as single agents and in combination with chemotherapy and radiation in the treatment of B-cell neoplasms. Obinutuzumab is an antibody with an altered glycosylation that enhances its ability to activate killer cells; it is also directed against CD20 and is of value in chronic lymphocytic leukemia. It seems to be more effective in this setting than rituximab.

The HER2/neu receptor overexpressed on epithelial cancers, especially breast cancer, was initially targeted by trastuzumab, with noteworthy activity in potentiating the action of chemotherapy in breast cancer as well as some evidence of single-agent activity. Trastuzumab also appears to interrupt intracellular signals derived from HER2/neu and to stimulate immune mechanisms. The anti-HER2 antibody pertuzumab, specifically targeting the domain of HER2/neu responsible for dimerization with other HER2 family members, is more specifically directed against HER2 signaling function and augments the action of trastuzumab.

EGF receptor (EGFR)-directed antibodies (such as cetuximab and panitumumab) have activity in colorectal cancer refractory to chemotherapy, particularly when used to augment the activity of an additional chemotherapy program, and in the primary treatment of head and neck cancers treated with radiation therapy. The mechanism of action is unclear. Direct effects on the tumor may mediate an antiproliferative effect as well as stimulate the participation of host mechanisms involving immune cell or complement-mediated response to tumor cell–bound antibody. Alternatively, the antibody may alter the release of paracrine factors promoting tumor cell survival.

The anti-VEGF antibody bevacizumab shows little evidence of antitumor effect when used alone, but when combined with chemotherapeutic agents, it improves tumor shrinkage and time to disease progression in colorectal and nonsquamous lung cancers. The mechanism for the effect is unclear and may relate to the capacity of the antibody to alter delivery and tumor uptake of the active chemotherapeutic agent. Ziv-aflibercept is not an antibody, but a solubilized VEGF receptor VEGF binding domain, and therefore may have a distinct mechanism of action with comparable side effects.

Unintended side effects of any antibody use include infusion-related hypersensitivity reactions, usually limited to the first infusion, which can be managed with glucocorticoid and/or antihistamine prophylaxis. In addition, distinct syndromes have emerged with different antibodies. Anti-EGFR antibodies produce an acneiform rash that poorly responds to glucocorticoid cream treatment. Trastuzumab (anti-HER2) can inhibit cardiac function, particularly in patients with prior exposure to anthracyclines. Bevacizumab has a number of side effects of medical significance, including hypertension, thrombosis, proteinuria, hemorrhage, and gastrointestinal perforations with or without prior surgeries; these adverse events also occur with small-molecule drugs modulating VEGFR function.

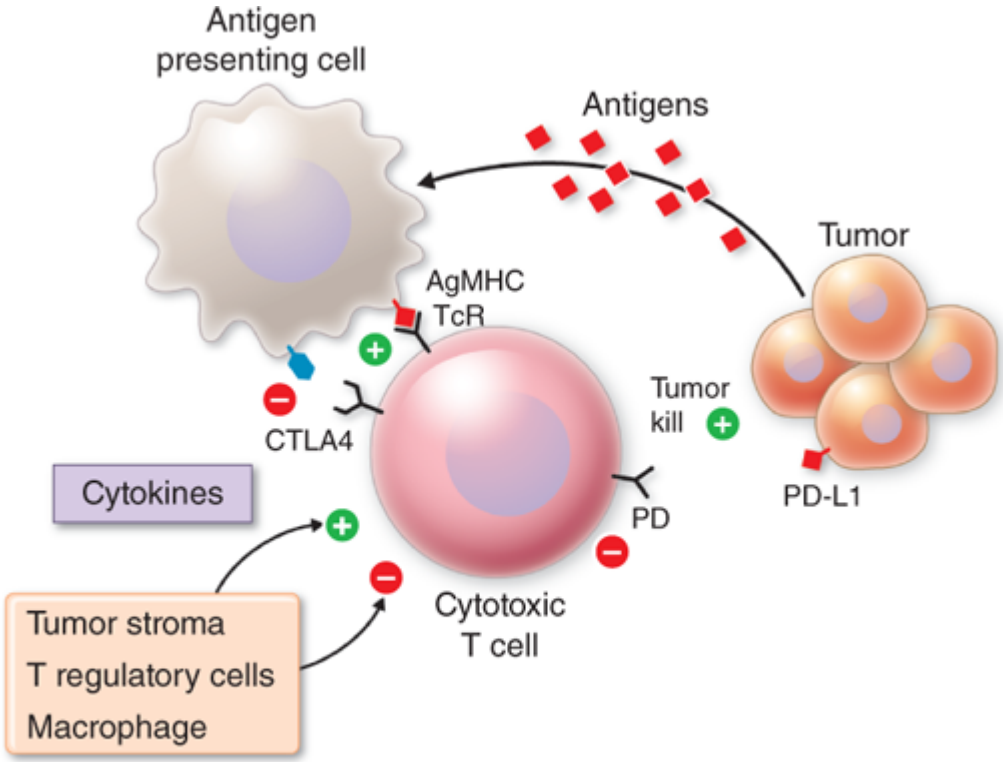
IMMUNOREGULATORY ANTIBODIES

Purely immunoregulatory antibodies stimulate immune responses to mediate tumor-directed cytotoxicity. First-generation approaches sought to activate complement and are exemplified by alemtuzumab against CD52; these are active in chronic lymphoid leukemia and T-cell malignancies. A more refined understanding of the tumor–host interface has defined that cytotoxic tumor-directed T cells are frequently inhibited by ligands upregulated in the tumor cells. The programmed death ligand 1 (PD-L1; also known as B7-homolog 1) was initially recognized as an entity that induced T cell death through a receptor present on T cells, termed the PD receptor **(Fig. 69-5)**, which physiologically exists to regulate the intensity of the immune response. The PD family of ligands and receptors also regulates macrophage function, present in tumor stroma. These actions raised the hypothesis that antibodies directed against the PD signaling axis (both anti-PD-L1 and anti-PD) might be useful in cancer treatment by allowing reactivation of the immune response against

tumors. Nivolumab, directed against the PD-1 receptor, is approved for use in renal cancer, metastatic melanoma, and non-small cell lung cancer, as well as in relapsed Hodgkin's disease. Pembrolizumab is approved for first-line treatment of metastatic non-small cell lung cancer whose tumors express the PD-L1 ligand. This development was a milestone in cancer therapeutics, replacing chemotherapy in this patient subset.

FIGURE 69-5

Tumors possess a microenvironment (tumor stroma) with immune cells including both helper T cells, suppressor T cells (both "regulatory" of other immune cell function), macrophages, and cytotoxic T cells. Cytokines found in the stroma and deriving from macrophages and regulatory T cells modulate the activities of cytotoxic T cells, which have the potential to kill tumor cells. Antigens released by tumor cells are taken up by Antigen Presenting Cells (APCs), also in the stroma. Antigens are processed by the APCs to peptides presented by the Major Histocompatibility Complex to T-cell antigen receptors, thus providing an (+) activation signal for the cytotoxic tumor cells to kill tumor cells bearing that antigen. Negative (–) signals inhibiting cytoxic T cell action include the CTLA4 receptor (on T cells), interacting with the B7 family of negative regulatory signals from APCs, and the PD receptor (on T cells), interacting with the PD-L1 (–) signal coming from tumor cells expressing the PD-1 ligand (PD-l1). As both CTLA4 and PD1 signals attenuate the anti-tumor T cell response, strategies which inhibit CTLA4 and PD1 function are a means of stimulating cytotoxic T cell activity to kill tumor cells. Cytokines from other immune cells and macrophages can provide both (+) and (–) signals for T cell action, and are under investigation as novel immunoregulatory therapeutics.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ipilimumab, an antibody directed against the anti-CTLA4 (cytotoxic T lymphocyte antigen 4), which is expressed on T cells (not tumor cells), responds to signals from antigen-presenting cells (Fig. 69-5), and also downregulates the intensity of the T-cell proliferative response to antigens derived from tumor cells. Indeed, manipulation of the CTLA4 axis was the first demonstration that purely immunoregulatory antibody strategies directed at T-cell physiology could be safe and effective in the treatment of cancer, although it acts at a very early stage in T-cell activation and can be considered somewhat nonspecific in its basis for T-cell stimulation. Ipilimumab alone or in combination with PD1-directed antibodies, is approved for initial treatment of metastatic melanoma.

Prominent activation of autoimmune hepatic, endocrine, cutaneous, neurologic, and gastrointestinal responses is a basis for adverse events with the use of ipilimumab and the PD-1-directed antibodies; the emergent use of glucocorticoids may be required to attenuate severe toxicities, which unfortunately can theoretically attenuate antitumor effect. Importantly for the general internist, these events may occur late after exposure to ipilimumab while the patient may otherwise be enjoying sustained control of tumor growth owing to the beneficial actions of ipilimumab.

Another class of immunoregulatory antibody is the “bispecific” antibody blinatumomab, which was constructed to have an anti-CD19 antigen combining site as one valency of an antibody with anti-CD3 binding site as the other valency. This antibody thus can bring T cells (with its anti-CD3 activity) close to B cells bearing the CD19 determinant. Blinatumomab is active in B-cell neoplasms such as acute lymphocytic leukemia, which may not have prominent expression of the CD20 targeted by rituximab.

ANTIBODY CONJUGATES

Conjugates of antibodies with drugs and isotopes have also been shown to be effective in the treatment of cancer and have the intent of increasing the therapeutic index of the drug or isotope by delivering the toxic “warhead” directly to the tumor cell or tumor microenvironment. Ado-trastuzumab is a conjugate of the HER2/neu-directed trastuzumab and a highly toxic microtubule targeted drug (emtansine), which by itself is too toxic for human use; the antibody-drug conjugate shows valuable activity in patients with breast cancer who have developed resistance to the “naked” antibody. Brentuximab vedotin is an anti-CD30 antibody drug conjugate with a distinct microtubule poison with activity in neoplasms such as Hodgkin’s lymphoma where the tumor cells frequently express CD30. Radioconjugates targeting CD20 on lymphomas have been approved for use (ibritumomab tiuxetan [Zevalin], using yttrium-90 or ^{131}I-tositumomab). Toxicity concerns have limited their use.

Cytokines

Only IFN-α and interleukin 2 (IL-2)-related molecules are in routine clinical use. The two recombinant interferons commercially available are IFN-α2a and α2b. IFN is not curative for any tumor but can induce partial responses in follicular lymphoma, hairy cell leukemia, CML, melanoma, and Kaposi’s sarcoma. It has been used in the adjuvant setting in stage II melanoma, multiple myeloma, and follicular lymphoma. It produces fever, fatigue, a flulike syndrome, malaise, myelosuppression, and depression and can induce clinically significant autoimmune disease.

IL-2 exerts its antitumor effects indirectly through augmentation of immune function. Its biologic activity is to promote the growth and activity of T cells and natural killer (NK) cells. High doses of IL-2 can produce tumor regression in certain patients with metastatic melanoma and renal cell cancer. About 2–5% of patients may experience complete remissions that are durable, unlike any other treatment for these tumors. IL-2 is associated with intravascular volume depletion, capillary leak syndrome, adult respiratory distress syndrome, hypotension, fever, chills, skin rash, and impaired renal and liver function. Patients may require blood pressure support and intensive care to manage the toxicity. However, once the agent is stopped, most of the toxicities reverse completely within 3–6 days. IL2 has been fused to translate in frame with a fragment of diphtheria toxin. A commercially available construct has activity against certain T-cell lymphomas. The drug's utility derives from the internalization of the targeted receptor and cleavage of the active drug or toxin moiety.

Immune Cell–Mediated Therapies

Tumors have a variety of means of avoiding the immune system: (1) they are often only subtly different from their normal counterparts; (2) they are capable of downregulating their major histocompatibility complex antigens, effectively masking them from recognition by T cells; (3) they are inefficient at presenting antigens to the immune system; (4) they can cloak themselves in a protective shell of fibrin to minimize contact with surveillance mechanisms; and (5) they can produce a range of soluble molecules, including potential immune targets, that can distract the immune system from recognizing the tumor cell or can kill or inactivate the immune effector cells. Prominent mediators of this effect are the PD receptors and their ligands described above. Some of the cell products initially polarize the immune response away from cellular immunity (shifting from T_H1 to T_H2 responses; Chap. 342) and ultimately lead to defects in T cells that prevent their activation and cytotoxic activity. A variety of strategies are being tested to overcome these barriers.

Cell-Mediated Immunity

The strongest evidence that the immune system can exert clinically meaningful antitumor effects comes from allogeneic bone marrow transplantation. Adoptively transferred T cells from the donor expand in the tumor-bearing host, recognize the tumor as being foreign, and can mediate impressive antitumor effects (graft-versus-tumor effects). Three types of experimental interventions are being developed to take advantage of the ability of T cells to kill tumor cells.

1. *Transfer of allogeneic T cells.* This occurs in three major settings: in allogeneic bone marrow transplantation; as purified lymphocyte transfusions following bone marrow recovery after allogeneic bone marrow transplantation; and as pure lymphocyte transfusions following immunosuppressive (nonmyeloablative) therapy (also called reduced intensity or minitransplants). In each of these settings, the effector cells are donor T cells that recognize the tumor as being foreign, probably through minor histocompatibility differences. The main risk of such therapy is the development of graft-versus-host

disease because of the minimal difference between the cancer and the normal host cells. This approach has been highly effective in certain hematologic cancers.

2. *Transfer of autologous T cells.* In this approach, the patient's own T cells are removed from the tumor-bearing host, manipulated in several ways in vitro, and given back to the patient. There are three major classes of autologous T-cell manipulation. First, tumor antigen–specific T cells can be developed and expanded to large numbers over many weeks ex vivo before administration. Second, the patient's T cells can be activated by exposure to polyclonal stimulators such as anti-CD3 and anti-CD28 after a short period ex vivo, and then amplified in the host after transfer by stimulation with IL-2, for example. Short periods removed from the patient permit the cells to overcome the tumor-induced T-cell defects, and such cells traffic and home to sites of disease better than cells that have been in culture for many weeks. In a third approach, genes that encode for a T-cell receptor specific for an antigen expressed by the tumor along with genes that facilitate T-cell activation can be introduced into subsets of a patient's T cells, which, after transfer back into the patient, allow homing of cytotoxic T cells to tumor cells expressing the antigen.

3. *Tumor vaccines aimed at boosting T-cell immunity.* The finding that mutant oncogenes that are expressed only intracellularly can be recognized as targets of T cell killing greatly expanded the possibilities for tumor vaccine development. However, major difficulties remain in getting the tumor-specific peptides presented in a fashion to prime the T cells. Tumors themselves are very poor at presenting their own antigens to T cells at the first antigen exposure (*priming*). Priming is best accomplished by professional antigen-presenting cells (dendritic cells). Thus, a number of experimental strategies are aimed at priming host T cells against tumor-associated peptides. Vaccine adjuvants such as granulocyte-macrophage colony-stimulating factor (GM-CSF) appear capable of attracting antigen-presenting cells to a skin site containing a tumor antigen. Purified antigen-presenting cells can be pulsed with tumor, its membranes, or particular tumor antigens and delivered as a vaccine. One such vaccine, Sipuleucel-T, is approved for use in patients with hormone-independent prostate cancer. In this approach, the patient undergoes leukapheresis, wherein mononuclear cells (that include antigen-presenting cells) are removed from the patient's blood. The cells are pulsed in a laboratory with an antigenic fusion protein comprising a protein frequently expressed by prostate cancer cells, prostate acid phosphatase, fused to GM-CSF, and matured to increase their capacity to present the antigen to immune effector cells. The cells are then returned to the patient in a well-tolerated treatment. Although no objective tumor response was documented in clinical trials, median survival was increased by about 4 months. Tumor cells can also be transfected with genes that attract antigen-presenting cells.

Another important vaccine strategy is directed at infectious agents whose action ultimately is tied to the development of human cancer. Hepatitis B vaccine in an epidemiologic sense prevents hepatocellular carcinoma, and a tetravalent human papillomavirus vaccine prevents infection by virus types currently accounting for 70% of cervical cancer. Unfortunately, these vaccines are ineffective at treating patients who have developed a virus-induced cancer.

SYSTEMIC RADIATION THERAPY

Although total-body irradiation has a role in preparing a patient to received allogeneic stem cells, and antibodies as described above can specifically target radioisotopes, systemically administered isotopes of iodide salts have an important role in the treatment of thyroid neoplasms, owing to the selective upregulation of the iodide transporter in the tumor cell compartment. Likewise, isotopes of samarium and radium have been found useful in the palliation of symptoms from advanced bony metastases of prostate cancer owing to their selective deposition at the tumor–bone matrix interface, thereby potentially affecting the function of both tumor and stromal cells in the progressive growth of the metastatic deposit.

RESISTANCE TO CANCER TREATMENTS

Resistance mechanisms to the conventional cytotoxic agents were initially characterized in the late twentieth century as defects in drug uptake, metabolism, or export by tumor cells. The *multidrug resistance* (*mdr*) gene defined in vitro in cell lines exposed to increasing concentrations of drugs led to the definition of a family of transport proteins that efficiently excrete the drug from the tumor cells; no clinically useful modulator of this process has yet emerged. Drug-metabolizing enzymes such as cytidine deaminase are upregulated in resistant tumor cells, and this is the basis for so-called “high-dose cytarabine” regimens in the treatment of leukemia. Another resistance mechanism defined during this era involved increased expression of a drug’s target, exemplified by amplification of the dihydrofolate reductase gene, in patients who had lost responsiveness to methotrexate, or mutation of topoisomerase II in tumors that relapsed after topoisomerase II modulator treatment.

A second class of resistance mechanisms involves loss of the cellular apoptotic mechanism activated after the engagement of a drug’s target by the drug. This occurs in a way that is heavily influenced by the biology of the particular tumor type. For example, decreased alkylguanine alkyltransferase expression defines a subset of glioblastoma patients with the prospect of enhanced benefit from treatment with temozolomide, but has no predictive value for benefit from temozolomide in epithelial neoplasms. Likewise, ovarian cancers resistant to platinating agents have decreased expression of the proapoptotic gene *bax*. These types of findings have prompted the idea that responsive tumors to chemotherapeutic agents are populated by cells that express drug-related cell death controlling genes, creating in effect a state of “synthetic lethality” with the drug **(Chap. 68)**. When drug is not present, absence or mutation in these genes is tolerated, but become lethal in the presence of the drug.

A third class of resistance mechanisms emerged from sequencing of the targets of agents directed at oncogenic kinases. Thus, patients with CML resistant to imatinib have acquired mutations in the ATP binding domain of $p210^{bcr\text{-}abl}$ in some cases, leading to the screening and design of agents with activity against the mutant proteins. Entirely analogous resistance mechanisms have emerged in patients with lung cancer treated with the EGFR antagonists gefitinib and erlotinib.

A final category of tumor resistance mechanisms to targeted agents includes the upregulation of alternate means of activating the pathway targeted by the agent. Thus melanomas initially responsive to *BRAF*V600E antagonists such as vemurafenib may reactivate raf signaling by upregulating isoforms that can bypass the

variant blocked by the drug. Likewise, inhibition of HER2/neu signaling in breast cancer cells can lead to the emergence of variants with distinct oncogenic signaling pathways such as PI3 kinase. The susceptibility of a tumor to different treatments as a function of its expression of potential drug targets or their mutational profile has led to efforts to define the dominant pathways driving a patient's tumor by genomic techniques including whole exome sequencing. The difficulty with applying such data to patient treatment is recognizing that these pathways may change during the natural history of a tumor and that different sites in a single patient may have tumors with different patterns of gene mutation.

SUPPORTIVE CARE DURING CANCER TREATMENT

MYELOSUPPRESSION

The common cytotoxic chemotherapeutic agents almost invariably affect bone marrow function. Titration of this effect determines the tolerated dose of the agent on a given schedule. The normal kinetics of blood cell turnover influences the sequence and sensitivity of each of the formed elements. Polymorphonuclear leukocytes (PMNs; $t_{1/2}$ = 6–8 h), platelets ($t_{1/2}$ = 5–7 days), and red blood cells (RBCs; $t_{1/2}$ = 120 days) have most, less, and least susceptibility, respectively, to usually administered cytotoxic agents. The nadir count of each cell type in response to classes of agents is characteristic. Maximal neutropenia occurs 6–14 days after conventional doses of anthracyclines, antifolates, and antimetabolites. Alkylating agents differ from each other in the timing of cytopenias. Nitrosoureas, DTIC, and procarbazine can display delayed marrow toxicity, first appearing 6 weeks after dosing.

Complications of myelosuppression result from the predictable sequelae of the missing cells' function. *Febrile neutropenia* refers to the clinical presentation of fever (one temperature ≥38.5°C or three readings ≥38°C but ≤38.5°C per 24 h) in a neutropenic patient with an uncontrolled neoplasm involving the bone marrow or, more usually, in a patient undergoing treatment with cytotoxic agents. Mortality from uncontrolled infection varies inversely with the neutrophil count. If the nadir neutrophil count is >1000/μL, there is little risk; if <500/μL, risk of death is markedly increased. Management of febrile neutropenia has conventionally included empirical coverage with antibiotics for the duration of neutropenia **(Chap. 70)**. Selection of antibiotics is governed by the expected association of infections with certain underlying neoplasms; careful physical examination (with scrutiny of catheter sites, dentition, mucosal surfaces, and perirectal and genital orifices by gentle palpation); chest x-ray; and Gram stain and culture of blood, urine, and sputum (if any) to define a putative site of infection. In the absence of any originating site, a broadly acting β-lactam with anti-*Pseudomonas* activity, such as ceftazidime, is begun empirically. The addition of vancomycin to cover potential cutaneous sites of origin (until these are ruled out or shown to originate from methicillin-sensitive organisms) or metronidazole or imipenem for abdominal or other sites favoring anaerobes reflects modifications tailored to individual patient presentations. Febrile neutropenic patients can be stratified broadly into two prognostic groups. The first, with expected short duration of neutropenia and no evidence of hypotension or abdominal or other localizing symptoms, may be expected to do well even with oral regimens, e.g., ciprofloxacin or moxifloxacin, or amoxicillin plus clavulanic acid. A less favorable prognostic group is patients with expected prolonged neutropenia, evidence of sepsis, and end

organ compromise, particularly pneumonia. Empirical addition of antifungal agents if fever and neutropenia persist for 7 days without identification of an adequately treated organism or site is frequent.

Transfusion of granulocytes has no role in the management of febrile neutropenia, owing to their exceedingly short half-life, mechanical fragility, and clinical syndromes of pulmonary compromise with leukostasis after their use. Instead, colony-stimulating factors (CSFs) are used to augment bone marrow production of PMNs. The American Society of Clinical Oncology has developed practice guidelines for the use of G-CSF and GM-CSF **(Table 69-7)**.

TABLE 69-7

Indications for the Clinical Use of G-CSF or GM-CSF

Preventive Uses
With the first cycle of chemotherapy (so-called *primary CSF administration*)
Not needed on a routine basis
Use if the probability of febrile neutropenia is ≥20%
Use if patient has preexisting neutropenia or active infection
Age >65 years treated for lymphoma with curative intent or other tumors treated by similar regimens
Poor performance status
Extensive prior chemotherapy
Dose-dense regimens in a clinical trial or with strong evidence of benefit
With subsequent cycles if febrile neutropenia has previously occurred (so-called *secondary CSF administration*)
Not needed after short-duration neutropenia without fever
Use if patient had febrile neutropenia in previous cycle
Use if prolonged neutropenia (even without fever) delays therapy
Therapeutic Uses
Afebrile neutropenic patients
No evidence of benefit
Febrile neutropenic patients
No evidence of benefit
May feel compelled to use in the face of clinical deterioration from sepsis, pneumonia, or fungal infection, but benefit unclear

In bone marrow or peripheral blood stem cell transplantation
Use to mobilize stem cells from marrow
Use to hasten myeloid recovery
In acute myeloid leukemia
G-CSF of minor or no benefit
GM-CSF of no benefit and may be harmful
In myelodysplastic syndromes
Not routinely beneficial
Use intermittently in subset with neutropenia and recurrent infection
What Dose and Schedule Should Be Used?
G-CSF: 5 mg/kg per day subcutaneously
GM-CSF: 250 mg/m^2 per day subcutaneously
Pegfilgrastim: one dose of 6 mg 24 h after chemotherapy
When Should Therapy Begin and End?
When indicated, start 24–72 h after chemotherapy
Continue until absolute neutrophil count is 10,000/μL
Do not use concurrently with chemotherapy or radiation therapy

Abbreviations: CSF, cerebrospinal fluid; G-CSF, granulocyte colony-stimulating factor; GM-CSF, granulocyte-macrophage colony-stimulating factor.

Source: From the American Society of Clinical Oncology: J Clin Oncol 24:3187, 2006.

Primary prophylaxis (i.e., shortly after completing chemotherapy to reduce the nadir) administers G-CSF to patients receiving cytotoxic regimens associated with a 20% incidence of febrile neutropenia. “Dose-dense” regimens, where cycling of chemotherapy is intended to be completed without delay of administered doses,

may also benefit, but such patients should be on a clinical trial. Administration of G-CSF in these circumstances has reduced the incidence of febrile neutropenia in several studies by about 50%. Most patients, however, receive regimens that do not have such a high risk of expected febrile neutropenia, and therefore most patients initially should not receive G-CSF or GM-CSF. Special circumstances—such as a documented history of febrile neutropenia with the regimen in a particular patient or categories of patients at increased risk, such as patients aged >65 years with aggressive lymphoma treated with curative chemotherapy regimens; extensive compromise of marrow by prior radiation or chemotherapy; or active, open wounds or deep-seated infection—may support primary treatment with G-CSF or GM-CSF. Administration of G-CSF or GM-CSF to afebrile neutropenic patients or to patients with low-risk febrile neutropenia is not recommended, and patients receiving concomitant chemoradiation treatment, particularly those with thoracic neoplasms, likewise are not generally recommended for treatment. In contrast, administration of G-CSF to high-risk patients with febrile neutropenia and evidence of organ compromise including sepsis syndrome, invasive fungal infection, concurrent hospitalization at the time fever develops, pneumonia, profound neutropenia ($<0.1 \times 10^9$/L), or age >65 years is reasonable.

Secondary prophylaxis refers to the administration of CSFs in patients who have experienced a neutropenic complication from a prior cycle of chemotherapy; dose reduction or delay may be a reasonably considered alternative. G-CSF or GM-CSF is conventionally started 24–72 h after completion of chemotherapy and continued until a PMN count of 10,000/μL is achieved, unless a “depot” preparation of G-CSF such as pegfilgrastim is used, where one dose is administered at least 14 days before the next scheduled administration of chemotherapy. Also, patients with myeloid leukemias undergoing induction therapy may have a slight reduction in the duration of neutropenia if G-CSF is commenced after completion of therapy, but the influence on long-term outcome has not been defined. GM-CSF probably has a more restricted utility than G-CSF, with its use currently limited to patients after autologous bone marrow transplants, although proper head-to-head comparisons with G-CSF have not been conducted in most instances. GM-CSF may be associated with more systemic side effects.

Dangerous degrees of thrombocytopenia do not frequently complicate the management of patients with solid tumors receiving cytotoxic chemotherapy (with the possible exception of certain carboplatin-containing regimens), but they are frequent in patients with certain hematologic neoplasms where marrow is infiltrated with tumor. Severe bleeding related to thrombocytopenia occurs with increased frequency at platelet counts <20,000/μL and is very prevalent at counts <5000/μL.

The precise “trigger” point at which to transfuse patients has been defined as a platelet count of 10,000/μL or less in patients without medical comorbidities that may increase the risk of bleeding. This issue is important not only because of the costs of frequent transfusion, but unnecessary platelet transfusions expose the patient to the risks of allosensitization and loss of value from subsequent transfusion owing to rapid platelet clearance, as well as the infectious and hypersensitivity risks inherent in any transfusion. Prophylactic transfusions to keep platelets >20,000/μL are reasonable in patients with leukemia who are stressed by fever or concomitant medical conditions (the threshold for transfusion is 10,000/μL in patients with solid tumors and no other bleeding diathesis or physiologic stressors such as fever or hypotension, a level that might also

be reasonably considered for leukemia patients who are thrombocytopenic but not stressed or bleeding). Careful review of medication lists to prevent exposure to nonsteroidal anti-inflammatory agents and maintenance of clotting factor levels adequate to support near-normal prothrombin and partial thromboplastin time tests are important in minimizing the risk of bleeding in the thrombocytopenic patient.

Anemia associated with chemotherapy can be managed by transfusion of packed RBCs. Transfusion is not undertaken until the hemoglobin falls to <80 g/L (8 g/dL), compromise of end organ function occurs, or an underlying condition (e.g., coronary artery disease) calls for maintenance of hemoglobin >90 g/L (9 g/dL). Randomized trials in certain tumors have raised the possibility that erythropoietin (EPO) use may promote tumor-related adverse events.

NAUSEA AND VOMITING

The most common side effect of chemotherapy administration is nausea, with or without vomiting. Nausea may be acute (within 24 h of chemotherapy), delayed (>24 h), or anticipatory of the receipt of chemotherapy. Patients may be likewise stratified for their risk of susceptibility to nausea and vomiting, with increased risk in young, female, heavily pretreated patients without a history of alcohol or drug use but with a history of motion or morning sickness. Antineoplastic agents vary in their capacity to cause nausea and vomiting. Highly emetogenic drugs (>90%) include mechlorethamine, streptozotocin, DTIC, cyclophosphamide at >1500 mg/m^2, and cisplatin; moderately emetogenic drugs (30–90% risk) include carboplatin, cytosine arabinoside (>1 mg/m^2), ifosfamide, conventional-dose cyclophosphamide, and anthracyclines; low-risk (10–30%) agents include 5FU, taxanes, etoposide, and bortezomib, with minimal risk (<10%) afforded by treatment with antibodies, bleomycin, busulfan, fludarabine, and vinca alkaloids.

Serotonin antagonists (5-HT_3) and neurokinin 1 (NK1) receptor antagonists are useful in “high-risk” chemotherapy regimens. The combination acts at both peripheral gastrointestinal and CNS sites that control nausea and vomiting. For example, the 5-HT_3 blocker dolasetron, 100 mg intravenously or orally; dexamethasone, 12 mg; and the NK1 antagonist aprepitant, 125 mg orally are combined on the day of administration of severely emetogenic regimens, with repetition of dexamethasone (8 mg) and aprepitant (80 mg) on days 2 and 3 for delayed nausea. Alternate 5-HT_3 antagonists include ondansetron, given as 0.15 mg/kg intravenously for three doses just before and at 4 and 8 h after chemotherapy; palonosetron at 0.25 mg over 30 s, 30 min before chemotherapy; and granisetron, given as a single dose of 0.01 mg/kg just before chemotherapy. Emesis from moderately emetic chemotherapy regimens may be prevented with a 5-HT_3 antagonist and dexamethasone alone for patients not receiving doxorubicin and cyclophosphamide combinations; the latter combination requires the 5-HT_3/dexamethasone/aprepitant on day 1, but aprepitant alone on days 2 and 3. Emesis from low-emetic-risk regimens may be prevented with 8 mg of dexamethasone alone or with non-5-HT_3, non-NK1 antagonist approaches including the following.

Antidopaminergic phenothiazines act directly at the chemoreceptor trigger zone (CTZ) in the brainstem medulla and include prochlorperazine (Compazine), 10 mg intramuscularly or intravenously, 10–25 mg

orally, or 25 mg per rectum every 4–6 h for up to four doses; and thiethylperazine, 10 mg by potentially all of the above routes every 6 h. Haloperidol is a butyrophenone dopamine antagonist given at 1 mg intramuscularly or orally every 8 h. Metoclopramide acts on peripheral dopamine receptors to augment gastric emptying and is used in high doses for highly emetogenic regimens (1–2 mg/kg intravenously 30 min before chemotherapy and every 2 h for up to three additional doses as needed); intravenous doses of 10–20 mg every 4–6 h as needed or 50 mg orally 4 h before and 8 and 12 h after chemotherapy are used for moderately emetogenic regimens. 5-9-Tetrahydrocannabinol (Marinol) is a rather weak antiemetic compared to other available agents, but it may be useful for persisting nausea and is used orally at 10 mg every 3–4 h as needed.

DIARRHEA

Regimens that include 5FU infusions and/or irinotecan may produce severe diarrhea. Similar to the vomiting syndromes, chemotherapy-induced diarrhea may be immediate or can occur in a delayed fashion up to 48–72 h after the drugs. Careful attention to maintained hydration and electrolyte repletion, intravenously if necessary, along with antimotility treatments such as “high-dose” loperamide, commenced with 4 mg at the first occurrence of diarrhea, with 2 mg repeated every 2 h until 12 h without loose stools, not to exceed a total daily dose of 16 mg. Octreotide (100–150 μg), a somatostatin analogue, or opiate-based preparations may be considered for patients not responding to loperamide.

MUCOSITIS

Irritation and inflammation of the mucous membranes particularly afflicting the oral and anal mucosa, but potentially involving the gastrointestinal tract, may accompany cytotoxic chemotherapy. Mucositis is due to damage to the proliferating cells at the base of the mucosal squamous epithelia or in the intestinal crypts. Topical therapies, including anesthetics and barrier-creating preparations, may provide symptomatic relief in mild cases. Palifermin or keratinocyte growth factor, a member of the fibroblast growth factor family, is effective in preventing severe mucositis in the setting of high-dose chemotherapy with stem cell transplantation for hematologic malignancies. It may also prevent or ameliorate mucositis from radiation.

ALOPECIA

Chemotherapeutic agents vary widely in causing alopecia, with anthracyclines, alkylating agents, and topoisomerase inhibitors reliably causing near-total alopecia when given at therapeutic doses. Antimetabolites are more variably associated with alopecia. Psychological support and the use of cosmetic resources are to be encouraged, and “chemo caps” that reduce scalp temperature to decrease the degree of alopecia are controversial during treatment with curative intent of neoplasms, such as leukemia or lymphoma, or in adjuvant breast cancer therapy. The richly vascularized scalp can certainly harbor micrometastatic or disseminated disease.

GONADAL DYSFUNCTION AND PREGNANCY

Cessation of ovulation and azoospermia reliably result from alkylating agent—and topoisomerase poison—containing regimens. The duration of these effects varies with age and sex. Sperm banking before treatment may be considered. Females experience amenorrhea with anovulation after alkylating agent therapy; egg preservation may be considered, but may delay inception of urgent treatment. Recovery of normal menses is frequent if treatment is completed before age 30 but unlikely to recover menses after age 35. Even those who regain menses usually experience premature menopause. Because the magnitude and extent of decreased fertility can be difficult to predict, patients should be counseled to maintain effective contraception, preferably by barrier means, during and after therapy. Resumption of efforts to conceive should be considered in the context of the patient's likely prognosis. Hormone replacement therapy should be undertaken in women who do not have a hormonally responsive tumor. For patients who have had a hormone-sensitive tumor primarily treated by a local modality, conventional practice would counsel against hormone replacement, but this issue is under investigation.

Chemotherapy agents have variable effects on the success of pregnancy. All agents tend to have increased risk of adverse outcomes when administered during the first trimester, and strategies to delay chemotherapy, if possible, until after this milestone should be considered if the pregnancy is to continue to term. Patients in their second or third trimester can be treated with most regimens for the common neoplasms afflicting women in their childbearing years, with the exception of antimetabolites, particularly antifolates, which have notable teratogenic or fetotoxic effects throughout pregnancy. The need for anticancer chemotherapy per se is infrequently a clear basis to recommend termination of a concurrent pregnancy, although each treatment strategy in this circumstance must be tailored to the individual needs of the patient.

Late effects of cancer and its treatment are reviewed in Chap. 91.

FURTHER READING

Jayson GC et al: Antiangiogenic therapy in oncology: Current status and future directions. Lancet 388:518, 2016. [PubMed: 26853587]

Maus MV et al: Antibody-modified T cells: CARs take the front seat for hematologic malignancies. Blood 123:2625, 2014. [PubMed: 24578504]

Rosenberg SA, Restifo NP: Adoptive cell transfer as personalized immunotherapy for human cancer. Science 348:62, 2015. [PubMed: 25838374]

Socinski MA, Villaruz LC, Ross J: Understanding mechanisms of resistance in the epithelial growth factor receptor in non-small cell lung cancer and the role of biopsy at progression. Oncologist 22:3, 2017.

Swanton C, Govindan R: Clinical implications of genomic discoveries in lung cancer. N Engl J Med 374:1864, 2016. [PubMed: 27168435]

Topalian SL et al: Immune checkpoint blockade: A common denominator approach to cancer therapy. Cancer Cell 27:450, 2015. [PubMed: 25858804]

Chapter 70: Infections in Patients with Cancer

Robert W. Finberg

INTRODUCTION

Infections are a common cause of death and an even more common cause of morbidity in patients with a wide variety of neoplasms. Autopsy studies show that most deaths from acute leukemia and half of deaths from lymphoma are caused directly by infection. With more intensive chemotherapy, patients with solid tumors have also become more likely to die of infection. Fortunately, an evolving approach to prevention and treatment of infectious complications of cancer has decreased infection-associated mortality rates and will probably continue to do so. This accomplishment has resulted from three major steps:

1. *Early treatment:* The practice of using "early empirical" antibiotics reduced mortality rates among patients with leukemia and bacteremia from 84% in 1965 to 44% in 1972. The mortality rate due to infection in febrile neutropenic patients dropped to <10% by 2013. This dramatic improvement is attributed to early intervention with appropriate antimicrobial therapy.

2. *Empirical treatment:* "Empirical" antifungal therapy has also lowered the incidence of disseminated fungal infection, with dramatic decreases in mortality rates. An antifungal agent is administered—on the basis of likely fungal infection—to neutropenic patients who, after 4–7 days of antibiotic therapy, remain febrile but have no positive cultures.

3. *Prophylaxis:* Use of antibiotics for afebrile neutropenic patients as broad-spectrum prophylaxis against infections has decreased both mortality and morbidity even further. The current approach to treatment of severely neutropenic patients (e.g., those receiving high-dose chemotherapy for leukemia or high-grade lymphoma) is based on initial prophylactic therapy at the onset of neutropenia, subsequent "empirical" antibacterial therapy targeting the organisms whose involvement is likely in light of physical findings (most often fever alone), and finally "empirical" antifungal therapy based on the known likelihood that fungal infection will become a serious issue after 4–7 days of broad-spectrum antibacterial therapy.

A physical predisposition to infection in patients with cancer (Table 70-1) can be a result of the neoplasm's production of a break in the skin. For example, a squamous cell carcinoma may cause local invasion of the epidermis, which allows bacteria to gain access to subcutaneous tissue and permits the development of cellulitis. The artificial closing of a normally patent orifice can also predispose to infection; for example, obstruction of a ureter by a tumor can cause urinary tract infection, and obstruction of the bile duct can

cause cholangitis. Part of the host's normal defense against infection depends on the continuous emptying of a viscus; without emptying, a few bacteria that are present as a result of bacteremia or local transit can multiply and cause disease.

TABLE 70-1

Disruption of Normal Barriers in Patients with Cancer That May Predispose Them to Infections

Type of Defense	Specific Lesion	Cells Involved	Organism	Cancer Association	Disease
Physical barrier	Breaks in skin	Skin epithelial cells	Staphylococci, streptococci	Head and neck, squamous cell carcinoma	Cellulitis, extensive skin infection
Emptying of fluid collections	Occlusion of orifices: ureters, bile duct, colon	Luminal epithelial cells	Gram-negative bacilli	Renal, ovarian, biliary tree, metastatic diseases of many cancers	Rapid, overwhelming bacteremia; urinary tract infection
Lymphatic function	Node dissection	Lymph nodes	Staphylococci, streptococci	Breast cancer surgery	Cellulitis
Splenic clearance of microorganisms	Splenectomy	Splenic reticuloendothelial cells	*Streptococcus pneumoniae, Haemophilus influenzae, Neisseria meningitidis, Babesia, Capnocytophaga canimorsus*	Hodgkin's disease, leukemia	Rapid, overwhelming sepsis

Type of Defense	Specific Lesion	Cells Involved	Organism	Cancer Association	Disease
Phagocytosis	Lack of granulocytes	Granulocytes (neutrophils)	Staphylococci, streptococci, enteric organisms, fungi	Acute myeloid and acute lymphocytic leukemias, hairy cell leukemia	Bacteremia
Humoral immunity	Lack of antibody	B cells	*S. pneumoniae, H. influenzae, N. meningitidis*	Chronic lymphocytic leukemia, multiple myeloma	Infections with encapsulated organisms, sinusitis, pneumonia
Cellular immunity	Lack of T cells	T cells and macrophages	*Mycobacterium tuberculosis, Listeria*, herpesviruses, fungi, intracellular parasites	Hodgkin's disease, leukemia, T cell lymphoma	Infections with intracellular bacteria, fungi, parasites; virus reactivation

A similar problem can affect patients whose lymph node integrity has been disrupted by radical surgery, particularly patients who have had radical node dissections. A common clinical problem following radical mastectomy is the development of cellulitis (usually caused by streptococci or staphylococci) because of lymphedema and/or inadequate lymph drainage. In most cases, this problem can be addressed by local measures designed to prevent fluid accumulation and breaks in the skin, but antibiotic prophylaxis has been used in refractory cases.

A life-threatening problem common to many cancer patients is the loss of the reticuloendothelial capacity to clear microorganisms after splenectomy, which may be performed as part of the management of hairy cell leukemia, chronic lymphocytic leukemia (CLL), and chronic myelogenous leukemia (CML) and in Hodgkin's disease. Even after curative therapy for the underlying disease, the lack of a spleen predisposes such patients to rapidly fatal infections. The loss of the spleen through trauma similarly predisposes the normal host to overwhelming infection throughout life. The splenectomized patient should be counseled about the risks of

infection with certain organisms, such as the protozoan *Babesia* **(Chap. 220)** and *Capnocytophaga canimorsus*, a bacterium carried in the mouths of animals **(Chaps. 136 and 153)**. Because encapsulated bacteria (*Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Neisseria meningitidis*) are the organisms most commonly associated with postsplenectomy sepsis, splenectomized persons should be vaccinated (and revaccinated; **Table 70-2** and **Chap. 118**) against the capsular polysaccharides of these organisms. Many clinicians recommend giving splenectomized patients a small supply of antibiotics effective against *S. pneumoniae*, *N. meningitidis*, and *H. influenzae* to avert rapid, overwhelming sepsis in the event that they cannot present for medical attention immediately after the onset of fever or other signs or symptoms of bacterial infection. A few tablets of amoxicillin/clavulanic acid (or levofloxacin if resistant strains of *S. pneumoniae* are prevalent locally) are a reasonable choice for this purpose.

TABLE 70-2

Vaccination of Cancer Patients Receiving Chemotherapy[a]

	Use in Indicated Patients		
Vaccine	**Intensive Chemotherapy**	**Hodgkin's Disease**	**Hematopoietic Stem Cell Transplantation**
Diphtheria-tetanus-pertussis[b]	Primary series and boosters as necessary	No special recommendation	3 doses given 6–12 months after transplantation
Poliomyelitis[c]	Complete primary series and boosters	No special recommendation	3 doses given 6–12 months after transplantation
Haemophilus influenzae type b conjugate	Primary series and booster for children	Single dose for adults	3 doses given 6–12 months after transplantation (separated by 1 month)
Human papillomavirus (HPV)	HPV vaccine is approved for males and females 9–26 years of age. Check Centers for Disease Control and Prevention (CDC) website (*www.cdc.gov/vaccines*) for updated recommendations.	HPV vaccine is approved for males and females 9–26 years of age. Check CDC website (*www.cdc.gov/vaccines*) for updated recommendations.	HPV vaccine is approved for males and females 9–26 years of age. Check CDC website (*www.cdc.gov/vaccines*) for updated recommendations.
Hepatitis A	As indicated for normal hosts on the basis of occupation and lifestyle	As indicated for normal hosts on the basis of occupation and lifestyle	As indicated for normal hosts on the basis of occupation and lifestyle
Hepatitis B	Same as for normal hosts	As indicated for normal hosts on the basis of occupation and lifestyle	3 doses given 6–12 months after transplantation

	Use in Indicated Patients		
Vaccine	**Intensive Chemotherapy**	**Hodgkin's Disease**	**Hematopoietic Stem Cell Transplantation**
Pneumococcal conjugate vaccine (PCV13) Pneumococcal polysaccharide vaccine (PPSV23)[d]	Finish series prior to chemotherapy if possible.	Patients with splenectomy should receive both PCV13 and PPSV23.	Three doses of PCV13, beginning 3–6 months after transplantation, are followed by a dose of PPSV23 at least 8 weeks later. A second PPSV23 dose can be given 5 years later.
Quadrivalent meningococcal vaccine[e]	Should be administered to splenectomized patients and to patients living in endemic areas, including college students in dormitories	Should be administered to splenectomized patients and to patients living in endemic areas, including college students in dormitories. An additional dose can be given after 5 years.	Should be administered to splenectomized patients and to patients living in endemic areas, including college students in dormitories. An additional dose can be given after 5 years.
Meningococcal B vaccine	See above.	See above.	See above (see *www.cdc.gov/vaccines* for updated recommendations).
Influenza	Seasonal immunization	Seasonal immunization	Seasonal immunization (A seasonal dose is recommended and can be given as early as 4 months after transplantation; if given <6 months after transplantation, an additional dose is recommended.)

	Use in Indicated Patients		
Vaccine	**Intensive Chemotherapy**	**Hodgkin's Disease**	**Hematopoietic Stem Cell Transplantation**
Measles/mumps/rubella	Contraindicated	Contraindicated during chemotherapy	After 24 months in patients without graft-versus-host disease
Varicella-zoster virus[f]	Contraindicated[g]	Contraindicated	Contraindicated (CDC recommends use on a case-by-case basis following reevaluation.)

[a]The latest recommendations by the Advisory Committee on Immunization Practices and the CDC guidelines can be found at *www.cdc.gov/vaccines*. [b]A single dose of TDaP (tetanus–diphtheria–acellular pertussis), followed by a booster dose of Td (tetanus-–diphtheria) every 10 years, is recommended for adults. [c]Live-virus vaccine is contraindicated; inactivated vaccine should be used. [d]Two types of vaccines are used to prevent pneumococcal disease. A conjugate vaccine active against 13 serotypes (13-valent pneumococcal conjugate vaccine, or PCV13) is currently administered in three separate doses to all children. A polysaccharide vaccine active against 23 serotypes (23-valent pneumococcal polysaccharide vaccine, or PPSV23) elicits titers of antibody lower than those achieved with the conjugate vaccine, and immunity may wane more rapidly. Because the ablative chemotherapy given to recipients of hematopoietic stem cell transplants (HSCTs) eradicates immunologic memory, revaccination is recommended for all such patients. Vaccination is much more effective once immunologic reconstitution has occurred; however, because of the need to prevent serious disease, pneumococcal vaccine should be administered 6–12 months after transplantation in most cases. Because PPSV23 includes serotypes not present in PCV13, HSCT recipients should receive a dose of PPSV23 at least 8 weeks after the last dose of PCV13. Although antibody titers from PPSV23 clearly decay, experience with multiple doses of PPSV23 is limited, as are data on the safety, toxicity, or efficacy of such a regimen. For this reason, the CDC currently recommends the administration of one additional dose of PPSV23 at least 5 years after the last dose to immunocompromised patients, including transplant recipients, as well as patients with Hodgkin's disease, multiple myeloma, lymphoma, or generalized malignancies. Beyond this single additional dose, further doses are not recommended at this time. [e]Meningococcal conjugate vaccine (MenACWY) is recommended for adults ≤55 years old, and meningococcal polysaccharide vaccine (MPSV4) is recommended for those ≥56 years old. [f]Includes both varicella vaccine for children and zoster vaccine for adults. [g]Contact the manufacturer for more information on use in children with acute lymphocytic leukemia.

The level of suspicion of infections with certain organisms should depend on the type of cancer diagnosed **(Table 70-3)**. Diagnosis of multiple myeloma or CLL should alert the clinician to the possibility of hypogammaglobulinemia. While immunoglobulin replacement therapy can be effective, in most cases

prophylactic antibiotics are a cheaper, more convenient method of eliminating bacterial infections in CLL patients with hypogammaglobulinemia. Patients with acute lymphocytic leukemia (ALL), patients with non-Hodgkin's lymphoma, and all cancer patients treated with high-dose glucocorticoids (or glucocorticoid-containing chemotherapy regimens) should receive antibiotic prophylaxis for *Pneumocystis* infection (Table 70-3) for the duration of their chemotherapy. In addition to exhibiting susceptibility to certain infectious organisms, patients with cancer are likely to manifest their infections in characteristic ways. For example, fever—generally a sign of infection in normal hosts—continues to be a reliable indicator in neutropenic patients. In contrast, patients receiving glucocorticoids and agents that impair T cell function and cytokine secretion may have serious infections in the absence of fever. Similarly, neutropenic patients commonly present with cellulitis without purulence and with pneumonia without sputum or even x-ray findings (see below).

TABLE 70-3

Infections Associated with Specific Types of Cancer

Cancer	Underlying Immune Abnormality	Organism(s) Causing Infection
Multiple myeloma	Hypogammaglobulinemia	*Streptococcus pneumoniae, Haemophilus influenzae, Neisseria meningitidis*
Chronic lymphocytic leukemia	Hypogammaglobulinemia	*S. pneumoniae, H. influenzae, N. meningitidis*
Acute myeloid or lymphocytic leukemia	Granulocytopenia, skin and mucous membrane lesions	Extracellular gram-positive and gram-negative bacteria, fungi
Hodgkin's disease	Abnormal T cell function	Intracellular pathogens (*Mycobacterium tuberculosis, Listeria, Salmonella, Cryptococcus, Mycobacterium avium*); herpesviruses
Non-Hodgkin's lymphoma and acute lymphocytic leukemia	Glucocorticoid chemotherapy, T and B cell dysfunction	*Pneumocystis*
Colon and rectal tumors	Local abnormalities[a]	*Streptococcus bovis* biotype 1 (bacteremia)
Hairy cell leukemia	Abnormal T cell function	Intracellular pathogens (*M. tuberculosis, Listeria, Cryptococcus, M. avium*)

[a]The reason for this association is not well defined.

The use of monoclonal antibodies that target B and T cells as well as drugs that interfere with lymphocyte signal transduction events is associated with reactivation of latent infections. The use of rituximab, the antibody to CD20 (a B cell surface protein), is associated with the development of reactivation tuberculosis as well as other latent viral infections, including hepatitis B and cytomegalovirus (CMV) infection. Like organ transplant recipients **(Chap. 138)**, patients with latent bacterial disease (like tuberculosis) and latent viral disease (like herpes simplex or zoster) should be carefully monitored for reactivation disease.

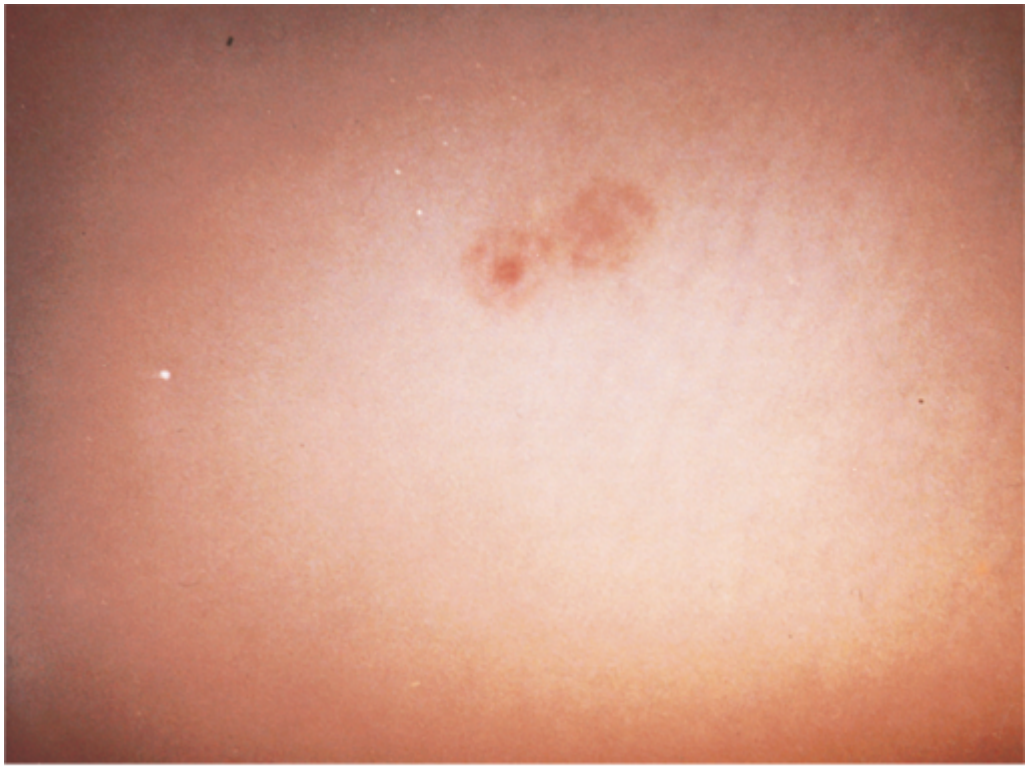
SYSTEM-SPECIFIC SYNDROMES

SKIN-SPECIFIC SYNDROMES

Skin lesions are common in cancer patients, and the appearance of these lesions may permit the diagnosis of systemic bacterial or fungal infection. While cellulitis caused by skin organisms such as *Streptococcus* or *Staphylococcus* is common, neutropenic patients—that is, those with <500 functional polymorphonuclear leukocytes (PMNs)/μL—and patients with impaired blood or lymphatic drainage may develop infections with unusual organisms. Innocent-looking macules or papules may be the first sign of bacterial or fungal sepsis in immunocompromised patients **(Fig. 70-1)**. In the neutropenic host, a macule progresses rapidly to ecthyma gangrenosum **(see Fig. A1-34)**, a usually painless, round, necrotic lesion consisting of a central black or gray-black eschar with surrounding erythema. Ecthyma gangrenosum, which is located in nonpressure areas (as distinguished from necrotic lesions associated with lack of circulation), is often associated with *Pseudomonas aeruginosa* bacteremia **(Chap. 159)** but may be caused by other bacteria.

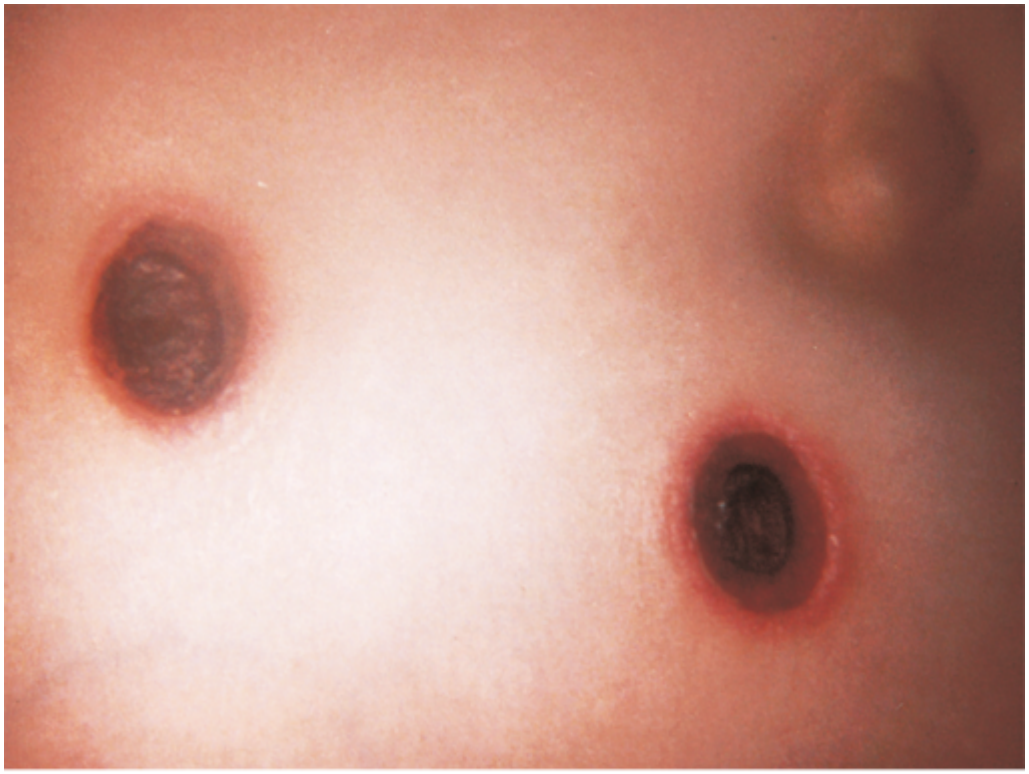
FIGURE 70-1

A. Papules related to *Escherichia coli* bacteremia in a patient with acute lymphocytic leukemia. ***B***. The same lesions on the following day.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Candidemia **(Chap. 211)** is also associated with a variety of skin conditions **(see Fig. A1-37)** and commonly presents as a maculopapular rash. Punch biopsy of the skin may be the best method for diagnosis.

Cellulitis, an acute spreading inflammation of the skin, is most often caused by infection with group A *Streptococcus* or *Staphylococcus aureus*, virulent organisms normally found on the skin **(Chap. 124)**. Although cellulitis tends to be circumscribed in normal hosts, it may spread rapidly in neutropenic patients. A tiny break in the skin may lead to spreading cellulitis, which is characterized by pain and erythema; in the affected patients, signs of infection (e.g., purulence) are often lacking. What might be a furuncle in a normal host may require amputation because of uncontrolled infection in a patient presenting with leukemia. A dramatic response to an infection that might be trivial in a normal host can mark the first sign of leukemia. Fortunately, granulocytopenic patients are likely to be infected with certain types of organisms **(Table 70-4)**; thus the selection of an antibiotic regimen is somewhat easier than it might otherwise be (see "Antibacterial Therapy," below). It is essential to recognize cellulitis early and to treat it aggressively. Patients who are neutropenic or who have previously received antibiotics for other reasons may develop cellulitis with unusual organisms (e.g., *Escherichia coli*, *Pseudomonas*, or fungi). Early treatment, even of innocent-looking lesions, is essential to prevent necrosis and loss of tissue. Debridement to prevent spread may sometimes be necessary early in the course of disease, but it can often be performed after chemotherapy, when the PMN count increases.

TABLE 70-4

Organisms Likely to Cause Infections in Granulocytopenic Patients

Gram-Positive Cocci	
Staphylococcus epidermidis	*Staphylococcus aureus*
Viridans *Streptococcus*	*Enterococcus faecalis*
Streptococcus pneumoniae	
Gram-Negative Bacilli	
Escherichia coli	*Serratia* spp.
Klebsiella spp.	*Acinetobacter* spp.[a]
Pseudomonas aeruginosa	*Stenotrophomonas* spp.
Enterobacter spp.	*Citrobacter* spp.
Non-*aeruginosa Pseudomonas* spp.[a]	
Gram-Positive Bacilli	
Diphtheroids	JK bacillus[a]
Fungi	
Candida spp.	*Mucor/Rhizopus*
Aspergillus spp.	

[a]Often associated with intravenous catheters.

Sweet syndrome, or *febrile neutrophilic dermatosis*, was originally described in women with elevated white blood cell (WBC) counts. The disease is characterized by the presence of leukocytes in the lower dermis, with edema of the papillary body. Ironically, this disease now is usually seen in neutropenic patients with cancer, most often in association with acute myeloid leukemia (AML) but also in association with a variety of other malignancies. Sweet syndrome usually presents as red or bluish-red papules or nodules that may coalesce

and form sharply bordered plaques (see Fig. A1-40). The edema may suggest vesicles, but on palpation the lesions are solid, and vesicles probably never arise in this disease. The lesions are most common on the face, neck, and arms. On the legs, they may be confused with erythema nodosum (see Fig. A1-39). The development of lesions is often accompanied by high fevers and an elevated erythrocyte sedimentation rate. Both the lesions and the temperature elevation respond dramatically to glucocorticoid administration. Treatment begins with high doses of glucocorticoids (prednisone, 60 mg/d) followed by tapered doses over the next 2–3 weeks.

Data indicate that *erythema multiforme* (see Fig. A1-24) with mucous membrane involvement is often associated with herpes simplex virus (HSV) infection and is distinct from Stevens-Johnson syndrome, which is associated with drugs and tends to have a more widespread distribution. Because cancer patients are both immunosuppressed (and therefore susceptible to herpes infections) and heavily treated with drugs (and therefore subject to Stevens-Johnson syndrome [see Fig. A2-4]), both of these conditions are common in this population.

Cytokines, which are used as adjuvants or primary treatments for cancer, can themselves cause characteristic rashes, further complicating the differential diagnosis. This phenomenon is a particular problem in bone marrow transplant recipients (Chap. 138), who, in addition to having the usual chemotherapy-, antibiotic-, and cytokine-induced rashes, are plagued by graft-versus-host disease.

CATHETER-RELATED INFECTIONS

Because IV catheters are commonly used in cancer chemotherapy and are prone to cause infection (Chap. 137), they pose a major problem in the care of patients with cancer. Some catheter-associated infections can be treated with antibiotics, whereas in others the catheter must be removed (Table 70-5). If the patient has a “tunneled” catheter (which consists of an entrance site, a subcutaneous tunnel, and an exit site), a red streak over the subcutaneous part of the line (the tunnel) is grounds for immediate device removal. Failure to remove catheters under these circumstances may result in extensive cellulitis and tissue necrosis.

TABLE 70-5

Approach to Catheter Infections in Immunocompromised Patients

Clinical Presentation or Isolated Pathogen	Catheter Removal	Antibiotics	Comments
Evidence of Infection, Negative Blood Cultures			
Exit-site erythema	Not necessary if infection responds to treatment	Usually, begin treatment for gram-positive cocci.	Coagulase-negative staphylococci are most common.
Tunnel-site erythema	Required	Treat for gram-positive cocci pending culture results.	Failure to remove the catheter may lead to necrosis of the involved area requiring skin grafts in the future.
Blood Culture–Positive Infections			
Coagulase-negative staphylococci	Line removal optimal but may be unnecessary if patient is clinically stable and responds to antibiotics	Usually, start with vancomycin. Linezolid, quinupristin/dalfopristin, and daptomycin are alternative agents.	If there are no contraindications to line removal, this course of action is optimal. If the line is removed, antibiotics may not be necessary.
Other gram-positive cocci (e.g., *Staphylococcus aureus, Enterococcus*); gram-positive rods (*Bacillus, Corynebacterium* spp.)	Recommended	Treat with antibiotics to which the organism is sensitive, with duration based on the clinical setting.	The incidence of metastatic infections following *S. aureus* infection and the difficulty of treating enterococcal infection make line removal the recommended course of action. In addition, gram-positive rods do not respond readily to antibiotics alone.

Clinical Presentation or Isolated Pathogen	Catheter Removal	Antibiotics	Comments
Gram-negative bacteria	Recommended	Use an agent to which the organism is shown to be sensitive.	Organisms like *Stenotrophomonas, Pseudomonas*, and *Burkholderia* are notoriously hard to treat, as are carbapenem-resistant organisms.
Fungi	Recommended	—	Fungal infections of catheters are extremely difficult to treat.

More common than tunnel infections are exit-site infections, often with erythema around the area where the line penetrates the skin. Most authorities (Chap. 142) recommend treatment (usually with vancomycin) for an exit-site infection caused by coagulase-negative *Staphylococcus*. Treatment of coagulase-positive staphylococcal infection is associated with a poorer outcome, and it is advisable to remove the catheter if possible. Similarly, most clinicians remove catheters associated with infections due to *P. aeruginosa* and *Candida* species, because such infections are difficult to treat and bloodstream infections with these organisms are likely to be deadly. Catheter infections caused by *Burkholderia cepacia*, *Stenotrophomonas* species, *Agrobacterium* species, *Acinetobacter baumannii*, *Pseudomonas* species other than *aeruginosa*, and carbapenem-resistant Enterobacteriaceae are likely to be very difficult to eradicate with antibiotics alone. Similarly, isolation of *Bacillus*, *Corynebacterium*, and *Mycobacterium* species should prompt removal of the catheter.

GASTROINTESTINAL TRACT–SPECIFIC SYNDROMES

Upper Gastrointestinal Tract Disease

• *Infections of the mouth* The oral cavity is rich in aerobic and anaerobic bacteria (Chap. 172) that normally live in a commensal relationship with the host. The antimetabolic effects of chemotherapy cause a breakdown of mucosal host defenses, leading to ulceration of the mouth and the potential for invasion by resident bacteria. Mouth ulcerations afflict most patients receiving cytotoxic chemotherapy and have been associated with viridans streptococcal bacteremia. *Candida* infections of the mouth are very common. Fluconazole is clearly effective in the treatment of both local infections (thrush) and systemic infections (esophagitis) due to *Candida albicans*. Other azoles (e.g., voriconazole) as well as echinocandins offer similar efficacy as well as activity against the fluconazole-resistant organisms that are associated with chronic fluconazole treatment (Chap. 211).

Noma (*cancrum oris*), commonly seen in malnourished children, is a penetrating disease of the soft and hard tissues of the mouth and adjacent sites, with resulting necrosis and gangrene. It has a counterpart in

immunocompromised patients and is thought to be due to invasion of the tissues by *Bacteroides*, *Fusobacterium*, and other normal inhabitants of the mouth. Noma is associated with debility, poor oral hygiene, and immunosuppression.

Viruses, particularly HSV, are a prominent cause of morbidity in immunocompromised patients, in whom they are associated with severe mucositis. The use of acyclovir, either prophylactically or therapeutically, is of value.

ESOPHAGEAL INFECTIONS

The differential diagnosis of esophagitis (usually presenting as substernal chest pain upon swallowing) includes herpes simplex and candidiasis, both of which are readily treatable.

Lower Gastrointestinal Tract Disease

Hepatic candidiasis **(Chap. 211)** results from seeding of the liver (usually from a gastrointestinal source) in neutropenic patients. It is most common among patients being treated for AML and usually presents symptomatically around the time neutropenia resolves. The characteristic picture is that of persistent fever unresponsive to antibiotics, abdominal pain and tenderness or nausea, and elevated serum levels of alkaline phosphatase in a patient with hematologic malignancy who has recently recovered from neutropenia. The diagnosis of this disease (which may present in an indolent manner and persist for several months) is based on the finding of yeasts or pseudohyphae in granulomatous lesions. Hepatic ultrasound or CT may reveal bull's-eye lesions. MRI scans reveal small lesions not visible by other imaging modalities. The pathology (a granulomatous response) and the timing (with resolution of neutropenia and an elevation in granulocyte count) suggest that the host response to *Candida* is an important component of the manifestations of disease. In many cases, although organisms are visible, cultures of biopsied material may be negative. The designation *hepatosplenic candidiasis* or *hepatic candidiasis* is a misnomer because the disease often involves the kidneys and other tissues; the term *chronic disseminated candidiasis* may be more appropriate. Because of the risk of bleeding with liver biopsy, diagnosis is often based on imaging studies (MRI, CT). Treatment should be directed to the causative agent (usually *C. albicans* but sometimes *Candida tropicalis* or other less common *Candida* species).

Typhlitis

Typhlitis (also referred to as necrotizing colitis, neutropenic colitis, necrotizing enteropathy, ileocecal syndrome, and cecitis) is a clinical syndrome of fever and right-lower-quadrant (or generalized abdominal) tenderness in an immunosuppressed host. This syndrome is classically seen in neutropenic patients after chemotherapy with cytotoxic drugs. It may be more common among children than among adults and appears to be much more common among patients with AML or ALL than among those with other types of cancer. Physical examination reveals right-lower-quadrant tenderness, with or without rebound tenderness. Associated diarrhea (often bloody) is common, and the diagnosis can be confirmed by the finding of a thickened cecal wall on CT, MRI, or ultrasonography. Plain films may reveal a right-lower-quadrant mass, but CT with contrast or MRI is a much more sensitive means of diagnosis. Although surgery is sometimes

attempted to avoid perforation from ischemia, most cases resolve with medical therapy alone. The disease is sometimes associated with positive blood cultures (which usually yield aerobic gram-negative bacilli), and therapy is recommended for a broad spectrum of bacteria (particularly gram-negative bacilli, which are likely to be found in the bowel flora).

Clostridium difficile–Induced Diarrhea

Patients with cancer are predisposed to the development of *C. difficile* diarrhea **(Chap. 129)** as a consequence of chemotherapy alone. Thus, they may test positive for *C. difficile* even without receiving antibiotics. Obviously, such patients are also subject to *C. difficile*–induced diarrhea as a result of antibiotic pressure. *C. difficile* should always be considered as a possible cause of diarrhea in cancer patients who have received either chemotherapy or antibiotics. New approaches to treatment of *C. difficile*–induced diarrhea and to prevention of *C. difficile* expansion as part of the gut microbiota may make this disease less troublesome in the future.

CENTRAL NERVOUS SYSTEM–SPECIFIC SYNDROMES

Meningitis

The presentation of meningitis in patients with lymphoma or CLL and in patients receiving chemotherapy (particularly with glucocorticoids) for solid tumors suggests a diagnosis of cryptococcal or listerial infection. As noted previously, splenectomized patients are susceptible to rapid, overwhelming infection with encapsulated bacteria (including *S. pneumoniae*, *H. influenzae*, and *N. meningitidis*). Similarly, patients who are antibody-deficient (e.g., those with CLL, those who have received intensive chemotherapy, or those who have undergone bone marrow transplantation) are likely to have infections caused by these bacteria. Other cancer patients, however, because of their defective cellular immunity, are likely to be infected with other pathogens (Table 70-3). Central nervous system (CNS) tuberculosis should be considered, especially in patients from countries where tuberculosis is highly prevalent in the population.

Encephalitis

The spectrum of disease resulting from viral encephalitis is expanded in immunocompromised patients. A predisposition to infections with intracellular organisms similar to those encountered in patients with AIDS **(Chap. 197)** is seen in cancer patients receiving (1) high-dose cytotoxic chemotherapy, (2) chemotherapy affecting T cell function (e.g., fludarabine), or (3) antibodies that eliminate T cells (e.g., anti-CD3, alemtuzumab, anti-CD52) or cytokine activity (anti–tumor necrosis factor agents or interleukin 1 receptor antagonists). Infection with varicella-zoster virus (VZV) has been associated with encephalitis that may be caused by VZV-related vasculitis. Chronic viral infections may also be associated with dementia and encephalitic presentations. A diagnosis of progressive multifocal leukoencephalopathy **(Chap. 133)** should be considered when a patient who has received chemotherapy (rituximab in particular) presents with dementia **(Table 70-6)**. Other abnormalities of the CNS that may be confused with infection include normal-pressure

hydrocephalus and vasculitis resulting from CNS irradiation. It may be possible to differentiate these conditions by MRI.

TABLE 70-6

Differential Diagnosis of Central Nervous System Infections in Patients with Cancer

	Underlying Predisposition	
Findings on CT or MRI	**Prolonged Neutropenia**	**Defects in Cellular Immunity[a]**
Mass lesions	*Aspergillus, Nocardia*, or *Cryptococcus* brain abscess	Toxoplasmosis, Epstein-Barr virus lymphoma (rare)
Diffuse encephalitis	Progressive multifocal leukoencephalopathy (JC virus)	Infection with varicella-zoster virus, cytomegalovirus, herpes simplex virus, human herpesvirus type 6, JC virus, *Listeria*

[a]High-dose glucocorticoid therapy, cytotoxic chemotherapy.

Brain Masses

Mass lesions of the brain most often present as headache with or without fever or neurologic abnormalities. Infections associated with mass lesions may be caused by bacteria (particularly *Nocardia*), fungi (particularly *Cryptococcus* or *Aspergillus*), or parasites (*Toxoplasma*). Epstein-Barr virus (EBV)–associated lymphoma may also present as single—or sometimes multiple—mass lesions of the brain. A biopsy may be required for a definitive diagnosis.

PULMONARY INFECTIONS

Pneumonia (Chap. 121) in immunocompromised patients may be difficult to diagnose because conventional methods of diagnosis depend on the presence of neutrophils. Bacterial pneumonia in neutropenic patients may present without purulent sputum—or, in fact, without any sputum at all—and may not produce physical findings suggestive of chest consolidation (rales or egophony).

In granulocytopenic patients with persistent or recurrent fever, the chest x-ray pattern may help to localize an infection and thus to determine which investigative tests and procedures should be undertaken and which therapeutic options should be considered (Table 70-7). In this setting, a simple chest x-ray is a screening tool; because the impaired host response results in less evidence of consolidation or infiltration, high-resolution CT is recommended for the diagnosis of pulmonary infections. The difficulties encountered in the

management of pulmonary infiltrates relate in part to the difficulties of performing diagnostic procedures on the patients involved. When platelet counts can be increased to adequate levels by transfusion, microscopic and microbiologic evaluation of the fluid obtained by endoscopic bronchial lavage is often diagnostic. Lavage fluid should be cultured for *Mycoplasma*, *Chlamydia*, *Legionella*, *Nocardia*, more common bacterial pathogens, fungi, and viruses. In addition, the possibility of *Pneumocystis* pneumonia should be considered, especially in patients with ALL or lymphoma who have not received prophylactic trimethoprim-sulfamethoxazole (TMP-SMX). The characteristics of the infiltrate may be helpful in decisions about further diagnostic and therapeutic maneuvers. Nodular infiltrates suggest fungal pneumonia (e.g., that caused by *Aspergillus* or *Mucor*). Such lesions may best be approached by visualized biopsy procedures. It is worth noting that while bacterial pneumonias classically present as lobar infiltrates in normal hosts, bacterial pneumonias in granulocytopenic hosts present with a paucity of signs, symptoms, or radiographic abnormalities; thus, the diagnosis is difficult.

TABLE 70-7

Differential Diagnosis of Chest Infiltrates in Immunocompromised Patients

	Cause of Pneumonia	
Infiltrate	**Infectious**	**Noninfectious**
Localized	Bacteria (including *Legionella*, mycobacteria)	Local hemorrhage or embolism, tumor
Nodular	Fungi (e.g., *Aspergillus* or *Mucor*), *Nocardia*	Recurrent tumor
Diffuse	Viruses (especially cytomegalovirus), *Chlamydia, Pneumocystis, Toxoplasma gondii*, mycobacteria	Congestive heart failure, radiation pneumonitis, drug-induced lung injury, lymphangitic spread of cancer

Aspergillus species **(Chap. 212)** can colonize the skin and respiratory tract or cause fatal systemic illness. Although this fungus may cause aspergillomas in a previously existing cavity or may produce allergic bronchopulmonary disease in some patients, the major problem posed by this genus in neutropenic patients is invasive disease, primarily due to *Aspergillus fumigatus* or *Aspergillus flavus*. The organisms enter the host following colonization of the respiratory tract, with subsequent invasion of blood vessels. The disease is likely to present as a thrombotic or embolic event because of this ability of the fungi to invade blood vessels. The risk of infection with *Aspergillus* correlates directly with the duration of neutropenia. In prolonged neutropenia, positive surveillance cultures for nasopharyngeal colonization with *Aspergillus* may predict the development of disease.

Patients with *Aspergillus* infection often present with pleuritic chest pain and fever, which are sometimes accompanied by cough. Hemoptysis may be an ominous sign. Chest x-rays may reveal new focal infiltrates or nodules. Chest CT may reveal a characteristic halo consisting of a mass-like infiltrate surrounded by an area of low attenuation. The presence of a "crescent sign" on chest x-ray or chest CT, in which the mass progresses to central cavitation, is characteristic of invasive *Aspergillus* infection but may develop as the lesions are resolving.

In addition to causing pulmonary disease, *Aspergillus* may invade through the nose or palate, with deep sinus penetration. The appearance of a discolored area in the nasal passages or on the hard palate should prompt a search for invasive *Aspergillus*. This situation is likely to require surgical debridement. Catheter infections with *Aspergillus* usually require both removal of the catheter and antifungal therapy.

Diffuse interstitial infiltrates suggest viral, parasitic, or *Pneumocystis* pneumonia. If the patient has a diffuse interstitial pattern on chest x-ray, it may be reasonable, while considering invasive diagnostic procedures, to institute empirical treatment for *Pneumocystis* with TMP-SMX and for *Chlamydia*, *Mycoplasma*, and *Legionella* with a quinolone or azithromycin. Noninvasive procedures, such as staining of induced sputum smears for *Pneumocystis*, serum cryptococcal antigen tests, and urine testing for *Legionella* antigen, may be helpful. Serum galactomannan and β-D-glucan tests may be of value in diagnosing *Aspergillus* infection, but their utility is limited by their lack of sensitivity and specificity. The presence of an elevated level of β-D-glucan in the serum of a patient being treated for cancer who is not receiving prophylaxis against *Pneumocystis* suggests the diagnosis of *Pneumocystis* pneumonia. Infections with viruses that cause only upper respiratory symptoms in immunocompetent hosts, such as respiratory syncytial virus (RSV), influenza viruses, and parainfluenza viruses, may be associated with fatal pneumonitis in immunocompromised hosts. CMV reactivation occurs in cancer patients receiving chemotherapy, but CMV pneumonia is most common among hematopoietic stem cell transplant (HSCT) recipients **(Chap. 138)**. Polymerase chain reaction testing now allows rapid diagnosis of viral pneumonia, which can lead to treatment in some cases (e.g., influenza). Multiplex studies that can detect a wide array of viruses in the lung and upper respiratory tract are now available and will lead to specific diagnoses of viral pneumonias.

Bleomycin is the most common cause of chemotherapy-induced lung disease. Other causes include alkylating agents (such as cyclophosphamide, chlorambucil, and melphalan), nitrosoureas (carmustine [BCNU], lomustine [CCNU], and methyl-CCNU), busulfan, procarbazine, methotrexate, and hydroxyurea. Both infectious and noninfectious (drug- and/or radiation-induced) pneumonitis can cause fever and abnormalities on chest x-ray; thus, the differential diagnosis of an infiltrate in a patient receiving chemotherapy encompasses a broad range of conditions (Table 70-7). The treatment of radiation pneumonitis (which may respond dramatically to glucocorticoids) or drug-induced pneumonitis is different from that of infectious pneumonia, and a biopsy may be important in the diagnosis. Unfortunately, no definitive diagnosis can be made in ~30% of cases, even after bronchoscopy.

Open-lung biopsy is the gold standard of diagnostic techniques. Biopsy via a visualized thoracostomy can replace an open procedure in many cases. When a biopsy cannot be performed, empirical treatment can be undertaken; a quinolone or an erythromycin derivative (azithromycin) and TMP-SMX are used in the case of

diffuse infiltrates, and an antifungal agent is administered in the case of nodular infiltrates. The risks should be weighed carefully in these cases. If inappropriate drugs are administered, empirical treatment may prove toxic or ineffective; either of these outcomes may be riskier than biopsy.

CARDIOVASCULAR INFECTIONS

Patients with Hodgkin's disease are prone to persistent infections by *Salmonella*, sometimes (and particularly often in elderly patients) affecting a vascular site. The use of IV catheters deliberately lodged in the right atrium is associated with a high incidence of bacterial endocarditis, presumably related to valve damage followed by bacteremia. Nonbacterial thrombotic endocarditis (marantic endocarditis) has been described in association with a variety of malignancies (most often solid tumors) and may follow bone marrow transplantation as well. The presentation of an embolic event with a new cardiac murmur suggests this diagnosis. Blood cultures are negative in this disease of unknown pathogenesis.

ENDOCRINE SYNDROMES

Infections of the endocrine system have been described in immunocompromised patients. *Candida* infection of the thyroid may be difficult to diagnose during the neutropenic period. It can be defined by indium-labeled WBC scans or gallium scans after neutrophil counts increase. CMV infection can cause adrenalitis with or without resulting adrenal insufficiency. The presentation of a sudden endocrine anomaly in an immunocompromised patient can be a sign of infection in the involved end organ.

MUSCULOSKELETAL INFECTIONS

Infection that is a consequence of vascular compromise, resulting in gangrene, can occur when a tumor restricts the blood supply to muscles, bones, or joints. The process of diagnosis and treatment of such infection is similar to that in normal hosts, with the following caveats:

1. *In terms of diagnosis*, a lack of physical findings resulting from a lack of granulocytes in the granulocytopenic patient should make the clinician more aggressive in obtaining tissue rather than more willing to rely on physical signs.

2. *In terms of therapy*, aggressive debridement of infected tissues may be required. However, it is usually difficult to operate on patients who have recently received chemotherapy, both because of a lack of platelets (which results in bleeding complications) and because of a lack of WBCs (which may lead to secondary infection). A blood culture positive for *Clostridium perfringens*—an organism commonly associated with gas gangrene—can have a number of meanings **(Chap. 149)**. *Clostridium septicum* bacteremia is associated with the presence of an underlying malignancy. Bloodstream infections with intestinal organisms such as *Streptococcus bovis* biotype 1 and *C. perfringens* may arise spontaneously from lower gastrointestinal lesions (tumor or polyps); alternatively, these lesions may be harbingers of invasive disease. The clinical setting must be considered in order to define the appropriate treatment for each case.

RENAL AND URETERAL INFECTIONS

Infections of the urinary tract are common among patients whose ureteral excretion is compromised (Table 70-1). *Candida*, which has a predilection for the kidney, can invade either from the bloodstream or in a retrograde manner (via the ureters or bladder) in immunocompromised patients. The presence of "fungus balls" or persistent candiduria suggests invasive disease. Persistent funguria (with *Aspergillus* as well as *Candida*) should prompt a search for a nidus of infection in the kidney.

Certain viruses are typically seen only in immunosuppressed patients. BK virus (polyomavirus hominis 1) has been documented in the urine of bone marrow transplant recipients and, like adenovirus, may be associated with hemorrhagic cystitis.

ABNORMALITIES THAT PREDISPOSE TO INFECTION

(Table 70-1)

THE LYMPHOID SYSTEM

It is beyond the scope of this chapter to detail how all the immunologic abnormalities that result from cancer or from chemotherapy for cancer lead to infections. Disorders of the immune system are discussed in other sections of this book. As has been noted, patients with antibody deficiency are predisposed to overwhelming infection with encapsulated bacteria (including *S. pneumoniae*, *H. influenzae*, and *N. meningitidis*). Infections that result from the lack of a functional cellular immune system are described in **Chap. 197**. It is worth mentioning, however, that patients undergoing intensive chemotherapy for any form of cancer will have not only defects due to granulocytopenia but also lymphocyte dysfunction, which may be profound. Thus, these patients—especially those receiving glucocorticoid-containing regimens or drugs that inhibit either T cell activation (calcineurin inhibitors or drugs like fludarabine, which affect lymphocyte function) or cytokine induction—should be given prophylaxis for *Pneumocystis* pneumonia.

Patients receiving treatment that eliminates B cells (e.g., with anti-CD20 antibodies or rituximab) are especially vulnerable to intercurrent viral infections. The incidence of progressive multifocal leukoencephalopathy (caused by JC virus) is elevated among these patients.

THE HEMATOPOIETIC SYSTEM

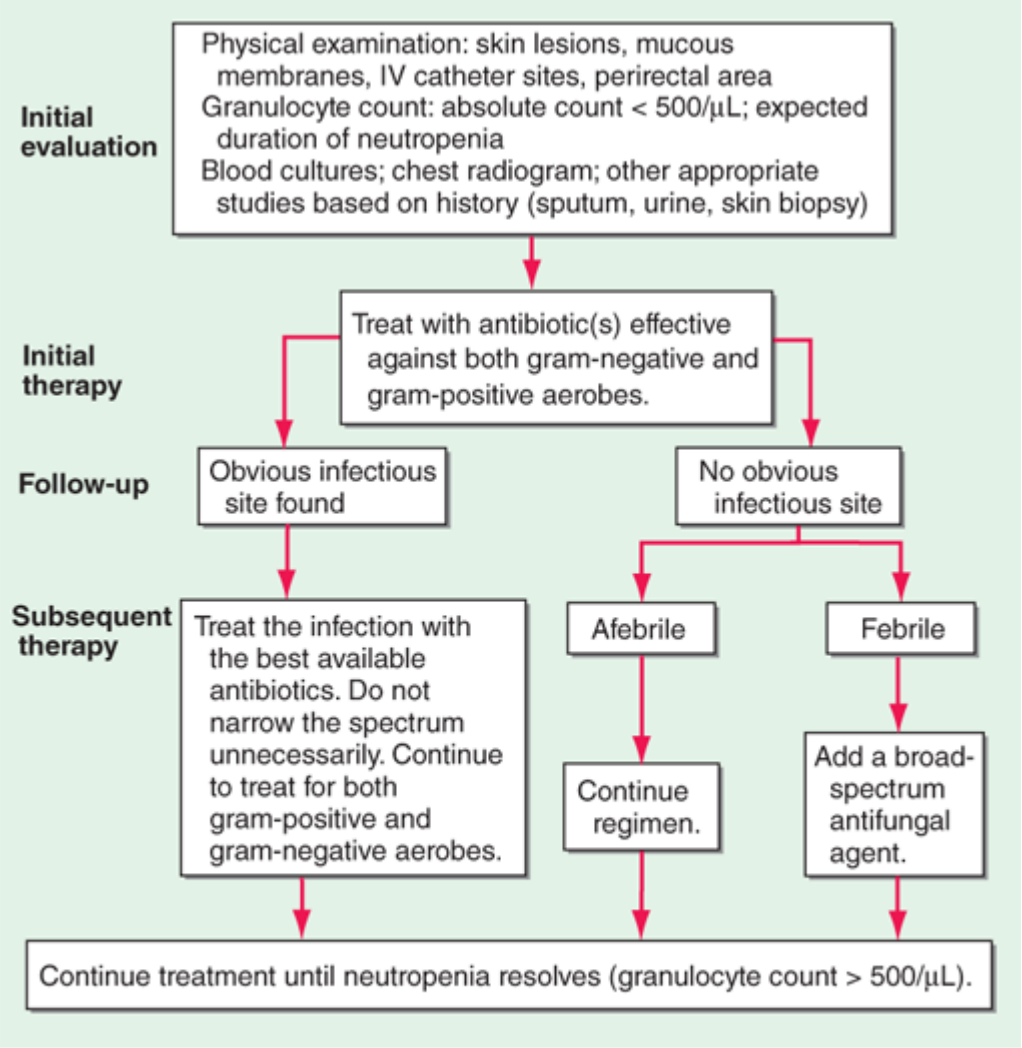

Initial studies in the 1960s revealed a dramatic increase in the incidence of infections (fatal and nonfatal) among cancer patients with a granulocyte count of <500/μL. The use of prophylactic antibacterial agents has reduced the number of bacterial infections, but 35–78% of febrile neutropenic patients being treated for hematologic malignancies develop infections at some time during chemotherapy. Aerobic pathogens (both gram-positive and gram-negative) predominate in all series, but the exact organisms isolated vary from center to center. Infections with anaerobic organisms are uncommon. Geographic patterns affect the types of

fungi isolated. Tuberculosis and malaria are common causes of fever in the developing world and may present in this setting as well.

Neutropenic patients are unusually susceptible to infection with a wide variety of bacteria; thus, antibiotic therapy should be initiated promptly to cover likely pathogens if infection is suspected. Indeed, early initiation of antibacterial agents is mandatory to prevent deaths. Like most immunocompromised patients, neutropenic patients are threatened by their own microbial flora, including gram-positive and gram-negative organisms found commonly on the skin and mucous membranes and in the bowel (Table 70-4). Because treatment with narrow-spectrum agents leads to infection with organisms not covered by the antibiotics used, the initial regimen should target all pathogens likely to be the initial causes of bacterial infection in neutropenic hosts. As noted in the algorithm shown in **Fig. 70-2**, administration of antimicrobial agents is routinely continued until neutropenia resolves— that is, the granulocyte count is sustained above 500/μL for at least 2 days. Fever may not resolve prior to granulocyte recovery. In some cases, patients remain febrile after resolution of neutropenia. In these instances, the risk of sudden death from overwhelming bacteremia is greatly reduced, and the following diagnoses should be seriously considered: (1) fungal infection, (2) bacterial abscesses or undrained foci of infection, and (3) drug fever (including reactions to antimicrobial agents as well as to chemotherapy or cytokines). In the proper setting, viral infection or graft-versus-host disease should be considered. In clinical practice, antibacterial therapy is usually discontinued when the patient is no longer neutropenic and all evidence of bacterial disease has been eliminated. Antifungal agents are then discontinued if there is no evidence of fungal disease. If the patient remains febrile, a search for viral diseases or unusual pathogens is conducted while unnecessary cytokines and other drugs are systematically eliminated from the regimen.

FIGURE 70-2

Algorithm for the diagnosis and treatment of fever and neutropenia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Infections in Cancer Patients

ANTIBACTERIAL THERAPY

Hundreds of antibacterial regimens have been tested for use in patients with cancer. The major risk of infection is related to the degree of neutropenia seen as a consequence of either the disease or the therapy. Many of the relevant studies have involved small populations in which the outcomes have generally been good, and most have lacked the statistical power to detect differences among the regimens studied. Each febrile neutropenic patient should be approached as a unique problem, with particular attention given to previous infections and recent antibiotic exposures. Several general guidelines are useful in the initial treatment of neutropenic patients with fever (Fig. 70-2):

1. In the initial regimen, it is necessary to use antibiotics active against both gram-negative and gram-positive bacteria (Table 70-4).

2. Monotherapy with an aminoglycoside or an antibiotic lacking good activity against gram-positive organisms (e.g., ciprofloxacin or aztreonam) is not adequate in this setting.

3. The agents used should reflect both the epidemiology and the antibiotic resistance pattern of the hospital.

4. If the pattern of resistance justifies its use, a single third-generation cephalosporin constitutes an appropriate initial regimen in many hospitals.

5. Most standard regimens are designed for patients who have not previously received prophylactic antibiotics. The development of fever in a patient who has received antibiotics affects the choice of subsequent therapy, which should target resistant organisms and organisms known to cause infections in patients being treated with the antibiotics already administered.

6. Randomized trials have indicated the safety of oral antibiotic regimens in the treatment of "low-risk" patients with fever and neutropenia. Outpatients who are expected to remain neutropenic for <10 days and who have no concurrent medical problems (such as hypotension, pulmonary compromise, or abdominal pain) can be classified as low risk and treated with a broad-spectrum oral regimen.

7. Several large-scale studies indicate that prophylaxis with a fluoroquinolone (ciprofloxacin or levofloxacin) decreases morbidity and mortality rates among afebrile patients who are anticipated to have neutropenia of long duration.

Commonly used antibiotic regimens for the treatment of febrile patients in whom prolonged neutropenia (>7 days) is anticipated include (1) ceftazidime or cefepime, (2) piperacillin/tazobactam, or (3) imipenem/cilastatin or meropenem. All three regimens have shown equal efficacy in large trials. All three are active against *P. aeruginosa* and a broad spectrum of aerobic gram-positive and gram-negative organisms. Imipenem/cilastatin has been associated with an elevated rate of *C. difficile* diarrhea, and many centers reserve carbapenem antibiotics for treatment of gram-negative bacteria that produce extended-spectrum β-lactamases; these limitations make carbapenems less attractive as an initial regimen. Despite the frequent involvement of coagulase-negative staphylococci, the initial use of vancomycin or its automatic addition to the initial regimen has not resulted in improved outcomes, and the antibiotic does exert toxic effects. For these reasons, only judicious use of vancomycin is recommended—for example, when there is good reason to suspect the involvement of coagulase-negative staphylococci (e.g., the appearance of erythema at the exit site of a catheter or a positive culture for methicillin-resistant *S. aureus* or coagulase-negative staphylococci). Because the sensitivities of bacteria vary from hospital to hospital, clinicians are advised to check their local sensitivities and to be aware that resistance patterns can change quickly, necessitating a change in approach to patients with fever and neutropenia. Similarly, infection control services should monitor for basic antibiotic resistance and for fungal infections. The appearance of a large number of *Aspergillus* infections, in

particular, suggests the possibility of an environmental source that requires further investigation and remediation.

The initial antibacterial regimen should be refined on the basis of culture results (Fig. 70-2). Blood cultures are the most relevant basis for selection of therapy; surface cultures of skin and mucous membranes may be misleading. In the case of gram-positive bacteremia or another gram-positive infection, it is important that the antibiotic be optimal for the organism isolated. Once treatment with broad-spectrum antibiotics has begun, it is not desirable to discontinue all antibiotics because of the risk of failing to treat a potentially fatal bacterial infection; the addition of more and more antibacterial agents to the regimen is not appropriate unless there is a clinical or microbiologic reason to do so. Planned progressive therapy (the serial, empirical addition of one drug after another without culture data) is not efficacious in most settings and may have unfortunate consequences. Simply adding another antibiotic for fear that a gram-negative infection is present is a dubious practice. The synergy exhibited by β-lactams and aminoglycosides against certain gram-negative organisms (especially *P. aeruginosa*) provides the rationale for using two antibiotics in this setting, but recent analyses suggest that efficacy is not enhanced by the addition of aminoglycosides, while toxicity may be increased. Mere "double coverage," with the addition of a quinolone or another antibiotic that is not likely to exhibit synergy, has not been shown to be of benefit and may cause additional toxicities and side effects. Cephalosporins can cause bone marrow suppression, and vancomycin is associated with neutropenia in some healthy individuals. Furthermore, the addition of multiple cephalosporins may induce β-lactamase production by some organisms; cephalosporins and double β-lactam combinations should probably be avoided altogether in *Enterobacter* infections.

ANTIFUNGAL THERAPY

Fungal infections in cancer patients are most often associated with neutropenia. Neutropenic patients are predisposed to the development of invasive fungal infections, most commonly those due to *Candida* and *Aspergillus* species and occasionally those caused by *Mucor, Rhizopus*, *Fusarium*, *Trichosporon*, *Bipolaris*, and others. Cryptococcal infection, which is common among patients taking immunosuppressive agents, is uncommon among neutropenic patients receiving chemotherapy for AML. Invasive candidal disease is usually caused by *C. albicans* or *C. tropicalis* but can be caused by *C. krusei*, *C. parapsilosis*, and *C. glabrata*.

For decades, it has been common clinical practice to add amphotericin B to antibacterial regimens if a neutropenic patient remains febrile despite 4–7 days of treatment with antibacterial agents. The rationale for this empirical addition is that it is difficult to culture fungi before they cause disseminated disease and that mortality rates from disseminated fungal infections in granulocytopenic patients are high. Before the introduction of newer azoles into clinical practice, amphotericin B was the mainstay of antifungal therapy. The insolubility of amphotericin B has resulted in the marketing of several lipid formulations that are less toxic than the amphotericin B deoxycholate complex. Echinocandins (e.g., caspofungin) are useful in the treatment of infections caused by azole-resistant *Candida* strains as well as in therapy for aspergillosis and have been shown to be equivalent to liposomal amphotericin B for the empirical treatment of patients with prolonged fever and neutropenia. Newer azoles have also been demonstrated to be effective in this setting. Although fluconazole is efficacious in the treatment of infections due to many *Candida* species, its use

against serious fungal infections in immunocompromised patients is limited by its narrow spectrum: it has no activity against *Aspergillus* or against several non-*albicans Candida* species. The broad-spectrum azoles (e.g., voriconazole and posaconazole) provide another option for the treatment of *Aspergillus* infections **(Chap. 212)**, including CNS infection. Clinicians should be aware that the spectrum of each azole is somewhat different and that no drug can be assumed to be efficacious against all fungi. *Aspergillus terreus* is resistant to amphotericin B. Although voriconazole is active against *Pseudallescheria boydii*, amphotericin B is not; however, voriconazole has no activity against *Mucor*. Posaconazole, which is administered orally, is useful as a prophylactic agent in patients with prolonged neutropenia. Studies in progress are assessing the use of these agents in combinations. **For a full discussion of antifungal therapy, see Chap. 206**.

ANTIVIRAL THERAPY

The availability of a variety of agents active against herpes-group viruses, including some new agents with a broader spectrum of activity, has heightened focus on the treatment of viral infections, which pose a major problem in cancer patients. Viral diseases caused by the herpes group are prominent. Serious (and sometimes fatal) infections due to HSV and VZV are well documented in patients receiving chemotherapy. CMV may also cause serious disease, but fatalities from CMV infection are more common in hematopoietic stem cell transplant recipients. The roles of human herpesvirus (HHV)-6, HHV-7, and HHV-8 (Kaposi's sarcoma–associated herpesvirus) in cancer patients are still being defined **(Chap. 190)**. EBV lymphoproliferative disease (LPD) can occur in patients receiving chemotherapy but is much more common among transplant recipients **(Chap. 138)**. While clinical experience is most extensive with acyclovir, which can be used therapeutically or prophylactically, a number of derivative drugs offer advantages over this agent **(Chap. 186)**.

In addition to the herpes group, several respiratory viruses (especially RSV) may cause serious disease in cancer patients. Although influenza vaccination is recommended (see below), it may be ineffective in this patient population. The availability of antiviral drugs with activity against influenza viruses gives the clinician additional options for the prophylaxis and treatment of these patients **(Chaps. 186 and 195)**.

OTHER THERAPEUTIC MODALITIES

Another way to address the problems posed by the febrile neutropenic patient is to replenish the neutrophil population. Although granulocyte transfusions may be effective in the treatment of refractory gram-negative bacteremia, they do not have a documented role in prophylaxis. Because of the expense, the risk of leukoagglutinin reactions (which has probably been decreased by improved cell-separation procedures), and the risk of transmission of CMV from unscreened donors (which has been reduced by the use of filters), granulocyte transfusion is reserved for patients whose condition is unresponsive to antibiotics. This modality is efficacious for documented gram-negative bacteremia refractory to antibiotics, particularly in situations where granulocyte numbers will be depressed for only a short period. The demonstrated usefulness of granulocyte colony-stimulating factor in mobilizing neutrophils and advances in preservation techniques may make this option more useful than in the past.

A variety of cytokines, including granulocyte colony-stimulating factor and granulocyte-macrophage colony-stimulating factor, enhance granulocyte recovery after chemotherapy and consequently shorten the period

of maximal vulnerability to fatal infections. The role of these cytokines in routine practice is still a matter of some debate. Most authorities recommend their use only when neutropenia is both severe and prolonged, and they should be used only in the appropriate setting (i.e., when stem cells are likely to be responsive) and not as an adjunct to antimicrobial agents. The cytokines themselves may have adverse effects, including fever, hypoxemia, and pleural effusions or serositis in other areas (Chap. 342).

Once neutropenia has resolved, the risk of infection decreases dramatically. However, depending on what drugs they receive, patients who continue on chemotherapeutic protocols remain at high risk for certain diseases. Any patient receiving more than a maintenance dose of glucocorticoids (e.g., in many treatment regimens for diffuse lymphoma) should also receive prophylactic TMP-SMX because of the risk of *Pneumocystis* infection; those with ALL should receive such prophylaxis for the duration of chemotherapy.

PREVENTION OF INFECTION IN CANCER PATIENTS

EFFECT OF THE ENVIRONMENT

Outbreaks of fatal *Aspergillus* infection have been associated with construction projects and materials in several hospitals. The association between spore counts and risk of infection suggests the need for a high-efficiency air-handling system in hospitals that care for large numbers of neutropenic patients. The use of laminar-flow rooms and prophylactic antibiotics has decreased the number of infectious episodes in severely neutropenic patients. However, because of the expense of such a program and the failure to show that it dramatically affects mortality rates, most centers do not routinely use laminar flow to care for neutropenic patients. Some centers use “reverse isolation,” in which health care providers and visitors to a patient who is neutropenic wear gowns and gloves. Since most of the infections these patients develop are due to organisms that colonize the patients’ own skin and bowel, the validity of such schemes is dubious, and limited clinical data do not support their use. Hand washing by all staff caring for neutropenic patients should be required to prevent the spread of resistant organisms.

The presence of large numbers of bacteria (particularly *P. aeruginosa*) in certain foods, especially fresh vegetables, has led some authorities to recommend a special “low-bacteria” diet. A diet consisting of cooked and canned food is satisfactory to most neutropenic patients and does not involve elaborate disinfection or sterilization protocols. However, there are no studies to support even this type of dietary restriction. Counseling of patients to avoid leftovers, deli foods, undercooked meat, and unpasteurized dairy products is recommended since these foods have been associated with outbreaks of listerial infection.

PHYSICAL MEASURES

Although few studies address this issue, patients with cancer are predisposed to infections resulting from anatomic compromise (e.g., lymphedema resulting from node dissections after radical mastectomy). Surgeons who specialize in cancer surgery can provide specific guidelines for the care of such patients, and patients benefit from common-sense advice about how to prevent infections in vulnerable areas.

IMMUNOGLOBULIN REPLACEMENT

Many patients with multiple myeloma or CLL have immunoglobulin deficiencies as a result of their disease, and all allogeneic bone marrow transplant recipients are hypogammaglobulinemic for a period after transplantation. However, current recommendations reserve intravenous immunoglobulin replacement therapy for those patients with severe, prolonged hypogammaglobulinemia (<400 mg of total IgG/dL) and a history of repeated infections. Antibiotic prophylaxis has been shown to be cheaper and is efficacious in preventing infections in most CLL patients with hypogammaglobulinemia. Routine use of immunoglobulin replacement is not recommended.

SEXUAL PRACTICES

The use of condoms is recommended for severely immunocompromised patients. Any sexual practice that results in oral exposure to feces is not recommended. Neutropenic patients should be advised to avoid any practice that results in trauma, as even microscopic cuts may result in bacterial invasion and fatal sepsis.

ANTIBIOTIC PROPHYLAXIS

Several studies indicate that the use of oral fluoroquinolones prevents infection and decreases mortality rates among severely neutropenic patients. Prophylaxis for *Pneumocystis* is mandatory for patients with ALL and for all cancer patients receiving glucocorticoid-containing chemotherapy regimens.

VACCINATION OF CANCER PATIENTS

In general, patients undergoing chemotherapy respond less well to vaccines than do normal hosts. Their greater need for vaccines thus leads to a dilemma in their management. Purified proteins and inactivated vaccines are almost never contraindicated and should be given to patients even during chemotherapy. For example, all adults should receive diphtheria–tetanus toxoid boosters at the indicated times as well as seasonal influenza vaccine. However, if possible, vaccination should not be undertaken concurrent with cytotoxic chemotherapy. If patients are expected to be receiving chemotherapy for several months and vaccination is indicated (e.g., influenza vaccination in the fall), the vaccine should be given midcycle—as far apart in time as possible from the antimetabolic agents that will prevent an immune response. The meningococcal and pneumococcal polysaccharide vaccines should be given to patients before splenectomy, if possible. The *H. influenzae* type b conjugate vaccine should be administered to all splenectomized patients.

In general, live virus (or live bacterial) vaccines should not be given to patients during intensive chemotherapy because of the risk of disseminated infection. Recommendations on vaccination are summarized in Table 70-2 (see *https://www.cdc.gov/vaccines/hcp/index.html* for updated recommendations).

FURTHER READING

Klastersky J et al: The MASCC Neutropenia, Infection and Myelosuppression Study Group evaluates recent new concepts for the use of granulocyte colony-stimulating factors for the prevention of febrile neutropenia. Support Care Cancer 21:1793, 2013. [PubMed: 23525960]

Pappas PG et al: Clinical practice guideline for the management of candidiasis: 2016 update by the Infectious Diseases Society of America. Clin Infect Dis 62:e1, 2016. [PubMed: 26679628]

Patterson TF et al: Practice guidelines for the diagnosis and management of aspergillosis: 2016 update by the Infectious Diseases Society of America. Clin Infect Dis 63:e1, 2016. [PubMed: 27365388]

Taur Y, Pamer EG: Microbiome mediation of infections in the cancer setting. Genome Med 8:40, 2016. [PubMed: 27090860]

WEBSITE

Prevention and Treatment of Cancer-Related Infections; National Comprehensive Cancer Network Clinical Practice Guidelines in Oncology Version 2.2016 (*https://www.nccn.org*)

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 71: Oncologic Emergencies

Rasim Gucalp; Janice P. Dutcher

INTRODUCTION

Emergencies in patients with cancer may be classified into three groups: pressure or obstruction caused by a space-occupying lesion, metabolic or hormonal problems (paraneoplastic syndromes, Chap. 89), and treatment-related complications.

STRUCTURAL-OBSTRUCTIVE ONCOLOGIC EMERGENCIES

SUPERIOR VENA CAVA SYNDROME

Superior vena cava syndrome (SVCS) is the clinical manifestation of superior vena cava (SVC) obstruction, with severe reduction in venous return from the head, neck, and upper extremities. Malignant tumors, such as lung cancer, lymphoma, and metastatic tumors, are responsible for the majority of SVCS cases. With the expanding use of intravascular devices (e.g., permanent central venous access catheters, pacemaker/defibrillator leads), the prevalence of benign causes of SVCS is increasing now, accounting for at least 40% of cases. Lung cancer, particularly of small-cell and squamous cell histologies, accounts for ~85% of all cases of malignant origin. In young adults, malignant lymphoma is a leading cause of SVCS. Hodgkin's lymphoma involves the mediastinum more commonly than other lymphomas but rarely causes SVCS. When SVCS is noted in a young man with a mediastinal mass, the differential diagnosis is lymphoma versus primary mediastinal germ cell tumor. Metastatic cancers to the mediastinal lymph nodes, such as testicular and breast carcinomas, account for a small proportion of cases. Other causes include benign tumors, aortic aneurysm, thyromegaly, thrombosis, and fibrosing mediastinitis from prior irradiation, histoplasmosis, or Behçet's syndrome. SVCS as the initial manifestation of Behçet's syndrome may be due to inflammation of the SVC associated with thrombosis.

Patients with SVCS usually present with neck and facial swelling (especially around the eyes), dyspnea, and cough. Other symptoms include hoarseness, tongue swelling, headaches, nasal congestion, epistaxis, hemoptysis, dysphagia, pain, dizziness, syncope, and lethargy. Bending forward or lying down may aggravate the symptoms. The characteristic physical findings are dilated neck veins; an increased number of collateral veins covering the anterior chest wall; cyanosis; and edema of the face, arms, and chest. Facial swelling and plethora are typically exacerbated when the patient is supine. More severe cases include proptosis, glossal and laryngeal edema, and obtundation. The clinical picture is milder if the obstruction is located above the azygos vein. Symptoms are usually progressive, but in some cases, they may improve as collateral circulation develops.

Signs and symptoms of cerebral and/or laryngeal edema, though rare, are associated with a poorer prognosis and require urgent evaluation. Seizures are more likely related to brain metastases than to cerebral edema from venous occlusion. Patients with small-cell lung cancer and SVCS have a higher incidence of brain metastases than those without SVCS.

Cardiorespiratory symptoms at rest, particularly with positional changes, suggest significant airway and vascular obstruction and limited physiologic reserve. Cardiac arrest or respiratory failure can occur, particularly in patients receiving sedatives or undergoing general anesthesia.

Rarely, esophageal varices may develop, particularly in the setting of SVC syndrome due to hemodialysis catheter. These are "downhill" varices based on the direction of blood flow from cephalad to caudad (in contrast to "uphill" varices associated with caudad to cephalad flow from portal hypertension). If the obstruction to the SVC is proximal to the azygous vein, varices develop in the upper one-third of the esophagus. If the obstruction involves or is distal to the azygous vein, varices occur in the entire length of the esophagus. Variceal bleeding may be a late complication of chronic SVCS.

SVC obstruction may lead to bilateral breast edema with bilateral enlarged breast. Unilateral breast dilation may be seen as a consequence of axillary or subclavian vein blockage.

The diagnosis of SVCS is a clinical one. The most significant chest radiographic finding is widening of the superior mediastinum, most commonly on the right side. Pleural effusion occurs in only 25% of patients, often on the right side. The majority of these effusions are exudative and occasionally chylous. However, a normal chest radiograph is still compatible with the diagnosis if other characteristic findings are present. Computed tomography (CT) provides the most reliable view of the mediastinal anatomy. The diagnosis of SVCS requires diminished or absent opacification of central venous structures with prominent collateral venous circulation. Magnetic resonance imaging (MRI) is increasingly being used to diagnose SVC obstruction with a 100% sensitivity and specificity, but dyspneic SVCS patients may have difficulty remaining supine for the entire imaging process. Invasive procedures, including bronchoscopy, percutaneous needle biopsy, mediastinoscopy, and even thoracotomy, can be performed by a skilled clinician without any major risk of bleeding. Endobronchial or esophageal ultrasound-guided needle aspiration may establish the diagnosis safely. For patients with a known cancer, a detailed workup usually is not necessary, and appropriate treatment may be started after obtaining a CT scan of the thorax. For those with no history of malignancy, a detailed evaluation is essential to rule out benign causes and determine a specific diagnosis to direct the appropriate therapy.

TREATMENT

TREATMENT

Superior Vena Cava Syndrome

The one potentially life-threatening complication of a superior mediastinal mass is tracheal obstruction. Upper airway obstruction demands emergent therapy. Diuretics with a low-salt diet, head elevation, and oxygen may produce temporary symptomatic relief. Glucocorticoids have a limited role except in the setting of mediastinal lymphoma masses.

Radiation therapy is the primary treatment for SVCS caused by non-small-cell lung cancer and other metastatic solid tumors. Chemotherapy is effective when the underlying cancer is small-cell carcinoma of the lung, lymphoma, or germ cell tumor. SVCS recurs in 10–30% of patients; it may be palliated with the use of intravascular self-expanding stents **(Fig. 71-1)**. Early stenting may be necessary in patients with severe symptoms; however, the prompt increase in venous return after stenting may precipitate heart failure and pulmonary edema. Other complications of stent placement include hematoma at the insertion site, SVC perforation, stent migration in the right ventricle, stent fracture, and pulmonary embolism.

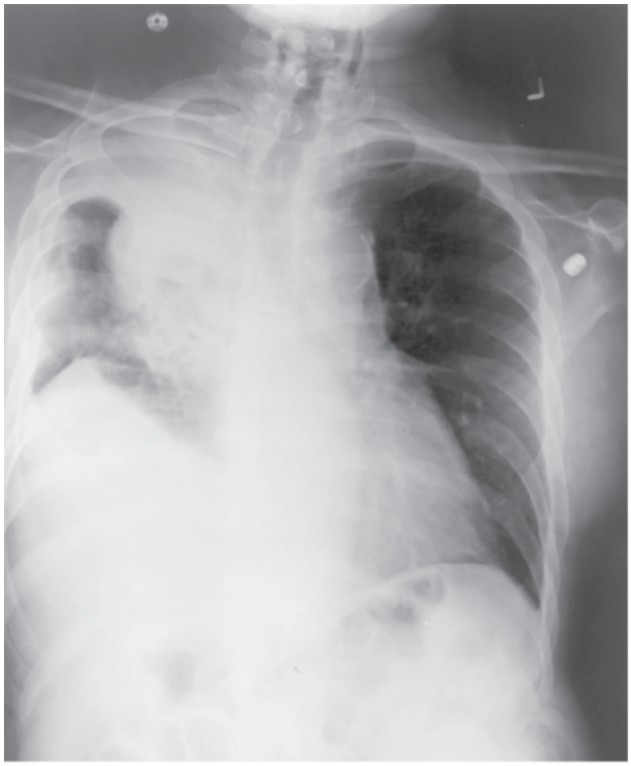
Clinical improvement occurs in most patients, although this improvement may be due to the development of adequate collateral circulation. The mortality associated with SVCS does not relate to caval obstruction but rather to the underlying cause.

SVCS AND CENTRAL VENOUS CATHETERS IN ADULTS

The use of long-term central venous catheters has become common practice in patients with cancer. Major vessel thrombosis may occur. In these cases, catheter removal should be combined with anticoagulation to prevent embolization. SVCS in this setting, if detected early, can be treated by fibrinolytic therapy without sacrificing the catheter. When managing patients with transvenous lead-related SVC syndrome, anticoagulation, local and systemic thrombolytic therapy, and surgical intervention can be effective therapy in select patients. Endovascular stenting has also been shown to be safe and promising, with minimal procedural or clinical complications. The role of anticoagulation after SVC stent placement is controversial.

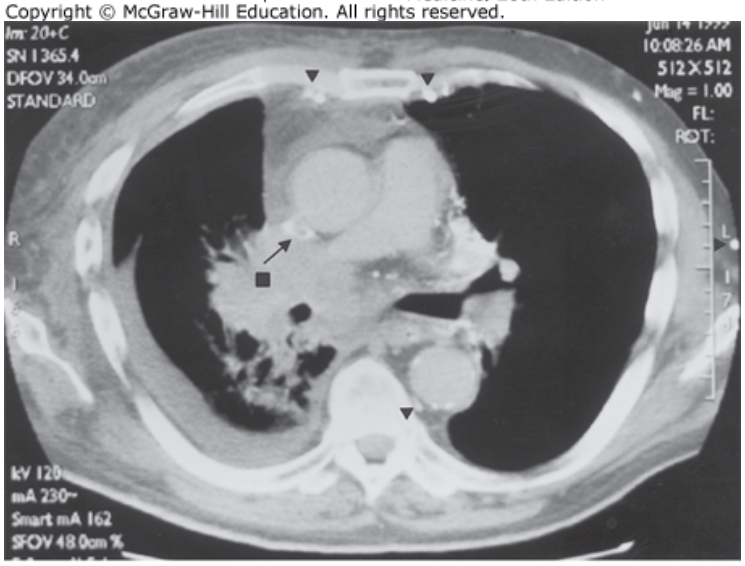
FIGURE 71-1

Superior vena cava syndrome (SVCS). ***A.*** Chest radiographs of a 59-year-old man with recurrent SVCS caused by non-small-cell lung cancer showing right paratracheal mass with right pleural effusion. ***B.*** Computed tomography of same patient demonstrating obstruction of the superior vena cava with thrombosis (*arrow*) by the lung cancer (*square*) and collaterals (*arrowheads*). ***C.*** Balloon angioplasty (*arrowhead*) with Wallstent (*arrow*) in same patient.



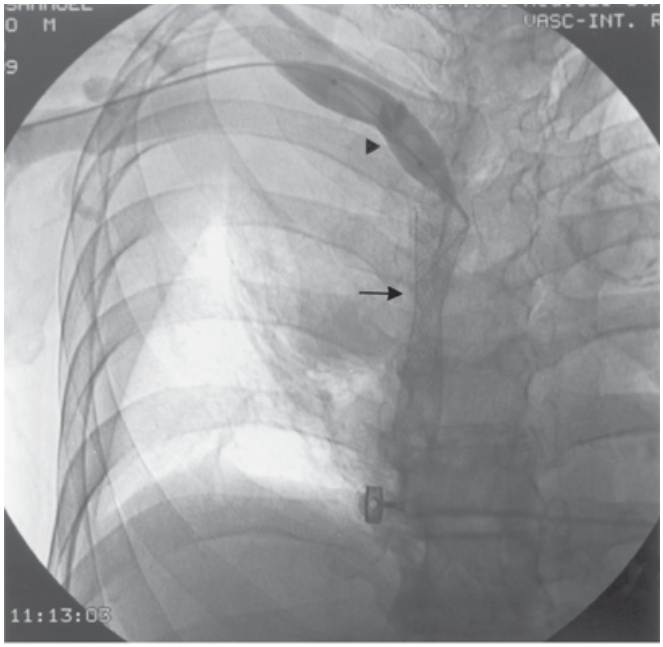
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PERICARDIAL EFFUSION/TAMPONADE

Malignant pericardial disease is found at autopsy in 5–10% of patients with cancer, most frequently with lung cancer, breast cancer, leukemias, and lymphomas. Cardiac tamponade as the initial presentation of extrathoracic malignancy is rare. The origin is not malignancy in ~50% of cancer patients with symptomatic pericardial disease, but it can be related to irradiation, drug-induced pericarditis, including chemotherapeutic agents such all-trans retinoic acid, arsenic trioxide, imatinib and other abl kinase inhibitors, hypothyroidism, idiopathic pericarditis, infection, or autoimmune diseases. Two types of radiation pericarditis occur: an acute inflammatory, effusive pericarditis occurring within months of irradiation, which usually resolves spontaneously, and a chronic effusive pericarditis that may appear up to 20 years after radiation therapy and is accompanied by a thickened pericardium.

Most patients with pericardial metastasis are asymptomatic. However, the common symptoms are dyspnea, cough, chest pain, orthopnea, and weakness. Pleural effusion, sinus tachycardia, jugular venous distention, hepatomegaly, peripheral edema, and cyanosis are the most frequent physical findings. Relatively specific diagnostic findings, such as paradoxical pulse, diminished heart sounds, pulsus alternans (pulse waves alternating between those of greater and lesser amplitude with successive beats), and friction rub are less common than with nonmalignant pericardial disease. Chest radiographs and electrocardiogram (ECG) reveal abnormalities in 90% of patients, but half of these abnormalities are nonspecific. Echocardiography is the most helpful diagnostic test. Pericardial fluid may be serous, serosanguineous, or hemorrhagic, and cytologic examination of pericardial fluid is diagnostic in most patients. Measurements of tumor markers in the pericardial fluid are not helpful in the diagnosis of malignant pericardial fluid. Pericardioscopy with targeted pericardial and epicardial biopsy may differentiate neoplastic and benign pericardial disease. A combination of cytology, pericardial and epicardial biopsy, and guided pericardioscopy gives the best diagnostic yield. CT scan of chest may also reveal the presence of a concomitant thoracic neoplasm. Cancer patients with pericardial effusion containing malignant cells on cytology have a very poor survival, ~7 weeks.

TREATMENT

TREATMENT

Pericardial Effusion/Tamponade

Pericardiocentesis with or without the introduction of sclerosing agents, the creation of a pericardial window, complete pericardial stripping, cardiac irradiation, or systemic chemotherapy are effective treatments. Acute pericardial tamponade with life-threatening

hemodynamic instability requires immediate drainage of fluid. This can be quickly achieved by pericardiocentesis. The recurrence rate after percutaneous catheter drainage is ~20%. Sclerotherapy (pericardial instillation of bleomycin, mitomycin C, or tetracycline) may decrease recurrences. Alternatively, subxiphoid pericardiotomy can be performed in 45 min under local anesthesia. Thoracoscopic pericardial fenestration can be employed for benign causes; however, 60% of malignant pericardial effusions recur after this procedure. In a subset of patients, drainage of the pericardial effusion is paradoxically followed by worsening hemodynamic instability. This so-called "postoperative low cardiac output syndrome" occurs in up to 10% of patients undergoing surgical drainage and carries poor short-term survival.

INTESTINAL OBSTRUCTION

Intestinal obstruction and reobstruction are common problems in patients with advanced cancer, particularly colorectal or ovarian carcinoma. However, other cancers, such as lung or breast cancer and melanoma, can metastasize within the abdomen, leading to intestinal obstruction. Metastatic disease from colorectal, ovarian, pancreatic, gastric, and occasionally breast cancer can lead to peritoneal carcinomatosis, with infiltration of the omentum and peritoneal surface, thus limiting bowel motility. Typically, obstruction occurs at multiple sites in peritoneal carcinomatosis. Melanoma has a predilection to involve the small bowel; this involvement may be isolated, and resection may result in prolonged survival. Intestinal pseudoobstruction is caused by infiltration of the mesentery or bowel muscle by tumor, involvement of the celiac plexus, or paraneoplastic neuropathy in patients with small-cell lung cancer. Paraneoplastic neuropathy is associated with IgG antibodies reactive to neurons of the myenteric and submucosal plexuses of the jejunum and stomach. Ovarian cancer can lead to authentic luminal obstruction or to pseudoobstruction that results when circumferential invasion of a bowel segment arrests the forward progression of peristaltic contractions.

The onset of obstruction is usually insidious. Pain is the most common symptom and is usually colicky in nature. Pain can also be due to abdominal distention, tumor masses, or hepatomegaly. Vomiting can be intermittent or continuous. Patients with complete obstruction usually have constipation. Physical examination may reveal abdominal distention with tympany, ascites, visible peristalsis, high-pitched bowel sounds, and tumor masses. Erect plain abdominal films may reveal multiple air-fluid levels and dilation of the small or large bowel. Acute cecal dilation to >12–14 cm is considered a surgical emergency because of the high likelihood of rupture. CT scan is useful in defining the extent of disease and the exact nature of the obstruction and differentiating benign from malignant causes of obstruction in patients who have undergone surgery for malignancy. Malignant obstruction is suggested by a mass at the site of obstruction or prior surgery, adenopathy, or an abrupt transition zone and irregular bowel thickening at the obstruction site. Benign obstruction is more likely when CT shows mesenteric vascular changes, a large volume of ascites, or a smooth transition zone and smooth bowel thickening at the obstruction site. In challenging patients with obstructive symptoms, particularly low-grade small-bowel obstruction (SBO), CT enteroclysis often can help establish the diagnosis by providing distention of small-bowel loops. In this technique, water-soluble contrast is infused through a nasoenteric tube into the duodenum or proximal small bowel followed by CT images. The prognosis for the patient with cancer who develops intestinal obstruction is poor; median survival is 3–4 months. About 25–30% of patients are found to have intestinal obstruction due to causes other than cancer. Adhesions from previous operations are a common benign cause. Ileus induced by vinca alkaloids, narcotics, or other drugs is another reversible cause.

TREATMENT

TREATMENT

Intestinal Obstruction

The management of intestinal obstruction in patients with advanced malignancy depends on the extent of the underlying malignancy, options for further antineoplastic therapy, estimated life expectancy, the functional status of the major organs, and the extent of the obstruction. The initial management should include surgical evaluation. Operation is not always successful and may lead to further complications with a substantial mortality rate (10–20%). Laparoscopy can diagnose and treat malignant bowel obstruction in some cases. Self-expanding metal stents placed in the gastric outlet, duodenum, proximal jejunum, colon, or rectum may palliate obstructive symptoms at those sites without major surgery. Patients known to have advanced intraabdominal malignancy should receive a prolonged course of conservative management, including nasogastric decompression. Percutaneous endoscopic or surgical gastrostomy tube placement is an option for palliation of nausea and vomiting, the so-called "venting gastrostomy." Treatment with antiemetics, antispasmodics, and analgesics may allow patients to remain outside the hospital. Octreotide may relieve obstructive

symptoms through its inhibitory effect on gastrointestinal secretion. Glucocorticoids have anti-inflammatory effects and may help the resolution of bowel obstruction. They also have antiemetic effects.

URINARY OBSTRUCTION

Urinary obstruction may occur in patients with prostatic or gynecologic malignancies, particularly cervical carcinoma; metastatic disease from other primary sites such as carcinomas of the breast, stomach, lung, colon, and pancreas; or lymphomas. Radiation therapy to pelvic tumors may cause fibrosis and subsequent ureteral obstruction. Bladder outlet obstruction is usually due to prostate and cervical cancers and may lead to bilateral hydronephrosis and renal failure.

Flank pain is the most common symptom. Persistent urinary tract infection, persistent proteinuria, or hematuria in patients with cancer should raise suspicion of ureteral obstruction. Total anuria and/or anuria alternating with polyuria may occur. A slow, continuous rise in the serum creatinine level necessitates immediate evaluation. Renal ultrasound is the safest and cheapest way to identify hydronephrosis. The function of an obstructed kidney can be evaluated by a nuclear scan. CT scan can reveal the point of obstruction and identify a retroperitoneal mass or adenopathy.

TREATMENT

TREATMENT

Urinary Obstruction

Obstruction associated with flank pain, sepsis, or fistula formation is an indication for immediate palliative urinary diversion. Internal ureteral stents can be placed under local anesthesia. Percutaneous nephrostomy offers an alternative approach for drainage. The placement of a nephrostomy is associated with a significant rate of pyelonephritis. In the case of bladder outlet obstruction due to malignancy, a suprapubic cystostomy can be used for urinary drainage. An aggressive intervention with invasive approaches to improve the obstruction should be weighed against the likelihood of antitumor response, and the ability to reverse renal insufficiency should be evaluated.

MALIGNANT BILIARY OBSTRUCTION

This common clinical problem can be caused by a primary carcinoma arising in the pancreas, ampulla of Vater, bile duct, or liver or by metastatic disease to the periductal lymph nodes or liver parenchyma. The most common metastatic tumors causing biliary obstruction are gastric, colon, breast, and lung cancers. Jaundice, light-colored stools, dark urine, pruritus, and weight loss due to malabsorption are usual symptoms. Pain and secondary infection are uncommon in malignant biliary obstruction. Ultrasound, CT scan, or percutaneous transhepatic or endoscopic retrograde cholangiography will identify the site and nature of the biliary obstruction.

TREATMENT

TREATMENT

Malignant Biliary Obstruction

Palliative intervention is indicated only in patients with disabling pruritus resistant to medical treatment, severe malabsorption, or infection. Stenting under radiographic control, surgical bypass, or radiation therapy with or without chemotherapy may alleviate the obstruction. The choice of therapy should be based on the site of obstruction (proximal vs distal), the type of tumor (sensitive to radiotherapy, chemotherapy, or neither), and the general condition of the patient. Stenting under radiographic or endoscopic control, surgical bypass, or radiation therapy with or without chemotherapy may alleviate the obstruction. Photodynamic therapy and radiofrequency ablation are promising endoscopic therapies for malignant biliary obstruction.

SPINAL CORD COMPRESSION

Malignant spinal cord compression (MSCC) is defined as compression of the spinal cord and/or cauda equina by an extradural tumor mass. The minimum radiologic evidence for cord compression is indentation of the theca at the level of clinical features. Spinal cord

compression occurs in 5–10% of patients with cancer. Epidural tumor is the first manifestation of malignancy in ~10% of patients. The underlying cancer is usually identified during the initial evaluation; lung cancer is the most common cause of MSCC.

Metastatic tumor involves the vertebral column more often than any other part of the bony skeleton. Lung, breast, and prostate cancers are the most frequent offenders. Multiple myeloma also has a high incidence of spine involvement. Lymphomas, melanoma, renal cell cancer, and genitourinary cancers also cause cord compression. The thoracic spine is the most common site (70%), followed by the lumbosacral spine (20%) and the cervical spine (10%). Involvement of multiple sites is most frequent in patients with breast and prostate carcinoma. Cord injury develops when metastases to the vertebral body or pedicle enlarge and compress the underlying dura. Another cause of cord compression is direct extension of a paravertebral lesion through the intervertebral foramen. These cases usually involve a lymphoma, myeloma, or pediatric neoplasm. Parenchymal spinal cord metastasis due to hematogenous spread is rare. Intramedullary metastases can be seen in lung cancer, breast cancer, renal cancer, melanoma, and lymphoma, and are frequently associated with brain metastases and leptomeningeal disease.

Expanding extradural tumors induce injury through several mechanisms. Expanding extradural tumors induce mechanical injury to axons and myelin. Compression compromises blood flow, leading to ischemia and/or infarction.

The most common initial symptom in patients with spinal cord compression is localized back pain and tenderness due to involvement of vertebrae by tumor. Pain is usually present for days or months before other neurologic findings appear. It is exacerbated by movement and by coughing or sneezing. It can be differentiated from the pain of disk disease by the fact that it worsens when the patient is supine. Radicular pain is less common than localized back pain and usually develops later. Radicular pain in the cervical or lumbosacral areas may be unilateral or bilateral. Radicular pain from the thoracic roots is often bilateral and is described by patients as a feeling of tight, band-like constriction around the thorax and abdomen. Typical cervical radicular pain radiates down the arm; in the lumbar region, the radiation is down the legs. *Lhermitte's sign*, a tingling or electric sensation down the back and upper and lower limbs upon flexing or extending the neck, may be an early sign of cord compression. Loss of bowel or bladder control may be the presenting symptom but usually occurs late in the course. Occasionally patients present with ataxia of gait without motor and sensory involvement due to involvement of the spinocerebellar tract.

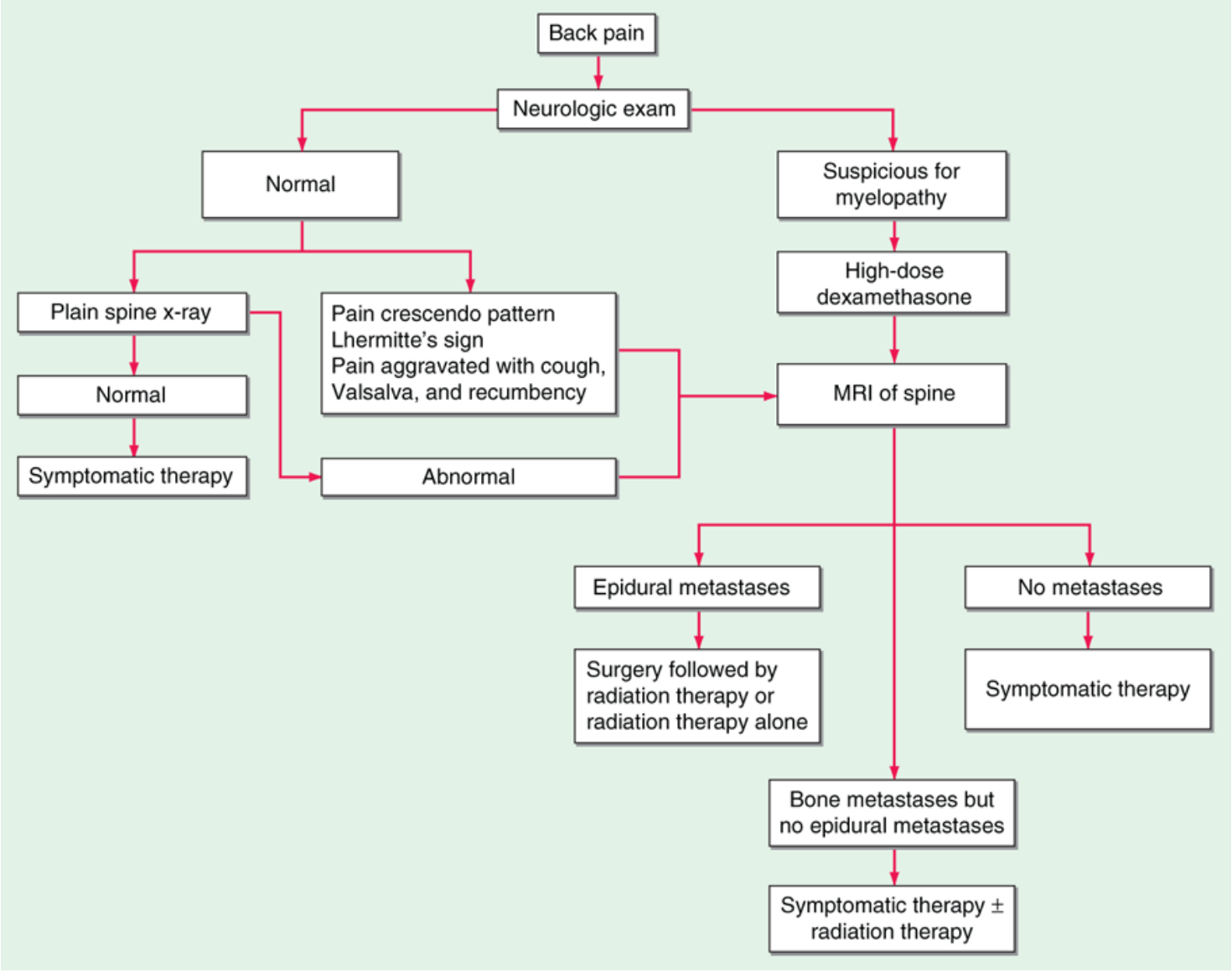
On physical examination, pain induced by straight leg raising, neck flexion, or vertebral percussion may help to determine the level of cord compression. Patients develop numbness and paresthesias in the extremities or trunk. Loss of sensibility to pinprick is as common as loss of sensibility to vibration or position. The upper limit of the zone of sensory loss is often one or two vertebrae below the site of compression. Motor findings include weakness, spasticity, and abnormal muscle stretching. An extensor plantar reflex reflects significant compression. Deep tendon reflexes may be brisk. Motor and sensory loss usually precedes sphincter disturbance. Patients with autonomic dysfunction may present with decreased anal tonus, decreased perineal sensibility, and a distended bladder. The absence of the anal wink reflex or the bulbocavernosus reflex confirms cord involvement. In doubtful cases, evaluation of postvoiding urinary residual volume can be helpful. A residual volume of >150 mL suggests bladder dysfunction. Autonomic dysfunction is an unfavorable prognostic factor. Patients with progressive neurologic symptoms should have frequent neurologic examinations and rapid therapeutic intervention. Other illnesses that may mimic cord compression include osteoporotic vertebral collapse, disk disease, pyogenic abscess or vertebral tuberculosis, radiation myelopathy, neoplastic leptomeningitis, benign tumors, epidural hematoma, and spinal lipomatosis.

Cauda equina syndrome is characterized by low back pain; diminished sensation over the buttocks, posterior-superior thighs, and perineal area in a saddle distribution; rectal and bladder dysfunction; sexual impotence; absent bulbocavernous, patellar, and Achilles' reflexes; and variable amount of lower-extremity weakness. This reflects compression of nerve roots as they form the cauda equina after leaving the spinal cord. The majority of cauda equine tumors are primary tumors of glial or nerve sheath origin; metastases are very rare.

Patients with cancer who develop back pain should be evaluated for spinal cord compression as quickly as possible **(Fig. 71-2)**. Treatment is more often successful in patients who are ambulatory and still have sphincter control at the time treatment is initiated. Patients should have a neurologic examination and plain films of the spine. Those whose physical examination suggests cord compression should receive dexamethasone starting immediately.

FIGURE 71-2

Management of cancer patients with back pain.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Erosion of the pedicles (the "winking owl" sign) is the earliest radiologic finding of vertebral tumor. Other radiographic changes include increased intrapedicular distance, vertebral destruction, lytic or sclerotic lesions, scalloped vertebral bodies, and vertebral body collapse. Vertebral collapse is not a reliable indicator of the presence of tumor; ~20% of cases of vertebral collapse, particularly those in older patients and postmenopausal women, are due not to cancer but to osteoporosis. Also, a normal appearance on plain films of the spine does not exclude the diagnosis of cancer. The role of bone scans in the detection of cord compression is not clear; this method is sensitive but less specific than spinal radiography.

The full-length image of the cord provided by MRI is the imaging procedure of choice. Multiple epidural metastases are noted in 25% of patients with cord compression, and their presence influences treatment plans. On T1-weighted images, good contrast is noted between the cord, cerebrospinal fluid (CSF), and extradural lesions. Owing to its sensitivity in demonstrating the replacement of bone marrow by tumor, MRI can show which parts of a vertebra are involved by tumor. MRI also visualizes intraspinal extradural masses compressing the cord. T2-weighted images are most useful for the demonstration of intramedullary pathology. Gadolinium-enhanced MRI can help to delineate intramedullary disease. MRI is as good as or better than myelography plus postmyelogram CT scan in detecting metastatic epidural disease with cord compression. Myelography should be reserved for patients who have poor MRIs or who cannot undergo MRI promptly. CT scan in conjunction with myelography enhances the detection of small areas of spinal destruction.

In patients with cord compression and an unknown primary tumor, a simple workup including chest radiography, mammography, measurement of prostate-specific antigen, and abdominal CT usually reveals the underlying malignancy.

TREATMENT

TREATMENT

Spinal Cord Compression

The treatment of patients with spinal cord compression is aimed at relief of pain and restoration/preservation of neurologic function (Fig. 71-2). Management of MSCC requires a multidisciplinary approach.

Radiation therapy plus glucocorticoids is generally the initial treatment of choice for most patients with spinal cord compression. The management decision of SCC involves assessment of neurologic (N), oncologic (O), mechanical (M), and systemic factors (S). NOMS was developed by Memorial Sloan Kettering Cancer Center (MSKCC) researchers to provide an algorithm for management of SCC. The neurologic assessment is based on the degree of epidural SCC, myelopathy, and/or functional radiculopathy. Oncologic assessment involves the radio-sensitivity of the tumor type. In patients with radio-resistant tumors, stereotactic body radiotherapy (SRS) is the preferred approach if radiation is appropriate. Safe delivery of SRS requires a 2- to 3-mm margin away from the spinal cord. Separation surgery followed by SRS is necessary in patients with high-grade SCC due to radio-resistant tumors. In patients with mechanical instability or retropulsion of bone fragments into the spinal canal or cord, a surgical approach is the treatment of choice. Systemic factors that need to be considered are the extent of disease and medical comorbidities that determine the patient's ability to tolerate planned therapy. Chemotherapy may have a role in patients with chemosensitive tumors who have had prior radiotherapy to the same region and who are not candidates for surgery. Patients who previously received radiotherapy for MSCC with an in-field tumor progression can be treated with reirradiation if they are not surgical candidates.

Patients with painful pathologic compression fractures without spinal instability may benefit from percutaneous vertebroplasty or kyphoplasty, the injection of acrylic cement into a collapsed vertebra to stabilize the fracture. Pain palliation is common, and local antitumor effects have been noted. Cement leakage may cause symptoms in ~10% of patients. Bisphosphonates and/or denosumab may be helpful in prevention of SCC in patients with bony involvement.

The histology of the tumor is an important determinant of both recovery and survival. Rapid onset and progression of signs and symptoms are poor prognostic features.

INCREASED INTRACRANIAL PRESSURE

About 25% of patients with cancer die with intracranial metastases. The cancers that most often metastasize to the brain are lung and breast cancers and melanoma. Brain metastases often occur in the presence of systemic disease, and they frequently cause major symptoms, disability, and early death. The initial presentation of brain metastases from a previously unknown primary cancer is common. Lung cancer is most commonly the primary malignancy. CT scans of the chest/abdomen and MRI of the brain as the initial diagnostic studies can identify a biopsy site in most patients.

The signs and symptoms of a metastatic brain tumor are similar to those of other intracranial expanding lesions: headache, nausea, vomiting, behavioral changes, seizures, and focal, progressive neurologic changes. Occasionally the onset is abrupt, resembling a stroke, with the sudden appearance of headache, nausea, vomiting, and neurologic deficits. This picture is usually due to hemorrhage into the metastasis. Melanoma, germ cell tumors, and renal cell cancers have a particularly high incidence of intracranial bleeding. The tumor mass and surrounding edema may cause obstruction of the circulation of CSF, with resulting hydrocephalus. Patients with increased intracranial pressure may have papilledema with visual disturbances and neck stiffness. As the mass enlarges, brain tissue may be displaced through the fixed cranial openings, producing various herniation syndromes.

MRI is superior to CT scan. Gadolinium-enhanced MRI is more sensitive than CT at revealing meningeal involvement and small lesions, particularly in the brainstem or cerebellum. The MRI of the brain shows brain metastases as multiple enhancing lesions of various sizes with surrounding areas of low-density edema.

Intracranial hypertension ("pseudotumor cerebri") secondary to tretinoin therapy for acute promyelocytic leukemia has been reported, as another cause of intracranial pressure in the setting of a malignancy.

TREATMENT

TREATMENT

Increased Intracranial Pressure

Dexamethasone is the best initial treatment for all symptomatic patients with brain metastases. Patients with multiple lesions should usually receive whole-brain radiation. Patients with a single-brain metastasis and with controlled extracranial disease may be treated with surgical excision followed by whole-brain radiation therapy, especially if they are aged <60 years. Radioresistant tumors should be resected if possible. Stereotactic radiosurgery (SRS) is recommended in patients with a limited number of brain metastases (one to four) who have stable, systemic disease or reasonable systemic treatment options and in patients who have a small number of metastatic lesions in whom whole-brain radiation therapy has failed. With a gamma knife or linear accelerator, multiple small, well-collimated beams of ionizing radiation destroy lesions seen on MRI. Some patients with increased intracranial pressure associated with hydrocephalus may benefit from shunt placement. If neurologic deterioration is not reversed with medical therapy, ventriculotomy to remove CSF or craniotomy to remove tumors or hematomas may be necessary.

Targeted agents and checkpoint inhibitors have significant activity in brain metastases from non-small-cell lung cancer, breast cancer, renal cancer, and melanoma.

NEOPLASTIC MENINGITIS

Tumor involving the leptomeninges is a complication of both primary central nervous system (CNS) tumors and tumors that metastasize to the CNS. The incidence is estimated at 3–8% of patients with cancer. Melanoma, breast and lung cancer, lymphoma (including AIDS-associated), and acute leukemia are the most common causes. Synchronous intraparenchymal brain metastases are evident in 11–31% of patients with neoplastic meningitis. Leptomeningeal seeding is frequent in patients undergoing resection of brain metastases or receiving stereotactic radiotherapy for brain metastases.

Patients typically present with multifocal neurologic signs and symptoms, including headache, gait abnormality, mental changes, nausea, vomiting, seizures, back or radicular pain, and limb weakness. Signs include cranial nerve palsies, extremity weakness, paresthesia, and decreased deep tendon reflexes.

Diagnosis is made by demonstrating malignant cells in the CSF; however, up to 40% of patients may have false-negative CSF cytology. An elevated CSF protein level is nearly always present (except in HTLV-1–associated adult T-cell leukemia). Patients with neurologic signs and symptoms consistent with neoplastic meningitis who have a negative CSF cytology should have the spinal tap repeated at least one more time for cytologic examination. MRI findings suggestive of neoplastic meningitis include leptomeningeal, subependymal, dural, or cranial nerve enhancement; superficial cerebral lesions; intradural nodules; and communicating hydrocephalus. Spinal cord imaging by MRI is a necessary component of the evaluation of nonleukemia neoplastic meningitis because ~20% of patients have cord abnormalities, including intradural enhancing nodules that are diagnostic for leptomeningeal involvement. Cauda equina lesions are common, but lesions may be seen anywhere in the spinal canal. The value of MRI for the diagnosis of leptomeningeal disease is limited in patients with hematopoietic malignancy. Radiolabeled CSF flow studies are abnormal in up to 70% of patients with neoplastic meningitis; ventricular outlet obstruction, abnormal flow in the spinal canal, or impaired flow over the cerebral convexities may affect distribution of intrathecal chemotherapy, resulting in decreased efficacy or increased toxicity. Radiation therapy may correct CSF flow abnormalities before use of intrathecal chemotherapy. Neoplastic meningitis can also lead to intracranial hypertension and hydrocephalus. Placement of a ventriculoperitoneal shunt may effectively palliate symptoms in these patients.

The development of neoplastic meningitis usually occurs in the setting of uncontrolled cancer outside the CNS; thus, prognosis is poor (median survival 10–12 weeks). However, treatment of the neoplastic meningitis may successfully alleviate symptoms and control the CNS spread.

TREATMENT

TREATMENT

Neoplastic Meningitis

Intrathecal chemotherapy, usually methotrexate, cytarabine, or thiotepa, is delivered by lumbar puncture or by an intraventricular reservoir (Ommaya). Among solid tumors, breast cancer responds best to therapy. Focal radiotherapy may have role in bulky disease, and in symptomatic or obstructive lesions. Targeted therapy such as systemically administered epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors (TKIs) in non-small-cell lung cancer may improve in subgroups of cancer patients with leptomeningeal spread. Patients with neoplastic meningitis from either acute leukemia or lymphoma may be cured of their CNS disease if the systemic disease can be eliminated.

SEIZURES

Seizures occurring in a patient with cancer can be caused by the tumor itself, by metabolic disturbances, by radiation injury, by cerebral infarctions, by chemotherapy-related encephalopathies, or by CNS infections. Metastatic disease to the CNS is the most common cause of seizures in patients with cancer. However, seizures occur more frequently in primary brain tumors than in metastatic brain lesions. Seizures are a presenting symptom of CNS metastasis in 6–29% of cases. Approximately 10% of patients with CNS metastasis eventually develop seizures. Tumors that affect the frontal, temporal, and parietal lobes are more commonly associated with seizures than are occipital lesions. Both early and late seizures are uncommon in patients with posterior fossa and sellar lesions. Seizures are common in patients with CNS metastases from melanoma and low-grade primary brain tumors. Very rarely, cytotoxic drugs such as etoposide, busulfan, ifosfamide, and chlorambucil cause seizures. Another cause of seizures related to drug therapy is reversible posterior leukoencephalopathy syndrome (RPLS). Chemotherapy, targeted therapy, and immunotherapies have been associated with the development of RPLS. RPLS occurs in patients undergoing allogeneic bone marrow or solid-organ transplantation. RPLS is characterized by headache, altered consciousness, generalized seizures, visual disturbances, hypertension, and symmetric posterior cerebral white matter vasogenic edema on CT/MRI. Seizures may begin focally but are typically generalized.

TREATMENT

TREATMENT

Seizures

Patients in whom seizures due to CNS metastases have been demonstrated should receive anticonvulsive treatment with phenytoin or levetiracetam. If this is not effective, valproic acid can be added. Prophylactic anticonvulsant therapy is not recommended. In postcraniotomy patients, prophylactic antiepileptic drugs should be withdrawn during the first week after surgery. Most antiseizure medications including phenytoin induce cytochrome P450 (CYP450), which alters the metabolism of many antitumor agents, including irinotecan, taxanes, and etoposide as well as molecular targeted agents, including imatinib, gefitinib, erlotinib, tipifarnib, sorafenib, sunitinib, temsirolimus, everolimus, and vemurafenib. Levetiracetam and topiramate are anticonvulsant agents not metabolized by the hepatic CYP450 system and do not alter the metabolism of antitumor agents. They have become the preferred drugs. Surgical resection and other antitumor treatments such as radiotherapy and chemotherapy may improve seizure control.

PULMONARY AND INTRACEREBRAL LEUKOSTASIS

Hyperleukocytosis and the leukostasis syndrome associated with it is a potentially fatal complication of acute leukemia (particularly myeloid leukemia) that can occur when the peripheral blast cell count is >100,000/mL. The frequency of hyperleukocytosis is 5–13% in acute myeloid leukemia (AML) and 10–30% in acute lymphoid leukemia; however, leukostasis is rare in lymphoid leukemia. At such high blast cell counts, blood viscosity is increased, blood flow is slowed by aggregates of tumor cells, and the primitive myeloid leukemic cells are capable of invading through the endothelium and causing hemorrhage. Brain and lung are most commonly affected. Patients with brain leukostasis may experience stupor, headache, dizziness, tinnitus, visual disturbances, ataxia, confusion, coma, or sudden death. On examination, papilledema, retinal vein distension, retinal hemorrhages, and focal deficit may be present. Pulmonary leukostasis may present as respiratory distress and hypoxemia and progress to respiratory failure. Chest radiographs may be normal but usually show interstitial or alveolar infiltrates. Hyperleukocytosis rarely may cause acute leg ischemia, renal vein thrombosis, myocardial ischemia, bowel infraction, and priapism. Arterial blood gas results should be interpreted cautiously. Rapid consumption of plasma oxygen by the markedly increased number of white blood cells can cause spuriously low arterial oxygen tension. Pulse oximetry is the most accurate way of assessing oxygenation in patients with hyperleukocytosis. Hydroxyurea can rapidly reduce a high blast cell count while the diagnostic workup is in progress. After the diagnosis is established, the patient should start quickly with effective

induction chemotherapy. Leukapheresis should be used in patients with symptoms of hyperleukocytosis. Patients with hyperleukocytosis are also at the risk for disseminated intravascular coagulation and tumor lysis syndrome. The clinician should monitor the patient for these complications and take preventive and therapeutic actions during induction therapy. Intravascular volume depletion and unnecessary blood transfusions may increase blood viscosity and worsen the leukostasis syndrome. Leukostasis is very rarely a feature of the high white cell counts associated with chronic lymphoid or chronic myeloid leukemia.

When acute promyelocytic leukemia is treated with differentiating agents like tretinoin and arsenic trioxide, cerebral or pulmonary leukostasis may occur as tumor cells differentiate into mature neutrophils. This complication can be largely avoided by using cytotoxic chemotherapy together with the differentiating agents.

HEMOPTYSIS

Hemoptysis may be caused by nonmalignant conditions, but lung cancer accounts for a large proportion of cases. Up to 20% of patients with lung cancer have hemoptysis some time in their course. Endobronchial metastases from carcinoid tumors, breast cancer, colon cancer, kidney cancer, and melanoma may also cause hemoptysis. The volume of bleeding is often difficult to gauge. Massive hemoptysis is defined as >200–600 mL of blood produced in 24 h. However, any hemoptysis should be considered massive if it threatens life. When respiratory difficulty occurs, hemoptysis should be treated emergently. The first priorities are to maintain the airway, optimize oxygenation, and stabilize the hemodynamic status. If the bleeding side is known, the patient should be placed in a lateral decubitus position, with the bleeding side down to prevent aspiration into the unaffected lung, and given supplemental oxygen. If large-volume bleeding continues or the airway is compromised, the patient should be intubated and undergo emergency bronchoscopy. If the site of bleeding is detected, either the patient undergoes a definitive surgical procedure or the lesion is treated with a neodymium:yttrium-aluminum-garnet (Nd:YAG) laser, argon plasma coagulation, or electrocautery. In stable patients, multidetector CT angiography delineates bronchial and nonbronchial systemic arteries and identifies the source of bleeding and underlying pathology with high sensitivity. Massive hemoptysis usually originates from the high-pressure bronchial circulation. Bronchial artery embolization is considered a first-line definite procedure for managing hemoptysis. Bronchial artery embolization may control brisk bleeding in 75–90% of patients, permitting the definitive surgical procedure to be done more safely if it is appropriate.

Embolization without definitive surgery is associated with rebleeding in 20–50% of patients. Recurrent hemoptysis usually responds to a second embolization procedure. A postembolization syndrome characterized by pleuritic pain, fever, dysphagia, and leukocytosis may occur; it lasts 5–7 days and resolves with symptomatic treatment. Bronchial or esophageal wall necrosis, myocardial infarction, and spinal cord infarction are rare complications. Surgery, as a salvage strategy, is indicated after failure of embolization and is associated with better survival when performed in a nonurgent setting.

Pulmonary hemorrhage with or without hemoptysis in hematologic malignancies is often associated with fungal infections, particularly *Aspergillus* sp. After granulocytopenia resolves, the lung infiltrates in aspergillosis may cavitate and cause massive hemoptysis. Thrombocytopenia and coagulation defects should be corrected, if possible. Surgical evaluation is recommended in patients with aspergillosis-related cavitary lesions.

Bevacizumab, an antibody to vascular endothelial growth factor (VEGF) that inhibits angiogenesis, has been associated with life-threatening hemoptysis in patients with non-small-cell lung cancer, particularly of squamous cell histology. Non-small-cell lung cancer patients with cavitary lesions or previous hemoptysis (≥2.5 mL) within the past 3 months have higher risk for pulmonary hemorrhage.

AIRWAY OBSTRUCTION

Airway obstruction refers to a blockage at the level of the mainstem bronchi or above. It may result either from intraluminal tumor growth or from extrinsic compression of the airway. The most common cause of malignant upper airway obstruction is invasion from an adjacent primary tumor, most commonly lung cancer, followed by esophageal, thyroid, and mediastinal malignancies including lymphomas. Extrathoracic primary tumors such as renal, colon, or breast cancer can cause airway obstruction through endobronchial and/or mediastinal lymph node metastases. Patients may present with dyspnea, hemoptysis, stridor, wheezing, intractable cough, postobstructive pneumonia, or hoarseness. Chest radiographs usually demonstrate obstructing lesions. CT scans reveal the extent of tumor. Cool, humidified oxygen, glucocorticoids, and ventilation with a mixture of helium and oxygen (Heliox) may provide temporary relief. If the obstruction is proximal to the larynx, a tracheostomy may be lifesaving. For more distal obstructions, particularly intrinsic

lesions incompletely obstructing the airway, bronchoscopy with mechanical debulking and dilation or ablational treatments including laser treatment, photodynamic therapy, argon plasma coagulation, electrocautery, or stenting can produce immediate relief in most patients (Fig. 71-3). However, radiation therapy (either external-beam irradiation or brachytherapy) given together with glucocorticoids may also open the airway. Symptomatic extrinsic compression may be palliated by stenting. Patients with primary airway tumors such as squamous cell carcinoma, carcinoid tumor, adenocystic carcinoma, or non-small-cell lung cancer, if resectable, should have surgery.

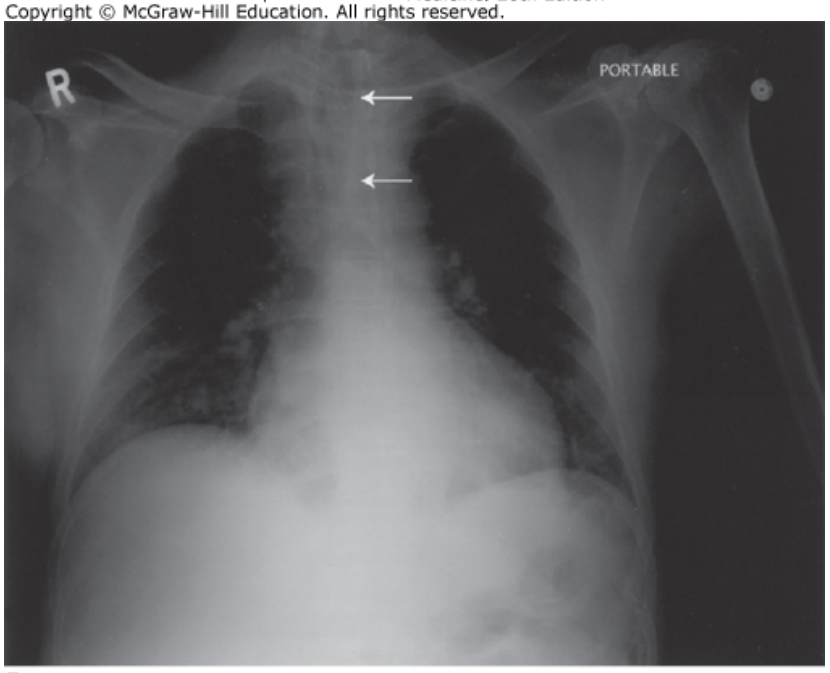
FIGURE 71-3

Airway obstruction. ***A.*** Computed tomography scan of a 62-year-old man with tracheal obstruction caused by renal carcinoma showing paratracheal mass with tracheal invasion/obstruction (*arrow*). ***B.*** Chest x-ray of same patient after stent (*arrows*) placement.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

METABOLIC EMERGENCIES

HYPERCALCEMIA

Hypercalcemia is the most common paraneoplastic syndrome. **Its pathogenesis and management are discussed fully in Chaps. 89 and 403.**

SYNDROME OF INAPPROPRIATE SECRETION OF ANTIDIURETIC HORMONE

Hyponatremia is a common electrolyte abnormality in cancer patients, and syndrome of inappropriate secretion of antidiuretic hormone (SIADH) is the most common cause among patients with cancer. **SIADH is discussed fully in Chaps. 89 and 374.**

LACTIC ACIDOSIS

Lactic acidosis is a rare and potentially fatal metabolic complication of cancer. Lactic acidosis associated with sepsis and circulatory failure is a common preterminal event in many malignancies. Lactic acidosis in the absence of hypoxemia may occur in patients with leukemia, lymphoma, or solid tumors. In some cases, hypoglycemia also is present. Extensive involvement of the liver by tumor is often present. In most cases, decreased metabolism and increased production by the tumor both contribute to lactate accumulation. Tumor cell overexpression of certain glycolytic enzymes and mitochondrial dysfunction can contribute to its increased lactate production. HIV-infected patients have an increased risk of aggressive lymphoma; lactic acidosis that occurs in such patients may be related either to the rapid growth of the tumor or from toxicity of nucleoside reverse transcriptase inhibitors. Symptoms of lactic acidosis include tachypnea, tachycardia, change of mental status, and hepatomegaly. The serum level of lactic acid may reach 10–20 mmol/L (90–180 mg/dL). Treatment is aimed at the underlying disease. *The danger from lactic acidosis is from the acidosis, not the lactate*. Sodium bicarbonate should be added if acidosis is very severe or if hydrogen ion production is very rapid and uncontrolled. Other treatment options include renal replacement therapy, such as hemodialysis, and thiamine replacement. The prognosis is poor regardless of the treatment offered.

HYPOGLYCEMIA

Persistent hypoglycemia is occasionally associated with tumors other than pancreatic islet cell tumors. Usually these tumors are large; tumors of mesenchymal origin, hepatomas, or adrenocortical tumors may cause hypoglycemia. Mesenchymal tumors are usually located in the retroperitoneum or thorax. Obtundation, confusion, and behavioral aberrations occur in the postabsorptive period and may precede the diagnosis of the tumor. These tumors often secrete incompletely processed insulin-like growth factor II (IGF-II), a hormone capable of activating insulin receptors and causing hypoglycemia. Tumors secreting incompletely processed big IGF-II are characterized by an increased IGF-II to IGF-I ratio, suppressed insulin and C-peptide level, and inappropriately low growth hormone and β-hydroxybutyrate concentrations. Rarely, hypoglycemia is due to insulin secretion by a non-islet cell carcinoma. The development of hepatic dysfunction from liver metastases and increased glucose consumption by the tumor can contribute to hypoglycemia. If the tumor cannot be resected, hypoglycemia symptoms may be relieved by the administration of glucose, glucocorticoids, recombinant growth hormone, or glucagon.

Hypoglycemia can be artifactual; hyperleukocytosis from leukemia, myeloproliferative diseases, leukemoid reactions, or colony-stimulating factor treatment can increase glucose consumption in the test tube after blood is drawn, leading to pseudohypoglycemia.

ADRENAL INSUFFICIENCY

In patients with cancer, adrenal insufficiency may go unrecognized because the symptoms, such as nausea, vomiting, anorexia, and orthostatic hypotension, are nonspecific and may be mistakenly attributed to progressive cancer or to therapy. Primary adrenal insufficiency may develop owing to replacement of both glands by metastases (lung, breast, colon, or kidney cancer; lymphoma), to removal of both glands, or to hemorrhagic necrosis in association with sepsis or anticoagulation. Impaired adrenal steroid synthesis occurs in patients being treated for cancer with mitotane, ketoconazole, or aminoglutethimide or undergoing rapid reduction in glucocorticoid therapy. Megestrol acetate, used to manage cancer and HIV-related cachexia, may suppress plasma levels of cortisol and adrenocorticotropic hormone (ACTH). Patients taking megestrol may develop adrenal insufficiency, and even those whose adrenal dysfunction is not symptomatic may have inadequate adrenal reserve if they become seriously ill. Paradoxically, some patients may develop Cushing's syndrome and/or hyperglycemia because of the glucocorticoid-like activity of megestrol acetate. Ipilimumab, an anti-CTLA-4 antibody used for treatment of malignant melanoma, may cause autoimmunity including autoimmune-like enterocolitis, hypophysitis, (leading to secondary adrenal insufficiency), hepatitis, and, rarely, primary adrenal insufficiency. Autoimmune hypophysitis may present with headache, visual field defects, and pituitary hormone deficiencies manifesting as hypopituitarism, adrenal insufficiency (including adrenal crisis), or hypothyroidism. Ipilimumab-associated hypophysitis symptoms occur at an average of 6–12 weeks after initiation of therapy. An MRI usually shows homogenous enhancement of pituitary gland. Early glucocorticoid

treatment and hormone replacement are the initial treatment. The role of high-dose glucocorticoids in the treatment of hypophysitis is not clear. High-dose glucocorticoids may not improve the frequency of pituitary function recovery. Autoimmune adrenalitis can also be observed with anti-CTLA-4 antibody. Pituitary dysfunction is usually permanent, requiring long term hormone replacement therapy. Other checkpoint inhibitors, monoclonal antibodies targeting program death-1 (PD-1), an inhibitory receptor expressed by T cells or one of its ligands (PD-L1) may cause hypophysitis infrequently (~1 %). Autoimmune adrenalitis is more frequent with use of PD/PD-L1 than with CTLA-4 inhibitors, but incidence is low. Cranial irradiation for childhood brain tumors may affect the hypothalamus-pituitary-adrenal axis, resulting in secondary adrenal insufficiency. Rarely, metastatic replacement causes primary adrenal insufficiency as the first manifestation of an occult malignancy. Metastasis to the pituitary or hypothalamus is found at autopsy in up to 5% of patients with cancer, but associated secondary adrenal insufficiency is rare.

Acute adrenal insufficiency is potentially lethal. Treatment of suspected adrenal crisis is initiated after the sampling of serum cortisol and ACTH levels (Chap. 379).

TREATMENT-RELATED EMERGENCIES

TUMOR LYSIS SYNDROME

Tumor lysis syndrome (TLS) is characterized by hyperuricemia, hyperkalemia, hyperphosphatemia, and hypocalcemia, and is caused by the destruction of a large number of rapidly proliferating neoplastic cells. Acidosis may also develop. Acute renal failure occurs frequently.

TLS is most often associated with the treatment of Burkitt's lymphoma, acute lymphoblastic leukemia, and other rapidly proliferating lymphomas, but it also may be seen with chronic leukemias and, rarely, with solid tumors. This syndrome has been seen in patients with chronic lymphocytic leukemia after treatment with nucleosides like fludarabine and is increased in frequency in lymphoid neoplasms treated with venetoclax, a bcl-2 antagonist. TLS has been observed with administration of glucocorticoids, hormonal agents such as letrozole and tamoxifen, and monoclonal antibodies such as rituximab and gemtuzumab. TLS usually occurs during or shortly (1–5 days) after chemotherapy. Rarely, spontaneous necrosis of malignancies causes TLS.

Hyperuricemia may be present at the time of chemotherapy. Effective treatment kills malignant cells and leads to increased serum uric acid levels from the turnover of nucleic acids. Owing to the acidic local environment, uric acid can precipitate in the tubules, medulla, and collecting ducts of the kidney, leading to renal failure. Lactic acidosis and dehydration may contribute to the precipitation of uric acid in the renal tubules. The finding of uric acid crystals in the urine is strong evidence for uric acid nephropathy. The ratio of urinary uric acid to urinary creatinine is >1 in patients with acute hyperuricemic nephropathy and <1 in patients with renal failure due to other causes.

Hyperphosphatemia, which can be caused by the release of intracellular phosphate pools by tumor lysis, produces a reciprocal depression in serum calcium, which causes severe neuromuscular irritability and tetany. Deposition of calcium phosphate in the kidney and hyperphosphatemia may cause renal failure. Potassium is the principal intracellular cation, and massive destruction of malignant cells may lead to hyperkalemia. Hyperkalemia in patients with renal failure may rapidly become life threatening by causing ventricular arrhythmias and sudden death.

The likelihood that TLS will occur in patients with Burkitt's lymphoma is related to the tumor burden and renal function. Hyperuricemia and high serum levels of lactate dehydrogenase (LDH >1500 U/L), both of which correlate with total tumor burden, also correlate with the risk of TLS. In patients at risk for TLS, pretreatment evaluations should include a complete blood count, serum chemistry evaluation, and urine analysis. High leukocyte and platelet counts may artificially elevate potassium levels ("pseudohyperkalemia") due to lysis of these cells after the blood is drawn. In these cases, plasma potassium instead of serum potassium should be followed. In pseudohyperkalemia, no electrocardiographic abnormalities are present. In patients with abnormal baseline renal function, the kidneys and retroperitoneal area should be evaluated by sonography and/or CT to rule out obstructive uropathy. Urine output should be watched closely.

TREATMENT

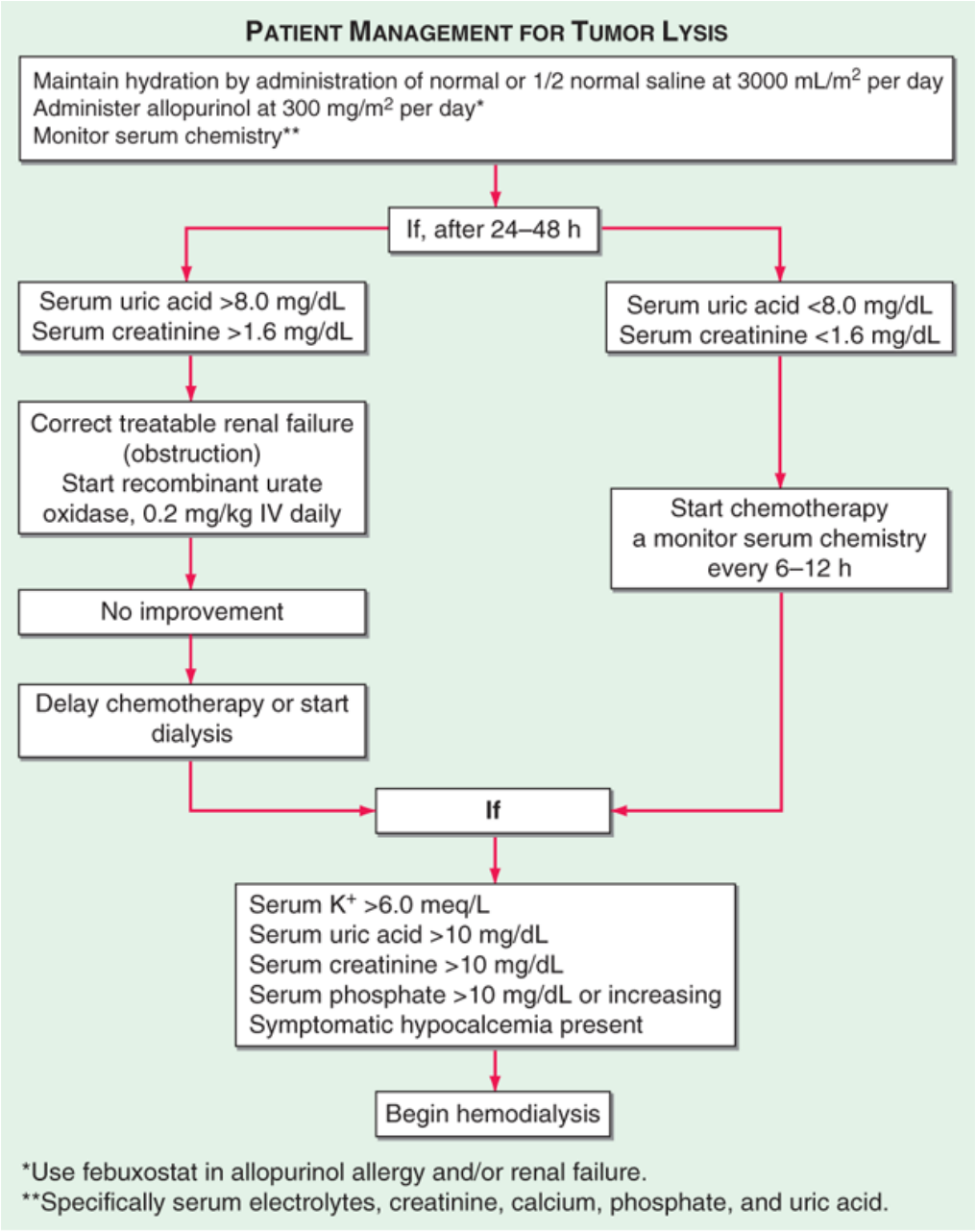
TREATMENT

Tumor Lysis Syndrome

Recognition of risk and prevention are the most important steps in the management of this syndrome (Fig. 71-4). The standard preventive approach consists of allopurinol and aggressive hydration. Urinary alkalization with sodium bicarbonate is no longer recommended. It increases uric acid solubility, but a high pH decreases the solubility of xanthine, hypoxanthine, and calcium phosphate, potentially increasing the likelihood of intratubular crystallization. Intravenous allopurinol may be given in patients who cannot tolerate oral therapy. Febuxostat, a potent nonpurine selective xanthine oxidase inhibitor, is indicated for treatment of hyperuricemia. It has less hypersensitivity reactions than allopurinol. Febuxostat does not require dosage adjustment in patients with mild to moderate renal impairment. Febuxostat achieved significantly superior serum uric acid control in comparison to allopurinol in patients with hematologic malignancies at intermediate to high TLS risk. In some cases, uric acid levels cannot be lowered sufficiently with the standard preventive approach. Rasburicase (recombinant urate oxidase) can be effective in these instances, particularly when renal failure is present. Urate oxidase is missing from primates and catalyzes the conversion of poorly soluble uric acid to readily soluble allantoin. Rasburicase acts rapidly, decreasing uric acid levels within hours; however, it may cause hypersensitivity reactions such as bronchospasm, hypoxemia, and hypotension. Rasburicase should also be administered to high-risk patients for TLS prophylaxis. Rasburicase is contraindicated in patients with glucose-6-phosphate dehydrogenase deficiency who are unable to break down hydrogen peroxide, an end product of the urate oxidase reaction. Rasburicase is known to cause ex vivo enzymatic degradation of uric acid in test tube at room temperature. This leads to spuriously low uric acid levels during laboratory monitoring of the patient with TLS. Samples must be cooled immediately to deactivate the urate oxidase. Despite aggressive prophylaxis, TLS and/or oliguric or anuric renal failure may occur. Dialysis is often necessary and should be considered early in the course. Hemodialysis is preferred. Hemofiltration offers a gradual, continuous method of removing cellular by-products and fluid.

FIGURE 71-4

Management of patients at high risk for the tumor lysis syndrome.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HUMAN ANTIBODY INFUSION REACTIONS

The initial infusion of human or humanized antibodies (e.g., rituximab, gemtuzumab, trastuzumab, alemtuzumab, panitumumab, brentuximab vedotin, blinatumomab) is associated with fever, chills, nausea, asthenia, and headache in up to half of treated patients. Bronchospasm and hypotension occur in 1% of patients. Severe manifestations including pulmonary infiltrates, acute respiratory distress syndrome (ARDS), and cardiogenic shock occur rarely. Laboratory manifestations include elevated hepatic aminotransferase levels, thrombocytopenia, and prolongation of prothrombin time. The pathogenesis is thought to be activation of immune effector processes (cells and complement) and release of inflammatory cytokines, such as tumor necrosis factor α, interferon gamma, interleukin 6, and interleukin 10 (cytokine release syndrome [CRS]). Although its origins are not completely understood, CRS is believed to be due to activation of a variety of cell types including monocytes/macrophages and T and B lymphocytes. Severe reactions from rituximab have occurred with high numbers (>50 × 10^9 lymphocytes) of circulating cells bearing the target antigen (CD20) and have been associated with a rapid fall in circulating tumor cells, mild electrolyte evidence of TLS, and, very rarely, death. In addition, increased liver enzymes, D-dimer, and LDH and prolongation of the prothrombin time may occur. Diphenhydramine, hydrocortisone, and acetaminophen can often prevent or suppress the infusion-related symptoms. If they occur, the infusion is stopped and restarted at

half the initial infusion rate after the symptoms have abated. Severe CRS may require intensive support for ARDS and resistant hypotension. Emerging clinical experience at several institutions has concluded that tocilizumab is an effective treatment for severe or life-threatening CRS. Tocilizumab prevents IL-6 binding to both cell-associated and soluble IL-6Rs and therefore inhibits both classical and trans-IL-6 signaling.

Adoptive transfer of chimeric antigen receptor (CAR)–engineered T cells is a promising therapy for cancers. The most common acute toxicity of CAR T cells is CRS. CAR T-cell–associated CRS may be associated cardiac dysfunction and neurotoxicity. The management includes supportive care and tocilizumab.

HEMOLYTIC-UREMIC SYNDROME

Hemolytic-uremic syndrome (HUS) and, less commonly, thrombotic thrombocytopenic purpura (TTP) **(Chap. 311)** may rarely occur after treatment with antineoplastic drugs, including mitomycin, gemcitabine, cisplatin, and bleomycin, and with VEGF inhibitors. Mitomycin and gemcitabine are the most common offenders. Unlike mitomycin, there is no clear-cut relationship between the cumulative dose of gemcitabine and risk of HUS. It occurs most often in patients with gastric, lung, colorectal, pancreatic, and breast carcinoma. In one series, 35% of patients were without evident cancer at the time this syndrome appeared. Secondary HUS/TTP has also been reported as a rare but sometimes fatal complication of BMT.

HUS usually has its onset 4–8 weeks after the last dose of chemotherapy, but it is not rare to detect it several months later. HUS is characterized by microangiopathic hemolytic anemia, thrombocytopenia, and renal failure. Dyspnea, weakness, fatigue, oliguria, and purpura are also common initial symptoms and findings. Systemic hypertension and pulmonary edema frequently occur. Severe hypertension, pulmonary edema, and rapid worsening of hemolysis and renal function may occur after a blood or blood product transfusion. Cardiac findings include atrial arrhythmias, pericardial friction rub, and pericardial effusion. Raynaud's phenomenon is part of the syndrome in patients treated with bleomycin.

Laboratory findings include severe to moderate anemia associated with red blood cell fragmentation and numerous schistocytes on peripheral smear. Reticulocytosis, decreased plasma haptoglobin, and an LDH level document hemolysis. The serum bilirubin level is usually normal or slightly elevated. The Coombs' test is negative. The white cell count is usually normal, and thrombocytopenia (<100,000/μL) is almost always present. Most patients have a normal coagulation profile, although some have mild elevations in thrombin time and in levels of fibrin degradation products. The serum creatinine level is elevated at presentation and shows a pattern of subacute worsening within weeks of the initial azotemia. The urinalysis reveals hematuria, proteinuria, and granular or hyaline casts, and circulating immune complexes may be present.

The basic pathologic lesion appears to be deposition of fibrin in the walls of capillaries and arterioles, and these deposits are similar to those seen in HUS due to other causes. These microvascular abnormalities involve mainly the kidneys and rarely occur in other organs. The pathogenesis of cancer treatment–related HUS is not completely understood, but probably the most important factor is endothelial damage. Primary forms of HUS/TTP are related to a decrease in processing of von Willebrand factor by a protease called ADAMTS13.

The case fatality rate is high; most patients die within a few months. There is no consensus on the optimal treatment for chemotherapy-induced HUS. Treatment modalities for HUS/TTP including immunocomplex removal (plasmapheresis, immunoadsorption, or exchange transfusion), antiplatelet/anticoagulant therapies, immunosuppressive therapies, and plasma exchange have varying degrees of success. The outcome with plasma exchange is generally poor, as in many other cases of secondary TTP. Rituximab is successfully used in patients with chemotherapy-induced HUS as well as in ADAMTS13-deficient TTP.

NEUTROPENIA AND INFECTION

These remain the most common serious complications of cancer therapy. **They are covered in detail in Chap. 70.**

PULMONARY INFILTRATES

Patients with cancer may present with dyspnea associated with diffuse interstitial infiltrates on chest radiographs. Such infiltrates may be due to progression of the underlying malignancy, treatment-related toxicities, infection, and/or unrelated diseases. The cause may be multifactorial; however, most commonly they occur as a consequence of treatment. Infiltration of the lung by malignancy has been

described in patients with leukemia, lymphoma, and breast and other solid cancers. Pulmonary lymphatics may be involved diffusely by neoplasm (pulmonary lymphangitic carcinomatosis), resulting in a diffuse increase in interstitial markings on chest radiographs. The patient is often mildly dyspneic at the onset, but pulmonary failure develops over a period of weeks. In some patients, dyspnea precedes changes on the chest radiographs and is accompanied by a nonproductive cough. This syndrome is characteristic of solid tumors. In patients with leukemia, diffuse microscopic neoplastic peribronchial and peribronchiolar infiltration is frequent but may be asymptomatic. However, some patients present with diffuse interstitial infiltrates, an alveolar capillary block syndrome, and respiratory distress. Thickening of bronchovascular bundles and prominence of peripheral arteries are CT findings suggestive of leukemic infiltration. In these situations, glucocorticoids can provide symptomatic relief, but specific chemotherapy should always be started promptly.

Several cytotoxic agents, such as bleomycin, methotrexate, busulfan, nitrosoureas, gemcitabine, mitomycin, vinorelbine, docetaxel, paclitaxel, fludarabine, pentostatin, and ifosfamide may cause pulmonary damage. The most frequent presentations are interstitial pneumonitis, alveolitis, and pulmonary fibrosis. Some cytotoxic agents, including methotrexate and procarbazine, may cause an acute hypersensitivity reaction. Cytosine arabinoside has been associated with noncardiogenic pulmonary edema. Administration of multiple cytotoxic drugs, as well as radiotherapy and preexisting lung disease, may potentiate the pulmonary toxicity. Supplemental oxygen may potentiate the effects of drugs and radiation injury. Patients should always be managed with the lowest F_{IO_2} that is sufficient to maintain hemoglobin saturation.

The onset of symptoms may be insidious, with symptoms including dyspnea, nonproductive cough, and tachycardia. Patients may have bibasilar crepitant rales, end-inspiratory crackles, fever, and cyanosis. The chest radiograph generally shows an interstitial and sometimes an intraalveolar pattern that is strongest at the lung bases and may be symmetric. A small effusion may occur. Hypoxemia with decreased carbon monoxide diffusing capacity is always present. Glucocorticoids may be helpful in patients in whom pulmonary toxicity is related to radiation therapy or to chemotherapy. Treatment is otherwise supportive.

Molecular targeted agents, imatinib, erlotinib, and gefitinib are potent inhibitors of tyrosine kinases. These drugs may cause interstitial lung disease (ILD). In the case of gefitinib, preexisting fibrosis, poor performance status, and prior thoracic irradiation are independent risk factors; this complication has a high fatality rate. In Japan, incidence of ILD associated with gefitinib was ~4.5% compared to 0.5% in the United States. Temsirolimus and everolimus, both esters a derivative of rapamycin, are agents that block the effects of mammalian target of rapamycin (mTOR), an enzyme that has an important role in regulating the synthesis of proteins that control cell division. It may cause ground-glass opacities in the lung with or without diffuse interstitial disease and lung parenchymal consolidation. Patients may be asymptomatic with only radiologic findings or may be symptomatic. Symptoms include cough, dyspnea, and/or hypoxemia, and sometimes patients present with systemic symptoms such as fever and fatigue. The incidence of everolimus-induced ILD also appears to be higher in Japanese patients. Treatment includes dose reduction or withdrawal and, in some cases, the addition of glucocorticoids.

The Food and Drug Administration (FDA)–approved immune checkpoint inhibitors of the PD-1 and PD-L1 pathway, including nivolumab, pembrolizumab, durvalumab, avelumab, and atezolizumab, enhance antitumor activity by blocking negative regulators of T cell function. Immune-mediated pneumonitis is rare (10%) but a life-threatening complication of these drugs. Pneumonitis symptoms include cough, shortness of breath, dyspnea, and fever, and often involve only asymptomatic radiographic changes. Pneumonitis shows ground-glass patchy lesions and/or disseminated nodular infiltrates, predominantly in the lower lobes. Treatment includes temporary or permanent withdrawal of drug and the addition of high-dose glucocorticoids.

Radiation pneumonitis and/or fibrosis are relatively frequent side effects of thoracic radiation therapy. It may be acute or chronic. Radiation-induced lung toxicity is a function of the irradiated lung volume, dose per fraction, and radiation dose. The larger the irradiated lung field, the higher is the risk for radiation pneumonitis. The use of concurrent chemoradiation, particularly regimens including paclitaxel, increases pulmonary toxicity. Radiation pneumonitis usually develops 2–6 months after completion of radiotherapy. The clinical syndrome, which varies in severity, consists of dyspnea, cough with scanty sputum, low-grade fever, and an initial hazy infiltrate on chest radiographs. The infiltrate and tissue damage usually are confined to the radiation field. The CT scan may show ground-glass opacities, consolidation, fibrosis, atelectatic cicatrization, pleural volume loss, or pleural thickening. The patients subsequently may develop a patchy alveolar infiltrate and air bronchograms, which may progress to acute respiratory failure that is sometimes fatal. A lung biopsy may be necessary to make the diagnosis. Asymptomatic infiltrates found incidentally after radiation therapy need not be treated. However, prednisone should be administered to patients with fever or other symptoms. The dosage

should be tapered slowly after the resolution of radiation pneumonitis, because abrupt withdrawal of glucocorticoids may cause an exacerbation of pneumonia. Delayed radiation fibrosis may occur years after radiation therapy and is signaled by dyspnea on exertion. Often it is mild, but it can progress to chronic respiratory failure. Therapy is supportive.

Classic radiation pneumonitis that leads to pulmonary fibrosis is due to radiation-induced production of local cytokines such as platelet-derived growth factor β, tumor necrosis factor, interleukins, and transforming growth factor β in the radiation field.

Stereotactic body radiation therapy (SBRT) is a radiotherapy treatment method that has been applied to the treatment of stage I lung cancers in medically inoperable patients. SBRT accurately delivers a high dose of irradiation in one or few treatment fractions to an image-defined lung mass. Most of the acute changes after SBRT occur later than 3 months after treatment, and the shape of the SBRT-induced injury conforms more tightly to the tumor.

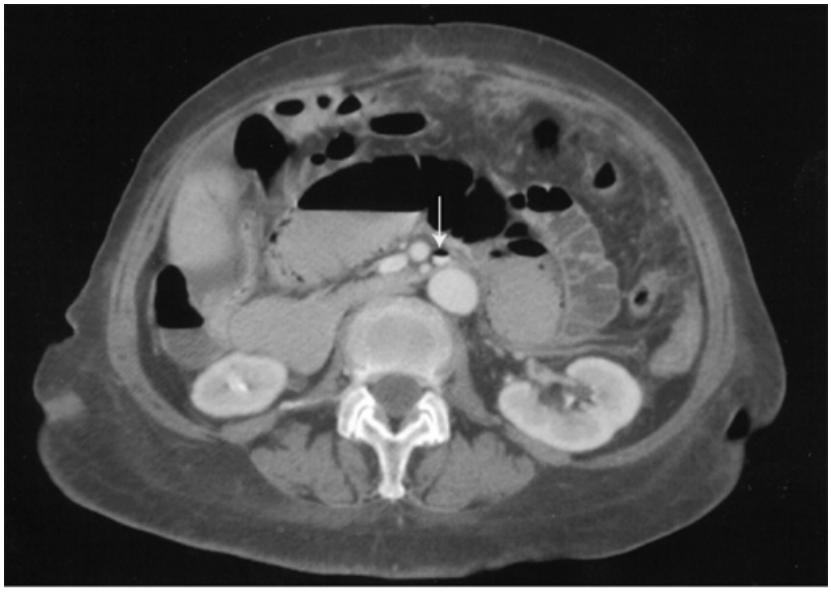
Pneumonia is a common problem in patients undergoing treatment for cancer (**see Chap 70**). In patients with pulmonary infiltrates who are afebrile, heart failure and multiple pulmonary emboli are in the differential diagnosis.

NEUTROPENIC ENTEROCOLITIS

Neutropenic enterocolitis (typhlitis) is the inflammation and necrosis of the cecum and surrounding tissues that may complicate the treatment of acute leukemia. Nevertheless, it may involve any segment of the gastrointestinal tract including small intestine, appendix, and colon. This complication has also been seen in patients with other forms of cancer treated with taxanes, 5-fluorouracil, irinotecan, vinorelbine, cisplatin, carboplatin, and high-dose chemotherapy **(Fig. 71-5)**. It also has been reported in patients with AIDS, aplastic anemia, cyclic neutropenia, idiosyncratic drug reactions involving antibiotics, and immunosuppressive therapies. The patient develops right lower quadrant abdominal pain, often with rebound tenderness and a tense, distended abdomen, in a setting of fever and neutropenia. Watery diarrhea (often containing sloughed mucosa) and bacteremia are common, and bleeding may occur. Plain abdominal films are generally of little value in the diagnosis; CT scan may show marked bowel wall thickening, particularly in the cecum, with bowel wall edema, mesenteric stranding, and ascites, and may help to differentiate neutropenic colitis from other abdominal disorders such as appendicitis, diverticulitis, and *Clostridium difficile*–associated colitis in this high-risk population. Patients with bowel wall thickness >10 mm on ultrasonogram have higher mortality rates. However, bowel wall thickening is significantly more prominent in patients with *C. difficile* colitis. Pneumatosis intestinalis is a more specific finding, seen only in those with neutropenic enterocolitis and ischemia. The combined involvement of the small and large bowel suggests a diagnosis of neutropenic enterocolitis. Rapid institution of broad-spectrum antibiotics, bowel rest, and nasogastric suction may reverse the process. Use of myeloid growth factors improved outcome significantly. Surgical intervention is reserved for severe cases of neutropenic enterocolitis with evidence of perforation, peritonitis, gangrenous bowel, or gastrointestinal hemorrhage despite correction of any coagulopathy.

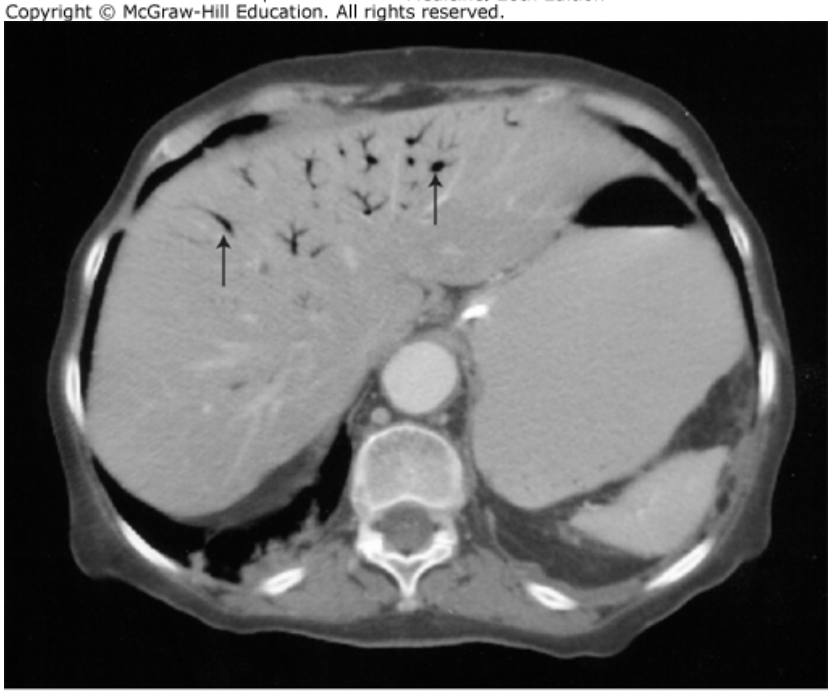
FIGURE 71-5

Abdominal computed tomography (CT) scans of a 72-year-old woman with neutropenic enterocolitis secondary to chemotherapy. *A.* Air in inferior mesenteric vein (*arrow*) and bowel wall with pneumatosis intestinalis. ***B.*** CT scan of upper abdomen demonstrating air in portal vein (*arrows*).



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

C. difficile colitis is increasing in incidence. Newer strains of *C. difficile* produce ~20 times more of toxins A and B compared to previously studied strains. *C. difficile* risk is also increased with chemotherapy. Antibiotic coverage for *C. difficile* should be added if pseudomembranous colitis cannot be excluded.

HEMORRHAGIC CYSTITIS

Hemorrhagic cystitis is characterized by diffuse bladder mucosal bleeding that develops secondary to chemotherapy (mostly cyclophosphamide or ifosfamide), radiation therapy, bone marrow transplantation (BMT), and/or opportunistic infections. Both cyclophosphamide and ifosfamide are metabolized to acrolein, which is a strong chemical irritant that is excreted in the urine. Prolonged contact or high concentrations may lead to bladder irritation and hemorrhage. Symptoms include gross hematuria, frequency, dysuria, burning, urgency, incontinence, and nocturia. The best management is prevention. Maintaining a high rate of urine flow minimizes exposure. In addition, 2-mercaptoethanesulfonate (mesna) detoxifies the metabolites and can be coadministered with

the instigating drugs. Mesna usually is given three times on the day of ifosfamide administration in doses that are each 20% of the total ifosfamide dose. If hemorrhagic cystitis develops, the maintenance of a high urine flow may be sufficient supportive care. If conservative management is not effective, irrigation of the bladder with a 0.37–0.74% formalin solution for 10 min stops the bleeding in most cases. *N*-Acetylcysteine may also be an effective irrigant. Prostaglandin (carboprost) can inhibit the process. In extreme cases, ligation of the hypogastric arteries, urinary diversion, or cystectomy may be necessary.

In the BMT setting, early-onset hemorrhagic cystitis is related to drugs in the treatment regimen (e.g., cyclophosphamide), and late-onset hemorrhagic cystitis is usually due to the polyoma virus BKV or adenovirus type 11. BKV load in urine alone or in combination with acute graft-versus-host disease correlates with development of hemorrhagic cystitis. Viral causes are usually detected by polymerase chain reaction (PCR)–based diagnostic tests. Treatment of viral hemorrhagic cystitis is largely supportive, with reduction in doses of immunosuppressive agents, if possible. No antiviral therapy is approved, although cidofovir is reported to be effective in a small series. Hyperbaric oxygen therapy has been used successfully in patients with BKV-associated and cyclophosphamide-induced hemorrhagic cystitis during hematopoietic stem cell transplantation, as well as in hemorrhagic radiation cystitis.

HYPERSENSITIVITY REACTIONS TO ANTINEOPLASTIC DRUGS

Many antineoplastic drugs may cause hypersensitivity reaction. These reactions are unpredictable and potentially life threatening. Most reactions occur during or within hours of parenteral drug administration. Taxanes, platinum compounds, asparaginase, etoposide, procarbazine, and biologic agents, including rituximab, bevacizumab, trastuzumab, gemtuzumab, cetuximab, and alemtuzumab, are more commonly associated with acute hypersensitivity reactions than are other agents. Hypersensitivity reactions to some drugs, such as taxanes, occur during the first or second dose administered. Hypersensitivity to platinum compounds occurs after prolonged exposure. Skin testing may identify patients with high risk for hypersensitivity after carboplatin exposure. Premedication with histamine H_1 and H_2 receptor antagonists and glucocorticoids reduces the incidence of hypersensitivity reaction to taxanes, particularly paclitaxel.

Despite premedication, hypersensitivity reactions may still occur. In these cases, rapid desensitization in the intensive care unit setting or re-treatment may be attempted with care, but the use of alternative agents may be required. Skin testing is used to assess the involvement of IgE in the reaction. Tryptase levels measured at the time of the reaction help to explain the mechanism of the reaction and its severity. Increased tryptase levels indicate underlying mast cell activation. Candidate patients for desensitization include those who have mild to severe hypersensitivity type I, with mast cell–mediated and IgE-dependent reactions occurring during a chemotherapy infusion or shortly thereafter.

FURTHER READING

Bauer R et al: Treatment of epileptic seizures in brain tumors: A critical review. Neurosurg Rev 37:381, 2014. [PubMed: 24760366]

Bodnar TW: Management of non-islet-cell tumor hypoglycemia: A clinical review. J Clin Endocrinol Metab 99:713, 2014. [PubMed: 24423303]

Jones GL et al: Guidelines for the management of tumour lysis syndrome in adults and children with haematological malignancies on behalf of the British Committee for Standards in Haematology. Br J Haematol 169:661, 2015. [PubMed: 25876990]

Laufer I et al: The NOMS framework: Approach to the treatment of spinal metastatic tumors. Oncologist 18:744, 2013. [PubMed: 23709750]

Lee DW et al: Current concepts in the diagnosis and management of cytokine release syndrome. Blood 124:188, 2014. [PubMed: 24876563]

Lin X, DeAngelis LM: Treatment of brain metastases. J Clin Oncol 33:3475, 2015. [PubMed: 26282648]

Mack F et al: Therapy of leptomeningeal metastasis in solid tumors. Cancer Treat Rev 43:83, 2016. [PubMed: 26827696]

Nishino M et al: Anti-PD-1 inhibitor-related pneumonitis in non-small cell lung cancer. Cancer Immunol Res 4:289, 2016. [PubMed: 26865455]

Rice TW et al: The superior vena cava syndrome: Clinical characteristics and evolving etiology. Medicine 85:37, 2006. [PubMed: 16523051]

Ruggiero A et al: Management of hyperleukocytosis. Curr Treat Options Oncol 17:7, 2016. [PubMed: 26820286]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 72: Cancer of the Skin

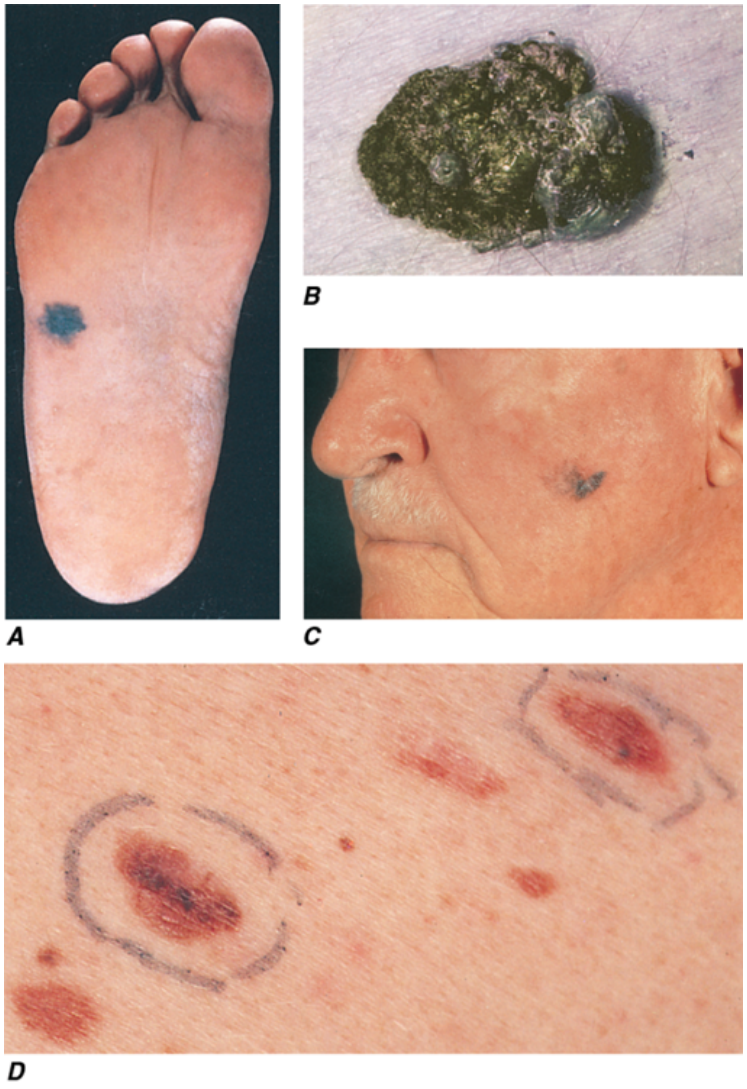
Brendan D. Curti; Sancy Leachman; Walter J. Urba

MELANOMA

Pigmented lesions are among the most common findings on skin examination. The challenge for the physician is to distinguish cutaneous melanomas, which account for the overwhelming majority of deaths resulting from skin cancer, from the remainder, which are usually benign. Cutaneous melanoma can occur in adults of all ages, even young individuals, and people of all colors; its location on the skin and its distinct clinical features often permit detection at a time when complete surgical excision leads to cure. Examples of malignant and benign pigmented lesions are shown in Fig. 72-1.

FIGURE 72-1

Atypical and malignant pigmented lesions. The most common melanoma is superficial spreading melanoma (not pictured). ***A***. Acral lentiginous melanoma is the most common melanoma in blacks, Asians, and Hispanics and occurs as an enlarging hyperpigmented macule or plaque on the palms and soles. ***B***. Nodular melanoma most commonly manifests as a rapidly growing, often ulcerated or crusted black nodule. ***C***. Lentigo maligna melanoma occurs on sun-exposed skin as a large, hyperpigmented macule or plaque with irregular borders and variable pigmentation. ***D***. Dysplastic nevi are benign, irregularly pigmented and shaped melanocytic hamartomas with some atypical cellular features and frequently associated with familial melanoma.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPIDEMIOLOGY

Melanoma is an aggressive malignancy of melanocytes, pigment-producing cells that originate from the neural crest and migrate to the skin, meninges, mucous membranes, upper esophagus, and eyes. Melanocytes in each of these locations have the potential for malignant transformation, but the vast majority arise in the skin. Melanomas can also arise in the mucosa of the head and neck (nasal cavity, paranasal sinuses, and oral cavity), the gastrointestinal tract, the CNS, the female genital tract (vulva, vagina), and the uveal tract of the eye. Cutaneous melanoma is predominantly a malignancy of white-skinned people (98% of cases), and the incidence correlates with latitude of residence, providing strong evidence for the role of sun exposure. Men are affected slightly more than women (1.3:1), and the median age at diagnosis is the late fifties. In 2016, >76,000 individuals in the United States were expected to develop melanoma, and ~10,130 were expected to die. Mortality rates begin to rise at age 55, with the greatest increase in men age >65 years. Of particular concern is the increase in incidence among women <40 years of age, an increase believed to be associated with a greater emphasis on tanned skin as a marker of beauty, the increased availability and use of indoor tanning beds, and exposure to intense ultraviolet (UV) light in childhood. The latest Surveillance, Epidemiology and End Results (SEER) Registry data reveal that from 2004 to 2013, the rate of new melanoma cases has risen 1.4% each year, while death rates have remained stable. This is in the context of a 5-year relative survival improvement from 93.1% to 93.3% overall, despite a 17.9% survival rate for those diagnosed with distant metastases. These statistics highlight the need to promote prevention and early detection.

GLOBAL CONSIDERATIONS

The incidence of both non-melanoma and melanoma skin cancers around the world has been increasing. Every year between 2 and 3 million people will get non-melanoma skin cancer and in 2012 there were 232,000 cases of melanoma. The highest incidence of melanoma is found in New Zealand and Australia consistent with Caucasians living in latitudes with increased UV exposure. The likelihood of developing melanoma is 25 per 100,000 in non-Hispanic whites, 4 per 100,000 in Hispanics, and 1 per 100,000 in African Americans.

Dark-skinned populations (such as those of India and Puerto Rico), blacks, and east Asians also develop melanoma, albeit at rates 10–20 times lower than those in whites. Cutaneous melanomas in these populations are more often diagnosed at a higher stage, and patients tend to have worse outcomes. Furthermore, in nonwhite populations, the frequency of acral (subungual, plantar, palmar) and mucosal melanomas is much higher. In China, about 20,000 new cases are reported each year and, in contrast to the United States where rates are stable, mortality is increasing. This may be due in part to the gap that remains in the diagnosis and treatment of melanoma between China and Western countries or to the fact that in Asians and dark-skinned populations, the melanomas that arise from the skin (comprising 50–70% of patients versus 90% in the West) arise from acral areas and the others from mucosal areas, all of which carry a poorer prognosis than cutaneous melanomas diagnosed in the West.

RISK FACTORS

Presence of Nevi

The risk of developing melanoma is related to genetic, environmental, and host factors. The strongest risk factors for melanoma are the presence of multiple benign or atypical nevi and a family or personal history of melanoma. The presence of >40 melanocytic nevi, common or dysplastic, is a marker for increased risk of melanoma. Nevi have been referred to as precursor lesions because they can transform into melanomas; however, the actual risk of transformation for any individual nevus is exceedingly low. About one-quarter of melanomas are histologically associated with nevi, but the majority arise de novo. The number of clinically atypical moles may vary from one to several hundred, and they usually differ from one another in appearance, although individuals can develop multiple similar atypical nevi (signature nevi). The borders are often hazy and indistinct, and the pigment pattern is more highly varied than that in benign acquired nevi. Individuals with clinically atypical moles and a strong family history of melanoma have been reported to have a >50% lifetime risk for developing melanoma and warrant close follow-up with a dermatologist. Of the 90% of patients whose disease is sporadic (i.e., who lack a family history of melanoma), ~40% have clinically atypical moles, compared with an estimated 5–10% of the population at large.

Congenital melanocytic nevi, which are classified as small (≤1.5 cm), medium (1.5–20 cm), and giant (>20 cm), can be precursors for melanoma. The risk is highest for the giant melanocytic nevus, also called the bathing trunk nevus, a rare malformation that affects 1 in 30,000–100,000 individuals. Since the lifetime risk of melanoma development is estimated to be as high as 6%, prophylactic excision early in life is prudent. This usually requires staged removal with coverage by split-thickness skin grafts. Surgery cannot remove all at-risk nevus cells, as some may penetrate into the muscles or central nervous system (CNS) below the nevus. Small- to medium-size congenital melanocytic nevi affect ~1% of persons; the lifetime risk of melanoma development in a typical nevus is low, estimated to be about 0.03% (1 in 3164) for men and 0.009% (1 in 10,800) for women. The management of small- to medium-size congenital melanocytic nevi remains controversial and is primarily based on histologic findings from biopsies of clinically atypical nevi.

Personal and Family History

Once diagnosed, patients with melanoma require a lifetime of surveillance because their risk of developing another melanoma is 10 times that of the general population. First-degree relatives have a twofold higher risk of developing melanoma than do individuals without a family history, but only 5–10% of all melanomas are truly familial. In familial melanoma, patients tend to be younger at first diagnosis, lesions are thinner, and multiple primary melanomas are common.

Genetic Susceptibility



Approximately 20–40% of cases of hereditary melanoma (0.2–2% of all melanomas) are due to germline mutations in the cell cycle regulatory gene cyclin-dependent kinase inhibitor 2A (*CDKN2A*). In fact, 70% of all cutaneous melanomas have mutations or deletions affecting the *CDKN2A* locus on chromosome 9p21. This locus encodes two distinct tumor-suppressor proteins from alternate reading frames: p16 and ARF ($p14^{ARF}$). The p16 protein inhibits CDK4/6-mediated phosphorylation and inactivation of the retinoblastoma (RB) protein, whereas ARF inhibits MDM2 ubiquitin-mediated degradation of p53. The end result of the loss of *CDKN2A* is inactivation of two critical tumor-suppressor pathways, RB and p53, which control entry of cells into the cell cycle. Several studies have shown an increased risk of pancreatic cancer among melanoma-prone families with *CDKN2A* mutations. A second high-risk locus for melanoma susceptibility, *CDK4*, is located on chromosome 12q13 and encodes the kinase inhibited by p16. *CDK4* mutations, which also inactivate the RB pathway, are much rarer than *CDKN2A* mutations. Germline mutations in the melanoma lineage-specific oncogene microphthalmia-associated transcription factor (*MITF*) and telomerase reverse transcriptase (TERT) mutations predispose to both familial and sporadic melanomas.

The melanocortin-1 receptor (*MC1R*) gene is a moderate-risk inherited melanoma susceptibility factor. Solar radiation stimulates the production of melanocortin (α-melanocyte-stimulating hormone [α-MSH]), the ligand for *MC1R*, which is a G-protein-coupled receptor that signals via cyclic AMP and regulates the amount and type of pigment produced. *MC1R* is highly polymorphic, and among its 80 variants are those that result in partial loss of signaling and lead to the production of red/yellow pheomelanins, which are not sun-protective and produce red hair, rather than brown/black eumelanins that are photoprotective. This red hair color (RHC) phenotype is associated with fair skin, red hair, freckles, increased sun sensitivity, and increased risk of melanoma. In addition to its weak UV shielding capacity relative to eumelanin, increased pheomelanin production in patients with

inactivating polymorphisms of *MC1R* also provides a UV-independent carcinogenic contribution to melanomagenesis via oxidative damage and reduced DNA damage repair.

A number of other more common, low-penetrance polymorphisms that have small effects on melanoma susceptibility include other genes related to pigmentation, nevus count, immune responses, DNA repair, metabolism, and the vitamin D receptor. Approximately 50% of the genetic risk for hereditary melanoma can be ascribed to previously identified melanoma predisposition genes, with ~40% of the risk being due to CDKN2A. The missing inherited risk is most likely due to the inheritance of additional modifier genes and/or shared environmental exposures.

PREVENTION AND EARLY DETECTION

Primary prevention of melanoma and nonmelanoma skin cancer (NMSC) is based on protection from the sun. Public health initiatives, such as the SunSmart program that started in Australia and now is operative in Europe and the United States, have demonstrated that behavioral change can decrease the incidence of NMSC and melanoma. Preventive measures should start early in life because damage from UV light begins early despite the fact that cancers develop years later. Some individuals tan compulsively. There is greater understanding of tanning addiction and the biology of cutaneous-neural connections that may give rise to this behavior. Compulsive tanners exhibit differences in dopamine binding and reactivity in reward pathways in the brain, such as the basal striatum, resulting in cutaneous secretion of β-endorphins after UV exposure. Identifying individuals with tanning addiction may be another method for preventive intervention. Regular use of broad-spectrum sunscreens that block UVA and UVB with a sun protection factor (SPF) of at least 30 and protective clothing should be encouraged. Avoidance of sunburns, tanning beds, and midday sun exposure is recommended.

Secondary prevention comprises education, screening, and early detection. Patients should be taught to recognize the clinical features of melanoma (ABCDEs; see below) and advised to report any change in a pigmented lesion. Brochures are available from the American Cancer Society, the American Academy of Dermatology, the National Cancer Institute, and the Skin Cancer Foundation. Self-examination at monthly intervals may enhance the likelihood of detecting change. Although the U.S. Preventive Services Task Force states that evidence is insufficient to recommend for or against skin cancer screening, a full-body skin exam seems to be a simple, practical way to approach reducing the mortality rate for skin cancer. Depending on the presence or absence of risk factors, strategies for early detection can be individualized. This is particularly true for patients with clinically atypical moles (dysplastic nevi) and those with a personal history of melanoma. For these individuals, surveillance should be performed by the dermatologist and include total-body photography and dermoscopy where appropriate. Individuals with three or more primary melanomas and families with at least one invasive melanoma and two or more cases of melanoma and/or pancreatic cancer among first- or second-degree relatives on the same side of the family may benefit from genetic testing. Precancerous and in situ lesions should be treated early. Early detection of small tumors allows the use of simpler treatment modalities with higher cure rates and lower morbidity.

DIAGNOSIS

The goal is to identify a melanoma before it invades and life-threatening metastases have occurred. Early detection may be facilitated by applying the ABCDEs: *a*symmetry (benign lesions are usually symmetric); *b*order irregularity (most nevi have clear-cut borders); *c*olor variegation (benign lesions usually have uniform light or dark pigment); *d*iameter >6 mm (the size of a pencil eraser); and *e*volving (any change in size, shape, color, or elevation or new symptoms such as bleeding, itching, and crusting). In addition, any nevus that appears atypical and different from the rest of the nevi on that individual (an “ugly duckling”) should be considered suspicious.

The entire skin surface, including the scalp and mucous membranes, as well as the nails should be examined in each patient. Bright room illumination is important, and a hand lens is helpful for evaluating variation in pigment pattern. Any suspicious lesions should be biopsied, evaluated by a specialist, or recorded by chart and/or photography for follow-up. A focused method for examining individual lesions, dermoscopy, employs low-level magnification of the epidermis with polarized light and may allow a more precise visualization of patterns of pigmentation than is possible with the naked eye. Additional technologies, including in vivo confocal microscopy, multi- and hyper-spectral imaging, optical coherence tomography, gene expression panels, tape stripping, and electrical conductance methods have been developed and are being refined for improved early detection of melanoma.

Biopsy

Any pigmented cutaneous lesion that has changed in size or shape or has other features suggestive of malignant melanoma is a candidate for biopsy. An excisional biopsy with 1- to 3-mm margins is suggested though excision can be accomplished tangentially or in a fusiform fashion. This facilitates pathologic assessment of the lesion, permits accurate measurement of thickness if the lesion is melanoma, and constitutes definitive treatment if the lesion is benign. For lesions that are large or on anatomic sites where excisional biopsy may not be feasible (such as the face, hands, and feet), an incisional biopsy through the most nodular or darkest area of the lesion is acceptable. Incisional biopsy does not appear to facilitate the spread of melanoma. For suspicious lesions, every attempt should be made to preserve the ability to assess the deep and peripheral margins and to perform immunohistochemistry. Shave, saucerization or tangential biopsies are an acceptable alternative, particularly if the suspicion of malignancy is low. They should be deep enough to include the deepest component of the entire lesion and any pigment at the base of the lesion should be removed and

included with the biopsy specimen. The biopsy should be read by a pathologist experienced in pigmented lesions, and the report should include Breslow thickness, mitotic rate, presence or absence of ulceration and lymphatic invasion, microsatellitosis and peripheral and deep margin status. Breslow thickness is the greatest thickness of a primary cutaneous melanoma measured on the slide from the top of the epidermal granular layer, or from the ulcer base, to the bottom of the tumor. To distinguish melanomas from benign nevi in challenging cases, fluorescence in situ hybridization (FISH) with multiple probes and comparative genome hybridization (CGH) can be helpful. Gene expression profiling assays have been developed to enhance diagnosis but are not yet widely applied.

CLASSIFICATION AND PATHOGENESIS

Clinical

The features of five major types of cutaneous melanoma are described in **Table 72-1**. In *superficial spreading melanoma*, *lentigo maligna melanoma*, and *acral lentiginous melanoma*, the lesion has a period of superficial (so-called radial) growth during which it increases in size but does not penetrate deeply. It is during this period that the melanoma is most capable of being cured by surgical excision. A fourth type—*nodular melanoma*—does not have a recognizable radial growth phase and usually presents as a deeply invasive lesion that is capable of early metastasis. Tumors that begin to penetrate deeply into the skin are in the so-called vertical growth phase. Melanomas with a radial growth phase are characterized by irregular and sometimes notched borders, variation in pigment pattern, and variation in color. A fifth type of melanoma, *desmoplastic melanoma*, is associated with a fibrotic response, neural invasion, and a greater tendency for local recurrence. Occasionally, melanomas appear clinically to be amelanotic, in which case the diagnosis is established microscopically after biopsy.

TABLE 72-1

Major Histologic Subtypes of Malignant Melanoma

Type	Site	Average Age at Diagnosis, Years	Appearance
Lentigo maligna melanoma	Sun-exposed surfaces, particularly malar region and temple	70	In flat portions, brown and tan predominate, but whitish gray occasionally present; in nodules, reddish brown, bluish gray, bluish black
Superficial spreading melanoma	Any site (more common on upper back and, in women, lower legs)	40–50	Brown mixed with bluish red, bluish black, reddish brown, and often whitish pink, and the border of lesion is at least in part visibly and/or palpably elevated
Nodular melanoma	Any	40–50	Reddish blue (purple) or bluish black; either uniform in color or mixed with brown or black
Acral lentiginous melanoma	Palm, sole, nail bed, mucous membrane	60	In flat portions, dark brown predominantly; in raised lesions (plaques), brown-black or blue-black predominantly
Desmoplastic melanoma	Any site (more common head and neck)	60	Highly variable, mimics other lesions; pigmentation is frequently absent

Although these subtypes are clinically and histopathologically distinct, this classification has minimal prognostic value and histologic subtype is not part of American Joint Committee on Cancer (AJCC) staging. Characterizing the genomic and mutational profiles of melanoma has become increasingly common and can reflect the mechanisms of tumorigenesis. These molecular classifications inform treatment and surveillance strategies.

Genomic

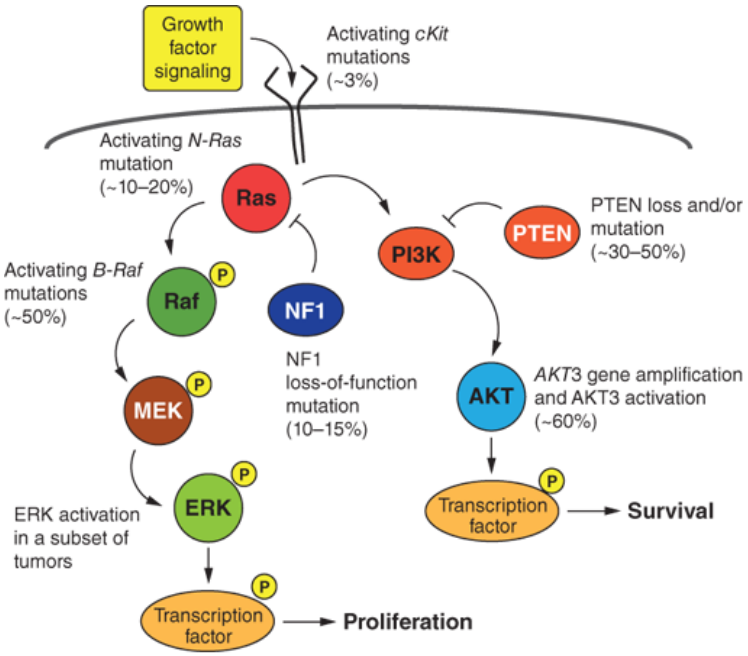
Considerable evidence from epidemiologic and molecular studies indicate that cutaneous melanomas arise via multiple causal pathways. There are both environmental and genetic components (susceptibility genes discussed earlier), and the major environmental factor in cutaneous melanomagenesis is sun exposure. The major effect of UV solar radiation is to cause genetic changes in the skin. However, it also impairs cutaneous immune function, increases the production of growth factors, and induces the formation of DNA-damaging reactive oxygen species that affect keratinocytes and melanocytes.

The advent of next-generation sequencing (NGS) has led to whole exome sequencing of hundreds of cutaneous melanomas derived from non-glabrous skin. This has revealed a very complex genetic landscape with genetic changes resulting from both germline (described earlier) and somatic mutations. Cutaneous melanomas have one of the highest somatic mutation rates (>10 mutations/Mb) compared to other cancers; the majority (76% primary tumors and 84% of metastatic melanomas) exhibit a mutation signature indicating UVR exposure. The mutation rate varies based on body site; melanomas arising in chronic sun-damaged skin harbor substantially more mutations than melanomas from non-sun-damaged skin.

Melanoma tumors can harbor thousands of mutations, but only a few are driver mutations; a mutation that is causally implicated in oncogenesis by virtue of a conferred growth advantage on the cancer cell. The driver mutations that have been identified for cutaneous melanoma are depicted in **Fig. 72-2**. As more melanomas are sequenced, more driver mutations have been identified. These mutations tend to be found in a smaller fraction of patients. Driver mutations often affect pathways that promote cell proliferation or inhibit normal pathways of apoptosis in response to DNA repair. They are often found in combination with mutations to the genetic susceptibility genes described earlier. The altered melanocytes accumulate DNA damage, and selection occurs for all the attributes that constitute the malignant phenotype: invasion, metastasis, and angiogenesis.

FIGURE 72-2

Major pathways involved in melanoma. The MAP kinase and PI3K/AKT pathways, which promote proliferation and inhibit apoptosis, respectively, are subject to mutations in melanoma. ERK, extracellular signal-regulated kinase; MEK, mitogen-activated protein kinase; NF-1; neurofibromatosis type 1 gene; PTEN, phosphatase and tensin homolog.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A recent report from the Cancer Genome Atlas (TCGA) has proposed a genomic classification of cutaneous melanoma based on the pattern of the most prevalent significantly mutated genes: BRAF, RAS, NF-1, and triple-WT (wild type). Distinct patterns of DNA mutations can vary with the site of origin and can be independent of the histologic subtype of the tumor. Thus, although the genetic landscape of melanoma is complex, and continues to evolve, the overall pattern of mutation, amplification, and loss of cancer genes indicates they have convergent effects on key biochemical pathways involved in proliferation, senescence, and apoptosis. An advantage of this classification is that these mutations can be used to select therapy.

The *p16* mutation that affects cell cycle arrest and the *ARF* mutation that results in defective apoptotic responses to genotoxic damage were described earlier. The proliferative pathways affected were the mitogen-activated protein (MAP) kinase and phosphatidylinositol 3' kinase/AKT pathways **(Fig. 72-2)**. *RAS* and *BRAF*, members of the MAP kinase pathway, which classically mediates the transcription of genes involved in cell proliferation and survival, undergo somatic mutation in melanoma and thereby generate potential therapeutic targets. *N-RAS* is mutated in ~20% of melanomas, and somatic activating *BRAF* mutations are found in most benign nevi and 40–50% of cutaneous melanomas. Neither mutation by itself appears to be sufficient to cause melanoma; thus, they often are accompanied by other mutations, such as TERT. The *BRAF* mutation is most commonly a point mutation (T→A nucleotide change) that results in a valine-to-glutamate amino acid substitution (V600E). V600E *BRAF* mutations are more common in younger patients and are present in most melanomas that arise on sites with intermittent sun exposure and are less common in melanomas from chronically sun-damaged skin. At present, *BRAF* mutations are the most important in therapeutic decision making in patients with advanced melanoma.

Melanomas also harbor mutations in *AKT* (primarily in *AKT3*) and *PTEN* (phosphatase and tensin homolog). *AKT* can be amplified, and *PTEN* may be deleted or undergo epigenetic silencing that leads to constitutive activation of the PI3K/AKT pathway and enhanced cell survival by antagonizing the intrinsic pathway of apoptosis. Loss of *PTEN*, which dysregulates AKT activity, and mutation of *AKT3* both prolong cell survival through inactivation of

BAD, BCL-2-antagonist of cell death, and activation of the forkhead transcription factor FOXO1, which leads to synthesis of prosurvival genes. A loss-of-function mutation in *NF1*, which can affect both MAP kinase and PI3K/AKT pathways, has been described in 10–15% of melanomas. In melanoma, these two signaling pathways (MAP kinase and PI3K/AKT) enhance tumorigenesis, chemoresistance, migration, and cell cycle dysregulation. Drugs that inhibit some of these pathways have been developed, and have proven to be effective therapeutic agents (see below).

PROGNOSTIC FACTORS

The most important prognostic factors for a newly diagnosed patient are incorporated in the staging classification **(Table 72-2)**. The best predictor of metastatic risk is Breslow thickness. The anatomic site of the primary is also prognostic; favorable sites are the forearm and leg (excluding the feet), and unfavorable sites include the scalp, hands, feet, and mucous membranes. In general, women with stage I or II disease have better survival than men, perhaps in part because of earlier diagnosis; women frequently have melanomas on the lower leg, where self-recognition is more likely and the prognosis is better. The effect of age is not straightforward. Older individuals, especially men >60, have worse prognoses, a finding that has been explained in part by a tendency toward later diagnosis (and thus thicker tumors) and in part by a higher proportion of acral melanomas in men. However, there is a greater risk of lymph node metastasis in young patients. Other important adverse factors recognized via the staging classification include high mitotic rate, presence of ulceration, microsatellite lesions and/or in-transit metastases, evidence of nodal involvement, elevated serum lactate dehydrogenase (LDH), and presence and site of distant metastases.

TABLE 72-2

Staging Criteria for Melanoma

Pathologic and TNM Stage	Thickness, mm	Ulceration	No. of Involved Lymph Nodes	Nodal Involvement	15-Year Survival Estimate (%)
0					98
Tis	In situ	No	0	None	
IA					92
T1a	<1	No, mitosis <1/mm	0	None	
IB					80
T1b	<1	Yes or mitosis >1/mm	0	None	
T2a	1.01–2	No	0	None	
IIA					62
T2b	1.01–2	Yes	0	None	
T3a	2.01–4	No	0	None	
IIB					51
T3b	2.01–4	Yes	0	None	
T4a	>4	No	0	None	
IIC					37
T4b	>4	Yes	0	None	
IIIA					68
N1a	T1-4a	No	1	Microscopic	
N2a	T1-4a	No	2 or 3	Microscopic	
IIIB					38
N1a	Any	Yes	1	Microscopic	
N2a	Any	Yes	2 or 3	Microscopic	
N1b	Any	Yes or no	1	Macroscopic	
N2b	Any	Yes or no	2 or 3	Macroscopic	
N2c	Any	Yes or no	In-transit metastases/satellites, no nodal involvement		
IIIC					22
N1b	Any	Yes or no	1	Macroscopic	

Pathologic and TNM Stage	Thickness, mm	Ulceration	No. of Involved Lymph Nodes	Nodal Involvement	15-Year Survival Estimate (%)
N2b	Any	Yes or no	2 or 3	Macroscopic	
N2c	Any	Yes or no	In-transit metastases/satellites, no nodal involvement		
N3	Any	Yes or no	4+ metastatic nodes, matted nodes or in-transit metastases/satellites, with metastatic nodes		
IV		Distant metastasis			<10
M1a		Skin, subcutaneous			
M1b		Lung			
M1c		Other visceral site			
		Elevated lactate dehydrogenase			

STAGING

Once the diagnosis of melanoma has been made, the tumor is staged to determine the prognosis and aid in treatment selection. The current melanoma staging criteria and estimated 15-year survival by stage are depicted in Table 72-2. The clinical stage is determined after the microscopic evaluation of the melanoma skin lesion and clinical and radiologic assessment. Pathologic staging also includes the microscopic evaluation of the regional lymph nodes obtained at sentinel lymph node biopsy or completion lymphadenectomy as indicated. All patients should have a complete history, with attention to symptoms that may suggest metastatic disease, such as malaise, weight loss, headaches, visual changes, and pain, and physical examination directed to the site of the primary melanoma, looking for persistent disease or for dermal or subcutaneous nodules that could represent satellite or in-transit metastases, and to the regional draining lymph nodes, CNS, liver, and lungs. A complete blood count (CBC), complete metabolic panel, and LDH should be performed. Although these rarely help uncover occult metastatic disease, a microcytic anemia would raise the possibility of bowel metastases, and the LDH, if elevated, should prompt a more extensive evaluation, including computed tomography (CT) scan or possibly a positron emission tomography (PET) (or CT/PET combined) scan. If signs or symptoms of metastatic disease are present, appropriate diagnostic imaging should be performed. At initial presentation, >80% of patients will have disease confined to the skin and a negative history and physical examination, in which case imaging is not indicated.

TREATMENT

TREATMENT

Melanoma

MANAGEMENT OF CLINICALLY LOCALIZED MELANOMA (STAGE I, II)

For a newly diagnosed cutaneous melanoma, wide surgical excision of the lesion with a margin of normal skin is necessary to remove all malignant cells and minimize possible local recurrence. The following margins are recommended for a primary melanoma: in situ, 0.5–1.0 cm; invasive up to 1 mm thick, 1 cm; >1.01–2 mm, 1–2 cm; and >2 mm, 2 cm. For lesions on the face, hands, and feet, strict adherence to these margins must give way to individual considerations about the constraints of surgery and minimization of morbidity. In all instances, however, inclusion of subcutaneous fat in the surgical specimen facilitates adequate thickness measurement and assessment of surgical margins by the pathologist. Topical imiquimod also has been used, particularly for lentigo maligna, in cosmetically sensitive locations.

Sentinel lymph node biopsy (SLNB) is a valuable staging tool that has replaced elective regional node dissection for the evaluation of regional nodal status. SLNB provides prognostic information and helps identify patients at high risk for relapse who may be candidates for adjuvant therapy. The initial (sentinel) draining node(s) from the primary site is (are) identified by injecting a blue dye and a radioisotope around the primary site. The sentinel node(s) then is (are) identified by inspection of the nodal basin for the blue-stained node and/or the node with high uptake of the radioisotope. The identified nodes are removed and subjected to careful histopathologic analysis with serial section using hematoxylin and eosin stains as well as immunohistochemical stains (e.g., S100, HMB45, and MelanA) to identify melanocytes.

Not every patient requires a SLNB. Patients whose melanomas are ≤0.75 mm thick have <5% risk of sentinel lymph node (SLN) disease and do not require a SLNB. Patients with tumors >1 mm thick generally undergo SLNB. For melanomas 0.76–1.0 mm thick, SLNB may be considered for lesions with high-risk features such as ulceration, high mitotic index, or lymphovascular invasion, but wide excision alone is the usual definitive therapy. Most other patients with clinically negative lymph nodes should undergo a SLNB. Patients whose SLNB is negative are spared a complete node dissection and its attendant morbidities, and can simply be followed, or based on the features of the primary melanoma, be considered for adjuvant therapy or a clinical trial. The current standard of care for all patients with a positive SLN is to perform a complete lymphadenectomy; however, complete lymph node dissection is not necessary for patients with lymph node micrometastases <1 mm. Patients with positive lymph nodes should be considered for adjuvant therapy with ipilimumab, interferon alpha or enrollment in a clinical trial.

MANAGEMENT OF REGIONALLY METASTATIC MELANOMA (STAGE III)

Melanomas may recur at the edge of the scar or graft, as satellite metastases, which are separate from but within 2 cm of the scar; as in-transit metastases, which are recurrences >2 cm from the primary lesion but not beyond the regional nodal basin; or, most commonly, as metastasis to a draining lymph node basin. Each of these presentations is managed surgically, following which there is the possibility of long-term disease-free survival. Isolated limb perfusion or infusion with melphalan and hyperthermia are options for patients with extensive cutaneous regional recurrences in an extremity. High complete response rates have been reported and significant palliation of symptoms can be achieved, but there is no change in overall survival. Other options for in-transit disease and distant skin and soft tissue metastases include topical immunotherapy and direct injection of melanoma lesions. Topical therapy with imiquimod has been useful for patients with low-volume dermal lesions. Historically, intralesional bacille Calmette-Guerin (BCG) has been used with high rates of regression of injected lesions and occasional regression of a distant, uninjected lesion. Talimogene laherparepvec is an engineered, oncolytic herpes simplex virus type 1 that is U.S. Food and Drug Administration (FDA) approved for injection of melanoma lesions that cannot be completely removed by surgery.

Patients rendered free of disease after surgery may be at high risk for a local or distant recurrence and should be considered for adjuvant therapy. Radiotherapy can reduce the risk of local recurrence after lymphadenectomy, but does not affect overall survival. Patients with large nodes (>3–4 cm), four or more involved lymph nodes, or extranodal spread on microscopic examination should be considered for radiation. Systemic adjuvant therapy is indicated primarily for patients with stage III disease, but high-risk, node-negative patients (>4 mm thick or ulcerated lesions), and patients with completely resected stage IV disease also may benefit.

Current treatment options include ipilimumab, interferon α2b (IFN-α2b) or investigational therapy. Ipilimumab is a fully human monoclonal antibody that blocks the immune checkpoint cytotoxic T-lymphocyte antigen-4 (CTLA-4) and augments antitumor immune responses. Treatment with ipilimumab 10 mg/kg IV every three weeks for four doses, then every three months for up to three years, improved survival of patients with high-risk stage III disease compared to placebo. IFN-α2b may be administered at high doses for one year or pegylated IFN can be administered at a lower dose for five years. The single study of ipilimumab documented a survival benefit whereas multiple trials of IFN have reported clear improvement in disease-free survival, but questionable improvement in overall survival. The two agents have not been compared directly. Ongoing clinical trials will address this issue as well as evaluate the potential value of other immunotherapies (e.g., PD-1/PD-L1 blocking agents) and targeted therapies in patients with BRAF mutated tumors in the adjuvant setting.

Both IFN and ipilimumab are accompanied by significant toxicity. For IFN, this may include a flulike illness, decline in performance status, and the development of depression. Side effects can be managed in most patients by appropriate treatment of symptoms, dose reduction, and treatment interruption. IFN may need to be discontinued prematurely because of unacceptable toxicity. The major side effects of ipilimumab are discussed below.

TREATMENT

TREATMENT

Metastatic Disease

At diagnosis, 84% patients with melanoma will have early-stage disease and 4% will present with metastases. Many others will develop metastases after initial therapy for loco-regional disease. The probability of recurrence is related to initial stage, ranging from <5% with stage IA to >90% for subsets of patients with stage IIIC disease at presentation. Patients with a history of melanoma who develop signs or symptoms suggesting recurrent disease should undergo restaging as described earlier. Distant metastases (stage IV) may involve any organ and commonly involve the skin and lymph nodes as well as viscera, bone, or the brain. The prognosis is better for patients with skin and subcutaneous metastases (M1a) than for lung (M1b) and worst for those with metastases to liver, bone, and brain (M1c). An elevated serum LDH is a poor prognostic factor and places the patient in stage M1c regardless of the site of the metastases (Table 72-2). Although historical data suggest that the 15-year survival of patients with melanoma is <10%; advances in targeted and immunotherapy have improved disease-free and overall survival, especially for patients with M1a and M1b disease.

The treatment for patients with stage IV melanoma has changed dramatically since 2011. FDA-approved agents include three immune T-cell checkpoint inhibitors, ipilimumab, nivolumab, and pembrolizumab, four oral agents that target the MAP kinase pathway: the BRAF inhibitors, vemurafenib and dabrafenib, the MEK inhibitors, trametinib and cobimetinib, and the oncolytic virus talimogene laherparepvec **(Table 72-3)**.

Surgery should be considered for patients with oligometastatic disease because they may experience long-term disease-free survival after metastasectomy. Patients with solitary metastases are the best candidates, but surgery can also be used for patients with metastases at more than one site if a complete resection of all sites can be achieved. Patients rendered free of disease can be considered for adjuvant therapy or a clinical trial because their risk of developing additional metastases is very high. Surgery can also be used as an adjunct to systemic therapy when for example, a few of many metastatic lesions prove resistant to immunotherapy.

IMMUNOTHERAPY

Interleukin 2 (IL-2 or aldesleukin) is used effectively to treat stage IV patients who have a good performance status. High-dose IL-2, which requires hospitalization in an intensive care unit–like setting, is administered by intravenous bolus doses over a 1-week cycle mainly at centers with experience managing IL-2-related toxicity. Treatment is continued until maximal benefit is achieved, usually 4–6 cycles distributed over 4–6 months to allow for recovery from toxicities between cycles. Long-term disease-free survival (probable cure) is observed in 5% of treated patients.

Checkpoint Blockade

Newer immunotherapies are based on an understanding of the control mechanisms of the normal immune response. Inhibitory receptors or checkpoints, including CTLA-4 and PD-1, are upregulated on T cells after engagement of the T-cell receptor by cognate tumor antigen in the context of the appropriate class I or II HLA molecules during the interaction between a T cell and antigen-presenting cell. An absolute requirement to ensure proper regulation of a normal immune response, the continued expression of inhibitory receptors during chronic infection (hepatitis, HIV) and in cancer patients leads to exhausted T cells with limited potential for proliferation, cytokine production, or cytotoxicity **(Fig. 72-3)**. Checkpoint blockade with an antagonistic monoclonal antibody results in improved T-cell function and eradication of tumor cells in preclinical animal models. Ipilimumab, a fully human IgG1 antibody that binds CTLA-4 and blocks inhibitory signals, was the first drug shown in a randomized trial to improve survival in patients with metastatic melanoma. A full course of therapy is four outpatient infusions of ipilimumab 3 mg/kg every 3 weeks. Although response rates are low (~10%), overall survival is improved.

Chronic T-cell activation also leads to induction of PD-1 on the surface of T cells. Expression of one of its ligands, PD-L1, on tumor cells can protect them from immune destruction (Fig. 72-3). Blockade of the PD-1:PD-L1 axis by IV administration of anti-PD-1 or anti-PD-L1 has substantial clinical activity in patients with advanced melanoma (and lung, renal, bladder and oral head and neck cancers as well as Hodgkin lymphoma) with significantly less toxicity than ipilimumab. The PD-1 blockers, nivolumab and pembrolizumab, have been approved to treat patients with advanced melanoma. Combination T-cell checkpoint therapy, blocking both inhibitory pathways with ipilimumab and nivolumab, leads to superior antitumor activity compared to treatment with either agent alone. Combined therapy with intravenous ipilimumab and nivolumab is administered in the outpatient setting every 3 weeks for 4 doses (induction), followed by nivolumab given every 2 weeks (maintenance) for up to one year. This regimen produces an objective response rate of 56% and enhanced survival compared to ipilimumab monotherapy. There may be subsets of patients, specifically those who have >5% expression of PD-1 on T cells in a melanoma biopsy sample, who derive a similar level of clinical benefit from nivolumab monotherapy.

The main benefit to patients from immune-based therapy is the durability of the responses achieved. The percentage of patients whose tumors regress following combination anti-CTLA-4 and anti-PD-1 immunotherapy is comparable to the response rate after targeted therapy (see below); however, the durability of immunotherapy-induced responses (>10 years in some cases with checkpoints and greater than 20 years in some patients after IL-2) appears to be superior to responses after targeted therapy and suggests that many of these patients have been cured.

T-cell checkpoint antibodies can also interfere with normal immune regulatory mechanisms, which may produce a novel spectrum of side effects. The most common immune-related adverse events were skin rash and diarrhea (sometimes severe, life-threatening colitis), but toxicity can involve most any organ (e.g., hypophysitis, hepatitis, nephritis, pneumonitis, myocarditis, neuritis). The severity and frequency of toxicity is greatest with combination T-cell checkpoint antibody therapy, followed by anti-CTLA-4 and then anti-PD-1 monotherapies. Vigilance, interruption of therapy and early intervention with steroids or other immunosuppressive agents, such as anti-tumor necrosis factor antibodies or mycophenolate mofetil, can mitigate toxicity and prevent permanent organ damage. The management of drug-induced toxicity with immunosuppressive agents does not appear to interfere with antitumor activity. The use of T-cell checkpoint antibodies for metastatic melanoma has become commonplace, but there is controversy about whether all patients need combined anti-CTLA-4 and anti-PD-1, whether biomarkers can be used to select patients who may benefit from anti-PD-1 alone and the best sequence of targeted and immunotherapy in patients who have a BRAF mutation. There is also a significant economic impact with the cost of combination anti-CTLA-4 and anti-PD-1, which must be placed in the context of the survival benefit.

TARGETED THERAPY

The high frequency of oncogenic mutations in the RAS-RAF-MEK-ERK pathway, which delivers proliferation and survival signals from the cell surface to the cytoplasm and nucleus, has led to the development of inhibitors to BRAF and MEK. RAF and MEK inhibitors of the MAP kinase pathway can induce regression of melanomas that harbor a *BRAF* mutation. Two BRAF inhibitors, vemurafenib and dabrafenib, have been approved for the treatment of patients whose stage IV melanomas harbor a mutation at position 600 in *BRAF*. Monotherapy with BRAF inhibitors has been supplanted with combined BRAF and MEK inhibition to address the rapid adaptation of the majority of melanomas that use MAP kinase pathway reactivation to facilitate growth when BRAF is inhibited. Combined therapy with BRAF and MEK inhibitors (dabrafenib and trametinib or vemurafenib with cobimetinib) improved progression-free survival compared to monotherapy with a BRAF inhibitor. The durability of responses following combined therapy is superior to monotherapy and survival is also enhanced. Long-term results of inhibition of the MAP kinase pathway are not yet available, but the major limitation of

both monotherapy and combined therapy appears to be the acquisition of resistance; the vast majority of patients relapse and eventually die. The mechanisms of resistance are diverse and reflect the genomic heterogeneity of melanoma; however, most instances involve reactivation of the MAPK pathway, often through RAS mutations or mutant BRAF amplification. Patients who develop resistance to BRAF and MEK inhibition are candidates for immunotherapy or clinical trials.

Targeted therapy is accompanied by manageable side effects that differ from those experienced during immunotherapy or chemotherapy. A class-specific side effect of BRAF inhibition is the development of numerous skin lesions, some of which are well-differentiated squamous cell skin cancers (SCC) (seen in up to 25% of patients). These hyperproliferative lesions are believed to be due to paradoxical activation of the MAPK pathway resulting from BRAF inhibitor-mediated changes in *BRAF*-wild type cells. The paradoxical activation is blocked by the MEK inhibitor, which explains why these lesions occur much less frequently during combined therapy. Patients should be co-managed with a dermatologist as these skin cancers will need excision. Metastases of the treatment-induced SCCs have not been reported, and BRAF and MEK inhibitors can be continued safely following simple excision. Cardiac and ocular toxicities, although infrequent, can occur with BRAF and MEK inhibitors and require medical evaluation and management.

Activating mutations in the c-kit receptor tyrosine kinase are found in a minority of cutaneous melanomas with chronic sun damage, but are more common in mucosal and acral lentiginous subtypes. Overall, the number of patients with *c-kit* mutations is exceedingly small, but when present, they are similar to those found in gastrointestinal stromal tumors; melanomas with activating *c-kit* mutations can have clinically meaningful responses to imatinib. The probability of objective response in patients whose melanomas harbor a *c-kit* mutation is 29%. *N-RAS* mutations occur in 15–20% of melanomas. At present, there are no effective targeted agents for these patients, but MEK inhibitors are being investigated in clinical trials.

CHEMOTHERAPY

No chemotherapy regimen has ever been shown to improve survival of patients with metastatic melanoma. The advances in immunotherapy and targeted therapy have relegated chemotherapy to the palliation of symptoms. Drugs with antitumor activity include dacarbazine (DTIC) or its orally administered analog temozolomide (TMZ), cisplatin and carboplatin, the taxanes (paclitaxel alone or albumin-bound), and carmustine (BCNU), which have reported response rates of 12–20%.

INITIAL APPROACH TO PATIENT WITH METASTATIC DISEASE

Upon diagnosis of stage IV disease, a sample of the patient's tumor should be submitted for molecular testing to determine whether a druggable mutation (e.g., *BRAF* and c-kit) is present. Analysis of a metastatic lesion biopsy (if possible) is preferred, but any sample will suffice because there is little discordance between primary and metastatic lesions. Treatment algorithms start with the tumor's *BRAF* status. For *BRAF* wild-type tumors, immunotherapy is recommended. For patients whose tumors harbor a *BRAF* mutation, initial therapy with either combination BRAF and MEK inhibitors or immunotherapy is acceptable. Combined therapy with BRAF and MEK inhibitors is favored for patients with rapidly growing and symptomatic disease when a *BRAF* mutation is present. The sequence of immunotherapy and targeted therapy that confers the greatest survival benefit in patients with minimally symptomatic melanoma is not yet known, but ongoing randomized phase III trials should answer this important question. Despite improvements in therapy, the majority of patients with metastatic melanoma are not cured so enrollment in a clinical trial is always an important consideration, even for previously untreated patients.

Since most patients with stage IV disease will eventually experience tumor progression despite therapy and many, because of extensive disease burden, poor performance status, or concomitant illness, will be poor candidates for therapy, the timely integration of palliative care and hospice should be a major focus of care. Future advances in the management of melanoma will likely include biomarkers to select the optimal combination and sequence of agents or to identify patients who are unlikely to respond to extant therapies and for whom clinical trials should be considered. New therapeutic agents could include T-cell co-stimulatory antibodies, engineered T cells, oncolytic viruses and possibly vaccines to prevent melanoma development or recurrence.

FOLLOW-UP

Skin examination and surveillance at least once a year are recommended for all patients with melanoma. Routine blood work and imaging for patients with stage IA–IIA disease is not recommended unless symptoms are present. In general, because there is no survival benefit to patients, routine surveillance diagnostic imaging is not recommended for patients with higher stage disease and imaging should be reserved for patients with signs or symptoms of recurrent disease. For stage-specific recommendations, please consult the National Comprehensive Cancer Network (NCCN) guidelines (see Further Reading).

TABLE 72-3

Treatment Options for Metastatic Melanoma

Surgery: Metastasectomy for small number of lesions

Immunotherapy:

Interleukin 2

Immune checkpoint blockade

- Anti-CTLA-4: ipilimumab
- Anti-PD-1: nivolumab, pembrolizumab
- Combined ipilimumab and nivolumab

Experimental

- Anti-PD-L1

Molecular targeted therapy:

BRAF inhibitor: vemurafenib, dabrafenib

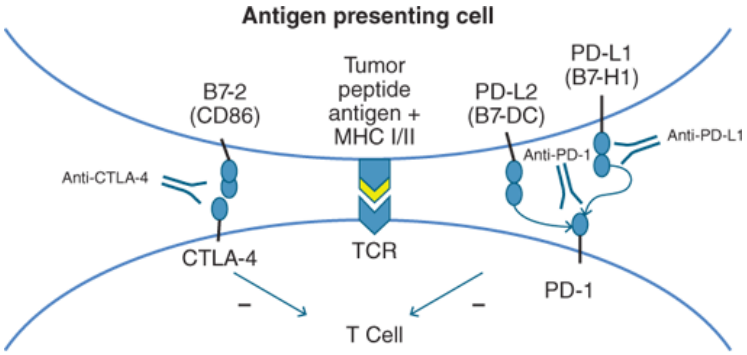
MEK inhibitor: trametinib, cobimetinib

Oncolytic virus: talimogene laherpaepvec

Chemotherapy: dacarbazine, temozolomide, paclitaxel, albumin-bound paclitaxel, carboplatin

FIGURE 72-3

Inhibitory regulatory pathways that influence T-cell function, memory and lifespan after engagement of the T-cell receptor by tumor peptide antigen presented by antigen-presenting cells in the context of MHC I/II. CTLA-4 and PD-1 are members of the CD28 family and their inhibitory effects can be mitigated by antagonistic antibodies to the receptors or ligand resulting in enhanced T-cell function and anti-tumor effects. TCR: T-cell receptor, MHC: Major Histocompatibility Complex, CTLA-4: Cytotoxic T-Lymphocyte Antigen-4, PD-1: Programmed Death-1, PD-L1: Programmed Death Ligand-1, PD-L2: Programmed Death Ligand-2.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

NONMELANOMA SKIN CANCER

Nonmelanoma skin cancer (NMSC) is the most common cancer in the United States. Although tumor registries do not routinely gather data on the incidence of basal cell and squamous cell skin cancers, it is estimated that the annual incidence is 1.5–2 million cases in the United States. Basal cell carcinomas (BCCs) account for 70–80% and squamous cell carcinomas (SCCs) ~20% of NMSCs, respectively. SCCs are more significant because they metastasize and account for 2400 deaths annually. There has also been an increase in the incidence of nonepithelial skin cancer, especially Merkel cell carcinoma, with nearly 5000 new diagnoses and 3000 deaths annually.

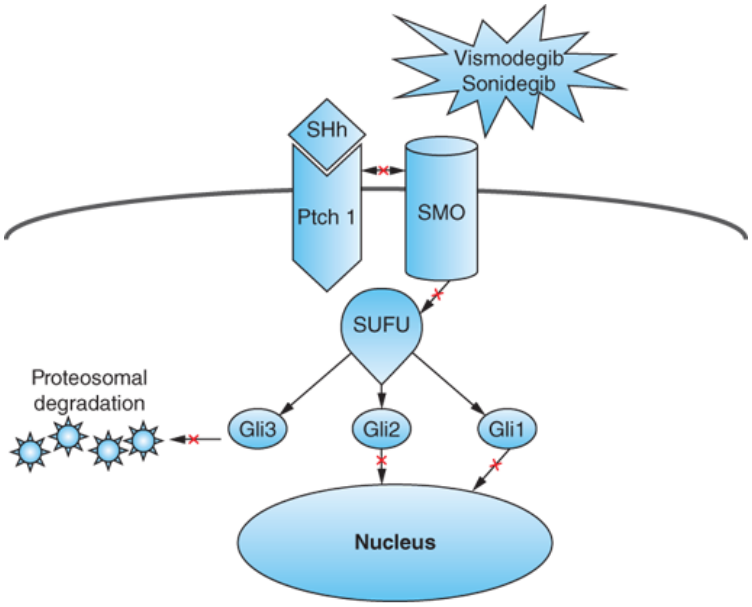
PATHOPHYSIOLOGY AND ETIOLOGY

The most significant cause of BCC and SCC is UV radiation, whether through direct exposure to sunlight or by artificial UV light sources (tanning beds). Both UVA and UVB light can induce DNA damage. The DNA damage can be repaired or lead to cell death. The mechanism for DNA repair involves excising damaged nucleotides. Inherited disorders of DNA repair, such as xeroderma pigmentosum, are associated with a greatly increased incidence of skin cancer and help to establish the link between UV-induced DNA damage, inadequate DNA repair, and skin cancer. The genes damaged most commonly by UV in BCC involve the hedgehog signaling pathway (Hh) and lead to basal cell proliferation. This is usually the result of loss of function of the tumor-suppressor patched homolog 1 (*PTCH1*), which normally inhibits the signaling of smoothened homolog (*SMO*). Aberrant *PTCH1* signaling is

propagated by the nuclear transcription factors Gli1 and Gli2, which are salient in the development of BCC. Two oral SMO inhibitors, vismodegib and sonidegib, have been approved by the FDA to treat advanced inoperable or metastatic BCC and locally advanced BCC that has recurred following surgery or RT, respectively **(Fig. 72-4)**. Vismodegib also reduces the incidence of BCC in patients with basal cell nevus syndrome who have *PTCH1* mutations, affirming the importance of Hh in the onset of BCC.

FIGURE 72-4

Inhibition of the hedgehog (Hh) pathway. The Hh pathway promotes gene transcription and is important in the pathogenesis of BCC. Normally, one of three Hh ligands (sonic [SHh], Indian, or desert) binds to patched homolog 1 (PTCH1), causing its degradation and release of smoothened homolog (SMO). SMO release represses another regulatory protein called suppressor of fused (SUFU). SUFU normally binds glioblastoma transcription factors Gli1, Gli2, and Gli3. SUFU repression allows Gli1 and Gli2 to translocate to the nucleus and promote gene transcription. Vismodegib and sonidegib are SMO antagonists. Antagonizing SMO decreases the interaction between SMO and PTCH1, resulting in decreased Hh pathway signaling, gene transcription, and cell division. The Hh pathway events inhibited by vismodegib and sonidegib are indicated in red.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In SCC, *p53* and *N-RAS* are commonly affected. There is a dose-response relationship between tanning bed use and the incidence of skin cancer. As few as four tanning bed visits per year confers a 15% increase in BCC and an 11% increase in SCC and melanoma. Tanning bed use as a teenager or young adult confers greater risk than comparable exposure in older individuals. Other associations include blond or red hair, blue or green eyes, a tendency to sunburn easily, and an outdoor occupation. The incidence of NMSC increases with decreasing latitude. Most tumors develop on sun-exposed areas of the head and neck. The risk of lip or oral SCC is increased with cigarette smoking and, like SCC of the ear, has a worse prognosis than that seen on other body sites. Human papillomaviruses and UV radiation may act as co-carcinogens.

Chronically immunosuppressed solid organ transplant recipients have a 65-fold increase in SCC and a 10-fold increase in BCC. The frequency of skin cancer is proportional to the level and duration of immunosuppression and the extent of sun exposure before and after transplantation. SCCs in this population also demonstrate higher rates of local recurrence, metastasis, and mortality. Tumor necrosis factor (TNF) antagonist therapy of inflammatory bowel disease and autoimmune disorders, such as rheumatoid and psoriatic arthritis, may also confer an increased risk of NMSC.

Other risk factors include HIV infection, ionizing radiation, thermal burn scars, and chronic ulcerations. Albinism, xeroderma pigmentosum, Muir-Torre syndrome, Rombo's syndrome, Bazex-Dupré-Christol syndrome, dyskeratosis congenita, and basal cell nevus syndrome (Gorlin syndrome) also increase the incidence of NMSC.

CLINICAL PRESENTATION

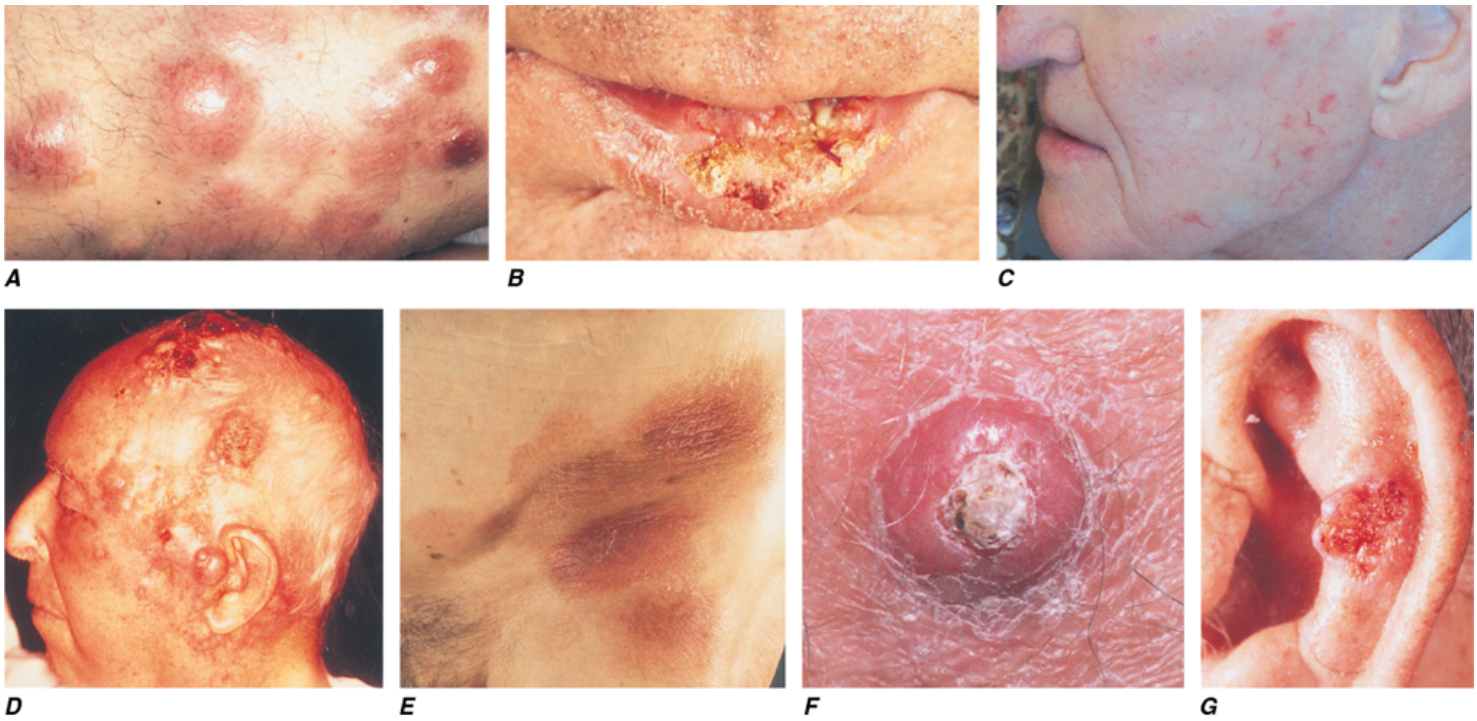
Basal Cell Carcinoma

BCC arises from epidermal basal cells. The least invasive of BCC subtypes, superficial BCC, consists of often subtle, erythematous scaling plaques that slowly enlarge and are most commonly seen on the trunk and proximal extremities **(Fig. 72-5)**. This BCC subtype may be confused with benign inflammatory dermatoses, especially nummular eczema and psoriasis or premalignant actinic keratoses. BCC also can present as a small, slowly growing pearly nodule, often with tortuous telangiectatic vessels on its surface, rolled borders, and a central crust (nodular BCC). The occasional presence of melanin in this variant of nodular BCC (pigmented BCC) may lead to confusion with melanoma. Morpheaform (fibrosing), infiltrative, and micronodular BCC, the most invasive and potentially aggressive subtypes, manifest as solitary, flat or slightly depressed, indurated whitish, yellowish,

or pink scar-like plaques. Borders are typically indistinct, and lesions can be subtle; thus, delay in treatment is common, and tumors can be more extensive than expected clinically.

FIGURE 72-5

Cutaneous neoplasms. ***A.*** Non-Hodgkin's lymphoma involves the skin with typical violaceous, "plum-colored" nodules. ***B.*** Squamous cell carcinoma is seen here as a hyperkeratotic crusted and somewhat eroded plaque on the lower lip. Sun-exposed skin in areas such as the head, neck, hands, and arms represent other typical sites of involvement. ***C.*** Actinic keratoses consist of hyperkeratotic erythematous papules and patches on sun-exposed skin. They arise in middle-aged to older adults and can undergo malignant transformation. ***D.*** Metastatic carcinoma to the skin is characterized by inflammatory, often ulcerated dermal nodules. ***E.*** Mycosis fungoides is a cutaneous T-cell lymphoma, and plaque-stage lesions are seen in this patient. ***F.*** Keratoacanthoma is a low-grade squamous cell carcinoma that presents as an exophytic nodule with central keratinous debris. ***G.*** This basal cell carcinoma shows central ulceration and a pearly, rolled telangiectatic tumor border.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Squamous Cell Carcinoma

Primary *cutaneous SCC* is a malignant neoplasm of keratinizing epidermal cells. SCC has a variable clinical course, ranging from indolent to rapid growth, with the potential to metastasize to regional and distant sites. Commonly, SCC appears as an ulcerated erythematous nodule or superficial erosion on sun-exposed skin of the head, neck, trunk, and extremities **(Fig. 72-5)**. It may also appear as a banal, firm, dome-shaped papule or rough-textured plaque. It is commonly mistaken for a wart or callous when the inflammatory response to the lesion is minimal. Visible overlying telangiectasias are uncommon, although dotted or coiled vessels are a hallmark of SCC when viewed through a dermatoscope. The margins of this tumor may be ill defined, and fixation to underlying structures may occur ("tethering").

A very rapidly growing but low-grade form of SCC, called keratoacanthoma (KA), typically appears as a large dome-shaped papule with a central keratotic crater. Some KAs regress spontaneously without therapy, but because progression to metastatic SCC has been documented, KAs should be treated in the same manner as other types of cutaneous SCC. KAs occur in 15–25% of patients receiving monotherapy with a BRAF inhibitor.

Actinic keratoses and *cheilitis* (actinic keratoses on the lip), both premalignant forms of SCC, present as hyperkeratotic papules on sun-exposed areas. Malignant transformation occurs in 0.25 to 20% of untreated lesions. SCC in situ, also called *Bowen's disease*, is the intraepidermal form of SCC and usually presents as a scaling, erythematous plaque. SCC in situ most commonly arises on sun-damaged skin, but can occur anywhere on the body. Bowen's disease occurring secondary to infection with human papillomavirus (HPV) can arise on skin with minimal or no prior sun exposure, such as the buttock or posterior thigh. Treatment of premalignant and in situ lesions reduces the subsequent risk of invasive disease.

NATURAL HISTORY

Basal Cell Carcinoma

The natural history of BCC is that of a slowly enlarging, locally invasive neoplasm. The degree of local destruction and risk of recurrence vary with the size, duration, location, and histologic subtype of the tumor. Location on the central face, ears, or scalp may portend a higher risk. Small nodular, pigmented, cystic, or superficial BCCs respond well to most treatments. Large lesions and micronodular, infiltrative, and morpheaform subtypes may be more aggressive. The metastatic potential of BCC is low (0.0028–0.1%) in immunocompetent patients, but the risk of recurrence or a new primary NMSC is about 40% over 5 years.

Squamous Cell Carcinoma

The natural history of SCC depends on tumor and host characteristics. Tumors arising on sun-damaged skin have a lower metastatic potential than do those on non-sun-exposed areas. Cutaneous SCC metastasizes in 0.3–5.2% of individuals, most frequently to regional lymph nodes. Tumors occurring on the lower lip and ear develop regional metastases in 13 and 11% of patients, respectively, whereas the metastatic potential of SCC arising in scars, chronic ulcerations, and genital or mucosal surfaces is higher. Recurrent SCC has a much higher potential for metastatic disease, approaching 30%. Large, poorly differentiated, deep tumors with perineural or lymphatic invasion, multifocal tumors, and those arising in immunosuppressed patients often behave aggressively.

TREATMENT

TREATMENT

Basal Cell and Squamous Cell Carcinoma

BASAL CELL CARCINOMA

Treatments used for BCC include electrodesiccation and curettage (ED&C), excision, cryosurgery, radiation therapy (RT), laser therapy, Mohs micrographic surgery (MMS), topical 5-fluorouracil, photodynamic therapy (PDT), and topical immunomodulators such as imiquimod. The choice of therapy depends on tumor characteristics including depth and location, patient age, medical status, and patient preference. ED&C remains the most commonly employed method for superficial, minimally invasive nodular BCCs and low-risk tumors (e.g., a small tumor of a less aggressive subtype in a favorable location). Wide local excision with standard margins is usually selected for invasive, ill-defined, and more aggressive subtypes of tumors, or for cosmetic reasons. MMS, a specialized type of surgical excision that provides the best method for tumor removal while preserving uninvolved tissue, is associated with cure rates >98%. It is the preferred modality for lesions that are recurrent, in high-risk or cosmetically sensitive locations (including recurrent tumors in these locations), and for which maximal tissue conservation is critical (e.g., the eyelids, lips, ears, nose, and digits). RT can cure patients not considered surgical candidates and can be used as a surgical adjunct in high-risk tumors. Imiquimod can be used to treat superficial and smaller nodular BCCs, although it is not FDA-approved for nodular BCC. Topical 5-fluorouracil therapy should be limited to superficial BCC. PDT, which uses selective activation of a photoactive drug by visible light, has been used in patients with numerous tumors. Intralesional therapy (5-fluorouracil or IFN) can also be employed. Like RT, it remains an option for selected patients who cannot or will not undergo surgery. Systemic therapy with an SMO inhibitor, vismodegib or sonidegib, is indicated for patients with metastatic or advanced BCC that has recurred after local therapy and who are not candidates for surgery or radiation. Targeted therapy with SMO antagonists does not cure patients with BCC, but induces regression in approximately 50% of patients with a median duration of response greater than 9 months.

SQUAMOUS CELL CARCINOMA

Therapy for cutaneous SCC should be based on the size, location, histologic differentiation, patient age, and functional status. Surgical excision and MMS are standard treatments. Cryosurgery and ED&C have been used for premalignant lesions and small, superficial, in situ primary tumors. Lymph node metastases are treated with surgical resection, RT, or both. Combination chemotherapy that includes cisplatin, and intralesional and systemic 5-fluorouracil, and cetuximab are also options for palliation in patients with advanced disease. SCC and keratoacanthomas that develop in patients receiving BRAF-targeted therapy should be excised, after which BRAF therapy can be continued.

PREVENTION

The general principles for prevention are those described for melanoma earlier. Unique strategies for NMSC include active surveillance for patients on immunosuppressive medications or BRAF-targeted therapy. Chemoprophylaxis using synthetic retinoids and immunosuppression reduction when possible may be useful in controlling new lesions and managing patients with multiple tumors. Field therapy with topical 5-FU, ingenol mebutate, or imiquimod can reduce transformation to SCC in patients with severe sun damaged skin and numerous premalignant actinic keratoses.

OTHER NONMELANOMA CUTANEOUS MALIGNANCIES

Neoplasms of cutaneous adnexae and sarcomas of fibrous, mesenchymal, fatty, and vascular tissues make up the remaining 1–2% of NMSCs.

Merkel cell carcinoma (MCC) is a neural crest–derived highly aggressive malignancy with mortality rates approaching 33% at 3 years. An oncogenic Merkel cell polyomavirus (MCPyV) is present in 80% of tumors and UV exposure also increases the incidence of this malignancy. In patients with MCPyV+ tumors, there is inactivation of tumor suppressor genes, specifically the p53 transcription factor and retinoblastoma protein (Rb). In addition, the viral

large T antigen is expressed on tumor cells and many patients have detectable cellular or humoral immune responses to polyoma viral proteins, although this immune response is insufficient to eradicate the malignancy. Survival depends on extent of disease: 90% survive with local disease, 52% with nodal involvement, but only 10% with distant disease. MCC incidence tripled over the last 20 years with an estimated 1600 cases per year in the United States. Immunosuppression increases the incidence and diminishes the prognosis compared to patients with no immunosuppression. MCC lesions typically present as an asymptomatic rapidly expanding bluish-red/violaceous tumor on sun-exposed skin of older white patients. Treatment is surgical excision with sentinel lymph node biopsy for accurate staging in patients with localized disease, often followed by adjuvant RT. Patients with extensive disease can be offered systemic chemotherapy; however, there is no survival benefit. Immunotherapy using anti-PD-1 (pembrolizumab) was associated with a 56% response rate with a progression-free survival at 6 months of 67%. Tumor regression occurred in MCPyV positive and negative tumors. A monoclonal antibody targeting anti-PD-L1 known as avelumab showed objective responses in 33% of patients with advanced MCC that was durable in 82% of the responders. The U.S. FDA approved avelumab for the treatment of patients with metastatic MCC in April 2017. Whenever possible a clinical trial should be considered for patients with this rare but aggressive NMSC.

Extramammary Paget's disease is an uncommon apocrine malignancy arising from stem cells of the epidermis that are characterized histologically by the presence of Paget cells. These tumors present as moist erythematous patches on anogenital or axillary skin of the elderly. Outcomes are generally good with surgery, and 5-year disease-specific survival is ~95% with localized disease. Advanced age and extensive disease at presentation confer diminished prognosis. RT or topical imiquimod can be considered for more extensive disease. Local management may be challenging because these tumors often extend far beyond clinical margins; surgical excision with MMS has the highest cure rates. Similarly, MMS is the treatment of choice in other rare cutaneous tumors with extensive subclinical extension such as *dermatofibromasarcoma protuberans*.

Kaposi's sarcoma (KS) is a soft tissue sarcoma of vascular origin that is induced by the human herpesvirus 8. The incidence of KS increased dramatically during the AIDS epidemic, but has now decreased tenfold with the institution of highly active antiretroviral therapy.

ACKNOWLEDGMENT

Steven Kolker, MD, provided valued feedback and suggested improvements to this chapter.

FURTHER READING

The Cancer Genome Atlas Network: Genomic classification of cutaneous melanoma. Cell 161:1681, 2015. [PubMed: 26091043]

Guo J et al: Chinese guidelines on the diagnosis and treatment of melanoma (2015 edition). Ann Transl Med 3:322, 2015. [PubMed: 26734632]

International Agency for Research on Cancer: GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012. Available from *http://globocan.iarc.fr/default.aspx.* Accessed December 19, 2016.

Larkin J et al: Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. N Engl J Med 373:23, 2015. [PubMed: 26027431]

Leiter U et al: Complete lymph node dissection versus no dissection in patients with sentinel lymph node biopsy positive melanoma (DeCOG-SLT): A multicentre, randomised, phase 3 trial. Lancet Oncol 17:757, 2016. [PubMed: 27161539]

National Comprehensive Cancer Network: NCCN clinical practice guidelines in oncology (NCCN guidelines): Melanoma. Available from *https://www.nccn.org/professionals/physician_gls/pdf/melanoma.pdf.* Accessed December 20, 2016.

Robert C et al: Improved overall survival in melanoma with combined debrafenib and trametinib. N Engl J Med 372:30, 2015. [PubMed: 25399551]

Shain AH, Bastian BC: From melanocytes to melanomas. Nat Rev Cancer 16:345, 2016. [PubMed: 27125352]

Wu YP et al: A systematic review of interventions to improve adherence to melanoma preventive behaviors for individuals at elevated risk. Prev Med 88:153, 2016. [PubMed: 27090434]

Zhang T et al: The genomic landscape of cutaneous melanoma. Pigment Cell Melanoma Res 29:266, 2016. [PubMed: 26833684]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library

Silverchair

Chapter 73: Head and Neck Cancer

Everett E. Vokes

INTRODUCTION

Epithelial carcinomas of the head and neck arise from the mucosal surfaces in the head and neck and typically are squamous cell in origin. This category includes tumors of the paranasal sinuses, the oral cavity, and the nasopharynx, oropharynx, hypopharynx, and larynx. Tumors of the salivary glands differ from the more common carcinomas of the head and neck in etiology, histopathology, clinical presentation, and therapy. They are rare and histologically highly heterogeneous. **Thyroid malignancies are described in Chap. 378.**

INCIDENCE AND EPIDEMIOLOGY

The number of new cases of head and neck cancers (oral cavity, pharynx, and larynx) in the United States was estimated at 48,330 in 2016, accounting for about 3% of adult malignancies; estimated deaths were 13,190. The worldwide incidence exceeds half a million cases annually. In North America and Europe, the tumors usually arise from the oral cavity, oropharynx, or larynx. The incidence of oropharyngeal cancers is increasing in recent years, especially in Western countries. Nasopharyngeal cancer is more commonly seen in the Mediterranean countries and in the Far East, where it is endemic in some areas.

ETIOLOGY AND GENETICS

Alcohol and tobacco use are the most significant risk factors for head and neck cancer, and when used together, they act synergistically. Smokeless tobacco is an etiologic agent for oral cancers. Other potential carcinogens include marijuana and occupational exposures such as nickel refining, exposure to textile fibers, and woodworking.

Some head and neck cancers have a viral etiology. Epstein-Barr virus (EBV) infection is frequently associated with nasopharyngeal cancer, especially in endemic areas of the Mediterranean and Far East. EBV antibody titers can be measured to screen high-risk populations and are under investigation to monitor treatment response. Nasopharyngeal cancer has also been associated with consumption of salted fish and in-door pollution.

In Western countries, the human papilloma virus (HPV) is associated with a rising incidence of tumors arising from the oropharynx, that is, the tonsillar bed and base of tongue. Over 50% of oropharyngeal tumors are caused by HPV in the United States, and in many urban centers this proportion is even higher. HPV 16 is the dominant viral subtype, although HPV 18 and other oncogenic subtypes are seen as well. Alcohol- and tobacco-related cancers, on the other hand, have decreased in incidence. HPV-related oropharyngeal cancer occurs in a younger patient population and is associated with increased numbers of sexual partners and oral sexual practices. It is associated with a better prognosis, especially for nonsmokers.

Dietary factors may contribute. The incidence of head and neck cancer is higher in people with the lowest consumption of fruits and vegetables. Certain vitamins, including carotenoids, may be protective if included in a balanced diet. Supplements of retinoids, such as *cis*-retinoic acid, have not been shown to prevent head and neck cancers (or lung cancer) and may increase the risk in active smokers. No specific risk factors or environmental carcinogens have been identified for salivary gland tumors.

HISTOPATHOLOGY, CARCINOGENESIS, AND MOLECULAR BIOLOGY

Squamous cell head and neck cancers are divided into well-differentiated, moderately well-differentiated, and poorly differentiated categories. Poorly differentiated tumors have a worse prognosis than well-differentiated tumors. For nasopharyngeal cancers, the less common differentiated squamous cell carcinoma is distinguished from non-keratinizing and undifferentiated carcinoma (lymphoepithelioma) that contains infiltrating lymphocytes and is commonly associated with EBV.

Salivary gland tumors can arise from the major (parotid, submandibular, sublingual) or minor salivary glands (located in the submucosa of the upper aerodigestive tract). Most parotid tumors are benign, but half of submandibular and sublingual gland tumors and most minor salivary gland tumors are malignant. Malignant tumors include mucoepidermoid and adenoid cystic carcinomas and adenocarcinomas.

The mucosal surface of the entire pharynx is exposed to alcohol- and tobacco-related carcinogens and is at risk for the development of a premalignant or malignant lesion. Erythroplakia (a red patch) or leukoplakia (a white patch) can be histopathologically classified as hyperplasia, dysplasia, carcinoma in

situ, or carcinoma. However, most head and neck cancer patients do not present with a history of premalignant lesions. Multiple synchronous or metachronous cancers can also be observed. In fact, over time, patients with treated early-stage head and neck cancer are at greater risk of dying from a second malignancy than from a recurrence of the primary disease.

Second head and neck malignancies are usually not therapy-induced; they reflect the exposure of the upper aerodigestive mucosa to the same carcinogens that caused the first cancer. These second primaries develop in the head and neck area, the lung, or the esophagus. Thus, computed tomography (CT) screening for lung cancer in heavy smokers who have already developed a head and neck cancer is recommended. Rarely, patients can develop a radiation therapy–induced sarcoma after having undergone prior radiotherapy for a head and neck cancer.

Much progress has been made in describing the molecular features of head and neck cancer. These features have allowed investigators to describe the genetic and epigenetic alterations and the mutational spectrum of these tumors. Early reports demonstrated frequent overexpression of the epidermal growth factor receptor (EGFR). Overexpression was shown to correlate with poor prognosis. However, it has not proved to be a good predictor of tumor response to EGFR inhibitors, which are active in only about 10–15% of patients as single agents. Complex genetic analyses, including those by The Cancer Genome Atlas project, have been performed. *p53* mutations are found frequently with other major affected oncogenic driver pathways including the mitotic signaling and Notch pathways and cell cycle regulation in HPV-negative tumors. HPV oncogenes act through direct inhibition of the *p53* and *RB* tumor-suppressor genes, thereby initiating the carcinogenic process. While overall mutation rates are similar in HPV-positive and carcinogen-induced tumors, the specific mutational signature of HPV-positive tumors differs with frequent alteration of the PI3K pathway and occasional mutations in KRAS. Overall, these alterations affect mitogenic signaling, genetic stability, cellular proliferation, and differentiation.

CLINICAL PRESENTATION AND DIFFERENTIAL DIAGNOSIS

Most tobacco-related head and neck cancers occur in patients older than age 60 years. HPV-related malignancies are frequently diagnosed in younger patients, usually in their forties or fifties, whereas EBV-related nasopharyngeal cancer can occur at all ages, including teenagers. The manifestations vary according to the stage and primary site of the tumor. Patients with nonspecific signs and symptoms in the head and neck area should be evaluated with a thorough otolaryngologic examination, particularly if symptoms persist longer than 2–4 weeks. Males are more frequently affected than women by head and neck cancers, including HPV-positive tumors.

Cancer of the nasopharynx typically does not cause early symptoms. However, it may cause unilateral serous otitis media due to obstruction of the eustachian tube, unilateral or bilateral nasal obstruction, or epistaxis. Advanced nasopharyngeal carcinoma causes neuropathies of the cranial nerves due to skull base involvement.

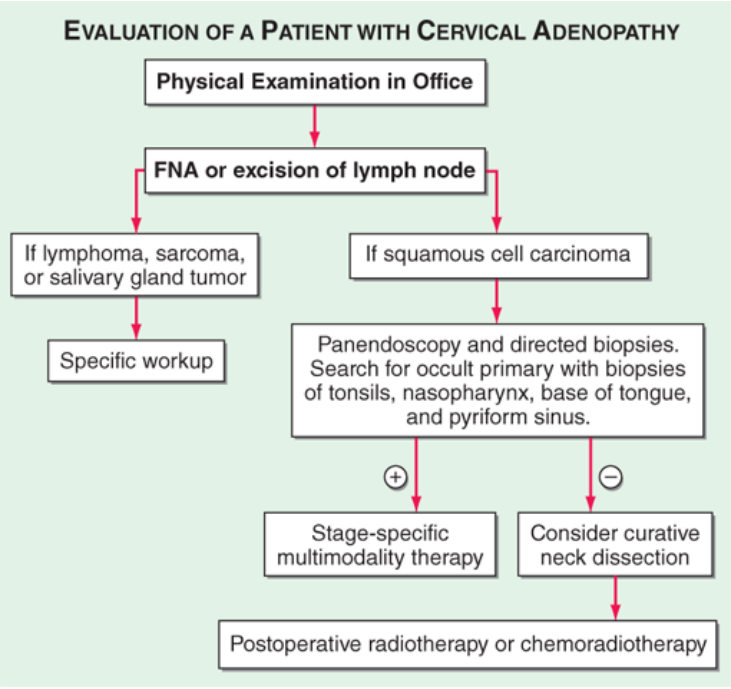
Carcinomas of the oral cavity present as non-healing ulcers, changes in the fit of dentures, or painful lesions and masses. Tumors of the tongue base or oropharynx can cause decreased tongue mobility and alterations in speech. Cancers of the oropharynx or hypopharynx rarely cause early symptoms, but they may cause sore throat and/or otalgia. HPV-related tumors frequently present with neck lymphadenopathy as the first sign.

Hoarseness may be an early symptom of laryngeal cancer, and persistent hoarseness requires referral to a specialist for indirect laryngoscopy and/or radiographic studies. If a head and neck lesion treated initially with antibiotics does not resolve in a short period, further workup is indicated; to simply continue the antibiotic treatment may be to lose the chance of early diagnosis of a malignancy.

Advanced head and neck cancers in any location can cause severe pain, otalgia, airway obstruction, cranial neuropathies, trismus, odynophagia, dysphagia, decreased tongue mobility, fistulas, skin involvement, and massive cervical lymphadenopathy, which may be unilateral or bilateral. Some patients have enlarged lymph nodes even though no primary lesion can be detected by endoscopy or biopsy; these patients are considered to have carcinoma of unknown primary (Fig. 73-1). Tonsillectomy and directed biopsies of the base of tongue can frequently identify a small primary tumor that frequently will be HPV-related. If the enlarged nodes are located in the upper neck and the tumor cells are of squamous cell histology, the malignancy probably arose from a mucosal surface in the head or neck. Tumor cells in supraclavicular lymph nodes may also arise from a primary site in the chest or abdomen.

FIGURE 73-1

Evaluation of a patient with cervical adenopathy without a primary mucosal lesion; a diagnostic workup. FNA, fine-needle aspiration.

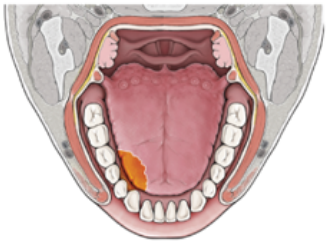
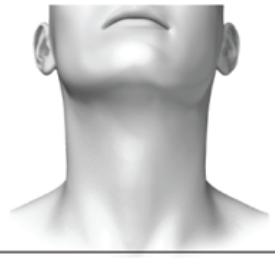
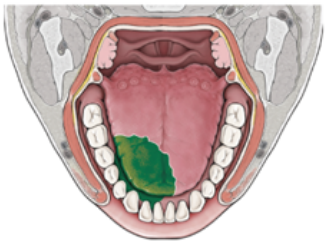
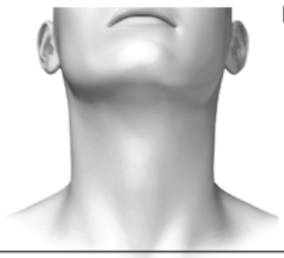
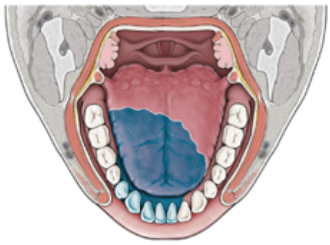
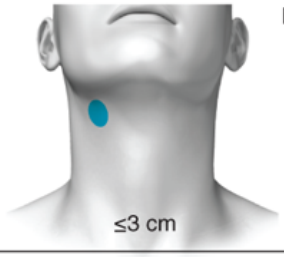
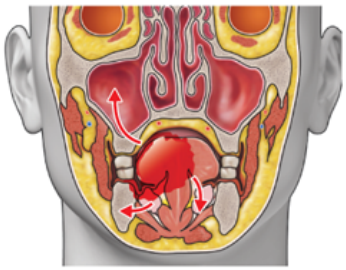
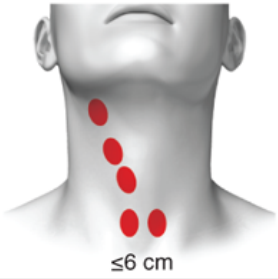
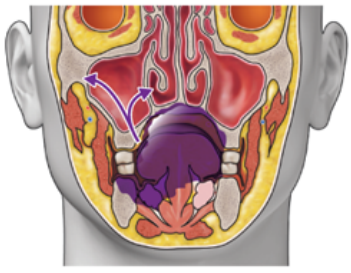
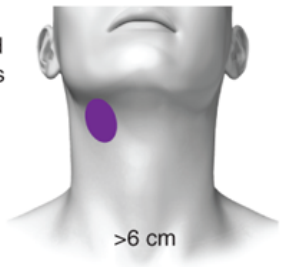


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The physical examination should include inspection of all visible mucosal surfaces and palpation of the floor of the mouth and of the tongue and neck. In addition to tumors themselves, leukoplakia (a white mucosal patch) or erythroplakia (a red mucosal patch) may be observed; these "premalignant" lesions can represent hyperplasia, dysplasia, or carcinoma in situ and require biopsy. Further examination should be performed by a specialist. Additional staging procedures include CT of the head and neck to identify the extent of the disease. Patients with lymph node involvement should have CT scan of the chest and upper abdomen to screen for distant metastases. In heavy smokers, the CT scan of the chest can also serve as a screening tool to rule out a second lung primary tumor. A positron emission tomography (PET) scan may also be administered and can help to identify or exclude distant metastases. CT and PET scans may also be useful in evaluating response to therapy. The definitive staging procedure is an endoscopic examination under anesthesia, which may include laryngoscopy, esophagoscopy, and bronchoscopy; during this procedure, multiple biopsy samples are obtained to establish a primary diagnosis, define the extent of primary disease, and identify any additional premalignant lesions or second primaries.

Head and neck tumors are classified according to the tumor-node-metastasis (TNM) system of the American Joint Committee on Cancer **(Fig. 73-2)**. This classification varies according to the specific anatomic subsite. In general, primary tumors are classified as T1 to T3 by increasing size, whereas T4 usually represents invasion of another structure such as bone, muscle, or root of tongue. Lymph nodes are staged by size, number, and location (ipsilateral vs contralateral to the primary). Distant metastases are found in <10% of patients at initial diagnosis and are more common in patients with advanced lymph node stage; microscopic involvement of the lungs, bones, or liver is more common, particularly in patients with advanced neck lymph node disease. Modern imaging techniques may increase the number of patients with clinically detectable distant metastases in the future. HPV-related oropharyngeal malignancies have consistently been shown to have a better prognosis, and in the upcoming 8th edition of the AJCC staging manual (active in 2018) a separate staging system that takes into account the more favorable outlook of these patients will be included. According to this system, patients with advanced nodal stage can still be considered to have an overall early stage (and associated good prognosis), especially if the patient is a non-smoker or has limited lifelong tobacco exposure.

FIGURE 73-2
Tumor-node-metastasis (TNM) staging system.

Definition of TNM				Stage groupings		
Stage I	T1 	Tumor ≤ 2 cm in greatest dimension without extraparenchymal extension	N0 	N0- No regional lymph node metastasis	T1	N0	M0
Stage II	T2 	Tumor ≥ 2 cm but not more than 4 cm in greatest dimension without extraparenchymal extension	N0 	N0- No regional lymph node metastasis	T2	N0	M0
Stage III	T3 	Tumor ≥ 4 cm and/or tumor having extraparenchymal extension	N1  ≤3 cm	N1- Metastasis in a single ipsilateral lymph node, ≤ 3 cm in greateast dimension	T3 T1 T2 T3	N0 N1 N1 N1	M0 M0 M0 M0
Stage IVA	T4a 	Tumor invades skin, mandible, ear canal, and or fascial nerve	N2  ≤6 cm	N2a- Metastasis in a single ipsilateral lymph node, >3 cm but ≤6 cm N2b- Metastasis in a multiple ipsilateral lymph node, none >6 cm N2c- Metastasis in a bilateral or contralateral lymph nodes, none >6 cm	T4a T4a T1 T2 T3 T4a	N0 N1 N2 N2 N2 N2	M0 M0 M0 M0 M0 M0
Stage IVB	T4b 	Tumor invades skull base and/or pterygoid plates and/or encases carotid artery	N3  >6 cm	N3- Metastasis in a lymph node >6 cm in greatest dimension	T4b Any T	Any N N3	M0 M0
Stage IVC		M1			Any T	Any N	M1

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In patients with lymph node involvement and no visible primary, the diagnosis should be made by lymph node excision (Fig. 73-1). If the results indicate squamous cell carcinoma, a panendoscopy should be performed, with biopsy of all suspicious-appearing areas and directed biopsies of common primary sites, such as the nasopharynx, tonsil, tongue base, and pyriform sinus. HPV-positive tumors especially can have small primary tumors that spread early to locoregional lymph nodes.

TREATMENT

TREATMENT

Head and Neck Cancer

Patients with head and neck cancer can be grossly categorized into three clinical groups: those with localized disease, those with locally or regionally advanced disease (lymph node positive), and those with recurrent and/or metastatic disease below the neck. Comorbidities associated with tobacco and alcohol abuse can affect treatment outcome and define long-term risks for patients who are cured of their disease.

LOCALIZED DISEASE

Nearly one-third of patients have localized disease, that is, T1 or T2 (stage I or stage II) lesions without detectable lymph node involvement or distant metastases. These patients are treated with curative intent by either surgery or radiation therapy. The choice of modality differs according to anatomic location and institutional expertise. Radiation therapy is often preferred for laryngeal cancer to preserve voice function, and surgery is preferred for small lesions in the oral cavity to avoid the long-term complications of radiation, such as xerostomia and dental decay. Randomized data suggest that a prophylactic staging neck dissection should be part of the surgical procedure to eliminate occult nodal metastatic disease. Overall 5-year survival is 60–90%. Most recurrences occur within the first 2 years following diagnosis and are usually local.

LOCALLY OR REGIONALLY ADVANCED DISEASE

Locally or regionally advanced disease—disease with a large primary tumor and/or lymph node metastases—is the stage of presentation for >50% of patients. Such patients can also be treated with curative intent, but not with surgery or radiation therapy alone. Combined-modality therapy including surgery, and/or radiation therapy, and chemotherapy is most successful. Chemotherapy can be administered as induction chemotherapy (chemotherapy before surgery and/or radiotherapy) or as concomitant (simultaneous) chemotherapy and radiation therapy. The latter is currently most commonly used and supported by the best evidence. Five-year survival rates exceed 50% in many trials, but part of this increased survival may be due to an increasing fraction of study populations with HPV-related tumors who carry a better prognosis. HPV testing of newly diagnosed tumors is now performed for most patients at the time of diagnosis, and clinical trials for HPV-related tumors are focused on exploring reductions in treatment intensity, especially radiation dose, in order to ameliorate long-term toxicities (fibrosis, swallowing dysfunction).

In patients with intermediate-stage tumors (stage III and early stage IV), concomitant chemoradiotherapy can be administered either as a primary treatment for patients with unresectable disease, to pursue an organ-preserving approach especially for patients with laryngeal cancer (omission of surgery), or in the postoperative setting for smaller resectable tumors.

Induction Chemotherapy

In this strategy, patients receive chemotherapy (current standard is a three-drug regimen of docetaxel, cisplatin, and fluorouracil [5-FU]) before surgery and radiation therapy. Most patients who receive three cycles show tumor reduction, and the response is clinically “complete” in up to half of patients. This “sequential” multimodality therapy allows for organ preservation in patients with laryngeal and hypopharyngeal cancer, and it has been shown to result in higher cure rates compared with radiotherapy alone.

Concomitant Chemoradiotherapy

With the concomitant strategy, chemotherapy and radiation therapy are given simultaneously rather than in sequence. Tumor recurrences from head and neck cancer develop most commonly locoregionally (in the head and neck area of the primary and draining lymph nodes). The concomitant approach is aimed at enhancing tumor cell killing by radiation therapy in the presence of chemotherapy (radiation enhancement) and is a conceptually attractive approach for bulky tumors. Toxicity (especially mucositis, grade 3 or 4 in 70–80%) is increased with concomitant chemoradiotherapy. However, meta-analyses of randomized trials document an improvement in 5-year survival of 8% with concomitant chemotherapy and radiation therapy. Results seem more favorable in recent trials as more active drugs or more intensive radiotherapy schedules are used. In addition, concomitant chemoradiotherapy produces better laryngectomy-free survival (organ preservation) than radiation therapy alone in patients with advanced larynx cancer. The use of radiation therapy together with cisplatin has also produced improved survival in patients with advanced nasopharyngeal cancer. The outcome of HPV-related cancers seems to be especially favorable following cisplatin-based chemoradiotherapy.

The success of concomitant chemoradiotherapy in patients with unresectable disease has led to the testing of a similar approach in patients with resected intermediate-stage disease as a postoperative therapy. Concomitant chemoradiotherapy produces a significant improvement over postoperative radiation therapy alone for patients whose tumors demonstrate higher risk features, such as extracapsular spread beyond involved lymph nodes, involvement of multiple lymph nodes, or positive margins at the primary site following surgery.

A monoclonal antibody to EGFR (cetuximab) increases survival rates when administered during radiotherapy. EGFR blockade results in radiation sensitization and has milder systemic side effects than traditional chemotherapy agents, although an acneiform skin rash is commonly observed. Nevertheless, the addition of cetuximab to current standard chemoradiotherapy regimens has failed to show further improvement in survival and is not recommended.

TREATMENT APPROACHES FOR HPV-RELATED HEAD AND NECK CANCERS

Given consistent observations of high survival rates for patients with advanced HPV-related oropharyngeal tumors using combined modality treatment strategies de-escalation protocols have attracted widespread interest. The goal here is to decrease the long-term morbidity resulting from high-dose radiation therapy, including extensive neck fibrosis, swallowing problems, and osteoradionecrosis of the jaw. Current studies are investigating the use of lower radiation doses, the use of induction chemotherapy and subsequent omission of chemotherapy or administration of significantly reduced chemoradiation doses in very good responders, and other strategies. In addition, there has been a resurgence of interest in surgical approaches using robotic surgery which allows better visualization of the base of tongue and tonsil. While technically feasible, this approach remains investigational at this time since a large number of patients with disease involving multiple lymph nodes disease will still require post-operative chemoradiotherapy thus negating the goal of treatment de-escalation. It is expected that distinct treatment guidelines from carcinogen-induced tumors will be defined in the coming years.

RECURRENT AND/OR METASTATIC DISEASE

Five to ten percent of patients present with metastatic disease and 30–50% of patients with locoregionally advanced disease experience recurrence, frequently outside the head and neck region. Patients with recurrent and/or metastatic disease are, with few exceptions, treated with palliative intent. Some patients may require local or regional radiation therapy for pain control, but most are given chemotherapy. Response rates to chemotherapy average only 30–50%; the durations of response are short, and the median survival time is 8–10 months. Therefore, chemotherapy provides transient symptomatic benefit. Drugs with single-agent activity in this setting include methotrexate, 5-FU, cisplatin, paclitaxel, and docetaxel. Combinations of cisplatin with 5-FU, carboplatin with 5-FU, and cisplatin or carboplatin with paclitaxel or docetaxel are frequently used.

EGFR-directed therapies, including monoclonal antibodies (e.g., cetuximab) and tyrosine kinase inhibitors (TKIs) of the EGFR signaling pathway (e.g., erlotinib or gefitinib), have single-agent activity of ~10%. Side effects are usually limited to an acneiform rash and diarrhea (for the TKIs). The addition of cetuximab to standard combination chemotherapy with cisplatin or carboplatin and 5-FU was shown to result in a significant increase in median survival. Drugs targeting specific mutations are under investigation, but no such strategy has yet been shown to be feasible in head and neck cancer.

IMMUNOTHERAPIES

Inhibitors of the immune suppressive lymphocyte-surface receptor PD-1 have shown activity in squamous cell cancers of the head and neck. A randomized trial evaluating the PD-1 inhibitor nivolumab vs traditional chemotherapy in second-line treatment of patients with current or metastatic disease showed a significant increase in survival time (7.5 vs 5.1 months) and 1-year survival rates with fewer severe treatment-related toxicities. Similarly, the PD-1 inhibitor pembrolizumab was shown to result in encouraging response rates and survival times in a single-arm phase II trial.

COMPLICATIONS

Complications from treatment of head and neck cancer are usually correlated to the extent of surgery and exposure of normal tissue structures to radiation. Currently, the extent of surgery has been limited or completely replaced by chemotherapy and radiation therapy as the primary approach. Acute complications of radiation include mucositis and dysphagia. Long-term complications include xerostomia, loss of taste, decreased tongue mobility, second malignancies, dysphagia, and neck fibrosis. The complications of chemotherapy vary with the regimen used but usually include myelosuppression, mucositis, nausea and vomiting, and nephrotoxicity (with cisplatin).

The mucosal side effects of therapy can lead to malnutrition and dehydration. Many centers address issues of dentition before starting treatment, and some place feeding tubes to ensure control of hydration and nutrition intake. About 50% of patients develop hypothyroidism from the treatment; thus, thyroid function should be monitored.

SALIVARY GLAND TUMORS

Most benign salivary gland tumors are treated with surgical excision, and patients with invasive salivary gland tumors are treated with surgery and radiation therapy. These tumors may recur regionally; adenoid cystic carcinoma has a tendency to recur along the nerve tracks. Distant metastases may occur as late as 10–20 years after the initial diagnosis. For metastatic disease, therapy is given with palliative intent, usually chemotherapy with doxorubicin and/or cisplatin. Identification of novel agents with activity in these tumors is a high priority. It is hoped that comprehensive genomic characterization of these rare tumors will facilitate these efforts.

FURTHER READING

D'Cruz AK et al: Elective versus therapeutic neck dissection in node-negative oral cancer. N Engl J Med 373:521, 2015.

Ferris RL et al: Nivolumab for recurrent squamous-cell carcinoma of the head and neck. N Engl J Med 375:1856, 2016. [PubMed: 27718784]

Forastiere AA et al: Long-term results of RTOG 91-11: A comparison of three nonsurgical treatment strategies to preserve the larynx in patients with locally advanced larynx cancer. J Clin Oncol 31:845, 2013. [PubMed: 23182993]

Gillison ML et al: Distinct risk factor profiles for human papillomavirus type 16-positive and human papillomavirus type 16-negative head and neck cancers. J Natl Cancer Inst 100:407, 2008. [PubMed: 18334711]

Hayes DN, Van Waes C, Seiwert TY: Genetic landscape of human papillomavirus-associated head and neck cancer and comparison to tobacco-related tumors. J Clin Oncol 33:3227, 2015. [PubMed: 26351353]

Kang H et al: Whole-exome sequencing of salivary gland mucoepidermoid carcinoma. Clin Cancer Res 23:283, 2017.

Vokes EE, Agrawal N, Seiwert TY: HPV-associated head and neck cancer. J Natl Cancer Inst 107:djv344, 2015. [PubMed: 26656751]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 74: Neoplasms of the Lung

Leora Horn; Christine M. Lovly

INTRODUCTION

Lung cancer, which was rare before 1900 with fewer than 400 cases described in the medical literature, is considered a disease of modern man. By the mid-twentieth century, lung cancer had become epidemic and firmly established as the leading cause of cancer-related death in North America and Europe, killing over three times as many men as prostate cancer and nearly twice as many women as breast cancer. Tobacco consumption is the primary cause of lung cancer, a reality firmly established in the mid-twentieth century and codified with the release of the U.S. Surgeon General's 1964 report on the health effects of tobacco smoking. Following the report, cigarette use started to decline in North America and parts of Europe, and with it, so did the incidence of lung cancer. Unfortunately, in many parts of the world cigarette use continues to increase, and along with it, the incidence of lung cancers is also rising. Although tobacco smoking remains the primary cause of lung cancer worldwide, approximately 60% of new lung cancers in the United States occur in former smokers (smoked ≥100 cigarettes per lifetime, quit ≥1 year), many of whom quit decades ago, or never smokers (smoked <100 cigarettes per lifetime). Moreover, one in five women and one in 12 men diagnosed with lung cancer have never smoked. Given the magnitude of the problem, it is incumbent that every internist has a general knowledge of lung cancer and its management.

EPIDEMIOLOGY

Lung cancer is the most common cause of cancer death among American men and women. Approximately 225,000 individuals will be diagnosed with lung cancer in the United States in 2017, and over 150,000 individuals will die from the disease. Lung cancer is uncommon below age 40, with rates increasing until age 80, after which the rate tapers off. The projected lifetime probability of developing lung cancer is estimated to be ~8% among males and ~6% among females. The incidence of lung cancer varies by racial and ethnic group, with the highest age-adjusted incidence rates among African Americans. The excess in age-adjusted rates among African Americans occurs only among men, but examinations of age-specific rates show that below age 50, mortality from lung cancer is more than 25% higher among African American than Caucasian women. Incidence and mortality rates among Hispanics and Native and Asian Americans are ~40–50% those of whites.

RISK FACTORS

Cigarette smokers have a 10-fold or greater increased risk of developing lung cancer compared to those who have never smoked. A large scale genomic study suggested that one genetic mutation is induced for every 15 cigarettes smoked. The risk of lung cancer is lower among persons who quit smoking than among those who continue smoking; former smokers have a ninefold increased risk of developing lung cancer compared to men who have never smoked versus the 20-fold excess in those who continue to smoke. The size of the risk reduction increases with the length of time the person has quit smoking, although generally even long-term former smokers have higher risks of lung cancer than those who never smoked. Cigarette smoking has been shown to increase the risk of all the major types of lung cancer. Environmental tobacco smoke (ETS) or second-hand smoke is also an established cause of lung cancer. The risk from ETS is less than from active smoking, with about a 20–30% increase in lung cancer observed among never smokers married for many years to smokers, in comparison to the 2000% increase among continuing active smokers.

Although cigarette smoking is the cause of the majority of lung cancers, several other risk factors have been identified, including occupational exposures to asbestos, arsenic, bischloromethyl ether, hexavalent chromium, mustard gas, nickel (as in certain nickel-refining processes), and polycyclic aromatic hydrocarbons. Occupational observations also have provided insight into possible mechanisms of lung cancer induction. For example, the risk of lung cancer among asbestos-exposed workers is increased primarily among those with underlying asbestosis, raising the possibility that the scarring and inflammation produced by this fibrotic nonmalignant lung disease may in many cases (although likely not in all) be the trigger for asbestos-induced lung cancer. Several other occupational exposures have been associated with increased rates of lung cancer, but the causal nature of the association is not as clear.

The risk of lung cancer appears to be higher among individuals with low fruit and vegetable intake during adulthood. This observation led to hypotheses that specific nutrients, in particular retinoids and carotenoids, might have chemopreventative effects for lung cancer. However, randomized trials failed to validate this hypothesis. In fact, studies found that the incidence of lung cancer was increased among smokers with supplementation. Ionizing radiation is also an established lung carcinogen, most convincingly demonstrated from studies showing increased rates of lung cancer among survivors of the atom bombs dropped on Hiroshima and Nagasaki and large excesses among workers exposed to alpha irradiation from radon in underground uranium mining. Prolonged exposure to low-level radon in homes might impart a risk of lung cancer equal or greater than that of ETS. Prior lung diseases such as chronic bronchitis, emphysema, and tuberculosis have been linked to increased risks of lung cancer as well.

Smoking Cessation

Given the undeniable link between cigarette smoking and lung cancer (not even addressing other tobacco-related illnesses), physicians must promote tobacco abstinence. Physicians also must help their patients who smoke to stop smoking. Smoking cessation, even well into middle age, can minimize an individual's subsequent risk of lung cancer. Stopping tobacco use before middle age avoids more than 90% of the lung cancer risk attributable to tobacco. However, there is little health benefit derived from just "cutting back." Importantly, smoking cessation can even be beneficial in individuals with an established diagnosis of lung cancer, as it is associated with improved survival, fewer side effects from therapy, and an overall improvement in quality of life. Moreover, smoking can alter the metabolism of many chemotherapy drugs, potentially adversely altering the toxicities and therapeutic benefits of the agents. Consequently, it is important to promote smoking cessation even *after* the diagnosis of lung cancer is established.

Physicians need to understand the essential elements of smoking cessation therapy. The individual must want to stop smoking and must be willing to work hard to achieve the goal of smoking abstinence. Self-help strategies alone only marginally affect quit rates, whereas individual and combined pharmacotherapies in combination with counseling can significantly increase rates of cessation. Therapy with an antidepressant (e.g., bupropion) and nicotine replacement therapy (varenicline, a $\alpha_4\beta_2$ nicotinic acetylcholine receptor partial agonist) are approved by the U.S. Food and Drug Administration (FDA) as first-line treatments for nicotine dependence. However, both drugs have been reported to increase suicidal ideation and must be used with caution. In a randomized trial, varenicline was shown to be more efficacious than bupropion or placebo. Prolonged use of varenicline beyond the initial induction phase proved useful in maintaining smoking abstinence. Clonidine and nortriptyline are recommended as second-line treatments. **(Chap. 448)**.

Inherited Predisposition to Lung Cancer

Exposure to environmental carcinogens, such as those found in tobacco smoke, induce or facilitate the transformation from bronchoepithelial cells to the malignant phenotype. The contribution of carcinogens on transformation is modulated by polymorphic variations in genes that affect aspects of carcinogen metabolism. Certain genetic polymorphisms of the P450 enzyme system, specifically CYP1A1, and chromosome fragility are associated with the development of lung cancer. These genetic variations occur at relatively high frequency in the population, but their contribution to an individual's lung cancer risk is generally low. However, because of their population frequency, the overall impact on lung cancer risk could be high. In addition, environmental factors, as modified by inherited modulators, likely affect specific genes by deregulating important pathways to enable the cancer phenotype.

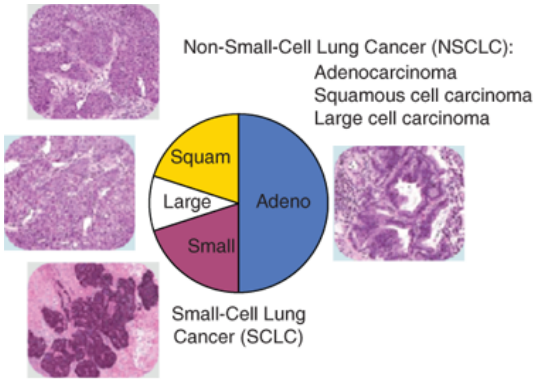
First-degree relatives of lung cancer probands have a two- to threefold excess risk of lung cancer and other cancers, many of which are not smoking-related. These data suggest that specific genes and/or genetic variants may contribute to susceptibility to lung cancer. However, very few such genes have yet been identified. Individuals with inherited mutations in *RB* (patients with retinoblastoma living to adulthood) and *p53* (Li-Fraumeni syndrome) genes may develop lung cancer. Common gene variants involved in lung cancer have been recently identified through large, collaborative, genome-wide association studies. These studies identified three separate loci that are associated with lung cancer (5p15, 6p21, and 15q25) and include genes that regulate acetylcholine nicotinic receptors and telomerase production. A rare germline mutation (T790M) involving the epidermal growth factor receptor (EGFR) maybe be linked to lung cancer susceptibility in never smokers. Likewise, a susceptibility locus on chromosome 6q greatly increases risk lung cancer risk among light and never smokers. Although progress has been made, there is a significant amount of work that remains to be done in identifying heritable risk factors for lung cancer. Currently no molecular criteria are suitable to select patients for more intense screening programs or for specific chemopreventative strategies.

PATHOLOGY

The World Health Organization (WHO) defines lung cancer as tumors arising from the respiratory epithelium (bronchi, bronchioles, and alveoli). The WHO classification system divides epithelial lung cancers into four major cell types: small-cell lung cancer (SCLC), adenocarcinoma, squamous cell carcinoma, and large-cell carcinoma; the latter three types are collectively known as non-small-cell carcinomas (NSCLCs) **(Fig. 74-1)**. Small-cell carcinomas consist of small cells with scant cytoplasm, ill-defined cell borders, finely granular nuclear chromatin, absent or inconspicuous nucleoli, and a high mitotic count. SCLC may be distinguished from NSCLC by the presence of neuroendocrine markers including CD56, neural cell adhesion molecule (NCAM), synaptophysin, and chromogranin. Adenocarcinomas possess glandular differentiation or mucin production and may show acinar, papillary, lepidic, or solid features or a mixture of these patterns. Squamous cell carcinomas of the lung are morphologically identical to extrapulmonary squamous cell carcinomas and cannot be distinguished by immunohistochemistry alone. Squamous cell tumors show keratinization and/or intercellular bridges that arise from bronchial epithelium. The tumor tends to consists of sheets of cells rather than the three-dimensional groups of cells characteristic of adenocarcinomas. Large-cell carcinomas comprise less than 10% of lung carcinomas. These tumors lack the cytologic and architectural features of small-cell carcinoma and glandular or squamous differentiation. Together these four histologic types account for ~90% of all epithelial lung cancers.

FIGURE 74-1

Traditional histologic view of lung cancer.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

All histologic types of lung cancer can develop in current and former smokers, although squamous and small-cell carcinomas are most commonly associated with heavy tobacco use. Through the first half of the twentieth century, squamous carcinoma was the most common subtype of NSCLC diagnosed in the United States. However, with the decline in cigarette consumption over the past six decades, adenocarcinoma has become the most frequent histologic subtype of lung cancer in the United States as both squamous carcinoma and small-cell carcinoma are on the decline. In lifetime never smokers or former light smokers (<10 pack-year history), women, and younger adults (<60 years), adenocarcinoma tends to be the most common form of lung cancer.

In addition to distinguishing between SCLC and NSCLC, because these tumors have quite different natural histories and therapeutic approaches (see below), it is necessary to classify if NSCLC is squamous or nonsquamous because of the recognition that some active chemotherapy agents perform quite differently in squamous carcinomas versus adenocarcinomas and the different recommendations for molecular testing. The revised 2011 classification system, developed jointly by the International Association for the Study of Lung Cancer, the American Thoracic Society, and the European Respiratory Society, provides an integrated approach to the classification of lung adenocarcinoma that includes clinical, molecular, radiographic, and pathologic information.

It is recognized that most lung cancers present in an advanced stage and are often diagnosed based on small biopsies or cytologic specimens, rendering clear histologic distinctions difficult if not impossible. This was addressed by the WHO 2015 revised classification of lung tumors. The distinction between squamous and nonsquamous lung cancer is viewed as critical to optimal therapeutic decision making, a diagnosis of *non-small-cell carcinoma, not otherwise specified* is no longer considered acceptable. This distinction can be achieved using a single marker for adenocarcinoma (thyroid transcription factor-1 or napsin-A) plus a squamous marker (p40 or p63) and/or mucin stains. If tissue is limited and a clear morphological pattern is evident, a diagnosis can be made without immunohistochemistry staining. Both classification systems recommend preservation of sufficient specimen material for appropriate molecular testing necessary to help guide therapeutic decision making (see below).

The terms *adenocarcinoma in situ* and *minimally invasive adenocarcinoma* are now recommended for small solitary adenocarcinomas (≤3 cm) with either pure lepidic growth (term used to describe single-layered growth of atypical cuboidal cells coating the alveolar walls) or predominant lepidic growth with ≤5 mm invasion. Individuals with these entities experience 100% or near 100% 5-year disease-free survival with complete tumor resection. *Invasive adenocarcinomas*, representing more than 70–90% of surgically resected lung adenocarcinomas, are now classified by their predominant pattern: lepidic, acinar, papillary, and solid patterns. Lepidic-predominant subtype has a favorable prognosis, acinar and papillary have an intermediate prognosis, and solid-predominant has a poor prognosis. The terms *signet ring* and *clear cell adenocarcinoma* have been eliminated from the variants of invasive lung adenocarcinoma, whereas the term *micropapillary*, a subtype with a particularly poor prognosis, has been added. Because of prognostic implications, squamous cell carcinoma has also been modified to consist of keratinizing, nonkeratinizing and basaloid, analogous to head and neck cancers.

IMMUNOHISTOCHEMISTRY

The diagnosis of lung cancer most often rests on the morphologic or cytologic features correlated with clinical and radiographic findings. Immunohistochemistry may be used to verify neuroendocrine differentiation within a tumor, with markers such as neuron-specific enolase (NSE), CD56 or NCAM, synaptophysin, chromogranin, and Leu7. Immunohistochemistry is also helpful in differentiating primary from metastatic adenocarcinomas; thyroid transcription factor-1 (TTF-1), identified in tumors of thyroid and pulmonary origin, is positive in over 70% of pulmonary adenocarcinomas and is a reliable indicator of primary lung cancer, provided a thyroid primary has been excluded. A negative TTF-1, however, does not exclude the possibility of a lung primary. TTF-1 is also positive in neuroendocrine tumors of pulmonary and extrapulmonary origin. Napsin-A (Nap-A) is an aspartic protease that plays an important role in maturation of surfactant B7 and is expressed in cytoplasm of type II pneumocytes. In several studies, Nap-A has been reported in >90% of primary lung adenocarcinomas. Notably, a combination of Nap-A and TTF-1 is useful in distinguishing primary lung adenocarcinoma (Nap-A positive, TTF-1 positive) from primary lung squamous cell carcinoma (Nap-A negative, TTF-1 negative) and primary SCLC (Nap-A negative, TTF-1 positive). Cytokeratins 7 and 20 used in combination can help narrow the differential diagnosis; nonsquamous NSCLC, SCLC, and mesothelioma may stain positive for CK7 and negative for CK20, whereas squamous cell lung cancer often will be both CK7 and CK20 negative. p63 is a useful marker for the detection of NSCLCs with squamous differentiation when used in cytologic pulmonary samples. Mesothelioma can be easily identified ultrastructurally, but it has historically been

difficult to differentiate from adenocarcinoma through morphology and immunohistochemical staining. Several markers in the last few years have proven to be more helpful including CK5/6, calretinin, and Wilms tumor gene-1 (*WT-1*), all of which show positivity in mesothelioma.

MOLECULAR PATHOGENESIS

Cancer is a disease involving dynamic changes in the genome. As proposed by Hanahan and Weinberg, virtually all cancer cells acquire six hallmark capabilities: self-sufficiency in growth signals, insensitivity to antigrowth signals, evading apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. The order in which these hallmark capabilities are acquired appears quite variable and can differ from tumor to tumor. Events leading to acquisition of these hallmarks can vary widely, although broadly, cancers arise as a result from accumulations of gain-of-function mutations in oncogenes and loss-of-function mutations in tumor-suppressor genes. Further complicating the study of lung cancer, the sequence of events that lead to disease is clearly different for the various histopathologic entities.

The exact cell of origin for lung cancers is not clearly defined. Whether one cell of origin leads to all histologic forms of lung cancer is unclear. However, for lung adenocarcinoma, evidence suggests that type II epithelial cells (or alveolar epithelial cells) have the capacity to give rise to tumors. For SCLC, cells of neuroendocrine origin have been implicated as precursors.

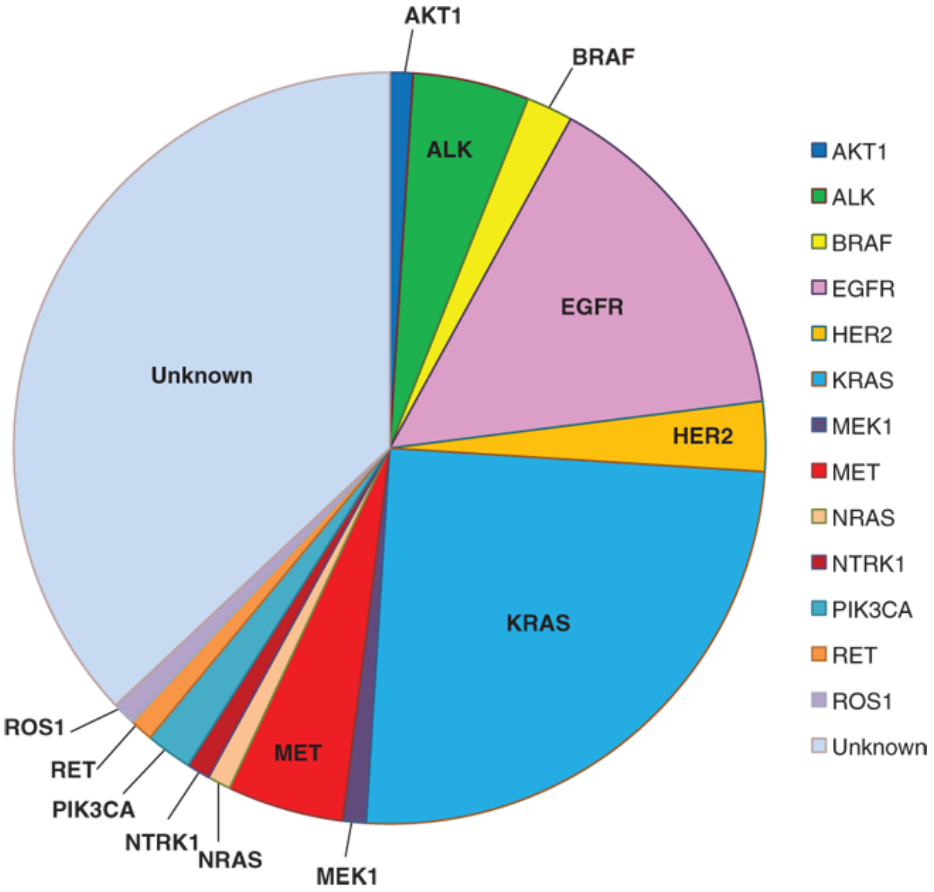
For cancers in general, one theory holds that a small subset of the cells within a tumor (i.e., “stem cells”) are responsible for the full malignant behavior of the tumor. As part of this concept, the large bulk of the cells in a cancer are “offspring” of these cancer stem cells. While clonally related to the cancer stem cell subpopulation, most cells by themselves cannot regenerate the full malignant phenotype. The stem cell concept may explain the failure of standard medical therapies to eradicate lung cancers, even when there is a clinical complete response. Disease recurs because therapies do not eliminate the stem cell component, which may be more resistant to chemotherapy or targeted therapy. Precise human lung cancer stem cells have yet to be identified.

Lung cancer cells harbor multiple chromosomal abnormalities, including mutations, amplifications, insertions, deletions, and translocations. One of the earliest sets of oncogenes found to be aberrant was the MYC family of transcription factors (*MYC, MYCN*, and *MYCL*). *MYC* is most frequently activated via gene amplification or transcriptional dysregulation in both SCLC and NSCLC. Currently, there are no MYC-specific drugs.

Among lung cancer histologies, adenocarcinomas have been the most extensively catalogued for recurrent genomic gains and losses as well as for somatic mutations **(Fig. 74-2)**. While multiple different kinds of aberrations have been found, a major class involves “driver mutations,” which are mutations that occur in genes encoding signaling proteins that when aberrant, drive initiation and maintenance of tumor cells. Importantly, driver mutations can serve as potential Achilles’ heels for tumors, if their gene products can be targeted appropriately. For example, one set of mutations involves the epidermal growth factor receptor (EGFR), which belongs to the ERBB (HER) family of proto-oncogenes, including *EGFR* (ERBB1), *HER2/neu* (ERBB2), *HER3* (ERBB3), and *HER4* (ERBB4). These genes encode cell-surface receptors consisting of an extracellular ligand-binding domain, a transmembrane structure, and an intracellular tyrosine kinase (TK) domain. The binding of ligand to receptor activates receptor dimerization and TK autophosphorylation, initiating a cascade of intracellular events, and leading to increased cell proliferation, angiogenesis, metastasis, and a decrease in apoptosis. Lung adenocarcinomas can arise when tumors express mutant EGFR. These same tumors display high sensitivity to small-molecule EGFR TK inhibitors (TKIs). Additional examples of driver mutations in lung adenocarcinoma include the GTPase *KRAS*, the serine-threonine kinase *BRAF*, and the lipid kinase *PIK3CA*. More recently, additional subsets of lung adenocarcinoma have been identified as defined by the presence of specific chromosomal rearrangements resulting in the aberrant activation of the tyrosine kinases *ALK, ROS1, NTRK* and *RET*. Notably, most driver mutations in lung cancer appear to be mutually exclusive, suggesting that acquisition of one of these mutations is sufficient to drive tumorigenesis. Although driver mutations have mostly been found in adenocarcinomas, three potential molecular targets recently have been identified in squamous cell lung carcinomas: *FGFR1* amplification, *DDR2* mutations, and *PIK3CA* mutations/*PTEN* loss **(Table 74-1)**.

FIGURE 74-2

Driver mutations in lung adenocarcinomas.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 74-1

Driver Mutations in Non-Small-Cell Lung Cancer (NSCLC)

Gene	Alteration	Frequency in NSCLC	Typical Histology
AKT1	Mutation	1%	Adenocarcinoma, squamous
ALK	Rearrangement	3–7%	Adenocarcinoma
BRAF	Mutation	1–3%	Adenocarcinoma
DDR2	Mutation	~4%	Squamous
EGFR	Mutation	10–35%	Adenocarcinoma
FGFR1	Amplification	~20%	Squamous
HER2	Mutation	2–4%	Adenocarcinoma
KRAS	Mutation	15–25%	Adenocarcinoma
MEK1	Mutation	1%	Adenocarcinoma
MET	Amplification	2–4%	Adenocarcinoma
NRAS	Mutation	1%	Adenocarcinoma
NTRK	Rearrangement	1–2%	Adenocarcinoma
PIK3CA	Mutation	1–3%	Squamous
PTEN	Mutation	4–8%	Squamous
ROS1	Rearrangement	1-2%	Adenocarcinoma

A large number of tumor-suppressor genes have also been identified that are inactivated during the pathogenesis of lung cancer. These include *TP53*, *RB1*, *RASSF1A*, *CDKN2A/B*, *LKB1* (*STK11*), and *FHIT*. Nearly 90% of SCLCs harbor mutations in *TP53* and *RB1*. Several tumor-suppressor genes on chromosome 3p appear to be involved in nearly all lung cancers. Allelic loss for this region occurs very early in lung cancer pathogenesis, including in histologically normal smoking-damaged lung epithelium.

EARLY DETECTION AND SCREENING

In lung cancer, clinical outcome is related to the stage at diagnosis, and hence, it is generally assumed that early detection of occult tumors will lead to improved survival. Early detection is a process that involves screening tests, surveillance, diagnosis, and early treatment. Screening refers to the use of tests across a healthy population in order to identify individuals who harbor asymptomatic disease. For a screening program to be successful, there must be a high burden of disease within the target population; the test must be sensitive, specific, accessible, and cost effective; and there must be effective treatment that can reduce mortality. With any screening procedure, it is important to consider the possible influence of *lead-time bias* (detecting the cancer earlier without an effect on survival), *length-time bias* (indolent cancers are detected on screening and may not affect survival, whereas aggressive cancers are likely to cause symptoms earlier in patients and are less likely to be detected), and *overdiagnosis* (diagnosing cancers so slow growing that they are unlikely to cause the death of the patient).

Because a majority of lung cancer patients present with advanced disease beyond the scope of surgical resection, there is understandable skepticism about the value of screening in this condition. Indeed, randomized controlled trials conducted in the 1960s to 1980s using screening chest x-rays (CXR), with or without sputum cytology, reported no impact on lung cancer–specific mortality in patients characterized as high risk (males age ≥45 years with a smoking history). These studies have been criticized for their design, statistical analyses, and outdated imaging modalities. The results of the more recently conducted Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO) are consistent with these earlier reports. Initiated in 1993, participants in the PLCO lung cancer screening trial received annual CXR screening for 4 years, whereas participants in the usual care group received no interventions other than their customary medical care. The diagnostic follow-up of positive screening results was determined by participants and their physicians. The

PLCO trial differed from previous lung cancer screening studies in that women and never smokers were eligible. The study was designed to detect a 10% reduction in lung cancer mortality in the interventional group. A total of 154,901 individuals between 55 and 74 years of age were enrolled (77,445 assigned to annual CXR screenings; 77,456 assigned to usual care). Participant demographics and tumor characteristics were well balanced between the two groups. Through 13 years of follow-up, cumulative lung cancer incidence rates (20.1 vs 19.2 per 10,000 person-years; rate ratio [RR], 1.05; 95% confidence interval [CI], 0.98–1.12) and lung cancer mortality (n = 1213 vs n = 1230) were identical between the two groups. The stage and histology of detected cancers in the two groups also were similar. These data corroborate previous recommendations *against* CXR screening for lung cancer.

In contrast to CXR, low-dose, noncontrast, thin-slice spiral chest computed tomography (LDCT) has emerged as an effective tool to screen for lung cancer. In nonrandomized studies conducted in the 1990s, LDCT scans were shown to detect more lung nodules and cancers than standard CXR in selected high-risk populations (e.g., age ≥60 years and a smoking history of ≥10 pack-years). Notably, up to 85% of the lung cancers discovered in these trials were classified as stage I disease and therefore considered potentially curable with surgical resection.

These data prompted the National Cancer Institute (NCI) to initiate the National Lung Screening Trial (NLST), a randomized study designed to determine if LDCT screening could reduce mortality from lung cancer in high-risk populations as compared with standard posterior anterior CXR. High-risk patients were defined as individuals between 55 and 74 years of age, with a ≥30 pack-year history of cigarette smoking; former smokers must have quit within the previous 15 years. Excluded from the trial were individuals with a previous lung cancer diagnosis, a history of hemoptysis, an unexplained weight loss of >15 lb in the preceding year, or a chest CT within 18 months of enrollment. A total of 53,454 persons were enrolled and randomized to annual screening yearly for three years (LDCT screening, n = 26,722; CXR screening, n = 26,732). Any noncalcified nodule measuring ≥4 mm in any diameter found on LDCT and CXR images with any noncalcified nodule or mass were classified as “positive.” Participating radiologists had the option of not calling a final screen positive if a noncalcified nodule had been stable on the three screening examinations. Overall, 39.1% of participants in the LDCT group and 16% in the CXR group had at least one positive screening result. Of those who screened positive, the false-positive rate was 96.4% in the LDCT group and 94.5% in the CXR group. This was consistent across all three rounds. In the LDCT group, 1060 cancers were identified compared with 941 cancers in the CXR group (645 vs 572 per 100,000 person-years; RR, 1.13; 95% CI, 1.03 to 1.23). Nearly twice as many early-stage 1A cancers were detected in the LDCT group compared with the CXR group (40% vs 21%). The overall rates of lung cancer death were 247 and 309 deaths per 100,000 participants in the LDCT and CXR groups, respectively, representing a 20% reduction in lung cancer mortality in the LDCT-screened population (95% CI, 6.8–26.7%; $p = 0.004$). Compared with the CXR group, the rate of death in the LDCT group from *any* cause was reduced by 6.7% (95% CI, 1.2–13.6; $p = 0.02$) **(Table 74-2)**. The number needed to screen (NNTS) to prevent one lung cancer death was calculated to be 320.

TABLE 74-2

Results of National Lung Screening Trial

	Event Number		Rates of Events per 100,000 Person-Years		Relative Risk (95% CI)	
	LDCT (n = 26,772)	CXR (n = 26,732)	LDCT	CXR	RR	*p* Value
Lung cancer mortality	356	443	247	309	0.80 (0.73–0.93)	.004
All-cause mortality	1877	2000	1303	1395	0.93 (0.86–0.99)	.02
Mortality not due to lung cancer	1521	1557	1056	1086	0.99 (0.95–1.02)	.51

Abbreviations: CI, confidence interval; CXR, chest x-ray; LDCT, low-dose computed tomography; RR, rate ratio.

Source: Modified from PB Bach et al: JAMA 307:2418, 2012.

LDCT screening for lung cancer comes with known risks including a high rate of false-positive results, false-negative results, potential for unnecessary follow-up testing, radiation exposure, overdiagnosis, changes in anxiety level and quality of life, and substantial financial costs. By far the biggest challenge confronting the use of CT screening is the high false-positive rate. False positives can have a substantial impact on patients through the expense and risk of unneeded further evaluation and emotional stress. The management of these patients usually consists of serial CT scans over time to see if the nodules grow, attempted fine-needle aspirates, or surgical resection. At $300 per scan (NCI estimated cost), the outlay for initial LDCT alone could run into the billions of dollars annually, an expense that only further escalates when factoring in various downstream expenditures an individual might incur in the assessment of positive findings. A formal cost-effectiveness analysis of the NLST demonstrated differences between sex, age, and current smoking status and the method of follow up. Despite some questions, low dose LDCT screening has been recommended for all patients meeting criteria for enrollment on NLST. When discussing the option of LDCT screening, use of absolute risks rather than relative risks is helpful because studies indicate the public can process absolute terminology more effectively than relative risk projections. A useful guide has been developed by the NCI to help patients and physicians assess the benefits and harms of LDCT screening for lung cancer **(Table 74-3)**.

TABLE 74-3

The Benefits and Harms of LDCT Screening for Lung Cancer Based on NLST Data

	LDCT	CXR
Benefits: How did CT scans help compared to CXR?		
4 in 1000 fewer died from lung cancer	13 in 1000	17 in 1000
5 in 1000 fewer died from all causes	70 in 1000	75 in 1000
Harms: What problems did CT scans cause compared to CXR?		
223 in 1000 had at least 1 false alarm	365 in 1000	142 in 1000
18 in 1000 had a false alarm leading to an invasive procedure	25 in 1000	7 in 1000
2 in 1000 had a major complication from an invasive procedure	3 in 1000	1 in 1000

Abbreviations: CXR, chest x-ray; LDCT, low-dose computed tomography; NLST, National Lung Screening Trial.

Source: Modified from S Woloshin et al: N Engl J Med 367:1677, 2012.

CLINICAL MANIFESTATIONS

Over half of all patients diagnosed with lung cancer present with locally advanced or metastatic disease at the time of diagnosis. The majority of patients present with signs, symptoms, or laboratory abnormalities that can be attributed to the primary lesion, local tumor growth, invasion or obstruction of adjacent structures, growth at distant metastatic sites, or a paraneoplastic syndrome (Tables 74-4 and 74-5). The prototypical lung cancer patient is a current or former smoker of either sex, usually in the seventh decade of life. A history of chronic cough with or without hemoptysis in a current or former smoker with chronic obstructive pulmonary disease (COPD) age 40 years or older should prompt a thorough investigation for lung cancer even in the face of a normal CXR. A persistent pneumonia without constitutional symptoms and unresponsive to repeated courses of antibiotics also should prompt an evaluation for the underlying cause. Lung cancer arising in a life-time never smoker is more common in women and East Asians. Such patients also tend to be younger than their smoking counterparts at the time of diagnosis. The clinical presentation of lung cancer in never smokers tends to mirror that of current and former smokers.

TABLE 74-4

Presenting Signs and Symptoms of Lung Cancer

Symptom and Signs	Range of Frequency
Cough	8–75%
Weight loss	0–68%
Dyspnea	3–60%
Chest pain	20–49%
Hemoptysis	6–35%
Bone pain	6–25%
Clubbing	0–20%
Fever	0–20%
Weakness	0–10%
Superior vena cava obstruction	0–4%
Dysphagia	0–2%
Wheezing and stridor	0–2%

Source: Reproduced with permission from MA Beckles: Chest 123:97, 2003.

TABLE 74-5

Clinical Findings Suggestive of Metastatic Disease

Symptoms elicited in history	• Constitutional: weight loss >10 lb • Musculoskeletal: pain • Neurologic: headaches, syncope, seizures, extremity weakness, recent change in mental status
Signs found on physical examination	• Lymphadenopathy (>1 cm) • Hoarseness, superior vena cava syndrome • Bone tenderness • Hepatomegaly (>13 cm span) • Focal neurologic signs, papilledema • Soft-tissue mass
Routine laboratory tests	• Hematocrit, <40% in men; <35% in women • Elevated alkaline phosphatase, GGT, SGOT, and calcium levels

Abbreviations: GGT, gamma-glutamyltransferase; SGOT, serum glutamic-oxaloacetic transaminase.

Source: Reproduced with permission from GA Silvestri et al: Chest 123(1 Suppl): 147S, 2003.

Patients with central or endobronchial growth of the primary tumor may present with cough, hemoptysis, wheeze, stridor, dyspnea, or postobstructive pneumonitis. Peripheral growth of the primary tumor may cause pain from pleural or chest wall involvement, dyspnea on a restrictive basis, and symptoms of a lung abscess resulting from tumor cavitation. Regional spread of tumor in the thorax (by contiguous growth or by metastasis to regional lymph nodes) may cause tracheal obstruction, esophageal compression with dysphagia, recurrent laryngeal paralysis with hoarseness, phrenic nerve palsy with elevation

of the hemidiaphragm and dyspnea, and sympathetic nerve paralysis with Horner's syndrome (enophthalmos, ptosis, miosis, and anhydrosis). Malignant pleural effusions can cause pain, dyspnea, or cough. Pancoast (or superior sulcus tumor) syndromes result from local extension of a tumor growing in the apex of the lung with involvement of the eighth cervical and first and second thoracic nerves, and present with shoulder pain that characteristically radiates in the ulnar distribution of the arm, often with radiologic destruction of the first and second ribs. Often Horner's syndrome and Pancoast syndrome coexist. Other problems of regional spread include superior vena cava syndrome from vascular obstruction; pericardial and cardiac extension with resultant tamponade, arrhythmia, or cardiac failure; lymphatic obstruction with resultant pleural effusion; and lymphangitic spread through the lungs with hypoxemia and dyspnea. In addition, lung cancer can spread transbronchially, producing tumor growth along multiple alveolar surfaces with impairment of gas exchange, respiratory insufficiency, dyspnea, hypoxemia, and sputum production. Constitutional symptoms may include anorexia, weight loss, weakness, fever, and night sweats. Apart from the brevity of symptom duration, these parameters fail to clearly distinguish SCLC from NSCLC or even from neoplasms metastatic to lungs.

Extrathoracic metastatic disease is found at autopsy in more than 50% of patients with squamous carcinoma, 80% of patients with adenocarcinoma and large-cell carcinoma, and more than 95% of patients with SCLC. Approximately one-third of patients present with symptoms as a result of distant metastases. Lung cancer metastases may occur in virtually every organ system, and the site of metastatic involvement largely determines other symptoms. Patients with brain metastases may present with headache, nausea and vomiting, seizures, or neurologic deficits. Patients with bone metastases may present with pain, pathologic fractures, or cord compression. The latter may also occur with epidural metastases. Individuals with bone marrow invasion may present with cytopenias or leukoerythroblastosis. Those with liver metastases may present with hepatomegaly, right upper quadrant pain, fever, anorexia, and weight loss. Liver dysfunction and biliary obstructions are rare. Adrenal metastases are common but rarely cause pain or adrenal insufficiency unless they are large.

Paraneoplastic syndromes are common in patients with lung cancer, especially those with SCLC, and may be the presenting finding or the first sign of recurrence. In addition, paraneoplastic syndromes may mimic metastatic disease and, unless detected, lead to inappropriate palliative rather than curative treatment. Often the paraneoplastic syndrome may be relieved with successful treatment of the tumor. In some cases, the pathophysiology of the paraneoplastic syndrome is known, particularly when a hormone with biological activity is secreted by a tumor. However, in many cases, the pathophysiology is unknown. Systemic symptoms of anorexia, cachexia, weight loss (seen in 30% of patients), fever, and suppressed immunity are paraneoplastic syndromes of unknown etiology or at least not well defined. Weight loss greater than 10% of total body weight is considered a bad prognostic sign. Endocrine syndromes are seen in 12% of patients; hypercalcemia resulting from ectopic production of parathyroid hormone (PTH), or more commonly, PTH-related peptide, is the most common life-threatening metabolic complication of malignancy, primarily occurring with squamous cell carcinomas of the lung. Clinical symptoms include nausea, vomiting, abdominal pain, constipation, polyuria, thirst, and altered mental status.

Hyponatremia may be caused by the syndrome of inappropriate secretion of antidiuretic hormone (SIADH) or possibly atrial natriuretic peptide (ANP) **(Chap. 89)**. SIADH resolves within 1–4 weeks of initiating chemotherapy in the vast majority of cases. During this period, serum sodium can usually be managed and maintained above 128 mEq/L via fluid restriction. Demeclocycline can be a useful adjunctive measure when fluid restriction alone is insufficient. Vasopressin receptor antagonists like tolvaptan also have been used in the management of SIADH. However, there are significant limitations to the use of tolvaptan including liver injury and overly rapid correction of the hyponatremia, which can lead to irreversible neurologic injury. Likewise, the cost of tolvaptan may be prohibitive (as high as $300 per tablet in some areas). Of note, patients with ectopic ANP may have worsening hyponatremia if sodium intake is not concomitantly increased. Accordingly, if hyponatremia fails to improve or worsens after 3–4 days of adequate fluid restriction, plasma levels of ANP should be measured to determine the causative syndrome.

Ectopic secretion of ACTH by SCLC and pulmonary carcinoids usually results in additional electrolyte disturbances, especially hypokalemia, rather than the changes in body habitus that occur in Cushing's syndrome from a pituitary adenoma **(Chap. 89)**. Treatment with standard medications, such as metyrapone and ketoconazole, is largely ineffective due to extremely high cortisol levels. The most effective strategy for management of the Cushing's syndrome is effective treatment of the underlying SCLC. Bilateral adrenalectomy may be considered in extreme cases.

Skeletal–connective tissue syndromes include clubbing in 30% of cases (usually NSCLCs) and hypertrophic primary osteoarthropathy in 1–10% of cases (usually adenocarcinomas). Patients may develop periostitis, causing pain, tenderness, and swelling over the affected bones and a positive bone scan. Neurologic-myopathic syndromes are seen in only 1% of patients but are dramatic and include the myasthenic Eaton-Lambert syndrome and retinal blindness with SCLC, whereas peripheral neuropathies, subacute cerebellar degeneration, cortical degeneration, and polymyositis are seen with all lung cancer types. Many of these are caused by autoimmune responses such as the development of anti-voltage-gated calcium channel antibodies in Eaton-Lambert syndrome. Patients with this disorder present with proximal muscle weakness, usually in the lower extremities, occasional autonomic dysfunction, and rarely, cranial nerve symptoms or involvement of the bulbar or respiratory muscles. Depressed deep tendon reflexes are frequently present. In contrast to patients with myasthenia gravis, strength improves with serial effort. Some patients who respond to chemotherapy will have resolution of the neurologic abnormalities. Thus, chemotherapy is the initial treatment of choice. Paraneoplastic encephalomyelitis and sensory neuropathies, cerebellar degeneration, limbic encephalitis, and brainstem encephalitis occur in SCLC in association with a variety of antineuronal antibodies such as anti-Hu, anti-CRMP5, and ANNA-3. Paraneoplastic cerebellar degeneration may be associated with anti-Hu, anti-Yo, or P/Q calcium channel autoantibodies. Coagulation or thrombotic or other hematologic manifestations occur in 1–8% of patients and include migratory venous thrombophlebitis (Trousseau's syndrome), nonbacterial thrombotic (marantic) endocarditis with arterial emboli, and disseminated intravascular coagulation with hemorrhage, anemia, granulocytosis, and leukoerythroblastosis. Thrombotic disease complicating cancer is usually a poor prognostic

sign. Cutaneous manifestations such as dermatomyositis and acanthosis nigricans are uncommon (1%), as are the renal manifestations of nephrotic syndrome and glomerulonephritis (≤1%).

DIAGNOSING LUNG CANCER

Tissue sampling is required to confirm a diagnosis in all patients with suspected lung cancer. In patients with suspected metastatic disease, a biopsy of a distant site of disease is preferred for tissue confirmation. Given the greater emphasis placed on molecular testing for NSCLC patients, a core biopsy is preferred to ensure adequate tissue for analysis. Tumor tissue may be obtained via minimally invasive techniques such as bronchial or transbronchial biopsy during fiberoptic bronchoscopy, by fine-needle aspiration (FNA) or percutaneous biopsy using image guidance, or via endobronchial ultrasound (EBUS)-guided biopsy. Depending on the location, lymph node sampling may occur via transesophageal endoscopic ultrasound-guided biopsy (EUS), EBUS, or blind biopsy. In patients with clinically palpable disease such as a lymph node or skin metastasis, a biopsy may be obtained. In patients with suspected metastatic disease, a diagnosis may be confirmed by percutaneous biopsy of a soft tissue mass, lytic bone lesion, bone marrow, pleural or liver lesion, or an adequate cell block obtained from a malignant pleural effusion. In patients with a suspected malignant pleural effusion, if the initial thoracentesis is negative, a repeat thoracentesis is warranted. Although the majority of pleural effusions are due to malignant disease, particularly if they are exudative or bloody, some may be parapneumonic. In the absence of distant disease, such patients should be considered for possible curative treatment.

The diagnostic yield of any biopsy depends on several factors including location (accessibility) of the tumor, tumor size, tumor type, and technical aspects of the diagnostic procedure including the experience level of the bronchoscopist and pathologist. In general, central lesions such as squamous cell carcinomas, small-cell carcinomas, or endobronchial lesions such as carcinoid tumors are more readily diagnosed by bronchoscopic examination, whereas peripheral lesions such as adenocarcinomas and large-cell carcinomas are more amenable to transthoracic biopsy. Diagnostic accuracy for SCLC versus NSCLC for most specimens is excellent, with lesser accuracy for subtypes of NSCLC.

Bronchoscopic specimens include bronchial brush, bronchial wash, bronchioloalveolar lavage, transbronchial FNA, and core biopsy. For more accurate histologic classification, mutation analysis, or investigational purposes, reasonable efforts (e.g., a core needle biopsy) should be made to obtain more tissue than what is contained in a routine cytology specimen obtained by FNA. Overall sensitivity for combined use of bronchoscopic methods is ~80%, and together with tissue biopsy, the yield increases to 85–90%. Like transbronchial core biopsy specimens, transthoracic core biopsy specimens are also preferred. Sensitivity is highest for larger lesions and peripheral tumors. In general, core biopsy specimens, whether transbronchial, transthoracic, or EUS-guided, are superior to other specimen types. This is primarily due to the higher percentage of tumor cells with fewer confounding factors such as obscuring inflammation and reactive nonneoplastic cells.

Sputum cytology is inexpensive and noninvasive but has a lower yield than other specimen types due to poor preservation of the cells and more variability in acquiring a good-quality specimen. The yield for sputum cytology is highest for larger and centrally located tumors such as squamous cell carcinoma and small-cell carcinoma histology. The specificity for sputum cytology averages close to 100%, although sensitivity is generally <70%. The accuracy of sputum cytology improves with increased numbers of specimens analyzed. Consequently, analysis of at least three sputum specimens is recommended.

STAGING LUNG CANCER

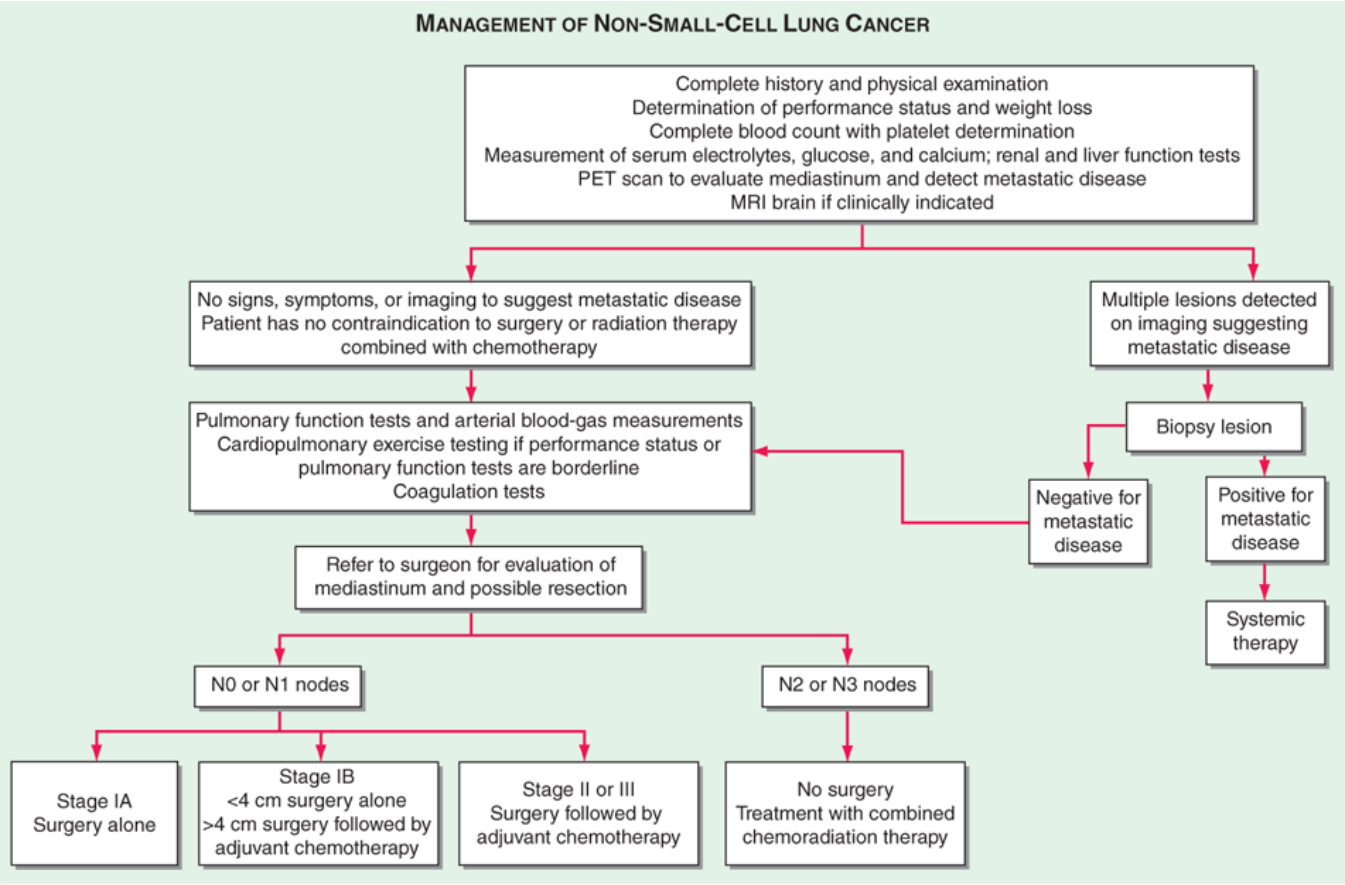
Lung cancer staging consists of two parts: first, a determination of the location of the tumor and possible metastatic sites (anatomic staging), and second, an assessment of a patient's ability to withstand various antitumor treatments (physiologic staging). All patients with lung cancer should have a complete history and physical examination, with evaluation of all other medical problems, determination of performance status, and history of weight loss. The most significant dividing line is between those patients who are candidates for surgical resection and those who are inoperable but will benefit from chemotherapy, radiation therapy, or both. Staging with regard to a patient's potential for surgical resection is principally applicable to NSCLC.

ANATOMIC STAGING OF PATIENTS WITH LUNG CANCER

The accurate staging of patients with NSCLC is essential for determining the appropriate treatment in patients with resectable disease and for avoiding unnecessary surgical procedures in patients with advanced disease **(Fig. 74-3)**. All patients with NSCLC should undergo initial radiographic imaging with CT scan, positron emission tomography (PET), or preferably CT-PET. PET scanning attempts to identify sites of malignancy based on glucose metabolism by measuring the uptake of ^{18}F-fluorodeoxyglucose (FDG). Rapidly dividing cells, presumably in the lung tumors, will preferentially take up ^{18}F-FDG and appear as a "hot spot." To date, PET has been mostly used for staging and detection of metastases in lung cancer and in the detection of nodules >15 mm in diameter. Combined ^{18}F-FDG PET-CT imaging has been shown to improve the accuracy of staging in NSCLC compared to visual correlation of PET and CT or either study alone. CT-PET has been found to be superior in identifying pathologically enlarged mediastinal lymph nodes and extrathoracic metastases. A standardized uptake value (SUV) of >2.5 on PET is highly suspicious for malignancy. False negatives can be seen in diabetes, in lesions <8 mm, and in slow-growing tumors (e.g., carcinoid tumors or well-differentiated adenocarcinoma). False positives can be seen in certain infections and granulomatous disease (e.g., tuberculosis). Thus, PET should never be used alone to diagnose lung cancer, mediastinal involvement, or metastases. Confirmation with tissue biopsy is required. For brain metastases, magnetic resonance imaging (MRI) is the most effective method. MRI can also be useful in selected circumstances, such as superior sulcus tumors to rule out brachial plexus involvement, but in general, MRI does not play a major role in NSCLC staging.

FIGURE 74-3

Algorithm for management of non-small-cell lung cancer. MRI, magnetic resonance imaging; PET, positron emission tomography.



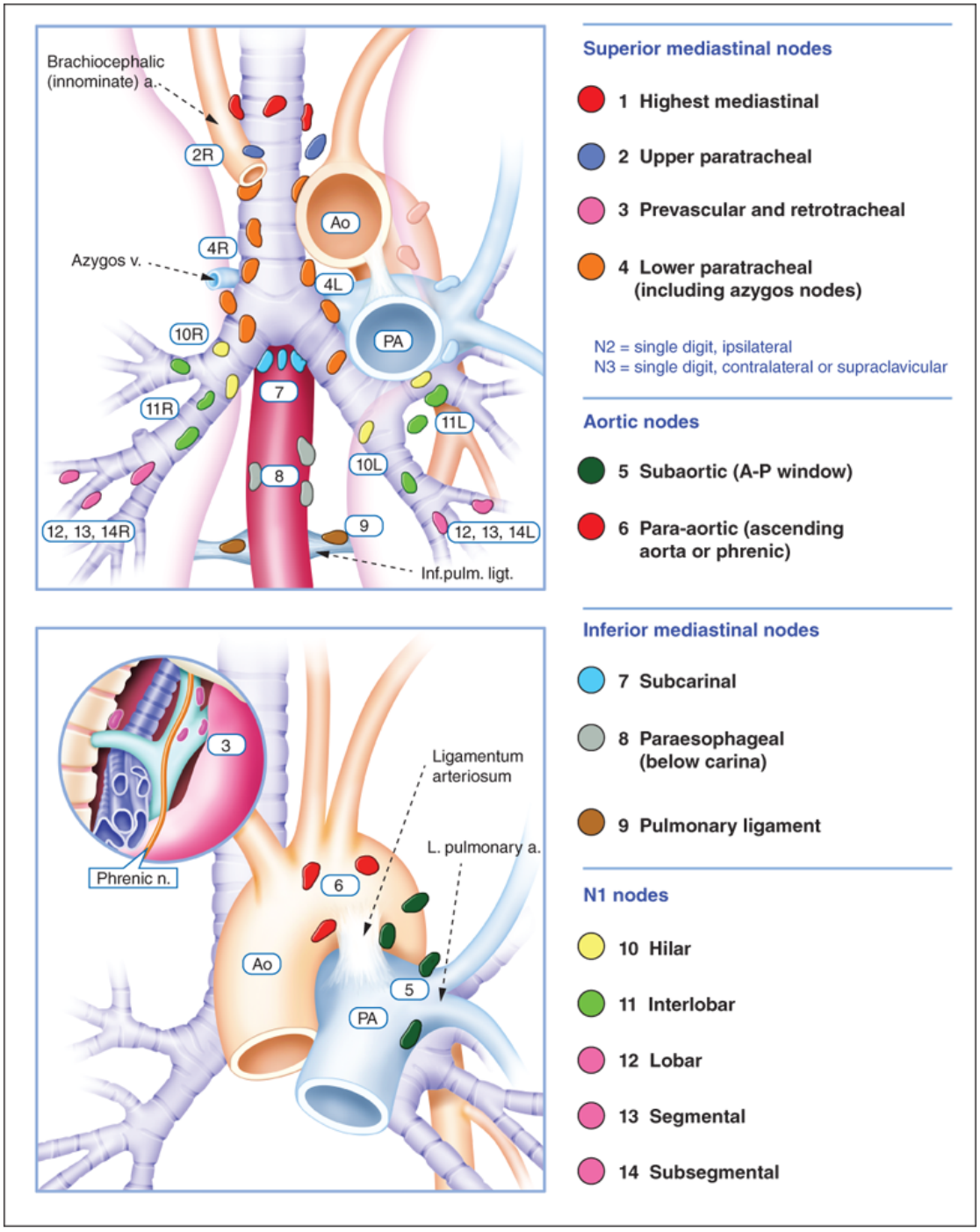
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In patients with NSCLC, the following are contraindications to potential curative resection: extrathoracic metastases, superior vena cava syndrome, vocal cord and, in most cases, phrenic nerve paralysis, malignant pleural effusion, cardiac tamponade, tumor within 2 cm of the carina (potentially curable with combined chemoradiotherapy), metastasis to the contralateral lung, metastases to supraclavicular lymph nodes, contralateral mediastinal node metastases (potentially curable with combined chemoradiotherapy), and involvement of the main pulmonary artery. In situations where it will make a difference in treatment, abnormal scan findings require tissue confirmation of malignancy so that patients are not precluded from having potentially curative therapy.

The best predictor of metastatic disease remains a careful history and physical examination. If signs, symptoms, or findings from the physical examination suggest the presence of malignancy, then sequential imaging starting with the most appropriate study should be performed. If the findings from the clinical evaluation are negative, then imaging studies beyond CT-PET are unnecessary and the search for metastatic disease is complete. More controversial is how one should assess patients with known stage III disease. Because these patients are more likely to have asymptomatic occult metastatic disease, current guidelines recommend a more extensive imaging evaluation including imaging of the brain with either CT scan or MRI. In patients in whom distant metastatic disease has been ruled out, lymph node status needs to be assessed via a combination of radiographic imaging and/or minimally invasive techniques such as those mentioned above and/or invasive techniques such as mediastinoscopy, mediastinotomy, thoracoscopy, or thoracotomy. Approximately one-quarter to one-half of patients diagnosed with NSCLC will have mediastinal lymph node metastases at the time of diagnosis. Lymph node sampling is recommended in all patients with enlarged nodes detected by CT or PET scan and in patients with large tumors or tumors occupying the inner third of the lung. The extent of mediastinal lymph node involvement is important in determining the appropriate treatment strategy: surgical resection followed by adjuvant chemotherapy versus combined chemoradiation alone (see below). A standard nomenclature for referring to the location of lymph nodes involved with lung cancer has evolved **(Fig. 74-4)**.

FIGURE 74-4

Lymph node stations in staging non-small-cell lung cancer. The International Association for the Study of Lung Cancer (IASLC) lymph node map, including the proposed grouping of lymph node stations into “zones” for the purposes of prognostic analyses. a., artery; Ao, aorta; Inf. pulm. ligt., inferior pulmonary ligament; n., nerve; PA, pulmonary artery; v., vein.

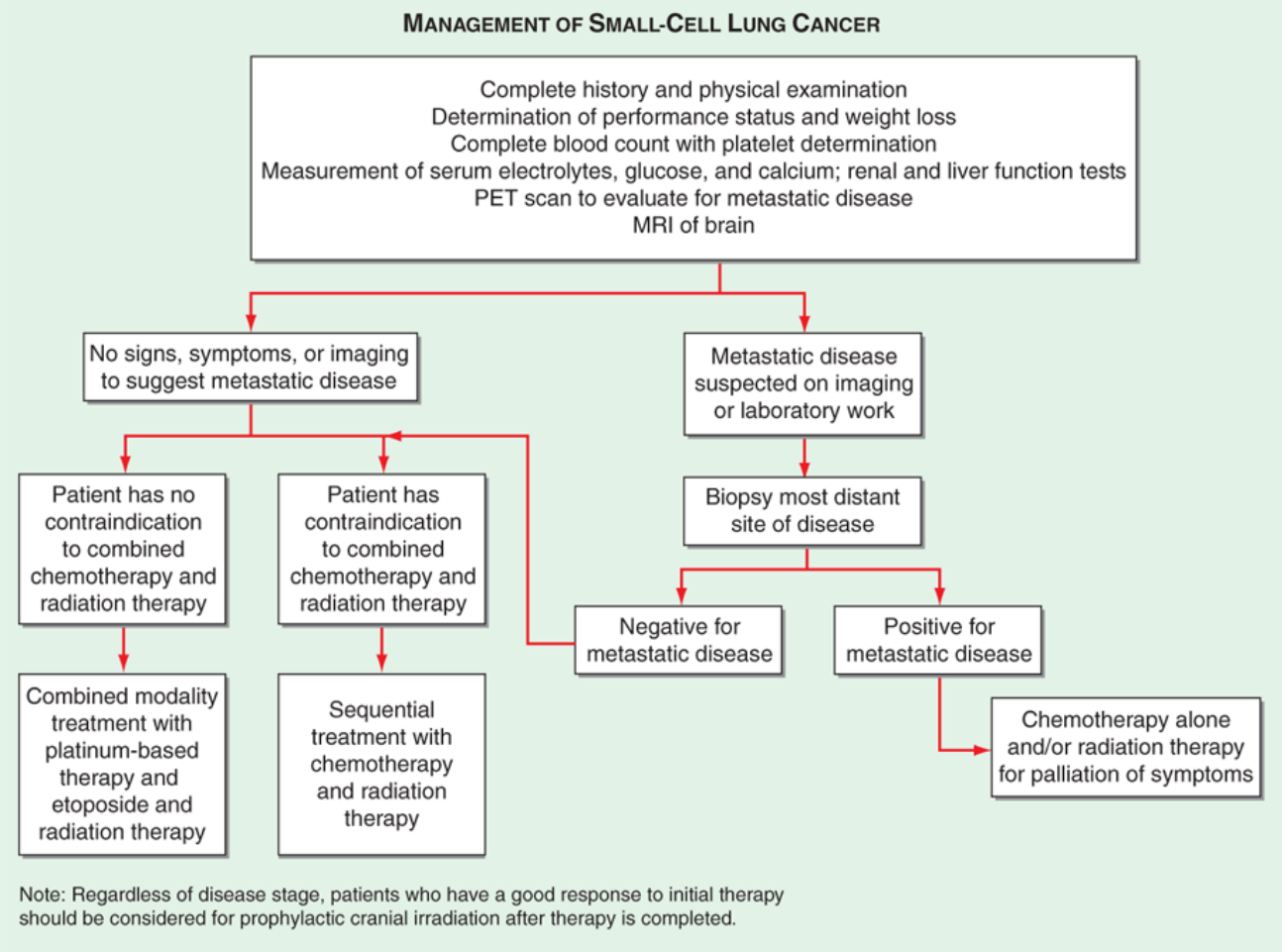


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In SCLC patients, current staging recommendations include a PET-CT scan and MRI of the brain (positive in 10% of asymptomatic patients) (Fig. 74-5). Bone marrow biopsies and aspirations are rarely performed now given the low incidence of isolated bone marrow metastases. Confirmation of metastatic disease, ipsilateral or contralateral lung nodules, or metastases beyond the mediastinum may be achieved by the same modalities recommended earlier for patients with NSCLC.

FIGURE 74-5

Algorithm for management of small-cell lung cancer. MRI, magnetic resonance imaging; PET, positron emission tomography.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

If a patient has signs or symptoms of spinal cord compression (pain, weakness, paralysis, urinary retention), a spinal CT or MRI scan and examination of the cerebrospinal fluid cytology should be performed. If metastases are evident on imaging, a neurosurgeon should be consulted for possible palliative surgical resection and/or a radiation oncologist should be consulted for palliative radiotherapy to the site of compression. If signs or symptoms of leptomeningitis develop at any time in a patient with lung cancer, an MRI of the brain and spinal cord should be performed, as well as a spinal tap, for detection of malignant cells. If the spinal tap is negative, a repeat spinal tap should be considered. There is currently no approved therapy for the treatment of leptomeningeal disease.

STAGING SYSTEM FOR NON-SMALL-CELL LUNG CANCER

The tumor-node-metastasis (TNM) international staging system provides useful prognostic information and is used to stage all patients with NSCLC. The various T (tumor size), N (regional node involvement), and M (presence or absence of distant metastasis) are combined to form different stage groups (Tables 74-6 and 74-7). The seventh edition of the TNM staging system went into effect in 2010 and developed using a much more robust database of more than 100,000 patients with lung cancer who were treated in multiple countries between 1990 and 2000. Data from 67,725 patients with NSCLC were then used to reevaluate the prognostic value of the TNM descriptors. In the current edition T1 tumors are divided into tumors ≤2 cm in size, as these patients were found to have a better prognosis compared to patients with tumors >2 cm but ≤3 cm. T2 tumors are divided into those that are >3 cm but ≤5 cm and those that are >5 cm but ≤7 cm. Tumors that are >7 cm are considered T3 tumors. T3 tumors also include tumors with invasion into local structures such as chest wall and diaphragm and additional nodules in the same lobe. T4 tumors include tumors of any size with invasion into mediastinum, heart, great vessels, trachea, or esophagus or multiple nodules in the ipsilateral lung. The eight edition of the TNM has been proposed and differences are outlined in Tables 74-6 and 74-7. The major changes are in the T and M staging while no changes have been made to the current classification of lymph node involvement (N). Patients with metastasis may be classified as M1a (malignant pleural or pericardial effusion, pleural nodules, or nodules in the contralateral lung) or M1b (distant metastasis; e.g., bone, liver, adrenal, or brain metastasis), M1b single metastasis to a single organ or M1c multiple metastases to a single organ or metastases to multiple organs. Based on these data, approximately one-third of patients have localized disease that can be treated with curative attempt (surgery or radiotherapy), one-third have local or regional disease that may or may not be amenable to a curative attempt, and one-third have metastatic disease at the time of diagnosis.

TABLE 74-6

Comparison of Seventh and Eight Edition TNM Staging Systems for Non-Small-Cell Lung Cancer

	TNM Staging System for Lung Cancer (7th Edition)	TNM Staging System for Lung Cancer (8th Edition)
Primary Tumor (T)		
T1	Tumor ≤3 cm diameter, surrounded by lung or visceral pleura, without invasion more proximal than lobar bronchus	T1 tumor ≤3 cm in diameter surrounds by lung or visceral pleural without evidence of main bronchus
T1a	Tumor ≤2 cm in diameter	Tumor <1 cm
T1b	Tumor >2 cm but ≤ 3 cm in diameter	Tumor ≥1 cm but ≤2 cm
T1c	N/A	Tumor >2 cm but ≤3 cm
T2	Tumor >3 cm but ≤7 cm, or tumor with any of the following features:	T2 tumor >3 cm but ≤5 cm or tumor with any of the following features that does not involve the entire lung
	Involves main bronchus ≥2 cm distal to carina	Involves main bronchus ≥2 cm distal to carina
	Invades visceral pleura	Invades visceral pleura
	Associated with atelectasis or obstructive pneumonitis that extends to the hilar region but does not involve the entire lung	Associate with atelectasis or obstructive pneumonitis that extends to the hilar region
T2a	Tumor >3 cm but ≤5 cm	Tumor > 3 cm but ≤4 cm
T2b	Tumor >5 cm but ≤7 cm	Tumor >4 cm but ≤5 cm
T3	Tumor >7 cm or any of the following:	>5 cm but ≤7 cm or any of the following:
	Directly invades any of the following: chest wall, diaphragm, phrenic nerve, mediastinal pleura, parietal pericardium, main bronchus <2 cm from carina (without involvement of carina)	Directly invades any of the following chest wall, diaphragm, phrenic nerve, mediastinal pleura, parietal pericardium, main bronchus <2 cm from carina (without involvement of carina)
	Atelectasis or obstructive pneumonitis of the entire lung	Atelectasis or obstructive pneumonitis of the entire lung
	Separate tumor nodules in the same lobe	
T4	Tumor of any size that invades the mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, vertebral body, carina, or with separate tumor nodules in a different ipsilateral lobe	7 cm or any of the following invades the mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, vertebral body, carina, or with separate tumor nodules in a different ipsilateral lobe
Regional Lymph Nodes (N)		
N0	No regional lymph node metastases	N0 No regional lymph node metastases
N1	Metastasis in ipsilateral peribronchial and/or ipsilateral hilar lymph nodes and intrapulmonary nodes, including involvement by direct extension	N1 Metastasis in ipsilateral peribronchial and/or ipsilateral hilar lymph nodes and intrapulmonary nodes, including involvement by direct extension
N2	Metastasis in ipsilateral mediastinal and/or subcarinal lymph node(s)	N2 Metastasis in ipsilateral mediastinal and/or subcarinal lymph node(s)
N3	Metastasis in contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular lymph node(s)	N3 Metastasis in contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular lymph node(s)

	TNM Staging System for Lung Cancer (7th Edition)	TNM Staging System for Lung Cancer (8th Edition)
Distant Metastasis (M)		
M0	No distant metastasis	M0 No distant metastasis
M1	Distant metastasis	Distant metastasis
M1a	Separate tumor nodule(s) in a contralateral lobe; tumor with pleural nodules or malignant pleural or pericardial effusion	M1 a separate nodule(s) in a contralateral tumor with pleural nodules or malignant pleural or pericardial effusion
M1b	Distant metastasis (in extrathoracic organs)	Single metastasis in a single organ
M1c		multiple metastases in a single organ or in several organs

Abbreviation: TNM, tumor-node-metastasis.

Source: Reproduced with permission from P Goldstraw et al: J Thorac Oncol 2:706, 2007.

TABLE 74-7

Comparison of Seventh and Eighth Edition TNM Staging Systems for Non-Small-Cell Lung Cancer

Stage Groupings Seventh Edition				Stage Groupings Eighth Edition			
Stage IA	T1a-T1b	N0	M0	Stage IA1 Stage IA2 Stage IA3	T1a T1b T1c	N0 N0 N0	M0 M0 M0
Stage IB	T2a	N0	M0	Stage IB	T2a	N0	M0
Stage IIA	T1a,T1b,T2a	N1	M0	Stage IIA T2bN0M0	T2b	N0	M0
	T2b	N0	M0				
Stage IIB	T2b	N1	M0	Stage IIB	T1a-T2b	N1	M0
	T3	N0	M0		T3	N0	M0
Stage IIIA	T1a,T1b,T2a,T2b	N2	M0	Stage IIIA	T1-2b	N2	M0
	T3	N1,N2	M0		T3	N1	M0
	T4	N0,N1	M0		T4	N0/N1	M0
Stage IIIB	T4	N2	M0	Stage IIIB	T1-2b	N3	M0
	Any T	N3	M0		T3/T4	N0/N1	M0
Stage IIIC	N/A				T3/T4	N3	M0
Stage IV	Any T	Any N	M1a or M1b	Stage IVA	Any T	Any N	M1a/M1b
				Stage IV B	Any T	Any N	M1c

STAGING SYSTEM FOR SMALL-CELL LUNG CANCER

In patients with SCLC, it is now recommended that both the Veterans Administration system and the American Joint Committee on Cancer/International Union Against Cancer seventh edition system (TNM) be used to classify the tumor stage. The Veterans Administration system is a distinct two-stage system dividing patients into those with limited- or extensive-stage disease. Patients with limited-stage disease (LD) have cancer that is confined to the ipsilateral hemithorax and can be encompassed within a tolerable radiation port. Thus, contralateral supraclavicular nodes, recurrent laryngeal nerve involvement, and superior vena caval obstruction can all be part of LD. Patients with extensive-stage disease (ED) have overt metastatic disease by imaging or physical examination. Cardiac tamponade, malignant pleural effusion, and bilateral pulmonary parenchymal involvement generally qualify disease as ED, because the involved organs cannot be encompassed safely or effectively within a single radiation therapy port. Sixty to 70% of patients are diagnosed with ED at presentation. The TNM staging system is preferred in the rare SCLC patient presenting with what appears to be clinical stage I disease (see above).

PHYSIOLOGIC STAGING

Patients with lung cancer often have other comorbid conditions related to smoking including cardiovascular disease and COPD. To improve their preoperative condition, correctable problems (e.g., anemia, electrolyte and fluid disorders, infections, cardiac disease, and arrhythmias) should be addressed, appropriate chest physical therapy should be instituted, and patients should be encouraged to stop smoking. Patients with a forced expiratory volume in 1 s (FEV_1) of greater than 2 L or greater than 80% of predicted can tolerate a pneumonectomy, and those with an FEV_1 greater than 1.5 L have adequate reserve for a lobectomy. In patients with borderline lung function but a resectable tumor, cardiopulmonary exercise testing could be performed as part of the physiologic evaluation. This test allows an estimate of the maximal oxygen consumption (Vo_{2max}). A Vo_{2max} <15 mL/(kg.min) predicts for a higher risk of postoperative complications. Patients deemed unable to tolerate lobectomy or pneumonectomy from a pulmonary functional standpoint may be candidates for more limited resections, such as wedge or anatomic segmental resection, although such procedures are associated with significantly higher rates of local recurrence and a trend toward decreased overall survival. All patients should be assessed for cardiovascular risk using American College of Cardiology and American Heart Association guidelines. A myocardial infarction within the past 3 months is a contraindication to thoracic surgery because 20% of patients will die of reinfarction. An infarction in the past 6 months is a relative contraindication. Other major contraindications include uncontrolled arrhythmias, an FEV_1 of less than 1 L, CO_2 retention (resting Pco_2 >45 mmHg), DL_{co} <40%, and severe pulmonary hypertension.

TREATMENT

TREATMENT

Non-Small-Cell Lung Cancer

The overall treatment approach to patients with NSCLC is shown in Fig. 74-3.

OCCULT AND STAGE 0 CARCINOMAS

Patients with severe atypia on sputum cytology have an increased risk of developing lung cancer compared to those without atypia. In the uncommon circumstance where malignant cells are identified in a sputum or bronchial washing specimen but the chest imaging appears normal (TX tumor stage), the lesion must be localized. More than 90% of tumors can be localized by meticulous examination of the bronchial tree with a fiberoptic bronchoscope under general anesthesia and collection of a series of differential brushings and biopsies. Surgical resection following bronchoscopic localization has been shown to improve survival compared to no treatment. Close follow-up of these patients is indicated because of the high incidence of second primary lung cancers (5% per patient per year).

SOLITARY PULMONARY NODULE AND “GROUND-GLASS” OPACITIES

A solitary pulmonary nodule is defined as an x-ray density completely surrounded by normal aerated lung with circumscribed margins, of any shape, usually 1–6 cm in greatest diameter. The approach to a patient with a solitary pulmonary nodule is based on an estimate of the probability of cancer, determined according to the patient’s smoking history, age, and characteristics on imaging **(Table 74-8)**. Prior CXRs and CT scans should be obtained if available for comparison. A PET scan may be useful if the lesion is greater than 7–8 mm in diameter. If no diagnosis is apparent, Mayo investigators reported that clinical characteristics (age, cigarette smoking status, and prior cancer diagnosis) and three radiologic characteristics (nodule diameter, spiculation, and upper lobe location) were independent predictors of malignancy. At present, only two radiographic criteria are thought to predict the benign nature of a solitary pulmonary nodule: lack of growth over a period >2 years and certain characteristic patterns of calcification. Calcification alone, however, does not exclude malignancy; a dense central nidus, multiple punctuate foci, and “bulls eye” (granuloma) and “popcorn ball” (hamartoma) calcifications are highly suggestive of a benign lesion. In contrast, a relatively large lesion, lack of or asymmetric calcification, chest symptoms, associated atelectasis, pneumonitis, or growth of the lesion revealed by comparison with an old x-ray or CT scan or a positive PET scan may be suggestive of a malignant process and warrant further attempts to establish a histologic diagnosis. An algorithm for assessing these lesions is shown in **Fig. 74-6**.

Since the advent of screening CTs, small “ground-glass” opacities (GGOs) have often been observed, particularly as the increased sensitivity of CTs enables detection of smaller lesions. Many of these GGOs, when biopsied, are found to be atypical adenomatous hyperplasia (AAH), adenocarcinoma in situ (AIS), or minimally invasive adenocarcinoma (MIA). AAH is usually a nodule of <5 mm and is minimally hazy, also called nonsolid or ground glass (i.e., hazy slightly increased attenuation, no solid component, and preservation of bronchial and vascular margins). On thin-section CT, AIS is usually a nonsolid nodule and tends to be slightly more opaque than AAH. MIA is mainly solid, usually with a small (<5 mm) central solid component. However, overlap exists among the

imaging features of the preinvasive and minimally invasive lesions in the lung adenocarcinoma spectrum. Lepidic adenocarcinomas are usually solid but may be nonsolid. Likewise, the small invasive adenocarcinomas also are usually solid but may exhibit a small nonsolid component.

MANAGEMENT OF STAGES I AND II NSCLC

Surgical Resection of Stage I and II NSCLC

Surgical resection, ideally by an experienced thoracic surgeon, is the treatment of choice for patients with clinical stage I and II NSCLC who are able to tolerate the procedure. Operative mortality rates for patients resected by thoracic or cardiothoracic surgeons are lower compared to general surgeons. Moreover, survival rates are higher in patients who undergo resection in facilities with a high surgical volume compared to those performing fewer than 70 procedures per year, even though the higher-volume facilities often serve older and less socioeconomic advantaged populations. The improvement in survival is most evident in the immediate postoperative period. The extent of resection is a matter of surgical judgment based on findings at exploration. In patients with stage IA NSCLC, lobectomy is superior to wedge resection with respect to rates of local recurrence. There is also a trend toward improvement in overall survival. In patients with comorbidities, compromised pulmonary reserve, and small peripheral lesions, a limited resection, wedge resection, and segmentectomy (potentially by video-assisted thoracoscopic surgery) may be reasonable surgical option. Pneumonectomy is reserved for patients with central tumors and should be performed only in patients with excellent pulmonary reserve. The 5-year survival rates are 60–80% for patients with stage I NSCLC and 40–50% for patients with stage II NSCLC.

Accurate pathologic staging requires adequate segmental, hilar, and mediastinal lymph node sampling. Ideally this includes a mediastinal lymph node dissection. On the right side, mediastinal stations 2R, 4R, 7, 8R, and 9R should be dissected; on the left side, stations 5, 6, 7, 8L, and 9L should be dissected. Hilar lymph nodes are typically resected and sent for pathologic review, although it is helpful to specifically dissect and label level 10 lymph nodes when possible. On the left side, level 2 and sometimes level 4 lymph nodes are generally obscured by the aorta. Although the therapeutic benefit of nodal dissection versus nodal sampling is controversial, a pooled analysis of three trials involving patients with stages I to IIIA NSCLC demonstrated a superior 4-year survival in patients undergoing resection and a complete mediastinal lymph node dissection compared with lymph node sampling. Moreover, complete mediastinal lymphadenectomy added little morbidity to a pulmonary resection for lung cancer when carried out by an experienced thoracic surgeon.

Radiation Therapy in Stages I and II NSCLC

There is currently no role for postoperative radiation therapy in patients following resection of stage I or II NSCLC with negative margins. However, patients with stage I and II disease who either refuse or are not suitable candidates for surgery should be considered for radiation therapy with *curative* intent. Stereotactic body radiation therapy (SBRT) is a technique used to treat patients with isolated pulmonary nodules (≤5 cm) who are not candidates for or refuse surgical resection. Treatment is typically administered in three to five fractions delivered over 1–2 weeks. In uncontrolled studies, disease control rates are >90%, and 5-year survival rates of up to 60% have been reported with SBRT. By comparison, survival rates typically range from 13 to 39% in patients with stage I or II NSCLC treated with standard external-beam radiotherapy. Cryoablation is another technique occasionally used to treat small, isolated tumors (i.e., ≤3 cm). However, very little data exist on long-term outcomes with this technique.

Chemotherapy in Stages I and II NSCLC

Although a landmark meta-analysis of cisplatin-based adjuvant chemotherapy trials in patients with resected stages I to IIIA NSCLC (the Lung Adjuvant Cisplatin Evaluation [LACE] Study) demonstrated a 5.4% improvement in 5-year survival for adjuvant chemotherapy compared to surgery alone, the survival benefit was seemingly confined to patients with stage II or III disease **(Table 74-9)**. By contrast, survival was actually worsened in stage IA patients with the application of adjuvant therapy. In stage IB, there was a modest improvement in survival of questionable clinical significance. Adjuvant chemotherapy was also detrimental in patients with poor performance status (Eastern Cooperative Oncology Group [ECOG] performance status = 2). These data suggest that adjuvant chemotherapy is best applied in patients with resected stage II or III NSCLC. There is no apparent role for adjuvant chemotherapy in patients with resected stage IA or IB NSCLC. A possible exception to the prohibition of adjuvant therapy in this setting is the stage IB patient with a resected lesion ≥4 cm. At present, targeted therapies and immunotherapies are not used in the adjuvant setting, unless given as part of a clinical trial.

As with any treatment recommendation, the risks and benefits of adjuvant chemotherapy should be considered on an individual patient basis. If a decision is made to proceed with adjuvant chemotherapy, in general, treatment should be initiated 6–12 weeks after surgery, assuming the patient has fully recovered, and should be administered for no more than four cycles. Although a cisplatin-based chemotherapy is the preferred treatment regimen, carboplatin can be substituted for cisplatin in patients who are unlikely to tolerate cisplatin for reasons such as reduced renal function, presence of neuropathy, or hearing impairment. No specific chemotherapy regimen is considered optimal in this setting, although platinum plus vinorelbine is most commonly used.

Neoadjuvant chemotherapy, which is the application of chemotherapy administered *before* an attempted surgical resection, has been advocated by some experts on the assumption that such an approach will more effectively extinguish occult micrometastases compared to postoperative chemotherapy. In addition, it is thought that preoperative chemotherapy might render an inoperable lesion resectable. With the exception of superior sulcus tumors, however, the role of neoadjuvant chemotherapy in stage I to III disease is not well defined. However, a meta-analysis of 15 randomized controlled trials involving more than 2300 patients with stage I to III NSCLC suggested there may be a modest 5-year survival benefit (i.e., ~5%) that is virtually identical to the survival benefit achieved with postoperative chemotherapy. Accordingly, neoadjuvant therapy may prove useful in selected cases (see below). A decision to use neoadjuvant chemotherapy should always be made in consultation with an experienced surgeon.

In should be noted that all patients with resected NSCLC are at high risk of recurrence, most of which occurs within 18–24 months of surgery, or developing a second primary lung cancer. Thus, it is reasonable to follow these patients with periodic imaging studies. Given the results of the NLST, periodic CT scans appear to be the most appropriate screening modality. Based on the timing of most recurrences, some guidelines recommend a contrasted chest CT scan every 6 months for the first 3 years after surgery, followed by yearly CT scans of the chest without contrast thereafter.

MANAGEMENT OF STAGE III NSCLC

Management of patients with stage III NSCLC usually requires a combined-modality approach. Patients with stage IIIA disease commonly are stratified into those with “nonbulky” or “bulky” mediastinal lymph node (N2) disease. Although the definition of “bulky” N2 disease varies somewhat in the literature, the usual criteria include the size of a dominant lymph node (i.e., >2–3 cm in short-axis diameter as measured by CT), groupings of multiple smaller lymph nodes, evidence of extracapsular nodal involvement, or involvement of more than two lymph node stations. The distinction between nonbulky and bulky stage IIIA disease is mainly used to select potential candidates for *upfront* surgical resection or for resection after neoadjuvant therapy. Many aspects of therapy of patients with stage III NSCLC remain controversial, and the optimal treatment strategy has not been clearly defined. Moreover, although there are many potential treatment options, none yields a very high probability of cure. Furthermore, because stage III disease is highly heterogeneous, no single treatment approach can be recommended for all patients. Key factors guiding treatment choices include the particular combination of tumor (T) and nodal (N) disease, the ability to achieve a complete surgical resection if indicated, and the patient’s overall physical condition and preferences. For example, in carefully selected patients with limited stage IIIA disease where involved mediastinal lymph nodes can be completed resected, initial surgery followed by postoperative chemotherapy (with or without radiation therapy) may be indicated. By contrast, for patients with clinically evident bulky mediastinal lymph node involvement, the standard approach to treatment is concurrent chemoradiotherapy. Nevertheless, in some cases, the latter group of patients may be candidates for surgery following chemoradiotherapy.

Absent and Nonbulky Mediastinal (N2, N3) Lymph Node Disease

For the subset of stage IIIA patients initially thought to have clinical stage I or II disease (i.e., pathologic involvement of mediastinal [N2] lymph nodes is *not* detected preoperatively), surgical resection is often the treatment of choice. This is followed by adjuvant chemotherapy in patients with microscopic lymph node involvement in a resection specimen. Postoperative radiation therapy (PORT) may also have a role for those with close or positive surgical margins. Patients with tumors involving the chest wall or proximal airways within 2 cm of the carina with hilar lymph node involvement (but not N2 disease) are classified as having T3N1 stage IIIA disease. They too are best managed with surgical resection, if technically feasible, followed by adjuvant chemotherapy if completely resected. Patients with tumors exceeding 7 cm in size also are now classified as T3 and are consider stage IIIA if tumor has spread to N1 nodes. The appropriate initial management of these patients involves surgical resection when feasible, provided the mediastinal staging is negative, followed by adjuvant chemotherapy for those who achieve complete tumor resection. Patients with T3N0 or T3N1 disease due to the presence of satellite nodules within the same lobe as the primary tumor also are candidates for surgery, as are patients with ipsilateral nodules in another lobe and negative mediastinal nodes (IIIA, T4N0 or T4N1). Although data regarding adjuvant chemotherapy in the latter subsets of patients are limited, it is often recommended.

Patients with T4N0-1 were reclassified as having stage IIIA tumors in the seventh edition of the TNM system. These patients may have involvement of the carina, superior vena cava, or a vertebral body and yet still be candidates for surgical resection in selected circumstances. The decision to proceed with an attempted resection must be made in consultation with an experienced thoracic surgeon often in association with a vascular or cardiac surgeon and an orthopedic surgeon depending on tumor location. However, if an incomplete resection is inevitable or if there is evidence of N2 involvement (stage IIIB), surgery for T4 disease is contraindicated. Most T4 lesions are best treated with chemoradiotherapy.

The role of PORT in patients with completely resected stage III NSCLC is controversial. To a large extent, the use of PORT is dictated by the presence or absence of N2 involvement and, to a lesser degree, by the biases of the treating physician. Using the Surveillance, Epidemiology, and End Results (SEER) database, a recent meta-analysis of PORT identified a significant increase in survival in patients with N2 disease but not in patients with N0 or N1 disease. An earlier analysis by the PORT Meta-analysis Trialist Group using an older database produced similar results.

Known Mediastinal (N2, N3) Lymph Node Disease

When pathologic involvement of mediastinal lymph nodes is documented preoperatively, a combined-modality approach is recommended assuming the patient is a candidate for treatment with curative intent. These patients are at high risk for both local and distant recurrence if managed with resection alone. For patients with stage III disease who are not candidates for initial surgical resection, *concurrent* chemoradiotherapy is most commonly used as the initial treatment. Concurrent chemoradiotherapy has been shown to produce superior survival compared to *sequential* chemoradiotherapy; however, it also is associated with greater host toxicities (including fatigue, esophagitis, and neutropenia). Therefore, for patients with a good performance status, concurrent chemoradiotherapy is the preferred treatment approach, whereas sequential chemoradiotherapy may be more appropriate for patients with a performance status that is not as good. For patients who are *not* candidates for a combined-modality treatment approach, typically due to a poor performance status or a comorbidity that makes chemotherapy untenable, radiotherapy alone may provide a modest survival benefit in addition to symptom palliation.

For patients with potentially resectable N2 disease, it remains uncertain whether surgery after neoadjuvant chemoradiotherapy improves survival. In an NCI-sponsored Intergroup randomized trial comparing concurrent chemoradiotherapy alone to concurrent chemoradiotherapy followed by attempted surgical resection, no survival benefit was observed in the trimodality arm compared to the bimodality therapy. In fact, patients subjected to a pneumonectomy had a worse survival outcome. By contrast, those treated with a lobectomy appeared to have a survival advantage based on a retrospective subset analysis. Thus, in carefully selected, otherwise healthy patients with nonbulky mediastinal lymph node involvement, surgery may be a

reasonable option if the primary tumor can be fully resected with a lobectomy. This is not the case if a pneumonectomy is required to achieve complete resection.

Superior Sulcus Tumors (Pancoast Tumors)

Superior sulcus tumors represent a distinctive subset of stage III disease. These tumors arise in the apex of the lung and may invade the second and third ribs, the brachial plexus, the subclavian vessels, the stellate ganglion, and adjacent vertebral bodies. They also may be associated with Pancoast syndrome, characterized by pain that may arise in the shoulder or chest wall or radiate to the neck. Pain characteristically radiates to the ulnar surface of the hand. Horner's syndrome (enophthalmos, ptosis, miosis, and anhydrosis) due to invasion of the paravertebral sympathetic chain may be present as well. Patients with these tumors should undergo the same staging procedures as all patients with stage II and III NSCLC. Neoadjuvant chemotherapy or combined chemoradiotherapy followed by surgery is reserved for those without N2 involvement. This approach yields excellent survival outcomes (>50% 5-year survival in patients with an R0 resection). Patients with N2 disease are less likely to benefit from surgery and can be managed with chemoradiotherapy alone. Patients presenting with metastatic disease can be treated with radiation therapy (with or without chemotherapy) for symptom palliation.

MANAGEMENT OF METASTATIC NSCLC

Approximately 40% of NSCLC patients present with advanced, stage IV disease at the time of diagnosis. In addition, a significant number of patients who first presented with early-stage NSCLC will eventually relapse with distant disease. Patients who have recurrent disease have a better prognosis than those presenting with metastatic disease at the time of diagnosis. Standard medical management, the judicious use of pain medications, and the appropriate use of radiotherapy and systemic therapy—which may compromise of traditional cytotoxic chemotherapy, targeted therapy, and immunotherapy depending on the specific diagnosis and molecular subtype—form the cornerstone of management. Systemic therapy palliates symptoms, improves the quality of life, and improves survival in patients with stage advanced NSCLC, particularly in patients with good performance status. Of note, the early application of palliative care in conjunction with chemotherapy is associated with improved survival and a better quality of life.

Cytotoxic Chemotherapy for Metastatic or Recurrent NSCLC

A landmark meta-analysis published in 1995 provided the earliest meaningful indication that chemotherapy could provide a survival benefit in metastatic NSCLC as opposed to supportive care alone. However, the survival benefit was seemingly confined to cisplatin-based chemotherapy regimens (hazard ratio 0.73; 27% reduction in the risk of death; 10% improvement in survival at 1 year). To date, platinum-based regimens remain the cornerstone of the cytotoxic chemotherapy regimens used for patients with metastatic NSCLC (**Table 74-10**). Several different platinum "doublet" regimens have been used—combining platinum (cisplatin or carboplatin) with another type of chemotherapy (for example, paclitaxel, docetaxel, pemetrexed, gemcitabine, or vinorelbine). Although specific tumor histology was once considered irrelevant to treatment choice in NSCLC, with the recent recognition that selected chemotherapy agents perform quite differently in squamous versus adenocarcinomas, accurate determination of histology has become essential. Specifically, in a landmark randomized phase III trial, patients with nonsquamous NSCLC were found to have an improved survival when treated with cisplatin and pemetrexed compared to cisplatin and gemcitabine. By contrast, patients with squamous carcinoma had an improved survival when treated with cisplatin and gemcitabine. This survival difference is thought to be related to the differential expression of thymidylate synthase (TS), one of the targets of pemetrexed, between tumor types. Squamous cancers have a much higher expression of TS compared to adenocarcinomas, accounting for their lower responsiveness to pemetrexed. By contrast, the activity of gemcitabine is not impacted by the levels of TS.

Maintenance Therapy for Metastatic NSCLC

Several large phase III randomized trials have failed to show a meaningful benefit for increasing the duration of platinum-doublet chemotherapy beyond four to six cycles. In fact, longer duration of platinum-doublet chemotherapy has been associated with increased toxicities and impaired quality of life. Maintenance chemotherapy in nonprogressing patients (patients with a complete response, partial response, or stable disease) is divided into two types of maintenance strategies: (1) switch maintenance therapy, where patients receive four to six cycles of platinum-based chemotherapy and are switched to an entirely different regimen; and (2) continuation maintenance therapy, where patients receive four to six cycles of platinum-based chemotherapy and then the platinum agent is discontinued but the agent it is paired with is continued **(Table 74-11)**. Two studies investigated switch maintenance single-agent chemotherapy with docetaxel or pemetrexed in nonprogressing patients following treatment with first-line platinum-based chemotherapy. Both trials randomized patients to immediate single-agent therapy versus observation and reported improvements in progression-free and overall survival. In both trials, a significant portion of patients in the observation arm did not receive therapy with the agent under investigation upon disease progression; 37% of study patients never received docetaxel in the docetaxel study and 81% of patients never received pemetrexed in the pemetrexed study. In the trial of maintenance docetaxel versus observation, survival was identical to the treatment group in the subset of patients who received docetaxel on progression, indicating this is an active agent in NSCLC. These data are not available for the pemetrexed study. Two additional trials evaluated switch maintenance therapy with erlotinib after platinum-based chemotherapy in patients with advanced NSCLC and reported an improvement in progression-free survival and overall survival in the erlotinib treatment group; however, erlotinib is not recommended in patients with EGFR wild type tumors. Bevacizumab, a monoclonal antibody against VEGF, has been shown to improve response rate, progression-free survival, and overall survival in patients with advanced disease when combined with chemotherapy. However, bevacizumab cannot be given to patients with squamous cell histology NSCLC because of their tendency to experience serious hemorrhagic effects. Currently, carboplatin/paclitaxel and bevacizumab or carboplatin/pemetrexed and bevacizumab are appropriate regimens for first-line treatment for stage IV nonsquamous NSCLC patients followed by maintenance bevacizumab or maintenance pemetrexed/bevacizumab respectively. Currently, maintenance pemetrexed following platinum-based chemotherapy in patients with advanced NSCLC is also approved by the U.S. FDA. Maintenance erlotinib is only approved in patients with EGFR mutations (see below). It should be noted that there are no

approved maintenance regimens for patients with squamous cell histology. Moreover, maintenance therapy is not without toxicity and, at this time, should be considered on an individual patient basis.

Targeted Therapies for Select Molecular Cohorts of NSCLC

As the efficacy of traditional cytotoxic chemotherapeutic agents plateaued in NSCLC, there was a critical need to define novel therapeutic treatment strategies. For a cohort of NSCLC patients, the presence of an oncogenic driver allows the use of oral therapies with significant tumor regression. These driver mutations occur in genes encoding signaling proteins that, when aberrant, drive initiation and maintenance of tumor cells. Importantly, driver mutations can serve as Achilles' heels for tumors, if their gene products can be targeted therapeutically with small-molecule inhibitors. For example, *EGFR* mutations have been detected in 10–15% of North American patients diagnosed with NSCLC. *EGFR* mutations are associated with younger age, light (<10 pack-year) and nonsmokers, and adenocarcinoma histology. Approximately 90% of these mutations are exon 19 deletions or exon 21 L858R point mutations within the EGFR TK domain, resulting in hyperactivation of both EGFR kinase activity and downstream signaling. Lung tumors that harbor activating mutations within the EGFR kinase domain display high sensitivity to small-molecule EGFR TKIs. Erlotinib, gefitinib, and afatinib are FDA-approved oral small-molecule TKIs that inhibit EGFR. Several large, international, phase III studies have demonstrated improved response rates, progression-free survival, and overall survival in patients with *EGFR* mutation–positive NSCLC patients treated with an EGFR TKI as compared with standard first-line chemotherapy regimens **(Table 74-12)**. A phase III trial also compared gefitinib to afatinib as first-line therapy in patients with *EGFR* mutation–positive NSCLC and demonstrated superior efficacy, with increasing toxicity for patients treated with afatinib. Unfortunately, all patients with *EGFR* mutation–positive NSCLC treated with EGFR TKIs eventually developed acquired resistance. Disease progression occurs usually around 12 months. Approximately 50% of patients have tumors that harbor a second site mutation, most commonly the T790M mutation occurring within exon 20. Osimertinib, a third generation mutant-selective EGFR TKI received approval in 2015 for patients who progress on erlotinib, gefitinib, or afatinib and whose tumors harbor the T790M mutation.

With the rapid pace of scientific discovery, additional driver mutations in lung cancer have been identified and targeted therapeutically with impressive clinical results. For example, chromosomal rearrangements involving the anaplastic lymphoma kinase (*ALK*) gene on chromosome 2 have been found in ~3–7% of NSCLC. The result of these *ALK* rearrangements is hyperactivation of the ALK TK domain. Similar to EGFR, *ALK* rearrangements are typically (but not exclusively) associated with younger age, light (<10 pack-year) and nonsmokers, and adenocarcinoma histology. Remarkably, *ALK* rearrangements were initially described in lung cancer in 2007, and by 2011, the first ALK inhibitor, crizotinib, received FDA approval for patients with lung tumors harboring *ALK* rearrangements. Two additional ALK inhibitors, ceritinib and alectinib, are currently approved in patients who progress on crizotinib. ALK testing may be performed via fluorescence in situ hybridization (FISH), immunohistochemistry (IHC), or next generation sequencing. *ROS1* fusions have been identified in ~1% of patients with NSCLC and similar to EGFR and ALK, *ROS* rearrangements are typically associated with younger age and light or never smoking status. Crizotinib, which inhibits both ALK and ROS1 kinases, was recently FDA approved for patients whose tumors harbor a *ROS1* fusion.

In addition to *EGFR*, *ALK*, and *ROS1* other driver mutations have been discovered with varying frequencies in NSCLC, including *KRAS*, *BRAF*, *PIK3CA*, *NRAS*, *AKT1*, *MET*, *MEK1* (*MAP2K1*), *NTRK*, and *RET*. Mutations within the *KRAS* GTPase are found in ~20% of lung adenocarcinomas. To date, however, no small-molecule inhibitors are available to specifically target mutant *KRAS*. Each of the other driver mutations occurs in less than 1–3% of lung adenocarcinomas. The great majority of the driver mutations are mutually exclusive, and there are ongoing clinical studies for their specific inhibitors. For example, the BRAF inhibitors dabrafenib and vemurafenib and the RET inhibitors cabozantinib and vandetanib have already demonstrated efficacy in patients with lung cancer harboring *BRAF* mutations or *RET* gene fusions, respectively. Most of these mutations are present in adenocarcinoma; however, mutations that may be linked to future targeted therapies in squamous cell carcinomas are emerging. In addition, there are active research efforts aimed at defining novel targetable mutations in lung cancer as well as defining mechanisms of acquired resistance to small-molecule inhibitors used in the treatment of patients with NSCLC.

Second-Line Therapy and Beyond

Second-line therapy for advanced NSCLC was almost never recommended until a seminal study in 2000 showed that docetaxel improved survival compared to supportive care alone. Little progress had been made in the second-line setting for NSCLC patients until the introduction of immunotherapy agents (see below) with only pemetrexed and docetaxel available as FDA-approved agents, and erlotinib recommended in patients with EGFR mutation–positive NSCLC who did not receive a first line EGFR TKI. Ramucirumab is a recombinant human IgG1 monoclonal antibody that targets VEGFR-2 and blocks the interaction of VEGF ligands and VEGFR-2. A phase III trial demonstrated a significant improvement in progression-free survival and overall survival when ramucirumab was combined with docetaxel as second-line therapy in patients who had progressed on platinum-based chemotherapy. Contrary to bevacizumab, ramucirumab was safe in patients with both squamous and nonsquamous NSCLC and is approved regardless of histology.

Immunotherapy

Immune checkpoint inhibitors are a novel class of agents that have significantly improved the quality of life and survival for a group of patients with advanced NSCLC. Immune checkpoint inhibitors work by blocking interactions between T cells and antigen presenting cells (APCs) or tumor cells that lead to T-cell inactivation. By inhibiting this interaction, the immune system is effectively upregulated and T cells become activated against tumor cells. Several large randomized phase III trials demonstrated superior overall survival for both the anti-PD1 antibodies, nivolumab and pembrolizumab and the anti-PD-L1 antibody atezolizumab compared to second-line docetaxel in patients with NSCLC who have progressed on platinum-based chemotherapy **(Table 74-13)**. Nivolumab and atezolizumab are approved as second-line therapy in patients who have progressed following platinum-based chemotherapy regardless of the presence of PD-L1 while pembrolizumab is approved in patients with tumors positive for PD-L1 expression in ≥1% of tumor cells. Pembrolizumab demonstrated superior efficacy to first-line platinum-based chemotherapy in patients with tumors expressing PD-L1 in greater than 50% of tumor cells, as

assessed with immunohistochemistry. A similarly designed study did not show efficacy when nivolumab was compared to chemotherapy; however, in this study patients with tumors expressing PD-L1 in greater than 1% of tumor cells were enrolled. Pembrolizumab is approved as first-line therapy in patients with tumors that are positive for PD-L1 expression in ≥50% of tumor cells. While PD-L1 has been identified as a biomarker that can predict response to immune checkpoint inhibitors, responses are observed in patients who do not appear to express the biomarker and not all PD-L1 positive patients respond to checkpoint inhibition. Complicating matter is that each checkpoint inhibitor is being developed in conjunction with its own antibody to assess PD-L1 expression and a large effort is underway to compare these tests. Further evaluation of these agents in both NSCLC and SCLC is ongoing in combination with already approved chemotherapy and targeted agents as well as other checkpoint inhibitors.

Supportive Care

No discussion of the treatment strategies for patients with advanced lung cancer would be complete without a mention of supportive care. Coincident with advances in chemotherapy and targeted therapy was a pivotal study that demonstrated that the early integration of palliative care with standard treatment strategies improved both quality of life and mood for patients with advanced lung cancer (**Chaps. 9 and 65**). Aggressive pain and symptom control is an important component for optimal treatment of these patients.

TABLE 74-8

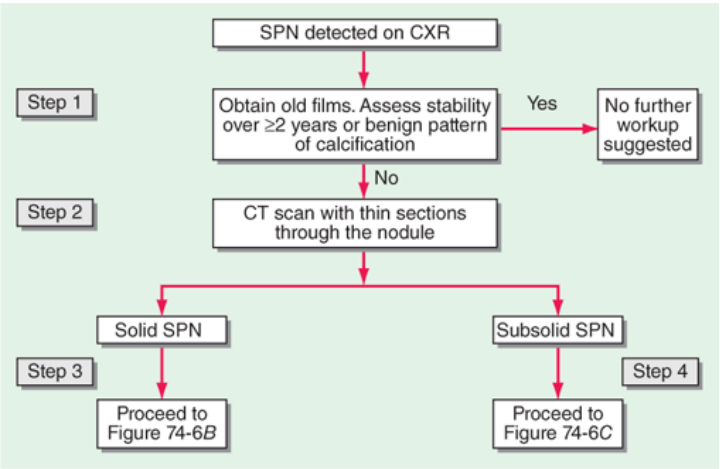
Assessment of Risk of Cancer in Patients with Solitary Pulmonary Nodules

	Risk		
Variable	**Low**	**Intermediate**	**High**
Diameter (cm)	<1.5	1.5–2.2	≥2.3
Age (years)	<45	45–60	>60
Smoking status	Never smoker	Current smoker (<20 cigarettes/d)	Current smoker (>20 cigarettes/d)
Smoking cessation status	Quit ≥7 years ago or quit	Quit <7 years ago	Never quit
Characteristics of nodule margins	Smooth	Scalloped	Corona radiata or spiculated

Source: Reproduced with permission from D Ost et al: N Engl J Med 348:2535, 2003.

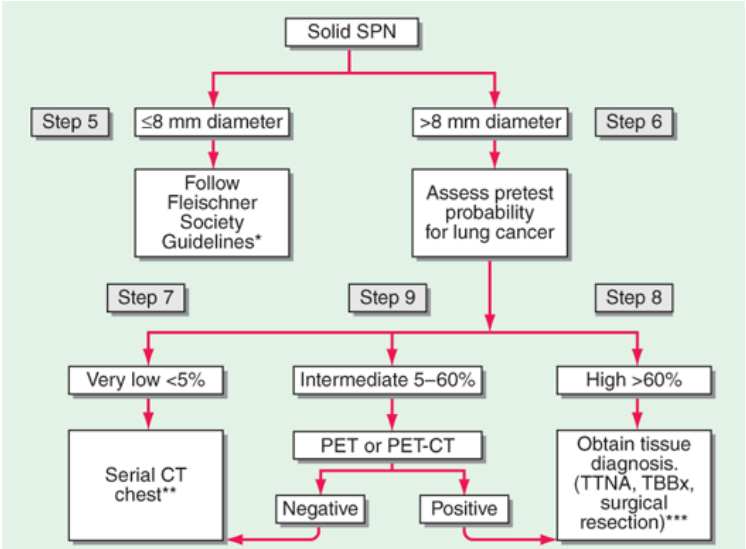
FIGURE 74-6

A. Algorithm for evaluation of solitary pulmonary nodule (SPN). ***B***. Algorithm for evaluation of solid SPN. ***C***. Algorithm for evaluation of semisolid SPN. CT, computed tomography; CXR, chest radiograph; GGN, ground-glass nodule; PET, positron emission tomography; TTBx, transbronchial biopsy; TTNA, transthoracic needle biopsy. *(Adapted from VK Patel et al: Chest 143:840, 2013.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



*Fleischner society guidelines; modified from: H. MacMahon, et al: *Radiology* 2005; 237;395–400

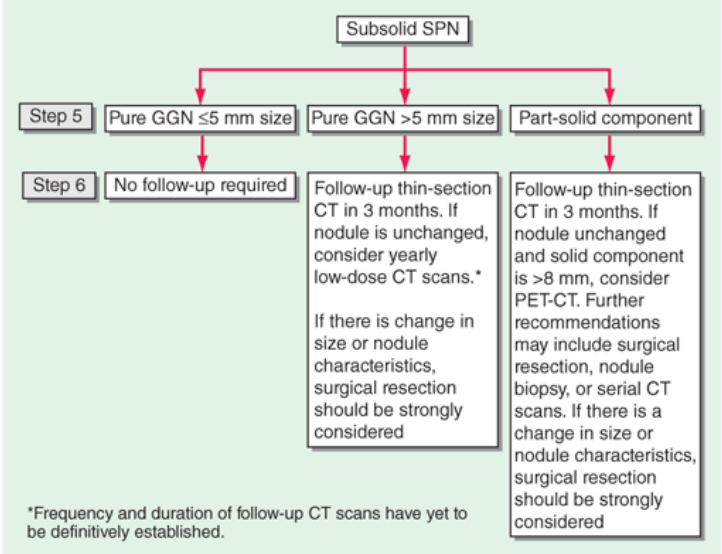
Nodule size (a):	Low-risk patient (b):	High-risk patient (c):
≤4 mm	No follow-up needed (d)	Follow-up at 12 months; if unchanged, no further follow-up
>4–≤6 mm	Follow-up CT at 12 months; if unchanged, no further follow-up	Follow-up CT at 6–12 months; then 18–24 months if no change
>6–≤8 mm	Follow-up CT at 6–12 months; then 18–24 months if no change	Follow-up CT at 3–6 months; then 9–12 and 24 months if no change
>8 mm	Follow-up CT at 3, 9, and 24 months; dynamic contrast-enhanced CT, PET, and/or biopsy	Same as low-risk patient

(a) Average of largest and smallest axial diameters of the nodule
(b) No smoking history and absence of other risk factors
(c) Previous or current smoking history or other risk factors
(d) Risk of malignancy (<0.1%) is substantially lower than for an asymptomatic smoker

**ACCP guidelines (see MK Gould et al: Chest 132(suppl 3):108s, 2007).
***Consider patient preference, severity of medical comorbidities, center specific expertise prior to tissue diagnosis.

B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



*Frequency and duration of follow-up CT scans have yet to be definitively established.

C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 74-9

Adjuvant Chemotherapy Trials in Non-Small-Cell Lung Cancer

Trial	Stage	Treatment	No. of Patients	5-Year Survival (%)	*p*
IALT	I–III	Cisplatin-based Control	932 835	44.5 40.4	< .03
BR10	IB–II	Cisplatin + vinorelbine Control	242 240	69 54	.03
ANITA	IB–IIIA	Cisplatin + vinorelbine Control	407 433	60 58	.017
ALPI	I–III	MVP Control	548 540	50 45	.49
BLT	I–III	Cisplatin-based Control	192 189	60 58	.90
CALGB	IB	Carboplatin + paclitaxel	173 171	59 57	.10

Abbreviations: ALPI, Adjuvant Lung Cancer Project Italy; ANITA, Adjuvant Navelbine International Trialist Association; BLT, Big Lung Trial; CALGB, Cancer and Lung Cancer Group B; IALT, International Adjuvant Lung Cancer Trial; MVP, mitomycin, vindesine, and cisplatin.

TABLE 74-10

First-Line Chemotherapy Trials for Metastatic Non-Small-Cell Lung Cancer

Trial	Regimen	No. of Patients	RR (%)	Median Survival (months)
ECOG1594	Cisplatin + paclitaxel	288	21	7.8
	Cisplatin + gemcitabine	288	22	8.1
	Cisplatin + docetaxel	289	17	7.4
	Carboplatin + paclitaxel	290	17	8.1
TAX-326	Cisplatin + docetaxel	406	32	11.3
	Cisplatin + vinorelbine	394	25	10.1
	Carboplatin + docetaxel	404	24	9.4
EORTC	Cisplatin + paclitaxel	159	32	8.1
	Cisplatin + gemcitabine	160	37	8.9
	Paclitaxel + gemcitabine	161	28	6.7
ILCP	Cisplatin + gemcitabine	205	30	9.8
	Carboplatin + paclitaxel	204	32	9.9
	Cisplatin + vinorelbine	203	30	9.5
SWOG	Cisplatin + vinorelbine	202	28	8.0
	Carboplatin + paclitaxel	206	25	8.0
FACS	Cisplatin + irinotecan	145	31	13.9
	Carboplatin + paclitaxel	145	32	12.3
	Cisplatin + gemcitabine	146	30	14.0
	Cisplatin + vinorelbine	145	33	11.4
Scagliotti	Cisplatin + gemcitabine	863	28	10.3
	Cisplatin + pemetrexed	862	31	10.3

Abbreviations: ECOG, Eastern Cooperative Oncology Group; EORTC, European Organization for Research and Treatment of Cancer; ILCP, Italian Lung Cancer Project; SWOG, Southwest Oncology Group; FACS, Follow-up After Colorectal Surgery.

TABLE 74-11

Maintenance Therapy Trials

			Survival	
Group	**CT**	**No. of Patients**	**OS (months)**	**PFS (months)**
Switch Maintenance				
Fidias	Immediate docetaxel Delayed docetaxel	153 156	12.3 9.7	5.7 2.7
Ciuleanu	Pemetrexed BSC	444 222	13.4 10.6	4.3 2.6
Paramount	Pemetrexed BSC	472 297	13.9 11.0	4.1 2.8
ATLAS	Bev + erlotinib Bev + placebo	384 384	15.9 13.9	4.8 3.8
SATURN	Erlotinib Placebo	437 447	12.3 11.1	2.9 2.6
Continuation Maintenance				
ECOG4599	Bev 15 mg/kg BSC	444 434	12.3 10.3	6.2 4.5
AVAiL	Bev 15 mg/kg Bev 7.5 mg/kg Placebo	351 345 347	13.4 13.6 13.1	6.5 6.7 6.1
POINTBREAK	Pemetrexed + Bev 15 mg/kg Bev 15 mg/kg			8.6 6.9

Abbreviations: Bev, bevacizumab; BSC, best supportive care; CT, chemotherapy; OS, overall survival; PFS, progression-free survival.

TABLE 74-12

Results of Phase III Trials Comparing Chemotherapy and First-Line EGFR TKI in *EGFR* Mutation–Positive Patients

Study	Therapy	No. of Patients	ORR (%)	PFS (months)
IPASS	CbP	129	47	6.3
	Gefitinib	132	71	9.3
EURTAC	CG	87	15	5.2
	Erlotinib	86	58	9.7
OPTIMAL	CG	72	36	4.6
	Erlotinib	82	83	13.1
NEJOO2	CG	114	31	5.4
	Gefitinib	114	74	10.8
WJTOG3405	CD	89	31	6.3
	Gefitinib	88	62	9.2
LUX LUNG 3	CP	115	23	6.9
	Afatinib	230	56	11.1
LUX LUNG 6	CG	122	23	5.6
	Afatinib	242	67	11.0

Abbreviations: CbP, carboplatin and paclitaxel; CD, cisplatin and docetaxel; CG, cisplatin and gemcitabine; CP, cisplatin and paclitaxel; CG, cisplatin and gemcitabine; ORR, overall response rate; PFS, progression-free survival.

TABLE 74-13

Results of Phase III Trials Comparing Chemotherapy and Immunotherapy in Patients with NSCLC

Study	Therapy	No. of Patients	OS (months)	PFS (months)
Checkmate 017	Docetaxel	137	6.0	2.8
Squamous	Nivolumab	135	9.2	3.5
Checkmate 057	Docetaxel	290	9.4	4.2
Nonsquamous	Nivolumab	292	12.2	2.3
Keynote 10	Docetaxel	212	8.5	4.0
PD-L1 ≥1%	Pembrolizumab 2 mg/kg Pembrolizumab 10 mg/kg	259 255	10.4 12.7	2.9 2.9
OAK	Docetaxel	425	10.3	2.8
	Atezolizumab	425	12.6	4.0
Keynote 24	Platinum-chemotherapy	116	NR	6.0
PD-L1 ≥50%	Pembrolizumab	73	NR	10.3
Checkmate 26	Platinum-chemotherapy	212	13.2	5.9
PD-L1 ≥1%	Nivolumab	211	14.4	4.2

Abbreviations: OS, overall survival; PFS, progression-free survival; Platinum-chemotherapy refers to first-line platinum-doublet chemotherapy.

TREATMENT

TREATMENT

Small-Cell Lung Cancer

The overall treatment approach to patients with SCLC is shown in Fig. 74-5.

SURGERY FOR LIMITED-DISEASE SMALL-CELL LUNG CANCER

SCLC is a highly aggressive disease characterized by its rapid doubling time, high growth fraction, early development of disseminated disease, and dramatic response to first-line chemotherapy and radiation. In general, surgical resection is *not* routinely recommended for patients because even patients with LD-SCLC still have occult micrometastases. However, the most recent American College of Chest Physicians Evidence-Based Clinical Practice Guidelines recommend surgical resection over nonsurgical treatment in SCLC patients with clinical stage I disease after a thorough evaluation for distant metastases and invasive mediastinal stage evaluation (grade 2C). After resection, these patients should receive platinum-based adjuvant chemotherapy (grade 1C). If the histologic diagnosis of SCLC is made in patients on review of a resected surgical specimen, such patients should receive standard SCLC chemotherapy as well.

CHEMOTHERAPY

Chemotherapy significantly prolongs survival in patients with SCLC. Four to six cycles of platinum-based chemotherapy with either cisplatin or carboplatin plus either etoposide or irinotecan has been the mainstay of treatment for nearly three decades and is recommended over other chemotherapy regimens irrespective of initial stage. Cyclophosphamide, doxorubicin (Adriamycin), and vincristine (CAV) may be an alternative for patients who are unable to tolerate a platinum-based regimen. Despite response rates to first-line therapy as high as 80%, the median survival ranges from 12 to 20 months for patients with LD and from 7 to 11 months for patients with ED. Regardless of disease extent, the majority of patients relapse and develop chemotherapy-resistant disease. Only 6–12% of patients with LD-SCLC and 2% of patients with ED-SCLC live beyond 5 years. The prognosis is especially poor for patients who relapse within the first 3 months of therapy; these patients are said to have *chemotherapy-resistant disease*. Patients are said to have *sensitive disease* if they relapse more than 3 months after their initial therapy and are thought to have a somewhat better overall survival. These patients also are thought to have the greatest potential benefit from second-line chemotherapy **(Fig. 74-7)**. Topotecan is the only FDA-approved agent for second-line therapy in

patients with SCLC. Topotecan has only modest activity and can be given either intravenously or orally. In one randomized trial, 141 patients who were not considered candidates for further IV chemotherapy were randomized to receive either oral topotecan or best supportive care. Although the response rate to oral topotecan was only 7%, overall survival was significantly better in patients receiving chemotherapy (median survival time, 26 weeks vs 14 weeks; $p = 0.01$). Moreover, patients given topotecan had a slower decline in quality of life than did those not receiving chemotherapy. Other agents with similar low levels of activity in the second-line setting include irinotecan, paclitaxel, docetaxel, vinorelbine, oral etoposide, and gemcitabine. Clearly novel treatments for this all too common disease are desperately needed.

THORACIC RADIATION THERAPY

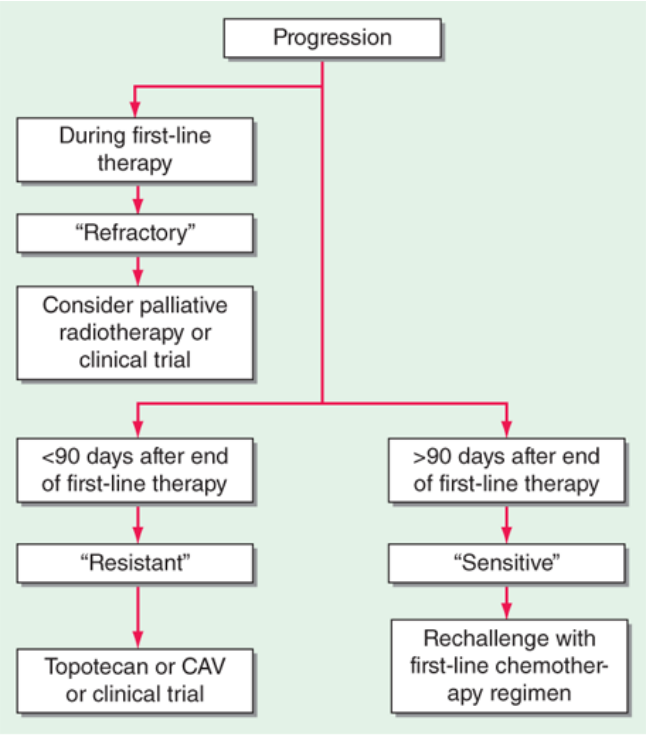
Thoracic radiation therapy (TRT) is a standard component of induction therapy for good performance status and limited-stage SCLC patients. Meta-analyses indicate that chemotherapy combined with chest irradiation improves 3-year survival by ~5% as compared with chemotherapy alone. The 5-year survival rate, however, remains disappointingly low at ~10–15%. Most commonly, TRT is combined with cisplatin and etoposide chemotherapy due to a superior toxicity profile as compared to anthracycline-containing chemotherapy regimens. As observed in locally advanced NSCLC, *concurrent* chemoradiotherapy is more effective than *sequential* chemoradiation but is associated with significantly more esophagitis and hematologic toxicity. Ideally TRT should be administered with the first two cycles of chemotherapy because later application appears slightly less effective. If for reasons of fitness or availability, this regimen cannot be offered, TRT should follow induction chemotherapy. With respect to fractionation of TRT, twice-daily 1.5-Gy fractioned radiation therapy has been shown to improve survival in LD-SCLC patients but is associated with higher rates of grade 3 esophagitis and pulmonary toxicity. Although it is feasible to deliver once-daily radiation therapy doses up to 70 Gy concurrently with cisplatin-based chemotherapy, there are no data to support equivalency of this approach compared with the 45-Gy twice-daily radiotherapy dose. Therefore, the current standard regimen of a 45-Gy dose administered in 1.5-Gy fractions twice daily for 30 days is being compared with higher-dose regimens in two phase III trials, one in the United States and one in Europe. Patients should be carefully selected for concurrent chemoradiation therapy based on good performance status and adequate pulmonary reserve. The role of radiotherapy in ED-SCLC is largely restricted to palliation of tumor-related symptoms such as bone pain and bronchial obstruction.

PROPHYLACTIC CRANIAL IRRADIATION

Prophylactic cranial irradiation (PCI) should be considered in all patients with either LD-SCLC or ED-SCLC who have responded well to initial therapy. A meta-analysis including seven trials and 987 patients with LD-SCLC who had achieved a complete remission after upfront chemotherapy yielded a 5.4% improvement in overall survival for patients treated with PCI. In patients with ED-SCLC who have responded to first-line chemotherapy, a prospective randomized phase III trial showed that PCI reduced the occurrence of symptomatic brain metastases and prolonged disease-free and overall survival compared to no radiation therapy. Long-term toxicities, including deficits in cognition, have been reported after PCI but are difficult to sort out from the effects of chemotherapy or normal aging.

FIGURE 74-7

Management of recurrent small-cell lung cancer (SCLC). CAV, cyclophosphamide, doxorubicin, and vincristine. *(Adapted with permission from JP van Meerbeeck et al: Lancet 378:1741, 2011.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THYMIC TUMORS

Thymic tumors are rare malignancies accounting for 0.5–1.5% of all malignancies in the United States with a higher incidence among Asian populations. They are particularly rare among children and young adults with incidence peaking in the fifth decade of life. There is no difference between sexes and no clear risk factors have been identified.

CLINICAL MANIFESTATIONS

The majority of thymic tumors occur in the anterior mediastinum. Approximately 40% of patients with mediastinal masses will be asymptomatic with an incidental finding on chest imaging. In patients presenting with an anterior mediastinal mass, if appropriate, serum beta-HCG (human chorionic gonadotropin) and α fetoprotein (AFP) should be sent to rule out a germ cell tumor. A patient with a sign or symptom of thymoma or thymic carcinoma may present with chest pain, dyspnea, cough or superior vena cava syndrome secondary to effects on adjacent organs or a paraneoplastic syndrome, most commonly myasthenia gravis, pure red cell aplasia or hypogammaglobulinemia. More rare paraneoplastic syndromes include limbic encephalitis, aplastic anemia, hemolytic anemia, and autoimmune disease such as Sjogrens syndrome, polymyositis, rheumatoid arthritis, ulcerative colitis among others.

STAGING

Given the rarity of the tumor, patients with suspected thymoma should be evaluated by a multidisciplinary team including a surgeon, medical and radiation oncologist as well as pathologist with experience in treating the disease. A CT scan of the chest with contrast is recommended to determine if the mass is resectable based on relationship to surrounding structures. An MRI with contrast may be performed if clinically indicated. A PET scan may be useful in the evaluation of a patient with thymic tumors although may be less useful in the staging of thymoma compared to thymic carcinoma. A core needle biopsy is considered standard of care for obtaining a histological diagnosis of an anterior mediastinal tumor. This may be obtained via CT or ultrasound imaging. However, in some circumstances a mediastinoscopy or open biopsy may be required.

Thymomas are commonly staged using the Masaoka system or the World Health Organization (WHO) staging system as described in **Table 74-14**. WHO type A, AB, and B1 tend to be more well-differentiated, type B2 and B3 are moderately differentiated, and C are poorly differentiated.

TABLE 74-14

Staging Thymic Tumors

Masaoka Stage	Definition
I	Grossly and microscopically encapsulated
IIA	Microscopic transcapsular invasion
IIB	Macroscopic invasion into surrounding tissue excluding pericardium, lung, and great vessels
III	Macroscopic invasion into neighboring organs of the lower neck or upper chest
IVA	Pleural or pericardial dissemination
IVB	Hematogenous or lymphatic dissemination to distal organs
WHO	
A	Tumor with few lymphocytes
AB	Tumor with features of type A and foci rich in lymphoctyes
B1	Tumor with features of normal epithelial cells with vesicular nuclei and distinct nucleoli and an abundant population of lymphocytes. Also known as cortical thymoma, lymphocyte-rich thymoma
B2	Thymoma with no or mild atypia with round or polygonal shaped cells with small component of lymphocytes
B3	Well differentiated thymic carcinoma with mild atypia
C	Thymic carcinoma with high atypia

TREATMENT

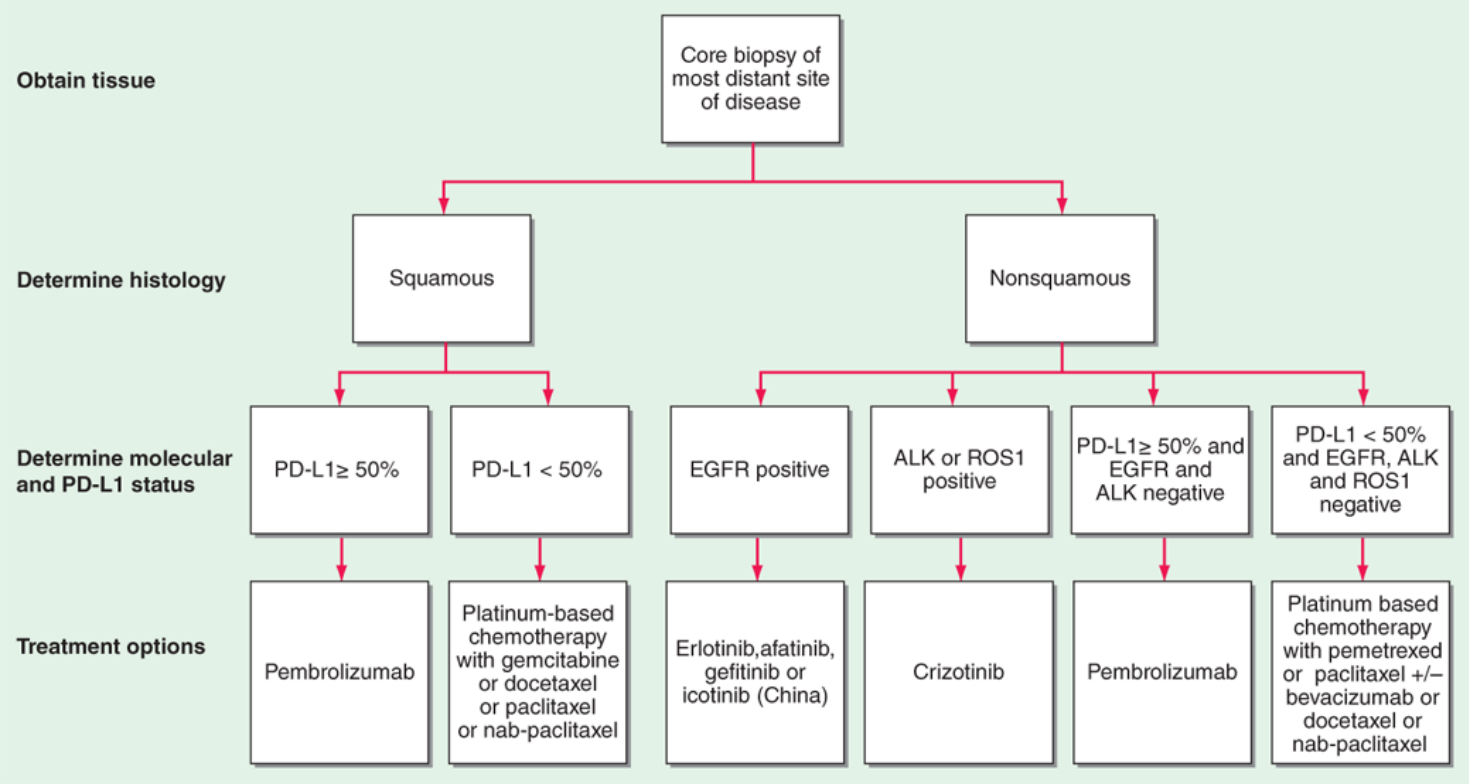
Surgical resection is the mainstay of treatment for patients with Masaoka type I and II thymic tumors. In patients with type III and IV who are potentially resectable thymic tumors, neoadjuvant chemotherapy may be given to decrease the tumor size and allow for a resection with negative margins. Surgery remains controversial and provides a limited role in the treatment of stage III and IV disease. No additional therapy may be required in patients with type I who have a resection with negative margins. Postoperative radiation therapy may be recommended based on extracapsular extension and the presence of positive margins in patients with type II or III thymic tumors or histological evaluation WHO B3 and C. Radiation therapy may be beneficial in patients with locally advanced disease (type III or IV) or in patients with symptoms secondary to compression of surrounding structures. Chemotherapy with cisplatin, doxorubicin, and cyclophosphamide (CAP) remains the mainstay of therapy in the neoadjuvant and adjuvant setting as well as first-line therapy in patients with metastatic thymoma, while carboplatin and paclitaxel are often employed in patients with thymic carcinoma. Limited additional agents are recommended based on small phase II trials as second-line therapy and beyond.

SUMMARY

The management of NSCLC has undergone major change in the past decade. To a lesser extent, the same is true for SCLC and thymic tumors. For patients with early-stage disease, advances in radiotherapy and surgical procedures as well as new systemic therapies have greatly improved prognosis in all diseases. For patients with advanced lung cancer, major progress in understanding tumor genetics and tumor immunology has led to the development of rational targets and specific inhibitors which have documented efficacy in specific subsets of NSCLC. Furthermore, increased understanding of how to activate the immune system to drive antitumor immunity has proven to be a successful therapeutic strategy for a subset of patients with advanced lung cancer. In **Fig. 74-8**, we propose an algorithm of the treatment approach for patient with stage IV NSCLC. However, the reality is that only a small subset of patients responds to immune checkpoint inhibitors and the majority of patients treated with targeted therapies or chemotherapy eventually develop resistance, which provides strong motivation for further research and enrollment of patients onto clinical trials in this rapidly evolving area.

FIGURE 74-8

Approach to first-line therapy in a patient with stage IV non-small-cell lung cancer (NSCLC).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ACKNOWLEDGMENT

David Johnson contributed to this chapter in the prior edition and material from that chapter has been retained here.

FURTHER READING

Antonio SJ et al: Nivolumab alone and nivolumab plus ipilimumab in recurrent small cell lung cancer (CheckMate 032): A multicenter, open-label, phase ½ trial. Lancet Oncol 17:883, 2016. [PubMed: 27269741]

Borghaei H et al: Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med 373:1627, 2015. [PubMed: 26412456]

Herbst RS et al: Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): A randomised controlled trial. Lancet 387:1540, 2016. [PubMed: 26712084]

Jänne PA et al: AZD9291 in EGFR inhibitor-resistant non-small-cell lung cancer. N Engl J Med 372:1689, 2015. [PubMed: 25923549]

Kris M et al: Using multiplexed assays of oncogenic drivers in lung cancers to selected targeted drugs. JAMA 311:1998, 2014. [PubMed: 24846037]

Mok TS et al: Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med 361:947, 2009. [PubMed: 19692680]

Shaw A et al: Ceritinib in ALK-rearranaged non-small cell lung cancer. N Engl J Med 370:1189, 2014. [PubMed: 24670165]

Shaw A et al: Crizotinib in ROS-1 rearranged non-small cell lung cancer. N Engl J Med 371:1963, 2014. [PubMed: 25264305]

Solomon B et al: First-line crizotinib versus chemotherapy in ALK-positive lung cancer. N Engl J Med 371:2167, 2014. [PubMed: 25470694]

Travis WD et al: The 2015 World Health Organization classification of lung tumors. J Thorac Oncol 10:1243, 2015. [PubMed: 26291008]

Chapter 75: Breast Cancer

Daniel F. Hayes; Marc E. Lippman

INTRODUCTION

Breast cancer is a malignant proliferation of epithelial cells lining the ducts or lobules of the breast. In the year 2017, ~247,000 cases of invasive and 61,000 cases of in situ breast cancer and 41,000 deaths will occur in the United States. In addition, ~2000 men will be diagnosed with breast cancer. Epithelial malignancies of the breast are the most common cause of cancer in women (excluding skin cancer), accounting for about one-third of all cancer in women. As a result of improved treatment and earlier detection, the mortality rate from breast cancer has begun to decrease very substantially in the United States. This chapter does not consider rare malignancies presenting in the breast, such as sarcomas and lymphomas, but focuses on the epithelial cancers.

EPIDEMIOLOGY AND RISK FACTORS

Breast cancer is principally a disease of older women. Seventy-five percent of all breast cancers occur in women aged >50 years. The female-to-male ratio is ~150:1. It is also a hormone-dependent disease. Women without functioning ovaries, or who experience an early menopause, and who never receive combination estrogen/progesterone replacement therapy, are much less likely to develop breast cancer than those who have a normal menstrual history. A log-log plot of incidence versus age for breast cancer shows two components: a straight-line increase with age but with a decrease in slope beginning at the age of menopause. Length of menstrual life—particularly the fraction occurring before first full-term pregnancy—is a substantial component of the total risk of breast cancer. Breast cancer risk is increased in women with early menarche, late first full-term pregnancy, and late menopause. These three factors account for 70–80% of the variation in breast cancer frequency in different countries. Also, duration of maternal nursing correlates with substantial risk reduction independent of either parity or age at first full-term pregnancy.

International variation and immigration statistics of incidence provide insight into hormonal carcinogenesis. A woman living to age 80 years in North America has one chance in nine of developing invasive breast cancer. Asian women have traditionally had only 1/5th to 1/10th the risk of breast cancer of women in North America or Western Europe. However, with shifts from agrarian to industrialized economic systems, and in immigrant populations, Asian women living in modern, Western-style environments have risks identical to those of their Western counterparts.

Presumably, these differences are secondary to menstrual, and associated intrinsic estrogen exposure, histories. However, differences in diets have also been implicated, although the role of diet in breast cancer etiology is controversial. While there are associative links between total caloric and fat intake and breast cancer risk, the exact role of fat in the diet is unproven and may actually intersect with menstrual history and estrogenic exposure.

Central obesity is both a risk factor for occurrence and recurrence of breast cancer. Moderate alcohol intake also increases the risk by an unknown mechanism. Folic acid supplementation appears to modify risk in women who use alcohol but is not additionally protective in abstainers. Recommendations favoring abstinence from alcohol must be weighed against other social pressures and the possible cardioprotective effect of moderate alcohol intake. Chronic low-dose aspirin use is associated with a decreased incidence of breast cancer. Depression is also associated with both occurrence and recurrence of breast cancer.

Exogenous use of female hormones also plays a role in breast cancer incidence. Oral contraceptive use causes a small increased risk of breast cancer. However, this risk is more than balanced by avoidance of an undesired pregnancy and a substantial protective effect against ovarian epithelial and endometrial cancers.

Hormone replacement therapy (HRT) with conjugated equine estrogens plus progestins increases the risk of breast cancer and adverse cardiovascular events, but decreases the risk of bone fractures and colorectal cancer. On balance, there appear to be more negative events with HRT; 6–7 years of HRT nearly doubled the risk of breast cancer. Of note, administration of conjugated estrogens alone (estrogen replacement therapy in women who have had hysterectomies) produces no significant increase in breast cancer incidence. Thus, there are serious concerns about long-term HRT, especially in combination with progestins, in terms of cardiovascular disease and breast cancer. No comparable safety data are available for other less potent forms of estrogen replacement, such as bioequivalent estrogen found in soy, and they should not be routinely used as substitutes. Rapid decrease in the number of women on HRT has already led to a coincident decrease in breast cancer incidence. HRT in women previously diagnosed with breast cancer, especially of the subtype that expresses estrogen receptors, increases recurrence rates.

In addition to the other factors, radiation is a risk factor in younger women. Women who have been exposed before age 30 years to radiation in the form of multiple fluoroscopies (200–300 cGy) or treatment for Hodgkin's disease (>3600 cGy) have a substantial increase in risk of breast cancer, whereas radiation exposure after age 30 years appears to have a minimal carcinogenic effect on the breast.

GENETIC CONSIDERATIONS



The genetics of breast cancer require an understanding of the distinction between inherited, germline genetic differences among individuals and acquired, somatic genetic changes within cancers. The former, often called single nucleotide polymorphisms (SNPs), if deleterious, may lead to higher susceptibility to developing cancer and/or to a patient's response to or toxicity from a given treatment (pharmacogenetics). Somatic genetic changes that are not inherited, including mutations, amplifications, deletions, translocations,

and others, are responsible for the malignant behavior of a cancer, including unrestrained proliferation, as well as extravasation from one site and migration and establishment of metastases into another.

In this regard, human breast cancer is a clonal disease. One or more transformed cells, which arise due to a combination of inherited germline susceptibility and environmentally driven somatic changes, are eventually able to express full malignant potential. Thus, breast cancer may exist for a long period as either a noninvasive disease or an invasive but nonmetastatic disease. These facts have significant clinical ramifications, including overdiagnosis of biologically nonmalignant but anatomically apparent cancers.

Germline Genetic Susceptibility

Although family history is an important risk factor, for most women the increased risk associated with a family member who has had breast cancer appears to be related to both a weak, and probably multi-gene germline susceptibility and/or similar exposure to environmental/life style risk factors. Not >10% of human breast cancers can be linked directly to single germline SNPs. However, when they are present, the relative and absolute risk for that individual's developing breast, and other, cancers in her lifetime are extraordinary.

Of these, the *BRCA1* and *2* genes are the best characterized and have the greatest clinical importance. *BRCA1* has been identified at the chromosomal locus 17q21; this gene encodes a zinc finger protein, and the protein product functions as a transcription factor and is involved in gene repair. Women who inherit a mutated allele of this gene from either parent have at least a 60–80% lifetime chance of developing breast cancer and about a 33% chance of developing ovarian cancer. The cancers that arise within a *BRCA1*-mutated patient are almost exclusively negative for estrogen and progesterone receptors (ER, PgR) and for human epidermal receptor 2 (HER2) (so-called "triple negative" breast cancers), and ~20% of women with triple negative breast cancers will be positive for deleterious germline BRCA1 SNPs. Nonetheless, the risk of breast cancer penetrance is variable within the *BRCA1*-affected population and is higher among women born after 1940, presumably due to promotional effects of hormonal factors. Men who carry a mutant allele of the gene have an increased incidence of prostate cancer and breast cancer.

BRCA2, which has been localized to chromosome 13q12, is also associated with an increased incidence of breast cancer in women. It should be noted that cancers that arise in *BRCA2* contexts are more likely to be ER positive, compared to those in *BRCA1* families, in which they are almost universally negative for ER, PgR, and HER2 expression. Of interest, men with *BRCA2* deleterious SNPs also have a higher risk of breast cancer, although most male breast cancer cases do not occur in *BRCA2*-mutated men, and the risk of breast cancer in men who do carry the *BRCA2* mutation is lower than that in women with this genetic abnormality.

Germline mutations in *BRCA1* and *BRCA2* can be readily detected in blood tests of normal circulating leucocytes. However, most experts do not recommend testing all women, since the rate of germline SNPS in this gene in the general population is quite low (well below 1%) and the tests are not 100% accurate. Further, it is not infrequent to identify variants of unknown significance (VUS) that may increase patient anxiety without a clear-cut set of recommendations about management. Consensus guidelines on who should be tested include any patient with a triple negative breast cancer and any patient with contralateral breast cancer

or who has a first-degree relative (mother, father, or sister) with breast cancer. Further, any man with breast cancer should also be tested. Some guidelines suggest testing any patient with breast cancer who is of Ashkenazi descent, since the incidence in this population of a specific founder *BRCA1* mutation (substitution of adenine for guanine at position 185) is ~2%. Patients with these mutations should be counseled appropriately.

Over the last 5 years, panels of germline genes have been offered in addition to *BRCA1* and *2*. These include genes that are known to be risk factors for breast cancer if the individual harbors deleterious SNPs, including *p53*, *PTEN*, and *PALB1*. However, several of the other genes included in these panels are less well-studied, and therefore it is less clear how to counsel affected individuals.

Somatic Genetic Changes in Breast Cancer

Abnormalities in these and other genes can also be acquired, leading to breast cancer and its specific behavior. The specific causes of these mutations in breast cancer are generally unknown. A *p53* mutation is present in ~40% of human breast cancers as an acquired defect. Acquired mutations in *PTEN* occur in ~10% of the cases. *BRCA1* mutation in sporadic primary breast cancer has not been reported. However, decreased expression of *BRCA1* messenger RNA (mRNA) (possibly via gene methylation) and abnormal cellular location of the BRCA1 protein have been found in some breast cancers. Loss of heterozygosity of *BRCA1* and *BRCA2* suggests that tumor-suppressor activity may be inactivated in sporadic cases of human breast cancer.

Approximately 80% of all breast cancers overexpress ER. Many of these cancers respond to antiestrogen treatments. Likewise, increased expression of the dominant oncogene *erbB2*, often due to amplification, occurs in approximately one-quarter of human breast cancer cases. The product of this gene, HER2, contributes to transformation of human breast epithelium. HER2 is the target of effective systemic therapy in adjuvant and metastatic disease settings.

A series of other acquired “driver” mutations has been identified in sporadic breast cancer by major sequencing consortia. Of interest, activating mutations in the gene that encodes for ER (*ESR1*) have been reported in ~20% of metastatic breast cancers after prior endocrine treatment, but almost never in untreated primary cancers. Similarly, activating mutations in *erb2* are reported in 3–5% of breast cancers. Both these findings may have therapeutic implications. Multiple academic and commercial entities are offering exon sequencing for these and many other possible mutations on either tumor biopsies or on circulating DNA shed from tumors. Unfortunately, most occur in no more than 5% of cases. Further, they are either not associated with any known targeted therapeutic agents, or the abnormalities are associated with response to an agent in another disease, but at present not in breast cancer. Therefore, while appealing, “personalized medicine” is for now more of a dream than a reality.

PREVENTION OF BREAST CANCER

One major reason to determine risk would be to develop and apply effective prevention strategies. These might either be lifestyle changes or surgical or pharmacologic interventions. At present, although diet and

exercise are certainly recommended approaches to healthy living, none has been proven to specifically decrease a woman's risk of breast cancer. Avoidance of combined estrogen/progestin HRT avoids their associated increased risk of breast cancer.

Prophylactic removal of the breasts is an effective, albeit usually unacceptable, preventive strategy. Retrospective and prospective registries have demonstrated that bilateral prophylactic mastectomies reduce the risk of breast cancer incidence and mortality by more than 95%. Because breasts are not encapsulated organs, some normal breast tissue is always left behind, and therefore women who elect to have prophylactic mastectomies should be counseled that they still have some risk of developing a new breast cancer. Because of its obvious adverse effect on sexuality, cosmesis, and breast-feeding, this approach is not considered appropriate for a woman of average risk.

As noted, cessation of menses and/or other means of reducing estrogen exposure, such as aromatase inhibition in postmenopausal women, and use of the selective estrogen receptor modulators (SERMs) tamoxifen and raloxifene are effective methods to lower breast cancer risk. So-called chemoprevention with SERMs or aromatase inhibition lowers risk of ER-positive breast cancer by approximately one-third to one-half, although it has no effect on the more lethal ER-negative breast cancers. Of interest, prophylactic bilateral oophorectomy and salpingo-oophorectomy, which is often performed in women with high genetic risk (such as those with inherited *BRCA1*/*2* deleterious SNPs), also reduces breast cancer risk.

SCREENING FOR BREAST CANCER

A recent review by the American Cancer Society (ACS) supports the perception that screening mammography reduces breast cancer mortality by one-quarter to one-third in women aged ≥50 years. The data for a relative reduction in breast cancer mortality for women between ages 40 and 50 years are almost as positive; however, since the incidence of breast cancer is much lower in younger women, the number of women whose lives are saved is much lower than in older women, and because they have denser breasts, and therefore there are more false-positive findings and the positive predictive factor is lower.

Further, screening mammography and early detection are more likely to identify tumors at a stage more appropriate for conservative local therapy. Better technology, including digitized mammography, routine use of magnified views, and greater skill in mammographic interpretation, have all improved the accuracy of mammography. Newer diagnostic techniques (magnetic resonance spectroscopy [MRI], positron emission tomography [PET], etc.) appear to have higher sensitivity, but their specificity is often lower. Further, many authors have raised concern about diagnosis of anatomically defined cancers that may be biologically insignificant, raising the specter of overdiagnosis and treatment. Since none of these newer technologies has been shown to be superior to mammography in regards to mortality reduction, screening of women with standard risk by any technique other than mammography is not recommended.

Screening with more sensitive but less specific techniques, in particular MRI, is recommended for women with genetic risk, such as *BRCA1* or *BRCA2* carriers or those with Li-Fraumeni, Cowden's, or Bannayan-Riley-Ruvalcaba syndromes; untested first-degree relatives of women with cancer; women with a history of

radiation therapy to the chest between ages 10 and 30 years; or women with a lifetime risk of breast cancer of at least 20%. In these women, the positive predictive value of MRI is higher because of the higher incidence of cancer, and, furthermore, many of them are considering prophylactic mastectomy as an alternative, and therefore the lower specificity and risk of a false positive finding has been considered more acceptable.

Research does not show a clear benefit of individual self-examination or by physical breast examinations done by a health professional. Because of this lack of evidence, regular clinical breast examination and breast self-examination are not recommended. Still, all women should be familiar with how their breasts normally look and feel and report any changes to a health care provider right away. Moreover, because the breasts are a common site of potentially fatal malignancy in women, examination of the breast is an essential part of the physical examination. Although breast cancer in men is unusual, unilateral lesions should be evaluated in the same manner as in women, with the recognition that gynecomastia in men can sometimes begin unilaterally and is often asymmetric.

EVALUATION OF BREAST MASSES IN MEN AND WOMEN

Virtually all breast cancer is diagnosed by biopsy of a nodule detected either on a mammogram or by palpation. Algorithms have been developed to enhance the likelihood of diagnosing breast cancer and reduce the frequency of unnecessary biopsy.

THE PALPABLE BREAST MASS

If a patient brings a breast abnormality to the attention of a health care giver, or if a lesion is appreciated during routine examination, proper attention needs to be given to ensure appropriate evaluation and treatment. Lesions with certain features are more likely to be cancerous. These include enigmatically, painless masses, and, more importantly, hard, irregular masses, especially if tethered or fixed to the underlying chest wall. In contrast, those that are cystic appearing on physical examination or are associated with pain, are less likely malignant. However, none of these is a terribly accurate positive or negative finding. Likewise, a negative mammogram in the presence of a persistent lump in the breast does not exclude malignancy. Any concerning, and persistent, breast finding should be referred to an experienced breast diagnostician.

In premenopausal women, lesions that are either equivocal or nonsuspicious on physical examination should be reexamined in 2–4 weeks, during the follicular phase of the menstrual cycle. Days 5–7 of the cycle are the best time for breast examination. A dominant mass in a postmenopausal woman or a dominant mass that persists through a menstrual cycle in a premenopausal woman should be referred to an experienced breast diagnostician for further evaluation, including biopsy if appropriate.

Several points are essential in pursuing these management decision trees. First, risk-factor analysis is not part of the decision structure. No constellation of risk factors, by their presence or absence, can be used to exclude biopsy. Second, fine-needle aspiration should be used only in centers that have proven skill in obtaining such specimens and analyzing them. The patient and physician must be aware of a 1% risk of false negatives. Third,

additional technologies such as MRI, ultrasound, and sestamibi imaging cannot be used to exclude the need for biopsy; although in unusual circumstances, they may provoke a biopsy.

THE ABNORMAL MAMMOGRAM

Diagnostic mammography, which is performed after a palpable abnormality has been detected, should not be confused with *screening mammography*, which is performed in an asymptomatic woman with no prediscovered abnormalities. Diagnostic mammography is aimed at evaluating the rest of the breast before biopsy is performed or occasionally is part of the triple-test strategy to exclude immediate biopsy.

Subtle abnormalities that are first detected by screening mammography should be evaluated carefully by compression or magnified views. These abnormalities include clustered, heterogeneous, linear, and branching microcalcifications; densities (especially if spiculated); and new or enlarging architectural distortion. For some nonpalpable lesions, ultrasound may be helpful either to identify cysts or to guide biopsy. If there is no palpable lesion and detailed mammographic studies are unequivocally benign, the patient should have routine follow-up appropriate to the patient's age. If a nonpalpable mammographic lesion has a low index of suspicion, mammographic follow-up in 3–6 months is reasonable. However, it cannot be stressed too strongly that in the presence of a breast lump a negative mammogram does not rule out cancer, and if it persists or enlarges during follow-up, the patient should be referred to an experienced breast diagnostician.

BREAST MASSES IN THE PREGNANT OR LACTATING WOMAN

During pregnancy, the breast grows under the influence of estrogen, progesterone, prolactin, and human placental lactogen. Lactation is suppressed by progesterone, which blocks the effects of prolactin. After delivery, lactation is promoted by the fall in progesterone levels, which leaves the effects of prolactin unopposed. The development of a dominant mass during pregnancy or lactation should never be attributed to hormonal changes. A dominant mass must be treated with the same concern in a pregnant woman as any other. Breast cancer develops in 1 in every 3000–4000 pregnancies. Stage for stage, breast cancer in pregnant patients is no different from premenopausal breast cancer in nonpregnant patients. However, pregnant women often have more advanced disease because the significance of a breast mass was not fully considered and/or because of endogenous hormone stimulation. Persistent lumps in the breast of pregnant or lactating women *cannot* be attributed to benign changes based on physical findings; such patients should be promptly referred for diagnostic evaluation.

BENIGN BREAST MASSES

Only ~1 in every 5–10 breast biopsies leads to a diagnosis of cancer, although the rate of positive biopsies varies in different countries and clinical settings. These differences may be related to interpretation, medico-legal considerations, and availability of mammograms. The vast majority of benign breast masses are due to "fibrocystic" changes, a descriptive term for small fluid-filled cysts and modest epithelial cell and fibrous tissue hyperplasia. The subset of women with ductal or lobular cell proliferation (~30% of patients),

particularly the small fraction (3%) with atypical hyperplasia, have a fourfold greater risk of developing breast cancer than those women who have not had a biopsy. The increase in the risk is about nine-fold for women in this category who also have an affected first-degree relative. Thus, careful follow-up of these patients is required. By contrast, patients with a benign biopsy without atypical hyperplasia are at little risk and may be followed routinely.

STAGING

Correct staging of breast cancer patients is of extraordinary importance. Not only does it permit an accurate prognosis, but in many cases, therapeutic decision making is based largely on the TNM (primary tumor, regional nodes, metastasis) classification. Comparison with historic series should be undertaken with caution, as the staging has changed several times in the past 20 years. The current staging is complex and results in significant changes in outcome by stage as compared with prior staging systems.

NONINVASIVE BREAST CANCER

Breast cancer develops as a series of molecular changes in the epithelial cells that lead to ever more malignant behavior. Increased use of mammography has led to more frequent diagnoses of noninvasive breast cancer. These lesions fall into two groups: ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (lobular neoplasia or LCIS). The management of both entities is controversial.

Ductal Carcinoma In Situ

Proliferation of cytologically malignant breast epithelial cells within the ducts is termed *ductal carcinoma in situ* (DCIS). Atypical hyperplasia may be difficult to differentiate from DCIS. At least one-third of patients with untreated DCIS develop invasive breast cancer within 5 years. However, many low-grade DCIS lesions do not appear to progress over many years; therefore, many patients are overtreated. Unfortunately, there is no reliable means of distinguishing patients who require treatment from those who may be safely observed.

For many years, the standard treatment for DCIS was mastectomy. Although no studies have compared breast-preserving therapy to mastectomy, the ~100% ten year survival rates with the former suggest that it is a satisfactory strategy. Breast-preserving surgery alone may also be acceptable. However, although survival was identical in the two arms of a randomized trial comparing wide excision plus or minus irradiation, the latter caused a substantial reduction in the local recurrence rate as compared with wide excision alone. Addition of tamoxifen or an aromatase inhibitor (AI) to any DCIS surgical/radiation therapy regimen further improves local control. However, in the largest trial comparing the two in DCIS, anastrozole did not improve distant disease-free or overall survival compared to tamoxifen.

Several prognostic features may help to identify patients at high risk for local recurrence after either lumpectomy alone or lumpectomy with radiation therapy, and therefore might provide an indication for mastectomy. These include extensive disease; age <40; and cytologic features such as necrosis, poor nuclear grade, and comedo subtype with overexpression of *erbB2*. In summary, it is reasonable to recommend breast-

preserving surgery for patients who have a localized focus of DCIS with clear margins followed by breast irradiation and tamoxifen or anastrozole. For patients with localized DCIS, axillary lymph node dissection is unnecessary.

More controversial is the question of what management is optimal when there is any degree of invasion. Because of a significant likelihood (10–15%) of axillary lymph node involvement even when the primary lesion shows only microscopic invasion, it is prudent to do at least a sentinel lymph node sampling for all patients with any degree of invasion. Further management is dictated by the presence of nodal spread.

Lobular Neoplasia

Proliferation of cytological malignant cells within the lobules is termed *lobular neoplasia* (LCIS). Nearly 30% of patients who have had adequate local excision, or incidentally discovered LCIS, or just a biopsy a needle biopsy of a suspicious area develop a subsequent breast cancer (usually infiltrating ductal carcinoma) over the next 15–20 years. Ipsilateral and contralateral cancers are equally common. Therefore, LCIS may be considered a premalignant condition with associated elevated risk of subsequent breast cancer, rather than a form of malignancy itself, and aggressive local management seems unreasonable. Management options include careful observation with routine mammography or chemoprevention with either a SERM or an AI (for postmenopausal women) for 5 years as well as concurrent and subsequent annual mammography and semiannual physical examinations. A third option, although no more effective and associated with substantial cosmetic, and perhaps emotional, morbidity is bilateral prophylactic mastectomy.

TREATMENT

TREATMENT

Breast Cancer

BIOLOGICAL CONSIDERATIONS

One of the most important advances in our understanding of breast cancer has been the appreciation that it can be classified by gene expression patterns into a series of subtypes.

1. **Luminal:** Luminal breast cancers are almost always positive for ER and negative for HER2 amplification. They are divided into two groups:

- **Luminal A:** Luminal A tumors have the highest levels of ER expression as well as of downstream ER-dependent genes, such as PgR. They are almost universally negative or low in HER2, and they have low proliferative thrust. They are usually low grade, are most likely to respond to endocrine therapy, and have a favorable prognosis. They appear to be less responsive to chemotherapy.

- **Luminal B:** Luminal B breast cancers are also of luminal epithelial origin, but with a gene expression pattern distinct from luminal A. They tend to be PgR negative and have evidence of higher proliferative activity. They also tend to express HER2, but not to the level of the so-called “HER2 amplified” cancers.

Their grade is more often higher than luminal A cancers. Prognosis is somewhat worse. They may be more sensitive to chemotherapy.

2. ***HER2* amplified:** These tumors have amplification of the *HER2* gene on chromosome 17q and frequently exhibit coamplification and overexpression of other genes adjacent to *HER2*. Historically the clinical prognosis of such tumors was poor. However, with the advent of trastuzumab and other targeted therapies, the clinical outcome of *HER2* positive patients is markedly improved compared to 20 or more years ago.

3. **Basal:** These ER/PgR–negative and *HER2*-negative tumors (so-called triple negative) are characterized by markers of basal/myoepithelial cells. They tend to be high grade, and express cytokeratins 5/6 and 17 as well as vimentin, p63, CD10, α-smooth muscle actin, and epidermal growth factor receptor (EGFR). Patients with *BRCA1* mutations also fall within this molecular subtype. They also have stem cell characteristics.

4. **Normal breast-like:** These tumors have a gene expression profile reminiscent of nonmalignant "normal" breast epithelium. Prognosis is similar to the luminal B group. This subtype is somewhat controversial and may represent contamination of the sample by normal mammary epithelium.

5. **Claudin-low:** These cancers are often triple negative but they have low expression of cell-cell junction proteins including E-cadherin. They are frequently associated with lymphocytic infiltration.

GENERAL TREATMENT CONSIDERATIONS

Treatment of breast cancer depends on whether the patient does or does not have evidence of distant (meaning outside the breast, chest wall, and regional lymph nodes) metastases, as detected by scintigraphic or radiologic imaging and biopsy. For patients with no evidence of detectable distant metastases, the goal of therapy is cure, or at least substantial survival prolongation, and is divided into primary and systemic considerations. Primary therapies consist of surgical and radiation treatments directed toward the breast and locoregional lymph nodes. These approaches are designed to excise and eliminate the cancer and sterilize unaffected breast tissue as appropriate. Adjuvant systemic treatments, consisting of antiestrogen (or endocrine), anti-HER2, and/or chemotherapies, are given to treat micrometastases that may have already escaped to distant sites but are not yet detectable.

All treatments for breast cancer are based on prognostic and predictive factors. Prognostic factors provide an indication of how likely a cancer will recur, either locally or in distant organs, in the future if a patient is not treated with the respective treatments. Predictive factors are used to determine if a given treatment is likely to work or not, assuming the patient's prognosis justifies treatment (or further treatment assuming the patient has been treated in some manner already).

Prognostic features guide both whether and what type of primary and adjuvant systemic treatments should be pursued. Anatomic prognostic features include visual and physical examination findings of locally advanced breast cancer (T4 lesions: skin erythema ["inflammatory"] or edema ["peau d'orange"], nodules, or

ulceration or tumor fixation to the chest wall). In patients without any of these findings, the most important prognostic features are tumor size and lymph node status (TN in the staging system). As discussed below, biologic features, such as histologic tumor grade as well as ER, PgR, and HER2, are also prognostic. Over the last decade, several multiparameter tests based on gene expression have been developed to determine prognosis in patients who have node-negative, ER-positive, and HER2-negative disease.

Predictive features are usually used to guide systemic therapies. These include ER for endocrine treatments and HER2 for anti-HER2 therapies, such as trastuzumab. There are no established predictive factors to predict response to radiation treatment. The issue of chemoresistance in luminal A cancers is under large-scale investigations.

EARLY-STAGE BREAST CANCER

Primary Therapies

Prior to 1980, the Halsted radical mastectomy, in which the breast, chest wall muscles, and complete axillary nodal contents were removed, was the standard treatment of choice for women with newly diagnosed breast cancer. In the 1980s, prospective randomized trials demonstrated that recurrence and survival rates were the same with the less disfiguring modified radical mastectomy, in which the chest wall muscles were preserved and only a sampling of axillary lymph nodes were removed.

In the same decade, breast-conserving treatments, consisting of the removal of the primary tumor by some form of surgical excision (designated as lumpectomy, quadrantectomy, or partial mastectomy), were shown to result in equal, if not slightly superior, to that associated with mastectomy. Several of these trials also demonstrated that the in-breast recurrence rate was quite high in the absence of breast radiation, while it was reduced substantially if radiation was provided. Therefore, for women undergoing breast conservation, postlumpectomy radiation is usually indicated, although it may be less necessary and withheld in older women with ER-positive, node-negative breast cancer, since their risk of subsequent in-breast recurrence is quite low with surgery and endocrine therapy only. When lumpectomy with negative tumor margins is achieved and radiation is delivered appropriately, breast conservation is associated with a recurrence rate in the breast of ≤5%.

Not all patients are candidates for breast-conserving therapy. Contraindications include large tumor to breast ratio, inability to achieve clear margins with adequate cosmesis after extensive surgery, multifocal cancers, extensive four-quadrant DCIS, and inability to receive radiation. The latter issue arises in women with dermal autoimmune disease (such as lupus erythematosus), prior radiation to the site, and/or lack of available radiation treatment facilities. Further, although not contraindicated, breast-conserving therapy may be less cosmetically acceptable than mastectomy with reconstruction if the nipple-areolar complex is involved with cancer and must be sacrificed. This is a personal choice, and some women prefer mastectomy, especially those with high genetic risks for second breast cancers.

For patients who do undergo mastectomy, postoperative chest wall and regional nodal radiation is also associated with an improvement in survival if they have a high risk of local-regional recurrence, such as

tumors ≥5 cm, four or more positive axillary lymph nodes, or postoperative positive margins. Postmastectomy radiation is not indicated in women with cancers <2 cm, negative lymph nodes, and negative margins. It is considered for women who fall into the areas between these (2–5 cm, one to three positive nodes, or close margins), and is usually recommended if a patient has one to three involved axillary lymph nodes.

At present, nearly one-third of women in the United States are managed by lumpectomy, and recent data suggest that the fraction of women treated with breast-conserving therapy is decreasing. It appears that many women still undergo mastectomy who could safely avoid this procedure and probably would if appropriately counseled.

Axillary node sampling or dissection is unnecessary in many cases. Sentinel lymph node mapping and biopsy (SLNB) is generally the standard of care for women with localized breast cancer and clinically negative axilla. If SLNB is negative, more extensive axillary surgery is not required, avoiding much of the risk of lymphedema following more extensive axillary dissections. Even in the presence of sentinel lymph node involvement, further axillary surgery may not be required for selected patients, such as older women and those with ER-positive cancers.

The survival of patients who have recurrence in the breast after proper treatment (adequate surgery and radiation if indicated) is somewhat worse than that of women who do not, but it is not worse than those who suffer local-regional recurrence after mastectomy. Thus, local-regional recurrence is a negative prognostic variable for long-term survival but not the *cause* of distant metastasis. Most patients should consult with a radiation oncologist before making a final decision concerning local therapy. However, a multimodality clinic in which the surgeon, radiation oncologist, medical oncologist, and other caregivers cooperate to evaluate the patient and develop a treatment plan is usually considered a major advantage by patients.

Adjuvant Systemic Therapies

The concept of adjuvant systemic therapy is based on the observation that cancer is a condition of genetic instability, and with increasing generations of cellular replication, genetic abnormalities accumulate. Although these occur randomly, and therefore may lead to sensitivity or resistance to therapies, the latter is of greater concern. Thus, as a consequence of accumulation of mutations to resistance, almost all patients with metastatic breast cancer are destined to die with, if not of their cancer.

However, treatment with the same therapies administered earlier, in the setting of micrometastatic disease only, has been repeatedly shown to be more effective than waiting until symptomatic, documented metastases occur. Put simply, the use of systemic therapy as an adjuvant to local management of breast cancer substantially improves survival. More than half of the women who would otherwise die of metastatic breast cancer remain disease-free and experience considerable survival advantaged when treated with the appropriate adjuvant systemic regimen. These data have grown more and more impressive with longer follow-up and more effective regimens.

PROGNOSTIC VARIABLES

As noted, prognostic factors help define who most likely needs, or perhaps more importantly does not need, adjuvant systemic therapy. The most important prognostic variables are provided by *tumor staging: tumor size (T), lymph node status (N) and detectable distant metastases (M)* (Table 75-1). *Histologic classification* of the tumor has also been used as a prognostic factor. Tumors with a poor nuclear grade have a higher risk of recurrence than tumors with a good nuclear grade. Semiquantitative measures such as the Elston score improve the reproducibility of this measurement. Importantly, there is no need to perform imaging for distant metastases in a patient with no signs or symptoms of widespread disease and who has a T3 or less tumor and fewer than four involved axillary lymph nodes.

TABLE 75-1

5-Year Survival Rate for Breast Cancer by Stage

Stage	5-Year Survival, %
0	99
I	92
IIA	82
IIB	65
IIIA	47
IIIB	44
IV	14

Source: Modified from data of the National Cancer Institute: Surveillance, Epidemiology, and End Results (SEER).

Adjuvant systemic therapy may not be needed at all for patients with very small (<1 cm) tumors and negative lymph nodes. However, there is no patient with invasive breast cancer who does not have some risk of subsequent distant metastases, and therefore who might not benefit at all. This consideration raises two issues: (1) the differences in odds of benefit and odds of toxicities of the various types of therapies and (2) the judgment between the patient and her caregiver regarding the calculated absolute benefit-risk ratio for specific types of adjuvant systemic treatments.

There are three types of adjuvant systemic therapies: (1) chemotherapy; (2) endocrine; and (3) anti-HER2 therapies. The decision whether to apply each of these depends on prognostic and predictive features as well as the combined judgment of the patient and caregiver. For example, a patient might be much more likely to accept endocrine therapy for a very small potential benefit than she would accept chemotherapy for the same

calculated advantage, since the former is much less often associated with either life-taking, life-threatening, or permanently life-changing toxicities than the latter. Thus, one has to consider prognostic and predictive factors for each type of therapy, separately.

The greatest controversy concerns the recommendation for adjuvant *chemotherapy*, since there is no good predictive factor for this class of treatments, and the decision must be made on prognosis alone. Large overview analyses suggest that chemotherapy reduces the risk of recurrence over the 10 years subsequent to primary diagnosis by approximately one-third. For patients with positive lymph nodes and/or features that render the cancer T4, the risk of distant recurrence (and thus not being cured) over that decade is 50% or higher. Therefore, a one-third reduction of at least 50% means that 15–20%, or more, women will be cured who would not have been in the absence of adjuvant chemotherapy. The life-taking, life-threatening, or permanently life-changing toxicities of adjuvant chemotherapy are ~1–2%, and therefore almost all medical oncologists would recommend adjuvant chemotherapy in this setting.

In contrast, there is rarely justification for adjuvant chemotherapy in most women with tumors <1 cm in size whose axillary lymph nodes are negative. However, this decision is very much weighed by the expression of ER and HER2. For example, the risk of recurrence of such a patient whose tumor is negative for ER, PgR, and HER2 (so-called triple-negative breast cancer) over the succeeding 10 years without any adjuvant is ~15%. If chemotherapy reduces this risk by approximately one-third or more, which is what large overview analyses suggest, then 5%, or perhaps even higher, of patients will be cured who would otherwise be destined to die of their disease. Likewise, a patient with ER and PgR-negative, but HER2-positive, disease has a slightly worse prognosis (risk of recurrence over 10 years is ~20%), and will benefit not only from the adjuvant chemotherapy but from anti-HER2 therapy as well, so that her potential absolute benefit is even higher. Many, but not all, clinicians would recommend adjuvant chemotherapy for such patients.

On the other hand, patients with ER-positive disease have a better prognosis than those with ER-negative breast cancer, and adjuvant endocrine therapy will further reduce the odds of recurrence by approximately one-half. Therefore, the same patient in the example above (<1 cm, node negative) but who has an ER-positive and HER2-negative cancer has a lower initial risk of recurrence (~10% over 10 years). Given the relatively low life-taking, life-threatening, or permanently life-changing toxicities, she is very likely to accept adjuvant endocrine therapy, further lowering her estimated risk of recurrence to ~5%. If chemotherapy reduces this risk by approximately one-third, no more than 1–2% of patients will benefit. This potential benefit is approximately the same as the number of patients who will suffer life-taking, life-threatening, or permanently life-changing toxicities. Thus, in this case, most clinicians would recommend adjuvant endocrine, but not chemotherapy.

These examples represent extremes. In the screening era, up to 30% of newly diagnosed patients have T2-3, node-negative, ER-positive cancers. These patients have an intermediate risk between the two extremes, and the calculated absolute benefit of adjuvant chemotherapy is ~3–5%. It is unclear if this small but real benefit is sufficient to justify adjuvant chemotherapy. Detection of breast cancer cells either in the circulation or bone marrow is associated with an increased relapse rate. However, the finding of bone marrow micrometastases

only portends a slightly worse prognosis, especially in node negative patients, and bone marrow biopsies are not recommended in patients with early stage disease.

The most exciting development in this area is the use of gene expression arrays to analyze patterns of tumor gene expression, especially for node-negative, ER-positive cancers. Several groups have independently defined gene sets that reliably predict disease-free and overall survival far more accurately than any single prognostic variable. The Oncotype DX® Recurrence Score (RS) analysis of 21 genes was the first such assay to be adopted. A number of retrospective and more recently prospective studies have documented its utility in identifying patients with node-negative, ER-positive breast cancer whose prognosis, assuming adequate adjuvant endocrine therapy, is so good that they can forego adjuvant chemotherapy. Basically, the 30–50% of patients with ER positive, node negative, but low RS, appear to have luminal A breast cancers, and they do not need chemotherapy, whereas those with high RS appear to have luminal B cancers and the benefits of adjuvant chemotherapy clearly outweigh the risks. For those with intermediate RS, the answer is still unclear and has been the focus of now completed, but as yet unreported, prospective trials.

More recently, other assays, including the Prosigna,® EndoPredict,® and Breast Cancer Index® have also been shown to have clinical utility in this setting. Only one of these tests should be ordered for a single patient, since they do not always give the same results and there are no data to determine which, in the case of discordance, might be "correct." Also, the use of such standardized risk assessment tools such as Adjuvant! Online (*www.adjuvantonline.com*) is very helpful. These tools are highly recommended in otherwise ambiguous circumstances.

Several *measures of tumor growth rate* correlate with early relapse, but their use is problematic due to analytical variability. Of these, assessment using immunochemical assays for the proliferation marker, Ki67, is the most widespread. However, there is substantial lab-to-lab variability and disagreement regarding optimal cut points. At present, in standard practice outside of a highly skilled laboratory, use of Ki67 is not recommended to make clinical decisions.

Molecular changes in the tumor are also useful. Tumors that overexpress *erbB2* (HER2/neu) have a worse prognosis, but expression of this gene for prognosis is most important in patients with ER-positive, node-negative disease. Indeed, patients with HER2-positive breast cancer are so likely to have a high RS that it is not recommended that the Oncotype DX,® or for that matter any of the other multiparameter assays, be ordered. HER2 should be performed on every breast cancer biopsy, however, because of its predictive role for anti-HER2 therapies.

Predictive Factors to Choose Adjuvant Systemic Therapy

The decision to recommend AST is also based on predictive factors; those that provide a prediction of the likelihood that a given class, or even specific drug within a class, will have activity or not. The two important predictive factors, which should be ordered in all breast cancer biopsies (primary or metastatic), are ER and HER2.

There is no detectable benefit in patients with ER-poor, or -negative, cancers, whereas adjuvant endocrine therapy reduces the risk of recurrence by one-half or more in patients with ER-rich cancers. ER is most commonly measured by counting the percent of positive cells within the cancer after immunohistochemical (IHC) staining. Endocrine therapy is recommended for any patient with ≥10% positive cells, whereas it is not for those whose cancers only have 0–1% staining. The evidence supporting benefit in 1–9% cases is weak, but given the potential benefit and relatively low toxicities of endocrine therapy, it is recommended for patients in this circumstance, with a low threshold for discontinuation if side effects are intolerable.

HER2 is the target for anti-HER2 therapies. Adjuvant trastuzumab therapy reduces the risk of distant recurrence by one-third or more, with associated substantial risk of dying of breast cancer. Most, if not all, of the large adjuvant trastuzumab trials have been performed in patients with HER2-"positive" breast cancer. HER2 status is determined using either IHC staining for protein overexpression, or fluorescent in situ hybridization (FISH) for gene amplification. IHC staining of 3+ (on a scale of 0–3+) is considered positive, whereas 0–1+ is considered negative. For cases with 2+ staining, reflex FISH analysis is recommended. FISH can either be used as the initial evaluation, or for additional evaluation in IHC 2+ cases. FISH results are considered positive if the ratio of HER2 to centromere signal on chromosome 17 is ≥2.0. There is no reason to do FISH if IHC is 3+ or 0–1+, nor is there reason to order IHC testing if FISH is ≥2.0. If note, preclinical studies and retrospective analyses of a few selected cases from the prospective randomized trials have suggested that perhaps trastuzumab might be effective in cases with IHC 1–2+ results. A large prospective randomized clinical trial addressing this issue is completed but not yet reported.

There are no reliable predictive factors for chemotherapy, in general or for specific types of chemotherapies. It has been hypothesized that chemotherapy may be more active in ER-negative and/or HER2-positive cancers. More recently, this issue has evolved to imply that luminal B cancers may be more chemosensitive, whereas luminal A cancers are perceived to be relatively chemoresistant. At present, none of the tests for intrinsic subtype should be used to determine whether to give chemotherapy or not, based on *prediction* of resistance in patients with poor *prognosis*, such as those with T4 or node-positive disease. Attempts to identify reliable predictive factors for individual classes of chemotherapeutic agents (such as anthracyclines, alkylating agents, or taxanes) have been unsuccessful. The platin salts (carbo-, cisplatin) may have higher activity in patients with triple-negative breast cancer and perhaps in patients with HER2-positive disease.

Adjuvant Regimens

If chemotherapy is indicated, it should include multiple agents, either in combination or as sequential single agents. If indicated, anti-HER2 therapy should include at least 1 year of trastuzumab, and preliminary data have supported addition of pertuzumab for at least three months. Endocrine therapy should be administered to patients with ER-positive breast cancer following completion of chemotherapy and administered for at least 5 years, and probably longer.

Endocrine Therapy

There are two proven endocrine therapy strategies: the SERM, tamoxifen, or estrogen ablation. In addition to being effective in preventing new cancers and reducing the risk of local-regional recurrences in patients with DCIS, tamoxifen reduces the risk of distant recurrence and death due to invasive breast cancer by ~40% over the decade following diagnosis. It is equally effective in pre- and postmenopausal women, although it may be slightly less effective in very young (<40 years) patients. Because tamoxifen is a SERM, it has mixed ER antagonism (in the breast and brain) and agonism (in the bone, liver, and uterus). Therefore, it is active against breast cancer in the prevention, adjuvant, and metastatic settings, but frequently causes hot flashes. The agonistic effect results in reduction of osteopenia/osteoporosis, especially in postmenopausal women, but it increases thrombosis and endometrial cancers due to this effect in the liver and uterus, respectively.

Estrogen depletion can be achieved surgically in premenopausal women by oophorectomy or ovarian suppression with a gonadotropin-releasing hormone super-agonist (GnRH agonist), such as goserelin, that results in tachyphylaxis of the pituitary. However, women with nonfunctioning ovaries, whether induced or by natural menopause, still produce small amounts of estrogen. Estrogen production in these women occurs by adrenal synthesis of estrogen precursors (testosterone, dehydroepiandrosterone [DHEA]) that are converted to estradiol and estrone by aromatase activity in peripheral fat and possible cancer cells. In postmenopausal women, circulating estrogen can be reduced to nearly imperceptible levels with the use of oral AIs. There are three such agents available (anastrozole, letrozole, and exemestane). Although there is no perceptible difference in activity or toxicity among the three AIs, they are all slightly more effective than tamoxifen.

It is recommended that all postmenopausal women with ER-positive breast cancer be treated for at least 3–5 years with an AI, unless there is a contraindication. The most common concern is the presence of severe osteoporosis, since this is the most frequent life-taking or life-threatening toxicity of the AIs. Likewise, ~15–20% of patients cannot tolerate the AIs due to musculoskeletal symptoms mimicking osteoarthritis and arthralgias. For both these groups of women, tamoxifen is a reasonable therapy, again assuming no contraindications exist. The most important of these is a past history of thrombosis, or high risk of cerebrovascular disease.

For premenopausal women, the decision of optimal endocrine therapy depends on prognosis and patient choice. Complete estrogen depletion is slightly more effective than tamoxifen alone, but it may also be associated with more bothersome side effects, such as hot flashes, vaginal dryness, and sexual dysfunction. Recent studies have suggested that complete estrogen depletion, consisting of either oophorectomy or chemical suppression of gonadotropins coupled with an AI, is indicated for women with worse prognosis, in particular node positivity. For those with more favorable prognosis, tamoxifen alone may be preferable. The AIs should not be administered to women with functioning, or dormant, ovaries, since the negative hypothalamic-pituitary feedback can result in a rebound hyperestrogenic production effect.

The duration of adjuvant endocrine treatment is unclear. Until recently, the standard recommendation was at least 5 years of therapy. Several studies have now demonstrated that although 5 years of adjuvant endocrine treatment clearly reduces the risk of recurrence during that time and for a few years after discontinuation, the annual risk of distant recurrence during the subsequent 15 years is 0.5–3%, depending on the initial T and N status. Further, so-called extended adjuvant endocrine therapy with either tamoxifen or an AI, for at least

years 6–10, continues to reduce this late risk of relapse. The decision of whether to continue adjuvant endocrine therapy or not after 5 years must therefore take into consideration initial risk (T, N, grade), current side effects and potential cumulative toxicities, and the patient's perception of the relative and absolute benefits and risks.

Chemotherapy

If adjuvant chemotherapy is indicated, as discussed above, one must consider the optimal regimen. Several studies, and a combined overview analysis, have demonstrated that multiple-agent chemotherapy is more effective than single agent. However, at least two studies have shown that sequential single-agent chemotherapy is as effective, and may be slightly less toxic, than simultaneous combination chemotherapy although it requires longer total duration to deliver. Administration of four to six cycles of chemotherapy appears to be optimal; one cycle is less effective than six, but more than six have generally increased toxicity without further efficacy.

Several chemotherapeutic agents have activity in the adjuvant setting. These include alkylating agents, (principally cyclophosphamide), anthracyclines (doxorubicin, epirubicin), antimetabolites (5-fluoruracil [5FU], capecitabine, methotrexate), and the taxanes (paclitaxel, docetaxel). Within classes, randomized trials have failed to demonstrate superiority of one agent versus another (e.g., doxorubicin vs epirubicin, or paclitaxel vs docetaxel). Dose escalation above an optimal dose is not more effective. The advantage of more frequent scheduling for most individual agents is unclear, but weekly or every other week paclitaxel is superior to every 3-week infusion, while, enigmatically, the opposite is true for its cousin, docetaxel. However, one benefit of a "dose dense" regimen (e.g., every 2 weeks with cytokine support vs every 3 weeks) is earlier completion of therapy.

These agents are usually combined within a single regimen. The oldest of these is cyclophosphamide, methotrexate, and 5FU (CMF) Addition of an anthracycline, or substitution of an anthracycline for the antimetabolite, improves outcomes slightly, albeit with slightly increased risk of heart failure and secondary leukemia. Addition of a taxane to an anthracycline-based regimen further reduces the chances of distant recurrence and death, albeit only modestly. Recent studies have suggested that addition of an anthracycline to a taxane-based regimen is also modestly more effective than a taxane plus cyclophosphamide alone.

Which regimen is appropriate for a patient must be individualized based on prognosis, comorbid conditions, and the perspective of the patient. For example, the modest relative improvement of giving an anthracycline, cyclophosphamide, and a taxane (AC-T) may not transfer to a sufficiently large absolute improvement in survival in a patient with a relative small (T2) tumor and negative nodes, whereas that same relative reduction in death may translate to a sufficiently large absolute benefit in a patient with a worse prognosis. Therefore, the former patient might best be served with a taxane/cyclophosphamide (TC) regimen alone, while the latter might wish to accept the added risk of congestive heart failure and leukemia associated with the anthracyclines.

Neoadjuvant treatment involves the administration of adjuvant systemic therapy, most commonly chemotherapy, before definitive surgery and radiation therapy. The objective partial and complete response rates of patients with breast cancer to neoadjuvant chemotherapy exceed 75%. Thus, many patients will be "downstaged" by neoadjuvant chemotherapy. In this circumstance, patients with locally advanced, inoperable cancers may become candidates for surgery, and a small fraction of patients who are not considered eligible for breast-conserving surgery may become so due to shrinkage of their cancer. However, overall survival has not been improved using this approach as compared with the same drugs given postoperatively.

Patients who achieve a pathologic complete remission after neoadjuvant chemotherapy have a substantially improved survival compared to those who do not. It is unknown if this observation implies that the latter group did not benefit, or just had a worse initial prognosis, yet still gained some benefit. Although it is appealing to consider treating patients who have not had a pathologic complete response with even more chemotherapy, no studies have demonstrated that doing so improves overall survival. It is possible that these patients have chemoresistant disease, and therefore more chemotherapy will not be of value. However, it is essential that all patients, regardless of response to neoadjuvant chemotherapy, receive adjuvant endocrine therapy if they have an ER-positive breast cancer and adjuvant anti-HER2 therapy if their cancer is HER2 positive.

The neoadjuvant setting also provides an appealing opportunity for the evaluation of new agents. For example, a second HER2-targeting antibody, pertuzumab, has been shown to provide increased rates of pathologic complete response when combined with trastuzumab in the neoadjuvant setting. However, this approach is controversial; it is not clear that demonstration of higher response rates in the neoadjuvant setting will translate into better overall survival. For example, neoadjuvant trials demonstrated that combination trastuzumab and lapatinib resulted in higher pathologic complete responses than trastuzumab alone, yet a classically performed adjuvant trial failed to demonstrate improved survival for this regimen.

Chemotherapy is associated with nausea, vomiting, and alopecia in ~100% of patients, although the former two are well controlled with modern antiemetics. More importantly, chemotherapy causes neutropenia and fever, with a risk of infection of ~1%. The neutropenia can be prevented in most patients with appropriate use of the growth factor filgrastim. Secondary myelodysplasia and leukemia occur in ~0.5–1% of patients treated with anthracyclines as well as with high cumulative doses of cyclophosphamide, usually occurring within 2–5 years of treatment. The anthracyclines cause cumulative dose-related congestive heart failure, which occurs in ~1% of patients treated with standard four to five cycles at 60 mg/m^2. Peripheral neuropathy is the major dose-limiting and life-changing toxicity of the taxanes. Neuropathy occurs during treatment in ~15–20% of patients, and permanent, chronic neuropathy persists in 3–5%.

Anti-HER2 Therapy

The emergence of therapies directed toward HER2 has been one of the great success stories of all oncology. Several trials have demonstrated that the humanized monoclonal antibody, trastuzumab, decreases both risk of recurrence and mortality in early-stage breast cancer. While trastuzumab administered after chemotherapy

is effective, the accumulated evidence suggests that it is optimally delivered concurrently with chemotherapy, particularly in association with a taxane. However, concurrent treatment with an anthracycline is generally avoided, since the main toxicity of trastuzumab is cardiac dysfunction, which appears more often when the agent is delivered simultaneously with doxorubicin. Therefore, if an anthracycline is to be used, it is most commonly given prior to administration of trastuzumab—for example as AC for four cycles followed by a taxane plus trastuzumab. In patients with reasonably favorable prognosis (T1 or 2, node negative), single-agent paclitaxel plus trastuzumab appears to be an adequate regimen.

Twelve months of trastuzumab therapy is optimal. Randomized trials have demonstrated no additional benefit beyond 12 months, whereas 6 months has been shown to be inferior to 12. Trastuzumab is administered intravenously weekly or every 3 weeks.

Other, anti-HER2 treatments that are effective in the metastatic setting are appealing candidates for adjuvant therapies. As noted, neoadjuvant studies have demonstrated that chemotherapy with the combination of trastuzumab and pertuzumab results in higher pathologic complete responses than trastuzumab alone. The U.S. Food and Drug Administration (FDA) has granted this combination with accelerated approval, but final approval for the combination is pending more clinically meaningful results (disease-free, overall survival) from now-completed, classic adjuvant trials. Although lapatinib did not add to trastuzumab therapy and single-agent adjuvant lapatinib is inferior to single agent trastuzumab, another anti-HER2 tyrosine kinase inhibitor, neratinib, is superior to no anti-HER2 therapy. Neratinib has not been compared to trastuzumab, either as a single agent or in combination. Ado-trastuzumab emtansine, an antibody-drug conjugate, has activity in the metastatic setting even in patients who have progressed on trastuzumab and is now being tested in the adjuvant setting.

Skeletal Strengthening Agents

Bone-strengthening agents that are commonly used to treat osteoporosis appear to have some, but limited, activity in preventing recurrent breast cancer, particularly in postmenopausal women. In an overview analysis of all trials addressing bisphosphonate therapy, improvement in overall survival was not significantly associated with any specific bisphosphonate class, treatment schedule, ER status, nodal status, tumor grade, or concomitant chemotherapy. No differences were seen in nonbreast cancer mortality. Bone fractures were reduced (relative risk [RR] 0.85, 95% confidence interval [CI] 0.75–0.97; 2 $p = 0.02$). At present, there is no clear consensus regarding routine use of bisphosphonates as an adjuvant therapy, although patients with advancing osteopenia or confirmed osteoporosis should be treated accordingly.

Novel Adjuvant Systemic Agents

Other exciting adjuvant strategies are being tested, such as poly–ADP ribose polymerase (PARP) inhibitors in patients with known germline *BRCA1* or *BRCA2* mutations or those with triple-negative cancers that share similar defects in DNA repair in their etiology. The remarkable results of immune checkpoint inhibitors in other cancers have led to studies of this approach in both metastatic and post-neoadjuvant chemotherapy settings but are still considered highly investigational.

Recommendations for adjuvant therapy are found in **Table 75-2**.

TABLE 75-2

Suggested Approaches to Adjuvant Systemic Therapy[a]

Nodal Status	Tumor Size	ER	HER2	Multi-parameter assay	Menstrual Status	Chemotherapy	Endocrine Therapy	anti-HER2 therapy
Positive	Any	Neg	Neg	Not indicated	Any	Multidrug	None	None
		Pos			Prem		Ovarian ablation + AI	
					Post		AI	
		Any	Pos		Any			Trastuzumab X12 mos; pertuzumab X12 weeks
Negative	<1 cm	Neg	Neg	Not indicated	Any	Consider multidrug	None	None
	≥1 cm	Neg	Neg		Any	Multidrug	None	None
	1–5 cm	Pos	Neg	Low RS	Prem	None	Tam	None
					Post		AI	None
				Intermed	Any	Consider multidrug	Tam (pre) or AI (post)	None
				Hi	Any	Multidrug	As for node pos	None
		Any	Pos	Not indicated	Any	Single-agent paclitaxel	As for node pos	Trastuzumab X12 mos

[a]Meant for guidance only. Each patient should be considered independently based on tumor and comorbidity status.

STAGE III BREAST CANCER

Between 10 and 25% of patients present with so-called locally advanced, or stage III, breast cancer at diagnosis. Many of these cancers are technically operable, whereas others, particularly cancers with chest wall involvement, inflammatory breast cancers, or cancers with large matted axillary lymph nodes, cannot be managed with surgery initially. As noted, neoadjuvant compared may be no more effective than postsurgical adjuvant chemotherapy in prolonging survival, but the advantages of downstaging and therefore facilitating local therapy are accepted. Radiotherapy either to the chest wall after mastectomy or to the breast after tumor excision is almost always recommended, as is regional lymph node treatment. Adjuvant anti-HER2 and endocrine therapies are also used, as appropriate. These patients should be managed in multimodality clinics to coordinate surgery, radiation therapy, and systemic chemo-, endocrine, and anti-HER2 therapies, as indicated. Such approaches produce long-term disease-free survival in ~30–50% of patients.

BREAST CANCER SURVIVORSHIP ISSUES

The odds of surviving breast cancer have increased dramatically over the last 35 years due to a combination of early detection and more effective therapies. Although detection bias improves case fatality rates, age-adjusted mortality rates (mortality/100,000 women in society/year) have declined by >30%. Therefore, while ~40,000 American women will die of metastatic breast cancer in 2016, >60,000 would have suffered breast cancer mortality without these advances. Thus, all clinicians, not just oncologists, need to be aware of survivorship issues in patients with previously diagnosed and treated breast cancer.

No special follow-up procedures, such as serial circulating tumor biomarkers or systemic radiographic/scintigraphic imaging, are indicated in an asymptomatic patient with no physical findings of recurrence. Although randomized trials have demonstrated slightly higher incidence of detection of metastases with lead times of 3–12 months by screening asymptomatic patients compared to no special follow-up, there is no evidence of improved overall survival. If anything, one of these studies suggested a worse quality of life due to higher anxiety levels associated with the testing, and toxicities associated with earlier treatment in patients who were otherwise doing well at that time. These recommendations are summarized in Table 75-3.

TABLE 75-3

Surveillance Guidelines for Breast Cancer Patients after Primary and Adjuvant Therapy during Routine Follow-up

Test	Frequency
Recommended	
History; eliciting symptoms; physical examination	q3–6 months × 3 years; q6–12 months × 2 years; then annually
Breast self-examination	Monthly
Mammography	Annually
Pelvic examination	Annually as per age-appropriate guidelines (particularly for patients on SERMs)
Patient education about symptoms of recurrence	Ongoing
Coordination of care Assessment of side effects if on endocrine therapy Echocardiography if on trastuzumab	Ongoing Ongoing Every 3 months; discontinue when trastuzumab therapy complete
Not Recommended (if asymptomatic)	
Complete blood count	
Serum chemistry studies	
Chest radiographs	
Bone scans	
Ultrasound examination of the liver	
Computed tomography of chest, abdomen, or pelvis	

Test	Frequency
Tumor markers CA 15-3, CA 27-29, CEA, CTC	
Transvaginal endometrial ultrasonography	

Abbreviations: CA, cancer antigen; CEA, carcinoembryonic antigen; CTC, circulating tumor cell; SERM, selective estrogen receptor modulator.

Source: Recommended Breast Cancer Surveillance Guidelines, ASCO Education Book, Fall, 1997. From J Clin Oncol 15:2149, 1997; with permission.

It is important to carefully assess and evaluate new symptoms, considering whether they might be due to the cancer, the treatment, or an unassociated condition. Judgment needs to be used to decide if blood tests or imaging are required, in order to avoid missing a lesion for which appropriate treatment would improve the patient's quality of life but to diminish overtesting, with associated inconvenience, anxieties, false positives, and cost. Serial echocardiography should be performed every 3 months for patients on adjuvant trastuzumab, but not after it is discontinued.

Likewise, there is no role for serial monitoring for long-term, life-threatening toxicities associated with chemotherapy, such as myelodysplastic syndromes or congestive heart failure, since these are quite uncommon and likely to cause obvious symptoms requiring proper evaluation if they occur.

For patients on endocrine therapy, quality-of-life issues may be critical, including hot flashes, sexual difficulties, musculoskeletal complaints, and risk of osteoporosis. Although estrogen therapy, given orally, transdermally, or transvaginally, effectively reduces these side effects, it should not be given to these patients, since it may counteract the efficacy of the endocrine therapy. Nonhormonal treatments, such as selected antidepressants for hot flashes and musculo-skeletal symptoms, and counseling and water-based lubricants for sexual issues can be quite helpful. It is important to screen bone density in patients on an AI more frequently than is recommended for the average postmenopausal woman, since total estrogen depletion results in enhanced risk of osteoporosis and fracture. All women should be counseled to take daily calcium and vitamin D replacement, and if osteoporosis is present or osteopenia is worsening, bone strengthening agents should be administered.

THERAPY OF METASTATIC DISEASE

About 15–20% of patients treated for localized breast cancer develop metastatic disease in the subsequent decade after diagnosis. Soft tissue, bony, and visceral (lung and liver) metastases all account for approximately one-third of sites of initial relapses. However, by the time of death, most patients will have bony involvement. Recurrences can appear at any time after primary therapy, but at least half occur >5 years

after initial therapy. This observation is particularly true in patients with ER-positive disease, for whom the risk of distant recurrence remains constant for as long as 20 years and is the basis for recommendation of extended adjuvant endocrine therapy. It is now clear that a variety of host factors can influence recurrence rates, including depression and central obesity, and these diseases should be managed as aggressively as possible.

For patients with no prior history of metastases, a biopsy of suspicious physical or radiographic lesions should be performed, both for confirmation that the lesion does, indeed, represent recurrent cancer and to reevaluate ER and HER2, which can differ between the primary and metastatic lesions in up to 15% of cases. One should not assume that an apparent abnormality is a breast cancer metastasis. Many benign conditions, such as tuberculosis, gallstones, sarcoidosis, or other nonmalignant diseases, can mimic a recurrent breast cancer and are of course treated much differently.

Although treatable, metastatic disease is rarely if ever cured. The median survival for all patients diagnosed with metastatic breast cancer is <3 years, but with remarkable variability depending on intrinsic subtype and effective treatments. Patients with triple-negative metastatic breast cancer have the shortest expected survival, while those with ER-positive disease can expect to live the longest. HER2 positivity was initially found to be a very poor prognostic factor in metastatic breast cancer, but the availability of several effective treatments has improved the expected survival rates to at least those of ER-positive patients, if not better.

In the absence of cure, the overall goal of treatment of metastatic disease is palliation, or, put simply, to “keep the patient feeling as well as she can for as long as she can.” A secondary goal is improved survival. It is important to point out that survival has not been improved by advocating more aggressive, or toxic, therapies, such as high-dose or combination chemotherapy, but rather by more selective and biologically based therapy, such as use of endocrine or anti-HER2 therapies in patients with ER- or HER2-positive breast cancers, respectively. Generally, a new treatment is continued until either progression or unacceptable toxicities are evident. These are both evaluated by serial history and physical examinations and periodic serologic evaluation for hematologic or hepatic abnormalities, as well as circulating tumor biomarker tests (assays for MUC1, such as CA15-3 or CA27.29, and for carcinoembryonic antigen or occasionally CA125). If all these evaluations fail to suggest progression, it is unlikely that imaging will contribute. However, if one or more of these suggest progression, whole-body imaging with either a PET/CT or a scintigraphic bone scan and dedicated CT are indicated. Brain imaging is not recommended unless the patient has some sort of central nervous system (CNS) symptom or finding.

The choice of therapy requires consideration of local therapy needs, specifically surgical approaches to particularly worrisome long-bone lytic lesions or isolated CNS metastases. New back pain in patients with breast cancer should be explored aggressively on an emergent basis; to wait for neurologic symptoms is a potentially catastrophic error. Metastatic involvement of endocrine organs can occasionally cause profound dysfunction, including adrenal insufficiency and hypopituitarism. Similarly, obstruction of the biliary tree or other impaired organ function may be better managed with a local therapy than with a systemic approach. Radiation as an adjunct to or instead of surgery is an important consideration for particularly symptomatic

disease in long or vertebral bones, local-regional recurrences, and CNS metastases. In many cases, systemic therapy can be withheld while the patient is managed with appropriate local therapy.

There is no evidence that aggressive local treatment, such as excision; radiation; radiofrequency ablation; or cryotherapy of metastases to the lung, liver, or other distant sites, improves survival. Although appealing, these strategies are associated with increased toxicity and cost and should be reserved for palliation.

Selection of the systemic therapy strategy depends on the overall medical condition of the patient, the hormone receptor and HER2 status of the tumor, and clinical judgment. Because therapy of systemic disease is palliative, the potential toxicities of therapies should be balanced against expected response rates. Several variables influence the response to systemic therapy. For example, the presence of ER and PgR is a strong indication for endocrine therapy, even for patients with limited visceral (lung/liver) disease. On the other hand, patients with short disease-free intervals or rapidly progressive visceral disease (liver and lung) with end-organ dysfunction, such as lymphangitic pulmonary disease, are unlikely to respond to endocrine therapy.

Many patients with bone-only or bone-dominant disease have a relatively indolent course. Because the goal of therapy is to maintain well-being for as long as possible, emphasis should be placed on avoiding the most hazardous complications of metastatic disease, including pathologic fracture of the axial skeleton and spinal cord compression. Under such circumstances, systemic chemotherapy has a modest effect, whereas radiation therapy may be effective for long periods. Other systemic treatments, such as strontium-89, may provide a palliative benefit without inducing objective responses. Patients with bone involvement should receive concurrent bone strengthening agents, such as bisphosphonates or the humanized monoclonal anti-RANK ligand antibody, denosumab.

Many patients are inappropriately treated with toxic regimens into their last days of life. Often, oncologists are unwilling to have the difficult conversations that are required with patients nearing the end of life, and not uncommonly, patients and families can pressure physicians into treatments with very little survival value. Palliative care consultation and realistic assessment of treatment expectations need to be reviewed with patients and families. We urge consideration of palliative care consultations for patients who have received at least two lines of therapy for metastatic disease.

Endocrine Therapy

ER-positive breast cancer will respond to endocrine therapy ~30–70% of the time. Potential endocrine therapies are summarized in **Table 75-4**. As in the adjuvant setting, one can choose among the SERM, tamoxifen, the AIs (anastrozole, letrozole, exemestane), or other strategies. Among the latter, the selective estrogen receptor downregulator (SERD), fulvestrant, has substantial activity. Early clinical studies with this drug were unexciting, but more recent studies have proven a very steep dose-response curve, and at higher levels (500 mg/month), it is as or more active than either tamoxifen or the AIs. Additive endocrine therapies, including treatment with progestins, and androgens, and enigmatically, pharmacologic doses of estrogens, are all active, but they may be associated with unacceptable side effects in many women. The mechanism of

action of these latter therapies is unknown. Cases in which tumors shrink in response to tamoxifen withdrawal (as well as withdrawal of pharmacologic doses of estrogens) have been reported, but with the advent of so many other therapies for metastatic disease, this strategy is rarely used in modern oncology.

TABLE 75-4

Endocrine Therapies for Breast Cancer

Therapy	Comments
Castration	For premenopausal women
Surgical	
LHRH agonists	
Antiestrogens	
Tamoxifen	Useful in pre- and postmenopausal women[a]
Fulvestrant	Responses in tamoxifen-resistant and aromatase inhibitor–resistant patients[a]
Aromatase inhibitors	Low toxicity; now first choice for metastatic disease[a]
High-dose progestogens	Common fourth-line choice after aromatase inhibitors, tamoxifen, and fulvestrant
Additive androgens or estrogens	Plausible fourth-line therapies; potentially toxic

[a]Consider retreatment with everolimus in combination for disease progression.

Abbreviation: LHRH, luteinizing hormone–releasing hormone.

The sequence of endocrine therapy is variable. Patients who respond to one endocrine therapy have at least a 50% chance of responding to a second endocrine therapy. It is not uncommon for patients to respond to two or three sequential endocrine therapies. In most postmenopausal patients, the initial endocrine therapy should be an AI rather than tamoxifen. As noted, AIs are not used in premenopausal women because their hypothalamus can respond to estrogen deprivation by producing gonadotropins that promote estrogen synthesis. Tamoxifen and fulvestrant are usually used in sequence after AI therapy. Combination endocrine therapies increase the chances of response initially, but they do not appear to increase the ultimate time to

chemotherapy use or overall survival. Combinations of chemotherapy with endocrine therapy are not useful, as summarized in Table 75-4.

At least two different targeted agents have been shown to enhance outcomes of patients with ER-positive metastatic breast cancer when combined with endocrine therapy. Addition of an inhibitor of the mammalian target of rapamycin (mTOR), everolimus, to the hormonal treatment can lead to AIs, tamoxifen, or fulvestrant improves time to progression, and this agent is now being explored as front-line therapy and in the adjuvant setting. Likewise, inhibitors of cyclin D kinase 4/6 (CDK4/6) (palbociclib, ribociclib, abemaciclib) have also been shown to substantially improve progression-free survival when combined either with an AI or fulvestrant. These agents are also being tested in the adjuvant setting. Data regarding overall survival benefits from the mTOR or CDK4/6 inhibitors are still pending, but addition of one or the other in combination with ET for women with ER-positive metastatic breast cancer is becoming the standard of care. These should not be given simultaneously but rather in sequence as appropriate, as summarized in **Table 75-5**.

TABLE 75-5

Common Agents Added to Endocrine Therapies for Metastatic Breast Cancer

Class	How administered	Agents	Common Toxicities
Anti-mTOR	Oral	Everolimus	Mucositis, diarrhea, rash
CDK4/6 inhibitors	Oral	Palbociclib, ribociclib, abemaciclib (not FDA approved)	Neutropenia; uncommon leukopenia, fatigue, and nausea

Abbreviations: CDK4/6, cyclin D kinase 4/6; FDA, Food and Drug Administration; mTOR, mammalian target of rapamycin.

Chemotherapy

Unlike many other epithelial malignancies, breast cancer responds to multiple chemotherapeutic agents, including anthracyclines, alkylating agents, taxanes, and antimetabolites. Multiple combinations of these agents have been found to improve response rates somewhat, but they have had little effect on duration of response or survival. Unless patients have rapidly progressive visceral (lung, liver) metastases with end-organ dysfunction, single-agent chemotherapy, used in sequence as one drug fails going on the next, is preferable. Given the significant toxicity of most drugs, the use of a single effective agent will minimize toxicity by sparing the patient exposure to drugs that would be of little value. No method to select the drugs most efficacious for a given patient has been demonstrated to be useful.

Most oncologists use either capecitabine or an anthracycline or a taxane for first-line chemotherapy, either in a patient with ER-positive disease that is refractory to endocrine therapy or for a patient with ER-negative

breast cancer. Within these general classes, it is not clear that one particular agent (such as doxorubicin vs epirubicin or paclitaxel vs docetaxel) is preferable, and the choice has to be balanced with individual needs. Objective responses in previously treated patients may also be seen with gemcitabine, vinorelbine, and oral etoposide, as well as a new class of agents, epothilones. Platinum-based agents have become far more widely used in both the adjuvant and advanced disease settings for some breast cancers, particularly those of the "triple-negative" subtype.

Anti-HER2 therapy

Treatment of patients with anti-HER2 metastatic breast cancer is one of the great success stories in the last 30 years of oncology. Initial use of a trastuzumab, either alone or with chemotherapy, was shown to improve response rate and survival for women with HER2-positive disease. Indeed, anecdotal reports of a few patients with remarkably sustained complete responses suggest that, on occasion, a few may be cured. Chronologically, the tyrosine kinase, lapatinib, was subsequently shown to be effective when added to chemotherapy after patients progressed on prior trastuzumab. Further, both continuation of trastuzumab after progression, in combination with the next chemotherapeutic regimen and combination of trastuzumab and lapatinib in patients who had progressed on trastuzumab are both superior to discontinuing the trastuzumab.

For patients who have become refractory to trastuzumab–based therapy, and more recently even in the upfront setting, other therapies have remarkably high activity. A novel antibody drug conjugate (ADC) that links trastuzumab to a cytotoxic agent, ado-trastuzumab emtansine, is active even in patients who have progressed on trastuzumab. More recently, the combination of chemotherapy and trastuzumab and pertuzumab has been shown to result in prolonged overall survival compared to trastuzumab alone. These recommendations are summarized in **Table 75-6**.

TABLE 75-6

Common Anti-HER2 Agents for Breast Cancer

Class	How Administered	Agents	Common Toxicities
Humanized monoclonal antibodies	IV	Trastuzumab, pertuzumab	Cardiac dysfunction GI (diarrhea)
Tyrosine Kinase Inhibitors	Oral	Lapatinib neratinib (not FDA approved)	Diarrhea, mucositis, rash
Antibody-drug conjugate	IV	Ado-trastuzumab emtansine	Peripheral neuropathy, thrombocytopenia

Other Therapies

Bevacizumab is an agent that targets the vascular endothelial growth factor (VEGF). Bevacizumab with paclitaxel or other chemotherapeutic agents modestly increases the response rate and response duration to paclitaxel, but without improvement in overall survival and with occasional major toxicities. After initial excitement and FDA approval, its use has been mostly abandoned in breast cancer. As in the metastatic setting, trials are ongoing testing the value of PARP (poly-ADP ribose polymerase) inhibitors in patients with known germline *BRCA1*/*2* mutations or cancers that have BRCA-like biologies. The excitement over immune check-point inhibitors has spread to metastatic breast cancer, especially of the triple-negative subtype, but at present there are no agents approved for it.

MALE BREAST CANCER

Breast cancer is ~1/150th as frequent in men as in women; ~2000 men developed breast cancer annually in the United States. Risk factors include inherited, deleterious SNPs in *BRCA2*, as well as Klinefelter's syndrome. Men with Klinefelter's syndrome have two or more copies of the X chromosome and have lower levels of and higher levels of estrogen. Other conditions of hyperestrogenism, such as in hepatic failure, are also associated with higher risk of male breast cancers. However, the vast majority of men who present with breast cancer have none of these conditions.

Breast cancer usually presents in men as a unilateral lump in the breast and is frequently not diagnosed promptly. Given the small amount of soft tissue and the unexpected nature of the problem, locally advanced presentations are somewhat more common. Although gynecomastia may initially be unilateral or asymmetric, any unilateral mass in a man aged >40 years should receive a careful workup including biopsy.

On the other hand, bilateral symmetric breast development rarely represents breast cancer and is almost invariably due to endocrine disease or a drug effect. It should be kept in mind, nevertheless, that the risk of cancer is much greater in men with gynecomastia; in such men, gross asymmetry of the breasts should arouse suspicion of cancer.

Approximately 90% of male breast cancers contain ERs, and it behaves similarly to that in a postmenopausal woman. When matched to female breast cancer by age and stage, its overall prognosis is identical. Male breast cancer is best managed by mastectomy and axillary lymph node dissection or SLNB, although some men prefer breast-conserving therapy. Patients with locally advanced disease or positive nodes should also be treated with irradiation, and ~60% of cases with metastatic disease respond to endocrine therapy. Tamoxifen is usually the agent of choice, and it is unknown if the AIs are effective in men. No randomized studies have evaluated adjuvant therapy for male breast cancer. Two historic experiences suggest that the disease responds well to adjuvant systemic therapy, and, if not medically contraindicated, the same criteria for the use of adjuvant therapy in women should be applied to men.

The sites of relapse and spectrum of response to chemotherapeutic drugs are virtually identical for breast cancers in either sex.

FURTHER READING

Colleoni M et al: Annual hazard rates of recurrence for breast cancer during 24 years of follow-up: Results from the International Breast Cancer Study Group Trials I to V. J Clin Oncol 34:927, 2016. [PubMed: 26786933]

Early Breast Cancer Trialists' Collaborative (EBCTCG) et al: Adjuvant bisphosphonate treatment in early breast cancer: Meta-analyses of individual patient data from randomised trials. Lancet 386:1353, 2015. [PubMed: 26211824]

Early Breast Cancer Trialists' Collaborative (ebctcg) et al: Aromatase inhibitors versus tamoxifen in early breast cancer: Patient-level meta-analysis of the randomised trials. Lancet 386:1341, 2015. [PubMed: 26211827]

Easton DF et al: Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med 372:2243, 2015. [PubMed: 26014596]

Le Tourneau C et al: Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): A multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol 16:1324, 2015. [PubMed: 26342236]

Oeffinger KC et al: Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society. JAMA 314:1599, 2015. [PubMed: 26501536]

Perez EA et al: Trastuzumab plus adjuvant chemotherapy for human epidermal growth factor receptor 2-positive breast cancer: Planned joint analysis of overall survival from NSABP B-31 and NCCTG N9831. J Clin Oncol 32:3744, 2014. [PubMed: 25332249]

Peto R et al: Comparisons between different polychemotherapy regimens for early breast cancer: Meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. Lancet 379:432, 2012. [PubMed: 22152853]

Runowicz CD et al: American Cancer Society/American Society of Clinical Oncology Breast Cancer Survivorship Care Guideline. J Clin Oncol 34:611–35, 2016. [PubMed: 26644543]

Sparano JA et al: Prospective validation of a 21-gene expression assay in breast cancer. N Engl J Med 373:2005, 2015. [PubMed: 26412349]

Chapter 76: Upper Gastrointestinal Tract Cancers

Robert J. Mayer

FIGURE 76-1

INTRODUCTION

Upper gastrointestinal cancers include malignancies arising in the esophagus, stomach, and small intestine.

ESOPHAGEAL CANCER

INCIDENCE AND ETIOLOGY

Cancer of the esophagus is an increasingly common and extremely lethal malignancy. The diagnosis was made in 16,940 Americans in 2017 and led to 15,690 deaths. Almost all esophageal cancers are either squamous cell carcinomas or adenocarcinomas; the two histologic subtypes have a similar clinical presentation but different causative factors.

Worldwide, squamous cell carcinoma is the more common cell type, having an incidence that rises strikingly in association with geographic location. It occurs frequently within a region extending from the southern shore of the Caspian Sea on the west to northern China on the east, encompassing parts of Iran, central Asia, Afghanistan, Siberia, and Mongolia. Familial increased risk has been observed in regions with high incidence, although gene associations are not yet defined. High-incidence “pockets” of the disease are also present in such disparate locations as Finland, Iceland, Curaçao, southeastern Africa, and northwestern France. In North America and western Europe, the disease is more common in blacks than whites and in males than females; it appears most often after age 50 and seems to be associated with a lower socioeconomic status. Such cancers generally arise in the cervical and thoracic portions of the esophagus.

A variety of causative factors have been implicated in the development of squamous cell cancers of the esophagus (Table 76-1). In the United States, the etiology of such cancers is primarily related to excess alcohol consumption and/or cigarette smoking. The relative risk increases with the amount of tobacco smoked or alcohol consumed, with these factors acting synergistically. The consumption of whiskey is linked to a higher incidence than the consumption of wine or beer. Squamous cell esophageal carcinoma has also been associated with the ingestion of nitrates, smoked opiates, and fungal toxins in pickled vegetables, as well as mucosal damage caused by such physical insults as long-term exposure to extremely hot tea, the ingestion of lye, radiation-induced strictures, and chronic achalasia. The presence of an esophageal web in

association with glossitis and iron deficiency (i.e., Plummer-Vinson or Paterson-Kelly syndrome) and congenital hyperkeratosis and pitting of the palms and soles (i.e., tylosis palmaris et plantaris) have each been linked with squamous cell esophageal cancer, as have dietary deficiencies of molybdenum, zinc, selenium, and vitamin A. Patients with head and neck cancer are at increased risk of squamous cell cancer of the esophagus.

TABLE 76-1

Some Etiologic Factors Associated with Squamous Cell Cancer of the Esophagus

Excess alcohol consumption
Cigarette smoking
Other ingested carcinogens
Nitrates (converted to nitrites)
Smoked opiates
Fungal toxins in pickled vegetables
Mucosal damage from physical agents
Hot tea
Lye ingestion
Radiation-induced strictures
Chronic achalasia
Host susceptibility
Esophageal web with glossitis and iron deficiency (i.e., Plummer-Vinson or Paterson-Kelly syndrome)
Congenital hyperkeratosis and pitting of the palms and soles (i.e., tylosis palmaris et plantaris)
? Dietary deficiencies of selenium, molybdenum, zinc, and vitamin A

For unclear reasons, the incidence of squamous cell esophageal cancer has decreased somewhat in both the black and white populations in the United States over the past 40 years, whereas the rate of adenocarcinoma

has risen sevenfold, particularly in white males (male-to-female ratio of 6:1). Whereas squamous cell cancers comprised the vast majority of esophageal cancers in the United States as recently as 40–50 years ago, >75% of esophageal tumors are now adenocarcinomas, with the incidence of this histologic subtype continuing to increase rapidly. Understanding the cause for this increase is the focus of current investigation.

Several strong etiologic associations have been observed to account for the development of adenocarcinoma of the esophagus **(Table 76-2)**. Such tumors arise in the distal esophagus in association with chronic gastric reflux, often in the presence of Barrett's esophagus (replacement of the normal squamous epithelium of the distal esophagus by columnar mucosa), which occurs more commonly in obese individuals. Adenocarcinomas arise within dysplastic columnar epithelium in the distal esophagus. Even before frank neoplasia is detectable, aneuploidy and *p53* mutations are found in the dysplastic epithelium. The value of proton-pump inhibitors in reducing the risk of cancer in individuals with chronic gastric reflux or Barrett's esophagus is uncertain. These adenocarcinomas behave clinically like gastric adenocarcinomas, although they are not associated with *Helicobacter pylori* infections. Approximately 15% of esophageal adenocarcinomas overexpress the *HER2/neu* gene.

TABLE 76-2

Some Etiologic Factors Associated with Adenocarcinoma of the Esophagus

Chronic gastroesophageal reflux
Obesity
Barrett's esophagus
Male sex
Cigarette smoking

CLINICAL FEATURES

About 5% of esophageal cancers occur in the upper third of the esophagus (cervical esophagus), 20% in the middle third, and 75% in the lower third. Squamous cell carcinomas and adenocarcinomas cannot be distinguished radiographically or endoscopically.

Progressive dysphagia and weight loss of short duration are the initial symptoms in the vast majority of patients. Dysphagia initially occurs with solid foods and gradually progresses to include semisolids and liquids. By the time these symptoms develop, the disease is already very advanced, because difficulty in swallowing does not occur until >60% of the esophageal circumference is infiltrated with cancer. Dysphagia may be associated with pain on swallowing (odynophagia), pain radiating to the chest and/or back,

regurgitation or vomiting, and aspiration pneumonia. The disease most commonly spreads to adjacent and supraclavicular lymph nodes, liver, lungs, pleura, and bone. Tracheoesophageal fistulas may develop, primarily in patients with upper and mid-esophageal tumors. As with other squamous cell carcinomas, hypercalcemia may occur in the absence of osseous metastases, probably from parathormone-related peptide secreted by tumor cells (Chap. 89).

DIAGNOSIS

Attempts at endoscopic and cytologic screening for carcinoma in patients with Barrett's esophagus, while effective as a means of detecting high-grade dysplasia, have not yet been shown to reduce the likelihood of death from esophageal adenocarcinoma. Esophagoscopy should be performed in all patients suspected of having an esophageal abnormality, to both visualize and identify a tumor and also to obtain histopathologic confirmation of the diagnosis. Because the population of persons at risk for squamous cell carcinoma of the esophagus (i.e., smokers and drinkers) also has a high rate of cancers of the lung and the head and neck region, endoscopic inspection of the larynx, trachea, and bronchi should also be carried out. A thorough examination of the fundus of the stomach (by retroflexing the endoscope) is imperative as well. The extent of tumor spread to the mediastinum and para-aortic lymph nodes should be assessed by computed tomography (CT) scans of the chest and abdomen and by endoscopic ultrasound. Positron emission tomography scanning provides a useful assessment of the presence of distant metastatic disease, offering accurate information regarding spread to mediastinal lymph nodes, which can be helpful in defining radiation therapy fields. Such scans, when performed sequentially, appear to provide a means of making an early assessment of responsiveness to preoperative chemotherapy and have increasingly been used in guiding a change in clinical management.

TREATMENT

TREATMENT

Esophageal Cancer

The prognosis for patients with esophageal carcinoma is poor. Approximately 10% of patients survive 5 years after the diagnosis; thus, management focuses on symptom control. Surgical resection of all gross tumor (i.e., total resection) is feasible in only 45% of cases, with residual tumor cells frequently present at the resection margins. Such esophagectomies have been associated with a postoperative mortality rate of ~5% due to anastomotic fistulas, subphrenic abscesses, and cardiopulmonary complications. Although debate regarding the comparative benefits of transthoracic versus transhiatal resections has continued, experienced thoracic surgeons are now favoring minimally invasive transthoracic esophagectomies. Endoscopic resections of superficial squamous cell cancers or adenocarcinomas are being examined but have not yet been shown to result in a similar likelihood of survival as observed with conventional surgical procedures. Similarly, the value of endoscopic ablation of dysplastic lesions in an area of Barrett's esophagus on reducing subsequent mortality from esophageal carcinoma is uncertain. Some experts have advocated fundoplication surgery (i.e., the removal of the gastroesophageal junction) as a means of cancer prevention in patients with

Barrett's esophagus; again, objective data are not yet available to fully assess the risks versus benefits of this invasive procedure. About 20% of patients who survive a total surgical resection live for 5 years. The evaluation of chemotherapeutic agents in patients with esophageal carcinoma has been hampered by ambiguity in the definition of "response" and the debilitated physical condition of many treated individuals, particularly those with squamous cell cancers. Nonetheless, significant reductions in the size of measurable tumor masses have been reported in 15–25% of patients given single-agent treatment and in 30–60% of patients treated with drug combinations that include a platinum form of chemotherapy. In the small subset of patients whose tumors overexpress the *HER2/neu* gene, the addition of the monoclonal antibody trastuzumab (Herceptin) appears to further enhance the likelihood of benefit, particularly in patients with gastroesophageal lesions. The use of the antiangiogenic agent bevacizumab (Avastin) seems to be of limited value in the setting of esophageal cancer. Combination chemotherapy and radiation therapy as the initial therapeutic approach, either alone or followed by an attempt at operative resection, seems to be beneficial. When administered along with radiation therapy, chemotherapy produces a better survival outcome than radiation therapy alone. The use of preoperative chemotherapy and radiation therapy followed by esophageal resection appears to prolong survival compared with surgery alone according to several randomized trials and a meta-analysis; some reports suggest that no additional benefit accrues when surgery is added if significant shrinkage of tumor has been achieved by the chemoradiation combination.

For the incurable, surgically unresectable patient with esophageal cancer, dysphagia, malnutrition, and the management of tracheoesophageal fistulas are major issues. Approaches to palliation include repeated endoscopic dilatation, the surgical placement of a gastrostomy or jejunostomy for hydration and feeding, endoscopic placement of an expansive metal stent to bypass the tumor, and radiation therapy.

TUMORS OF THE STOMACH

GASTRIC ADENOCARCINOMA

Incidence and Epidemiology

 For unclear reasons, the incidence and mortality rates for gastric cancer have decreased in the United States during the past 80 years, although the disease remains the third most frequent cause of worldwide cancer-related death. The mortality rate from gastric cancer in the United States has dropped in men from 28 to 7.4 per 100,000 persons, whereas in women, the rate has decreased from 27 to 2.4 per 100,000. Nonetheless, in 2017, 28,000 new cases of stomach cancer were diagnosed in the United States, and 10,960 Americans died of the disease. Although the incidence of gastric cancer has decreased worldwide, it remains high in such disparate geographic regions as Japan, China, Chile, and Ireland.

The risk of gastric cancer is greater among lower socioeconomic classes. Migrants from high- to low-incidence nations maintain their susceptibility to gastric cancer, whereas the risk for their offspring approximates that of the new homeland. These findings suggest that an environmental exposure, probably

beginning early in life, is related to the development of gastric cancer, with dietary carcinogens considered the most likely factor(s).

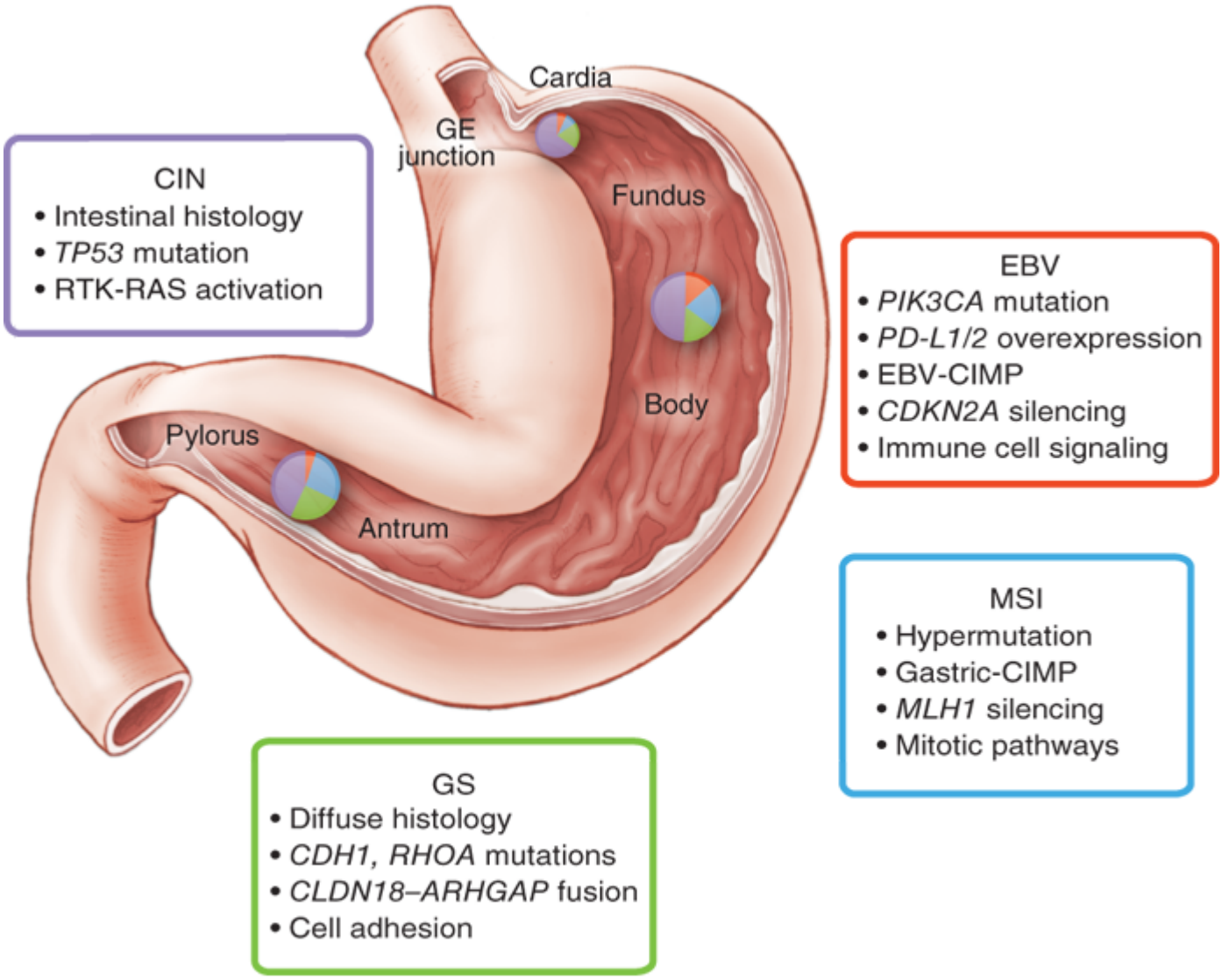
Pathology

About 85% of stomach cancers are adenocarcinomas, with 15% due to lymphomas, gastrointestinal stromal tumors (GISTs), and leiomyosarcomas. Gastric adenocarcinomas may be subdivided into two pathologically defined categories: a *diffuse type*, in which cell cohesion is absent, so that individual cells infiltrate and thicken the stomach wall without forming a discrete mass; and an *intestinal type*, characterized by cohesive neoplastic cells that form glandlike tubular structures. The diffuse carcinomas occur more often in younger patients, develop throughout the stomach (including the cardia), result in a loss of distensibility of the gastric wall (so-called *linitis plastica*, or "leather bottle" appearance), and carry a poorer prognosis. Diffuse cancers have defective intercellular adhesion, mainly as a consequence of loss of expression of E-cadherin. Intestinal-type lesions are frequently ulcerative, more commonly appear in the antrum and lesser curvature of the stomach, and are often preceded by a prolonged precancerous process, often initiated by *H. pylori* infection. Although the incidence of diffuse carcinomas is similar in most populations, the intestinal type tends to predominate in the high-risk geographic regions and is less likely to be found in areas where the frequency of gastric cancer is declining. Thus, different etiologic factor(s) are likely involved in these two subtypes. In the United States, ~30% of gastric cancers originate in the distal stomach, ~20% arise in the midportion of the stomach, and ~40% originate in the proximal third of the stomach. The remaining 10% involve the entire stomach.

Genomic profiling of gastric adenocarcinomas has led to subdividing the disease into four molecularly defined subgroups: chromosomally unstable tumors (50% of cases correlating with intestinal type histology), genomically stable tumors (20% of cases correlating with diffuse type histology), microsatellite unstable tumors (22% of cases), and Epstein-Barr virus (EBV) positive tumors (9% of cases) **(Fig. 76-1)**. Efforts to incorporate these molecular subtypes into clinical management are underway.

FIGURE 76-1

Molecular/genomic characterization of subtypes of gastric carcinomas. CIMP, CpG-island methylator phenotype; CIN, chromosomally unstable; EBV, Epstein-Barr virus-associated; GS, genomically stable; MSI, microsatellite instability-associated.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Etiology

The long-term ingestion of high concentrations of nitrates found in dried, smoked, and salted foods appears to be associated with a higher risk. The nitrates are thought to be converted to carcinogenic nitrites by bacteria (Table 76-3). Such bacteria may be introduced exogenously through the ingestion of partially decayed foods, which are consumed in abundance worldwide by the lower socioeconomic classes. Bacteria such as *H. pylori* may also contribute to this effect by causing chronic inflammatory atrophic gastritis, loss of gastric acidity, and bacterial growth in the stomach. Although the risk for developing gastric cancer is thought to be sixfold higher in people infected with *H. pylori*, it remains uncertain whether eradicating the bacteria after infection has already occurred actually reduces this risk. Loss of acidity may occur when acid-producing cells of the gastric antrum have been removed surgically to control benign peptic ulcer disease or when achlorhydria, atrophic gastritis, and even pernicious anemia develop in the elderly. Serial endoscopic examinations of the stomach in patients with atrophic gastritis have documented replacement of the usual gastric mucosa by intestinal-type cells. This process of intestinal metaplasia may lead to cellular atypia and

eventual neoplasia. Because the declining incidence of gastric cancer in the United States primarily reflects a decline in distal, ulcerating, intestinal-type lesions, it is conceivable that better food preservation and the availability of refrigeration for all socioeconomic classes have decreased the dietary ingestion of exogenous bacteria. *H. pylori* has not been associated with the diffuse, more proximal form of gastric carcinoma or with cancers arising at the gastroesophageal junction or in the distal esophagus. Approximately 10–15% of adenocarcinomas appearing in the proximal stomach, the gastroesophageal junction, and the distal esophagus overexpress the *HER2/neu* gene; individuals whose tumors demonstrate this overexpression benefit from treatment directed against this target (i.e., trastuzumab [Herceptin]).

TABLE 76-3

Nitrate-Converting Bacteria as a Factor in the Causation of Gastric Carcinomaa

Exogenous sources of nitrate-converting bacteria:
Bacterially contaminated food (common in lower socioeconomic classes, who have a higher incidence of the disease; diminished by improved food preservation and refrigeration)
Helicobacter pylori infection
Endogenous factors favoring growth of nitrate-converting bacteria in the stomach:
Decreased gastric acidity
Prior gastric surgery (antrectomy) (15- to 20-year latency period)
Atrophic gastritis and/or pernicious anemia
? Prolonged exposure to histamine H_2-receptor antagonists

[a]Hypothesis: Dietary nitrates are converted to carcinogenic nitrites by bacteria.

Several additional etiologic factors have been associated with gastric carcinoma. Gastric ulcers and adenomatous polyps have occasionally been linked, but data on a cause-and-effect relationship are unconvincing. The inadequate clinical distinction between benign gastric ulcers and small ulcerating carcinomas may, in part, account for this presumed association. The presence of extreme hypertrophy of gastric rugal folds (i.e., Ménétrier's disease), giving the impression of polypoid lesions, has been associated with a striking frequency of malignant transformation; such hypertrophy, however, does not represent the presence of true adenomatous polyps. Individuals with blood group A have a higher incidence of gastric cancer than persons with blood group O; this observation may be related to differences in the mucous secretion, leading to altered mucosal protection from carcinogens. A germline mutation in the E-cadherin

gene (*CDH1*), inherited in an autosomal dominant pattern and coding for a cell adhesion protein, has been linked to a high incidence of occult diffuse-type gastric cancers in young asymptomatic carriers in whom the endoscopic appearance of the gastric mucosa appears normal but foci of tumor are frequently present deeper in the stomach wall; this observation has led to a recommendation that they undergo a prophylactic gastrectomy. Carriers of this mutation are also at greater risk for the development of lobular breast cancer. Duodenal ulcers are not associated with gastric cancer.

Clinical Features

Gastric cancers, when superficial and surgically curable, usually produce no symptoms. As the tumor becomes more extensive, patients may complain of an insidious upper abdominal discomfort varying in intensity from a vague, postprandial fullness to a severe, steady pain. Anorexia, often with slight nausea, is very common but is not the usual presenting complaint. Weight loss may eventually be observed, and nausea and vomiting are particularly prominent in patients whose tumors involve the pylorus; dysphagia and early satiety may be the major symptoms caused by diffuse lesions originating in the cardia. There may be no early physical signs. A palpable abdominal mass indicates long-standing growth and predicts regional extension. Gastric carcinomas spread by direct extension through the gastric wall to the perigastric tissues, occasionally adhering to adjacent organs such as the pancreas, colon, or liver. The disease also spreads via lymphatics or by seeding of peritoneal surfaces. Metastases to intraabdominal and supraclavicular lymph nodes occur frequently, as do metastatic nodules to the ovary (Krukenberg's tumor), periumbilical region ("Sister Mary Joseph node"), or peritoneal cul-de-sac (Blumer's shelf palpable on rectal or vaginal examination); malignant ascites may also develop. The liver is the most common site for hematogenous spread of tumor.

The presence of iron-deficiency anemia in men and of occult blood in the stool in both sexes mandates a search for an occult gastrointestinal tract lesion. A careful assessment is of particular importance in patients with atrophic gastritis or pernicious anemia. Unusual clinical features associated with gastric adenocarcinomas include migratory thrombophlebitis, microangiopathic hemolytic anemia, diffuse seborrheic keratoses (so-called Leser-Trélat sign), and acanthosis nigricans.

Diagnosis

The use of double-contrast radiographic examinations has been supplanted by esophagogastroscopy and CT scanning for the evaluation of patients with epigastric complaints.

Gastric ulcers identified at the time of such endoscopic procedure may appear benign but merit biopsy in order to exclude a malignancy. Malignant gastric ulcers must be recognized before they penetrate into surrounding tissues, because the rate of cure of early lesions limited to the mucosa or submucosa is >80%. Because gastric carcinomas are difficult to distinguish clinically or endoscopically from gastric lymphomas, endoscopic biopsies should be made as deeply as possible, due to the submucosal location of lymphoid tumors.

The clinical staging system for gastric carcinoma is shown in **Table 76-4**.

TABLE 76-4

Staging System for Gastric Carcinoma

Stage	TNM	Features	Data from ACS in the United States	
			No. of Cases, %	5-Year Survival, %
0	$T_{is}N0M0$	Node negative; limited to mucosa	1	90
IA	T1N0M0	Node negative; invasion of lamina propria or submucosa	7	59
IB	T2N0M0 T1N1M0	Node negative; invasion of muscularis propria	10	44
II	T1N2M0 T2N1M0	Node positive; invasion beyond mucosa but within wall	17	29
		or		
	T3N0M0	Node negative; extension through wall		
IIIA	T2N2M0 T3N1-2M0	Node positive; invasion of muscularis propria or through wall	21	15
IIIB	T4N0-1M0	Node negative; adherence to surrounding tissue	14	9
IIIC	T4N2-3M0	>3 nodes positive; invasion of serosa or adjacent structures		
	T3N3M0	7 or more positive nodes; penetrates wall without invading serosa or adjacent structures		
IV	T4N2M0	Node positive; adherence to surrounding tissue	30	3
		or		

Stage	TNM	Features	Data from ACS in the United States	
			No. of Cases, %	5-Year Survival, %
	T1-4N0-2-M1	Distant metastases		

Abbreviations: ACS, American Cancer Society; TNM, tumor, node, metastasis.

TREATMENT

TREATMENT

Gastric Adenocarcinoma

Complete surgical removal of the tumor with resection of adjacent lymph nodes offers the only chance for cure. However, this is possible in less than a third of patients. A subtotal gastrectomy is the treatment of choice for patients with distal carcinomas, whereas total or near-total gastrectomies are required for more proximal tumors. The inclusion of extended lymph node dissection in these procedures appears to confer an added risk for complications without providing a meaningful enhancement in survival. The prognosis following complete surgical resection depends on the degree of tumor penetration into the stomach wall and is adversely influenced by regional lymph node involvement and vascular invasion, characteristics found in the vast majority of American patients. As a result, the probability of survival after 5 years for the 25–30% of patients able to undergo complete resection is ~20% for distal tumors and <10% for proximal tumors, with recurrences continuing for at least 8 years after surgery. In the absence of ascites or extensive hepatic or peritoneal metastases, even patients whose disease is believed to be incurable by surgery should be offered resection of the primary lesion. Reduction of tumor bulk is the best form of palliation and may enhance the probability of benefit from subsequent therapy. In high-incidence regions such as Japan and Korea, where the use of endoscopic screening programs has identified patients with superficial tumors, the use of laparoscopic gastrectomy has gained popularity. In the United States and western Europe, the use of this less invasive surgical approach remains investigational.

Gastric adenocarcinoma is a relatively radioresistant tumor, and the adequate control of the primary tumor requires doses of external-beam irradiation that exceed the tolerance of surrounding structures, such as bowel mucosa and spinal cord. As a result, the major role of radiation therapy in patients has been palliation of pain. Radiation therapy alone after a complete resection does not prolong survival. In the setting of surgically unresectable disease limited to the epigastrium, patients treated with 3500–4000 cGy did not live longer than similar patients not receiving radiotherapy; however, survival was prolonged slightly when 5-

fluorouracil (5-FU) plus leucovorin was given in combination with radiation therapy (3-year survival 50% vs 41% for radiation therapy alone). In this clinical setting, the 5-FU likely functions as a radiosensitizer.

The administration of combinations of cytotoxic drugs to patients with advanced gastric carcinoma has been associated with partial responses in 30–50% of cases; responders appear to benefit from treatment. Such drug combinations have generally included cisplatin combined with epirubicin or docetaxel and infusional 5-FU or capecitabine, or with either irinotecan or oxaliplatin. Despite the encouraging response rates, complete remissions are uncommon, the partial responses are transient, and the overall impact of multidrug therapy on survival has been limited; the median survival time for patients treated in this manner remains less than 12 months. As with adenocarcinomas arising in the esophagus, the addition of bevacizumab (Avastin) to chemotherapy regimens in treating gastric cancer appears to provide limited benefit. However, preliminary results utilizing another antiangiogenic compound—ramucirumab (Cyranza)—in the treatment of gastric cancer are encouraging, particularly when combined with paclitaxel. Additionally, initial experiences with checkpoint inhibitors (PD-1 and PD-2) have shown such immunotherapy to provide benefit to some patients. The administration of adjuvant chemotherapy alone following the complete resection of a gastric cancer has only minimally improved survival. However, combination chemotherapy administered before and after surgery (*perioperative treatment*) as well as postoperative chemotherapy combined with radiation therapy reduces the recurrence rate and prolongs survival.

PRIMARY GASTRIC LYMPHOMA

Primary lymphoma of the stomach is relatively uncommon, accounting for <15% of gastric malignancies and ~2% of all lymphomas. The stomach is, however, the most frequent extranodal site for lymphoma, and gastric lymphoma has increased in frequency during the past 35 years. The disease is difficult to distinguish clinically from gastric adenocarcinoma; both tumors are most often detected during the sixth decade of life; present with epigastric pain, early satiety, and generalized fatigue; and are usually characterized by ulcerations with a ragged, thickened mucosal pattern demonstrated by contrast radiographs or endoscopic appearance. The diagnosis of lymphoma of the stomach may occasionally be made through cytologic brushings of the gastric mucosa but usually requires a biopsy at gastroscopy or laparotomy. Failure of gastroscopic biopsies to detect lymphoma in a given case should not be interpreted as being conclusive, because superficial biopsies may miss the deeper lymphoid infiltrate. The macroscopic pathology of gastric lymphoma may also mimic adenocarcinoma, consisting of either a bulky ulcerated lesion localized in the corpus or antrum or a diffuse process spreading throughout the entire gastric submucosa and even extending into the duodenum. Microscopically, the vast majority of gastric lymphoid tumors are lymphomas of B-cell origin. Histologically, these tumors may range from well-differentiated, superficial processes (mucosa-associated lymphoid tissue [MALT]) to high-grade, large-cell lymphomas. Like gastric adenocarcinoma, infection with *H. pylori* increases the risk for gastric lymphoma in general and MALT lymphomas in particular. Large-cell lymphomas of the stomach spread initially to regional lymph nodes (often to Waldeyer's ring) and may then disseminate.

TREATMENT

TREATMENT

Primary Gastric Lymphoma

Primary gastric lymphoma is a far more treatable disease than adenocarcinoma of the stomach, a fact that underscores the need for making the correct diagnosis. Antibiotic treatment to eradicate *H. pylori* infection has led to regression of about 75% of gastric MALT lymphomas and should be considered before surgery, radiation therapy, or chemotherapy is undertaken in patients having such tumors. A lack of response to such antimicrobial treatment has been linked to a specific chromosomal abnormality, i.e., t(11;18). Responding patients should undergo periodic endoscopic surveillance because it remains unclear whether the neoplastic clone is eliminated or merely suppressed, although the response to antimicrobial treatment is quite durable. Subtotal gastrectomy, usually followed by combination chemotherapy, has led to 5-year survival rates of 40–60% in patients with localized high-grade lymphomas. The need for a major surgical procedure has been questioned, particularly in patients with preoperative radiographic evidence of nodal involvement, for whom chemotherapy (CHOP [cyclophosphamide, doxorubicin, vincristine, and prednisone]) plus rituximab is highly effective therapy. A role for radiation therapy is not defined because most recurrences develop at distant sites.

GASTRIC (NONLYMPHOID) SARCOMA

Leiomyosarcomas and GISTs make up 1–3% of gastric neoplasms. They most frequently involve the anterior and posterior walls of the gastric fundus and often ulcerate and bleed. Even those lesions that appear benign on histologic examination may behave in a malignant fashion. These tumors rarely invade adjacent viscera and characteristically do not metastasize to lymph nodes, but they may spread to the liver and lungs. The treatment of choice is surgical resection with 3 years of postoperative therapy to be considered following the removal of a GIST if the primary tumor demonstrates high-risk features. All such tumors should be analyzed for a mutation in the c-*kit* receptor. GISTs are unresponsive to conventional chemotherapy; yet ~50% of patients experience objective response and prolonged survival when treated with imatinib mesylate (Gleevec) (400–800 mg PO daily), a selective inhibitor of the c-*kit* tyrosine kinase. Many patients with GIST whose tumors have become refractory to imatinib subsequently benefit from sunitinib (Sutent) or regorafenib (Stivarga), other inhibitors of the c-*kit* tyrosine kinase.

TUMORS OF THE SMALL INTESTINE

Small-bowel tumors comprise <3% of gastrointestinal neoplasms. Because of their rarity and inaccessibility, a correct diagnosis is often delayed. Abdominal symptoms are usually vague and poorly defined, and conventional radiographic studies of the upper and lower intestinal tract often appear normal. Small-bowel tumors should be considered in the differential diagnosis in the following situations: (1) recurrent, unexplained episodes of crampy abdominal pain; (2) intermittent bouts of intestinal obstruction, especially in the absence of inflammatory bowel disease (IBD) or prior abdominal surgery; (3) intussusception in the adult; and (4) evidence of chronic intestinal bleeding in the presence of negative conventional and

endoscopic examination. A careful small-bowel barium study should be considered in such a circumstance; the diagnostic accuracy may be improved by infusing barium through a nasogastric tube placed into the duodenum (enteroclysis). Alternatively, capsule endoscopic procedures have been used.

BENIGN TUMORS

The histology of benign small-bowel tumors is difficult to predict on clinical and radiologic grounds alone. The symptomatology of benign tumors is not distinctive, with pain, obstruction, and hemorrhage being the most frequent symptoms. These tumors are usually discovered during the fifth and sixth decades of life, more often in the distal rather than the proximal small intestine. The most common benign tumors are adenomas, leiomyomas, lipomas, and angiomas.

Adenomas

These tumors include those of the islet cells and Brunner's glands as well as polypoid adenomas. *Islet cell adenomas* are occasionally located outside the pancreas; the associated syndromes are discussed in **Chap. 80**. *Brunner's gland adenomas* are not truly neoplastic but represent a hypertrophy or hyperplasia of submucosal duodenal glands. These appear as small nodules in the duodenal mucosa that secrete a highly viscous alkaline mucus. Most often, this is an incidental radiographic finding not associated with any specific clinical disorder.

Polypoid Adenomas

About 25% of benign small-bowel tumors are polypoid adenomas **(see Table 77-2)**. They may present as single polypoid lesions or, less commonly, as papillary villous adenomas. As in the colon, the sessile or papillary form of the tumor is sometimes associated with a coexisting carcinoma. Occasionally, patients with Gardner's syndrome develop premalignant adenomas in the small bowel; such lesions are generally in the duodenum. Multiple polypoid tumors may occur throughout the small bowel (and occasionally the stomach and colorectum) in the Peutz-Jeghers syndrome. The polyps are usually hamartomas (juvenile polyps) having a low potential for malignant degeneration. Mucocutaneous melanin deposits as well as tumors of the ovary, breast, pancreas, and endometrium are also associated with this autosomal dominant condition.

Leiomyomas

These neoplasms arise from smooth-muscle components of the intestine and are usually intramural, affecting the overlying mucosa. Ulceration of the mucosa may cause gastrointestinal hemorrhage of varying severity. Cramping or intermittent abdominal pain is frequently encountered.

Lipomas

These tumors occur with greatest frequency in the distal ileum and at the ileocecal valve. They have a characteristic radiolucent appearance and are usually intramural and asymptomatic, but on occasion cause bleeding.

Angiomas

While not true neoplasms, these lesions are important because they frequently cause intestinal bleeding. They may take the form of telangiectasia or hemangiomas. Multiple intestinal telangiectasias occur in a nonhereditary form confined to the gastrointestinal tract or as part of the hereditary Osler-Rendu-Weber syndrome. Vascular tumors may also take the form of isolated hemangiomas, most commonly in the jejunum. Angiography, especially during bleeding, is the best procedure for evaluating these lesions.

MALIGNANT TUMORS

While rare, small-bowel malignancies occur in patients with long-standing regional enteritis and celiac sprue as well as in individuals with AIDS. Malignant tumors of the small bowel are frequently associated with fever, weight loss, anorexia, bleeding, and a palpable abdominal mass. After ampullary carcinomas (many of which arise from biliary or pancreatic ducts), the most frequently occurring small-bowel malignancies are adenocarcinomas, lymphomas, carcinoid tumors, and leiomyosarcomas.

ADENOCARCINOMAS

The most common primary cancers of the small bowel are adenocarcinomas, accounting for ~50% of malignant tumors. These cancers occur most often in the distal duodenum and proximal jejunum, where they tend to ulcerate and cause hemorrhage or obstruction. Radiologically, they may be confused with chronic duodenal ulcer disease or with Crohn's disease if the patient has long-standing regional enteritis. The diagnosis is best made by endoscopy and biopsy under direct vision. Surgical resection is the treatment of choice with suggested postoperative adjuvant chemotherapy options generally following treatment patterns used in the management of colon cancer.

LYMPHOMAS

Lymphoma in the small bowel may be primary or secondary. A diagnosis of a primary intestinal lymphoma requires histologic confirmation in a clinical setting in which palpable adenopathy and hepatosplenomegaly are absent and no evidence of lymphoma is seen on chest radiograph, CT scan, or peripheral blood smear or on bone marrow aspiration and biopsy. Symptoms referable to the small bowel are present, usually accompanied by an anatomically discernible lesion. Secondary lymphoma of the small bowel consists of involvement of the intestine by a lymphoid malignancy extending from involved retroperitoneal or mesenteric lymph nodes (Chap. 104).

Primary intestinal lymphoma accounts for ~20% of malignancies of the small bowel. These neoplasms are non-Hodgkin's lymphomas; they usually have a diffuse, large-cell histology and are of T cell origin. Intestinal lymphoma involves the ileum, jejunum, and duodenum, in decreasing frequency—a pattern that mirrors the relative amount of normal lymphoid cells in these anatomic areas. The risk of small-bowel lymphoma is increased in patients with a prior history of malabsorptive conditions (e.g., celiac sprue), regional enteritis,

and depressed immune function due to congenital immunodeficiency syndromes, prior organ transplantation, autoimmune disorders, or AIDS.

The development of localized or nodular masses that narrow the lumen results in periumbilical pain (made worse by eating) as well as weight loss, vomiting, and occasional intestinal obstruction. The diagnosis of small-bowel lymphoma may be suspected from the appearance on contrast radiographs of patterns such as infiltration and thickening of mucosal folds, mucosal nodules, areas of irregular ulceration, or stasis of contrast material. The diagnosis can be confirmed by surgical exploration and resection of involved segments. Intestinal lymphoma can occasionally be diagnosed by peroral intestinal mucosal biopsy, but because the disease mainly involves the lamina propria, full-thickness surgical biopsies are usually required.

Resection of the tumor constitutes the initial treatment modality. While postoperative radiation therapy has been given to some patients following a total resection, most authorities favor short-term (three cycles) systemic treatment with combination chemotherapy. The frequent presence of widespread intraabdominal disease at the time of diagnosis and the occasional multicentricity of the tumor often make a total resection impossible. The probability of sustained remission or cure is ~75% in patients with localized disease but is ~25% in individuals with unresectable lymphoma. In patients whose tumors are not resected, chemotherapy may lead to bowel perforation.

A unique form of small-bowel lymphoma, diffusely involving the entire intestine, was first described in oriental Jews and Arabs and is referred to as *immunoproliferative small intestinal disease* (IPSID), *Mediterranean lymphoma*, or *α heavy chain disease*. This is a B-cell tumor. The typical presentation includes chronic diarrhea and steatorrhea associated with vomiting and abdominal cramps; clubbing of the digits may be observed. A curious feature in many patients with IPSID is the presence in the blood and intestinal secretions of an abnormal IgA that contains a shortened α heavy chain and is devoid of light chains. It is suspected that the abnormal α chains are produced by plasma cells infiltrating the small bowel. The clinical course of patients with IPSID is generally one of exacerbations and remissions, with death frequently resulting from either progressive malnutrition and wasting or the development of an aggressive lymphoma. The use of oral antibiotics such as tetracycline appears to be beneficial in the early phases of the disorder, suggesting a possible infectious etiology. Combination chemotherapy has been administered during later stages of the disease, with variable results. Results are better when antibiotics and chemotherapy are combined.

CARCINOID TUMORS

Carcinoid tumors arise from argentaffin cells of the crypts of Lieberkühn and are found from the distal duodenum to the ascending colon, areas embryologically derived from the midgut. More than 50% of intestinal carcinoids are found in the distal ileum, with most congregating close to the ileocecal valve. Most intestinal carcinoids are asymptomatic and of low malignant potential, but invasion and metastases may occur, leading to the carcinoid syndrome (Chap. 80).

LEIOMYOSARCOMAS

Leiomyosarcomas often are >5 cm in diameter and may be palpable on abdominal examination. Bleeding, obstruction, and perforation are common. Such tumors should be analyzed for the expression of mutant c-*kit* receptor (defining GIST), and in the presence of metastatic disease, justifying treatment with imatinib mesylate (Gleevec) or, in imatinib-refractory patients, sunitinib (Sutent) or regorafenib (Stivarga).

FURTHER READING

Bass AJ et al: The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513:202, 2014. [PubMed: 25079317]

Enzinger PC et al: CALGB 80403 (Alliance)/E1206: A randomized phase II study of three chemotherapy regimines plus cetuximab in metastatic esophageal and gastroesophageal junction cancers. J Clin Oncol 34:2736, 2016. [PubMed: 27382098]

Pennathur A et al: Oesophageal carcinoma. Lancet 381:400, 2013. [PubMed: 23374478]

Rustgi AK, El-Serag HB: Esophageal carcinoma. N Engl J Med 371:2499, 2014. [PubMed: 25539106]

Van Cutsem E et al: Gastric cancer. Lancet 388:2654, 2016. [PubMed: 27156933]

Van Hagen P et al: Preoperative chemoradiotherapy for esophageal or junctional cancer. N Engl N Med 366:2074, 2012.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 77: Lower Gastrointestinal Cancers

Robert J. Mayer

INTRODUCTION

Lower gastrointestinal cancers include malignant tumors of the colon, rectum, and anus.

COLORECTAL CANCER

INCIDENCE


Cancer of the large bowel is second only to lung cancer as a cause of cancer death in the United States: 135,430 new cases occurred in 2017, and 50,260 deaths were due to colorectal cancer. The incidence rate has decreased significantly during the past 25 years, likely due in large part to enhanced and more compliantly followed screening practices. Similarly, mortality rates in the United States have decreased by ~25%, resulting largely from earlier detection and improved treatment.

POLYPS AND MOLECULAR PATHOGENESIS

Most colorectal cancers, regardless of etiology, arise from adenomatous polyps. A polyp is a grossly visible protrusion from the mucosal surface and may be classified pathologically as a nonneoplastic hamartoma (e.g., *juvenile polyp*), a hyperplastic mucosal proliferation (*hyperplastic polyp*), or an adenomatous polyp. Only adenomas are clearly premalignant, and only a minority of adenomatous polyps evolve into cancer. Adenomatous polyps may be found in the colons of ~30% of middle-aged and ~50% of elderly people; however, <1% of polyps ever become malignant. Most polyps produce no symptoms and remain clinically undetected. Occult blood in the stool is found in <5% of patients with polyps.

A number of molecular changes are noted in adenomatous polyps and colorectal cancers that are thought to reflect a multistep process in the evolution of normal colonic mucosa to life-threatening invasive carcinoma. These developmental steps toward carcinogenesis include, but are not restricted to, point mutations in the K-*ras* protooncogene; hypomethylation of DNA, leading to gene activation; loss of DNA (*allelic loss*) at the site of a tumor-suppressor gene (the adenomatous polyposis coli [*APC*] gene) on the long arm of chromosome 5 (5q21); allelic loss at the site of a tumor-suppressor gene located on chromosome 18q (the deleted in colorectal cancer [*DCC*] gene); and allelic loss at chromosome 17p, associated with mutations in the *p53* tumor-suppressor gene **(see Fig. 67-2)**. Thus, the altered proliferative pattern of the colonic mucosa, which results in progression to a polyp and then to carcinoma, may involve the mutational activation of an oncogene followed by and coupled with the loss of genes that normally suppress tumorigenesis. It remains uncertain whether the genetic aberrations always occur in a defined order. Based on this model, however, cancer is believed to develop only in those polyps in which most (if not all) of these mutational events take place.

Clinically, the probability of an adenomatous polyp becoming a cancer depends on the gross appearance of the lesion, its histologic features, and its size. Polyps may be pedunculated (stalked) or sessile (flat-based), adenomatous or serrated. Invasive cancers develop more frequently in sessile, serrated (i.e. “flat”) polyps. Histologically, adenomatous polyps may be tubular, villous (i.e., papillary), or tubulovillous. Villous adenomas, most of which are sessile, become malignant more than three times as often as tubular adenomas. The likelihood that any polypoid lesion in the large bowel contains invasive cancer

is related to the size of the polyp, being negligible (<2%) in lesions <1.5 cm, intermediate (2–10%) in lesions 1.5–2.5 cm, and substantial (10%) in lesions >2.5 cm in size.

Following the detection of an adenomatous polyp, the entire large bowel should be visualized endoscopically because synchronous lesions are noted in about one-third of cases. Colonoscopy should then be repeated periodically, even in the absence of a previously documented malignancy, because such patients have a 30–50% probability of developing another adenoma and are at a higher-than-average risk for developing a colorectal carcinoma. Adenomatous polyps are thought to require >5 years of growth before becoming clinically significant; colonoscopy need not be carried out more frequently than every 3 years for the vast majority of patients.

ETIOLOGY AND RISK FACTORS

Risk factors for the development of colorectal cancer are listed in Table 77-1.

TABLE 77-1

Risk Factors for the Development of Colorectal Cancer

Diet: Animal fat
Hereditary syndromes
Polyposis coli
MYH-associated polyposis
Nonpolyposis syndrome (Lynch's syndrome)
Inflammatory bowel disease
Streptococcus bovis bacteremia
? Tobacco use

Diet

The etiology for most cases of large-bowel cancer appears to be related to environmental factors. The disease occurs more often in upper socioeconomic populations who live in urban areas. Mortality from colorectal cancer is directly correlated with per capita consumption of calories, meat protein, and dietary fat and oil as well as elevations in the serum cholesterol concentration and mortality from coronary artery disease. Geographic variations in incidence largely are unrelated to genetic differences, since migrant groups tend to assume the large-bowel cancer incidence rates of their adopted countries. Furthermore, population groups such as Mormons and Seventh Day Adventists, whose lifestyle and dietary habits differ somewhat from those of their neighbors, have significantly lower-than-expected incidence and mortality rates for colorectal cancer. The incidence of colorectal cancer has increased in Japan since that nation has adopted a more "Western" diet. At least three hypotheses have been proposed to explain the relationship to diet, none of which is fully satisfactory.

ANIMAL FATS

One hypothesis is that the ingestion of animal fats found in red meats and processed meat leads to an increased proportion of anaerobes in the gut microflora, resulting in the conversion of normal bile acids into carcinogens. This provocative hypothesis is supported by several reports of increased amounts of fecal anaerobes in the stools of patients with colorectal cancer. Diets high in animal (but not vegetable) fats are also associated with high serum cholesterol, which is also associated with enhanced risk for the development of colorectal adenomas and carcinomas.

INSULIN RESISTANCE

The large number of calories in Western diets coupled with physical inactivity has been associated with a higher prevalence of obesity. Obese persons develop insulin resistance with increased circulating levels of insulin, leading to higher circulating concentrations of insulin-like growth factor type I (IGF-I). This growth factor appears to stimulate proliferation of the intestinal mucosa.

FIBER

Contrary to prior beliefs, the results of randomized trials and case-controlled studies have *failed* to show any value for dietary fiber or diets high in fruits and vegetables in preventing the recurrence of colorectal adenomas or the development of colorectal cancer.

The weight of epidemiologic evidence, however, implicates diet as being the major etiologic factor for colorectal cancer, particularly diets high in animal fat and in calories.

HEREDITARY FACTORS AND SYNDROMES

Up to 25% of patients with colorectal cancer have a family history of the disease, suggesting a hereditary predisposition. Inherited large-bowel cancers can be divided into two main groups: the well-studied but uncommon polyposis syndromes and the more common nonpolyposis syndromes **(Table 77-2)**.

TABLE 77-2

Hereditable (Autosomal Dominant) Gastrointestinal Polyposis Syndromes

Syndrome	Distribution of Polyps	Histologic Type	Malignant Potential	Associated Lesions
Familial adenomatous polyposis	Large intestine	Adenoma	Common	None
Gardner's syndrome	Large and small intestines	Adenoma	Common	Osteomas, fibromas, lipomas, epidermoid cysts, ampullary cancers, congenital hypertrophy of retinal pigment epithelium
Turcot's syndrome	Large intestine	Adenoma	Common	Brain tumors
MYH-associated polyposis	Large intestine	Adenoma	Common	None
Nonpolyposis syndrome (Lynch's syndrome)	Large intestine (often proximal)	Adenoma	Common	Endometrial and ovarian tumors (most frequently) gastric, genitourinary, pancreatic, biliary cancers (less frequently)
Peutz-Jeghers syndrome	Small and large intestines, stomach	Hamartoma	Rare	Mucocutaneous pigmentation; tumors of the ovary, breast, pancreas, endometrium
Juvenile polyposis	Large and small intestines, stomach	Hamartoma, rarely progressing to adenoma	Rare	Various congenital abnormalities

Polyposis Coli

Polyposis coli (familial polyposis of the colon) is a rare condition characterized by the appearance of thousands of adenomatous polyps throughout the large bowel. It is transmitted as an autosomal dominant trait; the occasional patient with no family history probably developed the condition due to a spontaneous mutation. Polyposis coli is associated with a deletion in the long arm of chromosome 5 (including the *APC* gene) in both neoplastic (somatic mutation) and normal (germline mutation) cells. The loss of this genetic material (i.e., allelic loss) results in the absence of tumor-suppressor genes whose protein products would normally inhibit neoplastic growth. The presence of soft tissue and bony tumors, congenital hypertrophy of the retinal pigment epithelium, mesenteric desmoid tumors, and ampullary cancers in addition to the colonic polyps characterizes a subset of polyposis coli known as *Gardner's syndrome*. The appearance of malignant tumors of the

central nervous system accompanying polyposis coli defines *Turcot's syndrome*. The colonic polyps in all these conditions are rarely present before puberty but are generally evident in affected individuals by age 25. If the polyposis is not treated surgically, colorectal cancer will develop in almost all patients aged <40. Polyposis coli results from a defect in the colonic mucosa, leading to an abnormal proliferative pattern and impaired DNA repair mechanisms. Once the multiple polyps are detected, patients should undergo a total colectomy. Medical therapy with nonsteroidal anti-inflammatory drugs (NSAIDs) such as sulindac and selective cyclooxygenase-2 inhibitors such as celecoxib can decrease the number and size of polyps in patients with polyposis coli; however, this effect on polyps is only temporary, and the use of NSAIDs has not been shown to reduce the risk of cancer. Colectomy remains the primary therapy/prevention. The offspring of patients with polyposis coli, who often are prepubertal when the diagnosis is made in the parent, have a 50% risk for developing this premalignant disorder and should be carefully screened by annual flexible sigmoidoscopy until age 35. Proctosigmoidoscopy is a sufficient screening procedure because polyps tend to be evenly distributed from cecum to anus, making more invasive and expensive techniques such as colonoscopy or barium enema unnecessary. Testing for occult blood in the stool is an inadequate screening maneuver. If a causative germline *APC* mutation has been identified in an affected family member, an alternative method for identifying carriers is testing DNA from peripheral blood mononuclear cells for the presence of the specific *APC* mutation. The detection of such a germline mutation can lead to a definitive diagnosis before the development of polyps.

MYH-Associated Polyposis

MYH-associated polyposis (MAP) is a rare autosomal recessive syndrome caused by a biallelic mutation in the *MUT4H* gene. This hereditary condition may have a variable clinical presentation, resembling polyposis coli or colorectal cancer occurring in younger individuals without polyposis. Screening and colectomy guidelines for this syndrome are less clear than for polyposis coli, but annual to biennial colonoscopic surveillance is generally recommended starting at age 25–30.

Hereditary Nonpolyposis Colon Cancer

Hereditary nonpolyposis colon cancer (HNPCC), also known as *Lynch's syndrome*, is another autosomal dominant trait. It is characterized by the presence of three or more relatives with histologically documented colorectal cancer, one of whom is a first-degree relative of the other two; one or more cases of colorectal cancer diagnosed before age 50 in the family; and colorectal cancer involving at least two generations. In contrast to polyposis coli, HNPCC is associated with an unusually high frequency of cancer arising in the proximal large bowel. The median age for the appearance of an adenocarcinoma is <50 years, 10–15 years younger than the median age for the general population. Despite having a poorly differentiated, mucinous histologic appearance, the proximal colon tumors that characterize HNPCC have a better prognosis than sporadic tumors from patients of similar age. Families with HNPCC often include individuals with multiple primary cancers; the association of colorectal cancer with either ovarian or endometrial carcinomas is especially strong in women, and an increased appearance of gastric, small-bowel, genitourinary, pancreaticobiliary, and sebaceous skin tumors has been reported as well. It has been recommended that members of such families undergo annual or biennial colonoscopy beginning at age 25 years, with intermittent pelvic ultrasonography and endometrial biopsy for afflicted women; such a screening strategy has not yet been validated. HNPCC is associated with germline mutations of several genes, particularly *hMSH2* on chromosome 2 and *hMLH1* on chromosome 3. These mutations lead to errors in DNA replication and are thought to result in DNA instability because of defective repair of DNA mismatches resulting in abnormal cell growth and tumor development. Testing tumor cells through molecular analysis of DNA for "microsatellite instability" or immunohistochemical staining for deficiency in mismatch repair proteins in patients with colorectal cancer and a positive family history for colorectal or endometrial cancer may identify probands with HNPCC.

INFLAMMATORY BOWEL DISEASE

(Chap. 319) Large-bowel cancer is increased in incidence in patients with long-standing inflammatory bowel disease (IBD). Cancers develop more commonly in patients with ulcerative colitis than in those with granulomatous (i.e., Crohn's) colitis, but

this impression may result in part from the occasional difficulty of differentiating these two conditions. The risk of colorectal cancer in a patient with IBD is relatively small during the initial 10 years of the disease, but then appears to increase at a rate of ~0.5–1% per year. Cancer may develop in 8–30% of patients after 25 years. The risk is higher in younger patients with pancolitis.

Cancer surveillance strategies in patients with IBD are unsatisfactory. Symptoms such as bloody diarrhea, abdominal cramping, and obstruction, which may signal the appearance of a tumor, are similar to the complaints caused by a flare-up of the underlying disease. In patients with a history of IBD lasting ≥15 years who continue to experience exacerbations, the surgical removal of the colon can significantly reduce the risk for cancer and also eliminate the target organ for the underlying chronic gastrointestinal disorder. The value of such surveillance techniques as colonoscopy with mucosal biopsies and brushings for less symptomatic individuals with chronic IBD is uncertain. The lack of uniformity regarding the pathologic criteria that characterize dysplasia and the absence of data that such surveillance reduces the development of lethal cancers have made this costly practice an area of controversy.

OTHER HIGH-RISK CONDITIONS

Streptococcus bovis Bacteremia

For unknown reasons, individuals who develop endocarditis or septicemia from this fecal bacterium have a high incidence of occult colorectal tumors and, possibly, upper gastrointestinal cancers as well. Endoscopic or radiographic screening appears advisable.

Tobacco Use

Cigarette smoking is linked to the development of colorectal adenomas, particularly after >35 years of tobacco use. No biologic explanation for this association has yet been proposed.

PRIMARY PREVENTION

Several orally administered compounds have been assessed as possible inhibitors of colon cancer. The most effective class of chemopreventive agents is aspirin and other NSAIDs, which are thought to suppress cell proliferation by inhibiting prostaglandin synthesis. Regular aspirin use reduces the risk of colon adenomas and carcinomas as well as death from large-bowel cancer; such use also appears to diminish the likelihood for developing additional premalignant adenomas following successful treatment for a prior colon carcinoma. This effect of aspirin on colon carcinogenesis increases with the duration and dosage of drug use. Emerging data linking adequate plasma levels of vitamin D with reduced risk of adenomatous polyps and colorectal cancer appear promising. The value of vitamin D as a form of chemoprevention is under study. Antioxidant vitamins such as ascorbic acid, tocopherols, and β-carotene are ineffective at reducing the incidence of subsequent adenomas in patients who have undergone the removal of a colon adenoma. Estrogen replacement therapy has been associated with a reduction in the risk of colorectal cancer in women, conceivably by an effect on bile acid synthesis and composition or by decreasing synthesis of IGF-I.

SCREENING

The rationale for colorectal cancer screening programs is that the removal of adenomatous polyps will prevent colorectal cancer, and that earlier detection of localized, superficial cancers in asymptomatic individuals will increase the surgical cure rate. Such screening programs are particularly important for individuals with a family history of the disease in first-degree relatives. The relative risk for developing colorectal cancer increases to 1.75 in such individuals and may be even higher if the relative was afflicted before age 60. The prior use of rigid proctosigmoidoscopy as a screening tool was based on the observation that 60% of early lesions are located in the rectosigmoid. For unexplained reasons, however, the proportion of

large-bowel cancers arising in the rectum has been decreasing during the past several decades, with a corresponding increase in the proportion of cancers in the more proximal descending colon. As such, the potential for rigid proctosigmoidoscopy to detect a sufficient number of occult neoplasms to make the procedure cost-effective has been questioned.

Screening strategies for colorectal cancer that have been examined during the past several decades are listed in Table 77-3.

TABLE 77-3

Screening Strategies for Colorectal Cancer

Digital rectal examination
Stool testing
• Occult blood
• Fecal DNA
Imaging
• Contrast barium enema
• Virtual (i.e., computed tomography colonography)
Endoscopy
• Flexible sigmoidoscopy
• Colonoscopy

Many programs directed at the early detection of colorectal cancers have focused on digital rectal examinations and fecal occult blood (i.e., stool guaiac) testing. The digital examination should be part of any routine physical evaluation in adults aged >40 years, serving as a screening test for prostate cancer in men, a component of the pelvic examination in women, and an inexpensive maneuver for the detection of masses in the rectum. However, because of the proximal migration of colorectal tumors, its value as an overall screening modality for colorectal cancer has become limited. The development of the fecal occult blood test has greatly facilitated the detection of occult fecal blood. Unfortunately, even when performed optimally, the fecal occult blood test has major limitations as a screening technique. About 50% of patients with documented colorectal cancers have a negative fecal occult blood test, consistent with the intermittent bleeding pattern of these tumors. When random cohorts of asymptomatic persons have been tested, 2–4% have fecal occult blood-positive stools. Colorectal cancers have been found in <10% of these “test-positive” cases, with benign polyps being detected in an additional 20–30%. Thus, a colorectal neoplasm will not be found in most asymptomatic individuals with occult blood in their stool. Nonetheless, persons found to have fecal occult blood-positive stool routinely undergo further medical evaluation, including sigmoidoscopy and/or colonoscopy—procedures that are not only uncomfortable and expensive but also associated with a small risk for significant complications. The added cost of these studies would appear justifiable if the small number of patients found to have occult neoplasms because of fecal occult blood screening could be shown to have an improved prognosis and prolonged survival. Prospectively controlled trials have shown a statistically significant reduction in mortality rate from colorectal cancer for individuals undergoing annual stool guaiac screening. However, this benefit only emerged after >13 years of follow-up and was extremely expensive to achieve because all positive tests (most of which were falsely positive) were followed by

colonoscopy. Moreover, these colonoscopic examinations quite likely provided the opportunity for cancer prevention through the removal of potentially premalignant adenomatous polyps because the eventual development of cancer was reduced by 20% in the cohort undergoing annual screening.

With the appreciation that the carcinogenic process leading to the progression of the normal bowel mucosa to an adenomatous polyp and then to a cancer is the result of a series of molecular changes, investigators have examined fecal DNA for evidence of mutations associated with such molecular changes as evidence of the occult presence of precancerous lesions or actual malignancies. Such a strategy has been tested in >4000 asymptomatic individuals whose stool was assessed for occult blood and for 21 possible mutations in fecal DNA; these study subjects also underwent colonoscopy. Although the fecal DNA strategy suggested the presence of more advanced adenomas and cancers than did the fecal occult blood testing approach, the overall sensitivity, using colonoscopic findings as the standard, was <50%, diminishing enthusiasm for further pursuit of the fecal DNA screening strategy.

The use of imaging studies to screen for colorectal cancers has also been explored. Air contrast barium enemas had been used to identify sources of occult blood in the stool prior to the advent of fiberoptic endoscopy; the cumbersome nature of the procedure and inconvenience to patients limited its widespread adoption. The introduction of computed tomography (CT) scanning led to the development of virtual (i.e., CT) colonography as an alternative to the growing use of endoscopic screening techniques. Virtual colonography was proposed as being equivalent in sensitivity to colonoscopy and being available in a more widespread manner because it did not require the same degree of operator expertise as fiberoptic endoscopy. However, virtual colonography requires the same cathartic preparation that has limited widespread acceptance in association with endoscopic colonoscopy, is diagnostic but not therapeutic (i.e., patients with suspicious findings must undergo a subsequent endoscopic procedure for polypectomy or biopsy), and, in the setting of general radiology practices, appears to be less sensitive as a screening technique when compared with endoscopic procedures.

With the appreciation of the inadequacy of fecal occult blood testing alone, concerns about the practicality of imaging approaches, and the wider adoption of endoscopic examinations by the primary care community, screening strategies in asymptomatic persons have changed. At present, both the American Cancer Society and the National Comprehensive Cancer Network recommend either fecal occult blood testing annually coupled with flexible sigmoidoscopy every 5 years or colonoscopy every 10 years beginning at age 50 in asymptomatic individuals with no personal or family history of polyps or colorectal cancer. The recommendation for the inclusion of flexible sigmoidoscopy is strongly supported by the recently published results of three randomized trials performed in the United States, the United Kingdom, and Italy, involving >350,000 individuals, which consistently showed that periodic (even single) sigmoidoscopic examinations, after more than a decade of median follow-up, lead to an ~21% reduction in the development of colorectal cancer and a >25% reduction in mortality from the malignant disease. Less than 20% of participants in these studies underwent a subsequent colonoscopy. In contrast to the cathartic preparation required before colonoscopic procedures, which is only performed by highly trained specialists, flexible sigmoidoscopy requires only an enema as preparation and can be accurately performed by nonspecialty physicians or physician-extenders. The randomized screening studies using flexible sigmoidoscopy led to the estimate that ~650 individuals needed to be screened to prevent one colorectal cancer death; this contrasts with the data for mammography where the number of women needing to be screened to prevent one breast cancer death is 2500, reinforcing the efficacy of endoscopic surveillance for colorectal cancer screening. Presumably the benefit from the sigmoidoscopic screening is the result of the identification and removal of adenomatous polyps; it is intriguing that this benefit has been achieved using a technique that leaves the proximal half of the large bowel unvisualized.

It remains to be seen whether surveillance colonoscopy, which has gained increasing popularity in the United States for colorectal cancer screening, will prove to be more effective than flexible sigmoidoscopy. Ongoing randomized trials being conducted in Europe are addressing this issue. Although flexible sigmoidoscopy only visualizes the distal half of the large bowel, leading to the assumption that colonoscopy represents a more informative approach, colonoscopy has been reported as being less accurate for screening the proximal rather than the distal colon, perhaps due to technical considerations but also

possibly because of a greater frequency of serrated (i.e., “flat”) polyps in the right colon, which are more difficult to identify. At present, colonoscopy performed every 10 years has been offered as an alternative to annual fecal occult blood testing with periodic (every 5 years) flexible sigmoidoscopy. Colonoscopy has been shown to be superior to double-contract barium enema and also to have a higher sensitivity for detecting villous or dysplastic adenomas or cancers than the strategy using occult fecal blood testing and flexible sigmoidoscopy. Whether colonoscopy performed every 10 years beginning at age 50 is medically superior and economically equivalent to flexible sigmoidoscopy remains to be determined.

CLINICAL FEATURES

Presenting Symptoms

Symptoms vary with the anatomic location of the tumor. Because stool is relatively liquid as it passes through the ileocecal valve into the right colon, cancers arising in the cecum and ascending colon may become quite large without resulting in any obstructive symptoms or noticeable alterations in bowel habits. Lesions of the right colon commonly ulcerate, leading to chronic, insidious blood loss without a change in the appearance of the stool. Consequently, patients with tumors of the ascending colon often present with symptoms such as fatigue, palpitations, and even angina pectoris and are found to have a hypochromic, microcytic anemia indicative of iron deficiency. Because the cancers may bleed intermittently, a random fecal occult blood test may be negative. As a result, the unexplained presence of iron-deficiency anemia in any adult (with the possible exception of a premenopausal, multiparous woman) mandates a thorough endoscopic and/or radiographic visualization of the entire large bowel (Fig. 77-1).

FIGURE 77-1

Double-contrast air-barium enema revealing a sessile tumor of the cecum in a patient with iron-deficiency anemia and guaiac-positive stool. The lesion at surgery was a stage II adenocarcinoma.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Because stool becomes more formed as it passes into the transverse and descending colon, tumors arising there tend to impede the passage of stool, resulting in the development of abdominal cramping, occasional obstruction, and even

perforation. Radiographs of the abdomen often reveal characteristic annular, constricting lesions (“apple-core” or “napkin-ring”) (Fig. 77-2).

FIGURE 77-2

Annular, constricting adenocarcinoma of the descending colon. This radiographic appearance is referred to as an “apple-core” lesion and is always highly suggestive of malignancy.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

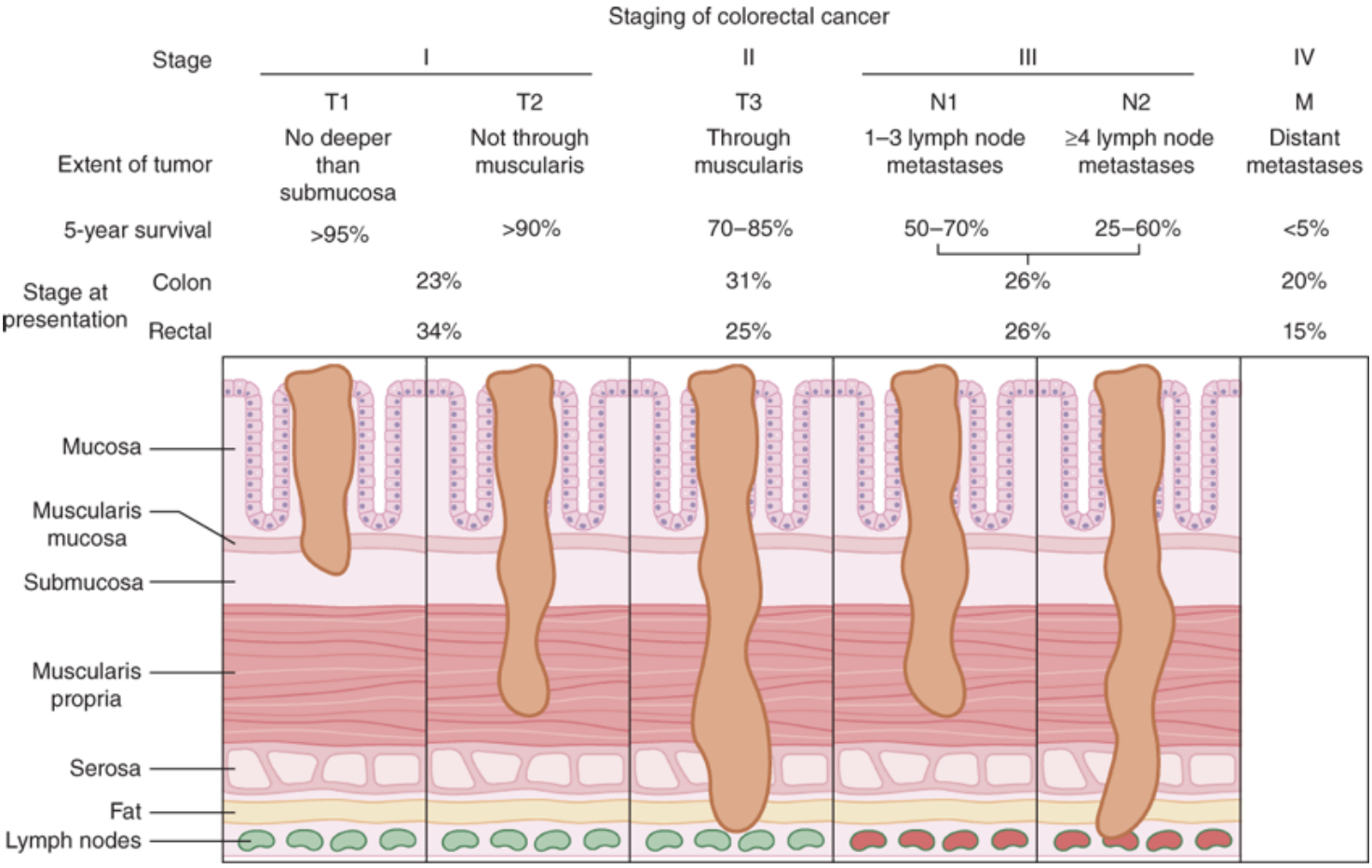
Cancers arising in the rectosigmoid are often associated with hematochezia, tenesmus, and narrowing of the caliber of stool; anemia is an infrequent finding. While these symptoms may lead patients and their physicians to suspect the presence of hemorrhoids, the development of rectal bleeding and/or altered bowel habits demands a prompt digital rectal examination and proctosigmoidoscopy.

Staging, Prognostic Factors, and Patterns of Spread

The prognosis for individuals having colorectal cancer is related to the depth of tumor penetration into the bowel wall and the presence of both regional lymph node involvement and distant metastases. These variables are incorporated into a TNM classification method, in which T represents the depth of tumor penetration, N the presence of lymph node involvement, and M the presence or absence of distant metastases (Fig. 77-3). Superficial lesions that do not involve regional lymph nodes and do not penetrate through the submucosa (T1) or the muscularis (T2) are designated as *stage I* (T1–2N0M0) disease; tumors that penetrate through the muscularis but have not spread to lymph nodes are *stage II* disease (T3-4N0M0); regional lymph node involvement defines *stage III* (TXN1-2M0) disease; and metastatic spread to sites such as liver, lung, or bone indicates *stage IV* (TXNXM1) disease. Unless gross evidence of metastatic disease is present, disease stage cannot be determined accurately before surgical resection and pathologic analysis of the operative specimens.

FIGURE 77-3

Staging and prognosis for patients with colorectal cancer.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Most recurrences after a surgical resection of a large-bowel cancer occur within the first 4 years, making 5-year survival a fairly reliable indicator of cure. The likelihood for 5-year survival in patients with colorectal cancer is stage-related (Fig. 77-3). That likelihood has improved during the past several decades when similar surgical stages have been compared. The most plausible explanation for this improvement is more thorough intraoperative and pathologic staging. In particular, more exacting attention to pathologic detail has revealed that the prognosis following the resection of a colorectal cancer is not related merely to the presence or absence of regional lymph node involvement; rather, prognosis may be more precisely gauged by the number of involved lymph nodes (one to three lymph nodes [“N1”] vs four or more lymph nodes [“N2”]) and the number of nodes examined. A minimum of 12 sampled lymph nodes is thought necessary to accurately define tumor stage, and the more nodes examined, the better. Other predictors of a poor prognosis after a total surgical resection include tumor penetration through the bowel wall into pericolic fat, poorly differentiated histology, perforation and/or tumor adherence to adjacent organs (increasing the risk for an anatomically adjacent recurrence), and venous invasion by tumor **(Table 77-4)**. Regardless of the clinicopathologic stage, a preoperative elevation of the plasma carcinoembryonic antigen (CEA) level predicts eventual tumor recurrence. The presence of specific chromosomal aberrations, particularly a mutation in the *b-raf* gene in tumor cells, appears to predict for a higher risk for metastatic spread. Conversely, the detection of microsatellite instability in tumor tissue indicates a more favorable outcome. Tumors arising in the left colon are associated with a better prognosis than those appearing in the right colon, likely due to differences in molecular patterns. In contrast to most other cancers, the prognosis in colorectal cancer is not influenced by the size of the primary lesion when adjusted for nodal involvement and histologic differentiation.

TABLE 77-4

Predictors of Poorer Outcomes Following Total Surgical Resection of Colorectal Cancer

Tumor spread to regional lymph nodes
Number of regional lymph nodes involved
Tumor penetration through the bowel wall
Poorly differentiated histology
Perforation
Tumor adherence to adjacent organs
Venous invasion
Preoperative elevation of CEA titer (>5 ng/mL)
Specific chromosomal deletion (e.g., mutation in the *b-raf* gene) Right-sided location of primary tumor

Abbreviation: CEA, carcinoembryonic antigen.

Cancers of the large bowel generally spread to regional lymph nodes or to the liver via the portal venous circulation. The liver represents the most frequent visceral site of metastasis; it is the initial site of distant spread in one-third of recurring colorectal cancers and is involved in more than two-thirds of such patients at the time of death. In general, colorectal cancer rarely spreads to the lungs, supraclavicular lymph nodes, bone, or brain without prior spread to the liver. A major exception to this rule occurs in patients having primary tumors in the distal rectum, from which tumor cells may spread through the paravertebral venous plexus, escaping the portal venous system and thereby reaching the lungs or supraclavicular lymph nodes without hepatic involvement. The median survival after the detection of distant metastases has increased during the last 30 years from 6–9 months (hepatomegaly, abnormal liver chemistries) to 27–30 months (small liver nodule initially identified by elevated CEA level and subsequent CT scan) with increasingly effective systemic therapy improving this prognosis further.

Efforts to use gene expression profiles to identify patients at risk of recurrence or those particularly likely to benefit from adjuvant therapy have not yet yielded practice-changing results. Despite a burgeoning literature examining a host of prognostic factors, pathologic stage at diagnosis remains the best predictor of long-term prognosis. Patients with lymphovascular invasion and high preoperative CEA levels are likely to have a more aggressive clinical course.

TREATMENT

TREATMENT

Colorectal Cancer

Total resection of tumor is the optimal treatment when a malignant lesion is detected in the large bowel. An evaluation for the presence of metastatic disease, including a thorough physical examination, biochemical assessment of liver function,

measurement of the plasma CEA level, and a CT scan of the chest, abdomen, and pelvis, should be performed before surgery. When possible, a colonoscopy of the entire large bowel should be performed to identify synchronous neoplasms and/or polyps. The detection of metastases should not preclude surgery in patients with tumor-related symptoms such as gastrointestinal bleeding or obstruction, but it often prompts the use of a less radical operative procedure. The necessity for a primary tumor resection in asymptomatic individuals with metastatic disease is an area of controversy. At the time of laparotomy, the entire peritoneal cavity should be examined, with thorough inspection of the liver, pelvis, and hemidiaphragm and careful palpation of the full length of the large bowel. Following recovery from a complete resection, patients should be observed carefully for 5 years by semiannual physical examinations and blood chemistry measurements. If a complete colonoscopy was not performed preoperatively, it should be carried out within the first several postoperative months. Some authorities favor measuring plasma CEA levels at 3-month intervals because of the sensitivity of this test as a marker for otherwise undetectable tumor recurrence. Subsequent endoscopic surveillance of the large bowel, probably at triennial intervals, is indicated, because patients who have been cured of one colorectal cancer have a 3–5% probability of developing an additional bowel cancer during their lifetime and a >15% risk for the development of adenomatous polyps. Anastomotic ("suture-line") recurrences are infrequent in colorectal cancer patients, provided the surgical resection margins were adequate and free of tumor. The value of periodic CT scans of the abdomen, assessing for an early, asymptomatic indication of tumor recurrence, while uncertain, has been recommended annually for the first 3 postoperative years.

Radiation therapy to the pelvis is recommended for patients with rectal cancer because it reduces the 20–25% probability of regional recurrences following complete surgical resection of stage II or III tumors, especially if they have penetrated through the serosa. This alarmingly high rate of local disease recurrence is believed to be due to the fact that the contained anatomic space within the pelvis limits the extent of the resection and because the rich lymphatic network of the pelvic side wall immediately adjacent to the rectum facilitates the early spread of malignant cells into surgically inaccessible tissue. The use of sharp rather than blunt dissection of rectal cancers (*total mesorectal excision*) appears to reduce the likelihood of local disease recurrence to ~10%. Radiation therapy, either pre- or postoperatively, further reduces the likelihood of pelvic recurrences but does not appear to prolong survival. Combining radiation therapy with 5-fluorouracil (5-FU)-based chemotherapy, preferably prior to surgical resection, lowers local recurrence rates and improves overall survival. Preoperative radiotherapy is indicated for patients with large, potentially unresectable rectal cancers; such lesions may shrink enough to permit subsequent surgical removal. Radiation therapy alone is not effective as the primary treatment of colon cancer.

Systemic therapy for patients with colorectal cancer has become more effective. 5-FU remains the backbone of treatment for this disease. Partial responses are obtained in 15–20% of patients. The probability of tumor response appears to be somewhat greater for patients with liver metastases when chemotherapy is infused directly into the hepatic artery, but intraarterial treatment is costly and toxic and does not appear to appreciably prolong survival. The concomitant administration of folinic acid (leucovorin [LV]) improves the efficacy of 5-FU in patients with advanced colorectal cancer, presumably by enhancing the binding of 5-FU to its target enzyme, thymidylate synthase. 5-FU is generally administered intravenously but may also be given orally in the form of capecitabine (Xeloda) with seemingly similar efficacy.

Irinotecan (CPT-11), a topoisomerase 1 inhibitor, has been added to 5-FU and LV (e.g., FOLFIRI) with resultant improvement in response rates and survival of patients with metastatic disease. The *FOLFIRI regimen* is as follows: irinotecan, 180 mg/m^2 as a 90-min infusion on day 1; LV, 400 mg/m^2 as a 2-h infusion during irinotecan administration; immediately followed by 5-FU bolus, 400 mg/m^2, and 46-h continuous infusion of 2.4–3 g/m^2 every 2 weeks. Diarrhea is the major side effect from irinotecan. Oxaliplatin, a platinum analogue, also improves the response rate when added to 5-FU and LV (FOLFOX) as initial treatment of patients with metastatic disease. The *FOLFOX regimen* is as follows: 2-h infusion of LV (400 mg/m^2 per day) followed by a 5-FU bolus (400 mg/m^2 per day) and 22-h infusion (1200 mg/m^2) every 2 weeks, together with oxaliplatin, 85 mg/m^2 as a 2-h infusion on day 1. Oxaliplatin frequently causes a dose-dependent sensory neuropathy that often but not always resolves following the cessation of therapy. FOLFIRI and FOLFOX are equal in efficacy. In metastatic disease, these regimens may produce median survivals of 2 years.

Monoclonal antibodies are also effective in patients with advanced colorectal cancer. Cetuximab (Erbitux) and panitumumab (Vectibix) are directed against the epidermal growth factor receptor (EGFR), a transmembrane glycoprotein involved in signaling pathways affecting growth and proliferation of tumor cells. Both cetuximab and panitumumab, when given alone, have been shown to benefit a small proportion of previously treated patients, and cetuximab appears to have therapeutic synergy with such chemotherapeutic agents as irinotecan, even in patients previously resistant to this drug; this suggests that cetuximab can reverse cellular resistance to cytotoxic chemotherapy. The antibodies are not effective in the ~65% subset of colon tumors that contain mutations in *ras* or *b-raf* genes. The use of both cetuximab and panitumumab can lead to an acne-like rash, with the development and severity of the rash being correlated with the likelihood of antitumor efficacy. Inhibitors of the EGFR tyrosine kinase such as erlotinib (Tarceva) or sunitinib (Sutent) do not appear to be effective in colorectal cancer.

Bevacizumab (Avastin) is a monoclonal antibody directed against the vascular endothelial growth factor (VEGF) and is thought to act as an antiangiogenesis agent. The addition of bevacizumab to irinotecan-containing combinations and to FOLFOX initially appeared to significantly improve the outcome observed with chemotherapy alone, but subsequent studies have suggested a more modest degree of benefit. The use of bevacizumab can lead to hypertension, proteinuria, and an increased likelihood of thromboembolic events.

Preliminary data suggest that the use of checkpoint inhibitors (i.e., PD-1 and PD-2) as immunotherapy is effective in the small subset of patients with metastatic colorectal cancer whose tumors are mismatch repair protein deficient (i.e., microsatellite unstable). Patients with solitary hepatic metastases without clinical or radiographic evidence of additional tumor involvement should be considered for partial liver resection, because such procedures are associated with 5-year survival rates of 25–30% when performed on selected individuals by experienced surgeons.

The administration of 5-FU and LV for 6 months after resection of tumor in patients with stage III disease leads to a 40% decrease in recurrence rates and 30% improvement in survival. The likelihood of recurrence has been further reduced when oxaliplatin has been combined with 5-FU and LV (e.g., FOLFOX), particularly in patients whose tumor has spread to 4 or more regional lymph nodes (N2). Unexpectedly, the addition of irinotecan to 5-FU and LV as well as the addition of either bevacizumab or cetuximab to FOLFOX did not significantly enhance outcome. Patients with stage II tumors do not appear to benefit appreciably from adjuvant therapy, with the use of such treatment generally restricted to those patients having biologic characteristics (e.g., perforated tumors, T4 lesions, lymphovascular invasion) that place them at higher likelihood for recurrence. The addition of oxaliplatin to adjuvant treatment for patients aged >70 and those with stage II disease does not appear to provide any therapeutic benefit.

In rectal cancer, the delivery of preoperative or postoperative combined-modality therapy (5-FU or capecitabine plus radiation therapy) reduces the risk of recurrence and increases the chance of cure for patients with stage II and III tumors, with the preoperative approach being better tolerated.

CANCERS OF THE ANUS

Cancers of the anus account for 1–2% of the malignant tumors of the large bowel. Most such lesions arise in the anal canal, the anatomic area extending from the anorectal ring to a zone approximately halfway between the pectinate (or dentate) line and the anal verge. Carcinomas arising proximal to the pectinate line (i.e., in the transitional zone between the glandular mucosa of the rectum and the squamous epithelium of the distal anus) are known as *basaloid*, *cuboidal*, or *cloacogenic* tumors; about one-third of anal cancers have this histologic pattern. Malignancies arising distal to the pectinate line have squamous histology, ulcerate more frequently, and constitute ~55% of anal cancers. The prognosis for patients with basaloid and squamous cell cancers of the anus is identical when corrected for tumor size and the presence or absence of nodal spread.

The development of anal cancer is associated with infection by human papillomavirus, the same organism etiologically linked to cervical cancer. The virus is sexually transmitted. The infection may lead to anal warts (condyloma acuminata), which may

progress to anal intraepithelial neoplasia and on to squamous cell carcinoma. The risk for anal cancer is increased among homosexual males, presumably related to anal intercourse. Anal cancer risk is increased in both men and women with AIDS, possibly because their immunosuppressed state permits more severe papillomavirus infection. Vaccination against human papilloma viruses appears to reduce the eventual risk for anal cancer. Anal cancers occur most commonly in middle-aged persons and are more frequent in women than men. At diagnosis, patients may experience bleeding, pain, sensation of a perianal mass, and pruritus.

Radical surgery (abdominal-perineal resection with lymph node sampling and a permanent colostomy) was once the treatment of choice for this tumor type. The 5-year survival rate after such a procedure was 55–70% in the absence of spread to regional lymph nodes and <20% if nodal involvement was present. An alternative therapeutic approach combining external beam radiation therapy with concomitant chemotherapy (5-FU and mitomycin C) has resulted in biopsy-proven disappearance of all tumor in >80% of patients whose initial lesion was <3 cm in size. Tumor recurrences develop in <10% of these patients, meaning that ~70% of patients with anal cancers can be cured with nonoperative treatment and without the need for a colostomy. Surgery should be reserved for the minority of individuals who are found to have residual tumor after being managed initially with radiation therapy combined with chemotherapy.

FURTHER READING

Brenner H et al: Colorectal cancer. Lancet 383;1490, 2014. [PubMed: 24225001]

Cao Y et al: Population-wide impact of long-term use of aspirin and the risk of cancer. JAMA Oncol 2;762, 2016. [PubMed: 26940135]

Lieberman D et al: Screening for colorectal cancer and evolving issues for physicians and patients. A review. JAMA 316;2135, 2016. [PubMed: 27893135]

Meyerhardt JA et al: Follow-up care, surveillance protocol, and secondary prevention measures for survivors of colorectal cancer: American Society of Clinical Oncology Clinical Practice Guidelines Endorsement. J Clin Oncol 31:4465, 2013. [PubMed: 24220554]

Petrelli F et al: Prognostic survival associated with left-sided vs right-sided colon cancer. A systemic review and meta-analysis. JAMA Oncol 3:211, 2017.

Shridhar R et al: Anal cancer: Current standards in care and recent changes in practice. CA Cancer J Clin 65:139, 2015. [PubMed: 25582527]

Strum WB: Colorectal adenomas. N Engl J Med 374:1065, 2016. [PubMed: 26981936]

Van Cutsem E et al: Fluorouracil, leucovorin, and irinotecan plus cetuximab treatment and *RAS* mutations in colorectal cancer. J Clin Oncol 33:692, 2015. [PubMed: 25605843]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 78: Tumors of the Liver and Biliary Tree

Josep M. Llovet

INTRODUCTION

The burden of cancer is increasing worldwide. Lung, breast, and colorectal cancers are the most commonly diagnosed while lung and liver cancers are the most common causes of cancer death. Liver cancer is the sixth most common cancer worldwide, the second leading cause of cancer-related deaths and one of the few neoplasms whose incidence and mortality rates have been steadily increasing. Liver cancer comprises a heterogeneous group of malignant tumors with different histologic features and unfavorable prognosis that range from hepatocellular carcinoma (HCC; 85–90% cases), intrahepatic cholangiocarcinoma (iCCA; 10%), and other malignancies accounting for <1% of tumors, such as fibrolamellar HCC, mixed HCC-iCCA, epithelioid hemangiothelioma, and the pediatric cancer hepatoblastoma. The burden of liver cancer is increasing globally in almost all countries, and it is estimated to reach one million cases by 2030.

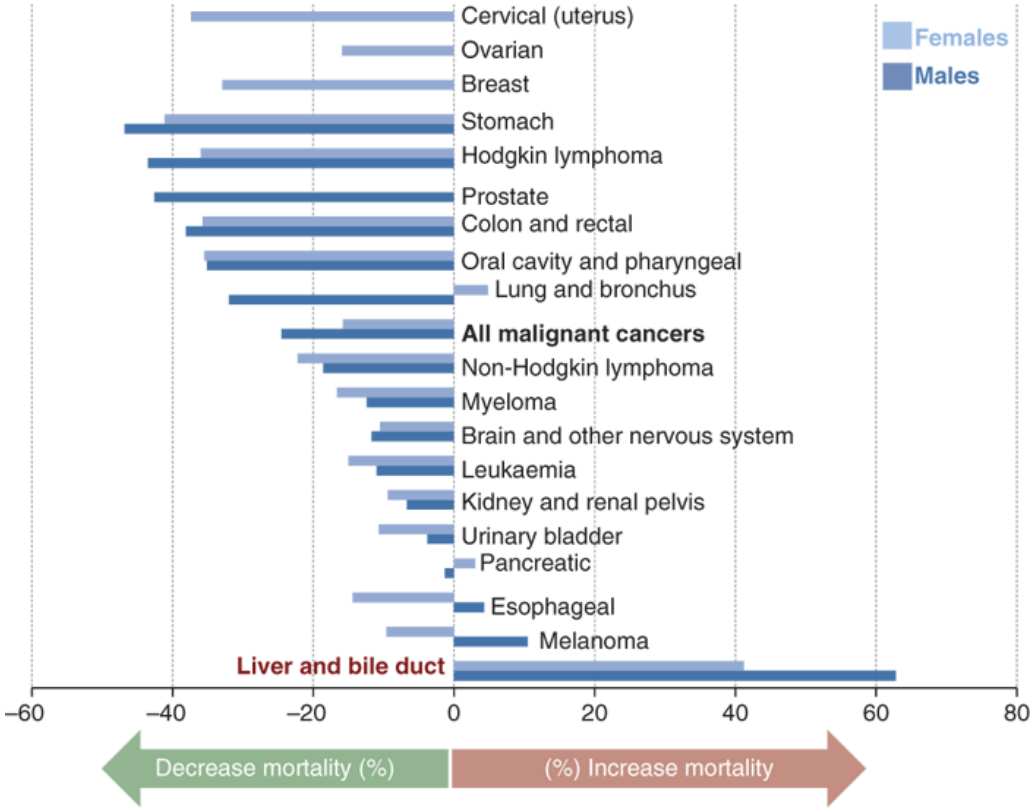
HEPATOCELLULAR CARCINOMA

EPIDEMIOLOGY AND RISK FACTORS

Overall, liver cancer accounts for 7% of all cancers (~850,000 new cases each year), and HCC represents 90% of primary liver cancers. The highest incidence rates of HCC occur in Asia and sub-Saharan Africa due to the high prevalence of hepatitis B virus (HBV) infection, with 20–35 cases per 100,000 inhabitants. Southern Europe, and now North America have intermediate incidence rates (10 cases per 100,000), whereas Northern and Western Europe have low incidence rates of less than 5 cases per 100,000 inhabitants. In the United States, liver cancer is ranked number one in terms of increased mortality during the past two decades **(Fig. 78-1)**, with an incidence of 35,000 cases per year. HCC has a strong male preponderance with a male to female ratio estimated to be 2.5. The incidence increases with age, reaching a peak at 65–70 years old. In Chinese and in black African populations (where vertical transmission of HBV occurs), the mean age is 40–50 years. By contrast, in Japan mean age in men is now around 75 years.

FIGURE 78-1

Mortality trends of patients with different malignancies in the United States between 1990 and 2009. Changes in cancer mortality across tumor types in the United States. Liver and bile duct cancer rank first in terms of increase mortality for both men and women. *(Reprinted with permission from JM Llovet et al: Nat Rev Clin Oncol 12:408, 2015.)*

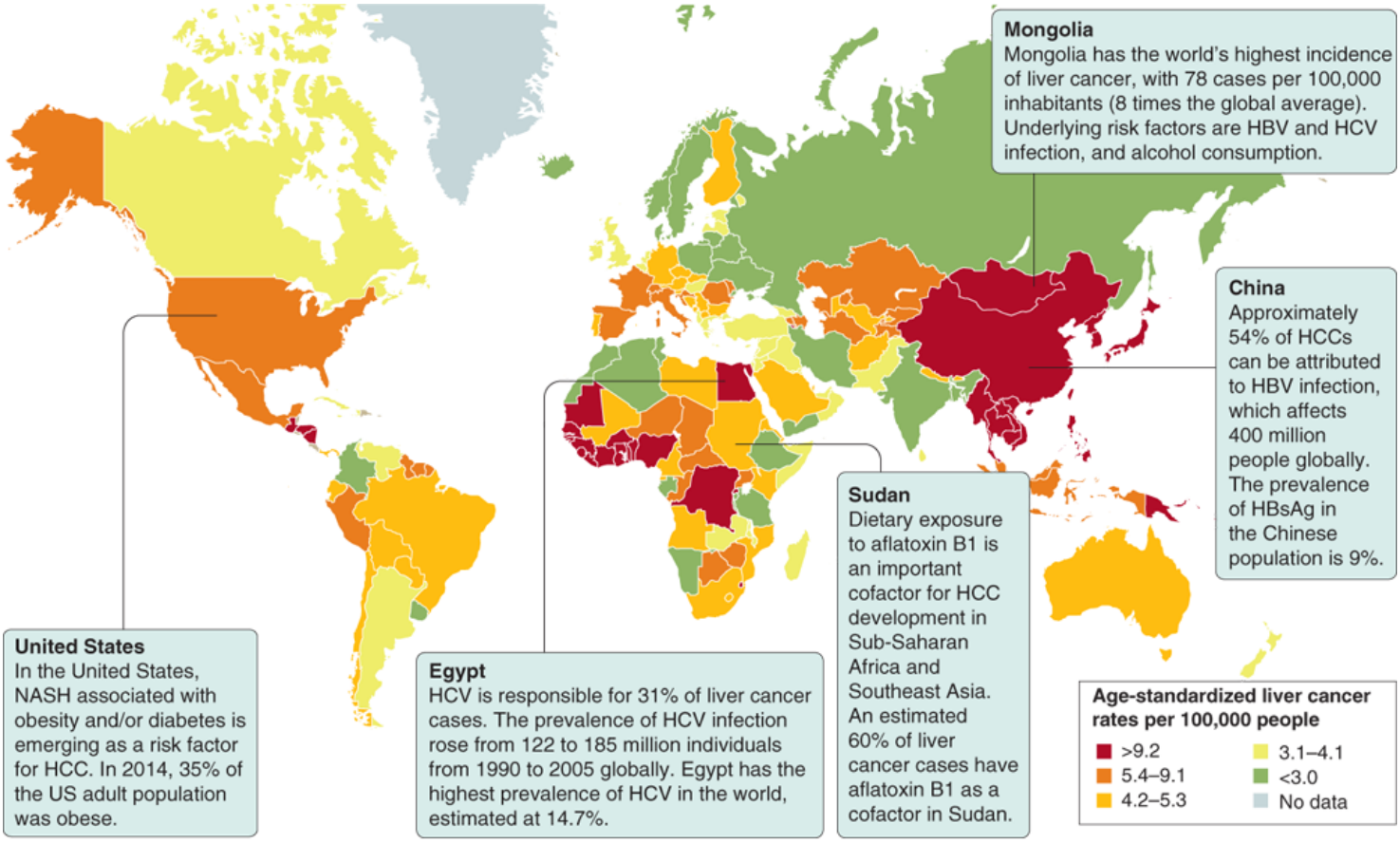


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The risk factors for HCC are well established **(Fig. 78-2)**. The main risk factor is cirrhosis—and associated chronic liver damage caused by inflammation and fibrosis—of any etiology, which underlies 80% of HCC cases worldwide and results from chronic infection by HBV or hepatitis C virus (HCV) infection, alcohol abuse, metabolic syndrome, and hemochromatosis (associated to *HFE1* gene germ-line mutations). Cirrhotic patients represent 1% of the human population and one-third of them will develop HCC during their lifetime. Long-term follow-up studies have established an annual risk of HCC development of 2% in HBV-infected cirrhotic patients and 3–7% in HCV-infected cirrhotic patients. HCC is less common in cirrhosis associated with alpha-1 antitrypsin deficiency, autoimmune hepatitis, Wilson's disease, and cholestatic liver disorders. Predictors of liver cancer development among cirrhotic patients have been associated with liver disease severity (platelet count of <100,000/mm^3, presence of portal hypertension), the degree of liver stiffness as measured by transient elastography, and liver gene signatures capturing the *cancer field effect*.

FIGURE 78-2

The global burden of hepatocellular carcinoma. Incidence of hepatocellular carcinoma (HCC) and risk factors. ASR, Age-standardized rate per 100,000 inhabitants. The main risk factors for HCC development are HBV infection (China), HCV infection (example: Egypt), alcohol intake, non-alcoholic steatohepatitis (United States), and aflatoxin B1 (Sudan). Mongolia has the highest incidence of HCC globally, with 78 cases per 100,000 inhabitants. *(Reprinted with permission from JM Llovet et al: Nat Rev Dis Primers 2:16018, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In terms of attributable risk fraction, HBV infection—a DNA virus that can cause insertional mutagenesis and affects 400 million people globally—accounts for 50% of HCC cases, and is the predominant cause in Asia and Africa. Among patients with HBV infection, a family history of HCC, HBeAg seropositivity, high viral load and genotype C are independent predictors of HCC development. Chronic treatments with effective antiviral HBV therapies are able to significantly decrease the risk of cancer. HCV infection—an RNA virus that affects 170 million people—is responsible for 30% of cases, and is the main cause of HCC in Europe and North America. Among patients with HCV infection, HCC occurs almost exclusively when relevant liver damage is present (either advanced fibrosis—Metavir F3 [Metavir is a scoring system for hepatic histology that grades fibrosis from 0 to 4 with higher numbers indicating more fibrosis]—or cirrhosis), particularly if associated with HCV genotype 1b. In addition, a polymorphism that activates *EGFR*, the EGF receptor, has been established as associated with HCV-HCC in several studies. Antiviral therapies with interferon regimes are able to prevent cirrhosis development and HCC occurrence. The impact of new direct-acting antiviral (DAA) regimes on HCC incidence has not yet been established.

Alcohol consumption and metabolic syndrome due to diabetes and obesity are responsible for ~20% of cases. Non-alcoholic steatohepatitis (NASH), related to metabolic syndrome, is now an emerging cause of HCC in developed countries. A *PNPLA3* polymorphism is strongly associated with fatty and alcoholic chronic liver diseases and HCC occurrence. Other co-factors contributing to HCC development are tobacco and aflatoxin B1, a fungal carcinogen present in food supplies that induces *TP53* mutations. Finally, infection with adeno-associated virus 2 is associated with HCC in individuals without cirrhosis. Aside from the associations described above, genome-wide association studies have not yet confirmed polymorphisms predisposing to HCC development.

There is a growing incidence of HCC world-wide. The growth in U.S. incidence results from the emergence of end-stage liver disease due to hepatitis C, the increase in HBV-related HCC among immigrants from endemic countries, and the accelerating prevalence of obesity and fatty liver disease. In Europe and Asia, the growth has been less prominent. In some countries the incidence is declining, like Italy and Japan, a country where the impact of HCV-related HCC was first noticed after World War II. Finally, the impact of universal infant vaccination against HBV has decreased the rate of HBV-related HCC in endemic countries, such as Taiwan.

MOLECULAR PATHOGENESIS

HCC development is a complex multistep process that starts with pre-cancerous cirrhotic nodules, so-called low-grade dysplastic nodules (LGDN) that evolve to high-grade dysplastic nodules (HGDN) that can transform into early-stage HCC. Molecular studies support the pivotal role of adult

hepatocytes as the cell of origin, either by directly transforming to HCC or by de-differentiating into hepatocyte precursor cells. Alternatively, progenitor cells also give rise to HCC with progenitor markers.

Genomic analysis has provided a clear picture of the main drivers responsible for HCC initiation and progression. This tumor results from the accumulation of around 35–40 somatic genomic alterations per tumor, among which 4–8 are considered driver cancer genes. HCC is a prototypical inflammation-associated cancer, where immune microenvironment and oxidative stress present in chronically damaged livers play pivotal roles in inducing mutations. In pre-neoplastic HGDN, mutations in telomere reverse transcriptase (*TERT*) gene (20% cases) and gains in 8q have been described. Oncogenic transformation occurs upon additional genomic hits including Wnt/β-catenin pathway activation, re-expression of fetal genes, deregulation of protein folding machinery and the response to oxidative stress. Genomic studies and next-generation sequencing conducted during the past decade enables a description of the landscape of mutations, signaling pathways and a molecular classification of the disease. Nonetheless, none of these data have yet translated into actual clinical benefits for molecularly defined tumor subgroups.

Molecular Drivers

The landscape of mutational drivers in HCC identified by deep-genome sequencing of ~1,000 samples is detailed in **Table 78-1**. The most common mutations are in the telomerase reverse transcriptase (*TERT*) promoter (56%), *TP53* (27%), *CTNNB1* (26%), *ARID2* (7%), *ARID1A* (6%), and *AXIN1* (5%) genes. These mutated genes participate in cell-cycle control and senescence -*TERT* or *TP53*-, in cell differentiation—*CTNNB1* and *AXIN-1*, and chromatin remodeling -*ARID2* and *ARID1A*-. Genes commonly mutated in other solid tumors such as *EGFR*, HER2, *PIK3CA, BRAF*, or *KRAS* are rarely mutated in HCC (<5%). Thus, the most prominent drivers are not currently targetable. Some risk factors have been associated with specific molecular aberrations. HBV integrates into the genome of driver genes, such as the *TERT* promoter, *MLL4*, and cyclin E1 (*CCNE1*). HCV infection and alcohol abuse have been significantly associated with *CTNNB1* mutations. *TP53* mutations are the most frequent alterations with a specific hotspot of mutation (R249S) in patients with aflatoxin B1 exposure.

TABLE 78-1

Molecular Aberrations Common in HCC[a]

Pathway	Target	Prevalence (%)
Mutations		
Telomere stability	*TERT* promoter	56
p53/cell cycle control	*TP53* *ATM* *RB1*	27 3 3
Wnt/β-catenin signaling	*CTNNB1* *AXIN1*	26 5
Chromatin remodeling	*ARID1A* *ARID2* *KMT2A* *KMT2C*	6 7 3 3
Ras/PI3K/MTOR pathway	*RPS6KA3* TSC1/TSC2	3 3
Oxidative stress	*NFE2L2* *KEAP1*	3 3
High-Level Focal Amplifications		
VEGF signaling	*VEGFA*	3
FGF signaling	*FGF19*	6
Cell-cycle control	*CCND1*	7
Target with Homozygous Deletion		
TP53/cell-cycle control	*CDKN2A* *TP53* *RB1*	5 4 4
Wnt/β-catenin signaling	*AXIN1*	3

[a]Recurrent mutations, focal amplifications or homozygous deletions in HCC based on next-generation sequencing analyses.

Studies assessing copy-number alterations in HCCs have consistently identified: (i) high level amplifications at 5–10% prevalence containing oncogenes in 11q13 (CCND1 and *FGF19*) and 6p21 (*VEGFA*), *TERT* focal amplification and homozygous deletion of *CDKN2A*; and (ii) common amplifications containing *MYC* (8q gain) and *MET* genes (focal gains 7q31). High-level amplifications of 11q13 include *CCND1* and *FGF19*, which have been demonstrated as prominent oncogenes in HCC and are potential therapeutic targets. Similarly, high-level gains of 6p21 containing more than four copies of *VEGFA* were identified in 4–8% of HCCs. *VEGFA* amplification can induce tumor proliferation by unleashing macrophage-mediated hepatocyte growth factor secretion.

Signaling Pathways

Several signaling pathways have been implicated in HCC progression and dissemination. Activation of these pathways can result from structural alterations (mutations and amplifications/losses), or epigenetic modifications. In brief, (a) TERT overexpression occurs in 90% of cases, particularly

related to promoter *TERT* mutations or amplifications; (b) inactivation of p53 and alterations of cell cycle are major defects in HCC, particularly in cases related to HBV infection; (c) Wnt/β-Catenin pathway activation occurs in 50% of cases, either as a result of β-catenin or *AXIN1* mutation, or overexpression of Frizzled receptors or inactivation of E-cadherin; (d) PI3K/PTEN/Akt/mTOR pathway is activated in 40–50% of HCCs due to mutation and focal deletion of the tuberous sclerosis complex (*TSC1*/*TSC2* genes, *PTEN* or ligand overexpression of EGF or IGF upstream signals; (e) Ras MAPK signaling is activated in half of early and almost all advanced HCCs, activation results from up-stream signaling by EGF, IGF, and MET activation, and from the epigenetic silencing of tumor suppressors such as *NORE1A* and *RASSF1A15*; (f) insulin-like growth factor receptor (IGFR) signaling is activated in 20% of cases through overexpression of the oncogenic ligand IGF2 or allelic loss affecting the tumor suppressor IGF2R; (g) dysregulation of the c-MET receptor and its ligand HGF, critical for hepatocyte regeneration after liver injury, are common events in advanced HCC (50%); (h) vascular endothelial growth factor (VEGF) signaling is the cornerstone of angiogenesis in HCC, along with activated angiogenic pathways such as Ang2 and FGF signaling; and (i) chromatin remodelling complexes and epigenetic regulators are frequently altered in HCC due to *ARID1A* and *ARID2* mutations. Several agents that target these different processes are currently being tested in Phase I-III trials.

Molecular Classes and Prognostic Gene Signatures

Genomic studies have revealed two molecular subclasses of HCC, each representing ~50% of patients. The proliferative subclass is enriched by activation of Ras, mTOR, and insulin-like growth factor (IGF) signaling and *FGF19* amplification, and is associated with HBV-related etiologies, overexpression of α-fetoprotein and poor outcomes (particularly those tumors enriched in progenitor cell markers). By contrast, the so called non-proliferative subclass contains a subtype characterized by *CTNNB1* mutations and better outcome. Another classification based upon immune status has been proposed. It defines an immune HCC class in ~25% of cases characterized by immune infiltrate with expression of PD1/PDL1, enrichment of T-cell activation, and better outcome. Direct translation of molecular subclasses into clinical decision-making is yet to be achieved.

PREVENTION AND EARLY DETECTION

Prevention

Primary prevention of HCC can be achieved by vaccination against HBV and effective treatment of HBV and HCV infection. Studies assessing the impact of universal vaccination against HBV infection started in Taiwan in 1984 have reported a significant decrease of the incidence of HCC. Nowadays, HBV vaccination is recommended to all newborns and high risk groups, following World Health Organization guidelines. Vaccination is also recommended in people with risk factors for acquiring HBV infection, such as health workers, travelers to areas where HBV-infection is prevalent, injecting drug users, and people with multiple sex partners.

Effective antiviral treatments for patients with chronic HBV infection—achieving undetectable viral titres (circulating HBV-DNA)—reduce the risk of HCC development. Evidence of this effect is supported by one randomized trial and several cohort studies. Regarding HCV infection, eradication of hepatitis C results in decreased HCC incidence. Anti-viral therapies achieving a sustained virological response (SVR) in patients with chronic hepatitis prevent the development of advanced stage disease and cirrhosis, hence resulting in a decreased risk of HCC development. However, once cirrhosis is established no high-level evidence suggests that SVR leads to HCC prevention. A meta-analysis of observational studies concluded that interferon-based regimens achieving SVR in patients with cirrhosis were associated with a substantially reduced risk of HCC development. Treatment of HCV has dramatically advanced with the new DAAs (drug antiviral agents) that yield >90% SVR rates after 12 weeks of treatment. A few observational studies with a short follow-up reported an HCC annual incidence of 3–5% in patients with cirrhosis following successful DAA therapy, an incidence similar to that of untreated patients, and higher than those observed with interferon-based therapies. Similarly, there is controversy on the effect of DAA-based SVR on HCC recurrence after curative therapies. Some studies suggest a 6-month recurrence rate higher than historical controls, thus emphasizing the need for large prospective studies. It is too early to estimate the effect of DAA therapy on the burden of HCC. Due to all these circumstances, surveillance remains recommended in patients with cirrhosis achieving SVR.

Additional putative chemopreventive agents have been proposed to reduce HCC incidence in at-risk populations, including statins and metformin. Nonetheless, the evidence is not strong enough to recommend using these therapies in at-risk patients. Finally, coffee consumption is associated with a reduced risk of HCC in population studies.

Surveillance

The aim of surveillance is to obtain a reduction in disease-related mortality. This is usually achieved through early detection that enhances the applicability and cost-effectiveness of curative therapies. United States and European guidelines recommend surveillance for patients at high risk for HCC on the basis of cost-effectiveness analyses. As a general rule, high-risk populations are considered those presenting an incidence cut-off > 1.5% for patients with cirrhosis and 0.2% for patients with chronic hepatitis B. However, the strength of evidence supporting surveillance is modest, and is based upon two randomized studies conducted in China and meta-analysis of observational studies. Overall, these studies conclude that surveillance identifies patients with smaller tumors who are more likely to undergo curative procedures. Because of lead time bias and length time bias it cannot be concluded that surveillance ultimately reduces HCC-related mortality.

Surveillance is recommended for cirrhotic patients owing to any cause, those with HCV-related advanced fibrosis (Metavir score of F3), and for patients with chronic HBV infection if Asian aged >40 years, African aged >20 years or family history of HCC. In terms of liver dysfunction, the presence of advanced cirrhosis (Child-Pugh class C) prevents potentially curative therapies from being employed, and thus surveillance is not recommended. As an exception, patients on the waiting list for liver transplantation, regardless of liver functional status, should be screened for HCC in order to detect tumors exceeding conventional criteria and to define priority policies for transplantation. Complex scoring systems to identify at-risk populations are not yet recommended by guidelines.

Ultrasonography every 6 months is the recommended method of surveillance. It has a sensitivity of 65–80% and a specificity of >90% for early detection. A 3-month interval does not enhance outcomes, and survival is lower with 12 month compared with 6 month intervals. A shorter follow-up interval (every 3–4 months) is recommended when a nodule of <1 cm has been detected. Computed tomography (CT) and magnetic resonance imaging (MRI) are not recommended as screening tools due to lack of data on accuracy, high cost and possible harm (i.e., radiation with CT). Exceptionally, these techniques can be considered in patients with obesity and fatty liver, where visualization with ultrasound is difficult. Accurate tumor biomarkers for early detection need to be developed. Use of alpha-fetoprotein (AFP) levels identifies patients with HCC with 60% sensitivity, but high false-positive results. One main limitation of AFP is that only a small proportion of early tumors (~20%) present with abnormal AFP serum levels. Combining AFP with ultrasound performed by experienced personnel only increase 6–8% the HCC detection rate. Nonetheless, testing AFP is still widely used and this remains an area of controversy. Particularly, testing AFP might be considered in special populations or health care environments when ultrasound is not available. The accuracy of other serum biomarkers proposed, such as des-γ carboxyprothrombin (DCP) and the L3 fraction of AFP (AFP-L3), in early detection is not known.

Despite the fact that surveillance is cost-effective in HCC, the global implementation of such programs is estimated to engage ~50% of the target population in Europe and ~30% in the United States. Public health policies encouraging the implementation of such programs should lead to an increase in early tumor detection.

Diagnosis

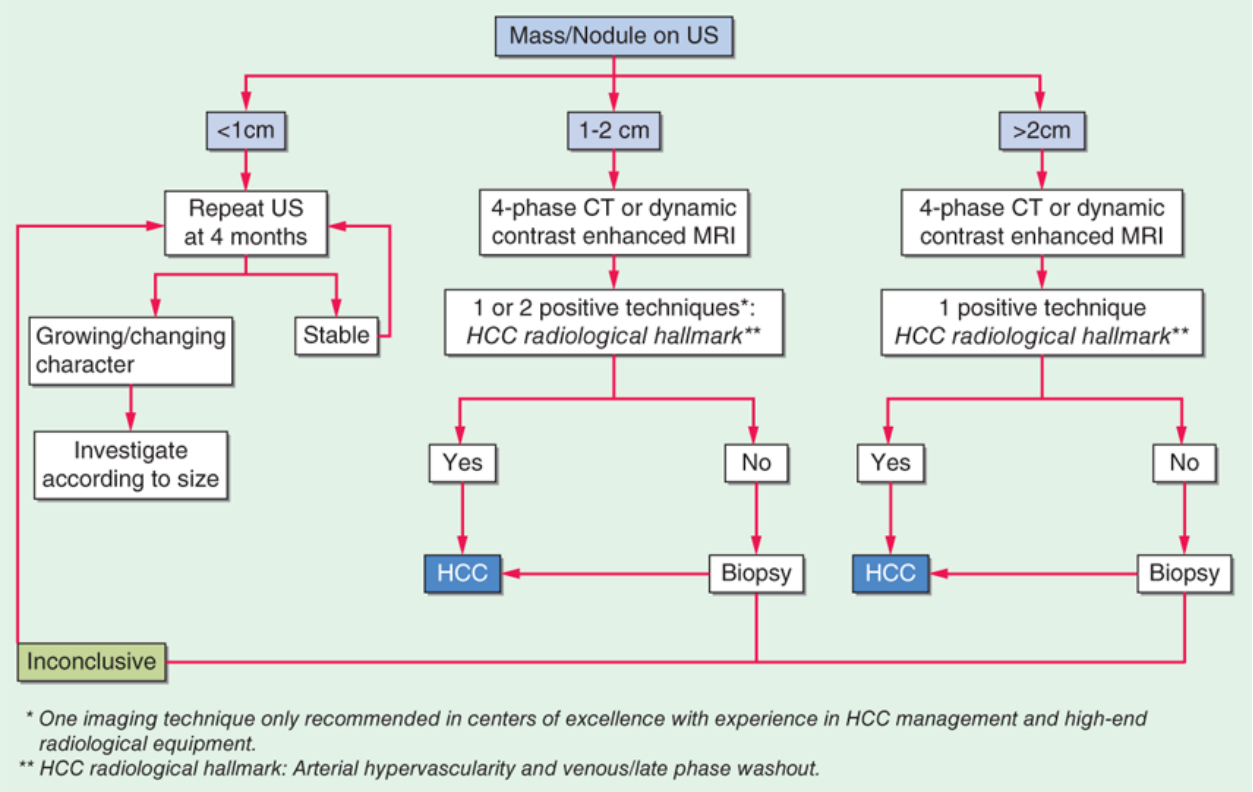
HCC is generally diagnosed at early or intermediate stages in Western countries, but at advanced stages in most Asian (except Japan) and African countries. A surveillance program yields early diagnosis in 70–80% of cases. At these stages the tumor is asymptomatic, and diagnosis can be made by non-invasive (radiological) or invasive (biopsy) approaches. Without surveillance, HCC is discovered either as a radiological finding or due to cancer-related symptoms. If symptoms are present the disease is already at an advanced stage with a median life expectancy <1 year. Symptoms include malaise, weight loss, anorexia, abdominal discomfort, or signs related to advanced liver dysfunction.

NON-INVASIVE (RADIOLOGICAL) DIAGNOSIS

Patients enrolled in a surveillance program are diagnosed by identification of a new liver nodule on abdominal ultrasound. Non-invasive criteria can only be applied to cirrhotic patients and are based on imaging techniques obtained by 4-phase multidetector CT scan (four phases are unenhanced, arterial, venous, and delayed) or dynamic contrast-enhanced MRI. A flow-chart of diagnosis and recall policy recommended by U.S. and European guidelines is summarized in **Fig. 78-3**. Radiological diagnosis is achieved with a high degree of confidence if the lesion is ≥2 cm in diameter and shows the *radiological hallmarks of HCC* by one imaging technique. Using contrast-enhanced imaging techniques, the typical hallmark of HCC consists of vascular uptake of the nodule in the arterial phase with washout in the portal venous or delayed phases. This radiological pattern captures the hypervascular nature characteristic of HCC. In these scenarios the diagnostic specificity is ~95–100% and a biopsy is not necessary. For lesions 1–2 cm in diameter, the radiological hallmarks of HCC define diagnosis, but need to be confirmed by two imaging techniques in non-specialized centers. Nodules <1 cm in size are unlikely to be HCC and would be very difficult to diagnose, and thus ultrasound follow-up at 3–4 months is recommended. MRI with liver specific contrast agents might help in the diagnosis of HCC, but the specificity of these agents is still suboptimal. Contrast-enhanced ultrasound (CEUS) and angiography are less accurate for HCC diagnosis. Positron emission tomography (PET)-scan performs poorly for early diagnosis. AFP levels ≥ 400 ng/dL are highly suspicious, but not diagnostic of HCC according to guidelines.

FIGURE 78-3

Recall diagnosis schedule for HCC (EASL). EASL, European Association for the Study of Liver Disease; HCC, hepatocellular carcinoma. *(Reprinted with permission from EASL-EORTC guidelines. J Hepatology 56: 908, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOLOGICAL DIAGNOSIS

Pathological diagnosis is required in two scenarios: (a) in patients without cirrhosis, and (b) if radiology is not typical in at least one of two imaging techniques (CT and MRI). This occurs mainly with early-stage HCC lesions. Biopsy is not an ideal gold standard, because of variation introduced by sampling and complications. Sensitivity of liver biopsies ranges between 70 and 90% for all tumor sizes, but decrease to <50% in tumors 1–2 cm in size. The risk of complications such as tumor seeding and bleeding after liver biopsy is ~3%. Biopsies should be assessed by an expert hepatopathologist. The use of special stains may help to resolve diagnostic uncertainties. Positive staining in two of four markers [glypican 3 (GPC3), glutamine synthetase, heat shock protein 70 (HSP70), and clathrin heavy chain] is highly specific for HCC. Gene expression blueprints (glypican 3, LYVE1, and survivin) are also able to differentiate high grade dysplastic nodules from early HCC. Additional staining can be considered to detect progenitor cell features (K19 and EpCAM) or assess neovascularization (CD34). A negative biopsy does not eliminate the diagnosis of HCC. A second biopsy is recommended in case of inconclusive findings, or growth or change in enhancement pattern identified during follow up. European guidelines advocate obtaining tissue samples in the setting of all research studies in HCC, even if radiological criteria are met.

TREATMENT

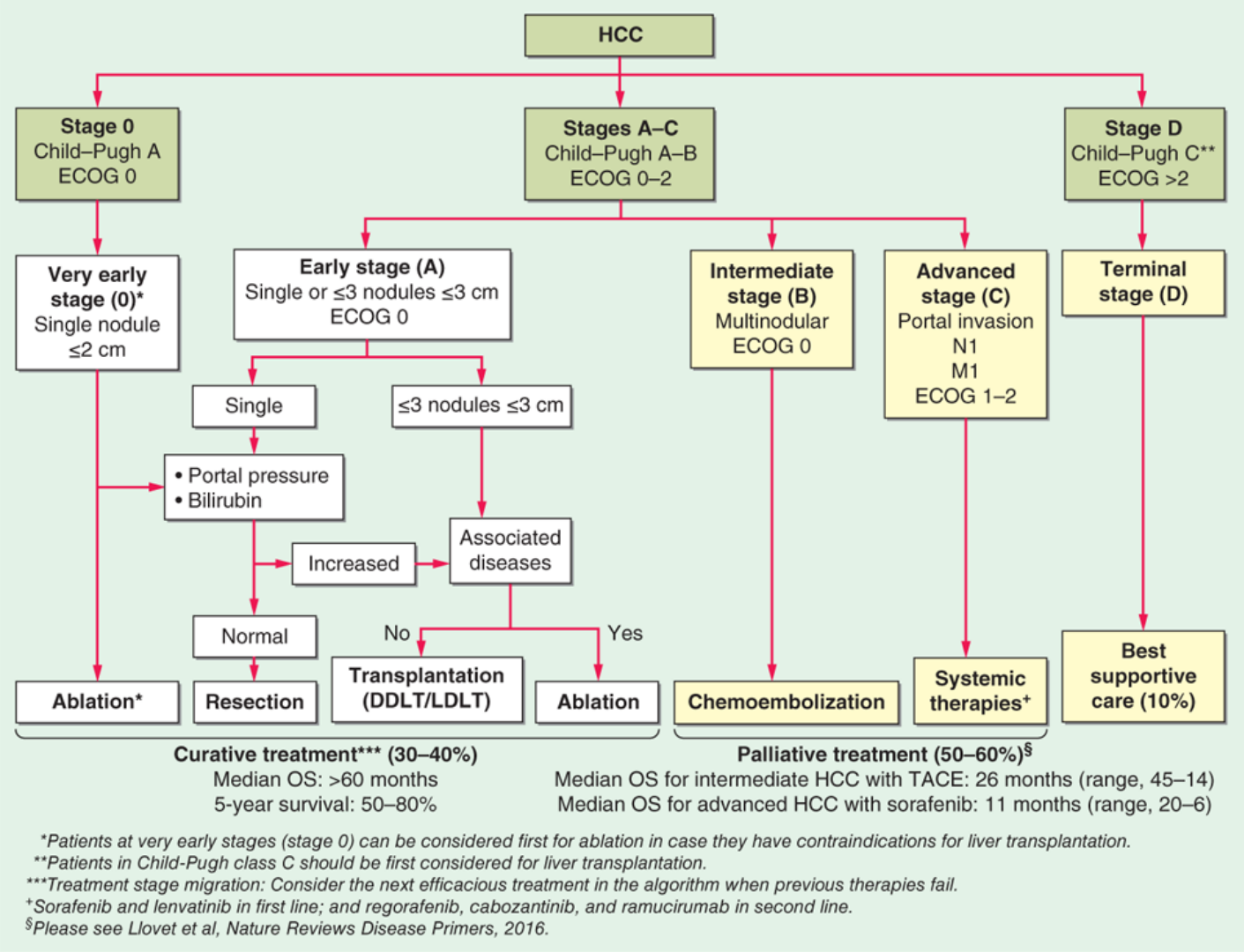
Staging Systems and Treatment Allocation

Staging systems are aimed at stratifying patients according to prognostic factors and outcome, and to allocate the best available therapies according to evidence. The most accepted staging system is the Barcelona-Clinic-Liver Cancer (BCLC) Classification, which is endorsed by U.S. and European clinical practice guidelines **(Fig. 78-4)**. This staging system defines five prognostic subclasses and allocates specific treatments for each stage. The BCLC staging system has been externally validated by numerous studies. It is an evolving system that allows incorporation of new therapies and treatment-dependent variables as new evidence emerges. Six treatments have been demonstrated to improve survival in HCC, five are adopted by guidelines and BCLC classification: surgical resection, liver transplantation, radiofrequency (RF) ablation, chemoembolization, and systemic therapies (sorafenib, regorafenib, lenvatinib, cabozantinib, ramucirumab). The BCLC system will also incorporate lenvatinib in first line and regorafenib as standard of care in patients with advanced HCC progressing on sorafenib as a consequence of a positive randomized controlled trial (RCT). The BCLC assigns each patient with a given treatment allocation. Treatment stage migration is also applied by this scheme, meaning that if patients are not candidates for the selected therapy, the next effective therapy at more advanced stages can be given.

FIGURE 78-4

BCLC staging system and therapeutic strategy. BCLC classification comprises five stages that select the best candidates for therapies according to evidence-based data. Patients with asymptomatic early tumors (stages 0–A) are candidates for radical therapies (resection, transplantation, or local ablation). Asymptomatic patients with multinodular HCC (stage B) are suitable for transcatheter arterial chemoembolization (TACE), whereas patients with advanced symptomatic tumors and/or an invasive tumoral pattern (stage C) are candidates to receive sorafenib. End-stage disease (stage D)

includes patients with poor prognosis that should be treated by best supportive care. BCLC, Barcelona Clinic Liver Cancer; DDLT, deceased donor liver transplantation; EASL, European Association for the Study of Liver Disease; ECOG, Eastern Cooperative Oncology Group Performance Status; EORTC, European Organisation for Research and Treatment of Cancer; GRADE, grading of recommendations assessment, development, and evaluation; HCC, hepatocellular carcinoma; LDLT, living donor liver transplantation; OS, overall survival; PEI, percutaneous ethanol injection; RF, radiofrequency ablation; TACE, transcatheter arterial chemoembolization. *(Reprinted with permission from JM Llovet et al: Nat Rev Dis Primers 2:16018, 2016; A Forner, JM Llovet, J Bruix: Hepatocellular carcinoma. Lancet 379:1245, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In HCC, three parameters are relevant for defining treatment strategy: tumor status, cancer-related symptoms, and liver dysfunction. The BCLC staging captures all three variables and allocates patients to treatments according to evidence. Since >80% of patients have two diseases, HCC and cirrhosis, a clear measurement of liver dysfunction should be in place. The prognosis of chronic liver disease is commonly assessed using the Child-Pugh score, which uses five clinical measures—total bilirubin, serum albumin, prothrombin time, ascites severity, and hepatic encephalopathy grade—to classify patients into one of three groups (A–C) of predicted survival rates. In brief, Child-Pugh's A reflects well-preserved liver function, Child's B moderate liver dysfunction with a median life expectancy of ~3 years and Child C severe liver dysfunction with life expectancy of ~1 year. At early BCLC stages more granular criteria to define patients with very-well preserved liver function (Child-Pugh's hyper-A class without portal hypertension) needs to be in place to select candidates for resection. Modifications of Child-Pugh scoring or model for end-stage liver disease (MELD) score have not been adopted for treatment allocation, except for prioritization on the waiting list for liver transplantation (MELD score). More sophisticated measures of liver dysfunction (i.e., assessment of portal hypertension) are recommended for preoperative assessment of candidates for resection. Performance status is assessed by Eastern Cooperative Oncology Group (ECOG) and presence of cancer-related symptoms (ECOG 1-2) is considered a sign of advanced stage. Patients with severe liver dysfunction (Child-Pugh's C class) or performance status impairment (ECOG 3-4) are offered supportive care management.

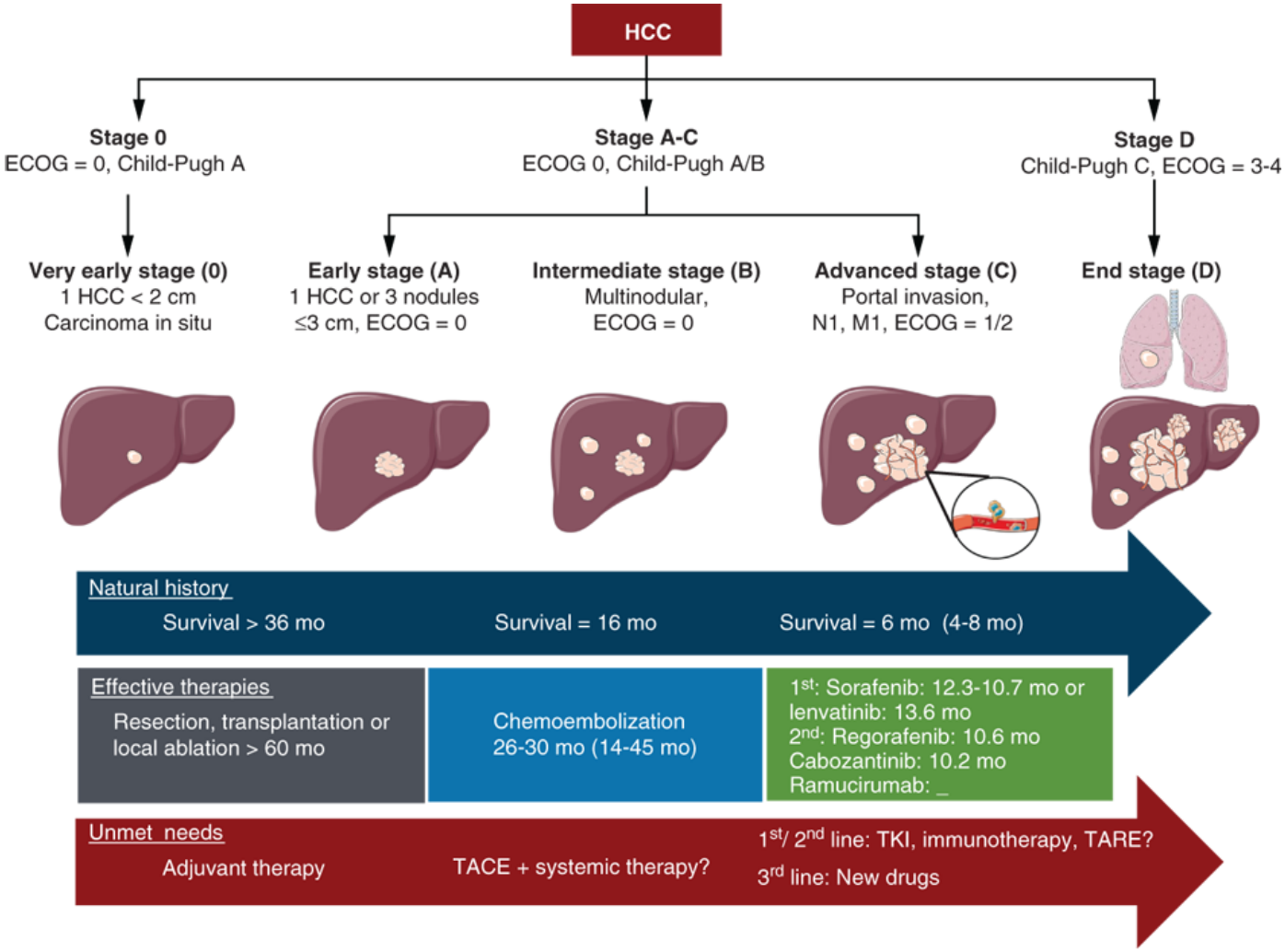
Considering all these prognostic/predictive variables and evidence-based treatment efficacy, five BCLC stages have been defined (Fig. 78-4). Patients with liver only neoplastic disease, no symptoms (ECOG 0) and with mild to moderate liver dysfunction (Child-Pugh A-B) can be classified as very early

(Stage 0) or early (Stage A) or intermediate (stage B) stages depending upon tumor size and number. Very early HCC (BCLC 0) is defined by single tumors ≤2 cm (if pathology is available they should be well-differentiated with absence of microvascular invasion or satellites). Early HCC (BCLC A) includes either single tumors or a maximum of three nodules of ≤3 cm in diameter. Intermediate stage (BCLC B) is defined by all other liver-only tumors. Conversely, HCC is considered at advanced stages (BCLC C) when patients present cancer-related symptoms (ECOG 1-2) or tumors with macrovascular invasion (of any type, including branch, hepatic, or portal vein), lymph node involvement, or extrahepatic spread. Finally, end-stage disease (BCLC D) is considered in cases of several impairment of quality of life/cancer-related symptoms (ECOG 3-4) or severe liver dysfunction (Child-Pugh C).

Around 40% of patients are diagnosed at Stages 0 and A, and hence are eligible for potentially curative therapies, resection, transplantation, or local ablation. These treatments provide median survival rates of 60 months and beyond, which are in sharp contrast with outcomes of 36 months reported in historical controls **(Fig. 78-5)**. No adjuvant therapy is recommended. Patients at intermediate stage (Stage B) with preserved liver function have a documented natural history of around 16 months. These patients benefit from transarterial chemoembolization (TACE) as reported in two randomized studies and one meta-analysis, and achieve an estimated survival of 25–30 months. None of the combination therapies with TACE have shown outcome advantages. Patients progressing on TACE or at advanced stage (Stage C) benefit from systemic sorafenib, which extends survival by ~3 months (from 7.9 to 10.7 months). Lenvatinib showed non inferiority results compared to sorafenib (13.6 months vs. 12.3 month, respectively). Regorafenib improves survival from 7.8 to 10.6 months in patients progressing on sorafenib (second-line advanced HCC). Therefore, these treatments have been adopted by guidelines and incorporated to the BCLC classification. Patients with end-stage disease (BCLC D) should be considered for nutritional and psychological support and proper management of pain.

FIGURE 78-5

Natural history, impact of therapies, and unmet needs in HCC.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although the BCLC establishes validated stages and treatment assignment according to evidence, clinical practice is not always aligned with this classification. In large cohort studies and surveys, only half of patients, or even less in Asia, are treated accordingly. Alternative staging or scoring systems have been proposed, but none of them has acquired global consensus. In contrast to BCLC, some proposed systems capture the standard of

practice in Asia, such as the Hong Kong classification or the Japan Integrated Staging score. These systems capture extended indications for resection and TACE applied in clinical practice in Asia. Other systems define prognostic stratification, such as the Cancer of the Liver Italian Program (CLIP) score, although they do not incorporate treatment allocation to distinct stages. Finally, the tumor-node-metastasis (TNM) staging system is not used in HCC since it does not incorporate the main prognostic variables related to liver function and performance status.

The BCLC system does not incorporate molecular classes or biomarkers. Some biomarkers (i.e., AFP at a cut-off of >200 or 400 ng/mL) or molecular classes/signatures have prognostic or biological significance. However, they are not ready for clinical application due to the lack of data on biomarker-based response to therapies. A few Phase III studies are currently conducted based upon biomarker-enriched populations (i.e., ramucirumab in AFP >400 ng/mL) or as proof-of-concept studies (i.e., FGFR4 inhibitors in patients with FGF19 amplification/overexpression). Ramucirumab trial has been positive, and thus biomarkers will be incorporated in the treatment allocation system.

Due to the complexities of HCC diagnosis and management, it is recommended to send patients to a referral center where all the armamentarium of therapies can be offered. In principle, patient management and outcome benefit from liver cancer multidisciplinary programs that include hepatologist, oncologist, hepatobiliary and transplant surgeons, interventional and body imaging radiologist, hepatopathologist, and specialized nurses.

SURGICAL THERAPIES

Resection

Surgical resection is the first-line option for non-cirrhotic patients at early-stage HCC (BCLC 0 or A) with solitary tumors (Fig. 78-4). In cirrhotic patients, ablation competes with resection for BCLC 0 tumors (<2 cm in diameter). Which is better is not defined. Cost-effectiveness approaches report a benefit for local ablation with RF. For single tumors >2 cm (BCLC A), resection remains the mainstay of treatment in patients with Child-Pugh's hyper-A class, those patients with normal bilirubin and absence of portal hypertension (portal hypertension is defined by hepatic venous pressure gradient ≥10 mmHg). Surrogate measures of portal hypertension are presence of esophageal varices or platelet count <100,000/mm^3 associated with splenomegaly. Anatomic resections following the functional segments of the liver are recommended to spare uninvolved liver parenchyma and to remove satellite tumors. Predictors of recurrence are tumor size, number, presence of microsatellites, or microvascular invasion at the specimen analysis. Macrovascular invasion, extrahepatic involvement, and liver dysfunction (Child-Pugh B-C) are major contraindications for resection.

ADJUVANT TREATMENTS

Tumor recurrence represents the major complication of resection (and local ablation) and occurs in 70% of cases at 5 years. Most of recurrences are intrahepatic metastases, but at least one third are considered *de novo* tumor, new clones developing in the cirrhotic carcinogenic field. The type of recurrence can only be defined by molecular studies. So far, no adjuvant therapies have proven to improve outcome or prevent recurrence after resection/ablation. Randomized trials testing adjuvant sorafenib, retinoids, chemotherapies or chemoembolization have been negative. Some trials testing adoptive immunotherapy or interferon showed positive results, but these treatments have not been adopted by guidelines of management due to weakness of the evidence or small magnitude of benefit. Therefore, the current recommendations are that in case of recurrence patients will be re-assessed by BCLC staging, and re-treated accordingly.

Liver Transplantation

Liver transplantation is the first treatment choice for cirrhotic patients with single tumors ≤5 cm and portal hypertension (including Child-Pugh's B and C) or with small multinodular tumors (≤3 nodules, each ≤3 cm) (Fig. 78-4). These so-called Milan criteria have been validated over the years and a meta-analysis reported 5-year and 10-year survival rates of ~70 and ~50%, respectively, similar to outcomes achieved in non-HCC transplantation indications. Perioperative mortality rates have been reduced to <3%. Transplantation simultaneously cures the tumor and the underlying cirrhosis, and it is associated with a low risk of recurrence, around 10–15% at 5 years. No immunosuppressive regimens or anti-tumor therapies after transplantation have demonstrated any preventive effect on recurrence. Milan criteria are integrated in the BCLC treatment strategy (BCLC 0 and A) and have also been adopted by the United Network for Organ Sharing (UNOS) pre-transplant staging for organ allocation in the United States (Stage T2). Aside from size and number, conventional contraindications for organ transplantation procedures (ABO incompatibility, co-morbidities, etc.) are applied in this setting.

Liver transplantation has a couple of important limitations, such as cost and donor availability, which limit this procedure to <5% of HCC cases. The scarcity of donors represents a major drawback of liver transplantation. Donor scarcity varies geographically, and deceased liver donation is almost zero in some Asian countries. Due to the shortage of donors, median waiting times in Western programs is ~6–12 months leading to 20% of candidates dropping off the list due to tumor progression before receiving the procedure. Predictors of drop-out are treatment failure, baseline AFP >400 ng/mL or steady increase of AFP level >15 ng/mL per month. Several strategies have been proposed to overcome this limitation. First, apply neo-adjuvant therapies in patients on the waiting list. Neo-adjuvant treatments testing TACE or RF ablation have been assessed in the setting of cohort and cost-effectiveness studies. In principle, the use of these therapies is recommended when the waiting time exceeds 6 months, even though impact on long-

term outcome is uncertain. Second, a priority policy has been established for patients enlisted. UNOS has implemented a scoring system based upon the dropout risk, giving priority to tumors 2–5 cm in size and multinodular tumors.

The Milan criteria are universally used as the basis for transplant eligibility, and adherence to them yields good post-transplant survival. Modest expansion of Milan criteria applying the "up-to-seven" criteria (i.e., those HCCs having the number 7 as the sum of the size of the largest tumor and the number of tumors) in patients without microvascular invasion achieves competitive outcomes. These pathologically-defined criteria are being used in clinical practice to predict the expected outcome after transplantation. Similarly, *down-staging to Milan criteria* is currently explored by several groups. Down-staging is defined as the reduction of HCC burden by loco-regional treatments to achieve Milan staging before transplantation. A few studies claim that down-staging lasting for >3 months achieves competitive outcomes, but robust long-term survival data is scarce, and thus it cannot yet be recommended. Down-staging policy is only endorsed by guidelines for patients outgrowing the Milan criteria while on the waiting list.

Since policies for enhancing organ donation have reached a ceiling during the past several years, alternatives to donation have emerged. Living donor liver transplantation represents a plausible alternative that accounts of ~5% of total transplantations performed globally. Outcomes reported are similar to those with deceased liver donors, and it is recommended as an alternative option in patients on a waiting list exceeding 6–7 months. The risks and benefits of this procedure should take into account both donor (death is estimated in 0.3%) and recipient, a concept known as *double equipoise*. Due to the complexity of this treatment, it must be restricted to centers of excellence in hepatobiliary surgery and transplantation.

LOCO-REGIONAL THERAPIES

Local Ablation

RF ablation is recommended as the primary ablative technique (Fig. 78-4). The energy generated by RF ablation (heating of tissue at 80°–100°C) induces coagulative necrosis of the tumor producing a *safety ring* in the peritumoral tissue, which might eliminate small-undetected satellites. Treatment consists of 1 or 2 sessions performed using a percutaneous approach, although in some instances ablation with laparoscopy is needed. RF ablation is more effective in response rate and time-to-recurrence compared with the once-conventional percutaneous ethanol injection. Long-term outcome of HCC patients treated by RF ablation have 5-year survival rates of ~60%. In tumors <2 cm, BCLC 0, RF ablation achieves complete responses in >90% of cases with good long-term outcome and is competitive with resection in terms of cost-effectiveness. For BCLC A cases, RF ablation is the first-line treatment for single tumors 2–5 cm or multinodular up to three nodules, each ≤3 cm in diameter, unsuitable for surgery.

The main limitation of RF ablation is that its failure rate increases in tumors >3 cm because of the heat loss due to perfusion-mediated tissue cooling within the area ablated. In tumors 3–5 cm in diameter, complete pathological tumor necrosis of <50% has been reported. Particularly, ~10–15% of tumors with difficult-to-treat locations, such as a subcapsular location or adjacent to the gallbladder, have a higher risk of incomplete ablation or major complications and can be approached by ethanol injection. Several approaches have been proposed to enhance the anti-tumor activity of RF ablation. The combination of RF ablation with either chemoembolization or with a heat-activated formulation of liposomal doxorubicin yielded good results in cohort studies. Other treatments, such as microwave ablation, high-intensity focused ultrasound or stereotactic body radiotherapy for small tumors are under investigation.

Chemoembolization

TACE is the most widely used primary treatment for unresectable HCC worldwide, and the first-line indication for patients with intermediate BCLC B stage (Fig. 78-4). Conventional chemoembolization (c-TACE) consists of the local hepatic artery administration of chemotherapy (either doxorubicin 50 mg/m^2 or cisplatin) mixed with an emulsion of lipiodol followed by obstruction of the feeding artery with sponge particles. c-TACE mainly benefits patients with liver-only disease, Child-Pugh A Class, or B without ascites, good performance status (ECOG 0), and absence of branch or trunk vascular invasion. Median survival is ~20 months (compared to 16 months for pooled control arms). The best randomized trial and subsequent Phase II studies have provided median survivals for TACE of 25–30 months in properly selected populations. Median objective response rates are of 50–70%. In randomized studies, the treatment is performed at a regular schedule of 0, 2, and 6 months (median number of sessions: 3), although no consensus has been established. TACE procedures should be stopped upon tumor progression or any other contraindication. Exceptionally, occurrence of a new small untreated nodule as the only progression feature can be considered for treatment. Around ~50% of patients present a limited postembolization syndrome of fever and abdominal pain related to ischemic injury and release of cytokines. Less than 5% of patients present major complications (liver abscess, ischemic cholecystitis, or liver failure) and in <2% of cases treatment-related death occurs.

Applicability of c-TACE in BCLC B patients is limited to half of cases, mostly as a result of the presence of liver failure (Child B, or ascites or encephalopathy), technical contraindications to the procedure (i.e., impaired portal-vein blood flow), or infiltrative/massive tumor burden (i.e., generally main tumor size >10 cm). Super-selective TACE minimize the ischemic insult to non-tumor tissue. According to guidelines, treatment-stage migration allows performing TACE on patients at early stages not suitable for surgical or ablative therapies. In selective studies, median survival rates of 5 year have been reported in patients with single HCC treated by supra-selective TACE. On the other hand, TACE performed beyond guidelines as a conventional practice to patients with formal contraindications (generally BCLC C) yields poor outcomes.

Drug-eluting beads chemoembolization (DEB-TACE) differs from c-TACE in the use of more standardized embolic spheres of regular size embedded with chemotherapy. This strategy ensures drug release over a 1-week period resulting in an enhancement of drug concentration within the tumor. DEB-TACE achieves similar anti-tumor activity (objective responses of ~60%) as c-TACE associated with significantly less systemic cytotoxic effects and better tolerance, but with no clear differences in clinical outcomes. Phase II and III studies have compared DEB-TACE with the combination of DEB-TACE with sorafenib or brivanib, a VEGF receptor inhibitor. Median survival in both arms of these international trials was 25–30 months.

Radioembolization and Other Intraarterial Therapies

Radioembolization using beads coated with yttrium-90 (Y-90)—an isotope that emits short-range β radiation—is the most promising alternative to TACE. Several Phase II studies reported objective responses and overall outcome with a safe profile. Due to the lack of Phase III trials, this treatment is currently not recommended in guidelines. Whether radioembolization might be effective in patients at an intermediate-stage not eligible for TACE needs to be studied. Radioembolization requires prevention of severe lung shunting and intestinal radiation before the procedure. Around 20% of patients present liver-related toxicity and 3% treatment-related death. Due to the minimally embolic effect of Y-90 microspheres, treatment can be safely used in patients with portal vein thrombosis, a setting where survival results in Phase II were encouraging and Phase III investigations in combination with sorafenib are ongoing. Head to head comparison of Y-90 vs sorafenib did not hit the primary end-point of overall survival.

TACE should be distinguished from other intraarterial therapies, such as chemo-lipiodolization, which involves the delivery of an emulsion of chemotherapy mixed with lipiodol, bland transcatheter embolization (TAE), where no chemotherapeutic agent is delivered, and intra-arterial chemotherapy, where no embolization is performed. None of these approaches is recommended due to the lack of survival benefit.

SYSTEMIC THERAPIES

Conventional systemic chemotherapy and radiotherapy have not produced survival advantages. Randomized studies also failed with anti-estrogen therapies and vitamin D derivatives. External beam liver-directed radiotherapy (stereotactic body radiotherapy) efficacy is currently being tested with and without sorafenib in Phase III trials. In 2007 a Phase III trial demonstrated survival benefits for patients with advanced stage disease treated with sorafenib, and more recently lenvatinib showed similar effects to sorafenib in first line treatment. A second multikinase inhibitor, regorafenib, has been shown to benefit patients progressing to sorafenib.

Molecular Targeted Therapies

Sorafenib is the standard of care systemic therapy for HCC. It is indicated for patients with well-preserved liver function (Child-Pugh A class) and with advanced tumors (i.e., BCLC C) or those tumors at intermediate stage (i.e., BCLC B) progressing upon loco-regional therapies (Fig. 78-4). A Phase III study comparing sorafenib vs placebo showed increased survival from 7.9 months to 10.7 months (HR 0.69; 31% reduction of risk of death). In this trial, 80% of patients were BCLC C and 20% BCLC B progressed to TACE. Overall, 35% of patients presented with macrovascular invasion and 50% with extrahepatic spread. A similar magnitude of benefit was observed in another positive Phase III study conducted in parallel in Asian patients, mostly with HBV-related HCC. Interestingly, objective responses account for 2% of patients assessed by RECIST criteria and ~10% assessed by the more refined modified RECIST (mRECIST) criteria. Patients with HCV-related HCC achieve significantly better outcomes with sorafenib, with a median survival of 14 months. No predictive biomarkers of responsiveness to sorafenib have been identified.

The recommended daily dose of sorafenib is 800 mg. Median treatment duration is about six months. Treatment is associated with manageable adverse events, such as diarrhea, hand-foot skin reactions, fatigue, and hypertension. Treatment-related liver failure or life-threatening complications are unusual. These toxicities lead to treatment discontinuation in 20% of patients and dose-reduction in up to half. Not all patients at advanced stages can receive sorafenib. It has been estimated that this therapy cannot be administered to around one-third of the targeted patients due to primary intolerance, advanced age, or liver failure (ascites or encephalopathy). Active vascular disease, either coronary or peripheral, is considered a formal contraindication.

The efficacy of sorafenib probably results from a balance between targeting cancer cells and the microenvironment by blocking up to 40 kinases, including anti-angiogenic (vascular endothelial growth factor receptor [VEGFR], platelet-derived growth factor receptor [PDGFR]), and anti-proliferative drivers (serine/threonine-protein kinase B-raf [BRAF] and mast/stem cell growth factor receptor [c-Kit]). Median time to progression on sorafenib is of 4–5 months in Phase III trials. Activation of MAPK14 signaling, IGF signaling, and enrichment in tumor-initiating cells is the main mechanisms of acquired resistance.

Several other agents have been tested with negative results in most of the cases **(Table 78-2)**. Recently, a phase III study comparing lenvatinib (an inhibitor of VEGFR, fibroblast growth factor recepctor [FGFR], PDGFR, RET, and c-Kit) with sorafenib has shown non-inferiority results in terms of overall survival (HR = 0.92). It is estimated that only half of the patients progressing on sorafenib can be considered for second-line therapies, and their median survival with no treatment is 7–8 months (obtained from patients allocated to the placebo arm).

TABLE 78-2

Phase III Trials Testing Molecular Therapies in Advanced HCC the Past 10 Years

	Drugs**	n	Median OS (month)	Hazard Ratio (p-value)	Median TTP (month)	Hazard Ratio (p-value)	Objective Response (%)
First-Line							
SHARP	Sorafenib	299	10.7	0.69	5.5	0.58	2.3
	Placebo	303	7.9	(<0.001)	2.8	(<0.001)	0.7
Asian-Pacific	Sorafenib	150	6.5	0.68	2.8	0.57	3.3
	Placebo	76	4.2	(0.01)	1.4	(<0.001)	1.3
Sunitinib	Sunitinib	530	7.9	1.3	4.1	1.13	6.6
	Sorafenib	544	10.2	(0.001)	3.8	(0.308)	6.1
BRISK-FL	Brivanib	577	9.5	1.06	4.2	1.01	12*
	Sorafenib	578	9.9	(0.31)	4.1	(0.853)	8.8*
LIGHT	Linifanib	514	9.1	1.04	5.4	0.76	10.1
	Sorafenib	521	9.8	(0.52)	4	(0.001)	6.1
SEARCH	Sorafenib + Erlotinib	362	9.5	0.92	3.2	1.13	6.6
	Sorafenib	358	8.5	(0.2)	4	(0.18)	3.9
Lenvatinib	Lenvatinib	478	13.6	0.92	8.9	0.63	24*
	Sorafenib	476	12.3	(--)	3.7	(<0.001)	9*
Second-Line							
BRISK-PS	Brivanib	263	9.4	0.89	4.2	0.56	9.9*
	Placebo	132	8.2	(0.33)	2.7	(<0.001)	1.5*
EVOLVE-1	Everolimus	362	7.6	1.05	3	0.93 (NS)	2.2
	Placebo	184	7.3	(0.68)	2.6		1.6
REACH	Ramucirumab	283	9.2	0.86	3.5	0.59	7.1
	Placebo	282	7.6	(0.13)	2.6	(<0.001)	0.7
RESOURCE	Regorafenib	379	10.6	0.63	3.2	0.44	10.6*
	Placebo	194	7.8	(<0.001)	1.5	(<0.001)	4.1*

Abbreviations: NS, not significant; OS, Overall Survival; TTP, Time-to-progression.

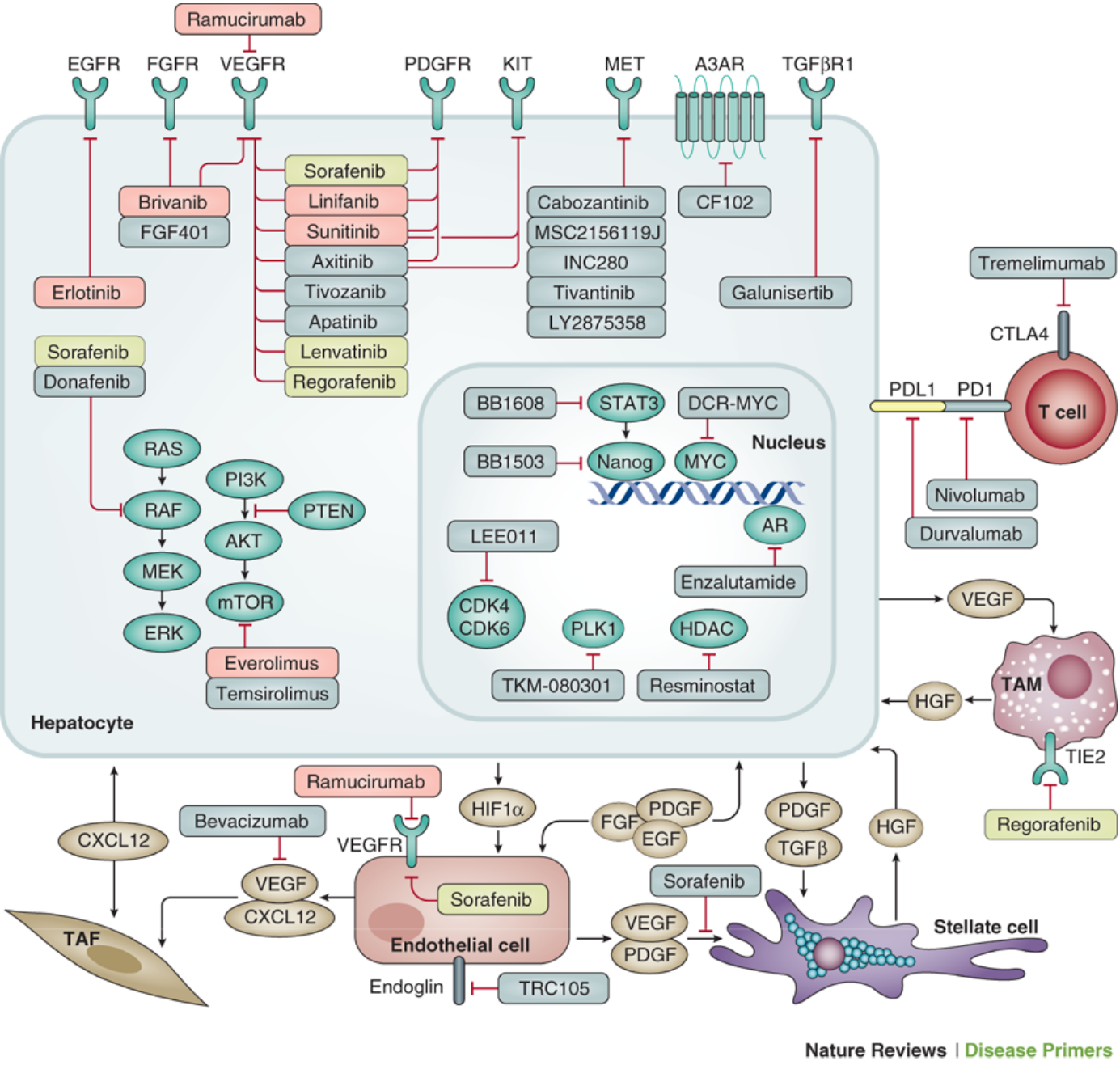
*Refers to Objective response by mRECIST criteria. Celestial Cabozantinib: n 470, Median OS (month) 10.2, Hazard ratio (p-value) 0.76, Median TTP (month) 5.5, Hazard ration (p-value) 0.40, Objective response (%) 4 Placebo: n 237, Median OS (month) 8.0, Hazard ration (p-value) 0.0049, Median TTP (month) 1.9, Hazard ratio (p-value) p<0.001, Objective response (%) 0.4.

**Ramucirumab vs placebo (postitive phase III in second line in patients with advanced HCC and AFP>400 Ng/mL).

Recently, lenvatinib showed similar efficacy compared to sorafenib in a phase III trial (median survival 13.6 months vs 12.3 months, respectively). A Phase III study comparing regorafenib (a more potent multikinase inhibitor than sorafenib, but targeting similar kinases) vs placebo in patients progressing to sorafenib has reported a benefit in survival from 7.8 to 10.6 months (HR: 0.62; 38% reduction of risk of death) (Fig. 78-5). Treatment improved survival in all patient subgroups. In this trial, 88% of patients were BCLC C and 12% BCLC B, and all had progressed on sorafenib. Around 30% of patients presented with macrovascular invasion, 70% with extrahepatic spread, and 45% with AFP >400 ng/dL. Response rate was 10% based upon mRECIST. Treatment was started at 160 mg/day (3 weeks on /1 week off). Median time on treatment was 3.5 months. Prevalence of toxicity (hand-foot reaction, fatigue, and hypertension) was higher compared with reported toxicity from sorafenib, but adverse events only led to treatment discontinuation in 10% of cases. Patients progressing after second-line therapy, along with those at a BCLC D stage should receive best supportive palliative care including management of pain, nutrition, and psychological support. Emerging therapies are shown in **Fig. 78-6**.

FIGURE 78-6

Effective and emerging therapies in HCC. Summary of treatments tested in Phase II -III clinical trials. Yellow boxes indicate drugs with positive Phase III studies, red boxes indicate drugs with negative results from Phase III trials and drugs in grey boxes have been tested in Phase II studies. *(Reprinted with permission from JM Llovet et al: Nat Rev Dis Primers 2:16018, 2016.)*



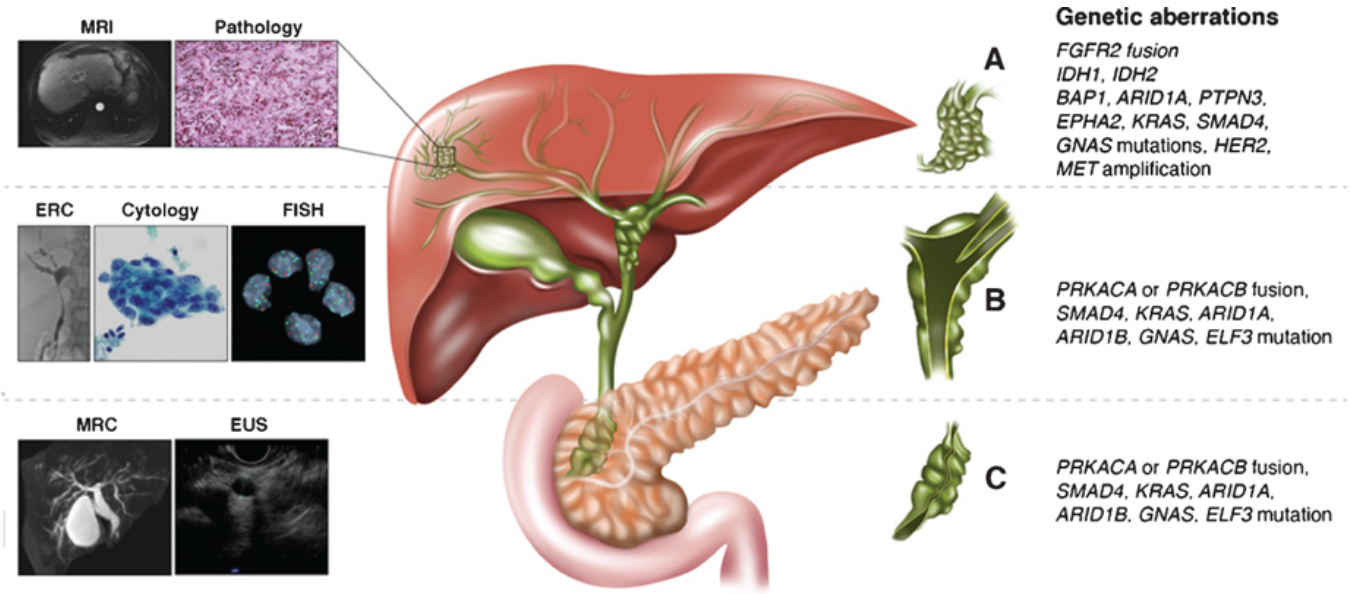
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CHOLANGIOCARCINOMA

Cholangiocarcinoma (CCA) is classified according to its anatomic location as intrahepatic (iCCA; ~30%), perihilar (pCCA; ~50%), and distal (dCCA: ~20%). The latter two are also known as extrahepatic cholangiocarcinomas (eCCA), with the second-order bile ducts acting as the separation point (Fig. 78-7). This classification is endorsed by the 7th edition of the American Joint Committee on Cancer (AJCC) Staging Manual. In addition, iCCA has been recognized as a distinct entity with specific *ad hoc* clinical practice guidelines. Treatment options beyond surgery are limited, and unlike most solid tumors, no molecular targeted therapies have been approved for its treatment. The three subtypes of CCA differ in their anatomic location, epidemiology and risk factors, cell of origin, pathogenesis, and treatment. iCCA originates from adult cholangiocytes, trans-differentiation of adult hepatocytes and hepatic progenitor cells, whereas eCCA arises from the biliary epithelium and peribiliary glands. Moreover, their mutational profile also differs. FGFR2 fusions and IDH1/2 mutations only occur in iCCA, whereas ERBB2 amplifications, APOBEC-associated mutation signatures, and PKA fusions occur in eCCA. Thus, clinical management and trials testing molecular therapies should be tailored according to each biological/anatomical subtype of CCA, as opposed to a common approach for all biliary tract cancers.

FIGURE 78-7

Classification of cholangiocarcinoma subtypes. The 7th edition AJCC/UICC TNM staging classification includes intrahepatic (iCCA, ***A***), perhiliar (pCCA, ***B***) and distal (dCCA, ***C***) tumors. *(Reprinted with permission from S Rizvi, GJ Gores: Hepatology 63:1356, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPIDEMIOLOGY, RISK FACTORS, AND MOLECULAR TRAITS

CCA is the second most common liver cancer following HCC, with a 5-year survival of 10%. iCCA has globally increasing incidence and mortality rates. The incidence of iCCA varies according to exposure to risk factors, ranging from 1–2 cases per 100,000 inhabitants in Europe and North America to the highest incidence in some areas of Southeast Asia, particularly in Thailand (>80 cases/100,000 inhabitants). The male/female ratio is 1.2. Overall, most cases occur with unknown risk factors. The classical risk factors for CCA development include primary sclerosing cholangitis (PSC), biliary duct cysts, hepatolithiasis, Caroli's disease. Parasitic biliary infestation with flukes (i.e., most common is *Opisthorchis viverrini* and *Clonorchis sinensis*), is a prevalent etiology in Asia that can be prevented with an antihelminth therapy, praziquantel. PSC is a clear risk factor for iCCA and pCCA development, with a lifetime incidence ranging from 5 to 10%. Surveillance in PSC patients is recommended with annual imaging techniques and CA 19.9 serum determination. Common risk factors for HCC, such as HBV and HCV infection and cirrhosis, have been associated to iCCA development.

Molecular classification and drivers. There is no established molecular classification of CCA. Genomic studies have provided insight on two subclasses of iCCA, a proliferation subclass—characterized by activation of oncogenic signaling pathways (including RAS and MET)—and an inflammation subclass, characterized by activation of inflammatory pathways, overexpression of cytokines, and STAT3 activation. The landscape of mutations discovered by whole exome sequencing techniques defines a distinct mutational fingerprint depending on etiology and CCA subtype (Fig. 78-7). iCCA mutation portrait is characterized by ~50–60% of tumors having at least one targetable driver including *FGFR2* fusion events (~25%), mutations in *IDH1-2* (15%), *KRAS* (15%), *BRAF* (5%) and *EGFR* (3%), and amplifications in *FGF19/CCDN1* (4%). While mutations in P53 (~30%) and *KRAS* (~25%) are more common in eCCA than in iCCA, some molecular drivers are specific for subypes, such as fusion of *PRKACA* or *PRKACB* for eCCA or *ERBB2* amplifications (~20%) for gallbladder cancer. Liver flukes-associated CCA have higher incidence of *TP53* and *SMAD4* mutations. Host genetic polymorphisms predisposing to CCA have not been established.

INTRAHEPATIC CHOLANGIOCARCINOMA

Diagnosis and Staging

Diagnosis of iCCA requires pathological confirmation. Guidelines are currently not recommending surveillance for early diagnosis, when patients are asymptomatic, since at-risk populations are ill-defined. Cirrhotic patients at risk of HCC development are enrolled in surveillance programs, and can benefit for early detection of iCCA. Otherwise, incidental diagnosis occurs due to cross-sectional imaging performed for other reasons. In most cases, iCCA is diagnosed at advanced stages where symptoms such as weight loss, malaise, abdominal discomfort, or jaundice are present. Pathological diagnosis of iCCA is based on the WHO criteria. Differential diagnosis should be established with metastatic adenocarcinoma and mixed iCCA-HCC tumors, which may require evaluation of markers such as Hep-Par-1, GPC3, HSP70, and glutamine synthetase markers. Imaging studies with CT/MRI are not accurate enough to establish iCCA non-invasive diagnosis. Dynamic CT scanning characterizes 80% of iCCA as liver mass-forming tumors with progressive contrast uptake from the arterial to the venous/delayed phase. MRI dynamic images also show peripheral enhancement in the arterial phase followed by progressive filling-in of the tumor. Atypical radiological behavior with arterial enhancement recapitulating HCC occurs in 10% of

cases. MRI with cholangiopancreatography (MRI/MRCP) is useful to visualize the ductal system and vascular structures. Guidelines do not recommend PET scan for diagnosis. Tumor biomarker carbohydrate antigen (CA) 19-9 at a cut-off level of 100 U/mL has prognostic significance, but lacks accuracy (sensitivity and specificity of ~60%) for early diagnosis.

Radiological criteria are inadequate for iCCA diagnosis in cirrhotic patients. However, in non-cirrhotic patients, guidelines endorse a presumed diagnosis of iCCA (i.e., venous phase contrast enhancement on dynamic CT/MRI) if resection is considered. Assessment of disease extent (venous or arterial invasion and extrahepatic disease) and resectability is best accomplished with CT and/or MRI studies. Doppler ultrasound is accurate in defining vascular invasion. Before surgery, PET scanning may be considered to rule out an occult primary or metastatic site.

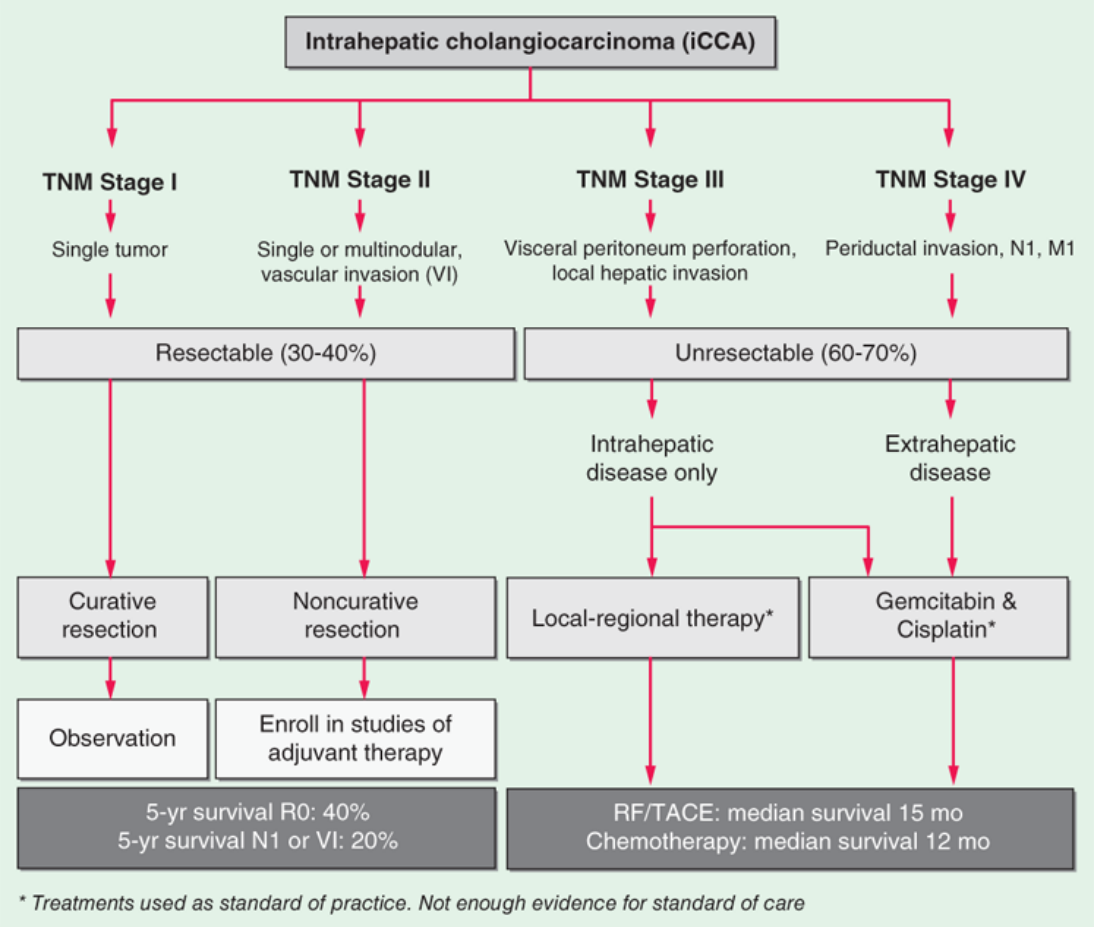
Staging system. The staging system for iCCA resected cases is based on the TNM staging as per the 7th edition of the AJCC/UICCA staging, which is a new system that has already been validated. T1 tumors are solitary without vascular invasion; T2 disease includes multiple tumors (e.g., multi-focal disease, satellitosis, intrahepatic metastasis), or with vascular invasion (microvascular or major vascular invasion); T3 tumors directly invade adjacent structures; and T4 disease includes tumors with any periductal-infiltrating component. Regional lymph node metastasis in the hilar, periduodenal, and peripancreatic nodes are considered N1 disease, while distant spread is considered M1 disease. TNM stages I, II, and III overlap with T status, whereas stage IV includes either periductal invasion or N1/M1 disease.

TREATMENT

After adopting the TNM staging system, the International Liver Cancer Association (ILCA) guidelines for management of iCCA proposed the treatment algorithm depicted in Fig. 78-8. Overall, most of the treatments endorsed have a modest level of evidence and, thus guidelines are providing physicians with recommendations as standards of practice rather than standards of care supported by robust evidence-based data. Surgical resection represents the sole curative treatment option in 30–40% of patients with a 5-year survival of 30%. The largest systematic review including ~4500 iCCA patients undergoing resection reported a median survival of 28 months. In non-cirrhotic individuals, the best candidates for resection are patients at TNM stage I-II, whereas in patients with cirrhosis liver function should be assessed as previously described for HCC. Preoperative disease assessment should discard vascular invasion, N1 and M1. Lymphadenectomy of regional nodes is recommended given its prognostic value. The main predictors of recurrence (~50–60% at 3 years) and survival are identified at the pathological examination, including presence of vascular invasion, lymph node metastases, and poor differentiation. There is no established adjuvant therapy. Liver transplantation remains controversial, and few studies have reported good outcomes for single tumors ≤2 cm.

FIGURE 78-8

Staging and treatment schedule for iCCA proposed by the International Liver Cancer Association. *(Reprinted with permission from J Bridgewater: J Hepatology 60:1268–1289, 2014.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Non-surgical candidates have a dismal life expectancy. Overall, patients at stage III might be considered for loco-regional therapies, such as chemoembolization or radioembolization, but the level of evidence is low, mostly based on cohort studies. A meta-analysis of 14 trials testing loco-regional therapies reported median survival times of 15 months. External beam radiation therapy is not recommended as standard therapy. At more advanced stages (stage IV) in patients ECOG 0-1, systemic chemotherapy with the combination of gemcitabine and cisplatin is considered the standard of practice yielding median survival rates of 11.7 months compared to 8 months for gemcitabine alone. This recommendation for first-line treatment of advanced tumors is based on a subgroup analysis of 80 iCCA patients included in a large randomized Phase III trial (n = 410, ABC trial-02) of patients with advanced biliary tract tumors.

No molecular targeted therapy has been proven effective for iCCA. Patient stratification based on molecular biomarkers is ongoing with FGFR2 aberrations and IDH1/2 mutations. Preliminary data of a Phase II trial testing BGJ398 in advanced iCCA harboring FGFR2 gene fusions reported ~20% objective response.

Mixed HCC-iCCA is a rare neoplasm accounting for <0.5% of all primary liver cancers. Diagnosis is based on pathology. The 2010 WHO classification defined two subtypes: the classical and the stem cell feature type. Molecular data has also characterized a third unique entity, cholangiolocellular carcinoma, with distinct molecular traits and better outcome. Due to its low incidence, the demographic features and clinical behavior of these tumors remain ill-defined. Survival rates are similar to iCCA, and until specific guidelines are available, they should be managed following the treatment algorithm of iCCA.

EXTRAHEPATIC CHOLANGIOCARCINOMA

Perihiliar (pCCA) and Distal Cholangiocarcinoma (dCCA)

The 7th edition AJCC/UICC TNM staging classification has established pCCA as tumors that arise between the second-order bile ducts up to the insertion of the cystic duct, whereas dCCA arise from this point to the ampulla of Vater (Fig. 78-7). Thus, dCCA can be difficult to distinguish from early pancreatic cancer. Both entities have a similar diagnostic approach. Acute onset of painless jaundice occurs in 90% of patients with pCCA, and 10% present with cholangitis. Primary biliary cholangitis with a cut-off for CA19.9 >129 U/mL is suspicious for CCA. Imaging assessment starts with CT and MRI; they have a good sensitivity and specificity (>85%) for detecting the degree of bile duct involvement, and hepatic and portal vein invasion. MRI-cholangiography is optimal for defining the extent of the bile duct lesion. Ruling out IgG-4 cholangiopathy by assessing serum IgG4 is mandatory. As a

second step, endoscopic retrograde cholangiography with brushing to explore cytology and fluorescence in situ hybridization (FISH)—for exploring polysomy—is recommended. FISH enhances the sensitivity of cytology from 20 to ~40%.

Diagnosis is based upon pathology. The treatment algorithm for pCCA indicates that in cases of a dominant stricture with positive cytology/biopsy or polysomy, a lymph node biopsy through endoscopic ultrasound should be obtained. pCCA with negative lymph node involvement is best treated by surgery, resection, or transplantation, the sole curative options. Resection entails hepatic and bile duct removal, Roux-en-Y-hepaticojejunostomy with regional lymphadenectomy. Bilobular involvement is considered a surgical contraindication. In few referral centers, unresectable single pCCA <3 cm without dissemination can be considered for liver transplantation with neoadjuvant chemoradiation. This procedure is associated with 5-year survival rates of ~70%. If lymph node involvement is present, systemic chemotherapy can be considered along with biliary tract stenting. Of note, the subgroup analysis of the Phase III ABC trial-02 did not identify differences between gemcitabine alone or in combination with cisplatin for pCCA. Surgical resection (Whipple procedure) is the primary option for management of dCCA, a procedure that achieves a median survival of 2 years and 5-year survival rates of ~25%. Main contraindications for resection are presence of distant lymph node involvement, metastases, or major vascular invasion. At the pathological examination, perineural invasion, lymph node metastasis, R0 resection (absence of residual tumor at pathological examination), and tumor differentiation are predictors of survival. Adjuvant therapy has not shown outcome benefits. There is no evidence of benefit of chemotherapy for unresectable cases. No molecular targeted therapies are available for these entities.

GALLBLADDER CANCER

Gallbladder cancer is the most common cancer of the biliary tract worldwide. The estimated cases of gallbladder cancer in the United States in 2016 are 11,400, more than CCA. The female:male ratio is 3:1. Cholelithiasis is the major risk factor, but <1% of patients with cholelithiasis develop this cancer. Gallbladder polyps at risk of transformation are those with ≥10 mm in diameter. Early cases are discovered incidentally at routine cholecystectomy. Clinical symptoms, such as jaundice, pain, and weight loss, are associated with advanced stages. Staging of gallbladder cancer follows the TNM classification. The most accurate technique to define staging and vascular and biliary tract invasion is the magnetic resonance cholangiopancreatography. CT and PET scan can be also useful for preoperative staging.

The mainstay of treatment is surgical, either simple or radical cholecystectomy (partial hepatectomy and regional lymph node dissection) for stage I or II disease, respectively. Only ~20% of patients are candidates for surgery with a curative intent. Survival rates are near 80–90% at 5 years for stage I, and range from 60 to 90% at 5 years for stage II. Regional nodal status and the depth of tumor invasion (T status) are the two most important prognostic factors. Adjuvant therapy has not proven effective. Gallbladder cancers at stage III and IV are considered unresectable. For patients with ECOG 0-1, chemotherapy with gemcitabine and cisplatin is the standard of practice based on data from the subgroup analysis including 181 patients with gallbladder cancer in the setting of two clinical trials. Overall, median survival is 10–12 months in advanced cases. Percutaneous transhepatic drainage is indicated in case of biliary obstruction. Radiotherapy is not effective.

OTHER MALIGNANT LIVER TUMORS

Fibrolamellar Hepatocellular Carcinoma (FLC)

FLC is a rare form of primary liver cancer that typically affects children and young adults (10–30 years of age) without background liver disease. FLC accounts for 0.85% of all primary hepatic malignancies in the United States, and its incidence rate is 0.02 cases per 100,000 inhabitants. FLC is considered a unique entity with a specific fusion oncogene *PRKACA-DNAJB1* present in 80–100% of cases. A few mutations have been described, all at a level of <10%. FLC has a better prognosis than HCC, probably due to the absence of cirrhosis and the earlier age of presentation. Surgical resection is the mainstay of treatment and indications are less restrictive than for HCC. A retrospective series of 575 FLC cases reported a median survival of 70 months after resection. At advanced stages, the expected outcome is <20 months. Chemotherapy is not effective and there is no standard of care.

Hepatoblastoma (HB)

HB is the most frequent primary liver tumor in children. The incidence of the disease is 1.5 cases per 1,000,000. Background liver disease is rare in these patients. WNT signaling plays a major role, with *CTNNB1* mutations (70%) as the most frequently reported molecular event. A gene signature is able to discriminate two molecular classes with distinct outcome. Resection followed by chemotherapy with doxorubicin is the mainstay treatment strategy. A study including 1605 patients randomized in eight clinical trials reported better outcome for patients with stage I-II of the PRETEXT classification (out of four stages), age <3 years, AFP >1,000 ng/mL, and absence of metastases. As opposed to HCC, low AFP indicates poor prognosis. Outcomes for best candidates after resection (stages I/II with small tumors, age <3 years and AFP >100 ng/mL) achieve 5-year disease-free survival of 90%, compared with worst candidates (metastatic disease and AFP <100 ng/mL) with 5-year disease-free survival of 20–30%.

BENIGN LIVER TUMORS

The most common benign liver tumors are hemangiomas, focal nodular hyperplasia (FNH), and hepatocellular adenomas (HCA). Most benign tumors are identified incidentally by abdominal ultrasound or other imaging techniques. *Hemangiomas* are present in ~5% of the general population, are diagnosed by ultrasound except in cirrhotic patients or oncology patients where contrast enhanced imaging (contrast enhanced ultrasound, CT, or MRI) is required. Conservative management is appropriate and follow-up is not recommended. Exceptionally, growing lesions causing symptoms by compression can be considered for resection. FNH is a benign tumor present in <2% of the population and occurring mostly in females aged 40–50 years. FNH is a polyclonal hepatocellular proliferation due to an arterial malformation. MRI has the highest diagnostic accuracy with a specificity of 100%, when typical imaging features are present (homogeneous enhancement in the arterial phase with a central scar). Atypical FNH requires biopsy for diagnosis. Treatment is not recommended since these tumors do not degenerate or cause complications. In exceptional cases of expanding symptomatic lesions, surgery is the treatment of choice.

Hepatic adenomas are clonal benign proliferations resulting from single gene driver mutations. HCA have a low prevalence of 0.001% of the population and are frequently diagnosed in women aged 35–40 years. The female:male ratio is 10:1, and the main risk factors are oral contraceptives in females and use of anabolic androgenic steroids in male body builders. HCA have the potential for hemorrhage and HCC development, particularly when sized >5 cm. Nowadays, there is a clear understanding of the molecular classification of HCA: (a) HCA with *CTNNB1* mutations (10–20%) are at-risk of HCC development and are present in men treated with androgens; (b) inflammatory adenomas (50–60%) are associated with single mutations (*Gp130*: 65%) and are more prevalent in females with obesity or diabetes; and (c) adenomas with inactivated *HNF-1A*. Diagnosis is based on MRI imaging, which is able to correlate with molecular subtypes in 80% of cases (Inflammatory and HNF-1A type). For defining HCA with *CTNNB1* mutations, biopsy is required. Upon diagnosis, discontinuation of oral contraceptives and weight loss is recommended. Resection is indicated in all cases of size >5 cm or men or *CTNNB1* mutation. For HCA <5 cm, 1-year follow-up is recommended. In case of active HCA bleeding, embolization followed by resection is the treatment of choice. The presence of multiple HCA is common, and guidelines endorse treating them based on the size of the main nodule.

FURTHER READING

Bridgewater J et al: Guidelines for the diagnosis and management of intrahepatic cholangiocarcinoma. J Hepatol 60:1268, 2014. [PubMed: 24681130]

Bruix J, Sherman M: Management of hepatocellular carcinoma. Hepatology 42:1208, 2005. [PubMed: 16250051]

Bruix J et al: Regorafenib for patients with hepatocellular carcinoma who progressed on sorafenib treatment (RESORCE): A randomized, double-blind, placebo-controlled, phase 3 trial. Lancet 389:56, 2017. [PubMed: 27932229]

Clavien PA et al: Recommendations for liver transplantation for hepatocellular carcinoma: An international consensus conference report. Lancet Oncol 13:e11, 2012. [PubMed: 22047762]

EASL-EORTC Clinical Practice Guidelines: Management of hepatocellular carcinoma. J Hepatol 2018.

Forner A et al: Hepatocellular carcinoma. Lancet 379:1245, 2012. [PubMed: 22353262]

Llovet JM et al: Sorafenib in advanced hepatocellular carcinoma. N Engl J Med 359:378, 2008. [PubMed: 18650514]

Llovet JM et al: Hepatocellular carcinoma. Nat Rev Disease Primers 2:16018, 2016. [PubMed: 27158749]

Mazzaferro V et al: Liver transplantation for the treatment of small hepatocellular carcinomas in patients with cirrhosis. N Engl J Med 334:693, 1996. [PubMed: 8594428]

Razumilava N, Gores G: Cholangiocarcinoma. Lancet 383:2168, 2014. [PubMed: 24581682]

Schulze K et al: Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet 47:505, 2015. [PubMed: 25822088]

Valle J et al: Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. N Engl J Med 362:1273, 2010. [PubMed: 20375404]

Zucman-Rossi J et al: The genetic landscape and biomarkers of hepatocellular carcinoma. Gastroenterology 149:1226, 2015. [PubMed: 26099527]

Chapter 79: Pancreatic Cancer

Daniel D. Von Hoff

INTRODUCTION

Pancreatic cancer is the third leading cause of death from cancer in the United States with >53,000 Americans diagnosed and >43,000 dying from the disease each year. Unfortunately, pancreatic cancer is projected to be the second leading cause of death from cancer in the United States by 2030. Worldwide pancreatic cancer is the eleventh most common cancer with 338,000 new patients diagnosed and >334,000 deaths (seventh cause of cancer deaths). Pancreatic cancer currently has the worst survival rate of any cancer with an overall 5-year survival (regardless of stage) of ~8.2%. However, that situation is changing because some advances have been made against the disease with some improvements in survival (see below) that may affect the 5-year survival statistics. In particular, knowledge about specific molecular subsets of the disease has become crucial so that one can provide the best possible care for their patients with pancreatic cancer.

EPIDEMIOLOGY

Pancreatic cancer accounts for 3.2% of all new cancer cases in the United States and for 7.2% of all deaths from cancer in the United States. The lifetime risk of developing pancreatic cancer is ~1.6%. The incidence of pancreatic cancer has been increasing between 0.5 and 1% per year. Pancreatic cancer is more common with increasing age and more common in men than in women. The 5-year survival rate for all stages has increased from 3% in 1975 to 8.2% in 2013. The latest information from the U.S. Surveillance, Epidemiology, and End Results (SEER) database predicts that the 5-year survival for patients with localized pancreatic cancer is about 31.5%, 11.5% for those with regional disease, and 2–5% for patients with advanced metastatic disease. Pancreatic cancer is more common in developed countries (although generally it tracks with the prevalence of smoking). The incidence is highest in North America and Western Europe followed by other areas in Europe, Australia, New Zealand, and South-Central Asia. Of note is that the population at greatest risk are women living in Scandinavian countries, while the lowest risk is seen for women living in middle Africa.

RISK FACTORS

Age is one of the greatest risk factors for pancreatic cancer with median age at diagnosis of 70 years (the disease is most frequently diagnosed in the 65–74 age group). The number of new cases per 100,00 persons and the number of deaths per 100,000 persons are higher for males and blacks of both sexes. Both the number of cases and the number of deaths per 100,000 people are lower for American Indian/Alaskan natives

and Asian Pacific Islanders. Both the number of cases and deaths are intermediate for the Hispanic population.

Environment

The greatest risk factor for pancreatic cancer is cigarette smoking. The risk correlates with the increased number of cigarettes smoked. It has been estimated that 30% of pancreatic cancer is caused by smoking. Exposure to cadmium as part of cigarette smoking or via exposure to welding, soldering, or dietary exposure has been weakly associated with an increased risk of pancreatic cancer.

Although dietary factors are often difficult to interpret, evidence suggests that high intakes of fat or meat (particularly well-done barbequed meat) are risk factors. High intakes of fruits and vegetables are associated with a decreased risk. Coffee and low-to-moderate alcohol consumption have been determined not to be associated with an increased risk for pancreatic cancers, while consumption of sugary fizzy drinks has been associated with an increased risk.

Microbiome

To date, there is no solid evidence of an association between *Helicobacter pylori* infection and pancreatic cancer. Some data link the oral microbiome associated with poor dentition to pancreatic cancer but the evidence is very thin.

Hereditary/Genetics

Hereditary factors may account for 10–16% of all pancreatic cancers. It is very important to recognize these factors for determining risk for family members of the patient affected with pancreatic cancer. (These family members should seek participation in an early detection program with genetic counseling, definition of risk and perhaps, if appropriate, periodic MRI screening of the abdomen, though this recommendation is not based on research data.) In addition, the identification of any of these germ-line mutations can lead to specific and effective new therapeutics for patients with these abnormalities in their tumors. **Table 79-1** identifies the various germ-line mutations along with their familial cancer syndromes where an increased risk for pancreatic cancer is known.

TABLE 79-1

Germ-line Mutations, Their Familial Cancer Syndrome, and Fold Risk of Pancreatic Cancer

Germ-Line Mutation	Familial Cancer Syndrome	Estimated Increased Risk (fold) of Pancreatic Cancer
BRCA2[a]	Familial breast/ovarian cancer	3.5–10
PALB2 (partner and localizer of BRCA2)	Familial breast cancer and others	~sixfold
p16/CDKN2A	Familial atypical multiple mole melanoma (FAMMM)	13–38
STKII (LKB1)	Peutz-Jeghers syndrome	132
PRSS1 or SPIN11[b]	Hereditary (familial) pancreatitis	53
ATM	Ataxia-telangiectasia	Not yet established
MLH1, MSH2, MSH6, PMS2	Heredity nonpolyposis colorectal syndrome or Lynch syndrome[c]	9–30

[a]Particularly common in individuals with Ashkenazi Jewish heritage.

[b]Forty percent chance of pancreatic cancer by the age of 70.

[c]Very important because this is associated with microsatellite instability, which is a marker for response to an anti-PD-1/PD-L1 agent.

Knowing the patient has a BRCA2 or PALB2 germ-line mutation or any of the above mutations should lead one to not only refer the patient's relatives to an early detection or high-risk individual clinic but also realize that for the BRCA2/PALB2 germ-line mutation patients consideration for treatment with a poly (ADP-ribose) polymerase (PARP) inhibitor should be entertained. Other germ-line mutations are under study to determine their increased risk of pancreatic cancer including *CDK4, FANCC, PALLD, APC, ATM, BMPR1A, BRCA1, EPCAM, MEN1, MLH1, MSH2, MSH6, NF1, PMS2, SMAD4, TP53, TSC1, TSC2,* and *VHL*. Some of these mutations are associated with pancreatic neuroendocrine tumors (Chap. 80).

In addition to the recognized genetic syndromes, other possible familial pancreatic cancer genes have not yet been discovered. For example, a family history of pancreatic cancer is associated with a 13-fold increase in

the disease. If you have one first-degree relative, the risk is increased 4.6-fold, 2 first-degree relatives 6.4-fold, and ≥3 first degree relatives a 32-fold increase. The risk is also increased if a relative developed pancreatic cancer at <55 years old.

Other Considerations

Most patients with pancreatic cancer relate that they have had developing symptoms over the last few years. However, Yachida and colleagues suggest that pancreatic cancer could be growing over a period of 21 years. Thus, there is a possibility for early detection of the disease.

Medical Conditions

Chronic pancreatitis that is nonfamilial is also associated with an increased risk of pancreatic cancer (2.3–16.5-fold increase). It is also increased in people with chronic pancreatitis associated with cystic fibrosis or tropical pancreatitis.

A clear association exists between diabetes mellitus and pancreatic cancer. Whether this is a causal association or whether the diabetes is the result of the cancer is not exactly clear. What is clear is that when a person presents with new onset diabetes, they should be considered at risk for having pancreatic cancer. The excessive insulin or insulin-like growth factors associated with adult onset diabetes and metabolic syndrome may promote pancreatic carcinogenesis.

Obesity is considered a possible risk factor for pancreatic cancer. A high body mass index (BMI) ≥30 is associated with a doubling of the risk of pancreatic cancer. Since obesity is a risk factor for diabetes, the contribution of obesity alone is unclear. Interestingly, patients with severe obesity who undergo a gastric bypass experience a reduction in the incidence of gastrointestinal (GI) cancer including pancreatic cancer by >30% in the first 3 years (along with a dramatic decrease in their Hgb A1c and blood glucose). Physical inactivity also has been associated with an increased risk in pancreatic cancer.

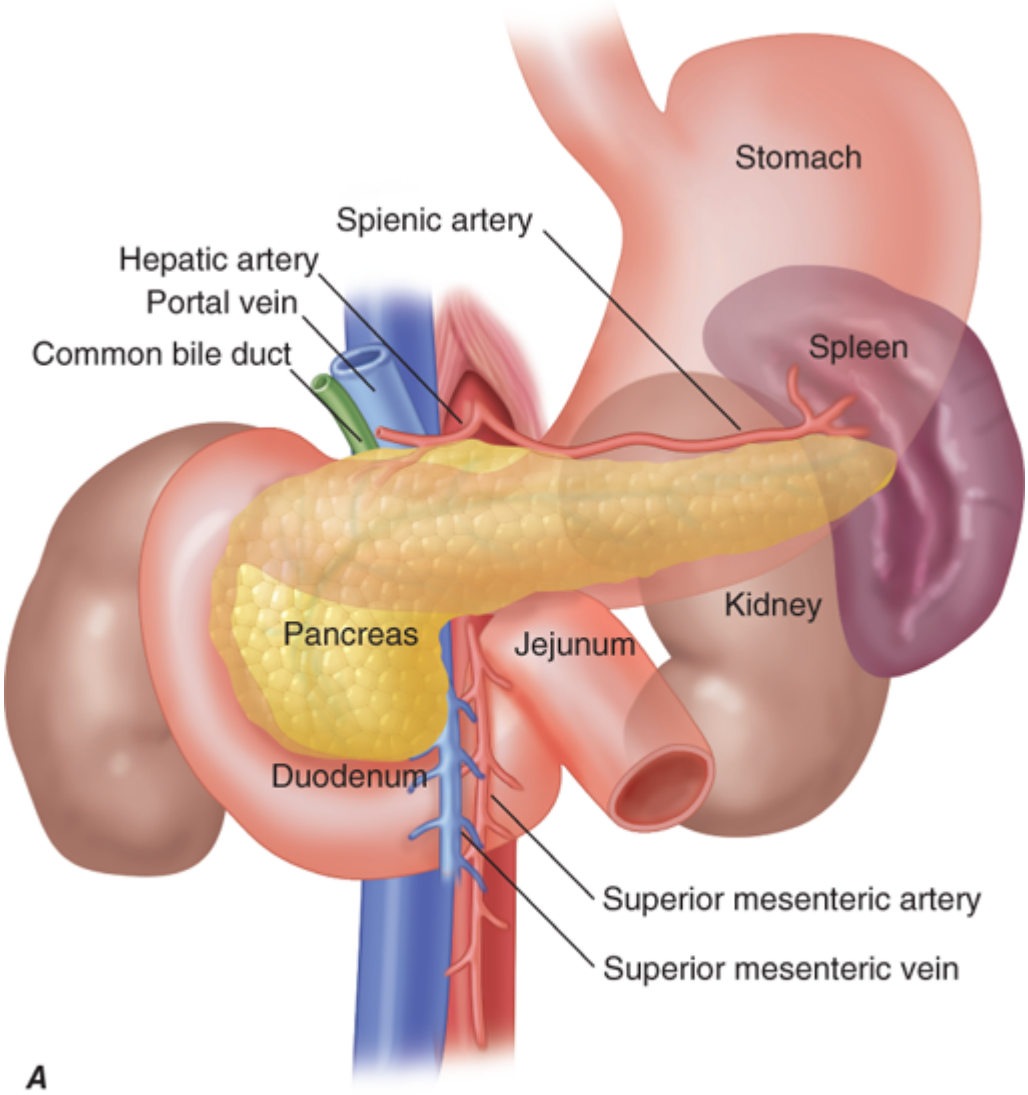
PATHOLOGY AND MOLECULAR CONSIDERATION

Location

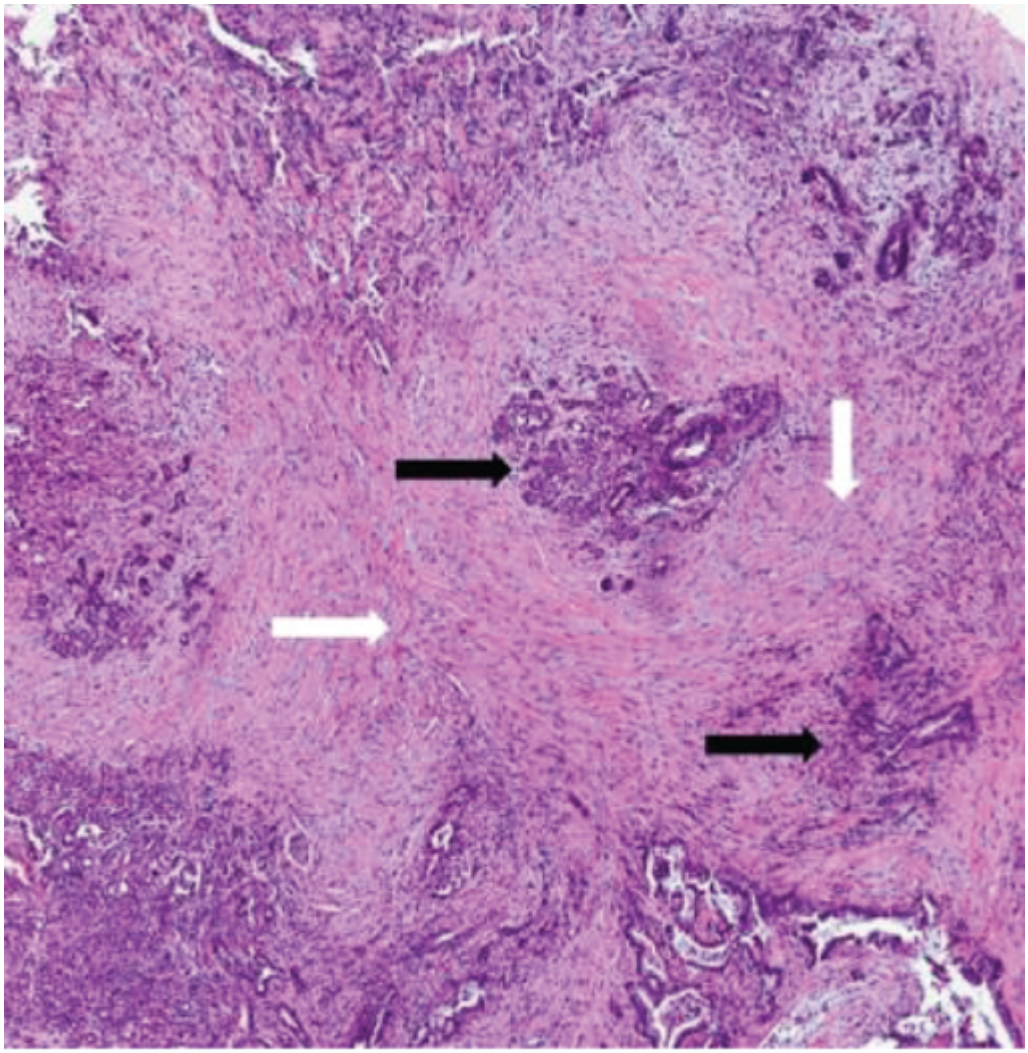
The posterior location of the pancreas in the abdomen is likely one of the issues that leads to a late diagnosis **(Fig. 79-1A)**.

FIGURE 79-1

A. Note the relationship of the pancreas to the major vessels of the retroperitoneum. ***B.*** Ductal adenocarcinoma of the pancreas (*black arrows*), with intense stromal component (*white arrows*). *(Part A is courtesy of Mary Kay Washington, MD, PhD Vanderbilt University. Part B is courtesy of Haiyong Han, PhD Translational Genomics Research Institute [TGen].)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pathology

Cancers of the pancreas can be divided into neoplasms of the endocrine pancreas (Chap. 80) and tumors of the exocrine pancreas. The most common neoplasm of the exocrine pancreas and most deadly is pancreatic infiltrating ductal adenocarcinoma. These tumors arise in the head, body, or tail of the pancreas and are characterized by infiltrating desmoplastic stromal reactions (Fig. 79-1B).

Other subtypes of nonneuroendocrine pancreatic cancers include acinar cell carcinoma (tumors of the exocrine enzyme producing cell); medullary carcinoma, adenosquamous, and other rare subtypes. Each of these are different in their behaviors and in their molecular characteristics and often require specific other types of treatment.

Molecular Characteristics

The molecular characteristics of pancreatic ductal adenocarcinoma reveal four genes that are commonly mutated or inactivated (sometimes referred to as the “4 horsemen”). The most common is *KRAS* (usually in codon12), which is seen in virtually 100% of pancreatic adenocarcinomas. In fact, with the deep sequencing now available, if a *KRAS* mutation is not detected in the patient’s tumor, one should consider the tumor being of a different origin (such as small bowel, gallbladder, or cholangiocarcinoma)—all of which could require different treatments. *p16/CDKn2A* is also noted in >90% of invasive pancreatic adenocarcinomas. *TP53* and *DPC4/MADH4* are mutated in about half of these tumors. As a reference point, the BRCA2 gene noted in Table 79-1 is mutated in 7–10% of pancreatic adenocarcinomas.

Precursor Lesions

Many pancreatic adenocarcinomas seem to arise from noninvasive epithelial precursor lesions. Detection of these could allow for early diagnosis of pancreatic cancer. These pancreatic intraepithelial neoplasias (PanINs) have varying degrees of dysplasia designated as PanINs 1–3 (and constitute a progression model for pancreatic cancer). Genetic alterations become more frequent as the PanIN grade increases (e.g., grade 3). Not all PanIN lesions progress to invasive malignancy. PanINs that are ≥1 cm are called *intraductal papillary neoplasms* and are usually noninvasive. If the intraductal tumor is in a branch duct, it is usually noninvasive; however, if the intraductal tumor is in a main duct and is large and nodular, it is more likely to have malignant behavior.

One other pancreatic tumor is the mucinous cystic neoplasm; they may be seen as incidental findings on scans. These lesions are less likely invasive (20%) unless they are large and have nodules in them.

CLINICAL FEATURES

History and Physical

The classic presentation for a patient with pancreatic cancer has been abdominal pain and weight loss with or without jaundice. The pain is midepigastric (sometimes described as a “boring-like” pain). Often the pain is in the back (due to retroperitoneal invasion of the splanchnic nerve plexus). The pain may be exacerbated by eating or lying flat. Other items of note in a history is lightening in stool color (steatorrhea also causes malodorous stools), or the onset of diabetes in the prior year. Jaundice, first detectable with a bilirubin of 2.5–3.0 mg/dL, is usually associated with tumor in the head of the pancreas. In some instances, depression is noted (with a higher subsequent number of suicides). Pruritis may be seen when the bilirubin reaches 6–8 mg/dL.

Physical signs include jaundice, signs of weight loss, a palpable gallbladder (Courvoisier’s sign), hepatomegaly, an abdominal mass, and even an enlarged spleen (usually indicating a portal vein thrombosis). Migratory superficial thrombophlebitis can also be seen (Trousseau’s syndrome). Signs of late disease include a lymph node palpable in the supraclavicular fossa (usually on the left where the thoracic duct enters the subclavian vein). This is clinically referred to as Virchow’s node. Occasionally one can palpate subcutaneous metastases in the periumbilical area referred to as a Sister Mary Joseph’s node—named after

one of the scrub nurses on the Mayo Clinic Operative Team who noted that when she prepped that area and felt those nodules, the patient often had peritoneal metastases.

The history and symptoms noted above may lead a person to see a physician; often CT and MRI scanning detects the disease before advanced disease symptoms appear.

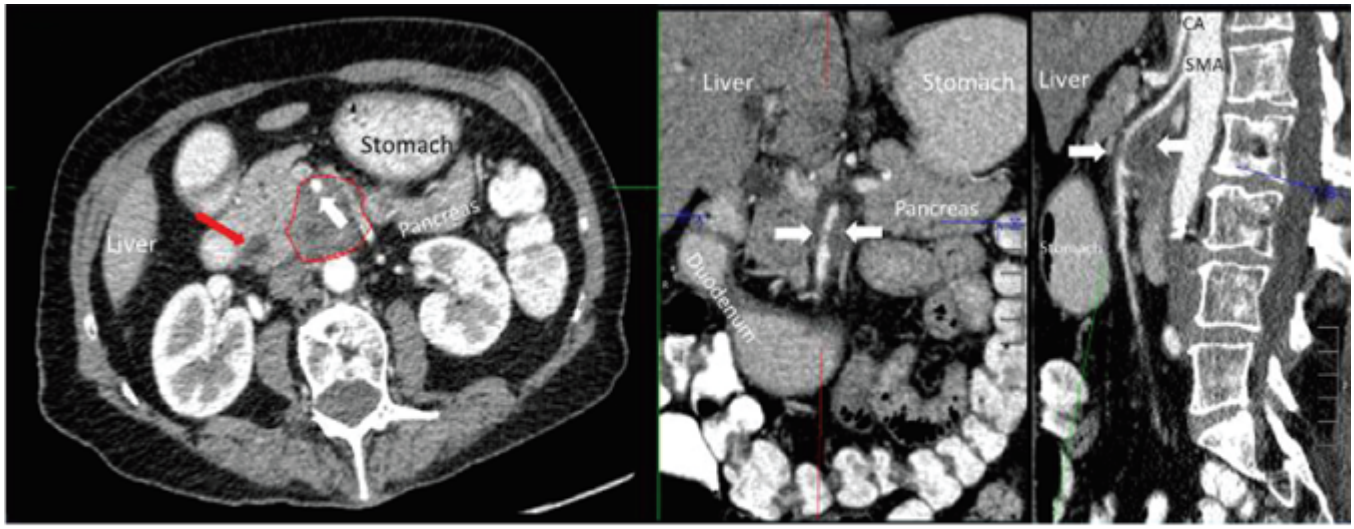
DIAGNOSTIC WORKUP

Imaging

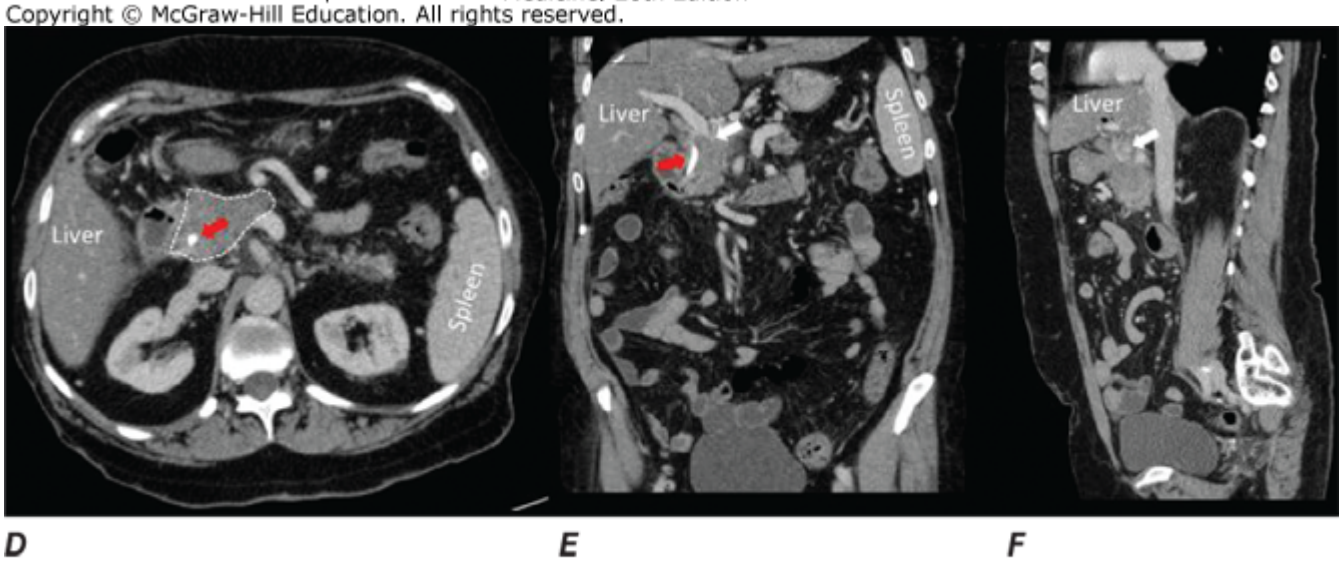
Diagnostic imaging plays a major role in diagnosing pancreatic cancer and other intraabdominal diseases. The best technique is the use of a dual-phase contrast-enhanced spiral CT using the pancreatic cancer protocol which allows arterial phase enhancement and portal venous phase enhancement. This special protocol can provide helpful prospective staging and assessment of resectability. **Figure 79-2** demonstrates such a CT scan (with vascular involvement). **Figure 79-3** demonstrates the use of an 18F glucose positron emission tomography (PET) scan.

FIGURE 79-2

Selected images from contrast-enhanced CT in patients with locally advanced adenocarcinoma of the pancreas. A high-quality contrast-enhanced CT scan (arterial phase in Panels ***A–C*** and portal venous phase in Panels ***D–F***) is required for optimal staging of pancreas cancer. Panel *A* demonstrates the typical features of adenocarcinoma of the pancreas on arterial phase axial CT scans (*dotted outline*) with tumor encasement of the superior mesenteric artery (*white arrow*). Note the dilatation of the common bile duct (*red arrow*). Panels ***B*** (magnified coronal) and ***C*** (sagittal) show reconstruction of CT images into additional orthogonal planes with exquisite details to confirm the unresectable nature of the tumor due to vascular encasement. Panel ***D*** demonstrates the typical features of adenocarcinoma of the pancreas on portal venous phase axial CT scans in a different subject. The dotted line outlines a pancreas cancer lesion in the pancreatic head, which is encasing the portal splenic confluence (*dotted outline*). Panels ***E*** (*white arrow*) and ***F*** show the pinched appearance of the portal splenic confluence by tumor abutment and invasion of the superior mesenteric vein (*white arrow*) on coronal and sagittal views. Note the presence of a stent in the common bile duct (*red arrow*) to help relieve biliary obstruction caused by the tumor. CA, celiac axis; SMA, superior mesenteric artery.



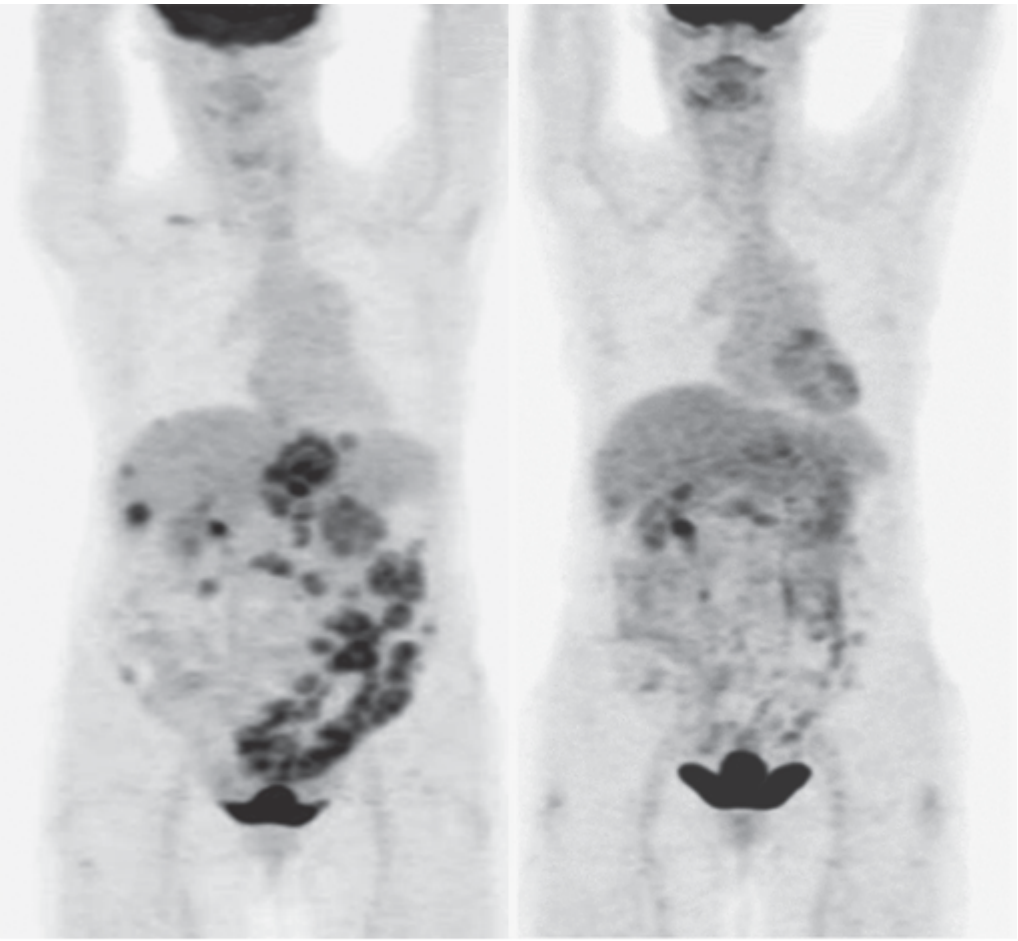
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 79-3.

PET scan demonstrating metastatic disease—baseline and after 6 weeks of chemotherapy with some resolution of liver metastases.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Histologic Diagnosis

A histologic (tissue) diagnosis is essential and should be obtained with a cutting biopsy needle (not a skinny needle with cytology). Misdiagnosis is more common based on only fine-needle aspirates. Obtaining a tissue diagnosis allows not only for accuracy but also for molecular testing for *KRAS* mutations, microsatellite instability, and other important molecular abnormalities. Those molecular abnormalities and others will be increasingly important as more targeted therapies are developed for patients with pancreatic cancer.

The core needle (16–18 gauge) biopsy can be obtained via endoscopic ultrasound-guided techniques for a tumor localized to the pancreas or, if there are liver lesions or Virchow's node, via percutaneous biopsy by interventional radiologists.

Serum Markers

Before treatment, a serum sample should be obtained for levels of CA19-9, carcinoembryonic antigen (CEA), or if both are negative, for CA125 (can be positive when the CA19-9 is negative due to the patient not being a Lewis antigen secretor). These markers are not useful for staging but can be useful in following the course of pancreatic cancer.

IMPORTANT IMMEDIATE CONSIDERATIONS IN PATIENT CARE

While the patient is being evaluated and staged, one must be alert for biliary tract obstruction (and the attendant risk for sepsis from the biliary tree). A stent can be placed (plastic if temporary or metal if needed longer) to relieve the jaundice and pruritus. If surgery is being contemplated, an early surgical consultation is in order as there are surgeons who will want to proceed to surgery without placement of a stent.

Patients with pancreatic cancer are often hypercoagulable and frequently have migratory thrombophlebitis (Trousseau's sign) as well as deep vein thrombosis with pulmonary emboli (a frequent cause of death). Appropriate examinations plus being alert to thromboses on the routine workup are mandatory so appropriate management can be put in place.

Control of pain or of any of the symptoms should be obtained if at all possible to help patients be as comfortable as possible for their decision-making. Sometimes simple approaches like the use of a replacement pancreatic enzyme (at good therapeutic doses) can relieve the bloating, cramping, and diarrhea these patients suffer from. Early involvement of a palliative care team can improve a patient's quality of life and sometimes even its length.

CLINICAL STAGING

The clinical staging of pancreatic cancer according to the American Joint Commission on cancer staging is presented in **Table 79-2**.

TABLE 79-2

Definition of Primary Tumor (T)

T CATEGORY	T CRITERIA
TX	Primary tumor cannot be assesed
T0	No evidence of primary tumor
Tis	Carcinoma in situ This includes high-grade pancreatic intraepithelial neoplasia (PanIn-3), intraductal papillary mucinous neoplasm with high-grade dysplasia, intraductal tubulopapillary neoplasm with high-grade dysplasia, and mucinous cystic neoplasm with high-grade dysplasia
T1 T1a T1b T1c	Tumor ≤2 cm in greatest dimension Tumor is ≤0.5 cm in greatest dimension Tumor >0.5 cm and <1 cm in greatest dimension Tumor 1–2 cm in greatest dimension
T2	Tumor >2 cm and ≤4 cm in greatest dimension
T3	Tumor >4 cm in greatest dimension
T4	Tumor involves celiac axis, superior mesenteric artery, and/or common hepatic artery, regardless of size
M CATEGORY	**M CRITERIA**
M0	No distant metastasis
M1	Distant metastasis
N CATEGORY	**N CRITERIA**
NX	Regional lymph nodes cannot be assessed
N0	No regional lymph node metastases

N1	Metastasis in one to three regional lymph nodes
N2	Metastasis in four or more regional lymph nodes

AJCC Prognostic Stage Groups

WHEN T IS…	AND N IS…	AND M IS…	THEN THE STAGE GROUP IS….
Tis	N0	M0	0
T1	N0	M0	IA
T1	N1	M0	IIB
T1	N2	M0	III
T2	N0	M0	IB
T2	N1	M0	IIB
T2	N2	M0	III
T3	N0	M0	IIA
T3	N1	M0	IIB
T3	N2	M0	III
T4	Any N	M0	III
Any T	Any N	M1	IV

Source: Adapted with permission from MB Amin et al (eds): *AJCC Cancer Staging Manual*, 8th ed. Springer, 2017.

Table 79-3 presents another clinical way to express extent of disease as well as therapeutic approaches (to be discussed later).

TABLE 79-3

Extent of Disease and Therapeutic Approach

Designation (Median Survival)	Therapeutic Approaches
1. Resectable (localized): (18–23 mo) • No encasement of celiac axis or superior mesenteric artery (SMA) • Patent superior mesenteric—portal veins • No extrapancreatic disease	Surgical option (or preoperative-neoadjuvant therapy first). Surgery is followed by postsurgery adjuvant therapy • Currently gemcitabine + capecitabine
2. Locally advanced: (6–10 mo) • Encasement of arteries • Venous occlusion (superior mesenteric vein [SMV] or portal) • No extrapancreatic disease	Either chemotherapy or chemotherapy + radiation therapy
3. Metastatic: (8.3–12.8 mo)	Systemic chemotherapy

For proper staging, some physicians believe that a laparoscopy either before or at the time of contemplated surgery is important. If metastatic disease is found at laparoscopy, one can avoid surgery that would not be helpful because disease is already advanced.

TREATMENT

TREATMENT

Resectable Disease

For patients with resectable disease (as defined in Table 79-3), the best option is surgery. Only a small percentage of patients are in this category (10–20%). The surgery for patients with tumors in the head or uncinate body of the pancreas is usually a pylorus-sparing pancreaticoduodenectomy (a modified Whipple procedure). For tumors in the body or tail, a distal pancreatomy is usually performed. Clinical and pathologic findings of the resection are defined as either, an Ro resection (no macroscopic or microscopic disease left after surgery); an R1 resection refers to residual disease likely left behind. Patients with smaller tumors and lymph node negative disease have a better survival (median of about 18–23 months with 5-year survival of about 20%).

Two approaches are being explored to try to improve on this.

(1) Postoperative adjuvant therapy. The standard of care is to use 6 months of adjuvant treatment with gemcitabine + capecitabine (referred to as the ESPAC4 trial). The median survival was 28 months (95% CI (23.5–31.5) for the combination of gemcitabine + capecitabine versus 25.5 months CI with (22.7–27.1) for the gemcitabine alone—hazard ratio 0.82 95% (0.6–0.98) ($p = 0.032$). Toxicities were manageable.

(2) A more experimental approach is the use of neoadjuvant chemotherapy (chemotherapy given before surgery) to try to shrink the tumor and normalize the patient's serum CA19-9 level. Studies of neoadjuvant chemotherapy are ongoing.

Locally Advanced Disease (30% of Patients)

For patients with locally advanced disease, the median survival is also quite poor (6–10 months) because many of the patients die with local problems (portal vein thrombosis with bleeding varices, obstruction, sepsis, etc.). The approach has been to try to reduce the bulk of the disease with use of radiation therapy plus chemotherapy or chemotherapy alone, hoping the disease could become resectable. No standard therapy has been agreed upon, but experimental approaches are applying some of the treatments that show promise in advanced metastatic disease.

Advanced Metastatic Disease (60% of Patients)

Only a few of the many phase III randomized trials in patients with advanced pancreatic cancer have led to meaningful increases in survival. We have learned that a regimen needs to have at least a 50% improvement in overall survival or 90% improvement in 1-year survival in a pilot trial to predict for any degree of success in large randomized phase III trials.

Patients with the best chance of receiving a benefit from treatment have a good performance status (functioning up and around at least 70% of the day), have a reasonable albumin level (≥3.0 g/dL), and a neutrophil/lymphocyte ratio of ≤5.0.

Single-agent gemcitabine achieves a median survival of 6 months and a 1-year survival rate of 18%. **Table 79-4** details three combination regimens that have further improved survival modestly. Median overall survival still ranges from 6 to 11 months. However, 1-year survival is now approaching 35% for these combination regimens with some long-term 4+ year survivors.

Liposomal irinotecan has been approved by the U.S. Food and Drug Administration (FDA) in combination with 5 fluorouracil + leucovorin for patients whose tumors have progressed on gemcitabine based on improved overall survival. PARP inhibitors have clinical activity against pancreatic cancers having mutations in BRCA2 or PALB2 (i.e., defective DNA repair proteins). In addition, tumors with microsatellite instability often have more mutations and such tumors appear to have a higher response rate to immunotherapy with checkpoint inhibitors, anti-PD-1 (pembrolizumab, nivolumab), and anti-PD-L1 antibodies.

Other Potential Factors Influencing Survival

Preclinical studies have suggested that vitamin D can inhibit the development and growth of cancer. In models of pancreatic cancer, synthetic analogs of vitamin D had an effect on both tumor cells and on the

tumor microenvironment. Clinical studies are conflicting as to whether circulating levels of plasma 25-hydroxyvitaminD (25[OH]D) affect the incidence of pancreatic cancer. However, patients with prediagnostic levels of 25(OH)D that are in the normal range have a longer survival than those who have reduced levels (35% lower hazard for death).

TABLE 79-4

Combination Chemotherapy Regimens that Have an Impact on Survival

Study Design (Author/Ref)	No. of Patients	Median Survival (Months)
Gemcitabine + erlotinib vs Gemcitabine, (Moore et al: J. Clin Oncol. 26:1960, 2007.)	569	6.24 vs 5.91 (HR 0.82; 95% CI 0.69–0.99, $p = 0.038$)
Folfirinox (folinic acid + 5FU + irinotecan + oxaliplatin) vs Gemcitabine, (Conroy et al: N Eng J Med 364:1817, 2011.)	342	11.1 vs 6.8 (HR 0.57; 95% CI 0.45–0.70, $p < 0.001$)
Nap-paclitaxel + gemcitabine vs gemcitabine, (Von Hoff et al: N Eng J Med 369:1691, 2013.)	861	8.5 vs 6.7 (HR 0.72; 95% CI 0.62–0.83, $p < 0.001$[a]

[a]The 2-year survival rate with this regimen is 9% and the 3+ year rate is 4%. Other studies have not reported on these parameters.

FUTURE DIRECTIONS

Death from pancreatic cancer is often due to progressive inanition. The metabolic consequences of this cancer are being examined. The tumor can be fatal at a modest level of tumor burden based on the profound metabolic effects. Other promising areas of investigation include addressing the florid stromal reaction around the tumor cells (believed to act as a physical barrier to drug delivery and as an immune sanctuary for the tumor cells). This attack on the stroma is being done with enzymatic (hyaluronidase) and other (antisuper enhancer genes) approaches. Also, utilization of hypomethylating and histone deacetylase inhibitors to correct epigenetic defects in the tumor microenvironment are under active study.

ACKNOWLEDGMENT

Thank you to Jennifer Byrne, BA, for assistance in the preparation of this chapter, and Drs. Elizabeth Washington, Ron Korn, and Haiyong Han and the American Joint Committee on Cancer for providing the figures.

FURTHER READING

Conroy T et al: FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. N Engl. J Med 364:1817, 2011. [PubMed: 21561347]

Hruban RJ et al: Genetic progression in the pancreatic ducts. Am J Pathol 156:1821, 2000. [PubMed: 10854204]

Ilic M, Ilic J: Epidemiology of pancreatic cancer. World J Gastroenterol 22:9694, 2016. [PubMed: 27956793]

Neoplolemus JP et al: Comparison of adjuvant gemcitabine and capecitabine with gemcitabine nanotherapy in patients with resected pancreatic cancer (ESPAC4) a multicenter, open labeled, randomized phase 3 trial. Lancet 389:1011, 2017. [PubMed: 28129987]

Rahib L et al: Evaluation of pancreatic cancer clinical trials and benchmarks for clinically meaningful future trials: A systemic review. JAMA Oncol 2:1209, 2016. [PubMed: 27270617]

Solomon S et al: Inherited pancreatic cancer syndromes. Cancer J 18:485, 2012. [PubMed: 23187834]

Von Hoff D et al: Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. N Engl J Med 369:1691, 2013. [PubMed: 24131140]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 80: Neuroendocrine Tumors of the Gastrointestinal Tract and Pancreas

Robert T. Jensen

GENERAL FEATURES OF GASTROINTESTINAL NEUROENDOCRINE TUMORS

Gastrointestinal (GI) neuroendocrine tumors (NETs) are tumors derived from the diffuse neuroendocrine system of the GI tract, which is composed of amine- and acid-producing cells with different hormonal profiles, depending on the site of origin. NETs of the GI tract share many features with other NETs throughout the body and were historically divided into GI-NETs (in the GI tract) (also frequently called *carcinoid tumors*) and pancreatic neuroendocrine tumors (pNETs), although in newer pathologic classifications they are all classified as NETs **(Table 80-1)**. These tumors originally were classified as APUDomas (for *a*mine *p*recursor *u*ptake and *d*ecarboxylation), as were pheochromocytomas, NETs in other locations, melanomas, and medullary thyroid carcinomas, because they share certain cytochemical features as well as various pathologic, biologic, and molecular features. It was originally proposed that APUDomas had a similar embryonic origin from neural crest cells, but it is now known that the peptide-secreting cells are not of neuroectodermal origin.

TABLE 80-1

Comparison of the Criteria for the Tumor Category in the ENETS AJCC TNM Classifications of Pancreatic and Appendiceal NETs (Top Panel) and the WHO/ENETS Grading and Classification (Bottom Panel)

A. TNM Classification		
	ENETS TNM	**AJCC/UICC TNM**
pNETs		
T1	Confined to pancreas, <2 cm	Confined to pancreas, <2 cm
T2	Confined to pancreas, 2–4 cm	Confined to pancreas, >2 cm
T3	Confined to pancreas, >4 cm, or invasion of duodenum or bile duct	Peripancreatic spread, but without major vascular invasion (truncus coeliacus, superior mesenteric artery)
T4	Invasion of adjacent organs or major vessels	Major vascular invasion
Appendiceal NETs		
T1	≤1 cm; invasion of muscularis propria	T1a, ≤1 cm; T1b, >1–2 cm
T2	≤2 cm and <3 mm invasion of subserosa/mesoappendix	>2–4 cm or invasion of cecum
T3	>2 cm or >3 mm invasion of subserosa/mesoappendix	>4 cm or invasion of ileum
T4	Invasion of peritoneum/other organs	Invasion of peritoneum/other organs

Abbreviations: AJCC, American Joint Committee on Cancer; ENETS, European Neuroendocrine Tumor Society; NET, neuroendocrine tumor; pNET, pancreatic neuroendocrine tumor; TNM, tumor, node, metastasis; UICC, International Union Against Cancer.

Source: Modified from DS Klimstra: Semin Oncol 40:23, 2013 and G Kloppel et al: Virchow Arch 456:595, 2010.

B. Grading			
CLASSIFICATION	**GRADE**	**MITOTIC COUNT (PER 10 HPF)**	**KI_{67} INDEX (%)**
NET	G1	<2	≤2
NET	G2	2–20	3–20
NEC (small cell and large cell)	G3	>20	>20

Abbreviations: HPF, high-power field; NEC, neuroendocrine carcinoma; NET, neuroendocrine tumor.

CLASSIFICATION/PATHOLOGY/TUMOR BIOLOGY OF NETs

NETs generally are composed of monotonous sheets of small round cells with uniform nuclei, and mitoses are uncommon. They can be frequently recognized on routine histology; however, these tumors are now recognized principally by their histologic staining patterns due to shared cellular proteins. Historically, silver staining was used, and tumors were classified as showing an argentaffin reaction if they took up and reduced silver or as being argyrophilic if they did not reduce it. Currently, immunocytochemical localization of chromogranins (A, B, C) and synaptophysin are routinely used. Chromogranins are acidic monomeric soluble proteins found in the large secretory granules. Synaptophysin is an integral membrane glycoprotein of 38,000 molecular weight found in small vesicles of neurons and NET. Chromogranin A is the most widely used.

Ultrastructurally, these tumors possess electron-dense neurosecretory granules and frequently contain small clear vesicles that correspond to synaptic vesicles of neurons. NETs synthesize numerous peptides, growth factors, and bioactive amines that may be ectopically secreted, giving rise to a specific clinical syndrome **(Table 80-2)**. The diagnosis of the specific syndrome requires the clinical features of the disease (Table 80-2) and cannot be made from the immunocytochemistry results alone. The presence or absence of a specific clinical syndrome also cannot be predicted from the immunocytochemistry alone.

TABLE 80-2

Gastrointestinal Neuroendocrine Tumor Syndromes

Name	Biologically Active Peptide(s) Secreted	Incidence (New Cases/10^6 Population/Year)	Tumor Location	Malignant, %	Associated with MEN 1, %	Main Symptoms/Signs
I. Established Specific Functional Syndromes						
A. Carcinoid syndrome due to GI-NET						
Carcinoid syndrome	Serotonin, possibly tachykinins, motilin, prostaglandins	0.5–2	Midgut (75–87%) Foregut (2–33%) Hindgut (1–8%) Unknown (2–15%)	95–100	Rare	Diarrhea (32–84%) Flushing (63–75%) Pain (10–34%) Asthma (4–18%) Heart disease (11–41%)
B. Well-established functional pNET syndromes						
Zollinger-Ellison syndrome	Gastrin	0.5–1.5	Duodenum (70%) Pancreas (25%) Other sites (5%)	60–90	20–25	Pain (79–100%) Diarrhea (30–75%) Esophageal symptoms (31–56%)
Insulinoma	Insulin	1–2	Pancreas (>99%)	<10	4–5	Hypoglycemic symptoms (100%)
VIPoma (Verner-Morrison syndrome, pancreatic cholera, WDHA)	Vasoactive intestinal peptide	0.05–0.2	Pancreas (90%, adult) Other (10%, neural, adrenal, periganglionic)	40–70	6	Diarrhea (90–100%) Hypokalemia (80–100%) Dehydration (83%)
Glucagonoma	Glucagon	0.01–0.1	Pancreas (100%)	50–80	1–20	Rash (67–90%) Glucose intolerance (38–87%) Weight loss (66–96%)
Somatostatinoma	Somatostatin	Rare	Pancreas (55%) Duodenum/jejunum (44%)	>70	45	Diabetes mellitus (63–90%) Cholelithiasis (65–90%) Diarrhea (35–90%)
GRFoma	Growth hormone–releasing hormone	Unknown	Pancreas (30%) Lung (54%) Jejunum (7%) Other (13%)	>60	16	Acromegaly (100%)
ACTHoma	ACTH	Rare	Pancreas (4–16% all ectopic Cushing's)	>95	Rare	Cushing's syndrome (100%)

Name	Biologically Active Peptide(s) Secreted	Incidence (New Cases/10^6 Population/Year)	Tumor Location	Malignant, %	Associated with MEN 1, %	Main Symptoms/Signs
pNET causing carcinoid syndrome	Serotonin, ?tachykinins	Rare (<100 cases)	Pancreas (<1% all carcinoids)	60–88	Rare	Same as carcinoid syndrome above
pNET causing hypercalcemia	PTHrP Others unknown	Rare	Pancreas (rare cause of hypercalcemia)	84	Rare	Abdominal pain due to hepatic metastases
II. Rare Specific Functional Syndromes						
pNET secreting renin	Renin	Rare	Pancreas	Unknown	No	Hypertension
pNET secreting luteinizing hormone	Luteinizing hormone	Rare	Pancreas	Unknown	No	Anovulation, virilization (female); reduced libido (male)
pNET secreting erythropoietin	Erythropoietin	Rare	Pancreas	100	No	Polycythemia
pNET secreting IGF-II	Insulin-like growth factor II	Rare	Pancreas	Unknown	No	Hypoglycemia
pNET secreting GLP-1	Glucagon-like peptide-1	Rare	Pancreas	Unknown	No	Hypoglycemia, diabetes
pNET secreting enteroglucagon	Enteroglucagon	Rare	Pancreas, small intestine	Unknown	Rare	Small intestinal hypertrophy, intestinal stasis, malabsorption
pNET secreting Cholecytokinin	Cholecytokinin	Rare	Pancreas	Unknown	No	Diarrhea, gallstones, peptic ulcer, weight loss
III. Possible Specific Functional pNET Syndromes						
pNET secreting calcitonin	Calcitonin	Rare	Pancreas (rare cause of hypercalcitonemia)	>80	16	Diarrhea (50%)
pNET secreting neurotensin	Neurotensin	Rare	Pancreas (100%)	Unknown	No	Motility disturbances, vascular symptoms
pNET secreting pancreatic polypeptide (PPoma)	Pancreatic polypeptide	1–2	Pancreas	>60	18–44	Watery diarrhea
pNET secreting ghrelin pNET secreting secretin	Ghrelin Secretin	Rare Rare	Pancreas Pancreas	Unknown Unknown	No unknown	Effects on appetite, body weight Watery diarrhea
IV. Nonfunctional Syndrome pNET						

Name	Biologically Active Peptide(s) Secreted	Incidence (New Cases/10^6 Population/Year)	Tumor Location	Malignant, %	Associated with MEN 1, %	Main Symptoms/Signs
PPoma/nonfunctional[a]	None	1–2	Pancreas (100%)	>60	18–44	Weight loss (30–90%) Abdominal mass (10–30%) Pain (30–95%)

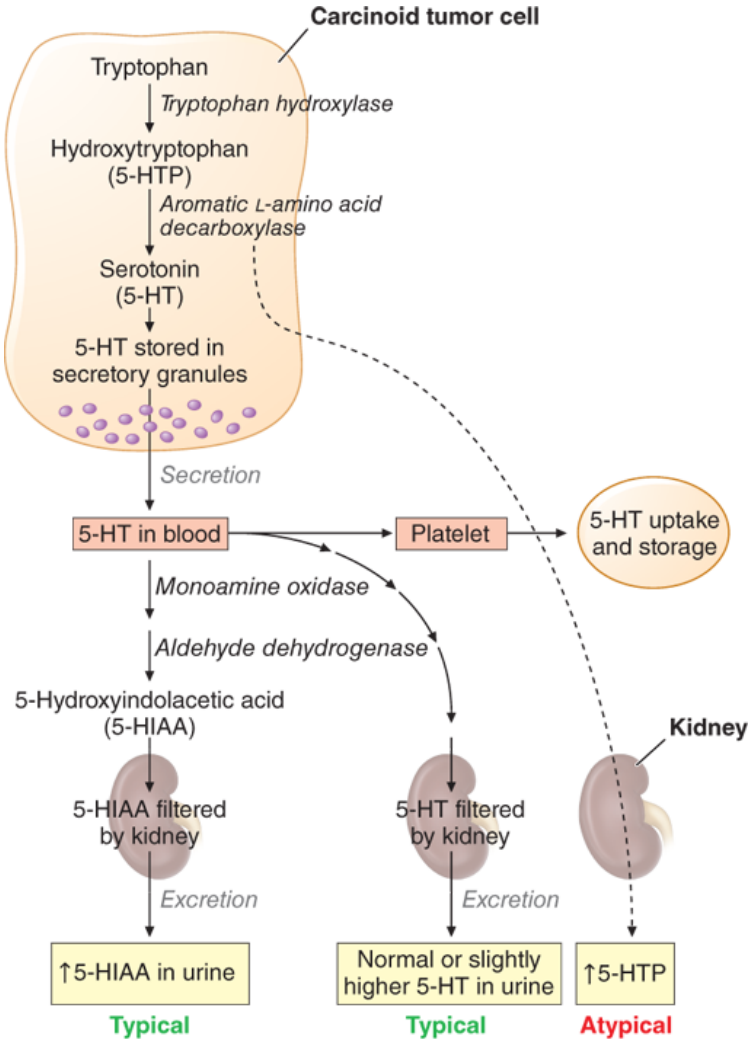
[a]Pancreatic polypeptide–secreting tumors (PPomas) are listed in two places because most authorities classify these as not associated with a specific hormonal syndrome (nonfunctional); however, rare cases of watery diarrhea proposed to be due to PPomas have been reported.

Abbreviations: ACTH, adrenocorticotropic hormone; GRFoma, growth hormone–releasing factor secreting pancreatic endocrine tumor; IGF-II, insulin-like growth factor II; MEN, multiple endocrine neoplasia; pNET, pancreatic neuroendocrine tumor; PPoma, tumor secreting pancreatic polypeptide; PTHrP, parathyroid hormone–related peptide; VIPoma, tumor secreting vasoactive intestinal peptide; WDHA, *w*atery *d*iarrhea, *h*ypokalemia, and *a*chlorhydria syndrome.

NETs of the GI tract (GI-NETs) have been classified according to their anatomic area of origin (i.e., foregut, midgut, hindgut) because tumors with similar areas of origin share functional manifestations, histochemistry, and secretory products **(Table 80-3)**. Foregut tumors generally have a low serotonin (5-HT) content, are argentaffin-negative but argyrophilic, occasionally secrete adrenocorticotropic hormone (ACTH) or 5-hydroxytryptophan (5-HTP), causing an atypical carcinoid syndrome **(Fig. 80-1)**; these are often multihormonal and may metastasize to bone. They may produce a clinical syndrome due to the secreted products. Midgut carcinoids are argentaffin-positive, have a high serotonin content, most frequently cause the typical carcinoid syndrome when they metastasize (Table 80-3, Fig. 80-1), release serotonin and tachykinins (substance P, neuropeptide K, substance K), rarely secrete 5-HTP or ACTH, and less commonly metastasize to bone. Hindgut carcinoids (rectum, transverse and descending colon) are argentaffin-negative, are often argyrophilic, rarely contain serotonin or cause the carcinoid syndrome (Fig. 80-1, Table 80-3), rarely secrete 5-HTP or ACTH, contain numerous peptides, and may metastasize to bone.

FIGURE 80-1

Synthesis, secretion, and metabolism of serotonin (5-HT) in patients with typical and atypical carcinoid syndromes. 5-HIAA, 5-hydroxyindolacetic acid.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 80-3

GI-NET (Carcinoid) Location, Frequency of Metastases, and Association with the Carcinoid Syndrome

Foregut			
Esophagus	<0.1	—	—
Stomach	4.6	10	9.5
Duodenum	2.0	—	3.4
Pancreas	0.7	71.9	20
Gallbladder	0.3	17.8	5
Bronchus, lung, trachea	27.9	5.7	13
Midgut			
Jejunum	1.8	58.4	9
Ileum	14.9		9
Meckel's diverticulum	0.5	—	13
Appendix	4.8	38.8	<1
Colon	8.6	51	5
Liver	0.4	32.2	—
Ovary	1.0	28	50
Testis	<0.1	—	50
Hindgut			
Rectum	13.6	3.9	—

Abbreviation: GI-NET, gastrointestinal neuroendocrine tumor.

Source: Location is from the PAN-SEER data (1973–1999), and incidence of metastases is from the SEER data (1992–1999), reported by IM Modlin et al: Cancer 97:934, 2003. Incidence of carcinoid syndrome is from 4349 cases studied from 1950 to 1971, reported by JD Godwin: Cancer 36:560, 1975.

However, the classification of GI NETs into foregut, midgut, or hindgut even though widely used, has not proved useful for prognostic or therapeutic purposes. More general classifications have been developed that allow NETs with similar features in different locations to be compared, have proven prognostic and tumor management value, and are now recommended in all recent guidelines and have become an essential requirement for management of these patients. The World Health Organization (WHO), European Neuroendocrine Tumor Society (ENETS), and the American Joint Committee on Cancer/International Union Against Cancer (AJCC/UICC) have developed classification systems (Table 80-1). Although there are some differences between these different classification systems, each uses similar information, and it is now recommended that the basic data underlying the classification be included in all standard pathology reports. These classification systems divide NETs from all sites into those that are well differentiated (low grade [G1] or intermediate grade [G2]) and those that are poorly differentiated (high grade [G3] divided into either small-cell carcinoma or large-cell neuroendocrine carcinoma [NEC]) (Table 80-1). In these classification systems, both pNETs and GI-NETs (carcinoids) are classified as NETs, and the old term of carcinoid is equivalent to well-differentiated NETs of the GI tract. These classification systems are based on not only the differentiation of the NET, but also a grading system assessing proliferative indices (Ki_{67} and the mitotic count) (Table 80-1). Based on these proliferative indices, NETs are classified as low grade (G1), intermediate grade (G2), or high grade (G3) (Table 80-1). In addition to the grading system, a TNM (TNM-tumor staging, T=primary size, N=regional lymph node involvement, M=distant metastases) classification has been proposed that is based on the level of tumor invasion, tumor size, and tumor extent (Table 80-1). Because of the proven prognostic value of these classification and grading systems, as well as the fact that NETs with different classifications/grades respond differently to treatments, these classification systems are now essential for the management of all NETs.

GI-NETs may or may not be associated with a specific functional syndrome (Table 80-2). In the case of pNETs the type of functional syndrome present is used to classify them into nine well-established specific functional syndromes (Table 80-2), seven additional very rare specific functional syndromes (less than five cases described), five possible specific functional syndromes (pNETs secreting calcitonin, neurotensin, pancreatic polypeptide [PP], ghrelin) (Table 80-2), and nonfunctional pNETs. Other functional hormonal syndromes due to nonpancreatic tumors (usually intra-abdominal in location) have been described only rarely and are not included in (Table 80-2). These include secretion by intestinal and ovarian tumors of peptide tyrosine tyrosine (PYY), which results in altered motility and constipation, and ovarian tumors secreting renin or aldosterone causing alterations in blood pressure or somatostatin causing diabetes or reactive hypoglycemia. Each of the functional syndromes listed in Table 80-2 is associated with symptoms due to the specific hormone released. In contrast, nonfunctional pNETs release no products that cause a specific clinical syndrome. “Nonfunctional” is a misnomer in the strict sense because those tumors frequently ectopically secrete a number of peptides (PP, chromogranin A, ghrelin, neurotensin, α subunits of human chorionic gonadotropin, and neuron-specific enolase); however, they cause no specific clinical syndrome. The symptoms caused by nonfunctional pNETs are entirely due to the tumor per se. pNETs frequently ectopically secrete PP (60–85%), neurotensin (30–67%), calcitonin (30–42%), and to a lesser degree, ghrelin (5–65%). Whereas a few studies have proposed that their secretion can cause a specific functional syndrome, most studies support the conclusion that their ectopic secretion is not associated with a specific clinical syndrome, and thus they are listed in Table 80-2 as possible clinical syndromes. Because a large proportion of nonfunctional pNETs (60–90%) secrete PP, these tumors are often referred to as PPomas (Table 80-2). pNETs can secrete secretin (secretinoma) producing watery diarrhea; however, only two possible cases are described.

GI-NETs (carcinoids) can occur in almost any GI tissue (Table 80-3); however, at present, most (70%) have their origin in one of three sites: bronchus, jejunoileum, or colon/rectum. In the past, GI-NET (carcinoids) most frequently were reported in the appendix (i.e., 40%); however, at present they account for <5% (Table 80-3). Overall, the GI tract is the most common site for NETs, accounting for 64%, with the respiratory tract a distant second at 28%. Both race and sex can affect the frequency as well as the distribution of GI-NETs (carcinoids). African Americans have a higher incidence of carcinoids. Race is particularly important for rectal carcinoids, which are found in 41% of Asians/Pacific Islanders with NETs compared to 32% of American Indians/Alaskan natives, 26% of African Americans, and 12% of white Americans. Females have a lower incidence of small intestinal and pancreatic carcinoids.

The term *pancreatic neuroendocrine* or *endocrine tumor*, although widely used and therefore retained here, is also a misnomer, strictly speaking, because these tumors can occur either almost entirely in the pancreas (insulinomas, glucagonomas, nonfunctional pNETs, pNETs causing hypercalcemia) or at both pancreatic and extrapancreatic sites (gastrinomas, VIPomas [vasoactive intestinal peptide], somatostatinomas, GRFomas [growth hormone–releasing factor]). pNETs are also called islet cell tumors; however, the use of this term is discouraged because it is not established that they originate from the islets, and many can occur at extrapancreatic sites.

In addition to these classification/grading systems, a number of other factors have been identified that provide important prognostic information that can guide treatment **(Table 80-4)**.

TABLE 80-4

Prognostic Factors in Neuroendocrine Tumors

I. Both GI-NETs (carcinoids) and pNETs
Symptomatic presentation (p <0.05)
Performance status (p <0.04)
Presence/extent of liver metastases (p <0.01)
Presence of lymph node metastases or lymph node positive ratio (p <0.001)
Development of bone or extrahepatic metastases (p <0.01)
Depth of invasion (p <0.001)
Rapid rate of tumor growth
Elevated serum alkaline phosphatase levels (p = 0.003)
Primary tumor site/site (p <0.005)
High serum chromogranin A level (p <0.01)
Presence of one or more circulating tumor cells (p <0.001)
Increased uptake on (18)F-FDG PET scanning
Low uptake (SUVmax) on (68)Ga-DOTANOC PET/CT scanning
Various histologic/molecular features
Tumor differentiation (p <0.001)
High growth indices (high Ki_{67} index, PCNA expression)
High mitotic counts (p <0.001)
Low progesterone receptor expression (p <0.001)
Necrosis present
Presence of cytokeratin 19 (p <0.02)
Vascular or perineural invasion (p <0.02)
Vessel density (low microvessel density, increased lymphatic density)
High CD10 metalloproteinase expression (in series with all grades of NETs)
Flow cytometric features (i.e., aneuploidy)
High VEGF expression (in low-grade or well-differentiated NETs only)
Abnormal expression of p53, Rb, SMADs
Loss of p27 expression(nuclear) (p <0.001)
WHO, ENETS, AJCC/UICC stage, and grade
Presence of a pNET rather than GI-NET associated with poorer prognosis (p = 0.0001)
Older age (p <0.01)
II. GI-NETs (carcinoids)
Location of primary: appendix < lung, rectum < small intestine < pancreas
Presence of carcinoid syndrome
Laboratory results (urinary 5-HIAA levels [p <0.01], plasma neuropeptide K [p <0.05], serum chromogranin A [p <0.01])
Presence of a second malignancy
Male sex (p <0.001)
Molecular findings (TGF-α expression [p <.05], chr 16q LOH or gain chr 4p [p <.05], gain in chr 14, loss of 3p13, loss of succinate dehydrogenase expression [ileal carcinoid], upregulation of Hoxc6), molecular profiling category (mutations, epigenetic changes, copy number-small intestinal NETs)
III. pNETs
Location of primary: duodenal (gastrinoma) better than pancreatic
Ha-*ras* oncogene or p53 overexpression
Female sex
MEN 1 syndrome absent
Presence of nonfunctional tumor (some studies, not all)
Various histologic features: IHC positivity for c-KIT, low cyclin B1 or ATM expression (p <0.01), loss of PTEN or of tuberous sclerosis-2 IHC, expression of fibroblast growth factor-13; high SSTR2 expression (p = .001);
Laboratory findings (increased chromogranin A in some studies; gastrinomas—increased gastrin level, increased CRP (p <0.001)

Molecular findings (increased HER2/*neu* expression [$p = 0.032$], chr 1q, 3p, 3q, or 6q LOH [$p = 0.0004$], EGF receptor overexpression [$p = 0.034$], gains in chr 7q, 17q, 17p, 20q; alterations in the VHL gene [deletion, methylation]; presence of FGFR4-G388R single-nucleotide polymorphism); loss of ATRX/DAXX or positive for alternative lengthening of telomeres ($p < 0.001$); high nuclear surviving expression ($p < 0.01$)
PHLDA3 LOH; altered miRNA expression (Inc miRNA-21, miRNA-196)

Abbreviations: 5-HIAA, 5-hydroxyindoleacetic acid; AJCC, American Joint Committee on Cancer; ATM, ataxia telangiectasia mutated kinase; ATRX, a-thalassemia/mental retardation X-linked; chr, chromosome; CRP, C-reactive protein; DAXX, death domain-associated protein; EGF, epidermal growth factor; FGFR, fibroblast growth factor receptor; GI-NET, gastrointestinal neuroendocrine tumor; IHC, immunohistochemistry; Ki_{67}, proliferation-associated nuclear antigen recognized by Ki_{67} monoclonal antibody; LOH, loss of heterozygosity; MEN, multiple endocrine neoplasia; NET, neuroendocrine tumors; PCNA, proliferating cell nuclear antigen; PHLDA3, Pleckstrin homology-like domain family A, member 3; pNET, pancreatic neuroendocrine tumor; PTEN, phosphatase and tensin homologue deleted from chromosome 10; Rb, retinoblastoma; SSTR2, somatostatin receptor subtype 2; TGF-α, transforming growth; factor α; TNM, tumor, node, metastasis; UICC, International Union Against Cancer; VEGF, vascular endothelial growth factor; WHO, World Health Organization.

The exact incidence of GI-NETs (carcinoids) or pNETs varies according to whether only symptomatic tumors or all tumors are considered. The incidence of clinically significant carcinoids is 7–13 cases/million population per year, whereas any malignant carcinoids at autopsy are reported in 21–84 cases/million population per year. The incidence of GI-NETs (carcinoids) is ~25–50 cases per million in the United States, which makes them less common than adenocarcinomas of the GI tract. However, their incidence has increased sixfold in the last 30 years. In an analysis of 35,825 GI-NETs (carcinoids) (2004) from the U.S. Surveillance, Epidemiology, and End Results (SEER) database which includes predominately malignant NETs, their incidence was 5.25/100,000 per year, and the 29-year prevalence was 35/100,000. Clinically significant pNETs have a prevalence of 10 cases/million population, with insulinomas, gastrinomas, and nonfunctional pNETs having an incidence of 0.5–2 cases/million population per year (Table 80-2). NF-pNETs are predominating, often making up 50–80% of the series, and increasingly found when asymptomatic. pNETs account for 1–10% of all tumors arising in the pancreas and 1.3% of tumors in the SEER database. VIPomas are 2–8 times less common, glucagonomas are 17–30 times less common, and somatostatinomas are the least common. In autopsy studies, 0.5–1.5% of all cases have a pNET; however, in <1 in 1000 cases was a functional tumor thought to occur.

Both GI-NETs (carcinoids) and pNETs commonly show malignant behavior (Tables 80-2 and 80-3). With pNETs, except for insulinomas in which <10% are malignant, 50–100% in different series are malignant. With GI-NETs (carcinoids), the percentage showing malignant behavior varies in different locations (Table 80-3). For the three most common sites of NET's occurrence, the incidence of metastases varies greatly from the jejunoileum (58%), lung/bronchus (6%), and rectum (4%) (Table 80-3). With both GI-NETs (carcinoids) and pNETs, a number of factors (Table 80-4) are important prognostic factors in determining survival and the aggressiveness of the tumor. Patients with pNETs (excluding insulinomas) generally have a poorer prognosis than do patients with GI-NETs (carcinoids). The presence of liver metastases is the single most important prognostic factor in single and multivariate analyses for both GI-NETs (carcinoids) and pNETs. Particularly important in the development of liver metastases is the size of the primary tumor. For example, with small intestinal carcinoids, which are the most common cause of the carcinoid syndrome due to metastatic disease in the liver (Table 80-2), metastases occur in 15–25% if the tumor is <1 cm in diameter, 58–80% if it is 1–2 cm in diameter, and >75% if it is >2 cm in diameter. Similar data exist for gastrinomas and other pNETs; the size of the primary tumor is an independent predictor of the development of liver metastases. The presence of lymph node metastases, their ratio or presence of extra-hepatic metastases; the depth of invasion; the rapid rate of growth; various histologic features (differentiation, mitotic rates, growth indices, vessel density, vascular endothelial growth factor [VEGF], CD10 metalloproteinase expression, abnormal expression of p53, retinoblastoma or SMAD, and low expression of p27 nuclear staining, low progesterone receptor expression); necrosis; presence of cytokeratin; elevated serum alkaline phosphatase levels; older age; presence of circulating tumor cells; increased uptake on ^{18}(F)-FDG-PET/CT scanning or low uptake (SUV/max) on ^{68}Ga-DOTANOC PET/CT scanning, and flow cytometric results, such as the presence of aneuploidy, are all important prognostic factors for the development of metastatic disease (Table 80-4). For patients with GI-NETs (carcinoids), additional associations with a worse prognosis include the development of the carcinoid syndrome (especially the development of carcinoid heart disease); male sex; the presence of a symptomatic tumor, a secondary malignancy, or greater increases in a number of tumor markers (5-hydroxyindolacetic acid [5-HIAA], neuropeptide K, chromogranin A), and the presence of various molecular features. With pNETs or gastrinomas, a worse prognosis is associated with female sex, overexpression of the Ha-*ras* oncogene or p53, the absence of Multiple Endocrine Neoplasia type 1 (MEN 1), presence of a NF-pNET, higher levels of various tumor markers (i.e., chromogranin A, gastrin, C-reactive protein), and presence of various histologic features (immunohistochemistry for c-KIT, low cyclin B1 or ATM, loss of PTEN/TSC-2, expression of fibroblast growth factor-13) and various molecular features (Table 80-4). The WHO, ENETs, and AJCC/UICC TNM classification systems and the grading systems (G1–G3) have important prognostic value and use in determining therapeutic management, that they are now generally routinely required.

A number of diseases due to various genetic disorders are associated with an increased incidence of NETs **(Table 80-5)**. Each one is caused by a loss of a possible tumor-suppressor gene. The most important is MEN 1, which is an autosomal dominant disorder due to a defect in a 10-exon gene on 11q13, which encodes for a 610-amino-acid nuclear protein, menin **(Chap. 381)**. Patients with MEN 1 develop hyperparathyroidism due to parathyroid hyperplasia in 95–100% of cases, pNETs in 80–100%, pituitary adenomas in 54–80%, adrenal adenomas in 27–36%, bronchial carcinoids in 8%, thymic carcinoids in 8% (predominately males), gastric carcinoids in 13–30% of patients with Zollinger-Ellison syndrome, skin tumors (angiofibromas [88%],

collagenomas [72%]), central nervous system (CNS) tumors (meningiomas, ependymomas, schwannomas [<8%]), and smooth-muscle tumors (leiomyomas, leiomyosarcomas [1–7%]). Among patients with MEN 1, 80–100% develop nonfunctional pNETs (most are microscopic with 0–13% large/symptomatic), and functional pNETs occur in 20–80% in different series, with a mean of 54% developing Zollinger-Ellison syndrome, 18% insulinomas, 3% glucagonomas, 3% VIPomas, and <1% GRFomas or somatostatinomas. MEN 1 is present in 20–25% of all patients with Zollinger-Ellison syndrome, 4% of patients with insulinomas, and a low percentage (<5%) of patients with other pNETs.

TABLE 80-5

Genetic Syndromes Associated with an Increased Incidence of Neuroendocrine Tumors (NETs) (GI-NETs [Carcinoids] or pNETs)

Syndrome	Location of Gene Mutation and Gene Product	NETs Seen/Frequency
Multiple endocrine neoplasia type 1 (MEN 1)	11q13 (encodes 610-amino-acid protein, menin)	80–100% develop pNETs (microscopic), 20–80% (clinical): (nonfunctional > gastrinoma > insulinoma)
		GI-NETs (Carcinoids): gastric (13–30%), bronchial/thymic (8%)
von Hippel–Lindau disease	3q25 (encodes 213-amino-acid protein)	12–17% develop pNETs (almost always nonfunctional)
von Recklinghausen's disease (neurofibromatosis 1 [NF-1])	17q11.2 (encodes 2485-amino-acid protein, neurofibromin)	0–10% develop pNETs, primarily duodenal somatostatinomas (usually nonfunctional)
		Rarely insulinoma, gastrinoma
Tuberous sclerosis	9q34 (TSCI) (encodes 1164-amino-acid protein, hamartin), 16p13 (TSC2) (encodes 1807-amino-acid protein, tuberin)	0–9% develop pNETs (nonfunctional and functional [insulinoma, gastrinoma])

Abbreviations: GI, gastrointestinal; PNETs, pancreatic neuroendocrine tumors.

Three phacomatoses associated with NETs are von Hippel–Lindau disease (VHL), von Recklinghausen's disease (neurofibromatosis type 1 [NF-1]), and tuberous sclerosis (Bourneville's disease) **(Table 80-5)**. VHL is an autosomal dominant disorder due to defects on chromosome 3p25, which encodes for a 213-amino-acid protein that interacts with the elongin family of proteins as a transcriptional regulator **(Chaps. 86, 309, 380, 381)**. In addition to cerebellar hemangioblastomas, renal cancer, and pheochromocytomas, 10–17% develop a pNET. Most are nonfunctional, although insulinomas and VIPomas have been reported. Patients with NF-1 (von Recklinghausen's disease) have defects in a gene on chromosome 17q11.2 that encodes for a 2845-amino-acid protein, neurofibromin, which functions in normal cells as a suppressor of the *ras* signaling cascade **(Chap. 86)**. Up to 10% of these patients develop an upper GI-NET (carcinoid), characteristically in the periampullary region (54%). Many are classified as somatostatinomas because they contain somatostatin immunocytochemically; however, they uncommonly secrete somatostatin and rarely produce a clinical somatostatinoma syndrome. NF-1 has rarely been associated with insulinomas and Zollinger-Ellison syndrome. NF-1 accounts for 48% of all duodenal somatostatinomas and 23% of all ampullary GI-NETs (carcinoids). Tuberous sclerosis is caused by mutations that alter either the 1164-amino-acid protein hamartin (TSC1) or the 1807-amino-acid protein tuberin (TSC2) **(Chap. 86)**. Both hamartin and tuberin interact in a pathway related to phosphatidylinositol 3-kinases and mammalian target of rapamycin (mTOR) signaling cascades. A few cases including nonfunctional and functional pNETs (insulinomas and gastrinomas) have been reported in these patients **(Table 80-6)**.

TABLE 80-6

Clinical Characteristics in Patients with Carcinoid Syndrome

	Percentage (Range)	
	At Presentation	**During Course of Disease**
Symptoms/signs		
Diarrhea	32–93%	68–100%
Flushing	23–100%	45–96%
Pain	10%	34%
Asthma/wheezing	4–14%	3–18%
Pellagra	0–7%	0–5%
None	12%	22%
Carcinoid heart disease present	11–40%	14–41%
Demographics		
Male	46–59%	46–61%
Age		
Mean	57 years	59.2 years
Range	25–79 years	18–91 years
Tumor location		
Foregut	5–14%	0–33%
Midgut	57–87%	60–100%
Hindgut	1–7%	0–8%
Unknown	2–21%	0–26%

A few other syndromes involving GI-NETs have been described due to various mutations, but no inherited cases have been reported. Mahvash disease is associated with the development of α-cell hyperplasia, hyperglucagonemia, without features of the glucagonoma syndrome, the development of NF pNETs, and is due to inactivating mutations of the human glucagon receptor. The polycythemia-paraganglioma-somatostatinoma (SSoma) syndrome involves childhood polycythemia, later the development of norepinephrine-producing paragangliomas (mean age 17 years) then SSomas (mean age 29 years) with gallbladder disease. In 28% of the patients, a somatic mutation in hypoxia-inducible factor 2 alphas is found.

Most GI NETs (carcinoids) are sporadic, although as discussed above gastric carcinoids occur (type 2) in MEN1 syndrome, duodenal carcinoids (SSomas) occur in NF-1 and a small percentage of patients (<3%) with small intestinal carcinoids have a familial form of the disease, which in one family was due to mutations in the inositol polyphosphate multikinase gene (IMPK).

Mutations in common oncogenes (*ras, myc, fos, src, jun*) or common tumor-suppressor genes (*p53*, retinoblastoma susceptibility gene) are not commonly found in either pNETs or GI-NETs (carcinoids). However, frequent (70%) gene amplifications in *MDM2, MDM4,* and *WIPI* inactivating the p53 pathway are noted in well-differentiated pNETs, and the retinoblastoma pathway is altered in the majority of pNETs. In addition to these genes, additional alterations that may be important in their pathogenesis include changes in the *MEN1* gene, *p16/MTS1* tumor-suppressor gene, and

DPC4/Smad4 gene; amplification of the HER-2/*neu* protooncogene; alterations in transcription factors (Hoxc6 [GI carcinoids]), growth factors, and their receptors; methylation of a number of genes that probably results in their inactivation; and deletions of unknown tumor-suppressor genes as well as gains in other unknown genes. The clinical antitumor activity of everolimus, an mTOR inhibitor, and sunitinib, a tyrosine kinase inhibitor (PDGFR, VEGFR1, VEGFR2, c-KIT, FLT-3), support the importance of the mTOR-AKT pathway and tyrosine kinase receptors in mediating growth of malignant NETs (especially pNETs). The importance of the mTOR pathway in pNET growth is further supported by the finding that a single-nucleotide polymorphism (FGFR4-G388R, in fibroblast growth factor receptor 4) affects selectivity to the mTOR inhibitor and can result in significantly higher risk of advanced pNET stage and liver metastases (Table 80-4). Comparative genomic hybridization, genome-wide allelotyping studies, and genome-wide single-nucleotide polymorphism analyses have shown that chromosomal losses and gains are common in pNETs and GI-NETs (carcinoids), but they differ between these two NETs, and some have prognostic significance (Table 80-4). Mutations in the *MEN1* gene are probably particularly important. Loss of heterozygosity at the MEN 1 locus on chromosome 11q13 is noted in 93% of sporadic pNETs (i.e., in patients without MEN 1) and in 26–75% of sporadic GI-NETs (carcinoids). Mutations in the *MEN1* gene are reported in 31–34% of sporadic gastrinomas. Exomic sequencing of sporadic pNETs found that the most frequently altered gene was *MEN1*, occurring in 44% of patients, followed by mutations in 43% of patients in genes encoding for two subunits of a transcription/chromatin remodeling complex consisting of DAXX (death-domain-associated protein) and ATRX (α-thalassemia/mental retardation syndrome X-linked) and in 15% of patients in the mTOR pathway. The presence of a number of these molecular alterations in pNETs or GI-NETs (carcinoids) correlates with tumor growth, tumor size, and disease extent or invasiveness and may have prognostic significance (Table 80-4).

GI NETs (carcinoids) have frequently loss of chromosome 18 (>60%) as well as losses on Chr 9p, 16q and chromosomal gains of 17q, 19p (57%) and lesser gains on 4q, 14q, and Chr 5, but the exact genes mediating possible effects on the tumor in these areas are still unclear. In contrast to pNETs, mutations in GI-NETs (carcinoids) are uncommon, and in one study of 180 small intestinal carcinoids, with exome and genome-sequencing analysis recurrent mutations were only observed in the CDKN1B gene (cyclin-dependent kinase inhibitor 1B [$p27^{Kip1}$]) in 8%. Integrative genomic analysis incorporating DNA methylation, show that small intestinal GI carcinoids commonly have epigenetic changes and three molecular subgroups with differing clinical course and outcomes have been identified (Table 80-4).

CHARACTERISTICS OF THE MOST COMMON GI-NETs (CARCINOIDS)

Appendiceal NETs (Carcinoids)

Appendiceal NETs (carcinoids) occur in 1 in every 200–300 appendectomies, usually in the appendiceal tip, have an incidence of 0.15/100,000 per year, comprise 2–5% of all GI-NETs (carcinoids), and comprise 32–80% of all appendiceal tumors. The mean age at diagnosis is 38–51 years. Most (i.e., >90%) are <1 cm in diameter without metastases in older studies, but more recently, 2–35% have had metastases (Table 80-3). In the SEER data of 1570 appendiceal carcinoids, 62% were localized, 27% had regional metastases, and 8% had distant metastases. The risk of metastases increases with size, with those <1 cm having a 0 to <10% risk of metastases and those >2 cm having a 25–44% risk. Besides tumor size, other important prognostic factors for metastases include basal NET location, invasion of mesoappendix, poor differentiation, advanced stage or WHO/ENETS classification, older age, and positive resection margins. The 5-year survival is 88–100% for patients with localized disease, 78–100% for patients with regional involvement, and 12–28% for patients with distal metastases. In patients with tumors <1 cm in diameter, the 5-year survival is 95–100%, whereas it is 29% if tumors are >2 cm in diameter. Most tumors are well-differentiated G1 tumors (87%) (Table 80-1), with the remainder primarily well-differentiated G2 tumors (13%); with poorly differentiated G3 tumors uncommon (<1%). Their percentage of the total number of carcinoids decreased from 43.9% (1950–1969) to 2.4% (1992–1999). Appendiceal goblet cell (GC) NETs (carcinoids)/carcinomas are a rare subtype (<5%) that are mixed adeno-neuroendocrine carcinomas. They are malignant and are thought to comprise a distinct entity; they frequently present with advanced disease and are recommended to be treated as adenocarcinomas, not carcinoid tumors.

SMALL INTESTINAL NETs (CARCINOIDS)

Small intestinal (SI) NETs (carcinoids) have a reported incidence of 0.67/100,000 in the United States, 0.32/100,000 in England, and 1.12/100,000 in Sweden and comprise >50% of all SI tumors. There is a male predominance (1.5:1), and race affects frequency, with a lower frequency in Asians and greater frequency in African Americans. The mean age of presentation is 52–63 years, with a wide range (1–93 years). Familial SI carcinoid families exist but are very uncommon. SI NETs (carcinoids) are frequently multiple; 9–18% occur in the jejunum, 70–80% are present in the ileum, and 70% occur within 6 cm (2.4 in.) of the ileocecal valve. Forty percent are <1 cm in diameter, 32% are 1–2 cm, and 29% are >2 cm. They are characteristically well differentiated; however, they are generally invasive, with 1.2% being intramucosal in location, 27% penetrating the submucosa, and 20% invading the muscularis propria. Metastases occur in a mean of 47–58% (range 20–100%). Liver metastases occur in 38%, to lymph nodes in 37% and more distant in 20–25%. They characteristically cause a marked fibrotic reaction, which can lead to intestinal obstruction. Tumor size is an important variable in the frequency of metastases. However, even small NETs (carcinoids) of the small intestine (<1 cm) have metastases in 15–25% of cases, whereas the proportion increases to 58–100% for tumors 1–2 cm in diameter. Carcinoids also occur in the duodenum, with 31% having metastases. Duodenal tumors <1 cm rarely metastasize, whereas 33% of those >2 cm had metastases. SI NETs (carcinoids) are the most common cause (60–87%) of the carcinoid syndrome (Table 80-6). Important prognostic factors are listed in Table 80-4, and particularly important are the tumor extent, proliferative index by grading, and stage (Table 80-1). SI NETs (carcinoid) are well differentiated in 99% of cases with 61–72% being G1 and 11–37% being G2. The

overall survival at 5 years is 55–75%; however, it varies markedly with disease extent, being 65–90% with localized disease, 66–72% with regional involvement, and 36–43% with distant disease.

Rectal NETs (Carcinoids)

Rectal NETs (carcinoids) comprise 27% of all GI-NETs (carcinoids) and 16% of all NETs and are increasing in frequency. In Europe they comprise 5–14% of all NETs and in some Asian series (Japan, China, Korea) they comprise 60–89% of all NETs. In the U.S. SEER data, they currently have an incidence of 0.86/100,000 per year (up from 0.2/100,000 per year in 1973) and represent 1–2% of all rectal tumors. They are found in ~1 in every 1500/2500 proctoscopies/colonoscopies or 0.05–0.07% of individuals undergoing these procedures. Nearly all occur between 4 and 13 cm above the dentate line. Most are small, with 66–80% being <1 cm in diameter, and rarely metastasize (5%). Tumors between 1 and 2 cm can metastasize in 5–30%, and those >2 cm, which are uncommon, in >70%. Most invade only to the submucosa (75%), with 2.1% confined to the mucosa, 10% to the muscular layer, and 5% to adjacent structures. Histologically, most are well differentiated (98%) with 72% ENETS/WHO grade G1 and 28% grade G2 (Table 80-1). Overall survival is 88%; however, it is very much dependent of the stage, with 5-year survival of 91% for localized disease, 36–49% for regional disease, and 20–32% for distant disease. Risk factors are listed in Table 80-4 and particularly include tumor size, depth of invasion, presence of metastases, differentiation, and recent TNM classification and grade.

Bronchial NETs (Carcinoids)

Bronchial NETs (carcinoids) comprise 25–33% of all well-differentiated NETs and 90% of all the poorly differentiated NETs found, likely due to a strong association with smoking. Their incidence ranges from 0.2 to 2/100,000 per year in the United States and European countries and is increasing at a rate of 6% per year. They are slightly more frequent in females and in whites compared with those of Hispanic/Asian/African descent, and are most commonly seen in the sixth decade of life, with a younger age of presentation for typical carcinoids (45 years) compared to atypical carcinoids (55 years).

A number of different classifications of bronchial GI-NETs (carcinoids) have been proposed. The principal factors used in classifying lung NETs include: morphology, presence or absence of necrosis, mitotic rate and size. In some studies, they are classified into four categories: typical carcinoid (also called bronchial carcinoid tumor, Kulchitsky cell carcinoma I [KCC-I]), atypical carcinoid (also called well-differentiated NEC [KC-II]), intermediate small-cell NEC, and small-cell neuroendocarcinoma (KC-III). Another proposed classification includes three categories of lung NETs: benign or low-grade malignant (typical carcinoid), low-grade malignant (atypical carcinoid), and high-grade malignant (poorly differentiated carcinoma of the large-cell or small-cell type). The WHO classification includes four general categories: typical carcinoid, atypical carcinoid, large-cell NEC, and small-cell carcinoma. The ratio of typical to atypical carcinoids is 8–10:1, with the typical carcinoids comprising 1–2% of lung tumors, atypical 0.1–0.2%, large-cell NETs 0.3%, and small-cell lung cancer 9.8% of all lung tumors. These different categories of lung NETs have different prognoses, varying from excellent for typical carcinoid to poor for small-cell NECs. The occurrence of large-cell and small-cell lung carcinoids, but not typical or atypical lung carcinoids, is related to tobacco use. The 5-year survival is very much influenced by the classification of the tumor, with survival of 92–100% for patients with a typical carcinoid, 61–88% with an atypical carcinoid, 13–57% with a large-cell neuroendocrine tumor, and 5% with a small-cell lung cancer. Typical/atypical lung carcinoids are generally well-differentiated with typical lung carcinoids sharing some homologies with G1 NETs, atypical sharing some homologies with G2 NETs of the GI tract, whereas small cell and large cell lung NECs are poorly differentiated and correspond to the G3 NEC category of the GI Tract (Table 80-1).

Gastric NET (Carcinoids)

Gastric NETs (carcinoids) account for 3 of every 1000 gastric neoplasms and 1.3–2% of all carcinoids, and their relative frequency has increased three- to fourfold over the last five decades (2.2% in 1950 to 9.6% in 2000–2007, SEER data). At present, it is unclear whether this increase is due to better detection with the increased use of upper GI endoscopy or to a true increase in incidence. Gastric NETs (carcinoids) are generally classified into three different categories, and this has important implications for pathogenesis, prognosis, and treatment. Each originates from gastric enterochromaffin-like (ECL) cells, one of the six types of gastric neuroendocrine cells, in the gastric mucosa. Two subtypes are associated with hypergastrinemic states, either chronic atrophic gastritis (type I) (70–80% of all gastric NETs [carcinoids]) or Zollinger-Ellison syndrome, which is almost always a part of the MEN 1 syndrome (type II) (5–6% of all cases). These tumors generally pursue a benign course, with type I uncommonly (<10%) associated with metastases, whereas type II tumors are slightly more aggressive, with 10–30% associated with metastases. Gastric carcinoids type 1 and type 2 are usually multiple, small, and infiltrate only to the submucosa. The third subtype of gastric NETs (carcinoids) (type III) (sporadic) occurs without hypergastrinemia (14–25% of all gastric carcinoids) and has an aggressive course, with 54–66% developing metastases. Sporadic carcinoids are usually single, large tumors; 50% have atypical histology, and they can be a cause of the carcinoid syndrome. Five-year survival is 99–100% in patients with type I, 60–90% in patients with type II, and 50% in patients with type III gastric NETs (carcinoids). Type 1 gastric carcinoids are usually grade G1 and well differentiated; type 2 are well-differentiated grade G1 or G2; and type 3 are characteristically NEC G3 with poor differentiation.

CLINICAL PRESENTATION OF NETs (CARCINOIDS)

GI/Lung NET (Carcinoid) Without the Carcinoid Syndrome

The age of patients at diagnosis ranges from 10 to 93 years, with a mean age of 63 years for the small intestine, 43–60 years for bronchial, and 66 years for the rectum. The presentation is diverse and is related to the site of origin and the extent of malignant spread. In the appendix, NETs (carcinoids) usually are found incidentally during surgery for suspected appendicitis. SI NETs (carcinoids) in the jejunoileum present with periodic abdominal pain (51%), intestinal obstruction with ileus/invagination (31%), an abdominal tumor (17%), or GI bleeding (11%). Because of the vagueness of the symptoms, the diagnosis usually is delayed ~2 years from onset of the symptoms, with a range up to 20 years. Duodenal, gastric, and rectal NETs (carcinoids) are most frequently found by chance at endoscopy. The most common symptoms of rectal carcinoids are melena/bleeding (39%), constipation (17%), and diarrhea (12%). Bronchial NETs (carcinoids) frequently are discovered as a lesion on a chest radiograph, and 31–43% of the patients are asymptomatic. Thymic NETs (carcinoids) present as anterior mediastinal masses, usually on chest radiograph or computed tomography (CT) scan. Ovarian and testicular NETs (carcinoids) usually present as masses discovered on physical examination or ultrasound. Metastatic NETs (carcinoids) in the liver frequently present as hepatomegaly in a patient who may have minimal symptoms and nearly normal liver function test results.

GI-NETs (CARCINOIDS) WITH SYSTEMIC SYMPTOMS DUE TO SECRETED PRODUCTS

GI/lung NETs (carcinoids) immunocytochemically can contain numerous GI peptides: gastrin, insulin, somatostatin, motilin, neurotensin, tachykinins (substance K, substance P, neuropeptide K), glucagon, gastrin-releasing peptide, vasoactive intestinal peptide (VIP), PP, ghrelin, other biologically active peptides (ACTH, calcitonin, growth hormone, GRF), prostaglandins, and bioactive amines (serotonin). These substances may or may not be released in sufficient amounts to cause symptoms. In various studies of patients with GI-NETs (carcinoids), elevated serum levels of PP were found in 43%, motilin in 14%, gastrin in 15%, and VIP in 6%. Foregut NETs (carcinoids) are more likely to produce various GI peptides than are midgut NETs (carcinoids). Ectopic ACTH production causing Cushing's syndrome is seen increasingly with foregut carcinoids (respiratory tract primarily) and, in some series, has been the most common cause of the ectopic ACTH syndrome, accounting for 64% of all cases. Acromegaly due to growth hormone–releasing factor release occurs with foregut NETs (carcinoids), as does the somatostatinoma syndrome, but rarely occurs with duodenal NETs (carcinoids). The most common systemic syndrome with GI-NETs (carcinoids) is the carcinoid syndrome, which is discussed in detail in the next section.

CARCINOID SYNDROME

Clinical Features

The cardinal features from a number of series at presentation as well as during the disease course are shown in Table 80-6.

Flushing and diarrhea are the two most common symptoms, occurring in a mean of 69–70% of patients initially and in up to 78% of patients during the course of the disease. The characteristic flush is of sudden onset; it is a deep red or violaceous erythema of the upper body, especially the neck and face, often associated with a feeling of warmth and occasionally associated with pruritus, lacrimation, diarrhea, or facial edema. Flushes may be precipitated by stress; alcohol; exercise; certain foods, such as cheese; or certain agents, such as catecholamines, pentagastrin; and serotonin reuptake inhibitors. Flushing episodes may be brief, lasting 2–5 min, especially initially, or may last hours, especially later in the disease course. Flushing usually is associated with metastatic midgut NETs (carcinoids) but can also occur with foregut NETs (carcinoids). With bronchial NETs (carcinoids), the flushes frequently are prolonged for hours to days, reddish in color, and associated with salivation, lacrimation, diaphoresis, diarrhea, and hypotension. The flush associated with gastric NETs (carcinoids) can also be reddish in color, but with a patchy distribution over the face and neck, although the classic flush seen with midgut NETs (carcinoids) can also be seen with gastric NETs (carcinoids). It may be provoked by food and have accompanying pruritus.

Diarrhea usually occurs with flushing (85% of cases). The diarrhea usually is described as watery, with 60% of patients having <1 L/d of diarrhea. Steatorrhea is present in 67%, and in 46%, it is >15 g/d (normal <7 g). Abdominal pain may be present with the diarrhea or independently in 10–34% of cases.

Cardiac manifestations occur initially in 11–40% (mean 26%) of patients with carcinoid syndrome and in 14–41% (mean 30%) at some time in the disease course. The cardiac disease is due to the formation of fibrotic plaques (composed of smooth-muscle cells, myofibroblasts, and elastic tissue) involving the endocardium, primarily on the right side, although lesions on the left side also occur occasionally (mean 11%, range 0–25), especially if a patent foramen ovale exists. The dense fibrous deposits are most commonly on the ventricular aspect of the tricuspid valve and less commonly on the pulmonary valve cusps. They can result in constriction of the valves, and pulmonic stenosis is usually predominant, whereas the tricuspid valve is often fixed open, resulting in regurgitation predominating. Overall, in patients with carcinoid heart disease, 90–100% have tricuspid insufficiency, 43–59% have tricuspid stenosis, 50–81% have pulmonary insufficiency, 25–59% have pulmonary stenosis, and 11% (0–25%) left-side lesions. Up to 80% of patients with cardiac lesions develop heart failure. Lesions on the left side are much less extensive, occur in 30% at autopsy, and most frequently affect the mitral valve. Up to 80% of patients with cardiac lesions have evidence of heart failure. At diagnosis in various series, 27–43% of patients are in New York Heart Association class I, 30–40% are in class II, 13–31% are in class III, and 3–12% are in class IV. At present, carcinoid heart disease is reported to be decreasing in frequency and severity, with mean occurrence in 20% of patients and occurrence in as few as 3–4% in some reports. Whether this decrease is due to the widespread use of somatostatin analogues, which control the release of bioactive agents thought involved in mediating the heart disease, is unclear.

Other clinical manifestations include wheezing or asthma-like symptoms (8–18%), pellagra-like skin lesions (2–25%), and impaired cognitive function. A variety of noncardiac problems due to increased fibrous tissue have been reported, including retroperitoneal fibrosis causing urethral obstruction, Peyronie's disease of the penis, intraabdominal fibrosis, and occlusion of the mesenteric arteries or veins.

Pathobiology

Carcinoid syndrome occurred in 8% of 8876 patients with GI-NETs (carcinoids), with a rate of 1.7–18.4% in different studies. It occurs only when sufficient concentrations of products secreted by the tumor reach the systemic circulation. In 91–100% of cases, this occurs after distant metastases to the liver. Rarely, primary GI-NETs (carcinoids) with nodal metastases with extensive retroperitoneal invasion, pNETs (carcinoids) with retroperitoneal lymph nodes, or NETs (carcinoids) of the lung, testis or ovary with direct access to the systemic circulation can cause the carcinoid syndrome without hepatic metastases. All GI-NETs (carcinoids) do not have the same propensity to metastasize and cause the carcinoid syndrome (Table 80-3). Midgut NETs (carcinoids) account for 57–67% of cases of carcinoid syndrome, foregut NETs (carcinoids) for 0–33%, hindgut for 0–8%, and an unknown primary location for 2–26% (Tables 80-3 and 80-6).

One of the main secretory products of GI-NETs (carcinoids) involved in the carcinoid syndrome is serotonin (5-HT) (Fig. 80-1), which is synthesized from tryptophan. Up to 50% of dietary tryptophan can be used in this synthetic pathway by tumor cells, and this can result in inadequate supplies for conversion to niacin; hence, some patients (2.5%) develop pellagra-like lesions. Serotonin has numerous biologic effects, including stimulating intestinal secretion with inhibition of absorption, stimulating increases in intestinal motility, and stimulating fibrogenesis. In various studies, 56–88% of all GI-NETs (carcinoids) were associated with serotonin overproduction; however, 12–26% of the patients did not have the carcinoid syndrome. In one study, platelet serotonin was elevated in 96% of patients with midgut NETs (carcinoids), 43% with foregut tumors, and 0% with hindgut tumors. In 90–100% of patients with the carcinoid syndrome, there is evidence of serotonin overproduction. Serotonin is thought to be predominantly responsible for the diarrhea. Patients with the carcinoid syndrome have increased colonic motility with a shortened transit time and possibly a secretory/absorptive alteration that is compatible with the known actions of serotonin in the gut mediated primarily through 5-HT_3 and, to a lesser degree, 5-HT_4 receptors. Serotonin receptor antagonists (especially 5-HT_3 antagonists) relieve the diarrhea in many, but not all, patients. A tryptophan 5-hydroxylase inhibitor, telotristat (LX-10310), which inhibits serotonin synthesis in peripheral tissues, caused a decrease in bowel movement frequency in 40–50% of patients with the carcinoid syndrome. Additional studies suggest that tachykinins may be important mediators of diarrhea in some patients. In one study, plasma tachykinin levels correlated with symptoms of diarrhea. Serotonin does not appear to be involved in the flushing in most patients because serotonin receptor antagonists do not relieve flushing. In patients with gastric carcinoids, the characteristic red, patchy pruritic flush is thought due to histamine release because H_1 and H_2 receptor antagonists can prevent it. Numerous studies have shown that tachykinins (substance P, neuropeptide K) are stored in GI-NETs (carcinoids) and released during flushing. However, some studies have demonstrated that octreotide can relieve the flushing induced by pentagastrin in these patients without altering the stimulated increase in plasma substance P, suggesting that other mediators must be involved in the flushing. A correlation between plasma tachykinin levels (but not substance P levels) and flushing has been reported. Prostaglandin release could be involved in mediating either the diarrhea or flush, but conflicting data exist. Both histamine and serotonin may be responsible for the wheezing as well as the fibrotic reactions involving the heart, causing Peyronie's disease and intraabdominal fibrosis.

The exact mechanism of the heart disease remains unclear, although increasing evidence supports a central role for serotonin. Patients with heart disease have higher plasma levels of neurokinin A, substance P, plasma atrial natriuretic peptide (ANP), pro-brain natriuretic peptide, chromogranin A, and activin A as well as higher urinary 5-HIAA excretion.

The valvular heart disease caused by the appetite-suppressant drugs dexfenfluramine and benfluorex is histologically indistinguishable from that observed in carcinoid disease. Furthermore, ergot-containing dopamine receptor agonists used for Parkinson's disease (pergolide, cabergoline) cause valvular heart disease that closely resembles that seen in the carcinoid syndrome. Furthermore, in animal studies, the formation of valvular plaques/fibrosis occurs after prolonged treatment with serotonin as well as in animals with a deficiency of the 5-HIAA transporter gene, which results in an inability to inactivate serotonin. Metabolites of fenfluramine, as well as the dopamine receptor agonists, have high affinity for serotonin receptor subtype 5-HT_{2B} receptors, whose activation is known to cause fibroblast mitogenesis. Serotonin receptor subtypes 5-$HT_{1B,1D,2A,2B}$ normally are expressed in human heart valve interstitial cells. High levels of 5-HT_{2B} receptors are known to occur in heart valves and occur in cardiac fibroblasts and cardiomyocytes. Studies of cultured interstitial cells from human cardiac valves have demonstrated that these valvulopathic drugs induce mitogenesis by activating 5-HT_{2B} receptors and stimulating upregulation of transforming growth factor β and collagen biosynthesis. These observations support the conclusion that serotonin overproduction by GI-NETs (carcinoids) is important in mediating the valvular changes, possibly by activating 5-HT_{2B} receptors in the endocardium. Both the magnitude of serotonin overproduction and prior chemotherapy are important predictors of progression of the heart disease, whereas patients with high plasma levels of ANP have a worse prognosis. Plasma connective tissue growth factor levels are elevated in many fibrotic conditions; elevated levels occur in patients with carcinoid heart disease and correlate with the presence of right ventricular dysfunction and the extent of valvular regurgitation in patients with GI-NETs (carcinoids).

Patients may develop either a typical or, rarely, an atypical carcinoid syndrome (Fig. 80-1). In patients with the typical form, which characteristically is caused by midgut NETs (carcinoids), the conversion of tryptophan to 5-HTP by tryptophan hydroxylase is the rate-limiting step (Fig. 80-1). Once 5-HTP is formed, it is rapidly converted to 5-HT and stored in secretory granules of the tumor or in platelets. A small amount remains in plasma and is converted to 5-HIAA, which appears in large amounts in the urine. These patients have an expanded serotonin pool size, increased blood and platelet serotonin, and increased urinary 5-HIAA. Some GI-NETs (carcinoids) cause an atypical carcinoid syndrome that is thought to be due to a deficiency in the enzyme dopa decarboxylase; thus, 5-HTP cannot be converted to 5-HT (serotonin), and 5-HTP is secreted into the bloodstream (Fig. 80-1). In these patients, plasma serotonin levels are normal but urinary levels may be increased because some 5-HTP is converted to 5-HT in the kidney. Characteristically, urinary 5-HTP and 5-HT are increased, but urinary 5-HIAA levels are only slightly elevated. Foregut carcinoids are the most likely to cause an atypical carcinoid syndrome; however, they also can cause a typical carcinoid syndrome.

One of the most immediate life-threatening complications of the carcinoid syndrome is the development of a carcinoid crisis. This is more common in patients who have intense symptoms or have greatly increased urinary 5-HIAA levels (i.e., >200 mg/d). The crisis may occur spontaneously; however, it is usually provoked by procedures such as anesthesia, chemotherapy, surgery, biopsy, endoscopy, or radiologic examinations such as during biopsies, hepatic artery embolization, and vessel catheterization. It can be provoked by stress or procedures as mild as repeated palpation of the tumor during physical examination. Patients develop intense flushing, diarrhea, abdominal pain, cardiac abnormalities including tachycardia, hypertension, or hypotension, and confusion or stupor. If not adequately treated, this can be a terminal event.

DIAGNOSIS OF THE CARCINOID SYNDROME AND GI-NETs (CARCINOIDS)

The diagnosis of carcinoid syndrome relies on measurement of urinary or plasma serotonin or its metabolites in the plasma or urine. The measurement of urinary 5-HIAA is used most frequently. False-positive elevations may occur if the patient is eating serotonin-rich foods such as bananas, pineapples, walnuts, pecans, avocados, or hickory nuts or is taking certain medications (cough syrup containing guaifenesin, acetaminophen, salicylates, serotonin reuptake inhibitors, or L-dopa). The normal range for daily urinary 5-HIAA excretion is 2–8 mg/d. Serotonin overproduction was noted in 92% of patients with carcinoid syndrome in one study, and in another study, 5-HIAA had 73% sensitivity and 100% specificity for carcinoid syndrome. Serotonin overproduction is *not* synonymous with the presence of clinical carcinoid syndrome because 12–26% of patients with serotonin overproduction do not have clinical evidence of the carcinoid syndrome.

Most physicians use only the urinary 24-h 5-HIAA excretion rate; even though a recent study shows an overnight urinary collection is just as accurate. Assessment of plasma and platelet serotonin levels and plasma 5-HIAA, if available, may provide additional information and/or substitute for the 24 h urinary 5-HIAA study. Platelet serotonin levels are more sensitive than urinary 5-HIAA but are not generally available. A single plasma 5-HIAA determination was found to have similar sensitivity/specificity to that with the 24-h urinary 5-HIAA assessment, suggesting this could replace the standard urinary collection because of its greater convenience and avoidance of incomplete or improper collections. It, however, could be affected by renal disease. Because patients with foregut NETs (carcinoids) may produce an atypical carcinoid syndrome, if this syndrome is suspected and the urinary 5-HIAA is minimally elevated or normal, other urinary metabolites of tryptophan, such as 5-HTP and 5-HT, should be measured (Fig. 80-1).

Flushing occurs in a number of other diseases, including systemic mastocytosis, chronic myeloid leukemia with increased histamine release, menopause, reactions to alcohol or glutamate, and side effects of chlorpropamide, calcium channel blockers, and nicotinic acid. None of these conditions causes increased urinary 5-HIAA.

The diagnosis of carcinoid tumor can be suggested by the carcinoid syndrome, recurrent abdominal symptoms in a healthy-appearing individual, or the discovery of hepatomegaly or hepatic metastases associated with minimal symptoms. Ileal NETs (carcinoids), which make up 25% of all clinically detected carcinoids, should be suspected in patients with bowel obstruction, abdominal pain, flushing, or diarrhea.

Serum chromogranin A levels are elevated in 56–100% of patients with GI-NETs (carcinoids), and the level correlates with tumor bulk. Serum chromogranin A levels are not specific for GI-NETs (carcinoids) because they are also elevated in patients with pNETs and other NETs. Furthermore, a major problem is caused by potent acid antisecretory drugs such as proton pump inhibitors (omeprazole and related drugs) because they almost invariably cause elevation of plasma chromogranin A levels; the elevation occurs rapidly (3–5 days) with continued use, and the elevated levels overlap with the levels seen in many patients with NETs. Plasma neuron-specific enolase levels are also used as a marker of GI-NETs (carcinoids) but are less sensitive than chromogranin A, being increased in only 17–47% of patients. Newer markers have been proposed including pancreastatin (a chromogranin A breakdown product), and activin A. The former is not affected by proton pump inhibitors; however, its sensitivity and specificity are not established. Plasma activin elevations are reported to correlate with the presence of cardiac disease with a sensitivity of 87% and specificity of 57%. Plasma levels of N-terminal pro brain natriuretic peptide moderately correlate with carcinoid heart disease severity.

TREATMENT

TREATMENT

Carcinoid Syndrome and Nonmetastatic Gastrointestinal Neuroendocrine Tumors (Carcinoids)

CARCINOID SYNDROME

Treatment includes avoiding conditions that precipitate flushing, dietary supplementation with nicotinamide, treatment of heart failure with diuretics, treatment of wheezing with oral bronchodilators, and control of the diarrhea with antidiarrheal agents such as loperamide and diphenoxylate. If patients still have symptoms, somatostatin analogues or less frequently, serotonin receptor antagonists, are the drugs of choice **(Fig. 80-2)**. An additional point dealt with in later sections, is the fact that most patients who develop the carcinoid syndrome have metastatic disease to the liver. Numerous antitumor therapies (liver-directed therapies, PRRT, surgery, chemotherapy/targeted drug therapies) also can ameliorate the severity of the carcinoid syndrome.

There are 14 subclasses of serotonin receptors, and antagonists for many are not available. The 5-HT_1 and 5-HT_2 receptor antagonists methysergide, cyproheptadine, and ketanserin have all been used to control the diarrhea but usually do not decrease flushing. The use of methysergide is limited because it can cause or enhance retroperitoneal fibrosis. Ketanserin diminishes diarrhea in 30–100% of patients. 5-HT_3 receptor antagonists (ondansetron, tropisetron, alosetron) can control diarrhea and nausea in up to 100% of patients and occasionally ameliorate the flushing. A combination of histamine H_1 and H_2 receptor antagonists (i.e., diphenhydramine and cimetidine or ranitidine) may control flushing in patients with foregut carcinoids. A phase 3 prospective, double-blind study provides evidence the peripheral tryptophan hydroxylase inhibitor, telotristat, will be useful to control the diarrhea in many of these patients.

Synthetic analogues of somatostatin (octreotide, lanreotide) are now the most widely used agents to control the symptoms of patients with carcinoid syndrome (Figs. 80-1 and 80-2). These drugs are effective at relieving symptoms and decreasing urinary 5-HIAA levels in patients with this syndrome. Octreotide-LAR (10–30 mg i.m., monthly) and lanreotide-SR/autogel (Somatuline) (60–120 mg sc-deep, monthly) (sustained-release formulations allowing monthly injections) (Fig. 80-2), control symptoms in 74 and 68% of patients, respectively, with carcinoid syndrome and show a biochemical response in 51 and 64%, respectively. Patients with mild to moderate symptoms usually are treated initially with octreotide 50–100 μg SC every 8 h or lower doses or low doses of the long-acting formulations, and then receive higher doses as needed of the long-acting monthly depot forms (octreotide-LAR or lanreotide-autogel). Forty percent of patients escape control after a median time of 4 months, and the depot dosage may have to be increased as well as supplemented with the shorter-acting formulation, SC octreotide. Pasireotide (SOM230) is a somatostatin analogue with broader selectivity (high-affinity somatostatin receptors [sst_1, sst_2, sst_3, sst_5]) than octreotide/lanreotide (sst_2, sst_5). In a phase II study of patients with refractory carcinoid syndrome, pasireotide controlled symptoms in 27%.

Carcinoid heart disease is associated with a decreased mean survival (3.8 years), and therefore, it should be sought for and carefully assessed in all patients with carcinoid syndrome. Transthoracic echocardiography remains a key element in establishing the diagnosis of carcinoid heart disease and determining the extent and type of cardiac abnormalities. Treatment with diuretics and somatostatin analogues can reduce the negative hemodynamic effects and secondary heart failure. It remains unclear whether long-term treatment with these drugs or with the tryptophan hydroxylase inhibitor, telotristat, when it becomes available, will decrease the progression of carcinoid heart disease. Balloon valvuloplasty for stenotic valves or cardiac valve surgery may be required.

To prevent as well as treat patients with carcinoid crises, somatostatin analogues are recommended, although there is controversy of how effective they are and what dosage should be use. To prevent carcinoid crises development, treatment with somatostatin analogues is recommended prior to the possible precipitating event such as surgery, anesthesia, chemotherapy, and stress. It is generally recommended that octreotide 150–250 μg SC every 6–8 h be used 24–48 h before anesthesia and then continued throughout the procedure. Another commonly used protocol is to use 100 μg/h by continuous infusion with or without a preoperative bolus.

Currently, sustained-release preparations of both octreotide (octreotide-LAR [long-acting release], 10, 20, 30 mg) and lanreotide (lanreotide-PR [prolonged release, lanreotide-autogel], 60, 90, 120 mg) are available and widely used because their use greatly facilitates long-term treatment. Octreotide-LAR (30 mg/month) gives a plasma level ≥1 ng/mL for 25 days, whereas this requires 3–6 injections a day of the non-sustained-release form. Lanreotide-autogel (Somatuline) is given every 4–6 weeks.

Short-term side effects occur in up to one-half of patients. Pain at the injection site and side effects related to the GI tract (59% discomfort, 15% nausea, diarrhea) are the most common. They are usually short-lived and do not interrupt treatment. Important long-term side effects include gallstone formation, steatorrhea, and deterioration in glucose tolerance. The overall incidence of gallstones/biliary sludge varies from 5 to 66%, with 7% having symptomatic disease that required surgical treatment in one study.

Interferon α is reported to be effective in controlling symptoms of the carcinoid syndrome either alone or combined with hepatic artery embolization. With interferon α alone, the clinical response rate is 30–70%, and with interferon α with hepatic artery embolization, diarrhea was controlled for 1 year in 43% and flushing was controlled in 86%. Side effects develop in almost all patients, with the most frequent being a flu-like syndrome (80–100%), followed by anorexia and fatigue, even though these frequently improve with continued treatment. Other more severe side effects include bone marrow toxicity, hepatotoxicity, autoimmune disorders, and rarely CNS side effects (depression, mental disorders, visual problems).

Hepatic artery embolization alone or with chemotherapy (chemoembolization) has been used to control the symptoms of carcinoid syndrome. Embolization alone is reported to control symptoms in up to 76% of patients, and chemoembolization (5-fluorouracil, doxorubicin, cisplatin, mitomycin) controls symptoms in 60–75% of patients. Hepatic artery embolization can have major side effects, including nausea, vomiting, pain, and fever. In two studies, 5–7% of patients died from complications of hepatic artery occlusion.

Other drugs have been used successfully in small numbers of patients to control the symptoms of carcinoid syndrome. Parachlorophenylanine can inhibit tryptophan hydroxylase and therefore the conversion of tryptophan to 5-HTP. However, its severe side effects, including psychiatric disturbances, make it intolerable for long-term use. α-Methyldopa inhibits the conversion of 5-HTP to 5-HT, but its effects are only partial.

Peptide radioreceptor therapy (PRRT; using radiotherapy with radiolabeled somatostatin analogues), cytoreductive surgery, the use of radiolabeled microspheres, and other methods for treatment of advanced metastatic disease can facilitate control of the carcinoid syndrome (see below).

GI-NETs (CARCINOIDS) (NONMETASTATIC)

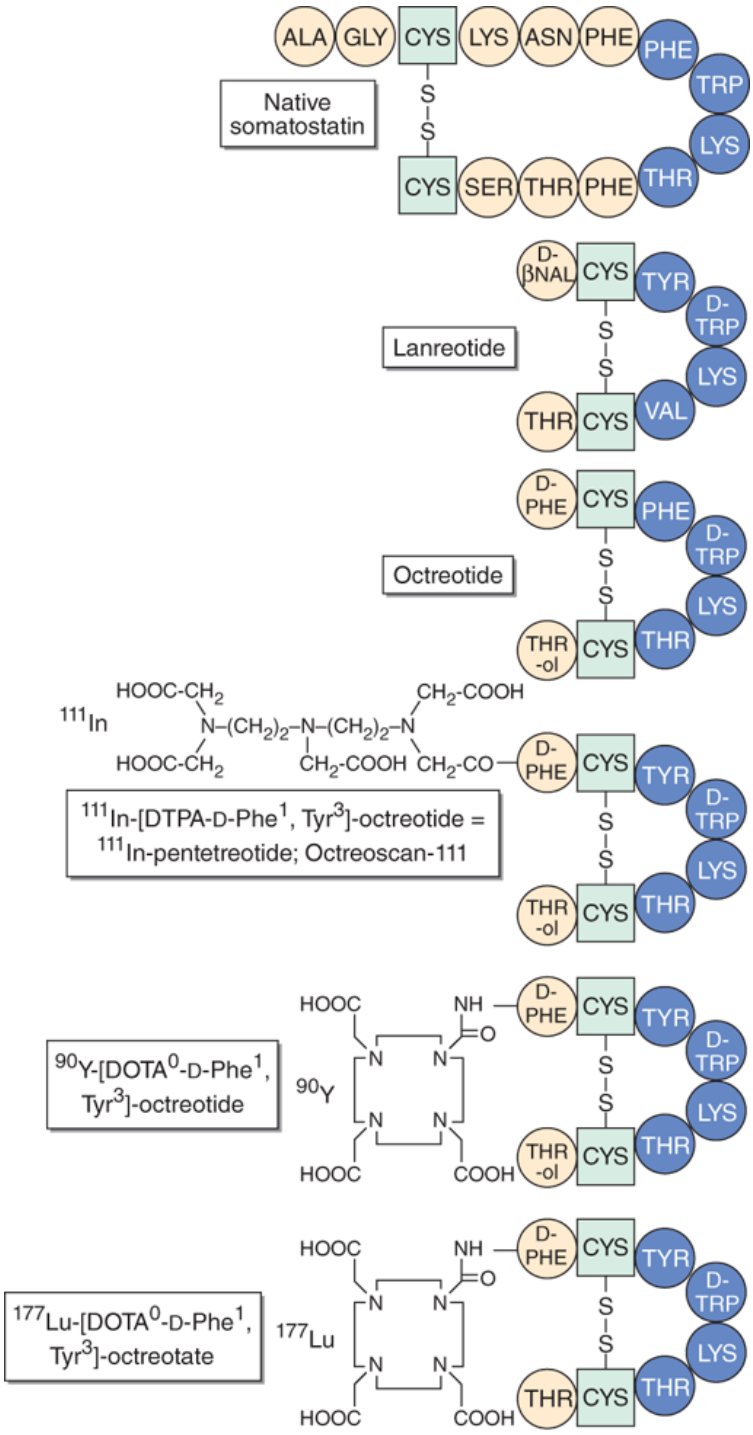
Surgery is the only potentially curative therapy. Because with most GI-NETs (carcinoids), the probability of metastatic disease increases with increasing size, in most guidelines, the therapeutic approach is determined accordingly. Furthermore, the grade of the tumor is having an increasingly important role in determining the therapeutic approach. With well-differentiated (G1/G2) GI-NETs the size of the primary NET plays an important role. With appendiceal NETs (carcinoids) <1 cm, simple appendectomy was curative in 103 patients followed for up to 35 years. With rectal NETs (carcinoids) <1 cm, local resection is curative. With SI NETs (carcinoids) <1 cm, there is not complete agreement. Because 15–69% of SI NETs (carcinoids) this size have metastases in different studies, most recommend a wide resection with en bloc resection of the adjacent lymph-bearing mesentery. If the tumor is >2 cm for rectal, appendiceal, or SI NETs (carcinoids), a full cancer operation should be done. This includes a right hemicolectomy for appendiceal NETs (carcinoids), an abdominoperineal resection or low anterior resection for rectal NETs (carcinoids), and an en bloc resection of adjacent lymph nodes for SI NETs (carcinoids). For appendiceal NETs (carcinoids) 1–2 cm in diameter, a simple appendectomy is proposed by some, whereas others favor a formal right hemicolectomy. For 1–2 cm rectal NETs (carcinoids), it is recommended that a wide, local, full-thickness excision be performed.

With well-differentiated (G1/G2) type I or II gastric NETs (carcinoids), which are usually <1 cm, endoscopic removal is recommended. In type I or II gastric carcinoids, if the tumor is >2 cm or if there is local invasion, some recommend total gastrectomy, whereas others recommend antrectomy in type I to reduce the hypergastrinemia, which has led to regression of the carcinoids in a number of studies. For types I and II gastric NETs (carcinoids) of 1–2 cm, there is no agreement, with some recommending endoscopic treatment followed by chronic somatostatin treatment and careful follow-up and others recommending surgical treatment. With type III gastric NETs (carcinoids) >2 cm, excision and regional lymph node clearance are recommended. Most tumors <1 cm are treated endoscopically. Type 1 and 2 gastric carcinoids tend to recur after endoscopic treatments so patients need to continue to be followed. Treatment of type 1 or 2 gastric carcinoids using a CCK_B (gastrin) receptor antagonist, netazepide (not yet FDA approved) decreased the size and number of gastric carcinoids. However, netazepide needed to be continued or they would return. Poorly differentiated G3 carcinoids of the GI tract are treated like G3 tumors in other locations, which involve primarily chemotherapy and will discussed in a later section on treatment of advanced/aggressive disease.

Resection of isolated or limited hepatic metastases may be beneficial and will be discussed in a later section on treatment of advanced disease.

FIGURE 80-2

Structure of somatostatin and synthetic analogues used for diagnostic or therapeutic indications.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PANCREATIC NEUROENDOCRINE TUMORS (pNETs)

Functional pNETs (F-pNETs) usually present clinically with symptoms due to the hormone-excess state (Table 80-2). Only late in the course of the disease does the tumor per se cause prominent symptoms such as abdominal pain. In contrast, all the symptoms due to nonfunctional pNETs (NF-pNET) are due to the tumor per se. The overall result of this is that some F-pNETs may present with severe symptoms with a small or undetectable primary tumor, whereas NF-pNETs usually present late in the disease course with large tumors, which are frequently metastatic. The mean delay between onset of continuous symptoms and diagnosis of a F-pNET syndrome is 4–7 years. Therefore, the diagnoses frequently are missed for extended periods.

TREATMENT

TREATMENT

Pancreatic Neuroendocrine Tumor (General Points)

Treatment of pNETs requires two different strategies. First, treatment must be directed at the hormone-excess state such as the gastric acid hypersecretion in gastrinomas or the hypoglycemia in insulinomas. Ectopic hormone secretion usually causes the presenting symptoms and can cause life-threatening complications. Second, with all the tumors except insulinomas, >50% are malignant (Table 80-2); therefore, treatment must also be directed against the tumor per se. Because in many patients these tumors are not surgically curable due to the presence of advanced disease at diagnosis, surgical resection for cure, which addresses both treatment aspects, is often not possible.

GASTRINOMA (ZOLLINGER-ELLISON SYNDROME)

A gastrinoma is an NET that secretes gastrin; the resultant hypergastrinemia causes gastric acid hypersecretion (Zollinger-Ellison syndrome [ZES]). The chronic hypergastrinemia results in marked gastric acid hypersecretion and growth of the gastric mucosa with increased numbers of parietal cells and proliferation of gastric ECL cells. The gastric acid hypersecretion characteristically causes peptic ulcer disease (PUD), often refractory and severe, as well as diarrhea. The most common presenting symptoms are abdominal pain (70–100%), diarrhea (37–73%), gastroesophageal reflux disease (GERD) (30–35%), and 10–20% of patients have diarrhea only. Although peptic ulcers may occur in unusual locations, most patients have a typical duodenal ulcer. Important observations that should suggest this diagnosis include PUD with diarrhea; PUD in an unusual location or with multiple ulcers; PUD refractory to treatment or persistent; PUD associated with prominent gastric folds; PUD associated with findings suggestive of MEN 1 (endocrinopathy, family history of ulcer or endocrinopathy, nephrolithiases); and PUD without *Helicobacter pylori* present. *H. pylori* is present in >90% of idiopathic peptic ulcers but is present in <50% of patients with gastrinomas. Chronic unexplained diarrhea also should suggest ZES.

Approximately 20–25% of patients with ZES have MEN 1 (MEN1/ZES), and in most cases, hyperparathyroidism is present before the ZES develops. In older studies it was generally reported that almost all MEN/ZES presented with the hyperparathyroidism, but in a number of recent series up to one-third of these patients present with the ZES, and while the hyperparathyroidism is present it may be mild and difficult to diagnose without appropriate testing. These patients are treated differently from those without MEN 1 (sporadic ZES); therefore, MEN 1 should be sought in all patients with ZES by family history and by measuring plasma ionized calcium and prolactin levels and plasma hormone levels (parathormone, growth hormone).

Most gastrinomas (50–90%) in sporadic ZES are present in the duodenum, followed by the pancreas (10–40%) and other intraabdominal sites (mesentery, lymph nodes, biliary tract, liver, stomach, ovary). Rarely, the tumor may involve extraabdominal sites (heart, lung cancer). In MEN 1/ZES the gastrinomas are also usually in the duodenum (80–100%), followed by the pancreas (0–20%), and are almost always multiple. About 60–90% of gastrinomas are malignant (Table 80-2) with metastatic spread to lymph nodes and liver. Distant metastases to bone occur in 12–30% of patients with liver metastases.

Diagnosis

The diagnosis of ZES requires the demonstration of inappropriate fasting hypergastrinemia, usually by demonstrating hypergastrinemia occurring with an increased basal gastric acid output (BAO) (hyperchlorhydria). More than 98% of patients with ZES have fasting hypergastrinemia, although in 40–60% the level may be elevated less than tenfold. Therefore, when the diagnosis is suspected, a fasting gastrin is usually the initial test performed. It is important to remember that potent gastric acid suppressant drugs such as proton pump inhibitors (PPIs) (omeprazole, esomeprazole, pantoprazole, lansoprazole, rabeprazole) can suppress acid secretion sufficiently to cause hypergastrinemia; because of their prolonged duration of action, these drugs have to be tapered or frequently discontinued for a week before the gastrin determination. Withdrawal of PPIs should be performed carefully because PUD complications can rapidly develop in some patients and is best done in consultation with GI units with experience in this area. The widespread use of PPIs can confound the diagnosis of ZES. First, by raising a false-positive diagnosis by causing hypergastrinemia in a patient being treated with idiopathic PUD (without ZES). Second, by leading to a false-negative diagnosis because at routine doses used to treat patients with idiopathic PUD, PPIs control symptoms in most ZES patients and thus mask the diagnosis. If ZES is suspected and the gastrin level is elevated, it is important to show that it is increased when gastric pH is ≤2.0 because physiologically hypergastrinemia secondary to achlorhydria (atrophic gastritis, pernicious anemia) is one of the most common causes of hypergastrinemia. Nearly all ZES patients have a fasting pH ≤2 when off antisecretory drugs. If the fasting gastrin is >1000 pg/mL (increased tenfold) and the pH is ≤2.0, which occurs in 40–60% of patients with ZES, the diagnosis of ZES is established after the possibility of retained antrum syndrome has been ruled out by history. In patients with hypergastrinemia with fasting gastrins < 1000 pg/mL (<tenfold increased) and gastric pH ≤2.0, other conditions, such as *H. pylori* infections, antral G-cell hyperplasia/hyperfunction, gastric outlet obstruction, and, rarely, renal failure, can masquerade as ZES. To establish the diagnosis in this group, a determination of BAO and a secretin provocative test should be done. In patients with ZES without previous gastric acid–reducing surgery, the BAO is usually (>90%) elevated (i.e., >15 mEq/h). The secretin provocative test is usually positive, with the criterion of a >120-pg/mL increase over the basal level having the highest sensitivity (94%) and specificity (100%). Unfortunately the diagnosis of ZES is becoming increasing more difficult. This is due not only to the widespread use of PPIs (leading to false-positive results as well as masking ZES presentation), but also recent studies demonstrate than many of the commercial gastrin kits that are used by most laboratories to measure fasting serum gastrin levels are not reliable. In one study, 7 of the 12 tested commercial gastrin kits inaccurately assessed the true serum concentration of gastrin primarily because the antibodies used had inappropriate specificity for the different circulating forms of gastrin and were not adequately validated. Both underestimation and overestimation of fasting serum gastrin levels occurred using these commercial kits. To circumvent this problem, it is either necessary to use one of the five reliable kits identified or, alternatively, to refer the

patient to a center with expertise in making the diagnosis in your area, or if this is not possible, to contact such a center and use the gastrin assay they recommend. An accurate gastrin assay is essential for accurate measurement of fasting serum gastrin level as well as for assessing gastrin levels during the secretin provocative test, and thus, the diagnosis of ZES cannot reliably be made without one.

TREATMENT

TREATMENT

Zollinger-Ellison Syndrome

Gastric acid hypersecretion in patients with ZES can be controlled in almost every case by oral gastric antisecretory drugs. Because of their long duration of action and potency, which allows dosing once or twice a day, the PPIs (H^+, K^+-ATPase inhibitors) are the drugs of choice. Histamine H_2-receptor antagonists are also effective, although more frequent dosing (q 4–8 h) and high doses are required. In patients with MEN1/ZES with hyperparathyroidism, correction of the hyperparathyroidism increases the sensitivity to gastric antisecretory drugs and decreases the basal acid output. Long-term treatment with PPIs (>15 years) has proved to be safe and effective, without development of tachyphylaxis. Although patients with ZES, especially those with MEN 1/ZES, more frequently develop gastric NETs (carcinoids), no data suggest that the long-term use of PPIs increases this risk in these patients. With long-term PPI use in ZES patients, vitamin B_{12} deficiency can develop; thus, vitamin B_{12} levels should be assessed during follow-up. Long-term PPI use may be associated with a number of side-effects including; an increased incidence of bone fractures; *Clostridium difficile* infections; dementia; hypomagnesemia; renal disease; and numerous drug interactions; however, at present, there is no report these are increased in ZES patients.

With the increased ability to control acid hypersecretion, >50% of patients who are not cured (>60% of patients) will die from tumor-related causes. At presentation, careful imaging studies are essential to localize the extent of the tumor to determine the appropriate treatment. A third of patients present with hepatic metastases, and in <15% of those patients, the disease is limited, so that surgical resection may be possible. Surgical short-term cure is possible in 60% of all patients without MEN 1/ZES or liver metastases (40% of all patients) and in 30% of patients long term. In patients with MEN 1/ZES, long-term surgical cure is rare without aggressive resection (i.e., Whipple resections), because the tumors are small, multiple, and frequently with lymph node metastases. At present the role of routine surgery for removal of the gastrinoma in MEN1/ZES patients is controversial for the above reason, with most guidelines recommending attempted gastrinoma resection only in MEN1/ZES patients with pNETs ≥1.5–2 cm in diameter. Surgical studies demonstrate that successful resection of the gastrinoma not only decreases the chances of developing liver metastases but also increases the disease-related survival rate. Therefore, all patients with gastrinomas without MEN 1/ZES or a medical condition that limits life expectancy should undergo surgery by a surgeon experienced in the treatment of these disorders.

INSULINOMAS

An insulinoma is an NET of the pancreas that is thought to be derived from beta cells that ectopically secrete insulin, which results in hypoglycemia. The average age of occurrence is 40–50 years old. The most common clinical symptoms are due to the effect of the hypoglycemia on the CNS (neuroglycemic symptoms) and include confusion, headache, disorientation, visual difficulties, irrational behavior, and even coma. Also, most patients have symptoms due to excess catecholamine release secondary to the hypoglycemia, including sweating, tremor, and palpitations. Characteristically, these attacks are associated with fasting.

Insulinomas are generally small (>90% are <2 cm) and usually not multiple (90%); only 5–15% are malignant, and they almost invariably occur only in the pancreas, distributed equally in the pancreatic head, body, and tail. They are associated with the MEN1 syndrome in 4%.

Insulinomas should be suspected in all patients with hypoglycemia, especially when there is a history suggesting that attacks are provoked by fasting, or with a family history of MEN 1. Insulin is synthesized as proinsulin, which consists of a 21-amino-acid α chain and a 30-amino-acid β chain connected by a 33-amino-acid connecting peptide (C peptide). In insulinomas, in addition to elevated plasma insulin levels, elevated plasma proinsulin levels are found, and C-peptide levels are elevated.

Diagnosis

The diagnosis of insulinoma requires the demonstration of an elevated plasma insulin level at the time of hypoglycemia. A number of other conditions may cause fasting hypoglycemia, such as the inadvertent or surreptitious use of insulin or oral hypoglycemic agents, severe liver disease, alcoholism, poor nutrition, and other extrapancreatic tumors. Furthermore, postprandial hypoglycemia can be caused by a number of conditions that confuse the diagnosis of insulinoma. Particularly important here is the increased occurrence of hypoglycemia after gastric bypass surgery for obesity, which is now widely performed. A new entity, insulinomatosis, was described that can cause hypoglycemia and mimic insulinomas. It occurs in 10% of patients with persistent hyperinsulinemic hypoglycemia and is characterized by the occurrence of multiple macro-/microadenomas expressing insulin, and it is not clear how to distinguish this entity from insulinoma preoperatively. The most reliable test to diagnose insulinoma is a fast up to 72 h with serum glucose, C-peptide, proinsulin, and insulin measurements every 4–8 h. If at any point the patient becomes symptomatic or glucose levels are

persistently <2.2 mmol/L (40 mg/dL), the test should be terminated, and repeat samples for the above studies should be obtained before glucose is given. Some 70–80% of patients will develop hypoglycemia during the first 24 h, and 98% by 48 h. In nonobese normal subjects, serum insulin levels should decrease to <43 pmol/L (<6 μU/mL) when blood glucose decreases to <2.2 mmol/L (<40 mg/dL) and the ratio of insulin to glucose is <0.3 (in mg/dL). In addition to having an insulin level >6 μU/mL when blood glucose is <40 mg/dL, some investigators also require an elevated C-peptide and serum proinsulin level, an insulin/glucose ratio >0.3, and a decreased plasma β-hydroxybutyrate level for the diagnosis of insulinomas. A commonly used set of criteria to make the diagnosis include: low blood glucose levels (≤2.2 mmol/L (≤40 mg/dL); concomitant insulin levels ≥6 U/L (≥ 36 pmol/L; ≥3 U/L by ICMA); C-peptide levels ≥200 pmol/L; proinsulin levels ≥5 pmol/L; β-hydroxybutyrate levels ≤2.7 mmol/L; and absence of sulfonylurea (metabolites) in the plasma and/or urine.

Surreptitious use of insulin or hypoglycemic agents may be difficult to distinguish from insulinomas. The combination of proinsulin levels (normal in exogenous insulin/hypoglycemic agent users), C-peptide levels (low in exogenous insulin users), antibodies to insulin (positive in exogenous insulin users), and measurement of sulfonylurea levels in serum or plasma will allow the correct diagnosis to be made. The diagnosis of insulinoma has been complicated by the introduction of specific insulin assays that do not also interact with proinsulin, as do many of the older radioimmunoassays (RIAs), and therefore give lower plasma insulin levels. The increased use of these specific insulin assays has resulted in increased numbers of patients with insulinomas having lower plasma insulin values (<6 μU/mL) than levels proposed to be characteristic of insulinomas by RIA. In these patients, the assessment of proinsulin and C-peptide levels at the time of hypoglycemia is particularly helpful for establishing the correct diagnosis. An elevated proinsulin level when the fasting glucose level is <45 mg/dL is sensitive and specific.

TREATMENT

TREATMENT

Insulinomas

Only 5–15% of insulinomas are malignant; therefore, after appropriate imaging (see below), surgery should be performed. In different studies, 75–100% of patients are cured by surgery. Before surgery, the hypoglycemia can be controlled by frequent small meals and the use of diazoxide (150–800 mg/d). Diazoxide is a benzothiadiazide whose hyperglycemic effect is attributed to inhibition of insulin release. Its side effects are sodium retention and GI symptoms such as nausea. Approximately 50–60% of patients respond to diazoxide. Other agents effective in some patients to control the hypoglycemia include verapamil and diphenylhydantoin. Long-acting somatostatin analogues such as octreotide and lanreotide are acutely effective in 40% of patients. However, octreotide must be used with care because it inhibits growth hormone secretion and can alter plasma glucagon levels; therefore, in some patients, it can worsen the hypoglycemia.

For the 5–15% of patients with malignant insulinomas, these drugs or somatostatin analogues are used initially. In a small number of patients with malignant tumors, mammalian target of rapamycin (mTOR) inhibitors (everolimus, rapamycin) are reported to control the hypoglycemia. If they are not effective, various antitumor treatments such as hepatic arterial embolization, chemoembolization, chemotherapy, and peptide receptor radiotherapy with radiolabeled somatostatin analogues (PRRT) have been used and can be effective, particularly PRRT.

Insulinomas, which are usually benign (>90%) and intrapancreatic in location, are increasingly resected using a laparoscopic approach, which has lower morbidity rates. This approach requires that the insulinoma be localized on preoperative imaging studies.

GLUCAGONOMAS

A glucagonoma is NET of the pancreas that secretes excessive amounts of glucagon, which causes a distinct syndrome characterized by dermatitis, glucose intolerance or diabetes, and weight loss. Glucagonomas principally occur between 45 and 70 years of age. The tumor is clinically heralded by a characteristic dermatitis (migratory necrolytic erythema) (67–90%), accompanied by glucose intolerance (40–90%), weight loss (66–96%), anemia (33–85%), diarrhea (15–29%), and thromboembolism (11–24%). The characteristic rash usually starts as an annular erythema at intertriginous and periorificial sites, especially in the groin or buttock. It subsequently becomes raised, and bullae form; when the bullae rupture, eroded areas form. The lesions can wax and wane. The development of a similar rash in patients receiving glucagon therapy suggests that the rash is a direct effect of the hyperglucagonemia. A characteristic laboratory finding is hypoaminoacidemia, which occurs in 26–100% of patients.

Glucagonomas are generally large tumors at diagnosis (5–10 cm). Some 50–80% occur in the pancreatic tail. From 50 to 82% have evidence of metastatic spread at presentation, usually to the liver. Glucagonomas are rarely extrapancreatic, usually occur singly, and <3% are associated with the MEN1 syndrome.

Two new entities have been described that can also cause hyperglucagonemia and may mimic glucagonomas. Mahvash disease is due to an inactivating mutation (homozygous P86S mutation) of the human glucagon receptor. It is associated with the development of α-cell hyperplasia, hyperglucagonemia, and the development of nonfunctioning pNETs. Subsequently other patients with other inactivating mutations of the human glucagon receptor have been described with similar findings, leading to the suggestion that a hepato-pancreatic feedback regulation of the cells,

possibly involving amino acids, may exist in humans. A second disease called *glucagon cell adenomatosis* can mimic glucagonoma syndrome clinically and is characterized by the presence of hyperplastic islets staining positive for glucagon instead of a single glucagonoma.

Diagnosis

The diagnosis is confirmed by demonstrating an increased plasma glucagon level. Characteristically, plasma glucagon levels exceed 1000 pg/mL (normal is <150 pg/mL) in 90%; 7% are between 500 and 1000 pg/mL, and 3% are <500 pg/mL. A trend toward lower levels at diagnosis has been noted in the last decade. A plasma glucagon level >1000 pg/mL is considered diagnostic of glucagonoma. Other diseases causing increased plasma glucagon levels include cirrhosis, diabetic ketoacidosis, celiac disease, renal insufficiency, acute pancreatitis, hypercorticism, hepatic insufficiency, severe stress, and prolonged fasting or familial hyperglucagonemia, as well as danazol treatment. With the exception of cirrhosis, these disorders do not increase plasma glucagon >500 pg/mL.

Necrolytic migratory erythema is not pathognomonic for glucagonoma and occurs in myeloproliferative disorders, hepatitis B infection, malnutrition, short-bowel syndrome, inflammatory bowel disease, zinc deficiency, and malabsorption disorders.

TREATMENT

TREATMENT

Glucagonomas

In 50–80% of patients, hepatic metastases are present, and so curative surgical resection is not possible. Surgical debulking in patients with advanced disease or other antitumor treatments may be beneficial as well as PRRT with radiolabeled somatostatin analogues (see below). Long-acting somatostatin analogues such as octreotide and lanreotide improve the skin rash in 75% of patients and may improve the weight loss, pain, and diarrhea, but usually do not improve the glucose intolerance.

SOMATOSTATINOMA SYNDROME

The somatostatinoma syndrome is due to an NET that secretes excessive amounts of somatostatin, which causes a distinct syndrome characterized by diabetes mellitus, gallbladder disease, diarrhea, and steatorrhea. There is no general distinction in the literature between a tumor that contains somatostatin-like immunoreactivity (somatostatinoma) and does (11–45%) or does not (55–90%) produce a clinical syndrome (somatostatinoma syndrome) by secreting somatostatin. In a review of 173 cases of somatostatinomas, only 11% were associated with the somatostatinoma syndrome. The mean age is 51 years. Somatostatinomas occur primarily in the pancreas and small intestine, and the frequency of the symptoms and occurrence of the somatostatinoma syndrome differ in each. Each of the usual symptoms is more common in pancreatic than in intestinal somatostatinomas: diabetes mellitus (95% vs 21%), gallbladder disease (94% vs 43%), diarrhea (92% vs 38%), steatorrhea (83% vs 12%), hypochlorhydria (86% vs 12%), and weight loss (90% vs 69%). The somatostatinoma syndrome occurs in 30–90% of pancreatic and 0–5% of SI somatostatinomas. In various series, 43% of all duodenal NETs contain somatostatin; however, the somatostatinoma syndrome is rarely present (<2%). Somatostatinomas occur in the pancreas in 56–74% of cases, with the primary location being the pancreatic head. The tumors are usually solitary (90%) and large (mean size 4.5 cm). Liver metastases are common, being present in 69–84% of patients. Somatostatinomas are rare in patients with MEN 1, occurring in only 0.65%.

The existence of a somatostatinoma syndrome (SSoma syndrome) has been called into question. This occurred because in one review of 821 patients with duodenal or pancreatic NETs, a proportion of which showed predominant somatostatin expression, none had the SSoma syndrome leading to the conclusion it is either very rare or non-existent. However, in other studies, a proportion of the patients with somatostatin positive pancreatic or duodenal NETs had number of the proposed features of the SSoma syndrome.

Somatostatin is a tetradecapeptide that is widely distributed in the CNS and GI tract, where it functions as a neurotransmitter or has paracrine and autocrine actions. It is a potent inhibitor of many processes, including release of almost all hormones, acid secretion, intestinal and pancreatic secretion, and intestinal absorption. Most of the clinical manifestations are directly related to these inhibitory actions.

Diagnosis

In most cases, somatostatinomas have been found by accident either at the time of cholecystectomy or during endoscopy. The presence of psammoma bodies in a duodenal tumor should particularly raise suspicion. Duodenal somatostatin-containing tumors are increasingly associated with von Recklinghausen's disease (NF-1) (Table 80-5). Most of these tumors (>98%) do not cause the SSoma syndrome. The diagnosis of the SSoma syndrome requires the demonstration of elevated plasma somatostatin levels.

TREATMENT

TREATMENT

Somatostatinomas

Pancreatic tumors are frequently (70–92%) metastatic at presentation, whereas 30–69% of SI somatostatinomas have metastases. Surgery is the treatment of choice for those without widespread hepatic metastases. Symptoms in patients with the SSoma syndrome are also improved by octreotide treatment.

VIPomas

VIPomas are NETs that secrete excessive amounts of vasoactive intestinal peptide (VIP), which causes a distinct syndrome characterized by large-volume diarrhea, hypokalemia, and dehydration. This syndrome also is called Verner-Morrison syndrome, pancreatic cholera, and WDHA syndrome for *w*atery *d*iarrhea, *h*ypokalemia, and *a*chlorhydria, which some patients develop. The mean age of patients with this syndrome is 49 years; however, it can occur in children, and when it does, it is usually caused by a ganglioneuroma or ganglioneuroblastoma.

The principal symptoms are large-volume diarrhea (100%) severe enough to cause hypokalemia (80–100%), dehydration (83%), hypochlorhydria (54–76%), and flushing (20%). The diarrhea is secretory in nature, persisting during fasting, and is almost always >1 L/d and in 70% is >3 L/d. In a number of studies, the diarrhea was intermittent initially in up to half the patients. Most patients do not have accompanying steatorrhea (16%), and the increased stool volume is due to increased excretion of sodium and potassium, which, with the anions, accounts for the osmolality of the stool. Patients frequently have hyperglycemia (25–50%) and hypercalcemia (25–50%).

VIP is a 28-amino-acid peptide that is an important neurotransmitter, ubiquitously present in the CNS and GI tract. Its known actions include stimulation of SI chloride secretion as well as effects on smooth-muscle contractility, inhibition of acid secretion, and vasodilatory effects, which explain most features of the clinical syndrome.

In adults, 80–90% of VIPomas are pancreatic in location, with the rest due to VIP-secreting pheochromocytomas, intestinal carcinoids, and rarely ganglioneuromas. These tumors are usually solitary, 50–75% are in the pancreatic tail, and 37–68% have hepatic metastases at diagnosis. In children <10 years old, the syndrome is usually due to ganglioneuromas or ganglioblastomas and is less often malignant (10%).

Diagnosis

The diagnosis requires the demonstration of an elevated plasma VIP level and the presence of large-volume diarrhea. A stool volume <700 mL/d is proposed to exclude the diagnosis of VIPoma. When the patient fasts, a number of diseases can be excluded that can cause marked diarrhea because the high volume of diarrhea is not sustained during the fast. Other diseases that can produce a secretory large-volume diarrhea include gastrinomas, chronic laxative abuse, carcinoid syndrome, systemic mastocytosis, rarely medullary thyroid cancer, diabetic diarrhea, sprue, and AIDS. Among these conditions, only VIPomas caused a marked increase in plasma VIP. Chronic surreptitious use of laxatives/diuretics can be particularly difficult to detect clinically. Hence, in a patient with unexplained chronic diarrhea, screens for laxatives should be performed; they will detect many, but not all, laxative abusers. Elevated plasma levels of VIP should not be the only basis of the diagnosis of VIPomas because they can occur with some diarrheal states including inflammatory bowel disease, post small bowel resection, and radiation enteritis. Furthermore, nesidioblastosis can mimic VIPomas by causing elevated plasma VIP levels, diarrhea, and even false-positive location in the pancreatic region on somatostatin receptor scintigraphy (SRS).

TREATMENT

TREATMENT

VIPomas

The most important initial treatment in these patients is to correct their dehydration, hypokalemia, and electrolyte losses with fluid and electrolyte replacement. These patients may require >5 L/d of fluid and >350 mEq/d of potassium. Because 37–68% of adults with VIPomas have metastatic disease in the liver at presentation, a significant number of patients cannot be cured surgically. In these patients, long-acting somatostatin analogues such as octreotide and lanreotide (Fig. 80-2) are the drugs of choice.

Octreotide/lanreotide will control the diarrhea short- and long-term in 75–100% of patients. In nonresponsive patients, the combination of glucocorticoids and octreotide/lanreotide has proved helpful in a small number of patients. Other drugs reported to be helpful in small numbers of patients include prednisone (60–100 mg/d), clonidine, indomethacin, phenothiazines, loperamide, lidamidine, lithium, propranolol, and metoclopramide. Treatment of advanced disease with cytoreductive surgery, embolization, chemoembolization, chemotherapy, radiotherapy, radiofrequency ablation (RFA), and peptide receptor radiotherapy may be helpful (see below). Control of the diarrhea in VIPoma patients using the tyrosine kinase inhibitor, sunitinib, has been described in case reports.

NONFUNCTIONAL PANCREATIC NEUROENDOCRINE TUMORS (NF-pNETs)

NF-pNETs are NETs that originate in the pancreas and either secretes no products or their products do not cause a specific clinical syndrome. Their symptoms are due entirely to the tumor per se. NF-pNETs secrete chromogranin A (90–100%), chromogranin B (90–100%), α-HCG (human chorionic gonadotropin) (40%), neuron-specific enolase (31%), and β-HCG (20%), and because 40–90% secrete PP, they are also often called PPomas. A

proportion also secrete ghrelin, neurotensin, calcitonin and other GI hormones/neurotransmitters, which are generally accepted as not causing a distinct clinical syndrome. Because the symptoms are due to the tumor mass, patients with NF-pNETs usually present late in the disease course with invasive tumors and hepatic metastases (64–92%), and the tumors are usually large (72% >5 cm). An increasing proportion of NF-PNETs are asymptomatic (up to 30–50%) and are found at screening for various nonspecific symptoms. NF-pNETs are usually solitary except in patients with MEN 1, in which case they are multiple. They occur primarily in the pancreatic head. Even though these tumors do not cause a functional syndrome, immunocytochemical studies show that they synthesize numerous peptides and cannot be distinguished from functional pNETs by immunocytochemistry. In MEN 1, 80–100% of patients have microscopic NF-pNETs, but they become large or symptomatic in a minority (0–13%) of cases. In VHL, 12–17% develop NF-pNETs, and in 4%, they are ≥3 cm in diameter.

The most common symptoms are abdominal pain (30–80%), jaundice (20–35%), and weight loss, fatigue, or bleeding. The average time from the beginning of symptoms to diagnosis is 5 years.

Diagnosis

The diagnosis is established by histologic confirmation in a patient without either the clinical symptoms or the elevated plasma hormone levels of one of the established syndromes. The principal difficulty in diagnosis is to distinguish an NF-pNET from a nonendocrine pancreatic tumor, which is more common, as well as from a F-pNET. Even though chromogranin A levels are elevated in almost every patient, this is not specific for this disease as it can be found in F-pNETs, GI-NETs (carcinoids), and other neuroendocrine disorders, as well in patients without any of these, but being treated with PPIs. Plasma PP elevations should strongly suggest the diagnosis in a patient with a pancreatic mass because it is usually normal in patients with pancreatic adenocarcinomas. Elevated plasma PP is not diagnostic of this tumor because it is elevated in a number of other conditions, such as chronic renal failure, old age, inflammatory conditions, alcohol abuse, pancreatitis, hypoglycemia, postprandially, and diabetes. A positive somatostatin receptor scan in a patient with a pancreatic mass should suggest the presence of pNET/NF-pNET rather than a nonendocrine tumor.

TREATMENT

TREATMENT

Nonfunctional Pancreatic Neuroendocrine Tumors (NF-pNETs)

Overall survival in patients with sporadic NF-pNET is 30–63% at 5 years, with a median survival of 6 years. Unfortunately, surgical curative resection can be considered only in a minority of these patients because 30–92% present with diffuse metastatic disease. Treatment needs to be directed against the tumor per se using the various modalities discussed below for advanced disease. Whereas the treatment of NF-pNETs in either MEN 1 patients or patients with VHL has remained controversial for a number of years, the treatment in sporadic cases has also become controversial. In these inherited disorders, most recommend surgical resection for any tumor >2–3 cm in diameter; however, there is controversy in patients with smaller NF-pNETs (≤1.5–2 cm), with most guidelines recommending careful surveillance of these patients. This approach is taken because patients with these inherited diseases are not curable without aggressive surgery with its associated mortality/morbidity, because of the multiplicity of the small NF-pNETs; studies show these patients with NF-NETs ≤2 cm have no increased mortality; and most are slow growing. Most of these are low- or intermediate-grade lesions, and <7% are malignant. Similarly in patients with sporadic NF-pNETs in the past almost all were operated on; however, because of the generally benign course of those that are asymptomatic and ≤2 cm in diameter, increasingly they are not operated, but followed closely. No consensus exists on this point with the result that some advocate a non-operative approach with careful, regular follow-up, whereas others recommend an operative or laparoscopic.

GRFomas

GRFomas are NETs that secrete excessive amounts of growth hormone–releasing factor (GRF) that cause acromegaly. GRF is a 44-amino-acid peptide, and 25–44% of pNETs have GRF immunoreactivity, although it is uncommonly secreted. GRFomas are lung tumors in 47–54% of cases, pNETs in 29–30%, SI carcinoids in 8–10%; and up to 12% occur at other sites. Patients have a mean age of 38 years, and the symptoms usually are due to either acromegaly or the tumor per se. The acromegaly caused by GRFomas is indistinguishable from classic acromegaly. The pancreatic tumors are usually large (>6 cm), and liver metastases are present in 39%. They should be suspected in any patient with acromegaly and an abdominal tumor, a patient with MEN 1 with acromegaly, or a patient without a pituitary adenoma with acromegaly or associated with hyperprolactinemia, which occurs in 70% of GRFomas. GRFomas are an uncommon cause of acromegaly. GRFomas occur in <1% of MEN 1 patients. The diagnosis is established by performing plasma assays for GRF and growth hormone. Most GRFomas have a plasma GRF level >300 pg/mL (normal <5 pg/mL men, <10 pg/mL women). Patients with GRFomas also have increased plasma levels of insulin-like growth factor type I (IGF-I) similar to those in classic acromegaly. Surgery is the treatment of choice if diffuse metastases are not present. Long-acting somatostatin analogues such as octreotide and lanreotide are the agents of choice, with 75–100% of patients responding.

OTHER RARE PANCREATIC NEUROENDOCRINE TUMOR SYNDROMES

Cushing's syndrome (ACTHoma) due to a pNET occurs in 4–16% of all ectopic Cushing's syndrome cases. It occurs in 5% of cases of sporadic gastrinomas, almost invariably in patients with hepatic metastases, and is an independent poor prognostic factor. Paraneoplastic hypercalcemia due to pNETs releasing parathyroid hormone–related peptide (PTHrP), a PTH-like material, or unknown factor, is rarely reported. The tumors are usually large, and liver metastases are usually present. Most (88%) appear to be due to release of PTHrP. pNETs occasionally can cause the carcinoid syndrome and this may occur without the presence of liver metastases. A number of very rare pNET syndromes involving a few cases (less than five) have been described; these include a renin-producing pNET in a patient presenting with hypertension; pNETs secreting luteinizing hormone, resulting in masculinization or decreased libido; a pNET secreting erythropoietin, resulting in polycythemia; pNETs secreting IGF-II, causing hypoglycemia; pNETs secreting enteroglucagon, causing small intestinal hypertrophy, colonic/SI stasis, and malabsorption and a pNET secreting cholecystokinin (CCKoma) which can mimic ZES clinically with patients presenting with severe peptic ulcer disease, diarrhea, weight loss, and gallstones, but with a normal fasting gastrin level (Table 80-2). A number of other possible functional pNETs have been proposed, but most authorities classify these as unclear or as a nonfunctional pNET because in each case numerous patients have been described with similar plasma hormone elevations that do not cause any symptoms. These include pNETs secreting calcitonin, neurotensin (neurotensinoma), PP (PPoma), and ghrelin (Table 80-2).

TUMOR LOCALIZATION

Localization of the primary tumor and knowledge of the extent of the disease are essential to the proper management of all GI-NETs (carcinoids) and pNETs. Without proper localization studies, it is not possible to determine whether the patient is a candidate for surgical resection (curative or cytoreductive) or requires antitumor treatment, to determine whether the patient is responding to antitumor therapies, whether postresection recurrent disease is present or to appropriately classify/stage the patient's disease to assess prognosis.

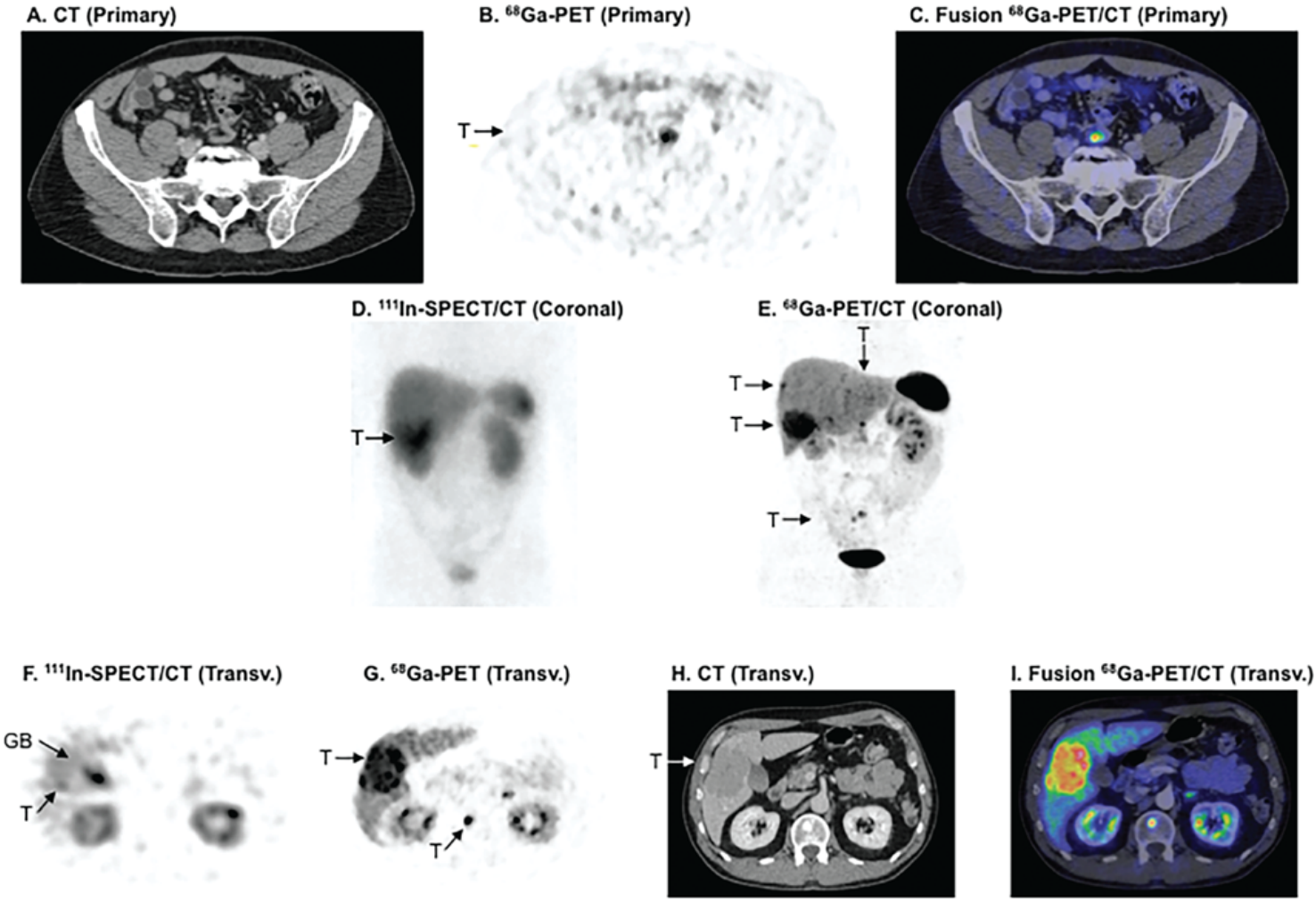
Numerous tumor localization methods are used in both types of NETs, including cross-sectional imaging studies (CT, magnetic resonance imaging [MRI], trans-abdominal ultrasound), selective angiography, SRS, and positron emission tomography. In pNETs, endoscopic ultrasound (EUS) and functional localization by measuring venous hormonal gradients are also reported to be useful. Bronchial carcinoids are usually detected by standard chest radiography and assessed by CT. Rectal, duodenal, colonic, and gastric carcinoids are usually detected by GI endoscopy. Because of their wide availability, CT and MRI are generally initially used to determine the location of the primary NETs and the extent of disease. NETs are hypervascular tumors, and with both MRI and CT, contrast enhancement is essential for maximal sensitivity, and it is recommended that generally triple-phase scanning be used. The ability of cross-sectional imaging and, to a lesser extent, SRS to detect NETs is a function of NET size. With CT and MRI, <10% of tumors <1 cm in diameter are detected, 30–40% of tumors 1–3 cm are detected, and >50% of tumors >3 cm are detected. Many primary GI-NETs (carcinoids) are small, as are insulinomas and duodenal gastrinomas, and are frequently not detected by cross-sectional imaging, whereas most other pNETs present late in the course of their disease and are large (>4 cm). Selective angiography is more sensitive, localizing 60–90% of all NETs; however, it is now used infrequently. For detecting liver metastases, CT and MRI are more sensitive than ultrasound, and with recent improvements, 5–25% of patients with liver metastases will be missed by CT and/or MRI.

pNETs, as well as GI-NETs (carcinoids), frequently (>80%) overexpress high-affinity sst in both the primary tumors and the metastases. Of the five types of somatostatin receptors (sst_{1-5}), radiolabeled octreotide binds with high affinity to sst_2 and sst_5, has a lower affinity for sst_3, and has a very low affinity for sst_1 and sst_4. Between 80 and 100% of well-differentiated (G1, G2 grades) GI-NETs (carcinoids) and pNETs possess sst_2, and many also have some of the other four sst subtypes. Interaction with these receptors can be used to treat these tumors as well as to localize NETs by using radiolabeled somatostatin analogues (i.e. somatostatin receptor imaging [SRI]). In contrast, only 50–70% of poorly differentiated (G3 grade) NETs have sst2 receptors. In the United States, (^{111}In-DTPA-D-Phe^1) octreotide (octreoscan) (Fig. 80-2) is still generally used with gamma camera detection using single-photon emission computed tomography (SPECT) imaging. Using gallium-68-labeled somatostatin analogues and positron emission tomography (^{68}Ga-PET/CT) detection has greater sensitivity than using ^{111}In-labeled somatostatin analogues (^{111}In-SPECT/CT) **(Fig. 80-3)**. It (NEWSPOT) is now approved for use in the United States. Because of its sensitivity and ability to localize tumor throughout the body, SRI is the initial imaging modality of choice for localizing both the primary tumor and metastatic NETs. SRI localizes tumor in 73–95% of patients with GI-NETs (carcinoids) and in 56–100% of patients with pNETs, except insulinomas. Insulinomas are usually small and have low densities of sst receptors, resulting in SRI being positive in only 12–50% of patients with insulinomas. SRS identifies >90–95% of patients with liver metastases due to NETs. Figure 80-3 shows an example with SRI of the increased sensitivity of ^{68}Ga-PET/CT over ^{111}In-SPECT-CT and CT scanning to localize both the primary NET and liver/bone metastases in a patient with a metastatic small intestinal carcinoid (GI-NET). Occasional false-positive responses with SRI can occur (12% in one study) because numerous other normal tissues as well as diseases can have high densities of sst receptors, including granulomas (sarcoid, tuberculosis, etc.), thyroid diseases (goiter, thyroiditis), activated lymphocytes (lymphomas, wound infections), splenunculi, increased osteoblastic activity, meningiomas, and increased physiological uptake in the pancreatic uncinate process (^{68}Ga-DOTATAE PET/CT). If liver metastases are identified by SRI (performed without hybrid CT), to plan the proper treatment, either a CT or an MRI (with contrast enhancement) is recommended to assess the size and exact location of the metastases, because SRI does not provide information on tumor size. For pNETs in the pancreas, EUS is highly sensitive, localizing 77–100% of insulinomas, which occur almost exclusively within the pancreas. EUS is less sensitive for extrapancreatic tumors. It is increasingly used in patients with MEN 1, and to a lesser extent VHL, to detect small pNETs not seen with other modalities or for serial pNET assessments to determine size changes or rapid growth in patients in whom surgery is deferred. EUS with cytologic evaluation also is used frequently to distinguish an NF-pNET from a

pancreatic adenocarcinoma or another nonendocrine pancreatic tumor. Not infrequently patients present with liver metastases due to an NET and the primary site is unclear. Occult small intestinal NETs (carcinoids) are increasingly detected by double-balloon enteroscopy or capsule endoscopy.

FIGURE 80-3

Enhanced sensitivity of ^{68}Ga-PET/CT to localize lesions in patient with a metastatic small intestinal carcinoid (GI-NET). Panels ***A–C*** show the ability of the ^{68}Ga-PET (transverse images) to localize the primary (T) when the CT is negative. Panels ***D*** and ***E*** (coronal views—maximum intensity projections) show the greater resolution and ability to localize more metastatic lesions (T) in the liver and bone of ^{68}Ga-PET than the ^{111}In-SPECT/CT scanning, which has been generally used in the United States until recently. Panels ***F-I*** (transverse images) show the increased sensitivity of ^{68}Ga-PET over the ^{111}In-SPECT/CT scanning in identifying the extent of the liver metastases as well as identifying bone metastasis. GB, gallbladder; T, tumor; Transv, transverse images. *(Results kindly provided by Prof. Anders Sundin, Department of Radiology, Uppsala University Hospital, Uppsala, Sweden.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Insulinomas frequently overexpress receptors for glucagon-like peptide-1 (GLP-1), and radiolabeled GLP-1 analogues have been developed that can detect occult insulinomas not localized by other imaging modalities. This study is only performed in a few specialty centers. Functional localization by measuring hormonal gradients is now uncommonly used with gastrinomas (after intra-arterial secretin injections) but is still frequently used in insulinoma patients in whom other imaging studies are negative (assessing hepatic vein insulin concentrations post-intra-arterial calcium injections). Functional localization measuring hormone gradients in insulinomas or gastrin gradients in gastrinoma is a sensitive method, being positive in 80–100% of patients. The intra-arterial calcium test may also allow differentiation of the cause of the hypoglycemia and indicate whether it is due to an insulinoma or a nesidioblastosis. The latter entity is becoming increasingly important because hypoglycemia after gastric bypass surgery for obesity is increasing in frequency, and it is primarily due to nesidioblastosis, although it can occasionally be due to an insulinoma.

PET and use of hybrid scanners such as CT and SRI has sensitivity because of the greater resolution of PET scanning. PET scanning with ^{18}F-fluoro-DOPA in patients with carcinoids or with ^{11}C-5-HTP in patients with pNETs or GI-NETs (carcinoids) has greater sensitivity than cross-sectional imaging studies and may be used increasingly in the future. PET/CT scanning using ^{18}F-FDG is receiving increasing attention in patients with NETs. It was initially thought that this would not be useful in NETs with the majority being well differentiated (G1, G2 grades, >85–98%) and having a low

proliferative rate. However, ^{18}F-FDG PET/CT can identify higher grade NETs, and is particularly helpful for imaging G3 NETs, which are more frequently negative with SRS. There ^{18}F-FDG positivity can not only provide imaging information on location and tumor size, but also prognostic information because the relative survival of patients with the different NET grades is G1>G2>G3.

TREATMENT

TREATMENT

Advanced and /or Aggressive Disease (Diffuse Metastatic Disease)

The single most important prognostic factor for survival is the presence of liver metastases **(Fig. 80-4A, B, D, and E)**. For patients with foregut carcinoids without hepatic metastases, the 5-year survival in one study was 95%, and with distant metastases, it was 20% (Fig. 80-4A, B). With gastrinomas, the 5-year survival without liver metastases is 98%; with limited metastases in one hepatic lobe, it is 78%; and with diffuse metastases, 16%. In a large study of 156 patients (67 pNETs, rest carcinoids), the overall 5-year survival rate was 77%; it was 96% without liver metastases, 73% with liver metastases, and 50% with distant disease.

The recent introduction and validation of the prognostic value of the different classification and grading systems (WHO, ENETS, the American Joint Committee on Cancer/International Union Against Cancer [AJCC/UICC]) are proving essential to stratify patients into different risk groups. A particular important prognostic factor is whether the NET is well differentiated (G1/G2) or poorly differentiated (<1% of all NETs) (G3) **(Fig. 80-4C, F)**. In various series overall, well-differentiated NETs, which are aggressive tumors, have a 5-year survival of 50–80%, whereas poorly differentiated NETs survival of only 0–15% at 5 years (Fig. 80-4C, F).

Therefore, treatment for advanced metastatic disease is an important challenge. A number of different modalities are reported to be effective, including cytoreductive surgery (surgically or by RFA), treatment with chemotherapy, somatostatin analogues, interferon α, hepatic embolization alone or with chemotherapy (chemoembolization), molecular targeted therapy, radiotherapy with radiolabeled beads/microspheres, PRRT, and liver transplantation.

SPECIFIC ANTITUMOR TREATMENTS

Cytoreductive surgery is considered if either all of the visible metastatic disease or at last 90% is thought resectable; however, unfortunately, this is possible in only the 9–22% of patients who present with limited hepatic metastases. Although no randomized studies have proven that it extends life, results from a number of studies suggest that it may increase survival; therefore, it is recommended, if possible. RFA can be applied to NET liver metastases if they are limited in number (usually <5) and size (usually <3.5 cm in diameter). It can be used at the time of surgery (either general or laparoscopic) or using radiologic guidance. Response rates are >80%, the responses can last up to 3 years, the morbidity rate is low, and this procedure may be particularly helpful in patients with F-pNETs that are difficult to control medically. Although RFA has not been established in a controlled trial, both the European and North American Neuroendocrine Tumor Society guidelines (ENETS, NANETS) state it can be an effective antitumor treatment for both refractory functional syndromes and for palliative treatment.

Although there are no controlled, long-term trials, palliative surgical resection of the small intestinal primary and surrounding tumor is generally recommended in most guidelines and expert opinion reviews for patients with midgut carcinoids with carcinoid syndrome, who almost invariably have unresectable live metastases. Systematic analysis of existing data supports the conclusion surgical resection of the primary prevents complications (obstruction, etc.) and also prolongs survival in some studies. At the time of this resection, a cholecystectomy is recommended to possible biliary complications from long-term somatostatin therapy.

Chemotherapy plays a different role in the treatment of patients with pNETs and GI-NETs (carcinoids). Chemotherapy continues to be widely used in the treatment of patients with advanced pNETs with moderate success (response rates 20–70%). However, in general, its results in patients with metastatic GI-NETs (carcinoids) have been disappointing, with response rates of 0–30% with various two- and three-drug combinations, and thus, it is infrequently used in these patients. An important distinction in patients with pNETs is whether the tumor is well differentiated (G1/G2) or poorly differentiated (G3). The chemotherapeutic approach is different for these two groups. The current regimen of choice for patients with well-differentiated pNETs is the combination of streptozotocin and doxorubicin with or without 5-fluorouracil. Streptozotocin is a glucosamine nitrourea compound originally found to have cytotoxic effects on pancreatic islets, and later in studies with doxorubicin with or without 5-fluorouracil, it produced response rates of 20–45% in advanced pNETs. Streptozotocin causes considerable morbidity, with 70–100% of patients developing side effects (most prominent being nausea/vomiting in 60–100% or leukopenia/thrombocytopenia) and 15–70% of patients developing some degree of renal dysfunction (primarily proteinuria and/or decreased creatinine clearance). The combination of temozolomide (TMZ) with capecitabine is receiving increased attention as a possible alternative to streptozotocin-based therapies. Experience is still limited with this protocol and it is being evaluated in a number of current studies; however, analysis of larger retrospective studies shows responded rates from 48 to 70%. The use of TMZ or another alkylating agent in advanced pNETs is supported by some, but not all, studies that show low levels of the DNA repair enzyme O^6-methylguanine DNA methyltransferase in pNETs correlate with the sensitivity to TMZ. Grade G3 NETs are primarily treated by chemotherapy (see below).

In addition to the effectiveness in controlling the functional hormonal state, long-acting somatostatin analogues such as octreotide and lanreotide are increasingly used for their antiproliferative effects. Whereas somatostatin analogues rarely decrease tumor size (i.e., 0–17%), these drugs have tumoristatic effects, stopping additional growth in 26–95% of patients with NETs. In a randomized, double-blind study in patients with metastatic midgut carcinoids (PROMID study), octreotide-LAR demonstrated a marked lengthening of time to progression (14.3 vs 6 months, $p = 0.000072$). This improvement was seen in patients with limited liver involvement. This study did not assess whether such treatment will extend survival. A double-blind, randomized, placebo-controlled, phase III study in patients with well-differentiated, metastatic, inoperable pNETs (45%) or GI-NETs (carcinoids) (55%) (CLARINET study) showed that monthly treatment with lanreotide-autogel reduced tumor progression or death by 53%. Somatostatin analogues can induce apoptosis in GI-NETs (carcinoids), which probably contributes to their tumoristatic effects. Treatment with somatostatin analogues is generally well-tolerated, with most side effects being mild and uncommonly leading to stopping the drug. Potential long-term side effects include diabetes/glucose intolerance, steatorrhea, and the development of gallbladder sludge/gallstones (10–80%), although only 1% of patients develop symptomatic gallbladder disease. Because of these phase III studies, somatostatin analogues are generally recommended as first-line treatment for patients with well-differentiated metastatic NETs.

Interferon α, similar to somatostatin analogues, is effective at controlling the hormonal excess symptoms of NETs and has antiproliferative effects in NETs, which primarily result in disease stabilization (30–80%), with a decrease in tumor size in <15% of patients. Interferon can inhibit DNA synthesis, block cell cycle progression in the G_1 phase, inhibit protein synthesis, inhibit angiogenesis, and induce apoptosis. Interferon α treatment results in side effects in the majority of patients, with the most frequent being a flu-like syndrome (80–100%), anorexia with weight loss, and fatigue. These side effects frequently decrease in severity with continued treatment. In addition, patients become accommodated to the symptoms. More serious side effects include hepatotoxicity (31%), hyperlipidemia (31%), bone marrow toxicity, thyroid disease (19%), and rarely CNS side effects (depression, mental/visual disorders). ENETS 2016 guidelines conclude that in patients with well-differentiated NETs that are slowly progressive, interferon α treatment should be considered if the tumor is somatostatin receptor negative or if somatostatin or targeted therapy (everolimus, sunitinib) treatment fails.

Molecular targeted medical treatment with either an mTOR inhibitor (everolimus) or a tyrosine kinase inhibitor (sunitinib) is now approved treatment in the United States and Europe for patients with metastatic unresectable pNET, each supported by a phase III, double-blind, prospective, placebo-controlled trial. Furthermore, a Phase 3 double-blind study (RADIANT-4) also demonstrated the effectiveness of everolimus in advanced, non-functional NETs of the lung or GI-tract. In this study involving patients with advanced, progressive well differentiated, NF-lung/GI-NETs, everolimus significantly (p <0.000001) improved progression-free survival and led to FDA approval for its use.

mTOR is a serine-threonine kinase that plays an important role in proliferation, cell growth, and apoptosis in both normal and neoplastic cells. Activation of the mTOR cascade is important in mediating NET cell growth. A number of mTOR inhibitors have shown promising antitumor activity in NETs including everolimus and temsirolimus, with the former undergoing two phase III trials (RADIANT-3/RADIANT-4) studies in patients with advance progressive NETs (RADIANT-3=pNETs, RADIANT-4=lung, GI NF-NETs). In the RADIANT-III study which involved 410 patients with advanced, progressive pNETs, everolimus caused significant improvement in progression-free survival (11 vs 4.6 months, $p < 0.001$) and increased by a factor of 3.7 the proportion of patients progression-free at 18 months (37% vs 9%). Everolimus treatment was associated with frequent side effects, causing a twofold increase in adverse events, with the most frequent being grade 1 or 2. Grade 3 or 4 side effects included hematologic, GI (diarrhea), stomatitis, or hypoglycemia occurring in 3–7% of patients. Most grade 3 or 4 side effects were controlled by dose reduction or drug interruption. Similar side-effects were found in the RADIANT-4 study. The ENETS 2016 guidelines conclude that everolimus, similar to sunitinib (below), can be considered as a first-line treatment in well-differentiated pNETs that are unresectable especially if somatostatin analogues are not an option. However, these guidelines recommended that somatostatin analogues be the initial treatment because of their low incidence of side-effects. In patients with GI-NETs, the ENET 2016 guidelines recommended that everolimus could be recommended as second-line therapy after somatostatin analogues.

Like other normal and neoplastic cells, NETs frequently possess multiple types of the 20 different tyrosine kinase (TK) receptors that are known and mediate the action of different growth factors. Numerous studies demonstrate that TK receptors in normal and neoplastic tissues as well as NETs are especially important in mediating cell growth, angiogenesis, differentiation, and apoptosis. Whereas a number of TK inhibitors show antiproliferative activity in NETs, only sunitinib has undergone a phase III controlled trial. Sunitinib is an orally active small-molecule inhibitor of TK receptors (PDGFRs, VEGFR-1, VEGFR-2, c-KIT, FLT-3). In a phase III study in which 171 patients with progressive, metastatic, nonresectable pNETs were treated with sunitinib (37.5 mg/d) or placebo, sunitinib treatment caused a doubling of progression-free survival (11.4 vs 4.5 months, $p < 0.001$), an increase in objective tumor response rate (9% vs 0%, $p = 0.007$), and an increase in overall survival. Sunitinib treatment was associated with an overall threefold increase in side effects, although most were grade 1 or 2. The most frequent grade 3 or 4 side effects were neutropenia (12%) and hypertension (9.6%), which were controlled by dose reduction or temporary interruption. There is no consensus regarding the order of sunitinib or everolimus use in patients with advanced, well-differentiated, progressive pNETs.

In patients with liver-predominant metastatic disease, a number of locoregional strategies have been used including: transarterial arterial embolization (TAE) alone or with chemotherapeutic agents (TACE); and selective internal radiation therapy (SIRT) or radioembolization. TACE/TAE can be effective because the blood supply to normal liver tissue is primarily from the portal vein whereas tumors receive 70–80% of their supply from the hepatic artery. Occlusion of selective branches of the hepatic artery is now generally performed radiologically. Contraindications include >50–75% liver

involvement by tumor, portal vein thrombosis, post-biliary reconstructive surgery, liver failure, and a poor performance rating. Results include a symptomatic response rate of 50–100%, and an objective response rate of 25–86% with a mean duration of response of 6–45 months. Complications include a postembolization syndrome with pain, nausea/vomiting and fever in 10–80% with <6% mortality. SIRT using yttrium-90 (^{90}Y) glass or resin microspheres is a relatively newer approach being evaluated in patients with unresectable NET liver metastases. The treatment requires careful evaluation for vascular shunting before treatment and a pretreatment angiogram to evaluate placement of the catheter and is generally is reserved for patients without extrahepatic metastatic disease and with adequate hepatic reserve. One of two types of ^{90}Y microspheres is used: either microspheres with a 20- to 60-μm diameter and 50 Bq/sphere (SIR-Spheres) or glass microspheres (TheraSpheres) with a 20- to 30-μm diameter and 2500 Bq/sphere. The ^{90}Y-microspheres are delivered to the liver by intraarterial injection from percutaneously placed catheters. The response rate varied from 50 to 61% (partial or complete), tumor stabilization occurred in 22–41%, 60–100% had symptomatic improvement, and overall survival varied from 25 to 70 months. Side effects include postembolization syndrome (pain, fever, nausea/vomiting [frequent]), which is usually mild, although grade 2 (43%) or grade 3 (1%) symptoms can occur; radiation-induced liver disease (<1%); and radiation pneumonitis (<1%). Contraindications to use include excess shunting to the GI tract or lung, inability to isolate the liver arterial supply, and inadequate liver reserve. Because of the limited data available in the ENETS 2012 guidelines, treatment with SIRTs is considered experimental.

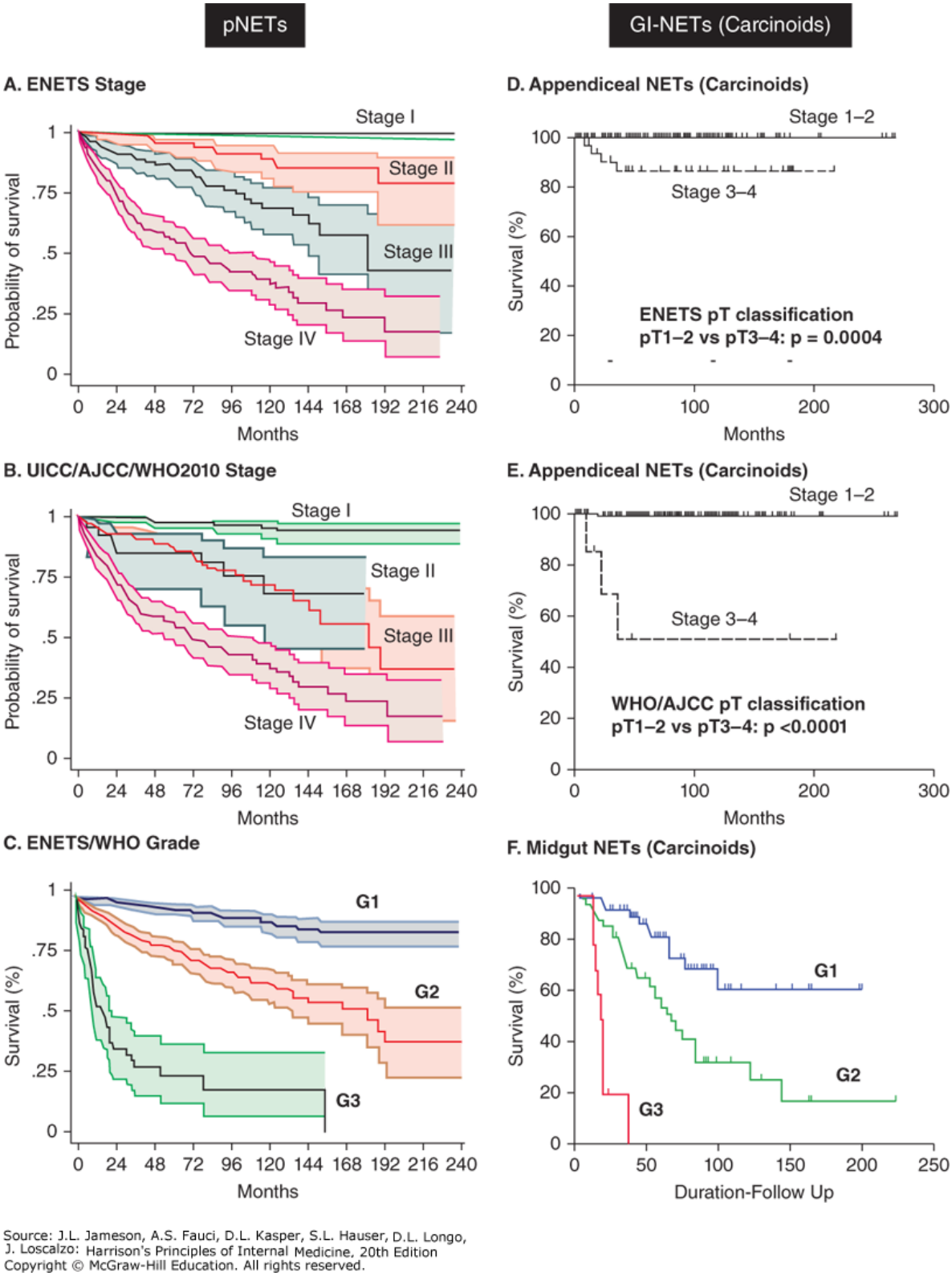
PRRT for NETs with radiolabeled somatostatin analogues is now being increasing considered for patients with advanced NETs. The success of this approach is based on the finding that somatostatin sst are overexpressed or ectopically expressed by 60–100% of all NETs, which allows the targeting of cytotoxic, radiolabeled somatostatin receptor ligands. Three different radionuclides have been used including: high doses of [^{111}In-DTPA-D-Phe^{1}] octreotide, which emits γ-rays, internal conversion, and Auger electrons; ^{90}yttrium, which emits high-energy β-particles coupled by a DOTA chelating group to octreotide or octreotate; and ^{177}lutetium-coupled analogues, which emit both (Fig. 80-2). At present, the ^{177}lutetium-coupled analogues are the most widely used and although not approved for general use in any country, they are frequently available in speciality centers on a special or compassionate basis. A double-blind, prospective, randomized trial (NETTER-1 Study) (using 177Lu-Dotatate [Lutathera]) has supported the efficacy and safety of this approach in patients with advanced inoperable, progressive midgut GI-NETs (carcinoids). In this trial, which included 229 patients with grade G1,2 metastatic midgut carcinoids, a marked increased in progressive-free survival ($p < 0.0001$) was seen with PRRT treatment with a ^{177}Lu-labeled-somatostatin-analog, with an acceptable safety profile and with a suggestion of an improved survival, although final survival analysis is not yet complete. In a number of retropective, non-blinded trials, ^{111}Indium-, ^{90}yttrium-, and ^{177}lutetium-labeled compounds caused tumor stabilization in patients with advanced, progressive NETs in 41–81%, 44–88%, and 23–51%, respectively, and a decrease in tumor size in 8–30%, 6–37%, and 38%, respectively, of patients. In one large study involving 504 patients with malignant NETs, ^{177}lutetium-labeled analogues produced a reduction of tumor size of >50% in 30% of patients (2% complete) and tumor stabilization in 51% of patients. The ENETS 2016, NANETS, Nordic 2010, and European Society for Medical Oncology (ESMO) guidelines list PRRT as an experimental or investigational treatment at present.

The use of liver transplantation has been abandoned for treatment of most metastatic tumors to the liver. However, for metastatic NETs, it is still a consideration by many centers although its use is controversial. An analysis of data from a number of centers showed that the overall 5-year survival is 47–58%, but varies widely in different studies from 36 to 97%; the 5-year disease-free survival was usually 20–30%, but varied from 9 to 77%, with a postoperative mortality <15%. With pNETs the 5-year survival rate varies from 30 to 50% and for GI-NETs from 60 to 90%. In various studies, important prognostic factors for a poor outcome include a major resection performed in addition at the time of the liver transplant; poor tumor differentiation; hepatomegaly; age >45 years; a primary NET in the duodenum or pancreas; the presence of extrahepatic metastatic disease or extensive liver involvement (>50%); Ki_{67} proliferative index >10%; and abnormal E-cadherin staining. The ENETS 2016 guidelines conclude that liver transplantation should be viewed as an option in highly selected patients, preferably in young patients with functional syndromes demonstrating early resistance to medical therapies.

The management and treatment of patients with G3 NETs (Ki_{67} >20) (WHO classification as NECs) has undergone a number of changes because of some important new insights. It is now realized that G3 NETs are heterogeneous and this has resulted in a proposal that they be divided into at least two categories; this division has important management ramifications because it is proposed they be treated differently. In reviews of G3 patients from a number of centers, a group of patients have G3 grading but with well-differentiated morphology (usually with a Ki_{67} 20–55) and it is proposed these be called G3 NET. These G3 NET patients have a better prognosis than poorly differentiated G3 tumors (usually with Ki_{67} >55), which are proposed to be called G3 NEC tumors. Pathology studies show that G3 NETs frequently have loss of ATRX/DAXX, whereas the G3 NEC poorly differentiated tumors have abnormal expression of p53, retinoblastoma and/or SMAD4. Most patients with G3 NETs have regional or distant metastases at the time of diagnosis and surgery is rarely curative, with the result that chemotherapy is usually recommended. This new subclassification has therapeutic implications because it is proposed to treat the G3 NET tumors similar to treatment for well-differentiated G2 tumors, whereas for G3 NEC tumors, treatment with cisplatin-based regimens with etoposide or other agents (vincristine, paclitaxel) is recommended. The response rates with this protocol are 40–70%; however, responses are generally short-lived (<12 months). This chemotherapy regimen can be associated with significant toxicity including GI toxicities (nausea, vomiting), myelosuppression.

FIGURE 80-4

Survival (Kaplan-Meier plots) of patients with pancreatic neuroendocrine tumors (pNETs; n = 1072) (***A–C***) or gastrointestinal neuroendocrine tumors (GI-NETs; carcinoids) (appendix, n = 138; midgut, n = 238) (***D–F***) stratified according to recent proposed classification and grading systems. *(Panels A–C are drawn from data in G Rindi et al: J Natl Cancer Inst 104:764, 2012; panels D and E are drawn from data in M Volante et al: Am J Surg Pathol 37:606, 2013; and panel F is drawn from data in MS Khan: Br J Cancer 108:1838, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Basuroy R et al: Neuroendocrine tumors. Gastroenterol Clin North Am 45:487, 2016. [PubMed: 27546845]

Cives M, Strosberg J: Treatment strategies for metastatic neuroendocrine tumors of the gastrointestinal tract. Curr Treat Options Oncol 18:1, 2017.

Ito T, Jensen RT: Molecular imaging in neuroendocrine tumors: Recent advances, controversies, unresolved issues, and roles in management. Curr Opin Endocrinol Diabetes Obes 24:15, 2017. [PubMed: 27875420]

Klimstra DS: Pathologic classification of neuroendocrine neoplasms. Hematol Oncol Clin North Am 30:1, 2016. [PubMed: 26614366]

Oberg K, Sundin A: Imaging of Neuroendocrine Tumors. Front Horm Res 45:142, 2016. [PubMed: 27002535]

O'Toole D et al: ENETS 2016 Consensus Guidelines for the Management of Patients with Digestive Neuroendocrine Tumors: An Update. Neuroendocrinology 103:113, 2016.

Pavel M et al: ENETS Consensus Guidelines Update for the Management of Distant Metastatic Disease of Intestinal, Pancreatic, Bronchial Neuroendocrine Neoplasms (NEN) and NEN of Unknown Primary Site. Neuroendocrinology 103:172, 2016. [PubMed: 26731013]

Strosberg J et al: Phase 3 Trial of ^{177}Lu-Dotatate for Midgut Neuroendocrine Tumors. N Engl J Med 376:125, 2017. [PubMed: 28076709]

Tamburrino D et al: Surgical management of neuroendocrine tumors. Best Pract Res Clin Endocrinol Metab 30:93, 2016. [PubMed: 26971846]

Yao JC et al: Everolimus for the treatment of advanced, non-functional neuroendocrine tumours of the lung or gastrointestinal tract (RADIANT-4): A randomised, placebo-controlled, phase 3 study. Lancet 387:968, 2016. [PubMed: 26703889]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 81: Renal Cell Carcinoma

Robert J. Motzer

FIGURE 81-1

INTRODUCTION

Renal cell carcinomas account for 90–95% of malignant neoplasms arising from the kidney. Notable features include diagnosis without symptoms, resistance to cytotoxic agents, infrequent responses to biologic response modifiers such as interleukin (IL)-2, robust activity of antiangiogenesis-targeted agents, and a variable clinical course for patients with metastatic disease, including anecdotal reports of spontaneous regression. The remaining 5–10% of malignant neoplasms arising from the kidney are transitional cell carcinomas (urothelial carcinomas) originating in the lining of the renal pelvis. See **Chap. 82** for transitional cell carcinomas.

EPIDEMIOLOGY

The incidence of renal cell carcinoma rose for three decades but has now reached a plateau of ~63,000 cases annually in the United States, resulting in >14,000 deaths per year. It is the ninth most common cancer overall in the United States, seventh most common in males, and tenth most common in females; the male-to-female ratio is 2:1. Though this malignancy may be diagnosed at any age, it is uncommon in those aged <45 years, and incidence peaks between the ages of 50 and 70 years. Many factors have been investigated as possible contributing causes; associations include cigarette smoking, obesity, and hypertension. Risk is also increased for patients who have acquired cystic disease of the kidney associated with end-stage renal disease and for those with tuberous sclerosis.

Most cases of renal cell carcinoma are sporadic, although familial forms have been reported **(Table 81-1)**. One such form is associated with von Hippel-Lindau (VHL) syndrome, an autosomal dominant disorder. Genetic studies identified the *VHL* gene on the short arm of chromosome 3. Approximately 35% of individuals with VHL disease develop clear cell renal cell carcinoma. Other *VHL*-associated neoplasms include retinal hemangioma, hemangioblastoma of the spinal cord and cerebellum, pheochromocytoma, and neuroendocrine tumors and cysts. Birt-Hogg-Dubé syndrome is a rare human autosomal dominant genetic disorder characterized by fibrofolliculomas (benign tumors arising in hair follicles), pulmonary cysts, and kidney tumors. The renal tumors are usually of the chromophobe type, but they can exist as hybrids with other cell types. This disorder is associated with mutations in the *FLCN* gene, which codes for folliculin.

TABLE 81-1

Hereditary Renal Cell Tumors

Syndrome	Chromosome(s)	Gene	Protein	Kidney Tumor Type	Additional Findings
von Hippel-Lindau syndrome	3p25	*VHL*	von Hippel-Lindau protein	Clear cell	Hemangioblastoma of the retina and central nervous system; pheochromocytoma; pancreatic and renal cysts; neuroendocrine tumors
Hereditary papillary RCC	7p31	*MET*	MET	Papillary (type I)	
Hereditary leiomyomatosis and RCC	1q42	*FH*	Fumarate hydratase	Papillary (non–type I)	Leiomyoma; uterine leiomyoma/leiomyosarcoma
Birt-Hogg-Dubé syndrome	17p11	*FLCN*	Folliculin	Chromophobe, oncocytoma	Facial fibrofolliculoma; pulmonary cysts
Tuberous sclerosis	9q34 16p13	*TSC1* *TSC2*	Hamartin Tuberin	Angiomyolipomas; lymphangioleiomyomatosis; rare RCC with variety of histologic appearances	Angiofibroma, subungual fibroma; cardiac rhabdomyoma; adenomatous small intestine polyps; pulmonary and renal cysts; cortical tuber; subependymal giant cell astrocytomas
Constitutional chromosome 3 translocations	3p13-14	Unknown	Unknown	Clear cell	

Abbreviation: RCC, renal cell carcinoma.

PATHOLOGY AND GENETICS

Renal cell neoplasia represents a heterogeneous group of tumors with distinct histopathologic, genetic, and clinical features ranging from benign to high-grade malignant (Table 81-2). They are classified on the basis of morphology and histology. Categories include clear cell carcinoma (70% of cases), papillary tumors (10%), chromophobe tumors (≤5%), oncocytomas (5–10%), collecting duct or Bellini duct tumors (<1%), and translocation carcinoma (<1%). Papillary tumors tend to be bilateral and multifocal. Chromophobe tumors have a more indolent clinical course, and oncocytomas are considered benign neoplasms. In contrast, Bellini duct carcinomas, which are thought to arise from the collecting ducts within the renal medulla, are rare but often very aggressive. Medullary carcinoma has histopathologic and clinical features similar to those of Bellini duct carcinoma, but it is associated with sickle cell trait.

TABLE 81-2

Classification of Epithelial Neoplasms Arising from the Kidney

Carcinoma Type	Characteristics Growth Pattern	Cell of Origin	Cytogenetics
Clear cell	Acinar or sarcomatoid	Proximal tubule	3p-, 5q+, 14q-
Papillary	Papillary or sarcomatoid	Proximal tubule	+7, +17, −Y
Chromophobe	Solid, tubular, or sarcomatoid	Distal tubules/cortical collecting duct	Whole arm losses (1, 2, 6, 10, 13, 17, and 21)
Oncocytic	Tumor nests	Cortical collecting duct	−1, −14, −Y; rearrangement involving 11q13; or normal karyotype
Collecting duct	Papillary or sarcomatoid	Medullary collecting duct	Variable or undetermined
MiTF Translocation[a]	Clear and papillary	Undetermined	Gene fusions involving Xp11 (*TFE3*) or t(6;11) (*MALAT1-TFEB*)

[a]Microphthalmia transcription factor gene family.

Clear cell tumors, the predominant histology, are found in >80% of patients who develop metastases. Clear cell tumors arise from the epithelial cells of the proximal tubules and usually show chromosome 3p deletions. Deletions of 3p21–26 (where the *VHL* gene maps) are identified in patients with familial as well as sporadic tumors. *VHL* encodes a tumor suppressor protein that is involved in regulating the transcription of vascular endothelial growth factor (VEGF), platelet-derived growth factor (PDGF), and a number of other hypoxia-inducible proteins. Inactivation of *VHL* leads to overexpression of these agonists of the VEGF and PDGF receptors, which promote tumor angiogenesis and tumor growth. Agents that inhibit proangiogenic growth factor activity show antitumor effects. Enormous genetic variability has been documented in tumors within individual patients. Although the tumors have a clear clonal origin and often contain *VHL* mutations in common, different portions of the primary tumor and different metastatic sites may have wide variation in genetic lesions they contain. This tumor heterogeneity may underlie the emergence of treatment resistance.

While *VHL* is the gene most frequently mutated in clear cell renal cell carcinoma (52% of cases), other genes are implicated as well: *PBRM1* in 40% of cases, *SETD2* in 15% of cases, and *BAP1* in 15% of cases. These three genes, all part of the chromatin remodeling/histone methylation pathway, are also located within a 50-Mb region on the short arm of chromosome 3p. Mutations in *BAP1* have been linked to shorter survival in renal cancer. In a subset of clear cell renal cell carcinomas, alterations have been found in components of the mammalian target of rapamycin (mTOR) pathway, spurring the study of mTOR inhibitors in renal cancer.

Approximately 10% of renal cell carcinomas are of the papillary subtype, where the most common copy-number events are gain of chromosome 7 (where MET is located) and chromosome 17. Alterations in MET are associated with type I papillary renal cell carcinoma, whereas type II papillary tumors are characterized by NFR2-antioxidant response element alterations. In the chromophobe subtype, which comprises ≤5% of cases of renal cell carcinoma, two mutations have been noted: TP53 in 32% of cases and PTEN in 9%.

CLINICAL PRESENTATION

The presenting signs and symptoms include hematuria, flank or abdominal pain, and a flank or abdominal mass. Other symptoms are fever, weight loss, anemia, and a varicocele. The tumor is most commonly detected as an incidental finding on a radiograph. Widespread use of radiologic cross-sectional imaging procedures (computed tomography [CT], ultrasound, magnetic resonance imaging [MRI]) contributes to earlier detection, including incidental renal masses detected during evaluation for other medical conditions. The increasing number of incidentally discovered low-stage tumors has contributed to an improved 5-year survival for patients with renal cell carcinoma and increased use of nephron-sparing surgery (partial nephrectomy). A spectrum of paraneoplastic syndromes has been associated with these malignancies, including erythrocytosis, hypercalcemia, nonmetastatic hepatic dysfunction (Stauffer's syndrome), and acquired dysfibrinogenemia. Erythrocytosis is noted at presentation in only ~3% of patients. Anemia, a sign of metastatic disease, is more common. Kidney cancer was called the "internist's tumor" since it was often discovered from the initial presentation of a paraneoplastic syndrome. This was more common before the era of modern imaging, as was initial presentation by the classic triad of hematuria, flank pain, and a palpable abdominal mass.

The standard evaluation of patients with suspected renal cell tumors includes a CT scan of the abdomen and pelvis, chest radiograph, urine analysis, and urine cytology. If metastatic disease is suspected from the chest radiograph, a CT of the chest is warranted. MRI is useful in evaluating the inferior vena cava

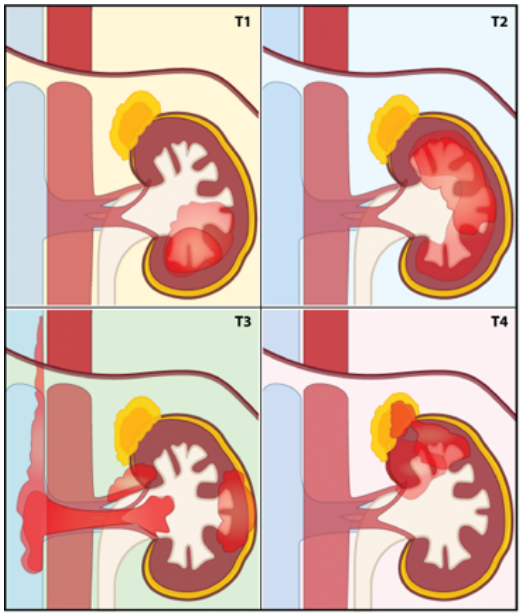
in cases of suspected tumor involvement or invasion by thrombus, or when intravenous contrast administration given with CT is prohibited by impaired renal function. In clinical practice, any solid renal masses should be considered malignant until proven otherwise; a definitive diagnosis is required. If no metastases are demonstrated, surgery is indicated, even if the renal vein or inferior vena cava is invaded. The differential diagnosis of a renal mass includes cysts, benign neoplasms (adenoma, angiomyolipoma, oncocytoma), inflammatory lesions (pyelonephritis or abscesses), and other primary or metastatic cancers. Other malignancies that may involve the kidney include transitional cell carcinoma of the renal pelvis, sarcoma, lymphoma, and Wilms' tumor. All of these are less common causes of renal masses than is renal cell cancer.

STAGING AND PROGNOSIS

Staging is based on the American Joint Committee on Cancer (AJCC) staging system **(Fig. 81-1)**. Stage I tumors are ≤7 cm in greatest diameter and confined to the kidney, stage II tumors are >7 cm and confined to the kidney, stage III tumors extend through the renal capsule but are confined to Gerota's fascia (IIIa) or involve a single hilar lymph node (N1), and stage IV disease includes tumors that have invaded adjacent organs or involve multiple lymph nodes or distant metastases. Sixty-five percent of patients present with stage I or II disease, 15–20% with stage III, and 15–20% with stage IV. The 5-year survival rate is currently 74% across all renal cell carcinomas, but it varies by stage: 81% for stage I, 74% for stage II, 53% for stage III, and 8% for stage IV.

FIGURE 81-1

Renal cell carcinoma staging. TNM, tumor, node, metastasis.



TNM	Involvement	Extent of Disease
T0	Primary not involved	
T1	≤ 7 cm	limited to kidney
T1a	≤ 4 cm	
T1b	≥ 4 cm	
T2	> 7 cm	limited to kidney
T2a	> 7 cm to ≤ 10 cm	
T2b	>10 cm	
T3	into major veins or perinephric tissues	not beyond Gerota's fascia
T3a	in renal vein or renal sinus fat	not beyond Gerota's fascia
T3b	into vena cava	below diaphragm
T3c	into vena cava	above diaphragm
T4	invasion beyond Gerota's fascia	including contiguous extensions & into ipsilateral adrenal gland

Regional

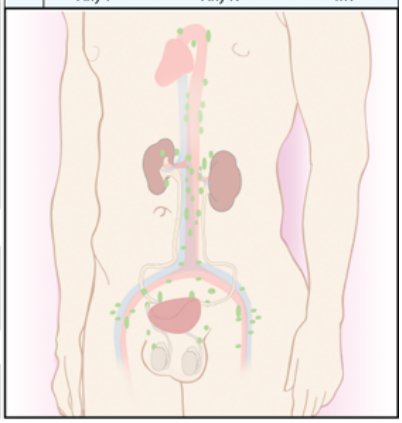
NX	Regional lymph nodes not assessed
N0	No lymph node involvment
N1	Regional lymph node involvment

Distant Metastases

M0	No distant metastases
M1	Distant metastases

Anatomic Stage / Prognostic Groups

I	T1	N0	M0
II	T2	N0	M0
III	T1 or T2,	N1	M0
	T3	N0 or N1	M0
IV	T4	Any N	M0
	Any T	Any N	M1



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prognostic risk models are helpful for counseling patients, and for anticipating survival rates when designing a clinical trial. The most widely used prognostic model, developed by investigators at Memorial Sloan Kettering Cancer Center, incorporated five factors shown to correlate with worse survival in advanced renal cell carcinoma: poor performance status, high serum lactate dehydrogenase, high serum calcium, low hemoglobin concentration, and <1-year interval from diagnosis to treatment. Patients with zero risk factors had significantly longer median survival (30 months) than those with one or two risk factors (14 months) and those with three to five risk factors (5 months).

TREATMENT

TREATMENT

Renal Cell Carcinoma

LOCALIZED TUMOR

The standard management for stage I or II tumors and selected cases of stage III disease is radical or partial nephrectomy. A radical nephrectomy involves en bloc removal of Gerota's fascia and its contents, including the kidney, the ipsilateral adrenal gland in some cases, and adjacent hilar lymph nodes. Open, laparoscopic, or robotic surgical techniques may be used to perform radical nephrectomy. The role of a regional lymphadenectomy is controversial. Extension into the renal vein or inferior vena cava (stage III disease) does not preclude resection even if cardiopulmonary bypass is required. If the tumor is resected, half of these patients have prolonged survival.

Nephron-sparing approaches via open or laparoscopic surgery may be appropriate for patients who have impaired renal function or only one kidney, depending on the size and location of the lesion. A nephron-sparing approach can also be used for patients with bilateral tumors. Partial nephrectomy

techniques are applied electively to resect small masses for patients with a normal contralateral kidney. Radical nephrectomy can lead to an increased risk for chronic kidney disease and is associated with increased risks of cardiovascular morbidity and mortality. When compared with radical nephrectomy, partial nephrectomy can achieve preserved renal function, and reduced frequency of late cardiovascular events.

Adjuvant therapy with interferon-α or radiation therapy following this surgery does not improve outcome, even in cases with a poor prognosis. Adjuvant trials with sunitinib, an orally administered antiangiogenesis inhibitor, do not consistently show a benefit in prolonging time to relapse following nephrectomy.

METASTATIC DISEASE

Surgery has a limited role for patients with metastatic disease. Long-term survival may occur in patients who relapse after nephrectomy in a solitary site that is removed. One indication for nephrectomy with metastases at initial presentation is to alleviate pain or hemorrhage of a primary tumor. Also, a cytoreductive nephrectomy before systemic treatment improves survival for carefully selected patients with stage IV tumors. The most common sites of distant metastases are the lungs, lymph nodes, liver, bone, and brain. These tumors may follow an unpredictable and protracted clinical course. It may be best to document progression before considering systemic treatment.

Radiation therapy is generally used for palliation of bone or brain metastases. The types of radiotherapy most commonly used are external beam therapy and stereotactic radiotherapy. In select cases, stereotactic ablative radiotherapy to a metastatic site may result in local control with relatively minimal toxicity.

Metastatic renal cell carcinoma is refractory to cytotoxic chemotherapy. Cytokine therapy with IL-2 or interferon-α produces regression in 10–15% of patients. IL-2 produces durable complete remission in a small proportion of cases. In general, cytokine therapy is considered unsatisfactory for most patients due to high levels of toxicity and the unpredictability of response.

The situation changed dramatically when two large-scale randomized trials established a role for antiangiogenic therapy, as predicted by the genetic studies. These trials separately evaluated two orally administered antiangiogenic agents, sorafenib and sunitinib, that inhibited receptor tyrosine kinase signaling through the VEGF and PDGF receptors. Both showed efficacy as second-line treatment following progression during cytokine treatment, resulting in approval by regulatory authorities for the treatment of metastatic renal cell carcinoma. A randomized phase III trial comparing sunitinib to interferon-α showed superior efficacy for sunitinib with an acceptable safety profile. This trial resulted in a change in the standard first-line treatment from interferon to sunitinib.

These were followed by eight new systemic agents for metastatic renal cell carcinoma **(Table 81-3)**: pazopanib, axitinib, cabozantinib, and lenvatinib, also tyrosine kinase inhibitors; the antiangiogenic bevacizumab that inhibits the VEGF ligand; the mTOR inhibitors temsirolimus and everolimus; and nivolumab that inhibits PD-1. While the improvements in 5-year renal cancer survival rates over the past decades (50% in the mid-1970s, 57% in the late 1980s, and 74% for 2005–2012) can be attributed to widespread imaging leading to earlier discovery of tumors, the new agents are likely playing a part as well.

TABLE 81-3

Approved Systemic Therapies for Metastatic Renal Cell Carcinoma

Class	Drug	First FDA Approval for RCC	Originally Approved for	Currently Used for
Cytokines	High-dose interleukin-2[a]	1992	Advanced RCC	Advanced RCC first-line
	Interferon-α	2009	Advanced RCC, in combination with bevacizumab	Advanced RCC, in combination with bevacizumab first-line
Antiangiogenic: tyrosine kinase inhibitors	Sorafenib	2005	Advanced RCC second-line	Advanced RCC third-line or later
	Sunitinib	2006	Advanced RCC second-line	Advanced RCC first-line
	Pazopanib	2009	Advanced RCC first-line or after cytokine therapy	Advanced RCC first-line
	Axitinib	2012	Advanced RCC second-line	Advanced RCC second-line
	Cabozantinib	2016	Advanced RCC second-line	Advanced RCC second- or third-line
	Lenvatinib	2016	Advanced RCC second-line in combination with everolimus	Advanced RCC second- or third-line in combination with everolimus
Antiangiogenic: VEGF ligand antibody	Bevacizumab	2009	Advanced RCC first-line in combination with interferon-α	Advanced RCC first-line in combination with interferon-α
mTOR inhibitors	Temsirolimus[b]	2007	Advanced RCC	Advanced RCC with poor prognosis features first-line
	Everolimus	2009	Metastatic RCC second-line	Advanced RCC third-line or later
PD-1 inhibitor	Nivolumab	2015	Advanced RCC second-line	Advanced RCC second- or third-line

[a]Option only for patients with good performance status, no significant comorbidity, and access to medical centers experienced with this agent. [b]Option for poor-risk patients.

Abbreviations: FDA, Food and Drug Administration; mTOR, mammalian target of rapamycin; PD-1, programmed cell death-1; RCC, renal cell carcinoma; VEGF, vascular endothelial growth factor.

Pazopanib was compared to sunitinib in a randomized first-line phase III trial. Efficacy was similar, and there was less fatigue and skin toxicity, resulting in better quality-of-life scores for pazopanib compared with sunitinib. Temsirolimus and everolimus show activity in patients with untreated poor-prognosis tumors and in sunitinib/sorafenib-refractory tumors. Patients benefit from the sequential use of axitinib and everolimus following progression with sunitinib or pazopanib first-line therapy. Nivolumab, cabozantinib, and lenvatinib plus everolimus were compared to everolimus in randomized trials and showed that patients lived longer with each of these agents compared to patients treated with everolimus.

Biomarkers are needed to select appropriate treatment for individual patients and to get quicker confirmation of whether treatment is working. However, though a number of predictive biomarker candidates have been tested in metastatic renal cell carcinoma patients receiving various systemic therapies, none have been validated for clinical use.

GLOBAL CONSIDERATIONS

Worldwide, ~340,000 patients are diagnosed every year with malignant tumors arising from the kidney, resulting in >140,000 deaths annually. Kidney cancer is the ninth most common cancer in men and the fourteenth most common cancer in women. Higher incidence is observed in developed countries, including the United States, Northern Europe, Eastern Europe, and Australia. Relatively low rates are reported in southeast Asia and Africa. The incidence of kidney cancer has been steadily increasing over the past four decades. Mortality trends have stabilized in Europe and the United States but not in less

developed countries. This is likely related to access to and availability of optimal therapies. Treatment guidelines for both localized and metastatic renal cancer are similar between U.S. and European documents, and contingent on the access to adequate health care and availability of targeted drugs to treat metastases.

FURTHER READING

Cancer Genome Atlas Research Network: Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499:43, 2013. [PubMed: 23792563]

Choueiri TK, Motzer RJ: Systemic therapy for metastatic renal cell carcinoma. N Engl J Med 376:354, 2017.

Gerlinger M et al: Intratumor heterogeneity and branched evolution revealed by multi-region sequencing. N Engl J Med 366:883, 2012. [PubMed: 22397650]

Gill IS et al: Clinical practice. Small renal mass. N Engl J Med 362:634, 2010.

Motzer RJ et al: Pazopanib versus sunitinib in metastatic renal-cell carcinoma. N Engl J Med 369:722, 2013. [PubMed: 23964934]

Motzer RJ et al: Nivolumab versus everolimus in advanced renal-cell carcinoma. N Engl J Med 373:1803, 2015. [PubMed: 26406148]

Srigley JR et al: The International Society of Urological Pathology (ISUP) Vancouver Classification of renal neoplasia. Am J Surg Pathol 37:1469, 2013. [PubMed: 24025519]

Znaor A et al: International variations and trends in renal cell carcinoma incidence and mortality. Eur Urol 67:531, 2015. [PubMed: 25454608]

Chapter 82: Cancer of the Bladder and Urinary Tract

Noah M. Hahn

GLOBAL CONSIDERATIONS


Within the United States, urothelial carcinoma of the bladder and urinary tract are most closely related to tobacco smoking history. However, within developing countries water supplies contaminated with arsenic or schistosomiasis parasites also are major carcinogenic contributors.

INTRODUCTION

Cancers of the urinary tract including the bladder, renal pelvis, ureter, and urethra occur frequently, and they represent the second most common class of genitourinary cancers. Bladder cancer alone represents the fifth most common cancer diagnosis annually in the United States with >76,000 new cases and 16,000 deaths every year. Because cancers of the renal pelvis are often lumped in with all kidney cancers, the true incidence and mortality from nonbladder urinary tract cancers are less precise. While less frequent than bladder cancer, an additional 20,000 new cases and 5000 deaths are estimated every year. While significant advances in therapy options and improvements in patient outcomes have rapidly occurred in many cancers in the past decade, progress in urinary tract cancers has lagged. Fortunately, an accelerated understanding of the molecular underpinnings of bladder and urinary tract cancer biology has led to a significant increase in clinical trials with the first U.S. Food and Drug Administration (FDA) approval of a new drug for advanced bladder and urinary tract cancers in over 25 years with many more expected to follow. This chapter reviews the established, current, and emerging evidence that serves as the basis for the rapidly evolving standards of care for patients with bladder and urinary tract cancers.

CLINICAL EPIDEMIOLOGY AND RISK FACTORS

Bladder cancer typically affects older patients with a median age at diagnosis of 73 years. Males are four times more frequently affected than females. Similarly, bladder cancer is more common in Caucasians than in Asian patients. Singular inheritable genetic risk factors are rare in patients with bladder or urinary tract cancers. Patients with defects in mismatch repair genes leading to microsatellite instability (*MLH1, MSH2, MSH6*, etc.) as part of the familial cancer Lynch syndrome are at particular risk of upper urinary tract cancers of the renal pelvis and ureter. Additionally, patients with Cowden disease (*PTEN* mutations) or retinoblastoma (*RB1* mutations) are at increased risk for developing bladder cancer.

Historically, associations have existed between environmental toxic exposures and higher rates of developing bladder cancer. Carcinogenic agents associated with increased risk of bladder cancer have included the aromatic amines benzidine and beta-naphthylamine that can be present in industrial dyes as well as arsenic that can be found in some drinking water supplies in underdeveloped countries. Other chemicals in the leather, paint, rubber, textiles, and printing industries have been associated with bladder cancer. More recently, associations with exposures to hair dyes and hair sprays in workers in the hairstyling field have been suggested. Additionally, much concern has been raised regarding use of the antidiabetic medication, pioglitazone, and bladder cancer risk. Extensive review of population data by leading bladder cancer experts has produced mixed associations. An association between chronic inflammatory states and the development of squamous bladder cancer clearly exists in underdeveloped countries in patients chronically infected with the parasitic disease schistosomiasis and in paraplegic patients with chronic indwelling catheters. Above and beyond each of these associations, however, smoking of tobacco products (cigarettes, cigars, pipes, etc.) has been and continues to remain the overwhelming leading risk factor for development of bladder cancer. Among new bladder cancer diagnoses, 90% of cases occur in current or former smokers. Toxicologists have estimated that over 70 confirmed carcinogenic toxins are present within tobacco smoke. It is estimated that one-third of bladder cancer cases could be prevented through simple modification of lifestyle choices, in particular cessation of smoking.

CLINICAL PRESENTATION AND DIAGNOSTIC WORKUP

Occasionally, patients will present with flank pain in association with an upper tract renal pelvis or ureter cancer or due to hydronephrosis in association with a bladder tumor obstructing the orifice of the ureter within the bladder. Only in rare cases do patients present with significant cachexia and widespread metastatic disease. For most patients, painless hematuria (either gross or microscopic) represents the initial manifestation of an underlying urinary tract cancer. In females, hematuria due to malignancy can often be mistaken for a urinary tract infection or menstrual bleeding. While treatment with antibiotics is warranted if a concurrent urinary tract infection is noted on initial urinalysis, persistent hematuria requires further workup. Painless hematuria in males is almost always abnormal and should be worked up. Initial investigations in patients of either sex should include urine cytology and visual examination of the bladder by cystoscopy. Cytology is successful in identifying cancer in only 50% of individuals with high-grade bladder cancers. In addition to urine cytology, radiographic evaluation of the kidneys and upper urinary tract by CT urogram should be performed. Because of the increased sensitivity and reduced IV contrast loads, CT urograms have largely replaced IV pyelograms as the preferred upper urinary tract imaging modality. A magnetic resonance (MR) urogram may be substituted in patients with poor renal function. Additional diagnostic testing of the urine to assess for cancer-associated chromosomal changes by fluorescent in situ hybridization, increased levels of nuclear mitotic proteins, increased bladder tumor–associated antigens, or higher levels of staining on cells shed by the bladder may identify some cancers missed by traditional cytology testing. However, they may also produce abnormal results in patients who do not have cancer. For now, these adjunct molecular tests are primarily utilized in detecting recurrent cancer in patients with a prior diagnosis of urinary tract cancer. Small tumors, particularly flat noninvasive tumors of the bladder, may be detected at higher rates with the use of blue light cystoscopy or narrow-band imaging cystoscopy. Both blue light and narrow-band

imaging cystoscopies are now used routinely in the initial workup and subsequent monitoring of patients with bladder cancer. For patients with no bladder abnormalities in whom upper tract tumors are suspected, visualization of the upper urinary tracts and renal pelvises should be performed by ureteroscopy or retrograde pyelography.

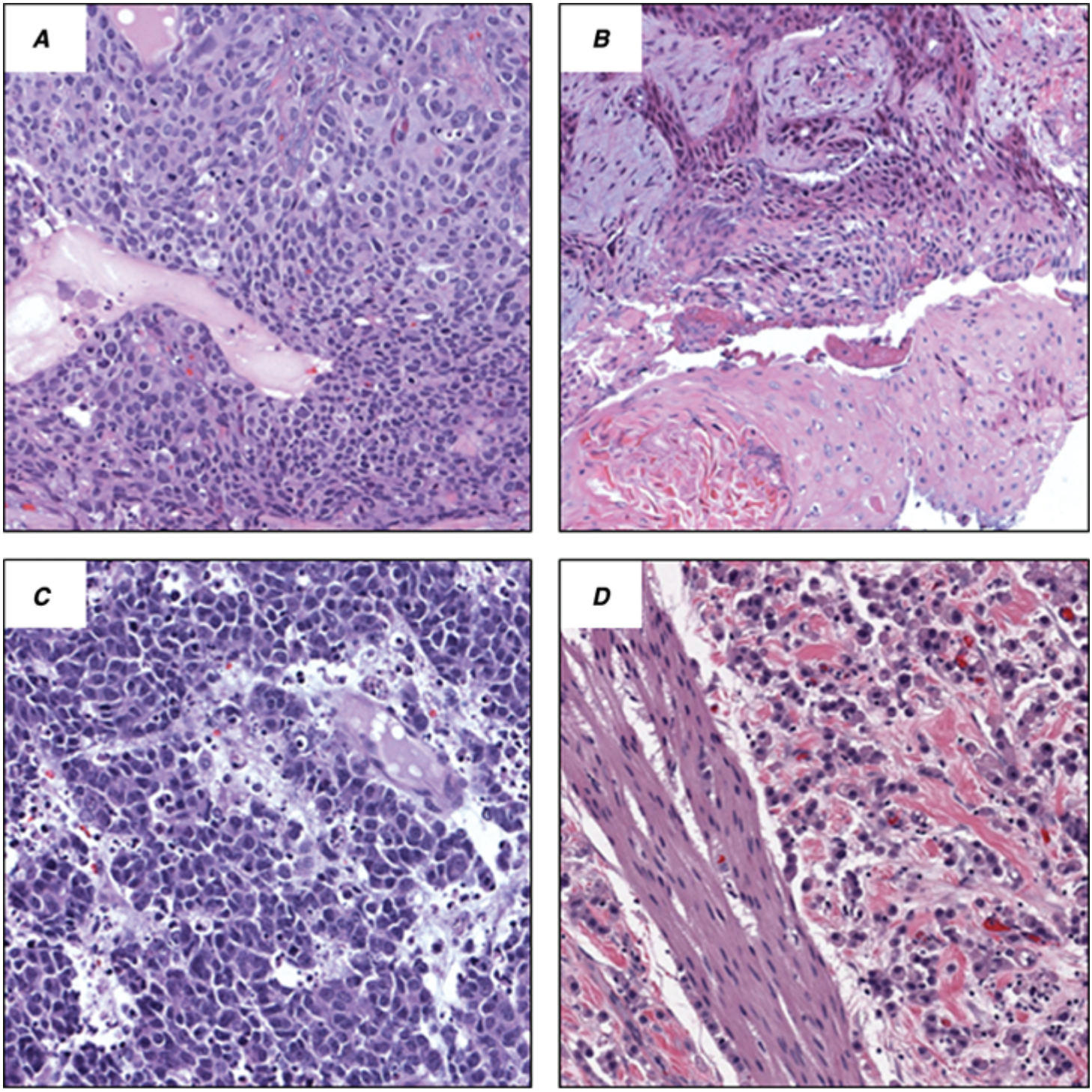
In all patients with abnormalities noted in the bladder or upper urinary tracts, complete endoscopic resection for histologic diagnosis and staging should be performed when possible via either transurethral resection of bladder tumor (TURBT) or endoscopic resection of upper tract tumors.

HISTOLOGY

Urothelial carcinoma, formerly referred to as *transitional cell carcinoma*, is the most common urinary tract cancer histology that is observed in ~90% of cases. Squamous, glandular, micropapillary, plasmacytoid, sarcomatoid, and other variant features can often be found in portions of urothelial carcinoma tumors; however, pure variant histologies are rare. The presence of some variant histologies including micropapillary and plasmacytoid has been associated with worse surgical outcomes compared to urothelial carcinoma. Nonurothelial variant histologies including squamous cell carcinoma, adenocarcinoma, small-cell carcinoma, and carcinosarcoma collectively account for ≤10% of urinary tract tumors. Examples of traditional urothelial carcinoma and some of the variant histologies are shown in Fig. 82-1.

FIGURE 82-1

Bladder and urinary tract cancer histologies. ***A.*** Urothelial carcinoma; ***B.*** squamous cell carcinoma; ***C.*** small-cell carcinoma; ***D.*** plasmacytoid variant. *(Courtesy of Alex Baras, MD, PhD, Johns Hopkins University Department of Pathology.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MOLECULAR BIOLOGY

Clinically, urothelial carcinoma of the bladder displays a biphasic phenotype characterized by (1) low-grade papillary tumors that frequently recur but rarely invade or metastasize and (2) high-grade sometimes flat tumors that invade early leading to lethal metastatic disease. In both of these phenotypes, loss of portions of chromosomes 9q and 9p by loss of heterozygosity analyses is an early molecular event, whose exact significance is not clear. Potential candidate regulatory genes in these genomic regions include *CDNK2A* and

TSC1. Early investigations have demonstrated that low-grade tumors are characterized by alterations in the *RAS/RAF* signaling pathway with activating *FGFR3* mutations or gene fusions present in 60–80% of patients. In contrast, the high-grade invasive phenotype is notable for early deleterious mutations in *TP53* and *RB1*, alterations in *CDH1*, and increased expression of *VEGFR2*. In urothelial carcinoma of the renal pelvis and ureter, 10–20% of cases may be associated with Lynch syndrome hereditary defects in the *MLH1*, *MSH2*, *or MSH6* mismatch repair genes leading to microsatellite instability and frequent DNA mutations. Testing for germline mutations in these genes is recommended in patients with upper urinary tract urothelial carcinoma under the age of 60 at diagnosis, with a first-degree relative with a Lynch syndrome–associated cancer diagnosed under the age of 50, or with two first-degree relatives with a Lynch syndrome–associated cancer regardless of the age at diagnosis.

As genomic analysis technologies have improved, so has our understanding of the molecular biology unique to urothelial carcinoma. In 2014, the initial bladder cancer results of The Cancer Genome Atlas (TCGA) project were published. This effort comprehensively analyzed gene mutations, fusions, expression, copy number variations, methylation, and microRNA across the genome of patients with bladder urothelial carcinoma treated with surgery. While this data set will continue to be analyzed for years to come, the initial findings include (1) genomic alterations in genes (e.g., *FGFR3*, *EGFR*, *ERBB2*, *ERBB3*, *PIK3CA*, *TSC1*, etc.) targetable by currently approved drugs or drugs in development in 69% of patients; (2) genomic alterations in chromatin modifying genes (*KDM6A*, *MLL2*, *CREBBP*, *EP300*, etc.) in 89% of patients; (3) hypermethylation of tumor suppressor genes in 34% of patients; and (4) the identification by RNA sequencing of four distinct intrinsic molecular subtypes (luminal 1, luminal 2, basal 3, and basal 4) closely resembling luminal and basal subclassifications of breast cancers. These initial bladder TCGA findings have led to clinical trial designs enriching for patients with specific gene mutation profiles as well as interrogation of candidate biomarkers according to intrinsic molecular subtypes.

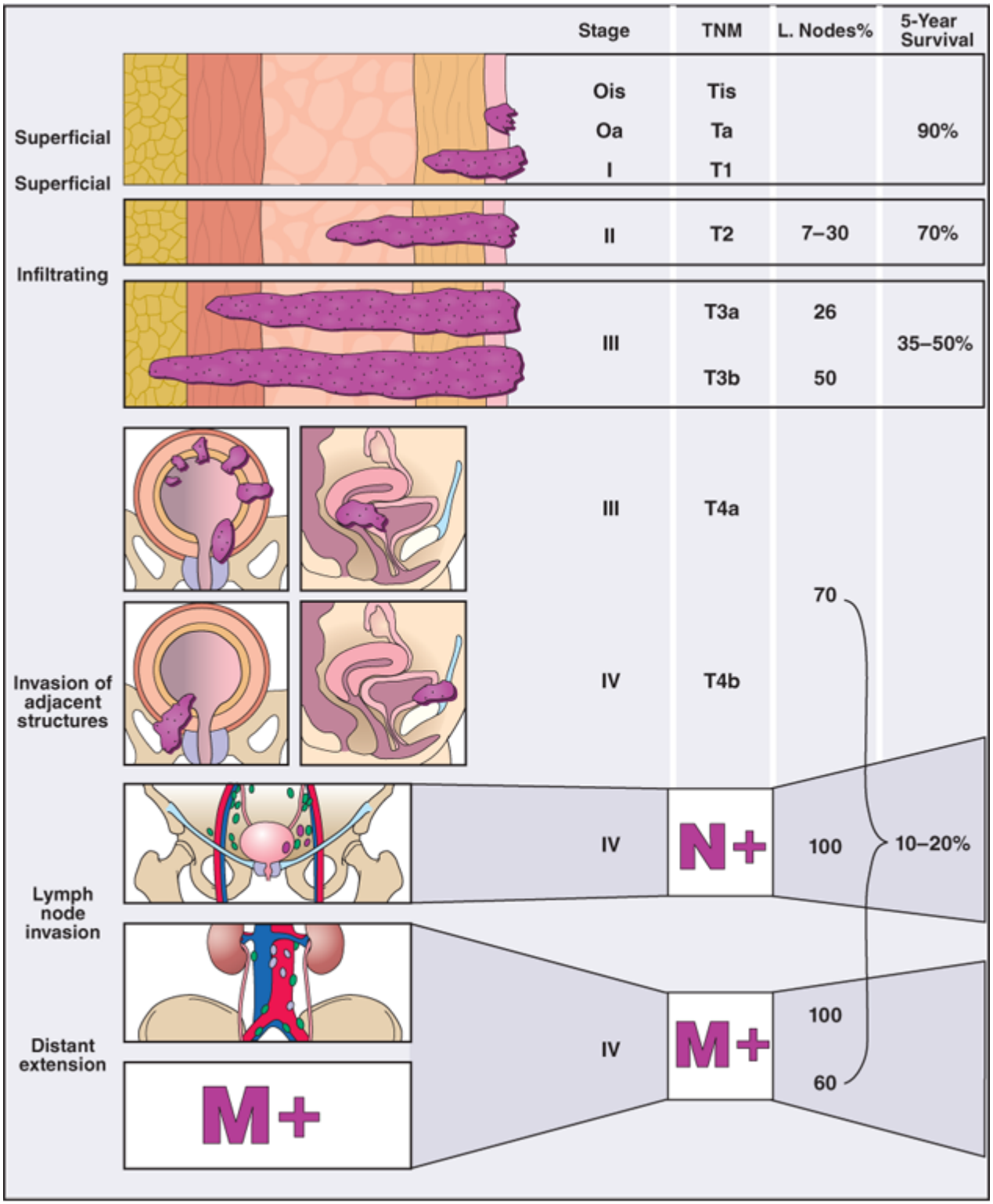
STAGING AND OUTCOMES BY STAGE

The staging of bladder cancer is dependent on the depth of invasion within the bladder wall, involvement of lymph nodes, and spread to surrounding and distant organs as depicted in **Fig. 82-2**. Approximately 75% of bladder cancer presents with non–muscle-invasive bladder cancer (NMIBC), 18% with disease invading into or through the muscular wall of the bladder, and only 3% presenting with metastatic spread to distant organs. NMIBC is defined by tumors that involve only the immediate epithelial layer of cells (carcinoma in situ [CIS] and Ta) or that only penetrate into the connective tissue below the urothelium (T1) but not into the muscular layer known as the *muscularis propria*. Muscle-invasive bladder cancer (MIBC) is defined by tumors that invade into the muscularis propria (T2), through the muscularis propria to involve the surrounding serosa (T3), or into immediately adjacent pelvic organs such as the rectum, prostate, vagina, or cervix (T4). Lymph node staging is classified according to involvement of a solitary node within the true pelvis (N1), multiple nodes involved in the true pelvis (N2), or involvement of the common iliac nodes (N3). Any disease that has spread beyond the true pelvis is considered metastatic (M1). The staging of bladder cancer is driven primarily by the T-stage of the tumor with stages 0a–III defined entirely by the T-stage in the absence of nodal or metastatic disease. Conversely, involvement of either nodal or distant metastases qualifies as stage IV

disease. Clinical outcomes of patients with bladder cancer correlate closely with staging at diagnosis with 5-year overall survival rates of 80% for disease confined to the bladder (stage I–II), 35–50% for disease that penetrates through the bladder (stage III), and only 10–20% for disease extending to surrounding organs, lymph nodes, or metastatic sites (stage IV).

FIGURE 82-2

Bladder cancer staging. TNM, tumor, node metastasis. *(Reprinted with permission from HI Scher, JE Rosenberg, RJ Motzer: Bladder and renal cell carcinomas, in DL Kasper et al [eds]: Harrison's Principles of Internal Medicine, 19th ed. New York, McGraw-Hill, Chap. 114, 2015.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT APPROACHES

Early-Stage Disease

For NMIBC, removal of all visible tumors by TURBT in the operating room is considered the mainstay of surgical treatment. Risk of recurrence can be classified as low, intermediate, or high depending on the

presence of features summarized in **Table 82-1**. For patients with low-risk disease meta-analyses have demonstrated a 12% reduction in early relapses when a single chemotherapy treatment of mitomycin C, epirubicin, or gemcitabine was instilled directly into the bladder (intravesical therapy) within 24 hours of the TURBT. For patients with intermediate- or high-risk tumors, weekly intravesical instillations for 6 consecutive weeks of the attenuated mycobacterium strain known as *Bacille-Calmette Guerin* (BCG) reduce the risk of recurrence at 12 months from 56 to 29%. In addition, BCG treatment has been shown to decrease the rate of progression to MIBC by 27%. Intravesical BCG is generally well tolerated. Side effects can include dysuria, urinary frequency, bladder spasms, hematuria, and, in rare cases (<5%), a systemic inflammatory response that can mimic disseminated BCG infection. Following a 6-week induction BCG schedule, additional maintenance BCG treatments given according to the Southwest Oncology Group schedule further reduce the risk of recurrent NMIBC compared to induction BCG alone. In patients with NMIBC that recurs long after initial BCG treatment, a repeat course of BCG can be considered. For patients with recurrence after a second induction course of BCG or with relapsed NMIBC within 6 months of initial BCG exposure, surgical removal of the entire bladder by cystectomy is recommended due to the high risk of progression to MIBC and potentially metastatic disease. For patients who are not fit enough for or who refuse cystectomy, non-BCG alternative intravesical agents (mitomycin C, gemcitabine, docetaxel, valrubicin) can achieve temporary tumor responses.

TABLE 82-1

Non–Muscle-Invasive Bladder Cancer Recurrence Risk Groups

Risk Group	Characteristics
Low risk	Initial tumor, solitary tumor, low grade, <3 cm, no CIS
Intermediate risk	All tumors not defined in the two adjacent categories (between the category of low and high risk)
High risk	Any of the following: • T1 tumor • High-grade • CIS • Multiple and recurrent and large (>3 cm) Ta low-grade tumors (all conditions must be met for this point on Ta low-grade tumors)

Abbreviation: CIS, carcinoma in situ.

In patients with urothelial carcinoma of the renal pelvis or ureter, endoscopic tissue acquisition and staging are more challenging than primary tumors located in the bladder. Tumors possessing all of the following are

considered low risk: solitary tumor, low grade, size <1 cm, no invasive component on imaging. Low-risk tumors can successfully be treated by laser ureteroscopic ablation or surgical resection and reanastomosis of the remaining ureter ends in tumors that cannot be successfully eradicated endoscopically.

Muscle-Invasive Disease

In patients with urothelial carcinoma of the bladder that invades into or through the muscularis propria but with no evidence of metastatic spread, more aggressive therapy options summarized in **Table 82-2** are required to achieve cure. In carefully selected patients with no evidence of CIS or hydronephrosis, bladder-sparing combined modality therapy with concurrent chemotherapy and radiation can achieve cure in ~65% of patients. Various chemotherapy regimens have been utilized in combination with radiation including cisplatin, carboplatin, 5-fluorouracil, mitomycin C, paclitaxel, and gemcitabine. It is important to note that a maximal debulking of all visible tumor by TURBT is required prior to initiation of combined modality therapy. In patients who achieve a complete response to combined modality therapy, regular cystoscopic monitoring of the bladder is required with salvage cystectomy offered to patients who develop MIBC in follow-up.

TABLE 82-2

Treatment Approaches to MIBC Patients

Treatment	Patient Selection	Clinical Outcomes
Bladder-sparing chemoradiation	No CIS, no hydronephrosis, maximal TURBT required	65% cure, 55% bladder intact, highly dependent on patient selection
Bladder-sparing partial cystectomy	Solitary tumors in dome of bladder are ideal	Variable, highly dependent on patient selection
Cystectomy	Any MIBC patient	50% cure with surgery alone, highly dependent on pathologic stage
Neoadjuvant cisplatin-based chemotherapy	Cisplatin-eligible MIBC patients	5–10% improvement in overall survival compared to cystectomy alone
Adjuvant cisplatin-based chemotherapy	Cisplatin-eligible high-risk postcystectomy MIBC patients (pT3-4, N+)	Similar improvement as neoadjuvant treatment, data less robust, many patients not suitable for adjuvant treatment

Abbreviations: CIS, carcinoma in situ; MIBC, muscle-invasive bladder cancer; TURBT, transurethral resection of bladder tumor.

In a similar fashion, bladder-sparing partial cystectomy can be performed in a very small subset of MIBC patients. The ideal patient for partial cystectomy is the patient with a solitary, clinical T2 urothelial carcinoma in the dome of the bladder. In such patients, the tumor and immediate surrounding urothelium can be resected with reconstruction of the remaining bladder to maintain near physiologic urinary function.

In the majority of patients, however, resection of the entire bladder is required. In males, a cystoprostatectomy with removal of the bladder, prostate, and pelvic lymph nodes is performed while in females an anterior exenteration with removal of the bladder, uterus, ovaries, cervix, and pelvic lymph nodes is performed. With the bladder removed, three options exist to re-route the urine outflow. In an ileostomy, the bilateral ureters are connected to a portion of ileum that is brought through an incision in the abdominal wall to create a stoma that drains urine into an affixed bag outside of the body. In a continent urinary reservoir or “Indiana pouch,” the ureters are connected to a portion of ileum that has been separated on both ends from the rest of the small bowel transit to form a urinary reservoir. The remaining small bowel is reanastomosed, and the urinary reservoir is brought up just beneath the abdominal wall muscles with patients catheterizing the urinary reservoir several times per day via a small stoma tract. Last, in a neobladder, the same urinary reservoir described previously is brought down into the pelvis and is anastomosed to the remaining urethra to provide the opportunity to the patient to void urine through the urethra. The choice of which urinary reconstruction to perform is affected not only by patient choice but also by anatomic tumor considerations and urologist experience with each procedure. Regardless of the type of surgery performed, all patients undergo a significant catabolic change in their metabolism following removal of the bladder. While many MIBC patients are affected by weight loss preoperatively, it is not uncommon for postcystectomy patients to lose an additional 10–15 lb in the first month postoperatively. In addition, patients can experience long-term nutritional changes such as low B_{12} levels due to alterations in small bowel physiology caused by all of the urinary diversion options.

Despite aggressive surgery, only half of patients undergoing cystectomy are cured by surgery alone. Therefore, many clinical trials have investigated the role of systemic chemotherapy before (neoadjuvant) or after (adjuvant) surgery. Meta-analyses have shown a 5–10% absolute overall survival advantage when combination chemotherapy regimens utilizing cisplatin have been used before surgery. A similar benefit exists with cisplatin-based combination chemotherapy given after surgery. However, the data in the adjuvant setting are based on smaller, older trials. Furthermore, in the postoperative setting, some patients may not recover sufficiently from their surgery within a time frame optimal for chemotherapy administration. Importantly, non–cisplatin-containing chemotherapy regimens have proven inferior to cisplatin-containing regimens. Therefore, if patients are not suitable candidates for cisplatin administration due to poor functional status or comorbidities (e.g., poor renal function), patients should proceed directly to surgery and forego neoadjuvant therapy.

For patients with high-risk urothelial carcinoma of the upper urinary tract, resection of the kidney and ureter (including the ureter bladder cuff) by nephroureterectomy is preferred. Segmental ureterectomy may be appropriate in patients with decreased renal function in which nephron-sparing outcomes are critical to prevent the need for dialysis. Similarly, in CIS patients, administration of BCG therapy via a nephrostomy

tube can be considered to preserve intact renal function. In retrospective series, the use of cisplatin-based neoadjuvant chemotherapy has been associated with a pathologic complete response at surgery of 14% in upper tract urothelial carcinoma patients. While adjuvant chemotherapy can be considered in patients with locally advanced stages (T3, T4, or node positive), due to the removal of a kidney at surgery and subsequent associated drop in renal function, many postoperative upper tract patients cannot receive cisplatin-based regimens.

Metastatic Disease

For patients with metastatic urothelial carcinoma regardless of primary tumor origin, systemic chemotherapy is the most established standard of care. In a randomized phase 3 clinical trial, the combination of methotrexate, vinblastine, doxorubicin, and cisplatin (MVAC) demonstrated an improvement in median overall survival from 8.2 to 12.5 months compared to single-agent cisplatin. In a head-to-head randomized phase 3 clinical trial, the combination of cisplatin and gemcitabine (CG) demonstrated similar overall survival compared to MVAC with a more favorable side-effect profile. Since 2000, treatment with either MVAC or CG has remained the standard for first-line treatment of patients with metastatic urothelial carcinoma with adequate renal function and functional status suitable for cisplatin therapy. For patients with lymph node only metastases and good functional status, cure is achieved in 15–20% of such patients. Unfortunately, only ~5% of metastatic patients fulfill both these criteria. For most patients, chemotherapy may prolong survival, but disease resistance proving lethal eventually develops. Furthermore, approximately half of patients with urothelial carcinoma have renal insufficiency, comorbidities, or frail functional status, and are not candidates for cisplatin treatment. In cisplatin-ineligible patients, carboplatin-based chemotherapy regimens are most often used with median overall survival rates decreased to 9.3 months.

Following front-line chemotherapy treatment, second-line chemotherapy regimens have shown modest 10–20% response rates, but no overall survival benefit. In recent years, exponential development of novel immunotherapy approaches has occurred for patients with urothelial carcinoma. The immune checkpoint targets programmed cell death protein 1 (PD-1) and programmed death ligand 1 (PD-L1) have demonstrated the most encouraging clinical benefits. In normal physiology, PD-1/PD-L1 are upregulated in response to inflammation to dampen and prevent an overactive inflammatory response. In cancers including urothelial carcinoma, however, PD-1/PD-L1 are often upregulated on the tumor surface or immune cells in the tumor microenvironment. Upregulated PD-1/PD-L1 in this situation serves as a mechanism of immune escape that facilitates tumor growth. Atezolizumab (an anti–PD-L1 antibody) was the first drug approved in the United States for metastatic urothelial carcinoma in over two decades based on a response rate of 15% in postplatinum patients. Subsequently, pembrolizumab (an anti–PD-1 antibody) demonstrated an improvement in overall survival from 7.4 to 10.3 months compared to standard second-line chemotherapy options. Multiple other PD-1/PD-L1 agents have demonstrated clinical responses in urothelial carcinoma. In addition, clinical trials investigating immunotherapy, chemotherapy, and radiation combinations are ongoing. Last, leveraging the molecular knowledge gained from the TCGA project, clinical trials are also investigating the role of molecularly targeted therapies in patients with metastatic urothelial carcinoma harboring specific genetic alterations predictive of clinical benefit (e.g., activating *FGFR3* mutations or gene

fusions). Collectively, these new emerging options for metastatic urothelial carcinoma patients offer hope for improved outcomes for patients with urothelial carcinoma of all stages in the future.

FURTHER READING

Babjuk M et al: EAU Guidelines on Non-Muscle-Invasive Urothelial Carcinoma of the Bladder: Update 2016. Eur Urol 2016; Available from *http://dx.doi.org.ezproxy1.library.usyd.edu.au/10.1016/j.eururo.2016.05.041*. Accessed December 3, 2016.

Cancer Genome Research Atlas Collaborators: Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 507:315, 2014. [PubMed: 24476821]

Howlader N et al: SEER Cancer Statistics Review, 1975–2013. Available from *https://seer.cancer.gov/statfacts/html/urinb.html.* Accessed December 3, 2016.

Kamat AM et al: Bladder cancer. Lancet 388:2796, 2016. [PubMed: 27345655]

Knowles MA, Hurst CD: Molecular biology of bladder cancer: New insights into pathogenesis and clinical diversity. Nat Rev Cancer 15:25, 2015. [PubMed: 25533674]

Rouprêt M et al: European Association of Urology Guidelines on Upper Urinary Tract Urothelial Cell Carcinoma: 2015 update. Eur Urol 68:868, 2015. [PubMed: 26188393]

Siegel RL et al: Cancer statistics, 2016. CA Cancer J Clin 66:7, 2016. [PubMed: 26742998]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 83: Benign and Malignant Diseases of the Prostate

Howard I. Scher; James A. Eastham

INTRODUCTION

Benign and malignant changes in the prostate increase with age. Autopsies of men in the eighth decade of life show hyperplastic changes in >90% and malignant changes in >70% of individuals. The high prevalence of these diseases among the elderly, who often have competing causes of morbidity and mortality, mandates a risk-adapted approach to diagnosis and treatment. This can be achieved by considering these diseases as a series of states. Each state represents a distinct clinical milestone for which therapy(ies) may be recommended based on disease extent, current symptoms, the risk of developing symptoms, or the risk of death from disease in relation to death from other causes within a given time frame. For benign proliferative disorders, symptoms of urinary frequency, infection, and potential for obstruction are weighed against the side effects and complications of medical or surgical intervention. For prostate malignancies, the likelihood that a clinically significant cancer is present in the gland and the concomitant risk of symptoms or death from cancer are balanced against the morbidities of the recommended treatments and preexisting comorbidities.

ANATOMY AND PATHOLOGY

The prostate is located in the pelvis and is surrounded by the rectum, the bladder, the periprostatic and dorsal vein complexes and neurovascular bundles that are responsible for erectile function, and the urinary sphincter that is responsible for passive urinary control. The prostate is composed of branching tubuloalveolar glands arranged in lobules surrounded by fibromuscular stroma. The acinar unit includes an epithelial compartment made up of epithelial, basal, and neuroendocrine cells and separated by a basement membrane, and a stromal compartment that includes fibroblasts and smooth-muscle cells. Prostate-specific antigen (PSA) and prostatic acid phosphatase (PAP) are produced in the epithelial cells. Both prostate epithelial cells and stromal cells express androgen receptors (ARs) and depend on androgens for growth. Testosterone, the major circulating androgen, is converted by the enzyme 5α-reductase to dihydrotestosterone in the gland.

The periurethral portion of the gland increases in size during puberty and after the age of 55 years due to the growth of nonmalignant cells in the transition zone of the prostate that surrounds the urethra. Most cancers develop in the peripheral zone, and cancers in this location may be palpated during a digital rectal examination (DRE).

PROSTATE CANCER

In 2017, ~161,360 prostate cancer cases were diagnosed and 26,730 men died from prostate cancer in the United States. The absolute number of prostate cancer deaths has decreased in the past 10 years, attributed by some to the widespread use of PSA-based detection strategies. However, the paradox of management is that although 1 in 6 men will eventually be diagnosed with prostate cancer, and the disease remains the second leading cause of cancer deaths in men, only 1 man in 30 with prostate cancer will die of his disease.

EPIDEMIOLOGY

Epidemiologic studies show that the risk of being diagnosed with prostate cancer increases 2.5-fold if one first-degree relative is affected and fivefold if two or more are affected. Current estimates are that 40% of early-onset and 5–10% of all prostate cancers are hereditary. Prostate cancer affects ethnic groups differently. Matched for age, African-American males have a higher incidence and present at a more advanced stage with higher-grade, more aggressive tumors. A high risk in families has been linked to the HPC1 (hereditary prostate cancer 1) susceptibility locus in *RNASEL*. Genome-wide association studies (GWAS) have identified >40 prostate cancer susceptibility loci that are estimated to explain up to 25% of prostate cancer risk. Among the genes implicated in variations in incidence and outcome are single-nucleotide polymorphisms (SNPs) in the vitamin D receptor in African-Americans and variants in the AR, CYP3A4, both involved in the deactivation of testosterone, as well as CYP17, which is involved in steroid biosynthesis.

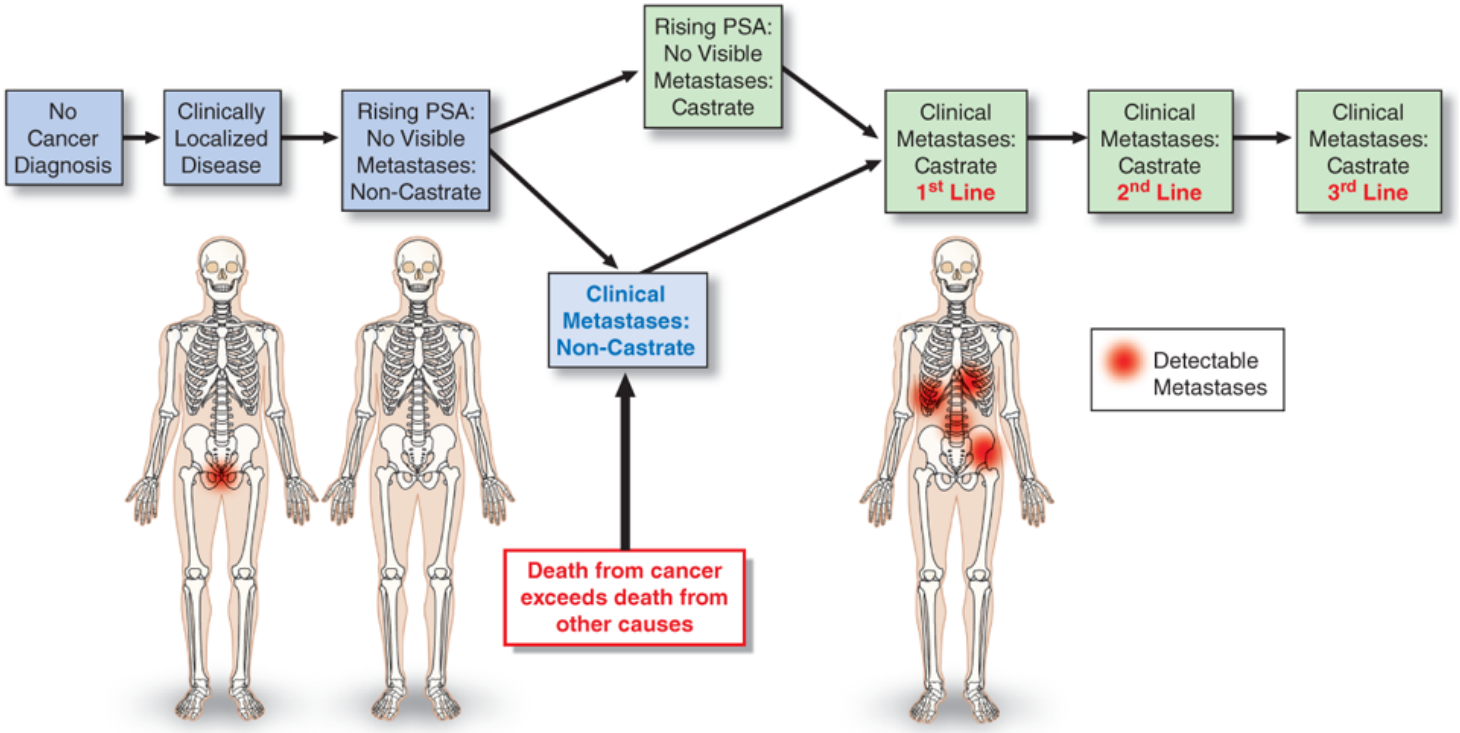
The prevalence of autopsy-detected cancers is similar around the world, while the incidence of clinical disease varies. Thus, environmental and dietary factors may play a role in prostate cancer growth and progression. High consumption of dietary fats, such as α-linoleic acid or polycyclic aromatic hydrocarbons that form when red meats are cooked, is believed to increase risk. Similar to breast cancer in Asian women, the risk of prostate cancer in Asian men increases when they move to Western environments. Protective factors include consumption of the isoflavonoid genistein (which inhibits 5α-reductase), cruciferous vegetables with isothiocyanate sulforaphane, lycopene found in tomatoes, and inhibitors of cholesterol biosynthesis (e.g., statin drugs). The development of prostate cancer is a multistep process. One early change is hypermethylation of the GSTP1 gene promoter, which leads to loss of function of a gene that detoxifies carcinogens. The finding that many prostate cancers develop adjacent to a lesion termed PIA (proliferative inflammatory atrophy) suggests a role for inflammation. Not smoking, regular exercise, and maintaining a healthy body weight may reduce the risk of progression.

DIAGNOSIS AND TREATMENT BY CLINICAL STATE

The prostate cancer continuum—from the appearance of a preneoplastic and invasive lesion that is localized to the gland, to a metastatic lesion causing symptoms and, ultimately, mortality—can span decades. To limit overdiagnosis of clinically insignificant cancers, and for disease management in general, competing risks are considered in the context of a series of clinical states (Fig. 83-1). The states are defined operationally on the basis of whether or not a cancer diagnosis has been established and, for those with a diagnosis, whether or not metastases are detectable on imaging studies and the measured level of testosterone in the blood. With this approach, an individual resides in only one state and remains in that state until he has progressed. At each assessment, the decision to offer treatment and the specific form of treatment are based on the risk posed by the cancer relative to competing causes of morbidity and mortality that may be present in that individual. It follows that the more advanced the disease, the greater the need for treatment.

FIGURE 83-1

Clinical states of prostate cancer. PSA, prostate-specific antigen.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

For those without a cancer diagnosis, the decision to undergo testing to detect a cancer is based on the individual's estimated life expectancy and, separately, the probability that a clinically significant cancer may be present. For those with a prostate cancer diagnosis, the clinical states model considers the probability of developing symptoms or dying from the disease. Thus, a patient with localized tumor that has been surgically removed remains in the state of localized disease as long as the PSA remains undetectable. The time within a state then becomes a measure of the efficacy of an intervention, though the effect may not be assessable for years. Because many men with active cancer are not at risk for developing metastases, symptoms, or death, the clinical states model allows a distinction between *cure*—the elimination of all cancer cells, the primary therapeutic objective of treatment for most cancers—and *cancer control*, by which the tempo of the illness is determined to be so slow or has been altered to the point where it is unlikely to cause symptoms, to metastasize, or to shorten a patient's life expectancy. Importantly,

from a patient standpoint, both outcomes can be considered equivalent therapeutically, assuming the patient has not experienced symptoms of the disease or the treatment needed to control it. Even when a recurrence is documented, immediate therapy is not always necessary. Rather, as at the time of diagnosis, the need for intervention is based on the tempo of the illness as it unfolds in the individual, relative to the risk-to-benefit ratio of the intervention being considered.

NO CANCER DIAGNOSIS

Prevention

No agent is currently approved for the prevention of prostate cancer. The results from several large double-blind, randomized chemoprevention trials have established 5α-reductase inhibitors (5ARI) as the predominant therapy to reduce the future risk of a prostate cancer diagnosis. The Prostate Cancer Prevention Trial (PCPT), in which men aged >55 years received placebo or the 5ARI finasteride, which inhibits the type 1 isoform, showed a 25% (95% confidence interval 19–31%) reduction in prostate cancer incidence from 24% with placebo to 18% with finasteride. In REDUCE (Reduction by Dutasteride of Prostate Cancer Events trial), a reduction in incidence from 25% with placebo to 20% with dutasteride was found ($p = 0.001$). Dutasteride inhibits both the type 1 and type 2 5ARI isoforms. While both studies met their endpoint, there was concern that most of the cancers that were prevented were low-risk and that there was a slightly higher rate of clinically significant cancers (those with higher Gleason score) in the treatment arm. Neither drug is approved for prostate cancer prevention. In comparison, the Selenium and Vitamin E Cancer Prevention Trial (SELECT), which enrolled African-American men aged ≥50 years and others aged ≥55 years, showed no difference in cancer incidence in patients receiving vitamin E (4.6%) or selenium (4.9%) alone or in combination (4.6%) relative to placebo (4.4%). A similar lack of benefit for vitamin E, vitamin C, and selenium was seen in the Physicians Health Study II.

Screening/Early Detection and Diagnosis

The need to pursue a diagnosis of prostate cancer must balance the benefit from detecting and treating clinically significant cancers that, left untreated, would adversely affect patients' quality and duration of life against the morbidity associated with overdiagnosis and overtreatment of clinically insignificant cancers that are highly prevalent in the general population. The balance is best approached through shared decision making between the patient and physician. Considerations for whether to pursue a diagnosis include symptoms, an abnormal DRE, or more typically, a change in or an elevated serum PSA. Genetic risk is also considered. The urologic history should focus on symptoms of outlet obstruction, continence, potency, or change in ejaculatory pattern.

PHYSICAL EXAMINATION

The DRE focuses on prostate size and consistency and abnormalities within or beyond the gland. Many cancers occur in the peripheral zone and may be palpated on DRE. Carcinomas are characteristically hard, nodular, and irregular, while induration may also be due to benign prostatic hyperplasia (BPH) or calculi. Overall, 20–25% of men with an abnormal DRE have prostate cancer.

PROSTATE-SPECIFIC ANTIGEN

PSA (kallikrein-related peptidase 3; *KLK3*) is a kallikrein-related serine protease that causes liquefaction of seminal coagulum. It is produced by both nonmalignant and malignant epithelial cells and, as such, is prostate-specific, not prostate cancer–specific. Serum levels may also increase from prostatitis and BPH. Serum levels are not significantly affected by DRE, but the performance of a cystoscopy or prostate biopsy can increase PSA levels up to tenfold for 8–10 weeks. PSA circulating in the blood is inactive and mainly occurs as a complex with the protease inhibitor α_1-antichymotrypsin and as free (unbound) PSA forms. The formation of complexes between PSA, α_2-macroglobulin, or other protease inhibitors is less significant. Free PSA is rapidly eliminated from the blood by glomerular filtration with an estimated half-life of 12–18 h. Elimination of PSA bound to α_1-antichymotrypsin is slow (estimated half-life of 1–2 weeks) as it too is largely cleared by the kidneys. Levels should be undetectable after about 6 weeks if the prostate has been completely removed (radical prostatectomy). Immunohistochemical staining for PSA can be used to establish a prostate cancer diagnosis.

PSA testing was approved by the U.S. Food and Drug Administration (FDA) in 1994 for early detection of prostate cancer, and the widespread use of the test has played a significant role in the proportion of men diagnosed with early-stage cancers: more than 70–80% of newly diagnosed cancers are clinically organ confined. The level of PSA in blood is strongly associated with the risk and outcome of prostate cancer. A single PSA measured at age 60 is associated (area under the curve [AUC] of 0.90) with lifetime risk of death from prostate cancer. Most (90%) prostate cancer deaths occur among men with PSA levels in the top quartile (>2 ng/mL), although only a minority of men with PSA >2 ng/mL will develop lethal prostate cancer. Despite this and mortality rate reductions reported from large randomized prostate cancer screening trials, routine use of the test remains controversial.

In 2012, the United States Preventive Services Task Force (USPSTF) published a review of the evidence for PSA-based screening for prostate cancer and made a clear recommendation against screening. By giving a grade of "D" in the recommendation statement that was based on this

review, the USPSTF concluded that “there is moderate or high certainty that this service has no net benefit or that the harms outweigh the benefits.” In 2013, the American Urological Association (AUA) updated their consensus statement regarding prostate cancer screening. They concluded that the quality of evidence for the benefits of screening was moderate for men aged 55–69 years. For men outside this age range, evidence was lacking for benefit, but the harms of screening, including overdiagnosis and overtreatment, remained. The AUA recommends shared decision-making for men aged 55–69 years considering PSA-based screening, a target age group for whom benefits may outweigh harms. Outside this age range, PSA-based screening as a routine was not recommended. The entire guideline is available at *http://www.auanet.org/guidelines/early-detection-of-prostate-cancer-(2013-reviewed-and-validity-confirmed-2015)*. As of 2017, the USPSTF has issued a draft of a revised recommendation with a grade of “C” for PSA-based prostate cancer screening for men aged 55–69. They recommend shared decision-making for men aged 55–69 and do not recommend screening for men aged ≥70; this is now roughly in agreement with the 2013 AUA guideline. The USPSTF notes that the increased use of active surveillance (observation with selective delayed treatment) for low-risk prostate cancer has reduced the risks of screening.

We believe that implementation of the following three guidelines will further improve PSA screening outcomes in the United States and will have a greater practical impact on men’s health than the USPSTF and AUA recommendations that are based almost solely on age. First, avoid PSA tests in men with little to gain. There is no rationale for recommending PSA screening in asymptomatic men with a short life expectancy. Hence, men aged >75 years should only be tested in special circumstances, such as higher than median PSAs measured before age 70 or excellent overall health. In addition, because a baseline PSA is a strong predictor of the future risk of lethal prostate cancer, men with low PSAs, for example <1 ng/mL, can undergo testing less frequently, perhaps every 5 years, with screening possibly ending at age 60 if the PSA remains at ≤1 ng/mL. Men with PSAs that are above age median but below biopsy thresholds can be counseled about their elevated risk and actively encouraged to return for regular screening and more comprehensive risk assessment. Second, do not treat those who do not need treatment. High proportions of men with screen-detected prostate cancer do not need immediate treatment and can be managed by active surveillance. Third, refer men who do need treatment to high-volume centers. Although it is clearly not feasible to restrict treatment exclusively to high-volume centers, shifting treatment trends so that more patients are treated at such centers by high-volume providers will improve cancer control and decrease complications. The goal of prostate cancer screening should be to maximize the benefits of PSA testing and minimize its harms. Following the three rules outlined here should continue to improve the ratio of harms to benefits from PSA screening.

The PSA criteria used to recommend a diagnostic prostate biopsy have evolved over time. However, based on the commonly used cut-point for prostate biopsy (a total PSA ≥4 ng/mL), most men with a PSA elevation do not have histologic evidence of prostate cancer at biopsy. In addition, many men with PSA levels below this cut-point harbor cancer cells in their prostate. Information from the Prostate Cancer Prevention Trial demonstrates that there is no PSA below which the risk of prostate cancer is zero. Thus, the PSA level establishes the likelihood that a man will harbor cancer if he undergoes a prostate biopsy. The goal is to increase the sensitivity of the test for younger men more likely to die of the disease and to reduce the frequency of detecting cancers of low malignant potential in elderly men more likely to die of other causes. Patients with symptomatic prostatitis should have a course of antibiotics before biopsy. However, the routine use of antibiotics in an asymptomatic man with an elevated PSA level is strongly discouraged.

SECOND-LINE SCREENING TESTS

The 4Kscore® Test (Opko Lab, Nashville, TN) measures four prostate-specific kallikreins (total PSA, free PSA, intact PSA, and human kallikrein 2). The results are combined with clinical information in an algorithm that estimates an individual’s percent risk of being found to harbor an aggressive prostate cancer should that individual opt for a prostate biopsy. The 4Kscore test has also been shown to identify the likelihood than an individual will develop aggressive prostate cancer, defined as high grade prostate cancer pathology and/or poor prostate cancer clinical outcomes, within 20 years.

Prostate Health Index (PHI™, Innovative Diagnostic Laboratory, Richmond, VA) is a blood test that estimates the risk of having prostate cancer. The PHI test is a combination of the free PSA, total PSA, and the [-2]proPSA isoform of free PSA. These three tests are combined in a formula that calculates the PHI score. The PHI score is a better predictor of prostate cancer than the total PSA test alone or the free PSA test alone.

PROSTATE BIOPSY

A diagnosis of cancer is established by an image-guided needle biopsy. Direct visualization by transrectal ultrasound (TRUS), magnetic resonance imaging (MRI), or fusion of the ultrasound and MRI images ensures that all areas of the gland including suspicious areas are sampled. Contemporary schemas advise an extended-pattern 12-core biopsy that includes sampling from the peripheral zone as well as a lesion-directed palpable nodule or suspicious image-guided sampling. Because a prostate biopsy is subject to sampling error, men with an abnormal PSA and negative biopsy are frequently advised to undergo additional testing which may include a 4Kscore Test, PHI, prostate MRI, and/or repeat biopsy.

PATHOLOGY

Each core of the biopsy is examined for the presence of cancer, and the amount of cancer is quantified based on the length of the cancer within the core and the percentage of the core involved. Of the cancers identified, >95% are adenocarcinomas; the rest are squamous or transitional

cell tumors or, rarely, carcinosarcomas. Metastases to the prostate are rare, but in some cases colon cancers or transitional cell tumors of the bladder invade the gland by direct extension.

When prostate cancer is diagnosed, a measure of histologic aggressiveness is assigned using the *Gleason grading system*, in which the dominant and secondary glandular histologic patterns are scored from 1 (well-differentiated) to 5 (undifferentiated) and summed to give a total score of 2–10 for each tumor. The most poorly differentiated area of tumor (i.e., the area with the highest histologic grade) often determines biologic behavior. The presence or absence of perineural invasion and extracapsular spread are also recorded.

Over the years, the Gleason grading system has undergone several changes. Currently, Gleason total scores 2–5 are no longer assigned and in practice the lowest total score is now assigned a 6, although the scale continues to range from 2 to 10. This leads to a logical yet incorrect assumption on the part of patients that their Gleason 6 cancer is in the middle of the scale, triggering the fear that their cancer is serious and the assumption that treatment is necessary despite Gleason score 6 actually being favorable risk. To address these issues, a new 5-grade group system has been developed:

Grade Group 1 (Gleason score ≤6)

Grade Group 2 (Gleason score 3+4 = 7)

Grade Group 3 (Gleason score 4+3 = 7)

Grade Group 4 (Gleason score 4+4 = 8)

Grade Group 5 (Gleason scores 9 and 10)

The new system simplifies the grading of prostate cancer, appropriately classifies the lowest risk as Grade Group 1 (rather than Gleason score 6), and accurately predicts prognosis.

PROSTATE CANCER STAGING

The TNM (tumor, nodes, metastasis) staging system includes categories for cancers that are identified solely on the basis of an abnormal PSA (T1c), those that are palpable but clinically confined to the gland (T2), and those that have extended outside the gland (T3 and T4) **(Table 83-1, Fig. 83-2)**. DRE alone is inaccurate in determining the extent of disease within the gland, the presence or absence of capsular invasion, involvement of seminal vesicles, and extension of disease to lymph nodes. Because of the inadequacy of DRE for staging, the TNM staging system was modified to include the results of imaging. Unfortunately, no single test has proven to accurately indicate the stage or the presence of organ-confined disease, seminal vesicle involvement, or lymph node spread.

TABLE 83-1

TNM Classification

TNM (tumor, nodes, metastasis) Staging System for Prostate Cancer[a]	
Tx	Primary tumor cannot be assessed
T0	No evidence of primary tumor
Localized Disease	
T1	Clinically inapparent tumor, neither palpable nor visible by imaging
T1a	Tumor incidental histologic finding in ≤5% of resected tissue; not palpable
T1b	Tumor incidental histologic finding in >5% of resected tissue
T1c	Tumor identified by needle biopsy (e.g., because of elevated PSA)
T2	Tumor confined within prostate[b]
T2a	Tumor involves half of one lobe or less
T2b	Tumor involves more than one half of one lobe, not both lobes
T2c	Tumor involves both lobes
Local Extension	
T3	Tumor extends through the prostate capsule[c]
T3a	Extracapsular extension (unilateral or bilateral)
T3b	Tumor invades seminal vesicles
T4	Tumor is fixed or invades adjacent structures other than seminal vesicles such as external sphincter, rectum, bladder, levator muscles, and/or pelvic wall
Metastatic Disease	
N1	Positive regional lymph nodes
M1	Distant metastases

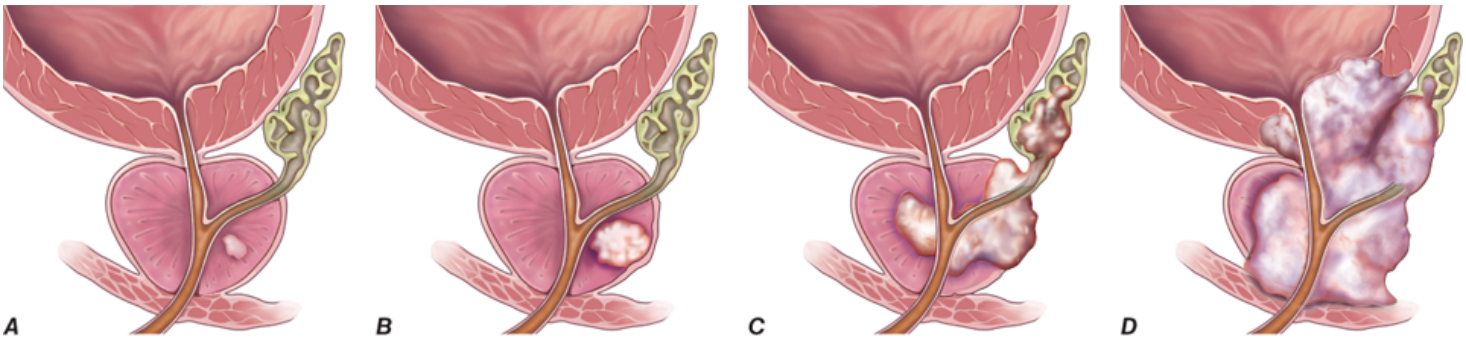
[a]Revised from SB Edge et al (eds): *AJCC Cancer Staging Manual*, 7th ed. New York, Springer, 2010. [b]Tumor found in one or both lobes by needle biopsy, but not palpable or reliably visible by imaging, is classified as T1c. [c]Invasion into the prostatic apex or into (but not beyond) the prostatic capsule is classified not as T3 but as T2.

Abbreviation: PSA, prostate-specific antigen.

FIGURE 83-2

T stages of prostate cancer. ***A.*** T1—Clinically inapparent tumor, neither palpable nor visible by imaging; ***B.*** T2—Tumor confined within prostate; ***C.*** T3—Tumor extends through prostate capsule and may invade the seminal vesicles; ***D.*** T4—Tumor is fixed or invades adjacent structures. Eighty percent of patients present with local disease (T1 and T2), which is associated with a 5-year survival rate of 100%. An additional 12% of patients present with regional disease (T3 and T4 without metastases), which is also associated with a 100% survival rate after 5 years. Four

percent of patients present with distant disease (T4 with metastases), which is associated with a 30% 5-year survival rate. (Three percent of patients are ungraded.) *(Data from AJCC, http://seer.cancer.gov/statfacts/html/prost.html. Figure © Memorial Sloan-Kettering Cancer Center Medical Graphics; used with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TRUS is the imaging technique most frequently used to assess the primary tumor, but its chief use is directing prostate biopsies, not staging. No TRUS finding consistently indicates cancer with certainty. Computed tomography (CT) lacks sensitivity and specificity to detect extraprostatic extension and is inferior to MRI in visualization of lymph nodes. In general, MRI is superior to CT to detect cancer in the prostate and to assess local disease extent. T1-weighted MRI produces a high signal in the periprostatic fat, periprostatic venous plexus, perivesicular tissues, lymph nodes, and bone marrow. T2-weighted MRI demonstrates the internal architecture of the prostate and seminal vesicles. Most cancers have a low signal, while the normal peripheral zone has a high signal, although the technique lacks sensitivity and specificity. MRI is also useful for the planning of surgery and radiation therapy.

Radionuclide bone scans (bone scintigraphy) are used to evaluate spread to osseous sites. This test is sensitive but relatively nonspecific because areas of increased uptake are not always related to metastatic disease. Healing fractures, arthritis, Paget's disease, and other conditions will also cause abnormal uptake. True-positive bone scans are uncommon when the PSA is <10 ng/mL unless the tumor is high-grade.

TREATMENT

TREATMENT

LOCALIZED DISEASE OR CLINICALLY LOCALIZED DISEASE

Cancers are those that appear to be nonmetastatic after staging studies are performed. Patients with clinically localized disease are managed by radical prostatectomy, radiation therapy, or active surveillance. Choice of therapy requires the consideration of several factors: the presence of symptoms, the probability that the untreated tumor will adversely affect the quality or duration of survival and thus require treatment, and the probability that the tumor can be cured by single-modality therapy directed at the prostate versus requiring both local and systemic therapy to achieve cure.

Data from the literature (such as the ProtecT trial) do not provide clear evidence for the superiority of any one form of local therapy relative to another. This is due to the lack of prospective randomized trials, referral bias and physician bias, variation in the experience of the treating teams, and differences in trial endpoints and the definitions of cancer control. Often, PSA relapse–free survival is used because an effect on metastatic progression or survival may not be apparent for years. For many patients, however, a PSA recurrence does not necessarily mean that the disease will cause symptoms or shorten survival. After radical surgery to remove all prostate tissue, PSA should become undetectable in the blood within 6 weeks. If PSA remains or becomes detectable after radical prostatectomy, the patient is considered to have persistent or recurrent disease. After radiation therapy, in contrast, PSA does not become undetectable because the remaining nonmalignant elements of the gland continue to produce PSA even if all cancer cells have been eliminated. Similarly, cancer control is not well defined for a patient managed by active surveillance because PSA levels may continue to rise in the absence of therapy. Other outcomes are time to objective progression (local or systemic), cancer-specific survival, and overall survival; however, these outcomes may take years to assess.

The more extensive the local disease, the higher the probability of regional lymph node involvement even when imaging studies are normal, the lower the probability of local control, and the higher the probability of systemic relapse. More important is that within the categories of T1, T2, and T3 disease are cancers with a range of prognoses. Some T3 tumors are curable with therapy directed solely at the prostate, and some T1 lesions have a high probability of systemic relapse that requires the integration of local and systemic therapy to achieve cure. For T1c cancers in particular, stage alone is inadequate to predict outcome and select treatment; other factors must be considered.

To better assess risk and guide treatment selection, many groups have developed prognostic models or nomograms that use a combination of the initial clinical T stage, biopsy Gleason score, the number of biopsy cores in which cancer is detected, and baseline PSA. Some use discrete cut-points (PSA <10 or ≥10 ng/mL; Gleason score of ≤6, 7, or ≥8); others employ nomograms that use PSA and Gleason score as continuous variables. More than 100 nomograms have been reported to predict (a) the probability that a clinically significant cancer is present, (b) disease extent (organ-confined vs non–organ-confined, node-negative or -positive), or (c) the probability of treatment success for specific local therapies using pretreatment variables. Considerable controversy exists over what constitutes “high risk” based on a predicted probability of success or failure. In these situations, nomograms and predictive models can only go so far. Exactly what probability of success or failure would lead a physician to recommend and a patient to seek alternative approaches is controversial. As an example, it may be appropriate to recommend radical surgery for a younger patient with a low probability of cure. Nomograms are being refined continually to incorporate additional clinical parameters, biologic determinants, and year of treatment, which can also affect outcomes, making treatment decisions a dynamic process.

The frequency of adverse events varies by treatment modality and the experience of the treating team. For example, following radical prostatectomy, incontinence rates range from 2 to 47% and impotence rates range from 25 to 89%. Part of the variability relates to how the complication is defined and whether the patient or physician is reporting the event. The time of the assessment is also important. After surgery, impotence is immediate but may reverse over time, while with radiation therapy impotence is not immediate but may develop over time. Of greatest concern to patients are the effects on continence, sexual potency, and bowel function.

Radical Prostatectomy

The goal of radical prostatectomy is to excise the cancer completely with a clear margin, to maintain continence by preserving the external sphincter, and to preserve potency by sparing the autonomic nerves in the neurovascular bundle. The procedure is advised for patients with a life expectancy of 10 years or more and is performed via a retropubic or perineal approach, or via a minimally invasive robotic-assisted or hand-held laparoscopic approach. Outcomes can be predicted using postoperative nomograms that consider pretreatment factors and the pathologic findings at surgery. PSA failure is usually defined as a value >0.1 or 0.2 ng/mL. Specific criteria to guide the choice of one approach over another are lacking. Minimally invasive approaches offer the advantage of a shorter hospital stay and reduced blood loss. Rates of cancer control, recovery of continence and recovery of erectile function are comparable. The individual surgeon rather than the surgical approach used is most important in determining outcomes after surgery.

Neoadjuvant hormonal treatment with gonadotropin-releasing hormone (GnRH) agonists/antagonists alone has also been explored in an attempt to improve the outcomes of surgery for high-risk patients using a variety of definitions. The results of several large trials testing 3 or 8 months of androgen depletion before surgery showed that serum PSA levels decreased by 96%, prostate volumes decreased by 34%, and margin positivity rates decreased from 41 to 17%. Unfortunately, these findings have not been shown to improve PSA relapse–free survival.

Factors associated with incontinence following radical prostatectomy include older age and urethral length, which impacts the ability to preserve the urethra beyond the apex and the distal sphincter. The skill and experience of the surgeon are also factors.

The likelihood of recovery of erectile function is associated with younger age, quality of erections before surgery, and the absence of damage to the neurovascular bundles. In general, erectile function begins to return about 6 months after surgery if neurovascular tissue has been preserved. Potency is reduced by half if at least one neurovascular bundle is sacrificed. Overall, with the availability of drugs such as sildenafil, intraurethral inserts of alprostadil, and intracavernosal injections of vasodilators, many patients recover satisfactory sexual function.

Radiation Therapy

Radiation therapy is given by external beam, by radioactive sources implanted into the gland, or by a combination of the two techniques.

External beam radiation therapy

Contemporary external beam intensity-modulated radiation therapy (IMRT) permits shaping of the dose, and allows the delivery of higher doses to the prostate and a dramatic reduction in normal tissue exposure compared to three-dimensional conformal treatment alone. These advances have enabled the safe administration of doses >80 Gy and resulted in higher local control rates and fewer side effects.

Cancer control after radiation therapy has been defined by various criteria, including a decline in PSA to <0.5 or 1 ng/mL, “nonrising” PSA values, and a negative biopsy of the prostate 2 years after completion of treatment. The current standard definition of biochemical failure (the Phoenix definition) is a rise in PSA by ≥2 ng/mL higher than the lowest PSA achieved. The date of failure is “at call” and not backdated.

Radiation dose is critical to the eradication of prostate cancer. In a representative study, a PSA nadir of <1.0 ng/mL was achieved in 90% of patients receiving 75.6 or 81.0 Gy vs 76 and 56% of those receiving 70.2 and 64.8 Gy, respectively. Positive biopsy rates at 2.5 years were 4% for those treated with 81 Gy vs 27 and 36% for those receiving 75.6 and 70.2 Gy, respectively.

More recently, hypofractionation schedules, utilizing fewer treatments of higher radiation doses, have been evaluated and shown to provide good cancer control rates based on post-treatment biopsies showing no evidence of cancer, with no apparent increase in treatment-related morbidity. Hypofractionated treatments can range from as few as 5 treatments to upwards of 26 treatments, both regimens representing substantial reductions in treatment length.

Multiple clinical trials have evaluated the use of androgen deprivation therapy (ADT) in combination with radiation. In patients with intermediate-risk prostate cancer, short-course ADT (6 months), when combined with external beam radiotherapy, has demonstrated significant improvements in overall survival. In patients with high-risk disease, longer courses of ADT (18–36 months) have proven superior to shorter courses and represent the current standard of care when combined with radiotherapy.

Neoadjuvant hormone therapy before radiation therapy is used to decrease the size of the prostate and, consequently, to reduce the exposure of normal tissues to full-dose radiation, to increase local control rates, and to decrease the rate of systemic failure. Short-term hormone therapy can reduce toxicities and improve local control rates, but long-term treatment (2–3 years) is needed to prolong the time to PSA failure and lower the risk of metastatic disease in men with high-risk cancers. The impact on survival has been less clear.

The Prostate Testing for Cancer and Treatment (ProtecT) trial investigated the effects of active monitoring, radical prostatectomy, and radical radiotherapy with hormones on patient-reported outcomes in men diagnosed with low- and intermediate-risk prostate cancer (about 75% with Gleason score 6 or Gleason Grade Group 1 cancer). Patient-reported outcomes among 1643 men who completed questionnaires before diagnosis, at 6 and 12 months, and annually thereafter were compared. Of the three treatments, prostatectomy had the greatest negative effect on sexual function and urinary continence, and although there was some recovery, these outcomes remained worse in the prostatectomy group than in the other groups throughout the trial. The negative effect of radiotherapy on sexual function was greatest at 6 months, but sexual function then recovered somewhat and was stable thereafter; radiotherapy had little effect on urinary continence. Sexual and urinary function declined gradually in the active-monitoring group. Bowel function was worse in the radiotherapy group at 6 months than in the other groups but then recovered somewhat, except for the increasing frequency of bloody stools; bowel function was unchanged in the other groups. Urinary voiding and nocturia were worse in the radiotherapy group at 6 months but then mostly recovered and were similar to the other groups after 12 months. Effects on quality of life mirrored the reported changes in function. No significant differences were observed among the groups in measures of anxiety, depression, or general health-related or cancer-related quality of life.

Brachytherapy

Brachytherapy is the direct implantation of radioactive sources (seeds) into the prostate. It is based on the principle that the deposition of radiation energy in tissues decreases as a function of the square of the distance from the source (Chap. 69). The goal is to deliver intensive irradiation to the prostate, minimizing the exposure of the surrounding tissues. The current standard technique achieves a more homogeneous dose distribution by placing seeds according to a customized template based on imaging assessment of the cancer and computer-optimized dosimetry. The implantation is performed transperineally as an outpatient procedure with real-time imaging.

Improvements in brachytherapy techniques have resulted in fewer complications and a marked reduction in local failure rates. In a series of 197 patients followed for a median of 3 years, 5-year actuarial PSA relapse–free survival for patients with pretherapy PSA levels of 0–4, 4–10, and >10 ng/mL were 98, 90, and 89%, respectively. In a separate report of 201 patients who underwent posttreatment biopsies, 80% were negative, 17% were indeterminate, and 3% were positive. The results did not change with longer follow-up. Nevertheless, many physicians feel that implantation is best reserved for patients with good or intermediate prognostic features.

Brachytherapy is well tolerated, although most patients experience urinary frequency and urgency that can persist for several months. Incontinence has been seen in 2–4% of cases. Higher complication rates are observed in patients who have undergone a prior transurethral resection of the prostate (TURP), while those with obstructive symptoms at baseline are at a higher risk for retention and persistent voiding symptoms. Proctitis has been reported in <2% of patients.

Active surveillance

Although prostate cancer is the most common form of cancer affecting men in the United States, patients are being diagnosed earlier and more frequently present with early-stage disease. Active surveillance, described previously as *watchful waiting* or *deferred therapy*, evolved from (1) studies that evaluated predominantly elderly men with well-differentiated tumors who demonstrated no clinically significant progression for protracted periods, (2) recognition of the contrast between incidence and disease-specific mortality, (3) the high prevalence of autopsy cancers, and (4) an effort to reduce overtreatment and treatment-related side effects. In practice, active surveillance is the treatment recommended to patients with cancers of low aggressiveness that can be safely monitored at fixed intervals with DREs, PSA measurements, imaging (usually prostate MRI), and repeat prostate biopsies as indicated until histopathologic or serologic changes correlative of progression warrant treatment with curative intent.

Case selection is critical, and determining clinical parameters predictive of cancer aggressiveness that can be used to reliably select men most likely to benefit from treatment by active surveillance is an area of intense study. In one prostatectomy series, it was estimated that 10–15% of those treated had “insignificant” disease. One set of criteria includes men with clinical T1c tumors that are biopsy Gleason grade 6 (Grade Group 1) involving 3 or fewer cores, each core having <50% involvement by tumor, and a PSA density of <0.15.

Concerns about active surveillance include the limited ability to predict pathologic findings by needle biopsy even when multiple cores are obtained, the recognized multifocality of the disease, and the possibility of a missed opportunity to cure the disease. Nomograms to help predict which patients can safely be managed by active surveillance continue to be refined, and as their predictive accuracy improves, it can be anticipated that more patients will be candidates.

RISING PSA AFTER DEFINITIVE LOCAL THERAPY

It includes patients in whom the sole manifestation of disease is a rising PSA after surgery and/or radiation therapy. By definition, there is no evidence of disease on imaging studies. For these patients, the central issue is whether the rise in PSA results from persistent disease in the primary site, systemic disease, or both. In theory, disease in the primary site may still be curable by additional local treatment.

The decision to recommend radiation therapy after prostatectomy is guided by the pathologic findings at surgery and an MRI of the prostate or prostate bed, as CT and radionuclide bone scan are typically uninformative. Others recommend that a biopsy of the urethrovesical anastomosis be obtained before considering radiation. New PET tracers such as C-11 choline, F-18 fluciclovine, (both FDA approved) and F-18 or Ga-68 PSMA (prostate-specific membrane antigen) are more sensitive and can detect low-volume disease in the prostate bed or other sites to better inform the decision to recommend additional local therapies. Detection rates, both in and outside the prostate bed, correlate with the absolute level of PSA. Factors that predict for response to salvage radiation therapy are a positive surgical margin, lower Gleason score in the radical prostatectomy specimen, long interval from surgery to PSA failure, slow PSA doubling time, and low (<0.5–1 ng/mL) PSA value at the time of radiation treatment. Radiation therapy is generally not recommended if the PSA was persistently elevated after surgery, which usually indicates that the disease had spread outside of the area of the prostate bed and is unlikely to be controlled with radiation therapy. As is the case for other disease states, nomograms to predict the likelihood of success are available.

For patients with a rising PSA after radiation therapy, salvage local therapy can be considered if the disease was “curable” at the outset, if persistent disease has been documented by a biopsy of the prostate or by PET or other imaging, and if no disease is detectable outside of the prostate bed or regional lymph nodes. Unfortunately, case selection is poorly defined in most series, and morbidities are significant. Options include salvage radical prostatectomy, salvage cryotherapy, salvage radiation therapy and salvage irreversible electroporation.

The rise in PSA after surgery or radiation therapy may indicate subclinical or micrometastatic disease with or without local recurrence. In these cases, the need for treatment depends, in part, on the estimated probability that the patient will show evidence of metastatic disease on a scan and in what time frame. That immediate therapy is not always required was shown in a series where patients received no systemic therapy until metastatic disease was documented. Overall, the median time to metastatic progression was 8 years, and 63% of the patients with rising PSA values remained free of metastases at 5 years. Factors associated with progression included the Gleason score of the radical prostatectomy specimen, time to recurrence, and PSA doubling time. For those with Gleason grade ≥8, the probability of metastatic progression was 37, 51, and 71% at 3, 5, and 7 years, respectively. If the time to recurrence was <2 years and PSA doubling time was long (>10 months), the proportion with metastatic disease at the same time intervals was 23, 32, and 53%, vs 47, 69, and 79% if the doubling time was short (<10 months). PSA doubling times are also prognostic for survival. In one series, all patients who succumbed to disease had PSA doubling times of ≤3 months. Most physicians advise treatment when PSA doubling times are ≤12 months. A difficulty with predicting the risk of metastatic spread, symptoms, or death from disease in the rising PSA state is that most patients receive some form of therapy before the development of metastases. Nevertheless, predictive models continue to be refined.

METASTATIC DISEASE: NONCASTRATE

The state of *noncastrate metastatic disease* includes men with metastases visible on an imaging study at the time of diagnosis or after local therapy(ies), and noncastrate levels of testosterone (>150 ng/dL). Symptoms of metastatic disease include pain from osseous spread, although many patients are asymptomatic despite extensive spread. Less common are symptoms related to marrow infiltration by tumor (myelophthisis), coagulopathy, or spinal cord compression.

Standard treatment is to deplete/lower androgens by medical or surgical means, the latter being the least acceptable to patients. A less frequently used treatment is to block androgen binding to the AR with antiandrogens. More than 90% of male hormones originate in the testes; <10% are synthesized in the adrenal gland **(Fig. 83-3)**. Survival benefits were shown for the combination of ADT plus docetaxel, and separately for ADT plus abiraterone plus prednisone in large scale randomized phase 3 trials.

Testosterone-Lowering Agents

Medical therapies that lower testosterone levels include the GnRH agonists/antagonists, 17,20-lyase inhibitors, CYP-17 inhibitors, and estrogens such as diethylstilbestrol (DES). The last have fallen out of favor due to the risk of vascular complications such as fluid retention, phlebitis, emboli, and stroke. GnRH agonists/antagonists (leuprolide acetate and goserelin acetate) initially produce a rise in luteinizing hormone and follicle-stimulating hormone followed by a downregulation of receptors in the pituitary gland, which effects a chemical castration. They were approved on the basis of randomized comparisons showing an improved safety profile (specifically, reduced cardiovascular toxicities) relative to DES, with equivalent potency. The initial rise in testosterone may result in a clinical flare of the disease and as such are relatively contraindicated in men with significant obstructive symptoms, cancer-related pain, or spinal cord compromise. Pure androgen antagonists such as bicalutamide can be used to prevent flare. GnRH antagonists such as degarelix achieve castrate levels of testosterone within 48 h without the initial rise in serum testosterone and may be associated with a lower risk of cardiovascular complications.

Agents that lower testosterone are associated with an androgen-deprivation syndrome that includes hot flushes, weakness, fatigue, loss of muscle mass, anemia, change in personality, and depression. Changes in lipids, obesity, and insulin resistance, along with an increased risk of diabetes and cardiovascular disease are also seen, along with a decrease in bone density that worsens over time and results in an increased risk of clinical fractures. This is a particular concern in men with preexisting osteopenia that results from hypogonadism or steroid or alcohol use, and which is significantly underappreciated. Baseline fracture risk can be assessed using the FRAX scale, and to minimize fracture risk patients are advised calcium and vitamin D supplementation, along with a bisphosphonate, RANK-ligand inhibitor (denosumab), or torimefene.

Antiandrogens

Nonsteroidal first-generation antiandrogens such bicalutamide and nilutamide block the ligand binding to the AR and were initially approved to block the flare associated with the rise in serum testosterone associated with GnRH agonist/antagonist therapy. When antiandrogens are given alone, testosterone levels increase, but relative to testosterone-lowering therapies, they cause fewer hot flushes, less of an effect on libido, less muscle wasting, fewer personality changes, and less bone loss. Gynecomastia remains a significant problem but can be prevented in part by tamoxifen or prophylactic breast irradiation.

Most reported randomized trials suggest that the cancer-specific outcomes are inferior when antiandrogens are used alone. Bicalutamide, even at a dose of 150 mg (three times the approved dose for use in combination in GnRH agonists), was associated with a shorter time to progression and inferior survival compared to surgical castration for patients with established metastatic disease.

Improving on the outcomes with ADT alone has been a focus of the field for decades. One approach was to combine a first-generation antiandrogen (flutamide, bicalutamide, or nilutamide) with a GnRH analogue or surgical orchiectomy, which has not been shown to be superior to ADT alone. As a result, use of these first-generation compounds is largely limited to the first 2–4 weeks of treatment, to protect against the flare.

More recently, significant improvements in time to progression and overall survival were reported in large-scale trials for the combination of ADT with docetaxel or with abiraterone acetate plus prednisone, relative to ADT alone. Docetaxel was the first systemic therapy shown to prolong life in metastatic castration-resistant prostate cancer (mCRPC) and was approved in 2004. Abiraterone acetate (a CYP-17 inhibitor shown to reduce androgen levels to the 1–2 ng/dl range) plus prednisone was approved for mCRPC in 2011. With docetaxel, the greatest benefit was seen for patients with “high-volume” disease defined as the presence of ≥4 lesions on radionuclide bone scan or visceral disease. For abiraterone acetate and prednisone, benefit was seen across disease states ranging from high-risk localized to metastatic disease.

Intermittent Androgen Deprivation Therapy (IADT)

One way to reduce the side effects of androgen depletion is to administer antiandrogens on an intermittent basis. This was proposed as a way to prevent the selection of cells that are resistant to androgen depletion. The hypothesis is that by allowing endogenous testosterone levels to rise, the cells that survive androgen depletion will induce a normal differentiation pathway. In this way, the surviving cells that are allowed to proliferate in the presence of androgen will retain sensitivity to subsequent androgen depletion. Applied in the clinic, androgen depletion is continued for 2–6 months beyond the point of maximal response. Once treatment is stopped, endogenous testosterone levels increase, and the symptoms associated with hormone treatment abate. PSA levels also begin to rise, and at some level treatment is restarted. With this approach, multiple cycles of regression and proliferation have been documented in individual patients. Unknown is whether the intermittent approach increases, decreases, or does not change the overall duration of sensitivity to androgen depletion. The approach is safe, but long-term data are needed to assess the course in men with low PSA levels. A trial to address this question is ongoing.

Outcomes of Androgen Deprivation

The anti–prostate cancer effects of the various androgen depletion strategies are similar, and the clinical course is predictable: an initial response, then a period of stability in which tumor cells are dormant and nonproliferative, followed after a variable period of time by a rise in PSA and regrowth that is visible on a scan as a castration-resistant lesion. Androgen depletion is not curative because cells that survive

castration are present when the disease is first diagnosed. Considered by disease manifestation, PSA levels return to normal in 60–70% of patients, and measurable disease regression occurs in 50%; improvements in bone scan occur in 25% of cases, but the majority remain stable. Duration of survival is inversely proportional to disease extent at the time androgen depletion is first started and the nadir level of PSA at 6 months. Patients with nadir values above a certain threshold have markedly inferior survival times and should be considered for alternative approaches.

An unresolved question remains on how early systemic therapies should be offered to patients: in the adjuvant setting after surgery or radiation treatment of the primary tumor; at the time that a PSA recurrence is documented; or wait until metastatic disease or symptoms of disease are manifest? Trials in support of early therapy have been largely underpowered relative to the reported benefit or have been criticized on methodologic grounds. One trial which showed a survival benefit for patients treated with radiation therapy and 3 years of ADT, relative to radiation alone, was criticized for the poor outcomes of the control group. Another trial showing a survival benefit for patients with positive lymph nodes who were randomized to immediate medical or surgical castration compared to observation ($p = .02$) was criticized because the confidence intervals around the 5- and 8-year survival distributions for the two groups overlapped.

METASTATIC DISEASE: CASTRATE

Castration-resistant prostate cancer (CRPC), disease that progresses while the measured levels of testosterone in the blood are 50 ng/mL or lower, can produce some of the most feared complications of the disease and is lethal for most men. The most common manifestation is a rising PSA, frequently co-occurring with progression in bone. Nodal and/or visceral spread is less frequent and symptoms may or may not be present. The bone- and PSA-dominant pattern limits the ability to assess treatment effects reliably because traditional bone imaging is inaccurate and no PSA-based outcome has been shown to be a true surrogate for survival. It is essential to define therapeutic objectives based on the manifestations of the disease in the individual. As such, for the patient with symptomatic bone disease, relief of pain can be more clinically relevant than lowering the PSA. Naturally, for all patients the central focus is delaying or preventing disease progression, symptom development, and death from cancer.

Through 2010, docetaxel was the only FDA-approved life-prolonging therapy for CRPC. Since then, our understanding of the biology of the disease has increased significantly, which in turn has led to improved therapies. In particular, it is now recognized that the majority of mCRPCs continue to express the AR, which in upwards of 50% of cases harbors a series of oncogenic changes including overexpression of the receptor itself and the enzymes in the androgen biosynthesis pathways. These oncogenic changes have been successfully targeted with the next-generation antiandrogen enzalutamide and the CYP-17 inhibitor abiraterone acetate (given in combination with prednisone), both of which have been proven to prolong life and are FDA-approved for use in CRPC in both the pre- and post-chemotherapy setting. More recently, the results of large-scale molecular profiling efforts have led to biologically based pathway-focused classification that showed a markedly higher than expected frequency of germline and somatic BRCA2 alterations, along with other genes in the DNA damage repair pathway, that has been successfully treated with poly ADP ribose polymerase (PARP) inhibitors. Other classes of therapy that have been approved based on a demonstrated survival benefit include the biologic agent sipuleucel-T, the second-generation taxane cabazitaxel, and the alpha-emitting bone targeting radiopharmaceutical radium-223. An intense of focus of current CRPC research is to understand the optimal sequence in which to utilize these agents to maximize benefit for the individual patient.

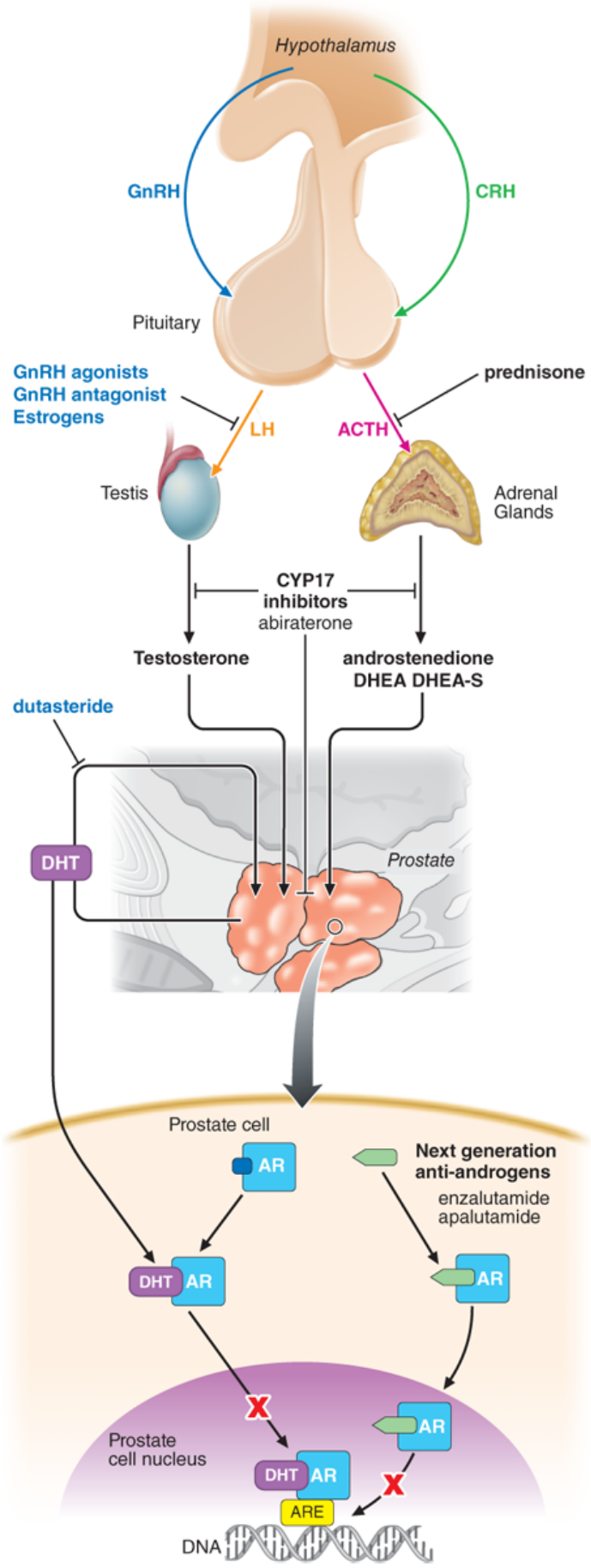
Pain Management

Pain secondary to osseous metastases is one of the most feared complications of the disease and a major cause of morbidity, worsened by the narcotics needed to control symptoms. Management requires accurate diagnoses because non-cancer etiologies including degenerative disease, spinal stenosis, and vertebral collapse secondary to bone loss are common. Neurologic symptoms, including those suggestive of base of skull disease or spinal cord compromise, require emergency evaluation because loss of function may be permanent if not addressed quickly. Neurologic symptoms or loss of function are best treated with external beam radiation, as are single sites of pain. Diffuse symptoms in the absence of neurologic deficits can be treated with bone-seeking radioisotopes such as radium-223 or the beta emitter ^{153}Sm-EDTMP, or mitoxantrone, or other systemic therapies such as abiraterone acetate, enzalutamide, and docetaxel. Radium-223 is indicated for patients with symptoms while ^{153}Sm-EDTMP and mitoxantrone are approved for the palliation of pain but not shown to prolong life. Abiraterone, enzalutamide, and docetaxel do not have a formal indication for pain, but were shown to palliate pain in the registration trials that led to their approval by showing a survival benefit.

Other bone-targeting agents, including bisphosphonates such as zoledronic acid and the RANK-ligand inhibitor denosumab, have been shown to reduce the frequency and development of skeletal complications including pain requiring analgesia, neurologic compromise from epidural extension of tumor, and/or the need for surgery or radiation therapy to treat symptomatic osseous disease. It is important to note that for all of these agents, the direct effect on the tumor is modest and benefits are seen without declines in PSA or improvements on imaging.

FIGURE 83-3

Sites of action of different hormone therapies.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

BENIGN DISEASE

BENIGN PROSTATIC HYPERPLASIA

BPH is a pathologic process that contributes to the development of lower urinary tract symptoms (LUTS) in men. LUTS, arising from lower urinary tract dysfunction, are further subdivided into obstructive symptoms (urinary hesitancy, straining, weak stream, terminal dribbling, prolonged voiding, incomplete emptying) and irritative symptoms (urinary frequency, urgency, nocturia, urge incontinence, small voided volumes). LUTS and other sequelae of BPH are not just due to a mass effect, but also likely due to a combination of the prostatic enlargement and age-related detrusor dysfunction.

Diagnostic Procedures and Treatment

LUTS symptoms are generally measured using a validated, reproducible index that is designed to determine disease severity and response to therapy—the American Urological Association's Symptom Index (AUASI), also adopted as the International Prostate Symptom Score (IPSS) **(Table 83-2)**. Serial AUASI is particularly useful in following patients as they are treated with various forms of therapy. Asymptomatic patients do not require treatment regardless of the size of the gland, while those with an inability to urinate, gross hematuria, recurrent infection, or bladder stones may require surgery. In patients with symptoms, uroflowmetry can identify those with normal flow rates who are unlikely to benefit from treatment, and bladder ultrasound can identify those with high postvoid residuals who may need intervention. Pressure-flow (urodynamic) studies detect primary bladder dysfunction. Cystoscopy is recommended if hematuria is documented and to assess the urinary outflow tract before surgery. Imaging of the upper tracts is advised for patients with hematuria, a history of calculi, or prior urinary tract problems.

TABLE 83-2

AUA Symptom Index

	AUA Symptom Score (Circle 1 Number on Each Line)					
Questions to Be Answered	**Not at All**	**Less Than 1 Time in 5**	**Less Than Half the Time**	**About Half the Time**	**More Than Half the Time**	**Almost Always**
Over the past month, how often have you had a sensation of not emptying your bladder completely after you finished urinating?	0+	1	2	3	4	5
Over the past month, how often have you had to urinate again less than 2 h after you finished urinating?	0	1	2	3	4	5
Over the past month, how often have you found you stopped and started again several times when you urinated?	0	1	2	3	4	5
Over the past month, how often have you found it difficult to postpone urination?	0	1	2	3	4	5
Over the past month, how often have you had a weak urinary stream?	0	1	2	3	4	5
Over the past month, how often have you had to push or strain to begin urination?	0	1	2	3	4	5
Over the past month, how many times did you most typically get up to urinate from the time you went to bed at night until the time you got up in the morning?	(None)	(1 time)	(2 times)	(3 times)	(4 times)	(5 times)
Sum of 7 circled numbers (AUA Symptom Score): ____						

Abbreviation: AUA, American Urological Association.

Source: MJ Barry et al: J Urol 148:1549, 1992. Used with permission.

Symptomatic relief is the most common reason men seek treatment for BPH, and therefore symptomatic relief is usually the goal of therapy for BPH. Alpha-adrenergic receptor antagonists are thought to treat the dynamic aspect of BPH by reducing sympathetic tone of the bladder outlet, thereby decreasing resistance and improving urinary flow. 5ARIs are thought to treat the static aspect of BPH by reducing prostate volume and having a similar, albeit delayed effect. They have also proven to be beneficial in the prevention of BPH progression, as measured by prostate volume, the risk of developing acute urinary retention, and the risk of having BPH-related surgery. The use of an alpha-adrenergic receptor antagonist and a 5ARI as combination therapy seeks to provide symptomatic relief while preventing progression of BPH.

Another class of medications that has shown improvement in LUTS secondary to BPH is phosphodiesterase-5 (PDE5) inhibitors, used currently in the treatment of erectile dysfunction. All four of the PDE5 inhibitors available in the United States, sildenafil, vardenafil, tadalafil, and avanafil, appear to be effective in the treatment of LUTS secondary to BPH. The use of PDE5 inhibitors is not without controversy, however, given the fact that short-active phosphodiesterase inhibitors such as sildenafil need to be dosed separately from alpha blockers such as tamsulosin because of potential hypotensive effects.

Symptoms due to BPH often coexist with symptoms due to overactive bladder, and the most common pharmacologic agents for the treatment of overactive bladder symptoms are anticholinergics. This has led to multiple studies evaluating the efficacy of anticholinergics for the treatment of LUTS secondary to BPH.

Surgical therapy is now considered second-line therapy and is usually reserved for patients after a trial of medical therapy. The goal of surgical therapy is to reduce the size of the prostate, effectively reducing resistance to urine flow. Surgical approaches include TURP, transurethral incision, or removal of the gland via a retropubic, suprapubic, or perineal approach. Also used are transurethral ultrasound-guided laser-induced prostatectomy (TULIP), stents, and hyperthermia.

FURTHER READING

Barry MJ, Simmons LH: Prevention of prostate cancer morbidity and mortality: Primary prevention and early detection. Med Clin North Am 101:787, 2017. [PubMed: 28577627]

Buyyounouski MK et al: Prostate cancer—major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin 67:245, 2017. [PubMed: 28222223]

Fizazi K et al: Abiraterone plus prednisone in metastatic, castration-sensitive prostate cancer. N Engl J Med 377:352, 2017. [PubMed: 28578607]

Gartrell BA et al: Metastatic prostate cancer and the bone: Significance and therapeutic options. Eur Urol 68:850, 2015. [PubMed: 26153564]

Higano CS: Intermittent versus continuous androgen deprivation therapy. J Natl Compr Canc Netw 12:727, 2014. [PubMed: 24812139]

James ND et al: Abiraterone for prostate cancer not previously treated with hormone therapy. N Engl J Med 377:338, 2017.

McGinley KF et al: Prostate cancer in men of African origin. Nat Rev Urol 13:99, 2016. [PubMed: 26718455]

Pritchard CC et al: Inherited DNA-repair gene mutations in men with metastatic prostate cancer. N Engl J Med 375:443, 2016. [PubMed: 27433846]

Robinson D et al: Integrated clinical genomics of advanced prostate cancer. Cell 161:1215, 2015. [PubMed: 26000489]

Sweeney CJ et al: Chemohormonal therapy in metastatic hormone-sensitive prostate cancer. N Engl J Med 373:737, 2015. [PubMed: 26244877]

Chapter 84: Testicular Cancer

David J. Vaughn

FIGURE 84-1

INTRODUCTION

Testicular germ cell tumors (GCTs) represent 95% of all testicular neoplasms. Non-germ cell tumors of the testis are much less common. Approximately 5% of GCTs arise in extragonadal locations including the mediastinum, retroperitoneum, and pineal gland. Treatment for testicular GCTs is determined by pathology and stage. The development of effective chemotherapy for this disease represents a landmark achievement in oncology. About 95% of newly diagnosed patients with testicular GCTs will be cured. For this reason, testicular cancer has been called “a model for a curable neoplasm.”

INCIDENCE

In 2016, ~8700 cases of testicular GCTs will be diagnosed in the United States, with <400 deaths. These tumors are diagnosed most commonly in men between 20 and 40 years. It has recently been reported that the incidence of GCTs is increasing in men 50 years and older.

GLOBAL CONSIDERATIONS

The incidence of testicular GCTs appears to be increasing worldwide. The disease has the highest incidence in Scandinavia, Western Europe, and Australia/New Zealand. Africa and Asia have the lowest incidence. The incidence in the United States and the United Kingdom is intermediate. While there does not appear to be a distinct biology related to geography, several countries have reported a migration to earlier stage disease in part related to public awareness and earlier diagnosis.

EPIDEMIOLOGY

GCTs are predominantly seen in young Caucasian men. The disease is much less commonly seen in African Americans. Although most patients with GCTs do not have a family history of this disease, there are rare familial cases. Interestingly, the risk of GCT is higher in male siblings and cousins than in offspring of the patient. Although epidemiological studies have been performed attempting to identify a relationship with environmental exposures, no conclusive causal links have been established.

Risk Factors

The strongest risk factors for testicular GCT include a prior history of the disease, cryptorchidism, and a history of testicular intratubular germ cell neoplasia (ITGCN). Patients with a prior history of testicular GCT have a 2% risk of developing a contralateral GCT. These are more commonly metachronous than synchronous. Men with cryptorchidism have approximately a four- to sixfold increased risk of developing testicular GCT. Orchidopexy before puberty decreases but does not eliminate this risk. Interestingly, the contralateral descended testis is also at risk for this disease. Men undergoing infertility evaluation in which a testicular biopsy demonstrates ITGCN have a 50% risk of developing GCT. Although scrotal ultrasound of patients with testicular GCT may demonstrate testicular microcalcifications that may be related to ITGCN, the significance of testicular microcalcifications in the general population is unclear.

BIOLOGY

The primordial germ cell is the cell of origin for GCTs. All malignant GCTs arise from ITGCN. The molecular events that result in the development of ITGCN and subsequent malignant GCT have not been fully determined. However, genetic analysis of GCTs have demonstrated an excess copy number of isochromosome 12p (i[12p]) in most cases. Several genome-wide association studies have identified independent loci associated with testicular GCT risk. The strongest of these is the *KITLG* (KIT ligand) locus on chromosome 12.

PATHOLOGY

GCTs are either seminomas or nonseminomas. For a tumor to be considered a seminoma, it must be 100% seminoma. Any mixed GCT is best approached as a nonseminomatous GCT (NSGCT). Seminomas represent ~50% of cases. Seminomas arise most commonly in patients in the fourth decade of life. Seminomas may contain syncytiotrophoblastic cells which may secrete β human chorionic gonadotropin (HCG). Seminomas do not secrete α fetoprotein (AFP). Seminomas are exquisitely sensitive to both chemotherapy and radiation therapy. Seminomas are believed to be a common precursor that subsequently differentiates into the NSGCT subtypes. NSGCTs are most commonly diagnosed in the third decade of life. The histologic subtypes include embryonal carcinoma, yolk sac tumor, choriocarcinoma, and teratoma. Embryonal carcinoma is the most undifferentiated NSGCT subtype with the potential to differentiate into the other subtypes. Embryonal carcinoma may secrete AFP, HCG, both, or neither. Yolk sac tumor (also referred to as endodermal sinus tumor) often secretes AFP. Choriocarcinoma is an aggressive subtype, often secreting HCG at very high levels. These NSGCT subtypes are all considered chemotherapy sensitive. Teratoma is composed of somatic cell types that are derived from two or more germinal layers (endoderm, mesoderm, and ectoderm). Teratomas are classified as mature, in which cell types resemble normal adult somatic tissue; immature, in which cell types resemble fetal somatic tissue; and malignant, in which the cell types have undergone malignant transformation into the malignant counterpart of the somatic tissue. Teratomas are chemotherapy resistant and must be approached surgically.

INITIAL PRESENTATION

Signs and Symptoms

Although a painless testicular mass is pathognomonic of a GCT, most patients present with testicular swelling, firmness, discomfort, or a combination of these. The differential diagnosis may include epididymitis or orchitis and a trial of antibacterials may be considered. Patients with retroperitoneal metastases may complain of back or flank pain. Patients may have cough, shortness of breath, or hemoptysis as a result of lung metastases. In patients with elevation of serum HCG, gynecomastia may be present. Diagnostic delay is not uncommon, and may be associated with a more advanced stage at diagnosis.

Physical Examination

Careful examination of the affected testis and the contralateral normal testis should be performed. Many tumors will have a hard consistency to palpation. Some patients may show testicular atrophy. Evaluation for supraclavicular lymphadenopathy, gynecomastia, and abdominal mass should be performed. Inguinal lymphadenopathy is rare. Most patients with lung metastases will have normal auscultation of the lungs.

Diagnostic Testing

If a firm testicular mass is identified, a scrotal ultrasound should be performed. Patients with suspected epididymitis or orchitis who do not respond to antibiotics should also undergo scrotal ultrasound. Scrotal ultrasound should include both testicles. On ultrasound, a testicular GCT is hypoechoic and may be multifocal. A solid mass identified on ultrasound should be considered malignant until otherwise proven. Transscrotal aspiration or biopsy of a testicular mass should never be performed. Such scrotal violation may result in tumor seeding of the scrotum or inguinal lymph nodes.

Serum Tumor Markers

Serum AFP, HCG, and lactate dehydrogenase (LDH) should be measured in patients suspected of testicular GCT. AFP is elevated in ~60–70% of patients who present with NSGCTs. Seminomas never secrete AFP. A patient with a seminoma with elevation of AFP should be approached as having a NSGCT. The half-life of AFP is 5–7 days. A falsely elevated AFP may be seen in patients with hepatic disease or a condition called hereditary persistence of AFP in which patients may have baseline AFP levels that are mildly elevated. HCG may be elevated in both NSGCTs as well as seminomas. Patients with choriocarcinoma may have markedly elevated levels of HCG. The half-life for HCG is 24–36 h. False-positive elevation of HCG may be seen secondary to hypogonadism, marijuana use, or as a result of interfering substances measured by the assay. LDH is a nonspecific marker for GCT. Its principal use is to help in the assessment of the risk classification of a patient with metastatic disease. Although elevation of serum tumor markers support the diagnosis of a testicular GCT, it should be remembered that most patients with a seminoma and up to a third of patients with NSGCTs do not have elevated levels.

INITIAL MANAGEMENT

Inguinal Orchiectomy

Prompt referral to urology should be performed if a testicular GCT is suspected. The initial treatment for most patients suspected of having a testicular GCT is radical inguinal orchiectomy with removal of the testicle and spermatic cord to the level of the internal inguinal ring. In patients who present with metastatic disease and the diagnosis of GCT is certain, orchiectomy may be deferred until completion of chemotherapy. Pathologic examination of the entire testicle is important, since testicular GCTs may be multifocal. Given the rarity of this cancer, review by an experienced pathologist is essential for accurate tumor classification. Serum tumor markers should be obtained before and after orchiectomy.

Staging

The staging of testicular GCT is based upon an understanding of the pattern of spread. The initial spread is by the lymphatic route to the retroperitoneal lymph nodes. A left-sided testicular GCT spreads first to the primary landing zone of left paraaortic lymph nodes inferior to the left renal vessels. A right-sided testicular GCT spreads first to the primary landing zone of the aortocaval nodes inferior to the right renal vessels. Nodal metastases may extend into the iliac regions. If scrotal violation occurred, inguinal lymph node metastases may be seen. Subsequent lymphatic spread is to the retrocrural, mediastinal, and supraclavicular lymph nodes. Hematogenous spread to the lung is the next most common site of metastasis. Metastases to the liver, bone, and brain are less commonly seen. Patients with newly diagnosed testicular GCTs should undergo computed tomography (CT) scan of the abdomen and pelvis. Chest x-ray should be performed. CT scan of the chest is performed if retroperitoneal metastases are present or if lung nodules are identified on chest x-ray. Bone scan and magnetic resonance imaging (MRI) of the brain are not routinely performed unless clinically indicated. Positron emission tomography (PET) has little role in the initial staging of testicular GCTs.

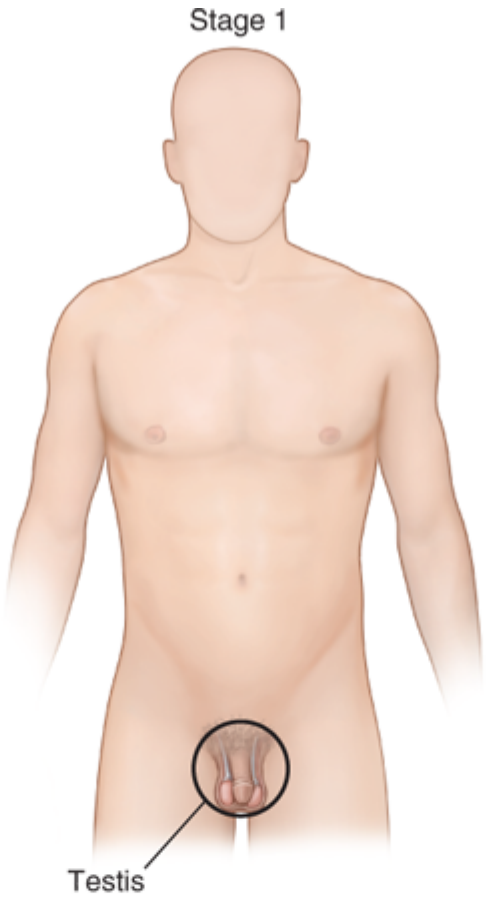
The American Joint Committee on Cancer tumor/node/metastasis (TNM) staging classification is used. There are three main stages of testicular GCT. Stage I is limited to the testis; stage II involves the retroperitoneal lymph nodes; stage III includes lymph node involvement beyond the retroperitoneum and/or distant metastatic disease.

STAGE-BASED MANAGEMENT

Treatment of testicular GCT is based upon two factors: (1) whether the tumor is seminoma or NSGCT, and (2) the stage of the patient. This is summarized in Fig. 84-1.

FIGURE 84-1

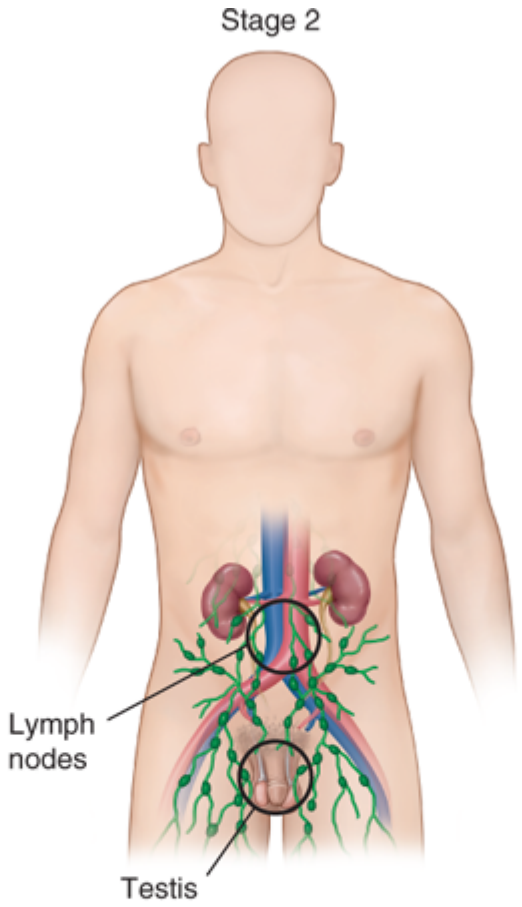
Stage-based management of testicular GCT.



	Seminoma	NSGCT
Stage IA pT1: Testis only, no vascular/lymphatic invasion	Active surveillance; or, Adjuvant carboplatin x 1 cycle; or, Adjuvant para-aortic RT	Active surveillance; or, Nerve sparing RPLND
Stage IB pT2: Testis only, with vascular/lymphatic invasion or extension through tunica albuginea into tunica vaginalis	Active surveillance; or, Adjuvant carboplatin x 1 cycle; or, Adjuvant para-aortic RT	Adjuvant BEP x 1 cycle; or, Active surveillance; or, Nerve sparing RPLND
Stage IS Elevated serum tumor markers post-orchiectomy	BEP x 3 cycles; or, EP x 4 cycles	BEP x 3 cycles; or, EP x 4 cycles

A

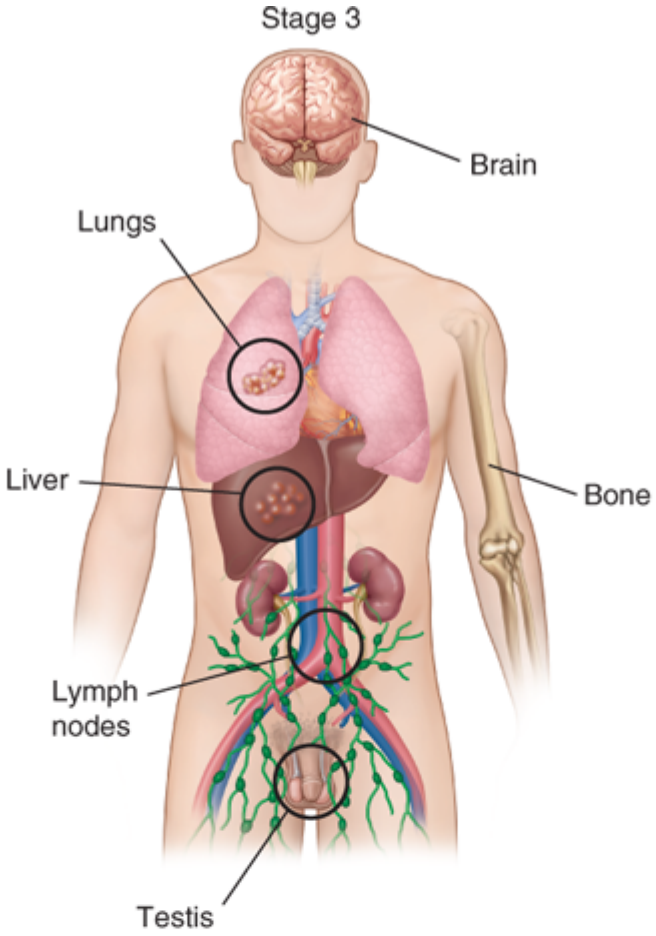
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



	Seminoma	NSGCT
Stage IIA N1: nodes ≤ 2 cm	Para-aortic and ipsilateral iliac RT; or, BEP x 3 cycles or EP x 4 cycles	Nerve-sparing RPLND; or, BEP x 3 cycles or EP x 4 cycles
Stage IIB N2: nodes > 2 to 5 cm	BEP x 3 cycles or EP x 4 cycles; or, Para-aortic and ipsilateral iliac RT	BEP x 3 cycles or EP x 4 cycles +/– post-chemotherapy RPLND
Stage IIC N3: nodes > 5 cm	BEP x 3 cycles or EP x 4 cycles	BEP x 3 cycles or EP x 4 cycles +/– post-chemotherapy RPLND

B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



	Seminoma	NSGCT
Stage IIIA (good-risk)	BEP x 3 cycles; or, EP x 4 cycles	BEP x 3 cycles; or, EP x 4 cycles; +/– Post-chemotherapy surgery
Stage IIIB (intermediate-risk)	BEP x 4 cycles; or, VIP x 4 cycles	BEP x 4 cycles; or, VIP x 4 cycles +/– Post-chemotherapy surgery
Stage IIIC (poor-risk)	N/A	BEP x 4 cycles; or, VIP x 4 cycles +/– Post-chemotherapy surgery

Abbreviations: BEP, bleomycin, etoposide, cisplatin; EP, etoposide, cisplatin; NSGCT, non-seminomatous germ cell tumor; RPLND, retroperitoneal lymph node dissection; RT, radiation therapy; VIP, etoposide, ifosfamide, cisplatin.

C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Stage I

SEMINOMA

About 70% of newly diagnosed patients with seminoma present with stage I disease. This is defined as no evidence of metastatic disease on imaging of the chest, abdomen, and pelvis. If pre-orchiectomy serum HCG

is elevated, this must normalize post-orchiectomy to be considered stage I. Approximately 15% of patients with stage I seminoma have metastatic disease at the microscopic level, usually in the retroperitoneum. Historically, patients with stage I seminoma were treated with a course of adjuvant radiation therapy to the paraaortic lymph nodes. While still an option, this is not usually performed because of concerns for late radiation-induced secondary malignancies. Active surveillance is the most common approach elected by these patients following orchiectomy. With active surveillance, interval physical examination and CT scan of the abdomen are performed. For the 15% of patients that develop metastatic disease during active surveillance, treatment with definitive radiation therapy or chemotherapy is curative in nearly all. A third option for clinical stage I seminoma is adjuvant chemotherapy with carboplatin monotherapy for one cycle. While effective in decreasing the risk of recurrence, it should be remembered that most patients are cured by orchiectomy alone, and therefore the additional treatment is unnecessary. In addition, long-term data on toxicity and efficacy are not available.

NSGCTs

About 40% of newly diagnosed patients with NSGCTs present with stage I disease. Because NSGCTs have an increased potential for invasion and metastasis, spread to the retroperitoneum and beyond is more common than with seminoma. If pre-orchiectomy serum tumor markers are elevated, these must normalize post-orchiectomy to be considered stage I. Patients with persistently elevated or rising serum tumor markers after orchiectomy have stage IS disease and should be treated with cisplatin-based chemotherapy. If the tumor is pT1, defined as limited to testis and epididymis with no vascular or lymphatic invasion and no invasion into tunica vaginalis, the risk of recurrence is approximately 20%. However, if the tumor is pT2, defined as limited to testis and epididymis with vascular or lymphatic invasion, or tumor extension into tunica vaginalis, the risk of recurrence is ~50%. Historically, a prophylactic retroperitoneal lymph node dissection (RPLND) was performed. This surgery is not only diagnostic, but also therapeutic. In fact, most patients who undergo prophylactic RPLND will never require chemotherapy. While still an option, this approach subjects many patients to unnecessary major abdominal surgery. RPLND is also associated with a small risk of retrograde ejaculation due to nerve injury, and nerve sparing techniques have been developed. Active surveillance is frequently performed especially for patients with pT1 disease. Most patients who relapse will be treated with cisplatin-based chemotherapy and achieve cure rates approaching 100%. Active surveillance can also be employed for patients with pT2 disease, although the risk of progression is significantly higher. For this reason, some advocate adjuvant cisplatin-based chemotherapy such as BEP for one cycle for patients with pT2 disease. Other centers favor a prophylactic RPLND. Almost all patients who present with stage I NSGCTs will achieve cure.

Stage II

seminoma

Approximately 15–20% of newly diagnosed patients with seminoma present with stage II disease. Patients are subgrouped into IIA, IIB, or IIC based upon the size of the retroperitoneal nodes (2 cm or less, more than 2 to 5 cm, or >5 cm, respectively). Patients with stage IIA disease are usually treated with “dogleg” radiation

therapy which includes the paraaortic and ipsilateral iliac nodes. Cisplatin-based chemotherapy may also be considered. Stage IIB disease is treated with cisplatin-based chemotherapy or, in select patients, radiation therapy. Most patients treated with radiation therapy that relapse will subsequently be cured with cisplatin-based chemotherapy. For patients with stage IIC disease, cisplatin-based chemotherapy should be used.

NSGCTs

Approximately 15% of newly diagnosed patients with NSGCTs present with clinical stage II disease. Patients with stage IIA disease may be treated with primary RPLND. Alternatively, these patients may be treated with cisplatin-based chemotherapy. Patients with stage IIB and IIC disease are best initially managed with cisplatin-based chemotherapy.

Stage III

Patients who present with stage III GCT (seminoma or NSGCT) are treated with cisplatin-based chemotherapy. These patients are classified into good-, intermediate-, or poor-risk categories using the International Germ Cell Consensus Classification system, which is based upon clinical factors including histology, site of primary, the presence of nonpulmonary visceral metastatic disease, and the level of post-orchiectomy serum tumor markers **(Table 84-1)**. Most patients with stage III GCT present with good-risk disease; >90% will be cured. The remainder present with intermediate-risk or poor-risk disease, associated with 5-year survival rates of ~80% and 50%, respectively. Select patients with rapidly progressive metastatic disease and life-threatening symptoms such as hemoptysis in whom there is a high clinical suspicion of GCT should emergently initiate cisplatin-based chemotherapy, even without a tissue diagnosis.

TABLE 84-1

International Germ Cell Consensus Classification System

Risk Group	Seminoma	NSGCT
Good	Any primary site; and normal AFP, any HCG, any LDH; and nonpulmonary visceral metastases absent	Gonadal or retroperitoneal primary; and nonpulmonary visceral metastases absent; and AFP <1000 ng/mL; and HCG <5000 mIU/mL; and LDH <1.5 × ULN
Intermediate	Any primary site; and normal AFP, any HCG, any LDH; and nonpulmonary visceral metastases present	Gonadal or retroperitoneal primary; and nonpulmonary visceral metastases absent; and one of the following: AFP 1000–10,000 ng/mL HCG 5000–50,000 mIU/mL LDH 1.5–10 × ULN
Poor	N/A	Mediastinal primary; or nonpulmonary visceral metastases present; or one of the following: AFP >10,000 ng/mL HCG >50,000 mIU/mL LDH >10 × ULN

Abbreviations: AFP, α fetoprotein; HCG, human chorionic gonadotropin; LDH, lactate dehydrogenase; N/A, not applicable; NSGCT, nonseminomatous germ cell tumor; ULN, upper limit normal. Nonpulmonary visceral metastases include liver, bone, and brain.

Source: International Germ Cell Cancer Collaborative Group: International Germ-Cell Consensus Classification: A prognostic factor based staging system for metastatic germ cell tumors. J Clin Oncol 15:594, 1997.

Chemotherapy

The development of cisplatin-based chemotherapy represents an important advance in cancer medicine. Through a series of carefully performed clinical trials with the aim of maximizing cure while minimizing the extent of treatment, the chemotherapy approach to the treatment of these patients has been standardized. Patients with good-risk metastatic GCT are treated with either three cycles of bleomycin, etoposide, cisplatin (BEP) or four cycles of etoposide, cisplatin (EP). Patients with intermediate- and poor-risk metastatic disease

are treated with either four cycles of BEP or four cycles of etoposide, ifosfamide, cisplatin (VIP). Maintaining dose and schedule is important, as dose modifications and delays have been associated with inferior outcomes. Serum tumor markers should be monitored throughout treatment and should normalize during or after treatment. Cisplatin-based chemotherapy is associated with myelosuppression, nausea and vomiting, and alopecia. Cisplatin may result in nephrotoxicity, ototoxicity, and peripheral neuropathy. Bleomycin may result in pulmonary toxicity and risk factors for this include age greater than 40, renal failure, tobacco use, and the cumulative dose of bleomycin received. For patients at increased risk of bleomycin-induced pneumonitis, non-bleomycin-containing regimens as noted above may be given. Cisplatin-based chemotherapy is also associated with sterility. Approximately 30% of newly diagnosed testicular GCT patients have severe oligospermia or azospermia. For the remainder with normal baseline spermatogenesis who receive cisplatin-based chemotherapy, all will be azospermic at the completion of therapy. Approximately 80% of these patients will recover spermatogenesis over a period of several years. For this reason, pre-chemotherapy sperm banking should be offered to all patients treated with chemotherapy.

Post-Chemotherapy Surgery

Upon completion of cisplatin-based chemotherapy, many patients with normalized serum tumor markers will have radiographic evidence of residual masses. In approximately half of patients with NSGCT, the residual mass is composed of necrosis and/or fibrosis. About 40% will have residual teratoma and only 10% will have residual viable non-teratomatous GCT. Unfortunately, radiographic imaging cannot accurately differentiate between these entities. For this reason NSGCT patients with residual masses after chemotherapy undergo resection of all sites of disease. This most commonly includes a post-chemotherapy RPLND. However, thoracotomy and neck dissection are required in some patients. If the patients are found to have residual necrosis or teratoma, no additional therapy is required. However, for patients with residual viable non-teratomatous GCT, two additional cycles of chemotherapy are frequently administered. It should be noted that in most centers patients with minimal residual tumors defined as retroperitoneal lymph nodes of <10 mm in short axis will forego post-chemotherapy RPLND. Patients who experience normalization of serum tumor markers with first-line chemotherapy but have enlarging tumors, most often cystic masses in the retroperitoneum, may have “growing teratoma syndrome.” These patients are best approached with surgery.

For patients with metastatic seminoma, ~20% of residual masses harbor viable tumor; the remainder have only necrosis. Patients with residual masses 3 cm or less may be observed without surgery. For patients with residual masses >3 cm, FDG-PET may be used to distinguish necrosis from viable seminoma and identify patients who should be considered for post-chemotherapy surgery.

RELAPSED DISEASE

Approximately 20–30% of patients with metastatic GCTs treated with cisplatin-based chemotherapy will not achieve durable disease control. Most of these patients will experience disease progression within 2 years following completion of chemotherapy. The International Prognostic Factors Study Group developed a risk

stratification classification system for patients in first relapse. Contributors to a worsened prognosis include NSGCT histology, extragonadal primary, incomplete response to first-line chemotherapy, time to relapse of 3 months or less, level of serum tumor markers at relapse, and the presence of nonpulmonary visceral metastatic disease.

Patients in first relapse may be treated with either conventional-dose salvage chemotherapy or high-dose salvage chemotherapy with autologous stem cell rescue. There is controversy concerning which approach is optimal. Some institutions advocate for risk stratification, with more favorable prognosis patients receiving conventional-dose chemotherapy and worse prognosis patients receiving high-dose chemotherapy. The most commonly utilized conventional-dose regimen includes paclitaxel, ifosfamide, and cisplatin (TIP). In one study of TIP in patients with more favorable risk disease, approximately two-thirds experienced 2-year progression-free survival. High-dose chemotherapy consists of initial salvage therapy followed by stem cell harvest and then two or three cycles of high-dose carboplatin and etoposide (CE) with stem cell rescue. The largest series of patients treated with high-dose chemotherapy was reported by researchers at Indiana University where this approach is considered standard for most patients in first relapse regardless of risk classification. In their study, ~70% of patients in first relapse achieved durable progression-free survival. A large retrospective analysis has compared conventional-dose salvage chemotherapy to high-dose salvage chemotherapy in patients in first relapse. This study reports a more favorable outcome with high-dose salvage chemotherapy across nearly all risk groups. However given the retrospective nature of this study and the controversy concerning optimal approaches, an international randomized trial comparing conventional dose chemotherapy (TIP) to high-dose chemotherapy with autologous stem cell rescue (TI-CE) has been initiated.

Some patients who experience disease progression after conventional-dose salvage chemotherapy may successfully be treated with high-dose salvage chemotherapy with autologous stem cell rescue. Patients with disease progression after high-dose salvage chemotherapy may be treated with subsequent chemotherapy regimens that include gemcitabine/oxaliplatin, gemcitabine/paclitaxel, epirubicin/cisplatin, and oral etoposide. While these patients may benefit from third-line chemotherapy, few will achieve durable disease control. Select patients with relapsed but resectable disease may be candidates for salvage or so-called “desperation” surgery.

Patients who experience disease progression >2 years after chemotherapy are considered to have “late relapse.” Late relapse appears to have a different biology than early relapse. These patients tend to have more chemotherapy-resistant disease. Patients with late relapse usually have NSGCT with elevation of serum AFP. Many of these patients recur in the retroperitoneum many years after first-line chemotherapy, and this likely represents residual retroperitoneal disease that was not controlled after first-line therapy. These patients are best approached with salvage surgery.

EXTRAGONADAL GCTs

Approximately 5% of patients who present with GCTs have extragonadal primaries. These mainly originate in the mediastinum or retroperitoneum. Patients suspected of extragonadal GCT should undergo scrotal ultrasound to exclude a gonadal primary. Extragonadal seminomas have a similar excellent prognosis as their gonadal counterparts and are approached the same. Mediastinal NSGCTs are classified as poor-risk and are treated with either four cycles of BEP or four cycles of VIP. These patients frequently require post-chemotherapy thoracic surgery for residual disease. For this reason, some advocate avoiding bleomycin in this patient population. Klinefelter's syndrome is associated with an increased risk of mediastinal NSGCTs. Rarely, mediastinal NSGCTs are associated with hematologic disorders including acute myelogenous leukemia. NSGCTs arising in the retroperitoneum do not have a worse prognosis than their gonadal counterparts. Many patients who present with extragonadal GCTs will undergo core needle biopsy for diagnosis. However, select patients with extragonadal tumors and definitive elevation of serum tumor markers may initiate chemotherapy without a tissue diagnosis.

Cancers of unknown primary are defined as histologically proven metastatic malignancy in which the primary site is not obvious. A subgroup of patients with cancer of unknown primary have occult GCTs. Male gender, age <65 years, midline tumors, and nonsmoking status increase the likelihood of this presentation. Pathology may demonstrate a poorly differentiated malignant neoplasm. Immunohistochemical staining is used to exclude lymphoma. Tumor may be analyzed by FISH for i(12p) which confirms the diagnosis. Even if the diagnosis is not certain, patients should be treated with cisplatin-based chemotherapy, which will cure up to 20% of this patient group.

TESTICULAR NON-GERM CELL TUMORS

Rarely, patients may develop testicular non-GCTs. These include lymphoma, most commonly occurring in men over the age of 50; sex cord stromal tumors including Leydig cell tumors and Sertoli cell tumors; mesothelioma of the tunica vaginalis; and, paratesticular sarcoma. Metastasis to the testis is rare, most commonly occurring in patients with advanced prostate cancer and melanoma.

SURVIVORSHIP AND LATE EFFECTS

Because most patients with testicular GCT will experience long-term survival, survivorship care is important. Since many of these patients will be followed by primary care physicians, an understanding of the physical, psychological, and social late effects is important. Late effects are defined as health problems that occur months or years after a disease is diagnosed or after treatment has ended. Late effects may be related to the underlying cancer or to the treatment the patient received. In long-term survivors of testicular GCT, increased cardiovascular risk and increased secondary malignancies have been reported. Patients treated with cisplatin-based chemotherapy have an increased risk of hypertension, hyperlipidemia, metabolic syndrome, and cardiovascular events. Patients treated with high cumulative doses of etoposide (such as patients who receive standard chemotherapy, relapse, and then receive salvage high dose chemotherapy) may experience up to a 1–2% risk of developing acute myelogenous leukemia, typically 2–3 years after completing therapy

and associated with an 11q23 translocation. Patients treated with radiation therapy, cisplatin-based chemotherapy, or both have an increased risk of developing secondary solid malignancies.

FURTHER READING

Einhorn LH et al: High-dose chemotherapy and stem-cell rescue for metastatic germ cell tumors. N Engl J Med 357:340, 2007. [PubMed: 17652649]

Feldman DR et al: Medical treatment of advanced testicular cancer. JAMA 299:272, 2008.

Hanna NH, Einhorn LH: Testicular cancer-discoveries and updates. N Engl J Med 371:2005, 2014. [PubMed: 25409373]

International Germ Cell Cancer Collaborative Group: International Germ-Cell Consensus Classification: A prognostic factor based staging system for metastatic germ cell tumors. J Clin Oncol 15:594, 1997. [PubMed: 9053482]

International Prognostic Factors Study Group et al: Prognostic factors in patients with metastatic germ cell tumors who experienced treatment failure with cisplatin-based first-line chemotherapy. J Clin Oncol 28:4906, 2010. [PubMed: 20956623]

Kanetsky PA et al: Common variation in KITGL and at 5q31.3 predisposes to testicular germ cell cancer. Nat Genet 41:811, 2009. [PubMed: 19483682]

Kollmansberger C et al: Patterns of relapse in patients with clinical stage 1 testicular cancer managed with active surveillance. J Clin Oncol 33:51, 2015. [PubMed: 25135991]

Lorch A et al: Conventional-dose versus high-dose chemotherapy as first salvage treatment in male patients with metastatic germ cell tumors: Evidence from a large international database. J Clin Oncol 29:2178, 2011. [PubMed: 21444870]

Travis LB et al: Testicular cancer survivorship: Research strategies and recommendations. J Natl Cancer Inst 102:1114, 2010. [PubMed: 20585105]

Chapter 85: Gynecologic Malignancies

David Spriggs

OVARIAN CANCER

INCIDENCE AND PATHOLOGY

Ovarian cancer remains a leading cause of cancer deaths in American women, ranking behind lung, breast, colon, and pancreatic cancers. The ovary is responsible for the hormone and egg production. Between menarche (11–13 years) and menopause (45–55 years), the ovary is responsible for follicle maturation associated with egg maturation, ovulation, and cyclical sex steroid hormone production. These complex biologic functions are linked to stromal and germ cells within the ovary. These cells can be broadly grouped into stromal cells and ovarian germ cells and the enveloping epithelial cells. Malignancies arising in each group include multiple histological variants with unique neoplastic behaviors. Epithelial tumors are the most common histological variant of ovarian neoplasms; they may be benign (50%), frankly malignant (33%), or of borderline malignancy of low malignant potential (16%). In adnexal masses detected by imaging or physical examination, age influences risk of malignancy; tumors in younger women are more likely benign. In the malignant group, the most common tumors are epithelial. In the group of the ovarian epithelial, malignancies are the serous tumors (60–70%); mucinous tumors (10%), endometrioid (10–15%), and clear cell (10–15%), tumors. The distribution of histologic types varies in different parts of the world. Less common stromal tumors arise from the ancillary, supportive cells such as steroid hormone-producing cells and likewise have different phenotypes and clinical presentations. Most stromal tumors do not produce estrogen, but ectopic hormone production can be seen in certain subtypes. Tumors arising in the ovarian germ cell lineage are generally similar in biology and behavior to testicular tumors in males, although their intraperitoneal location alters some metastatic behaviors (Chap. 84). Ovarian tissue may also host metastatic tumors arising from breast, colon, gastric, and pancreatic primaries. Bilateral ovarian masses from metastatic mucin-secreting gastrointestinal cancers are termed *Krukenberg tumors*. A survey of other potential primaries is commonly required during the diagnostic workup of ovarian masses.

OVARIAN CANCER OF EPITHELIAL ORIGIN

Epidemiology

An American woman has ~1 in 72 lifetime risk (1.6%) of developing ovarian cancer, with the majority of affected women developing epithelial tumors. In 2017, 22,440 cases of ovarian cancer with 14,195 deaths are

expected in the United States. Sporadic (not familial) epithelial tumors of the ovary have a peak incidence in women in their fifties and sixties, although age at presentation ranges from the third decade to the eighties and nineties. Ovarian cancer risk has been linked to an interactive mixture of epidemiologic, environmental, and genetic factors. Nulliparity, obesity, diet, infertility treatments, and possibly hormone replacement therapy have all been linked to an increase in risk. Protective factors include the use of oral contraceptives, multiparity, tubal ligation, aspirin use, and breast-feeding. Other epidemiologic factors such as the use of perineal talcum agents remain controversial. The mechanisms underlying the various protective factors are largely unknown, but theories include suppression of ovulation, modulation of gonadotropins and progestins, and perhaps reduction of ovarian inflammation and damage associated with the repair of the ovarian cortex associated with ovulation.

GENETICS AND PATHOGENESIS

Ovarian cancers are divided into type 1 cancers and a more aggressive type 2 variant. The type 1 cancers are characterized by low-grade histology and more indolent behavior. These tumors include the low malignant potential tumors, low-grade endometrial and mucinous histologies, and clear cell cancers. Genetic alterations commonly include mutations in *KRAS, BRAF, PTEN*, and *PIK3CA*. In contrast, studies have implicated serial genetic changes in the fallopian tube as the actual site of origin for most type 2 serous epithelial ovarian cancers. These aggressive tumors are more common and linked to losses in *TP53* and DNA repair capacity. Carcinoma in situ has been identified in the tubal epithelium with early losses in *TP53* and the *BRCA1/BRCA2* genes characterizing early tubal intraepithelial cancers. Following these two early genetic events, additional mutations in these transformed cells lead to tumor cell shedding, metastasis, and invasion. These type 2, poorly differentiated “ovarian” cancer cells can then spread to the ovaries, and the peritoneal cavity, aided by the ovarian cancer cell’s affinity for mesothelial lining cells.

In work done as part of the Tumor Genome Atlas, type 2, serous ovarian cancer is principally a disease characterized by amplifications and deletions rather than point mutations. Damage to the tumor suppressor gene *TP53* occurs in >95% of serous ovarian cancers. Damage to homologous DNA repair genes including *BRCA1 and BRCA2* was also common in these tumors. Low prevalence but statistically recurrent somatic mutations in seven other genes including *NF1, RB1*, and *CDK12* were also seen. The most common heritable abnormality linked to ovarian cancer is a germ-line mutation in either *BRCA1* (chromosome 17q12-21) or *BRCA2* (chromosome 13q12-13). These genes are important parts of the homologous DNA repair machinery for double-stranded DNA break repair. Individuals inheriting a single copy of a mutant allele (these act as autosomal dominant genes) have an increased lifetime risk of breast (46–87% for *BRCA1*; 38–84% for *BRCA2*) and ovarian cancer (39–63% for *BRCA1*; 16.5–27% for *BRCA2*). Many of these women have a family history that includes multiple cases of breast and/or ovarian cancer of at an early age. Male breast cancer, pancreatic cancer, and prostate cancer are also linked to familial *BRCA2* mutations. The most common malignancy in these women is breast carcinoma, although women harboring germ-line *BRCA1* mutations have a marked increased risk of developing ovarian malignancies in their forties and fifties. Women harboring a mutation in *BRCA2* have a lower penetrance of ovarian cancer with onset typically in their fifties or sixties. Other uncommon germ-line mutation of other genes encoding proteins linked to homologous DNA repair (e.g.,

PALB2) can also contribute to cancer risk although the frequency mutation and magnitude of risk increment is much lower and not well defined. Screening studies, even in the *mBRCA1/mBRCA2* families, suggest that any of the available screening techniques, including serial evaluation of the CA-125 tumor marker and transvaginal ultrasound, are insufficient to reliably detect early-stage ovarian cancer. Uniform germ-line *BRCA1/BRAC2* testing is recommended for all incident epithelial ovarian cancers to detect probands and identify relatives at risk. Women with these high-risk germ-line mutations are advised to undergo prophylactic removal of fallopian tubes and ovaries after completing childbearing and ideally before age 40. Early prophylactic salpingo-oophorectomy is highly protective. Salpingo-oophorectomy also appears to protect these women from subsequent breast cancer (risk reduction 50%). Prophylactic salpingectomy is almost certainly a key part of any surgical prophylaxis strategy for ovarian cancer, but the benefits of oophorectomy on either ovarian or breast cancer risk have not yet been clearly defined. Although less common, ovarian cancer is also another form of cancer (along with colorectal and endometrial cancer) that may develop in women with Lynch syndrome, type II, caused by mutations in one of the DNA mismatch repair genes (*MSH2, MLH1, MLH6, PMS1, PMS2*). Ovarian cancer may appear in women <50 years of age in this syndrome.

Presentation

Neoplasms of the ovary tend to be painless unless they undergo torsion. Symptoms are therefore typically related to compression of local organs or due to symptoms from metastatic disease. Women with tumors localized to the ovary sometimes do have an increased incidence of symptoms including pelvic discomfort, bloating, and perhaps changes in a woman's typical urinary or bowel pattern. Unfortunately, these symptoms are common in primary care and are frequently dismissed by either the woman or her health care team until later stages of disease. The pathogenic factors and timing of spread beyond the ovary are still not well understood. The most common symptoms at presentation include a period of progressive complaints that typically include some combination of nausea, early satiety, bloating, indigestion, constipation, and abdominal pain. Signs include the rapid increase in abdominal girth due to the accumulation of ascites that typically alerts the patient and her physician that the concurrent gastrointestinal symptoms are likely associated with malignant pathology. Radiologic evaluation typically demonstrates a complex adnexal mass with ascites, carcinomatosis, with pelvic, para-aortic and mesenteric adenopathy in advanced disease. Positron emission tomography (PET) scans are generally not required. Laboratory evaluation demonstrates a markedly elevated CA-125, a shed mucin (MUC16) associated with, but not specific for, ovarian cancer. Ovarian cancers are divided into four stages, with stage I tumors confined to the ovary, stage II malignancies confined to the pelvis, and stage III confined to the peritoneal cavity and retroperitoneal nodes **(Table 85-1)**. These three stages are subdivided, with the most common presentation, stage IIIc, defined as tumors with bulky intraperitoneal disease or positive lymph node involvement. About 70% of women present with stage III disease. Stage IV disease includes women with parenchymal metastases (liver, lung, spleen) or, alternatively, abdominal wall or pleural disease. The 30% not presenting with stage III disease are roughly evenly distributed among the other stages.

TABLE 85-1

Staging and Survival in Gynecologic Malignancies

Stage	Ovarian	5-Year Survival, %	Endometrial	5-Year Survival, %	Cervix	5-Year Survival, %
0	—		—		Carcinoma in situ	100
I	Confined to ovary	88–95	Confined to corpus	>90	Confined to uterus	85
II	Confined to pelvic organs	70–80	Involves corpus and cervix	~75	Invades beyond uterus but not to pelvic wall	65
III	Intraabdominal spread to omentum, diaphragm or lymph nodes	20–40	Extends outside the uterus but not outside the true pelvis	45–60	Extends to pelvic wall and/or lower third of vagina, or hydronephrosis	35
IV	Spread outside abdominal cavity, parenchymal spread + pleural effusion cytology or extra-abdominal lymph nodes (inguinal, thoracic or supraclavicular)	17	Extends outside the true pelvis or involves the bladder or rectum	~20	Invades mucosa of bladder or rectum or extends beyond the true pelvis	7

Screening

Ovarian cancer is a highly lethal condition, curable in early stages, and seldom curable in advanced stages; hence, screening is of considerable interest. Early-stage tumors often secrete excessive amounts of normal

proteins that can be measured in the serum such as CA-125, mesothelin, and HE-4. Nevertheless, the incidence of ovarian cancer in the middle-aged female population is very low, with only ~1 in 2000 women between the ages of 50 and 60 carrying an asymptomatic and undetected tumor. Thus, effective screening techniques must be both sensitive and highly specific so to minimize the number of false positives. Panels of serum markers have not improved on CA-125 alone, although risk assessment by algorithms with multiple CA-125 is in advanced testing. No other screening strategies have been successful to date. Some large studies have suggested that low specificity screening might even worsen mortality in the screened population. Screening for ovarian cancer is currently not recommended outside of a clinical trial.

TREATMENT

TREATMENT

TREATMENT

Ovarian Cancer

TREATMENT

Epithelial ovarian cancer can be divided into distinct “disease states” for the purpose of treatment selection as shown in **Fig. 85-1**. Surgery by a skilled gynecologic oncologist remains the mainstay of initial therapy for ovarian cancer. The amount of residual visible cancer at the end of a primary operation is strongly predictive of outcome, and is paired with histology, grade, and stage to determine prognosis and treatment. In women presenting with a localized ovarian mass, the principal diagnostic and therapeutic maneuver is abdominal surgery to determine if the tumor is benign or malignant. In the event that the tumor is malignant, the surgical specimen will determine if the tumor arises in the ovary or is a site of metastatic disease. Metastatic disease to the ovary can be seen from primary tumors of the colon, appendix, stomach (Krukenberg tumors), and breast. Needle biopsy is contraindicated to avoid malignant contamination of the peritoneal cavity with malignant cells. Typically, women undergo laparoscopic evaluation and unilateral salpingo-oophorectomy for diagnostic purposes. If pathology reveals a primary ovarian malignancy or disseminated disease is present, then the procedure should be followed by a total hysterectomy, removal of the remaining tube and ovary, omentectomy, and pelvic node sampling along with biopsies of the peritoneal cavity and diaphragms. This extensive surgical procedure is performed because ~30% of tumors that by visual inspection appear to be confined to the ovary have already disseminated to the peritoneal cavity and/or surrounding lymph nodes. As with axillary dissections in breast cancer, node sampling is diagnostic but full lymphadenectomy appears to provide little or no additional therapeutic advantage over nodal sampling. The desired outcome of an ovarian cancer surgery is always an “R0” resection, with no visible residual cancer. The less favorable “optimal resection” (no disease greater than 1 cm in size) is still clinically useful and the prognosis of those patients is much better than the patients who are left with >1 cm disease at the end of surgery. These “suboptimally debulked” patients derive very little benefit from their surgery. Patients without gross residual disease after resection have a median survival in excess of 60 months, compared to 28–42 months for those left with macroscopic tumor.

After appropriate surgical treatment, primary chemotherapy will consist of combination treatment with paclitaxel and carboplatin. Primary chemotherapy can be delivered intravenously or alternatively, some

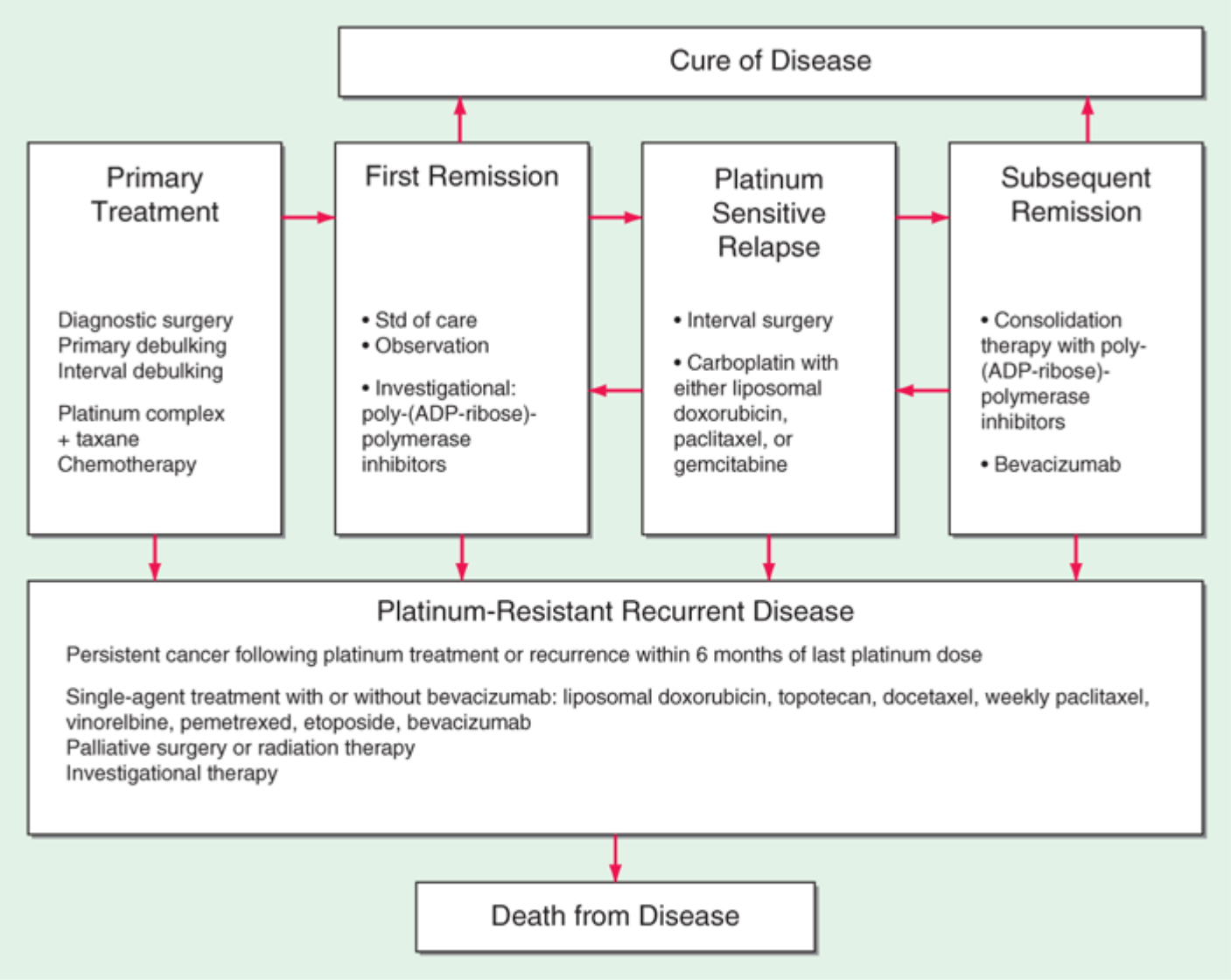
therapy can be directly administered into the peritoneal cavity via an indwelling catheter. Several randomized studies have demonstrated improved survival with intraperitoneal therapy, but this approach is technically difficult and is increasingly replaced by carboplatin and dose dense (weekly) paclitaxel, which appears to offer similar results in some studies.

With optimal debulking surgery and platinum-based chemotherapy (usually carboplatin dosed to an area under the curve [AUC] of 6.0 plus paclitaxel 175 mg/m^2 by 3-h infusion in monthly cycles), 70% of women who present with advanced-stage tumors respond, and 40–50% experience a complete remission with normalization of their CA-125, CT scans, and physical examination. These patients are sometimes enrolled in consolidation trials to extend remission and increase likelihood of cure. New immunotherapies and poly–ADP ribose polymerase inhibitors (PARPi) such as olaparib are in active testing in these patients because less than half of the complete responders are cured. Disease recurs within 1–4 years from the completion of their primary therapy. CA-125 levels often increase as a first sign of relapse and CT scan findings are confirmatory. Recurrent disease is managed, but rarely cured, with additional surgery and variety of chemotherapeutic agents. Eventually all of these women develop chemotherapy-refractory disease and refractory ascites, poor bowel motility, and obstruction or tumor-infiltrated aperistaltic bowel are all common premorbid events. Limited surgery to relieve intestinal obstruction, localized radiation therapy to relieve pressure or pain from masses, or palliative chemotherapy may be helpful. Agents with >15% response rates include gemcitabine, topotecan, liposomal doxorubicin, and bevacizumab.

Five-year survival correlates with the stage of disease: stage I, 90–95%; stage II, 70–80%; stage III, 25–40%; stage IV, 10–15% (Table 85-1). Prognosis is also influenced by histologic grade: 5-year survival is 88% for well-differentiated tumors, 58% for moderately differentiated tumors, and 27% for poorly differentiated tumors. Histologic type has less influence on outcome.

FIGURE 85-1

Disease states model of epithelial ovarian cancer and its treatment. Each box represents a relatively homogeneous group of patients that share a palette of potential treatment choices and have a similar prognosis. The arrows indicate that a single patient may move from one state to another during the course of her illness and the choice of treatments will become different in her new disease state.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

UNCOMMON OVARIAN TUMORS

Low Malignant Potential Tumors (Borderline Tumors)

These type 1 tumors are found in younger women (ages 40–50), indolent in behavior and few of these patients will succumb to their tumors (10 years survival may approach 98%) although recurrence is not uncommon. Certain features like micropapillary histology and microinvasion are linked to a more aggressive behavior. Patients with tumors of low malignant potential are managed primarily by surgery; chemotherapy and radiation therapy do not substantially alter survival.

Stromal Tumors

Approximately 7% of ovarian neoplasms are stromal tumors, with ~1800 cases expected each year in the United States. Ovarian stromal tumors or sex cord tumors are most common in women in their fifties or sixties, but tumors can present at any age. These tumors arise from the mesenchymal components of the

ovary, including both steroid-producing cells and fibroblasts. Most of these tumors are indolent tumors with limited metastatic potential and present as unilateral solid masses. These tumors primarily are discovered by the detection of an abdominal mass sometimes with abdominal pain due to ovarian torsion, intratumoral hemorrhage, or rupture. Rarely, stromal tumors can produce estrogen and present with breast tenderness as well as precocious puberty in children, menstrual disturbances in reproductively active women, or postmenopausal bleeding. In some women, estrogen-associated secondary malignancies, such as endometrial or breast cancer, may present as synchronous malignancies. Sertoli-Leydig tumors often present with hirsutism, virilization due to increased production androgens. Hormonally inert tumors include fibroma that presents as a solitary mass often in association with ascites and occasionally hydrothorax also known as Meigs' syndrome. A subset of these tumors present in individuals with a variety of inherited disorders that predispose them to mesenchymal neoplasia including Ollier's disease (juvenile granulosa cell tumors) and Peutz-Jeghers syndrome (ovarian sex cord tumors). The treatment of these tumors is almost exclusively by surgical resection. Chemotherapy with carboplatin and paclitaxel is generally reserved for either unresectable or multiply recurrent tumors.

Germ Cell Tumors of the Ovary

Germ cell tumors, like their counterparts in the testis, are cancers of germ cells. These totipotent cells contain the programming for differentiation to essentially all tissue types, and hence the germ cell tumors include a histologic menagerie of bizarre tumors, including benign teratomas (dermoid cysts) and a variety of malignant tumors, such as dysgerminoma, immature teratomas, yolk sac malignancies, and choriocarcinomas. Benign teratoma (or dermoid cyst) is the most common germ cell neoplasm of the ovary and often presents in young woman. These tumors include a complex mixture of differentiated tissue including tissues from all three germ layers. In older women these differentiated tumors can develop malignant transformation, most commonly squamous cell carcinomas. Malignant germ cell tumors include dysgerminomas, yolk sac tumors, immature teratomas, as well as embryonal and choriocarcinomas. Germ cell tumors can present at all ages, but the peak age of presentation tends to be in adolescents. Typically these tumors will become large ovarian masses, which eventually present as palpable low abdominal or pelvic masses. Like sex cord tumors, torsion or hemorrhage may present urgently or emergently as acute abdominal pain. Some of germ cell tumors produce elevated levels of human chorionic gonadotropin (hCG) or α fetoprotein (AFP). Unlike epithelial ovarian cancer, these tumors have a higher proclivity for nodal or hematogenous metastases. Germ cell tumors typically present in women who are of childbearing age, and because bilateral tumors are uncommon (except in dysgerminoma, 10–15%), the typical treatment is unilateral oophorectomy or salpingo-oophorectomy with lymph node sampling. Most commonly, women with advanced malignant germ cell tumors typically receive bleomycin, etoposide, and cisplatin (BEP) chemotherapy, in an analogous fashion to the treatment of testicular cancers. In the majority of these women, even those with advanced-stage disease, cure is expected. Dysgerminoma is the ovarian counterpart of testicular seminoma and is highly curable. Although the tumor is highly radiation-sensitive, radiation produces infertility in many patients. BEP chemotherapy is as effective or more so without causing infertility.

FALLOPIAN TUBE CANCER

Transport of the egg to the uterus occurs through the fallopian tube, with the distal ends of these tubes composed of fimbriae that drape about the ovarian surface and capture the egg as it erupts from the ovarian cortex. As described previously, the majority of type 2 ovarian cancers are now thought to arise from the tubal epithelium. As might be expected, fallopian tube malignancies are typically serous histology and share the biology and recommended treatment as serous ovarian cancer. These tumors often present as clinically isolated adnexal masses, but like ovarian cancer, these tumors spread relatively early throughout the peritoneal cavity. Fallopian tubal cancers have a natural history and treatment that is essentially identical to ovarian cancer (Table 85-1).

CERVICAL CANCER

ETIOLOGY AND GENETICS

Cervical cancer is the second most common and the most lethal malignancy in women worldwide. Infection with high-risk strains of human papillomavirus (HPV) is the primary neoplastic-initiating event in the vast majority of women with invasive cervical cancer. This double-strand DNA virus infects epithelium near the transformation zone of the cervix where underlying columnar epithelium becomes squamous epithelium. More than 60 types of HPV are known, with ~20 types having the ability to generate high-grade dysplasia and malignancy. HPV16 and 18 are the types most frequently associated with high-grade dysplasia, but types 31, 33, 35, 52, and 58 are also considered to be high-risk variants. The large majority of sexually active adults are exposed to HPV, and most women clear the infection without specific intervention. The 8-kilobase HPV genome encodes seven early genes, most notably *E6* and *E7*, which can bind to *RB* and *p53*, respectively. High-risk types of HPV encode *E6* and *E7* molecules that are particularly effective at inhibiting the normal cell cycle checkpoint functions of these regulatory proteins, leading to immortalization but not full transformation of cervical epithelium. A minority of women will fail to clear the infection with subsequent HPV integration into the host genome. Over as little as a few months to several years, some of these persistently infected women develop worsening dysplasia, a premalignant condition that, untreated, can progress to cervical carcinoma. Complete transformation to cancer occurs over a period years and almost certainly requires the acquisition of other poorly defined genetic mutations within the infected and immortalized epithelium.

Approximately 528,000 new cases of cervical cancer were reported in 2012 worldwide with approximately an estimated 266,000 deaths. Cancer incidence is particularly high in women residing in central and South America, the Caribbean, and southern and eastern Africa. Mortality rate is disproportionately high in Africa. In the United States, an estimated 12,800 women will be diagnosed with cervical cancer this year, and 4210 women will die of the disease. Efforts in developed countries have looked at high-technology screening techniques for HPV involving polymerase chain reaction (PCR) and other molecular technologies.

In the integrated genomic characterization of cervical cancer by the Cancer Genome Atlas (TCGA), integration of HPV sequences was found in all of the HPV18 linked cancers and over 3 quarters of the HPV16 cancers. The cervical tumors also showed a characteristic APOBEC (apolipoprotein B mRNA editing enzyme, catalytic

polypeptide-like; a family of cytidine deaminases that edit DNA and are endogenous mutagenic enzymes) pattern of mutagenesis with *ERBB3, CASP8*, and *TGFRB2* identified as significantly mutated genes presumably linked to progression from dysplasia to carcinoma. Amplification of immune targets PD-L1 and PD-L2 was also seen, which may suggest vulnerability to immunotherapy. In the much smaller number of HPV negative cancers, mutations in the common oncogenes *KRAS, ARID1A*, and *PTEN* were commonly seen. The clinical behavior of these cancers is likely to be different.

HPV INFECTION AND PREVENTION

The Pap smear is the primary detection method for asymptomatic preinvasive cervical dysplasia of squamous epithelial lining during a gynecologic examination. Because of the delay between dysplasia and frank cervical cancer is years long; annual (or longer) screening and prevention strategies that detect precancerous dysplasia and carcinoma in situ can be implemented successfully. Annual or biannual cervical scraping for cytology (Pap Smear) is highly effective in reducing the incidence of cervical cancer by early detection and subsequent surgical treatment. Although no randomized trial data demonstrate the utility of Pap smears, the dramatic drop in cervical cancer incidence and death in developed countries employing wide-scale screening provides strong evidence for its effectiveness. The incorporation of HPV testing by PCR or other molecular techniques increases the sensitivity of detecting cervical pathology but at the cost of lower sensitivity in that it identifies many women with transient infections who require no specific medical intervention. Unfortunately, both the collection of a Pap smear and its cytological evaluation require infrastructure beyond the means of many middle- and low-income countries. High-throughput, low-technology prevention strategies are needed to identify and treat women bearing high-risk but treatable cervical dysplasia.

A primary prevention strategy relies on HPV vaccines. Currently approved vaccines include the recombinant proteins to the late proteins, L1 and L2 of HPV-16 and -18 as well as other, less common cancer causing isotypes 11, 31, 33, 45, 52, and 58. Vaccination of girls aged 11–13 years with two injections (one year apart) before the initiation of sexual activity dramatically reduces the rate of high-risk HPV infection and subsequent dysplasia. There is also partial protection against other HPV types, although vaccinated women are still at risk for HPV infection and still benefit from standard Pap smear screening.

CLINICAL PRESENTATIONS

Risk Factors

Clinical risk factors include many HPV infection–linked features: a high number of sexual partners, early age of first intercourse, and history of venereal disease. Smoking is a cofactor; heavy smokers have a higher risk of dysplasia with HPV infection. HIV infection, especially when associated with low CD4+ T-cell counts, is associated with a higher rate of high-grade dysplasia and likely a shorter latency period between infection and invasive disease. Histologically, the majority of cervical malignancies are squamous cell carcinomas associated with HPV, but adenocarcinomas are also HPV-related, and both arise in transitional zone of the endocervical canal; the lesions in the canal or cervical glands may not be seen by visual inspection of the

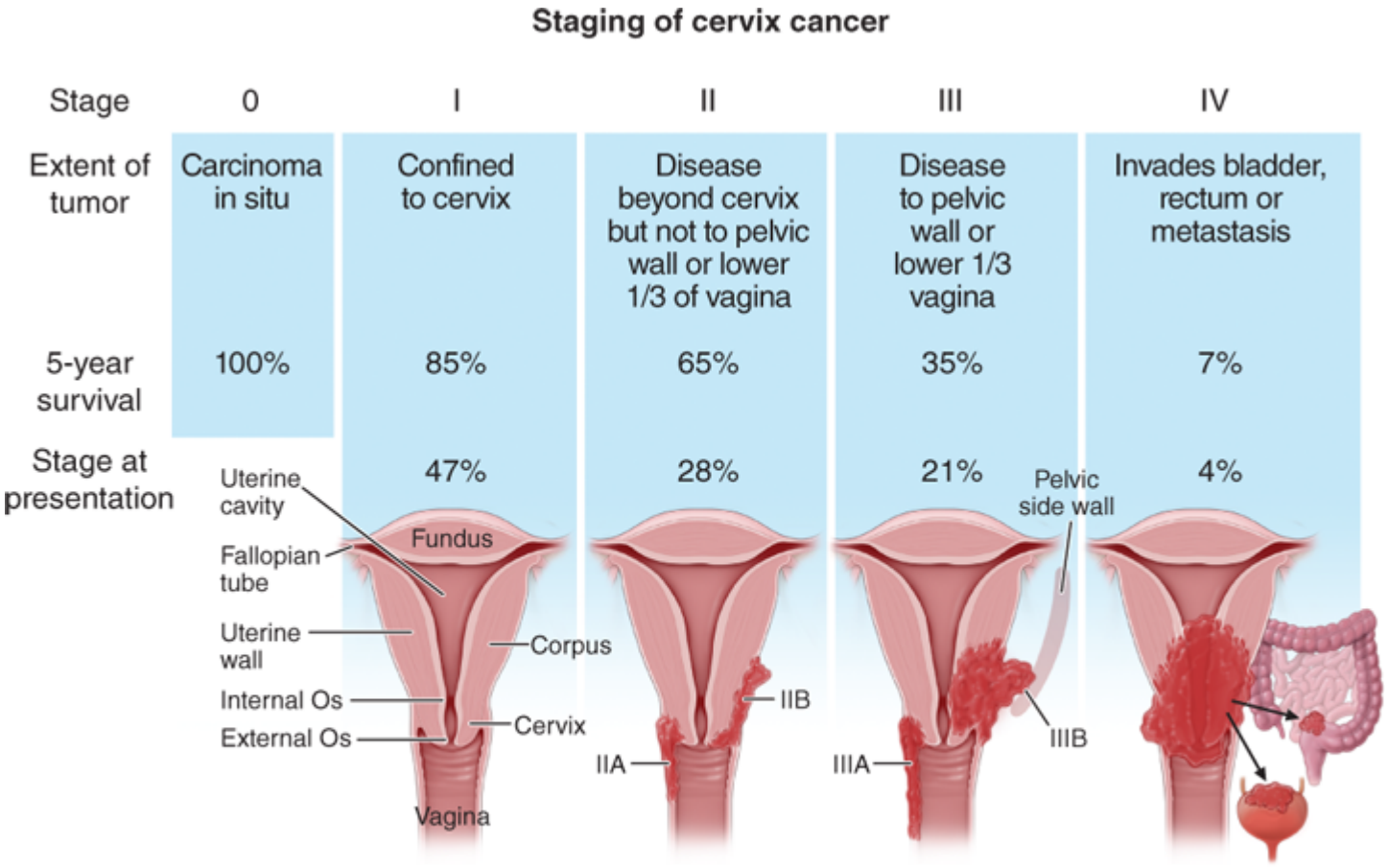
cervix and can be missed by Pap smear screening. Uncommon malignancies including carcinoids, small cell carcinomas, sarcomas, and lymphomas are also found but are linked to HPV infection.

Diagnosis of Cervical Cancer

Early cancer of the cervix is asymptomatic and this underlies the recommendations for routine gynecologic care. Larger, invasive carcinomas often have symptoms or signs including postcoital spotting or intermenstrual cycle bleeding or menometrorrhagia. Foul-smelling or persistent yellow discharge may also be seen. Presentations that include pelvic or sacral pain suggest lateral extension of the tumor into pelvic nerve plexus by either the primary tumor or a pelvic node and are signs of advanced-stage disease. Likewise, flank pain from hydronephrosis from ureteral compression or deep venous thrombosis from iliac vessel compression suggests either extensive nodal disease or direct extension of the primary tumor to the pelvic sidewall. The most common finding upon physical examination is a visible tumor on the cervix. Larger tumors may be identified by inspection and biopsied directly. Staging of cervical cancer is performed by clinical examination. Stage I cervical tumors are confined to the cervix, whereas stage II tumors extend into the upper vagina or paracervical soft tissue **(Fig. 85-2)**. Stage III tumors extend to the lower vagina or the pelvic sidewalls, whereas stage IV tumors invade the bladder or rectum or have spread to distant sites. While radiographic studies are not part of the formal clinical staging of cervical cancer, treatment planning requires them for appropriate therapy. CT can detect hydronephrosis indicative of pelvic sidewall disease but is not accurate at evaluating other pelvic structures. MRI is more accurate at estimating uterine extension and paracervical extension of disease into soft tissues typically bordered by broad and cardinal ligaments that support the uterus in the central pelvis. Very small stage I cervical tumors can be treated with a variety of surgical procedures. In young women desiring to maintain fertility, radical trachelectomy removes the cervix with subsequent anastomosis of the upper vagina to the uterine corpus; however, subsequent pregnancies may be more problematic. Large stage I cervical tumors (4 cm) confined to the cervix and all stage II–IV patients are treated with radiation therapy in combination with cisplatin-based chemotherapy. This multimodality treatment can offer the patient with advanced stage disease, a 40–80% of cure depending on the clinical circumstances. Platinum agents (cisplatin or carboplatin) combined with paclitaxel and bevacizumab are generally considered as the best palliative choice for metastatic cervical cancer patients. Secondary chemotherapy confers minimal improvement in most patients.

FIGURE 85-2

Anatomic display of the stages of cervix cancer defined by location, extent of tumor, frequency of presentation, and 5-year survival. *(From MV Seiden: Gynecologic malignancies, in Harrison's Principles of Internal Medicine, 19th ed. New York, McGraw-Hill.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

UTERINE CANCER

EPIDEMIOLOGY

Several different tumor types arise in uterine corpus. Most tumors arise in the glandular lining and are endometrial adenocarcinomas. Benign (leiomyomas) and malignant tumors (leiomyosarcomas) can also arise in the uterine smooth muscle and have very different clinical features. The endometrioid histologic subtype of endometrial cancer is the most common gynecologic malignancy in the United States. In 2017, over 60,000 new corpus cancers of uterus are projected for American women, but the surgical cure rate is high, and about 10,920 deaths from uterine cancers are predicted. Development of these tumors is a multistep process with estrogen playing an important early role in driving endometrial gland proliferation. Relative overexposure to this class of hormones is the principal risk factor for the subsequent development of endometrioid tumors. In contrast, progestins drive glandular maturation and are protective. Hence, women with high endogenous or pharmacologic exposure to estrogens, especially if unopposed by progesterone, are at higher risk for endometrial cancer. Obese women, women treated with postmenopausal estrogens or women with estrogen-producing tumors are at higher risk for endometrial cancer. In addition, long-term treatment with tamoxifen, which has antiestrogenic effects in breast tissue but can show weak estrogenic effects in uterine epithelium, is associated with an increased risk of endometrial cancer.

Genetics



Women with a germ-line mutation in one of the series of DNA mismatch repair genes associated with the Lynch syndrome, also known as hereditary nonpolyposis colon cancer (HNPCC) syndrome, are at increased risk for endometrioid endometrial carcinoma. These individuals have germ-line mutations in *MSH2, MLH1*, and in rare cases *PMS1* and *PMS2*. Individuals who carry these mutations typically have a family history of cancer and are at markedly increased risk for colon cancer and modestly increased risk for ovarian cancer and a variety of other tumors. Middle-aged women with HNPCC carry a 4% annual risk of endometrial cancer and a relative overall risk of approximately 200-fold as compared to age-matched women without HNPCC. In sporadic cancers, secondary events such as mutation of the *PI3K* gene or the loss of the *PTEN* tumor suppressor gene likely serve as secondary “hits” in the carcinogenesis of estrogenic excess. The molecular events that underlie less common endometrial cancers such as clear cell and papillary serous tumors of the uterine corpus are less well understood.

PATHOLOGY

Approximately 75–80% of endometrial cancers are adenocarcinomas and have been characterized as type 1 (estrogen-linked) endometrial cancers and type 2 cancers that have less clear associations with estrogens (clear cell cancers, serous cancers, and mucinous cancers). Endometrial serous cancers show TP53 functional loss and behave clinically like ovarian cancers. Serous endometrial cancers are marked by a much higher risk of distant recurrence and a lower risk for locoregional spread. Prognosis depends on stage, histologic grade, and depth of myometrial invasion.

CLINICAL PRESENTATION

The majority of women with tumors of the uterine corpus present with postmenopausal vaginal bleeding due to shedding of the malignant endometrial lining. Premenopausal women often will present with atypical bleeding between typical menstrual cycles. These signs typically bring a woman to the attention of a health care professional, and the majority of women present with early-stage disease in which the tumor is confined to the uterine corpus and the consequent high cure rate. Diagnosis is typically established by endometrial biopsy. Epithelial tumors may spread to pelvic or para-aortic lymph nodes. Serous tumors tend to have patterns of spread much more reminiscent of ovarian cancer, and patients may present with omental/peritoneal disease and sometimes ascites. Some women with endometrial cancer have a history of endometriosis. Some women presenting with uterine sarcomas will present with pelvic pain. Sarcomas commonly are found by detection of symptomatic large pelvic masses that may or may not be associated with dysfunctional bleeding.

TREATMENT

TREATMENT

Uterine Cancer

Most women with endometrial cancer have disease that is localized to the uterus (75% are stage I, Table 85-1), and definitive treatment typically involves a hysterectomy with removal of the ovaries and fallopian tubes. The resection of lymph nodes does not improve outcome, but sentinel node resection does provide staging and prognostic information. Node involvement defines stage IIIC disease. Tumor grade and depth of invasion are the two key prognostic variables in early-stage tumors, and women with low-grade and/or minimally invasive tumors (<50% myometrial penetration) are typically observed after definitive surgical therapy. Patients with high-grade tumors or tumors that are deeply invasive (stage IB) are at higher risk for pelvic recurrence or recurrence at the vaginal cuff, which is typically prevented by intravaginal brachytherapy.

Women with regional metastases or metastatic disease (3% of patients) with low-grade tumors can be treated with progesterone or tamoxifen. Poorly differentiated tumors lack hormone receptors and are typically resistant to hormonal manipulation. The role of adjuvant chemotherapy in stage I–II disease is currently under investigation but is usually employed for advanced stage (III–IV) cancer and most tumors with serous histology. Carboplatin and paclitaxel combinations are the current standard of care. Chemotherapy for metastatic disease is delivered with palliative intent. Potentially active drugs include bevacizumab, mTOR inhibitors (e.g., temsirolimus). Patients with advanced cancer and known mismatch repair deficits may respond particularly well to immunotherapy with antagonists of the PD1/PDL1 axis.

Chemotherapy of leiomyosarcomas of the uterus with docetaxel/gemcitabine, ifosfamide/doxorubicin, and trabectedin can have substantial benefit. Carcinosarcomas of the uterus contain both mesenchymal and epithelial components but will often respond to paclitaxel and platinum complex therapy.

GESTATIONAL TROPHOBLASTIC TUMORS



Gestational trophoblastic diseases represent a spectrum of neoplasia from benign hydatidiform mole to choriocarcinoma due to persistent trophoblastic disease associated most commonly with molar pregnancy but occasionally seen after normal gestation. The most common presentations of trophoblastic tumors are partial and complete molar pregnancies. These represent ~1 in 1500 conceptions in developed Western countries. The incidence widely varies globally, with areas in Southeast Asia having a much higher incidence of molar pregnancy. Regions with high molar pregnancy rates are often associated with diets low in carotene and animal fats.

RISK FACTORS

Trophoblastic tumors result from the outgrowth or persistence of placental tissue. They arise most commonly in the uterus but can also arise in other sites such as the fallopian tubes due to ectopic pregnancy. Risk factors include poorly defined dietary and environmental factors as well as conceptions at the extremes of reproductive age, with the incidence particularly high in females conceiving younger than age 16 or older than age 50. In older women, the incidence of molar pregnancy might be as high as one in three, likely due to increased risk of abnormal fertilization of the aged ova. Most trophoblastic neoplasms are associated with

complete moles, diploid tumors with all genetic material from the paternal donor (known as uniparental disomy). This is thought to occur when a single sperm fertilizes an enucleate egg that subsequently duplicates the paternal DNA. Trophoblastic proliferation occurs with exuberant villous stroma. If pseudopregnancy extends out past the 12th week, fluid progressively accumulates within the stroma leading to “hydropic changes.” There is no fetal development in complete moles.

Partial moles arise from the fertilization of an egg with two sperm; hence two-thirds of genetic material is paternal in these triploid tumors. Hydropic changes are less dramatic, and fetal development can often occur through late first trimester or early second trimester at which point spontaneous abortion is common. Laboratory findings will include excessively high hCG and high AFP. The risk of persistent gestational trophoblastic disease after partial mole is ~5%. Complete and partial moles can be noninvasive or invasive. Myometrial invasion occurs in no more than one in six complete moles and a lower portion of partial moles.

PRESENTATION OF INVASIVE TROPHOBLASTIC DISEASE

The clinical presentation of molar pregnancy is changing in developed countries due to the early detection of pregnancy with home pregnancy kits and the very early use of Doppler and ultrasound to evaluate the early fetus and uterine cavity for evidence of a viable fetus. Thus, in these countries, the majority of women presenting with trophoblastic disease have their moles detected early and have typical symptoms of early pregnancy including nausea, amenorrhea, and breast tenderness. With uterine evacuation of early complete and partial moles, most women experience spontaneous remission of their disease as monitored by serial hCG levels. These women require no chemotherapy. Patients with persistent elevation of hCG or rising hCG postevacuation have persistent or actively growing gestational trophoblastic disease and require therapy. Most series suggest that between 15 and 25% of women will have evidence of persistent gestational trophoblastic disease after molar evacuation.

In women who lack access to prenatal care, presenting symptoms can be life threatening including the development of preeclampsia or even eclampsia. Hyperthyroidism can also be seen. Evacuation of large moles can be associated with life-threatening complications including uterine perforation, volume loss, high-output cardiac failure, and adult respiratory distress syndrome (ARDS).

For women with evidence of rising hCG or radiologic confirmation of metastatic or persistent regional disease, prognosis can be estimated through a variety of scoring algorithms that identify those women at low, intermediate, and high risk for requiring multiagent chemotherapy. In general, women with widely metastatic nonpulmonary disease, very elevated hCG, and prior normal antecedent term pregnancy are considered at high risk and typically require multiagent chemotherapy at an expert center for cure. Even very advanced gestational trophoblastic disease is almost uniformly curable when managed by an expert in this rare malignancy.

TREATMENT

TREATMENT

Invasive Trophoblastic Disease

Management of invasive trophoblastic disease should be 100% curative and complex patients should be managed by clinicians experienced in this disease. The management for a persistent and rising hCG postevacuation of a molar conception is typically chemotherapy, although surgery can play an important role for chemotherapy-resistant disease that is isolated in the uterus (especially if childbearing is complete) or to control hemorrhage. For women wishing to maintain fertility or with metastatic disease, the preferred treatment is chemotherapy. Trophoblastic disease is exquisitely sensitive to chemotherapy and guided by serial serum hCG testing, successful, curative treatment is the rule. Single-agent treatment with methotrexate or actinomycin D cures 90% of women with low-risk disease. Patients with high-risk disease (very high hCG levels, presentation 4 or more months after pregnancy, brain or liver metastases, failure of methotrexate therapy) are typically treated with multiagent chemotherapy (etoposide, methotrexate, and actinomycin D alternating with cyclophosphamide and vincristine [EMA-CO]), which is typically curative even in those women with extensive metastatic disease. Cisplatin and etoposide alternating with etoposide/methotrexate/actinomycin D is used for the highest risk patients. In the highest-risk patients with liver lung and brain metastases, hemorrhage from the rich tumor vasculature is a major risk during chemotherapy initiation. Cured women may get pregnant again without evidence of increased fetal or maternal complications.

ACKNOWLEDGMENT

Michael V. Seiden was the author of this chapter in the 19th edition. Material from his chapter has been included here.

FURTHER READING

Brown J et al: 15 years of progress in gestational trophoblastic disease: Scoring, standardization and salvage. Gynecol Oncol 49:241, 2017.

Chuang LT et al: Management and care of women with invasive cervical cancer: American Society of Clinical Oncology Resource Stratified Clinical Practice Guideline. J Global Oncol 2:311, 2016.

Hartmann LC, Lindor NM: The role of risk reducing surgery in hereditary breast and ovarian cancer. N Engl J Med 374:454, 2016. [PubMed: 26840135]

Joura EA et al: A 9 valent HPV vaccine against infection and intraepithelial neoplasia in women. N Engl J Med 372:711, 2015. [PubMed: 25693011]

Morice P et al: Endometrial cancer. Lancet 2016, 387:1094, 2016.

Chapter 86: Primary and Metastatic Tumors of the Nervous System

Lisa M. DeAngelis; Patrick Y. Wen

INTRODUCTION

Primary brain tumors are diagnosed in ~78,000 people each year in the United States. At least 25,000 are malignant, and most of these are gliomas. Meningiomas account for 35%, vestibular schwannomas 10%, and central nervous system (CNS) lymphomas ~2%. Brain metastases are three times more common than all primary brain tumors combined and are diagnosed in ~150,000 people each year. Metastases to the leptomeninges and epidural space of the spinal cord each occur in ~3–5% of patients with systemic cancer and are also a major cause of neurologic disability.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Primary and Metastatic Tumors of the Nervous System

CLINICAL FEATURES

Brain tumors of any type can present with a variety of symptoms and signs that fall into two categories: general and focal; patients often have a combination of the two (Table 86-1). General or nonspecific symptoms include headache, with or without nausea or vomiting, cognitive difficulties, personality change, and gait disorder. Generalized symptoms arise when the enlarging tumor and its surrounding edema cause an increase in intracranial pressure or compression of cerebrospinal fluid (CSF) circulation leading to hydrocephalus. The classic brain tumor headache predominates in the morning and improves during the day, but this pattern is seen only in a minority of patients. Headaches are often holocephalic but can be ipsilateral to the side of a tumor. Occasionally, headaches have features of a typical migraine with unilateral throbbing pain associated with visual scotoma. Personality changes may include apathy and withdrawal from social situations, mimicking depression. Focal or lateralizing findings include hemiparesis, aphasia, or visual field defect. Lateralizing symptoms are typically subacute and progressive; language difficulties may be mistaken for confusion. Seizures are common, occurring in ~25% of patients with brain metastases or malignant gliomas and are the presenting symptom in up to 90% of patients with a low-grade glioma. All seizures that arise from a brain tumor will have a focal onset whether or not it is apparent clinically.

NEUROIMAGING

Cranial magnetic resonance imaging (MRI) is the preferred diagnostic test for any patient suspected of having a brain tumor and should be performed with gadolinium contrast administration. Computed tomography (CT) scan should be reserved for those patients unable to undergo MRI. Malignant brain tumors—whether primary or metastatic—typically enhance with gadolinium, have central areas of necrosis, and are surrounded by edema of the neighboring white matter. Low-grade gliomas usually do not enhance with gadolinium and are best appreciated on fluid-attenuated inversion recovery (FLAIR) MRIs. Meningiomas have a typical appearance on MRI because they are dural-based enhancing tumors with a dural tail and compress but do not invade the brain. Dural metastases or a dural lymphoma can have a similar appearance. Imaging is characteristic for many primary and metastatic tumors and sometimes will suffice to establish a diagnosis when the location precludes surgical intervention (e.g., brainstem glioma). Functional MRI is useful in presurgical planning to define eloquent sensory, motor, or language cortex. Positron emission tomography (PET) is useful in determining the metabolic activity of the lesions seen on MRI; MR perfusion and spectroscopy can provide information on blood flow or tissue composition. These techniques may help distinguish tumor progression from necrotic tissue as a consequence of treatment with radiation and chemotherapy. Neuroimaging is the only test necessary to diagnose a brain tumor. Laboratory tests are rarely useful, although patients with metastatic disease may have elevation of a serum tumor marker (e.g., β human chorionic gonadotropin [β-hCG] from testicular cancer). Additional testing such as cerebral angiogram, electroencephalogram (EEG), or lumbar puncture is rarely indicated or helpful.

TABLE 86-1

Symptoms and Signs at Presentation of Brain Tumors

	High-Grade Glioma (%)	Low-Grade Glioma (%)	Meningioma (%)	Metastases (%)
Generalized				
Impaired cognitive function	50	10	30	60
Hemiparesis	40	10	36	60
Headache	50	40	37	50
Lateralizing				
Seizures	20	70+	17	18
Aphasia	20	<5	—	18
Visual field deficit	—	—	—	7

TREATMENT

TREATMENT

Brain Tumors

Therapy of any intracranial malignancy requires both symptomatic and definitive treatments. Definitive treatment is based on the specific tumor type and includes surgery, radiotherapy, and chemotherapy. However, symptomatic treatments apply to brain tumors of any type. Most high-grade malignancies are accompanied by substantial surrounding edema, which contributes to neurologic disability and raised intracranial pressure. Glucocorticoids are highly effective at reducing perilesional edema and improving neurologic function, often within hours of administration. Dexamethasone has been the glucocorticoid of choice because of its relatively low mineralocorticoid activity; initial doses are 8–16 mg/d. Glucocorticoids rapidly ameliorate symptoms and signs, but their long-term use causes substantial toxicity including insomnia, weight gain, diabetes mellitus, steroid myopathy, and personality changes. Consequently, a taper is indicated as definitive treatment is administered and the patient improves.

Patients with brain tumors who present with seizures require antiepileptic drug therapy. There is no role for prophylactic antiepileptic drugs in patients who have not had a seizure. The agents of choice are those drugs that do not induce the hepatic microsomal enzyme system. These include levetiracetam, topiramate, lamotrigine, valproic acid, and lacosamide **(Chap. 418)**. Other drugs, such as phenytoin and carbamazepine, are used less frequently because they are potent enzyme inducers that can interfere with both glucocorticoid and chemotherapy metabolism. Venous thromboembolic disease occurs in 20–30% of patients with high-grade gliomas or brain metastases. Prophylactic anticoagulants should be used during hospitalization and in nonambulatory patients. Those who have had either a deep vein thrombosis or pulmonary embolus can receive therapeutic doses of anticoagulation safely and without increasing the risk for hemorrhage into the tumor. Inferior vena cava filters are reserved for patients with absolute contraindications to anticoagulation such as recent craniotomy.

PRIMARY BRAIN TUMORS

EPIDEMIOLOGY

No underlying cause has been identified for the majority of primary brain tumors. The only established risk factors are exposure to ionizing radiation (meningiomas, gliomas, and schwannomas) and immunosuppression (primary CNS lymphoma). There is no proven evidence for any association with exposure to electromagnetic fields including cellular telephones, head injury, foods containing *N*-nitroso compounds, or occupational risk factors. A small minority of patients have a family history of brain tumors. Some of these familial cases are associated with genetic syndromes **(Table 86-2)**.

TABLE 86-2

Genetic Syndromes Associated with Primary Brain Tumors

Syndrome	Inheritance	Gene/Protein	Associated Tumors
Cowden's syndrome	AD	Mutations of *PTEN* (ch10p23)	Dysplastic cerebellar gangliocytoma (Lhermitte-Duclos disease), meningioma, astrocytoma
			Breast, endometrial, thyroid cancer, trichilemmomas
Familial schwannomatosis	Sporadic Hereditary	Mutations in *INI1/SNF5* (ch22q11)	Schwannomas, gliomas
Gardner's syndrome	AD	Mutations in *APC* (ch5q21)	Medulloblastoma, glioblastoma, craniopharyngioma
			Familial polyposis, multiple osteomas, skin and soft tissue tumors
Gorlin syndrome (basal cell nevus syndrome)	AD	Mutations in *Patched 1* gene (ch9q22.3)	Medulloblastomas
			Basal cell carcinoma
Li-Fraumeni syndrome	AD	Mutations in *p53* (ch17p13.1)	Gliomas, medulloblastomas
			Sarcomas, breast cancer, leukemias, others
Multiple endocrine neoplasia 1 (Werner's syndrome)	AD	Mutations in *Menin* (ch11q13)	Pituitary adenoma, malignant schwannomas
			Parathyroid and pancreatic islet cell tumors
NF1	AD	Mutations in *NF1*/neurofibromin (ch17q12-22)	Schwannomas, astrocytomas, optic nerve gliomas, meningiomas
			Neurofibromas, neurofibrosarcomas, others

Syndrome	Inheritance	Gene/Protein	Associated Tumors
NF2	AD	Mutations in *NF2*/merlin (ch22q12)	Bilateral vestibular schwannomas, astrocytomas, multiple meningiomas, ependymomas
TSC (Bourneville disease)	AD	Mutations in *TSC1/TSC2* (ch9q34/16)	Subependymal giant-cell astrocytoma, ependymomas, glioma, ganglioneuroma, hamartoma
Turcot syndrome	AD	Mutations in *APC*[a] (ch5)	Gliomas, medulloblastomas
	AR	hMLH1 (ch3p21)	Adenomatous colon polyps, adenocarcinoma
VHL	AD	Mutations in *VHL* gene (ch3p25)	Hemangioblastomas
			Retinal angiomas, renal cell carcinoma, pheochromocytoma, pancreatic tumors and cysts, endolymphatic sac tumors of the middle ear

[a]Various DNA mismatch repair gene mutations may cause a similar clinical phenotype, also referred to as Turcot syndrome, in which there is a predisposition to nonpolyposis colon cancer and brain tumors.

Abbreviations: AD, autosomal dominant; APC, adenomatous polyposis coli; AR, autosomal recessive; ch, chromosome; NF, neurofibromatosis; PTEN, phosphatase and tensin homologue; TSC, tuberous sclerosis complex; VHL, von Hippel–Lindau.

MOLECULAR PATHOGENESIS

As with other neoplasms, brain tumors arise as a result of a multistep process driven by the sequential acquisition of genetic alterations. These include loss of tumor-suppressor genes (e.g., *p53*, cyclin-dependent kinase inhibitor 2A and 2B [*CDKN2A/B*], and phosphatase and tensin homolog on chromosome 10 [*PTEN*]) and amplification and overexpression of protooncogenes such as the epidermal growth factor receptor (*EGFR*) and the platelet-derived growth factor receptors (*PDGFR*). The accumulation of these genetic abnormalities results in uncontrolled cell growth and tumor formation.

Important progress has been made in understanding the molecular pathogenesis of several types of brain tumors, including glioblastoma and medulloblastoma, allowing them to be separated into different subtypes

with different prognoses. This has led the World Health Organization (WHO) to issue an update on the classification of CNS tumors in 2016 that for the first time incorporates molecular parameters in addition to traditional histology into the diagnosis of brain tumors.

INTRINSIC "MALIGNANT" TUMORS

DIFFUSE GIOMAS

Gliomas are the most common type of malignant primary brain tumor and are derived, based on their presumed lineage, into astrocytomas and oligodendrogliomas. These tumors are classified based on two highly recurrent molecular alterations, isocitrate dehydrogenase (*IDH*) mutations and 1p/19q codeletion, in addition to more conventional histopathologic parameters. Most lower-grade astrocytomas have *IDH* mutations but intact 1p/19q, and often mutations in *ATRX* and *p53*. Oligodendrogliomas usually have *IDH* mutations and codeletion of 1p/19q.

ASTROCYTOMAS

These are infiltrative tumors with a presumptive glial cell of origin. WHO classifies astrocytomas into four prognostic grades based on histologic features: grade I (pilocytic astrocytoma, subependymal giant cell astrocytoma); grade II (astrocytoma); grade III (anaplastic astrocytoma); and grade IV (glioblastoma). Grades I and II are considered low-grade, and grades III and IV high-grade, astrocytomas.

Low-Grade Astrocytoma

GRADE I ASTROCYTOMAS

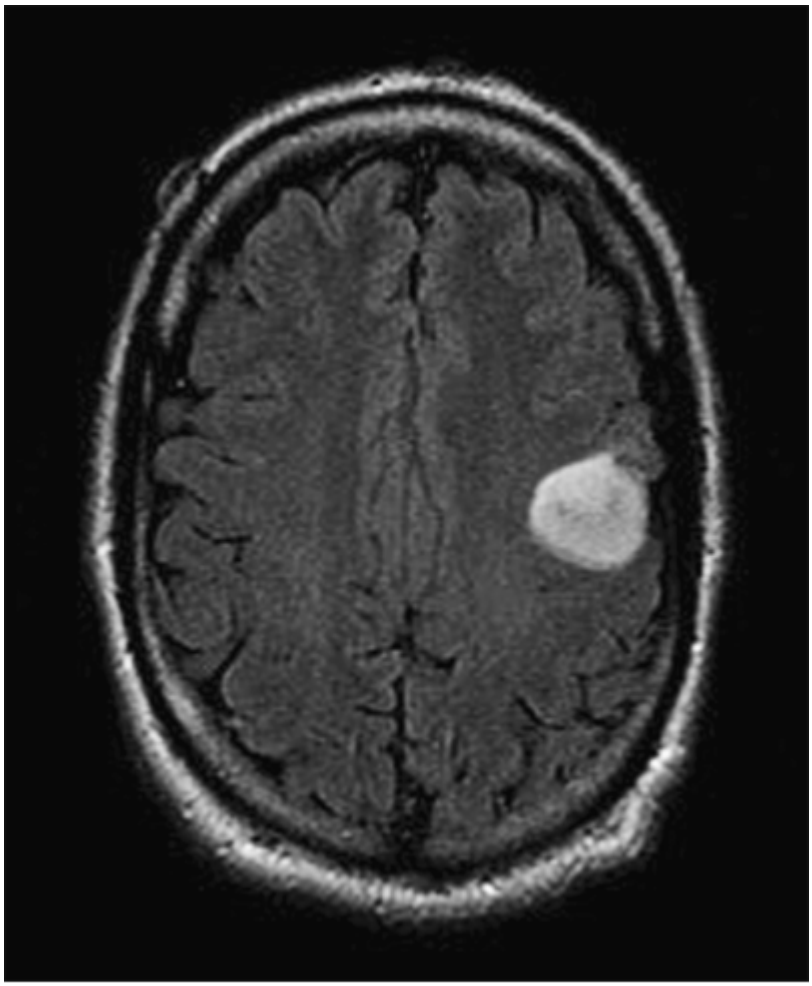
Pilocytic astrocytomas (WHO grade I) are the most common tumor of childhood. They occur typically in the cerebellum but may also be found elsewhere in the neuraxis, including the optic nerves and brainstem. Frequently they appear as cystic lesions with an enhancing mural nodule. Often they have *BRAF* fusions or mutations. These are well-demarcated lesions that are potentially curable if they can be resected completely. Giant-cell subependymal astrocytomas are usually found in the ventricular wall of patients with tuberous sclerosis. They often do not require intervention but can be treated surgically or with inhibitors of the mammalian target of rapamycin (mTOR).

GRADE II ASTROCYTOMAS

These are infiltrative tumors that usually present with seizures in young adults. They appear as nonenhancing tumors with increased T2/FLAIR signal **(Fig. 86-1)**. If feasible, patients should undergo maximal surgical resection, although complete resection is rarely possible because of the invasive nature of the tumor. In patients at higher risk for recurrence (subtotal resection or above the age of 40 years), there is evidence that radiation therapy (RT) followed by PCV (procarbazine, cyclohexylchloroethylnitrosourea [CCNU], and vincristine) chemotherapy may possibly be of benefit. The tumor transforms to a malignant astrocytoma in most patients, leading to variable survival with a median of ~5–10 years. The minority of grade II astrocytomas without *IDH* mutations have a worse prognosis.

FIGURE 86-1

Fluid-attenuated inversion recovery (FLAIR) MRI of a left frontal low-grade astrocytoma. This lesion did not enhance.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

High-Grade Astrocytoma

GRADE III (ANAPLASTIC) ASTROCYTOMA

These account for ~15–20% of high-grade astrocytomas. They generally present in the fourth and fifth decades of life as variably enhancing tumors. Treatment is the same as for glioblastoma, consisting of maximal safe surgical resection followed by RT and adjuvant temozolomide alone or RT with concurrent and adjuvant temozolomide.

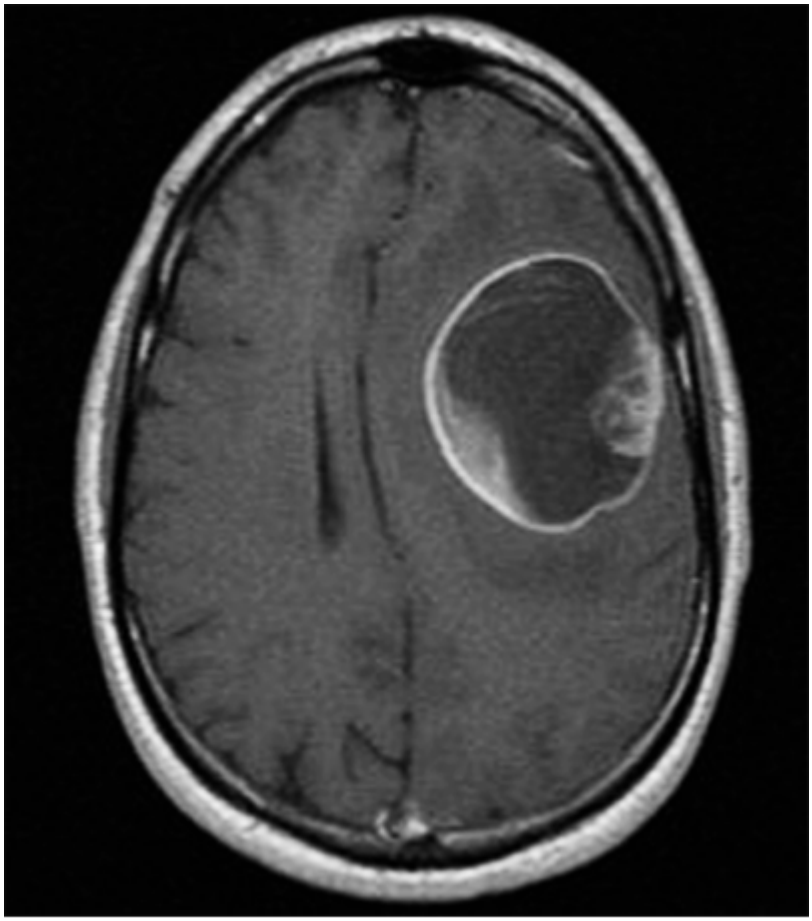
GRADE IV ASTROCYTOMA (GLIOBLASTOMA)

Glioblastoma accounts for the majority of high-grade astrocytomas. Approximately 10% of glioblastoma have *IDH* mutations. These tend to arise from lower-grade tumors (secondary glioblastomas) and have a better prognosis. They are the most common malignant primary brain tumor, with >12,000 cases diagnosed each year in the United States. Patients usually present in the sixth and seventh decades of life with headache, seizures, or focal neurologic deficits. The tumors appear as ring-enhancing masses with central necrosis and surrounding edema **(Fig. 86-2)**. These are highly infiltrative tumors, and the areas of increased

T2/FLAIR signal surrounding the main tumor mass contain invading tumor cells. Treatment involves maximal surgical resection followed by partial-field external-beam RT (6000 cGy in thirty 200-cGy fractions) with concomitant temozolomide, followed by 6–12 months of adjuvant temozolomide. With this regimen, median survival is increased to 14.6–18 months compared to only 12 months with RT alone, and 5-year survival is ~10%. Patients whose tumor contains the DNA repair enzyme O^6-methylguanine-DNA methyltransferase (MGMT) are relatively resistant to temozolomide and have a worse prognosis compared to those whose tumors contain low levels of MGMT as a result of silencing of the MGMT gene by promoter hypermethylation. Implantation of biodegradable polymers containing carmustine chemotherapy into the tumor bed after resection of the tumor, or addition of tumor treating fields (scalp electrodes delivering low intensity electric currents), produces a modest improvement in survival.

FIGURE 86-2

Postgadolinium T1 MRI of a large cystic left frontal glioblastoma.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

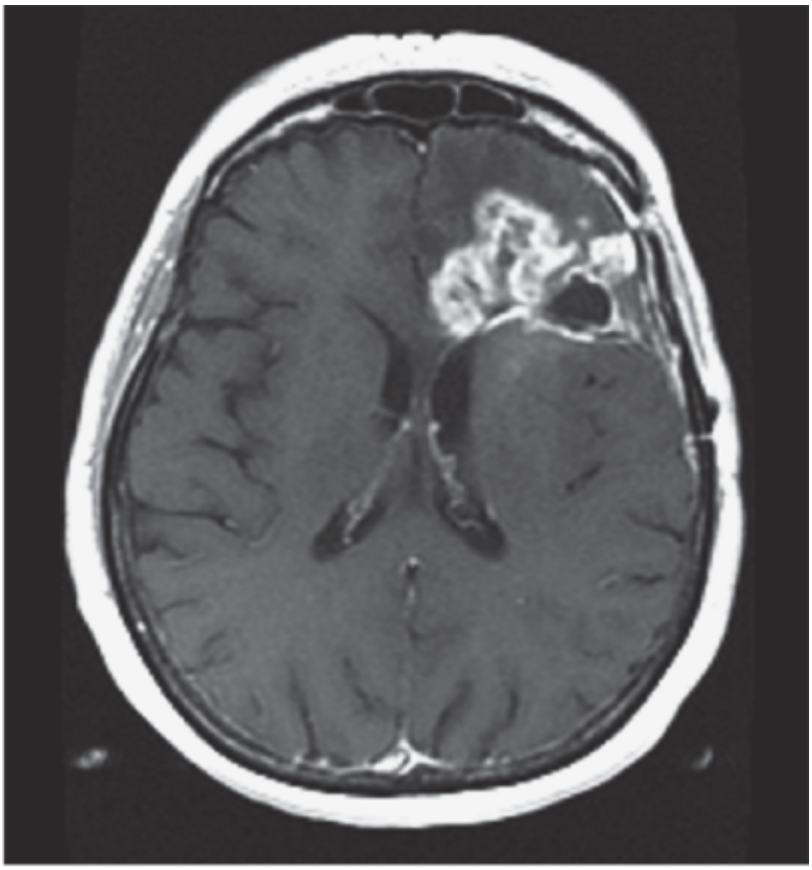
For elderly patients aged >65–70 years, a hypofractionated RT regimen of 40 Gy over 3 weeks with temozolomide is well-tolerated and likely leads to similar outcomes as the 6-week standard RT regimen.

Despite optimal therapy, glioblastomas invariably recur. Treatment options for recurrent disease may include reoperation, carmustine wafers, and alternate chemotherapeutic regimens. Reirradiation is rarely helpful. Bevacizumab, a humanized vascular endothelial growth factor (VEGF) monoclonal antibody, has activity in

recurrent glioblastoma, increasing progression-free survival but not overall survival, and reducing peritumoral edema and glucocorticoid use **(Fig. 86-3)**. Treatment decisions for patients with recurrent glioblastoma must be made on an individual basis, taking into consideration such factors as previous therapy, time to relapse, performance status, and quality of life. Whenever feasible, patients with recurrent disease should be enrolled in clinical trials. Novel therapies undergoing evaluation in patients with glioblastoma include targeted molecular agents directed at receptor tyrosine kinases and signal transduction pathways; immunotherapy; oncolytic viruses; antiangiogenic agents; chemotherapeutic agents that cross the blood-brain barrier more effectively than currently available drugs; and infusion of radiolabeled drugs and targeted toxins into the tumor and surrounding brain by means of convection-enhanced delivery.

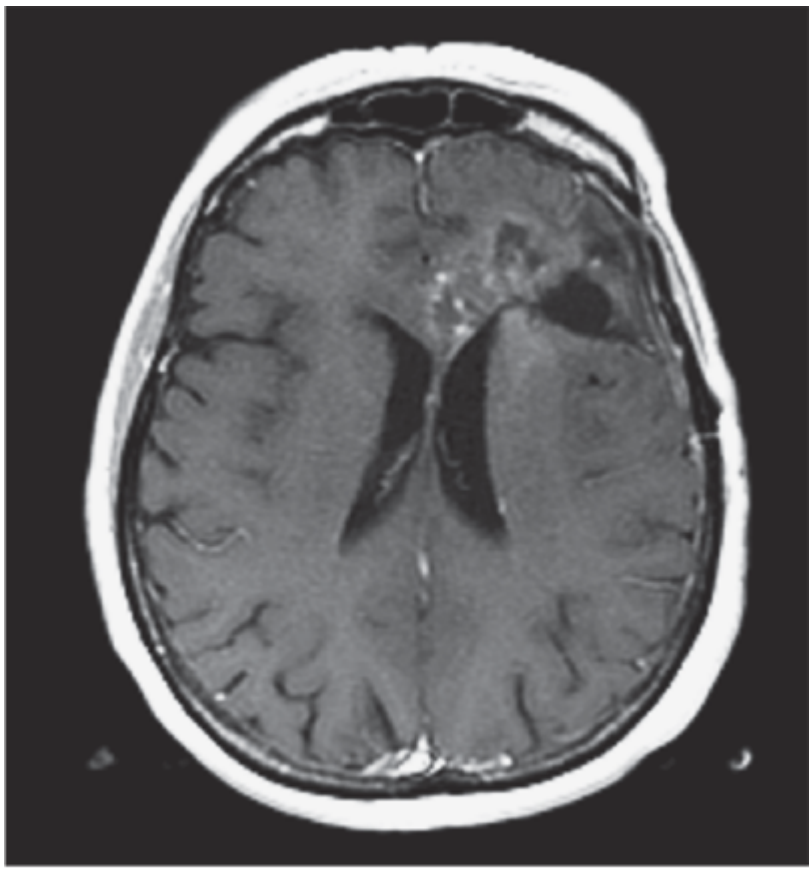
FIGURE 86-3

Postgadolinium T1 MRI of a recurrent glioblastoma before (***A***) and after (***B***) administration of bevacizumab. Note the decreased enhancement and mass effect.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The most important adverse prognostic factors in patients with glioblastomas are older age, absence of *IDH* mutations, unmethylated MGMT promoter, poor Karnofsky performance status, and unresectable tumor.

Gliosarcomas are a variant of glioblastoma containing both an astrocytic and a sarcomatous component and are treated in the same way as glioblastomas.

OLIGODENDROGLIOMA

Oligodendrogliomas account for ~15–20% of gliomas. They are characterized by codeletion of 1p/19q and usually have *IDH* mutations. Oligodendrogliomas are classified by the WHO into oligodendrogliomas (grade II) or anaplastic oligodendrogliomas (AOs) (grade III). Oligodendrogliomas have distinctive pathologic features such as perinuclear clearing—giving rise to a “fried-egg” appearance—and a reticular pattern of blood vessel growth. Some tumors have both an oligodendroglial as well as an astrocytic component. With molecular testing, it is now clear that almost all these mixed tumors (oligoastrocytomas) are genetically either astrocytomas or oligodendrogliomas. As a result, the diagnosis of oligoastrocytoma is now rarely made unless molecular testing is not available.

Grade II oligodendrogliomas are generally more responsive to therapy and have a better prognosis than pure astrocytic tumors. These tumors present similarly to grade II astrocytomas in young adults. The tumors are nonenhancing and often partially calcified. They should be treated with surgery and, in patients with residual

disease or aged >40 years, RT and chemotherapy. Patients with oligodendrogliomas have a median survival in excess of 10 years.

AOs present in the fourth and fifth decades as variably enhancing tumors. They are more responsive to therapy than grade III astrocytomas. Treatment involves maximal safe resection followed by RT and PCV or temozolomide chemotherapy. Median survival of patients with AO is in excess of 10 years.

EPENDYMOMAS

Ependymomas are tumors derived from ependymal cells that line the ventricular surface. They account for ~5% of childhood tumors and frequently arise from the wall of the fourth ventricle in the posterior fossa. Although adults can have intracranial ependymomas, they occur more commonly in the spine, especially in the filum terminale of the spinal cord where they have a myxopapillary histology. Ependymomas that can be completely resected are potentially curable. Partially resected ependymomas will recur and require irradiation. The less common anaplastic ependymoma is more aggressive and is treated with resection and RT; chemotherapy has limited efficacy. Subependymomas are slow-growing benign lesions arising in the wall of ventricles that often do not require treatment.

OTHER LESS COMMON GLIOMAS

Gangliogliomas and pleomorphic xanthoastrocytomas occur in young adults. They behave as more indolent forms of grade I gliomas and are usually treated with surgery. Frequently they will have *BRAFV600E* mutations. Brainstem gliomas usually occur in children or young adults. Despite treatment with RT and chemotherapy, the prognosis is poor, with a median survival of only 1 year.

PRIMARY CENTRAL NERVOUS SYSTEM LYMPHOMA

Primary central nervous system lymphoma (PCNSL) is a rare non-Hodgkin lymphoma accounting for <3% of primary brain tumors. For unclear reasons, its incidence is increasing, particularly in immunocompetent, older individuals.

PCNSL in immunocompetent patients is usually a diffuse large B-cell lymphoma. Immunocompromised patients, especially those infected with the human immunodeficiency virus (HIV) or organ transplant recipients, are at risk for PCNSL that is typically large cell with immunoblastic and more aggressive features. Epstein-Barr virus (EBV) plays an important role in the pathogenesis of PCNSL in this population. These patients are usually severely immunocompromised, with CD4 counts of <50/mL.

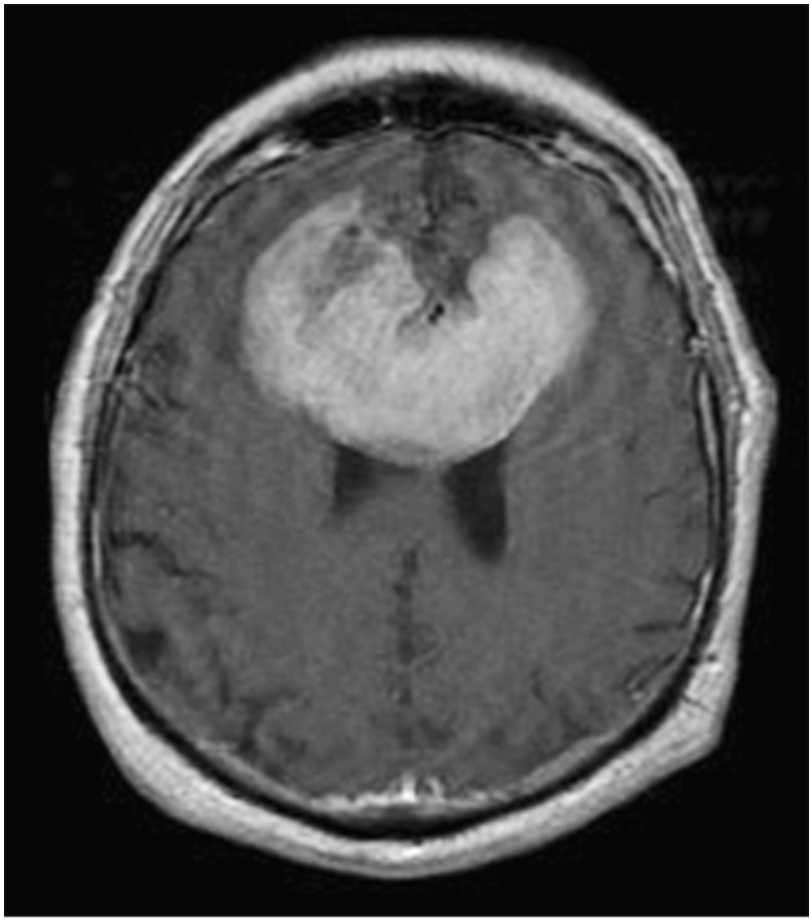
Immunocompetent patients with PCNSL are older (median 60 years) compared to those with HIV-related PCNSL (median 31 years). PCNSL usually presents as a mass lesion, with neuropsychiatric symptoms, lateralizing signs, or seizures. Ocular and leptomeningeal involvement each occur in 15–20% of patients.

On contrast-enhanced MRI, PCNSL usually appears as a densely enhancing tumor **(Fig. 86-4)**. Immunocompetent patients have solitary lesions more often than immunosuppressed patients. Frequently

there is involvement of the basal ganglia, corpus callosum, or periventricular region. Stereotactic biopsy is necessary to obtain a histologic diagnosis. Whenever possible, glucocorticoids should be withheld until after the biopsy has been obtained because they have a cytolytic effect on lymphoma cells and may lead to nondiagnostic tissue. In addition, patients should be tested for HIV, and the extent of disease should be assessed by performing PET or CT of the body, MRI of the spine, CSF analysis, and slit-lamp examination of the eye. Bone marrow biopsy and testicular ultrasound are occasionally performed.

FIGURE 86-4

Postgadolinium T1 MRI demonstrating a large bifrontal primary central nervous system lymphoma (PCNSL). The periventricular location and diffuse enhancement pattern are characteristic of lymphoma.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Primary Central Nervous System Lymphoma

PCNSL is more sensitive to glucocorticoids, chemotherapy, and RT than other primary brain tumors. Durable complete responses and long-term survival are possible with these treatments. High-dose methotrexate, a folate antagonist that interrupts DNA synthesis, produces response rates ranging from 35 to 80% and median

survival of up to 50 months. The combination of methotrexate with other chemotherapeutic agents such as cytarabine increases the response rate to 70–100%. The addition of whole-brain RT to methotrexate-based chemotherapy prolongs progression-free survival but not overall survival, but it is associated with delayed neurotoxicity, especially in patients aged >60 years. As a result, full-dose RT is frequently omitted, but there may be a role for reduced-dose RT. The anti-CD20 monoclonal antibody rituximab has activity in PCNSL and is often incorporated into the chemotherapy regimen. For some patients, high-dose chemotherapy with autologous stem cell rescue may offer the best chance of preventing relapse. At least 50% of patients will eventually develop recurrent disease. Treatment options include RT for patients who have not had prior irradiation, re-treatment with methotrexate, as well as other agents such as temozolomide, rituximab, procarbazine, topotecan, and pemetrexed. High-dose chemotherapy with autologous stem cell rescue may be appropriate in selected patients with relapsed disease.

PCNSL IN IMMUNOCOMPROMISED PATIENTS

PCNSL in immunocompromised patients often produces multiple ring-enhancing lesions that can be difficult to differentiate from metastases or infections such as toxoplasmosis. The diagnosis is usually established by examination of the CSF for cytology and EBV DNA, toxoplasmosis serologic testing, brain PET imaging for hypermetabolism of the lesions which, although nonspecific, can be consistent with tumor, and, if necessary, brain biopsy. Since the advent of highly active antiretroviral drugs, the incidence of HIV-related PCNSL has declined. These patients are preferably treated with high-dose methotrexate-based regimens and initiation of highly active antiretroviral therapy; whole-brain RT is reserved for those who cannot tolerate systemic chemotherapy. In organ transplant recipients, reduction of immunosuppression may improve outcome.

MEDULLOBLASTOMAS

Medulloblastomas are the most common malignant brain tumor of childhood, accounting for ~20% of all primary CNS tumors among children. They arise from granule cell progenitors or from multipotent progenitors from the ventricular zone. Approximately 5% of children with medulloblastoma have an inherited syndrome, such as Gorlin, Turcot, or Li-Fraumeni, which predisposes to the development of medulloblastoma. Histologically, medulloblastomas are highly cellular tumors with abundant dark staining, round nuclei, and rosette formation (Homer-Wright rosettes). In the 2016 WHO pathologic classification, they have been divided into four molecular subgroups: (1) WNT-activated (primarily affects children and has the best outcome); (2) SHH-activated (affects adults, infants, and children with the younger patients having the better outcome and adults doing poorly); (3) non-WNT/non-SHH, group 3 (frequently has disseminated CNS disease at diagnosis and has the worst outcome); and (4) non-WNT/non-SHH, group 4 (30% have metastases at diagnosis, but 5-year progression-free survival is 95%). Regardless of subtype, patients present with headache, ataxia, and signs of brainstem involvement. On MRI they appear as densely enhancing tumors in the posterior fossa, sometimes associated with hydrocephalus. Treatment involves maximal surgical resection, craniospinal irradiation, and chemotherapy with agents such as cisplatin, lomustine, cyclophosphamide, and vincristine. Approximately 70% of patients overall have long-term survival but usually at the cost of significant neurocognitive impairment. A major goal of current research is to improve

survival while minimizing long-term complications, and clinical trials are now being designed for specific molecular subgroups.

PINEAL REGION TUMORS

A large number of tumors can arise in the region of the pineal gland. These typically present with headache, visual symptoms, and hydrocephalus. Patients may have Parinaud's syndrome characterized by impaired upgaze and accommodation. Some pineal tumors such as pineocytomas and benign teratomas can be treated by surgical resection. Germinomas respond to irradiation, whereas pineoblastomas and nongerminomatous germ cell tumors require craniospinal radiation and chemotherapy.

EXTRINSIC "BENIGN" TUMORS

MENINGIOMAS

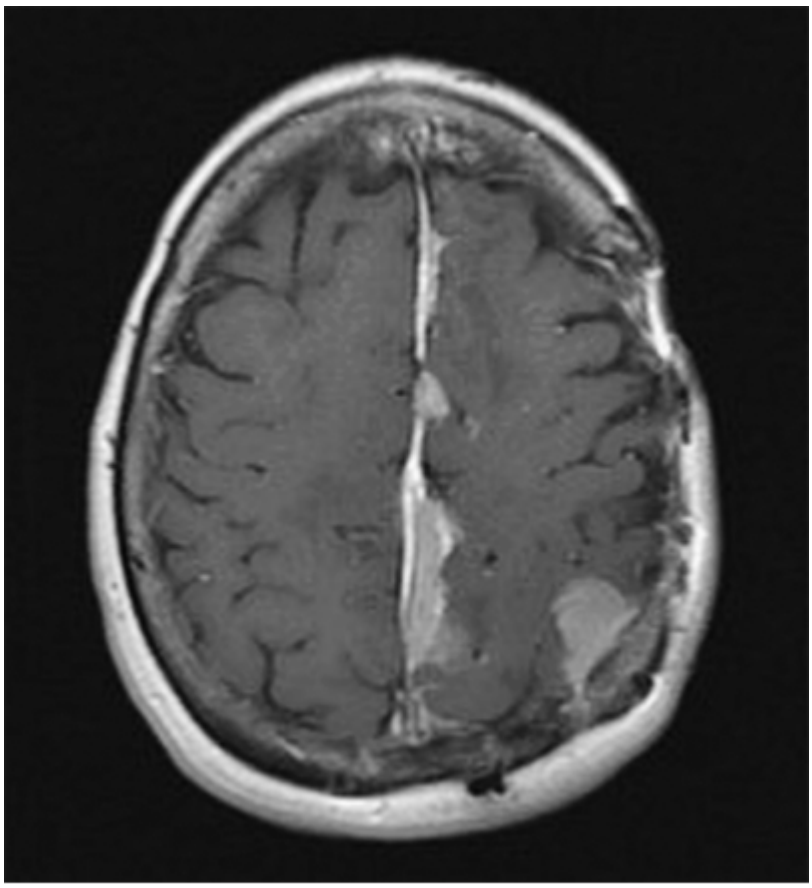
Meningiomas are diagnosed with increasing frequency as more people undergo neuroimaging for various indications. They are now the most common primary brain tumor, accounting for ~35% of the total. Their incidence increases with age. They tend to be more common in women and in patients with neurofibromatosis type 2 (NF2). They also occur more commonly in patients with a past history of cranial irradiation.

Meningiomas arise from the dura mater and are composed of neoplastic meningothelial (arachnoidal cap) cells. They are most commonly located over the cerebral convexities, especially adjacent to the sagittal sinus, but they can also occur in the skull base and along the dorsum of the spinal cord. Meningiomas are classified by the WHO into three histologic grades of increasing aggressiveness: grade I (benign), grade II (atypical), and grade III (malignant).

Many meningiomas are found incidentally following neuroimaging for unrelated reasons. They can also present with headaches, seizures, or focal neurologic deficits. On imaging studies they have a characteristic appearance usually of a densely enhancing extra-axial tumor arising from the dura **(Fig. 86-5)**. Typically they have a dural tail, consisting of thickened, enhanced dura extending like a tail from the mass. The main differential diagnosis of meningioma is a dural metastasis.

FIGURE 86-5

Postgadolinium T1 MRI demonstrating multiple meningiomas along the falx and left parietal cortex.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

If the meningioma is small and asymptomatic, no intervention is necessary and the lesion can be observed with serial MRI studies. Larger, symptomatic lesions should be resected. If complete resection is achieved, the patient is cured. Incompletely resected tumors tend to recur, although the rate of recurrence can be very slow with grade I tumors. Tumors that cannot be resected, or can only be partially removed, may benefit from external-beam RT or stereotactic radiosurgery (SRS). These treatments may also be helpful in patients whose tumor has recurred after surgery. Hormonal therapy and chemotherapy are currently unproven.

Rarer tumors that resemble meningiomas include hemangiopericytomas and solitary fibrous tumors. Since they share similar molecular alterations, the 2016 WHO classification introduced the combined term solitary fibrous tumor/hemangiopericytoma for this entity. These tumors are treated with surgery and RT but have a higher propensity to recur locally or metastasize systemically.

SCHWANNOMAS

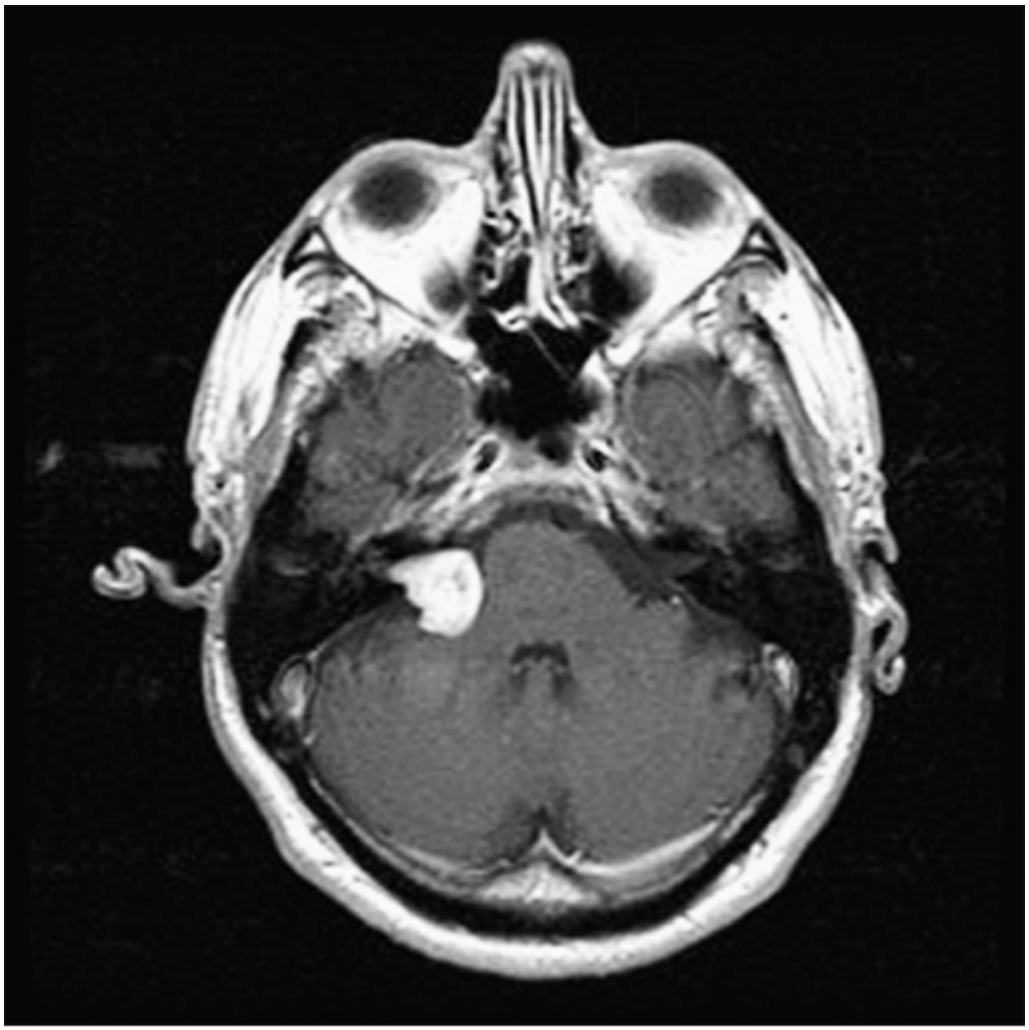
These are generally benign tumors arising from the Schwann cells of cranial and spinal nerve roots. The most common schwannomas, termed *vestibular schwannomas* or *acoustic neuromas*, arise from the vestibular portion of the eighth cranial nerve and account for ~9% of primary brain tumors. Patients with NF2 have a high incidence of vestibular schwannomas that are frequently bilateral. Schwannomas arising from other cranial nerves, such as the trigeminal nerve (cranial nerve V), occur with much lower frequency.

Neurofibromatosis type 1 (NF1) is associated with an increased incidence of schwannomas of the spinal nerve roots.

Vestibular schwannomas may be found incidentally on neuroimaging or present with progressive unilateral hearing loss, dizziness, tinnitus, or, less commonly, symptoms resulting from compression of the brainstem and cerebellum. On MRI they appear as densely enhancing lesions, enlarging the internal auditory canal and often extending into the cerebellopontine angle **(Fig. 86-6)**. The differential diagnosis includes meningioma. Very small, asymptomatic lesions can be observed with serial MRIs. Larger lesions should be treated with surgery or SRS. The optimal treatment will depend on the size of the tumor, symptoms, and the patient's preference. In patients with small vestibular schwannomas and relatively intact hearing, early surgical intervention increases the chance of preserving hearing.

FIGURE 86-6

Postgadolinium MRI of a right vestibular schwannoma. The tumor can be seen to involve the internal auditory canal.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PITUITARY TUMORS

These are discussed in detail in **Chap. 373**.

CRANIOPHARYNGIOMAS

Craniopharyngiomas are rare, usually suprasellar, partially calcified, solid, or mixed solid-cystic benign tumors that arise from remnants of Rathke's pouch. They have a bimodal distribution, occurring predominantly in children but also between the ages of 55 and 65 years. They present with headaches, visual impairment, and impaired growth in children and hypopituitarism in adults. Treatment involves surgery, RT, or a combination of the two.

OTHER BENIGN TUMORS

Dysembryoplastic Neuroepithelial Tumors (DNTs)

These are benign, supratentorial tumors, usually in the temporal lobe. They typically occur in children and young adults with a long-standing history of seizures. Surgical resection is curative.

Epidermoid Cysts

These consist of squamous epithelium surrounding a keratin-filled cyst. They are usually found in the cerebellopontine angle and the intrasellar and suprasellar regions. They may present with headaches, cranial nerve abnormalities, seizures, or hydrocephalus. MRI demonstrates an extra-axial lesion with characteristics that are similar to CSF but have restricted diffusion. Treatment involves surgical resection.

Dermoid Cysts

Like epidermoid cysts, dermoid cysts arise from epithelial cells that are retained during closure of the neural tube. They contain both epidermal and dermal structures such as hair follicles, sweat glands, and sebaceous glands. Unlike epidermoid cysts, these tumors usually have a midline location. They occur most frequently in the posterior fossa, especially the vermis, fourth ventricle, and suprasellar cistern. On MRI, dermoid cysts resemble lipomas, demonstrating T1 hyperintensity and variable signal on T2. Symptomatic dermoid cysts can be treated with surgery.

Colloid Cysts

These usually arise in the anterior third ventricle and may present with headaches, hydrocephalus, and, very rarely, sudden death. Surgical resection is curative, or a third ventriculostomy may relieve the obstructive hydrocephalus and be sufficient therapy.

NEUROCUTANEOUS SYNDROMES (PHAKOMATOSES)

A number of genetic disorders are characterized by cutaneous lesions and an increased risk of brain tumors. Most of these disorders have an autosomal dominant inheritance with variable penetrance.

NEUROFIBROMATOSIS TYPE 1 (NF1) (von RECKLINGHAUSEN'S DISEASE)

NF1 is an autosomal dominant disorder with variable penetrance and an incidence of ~1 in 2600–3000. Approximately one-half of cases are familial; the remainder are caused by new mutations arising in patients with unaffected parents. The *NF1* gene on chromosome 17q11.2 encodes neurofibromin, a guanosine triphosphatase (GTPase)-activating protein (GAP) that modulates signaling through the RAS pathway. Mutations of *NF1* result in a large number of nervous system tumors including neurofibromas, plexiform neurofibromas, optic nerve gliomas, astrocytomas, and meningiomas. In addition to neurofibromas, which appear as multiple, soft, rubbery cutaneous tumors, other cutaneous manifestations of NF1 include café-au-lait spots and axillary freckling. NF1 is also associated with hamartomas of the iris termed Lisch nodules, pheochromocytomas, pseudoarthrosis of the tibia, scoliosis, epilepsy, and mental retardation.

NEUROFIBROMATOSIS TYPE 2 (NF2)

NF2 is less common than NF1, with an incidence of 1 in 25,000–40,000. It is an autosomal dominant disorder with full penetrance. As with NF1, approximately one-half of cases arise from new mutations. The *NF2* gene on 22q encodes a cytoskeletal protein, merlin (moesin, ezrin, radixin-like protein) that functions as a tumor suppressor. NF2 is characterized by bilateral vestibular schwannomas in >90% of patients, multiple meningiomas, and spinal ependymomas and astrocytomas. Treatment of bilateral vestibular schwannomas can be challenging because the goal is to preserve hearing for as long as possible. These patients may also have diffuse schwannomatosis that may affect the cranial, spinal, or peripheral nerves; posterior subcapsular lens opacities; and retinal hamartomas.

TUBEROUS SCLEROSIS (BOURNEVILLE DISEASE)

This is an autosomal dominant disorder with an incidence of ~1 in 5000–10,000 live births. It is caused by mutations in either the *TSC1* gene, which maps to chromosome 9q34 and encodes a protein termed hamartin, or the *TSC2* gene, which maps to chromosome 16p13.3 and encodes the protein tuberin. Hamartin forms a complex with tuberin, which inhibits cellular signaling through mTOR, and acts as a negative regulator of the cell cycle. Patients with tuberous sclerosis may have seizures, mental retardation, adenoma sebaceum (facial angiofibromas), shagreen patch, hypomelanotic macules, periungual fibromas, renal angiomyolipomas, and cardiac rhabdomyomas. These patients have an increased incidence of subependymal nodules, cortical tubers, and subependymal giant-cell astrocytomas (SEGAs). Patients frequently require anticonvulsants for seizures. SEGAs do not always require therapeutic intervention, but the most effective therapy is with the mTOR inhibitors sirolimus or everolimus, which often decrease seizures as well as SEGA size.

TUMORS METASTATIC TO THE BRAIN

Brain metastases arise from hematogenous spread and frequently originate from a lung primary or are associated with pulmonary metastases. Most metastases develop at the gray matter–white matter junction in

the watershed distribution of the brain where intravascular tumor cells lodge in terminal arterioles. The distribution of metastases in the brain approximates the proportion of blood flow such that ~85% of all metastases are supratentorial and 15% occur in the posterior fossa. The most common sources of brain metastases are lung and breast carcinomas; melanoma has the greatest propensity to metastasize to the brain, being found in 80% of patients at autopsy (Table 86-3). Other tumor types such as ovarian and esophageal carcinoma rarely metastasize to the brain. Prostate and breast cancers also have a propensity to metastasize to the dura and can mimic meningioma. Leptomeningeal metastases are common from hematologic malignancies and also breast and lung cancers. Spinal cord compression primarily arises in patients with prostate and breast cancer, tumors with a strong propensity to metastasize to the axial skeleton.

TABLE 86-3

Frequency of Nervous System Metastases by Common Primary Tumors

	Brain (%)	LM (%)	ESCC (%)
Lung	41	17	15
Breast	19	57	22
Melanoma	10	12	4
Prostate	1	1	10
GIT	7	—	5
Renal	3	2	7
Lymphoma	<1	10	10
Sarcoma	7	1	9
Other	11	—	18

Abbreviations: ESCC, epidural spinal cord compression; GIT, gastrointestinal tract; LM, leptomeningeal metastases.

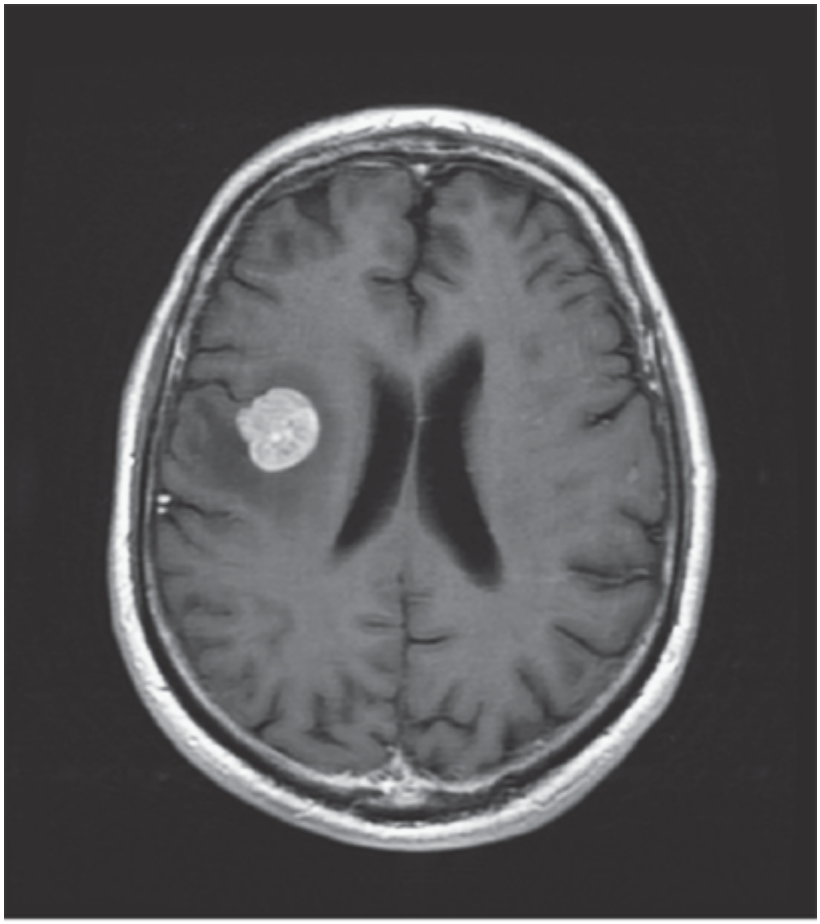
DIAGNOSIS OF METASTASES

Brain metastases are best visualized on MRI, where they usually appear as well-circumscribed lesions (Fig. 86-7). The amount of perilesional edema can be highly variable, with large lesions causing minimal edema

and sometimes very small lesions causing extensive edema. Enhancement may be in a ring pattern or diffuse. Occasionally, intracranial metastases will hemorrhage; although melanoma, thyroid, and kidney cancer have the greatest propensity to hemorrhage, the most common cause of a hemorrhagic metastasis is lung cancer because it accounts for the majority of brain metastases. The radiographic appearance of brain metastasis is nonspecific, and similar-appearing lesions can occur with infection including brain abscesses and also with demyelinating lesions, sarcoidosis, radiation necrosis in a previously treated patient, or a primary brain tumor that may be a second malignancy in a patient with systemic cancer. Biopsy is rarely necessary for diagnosis because imaging alone in the appropriate clinical situation usually suffices. However, in ~10% of patients, a systemic cancer may present with a brain metastasis, and if there is not an easily accessible systemic site to biopsy, a brain lesion must be removed for diagnostic purposes.

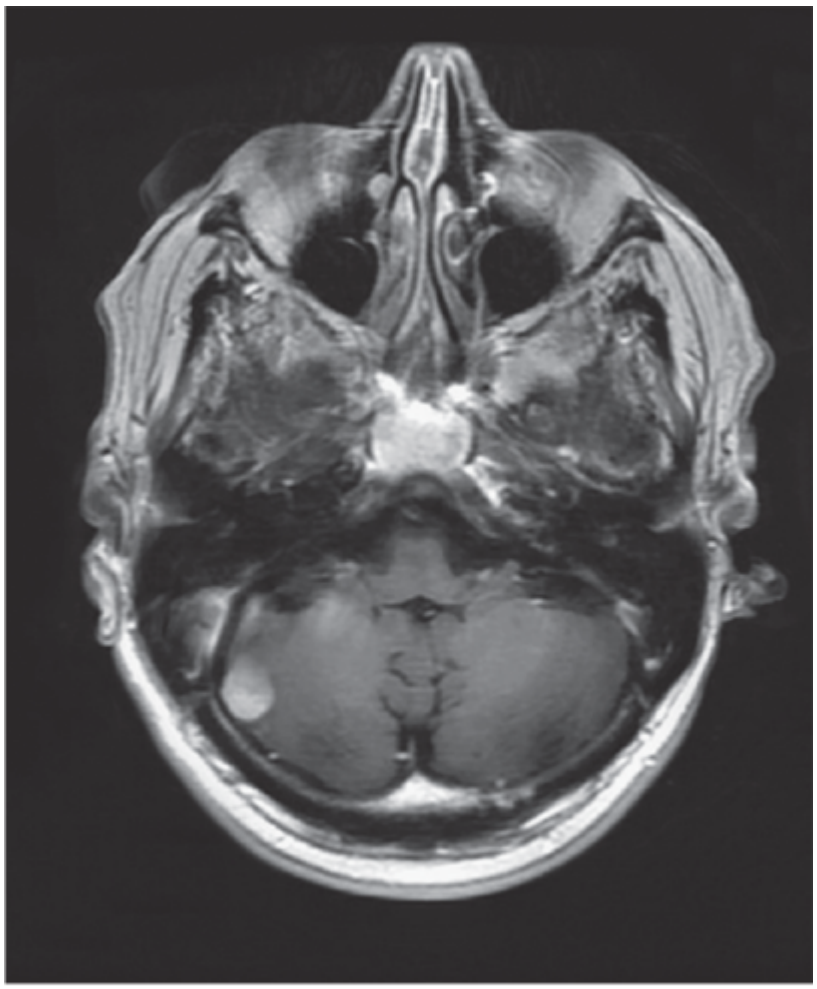
FIGURE 86-7

Postgadolinium T1 MRI of multiple brain metastases from non-small-cell lung cancer involving the right frontal (***A***) and right cerebellar (***B***) hemispheres. Note the diffuse enhancement pattern and absence of central necrosis.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Tumors Metastatic to the Brain

DEFINITIVE TREATMENT

The number and location of brain metastases often determine the therapeutic options. The patient's overall condition and current or potential control of systemic disease are also major determinants. Brain metastases are single in approximately one-half of patients and multiple in the other half.

RADIATION THERAPY

The standard treatment for brain metastases has previously been whole-brain radiotherapy (WBRT) usually administered to a total dose of 3000 cGy in 10 fractions. This affords rapid palliation, and ~80% of patients improve with glucocorticoids and RT. However, it is not curative, is associated with neurocognitive toxicity, and produces median survival of only 4–6 months. If feasible, SRS has become the primary radiation oncology approach to brain metastases. It can be delivered through a variety of equally effective techniques

including the gamma knife, linear accelerator, proton beam, or CyberKnife, all of which can deliver highly focused doses of RT, usually in a single fraction. SRS can effectively sterilize the visible lesions and afford local disease control in 80–90% of patients. Some patients have been cured of their brain metastases using SRS, whereas this is distinctly rare with WBRT. Traditionally SRS was used only for patients with 1–3 metastases, but recent data suggest that SRS can effectively treat up to 10 lesions. It is, however, confined to lesions of ≤3 cm and is most effective in metastases of ≤1 cm. The addition of WBRT to SRS improves disease control in the nervous system but does not prolong survival and thus is rarely employed.

SURGERY

Randomized controlled trials have demonstrated that surgical extirpation of a single brain metastasis followed by WBRT is superior to WBRT alone. Removal of two lesions or a single symptomatic mass, particularly if compressing the ventricular system, can also be useful. This is particularly important in patients who have highly radioresistant lesions such as renal carcinoma. Surgical resection can produce rapid amelioration of symptoms, improve control of edema, and result in prolonged survival. WBRT administered after complete resection of a brain metastasis improves disease control but does not prolong survival. Some centers administer focal RT or even SRS to a resected cavity, especially if there is concern that tumor has been left behind.

CHEMOTHERAPY

Chemotherapy is becoming increasingly useful for brain metastases. Metastases from tumor types that are highly chemosensitive, such as germ cell tumors or small-cell lung cancer, may respond to chemotherapeutic regimens chosen according to the underlying malignancy. Increasingly, there are data demonstrating responsiveness of brain metastases to chemotherapy including targeted therapy, such as for patients with lung cancer harboring EGFR mutations that sensitize them to EGFR inhibitors. Immunotherapy may also be effective against those primary tumors that are sensitive to this approach, such as melanoma. Antiangiogenic agents such as bevacizumab are also effective in the treatment of CNS metastases in those primary tumors for which it is approved.

LEPTOMENINGEAL METASTASES

Leptomeningeal metastases are also described as carcinomatous meningitis, meningeal carcinomatosis, or in the case of specific tumors, leukemic or lymphomatous meningitis. Among the hematologic malignancies, acute leukemias most commonly metastasize to the subarachnoid space, followed in frequency by aggressive diffuse lymphomas. Among solid tumors, breast and lung carcinomas and melanoma most frequently spread in this fashion. Tumor cells reach the subarachnoid space via the arterial circulation or occasionally through retrograde flow in venous systems that drain metastases along the bony spine or cranium. In addition, leptomeningeal metastases may develop as a direct consequence of prior brain metastases and occur in almost 40% of patients who have a metastasis resected from the cerebellum.

CLINICAL FEATURES

Leptomeningeal metastases are characterized by multilevel symptoms and signs along the neuraxis. Combinations of lumbar and cervical radiculopathies, cranial neuropathies, seizures, confusion, and encephalopathy from hydrocephalus or raised intracranial pressure can be present. Focal deficits such as hemiparesis or aphasia are rarely due to leptomeningeal metastases unless there is direct brain infiltration. New-onset limb pain in patients with breast cancer, lung cancer, or melanoma should prompt consideration of leptomeningeal spread.

LABORATORY AND IMAGING DIAGNOSIS

Leptomeningeal metastases are particularly challenging to diagnose because identification of tumor cells in the subarachnoid compartment may be elusive. MRI can be definitive when there are clear tumor nodules adherent to the cauda equina or spinal cord, enhancing cranial nerves, or subarachnoid enhancement on brain imaging **(Fig. 86-8)**. Imaging is diagnostic in ~75% of patients and is more often positive in patients with solid tumors. Demonstration of tumor cells in the CSF is definitive and often considered the gold standard. However, CSF cytologic examination is positive in only 50% of patients on the first lumbar puncture and still misses 10% after three CSF samples. New technologies, such as rare cell capture, enhance identification of tumor cells in the CSF. CSF cytologic examination is most useful in hematologic malignancies, especially when combined with flow cytometry to identify a clonal population. Accompanying CSF abnormalities include an elevated protein concentration and an elevated white count; hypoglycorrhachia is noted in <25% of patients but is useful when present. Identification of tumor markers may be helpful in some solid tumors.

FIGURE 86-8

Postgadolinium MRI images of extensive leptomeningeal metastases from breast cancer. Nodules along the dorsal surface of the spinal cord (***A***) and cauda equina (***B***) are seen.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Leptomeningeal Metastases

The treatment of leptomeningeal metastasis is palliative because there is no curative therapy. RT to the symptomatically involved areas, such as skull base for cranial neuropathy, can relieve pain and sometimes improve function. Whole-neuraxis RT is avoided because it has significant toxicity with myelosuppression and gastrointestinal irritation as well as limited effectiveness. Systemic chemotherapy with agents that can penetrate the blood-CSF barrier may be helpful. Alternatively, intrathecal chemotherapy can be effective, particularly in hematologic malignancies. This is optimally delivered through an intraventricular cannula (Ommaya reservoir) rather than by lumbar puncture. Few drugs can be delivered safely into the subarachnoid space, and they have a limited spectrum of antitumor activity, perhaps accounting for the relatively poor response to this approach. In addition, impaired CSF flow dynamics can compromise intrathecal drug delivery. Surgery has a limited role in leptomeningeal metastasis; a ventriculoperitoneal

shunt can relieve raised intracranial pressure; however, it compromises delivery of chemotherapy into the CSF.

EPIDURAL METASTASIS

Epidural metastasis occurs in 3–5% of patients with a systemic malignancy and causes neurologic compromise by compressing the spinal cord or cauda equina. The most common cancers that metastasize to the epidural space are those malignancies that spread to bone, such as breast and prostate. Lymphoma can cause bone involvement and compression, but it can also invade an intervertebral foramen and cause spinal cord compression without bone destruction. The thoracic spine is affected most commonly, followed by the lumbar and then cervical spine.

CLINICAL FEATURES

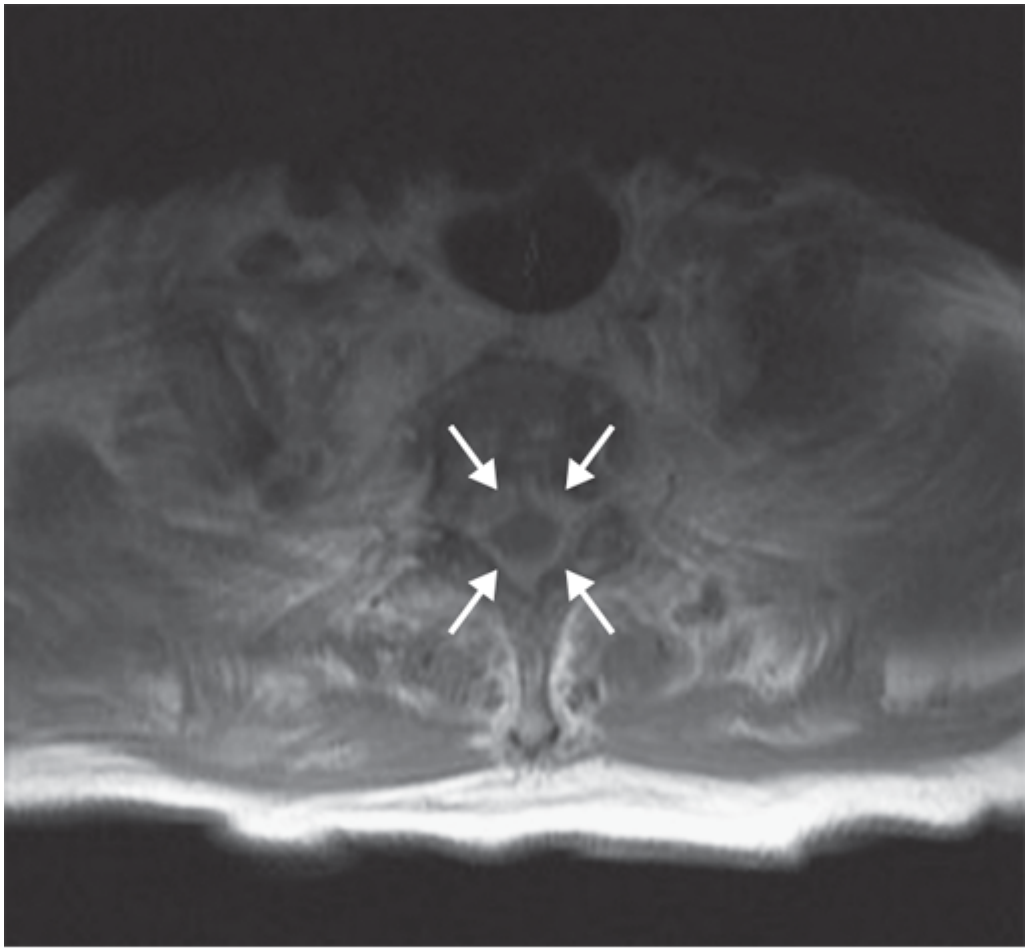
Back pain is the presenting symptom of epidural metastasis in virtually all patients; the pain may precede neurologic findings by weeks or months. The pain is usually exacerbated by lying down; by contrast, arthritic pain is often relieved by recumbency. Leg weakness is seen in ~50% of patients, as is sensory dysfunction. Sphincter problems are present in ~25% of patients at diagnosis.

DIAGNOSIS

Diagnosis is established by imaging, preferably with an MRI of the entire spine (Fig. 86-9). Contrast is not required to identify bony or epidural lesions. Any patient with cancer who has severe back pain should undergo an MRI. Plain films, bone scans, or even CT scans may show bone metastases, but only MRI can reliably delineate epidural tumor. For patients unable to have an MRI, CT myelography should be performed to outline the epidural space. The differential diagnosis of epidural tumor includes epidural abscess, acute or chronic hematomas, epidural lipomatosis and rarely, extramedullary hematopoiesis.

FIGURE 86-9

Postgadolinium T1 MRI showing circumferential epidural tumor around the thoracic spinal cord from esophageal cancer.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Epidural Metastasis

Epidural metastasis requires immediate treatment. A randomized controlled trial demonstrated the superiority of surgical resection followed by RT compared to RT alone. However, patients must be able to tolerate surgery, and the surgical procedure of choice is a complete removal of the mass, which is typically anterior to the spinal canal, necessitating an extensive approach and resection. Otherwise, RT is the mainstay of treatment and can be used for patients with radiosensitive tumors, such as lymphoma, or for those unable to undergo surgery. SRS is increasingly being used, especially for radioresistant tumor types or for re-irradiation. Chemotherapy is rarely used for epidural metastasis unless the patient has minimal to no neurologic deficit and a highly chemosensitive tumor such as lymphoma or germinoma. Patients generally fare well if treated before there is a severe neurologic deficit. Recovery from paraparesis is better after surgery than with RT alone, but survival is often short due to widespread metastatic tumor.

NEUROLOGIC TOXICITY OF THERAPY

TOXICITY FROM RADIOTHERAPY

RT can cause a variety of toxicities in the CNS. These are usually described based on their relationship in time to the administration of RT: acute (occurring within days of RT), early delayed (months), or late delayed (years). In general, the acute and early delayed syndromes resolve and do not result in persistent deficits, whereas the late delayed toxicities are usually permanent and sometimes progressive.

Acute Toxicity

Acute cerebral toxicity may occur during the course of RT to the brain. RT can cause a transient disruption of the blood-brain barrier, resulting in edema and elevated intracranial pressure. This is usually manifest as headache, lethargy, nausea, and vomiting, and can be both prevented and treated with the administration of glucocorticoids. There is no acute RT toxicity that affects the spinal cord.

Early Delayed Toxicity

Early delayed toxicity is usually apparent weeks to months after completion of cranial irradiation and is likely due to focal demyelination. Clinically it may be asymptomatic or take the form of worsening or reappearance of a preexisting neurologic deficit. At times a contrast-enhancing lesion can be seen on MRI/CT that can mimic the tumor for which the patient received the RT. For patients with a malignant glioma, this has been described as “pseudoprogression” because it mimics tumor recurrence on MRI but actually represents inflammation and necrotic debris engendered by effective therapy. This is seen with increased frequency when chemotherapy, particularly temozolomide, is given concurrently with RT. Pseudoprogression can resolve on its own or, if very symptomatic, may require resection.

In the spinal cord, early delayed RT toxicity is manifest as a Lhermitte symptom with paresthesias of the limbs or along the spine when the patient flexes the neck. Although frightening, it is benign, resolves on its own, and does not portend more serious problems.

Late Delayed Toxicity

Late delayed toxicities are the most serious because they are often irreversible and cause severe neurologic deficits. In the brain, late toxicities can take several forms, the most common of which include radiation necrosis and leukoencephalopathy. Radiation necrosis is a focal mass of necrotic tissue that is contrast enhancing on CT/MRI and may be associated with significant edema. This may appear identical to pseudoprogression but is seen months to years after RT and is always symptomatic. Clinical symptoms and signs include seizures and findings referable to the location of the necrotic mass. The necrosis is caused by the effect of RT on cerebral vasculature with fibrinoid necrosis and occlusion of blood vessels. It can mimic tumor radiographically, but unlike tumor it is typically hypometabolic on a PET scan and has reduced perfusion on perfusion MR sequences. It may require resection for diagnosis and treatment unless it can be managed with glucocorticoids. There are reports of improvement with hyperbaric oxygen or bevacizumab, but symptomatic benefit does not always accompany radiographic improvement.

Leukoencephalopathy is seen most commonly after WBRT as opposed to focal RT. On T2 or FLAIR MR sequences, there is diffusely increased signal seen throughout the hemispheric white matter, often bilaterally and symmetrically. There tends to be a periventricular predominance that may be associated with atrophy and ventricular enlargement. Clinically, patients develop cognitive impairment, a gait disorder, and later urinary incontinence, all of which can progress over time. These symptoms mimic those of normal pressure hydrocephalus, and placement of a ventriculoperitoneal shunt can improve function in some patients but does not reverse the deficits completely. Increased age is a risk factor for leukoencephalopathy but not for radiation necrosis. Necrosis appears to depend on an as yet unidentified predisposition.

Other late neurologic toxicities include endocrine dysfunction if the pituitary or hypothalamus was included in the RT port. An RT-induced neoplasm can occur many years after therapeutic RT for either a prior CNS or a head and neck tumor; accurate diagnosis requires surgical resection or biopsy. In addition, RT causes accelerated atherosclerosis, which can cause stroke either from intracranial vascular disease or carotid plaque from neck irradiation.

The peripheral nervous system is relatively resistant to RT toxicities. Peripheral nerves are rarely affected by RT, but the plexus is more vulnerable. Plexopathy develops more commonly in the brachial than in the lumbosacral distribution. It must be differentiated from tumor progression in the plexus, which is usually visualized by CT/MRI or PET scan demonstrating tumor infiltrating the region. Clinically, tumor progression is usually painful, whereas RT-induced plexopathy is painless. Radiation plexopathy is also more commonly associated with lymphedema of the affected limb. Sensory loss and weakness are seen in both.

TOXICITY FROM CHEMOTHERAPY

Neurotoxicity is second to myelosuppression as the dose-limiting toxicity of chemotherapeutic agents **(Table 86-4)**. Chemotherapy causes peripheral neuropathy from a number of commonly used agents, and the type of neuropathy can vary depending on the drug. Vincristine causes paresthesias but little sensory loss and is associated with motor dysfunction, autonomic impairment (frequently ileus), and, rarely, cranial nerve compromise. Cisplatin causes large fiber sensory loss resulting in sensory ataxia but little cutaneous sensory loss and no weakness. The taxanes also cause a predominately sensory neuropathy. Agents such as bortezomib and thalidomide also cause neuropathy.

TABLE 86-4

Neurologic Signs Caused by Agents Commonly Used in Patients with Cancer

Acute encephalopathy (delirium)	Seizures
Methotrexate (high-dose IV, IT)	Methotrexate
Cisplatin	Etoposide (high-dose)
Vincristine	Cisplatin
Asparaginase	Vincristine
Procarbazine	Asparaginase
5-Fluorouracil (± levamisole)	Nitrogen mustard
Cytarabine (high-dose)	Carmustine
Nitrosoureas (high-dose or arterial)	Dacarbazine (intraarterial or high-dose)
Ifosfamide	Busulfan (high-dose)
Etoposide (high-dose)	Myelopathy (IT drugs)
Bevacizumab (PRES)	Methotrexate
Chronic encephalopathy (dementia)	Cytarabine
Methotrexate	Thiotepa
Carmustine	Peripheral neuropathy
Cytarabine	Vinca alkaloids
Fludarabine	Cisplatin
Visual loss	Procarbazine
Tamoxifen	Etoposide
Gallium nitrate	Teniposide
Cisplatin	Cytarabine
Fludarabine	Taxanes
Cerebellar dysfunction/ataxia	Suramin
5-Fluorouracil (± levamisole)	Bortezomib
Cytarabine	
Procarbazine	

Abbreviations: IT, intrathecal; IV, intravenous; PRES, posterior reversible encephalopathy syndrome.

Encephalopathy and seizures are common toxicities from chemotherapeutic drugs. Ifosfamide can cause a severe encephalopathy, which is reversible with discontinuation of the drug and the use of methylene blue for severely affected patients. Fludarabine also causes a severe global encephalopathy that may be permanent. Bevacizumab and other anti-VEGF agents can cause posterior reversible encephalopathy syndrome. Cisplatin can cause hearing loss and less frequently vestibular dysfunction. Immunotherapy with monoclonal antibodies such as ipilimumab or nivolumab can cause an autoimmune hypophysitis, Guillain-Barré syndrome, or an autoimmune encephalitis.

FURTHER READING

Buckner JC et al: Radiation plus procarbazine, CCNU, and Vincristine in low-grade glioma. N Engl J Med 374:1344, 2016. [PubMed: 27050206]

Cancer Genome Atlas Research Network et al: Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med 372:2481, 2015. [PubMed: 26061751]

Louis DN et al: The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A summary. Acta Neuropathol 131:803, 2016. [PubMed: 27157931]

Morikawa A et al: Characteristics and outcomes of patients with breast cancer with leptomeningeal metastasis. Clin Breast Cancer 17:23, 2017. [PubMed: 27569275]

Omuro A et al: R-MVP followed by high-dose chemotherapy with TBC and autologous stem-cell transplant for newly diagnosed primary CNS lymphoma. Blood 125:1403, 2015. [PubMed: 25568347]

Pentsova EI et al: Evaluating cancer of the central nervous system through next-generation sequencing of cerebrospinal fluid. J Clin Oncol 34:2404, 2016. [PubMed: 27161972]

Ramaswamy V et al: Risk stratification of childhood medulloblastoma in the molecular era: The current consensus. Acta Neuropathol 131:821, 2016. [PubMed: 27040285]

Thibault I et al: Response assessment after stereotactic body radiotherapy for spinal metastasis: A report from the SPIne response assessment in Neuro-Oncology (SPINO) group, Lancet Oncol 16:e595, 2015. [PubMed: 26678212]

Tsakonas G et al: Management of brain metastasized non-small cell lung cancer (NSCLC)—From local treatment to new systemic therapies. Cancer Treat Rev 54:122, 2017. [PubMed: 28254730]

Yamada Y et al: The impact of histology and delivered dose on local control of spinal metastases treated with stereotactic radiosurgery. Neurosurg Focus 42:E6, 2017. [PubMed: 28041329]

Chapter 87: Soft Tissue and Bone Sarcomas and Bone Metastases

Shreyaskumar R. Patel

INTRODUCTION

Sarcomas are rare (<1% of all malignancies) mesenchymal neoplasms that arise in bone and soft tissues. These tumors are usually of mesodermal origin, although a few are derived from neuroectoderm, and they are biologically distinct from the more common epithelial malignancies. Sarcomas affect all age groups; 15% are found in children <15 years of age, and 40% occur after age 55 years. Sarcomas are one of the most common solid tumors of childhood and are the fifth most common cause of cancer deaths in children. Sarcomas may be divided into two groups, those derived from bone and those derived from soft tissues.

SOFT TISSUE SARCOMAS

Soft tissues include muscles, tendons, fat, fibrous tissue, synovial tissue, vessels, and nerves. Approximately 60% of soft tissue sarcomas arise in the extremities, with the lower extremities involved three times as often as the upper extremities. Thirty percent arise in the trunk, the retroperitoneum accounting for 40% of all trunk lesions. The remaining 10% arise in the head and neck.

INCIDENCE

Approximately 12,310 new cases of soft tissue sarcomas occurred in the United States in 2016. The annual age-adjusted incidence is 3 per 100,000 population, but the incidence varies with age. Soft tissue sarcomas constitute 0.7% of all cancers in the general population and 6.5% of all cancers in children.

EPIDEMIOLOGY

Malignant transformation of a benign soft tissue tumor is extremely rare, with the exception that malignant peripheral nerve sheath tumors (neurofibrosarcoma, malignant schwannoma) can arise from neurofibromas in patients with neurofibromatosis. Several etiologic factors have been implicated in soft tissue sarcomas.

Environmental Factors

Trauma or previous injury is rarely involved, but sarcomas can arise in scar tissue resulting from a prior operation, burn, fracture, or foreign body implantation. Chemical carcinogens such as polycyclic hydrocarbons, asbestos, and dioxin may be involved in the pathogenesis.

Iatrogenic Factors

Sarcomas in bone or soft tissues occur in patients who are treated with radiation therapy. The tumor nearly always arises in the irradiated field. The risk increases with time.

Viruses

Kaposi's sarcoma (KS) in patients with HIV type 1, classic KS, and KS in HIV-negative homosexual men is caused by human herpesvirus (HHV) 8 (Chap. 190). No other sarcomas are associated with viruses.

Immunologic Factors

Congenital or acquired immunodeficiency, including therapeutic immunosuppression, increases the risk of sarcoma.

GENETIC CONSIDERATIONS



Li-Fraumeni syndrome is a familial cancer syndrome in which affected individuals have germline abnormalities of the tumor-suppressor gene *p53* and an increased incidence of soft tissue sarcomas and other malignancies, including breast cancer, osteosarcoma, brain tumor, leukemia, and adrenal carcinoma (Chap. 67). Neurofibromatosis 1 (*NF-1*, peripheral form, von Recklinghausen's disease) is characterized by multiple neurofibromas and café-au-lait spots. Neurofibromas occasionally undergo malignant degeneration to become malignant peripheral nerve sheath tumors. The gene for *NF-1* is located in the pericentromeric region of chromosome 17 and encodes neurofibromin, a tumor-suppressor protein with guanosine 5′-triphosphate (gtp)ase-activating activity that inhibits ras function (Chap. 86). Germline mutation of the *Rb-1* locus (chromosome 13q14) in patients with inherited retinoblastoma is associated with the development of osteosarcoma in those who survive the retinoblastoma and of soft tissue sarcomas unrelated to radiation therapy. Other soft tissue tumors, including desmoid tumors, lipomas, leiomyomas, neuroblastomas, and paragangliomas, occasionally show a familial predisposition.

Ninety percent of synovial sarcomas contain a characteristic chromosomal translocation t(X;18)(p11;q11) involving a nuclear transcription factor on chromosome 18 called *SYT* and two breakpoints on X. Patients with translocations to the second X breakpoint (*SSX2*) may have longer survival than those with translocations involving *SSX1*.

Insulin-like growth factor (IGF) type II is produced by some sarcomas and may act as an autocrine growth factor and as a motility factor that promotes metastatic spread. IGF-II stimulates growth through IGF-I receptors, but its effects on motility are through different receptors. If secreted in large amounts, IGF-II may

produce hypoglycemia (Chaps. 89 and 399). A large international sarcoma kindred study including 1162 patients and 6545 Caucasian controls revealed that about half the patients with sarcoma have putatively pathogenic monogenic and polygenic variation in previously reported and new cancer genes, some of them representing therapeutically actionable targets. These patients were diagnosed with sarcoma at an earlier age compared to controls.

CLASSIFICATION

Approximately 20 different groups of sarcomas are recognized on the basis of the pattern of differentiation toward normal tissue. For example, rhabdomyosarcoma shows evidence of skeletal muscle fibers with cross-striations; leiomyosarcomas contain interlacing fascicles of spindle cells resembling smooth muscle; and liposarcomas contain adipocytes. When precise characterization of the group is not possible, the tumors are called *unclassified sarcomas*. All of the primary bone sarcomas can also arise from soft tissues (e.g., extraskeletal osteosarcoma). The entity *malignant fibrous histiocytoma* (*MFH*) includes many tumors previously classified as fibrosarcomas or as pleomorphic variants of other sarcomas and is characterized by a mixture of spindle (fibrous) cells and round (histiocytic) cells arranged in a storiform pattern with frequent giant cells and areas of pleomorphism. As immunohistochemical suggestion of differentiation, particularly myogenic differentiation, may be found in a significant fraction of these patients, many are now characterized as poorly differentiated leiomyosarcomas, and the terms *undifferentiated pleomorphic sarcoma* (UPS) and *myxofibrosarcoma* are replacing MFH and myxoid MFH.

For purposes of treatment, most soft tissue sarcomas can be considered together. However, some specific tumors have distinct features. For example, *liposarcoma* can have a spectrum of behaviors. Pleomorphic liposarcomas and dedifferentiated liposarcomas behave like other high-grade sarcomas; in contrast, well-differentiated liposarcomas (better termed *atypical lipomatous tumors*) lack metastatic potential, and myxoid liposarcomas metastasize infrequently, but, when they do, they have a predilection for unusual metastatic sites containing fat, such as the retroperitoneum, mediastinum, and subcutaneous tissue. Rhabdomyosarcomas, Ewing's sarcoma, and other small-cell sarcomas tend to be more aggressive and are more responsive to chemotherapy than other soft tissue sarcomas.

Gastrointestinal stromal tumors (GISTs), previously classified as gastrointestinal leiomyosarcomas, are now recognized as a distinct entity within soft tissue sarcomas. Its cell of origin resembles the interstitial cell of Cajal, which controls peristalsis. The majority of malignant GISTs have activating mutations of the c-*kit* gene that result in ligand-independent phosphorylation and activation of the KIT receptor tyrosine kinase, leading to tumorigenesis. Approximately 5–10% of tumors will have a mutation in the platelet-derived growth factor receptor α (*PDGFRA*). GISTs that are wild type for both *KIT* and *PDGFRA* mutations may show mutations in *SDH* B, C, or D and may be driven by the IGF-I pathway.

DIAGNOSIS

The most common presentation is an asymptomatic mass. Mechanical symptoms referable to pressure, traction, or entrapment of nerves or muscles may be present. All new and persistent or growing masses

should be biopsied, either by a small incision or by a cutting needle (core-needle biopsy) placed so that it can be encompassed in the subsequent excision without compromising a definitive resection. Lymph node metastases occur in 5%, except in synovial and epithelioid sarcomas, clear-cell sarcoma (melanoma of the soft parts), angiosarcoma, and rhabdomyosarcoma, where nodal spread may be seen in 17%. The pulmonary parenchyma is the most common site of metastases. Exceptions are GISTs, which metastasize to the liver; myxoid liposarcomas, which seek fatty tissue; and clear-cell sarcomas, which may metastasize to bones. Central nervous system metastases are rare, except in alveolar soft part sarcoma.

Radiographic Evaluation

Imaging of the primary tumor is best with plain radiographs and magnetic resonance imaging (MRI) for tumors of the extremities or head and neck and by computed tomography (CT) for tumors of the chest, abdomen, or retroperitoneal cavity. A radiograph and CT scan of the chest are important for the detection of lung metastases. Other imaging studies may be indicated, depending on the symptoms, signs, or histology.

STAGING AND PROGNOSIS

The histologic grade, relationship to fascial planes, and size of the primary tumor are the most important prognostic factors. The current American Joint Committee on Cancer (AJCC) staging system is shown in **Table 87-1**. Prognosis is related to the stage. Cure is common in the absence of metastatic disease, but a small number of patients with metastases can also be cured. Historically, most patients with stage IV disease used to die within 12 months, but with availability of multiple lines of treatments, median survival in second-line and beyond ranges from 13 to 14 months, and some patients may live with stable or slowly progressive disease for many years.

TABLE 87-1

American Joint Committee on Cancer Staging System for Sarcomas

HISTOLOGIC GRADE (G)	TUMOR SIZE (T)	NODE STATUS (N)	METASTASES (M)
Well differentiated (G1)	≤5 cm (T1)	Not involved (N0)	Absent (M0)
Moderately differentiated (G2)	>5 cm (T2)	Involved (N1)	Present (M1)
Poorly differentiated (G3)	Superficial fascial involvement (Ta)		
Undifferentiated (G4)	Deep fascial involvement (Tb)		
DISEASE STAGE	**5-YEAR SURVIVAL, %**		
Stage I	98.8		
A: G1,2; T1a,b; N0; M0			
B: G1,2; T2a; N0; M0			
Stage II	81.8		
A: G1,2; T2b; N0; M0			
B: G3,4; T1; N0; M0			
C: G3,4; T2a; N0; M0			
Stage III: G3,4; T2b; N0; M0	51.7		
Stage IV	<20		
A: any G; any T; N1; M0			
B: any G; any T; any N; M1			

TREATMENT

TREATMENT

Soft Tissue Sarcomas

AJCC stage I patients are adequately treated with surgery alone. Stage II patients are considered for adjuvant radiation therapy. Stage III patients may benefit from adjuvant chemotherapy. Stage IV patients are managed primarily with chemotherapy, with or without other modalities.

SURGERY

Soft tissue sarcomas tend to grow along fascial planes, with the surrounding soft tissues compressed to form a pseudocapsule that gives the sarcoma the appearance of a well-encapsulated lesion. This is invariably deceptive because “shelling out,” or marginal excision, of such lesions results in a 50–90% probability of local recurrence. Wide excision with a negative margin, incorporating the biopsy site, is the standard surgical procedure for local disease. The adjuvant use of radiation therapy and/or chemotherapy improves the local control rate and permits the use of limb-sparing surgery with a local control rate (85–90%) comparable to that achieved by radical excisions and amputations. Limb-sparing approaches are indicated except when negative margins are not obtainable, when the risks of radiation are prohibitive, or when neurovascular structures are involved so that resection will result in serious functional consequences to the limb.

RADIATION THERAPY

External-beam radiation therapy is an adjuvant to limb-sparing surgery for improved local control. Preoperative radiation therapy allows the use of smaller fields and smaller doses but results in a higher rate of wound complications. Postoperative radiation therapy must be given to larger fields, because the entire surgical bed must be encompassed, and in higher doses to compensate for hypoxia in the operated field. This results in a higher rate of late complications. Brachytherapy or interstitial therapy, in which the radiation source is inserted into the tumor bed, is comparable in efficacy (except in low-grade lesions), less time-consuming, and less expensive.

ADJUVANT CHEMOTHERAPY

Chemotherapy is the mainstay of treatment for Ewing’s primitive neuroectodermal tumors (PNET) and rhabdomyosarcomas. Meta-analysis of 14 randomized trials revealed a significant improvement in local control and disease-free survival in favor of doxorubicin-based chemotherapy. Overall survival improvement was 4% for all sites and 7% for the extremity site. An updated meta-analysis including four additional trials with doxorubicin and ifosfamide combination has reported a statistically significant 6% survival advantage in favor of chemotherapy. A chemotherapy regimen including an anthracycline and ifosfamide with growth factor support improved overall survival by 19% for high-risk (high-grade, ≥5 cm primary, or locally recurrent) extremity soft tissue sarcomas. Long-term follow-up of a trial evaluating neo-adjuvant use of the same combination confirms survival advantage and reports a 10-year survival of 61%.

ADVANCED DISEASE

Metastatic soft tissue sarcomas are largely incurable, but up to 20% of patients who achieve a complete response become long-term survivors. The therapeutic intent, therefore, is to produce a complete remission with chemotherapy (<10%) and/or surgery (30–40%). Surgical resection of metastases, whenever possible, is an integral part of the management. Some patients benefit from repeated surgical excision of metastases. The two most active chemotherapeutic agents are doxorubicin and ifosfamide. These drugs show a steep dose-response relationship in sarcomas. Gemcitabine with or without docetaxel has become an established second-line regimen and is particularly active in patients with undifferentiated pleomorphic sarcoma (UPS) and leiomyosarcomas. Dacarbazine also has some modest activity. Taxanes have selective activity in angiosarcomas, and vincristine, etoposide, and irinotecan are effective in rhabdomyosarcomas and Ewing's sarcomas. Pazopanib, an inhibitor of the vascular endothelial growth factor, platelet-derived growth factor (PDGF), and c-kit is now approved for patients with advanced soft tissue sarcomas excluding liposarcomas after failure of chemotherapy. Two additional chemotherapy drugs have gained approval from the Food and Drug Administration (FDA). Trabectedin was compared to dacarbazine in a large phase 3 randomized study in advanced leiomyosarcomas and liposarcomas after failure of an anthracycline, and resulted in significant improvement in progression-free survival. Eribulin was also tested in a similar trial and showed improvement in survival, predominantly in the liposarcoma subgroup and is therefore now approved for that subset. Imatinib targets the KIT and PDGF tyrosine kinase activity and is standard therapy for advanced/metastatic GISTs and dermatofibrosarcoma protuberans. Imatinib is now also indicated as adjuvant therapy for completely resected primary GISTs. Three years of adjuvant imatinib appear to be superior to 1 year of therapy for high-risk GISTs, although the optimal treatment duration remains unknown. Sunitinib and regorafenib are approved for second and third line use respectively in metastatic GIST after failure of or intolerance to imatinib.

Plexiform neurofibromas occurring in neurofibromatosis can be disfiguring and compromise function, particularly when they involve joints. These tumors are characterized by increased RAS-MAPK signaling and may respond to inhibiting MEK1/2 function with selumetinib.

BONE SARCOMAS

INCIDENCE AND EPIDEMIOLOGY

Bone sarcomas are rarer than soft tissue sarcomas; they accounted for only 0.2% of all new malignancies and 3010 new cases in the United States in 2016. Several benign bone lesions have the potential for malignant transformation. Enchondromas and osteochondromas can transform into chondrosarcoma; fibrous dysplasia, bone infarcts, and Paget's disease of bone can transform into either UPS or osteosarcoma.

CLASSIFICATION

Benign Tumors

The common benign bone tumors include enchondroma, osteochondroma, chondroblastoma, and chondromyxoid fibroma, of cartilage origin; osteoid osteoma and osteoblastoma, of bone origin; fibroma and

desmoplastic fibroma, of fibrous tissue origin; hemangioma, of vascular origin; and giant-cell tumor, of unknown origin.

Malignant Tumors

The most common malignant tumors of bone are plasma cell tumors **(Chap. 107)**. The four most common malignant nonhematopoietic bone tumors are osteosarcoma, chondrosarcoma, Ewing's sarcoma, and UPS. Rare malignant tumors include chordoma (of notochordal origin), malignant giant-cell tumor, adamantinoma (of unknown origin), and hemangioendothelioma (of vascular origin).

Musculoskeletal Tumor Society Staging System

Sarcomas of bone are staged according to the Musculoskeletal Tumor Society staging system based on grade and compartmental localization. A Roman numeral reflects the tumor grade: stage I is low grade, stage II is high grade, and stage III includes tumors of any grade that have lymph node or distant metastases. In addition, the tumor is given a letter reflecting its compartmental localization. Tumors designated A are intracompartmental (i.e., confined to the same soft tissue compartment as the initial tumor), and tumors designated B are extracompartmental (i.e., extending into the adjacent soft tissue compartment or into bone). The tumor-node-metastasis (TNM) staging system is shown in **Table 87-2**.

TABLE 87-2

Staging System for Bone Sarcomas

Primary tumor (T)		TX	Primary tumor cannot be assessed	
		T0	No evidence of primary tumor	
		T1	Tumor ≤8 cm in greatest dimension	
		T2	Tumor >8 cm in greatest dimension	
		T3	Discontinuous tumors in the primary bone site	
Regional lymph nodes (N)		NX	Regional lymph nodes cannot be assessed	
		N0	No regional lymph node metastasis	
		N1	Regional lymph node metastasis	
Distant metastasis (M)		MX	Distant metastasis cannot be assessed	
		M0	No distant metastasis	
		M1	Distant metastasis	
		M1a	Lung	
		M1b	Other distant sites	
Histologic grade (G)		GX	Grade cannot be assessed	
		G1	Well differentiated—low grade	
		G2	Moderately differentiated—low grade	
		G3	Poorly differentiated—high grade	
		G4	Undifferentiated—high grade (Ewing's is always classed G4)	
Stage Grouping				
Stage IA	T1	N0	M0	G1,2 low grade

Stage IB	T2	N0	M0	G1,2 low grade
Stage IIA	T1	N0	M0	G3,4 high grade
Stage IIB	T2	N0	M0	G3,4 high grade
Stage III	T3	N0	M0	Any G
Stage IVA	Any T	N0	M1a	Any G
Stage IVB	Any T	N1	Any M	Any G
	Any T	Any N	M1b	Any G

OSTEOSARCOMA

Osteosarcoma, accounting for almost 45% of all bone sarcomas, is a spindle cell neoplasm that produces osteoid (unmineralized bone) or bone. Approximately 60% of all osteosarcomas occur in children and adolescents in the second decade of life, and ~10% occur in the third decade of life. Osteosarcomas in the fifth and sixth decades of life are frequently secondary to either radiation therapy or transformation in a preexisting benign condition, such as Paget's disease. Males are affected 1.5–2 times as often as females. Osteosarcoma has a predilection for metaphyses of long bones; the most common sites of involvement are the distal femur, proximal tibia, and proximal humerus. The classification of osteosarcoma is complex, but 75% of osteosarcomas fall into the "classic" category, which include osteoblastic, chondroblastic, and fibroblastic osteosarcomas. The remaining 25% are classified as "variants" on the basis of (1) clinical characteristics, as in the case of osteosarcoma of the jaw, postradiation osteosarcoma, or Paget's osteosarcoma; (2) morphologic characteristics, as in the case of telangiectatic osteosarcoma, small-cell osteosarcoma, or epithelioid osteosarcoma; or (3) location, as in parosteal or periosteal osteosarcoma. Diagnosis usually requires a synthesis of clinical, radiologic, and pathologic features. Patients typically present with pain and swelling of the affected area. A plain radiograph reveals a destructive lesion with a moth-eaten appearance, a spiculated periosteal reaction (sunburst appearance), and a cuff of periosteal new bone formation at the margin of the soft tissue mass (Codman's triangle). A CT scan of the primary tumor is best for defining bone destruction and the pattern of calcification, whereas MRI is better for defining intramedullary and soft tissue extension. A chest radiograph and CT scan are used to detect lung metastases. Metastases to the bony skeleton should be imaged by a bone scan or by fluorodeoxyglucose positron emission tomography (FDG-PET). Almost all osteosarcomas are hypervascular and PET-avid. Pathologic diagnosis is established either with a core-needle biopsy, where feasible, or with an open biopsy with an appropriately placed incision that does not compromise future limb-sparing resection. Most osteosarcomas are high-grade. The most important prognostic factor for long-term survival is response to chemotherapy.

Preoperative chemotherapy followed by limb-sparing surgery (which can be accomplished in >80% of patients) followed by postoperative chemotherapy is standard management. The effective drugs are doxorubicin, ifosfamide, cisplatin, and high-dose methotrexate with leucovorin rescue. The various combinations of these agents that have been used have all been about equally successful. Long-term survival rates in extremity osteosarcoma range from 60 to 80%. Osteosarcoma is radioresistant; radiation therapy has no role in the routine management. UPS is considered a part of the spectrum of osteosarcoma and is managed similarly.

CHONDROSARCOMA

Chondrosarcoma, which constitutes ~20–25% of all bone sarcomas, is a tumor of adulthood and old age with a peak incidence in the fourth to sixth decades of life. It has a predilection for the flat bones, especially the shoulder and pelvic girdles, but can also affect the diaphyseal portions of long bones. Chondrosarcomas can arise de novo or as a malignant transformation of an enchondroma or, rarely, of the cartilaginous cap of an osteochondroma. Chondrosarcomas have an indolent natural history and typically present as pain and swelling. Radiographically, the lesion may have a lobular appearance with mottled or punctate or annular calcification of the cartilaginous matrix. It is difficult to distinguish low-grade chondrosarcoma from benign lesions by x-ray or histologic examination. The diagnosis is therefore influenced by clinical history and physical examination. A new onset of pain, signs of inflammation, and progressive increase in the size of the mass suggest malignancy. The histologic classification is complex, but most tumors fall within the classic category. Like other bone sarcomas, high-grade chondrosarcomas spread to the lungs. Most chondrosarcomas are resistant to chemotherapy, and surgical resection of primary or recurrent tumors, including pulmonary metastases, is the mainstay of therapy. This rule does not hold for two histologic variants. Dedifferentiated chondrosarcoma has a high-grade osteosarcoma or a malignant fibrous histiocytoma component that responds to chemotherapy. Mesenchymal chondrosarcoma, a rare variant composed of a small-cell element, also is responsive to systemic chemotherapy and is treated like Ewing's sarcoma.

EWING'S SARCOMA

Ewing's sarcoma, which constitutes ~10–15% of all bone sarcomas, is common in adolescence and has a peak incidence in the second decade of life. It typically involves the diaphyseal region of long bones and also has an affinity for flat bones. The plain radiograph may show a characteristic "onion peel" periosteal reaction with a generous soft tissue mass, which is better demonstrated by CT or MRI. This mass is composed of sheets of monotonous, small, round, blue cells and can be confused with lymphoma, embryonal rhabdomyosarcoma, and small-cell carcinoma. The presence of p30/32, the product of the *mic-2* gene (which maps to the pseudoautosomal region of the X and Y chromosomes), is a cell-surface marker for Ewing's sarcoma (and other members of the Ewing's family of tumors, previously also called PNETs). Most PNETs arise in soft tissues; they include peripheral neuroepithelioma, Askin's tumor (chest wall), and esthesioneuroblastoma. Glycogen-filled cytoplasm detected by staining with periodic acid–Schiff is also characteristic of Ewing's sarcoma cells. The classic cytogenetic abnormality associated with this disease is a

reciprocal translocation of the long arms of chromosomes 11 and 22, t(11;22), which creates a chimeric gene product of unknown function with components from the *fli-1* gene on chromosome 11 and *ews* on 22. This disease is very aggressive, and it is therefore considered a systemic disease. Common sites of metastases are lung, bones, and bone marrow. Systemic chemotherapy is the mainstay of therapy, often being used before surgery. Doxorubicin, cyclophosphamide or ifosfamide, etoposide, vincristine, and dactinomycin are active drugs. Topotecan or irinotecan in combination with an alkylating agent is often used in relapsed patients. Targeted therapy with an anti-IGF-I receptor antibody in combination with an inhibitor of mammalian target of rapamycin (mTOR) has shown promising activity in refractory cases. Local treatment for the primary tumor includes surgical resection, usually with limb salvage or radiation therapy. Patients with lesions below the elbow and below the mid-calf have a 5-year survival rate of 80% with effective treatment. Ewing's sarcoma at first presentation is a curable tumor, even in the presence of obvious metastatic disease, especially in children <11 years old.

TUMORS METASTATIC TO BONE

Bone is a common site of metastasis for carcinomas of the prostate, breast, lung, kidney, bladder, and thyroid and for lymphomas and sarcomas. Prostate, breast, and lung primaries account for 80% of all bone metastases. Metastatic tumors of bone are more common than primary bone tumors. Tumors usually spread to bone hematogenously, but local invasion from soft tissue masses also occurs. In descending order of frequency, the sites most often involved are the vertebrae, proximal femur, pelvis, ribs, sternum, proximal humerus, and skull. Bone metastases may be asymptomatic or may produce pain, swelling, nerve root or spinal cord compression, pathologic fracture, or myelophthisis (replacement of the marrow). Symptoms of hypercalcemia may be noted in cases of bony destruction.

Pain is the most frequent symptom. It usually develops gradually over weeks, is usually localized, and often is more severe at night. When patients with back pain develop neurologic signs or symptoms, emergency evaluation for spinal cord compression is indicated **(Chap. 71)**. Bone metastases exert a major adverse effect on quality of life in cancer patients.

Cancer in the bone may produce osteolysis, osteogenesis, or both. Osteolytic lesions result when the tumor produces substances that can directly elicit bone resorption (vitamin D–like steroids, prostaglandins, or parathyroid hormone–related peptide) or cytokines that can induce the formation of osteoclasts (interleukin 1 and tumor necrosis factor). Osteoblastic lesions result when the tumor produces cytokines that activate osteoblasts. In general, purely osteolytic lesions are best detected by plain radiography, but they may not be apparent until they are >1 cm. These lesions are more commonly associated with hypercalcemia and with the excretion of hydroxyproline-containing peptides indicative of matrix destruction. When osteoblastic activity is prominent, the lesions may be readily detected using radionuclide bone scanning (which is sensitive to new bone formation), and the radiographic appearance may show increased bone density or sclerosis. Osteoblastic lesions are associated with higher serum levels of alkaline phosphatase and, if extensive, may produce hypocalcemia. Although some tumors may produce mainly osteolytic lesions (e.g.,

kidney cancer) and others mainly osteoblastic lesions (e.g., prostate cancer), most metastatic lesions produce both types of lesion and may go through stages where one or the other predominates.

In older patients, particularly women, it may be necessary to distinguish metastatic disease of the spine from osteoporosis. In osteoporosis, the cortical bone may be preserved, whereas cortical bone destruction is usually noted with metastatic cancer.

TREATMENT

TREATMENT

Metastatic Bone Disease

Treatment of metastatic bone disease depends on the underlying malignancy and the symptoms. Some metastatic bone tumors are curable (lymphoma, Hodgkin's disease), and others are treated with palliative intent. Pain may be relieved by local radiation therapy. Hormonally responsive tumors are responsive to hormone inhibition (antiandrogens for prostate cancer, antiestrogens for breast cancer). Strontium-89, samarium-153, and radium-223 are bone-seeking radionuclides that can exert antitumor effects and relieve symptoms. Denosumab, a monoclonal antibody that binds to RANK ligand, inhibits osteoclastic activity and increases bone mineral density. Bisphosphonates such as pamidronate may relieve pain and inhibit bone resorption, thereby maintaining bone mineral density and reducing risk of fractures in patients with osteolytic metastases from breast cancer and multiple myeloma. Careful monitoring of serum electrolytes and creatinine is recommended. Monthly administration prevents bone-related clinical events and may reduce the incidence of bone metastases in women with breast cancer. When the integrity of a weight-bearing bone is threatened by an expanding metastatic lesion that is refractory to radiation therapy, prophylactic internal fixation is indicated. Overall survival is related to the prognosis of the underlying tumor. Bone pain at the end of life is particularly common; an adequate pain relief regimen including sufficient amounts of narcotic analgesics is required. **The management of hypercalcemia is discussed in Chap. 403**.

FURTHER READING

Ballinger ML et al: Monogenic and polygenic determinants of sarcoma risk: An international genetic study. Lancet Oncol 17:1261, 2016. [PubMed: 27498913]

Demetri GD et al: Long-term results of a randomized phase II study of trabectedin by two different dose and schedule regimens in patients with advanced liposarcoma or leiomyosarcoma after failure of prior anthracyclines and ifosfamide. J Clin Oncol 34:786, 2016. [PubMed: 26371143]

Dombi E et al. Activity of selumetinib in neurofibromatosis type 1 plexiform neurofibromas. N Engl J Med 376:2550, 2016.

Gronchi A et al: Short, full-dose adjuvant chemotherapy (CT) in high-risk adult soft tissue sarcomas (STS): Long-term follow-up of a randomized clinical trial from the Italian Sarcoma Group and the Spanish Sarcoma Group. Ann Oncol 27:1, 2016.

Nathenson MJ, Sausville E: Looking for answers: Current status of adjuvant treatment for localized soft tissue sarcomas. Cancer Chemother Pharmacol 78:895, 2016. [PubMed: 27206640]

Ratan R, Patel SR: Chemotherapy for soft tissue sarcoma. Cancer 122:2952, 2016. [PubMed: 27434055]

Schoffski P et al: Eribulin versus dacarbazine in previously treated patients with advanced liposarcoma or leiomyosarcoma: A randomised, open-label, multicentre, phase 3 trial. Lancet 387:1629, 2016. [PubMed: 26874885]

Wagner MJ et al: Chemotherapy for bone sarcoma in adults. J Oncol Pract 12:208, 2016. [PubMed: 26962160]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 88: Carcinoma of Unknown Primary

Gauri R. Varadhachary; James L. Abbruzzese

INTRODUCTION

Carcinoma of unknown primary (CUP) is a biopsy-proven malignancy for which the anatomic site of origin remains unidentified after a focused search. CUP is one of the 10 most frequently diagnosed cancers worldwide, accounting for 3–5% of all cancers. Most investigators limit CUP to epithelial and undifferentiated cancers and do not include lymphomas, metastatic melanomas, and metastatic sarcomas because these cancers have specific histology- and stage-based treatments that guide management even in the absence of a primary cancer.

The emergence of sophisticated imaging, robust immunohistochemistry (IHC), and genomic and proteomic tools has challenged the “unknown” designation. Additionally, effective targeted therapies in several cancers have moved the paradigm from empiricism to considering a personalized approach to CUP management. The reasons cancers present as CUP remain unclear. One hypothesis is that the primary tumor either regresses after seeding the metastasis or remains so small that it is not detected. It is possible that CUP falls on the continuum of cancer presentation where the primary has been contained or eliminated by the natural body defenses. Alternatively, CUP may represent a specific malignant event that results in an increase in metastatic spread or survival relative to the primary. Whether the CUP metastases truly define a clone that is genetically and phenotypically unique to this diagnosis remains to be determined. Of note, the incidence of intrahepatic cholangiocarcinoma (ICC) is increasing whereas the incidence of CUP is declining during this same time period. Because the liver is a common site of CUP presentation, ICC can be misdiagnosed as CUP. Improvements in diagnostic technologies and awareness among clinicians to differentiate the two are contributing to an increased recognized incidence of ICC.

CUP BIOLOGY

Studies looking for unique signature abnormalities in CUP tumors have not been positive. Abnormalities in chromosomes 1 and 12 and other complex cytogenetic abnormalities have been reported. Aneuploidy has been described in 70% of CUP patients with metastatic adenocarcinoma or undifferentiated carcinoma. The overexpression of various genes, including *Ras*, *bcl-2* (40%), *her-2* (11%), and *p53* (26–53%), has been identified in CUP samples, but they have no effect on response to therapy or survival. The extent of angiogenesis in CUP relative to that in metastases from known primaries has also been evaluated, but no consistent findings have emerged. Current focus is on comprehensive genomic profiling that may identify targeted therapeutic approaches to improve outcomes for this disease as discussed below. Additionally, ongoing profiling efforts may also provide insights into CUP biology through recognition of molecular aberrations that especially drive metastatic growth.

CLINICAL EVALUATION

Initial CUP evaluation has two goals: search for the primary tumor based on pathologic evaluation of the metastases and determine the extent of disease. Obtaining a thorough medical history from CUP patients is essential, including paying particular attention to previous surgeries, removed lesions, and family medical history to assess potential hereditary cancers. Adequate physical examination, including a digital rectal examination in men and breast and pelvic examinations in women, should be performed based on clinical presentation.

Role of Serum Tumor Markers and Cytogenetics

Most tumor markers, including CEA, CA-125, CA 19-9, and CA 15-3, when elevated, are nonspecific and not helpful in determining the primary tumor site. Men who present with adenocarcinoma and predominant osteoblastic metastasis should undergo a prostate-specific antigen (PSA) test. In patients with undifferentiated or poorly differentiated carcinoma (especially with a midline tumor), elevated β-human chorionic gonadotropin (β-hCG) and α fetoprotein (AFP) levels suggest the possibility of an extragonadal germ cell (testicular) tumor. AFP should also be considered in patients with a potential diagnosis of hepatoma. With the availability of IHC, cytogenetic studies are rarely needed.

Role of Imaging Studies

In the absence of contraindications, a baseline IV contrast computed tomography (CT) scan of the chest, abdomen, and pelvis is the standard of care. This helps to search for the primary tumor, evaluate the extent of disease, and select the most accessible biopsy site. Older studies suggested that the primary tumor site is detected in 20–35% of patients who undergo a CT scan of the abdomen and pelvis, although by current definition, these patients do not have CUP. With precise imaging and reporting, latent primary cancers, defined as appearance of a new primary cancer after a latent period of several months to

years, is uncommon and seen in ≤5% of CUP patients, usually in patients with very indolent presentations and/or highly responsive metastatic cancers that allows a latent primary to emerge (grow) over time.

Mammography should be performed in all women who present with metastatic adenocarcinoma, especially in those with adenocarcinoma and isolated axillary lymphadenopathy. Magnetic resonance imaging (MRI) of the breast is a follow-up modality in patients with axillary adenopathy and suspected occult primary breast carcinoma following a negative mammography and ultrasound. The results of these imaging modalities can influence surgical management; a negative MRI of the breast result predicts a low tumor yield at mastectomy.

A conventional workup for a squamous cell carcinoma and cervical CUP (neck lymphadenopathy with no known primary tumor) includes a CT scan or MRI and invasive studies, including indirect and direct laryngoscopy, bronchoscopy, and upper endoscopy. Ipsilateral (or bilateral) staging tonsillectomy has been recommended for these patients. 18-Fluorodeoxyglucose positron emission tomography (18-FDG-PET) scans are useful in this patient population and may help guide the biopsy; determine the extent of disease; facilitate the appropriate treatment, including planning radiation fields; and help with disease surveillance. A smaller radiation field encompassing the primary (when found) and metastatic adenopathy decreases the risk of chronic xerostomia. Several studies have evaluated the utility of PET in patients with squamous cervical CUP, and head and neck primary tumors were identified in ~21–30%.

The diagnostic contribution of PET to the evaluation of other CUP presentations (outside of the neck adenopathy indication) remains controversial and is not routinely recommended. PET-CT can be helpful for patients who are candidates for surgical intervention for solitary metastatic disease because the identification of disease in addition to the solitary metastatic site may affect surgical planning.

Invasive studies, including upper endoscopy, colonoscopy, and bronchoscopy, should be limited to symptomatic patients or those with laboratory, imaging, or pathologic abnormalities that suggest that these techniques will result in a high yield in finding a primary cancer.

Role of Pathologic Studies

A detailed pathologic examination of the most accessible biopsied tissue specimen is mandatory in CUP patients. Pathologic evaluation typically consists of hematoxylin and eosin stains and immunohistochemical tests.

LIGHT MICROSCOPY EVALUATION

Adequate tissue obtained preferably by excisional biopsy or core-needle biopsy (instead of only a fine-needle aspiration) is stained with hematoxylin and eosin and subjected to light microscopic examination. On light microscopy, 60–65% of CUP is adenocarcinoma, and 5% is squamous cell carcinoma. The remaining 30–35% is poorly differentiated adenocarcinoma, poorly differentiated carcinoma, or poorly differentiated neoplasm. A small percentage of lesions are diagnosed as neuroendocrine cancers (2%), mixed tumors (adenosquamous or sarcomatoid carcinomas), or undifferentiated neoplasms (Table 88-1).

TABLE 88-1

Major Histologies in Carcinoma of Unknown Primary

Histology	Proportion, %
Well to moderately differentiated adenocarcinoma	60
Squamous cell cancer	5
Poorly differentiated adenocarcinoma, poorly differentiated carcinoma	30
Neuroendocrine	2
Undifferentiated malignancy	3

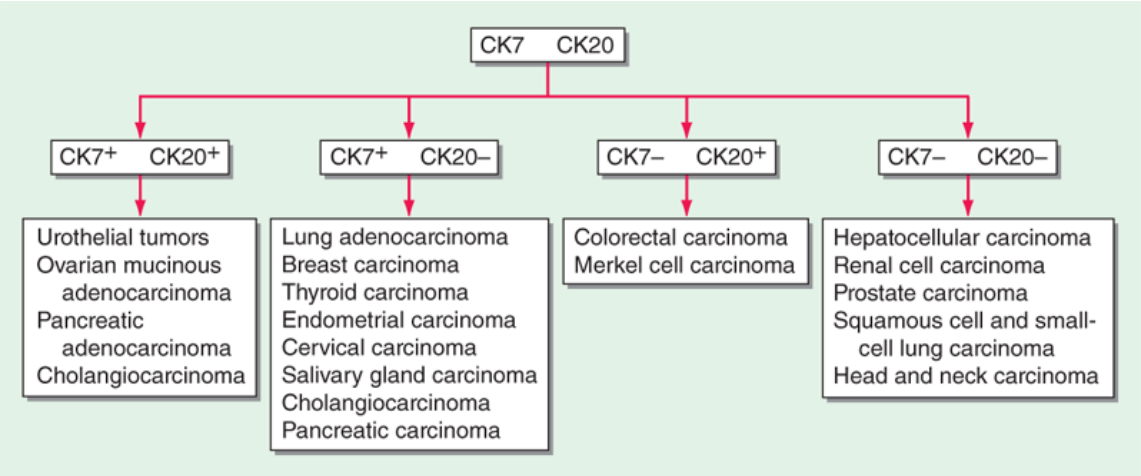
ROLE OF IMMUNOHISTOCHEMICAL ANALYSIS

Immunohistochemical stains are peroxidase-labeled antibodies against specific tumor antigens that are used to define tumor lineage. The number of available immunohistochemical stains is ever-increasing and unfortunately, we lack a tiered and uniform approach to tissue evaluation in the CUP setting. For CUP cases, more is not necessarily better, and immunohistochemical stains should be used in conjunction with the patient's clinical presentation and imaging studies to select the best therapy. Communication between the clinician and pathologist is essential. No stain is 100% specific, and overinterpretation should be avoided. PSA and thyroglobulin tissue markers, which are positive in prostate and thyroid cancer, respectively, are the most specific of the current marker panel. However, these cancers rarely present as CUP, so the yield of these tests may be low. Figure 88-1 delineates a simple algorithm for immunohistochemical staining in CUP cases. Table 88-2 lists additional tests that may be useful to further define the tumor lineage. A

more comprehensive algorithm may improve the diagnostic accuracy but can make the process complex and increase cost. With the use of immunohistochemical markers, electron microscopic analysis, which is time-consuming and expensive, is rarely needed.

FIGURE 88-1

Approach to cytokeratin (CK7 and CK20) markers used in adenocarcinoma of unknown primary.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 88-2

Select Immunohistochemical Stains Useful in the Diagnosis of CUP

Likely Primary Profile	Commonly Considered IHC to Assist in Differential Diagnosis of CUP[a]
Breast	ER, GCDFP-15, mammaglobin, Her-2/neu, GATA3
Ovarian/mullerian	ER, *WT1* gene, CK7, PAX8, PAX2
Lung adenocarcinoma	TTF-1; nuclear staining, napsin A, SP-A1
Germ cell	β-HCG, AFP, OCT3/4, CKIT, CD30 (embryonal), SALL4
Prostate	PSA, α-methylacyl CoA racemase/P504S (AMACR/P504S), P501S (prostein), and PSMA, NKX3-1
Intestinal	CK7, CK20, CDX-2, CEA
Neuroendocrine	Chromogranin, synaptophysin, CD56
Sarcoma	Desmin (desmoid tumors), factor VIII (angiosarcomas), CD31, smooth muscle actin (leiomyosarcoma), MyoD1 (rhabdomyosarcoma)
Renal	RCC, CD10, PAX8, CD10
Hepatocellular carcinoma	Hep Par-1, Arg-1, glypican-3
Melanoma	S100, SOX-10, vimentin, HMB-45, tyrosinase and melan-A
Urothelial	CK7, CK20, thrombomodulin, uroplakin III
Mesothelioma	Calretinin, *WT1*, D2-40, mesothelin
Lymphoma	LCA, CD3, CD4, CD5, CD20, CD45
SCC	p63, p40 (lung SCC), CK5/6

[a]Patterns emerging from coexpression of stains are better than individual stains to suggest putative primary site. Even with optimization, no IHC panel is 100% sensitive or specific (e.g., ovarian mucinous carcinoma can exhibit positivity with intestinal markers).

Abbreviations: AFP, α fetoprotein; Arg-1, arginase-1; β-hCG, β-human chorionic gonadotropin; CEA, carcinoembryonic antigen; CUP, carcinoma of unknown primary; ER, estrogen receptor; GCDFP-15, gross cystic disease fibrous protein-15; IHC, immunohistochemistry; LCA, leukocyte common antigen; PSA, prostate-specific antigen; PSMA, prostate-specific membrane antigen; SCC, squamous cell carcinoma; SP-A1, surfactant protein A precursor; TTF, thyroid transcription factor; WT, Wilms' tumor.

There are >20 subtypes of cytokeratin (CK) intermediate filaments with different molecular weights and differential expression in various cell types and cancers. Monoclonal antibodies to specific CK subtypes have been used to help classify tumors according to their site of origin; commonly used CK stains in adenocarcinoma CUP are CK7 and CK20. CK7 is found in tumors of the lung, ovary, endometrium, breast, and upper gastrointestinal tract including pancreaticobiliary cancers, whereas CK20 is normally expressed in the gastrointestinal epithelium, urothelium, and Merkel cells. The nuclear CDX-2 transcription factor, which is the product of a homeobox gene necessary for intestinal organogenesis, is often used to aid in the diagnosis of gastrointestinal adenocarcinomas.

Thyroid transcription factor 1 (TTF-1) nuclear staining is frequently positive in lung and thyroid cancers. Approximately 68% of adenocarcinomas and 25% of squamous cell lung cancers stain positive for TTF-1, which helps differentiate a lung primary tumor from metastatic adenocarcinoma in a pleural effusion, the mediastinum, or the lung parenchyma.

Gross cystic disease fibrous protein-15, a 15-kDa monomer protein, is a marker of apocrine differentiation that is detected in 62–72% of breast carcinomas. GATA3 is being increasingly used in the CUP setting when there is concern for a breast primary, and can be particularly useful as a marker for metastatic breast carcinoma, especially triple-negative and metaplastic carcinomas, which lack specific endocrine markers of mammary origin. UROIII, high-molecular-weight cytokeratin, thrombomodulin, and CK20 are the markers used to diagnose lesions of urothelial origin.

IHC performs the best when used in groups that give rise to patterns that are strongly indicative of certain profiles. For example, the TTF-1/CK7+ and CK20+/CDX-2+/CK7− phenotypes have been reported as very suggestive of lung and lower gastrointestinal cancer profiles, respectively. Despite their practical utility, these patterns have not been validated prospectively in CUP patients. IHC is not without its limitations; several factors affect tissue antigenicity (antigen retrieval, specimen processing, and fixation), interpretation of stains in tumor (nuclear, cytoplasmic, membrane) versus normal tissue, inter- and intraobserver variability, variable performance of different antibodies said to recognize the same antigen, and tissue heterogeneity and inadequacy (given small biopsy sizes). Communication with the pathologist is critical to determine if additional tissue will be beneficial in the pathologic evaluation. Pathologic features should supersede clinical or radiologic findings when considering testing for biomarkers of therapeutic response (e.g., epidermal growth factor receptor [EGFR], Alk mutations, human epidermal growth factor receptor 2 [HER-2]).

ROLE OF CANCER CLASSIFIER MOLECULAR PROFILING

In the absence of a known primary, developing therapeutic strategies for CUP is challenging. The current diagnostic yield with imaging and immunochemistry is ~20–30% for CUP patients. To reduce diagnostic uncertainty, sophisticated molecular analytics have been applied to CUP samples. These include gene expression profiling, messenger RNA (mRNA), microRNA, and epigenetic profiling to classify the CUP cancer.

Gene expression profiles are most commonly generated using quantitative reverse transcriptase polymerase chain reaction (RT-PCR) or DNA microarray. Neural network programs are then used to develop predictive algorithms from the gene expression profiles. Typically, a training set of gene profiles from known cancers (preferably from metastatic sites) is used to train the software. Comprehensive gene expression databases that have become available for common malignancies are then applied to CUP samples and the program can then be used to predict the putative origin of a CUP sample.

mRNA- or microRNA-based tissue of origin cancer classifier assays have also been studied in prospective and retrospective CUP trials. More recently, a classifier based on microarray DNA methylation signatures has been studied and validated in known cancers. The DNA methylation profiling predicted a primary cancer in 87% of the 216 CUP patients.

Despite the sophistication of the cancer classifier molecular assays, most of the CUP studies have evaluated assay *performance*, although the challenge with validating the accuracy of an assay for CUP is that, by definition, the primary cancer diagnosis cannot be verified. Thus, current estimates of tissue of origin test accuracy have relied on indirect metrics, including comparison with pathology/IHC, clinical presentation, appearance of latent primaries, and autopsies. Using these measures, the assays suggest a plausible primary in ~70–80% of patients studied. The only outcomes-based study is a single-arm study reporting a median survival of 12.5 months for patients who received assay-directed site-specific therapy. Firm conclusions of therapeutic impact cannot be drawn from this study given the nonrandomized design, statistical biases, confounding variables including use of subsequent lines of (empiric) therapy, and the heterogeneity of the CUP cancers. Additional studies are needed to better understand the clinical influence of tissue of origin profiling tools and how these assays complement IHC and help guide therapy.

ROLE OF NEXT GENERATION SEQUENCING

A significant push is being made toward personalized medicine across all cancer types with the goal of identifying driver mutation(s) in a patient who can be treated with targeted agents independent of the site of origin. A retrospective study of 200 CUP tumor specimens reported on genomic alterations (GA) using the hybrid-capture–based FoundationOne assay. The authors reported that a large number of CUP samples (85%) harbored at least one clinically relevant GA with the potential to influence and personalize therapy. The mean number of GAs was 4.2 per tumor, and the most common GAs included *TP53* (55%), *KRAS* (20%), *CDKN2A* (19%), and *ARID1A* (11%). The adenocarcinoma CUP tumors were more frequently driven by GAs in the receptor tyrosine kinase (RTK)/Ras/mitogen-activated protein kinase (MAPK) signaling pathway than nonadenocarcinoma CUP tumors.

Ongoing histology and cellular-context agnostic prospective clinical trials are studying the presence of actionable mutations and matching patients to the right targeted drug. Should this approach eventually be appropriately validated, CUP would be a natural fit for GA-based targeted therapy independent of tumor site.

TREATMENT

TREATMENT

Carcinoma of Unknown Primary

GENERAL CONSIDERATIONS

The treatment of CUP continues to evolve, albeit slowly. The median survival duration of most patients with disseminated CUP is ~6–10 months. Systemic chemotherapy is the primary treatment modality in most patients with disseminated disease, but the careful integration of surgery, radiation therapy, and even periods of observation is important in the overall management of this condition (Figs. 88-2 and 88-3). Prognostic factors include performance status, site and number of metastases, response to chemotherapy, and serum lactate dehydrogenase (LDH) levels. Culine and colleagues developed a prognostic model using performance status and serum LDH levels, which allowed the assignment of patients into two subgroups with divergent outcomes. Future prospective trials using this prognostic model are warranted. Clinically, some CUP diagnoses fall into a favorable prognostic subset. Others, including those with disseminated CUP, do not and have a more unfavorable prognosis.

TREATMENT OF FAVORABLE CUP SUBSETS

Women with Isolated Axillary Adenopathy

Women with isolated axillary adenopathy with adenocarcinoma or carcinoma are usually treated for stage II or III breast cancer based on pathologic findings. These patients should undergo a breast MRI if mammogram and ultrasound are negative. Radiation therapy to the ipsilateral breast is indicated if the MRI of the breast is positive. Chemotherapy and/or hormonal therapy are indicated based on patient's age (premenopausal or postmenopausal), nodal disease bulk, and hormone receptor status **(Chap. 75)**. It is important to verify that the pathology suggests a breast cancer profile (morphology, immunohistochemical breast markers including estrogen receptor, mammaglobin, GCDFP-15, GATA3, HER-2 gene expression) before embarking on a breast cancer therapeutic program.

WOMEN WITH PERITONEAL CARCINOMATOSIS

The term *primary peritoneal papillary serous carcinoma* (PPSC) has been used to describe CUP with carcinomatosis with the pathologic and laboratory (elevated CA-125 antigen) characteristics of ovarian cancer but no ovarian primary tumor identified on transvaginal sonography or laparotomy. Studies suggest that ovarian cancer and PPSC, which are both of müllerian origin, have similar gene expression profiles. Similar to patients with ovarian cancer, patients with PPSC are candidates for cytoreductive surgery, followed by adjuvant taxane- and platinum-based chemotherapy. In one retrospective study of 258 women with peritoneal carcinomatosis who had undergone cytoreductive surgery and chemotherapy, 22% of patients had a complete response to chemotherapy; the median survival duration was 18 months (range 11–24 months). However, not all peritoneal carcinomatosis in women is PPSC. Careful pathologic evaluation can help diagnose a colon cancer profile (CDX-2+, CK-20+, CK7−) or a pancreaticobiliary cancer or even a mislabeled peritoneal mesothelioma (calretinin positive).

POORLY DIFFERENTIATED CARCINOMA WITH MIDLINE ADENOPATHY

Men with poorly differentiated or undifferentiated carcinoma that presents with midline adenopathy should be evaluated for extragonadal germ cell malignancy **(Chap. 84)**. If diagnosed and treated as such, they often experience a good response to treatment with platinum-based combination chemotherapy. Response rates of >50% have been noted, and long-term survival rates of 10–15% long have been reported. Older patients, especially smokers, who present with mediastinal adenopathy are more likely to have a lung or head-and-neck cancer profile.

NEUROENDOCRINE CARCINOMA

Low-grade neuroendocrine tumor (NET) often has an indolent course, and treatment decisions are based on symptoms and tumor bulk **(Chap. 80)**. Urine 5-HIAA and serum chromogranin may be elevated and can be followed as markers. Often, the patient is treated with somatostatin analogues alone for hormone-related symptoms (diarrhea, flushing, nausea). Specific local therapies or systemic therapy would only be indicated if the patient is symptomatic with local pain secondary to significant growth of the metastasis or the hormone-related symptoms are not controlled with endocrine therapy. Novel therapy options have demonstrated benefit in patients with low-grade NET, including sunitinib (which targets the vascular endothelial growth factor pathway) and everolimus (which inhibits the mammalian target of rapamycin). Patients with high-grade NET are treated with platinum-based doublet therapy; 20–25% show a complete response, and up to 10% patients with limited/oligo presentations survive for >5 years.

SQUAMOUS CELL CARCINOMA PRESENTING AS NECK ADENOPATHY

Patients with early-stage squamous cell carcinoma involving the cervical lymph nodes are candidates for node dissection and radiation therapy, which can result in long-term survival. The role of chemotherapy in these patients is undefined, although chemoradiation therapy or induction chemotherapy is often used and is beneficial in bulky N2/N3 lymph node disease.

SOLITARY METASTATIC SITE

Patients with solitary metastases can also experience good treatment outcomes. Some patients who present with locoregional disease are candidates for aggressive trimodality management—both prolonged disease-free survival, and, occasionally, cure is possible.

MEN WITH BLASTIC SKELETAL METASTASES AND ELEVATED PSA

Blastic bone-only metastasis is a rare presentation, and elevated serum PSA or tumor staining with PSA may provide confirmatory evidence of prostate cancer in these patients **(Chap. 83)**. Those with elevated levels are candidates for hormonal therapy for prostate cancer, although it is important to rule out other primary tumors (lung most common).

MANAGEMENT OF DISSEMINATED CUP

Patients who present with liver, brain, and adrenal metastatic disease usually have a poor prognosis. Patients with peritoneal carcinomatosis secondary to metastatic adenocarcinoma have a broad differential diagnosis, which includes mainly gastrointestinal cancers including gastric, appendiceal, colon, and pancreaticobiliary cancers.

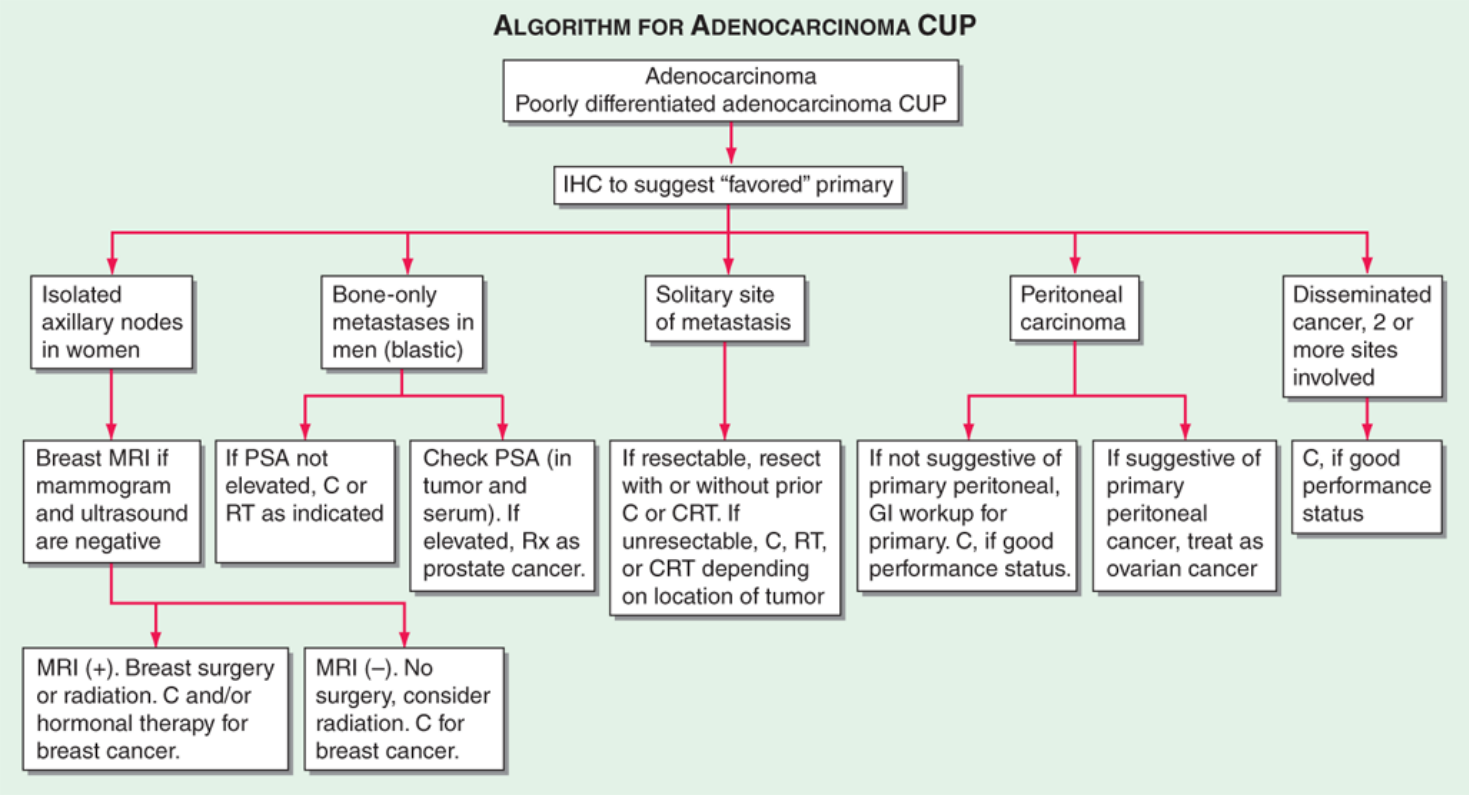
Traditionally, platinum-based combination chemotherapy regimens have been used to treat CUP. Several broadly used regimens have been studied in the last two decades; these include paclitaxel-carboplatin, gemcitabine-cisplatin, gemcitabine-oxaliplatin, and irinotecan and fluoropyrimidine-based therapies. These chemotherapeutic agents used as empiric regimens have shown a response rate of 25–40%, and their use obtains median survival times of 6–13 months.

Outside of favorable subsets, there is a small group of patients with a "definitive" IHC profile. These patients usually have a single diagnosis based on their clinicopathologic presentation and are often treated for the putative primary tumor. This does not guarantee a response, although it increases the probability of response when select drugs are chosen from a class of drugs known to be effective in that cancer type. Patients who do not fall into those

categories are candidates for broad-spectrum platinum-based regimens, clinical trials, and additional trial-based genomic and proteomic tests. Today, we do not have many effective drugs for several CUP cancer profiles, and treatments overlap for some cancers. However, as novel therapies are developed for known cancers, these agents will likely impact management of CUP patients.

FIGURE 88-2

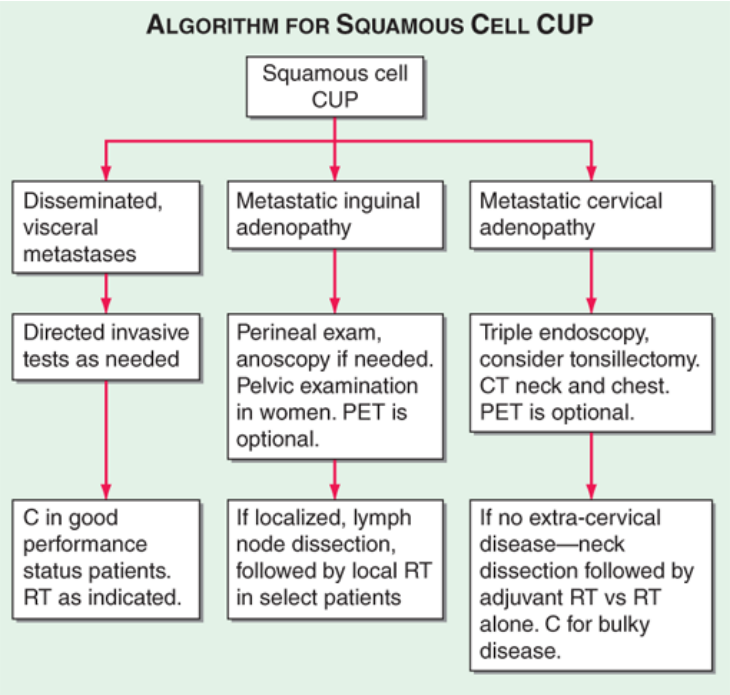
Treatment algorithm for adenocarcinoma and poorly differentiated adenocarcinoma of unknown primary (CUP). C, chemotherapy; CRT, chemoradiation; GI, gastrointestinal; IHC, immunohistochemistry; MRI, magnetic resonance imaging; PSA, prostate-specific antigen; RT, radiation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 88-3

Treatment algorithm for squamous cell carcinoma of unknown primary (CUP). C, chemotherapy; CT, computed tomography; PET, positron emission tomography; RT, radiation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SUMMARY

Patients with CUP should undergo a directed diagnostic search for the primary tumor on the basis of clinical and pathologic data. Subsets of patients have prognostically favorable disease, as defined by clinical or histologic criteria, and may substantially benefit from aggressive treatment and prolonged survival can be expected. However, for most patients who present with advanced CUP, the prognosis remains poor with early resistance to available cytotoxic therapy. The current focus has shifted away from empirical chemotherapeutic trials to understanding the metastatic phenotype, tissue of origin profiling in select patients, and next-generation sequencing to evaluate actionable mutations in CUP patients.

FURTHER READING

Fizazi K et al: Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 26(Suppl 5):v133, 2015. [PubMed: 26314775]

Hainsworth JD et al: Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: A prospective trial of the Sarah Cannon Research Institute. J Clin Oncol 31:217, 2013. [PubMed: 23032625]

National Comprehensive Cancer Network: Occult Primary (Cancer of Unknown Primary) version 2.2017, October 2016. *http://www.nccn.org/professionals/physician_gls/pdf/occult.pdf*.

Oien KA, Dennis JL: Diagnostic work-up of carcinoma of unknown primary: From immunohistochemistry to molecular profiling. Ann Oncol 23:271, 2012.

Ross JS et al: Comprehensive genomic profiling of carcinoma of unknown primary site: New routes to targeted therapies. JAMA Oncol 1:40, 2015. [PubMed: 26182302]

Rudmik L et al: Clinical utility of PET/CT in the evaluation of head and neck squamous cell carcinoma with an unknown primary: A prospective clinical trial. Head Neck 33:935, 2011. [PubMed: 21674668]

Sebastian MS et al: Epigenetic profiling to classify cancer of unknown primary: A multicentre, retrospective analysis. Lancet Oncol 17:1386–1395, 2016. [PubMed: 27575023]

Varadhachary GR, Raber MN: Cancer of unknown primary site. N Engl J Med 371:757, 2014. [PubMed: 25140961]

Chapter 89: Paraneoplastic Syndromes: Endocrinologic/Hematologic

J. Larry Jameson; Dan L. Longo

INTRODUCTION

Neoplastic cells can produce a variety of products that can stimulate hormonal, hematologic, dermatologic, rheumatologic, renal, and neurologic responses. *Paraneoplastic syndromes* is the term used to refer to the disorders that accompany benign or malignant tumors but are not directly related to mass effects or invasion. Tumors of neuroendocrine origin, such as small-cell lung carcinoma (SCLC) and carcinoids are common causes of paraneoplastic syndromes, and produce a wide array of peptide hormones and antibodies. However, almost every type of tumor has the potential to produce hormones or to induce cytokine and immunologic responses. Careful studies of the prevalence of paraneoplastic syndromes indicate that they are more common than is generally appreciated. The signs, symptoms, and metabolic alterations associated with paraneoplastic disorders may be overlooked in the context of a malignancy and its treatment. Consequently, atypical clinical manifestations in a patient with cancer should prompt consideration of a paraneoplastic syndrome. The most common hormonal and hematologic syndromes associated with underlying neoplasia will be discussed here.

ENDOCRINE PARANEOPLASTIC SYNDROMES

Etiology

Hormones can be produced from eutopic or ectopic sources. *Eutopic* refers to the expression of a hormone from its normal tissue of origin, whereas *ectopic* refers to hormone production from an atypical tissue source. For example, adrenocorticotropic hormone (ACTH) is expressed eutopically by the corticotrope cells of the anterior pituitary, but it can be expressed ectopically in SCLC. Many hormones are produced at low levels from a wide array of tissues in addition to the classic endocrine source. Thus, ectopic expression is often a quantitative change rather than an absolute change in tissue expression. Nevertheless, the term *ectopic expression* is firmly entrenched and conveys the abnormal physiology associated with hormone production by neoplastic cells. In addition to high levels of hormones, ectopic expression is often characterized by abnormal regulation of hormone production (e.g., defective feedback control) and peptide processing (resulting in large, unprocessed precursors).

A diverse array of molecular mechanisms has been suggested to explain ectopic hormone production. In rare instances, genetic rearrangements account for aberrant hormone expression. For example, translocation of the parathyroid hormone (*PTH*) gene can result in high levels of PTH expression in tissues other than the parathyroid gland because the genetic rearrangement brings the *PTH* gene under the control of atypical regulatory elements. A related phenomenon is well documented in many forms of leukemia and lymphoma, in which somatic genetic rearrangements confer a growth advantage and alter cellular differentiation and function. Although genetic rearrangements cause selected cases of ectopic hormone production, this mechanism is rare, as many tumors are associated with excessive production of numerous peptides. Cellular dedifferentiation probably underlies most cases of ectopic hormone production. Many cancers are poorly differentiated, and certain tumor products, such as human chorionic gonadotropin (hCG), PTH–related protein (PTHrP), and α fetoprotein, are characteristic of gene expression at earlier developmental stages. In contrast, the propensity of certain cancers to produce particular hormones (e.g., squamous cell carcinomas produce PTHrP) suggests that dedifferentiation is partial or that selective pathways are derepressed. These expression profiles probably reflect epigenetic modifications that alter transcriptional repression, microRNA expression, and other pathways that govern cell differentiation.

In SCLC, the pathway of differentiation has been relatively well defined. The neuroendocrine phenotype is dictated in part by the basic-helix-loop-helix (bHLH) transcription factor human achaete-scute homologue 1 (hASH-1), which is expressed at abnormally high levels in SCLC associated with ectopic ACTH. The activity of hASH-1 is inhibited by hairy enhancer of split 1 (HES-1) and by Notch proteins, which also are capable of inducing growth arrest. Thus, abnormal expression of these developmental transcription factors appears to provide a link between cell proliferation and differentiation.

Ectopic hormone production might be considered merely epiphenomenon associated with cancer if it did not result in clinical manifestations. Excessive and unregulated production of hormones such as ACTH, PTHrP, and vasopressin can lead to substantial morbidity and complicate the cancer treatment plan. Moreover, the paraneoplastic endocrinopathies may be a presenting clinical feature of underlying malignancy and prompt the search for an unrecognized tumor.

A large number of paraneoplastic endocrine syndromes have been described, linking overproduction of particular hormones with specific types of tumors. However, certain recurring syndromes emerge from this group (Table 89-1). The most common paraneoplastic endocrine syndromes include hypercalcemia from overproduction of PTHrP and other factors, hyponatremia from excess vasopressin, and Cushing's syndrome from ectopic ACTH.

TABLE 89-1

Paraneoplastic Syndromes Caused by Ectopic Hormone Production

Paraneoplastic Syndrome	Ectopic Hormone	Typical Tumor Types[a]
Common		
Hypercalcemia of malignancy	Parathyroid hormone–related protein (PTHrP)	Squamous cell (head and neck, lung, skin), breast, genitourinary, gastrointestinal
	1,25-dihydroxyvitamin D	Lymphomas
	Parathyroid hormone (PTH) (rare)	Lung, ovary
	Prostaglandin E_2 (PGE_2) (rare)	Renal, lung
Syndrome of inappropriate antidiuretic hormone secretion (SIADH)	Vasopressin	Lung (squamous, small cell), gastrointestinal, genitourinary, ovary
Cushing's syndrome	Adrenocorticotropic hormone (ACTH)	Lung (small cell, bronchial carcinoid, adenocarcinoma, squamous), thymus, pancreatic islet, medullary thyroid carcinoma, pheochromocytoma
	Corticotropin-releasing hormone (CRH) (rare)	Pancreatic islet, carcinoid, lung, prostate
	Ectopic expression of gastric inhibitory peptide (GIP), luteinizing hormone (LH)/human chorionic gonadotropin (hCG), other G protein–coupled receptors (rare)	Macronodular adrenal hyperplasia
Less Common		

Paraneoplastic Syndrome	Ectopic Hormone	Typical Tumor Types[a]
Non–islet cell hypoglycemia	Insulin-like growth factor type II (IGF-II)	Mesenchymal tumors, sarcomas, adrenal, hepatic, gastrointestinal, kidney, prostate
	Insulin (rare)	Cervix (small-cell carcinoma)
Male feminization	hCG[b]	Testis (embryonal, seminomas), germinomas, choriocarcinoma, lung, hepatic, pancreatic islet
Diarrhea or intestinal hypermotility	Calcitonin[c]	Lung, colon, breast, medullary thyroid carcinoma
	Vasoactive intestinal peptide (VIP)	Pancreas, pheochromocytoma, esophagus
Rare		
Oncogenic osteomalacia	Phosphatonin (fibroblast growth factor 23 [FGF23])	Hemangiopericytomas, osteoblastomas, fibromas, sarcomas, giant cell tumors, prostate, lung
Acromegaly	Growth hormone–releasing hormone (GHRH)	Pancreatic islet, bronchial, and other carcinoids
	Growth hormone (GH)	Lung, pancreatic islet
Hyperthyroidism	Thyroid-stimulating hormone (TSH)	Hydatidiform mole, embryonal tumors, struma ovarii
Hypertension	Renin	Juxtaglomerular tumors, kidney, lung, pancreas, ovary
Consumptive hypothyroidism	Type 3 deiodinase	Hepatic hemangiomas

[a]Only the most common tumor types are listed. For most ectopic hormone syndromes, an extensive list of tumors has been reported to produce one or more hormones. [b]hCG is produced eutopically by trophoblastic tumors. Certain tumors produce disproportionate amounts of the hCG α or hCG β subunit. High levels of hCG rarely cause hyperthyroidism because of weak binding to the TSH receptor. [c]Calcitonin is produced eutopically by medullary thyroid carcinoma and is used as a tumor marker.

HYPERCALCEMIA CAUSED BY ECTOPIC PRODUCTION OF PTHRP

(See also Chap. 403)

Etiology

Humoral hypercalcemia of malignancy (HHM) occurs in up to 20% of patients with cancer. HHM is most common in cancers of the lung, head and neck, skin, esophagus, breast, and genitourinary tract and in multiple myeloma and lymphomas. There are several distinct humoral causes of HHM, but it is caused most commonly by overproduction of PTHrP. In addition to acting as a circulating humoral factor, bone metastases (e.g., breast, multiple myeloma) may produce PTHrP, leading to local osteolysis and hypercalcemia. PTHrP may also affect the initiation and progression of tumors by acting through prosurvival and chemokine pathways.

PTHrP is structurally related to PTH and binds to the PTH receptor, explaining the similar biochemical features of HHM and hyperparathyroidism. PTHrP plays a key role in skeletal development and regulates cellular proliferation and differentiation in other tissues, including skin, bone marrow, breast, and hair follicles. The mechanism of PTHrP induction in malignancy is incompletely understood; however, tumor-bearing tissues commonly associated with HHM normally produce PTHrP during development or cell renewal. PTHrP expression is stimulated by hedgehog pathways and Gli transcription factors that are active in many malignancies. Transforming growth factor β (TGF-β), which is produced by many tumors, also stimulates PTHrP. Mutations in certain oncogenes, such as *Ras*, also can activate PTHrP expression, as does loss of the tumor suppressor, p53. In addition to its role in HHM, the PTHrP pathway may also provide a potential target for therapeutic intervention to impede cancer growth.

Another relatively common cause of HHM is excess production of 1,25-dihydroxyvitamin D. Like granulomatous disorders associated with hypercalcemia, lymphomas can produce an enzyme that converts 25-hydroxyvitamin D to the more active 1,25-dihydroxyvitamin D, leading to enhanced gastrointestinal calcium absorption. Other causes of HHM include tumor-mediated production of osteolytic cytokines and inflammatory mediators.

Clinical Manifestations

The typical presentation of HHM is a patient with a known malignancy who is found to be hypercalcemic on routine laboratory tests. Less often, hypercalcemia is the initial presenting feature of malignancy. Particularly

when calcium levels are markedly increased (>3.5 mmol/L [>14 mg/dL]), patients may experience fatigue, mental status changes, dehydration, or symptoms of nephrolithiasis.

Diagnosis

Features that favor HHM, as opposed to primary hyperparathyroidism, include known malignancy, recent onset of hypercalcemia, and very high serum calcium levels. Like hyperparathyroidism, hypercalcemia caused by PTHrP is accompanied by hypercalciuria and hypophosphatemia. Patients with HHM typically have metabolic alkalosis rather than hyperchloremic acidosis, as is seen in hyperparathyroidism. Measurement of PTH is useful to exclude primary hyperparathyroidism; the PTH level should be suppressed in HHM. An elevated PTHrP level confirms the diagnosis, and it is increased in ~80% of hypercalcemic patients with cancer. 1,25-Dihydroxyvitamin D levels may be increased in patients with lymphoma.

TREATMENT

TREATMENT

Humoral Hypercalcemia of Malignancy

The management of HHM begins with removal of excess calcium in the diet, medications, or IV solutions. Saline rehydration (typically 200–500 mL/h) is used to dilute serum calcium and promote calciuresis; exercise caution in patients with cardiac, hepatic, or renal insufficiency. Forced diuresis with furosemide (20–80 mg IV in escalating doses) or other loop diuretics can enhance calcium excretion but provides relatively little value except in life-threatening hypercalcemia. When used, loop diuretics should be administered only after complete rehydration and with careful monitoring of fluid balance. Oral phosphorus (e.g., 250 mg Neutra-Phos 3–4 times daily) should be given until serum phosphorus is >1 mmol/L (>3 mg/dL). Bisphosphonates such as pamidronate (60–90 mg IV), zoledronate (4–8 mg IV), and etidronate (7.5 mg/kg per day PO for 3–7 consecutive days) can reduce serum calcium within 1–2 days and suppress calcium release for several weeks. Bisphosphonate infusions can be repeated, or oral bisphosphonates can be used for chronic treatment. Dialysis should be considered in severe hypercalcemia when saline hydration and bisphosphonate treatments are not possible or are too slow in onset. Previously used agents such as calcitonin and mithramycin have little utility now that bisphosphonates are available. Calcitonin (2–8 U/kg SC every 6–12 h) should be considered when rapid correction of severe hypercalcemia is needed. Hypercalcemia associated with lymphomas, multiple myeloma, or leukemia may respond to glucocorticoid treatment (e.g., prednisone 40–100 mg PO in four divided doses).

ECTOPIC VASOPRESSIN: TUMOR-ASSOCIATED SIADH

(See also Chap. 49)

Etiology

Vasopressin is an antidiuretic hormone normally produced by the posterior pituitary gland. Ectopic vasopressin production by tumors is a common cause of the syndrome of inappropriate antidiuretic hormone (SIADH), occurring in at least half of patients with SCLC. SIADH also can be caused by a number of nonneoplastic conditions, including central nervous system (CNS) trauma, infections, and medications (Chap. 374). Compensatory responses to SIADH, such as decreased thirst, may mitigate the development of hyponatremia. However, with prolonged production of excessive vasopressin, the osmostat controlling thirst and hypothalamic vasopressin secretion may become reset. In addition, intake of free water, orally or intravenously, can quickly worsen hyponatremia because of reduced renal diuresis.

Tumors with neuroendocrine features, such as SCLC and carcinoids, are the most common sources of ectopic vasopressin production, but it also occurs in other forms of lung cancer and with CNS lesions, head and neck cancer, and genitourinary, gastrointestinal, and ovarian cancers. The mechanism of activation of the vasopressin gene in these tumors is unknown but often involves concomitant expression of the adjacent oxytocin gene, suggesting derepression of this locus.

Clinical Manifestations

Most patients with ectopic vasopressin secretion are asymptomatic and are identified because of the presence of hyponatremia on routine chemistry testing. Symptoms may include weakness, lethargy, nausea, confusion, depressed mental status, and seizures. The severity of symptoms reflects the rapidity of onset as well as the severity of hyponatremia. Hyponatremia usually develops slowly but may be exacerbated by the administration of IV fluids or the institution of new medications.

Diagnosis

The diagnostic features of ectopic vasopressin production are the same as those of other causes of SIADH (Chaps. 49 and 374). Hyponatremia and reduced serum osmolality occur in the setting of an inappropriately normal or increased urine osmolality. Urine sodium excretion is normal or increased unless volume depletion is present. Other causes of hyponatremia should be excluded, including renal, adrenal, or thyroid insufficiency. Physiologic sources of vasopressin stimulation (CNS lesions, pulmonary disease, nausea), adaptive circulatory mechanisms (hypotension, heart failure, hepatic cirrhosis), and medications, including many chemotherapeutic agents, also should be considered as possible causes of hyponatremia. Vasopressin measurements are not usually necessary to make the diagnosis.

TREATMENT

TREATMENT

Ectopic Vasopressin: Tumor-Associated SIADH

Most patients with ectopic vasopressin production develop hyponatremia over several weeks or months. The disorder should be corrected gradually unless mental status is altered or there is risk of seizures. Treatment of the underlying malignancy may reduce ectopic vasopressin production, but this response is slow if it

occurs at all. Fluid restriction to less than urine output, plus insensible losses, is often sufficient to correct hyponatremia partially. However, strict monitoring of the amount and types of liquids consumed or administered intravenously is required for fluid restriction to be effective. Salt tablets and saline are not helpful unless volume depletion is also present. Demeclocycline (150–300 mg orally three to four times daily) can be used to inhibit vasopressin action on the renal distal tubule, but its onset of action is relatively slow (1–2 weeks). The vaptan class of drugs act by inhibiting vasopressin receptors (V_{1A}, V_{1B}, V_2). Conivaptan, a nonpeptide V_2-receptor antagonist, can be administered either PO (20–120 mg bid) or IV (10–40 mg) and is particularly effective when used in combination with fluid restriction in euvolemic hyponatremia. Tolvaptan (15 mg PO daily) is another vasopressin antagonist. The dose can be increased to 30–60 mg/d based on response. Severe hyponatremia (Na <115 meq/L) or mental status changes may require treatment with hypertonic (3%) or normal saline infusion together with furosemide to enhance free water clearance. The rate of sodium correction should be slow (0.5–1 meq/L per hour) to prevent rapid fluid shifts and the possible development of central pontine myelinolysis.

CUSHING'S SYNDROME CAUSED BY ECTOPIC ACTH PRODUCTION

(See also Chap. 379)

Etiology

Ectopic ACTH production accounts for 10–20% of cases of Cushing's syndrome. The syndrome is particularly common in neuroendocrine tumors. SCLC is the most common cause of ectopic ACTH, followed by bronchial and thymic carcinoids, islet cell tumors, other carcinoids, and pheochromocytomas. Ectopic ACTH production is caused by increased expression of the proopiomelanocortin (*POMC*) gene, which encodes ACTH, along with melanocyte-stimulating hormone (MSH), β lipotropin, and several other peptides. In many tumors, there is abundant but aberrant expression of the *POMC* gene from an internal promoter, proximal to the third exon, which encodes ACTH. However, because this product lacks the signal sequence necessary for protein processing, it is not secreted. Increased production of ACTH arises instead from less abundant, but unregulated, *POMC* expression from the same promoter site used in the pituitary. Because tumors lack many of the enzymes needed to process the POMC polypeptide, it is typically released as multiple large, biologically inactive fragments along with relatively small amounts of fully processed, active ACTH.

Rarely, corticotropin-releasing hormone (CRH) is produced by pancreatic islet cell tumors, SCLC, medullary thyroid cancer, carcinoids, or prostate cancer. When levels are high enough, CRH can cause pituitary corticotrope hyperplasia and Cushing's syndrome. Tumors that produce CRH sometimes also produce ACTH, raising the possibility of a paracrine mechanism for ACTH production.

A distinct mechanism for ACTH-independent Cushing's syndrome involves ectopic expression of various G protein–coupled receptors in the adrenal nodules. Ectopic expression of the gastric inhibitory peptide (GIP) receptor is the best-characterized example of this mechanism. In this case, meals induce GIP secretion, which inappropriately stimulates adrenal growth and glucocorticoid production.

Clinical Manifestations

The clinical features of hypercortisolemia are detected in only a fraction of patients with documented ectopic ACTH production. Patients with ectopic ACTH syndrome generally exhibit less marked weight gain and centripetal fat redistribution, probably because the exposure to excess glucocorticoids is relatively brief and because cachexia reduces the propensity for weight gain and fat deposition. The ectopic ACTH syndrome is associated with several clinical features that distinguish it from other causes of Cushing's syndrome (e.g., pituitary adenomas, adrenal adenomas, iatrogenic glucocorticoid excess). The metabolic manifestations of ectopic ACTH syndrome are dominated by fluid retention and hypertension, hypokalemia, metabolic alkalosis, glucose intolerance, and occasionally steroid psychosis. The very high ACTH levels often cause increased pigmentation, reflecting increased activity of MSH derived from the POMC precursor peptide. The extraordinarily high glucocorticoid levels in patients with ectopic sources of ACTH can lead to marked skin fragility and easy bruising. In addition, the high cortisol levels often overwhelm the renal 11β-hydroxysteroid dehydrogenase type II enzyme, which normally inactivates cortisol and prevents it from binding to renal mineralocorticoid receptors. Consequently, in addition to the excess mineralocorticoids produced by ACTH stimulation of the adrenal gland, high levels of cortisol exert activity through the mineralocorticoid receptor, leading to severe hypokalemia.

Diagnosis

The diagnosis of ectopic ACTH syndrome is usually not difficult in the setting of a known malignancy. Urine-free cortisol levels fluctuate but are typically greater than two to four times normal, and the plasma ACTH level is usually >22 pmol/L (>100 pg/mL). A suppressed ACTH level excludes this diagnosis and indicates an ACTH-independent cause of Cushing's syndrome (e.g., adrenal or exogenous glucocorticoid). In contrast to pituitary sources of ACTH, most ectopic sources of ACTH do not respond to glucocorticoid suppression. Therefore, high-dose dexamethasone (8 mg PO) suppresses 8:00 A.M. serum cortisol (50% decrease from baseline) in ~80% of pituitary ACTH-producing adenomas but fails to suppress ectopic ACTH in ~90% of cases. Bronchial and other carcinoids are well-documented exceptions to these general guidelines, as these ectopic sources of ACTH may exhibit feedback regulation indistinguishable from pituitary adenomas, including suppression by high-dose dexamethasone, and ACTH responsiveness to adrenal blockade with metyrapone. If necessary, petrosal sinus catheterization can be used to evaluate a patient with ACTH-dependent Cushing's syndrome when the source of ACTH is unclear. After CRH stimulation, a 3:1 petrosal sinus:peripheral ACTH ratio strongly suggests a pituitary ACTH source. Imaging studies (computed tomography or magnetic resonance imaging) are also useful in the evaluation of suspected carcinoid lesions, allowing biopsy and characterization of hormone production using special stains. If available, positron emission tomography or octreotide scanning may identify some sources of ACTH production.

TREATMENT

TREATMENT

Cushing's Syndrome Caused by Ectopic ACTH Production

The morbidity associated with the ectopic ACTH syndrome can be substantial. Patients may experience depression or personality changes because of extreme cortisol excess. Metabolic derangements, including diabetes mellitus and hypokalemia, can worsen fatigue. Poor wound healing and predisposition to infections can complicate the surgical management of tumors, and opportunistic infections caused by organisms such as *Pneumocystis carinii* and mycoses are often the cause of death in patients with ectopic ACTH production. These patients likely have increased risk of venous thromboembolism reflecting the combination of malignancy and altered coagulation factor profiles. Depending on prognosis and treatment plans for the underlying malignancy, measures to reduce cortisol levels are often indicated. Treatment of the underlying malignancy may reduce ACTH levels but is rarely sufficient to reduce cortisol levels to normal. Adrenalectomy is not practical for most of these patients but should be considered during surgery for the malignancy or if the underlying tumor is not resectable and the prognosis is otherwise favorable (e.g., carcinoid). Medical therapy with ketoconazole (300–600 mg PO bid), metyrapone (250–500 mg PO every 6 h), mitotane (3–6 g PO in four divided doses, tapered to maintain low cortisol production), or other agents that block steroid synthesis or action is often the most practical strategy for managing the hypercortisolism associated with ectopic ACTH production. Glucocorticoid replacement should be provided to prevent adrenal insufficiency **(Chap. 379)**. Unfortunately, many patients eventually progress despite medical blockade.

TUMOR-INDUCED HYPOGLYCEMIA CAUSED BY EXCESS PRODUCTION OF IGF-II

(See also Chap. 399) Mesenchymal tumors, hemangiopericytomas, hepatocellular tumors, adrenal carcinomas, and a variety of other large tumors have been reported to produce excessive amounts of insulin-like growth factor type II (IGF-II) precursor, which binds weakly to insulin receptors and more strongly to IGF-I receptors, leading to insulin-like actions. The gene encoding IGF-II resides on chromosome 11p15, a locus that is normally imprinted (that is, expression is exclusively from a single parental allele). Biallelic expression of the IGF-II gene occurs in a subset of tumors, suggesting loss of methylation and loss of imprinting as a mechanism for gene induction. In addition to increased IGF-II production, IGF-II bioavailability is increased due to complex alterations in circulating binding proteins. Increased IGF-II suppresses growth hormone (GH) and insulin, resulting in reduced IGF binding protein 3 (IGFBP-3), IGF-I, and acid-labile subunit (ALS). The reduction in ALS and IGFBP-3, which normally sequester IGF-II, causes it to be displaced to a small circulating complex that has greater access to insulin target tissues. For this reason, circulating IGF-II levels may not be markedly increased despite causing hypoglycemia. In addition to IGF-II–mediated hypoglycemia, tumors may occupy enough of the liver to impair gluconeogenesis.

In most cases, a tumor causing hypoglycemia is clinically apparent (usually >10 cm in size), and hypoglycemia develops in association with fasting. The diagnosis is made by documenting low serum glucose and suppressed insulin levels in association with symptoms of hypoglycemia. Serum IGF-II levels may not be increased (IGF-II assays may not detect IGF-II precursors). Increased IGF-II mRNA expression is found in most of these tumors. Any medications associated with hypoglycemia should be eliminated. Treatment of the underlying malignancy, if possible, may reduce the predisposition to hypoglycemia.

Frequent meals and IV glucose, especially during sleep or fasting, are often necessary to prevent hypoglycemia. Glucagon and glucocorticoids have also been used to enhance glucose production.

HUMAN CHORIONIC GONADOTROPIN

hCG is composed of α and β subunits and can be produced as intact hormone, which is biologically active, or as uncombined biologically inert subunits. Ectopic production of intact hCG occurs most often in association with testicular embryonal tumors, germ cell tumors, extragonadal germinomas, lung cancer, hepatoma, and pancreatic islet tumors. Eutopic production of hCG occurs with trophoblastic malignancies. hCG α subunit production is particularly common in lung cancer and pancreatic islet cancer. In men, high hCG levels stimulate steroidogenesis and aromatase activity in testicular Leydig cells, resulting in increased estrogen production and the development of gynecomastia. Precocious puberty in boys or gynecomastia in men should prompt measurement of hCG and consideration of a testicular tumor or another source of ectopic hCG production. Most women are asymptomatic. hCG is easily measured. Treatment should be directed at the underlying malignancy.

ONCOGENIC OSTEOMALACIA

Hypophosphatemic oncogenic osteomalacia, also called tumor-induced osteomalacia (TIO), is characterized by markedly reduced serum phosphorus and renal phosphate wasting, leading to muscle weakness, bone pain, and osteomalacia. Serum calcium and PTH levels are normal, and 1,25-dihydroxyvitamin D is low. Oncogenic osteomalacia is usually caused by benign mesenchymal tumors, such as hemangiopericytomas, fibromas, and giant cell tumors, often of the skeletal extremities or head. It has also been described in sarcomas and in patients with prostate or lung cancer. Resection of the tumor reverses the disorder, confirming its humoral basis. The circulating phosphaturic factor was originally called *phosphatonin*—a factor that inhibits renal tubular reabsorption of phosphate and renal conversion of 25-hydroxyvitamin D to 1,25-dihydroxyvitamin D. Phosphatonin has been identified as fibroblast growth factor 23 (FGF23). FGF23 levels are increased in some, but not all, patients with osteogenic osteomalacia. FGF23 forms a ternary complex with the klotho protein and renal FGF receptors to reduce renal phosphate reabsorption. Treatment involves removal of the tumor, if possible, and supplementation with phosphate and vitamin D. Octreotide treatment reduces phosphate wasting in some patients with tumors that express somatostatin receptor subtype 2. Octreotide scans may also be useful in detecting these tumors.

CONSUMPTIVE HYPOTHYROIDISM

Newborns with hepatic hemangiomas can develop a rare form of hypothyroidism caused by overexpression of type 3 deiodinase (D3), an enzyme that degrades and inactivates T4 and T3. The very high expression of D3, and consumption of thyroid hormones, apparently outstrips the thyroid gland's rate of hormone production. The disorder is characterized by low T4, low T3, high TSH, and markedly elevated reverse T3 (rT3), reflecting the degradation of T4 to rT3. In addition to treating the underlying hemangioma (rarely other tumor types), patients are treated with l-thyroxine replacement, titrated to normalize TSH. Steroids and

propranolol may provide benefit, perhaps by inhibiting growth factor pathways thought to stimulate D3 production.

HEMATOLOGIC SYNDROMES

The elevation of granulocyte, platelet, and eosinophil counts in most patients with myeloproliferative disorders is caused by the proliferation of the myeloid elements due to the underlying disease rather than to a paraneoplastic syndrome. The paraneoplastic hematologic syndromes in patients with solid tumors are less well characterized than are the endocrine syndromes because the ectopic hormone(s) or cytokines responsible have not been identified in most of these tumors **(Table 89-2)**. The extent of the paraneoplastic syndromes parallels the course of the cancer.

TABLE 89-2

Paraneoplastic Hematologic Syndromes

Syndrome	Proteins	Cancers Typically Associated with Syndrome
Erythrocytosis	Erythropoietin	Renal cancers, hepatocarcinoma, cerebellar hemangioblastomas
Granulocytosis	G-CSF, GM-CSF, IL-6	Lung cancer, gastrointestinal cancer, ovarian cancer, genitourinary cancer, Hodgkin's disease
Thrombocytosis	IL-6	Lung cancer, gastrointestinal cancer, breast cancer, ovarian cancer, lymphoma
Eosinophilia	IL-5	Lymphoma, leukemia, lung cancer
Thrombophlebitis	Unknown	Lung cancer, pancreatic cancer, gastrointestinal cancer, breast cancer, genitourinary cancer, ovarian cancer, prostate cancer, lymphoma

Abbreviations: G-CSF, granulocyte colony-stimulating factor; GM-CSF, granulocyte-macrophage colony-stimulating factor; IL, interleukin.

ERYTHROCYTOSIS

Ectopic production of erythropoietin by cancer cells causes most paraneoplastic erythrocytosis. The ectopically produced erythropoietin stimulates the production of red blood cells (RBCs) in the bone marrow and raises the hematocrit. Other lymphokines and hormones produced by cancer cells may stimulate erythropoietin release but have not been proved to cause erythrocytosis.

Most patients with erythrocytosis have an elevated hematocrit (>52% in men, >48% in women) that is detected on a routine blood count. Approximately 3% of patients with renal cell cancer, 10% of patients with hepatoma, and 15% of patients with cerebellar hemangioblastomas have erythrocytosis. In most cases, the erythrocytosis is asymptomatic.

Patients with erythrocytosis due to a renal cell cancer, hepatoma, or CNS cancer should have measurement of red cell mass. If the red cell mass is elevated, the serum erythropoietin level should be measured. Patients with an appropriate cancer, elevated erythropoietin levels, and no other explanation for erythrocytosis (e.g., hemoglobinopathy that causes increased O_2 affinity; **Chap. 59**) have the paraneoplastic syndrome.

TREATMENT

TREATMENT

Erythrocytosis

Successful resection of the cancer usually resolves the erythrocytosis. If the tumor cannot be resected or treated effectively with radiation therapy or chemotherapy, phlebotomy may control any symptoms related to erythrocytosis.

GRANULOCYTOSIS

Approximately 30% of patients with solid tumors have granulocytosis (granulocyte count >8000/μL). In about half of patients with granulocytosis and cancer, the granulocytosis has an identifiable nonparaneoplastic etiology (infection, tumor necrosis, glucocorticoid administration, etc.). The other patients have proteins in urine and serum that stimulate the growth of bone marrow cells. Tumors and tumor cell lines from patients with lung, ovarian, and bladder cancers have been documented to produce granulocyte colony-stimulating factor (G-CSF), granulocyte-macrophage colony-stimulating factor (GM-CSF), and/or interleukin 6 (IL-6). However, the etiology of granulocytosis has not been characterized in most patients.

Patients with granulocytosis are nearly all asymptomatic, and the differential white blood cell count does not have a shift to immature forms of neutrophils. Granulocytosis occurs in 40% of patients with lung and gastrointestinal cancers, 20% of patients with breast cancer, 30% of patients with brain tumors and ovarian cancers, 20% of patients with Hodgkin's disease, and 10% of patients with renal cell carcinoma. Patients with advanced-stage disease are more likely to have granulocytosis than are those with early-stage disease.

Paraneoplastic granulocytosis does not require treatment. The granulocytosis resolves when the underlying cancer is treated.

THROMBOCYTOSIS

Some 35% of patients with thrombocytosis (platelet count >400,000/μL) have an underlying diagnosis of cancer. IL-6, a candidate molecule for the etiology of paraneoplastic thrombocytosis, stimulates the

production of platelets in vitro and in vivo. Some patients with cancer and thrombocytosis have elevated levels of IL-6 in plasma. Another candidate molecule is thrombopoietin, a peptide hormone that stimulates megakaryocyte proliferation and platelet production. The etiology of thrombocytosis has not been established in most cases.

Patients with thrombocytosis are nearly all asymptomatic. Thrombocytosis is not clearly linked to thrombosis in patients with cancer. Thrombocytosis is present in 40% of patients with lung and gastrointestinal cancers; 20% of patients with breast, endometrial, and ovarian cancers; and 10% of patients with lymphoma. Patients with thrombocytosis are more likely to have advanced-stage disease and have a poorer prognosis than do patients without thrombocytosis. In ovarian cancer, IL-6 has been shown to directly promote tumor growth. Paraneoplastic thrombocytosis does not require treatment other than treatment of the underlying tumor.

EOSINOPHILIA

Eosinophilia is present in ~1% of patients with cancer. Tumors and tumor cell lines from patients with lymphomas or leukemia may produce IL-5, which stimulates eosinophil growth. Activation of IL-5 transcription in lymphomas and leukemias may involve translocation of the long arm of chromosome 5, to which the genes for IL-5 and other cytokines map.

Patients with eosinophilia are typically asymptomatic. Eosinophilia is present in 10% of patients with lymphoma, 3% of patients with lung cancer, and occasional patients with cervical, gastrointestinal, renal, and breast cancer. Patients with markedly elevated eosinophil counts (>5000/μL) can develop shortness of breath and wheezing. A chest radiograph may reveal diffuse pulmonary infiltrates from eosinophil infiltration and activation in the lungs.

TREATMENT

TREATMENT

Eosinophilia

Definitive treatment is directed at the underlying malignancy: Tumors should be resected or treated with radiation or chemotherapy. In most patients who develop shortness of breath related to eosinophilia, symptoms resolve with the use of oral or inhaled glucocorticoids. IL-5 antagonists exist but have not been evaluated in this clinical setting.

THROMBOPHLEBITIS AND DEEP VENOUS THROMBOSIS

Deep venous thrombosis and pulmonary embolism are the most common thrombotic conditions in patients with cancer. Migratory or recurrent thrombophlebitis may be the initial manifestation of cancer. Nearly 15% of patients who develop deep venous thrombosis or pulmonary embolism have a diagnosis of cancer (Chap.

113). The coexistence of peripheral venous thrombosis with visceral carcinoma, particularly pancreatic cancer, is called *Trousseau's syndrome*.

Pathogenesis

Patients with cancer are predisposed to thromboembolism because they are often at bed rest or immobilized, and tumors may obstruct or slow blood flow. Postoperative deep venous thrombosis is twice as common in cancer patients who undergo surgery. Chronic IV catheters also predispose to clotting. In addition, clotting may be promoted by release of procoagulants or cytokines from tumor cells or associated inflammatory cells or by platelet adhesion or aggregation. The specific molecules that promote thromboembolism have not been identified.

Chemotherapeutic agents, particularly those associated with endothelial damage, can induce venous thrombosis. The annual risk of venous thrombosis in patients with cancer receiving chemotherapy is about 11%, sixfold higher than the risk in the general population. Bleomycin, L-asparaginase, nitrogen mustard, thalidomide analogues, cisplatin-based regimens, and high doses of busulfan and carmustine are all associated with an increased risk.

In addition to cancer and its treatment causing secondary thrombosis, primary thrombophilic diseases may be associated with cancer. For example, the antiphospholipid antibody syndrome is associated with a wide range of pathologic manifestations (Chap. 350). About 20% of patients with this syndrome have cancers. Among patients with cancer and antiphospholipid antibodies, 35–45% develop thrombosis.

Clinical Manifestations

Patients with cancer who develop deep venous thrombosis usually develop swelling or pain in the leg, and physical examination reveals tenderness, warmth, and redness. Patients who present with pulmonary embolism develop dyspnea, chest pain, and syncope, and physical examination shows tachycardia, cyanosis, and hypotension. Some 5% of patients with no history of cancer who have a diagnosis of deep venous thrombosis or pulmonary embolism will have a diagnosis of cancer within 1 year. The most common cancers associated with thromboembolic episodes include lung, pancreatic, gastrointestinal, breast, ovarian, and genitourinary cancers; lymphomas; and brain tumors. Patients with cancer who undergo surgical procedures requiring general anesthesia have a 20–30% risk of deep venous thrombosis.

Diagnosis

The diagnosis of deep venous thrombosis in patients with cancer is made by impedance plethysmography or bilateral compression ultrasonography of the leg veins. Patients with a noncompressible venous segment have deep venous thrombosis. If compression ultrasonography is normal and there is a high clinical suspicion for deep venous thrombosis, venography should be done to look for a luminal filling defect. Elevation of D-dimer is not as predictive of deep venous thrombosis in patients with cancer as it is in patients

without cancer; elevations are seen in people over age 65 years without concomitant evidence of thrombosis, probably as a consequence of increased thrombin deposition and turnover in aging.

Patients with symptoms and signs suggesting a pulmonary embolism should be evaluated with a chest radiograph, electrocardiogram, arterial blood gas analysis, and ventilation-perfusion scan. Patients with mismatched segmental perfusion defects have a pulmonary embolus. Patients with equivocal ventilation-perfusion findings should be evaluated as described above for deep venous thrombosis in their legs. If deep venous thrombosis is detected, they should be anticoagulated. If deep venous thrombosis is not detected, they should be considered for a pulmonary angiogram.

Patients without a diagnosis of cancer who present with an initial episode of thrombophlebitis or pulmonary embolus need no additional tests for cancer other than a careful history and physical examination. In light of the many possible primary sites, diagnostic testing in asymptomatic patients is wasteful. However, if the clot is refractory to standard treatment or is in an unusual site or if the thrombophlebitis is migratory or recurrent, efforts to find an underlying cancer are indicated.

TREATMENT

TREATMENT

Thrombophlebitis and Deep Venous Thrombosis

Patients with cancer and a diagnosis of deep venous thrombosis or pulmonary embolism should be treated initially with IV unfractionated heparin or low-molecular-weight heparin for at least 5 days, and warfarin should be started within 1 or 2 days. The warfarin dose should be adjusted so that the international normalized ratio (INR) is 2–3. Patients with proximal deep venous thrombosis and a relative contraindication to heparin anticoagulation (hemorrhagic brain metastases or pericardial effusion) should be considered for placement of a filter in the inferior vena cava (Greenfield filter) to prevent pulmonary embolism. Warfarin should be administered for 3–6 months. An alternative approach is to use low-molecular-weight heparin for 6 months. The new oral anticoagulants (factor Xa and thrombin inhibitors) are attractive because they do not require close monitoring of the prothrombin time and are not affected by dietary factors. However, data on their use in patients with cancer are not yet mature. Patients with cancer who undergo a major surgical procedure should be considered for heparin prophylaxis or pneumatic boots. Breast cancer patients undergoing chemotherapy and patients with implanted catheters should be considered for prophylaxis. Guidelines recommend that hospitalized patients with cancer and patients receiving a thalidomide analogue receive prophylaxis with low-molecular-weight heparin or low-dose aspirin. Use of prophylaxis routinely during chemotherapy is controversial and not recommended by the American Society of Clinical Oncology.

MISCELLANEOUS REMOTE EFFECTS OF CANCER

Patients with cancer can develop paraneoplastic autoimmune disorders (e.g., thrombocytopenia) and dysfunction of organs not directly invaded or involved with the cancer (rheumatologic and renal abnormalities are among the most frequent). The pathogenesis of these disorders is undefined, but often the conditions reverse if the tumor is removed or successfully treated.

Cutaneous paraneoplastic syndromes are discussed in Chap. 54. Neurologic paraneoplastic syndromes are discussed in Chap. 90.

FURTHER READING

Chong WH et al: Tumor-induced osteomalacia. Endocr Relat Cancer 18:R53, 2012.

De Groot JWB et al: Non-islet cell tumour-induced hypoglycaemia: A review of the literature including two new cases. Endocr Relat Cancer 14:979, 2007. [PubMed: 18045950]

Ellison DH, Berl T: The syndrome of inappropriate antidiuresis. N Engl J Med 356:2064, 2007. [PubMed: 17507705]

Isidori AM et al: The ectopic adrenocorticotropin syndrome: Clinical features, diagnosis, management and long-term follow-up. J Clin Endocrinol Metab 91:371, 2006. [PubMed: 16303835]

Jones PA, Baylin SB: The fundamental role of epigenetic events in cancer. Nat Rev Genet 3:415, 2002. [PubMed: 12042769]

Lin RJ et al: Paraneoplastic thrombocytosis: The secrets of tumor self-promotion. Blood 124:184, 2014. [PubMed: 24868077]

Mundy GR, Edwards JR: PTH-related peptide (PTHrP) in hypercalcemia. J Am Soc Nephrol 19:672, 2008. [PubMed: 18256357]

Rikhof B et al: The insulin-like growth factor system and sarcomas. J Pathol 217:469, 2009. [PubMed: 19148905]

Shaikh A et al: Regulation of phosphate homeostasis by the phosphatonins and other novel mediators. Pediatr Nephrol 23:1203, 2008. [PubMed: 18288501]

Stewart AF: Clinical practice. Hypercalcemia associated with cancer. N Engl J Med 352:373, 2005. [PubMed: 15673803]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 90: Paraneoplastic Neurologic Syndromes and Autoimmune Encephalitis

Josep Dalmau; Myrna R. Rosenfeld

INTRODUCTION

Paraneoplastic neurologic disorders (PNDs) are cancer-related syndromes that can affect any part of the nervous system (Table 90-1). They are caused by mechanisms other than metastasis or by any of the complications of cancer such as coagulopathy, stroke, metabolic and nutritional conditions, infections, and side effects of cancer therapy. In 60% of patients, the neurologic symptoms precede the cancer diagnosis. Clinically disabling PNDs occur in 0.5–1% of all cancer patients, but they affect 2–3% of patients with neuroblastoma or small-cell lung cancer (SCLC) and 30–50% of patients with thymoma.

TABLE 90-1

Paraneoplastic Syndromes of the Nervous System

Classic Syndromes: Usually Occur with Cancer Association	Nonclassic Syndromes: May Occur with and without Cancer Association
Encephalomyelitis	Brainstem encephalitis
Limbic encephalitis	Stiff-person syndrome
Cerebellar degeneration (adults)	Progressive encephalomyelitis with rigidity and myoclonus
Opsoclonus-myoclonus	Necrotizing myelopathy
Subacute sensory neuronopathy	Motor neuron disease
Gastrointestinal paresis or pseudo-obstruction	Guillain-Barré syndrome
Dermatomyositis (adults)	Subacute and chronic mixed sensory-motor neuropathies
Lambert-Eaton myasthenic syndrome	Neuropathy associated with plasma cell dyscrasias and lymphoma
Cancer- or melanoma-associated retinopathy	Vasculitis of nerve
	Pure autonomic neuropathy
	Acute necrotizing myopathy
	Polymyositis
	Optic neuropathy
	BDUMP
	Peripheral nerve hyperexcitability (neuromyotonia)
	Myasthenia gravis

Abbreviation: BDUMP, bilateral diffuse uveal melanocytic proliferation.

PATHOGENESIS

Most PNDs are mediated by immune responses triggered by neuronal proteins (onconeuronal antigens) expressed by tumors. In PNDs of the central nervous system (CNS), many antibody-associated immune responses have been identified **(Table 90-2)**. These antibodies react with both neurons and the patient's tumor, and their detection in serum or cerebrospinal fluid (CSF) usually predicts the presence of cancer. When the antigens are intracellular, most syndromes are associated with extensive infiltrates of CD4+ and CD8+ T cells, microglial activation, gliosis, and variable neuronal loss. The infiltrating T cells are often in close contact with neurons undergoing degeneration, suggesting a primary pathogenic role. T cell–mediated cytotoxicity may contribute directly to cell death in these PNDs. Thus both humoral and cellular immune mechanisms participate in the pathogenesis of many PNDs. This complex immunopathogenesis may underlie the resistance of many of these conditions to therapy.

TABLE 90-2

Antibodies to Intracellular Antigens, Syndromes, and Associated Cancers

Antibody	Associated Neurologic Syndrome(s)	Tumors
Anti-Hu (ANNA1)	Encephalomyelitis, subacute sensory neuronopathy	SCLC
Anti-Yo (PCA1)	Cerebellar degeneration	Ovary, breast
Anti-Ri (ANNA2)	Cerebellar degeneration, opsoclonus, brainstem encephalitis	Breast, gynecologic, SCLC
Anti-CRMP5 (CV2)	Encephalomyelitis, chorea, optic neuritis, uveitis, peripheral neuropathy	SCLC, thymoma, other
Anti-Ma proteins	Limbic, hypothalamic, brainstem encephalitis	Testicular (Ma2), other (Ma)
Anti-amphiphysin[a]	Stiff-person syndrome, encephalomyelitis	Breast, SCLC
Recoverin, bipolar cell antibodies, others[b]	Cancer-associated retinopathy (CAR) Melanoma-associated retinopathy (MAR)	SCLC (CAR), melanoma (MAR)
Anti-GAD	Stiff-person, cerebellar syndrome, limbic encephalitis	Infrequent tumor association (thymoma and several cancers)

[a]Amphiphysin is likely exposed to the cell surface during synaptic vesicle endocytosis. [b]A variety of target antigens have been identified.

Abbreviations: CRMP, collapsing response-mediator protein; SCLC, small-cell lung cancer.

In contrast to the disorders associated with immune responses against intracellular antigens, those associated with antibodies to antigens expressed on the neuronal cell surface of the CNS or at the neuromuscular junction are more responsive to immunotherapy **(Table 90-3, Fig. 90-1)**. These disorders occur with and without a cancer association and may affect children and young adults, and there is evidence demonstrating a pathogenic role of the antibodies.

TABLE 90-3

Antibodies to Cell Surface or Synaptic Antigens, Syndromes, and Associated Tumors

Antibody	Neurologic Syndrome	Tumor Type When Associated
Anti-AChR (muscle)[a]	Myasthenia gravis	Thymoma
Anti-AChR (neuronal)[a]	Autonomic ganglionopathy	SCLC
Anti-VGCC[a]	LEMS, cerebellar degeneration	SCLC
Anti-NMDAR[a]	Anti-NMDAR encephalitis	Teratoma in young women (children and men rarely have tumors)
Anti-LGI1[b]	Limbic encephalitis, hyponatremia, faciobrachial tonic or dystonic seizures	Rarely thymoma
Anti-Caspr2[b]	Morvan's syndrome, neuromyotonia, limbic encephalitis	Thymoma, prostate cancer
Anti-$GABA_BR$[c]	Limbic encephalitis, seizures	SCLC, neuroendocrine
Anti-$GABA_AR$[a]	Encephalitis with prominent seizures and status epilepticus; less often opsoclonus and stiff-person syndrome	Thymoma in ~30% of patients
Anti-AMPAR[a]	Limbic encephalitis with relapses	SCLC, thymoma, breast
Glycine receptor	PERM, stiff-person syndrome	Rarely, thymoma, lung, Hodgkin lymphoma
Anti-DPPX[a]	Agitation, myoclonus, tremor, seizures, hyperekplexia, encephalomyelitis with rigidity	No cancer, but frequent diarrhea or cachexia suggesting paraneoplasia
Anti-Neurexin 3alpha	Autoimmune encephalitis without distinctive features	No cancer association
Anti-Dopamine-2R	Basal ganglia encephalitis	No cancer association

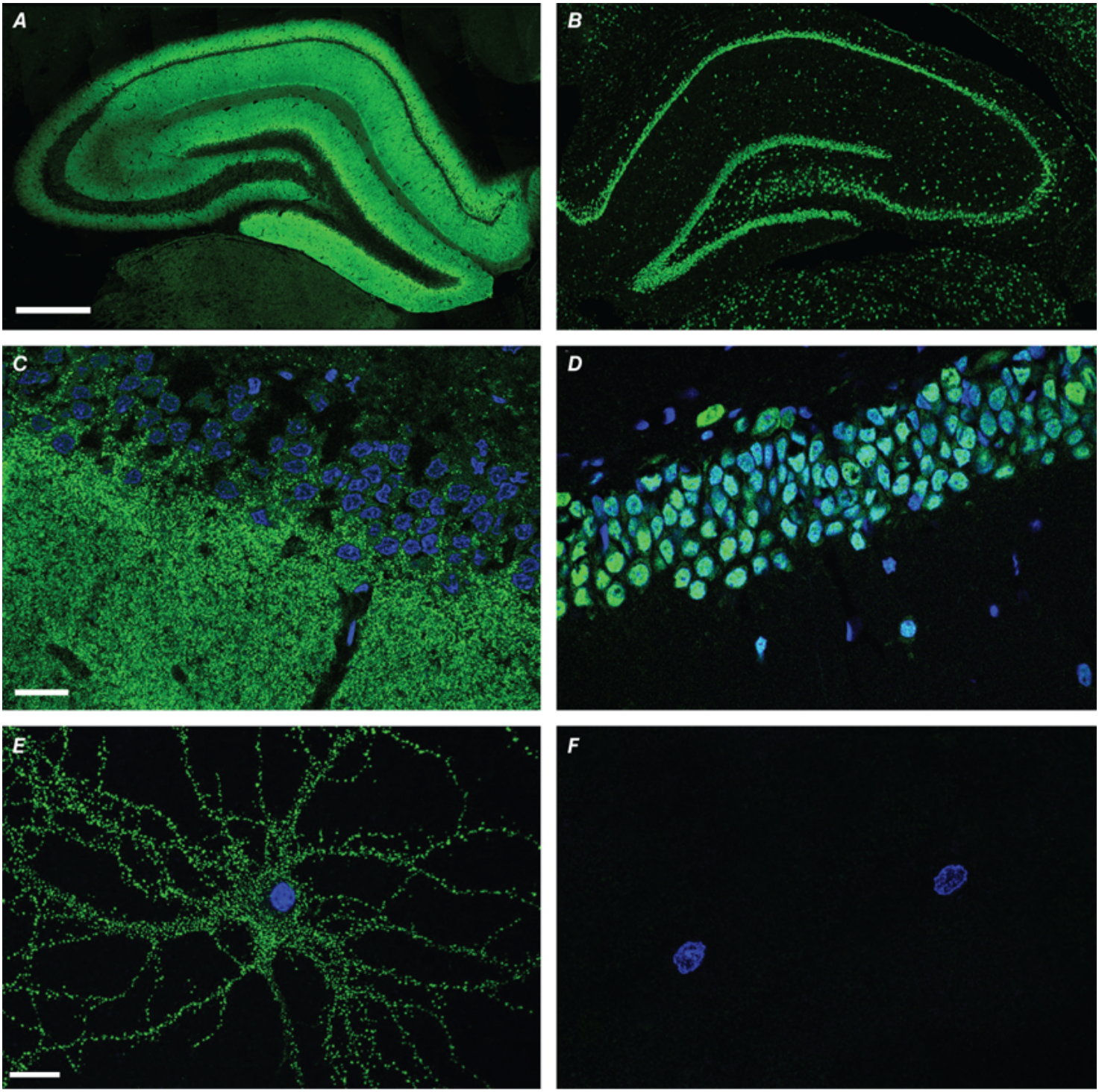
Antibody	Neurologic Syndrome	Tumor Type When Associated
Anti-Tr (DNER)	Cerebellar syndrome	Hodgkin lymphoma, or no tumor
Anti-mGluR1	Cerebellar syndrome	Hodgkin lymphoma, or no tumor
Anti-mGluR5	Autoimmune encephalitis without distinctive features	Hodgkin lymphoma, or no tumor
IgLON5	NREM and REM sleep disorder, and brainstem dysfunction	No tumor association

[a]A direct pathogenic role of these antibodies has been demonstrated in cultured neurons or animal models. [b]Previously named voltage-gated potassium channel antibodies (VGKC); currently included under the term VGKC-complex proteins. Of note, the significance of antibodies to VGKC-complex proteins other than LGI1 and Caspr2 is uncertain (the antigens are unknown, and the response to immunotherapy is variable). [c]This antibody is strongly suspected to be pathogenic.

Abbreviations: AChR, acetylcholine receptor; AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor; Caspr2, contactin-associated protein-like 2; DNER, delta/notch-like epidermal growth factor-related receptor; DPPX, dipeptidyl-peptidase-like protein-6; $GABA_BR$, γ-aminobutyric acid B receptor; GAD, glutamic acid decarboxylase; mGluR, metabotropic glutamate receptor; LEMS, Lambert-Eaton myasthenic syndrome; LGI1, leucine-rich glioma-inactivated 1; NMDAR, *N*-methyl-D-aspartate receptor; NREM, non-rapid eye movement; PERM, progressive encephalomyelitis with rigidity and myoclonus; REM, rapid eye movement; SCLC, small-cell lung cancer; VGCC, voltage-gated calcium channel.

FIGURE 90-1

Comparison of brain and neuronal reactivity of antibodies against a neuronal cell surface, and an intracellular antigen. Coronal section of rat hippocampus immunolabeled with antibodies against a cell surface antigen (*N*-methyl-D-aspartate; NMDA receptor) from a patient with anti-NMDA receptor encephalitis (***A***), compared with the antibodies against an intracellular antigen (Hu) from a patient with small-cell lung cancer and paraneoplastic encephalitis (***B***). Magnification of the reactivity is shown in ***C*** and ***D*** respectively. Compared with the NMDA receptor antibodies that show intense reactivity with the neuropil of hippocampus, the Hu antibodies only show intracellular reactivity after tissue permeabilization. In cultures of dissociated rat hippocampal neurons, only the NMDA receptor antibodies react with the antigen in live non-permeabilized neurons (***E***). The Hu antibodies do not reach the target intracellular antigen in live neurons (***F***). Scale bars: ***A, B*** = 500 µm; ***C, D*** = 20 µm; ***E, F*** = 10 µm. *(From J Dalmau et al: Physiol Rev 97:839, 2017, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Other PNDs are likely immune-mediated, although their antigens are unknown. These include several syndromes of inflammatory neuropathies and myopathies. In addition, many patients with typical PND syndromes are antibody-negative.

For still other PNDs, the cause remains quite obscure. These include, among others, several neuropathies that occur in the terminal stages of cancer and a number of neuropathies associated with plasma cell dyscrasias or lymphoma without evidence of inflammatory infiltrates or deposits of immunoglobulin, cryoglobulin, or amyloid.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Three key concepts are important for the diagnosis and management of PNDs. First, it is common for symptoms to appear before the presence of a tumor is known; second, the neurologic syndrome usually develops rapidly, producing severe deficits in a short period of time; and third, there is evidence that prompt tumor control improves the neurologic outcome. Therefore, the major concern of the physician is to recognize a disorder promptly as paraneoplastic and to identify and treat the tumor.

PND OF THE CENTRAL NERVOUS SYSTEM AND DORSAL ROOT GANGLIA

When symptoms involve brain, spinal cord, or dorsal root ganglia, the suspicion of PND is usually based on a combination of clinical, radiologic, and CSF findings. Presence of antineuronal antibodies (Tables 90-2 and 90-3) may help in the diagnosis, but only 60–70% of PNDs of the CNS and <20% of those involving the peripheral nervous system have neuronal or neuromuscular junction antibodies that can be used as diagnostic tests.

Magnetic resonance imaging (MRI) and CSF studies are important to rule out neurologic complications due to the direct spread of cancer, particularly metastatic and leptomeningeal disease. In most PNDs, the MRI findings are nonspecific. Paraneoplastic limbic encephalitis is usually associated with characteristic MRI abnormalities in the mesial temporal lobes (see below), but similar findings can occur with other disorders (e.g., nonparaneoplastic autoimmune limbic encephalitis and human herpesvirus type 6 [HHV-6] encephalitis) **(Fig. 90-2)**. The CSF profile of patients with PND of the CNS or dorsal root ganglia typically consists of mild to moderate pleocytosis (<200 mononuclear cells, predominantly lymphocytes), an increase in the protein concentration, and a variable presence of oligoclonal bands. There are no specific electrophysiologic tests that are diagnostic of PND. Moreover, a biopsy of the affected tissue is often difficult to obtain, and although useful to rule out other disorders (e.g., metastasis) the pathologic findings are not specific for PND.

PND OF NERVE AND MUSCLE

If symptoms involve peripheral nerve, neuromuscular junction, or muscle, the diagnosis of a specific PND is usually established on clinical, electrophysiologic, and pathologic grounds. The clinical history, accompanying symptoms (e.g., anorexia, weight loss), and type of syndrome dictate the studies and degree of effort needed to demonstrate a neoplasm. For example, the frequent association of Lambert-Eaton myasthenic syndrome (LEMS) with SCLC should lead to a chest and abdomen computed tomography (CT) or body positron emission tomography (PET) scan and, if negative, periodic tumor screening for at least 3 years after the neurologic diagnosis. In contrast, the weak association of polymyositis with cancer calls into question, the need for repeated cancer screenings in this situation. Serum and urine immunofixation studies should be considered in patients with peripheral neuropathy of unknown cause; detection of a monoclonal gammopathy suggests the need for additional studies to uncover a B cell or plasma cell malignancy. In paraneoplastic neuropathies, diagnostically useful antineuronal antibodies are limited to CRMP5 and Hu (ANNA1).

For any type of PND, if antineuronal antibodies are negative, the diagnosis relies on the demonstration of cancer and the exclusion of other cancer-related or independent neurologic disorders. Combined CT and PET scans often uncover tumors undetected by other tests. For germ cell tumors of the testis and teratomas of the ovary, ultrasound and CT or MRI of the abdomen and pelvis may reveal tumors undetectable by PET.

FIGURE 90-2

Fluid-attenuated inversion recovery sequence magnetic resonance imaging of a patient with limbic encephalitis and LGI1 antibodies. Note the abnormal hyperintensity involving the medial aspect of the temporal lobes.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SPECIFIC PARANEOPLASTIC NEUROLOGIC SYNDROMES

PARANEOPLASTIC ENCEPHALOMYELITIS AND FOCAL ENCEPHALITIS WITH ANTIBODIES AGAINST INTRACELLULAR NEURONAL PROTEINS

The term *encephalomyelitis* describes an inflammatory process with multifocal involvement of the nervous system, including brain, brainstem, cerebellum, and spinal cord. It is often associated with dorsal root ganglia and autonomic dysfunction. For any given patient, the clinical manifestations are determined by the areas predominantly involved, but pathologic studies almost always reveal abnormalities beyond the symptomatic regions. Several clinicopathologic syndromes may occur alone or in combination: (1) *cortical encephalitis*, which may present as “epilepsia partialis continua”; (2) *limbic encephalitis*, characterized by confusion, depression, agitation, anxiety, severe deficits in forming new memories, partial complex seizures, and sometimes dementia (the MRI usually shows unilateral or bilateral medial temporal lobe abnormalities, best seen with T2 and fluid-attenuated inversion recovery sequences); (3) *brainstem encephalitis*, resulting in eye movement disorders (nystagmus, opsoclonus, supranuclear or nuclear paresis), cranial nerve paresis, dysarthria, dysphagia, and central autonomic dysfunction; (4) *cerebellar gait and limb ataxia*; (5) *myelitis*, which may cause lower or upper motor neuron symptoms, myoclonus, muscle rigidity, and spasms; and (6) *autonomic dysfunction* as a result of involvement of the neuraxis at multiple levels, including hypothalamus, brainstem, and autonomic nerves (see “Paraneoplastic Peripheral Neuropathies,” below). Cardiac arrhythmias, postural hypotension, and central hypoventilation are frequent causes of death in patients with encephalomyelitis.

Paraneoplastic encephalomyelitis and focal encephalitis are usually associated with SCLC, but many other cancers have been implicated. Patients with SCLC and these syndromes usually have anti-Hu antibodies in serum and CSF. Anti-CRMP5 antibodies occur less frequently; some of these patients may develop chorea, uveitis, or optic neuritis. Antibodies to Ma proteins are associated with limbic, hypothalamic, and brainstem encephalitis and occasionally with cerebellar symptoms;

some patients develop hypersomnia, cataplexy, and severe hypokinesia. MRI abnormalities are frequent, including those described with limbic encephalitis and variable involvement of the hypothalamus, basal ganglia, or upper brainstem. The oncologic associations of these antibodies are shown in Table 90-2.

TREATMENT

TREATMENT

Encephalomyelitis and Focal Encephalitis

Most types of paraneoplastic encephalitis and encephalomyelitis in which the antigens are intracellular respond poorly to treatment. Stabilization of symptoms or partial neurologic improvement may occur, particularly if there is a satisfactory response of the tumor to treatment. Controlled trials of therapy are lacking, but many reports and the opinion of experts suggest that therapies aimed to remove antibodies against intracellular antigens, such as intravenous immunoglobulin (IVIg) or plasma exchange, usually fail. The main concern should be to treat the tumor and consider immunotherapies, such as cyclophosphamide or tacrolimus, aimed at controlling pathogenic cytotoxic T-cell responses. Approximately 30% of patients with anti-Ma2-associated encephalitis respond to treatment of the tumor (usually a germ cell neoplasm of the testis) and immunotherapy.

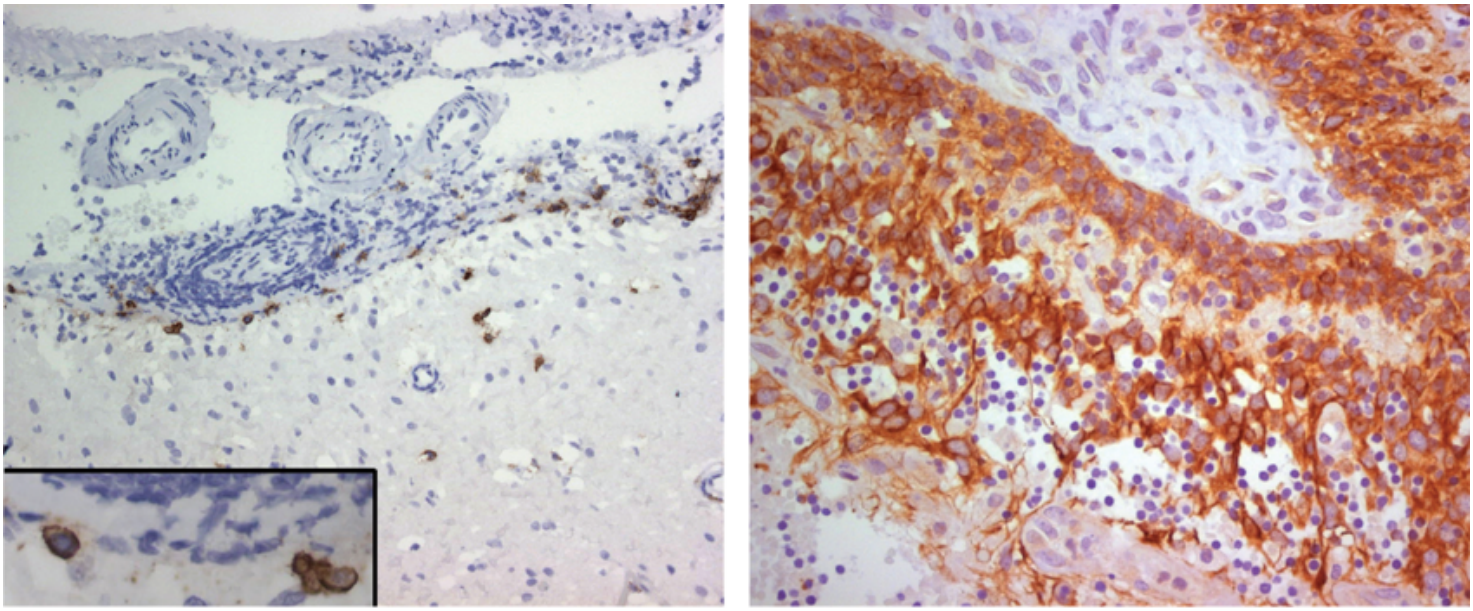
ENCEPHALITIDES WITH ANTIBODIES TO CELL-SURFACE OR SYNAPTIC PROTEINS

These disorders are important for three reasons: (1) they can occur with and without tumor association, (2) some syndromes predominate in young individuals and children, and (3) despite the severity of the symptoms patients usually respond to treatment of the tumor, if found, and immunotherapy (e.g., glucocorticoids, IVIg, plasma exchange, rituximab, or cyclophosphamide) **(Table 90-3)**.

Encephalitis with N*-methyl-*D*-aspartate (NMDA) receptor antibodies* (Fig. 90-1) usually occurs in young women and children, but men and older patients of both sexes can be affected. The disorder has a characteristic pattern of symptom progression that includes a prodrome resembling a viral process, followed in a few days by the onset of severe psychiatric symptoms, sleep dysfunction (usually insomnia), reduced verbal output, memory loss, seizures, decreased level of consciousness, abnormal movements (orofacial, limb, and trunk dyskinesias, dystonic postures), autonomic instability, and frequent hypoventilation. Monosymptomatic episodes, such as pure psychosis, occur in 4% of the patients. Clinical relapses occur in 12–24% of patients (12% during the first 2 years after initial presentation). Most patients have intrathecal synthesis of antibodies, likely by infiltrating plasma cells in brain and meninges **(Fig. 90-3A)**. The syndrome may be misdiagnosed as a viral or idiopathic encephalitis, neuroleptic malignant syndrome, or encephalitis lethargica, and some patients are initially evaluated by psychiatrists with the suspicion of acute psychosis. The detection of an associated teratoma is dependent on age and gender: 46% of female patients aged ≥12 years have uni- or bilateral ovarian teratomas whereas <7% of girls aged <12 have a teratoma **(Fig. 90-3B)**. In male patients, the detection of a tumor is rare. Patients aged >45 years are more frequently male; about 20% of these patients have tumors (e.g., cancer of the breast, ovary, or lung).

FIGURE 90-3

Pathologic findings in anti–*N*-methyl-D-aspartate (NMDA) receptor encephalitis. Infiltrates of plasma cells (brown cells; stained for CD138) in the meninges and brain of a patient (***A***); the inset is a magnification of some plasma cells. ***B.*** Neurons and neuronal processes in the teratoma of a patient (brown cells; stained with MAP2); these neurons express NMDA receptors (not shown). *(Reproduced in part from E Martinez-Hernandez et al: Neurology 77:589, 2011, with permission.)*



A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Approximately 20% of patients with herpes simplex encephalitis develop a form of autoimmune encephalitis that in children usually associates with abnormal movements (choreoathetosis post-herpes simplex encephalitis) and in adults with cognitive and psychiatric symptoms. This disorder develops a few weeks after the viral infection has resolved, associates with new synthesis of antibodies against the NMDA receptor and other neuronal cell surface proteins, and is usually less responsive to immunotherapy than anti-NMDA receptor encephalitis.

Encephalitis with leucine-rich glioma-inactivated 1 (LGI1) antibodies predominates in patients aged >50 years (65% male) and frequently presents with memory loss and seizures (limbic encephalopathy), along with hyponatremia and sleep dysfunction. In some patients, the encephalitis is preceded by or occurs with myoclonic-like movements called faciobrachial dystonic seizures. Less than 10% of patients have thymoma.

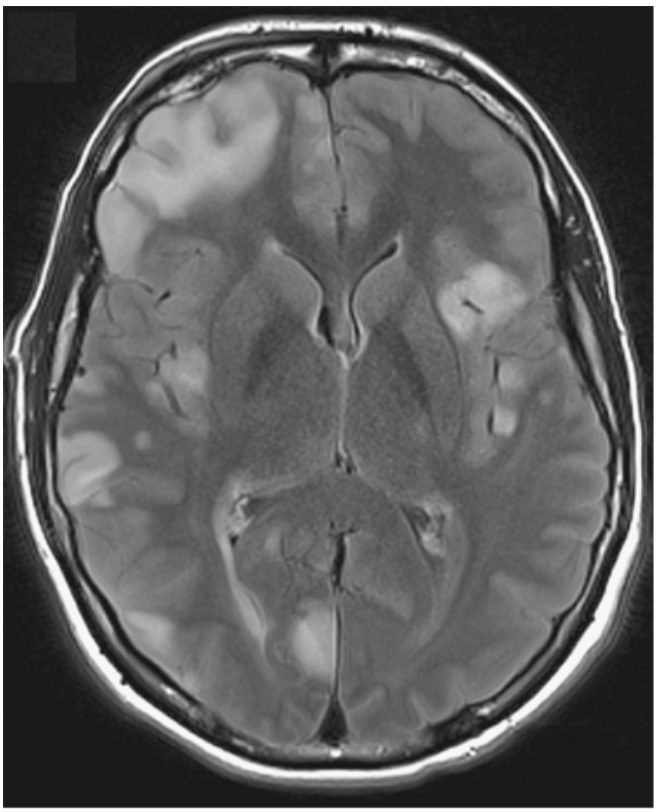
Encephalitis with contactin-associated protein-like 2 (Caspr2) antibodies predominates in patients aged >50 years and is associated with Morvan's syndrome (encephalitis, insomnia, confusion, hallucinations, autonomic dysfunction, and neuromyotonia), or a form of encephalitis with 3 or more of the following core symptoms: encephalopathy, cerebellar symptoms, peripheral nervous system hyperexcitability, dysautonomia, insomnia, neuropathic pain or weight loss. About 20% of patients have thymoma; in patients with Morvan's syndrome the frequency of thymoma appears to be higher.

Encephalitis with γ-aminobutyric acid type B ($GABA_B$) receptor antibodies is usually associated with limbic encephalitis and seizures. In rare instances, patients develop cerebellar symptoms and opsoclonus. Fifty percent of patients have SCLC or a neuroendocrine tumor of the lung. Patients may have additional antibodies to glutamic acid decarboxylase (GAD), which are of unclear significance. Other antibodies to nonneuronal proteins are often found in these patients as well as in patients with α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor antibodies, indicating a general tendency to autoimmunity.

Encephalitis with $GABA_A$ receptor antibodies may affect children and adults, and associates with prominent seizures and status epilepticus, often requiring pharmacologically induced coma. In ~80% of patients, the brain MRI shows multifocal, asynchronous, cortical-subcortical T2/FLAIR abnormalities predominantly involving temporal and frontal lobes, but also basal ganglia and other regions (Fig. 90-4). Most patients do not have an underlying tumor, but some may have thymoma.

FIGURE 90-4

Fluid-attenuated inversion recovery sequence magnetic resonance imaging of a patient with autoimmune encephalitis and GABAA receptor antibodies. Note the abnormal hyperintensity involving cortical and subcortical brain regions, predominantly in frontal and temporal lobes. (*Reprinted from M Spatola et al: Neurology 88:1012, 2017, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Encephalitis with AMPA receptor antibodies affects middle-aged women, who develop acute limbic dysfunction or, less frequently, prominent psychiatric symptoms; 70% of the patients have an underlying tumor in the lung, breast, or thymus. Neurologic relapses may occur; these also respond to immunotherapy and are not necessarily associated with tumor recurrence.

Encephalitis with glycine receptor (GlyR) antibodies has been described in adults with progressive encephalomyelitis with rigidity and myoclonus (PERM) and stiff-person spectrum of symptoms (with or without GAD antibodies). The disorder usually occurs without tumor association, although some patients have lung cancer, thymoma, or Hodgkin's lymphoma.

Encephalitis with dipeptidyl-peptidase-like protein-6 (or DPPX) antibodies results in symptoms of CNS hyperexcitability including agitation, hallucinations, paranoid delusions, tremor, myoclonus, nystagmus, seizures, and hyperekplexia. Some patients develop progressive encephalomyelitis with rigidity and myoclonus. Diarrhea, other gastrointestinal symptoms, and substantial loss of weight often suggest the presence of an underlying tumor, but in most patients no tumor is identified.

Encephalitis with metabotropic glutamate receptor 5 (mGluR5) antibodies is characterized by the development of encephalopathy associated with Hodgkin lymphoma (Ophelia syndrome). Patients show confusion, agitation, memory loss, delusions, paranoid ideation, hallucinations, psychosis, or seizures that are highly responsive to immunotherapy and tumor treatment.

Encephalitis with antibodies against neurexin 3 alpha does not have distinctive clinical features; the experience is limited and the disorder does not appear to associate with cancer.

Encephalitis with antibodies against the dopamine-2 receptor has been reported in some patients with abnormal movements, gait disturbance, psychiatric symptoms, and evidence of basal ganglia encephalitis. The disorder is rare and does not appear to associate with cancer.

Anti-IgLON5 disease is a chronic or subacute encephalopathy that characteristically associates with non-REM and REM parasomnia, brainstem dysfunction, breathing difficulty, and obstructive sleep apnea. It does not associate with cancer, but shows a strong association with HLA-DRB1*10:01 and HLA-DQB1*05:01. The response to immunotherapy is poor. Neuropathological studies show a neuronal tauopathy predominantly involving hypothalamus and tegmentum of the brainstem.

With the exception of patients with anti-IgLON 5 disease, who rarely respond to treatment, most patients with autoimmune or paraneoplastic encephalopathies associated with antibodies against cell surface or synaptic proteins respond to immunotherapy and treatment of the tumor (if appropriate). Although there are no standardized treatment protocols, the most frequent approach employs progressive escalation of immunotherapy using first a combination of glucocorticoids, IVIg or plasma exchange, and if there is no response, either rituximab or cyclophosphamide.

PARANEOPLASTIC CEREBELLAR DEGENERATION

This disorder is often preceded by a prodrome that may include dizziness, oscillopsia, blurry or double vision, nausea, and vomiting. A few days or weeks later, patients develop dysarthria, gait and limb ataxia, and variable dysphagia. The examination usually shows downbeating nystagmus and, rarely, opsoclonus. Brainstem dysfunction, upgoing toes, or a mild neuropathy may occur. Early in the course, MRI studies are usually normal; later, the MRI reveals cerebellar atrophy. The disorder results from extensive degeneration of Purkinje cells, with variable involvement of other cerebellar cortical neurons, deep cerebellar nuclei, and spinocerebellar tracts. The tumors more frequently involved are SCLC, cancer of the breast and ovary, and Hodgkin's lymphoma.

Anti-Yo (PCA1) antibodies in patients with breast or gynecologic cancers typically associate with prominent or pure cerebellar degeneration. A variable degree of cerebellar dysfunction can be associated with virtually any of the antibodies and PND of the CNS shown in Table 90-2. A number of single case reports have described neurologic improvement after tumor removal, plasma exchange, IVIg, cyclophosphamide, rituximab, or glucocorticoids. However, most patients with paraneoplastic cerebellar degeneration and any of the antibodies shown in Table 90-2 do not improve with treatment.

A cerebellar syndrome can also occur with antibodies against cell-surface or synaptic proteins, including P/Q-type voltage-gated calcium channels (VGCC), Tr (DNER), or mGluR1 (Table 90-3). The frequency and type of tumor association varies with the type of antibody. The cerebellar syndrome of patients with mGluR1 antibodies is highly responsive to treatment of the tumor and immunotherapy, whereas the syndrome of patients with Tr or VGCC antibodies is less treatment responsive.

PARANEOPLASTIC OPSOCLONUS-MYOCLONUS SYNDROME

Opsoclonus is a disorder of eye movement characterized by involuntary, chaotic saccades that occur in all directions of gaze; it is frequently associated with myoclonus and ataxia. Opsoclonus-myoclonus may be cancer-related or idiopathic. When the cause is paraneoplastic, the tumors involved are usually cancer of the lung and breast in adults, neuroblastoma in children, and ovarian teratoma in adolescents and young women. The pathologic substrate of opsoclonus-myoclonus is unclear, but studies suggest that disinhibition of the fastigial nucleus of the cerebellum is involved. Most patients do not have antineuronal antibodies. A small subset of patients with ataxia, opsoclonus, and other eye-movement disorders develop anti-Ri antibodies; these patients may also develop muscle rigidity, laryngeal spasms, and autonomic dysfunction. The tumors most frequently involved in anti-Ri-associated syndromes are breast and ovarian cancer. If the tumor is not successfully treated, the syndrome in adults often progresses to encephalopathy, coma, and death. In addition to treating the tumor, symptoms may respond to immunotherapy (glucocorticoids, plasma exchange, and/or IVIg).

At least 50% of children with opsoclonus-myoclonus have an underlying neuroblastoma. Hypotonia, ataxia, behavioral changes, and irritability are frequent accompanying symptoms. Neurologic symptoms often improve with treatment of the tumor and glucocorticoids, adrenocorticotropic hormone (ACTH), plasma exchange, IVIg, rituximab, or cyclophosphamide. Many patients are left with psychomotor retardation and behavioral and sleep problems.

PARANEOPLASTIC SYNDROMES OF THE SPINAL CORD

The number of reports of paraneoplastic spinal cord syndromes, such as *subacute motor neuronopathy* and *acute necrotizing myelopathy*, has decreased in recent years. This may represent a true decrease in incidence, due to improved and prompt oncologic interventions, or the identification of nonparaneoplastic etiologies. Some patients with cancer develop *upper* or *lower motor neuron dysfunction* or both, resembling amyotrophic lateral sclerosis. It is unclear whether these disorders have a paraneoplastic etiology or simply coincide with the presence of cancer. There are isolated case reports of cancer patients with motor neuron dysfunction who had neurologic improvement after tumor treatment. A search for lymphoma should be undertaken in patients with a rapidly progressive motor neuron syndrome and a monoclonal protein in serum or CSF.

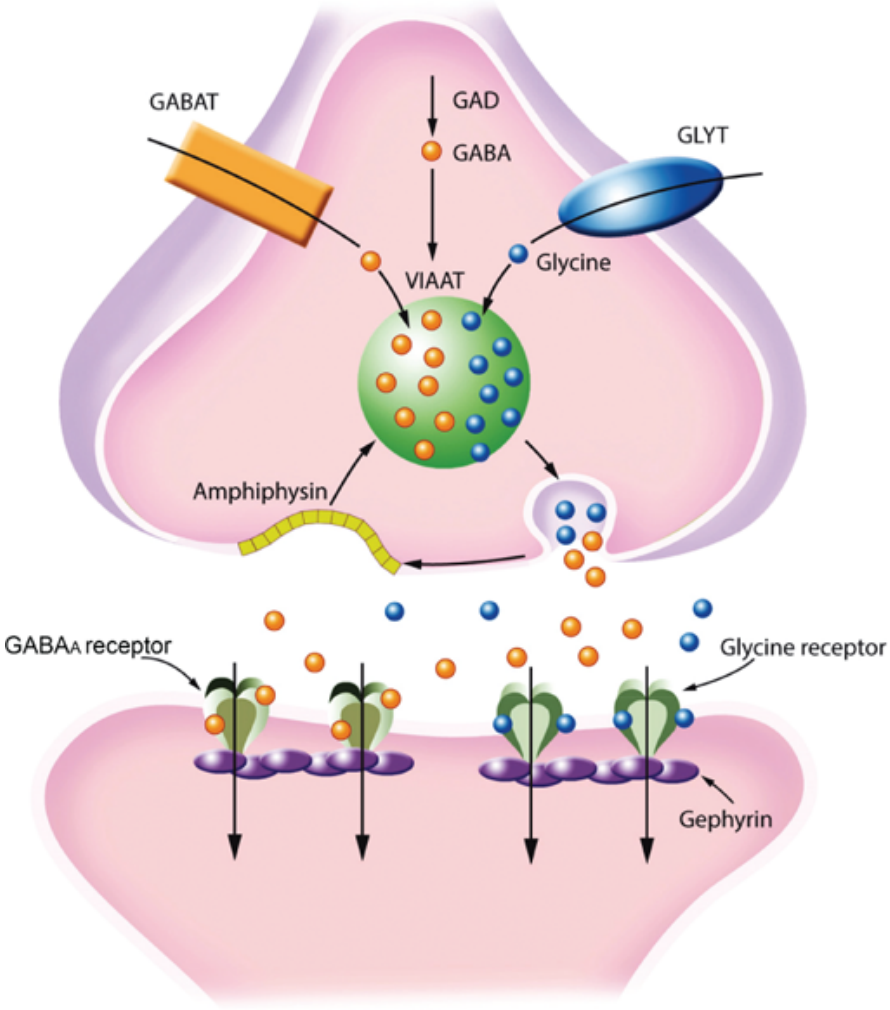
Paraneoplastic myelitis may present with upper or lower motor neuron symptoms, segmental myoclonus, and rigidity, and can be the first manifestation of encephalomyelitis. *Neuromyelitis optica (NMO) with aquaporin 4 antibodies* may occur in rare instances as a paraneoplastic manifestation of a cancer.

PARANEOPLASTIC STIFF-PERSON SYNDROME

This disorder is characterized by progressive muscle rigidity, stiffness, and painful spasms triggered by auditory, sensory, or emotional stimuli. Rigidity mainly involves the lower trunk and legs, but it can affect the upper extremities and neck. Sometimes, only one extremity is affected (*stiff-limb syndrome*). Symptoms improve with sleep and general anesthetics. Electrophysiologic studies demonstrate continuous motor unit activity. The associated antibodies target proteins (GAD, amphiphysin) involved in the function of inhibitory synapses using γ-aminobutyric acid (GABA) or glycine as neurotransmitters. The presence of amphiphysin antibodies usually indicates a paraneoplastic etiology related to SCLC and breast cancer. By contrast, GAD antibodies may occur in some cancer patients but are much more frequently present in the nonparaneoplastic disorder. GlyR antibodies may occur in some patients with stiff-person syndrome; these antibodies are also detectable in patients with PERM **(Fig. 90-5)**.

FIGURE 90-5

Autoantibodies in stiff-person spectrum disorders. Schematic representation of an inhibitory synapse including the three main targets of autoantibodies in patients with stiff-person spectrum disorders: the presynaptic proteins GAD and amphiphysin, and the postsynaptic glycine receptor (GlyR). *(From J Dalmau et al: Physiol Rev 97:839, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Optimal treatment of stiff-person syndrome requires therapy of the underlying tumor, glucocorticoids, and symptomatic use of drugs that enhance GABA-ergic transmission (diazepam, baclofen, sodium valproate, tiagabine, vigabatrin). IVIg and plasma exchange are transiently effective in some patients, and there are reports of responses to rituximab in patients who did not respond to other treatments.

PARANEOPLASTIC SENSORY NEURONOPATHY OR DORSAL ROOT GANGLIONOPATHY

This syndrome is characterized by sensory deficits that may be symmetric or asymmetric, painful dysesthesias, radicular pain, and decreased or absent reflexes. All modalities of sensation and any part of the body including face and trunk can be involved. Specialized sensations such as taste and hearing can also be affected. Electrophysiologic studies show decreased or absent sensory nerve potentials with normal or near-normal motor conduction velocities. Symptoms result from an inflammatory, likely immune-mediated, process that targets the dorsal root ganglia, causing neuronal loss and secondary degeneration of the posterior columns of the spinal cord. The dorsal and, less frequently, the anterior nerve roots and peripheral nerves may also be involved. This disorder often precedes or is associated with encephalomyelitis and autonomic dysfunction and has the same immunologic and oncologic associations (Hu antibodies, SCLC).

As with anti-Hu-associated encephalomyelitis, the therapeutic approach focuses on prompt treatment of the tumor and use of immune suppressants such as cyclophosphamide or tacrolimus. Glucocorticoids occasionally produce clinical stabilization or improvement. The benefit of IVIg and plasma exchange is not proven.

PARANEOPLASTIC PERIPHERAL NEUROPATHIES

These disorders may develop any time during the course of the neoplastic disease. Neuropathies occurring at late stages of cancer or lymphoma usually cause mild to moderate sensorimotor deficits due to axonal degeneration of unclear etiology. These neuropathies are often masked by concurrent neurotoxicity from chemotherapy and other cancer therapies. In contrast, the neuropathies that develop in the early stages of cancer frequently show a rapid progression, sometimes with a relapsing and remitting course, and evidence of inflammatory infiltrates and axonal loss or demyelination. If demyelinating features predominate **(Chaps. 438 and 439)**, IVIg, plasma exchange, or glucocorticoids may improve symptoms. Occasionally anti-CRMP5 antibodies are present; detection of anti-Hu suggests concurrent dorsal root ganglionitis.

Guillain-Barré syndrome and *brachial plexitis* have occasionally been reported in patients with lymphoma, but there is no clear evidence of a paraneoplastic association **(Chap. 439)**.

Monoclonal gammopathies associated with multiple myeloma, cryoglobulinemia, amyloidosis, Waldenstrom macroglobulinemia, or POEMS (polyneuropathy, organomegaly, endocrinopathy, M-protein spike, and skin manifestations) syndrome among others, may cause neuropathy by a variety of mechanisms, including compression of roots and plexuses by metastasis to vertebral bodies and pelvis, by deposits of amyloid in peripheral nerves, or through a direct interaction of the abnormal immunoglobulin with peripheral nerve antigens. In other patients, the mechanisms underlying the neuropathy remain unknown. Neuropathies more often occur with IgM gammopathies than with IgG and IgA isotypes. The phenotype of the neuropathy and likelihood of improvement with successful treatment of the gammopathy is dependent on the underlying hematologic disorder.

Vasculitis of the nerve and muscle causes a painful symmetric or asymmetric distal axonal sensorimotor neuropathy with variable proximal weakness. It predominantly affects elderly men and is associated with an elevated erythrocyte sedimentation rate and increased CSF protein concentration. SCLC and lymphoma are the primary tumors involved. Glucocorticoids and cyclophosphamide often result in neurologic improvement.

Peripheral nerve hyperexcitability (*neuromyotonia,* or *Isaacs' syndrome*) is characterized by spontaneous and continuous muscle fiber activity of peripheral nerve origin. Clinical features include cramps, muscle twitching (fasciculations or myokymia), stiffness, delayed muscle relaxation (pseudomyotonia), and spontaneous or evoked carpal or pedal spasms. The involved muscles may be hypertrophic, and some patients develop paresthesias and hyperhidrosis. CNS dysfunction, including mood and behavioral changes, sleep disorder, hallucinations, and autonomic symptoms may occur. The electromyogram (EMG) shows fibrillations; fasciculations; and doublet, triplet, or multiplet single-unit (myokymic) discharges that have a high intraburst frequency. Some patients have Caspr2 antibodies usually in the context of Morvan's syndrome, but most patients with isolated neuromyotonia are antibody negative. The disorder often occurs without cancer; if paraneoplastic, benign and malignant thymomas and SCLC are the usual tumors. Phenytoin, carbamazepine, and plasma exchange improve symptoms.

Paraneoplastic autonomic neuropathy usually develops as a component of other disorders, such as LEMS and encephalomyelitis. It may rarely occur as a pure or predominantly autonomic neuropathy with cholinergic or adrenergic dysfunction at the pre- or postganglionic levels. Patients can develop several life-threatening complications, such as gastrointestinal paresis with pseudo-obstruction, cardiac dysrhythmias, and postural hypotension. Other clinical features include abnormal pupillary responses, dry mouth, anhidrosis, erectile dysfunction, and problems in sphincter control. The disorder occurs in association with several tumors, including SCLC, cancer of the pancreas or testis, carcinoid tumors, and lymphoma. Because autonomic symptoms can be the presenting feature of encephalomyelitis, serum anti-Hu and anti-CRMP5 antibodies should be sought. Antibodies to ganglionic (alpha3-type) neuronal acetylcholine receptors are the cause of autoimmune autonomic ganglionopathy, a disorder that frequently occurs without cancer association **(Chap. 432)**.

LAMBERT-EATON MYASTHENIC SYNDROME

LEMS is discussed in Chap. 440.

MYASTHENIA GRAVIS

Myasthenia gravis is discussed in Chap. 440.

POLYMYOSITIS-DERMATOMYOSITIS

Polymyositis and dermatomyositis are discussed in detail in Chap. 358.

ACUTE NECROTIZING MYOPATHY

Patients with this syndrome develop myalgias and rapid progression of weakness involving the extremities, neck, pharyngeal, respiratory, and sometimes cardiac muscles. Serum muscle enzymes are elevated, and muscle biopsy shows extensive necrosis with minimal or absent inflammation and sometimes deposits of complement. The disorder may occur without a cancer association (sometimes as a result of statin exposure, connective tissue disease, or HIV) or with cancer. Patients with antibodies against 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMGCR) and seronegative patients are more likely to have an underlying cancer than those with antibodies against signal recognition particle. No specific type of cancer has been found to predominate. Successful treatment of the tumor and aggressive immunotherapy (glucocorticoids, IVIg, and steroid-sparing immunosuppressants) may lead to complete or substantial recovery.

PARANEOPLASTIC VISUAL SYNDROMES

This group of disorders involves the retina and, less frequently, the uvea and optic nerves. The term *cancer-associated retinopathy* is used to describe paraneoplastic cone and rod dysfunction characterized by photosensitivity, progressive loss of vision and color perception, central or ring scotomas, night blindness, and attenuation of photopic and scotopic responses in the electroretinogram (ERG). The most commonly associated tumor is SCLC. *Melanoma-associated retinopathy* affects patients with metastatic cutaneous melanoma. Patients develop acute onset of night blindness and shimmering, flickering, or pulsating photopsias that often progress to visual loss. The ERG shows reduced b-waves with normal dark adapted a-waves. Paraneoplastic optic neuritis and uveitis can develop in association with encephalomyelitis. Patients with paraneoplastic uveitis and optic neuritis may harbor CRMP5 antibodies.

Some paraneoplastic retinopathies are associated with serum antibodies that specifically react with the subset of retinal cells undergoing degeneration, supporting an immune-mediated pathogenesis (Table 90-2). Paraneoplastic retinopathies rarely show substantial improvement after treatment of the tumor and immunotherapy, however, stabilization of symptoms and partial responses to a variety of immunotherapies (glucocorticoids, plasma exchange, IVIg, rituximab or alemtuzumab) have been reported.

FURTHER READING

Dalmau J, Rosenfeld MR: Paraneoplastic syndromes of the CNS. Lancet Neurol 7:327, 2008. [PubMed: 18339348]

Dalmau J et al: Autoantibodies to synaptic receptors and neuronal cell-surface proteins in autoimmune diseases of the central nervous system. Physiol Rev 97:839, 2017. [PubMed: 28298428]

Darnell RB, Posner J: *Paraneoplastic Syndromes*. New York, Oxford University Press, 2011.

Didelot A, Honnorat J: Paraneoplastic disorders of the central and peripheral nervous systems. Handb Clin Neurol 121:1159, 2014. [PubMed: 24365410]

Graus F et al: Clinical approach to diagnosis of autoimmune encephalitis. Lancet Neurol 15:391, 2016. [PubMed: 26906964]

Chapter 91: Late Consequences of Cancer and Its Treatment

Carl E. Freter; Dan L. Longo

INTRODUCTION

There are over 10 million American cancer survivors. The vast majority of these will bear some mark of their cancer and its treatment, and a large proportion will experience long-term consequences including medical problems, psychosocial dysfunction, economic hardship, sexual dysfunction, and discrimination regarding employment and insurance. Many of these problems are directly related to cancer treatment. As patients survive longer from more types of malignancies, we are increasingly recognizing the biologic toll our very imperfect therapies take in terms of morbidity and mortality. The human face of these consequences of therapy confronts the cancer specialist who treats them every day. Although long-term survivors of childhood leukemias, Hodgkin's lymphoma, and testicular cancer, as examples, have taught us much about the consequences of cancer treatment, we keep learning more as patients survive longer with newer therapies. Newer "targeted" chemotherapy drugs have their own, often unique, long-term toxicities about which we remain in a learning process. Cancer "survivorship" clinics are increasing to expressly follow patients for long-term toxicities of cancer treatment.

The pace of developing therapies that mitigate treatment-related consequences has been slow, partly due to an understandable aversion to alter regimens that work and partly due to a lack of new, effective, less toxic therapeutic agents with less "collateral damage" to replace known agents with known toxicities. The types of damage from cancer treatment vary. Often, a final common pathway is irreparable damage to DNA. Surgery can create dysfunction, including blind gut loops with absorption problems and loss of function of removed body parts. Radiation may damage end-organ function, for example, loss of potency in prostate cancer patients, pulmonary fibrosis, and neurocognitive impairment, and may act as a direct carcinogen. Cancer chemotherapy can be a direct carcinogen and has a kaleidoscope of other toxicities discussed in this chapter. Table 91-1 lists the late effects of cancer treatment.

TABLE 91-1

Late Effects of Cancer Therapy

Surgical Procedure	**Effect**
Amputation	Functional loss
Lymph node dissection	Risk of lymphedema
Ostomy	Psychosocial impact
Splenectomy	Risk of sepsis
Adhesions	Risk of obstruction
Bowel anastomoses	Malabsorption syndromes
Radiation Therapy	**Effect**
Organ	
Bone	Premature termination of growth, osteonecrosis
Soft tissues	Atrophy, fibrosis
Brain	Neuropsychiatric deficits, cognitive dysfunction
Thyroid	Hypothyroidism, Graves' disease, cancer
Salivary glands	Dry mouth, caries, dysgeusia
Eyes	Cataracts
Heart	Pericarditis, myocarditis, coronary artery disease
Lung	Pulmonary fibrosis
Kidney	Decreased function, hypertension
Liver	Decreased function
Intestine	Malabsorption, stricture

Gonads	Infertility, premature menopause	
Any	Secondary neoplasia	
Chemotherapy		**Effect**
Organ	**Drug**	
Bone	Glucocorticoids	Osteoporosis, avascular necrosis
Brain	Methotrexate, cytarabine, others	Neuropsychiatric deficits, cognitive decline?
Peripheral nerves	Vincristine, platinum, taxanes	Neuropathy, hearing loss
Eyes	Glucocorticoids	Cataracts
Heart	Anthracyclines, trastuzumab	Cardiomyopathy
Lung	Bleomycin	Pulmonary fibrosis
	Methotrexate	Pulmonary hypersensitivity
Kidney	Platinum, others	Decreased function, hypomagnesemia
Liver	Various	Altered function
Gonads	Alkylating agents, others	Infertility, premature menopause
Bone marrow	Various	Aplasia, myelodysplasia, secondary leukemia

The first goal of therapy is to eradicate or control the malignancy. Late treatment consequences are, indeed, testimony to the increasing success of such treatment. Their occurrence sharply underlines the necessity to develop more effective therapies with less long-term morbidity and mortality. At the same time, a sense of perspective and relative risk is necessary; fear of long-term complications should not prevent the application of effective, particularly curative, cancer treatment.

CARDIOVASCULAR DYSFUNCTION

CHEMOTHERAPEUTIC AGENTS

Cardiovascular toxicity of cancer chemotherapeutic agents includes dysrhythmias, cardiac ischemia, cardiomyopathic congestive heart failure (CHF), pericardial disease, and peripheral vascular disease. Because these cardiac toxicities are difficult to distinguish from disease that is not associated with cancer treatment, clear etiologic implication of cancer chemotherapeutic agents may be difficult. Cardiovascular complications occurring in an unexpected clinical setting in patients who have undergone cancer therapy are often important in raising suspicion. Dose-dependent myocardial toxicity of anthracyclines with characteristic myofibrillar dropout is pathologically pathognomonic on endomyocardial biopsy. Anthracycline cardiotoxicity occurs through a root mechanism of chemical free radical damage. Fe^{3+}-doxorubicin complexes damage DNA, nuclear and cytoplasmic membranes, and mitochondria. About 5% of patients receiving >450–550 mg/m^2 of doxorubicin will develop CHF. Cardiotoxicity in relation to the dose of anthracycline is clearly not a step function, but rather a continuous function, and occasional patients are seen with CHF at substantially lower doses. Advanced age, other concomitant cardiac disease, hypertension, diabetes, and thoracic radiation therapy are all important cofactors in promoting anthracycline-associated CHF. The risk of cardiac failure appears to be substantially lower when doxorubicin is administered by continuous infusion. Anthracycline-related CHF is difficult to reverse and has a mortality rate as high as 50%, making prevention crucial. Some anthracyclines such as mitoxantrone are associated with less cardiotoxicity, and continuous-infusion regimens and liposomally encapsulated doxorubicin are associated with less cardiotoxicity. Dexrazoxane, an intracellular iron chelator, may limit anthracycline toxicity, but the concern of limiting chemotherapeutic efficacy has somewhat limited its use. Monitoring patients for cardiac toxicity typically involves periodic gated nuclear cardiac blood pool ejection fraction testing (multigated acquisition scan [MUGA]) or cardiac ultrasonography. More recently, cardiac magnetic resonance imaging (MRI) has been used, but MRI is not standard or widespread. Testing is performed more frequently at higher cumulative doses, with additional risk factors, and certainly for any newly developing CHF or other symptoms of cardiac dysfunction.

After anthracyclines, trastuzumab is the next most frequent cardiotoxic drug currently in use. Trastuzumab is frequently used as adjuvant breast cancer therapy, sometimes in conjunction with anthracyclines, which is believed to result in additive or possibly synergistic toxicity. In contrast to anthracyclines, cardiotoxicity is not dose-related, is usually reversible, is not associated with pathologic changes of anthracyclines on cardiac myofibrils, and has a different biochemical mechanism inhibiting intrinsic cardiac repair mechanisms. Toxicity is typically routinely monitored every 3–4 doses using functional cardiac testing as mentioned earlier for anthracyclines.

Other cardiotoxic drugs include lapatinib, phosphoramide mustards (cyclophosphamide), ifosfamide, interleukin 2, ponatinib, imatinib, and sunitinib.

RADIATION THERAPY

Radiation therapy that includes the heart can cause interstitial myocardial fibrosis, acute and chronic pericarditis, valvular disease, and accelerated premature atherosclerotic coronary artery disease. Repeated or high (>6000 cGy) radiation doses are associated with greater risk, as is concomitant or distant cardiotoxic

cancer chemotherapy exposure. Symptoms of acute pericarditis, which peaks about 9 months after treatment, include dyspnea, chest pain, and fever. Chronic constrictive pericarditis may develop 5–10 years following radiation therapy. Cardiac valvular disease includes aortic insufficiency from fibrosis or papillary muscle dysfunction resulting in mitral regurgitation. A threefold increased risk of fatal myocardial infarction is associated with mantle field radiation with accelerated coronary artery disease. Carotid radiation similarly increases the risk of embolic stroke.

TREATMENT

TREATMENT

Chemotherapeutic/Radiation-Induced Cardiovascular Disease

Therapy for chemotherapeutic/radiation-induced cardiovascular disease is essentially the same as therapy for disease not associated with cancer treatment. Discontinuation of the offending agent is the first step. Diuretics, fluid and sodium restriction, and antiarrhythmic agents are often useful for acute symptoms. Afterload reduction with angiotensin-converting enzyme (ACE) inhibitors or, in some cases, β-adrenergic blockers (carvedilol) often is of significant benefit, and digitalis may be helpful as well.

A hybrid discipline of “cardio-oncology” has been developing in clinics to expressly follow chemotherapy-treated patients for cardiotoxicity. The goals are early intervention using more sensitive techniques, management of cardiotoxicity before it becomes symptomatic, and using clinical trials to identify cardioprotective strategies.

PULMONARY DYSFUNCTION

CHEMOTHERAPEUTIC AGENTS

Bleomycin generates activated free radical oxygen species and causes pneumonitis associated with a radiographic or interstitial ground-glass appearance diffusely throughout both lungs, often worse in the lower lobes. A nonproductive cough with or without fever may be an early sign. This toxicity is dose-related and dose-limiting. The diffusion capacity of the lungs for carbon dioxide (DL_{CO}) is a sensitive measure of toxicity and recovery, and a baseline value is generally obtained for future comparison prior to bleomycin therapy. Additive or synergistic risk factors include age, prior lung disease, and concomitant use of other chemotherapy, lung irradiation, and high concentrations of inspired oxygen. Other chemotherapeutic agents notable for pulmonary toxicity include mitomycin, nitrosoureas, doxorubicin with radiation, gemcitabine combined with weekly docetaxel, methotrexate, and fludarabine. High-dose alkylating agents, cyclophosphamide, ifosfamide, and melphalan are frequently used in the hematopoietic stem cell transplant setting, often with whole-body radiation. This therapy may result in severe pulmonary fibrosis and/or pulmonary venoocclusive disease.

RADIATION THERAPY

Risk factors for radiation pneumonitis include advanced age, poor performance status, preexisting compromised pulmonary function, and radiation volume and dose. The dose "threshold" is thought to be in the range of 5–20 Gy. Hypoxemia and dyspnea on exertion are characteristic. Fine, high-pitched "Velcro rales" may be an accompanying physical finding, and fever, cough, and pleuritic chest pain are common symptoms. The DL_{CO} is the most sensitive measure of pulmonary functional impairment, and ground-glass infiltrates often correspond with relatively sharp edges to the irradiated volume, although the pneumonitis may progress beyond the field and even occasionally involve the contralateral unirradiated lung.

TREATMENT

TREATMENT

Pulmonary Dysfunction

Chemotherapy- and radiation-induced pneumonitis is generally very corticosteroid responsive, except in the case of nitrosoureas. Prednisone 1 mg/kg is often used to control acute symptoms and pulmonary dysfunction with a generally slow taper. Prolonged glucocorticoid therapy requires gastrointestinal protection with proton pump inhibitors, management of hyperglycemia, heightened infection management, and treatment of steroid-induced osteoporosis. Antibiotics, bronchodilators, oxygen in only necessary doses, and diuretics may all play an important role in management of pneumonitis, and consultation with a pulmonologist should be routinely undertaken. Amifostine has been studied as a pulmonary radioprotectant, with inconclusive results, and is associated with skin rash, fatigue, and nausea; hence, it is not considered standard therapy at this time. Transforming growth factor β (TGF-β) is believed to be a major inducer of radiation fibrosis and represents a therapeutic target for development of anti-TGF-β therapies.

NEUROLOGIC DYSFUNCTION

CHEMOTHERAPEUTIC AGENTS

Chemotherapy- and radiation-induced neurologic dysfunction is unfortunately increasing in both incidence and severity as a result of improved supportive care leading to more aggressive regimens and longer cancer survivorship allowing the development of late toxicity. Direct effects on myelin, glial cells, and neurons have all been implicated, with alterations in cellular cytoskeleton, axonal transport, and cellular metabolism as mechanisms.

Vinca alkaloids produce a characteristic "stocking-glove" neuropathy with numbness and tingling advancing to loss of motor function, which is highly dose related. Distal sensorimotor polyneuropathy prominently involves loss of deep tendon reflexes with initially loss of pain and temperature sensation, followed by proprioceptive and vibratory loss. This requires careful patient history and physical examination by experienced oncologists to decide when the drug must be stopped due to toxicity. Milder toxicity often slowly completely resolves. Vinca alkaloids may sometimes be associated with jaw claudication, autonomic neuropathy, ileus, cranial nerve palsies, and, in severe cases, encephalopathy, seizures, and coma.

Cisplatin is associated with sensorimotor neuropathy and hearing loss, especially at doses >400 mg/m^2, requiring audiometry in patients with preexisting hearing compromise. Carboplatin is often substituted in such cases given its lesser effect on hearing.

Many of the agents that target kinase enzymes in tumor cells and 5-fluorouracil congeners produce dysesthesias and painful hands and feet known as hand-foot syndrome or palmar-plantar erythrodysesthesia. Symptoms usually abate when the agent is stopped.

Neurocognitive dysfunction has been well described in childhood survivors of acute lymphoblastic leukemia (ALL) treatment, including intrathecal methotrexate or cytosine arabinoside in conjunction with prophylactic cranial irradiation. Methotrexate alone may cause acute leukoencephalopathy characterized by somnolence and confusion that is often reversible. Acute toxicity is dose related, especially at doses >3 g/m^2, with younger patients being at greater risk. Subacute methotrexate toxicity occurs weeks after therapy and is often ameliorated with glucocorticoid therapy. Chronic methotrexate toxicity (leukoencephalopathy) develops months or years after treatment and is characterized clinically as progressive loss of cognitive function and focal neurologic signs, which are irreversible, promoted by synchronous or metachronous radiation therapy, and more pronounced at a younger age.

Neurocognitive decline following chemotherapy alone occurs notably in breast cancer patients receiving adjuvant chemotherapy; this has been referred to as “chemo brain.” It is clinically associated with impaired memory, learning, attention, and speed of information processing. There is no clear mechanistic explanation for its cause and no clearly effective therapy. This entity is justifiably attracting more attention and clearly needs to be studied to develop effective therapy or prophylaxis.

Many cancer patients experience intrusive or debilitating concerns about cancer recurrence following successful therapy. In addition, these patients may experience job, insurance, stress, relationship, financial, and sexual difficulties. Oncologists need to ask about and address these issues explicitly with patients and provide appropriate counseling or support systems. Suicidal ideation and suicide have an increased incidence in cancer patients and survivors.

RADIATION THERAPY

Acute radiation central nervous system (CNS) toxicity occurs within weeks; is characterized by nausea, drowsiness, hypersomnia, and ataxia; and is most often associated with recovery. Early delayed toxicity occurring weeks to 3 months following therapy is associated with similar symptoms as acute toxicity and is pathologically associated with reversible demyelination. Chronic, late radiation injury occurs 9 months to up to 10 years following therapy. Focal necrosis is a common pathologic finding, and glucocorticoid therapy may be helpful. Diffuse radiation injury is associated with global CNS neurologic dysfunction and diffuse white matter changes on computed tomography (CT) or MRI. Pathologically, small vessel changes are prominent. Glucocorticoids may be symptomatically useful but do not alter the course. Necrotizing encephalopathy is the most severe form of radiation injury and almost always is associated with chemotherapy, notably methotrexate.

Cranial radiation may also be associated with an array of endocrine abnormalities with disruption of normal pituitary/hypothalamic axis function, and a high index of suspicion needs to be maintained to identify and treat this toxicity.

Radiation-associated spinal cord injury (myelopathy) is highly dose-dependent and rarely occurs with modern radiation therapy. An early, self-limited form involving electric sensations down the spine on neck flexion (Lhermitte's sign) is seen 6–12 weeks after treatment and generally resolves over weeks. Peripheral nerve toxicity is quite rare owing to relative radiation resistance.

HEPATIC DYSFUNCTION

CHEMOTHERAPEUTIC AGENTS

Long-term hepatic damage from standard chemotherapy regimens is rare. Long-term methotrexate or high-dose chemotherapy alone or with radiation therapy, for example, in preparative regimens for bone marrow transplantation, may result in venoocclusive disease of the liver. This potentially lethal complication classically presents with anicteric ascites, elevated alkaline phosphatase, and hepatosplenomegaly. Pathologically, there is venous congestion, epithelial cell proliferation, and hepatocyte atrophy progressing to frank fibrosis. Frequent monitoring of liver function tests during any chemotherapy is necessary to avoid both idiosyncratic and expected toxicities.

Certain nucleoside drugs have been associated with hepatic dysfunction; however, this complication is rare in oncology.

RADIATION THERAPY

Hepatic radiation damage depends on dose, volume, fractionation, preexisting liver disease, and synchronous or metachronous chemotherapy. In general, radiation doses to the liver >1500 cGy can produce hepatic dysfunction with a steep dose-injury curve. Radiation-induced liver disease closely mimics hepatic venoocclusive disease.

RENAL/BLADDER DYSFUNCTION

Cisplatin produces reversible decrements in renal function, but may also produce severe irreversible toxicity in the presence of renal disease and may predispose to accentuated damage with subsequent renal insults. Cyclophosphamide and ifosfamide, as prodrugs primarily activated in the liver, have cleavage products (acrolein) that can produce hemorrhagic cystitis. This can be prevented with the free radical scavenger MESNA (mercaptoethane sulfonate), which is required for ifosfamide administration. Hemorrhagic cystitis caused by these agents may predispose to bladder cancer.

REPRODUCTIVE AND ENDOCRINE DYSFUNCTION

CHEMOTHERAPEUTIC AGENTS

Alkylating agents are associated with the highest rates of male and female infertility, which is directly dependent on age, dose, and duration of treatment. The age at treatment is an important determinant of fertility outcome, with prepubertal patients having the highest tolerance. Ovarian failure is age related, and females who resume menses after treatment are still at increased risk for premature menopause. Males generally have reversible azoospermia during lower intensity alkylator chemotherapy, and long-term infertility is associated with doses of cyclophosphamide >9 g/m^2 and with high-intensity therapy, such as that used in hematopoietic stem cell transplantation. Males undergoing potentially sterilizing chemotherapy should be offered sperm banking. Gonadotropin-releasing hormone (GnRH) analogs remain experimental to preserve ovarian function. Assisted reproductive technologies can be helpful to couples with chemotherapy-induced infertility.

RADIATION THERAPY

Testicles and ovaries in prepubertal patients are less sensitive to radiation damage; spermatogenesis is affected by low doses of radiation, and complete azoospermia occurs at 600–700 cGy. Leydig cell dysfunction, in contrast, occurs at <2000 cGy, and hence, endocrine function is lost at much higher radiation doses than spermatogenesis. Erectile dysfunction occurs in up to 80% of men treated with external-beam radiation therapy for prostate cancer. Sildenafil may be useful in reversing erectile dysfunction. Ovarian function damage with radiation is age related and occurs at doses of 150–500 cGy. Premature induction of menopause can have serious medical and psychological sequelae. Hormone replacement therapy is often contraindicated (as in estrogen receptor–positive breast cancer). Attention must be paid to maintenance of bone mass with calcium and vitamin D supplements and oral bisphosphonates, and bone mass should be monitored using bone density determinations. Paroxetine, clonidine, pregabalin, and other drugs may be useful in symptomatically controlling hot flashes.

Long-term survivors of childhood cancer (e.g., ALL) who have received cranial radiation may have altered leptin biology and growth hormone deficiency, leading to obesity and reduced strength, exercise tolerance, and bone density.

Radiation therapy to the neck (e.g., in Hodgkin's lymphoma) may lead to hypothyroidism, Graves' disease, thyroiditis, and thyroid malignancies. Thyroid-stimulating hormone (TSH) is followed routinely in such patients to prevent hypothyroidism, and to suppress persistently elevated levels of TSH, which may cause or drive thyroid cancer.

IMMUNOTHERAPY

The development of treatments that inhibit immune checkpoints like CTLA-4 and PD-1 has led to significant advances in treating a number of malignancies. However, late side effects from these agents may occur. The most serious chronic toxicities include the breaking of self-tolerance and the autoimmune destruction of certain endocrine organs, particularly the thyroid and the adenohypophysis (anterior pituitary). Patients with

autoimmune thyroiditis or hypophysitis require life-long hormone replacement. Early recognition is important.

OCULAR COMPLICATIONS

Cataracts may be caused by glucocorticoids, depending on duration and dose; radiation therapy; and uncommonly, tamoxifen. Orbital radiation therapy may cause blindness.

ORAL COMPLICATIONS

Radiation therapy can produce xerostomia (dry mouth), with an attendant increase in caries and poor dentition. Taste and appetite may be suppressed. Bisphosphonate use may result in osteonecrosis of the jaw.

RAYNAUD'S PHENOMENON

Up to 40% of patients treated with bleomycin may develop Raynaud's phenomenon as a result of an unknown mechanism.

SECOND MALIGNANCIES

Second malignancies in patients cured of cancer are a major cause of death, and treated cancer patients must be monitored for their occurrence. The induction of second malignancies is governed by the complex interplay of a number of factors including age, gender, environmental exposures, genetic susceptibility, and cancer treatment itself. In a number of settings, the events leading to the primary cancer themselves increase the risk of second malignancies. Patients with lung cancer are at increased risk of esophageal and head and neck cancers, and vice versa, due to shared risk factors including alcohol and tobacco abuse. Indeed, the risk of developing a second primary head and neck, esophageal, or lung cancer is also increased in these patients. Patients with breast cancer are at increased risk of breast cancer in the opposite breast. Patients with Hodgkin's lymphoma are at risk for non-Hodgkin's lymphomas. Genetic cancer syndromes (e.g., multiple endocrine neoplasia or Li-Fraumeni, Lynch's, Cowden's, and Gardner's syndromes) are examples of genetically based second malignancies of specific types. Cancer treatment itself does not appear to be responsible for the risk of these secondary malignancies. Deficient DNA repair can greatly increase the risk of cancers from DNA-damaging agents, as in ataxia-telangiectasia. Importantly, the risk of treatment-related second malignancies is at least additive and often synergistic with combined chemotherapy and radiation therapy, and hence for such combined-therapy treatment approaches, it is important to establish the necessity of each in the treatment program. All of these patients require special surveillance or, in some cases, prophylactic surgery as part of appropriate treatment and follow-up.

CHEMOTHERAPEUTIC AGENTS

Chemotherapy is significantly associated with two fatal second malignancies: acute leukemia and myelodysplastic syndromes. Two types of leukemia have been described; in patients treated with alkylating agents, acute myeloid leukemia is associated with deletions in chromosome 5 or 7. The lifetime risk is about 1–5%, is increased by radiation therapy, and increases with age. The incidence of these leukemias peaks at 4–6 years, with risk returning close to baseline at 10 years. The other type of acute myeloid leukemia is related to therapy with topoisomerase inhibitors, is associated with chromosome 10q23 translocations, has an incidence of <1%, and generally occurs 1.5–3 years after treatment. Both of these acute myeloid leukemias are refractory to treatment and have a high mortality. The development of myelodysplastic syndromes is increased following chemotherapy, and these are often associated with leukemic progression and a dismal prognosis. A fraction of the population (those with or without cancer) develops clonal hematopoiesis and the percentage increases with age. In such patients, the hematopoietic stem cells carry mutations that are associated with myeloid malignancy despite normal blood counts. It is thought that the presence of these genetic lesions may predispose patients to develop myeloid malignancies.

RADIATION THERAPY

Patients receiving radiation have an increasing and lifelong risk of second malignancies that is 1–2% in the second decade following treatment but increases to >25% after 25 years. These malignancies include cancers of the thyroid and breast, sarcomas, gastric cancer, and CNS cancers, which often tend to be aggressive, occur in or adjacent to a radiation field, and have a poor prognosis. An example of organ-, age-, and sex-dependent radiation-induced secondary malignancy is breast cancer, in which the risk is small with radiation in women aged <30 years but increases about twentyfold over baseline in women aged >30 years. A 25-year-old woman treated with mantle radiation for Hodgkin’s lymphoma has a 29% actuarial risk of developing breast cancer by age 55.

HORMONAL THERAPY

Treatment of breast cancer with tamoxifen for 5 years or longer is associated with a 1–2% risk of endometrial cancer. Surveillance is generally effective at finding these cancers at an early stage. The risk of mortality from tamoxifen-induced endometrial cancer is low compared to the benefit of tamoxifen as adjuvant therapy for breast cancer.

IMMUNOSUPPRESSIVE THERAPY

Immunosuppressive therapy, as used in allogeneic bone marrow transplantation, particularly with T-cell depletion using antithymocyte globulin or other means, increases the risk of Epstein Barr virus–associated B-cell lymphoproliferative disorder. The incidence at 10 years after T-cell depletion is 9–12%. Discontinuing immunosuppressive therapy, if possible, is often associated with complete disease regression.

RECOMMENDATIONS FOR FOLLOW-UP

All former cancer patients should be followed indefinitely. This is most often done by oncologists, but demographic changes suggest that more primary care physicians will need to be trained in the follow-up of treated cancer patients in remission. Cancer patients need to be educated about signs and symptoms of recurrence and potentially adverse effects related to therapy. Localized pain or palpable abnormality in a previously radiated field should prompt radiographic evaluation. Screening tests, when available and validated, should be used on a routine and regular basis (e.g., mammography and Pap smear), particularly in patients receiving radiation to specific organs. Annual mammography should start no later than 10 years after breast radiation. Patients receiving radiation fields encompassing thyroid tissue should have regular thyroid examinations and TSH testing. Patients treated with alkylating agents or topoisomerase inhibitors should have a complete blood count every 6–12 months, and cytopenias, abnormal cells on peripheral smear, or macrocytosis should be evaluated with bone marrow biopsy and aspirate, to include cytogenetics, flow cytometry, or fluorescence in situ hybridization (FISH) studies as appropriate.

As the population of cancer survivors lives longer and grows, cancer survivorship has become an increasingly recognized subject, and the Institute of Medicine and National Research Council have published a monograph entitled *From Cancer Patient to Cancer Survivor: Lost in Transition.* The monograph proposes a plan that would inform clinicians caring for cancer survivors in complete detail of their previous treatments, complications thereof, signs and symptoms of late effects, and recommended screening and follow-up procedures. **Table 91-2** lists long-term treatment effects by cancer type.

TABLE 91-2

Long-Term Treatment Effects by Cancer Type

Cancer Type	Late Effects
Pediatric cancers	Majority have at least one late effect 30% with moderate/severe problems Cardiovascular: radiation, anthracyclines Lungs: radiation Skeletal abnormalities: radiation Psychological, cognitive, and sexual problems Second neoplasms significant cause of death
Hodgkin's lymphoma	Thyroid dysfunction: radiation Premature coronary artery disease: radiation Gonadal dysfunction: chemotherapy Postsplenectomy sepsis Myelodysplasia Acute myeloid leukemia Non-Hodgkin's lymphomas Breast cancer, lung cancer, and melanoma Fatigue, psychological and sexual problems Peripheral neuropathy
Non-Hodgkin's lymphoma	Myelodysplasia Acute leukemia Bladder cancer Peripheral neuropathy
Acute leukemia	Second malignancies: hematologic, solid tumors Neuropsychiatric dysfunction Subnormal growth Thyroid abnormalities Infertility
Bone marrow stem cell transplantation	Infertility Graft-versus-host disease (allogeneic transplant) Psychosexual dysfunction
Head and neck cancer	Poor dentition, dry mouth, poor nutrition: radiation

Cancer Type	Late Effects
Breast cancer	Tamoxifen: endometrial cancer, blood clots Aromatase inhibitors: osteoporosis, arthritis Cardiomyopathy: anthracycline ± radiation, trastuzumab Acute leukemia Hormone deficiency symptoms: hot flashes, vaginal dryness, dyspareunia Psychosocial dysfunction “Chemo brain”
Testicular cancer	Raynaud’s phenomenon Renal dysfunction Pulmonary dysfunction Retrograde ejaculation: surgery 15% sexual dysfunction
Colon cancer	Major risk is second colon cancer Quality of life high in survivors
Prostate cancer	Impotence Urinary incontinence (0–15%) Chronic proctitis, prostatitis/cystitis: radiation

OUTLOOK

Clearly, the challenge for the future is to combine chemotherapy, targeted agents, biologic therapies, radiation, and surgery to produce better outcomes with less toxicity, including late effects of therapy. This is easily said but less easily accomplished. As treatment becomes more effective in new patient populations (ovarian, bladder, anal, and laryngeal cancers, for example), we will expect to discover new populations at risk for late effects. These populations will need to be followed carefully, so that such effects are recognized and treated. Cancer survivors represent an underused resource for prevention studies. Childhood cancer survivors, especially, suffer multiple chronic health impairments. The incidence of these late treatment consequences appears to have no plateau with age, throwing in stark relief the necessity of close monitoring and therapies with fewer late consequences of treatment.

FURTHER READING

Bhave M et al: Cardiovascular toxicity of biologic agents for cancer therapy. Oncology (Williston Park) 28:482, 2014. [PubMed: 25134321]

Herrmann J et al: Vascular toxicities of cancer therapies: The old and the new—An evolving avenue. Circulation 133:1272, 2016. [PubMed: 27022039]

Moslehi JJ: Cardiovascular toxic effects of targeted cancer therapies. N Engl J Med 375:1457, 2016. [PubMed: 27732808]

Postow MA et al: Immune-related events associated with immune checkpoint blockade. N Engl J Med 378:158, 2018.

Shahrokni A et al: Long-term toxicity of cancer treatment in older patients. Clin Geriatr Med 32:63, 2016. [PubMed: 26614861]

Vichaya EG et al: Mechanisms of chemotherapy-induced behavioral toxicities. Front Neurosci 9:1, 2015. [PubMed: 25653585]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 92: Hematopoietic Stem Cells

David T. Scadden; Dan L. Longo

INTRODUCTION

All of the cell types in the peripheral blood and some cells in every tissue of the body are derived from hematopoietic (*hemo*: blood; *poiesis*: creation) stem cells. If the hematopoietic stem cell is damaged and can no longer function (e.g., due to a nuclear accident), a person would survive 2–4 weeks in the absence of extraordinary support measures. With the clinical use of hematopoietic stem cells, tens of thousands of lives are saved each year **(Chap. 110)**. Stem cells produce hundreds of billions of blood cells daily from a stem cell pool that is estimated to be only in the tens of thousands. How stem cells do this, how they persist for many decades despite the production demands, and how they may be better used in clinical care are important issues in medicine.

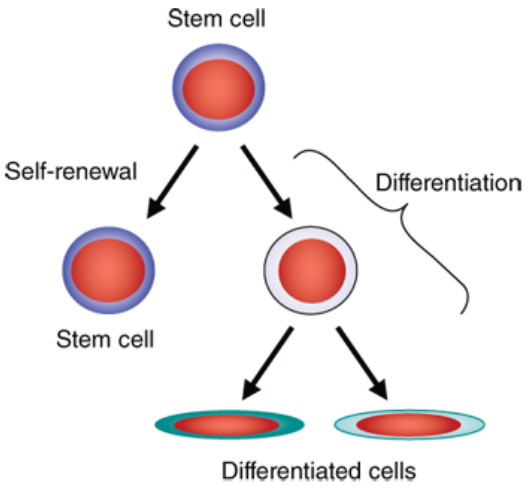
The study of blood cell production has become a paradigm for how other tissues may be organized and regulated. Basic research in hematopoiesis includes defining stepwise molecular changes accompanying functional changes in maturing cells, aggregating cells into functional subgroups, and demonstrating hematopoietic stem cell regulation by a specialized microenvironment; these concepts are worked out in hematology, but they offer models for other tissues. Moreover, these concepts may not be restricted to normal tissue function but extend to malignancy. Stem cells are rare cells among a heterogeneous population of cell types, and their behavior is assessed mainly in experimental animal models involving reconstitution of hematopoiesis. Thus, much of what we know about stem cells is imprecise and based on inferences from genetically manipulated animals.

CARDINAL FUNCTIONS OF HEMATOPOIETIC STEM CELLS

All stem cell types have two cardinal functions: self-renewal and differentiation **(Fig. 92-1)**. Stem cells exist to generate, maintain, and repair tissues. They function successfully if they can replace a wide variety of shorter-lived mature cells over prolonged periods. The process of self-renewal (see below) assures that a stem cell population can be sustained over time. Without self-renewal, the stem cell pool would become exhausted and tissue maintenance would not be possible. The process of differentiation leads to production of the effectors of tissue function: mature cells. Without proper differentiation, the integrity of tissue function would be compromised and organ failure or neoplasia would ensue.

FIGURE 92-1

Signature characteristics of the stem cell. Stem cells have two essential features: the capacity to differentiate into a variety of mature cell types and the capacity for self-renewal. Intrinsic factors associated with self-renewal include expression of *Bmi-1, Gfi-1, PTEN, STAT5, Tel/Atv6, p21, p18, MCL-1, Mel-18, RAE28*, and *HoxB4*. Extrinsic signals for self-renewal include Notch, Wnt, SHH, angiogenin, and Tie2/Ang-1. Based mainly on murine studies, hematopoietic stem cells express the following cell surface molecules: CD34, Thy-1 (CD90), c-Kit receptor (CD117), CD133, CD164, and c-Mpl (CD110, also known as the thrombopoietin receptor).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In the blood, mature cells have variable average life spans, ranging from hours for mature neutrophils to a few months for red blood cells to many years for memory lymphocytes. However, the stem cell pool is the central, durable source of all blood and immune cells, maintaining a capacity to produce a broad range of cells from a single cell source, yet keeping itself vigorous over decades of life. As an individual stem cell divides, it has the capacity to accomplish one of three division outcomes: two stem cells, two cells destined for differentiation, or one stem cell and one differentiating cell. The former two outcomes are the result of symmetric cell division, whereas the latter indicates a different outcome for the two daughter cells—an event termed *asymmetric cell division*. The relative balance for these types of outcomes may change during development and under particular kinds of demands on the stem cell pool.

DEVELOPMENTAL BIOLOGY OF HEMATOPOIETIC STEM CELLS

During development, blood cells are produced at different sites. Initially, the yolk sac provides oxygen-carrying red blood cells and many of the macrophage-like cells that are resident in tissues: cells like microglia in the brain. The placenta and several sites of intraembryonic blood cell production then become involved in sequential order. These move from the genital ridge at a site where the aorta, gonadal tissue, and mesonephros are emerging to the fetal liver and then, in the second trimester, to the bone marrow and spleen. As the location of stem cells changes, the cells they produce also change. The yolk sac provides red cells expressing embryonic hemoglobins and tissue-resident macrophages. Intraembryonic sites of hematopoiesis generate stem cells, red cells, platelets, and the circulating cells of innate immunity. The production of the cells of adaptive immunity occurs when the bone marrow is colonized and the thymus forms. Stem cell proliferation remains high, even in the bone marrow, until shortly after birth, when it appears to dramatically decline. The cells in the bone marrow are thought to arrive by the bloodborne transit of cells from the fetal liver after calcification of the long bones has begun. The presence of stem cells in the circulation is not unique to a time window in development, however, as hematopoietic stem cells circulate throughout life. The time that stem cells spend freely circulating appears to be brief (measured in minutes in the mouse), but the stem cells that do circulate are functional and can be used for transplantation. The number of stem cells that circulate can be increased in a number of ways to facilitate harvest and transfer to the same or a different host.

MOBILITY OF HEMATOPOIETIC STEM CELLS

Cells entering and exiting the bone marrow do so through a series of molecular interactions. Circulating stem cells (through CD162 and CD44) engage the lectins (carbohydrate binding proteins) P- and E-selectin on the endothelial surface to slow the movement of the cells to a rolling phenotype. Stem cell integrins are then activated and accomplish firm adhesion between the stem cell and vessel wall, with a particularly important role for stem cell VCAM-1 engaging endothelial VLA-4. The chemokine CXCL12 (SDF1) interacting with stem cell CXCR4 receptors and ionic calcium interacting with the calcium sensing receptor appear to be important in the process of stem cells getting from the circulation to where they engraft in the bone marrow. This is particularly true in the developmental move from fetal liver to bone marrow.

However, the role for CXCR4 in adults appears to be more related to retention of stem cells in the bone marrow rather than the process of getting them there. Interrupting that retention process through either specific molecular blockers of the CXCR4/CXCL12 interaction, cleavage of CXCL12, or downregulation of the CXCR4 receptor can all result in the release of stem cells into the circulation. This process is an increasingly important aspect of recovering stem cells for therapeutic use as it has permitted the harvesting process to be done by leukapheresis rather than bone marrow punctures in the operating room. Granulocyte colony-stimulating factor and plerixafor, a macrocyclic compound that can block CXCR4, are both used clinically to mobilize marrow hematopoietic stem cells for transplant. Refining our knowledge of how stem cells get into and out of the bone marrow may improve our ability to obtain stem cells and make them more efficient at finding their way to the specific sites for blood cell production, the so-called stem cell niche.

HEMATOPOIETIC STEM CELL MICROENVIRONMENT

The concept of a specialized microenvironment, or stem cell niche, was first proposed to explain why cells derived from the bone marrow of one animal could be used in transplantation and again be found in the bone marrow of the recipient. This niche is more than just a housing site for stem cells, however. It is an anatomic location where regulatory signals are provided that allow the stem cells to thrive, to expand if needed, and to provide varying amounts of descendant daughter cells. In addition, unregulated growth of stem cells may be problematic based on their undifferentiated state and self-renewal capacity. Thus, the niche must also regulate the number of stem cells produced. In this manner, the niche has the dual function of serving as a site of nurture but imposing limits for stem cells: in effect, acting as both a nutritive and constraining home.

The niche for blood stem cells changes with each of the sites of blood production during development, but for most of human life it is located in the bone marrow. Within the bone marrow, the perivascular space particularly in regions of trabecular bone serves as a niche. The mesenchymal and endothelial cells of the marrow microvessels produce kit ligand and CXCL12, both known to be important for hematopoietic stem cells. Other cell types, such as sympathetic neurons, nonmyelinating Schwann cells, macrophages, megakaryocytes, osteoclasts, and osteoblasts, have been shown to regulate stem cells, some by direct and others by indirect effects. Extracellular matrix proteins like osteopontin and heparan sulfates also affect stem cell function. The endosteal region appears to be particularly important for transplanted cells, in part because many of the mesenchymal cells and sinusoidal blood vessels of the central marrow are disrupted by the conditioning regimens used to prepare a patient for transplantation. The functioning of the niche as a supportive context for stem cells is of obvious importance for maintaining hematopoiesis and in transplantation. An active area of study involves determining whether the niche is altered in disease and whether drugs can modify niche function to improve transplantation or normal stem cell function in hematologic disease.

EXCESS CAPACITY OF HEMATOPOIETIC STEM CELLS

In the absence of disease, one never runs out of hematopoietic stem cells. Indeed, serial transplantation studies in mice suggest that sufficient stem cells are present to reconstitute several animals in succession, with each animal having normal blood cell production. The fact that allogeneic stem cell transplant recipients also never run out of blood cells in their life span, which can extend for decades, argues that even the limiting numbers of stem cells provided to them are sufficient. How stem cells respond to different conditions to increase or decrease their mature cell production remains poorly understood. Clearly, negative feedback mechanisms affect the level of production of most of the cells, leading to the normal tightly regulated blood cell counts. However, many of the regulatory mechanisms that govern production of more mature progenitor cells do not apply or apply differently to stem cells. Similarly, most of the molecules shown to be able to change the size of the stem cell pool have little effect on more mature blood cells. For example, the growth factor erythropoietin, which stimulates red blood cell production from more mature precursor cells, has no effect on stem cells. Similarly, granulocyte colony-stimulating factor drives the rapid proliferation of granulocyte precursors but has little or no effect on the cell cycling of stem cells. Rather, it changes the location of stem cells by indirect means, altering molecules such as CXCL12 that tether stem cells to their niche. Molecules shown to be important for altering the proliferation, self-renewal, or survival of stem cells, such as cyclin-dependent kinase inhibitors, transcription factors like *Bmi-1*, microRNA-processing enzymes like Dicer, or even metabolic regulators like pyruvate kinase isoforms have little or different effects on progenitor cells. Hematopoietic stem cells have governing mechanisms that are distinct from the cells they generate.

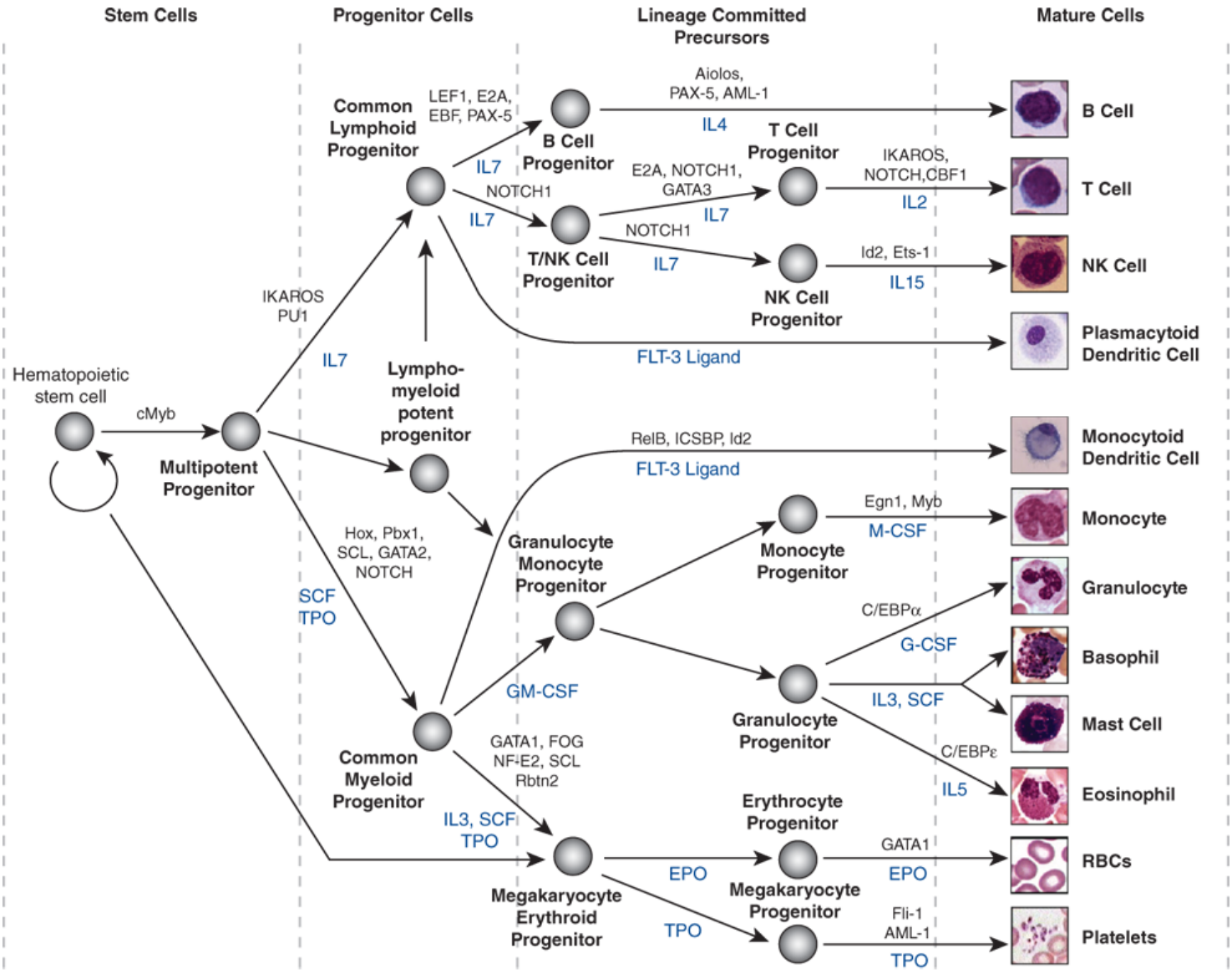
HEMATOPOIETIC STEM CELL DIFFERENTIATION

Hematopoietic stem cells sit at the base of a branching hierarchy of cells culminating in the many mature cell types that compose the blood and immune system (Fig. 92-2). The maturation steps leading to terminally differentiated and functional blood cells take place both as a consequence of intrinsic changes in gene expression and niche-directed and cytokine-directed changes in the cells. Our knowledge of the details remains incomplete. As stem cells mature to progenitors, precursors, and, finally, mature effector cells, they undergo a series of functional changes. These include the obvious acquisition of functions defining mature blood cells, such as phagocytic capacity or hemoglobin synthesis. They also include the progressive loss of plasticity (i.e., the ability to become other cell types). For example, the myeloid progenitor can make all cells in the myeloid series but none in the lymphoid series. As common myeloid progenitors mature, they become precursors for either monocytes and granulocytes or erythrocytes and megakaryocytes, but not both. Some amount of reversibility of this process may exist early in the differentiation cascade, but that is lost beyond a distinct stage in normal physiologic conditions. With genetic interventions, however, blood cells, like other somatic cells, can be reprogrammed to become a variety of cell types.

FIGURE 92-2

Hierarchy of hematopoietic differentiation. *Stem cells* are multipotent cells that are the source of all descendant cells and have the capacity to provide either long-term (measured in years) or short-term (measured in months) cell production. *Progenitor cells* have a more limited spectrum of cells they can produce and are generally a shorter-lived, highly proliferative population also known as transient amplifying cells. *Precursor cells* are cells committed to a single blood cell lineage but with a continued ability to proliferate; they do not have all the features of a fully mature cell. *Mature cells* are the terminally differentiated product of the differentiation process and are the effector cells of specific activities of the blood and

immune system. Progress through the pathways is mediated by alterations in gene expression. The regulation of the differentiation by soluble factors and cell-cell communications within the bone marrow niche are still being defined. The transcription factors that characterize particular cell transitions are illustrated on the *arrows*; the soluble factors that contribute to the differentiation process are in *blue*. This picture is a simplification of the process. Active research is revealing multiple discrete cell types in the maturation of B cells and T cells and has identified cells that are biased toward one lineage or another (rather than uncommitted) in their differentiation. EPO, erythropoietin; RBC, red blood cell; SCF, stem cell factor; TPO, thrombopoietin.

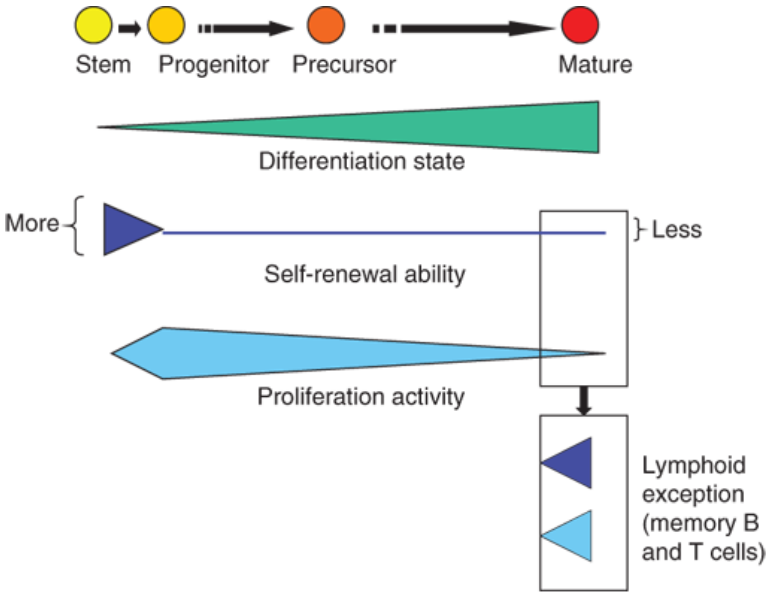


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

As cells differentiate, they may also lose proliferative capacity **(Fig. 92-3)**. Mature granulocytes are incapable of proliferation and only increase in number by increased production from precursors. The exceptions to the rule are some tissue-resident macrophages, which appear capable of proliferation, and lymphoid cells. Lymphoid cells retain the capacity to proliferate but have linked their proliferation to the recognition of particular proteins or peptides by specific antigen receptors on their surface. Like many tissues with short-lived mature cells such as the skin and intestine, blood cell proliferation is largely accomplished by a more immature progenitor population. In general, cells within the highly proliferative progenitor cell compartment are also relatively short-lived, making their way through the differentiation process in a defined molecular program involving the sequential activation of particular sets of genes. For any particular cell type, the differentiation program is difficult to speed up. The time it takes for hematopoietic progenitors to become mature cells is ~10–14 days in humans, evident clinically by the interval between cytotoxic chemotherapy and blood count recovery in patients.

FIGURE 92-3

Relative function of cells in the hematopoietic hierarchy. The boxes represent distinct functional features of cells in the myeloid (*upper box*) versus lymphoid (*lower box*) lineages.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although hematopoietic stem cells are generally thought to have the capacity to form all cells of the blood, it is becoming clear that individual stem cells may not be equal in their differentiation potential. That is, some stem cells are “biased” to become mature cells of a particular type. In addition, the general concept of cells having a binary choice of lymphoid or myeloid differentiation is not entirely accurate. A cell population with limited megakaryocytic and erythroid or myeloid (monocyte and granulocyte) and lymphoid potential is now added to the commitment steps stem cells may undergo.

SELF-RENEWAL

The hematopoietic stem cell must balance its three potential fates: apoptosis, self-renewal, and differentiation. The proliferation of cells is generally not associated with the ability to undergo a self-renewing division except among memory T and B cells and among stem cells. Self-renewal capacity has generally been regarded as giving way to differentiation as the only option after cell division when cells leave the stem cell compartment, unless they become memory lymphocytes. However, emerging data suggest that some myeloid committed progenitors may have self-renewing potential in vivo, providing long-term production of cells. Stem cells all have self-renewing capacity by definition, and they have an additional feature characterizing their proliferation machinery. Stem cells in many mature adult tissues are heterogeneous with some being deeply quiescent, serving as a deep reserve, whereas others are more proliferative and replenish the short-lived progenitor population. In the hematopoietic system, stem cells are generally cytokine-resistant, remaining dormant even when cytokines drive bone marrow progenitors to proliferation rates measured in hours. Stem cells, in contrast, are thought to divide at far longer intervals, measured in months to years, for the most quiescent cells. This quiescence is difficult to overcome in vitro, limiting the ability to effectively expand human hematopoietic stem cells. The process may be controlled by particularly high levels of cyclin-dependent kinase inhibitors like p57 or CDKN1c that restrict entry of stem cells into the cell cycle, blocking the G_1-S transition. Exogenous signals from the niche also appear to enforce quiescence, including angiogenin, interleukin-18, and perhaps angiopoietin 1.

The regulation of stem cell proliferation also appears to change with age. In mice, the cyclin-dependent kinase inhibitor p16INK4a accumulates in stem cells in older animals and is associated with a change in stem cell functions, including cell cycling. Lowering expression of p16INK4a in older animals improves stem cell cycling and capacity to reconstitute hematopoiesis in adoptive hosts, making them similar to younger animals. Mature cell numbers are unaffected. Therefore, molecular events governing the specific functions of stem cells are being gradually made clear and offer the potential of new approaches to changing stem cell function for therapy. One critical stem cell function that remains poorly defined is the molecular regulation of self-renewal.

For medicine, self-renewal is perhaps the most important function of stem cells because it is critical in regulating the number of stem cells. Stem cell number is a key limiting parameter for both autologous and allogeneic stem cell transplantation. Were we to have the ability to use fewer stem cells or expand limited numbers of stem cells ex vivo, it might be possible to reduce the morbidity and expense of stem cell harvests and enable use of other stem cell sources. Specifically, umbilical cord blood is a rich source of stem cells. However, the volume of cord blood units is extremely small, and therefore, the total number of hematopoietic stem cells that can be obtained in any single cord blood unit is generally only sufficient to transplant an individual of <40 kg. This limitation restricts what would otherwise be an extremely promising source of stem cells. Two features of cord blood stem cells are particularly important: (1) They are derived from a diversity of individuals that far exceeds the adult donor pool and

therefore can overcome the majority of immunologic cross-matching obstacles. (2) Cord blood stem cells have a large number of T cells associated with them, but (paradoxically) they appear to be associated with a lower incidence of graft-versus-host disease when compared with similarly mismatched stem cells from other sources. If stem cell expansion by self-renewal could be achieved, the number of cells available might be sufficient for use in larger adults. An alternative approach to this problem is to improve the efficiency of engraftment of donor stem cells. Graft engineering is exploring methods of adding cell components that may enhance engraftment. Furthermore, at least some data suggest that depletion of host NK (natural killer) cells may lower the number of stem cells necessary to reconstitute hematopoiesis.

Some limited understanding of self-renewal exists and, intriguingly, implicates gene products that are associated with the chromatin state, a high-order organization of chromosomal DNA that influences transcription. These include members of the polycomb family, a group of zinc finger–containing transcriptional regulators that interact with the chromatin structure, contributing to the accessibility of groups of genes for transcription. One member, *Bmi-1*, is important in enabling hematopoietic stem cell self-renewal through modification of cell cycle regulators such as the cyclin-dependent kinase inhibitors. In the absence of *Bmi-1* or of the transcriptional regulator, Gfi-1, hematopoietic stem cells decline in number and function. In contrast, dysregulation of *Bmi-1* has been associated with leukemia; it may promote leukemic stem cell self-renewal when it is overexpressed. Other transcription regulators have also been associated with self-renewal, particularly homeobox, or “hox,” genes. These transcription factors are named for their ability to govern large numbers of genes, including those determining body patterning in invertebrates. HoxB4 is capable of inducing extensive self-renewal of stem cells through its DNA-binding motif. Other members of the hox family of genes have been noted to affect normal stem cells, but they are also associated with leukemia. External signals that may influence the relative self-renewal versus differentiation outcomes of stem cell cycling include specific Wnt ligands. Intracellular signal transducing intermediates are also implicated in regulating self-renewal. They include PTEN, an inhibitor of the AKT pathway, and STAT5, both of which are downstream of activated growth factor receptors and necessary for normal stem cell functions including self-renewal, at least in mouse models. The connections between these molecules remain to be defined, and their role in physiologic regulation of stem cell self-renewal is still poorly understood.

CANCER IS SIMILAR TO AN ORGAN WITH SELF-RENEWING CAPACITY

The relationship of stem cells to cancer is an important evolving dimension of adult stem cell biology. Cancer may share principles of organization with normal tissues. Cancer cells are heterogeneous even within a given patient and may have a hierarchical organization of cells with a base of stem-like cells capable of the signature stem cell features: self-renewal and differentiation. These stem-like cells might be the basis for perpetuation of the tumor and represent a slowly dividing, rare population with distinct regulatory mechanisms, including a relationship with a specialized microenvironment. A subpopulation of self-renewing cells has been defined for some, but not all, cancers. A more sophisticated understanding of the stem cell organization of cancers may lead to improved strategies for developing new therapies for the many common and difficult-to-treat types of malignancies that have been relatively refractory to interventions aimed at dividing cells.

Does the concept of cancer stem cells provide insight into the cellular origin of cancer? The fact that some cells within a cancer have stem cell–like properties does not necessarily mean that the cancer arose in the stem cell itself. Rather, more mature cells could have acquired the self-renewal characteristics of stem cells. Any single genetic event is unlikely to be sufficient to enable full transformation of a normal cell to a frankly malignant one. Rather, cancer is a multistep process, and for the multiple steps to accumulate, the cell of origin must be able to persist for prolonged periods. It must also be able to generate large numbers of daughter cells. The normal stem cell has these properties and, by virtue of its having intrinsic self-renewal capability, may be more readily converted to a malignant phenotype. This hypothesis has been tested experimentally in the hematopoietic system. Taking advantage of the cell-surface markers that distinguish hematopoietic cells of varying maturity, stem cells, progenitors, precursors, and mature cells can be isolated. Powerful transforming gene constructs were placed in these cells, and it was found that the cell with the greatest potential to produce a malignancy was dependent on the transforming gene. In some cases, it was the stem cell, but in others, the progenitor cell functioned to initiate and perpetuate the cancer. This shows that cells can acquire stem cell–like properties in malignancy.

WHAT ELSE CAN HEMATOPOIETIC STEM CELLS DO?

Some experimental data have suggested that hematopoietic stem cells or other cells mobilized into the circulation by the same factors that mobilize hematopoietic stem cells are capable of playing a role in healing the vascular and tissue damage associated with stroke and myocardial infarction. These data are controversial, and the applicability of a stem cell approach to nonhematopoietic conditions remains experimental. However, reprogramming technology offers the potential for using the readily obtained hematopoietic stem cell as a source for cells with other capabilities.

The stem cell, therefore, represents a true dual-edged sword. It has tremendous healing capacity and is essential for life. Uncontrolled, it can threaten the life it maintains. Understanding how stem cells function, the signals that modify their behavior, and the tissue niches that modulate stem cell responses to injury and disease are critical for more effectively developing stem cell–based medicine. That aspect of medicine will include

the use of the stem cells and the use of drugs to target stem cells to enhance repair of damaged tissues. It will also include the careful balance of interventions to control stem cells where they may be dysfunctional or malignant.

FURTHER READING

Hofer T, Rodewald HR: Output without input: The lifelong productivity of hematopoietic stem cells. Curr Opin Cell Biol 43:69, 2016. [PubMed: 27620508]

Hoggatt J, Kfoury Y, Scadden DT: Hematopoietic stem cell niche in health and disease. Annu Rev Pathol 11:555, 2016. [PubMed: 27193455]

Notta F et al: Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science 351:aab2116, 2016. [PubMed: 26541609]

Wahlster L, Daley GQ: Progress towards generation of human haematopoietic stem cells. Nat Cell Biol 18:1111, 2016. [PubMed: 27723718]

Chapter 93: Iron Deficiency and Other Hypoproliferative Anemias

John W. Adamson

INTRODUCTION

Anemias associated with normocytic and normochromic red cells and an inappropriately low reticulocyte response (reticulocyte index <2–2.5) are *hypoproliferative anemias*. This category includes early iron deficiency (before hypochromic microcytic red cells develop), acute and chronic inflammation (including many malignancies), renal disease, hypometabolic states such as protein malnutrition and endocrine deficiencies, and anemias from marrow damage. **Marrow damage states are discussed in Chap. 98.**

Hypoproliferative anemias are the most common anemias, and in the clinic, iron deficiency anemia is the most common of these followed by the anemia of inflammation. The anemia of inflammation, similar to iron deficiency, is related in part to abnormal iron metabolism. The anemias associated with renal disease, inflammation, cancer, and hypometabolic states are characterized by a suboptimal erythropoietin response to the anemia.

IRON METABOLISM

Iron is a critical element in the function of all cells, although the amount of iron required by individual tissues varies during development. At the same time, the body must protect itself from free iron, which is highly toxic in that it participates in chemical reactions that generate free radicals such as singlet O_2 or OH^-.

Consequently, elaborate mechanisms have evolved that allow iron to be made available for physiologic functions while at the same time conserving this element and handling it in such a way that toxicity is avoided.

The major role of iron in mammals is to carry O_2 as part of hemoglobin. O_2 is also bound by myoglobin in muscle. Iron is a critical element in iron-containing enzymes, including the cytochrome system in mitochondria. Iron distribution in the body is shown in **Table 93-1**. Without iron, cells lose their capacity for electron transport and energy metabolism. In erythroid cells, hemoglobin synthesis is impaired, resulting in anemia and reduced O_2 delivery to tissue.

TABLE 93-1

Body Iron Distribution

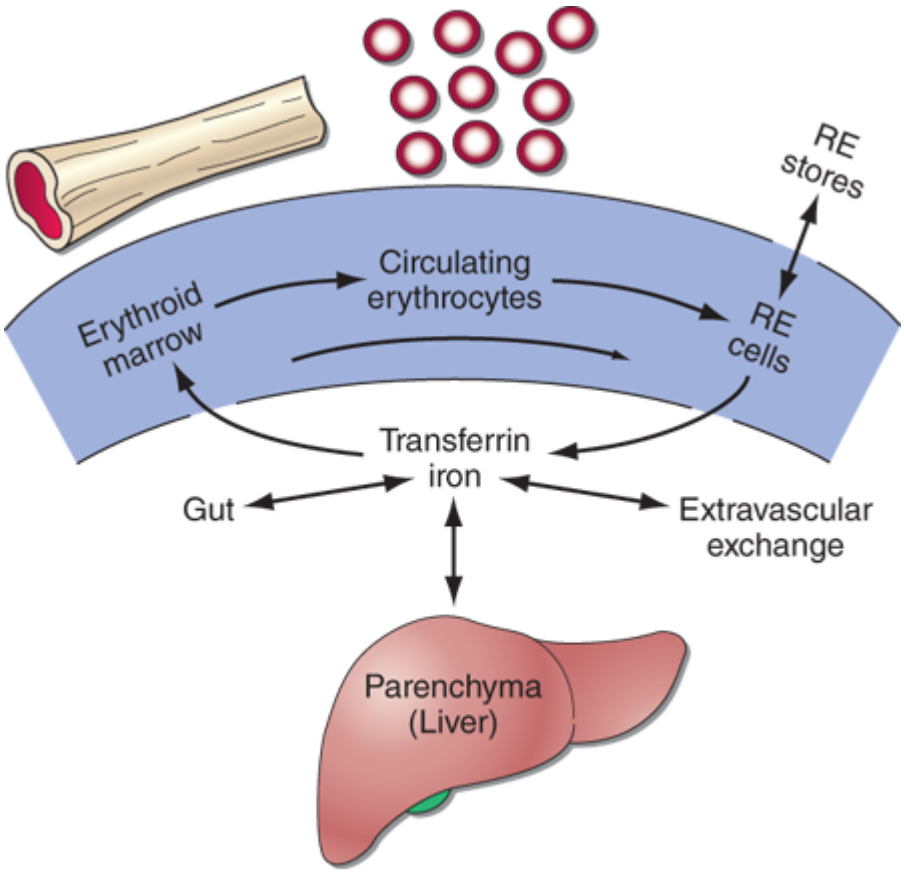
	Iron Content, mg	
	Adult Male, 80 kg	**Adult Female, 60 kg**
Hemoglobin	2500	1700
Myoglobin/enzymes	500	300
Transferrin iron	3	3
Iron stores	600–1000	0–300

THE IRON CYCLE IN HUMANS

Figure 93-1 outlines the major pathways of internal iron exchange in humans. Iron absorbed from the diet or released from stores circulates in the plasma bound to *transferrin*, the iron transport protein. Transferrin is a bilobed glycoprotein with two iron-binding sites. Transferrin that carries iron exists in two forms—*monoferric* (one iron atom) or *diferric* (two iron atoms). The turnover (half-clearance time) of transferrin-bound iron is very rapid—typically 60–90 min. Because almost all of the iron transported by transferrin is delivered to the erythroid marrow, the clearance time of transferrin-bound iron from the circulation is affected most by the plasma iron level and the erythroid marrow activity. When erythropoiesis is markedly stimulated, the pool of erythroid cells requiring iron increases, and the clearance time of iron from the circulation decreases. The half-clearance time of iron in the presence of iron deficiency is as short as 10–15 min. With suppression of erythropoiesis, the plasma iron level typically increases, and the half-clearance time may be prolonged to several hours. Normally, the iron bound to transferrin turns over 6–8 times per day. Assuming a normal plasma iron level of 80–100 μg/dL, the amount of iron passing through the transferrin pool is 20–24 mg/d.

FIGURE 93-1

Internal iron exchange. Normally 80% of iron passing through the plasma transferrin pool is recycled from senescent red cells. Absorption of ~1 mg/d is required from the diet in men, and 1.4 mg/d in women to maintain homeostasis. As long as transferrin saturation is maintained between 20 and 60% and erythropoiesis is not increased, use of iron stores is not required. However, in the event of blood loss, dietary iron deficiency, or inadequate iron absorption, up to 40 mg/d of iron can be mobilized from stores. RE, reticuloendothelial.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The iron-transferrin complex circulates in the plasma until it interacts with specific *transferrin receptors* on the surface of marrow erythroid cells. Diferric transferrin has the highest affinity for transferrin receptors; apotransferrin (not carrying iron) has very little affinity. Although transferrin receptors are found on cells in many tissues within the body—and all cells at some time during development will display transferrin receptors—the cell having the greatest number of receptors (300,000–400,000/cell) is the developing erythroblast.

Once the iron-bearing transferrin interacts with its receptor, the complex is internalized via clathrin-coated pits and transported to an acidic endosome, where the iron is released at the low pH. The iron is then made available for heme synthesis while the transferrin-receptor complex is recycled to the surface of the cell, where the bulk of the transferrin is released back into circulation and the transferrin receptor re-anchors into the cell membrane. At this point a certain amount of the transferrin receptor protein may be released into circulation and can be measured as soluble transferrin receptor protein. Within the erythroid cell, iron in excess of the amount needed for hemoglobin synthesis binds to a storage protein, *apoferritin*, forming *ferritin*. This mechanism of iron exchange also takes place in other cells of the body expressing transferrin receptors, especially liver parenchymal cells where the iron can be incorporated into heme-containing enzymes or stored. The iron incorporated into hemoglobin subsequently enters the circulation as new red cells are released from the bone marrow. The iron is then part of the red cell mass and will not become available for reutilization until the red cell dies.

In a normal individual, the average red cell life span is 120 days. Thus, 0.8–1% of red cells is replaced each day. At the end of its life span, the red cell is recognized as senescent by the cells of the *reticuloendothelial (RE) system*, and the red cell undergoes phagocytosis. Once within the RE cell, the ingested hemoglobin is broken down, the globin and other proteins are returned to the amino acid pool, and the iron is shuttled back to the surface of the RE cell, where it is presented to circulating transferrin. It is the efficient and highly conserved recycling of iron from senescent red cells that supports steady-state (and even mildly accelerated) erythropoiesis.

Because each milliliter of red cells contains 1 mg of elemental iron, the amount of iron needed to replace those red cells lost through senescence amounts to 20 mg/d (assuming an adult with a red cell mass of 2 L). Any additional iron required for daily red cell production comes from the diet. Normally, an adult male will need to absorb at least 1 mg of elemental iron daily to meet needs, while females in the childbearing years will need to absorb an average of 1.4 mg/d. However, to achieve a maximum proliferative erythroid marrow response to anemia, additional iron must be available. With markedly stimulated erythropoiesis, demands for iron are increased by as much as six- to eightfold. With extravascular hemolytic anemia, the rate of red cell destruction is increased, but the iron recovered from the red cells is efficiently reutilized for hemoglobin synthesis. In contrast, with intravascular hemolysis or blood loss anemia, the rate of red cell production is limited by the amount of iron that can be mobilized from stores. Typically, the rate of mobilization under these circumstances will not support red cell production more than 2.5 times normal. If the delivery of iron to the stimulated marrow is suboptimal, the marrow's proliferative response is blunted, and hemoglobin synthesis is impaired. The result is a hypoproliferative marrow accompanied by microcytic, hypochromic anemia.

Whereas blood loss or hemolysis places a demand on the iron supply, inflammatory conditions interfere with iron release from stores and can result in a rapid decrease in the serum iron (see below).

NUTRITIONAL IRON BALANCE

The balance of iron in humans is tightly controlled and designed to conserve iron for reutilization. There is no regulated excretory pathway for iron, and the only mechanisms by which iron is lost are blood loss (via gastrointestinal bleeding, menses, or other forms of bleeding) and the loss of epithelial cells from the skin, gut, and genitourinary tract. Normally, the only route by which iron comes into the body is via absorption from food or from medicinal iron taken orally. Iron may also enter the body through red cell transfusions or injection of iron complexes. The margin between the amount of iron available for absorption and the requirement for iron in growing infants and the adult female is narrow; this accounts for the great prevalence of iron deficiency worldwide—currently estimated at more than one billion people.

The amount of iron required from the diet to replace losses averages ~10% of body iron content a year in men and 15% in women of childbearing age. Dietary iron content is closely related to total caloric intake (~6 mg of elemental iron per 1000 calories). Iron bioavailability is affected by the nature of the foodstuff, with heme iron (e.g., red meat) being most readily absorbed. In the United States, the average iron intake in an adult male is 15 mg/d with 6% absorption; for the average female, the intake is 11 mg/d with 12%

absorption. An individual with iron deficiency can increase iron absorption to ~20% of the iron present in a meat-containing diet but only 5–10% of the iron in a vegetarian diet. As a result, one-third of the female population in the United States has virtually no iron stores. Vegetarians are at an additional disadvantage because certain foodstuffs that include phytates and phosphates reduce iron absorption by ~50%. When ionizable iron salts are given together with food, the amount of iron absorbed is reduced. When the percentage of iron absorbed from individual food items is compared with the percentage for an equivalent amount of ferrous salt, iron in vegetables is only about one-twentieth as available, egg iron one-eighth, liver iron one-half, and heme iron one-half to two-thirds.

Infants, children, and adolescents may be unable to maintain normal iron balance because of the demands of body growth and lower dietary intake of iron. During the last two trimesters of pregnancy, daily iron requirements increase to 5–6 mg, and iron supplements are strongly recommended for pregnant women in developed countries.

Iron absorption takes place largely in the duodenum and proximal small intestine and is a carefully regulated process. For absorption, iron must be taken up by the luminal cell. That process is facilitated by the acidic contents of the stomach, which maintains the iron in solution. At the brush border of the absorptive cell, the ferric iron is converted to the ferrous form by a ferrireductase. Transport across the membrane is accomplished by divalent metal transporter type 1 (DMT-1, also known as natural resistance macrophage-associated protein type 2 [Nramp 2] or DCT-1). DMT-1 is a general cation transporter. Once inside the gut cell, iron may be stored as ferritin or transported through the cell to be released at the basolateral surface to plasma transferrin through the membrane-embedded iron exporter, ferroportin. The function of ferroportin is negatively regulated by hepcidin, the principal iron regulatory hormone. In the process of release, iron interacts with another ferroxidase, hephaestin, which oxidizes the iron to the ferric form for transferrin binding. Hephaestin is similar to ceruloplasmin, the copper-carrying protein.

Iron absorption is influenced by a number of physiologic states. Erythroid hyperplasia stimulates iron absorption even in the face of normal or increased iron stores, and hepcidin levels are inappropriately low. Thus, patients with anemias associated with high levels of ineffective erythropoiesis absorb excess amounts of dietary iron. The molecular mechanism underlying this is the production of erythroferrone *(ERFE)* by developing erythroblasts. ERFE suppresses hepcidin production and, over time, this may lead to iron overload and tissue damage. In iron deficiency, hepcidin levels are also low and iron is much more efficiently absorbed; the contrary is true in states of secondary iron overload. The normal individual can reduce iron absorption in situations of excessive intake or medicinal iron intake; however, while the percentage of iron absorbed goes down, the absolute amount goes up. This accounts for the acute iron toxicity occasionally seen when children ingest large numbers of iron tablets. Under these circumstances, the amount of iron absorbed exceeds the transferrin binding capacity of the plasma, resulting in free iron that affects critical organs such as cardiac muscle cells.

IRON-DEFICIENCY ANEMIA

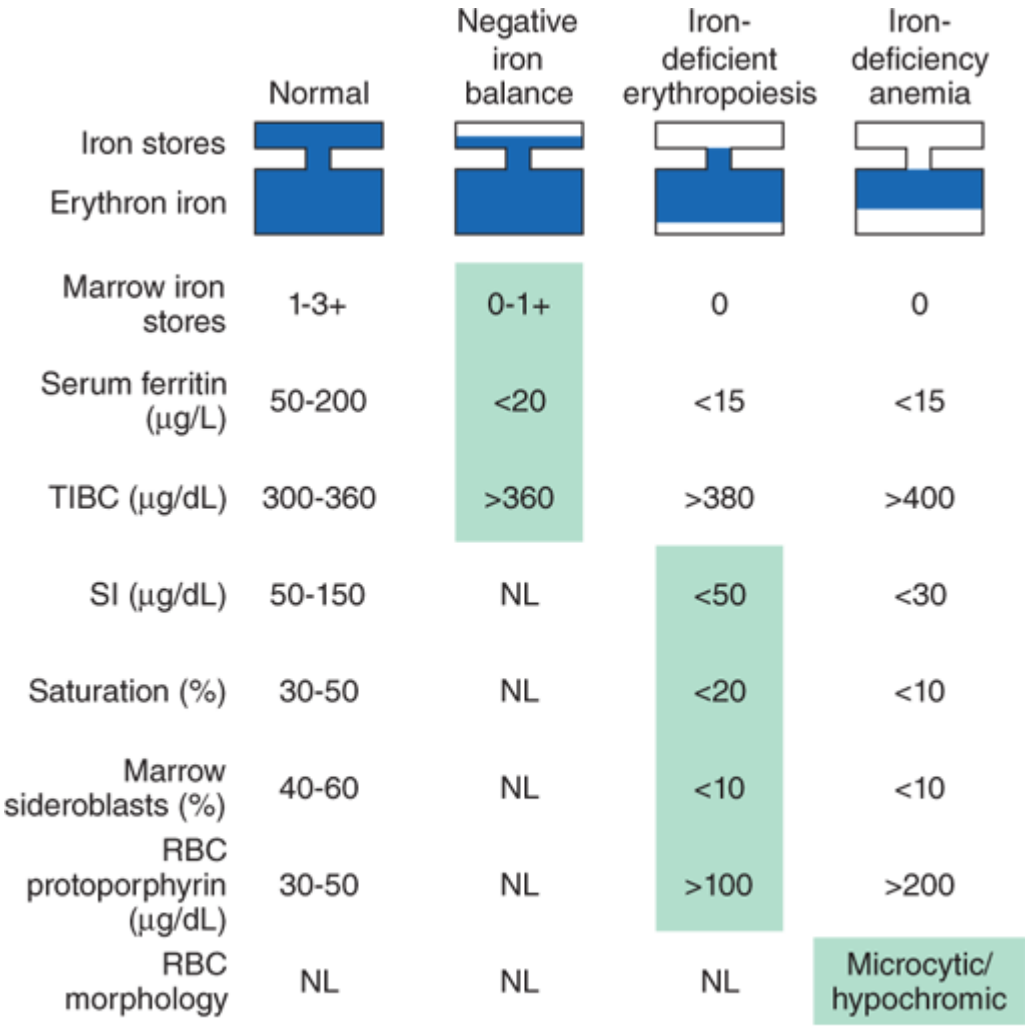
Iron deficiency is one of the most prevalent forms of malnutrition. Globally, 50% of anemia is attributable to iron deficiency and accounts for approximately nearly a million deaths annually worldwide. Africa and parts of Asia bear 71% of the global mortality burden; North America represents only 1.4% of the total morbidity and mortality associated with iron deficiency.

STAGES OF IRON DEFICIENCY

The progression to iron deficiency can be divided into three stages **(Fig. 93-2)**. The first stage is *negative iron balance*, in which the demands for (or losses of) iron exceed the body's ability to absorb iron from the diet. This stage results from a number of physiologic mechanisms, including blood loss, pregnancy (in which the demands for red cell production by the fetus outstrip the mother's ability to provide iron), rapid growth spurts in the adolescent, or inadequate dietary iron intake. Blood loss in excess of 10–20 mL of red cells per day is greater than the amount of iron that the gut can absorb from a normal diet. Under these circumstances, the iron deficit must be made up by mobilization of iron from RE storage sites. During this period, iron stores—reflected by the serum ferritin level or the appearance of stainable iron on bone marrow aspirations—decrease. As long as iron stores are present and can be mobilized, the serum iron, total iron-binding capacity (TIBC), and red cell protoporphyrin levels remain within normal limits. At this stage, red cell morphology and indices are normal.

FIGURE 93-2

Laboratory studies in the evolution of iron deficiency. Measurements of marrow iron stores, serum ferritin, and total iron-binding capacity (TIBC) are sensitive to early iron-store depletion. Iron-deficient erythropoiesis is recognized from additional abnormalities in the serum iron (SI), percent transferrin saturation, the pattern of marrow sideroblasts, and the red blood cell (RBC) protoporphyrin level. Patients with iron-deficiency anemia demonstrate all the same abnormalities plus hypochromic microcytic anemia. *(From RS Hillman, CA Finch: The Red Cell Manual, 7th ed. Philadelphia, F.A. Davis and Co, 1996, with permission.)*



	Normal	Negative iron balance	Iron-deficient erythropoiesis	Iron-deficiency anemia
Marrow iron stores	1-3+	0-1+	0	0
Serum ferritin (μg/L)	50-200	<20	<15	<15
TIBC (μg/dL)	300-360	>360	>380	>400
SI (μg/dL)	50-150	NL	<50	<30
Saturation (%)	30-50	NL	<20	<10
Marrow sideroblasts (%)	40-60	NL	<10	<10
RBC protoporphyrin (μg/dL)	30-50	NL	>100	>200
RBC morphology	NL	NL	NL	Microcytic/ hypochromic

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

When iron stores become depleted, the serum iron begins to fall. Gradually, the TIBC increases, as do red cell protoporphyrin levels. By definition, marrow iron stores are absent when the serum ferritin level is <15 μg/L. As long as the serum iron remains within the normal range, hemoglobin synthesis is unaffected despite the dwindling iron stores. Once the transferrin saturation falls to 15–20%, hemoglobin synthesis becomes impaired. This is a period of *iron-deficient erythropoiesis*. Careful evaluation of the peripheral blood smear reveals the first appearance of microcytic cells, and if the laboratory technology is available, one finds hypochromic reticulocytes in circulation. Gradually, the hemoglobin begins to fall, reflecting *iron-deficiency anemia*. The transferrin saturation at this point is <10–15%.

When moderate anemia is present (hemoglobin 10–13 g/dL), the bone marrow remains hypoproliferative. With more severe anemia (hemoglobin 7–8 g/dL), hypochromia and microcytosis become more prominent, target cells and misshapen red cells (poikilocytes) appear on the blood smear as cigar- or pencil-shaped forms, and the erythroid marrow becomes increasingly ineffective. Consequently, with severe prolonged iron-deficiency anemia, erythroid hyperplasia of the marrow develops, rather than hypoproliferation.

CAUSES OF IRON DEFICIENCY

Conditions that increase demand for iron, increase iron loss, or decrease iron intake or absorption can produce iron deficiency **(Table 93-2)**.

TABLE 93-2

Causes of Iron Deficiency

Increased Demand for Iron
Rapid growth in infancy or adolescence Pregnancy Erythropoietin therapy
Increased Iron Loss
Chronic blood loss Menses Acute blood loss Blood donation Phlebotomy as treatment for polycythemia vera
Decreased Iron Intake or Absorption
Inadequate diet Malabsorption from disease (sprue, Crohn's disease) Malabsorption from surgery (gastrectomy and some forms of bariatric surgery) Acute or chronic inflammation

CLINICAL PRESENTATION OF IRON DEFICIENCY

Certain clinical conditions carry an increased likelihood of iron deficiency. Pregnancy, adolescence, periods of rapid growth, and an intermittent history of blood loss of any kind should alert the clinician to possible iron deficiency. A cardinal rule is that the appearance of iron deficiency in an adult male or post-menopausal female means gastrointestinal blood loss until proven otherwise. Signs related to iron deficiency depend on the severity and chronicity of the anemia in addition to the usual signs of anemia—fatigue, pallor, and reduced exercise capacity. *Cheilosis* (fissures at the corners of the mouth) and *koilonychia* (spooning of the fingernails) are signs of advanced tissue iron deficiency. The diagnosis of iron deficiency is typically based on laboratory results.

LABORATORY IRON STUDIES

Serum Iron and Total Iron-Binding Capacity

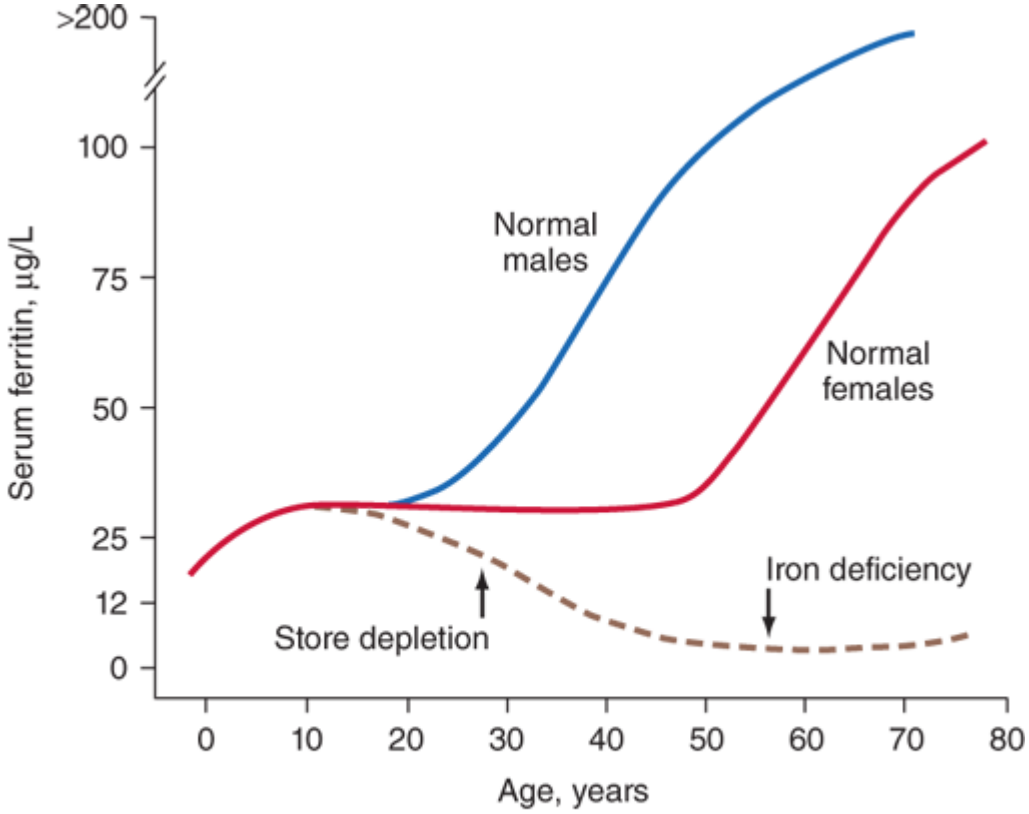
The serum iron level represents the amount of circulating iron bound to transferrin. The TIBC is an indirect measure of the circulating transferrin. The normal range for the serum iron is 50–150 µg/dL; the normal range for TIBC is 300–360 µg/dL. Transferrin saturation, which is normally 25–50%, is obtained by the following formula: serum iron × 100 ÷ TIBC. Iron-deficiency states are associated with saturation levels <20%. There is a diurnal variation in the serum iron. A transferrin saturation >50% indicates that a disproportionate amount of the iron bound to transferrin is being delivered to nonerythroid tissues. If this persists for an extended time, tissue iron overload may occur.

Serum Ferritin

Free iron is toxic to cells, and the body has established an elaborate set of protective mechanisms to bind iron in various tissue compartments. Within cells, iron is stored complexed to protein as ferritin or hemosiderin. Apoferritin binds to free ferrous iron and stores it in the ferric state. As ferritin accumulates within cells of the RE system, protein aggregates are formed as hemosiderin. Iron in ferritin or hemosiderin can be extracted for release by the RE cells, although hemosiderin is less readily available. Under steady-state conditions, the serum ferritin level correlates with total body iron stores; thus, the serum ferritin level is the most convenient laboratory test to estimate iron stores. The normal value for ferritin varies according to the age and gender of the individual **(Fig. 93-3)**. Adult males have serum ferritin values averaging 100 µg/L, while adult females have levels averaging 30 µg/L. As iron stores are depleted, the serum ferritin falls to <15 µg/L. Such levels are diagnostic of absent body iron stores.

FIGURE 93-3

Serum ferritin levels as a function of sex and age. Iron store depletion and iron deficiency are accompanied by a decrease in serum ferritin level below 20 µg/L. *(From RS Hillman et al: Hematology in Clinical Practice, 5th ed. New York, McGraw-Hill, 2011, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Evaluation of Bone Marrow Iron Stores

Although RE iron stores can be estimated from the iron stain of a bone marrow aspirate or biopsy, the measurement of serum ferritin has largely supplanted these procedures for determination of storage iron (Table 93-3). The serum ferritin level is a better indicator of iron overload than the marrow iron stain. However, in addition to storage iron, the marrow iron stain provides information about the effective delivery of iron to developing erythroblasts. Normally, when the marrow smear is stained for iron, 20–40% of developing erythroblasts—called *sideroblasts*—will have visible ferritin granules in their cytoplasm. This represents iron in excess of that needed for hemoglobin synthesis. In states in which release of iron from storage sites is blocked, RE iron will be detectable, and there will be few or no sideroblasts. In the myelodysplastic syndromes, mitochondrial dysfunction can occur, and accumulation of iron in mitochondria appears in a necklace fashion around the nucleus of the erythroblast. Such cells are referred to as *ring sideroblasts*.

TABLE 93-3

Iron Store Measurements

Iron Stores	Marrow Iron Stain, 0–4+	Serum Ferritin, µg/L
0	0	<15
1–300 mg	Trace to 1+	15–30
300–800 mg	2+	30–60
800–1000 mg	3+	60–150
1–2 g	4+	>150
Iron overload	—	>500–1000

Red Cell Protoporphyrin Levels

Protoporphyrin is an intermediate in the pathway to heme synthesis. Under conditions in which heme synthesis is impaired, protoporphyrin accumulates within the red cell. This reflects an inadequate iron supply to erythroid precursors to support hemoglobin synthesis. Normal values are <30 µg/dL of red cells. In iron deficiency, values >100 µg/dL are seen. The most common causes of increased red cell protoporphyrin levels are absolute or relative iron deficiency and lead poisoning.

Serum Levels of Transferrin Receptor Protein

Because erythroid cells have the highest numbers of transferrin receptors of any cell in the body, and because transferrin receptor protein (TRP) is released by cells into the circulation, serum levels of TRP reflect the total erythroid marrow mass. Another condition in which TRP levels are elevated is absolute iron deficiency. Normal values are 4–9 µg/L determined by immunoassay. This laboratory test is becoming increasingly available and, along with the serum ferritin, has been proposed to distinguish between iron deficiency and the anemia of inflammation (see below).

DIFFERENTIAL DIAGNOSIS

Other than iron deficiency, only three conditions need to be considered in the differential diagnosis of a hypochromic microcytic anemia (Table 93-4). The first is an inherited defect in globin chain synthesis: the thalassemias. These are differentiated from iron deficiency most readily by serum iron values; normal or increased serum iron levels and transferrin saturation are characteristic of the thalassemias. In addition, the

red blood cell distribution width (RDW) index is generally normal in thalassemia and elevated in iron deficiency.

TABLE 93-4

Diagnosis of Microcytic Anemia

Tests	Iron Deficiency	Inflammation	Thalassemia	Sideroblastic Anemia
Smear	Micro/hypo	Normal micro/hypo	Micro/hypo with targeting	Variable
Serum iron (µg/dL)	<30	<50	Normal to high	Normal to high
TIBC (µg/dL)	>360	<300	Normal	Normal
Percent saturation	<10	10–20	30–80	30–80
Ferritin (µg/L)	<15	30–200	50–300	50–300
Hemoglobin pattern on electrophoresis	Normal	Normal	Abnormal with β thalassemia; can be normal with α thalassemia	Normal

Abbreviation: TIBC, total iron-binding capacity.

The second condition is the anemia of inflammation (AI; also referred to as the anemia of chronic disease) with inadequate iron supply to the erythroid marrow. The distinction between true iron-deficiency anemia and AI is among the most common diagnostic problems encountered by clinicians (see below). Usually, AI is normocytic and normochromic. The iron values usually make the differential diagnosis clear, as the ferritin level is normal or increased and the percent transferrin saturation and TIBC are typically below normal.

Finally, the myelodysplastic syndromes represent the third and least common condition. Occasionally, patients with myelodysplasia have impaired hemoglobin synthesis with mitochondrial dysfunction, resulting in impaired iron incorporation into heme. The iron values again reveal normal stores and more than an adequate supply to the marrow, despite the microcytosis and hypochromia.

TREATMENT

TREATMENT

Iron-Deficiency Anemia

The severity and cause of iron-deficiency anemia will determine the appropriate approach to treatment. As an example, symptomatic elderly patients with severe iron-deficiency anemia and cardiovascular instability may require red cell transfusions. Younger individuals who have compensated for their anemia can be treated more conservatively with iron replacement. The foremost issue for the latter patient is the precise identification of the cause of the iron deficiency.

For the majority of cases of iron deficiency (pregnant women, growing children and adolescents, patients with infrequent episodes of bleeding, and those with inadequate dietary intake of iron), oral iron therapy will suffice. For patients with unusual blood loss or malabsorption, specific diagnostic tests and appropriate therapy take priority. Once the diagnosis of iron-deficiency anemia and its cause is made, there are three major therapeutic approaches.

RED CELL TRANSFUSION

Transfusion therapy is reserved for individuals who have symptoms of anemia, cardiovascular instability, and continued and excessive blood loss from whatever source and who require immediate intervention. The management of these patients is less related to the iron deficiency than it is to the consequences of the severe anemia. Not only do transfusions correct the anemia acutely, but the transfused red cells provide a source of iron for reutilization, assuming they are not lost through continued bleeding. Transfusion therapy will stabilize the patient while other options are reviewed.

ORAL IRON THERAPY

In the asymptomatic patient with established iron-deficiency anemia and an intact gastrointestinal tract, treatment with oral iron is usually adequate. Multiple preparations are available, ranging from simple iron salts to complex iron compounds designed for sustained release throughout the small intestine (Table 93-5). Although the various preparations contain different amounts of iron, they are generally all absorbed well and are effective in treatment. Some come with other compounds designed to enhance iron absorption, such as ascorbic acid. It is not clear whether the benefits of such compounds justify their costs. Typically, for iron replacement therapy, up to 200 mg of elemental iron per day is given, usually as three or four iron tablets (each containing 50–65 mg elemental iron) given over the course of the day. Ideally, oral iron preparations should be taken on an empty stomach, since food may inhibit iron absorption. Some patients with gastric disease or prior gastric surgery require special treatment with iron solutions because the retention capacity of the stomach may be reduced. The retention capacity is necessary for dissolving the shell of the iron tablet before the release of iron. A dose of 200 mg of elemental iron per day should result in the absorption of iron up to 50 mg/d. This supports a red cell production level of two to three times normal in an individual with a normally functioning marrow and appropriate erythropoietin stimulus. However, as the hemoglobin level rises, erythropoietin stimulation decreases, and the amount of iron absorbed is reduced. The goal of therapy in individuals with iron-deficiency anemia is not only to repair the anemia, but also to provide stores of at least 0.5–1 g of iron. Sustained treatment for a period of 6–12 months after correction of the anemia will be necessary to achieve this.

Of the complications of oral iron therapy, gastrointestinal distress is the most prominent and is seen in at least 15–20% of patients. Abdominal pain, nausea, vomiting, or constipation may lead to noncompliance. Although small doses of iron or iron preparations with delayed release may help somewhat, the gastrointestinal side effects are a major impediment to the effective treatment of a number of patients.

The response to iron therapy varies, depending on the erythropoietin stimulus and the rate of absorption. Typically, the reticulocyte count should begin to increase within 4–7 days after initiation of therapy and peak at 1–1½ weeks. The absence of a response may be due to poor absorption, noncompliance (which is common), or a confounding diagnosis. A useful test in the clinic to determine the patient's ability to absorb iron is the *iron tolerance test*. Two iron tablets are given to the patient on an empty stomach, and the serum iron is measured serially over the subsequent 2–3 h. Normal absorption will result in an increase in the serum iron of at least 100 µg/dL. If iron deficiency persists despite adequate treatment, it may be necessary to switch to parenteral iron therapy.

PARENTERAL IRON THERAPY

Intravenous iron can be given to patients who are unable to tolerate oral iron; whose needs are relatively acute; or who need iron on an ongoing basis, usually due to persistent gastrointestinal or menstrual blood loss. Parenteral iron use has been increasing rapidly in the last several years with the recognition that recombinant erythropoietin (EPO) therapy induces a large demand for iron—a demand that frequently cannot be met through the physiologic release of iron from RE sources or oral iron absorption. The safety of parenteral iron has been a concern largely driven by the high adverse reaction rate to high-molecular-weight iron dextran. The newer iron complexes that are available, such as ferumoxytol (Feraheme), sodium ferric gluconate (Ferrlecit), iron sucrose (Venofer), low-molecular-weight (LMW) iron dextran (InFed), and ferric carboxymaltose (Injectafer), have much lower rates of adverse effects. Ferumoxytol delivers 510 mg of iron per injection; ferric gluconate 125 mg per injection; LMW iron dextran up to 1500 mg; ferric carboxymaltose 750 mg per injection and iron sucrose 200 mg per injection.

Parenteral iron is used in two ways: one is to administer the total dose of iron required to correct the hemoglobin deficit and provide the patient with at least 500 mg of iron stores; the second is to give repeated small doses of parenteral iron over a protracted period. The latter approach is common in dialysis centers, where it is not unusual for 100 mg of elemental iron to be given weekly for 10 weeks to augment the response to recombinant EPO therapy. The amount of iron needed by an individual patient is calculated by the following formula:

$$\text{Body weight (kg)} \times 2.3 \times (15 - \text{patient's hemoglobin, g/dL}) + 500 \text{ or } 1000 \text{ mg (for stores)}.$$

In administering any intravenous iron preparation, anaphylaxis is a concern. Anaphylaxis is much rarer with the newer preparations. The factors that have correlated with an anaphylactic-like reaction include a history of multiple allergies or a prior allergic reaction to an iron preparation. Generalized symptoms appearing several days after the infusion of a large dose of iron can include arthralgias, skin rash, and low-grade fever. These may be dose-related, but they do not preclude the further use of parenteral iron in the patient. To

date, patients with sensitivity to one iron preparation have been safely treated with other parenteral iron preparations. If a large dose of LMW iron dextran is to be given (>100 mg), the iron preparation should be diluted in 5% dextrose in water or 0.9% NaCl solution. The iron solution can then be infused over a 60- to 90-min period (for larger doses) or at a rate convenient for the attending nurse or physician. Although a test dose (25 mg) of parenteral LMW iron dextran is recommended, in reality a slow infusion of a larger dose of parenteral iron solution will afford the same kind of early warning as a separately injected test dose. Early in the infusion of iron, if chest pain, wheezing, a fall in blood pressure, or other systemic symptoms occur, the infusion of iron should be stopped immediately.

TABLE 93-5

Oral Iron Preparations

Generic Name	Tablet (Iron Content), mg	Elixir (Iron Content), mg in 5 mL
Ferrous sulfate	325 (65)	300 (60)
	195 (39)	90 (18)
Extended release	525 (105)	
Ferrous fumarate	325 (107)	
	195 (64)	100 (33)
Ferrous gluconate	325 (39)	300 (35)
Polysaccharide iron	150 (150)	100 (100)
	50 (50)	

OTHER HYPOPROLIFERATIVE ANEMIAS

In addition to mild to moderate iron-deficiency anemia, the hypoproliferative anemias can be divided into four categories: (1) chronic inflammation, (2) renal disease, (3) endocrine and nutritional deficiencies (hypometabolic states), and (4) marrow damage (Chap. 98). With chronic inflammation, renal disease, or hypometabolism, endogenous EPO production is inadequate for the degree of anemia observed. For the anemia of chronic inflammation, the erythroid marrow also responds inadequately to stimulation, due in part to defective *iron reutilization*. As a result of the lack of adequate EPO stimulation, an examination of the peripheral blood smear will disclose only an occasional polychromatophilic (“shift”) reticulocyte. In cases of

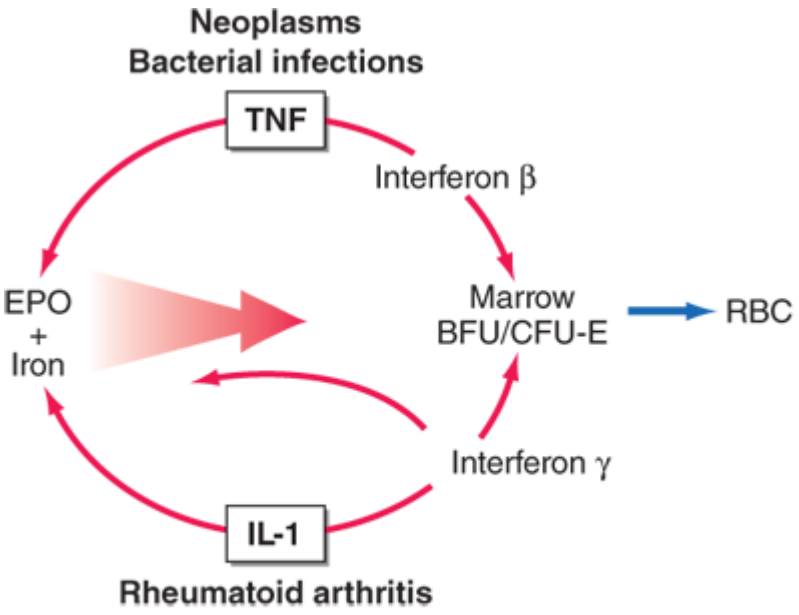
iron deficiency or marrow damage, appropriate elevations in endogenous EPO levels are typically found, and shift reticulocytes will be present on the blood smear.

ANEMIA OF ACUTE AND CHRONIC INFLAMMATION/INFECTION (AI)

AI—which encompasses inflammation, infection, tissue injury, and conditions (such as cancer) associated with the release of proinflammatory cytokines—is one of the most common forms of anemia seen clinically. It is the most important anemia in the differential diagnosis of iron deficiency because many of the features of the anemia are brought about by inadequate iron delivery to the marrow, despite the presence of normal or increased iron stores. This is reflected by a low serum iron, increased red cell protoporphyrin, a hypoproliferative marrow, transferrin saturation in the range of 15–20%, and a normal or increased serum ferritin. The serum ferritin values are often the most distinguishing features between true iron-deficiency anemia and the iron-restricted erythropoiesis associated with inflammation. Typically, serum ferritin values increase threefold over basal levels in the face of inflammation. These changes are due to the effects of inflammatory cytokines and hepcidin, the key iron regulatory hormone, acting at several levels of erythropoiesis (Fig. 93-4).

FIGURE 93-4

Suppression of erythropoiesis by inflammatory cytokines. Through the release of tumor necrosis factor (TNF) and interferon β (IFN-β), neoplasms and bacterial infections suppress erythropoietin (EPO) production and the proliferation of erythroid progenitors (erythroid burst-forming units and erythroid colony-forming units [BFU/CFU-E]). The mediators in patients with vasculitis and rheumatoid arthritis include interleukin 1 (IL-1) and IFN-γ. The *red arrows* indicate sites of inflammatory cytokine inhibitory effects. RBC, red blood cell.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Interleukin 1 (IL-1) directly decreases EPO production in response to anemia. IL-1, acting through accessory cell release of interferon γ (IFN-γ), suppresses the response of the erythroid marrow to EPO—an effect that can be overcome by EPO administration in vitro and in vivo. In addition, tumor necrosis factor (TNF), acting

through the release of IFN-β by marrow stromal cells, also suppresses the response to EPO. Hepcidin, made by the liver, is increased in inflammation via an IL-6 mediated pathway, and acts to suppress iron absorption and iron release from storage sites. The overall result is a chronic hypoproliferative anemia with classic changes in iron metabolism. The anemia is further compounded by a mild to moderate shortening in red cell survival.

With chronic inflammation, the primary disease will determine the severity and characteristics of the anemia. For example, many patients with cancer also have anemia that is typically normocytic and normochromic. In contrast, patients with long-standing active rheumatoid arthritis or chronic infections such as tuberculosis will have a microcytic, hypochromic anemia. In both cases, the bone marrow is hypoproliferative, but the differences in red cell indices reflect differences in the availability of iron for hemoglobin synthesis. Occasionally, conditions associated with chronic inflammation are also associated with chronic blood loss. Under these circumstances, the measurement of soluble transferrin protein may be necessary to rule out absolute iron deficiency. However, the administration of iron in this case will correct the iron deficiency component of the anemia and leave the inflammatory component unaffected.

The anemia associated with acute infection or inflammation is typically mild but becomes more pronounced over time. Acute infection can produce a decrease in hemoglobin levels of 2–3 g/dL within 1 or 2 days; this is largely related to the hemolysis of red cells near the end of their natural life span. The fever and cytokines released exert a selective pressure against cells with more limited capacity to maintain the red cell membrane. In most individuals, the mild anemia is reasonably well tolerated, and symptoms, if present, are associated with the underlying disease. Occasionally, in patients with preexisting cardiac disease, moderate anemia (hemoglobin 10–11 g/dL) may be associated with angina, exercise intolerance, and shortness of breath. The erythropoietic profile that distinguishes the anemia of inflammation from the other causes of hypoproliferative anemias is shown in Table 93-6.

TABLE 93-6

Diagnosis of Hypoproliferative Anemias

Tests	Iron Deficiency	Inflammation	Renal Disease	Hypometabolic States
Anemia	Mild to severe	Mild	Mild to severe	Mild
MCV (fL)	60–90	80–90	90	90
Morphology	Normo-microcytic	Normocytic	Normocytic	Normocytic
SI (µg/dL)	<30	<50	Normal	Normal
TIBC (µg/dL)	>360	<300	Normal	Normal
Saturation (%)	<10	10–20	Normal	Normal
Serum ferritin (µg/L)	<15	30–200	115–150	Normal
Iron stores	0	2–4+	1–4+	Normal

Abbreviations: MCV, mean corpuscular volume; SI, serum iron; TIBC, total iron-binding capacity.

ANEMIA OF CHRONIC KIDNEY DISEASE (CKD)

Progressive CKD is usually associated with a moderate to severe hypoproliferative anemia; the level of the anemia correlates with the stage of CKD. Red cells are typically normocytic and normochromic, and reticulocytes are decreased. The anemia is primarily due to a failure of EPO production by the diseased kidney and a reduction in red cell survival. In certain forms of acute renal failure, the correlation between the anemia and renal function is weaker. Patients with the hemolytic-uremic syndrome increase erythropoiesis in response to the hemolysis, despite renal failure. Polycystic kidney disease also shows a smaller degree of EPO deficiency for a given level of renal failure. By contrast, patients with diabetes or myeloma have more severe EPO deficiency for a given level of renal failure.

Assessment of iron status provides information to distinguish the anemia of CKD from the other forms of hypoproliferative anemia (Table 93-6) and to guide management. Patients with the anemia of CKD usually present with normal serum iron, TIBC, and ferritin levels. However, those maintained on chronic hemodialysis may develop iron deficiency from blood loss through the dialysis procedure. Iron must be replenished in these patients to ensure an adequate response to EPO therapy (see below).

ANEMIA IN HYPOMETABOLIC STATES

Patients who are starving, particularly for protein, and those with a variety of endocrine disorders that produce lower metabolic rates, may develop a mild to moderate hypoproliferative anemia. The release of EPO from the kidney is sensitive to the need for O_2, not just O_2 levels. Thus, EPO production is triggered at lower levels of blood O_2 content in disease states (such as hypothyroidism and starvation) where metabolic activity, and thus O_2 demand, is decreased.

Endocrine Deficiency States

The difference in the levels of hemoglobin between men and women is related to the effects of androgen and estrogen on erythropoiesis. Testosterone and anabolic steroids augment erythropoiesis; castration and estrogen administration to males decrease erythropoiesis. Patients who are hypothyroid or have deficits in pituitary hormones also may develop a mild anemia. Pathogenesis may be complicated by other nutritional deficiencies because iron and folic acid absorption can be affected by these disorders. Usually, correction of the hormone deficiency reverses the anemia.

Anemia may be more severe in Addison's disease, depending on the level of thyroid and androgen hormone dysfunction; however, anemia may be masked by decreases in plasma volume. Once such patients are given cortisol and volume replacement, the hemoglobin level may fall rapidly. Mild anemia complicating hyperparathyroidism may be due to decreased EPO production as a consequence of the renal effects of hypercalcemia or to impaired proliferation of erythroid progenitors.

Protein Starvation

Decreased dietary intake of protein may lead to mild to moderate hypoproliferative anemia; this form of anemia may be prevalent in the elderly. The anemia can be more severe in patients with a greater degree of starvation. In marasmus, where patients are both protein- and calorie-deficient, the release of EPO is impaired in proportion to the reduction in metabolic rate; however, the degree of anemia may be masked by volume depletion and becomes apparent after refeeding. Deficiencies in other nutrients (iron, folate) may also complicate the clinical picture but may not be apparent at diagnosis. Changes in the erythrocyte indices on refeeding should prompt evaluation of iron, folate, and B_{12} status.

Anemia in Liver Disease

A mild hypoproliferative anemia may develop in patients with chronic liver disease from nearly any cause. The peripheral blood smear may show spur cells and stomatocytes from the accumulation of excess cholesterol in the membrane from a deficiency of lecithin-cholesterol acyltransferase. Red cell survival is shortened, and the production of EPO is inadequate to compensate. In alcoholic liver disease, nutritional deficiencies are common and complicate the management. Folate deficiency from inadequate intake, as well as iron deficiency from blood loss and inadequate intake, can alter the red cell indices.

ANEMIA IN AGING

Anemia is common in people over age 65 years. It has been estimated to affect about 11% of community living older adults and up to 40% of nursing home residents. In at least one-third of these anemic people, a cause for the anemia is not found. Patients with the unexplained anemia of aging do not have nutrient deficiency or renal dysfunction and while older people can have an increase in systemic inflammatory cytokines (the inflammation of aging), the levels are not high enough to mimic the anemia of chronic inflammation. If hepcidin levels are elevated at all, they are minimally so.

Investigations into the cause(s) of this form of anemia have noted that erythropoietin levels are generally in the normal range, that is, they are inappropriately low for the hemoglobin level. In general, in older people who maintain a normal hemoglobin level, erythropoietin levels increase with age. This compensatory increase to maintain normal oxygen delivery seems to be due to a relative resistance to erythropoietin stimulation; studies of red cell life span in older people have not noted a decrease in red cell survival. More data on mechanism are needed.

The importance of this unexplained anemia of aging is that low hemoglobin levels are associated with increases in falls, hospitalizations, development of frailty, and mortality. It is not clear whether reversing the anemia would influence these increased risks.

TREATMENT

TREATMENT

Hypoproliferative Anemias

Many patients with hypoproliferative anemias experience recovery of normal hemoglobin levels when the underlying disease is appropriately treated. For those in whom such reversals are not possible—such as patients with end-stage kidney disease, cancer, and chronic inflammatory diseases—symptomatic anemia requires treatment. The two major forms of treatment are transfusions and EPO.

TRANSFUSIONS

Thresholds for transfusion should be determined based on the patient's symptoms. In general, patients without serious underlying cardiovascular or pulmonary disease can tolerate hemoglobin levels above 7–8 g/dL and do not require intervention until the hemoglobin falls below that level. Patients with more physiologic compromise may need to have their hemoglobin levels kept above 11 g/dL. Usually, a unit of packed red cells increases the hemoglobin level by 1 g/dL. Transfusions are associated with certain infectious risks (Chap. 109), and chronic transfusions can produce iron overload. Importantly, the liberal use of blood has been associated with increased morbidity and mortality, particularly in the intensive care setting. Therefore, in the absence of documented tissue hypoxia, a conservative approach to the use of red cell transfusions is preferable.

ERYTHROPOIETIN

EPO is particularly useful in anemias in which endogenous EPO levels are inappropriately low, such as CKD or AI. Iron status must be evaluated and iron replaced to obtain optimal effects from EPO. In patients with CKD, the usual dose of EPO is 50–150 U/kg three times a week intravenously. Hemoglobin levels of 10–12 g/dL are usually reached within 4–6 weeks if iron levels are adequate; 90% of these patients respond. Once a target hemoglobin level is achieved, the EPO dose can be decreased. A decrease in hemoglobin level occurring in the face of EPO therapy usually signifies the development of an infection or iron depletion. Aluminum toxicity and hyperparathyroidism can also compromise the response to EPO. When an infection intervenes, it is best to interrupt the EPO therapy and rely on transfusions to correct the anemia until the infection is adequately treated. The dose of EPO needed to correct chemotherapy-induced anemia in patients with cancer is higher, up to 300 U/kg three times a week, and only ~60% of patients respond. Because of evidence that there is an increased risk of thromboembolic complications and tumor progression with EPO administration, the risks and benefits of using EPO in such patients must be weighed carefully, and the target hemoglobin should be that necessary to avoid transfusions.

Longer-acting preparations of EPO can reduce the frequency of injections. Darbepoetin alfa, a molecularly modified EPO with additional carbohydrate, has a half-life in the circulation that is three to four times longer than recombinant human EPO, permitting weekly or every other week dosing.

Orally bioavailable EPO mimetics that act to increase the biological half-life of active hypoxia-induced factor (HIF) are demonstrating activity to increase hemoglobin levels in patients with chronic renal disease and other settings.

FURTHER READING

Andrews N: Forging a field. The golden age of iron biology. Blood 112:219, 2008. [PubMed: 18606887]

Auerbach M, Adamson J: How we diagnose and treat iron deficiency anemia. Am J Hematol 91:31, 2016. [PubMed: 26408108]

Drüeke T, Parfrey P: Summary of the KDIGO guideline on anemia and comment: Reading between the (guide)line(s). Kidney Int 82:952, 2012. [PubMed: 22854645]

Kautz L et al: Identification of erythroferrone as an erythroid regulator of iron metabolism. Nature Genet 46:678, 2014. [PubMed: 24880340]

Krayenbuehl P-A et al: Intravenous iron for the treatment of fatigue in non-anemic, premenopausal woman with low serum ferritin concentration. Blood 118:3222, 2011. [PubMed: 21705493]

Punnonen K, Kerttu I, Rajamaki A: Serum transferrin receptor and its ratio to serum ferritin in the diagnosis of iron deficiency. Blood 89:1052; 1997. [PubMed: 9028338]

Weiss G, Goodnough L: Medical Progress: Anemia of chronic disease. N Engl J Med 352:1011, 2005.
[PubMed: 15758012]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 94: Disorders of Hemoglobin

Edward J. Benz, Jr.

INTRODUCTION

Hemoglobin is critical for normal oxygen delivery to tissues; it is also present in erythrocytes in such high concentrations that it can alter red cell shape, deformability, and viscosity. Hemoglobinopathies are disorders affecting the structure, function, or production of hemoglobin. These conditions are usually inherited and range in severity from asymptomatic laboratory abnormalities to death in utero. Different forms may present as ineffective erythropoiesis, hemolytic anemia, erythrocytosis, cyanosis, or vasoocclusive stigmata.

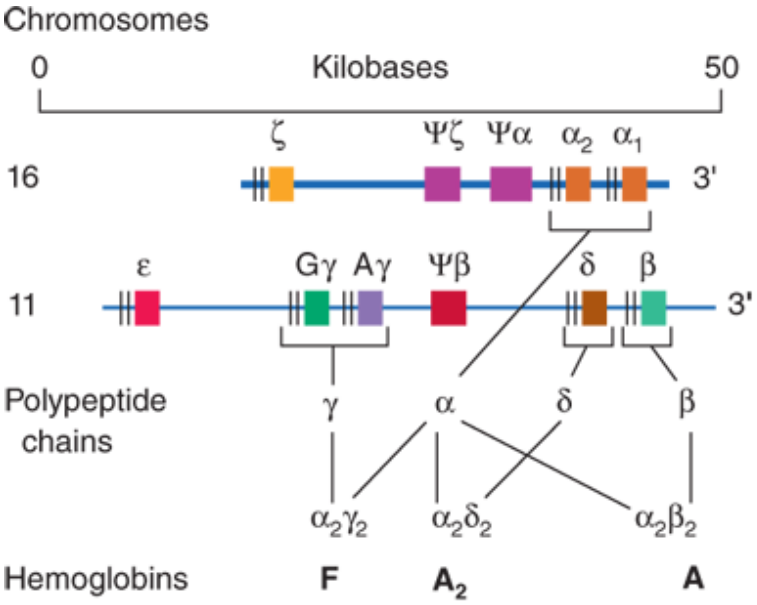
PROPERTIES OF THE HUMAN HEMOGLOBINS

HEMOGLOBIN STRUCTURE

Different hemoglobins are produced during embryonic, fetal, and adult life **(Fig. 94-1)**. Each consists of a tetramer of globin polypeptide chains: a pair of α-like chains 141 amino acids long and a pair of β-like chains 146 amino acids long. The major adult hemoglobin, HbA, has the structure $\alpha_2\beta_2$. HbF ($\alpha_2\gamma_2$) predominates during most of gestation, and HbA_2 ($\alpha_2\delta_2$) is minor adult hemoglobin. Embryonic hemoglobins need not be considered here.

FIGURE 94-1

The globin genes. The α-like genes (α, ζ) are encoded on chromosome 16; the β-like genes (β, γ, δ, ε) are encoded on chromosome 11. The ζ and ε genes encode embryonic globins.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Each globin chain enfolds a single heme moiety, consisting of a protoporphyrin IX ring complexed with a single iron atom in the ferrous state (Fe^{2+}). Each heme moiety can bind a single oxygen molecule; a molecule of hemoglobin can transport up to four oxygen molecules.

Each globin chain has a highly helical *secondary structure*. Their globular *tertiary structures* cause the exterior surfaces to be rich in polar (hydrophilic) amino acids that enhance solubility, and the interior to be lined with nonpolar groups, forming a hydrophobic pocket into which heme is inserted. The tetrameric *quaternary structure* of HbA contains two αβ dimers. Numerous tight interactions (i.e., $\alpha_1\beta_1$ contacts) hold the α and β chains together. The complete tetramer is held together by interfaces (i.e., $\alpha_1\beta_2$ contacts) between the α-like chain of one dimer and the non-α chain of the other dimer.

The hemoglobin tetramer is highly soluble, but individual globin chains are insoluble. Unpaired globin precipitates, forming inclusions that damage the erythroblast and can trigger apoptosis. Normal globin chain synthesis is balanced so that each newly synthesized α or non-α globin chain will have an available partner with which to pair.

Solubility and reversible oxygen binding are the key properties deranged in hemoglobinopathies. Both depend most on the hydrophilic surface amino acids, the hydrophobic amino acids lining the heme pocket, a key histidine in the F helix, and the amino acids forming the $\alpha_1\beta_1$ and $\alpha_1\beta_2$ contact points. Mutations in these strategic regions tend to be the ones that alter oxygen affinity or solubility.

FUNCTION OF HEMOGLOBIN

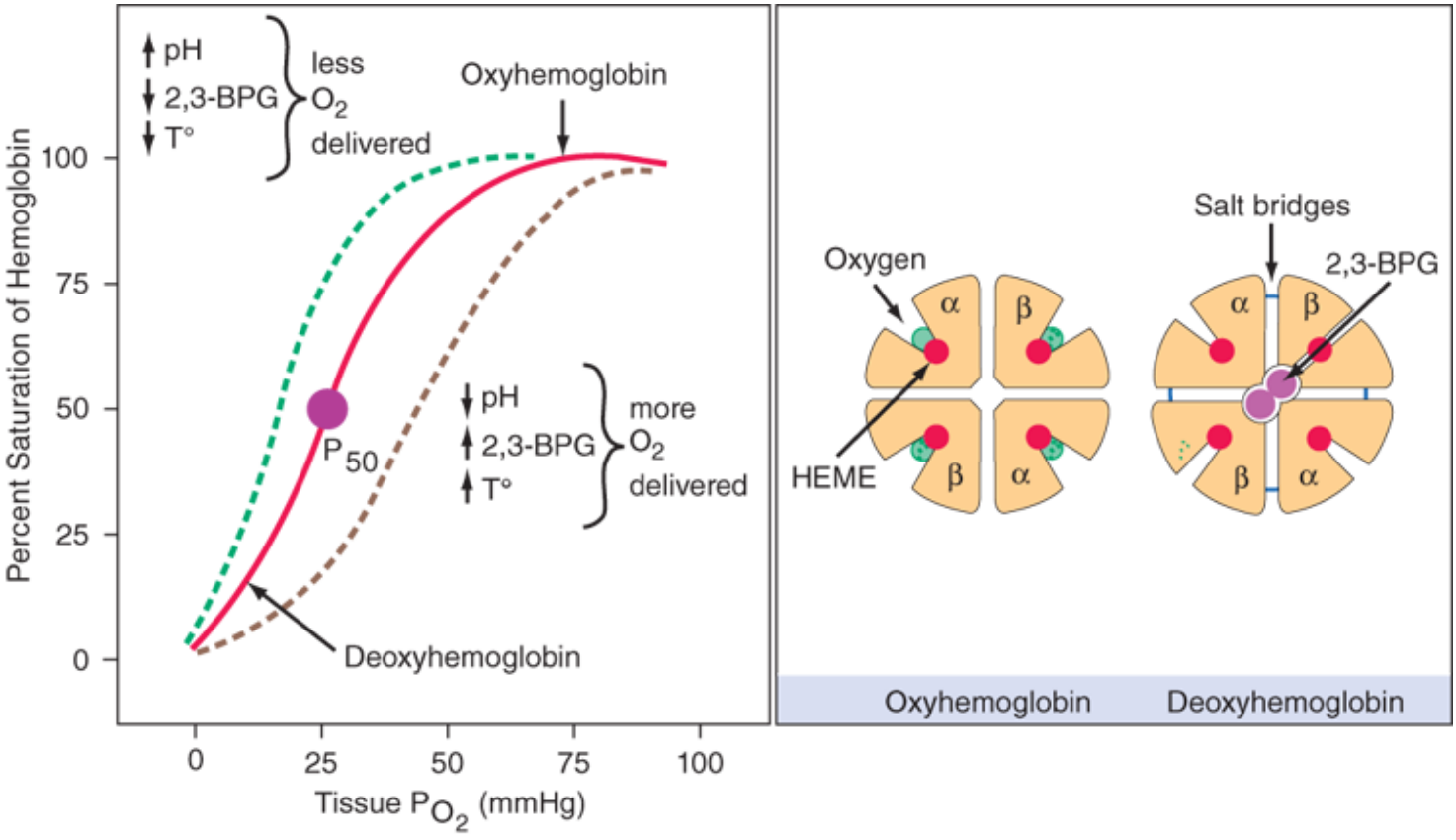
To support oxygen transport, hemoglobin must bind O_2 efficiently at the partial pressure of oxygen (Po_2) of the alveolus, retain it in the circulation, and release it to tissues at the Po_2 of tissue capillary beds. Oxygen acquisition and delivery over a relatively narrow range of oxygen tensions depend on a property inherent in the

tetrameric arrangement of heme and globin subunits within the hemoglobin molecule called *cooperativity* or *heme-heme interaction*.

At low oxygen tensions, the hemoglobin tetramer is fully deoxygenated **(Fig. 94-2)**. Oxygen binding begins slowly as O_2 tension rises. However, as soon as some oxygen has been bound by the tetramer, an abrupt increase occurs in the slope of the curve. Thus, hemoglobin molecules that have bound some oxygen develop a higher oxygen affinity, greatly accelerating their ability to combine with more oxygen. This S-shaped oxygen equilibrium curve (Fig. 94-2), along which substantial amounts of oxygen loading *and unloading* can occur over a narrow range of oxygen tensions, is physiologically more useful than the high-affinity hyperbolic curve of individual monomers.

FIGURE 94-2

Hemoglobin-oxygen dissociation curve. The hemoglobin tetramer can bind up to four molecules of oxygen in the iron-containing sites of the heme molecules. As oxygen is bound, 2,3-bisphosphoglycerate (2,3-BPG) and carbon dioxide (CO_2) are expelled. Salt bridges are broken, and each of the globin molecules changes its conformation to facilitate oxygen binding. Oxygen release to the tissues is the reverse process, with salt bridges being formed and 2,3-BPG and CO_2 bound. Deoxyhemoglobin does not bind oxygen efficiently until the cell returns to conditions of higher pH, the most important modulator of O_2 affinity (Bohr effect). When acid is produced in the tissues, the dissociation curve shifts to the right, facilitating oxygen release and CO_2 binding. Alkalosis has the opposite effect, reducing oxygen delivery.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Oxygen affinity is modulated by several factors. The Bohr effect is the ability of hemoglobin to deliver more oxygen to tissues at low pH. It arises from the stabilizing action of protons on deoxyhemoglobin, which binds protons more readily than oxyhemoglobin because the latter is a weaker acid (Fig. 94-2). Thus, hemoglobin has a lower oxygen affinity at low pH. The major small molecule that alters oxygen affinity in humans is 2,3-bisphosphoglycerate (2,3-BPG; formerly 2,3-DPG), which lowers oxygen affinity when bound to hemoglobin. HbA has a reasonably high affinity for 2,3-BPG. HbF does not bind 2,3-BPG, so it tends to have a higher oxygen affinity in vivo. Hemoglobin also binds nitric oxide reversibly; this interaction influences vascular tone, but its clinical relevance remains incompletely understood. Normal oxygen transport thus depends on the tetrameric structure of the proteins, the proper arrangement of hydrophilic and hydrophobic amino acids, and interaction with protons or 2,3-BPG.

DEVELOPMENTAL BIOLOGY OF HUMAN HEMOGLOBINS

Red cells, first appearing at about 6 weeks after conception, contain the embryonic hemoglobins Hb Portland ($\zeta_2\gamma_2$), Hb Gower I ($\zeta_2\varepsilon_2$), and Hb Gower II ($\alpha_2\varepsilon_2$). At 10–11 weeks, fetal hemoglobin (HbF; $\alpha_2\gamma_2$) becomes predominant. The switch to nearly exclusive synthesis of adult hemoglobin (HbA; $\alpha_2\beta_2$) occurs at about 38 weeks (Fig. 94-1). Fetuses and newborns therefore require α-globin but not β-globin for normal gestation. A major advance in understanding the HbF to HbA transition has been the demonstration that transcription factors such as Bcl11a play a pivotal role in its regulation, and that access to these factors depend on complex chromatin changes that render accessible binding sites for these factors. These “looping out” phenomena are in turn dependent on multiple protein factors.

Small amounts of HbF are produced during postnatal life. A few red cell clones called *F cells* are progeny of a small pool of immature committed erythroid precursors (BFU-e) that retain the ability to produce HbF. Profound erythroid stresses, such as severe hemolytic anemias, bone marrow transplantation, or cancer chemotherapy, cause more of the F-potent BFU-e to be recruited. HbF levels thus tend to rise in some patients with sickle cell anemia or thalassemia. This phenomenon probably explains the ability of hydroxyurea to increase levels of HbF in adults. Agents such as butyrate and histone deacetylase inhibitors can also activate fetal globin genes partially after birth.

GENETICS AND BIOSYNTHESIS OF HUMAN HEMOGLOBIN



The human hemoglobins are encoded in two tightly linked gene clusters; the α-like globin genes are clustered on chromosome 16 and the β-like genes on chromosome 11 (Fig. 94-1). The α-like cluster consists of two α-globin genes and a single copy of the ζ gene. The non-α gene cluster consists of a single ε gene, the Gγ and Aγ fetal globin genes, and the adult δ and β genes.

Important regulatory sequences flank each gene. Immediately upstream are typical promoter elements needed for the assembly of the transcription initiation complex. Sequences in the 5′ flanking region of the γ and the β genes appear to be crucial for the correct developmental regulation of these genes, whereas elements that function like classic enhancers and silencers are in the 3′ flanking regions. The locus control region (LCR)

elements located far upstream appear to control the overall level of expression of each cluster. These elements achieve their regulatory effects by interacting with *trans*-acting transcription factors. Some of these factors are ubiquitous (e.g., Sp1 and YY1), while others are more or less limited to erythroid cells or hematopoietic cells (e.g., GATA-1, NFE-2, and EKLF). The LCR controlling the α-globin gene cluster is modulated by a SWI/SNF-like protein called *ATRX*; this protein appears to influence chromatin remodeling and DNA methylation. The association of α thalassemia with mental retardation and myelodysplasia in some families appears to be related to mutations in the ATRX pathway. This pathway also modulates genes specifically expressed during erythropoiesis, such as those that encode the enzymes for heme biosynthesis. Normal red blood cell (RBC) differentiation requires the coordinated expression of the globin genes with the genes responsible for heme and iron metabolism. RBC precursors contain a protein, α-hemoglobin-stabilizing protein (AHSP), that enhances the folding and solubility of α globin, which is otherwise easily denatured, leading to insoluble precipitates. These precipitates play an important role in the thalassemia syndromes and certain unstable hemoglobin disorders.

CLASSIFICATION OF HEMOGLOBINOPATHIES

There are five major classes of hemoglobinopathies (**Table 94-1**). *Structural hemoglobinopathies* occur when mutations alter the amino acid sequence of a globin chain, altering the physiologic properties of the variant hemoglobins and producing the characteristic clinical abnormalities. The most clinically relevant variant hemoglobins polymerize abnormally, as in sickle cell anemia, or exhibit altered solubility or oxygen-binding affinity. *Thalassemia syndromes* arise from mutations that impair production or translation of globin mRNA, leading to deficient globin chain biosynthesis. Clinical abnormalities are attributable to the inadequate supply of hemoglobin and the imbalances in the production of individual globin chains, leading to premature destruction of erythroblasts and RBC. *Thalassemic hemoglobin variants* combine features of thalassemia (i.e., abnormal globin biosynthesis) and of structural hemoglobinopathies (i.e., an abnormal amino acid sequence). *Hereditary persistence of fetal hemoglobin* (HPFH) is characterized by synthesis of high levels of fetal hemoglobin in adult life. *Acquired hemoglobinopathies* include modifications of the hemoglobin molecule by toxins (e.g., acquired methemoglobinemia) and clonal abnormalities of hemoglobin synthesis (e.g., high levels of HbF production in preleukemia and α thalassemia in myeloproliferative disorders).

TABLE 94-1

Classification of Hemoglobinopathies

I. Structural hemoglobinopathies—hemoglobins with altered amino acid sequences that result in deranged function or altered physical or chemical properties
  A. Abnormal hemoglobin polymerization—HbS, hemoglobin sickling
  B. Altered O_2 affinity
    1. High affinity—polycythemia
    2. Low affinity—cyanosis, pseudoanemia
  C. Hemoglobins that oxidize readily
    1. Unstable hemoglobins—hemolytic anemia, jaundice
    2. M hemoglobins—methemoglobinemia, cyanosis
II. Thalassemias—defective biosynthesis of globin chains
  A. α thalassemias
  B. β thalassemias
  C. δβ, γδβ, αβ thalassemias
III. Thalassemic hemoglobin variants—structurally abnormal Hb associated with coinherited thalassemic phenotype
  A. HbE
  B. Hb Constant Spring
  C. Hb Lepore
IV. Hereditary persistence of fetal hemoglobin—persistence of high levels of HbF into adult life
V. Acquired hemoglobinopathies
  A. Methemoglobin due to toxic exposures
  B. Sulfhemoglobin due to toxic exposures
  C. Carboxyhemoglobin
  D. HbH in erythroleukemia
  E. Elevated HbF in states of erythroid stress and bone marrow dysplasia

EPIDEMIOLOGY



Hemoglobinopathies are especially common in areas in which malaria is endemic. This clustering of hemoglobinopathies is assumed to reflect a selective survival advantage for the abnormal RBC, which presumably provides a less hospitable environment during the obligate RBC stages of the parasitic life cycle. Very young children with α thalassemia are *more* susceptible to infection with the nonlethal *Plasmodium vivax*. Thalassemia might then favor a natural protection against infection with the more lethal *Plasmodium falciparum*.

Thalassemias are the most common genetic disorders in the world, affecting nearly 200 million people worldwide. About 15% of African Americans are silent carriers for α thalassemia; α thalassemia trait (minor) occurs in 3% of African American and in 1–15% of persons of Mediterranean origin. β thalassemia has a 10–15% incidence in individuals from the Mediterranean and Southeast Asia and 0.8% in African Americans. The number of severe cases of thalassemia in the United States is about 1000. Sickle cell disease is the most common structural hemoglobinopathy, occurring in heterozygous form in ~8% of African Americans and in homozygous form in 1 in 400. Between 2 and 3% of African Americans carry a hemoglobin C allele. The chronic nature of hemoglobinopathies and their requirement for complex, resource intensive care pose increasing major public health challenges in regions with emerging economies, as children increasingly survive the traditional causes of childhood mortality.

INHERITANCE AND ONTOGENY

Hemoglobinopathies are autosomal codominant traits—thus, compound heterozygotes who inherit a different abnormal mutant allele from each parent exhibit composite features of each. For example, patients inheriting sickle β thalassemia exhibit features of β thalassemia and sickle cell anemia. The α chain is present in HbA, HbA_2, and HbF; α-chain mutations thus cause abnormalities in all three. The α-globin hemoglobinopathies are symptomatic in utero and after birth because normal function of the α-globin gene is required throughout gestation and adult life. In contrast, infants with β-globin hemoglobinopathies tend to be asymptomatic until 3–9 months of age, when HbA has largely replaced HbF. Prevention or partial reversion of the switch should thus be an effective therapeutic strategy for β-chain hemoglobinopathies.

DETECTION AND CHARACTERIZATION OF HEMOGLOBINOPATHIES–GENERAL METHODS

While electrophoretic techniques are still used for hemoglobin analysis in some settings, high-performance liquid chromatography (HPLC) has largely supplanted electrophoresis in most reference laboratories. Some important variants can be missed by these methods because they co-migrate with normal hemoglobins. Complete characterization, including amino acid sequencing or genotyping by direct DNA analysis, is readily available from several reference laboratories and should be requested if clinical suspicion is high, or when HPLC fails to yield definitive answers.

Quantitation of the hemoglobin profile is often desirable. HbA_2 is frequently elevated in β thalassemia trait and depressed in iron deficiency. HbF is elevated in HPFH, some β thalassemia syndromes, and occasional periods of erythroid stress or marrow dysplasia. For characterization of sickle cell trait, sickle thalassemia syndromes, or HbSC disease, and for monitoring the progress of exchange transfusion therapy to lower the percentage of circulating HbS, quantitation of individual hemoglobins is also required. In most laboratories, quantitation is performed only if the test is specifically ordered.

Functional assays for hemoglobin sickling, solubility, or oxygen affinity should also be performed, as dictated by the clinical presentation. The best sickling assays involve measurement of the degree to which the hemoglobin sample becomes insoluble, or gelated, as it is deoxygenated (i.e., sickle solubility test). Unstable

hemoglobins can be detected by their precipitation in isopropanol or after heating to 50°C. High-O_2 affinity and low-O_2 affinity variants can be detected by quantitating the P_{50}, the partial pressure of oxygen at which the hemoglobin sample becomes 50% saturated with oxygen. Direct tests for the percent carboxyhemoglobin and methemoglobin, using spectrophotometric techniques, can readily be obtained from most clinical laboratories on an urgent basis.

Laboratory evaluation remains an adjunct, rather than the sole diagnostic aid. Diagnosis is best established by recognition of a characteristic history, physical findings, peripheral blood smear morphology, and abnormalities of the complete blood cell count (e.g., profound microcytosis with minimal anemia in thalassemia trait).

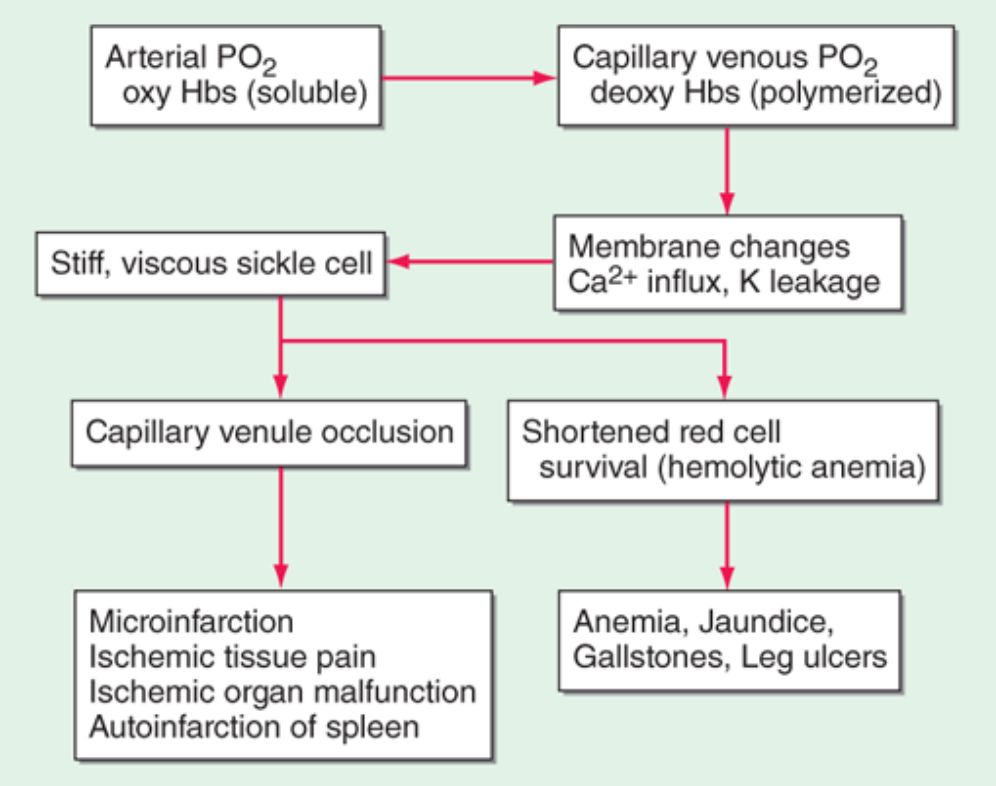
STRUCTURALLY ABNORMAL HEMOGLOBINS

SICKLE CELL SYNDROMES

The sickle cell syndromes are caused by a mutation in the β-globin gene that changes the sixth amino acid from glutamic acid to valine. HbS ($\alpha_2\beta_2^{6\ Glu\rightarrow Val}$) polymerizes reversibly when deoxygenated to form a gelatinous network of fibrous polymers that stiffen the RBC membrane, increase viscosity, and cause dehydration due to potassium leakage and calcium influx **(Fig. 94-3)**. These changes also produce the sickle shape **(Fig. 94-4)**. Sickled cells lose the pliability needed to traverse small capillaries. They possess altered “sticky” membranes that are abnormally adherent to the endothelium of small venules. These abnormalities provoke unpredictable episodes of microvascular vasoocclusion and premature RBC destruction (hemolytic anemia) in the liver and spleen. The rigid adherent cells also clog small capillaries and venules, causing tissue ischemia, acute pain, and gradual end-organ damage. This venoocclusive component usually dominates the clinical course. Prominent manifestations include episodes of ischemic pain (i.e., painful crises) and ischemic malfunction or frank infarction in the spleen, central nervous system, bones, joints, liver, kidneys, and lungs (Fig. 94-3).

FIGURE 94-3

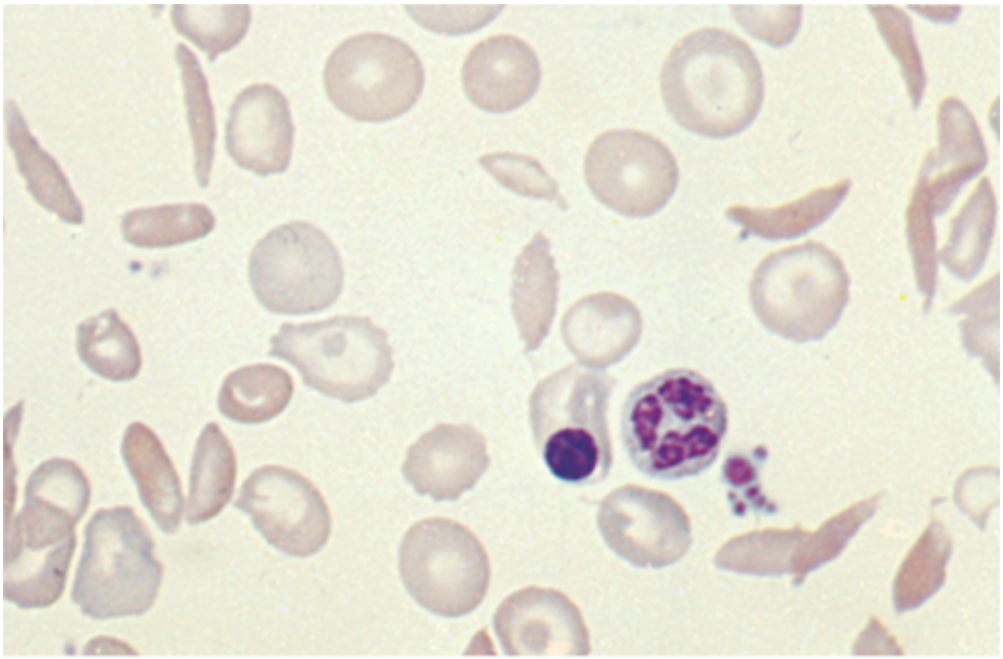
Pathophysiology of sickle cell crisis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 94-4

Sickle cell anemia. The elongated and crescent-shaped red blood cells seen on this smear represent circulating irreversibly sickled cells. Target cells and a nucleated red blood cell are also seen.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Several sickle syndromes occur as the result of inheritance of HbS from one parent and another hemoglobinopathy, such as β thalassemia or HbC ($\alpha_2\beta_2^{6\ Glu\rightarrow Lys}$), from the other parent. The prototype

disease, sickle cell anemia, is the homozygous state for HbS (Table 94-2).

TABLE 94-2

Clinical Features of Sickle Hemoglobinopathies

Condition	Clinical Abnormalities	Hemoglobin Level, g/L (g/dL)	MCV, fL	Hemoglobin Electrophoresis
Sickle cell trait	None; rare painless hematuria	Normal	Normal	HbS/A: 40/60
Sickle cell anemia	Vasoocclusive crises with infarction of spleen, brain, marrow, kidney, lung; aseptic necrosis of bone; gallstones; priapism; ankle ulcers	70–100 (7–10)	80–100	HbS/A: 100/0 HbF: 2–25%
S/β° thalassemia	Vasoocclusive crises; aseptic necrosis of bone	70–100 (7–10)	60–80	HbS/A: 100/0 HbF: 1–10%
S/β+ thalassemia	Rare crises and aseptic necrosis	100–140 (10–14)	70–80	HbS/A: 60/40
Hemoglobin SC	Rare crises and aseptic necrosis; painless hematuria	100–140 (10–14)	80–100	HbS/A: 50/0 HbC: 50%

Clinical Manifestations of Sickle Cell Anemia

Most patients with sickling syndromes suffer from hemolytic anemia, with hematocrits from 15 to 30%, and significant reticulocytosis. Anemia was once thought to exert protective effects against vasoocclusion by reducing blood viscosity. However, natural history and drug therapy trials suggest that an *increase* in the hematocrit and feedback inhibition of reticulocytosis might be beneficial, even at the expense of slightly increased blood viscosity. The role of adhesive reticulocytes in promoting vasoocclusion might account for these paradoxical effects.

Granulocytosis is common. The white count can fluctuate substantially and unpredictably during and between painful crises, infectious episodes, and other intercurrent illnesses. Granulocytes, platelets, and mononuclear inflammatory cells, and the inflammatory mediators that they release at the sites of vasoocclusion, are being increasingly appreciated as key contributors to the initiation and aggravation of the morbidity associated with vasoocclusive crises.

Vasoocclusion causes protean manifestations. Intermittent episodes of vasoocclusion in connective and musculoskeletal structures produce ischemia manifested by acute pain and tenderness, fever, tachycardia, and anxiety. These recurrent episodes, called *painful crises*, are the most common clinical manifestation. Their frequency and severity vary greatly. Pain can develop almost anywhere in the body and may last from a few hours to 2 weeks. Repeated crises requiring hospitalization (>3 episodes per year) correlate with reduced survival in adult life, suggesting that these episodes are associated with accumulation of chronic end-organ damage. Provocative factors include infection, fever, excessive exercise, anxiety, abrupt changes in temperature, hypoxia, or hypertonic dyes.

Repeated microinfarction can destroy tissues having microvascular beds prone to sickling. Thus, splenic function is frequently lost within the first 18–36 months of life, causing susceptibility to infection, particularly by pneumococci. Acute venous obstruction of the spleen (*splenic sequestration crisis*), a rare occurrence in early childhood, may require emergency transfusion and/or splenectomy to prevent trapping of the entire arterial output in the obstructed spleen. Occlusion of retinal vessels can produce hemorrhage, neovascularization, and eventual detachments. Renal papillary necrosis invariably produces isosthenuria. More widespread renal necrosis leads to renal failure in adults, a common late cause of death. Bone and joint ischemia can lead to aseptic necrosis, especially of the femoral or humeral heads, chronic arthropathy, and unusual susceptibility to osteomyelitis, which may be caused by organisms, such as *Salmonella*, rarely encountered in other settings. The *hand-foot syndrome* is caused by painful infarcts of the digits and dactylitis. Stroke is especially common in children; a small subset tends to suffer repeated episodes. Stroke is less common in adults and is often hemorrhagic. A particularly painful complication in males is priapism, due to infarction of the penile venous outflow tracts; permanent impotence is a frequent consequence. Chronic lower leg ulcers probably arise from ischemia and superinfection in the distal circulation.

Acute chest syndrome is a distinctive manifestation characterized by chest pain, tachypnea, fever, cough, and arterial oxygen desaturation. It can mimic pneumonia, pulmonary emboli, bone marrow infarction and embolism, myocardial ischemia, or in situ lung infarction. Acute chest syndrome is thought to be due to in situ sickling within the lung and/or bone marrow microemboli, producing pain and temporary pulmonary dysfunction. Often it is difficult or impossible to distinguish among other possibilities. Pulmonary infarction and pneumonia are the most frequent underlying or concomitant conditions in patients with this syndrome. Repeated episodes of acute chest pain correlate with reduced survival. Acutely, reduction in arterial oxygen saturation is especially ominous because it promotes sickling on a massive scale. Chronic acute or subacute pulmonary crises lead to pulmonary hypertension and cor pulmonale, an increasingly common cause of death as patients survive longer. A possible role played by free plasma HbS in scavenging nitrogen dioxide (NO_2), thus raising pulmonary vascular tone led to trials of sildenafil to restore NO_2 levels. These were terminated because of adverse effects.

Chronic subacute central nervous system damage in the absence of an overt stroke is a distressingly common phenomenon beginning in early childhood. Modern functional imaging techniques have pinpointed circulatory dysfunction due to a likely CNS sickle vasculopathy; these changes correlate with an array of cognitive and behavioral abnormalities in children and young adults. It is important to be aware of these often subtle changes because they can complicate clinical management or be misinterpreted as “difficult patient” behaviors.

Sickle cell syndromes are remarkable for their clinical heterogeneity. Some patients remain virtually asymptomatic into or even through adult life, while others suffer repeated crises requiring hospitalization from early childhood. Patients with sickle thalassemia and sickle-HbE tend to have similar, slightly milder symptoms, perhaps because of the ameliorating effects of the presence of other hemoglobins within the RBC. Hemoglobin SC disease, one of the more common variants of sickle cell anemia, is frequently marked by lesser degrees of hemolytic anemia and a greater propensity for the development of retinopathy and aseptic necrosis of bones. In most respects, however, the clinical manifestations resemble sickle cell anemia. Some rare hemoglobin variants actually aggravate the sickling phenomenon.

The clinical variability in different patients inheriting the same disease-causing mutation (sickle hemoglobin) has made sickle cell disease the focus of efforts to identify modifying genetic polymorphisms in other genes that might account for the heterogeneity. The complexity of the data obtained thus far has dampened the expectation that genome-wide analysis will yield individualized profiles that predict a patient's clinical course. Nevertheless, a number of interesting patterns have emerged from these modifying gene analyses. For example, genes affecting the inflammatory response or cytokine expression appear to be modifying candidates. Genes that affect transcriptional regulation of lymphocytes may also be involved.

Clinical Manifestations of Sickle Cell Trait

Sickle cell trait is often asymptomatic. Anemia and painful crises are rare. An uncommon but highly distinctive symptom is painless hematuria often occurring in adolescent males, probably due to papillary necrosis. Isosthenuria is a more common manifestation of the same process. Sloughing of papillae with urethral obstruction has been reported, as have isolated cases of massive sickling or sudden death due to exposure to high altitudes or extremes of exercise and dehydration. Avoidance of dehydration or extreme physical stress should be advised.

Diagnosis

Sickle cell syndromes are suspected on the basis of hemolytic anemia, RBC morphology (**Fig. 94-4**), and intermittent episodes of ischemic pain. Diagnosis is confirmed by hemoglobin electrophoresis, mass spectroscopy, and the sickling tests already discussed. Thorough characterization of the exact hemoglobin profile of the patient is important, because sickle thalassemia and hemoglobin SC disease have distinct prognoses or clinical features. Diagnosis is usually established in childhood, but occasional patients, often with compound heterozygous states, do not develop symptoms until the onset of puberty, pregnancy, or early adult life. Genotyping of family members and potential parental partners is critical for genetic counseling. Details of the childhood history establish prognosis and need for aggressive or experimental therapies. Factors associated with increased morbidity and reduced survival include more than three crises requiring hospitalization per year, chronic neutrophilia, a history of splenic sequestration or hand-foot syndrome, and second episodes of acute chest syndrome. Patients with a history of cerebrovascular accidents are at higher risk for repeated episodes and require partial exchange transfusion and especially close monitoring using Doppler carotid flow measurements. Patients with severe or repeated episodes of acute chest syndrome may need lifelong transfusion support, using partial exchange transfusion, if possible.

TREATMENT

Sickle Cell Syndromes

Patients with sickle cell syndromes require ongoing continuity of care. Familiarity with the pattern of symptoms provides the best safeguard against excessive use of the emergency room, hospitalization, and habituation to addictive narcotics. Additional preventive measures include regular slit-lamp examinations to monitor development of retinopathy; antibiotic prophylaxis appropriate for splenectomized patients during dental or other invasive procedures; and vigorous oral hydration during or in anticipation of periods of extreme exercise, exposure to heat or cold, emotional stress, or infection. Pneumococcal and *Haemophilus influenzae* vaccines are less effective in splenectomized individuals. Thus, patients with sickle cell anemia should be vaccinated early in life.

The management of an acute painful crisis includes vigorous but careful hydration, thorough evaluation for underlying causes (such as infection), and aggressive analgesia administered by a standing order and/or patient-controlled analgesia (PCA) pump. Morphine (0.1–0.15 mg/kg every 3–4 h) should be used to control severe pain. Bone pain may respond as well to ketorolac (30–60 mg initial dose, then 15–30 mg every 6–8 h). Inhalation of nitrous oxide can provide short-term pain relief, but great care must be exercised to avoid hypoxia and respiratory depression. Nitrous oxide also elevates O_2 affinity, reducing O_2 delivery to tissues. Its use should be restricted to experts. Blockade of the activities of adhesive molecules (e.g., P-selectin) or inflammatory mediators are being used to shorten crises and reduce crisis pain. Many crises can be managed at home with oral hydration and oral analgesia. Use of the emergency room should be reserved for especially severe symptoms or circumstances in which other processes, for example, infection, are strongly suspected. Nasal oxygen should be used as appropriate to protect arterial saturation. Most crises resolve in 1–7 days. Use of blood transfusion should be reserved for extreme cases: transfusions do not shorten the duration of the crisis.

No tests are definitive to diagnose acute painful crisis. Critical to good management is an approach that recognizes that most patients reporting crisis symptoms do indeed have crisis or another significant medical problem. Diligent diagnostic evaluation for underlying causes is imperative, even though these are found infrequently. In adults, the possibility of aseptic necrosis or sickle arthropathy must be considered, especially if pain and immobility become repeated or chronic at a single site. Nonsteroidal anti-inflammatory agents are often effective for sickle cell arthropathy.

Acute chest syndrome is a medical emergency that may require management in an intensive care unit. Hydration should be monitored carefully to avoid the development of pulmonary edema, and oxygen therapy should be especially vigorous for protection of arterial saturation. Diagnostic evaluation for pneumonia and pulmonary embolism should be especially thorough, since these may occur with atypical symptoms. Critical interventions are transfusion to maintain a hematocrit >30, and emergency exchange transfusion if arterial saturation drops to <90%. As patients with sickle cell syndrome increasingly survive into their fifth and sixth decades, end-stage renal failure and pulmonary hypertension are becoming increasingly prominent causes of

end-stage morbidity. A sickle cell cardiomyopathy and/or premature coronary artery disease may compromise cardiac function in later years. Sickle cell patients have received kidney transplants, but they often experience an increase in the frequency and severity of crises, possibly due to increased infection as a consequence of immunosuppression.

The most significant advance in the therapy of sickle cell anemia has been the introduction of hydroxyurea as a mainstay of therapy for patients with severe symptoms. Hydroxyurea (10–30 mg/kg per day) increases fetal hemoglobin and may also exert beneficial effects on RBC hydration, vascular wall adherence, and suppression of the granulocyte and reticulocyte counts; dosage is titrated to maintain a white cell count between 5000 and 8000/μL. White cells and reticulocytes may play a major role in the pathogenesis of sickle cell crisis, and their suppression may be an important side benefit of hydroxyurea therapy.

Hydroxyurea should be considered in patients experiencing repeated episodes of acute chest syndrome or with more than three crises per year requiring hospitalization. The utility of this agent for reducing the incidence of other complications (priapism, retinopathy) is under evaluation, as are the long-term side effects. Clinical experience is now sufficient to state that the risk of bone marrow dyscrasias or other neoplasms is minimal. Hydroxyurea offers broad benefits to most patients whose disease is severe enough to impair their functional status; it likely improves survival. HbF levels increase in most patients within a few months.

The antitumor drug 5-azacytidine was the first agent found to elevate HbF. It never achieved widespread use because of concerns about acute toxicity and carcinogenesis. However, low doses of the related agent 5-deoxyazacytidine (decitabine) can elevate HbF with more acceptable toxicity.

Bone marrow transplantation can provide definitive cures but is known to be effective and safe only in children. Clinical trials studying partially myeloablative conditioning regimens (“mini” transplants) are likely to support more widespread use in older patients, and should be considered in adults with significant morbidity. Prognostic features justifying bone marrow transplant are the presence of repeated crises early in life, a high neutrophil count, or the development of hand-foot syndrome. Children at risk for stroke can now be identified through the use of Doppler ultrasound techniques. Prophylactic exchange transfusion appears to substantially reduce the risk of stroke in this population. Children who do suffer a cerebrovascular accident should be maintained for at least 3–5 years on a program of vigorous exchange transfusion, as the risk of second strokes is extremely high. Hydroxyurea therapy could eventually supplant indefinite chronic transfusion in patients responding favorably.

Gene therapy for sickle cell anemia is being intensively pursued, but no safe measures are currently available. The development of newer methods of direct gene correction in situ (e.g., zinc finger nucleases, or “CRISPR” [clustered regularly interspaced short palindromic repeats] technology) could well find clinical use in these patients. Experimental methods of derepressing HbF by manipulating Bcl11a or chromatin looping are also being explored.

UNSTABLE HEMOGLOBINS

Amino acid substitutions that reduce solubility or increase susceptibility to oxidation result in unstable hemoglobins that precipitate, forming inclusion bodies injurious to the RBC membrane. Representative mutations are those that interfere with contact points between the α and β subunits (e.g., Hb Philly [$\beta^{35Tyr\rightarrow Phe}$]), alter the helical segments (e.g., Hb Genova [$\beta^{28Leu\rightarrow Pro}$]), or disrupt interactions of the hydrophobic pockets of the globin subunits with heme (e.g., Hb Köln [$\beta^{98Val\rightarrow Met}$]) **(Table 94-3)**. The inclusions, called *Heinz bodies*, are clinically detectable by staining with supravital dyes such as crystal violet. Removal of these inclusions by the spleen generates pitted, rigid cells that have shortened life spans, producing hemolytic anemia of variable severity, sometimes requiring chronic transfusion support. Splenectomy may be needed to correct the anemia. Leg ulcers and premature gallbladder disease due to bilirubin loading are frequent stigmata.

TABLE 94-3

Representative Abnormal Hemoglobins with Altered Synthesis or Function

Designation	Mutation	Population	Main Clinical Effects[a]
Sickle or S	$\beta^{6Glu\rightarrow Val}$	African	Anemia, ischemic infarcts
C	$\beta^{6Glu\rightarrow Lys}$	African	Mild anemia; interacts with HbS
E	$\beta^{26Glu\rightarrow Lys}$	Southeast Asian	Microcytic anemia, splenomegaly, thalassemic phenotype
Köln	$\beta^{98Val\rightarrow Met}$	Sporadic	Hemolytic anemia, Heinz bodies when splenectomized
Yakima	$\beta^{99Asp\rightarrow His}$	Sporadic	Polycythemia
Kansas	$\beta^{102Asn\rightarrow Lys}$	Sporadic	Mild anemia
M Iwata	$\alpha^{87His\rightarrow Tyr}$	Sporadic	Methemoglobinemia

[a]See text for details.

Unstable hemoglobins occur sporadically, often by spontaneous new mutations. Heterozygotes are often symptomatic because a significant Heinz body burden can develop even when the unstable variant accounts for only a portion of the total hemoglobin. Symptomatic unstable hemoglobins tend to be β-globin variants, because sporadic mutations affecting only one of the four α globins alleles would generate only 20–30% abnormal hemoglobin.

HEMOGLOBINS WITH ALTERED OXYGEN AFFINITY

High-affinity hemoglobins (e.g., Hb Yakima [$\beta^{99Asp \rightarrow His}$]) bind oxygen more readily but deliver less O_2 to tissues at normal capillary Po_2 levels (Fig. 94-2). Mild tissue hypoxia ensues, stimulating RBC production and erythrocytosis (Table 94-3). In extreme cases, the hematocrits can rise to 60–65%, increasing blood viscosity and producing typical symptoms (headache, somnolence, or dizziness). Phlebotomy may be required. Typical mutations alter interactions within the heme pocket or disrupt the Bohr effect or salt-bond site. Mutations that impair the interaction of HbA with 2,3-BPG can increase O_2 affinity because 2,3-BPG binding lowers O_2 affinity.

Low-affinity hemoglobins (e.g., Hb Kansas [$\beta^{102Asn \rightarrow Lys}$]) bind sufficient oxygen in the lungs, despite their lower oxygen affinity, to achieve nearly full saturation. At capillary oxygen tensions, they lose sufficient amounts of oxygen to maintain homeostasis at a low hematocrit (Fig. 94-2) (*pseudoanemia*). Capillary hemoglobin desaturation can also be sufficient to produce clinically apparent cyanosis. Despite these findings, patients usually require no specific treatment.

METHEMOGLOBINEMIAS

Methemoglobin is generated by oxidation of the heme iron moieties to the ferric state, causing a characteristic bluish-brown muddy color resembling cyanosis. Methemoglobin has such high oxygen affinity that virtually no oxygen is delivered. Levels >50–60% are often fatal.

Congenital methemoglobinemia arises from globin mutations that stabilize iron in the ferric state (e.g., HbM Iwata [$\alpha^{87His \rightarrow Tyr}$], Table 94-3) or from mutations that impair the enzymes that reduce methemoglobin to hemoglobin (e.g., methemoglobin reductase, NADP diaphorase). Acquired methemoglobinemia is caused by toxins that oxidize heme iron, notably nitrate and nitrite-containing compounds, including drugs commonly used in cardiology and anesthesiology.

DIAGNOSIS AND MANAGEMENT OF PATIENTS WITH UNSTABLE HEMOGLOBINS, HIGH-AFFINITY HEMOGLOBINS, AND METHEMOGLOBINEMIA

Unstable hemoglobin variants should be suspected in patients with nonimmune hemolytic anemia, jaundice, splenomegaly, or premature biliary tract disease. Severe hemolysis usually presents during infancy as neonatal jaundice or anemia. Milder cases may present in adult life with anemia or only as unexplained reticulocytosis, hepatosplenomegaly, premature biliary tract disease, or leg ulcers. Because spontaneous mutation is common, family history of anemia may be absent. The peripheral blood smear often shows anisocytosis, abundant cells with punctate inclusions, and irregular shapes (i.e., poikilocytosis).

The two best tests for diagnosing unstable hemoglobins are the Heinz body preparation and the isopropanol or heat stability test. Many unstable Hb variants are electrophoretically silent. A normal electrophoresis does not rule out the diagnosis. HPLC or direct gene analysis will provide a definitive diagnosis.

Severely affected patients may require transfusion support for the first 3 years of life, because splenectomy before age 3 is associated with a significantly higher immune deficit. Splenectomy is usually effective thereafter, but occasional patients may require lifelong transfusion support. After splenectomy, patients can develop

cholelithiasis and leg ulcers, hypercoagulable states, and susceptibility to overwhelming sepsis. Splenectomy should thus be avoided or delayed unless it is the only alternative. Precipitation of unstable hemoglobins is aggravated by oxidative stress, for example, infection and certain antimalarial drugs, which should be avoided where possible.

High-O_2 affinity hemoglobin variants should be suspected in patients with erythrocytosis. The best test for confirmation is measurement of the P_{50}. A high-O_2 affinity hemoglobin causes a significant left shift (i.e., lower numeric value of the P_{50}); confounding conditions, for example, tobacco smoking or carbon monoxide exposure, can also lower the P_{50}.

High-affinity hemoglobins are often asymptomatic; rubor or plethora may be telltale signs. When the hematocrit approaches 60%, symptoms of high blood viscosity and sluggish flow (headache, lethargy, dizziness, etc.) may be present. These persons may benefit from judicious phlebotomy. Erythrocytosis represents an appropriate attempt to compensate for the impaired oxygen delivery by the abnormal variant. Overzealous phlebotomy may stimulate increased erythropoiesis or aggravate symptoms by thwarting this compensatory mechanism. The guiding principle of phlebotomy should be to improve oxygen delivery by reducing blood viscosity and increasing blood flow rather than restoration of a normal hematocrit. Phlebotomy-induced modest iron deficiency may aid in control.

Low-affinity hemoglobins should be considered in patients with cyanosis or a low hematocrit with no other reason apparent after thorough evaluation. The P_{50} test confirms the diagnosis. Counseling and reassurance are the interventions of choice.

Acquired methemoglobinemia should be suspected in patients with hypoxic symptoms who appear cyanotic but have a Pa_{O_2} sufficiently high that hemoglobin should be fully saturated with oxygen. A history of nitrite or other oxidant ingestions may not always be available; some exposures may be inapparent to the patient, and others may result from suicide attempts. The characteristic muddy appearance of freshly drawn blood can be a critical clue. The best diagnostic test is methemoglobin assay, which is usually available on an emergency basis.

Methemoglobinemia often causes symptoms of cerebral ischemia at levels >15%; levels >60% are usually lethal. Intravenous injection of 1 mg/kg of methylene blue is effective emergency therapy. Milder cases and follow-up of severe cases can be treated orally with methylene blue (60 mg three to four times each day) or ascorbic acid (300–600 mg/d). Congenital methemoglobinemia does not usually require treatment other than avoidance of oxidative drugs or agents. Patient and provider awareness is essential in order that inappropriate evaluations for cyanosis be avoided.

THALASSEMIA SYNDROMES

The thalassemia syndromes are inherited disorders of α- or β-globin biosynthesis. The reduced supply of globin diminishes production of hemoglobin tetramers, causing hypochromia and microcytosis. Unbalanced accumulation of α and β subunits occurs because the synthesis of the unaffected globins proceeds at a normal rate. Unbalanced chain accumulation dominates the clinical phenotype. Clinical severity varies widely,

depending on the degree to which the synthesis of the affected globin is impaired, altered synthesis of other globin chains, and coinheritance of other abnormal globin alleles.

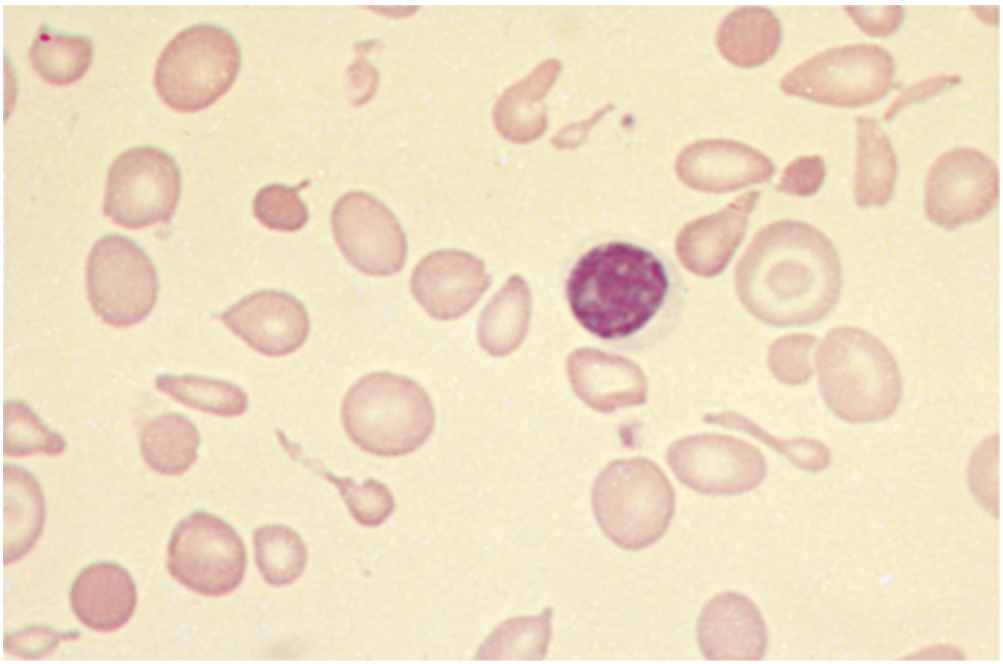
CLINICAL MANIFESTATIONS OF β THALASSEMIA SYNDROMES

Mutations causing thalassemia can affect any step in the pathway of globin gene expression: transcription, processing of the mRNA precursor, translation, and posttranslational metabolism of the β-globin polypeptide chain. The most common forms arise from mutations that derange splicing of the mRNA precursor or prematurely terminate translation of the mRNA.

Hypochromia and microcytosis characterize all forms of β thalassemia because of the reduced amounts of hemoglobin tetramers (Fig. 94-5). In heterozygotes (β thalassemia trait), this is the only abnormality seen. Anemia is minimal. In more severe homozygous states, unbalanced α- and β-globin accumulation causes accumulation of highly insoluble unpaired α chains. They form toxic inclusion bodies that kill developing erythroblasts in the marrow. Few of the proerythroblasts beginning erythroid maturation survive. The surviving RBCs bear a burden of inclusion bodies that are detected in the spleen, shortening the RBC life span and producing severe hemolytic anemia. The resulting profound anemia stimulates erythropoietin release and compensatory erythroid hyperplasia, but the marrow response is sabotaged by the ineffective erythropoiesis. Anemia persists. Erythroid hyperplasia can become exuberant and produce masses of extramedullary erythropoietic tissue in the liver and spleen.

FIGURE 94-5

a Thalassemia intermedia. Microcytic and hypochromic red blood cells are seen that resemble the red blood cells of severe iron-deficiency anemia. Many elliptical and teardrop-shaped red blood cells are noted.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Massive bone marrow expansion deranges growth and development. Children develop characteristic "chipmunk" facies due to maxillary marrow hyperplasia and frontal bossing. Thinning and pathologic fracture

of long bones and vertebrae may occur due to cortical invasion by erythroid elements and profound growth retardation. Hemolytic anemia causes hepatosplenomegaly, leg ulcers, gallstones, and high-output congestive heart failure. The conscription of caloric resources to support erythropoiesis leads to inanition, susceptibility to infection, endocrine dysfunction, and, in the most severe cases, death during the first decade of life. Chronic transfusions with RBCs improve oxygen delivery, suppress the excessive ineffective erythropoiesis, and prolong life, but the inevitable side effects, notably iron overload, often prove fatal by age 30 years.

Severity is highly variable. Known modulating factors are those that ameliorate the burden of unpaired α-globin inclusions. Alleles associated with milder synthetic defects and coinheritance of α thalassemia trait reduce clinical severity by reducing accumulation of excess α globin. HbF persists to various degrees in β thalassemias. γ-Globin gene chains can substitute for β chains, generating more hemoglobin and reducing the burden of α-globin inclusions. The terms *β thalassemia major* and *β thalassemia intermedia* are used to reflect the clinical heterogeneity. Patients with β thalassemia major require intensive transfusion support to survive. Patients with β thalassemia intermedia have a somewhat milder phenotype and can survive without transfusion. The terms *β thalassemia minor* and *β thalassemia trait* describe asymptomatic heterozygotes for β thalassemia.

ALPHA THALASSEMIA SYNDROMES

The four classic α thalassemias, most common in Asians, are α thalassemia-2 trait, in which one of the four α-globin loci is deleted; α thalassemia-1 trait, with two deleted loci; HbH disease, with three loci deleted; and hydrops fetalis with Hb Barts, with all four loci deleted (**Table 94-4**). Nondeletion forms of α thalassemia also exist.

TABLE 94-4

The α Thalassemias

Condition	Hemoglobin A, %	Hemoglobin H (β4), %	Hemoglobin Level, g/L (g/dL)	MCV, fL
Normal	97	0	150 (15)	90
Silent thalassemia: −α/αα	98–100	0	150 (15)	90
Thalassemia trait:−α/−α homozygous α-thal-2[a]	85–95	Rare red blood cell inclusions	120–130 (12–13)	70–80
or				
−−/αα heterozygous α-thal-1[a]				
Hemoglobin H disease: −−/−α heterozygous α-thal-1/α-thal-2	70–95	5–30	60–100 (6–10)	60–70
Hydrops fetalis: −−/−− homozygous α-thal-1	0	5–10[b]	Fatal in utero or at birth	

[a]When both α alleles on one chromosome are deleted, the locus is called α-thal-1; when only a single α allele on one chromosome is deleted, the locus is called α-thal-2. [b]90–95% of the hemoglobin is hemoglobin Barts (tetramers of γ chains).

α Thalassemia-2 trait is an asymptomatic, silent carrier state. *α Thalassemia-1 trait* resembles β thalassemia minor. Offspring doubly heterozygous for α thalassemia-2 and α thalassemia-1 exhibit a more severe phenotype called *HbH disease*. Heterozygosity for a deletion that removes both genes from the same chromosome (*cis* deletion) is common in Asians and in those from the Mediterranean region, as is homozygosity for α thalassemia-2 (*trans* deletion). Both produce asymptomatic hypochromia and microcytosis.

In *HbH disease*, HbA production is only 25–30% normal. Fetuses accumulate some unpaired γ chains (Hb Barts; γ-chain tetramers). In adults, unpaired β chains accumulate and are soluble enough to form β_4 tetramers called HbH. HbH forms few inclusions in erythroblasts and precipitates in circulating RBC. Patients with HbH disease have thalassemia intermedia characterized by moderately severe hemolytic anemia but milder ineffective erythropoiesis. Survival into mid-adult life without transfusions is common.

The homozygous state for the α thalassemia-1 *cis* deletion (hydrops fetalis) causes total absence of α-globin synthesis. No physiologically useful hemoglobin is produced beyond the embryonic stage. Excess γ globin

forms tetramers called *Hb Barts* (γ_4), which has a very high oxygen affinity. It delivers almost no O_2 to fetal tissues, causing tissue asphyxia, edema (hydrops fetalis), congestive heart failure, and death in utero. α Thalassemia-2 trait is common (15–20%) among people of African descent. The *cis* α thalassemia-1 deletion is almost never seen, however. Thus, α thalassemia-2 and the *trans* form of α thalassemia-1 are very common, but HbH disease and hydrops fetalis are rare.

It has been known for some time that some patients with myelodysplasia or erythroleukemia produce RBC clones containing HbH. This phenomenon is due to mutations in the ATRX pathway that affect the LCR of the α-globin gene cluster.

DIAGNOSIS AND MANAGEMENT OF THALASSEMIAS

The diagnosis of β-thalassemia major is readily made during childhood on the basis of severe anemia accompanied by the characteristic signs of massive ineffective erythropoiesis: hepatosplenomegaly, profound microcytosis, a characteristic blood smear (Fig. 94-5), and elevated levels of HbF, HbA_2, or both. Many patients require chronic hypertransfusion therapy designed to maintain a hematocrit of at least 27–30% so that erythropoiesis is suppressed. Splenectomy is required if the annual transfusion requirement (volume of RBCs per kilogram of body weight per year) increases by >50%. Folic acid supplements may be useful. Vaccination with Pneumovax in anticipation of eventual splenectomy is advised, as is close monitoring for infection, leg ulcers, and biliary tract disease. Many patients develop endocrine deficiencies as a result of iron overload. Early endocrine evaluation is required for glucose intolerance, thyroid dysfunction, and delayed onset of puberty or secondary sexual characteristics.

Patients with β thalassemia intermedia exhibit similar stigmata but can survive without chronic hypertransfusion. Management is particularly challenging because a number of factors can aggravate the anemia, including infection, onset of puberty, and development of splenomegaly and hypersplenism. Some patients may eventually benefit from splenectomy. The expanded erythron can cause absorption of excessive dietary iron and hemosiderosis, even without transfusion. Some patients eventually become transfusion dependent.

β Thalassemia minor (i.e., thalassemia trait) usually presents as profound microcytosis and hypochromia with target cells, but only minimal or mild anemia. The mean corpuscular volume is rarely >75 fL; the hematocrit is rarely <30–33%. Hemoglobin analysis classically reveals an elevated HbA_2 (3.5–7.5%), but some forms are associated with normal HbA_2 and/or elevated HbF. Genetic counseling and patient education are essential. Patients with β thalassemia trait should be warned that their blood picture resembles iron deficiency and can be misdiagnosed. They should eschew empirical use of iron, yet iron deficiency requiring replacement therapy can develop during pregnancy or from chronic bleeding.

Persons with α thalassemia trait may exhibit mild hypochromia and microcytosis usually without anemia. HbA_2 and HbF levels are normal. Affected individuals usually require only genetic counseling. HbH disease resembles β thalassemia intermedia, with the added complication that the HbH molecule behaves like moderately unstable hemoglobin. Patients with HbH disease should undergo splenectomy if excessive anemia or a

transfusion requirement develops. Oxidative drugs should be avoided. Iron overload leading to death can occur in more severely affected patients.

PREVENTION

Antenatal diagnosis of thalassemia syndromes is now widely available. DNA diagnosis is based on polymerase chain reaction (PCR) amplification of fetal DNA, obtained by amniocentesis or chorionic villus biopsy followed by hybridization to allele-specific oligonucleotide probes or direct DNA sequencing.

THALASSEMIC STRUCTURAL VARIANTS

Thalassemic structural variants are characterized by both defective synthesis and abnormal structure.

HEMOGLOBIN LEPORE

Hb Lepore [$\alpha_2(\delta\beta)_2$] arises by an unequal crossover and recombination event that fuses the proximal end of the δ-gene with the distal end of the closely linked β-gene. It is common in the Mediterranean basin. The resulting chromosome contains only the fused δβ gene. The Lepore (δβ) globin is synthesized poorly because the fused gene is under the control of the weak δ-globin promoter. Hb Lepore alleles have a phenotype like β thalassemia, except for the added presence of 2–20% Hb Lepore. Compound heterozygotes for Hb Lepore and a classic β thalassemia allele may also have severe thalassemia.

HEMOGLOBIN E

 HbE (i.e., $\alpha_2\beta_2^{26Glu\rightarrow Lys}$) is extremely common in Cambodia, Thailand, and Vietnam. The gene has become far more prevalent in the United States as a result of immigration of Asian persons, especially in California, where HbE is the most common variant detected. HbE is mildly unstable but not enough to affect RBC life span significantly. Heterozygotes resemble individuals with a mild β-thalassemia trait. HbE homozygosity is a condition associated with mildly asymptomatic microcytosis, hypochromia, and hemoglobin levels rarely <100 g/L (<10 g/dL).

Compound heterozygotes for HbE and a β thalassemia gene can have β thalassemia intermedia or β-thalassemia major, depending on the severity of the coinherited thalassemic gene.

The β^E allele contains a single base change in codon 26 that causes the amino acid substitution. This mutation also activates a cryptic RNA splice site, generating a structurally abnormal globin mRNA that cannot be translated, from about 50% of the initial pre-mRNA molecules. The remaining 40–50% are normally spliced and generate functional mRNA that is translated into β^E-globin because the mature mRNA carries the base change that alters codon 26.

Genetic counseling of the persons at risk for HbE should focus especially on the interaction of HbE with β thalassemia, because HbE homozygosity is a condition associated with microcytosis, hypochromia, and hemoglobin levels rarely below 100 g/L (10 g/dL) and is usually asymptomatic.

HEREDITARY PERSISTENCE OF FETAL HEMOGLOBIN

HPFH is characterized by continued synthesis of high levels of HbF in adult life. No deleterious effects are apparent, even when all of the hemoglobin produced is HbF. These rare patients demonstrate convincingly that prevention or reversal of the fetal to adult hemoglobin switch would provide effective therapy for sickle cell anemia and β thalassemia.

ACQUIRED HEMOGLOBINOPATHIES

The two most important acquired hemoglobinopathies are carbon monoxide poisoning and methemoglobinemia (see above). Carbon monoxide has a higher affinity for hemoglobin than does oxygen; it can replace oxygen and diminish O_2 delivery. Chronic elevation of carboxyhemoglobin levels to 10 or 15%, as occurs in smokers, can lead to secondary polycythemia. Carboxyhemoglobin is cherry red in color and masks the development of cyanosis usually associated with poor O_2 delivery to tissues.

Abnormalities of hemoglobin biosynthesis have also been described in blood dyscrasias. In some patients with myelodysplasia, erythroleukemia, or myeloproliferative disorders, elevated HbF or a mild form of HbH disease may also be seen. The abnormalities are not severe enough to alter the course of the underlying disease.

TREATMENT

TREATMENT

Transfusional Hemosiderosis

Chronic blood transfusion can lead to bloodborne infection, alloimmunization, febrile reactions, and lethal iron overload **(Chap. 109)**. A unit of packed RBCs contains 250–300 mg iron (1 mg/mL). The iron assimilated by a single transfusion of 2 units of packed RBCs is thus equal to a 1- to 2-year oral intake of iron. Iron accumulates in chronically transfused patients because no mechanisms exist for increasing iron excretion: an expanded erythron causes especially rapid development of iron overload because accelerated erythropoiesis promotes excessive absorption of dietary iron. Vitamin C should not be supplemented because it generates free radicals in iron excess states.

Patients who receive >100 units of packed RBCs usually develop hemosiderosis. The ferritin level rises, followed by early endocrine dysfunction (glucose intolerance and delayed puberty), cirrhosis, and cardiomyopathy. Liver biopsy shows both parenchymal and reticuloendothelial iron. The superconducting quantum-interference device (SQUID) is accurate at measuring hepatic iron but not widely available. Cardiac toxicity is often insidious. Early development of pericarditis is followed by dysrhythmia and pump failure. The onset of heart failure is ominous, often presaging death within a year **(Chap. 407)**.

The decision to start long-term transfusion support should also prompt one to institute therapy with iron-chelating agents. Deferoxamine (Desferal) is for parenteral use. Its iron-binding kinetics require chronic slow infusion via a metering pump. The constant presence of the drug improves the efficiency of chelation and

protects tissues from occasional releases of the most toxic fraction of iron—low-molecular-weight iron—which may not be sequestered by protective proteins.

Deferoxamine is relatively nontoxic. Occasional cataracts, deafness, and local skin reactions, including urticaria, occur. Skin reactions can usually be managed with antihistamines. Negative iron balance can be achieved, even in the face of a high transfusion requirement, but this alone does not prevent long-term morbidity and mortality in chronically transfused patients. Irreversible end-organ deterioration develops at relatively modest levels of iron overload, even if symptoms do not appear for many years thereafter. To enjoy a significant survival advantage, chelation must begin before 5–8 years of age in β-thalassemia major.

Deferasirox is an oral iron-chelating agent. Single daily doses of 20–30 mg/kg deferasirox produced reductions in liver iron concentration comparable to deferoxamine in long-term transfused adult and pediatric patients. Deferasirox produces some elevations in liver enzymes and slight but persistent increases in serum creatinine, without apparent clinical consequence. Other toxicities are similar to those of deferoxamine. Its toxicity profile is acceptable, although long-term effects are still being evaluated.

EXPERIMENTAL THERAPIES

BONE MARROW TRANSPLANTATION, GENE THERAPY, AND MANIPULATION OF HbF

Bone marrow transplantation provides stem cells able to express normal hemoglobin; it has been used in a large number of patients with β thalassemia and a smaller number of patients with sickle cell anemia. Early in the course of disease, before end-organ damage occurs, transplantation is curative in 80–90% of patients. In highly experienced centers, the treatment-related mortality is <10%. Because survival into adult life is possible with conventional therapy, the decision to transplant is best made in consultation with specialized centers.

Gene therapy of thalassemia and sickle cell disease has proved to be an elusive goal, but experimental advances are raising expectations. New lentivirus vectors appear to be capable of achieving stable synthesis of anti-sickling hemoglobin variants in some cases without insertional mutagenesis that has been seen with other vectors.

Reestablishing high levels of fetal hemoglobin synthesis should ameliorate the symptoms of β-chain hemoglobinopathies. Cytotoxic agents such as hydroxyurea and cytarabine promote high levels of HbF synthesis, probably by stimulating proliferation of the primitive HbF-producing progenitor cell population (i.e., F cell progenitors). Unfortunately, this regimen has not yet been effective in β thalassemia. Butyrates stimulate HbF production, but only transiently. Pulsed or intermittent administration has been found to sustain HbF induction in the majority of patients with sickle cell disease. It is unclear whether butyrates will have similar activity in patients with β thalassemia.

APLASTIC AND HYPOPLASTIC CRISIS IN PATIENTS WITH HEMOGLOBINOPATHIES

Patients with hemolytic anemias sometimes exhibit an alarming decline in hematocrit during and immediately after acute illnesses. Bone marrow suppression occurs in almost everyone during acute and chronic inflammatory illnesses. In patients with short RBC life spans, suppression can affect RBC counts more dramatically. These hypoplastic crises are usually transient and self-correcting before intervention is required.

Aplastic crisis refers to a profound cessation of erythroid activity in patients with chronic hemolytic anemias. It is associated with a rapidly falling hematocrit. Episodes are usually self-limited. Aplastic crises are caused by infection with a particular strain of parvovirus, B19A. Children infected with this virus usually develop permanent immunity. Aplastic crises do not often recur and are rarely seen in adults. Management requires close monitoring of the hematocrit and reticulocyte count. If anemia becomes symptomatic, transfusion support is indicated. Most crises resolve spontaneously within 1–2 weeks.

FURTHER READING

Benz EJ Jr, Ebert BL: Hemoglobin structural variants associated with hemolytic anemia, altered oxygen affinity, and methemoglobinemias, in *Hematology Basic Principles and Practice,* 6th ed, Hoffman R et al (eds). Philadelphia, Elsevier, 573–581, 2013.

Costa FF, Fertrin KY: Clinical manifestations and treatment of adult sickle cell disease, in *Sickle Cell Anemia*: *From Basic Science to Clinical Practice*, Costa FF, Conran N (eds). Cham, Switzerland, Springer, 285, 2016.

Higgs DR, Engel JD, Stamatoyannopoulos G: Thalassaemia. Lancet 379:373, 2012. [PubMed: 21908035]

Howard J et al: Guideline on the management of acute chest syndrome in sickle cell disease. Br J Haematol 169:492, 2015. [PubMed: 25824256]

Jorge SE et al: Hemoglobin: Structure, synthesis and oxygen transport, in *Sickle Cell Anemia*: *From Basic Science to Clinical Practice*, Costa FF, Conran N (eds). Cham, Switzerland, Springer, 1, 2016.

Maserejian N et al: Beta-thalassemia treatment and complications in two large US insured population databases. Blood 126:3278, 2015.

Robinson TM, Fuchs EJ: Allogeneic stem cell transplantation for sickle cell disease. Curr Opin Hematol 23:524, 2016. [PubMed: 27496639]

Steinberg MH et al: *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management*. Cambridge, Massachusetts, Cambridge University Press, 2001.

Steinberg MH et al: Pathobiology of the human erythrocyte and its hemoglobins, in *Hematology Basic Principles and Practice*, 6th ed, Hoffman R et al (eds). Philadelphia, Elsevier, 2013, pp. 406–418.

Steven A et al: Neurologic and head and neck manifestations of sickle cell disease. Hematol Oncol Clin North Am 30:779, 2016.

Weatherall DJ, Clegg JB: *The Thalassaemia Syndromes*, 4th ed. Oxford, United Kingdom, Blackwell Science Limited, 2001.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 95: Megaloblastic Anemias

A. Victor Hoffbrand

INTRODUCTION

The megaloblastic anemias are a group of disorders characterized by the presence of distinctive morphologic appearances of the developing red cells in the bone marrow. The marrow is usually hypercellular and the anemia is based on ineffective erythropoiesis. The cause is usually a deficiency of either cobalamin (vitamin B_{12}) or folate, but megaloblastic anemia may occur because of genetic or acquired abnormalities that affect the metabolism of these vitamins or because of defects in DNA synthesis not related to cobalamin or folate (**Table 95-1**). Cobalamin and folate absorption and metabolism are described next, followed by the biochemical basis, clinical and laboratory features, causes, and treatment of megaloblastic anemia.

TABLE 95-1

Causes of Megaloblastic Anemia

Cobalamin deficiency or abnormalities of cobalamin metabolism **(see Tables 95-3, 95-4)**
Folate deficiency or abnormalities of folate metabolism **(see Table 95-5)**
Therapy with antifolate drugs (e.g., methotrexate)
Independent of either cobalamin or folate deficiency and refractory to cobalamin and folate therapy:
Some cases of acute myeloid leukemia, myelodysplasia
Therapy with drugs interfering with synthesis of DNA (e.g., cytosine arabinoside, hydroxyurea, 6-mercaptopurine, azidothymidine [AZT])
Orotic aciduria (responds to uridine)
Thiamine-responsive

COBALAMIN

Cobalamin (vitamin B_{12}) exists in a number of different chemical forms. All have a cobalt atom at the center of a corrin ring. In nature, the vitamin is mainly in the 2-deoxyadenosyl (ado) form, which is located in mitochondria. It is the cofactor for the enzyme L-methylmalonyl coenzyme A (CoA) mutase. The other major natural cobalamin is methylcobalamin, the form in human plasma and in cell cytoplasm. It is the cofactor for methionine synthase. There are also minor amounts of hydroxocobalamin to which methyl- and adocobalamin are converted rapidly by exposure to light.

DIETARY SOURCES AND REQUIREMENTS

Cobalamin is synthesized solely by microorganisms. Ruminants obtain cobalamin from the foregut, but the only source for humans is food of animal origin, for example, meat, fish, and dairy products. Vegetables, fruits, and other foods of nonanimal origin are free from cobalamin unless they are contaminated by bacteria. A normal Western diet contains 5–30 μg of cobalamin daily. Adult daily losses (mainly in the urine and feces) are 1–3 μg (~0.1% of body stores), and because the body does not have the ability to degrade cobalamin, daily requirements are also about 1–3 μg. Body stores are of the order of 2–3 mg, sufficient for 3–4 years if supplies are completely cut off.

ABSORPTION

Two mechanisms exist for cobalamin absorption. One is passive, occurring equally through buccal, duodenal, and ileal mucosa; it is rapid but extremely inefficient, with <1% of an oral dose being absorbed by this process. The normal physiologic mechanism is active; it occurs through the ileum and is efficient for small (a few micrograms) oral doses of cobalamin, and it is mediated by gastric intrinsic factor (IF). Dietary cobalamin is released from protein complexes by enzymes in the stomach, duodenum, and jejunum; it combines rapidly with a salivary glycoprotein that belongs to the family of cobalamin-binding proteins known as haptocorrins (HCs). In the intestine, the HC is digested by pancreatic trypsin and the cobalamin is transferred to IF.

IF (gene at chromosome 11q13) is produced in the gastric parietal cells of the fundus and body of the stomach, and its secretion parallels that of hydrochloric acid. Normally, there is a vast excess of IF. The IF-cobalamin complex passes to the ileum, where IF attaches to a specific receptor (cubilin) on the microvillus membrane of the enterocytes. Cubilin also is present in yolk sac and renal proximal tubular epithelium. Cubilin appears to traffic by means of amnionless (AMN), an endocytic receptor protein that directs sublocalization and endocytosis of cubilin with its ligand IF-cobalamin complex. The cobalamin-IF complex enters the ileal cell, where IF is destroyed. After a delay of about 6 h, the cobalamin appears in portal blood attached to transcobalamin (TC) II.

Between 0.5 and 5 µg of cobalamin enter the bile each day. This binds to IF, and a major portion of biliary cobalamin normally is reabsorbed together with cobalamin derived from sloughed intestinal cells. Because of the appreciable amount of cobalamin undergoing enterohepatic circulation, cobalamin deficiency develops more rapidly in individuals who malabsorb cobalamin than it does in vegans, in whom reabsorption of biliary cobalamin is intact.

TRANSPORT

Two main cobalamin transport proteins exist in human plasma; they both bind cobalamin—one molecule for one molecule. One HC, also known as TC I, is closely related to other cobalamin-binding HCs in milk, gastric juice, bile, saliva, and other fluids. The gene *TCNL* is at chromosome 11q11-q12.3. These HCs differ from each other only in the carbohydrate moiety of the molecule. TC I is derived primarily from the specific granules in neutrophils. Normally, it is about two-thirds saturated with cobalamin, which it binds tightly. TC I does not enhance cobalamin entry into tissues. Glycoprotein receptors on liver cells are involved in the removal of TC I from plasma, and TC I may play a role in the transport of cobalamin analogues (which it binds more effectively than IF) to the liver for excretion in bile.

The other major cobalamin transport protein in plasma is transcobalamin, also known as TC II. The gene is on chromosome 22q11-q13.1. As for IF and HC, there are nine exons. The three proteins are likely to have a common ancestral origin. TC II is synthesized by liver and by other tissues, including macrophages, ileum, and vascular endothelium. It normally carries only 20–60 ng of cobalamin per liter of plasma and readily gives up cobalamin to marrow, placenta, and other tissues, which it enters by receptor-mediated endocytosis involving the TC II receptor and megalin (encoded by the *LRP-2* gene). The TC II cobalamin is internalized by endocytosis via clathrin-coated pits; the complex is degraded, but the receptor probably is recycled to the cell membrane as is the case for transferrin. Export of “free” cobalamin is via the ATP-binding cassette drug transporter alias multidrug resistance protein 1.

FOLATE

DIETARY FOLATE

Folic (pteroylglutamic) acid is a yellow, crystalline, water-soluble substance. It is the parent compound of a large family of natural folate compounds, which differ from it in three respects: (1) they are partly or completely reduced to dihydrofolate (DHF) or tetrahydrofolate (THF) derivatives, (2) they usually contain a single carbon unit **(Table 95-2)**, and (3) 70–90% of natural folates are folate-polyglutamates.

TABLE 95-2

Biochemical Reactions of Folate Coenzymes

Reaction	Coenzyme Form of Folate Involved	Single Carbon Unit Transferred	Importance
Formate activation	THF	−CHO	Generation of 10-formyl-THF
Purine synthesis			
Formation of glycinamide ribonucleotide	5,10-Methylene-THF	−CHO	Formation of purines needed for DNA, RNA synthesis, but reactions probably not rate-limiting
Formylation of aminoimidazole carboxamide ribonucleotide (AICAR)	10-Formyl (CHO)THF		
Pyrimidine synthesis			
Methylation of deoxyuridine monophosphate (dUMP) to thymidine monophosphate (dTMP)	5,10-Methylene-THF	$-CH_3$	Rate limiting in DNA synthesis Oxidizes THF to DHF Some breakdown of folate at the C-9–N-10 bond
Amino acid interconversion			
Serine–glycine interconversion	THF	$=CH_2$	Entry of single carbon units into active pool
Homocysteine to methionine	5-Methyl(M)THF	$-CH_3$	Demethylation of 5-MTHF to THF; also requires cobalamin, flavine adenine dinucleotide, ATP, and adenosylmethionine
Forminoglutamic acid to glutamic acid in histidine catabolism	THF	−HN−CH=	

Abbreviations: DHF, dihydrofolate; THF, tetrahydrofolate.

Most foods contain some folate. The highest concentrations are found in liver, yeast, spinach, other greens, and nuts (>100 µg/100 g). The total folate content of an average Western diet is ~250 µg daily, but the amount varies widely according to the type of food eaten and the method of cooking. Folate is easily destroyed by heating, particularly in large volumes of water. Total body folate in the adult is ~10 mg, with the liver containing the largest store. Daily adult requirements are ~100 µg, and so stores are sufficient for only 3–4 months in normal adults and severe folate deficiency may develop rapidly.

ABSORPTION

Folates are absorbed rapidly from the upper small intestine. The absorption of folate polyglutamates is less efficient than that of monoglutamates; on average, ~50% of food folate is absorbed. Polyglutamate forms are hydrolyzed to the monoglutamate derivatives either in the lumen of the intestine or within the mucosa. All dietary folates are converted to 5-methylTHF (5-MTHF) within the small intestinal mucosa before entering portal plasma. The monoglutamates are actively transported across the enterocyte by a proton-coupled folate transporter (PCFT, SCL46A1). This is situated at the apical brush border and is most active at pH 5.5, which is about the pH of the duodenal and jejunal surface. Genetic mutations of this protein underlie hereditary malabsorption of folate (see below). Pteroylglutamic acid at doses >400 µg is absorbed largely unchanged and converted to natural folates in the liver. Lower doses are converted to 5-MTHF during absorption through the intestine.

About 60–90 µg of folate enters the bile each day and is excreted into the small intestine. Loss of this folate, together with the folate of sloughed intestinal cells, accelerates the speed with which folate deficiency develops in malabsorption conditions.

TRANSPORT

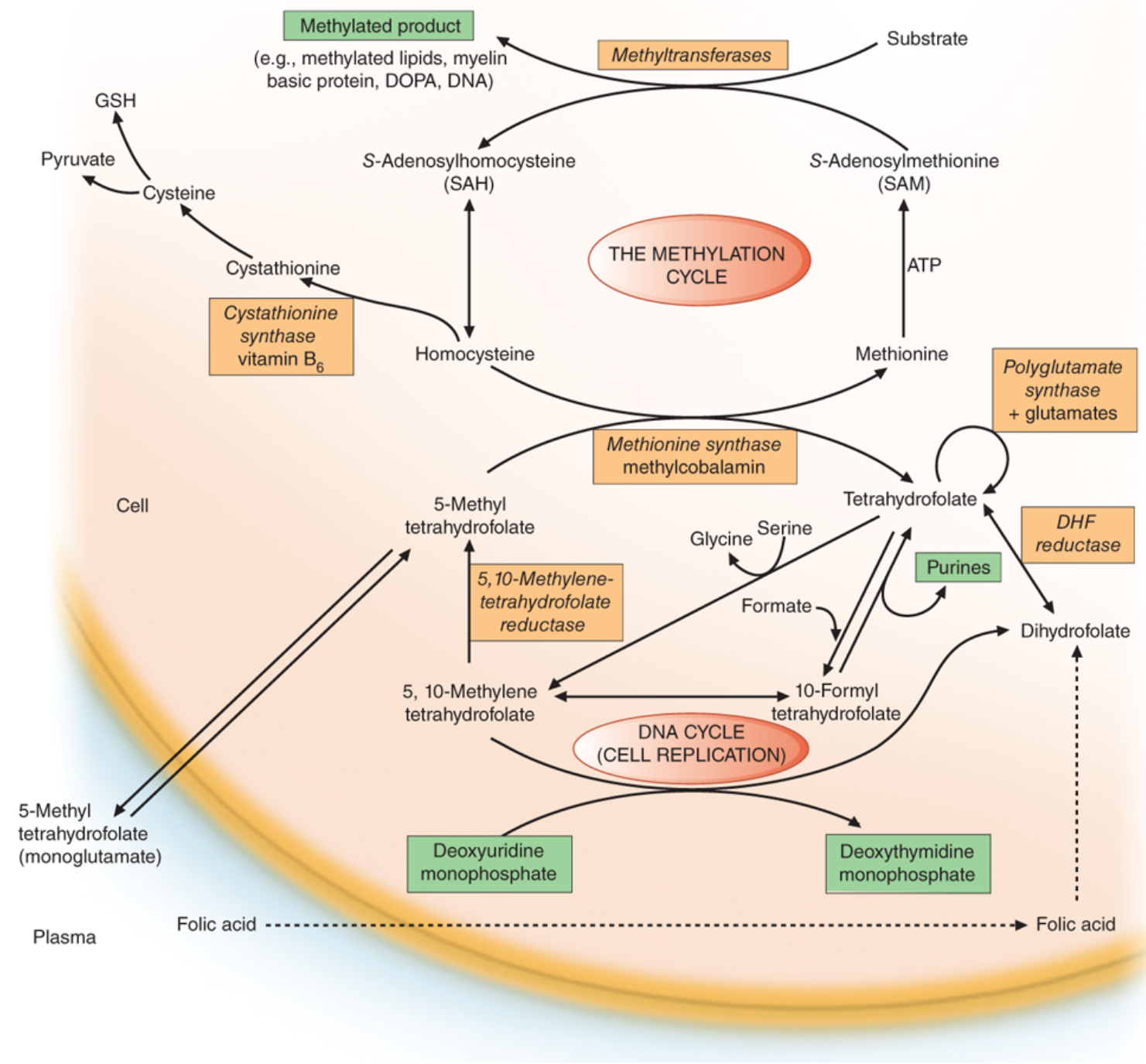
Folate is transported in plasma; about one-third is loosely bound to albumin, and two-thirds is unbound. In all body fluids (plasma, cerebrospinal fluid, milk, bile), folate is largely, if not entirely, 5-MTHF in the monoglutamate form. Three types of folate-binding protein are involved. A reduced folate transporter (RFC, SLC19A1) is the major route of delivery of plasma folate (5-MTHF) to cells. Two folate receptors, FR2 and FR3 embedded in the cell membrane by a glycosyl phosphatidylinositol anchor, transport folate into the cell via receptor-mediated endocytosis. The third protein, PCFT, transports folate at low pH from the vesicle to the cell cytoplasm. The reduced folate transporter also mediates uptake of methotrexate by cells.

BIOCHEMICAL FUNCTIONS

Folates (as the intracellular polyglutamate derivatives) act as coenzymes in the transfer of single-carbon units (**Fig. 95-1** and Table 95-2). Two of these reactions are involved in purine synthesis and one in pyrimidine synthesis necessary for DNA and RNA replication. Folate is also a coenzyme for methionine synthesis, in which methylcobalamin is also involved and in which THF is regenerated. THF is the acceptor of single carbon units newly entering the active pool via conversion of serine to glycine. Methionine, the other product of the methionine synthase reaction, is the precursor for *S*-adenosylmethionine (SAM), the universal methyl donor involved in >100 methyltransferase reactions (Fig. 95-1).

FIGURE 95-1

The role of folates in DNA synthesis and in formation of *S*-adenosylmethionine (SAM), which is involved in numerous methylation reactions. DHF, dihydrofolate; GSH, glutathione. *(Reprinted from AV Hoffbrand et al [eds]: Postgraduate Haematology, 5th ed. Oxford, UK, Blackwell Publishing, 2005; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

During thymidylate synthesis, 5,10-methylene-THF is oxidized to DHF. The enzyme DHF reductase converts this to THF. The drugs methotrexate, pyrimethamine, and (mainly in bacteria) trimethoprim inhibit DHF reductase and so prevent formation of active THF coenzymes from DHF. A small fraction of the folate coenzyme is not recycled during thymidylate synthesis but is degraded at the C9-N10 bond.

BIOCHEMICAL BASIS OF MEGALOBLASTIC ANEMIA

The common feature of all megaloblastic anemias is a defect in DNA synthesis that affects rapidly dividing cells in the bone marrow. All conditions that give rise to megaloblastic changes have in common a disparity in the rate of synthesis or availability of the four immediate precursors of DNA: the deoxyribonucleoside triphosphates (dNTPs)—dA(adenine)TP and dG(guanine)TP (purines), dT(thymine)TP, and dC(cytosine)TP (pyrimidines). In deficiencies of either folate or cobalamin, there is failure to convert deoxyuridine monophosphate (dUMP) to deoxythymidine monophosphate (dTMP), the precursor of dTTP (Fig. 95-1). This is the case because folate is needed as the coenzyme 5,10-methylene-THF polyglutamate for conversion of dUMP to dTMP; the availability of 5,10-methylene-THF is reduced in either cobalamin or folate deficiency. DNA replication from multiple origins along the chromosome is slower than normal during mitosis, and there is failure of joining up

the incomplete replicons with resulting single stranded DNA breaks. An alternative theory for megaloblastic anemia in cobalamin or folate deficiency is misincorporation of uracil into DNA because of the accumulation of deoxyuridine triphosphate (dUTP) at the DNA replication fork as a consequence of the block in conversion of dUMP to dTMP.

COBALAMIN-FOLATE RELATIONS

Folate is required for many reactions in mammalian tissues. Only two reactions in the body are known to require cobalamin. Methylmalonyl-CoA isomerization requires adocobalamin, and the methylation of homocysteine to methionine requires both methylcobalamin and 5-MTHF (Fig. 95-1). This reaction is the first step in the pathway by which 5-MTHF, which enters bone marrow and other cells from plasma, is converted into all the intracellular folate coenzymes. The coenzymes are all polyglutamated (the larger size aiding retention in the cell), but the enzyme folate polyglutamate synthase can use only THF, not MTHF, as substrate. In cobalamin deficiency, MTHF accumulates in plasma, and intracellular folate concentrations fall due to failure of formation of THF, the substrate on which folate polyglutamates are built. This has been termed *THF starvation*, or the *methylfolate trap*.

This theory explains the abnormalities of folate metabolism that occur in cobalamin deficiency (high serum folate, low cell folate, positive purine precursor aminoimidazole carboxamide ribonucleotide [AICAR] excretion; Table 95-2) and also why the anemia of cobalamin deficiency responds to folic acid in large doses.

CLINICAL FEATURES

Many symptomless patients are detected through the finding of a raised mean corpuscular volume (MCV) on a routine blood count. The main clinical features in more severe cases are those of anemia. Anorexia is usually marked, and there may be weight loss, diarrhea, or constipation. Glossitis, angular cheilosis, a mild fever in more severely anemic patients, jaundice (unconjugated), and reversible melanin skin hyperpigmentation also may occur with a deficiency of either folate or cobalamin. Thrombocytopenia sometimes leads to bruising, and this may be aggravated by vitamin C deficiency or alcohol in malnourished patients. The anemia and low leukocyte count may predispose to infections, particularly of the respiratory and urinary tracts. Cobalamin deficiency has also been associated in a few studies with impaired bactericidal function of phagocytes and with osteoporosis.

Neurologic Manifestations

Vitamin B_{12} is needed for the myelination of the central nervous system. Its deficiency may cause a bilateral peripheral neuropathy or degeneration (demyelination) of the cervical and thoracic posterior and lateral (pyramidal) tracts of the spinal cord and, less frequently, of the cranial nerves and of the white matter of the brain. Optic atrophy and cerebral symptoms including dementia, depression, psychotic symptoms, and cognitive impairment may be prominent. There may also be anosmia and loss of taste. MRI may show the “spongy” degeneration of the cord.

The patient, more frequently male, typically presents with paresthesias, muscle weakness, or difficulty in walking but sometimes with the dementia, psychotic disturbances, or visual impairment. There is usually loss of proprioception and vibration sensation with positive Romberg and Lhermitte signs. Gait may be ataxic with spasticity (hyperreflexia). Autonomic nervous dysfunction can result in postural hypotension, impotence, and incontinence.

Long-term nutritional cobalamin deficiency in infancy leads to poor brain development and impaired intellectual development. In infancy there may be feeding difficulties, lethargy, and coma. Convulsions and myoclonus have been described. An important clinical problem is the non-anemic patient with neurologic or psychiatric abnormalities and a low or borderline serum cobalamin level. In such patients, it is necessary to try to establish whether there is significant cobalamin deficiency, for example, by careful examination of the blood film, tests for pernicious anemia (PA) by serum gastrin level and for antibodies to IF or parietal cells, along with serum methylmalonic acid (MMA) measurement if available. A trial of cobalamin therapy for at least 3 months will usually also be needed to determine whether the symptoms improve.

The biochemical basis for cobalamin neuropathy remains obscure. Its occurrence in the absence of methylmalonic aciduria in TC II deficiency suggests that the neuropathy is related to the defect in homocysteine-methionine conversion. Accumulation of *S*-adenosylhomocysteine in the brain, resulting in inhibition of transmethylation reactions, has been suggested. Folate deficiency has been suggested to cause organic nervous disease, but this is uncertain, although methotrexate injected into the cerebrospinal fluid may cause brain or spinal cord damage.

Psychiatric disturbance as discussed above is common in both folate and cobalamin deficiencies. This, like the neuropathy, has been attributed to a failure of the synthesis of SAM, which is needed in methylation of biogenic amines (e.g., dopamine) as well as that of proteins, phospholipids, and neurotransmitters in the brain (Fig. 95-1). Associations between lower serum folate or cobalamin levels and higher homocysteine levels and the development of decreased cognitive function and dementia in Alzheimer's disease have been reported. A meta-analysis of randomized, placebo-controlled trials of homocysteine-lowering B-vitamin supplementation of individuals with and without cognitive impairment, however, showed that supplementation with vitamin B_{12}, vitamin B_6, and folic acid alone or in combination did not improve cognitive function. It is unknown whether prolonged treatment with these B vitamins can reduce the risk of dementia in later life.

GENERAL TISSUE EFFECTS OF COBALAMIN AND FOLATE DEFICIENCIES

Epithelial Surfaces

After the marrow, the next most frequently affected tissues are the epithelial cell surfaces of the mouth (with glossitis), stomach, and small intestine and the respiratory, urinary, and female genital tracts. The cells show macrocytosis, with increased numbers of multinucleate and dying cells. The deficiencies may cause cervical smear abnormalities.

Complications of Pregnancy

The gonads are also affected, and infertility is common in both men and women with severe deficiency of either vitamin. Maternal folate deficiency has been implicated as a cause of prematurity, and both folate deficiency and cobalamin deficiency have been implicated in recurrent fetal loss and neural tube defects, as discussed below.

Neural Tube Defects

Folic acid supplements at the time of conception and in the first 12 weeks of pregnancy reduce by ~70% the incidence of neural tube defects (NTDs) (anencephaly, meningomyelocele, encephalocele, and spina bifida) in the fetus. Most of this protective effect can be achieved by taking folic acid, 0.4 mg daily, at the time of conception.

The incidence of cleft palate and harelip also can be reduced by prophylactic folic acid. There is no clear simple relationship between maternal folate status and these fetal abnormalities, although overall the lower the maternal folate, the greater the risk to the fetus. NTDs also can be caused by antifolate and antiepileptic drugs.

An underlying maternal folate metabolic abnormality has also been postulated. One abnormality has been identified: reduced activity of the enzyme 5,10-methylene-THF reductase (MTHFR) (Fig. 95-1) caused by a common C677T polymorphism in the *MTHFR* gene. In one study, the prevalence of this polymorphism was found to be higher than in controls in the parents of NTD fetuses and in the fetuses themselves: homozygosity for the TT mutation was found in 13% of cases compared with 5% of control subjects. The polymorphism codes for a thermolabile form of MTHFR. The homozygous state results in a lower mean serum and red cell folate level compared with control subjects, as well as significantly higher serum homocysteine levels. Tests for mutations in other enzymes possibly associated with NTDs, for example, methionine synthase and serine–glycine hydroxymethylase, have been negative. Serum vitamin B_{12} levels are also lower in the sera of mothers of NTD infants than in controls. In addition, maternal TC II receptor polymorphisms are associated with increased risk of NTD births. There are, however, no studies showing dietary fortification with vitamin B_{12} reduces the incidence of NTDs.

Cardiovascular Disease

Children with severe homocystinuria (blood levels ≥100 µmol/L) due to deficiency of one of three enzymes, methionine synthase, MTHFR, or cystathionine synthase (Fig. 95-1), have vascular disease, for example, ischemic heart disease, cerebrovascular disease, or pulmonary embolus, as teenagers or in young adulthood. Lesser degrees of raised serum homocysteine and low levels of serum folate and homozygous inherited mutations of MTHFR have been found to be associated with cerebrovascular, peripheral vascular, and coronary heart disease and with deep vein thrombosis. Prospective randomized trials of lowering homocysteine levels with supplements of folic acid, vitamin B_{12}, and vitamin B_6 against placebo over a 5-year period in patients with vascular disease or diabetes have not, however, shown a reduction of first event fatal or nonfatal myocardial infarction, nor have these supplements reduced the risk of recurrent cardiovascular disease after an acute myocardial infarct. Meta-analysis showed an 18% reduction in strokes. The benefit for stroke prevention has been confirmed by a large (>20,000 subjects) randomized prospective study in hypertensive subjects in China. This showed a significant reduction in the first incidence of stroke in subjects receiving enalapril and folic acid compared to enalapril alone. This was especially marked in the subjects commencing the prospective trial

with the lowest serum folate levels. Venous thrombosis has been reported to be more frequent in folate– or vitamin B_{12}– deficient subjects than in controls and to occur at unusual sites such as cerebral venous sinuses. This was ascribed to raised plasma homocysteine levels in folate or vitamin B_{12} deficiency.

Malignancy

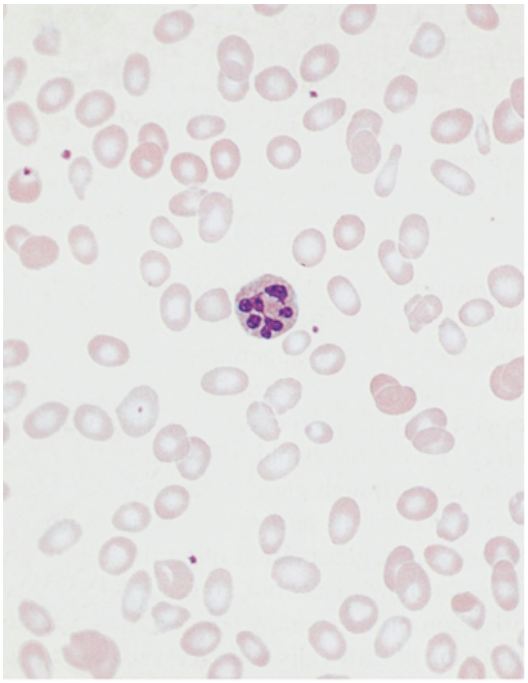
Prophylactic folic acid in pregnancy has been found in some but not all studies to reduce the subsequent incidence of acute lymphoblastic leukemia (ALL) in childhood. A significant negative association has also been found with the *MTHFR* C677T polymorphism and leukemias with mixed lineage leukemia (MLL) translocations, but a positive association with hyperdiploidy in infants with ALL or acute myeloid leukemia or with childhood ALL. A second polymorphism in the *MTHFR* gene, A1298C, is also strongly associated with hyperdiploid leukemia. There are various positive and negative associations between polymorphisms in folate-dependent enzymes and the incidence of adult ALL. The C677T polymorphism is thought to lead to increased thymidine pools and "better quality" of DNA synthesis by shunting one-carbon groups toward thymidine and purine synthesis. This may explain its reported association with a lower risk for colorectal cancer. Most but not all studies suggest that prophylactic folic acid also protects against colon adenomas. Other tumors that have been associated with folate polymorphisms or status include follicular lymphoma, breast cancer, and gastric cancer. A meta-analysis of 50,000 individuals given folic acid (0.5–40 mg daily) or placebo in cardiovascular or colon adenoma prevention trials found that folic acid supplementation did not significantly increase or decrease the overall incidence of cancer or of any site-specific cancer during a weighted average scheduled treatment duration of 5.7 years. Because folic acid may "feed" tumors, it probably should be avoided in those with established tumors unless there is severe megaloblastic anemia due to folate deficiency.

HEMATOLOGIC FINDINGS

PERIPHERAL BLOOD

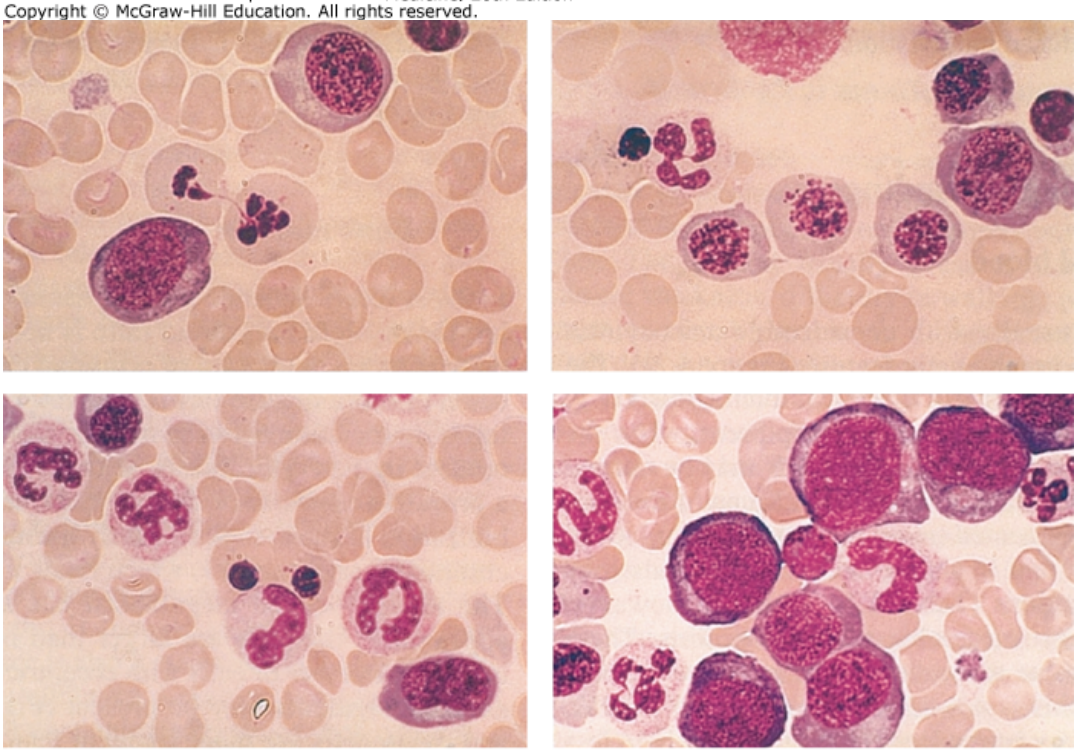
Oval macrocytes, usually with considerable anisocytosis and poikilocytosis, are the main feature **(Fig. 95-2A)**. The MCV is usually >100 fL unless a cause of microcytosis (e.g., iron deficiency or thalassemia trait) is present. Some of the neutrophils are hypersegmented (more than five nuclear lobes). There may be leukopenia due to a reduction in granulocytes and lymphocytes, but this is usually $>1.5 \times 10^9$/L; the platelet count may be moderately reduced, rarely to $<40 \times 10^9$/L. The severity of all these changes parallels the degree of anemia. In a nonanemic patient, the presence of a few macrocytes and hypersegmented neutrophils in the peripheral blood may be the only indication of the underlying disorder.

FIGURE 95-2
A. The peripheral blood in severe megaloblastic anemia. ***B.*** The bone marrow in severe megaloblastic anemia. *(Reprinted from AV Hoffbrand et al [eds]: Postgraduate Haematology, 5th ed. Oxford, UK, Blackwell Publishing, 2005; with permission.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

BONE MARROW

In a severely anemic patient, the marrow is hypercellular with an accumulation of primitive cells due to selective death by apoptosis of more mature forms. The erythroblast nucleus maintains a primitive appearance despite maturation and hemoglobinization of the cytoplasm. The cells are larger than normoblasts, and an increased number of cells with eccentric lobulated nuclei or nuclear fragments may be present (Fig. 95-2B). Giant and abnormally shaped metamyelocytes and enlarged hyperpolyploid megakaryocytes are characteristic. In severe cases, the accumulation of primitive cells may mimic acute myeloid leukemia, whereas in less anemic patients, the changes in the marrow may be

difficult to recognize. The terms *intermediate*, *mild*, and *early* have been used. The term *megaloblastoid* does not mean mildly megaloblastic. It is used to describe cells with both immature-appearing nuclei and defective hemoglobinization and is usually seen in myelodysplasia.

CHROMOSOMES

Bone marrow cells, transformed lymphocytes, and other proliferating cells in the body show a variety of changes, including random breaks, reduced contraction, spreading of the centromere, and exaggeration of secondary chromosomal constrictions and overprominent satellites. Similar abnormalities may be produced by antimetabolite drugs (e.g., cytosine arabinoside, hydroxyurea, and methotrexate) that interfere with either DNA replication or folate metabolism and that also cause megaloblastic appearances.

INEFFECTIVE HEMATOPOIESIS

There is an accumulation of unconjugated bilirubin in plasma due to the death of nucleated red cells in the marrow (ineffective erythropoiesis). Other evidence for this includes raised urine urobilinogen, reduced haptoglobins and positive urine hemosiderin, and a raised serum lactate dehydrogenase. A weakly positive direct antiglobulin test due to complement can lead to a false diagnosis of autoimmune hemolytic anemia.

CAUSES OF COBALAMIN DEFICIENCY

Cobalamin deficiency is usually due to malabsorption. The only other cause is inadequate dietary intake.

INADEQUATE DIETARY INTAKE

Adults

Dietary cobalamin deficiency arises in vegans who omit meat, fish, eggs, cheese, and other animal products from their diet. The largest group in the world consists of Hindus, and it is likely that many millions of Indians are at risk of deficiency of cobalamin on a nutritional basis. Subnormal serum cobalamin levels are found in up to 50% of randomly selected, young, adult Indian vegans, but the deficiency usually does not progress to megaloblastic anemia since the diet of most vegans is not totally lacking in cobalamin and the enterohepatic circulation of cobalamin is intact. Dietary cobalamin deficiency may also arise rarely in non-vegetarian individuals who exist on grossly inadequate diets because of poverty or psychiatric disturbance.

Infants

Cobalamin deficiency has been described in infants born to severely cobalamin-deficient mothers. These infants develop megaloblastic anemia at about 3–6 months of age, presumably because they are born with low stores of cobalamin and because they are fed breast milk with low cobalamin content. The babies have also shown growth retardation, impaired psychomotor development, and other neurologic sequelae. MRI shows delayed myelination and atrophy.

GASTRIC CAUSES OF COBALAMIN MALABSORPTION

See Tables 95-3 and 95-4.

TABLE 95-3

Causes of Cobalamin Deficiency Sufficiently Severe to Cause Megaloblastic Anemia

Nutritional	**Vegans**
Malabsorption	Pernicious anemia
Gastric causes	Congenital absence of intrinsic factor or functional abnormality
	Total or partial gastrectomy
Intestinal causes	Intestinal stagnant loop syndrome: jejunal diverticulosis, ileocolic fistula, anatomic blind loop, intestinal stricture, etc.
	Ileal resection and Crohn’s disease
	Selective malabsorption with proteinuria
	Tropical sprue
	Transcobalamin II deficiency
	Fish tapeworm

TABLE 95-4

Malabsorption of Cobalamin May Occur in the Following Conditions but Is Not Usually Sufficiently Severe and Prolonged to Cause Megaloblastic Anemia

Gastric causes
Simple atrophic gastritis (food cobalamin malabsorption)
Zollinger-Ellison syndrome
Gastric bypass or bariatric surgery
Use of proton pump inhibitors
Intestinal causes
Gluten-induced enteropathy
Severe pancreatitis
HIV infection
Radiotherapy
Graft-versus-host disease
Deficiencies of cobalamin, folate, protein, ?riboflavin, ?nicotinic acid
Therapy with colchicine, para-aminosalicylate, neomycin, slow-release potassium chloride, anticonvulsant drugs,?metformin*, cytotoxic drugs
Alcohol

*It is now thought that metformin lowers serum vitamin B_{12} level by lowering the level of TC I.

Formerly, the pathogenesis of B_{12} malabsorption was distinguishable based on the results of a Schilling test in which a radioactive form of B_{12} was administered orally and its appearance in the urine was a sign of absorption. Radioactive B_{12} is no longer available, and Schilling tests are no longer performed. Other approaches to the differential diagnosis of B_{12} malabsorption are now employed.

Pernicious Anemia

PA may be defined as a severe lack of IF due to gastric atrophy. It is a common disease in northern Europeans but occurs in all countries and ethnic groups. It is more frequent in people of African than Asian ancestry. The overall incidence is about 120 per 100,000 population in the United Kingdom (UK). The ratio of incidence in men and women among whites is ~1:1.6, and the median age of onset is 70–80 years, with only 10% of patients being <40 years of age. However, in some ethnic groups, notably black individuals and Latin Americans, the age at onset of PA is generally lower. The disease occurs more commonly than by chance in close relatives and in persons with other organ-specific autoimmune diseases, for example, thyroid diseases, vitiligo, hypoparathyroidism, Type 1 diabetes, and Addison's disease. It is also associated with hypogammaglobulinemia, with premature graying or blue eyes, and persons of blood group A. An association with human leukocyte antigen (HLA) 3 has been reported in some but not all series and, in those with endocrine disease, with HLA-B8, -B_{12}, and -BW15. Life expectancy is normal in women once regular treatment has begun. Men had a slightly subnormal life expectancy as a result of a higher incidence of carcinoma of the stomach than in control subjects, but current data on their life expectancy are unavailable. Gastric output of hydrochloric acid, pepsin, and IF is severely reduced. The serum gastrin level is raised, and serum pepsinogen I levels are low.

Gastric Biopsy

A single endoscopic examination is recommended if PA is diagnosed. Gastric biopsy usually shows atrophy of all layers of the body and fundus, with loss of glandular elements, an absence of parietal and chief cells and replacement by mucous cells, a mixed inflammatory cell infiltrate, and perhaps intestinal metaplasia. The infiltrate of plasma cells and lymphocytes contains an excess of CD4 cells. These are directed against gastric H/K-ATPase. The antral mucosa is usually well preserved. *Helicobacter pylori* infection occurs infrequently in PA, but it has been suggested that *H. pylori* gastritis occurs at an early phase of atrophic gastritis and presents in younger patients as iron-deficiency anemia but in older patients as PA. *H. pylori* is suggested to stimulate an autoimmune process directed against parietal cells, with the *H. pylori* infection then being gradually replaced, in some individuals, by an autoimmune process.

Serum Antibodies

Two types of IF immunoglobulin G antibody may be found in the sera of patients with PA. One, the “blocking,” or type I, antibody, prevents the combination of IF and cobalamin, whereas the “binding,” or type II, antibody prevents attachment of IF to ileal mucosa. Type I occurs in the sera of ~55% of patients, and type II in 35%. IF antibodies cross the placenta and may cause temporary IF deficiency in a newborn infant. Patients with PA also show cell-mediated immunity to IF. Type I antibody has been detected rarely in the sera of patients without PA but with thyrotoxicosis, myxedema, Hashimoto’s disease, or diabetes mellitus and in relatives of PA patients. IF antibodies also have been detected in gastric juice in ~80% of PA patients. These gastric antibodies may reduce absorption of dietary cobalamin by combining with small amounts of remaining IF.

Parietal cell antibody is present in the sera of almost 90% of adult patients with PA but is frequently present in other subjects. Thus, it occurs in as many as 16% of randomly selected female subjects age >60 years. The parietal cell antibody is directed against the α and β subunits of the gastric proton pump (H^+, K^+-ATPase).

JUVENILE PERNICIOUS ANEMIA

This usually occurs in older children and resembles PA of adults. Gastric atrophy, achlorhydria, and serum IF antibodies are all present, although parietal cell antibodies are usually absent. About one-half of these patients show an associated endocrinopathy such as autoimmune thyroiditis, Addison’s disease, or hypoparathyroidism; in some, mucocutaneous candidiasis occurs.

CONGENITAL INTRINSIC FACTOR DEFICIENCY OR FUNCTIONAL ABNORMALITY

An affected child usually presents with megaloblastic anemia in the first to third year of life; a few have presented as late as the second decade. The child usually has no demonstrable IF but has a normal gastric mucosa and normal secretion of acid. The inheritance is autosomal recessive. Parietal cell and IF antibodies are absent. Variants have been described in which the child is born with IF that can be detected immunologically but is unstable or functionally inactive, unable to bind cobalamin or to facilitate its uptake by ileal receptors.

GASTRECTOMY

After total gastrectomy, cobalamin deficiency is inevitable, and prophylactic cobalamin therapy should be commenced immediately after the operation. After partial gastrectomy, 10–15% of patients also develop this deficiency. The exact incidence and time of onset are most influenced by the size of the resection and the preexisting size of cobalamin body stores.

FOOD COBALAMIN MALABSORPTION

Failure of release of cobalamin from binding proteins in food is believed to be responsible for this condition, which is more common in the elderly. It is associated with low serum cobalamin levels, with or without raised serum levels of MMA and homocysteine. Typically, these patients have normal cobalamin absorption, as measured with crystalline cobalamin, but show malabsorption when a modified test using food-bound cobalamin is used. It is usually due to mild forms of atrophic gastritis or therapy with proton pump inhibitors. Bariatric surgery is likely to be an increasing cause of this form of B_{12} malabsorption and deficiency. The frequency of progression to severe cobalamin deficiency and the reasons for this progression are not clear.

INTESTINAL CAUSES OF COBALAMIN MALABSORPTION

Intestinal Stagnant Loop Syndrome

Malabsorption of cobalamin occurs in a variety of intestinal lesions in which there is colonization of the upper small intestine by fecal organisms. This may occur in patients with jejunal diverticulosis, enteroanastomosis, or an intestinal stricture or fistula or with an anatomic

blind loop due to Crohn's disease, tuberculosis, or an operative procedure.

Ileal Resection

Removal of ≥1.2 m of terminal ileum causes malabsorption of cobalamin. In some patients after ileal resection, particularly if the ileocecal valve is incompetent, colonic bacteria may contribute further to the onset of cobalamin deficiency.

Selective Malabsorption of Cobalamin with Proteinuria (Imerslund's Syndrome; Imerslund-Gräsbeck Syndrome; Congenital Cobalamin Malabsorption; Autosomal Recessive Megaloblastic Anemia; MGA1)

This autosomally recessive disease is the most common cause of megaloblastic anemia due to cobalamin deficiency in infancy in Western countries. More than 200 cases have been reported with familial clusters in Finland, Norway, the Middle East, and North Africa. The patients secrete normal amounts of IF and gastric acid but are unable to absorb cobalamin. In Finland, impaired synthesis, processing, or ligand binding of cubilin due to inherited mutations is found. In Norway, mutation of the gene for *AMN* has been reported. Other tests of intestinal absorption are normal. Over 90% of these patients show nonspecific proteinuria, but renal function is otherwise normal, and renal biopsy has not shown any consistent renal defect. A few have shown aminoaciduria and congenital renal abnormalities, such as duplication of the renal pelvis.

Tropical Sprue

Nearly all patients with acute and subacute tropical sprue show malabsorption of cobalamin; this may persist as the principal abnormality in the chronic form of the disease, when the patient may present with megaloblastic anemia or neuropathy due to cobalamin deficiency. Absorption of cobalamin usually improves after antibiotic therapy and, in the early stages, folic acid therapy.

Fish Tapeworm Infestation

The fish tapeworm (*Diphyllobothrium latum*) lives in the small intestine of humans and accumulates cobalamin from food, rendering the cobalamin unavailable for absorption. Individuals acquire the worm by eating raw or partly cooked fish. Infestation is common around the lakes of Scandinavia, Germany, Japan, North America, and Russia. Megaloblastic anemia or cobalamin neuropathy occurs only in those with a heavy infestation.

Gluten-Induced Enteropathy

Malabsorption of cobalamin occurs in ~30% of untreated patients (presumably those in whom the disease extends to the ileum). Cobalamin deficiency is not severe in these patients and is corrected with a gluten-free diet.

Severe Chronic Pancreatitis

In this condition, lack of trypsin is thought to cause dietary cobalamin attached to gastric non-IF (R) binder to be unavailable for absorption. It also has been proposed that in pancreatitis, the concentration of calcium ions in the ileum falls below the level needed to maintain normal cobalamin absorption.

HIV Infection

Serum cobalamin levels tend to fall in patients with HIV infection and are subnormal in 10–35% of those with AIDS. Malabsorption of cobalamin not corrected by IF has been shown in some, but not all, patients with subnormal serum cobalamin levels. Cobalamin deficiency sufficiently severe to cause megaloblastic anemia or neuropathy is rare.

Zollinger-Ellison Syndrome

Malabsorption of cobalamin has been reported in the Zollinger-Ellison syndrome. It is thought that there is a failure to release cobalamin from R-binding protein due to inactivation of pancreatic trypsin by high acidity, as well as interference with IF binding of cobalamin.

Radiotherapy

Both total-body irradiation and local radiotherapy to the ileum (e.g., as a complication of radiotherapy for carcinoma of the cervix) may cause malabsorption of cobalamin.

Graft-versus-Host Disease

This commonly affects the small intestine. Malabsorption of cobalamin due to abnormal gut flora, as well as damage to ileal mucosa, is common.

Drugs

The drugs that have been reported to cause malabsorption of cobalamin are listed in Table 95-4. However, megaloblastic anemia due to these drugs is rare. It has been suggested that metformin lowers serum B_{12} by lowering TC I level rather than causing malabsorption of B_{12}.

ABNORMALITIES OF COBALAMIN METABOLISM

Congenital Transcobalamin II Deficiency or Abnormality

Infants with TC II deficiency usually present with megaloblastic anemia within a few weeks of birth. Serum cobalamin and folate levels are normal, but the anemia responds to massive (e.g., 1 mg three times weekly) injections of cobalamin. Some cases show neurologic complications. The protein may be present but functionally inert. Genetic abnormalities found include mutations of an intra-exonic cryptic splice site, extensive deletion, single nucleotide deletion, nonsense mutation, and an RNA editing defect. Malabsorption of cobalamin occurs in all cases, and serum immunoglobulins are usually reduced. Failure to institute adequate cobalamin therapy or treatment with folic acid may lead to neurologic damage.

Congenital Methylmalonic Acidemia and Aciduria

Infants with this abnormality are ill from birth with vomiting, failure to thrive, severe metabolic acidosis, ketosis, and mental retardation. Anemia, if present, is normocytic and normoblastic. The condition may be due to a functional defect in either mitochondrial methylmalonyl-CoA mutase or its cofactor adocobalamin. Mutations in the methylmalonyl- CoA mutase are not responsive, or only poorly responsive, to treatment with cobalamin. A proportion of infants with failure of adocobalamin synthesis respond to cobalamin in large doses. Some children have combined methylmalonic aciduria and homocystinuria due to defective formation of both cobalamin coenzymes. This usually presents in the first year of life with feeding difficulties, developmental delay, microcephaly, seizures, hypotonia, and megaloblastic anemia.

Acquired Abnormality of Cobalamin Metabolism: Nitrous Oxide Inhalation

Nitrous oxide (N_2O) irreversibly oxidizes methylcobalamin to an inactive precursor; this inactivates methionine synthase. Megaloblastic anemia has occurred in patients undergoing prolonged N_2O anesthesia (e.g., in intensive care units). A neuropathy resembling cobalamin neuropathy has been described in dentists and anesthetists who are exposed repeatedly to N_2O. Methylmalonic aciduria does not occur as adocobalamin is not inactivated by N_2O.

CAUSES OF FOLATE DEFICIENCY

(Table 95-5)

TABLE 95-5

Causes of Folate Deficiency

Dietary[a]
Particularly in: old age, infancy, poverty, alcoholism, chronic invalids, and the psychiatrically disturbed; may be associated with scurvy or kwashiorkor
Malabsorption
Major causes of deficiency
Tropical sprue, gluten-induced enteropathy in children and adults, and in association with dermatitis herpetiformis, specific malabsorption of folate, intestinal megaloblastosis caused by severe cobalamin or folate deficiency
Minor causes of deficiency
Extensive jejunal resection, Crohn's disease, partial gastrectomy, congestive heart failure, Whipple's disease, scleroderma, amyloid, diabetic enteropathy, systemic bacterial infection, lymphoma, sulfasalazine (Salazopyrin)
Excess utilization or loss
Physiologic
Pregnancy and lactation, prematurity
Pathologic
Hematologic diseases: chronic hemolytic anemias, sickle cell anemia, thalassemia major, myelofibrosis
Malignant diseases: carcinoma, lymphoma, leukemia, myeloma
Inflammatory diseases: tuberculosis, Crohn's disease, psoriasis, exfoliative dermatitis, malaria
Metabolic disease: homocystinuria
Excess urinary loss: congestive heart failure, active liver disease
Hemodialysis, peritoneal dialysis
Antifolate drugs[b]
Anticonvulsant drugs (phenytoin, primidone, barbiturates), sulfasalazine
Nitrofurantoin, tetracycline, antituberculosis (less well documented)
Mixed causes
Liver diseases, alcoholism, intensive care units

[a]In severely folate-deficient patients with causes other than those listed under Dietary, poor dietary intake is often present. [b]Drugs inhibiting dihydrofolate reductase are discussed in the text.

NUTRITIONAL

Dietary folate deficiency is common. Indeed, in most patients with folate deficiency a nutritional element is present. Certain individuals are particularly prone to have diets containing inadequate amounts of folate (Table 95-5). In the United States and other countries where fortification of the diet with folic acid has been adopted, the prevalence of folate deficiency has dropped dramatically and is now almost restricted to high-risk groups with increased folate needs. Nutritional folate deficiency occurs in kwashiorkor and scurvy and in infants with repeated infections or those who are fed solely on goats' milk, which has a low folate content.

MALABSORPTION

Malabsorption of dietary folate occurs in tropical sprue and in gluten-induced enteropathy. In the rare congenital recessive syndrome of selective malabsorption of folate due to mutation of the PCFT, there is an associated defect of folate transport into the cerebrospinal fluid, and these patients show megaloblastic anemia, which responds to physiologic doses of folic acid given parenterally but not orally. They also show mental retardation, convulsions, and other central nervous system abnormalities. Minor degrees of malabsorption may also occur after jejunal resection or partial gastrectomy, in Crohn's disease, and in systemic infections, but in these conditions, if severe deficiency occurs, it is usually largely due to poor nutrition. Malabsorption of folate has been described in patients receiving sulfasalazine (Salazopyrin), cholestyramine, and triamterene.

EXCESS UTILIZATION OR LOSS

Pregnancy

Folate requirements are increased by 200–300 µg to ~400 µg daily in a normal pregnancy, partly because of transfer of the vitamin to the fetus but mainly because of increased folate catabolism due to cleavage of folate coenzymes in rapidly proliferating tissues. Megaloblastic anemia due to this deficiency is prevented by prophylactic folic acid therapy. It occurred in 0.5% of pregnancies in the UK and other Western countries before prophylaxis with folic acid, but the incidence is much higher in countries where the general nutritional status is poor.

Prematurity

A newborn infant, whether full term or premature, has higher serum and red cell folate concentrations than does an adult. However, a newborn infant's demand for folate has been estimated to be up to 10 times that of adults on a weight basis, and the neonatal folate level falls rapidly to the lowest values at about 6 weeks of age. The falls are steepest and are liable to reach subnormal levels in premature babies, a number of whom develop megaloblastic anemia responsive to folic acid at about 4–6 weeks of age. This occurs particularly in the smallest babies (<1500 g birth weight) and those who have feeding difficulties or infections or have undergone multiple exchange transfusions. In these babies, prophylactic folic acid should be given.

Hematologic Disorders

Folate deficiency frequently occurs in chronic hemolytic anemia, particularly in sickle cell disease, autoimmune hemolytic anemia, and congenital spherocytosis. In these and other conditions of increased cell turnover (e.g., myelofibrosis, malignancies), folate deficiency arises because it is not completely reutilized after performing coenzyme functions.

Inflammatory Conditions

Chronic inflammatory diseases such as tuberculosis, rheumatoid arthritis, Crohn's disease, psoriasis, exfoliative dermatitis, bacterial endocarditis, and chronic bacterial infections cause deficiency by reducing the appetite and increasing the demand for folate. Systemic infections also may cause malabsorption of folate. Severe deficiency is virtually confined to the patients with the most active disease and the poorest diet.

Homocystinuria

This is a rare metabolic defect in the conversion of homocysteine to cystathionine. Folate deficiency occurring in most of these patients may be due to excessive utilization because of compensatory increased conversion of homocysteine to methionine.

Long-Term Dialysis

Because folate is only loosely bound to plasma proteins, it is easily removed from plasma by dialysis. In patients with anorexia, vomiting, infections, and hemolysis, folate stores are particularly likely to become depleted. Routine folate prophylaxis is now given.

Congestive Heart Failure, Liver Disease

Excess urinary folate losses of >100 µg per day may occur in some of these patients. The explanation appears to be release of folate from damaged liver cells.

ANTIFOLATE DRUGS

A large number of epileptics who are receiving long-term therapy with phenytoin or primidone, with or without barbiturates, develop low serum and red cell folate levels. The exact mechanism is unclear. Alcohol may also be a folate antagonist, as patients who are drinking spirits may develop megaloblastic anemia that will respond to normal quantities of dietary folate or to physiologic doses of folic acid only if alcohol is withdrawn. Macrocytosis of red cells is associated with chronic alcohol intake even when folate levels are normal. Inadequate folate intake is the major factor in the development of deficiency in spirit-drinking alcoholics. Beer is relatively folate-rich in some countries, depending on the technique used for brewing.

The drugs that inhibit DHF reductase include methotrexate, pyrimethamine, and trimethoprim. Methotrexate has the most powerful action against the human enzyme, whereas trimethoprim is most active against the bacterial enzyme and is likely to cause megaloblastic anemia only when used in conjunction with sulfamethoxazole in patients with preexisting folate or cobalamin deficiency. The activity of pyrimethamine is intermediate. The antidote to these drugs is folinic acid (5-formyl-THF).

CONGENITAL ABNORMALITIES OF FOLATE METABOLISM

Some infants with congenital defects of folate enzymes (e.g., cyclohydrolase or methionine synthase) have had megaloblastic anemia.

DIAGNOSIS OF COBALAMIN AND FOLATE DEFICIENCIES

The diagnosis of cobalamin or folate deficiency has traditionally depended on the recognition of the relevant abnormalities in the peripheral blood and analysis of the blood levels of the vitamins.

COBALAMIN DEFICIENCY

Serum Cobalamin

This is measured by an automated enzyme-linked immunosorbent assay (ELISA) or competitive-binding luminescence assay (CBLA). Normal serum levels range from 118–148 pmol/L (160–200 ng/L) to ~738 pmol/L (1000 ng/L). In patients with megaloblastic anemia due to cobalamin deficiency, the level is usually <74 pmol/L (100 ng/L). In general, the more severe the deficiency, the lower is the serum cobalamin level. In patients with spinal cord damage due to the deficiency, levels are very low even in the absence of anemia. Values between 74 and 148 pmol/L (100 and 200 ng/L) are regarded as borderline. They may occur, for instance, in pregnancy, in patients with megaloblastic anemia due to folate deficiency. They may also be due to heterozygous, homozygous, or compound heterozygous mutations of the gene *TCN1* that codes for HC (transcobalamin I). There is then no clinical or hematologic abnormality. The serum cobalamin level is sufficiently robust, cost-effective, and most convenient to rule out cobalamin deficiency in the vast majority of patients suspected of having this problem. However, problems have arisen with commercial CBLA assays involving IF in PA patients with intrinsic antibodies in serum. These antibodies may cause false normal serum vitamin B_{12} levels in up to 50% of cases tested. Where clinical indications of PA are strong, a normal serum vitamin B_{12} does not rule out the diagnosis. Serum MMA levels will be elevated in untreated PA (see below).

Folate deficiency, transcobalamin 1 (HC) deficiency, oral contraceptives, and multiple myeloma have all been associated with low serum B_{12} levels that do not indicate B_{12} deficiency. On the other hand, high serum B_{12} levels are usually due to raised serum transcobalamin 1 levels and can be due to the presence of liver, renal, or myeloproliferative diseases or to cancer of the breast, colon, or liver.

Serum Methylmalonate and Homocysteine

In patients with cobalamin deficiency sufficient to cause anemia or neuropathy, the serum MMA level is raised. Sensitive methods for measuring MMA and homocysteine in serum have been introduced and recommended for the early diagnosis of cobalamin deficiency, even in the absence

of hematologic abnormalities or subnormal levels of serum cobalamin. Serum MMA levels fluctuate, however, in patients with renal failure. Mildly elevated serum MMA and/or homocysteine levels occur in up to 30% of apparently healthy volunteers, with serum cobalamin levels up to 258 pmol/L (350 ng/L) and normal serum folate levels; 15% of elderly subjects, even with cobalamin levels >258 pmol/L (>350 ng/L), have this pattern of raised metabolite levels. These findings bring into question the exact cutoff points for normal MMA and homocysteine levels. It is also unclear at present whether these mildly raised metabolite levels have clinical consequences.

Serum homocysteine is raised in both early cobalamin and folate deficiency but may be raised in other conditions, for example, chronic renal disease, alcoholism, smoking, pyridoxine deficiency, hypothyroidism, and therapy with steroids, cyclosporine, and other drugs. Levels are also higher in serum than in plasma, in men than in premenopausal women, in women taking hormone replacement therapy or in oral contraceptive users, and in elderly persons and patients with several inborn errors of metabolism affecting enzymes in trans-sulfuration pathways of homocysteine metabolism. Thus, homocysteine levels must be carefully interpreted for diagnosis of cobalamin or folate deficiency.

Tests for the Cause of Cobalamin Deficiency

Only vegans, strict vegetarians, or people living on a totally inadequate diet will become vitamin B_{12} deficient because of inadequate intake.

Studies of cobalamin absorption once were widely used, but difficulty in obtaining radioactive cobalamin and ensuring that IF preparations are free of viruses has made these tests obsolete. Tests to diagnose PA include serum gastrin, which is raised; serum pepsinogen I, which is low in PA (90–92%) but also in other conditions; and gastric endoscopy. Tests for IF and parietal cell antibodies are also used, as well as tests for individual intestinal diseases.

FOLATE DEFICIENCY

Serum Folate

This is also measured by an ELISA technique. In most laboratories, the normal range is from 11 nmol/L (2 µg/L) to ~82 nmol/L (15 µg/L). The serum folate level is low in all folate-deficient patients. It also reflects recent diet. Because of this, serum folate may be low before there is hematologic or biochemical evidence of deficiency. Serum folate rises in severe cobalamin deficiency because of the block in conversion of MTHF to THF inside cells; raised levels have also been reported in the intestinal stagnant loop syndrome due to absorption of bacterially synthesized folate.

Red Cell Folate

The red cell folate assay is a valuable test of body folate stores. It is less affected than the serum assay by recent diet and traces of hemolysis. In normal adults, concentrations range from 880 to 3520 µmol/L (160–640 µg/L) of packed red cells. Subnormal levels occur in patients with megaloblastic anemia due to folate deficiency but also in nearly two-thirds of patients with severe cobalamin deficiency. False-normal results may occur if a folate-deficient patient has received a recent blood transfusion or if a patient has a raised reticulocyte count. Serum homocysteine assay is discussed earlier.

Tests for the Cause of Folate Deficiency

The diet history is important. Tests for transglutaminase antibodies are performed to confirm or exclude celiac disease. If positive, duodenal biopsy is needed. An underlying disease causing increased folate breakdown should also be excluded.

Treatment

TREATMENT

Cobalamin and Folate Deficiency

It is usually possible to establish which of the two deficiencies, folate or cobalamin, is the cause of the anemia and to treat only with the appropriate vitamin. In patients who enter the hospital severely ill, however, it may be necessary to treat with both vitamins in large doses once blood samples have been taken for cobalamin and folate assays and a bone marrow biopsy has been performed (if deemed necessary). Transfusion is usually unnecessary and inadvisable. If it is essential, packed red cells should be given slowly, one or two units only, with the usual treatment for heart failure if present. Potassium supplements have been recommended to obviate the danger of the hypokalemia but are

not necessary. Occasionally, an excessive rise in platelets occurs after 1–2 weeks of therapy. Antiplatelet therapy, for example, aspirin, should be considered if the platelet count rises to >800 × 10^9/L.

COBALAMIN DEFICIENCY

It is usually necessary to treat patients who have developed cobalamin deficiency with lifelong regular cobalamin injections. In the UK, the form used is hydroxocobalamin; in the United States, cyanocobalamin. In a few instances, the underlying cause of cobalamin deficiency can be permanently corrected, for example, fish tapeworm, tropical sprue, or an intestinal stagnant loop that is amenable to surgery. The indications for starting cobalamin therapy are a well-documented megaloblastic anemia or other hematologic abnormalities and neuropathy due to the deficiency. Patients with borderline serum cobalamin levels but no hematologic or other abnormality may be followed to make sure that the cobalamin deficiency does not progress (see below). If malabsorption of cobalamin or rises in serum MMA levels have been demonstrated, however, these patients also should be given regular maintenance cobalamin therapy. Cobalamin should be given routinely to all patients who have had a total gastrectomy or ileal resection. Patients who have undergone gastric reduction for control of obesity or who are receiving long-term treatment with proton pump inhibitors should be screened and, if necessary, given cobalamin replacement.

Replenishment of body stores should be complete with six 1000-µg IM injections of hydroxocobalamin given at 3- to 7-day intervals. More frequent doses are usually used in patients with cobalamin neuropathy, but there is no evidence that they produce a better response. Allergic reactions are rare and may require desensitization or antihistamine or glucocorticoid cover. For maintenance therapy, 1000 µg hydroxocobalamin IM once every 3 months is satisfactory. Because of the poorer retention of cyanocobalamin, protocols generally use higher and more frequent doses, for example, 1000 µg IM, monthly, for maintenance treatment.

Because a small fraction of cobalamin can be absorbed passively through mucous membranes even when there is complete failure of physiologic IF-dependent absorption, large daily oral doses (1000–2000 µg) of cyanocobalamin are used in PA for replacement (especially in Canada and Sweden) and maintenance of normal cobalamin status in, for example, food malabsorption of cobalamin. Sublingual therapy has also been proposed for those in whom injections are difficult because of a bleeding tendency and who may not tolerate oral therapy. If oral therapy is used, it is important to monitor compliance, particularly with elderly, forgetful patients. This author prefers parenteral therapy for initial treatment, particularly in severe anemia or if a neuropathy is present, and for maintenance in PA. Oral B_{12} therapy even with low doses of 50 µg daily may have a larger role in treating food malabsorption of B_{12}.

For treatment of patients with subnormal serum vitamin B_{12} levels with a normal MCV and no hypersegmentation of neutrophils, a negative IF antibody test in the absence of tests of B_{12} absorption is problematic. Some (perhaps 15%) cases may be due to TC I (HC) deficiency. Homocysteine and/or MMA measurements may help, but in the absence of these tests and with otherwise normal gastrointestinal function, repeat serum B_{12} assay after 6–12 months may help one decide whether to start cobalamin therapy.

Vitamin B_{12} injections are used in a wide variety of diseases, often neurologic, despite normal serum B_{12} and folate levels and a normal blood count and in the absence of randomized, double-blind, controlled trials. These conditions include multiple sclerosis and chronic fatigue syndrome/myalgic encephalomyelitis (ME). It seems probable that any benefit is due to the placebo effect of a usually painless, pink injection. In ME, oral B_{12} therapy, despite providing equally large amounts of B_{12}, has not been beneficial, supporting the view of the effect of the injections being placebo only.

FOLATE DEFICIENCY

Oral doses of 5–15 mg folic acid daily are satisfactory, as sufficient folate is absorbed from these extremely large doses even in patients with severe malabsorption. The length of time therapy must be continued depends on the underlying disease. It is customary to continue therapy for about 4 months, when all folate-deficient red cells will have been eliminated and replaced by new folate-replete populations.

Before large doses of folic acid are given, cobalamin deficiency must be excluded and, if present, corrected; otherwise cobalamin neuropathy may develop despite a response of the anemia of cobalamin deficiency to folate therapy. Studies in the United States, however, suggest that there is no increase in the proportion of individuals with low serum cobalamin levels and no anemia since food fortification with folic acid, but it is unknown if there has been a change in incidence of cobalamin neuropathy.

Long-term folic acid therapy is required when the underlying cause of the deficiency cannot be corrected and the deficiency is likely to recur, for example, in chronic dialysis or hemolytic anemias. It may also be necessary in gluten-induced enteropathy that does not respond to a gluten-free diet. Where mild but chronic folate deficiency occurs, it is preferable to encourage improvement in the diet after correcting the deficiency

with a short course of folic acid. In any patient receiving long-term folic acid therapy, it is important to measure the serum cobalamin level at regular (e.g., once-yearly) intervals to exclude the coincidental development of cobalamin deficiency.

Folinic Acid (5-Formyl-THF)

This is a stable form of fully reduced folate. It is given orally or parenterally to overcome the toxic effects of methotrexate or other DHF reductase inhibitors, for example, trimethoprim or cotrimoxazole.

PROPHYLACTIC FOLIC ACID

Prophylactic folic acid is used in chronic dialysis patients and in parenteral feeds. Prophylactic folic acid has been used to reduce homocysteine levels to prevent cardiovascular disease and for cognitive function in the elderly, but there are no firm data to show any benefit.

Pregnancy

In over 70 countries (but none in Europe), food is fortified with folic acid (in grain or flour) to reduce the risk of NTDs. Nevertheless, folic acid, 400 µg daily, should be given as a supplement before and throughout pregnancy to prevent megaloblastic anemia and reduce the incidence of NTDs, even in countries with fortification of the diet. The levels of fortification provide up to 400 µg daily on average in Chile, but in most countries, it is nearer to 200 µg, so periconceptual folic acid is still needed. Most if not all the folic acid used in fortification and eaten over three meals a day will be converted during absorption to methyltetrahydrofolate. This compound will not correct the anemia in B_{12} deficiency.

Studies in early pregnancy show significant lack of compliance with the folic acid supplements, emphasizing the benefit of food fortification. Supplemental folic acid reduces the incidence of birth defects in babies born to diabetic mothers. In women who have had a previous fetus with an NTD, 5 mg daily is recommended when pregnancy is contemplated and throughout the subsequent pregnancy.

Infancy and Childhood

The incidence of folate deficiency is so high in the smallest premature babies during the first 6 weeks of life that folic acid (e.g., 1 mg daily) should be given routinely to those weighing <1500 g at birth and to larger premature babies who require exchange transfusions or develop feeding difficulties, infections, or vomiting and diarrhea.

The World Health Organization currently recommends routine supplementation with iron and folic acid in children in countries where iron deficiency is common and child mortality, largely due to infectious diseases, is high. However, some studies suggest that in areas where malaria rates are high, this approach may increase the incidence of severe illness and death. Even where malaria is rare, there appears to be no survival benefit.

MEGALOBLASTIC ANEMIA NOT DUE TO COBALAMIN OR FOLATE DEFICIENCY OR ALTERED METABOLISM

This may occur with many antimetabolic drugs (e.g., hydroxyurea, cytosine arabinoside, 6-mercaptopurine) that inhibit DNA replication. Antiviral nucleoside analogues used in treatment of HIV infection may also cause macrocytosis and megaloblastic marrow changes. In the rare disease orotic aciduria, two consecutive enzymes in purine synthesis are defective. The condition responds to therapy with uridine, which bypasses the block. In thiamine-responsive megaloblastic anemia, there is a genetic defect in the high-affinity thiamine transport (*SLC19A2*) gene. This causes defective RNA ribose synthesis through impaired activity of transketolase, a thiamine-dependent enzyme in the pentose cycle. This leads to reduced nucleic acid production. It may be associated with diabetes mellitus and deafness and the presence of many ringed sideroblasts in the marrow. The explanation is unclear for megaloblastic changes in the marrow in some patients with acute myeloid leukemia and myelodysplasia.

FURTHER READING

Bunn HF: Vitamin B_{12} and pernicious anemia: The dawn of molecular medicine. N Engl J Med 370:773, 2014. [PubMed: 24552327]

Carmel R: Subclinical cobalamin deficiency. Curr Opin in Gastro 28:151, 2012.

Devalia V et al: Guidelines for the diagnosis and treatment of cobalamin and folate disorders. Br J Haematol 166:496, 2014. [PubMed: 24942828]

Hesdorffer CS, Longo DL: Drug-induced megaloblastic anemia. N Engl J Med 373:1649, 2015. [PubMed: 26488695]

Huo Y et al: Efficacy of folic acid therapy in the primary prevention of stroke among adults with hypertension in China: The CSPPT randomized clinical trial. JAMA 313:1325, 2015. [PubMed: 25771069]

Morris JK et al: Prevention of neural tube defects in the UK: A missed opportunity. Arch Dis Child 101:604, 2016. [PubMed: 26681697]

Shipton MJ, Thachil J: Vitamin B_{12} deficiency—A 21st century perspective. Clin Med 15:145, 2015.

Stabler SP: Vitamin B_{12} deficiency. N Engl J Med 368:149, 2013. [PubMed: 23301732]

Vollset SE et al: Effects of folic acid supplementation on overall and site-specific cancer incidence during randomised trials: Meta-analyses of data on 50 000 individuals. Lancet 381:1029, 2013. [PubMed: 23352552]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 96: Hemolytic Anemias

Lucio Luzzatto

DEFINITIONS

A finite life span is a distinct characteristic of red cells. Hence, a logical, time-honored classification of anemias is in three groups: (1) decreased production of red cells, (2) increased destruction of red cells, and (3) acute blood loss. Decreased production is covered in **Chaps. 93, 94, and 98**; acute blood loss in **Chap. 97**; increased destruction is covered in this chapter.

All patients who are anemic as a result of either increased destruction of red cells or acute blood loss have one important element in common: the anemia results from overconsumption of red cells from the peripheral blood, whereas the supply of cells from the bone marrow is normal (indeed, it is usually increased). On the other hand, these two groups differ in that the consequences of physical loss of red cells from the bloodstream or *from* the body itself, as in acute hemorrhage, is fundamentally different from destruction of red cells *within* the body, as in hemolytic anemias (HAs).

With respect to primary etiology, HAs may be *inherited* or *acquired*; from a clinical point of view, they may be more *acute* or more *chronic*, and they may vary from mild to very severe; the site of hemolysis may be predominantly *intravascular* or *extravascular.* With respect to mechanisms, HAs may be due to *intracorpuscular* causes or to *extracorpuscular* causes **(Table 96-1)**. But before reviewing the individual types of HA, it is appropriate to consider what general features they have in common, in terms of clinical aspects and of pathophysiology.

TABLE 96-1

Classification of Hemolytic Anemias[a]

	Intracorpuscular Defects	Extracorpuscular Factors
Inherited	Hemoglobinopathies Enzymopathies Membrane-cytoskeletal defects	Familial (atypical) hemolytic-uremic syndrome
Acquired	Paroxysmal nocturnal hemoglobinuria (PNH)	Mechanical destruction (microangiopathic) Toxic agents Drugs Infectious Autoimmune

[a]Hereditary causes correlate with intracorpuscular defects because these defects are due to inherited mutations; the one exception is PNH because the defect is due to an acquired somatic mutation. Similarly, acquired causes correlate with extracorpuscular factors because mostly these factors are exogenous; the one exception is familial hemolytic-uremic syndrome (HUS; often referred to as atypical HUS) because here an inherited abnormality allows complement activation to be excessive, with bouts of production of membrane attack complex capable of destroying normal red cells. Interestingly, in both PNH and aHUS hemolysis is complement-mediated.

GENERAL CLINICAL AND LABORATORY FEATURES

The clinical presentation of a patient with anemia is greatly influenced in the first place by whether the onset is abrupt or gradual, and HAs are no exception. A patient with autoimmune HA or with favism may be a medical emergency, whereas a patient with mild hereditary spherocytosis (HS) or with cold agglutinin disease (CAD) may be diagnosed after years. This is due in large measure to the remarkable ability of the body to adapt to anemia when it is slowly progressing **(Chap. 59)**.

What differentiates HAs from other anemias is that the patient has signs and symptoms arising directly from hemolysis **(Table 96-2)**. At the clinical level, the main sign is *jaundice*; in addition, the patient may report discoloration of the urine. In many cases of HA, the spleen is enlarged because it is a preferential site of hemolysis; and in some cases, the liver may be enlarged as well. In all severe congenital forms of HA, there may be also skeletal

changes due to overactivity of the bone marrow: they are never as severe as in thalassemia major because there is less ineffective erythropoiesis, or none at all.

TABLE 96-2

Features Common to Most Patients with a Hemolytic Disorder

General examination	Jaundice, pallor
Other physical findings	Spleen may be enlarged; bossing of skull in severe congenital cases
Hemoglobin level	From normal to severely reduced
MCV, MCH	Usually increased
Reticulocytes	Usually Increased
Bilirubin	Almost always increased (mostly unconjugated)
LDH	Increased (up to 10× normal with intravascular hemolysis)
Haptoglobin	Reduced to absent if hemolysis is at least in part intravascular)

Abbreviations: LDH, lactate dehydrogenase; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume.

The laboratory features of HA are related to (i) hemolysis per se, (ii) the erythropoietic response of the bone marrow. In most cases hemolysis is largely extravascular, and it produces an increase in unconjugated bilirubin and aspartate aminotransferase (AST) in the serum; urobilinogen will be increased in both urine and stool. If hemolysis is mainly intravascular, the telltale sign is hemoglobinuria (often associated with hemosiderinuria); in the serum there is free hemoglobin, lactate dehydrogenase (LDH) is increased, and haptoglobin is reduced. In contrast, the serum bilirubin level may be normal or only mildly elevated. The main sign of the erythropoietic response by the bone marrow is an increase in reticulocytes (a test all too often neglected in the initial workup of a patient with anemia). Usually the increase will be reflected in both the percentage of reticulocytes (the more commonly quoted figure) and in the absolute reticulocyte count (the more definitive parameter). The increased number of reticulocytes is associated with an increased mean corpuscular volume (MCV) in the blood count. On the blood smear, this is reflected in the presence of macrocytes; there is also polychromasia, and sometimes one sees nucleated red cells. In most cases, a bone marrow aspirate is not necessary in the diagnostic workup; if it is done, it will show erythroid hyperplasia. In practice, once an HA is suspected, specific tests will usually be required for a definitive diagnosis of a specific type of HA.

GENERAL PATHOPHYSIOLOGY

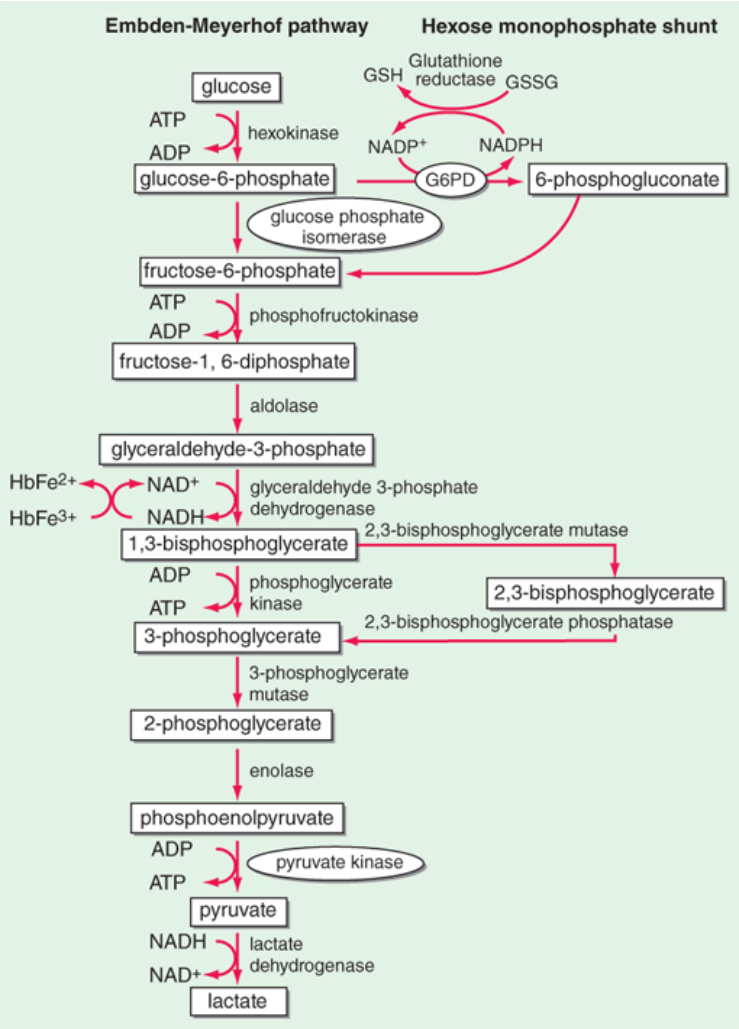
The mature red cell is the product of a developmental pathway that brings the phenomenon of differentiation to an extreme. An orderly sequence of events produces synchronous changes, whereby the gradual accumulation of a huge amount of hemoglobin in the cytoplasm (to a final level of 340 g/L, i.e., about 5 m*M*) goes hand in hand with the gradual loss of cellular organelles and of biosynthetic abilities. In the end, the erythroid cell undergoes a process that has features of apoptosis, including nuclear pyknosis and eventually extrusion of the nucleus. However, the final result is more altruistic than suicidal; the cytoplasmic body, instead of disintegrating, is now able to provide oxygen to all cells in the human organism for some remaining 120 days of the red cell life span.

As a result of this unique process of differentiation and maturation, intermediary metabolism is drastically curtailed in mature red cells **(Fig. 96-1)**; for instance, cytochrome-mediated oxidative phosphorylation has been lost with the loss of mitochondria (through a process of physiologic autophagy); therefore, there is no backup to anaerobic glycolysis, which in the red cell is the only provider of adenosine triphosphate (ATP). Also, the capacity of making protein has been lost with the loss of ribosomes. This places the cell's limited metabolic apparatus at risk, because if any protein component deteriorates, it cannot be replaced, as it would be in most other cells; and in fact the activity of most enzymes gradually decreases as red cells age. At the same time, during their long time in circulation, various red cell components inevitably accumulate damage; in senescent red cells, the membrane protein band 3 molecules (see below and Fig. 96-1), having bound hemichromes on their intracellular domains, tend to cluster. Now they bind anti–band 3 IgG antibodies (present in most people) and C3 complement fragments; thus they become opsonized and are eventually removed by phagocytosis in the reticuloendothelial system.

FIGURE 96-1

Red blood cell (RBC) metabolism. The Embden-Meyerhof pathway (glycolysis) generates ATP required for cation transport and for membrane maintenance. The generation of NADH maintains hemoglobin iron in a reduced state. The hexose monophosphate shunt generates NADPH that is used

to reduce glutathione, which protects the red cell against oxidant stress; the 6-phosphogluconate, after decarboxylation, can be recycled via pentose sugars to glycolysis. Regulation of the 2,3-bisphosphoglycerate level is a critical determinant of oxygen affinity of hemoglobin. Enzyme deficiency states in order of prevalence: glucose 6-phosphate dehydrogenase (G6PD) > pyruvate kinase > glucose-6-phosphate isomerase > rare deficiencies of other enzymes in the pathway. The more common enzyme deficiencies are encircled.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Another consequence of the relative simplicity of red cells is that they have a very limited range of ways to manifest distress under hardship; in essence, any sort of metabolic failure will eventually lead either to structural damage to the membrane or to failure of the cation pump. In either case, the life span of the red cell is reduced, which is the definition of a *hemolytic disorder*. If the rate of red cell destruction exceeds the capacity of the bone marrow to produce more red cells, the hemolytic disorder will manifest as HA.

Thus, the essential pathophysiologic process common to all HAs is an increased red cell turnover; and in many HAs, this is due at least in part to an acceleration of the senescence process described above. The gold standard for proving that the life span of red cells is reduced (compared to the normal value of about 120 days) is a *red cell survival* study, which can be carried out by labeling the red cells with ^{51}Cr and measuring the fall in radioactivity over several days or weeks (this classic test can now be replaced by a methodology using the non-radioactive isotope ^{15}N). If the hemolytic event is transient, it does not usually cause any long-term consequences, except for an increased requirement for erythropoietic factors, particularly folic acid. However, if hemolysis is recurrent or persistent, the increased bilirubin production favors the formation of gallstones. If a considerable proportion of hemolysis takes place in the spleen, as is often the case, splenomegaly may become increasingly a feature, and hypersplenism may develop, with consequent neutropenia and/or thrombocytopenia.

The increased red cell turnover also has metabolic consequences. In normal subjects, the iron from effete red cells is very efficiently recycled by the body; however, with chronic intravascular hemolysis, the persistent hemoglobinuria will cause considerable iron loss, needing replacement. With chronic extravascular hemolysis, the opposite problem, iron overload, is more common, especially if the patient needs frequent blood transfusions; however, if erythropoiesis is massively increased, the hepcidin-mediated regulation of iron absorption may be disturbed, to the extent that iron overload may set in even without blood transfusion. In the long run, in the absence of iron-chelation therapy iron overload will cause secondary

hemochromatosis; this will cause damage particularly to the liver, eventually leading to cirrhosis; and to the heart muscle, eventually causing heart failure.

Compensated Hemolysis versus Hemolytic Anemia

Red cell destruction is a potent stimulus for erythropoiesis, which is mediated by erythropoietin (EPO) produced by the kidney. This mechanism is so effective that in many cases the increased output of red cells from the bone marrow can fully balance an increased destruction of red cells. In such cases, we say that hemolysis is *compensated*. The pathophysiology of compensated hemolysis is similar to what we have just described, except there is no anemia. This notion is important from the diagnostic point of view, because a patient with a hemolytic condition, even an inherited one, may present without anemia; and it is also important from the point of view of management because compensated hemolysis may become "decompensated," i.e., anemia may suddenly appear in certain circumstances, for instance in pregnancy, folate deficiency, or renal failure interfering with adequate EPO production. Another general feature of chronic HAs is seen when any intercurrent condition, such as an acute infection, depresses erythropoiesis. When this happens, in view of the increased rate of red cell turnover, the effect will be predictably much more marked than in a person who does not have hemolysis. The most dramatic example is infection by parvovirus B19, which may cause a rather precipitous fall in hemoglobin—an occurrence sometimes referred to as *aplastic crisis*.

INHERITED HEMOLYTIC ANEMIAS

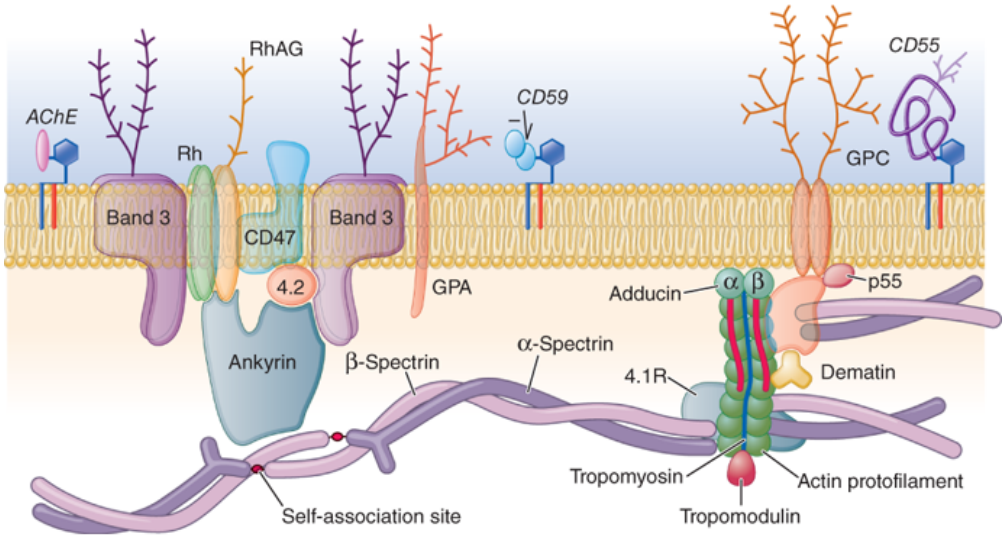
There are three essential components in the red cell: (1) hemoglobin, (2) the membrane-cytoskeleton complex, and (3) the metabolic machinery necessary to keep hemoglobin and the membrane-cytoskeleton complex in working order. Diseases caused by inherited abnormalities of hemoglobin, or hemoglobinopathies, are covered in **Chap. 94**. Here we will deal with diseases of the other two components.

Hemolytic Anemias due to Abnormalities of the Membrane-Cytoskeleton Complex

The detailed architecture of the red cell membrane is complex, but its basic design is relatively simple **(Fig. 96-2)**. The lipid bilayer incorporates phospholipids and cholesterol, and it is spanned by a number of proteins that have their hydrophobic transmembrane domain(s) embedded in the membrane; most of these proteins also extend to both the outside (extracellular domains) and the inside of the cell (cytoplasmic domains). Other proteins are tethered to the membrane through a glycosylphosphatidylinositol (GPI) anchor; these have only an extracellular domain, and they include ion channels, receptors for complement components, and receptors for other ligands. The most abundant red cell membrane proteins are glycophorins and the so-called *band 3*, an anion transporter. The extracellular domains of many of these proteins are heavily glycosylated, and they carry antigenic determinants that correspond to blood groups. Underneath the membrane, and tangential to it, is a network of other proteins that make up the cytoskeleton. The main cytoskeletal protein is spectrin, the basic unit of which is a dimer of α-spectrin and β-spectrin. The membrane is physically linked to the cytoskeleton by a third set of proteins (including ankyrin and the so-called *band 4.1* and *band 4.2*), which thus make these two structures intimately connected to each other.

FIGURE 96-2

The red cell membrane. In this figure, one sees, within the lipid bilayer, several membrane proteins, of which band 3 (anion exchanger 1 [AE1]) is the most abundant; the α-β spectrin dimers that associate to form most of the cytoskeleton; and several proteins (e.g., ankyrin) that connect the membrane to the cytoskeleton. In addition, as examples of glycosylphosphatidylinositol (GPI)-linked proteins, one sees acetylcholinesterase (AChE) and the two complement-regulatory proteins CD59 and CD55. The (non-realistic) shapes of the protein moieties of the GPI-linked proteins are meant to indicate that each one of them is different from others and that, unlike with the other membrane proteins shown, the entire polypeptide chain is extracellular. Branched lines symbolize carbohydrate moiety of proteins. The molecules are obviously not drawn to the same scale. Additional explanations can be found in the text. *(From N Young et al: Clinical Hematology. Copyright Elsevier, 2006; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The membrane-cytoskeleton complex has essentially three functions: It is an envelope for the red cell cytoplasm, it maintains the normal red cells shape, it provides highly specific cross-membrane transport of electrolytes and of metabolites such as glucose. In the membrane-cytoskeleton complex the individual components are so intimately integrated with each other that an abnormality of almost any of them will be disturbing or disruptive, causing structural or functional failure, which results ultimately in hemolysis. These abnormalities are almost invariably inherited mutations; thus, diseases of the membrane-cytoskeleton complex belong to the category of inherited HAs. Before the red cells lyse, they often exhibit more or less specific morphologic changes that alter the normal biconcave disk shape. Thus, the majority of the diseases in this group have been known for over a century as hereditary spherocytosis (HS) and *hereditary elliptocytosis* (HE; as well as more rare ones like *stomatocytosis*, *xerocytosis*, etc). Now that their molecular basis has been elucidated, it has emerged **(see Table 96-3)** that, although we are dealing with monogenic disorders, there is no one-to-one correlation between a certain gene and a certain disorder. Rather, what has been regarded as a single disorder (e.g., HS) can arise through mutation of one of several genes; conversely, what have been regarded as different disorders can arise through different mutations of the very same gene **(Fig. 96-3)**.

TABLE 96-3

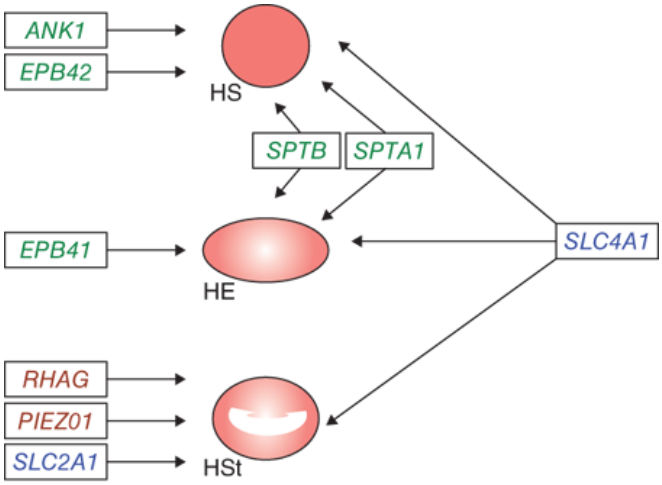
Inherited Diseases of the Red Cell Membrane-Cytoskeleton Complex

Gene	Chromosomal Location	Protein Produced	Disease(s) with Certain Mutations (Inheritance)	Comments
SPTA1	1q22-q23	α-Spectrin	HS (recessive)	Rare
			HE (dominant)	Mutations of this gene account for about 65% of HE. More severe forms may be due to coexistence of an otherwise silent mutant allele.
SPTB	14q23-q24.1	β-Spectrin	HS (dominant)	Rare
			HE (dominant)	Mutations of this gene account for about 30% of HE, including some severe forms.
ANK1	8p11.2	Ankyrin	HS (dominant)	May account for majority of HS.
SLC4A1	17q21	Band 3; also known as AE (anion exchanger) or AE1	HS (dominant)	Mutations of this gene may account for about 25% of HS.
			Southeast Asia ovalocytosis (dominant)	Polymorphic mutation (deletion of 9 amino acids); in heterozygotes clinically asymptomatic and protective against *Plasmodium falciparum.*
			Stomatocytosis	Certain specific missense mutations shift protein function from anion exchanger to cation conductance.
EPB41	1p33-p34.2	Band 4.1	HE (dominant)	Mutations of this gene account for about 5% of HE, mostly with prominent morphology but little/no hemolysis in heterozygotes; severe hemolysis in homozygotes.
EPB42	15q15-q21	Band 4.2	HS (recessive)	Mutations of this gene account for about 3% of HS.
RHAG	6p21.1-p11	Rhesus-associated glycoprotein	Chronic nonspherocytic hemolytic anemia (recessive)	Very rare; associated with total loss of all Rh antigens. One specific mutation in this gene causes overhydrated stomatocytosis.
PIEZO1	16q23-q24	PIEZO1 (mechanosensitive cation channel)	Dehydrated hereditary stomatocytosis (dominant)	Also known as xerocytosis with pseudohyperkalemia. Patients may present with perinatal edema.
KCNN4	19q13.31	KCNN4 Intermediate conductance calcium-activated potassium channel protein 4	Dehydrated hereditary stomatocytosis (dominant)	Clinical presentation similar to that of *PIEZO1* mutants.
SLC2A1	1p34.2,	GLUT1 glucose transporter	Over-hydrated hereditary stomatocytosis	Associated with serious neurological manifestations.

Abbreviations: HE, hereditary elliptocytosis; HS, hereditary spherocytosis.

FIGURE 96-3

Hereditary spherocytosis (HS), hereditary elliptocytosis (HE), and hereditary stomatocytosis (HSt) are three morphologically distinct forms of congenital hemolytic anemia. It has emerged that each one can arise from mutation of one of several genes and that different mutations of the same gene can give one or another form. (See also Table 96-3.)



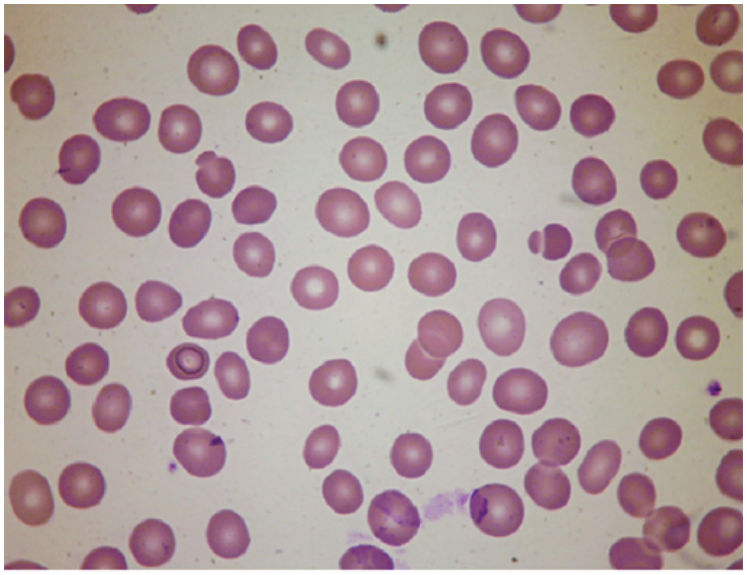
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HEREDITARY SPHEROCYTOSIS

This is a relatively common type of genetically determined HA, with an estimated frequency of at least 1 in 5000. Its identification is credited to Minkowksy and Chauffard, who, at the end of the nineteenth century, reported families who had the presence of numerous spherocytes in the peripheral blood (**Fig. 96-4A**). In vitro studies revealed that the red cells were abnormally susceptible to lysis in hypotonic media; indeed, the presence of *osmotic fragility* became the main diagnostic test for HS. Today we know that HS, thus defined, is genetically heterogeneous; i.e., it can arise from a variety of mutations in one of several genes (Table 96-3). It has been also recognized that the inheritance of HS is not always autosomal dominant (with the patient being heterozygous); indeed, some of the most severe forms are instead autosomal recessive (with the patient being homozygous).

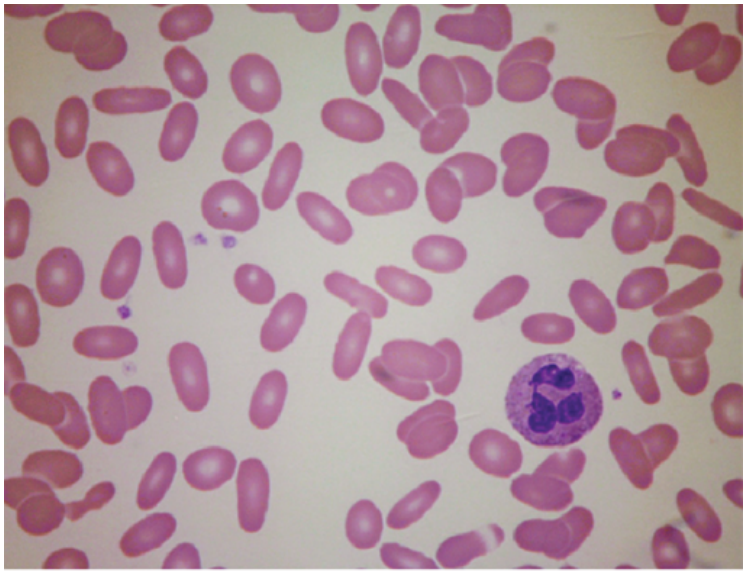
FIGURE 96-4

Peripheral blood smear from patients with membrane-cytoskeleton abnormalities. ***A.*** Hereditary spherocytosis. ***B.*** Hereditary elliptocytosis, heterozygote. ***C.*** Elliptocytosis, with both alleles of the α-spectrin gene mutated.



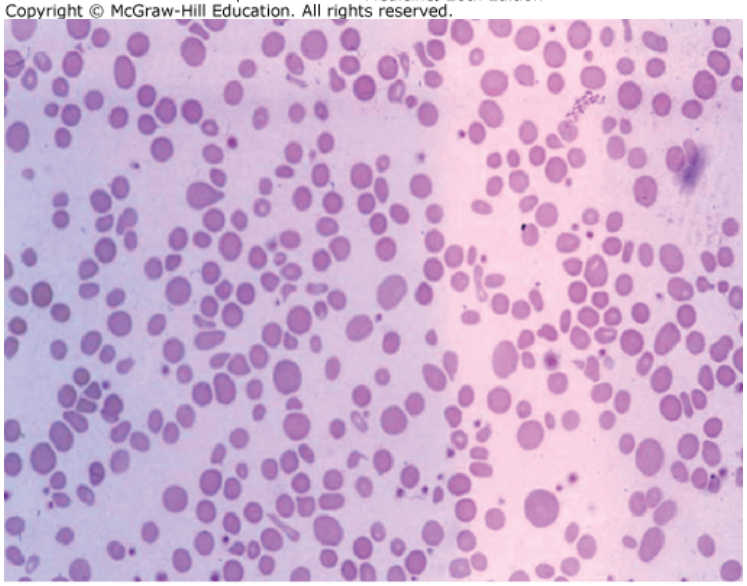
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical Presentation and Diagnosis

The spectrum of clinical severity of HS is broad. Severe cases may present in infancy with severe anemia, whereas mild cases may present in young adults or even later in life. The main clinical findings are jaundice, an enlarged spleen, and often gallstones; indeed, it may be the finding of gallstones in a young person that triggers diagnostic investigations.

The variability in clinical manifestations that is observed among patients with HS is largely due to the different underlying molecular lesions (Table 96-3). Not only are mutations of several genes involved: even different mutations of the same gene can give very different clinical manifestations. In milder cases, hemolysis is often compensated (see above), but changes in clinical expression may be seen even in the same patient because intercurrent conditions (e.g., pregnancy, infection) may cause decompensation. The anemia is usually normocytic, with the characteristic morphology that gives the disease its name. An increased mean corpuscular hemoglobin concentration (MCHC >34) on an ordinary blood count report should raise the suspicion of HS, because HS is almost the only condition in which this abnormality occurs. It has been apparent for a long time that the spleen plays a special role in HS through a dual mechanism. On one hand, like in many other HAs, the spleen itself is a major site of destruction; on the other hand, transit through the splenic circulation makes the defective red cells more spherocytic and, therefore, accelerates their demise, even though that may take place elsewhere.

When there is a family history, it is usually easy to make a diagnosis based on features of HA and typical red cell morphology. However, there may be no family history for at least two reasons. First, the patient may have a de novo mutation, i.e., a mutation that has taken place in a germ cell of one of his parents or early after zygote formation. Second, the patient may have a recessive form of HS (Table 96-3). In such cases, more extensive laboratory

investigations are required, including osmotic fragility, the acid glycerol lysis test, the eosin-5′-maleimide (EMA)–binding test, and SDS-gel electrophoresis of membrane proteins; these tests are usually carried out in laboratories with special expertise in this area. Sometimes a definitive diagnosis can be obtained only by molecular studies demonstrating a mutation in one of the genes underlying HS (Table 96-3).

Treatment

TREATMENT

Hereditary Spherocytosis

We do not have a causal treatment for HS; i.e., no way has yet been found to correct the basic defect in the membrane-cytoskeleton structure. Given the special role of the spleen in HS (see above), it has long been thought that an almost obligatory therapeutic measure was splenectomy. Because this operation may have more than trivial consequences, today we have more articulate recommendations, based on disease severity, as follows. In mild cases, avoid splenectomy; in moderate cases, delay splenectomy until puberty; in severe cases, proceed with splenectomy at the age of 4–6. It is also helpful, whenever possible, to know about the outcome of splenectomy in the patient's affected relatives. Antipneumococcal vaccination before splenectomy is imperative, whereas penicillin prophylaxis after splenectomy is controversial. Along with splenectomy, cholecystectomy should not be carried out automatically; it should be carried out, usually by the laparoscopic approach, when clinically indicated.

HEREDITARY ELLIPTOCYTOSIS

HE is at least as heterogeneous as HS, both from the genetic point of view (Table 96-3, Fig. 96-3) and from the clinical point of view. Again, it is the shape of the red cells **(Fig. 96-4B)** that gives the name to the condition, but there is no direct correlation between the elliptocytic morphology and clinical severity. In fact, some mild or even asymptomatic cases may have nearly 100% elliptocytes (or ovalocytes); whereas in severe cases all kinds of bizarre poikilocytes may predominate. Clinical features and recommended management are similar to those outlined above for HS. Although the spleen may not have the specific role it has in HS, in severe cases, splenectomy may be beneficial. The prevalence of HE causing clinical disease is similar to that of HS. However, one particular in-frame deletion of nine amino acids in the *SLC4A1* gene encoding band 3, which underlies the so-called *Southeast Asia ovalocytosis* (SAO), has a frequency of up to 7% in certain populations, presumably as a result of malaria selection; it is asymptomatic in heterozygotes and probably lethal in homozygotes.

Disorders of Cation Transport

These rare conditions with autosomal dominant inheritance are characterized by increased intracellular sodium in red cells, with concomitant loss of potassium; indeed, they are sometimes discovered through the incidental finding, in a blood test, of a high serum K^+ (*pseudohyperkalemia*). In patients from some families, the cation transport disturbance is associated with gain of water; as a result, the red cells are overhydrated (low MCHC), and on a blood smear, the normally round-shaped central pallor is replaced by a linear-shaped central pallor, which has earned this disorder the name *stomatocyto*sis (Fig. 96-3). In patients from other families, instead, the red cells are dehydrated (high MCHC), and their consequent rigidity has earned this disorder the name *xerocytosis*. One would surmise that in these disorders the primary defect may be in a cation transporter; indeed, xerocytosis results from mutations in *PIEZO1*. In other patients with stomatocytosis, mutations are found in other genes also related to solute transport (Table 96-3), including *SLC4A1* (encoding band 3), the Rhesus gene *RHAG*, and the glucose transporter gene *SLC2A1* responsible for a special form called cryohydrocytosis, an unusual variant in which the red cells swell and burst when they are cooled. In vivo hemolysis can vary from relatively mild to quite severe. From the practical point of view, it is important to know that in stomatocytosis splenectomy is strongly contraindicated because it has been followed in a significant proportion of cases by severe thromboembolic complications.

Enzyme Abnormalities

When there is an important defect in a component of the membrane cytoskeleton complex, hemolysis is a direct consequence of the fact that the very structure of the red cell is compromised. Instead, when one of the enzymes is defective, the consequences will depend on the precise role of that enzyme in the metabolic machinery of the red cell. This machinery has two main functions: (1) to provide energy in the form of ATP and (2) to prevent oxidative damage to hemoglobin and to other proteins by providing sufficient reductive potential; the key molecule for this is NADPH.

ABNORMALITIES OF THE GLYCOLYTIC PATHWAY

Because red cells, in the course of their differentiation, have sacrificed not only their nucleus and their ribosomes but also their mitochondria, they rely exclusively on the anaerobic portion of the glycolytic pathway for producing ATP, most of which is required by the red cell for cation transport against a concentration gradient across the membrane. If this fails due to a defect of any of the enzymes of the glycolytic pathway **(Table 96-4)**, the result will be hemolytic disease.

TABLE 96-4

Red Cell Enzyme Abnormalities Causing Hemolysis

	Enzyme (Acronym)	Gene Symbol; Chromosomal Location	Prevalence of Enzyme Deficiency (Rank)	Clinical Manifestations Extra-Red Cell	Comments
Glycolytic Pathway					
	Hexokinase (HK)	*HK1*; 10q22	Very rare		May benefit from splenectomy; BMT[c]
	Glucose 6-phosphate isomerase (G6PI)	*GPI*; 19q31.1	Rare (4)[a]	NM, CNS	May benefit from splenectomy
	Phosphofructokinase (PFK)[b]	*PFKM*; 12q13	Very rare	Myopathy; myoglobinuria	
	Aldolase	*ALDOA*; 16q22-24	Very rare	Myopathy	
	Triose phosphate isomerase (TPI)	*TPI1*; 12p13.31	Very rare	CNS (severe), NM	
	Glyceraldehyde 3-phosphate dehydrogenase (GAPD)	*GAPDH*; 12p13.31-	Very rare	Myopathy	
	Bisphosphoglycerate mutase (DPGM)	*BPGM*; 7q33	Very rare		Erythrocytosis rather than hemolysis
	Phosphoglycerate kinase (PGK)	*PGK1*; Xq21.1	Very rare	CNS, NM	May benefit from splenectomy; BMT[c]
	Pyruvate kinase (PK)	*PKLR*; 1q22	Rare (2)[a]		May benefit from splenectomy; BMT[c]
Redox					
	Glucose 6-phosphate dehydrogenase (G6PD)	*G6PD*; Xq28	Common (1)[a]	Very rarely granulocytes	In almost all cases, only AHA from exogenous trigger
	Glutathione synthase	*GSS*; 20q11.22	Very rare	CNS	
	Glutathione reductase	*GSR:* 8p12	Very rare	Cataracts	AHA from exogenous trigger (favism)
	γ-Glutamylcysteine synthase	*GCLC*; 6p12.1	Very rare	CNS	Mutations affect catalytic subunit
	Cytochrome b5 reductase	*CYB5R3*: 22q13.2	Rare	CNS	Methemoglobinemia rather than hemolysis
Nucleotide Metabolism					
	Adenylate kinase (AK)	*AK1*; 9q34.11	Very rare	CNS	May benefit from splenectomy

	Enzyme (Acronym)	Gene Symbol; Chromosomal Location	Prevalence of Enzyme Deficiency (Rank)	Clinical Manifestations Extra-Red Cell	Comments
	Pyrimidine 5' nucleotidase (P5N)	*NTSC3A*; 7p14.3	Rare (3)[a]		May benefit from splenectomy

[a]The numbers from (1) to (4) indicate the ranking order of these enzymopathies in terms of frequency. [b]PFK deficiency is associated with increased glycogen in muscle, and it is also known as glycogen storage disease type VII or Tarui's disease. [c]Occasional report of successful treatment of the hematologic manifestations by BMT.

Abbreviations: AHA, acquired hemolytic anemia; CNS, central nervous system; NM, neuromuscular.

Pyruvate Kinase Deficiency

Abnormalities of the glycolytic pathway are all inherited and all rare. Among them, deficiency of pyruvate kinase (PK) is the least rare, with an estimated prevalence in most populations of the order of 1:10,000. However, recently, a polymorphic PK mutation (E277K) was found in some African populations, with heterozygote frequencies of 1–7%, suggesting that this may be another malaria-related polymorphism. The clinical picture of homozygous (or biallelic) PK deficiency is that of an HA that often presents in the newborn with neonatal jaundice; the jaundice persists, and it is often associated with reticulocytosis. The anemia is of variable severity; sometimes it is so severe as to require regular blood transfusion treatment, whereas sometimes it is mild, bordering on a nearly compensated hemolytic disorder. As a result, the diagnosis may be delayed: in some cases it is made, for instance, in a young woman during her first pregnancy, when the anemia may get worse. The delay in diagnosis may be caused in part by the fact that the anemia is often remarkably well tolerated, because the metabolic block at the last step in glycolysis causes an increase in 2,3-bisphosphoglycerate (or DPG; Fig. 96-1), a major effector of the hemoglobin-oxygen dissociation curve; thus, the oxygen delivery to the tissues is enhanced, a remarkable compensatory feat.

Treatment

TREATMENT

Pyruvate Kinase Deficiency

The management of PK deficiency is mainly supportive. In view of the marked increase in red cell turnover, oral folic acid supplements should be given constantly. Blood transfusion should be used as necessary, and iron chelation may be required even in some patients who, though not receiving blood transfusion, may be developing iron overload (see "General Pathophysiology" above). In patients who have more severe disease splenectomy may be beneficial, as the anemia improves (paradoxically, reticulocytes often increase considerably). There is a single case report of curative treatment of PK deficiency by bone marrow transplantation (BMT) from an HLA-identical PK-normal sibling. This seems a viable option for severe cases when a sibling donor is available. Prenatal diagnosis has been carried out in a mother who had already had an affected child. A clinical trial of a small molecule that is a specific PK ligand and may increase the stability and/or catalytic efficiency of mutant PK is currently on-going. Rescue of inherited PK deficiency through lentiviral-mediated human PK gene transfer has been successful in mice.

Other Glycolytic Enzyme Abnormalities

All of these defects are rare to very rare (Table 96-4), and most of them cause HA with varying degrees of severity. It is not unusual for the presentation to be in the guise of severe neonatal jaundice, which may require exchange transfusion; if the anemia is less severe, it may present later in life, or it may even remain asymptomatic and be detected incidentally when a blood count is done for unrelated reasons. The spleen is often enlarged. When other systemic manifestations occur, they can involve the central nervous system (sometimes entailing severe mental retardation, particularly in the case of triose phosphate isomerase deficiency), the neuromuscular system, or both (see Table 96-4). This is not altogether surprising, if we consider that these are housekeeping genes, i.e., expressed in all tissues. The *diagnosis* of HA is usually not difficult, thanks to the triad of normomacrocytic anemia, reticulocytosis, and hyperbilirubinemia. Enzymopathies should be considered in the differential diagnosis of any chronic Coombs-negative HA. Unlike with membrane disorders, in most cases of glycolytic enzymopathies morphologic abnormalities are conspicuous by their absence. A definitive diagnosis can be made only by demonstrating the deficiency of an individual enzyme by quantitative assays; these are carried out in only a few specialized laboratories. If a particular molecular abnormality is already known in the family, then one could test directly for that defect at the DNA level, thus bypassing the need for enzyme assays. Of course the time may be getting nearer when a patient will present with her or his exome already sequenced, and we will need to concentrate on which genes to look up within the file. The principles for the management of these conditions are similar as for PK deficiency. In isolated cases of glycolytic enzyme abnormalities BMT has been carried out successfully: although unfortunately non-hematologic manifestations, if any, are not reversed.

ABNORMALITIES OF REDOX METABOLISM

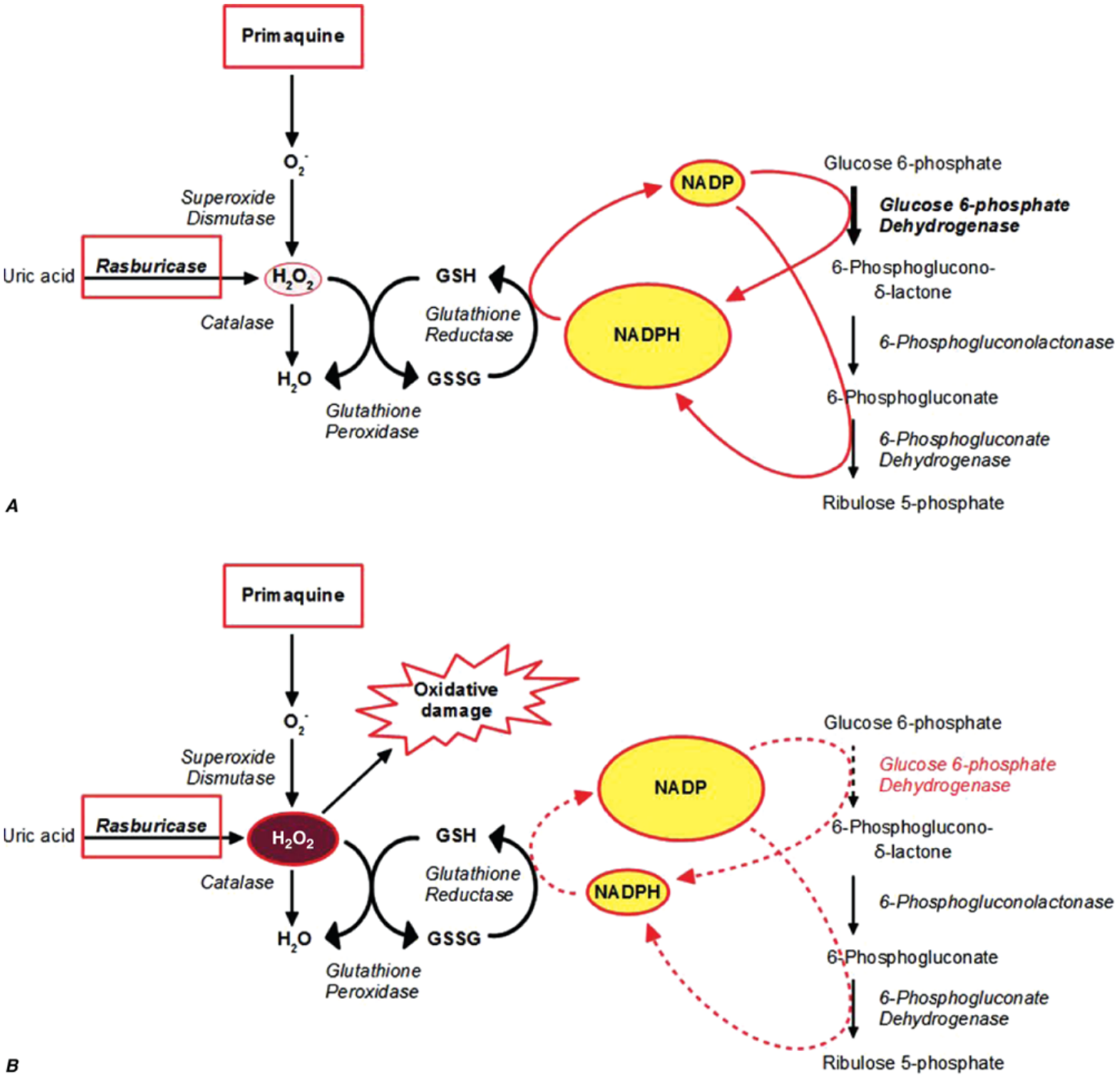
Glucose 6-Phosphate Dehydrogenase (G6PD) Deficiency

G6PD is a housekeeping enzyme critical in the redox metabolism of all aerobic cells (Fig. 96-1). In red cells, its role is even more critical because it is the only source of NADPH, which directly and via glutathione (GSH) defends these cells against oxidative stress **(Fig. 96-5)**. G6PD deficiency-related HA is a

prime example of an HA due to interaction between an intracorpuscular cause and an extracorpuscular cause because in the majority of cases hemolysis is triggered by an exogenous agent. Although a decrease in G6PD activity is present in most tissues of G6PD-deficient subjects, in other cells the decrease is much less pronounced than in red cells, and it does not seem to impact on clinical expression.

FIGURE 96-5

The role of G6PD in protecting red cells from oxidative damage. ***A.*** In G6PD-normal red cells, G6PD and 6-phosphogluconate dehydrogenase – two of the enzymes of the pentose phosphate pathway—provide ample supply of NADPH, which in turn regenerates GSH when this is oxidized by reactive oxygen species (e.g., O_2^- and H_2O_2). O_2^- is one of the most reactive oxygen species that can be generated from the metabolism of pro-oxidant compounds such as primaquine; rasburicase, on the other hand, produces directly hydrogen peroxide in equimolar amount to uric acid degraded. ***B.*** In G6PD-deficient red cells, where the enzyme activity is reduced, NADPH production is limited, and it may not be sufficient to cope with the excess of reactive oxygen species generated in the presence of pro-oxidant compounds. This diagram also explains why a defect in glutathione reductase has very similar consequences to G6PD deficiency.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GENETIC CONSIDERATIONS



The *G6PD* gene is X-linked, and this has important implications. First, because males have only one *G6PD* gene (i.e., they are hemizygous for this gene), they must be either normal or G6PD deficient. By contrast, females, who have two *G6PD* genes, can be either normal or deficient (homozygous) or intermediate (heterozygous). Second, as a result of the phenomenon of X chromosome inactivation, heterozygous females are genetic mosaics, with a highly variable ratio of G6PD-normal to G6PD-deficient cells and an equally variable degree of clinical expression; some heterozygotes can be just as affected as hemizygous males. The enzymatically active form of G6PD is either a dimer or a tetramer of a single protein subunit of 514 amino acids. G6PD-deficient subjects have been found invariably to have mutations in the coding region of the *G6PD* gene (Fig. 96-5). Almost all of the nearly 200 different mutations known are single missense point mutations, entailing single amino acid replacements in the G6PD protein. In most cases, these mutations cause G6PD deficiency by decreasing the in vivo stability of the protein; thus, the physiologic decrease in G6PD activity that takes place with red cell aging is greatly accelerated. In some cases, an amino acid replacement can also affect the catalytic function of the enzyme.

Among these mutations, those underlying chronic nonspherocytic hemolytic anemia (CNSHA; see below) are a discrete subset. This much more severe clinical phenotype can be ascribed in some cases to adverse qualitative changes (for instance, a decreased affinity for the substrate, glucose 6-phosphate) or simply to the fact that the enzyme deficit is more extreme because of a more severe instability of the enzyme. For instance, a cluster of mutations map at or near the dimer interface, and clearly they compromise severely the formation of the dimer.

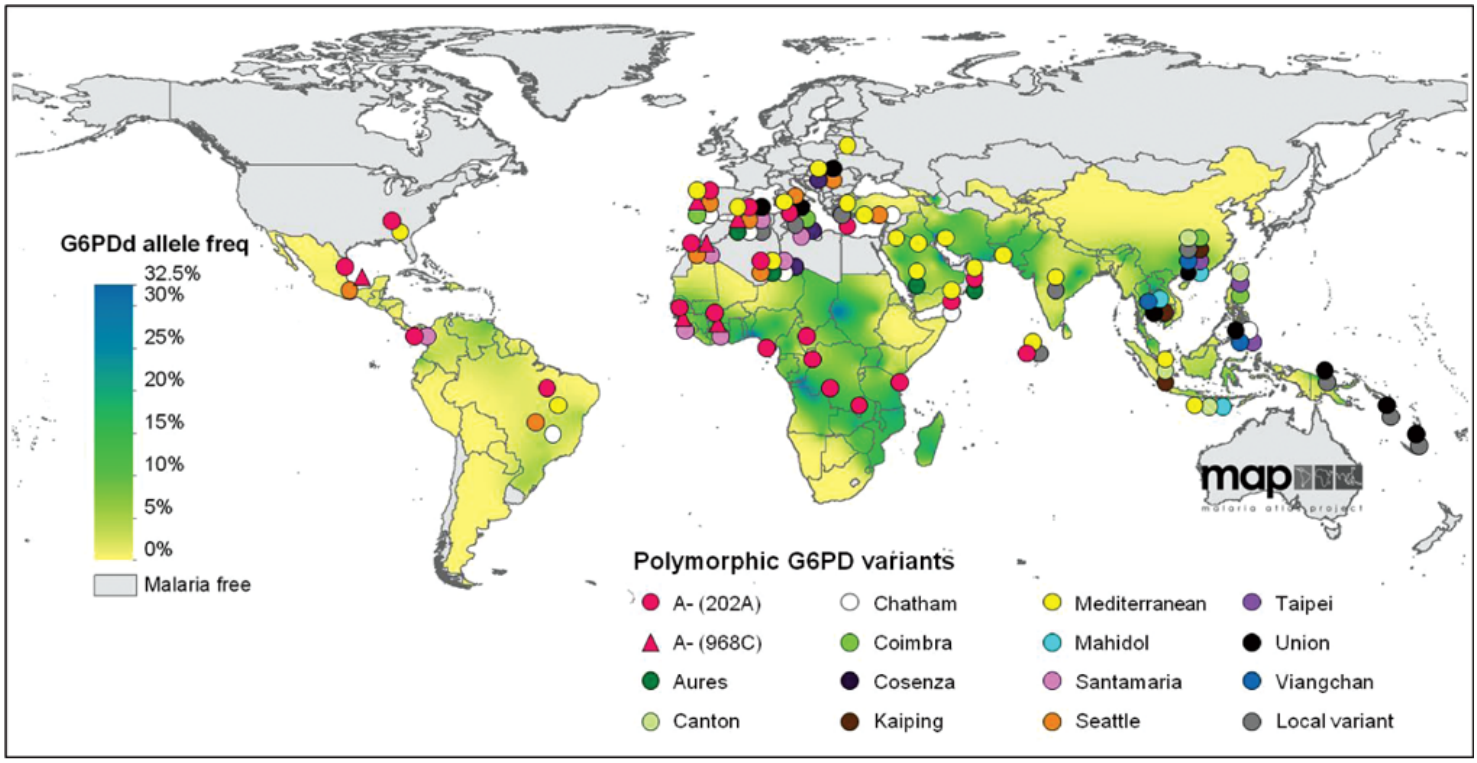
Epidemiology



G6PD deficiency is widely distributed in tropical and subtropical parts of the world (Africa, Southern Europe, the Middle East, Southeast Asia, and Oceania) **(Fig. 96-6)** and wherever people from those areas have migrated. A conservative estimate is that at least 400 million people have a *G6PD* deficiency gene. In several of these areas, the frequency of a *G6PD* deficiency gene may be as high as 20% or more. It would be quite extraordinary for a trait that causes significant pathology to spread widely and reach high frequencies in many populations without conferring some biologic advantage. Indeed, G6PD is one of the best-characterized examples of genetic polymorphisms in the human species. Clinical field studies and in vitro experiments strongly support the view that G6PD deficiency has been selected by *Plasmodium falciparum* malaria because it confers a relative resistance against this highly lethal infection. As in other cases of balanced polymorphism, it is heterozygotes, therefore females, who are protected. Different G6PD variants underlie G6PD deficiency in different parts of the world. Some of the more widespread variants are G6PD Mediterranean on the shores of that sea, in the Middle East, and in India; G6PD A– in Africa, in the Middle East, and in Southern Europe; G6PD Vianchan and G6PD Mahidol in Southeast Asia; G6PD Canton in China; and G6PD Union worldwide. The heterogeneity of polymorphic G6PD variants is proof of their independent origin, and it supports the notion that they have been selected by a common environmental agent, in keeping with the concept of convergent evolution (Fig. 96-6).

FIGURE 96-6

Epidemiology of glucose 6-phosphate dehydrogenase (G6PD) deficiency throughout the world. Color shades on the map indicate the median allele frequency of G6PD deficiency in malaria endemic and malaria-eliminating countries, according to a geostatistical model. Each colored circle illustrates the geographic distribution of one polymorphic G6PD allele present in more than one population. Dark grey circles indicate “local” polymorphic variants that have been detected only in one population. *(From L Luzzatto et al: Hematology/Oncology Clinics of North America, 30:373, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical Manifestations

The vast majority of people with G6PD deficiency remain clinically asymptomatic throughout their lifetime; however, all of them have an increased risk of developing neonatal jaundice (NNJ) and a risk of developing acute HA (AHA) when challenged by a number of oxidative agents. NNJ related to G6PD deficiency is very rarely present at birth; the peak incidence of clinical onset is between day 2 and day 3, and in most cases, the anemia is not severe. However, NNJ can be very severe in some G6PD-deficient babies, especially in association with prematurity, infection, and/or environmental factors (such as naphthalene-camphor balls, which may be used in babies' bedding and clothing); and the risk of severe NNJ is also increased by the coexistence of a monoallelic or biallelic mutation in the uridyl transferase gene (*UGT1A1*; the same mutations are associated with Gilbert's syndrome). If inadequately managed, NNJ associated with G6PD deficiency can produce kernicterus and permanent neurologic damage.

AHA can develop as a result of three types of triggers: (1) fava beans, (2) infections, and (3) drugs **(Table 96-5)**. Typically, a hemolytic attack starts with malaise, weakness, and abdominal or lumbar pain. After an interval of several hours to 2–3 days, the patient develops jaundice and often dark urine. The onset can be extremely abrupt, especially with favism in children. The anemia is moderate to extremely severe, usually normocytic and normochromic, and due partly to intravascular hemolysis; hence, it is associated with hemoglobinemia, hemoglobinuria, high LDH, and low or absent plasma haptoglobin. The blood film shows anisocytosis, polychromasia, and spherocytes; in addition, the most typical feature of G6PD deficiency is the presence of bizarre poikilocytes, with red cells that appear to have unevenly distributed hemoglobin ("hemighosts") and red cells that appear to have had parts of them bitten away ("bite cells" or "blister cells") **(Fig. 96-7)**. A classical test, now rarely carried out, is supravital staining with methyl violet, which, if done promptly, reveals the presence of Heinz bodies (consisting of precipitates of denatured hemoglobin and hemichromes), which are regarded as a signature of oxidative damage to red cells (they are also seen with unstable hemoglobins). Not only LDH is high; also unconjugated bilirubin is high, indicating that there is also extravascular hemolysis. The most serious threat from AHA in adults is the development of acute renal failure (this is exceedingly rare in children). Once the threat of acute anemia is over and in the absence of comorbidity, full recovery from AHA associated with G6PD deficiency is the rule.

TABLE 96-5

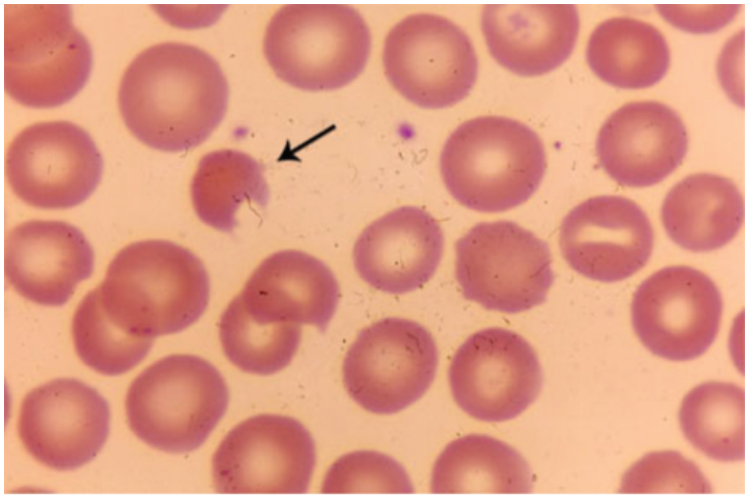
Drugs That Carry Risk of Clinical Hemolysis in Persons with Glucose 6-Phosphate Dehydrogenase Deficiency

	Definite Risk	Possible Risk	Doubtful Risk
Antimalarials	Primaquine	Chloroquine	Quinine
	Dapsone/chlorproguanil[a]		
Sulphonamides/sulphones	Sulfamethoxazole	Sulfasalazine	Sulfisoxazole
	Others	Sulfadimidine	Sulfadiazine
	Dapsone		
Antibacterial/antibiotics	Cotrimoxazole	Ciprofloxacin	Chloramphenicol
	Nalidixic acid	Norfloxacin	*p*-Aminosalicylic acid
	Nitrofurantoin		
	Niridazole		
Antipyretic/analgesics	Acetanilide	Acetylsalicylic acid high dose (>3 g/d)	Acetylsalicylic acid (<3 g/d)
	Phenazopyridine		Acetaminophen
			Phenacetin
Other	Naphthalene	Vitamin K analogues	Doxorubicin
	Methylene blue	Ascorbic acid (>1 g)	Probenecid
	Rasburicase		

[a]Marketed as Lapdap from 2003 to 2008.

FIGURE 96-7

Peripheral blood smear from a glucose 6-phosphate dehydrogenase (G6PD)-deficient boy experiencing hemolysis. Note the red cells that are misshapen and called "bite" cells. *(From MA Lichtman et al: Lichtman's Atlas of Hematology: http://www.accessmedicine.com.ezproxy1.library.usyd.edu.au. Copyright © The McGraw-Hill Companies, Inc. All rights reserved.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

It was primaquine (PQ)-induced AHA that led to the discovery of G6PD deficiency, but this drug has not been very prominent subsequently because it is not necessary for the treatment of life-threatening *P. falciparum* malaria. Today there is a revival of interest in PQ because it is the only effective agent for eliminating the gametocytes of *P. falciparum* (thus preventing further transmission) and for eliminating the hypnozoites of *Plasmodium vivax* (thus preventing endogenous relapse). In countries aiming to eliminate malaria, there may be a call for mass administration of PQ; this ought to be associated with G6PD testing. At the other end of the historic spectrum, the latest addition to the list of potentially hemolytic drugs (Table 96-5) is rasburicase; again G6PD testing ought to be made mandatory before giving this drug because fatal cases have been reported in newborns with kidney injury and in adults with tumor lysis syndrome.

Although drug-induced AHA has been prominent in the study of G6PD deficiency, the commonest clinical manifestations are in fact NNJ and favism, both of which are of public health importance in many populations. Contrary to beliefs that are still widespread, fava bean pollen inhalation does not cause favism, and other beans are safe.

A very small minority of subjects with G6PD deficiency have CNSHA of variable severity. The patient is nearly always a male, usually with a history of NNJ, who may present with anemia, unexplained jaundice, or gallstones later in life. The spleen may be enlarged. The severity of anemia ranges in different patients from borderline to transfusion dependent. The anemia is usually normomacrocytic, with reticulocytosis. Bilirubin and LDH are increased. Although hemolysis is, by definition, chronic in these patients, they are also vulnerable to acute oxidative damage, and therefore the same agents that can cause AHA in people with the ordinary type of G6PD deficiency will cause severe exacerbations in people with CNSHA associated with G6PD deficiency. In some cases of CNSHA, the deficiency of G6PD is so severe in granulocytes that it becomes rate-limiting for their oxidative burst, with consequent increased susceptibility to some bacterial infections.

Laboratory Diagnosis

The suspicion of G6PD deficiency can be confirmed by semi-quantitative methods often referred to as screening tests, which are suitable for population studies and can correctly classify male subjects, in the steady state, as G6PD normal or G6PD-deficient. However, in clinical practice, a diagnostic test is usually needed when the patient has had a hemolytic attack; this implies that the oldest, most G6PD-deficient red cells have been selectively destroyed, and young red cells, having higher G6PD activity, are being released into the circulation. Under these conditions, only a quantitative test can give a definitive result. In males, this test will identify normal hemizygotes and G6PD-deficient hemizygotes; among females, some heterozygotes will be missed, but those who are at most risk of hemolysis will be identified. Of course, G6PD deficiency also can be diagnosed by DNA testing. Currently easy-to-use “point of care” tests for G6PD deficiency are becoming available, geared especially to the prospect of mass administration of PQ or other anti-malarials.

Treatment

TREATMENT

G6PD Deficiency

The AHA of G6PD deficiency is largely preventable by avoiding exposure to triggering factors of previously screened subjects. Of course, the practicability and cost-effectiveness of screening depend on the prevalence of G6PD deficiency in each individual community. Favism is entirely preventable in G6PD-deficient subjects by not eating fava beans. Drug-induced hemolysis can be prevented by testing for G6PD deficiency before prescribing; in many cases one can use alternative drugs. When AHA develops and once its cause is recognized, no specific treatment is needed in most cases. However, if the anemia is severe, it may be a medical emergency, especially in children, requiring immediate action, including blood transfusion. This has been the case with an antimalarial drug combination containing dapsone (called Lapdap, introduced in 2003) that has caused severe acute hemolytic episodes in children with malaria in several African countries; after a few years, the drug was taken off the market. If there is acute renal

failure, hemodialysis may be necessary, but if there is no previous kidney disease, recovery is the rule. The management of NNJ associated with G6PD deficiency is no different from that of NNJ due to other causes.

In cases with cnsha, if the anemia is not severe, regular folic acid supplements and regular hematologic surveillance will suffice. It will be important to avoid exposure to potentially hemolytic drugs, and blood transfusion may be indicated when exacerbations occur, mostly in concomitance with intercurrent infection. In rare patients, regular blood transfusions may be required, in which case appropriate iron chelation should be instituted. Unlike in HS, there is no evidence of selective red cell destruction in the spleen; however, in practice, splenectomy has proven beneficial in severe cases.

Other Abnormalities of the Redox System

As mentioned previously, GSH is a key player in the defense against oxidative stress. Inherited defects of GSH metabolism are exceedingly rare, but each one can give rise to chronic HA (Table 96-4). A rare, peculiar, severe but usually self-limited HA occurring in the first month of life, called *infantile poikilocytosis*, may be associated with deficiency of glutathione peroxidase (GSHPX) due not to an inherited abnormality, but to transient nutritional deficiency of selenium, an element essential for the activity of GSHPX.

PYRIMIDINE 5′-NUCLEOTIDASE (P5N) DEFICIENCY

P5N is a key enzyme in the catabolism of nucleotides arising from the degradation of nucleic acids that takes place in the final stages of erythroid cell maturation. How exactly its deficiency causes HA is not well understood, but a highly distinctive feature of this condition is a morphologic abnormality of the red cells known as *basophilic stippling*. The condition is rare, but it probably ranks third in frequency among red cell enzyme defects (after G6PD deficiency and PK deficiency). The anemia is lifelong, of variable severity, and may benefit from splenectomy.

Familial (Atypical) Hemolytic-Uremic Syndrome (aHUS)

This term is used to designate a group of rare disorders, mostly affecting children, characterized by microangiopathic HA with presence of fragmented erythrocytes in the peripheral blood smear, thrombocytopenia (usually mild), and acute renal failure. (The word *atypical* in this phrase should be consigned to history: it was introduced originally to distinguish this condition from the hemolytic-uremic syndrome [HUS] caused by infection with *Escherichia coli* producing the Shiga toxin, regarded as *typical*.) The genetic basis of atypical HUS (aHUS) has been elucidated. Studies of >100 families have revealed that those family members who developed HUS had mutations in any one of several genes encoding complement regulatory proteins: complement factor H (*CFH*), CD46 or membrane cofactor protein (*MCP*), complement factor I (*CFI*), complement component C3, complement factor B (*CFB*), thrombomodulin, and others. Thus, whereas all other inherited HAs are due to intrinsic red cell abnormalities, this group is unique in that hemolysis results from an inherited defect external to red cells (Table 96-1). Because the regulation of the complement cascade has considerable redundancy, in the steady state, any of the above abnormalities can be tolerated. However, when an intercurrent infection or some other trigger briskly activates complement the deficiency of one of the complement regulators becomes critical. Endothelial cells get damaged, especially in the kidney; at the same time, and partly as a result of this, there will be brisk hemolysis (thus, the more common Shiga toxin–related HUS **(Chap. 161)** can be regarded as a phenocopy of aHUS). aHUS is a severe disease, with up to 15% mortality in the acute phase and up to 50% of cases progressing to end-stage renal disease (ESRD). Not infrequently, aHUS undergoes spontaneous remission; but because its basis is an inherited abnormality, it is not surprising that, given renewed exposure to a trigger, the syndrome will tend to recur; when it does, the prognosis is always serious. The traditional treatment has been plasma exchange, which will supply the deficient complement regulator. This has changed since the introduction of the anti-C5 complement inhibitor eculizumab (see "Paroxysmal Nocturnal Hemoglobinuria") was found to greatly ameliorate the microangiopathic picture, with improvement in platelet counts and in renal function, thus abrogating the need for plasma exchange, which is not always effective and not free of complications. Since the basis of aHUS is genetic, and even after complete remission relapses are always possible, there is a rationale for continuing eculizumab indefinitely, especially in order to prevent ESRD. Patients who relapsed after discontinuing eculizumab have responded again. Discontinuation of eculizumab might be reasonable especially in patients heterozygous for a MCP mutation. However, there is no evidence base at the moment for balancing the pros and cons of lifetime eculizumab (a very expensive drug).

ACQUIRED HEMOLYTIC ANEMIA

Mechanical Destruction of Red Cells

Although red cells are characterized by the remarkable deformability that enables them to squeeze through capillaries narrower than themselves for thousands of times in their lifetime, there are at least two situations in which they succumb to shear, if not to wear and tear; the result is intravascular hemolysis, resulting in hemoglobinuria **(Table 96-6)**. One situation is acute and self-inflicted, *march hemoglobinuria*. Why sometimes a marathon runner may develop this complication, whereas on another occasion, this does not happen, we do not know (perhaps her or his footwear needs attention). A similar syndrome may develop after prolonged barefoot ritual dancing or intense playing of bongo drums. The other situation is chronic and iatrogenic (it has been called *microangiopathic hemolytic anemia*). It takes place in patients with prosthetic heart valves, especially when paraprosthetic regurgitation is present. If the hemolysis consequent on mechanical trauma to the red cells is mild, and if the supply of iron is adequate, the loss may be largely compensated; if more than mild anemia develops, reintervention to correct regurgitation may be required.

TABLE 96-6

Diseases and Clinical Situations in Which Hemolysis Is Largely Intravascular

	Onset/Time Course	Main Mechanism	Appropriate Diagnostic Procedure	Comments
Mismatched blood transfusion	Abrupt	Nearly always ABO incompatibility	Repeat cross-match	
Paroxysmal nocturnal hemoglobinuria (PNH)	Chronic with acute exacerbations	Complement (C)-mediated destruction of CD59(−) red cells	Flow cytometry to display a CD59(−) red cell population	Exacerbations due to C activation through any pathway
Paroxysmal cold hemoglobinuria (PCH)	Acute	Immune lysis of normal red cells	Test for Donath-Landsteiner antibody	Often triggered by viral infection
Septicemia	Very acute	Exotoxins produced by *Clostridium perfringens*	Blood cultures	Other organisms may be responsible
Microangiopathic	Acute or chronic	Red cell fragmentation	Red cell morphology on blood smear	Different causes ranging from endothelial damage to hemangioma to leaky prosthetic heart valve
March hemoglobinuria	Abrupt	Mechanical destruction	Targeted history taking	
Favism	Acute	Destruction of older fraction of G6PD-deficient red cells	G6PD assay	Triggered by ingestion of large dish of fava beans[a]

[a]The trigger of acute hemolytic anemia, often with hemoglobinuria, can be infection or a drug (see Table 96-5) rather than fava beans.

Abbreviation: G6PD, glucose 6-phosphate dehydrogenase.

Infection

By far the most frequent infectious cause of HA, in endemic areas, is malaria **(Chap. 219)**. In other parts of the world, the most frequent direct cause is probably Shiga toxin–producing *E. coli* O157:H7, now recognized as the main etiologic agent of HUS, which is more common in children than in adults **(Chap. 156)**. Life-threatening intravascular hemolysis, due to a toxin with lecithinase activity, occurs with *Clostridium perfringens* sepsis, particularly following open wounds, septic abortion, or as a disastrous accident due to a contaminated blood unit. Rarely, and if at all in children, HA is seen with sepsis or endocarditis from a variety of organisms. In addition, bacterial and viral infections can cause HA by indirect mechanisms (see previous section on G6PD deficiency and Table 96-6).

Immune Hemolytic Anemias

These can arise through at least two distinct mechanisms. First, when an antibody directed against a certain molecule (e.g., a drug) reacts with that molecule, red cells may get caught in the reaction (the so-called *innocent bystander* mechanism: see section below on Hemolytic Anemia from Toxic Agents and Drugs), whereby they are damaged or destroyed. Second, and more frequently, a true auto-antibody is directed against a red cell antigen, i.e., a molecule present on the surface of red cells.

AUTOIMMUNE HEMOLYTIC ANEMIA (AIHA)

This latter mechanism is common to a group of rare disorders (AIHA), with an estimated incidence in the United States of about $2/10^5$/year. AIHA can be serious, since even with appropriate management the mortality is of the order of 5–10%.

Clinical Features

The onset is often abrupt and can be dramatic. The hemoglobin level may drop, within days, to as low as 4 g/dL; the massive red cell removal will produce jaundice; and sometimes the spleen is enlarged. When this triad is present, the suspicion of AIHA must be high. When hemolysis is (in part) intravascular, the telltale sign will be hemoglobinuria, which the patient may report or about which we must enquire or test for.

There are few situations in hematology where one laboratory test is so informative as the direct antiglobulin test developed in 1945 by R. R. A. Coombs, and known since then by this name. The currently recommended version of this test uses in the first instance a "broad spectrum" reagent: i.e., one that will detect not only immunoglobulins (Ig) but also complement (C) components (usually C3 fragments) bound to the surface of the patient's red cells. If the test is positive (and barring special circumstances such as previous blood transfusion), it is practically diagnostic of AIHA; and one can then determine, by using specific reagents, whether Ig or C or both are implicated. The sensitivity of the Coombs test varies depending on the techniques that are used: in general, the test is positive if there are an average of at least 400 molecules of Ig and/or C on each red cell; but with more advanced techniques the sensitivity can be pushed to as low as 40 molecules per red cell: therefore liaison with a specialized laboratory is desirable. In the past the diagnosis of "Coombs-negative AIHA" was regarded as a last resort, but it is important to know that a patient with this label may have severe AIHA, because if the antibody is powerful (high affinity/avidity), few molecules may be sufficient to opsonize red cells. Based on the Coombs test findings as well as on the thermal characteristics and the antigenic specificities of the auto-antibodies (**Table 96-7**), AIHA has been classified into subtypes.

TABLE 96-7

Classification of Acquired Immune Hemolytic Anemias

Clinical Setting	Type of Antibody	
	Cold, Mostly IgM, Optimal Temperature 4°C–30°C	**Warm, Mostly IgG, Optimal Temperature 37°C; or Mixed**
Primary	CAD	AIHA (idiopathic)
Secondary to viral infection	EBV CMV Other	HIV Viral vaccines
Secondary to other infection	Mycoplasma infection: paroxysmal cold hemoglobinuria	
Secondary to/associated with other disease	CAD in: Waldenström's disease Lymphoma	AIHA in: SLE CLL Other malignancy Chronic inflammatory disorders (e.g., IBD) After allogeneic HSCT After immune check-point modulating drugs
Secondary to drugs: drug-induced immune hemolytic anemia	Small minority (e.g., with lenalidomide)	Majority: currently most common culprit drugs are cefotetan, ceftriaxone, piperacillin
	Drug-dependent: antibody destroys red cells only when drug present (e.g., rarely penicillin)	
	Drug-independent: antibody can destroy red cells even when drug no longer present (e.g., methyldopa)	

Abbreviations: AIHA, autoimmune hemolytic anemia; CAD, cold agglutinin disease; CLL, chronic lymphocytic leukemia; CMV, cytomegalovirus; EBV, Epstein-Barr virus; HIV, human immunodeficiency virus; HSCT, hematopoietic stem cell transplantation; IBD, inflammatory bowel disease; SLE, systemic lupus erythematosus.

Warm Antibody AIHA

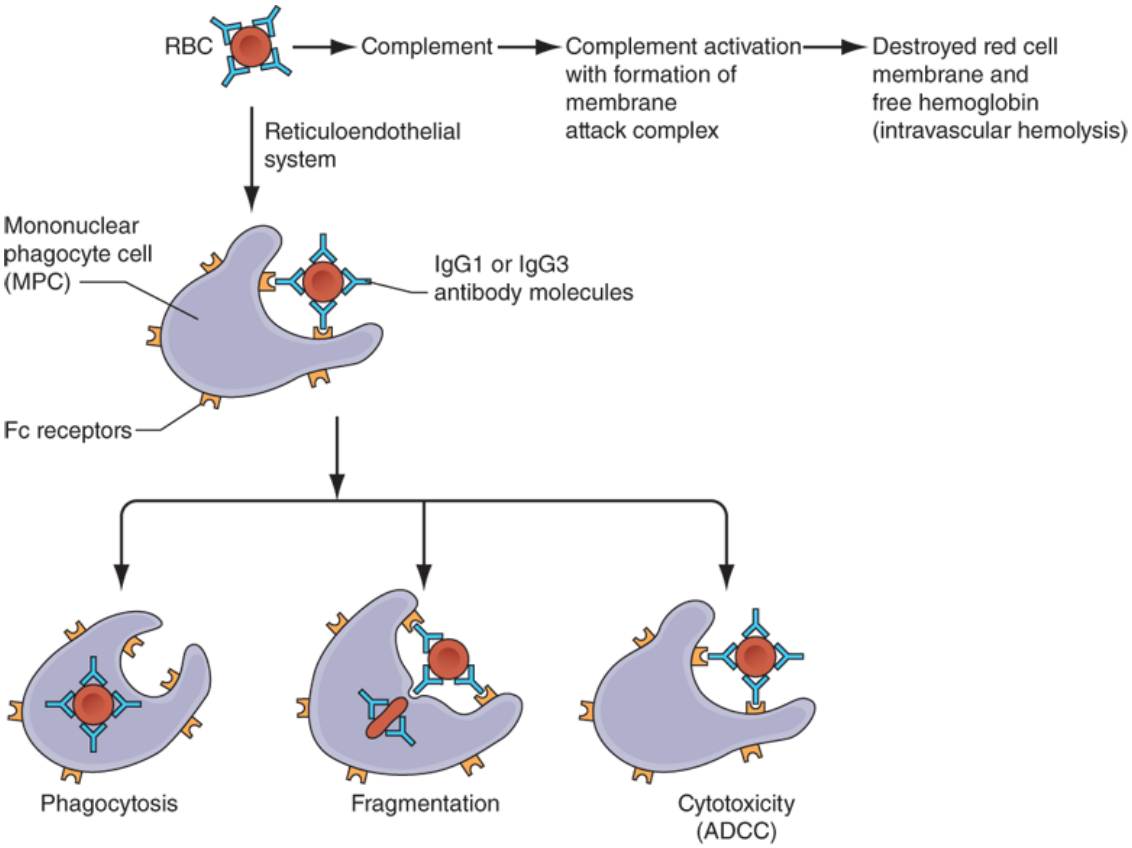
This is the more common type of AIHA. As the name suggests, the auto-antibody reacts best at 37°C: it will often react with most red cells, but it is usually Rhesus-specific (sometimes specifically anti-e). Warm antibody AIHA may be seen in isolation (and it is then called *idiopathic*) or as part of a systemic auto-immune disorder such as systemic lupus erythematosus (SLE: sometimes AIHA may be the first manifestation that leads to a diagnosis of SLE). Like all auto-immune diseases, AIHA must arise from a dysregulation of immunity. It is therefore not surprising that it is increasingly being recognized in chronic lymphocytic leukemia (CLL), whether treated or untreated; after BMT; and after solid organ transplantation entailing immuno-

suppressive treatment. Recently, warm antibody AIHA has also occurred as a side effect of the use of immune checkpoint inhibitors, such as nivolumab, in patients with various types of cancer.

Once a red cell is coated by an autoantibody it will be destroyed by one or more mechanisms. In most cases, the Fc portion of the antibody will be recognized by the Fc receptor of macrophages, and this will trigger erythrophagocytosis. Thus, destruction of red cells will take place wherever macrophages are abundant, i.e., in the spleen, liver, and bone marrow (*extravascular hemolysis* **see Fig. 96-8**). Because of the special anatomy of the spleen, this organ is particularly efficient in trapping antibody-coated red cells, and often this is the predominant site of red cell destruction. In some cases, the nature of the antibody is such (usually an IgM antibody) that the antigen-antibody complex on the surface of red cells is able to activate complement (C); as a result, a large amount of membrane attack complex (MAC) will form, and the red cells may be destroyed directly (*intravascular hemolysis*).

FIGURE 96-8

Mechanism of antibody-mediated immune destruction of red blood cells (RBCs). The three bottom images illustrate three different modalities of extravascular hemolysis. ADCC, antibody-dependent cell-mediated cytotoxicity. *(From N Young et al: Clinical Hematology. Philadelphia, Elsevier, 2006; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The hematological picture of AIHA includes in most cases reticulocytosis, as the bone marrow responds to anemia: but in some cases reticulocytes may not be increased because they themselves are attacked by the auto-antibody, and this may signify the disease is more severe. In some cases AIHA can be associated, on first presentation or subsequently, with autoimmune thrombocytopenia (Evans' syndrome): this too usually signals severe disease. Evans' syndrome may be a manifestation of common variable immune deficiency, and in children it may suggest one of several primary immune deficiency syndromes.

Treatment

TREATMENT

Warm Antibody Autoimmune Hemolytic Anemia

Severe acute AIHA can be a medical emergency. The immediate treatment almost invariably includes transfusion of red cells. This may pose a special problem because many or all of the blood units cross-matched may be incompatible. In these cases, it is often correct, if paradoxical, to transfuse ABO-matched but incompatible blood: the rationale being that the transfused red cells will be destroyed no less—but no more—than the patient's own red cells, and in the meantime the patient stays alive. A situation like this requires close liaison and understanding between the clinical unit treating the patient and the blood transfusion/serology lab. Whenever the anemia is not immediately life threatening, blood transfusion should be withheld

(because compatibility problems may increase with each unit of blood transfused), and medical treatment started immediately with prednisone (1 mg/kg per day), which will produce a remission promptly in at least one-half of patients. Rituximab (anti-CD20), previously regarded as second-line treatment, is increasingly being used at a relatively low dose (100 mg/wk × 4), together with prednisone as part of first-line treatment. It is especially encouraging that this approach seems to reduce the rate of relapse, a common occurrence in AIHA. For patients who do relapse or are refractory to medical treatment, one may have to consider splenectomy: this procedure does not cure the disease, but it can produce significant benefit by removing a major site of hemolysis, thus improving the anemia and/or reducing the need for other therapies (e.g., the dose of prednisone); of course splenectomy is not free of risk, as it entails increased risk of sepsis and of thrombosis. Since the introduction of rituximab, azathioprine, cyclophosphamide, cyclosporine, and intravenous immunoglobulin have become second- or third-line agents. In very rare severe refractory cases, one may have to consider myelo-immuno-ablative chemotherapy followed by rescue with either autologous or allogeneic hematopoietic stem cell transplantation.

PAROXYSMAL COLD HEMOGLOBINURIA (PCH)

PCH is a rather rare form of AIHA occurring mostly in children, usually triggered by a viral infection, usually self-limited, and characterized by the involvement of the so-called Donath-Landsteiner antibody. In vitro, this antibody has unique serologic features; it has anti-P specificity and binds to red cells only at a low temperature (optimally at 4°C), but when the temperature is shifted to 37°C, lysis of red cells takes place in the presence of complement. Consequently, in vivo there is intravascular hemolysis, resulting in hemoglobinuria. Clinically the differential diagnosis must include other causes of hemoglobinuria (Table 96-6), but the presence of the Donath-Landsteiner antibody will prove PCH. Active supportive treatment, including blood transfusion, may be needed to control the anemia; subsequently, recovery is the rule.

COLD AGGLUTININ DISEASE

This designation is used for a form of AIHA that usually affects the elderly and has special clinical and pathologic features. First, CAD is characteristically a chronic condition—in contrast to the abrupt onset of warm antibody AIHA. Second, the term *cold* refers to the fact that the autoantibody involved reacts with red cells poorly or not at all at 37°C, whereas it reacts strongly at lower temperatures. As a result, hemolysis is more prominent the more the body is exposed to the cold. The antibody is usually IgM; usually it has an anti-I specificity (the I antigen is present on the red cells of almost everybody), and it may have a very high titer (1:100,000 or more has been observed). Third, the antibody is produced by an expanded B lymohocyte clone (a low-grade mature B cell lymphoma): and sometimes the antibody concentration in the serum is high enough to show up as a spike in plasma protein electrophoresis, i.e., as a monoclonal gammopathy. Indeed, since we are dealing with a clonal disease and the antibody is IgM, CAD must be regarded as a form of Waldenström macroglobulinemia (**see Chap. 107**). The unique biologic properties of the IgM produced in a patient with CAD give the clinical picture of HA, often well before other clinical features of WM become manifest.

In mild forms of CAD, avoidance of exposure to cold may be all that is needed to enable the patient to have a reasonably comfortable quality of life; but in more severe forms, the management of CAD is not easy. Plasma exchange will remove antibody and is, therefore, in theory, a rational approach, but it is laborious and must be carried out at frequent intervals if it is to be beneficial. The management of CAD has changed significantly with the advent of the anti-CD20 antibody rituximab: up to 60% of patients respond. If remission is followed by relapse a new course of rituxumab may be again effective, and remissions may be more durable with a rituximab-fludarabine combination. Therefore, even in the absence of a formal trial, rituximab has become de facto first-line treatment: especially since previously used immunosuppressive/cytotoxic agents such as azathioprine or cyclophosphamide, although they can reduce the antibody titer, have limited clinical efficacy and, in view of the chronic nature of CAD, their side effects may prove unacceptable. Unlike in AIHA, prednisone and splenectomy are ineffective. In terms of supportive treatment blood transfusion may be helpful—in spite of the fact that red cells from the donor, being I-positive, will survive no longer than those of the patient: both the blood bag and the patient's extremities must be kept warm during transfusion.

Hemolytic Anemia from Toxic Agents and Drugs

A number of chemicals with oxidative potential, whether medicinal or not, can cause hemolysis even in people who are not G6PD-deficient (for which see above). Examples are hyperbaric oxygen (or 100% oxygen), nitrates, chlorates, methylene blue, dapsone, cisplatin, and numerous aromatic (cyclic) compounds. Other chemicals may be hemolytic through nonoxidative, largely unknown mechanisms; examples include arsine, stibine, copper, and lead. The HA caused by lead poisoning is characterized by basophilic stippling; it is in fact a phenocopy of that seen in P5N deficiency (see above), suggesting it is mediated at least in part by lead inhibiting this enzyme.

In these cases, hemolysis appears to be mediated by a direct chemical action on red cells. But drugs can cause hemolysis through at least two other mechanisms. (1) A drug can behave as a hapten and induce antibody production; in rare subjects, this happens, for instance, with penicillin. Upon a subsequent exposure, red cells are caught, as innocent bystanders, in the reaction between penicillin and anti-penicillin antibodies. Hemolysis will subside as soon as penicillin administration is stopped. (2) A drug can trigger, perhaps through mimicry, the production of an antibody against a red cell antigen. The best known example is methyldopa, an antihypertensive agent no longer in use, which in a small fraction of patients stimulated the production of the Rhesus antibody anti-e. In patients who have this antigen, the anti-e is a true autoantibody, which then causes true AIHA (see above). Usually this will gradually subside once methyldopa is discontinued.

Severe intravascular hemolysis can be caused by the venom of certain snakes (cobras and vipers), and HA can also follow spider bites.

Paroxysmal Nocturnal Hemoglobinuria (PNH)

PNH is an acquired chronic HA characterized by persistent intravascular hemolysis with occasional or frequent recurrent exacerbations. In addition to (i) hemolysis, there may be (ii) pancytopenia and (iii) a distinct tendency to venous thrombosis. This triad makes PNH a truly unique clinical condition; however, when not all of these three features are manifest on presentation, the diagnosis is often delayed, although it can always be made by appropriate laboratory investigations (see below).

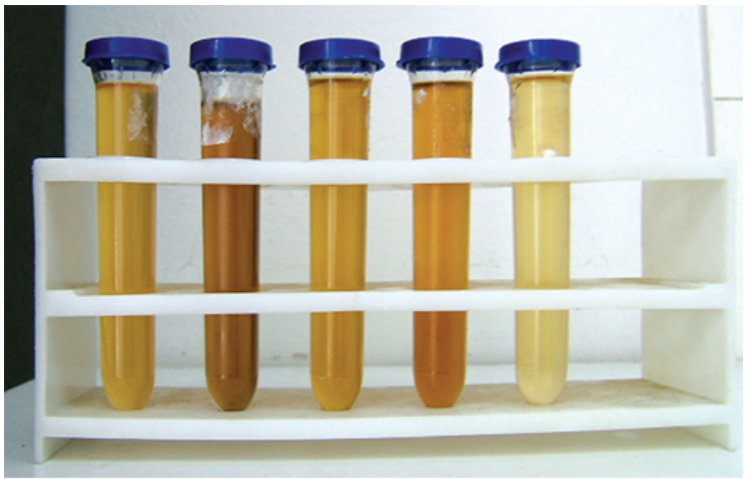
PNH is encountered in all populations throughout the world, but it is a rare disease, with an estimated prevalence of ~5 per million (it may be somewhat less rare in Southeast Asia and in the Far East). PNH has about the same frequency in men and women. PNH is not inherited, and it has never been reported as a congenital disease, but it can present in small children or as late as in the seventies, although most patients are young adults.

CLINICAL FEATURES

When seeking medical attention, the patient may report that one morning, she or he "passed blood instead of urine" **(Fig. 96-9)**. This distressing or frightening event may be regarded as the classic presentation; however, more frequently, this symptom is not noticed or not reported. Indeed, the patient often presents simply as a problem in the differential diagnosis of *anemia*, whether symptomatic or discovered incidentally. Sometimes the anemia is associated from the outset with neutropenia, thrombocytopenia, or both, thus signaling an element of bone marrow failure (see below). Some patients may present with recurrent attacks of severe abdominal pain eventually found to be related to thrombosis in abdominal veins, or attributable to NO depletion associated with intravascular hemolysis. When thrombosis affects the hepatic vein it may produce acute hepatomegaly and ascites, i.e., a full-fledged Budd-Chiari syndrome, which, in the absence of liver disease, ought to raise the suspicion of PNH.

FIGURE 96-9

Consecutive urine samples from a patient with paroxysmal nocturnal hemoglobinuria (PNH). The variation in the severity of hemoglobinuria within hours is probably unique to this condition.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The *natural history* of PNH can extend over decades. In the past, with supportive treatment only, the median survival was estimated to be about 10–20 years; with the most common cause of death being venous thrombosis, followed by infection secondary to severe neutropenia and hemorrhage secondary to severe thrombocytopenia. Rarely (estimated 1–2% of all cases), PNH may terminate in acute myeloid leukemia. On the other hand, full spontaneous recovery from PNH has been documented, albeit rarely.

LABORATORY INVESTIGATIONS AND DIAGNOSIS

The most consistent blood finding is anemia, which may range from mild to moderate to very severe. The anemia is usually normo-macrocytic, with unremarkable red cell morphology. If the MCV is high, it is usually largely accounted for by reticulocytosis, which may be quite marked (up to 20%, or up to 400,000/μL). The anemia may become microcytic if the patient is allowed to become iron-deficient as a result of chronic iron loss through hemoglobinuria. Unconjugated bilirubin is mildly or moderately elevated; LDH is typically markedly elevated (values in the thousands are common); and haptoglobin is usually undetectable. All of these findings make the diagnosis of HA compelling. Hemoglobinuria may be overt in a random urine sample; if it is not, it may be helpful to obtain serial urine samples (Fig. 96-9) because hemoglobinuria can vary dramatically from day to day and even from hour to hour. The bone marrow is usually cellular, with marked to massive erythroid hyperplasia, often with mild to moderate dyserythropoietic

features (these overlap with those seen in myelodysplastic syndromes, but PNH remains a separate entity). At some stage of the disease, the marrow may become hypocellular or even frankly aplastic (see below).

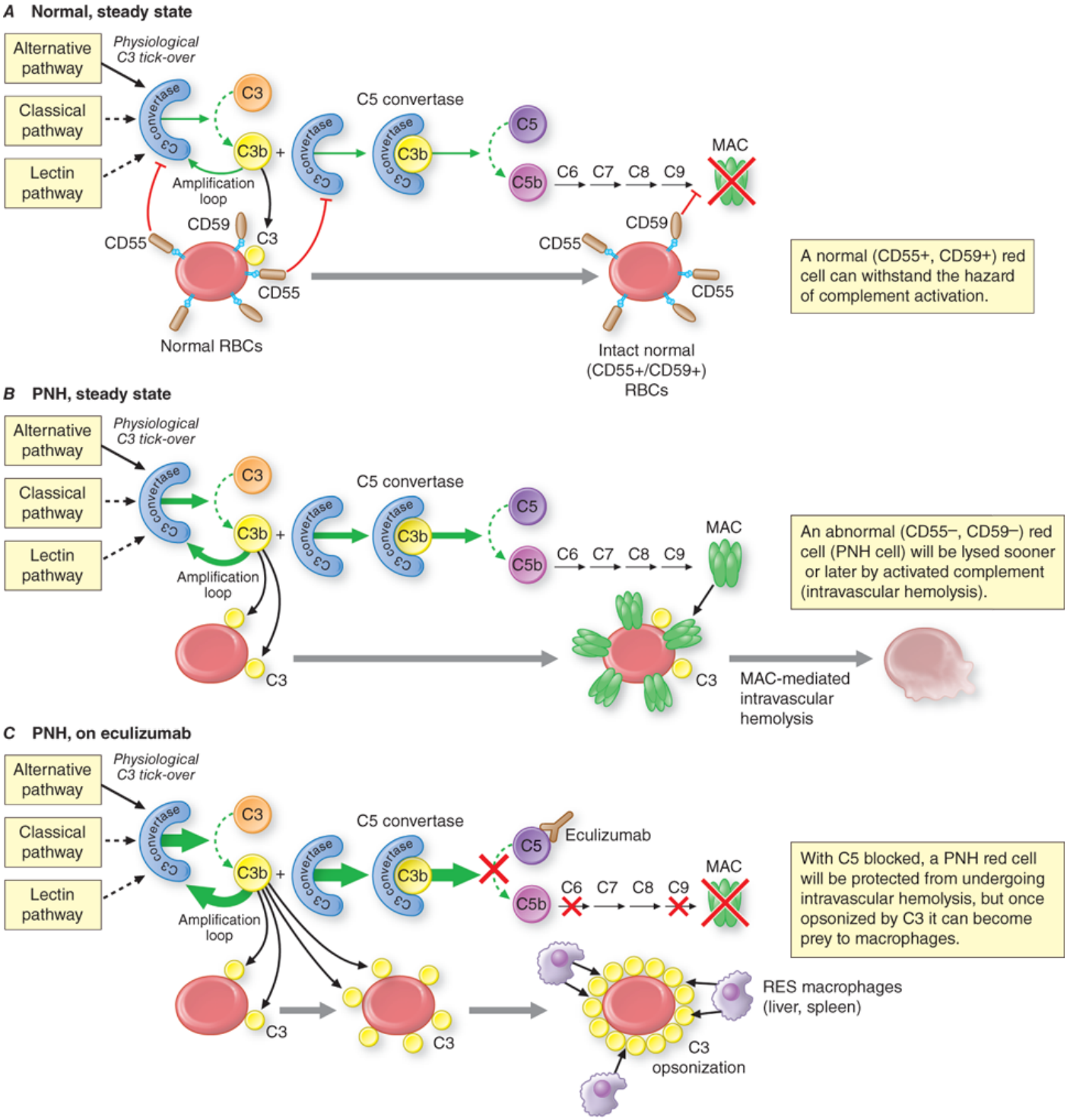
The definitive diagnosis of PNH must be based on the demonstration that a substantial proportion of the patient's red cells have an increased susceptibility to complement (C), due to the deficiency on their surface of proteins (particularly CD59 and CD55) that normally protect the red cells from activated C. The sucrose hemolysis test is unreliable; in contrast, the acidified serum (Ham) test is highly reliable but is carried out only in a few labs. The gold standard today is flow cytometry, which can be carried out on granulocytes as well as on red cells, and has a very high sensitivity. In PNH, characteristically, one sees a bimodal distribution of cells, with a discrete population that is CD59 and CD55 negative. Although very small populations of CD59(-) cells are of interest in terms of pathophysiology (particularly of aplastic anemia [AA]), no patient should be diagnosed with PNH unless their proportion is substantial: in first approximation at least 5% of the total red cells and at least 20% of the total granulocytes.

PATHOPHYSIOLOGY

Hemolysis in PNH is mainly intravascular and is due to an intrinsic abnormality of the red cell, which makes it exquisitely sensitive to activated C, whether C is activated through the alternative pathway or through an antigen-antibody reaction (classic pathway). The former mechanism is mainly responsible for chronic hemolysis in PNH; the latter explains why the hemolysis can be dramatically exacerbated in the course of a viral or bacterial infection. Hypersusceptibility to C is due to deficiency in the red cell membrane of several protective proteins **(Fig. 96-10)**, among which CD59 is the most important because it is able to hinder the insertion into the membrane of C9 polymers. The molecular basis for the deficiency of these proteins has been pinpointed not to a defect in any of the respective genes, but rather to the shortage of a unique glycolipid molecule, GPI (Fig. 96-2), which, through a peptide bond, anchors these proteins to the surface membrane of cells. The shortage of GPI is due in turn to a somatic mutation in an X-linked gene, called *PIGA*, required for an early step in GPI biosynthesis. As a result, the patient's marrow is a mosaic of mutant and nonmutant cells, and the peripheral blood always contains both GPI-negative (PNH) cells and GPI-positive (non-PNH) cells. Thrombosis is one of the most immediately life-threatening complications of PNH and yet one of the least understood in its pathogenesis. It could be that deficiency of CD59 on the PNH platelet causes inappropriate platelet activation; however, other mechanisms are possible.

FIGURE 96-10

The complement cascade and the fate of red cells. ***A.*** Normal red cells are protected from complement activation and subsequent hemolysis by CD55 and CD59. These two proteins, being GPI-linked, are missing from the surface of PNH red cells as a result of a somatic mutation of the X-linked *PIG-A* gene that encodes a protein required for an early step of the GPI molecule biosynthesis. ***B.*** In the steady state, PNH erythrocytes suffer from spontaneous (tick-over) complement activation, with consequent intravascular hemolysis through formation of the membrane attack complex (MAC); when extra complement is activated through the classical pathway, an exacerbation of hemolysis will result. ***C.*** On eculizumab, PNH erythrocytes are protected from hemolysis from the inhibition of C5 cleavage; however, upstream complement activation may lead to C3 opsonization and possible extravascular hemolysis. GPI, glycosylphosphatidylinositol; PNH, paroxysmal nocturnal hemoglobinuria. *(From L Luzzatto et al: Haematologica 95:523, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

BONE MARROW FAILURE (BMF) AND RELATIONSHIP BETWEEN PNH AND APLASTIC ANEMIA (AA)

It is not unusual that patients with firmly established PNH have a previous history of AA, sometimes well-documented; indeed, BMF preceding overt PNH is probably the rule rather than the exception. On the other hand, sometimes a patient with PNH becomes less hemolytic and more pancytopenic and ultimately has the clinical picture of AA. The relationship between PNH and AA manifested in the clinical course of patients may reflect a close link in pathogenesis. AA is thought to be an organ-specific autoimmune disease, in which T cells cause damage to hematopoietic stem cells via an as yet unidentified molecular target. The same may be true of PNH, and in this condition the target might be the GPI molecule itself. This would explain why GPI-negative (PNH) stem cells are spared; *PIGA* mutations can be demonstrated in normal people. Thus, PNH results from the combined action of two factors: failure of normal hematopoiesis and massive expansion of a PNH clone. There is evidence from mouse models that PNH stem cells do not expand on their own, and there is evidence from human patients that expansion is associated with negative selection against GPI-positive cells by GPI-specific T cells. Thus, PNH is a prime example of a clonal disease that is not malignant.

TREATMENT

TREATMENT

Paroxysmal Nocturnal Hemoglobinuria

Until 10 years ago there were essentially two treatment options for PNH: either allogeneic BMT, providing a definitive cure at the cost of non-negligible risks; or continued supportive treatment for what, unlike other acquired HAs, may be a lifelong condition. A major advance has been the introduction in 2007 of a humanized monoclonal antibody, eculizumab, which binds to the complement component C5 near the site that, when cleaved, will trigger the distal part of the complement cascade leading to formation of the MAC. With C5 blocked, the patient is relieved of intravascular hemolysis and of its attendant consequences, including hemoglobinuria. In the majority of those patients who needed regular blood transfusion, the transfusion requirement is either abolished or significantly reduced. For many PNH patients, eculizumab has meant a real improvement in the quality of life, as well as a decrease in complications, particularly thrombosis. At the same time, it is important to know that in patients on eculizumab the PNH red cells, now protected from being lysed through the MAC, do still bind C3 fragments and thus become opsonized. Therefore hemolysis continues, but it is now extravascular. The extent to which this happens depends in part on a genetic polymorphism of the complement receptor CR1. Those patients who, on eculizumab, are still receiving blood transfusion are at risk of iron overload. Based on its half-life, eculizumab must be administered intravenously every 14 days: a trial of a long-lived anti-C5 antibody is currently under way, and other complement inhibitors are under experimentation.

Eculizumab is very expensive and therefore not accessible to patients in many parts of the world. Therefore, the management of PNH by supportive treatment is still very important. Folic acid supplements (at least 3 mg/d) are mandatory; the serum iron should be checked periodically, and iron supplements should be administered as appropriate. Transfusion of filtered red cells should be used whenever necessary, which, for some patients, means quite frequently. Long-term glucocorticoids are not indicated because there is no evidence that they have any effect on chronic hemolysis; in fact, they are contraindicated because their side effects are considerable. A short course of prednisone may be useful when an inflammatory process exacerbates hemolysis. Any patient who has had venous thrombosis or who has a genetically determined thrombophilic state in addition to PNH should be on regular anticoagulant prophylaxis. With thrombotic complications that do not resolve otherwise, thrombolytic treatment with tissue plasminogen activator may be indicated.

Where eculizumab is available the proportion of PNH patients receiving BMT has decreased significantly. However, when an HLA-identical sibling is available, BMT should be taken into consideration for any young patient with severe PNH; and for patients with the so-called PNH-AA syndrome, since eculizumab has no effect on BMF. For these patients immunosuppressive treatment with antithymocyte globulin and cyclosporine A may be an alternative, and it may be compatible with concurrent administration of eculizumab.

FURTHER READING

Andolfo I et al: New insights on hereditary erythrocyte membrane defects. Haematologica 101:1284, 2016. [PubMed: 27756835]

Barcellini W: Immune hemolysis: Diagnosis and treatment recommendations. Semin Hematol 52:304, 2015. [PubMed: 26404442]

Dacie J: *The Haemolytic Anaemias*. London, Churchill Livingstone, 1985.

Grace RF et al: Erythrocyte pyruvate kinase deficiency: 2015 status report. Am J Hematol 90:825, 2015. [PubMed: 26087744]

Hill A et al: Thrombosis in paroxysmal nocturnal hemoglobinuria. Blood 121:4985; quiz 5105, 2013.

Kalfa TA: Warm antibody autoimmune hemolytic anemia. Hematology Am Soc Hematol Educ Program 2016:690, 2016. [PubMed: 27913548]

Loirat C et al: An international consensus approach to the management of atypical hemolytic uremic syndrome in children. Pediatr Nephrol 31:15, 2016. [PubMed: 25859752]

Luzzatto L et al: Management of paroxysmal nocturnal haemoglobinuria: A personal view. Br J Haematol 153:709, 2011. [PubMed: 21517820]

Luzzatto L et al: Glucose-6-phosphate dehydrogenase deficiency. Hematol Oncol Clin North Am 30:373, 2016. [PubMed: 27040960]

Uyoga S et al: Glucose-6-phosphate dehydrogenase deficiency and the risk of malaria and other diseases in children on the coast of Kenya: A case-control and a cohort study. Lancet Haematology 2:e437, 2015. [PubMed: 26686045]

Chapter 97: Anemia Due to Acute Blood Loss

Dan L. Longo

INTRODUCTION

Blood loss causes anemia by two main mechanisms: (1) by the direct loss of red cells; and (2) if the loss of blood is protracted, it will gradually deplete iron stores, eventually resulting in iron deficiency. The latter type of anemia is covered in Chap. 93; here, we are concerned with the former type, that is, *posthemorrhagic anemia*, which follows *acute* blood loss. This can be *external* (e.g., after trauma or obstetric hemorrhage) or *internal* (e.g., from bleeding in the gastrointestinal tract, rupture of the spleen, rupture of an ectopic pregnancy, subarachnoid hemorrhage). In any of these cases, after the sudden loss of a large amount of blood, there are three clinical/pathophysiologic stages. (1) At first, the dominant feature is hypovolemia, which poses a threat particularly to organs that normally have a high blood supply, like the brain and the kidneys; therefore, loss of consciousness and acute renal failure are major threats. It is important to note that at this stage an ordinary blood count will not show anemia because the hemoglobin concentration is not affected. (2) Next, as an emergency response, baroreceptors and stretch receptors will cause release of vasopressin and other peptides, and the body will shift fluid from the extravascular to the intravascular compartment, producing hemodilution; thus, the hypovolemia gradually converts to anemia. The degree of anemia will reflect the amount of blood lost. If after 3 days the hemoglobin is, for example, 7 g/dL, it means that about half of the entire blood has been lost. (3) Provided bleeding does not continue, the bone marrow response will gradually ameliorate the anemia. In this phase of the process, the reticulocyte count and erythropoietin levels will be elevated.

The diagnosis of acute posthemorrhagic anemia (APHA) is usually straightforward, although sometimes internal bleeding episodes (e.g., after a traumatic injury), even when large, may not be immediately obvious. Whenever an abrupt fall in hemoglobin has taken place, whatever history is given by the patient, APHA should be suspected. Supplementary history may have to be obtained by asking the appropriate questions, and appropriate investigations (e.g., a sonogram or an endoscopy) may have to be carried out.

TREATMENT

TREATMENT

Anemia Due to Acute Blood Loss

With respect to treatment, a two-pronged approach is imperative. (1) In many cases, the blood lost needs to be replaced promptly. Unlike with many chronic anemias, when finding and correcting the cause of the anemia is the first priority and blood transfusion may not be even necessary because the body is adapted to the anemia, with acute blood loss the reverse is true; because the body is not adapted to the anemia, blood transfusion takes priority. (2) While the emergency is being confronted, it is imperative to stop the hemorrhage and to eliminate its source.

In an acute hemorrhage situation, plasma may be preferred to saline for volume expansion since dilution of clotting factors with crystalloid may interfere with hemostasis.

A special type of APHA is blood loss during and immediately after surgery, which can be substantial (e.g., up to 2 L in the case of a radical prostatectomy). Of course with elective surgical procedures, the patient's own stored blood may be available (through preoperative autologous blood donation), and in any case, blood loss ought to have been carefully monitored/measured. The fact that this blood loss is iatrogenic dictates that ever more effort should be invested in optimizing its management. **The special features of transfusion medicine are discussed in Chap. 109.**

A Holy Grail of emergency medicine for a long time has been the idea of a blood substitute that would be universally available, suitable for all recipients, easy to store and to transport, safe, and as effective as blood itself. Two main paths have been pursued: (1) fluorocarbon synthetic chemicals that bind oxygen reversibly, and (2) artificially modified hemoglobins, known as hemoglobin-based oxygen carriers (HBOCs). Although there are numerous anecdotal reports of the use of both approaches in humans, and although HBOCs have reached the stage of phase 2–3 clinical trials, no "blood substitute" has yet become standard treatment.

FURTHER READING

Martini WZ: Coagulation complications following trauma. Mil Med Res 3:35, 2016. [PubMed: 27895932]

Moradi S et al: Artificial blood substitutes: First steps on the long route to clinical utility. Clin Med Insights Blood Disord 9:33, 2016. [PubMed: 27812292]

Chapter 98: Bone Marrow Failure Syndromes Including Aplastic Anemia and Myelodysplasia

Neal S. Young

INTRODUCTION

The hypoproliferative anemias are normochromic, normocytic, or macrocytic and are characterized by a low reticulocyte count. Hypoproliferative anemia is also a prominent feature of hematologic diseases that are described as bone marrow failure states; these include aplastic anemia, myelodysplastic syndrome (MDS), pure red cell aplasia (PRCA), and myelophthisis. Anemia in these disorders is often not a solitary or even the major hematologic finding. More frequent in bone marrow failure is *pancytopenia*: anemia, leukopenia, and thrombocytopenia. Low blood counts in the marrow failure diseases result from deficient hematopoiesis, as distinguished from blood count depression due to peripheral destruction of red cells (hemolytic anemias), platelets (idiopathic thrombocytopenic purpura [ITP] or due to splenomegaly), and granulocytes (as in the immune leukopenias). Marrow damage and dysfunction also may be secondary to infection, inflammation, or cancer.

Hematopoietic failure syndromes are classified by dominant morphologic features of the bone marrow (Table 98-1). Although practical distinction among these syndromes usually is clear, some processes are so closely related that the diagnosis may be complex. Separation between aplastic anemia and hypocellular MDS can be particularly difficult. Further, identification of constitutional genetic risk factors has blurred the distinction between acquired and inherited marrow failure. Patients may seem to suffer from two or three related diseases simultaneously, or one diagnosis may appear to evolve into another. Many of these syndromes share an immune-mediated mechanism of marrow destruction and some element of genomic instability resulting in a higher rate of malignant transformation.

TABLE 98-1

Differential Diagnosis of Pancytopenia

Pancytopenia with Hypocellular Bone Marrow	
Acquired aplastic anemia	
Constitutional aplastic anemia (Fanconi anemia, dyskeratosis congenita, and others)	
Hypocellular myelodysplastic syndrome	
Rare aleukemic leukemia	
Some acute lymphoid leukemia	
Rare lymphomas of bone marrow	
Copper deficiency	
Pancytopenia with Cellular Bone Marrow	
Primary bone marrow diseases	Secondary to systemic diseases
Myelodysplastic syndromes	Systemic lupus erythematosus
Paroxysmal nocturnal hemoglobinuria (PNH)	Hypersplenism B_{12}, folate deficiency Copper deficiency
Myelofibrosis	Alcohol
Aleukemic leukemia	HIV infection
Myelophthisis	Brucellosis
Bone marrow lymphoma	Sarcoidosis
Hairy cell leukemia	Tuberculosis
	Leishmaniasis Sepsis

Hypocellular Bone Marrow ± Pancytopenia
Q fever
Legionnaires' disease
Anorexia nervosa, starvation
Mycobacterium

It is important that the internist and general practitioner recognize the marrow failure syndromes, as their prognosis may be poor if the patient is untreated; effective therapies are often available but sufficiently complicated in their choice and delivery so as to warrant the care of a hematologist or oncologist. The identification of genes in which pathogenic mutations are etiologic for anemia, leukopenia, and thrombocytopenia, and the commercial availability of assays based on sophisticated DNA sequencing and functional consequences of mutations, should allow precise diagnosis early, by the primary care internist and pediatrician as well as by a specialist.

APLASTIC ANEMIA

DEFINITION

Aplastic anemia is pancytopenia with bone marrow hypocellularity. Acquired aplastic anemia is distinguished from iatrogenic aplasia, marrow hypocellularity after intensive cytotoxic chemotherapy for cancer, and from usually accidental physical and chemical injury, as in radiation poisoning. Aplastic anemia can also be constitutional. Genetic diseases such as Fanconi anemia and dyskeratosis congenita usually (but not always) present in early childhood and have typical physical anomalies. Telomere diseases (**see Chap. 470**) and hematologic manifestations of mutations in genes such as *GATA2, RUNX1*, and *MPL* can present as marrow failure in normal-appearing adults. Acquired aplastic anemia is often stereotypical in its manifestations, with the abrupt onset of low blood counts in a previously well young adult; seronegative hepatitis or a course of an incriminated medical drug may precede the onset. The diagnosis in these instances is uncomplicated. Sometimes blood count depression is moderate or incomplete, resulting in anemia, leukopenia, and thrombocytopenia in some combination. Aplastic anemia is related to both paroxysmal nocturnal hemoglobinuria (PNH; **Chap. 96**) and to MDS, and a clear distinction among these disorders may not be possible.

EPIDEMIOLOGY

The incidence of acquired aplastic anemia in Europe and Israel is two cases per million persons annually. In Thailand and China, rates of five to seven per million have been established. Men and women are affected

with equal frequency, but the age distribution is biphasic, with the major peak in the teens and twenties and a second rise in older adults.

ETIOLOGY

The origins of aplastic anemia have been inferred from several recurring clinical associations (Table 98-2); unfortunately, these relationships are not reliable in an individual patient and may not be etiologic. In addition, although most cases of aplastic anemia are idiopathic, little other than history separates these cases from those with a presumed etiology such as a drug exposure.

TABLE 98-2

Classification of Aplastic Anemia and Single Cytopenias

Acquired	Inherited/Constitutional
Aplastic Anemia	
Secondary	Fanconi anemia
Radiation	Dyskeratosis congenita/telomere disease
Drugs and chemicals	Shwachman-Diamond syndrome
Regular effects	Familial aplastic anemia/leukemia predisposition syndromes: *GATA2, RUNX1, CTLA4*, and others
Idiosyncratic reactions	
Viruses	
Epstein-Barr virus (infectious mononucleosis)	
Hepatitis (non-A, non-B, non-C hepatitis)	Nonhematologic syndromes (Down, Dubowitz, Seckel)
Parvovirus B19 (transient aplastic crisis, pure red cell aplasia [PRCA])	
HIV-1 (AIDS)	
Immune diseases	
Eosinophilic fasciitis	
Hypoimmunoglobulinemia	
Large granular lymphocytosis (LGL)	
Thymoma/thymic carcinoma	
Graft-versus-host disease in immunodeficiency	

Acquired	Inherited/Constitutional
Paroxysmal nocturnal hemoglobinuria (PNH)	
Pregnancy	
Idiopathic	
Cytopenias	
PRCA (see Table 98-4)	Congenital PRCA (Diamond-Blackfan anemia)
Neutropenia/agranulocytosis	
Idiopathic	Kostmann syndrome
Drugs, toxins	Shwachman-Diamond syndrome
LGL Pure white cell aplasia (+/– thymoma)	Reticular dysgenesis
Thrombocytopenia	
Drugs, toxins	Amegakaryocytic thrombocytopenia
Acquired amegakaryocytic thrombocytopenia	Thrombocytopenia with absent radii Other rare germline mutations

Radiation

Marrow aplasia is a major acute sequela of radiation. Radiation damages DNA; tissues dependent on active mitosis are particularly susceptible. Nuclear accidents involve not only power plant workers but also employees of hospitals, laboratories, and industry (food sterilization, metal radiography, etc.), as well as innocents exposed to stolen, misplaced, or misused sources. Whereas the radiation dose can be approximated from the rate and degree of decline in blood counts, dosimetry by reconstruction of the exposure can help to estimate the patient's prognosis and also to protect medical personnel from contact with radioactive tissue and excreta. MDS and leukemia, but probably not aplastic anemia, are late effects of radiation.

Chemicals

Benzene is a notorious cause of bone marrow failure: epidemiologic, clinical, and laboratory data link benzene to aplastic anemia, acute leukemia, and blood and marrow abnormalities. For leukemia, incidence is correlated with cumulative exposure, but susceptibility must also be important because only a minority of even heavily exposed workers develop myelotoxicity. The employment history is important, especially in industries where benzene is used for a secondary purpose, usually as a solvent. Benzene-related blood diseases have declined with regulation of industrial exposure. Although benzene is no longer generally available as a household solvent, exposure to its metabolites occurs in the normal diet and in the environment. The association between marrow failure and other chemicals is much less well substantiated.

Drugs

Many chemotherapeutic drugs have marrow suppression as a major toxicity; effects are dose dependent and will occur in all recipients (Table 98-3). In contrast, idiosyncratic reactions to a large and diverse group of drugs may lead to aplastic anemia without a clear dose-response relationship. A large international study in Europe in the 1980s quantitated drug relationships, especially for nonsteroidal analgesics, sulfonamides, thyrostatic drugs, some psychotropics, penicillamine, allopurinol, and gold. Association does not equal causation: a drug may have been used to treat the first symptoms of bone marrow failure (antibiotics for fever or a preceding viral illness) or provoked the first symptom of a preexisting disease (petechiae by nonsteroidal anti-inflammatory agents administered to the thrombocytopenic patient). In the context of total drug use, idiosyncratic reactions, although individually devastating, are rare events. Risk estimates are usually lower when determined in population-based studies. Furthermore, the low absolute risk is also made more obvious: even a 10- or 20-fold increase in risk translates, in a rare disease, to just a handful of drug-induced aplastic anemia cases among hundreds of thousands of exposed persons.

TABLE 98-3

Some Drugs and Chemicals Associated with Aplastic Anemia

Agents that regularly produce marrow depression as major toxicity in commonly used doses or normal exposures:
Cytotoxic drugs used in cancer chemotherapy: *alkylating agents*, *antimetabolites*, *antimitotics*, some antibiotics
Agents that frequently but not inevitably produce marrow aplasia:
Benzene
Agents associated with aplastic anemia but with a relatively low probability:
Chloramphenicol
Insecticides
Antiprotozoals: *quinacrine* and chloroquine, mepacrine
Nonsteroidal anti-inflammatory drugs (including *phenylbutazone*, indomethacin, ibuprofen, sulindac, aspirin)
Anticonvulsants (*hydantoins*, *carbamazepine*, phenacemide, felbamate)
Heavy metals (*gold*, arsenic, bismuth, mercury)
Sulfonamides: some antibiotics, antithyroid drugs (methimazole, methylthiouracil, propylthiouracil), antidiabetes drugs (tolbutamide, chlorpropamide), carbonic anhydrase inhibitors (acetazolamide and methazolamide)
Antihistamines (*cimetidine*, chlorpheniramine)
D-Penicillamine
Estrogens (in pregnancy and in high doses in animals)
Agents whose association with aplastic anemia is more tenuous:
Other antibiotics (streptomycin, tetracycline, methicillin, mebendazole, trimethoprim/sulfamethoxazole, flucytosine)

Sedatives and tranquilizers (chlorpromazine, prochlorperazine, piperacetazine, chlordiazepoxide, meprobamate, methyprylon)
Allopurinol
Methyldopa
Quinidine
Lithium
Guanidine
Potassium perchlorate
Thiocyanate
Carbimazole

Note: Terms set in italics show the most consistent association with aplastic anemia.

Infections

Posthepatitis marrow failure accounts for ~5% of etiologies in most series. Patients are usually young men who have recovered from a bout of liver inflammation 1–2 months earlier; the subsequent pancytopenia is very severe. The hepatitis is seronegative (non-A, non-B, non-C); intensive laboratory efforts including deep sequencing have not disclosed an infectious agent. Fulminant liver failure in childhood also follows seronegative hepatitis, and marrow failure occurs at a high rate in these patients. Aplastic anemia can rarely follow infectious mononucleosis. Parvovirus B19 does not usually cause generalized bone marrow failure. Transient, mild blood count depression is frequent in the course of many viral and bacterial infections.

Immunologic Diseases

Aplasia is a major consequence and the inevitable cause of death in *transfusion-associated graft-versus-host disease* (GVHD) that can occur after infusion of nonirradiated blood products to an immunodeficient recipient. Aplastic anemia is strongly associated with the rare collagen vascular syndrome eosinophilic fasciitis that is characterized by painful induration of subcutaneous tissues **(Chap. 353)**. Thymoma and hypoimmunoglobulinemia are occasional associations with aplastic anemia. Pancytopenia with marrow hypoplasia can also occur in systemic lupus erythematosus (SLE).

Pregnancy

Aplastic anemia very rarely may occur and recur during pregnancy and resolve with delivery or with spontaneous or induced abortion.

Paroxysmal Nocturnal Hemoglobinuria

An acquired mutation in the *PIG-A* gene in a hematopoietic stem cell is required for the development of PNH, but *PIG-A* mutations probably occur commonly in normal individuals. If the PIG-A mutant stem cell proliferates, the result is a clone of progeny deficient in glycosylphosphatidylinositol-linked cell surface membrane proteins **(Chap. 96)**. Small clones of deficient cells can be detected by sensitive flow cytometry tests in one-half or more of patients with aplastic anemia at the time of presentation. Functional studies of bone marrow from PNH patients, even those with mainly hemolytic manifestations, show evidence of defective hematopoiesis. Patients with an initial clinical diagnosis of PNH, especially younger individuals, may later develop frank marrow aplasia and pancytopenia; patients with an initial diagnosis of aplastic anemia may suffer later from hemolytic PNH years after recovery of blood counts.

Constitutional Disorders

Fanconi anemia, an autosomal recessive disorder, manifests as congenital developmental anomalies, progressive pancytopenia, and an increased risk of malignancy. Chromosomes in Fanconi anemia are susceptible to DNA cross-linking agents, the basis for a diagnostic assay. Patients with Fanconi anemia typically have short stature, café au lait spots, and anomalies involving the thumb, radius, and genitourinary tract. At least 17 different genetic defects (all but one with an identified gene) have been defined; the most common, type A Fanconi anemia, is due to a mutation in *FANCA*. Most of the Fanconi anemia gene products form a protein complex that activates FANCD2 by monoubiquitination to play a role in the cellular response to DNA damage and especially interstrand cross-linking.

Dyskeratosis congenita is characterized by the triad of mucous membrane leukoplasia, dystrophic nails, reticular hyperpigmentation, and with the development of aplastic anemia in childhood **(Chapter 470)**. Dyskeratosis congenita is due to mutations in genes of the telomere repair complex, which acts to maintain telomere length in replicating cells: the X-linked variety is due to mutations in the *DKC1* (*dyskerin*) gene; the more unusual autosomal dominant type is due to mutation in *TERC*, which encodes an RNA template, and *TERT*, which encodes the catalytic reverse transcriptase, telomerase. Mutations can also occur in genes like *TNF2* that encode shelterin proteins, which bind telomere DNA.

In Shwachman-Diamond syndrome, presentation is early in life with neutropenia with pancreatic insufficiency and malabsorption; most patients have compound heterozygous mutations in *SBDS* that may affect both ribosomal biogenesis (as in Diamond-Blackfan anemia; see below) and marrow stroma function.

While these constitutional syndromes can on occasion present in adults, genetic mutations are also risk factors for bone marrow failure. In the recently recognized telomeropathies (see Chap. 470), mutations in *TERT* and *TERC* have subtle effects on hematopoietic function. Typical presentations include moderate aplastic anemia, which can be chronic and not progressive, and isolated macrocytic anemia or

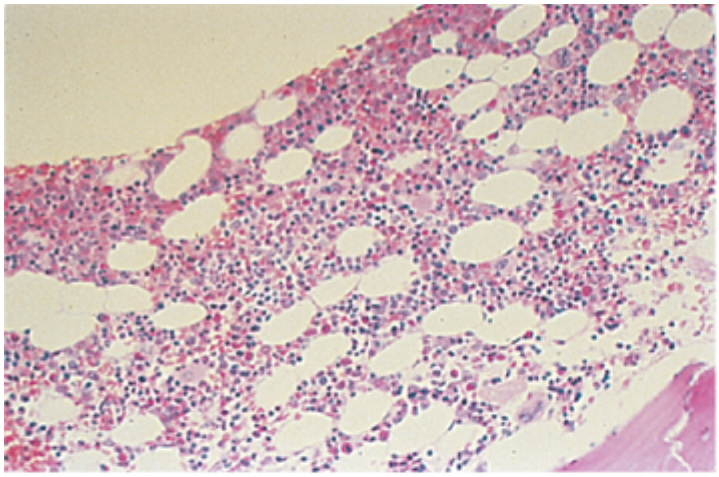
thrombocytopenia. Physical anomalies are usually not present, but early hair graying is a clue to the diagnosis. A family history may disclose pulmonary fibrosis and hepatic cirrhosis. Variable penetrance means that *TERT* and *TERC* mutations represent risk factors for marrow failure, as family members with the same mutations may have normal or only slight hematologic abnormalities but more subtle evidence of (compensated) hematopoietic insufficiency.

PATHOPHYSIOLOGY

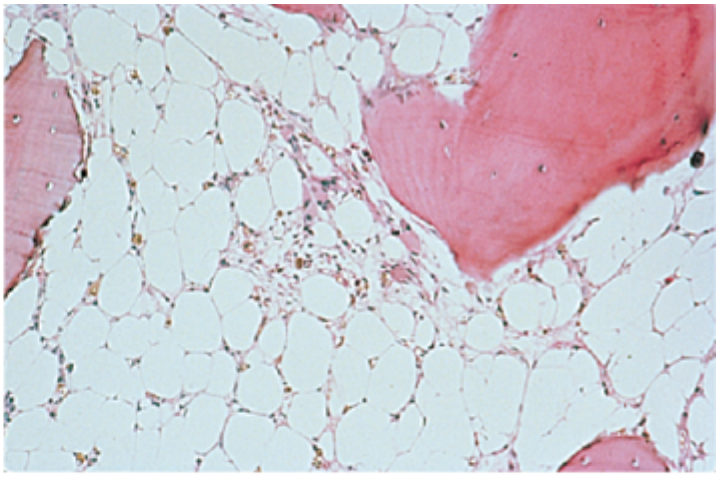
Bone marrow failure results from severe damage to the hematopoietic cell compartment. In aplastic anemia, replacement of the bone marrow by fat is apparent in the morphology of the biopsy specimen **(Fig. 98-1)** and magnetic resonance imaging (MRI) of the spine. Cells bearing the CD34 antigen, a marker of early hematopoietic cells, are greatly diminished, and in functional studies, committed and primitive progenitor cells are virtually absent; in vitro assays have suggested that the stem cell pool is reduced to ≤1% of normal in severe disease at the time of presentation.

FIGURE 98-1

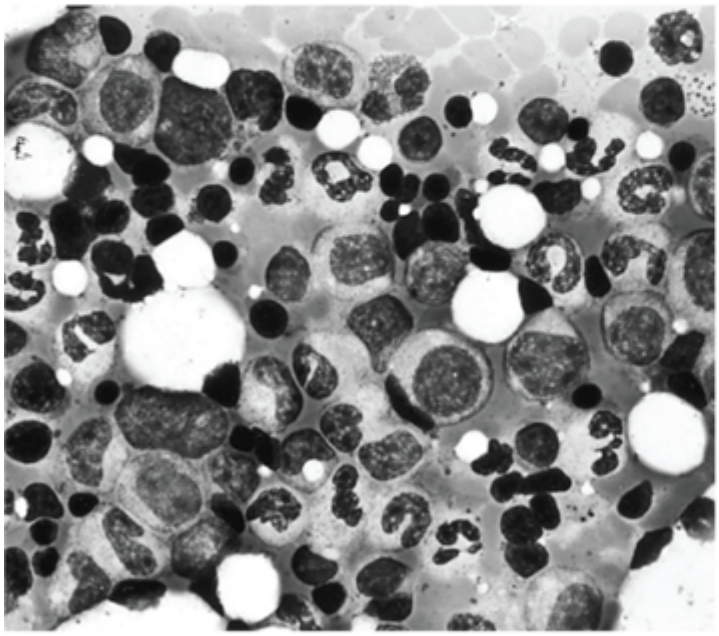
Normal and aplastic bone marrow. ***A.*** Normal bone marrow biopsy. ***B.*** Normal bone marrow aspirate smear. The marrow is normally 30–70% cellular, and there is a heterogeneous mix of myeloid, erythroid, and lymphoid cells. ***C.*** Aplastic anemia biopsy. ***D.*** Marrow smear in aplastic anemia. The marrow shows replacement of hematopoietic tissue by fat and only residual stromal and lymphoid cells.



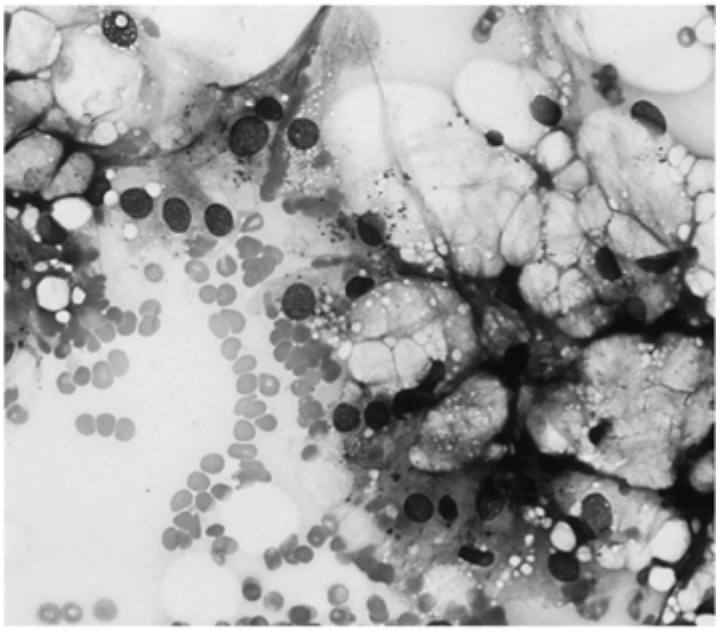
A



C



B



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Constitutional Genetic Syndromes

An intrinsic stem cell defect exists for the constitutional aplastic anemias: cells from patients with Fanconi anemia exhibit chromosome damage and death on exposure to certain chemical agents. Telomeres are short in some patients with aplastic anemia, due to heterozygous mutations in genes of the telomere repair complex. Telomeres may also shorten physiologically in acquired marrow failure due to replicative demands on a limited stem cell pool.

Chemical and Drug Injury

Extrinsic damage to the marrow follows massive physical or chemical insults such as high doses of radiation and toxic chemicals. For the more common idiosyncratic reaction to modest doses of medical drugs, altered

drug metabolism has been invoked as a mechanism. The metabolic pathways of many drugs and chemicals, especially if they are polar and have limited water solubility, involve enzymatic degradation to highly reactive electrophilic compounds; these intermediates are toxic because of their propensity to bind to cellular macromolecules. For example, derivative hydroquinones and quinolones are responsible for benzene-induced tissue injury. Excessive generation of toxic intermediates or failure to detoxify the intermediates may be genetically determined and apparent only on specific drug challenge; the complexity and specificity of the pathways imply multiple susceptibility loci and would provide an explanation for the rarity of idiosyncratic drug reactions.

Immune-Mediated Stem Cell Destruction

The recovery of marrow function in some patients prepared for bone marrow transplantation with antilymphocyte globulin first suggested that aplastic anemia might be immune mediated. Consistent with this hypothesis was the frequent failure of simple bone marrow transplantation from a syngeneic twin, without conditioning cytotoxic chemotherapy, which also argued both *against* simple stem cell absence as the cause and *for* the presence of a host factor producing marrow failure. Laboratory data support an important role for the immune system in aplastic anemia. Blood and bone marrow cells of patients can suppress normal hematopoietic progenitor cell growth, and removal of T cells from aplastic anemia bone marrow improves colony formation in vitro. Increased numbers of activated cytotoxic T cell clones are observed in aplastic anemia patients and usually decline with successful immunosuppressive therapy; type 1 cytokines are implicated; and interferon γ (IFN-γ) induces Fas expression on CD34 cells, leading to apoptotic cell death. The early immune system events in aplastic anemia are not well understood, but an oligoclonal, T cell response implies antigenic stimulus. The rarity of aplastic anemia despite common exposures (medicines, seronegative hepatitis) suggests that genetically determined features of the immune response can convert a normal physiologic response into a sustained abnormal autoimmune process, including polymorphisms in histocompatibility antigens, cytokine genes, and genes that regulate T cell polarization and effector function.

CLINICAL FEATURES

History

Aplastic anemia can appear abruptly or insidiously. Bleeding is the most common early symptom; a complaint of days to weeks of easy bruising, oozing from the gums, nose bleeds, heavy menstrual flow, and sometimes petechiae will have been noticed. With thrombocytopenia, massive hemorrhage is unusual, but small amounts of bleeding in the central nervous system can result in catastrophic intracranial or retinal hemorrhage. Symptoms of anemia are also frequent, including lassitude, weakness, shortness of breath, and a pounding sensation in the ears. Infection is an unusual first symptom in aplastic anemia (unlike in agranulocytosis, where pharyngitis, anorectal infection, or frank sepsis occurs early). Patients often feel and look remarkably well despite drastically reduced blood counts. Systemic complaints and weight loss should point to other etiologies of pancytopenia. Prior medical drug use, chemical exposure, and preceding viral illnesses must often be elicited with directed questioning. A family history of hematologic diseases or blood

abnormalities, of pulmonary or liver fibrosis, or of early hair graying points to a telomeropathy; a family history of unusual infections and warts to *GATA2* deficiency.

Physical Examination

Petechiae and ecchymoses are typical, and retinal hemorrhages may be present. Pelvic and rectal examinations can often be deferred but, when performed, should be undertaken with great gentleness to avoid trauma; these may show bleeding from the cervical os and blood in the stool. Pallor of the skin and mucous membranes is common. Infection on presentation is unusual but may occur if the patient has been symptomatic for a few weeks. Lymphadenopathy and splenomegaly are highly atypical of aplastic anemia. Café au lait spots and short stature suggest Fanconi anemia; peculiar nails and leukoplakia suggest dyskeratosis congenita; early graying (and use of hair dyes to mask it!) suggest a telomerase defect.

LABORATORY STUDIES

Blood

The smear shows large erythrocytes and a paucity of platelets and granulocytes. Mean corpuscular volume (MCV) is commonly increased. Reticulocytes are absent or few, and lymphocyte numbers may be normal or reduced. The presence of immature myeloid forms suggests leukemia or MDS; nucleated red blood cells (RBCs) suggest marrow fibrosis or tumor invasion; abnormal platelets suggest either peripheral destruction or MDS.

Bone Marrow

The bone marrow is usually readily aspirated but dilute on smear, and the fatty biopsy specimen may be grossly pale on withdrawal; a “dry tap” instead suggests fibrosis or myelophthisis. In severe aplasia, the smear of the aspirated specimen shows only red cells, residual lymphocytes, and stromal cells; the biopsy (which should be >1 cm in length) is superior for determination of cellularity and shows mainly fat under the microscope, with hematopoietic cells occupying <25% of the marrow space; sometimes, the biopsy is virtually all fat. The correlation between marrow cellularity and disease severity is imperfect; patients with moderate disease by blood counts can have empty iliac crest biopsies, whereas “hot spots” of hematopoiesis may be seen in severe cases. Residual hematopoietic cells should have normal morphology, except for mildly megaloblastic erythropoiesis; megakaryocytes are greatly reduced and usually absent. Granulomas may indicate an infectious etiology of the marrow failure.

Ancillary Studies

Chromosome breakage studies of peripheral blood using diepoxybutane or mitomycin C should be performed on children and younger adults to exclude Fanconi anemia. Very short telomere length strongly suggests the presence of a telomerase or shelterin mutation, which can be pursued by family studies and nucleotide sequencing. Chromosome studies of bone marrow cells are often revealing in MDS, but should be

negative in typical aplastic anemia. Flow cytometry offers a sensitive diagnostic test for PNH. Serologic studies may show evidence of viral infection, such as Epstein-Barr virus and HIV. Posthepatitis aplastic anemia is seronegative. Occasionally MRI may be helpful to assess the fat content of vertebrae in order to distinguish aplasia from MDS.

DIAGNOSIS

The diagnosis of aplastic anemia is usually straightforward, based on the combination of pancytopenia with a fatty bone marrow. Aplastic anemia is a disease of the young and should be a leading diagnosis in the pancytopenic adolescent or young adult. When pancytopenia is secondary, the primary diagnosis is usually obvious from either history or physical examination: the massive spleen of alcoholic cirrhosis, the history of metastatic cancer or SLE, or miliary tuberculosis on chest radiograph (Table 98-1).

Diagnostic problems can occur with atypical presentations and among related hematologic diseases. Patients with bone marrow hypocellularity may have depression of only one or two of three blood lines, with later progression to pancytopenia. The most important differential diagnosis is between acquired and constitutional aplastic anemia, and between aplastic anemia and MDS. The bone marrow in constitutional aplastic anemia is usually morphologically indistinguishable from the aspirate in acquired disease (an exception is *GATA2* deficiency with it characteristic megakaryocyte atypia). The diagnosis can be suggested by family history, abnormal blood counts since childhood, or the presence of associated physical anomalies. Many laboratories now offer genomic sequencing of a panel of etiologic germline genes, allowing for precision diagnosis. Aplastic anemia may be difficult to distinguish from the hypocellular variety of MDS: MDS is favored by finding morphologic abnormalities, particularly of megakaryocytes and myeloid precursor cells, and typical cytogenetic abnormalities. Somatic mutations can be detected on genomic screening of genes recurrently mutated in MDS, acute myeloid leukemia (AML), and myeloproliferative neoplasms (see below).

PROGNOSIS

The natural history of severe aplastic anemia is rapid deterioration and death. Historically, provision first of RBC and later of platelet transfusions and effective antibiotics were of some benefit, but few patients show spontaneous recovery. The major prognostic determinant is the blood count. Severe disease historically has been defined by the presence of two of three parameters: absolute neutrophil count <500/μL, platelet count <20,000/μL, and corrected reticulocyte count <1% (or absolute reticulocyte count <60,000/μL). In the era of effective immunosuppressive therapies, absolute numbers of reticulocytes (>25,000/μL) and lymphocytes (>1000/μL) may be better predictors of response to treatment and long-term outcome.

TREATMENT

TREATMENT

Aplastic Anemia

Severe acquired aplastic anemia can be cured by replacement of the absent hematopoietic cells (and the immune system) by stem cell transplant, or it can be ameliorated by suppression of the immune system to allow recovery of the patient's residual bone marrow function. Glucocorticoids are not of value as primary therapy. Suspect exposures to drugs or chemicals should be discontinued; however, spontaneous recovery of severe blood count depression is rare, and a waiting period before beginning treatment may not be advisable unless the blood counts are only modestly depressed.

HEMATOPOIETIC STEM CELL TRANSPLANTATION

This is the best therapy for the younger patient with a fully histocompatible sibling donor (**Chap. 110**). Human leukocyte antigen (HLA) typing should be ordered as soon as the diagnosis of aplastic anemia is established in a child or younger adult. In transplant candidates, transfusion of blood from family members should be avoided so as to prevent sensitization to histocompatibility antigens. In general, limited numbers of blood products probably do not greatly affect outcome. For allogeneic transplant from fully matched siblings, long-term survival rates for children are ~90%. Transplant morbidity and mortality are increased among adults, due to the higher risk of chronic GVHD and serious infections.

Most patients do not have a suitable sibling donor. Occasionally, a full phenotypic match can be found within the family and serve as well. Far more available are other alternative donors, either unrelated but histocompatible volunteers or closely but not perfectly matched family members. High-resolution matching at HLA and more effective conditioning regimens and GVHD prophylaxis have led to improved survival rates in patients who proceed to alternative donor transplant. Survival is equivalent between matched unrelated and conventional sibling donors, although complication rates (mainly graft-versus-host disease and infection) are higher using unrelated donors. Cord blood can be a source of stem cells especially for children. Transplantation from an HLA haploidentical family donor is increasingly popular: a donor is almost always quickly available, and post-transplant cyclophosphamide appears to be effective in preventing graft-versus-host disease. Transplant protocols for marrow failure now usually do not include radiation in order to avoid late occurrence of cancer.

IMMUNOSUPPRESSION

The standard regimen of antithymocyte globulin (ATG) in combination with cyclosporine induces hematologic recovery (independence from transfusion and a leukocyte count adequate to prevent infection) in 60–70% of patients. Children do especially well, whereas older adult patients can suffer complications due to the presence of comorbidities. An early robust hematologic response correlates with long-term survival. Improvement in granulocyte number is generally apparent within 2 months of treatment. Most recovered patients continue to have some degree of blood count depression, the MCV remains elevated, and bone marrow cellularity returns toward normal very slowly if at all. Relapse (recurrent pancytopenia) is frequent, often occurring as cyclosporine is discontinued; most, but not all, patients respond to reinstitution of immunosuppression, but some responders become dependent on continued cyclosporine administration. Development of MDS, with typical marrow morphologic, but more often cytogenetic abnormalities, occurs in ~15% of treated patients, usually but not invariably associated with a return of pancytopenia, and some patients develop leukemia. A laboratory diagnosis of PNH can generally be made at the time of presentation

of aplastic anemia by flow cytometry; recovered patients may have frank hemolysis if the PNH clone expands. Bone marrow examinations should be performed if there is an unfavorable change in blood counts.

Horse ATG is administered as intravenous infusions over 4 days; rabbit ATG is much less effective, perhaps because it reduces T-regulatory cell numbers in patients. Serum sickness, a flulike illness with a characteristic cutaneous eruption and arthralgia, often develops ~10 days after initiating treatment. Methylprednisolone is administered with ATG to ameliorate the immune consequences of heterologous protein infusion. (Excessive or extended glucocorticoid therapy is associated with avascular joint necrosis.) Cyclosporine is administered orally at an initial high dose, with subsequent adjustment according to blood levels. Its most important side effects are nephrotoxicity, hypertension, and seizures.

Most patients with aplastic anemia lack a suitable marrow donor, and immunosuppression is the treatment of choice. Overall survival is equivalent with transplantation and immunosuppression. However, successful transplant cures marrow failure, whereas patients who recover adequate blood counts after immunosuppression remain at risk of relapse and malignant evolution. Because of excellent results in children and younger adults, allogeneic transplant should be performed if a suitable sibling donor is available. Increasing age and the severity of neutropenia are the most important factors weighing in the decision between transplant and immunosuppression in adults who have a matched family donor: older patients do better with ATG and cyclosporine, whereas transplant is preferred if neutrocytopenia is profound.

ELTROMBOPAG

Hematopoietic growth factors (HGFs) such as erythropoietin (EPO) and granulocyte colony-stimulating factor (G-CSF) are not effective in aplastic anemia, probably because endogenous blood levels in patients are extremely high. Circulating thrombopoietin is also elevated, but a thrombopoietin mimetic has shown unexpected activity, and eltrombopag now is approved for use in patients with refractory aplastic anemia. Robust, trilineage responses suggest action on the stem cell. Eltrombopag added to first-line immunosuppression with horse ATG markedly increased overall and complete response rates.

That outcomes following both stem cell transplant and immunosuppression have improved with time has complicated development of consensus algorithms. Children with histocompatible siblings should be offered transplant, but some have advocated for “upfront” transplants from well-matched unrelated donors. Transplant has also been extended to older patients, including from unrelated matched donors and haploidentical donors. Conversely, immunosuppression combined with stem cell stimulation may lead to responses within a few months in almost all patients and can be instituted at diagnosis. Even heavily transfused and infected patients in whom immunosuppression has failed can be salvaged by stem cell transplant later.

ANDROGENS

The effectiveness of androgens has not been verified in controlled trials, but occasional patients will respond or even demonstrate blood count dependence on continued therapy. Sex hormones upregulate telomerase gene activity in vitro, which is possibly also their mechanism of action in improving marrow function. For

patients with moderate disease, especially if a telomere gene defect is present, a 3- to 4-month trial may improve all blood counts (**Chap. 470**).

SUPPORTIVE CARE

Meticulous medical attention is required so that the patient may survive to benefit from definitive therapy or, having failed treatment, to maintain a reasonable existence in the face of pancytopenia. First and most important, infection in the presence of severe neutropenia must be aggressively treated by prompt institution of parenteral, broad-spectrum antibiotics. Therapy is empirical and must not await results of culture, although specific foci of infection such as oropharyngeal or anorectal abscesses, pneumonia, sinusitis, and typhlitis (necrotizing colitis) should be sought on physical examination and with radiographic studies. When indwelling plastic catheters become contaminated, vancomycin should be added. Persistent or recrudescent fever implies fungal disease: *Candida* and *Aspergillus* are common, especially after several courses of antibacterial antibiotics. A major reason for the improved prognosis in aplastic anemia has been the development of better antifungal drugs and the timely institution of such therapy when infection is suspected. Granulocyte transfusions using G-CSF–mobilized peripheral blood can be effective when infections are overwhelming or refractory. Hand washing, the single best method of preventing the spread of infection, remains a neglected practice. Nonabsorbed antibiotics for gut decontamination are poorly tolerated and not of proven value, nor does reverse isolation reduce mortality from infections.

Both platelet and erythrocyte numbers can be maintained by transfusion. Alloimmunization historically limited the usefulness of platelet transfusions and is now minimized by several strategies, including use of single donors to reduce exposure and physical or chemical methods to diminish leukocytes in the product; HLA-matched platelets are usually effective in patients refractory to random donor products. Inhibitors of fibrinolysis such as aminocaproic acid have not been shown to relieve mucosal oozing; the use of low-dose glucocorticoids to induce “vascular stability” is unproven and not recommended. With prophylactic platelet transfusions, the goal is to maintain the platelet count >10,000/μL (oozing from the gut increases precipitously at counts <5000/μL). Menstruation should be suppressed either by oral estrogens or nasal follicle-stimulating hormone/luteinizing hormone antagonists. Aspirin and other nonsteroidal anti-inflammatory agents must be avoided in the presence of thrombocytopenia.

RBCs should be transfused so as to allow patient a normal level of activity, usually at a hemoglobin value of 70 g/L (90 g/L if there is underlying cardiac or pulmonary disease); a regimen of 2 units every 2 weeks will replace normal losses in a patient without a functioning bone marrow. In chronic anemia, the iron chelators, deferoxamine and deferasirox, should be added at approximately the fiftieth transfusion to avoid secondary hemochromatosis.

PURE RED CELL APLASIA

Other more restricted forms of marrow failure occur, in which only a single circulating cell type is affected and the marrow shows corresponding absence or decreased numbers of specific precursor cells: aregenerative anemia as in PRCA (see below), thrombocytopenia with amegakaryocytosis (**Chap. 111**), and

neutropenia without marrow myeloid cells in agranulocytosis (Chap. 60). In general, and in contrast to aplastic anemia and MDS, the unaffected lineages appear quantitatively and qualitatively normal. Agranulocytosis, the most frequent of these syndromes, is usually a complication of medical drug use (with agents similar to those related to aplastic anemia), either by a mechanism of direct chemical toxicity or by immune destruction. Agranulocytosis has an incidence similar to aplastic anemia but is especially frequent among older adults and in women. The syndrome should resolve with discontinuation of exposure, but significant mortality is attached to neutropenia in the older and often previously unwell patient. Both pure white cell aplasia (agranulocytosis without incriminating drug exposure) and amegakaryocytic thrombocytopenia are exceedingly rare and, like PRCA, appear to be due to destructive antibodies or lymphocytes and can respond to immunosuppressive therapies. In all of the single-lineage failure syndromes, progression to pancytopenia or leukemia is unusual.

DEFINITION AND DIFFERENTIAL DIAGNOSIS

PRCA is characterized by anemia, reticulocytopenia, and absent or rare erythroid precursor cells in the bone marrow. The classification of PRCA is shown in Table 98-4. In adults, PRCA is acquired. An identical syndrome can occur constitutionally: Diamond-Blackfan anemia, or congenital PRCA, is diagnosed at birth or in early childhood and often responds to glucocorticoid treatment; mutations in ribosome protein genes are etiologic. Temporary red cell failure occurs in transient aplastic crisis of hemolytic anemias due to acute parvovirus infection (Chap. 192) and in transient erythroblastopenia of childhood, which occurs in normal children.

TABLE 98-4

Classification of Pure Red Cell Aplasia

Self-limited
Transient erythroblastopenia of childhood
Transient aplastic crisis of hemolysis (acute B19 parvovirus infection)
Fetal red blood cell aplasia
Nonimmune hydrops fetalis (in utero B19 parvovirus infection)
Hereditary pure red cell aplasia
Congenital pure red cell aplasia (Diamond-Blackfan anemia)
Acquired pure red cell aplasia
Cancer
Thymoma
Lymphoid malignancies (and more rarely other hematologic diseases)
Paraneoplastic to solid tumors
Connective tissue disorders with immunologic abnormalities
Systemic lupus erythematosus, juvenile rheumatoid arthritis, rheumatoid arthritis
Multiple endocrine gland insufficiency
Viruses
Persistent B19 parvovirus, hepatitis, adult T cell leukemia virus, Epstein-Barr virus
Pregnancy
Drugs
Especially phenytoin, azathioprine, chloramphenicol, procainamide, isoniazid

Antibodies to erythropoietin
Idiopathic

CLINICAL ASSOCIATIONS AND ETIOLOGY

PRCA has important associations with immune system diseases. A minority of cases occur with a thymoma. More frequently, red cell aplasia can be the major manifestation of large granular lymphocytosis or complicate chronic lymphocytic leukemia. Some patients may be hypogammaglobulinemic. Occasionally (as compared to agranulocytosis), PRCA can be due to an idiosyncratic drug reaction. Subcutaneous administration of EPO has provoked PRCA mediated by neutralizing antibodies to the hormone.

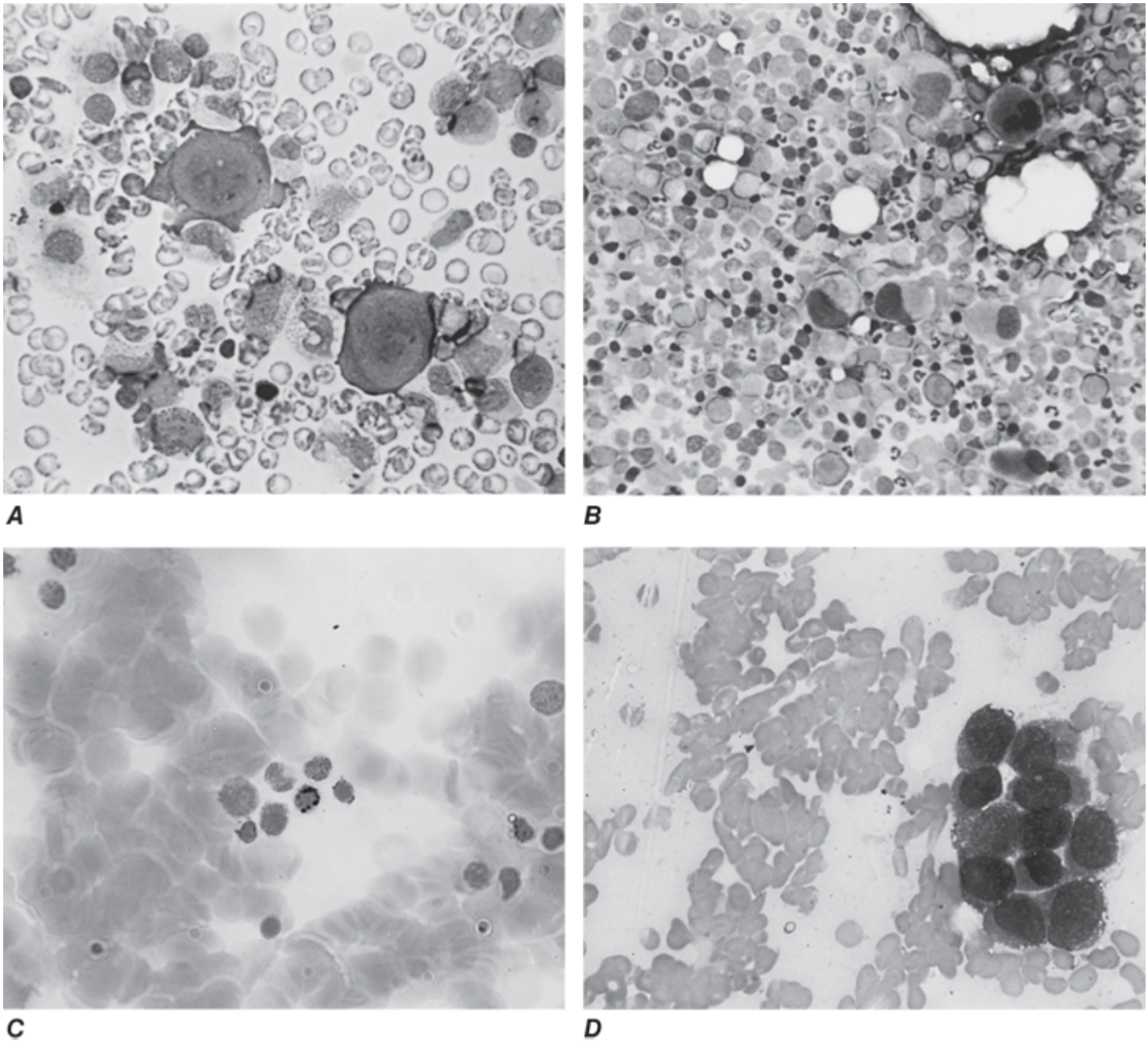
Antibodies to RBC precursors are frequently present in the blood, but T cell inhibition is probably the more common immune mechanism. Cytotoxic lymphocyte activity restricted by histocompatibility locus or specific for human T-cell leukemia/lymphoma virus I–infected cells and natural killer cell activity inhibitory of erythropoiesis have been demonstrated in particularly well-studied individual cases.

PERSISTENT PARVOVIRUS B19 INFECTION

Chronic parvovirus infection is an important, treatable cause of red cell aplasia. This common virus causes a benign exanthem of childhood (fifth disease) and a polyarthralgia/arthritis syndrome in adults. In patients with underlying hemolysis (or any condition that increases demand for RBC production), parvovirus infection can cause a transient aplastic crisis and an abrupt but temporary worsening of the anemia due to failed erythropoiesis. In normal individuals, acute infection is resolved by production of neutralizing antibodies to the virus, but in the setting of congenital, acquired, or iatrogenic immunodeficiency, persistent viral infection may occur. The bone marrow shows red cell aplasia and the presence of giant pronormoblasts **(Fig. 98-2)**, which is the cytopathic sign of B19 parvovirus infection. Viral tropism for human erythroid progenitor cells is due to its use of erythrocyte P antigen as a cellular receptor for entry. Direct cytotoxicity of virus causes anemia if demands on erythrocyte production are high; in normal individuals, the temporary cessation of red cell production is not clinically apparent, and skin and joint symptoms are mediated by immune complex deposition.

FIGURE 98-2

Pathognomonic cells in marrow failure syndromes. ***A.*** Giant pronormoblast, the cytopathic effect of B19 parvovirus infection of the erythroid progenitor cell. ***B.*** Uninuclear megakaryocyte and microblastic erythroid precursors typical of the 5q–myelodysplasia syndrome. ***C.*** Ringed sideroblast showing perinuclear iron granules. ***D.*** Tumor cells present on a touch preparation made from the marrow biopsy of a patient with metastatic carcinoma.



A B C D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Pure Red Cell Aplasia

History, physical examination, and routine laboratory studies may disclose an underlying disease or a drug exposure. Thymoma should be sought by radiographic procedures; tumor excision is indicated, but anemia does not necessarily improve with surgery. The diagnosis of parvovirus infection requires detection of viral DNA sequences in the blood (IgG and IgM antibodies are commonly absent). The presence of erythroid colonies has been considered predictive of response to immunosuppressive therapy in idiopathic PRCA.

Red cell aplasia is compatible with long-term survival with supportive care alone: a combination of erythrocyte transfusions and iron chelation. For persistent B19 parvovirus infection, almost all patients respond to intravenous immunoglobulin therapy. The majority of patients with idiopathic PRCA respond favorably to immunosuppression: glucocorticoids, cyclosporine, ATG, azathioprine, and cyclophosphamide are effective.

MYELODYSPLASIA

DEFINITION

The MDS are a heterogeneous group of hematologic disorders broadly characterized by both (1) cytopenias due to bone marrow failure and (2) a high risk of development of AML. Anemia, often with thrombocytopenia and neutropenia, occurs with dysmorphic (abnormal appearing) and usually cellular bone marrow, which is evidence of ineffective blood cell production. In patients with “low-risk” MDS, marrow failure dominates the clinical course. In other patients, myeloblasts are present at diagnosis, chromosomes are abnormal, and the “high risk” is due to leukemic progression. MDS may be fatal due, most often, to complications of pancytopenia, or to progression to leukemia, but a large proportion of patients will die of concurrent disease, the comorbidities typical in an elderly population. A useful nosology of these often confusing entities was first developed by the French-American-British Cooperative Group in 1983. Five subtypes were defined then: refractory anemia (RA), refractory anemia with ringed sideroblasts (RARS), refractory anemia with excess blasts (RAEB), refractory anemia with excess blasts in transformation (RAEB-t), and chronic myelomonocytic leukemia (CMML). The World Health Organization (WHO) classification (2002) recognized that the distinction between RAEB-t and AML is arbitrary and grouped them together as acute leukemia, and clarified that CMML behaves as a myeloproliferative disease. The current WHO classification of 2016 is complicated **(Table 98-5)**: blast percentage remains critical in defining MDS categories; erythroid predominant leukemias are now largely regarded as MDS; defining cytogenetic abnormalities are reaffirmed; and a single somatic mutation, in *SF3B1*, is now a feature of sideroblastic anemias. Identification of somatically mutated genes and their correlation with clinical outcomes will be increasingly important in defining classification, prognosis, and targeting therapy.

TABLE 98-5

World Health Organization (WHO) Classification of Myelodysplastic Syndromes (MDS)/Neoplasms

Name	Ring Sideroblasts	Myeloblasts	Karyotype
MDS with single lineage dysplasia (MDS-SLD)	<15% (<5%)[a]	BM <5%, PB <1%, no Auer rods	Any, unless fulfills all criteria for MDS with isolated del(5q)
MDS with multilineage dysplasia (MDS-MLD)	<15% (<5%)[a]	BM <5%, PB <1%, no Auer rods	Any, unless fulfills all criteria for MDS with isolated del(5q)
MDS with ring sideroblasts (MDS-RS)			
MDS-RS with single lineage dysplasia (MDS-RS-SLD)	≥15% / ≥5%[a]	BM <5%, PB <1%, no Auer rods	Any, unless fulfills all criteria for MDS with isolated del(5q)
MDS-RS with multilineage dysplasia (MDS-RS-MLD)	≥15% / ≥5%[a]	BM <5%, PB <1%, no Auer rods	Any, unless fulfills all criteria for MDS with isolated del(5q)
MDS with isolated del(5q)	None or any	BM <5%, PB <1%, no Auer rods	del(5q) alone or with 1 additional abnormality except −7 or del (7q)
MDS with excess blasts (MDS-EB)			
MDS-EB-1	None or any	BM 5–9% or PB 2–4%, no Auer rods	Any
MDS-EB-2	None or any	BM 10–19% or PB 5–19% or Auer rods	Any
MDS, unclassifiable (MDS-U) • with 1% blood blasts • with single lineage dysplasia and pancytopenia • based on defining cytogenetic abnormality	None or any None or any 15%	BM <5%, PB=1%, no Auer rods BM <5%, PB=1%, no Auer rods BM <5%, PB=1%, no Auer rods	Any Any MDS-defining abnormality

Name	Ring Sideroblasts	Myeloblasts	Karyotype
MDS, unclassifiable (MDS-U) • with 1% blood blasts • with single lineage dysplasia and pancytopenia • based on defining cytogenetic abnormality	None or any None or any 15%	BM <5%, PB=1%, no Auer rods BM <5%, PB=1%, no Auer rods BM <5%, PB=1%, no Auer rods	Any Any MDS-defining abnormality
Refractory cytopenia of childhood	None	BM <5%, PB <2%	Any

[a]If *SF3B1* mutation is present.

Abbreviations: BM, bone marrow; PB, peripheral blood.

The diagnosis of MDS can be a challenge, even for the expert, because sometimes subtle clinical and pathologic features must be distinguished, and precise diagnostic categorization requires a hematopathologist knowledgeable in the latest classification scheme. Unfortunately, agreement among pathologists on morphologic features and classification is imperfect; changes in the appearance of megakaryocytes are more reliable than loss of granules in neutrophil precursors or dyserythropoiesis. Further, dysplastic changes can be observed in normal individuals, and they can occur with vitamin deficiencies and as drug effects. Genomic testing is increasingly routine and can be difficult to interpret, as in differences between somatic and germline mutations, pathogenic mutations versus those of unknown significance (clonal hematopoiesis increases in frequency with age and involves genetic changes that may be clinically silent or convey an increased risk of hematologic malignancy), and clone size and changes over time. It is important that the internist and primary care physician be sufficiently familiar with MDS to expedite referral to a hematologist because many new therapies are now available to improve hematopoietic function and the judicious use of supportive care can improve the patient's quality of life.

EPIDEMIOLOGY

MDS is a disease of the elderly; the mean age at onset is older than 70 years. There is a slight male preponderance. MDS is a relatively common form of bone marrow failure, with reported incidence rates of 35 to >100 per million persons in the general population and 120 to >500 per million in older adults. Estimates of incidence in the United States range from 30,000 to 40,000 new cases annually and a prevalence of 60,000–120,000 in the population. MDS is rare in children, in whom it often has an identifiable genetic basis.

Secondary or therapy-related MDS is not age related. Rates of MDS have increased over time, due to better recognition of the syndrome by physicians, and an aging population.

ETIOLOGY AND PATHOPHYSIOLOGY

MDS is associated with environmental exposures such as radiation and benzene; other risk factors have been reported inconsistently. Secondary, therapy-related MDS occurs as a late toxicity of cancer treatment, radiation and the radiomimetic alkylating agents such as busulfan, nitrosourea, or procarbazine (with a latent period of 5–7 years), or the DNA topoisomerase inhibitors (2-year latency). Acquired aplastic anemia, Fanconi anemia, and other constitutional marrow failure diseases can evolve into MDS; occasionally, MDS in adults is recognized as due to germline *GATA2, RUNX1,* or telomere repair gene mutations. The typical MDS patient does not have a suggestive environmental exposure history or a preceding hematologic disease. MDS is a disease of aging, suggesting random cumulative intrinsic and environmental damage to marrow cells.

MDS is a clonal hematopoietic stem cell disorder characterized by disordered cell proliferation and impaired differentiation, resulting in cytopenias and risk of progression to leukemia. Both chromosomal and genetic instability have been implicated: both are aging-related. Cytogenetic abnormalities are found in approximately one-half of patients, and some of the same specific lesions are also seen in leukemia; aneuploidy (chromosome loss or gain) is more frequent than translocations. Accelerated telomere attrition may destabilize the genome in marrow failure and predispose to acquisition of chromosomal lesions. Cytogenetic abnormalities are not random (loss of all or part of 5, 7, and 20, trisomy of 8) and may be related to etiology (11q23 following topoisomerase II inhibitors). The type and number of cytogenetic abnormalities strongly correlate with the probability of leukemic transformation and survival.

Genomics has illuminated the role of mutations in the pathophysiology of MDS. Recurrent somatic mutations, acquired in the abnormal marrow cells and absent in the germline, have been identified in about 100 genes. Many of the same genes are also mutated in AML without MDS, whereas others are distinctive in subtypes of MDS. A prominent example of the latter is the discovery of mutations in genes of the RNA splicing machinery, especially *SF3B1*, which strongly associates with sideroblastic anemia. Some mutations correlate with prognosis: spliceosome defects with favorable outcome, and mutations in *EZH2, TP53, RUNX1*, and *ASXL1* with poor outcome. Mutations and cytogenetic abnormalities are not independent: *TP53* mutations associate with complex cytogenetic abnormalities and *TET2* mutations with normal cytogenetics. Correlation and exclusion in the pattern of mutations indicate a functional genomic architecture. Analysis of deep sequencing results in patients whose MDS evolved to AML has shown clonal succession, with founder clones acquiring additional mutations to produce clonal dominance. Furthermore, the prevalence of abnormal cells by morphology underestimates bone marrow involvement by MDS clones, as cells normal in appearance are derived from the abnormal clones. Both presenting and evolving hematologic manifestations result from the accumulation of multiple genetic lesions: loss of tumor-suppressor genes, activating oncogene, epigenetic pathways that affect mRNA processing and methylation status, or other harmful alterations. Pathophysiology has been linked to mutations and chromosome abnormalities in some specific MDS syndromes. The 5q– deletion leads to heterozygous loss of a ribosomal protein gene (ribosomal protein gene mutations cause

Diamond-Blackfan anemia, like much MDS characterized by deficient erythropoiesis). An immune pathophysiology may underlie trisomy 8 MDS; selected younger MDS patients can respond to immunosuppressive therapy as administered for aplastic anemia. However, in general for MDS, the role of the immune system and its cells and cytokines; the role of the hematopoietic stem cell niche, the microenvironment, and cell–cell interactions; the fate of normal cells in the Darwinian competitive environment of the dysplastic marrow; and how mutant cells produce marrow failure in MDS are still not completely understood.

CLINICAL FEATURES

Anemia dominates the early course. Most symptomatic patients complain of the gradual onset of fatigue and weakness, dyspnea, and pallor, but at least one-half of patients are asymptomatic, and their MDS is discovered only incidentally on routine blood counts. Previous chemotherapy or radiation exposure is an important historic fact. Fever and weight loss should point to a myeloproliferative rather than myelodysplastic process. MDS in childhood is rare and, when diagnosed, implicates an underlying genetic disease. Children with Down syndrome are susceptible to MDS as well as leukemia. A family history may indicate a hereditary form of sideroblastic anemia, Fanconi anemia, or a telomeropathy. Inherited *GATA2* mutations, as in the MonoMAC syndrome (with increased susceptibility to viral, mycobacteria, and fungal infections, as well as deficient numbers of monocytes, natural killer cells, and B lymphocytes), also can cause MDS in young patients.

The physical examination is remarkable for signs of anemia; approximately 20% of patients have splenomegaly. Some unusual skin lesions, including Sweet's syndrome (febrile neutrophilic dermatosis), occur with MDS. Accompanying autoimmune syndromes are not infrequent. In the younger patient, stereotypical anomalies point to a constitutional syndrome (short stature, abnormal thumbs in Fanconi anemia; early graying in the telomeropathies; cutaneous warts in *GATA2* deficiency).

LABORATORY STUDIES

Blood

Anemia is present in most cases, either alone or as part of bi- or pancytopenia; isolated neutropenia or thrombocytopenia is more unusual. Macrocytosis is common, as in most marrow failure disease. Platelets also are large and lack granules. In functional studies, they may show marked abnormalities, and patients may have bleeding symptoms despite seemingly adequate numbers. Neutrophils are hypogranulated; have hyposegmented, ringed, or abnormally segmented nuclei; contain Döhle bodies; and may be functionally deficient. Circulating myeloblasts usually correlate with marrow blast numbers, and their quantity is important for classification and prognosis. The total white blood cell count (WBC) is usually normal or low, except in chronic myelomonocytic leukemia. As in aplastic anemia, MDS can be associated with a clonal population of PNH cells. Genetic testing is commercially available for constitutional syndromes.

Bone Marrow

The bone marrow is usually normal or hypercellular, but in about 20% of cases it is sufficiently hypocellular to lead to confusion with aplastic anemia. No single characteristic feature of marrow morphology distinguishes MDS, but the following are commonly observed: dyserythropoietic changes (especially nuclear abnormalities) and ringed sideroblasts in the erythroid lineage; hypogranulation and hyposegmentation in granulocytic precursors, with an increase in myeloblasts; and megakaryocytes showing reduced numbers of or disorganized nuclei. Megaloblastic nuclei and defective hemoglobinization in the erythroid lineage are common. Prognosis strongly correlates with the proportion of marrow blasts. Cytogenetic analysis and fluorescent in situ hybridization can identify chromosomal abnormalities.

DIFFERENTIAL DIAGNOSIS

Deficiencies of vitamin B_{12} or folate should be excluded by appropriate blood tests; vitamin B_6 deficiency can be assessed by a therapeutic trial of pyridoxine if the bone marrow shows ringed sideroblasts. Copper deficiency can lead to cytopenias and dysplastic marrows of varying cellularity. Marrow dysplasia can be observed in acute viral infections, drug reactions, or chemical toxicity but should be transient. More difficult are the distinctions between hypocellular MDS and aplasia or between RA with excess blasts and early acute leukemia. The WHO considers 20% blasts in the marrow as the criterion that separates AML from MDS. In young patients, underlying, predisposing genetic diseases should be considered and appropriate genomic testing performed (see above).

PROGNOSIS

The median survival varies greatly from years for patients with 5q– or sideroblastic anemia to a few months in RA with excess blasts or severe pancytopenia associated with monosomy 7. The International Prognostic Scoring System, revised in 2012 (IPSS; Table 98-6) assists in making predictions. Even “low-risk” MDS has significant morbidity and mortality. More refined (and also more complicated) prognostic scoring systems can separate out Int-1 risk patients who have relatively poor prognoses. Prognostic systems have been developed based on survival from diagnosis, but prognosis changes over time, and hazards ratios for survival and leukemic transformation converge over time among risk categories, consistent with dynamic changes in clonal architecture.

TABLE 98-6

International Prognostic Scoring System (IPSS)

1. New marrow blast categories
   ≤2%, >2%–<5%, 5–10%, >10–30%
2. Refined cytogenetic abnormalities and risk groups
   16 (vs 6) specific abnormalities, 5 (vs 3) subgroups[a]
3. Evaluation of depth of cytopenias[b]
   Clinically and statistically relevant cutpoints used
4. Inclusion of differentiating features
   Age, performance status, serum ferritin, LDH; β_2-microglobulin
5. Prognostic model with 5 (vs 4) risk categories
   Improved predictive power

[a]Good, normal, –Y, del(5q), del (20q); poor, complex (≥3 abnormalities) or chromosome 7 abnormalities; intermediate, all other abnormalities. [b]Cytopenias defined as hemoglobin <100 g/L, platelet count <100,000/μL, and absolute neutrophil count <1500/μL.

Abbreviation: LDH, lactate dehydrogenase.

Most patients die as a result of complications of pancytopenia and not due to leukemic transformation; perhaps one-third succumb to diseases unrelated to their MDS. Precipitous worsening of pancytopenia, acquisition of new chromosomal abnormalities on serial cytogenetic determination, increase in the number of blasts, and marrow fibrosis are all poor prognostic indicators. The outlook in therapy-related MDS, regardless of type, is extremely poor, and most patients will progress within a few months to refractory AML.

TREATMENT

TREATMENT

Myelodysplasia

Historically, therapy of MDS has been unsatisfactory, but several drugs may not only improve blood counts but delay onset of leukemia and improve survival. The choice of therapy for an individual patient, administration of treatment, and management of toxicities are complicated and require hematologic expertise.

Only hematopoietic stem cell transplantation offers cure of MDS. The survival rate in selected patient cohorts is ~50% at 3 years but improving. Results using unrelated matched donors are now similar to those with siblings, and patients in their fifties and sixties have been successfully transplanted. Nevertheless, treatment-

related mortality and morbidity increase with recipient age. The transplant conundrum is that the high-risk patient (by IPSS score and presence of monosomal karyotype), for whom the procedure is most obviously indicated, has a high probability of a poor outcome from transplant-related mortality or disease relapse, whereas the low-risk patient, who is more likely to tolerate transplant, also may do well for years with less aggressive therapies. In practice, only a small proportion of MDS patients undergo transplantation.

MDS has been regarded as particularly refractory to cytotoxic chemotherapy regimens, and as in AML in the older adult, drug toxicity is frequent and often fatal, and remissions if achieved are brief. Low doses of cytotoxic drugs have been administered for their "differentiation" potential, and from this experience, drug therapies have emerged based on pyrimidine analogues. These drugs are classified as epigenetic modulators, believed to act through a demethylating mechanism to alter gene regulation and allow differentiation to mature blood cells from the abnormal MDS stem cell. The hypomethylating agents azacitidine and decitabine are frequently used in bone marrow failure clinics. Azacitidine improves blood counts and survival in MDS, compared to best supportive care. Azacitidine is usually administered subcutaneously, daily for 7 days, at 4-week intervals, for at least four cycles before assessing for response. Overall, generally improved blood counts with a decrease in transfusion requirements occurred in ~50% of patients in published trials. Response is dependent on continued drug administration, and most patients eventually become refractory to drug intervention and experience recurrent cytopenias or progression to AML. Decitabine is closely related to azacitidine and more potent; 30–50% of patients show responses in blood counts, with a duration of response of almost a year. Decitabine is usually administered by continuous intravenous infusion in regimens of varying doses and durations of 3–10 days in repeating cycles. The major toxicity of azacitidine and decitabine is myelosuppression, leading to worsening blood counts. Hypomethylating agents are frequently used in the high-risk patient who is not a candidate for stem cell transplant. In the lower risk patient, they are also effective, but alternative therapies should be considered.

Lenalidomide, a thalidomide derivative with a more favorable toxicity profile, is particularly effective in reversing anemia in MDS patients with 5q– syndrome; not only do a high proportion of these patients become transfusion independent with normal or near-normal hemoglobin levels, but their cytogenetics also become normal. The drug has many biologic activities, and it is unclear which is critical for clinical efficacy. Lenalidomide is administered orally. Most patients will improve within 3 months of initiating therapy. Toxicities include myelosuppression (worsening thrombocytopenia and neutropenia, necessitating blood count monitoring) and an increased risk of deep vein thrombosis and pulmonary embolism.

Immunosuppression also may produce sustained independence from transfusion and improve survival. ATG, cyclosporine, and the anti-CD52 monoclonal antibody alemtuzumab are especially effective in younger MDS patients (<60 years old) with more favorable IPSS scores and who bear the histocompatibility antigen HLA-DR15.

HGFs can improve blood counts but, as in most other marrow failure states, have been most beneficial to patients with the least severe pancytopenia. EPO alone or in combination with G-CSF can improve hemoglobin levels, particularly in those with low serum EPO levels who have no or a modest need for transfusions. Survival may be enhanced by EPO and amelioration of anemia. G-CSF treatment alone failed to

improve survival in a controlled trial. Thrombopoietin mimetics appear to improve platelet counts in some MDS patients, with no clear evidence that they increase the rate of leukemic transformation.

The same principles of supportive care described for aplastic anemia apply to MDS. Many patients will be anemic for years. RBC transfusion support should be accompanied by iron chelation to prevent secondary hemochromatosis.

MYELOPHTHISIC ANEMIAS

Fibrosis of the bone marrow **(see Fig. 96-2)**, usually accompanied by a characteristic blood smear picture called *leukoerythroblastosis*, can occur as a primary hematologic disease, called *myelofibrosis* or *myeloid metaplasia* **(Chap. 99)**, and as a secondary process, called *myelophthisis*. Myelophthisis, or secondary myelofibrosis, is reactive. Fibrosis can be a response to invading tumor cells, usually an epithelial cancer of breast, lung, or prostate origin or neuroblastoma. Marrow fibrosis may occur with infection of mycobacteria (both *Mycobacterium tuberculosis* and *Mycobacterium avium*), fungi, or HIV and in sarcoidosis. Intracellular lipid deposition in Gaucher disease and obliteration of the marrow space related to absence of osteoclast remodeling in congenital osteopetrosis also can produce fibrosis. Secondary myelofibrosis is a late consequence of radiation therapy or treatment with radiomimetic drugs. Usually the infectious or malignant underlying processes are obvious. Marrow fibrosis can also be a feature of a variety of hematologic syndromes, especially chronic myeloid leukemia, multiple myeloma, lymphomas, myeloma, and hairy cell leukemia.

The pathophysiology has three distinct features: proliferation of fibroblasts in the marrow space (myelofibrosis); the extension of hematopoiesis into the long bones and into extramedullary sites, usually the spleen, liver, and lymph nodes (myeloid metaplasia); and ineffective erythropoiesis. The etiology of the fibrosis is unknown but most likely involves dysregulated production of growth factors: platelet-derived growth factor and transforming growth factor β have been implicated. Abnormal regulation of other hematopoietins would lead to localization of blood-producing cells in nonhematopoietic tissues and uncoupling of the usually balanced processes of stem cell proliferation and differentiation. Myelofibrosis is remarkable for pancytopenia despite very large numbers of circulating hematopoietic progenitor cells.

Anemia is dominant in secondary myelofibrosis, usually normocytic and normochromic. The diagnosis is suggested by the characteristic leukoerythroblastic smear **(see Fig. 96-1)**. Erythrocyte morphology is highly abnormal, with circulating nucleated RBCs, teardrops, and shape distortions. WBC numbers are often elevated, sometimes mimicking a leukemoid reaction, with circulating myelocytes, promyelocytes, and myeloblasts. Platelets may be abundant and are often of giant size. Inability to aspirate the bone marrow, the characteristic “dry tap,” can allow a presumptive diagnosis in the appropriate setting before the biopsy is decalcified.

The course of secondary myelofibrosis is determined by its etiology, usually a metastatic tumor or an advanced hematologic malignancy. Treatable causes must be excluded, especially tuberculosis and fungus. Transfusion support can relieve symptoms.

FURTHER READING

Arber DA et al: The 2016 revision to the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia. Blood 127:2391, 2016. [PubMed: 27069254]

Bacigalupo A, Sica S: Alternative donor transplants for severe aplastic anemia: Current experience. Semin Hematol 53:1159, 2016.

Deeg HJ: Hematopoietic cell transplantation for myelodysplastic syndrome. Am Soc Clin Oncol Educ Book e375, 2015.

Desmond R et al: Eltrombopag restores trilineage hematopoiesis in refractory severe aplastic anemia that can be sustained on discontinuation of drug. Blood 123:1818, 2014. [PubMed: 24345753]

Garcia-Manero G: Myelodysplastic syndromes: 2015 Update on diagnosis, risk-stratification and management. Am J Hematol 90:831, 2016.

Greenberg PL et al: Revised international prognostic scoring system for myelodysplastic syndromes. Blood 120:2454, 2012. [PubMed: 22740453]

Komrokji RS, List AF: Short- and long-term benefits of lenalidomide treatment in patients with lower-risk del(5q) myelodysplastic syndromes. Ann Oncol 27:62, 2016. [PubMed: 26504152]

Mossner M et al: Mutational hierarchies in myelodysplastic syndromes dynamically adapt and evolve upon therapy response and failure. Blood 128:1246, 2016. [PubMed: 27268087]

Sloand EM et al: Alemtuzumab treatment of intermediate-1 myelodysplasia patients is associated with sustained improvement in blood counts and cytogenetic remissions. J Clin Oncol 28:5166, 2010. [PubMed: 21041705]

Yoshizato T et al: Somatic mutations and clonal hematopoiesis in aplastic anemia. N Engl J Med 373:35, 2015. [PubMed: 26132940]

Chapter 99: Polycythemia Vera and Other Myeloproliferative Neoplasms

Jerry L. Spivak

INTRODUCTION

The World Health Organization (WHO) classification of the chronic myeloproliferative neoplasms (MPNs) includes eight disorders, some of which are rare or poorly characterized (**Table 99-1**) but all of which share an origin in a hematopoietic cell; overproduction of one or more of the formed elements of the blood without significant dysplasia; and a predilection to extramedullary hematopoiesis, myelofibrosis, and transformation at varying rates to acute leukemia. Within this broad classification, however, significant phenotypic heterogeneity exists. Some diseases such as chronic myelogenous leukemia (CML), chronic neutrophilic leukemia (CNL), and chronic eosinophilic leukemia (CEL) express primarily a myeloid phenotype, whereas in other diseases, such as polycythemia vera (PV), primary myelofibrosis (PMF), and essential thrombocytosis (ET), erythroid or megakaryocytic hyperplasia predominates. The latter three disorders, in contrast to the former three, also appear capable of transforming into each other.

TABLE 99-1

World Health Organization Classification of Chronic Myeloproliferative Neoplasms

Chronic myeloid leukemia, bcr-abl–positive
Chronic neutrophilic leukemia
Chronic eosinophilic leukemia, not otherwise specified
Polycythemia vera
Primary myelofibrosis
Essential thrombocytosis
Mastocytosis
Myeloproliferative neoplasms, unclassifiable

Such phenotypic heterogeneity has a genetic basis; CML is the consequence of the balanced translocation between chromosomes 9 and 22 (t[9;22][q34;11]); CNL has been associated with a t(15;19) translocation; and CEL occurs with a deletion or balanced translocations involving the *PDGFRα* gene. By contrast, PV, PMF, and ET are characterized by driver mutations that directly or indirectly constitutively activate JAK2, a tyrosine kinase essential for the function of the erythropoietin and thrombopoietin receptors and also utilized by the granulocyte colony-stimulating factor receptor. This important distinction is reflected in the natural histories of CML, CNL, and CEL, which are usually measured in years, with a high rate of leukemic transformation. The natural histories of PV, PMF, and ET, by contrast, are usually measured in decades, and transformation to acute leukemia is uncommon in the absence of chemotherapy. This chapter focuses only on PV, PMF, and ET because their clinical features and driver mutation overlap are substantial, though their disease duration varies.

The other chronic MPNs will be discussed in Chaps. 101 and 106.

POLYCYTHEMIA VERA

PV is a clonal hematopoietic stem cell disorder in which phenotypically normal red cells, granulocytes, and platelets accumulate in the absence of a recognizable physiologic stimulus. The most common of the MPN, PV occurs in 2.5 per 100,000 persons, sparing no adult age group and increasing with age to rates over 10/100,000. Familial transmission is infrequent, and women predominate among sporadic cases.

ETIOLOGY



Nonrandom chromosome abnormalities such as deletion 20q and deletion 13q or trisomy 9 occur in up to 30% of untreated PV patients, but unlike CML, no consistent cytogenetic abnormality has been associated with the disorder. However, a mutation in the autoinhibitory pseudokinase domain of the tyrosine kinase JAK2 that replaces valine with phenylalanine (V617F), causing constitutive kinase activation—appears to have a central role in PV pathogenesis.

JAK2 is a member of an evolutionarily well-conserved, nonreceptor tyrosine kinase family and serves as the cognate tyrosine kinase for the erythropoietin and thrombopoietin receptors. It also functions as an obligate chaperone for these receptors in the Golgi apparatus and is responsible for their cell-surface expression. The conformational change induced in the erythropoietin and thrombopoietin receptors following binding to their respective cognate ligands, erythropoietin or thrombopoietin, leads to JAK2 autophosphorylation, receptor phosphorylation, and phosphorylation of proteins involved in cell proliferation, differentiation, and resistance to apoptosis. Transgenic animals lacking *JAK2* die as embryos from severe anemia. Constitutive activation of JAK2, on the other hand, explains the erythropoietin hypersensitivity, erythropoietin-independent erythroid colony formation, rapid terminal differentiation, increased Bcl-X_L expression, and apoptosis resistance in the absence of erythropoietin that characterize the in vitro behavior of PV erythroid progenitor cells.

More than 95% of PV patients express this mutation, as do ~50% of PMF and ET patients. Importantly, the *JAK2* gene is located on the short arm of chromosome 9, and loss of heterozygosity on chromosome 9p involving the segment containing the *JAK2* locus over time due to mitotic recombination (uniparental disomy), is the most common cytogenetic abnormality in PV. Loss of heterozygosity in this region leads to homozygosity for *JAK2* V617F and occurs in ~60% of PV patients and to a lesser extent in PMF but is rare in ET. Most PV patients who do not express *JAK2* V617F express a mutation in exon 12 of the gene and are not clinically different from those who do, with the exception of a higher frequency of isolated erythrocytosis, nor do *JAK2* V617F heterozygotes differ clinically from homozygotes. Importantly, the predisposition to acquire *JAK2* mutations appears to be associated with a specific *JAK2* gene haplotype, GGCC. *JAK2* V617F is the basis for many of the phenotypic and biochemical characteristics of PV such as increased blood cell production and increased inflammatory cytokine production; however, it cannot solely account for the entire PV phenotype and is probably not the initiating lesion in any of the MPN. First, PV patients with the same phenotype and documented clonal disease can have mutations in *LNK*, a JAK2 inhibitor, or rarely, calreticulin (*CALR*), an ER chaperone. Second, ET and PMF patients have the same mutation but different clinical phenotypes. Third, familial PV can occur without the mutation, even when other members of the same family express it. Fourth, not all the cells of the malignant clone express *JAK2* V617F. Fifth, inhibition of *JAK2* V617F-expressing hematopoietic progenitor cells by the nonspecific JAK1/2 kinase inhibitor, ruxolitinib, does not affect the behavior of the involved hematopoietic stem cells. Finally, in some *JAK2* V617F–positive PV or ET patients, acute leukemia can occur in a *JAK2* V617F–negative progenitor cell, suggesting the presence of an ancestral precursor cell.

CLINICAL FEATURES

Isolated thrombocytosis, leukocytosis, or splenomegaly may be the initial presenting manifestation of PV, but most often the disorder is first recognized by the incidental discovery of a high hemoglobin, hematocrit, or red cell count. With the exception of aquagenic pruritus, no symptoms distinguish PV from other causes of erythrocytosis.

Uncontrolled erythrocytosis causes hyperviscosity, leading to neurologic symptoms such as vertigo, tinnitus, headache, visual disturbances, and transient ischemic attacks (TIA). Systolic hypertension is also a feature of the red cell mass elevation. In some patients, venous or arterial thrombosis may be the presenting manifestation of PV. Any vessel can be affected; but cerebral, cardiac, and mesenteric vessels are most commonly involved. Hepatic venous thrombosis (Budd-Chiari syndrome) is particularly common in young women and may be catastrophic if sudden and complete obstruction of the hepatic vein occurs. Indeed, PV should be suspected in any patient who develops hepatic vein thrombosis. Digital ischemia, easy bruising, epistaxis, acid-peptic disease, or gastrointestinal hemorrhage may occur due to vascular stasis or thrombocytosis. In the latter instance, absorption and proteolysis of high molecular weight von Willebrand multimers by the large platelet mass causes acquired von Willebrand disease. Erythema, burning, and pain in the extremities, a symptom complex known as erythromelalgia, is another complication of thrombocytosis in PV due to increased platelet stickiness. Given the large turnover of hematopoietic cells, hyperuricemia with secondary gout, uric acid stones, and symptoms due to hypermetabolism can also complicate the disorder.

DIAGNOSIS

When PV presents with erythrocytosis in combination with leukocytosis, thrombocytosis, or splenomegaly or any combination of these, the diagnosis is apparent. However, when patients present with an elevated hemoglobin, hematocrit, or red cell count alone, the diagnostic evaluation is more complex because of the many diagnostic possibilities (**Table 99-2**). Furthermore, unless the hemoglobin level is ≥20 g/dL (hematocrit ≥60%), it is not possible to distinguish true erythrocytosis from disorders causing plasma volume contraction. This is because uniquely in PV, in contrast to other causes of true erythrocytosis, there is expansion of the plasma volume, which can mask the elevated red cell mass, particularly in women; thus, red cell mass and plasma volume determinations are necessary to establish the presence of an absolute erythrocytosis and distinguish this from relative erythrocytosis due to a reduction in plasma volume alone (also known as *stress* or *spurious erythrocytosis* or *Gaisböck's syndrome*). **Figure 59-18** illustrates a diagnostic algorithm for the evaluation of suspected erythrocytosis. Assay for *JAK2* mutations in the presence of a normal arterial oxygen saturation provides an alternative diagnostic approach to erythrocytosis when red cell mass and plasma volume determinations are not available; a normal serum erythropoietin level does not exclude the presence of PV, but an elevated erythropoietin level is more consistent with a secondary cause for the erythrocytosis.

TABLE 99-2

Causes of Erythrocytosis

Relative Erythrocytosis	
Hemoconcentration secondary to dehydration, diuretics, ethanol abuse, androgens, or tobacco abuse	
Absolute Erythrocytosis	
Hypoxia	**Tumors**
Carbon monoxide intoxication	Hypernephroma
High-oxygen-affinity hemoglobins	Hepatoma
High altitude	Cerebellar hemangioblastoma
Pulmonary disease	Uterine myoma
Right-to-left cardiac or vascular shunts	Adrenal tumors
Sleep apnea syndrome	Meningioma
Hepatopulmonary syndrome	Pheochromocytoma
Renal Disease	**Drugs**
Renal artery stenosis	Androgens
Focal sclerosing or membranous glomerulonephritis	Recombinant erythropoietin
Postrenal transplantation	**Familial (with normal hemoglobin function)**
Renal cysts	Erythropoietin receptor mutations
Bartter's syndrome	*VHL* mutations (Chuvash polycythemia)
	2,3-BPG mutation
	PHD2 and HIF2α mutations
	Polycythemia vera

Abbreviations: 2,3-BPG, 2,3-bisphosphoglycerate; VHL, von Hippel-Lindau.

Other laboratory studies that may aid in diagnosis include the red cell count, mean corpuscular volume, and red cell distribution width (RDW), particularly when the hematocrit or hemoglobin levels are less than 60% or 20 g/dL, respectively. Only three situations cause microcytic erythrocytosis: β-thalassemia trait, hypoxic erythrocytosis, and PV. With β-thalassemia trait, the RDW is usually normal, whereas with hypoxic erythrocytosis and PV, the RDW may be elevated due to associated iron deficiency. Today, however, the assay for *JAK2* V617F has superseded other tests for establishing the diagnosis of PV. Of course, in patients with associated acid-peptic disease, occult gastrointestinal bleeding may lead to a presentation with hypochromic, microcytic anemia, masking the presence of PV.

A bone marrow aspirate and biopsy provide no specific diagnostic information because these may be normal or indistinguishable from ET or PMF. Similarly, no specific cytogenetic abnormality is associated with the

disease, and the absence of a cytogenetic marker does not exclude the diagnosis.

COMPLICATIONS

Many of the clinical complications of PV relate directly to the increase in blood viscosity associated with red cell mass elevation and indirectly to the increased turnover of red cells, leukocytes, and platelets with the attendant increase in uric acid and inflammatory cytokine production. The latter appears to be responsible for constitutional symptoms. Peptic ulcer disease can also be due to *Helicobacter pylori* infection, the incidence of which is increased in PV, while the pruritus associated with this disorder may be a consequence of mast cell activation by JAK2 V617F. A sudden increase in spleen size can be associated with painful splenic infarction. Myelofibrosis appears to be part of the natural history of the disease but is a reactive, reversible process that does not itself impede hematopoiesis and by itself has no prognostic significance. In ~15% of patients, however, myelofibrosis is associated with hematopoietic stem cell failure, manifested by substantial extramedullary hematopoiesis in the liver and spleen and transfusion-dependent anemia. The organomegaly can cause significant mechanical discomfort, portal hypertension, and progressive cachexia. Although the incidence of acute myeloid leukemia is increased in PV, the incidence of acute leukemia in patients not exposed to chemotherapy or radiation therapy is low. Interestingly, chemotherapy, including hydroxyurea, has been associated with acute leukemia in *JAK2* V617F–negative stem cells in some PV patients. *Erythromelalgia* is a curious syndrome of unknown etiology associated with thrombocytosis, primarily involving the lower extremities and usually manifested by erythema, warmth, and pain of the affected appendage and occasionally digital infarction. It occurs with a variable frequency and is usually responsive to salicylates. Some of the central nervous system symptoms observed in patients with PV, such as ocular migraine, appear to represent a variant of erythromelalgia.

Left uncontrolled, erythrocytosis can lead to thrombosis involving vital organs such as the liver, heart, brain, or lungs. Patients with massive splenomegaly are particularly prone to thrombotic events because the associated increase in plasma volume masks the true extent of the red cell mass elevation measured by the hematocrit or hemoglobin level. A “normal” hematocrit or hemoglobin level in a PV patient with massive splenomegaly should be considered indicative of an elevated red cell mass until proven otherwise.

TREATMENT

TREATMENT

Polycythemia Vera

PV is generally an indolent disorder, the clinical course of which is measured in decades, and its management should reflect its tempo. Thrombosis due to erythrocytosis is the most significant complication and often the presenting manifestation; maintenance of the hemoglobin level at ≤140 g/L (14 g/dL; hematocrit <45%) in men and ≤120 g/L (12 g/dL; hematocrit <42%) in women is mandatory to avoid thrombotic complications. Phlebotomy serves initially to reduce hyperviscosity by reducing the red cell mass to normal while further expanding the plasma volume. Periodic phlebotomies thereafter serve to maintain the red cell mass within

the normal range and induce a state of iron deficiency that prevents accelerated reexpansion of the red cell mass. In most PV patients, once an iron-deficient state is achieved, phlebotomy is usually only required at 3-month intervals. Neither phlebotomy nor iron deficiency increases the platelet count relative to the effect of the disease itself, and neither thrombocytosis nor leukocytosis are correlated with thrombosis in PV, in contrast to the strong correlation between erythrocytosis and thrombosis. The use of salicylates to prevent thrombosis in PV patients is not only potentially harmful if the red cell mass is not controlled by phlebotomy, but is also an unproven remedy. Anticoagulants are indicated when a thrombosis has occurred and can be difficult to monitor if the red cell mass is substantially elevated owing to the artifactual imbalance between the test tube anticoagulant and plasma that occurs when blood from these patients is assayed for prothrombin or partial thromboplastin activity. Asymptomatic hyperuricemia (<10 mg/dL) requires no therapy, but allopurinol should be administered to avoid further elevation of the uric acid when chemotherapy is used to reduce splenomegaly or leukocytosis or to treat pruritus. Generalized pruritus intractable to antihistamines or antidepressants such as doxepin can be a major problem in PV; the JAK1/2 inhibitor, ruxolitinib, pegylated interferon α (IFN-α), psoralens with ultraviolet light in the A range (PUVA) therapy, and hydroxyurea are other methods of palliation. Asymptomatic thrombocytosis requires no therapy unless the platelet count is sufficiently high to cause bleeding due to acquired von Willebrand's disease, but bleeding in this situation is not usually spontaneous and is responsive to ε-aminocaproic acid. Symptomatic splenomegaly can be treated with either ruxolitinib or pegylated IFN-α. Pegylated IFN-α has the advantage over recombinant IFN-α of being better tolerated and requiring only weekly administration and produced complete hematologic and molecular remissions in ~20% of PV patients; its role in this disorder is currently under investigation. Anagrelide, a phosphodiesterase inhibitor, can reduce the platelet count and, if tolerated, is preferable to hydroxyurea because it lacks marrow toxicity and is also protective against venous thrombosis while hydroxyurea is not. A reduction in platelet number may be necessary for the treatment of erythromelalgia or ocular migraine if salicylates are not effective or if the platelet count is sufficiently high to increase the risk of hemorrhage, but only to the degree that symptoms are alleviated. Alkylating agents and radioactive sodium phosphate (^{32}P) are leukemogenic in PV, and their use should be avoided. If a cytotoxic agent must be used, hydroxyurea is preferred, but this drug does not prevent either thrombosis or myelofibrosis in PV, is itself leukemogenic, and should be used for as short a time as possible. Previously, PV patients with massive splenomegaly unresponsive to reduction by chemotherapy or interferon required splenectomy. However, with the introduction of the nonspecific JAK2 inhibitor ruxolitinib, it has been possible in the majority of patients with PV complicated by myelofibrosis and myeloid metaplasia to reduce spleen size while at the same time alleviating constitutional symptoms and pruritus due to cytokine release and reducing the phlebotomy requirement. Ruxolitinib has also been demonstrated in a phase three clinical trial to be effective in PV patients without myelofibrosis who are intolerant or refractory to hydroxyurea or best available supportive therapy. In some patients with end-stage disease, pulmonary hypertension may develop due to fibrosis or extramedullary hematopoiesis. A role for bone marrow transplantation, either allogeneic or haploidentical, in PV has not been defined.

Most patients with PV can live long lives without functional impairment when their red cell mass is effectively managed with phlebotomy alone. Chemotherapy is never indicated to control the red cell mass unless venous access is inadequate.

PRIMARY MYELOFIBROSIS

Chronic PMF (other designations include *idiopathic myelofibrosis*, *agnogenic myeloid metaplasia*, or *myelofibrosis with myeloid metaplasia*) is a clonal hematopoietic stem cell disorder associated with mutations in *JAK2, MPL* or *CALR* and characterized by marrow fibrosis, extramedullary hematopoiesis, and splenomegaly. PMF is the least common MPN, and establishing its diagnosis in the absence of a specific clonal marker is difficult because myelofibrosis and splenomegaly are also features of both PV and CML. Furthermore, myelofibrosis and splenomegaly also occur in a variety of benign and malignant disorders (**Table 99-3**), many of which are amenable to specific therapies not effective in PMF. In contrast to the other MPN and so-called acute or malignant myelofibrosis, which can occur at any age, PMF primarily afflicts men in their sixth decade or later.

TABLE 99-3

Disorders Causing Myelofibrosis

Malignant	Nonmalignant
Acute leukemia (lymphocytic, myelogenous, megakaryocytic)	HIV infection
Chronic myeloid leukemia	Hyperparathyroidism
Hairy cell leukemia	Renal osteodystrophy
Hodgkin's disease	Systemic lupus erythematosus
Primary myelofibrosis	Tuberculosis
Lymphoma	Vitamin D deficiency
Multiple myeloma	Thorium dioxide exposure
Myelodysplasia	Gray platelet syndrome
Metastatic carcinoma	
Polycythemia vera	
Systemic mastocytosis	

ETIOLOGY

Nonrandom chromosome abnormalities such as 9p, 20q−, 13q−, trisomy 8 or 9, or partial trisomy 1q are common in PMF, but no cytogenetic abnormality specific to the disease has been identified. *JAK2* V617F is present in ~50% of PMF patients, and mutations in the thrombopoietin receptor, *MPL*, occur in about 8%. Most of the rest have mutations in the calreticulin gene (*CALR*) that alter the carboxy-terminal portion of the protein, permitting it to bind and activate MPL. The degree of myelofibrosis and the extent of extramedullary hematopoiesis are not related. Fibrosis in this disorder is associated with overproduction of transforming growth factor β and tissue inhibitors of metalloproteinases, while osteosclerosis is associated with overproduction of osteoprotegerin, an osteoclast inhibitor. Marrow angiogenesis occurs due to increased

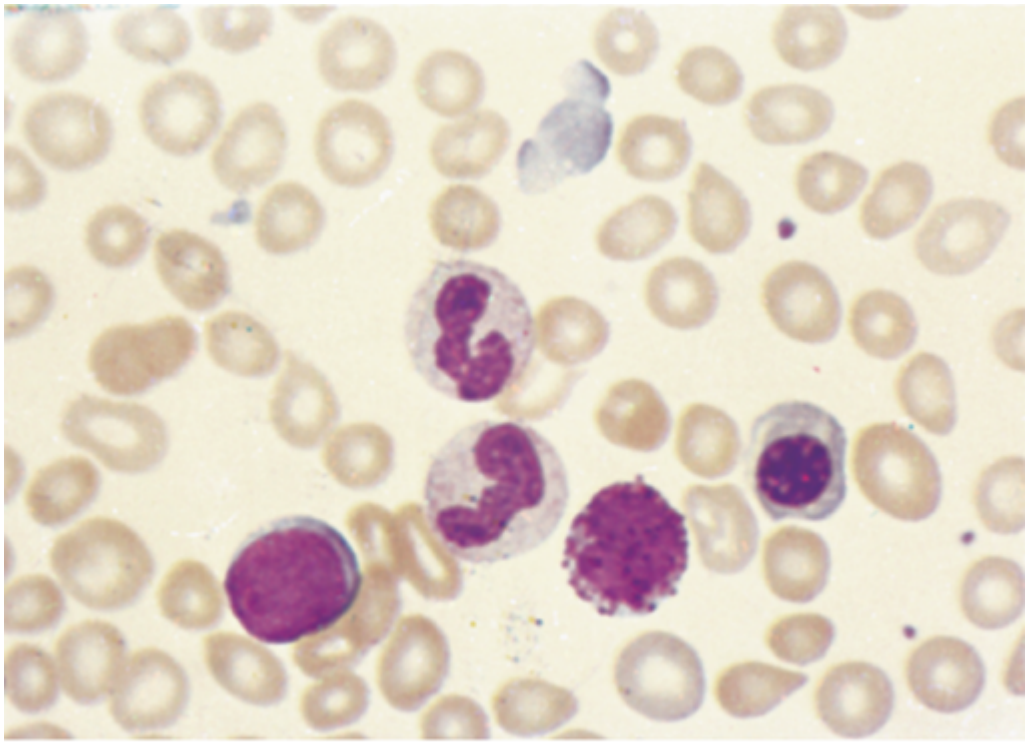
production of vascular endothelial growth factor. Importantly, fibroblasts in PMF are polyclonal and not part of the neoplastic clone but can be induced by it to produce inflammatory cytokines.

CLINICAL FEATURES

No signs or symptoms are specific for PMF. Many patients are asymptomatic at presentation, and the disease is usually detected by the discovery of splenic enlargement and/or abnormal blood counts during a routine examination. In contrast to its companion MPN, night sweats, fatigue, and weight loss are common presenting complaints. A blood smear will show the characteristic features of extramedullary hematopoiesis: teardrop-shaped red cells, nucleated red cells, myelocytes, and promyelocytes; myeloblasts may also be present **(Fig. 99-1)**. Anemia, usually mild initially, is common, whereas the leukocyte and platelet counts are either normal or increased, but either can be depressed. Mild hepatomegaly may accompany the splenomegaly but is unusual in its absence; isolated lymphadenopathy should suggest another diagnosis. Both serum lactate dehydrogenase and alkaline phosphatase levels can be elevated. Marrow is usually inaspirable due to the myelofibrosis **(Fig. 99-2)**, and bone x-rays may reveal osteosclerosis. Exuberant extramedullary hematopoiesis can cause ascites; portal, pulmonary, or intracranial hypertension; intestinal or ureteral obstruction; pericardial tamponade; spinal cord compression; or skin nodules. Splenic enlargement can be sufficiently rapid to cause splenic infarction with fever and pleuritic chest pain. Hyperuricemia and secondary gout may ensue.

FIGURE 99-1

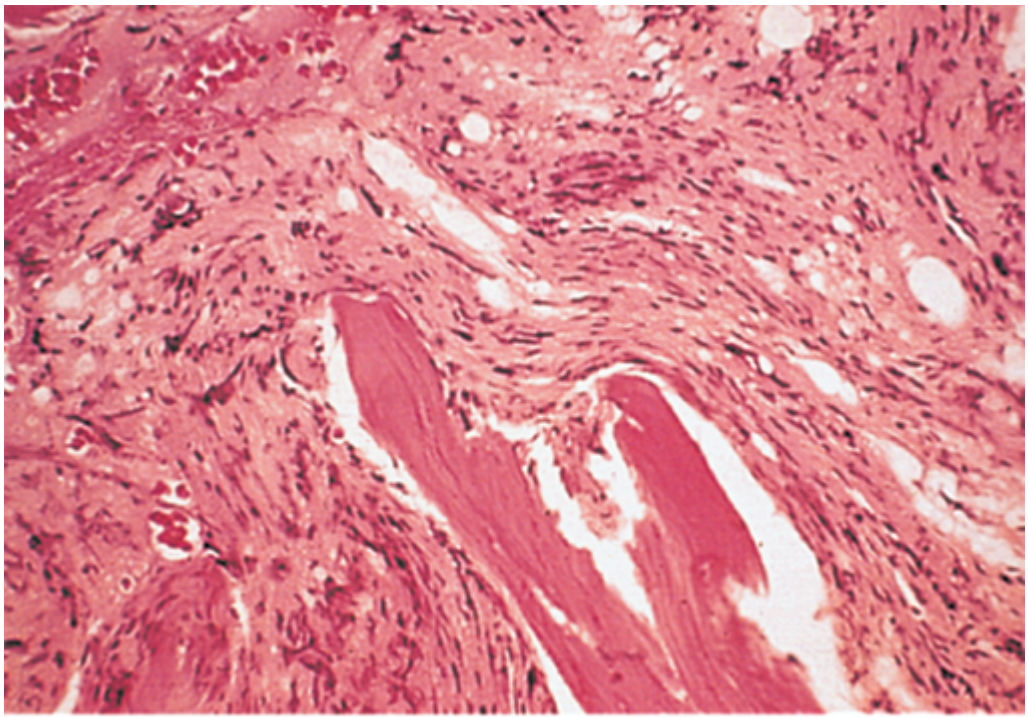
Teardrop-shaped red blood cells indicative of membrane damage from passage through the spleen, a nucleated red blood cell, and immature myeloid cells indicative of extramedullary hematopoiesis are noted. This peripheral blood smear is related to any cause of extramedullary hematopoiesis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 99-2

This marrow section shows the marrow cavity replaced by fibrous tissue composed of reticulin fibers and collagen. When this fibrosis is due to a primary hematologic process, it is called *myelofibrosis*. When the fibrosis is secondary to a tumor or a granulomatous process, it is called *myelophthisis*.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS

While the clinical picture described above is characteristic of PMF, all of these clinical features can be observed in PV or CML. Massive splenomegaly commonly masks erythrocytosis in PV, and reports of intraabdominal thrombosis in PMF most likely represent instances of unrecognized PV. In some PMF patients, erythrocytosis has developed during the course of the disease. Furthermore, because many other disorders have features that overlap with PMF but respond to distinctly different therapies, the diagnosis of PMF is one of exclusion, which requires that the disorders listed in Table 99-3 be ruled out.

The presence of teardrop-shaped red cells, nucleated red cells, myelocytes, and promyelocytes establishes the presence of extramedullary hematopoiesis, while the presence of leukocytosis, thrombocytosis with large and bizarre platelets, and circulating myelocytes suggests the presence of an MPN as opposed to a secondary form of myelofibrosis (Table 99-3). Marrow is usually inaspirable due to increased marrow reticulin, but marrow biopsy will reveal a hypercellular marrow with trilineage hyperplasia and, in particular, increased numbers of megakaryocytes in clusters and with large, dysplastic nuclei. However, there are no characteristic bone marrow morphologic abnormalities that distinguish PMF from the other MPN. Splenomegaly due to extramedullary hematopoiesis may be sufficiently massive to cause portal hypertension and variceal formation. In some patients, exuberant extramedullary hematopoiesis can dominate the clinical picture. An intriguing feature of PMF is the occurrence of autoimmune abnormalities such as immune complexes, antinuclear antibodies, rheumatoid factor, or a positive Coombs' test. Whether these represent a host reaction to the disorder or are involved in its pathogenesis is unknown. Cytogenetic analysis of the blood is useful both to exclude CML and for prognostic purposes because the development of complex karyotype abnormalities portends a poor prognosis in PMF. For unknown reasons, the number of circulating CD34+ cells is markedly increased in PMF (>15,000/μL) compared to the other MPN, unless they too develop extramedullary hematopoiesis.

Importantly, ~50% of PMF patients, like patients with its companion MPN, express the *JAK2* V617F mutation, often as homozygotes. Such patients are usually older and have higher hematocrits than patients with *MPL* (8%) or *CALR* (30%) mutations; PMF patients expressing an *MPL* mutation tend to be more anemic and have lower leukocyte counts than *JAK2* V617F-positive patients. Somatic mutations (due to deletions [type 1] or insertions [type 2]) in exon 9 of *CALR* have been found in a majority of patients with PMF who lack mutations in either *JAK2* or *MPL*. In some studies, type 1 mutations, the most common CALR mutation in PMF, had a survival advantage compared to *JAK2* or *MPL* mutations but not with respect to leukemic transformation. PMF patients who lack a known driver mutation have the worst prognosis.

COMPLICATIONS

Survival in PMF varies according to specific risk factors at diagnosis **(Tables 99-4 and 99-5)** but is shorter than in PV and ET patients. The natural history of PMF is one of increasing marrow failure with transfusion-dependent anemia and increasing organomegaly due to extramedullary hematopoiesis. As with CML, PMF can evolve from a chronic to an accelerated phase with constitutional symptoms and increasing marrow failure. About 10% of patients spontaneously transform to an aggressive form of acute leukemia for which

therapy is usually ineffective. Additional important prognostic factors for disease acceleration during the course of PMF include the presence of complex cytogenetic abnormalities, thrombocytopenia, and transfusion-dependent anemia. Mutations in the *ASXL1*, *EZH2*, *SRSF2*, and *IDH1/2* genes have been identified as risk factors for early death or transformation to acute leukemia and may prove to be more useful for PMF risk assessment than clinical scoring systems.

TABLE 99-4

Three Current Scoring Systems for Estimating Prognosis in PMF Patients

Risk Factor	IPSS (2009)[a]	DIPSS (2010)[b]	DIPSS Plus (2011)[c]
Anemia (<10 g/dL)	X	X	X
Leukocytosis (>25,000/μL)	X	X	X
Peripheral blood blasts (≥1%)	X	X	X
Constitutional symptoms	X	X	X
Age (>65 years)	X	X	X
Unfavorable karyotype			X
Platelet count (<100,000/μL)			X
Transfusion dependence			X

[a]Blood 113:2895, 2009. [b]Blood 115:1703, 2010. [c]J Clin Oncol 29:392, 2011.

Note: The Dynamic International Prognostic Scoring System (DIPSS) was developed to determine if the International Prognostic Scoring System (IPSS) risk factors identified as important for survival at the time of primary myelofibrosis (PMF) diagnosis could also be used for risk stratification following their acquisition during the course of the disease. One point is assigned to each risk factor for IPSS scoring. For DIPSS, the same is true, but anemia is assigned 2 points. The DIPSS Plus scoring system represents recognition that the addition of unfavorable karyotype, thrombocytopenia, and transfusion dependence improved the DIPSS risk stratification system for which additional points are assigned (Table 99-5). More recent studies suggest that mutational analysis of the *ASXL1, EZH2, SRSF2*, and *IDH1/2* genes further improves risk stratification for survival and leukemic transformation (Leukemia 27:1861, 2013). These prognostic scoring systems are not accurate for risk assessment in PV or ET patients who have developed myelofibrosis (Haematologica 99:e55, 2014).

TABLE 99-5

IPSS and DIPSS Risk Stratification Systems

Risk Categories[a]	Number of Risk Factors		
	IPSS	DIPSS	DIPSS PLUS
Low	0	0	0
Intermediate-1	1	1–2	1
Intermediate-2	2	3–4	2–3
High	≥3	>4	4–6

[a]The corresponding survival curves for each risk category can be found in the references cited in the footnotes of Table 99-4.

Abbreviations: DIPSS, Dynamic International Prognostic Scoring System; IPSS, International Prognostic Scoring System.

TREATMENT

TREATMENT

Primary Myelofibrosis

No specific therapy exists for PMF. The causes for anemia are multifarious and include ineffective erythropoiesis uncompensated by splenic extramedullary hematopoiesis, hemodilution due to splenomegaly, splenic sequestration, blood loss secondary to thrombocytopenia or portal hypertension, folic acid deficiency, systemic inflammation, and autoimmune hemolysis. Neither recombinant erythropoietin nor androgens such as danazol have proven to be consistently effective as therapy for anemia. Erythropoietin may worsen splenomegaly and will be ineffective if the serum erythropoietin level is >125 mU/L. Given the inflammatory milieu that characterizes PMF, glucocorticoids can ameliorate anemia as well as constitutional symptoms such as fever, chills, night sweats, anorexia, and weight loss, and combining these with low-dose thalidomide has proved effective as well. Thrombocytopenia can be due to impaired marrow function, splenic sequestration, or autoimmune destruction and may also respond to low-dose thalidomide and prednisone.

Splenomegaly is by far the most distressing and intractable problem for PMF patients, causing abdominal pain, portal hypertension, easy satiety, and cachexia, whereas surgical removal of a massive spleen is

associated with significant postoperative complications including mesenteric venous thrombosis, hemorrhage, rebound leukocytosis and thrombocytosis, and hepatic extramedullary hematopoiesis with no amelioration of either anemia or thrombocytopenia when present. For unexplained reasons, splenectomy also increases the risk of blastic transformation. Splenic irradiation is, at best, temporarily palliative and associated with a significant risk of neutropenia, infection, and subsequent operative hemorrhage if splenectomy is attempted. Allopurinol can control significant hyperuricemia, and bone pain can be alleviated by local irradiation. Pegylated IFN-α can ameliorate fibrosis in early PMF, but in advanced disease, it may exacerbate the bone marrow failure. The JAK2 inhibitor, ruxolitinib, has proved effective in reducing splenomegaly and alleviating constitutional symptoms in a majority of advanced PMF patients while also prolonging survival, although it does not significantly influence the *JAK2* V617F neutrophil allele burden. Although anemia and thrombocytopenia are its major side effects, these are dose-dependent, and with time, anemia stabilizes and thrombocytopenia may improve. Allogeneic bone marrow transplantation is the only curative treatment for PMF and should be considered in younger patients and older patients with high risk disease; nonmyeloablative conditioning regimens may permit hematopoietic cell transplantation to be extended to older individuals, and is currently under investigation.

ESSENTIAL THROMBOCYTOSIS

ET (other designations include *essential thrombocythemia*, *idiopathic thrombocytosis*, *primary thrombocytosis*, and *hemorrhagic thrombocythemia*) is a clonal hematopoietic stem cell disorder associated with mutations in *JAK2* (V617F), *MPL*, and *CALR* and manifested clinically by overproduction of platelets without a definable cause. ET has an incidence of 1–2/100,000 and a distinct female predominance in association with *JAK2* V617F, *CALR*, and *MPL* mutations. Canonical MPN driver mutations distinguish 90% of ET patients from the more common nonclonal, reactive forms of thrombocytosis **(Table 99-6);** mutation-negative ET patients may have an hereditary form of thrombocytosis. Once considered a disease of the elderly and responsible for significant morbidity due to hemorrhage or thrombosis, it is now clear that ET can occur at any age in adults and often without symptoms or disturbances of hemostasis. There is an unexplained female predominance in contrast to PMF or the reactive forms of thrombocytosis where no sex difference exists. Because no specific clonal marker is available, clinical and laboratory criteria have been proposed to distinguish ET from other MPN, which may also present with initially with isolated thrombocytosis but have differing prognoses and therapies (Table 99-6). These criteria are useful in identifying disorders such as CML, PV, PMF, or myelodysplasia, which can masquerade as ET. Furthermore, as with “idiopathic” erythrocytosis, nonclonal benign forms of thrombocytosis exist (such as hereditary overproduction of thrombopoietin and those with noncanonical *JAK2* driver mutations) that are not widely recognized because we currently lack diagnostic assays. Approximately 55% of ET patients express *JAK2* V617F, 36% *CALR* (both type 1 and type 2) and 4% *MPL* mutations. ET patients lacking a canonical MPN driver mutation usually have a benign prognosis.

TABLE 99-6

Causes of Thrombocytosis

Tissue inflammation: collagen vascular disease, inflammatory bowel disease	Hemorrhage
Malignancy	Iron-deficiency anemia
Infection	Surgery
Myeloproliferative disorders: polycythemia vera, primary myelofibrosis, essential thrombocytosis, chronic myelogenous leukemia	Rebound: Correction of vitamin B_{12} or folate deficiency, post-ethanol abuse
Myelodysplastic disorders: 5q−syndrome, idiopathic refractory sideroblastic anemia	Hemolysis
Postsplenectomy or hyposplenism	Familial: Thrombopoietin overproduction, *JAK2* or *MPL* mutations

ETIOLOGY

Megakaryocytopoiesis and platelet production depend on thrombopoietin and its receptor MPL. As in the case of early erythroid and myeloid progenitor cells, early megakaryocytic progenitors require the presence of interleukin 3 (IL-3) and stem cell factor for optimal proliferation in addition to thrombopoietin. Their subsequent terminal development is also enhanced by the chemokine stromal cell-derived factor 1 (SDF-1). Interestingly, terminal megakaryocyte maturation and platelet production do not require thrombopoietin.

Megakaryocytes are unique amongst hematopoietic progenitor cells because reduplication of their genome is endomitotic rather than mitotic and promoted by thrombopoietin. Unlike erythropoietin, thrombopoietin is produced primarily in the liver and to a lesser extent in other organs, most importantly the bone marrow where it functions to maintain hematopoietic stem cells quiescent in the endosteal niche; once released, thrombopoietin promotes the proliferation of these cells in the sinusoidal niche. An inverse correlation exists between the platelet count and plasma thrombopoietin. However, again unlike erythropoietin, thrombopoietin is only constitutively produced and the plasma thrombopoietin level is controlled by the size of the megakaryocyte progenitor cell pool. Also, in contrast to erythropoietin, but like its myeloid counterparts, granulocyte and granulocyte-macrophage colony-stimulating factors, thrombopoietin not only

enhances the proliferation of its target cells but also enhances the reactivity of their end-stage product, the platelet.

The clonal nature of ET was established by analysis of glucose-6-phosphate dehydrogenase isoenzyme expression in patients hemizygous for this gene. Although thrombocytosis is its principal manifestation, like the other MPN, a hematopoietic stem cell is involved in ET. Furthermore, a number of families have been described in which ET was inherited, in one instance as an autosomal dominant trait. In addition to ET, PMF and PV have also been observed in such kindreds.

CLINICAL FEATURES

Clinically, ET is most often identified incidentally when a platelet count is obtained during the course of a routine medical evaluation. Occasionally, review of previous blood counts will reveal that an elevated platelet count was present but overlooked for many years. No symptoms or signs are specific for ET, but these patients can have hemorrhagic and thrombotic tendencies expressed as easy bruising for the former and microvascular occlusive events for the latter such as erythromelalgia, ocular migraine, or a TIA. Physical examination is generally unremarkable except occasionally for mild splenomegaly. Significant splenomegaly is indicative of another MPN, in particular PV, PMF, or CML.

Anemia is unusual, but a mild neutrophilic leukocytosis is not. The blood smear is most remarkable for the number of platelets present, some of which may be very large. The large mass of circulating platelets may prevent the accurate measurement of serum potassium due to release of platelet potassium upon blood clotting. This type of hyperkalemia is a test tube artifact and not associated with electrocardiographic abnormalities. Similarly, arterial oxygen measurements can be inaccurate unless thrombocythemic blood is collected on ice. The prothrombin and partial thromboplastin times are normal, whereas abnormalities of platelet function such as a prolonged bleeding time and impaired platelet aggregation can be present. However, despite much study, no platelet function abnormality is characteristic of ET, and no platelet function test predicts the risk of clinically significant bleeding or thrombosis.

The elevated platelet count may hinder marrow aspiration, but marrow biopsy usually reveals megakaryocyte hypertrophy and hyperplasia, as well as an overall increase in marrow cellularity. If marrow reticulin is increased, another diagnosis should be considered. The absence of stainable iron demands an explanation because iron deficiency alone can cause thrombocytosis, and absent marrow iron in the presence of marrow hypercellularity is a feature of PV.

Nonrandom cytogenetic abnormalities occur in ET but are uncommon, and no specific or consistent abnormality is notable, even those involving chromosomes 3 and 1, where the genes for thrombopoietin and its receptor, MPL, respectively, are located.

DIAGNOSIS

Thrombocytosis is encountered in a broad variety of clinical disorders (Table 99-6), in many of which inflammatory cytokine production is increased. The absolute level of the platelet count is not a useful diagnostic aid for distinguishing between benign and clonal causes of thrombocytosis. About 55% of ET patients express the *JAK2* V617F mutation. When *JAK2* V617F is absent, cytogenetic evaluation is mandatory to determine if the thrombocytosis is due to CML or a myelodysplastic disorder such as the 5q− syndrome. Because the bcr-abl translocation can be present in the absence of the Ph chromosome, and because bcr-abl reverse transcriptase polymerase chain reaction is associated with false-positive results, fluorescence in situ hybridization (FISH) analysis for bcr-abl is the preferred assay in patients with thrombocytosis in whom a cytogenetic study for the Ph chromosome is negative. *CALR* (type 1 or type 2) are present in 36%, and *MPL* mutations are present in 4% respectively, of ET patients who do not have a *JAK2* mutation. Anemia and ringed sideroblasts are not features of ET, but they are features of idiopathic refractory sideroblastic anemia, and in some of these patients, the thrombocytosis occurs in association with expression of *JAK2* V617F, *CALR*, or an *MPL* mutation. Significant splenomegaly should suggest the presence of another MPN, and in this setting, a red cell mass determination should be performed because splenomegaly can mask the presence of erythrocytosis. Importantly, what appears to be ET can evolve into PV (usually in women with *JAK2* V617F) or PMF (usually in men with type 1 CALR mutations) after a period of many years due to clonal evolution or succession. There is sufficient overlap of the *JAK2* V617F neutrophil allele burden between ET and PV that this cannot be used as a distinguishing diagnostic feature; only a red cell mass and plasma volume determination can distinguish PV from ET, and importantly in this regard, 64% of *JAK2* V617F–positive ET patients in one study actually were found to have PV when red cell mass and plasma volume determinations were performed. Claims that ET and PV form a biological continuum are unfounded as these disorders have different gene expression profiles and different natural histories.

COMPLICATIONS

Perhaps no other condition in clinical medicine has caused otherwise astute physicians to intervene inappropriately more often than thrombocytosis, particularly if the platelet count is >1 × 10^6/μL. It is commonly believed that a high platelet count causes thrombosis; however, no controlled clinical study has ever established this association, and in patients younger than age 60 years, the incidence of thrombosis was not greater in patients with thrombocytosis than in age-matched controls, and tobacco use appears to be the most important risk factor for thrombosis in ET patients.

To the contrary, very high platelet counts are associated primarily with hemorrhage due to acquired von Willebrand's disease. This is not meant to imply that an elevated platelet count cannot cause symptoms in an ET patient, but rather that the focus should be on the patient, not the platelet count. For example, some of the most dramatic neurologic problems in ET are migraine-related and respond only to lowering of the platelet count, whereas other symptoms such as erythromelalgia respond simply to platelet cyclooxygenase-1 inhibitors such as aspirin or ibuprofen, without a reduction in platelet number. Still others may represent an interaction between an atherosclerotic vascular system and a high platelet count, and others may have no relationship to the platelet count whatsoever. Recognition that PV can present with thrombocytosis alone as

well as the discovery of previously unrecognized causes of hypercoagulability (**Chap. 113**) make the older literature on the complications of thrombocytosis unreliable.

TREATMENT

TREATMENT

Essential Thrombocytosis

Survival of ET patients is not different than the general population regardless of their driver mutation. An elevated platelet count in an asymptomatic patient without cardiovascular risk factors or tobacco use requires no therapy. Indeed, before any therapy is initiated in a patient with thrombocytosis, the cause of symptoms must be clearly identified as due to the elevated platelet count. When the platelet count rises above $1 \times 10^6/\mu L$, a substantial quantity of high-molecular-weight von Willebrand multimers are removed from the circulation and destroyed by the enlarged platelet mass, resulting in an acquired form of von Willebrand's disease. This can be identified by a reduction in ristocetin cofactor activity. In this situation, aspirin could promote hemorrhage. Bleeding in this situation is rarely spontaneous and usually responds to ε-aminocaproic acid, which can be given prophylactically before and after elective surgery. Plateletpheresis is at best a temporary and inefficient remedy that is rarely required. Importantly, ET patients treated with ^{32}P or alkylating agents are at risk of developing acute leukemia without any proof of benefit; combining either therapy with hydroxyurea increases this risk. If platelet reduction is deemed necessary on the basis of symptoms refractory to salicylates alone, pegylated IFN-α, the quinazoline derivative, anagrelide, or hydroxyurea can be used to reduce the platelet count, but none of these is uniformly effective or without significant side effects. Hydroxyurea and aspirin are more effective than anagrelide and aspirin for prevention of TIA because hydroxyurea is an NO donor, but not more effective for the prevention of other types of arterial thrombosis and actually less effective for venous thrombosis. The risk of gastrointestinal bleeding is also higher when aspirin is combined with anagrelide. Normalizing the platelet count does not prevent either arterial or venous thrombosis. Pegylated interferon can produce a complete molecular remission in some ET patients, but a role for it or ruxolitinib in ET management has not yet been established.

As more clinical experience is acquired, ET appears more benign than previously thought. Evolution to acute leukemia is more likely to be a consequence of therapy than of the disease itself. In managing patients with thrombocytosis, the physician's first obligation is to do no harm.

FURTHER READING

Alvarez-Larran A et al: Antiplatelet therapy versus observation in low-risk essential thrombocythemia with CALR mutation. Haematologica 101:926, 2016. [PubMed: 27175028]

Lamy T et al: Inapparent polycythemia vera: An unrecognized diagnosis. Am J Med 102:14, 1997. [PubMed: 9209196]

Marchioli R et al: Cardiovascular events and intensity of treatment in polycythemia vera. N Engl J Med 368:22, 2013. [PubMed: 23216616]

Vannucchi AM et al: Ruxolitinib versus standard therapy for the treatment of polycythemia vera. N Engl J Med 372:426, 2015. [PubMed: 25629741]

Verstovsek S et al: A double-blind, placebo-controlled trial of ruxolitinib for myelofibrosis. N Engl J Med 366:799, 2012. [PubMed: 22375971]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 100: Acute Myeloid Leukemia

William Blum; Clara D. Bloomfield

INCIDENCE

Acute myeloid leukemia (AML) is a neoplasm characterized by infiltration of the blood, bone marrow, and other tissues by proliferative, clonal, poorly differentiated cells of the hematopoietic system. These leukemias comprise a spectrum of malignancies that, untreated, are uniformly fatal. In 2016, the estimated number of new AML cases in the United States was 19,950, comprising ~1.2% of all cancer cases. AML is the most common acute leukemia in older patients, with a median age at diagnosis of 67 years. Long-term survival is infrequent; U.S. registry data report that only 27% of patients survive 5 years.

ETIOLOGY

Most cases of AML are idiopathic. Genetic predisposition, radiation, chemical/other occupational exposures, and drugs have been implicated in the development of AML, but AML cases with established etiology are relatively rare. No direct evidence suggests a viral etiology. Genome sequencing studies suggest that most cases of AML arise from a limited number of mutations that accumulate with advancing age. Indeed, genome sequencing is providing paradigm-shifting advances in our understanding of leukemogenesis. The Cancer Genome Atlas (TCGA) and other databases demonstrate that blood cells from up to 5–6% of normal individuals aged >70 years contain potentially “premalignant” mutations that are associated with clonal expansion. The additional insults that subsequently direct “premalignant” blood cells to leukemia are quite heterogeneous and still poorly understood.

Genetic Predisposition

Myeloid neoplasms typically occur sporadically in adults; inherited predisposition is rare. Yet, it is clear that myeloid neoplasms with germline predisposition represent an important and growing subset of disease. Germline mutations associated with increased risk of developing a myeloid neoplasm include *CEBPA*, *DDX41*, *RUNX1*, *ANKRD26*, *ETV6*, and *GATA2* (**Table 100-1**). Likewise, myeloid neoplasms with germline predisposition are a feature of several well-described clinical syndromes, including bone marrow failure disorders (e.g., Fanconi anemia, Shwachman-Diamond syndrome, Diamond-Blackfan anemia), and telomere biology disorders (e.g., dyskeratosis congenita). As new mutations and associations are added to a rapidly growing list, it is increasingly clear that genetic predisposition plays a larger role than has been previously understood.

TABLE 100-1

WHO 2016 Classification of Myeloid Neoplasms with Germline Predisposition

Classification[a]
Myeloid neoplasms with germline predisposition without a preexisting disorder or organ dysfunction
Acute myeloid leukemia with germline *CEBPA* mutation
Myeloid neoplasms with germline *DDX41* mutation[b]
Myeloid neoplasms with germline predisposition and preexisting platelet disorders
Myeloid neoplasms with germline *RUNX1* mutation[b]
Myeloid neoplasms with germline *ANKRD26* mutation[b]
Myeloid neoplasms with germline *ETV6* mutation[b]
Myeloid neoplasms with germline predisposition and other organ dysfunction
Myeloid neoplasms with germline *GATA2* mutation
Myeloid neoplasms associated with bone marrow failure syndromes
Myeloid neoplasms associated with telomere biology disorders
Myeloid neoplasms associated with Noonan syndrome
Myeloid neoplasms associated with Down syndrome[b]

[a]Recognition of familial myeloid neoplasms requires that physicians take a thorough patient and family history to assess for typical signs and symptoms of known syndromes, including data on malignancies and previous bleeding episodes. Molecular genetic diagnostics is guided by a detailed patient and family history. Diagnostics should be performed in close collaboration with a genetic counselor; patients with a suspected heritable myeloid neoplasm, who test negative for known predisposition genes, should ideally be entered on a research study to facilitate new syndrome discovery. [b]Lymphoid neoplasms also reported.

Source: Adapted from Peterson L et al: Myeloid Neoplasms with Germline Predisposition, in *World Health Organization Classification of Tumours of Haematopoietic and Lymphoid Tissues*, update to 4th ed. IARC, 2017.

Several genetic syndromes with somatic cell chromosome aneuploidy, such as Down syndrome with trisomy 21, are associated with an increased incidence of AML. Down syndrome–associated AML in young children (<4 years) is typically of the acute megakaryocytic subtype and is associated with mutation in the *GATA1* gene. Such patients have excellent clinical outcomes but require dose modification of chemotherapy due to high treatment-related toxicities. Inherited diseases with defective DNA repair (e.g., Fanconi anemia, Bloom syndrome, and ataxia-telangiectasia) are also associated with AML. Each syndrome is associated with unique clinical features and atypical toxicities with chemotherapy, requiring expert care. Congenital neutropenia (Kostmann syndrome), due to mutations in the genes encoding the granulocyte colony-stimulating factor receptor and neutrophil elastase, is another disorder that may evolve into AML.

Chemical, Radiation, and Other Exposures

Anticancer drugs are the leading cause of therapy-associated AML. Alkylating agent–associated leukemias occur on average 4–6 years after exposure, and affected individuals often have multilineage dysplasia and monosomy/aberrations in chromosomes 5 and 7. Topoisomerase II inhibitor–associated leukemias occur 1–3 years after exposure, and affected individuals often have AML with monocytic features and aberrations involving chromosome 11q23. Exposure to ionizing radiation, benzene, chloramphenicol, phenylbutazone, and other drugs can uncommonly result in bone marrow failure that may evolve into AML.

CLASSIFICATION

The current categorization of AML uses the World Health Organization (WHO) classification **(Table 100-2)**, which defines biologically distinct groups based on cytogenetic and molecular abnormalities in addition to clinical features and light microscope morphology. Myeloid neoplasms with germline predisposition, as introduced above, are included as a new and important feature of this classification (Table 100-1). The WHO classification enables the identification of subsets of disease that may now (or in the future) be treated differently and advances the care of AML patients by enhancing recognition of the molecular basis of the disease from the time of diagnosis. Marrow (or blood) blast count of ≥20% is required to establish the diagnosis of AML, except for AML with the recurrent genetic abnormalities t(15;17), t(8;21), inv(16), or t(16;16).

TABLE 100-2

WHO 2016 Classification of Acute Myeloid Leukemia and Related Neoplasms

Acute myeloid leukemia (AML) with recurrent genetic abnormalities
AML with t(8;21)(q22;q22); RUNX1-RUNX1T1
AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11*
Acute promyelocytic leukemia with *PML-RARA*
AML with t(9;11)(p21.3;q23.3); *MLLT3-KMT2A*
AML with t(6;9)(p23;q34.1); *DEK-NUP214*
AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); *GATA2, MECOM*
AML (megakaryoblastic) with t(1;22)(p13.3;q13.3); *RBM15-MKL1*
Provisional entity: AML with BCR-ABL1
AML with mutated *NPM1*
AML with biallelic mutations of *CEBPA*
Provisional entity: AML with mutated RUNX1
AML with myelodysplasia-related changes
Therapy-related myeloid neoplasms
AML, not otherwise specified (NOS)
AML with minimal differentiation
AML without maturation
AML with maturation
Acute myelomonocytic leukemia
Acute monoblastic/monocytic leukemia
Pure erythroid leukemia
Acute megakaryoblastic leukemia
Acute basophilic leukemia
Acute panmyelosis with myelofibrosis
Myeloid sarcoma
Myeloid proliferations related to Down syndrome
Transient abnormal myelopoiesis (TAM)
Myeloid leukemia associated with Down syndrome

Note: Marrow blast count of ≥20% is required, except for AML with the recurrent genetic abnormalities t(15;17), t(8;21), inv(16), or t(16;16).

Source: Adapted from Arber DA et al: Acute myeloid leukaemia (AML) with recurrent genetic abnormalities, in *World Health Organization Classification of Tumours of Haematopoietic and Lymphoid Tissues*, update to 4th ed. IARC, 2017.

Clinical Features

Even with advances in molecular biology, recognizing clinical features remains important in understanding AML. For example, therapy-related AML is a distinct entity that develops following prior chemotherapy (e.g., alkylating agents, topoisomerase II inhibitors) or ionizing radiation. AML with myelodysplasia-related changes is recognized based in part on morphology but also on a medical history of an antecedent myelodysplastic syndrome (MDS) or myelodysplastic/myeloproliferative neoplasm. These clinical features contribute to AML prognosis and have therefore been included in the WHO classification.

Genetic Findings



Subtypes of AML are recognized based on the presence or absence of specific, recurrent cytogenetic, and/or genetic abnormalities. For example, the diagnosis of *acute promyelocytic leukemia* (APL) is based on the presence of either the t(15;17)(q22;q12) cytogenetic rearrangement or the *PML-RARA* fusion product of the translocation. Similarly, core binding factor (CBF) AML is designated based on the presence of t(8;21)(q22;q22), inv(16)(p13.1q22), or t(16;16)(p13.1;q22) or the respective fusion products *RUNX1-RUNX1T1* and *CBFB-MYH11*. Each of these groups identifies patients with favorable clinical outcomes when appropriately treated.

Several cytogenetic or genetic AML subtypes often associate with a specific morphologic appearance, such as a complex karyotype and AML with myelodysplasia-related changes. Patients with such changes typically fare poorly with standard treatments. However, only one cytogenetic abnormality is invariably associated with specific morphologic features: t(15;17)(q22;q12) with APL. Other cytogenetic and genetic findings may be commonly but not

invariably associated with a morphological description, highlighting the necessity of genetic and cytogenetic testing beyond simple morphology to most accurately diagnose AML. Several chromosomal abnormalities often associate primarily with one morphologic/immunophenotypic group. Examples include inv(16)(p13.1q22) with AML with abnormal bone marrow eosinophils; t(8;21)(q22;q22) with slender Auer rods, expression of CD19, and increased normal eosinophils; and t(9;11)(p22;q23), and other translocations involving 11q23, with monocytic features. Recurring chromosomal abnormalities in AML may also be loosely associated with specific clinical characteristics. More commonly associated with younger age are t(8;21) and t(15;17), and with older age, del(5q) and del(7q). Myeloid sarcomas are associated with t(8;21); disseminated intravascular coagulation (DIC) is associated with t(15;17). 11q23 aberrations and monocytic leukemia are associated with extramedullary sites of involvement at presentation, especially gingival hypertrophy.

The WHO classification also incorporates molecular abnormalities by recognizing fusion genes or specific genetic mutations with a role in leukemogenesis. As a classic example, t(15;17) results in the fusion gene *PML-RARA* that encodes a chimeric protein, promyelocytic leukemia (Pml)–retinoic acid receptor α (Rarα), which is formed by the fusion of the retinoic acid receptor α (*RARA*) gene from chromosome 17 and the promyelocytic leukemia (*PML*) gene from chromosome 15. The *RARA* gene encodes a member of the nuclear hormone receptor family of transcription factors. *PML* is important in many cellular processes, including cell growth control, apoptosis, and senescence; its effects are mediated at least in part by nuclear bodies that store a myriad of proteins/enzymes that are involved in these functions. The PML-RARA fusion protein suppresses gene transcription and blocks differentiation beyond the promyelocyte stage. Pharmacologic concentrations of the Rarα ligand, achieved with the drug all-*trans*-retinoic acid (tretinoin, ATRA), relieve the block and promote hematopoietic cell differentiation. However, the effects of ATRA are not primarily from direct restoration of gene transactivation via RA signaling. Rather, drug treatment induces degradation of the fusion protein. Mechanistic work has demonstrated that the RARA fusion partner PML is far more important in the pathobiology than was initially understood. PML-RARA disturbs nuclear body assembly. This impairs many PML functions, culminating in enhanced self-renewal of leukemic cells. ATRA and arsenic trioxide (ATO) both induce PML-RARA degradation (by different mechanisms), leading to reformation of PML nuclear bodies (or enhanced nuclear body activity). Restored PML functions include the activation of p53 which triggers senescence in leukemic cells. Clinical therapy with ATRA and ATO has revolutionized the care of APL patients (see "Treatment of Acute Promyelocytic Leukemia" section).

Similar examples of molecular subtypes included in the category of AML with recurrent genetic abnormalities are those characterized by the leukemogenic fusion genes *RUNX1-RUNX1T1*, *CBFB-MYH11*, *MLLT3-KMT2A*, and *DEK-NUP214*, resulting, respectively, from t(8;21), inv(16) or t(16;16), t(9;11), and t(6;9)(p23;q34).

The WHO classification of AML continues to expand as knowledge of specific genetic or cytogenetic aberrations grows. Several AML subtypes are defined by the presence of genetic mutations rather than chromosomal aberrations including, for example, *AML with mutated nucleophosmin (nucleolar phosphoprotein B23, numatrin)* (*NPM1*) and *AML with biallelic mutated CEBPA*, respectively. Both entities are associated with more favorable clinical outcomes, though the *NPM1* prognostic impact is affected by coexisting mutation in fms-related tyrosine kinase 3 (*FLT3*). Activating mutations of *FLT3* are present in ~30% of adult AML patients, primarily due to internal tandem duplications (ITD) in the juxtamembrane domain that have negative prognostic impact. In contrast, point mutations of the activating loop of the kinase (called tyrosine kinase domain [TKD] mutations) have uncertain prognostic impact. Aberrant activation of the *FLT3*-encoded protein provides increased proliferation and antiapoptotic signals to the myeloid progenitor cell. *FLT3*-ITD, the more common of the *FLT3* mutations, occurs preferentially in patients with cytogenetically normal AML (CN-AML). The importance of identifying *FLT3*-ITD at diagnosis relates to the fact that is it useful not only as a prognosticator but also may predict response to specific treatment such as a tyrosine kinase inhibitor (TKI); several TKI are currently in clinical investigation (e.g., midostaurin, quizartinib, gilteritinib, crenolanib, sorafenib). The *FLT3* allelic ratio (of the number of mutated alleles to wild type alleles) provides information beyond the mere presence or absence of the mutation. The ratio is affected by several mutational scenarios such as one mutated gene and one wild type gene, or one mutated gene with no (deleted) wild type gene, and the ratio of malignant to nonmalignant cells in the sample. The allelic ratio affects the prognostic impact of the *FLT3*-ITD mutation; patients with *FLT3*-ITD "low" allelic ratio (<0.5) fare better. Accordingly, mutated *NPM1* without *FLT3*-ITD or with *FLT3*-ITD^{low} are both viewed as favorable-risk by the European LeukemiaNet (ELN) risk stratification schema **(Table 100-3)**. Conversely, *FLT3*-ITD^{high} is known to have an adverse prognostic impact; patients with mutated *NPM1* and *FLT3*-ITD with an allelic ratio >0.5 are thus intermediate-risk by ELN stratification. Involving a different tyrosine kinase, AML with *BCR-ABL1* fusion is a new WHO provisional entity, to recognize rare cases that may benefit from *BCR-ABL* TKI therapy (Table 100-2).

TABLE 100-3

2017 European LeukemiaNet Risk Stratification by Genetics for Acute Myeloid Leukemia[a]

Risk Category[b]	Genetic Abnormality
Favorable	t(8;21)(q22;q22); *RUNX1-RUNX1T1* inv(16)(p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11* Mutated *NPM1* without *FLT3*-ITD or with *FLT3*-ITD$^{low(c)}$ Biallelic mutated *CEBPA*
Intermediate	Mutated *NPM1* and *FLT3*-ITD$^{high(c)}$ Wild type *NPM1* without *FLT3*-ITD or with *FLT3*-ITD$^{low(c)}$ (w/o adverse-risk genetic lesions) t(9;11)(p21.3;q23.3); *MLLT3-KMT2A*[d] Cytogenetic abnormalities not classified as favorable or adverse
Adverse	t(6;9)(p23;q34.1); *DEK-NUP214* t(v;11q23.3); *KMT2A* rearranged t(9;22)(q34.1;q11.2); *BCR-ABL1* inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); *GATA2, MECOM(EVI1)* –5 or del(5q); –7; –17/abn(17p) Complex karyotype,[e] monosomal karyotype[f] Wild type *NPM1* and *FLT3*-ITD$^{high(c)}$ Mutated *RUNX1*[g] Mutated *ASXL1*[g] Mutated *TP53*[h]

[a]This table excludes acute promyelocytic leukemia. Frequencies, response rates, and outcome measures should be reported by risk category, and, if sufficient numbers are available, by specific genetic lesions indicated. [b]Prognostic impact of a marker is treatment-dependent and may change with new therapies. [c]Low, low allelic ratio (<0.5); high, high allelic ratio (≥0.5); semiquantitative assessment of *FLT3*-ITD allelic ratio (using DNA fragment analysis) is determined as ratio of the area under the curve (AUC) "*FLT3*-ITD" divided by AUC "*FLT3*-wild type"; recent studies indicate that acute myeloid leukemia with *NPM1* mutation and *FLT3*-ITD low allelic ratio may also have a more favorable prognosis and patients should not routinely be assigned to allogeneic hematopoietic-cell transplantation. [d]The presence of t(9;11)(p21.3;q23.3) takes precedence over rare, concurrent adverse-risk gene mutations. [e]Three or more unrelated chromosome abnormalities in the absence of one of the World Health Organization-designated recurring translocations or inversions, i.e., t(8;21), inv(16) or t(16;16), t(9;11), t(v;11)(v;q23.3), t(6;9), inv(3) or t(3;3); AML with *BCR-ABL1*. [f]Defined by the presence of one single monosomy (excluding loss of X or Y) in association with at least one additional monosomy or structural chromosome abnormality (excluding core-binding factor AML). [g]These markers should not be used as an adverse prognostic marker if they co-occur with favorable-risk AML subtypes. [h]*TP53* mutations are significantly associated with AML with complex and monosomal karyotype.

Source: Adapted from Döhner H et al: Diagnosis and management of acute myeloid leukemia in adults: 2017 recommendations from an international expert panel, on behalf of the European LeukemiaNet. Blood 129:424, 2017.

Immunophenotypic Findings

The immunophenotype of human leukemia cells can be studied by multiparameter flow cytometry after the cells are labeled with monoclonal antibodies to cell-surface antigens. This can be important in quickly distinguishing AML from acute lymphoblastic leukemia and for identifying some subtypes of AML. For example, AML with minimal differentiation, characterized by immature morphology and no lineage-specific cytochemical reactions, may be diagnosed by flow-cytometric demonstration of the myeloid-specific antigens cluster designation (CD) 13 and/or 117. Similarly, acute megakaryoblastic leukemia can often be diagnosed only by expression of the platelet-specific antigens CD41 and/or CD61. Although flow cytometry is widely used, and in some cases essential for the diagnosis of AML, it has only a supportive role in establishing the different subtypes of AML through the WHO classification. Increasingly, multiparameter flow cytometry is used for the measurement of minimal residual disease (MRD) after remission is achieved.

PROGNOSTIC FACTORS

Several factors predict outcome of AML patients treated with chemotherapy; they should be used for risk stratification and treatment guidance.

Chromosome findings at diagnosis currently provide the most important independent prognostic information. Several reports have categorized patients as having favorable, intermediate, or adverse cytogenetic risk based on the presence of structural and/or numerical aberrations. Patients with t(15;17) have a very good prognosis (~85% cured), and those with t(8;21) and inv(16) have a good prognosis (~55% cured), whereas those with no cytogenetic abnormality have an intermediate outcome risk (~40% cured). Patients with a complex karyotype, t(6;9), inv(3), or –7 have a very poor prognosis. Another cytogenetic subgroup, the monosomal karyotype, has been suggested to adversely impact the outcome of AML patients other than those with t(15;17), t(8;21), or inv(16) or t(16;16). The monosomal karyotype subgroup is defined by the presence of at least two autosomal monosomies (loss of chromosomes other than Y or X) or a single autosomal monosomy with additional structural abnormalities.

For patients lacking prognostic cytogenetic abnormalities, such as those with CN-AML, testing for several mutated genes can help to risk-stratify. In addition to the *NPM1* mutation and/or *FLT3*-ITD as described above, biallelic *CEBPA* mutations have prognostic value. Such mutations predict favorable outcome. Given the proven prognostic importance of *NPM1*, *CEBPA*, and *FLT3*, molecular assessment of these genes at diagnosis has been incorporated into AML management guidelines by the National Comprehensive Cancer Network (NCCN) and the ELN. The same markers help to define genetic groups in the ELN standardized reporting system, which is based on both cytogenetic and molecular abnormalities and is used for comparing clinical features/treatment response among subsets of patients reported across different clinical studies (Table 100-3). These genetic groups should be used for risk stratification and treatment guidance.

In addition to *NPM1* and *CEBPA* mutations and *FLT3*-ITD, molecular aberrations in other genes may be routinely used for prognostication in the future (**Table 100-4**). Among these mutated genes are those encoding receptor tyrosine kinases (e.g., v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog [*KIT*]), transcription factors (i.e., *RUNX1* and Wilms tumor 1 [*WT1*]), and epigenetic modifiers (i.e., additional sex combs like transcriptional regulator 1 [*ASXL1*], DNA (cytosine-5-)-methyltransferase 3 alpha [*DNMT3A*], isocitrate dehydrogenase 1 [NADP+], soluble [*IDH1*], isocitrate dehydrogenase 2 (NADP+), mitochondrial [*IDH2*], lysine (K)–specific methyltransferase 2A [*KMT2A*, also known as *MLL*], and tet methylcytosine dioxygenase 2 [*TET2*]). Although *KIT* mutations are almost exclusively present in CBF AML and impact adversely the outcome, the remaining markers have been reported primarily in CN-AML. These gene mutations have been shown to be associated with outcome in multivariable analyses independent of other prognostic factors. However, for some of them, data remain unclear on the prognostic impact due to conflicting reports (e.g., *TET2*, *IDH1, IDH2*). Increasingly, novel drugs that inhibit/modulate aberrant pathways activated by some of these genes (e.g., *IDH1, IDH2, KMT2A*, among others) are being incorporated into clinical trials to treat AML.

TABLE 100-4

Molecular Prognostic Markers in AML[a]

Gene Symbol	Gene Location	Prognostic Impact
Genes Included in the WHO Classification and ELN Reporting System		
NPM1 mutations	5q35.1	Favorable
CEBPA mutations	19q13.1	Favorable
FLT3-ITD	13q12	Depends on allelic ratio and *NPM1* mutational status
Genes Encoding Receptor Tyrosine Kinases		
KIT mutation	4q12	Adverse
FLT3-TKD	13q12	Unclear
Genes Encoding Transcription Factors		
RUNX1 mutations	21q22.12	Adverse
WT1 mutations	11p13	Adverse
Genes Encoding Epigenetic Modifiers		
ASXL1 mutations	20q11.21	Adverse
DNMT3A mutations	2p23.3	Adverse
IDH mutations (*IDH1* and *IDH2*)	2q34 & 15q26.1	Adverse
KMT2A-PTD	11q23	Adverse
TET2 mutations	4q24	Adverse
Deregulated Genes		
BAALC overexpression	8q22.3	Adverse
ERG overexpression	21q22.3	Adverse
MN1 overexpression	22q12.1	Adverse
EVI1 overexpression	3q26.2	Adverse
Deregulated MicroRNAs		
miR-155 overexpression	21q21.3	Adverse
miR-3151 overexpression	8q22.3	Adverse
miR-181a overexpression	1q32.1 and 9q33.3	Favorable

[a]This table excludes acute promyelocytic leukemia.

Abbreviations: AML, acute myeloid leukemia; ELN, European LeukemiaNet; ITD, internal tandem duplication; PTD, partial tandem duplication; TKD, tyrosine kinase domain; WHO, World Health Organization.

In addition to gene mutations, deregulation of the expression levels of coding genes and of short noncoding RNAs (microRNAs) also provide prognostic information (Table 100-4). Overexpression of genes such as brain and acute leukemia, cytoplasmic (*BAALC*), v-ets avian erythroblastosis virus E26 oncogene homologue (avian) (*ERG*), meningioma (disrupted in balanced translocation) 1 (*MN1*), and MDS1 and EVI1 complex locus (*MECOM*, also known as *EVI1*) predict poor outcome, especially in CN-AML. Similarly, deregulated expression levels of microRNAs, naturally occurring noncoding RNAs that regulate the expression of proteins via degradation or translational inhibition of their target coding RNAs, have also been associated with prognosis in AML. Overexpression of *miR-155* and *miR-3151* predicts unfavorable outcome in CN-AML, whereas overexpression of *miR-181a* predicts favorable outcome both in CN-AML and cytogenetically abnormal AML.

Because prognostic molecular markers in AML are not mutually exclusive and often occur concurrently (>80% patients have at least two or more prognostic gene mutations), the likelihood that distinct marker combinations may be more informative than single markers is increasingly clear.

Epigenetic changes (e.g., DNA methylation and/or post-translational histone modification) and microRNAs are often involved in deregulation of genes involved in hematopoiesis, contribute to leukemogenesis and may associate with the previously discussed prognostic gene mutations. These changes have been shown to provide biologic insights into leukemogenic mechanisms and also independent prognostic information. Indeed, it is anticipated that with the enormous progress made in DNA and RNA sequencing technology, additional genetic and epigenetic aberrations will soon be discovered, further improving classification and risk-stratification in AML patients.

In addition to cytogenetics and molecular aberrations, several other factors are associated with outcome in AML. Age at diagnosis is one of the most important risk factors. Advancing age is associated with a poor prognosis for two reasons: (1) its influence on the ability to survive induction therapy due to coexisting medical comorbidities, and (2) with each successive decade of age, a greater proportion of patients have intrinsically more resistant disease. A prolonged symptomatic interval with cytopenias preceding AML diagnosis, or a history of antecedent hematologic disorders including MDS or myeloproliferative neoplasms, is often found in older patients. Cytopenia is a clinical feature associated with a lower complete remission (CR) rate and shorter survival time. The CR rate is lower in patients who have had anemia, leukopenia, and/or thrombocytopenia for >3 months before the diagnosis of AML, when compared to those without such a history. Responsiveness to chemotherapy declines as the duration of the antecedent disorder increases. Likewise, AML developing after treatment with cytotoxic agents for other malignancies is usually difficult to treat successfully. In addition, older patients less frequently harbor favorable cytogenetic abnormalities (i.e., t[8;21], inv[16], and t[16;16]) and more frequently harbor adverse cytogenetic (e.g., complex and monosomal karyotypes) and/or molecular (e.g., *ASXL1*, *p53*) abnormalities.

Other factors independently associated with worse outcome are a poor performance status that influences ability to survive induction therapy and a high presenting leukocyte count that in some series is an adverse prognostic factor for attaining a CR. Among patients with hyperleukocytosis (>100,000/μL), early central nervous system bleeding and pulmonary leukostasis contribute to poor outcomes.

Following administration of therapy, achievement of CR is associated with better outcome and longer survival. CR is defined after examination of both blood and bone marrow and essentially represents both eradication of detectable leukemia and restoration of normal hematopoiesis. The blood neutrophil count must be ≥1000/μL and the platelet count ≥100,000/μL. Hemoglobin concentration is not considered in determining CR. Circulating blasts should be absent. Although rare blasts may be detected in the blood during marrow regeneration, they should disappear on successive studies. At CR, the bone marrow should contain <5% blasts, and Auer rods should be absent. Extramedullary leukemia should not be present.

CLINICAL PRESENTATION

Symptoms

Patients with AML usually present with nonspecific symptoms that begin gradually, or abruptly, and are the consequence of anemia, leukocytosis, leukopenia/leukocyte dysfunction, or thrombocytopenia. Nearly half have symptoms for ≤3 months before the leukemia is diagnosed.

Fatigue is a frequent first symptom among AML patients. Anorexia and weight loss are common. Fever with or without an identifiable infection is the initial symptom in ~10% of patients. Signs of abnormal hemostasis (bleeding, easy bruising) are common. Bone pain, lymphadenopathy, nonspecific cough, headache, or diaphoresis may also occur.

Rarely, patients may present with symptoms from a myeloid sarcoma (a tumor mass consisting of myeloid blasts occurring at anatomic sites other than bone marrow). Sites involved are most commonly the skin, lymph node, gastrointestinal tract, soft tissue, and testis. This rare presentation, often characterized by chromosome aberrations (e.g., monosomy 7, trisomy 8, 11q23 rearrangement, inv[16], trisomy 4, t[8;21]), may precede or coincide with blood and/or marrow involvement by AML. Patients who present with isolated myeloid sarcoma typically develop blood and/or marrow involvement quickly thereafter and cannot be cured with local therapy (radiation or surgery) alone.

Physical Findings

Fever, infection, and hemorrhage are often found at the time of diagnosis; splenomegaly, hepatomegaly, lymphadenopathy, and “bone pain” may also be present less commonly. Hemorrhagic complications are most commonly and, classically, found in APL. APL patients often present with DIC-associated minor hemorrhage but may have significant gastrointestinal bleeding, intrapulmonary hemorrhage, or intracranial hemorrhage. Likewise, thrombosis is another less frequent but well recognized clinical feature of DIC in APL. Bleeding associated with coagulopathy may also occur in monocytic AML and with

extreme degrees of leukocytosis or thrombocytopenia in other morphologic subtypes. Retinal hemorrhages are detected in 15% of patients. Infiltration of the gingiva, skin, soft tissues, or meninges with leukemic blasts at diagnosis is characteristic of the monocytic subtypes and those with 11q23 chromosomal abnormalities.

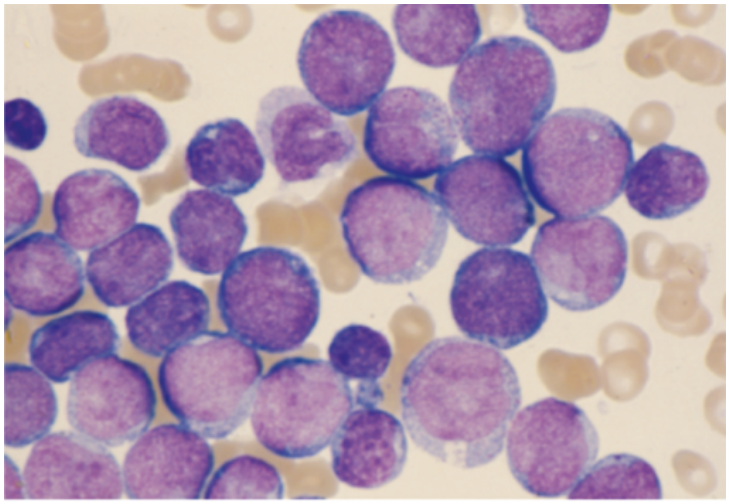
Hematologic Findings

Anemia is usually present at diagnosis though it is not typically severe. The anemia is usually normocytic normochromic. Decreased erythropoiesis in the setting of AML often results in a reduced reticulocyte count, and red blood cell (RBC) survival is decreased by accelerated destruction. Active blood loss may rarely contribute to the anemia.

The median presenting leukocyte count is ~15,000/μL. Lower presenting leukocyte counts are more typical of older patients and those with antecedent hematologic disorders. Between 25 and 40% of patients have counts <5000/μL, and 20% have counts >100,000/μL. Fewer than 5% have no detectable leukemic cells in the blood. In AML, the cytoplasm often contains primary (nonspecific) granules, and the nucleus shows fine, lacy chromatin with one or more nucleoli characteristic of immature cells. Abnormal rod-shaped granules called Auer rods are not uniformly present, but when they are, AML is virtually certain (Fig. 100-1).

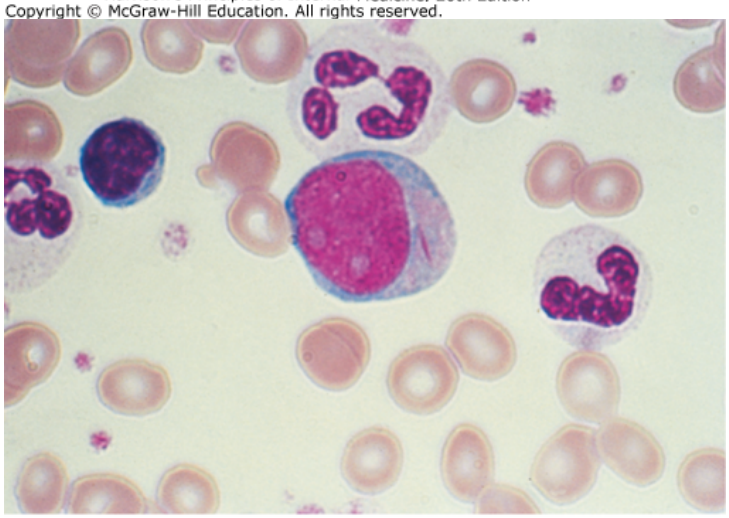
FIGURE 100-1

Morphology of acute myeloid leukemia (AML) cells. ***A.*** Uniform population of primitive myeloblasts with immature chromatin, nucleoli in some cells, and primary cytoplasmic granules. ***B.*** Leukemic myeloblast containing an Auer rod. ***C.*** Promyelocytic leukemia cells with prominent cytoplasmic primary granules. ***D.*** Peroxidase stain shows dark blue color characteristic of peroxidase in granules in AML.



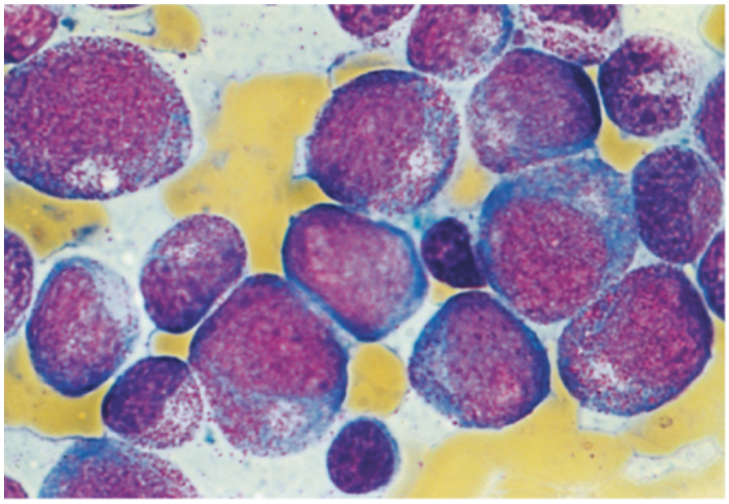
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Platelet counts <100,000/μL are found at diagnosis in ~75% of patients, and ~25% have counts <25,000/μL. Both morphologic and functional platelet abnormalities can be observed, including large and bizarre shapes with abnormal granulation and inability of platelets to aggregate or adhere normally to one another.

Pretreatment Evaluation

Once the diagnosis of AML is suspected, thorough evaluation and initiation of appropriate therapy should follow. In addition to clarifying the subtype of leukemia, initial studies should evaluate the overall functional integrity of the major organ systems, including the cardiovascular, pulmonary, hepatic, and renal systems (Table 100-5). Factors that have prognostic significance, either for achieving CR or for predicting CR duration, should also be assessed before initiating treatment, including cytogenetics and molecular markers. Leukemic cells should be obtained from all patients and cryopreserved for future investigational testing as well as potential future use as new diagnostics and therapeutics become available. All patients should be evaluated for infection.

TABLE 100-5

Initial Diagnostic Evaluation and Management of Adult Patients with AML

History
Increasing fatigue or decreased exercise tolerance (anemia)
Excess bleeding or bleeding from unusual sites (DIC, thrombocytopenia)
Fevers or recurrent infections (neutropenia)
Headache, vision changes, nonfocal neurologic abnormalities (CNS leukemia or bleed)
Early satiety (splenomegaly)
Family history of AML (Fanconi, Bloom, or Kostmann syndromes or ataxia-telangiectasia)
History of cancer (exposure to alkylating agents, radiation, topoisomerase II inhibitors)
Occupational exposures (radiation, benzene, petroleum products, paint, smoking, pesticides)
Physical Examination
Performance status (prognostic factor)
Ecchymosis and oozing from IV sites (DIC, possible acute promyelocytic leukemia)
Fever and tachycardia (signs of infection)
Papilledema, retinal infiltrates, cranial nerve abnormalities (CNS leukemia)
Poor dentition, dental abscesses
Gum hypertrophy (leukemic infiltration, most common in monocytic leukemia)
Skin infiltration or nodules (leukemia infiltration, most common in monocytic leukemia)
Lymphadenopathy, splenomegaly, hepatomegaly
Back pain, lower extremity weakness (spinal granulocytic sarcoma, most likely in t[8;21] patients)
Laboratory and Radiologic Studies
CBC with manual differential cell count
Chemistry tests (electrolytes, creatinine, BUN, calcium, phosphorus, uric acid, hepatic enzymes, bilirubin, LDH, amylase, lipase)
Clotting studies (prothrombin time, partial thromboplastin time, fibrinogen, D-dimer)
Viral serologies (CMV, HSV-1, varicella-zoster)
RBC type and screen
HLA typing for potential allogeneic HCT
Bone marrow aspirate and biopsy (morphology, cytogenetics, flow cytometry, molecular studies for NPM1 and CEBPA mutations and FLT3-ITD)
Cryopreservation of viable leukemia cells
Myocardial function (echocardiogram or MUGA scan)

PA and lateral chest radiograph
Placement of central venous access device
Interventions for Specific Patients
Dental evaluation (for those with poor dentition)
Lumbar puncture (for those with symptoms of CNS involvement)
Screening spine MRI (for patients with back pain, lower extremity weakness, paresthesias)
Social work referral for patient and family psychosocial support
Counseling for All Patients
Provide patients with information regarding their disease and genetic risks, sperm banking or menstrual suppression, financial counseling, and support group contact

Abbreviations: AML, acute myeloid leukemia; BUN, blood urea nitrogen; CBC, complete blood count; CMV, cytomegalovirus; CNS, central nervous system; DIC, disseminated intravascular coagulation; HLA, human leukocyte antigen; HCT, hematopoietic stem cell transplantation; HSV, herpes simplex virus; IV, intravenous; LDH, lactate dehydrogenase; MRI, magnetic resonance imaging; MUGA, multigated acquisition; PA, posteroanterior; RBC, red blood (cell) count.

Most patients are anemic and thrombocytopenic at presentation. Replacement of the appropriate blood components, if necessary, should begin promptly. Because qualitative platelet dysfunction or the presence of an infection may increase the likelihood of bleeding, evidence of hemorrhage justifies the immediate use of platelet transfusion, even if the platelet count is only moderately decreased.

About 50% of patients have a mild to moderate elevation of serum uric acid at presentation. Only 10% have marked elevations, but renal precipitation of uric acid and the nephropathy that may result is a serious but uncommon complication. The initiation of chemotherapy may aggravate hyperuricemia, and patients are usually started immediately on allopurinol and hydration at diagnosis. Rasburicase (recombinant uric oxidase) is also useful for treating uric acid nephropathy and often can normalize the serum uric acid level within hours with a single dose of treatment, although its expense suggests that limiting its use to patients with severe hyperuricemia and/or kidney injury may be prudent. The presence of high concentrations of lysozyme, a marker for monocytic differentiation, may be etiologic in renal tubular dysfunction for a minority of patients.

Treatment

TREATMENT

Acute Myeloid Leukemia

Treatment of the newly diagnosed patient with AML is usually divided into two phases, induction and postremission management (consolidation) **(Fig. 100-2)**. The initial goal is to induce CR. Once CR is obtained, further therapy must be given to prolong survival and achieve cure. The initial induction treatment and subsequent postremission therapy are chosen based on the patient's age, overall fitness, and cytogenetic/molecular risk. Intensive therapy with cytarabine and anthracyclines in younger patients (<60 years) increases the cure rate of AML. In older patients, the benefit of intensive therapy is controversial in all but favorable-risk patients; novel approaches for selecting patients predicted to be responsive to treatment and new therapies are being pursued.

INDUCTION CHEMOTHERAPY

The most commonly used induction regimens (for patients other than those with APL) consist of combination chemotherapy with cytarabine and an anthracycline (e.g., daunorubicin, idarubicin). Cytarabine is a cell cycle S-phase–specific antimetabolite that becomes phosphorylated intracellularly to an active triphosphate form that interferes with DNA synthesis. Anthracyclines are DNA intercalators. Their primary mode of action is thought to be inhibition of topoisomerase II, leading to DNA breaks.

In adults, cytarabine used at standard dose (100–200 mg/m^2) is administered as a continuous intravenous infusion for 7 days. With cytarabine, anthracycline therapy generally consists of daunorubicin (60–90 mg/m^2) or idarubicin (12 mg/m^2) intravenously on days 1, 2, and 3 (the 7 and 3 regimen). Other agents can be added (e.g., cladribine) when 60 mg/m^2 of daunorubicin is used. With the 7 and 3 regimen, it is now clearly established that 45 mg/m^2 dosing of daunorubicin results in inferior outcomes; patients should receive higher doses as described. Patients failing remission after one induction are offered reinduction with the same (or slightly modified) therapy.

In older patients (age ≥60–65 years), the outcome is generally poor due to a higher frequency of resistant disease and increased rate of treatment-related mortality. This is especially true in patients with prior hematologic disorders (MDS or myeloproliferative neoplasms), therapy-related AML, or cytogenetic and genetic abnormalities that adversely impact on clinical outcome. All older patients should be considered for clinical trials, but in particular older patients in the adverse-risk groups delineated above should be offered investigational approaches when possible. Conventional therapy for older patients is similar to that for younger: the 7 and 3 regimen with standard-dose cytarabine and idarubicin (12 mg/m^2), or daunorubicin (60 mg/m^2, or 90 mg/m^2 for those <65 years). For patients aged >65 years, high-dose daunorubicin (90 mg/m^2) has increased toxicity and is not recommended. Older patients and those with adverse-risk genetics may receive lower intensity therapy with a hypomethylating agent (decitabine or azacitidine), clofarabine, or preferably investigational therapy **(Table 100-6)**.

With the 7 and 3 regimen, 60–80% of younger and 33–60% of older patients (among those who are candidates for intensive therapy) with primary AML achieve CR. Of patients who do not achieve CR, most have drug-resistant leukemia, although induction death is more frequent with advancing age and medical comorbidity. Patients with refractory disease after induction should be considered for salvage treatments, preferentially on clinical trials, before receiving allogeneic hematopoietic stem cell transplantation (HCT) that is usually reserved for patients in or near CR. However, fit younger patients with primary refractory disease have ~15–20% cure rates with allogeneic HCT (after myeloablative conditioning); for this reason early consideration of future allogeneic HCT feasibility (including HLA typing, donor search, etc.) should be part of the initial induction approach for most AML patients.

POSTREMISSION THERAPY

Induction of a durable first CR (CR1) is critical to long-term survival in AML. However, without further therapy virtually all patients relapse. Thus, postremission therapy is designed to eradicate residual (typically undetectable) leukemic cells to prevent relapse and prolong survival. As for induction, the type of postremission therapy in AML is selected for each individual patient based on age, fitness, and cytogenetic/molecular risk.

The choice between consolidation with chemotherapy or transplantation is complex and based on age, risk, and practical considerations. In younger patients receiving chemotherapy, postremission therapy with intermediate/high-dose cytarabine for two to four cycles is standard practice. Higher doses of cytarabine during post remission therapy appear more effective than standard doses (as are used in induction), at least for those who do not have adverse-risk genetics. Recent studies suggest that the long-standing practice of high-dose cytarabine (3 g/m^2, every 12 h on days 1, 3, and 5) may not improve survival over intermediate-dose cytarabine (IDAC, 1-1.5 g/m^2) for such patients. Thus, the ELN has recommended IDAC at 1–1.5 g/m^2, every 12 h, on days 1–3, as the optimal postremission chemotherapy approach for favorable and intermediate-risk younger patients, for two to four cycles. While high-dose cytarabine may not be necessary, it is important to note that younger favorable-risk patients have worse outcomes when doses below 1g/m^2 are used. In contrast to favorable-risk, intermediate- and adverse-risk patients should be considered for allogeneic HCT CR1 when feasible (see transplant discussion below). As older patients have increased toxicities with higher doses of cytarabine, ELN recommends relatively attenuated cytarabine doses (0.5–1g/m^2, every 12 h, on days 1–3) in favorable-risk older patients. There is no clear value for intensive postremission therapy in non–favorable-risk older patients; allogeneic HCT in CR1 (up to age 75 years) or investigational therapy is recommended. Indeed, postremission therapy is an appropriate setting for introduction of new agents in both older and younger patients (Table 100-6).

Allogeneic HCT is the best relapse-prevention strategy currently available for AML. Allogeneic HCT is probably best understood as an opportunity for immunotherapy; residual leukemia cells potentially elicit an immunologic response from donor immune cells, the so-called graft-versus-leukemia (GVL) effect. The benefit of GVL in relapse risk reduction, unfortunately, is offset somewhat by increased morbidity and mortality from complications of allogeneic HCT including graft-versus-host disease (GVHD). Given that relapsed AML is typically resistant to chemotherapy, allogeneic HCT in CR1 is a favored strategy. It is recommended in patients age <75 years who do not have favorable-risk disease and who have a human leukocyte antigen (HLA)–matched donor (related or unrelated). We recommend allogeneic HCT in CR1 for patients with intermediate-risk disease (Table 100-3). However, considerable debate exists regarding whether allogeneic HCT in CR1 is a requirement for younger patients with intermediate-risk AML, as one large series from the Medical Research Council reported that such patients have similar outcomes if transplanted only after relapse (and achievement of CR2), sparing some the long-term morbidity of transplantation. That said, allogeneic HCT is generally recommended as soon as possible after CR1 is achieved unless the patient is in a favorable-risk group. Selected adverse-risk patients without HLA-matched donors are considered for alternative donor transplants (e.g., HLA-mismatched unrelated, haploidentical related, and umbilical cord blood) even in CR1. Notably, more effective methods of in vivo T cell depletion (i.e., posttransplant cyclophosphamide following haploidentical transplantation) have broadened the availability of potential allogeneic HCT donors. Now, virtually any patient with a healthy parent or child (i.e., haploidentical) has an available donor suitable for allogeneic HCT if desired. Long-term outcomes with conventional chemotherapy for older patients are dismal; transplantation for such patients is expanding. For older patients, nonrandomized data demonstrate benefit for older patients in CR1 treated with reduced-intensity conditioning regimens and allogeneic HCT. Prospective data suggest that 40% of older patients in CR1 who are candidates for allogeneic HCT may be cured.

Trials comparing allogeneic HCT with intensive chemotherapy or autologous HCT have shown improved duration of remission with allogeneic HCT. However, the relapse risk reduction that is observed with allogeneic HCT is partially offset by the increase in fatal treatment-related toxicity (GVHD, organ toxicity). Despite this, there is no debate that patients with adverse-risk AML have improved long-term survival with early allogeneic HCT. Alternatively, high-dose chemotherapy with autologous HCT rescue is another postremission approach in non-adverse risk subsets. Autologous HCT patients receive their own stem cells (collected during remission and cryopreserved), following administration of myeloablative chemotherapy. The toxicity is relatively low with autologous HCT (5% mortality rate), but the relapse rate is higher than with allogeneic HCT, due to the absence of the GVL effect. Favorable and intermediate-risk patients may benefit from autologous HCT more so than adverse-risk patients. Practically speaking, however, autologous HCT in AML

patients is less frequently employed currently due to enhanced relapse risk reduction seen with allogeneic HCT and the growing use of HLA mismatched donors (in novel transplantation approaches).

Prognostic factors help to select the appropriate postremission therapy in patients in CR1. Our approach includes allogeneic HCT in first CR for patients without favorable cytogenetics or genotype (e.g., patients who do not have CEBPA biallelic mutations or NPM1 mutations without FLT3-ITD/with *FLT3*-ITD^{low}). Patients with adverse-risk disease should proceed to allogeneic HCT at CR1 if possible. The decision for allogeneic HCT for younger intermediate-risk patients is complex and individualized as described above; we recommend it when an HLA-matched donor is available. Subsets of patients may benefit from targeted therapy given during remission; emerging data demonstrate survival benefit from incorporation of the FLT3 inhibitor midostaurin, for example, into induction and postremission therapies for patients with *FLT3* mutated AML. On April 28, 2017, the U.S. Food and Drug Administration (FDA) approved midostaurin (RYDAPT, Novartis Pharmaceuticals Corp.) for the treatment of adult patients with newly diagnosed AML who are *FLT3* mutation-positive (either ITD or TKD+), in combination with standard cytarabine and daunorubicin induction and cytarabine consolidation. Allogeneic transplantation in CR1 is still recommended for these patients.

For patients in morphologic CR, measurement of MRD remains a very important and challenging research area. Cytogenetics are a mainstay of disease assessment, and persistence of abnormal karyotype (in spite of morphologic CR) is clearly associated with poor clinical outcomes. Immunophenotyping to detect minute populations of blasts or sensitive molecular assays (e.g., reverse transcriptase polymerase chain reaction [RT-PCR]) to detect AML-associated molecular abnormalities (e.g., NPM1 mutation, the CBF AML RUNX1/RUNX1T1 and CBFB/MYH11 transcripts, the APL PML/RARA transcript) can be performed to assess whether MRD is present at sequential time points during or after treatment. Whether emerging next-generation sequencing or serial quantitative assessment using flow or RT-PCR, performed during remission, can effectively direct successful subsequent therapy and improve clinical outcome remains to be determined. Currently, no consensus exists for the optimal MRD measurement technique, or its application. Data suggest that MRD measurement can in some settings be a reliable discriminator between patients who will continue in CR or relapse, but whether subsequent therapy (i.e., allogeneic HCT or additional chemotherapy) can effectively eradicate disease in such patients is not yet clear. In the subset of patients with APL, serial RT-PCR (for the *PML/RARA* transcript) is a very useful and reliable tool to detect early relapse and direct initiation of reinduction therapy prior to onset of overt relapse.

SUPPORTIVE CARE

Measures geared to supporting patients through several weeks of neutropenia and thrombocytopenia are critical to successful AML therapy. Patients with AML should be treated in centers expert in providing supportive care. Multilumen central venous catheters should be inserted as soon as newly diagnosed AML patients have been stabilized. They should be used thereafter for administration of intravenous medications/chemotherapy and transfusions, as well as for blood drawing instead of venipuncture.

Adequate and prompt blood bank support is critical to therapy of AML. Platelet transfusions should be given as needed to maintain a platelet count ≥10,000/μL. The platelet count should be kept at higher levels in febrile patients and during episodes of active bleeding or DIC. Patients with poor posttransfusion platelet count increments may benefit from administration of platelets from HLA-matched donors. RBC transfusions should be administered to keep the hemoglobin level >70–80 g/L (7–8 g/dL) in the absence of active bleeding, DIC, or congestive heart failure, which require higher hemoglobin levels. Blood products leukodepleted by filtration should be used to avert or delay alloimmunization as well as febrile reactions. Blood products may also be irradiated to prevent transfusion-associated GVHD. Cytomegalovirus (CMV)–negative blood products should be used for CMV-seronegative patients who are potential candidates for allogeneic HCT; fortunately white blood cell filtration is quite effective at reducing CMV exposure as well.

Neutropenia (neutrophils <500/μL or <1000/μL and predicted to decline to <500/μL over the next 48 h) can be part of the initial presentation and/or a side effect of the chemotherapy treatment in AML patients. Thus, infectious complications remain the major cause of morbidity and death during induction and postremission chemotherapy for AML. Antibacterial (i.e., quinolones) and antifungal (i.e., posaconazole) prophylaxis, especially in conjunction with regimens that cause mucositis, is beneficial. For patients who are herpes simplex virus or varicella-zoster seropositive, antiviral prophylaxis should be initiated (e.g., acyclovir, valacyclovir).

Fever develops in most patients with AML, but infections are documented in only half of febrile patients. Early initiation of empirical broad-spectrum antibacterial and antifungal antibiotics has significantly reduced the number of patients dying of infectious complications **(Chap. 70)**. An antibiotic regimen adequate to treat gram-negative organisms should be instituted at the onset of fever in a neutropenic patient after clinical evaluation, including a detailed physical examination with inspection of the indwelling catheter exit site and a perirectal examination (for perirectal abscess), as well as procurement of cultures and radiographs aimed at documenting the source of fever. Specific antibiotic regimens should be based on institutional antibiotic sensitivity data obtained from where the patient is being treated. Acceptable regimens for empiric antibiotic therapy include monotherapy with imipenem-cilastatin, meropenem, piperacillin/tazobactam, or an extended-spectrum antipseudomonal cephalosporin (cefepime or ceftazidime). The combination of an aminoglycoside with an antipseudomonal penicillin (e.g., piperacillin) or an aminoglycoside in combination with an extended-spectrum antipseudomonal cephalosporin should be considered in complicated or resistant cases. Aminoglycosides should be avoided, if possible, in patients with renal insufficiency. Empirical vancomycin should be added in neutropenic patients with catheter-related infections, blood cultures positive for gram-positive bacteria before final identification and susceptibility testing, hypotension or shock, or known colonization with penicillin/cephalosporin-resistant pneumococci or methicillin-resistant *Staphylococcus aureus*. In special situations where decreased susceptibility to vancomycin, vancomycin-resistant organisms, or vancomycin toxicity is documented, other options including linezolid, daptomycin, and quinupristin/dalfopristin need to be considered.

Caspofungin (or a similar echinocandin), voriconazole, isavuconazonium, or liposomal amphotericin B should be considered for antifungal treatment if fever persists for 4–7 days following initiation of empiric antibiotic therapy. Amphotericin B has long been used for antifungal therapy. Although liposomal formulations have improved the toxicity profile of this agent, its use has been limited to situations with high risk of or documented mold infections. Caspofungin has been approved for empiric antifungal treatment. Voriconazole has also been shown to be equivalent in efficacy and less toxic than amphotericin B; isavuconazonium may also be effective with fewer drug-drug interactions. Antibacterial and antifungal antibiotics should be continued until patients are no longer neutropenic, regardless of whether a specific source has been found for the fever. Unfortunately, this practice likely contributes to development of resistance and increased incidence of nosocomial infections such as *Clostridium difficile* colitis, so great care should be taken preferably in hospital-wide antibiotic surveillance and isolation strategies to reduce these complications. Recombinant hematopoietic growth factors have a limited role in AML; myeloid growth factors may be useful in the postremission setting but are not recommended in induction or for "palliative" care for patients not in remission.

TREATMENT FOR REFRACTORY OR RELAPSED AML

In patients who relapse after achieving CR, the length of first CR is predictive of response to salvage chemotherapy treatment; patients with longer first CR (>12 months) generally relapse with drug-sensitive disease and have a higher chance of attaining a CR, even with the same chemotherapeutic agents used for first remission induction. Whether initial CR was achieved with one or two courses of chemotherapy and the type of postremission therapy may also predict achievement of second CR. Similar to patients with refractory disease, patients with relapsed disease are rarely cured by salvage chemotherapy treatments. Therefore, patients who eventually achieve a second CR and are eligible for allogeneic HCT should be transplanted. However, there is no consensus on optimal treatment for patients who relapse after allogeneic HCT; outcomes in this setting are very poor.

Because achievement of a second CR with routine salvage therapies is relatively uncommon, especially in patients who relapse rapidly after achievement of first CR (<12 months), these patients and those lacking HLA-compatible donors or who are not candidates for allogeneic HCT should be considered for innovative approaches on clinical trials. Many new agents are in current testing (Table 100-6). The discovery of novel gene mutations and mechanisms of leukemogenesis that might represent actionable therapeutic targets has prompted the development of many new targeting agents. In addition to kinase inhibitors for *FLT3*-mutated AML, other compounds targeting the aberrant activity of mutant proteins (e.g., IDH1/2 inhibitors) and numerous other biologic mechanisms are being tested in clinical trials. Furthermore, approaches with antibodies targeting markers commonly expressed on leukemia blasts (e.g., CD33) or leukemia-initiating cells (e.g., CD123) are also under investigation. Once these compounds have demonstrated safety and activity as single agents, investigation of combinations with other molecular targeting compounds and/or chemotherapy should be pursued.

TREATMENT OF ACUTE PROMYELOCYTIC LEUKEMIA

APL is a highly curable AML subtype, and ~85% of these patients achieve long-term survival with current approaches. APL has long been shown to be responsive to cytarabine and daunorubicin, but in the past patients who were treated with these drugs alone frequently died from DIC induced by the release of granule components by the chemotherapy-treated leukemia cells. However, the prognosis of APL patients has changed dramatically with the introduction of tretinoin (ATRA), an oral drug that induces the differentiation of leukemic cells bearing the t(15;17), where disruption of the *RARA* gene encoding a retinoid acid receptor occurs. ATRA decreases the frequency of DIC but often produces another complication called the APL (differentiation) syndrome. Occurring within the first 3 weeks of treatment, it is characterized by fever, fluid retention, dyspnea, chest pain, pulmonary infiltrates, pleural and pericardial effusions, and hypoxemia. The syndrome is related to adhesion of differentiated neoplastic cells to the pulmonary vasculature endothelium. Glucocorticoids, chemotherapy, and/or supportive measures can be effective for management of the APL syndrome. Temporary discontinuation of ATRA is necessary in cases of severe APL syndrome (i.e., patients developing renal failure or requiring admission to the intensive care unit due to respiratory distress). The mortality rate of this syndrome is ~10%. APL syndrome may also occur, less commonly, with ATO in APL.

In low-risk APL (low leukocyte count at presentation), ATRA (45 mg/m^2/d) plus ATO (0.15 mg/kg/d) was recently compared to ATRA plus concurrent idarubicin chemotherapy. ATRA/ATO was superior and is the new standard of care for such patients. CR rates in low-risk disease approach 100%, with excellent long-term survival. Notably, patients with high-risk APL (high leukocyte count) must be uniquely treated, as they require immediate cytoreduction with chemotherapy due to life-threatening APL syndrome often with rapidly rising leukocyte count after initiation of ATRA. High-risk patients are at increased risk for induction death due to this syndrome as well as increased frequency of hemorrhagic complications (related to DIC).

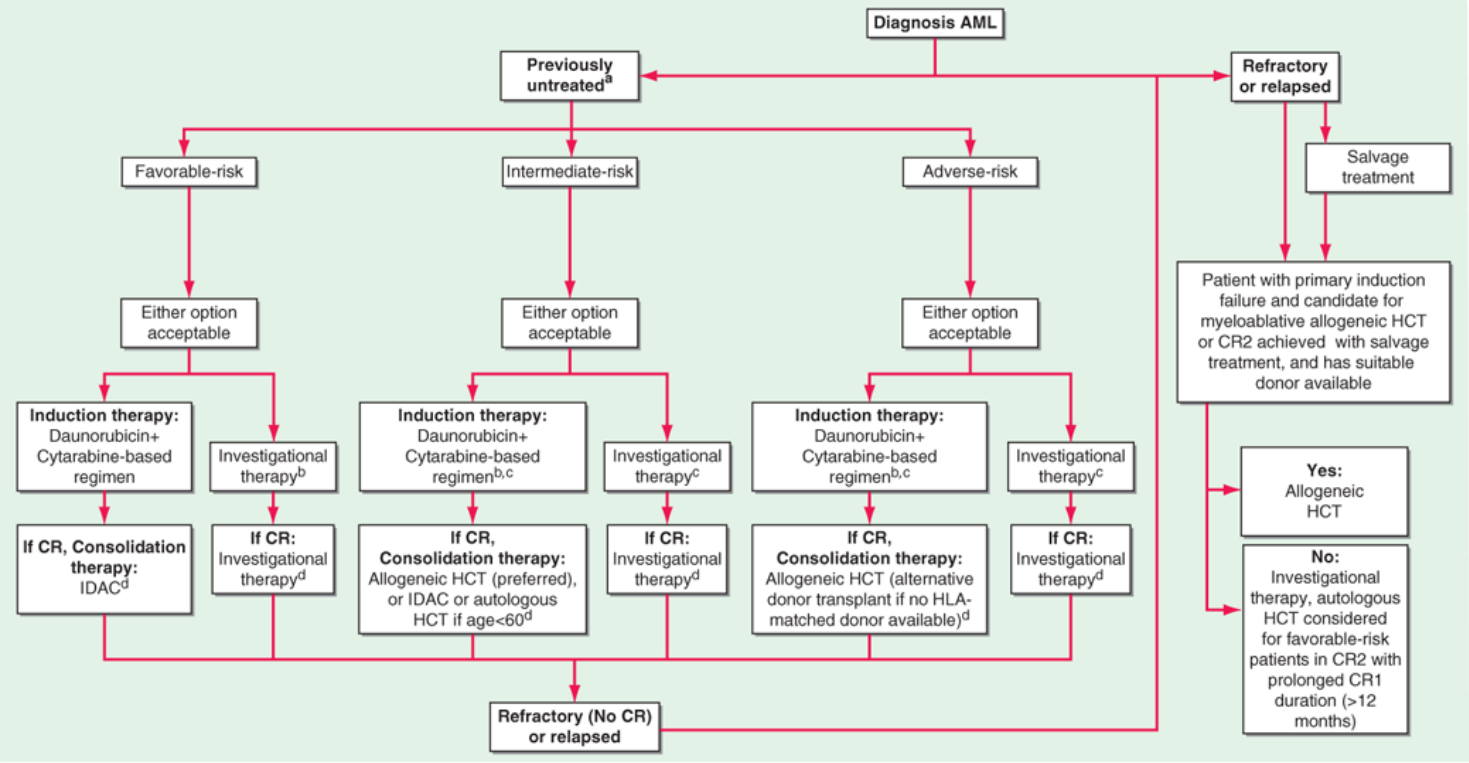
Assessment of residual disease by RT-PCR amplification of the t(15;17) chimeric gene product PML-RARA following the final cycle of treatment is important. Disappearance of the signal is associated with long-term disease-free survival; its persistence or reemergence invariably predicts relapse. Sequential monitoring of RT-PCR for PML-RARA is now considered standard for postremission monitoring of APL, at least in high-risk patients.

Patients in molecular, cytogenetic, or clinical relapse should be salvaged with ATO with or without ATRA; in patients who were treated with ATRA plus chemotherapy in the frontline setting, ATO-based therapy at relapse produces meaningful responses in up to 85% of patients. Though experience with relapsed APL in patients who received ATO during initial induction is limited (given that few relapses occur in low-risk patients, and widespread use of ATO during first-line therapy is relatively new), ATO remains the preferred reinduction therapy for patients who relapse. Achievement of CR2 should be followed by consolidation with autologous HCT (for patients who achieve RT-PCR negative status). In the minority who do not achieve negative RT-PCR or who relapse again, allogeneic HCT may still be potentially curative.

FIGURE 100-2

Flowchart for the therapy of newly diagnosed acute myeloid leukemia (AML). [a]Risk stratification according to the European LeukemiaNet (see Table 100-3). [b]Younger patients (<60–65 years) should routinely be offered investigational therapy on a backbone of standard chemotherapy for induction and consolidation. [c]Older patients, especially those >65 years or with adverse-risk disease, or those who are unfit for intensive daunorubicin + cytarabine regimens, may be considered for investigational therapy alone or in combination with lower intensity chemotherapy regimens (azacitidine, decitabine). [d]Investigational therapy as maintenance should be considered if available (after consolidation for younger patients and older patients with favorable-risk disease, and for all other older patients after induction).

For all forms of AML except acute promyelocytic leukemia (APL), standard induction therapy includes a regimen based on a 7-day continuous infusion of cytarabine (100–200 mg/m^2/d) and a 3-day course of daunorubicin (60–90 mg/m^2/d) with or without additional drugs. Idarubicin (12 mg/m^2/d) could be used in place of daunorubicin (not shown). The value of postremission/consolidation therapy for older patients (>60 years) who do not have favorable-risk disease is uncertain. Patients who achieve complete remission (CR) undergo postremission consolidation therapy, including sequential courses of intermediate-risk cytarabine, allogeneic HCT, autologous HCT, or novel therapies, based on their predicted risk of relapse (i.e., risk-stratified therapy). Patients with APL (see text for treatment) usually receive tretinoin and arsenic trioxide–based regimens with or without anthracycline-based chemotherapy and possibly maintenance with tretinoin. HCT, hematopoietic cell transplantation; HLA, human leukocyte antigen; IDAC, intermediate dose cytarabine.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 100-6

Novel Therapies in Clinical Development in Acute Myeloid Leukemia

Protein kinase inhibitors	• FLT3 inhibitors (midostaurin, quizartinib, gilteritinib, crenolanib, sorafenib) • KIT inhibitors • PI3K/AKT/mTOR inhibitors • Aurora and polo-like kinase inhibitors, CDK4/6 inhibitors, CHK1, WEE1, and MPS1 inhibitors • SRC and HCK inhibitors • Syk inhibitors
Epigenetic modulators	• New DNA methyltransferase inhibitors (SGI-110) • Histone deacetylase (HDAC) inhibitors • IDH1 and IDH2 inhibitors • DOT1L inhibitors • BET-bromodomain inhibitors
Chemotherapeutic agents	• CPX-351 (especially in secondary AML) • Vosaroxin • Nucleoside analogues
Mitochondrial inhibitors	• Bcl-2, Bcl-xL, and Mcl-1 inhibitors • Caseinolytic protease inhibitors
Therapies targeting oncogenic proteins	• Fusion transcripts targeting • EVI1 targeting • NPM1 targeting • Hedgehog inhibitors (glasdegib)
Antibodies and immunotherapies	• Monoclonal antibodies against CD33, CD44, CD47, CD123, CLEC12A • Immunoconjugates (e.g., gemtuzumab ozogamicin, SGN33A) • Bispecific T-cell engagers (BiTEs) and dual affinity re-targeting molecules (DARTs) • Chimeric antigen-receptor (CAR) T cells or genetically engineered T-cell receptor (TCR) T cells • Immune checkpoint inhibitors (PD-1/PD-L1, CTLA-4) • Anti-KIR antibody (lirilumab) • Vaccines (e.g., WT1)
Therapies targeting AML environment	• CXCR4 and CXCL12 antagonists • Anti-angiogenic therapies

Source: Adapted from Döhner H et al: Diagnosis and management of acute myeloid leukemia in adults: 2017 recommendations from an international expert panel, on behalf of the European LeukemiaNet. Blood. 129:424, 2017.

FURTHER READING

Ablain J et al: Activation of a promyelocytic leukemia–tumor protein 53 axis underlies acute promyelocytic leukemia cure. Nature Medicine 20: 167, 2014. [PubMed: 24412926]

Arber DA et al: Acute myeloid leukaemia (AML) with recurrent genetic abnormalities, in *World Health Organization Classification of Tumours of Haematopoietic and Lymphoid Tissues*, update to 4th ed. In Press.

Burnett AK et al: Arsenic trioxide and all-trans retinoic acid treatment for acute promyelocytic leukaemia in all risk groups (AML17): Results of a randomised, controlled, phase 3 trial. Lancet Oncol 16:1295, 2015. [PubMed: 26384238]

Cancer Genome Atlas Research Network: Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med 368:2059, 2003.

Döhner H et al: Diagnosis and management of acute myeloid leukemia in adults: 2017 recommendations from an international expert panel, on behalf of the European LeukemiaNet. Blood. 129:424, 2017.

Döhner H et al: Review article: Acute myeloid leukemia. N Engl J Med 373:1136, 2015. [PubMed: 26376137]

Fernandez HF et al: Anthracycline dose intensification in acute myeloid leukemia. N Engl J Med 361:1249, 2009. [PubMed: 19776406]

Lo-Coco F et al: Retinoic acid and arsenic trioxide for acute promyelocytic leukemia. N Engl J Med 369:111, 2013. [PubMed: 23841729]

Löwenberg B et al: Cytarabine dose for acute myeloid leukemia. N Engl J Med 364:1027, 2011. [PubMed: 21410371]

Papaemmanuil E et al: Genomic classification and prognosis in acute myeloid leukemia. N Engl J Med 374:2209, 2016. [PubMed: 27276561]

Peterson L et al: Myeloid neoplasms with germline predisposition, in *World Health Organization Classification of Tumours of Haematopoietic and Lymphoid Tissues*, update to 4th ed. In Press.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 101: Chronic Myeloid Leukemia

Hagop Kantarjian; Jorge Cortes

INTRODUCTION

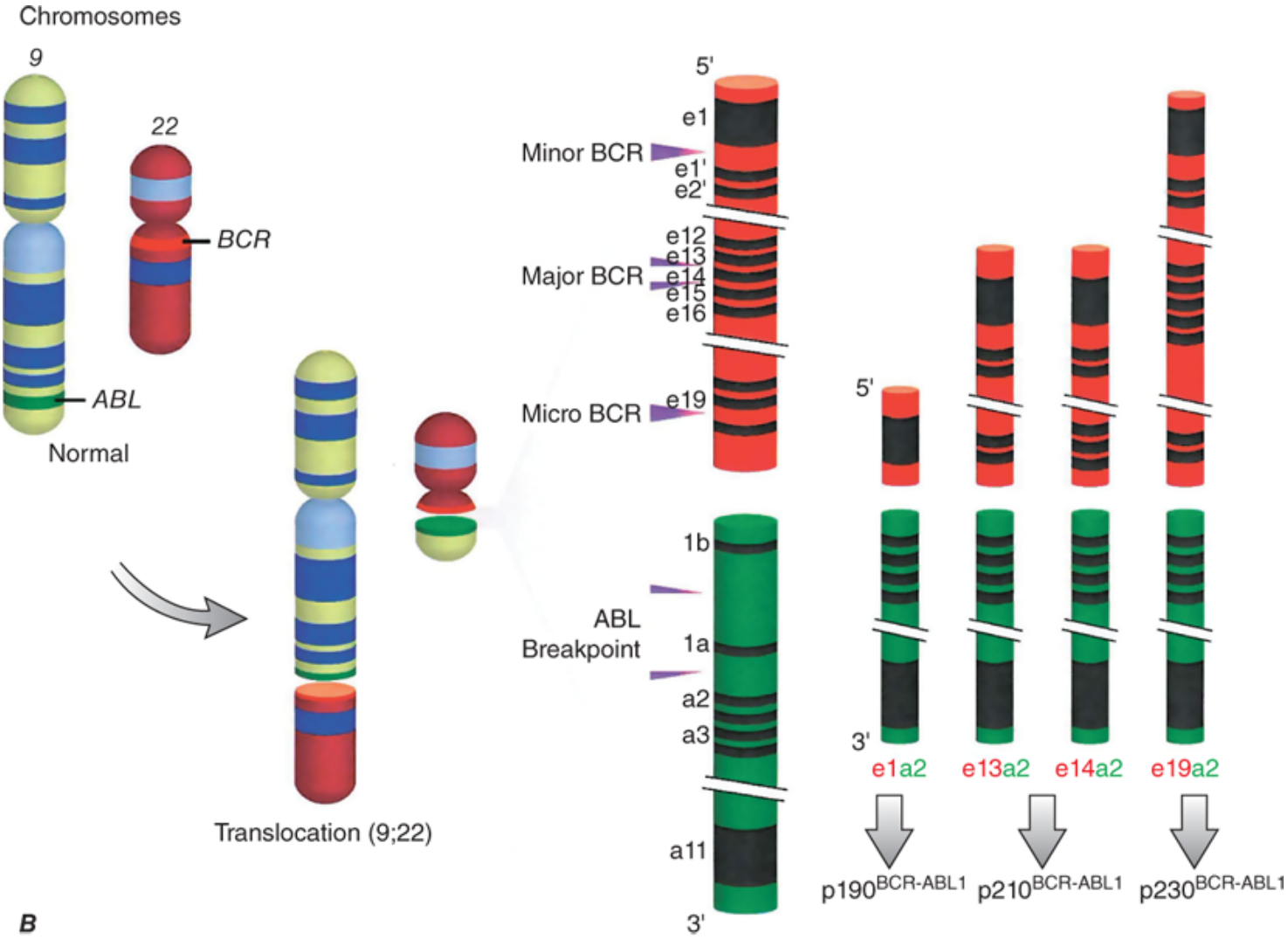
Chronic myeloid leukemia (CML) is a clonal hematopoietic stem cell disorder. The disease is driven by the *BCR-ABL1* chimeric gene product, that codes for a constitutively active tyrosine kinase, resulting from a reciprocal balanced translocation between the long arms of chromosomes 9 and 22, t(9;22)(q34.1;q11.2), known as the Philadelphia chromosome (Ph) **(Fig. 101-1)**. Untreated, the course of CML is typically biphasic or triphasic, with an early indolent or chronic phase, followed often by an accelerated phase and a terminal blastic phase. Before the era of selective BCR-ABL1 tyrosine kinase inhibitors (TKIs), the median survival in CML was 3–7 years, and the 10-year survival rate was 30% or less. Introduced into standard CML therapy in 2000, TKIs have revolutionized the treatment, natural history, and prognosis of CML. Today, the estimated 10-year survival rate with imatinib mesylate, the first BCR-ABL1 TKI approved, is 85%. Allogeneic stem cell transplantation (SCT), a curative approach but one that involves more risks, is now more often offered as second- or third-line therapy after failure of TKIs.

FIGURE 101-1

A. The Philadelphia (Ph) chromosome cytogenetic abnormality. ***B.*** Breakpoints in the long arms of chromosome 9 (*ABL* locus) and chromosome 22 (*BCR* regions) result in at least three different BCR-ABL1 oncoprotein messages, $p210^{BCR\text{-}ABL1}$ (most common message in chronic myeloid leukemia [CML]), $p190^{BCR\text{-}ABL1}$ (present in two-thirds of patients with Ph-positive acute lymphocytic leukemia; rare in CML), and $p230^{BCR\text{-}ABL1}$ (rare in CML and associated with an indolent course). Other rearrangements (e.g., b14a3) are less common. *(© 2013 The University of Texas MD Anderson Cancer Center.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

INCIDENCE AND EPIDEMIOLOGY

CML accounts for ~15% of all cases of leukemia. There is a slight male preponderance (male:female ratio 1.6:1). The median age at diagnosis is 55–65 years. It is uncommon in children; only 3% of patients with CML are younger than 20 years although in recent years a higher proportion of young patients seem to be diagnosed. CML incidence increases slowly with age, with a steeper increase after the age of 40–50 years. The annual incidence of CML is 1.5 cases per 100,000 individuals. In the United States this translates into about 8000 new cases per year. The incidence of CML has not changed over several decades. By extrapolation, the worldwide annual incidence of CML is about 100,000–120,000 cases. With a median survival of 6 years before 2000, the disease prevalence in the United States was 25,000–30,000 cases. With TKI therapy, the annual mortality has been reduced from 10–20% to about 2%. Therefore, the prevalence of CML in the United States is expected to continue to increase (approximately 100,000 in 2016). The worldwide prevalence will depend on the treatment penetration of TKIs and their effect on reduction of worldwide annual mortality. Ideally, with full TKI treatment penetration, the worldwide prevalence should plateau at 35 times the incidence, or around 3–4 million patients.

ETIOLOGY

There are no familial associations in CML. The risk of developing CML is not increased in monozygotic twins or in relatives of patients. No etiologic agents are incriminated, and no associations exist with exposures to benzene or other toxins, fertilizers, insecticides, or viruses. CML is not a frequent secondary leukemia following therapy of other cancers with alkylating agents and/or radiation. Exposure to ionizing radiation (e.g., nuclear accidents, radiation treatment for ankylosing spondylitis or cervical cancer) has increased the risk of CML, which peaks at 5–10 years after exposure and is dose-related. The median time to development of CML among atomic bomb survivors was 6.3 years. Following the Chernobyl accident, the incidence of CML did not increase, suggesting that larger dose exposures of radiation are required to cause CML. Because of adequate protection, the risk of CML is not increased in individuals working in the nuclear industry or among radiologists in recent times.

PATHOPHYSIOLOGY

The t(9;22)(q34.1;q11.2) is present in >90% of classical CML cases. It results from a balanced reciprocal translocation between the long arms of chromosomes 9 and 22. It is present in hematopoietic cells (myeloid, erythroid, megakaryocytes, and monocytes; less often mature B lymphocytes; rarely mature T lymphocytes, but not stromal cells), but not in other cells in the human body. As a result of the translocation, DNA sequences from the cellular oncogene *ABL1* are translocated next to the major breakpoint cluster region (*BCR*) gene on chromosome 22, generating a hybrid oncogene, *BCR-ABL1*. This fusion gene typically encodes for a novel oncoprotein of molecular weight 210 kDa, referred to as $p210^{BCR\text{-}ABL1}$ (Fig. 101-1B). This BCR-ABL1 oncoprotein exhibits constitutive kinase activity that leads to excessive proliferation and reduced apoptosis of CML cells, endowing them with a growth advantage over their normal counterparts. Over time, normal hematopoiesis is suppressed, but normal stem cells can persist and reemerge following effective therapy, for example with TKIs. In most instances of Ph-positive acute lymphoblastic leukemia (ALL) and in rare cases of CML, the breakpoint in *BCR* is more centromeric, in a region called the minor *BCR* region (*mBCR*). As a result, a shorter sequence of *BCR* is fused to *ABL1*, with a consequent smaller BCR-ABL1 oncoprotein, $p190^{BCR\text{-}ABL1}$. When occurring in Ph-positive CML, this translocation may predict for a worse outcome. A third rarer breakpoint in *BCR* occurs telomeric to the major *BCR* region and is called *micro-BCR* (μ-BCR). It juxtaposes a larger fragment of the *BCR* gene to *ABL1* and produces a larger $p230^{BCR\text{-}ABL1}$ oncoprotein, which is associated with a more indolent CML course. Other rearrangements, such as b14a3, occur much less frequently.

The constitutive activation of *BCR-ABL1* results in autophosphorylation and activation of multiple downstream pathways that affect gene transcription, apoptosis, stromal adherence, skeletal organization, and degradation of inhibitory proteins. These transduction pathways may involve RAS, mitogen-activated protein (MAP) kinases, signal transducers and activators of transcription (STAT), phosphatidylinositol-3-kinase (PI3k), MYC, and others. These interactions are mostly mediated through tyrosine phosphorylation and require binding of BCR-ABL1 to adapter proteins such as GRB-2, CRK, CRK-like (CRK-L) protein, and Src homology containing proteins (SHC). Most BCR-ABL1 TKIs bind to the *BCR-ABL1* ATP-binding domain, preventing the activation of transformation pathways and inhibiting downstream signaling. As a result, proliferation of CML cells is inhibited and apoptosis induced, allowing the reemergence of normal hematopoiesis. A plethora of signaling pathways have been implicated in BCR-ABL1-mediated cellular transformation. The emerging picture is a complex and redundant transformation network. An additional layer of complexity is related to differences in signal transduction between CML-differentiated cells and early progenitors. Beta-catenin, Wnt1, Foxo3a, transforming growth factor β, interleukin-6, PP2A, SIRT1, and others have been implicated in CML stem cell survival.

Experimental models have established the causal relationship between the *BCR-ABL1* rearrangement and the development of CML. In animal models, expression of *BCR-ABL1* in normal hematopoietic cells produced CML-like disorders or lymphoid leukemia, demonstrating the leukemogenic potential of *BCR-ABL1* as a single oncogenic abnormality. Other models however suggest the need for a “second hit.”

The cause of the *BCR-ABL1* molecular rearrangement is unknown. Molecular techniques that detect *BCR-ABL1* at a level of 1 in 10^8 identify this molecular abnormality in the blood of up to 25% of normal adults and 5% of infants, but 0% of cord blood samples. This suggests that *BCR-ABL1* is not sufficient to cause overt CML in the overwhelming majority of individuals in whom it occurs. Because CML develops in only 1.5 of 100,000 individuals annually, it is evident that additional molecular events or poor immune recognition of the rearranged cells are needed to cause overt CML.

CML is defined by the presence of *BCR-ABL1* fusion gene in a patient with a myeloproliferative neoplasm. In some patients with a typical morphologic picture of CML, the Ph chromosome is not detectable by standard G-banding karyotype, but fluorescence in situ hybridization (FISH) and/or molecular studies (polymerase chain reaction [PCR]) detect *BCR-ABL1*. These patients have a course similar to Ph-positive CML and respond to TKI therapy. Many of the remaining patients have atypical morphologic or clinical features and belong to other diagnostic groups, such as atypical CML, chronic myelomonocytic leukemia, and myelodysplastic-myeloproliferative neoplasms (MDS-MPN). These individuals do not respond to TKI therapy and have a poor prognosis with a median survival of about 2–3 years. Detection of mutations in the granulocyte colony-stimulating factor receptor (*CSF3R*) in chronic neutrophilic leukemia (90% of cases) and in some cases of atypical CML, of mutations in *SETBP1* in atypical CML (25% of cases), and of mutations in *SF3B1* in MDS-MPN with ringed sideroblasts and marked thrombocytosis (MDS-MPD-RST; 50–70% of cases, associated with longer median survival of 7 years vs 3.3 years with wild-type *SF3B1*), confirmed that they are distinct molecular and biologic entities.

The events associated with the transition of CML from a chronic to accelerated-blastic phase are poorly understood. They are often associated with characteristic chromosomal abnormalities such as a double Ph, trisomy 8, isochromosome 17 or deletion of 17p (loss of *TP53*), 20q−, and others. Molecular events associated with transformation include mutations in *TP53*, retinoblastoma 1 (*RB1*), myeloid transcriptions factors like *RUNX1*, and cell cycle regulators like *p16*. A plethora of other mutations or functional abnormalities have been implicated in blastic transformation, but no unifying theme has emerged other than the fact that *BCR-ABL1* itself induces genetic instability that favors the acquisition of additional molecular events and eventually to blastic transformation. In this frame of thinking, one critical effect of TKIs is their ability to stabilize the CML genome, leading to a much reduced transformation rate. In particular, the previously observed sudden blastic transformations (i.e., abrupt transformation to blastic phase in a patient who had been in cytogenetic response) have become uncommon, occurring rarely in younger patients in the first 1–2 years of TKI therapy (usually sudden lymphoid blastic transformations). Sudden transformations beyond the third year of TKI therapy are rare in patients who continue on TKI therapy. Moreover, initial experience suggests that the course of CML has become significantly more indolent, even without cytogenetic responses, in patients on TKI-based therapy compared to previous experience with hydroxyurea/busulfan.

Among patients developing resistance to TKIs, several resistance mechanisms have been observed. The most clinically relevant one is the development of different *ABL1* kinase domain mutations that may prevent the binding of TKIs to the catalytic site (ATP-binding site) of the kinase or maintain the kinase activity despite the presence of a TKI. More than 100 *BCR-ABL1* mutations have now been described, many of which confer relative or absolute resistance to imatinib. This has resulted in the development of second-generation TKIs (i.e., dasatinib, nilotinib, bosutinib) and of a third-generation TKI (ponatinib) with significant efficacy against T315I, a “gatekeeper” mutation that prevents binding of and causes resistance to all other TKIs.

CLINICAL PRESENTATION

The presenting signs and symptoms in CML depend on the availability of and access to health care, including physical examinations and screening tests. In the United States, because of the wider access to health care screening and physical examinations, 50–60% of patients are diagnosed on routine blood tests and have minimal symptoms at presentation, such

as fatigue. In geographic locations where access to health care is more limited, patients often present with high CML burden including splenomegaly, anemia, and related symptoms (abdominal pain, weight loss, fatigue), which translate into a higher frequency of high-risk CML. Presenting findings in patients diagnosed in the United States are shown in **Table 101-1**.

TABLE 101-1

Presenting Signs and Symptoms of Newly Diagnosed Philadelphia Chromosome–Positive Chronic Myeloid Leukemia in Chronic Phase

Parameter	Percentage
Age ≥60 years (median)	40–50 (55–65)
Female gender	35–45
Splenomegaly	30
Hepatomegaly	5–10
Lymphadenopathy	5
Other extramedullary disease	2
Hemoglobin <10 g/dL	10–15
Platelets	
>450 × 10^9 cells/L	30–35
<100 × 10^9 cells/L	3–5
White blood cells ≥50 × 10^9 cells/L	35–40
Marrow	
≥5% blasts	5
≥5% basophils	10–15
Peripheral blood	
≥3% blasts	8–10
≥7% basophils	10
Cytogenetic clonal evolution other than the Philadelphia chromosome	4–5
Sokal risk	
Low	60–65
Intermediate	25–30

Parameter	Percentage
High	10

Symptoms

Most patients with CML (90%) present in the indolent or chronic phase. Depending on the timing of diagnosis, patients are often asymptomatic (if the diagnosis is discovered during health care screening tests). Common symptoms, when present, are manifestations of anemia and splenomegaly. These may include fatigue, malaise, weight loss (if high leukemia burden), or early satiety and left upper quadrant pain or masses (from splenomegaly). Less common presenting findings include thrombotic or hyperviscosity-related events (from severe leukocytosis or thrombocytosis). These include priapism, cardiovascular complications, myocardial infarction, venous thrombosis, visual disturbances, dyspnea and pulmonary insufficiency, drowsiness, loss of coordination, confusion, or cerebrovascular accidents. Manifestations of bleeding diatheses include retinal hemorrhages, gastrointestinal bleeding, and others. Patients who present with, or progress to, the accelerated or blastic phases frequently have additional symptoms including unexplained fever, significant weight loss, severe fatigue, bone and joint pain, bleeding and thrombotic events, and infections.

Physical Findings

Splenomegaly is the most common physical finding, occurring in 20–70% of patients depending on health care screening frequency. Other less common findings include hepatomegaly (5–10%), lymphadenopathy (5–10%), and extramedullary disease (skin or subcutaneous lesions). The latter indicates CML transformation if a biopsy confirms predominance of blasts. Other physical findings are manifestations of complications of high tumor burden described earlier (e.g., cardiovascular, cerebrovascular, bleeding). High basophil counts may be associated with histamine overproduction causing pruritus, diarrhea, flushing, and even gastrointestinal ulcers.

Hematologic and Marrow Findings

In untreated CML, leukocytosis ranging from 10–500 × 10^9/L is common. The peripheral blood differential shows left-shifted hematopoiesis with predominance of neutrophils and the presence of bands, myelocytes, metamyelocytes, promyelocytes, and blasts (usually ≤5%). Basophils and/or eosinophils are frequently increased. Thrombocytosis is common, but thrombocytopenia is rare and, when present, suggests a worse prognosis, disease acceleration, or an unrelated etiology. Anemia is present in one-third of patients. Cyclic oscillations of counts are noted in 10–20% of patients without treatment. Biochemical abnormalities include a low leukocyte alkaline phosphatase score and high levels of vitamin B_{12}, uric acid, lactic dehydrogenase, and lysozyme. The presence of unexplained and sustained leukocytosis, with or without splenomegaly, should lead to a marrow examination and cytogenetic analysis.

The bone marrow is hypercellular with marked myeloid hyperplasia and a high myeloid-to-erythroid ratio of 15–20:1. Marrow blasts are 5% or less; when higher, they carry a worse prognosis or represent transformation to accelerated phase (if they are ≥15%). Increased reticulin fibrosis (by Snook's silver stain) is common, with 30–40% of patients demonstrating grade 3–4 reticulin fibrosis. This was considered adverse in the pre-TKI era. With TKI therapy, reticulin fibrosis resolves in most patients and is not an indicator of poor prognosis. Collagen fibrosis (Wright-Giemsa stain) is rare at diagnosis. Disease progression with a "spent phase" of myelofibrosis (myelophthisis, or burnt-out marrow) was common with busulfan therapy (20–30%) but is rare with TKI therapy.

Cytogenetic and Molecular Findings

The diagnosis of CML is straightforward and depends on documenting the t(9;22)(q34.1;q11.2), which is identified by G-banding in 90% of cases. This is known as the Philadelphia chromosome (initially identified in Philadelphia as a minute chromosome), later identified to be chromosome 22 (Fig. 101-1). Some patients may have complex translocations (variant Ph) involving three or more chromosomes including chromosomes 9 and 22 and one or more additional chromosomes. Others may have a "masked Ph," involving translocations between chromosome 9 and a chromosome other than 22 (but molecularly showing the *BCR-ABL1* rearrangement). The prognosis of these patients and their response to TKI therapy are similar to those in patients with Ph. About 5–10% of patients may have additional chromosomal abnormalities in the Ph-positive cells. These usually involve trisomy 8, a double Ph, isochromosome 17 or 17p deletion, 20q–, or others. This is referred to as cytogenetic clonal evolution and was historically a sign of adverse prognosis, particularly when trisomy 8, double Ph, or chromosome 17 abnormalities were noted. A less common abnormality involving chromosome 3q26.2 also carries a poor prognosis.

Techniques such as FISH and PCR are now used to aid in the diagnosis of CML. They are more sensitive approaches to estimate the CML burden in patients on TKI therapy. They can be done on peripheral blood, and thus are more convenient to patients. Patients with CML at diagnosis should have a FISH analysis to quantify the percentage of Ph-positive cells, if FISH is used to replace marrow cytogenetic analysis in monitoring response to therapy. FISH will not detect additional chromosomal abnormalities (clonal evolution); thus, a cytogenetic analysis is usually recommended at the time of diagnosis. The *BCR-ABL1* rearrangement is usually one of two variants: e13a2 (formerly b2a2) and e14a2 (formerly b3a2). About 2–5% of patients may have other RNA fusion types (e.g., e1a2, e13a3, or e14a3). In these patients, the routine real-time PCR primers may not amplify the *BCR-ABL1* transcripts, thus leading to false-negative results. Therefore, molecular studies at diagnosis are important to document the type and presence of *BCR-ABL1* transcripts to avoid erroneously "undetectable" *BCR-ABL1* transcripts on follow-up studies, with the false impression of a complete molecular response. The presence of the Philadelphia chromosome with "negative" PCR with standard methodology should prompt investigation of atypical transcripts.

Both FISH and PCR studies can be falsely positive at low levels or falsely negative because of technical issues. Therefore, a diagnosis of CML must always rely on a marrow analysis with routine cytogenetics. The diagnostic bone marrow confirms the presence of the Ph chromosome, detects clonal evolution, that is, chromosomal abnormalities in the Ph-positive cells (which may be prognostic), and also quantifies the percentage of marrow blasts and basophils. In 10% of patients, the percentage of marrow blasts and basophils can be significantly higher than in the peripheral blood, conferring poorer prognosis or even representing disease transformation.

Monitoring patients on TKI therapy by cytogenetics, FISH, and molecular studies has become an important standard practice to assess response to therapy, emphasize compliance, evaluate possible treatment resistance, identify the need to change TKI therapy, and determine the need to assess for kinase domain mutations. It is thus important to recognize the comparability of these measures in monitoring response. A partial cytogenetic response is defined as the presence of 35% or less Ph-positive metaphases by routine cytogenetic analysis. This is roughly equivalent to *BCR-ABL1* transcripts by the International Scale (IS) of 10% or less. A complete cytogenetic response refers to the absence of Ph-positive metaphases (0% Ph positivity). This is approximately equivalent to *BCR-ABL1* transcripts (IS) of 1% or less. A major molecular response refers to *BCR-ABL1* transcripts (IS) ≤0.1%, or roughly a 3-log or greater reduction of *BCR-ABL1* transcripts from a standardized basline. A molecular response MR4.5 refers to *BCR-ABL1* transcripts (IS) ≤0.0032%, roughly equivalent to a 4.5-log reduction or greater of transcripts.

Findings in CML Transformation

Progression of CML is usually associated with leukocytosis resistant to therapy, increasing anemia, fever and constitutional symptoms, and increased blasts and basophils in the peripheral blood or marrow. Criteria of accelerated-phase CML, historically associated with median survival of <1.5 years, include the presence of 15% or more peripheral blasts, 30% or more peripheral blasts plus promyelocytes, 20% or more peripheral basophils, cytogenetic clonal evolution (presence of chromosomal abnormalities in addition to Ph), and thrombocytopenia $<100 \times 10^9$/L (unrelated to therapy). About 5–10% of patients present with de novo accelerated phase or blastic phase. The prognosis of de novo accelerated phase with TKI therapy has improved significantly, with an estimated 8-year survival rate of 75%. The median survival of accelerated phase evolving from chronic phase has also improved from a historical median survival of 18 months to an estimated 4-year survival rate of 70% on TKI therapy. Therefore, the criteria for accelerated-phase CML should be revisited because most clinical criteria defining accelerated phase have lost much of their prognostic significance. Blastic-phase CML is defined by the presence of 30% or more peripheral or marrow blasts or the presence of sheets of blasts in extramedullary disease (usually skin, soft tissues, or lytic bone lesions). Blastic-phase CML is commonly myeloid (60%) but can present uncommonly as erythroid, promyelocytic, monocytic, or megakaryocytic. Lymphoid blastic phase occurs in about 25% of patients. Lymphoblasts are terminal deoxynucleotide transferase positive and peroxidase negative (although occasionally with low positivity up to 3–5%) and express lymphoid markers (CD10, CD19, CD20, CD22). However, they also often express myeloid markers (50–80%), resulting in diagnostic challenges. This is important because, unlike other morphologic blastic phases, lymphoid blastic-phase CML is quite responsive to anti-ALL-type chemotherapy (e.g., hyper-CVAD [cyclophosphamide, vincristine, doxorubicin, and dexamethasone]) in combination with TKIs (complete response rates 60–70%; median survival 2–3 years).

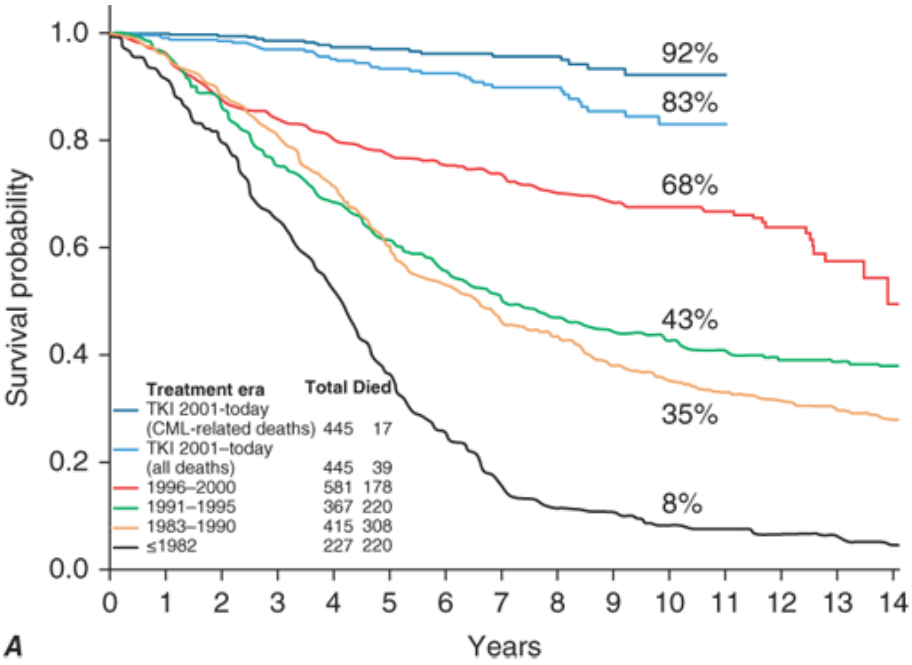
PROGNOSIS AND CML COURSE

Before the imatinib era, the annual mortality in CML was 10% in the first 2 years and 15–20% thereafter. The median survival time in CML was 3–7 years (with hydroxyurea-busulfan and interferon α). Without a curative option of allogeneic SCT, the course of CML was toward transformation to, and death from, accelerated or blastic phases for most patients as the rate of complete cytogenetic response with interferon was low. Even apparent disease stability was unpredictable, with some patients demonstrating sudden transformation to a blastic phase. With imatinib therapy, the annual mortality in CML has decreased to 2% in the first 16 years of observation. More than half of the deaths are from factors other than CML, such as aging-related comorbidities, accidents, suicides, other cancers, and other medical conditions (e.g., infections, surgical procedures). The estimated 10-year survival rate is 85%, or 93% if only CML-related deaths are considered **(Fig. 101-2)**. The course of CML has also become quite predictable. In the first 2 years of TKI therapy, rare sudden transformations are still reported (1–2%), usually lymphoid blastic transformations that respond to combinations of chemotherapy and TKIs followed by allogeneic SCT. These may be explained by the intrinsic mechanisms of sudden transformation already existing in the CML clones before the start of therapy that were not amenable to TKI inhibition, in particular imatinib. Second-generation TKIs (nilotinib, dasatinib) used as frontline therapy have reduced the incidence of transformation in the first 2–3 years from 6–8% with imatinib to 2–5% with nilotinib or dasatinib. Disease transformation to accelerated or blastic phase is rare with continued TKI therapy, estimated at <1% annually in years 4–10 of follow-up on the original imatinib trials. Patients usually develop resistance in the form of cytogenetic resistance or relapse, followed by hematologic relapse and subsequent transformation, rather than the previously feared sudden transformations without the warning signals of cytogenetic-hematologic relapse.

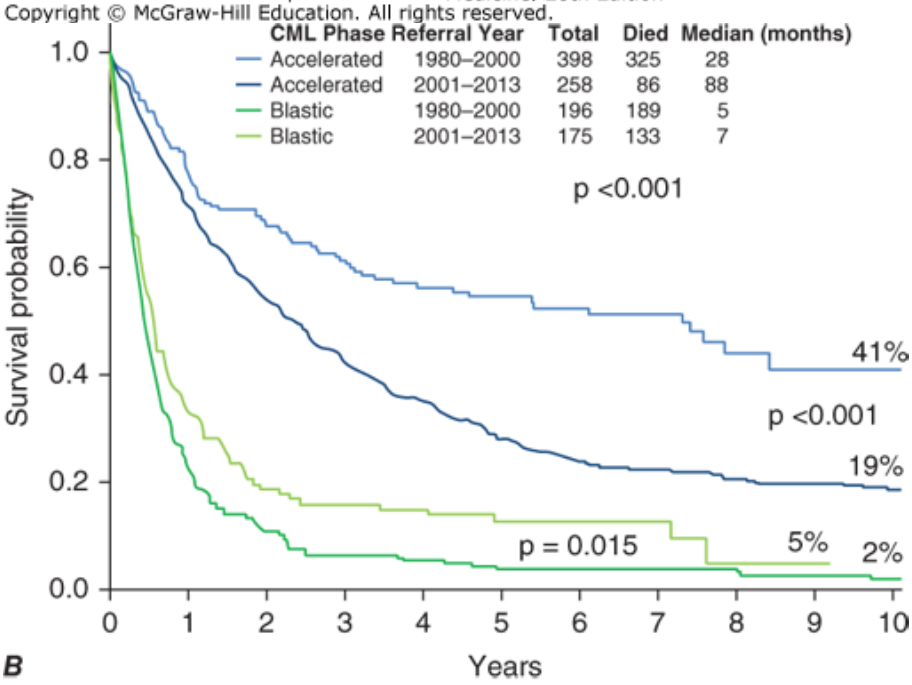
FIGURE 101-2

A. Survival in newly diagnosed chronic-phase chronic myeloid leukemia (CML) by era of therapy (MD Anderson Cancer Center experience from 1965 to present). Causes of non-CML deaths in 22 patients were other cancers (n = 7), postsurgical complications (n = 3), car accident (n = 2), suicide (n = 1), neurologic events (n = 3), cardiac (n = 3), pneumonia (n = 1), and unknown (n = 2). ***B.*** Survival in patients with accelerated- and blastic-phase CML referred to MD Anderson Cancer Center by

era of therapy, demonstrating the significant survival benefit in the tyrosine kinase inhibitor (TKI) era in accelerated-phase CML but the modest benefit in blastic-phase CML. Referred cases included de novo and post-chronic-phase transformations.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Before the imatinib era, several pretreatment prognostic factors predicted for worse outcome in CML and have been incorporated into prognostic models and staging systems. These have included older age, significant splenomegaly, anemia, thrombocytopenia or thrombocytosis, high percentages of blasts and basophils (and/or eosinophils), marrow fibrosis, deletions in the long arm of chromosome 9, clonal evolution, and others. Different risk models and staging systems, derived from multivariate analyses, were proposed to define different risk groups. As with the introduction of cisplatin into testicular cancer therapy, the introduction of TKIs into CML therapy has decreased or, in some instances,

eliminated the prognostic impact of most of these prognostic factors and the significance of the CML models (e.g., Sokal, Hasford, European Treatment and Outcome Study [EUTOS]). Treatment-related prognostic factors have emerged as the most important prognostic factors in the era of imatinib therapy. Achievement of complete cytogenetic response has become the major therapeutic endpoint and is the only endpoint associated with improvement in survival. Achievement of a major molecular response is associated with decreased risk of events (relapse) and CML transformation, but has not been associated with survival prolongation among patients with complete cytogenetic response. This may be due to the efficacy of salvage TKI therapies, which are and should be implemented at the first evidence of cytogenetic relapse. Achievement of undetectable *BCR-ABL1* transcripts, particularly when sustained (>2–3 years), may offer the possibility of treatment-free remission (molecular cure rather than functional cure) in the context of investigational trials, and may allow temporary therapy interruption in women pursuing pregnancy. The lack of achievement of major or "complete" molecular responses should not be considered as "failure" of a particular TKI therapy and/or an indication to change the TKI or to consider allogeneic SCT.

Long-term updates of randomized trials suggest that second generation TKIs and imatinib are more similarly effective in lower-risk CML, but second generation TKIs may offer a greater therapeutic advantage for patients with high-risk CML.

TREATMENT

TREATMENT

Chronic Myeloid Leukemia

With TKI therapy, the estimated 10-year survival in CML is 85%. Since 2001, six agents have been approved by the U.S. Food and Drug Administration (FDA) for the treatment of CML. These include five oral TKIs: imatinib (Gleevec, Glivec), nilotinib (Tasigna), dasatinib (Sprycel), bosutinib (Bosulif), and ponatinib (Iclusig). Imatinib 400 mg orally daily, nilotinib 300 mg orally twice a day (on an empty stomach), dasatinib 100 mg orally daily, and bosutinib 400 mg orally daily are approved for frontline therapy of CML. All four are also approved for salvage therapy (nilotinib 400 mg twice daily; bosutinib 500 mg daily), in addition to ponatinib (45 mg daily). Because of concerns of arterio-occlusive events with ponatinib, the dose is often reduced to 15–30 mg daily after a response is achieved. Imatinib, dasatinib (140 mg daily), bosutinib, and ponatinib are also approved for the treatment of CML in transformation (accelerated and blastic phase), whereas nilotinib is only approved for chronic and accelerated phase. Dasatinib, nilotinib, and bosutinib are referred to as second-generation TKIs; ponatinib is referred to as a third-generation TKI as it is the only currently approved TKI that has clinical activity in the setting of T315I mutation (in addition to unmutated BCR-ABL1 and BCR-ABL1 with other common mutations). Nilotinib is similar in structure to imatinib but 30 times more potent. Dasatinib and bosutinib inhibit SRC family of kinases in addition to ABL1, with dasatinib reported to be 300 times more potent and bosutinib 30–50 times more potent than imatinib. In contrast to all other TKI, bosutinib has no activity against c-Kit or PDGFR. Ponatinib is effective against wild-type, and mutant *BCR-ABL1* clones. It is currently the only available BCR-ABL1 TKI active against T315I, a gatekeeper mutation resistant to the other four TKIs **(Table 101-2)**. Ponatinib also inhibits VEGFR which may be related to the high incidence of hypertension observed with this agent **(Table 101-2)**. The sixth approved agent is omacetaxine (Synribo), a protein synthesis inhibitor with presumed more selective inhibition of the synthesis of the BCR-ABL1 oncoprotein. It is approved for the treatment of chronic and accelerated phase CML after failure of two or more TKIs, at 1.25 mg/m^2 subcutaneously twice a day for 14 days for induction and for 7 days for consolidation-maintenance. The main adverse event of omacetaxine is prolonged myelosuppression: omacetaxine 5–7 days induction and 2–5 days maintenance, perhaps combined with a TKI, may be equally effective and less toxic **(Table 101-2)**.

Imatinib, nilotinib, and dasatinib are all acceptable frontline therapies in CML. The long-term results of imatinib are very favorable. The 8-year follow-up results show a cumulative complete cytogenetic response rate (occurring at least once) of

83%, with 60–65% of patients being in complete cytogenetic response at 5-year follow-up. The estimated 10-year survival rate is 85%. Among patients continuing on imatinib, the annual rate of transformation to accelerated-blastic phase in years 4–8 is <1%. In two randomized studies, one comparing nilotinib 300 mg twice daily or 400 mg twice daily with imatinib (ENESTnd) and the other comparing dasatinib 100 mg daily with imatinib (DASISION), the second-generation TKIs were associated with better outcomes in early surrogate endpoints, including higher rates of complete cytogenetic responses (85–87% vs 77–82%), major molecular responses (5-year rates 76–77% vs 60–64%), and MR4.5 (5-year rates 42–53% vs 31–33%), with lower rates of transformation to accelerated and blastic phase (2–5% vs 7%). However, neither study has shown a survival benefit with second-generation TKIs (with a minimum follow-up times of 5–6 years). This may be because the rate of complete cytogenetic response is ultimately similarly high with either agent, and also because sequential therapy with TKIs (following close observation and treatment change at progression) provides highly effective therapy for most patients that allows adequate long-term outcome despite relapse or intolerance after initial therapy.

Salvage therapy in chronic phase with dasatinib, nilotinib, bosutinib, or ponatinib is associated with complete cytogenetic response rates of 30–60%, depending on the salvage status (cytogenetic vs hematologic relapse), prior response to other TKIs, and the mutations at the time of relapse. Complete cytogenetic responses are generally durable, particularly in the absence of clonal evolution and mutations. Ponatinib is the only TKI active in the setting of T315I mutation, with complete cytogenetic response rates of 50–70% among patients who have received 2 or more TKI. The estimated 5-year survival rates with new TKIs as salvage are 70–75% (compared with <50% before their availability). For example, with dasatinib salvage after imatinib failure in chronic-phase CML, the estimated 7-year rate of major molecular was 46%, the estimated 7-year survival rate was 65%, and progression-free survival rate was 42%. Thus, TKIs in the salvage setting have already reduced the annual mortality from the historical rate of 10–15% to ≤5%.

The goal of CML therapy is viewed differently in the context of research versus standard practice. In current practice, functional cure, which can be considered when the relative survival is similar to that of the general population, is the current goal of therapy. CML is now considered an indolent disease, which, with appropriate continuous TKI therapy, treatment compliance, careful monitoring, and early change to other TKIs as indicated, can be associated with close to normal survival. Therefore, in standard practice, achievement and maintenance of a complete cytogenetic response are the aims of therapy because complete cytogenetic response is the only outcome associated with survival prolongation. Lack of achievement of a major molecular response (protects against events; associated with longer event-free survival) or of negative *BCR-ABL1* transcripts (offers the potential of TKI interruption on investigational studies) should not be considered indications to change TKI therapy or to consider allogeneic SCT. A general practice rule is to continue the particular TKI chosen at the most tolerable dose schedule not associated with grade 3–4 side effects or with bothersome chronic side effects, for as long as possible, until either cytogenetic relapse or the persistence of unacceptable side effects. These two factors (i.e., cytogenetic relapse and intolerable side effects as judged by the patient and treating physician) are the indicators of “failure” of a particular TKI therapy. A second emerging general practice rule is that patients with CML should always receive daily TKI therapy throughout their lifetime (chronic, transformation), either alone (chronic) or in combinations (possibly for those in transformation although combinations not formally approved), except perhaps in situations of “molecular cure” (elective discontinuation of TKI with close observation if *BCR-ABL* transcripts undetectable are sustained for >2–3 years) or after allogeneic SCT with undetectable disease.

Because of the increasing prevalence of CML (cost of TKI therapy) and the emerging evidence of possible organ toxicities with long-term use (e.g., renal with imatinib, arterio-occlussive with nilotinib, dasatinib, and ponatinib), a goal of therapy of increasing interest in CML is to achieve eradication of the disease (molecular cure) that is prolonged and durable, with recovery of non-neoplastic, non-clonal hematopoiesis off TKI therapy. The first step toward this aim is to obtain the highest rates of undetectable *BCR-ABL1* transcripts lasting for at least 2 or more years. This is currently achievable in about 25–30% of patients treated with imatinib and in 40–45% of patients treated with second-generation TKIs. As a result, molecular cures (off TKI therapy) are estimated to be about 15% post-imatinib therapy and 20–25% post–second-generation TKIs.

Recommendations provided by the National Comprehensive Cancer Network (NCCN) and by the European LeukemiaNet (ELN) propose optimal/expected, suboptimal/warning, and failure response scenarios at different time points of TKI treatment duration. Unfortunately, they may have been misinterpreted in current practice, because oncologists often report that their aim of treatment is the achievement of major molecular response and disease eradication. Significantly, a substantial proportion of oncologists consider a change of TKI therapy in a patient in complete cytogenetic response if they note "loss of major molecular response" (increase of *BCR-ABL1* transcripts ([IS] from ≤0.1 to >0.1%). This perception may be the result of confusion regarding the aims of the NCCN and ELN guidelines, which have been updated often as a result of maturing data and have multiple treatment endpoint considerations. Although such endpoints may have been suggested as possible criteria for failure or suboptimal response, it is important to emphasize that no randomized study has yet shown that a change of TKI treatment in patients with complete cytogenetic response because of a loss of major molecular response, versus changing at the time of cytogenetic relapse, has been shown to improve survival or other long-term outcome. This is likely because of the high efficacy of salvage TKI therapy at the time of cytogenetic relapse.

Side effects of TKIs are generally mild to moderate, although with long-term TKI therapy, they could affect the patient's quality of life. Serious side effects occur in <5–10% of patients. With imatinib therapy, common mild to moderate side effects include fluid retention, weight gain, nausea, diarrhea, skin rashes, periorbital edema, bone or muscle aches, fatigue, and others (rates of 10–20%). In general, second-generation TKIs are associated with lower rates of these bothersome adverse events. However, dasatinib 100 mg daily is associated with higher rates of myelosuppression (20–30%), particularly thrombocytopenia, with pleural (10–25%) or pericardial effusions (≤5%), and with pulmonary hypertension (<5%). Nilotinib is associated with higher rates of hyperglycemia (10–20%), pruritus and skin rashes, hyperbilirubinemia (typically among patients with Gilbert's syndrome and mostly of no clinical consequences), and headaches. Nilotinib is also associated with occasional instances of pancreatitis (<5%). Bosutinib is associated with higher rates of liver toxicity and of early and self-limited gastrointestinal adverse events, particularly diarrhea (70–85%). Ponatinib is associated with higher rates of skin rashes (10–15%), pancreatitis (10%), elevations of amylase/lipase (10%), and systemic hypertension (50–60%; severe in 20%). Arterio-occlusive events (cardiovascular, cerebrovascular, and peripheral arterial) have been reported with most TKI. The incidence appears to be highest with ponatinib, but both nilotinib and dasatinib are associated with these events at an incidence significantly higher than imatinib. Nilotinib and dasatinib may cause prolongation of the QTc interval; therefore, they should be evaluated cautiously in patients with prolonged QTc interval on electrocardiogram (>470–480 ms), and drugs given for other medical conditions should have relatively smaller or no effects on QTc. These side effects can often be dose-dependent and are generally reversible with treatment interruptions and dose reductions. Dose reductions can be individualized. However, the lowest estimated effective doses of TKIs (from different studies and treatment practices) are imatinib 200–300 mg daily; nilotinib 150–200 mg twice daily; dasatinib 20 mg daily; bosutinib 300 mg daily; and ponatinib 15 mg daily.

With long-term follow-up, rare but clinically relevant serious toxicities are emerging. Renal dysfunction and occasionally renal failure (creatinine elevations >2–3 mg/dL) are observed in 2–3% of patients, more frequently with imatinib and bosutinib than other TKI, and usually reverse with TKI discontinuation and/or dose reduction. Rarely, patients may develop TKI-related peripheral neuropathy or even central neurotoxicities that are misdiagnosed as dementia or Alzheimer's disease; they may reverse slowly after TKI discontinuation. Pulmonary hypertension has been reported with dasatinib (<1–2%) and should be considered in a patient with shortness of breath and a normal chest x-ray (echocardiogram with emphasis on measurement of pulmonary artery pressure). This may be reversible with dasatinib discontinuation and occasionally the use of sildenafil citrate. Systemic hypertension has been observed more often with ponatinib. Hyperglycemia and occasionally diabetes have been noted more frequently with nilotinib. Finally, mid- and small-vessel arterio-occlusive and vasospastic events have been reported at low but significant rates with nilotinib and ponatinib and should be considered possibly TKI-related and represent indications to interrupt or reduce the dose of the TKI. These events include angina, coronary artery disease, myocardial infarction, peripheral arterial occlusive disease, transient

ischemic attacks, cerebral vascular accidents, Raynaud's phenomenon, and accelerated atherosclerosis. Although these events are uncommon (<5%) (10-year cumulative rates 10% with nilotinib 300 mg BID, 16% with 400 mg BID, compared with 2.5% with imatinib), they are clinically significant for the patient's long-term prognosis and occur at significantly higher rates than in the general population. Serious arterio-occlusive and vasospastic events are more common with ponatinib 45 mg daily (5-year rates 20%).

Discontinuation of TKIs and Treatment-Free Remissions or "Molecular Cures"

Several studies have confirmed that TKI discontinuation among patients who achieve undetectable *BCR-ABL1* transcripts for longer than 2–3 years can result in treatment-free remission rates of 40–60%. Since the incidence of durable undetectable *BCR-ABL1* transcripts is 20–45%, about 13–22% of all patients with CML on TKI therapy may achieve treatment-free remission status or molecular cure. This approach is still considered investigational, but may be ready for community practice provided it is done under optimal conditions. These include the following: patients must have low Sokal-risk CML in first chronic phase (no evidence of transformation), with history of quantifiable *BCR-ABL1* transcripts (e13a2, e14a2), on long-term TKI therapy (5 to 8+ years), with documented undetectable *BCR-ABL1* transcripts for >2–3 years (assessed every 6 months during this timespan and with a PCR with adequate sensitivity), and should be monitored at referral centers that offer rigorous testing of residual CML disease. Patients must also be compliant to frequent monitoring (PCR studies every 1–2 months for the first 6 months, then every 2 months until 2 years and every 3–6 months thereafter).

ALLOGENEIC STEM CELL TRANSPLANT

Allogeneic SCT, a curative modality in CML, is associated with long-term survival rates of 40–60% when implemented in the chronic phase. It is associated with early (1-year) mortality rates of 5–30%. Although the 5- to 10-year survival rates were reported to be around 50–60% (and considered as cure rates), about 10–15% of patients die in the subsequent 1–2 decades from subtle long-term complications of the transplant (rather than from CML relapse). These are related to chronic graft-versus-host disease (GVHD), organ dysfunction, development of second cancers, and hazard ratios for mortality higher than in the normal population. Other significant morbidities include infertility, chronic immune-mediated complications, cataracts, hip necrosis, and other morbidities affecting quality of life. The cure and early mortality rates in chronic-phase CML are also associated with several factors: patient age, duration of chronic phase, whether the donor is related or unrelated, degree of matching, preparative regimen, and others. In accelerated-phase CML, the cure rates with allogeneic SCT are 20–40%, depending on the definition of accelerated disease. Patients with clonal evolution as the only criterion have cure rates of up to 40–50%. Patients undergoing allogeneic SCT in second chronic phase have cure rates of 40–50%. The cure rates with allogeneic SCT in blastic phase CML are ≤15%. Post–allogeneic SCT strategies are now implemented in the setting of molecular or cytogenetic relapse or in hematologic relapse/transformation. These include the use of TKIs for prevention or treatment of relapse, donor lymphocyte infusions, and second allogeneic SCTs, among others. TKIs appear to be highly successful at reinducing cytogenetic/molecular remissions in the setting of cytogenetic or molecular relapse after allogeneic SCT.

Choice and Timing of Allogeneic SCT

Allogeneic SCT was considered first-line CML therapy before 2000. The maturing positive experience with TKIs has now relegated its use to after first-line TKI failures. An important question is the optimal timing and sequence of TKIs and allogeneic SCT (whether allogeneic SCT should be used as second- or third-line therapy). Among patients who present with or evolve to blastic phase, combinations of chemotherapy and TKIs should be used to induce remission, followed by allogeneic SCT as soon as possible. The same applies to patients who evolve from chronic to accelerated phase. Patients with de novo accelerated-phase CML may do well with long-term TKI therapy (estimated 8-year survival rate 75%); the timing of allogeneic SCT depends on their optimal response to TKI (achievement of complete cytogenetic response). Among patients who relapse in chronic phase, the treatment sequence depends on several factors: (1) patient age and

availability of appropriate donors; (2) risk of allogeneic SCT; (3) presence or absence of clonal evolution and mutations; (4) patient's prior history and comorbidities; and (5) patient and physician preferences **(Table 101-3)**. Patients with T315I mutations at relapse should be offered ponatinib and considered for allogeneic SCT particularly in blastic phase and perhaps also in accelerated phase (because of the short follow-up with ponatinib). Patients with mutations involving Y253H, E255K/V, and F359V/C/I respond better to dasatinib or bosutinib. Patients with mutations involving V299L, T315A, and F317L/F/I/C respond better to nilotinib. Comorbidities such as diabetes, hypertension, pulmonary hypertension, chronic lung disease, cardiac conditions, and pancreatitis may influence the choice for or against a particular TKI. Patients with clonal evolution, unfavorable mutations, or lack of major/complete cytogenetic response within 1 year of salvage TKI therapy have short remission durations and should consider allogeneic SCT as more urgent in the setting of salvage. Patients without clonal evolution or mutations at relapse and who achieve a complete cytogenetic response with TKI salvage, have long-lasting complete remissions and may delay the option of allogeneic SCT to third-line therapy. Finally, older patients (age 65–70 years or older) and those with high risk of mortality with allogeneic SCT may forgo this curative option for several years of disease control in chronic phase with or without cytogenetic response (Table 101-3). In emerging nations, where generic imatinib is now available at the annual price of $400–3000, frontline imatinib is a cost-effective therapy. However, second-line therapy with allogeneic SCT, a one-time curative option with a cost of $20,000–100,000, may be considered (in preference to second-generation TKIs—annual cost above $40,000–100,000) as a more cost-effective national health-care strategy in CML. Table 101-3 summarizes a general guidance to the choice of TKIs versus allogeneic SCT.

MONITORING THERAPY IN CML

Achievement of complete cytogenetic response by 12 months of imatinib therapy and its persistence later, the only consistent prognostic factor associated with survival, is now the main therapeutic endpoint in CML. Failure to achieve a complete cytogenetic response by 12 months or occurrence of later cytogenetic or hematologic relapse are considered as treatment failure and an indication to change therapy. Because salvage therapy with other TKIs re-establishes good outcome, it is important to ensure patient compliance to continued TKI therapy and change therapy when cytogenetic relapse is confirmed unless this is related to non-adherence. Patients on frontline imatinib therapy should be closely monitored until documentation of complete cytogenetic response, at which time they can be monitored every 6 months with peripheral blood FISH and PCR studies (to check for concordance of results), or more frequently if there are concerns about changes in *BCR-ABL1* transcripts (e.g., every 3 months). Monitoring by PCR only is reasonable in patients who are in major molecular response. Cytogenetic relapse on imatinib is an indication of treatment failure and need to change TKI therapy. Mutational analysis in this instance helps in the selection of the next TKI and identifies mutations in 30–50% of patients. Mutational studies by standard Sanger sequencing (which is the technique currently available in most clinical laboratories) in patients in complete cytogenetic response (in whom there may be concerns of increasing *BCR-ABL1* transcripts) identify mutations in ≤5% and are therefore not indicated. Earlier response has been identified as a prognostic factor for long-term outcome, including achievement of partial cytogenetic response (*BCR-ABL1* transcripts ≤10%) by 3–6 months of therapy. Failure to achieve such a response has been associated with significantly worse survival.

The use of second-generation TKIs (nilotinib, dasatinib) as frontline therapy changed the monitoring approach slightly. Patients are expected to achieve major cytogenetic response (or *BCR-ABL1* transcripts ≤10%) by 3–6 months of therapy. Failure to do so is associated with worse event-free survival, transformation rates, and survival. However, the 3- to 5-year estimated survival among such patients is still high, around 80–90%, which is better than what would be anticipated if such patients were offered allogeneic SCT at that time. Changes of therapy for patients with "slow" response have not been proven to be of long-term benefit compared to changes when more obvious signs of resistance appear. Thus, this adverse response to therapy is considered a warning signal, but it is not known whether changing therapy to other TKIs at that time would improve longer-term outcome.

TREATMENT OF ACCELERATED AND BLASTIC PHASES

Patients in accelerated or blastic phase may receive therapy with TKIs, preferably second- or third-generation TKIs (dasatinib, nilotinib, bosutinib, ponatinib), alone or in combination with chemotherapy, to reduce the CML burden, before undergoing allogeneic SCT. Response rates (major hematologic) with single-agent TKIs range from 30 to 50% in accelerated phase and from 20 to 30% in blastic phase. Cytogenetic responses, particularly complete cytogenetic responses, are uncommon (10–30%) and transient in blastic phase. Studies of TKIs in combination with chemotherapy are ongoing; the general experience suggests that combined TKI-chemotherapy strategies increase the response rates and their durability and improve survival. This is particularly true in CML lymphoid blastic phase, where the combination of anti-ALL chemotherapy with TKIs results in complete response rates of 60–70% and median survival times of 2–3 years (compared with historical response rates of 40–50% and median survival times of 12–18 months). This allows many patients to undergo allogeneic SCT in a state of minimal CML burden or second chronic phase, which are associated with higher probability of long-term survival. In CML nonlymphoid blastic phase, anti-AML chemotherapy combined with TKIs results in CR rates of 30–50% and median survival times of 9–12 months (compared with historical response rates of 20–30% and median survival times of 3–5 months). In accelerated phase, response to single TKIs is significant in conditions where "softer" accelerated phase criteria are considered (e.g., clonal evolution alone, thrombocytosis alone, significant splenomegaly or resistance to hydroxyurea, but without evidence of high blast and basophil percentages). In accelerated phase, combinations frequently include TKIs with low-intensity chemotherapy such as low-dose cytarabine, low-dose idarubicin, decitabine, interferon α, hydroxyurea, or others.

OTHER TREATMENTS AND SPECIAL THERAPEUTIC CONSIDERATIONS

Interferon α

Interferon α is considered in combination with TKIs (an investigational approach), sometimes after CML failure on TKIs, occasionally in patients during pregnancy, or as part of investigational strategies with TKIs to eradicate residual molecular disease.

Chemotherapeutic Agents

Hydroxyurea remains a safe and effective agent (at daily doses of 0.5–10 g) to reduce initial CML burden, as a temporary measure in between definitive therapies, or in combination with TKIs to sustain complete hematologic or cytogenetic responses. Busulfan is often used in allogeneic SCT preparative regimens. Because of its side effects (delayed myelosuppression, Addison-like disease, pulmonary and cardiac fibrosis, myelofibrosis), it is now only rarely used in the chronic management of CML. Low-dose cytarabine, decitabine, anthracyclines, 6-mercaptopurine, 6-thioguanine, thiotepa, anagrelide, and other agents are sometimes useful in different CML settings to control the disease burden.

Others

Splenectomy is now seldom considered to alleviate symptoms of massive splenomegaly and/or hypersplenism. Splenic irradiation is rarely used, if at all, because of the postirradiation adhesions and complications. Leukapheresis is occasionally used in patients presenting with extreme leukocytosis and leukostatic complications. Single doses of high-dose cytarabine or high doses of hydroxyurea, with tumor lysis management, may be as effective and less cumbersome.

Special Considerations

Women with CML who become pregnant should discontinue TKI therapy immediately. Among 125 babies delivered to women with CML who discontinued imatinib therapy as soon as the pregnancy was known, three babies were born with neurologic, skeletal, and renal malformations, suggesting the teratogenicity of imatinib known from animal studies. A similar experience has been reported with dasatinib, the incidence of malformations was reported to be higher, 10–12%. There are no or little data with other TKIs. Control of CML during pregnancy can be managed with leukapheresis for severe symptomatic leukocytosis in the first trimester and with hydroxyurea subsequently until delivery. There are case reports of

successful pregnancies and deliveries of normal babies with interferon α therapy and registry studies in essential thrombocytosis of its safety, but interferon α can be antiangiogenic and may increase the risk of spontaneous abortions.

Approximately 10–15% of patients on TKI therapy may develop chromosomal abnormalities in the Ph-negative cells. These may involve loss of chromosome Y, trisomy 8, 20q–, chromosome 5 or 7 abnormalities, and others. Most chromosomal abnormalities disappear spontaneously and may be indicative of the genetic instability of the hematopoietic stem cells that predisposes the patient to develop CML in the first place. Rarely (in <1% of instances), abnormalities involving chromosomes 5 or 7 may be truly clonal and evolve into myelodysplastic syndrome or acute myeloid leukemia. This is thought to be part of the natural course of patients in whom CML was suppressed and who live long enough to develop other hematologic malignancies.

TABLE 101-2

Medical Therapeutic Options in Chronic Myeloid Leukemia

Agent (Brand Name)	Approved Indications	Dose Schedule	Notable Toxicities
Imatinib mesylate (Gleevec)	All phases	400 mg daily	See text
Dasatinib (Sprycel)	All phases	First-line: 100 mg daily Salvage: 100 mg daily in chronic phase; 140 mg daily in transformation	Myelosuppression; pleural and pericardial effusions; pulmonary hypertension
Nilotinib (Tasigna)	All phases except blastic phase	First-line: 300 mg twice daily Salvage: 400 mg twice daily	Diabetes; arterio-occlusive disease; pancreatitis
Bosutinib (Bosulif)	All phases except frontline	500 mg daily	Diarrhea, liver toxicity
Ponatinib (Iclusig)	Optimal TKI if T315I mutation Failure of ≥2 tyrosine kinase inhibitors	45 mg daily (may consider lower starting doses in the future, e.g., 30 mg daily)	Skin rashes (10–20%); pancreatitis (5%); arterio-occlusive disease (10–20%); systemic hypertension (10–15%)
Omacetaxine mepesuccinate (Synribo)	Failure ≥2 tyrosine kinase inhibitors	1.25 mg/m^2 subcutaneously twice daily for 14 days of induction; 7 days of maintenance every month (may consider shorter dose schedules, 7 days of induction, 2–5 days of maintenance)	Myelosuppression

TABLE 101-3

General Suggestions Regarding the Use of Tyrosine Kinase Inhibitors (TKIs) and Allogeneic Stem Cell Transplantation (SCT) in Chronic Myeloid Leukemia (CML)

CML Phase	Use of TKI	Consideration of Allogeneic SCT
Accelerated or blastic	Interim therapy to achieve minimal CML burden	As soon as possible (exception: de novo accelerated phase)
T315I mutation	Ponatinib to achieve minimal CML burden	Depends on longer term follow-up results of ponatinib efficacy
Imatinib failure in chronic phase; no clonal evolution, no mutations, good initial response; no T315I	Second-line tyrosine kinase inhibitors long-term	Third-line after second-line TKI failures
Clonal evolution or mutations, or no cytogenetic response to second-line TKI	Interim therapy with alternative second generation TKI or ponatinib to achieve minimal CML burden	Second-line
Older patients (≥65–70 years) after imatinib failure in chronic phase	Salvage TKIs as longer-term therapy	May forgo allogeneic SCT in favor of good quality of life and survival in chronic phase
Imatinib failure; emerging nation	–	Second-line: curative, one-time cost $20,000–100,000 (versus >$40,000–100,000/year with TKI)

Note: Mutations involving Y253H, E255K/V, or F359V/C/I: prefer dasatinib or bosutinib. Mutations involving V299L, T315A, or F317L/F/I/C: prefer nilotinib.

GLOBAL ASPECTS OF CML

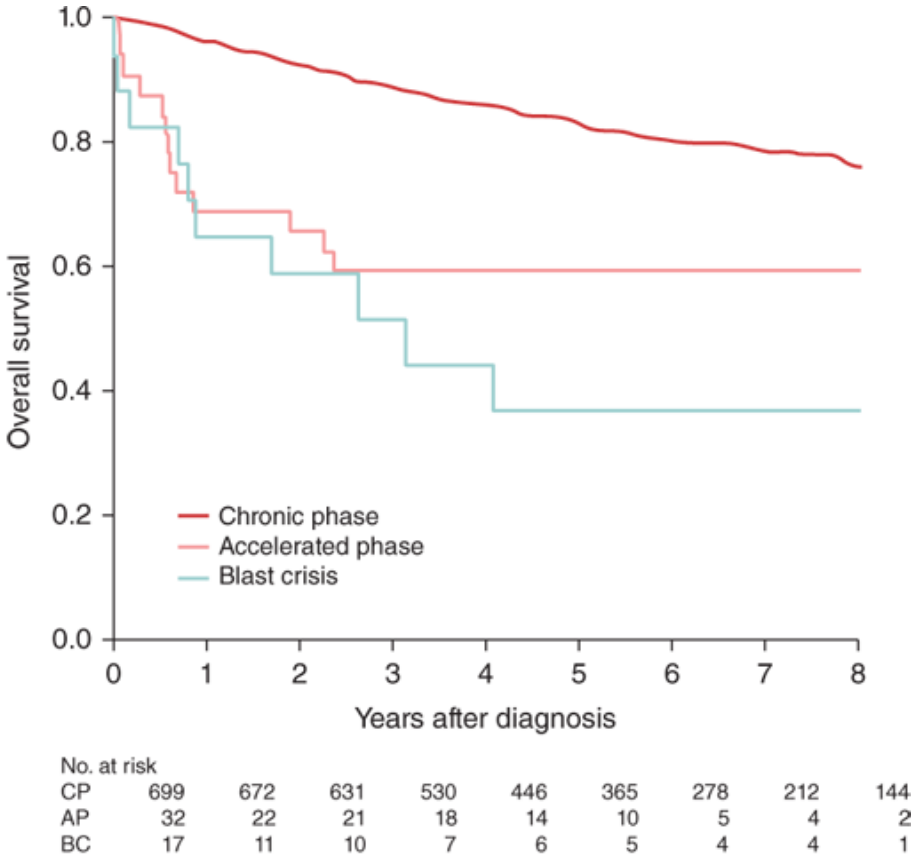
Routine physical examinations and blood tests in the United States and advanced countries result in early detection of CML in most patients. About 50–70% of patients with CML are diagnosed incidentally, and high-risk CML as defined by prognostic models (e.g., Sokal risk groups) is found in only 10% of patients. This is not the same situation in emerging nations where most patients are diagnosed following evaluation for symptoms and many present with high tumor burden, such as massive splenomegaly, and advanced phases of CML (high-risk CML documented in 20–30%). Therefore, the prognosis of such patients on TKI therapy may be worse than the published experience.

The high cost of TKI therapies (annual costs of $90,000–140,000 in the United States; lower but variable in the rest of the world) makes the general affordability of such treatments difficult. Although TKI treatment penetration is high in nations where cost of therapy is not an issue (e.g., Sweden, European Union), it may be less so in other nations, even in advanced ones like the United States, where out-of-pocket expenses may be prohibitive to a subset of patients. Based on the sales of imatinib worldwide and charity-free drug supplies, it is estimated that <30% of patients are treated with imatinib (or other

TKIs) consistently. Although the estimated 10-year survival in CML is 85% in single-institution studies (e.g., MD Anderson Cancer Center), in national studies in countries with TKI affordability (Sweden) (**Figs. 101-2** and **101-3**) or in company-sponsored studies (where all patients have access to TKIs throughout their care), the estimated 10-year survival worldwide, even 16 years after the introduction of TKI therapies, is likely to be <50%. The Surveillance, Epidemiology, and End Results (SEER) data from the United States report an estimated 5-year survival rate of 60% in the era of TKIs.

FIGURE 101-3

Survival in chronic (CP), accelerated (AP), and blastic crisis (BC) phases of chronic myeloid leukemia (CML) in the population-based Swedish national registry study. The accelerated- and blastic-phase cases are de novo presentations. The favorable outcome with de novo blastic phase may be due to use of 20% blasts or more to define blastic phase. *(With permission from Dr. Martin Hoglund, Swedish CML Registry, 2013.)*



No. at risk

CP	699	672	631	530	446	365	278	212	144
AP	32	22	21	18	14	10	5	4	2
BC	17	11	10	7	6	5	4	4	1

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The current high cost of TKI therapies poses two additional considerations. The first are the treatment pathways and guidelines in nations where TKIs may not be affordable by patients or the health care system. In these conditions, there are trends of pathways advocating allogeneic SCT as frontline or second-line therapy (i.e., after imatinib failure; as a one time cost of $20,000–100,000) despite the associated mortality and morbidities. The second is the choice of frontline TKI therapy once imatinib becomes available in generic forms, hopefully at much lower annual prices, e.g., $2000–10,000 per year (currently $400 per year in India). This will depend on the maturing data in randomized studies of second-generation TKIs versus imatinib in relation to important long-term outcome endpoints, particularly survival, but also event-free survival, transformation-free survival, and treatment-free remission.

FURTHER READING

Baccarani M et al: European LeukemiaNet recommendations for the management of chronic myeloid leukemia. Blood 122:872, 2013.

Cortes J, Kantarjian H: How I treat newly diagnosed chronic phase CML. Blood 120:1390, 2012.

Cortes JE et al: Ponatinib in refractory Philadelphia chromosome-positive leukemias. N Engl J Med 367:2075, 2012.

Cortes JE et al: Final 5-year study results of DASISION: The dasatinib versus imatinib study in treatment-naïve chronic myeloid leukemia patients trial. J Clin Oncol 34:2333, 2016.

Gambacorti-Passerini C et al: Long-term efficacy and safety of bosutinib in patients with advanced leukemia following resistance/intolerance to imatinib and other tyrosine kinase inhibitors. Am J Hematol 90:755, 2015.

Hochhaus A et al: Long-term benefits and risks of frontline nolitinib vs imatinib for chronic myeloid leukemia in chronic phase: 5-year update of the randomized ENESTnd trial. Leukemia 30:1044, 2016.

Hochhaus A et al: Long-term outcomes of imatinib treatment for chronic myeloid leukemia. N Engl J Med 376; 917, 2017. [PubMed: 28273028]

Kantarjian HM et al: Nilotinib is effective in patients with chronic myeloid leukemia in chronic phase after imatinib resistance or intolerance: 24-month follow-up results. Blood 117:1141–1145, 2013.

Rousselot P et al: Loss of major molecular response as a trigger for restarting tyrosine kinase inhibitor therapy in patients with chronic-phase chronic myelogenous leukemia who have stopped imatinib after durable undetectable disease. J Clin Oncol 32:424–430, 2014.

Sasaki K et al: Conditional survival in patients with chronic myeloid leukemia in chronic phase in the era of tyrosine kinase inhibitors. Cancer 122:238–248, 2016.

Shah NP et al: Dasatinib in imatinib-resistant or –intolerant chronic-phase, chronic myeloid leukemia patients: 7-year follow-up of study CA180-034. Am J Hematol 9:869–874, 2016.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 102: Acute Lymphoid Leukemia

Dieter Hoelzer

FIGURE 102-1

INTRODUCTION

In acute lymphoblastic leukemia (ALL), the malignant clone arises from hematopoietic progenitors in the bone marrow or lymphatic system resulting in an increase of immature nonfunctioning leukemic cells. Infiltration of bone marrow leads to anemia, granulocytopenia, and thrombocytopenia with the clinical manifestations of fatigue, weakness, infection, and hemorrhages. These symptoms are more often the reason a patient first seeks medical advice rather than consequences of tumor bulk, such as lymph node enlargement, hepatosplenomegaly caused by leukemic infiltration, or symptoms of the central nervous system (meningeosis leukemica).

The diagnosis of ALL is made by morphology from smears of peripheral blood or bone marrow. The classification includes cytochemistry, immunological markers, cytogenetic, and molecular genetic analysis.

INCIDENCE AND AGE

ALL is the most frequent neoplastic disease in children with an early peak at the age of 3–4 years. The incidence in adults ranges from 0.7 to 1.8/100,000 per year, being somewhat higher in younger adults (1–1.5 for the age group 15–24 years) and decreasing thereafter, only to increase again in elderly people to 2.3 for age >65 years. The frequency of immunological, cytogenetic, and genetic subtypes changes with age.

ETIOLOGY

The etiology of acute leukemias is unknown. There are, however, internal and external factors that influence the incidence of leukemia. In ALL, inheritance of certain diseases and exposure to ionizing radiation or to chemicals, including prior chemotherapy, are associated with an increased risk of developing leukemia, but less than in acute myeloid leukemia (AML).

CONGENITAL DISORDERS

Patients with some rare congenital chromosomal abnormalities have a higher risk of development of acute leukemia; e.g., Klinefelter's syndrome, Fanconi's anemia, Bloom's syndrome, ataxia telangiectasia, and neurofibromatosis. There is a twentyfold increased incidence of leukemia in patients with Down syndrome, in whom ALL is increased in childhood or AML at an older age. Genetic predisposition may play a part in acute leukemia even, when not associated with another inherited disease, as the identical twin of a leukemic child has a fivefold risk of developing acute leukemia.

INFECTIOUS AGENTS

Human T-cell leukemia virus I (HTLV-I), endemic in Japan and the Caribbean, is the etiological agent for adult T-cell leukemia/lymphoma, an aggressive adult T-cell leukemia **(see Chap. 196)**. In the endemic African type of Burkitt's lymphoma, the Epstein-Barr virus, a DNA virus of the herpes family, has been implicated as a potential causative agent.

DIAGNOSIS AND CLASSIFICATION

The diagnosis of acute leukemia is made by examination of the peripheral blood and bone marrow. Classification of the patient's disease also requires cytochemical stains, assessment of expression of immunological markers, cytogenetic analysis, and molecular markers. The

major aim of classification is to distinguish between AML and ALL because of the different treatment approaches and drug sensitivities. The immunological markers are the major criteria to subdivide ALL into B-cell lineage or T-cell lineage (T-ALL) leukemias. Cytogenetic and molecular evaluation provide further identification of ALL subgroups.

PERIPHERAL BLOOD

Peripheral blood counts and a differential count from a Wright-Giemsa-stained blood smear are essential at the time of presentation. The white blood cell count in about 40% of ALL patients is reduced or normal (**Table 102-1**). Thus, in the frequently used automatic blood cell counter, the disease may not be detected. One-third of the patients have a moderately increased initial white blood cell count, between 10×10^9 and 50×10^9/L. Leukemic blast cells (LBC) in the peripheral blood are largely responsible for the rise in white blood cell count, but it is noteworthy that in 8% of the ALL patients, no circulating leukemic blast cells are observed.

TABLE 102-1

Laboratory Values at Diagnosis of Acute Lymphoblastic Leukemia

		ALL
n		**1273***
Initial white blood cell count ($\times 10^9$/L)	<10 10–50 >50	41% 31% 28%
Neutrophils	<50–100 <100,000	12% 16%
Platelets ($\times 10^9$/L)	<20 21–40 41–100 >100	22% 22% 29% 27%
Hemoglobin (g/dL)	<7 7–9 >9	20% 33% 47%
Leukemic blasts in PB	0% 25–75% >75%	8% 34% 36%
Leukemic blasts in BM	<50% 51–90% >90%	4% 25% 71%

Source: Data from three consecutive German Multicenter Trials for Adult ALL (GMALL).

Peripheral blood observation shows characteristic anemia, thrombocytopenia, and neutropenia. The reduction in the level of hemoglobin is usually mild to moderate, but nearly one-third of the patients have hemoglobin levels <7–8 g/dL. A platelet count below the critical number of 20×10^9/L is seen in one-fifth of the ALL patients. The proportion of patients with granulocyte count $<0.5 \times 10^9$/L usually associated with high risk of infection was only one-fifth in adult ALL series.

BONE MARROW EXAMINATION

Bone marrow aspirates are important for immunological, cytogenetic, and genomic markers. Direct smears from the bone marrow are essential to confirm the diagnosis of acute leukemia and to distinguish between AML and ALL. The bone marrow is usually heavily packed with leukemic blast cells comprising >90% of the nucleated cells in ~70% of patients. The normal hemopoietic elements are greatly reduced or absent. A biopsy of the bone marrow will further demonstrate marked hypercellularity with replacement of fat spaces and normal elements by infiltration with leukemic cells.

LUMBAR PUNCTURE

The examination of the cerebrospinal fluid is an essential routine diagnostic measure for ALL. There are different opinions as to when the first lumbar puncture should be done. One procedure is to delay the examination until remission is achieved in order to avoid seeding the central nervous system (CNS) by circulating leukemic blast cells from the peripheral blood. On the other hand, early recognition of CNS disease will lead to immediate CNS-specific therapy, which is required for such patients. Thus, other clinicians prefer to perform the lumbar puncture before treatment starts. This procedure is restricted to patients with an adequate platelet count (>20×10^9/L), an absence of manifest clinical hemorrhage, and without a high white blood cell count. For safety reasons, all patients should receive intrathecal methotrexate at the first lumbar puncture.

MORPHOLOGICAL SUBTYPES IN ALL

Three morphologic subgroups of acute lymphoblastic leukemia are distinguished by the French-American-British classification. Whereas the distinction between L1 and L2 morphology has no clinical consequences, the detection of L3 ALL is of clinical and prognostic relevance. It is observed in up to 5% of adult patients and should be distinguished as it is indicative of mature B-ALL, usually termed Burkitt's leukemia, with distinct treatment options. A surface marker confirmation should be obtained.

IMMUNOLOGICAL SUBTYPES OF ALL

A series of monoclonal antibodies is employed to identify antigens expressed on the surface of normal or leukemic cells. The main aim of the immunological classification is to subdivide ALLs according to the presence or absence of B-cell or T-cell markers, or B-phenotypic/hybrid acute leukemia. A marker is positive if >20% of the cells are stained with the monoclonal antibody (Table 102-2).

TABLE 102-2

Immunological, Cytogenetic, Molecular, and Clinical Characteristics of Adult ALL

Subtypes	Marker	Incidence	Frequent Cytogenetic Aberrations	Fusion Transcripts and Mutations	Clinical Characteristics	Relapse Kinetics and Localization
B-lineage ALL	HLA-DR+, TdT+, CD19+ and/or CD79a+ and/or CD22+	76%				
Pro B-ALL	No additional differentiation markers, Freq. myeloid coexpression (>50%)	12%	t(4;11) (q21;q23)	70% ALL1-AF4 (20% Flt3 in MLL+)	High WBC (>100.000/mL) (26%)	Mainly BM (>90%)
Common ALL	CD10	49%	t(9;22)(q34;q11) del(6q)	33% BCR–ABL (30–50% in c/preB)	Higher age >50 years (24%)	Mainly BM (>90%) Prolonged relapse kinetics (up to 5–7 years)
Pre-B-ALL	CD10±, cyIg+	11%	t(9;22)(q34;q11) t(1;19)(q23;p13)	4% t(1;19)/ PBX-E2A		
Mature B-ALL	CD10 ±, sIg+	4%	t(8;14)(q24;q32) t(2;8)(p12;q24) t(8;22)(q24;q11)		Higher age >55 years (27%) Frequent organ involvement (32%) and CNS-involvement (13%)	Frequent CNS (10%) Short relapse kinetics (up to 1–1.5 years)
T-lineage ALL	cyCD3 or sCD3	24%			Younger age (90% <50 years) Frequent mediastinal tumors (60%) Frequent CNS involvement (8%) High WBC (>50/mL) (46%)	Frequent CNS (10%)/ extramedullary (6%) Intermediate relapse kinetics (up to 3–4 years)

Subtypes	Marker	Incidence	Frequent Cytogenetic Aberrations	Fusion Transcripts and Mutations	Clinical Characteristics	Relapse Kinetics and Localization
Early Pro/Pre T-ALL Cortical T-ALL Mature T-ALL	No additional differentiation markers, mostly CD2-CD1a+, sCD3± sCD3+, CD1a-	6% 12% 6%	t(10;14)(q24;q11) t(11;14)(p13;q11)	5% HOX11-TCR <5% LMO/TCR 2% SIL-TAL1 In T-ALL 4% NUP213-ABL1 33% HOX11b 5% HOX11L2b 50% Notch1b		

B-Cell Lineage

More than 70% of adult ALLs are of B-cell origin, and the most frequent immunological subtype, common ALL, is characterized by the presence of ALL antigen, a glycoprotein (gp100/CD10). Common ALL blast cells do not carry markers of mature B cells such as cytoplasmic immunoglobulins or surface membrane immunoglobulins. Pre-B-ALL (early B-ALL) is characterized by the expression of cytoplasmic immunoglobulin, being negative in common ALL but is otherwise identical with all other cell markers. Very rarely, the common ALL antigen may be absent in this subtype. Mature B-ALL comprises about 3–4% of adult ALL patients. The blast cells express surface antigens of mature B cells, including the sIgM. Common ALL antigens may also be present and also occasionally cytoplasmic immunoglobulin. Pro-B-ALL (also termed *early B-precursor ALL*) is a leukemia that was formerly termed *non-T*, *non-B-ALL*, or *null-ALL* as neither T-cell nor B-cell features could be demonstrated. This subtype is Tdt (terminal) deoxynucleotidyl transferase, and CD19 positive and forms about 11% of adult ALL.

T-Cell Lineage

Approximately 25% of adult ALL belongs to the T-cell lineage. All cases express the T-cell antigen (gp40, CD7) and cytoplasmatic or surface CD3. They may, according to their step of T-cell differentiation, express other T-cell antigens, e.g., the E-rosette receptor (CD2) and/or the cortical thymocyte antigen T6 (CD1). A minority of T-ALL blast cells may also express common ALL antigen together with other T-cell antigens. According to these markers, it is possible to distinguish a pro-T-ALL (also termed *early T-precursor ALL*), cortical, or thymic T-ALL and a mature T-ALL expressing different stages of differentiation.

Biphenotypic or Mixed Leukemias

Biphenotypic leukemias are defined as those in which markers of lymphoid and myeloid lineages are coexpressed on the same leukemic cells without the typical phenotype of either ALL or AML. Bilineage leukemias are those with two populations of blast cells with either lymphoid or myeloid antigens. They must be differentiated from AML by coexpression of lymphoid markers and from ALL by coexpression of myeloid markers.

CYTOGENETIC AND MOLECULAR ANALYSIS

Cytogenetic analysis should be performed in all cases of acute leukemia. The demonstration of a specific karyotype may be required for the confirmation of the diagnosis, but chromosome abnormalities may be also important independent prognostic variables for disease-free survival or may lead to a specific targeted therapy.

DIAGNOSTICS

The diagnostic techniques are standard cytogenetics, fluorescence in situ hybridization, and reverse transcriptase polymerase chain reaction. These methods allow the detection of Ph+ ALL, with the chromosomal translocation t(9;22)(q34;q11), and the detection of the corresponding *BCR-ABL1* gene rearrangement. Further ALL entities that have been identified are t(4;11)(q21;q23)/*MLL-AFA4*, abn11q23/*MLL*, and t(1;19) (q23;p13)/*PBX-E2A*.

Gene expression profiling, single nucleotide polymorphism array analysis, array-comparative genomic hybridization, and next generation sequencing recognize newly defined ALL entities with poor prognosis: Ph-like ALL and early T-precursor ALL.

NEW ENTITIES

Ph-Like

ALL, also called *BCR-ABL1*-like ALL, is characterized by genetic lesions similar to Ph+ ALL, associated with *IKZF1* deletion, *CLRF2* overexpression, and tyrosine kinase activating rearrangements involving *ABL1*, *JAK2*, *PDGFRB*, and several others. The frequency is 10% in children and 25–30% in young adults, but does not increase further with age. Treatment could be directed at the underlying genetic pattern with BCR-ABL inhibitors (e.g., dasatinib) or JAK2 inhibitors (e.g., ruxolitinib).

Early T-Precursor ALL

(ETP-ALL) is characterized by lack of CD1a and CD8, weak CD5 expression, at least one myeloid/stem cell marker, a specific transcriptional profile and the possible involvement of several critical genes. No new treatment approaches are currently available for this subtype, and thus hematopoietic stem cell transplantation in first complete remission is the preferred option.

MINIMAL RESIDUAL DISEASE (MRD)

MRD is the detection of residual leukemic cells, not recognizable by light microscopy.

Methods for determining MRD are based on the detection of leukemia-specific aberrant immunophenotypes by flow cytometry, the evaluation of leukemia-specific rearranged immunoglobulin or T-cell receptor sequences by real-time quantitative polymerase chain reaction, or the detection of fusion genes associated with chromosomal abnormalities (e.g., BCR-ABL, MLL-AF4). The detection limit with these methods is 10^{-3}–10^{-5} (0.1–0.001%). The phenotypic aberrations are unique to each patient with ALL and can be detected in up to 95% of individuals. Methods for MRD evaluation and standardization of MRD quantification have been extensively described.

MRD RESPONSE AND TERMINOLOGY

Molecular response can be evaluated only for patients in complete cytological remission, with one marker or more for MRD analysis and samples available at diagnosis and followed at specific time points during the course of disease. Results are classified as presented in **Table 102-3**.

TABLE 102-3

Response Parameters According to MRD

Terminology	Definition
Complete (hematologic) remission	Leukemic cells not detectable by light microscopy (<5% blast cells in bone marrow)
Complete molecular remission MRD-negativity	Patient in complete remission, MRD not detectable, ≤0.01% = ≤1 leukemia blast cell in 10,000
Molecular failure/MRD-positivity	Patient in complete hematologic remission but not in molecular complete remission >0.01%
Molecular relapse/MRD-positivity	Patient still in complete remission had prior molecular complete remission Leukemic blast cells detectable in bone marrow not detectable (<5%)
Hematologic relapse	>5% ALL cells in bone marrow/blood

MOLECULAR RESPONSE AFTER INDUCTION THERAPY AND IMPACT ON OUTCOME

Achievement of molecular complete response/molecular remission is the most relevant independent prognostic factor for disease-free survival and overall survival. Patients with molecular complete remission after induction therapy had significantly superior outcome in several studies, with a disease-free survival of 54–74%, compared to 17–40% for MRD-positive patients. Patients with molecular failure after induction therapy should proceed to a targeted therapy to reduce the tumor load, to be followed by immediate allogeneic hematopoietic stem cell transplant.

PROGNOSTIC FACTORS, RISK STRATIFICATION, AND MRD

The aim of identification of prognostic parameters at diagnosis, which include age, white blood cell count, specific immunophenotypes, and cytogenetic and genetic aberrations, is to stratify patients into risk groups: standard-risk patients without any poor-risk factors, with a good chance of cure by chemotherapy, and high-risk patients with one or more of those risk factors. High-risk patients are most often candidates for a stem cell transplant in first complete remission.

WILL MINIMAL RISK DISEASE EVALUATION REPLACE PRETHERAPEUTIC RISK FACTORS?

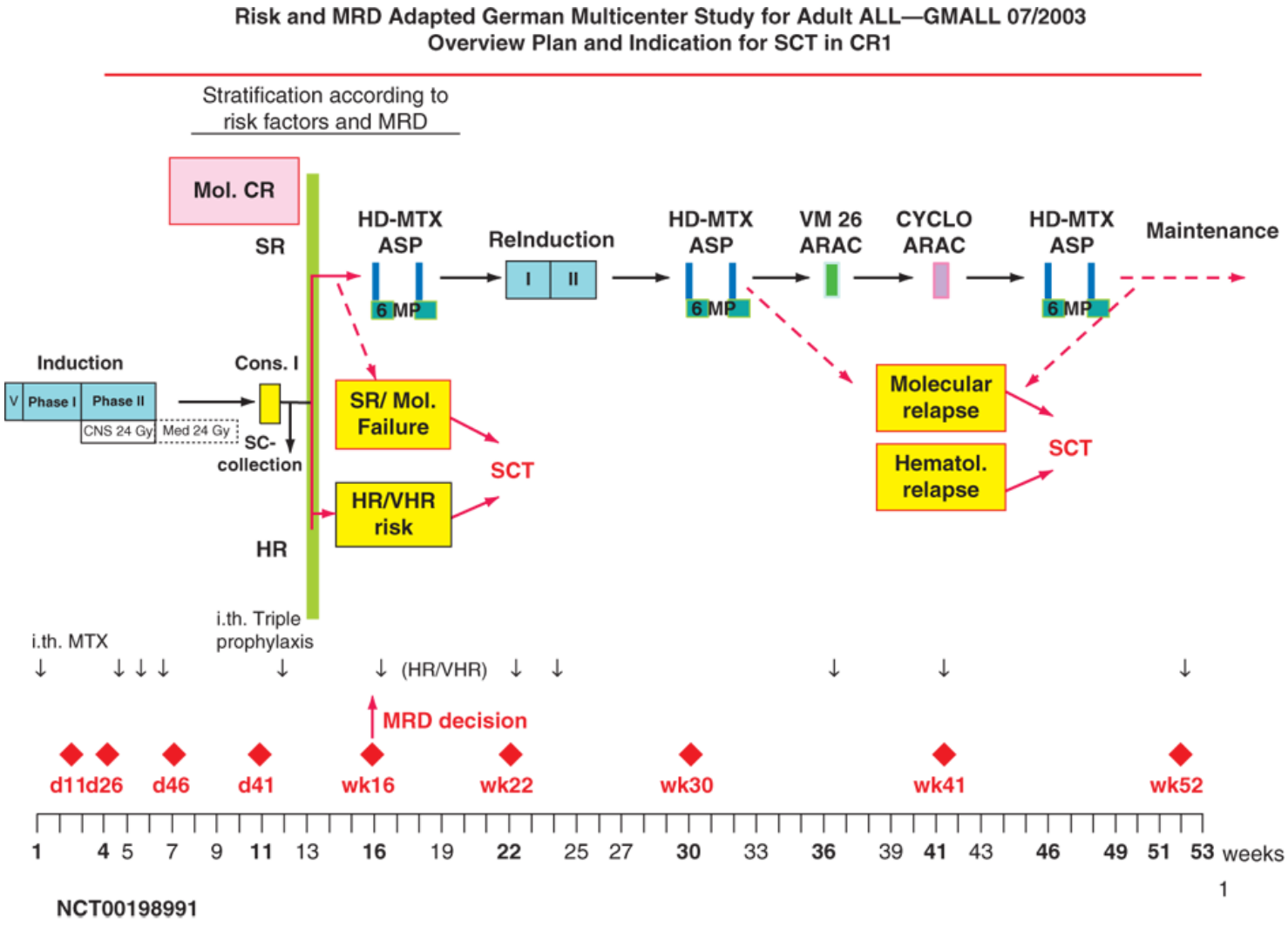
The question arises as to whether the evaluation of MRD overcomes all of those pretherapeutic risk factors, or whether they should be combined. A practical approach is to enter the conventional prognostic factors and MRD into a decision algorithm. Thereby defined standard-risk patients who are highly likely to achieve molecular remission (about 90–95%) will remain as standard-risk patients, whereas those who are MRD-positive will be defined accordingly as high-risk patients. Clinically defined high-risk patients are potential candidates for a stem cell transplant in first complete remission. However, it is not clear how to proceed if they achieve a complete molecular remission, since some studies suggest a lack of benefit from transplant in those who are MRD negative. If MRD information is not available, the risk stratification should rely on clinical risk factors evaluated at diagnosis. Unfortunately, 20–30% of adult ALL patients who are MRD-negative after induction will relapse. Potential reasons include loss of sensitivity, evolution of leukemic subclones, and extramedullary origin of disease.

TREATMENT PRINCIPLES

The goal of induction therapy is the achievement of a complete remission, or even better, a molecular complete remission, mostly evaluated within 6–16 weeks of starting chemotherapy. With current regimens (see Fig. 102-1), the complete remission rate has increased to 80–90%, being higher for standard-risk patients (≥90%), and lower for high-risk patients (~80%). The outcome of ALL is strictly related to the age of the patient, and treatment protocols considering the age of an individual patient have emerged.

FIGURE 102-1

An example treatment schema for adult ALL. A variety of effective treatment programs are available. Each generally includes an induction phase followed by a consolidation phase and concludes with a maintenance phase. 6-MP, 6-mercaptopurine; arac, cytosine arabinoside, or cytarabine; ASP, L-asparaginase; cyclo, cyclophosphamide; HD-MTX, systemic high-dose methotrexate; HR, high risk; i.th. MTX, intrathecal methotrexate; SCT, stem cell transplantation; SR, standard risk; VM26, teniposide.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT IN ADULT PATIENTS WITH ACUTE LYMPHOBLASTIC LEUKEMIA

Pediatric-Inspired Therapies

Pediatric-inspired therapies for adolescents and young adults provide increased drug intensity at several stages of treatment, including larger cumulative doses of drugs such as glucocorticoids, vincristine, L-asparaginase, and consequent central nervous system-directed therapy, which should be strictly adhered to, thereby reducing the role of stem cell transplant in such cases. In a 2012 meta-analysis of 11 trials, including 2489 adolescents and young adults, pediatric-inspired regimens were superior to conventional adult chemotherapy. However, none of the trials was a randomized comparison. Table 102-4 gives the outcome of recent studies with pediatric-inspired regimens for adolescents and young adults with a median age of 27 years. The complete remission rate was very high with 93% (85–98%), and the overall survival rate of 70% (60–78%) was very encouraging. Survival rates at ≥5 years were 70% (67–78%) compared to 34–41% with the former protocols.

TABLE 102-4

Outcome of Adult ALL

Treatment Approaches	Time Period	Number Studies	Number Patients	Age* Years	CR* Rate	Early* Death	Overall* Survival
Pediatric-inspired for adult ALL	2008–2015	6	832	**27** (15–60)	**93%** (85–98%)	**5%** (1–7%)	**70%** (60–78%)
Adult protocols	1998–2016	20	7961	**32.7** (12–92)	**84%** (74–93%)	**7%** (1–10%)	**36%** (27–60%)
Elderly protocols	1996–2016			**62**			
Palliative		4	94	60–91	**43%**	**24%**	7 months
Intensive		12	519	60–92	**56%**	**23%**	**14%**
Age-specific		11	653	55–85	**73%**	**13%**	**42%**

*Weighted mean and range.

Adult ALL

The treatment results for adult ALL patients have moderately improved (Table 102-4). The overall survival is 36% with a wide variation from 27 to 60% due to differences in the intensity of the chemotherapy regimen and the outcome of stem cell transplantation.

In several current multicenter prospective trials, the overall survival rate for standard-risk adult ALL patients is now 50–70% with chemotherapy alone. Overall survival for high-risk patients increased from 20–30% to >50% when they received an allogeneic stem cell transplant in first complete remission.

Elderly ALL

Since palliative treatments or intensive chemotherapy regimens have failed, with either low complete remission rates or high early death rates, short elderly specific ALL protocols have been initiated, with less intensive therapy (avoiding anthracyclines and alkylating agents). With these protocols, the complete remission rate was increased to 73%, early death could be reduced to 13% (0–36%), and overall survival was 42%.

MAINTENANCE

Maintenance therapy usually consists of 6-mercaptopurine and methotrexate, a strategy transferred from childhood ALL. The duration of maintenance therapy is between 2 and 2.5 years. The potential effect of further intensification cycles of maintenance therapy remains unclear. In a large multicenter Italian study after intensive consolidation treatment, patients were randomly assigned to postconsolidation therapy with conventional maintenance or to intensified maintenance with additional alternating treatment courses of different intensity. There was no difference in the survival rate at 10 years between the treatment groups, which may suggest that, after adequate induction and consolidation therapy, the intensity of the maintenance therapy has no influence on survival.

Maintenance therapies should be adapted to immunological subtypes of ALL. In mature B-ALL, maintenance is not required. In T-ALL and B-Lineage ALL with relapses up to 2.5 years, maintenance therapy is necessary. In Ph-positive ALL, maintenance should include a BCR/ABL tyrosine kinase inhibitor (TKI), most likely the one that was used during induction and consolidation therapy. It is now also standard to give a TKI after allogeneic stem cell transplant in Ph-positive ALL as a maintenance therapy. The duration of maintenance therapy with a TKI is also 2–2.5 years and may be guided by MRD evaluation.

PROPHYLAXIS OF CENTRAL NERVOUS SYSTEM LEUKEMIA

Without some form of prophylactic CNS-directed therapy, in very early studies without any intensive systemic chemotherapy, around 30% of adults with ALL developed CNS leukemia. Prophylactic CNS therapy in ALL is essential for several reasons: CNS leukemia is more easily prevented than treated; once CNS leukemia has developed, it is generally followed by systemic relapse shortly after; and effective CNS prophylaxis also prevents systemic relapse.

Several treatment options are available for prevention of CNS relapse: intrathecal (i.th.) therapy, cranial radiation therapy (CRT), and systemic high-dose or intrathecal therapy is usually based on methotrexate as single drug, but combinations with cytosine arabinoside and/or glucocorticoids are used in some studies. The route of application is generally lumbar puncture. CRT (18–24 Gy in 12 fractions >16 days) may be administered with or without parallel intrathecal therapy. Systemic high-dose chemotherapy may include methotrexate or cytosine arabinoside since both drugs reach cytotoxic drug levels in the cerebrospinal fluid (CSF) and showed efficacy in overt CNS leukemia. A liposomal preparation of cytosine arabinoside has been introduced for the treatment of CNS in patients with ALL or lymphoma; this formulation is more effective since it has a half-life of >2 weeks, compared to only a few hours for the other used intrathecal drugs.

In Ph-positive ALL, tyrosine kinase inhibitors are now an essential part of the treatment strategy. TKIs are equally effective at crossing the blood-brain barrier; dasatinib and ponatinib do, whereas imatinib and nilotinib do not.

In several studies with combined modalities of CNS prophylaxis, the CNS relapse rate was ≤5%. In trials with early high-dose chemotherapy and intrathecal therapy with or without CNS irradiation, the CNS relapse rate was as low as 2%.

THERAPY OF CNS DISEASE

About 5–10% of adult patients present with manifestations of CNS leukemia. The incidence is correlated with the immunological subtype and is higher in mature B-ALL up to 10–15% and in T-ALL up to 10%. For the treatment of CNS leukemia, the same treatment measures as those used for CNS prophylaxis are employed, either intrathecal MTX alone or in combination with cytosine arabinoside or hydrocortisone. The intrathecal therapy is given 2–3 times per week, continued for >2 or 3 weeks until two consecutive CSF examinations show no evidence of leukemic infiltration. When adult ALL patients with initial CNS leukemia are treated adequately, leukemia-free survival and CNS relapse rate are not inferior to those without CNS involvement.

Relapse in the CNS is difficult to treat. Mostly it occurs synchronously with bone marrow relapse, and if leukemic blasts are not seen morphologically, MRD is positive in nearly all cases. This requires a local as well as a systemic therapy. The outcome after CNS relapse is still dismal, and an allogeneic stem cell transplantation is the most promising option to achieve a remission.

STEM CELL TRANSPLANTATION

Hematopoietic stem cell transplantation is an essential part in the treatment strategy of adult ALL. As a stem cell source, peripheral blood cells are increasingly used compared to bone marrow. Also with regard to the donors, there is a shift from sibling donors to matched unrelated donors or haploidentical transplants from relatives. Indications for stem cell transplantation in first remission are controversial. However, in most studies, it is recommended for patients with persistent MRD and all high-risk patients either defined by conventional clinical prognostic factors or by MRD positivity. High-risk patients have a survival rate of ≥50% if transplanted in first remission. For standard risk patients with sustained molecular remission, allogeneic stem cell transplantation is not recommended in first remission. Autologous stem cell transplantation in first remission is restricted to a few disease entities, e.g., in Ph+ ALL, and should be done only in patients who are MRD negative or older patients.

For all relapsed adult ALL patients, an allogeneic stem cell transplant is the only curative option to date, and it is recommended to all patients in second or later complete remission. The potential advantages of stem cell transplant short treatment duration, favorable outcome in some trials must be balanced against the disadvantages; mortality of about 20%, morbidity, late complications, reduced quality of life, and has to be assessed in relation to the improved outcome with targeted therapies.

TARGETED THERAPIES

Substantial progress in adult acute lymphoblastic leukemia has been made in the last decade by the introduction of targeted therapies, either with tyrosine kinase inhibitors or by immunotherapeutic approaches (Table 102-5).

TABLE 102-5

Targeted Therapies in Ph-Positive Adult ALL

Tyrosine Kinase Inhibitors
Ph/BCR-ABL+ ALL TKIs Imatinib, Dasatinib, Nilotinib, Bosutinib, Ponatinib **Ph-/bcr-abl-like ALL** ABL1, ABL2; Dasatinib, JAK2; Ruxolitinib
Immunologic Approaches (see Table 102-6)
Antibodies directed leukemia surface antigens Monovalent antibodies Bivalent antibodies against the tumor and CD3 (e.g., blinatumomab) **Adoptive cellular therapy** T cells engineered to kill leukemic cells

TABLE 102-6

Expression of Antigens in B-Cell Lineage ALL for Potential Antibody Therapy

Surface Antigen	ALL Subtype	Expression on LBC	Monoclonal Antibody
CD20	Burkitt lymphoma/leukemia B-precursor	86–100% 30–40%	Rituximab Ofatumomab
CD22	B-precursor Mature B-ALL	93–98% ~100%	Inotuzumab Epratuzumab Moxetumomab pasudotox
CD19	B-precursor Mature B-ALL	95–<100% 94–<100%	T-cell-activating therapies Blinatumomab Bispecific CD3/CD19 CAR T cells (Chimeric antigen receptor modified T cells)

Abbreviation: LBC, leukemic blast count.

Source: D Hoelzer: Hematology Am Soc Hematol Educ Program 2011:243, 2011.

TYROSINE KINASE INHIBITORS IN PHILADELPHIA POSITIVE ACUTE LYMPHOBLASTIC LEUKEMIA

Patients with Ph+ ALL constitute ~25% of adult B-lineage ALL, with the incidence increasing to about 50% among elderly patients. In the "pre-Imatinib era," complete remission rates were 60–70%, the survival in those patients treated with chemotherapy alone was ~10%, and after allogeneic stem cell transplant it was ~30%. The results improved substantially when the first-generation TKI imatinib became available, with complete remission rates of 80–90%, but particularly the rate of molecular remissions (*BCR-ABL*-negativity) increased from 5 to 50% or higher, and the 5-year survival increased to ≥50–60%.

Treating adult Ph+ ALL with an allogeneic stem cell transplant in first complete remission is still the best treatment option. However, patients may receive only chemotherapy plus a TKI as primary treatment, not undergoing a stem cell transplant in first remission. Thus, a Ph+ group with lower relapse risk could probably be identified by MRD response, absence of additional chromosomal abnormalities, or *IKZF1* gene deletion. Faster and deeper molecular responses are achieved with second-generation TKIs (dasatinib, nilotinib), but it is not clear whether this translates into a survival benefit. A third-generation TKI is ponatinib, which targets the T315I and other resistant mutations either present at diagnosis, or developing after treatment with other TKIs.

IMMUNOTHERAPEUTIC APPROACHES

Immunologically based treatments with monoclonal antibodies or activated T cells are showing encouraging antitumor effects in patients with ALL.

B-lineage blast cells express a variety of specific antigens, such as CD19, CD20, and CD22. Monoclonal antibodies have been developed to target these antigens.

Anti-CD20

The anti-CD20 monoclonal antibody rituximab has substantially improved the outcome of patients with de novo Burkitt leukemia/lymphoma. With repeated short cycles of intensive chemotherapy combined with rituximab, the overall survival of such patients increased to >80% compared to earlier results of <60% without rituximab.

Anti-CD22

Monoclonal antibodies directed against CD22, linked to cytotoxic agents, such as calicheamicin (inotuzumab ozogamicin), or to plant or bacterial toxins (epratuzumab) are being explored in refractory/relapsed adult ALL. In a trial of patients with relapsed/refractory ALL treated with inotuzumab ozogamicin, the complete response rate (including responses without blood cell count recovery) was 66%, and of those, 78% achieved a molecular complete remission. When inotuzumab is combined with less intensive chemotherapy (anthracyclines and alkylating agents) encouraging results were obtained in elderly patients (>60 years) as first-line therapy.

Anti-CD19

Targeting CD19 is of great interest, as this antigen is expressed in all B-lineage cells, most likely including early lymphoid precursor cells. A promising approach is the bi-specific antibody blinatumomab, which combines single chain antibodies to CD19 and CD3, such that T cells are brought into proximity with and lyse the CD19-bearing B cells.

This antibody was effective in patients with positive MRD, and 80% converted to MRD-negativity; some patients had promising survival durations even without stem cell transplantation. In a trial of adult patients with refractory/relapsed ALL treated with blinatumomab, the rate of complete remissions was 43%, and the MRD response rate was 82%.

Chimeric Antigen Receptor (CAR) T cells

The adoptive transfer of CAR-modified T cells directed against CD19 is a promising approach to the treatment of CD19+ childhood or adult ALL. Complete response rates in adults ranged from 67 to 91% with an MRD negativity in 60–81% of the complete responders. The rate of allogeneic stem cell transplantation after CAR T cells varied from about 10 to 50%. Since patients who underwent an allogeneic stem cell transplant after CAR T-cell therapy had a similar outcome as the nontransplanted patients, the value of the transplant is unclear. CAR T-cell therapy can be highly toxic. The accompanying cytokine release syndrome related to systemic immune activation produces fever, hypotension, confusion, and delirium. These effects appear in the first week of therapy and generally abate, but severe neurotoxicity may be slow to recover. The CD19 negative relapse rate after CAR T-cell therapy is 10–20%, and those patients have limited treatment options.

CONCLUSION AND FUTURE DIRECTIONS

Progress in ALL has led to the recognition of better defined ALL subentities, the importance of MRD as a prognostic factor and therapeutic target, and the value of new targeted therapies. Treatment outcome of adult ALL has improved with about half of the patients surviving >5 years and those surviving 5 years are most likely cured. Newer options, such as less intensive chemotherapy, reduction of stem cell transplantation, and incorporation of targeted therapies are promising options to reduce toxicities and improve the life quality.

FURTHER READING

Bassan R, Hoelzer D: Modern therapy of acute lymphoblastic leukemia. J Clin Oncol 29:532, 2011.

Campana D: Minimal residual disease in acute lymphoblastic leukemia. Hematology Am Soc Hematol Educ Program 2010:7, 2010.

Coustan-Smith E et al: Early T-cell precursor leukaemia: A subtype of very high-risk acute lymphoblastic leukaemia. Lancet Oncol 10:147, 2009.

Gökbuget N et al: Adult patients with acute lymphoblastic leukemia and molecular failure display a poor prognosis and are candidates for stem cell transplantation and targeted therapies. Blood 120:1868, 2012.

Hoelzer D et al: Improved outcome of adult Burkitt lymphoma/leukemia with rituximab and chemotherapy: Report of a large prospective multicenter trial. Blood 124:3870, 2014.

Iacobucci I, Mullighan C: Genetic basis of acute lymphoblastic leukemia. J Clin Oncol 35:975, 2017.

Kantarjian HM: Inotuzumab ozogamicin versus standard therapy for acute lymphoblastic leukemia. N Engl J Med 375:740, 2016.

Park JH et al: CD19-targeted CAR T-cell therapeutics for hematologic malignancies: Interpreting clinical outcomes to date. Blood 127:3312, 2016.

Roberts KG et al: Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. N Engl J Med 371:1005, 2014.

Chapter 103: Chronic Lymphocytic Leukemia

Jennifer A. Woyach; John C. Byrd

INTRODUCTION

Chronic lymphocytic leukemia (CLL) is a monoclonal proliferation of mature B lymphocytes defined by an absolute number of malignant cells in the blood (5×10^9/mL). The presence of malignant B cells under this count in the blood without nodal, spleen, or liver involvement and absent cytopenias is a precursor of this disease called *monoclonal B-cell lymphocytosis* (MBL) with ~1–2% chance per year of progressing to overt CLL. CLL is a heterogeneous disease in terms of natural history, with some patients presenting asymptomatically and never requiring therapy, whereas others present with symptomatic disease, require multiple lines of therapy, and eventually die of their disease. Over the past 10–15 years, the understanding of CLL origin and biology has grown exponentially, leading first to more refined disease definition, prognostic markers, and, subsequently, introduction of novel therapies that have significantly changed the natural history of this disease. In this chapter, we review the epidemiology, biology, and management of CLL, with a focus on new knowledge that is currently changing standards of care.

EPIDEMIOLOGY

CLL is primarily a disease of older adults, with a median age at diagnosis of 71 and an age-adjusted incidence of 4.5/100,000 people in the United States. The prevalence of CLL has increased over the past decades due to improvements in therapy for this disease and also survival of older patients from other medical ailments. In 1980, the 5-year overall survival of patients was 69%, and this increased to 87.9% in 2007 and is likely even higher today. The male:female ratio is 2:1; however, as patients age, the ratio becomes more even, and over the age of 80, the incidence is equal between men and women. The disease is most common in Caucasians, less common in Hispanic and African Americans, and is rare in the Asian population.

Unlike many other malignancies, there have been no definitive links between CLL and exposures. Indeed, CLL is one of the only types of leukemia not linked to radiation exposure. Agent Orange exposure has been implicated, and CLL is thus a service-connected condition for those who were exposed to Agent Orange in the Vietnam conflict.

CLL is one of the most familial-associated malignancies, and the first-degree relative of a CLL patient has an 8.5-fold elevated risk of developing CLL than the general population. MBL is also more common in families with two first-degree relatives having CLL, further supporting a genetic predisposition of this disease. Despite

this, specific genes conferring risk in the familial setting outside of specific families have been difficult to identify. In genome-wide association studies (GWAS), ~30 SNPs have been identified, which is estimated to account for 19% of the familial risk of CLL. Genes involved in apoptosis, telomere function, B-cell receptor (BCR) activation, and B cell differentiation have all been implicated in GWAS. Variants in shelterin complex proteins involved in telomere maintenance such as POT1 have been identified in a small number of families.

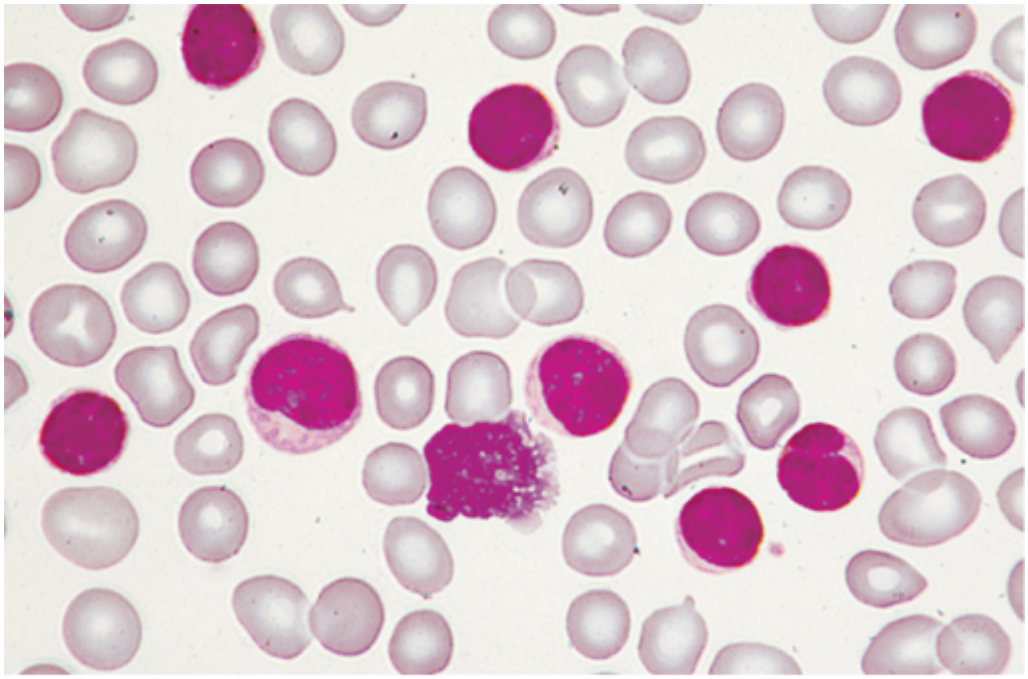
BIOLOGY AND PATHOPHYSIOLOGY

CELL OF ORIGIN

The cell of origin in CLL has not definitively been established. The morphology, immunophenotype, and gene expression pattern of CLL cells are that of a mature B cell (**Fig 103-1**), and so it has been presumed that the initiating cell is a mature lymphocyte, perhaps memory B cells. However, many facets of CLL biology do not support this idea, including antigen-binding characteristics of CLL cells and the presence of stereotyped BCRs. Other possibilities include a stepwise process including a series of transforming events at various stages of B-cell development, potentially including de-differentiation of more mature cells. The self-renewing, multipotent hematopoietic stem cell (HSC) might also be the originating cell of CLL, postulated based on transplant studies in mice showing clonal leukemic cell development with same or different characteristics from donor leukemia after transplantation of HSC. More work will be required to elucidate the origins of CLL.

FIGURE 103-1

Chronic lymphoid leukemia in the peripheral blood. *(From Williams Hematology, 7th ed, in M Lichtman et al [eds]: New York, McGraw-Hill, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

B-CELL RECEPTOR SIGNALING IN CLL

Perhaps the most important advancement in CLL biology is the understanding of the role of BCR signaling in the disease. CLL has distinct BCR signaling as compared to normal B cells, which is characterized by low-level IgM expression, variable response to antigen stimulation, and tonic activation of antiapoptotic signaling pathways that promote tumor survival. CLL cells by gene expression profiling share many features with antigen-activated mature B cells, suggesting a role for activation of BCR signaling in the disease pathogenesis. Tissue-based microarrays have revealed upregulation of BCR pathway genes in the lymph nodes and bone marrow compared to the peripheral blood, suggesting a particular importance of this pathway in microenvironmental homing.

Fitting with the role of BCR signaling in CLL, one of the most influential prognostic factors identified in this disease is the mutational status of the immunoglobulin heavy chain variable (IGHV) region. During normal B-cell maturation, the variable regions of the immunoglobulin heavy chain undergo somatic hypermutation. In CLL, ~60% of patients have IGVH that is ≥2% mutated from germline. This may indicate a more mature, postgerminal center progenitor, and is typically associated with a more indolent disease course. Conversely, ~40% of patients will have IGVH <2% mutated from germline, which is associated with more rapid progression of disease and short survival. Unfavorable biologic properties including enhanced telomerase activity, overexpression of activation-induced cytidine deaminase, increased nuclear factor-κB (NF-κB) activity, high-risk genomic features, and clonal evolution are also associated with IGHV unmutated disease.

Because IGHV sequencing was initially cumbersome to perform, a number of surrogate factors have been identified; however, none yet have been shown to be equal or superior to IGVH sequencing. The most prevalent of these surrogate markers are Zap-70 expression, *ZAP-70* methylation, and surface CD38 expression. Zap-70 protein is a normal intracellular T-cell signaling protein that is aberrantly expressed in most IGHV unmutated CLL cells. CD38 is a marker that is also more highly expressed on the surface of IGHV unmutated CLL cells. Both these prognostic factors are widely used but limited in their applicability. Zap-70 protein status is difficult to measure by flow cytometry, and it has low reproducibility. Measurement of methylation status of the *ZAP-70* promoter is much more precise but not widely available. CD38 expression is easier to measure by flow cytometry but not as highly predictive of outcomes and can change during the course of disease.

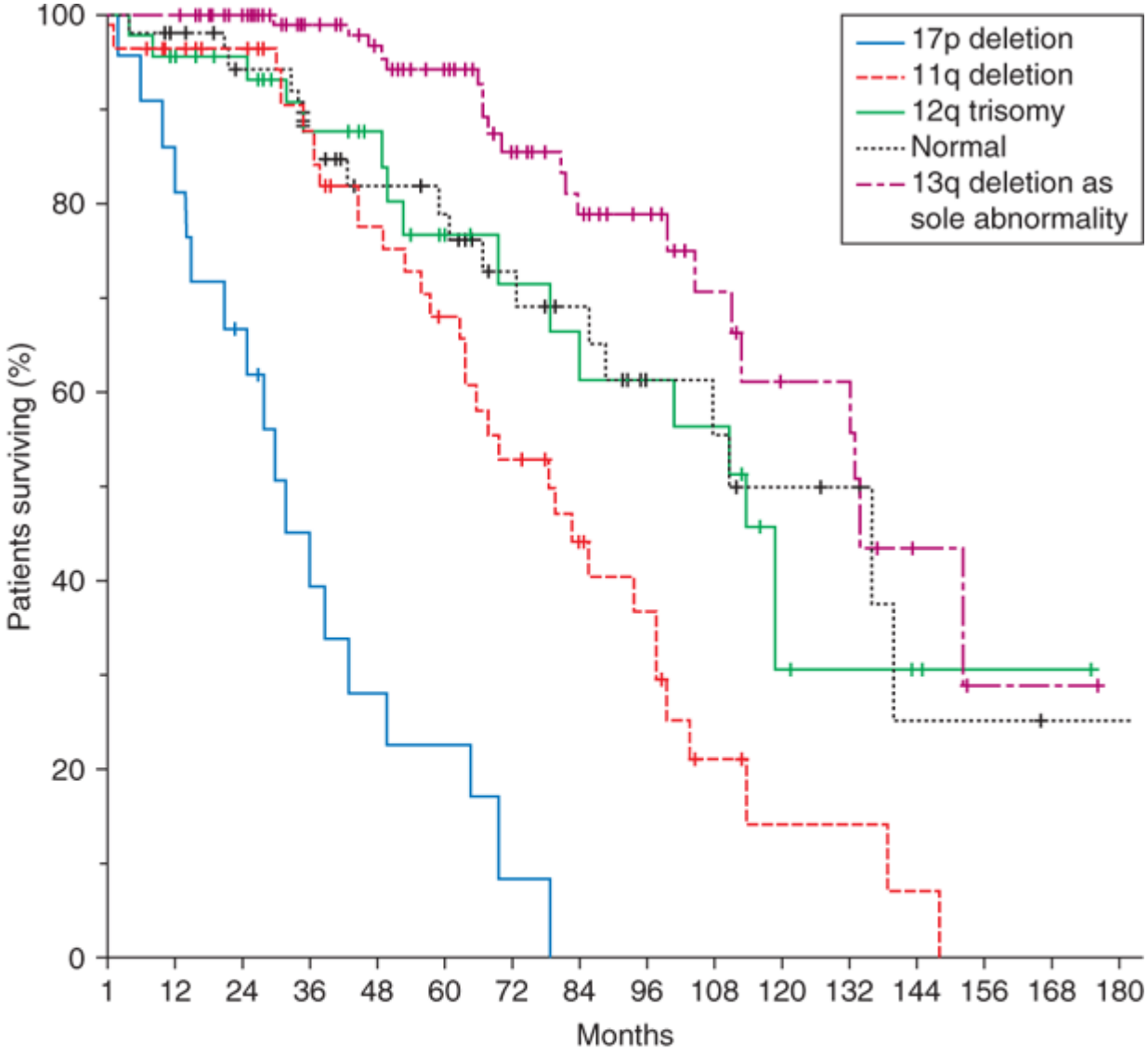
CYTOGENETIC ABNORMALITIES

Besides IGHV mutational status, recurrent cytogenetic abnormalities are the most robust prognostic factor clinically available in CLL. These abnormalities are typically identified by fluorescent in situ hybridization (FISH) analysis; however, stimulated metaphase karyotype has a role as well. The most well-characterized abnormalities include del(13)(q14.3), trisomy 12, del(11)(q22.3), and del(17)(p13.1) **(Fig. 103-2)**. The presence of sole del(13)(q14.3) is associated with more indolent disease, prolonged survival, and good response to traditional therapies. Usually this abnormality is not seen on banded karyotype analysis, and when present on karyotype, it indicates a larger deletion involving the retinoblastoma gene, which negates the favorable prognosis associated with this marker. Trisomy 12 has a more intermediate prognosis. The del(11)(q23.3)

results in deletion of the *ATM* gene and is associated with bulky lymphadenopathy and aggressive disease in young patients, with inferior prognosis, more rapid progression to symptomatic disease, and shorter survival. The del(17)(p13.1) results in loss of one allele of the tumor suppressor TP53 and is associated with the poorest prognosis in CLL with rapid disease progression, poor response to traditional therapies, and shorter survival. Other abnormalities have been shown to be important in smaller studies but are not routinely performed at all centers. Finally, complex karyotype (three or more abnormalities) on stimulated metaphase karyotype analysis has significant adverse impact on time to treatment and overall survival.

FIGURE 103-2

Outcomes among CLL patients with various cytogenetic abnormalities. *(From Döhner H et al: N Engl J Med 343:1910, 2000.)*



No. AT Risk

17p deletion	23	18	13	8	5	4	1	0	0	0	0	0	0	0	0	0
11q deletion	56	53	47	43	33	27	20	15	10	4	2	2	1	0	0	0
12q trisomy	47	44	41	29	24	17	14	13	12	11	4	3	2	1	1	0
Normal	57	51	45	37	30	27	20	17	12	11	6	5	2	2	1	1
13q deletion as sole abnormality	117	117	106	91	80	63	45	36	24	16	12	11	3	1	1	0

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clonal evolution, or acquisition of cytogenetic or molecular abnormalities, is common in CLL, especially in patients with IGHV unmutated CLL. Because the cytogenetics of patients can change even in the absence of therapy, it is recommended that FISH +/− cytogenetics are checked before every line of therapy, mostly to evaluate acquisition of del(17)(p13.1).

GENE MUTATIONS AND MIR ALTERATIONS

Compared with many other malignancies, the genome in CLL is relatively simple, with an average CLL genome carrying ~20 nonsynonymous alterations and ~5 structural abnormalities. And, unlike many other hematologic malignancies, there is no unifying genetic lesion, and most recurrent genetic driving mutations exist at frequencies of <5%. Whole genome and whole exome sequencing have identified the most common mutations in CLL to be in *SF3B1*, *NOTCH1*, *MYD88*, *ATM*, and *TP53* (**Table 103-1**). Most of the identified mutations in these genes are common among different malignancies, and with the exception of MYD88, they are generally subclonal drivers identified with much higher frequency in IGHV unmutated disease.

TABLE 103-1

Recurrent Mutations in CLL

Gene	Frequency of Mutations (%)
SF3B1	8–14
TP53	5–13
NOTCH1	10–13
MYD88	4–8
ATM	8–11
BIRC3	<5
XPO1	<5
FBXW7	<5
POT1	<5
BRAF	<5
EGR2	<5
IKZF3	<5

Abbreviation: CLL, chronic lymphocytic leukemia.

NOTCH1 mutations are present in ~15% of CLL patients and are commonly associated with trisomy 12. Although multiple different mutations are seen, most are located within the PEST (proline, glutamic acid,

serine, and threonine) domain and result in constitutive NOTCH signaling. *NOTCH1* mutations have been associated with lower sensitivity to CD20 antibody therapy and increased risk of transformation to aggressive diffuse large B-cell lymphoma (DLBCL; Richter's transformation).

SF3B1 is a component of the RNA spliceosome and is mutated in 10–15% of CLL cases. Mutations appear to be associated with intermediate-risk disease, and, functionally, SF3B1 may be important in the response to DNA damage.

Mutations of the tumor suppressor *TP53* are found in ~5% of CLL in previously untreated early stage disease and up to 40% in later stages. Seventy percent of the time these mutations coexist with del(17)(p13.1), effectively eliminating TP53 function. As expected, and consistent with other malignancies, TP53 mutations are associated with a poor prognosis and expected lack of response to DNA-damaging therapies.

ATM mutations, which are heterogeneous and occur throughout the gene, occur in 10–15% of CLL patients. *ATM* mutations often coexist with del(11)(q22.3), eliminating *ATM* on the alternate allele. Similar to *TP53*, mutations in *ATM* tend to result in impaired response to DNA damage, which can reduce responsiveness to chemotherapy.

In contrast to the aforementioned mutations, those in MYD88 tend to occur in IGHV mutated CLL and be associated with a more indolent prognosis. This gene is involved in Toll-like receptor signaling, and the most common mutation, L265P, results in constitutive activation and NF-κB activity.

Along with abnormalities in coding genes, it has become apparent that noncoding genes such as microRNAs are recurrently altered in CLL. The most common cytogenetic abnormality, del(13)(q14.3) results in loss of the miR15/16 cluster, which is important in the pathogenesis of CLL. In normal cells, miR15A/miR16A inhibit antiapoptotic gene expression (including BCL2, CCND1&3, and CDK6), and this specific deletion allows for overexpression of these genes and thus increased cell survival. Loss of other miR expression such as mir-181a leads to overexpression of proteins such as the antiapoptotic gene MCL-1 and TCL1. Overexpression of miR-155, an onco-miR associated with B-cell transformation, has also been documented in the majority of CLL patients.

IMMUNOLOGY

CLL is characterized by dysregulation of the normal immune system in addition to the malignant immune cells. Besides numerical abnormalities due to bone marrow dysfunction, even in the early stages of disease there are skewed ratios of immune cells and functional abnormalities. Innate immune system defects associated with CLL include reduced complement proteins and activity, qualitative neutrophil defects, and functional defects of natural killer cells.

More focus has been placed on the impairments in the adaptive immune system in this disease. Within the CD4+ T-cell compartment, a qualitative defect is noted similar to chronic antigen stimulation inducing a phenotype of T-cell exhaustion typical of what is seen in chronic viral infections such as hepatitis. This has been demonstrated to lead to impaired T-cell cytotoxic capacity and reduced proliferative ability.

Additionally, there are physical changes in the T-cell cytoskeleton that causes impaired immune synapse formation with antigen presenting cells. In addition to a lack of capacity to respond to pathogens, the T-cell defect in CLL also likely leads to tumor cell tolerance. During the course of the disease, the polarization of the CD4+ T cells shifts from a Th1 (cytotoxic) phenotype to a Th2 phenotype, which leads to expansion of immunosuppressive cytokines such as IL-10. Additionally, in the later stage of disease T regulatory cells are expanded, which contributes to an immunosuppressive phenotype.

Other components of the immune microenvironment are altered as well to form a more supportive environment for the malignant cells. M2 monocytes have been shown to differentiate into a type of tumor-associated macrophage known as a *nurse-like cell* in CLL. These cells promote survival by secreting chemokines and cytokines that increase migration and activation.

The humoral immune system in CLL is also dysregulated, as is expected for a malignancy that results in very few normal B cells. Hypogammaglobulinemia is very common and affects all subclasses of immunoglobulins, occurring in ~85% of patients at some time in their disease course, and is more common as disease progresses. A correlation between low IgG and IgA and infection risk has been established, but isolated IgM reduction does not seem to be associated with excess infection risk. Also, CLL cells can secrete monoclonal IgM in a small number of cases, and this can correlate with disease progression.

CLINICAL PRESENTATION AND DIAGNOSIS OF CLL

CLINICAL PRESENTATION AND DIAGNOSIS

The presentation of CLL most commonly occurs as an incidental diagnosis made at the time of medical evaluation for another cause. In this regard, CLL is most commonly diagnosed on routine blood work demonstrating an elevated lymphocyte count in asymptomatic individuals, although some patients present with symptoms and require early therapy. When noting either an elevated total white blood cell (WBC) count with lymphocytic predominance or a normal WBC with a differential showing a lymphocytosis, the next step is to perform flow cytometry on the peripheral blood. In CLL, this will reveal the typical immunophenotype that includes the typical B-cell markers CD19, CD20, CD22, CD23, the T-cell marker CD5 (CD5 is also expressed on the B1 subset of B cells that typically has unmutated immunoglobulin and responds to antigens independent of cognate T-cell help), and dim surface immunoglobulin of either kappa or lambda type (**Table 103-2**). Atypical phenotypes can be seen as well and usually can be differentiated on the basis of morphology, cytogenetics, or clinical presentation. In cases in which the clonal B cell count based on flow cytometry is $\geq 5 \times 10^9$/L, no further workup is needed to confirm the diagnosis of CLL.

TABLE 103-2

Typical Immunophenotype of CLL Compared with Other B-Cell Malignancies

	CD5	CD10	CD19	CD20	CD23	Cyclin D1	Surface Ig
Disease							
CLL	+	−	+	+ (dim)	+	−	+ (dim)
Mantle cell lymphoma	+	−	+	+ (mod/bright)	−	+	+ (mod/bright)
Marginal zone lymphoma	−/+	−	+	+ (mod/bright)	−/+	−	+ (mod/bright)
Follicular lymphoma	−	+	+	+	+	−	

Abbreviation: CLL, chronic lymphocytic leukemia.

Some patients will present with a small clonal proliferation of CLL cells in the peripheral blood but will also have lymphadenopathy or splenomegaly. In these cases, the likely diagnosis is small lymphocytic lymphoma (SLL), a semantic designation from CLL that denotes a primarily tissue-based disease rather than bone marrow/blood-based disease. The genetic and molecular features of SLL are identical to those of CLL. The retention of the cells in tissues may be related to the expression of particular adhesion molecule. Thus, SLL patients are managed identically to CLL, and often in the later stages of disease these patients will often have blood and bone marrow involvement as well.

MONOCLONAL B-CELL LYMPHOCYTOSIS

Patients who do not meet the diagnostic criteria for CLL based on quantification of clonal B cells in the peripheral blood and who do not have associated signs of CLL including lymphadenopathy, organomegaly, or cytopenias have a disorder known as monoclonal B-cell lymphocytosis (MBL), which is now thought to precede every case of CLL. Analogous to monoclonal gammopathy of uncertain significance (MGUS) in myeloma, not all MBL progresses to CLL. MBL is initially characterized by a CLL-like immunophenotype in ~75% of cases but can also be atypical (CD23 negative or bright CD20) or CD5 negative. More relevant for prognosis is characterization by count, with low-count MBL defining those patients with $<0.5 \times 10^9$ clonal B cells/L, and high-count MBL defining those with $>0.5 \times 10^9$ but $<5 \times 10^9$/L. Patients with low-count MBL have a negligible rate of progression to CLL, whereas those with high count progress to overt CLL at a rate of 1–2% per year, warranting continued monitoring. Population-based studies have estimated the prevalence of MBL

up to ~12% in the general population, where it is most common in elderly men. It is especially common in first-degree relatives of CLL patients, where the frequency is ~18%.

Although the risk of MBL progression is relatively low, it has become apparent that patients still experience complications that suggest an immune dysfunction in MBL that is similar to that seen with CLL. Rates of serious infections requiring hospitalization appear to be significantly increased in MBL, similar to the rates seen in CLL. In a case-control study, patients with MBL had a 16% chance of hospitalization over a 4-year time period, compared with 18.4% in patients with newly diagnosed CLL. Secondary cancers also appear to be increased in MBL. These data suggest that monitoring for patients with MBL should focus on vaccinations and age-appropriate cancer screening, as the probability of complications appears to be higher than the risk of progression in most of these patients. Follow-up for patients with MBL can occur with the primary care physician as this does not represent a malignancy, whereas CLL is mostly comanaged with both a primary care physician and a hematologist.

COMPLICATIONS OF CLL

A significant amount of morbidity and mortality related to CLL is due to complications of the disease. In general, complications besides disease progression include infections, secondary cancers, autoimmune complications, and transformation to a more aggressive clonally related lymphoma.

INFECTIONS

Infections are a leading cause of both disease-related morbidity and death in patients with CLL, with ~30–50% of deaths in CLL patient attributed to infection. Owing to the immune dysfunction associated with the disease, patients are at risk for both typical and atypical infections. Besides this baseline risk of infections, most CLL therapies can increase infection risk. For many nucleoside analog-based chemotherapy regimens used in CLL, prophylaxis for *Pneumocystis* pneumonia is indicated for at least 6 months following therapy to allow recovery of functional T cells. Viral prophylaxis is also indicated for many chemotherapy regimens and for patients with a history of varicella zoster to diminish reactivation and morbidity from this virus.

Because of the abnormalities in cellular and humoral immunity, vaccine responses in CLL are limited in many patients, especially in the later stages of disease. In one study, one dose of 13-valent pneumococcal vaccine produced an adequate immune response in only 58% of patients compared with 100% in age-matched controls. Despite the known limitations, vaccination against influenza and pneumococcal pneumonia is recommended in CLL. Live vaccines, such as the varicella zoster vaccine, should be avoided because of the small risk of viral reactivation with an immunocompromised host.

As discussed earlier, hypogammaglobulinemia is common in CLL and can be associated with significant risk for infections, primarily of mucocutaneous etiology such as sinusitis and bronchitis. In addition, women can have frequent urinary tract infections. While administration of prophylactic intravenous immunoglobulin (IVIg) has not been shown to improve survival, it has been shown to reduce the number of minor or moderate bacterial infections, and thus is indicated in patients with hypogammaglobulinemia who suffer from

recurrent infections or have pulmonary bronchiectasis. It is also our practice to administer at least one dose immunoglobulin to CLL patients who develop influenza with coexisting hypogammaglobulinemia to diminish risk of post-influenza pneumococcal pneumonia. IVIg is probably indicated in patients who have been hospitalized for a serious infection and in those whose IgG level is <300 mg/dL.

SECONDARY MALIGNANCIES

Multiple population-based studies have shown that patients with CLL are at an elevated risk to develop other cancers, with a rate up to three times that of the general population, even in the absence of cytotoxic chemotherapy. The most common types of cancers seen in CLL are skin cancers, prostate, and breast cancers, although other cancers are seen as well. Skin cancers are particularly common, with a rate of 8- to 15-fold higher than the general population, and may behave more aggressively. All CLL patients should be counseled on the use of sunscreen while outdoors and should undergo preventative skin examinations.

In one single-center study, older age at CLL diagnosis, male sex, high β_2 microglobulin, high lactate dehydrogenase (LDH), and chronic kidney disease were associated with excess risk of other cancers; other CLL-specific risk factors have not shown association with other cancer risk.

While cancer risk is higher, there are no specific recommendations for increased cancer screening in CLL patients. Age- and sex-appropriate screenings should be recommended.

Conflicting data exist regarding the risk of cancers following CLL-specific therapy. Chemoimmunotherapy, in particular alkylator-containing regimens, seems to be associated with an increased risk for secondary cancers.

AUTOIMMUNE COMPLICATIONS

Autoimmune complications are frequent in CLL. Most commonly, these include autoimmune cytopenias, but autoimmune complications of other organs including glomerulonephritis, vasculitis, and neuropathies have also been reported. Of the autoimmune cytopenias, the most common is autoimmune hemolytic anemia (AIHA), which is an antibody-mediated destruction of autologous red blood cells (RBCs). Second most common is immune thrombocytopenia (ITP), which shares some features with AIHA and has a similar mechanism targeting platelets. These two syndromes may occur in isolation, sequentially in the same patient, or present in combination as Evan's syndrome. Pure red cell aplasia (PRCA) and autoimmune granulocytopenia (AIG) are comparatively rare and can occur alone or in combination with other AIC. It is difficult to tease out whether autoimmune cytopenias lead to worse prognosis in CLL because of various complicating factors. However, it is clear that these can lead to significant morbidity, both due to the process itself and due to therapies required for management.

AIHA usually presents as an isolated anemia with an elevated reticulocyte count and features of hemolysis including elevated bilirubin and LDH, and low haptoglobin. Detection of a warm IgG antibody on the surface of RBCs with a Coombs test can help solidify the diagnosis, although Coombs-negative cases can occur.

Immediate therapy is almost always necessary, and consists of transfusion and immunosuppression. Glucocorticoids are often used for initial therapy, although in most cases additional treatment is needed due to either poor response or recurrence with taper of steroid dosing. Rituximab can be successful, and therapy directed toward the underlying CLL is often effective in more resistant cases. Transfusion of blood in cases of robust AIHA must be initiated with caution as transfusion reactions can be seen due to poorly matched blood.

ITP can be more difficult to diagnose, as it may be difficult to differentiate from progression of disease due to the lack of laboratory tests that identify platelet destruction from this mechanism. Signs that point toward ITP include isolated thrombocytopenia and rapid decline in platelet levels in the absence of an alternative etiology. A bone marrow biopsy showing normal or increased megakaryocytes can be used to confirm the diagnosis but is often not necessary. In CLL, treatment for ITP is usually instituted when platelet levels drop to 20–30,000 or if there is evidence of bleeding complications or need for invasive procedures. Like AIHA, initial therapy consists of glucocorticoids and IVIG, with rituximab also being an effective method to induce long-term remissions. Also, the thrombopoietin receptor agonists romiplostim and eltrombopag are effective in secondary ITP. In many cases, ITP can be successfully treated without treating the underlying CLL. In cases in which anemia or thrombocytopenia appear, it is important to investigate the mechanism as the approach to therapy of autoimmune cytopenias in CLL differs from cytopenias due to marrow replacement.

RICHTER'S TRANSFORMATION

One of the most devastating complications of CLL is Richter's transformation, transformation of CLL to an aggressive lymphoma, most commonly DLBCL. The World Health Organization also recognizes Hodgkin's lymphoma (HL) as a variant of Richter's transformation; other aggressive lymphomas are rarely identified. Some older series have included prolymphocytic transformation in this category, although this has much less prognostic impact on long-term outcome. The prevalence of Richter's transformation is difficult to estimate based on previous studies, but one prospective observational study estimated a rate of 0.5% per year for DLBCL and 0.05% per year for HL. Risk factors for development include bulky lymphadenopathy, *NOTCH1* mutations, del(17)(p13.1), and a specific stereotyped IGVH usage. Lymphomas arising in the setting of CLL can either be clonally related or unrelated to the initial CLL, with prognosis significantly better for clonally unrelated lymphomas. In addition, patients with Hodgkin's transformation have improved outcome, particularly in the absence of prior fludarabine treatment.

Clinical signs of Richter's transformation include rapid progression in adenopathy, often in a specific area, and constitutional symptoms including fatigue, night sweats, fever, and weight loss. LDH is usually high. In suspected cases, the first step is ^{18}FDG-PET/CT (fluorodeoxyglucose–positron emission tomography combined with computed tomography) scan to localize an area for biopsy. Standardized uptake values (SUV) <5 is consistent with CLL and can rule out Richter's transformation in many cases. SUV >5 are suspicious for Richter's transformation, with SUV ≥10 very concerning. Excisional biopsy is diagnostic. Needle biopsy should be discouraged.

Therapy for DLBCL Richter's transformation usually involves combination chemoimmunotherapy. Outcomes are poor with median survivals of 6–16 months in most series for clonally related Richter's versus ~5 years for clonally unrelated. This highlights an area of unmet need in CLL therapy and an area of active investigation. For fit patients who achieve a response with therapy, stem cell transplantation has the possibility to induce long-term remissions and should be explored. Patients with Hodgkin's disease can be treated according to the algorithm for this disease, with many individuals being cured.

WORKUP OF CLL AND APPROACH TO THERAPY

WORKUP AND STAGING

Workup of a patient with new diagnosis of CLL based on typical immunophenotyping includes a detailed history of infectious history; family history of CLL; and careful physical examination with attention to the lymph nodes, spleen, and liver. In patients desiring to know the expected natural history of their CLL, prognostic testing using FISH and stimulated karyotype as well as IGHV sequencing can be performed. Imaging with CT scan is usually not necessary unless there are symptoms and concern for intra-abdominal nodes out of proportion to peripheral nodes. Bone marrow biopsy is not undertaken until therapy is initiated except in cases of unexplained cytopenias.

STAGING

There are two widely used staging systems in CLL: The Rai staging system is used more commonly in the United States, whereas the Binet system is more commonly used in Europe. Both characterize CLL on the basis of disease bulk and marrow failure (**Table 103-3**). Both rely on physical examination and laboratory studies and do not require imaging or bone marrow analysis. While the initial staging systems could reliably predict survival in CLL, with the changes in therapy since the original description of the stages, the impact of initial stage on survival is not as clear. Cytogenetic and genomic testing can help refine outcome of these staging tests. An international collaboration integrated both clinical and genomic staging to better predict outcome at diagnosis and time of initial treatment.

TABLE 103-3

Staging of CLL

Rai Staging System	
Low risk (stage 0)	Lymphocytosis only
Intermediate risk (stage I/II)	Lymphocytosis with lymphadenopathy, with or without splenomegaly or hepatomegaly
High risk (stage III/IV)	Lymphocytosis with anemia or thrombocytopenia due to bone marrow involvement
Binet Staging System	
A	<3 areas of lymphadenopathy
B	≥3 areas of lymphadenopathy
C	Hemoglobin ≤10 g/dL and/or platelets <100,000/μL

Abbreviation: CLL, chronic lymphocytic leukemia.

CRITERIA FOR THE INITIATION OF THERAPY

Currently, a watchful waiting strategy is used for most patients with CLL, with therapy reserved for patients with symptomatic disease. This recommendation is based on multiple trials showing no survival advantage with earlier therapy, although this question is currently being revisited with novel targeted therapies.

With the exception of patients participating on early intervention studies in CLL, disease-related symptoms that require the initiation of therapy are outlined in Table 103-4. Except for the rare patient who presents with disease requiring urgent therapy, most times these symptoms can be monitored over short periods to determine relatedness to CLL and need for therapy.

TABLE 103-4

Criteria for the Initiation of Therapy

Symptoms Indicating Need for Therapy in CLL
Evidence of progressive marrow failure (worsening of anemia or thrombocytopenia not due to autoimmune destruction)
Massive (≥6 cm below costal margin), progressive, or symptomatic splenomegaly
Massive (≥10 cm), progressive, or symptomatic lymphadenopathy
Autoimmune anemia or thrombocytopenia not responsive to standard therapy
Constitutional symptoms (one or more of the following: unintentional weight loss ≥10% over 6 months, significant fatigue, fevers ≥100.5°C for 2+ weeks without infection, night sweats for >1 month without infection)

Abbreviation: CLL, chronic lymphocytic leukemia.

INITIAL THERAPY FOR CLL

Chemoimmunotherapy and Monoclonal Antibody Therapy

Chemotherapy and chemoimmunotherapy are the standard therapies for CLL. For patients who are young (≤65 years), the gold standard for therapy is a combination of the nucleoside analogue fludarabine, the alkylator cyclophosphamide, and the anti-CD20 monoclonal antibody rituximab (FCR). In phase III study, this combination produced an overall response rate (ORR) of 93% with a complete response (CR) rate of 44%. Median progression-free survival (PFS) is almost 5 years. Substitution of bendamustine for fludarabine and cyclophosphamide or addition of other chemotherapy-based treatments to FCR have not improved outcome. A subset of patients treated with the FCR regimen has durable responses over 10 years. This group is primarily composed of those patients with mutated IGVH and good cytogenetic risk. However, despite the efficacy of this regimen, short- and long-term toxicities limit its adaptability to many patients with IGHV-mutated disease. Short-term toxicities are mostly related to myelosuppression and include neutropenia and infection. Long-term cytopenias are less common, but they do occur. Also, there is about a 3–5% risk of therapy-related myeloid neoplasm with this regimen that is almost always fatal. Trials in the future will need to focus on the superiority of FCR versus targeted regimens for patients who may be cured by chemoimmunotherapy.

For older patients or those with multiple comorbidities, FCR is not an appropriate option due to toxicities. For these patients, the alkylator chlorambucil in combination with the anti-CD20 antibody obinutuzumab or bendamustine with rituximab are appropriate options. While neither produces remissions as durable as FCR, both can induce CRs and remissions of 2–3 years in many patients. Toxicities with these regimens mostly relate to myelosuppression, but neither is as immunosuppressive as FCR.

Monoclonal antibodies given alone or in combination with chemotherapy were the first targeted therapies to be successful in CLL. Anti-CD20 monoclonal antibodies including rituximab, ofatumumab, and obinutuzumab are all used in this disease. Rituximab was the first agent to show a survival advantage in CLL, where the combination of FCR improved survival over the combination chemotherapy regimen FC. Alone, this antibody has shown modest activity, but activity is improved with higher doses and increased frequency of administration. Both ofatumumab and obinutuzumab are effective as single agents, but it is likely that monoclonal antibodies will be most widely used in combination. Current trials are focused on combining anti-CD20 antibodies with therapies that target BCR signaling or antiapoptotic proteins. Antibodies against other targets are also being developed, including CD19, BAFF receptor, and CD37.

B-Cell Receptor Signaling Inhibitors

Three specific targets were the first identified: spleen tyrosine kinase, phosphoinositide-3-kinase (PI3K), and Bruton's tyrosine kinase (BTK). Therapeutics directed at the latter two targets have moved forward through clinical trials and are now utilized in clinical practice.

Idelalisib is a reversible, p110 delta isoform-specific PI3K inhibitor. Because the delta isoform is specific for B lymphocytes, this agent has selective effects on the CLL cells with relative sparing of other hematopoietic cells. The definitive phase III study of the idelalisib plus rituximab regimen assessed this combination versus placebo plus rituximab in 220 patients and showed an ORR of 77% (vs 15% with rituximab plus placebo) with improved progression-free and overall survival. Toxicity specific to idelalisib included elevated alanine transaminase (ALT) and aspartate transaminase (AST) in the first 3 months of therapy and diarrhea, colitis, pneumonitis, and rash later (9+ months) into treatment. These side effects appear to be more common in younger patients given idelalisib earlier in the course of their disease.

Ibrutinib is a relatively selective, irreversible inhibitor of BTK. This target is attractive because, unlike other kinases in the BCR pathway, BTK does not have natural redundancy and is selective for B cells, so inhibition leads to a B-cell–specific phenotype. As initial therapy, ibrutinib was compared with chlorambucil, and there was an 84% lower risk of progression or death with ibrutinib, with 90% of ibrutinib-treated patients alive and progression-free at 18 months. In the relapsed phase III study, ibrutinib was compared to the CD20 monoclonal antibody ofatumumab, and there was a 78% reduction in the risk of progression or death with ibrutinib. Side effects distinct to ibrutinib include rash, diarrhea, dyspepsia, increased risk of bleeding (particularly when on anticoagulation therapy or with surgery), and atrial fibrillation. Second-generation BTK inhibitors with more specificity such as acalabrutinib are in clinical trials and may diminish these side effects. Although direct comparison of agents targeting p110 delta PI3 kinase and BTK has not occurred, BTK inhibitors appear to induce more durable remissions.

The success of ibrutinib and idelalisib has generated significant interest in other molecules targeting PI3K, BTK, and other members of the BCR signaling pathway. One issue with most drugs targeting this pathway in CLL is that while durable responses are common, CRs are not, which leads to the recommendation for indefinite therapy with these molecules. Combination clinical trials are currently underway to determine whether combinations with other active agents might allow discontinuation of drug in some settings.

Antiapoptotic Therapies

BCL2 is another promising target in CLL. Venetoclax is an orally bioavailable, selective BCL2 inhibitor. It is currently Food and Drug Administration (FDA) approved for marketing in patients with relapsed or refractory CLL who have the del(17)(p13.1). In a phase I study, the ORR with this agent in relapsed/refractory CLL was 79%, with 69% of patients on the recommended phase II dose being progression-free at 15 months. Unlike the BCR signaling antagonists, venetoclax is able to induce very deep responses including CRs with minimal residual disease negativity in a subset of patients. Distinct toxicities associated with venetoclax include neutropenia, diarrhea, and acute tumor lysis syndrome. Tumor lysis syndrome risk can be mitigated with stepped-up dosing at the beginning of treatment.

Immune Therapies

Current immune therapies include allogenic stem cell transplantation, chimeric antigen receptor (CAR) T-cell therapy, and oral immunomodulatory agents such as lenalidomide.

Stem cell transplantation is currently considered the only standard curative approach to CLL. Because most CLL patients are older and many have significant comorbidities, myeloablative transplants incur extensive morbidity and mortality, making them prohibitive in many individuals. Reduced intensity conditioning (RIC) allogeneic transplants have been successfully incorporated into the treatment of patients up to ~75 years in age but still have a ≥50% frequency of chronic graft-versus-host disease.

ASSESSING RESPONSE TO THERAPY AND MINIMAL RESIDUAL DISEASE IN CLL

Following the completion of therapy or during therapy for indefinite targeted agents, response is initially assessed using physical examination and laboratory studies (**Table 103-5**). If residual disease is not detected using these methodologies, CT scans are used to assess response. Bone marrow biopsies with flow cytometry are indicated if no disease is detected to confirm CR.

TABLE 103-5

Response Criteria in CLL

	Lymphocyte Count	Lymph Nodes[a]	Spleen/Liver Size[b]	Bone Marrow[c]	Peripheral Blood Counts
CR	<4000/μL	None >1.5 cm	Not palpable	Normocellular, <30% lymphocytes, no B lymphoid nodules	• Platelet count >100,000/μL • Hemoglobin >11 g/dL • Neutrophils >1500/μL
PR	Decrease ≥50% from baseline	Decrease ≥50% from baseline	Decrease ≥50% from baseline	Infiltrate ≤50% of baseline	One of the following: • Platelet count >100,000/μL or ≥50% from baseline • Hemoglobin >11 g/dL or ≥50% from baseline • Neutrophils >1500/μL or ≥50% from baseline
Stable disease	Not meeting CR/PR/PD criteria	Not meeting CR/PR/PD criteria	Not meeting CR/PR/PD criteria	Not meeting CR/PR/PD criteria	Not meeting CR/PR/PD criteria
PD	Increase ≥50%	Increase ≥50%	Increase ≥50%		• Platelet count ≤50% of baseline due to CLL • Hemoglobin decrease >2 g/dL due to CLL

[a]Refers to sum of the products of multiple lymph nodes evaluated by CT scan. [b]Based on physical examination. [c]Bone marrow only required to confirm CR.

Abbreviations: CLL, chronic lymphocytic leukemia; CR, complete response; PD, progressive disease; PR, partial response.

It has been established in various malignancies that complete tumor eradication is associated with longer survival. In CLL, if no malignant cells can be detected in the bone marrow down to a level of 1 CLL cell in 10^4 leukocytes (0.01%), the patient is said to be negative for minimal residual disease (MRD). Following combination chemoimmunotherapy, eradication of MRD correlates with long-term survival and potentially cure in a subset of patients receiving FCR chemoimmunotherapy. It has yet to be established whether MRD negativity in the setting of targeted therapies is a meaningful endpoint.

CONCLUSION

CLL is treated only when it becomes symptomatic. At the time of therapy chemoimmunotherapy in a small subset of patients is potentially curative. In the majority of remaining individuals with symptomatic CLL, targeted therapy directed at BTK greatly improves survival and also reverses the immune deficiency associated with the disease.

FURTHER READING

Burger JA et al: Ibrutinib as initial therapy for patients with chronic lymphocytic leukemia. N Engl J Med 373:2425, 2015.

Byrd JC et al: Targeting BTK with ibrutinib in relapsed chronic lymphocytic leukemia. N Engl J Med 369:32, 2013.

Hallek M et al: Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: A report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. Blood 111:5446, 2008.

Landau DA et al: Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell 152:714–26, 2013.

Oakes CC et al: DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. Nat Genet 48:253, 2106.

Puente XS et al: Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 475:101, 2011.

Roberts AW et al: Targeting BCL2 with venetoclax in relapsed and refractory CLL. N Engl J Med 374:311, 2016.

Thompson PA et al: Fludarabine, cyclophosphamide, and rituximab treatment achieves long-term disease-free survival in IGHV-mutated chronic lymphocytic leukemia. Blood 127:303, 2016.

Woyach JA et al: Resistance mechanisms for the Bruton's tyrosine kinase inhibitor ibrutinib. N Engl J Med 370:2286, 2014.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 104: Non-Hodgkin's Lymphoma

Caron A. Jacobson; Dan L. Longo

INTRODUCTION

Non-Hodgkin's lymphomas (NHL) are cancers of mature B, T, and NK cells. They were distinguished from Hodgkin lymphoma (HL) upon recognition of the Reed-Sternberg (RS) cell, and differ from HL with respect to their biologic and clinical characteristics. Whereas ~80–85% of patients with HL will be cured of their lymphoma by chemotherapy with or without radiotherapy, the prognosis and natural history of NHL tends to be more variable. NHL can be classified as either a mature B-NHL, or a mature T/NK-NHL depending on whether the cancerous lymphocyte is a B, T, or NK-cell, respectively. Within each category are lymphomas that grow quickly and behave aggressively, as well as lymphomas that are more indolent, or slow growing in nature. For a list of the World Health Organization (WHO) classification of lymphoid neoplasms, see **Table 104-1**.

TABLE 104-1

WHO Classification of Lymphoid Malignancies

B Cell	T Cell
Mature (peripheral) B-cell neoplasms Lymphoplasmacytic lymphoma (Waldenstrom's macroglobulinemia) Hairy cell leukemia Splenic marginal zone B-cell lymphoma Extranodal marginal zone B-cell lymphoma of MALT type Nodal marginal zone B-cell lymphoma Follicular lymphoma Mantle cell lymphoma Diffuse large B-cell lymphoma (including subtypes) High grade B-cell lymphoma with MYC and BCL2 and/or BCL6 rearrangements High grade B-cell lymphoma NOS Burkitt's lymphoma/Burkitt's cell leukemia Primary mediastinal large B-cell lymphoma Plasmablastic lymphoma Primary effusion lymphoma HHV8+ DLBCL NOS Intravascular large B-cell lymphoma ALK+ large B-cell lymphoma	Mature (peripheral) T-cell neoplasms T-cell granular lymphocytic leukemia Adult T-cell leukemia/lymphoma (HTLV-1+) Extranodal NK/T-cell lymphoma, nasal type Enteropathy-associated T-cell lymphoma Hepatosplenic T-cell lymphoma Subcutaneous panniculitis-like T-cell lymphoma Mycosis fungoides Sezary syndrome Peripheral T-cell lymphoma, NOS Angioimmunoblastic T-cell lymphoma Anaplastic large cell lymphoma, ALK+ Anaplastic large cell lymphoma, ALK-

Abbreviations: HTLV, human T-cell lymphotropic virus; MALT, mucosa-associated lymphoid tissue; NK, natural killer; WHO, World Health Organization.

Source: Adapted from SH Swerdlow et al: *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*, 5th ed. IARC, 2016.

EPIDEMIOLOGY AND ETIOLOGY

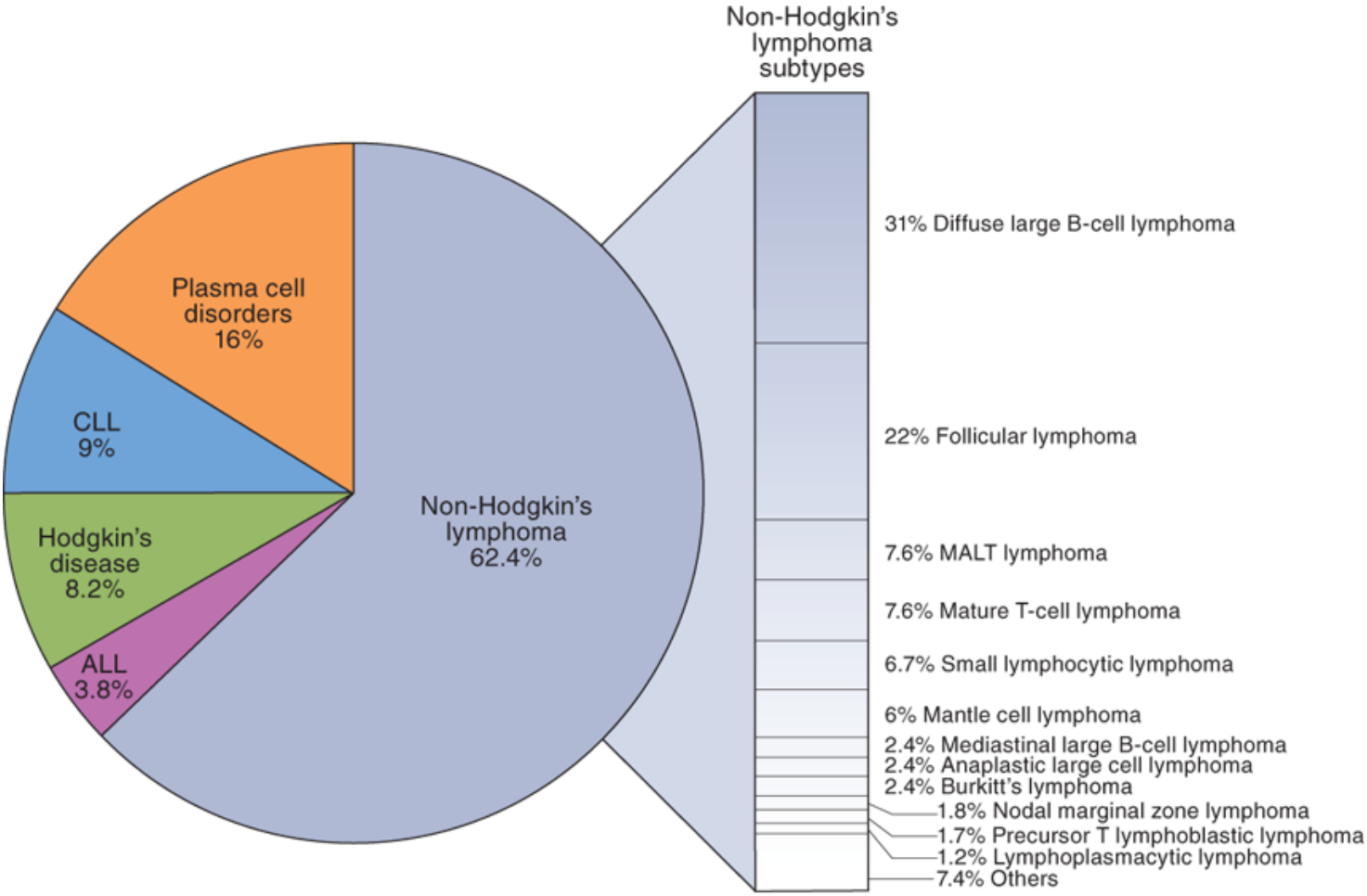
In 2017 over 72,000 new cases of NHL were diagnosed in the United States, about 4% of all new cancers in both males and females making it the eighth and ninth most common cause of cancer-related death in women and men, respectively. The incidence is nearly 10 times the incidence of Hodgkin's lymphoma. There is a slight male-to-female predominance and a higher incidence for Caucasians than for African Americans. The incidence rises steadily with age, especially after age 40, but lymphomas are also among the most common malignancies in adolescent and young adult patients. The incidence of NHL has nearly doubled over the last 20–40 years, and continues to rise by 1.5–2% each year. Patients with both primary and secondary immunodeficiency states are predisposed to developing non-NHL. These include patients with HIV infection; patients who have undergone organ transplantation; and patients with inherited immune deficiencies and autoimmune conditions. The 5-year survival rates for NHL is 72% for Caucasians and 63% for African Americans.

The incidence of NHL and the patterns of expression of the various subtypes differ geographically and across age groups. T-cell lymphomas are more common in Asia than in Western countries, while certain subtypes of B-cell lymphomas such as follicular lymphoma (FL) are more common in Western countries. A specific subtype of non-Hodgkin's lymphoma known as the angiocentric nasal T/natural killer- (NK-) cell lymphoma has a striking geographic occurrence, being most frequent in Southern

Asia and parts of Latin America. Another subtype of non-Hodgkin's lymphoma associated with infection by human T-cell lymphotropic virus (HTLV) 1 is seen particularly in southern Japan and the Caribbean. Likewise, there are differences in the age-dependent incidence of NHL by histologic subtype, with aggressive lymphomas like diffuse large B-cell lymphoma (DLBCL) and Burkitt's lymphoma (BL) being the most common entities in children, and DLBCL and indolent lymphomas including FL being the most common forms in adults. The relative frequencies of the various types of lymphoid malignancies, including Hodgkin's lymphoma, plasma cell disorders, and lymphoid leukemias is shown in **Fig. 104-1.**

FIGURE 104-1

Relative frequency of lymphoid malignancies. ALL, acute lymphoid leukemia; CLL, chronic lymphoid leukemia; MALT, mucosa-associated lymphoid tissue.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A number of environmental factors have been implicated in the occurrence of non-Hodgkin's lymphoma, including infectious agents, chemical exposures, and medical treatments. Several studies have demonstrated an association between exposure to agricultural chemicals and an increased incidence of non-Hodgkin's lymphoma. Patients treated for Hodgkin's lymphoma can develop non-Hodgkin's lymphoma; it is unclear whether this is a consequence of the Hodgkin's lymphoma or its treatment, especially radiation.

Several NHL are associated with infectious agents **(Table 104-2)**. Epstein-Barr Virus (EBV) is associated with the development of Burkitt's lymphoma in Central Africa and the occurrence of aggressive NHL in immunosuppressed patients in Western countries. The majority of primary central nervous system (CNS) lymphomas are associated with EBV. EBV infection is strongly associated with the occurrence of extranodal nasal NK/T-cell lymphomas in Asia and South America. HTLV-1 infects T cells and leads directly to the development of adult T-cell lymphoma (ATL) in a small percentage of patients infected as babies through ingestion of breast milk of infected mothers. The median age of patients with ATL is ~56 years; thus, HTLV-1 demonstrates a long

latency from infection to oncogenesis (Chap. 196). Infection with HIV predisposes to the development of aggressive, B-cell non-Hodgkin's lymphoma. This may be through overexpression of interleukin 6 by infected macrophages. Infection of the stomach by the bacterium *Helicobacter pylori* induces the development of gastric MALT (mucosa-associated lymphoid tissue) lymphomas. This association is supported by evidence that patients treated with antibiotics to eradicate *H. pylori* have regression of their MALT lymphoma. The bacterium does not transform lymphocytes to produce the lymphoma; instead, a vigorous immune response is made to the bacterium, and the chronic antigenic stimulation leads to the neoplasia. MALT lymphomas of the skin may be related to *Borrelia* sp. infections in Europe, those of the eyes to *Chlamydophila psittaci*, and those of the small intestine to *Campylobacter jejuni*. Chronic hepatitis C virus infection has been associated with the development of lymphoplasmacytic lymphoma and splenic marginal zone lymphoma (MZL). Human herpesvirus 8 is associated with primary effusion lymphoma in HIV-infected persons and multicentric Castleman's disease, a diffuse lymphadenopathy associated with systemic symptoms of fever, malaise, and weight loss.

TABLE 104-2

Infectious Agents Associated with the Development of Lymphoid Malignancies

Infectious Agent	Lymphoid Malignancy
Epstein-Barr virus	Burkitt's lymphoma
	Post–organ transplant lymphoma
	Primary CNS diffuse large B-cell lymphoma
	Hodgkin's lymphoma
	Extranodal NK/T-cell lymphoma, nasal type
HTLV-1	Adult T-cell leukemia/lymphoma
HIV	Diffuse large B-cell lymphoma
	Burkitt's lymphoma
Hepatitis C virus	Lymphoplasmacytic lymphoma
Helicobacter pylori	Gastric MALT lymphoma
Human herpesvirus 8	Primary effusion lymphoma
	Multicentric Castleman's disease

Abbreviations: CNS, central nervous system; HIV, human immunodeficiency virus; HTLV, human T cell lymphotropic virus; MALT, mucosa-associated lymphoid tissue; NK, natural killer.

In addition to infectious agents, a number of other diseases or exposures may predispose to developing lymphoma (Table 104-3). Diseases of inherited and acquired immunodeficiency as well as autoimmune diseases are associated with an increased incidence of lymphoma. The association between immunosuppression and induction of NHLs is compelling since if the immunosuppression can be reversed, a percentage of these lymphomas regress spontaneously. The incidence of NHL is nearly

hundredfold increased for patients undergoing organ transplantation necessitating chronic immunosuppression, and is greatest in the first year post-transplant. About 30% of these arise as a polyclonal B-cell proliferation that evolves into a clonal B-cell malignancy. The NHLs that occur in the context of immunosuppression or immunodeficiency, including HIV infection, are frequently associated with EBV. Histologically, DLBCLs are most frequently associated with immunosuppression and autoimmune diseases, although almost all histologies can be seen, especially MALT lymphomas in the context of autoimmune diseases like Sjogren's and Hashimoto's thyroiditis. The rare inherited immunodeficiency diseases X-linked lymphoproliferative syndrome, Wiskott-Aldrich syndrome, Chédiak-Higashi syndrome, ataxia telangiectasia, and common variable immunodeficiency syndrome are complicated by highly aggressive lymphomas. The elevated incidence of lymphoma in iatrogenic immunosuppression, AIDS, and autoimmune disease argues strongly for immune dysregulation contributing in the pathogenesis of some lymphomas. An increased risk of NHL has been observed in first-degree relatives with NHL, Hodgkin's lymphoma or chronic lymphocytic leukemia (CLL). In large databases studies, about 9% of patients with lymphoma or CLL have a first-degree relative with a lymphoproliferative disorder.

TABLE 104-3

Diseases or Exposures Associated with Increased Risk of Development of Malignant Lymphoma

Inherited immunodeficiency disease	Autoimmune disease
Klinefelter's syndrome	Sjögren's syndrome
Chédiak-Higashi syndrome	Celiac sprue
Ataxia-telangiectasia syndrome	Rheumatoid arthritis and systemic lupus erythematosus
Wiskott-Aldrich syndrome	Chemical or drug exposures
Common variable immunodeficiency disease	Phenytoin
Acquired immunodeficiency diseases	Dioxin, phenoxy herbicides
Iatrogenic immunosuppression	Radiation
HIV-1 infection	Prior chemotherapy and radiation therapy
Acquired hypogammaglobulinemia	

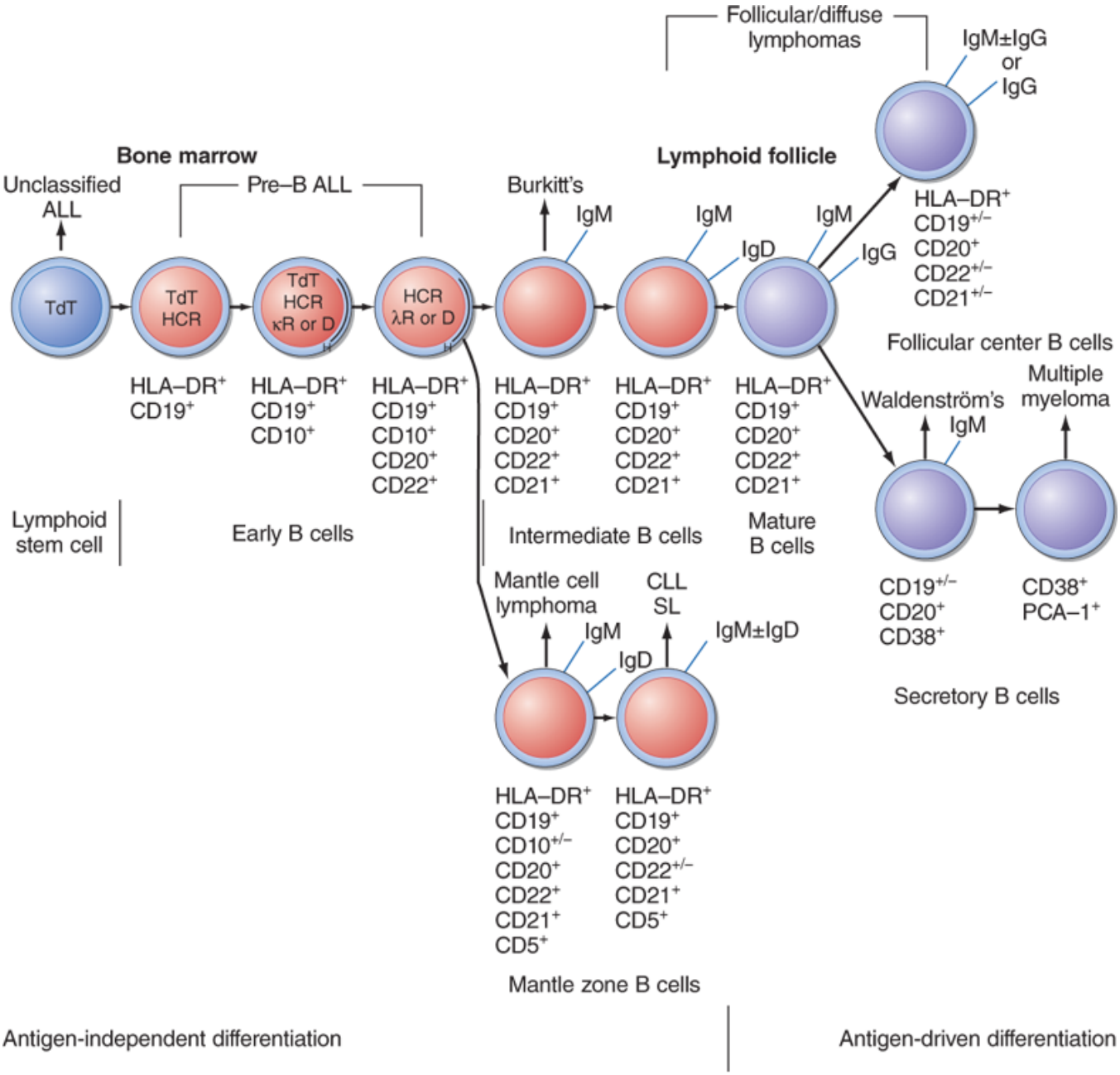
IMMUNOLOGY

All lymphoid cells are derived from a common hematopoietic progenitor that gives rise to lymphoid, myeloid, erythroid, monocyte, and megakaryocyte lineages. Through the ordered and sequential activation of a series of transcription factors, the cell first becomes committed to the lymphoid lineage and then gives rise to B and T cells.

About 90% of all lymphomas are of B-cell origin. A cell becomes committed to B-cell development when it expresses the master B lineage transcription factor PAX5, which ultimately results in a transcriptional program that leads to the rearrangement of its immunoglobulin genes, which involves chromosomal recombination as well as somatic hypermutation to create an immunoglobulin gene that is unique to that B cell. The sequence of cellular changes, including changes in cell-surface phenotype that characterizes normal B-cell development is shown in **Fig. 104-2.** Most B-cell lymphomas arise following the process of immunoglobulin gene recombination and somatic hypermutation, which leads to class switching and affinity maturation of the mature immunoglobulin, respectively, suggesting that it is the error-prone nature of these genetic events that contributes to oncogenesis. Certainly the frequency of chromosomal translocations that result in the activation of an oncogene or the inactivation of a tumor suppressor gene in B-cell NHL may be the result of these normal cellular processes gone awry (see below). In addition, the key roles of the transcription factors MYC and BCL6 and the anti-apoptotic protein BCL2 in the process of B-cell development explain why the genes encoding these proteins are commonly mutated in B-cell lymphomas.

FIGURE 104-2

Pathway of normal B-cell differentiation and relationship to B-cell lymphomas. HLA-DR, CD10, CD19, CD20, CD21, CD22, CD5, and CD38 are cell markers used to distinguish stages of development. Terminal transferase (TdT) is a cellular enzyme. Immunoglobulin heavy chain gene rearrangement (HCR) and light chain gene rearrangement or deletion (κR or D, λR or D) occur early in B-cell development. The approximate normal stage of differentiation associated with particular lymphomas is shown. ALL, acute lymphoid leukemia; CLL, chronic lymphoid leukemia; SL, small lymphocytic lymphoma.



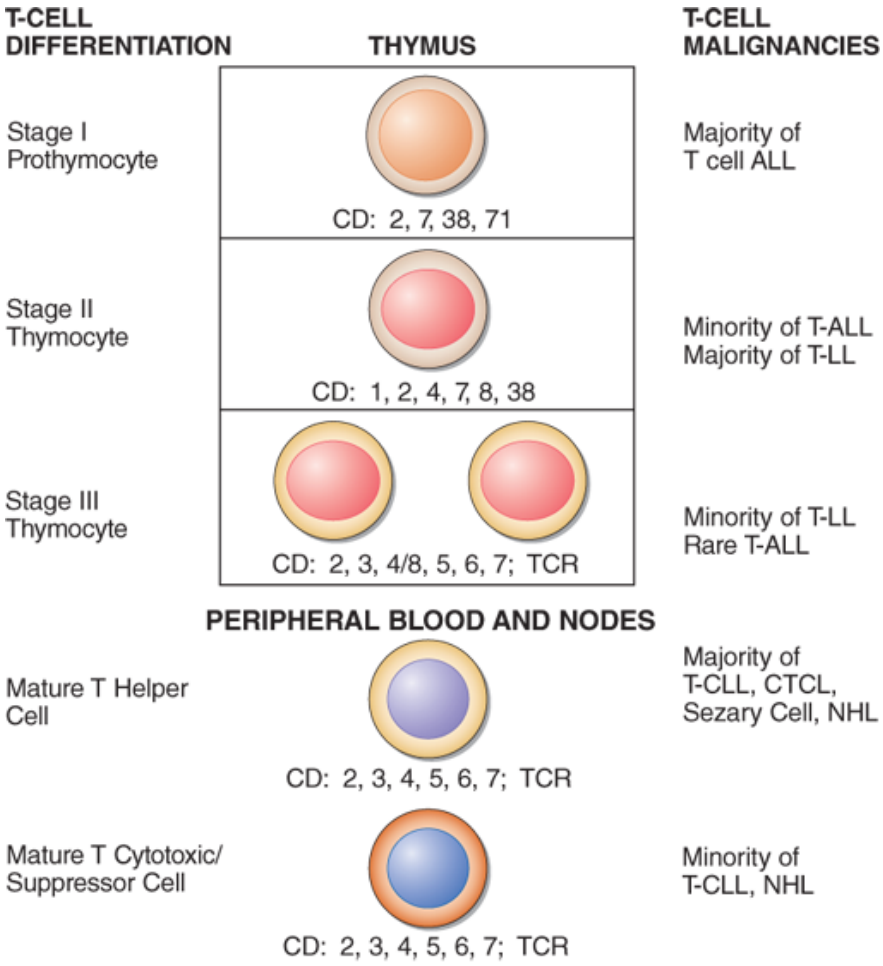
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A cell becomes committed to T-cell differentiation upon migration to the thymus and rearrangement of T-cell receptor (TCR) genes. This requires the expression of the T-cell master regulatory transcription factor, NOTCH-1. As in B cells, the development of the mature TCR involves the rearrangement and recombination of the TCR loci, which is error-prone and potentially oncogenic. The sequence of the events that characterize T-cell development is depicted in **Fig. 104-3.**

FIGURE 104-3

Pathway of normal T-cell differentiation and relationship to T-cell lymphomas. CD1, CD2, CD3, CD4, CD5, CD6, CD7, CD8, CD38, and CD71 are cell markers used to distinguish stages of development. T-cell antigen receptors (TCR) rearrange in the thymus,

and mature T cells emigrate to nodes and peripheral blood. ALL, acute lymphoid leukemia; T-ALL, T-cell ALL; T-LL, T-cell lymphoblastic lymphoma; T-CLL, T-cell chronic lymphoid leukemia; CTCL, cutaneous T-cell lymphoma; NHL, non-Hodgkin's lymphoma.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although lymphoid malignancies often retain the cell-surface phenotype of lymphoid cells at particular stages of differentiation, this information is of little clinical or prognostic consequence. The so-called stage of differentiation of a malignant lymphoma does not predict its natural history. The antigen footprint, or immunophenotype, of the cell, however, is valuable diagnostically as it allows for the distinguishing of specific NHL subtypes. It can be detected by flow cytometry of single cell suspension from blood, bone marrow, body fluid or disaggregated tissue using fluorescently labeled antibodies against these antigens, or by immunohistochemical staining of paraffin-embedded tissue sections with enzyme-linked antibodies against these antigens followed by a colorimetric reaction.

As already mentioned, malignancies of lymphoid cells are associated with recurring genetic abnormalities including chromosomal translocations and genetic mutations that may in part be the result of aberrant immunoglobulin or TCR development. While specific genetic abnormalities have not been identified for all subtypes of lymphoid malignancies, it is presumed that they exist. As previously discussed, B cells are even more susceptible to acquiring mutations during their maturation in germinal centers; the generation of antibody of higher affinity requires the introduction of mutations into the variable region genes in the germinal centers. Given this, other nonimmunoglobulin genes, e.g., *bcl-6*, may acquire mutations as well. Likewise, many lymphomas contain balanced chromosomal translocations involving the antigen receptor genes; immunoglobulin genes on chromosomes 2, 14, and 22 in B cells; and T-cell antigen receptor genes on chromosomes 7 and 14 in T cells. The rearrangement of chromosome segments to generate mature antigen receptors must create a site of vulnerability to aberrant recombination. Examples of this type of event include the (8;14)(q24;q32) translocation in BL, involving the *MYC* proto-

oncogene and the IgH gene; the (14;18)(q32;q32) translocation in FL, involving the *BCL2* proto-oncogene and the IgH gene; and the (11;14) (q13;q32) translocation in mantle cell lymphoma (MCL), involving the gene encoding cyclin D1 (*CCDN1*) and the IgH gene. Less commonly, chromosomal translocations produce fusion genes that encode chimeric oncogenic proteins. Examples of this include the (2;5)(p23;q35) translocation involving the *ALK* and *NPM1* genes in anaplastic large cell lymphoma (ALCL) and the t(11;18)(q21;q21) translocation involving the *API2* and *MLT* genes in MALT lymphoma. Table 104-4 presents the most common translocations and associated oncogenes for various subtypes of lymphoid malignancies.

TABLE 104-4

Genetic Features of B- and T-Cell Lymphomas

Genetic feature	Genes	Lymphoma
t(8;14) **t(2;8)** **t(8;22)**	*MYC/IgH* *MYC/Igκ* *MYC/Igλ*	Burkitt's lymphoma
t(11;14)	*BCL1 (CCND1)/IgH*	Mantle cell lymphoma; multiple myeloma
t(14;18) **t(3;14)**	*BCL2/IgH* *BCL6/IgH*	Follicular lymphoma, diffuse large B-cell lymphoma (DLBCL)
t(11;18) **t(1;14)** **t(14;18)** **t(3;14)**	*API2/MALT1* *BCL10/IgH* *MALT1/IgH* *FOXP1/IgH*	MALT lymphoma
Trisomy 3 **7q21 deletion**	*Unknown* *CDK6*	Splenic marginal zone lymphoma
t(9;14) **6q21 deletion**	*PAX5/IgH* *Unknown*	Lymphoplasmacytic lymphoma
inv(14) **t(14;14)**	*TCRα/TCL1*	Peripheral T-cell lymphoma, NOS; T-PLL
t(2;5) **t(1;2)** **t(2;3)** **t(2;17)** **inv(2)**	*NPM1/ALK* *TPM3/ALK* *TFG/ALK* *CTLC/ALK* *ATIC/ALK*	Anaplastic large cell lymphoma (ALCL)
Trisomy 3 **Trisomy 5**	*Unknown* *Unknown*	Angioimmunoblastic T-cell lymphoma
Isochromosome 7q	*Unknown*	Hepatosplenic T-cell lymphoma

Abbreviation: MALT, mucosa-associated lymphoid tissue.

Gene profiling using array technology allows the simultaneous assessment of the expression of thousands of genes. This technology provides the possibility to identify new genes with pathologic importance in lymphomas, the identification of patterns of gene expression with diagnostic and/or prognostic significance, and the identification of new therapeutic targets. Recognition of patterns of gene expression is complicated and requires sophisticated mathematical techniques. Early successes using this technology in lymphoma include the identification of previously unrecognized subtypes of DLBCL whose gene expression patterns resemble either those of follicular, or germinal center B (GCB) cells or activated peripheral blood B cells (ABC). Patients whose lymphomas have a GCB-like pattern of gene expression have a considerably better prognosis than those whose lymphomas have a pattern resembling ABCs. This improved prognosis is independent of other known prognostic factors. Similar information is being generated in FL and MCL. The challenge remains to provide information from such techniques in a clinically useful time frame.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Regardless of the type of lymphoid malignancy, the initial evaluation of the patient should include performance of a careful history and physical examination. These will help confirm the diagnosis, identify those manifestations of the disease that might require prompt attention, and aid in the selection of further studies to optimally characterize the patient's status to allow the best choice of therapy. It is difficult to overemphasize the importance of a carefully done history and physical examination. They might provide observations that lead to reconsidering the diagnosis, provide hints at etiology, clarify the stage, and allow the physician to establish rapport with the patient that will make it possible to develop and carry out a therapeutic plan.

The duration of symptoms and pace of symptomatic progression are important in distinguishing aggressive from more indolent lymphomas, as are the presence or absence of "B" symptoms, such as fevers, night sweats, or unexplained weight loss. Patients should be asked about localizing symptoms that may point towards lymphomatous involvement of specific sites, such as the chest, abdomen, or CNS. Comorbid diagnoses that may impact therapy or monitoring on therapy should be reviewed and acknowledged, including a history of diabetes or congestive heart failure. A physical examination should pay close attention to all the peripherally accessible sites of lymph nodes, the liver and spleen size, Waldeyer's ring, whether there is a pleural or pericardial effusion or abdominal ascites, whether there is an abdominal, testicular or breast mass, and whether there is cutaneous involvement as all of these findings may influence further evaluation and disease management.

Laboratory studies should include a complete blood count, routine chemistries, liver function tests, and serum protein electrophoresis to document the presence of circulating monoclonal paraproteins. The serum β-2 microglobulin level and serum lactate dehydrogenase (LDH) are important independent prognostic factors in NHL. Staging of certain diseases may involve a bone marrow biopsy; results of other laboratory and staging studies may also warrant a marrow evaluation. A lumbar puncture for evaluation of lymphomatous involvement may be indicated in the setting of concerning neurologic signs or symptoms, or diseases that are high risk for CNS involvement. The latter may include disease involving the paranasal sinuses, testes, breast, kidneys, adrenal glands, and epidural space, as well as highly aggressive histologies like BL. Since HIV and hepatitis B and C infection can be risk factors for developing NHL, and since treatment for some NHL can result in the potentially life threatening reactivation of hepatitis B, patients with a new diagnosis of NHL should be screened for these viruses as well.

Lymphoma histology and clinical presentation dictates which imaging studies should be ordered. Chest, abdominal, and pelvic computed tomography (CT) scans are essential for accurate staging to assess lymphadenopathy for indolent lymphomas, whereas positron emission tomography (PET) using ^{18}F-fluorodeoxyglucose (FDG PET) is useful for aggressive lymphomas, including BL, DLBCL, plasmablastic lymphoma, and the aggressive T-cell NHLs. It is highly sensitive for detecting both nodal and extranodal sites involved by NHL. The intensity of FDG avidity, or SUV, correlates with histologic aggressiveness, and may be useful in cases when disease transformation of an indolent lymphoma to a diffuse aggressive lymphoma is suspected. PET

scanning can also differentiate between treated disease and active disease at the end of therapy in patients with residual masses on CT scans. Consensus recommendations regarding PET scanning were published as a result of an International Harmonization Project, and state that PET should only be used for DLBCL and Hodgkin's lymphoma, that scanning during therapy should only be done as part of clinical trials, and that the end-of-treatment scan should not be done before 3 weeks but preferably 6–8 weeks after chemotherapy and 8–12 weeks after radiation or chemoradiotherapy. There is no evidence that long-term follow-up should include PET scanning. More recently, though, PET scan results at the end of therapy for FL have been associated with prognosis, with patients with residual PET avid disease at the end of treatment having a poorer prognosis than those who are PET negative, and so it may be used for this prognostic purpose. Finally, magnetic resonance imaging (MRI) is useful in detecting bone, bone marrow and CNS disease in the brain and spinal cord. The staging evaluation is outlined in **Table 104-5**.

The Ann Arbor staging system developed in 1971 for HL was adapted for staging NHLs (**Table 104-6**). This staging system focuses on the number of tumor sites (nodal and extranodal), location, and the presence or absence of systemic, or B, symptoms. Table 104-6 summarizes the essential features of the Ann Arbor system.

This anatomic based system is less useful in NHL, which disseminates widely, not in an ordered stepwise fashion. A majority of patients with NHL have advanced stage disease at diagnosis. Apart from early-stage disease limited to a radiation field where local therapy with radiation is an option, all other disease is treated the same regardless of stage. Histology and clinical parameters at presentation are more important than stage with respect to prognosis. The International Prognostic Index (IPI) is perhaps the best predictor of outcome (**Table 104-7**). The IPI was developed based on the analysis of over 2000 patients with aggressive NHLs treated with an anthracycline containing regimen. Age (≤60 vs >60); serum LDH (≤normal vs >normal); performance status (0 or 1 vs 2–4); stage (I or II vs III or IV); and extranodal involvement (<1 site vs >1 site) were identified as independently prognostic for overall survival (OS). A point is awarded for each risk factor and then summed, defining four risk groups: low-risk (0 or 1); low-intermediate (2); high-intermediate (3); and high (4–5). The 5-year OS rates for patients with scores of 0 to 1, 2, 3, and 4–5 were 73, 51, 43, and 26%, respectively. The age-adjusted IPI separates patients ≤60 from patients >60. For the age adjusted IPI, only stage, LDH, and performance status were important. Younger patients with 0, 1, 2, or 3 risk factors had 5-year survival rates of 83%, 69%, 46%, and 32%, compared to 56%, 44%, 37%, and 21% for older patients. When factoring in the introduction and clinical benefit of rituximab, the 4-year progression-free survival is 94%, 80%, and 53% for 0 and 1, 2, or 3 or more risk factors, respectively.

The follicular lymphoma prognostic index (FLIPI) is a similar predictive model for FL, derived from the analysis of over 4000 patients. Age >60, stage III/IV disease, the presence of >4 nodal sites, an elevated serum LDH concentration and a hemoglobin <12 were identified as independent prognostic variables, and summation of each variable identified three risk groups. The median 10-year survival rates for patients with zero to one (low-risk), two (intermediate-risk), or three or more (high-risk) of these adverse factors were 71, 51, and 36%, respectively. Similar disease-specific IPIs have been developed for MCL and peripheral T-cell lymphoma (PTCL) as well. These prognostic indices take into account the proliferative index and cell surface markers, respectively.

Finally, as mentioned previously, gene expression profiling has identified DLBCLs with differential prognoses: GCB and ABC, where GCB-like DLBCL is associated with a significantly better OS. A more readily accessible immunohistochemical algorithm has been developed, based on the presence of absence of CD10, BCL6, and MUM1 that correlates closely with gene expression profiles and can differentiate the majority of GCB from non-GCB-like DLBCL. These profiles have prognostic importance but to date do not alter treatment recommendations for the primary treatment of DLBCL. Current clinical trials do stratify by DLBCL subtype, and it appears that agents like the Bruton's tyrosine kinase (BTK) inhibitor ibrutinib and lenalidomide are most active in non-GCB DLBCL in the relapsed setting. Treatment may then be differentiated by these subtypes in the future.

TABLE 104-5

Staging Evaluation for Non-Hodgkin's Lymphoma

Physical examination
Documentation of B symptoms
Laboratory evaluation
Complete blood counts
Liver function tests
Uric acid
Calcium
Serum protein electrophoresis
Serum β_2-microglobulin
Chest radiograph
CT scan of abdomen, pelvis, and usually chest
Bone marrow biopsy
Lumbar puncture in lymphoblastic, Burkitt's, and diffuse large B-cell lymphoma with positive marrow biopsy
Gallium scan (SPECT) or PET scan in large cell lymphoma

Abbreviations: CT, computed tomography; PET, positron emission tomography; SPECT, single-photon emission computed tomography.

TABLE 104-6

Ann Arbor Staging for Lymphoma*

Stage	Description
I	Involvement of a single lymph node region (I) or single extra-nodal site (IE)
II	Involvement of two or more lymph node regions or lymphatic structures on the same side of the diaphragm alone (II) or with involvement of limited, contiguous, extralymphatic organ or tissue (IIE)
III	Involvement of lymph node regions on both sides of the diaphragm (III), which may include the spleen (IIIS), or limited, contiguous, extralymphatic organ or tissue (IIIE), or both (IIIES)
IV	Diffuse or disseminated foci of involvement of one or more extralymphatic organs or tissues, with or without associated lymphatic involvement

*All stages are further subdivided according to the absence (A) or presence (B) of systemic B symptoms including fevers, night sweats, and/or weight loss (>10% of body weight over 6 months prior to diagnosis).

TABLE 104-7

International Prognostic Index for NHL

Five Clinical Risk Factors	
Age ≥60 years	
Serum lactate dehydrogenase levels elevated	
Performance status ≥2 (ECOG) or ≤70 (Karnofsky)	
Ann Arbor stage III or IV	
>1 site of extranodal involvement	
For Diffuse Large B-cell Lymphoma	
0, 1 factor = low risk	35% of cases; 5-year survival, 73%
2 factors = low-intermediate risk	27% of cases; 5-year survival, 51%
3 factors = high-intermediate risk	22% of cases; 5-year survival, 43%
4, 5 factors = high risk	16% of cases; 5-year survival, 26%
For Diffuse Large B-cell Lymphoma Treated With R-CHOP	
0 factor = good	10% of cases; 4-year survival, 94%
1, 2 factors = intermediate	45% of cases; 4-year survival, 80%
3, 4, 5 factors = poor	45% of cases; 4-year survival, 53%

Abbreviations: ECOG, Eastern Cooperative Oncology Group; R-CHOP, rituximab, cyclophosphamide, doxorubicin, vincristine, prednisone.

CLINICAL FEATURES, TREATMENT, AND PROGNOSIS OF SPECIFIC NHL

MATURE B-CELL NEOPLASMS

B-cell NHLs can be characterized into two broad groups—those that behave aggressively, require immediate or urgent treatment with combination chemotherapy regimens, and are potentially curable, and those that are more indolent in nature, can be observed and treated only when they cause symptoms or signs of organ function impairment, are very responsive to therapy, but are not ultimately curable in the vast majority of cases. Among the aggressive diseases, the most common are NHL and DLBCL; and the most rapidly prolific are NHL and BL. FL is the second most common NHL and the most common indolent NHL. Other indolent NHLs include MZL, lymphoplasmacytic lymphoma (LPL), and hairy cell leukemia (HCL). MCL is an intermediate grade lymphoma that shares some characteristics with the aggressive lymphomas (fairly urgent need for

treatment and aggressive upfront combination chemotherapy regimens), but like the indolent lymphomas, it is not readily curable with conventional dose therapies.

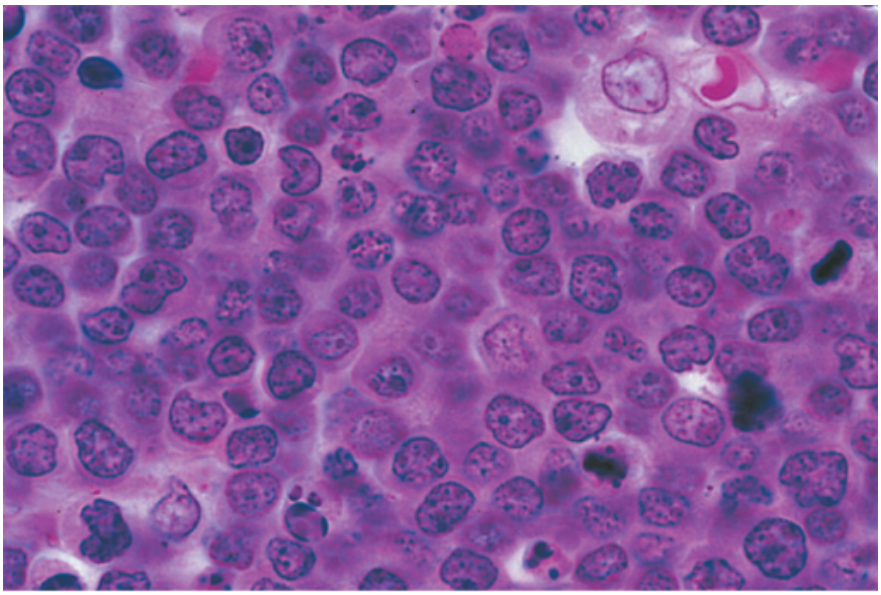
Burkitt's Lymphoma

Burkitt's lymphoma/leukemia is a rare disease in adults in the United States, making up <1% of NHL, but it makes up ~30% of childhood non-Hodgkin's lymphoma. It is one of the fastest growing neoplasms, with a doubling time of <24 h. In general it is a pediatric tumor that has three major clinical presentations. The endemic (African) form presents as a jaw or facial bone tumor that spreads to extranodal sites including ovary, testis, kidney, breast, and especially to the bone marrow and meninges. The non-endemic form has an abdominal presentation with massive disease, ascites, and renal, testis, and/or ovarian involvement, and, like the endemic form, also spreads to the bone marrow and CNS. Immunodeficiency-related cases more often involve lymph nodes and may present as acute leukemia. BL has a male predominance and is typically seen in patients <35 years of age.

On biopsy, there is a monotonous infiltration of medium-sized cells with round nuclei, multiple nucleoli, and basophilic cytoplasm with vacuoles. The proliferation rate is ~100%, and tingible body macrophages give rise to the classic "starry sky" appearance of this tumor **(Fig. 104-4)**. Tumor cells are positive for B-cell antigens CD19, CD20, and surface immunoglobulin. They are also uniformly positive for CD10 and BCL6 but negative for BCL2. Endemic BLs are EBV positive, whereas the majority of non-endemic BLs are EBV negative. BL is associated with a translocation involving MYC on chromosome 8q24 in >95% of the cases. The most common partners are chromosomes 14, 2, or 22, rearrangements that produce fusions of MYC with either the IgH (80%), kappa (15%), or lambda (5%) light chain genes, respectively.

FIGURE 104-4

Burkitt's lymphoma. The neoplastic cells are homogeneous, medium-sized B cells with frequent mitotic figures, a morphologic correlate of high growth fraction. Reactive macrophages are scattered through the tumor, and their pale cytoplasm in a background of blue-staining tumor cells gives the tumor a so-called starry sky appearance.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

While exquisitely chemosensitive, it is imperative that treatment for BL be initiated quickly given the rapid doubling time and high morbidity of this disease. There are several effective intensive combination chemotherapy regimens, all of which incorporate high doses of cyclophosphamide. Prophylactic therapy to the CNS is mandatory. Cure can be expected in 70–80% of patients when treated promptly and correctly. Salvage therapy has been generally ineffective in patients whose disease

progresses after upfront therapy, emphasizing the importance of the initial treatment approach and referral to a tertiary cancer center with experience treating this disease.

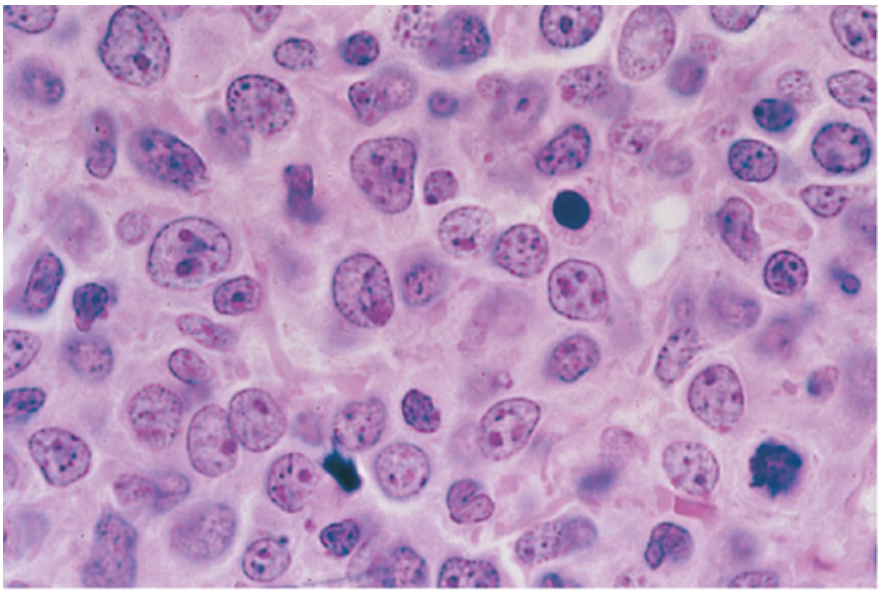
Diffuse Large B-Cell Lymphoma

Diffuse large B-cell lymphoma (DLBCL) is the most common histologic subtype of NHL diagnosed, representing about one-third of all cases. Previously felt to be “one disease” it is now recognized as a heterogeneous collection of multiple entities. It is slightly more common in Caucasians and men, and the median age at diagnosis is 64. The relative risk of DLBCL is higher amongst people with affected first-degree relatives (RR 3.5-fold), and patients with congenital or acquired immunodeficiency, patients on immunosuppression, and patients with autoimmune disorders also have a higher risk of developing DLBCL, often EBV-related. The majority of patients present with advanced stage disease, with only 30–40% of patients having stage I or II disease; about 40% of patients will have “B” symptoms, and 50% of patients will have an elevated LDH. Up to 40% of patients will have involvement of non-lymph node sites including bone marrow, CNS, GI track, thyroid, liver, and skin. Patients with extensive bone marrow involvement, or involvement of the testes, breast, kidney, adrenal gland, paranasal sinus, or epidural space are at increased risk of CNS dissemination.

The tumor consists of a diffuse proliferation of large, atypical lymphocytes with a high proliferative index **(Fig. 104-5)**. These cells typically express the B-cell antigens CD19, CD20, and CD79a. Expression of CD10 and BCL6 is consistent with the tumor cell being of germinal center origin (GCB), while the expression of MUM1 corresponds with the non-GC or activated B cell (ABC) subtype. BCL2 is overexpressed in anywhere from 25 to 80% of DLBCL, whereas BCL6 is positive in more than two-thirds of cases, either as the result of translocations, gain of copy number, or promoter mutations. MYC is rearranged in 10% of DLBCLs, and ~20% of MYC-rearranged cases have concurrent BCL2 or BCL6 rearrangements, a combination referred to as “double- hit lymphoma.” These double-hit lymphomas are associated with an extremely poor prognosis with a median OS of only 12–18 months. Amplification and/or overexpression of MYC independent of rearrangements or amplification has also been described and is also associated with a poor, albeit better, prognosis.

FIGURE 104-5

Diffuse large B-cell lymphoma. The neoplastic cells are heterogeneous but predominantly large cells with vesicular chromatin and prominent nucleoli.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Combination chemotherapy offers potentially curative therapy for DLBCL, regardless of the stage. The addition of the anti-CD20 antibody rituximab to cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) improved survival beyond CHOP

alone and is the standard first-line chemotherapy for this disease. For patients with early stage disease localized to a radiation field, treatment options include full course chemotherapy with R-CHOP every 3 weeks for 6 cycles, or abbreviated chemotherapy for 3–4 cycles followed by involved field radiotherapy. For advanced stage DLBCL, therapy is with a full course of chemotherapy. On average, about 60–65% of patients with DLBCL can be expected to be cured with this approach, and the likelihood of cure is predicted by the IPI, gene expression profile cell of origin, and/or MYC cytogenetics and expression. Several studies have investigated alternative anthracycline-containing chemotherapy regimens and/or consolidation autologous stem cell transplantation in first remission for higher-risk disease without improvement over R-CHOP alone. Dose adjusted R-EPOCH (rituximab, infusional etoposide/vincristine/adriamycin, cyclophosphamide, prednisone) is one such regimen. Although this regimen is no better than R-CHOP for DLBCL in general, it is often used to treat primary mediastinal large B-cell lymphoma and double-hit DLBCL based on results from phase 2 and retrospective studies, respectively. CNS prophylaxis with either intrathecal chemotherapy or high-dose systemic methotrexate and leucovorin rescue should be considered for patients with high risk of CNS dissemination. This includes patients with primary testicular involvement and breast involvement, as well as patients with several IPI risk factors and diffuse bone marrow involvement, renal involvement, or adrenal involvement. The use of CNS prophylaxis for disease involving the paranasal sinuses or the epidural space is less clear but may be considered.

Over one-third of patients will either have primary refractory disease or disease that relapses after first-line chemotherapy. These patients may still be cured with salvage chemotherapy regimens followed by autologous stem cell transplantation. Patients with a poor performance status or advanced age that are not candidates for such an approach, however, are often managed with palliative intentions. Radiation to symptomatic areas of disease can be transiently helpful. Less intensive chemotherapy with drugs like gemcitabine, cytarabine, or bendamusine can help control disease and symptoms for a limited period of time. These patients should be referred for clinical trials when applicable. For patients in whom more aggressive therapy is an option, treatment is with combination chemotherapy using various combinations of drugs primarily in order to identify patients with chemosensitive disease. Patients with chemosensitive disease have the greatest likelihood of benefiting from high-dose chemotherapy and autologous stem cell transplant, which improves response duration and survival over salvage chemotherapy alone and leads to long-term disease free survival in about 40–50% of patients. For patients with chemorefractory disease, clinical trials or palliative therapy or clinical trials should be considered, with a goal of achieving a disease response sufficient for allogeneic stem cell transplant. Several new agents have shown some promise in patients with relapsed DLBCL, including ibrutinib, particularly in the ABC cell of origin subtype, lenalidomide, and everolimus.

Chimeric antigen receptor T cells (CAR-T cells) are an investigational immunotherapy approach for treating malignancies that have had early success in CLL and B-cell acute lymphoblastic leukemia, as well as B-cell NHL. This strategy uses T cells collected from a patient that are genetically modified to express a receptor that will bind to a surface antigen expressed on the patient's own tumor cells. In the case of B-cell malignancies, CD19 has been targeted most commonly. After infusion, autologous CAR-T cells home to sites of disease and also persist over time. The CARs consist of an extracellular antigen recognition domain (typically a single chain Fv variable fragment from a monoclonal antibody) linked via a transmembrane domain to an intracellular signaling domain (usually the CD3ζ endodomain), resulting in the redirection of T-cell specificity toward target antigen-positive cells, and one or more costimulatory domains including CD28, 4-1BB, or OX40 to enhance cytokine secretion and effector cell expansion, and prevent activation-induced apoptosis and immune suppression by tumor-related metabolites. Anti-CD19 CAR-T cells have been approved for the treatment of relapsed/refractory DLBCL following two prior systemic therapies. This would include patients with chemoinsensitive disease following second-line salvage chemotherapy for whom autologous stem cell transplant is not an option, or for patients who relapse after autologous stem cell transplant. In this setting, the response rate of CAR-T cells is over 80%, with over 50% of patients achieving a complete response. These responses appear to be durable, with 70% of complete responders still in remission past 1 year of therapy.

Other large B-cell lymphomas include intravascular large B-cell lymphoma, T-cell/histiocyte–rich large B-cell lymphoma, EBV-positive DLBCL of the elderly, and ALK-positive large B-cell lymphoma. Patients with the latter two diseases tend to have a poor prognosis, whereas the addition of rituximab to CHOP chemotherapy has dramatically improved outcomes with intravascular

large B-cell lymphoma, and the outcomes in T-cell/histiocyte-rich large B-cell lymphoma are similar to DLBCL. R-CHOP remains the treatment of choice for each of these lymphomas.

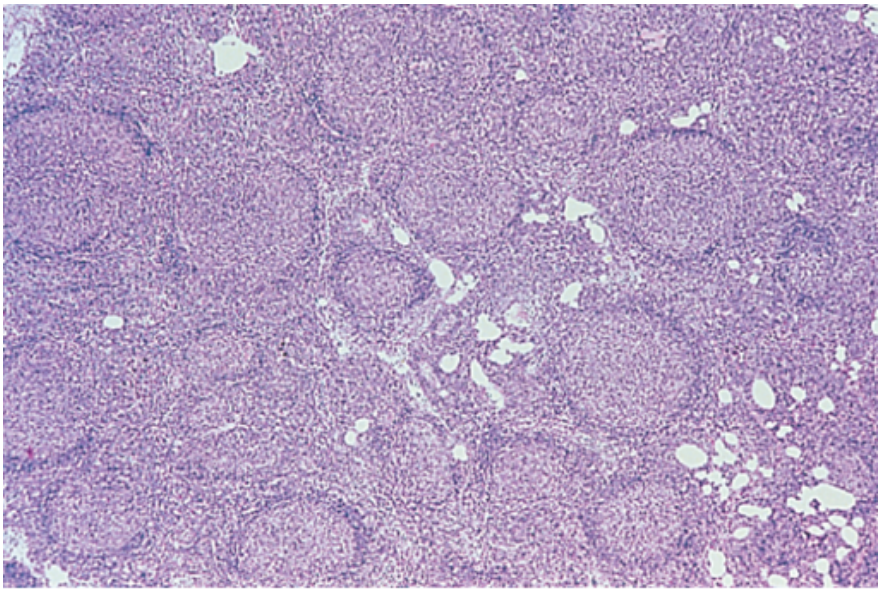
Follicular Lymphoma

FLs are the second leading NHL diagnosis in the United States and Europe and makes up 22% of NHL worldwide and at least 30% of NHL diagnosed in the United States. This type of lymphoma can be diagnosed accurately on morphologic findings alone and has been the diagnosis in the majority of patients in therapeutic trials for “low-grade” lymphoma in the past.

Evaluation of an adequate biopsy by an expert hematopathologist is sufficient to make a diagnosis of FL. The tumor is composed of small cleaved and large cells in varying proportions organized in a follicular pattern of growth **(Fig. 104-6)**. Confirmation of B-cell immunophenotype (monoclonal immunoglobulin light chain, CD19, CD20, CD10 and BCL6 positive, and CD5 and CD23 negative) and the existence of the t(14;18) and abnormal expression of BCL-2 protein are confirmatory. While >85% of FL will harbor a t(14;18) and overexpress the antiapoptotic protein BCL2, this genetic event is necessary but not sufficient for malignant transformation of the B lymphocytes and multiple genetic events are required for the development of FL. Studies have identified the most common recurrent genetic events in FL, and they included mutations in several epigenetic modifying genes, including *MLL2, EZH2, CREBBP*, and *EP300*. The major differential diagnosis is between lymphoma and reactive follicular hyperplasia. The coexistence of DLBCL must be considered. Patients with FL are often subclassified, or graded, into those with predominantly small cells, those with a mixture of small and large cells, and those with predominantly large cells. The WHO Classification adopted grading from 1 to 3 based on the number of centroblasts, or large cells, counted per high power field (hpf): grade I, from 0 to 5 centroblasts/hpf; grade II, from 6 to 15 centroblasts/hpf; and grade III, >15 centroblasts/hpf. Grade III has been subdivided into grade IIIa, in which centrocytes predominate, and grade IIIb, in which there are sheets of centroblasts. While this distinction cannot be made simply or very reproducibly, these subdivisions do have prognostic significance. Patients with FL with predominantly large cells have a higher proliferative fraction, progress more rapidly, and have a shorter OS with simple chemotherapy regimens. Grade IIIb FL is an aggressive disease and considered most similar to DLBCL and treated as such with curative intent.

FIGURE 104-6

Follicular lymphoma. The normal nodal architecture is effaced by nodular expansions of tumor cells. Nodules vary in size and contain predominantly small lymphocytes with cleaved nuclei along with variable numbers of larger cells with vesicular chromatin and prominent nucleoli.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The most common presentation for FL is with new, painless lymphadenopathy. Multiple sites of lymphoid involvement are typical, and unusual sites such as epitrochlear nodes are sometimes seen. However, essentially any organ can be involved, and extranodal presentations do occur. Most patients do not have an elevated LDH or fevers, night sweats, or weight loss, although histologic transformation to DLBCL does occur at a rate of ~3% per year and can be associated with these signs/symptoms. As discussed previously, prognosis is best predicted by the FLIPI. Staging is typically done with CT scans of the chest, abdomen and pelvis, as well as the neck if neck disease is suspected, although PET/CT scans can be helpful in cases where disease transformation is suspected, as transformed disease will be more FDG avid than indolent disease, or for confirmation of early-stage disease, where definitive local therapy with radiation may be considered.

Although FL is highly sensitive to chemotherapy and radiotherapy, these therapies are usually not ultimately curative, except in the setting of early-stage disease. If the disease can be encompassed in a radiation field, involved field radiotherapy at a dose of 24–30 Gy may be curative, with 5-, 10-, and 15-year freedom from treatment failure of 72%, 46%, and 39%, and an overall 5-, 10-, and 15-year survival rates of 93%, 75%, and 62%, respectively. If radiation therapy would not be tolerated, or if a patient prefers not to receive radiation, observation is a reasonable alternative with a median time to treatment not reached at 7 years of follow-up in one study. Many of these patients are diagnosed incidentally or at a time when their lymphoma is not causing symptoms or signs of organ function impairment. Numerous studies have shown that treating patients with asymptomatic disease does not improve survival compared with a program of close observation with treatment reserved for symptomatic disease progression or organ dysfunction. Thus, asymptomtic patients should be observed. When treatment is indicated, there are a variety of treatment options, ranging from the use of the monoclonal antibody against CD20, rituximab, alone or its use in combination with chemotherapy. Treatment decisions are often determined by the indication for treatment and/or by the volume of disease being treated. For patients requiring therapy for inflammatory or autoimmune phenomenon thought to be driven by FL, or for patients with low volume disease, single-agent rituximab is associated with a response rate of ~70% and a median response duration of >2 years. This response duration is improved with the addition of maintenance rituximab following a favorable response to rituximab induction therapy. For patients with a larger volume of disease at the time of treatment initiation, the addition of rituximab to chemotherapy like cyclophosphamide, doxorubicin, vincristine and prednisone (CHOP) or cyclophosphamide, vincristine, and prednisone (CVP) has improved survival in this disease. The combination of bendamustine and rituximab (BR) has been compared to R-CHOP and results in longer response duration and less toxicity. BR then has become the standard of care for the first-line therapy of medium to high-volume FL. Similarly, the addition of maintenance rituximab following a good response to R-CHOP or R-CVP improves response duration when used in newly treated FL patients.

In patients with FL, the disease nearly always recurs following therapy, after which retreatment is again reserved for symptomatic disease or disease interfering with organ function. Single agent rituximab or alternative chemotherapy regimens can again be employed. Both autologous and allogeneic hematopoietic stem cell transplantation yield high complete response rates in patients with relapsed FL, and long-term remissions can occur in 40 and 60% of patients, respectively. The latter is associated with considerable treatment-related morbidity and mortality and so is usually reserved for patients with multiply relapsed FL that is no longer responsive to chemotherapy. More targeted oral therapies like lenalidomide and the PI3 kinase inhibitor idelalisib are active in both untreated and relapsed FL. On average, most patients will live with FL for 15–20 years, a number that is increasing given our improved understanding of the genetics and microenvironment of FL, and the increasing number of drugs and therapies being tested in this disease. However, in addition to a high risk FLIPI, patients who do not have a complete metabolic response by PET/CT scanning to their primary therapy, and patients who relapse within 2 years of the completion of their primary chemotherapy tend to do poorly with chemotherapy.

Patients with FL have a high rate of histologic transformation to DLBCL (~3% per year). This is recognized ~40% of the time during the course of the illness by repeat biopsy and is present in almost all patients at autopsy. This transformation is usually heralded by rapid growth of lymph nodes—often localized—and the development of systemic symptoms such as fevers, sweats, and weight loss. When this happens in patients who have had previously untreated FL, treatment with R-CHOP chemotherapy, as for DLBCL, can be curative for the aggressive component while the FL may eventually recur. In patients with previously

treated FL that transforms to DLBCL, prognosis is poor, and successful therapy with an aggressive combination chemotherapy regimen should be consolidated with an autologous stem cell transplant. Finally, as discussed previously, grade 3B FL is more similar to DLBCL than it is to FL and should be treated as such.

Marginal Zone Lymphoma

The second most common indolent B-cell NHL is MZL. There are three main types: splenic MZL, extranodal MZL of MALT, and nodal MZL.

Nodal MZL most closely resembles FL clinically, and much of the way we manage and treat it is based on studies done in FL. Tumor biopsies in this disease show parafollicular and perivascular infiltration by monocytoid-appearing atypical lymphocytes with folded nuclear contours that are positive for CD19, CD20, and CD79a but negative for CD10 and largely negative for CD5. Some cases can have plasmacytoid differentiation and can be associated with a monoclonal expression of kappa or lambda light chains and with small monoclonal immunoglobulin spikes.

Splenic MZL is largely a disease of older Caucasian patients; infection with hepatitis C is a risk factor for this disease and treatment of hepatitis C can result in regression of the lymphoma. Patients present with a lymphocytosis with or without cytopenias and splenomegaly. Bone marrow involvement is common. Diagnosis can be made by flow cytometry of the peripheral blood; malignant lymphocytes will be positive for surface immunoglobulin, CD19, and CD20 and will generally lack CD5 and CD10. On peripheral smear they have small nuclei and abundant cytoplasm with “shaggy” or villous projections. It can be differentiated from hairy cell leukemia by the absence of CD25, CD103, and annexin A1. Recurrent cytogenetic abnormalities include trisomy 3 and abnormalities of chromosome 7q. Therapy is indicated for symptomatic disease or significant cytopenias. Splenectomy is reasonable for selected patients with excellent relief of symptoms and cytopenias. Splenectomy is associated with an overall response rate of 85% and an estimated progression-free and OS at 5 years of 58 and 77%, respectively. Single-agent rituximab can improve splenomegaly and cytopenias in >90% of patients. In a study of induction with weekly rituximab followed by maintenance, the response rate was 95%, with overall and progression-free survival at 5 years of 92 and 73%, respectively. Other options for therapy at relapse are similar to those used for FL and include retreatment with rituximab, alkylating agents, and purine analogues in combination with rituximab. The survival rate of patients is in excess of 70% at 10 years.

MALT lymphoma is an MZL lymphoma of extranodal tissue, most commonly the stomach but other common sites include the skin, salivary glands, lung, small bowel, ocular adnexa, breasts, bladder, thyroid, dura, and synovium. It is associated with states of chronic inflammation either due to autoimmune diseases like Sjogren’s syndrome or Hashimoto’s thyroiditis, or chronic infections with organisms like *Helicobacter pylori* (gastric), *Borrelia burgdorferi* (skin), *Chlamydia psittaci* (conjunctiva), *Campylobacter jejuni* (intestines), and hepatitis C virus. The essential pathologic feature of MALT lymphoma is the presence of lymphepithelial lesions, which result from invasion of mucosal glands and crypts by the neoplastic lymphocytes. These cells are positive for CD19, CD20, and CD79a and negative for CD5 and CD10. Recurrent cytogentic abnormalities include t(11;18), t(14;18), t(1;14), t(3;14), and trisomy 8. The t(11;18) is most common, occurring in up to 50% of MALT lymphomas. It results in the fusion of the apoptosis inhibitor 2 (API2) gene and the MALT1 gene, resulting in activation of nuclear factor κB (NFκB). Unlike other indolent B-cell lymphomas, MALT lymphomas present most commonly with stage I or II disease. In these cases, radiation therapy may be curative. Alternatively, patients may respond to antibiotics for the associated underlying infection. Treatment of symptomatic or organ impairing relapsed, refractory, or advanced stage disease is similar to approaches used in FL with chemotherapy, immunotherapy, or chemoimmunotherapy.

Lymphoplasmacytic Lymphoma

About 1% of all NHLs will be lymphoplasmacytic lymphomas, which are indolent B-cell NHLs with lymphoplasmacytic differentiation, most commonly associated with a monoclonal IgM paraprotein. Nearly all patients will have stage IV disease at diagnosis with bone marrow involvement. Patients with high levels of circulating IgM paraproteins constitute a specific entity

known as Waldenstrom macroglobulinemia and can have symptoms due to hyperviscosity as a result of the circulating IgM. Activating mutations in MYD88, an adaptor protein that is involved in signaling downstream of the Ig receptor leading to NFκB activation, is present in >90% of cases. Tumor biopsies are notable for proliferation of small lymphocytes, lymphoplasmacytic cells, and plasma cells, and malignant lymphocytes are positive for CD19, CD20, and surface IgM but generally negative for CD5 and CD10. Like the other indolent NHLs, treatment is indicated for disease that causes symptoms or interferes with organ function; hyperviscosity related to elevated serum IgM and paraneoplastic neuropathy are additional indications for therapy. Single-agent rituximab may be useful for low-volume disease, but can be associated with a transient rise in serum IgM concentrations that can cause or exacerbate hyperviscosity. Chemoimmunotherapy with regimens like BR and rituximab, cyclophosphamide, and dexamethasone are active, as are myeloma therapies such as bortezomib. Given that 85% of IgM remains intravascular, acute relief of hyperviscosity symptoms can be obtained by plasmapheresis. For recurrent disease, one can often utilize agents that were previously used. For patients with more refractory LPL, the mTOR inhibitor everolimus and the oral BTK ibrutinib are active. Selected patients with relapsed disease are considered for high-dose therapy with ASCT or alloSCT. The results seen are similar to that of other indolent lymphomas.

Mantle Cell Lymphoma

MCL comprises about 6% of NHLs. It is an intermediate grade lymphoma that like the indolent B-cell NHLs is not curable with conventional therapies, but like the aggressive lymphomas often requires more aggressive chemoimmunotherapy regimens with or without an autologous stem cell transplant to achieve a reasonable response duration. This therapy is not curative, however, and median survival with this disease is on the order of 5–10 years. An exception to this is a more indolent, SOX11 variant that often presents with circulating disease with splenomegaly but without significant lymphadenopathy and with a low Ki67 (<10%). This subset behaves more like the indolent B-cell NHLs and can be observed until treatment is indicated by symptoms or organ function impairment. Similarly, there is a blastic variant with a high Ki67 index that is associated with a poor prognosis and a median OS of only 18 months. For other patients, prognosis is best predicted by the biologic mantle cell prognostic index (MIPI), which factors in age, performance status, LDH, white blood cell count, and Ki67 expression to determine a risk group. This disease is more common in men, and the average age of diagnosis is 63. Over two-thirds of patients will have stage IV disease, mostly with bone marrow and peripheral blood involvement, at the time of diagnosis. Other common extranodal sites of involvement include the GI tract where diffuse lymphomatous polyposis may be seen.

The pathognomonic cytogentic finding in MCL is t(11;14), which brings the gene for the cell cycle control protein cyclin D1 under the control of the immunoglobulin heavy chain gene promoter on chromosome 14. This translocation is present in >90% of cases. The remaining cases usually overexpress cyclin D2, cyclin D3, or cyclin E. Tumor cells also are positive for B-cell markers CD19 and CD20, as well as CD5. They usually lack CD10 and CD23.

Therapies for MCL are evolving. Patients with localized disease might be treated with combination chemotherapy followed by radiotherapy; however, these patients are exceedingly rare. Similarly, patients with the indolent variant can be observed until disease progresses to cause symptoms or signs of organ function impairment. For the usual presentation with disseminated disease, standard lymphoma treatments like R-CHOP have been unsatisfactory, with the minority of patients achieving complete remission. The addition of high dose cytarabine to an R-CHOP-like backbone with or without consolidation autologous stem cell transplantation in first remission has improved progression-free survival, but it has not elicited cures in this disease. These include the Nordic regimens and R-HyperCVAD (rituximab, cyclophosphamide, vincristine, doxorubicin, dexamethasone, cytarabine, and methotrexate). BR has activity in this disease and is more effective and better tolerated than R-CHOP. Newer studies with short follow-up suggest that strategies that combine BR with cytarabine with or without autologous stem cell transplant may be effective and well tolerated. Maintenance rituximab, following a good response to induction chemotherapy or after autologous stem cell transplant, also improves outcomes over observation alone. For relapsed disease, the BTK inhibitor ibrutinib has single-agent activity with a response rate of almost 70% but a response duration of only 18 months. Drugs like lenalidomide, bortezomib, and temsirolimus can similarly induce transient partial responses. Appropriate

patients who respond to salvage therapy should be considered for allogeneic stem cell transplant, which can lead to long-term disease-free survival in 30–50% of patients.

MATURE (PERIPHERAL) T-CELL DISORDERS

Mature T-cell disorders include cutaneous lymphomas, like mycosis fungoides, and the PTCLs, some of which are distinguished based on specific clinical presentations or contexts or by molecular or biologic features, but many of which fall into the category of PTCL not otherwise specified (NOS). T-cell NHLs are significantly more rare than the B-cell NHLs, and as such, our understanding of their biology is less advanced, and our therapies are less well developed. While some T-cell lymphomas, like mycosis fungoides, can behave indolently and some, like ALK-positive ALCL, can be cured with chemotherapy, the majority are associated with a poor prognosis. The advent of genomic technologies is enhancing our ability to understand the genetic and biologic basis of these neoplasms.

Mycosis Fungoides

Mycosis fungoides is also known as cutaneous T-cell lymphoma. This lymphoma is more often seen by dermatologists than internists. The median age of onset is in the mid-fifties, and the disease is more common in males and in blacks.

Mycosis fungoides is an indolent lymphoma with patients often having several years of eczematous or dermatitic skin lesions before the diagnosis is finally established. The skin lesions progress from patch stage to plaque stage to cutaneous tumors. Early in the disease, biopsies are often difficult to interpret, and the diagnosis may only become apparent by observing the patient over time. Adenopathy may reflect involvement with mycosis fungoides or be read as dermatopathic change. In advanced stages, the lymphoma can spread to lymph nodes and visceral organs. Patients with this lymphoma may develop generalized erythroderma and circulating tumor cells, called *Sézary's syndrome*.

Rare patients with localized early-stage mycosis fungoides can be cured with radiotherapy, often total-skin electron beam irradiation. More advanced disease has been treated with topical glucocorticoids, topical nitrogen mustard, phototherapy, psoralen with ultraviolet A (PUVA), extracorporeal photopheresis, retinoids (bexarotene), electron beam radiation, interferon, antibodies, fusion toxins, histone deacetylase inhibitors, and systemic cytotoxic therapy. Unfortunately, these treatments are palliative.

Peripheral T-Cell Lymphoma, Not Otherwise Specified

PTCLs include a number of entities, which constitute 15% of all NHLs in adults. PTCL, NOS, comprising 6% of all NHLs, is the term used for cases that are not other entities defined in the WHO classification. Named varieties include ALCL, angioimmunoblastic T-cell lymphoma (AITL), hepatosplenic T-cell lymphoma, enteropathy-associated T-cell lymphoma, and subcutaneous panniculitis T-cell lymphoma. PTCL NOS is a disease of older individuals, with a median age at presentation of 65, and the majority of patients will have advanced-stage disease at diagnosis, with involvement of the bone marrow, liver, spleen, and skin being common. Associated "B" symptoms and pruritis are also common. These lymphomas can be associated with a reactive eosinophilia as well as hemophagocytic syndrome. The IPI has been applied to PTCL NOS and provides some assessment of outcomes, but even the low-risk group has a median OS of just >2 years.

This diagnostic category is a collection of heterogeneous lymphomas that vary widely and lack typical findings of other specific PTCL subgroups. Because of this heterogeneity, histology, immunophenotype, and genetics are variable. Often lymph nodes are effaced by atypical lymphoid cells of various sizes, sometimes associated with vascular proliferation or an infiltrate of eosinophils and/or macrophages. As most of these lymphomas behave aggressively, note is often made of mitotic and apoptotic figures as well as geographic necrosis. The cells often are positive for CD3, and the majority of PTCL NOS is positive for CD4 rather than CD8, but some are negative for both markers. There can be loss of more mature T-cell markers like CD5 and

CD7, and this is associated with a more aggressive course. There are some recurrent translocations, including t(7;14), t(11;14), inv(14), and t(14;14), all of which involve the TCR genes.

The most common primary therapy for PTCL NOS involves a CHOP-like chemotherapy backbone—either CHOP alone or CHOP in combination with etoposide, or CHOEP. The latter may provide the most benefit to younger patients and patients with more favorable disease risk factors. Autologous stem cell transplant has been investigated for patients in their first remission and does seem to improve PFS in certain contexts. Drugs like gemcitabine, bendamustine, and pralatrexate have activity in relapsed disease, as do the histone deacetylase inhibitors romidepsin and belinostat. All of these agents are associated with transient responses in a minority of patients. Patients should be considered for clinical trials. For patients who do achieve remission, reduced intensity allogeneic stem cell transplantation can yield long-term non-relapse survival on the order of 40–50%.

Angioimmunoblastic T-Cell Lymphoma

AITL constitutes about 20% of T-cell NHL and about 4% of all NHL diagnosed. Patients present with a variety of signs and symptoms, most often including lymphadenopathy, hepatosplenomegaly, B symptoms, rash, polyarthritis, and hemolytic anemia. Over 80% of patients have advanced stage disease at diagnosis, and bone marrow involvement is common. Polyclonal hypergammaglobulinemia is common, as are elevated LDH, eosinophilia, a positive Coombs test, and opportunistic infections.

On biopsy, lymph nodes are effaced by a polymorphous infiltrate of lymphoctyes, ranging in size and shape, and of immunoblasts. The neoplastic lymphocytes are positive for CD3 as well as CXCL13, PD-1, CD10, and BCL6, most closely resembling CD4-positive follicular helper T cells. There is an expanded follicular dendritic cell network surrounding tumor cells. Scattered immunoblasts are often EBV positive and may give rise to secondary EBV-positive B-cell lymphomas at a later time. Genetic analysis of this disease has revealed recurrent mutations in *TET2* (76%), *DNMT3* (33%), and *IDH2* (20%).

There is a subset of AITL that can remit with immunosuppression with agents like glucocorticoids or methotrexate. Most patients, however, will need combination chemotherapy with regimens like those used in PTCL NOS. Median response duration is short, median OS is only 15–36 months. Treatment of relapsed disease is similar to that of relapsed PTCL NOS.

Anaplastic Large Cell Lymphoma

ALCL is the next most common T-cell lymphoma after AITL but is more common in children, accounting for up to 10% of pediatric lymphomas. Approximately 40–60% of cases with harbor t(2;5) which fuses a portion of the nucleolar protein nucleophosmin-1 (NPMI) gene to a part of the anaplastic lymphoma kinase (ALK) gene, the product of which has constitutive tyrosine kinase activity. These patients have a much more favorable prognosis compared to ALK-negative ALCL, akin to that of DLBCL. There is an additional, more indolent and favorable subtype that occurs in the breast tissue of patients with breast implants, and there is a cutaneous variant. In general, this is a disease that is more common in men. ALK-positive disease is a disease of younger patients with a median age at diagnosis of 34 years, whereas the median age at diagnosis of ALK negative patients is 58. With the exception of the cutaneous variant and the variant associated with breast implants, most patients present with rapidly growing lymphadenopathy with or without extranodal involvement; B symptoms are common.

Most cases of ALCL involve large atypical lymphocytes with a horseshoe-shaped nuclei with prominent nucleoli (“hallmark” cells). Tumor cells tend to be localized within the lymph node sinuses, and almost all are positive for CD30 but negative for CD15. A majority will also express CD3, CD25, CD43, and CD4. ALK rearranged ALCL can be diagnosed by FISH cytogenetics for t(2;5) or by immunohistochemical staining for ALK.

ALCL is generally treated with CHOP, although like PTCL NOS, CHOEP may benefit younger patients, particularly with ALK+ disease. Overall, ALCL has a better prognosis than PTCL, and this is particularly true for ALK-positive disease, which has an 8-year OS of 82 versus 49% for ALK-negative disease. Relapsed ALK-positive ALCL is treated similarly to relapsed DLBCL, with salvage combination chemotherapy to identify chemotherapy sensitivity followed by autologous stem cell transplant. For patients with chemoinsensitive diseaese or for ALK- negative disease, the conjugated anti-CD30 antibody to monomethyl

aurostatin E (MMAE) brentuximab is highly active with a response rate of 86% and a complete response rate of 57%. It is currently being investigated in combination with cyclophosphamide, adriamycin, and prednisone for the primary treatment of disease. The ALK inhibitors, including crizotinib, are active in refractory ALK-positive ALCL with excellent outcomes.

Other PTCL Subtypes

Enteropathy-associated T-cell lymphoma, hepatosplenic T-cell lymphoma, and subcutaneous panniculitis-like T-cell lymphoma are other less common PTCL subtypes. Enteropathy-type intestinal T-cell lymphoma is a rare disorder. Type I occurs in patients with a history of gluten-sensitive enteropathy and is associated with HLADQA1*0501, DQB1*0201; a gluten-free diet can prevent the development of this lymphoma. Type II is not associated with celiac disease and may be a separate disease entity. Patients are frequently cachectic and sometimes present with intestinal perforation. The prognosis is poor with a median survival of 10 months. Therapy is often with combination chemotherapy, including high-dose methotrexate, and autologous stem cell transplant in first remission. *Hepatosplenic* γδ *T-cell lymphoma* is a systemic illness that presents with sinusoidal infiltration of the liver, spleen, and bone marrow by malignant T cells. Tumor masses generally do not occur. The disease is associated with systemic symptoms and is often difficult to diagnose. Recurrent genetic events include isochromosome 7q and trisomy 8. Treatment outcome is poor, but regimens that include ifosphamide, like ifosphamide, carboplatin, etoposide (ICE) or ifosphamide, etoposide, cytarabine (IVAC), are associated with better outcomes in small series of patients. Responding patients should be considered for allogeneic stem cell transplantation. *Subcutaneous panniculitis-like T-cell lymphoma* is a rare disorder that is often confused with panniculitis. Patients present with multiple subcutaneous nodules, which progress and can ulcerate. There is a more indolent form that tends to express α/β TCRs and can be managed with immune suppression, whereas lymphomas that express γ/δ TCRs are more aggressive and are associated with a worse prognosis and coincident hemophagocytic syndrome. This is a disease of young men in their fifth and sixth decades of life. Patients with aggressive disease are managed with multiagent chemotherapy, and responding patients should be considered for allogeneic stem cell transplantation.

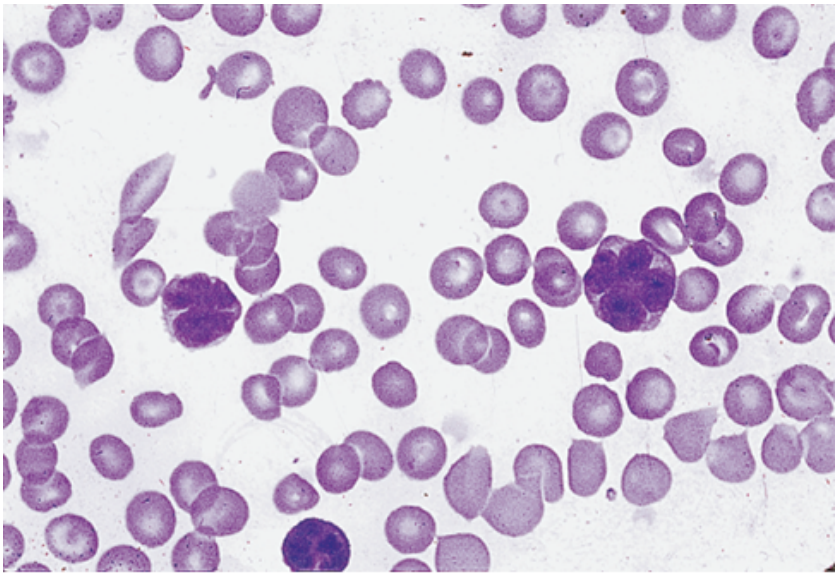
Adult T-Cell Leukemia/Lymphoma

Adult T-cell leukemia/lymphoma (ATLL) is a disease that is most prevalent in Japan and the Caribbean basin. It is a neoplasm that is driven by HTLV-1, often contracted through the breast milk of infected mothers. The average age at diagnosis is 60, so there is a long latency between viral infection and viral transformation, and only 4% of infected patients will develop the disease. This suggests that HTLV-1 may not be sufficient to cause the malignant phenotype. There are four disease variants: acute (60% of patients), lymphomatous (20% of patients), chronic (15% of patients), and smoldering (5% of patients). Prognosis varies across these groups with with median survivals of 6 months, 10 months, 24 months, and not yet reached, respectively. Presentation depends on the subtype, but most commonly patients present with circulating disease and bone marrow involvement, hypercalcemia, lytic bone lesions, lymphadenopathy, heptosplenomegaly, skin lesions, and opportunistic infections.

The pathognomonic finding is the malignant “flower cell” that is positive for CD4 and CD25, as well as CD2, CD3, and CD5 but lacking CD7 **(Fig. 104-7)**. Combination chemotherapy is generally used, but for patients fortunate enough to respond, response durations are very short. Other active agents in this disease include the anti-retroviral agent zidovudine, interferon α, and arsenic. In any patients who do respond to therapy, allogeneic stem cell transplant should be considered.

FIGURE 104-7

Adult T-cell leukemia/lymphoma. Peripheral blood smear showing leukemia cells with typical “flower-shaped” nucleus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Extranodal NK/T-Cell Lymphoma, Nasal Type

Extranodal NK/T-cell lymphoma, nasal type, is a lymphoma that is associated with EBV infection in nearly all cases and more common in Asia and native populations in Peru. It usually presents with a mass and obstructive symptoms in the upper aerodigestive tract with occasional extranodal sites, but over two-thirds of patients will have localized disease. It is more common in men, and the median age at diagnosis is 60. This disease has its own prognostic score, which takes into account the presence or absence of B symptoms, disease stage, whether LDH is elevated, and whether there is lymph node involvement. EBV viral load at diagnosis and end of therapy is also predictive.

Treatment for early stage disease is usually with combined modality therapy of chemotherapy (commonly using etoposide, ifosfamide, cisplating, and dexamethasone) and intensity-modulated radiation therapy (50–55 Gy), and patients with localized disease involving the nasal passages do quite well, with 3-year OS of ~85%. Patients with more advanced stage disease do poorly with disseminated extranodal relapse occurring frequently, and median OS is only 4.3 months. The most commonly used regimen for the regimens is the SMILE regimen (dexamethasone, methotrexate, ifosfamide, L-asparaginase, and etoposide).

FURTHER READING

Al-Zahrani M, Savage KJ: Peripheral T-cell lymphoma, not otherwise specified: A review of current disease understanding and therapeutic approaches. Hematol Oncol Clin North Am 31:189, 2017.

Bachy E, Salles G: Treatment approach to newly diagnosed diffuse large B-cell lymphoma. Semin Hematol 52:107, 2015.

Broccoli A, Zinzani PL: Angioimmunoblastic T-cell lymphoma. Hematol Oncol Clin North Am 31:223, 2017.

Cheah CY et al: Mantle cell lymphoma. J Clin Oncol 34:1256, 2016.

Du MQ: MALT lymphoma: Genetic abnormalities, immunological stimulation and molecular mechanism. Best Pract Res Clin Haematol 30:13, 2017.

Dunleavy K et al: Update on Burkitt lymphoma. Hematol Oncol Clin North Am 30:1333, 2016.

Jacobson CA, Freedman AS: Is there a best initial treatment for a new patients with low grade follicular lymphoma? Curr Hematol Malig Rep 11:218, 2016.

Sesques P, Johnson NA: Approach to the diagnosis and treatment of high-grade B-cell lymphomas with MYC and BCL2 and/or BCL6 rearrangements. Blood 129:280, 2017.

Zucca E, Bertoni F: The spectrum of MALT lymphoma at different sites: Biological and therapeutic relevance. Blood 127:2082, 2016.

Chapter 105: Hodgkin's Lymphoma

Caron A. Jacobson; Dan L. Longo

FIGURE 105-1

INTRODUCTION

Hodgkin's lymphoma (HL) is a malignancy of mature B lymphocytes. It represents ~10% of all lymphomas diagnosed each year. The majority of HL diagnoses are classical HL (cHL), but there is a second subtype of HL, nodular lymphocyte predominant HL (NLPHL). While this diagnosis does resemble cHL morphologically in certain respects, there is some evidence that it is more related to the indolent B-cell NHLs biologically than it is to cHL. The majority of this chapter will be specific to cHL, with a discussion of NLPHL at the end.

Classical HL is one of the success stories of modern oncology. Until the advent of extended-field radiotherapy in the mid-twentieth century, it was a highly fatal disease of young people. Radiation therapy cured some patients with early stage disease, and the introduction of multi-agent chemotherapy in the 1970s resulted in further improved cure rates, both for patients with early and advanced stage disease. Cure rates now are >85%. The new challenge in the treatment of HL is late therapy-related toxicity, including a high rate of secondary malignancies and cardiovascular disease. Current clinical trials are aimed at minimizing this risk while preserving efficacy.

EPIDEMIOLOGY AND ETIOLOGY

HL is of B-cell origin. The incidence of HL appears fairly stable, with 8260 new cases diagnosed in 2017 in the United States. HL is more common in whites than in blacks and more common in males than in females. A bimodal distribution of age at diagnosis has been observed, with one peak incidence occurring in patients in their twenties and the other in those in their eighties. Some of the late age peak may be attributed to confusion among entities with similar appearance such as anaplastic large cell lymphoma and T-cell/histiocyte–rich B-cell lymphoma. There are four distinct subtypes of classical Hodgkin's lymphoma (cHL) that are differentiated based on their histopathologic features (**Table 105-1**): nodular sclerosis, mixed cellularity, lymphocyte-rich, and lymphocyte-depleted. Patients in the younger age groups diagnosed in the United States largely have the nodular sclerosing subtype of HL. Elderly patients, patients infected with HIV, and patients in Third World countries more commonly have mixed-cellularity HL or lymphocyte-depleted HL. Together, nodular sclerosis and mixed cellularity types account for nearly 95% of cases. Infection by HIV is a risk factor for developing HL. In addition, an association between infection by Epstein-Barr virus (EBV) and HL has been suggested. A monoclonal or oligoclonal proliferation of EBV-infected cells in 20–40% of the

patients with HL has led to proposals for this virus having an etiologic role in HL. However, the matter is not settled definitively. Viral oncogenesis appears to play a greater role in HIV-related cHL: EBV can be detected in nearly all cases of HIV-associated cHL, compared to only one-third of cases of non-HIV-associated cHL. Reed-Sternberg (HRS) cells are the malignant cells in HL. HRS cells in HIV-associated cHL express the EBV-transforming protein latent membrane protein 1 (LMP-1), and the EBV genomes from multiple disease sites in the same HIV-associated cHL patient are episomal and clonal, suggesting that EBV is directly involved in early lymphomagenesis.

TABLE 105-1

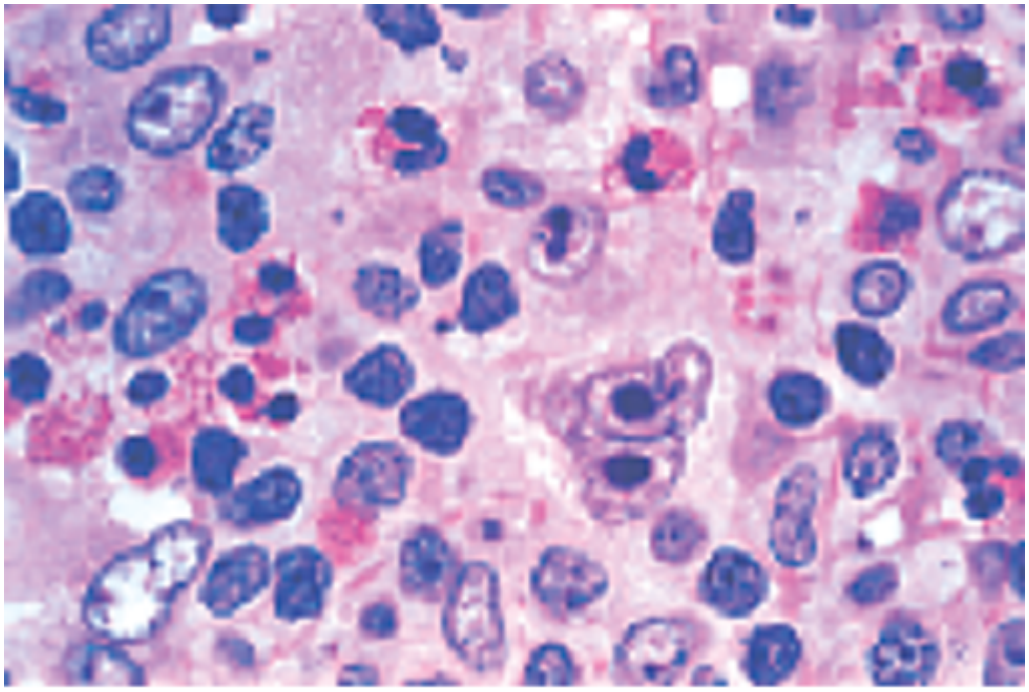
WHO Classification of Hodgkin's Lymphoma

Nodular lymphocyte predominant Hodgkin's lymphoma
Classical Hodgkin's lymphoma
Nodular sclerosis
Lymphocyte-rich
Mixed cellularity
Lymphocyte-depleted

Histologically, the HRS cell is diagnostic of cHL **(Fig. 105-1)**. These cells are large cells with abundant cytoplasm with bilobed and/or multiple nuclei. By immunohistochemistry they are often PAX-5 positive but have low to no expression of other B-cell antigens like CD19 and CD20. They express CD15 and CD30 in 85 and 100% of cases, respectively. These cells, though, comprise <1% of the tumor cellularity, with the majority of the tumor made up of a surrounding inflammatory infiltrate of polyclonal lymphocytes, eosinophils, neutrophils, macrophages, plasma cells, fibroblasts, and collagen. The HRS cell interacts with its microenvironment via cell-cell contact and elaboration of growth factors and cytokines, which results in a surrounding cellular milieu that protects it from host immune attack. The surrounding environmental cells likewise support the HRS cells via cell-cell signaling and cytokine production which provides signals that promote proliferation and survival of the HRS cell itself. Interestingly, 97% of HRS cells in cHL harbor genetic aberrations in the PD-L1 locus on chromosome 9p24.1, resulting in overexpression of PD-L1, the ligand for the inhibitory PD-1 receptor on immune cells. This is one mechanism whereby the HRS cell may be able to avoid immune destruction in its inflammatory microenvironment and may contribute to the generalized immune suppression in HL patients.

FIGURE 105-1

Hodgkin's disease: A classic Reed-Sternberg (RS) cell is present near the center of the field. RS cells are large cells with a bilobed nucleus and prominent nucleoli surrounded by a pleiomorphic cellular infiltrate.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Classic Hodgkin's Lymphoma

Most patients with cHL present with palpable lymphadenopathy that is nontender; in most patients, these lymph nodes are in the neck, supraclavicular area, and axilla. More than half of the patients will have mediastinal adenopathy at diagnosis, and this is sometimes the initial manifestation. Subdiaphragmatic presentation of cHL is unusual and more common in older males. One-third of patients present with fevers, night sweats, and/or weight loss, or "B" symptoms. Occasionally, HL can present as a fever of unknown origin. This is more common in older patients who are found to have mixed-cellularity HL in an abdominal site. Rarely, the fevers persist for days to weeks, followed by afebrile intervals and then recurrence of the fever. This pattern is known as *Pel-Ebstein* fever. HL can occasionally present with unusual manifestations. These include severe and unexplained itching, cutaneous disorders such as erythema nodosum and ichthyosiform atrophy, paraneoplastic cerebellar degeneration and other distant effects on the CNS, nephrotic syndrome, immune hemolytic anemia and thrombocytopenia, hypercalcemia, and pain in lymph nodes on alcohol ingestion.

Evaluation of patients with HL will typically begin with a careful history and physical examination. Patients should be asked about the presence or absence of "B" symptoms. Comorbid diagnoses that may impact therapy should be reviewed, including a history of pulmonary disease and congestive heart failure given the use of chemotherapy drugs that can cause both lung and heart toxicity. A physical examination should pay attention to the peripherally accessible sites of lymph nodes and to the liver and spleen size. Laboratory

evaluation should include a complete blood count with differential; erythrocyte sedimentation rate; chemistry studies reflecting major organ function including serum albumin; and HIV and hepatitis virus testing. A PET/CT scan is used for staging, and is more accurate than a bone marrow biopsy for evaluation of bone marrow involvement as the bone marrow involvement in cHL tends to be patchy and therefore potentially missed on a unilateral bone marrow biopsy. The initial evaluation of a patient with HL or NHL is similar. In both situations, the determination of an accurate anatomic stage is an important part of the evaluation. Staging is done using the Ann Arbor staging system (**Table 105-2**).

The diagnosis of HL is established by review of an adequate biopsy specimen by an expert hematopathologist. HL is a tumor characterized by rare neoplastic cells of B-cell origin (immunoglobulin genes are rearranged but not expressed) in a tumor mass that is largely polyclonal inflammatory infiltrate, probably a reaction to cytokines produced by the tumor cells. The differential diagnosis of a lymph node biopsy suspicious for HL includes inflammatory processes, mononucleosis, NHL, phenytoin-induced adenopathy, and nonlymphomatous malignancies.

Staging for cHL is anatomically based given the propensity of the disease to march from one lymph node group to the next group, often contiguous to the first. Staging is important for selecting therapy of appropriate intensity, but the outcome of optimal therapy for all the stages is excellent. Patients are stratified based on whether they have early stage, stage I or II, or advanced stage, stage III or IV, disease. Patients with early stage disease have a better prognosis overall but are further classified as favorable or unfavorable based on a variety of factors. These factors vary from study to study but include bulky disease, number of lymph node areas involved, an elevated ESR (>30 if B symptoms are present; >50 if B symptoms are absent), and age. Prognosis in advanced stage disease is best predicted by the International Prognostic Score (IPS), which ascribes a point for male sex, older age (>45 years), stage IV disease, serum albumin <4 g/dL, hemoglobin <10.5 g/dL, white blood cell count ≥15,000/μL, and a lymphocyte count <600/μL and/or <8% of white blood cell count. Five-year progression-free survival ranges from 88% for patients with no risk factors, to 62% for patients with four or more factors, but very few patients have multiple risk factors.

TABLE 105-2

The Ann Arbor Staging System for Hodgkin's Lymphoma

Stage	Definition
I	Involvement of a single lymph node region or lymphoid structure (e.g., spleen, thymus, Waldeyer's ring)
II	Involvement of two or more lymph node regions on the same side of the diaphragm (the mediastinum is a single site; hilar lymph nodes should be considered "lateralized" and, when involved on both sides, constitute stage II disease)
III	Involvement of lymph node regions or lymphoid structures on both sides of the diaphragm
III_1	Subdiaphragmatic involvement limited to spleen, splenic hilar nodes, celiac nodes, or portal nodes
III_2	Subdiaphragmatic involvement includes paraaortic, iliac, or mesenteric nodes plus structures in III_1
IV	Involvement of extranodal site(s) beyond that designated as "E"
	More than one extranodal deposit at any location
	Any involvement of liver or bone marrow
A	No symptoms
B	Unexplained weight loss of >10% of the body weight during the 6 months before staging investigation
	Unexplained, persistent, or recurrent fever with temperatures >38°C during the previous month
	Recurrent drenching night sweats during the previous month
E	Localized, solitary involvement of extralymphatic tissue, excluding liver and bone marrow

TREATMENT

TREATMENT

Classic Hodgkin's Lymphoma

The overwhelming majority of patients with HL will be cured with either chemotherapy alone, or a combination of chemotherapy and radiation therapy. It has long been appreciated that patients with advanced stage disease do not benefit from the addition of radiation therapy to chemotherapy and are thus treated with chemotherapy alone. For early stage disease, however, treatment with combined modality therapy has been associated with a small decrease in risk of relapse but with an increased risk of late toxicity including secondary malignancies, thyroid disease, and premature cardiovascular disease and stroke resulting in minimal or no improvement in long-term survival. Much of this risk can be attributed to radiation therapy. Thus, investigation into the treatment of early stage HL at present is aimed at trying to maximize treatment outcome without using radiotherapy. This is an area of controversy in the treatment of HL.

EARLY STAGE DISEASE

The most common chemotherapy regimen used to treat HL in the United States is ABVD (adriamycin, bleomycin, vinblastine, and dacarbazine). This regimen is given every other week, with each cycle including two treatments. In patients with low-risk, or favorable disease, the use of 4–6 cycles of ABVD alone, without radiation therapy, results in progression-free and overall survivals of 88–92% and 97–100% at 5–7 years. This may be associated with a slightly increased risk of relapse when compared with abbreviated chemotherapy (ABVD x4 cycles) followed by involved field radiation therapy (30 Gy), but with no difference in overall survival owing to the excellent salvage strategies used for relapsed HL and to the late toxicities seen following radiation therapy to the chest. German studies have examined a very abbreviated chemotherapy regimen (ABVD x2 cycles) and low-dose radiation (20 Gy) for particularly good risk disease with two or fewer lymph node areas involved and found that this was equally effective to standard combined modality therapy of ABVD x4 cycles and 30 Gy of radiation, though long-term follow-up is not yet available to assess the impact of the lower radiotherapy dose on late toxicities. Finally, the use of an early interim PET/CT scan can aid decisions on the duration and extent of therapy. In one study, a negative PET/CT scan after 3 cycles of ABVD predicted for excellent outcomes with no additional therapy; in another, a negative PET/CT scan after 2 cycles of ABVD predicted for good outcomes with 2 additional cycles of ABVD alone, without radiation therapy.

For unfavorable risk disease, the omission of radiation therapy following chemotherapy is associated with a more significant increased risk of relapse compared to favorable risk disease, but again with no change in overall survival. For these patients, treatment options would include ABVD x4 cycles followed by involved field radiation therapy or ABVD alone for 6 cycles. Treatment decisions are often based on the extent of the radiation field and the unfavorable risk factor, with patients with non-bulky disease being candidates for chemotherapy alone if radiation would be contraindicated for another reason. Combined modality therapy has typically been used for patients with bulky disease, although patients with bulky disease who have a negative PET/CT scan after chemotherapy may not benefit from additional radiation therapy.

Alternative chemotherapy regimens to ABVD have been developed and include the Stanford V regimen and escalated BEACOPP. Neither of these regimens has resulted in improved outcomes in patients with early

stage disease.

ADVANCED STAGE DISEASE

Patients with advanced stage disease do not benefit from the addition of radiation therapy after a complete response to chemotherapy alone and should be treated with chemotherapy alone. The most common regimen used in the United States is ABVD x6 cycles. Again, Stanford V and escalated BEACOPP have been evaluated in advanced stage disease and are not associated with an improvement in overall survival but are associated with increased toxicity. The small fraction of patients who do not achieve complete remission with chemotherapy alone (partial responders with persistent PET scan positivity account for <10% of patients) may benefit from the addition of involved field radiotherapy.

Newer drugs have been developed for the treatment of relapsed HL (see "Relapsed Disease," below). These include the antibody drug conjugate brentuximab, which is an antibody against CD30 conjugated to the microtubule inhibitor MMAE. This drug has been combined with adriamycin, bleomycin, and dacarbazine in early phase studies for advanced stage HL with favorable efficacy compared to historical controls. We await the data from the randomized trial of AVD+brentuximab compared to ABVD. Drugs that target the PD-1/PD-L1 axis have been developed in an attempt to boost the host immune recognition of tumors. This was particularly attractive in HL given the overexpression of PD-L1 on the HRS cell surface. In the setting of relapsed disease, these drugs, which include pembrolizumab and nivolumab, have very high response rates and are associated with durable responses. These are now being tested in conjunction with chemotherapy both as salvage therapy for relapsed disease and in previously untreated patients.

RELAPSED DISEASE

Patients who relapse after primary therapy of Hodgkin's lymphoma can frequently still be cured. Patients who relapse after an effective chemotherapy regimen are usually not curable with subsequent chemotherapy administered at standard doses. Alternative salvage chemotherapy administered at standard doses, then, is given in order to document sensitivity to chemotherapy and to achieve maximum reduction of tumor mass. For patients who respond completely or nearly so, autologous bone marrow transplantation can cure over half of patients. Standard salvage chemotherapy regimens include ICE (ifosfamide, carboplatin, etoposide) or GND (gemcitabine, navelbine, doxil). For patients with early stage disease who do not respond sufficiently to salvage chemotherapy, radiation therapy can be very effective to achieve a remission; whether to consolidate such a remission with an autologous stem cell transplant is debated. For patients with advanced stage disease in whom salvage chemotherapy fails, the antibody drug conjugate brentuximab vedotin, a CD30-directed antibody linked to the microtubule toxin MMAE, is active and can be tried as a bridge to allogeneic transplant. This immunotoxin is also being combined with chemotherapy for use in both the first-line salvage setting and for the upfront treatment of both early- and advanced-stage disease. The anti-PD-1 immune checkpoint inhibitors, nivolumab and pembrolizumab, have efficacy in relapsed HL, and many responses are durable.

SURVIVORSHIP

Because of the very high cure rate in patients with HL, long-term complications have become a major focus for clinical research. In fact, in some series of patients with early-stage disease, more patients died from late complications of therapy than from HL itself. This is particularly true in patients with localized disease. The most serious late side effects include second malignancies and cardiac injury. Patients are at risk for the development of acute leukemia in the first 10 years after treatment with combination chemotherapy regimens that contain alkylating agents plus radiation therapy. The risk for development of acute leukemia is greater after MOPP-like and BEACOPP-like regimens than with ABVD. The risk of development of acute leukemia after treatment for HL is also related to the number of exposures to potentially leukemogenic agents (i.e., multiple treatments after relapse) and the age of the patient being treated, with those aged >60 years at particularly high risk. The development of carcinomas as a complication of treatment for HL is a major problem. These tumors usually occur ≥10 years after treatment and are associated with use of radiotherapy. For this reason, young women treated with thoracic radiotherapy for HL should institute screening mammograms 5–10 years after treatment, and all patients who receive thoracic radiotherapy for HL should be discouraged from smoking. Mediastinal radiation also accelerates coronary artery disease, and patients should be encouraged to minimize risk factors for coronary artery disease such as smoking and elevated cholesterol levels. Cervical radiation therapy increases the risk of carotid atherosclerosis and stroke and thyroid disease, including cancer.

A number of other late side effects from the treatment of HL are well known. Patients who receive thoracic radiotherapy are at very high risk for the eventual development of hypothyroidism and should be observed for this complication; intermittent measurement of thyrotropin should be made to identify the condition before it becomes symptomatic. Lhermitte's syndrome occurs in ~15% of patients who receive thoracic radiotherapy. This syndrome is manifested by an "electric shock" sensation into the lower extremities on flexion of the neck. Because of the young age at which HL is often diagnosed, infertility is a concern for patients undergoing treatment for HL. Chemotherapy regimens containing alkylating agents induce permanent infertility in nearly all men. The risk of permanent infertility in women treated with alkylating agent-containing chemotherapy is age-related, with younger women more likely to recover fertility. Infertility is very rare after treatment with ABVD.

NODULAR LYMPHOCYTE-PREDOMINANT HODGKIN LYMPHOMA

NLPHL is now recognized as an entity distinct from cHL. Previous classification systems recognized that biopsies from a small subset of patients diagnosed as having HL contained a predominance of small lymphocytes and rare Reed-Sternberg-like cells; tumors have a nodular growth pattern and a clinical course that varied from that of patients with cHL. This is an unusual clinical entity and represents <5% of cases of HL and defines NLPHL.

NLPHL has a number of characteristics that suggest its relationship to NHL, rather than cHL, however. The HRS-like cell, or L&H (lymphocyte and histiocyte) or "popcorn" cell, is a clonal proliferation of B cells that are positive for B cell markers CD45, CD79a, CD20, CD19, and BCL2. They do not express two markers normally found on HRS cells, CD30 and CD15. This lymphoma tends to have a chronic, relapsing course and sometimes transforms to diffuse large B-cell lymphoma, including a specific subtype of diffuse large B-cell lymphoma

known as T cell/histiocyte-rich B-cell lymphoma, which shares an immunophenotype with the L&H cell. This natural history most closely resembles that of the indolent B cell NHLs outlined in **Chaps. 104 and 106.**.

Patients with NLPHL are more commonly male (75%). Like cHL, the age distribution of patients with this disease has two peaks, but unlike cHL these peaks include children and adults ages 30–40 years, respectively. The majority of patients diagnosed have stage I or II disease (75%), with a minority having advanced stage disease at diagnosis. B symptoms are uncommon.

Patients with early stage disease at diagnosis should be treated with definitive radiotherapy. This is associated with a 15-year non-relapse survival of 82%. The treatment of patients with advanced stage NLPHL is controversial. Some clinicians favor no treatment of asymptomatic disease and merely close follow-up, akin to the indolent B cell NHLs. For patients who need therapy due to symptoms or signs of organ function impairment, both cHL regimens and B-cell NHL regimens have been used, including ABVD and R-CHOP. A single institution experience with R-CHOP resulted in a 100% response rate in a small group of patients without a single relapse with 42 months follow-up. Although this is short follow-up for an indolent disease, some believe R-CHOP may be curative in this disease and advocate treating with advanced stage disease at diagnosis, regardless of symptoms or organ function.

FURTHER READING

Alperovich A, Younes A: Targeting CD30 using brentuximab vedotin in the treatment of Hodgkin lymphoma. Cancer J 22:23, 2016.

Ansell SM: Hodgkin lymphoma: 2016 update on diagnosis, risk-stratification, and management. Am J Hematol 91:434, 2016.

Armand P et al: Programmed death-1 blockade with pembrolizumab in patinets with classical Hodgkin lymphoma after brentuximab vedotin failure. J Clin Oncol 34:3733, 2016.

Ng AK, van Leeuwen FE: Hodgkin lymphoma: Late effects of treatment and guidelines for surveillance. Semin Hematol 53:209, 2016.

Radford J et al: Results of a trial of PET-directed therapy for early-stage Hodgkin's lymphoma. N Engl J Med 372:1598, 2015.

Rashidi A et al: Allogeneic hematopoietic stem cell transplantation in Hodgkin lymphoma: A systemic review and meta-analysis. Bone Marrow Transplant 51:521, 2016.

Chapter 106: Less Common Hematologic Malignancies

Ayalew Tefferi; Dan L. Longo

INTRODUCTION

The most common lymphoid malignancies are discussed in **Chaps. 102**, **103**, **104**, **105** and **107**, myeloid leukemias in **Chaps. 100** and **101**, myelodysplastic syndromes (MDS) in **Chap. 98**, and myeloproliferative syndromes in **Chap. 99**. This chapter will focus on the more unusual forms of hematologic malignancy. The diseases discussed here are listed in **Table 106-1**. Each of these entities accounts for <1% of hematologic neoplasms.

TABLE 106-1

Unusual Lymphoid and Myeloid Malignancies

Lymphoid
Mature B-cell neoplasms
B-cell prolymphocytic leukemia
Splenic marginal zone lymphoma
Hairy cell leukemia
Nodal marginal zone B-cell lymphoma
Mediastinal large B-cell lymphoma
Intravascular large B-cell lymphoma
Primary effusion lymphoma
Lymphomatoid granulomatosis
Mature T-cell and natural killer (NK) cell neoplasms
T-cell prolymphocytic leukemia
T-cell large granular lymphocytic leukemia
Aggressive NK cell leukemia
Extranodal NK/T-cell lymphoma, nasal type
Enteropathy-type T-cell lymphoma
Hepatosplenic T-cell lymphoma
Subcutaneous panniculitis-like T-cell lymphoma
Blastic NK cell lymphoma
Primary cutaneous CD30+ T-cell lymphoma

Angioimmunoblastic T-cell lymphoma
Myeloid
Chronic neutrophilic leukemia
Chronic eosinophilic leukemia/hypereosinophilic syndrome
Histiocytic and Dendritic Cell Neoplasms
Histiocytic sarcoma
Langerhans cell histiocytosis
Langerhans cell sarcoma
Interdigitating dendritic cell sarcoma
Follicular dendritic cell sarcoma
Mast Cells
Mastocytosis
Cutaneous mastocytosis
Systemic mastocytosis
Mast cell sarcoma
Extracutaneous mastocytoma

RARE LYMPHOID MALIGNANCIES

All the lymphoid tumors discussed here are mature B-cell or T-cell natural killer (NK) cell neoplasms.

MATURE B-CELL NEOPLASMS

B-Cell Prolymphocytic Leukemia (B-PLL)

This is a malignancy of medium-sized (about twice the size of a normal small lymphocyte), round lymphocytes with a prominent nucleolus and light blue cytoplasm on Wright's stain. It dominantly affects the blood, bone marrow (BM), and spleen and usually does not cause adenopathy. The median age of affected patients is 70 years, and men are more often affected than women (male-to-female ratio is 1.6). This entity is distinct from chronic lymphoid leukemia (CLL) and does not develop as a consequence of that disease.

Clinical presentation is generally from symptoms of splenomegaly or incidental detection of an elevated white blood cell (WBC) count. The clinical course can be rapid. The cells express surface IgM (with or without IgD) and typical B-cell markers (CD19, CD20, CD22). CD23 is absent, and about one-third of cases express CD5. The CD5 expression along with the presence of the t(11;14) translocation in 20% of cases leads to confusion in distinguishing B-PLL from the leukemic form of mantle cell lymphoma. No reliable criteria for the distinction have emerged and gene expression studies suggest a close relationship between mantle cell lymphoma and B-PLL and significant differences with CLL. About half of patients have mutation or loss of p53, and deletions have been noted in 11q23 and 13q14. Nucleoside analogues like fludarabine and cladribine and combination chemotherapy (cyclophosphamide, doxorubicin, vincristine, and prednisone [CHOP]) have produced responses. CHOP plus rituximab may be more effective than CHOP alone, but the disease is sufficiently rare that large series have not been reported. Splenectomy can produce palliation of symptoms but appears to have little or no impact on the course of the disease. BM transplantation may be curative. Imatinib may also have activity.

Splenic Marginal Zone Lymphoma (SMZL)

This tumor of mainly small lymphocytes originates in the marginal zone of the spleen white pulp, grows to efface the germinal centers and mantle, and invades the red pulp. Splenic hilar nodes, BM, and peripheral blood (PB) may be involved. The circulating tumor cells have short surface villi and are called villous lymphocytes. **Table 106-2** shows differences in tumor cells of a number of neoplasms of small lymphocytes that aid in the differential diagnosis. SMZL cells express surface immunoglobulin and CD20, but are negative for CD5, CD10, CD43, and CD103. Lack of CD5 distinguishes SMZL from CLL, and lack of CD103 separates SMZL from hairy cell leukemia.

TABLE 106-2

Immunophenotype of Tumors of Small Lymphocytes

	CD5	CD20	CD43	CD10	CD103	sIg	CyclinD1
Follicular lymphoma	neg	pos	pos	pos	neg	pos	neg
Chronic lymphoid leukemia	pos	pos	pos	neg	neg	pos	neg
B-cell prolymphocytic leukemia	pos	pos	pos	neg	neg	pos	pos
Mantle cell lymphoma	pos	pos	pos	neg	neg	pos	pos
Splenic marginal zone lymphoma	neg	pos	neg	neg	neg	pos	neg
Hairy cell leukemia	neg	pos	?	neg	pos	pos	neg

Abbreviations: neg, negative; pos, positive.

The median age of patients with SMZL is mid-fifties, and men and women are equally represented. Patients present with incidental or symptomatic splenomegaly or incidental detection of lymphocytosis in the PB with villous lymphocytes. Autoimmune anemia or thrombocytopenia may be present. The immunoglobulin produced by these cells contains somatic mutations that reflect transit through a germinal center, and ongoing mutations suggest that the mutation machinery has remained active. About 40% of patients have either deletions or translocations involving 7q21, the site of the *FLNC* gene (filamin Cγ, involved in cross-linking actin filaments in the cytoplasm). *NOTCH2* mutations are seen in 25% of patients. Chromosome 8p deletions may also be noted. The genetic lesions typically found in extranodal marginal zone lymphomas (e.g., trisomy 3 and t[11;18]) are uncommon in SMZL.

The clinical course of disease is generally indolent with median survivals exceeding 10 years. Patients with elevated lactate dehydrogenase (LDH) levels, anemia, and hypoalbuminemia generally have a poorer prognosis. Long remissions can be seen after splenectomy. Rituximab is also active. A small fraction of patients undergo histologic progression to diffuse large B-cell lymphoma with a concomitant change to a more aggressive natural history. Experience with combination chemotherapy in SMZL is limited.

Hairy Cell Leukemia

Hairy cell leukemia is a tumor of small lymphocytes with oval nuclei, abundant cytoplasm, and distinctive membrane projections (hairy cells). Patients have splenomegaly and diffuse BM involvement. While some circulating cells are noted, the clinical picture is dominated by symptoms from the enlarged spleen and pancytopenia. The mechanism of the pancytopenia is not completely clear and may be mediated by both

inhibitory cytokines and marrow replacement. The marrow has an increased level of reticulin fibers; indeed, hairy cell leukemia is a common cause of inability to aspirate BM or so-called “dry tap” (**Table 106-3**). Monocytopenia is profound and may explain a predisposition to atypical mycobacterial infection that is observed clinically. The tumor cells have strong expression of CD22, CD25, and CD103; soluble CD25 level in serum is an excellent tumor marker for disease activity. The cells also express tartrate-resistant acid phosphatase. The immunoglobulin genes are rearranged and mutated, indicating the influence of a germinal center. No specific cytogenetic abnormality has been found, but most cases contain the activating *BRAF* mutation V600E.

TABLE 106-3

Differential Diagnosis of “Dry Tap”—Inability to Aspirate Bone Marrow

Dry taps occur in about 4% of attempts and are associated with:	
Metastatic carcinoma infiltration	17%
Chronic myeloid leukemia	15%
Myelofibrosis	14%
Hairy cell leukemia	10%
Acute leukemia	10%
Lymphomas, Hodgkin’s disease	9%
Normal marrow	Rare

The median age of affected patients is mid-fifties, and the male-to-female ratio is 5:1. Treatment options are numerous. Splenectomy is often associated with prolonged remission. Nucleosides including cladribine and deoxycoformycin are highly active but are also associated with further immunosuppression and can increase the risk of certain opportunistic infections. However, after brief courses of these agents, patients usually obtain very durable remissions during which immune function spontaneously recovers. Interferon α is also an effective therapy but is not as effective as nucleosides. Chemotherapy-refractory patients have responded to vemurafenib, a BRAF inhibitor. It is not yet clear if vemurafenib can induce long-term remissions without continuous treatment.

Nodal Marginal Zone B-Cell Lymphoma

This rare node-based disease bears an uncertain relationship with extranodal marginal zone lymphomas, which are often mucosa-associated and are called mucosa-associated lymphoid tissue (MALT) lymphomas,

and SMZLs. Patients may have localized or generalized adenopathy. The neoplastic cell is a marginal zone B cell with monocytoid features and has been called monocytoid B-cell lymphoma in the past. Up to one-third of the patients may have extranodal involvement, and involvement of the lymph nodes can be secondary to the spread of a mucosal primary lesion. In authentic nodal primaries, the cytogenetic abnormalities associated with MALT lymphomas (trisomy 3 and t[11;18]) are very rare. The clinical course is indolent. Patients often respond to combination chemotherapy, although remissions have not been durable. Few patients have received CHOP plus rituximab, which is likely to be an effective approach to management.

Mediastinal (Thymic) Large B-Cell Lymphoma

This entity was originally considered a subset of diffuse large B-cell lymphoma; however, additional study has identified it as a distinct entity with its own characteristic clinical, genetic, and immunophenotypic features. This is a disease that can be bulky in size but usually remains confined to the mediastinum. It can be locally aggressive, including progressing to produce a superior vena cava obstruction syndrome or pericardial effusion. About one-third of patients develop pleural effusions, and 5–10% can disseminate widely to kidney, adrenal, liver, skin, and even brain. The disease affects women more often than men (male-to-female ratio is 1:2–3), and the median age is 35–40 years.

The tumor is composed of sheets of large cells with abundant cytoplasm accompanied by variable, but often abundant, fibrosis. It is distinguished from nodular sclerosing Hodgkin's disease by the paucity of normal lymphoid cells and the absence of lacunar variants of Reed-Sternberg cells. However, more than one-third of the genes that are expressed to a greater extent in primary mediastinal large B-cell lymphoma than in usual diffuse large B-cell lymphoma are also overexpressed in Hodgkin's disease, suggesting a possible pathogenetic relationship between the two entities that affect the same anatomic site. Tumor cells may overexpress *MAL*. The genome of tumor cells is characterized by frequent chromosomal gains and losses. The tumor cells in mediastinal large B-cell lymphoma express CD20, but surface immunoglobulin and HLA class I and class II molecules may be absent or incompletely expressed. Expression of lower levels of class II HLA identifies a subset with poorer prognosis. The cells are CD5 and CD10 negative but may show light staining with anti-CD30. The cells are CD45 positive, unlike cells of classical Hodgkin's disease.

Methotrexate, leucovorin, doxorubicin, cyclophosphamide, vincristine, prednisone, and bleomycin (MACOP-B) and rituximab plus CHOP are effective treatments, achieving 5-year survival of 75–87%. Dose-adjusted therapy with prednisone, etoposide, vincristine, cyclophosphamide, and doxorubicin (EPOCH) plus rituximab has produced 5-year survival of 97%. A role for mediastinal radiation therapy has not been definitively demonstrated, but it is frequently used, especially in patients whose mediastinal area remains positron emission tomography–avid after 4–6 cycles of chemotherapy.

Intravascular Large B-Cell Lymphoma

This is an extremely rare form of diffuse large B-cell lymphoma characterized by the presence of lymphoma in the lumen of small vessels, particularly capillaries. It is also known as malignant angioendotheliomatosis or angiotropic large cell lymphoma. It is sufficiently rare that no consistent picture has emerged to define a

clinical syndrome or its epidemiologic and genetic features. It is thought to remain inside vessels because of a defect in adhesion molecules and homing mechanisms, an idea supported by scant data suggesting absence of expression of β-1 integrin and ICAM-1. Patients commonly present with symptoms of small-vessel occlusion, skin lesions, or neurologic symptoms. The tumor cell clusters can promote thrombus formation. In general, the clinical course is aggressive and the disease is poorly responsive to therapy. Often a diagnosis is not made until very late in the course of the disease.

Primary Effusion Lymphoma

This entity is another variant of diffuse large B-cell lymphoma that presents with pleural effusions, usually without apparent tumor mass lesions. It is most common in the setting of immune deficiency disease, especially AIDS, and is caused by human herpes virus 8 (HHV-8)/Kaposi's sarcoma herpes virus (KSHV). It is also known as *body cavity–based lymphoma*. Some patients have been previously diagnosed with Kaposi's sarcoma. It can also occur in the absence of immunodeficiency in elderly men of Mediterranean heritage, similar to Kaposi's sarcoma but even less common.

The malignant effusions contain cells positive for HHV-8/KSHV, and many are also co-infected with Epstein-Barr virus. The cells are large with large nuclei and prominent nucleoli that can be confused with Reed-Sternberg cells. The cells express CD20 and CD79a (immunoglobulin-signaling molecule), although they often do not express immunoglobulin. Some cases aberrantly express T-cell markers such as CD3 or rearranged T-cell receptor genes. No characteristic genetic lesions have been reported, but gains in chromosome 12 and X material has been seen, similar to other HIV-associated lymphomas. The clinical course is generally characterized by rapid progression and death within 6 months. CHOP plus lenalidomide or bortezomib may produce responses. HAART therapy for HIV should be maintained during treatment.

Lymphomatoid Granulomatosis

This is an angiocentric, angiodestructive lymphoproliferative disease comprised by neoplastic Epstein-Barr virus–infected monoclonal B cells accompanied and outnumbered by a polyclonal reactive T-cell infiltrate. The disease is graded based on histologic features such as cell number and atypia in the B cells. It is most often confused with extranodal NK/T-cell lymphoma, nasal type, which can also be angiodestructive and is Epstein-Barr virus–related. The disease usually presents in adults (males > females) as a pulmonary infiltrate. Involvement is often entirely extranodal and can include kidney (32%), liver (29%), skin (25%), and brain (25%). The disease often but not always occurs in the setting of immune deficiency.

The disease can be remitting and relapsing in nature or can be rapidly progressive. The course is usually predicted by the histologic grade. The disease is highly responsive to combination chemotherapy and is curable in most cases. Some investigators have claimed that low-grade disease (grade I and II) can be treated with interferon α.

MATURE T-CELL AND NK CELL NEOPLASMS

T-Cell Prolymphocytic Leukemia

This is an aggressive leukemia of medium-sized prolymphocytes involving the blood, marrow, nodes, liver, spleen, and skin. It accounts for 1–2% of all small lymphocytic leukemias. Most patients present with elevated WBC count (often >100,000/μL), hepatosplenomegaly, and adenopathy. Skin involvement occurs in 20%. The diagnosis is made from PB smear, which shows cells about 25% larger than those in small lymphocytes, with cytoplasmic blebs and nuclei that may be indented. The cells express T-cell markers like CD2, CD3, and CD7; two-thirds of patients have cells that are CD4+ and CD8–, and 25% have cells that are CD4+ and CD8+. T-cell receptor β chains are clonally rearranged. In 80% of patients, inversion of chromosome 14 occurs between q11 and q32. Ten percent have t(14;14) translocations that bring the T-cell receptor alpha/beta gene locus into juxtaposition with oncogenes *TCL1* and *TCL1b* at 14q32.1. Chromosome 8 abnormalities are also common. Deletions in the *ATM* gene are also noted. Activating *JAK3* mutations have also been reported.

The course of the disease is generally rapid, with median survival of about 12 months. Responses have been seen with the anti-CD52 antibody, alemtuzumab, nucleoside analogs, and CHOP chemotherapy. Histone deacetylase inhibitors like vorinostat and romidepsin may also have activity. Small numbers of patients with T-cell prolymphocytic leukemia have also been treated with high-dose therapy and allogeneic BM transplantation after remission has been achieved with conventional-dose therapy.

T-Cell Large Granular Lymphocytic Leukemia

T-cell large granular lymphocytic leukemia (LGL leukemia) is characterized by increases in the number of LGLs in the PB (2000–20,000/μL) often accompanied by severe neutropenia, with or without concomitant anemia. Patients may have splenomegaly and frequently have evidence of systemic autoimmune disease, including rheumatoid arthritis, hypergammaglobulinemia, autoantibodies, and circulating immune complexes. BM involvement is mainly interstitial in pattern, with fewer than 50% lymphocytes on differential count. Usually the cells express CD3, T-cell receptors, and CD8; NK-like variants may be CD3–. The leukemic cells often express Fas and Fas ligand.

The course of the disease is generally indolent and dominated by the neutropenia. Paradoxically, immunosuppressive therapy with cyclosporine, methotrexate, or cyclophosphamide plus glucocorticoids can produce an increase in granulocyte counts. Nucleosides have been used anecdotally. Occasionally the disease can accelerate to a more aggressive clinical course.

Aggressive NK Cell Leukemia

NK neoplasms are very rare, and they may follow a range of clinical courses from very indolent to highly aggressive. They are more common in Asians than whites, and the cells frequently harbor a clonal Epstein-Barr virus episome. The PB white count is usually not greatly elevated, but abnormal large lymphoid cells with granular cytoplasm are noted. The aggressive form is characterized by symptoms of fever and laboratory abnormalities of pancytopenia. Hepatosplenomegaly is common; node involvement is less common. Patients may have hemophagocytosis, coagulopathy, or multiorgan failure. Serum levels of Fas ligand are elevated.

The cells express CD2 and CD56 and do not have rearranged T-cell receptor genes. Deletions involving chromosome 6 are common. The disease can be rapidly progressive. Some forms of NK neoplasms are more indolent. They tend to be discovered incidentally with LGL lymphocytosis and do not manifest the fever and hepatosplenomegaly characteristic of the aggressive leukemia. The cells are also CD2 and CD56 positive, but they do not contain clonal forms of Epstein-Barr virus and are not accompanied by pancytopenia or autoimmune disease.

Extranodal NK/T-Cell Lymphoma, Nasal Type

Like lymphomatoid granulomatosis, extranodal NK/T-cell lymphoma tends to be an angiocentric and angiodestructive lesion, but the malignant cells are not B cells. In most cases, they are CD56+ Epstein-Barr virus–infected cells; occasionally they are CD56–Epstein-Barr virus–infected cytotoxic T cells. They are most commonly found in the nasal cavity. Historically, this illness was called lethal midline granuloma, polymorphic reticulosis, and angiocentric immunoproliferative lesion. This form of lymphoma is prevalent in Asia, Mexico, and Central and South America; it affects males more commonly than females. When it spreads beyond the nasal cavity, it may affect soft tissue, the gastrointestinal tract, or the testis. In some cases, hemophagocytic syndrome (HPS) may influence the clinical picture. Patients may have B symptoms. Many of the systemic manifestations of disease are related to the production of cytokines by the tumor cells and the cells responding to their signals. Deletions and inversions of chromosome 6 are common.

Many patients with extranodal NK/T-cell lymphoma, nasal type have excellent antitumor responses with combination chemotherapy regimens, particularly those with localized disease. Radiation therapy is often used after completion of chemotherapy. Four risk factors have been defined, including B symptoms, advanced stage, elevated LDH, and regional lymph node involvement. Patient survival is linked to the number of risk factors: 5-year survival is 81% for zero risk factors, 64% for one risk factor, 32% for two risk factors, and 7% for three or four risk factors. Combination regimens without anthracyclines have been touted as superior to CHOP, but data are sparse. High-dose therapy with stem cell transplantation has been used, but its role is unclear.

Enteropathy-Type T-Cell Lymphoma

Enteropathy-type T-cell lymphoma is a rare complication of longstanding celiac disease. It most commonly occurs in the jejunum or the ileum. In adults, the lymphoma may be diagnosed at the same time as celiac disease, but the suspicion is that the celiac disease was a longstanding precursor to the development of lymphoma. The tumor usually presents as multiple ulcerating mucosal masses, but may also produce a dominant exophytic mass or multiple ulcerations. The tumor expresses CD3 and CD7 nearly always and may or may not express CD8. The normal-appearing lymphocytes in the adjacent mucosa often have a similar phenotype to the tumor. Most patients have the HLA genotype associated with celiac disease, HLA DQA1*0501 or DQB1*0201.

The prognosis of this form of lymphoma is typically (median survival is 7 months) poor, but some patients have a good response to CHOP chemotherapy. Patients who respond can develop bowel perforation from

responding tumor. If the tumor responds to treatment, recurrence may develop elsewhere in the celiac disease–affected small bowel.

Hepatosplenic T-Cell Lymphoma

Hepatosplenic T-cell lymphoma is a malignancy derived from T cells expressing the gamma/delta T-cell antigen receptor that affects mainly the liver and fills the sinusoids with medium-size lymphoid cells. When the spleen is involved, dominantly the red pulp is infiltrated. It is a disease of young people, especially young people with an underlying immunodeficiency or with an autoimmune disease that demands immunosuppressive therapy. The use of thiopurine and infliximab is particularly common in the history of patients with this disease. The cells are CD3+ and usually CD4– and CD8–. The cells may contain isochromosome 7q, often together with trisomy 8. The lymphoma has an aggressive natural history. Combination chemotherapy may induce remissions, but most patients relapse. Median survival is about 2 years. The tumor does not appear to respond to reversal of immunosuppressive therapy.

Subcutaneous Panniculitis-Like T-Cell Lymphoma

Subcutaneous panniculitis-like T-cell lymphoma involves multiple subcutaneous collections of neoplastic T cells that are usually cytotoxic cells in phenotype (i.e., contain perforin and granzyme B and express CD3 and CD8). The rearranged T-cell receptor is usually alpha/beta-derived, but occasionally the gamma/delta receptors are involved, particularly in the setting of immunosuppression. The cells are negative for Epstein-Barr virus. Patients may have a HPS in addition to the skin infiltration; fever and hepatosplenomegaly may also be present. Nodes are generally not involved. Patients frequently respond to combination chemotherapy, including CHOP. When the disease is progressive, the HPS can be a component of a fulminant downhill course. Effective therapy can reverse the HPS.

Blastic NK Cell Lymphoma

The neoplastic cells express NK cell markers, especially CD56, and are CD3 negative. They are large blastic-appearing cells and may produce a leukemia picture, but the dominant site of involvement is the skin. Morphologically, the cells are similar to the neoplastic cells in acute lymphoid and myeloid leukemia. No characteristic chromosomal abnormalities have been described. The clinical course is rapid, and the disease is largely unresponsive to typical lymphoma treatments.

Primary Cutaneous CD30+ T-Cell Lymphoma

This tumor involves the skin and is composed of cells that appear similar to the cells of anaplastic T-cell lymphoma. Among cutaneous T-cell tumors, about 25% are CD30+ anaplastic lymphomas. If dissemination to lymph nodes occurs, it is difficult to distinguish between the cutaneous and systemic forms of the disease. The tumor cells are often CD4+, and the cells contain granules that are positive for granzyme B and perforin in 70% of cases. The typical t(2;5) of anaplastic T-cell lymphoma is absent; indeed, its presence should prompt a closer look for systemic involvement and a switch to a diagnosis of anaplastic T-cell lymphoma.

This form of lymphoma has sporadically been noted as a rare complication of silicone on saline breast implants. The natural history of breast implant associated lymphoma is generally indolent. Cutaneous CD30+ T-cell lymphoma often responds to therapy. The anti-CD30 immunotoxin conjugate, brentuximab vedotin, is active. Radiation therapy can be effective, and surgery can also produce long-term disease control. Five-year survival exceeds 90%.

Angioimmunoblastic T-Cell Lymphoma

Angioimmunoblastic T-cell lymphoma is a systemic disease that accounts for about 15% of all T-cell lymphomas. Patients frequently have fever, advanced stage, diffuse adenopathy, hepatosplenomegaly, skin rash, polyclonal hypergammaglobulinemia, and a wide range of autoantibodies including cold agglutinins, rheumatoid factor, and circulating immune complexes. Patients may have edema, arthritis, pleural effusions, and ascites. The nodes contain a polymorphous infiltrate of neoplastic T cells and nonneoplastic inflammatory cells together with proliferation of high endothelial venules and follicular dendritic cells (FDCs). The most common chromosomal abnormalities are trisomy 3, trisomy 5, and an extra X chromosome. Aggressive combination chemotherapy can induce regressions. The underlying immune defects make conventional lymphoma treatments more likely to produce infectious complications.

RARE MYELOID MALIGNANCIES

The World Health Organization (WHO) system uses PB counts and smear analysis, BM morphology, cytogenetic and molecular genetic tests in order to classify myeloid malignancies into several major categories **(Table 106-4)**. Amongst them, acute myeloid leukemia (AML) is discussed in **Chap. 100**, MDS in **Chap. 98**, chronic myeloid leukemia (CML) in **Chap. 101**, and *JAK2* mutation-enriched myeloproliferative neoplasms (MPN) in **Chap. 99**. In this chapter, we focus on the rest (listed in Table 106-4) including chronic neutrophilic leukemia (CNL), "atypical CML, *BCR-ABL1* negative (aCML)," chronic myelomonocytic leukemia (CMML), juvenile myelomonocytic leukemia (JMML), "chronic eosinophilic leukemia, not otherwise specified (CEL-NOS)," mastocytosis, "MPN, unclassifiable (MPN-U)," MDS/MPN, unclassifiable (MDS/MPN-U), "MDS/MPN with ring sideroblasts and thrombocytosis (MDS/MPN-RS-T)," and "myeloid/lymphoid neoplasms with eosinophilia and rearrangements of *PDGFRA, PDGFRB*, *FGFR1* or with *PCM1-JAK2.*" This chapter also includes histiocytic and DC neoplasms, transient myeloproliferative disorders (TMD) as well as a broader discussion on primary eosinophilic disorders including hypereosinophilic syndrome (HES).

TABLE 106-4

World Health Organization Classification of Myeloid Malignancies

1. **Acute myeloid leukemia (AML) and related precursor neoplasms**
2. **Myeloproliferative neoplasms (MPN)**
   - **2.1.** Chronic myeloid leukemia (CML), *BCR-ABL1* positive
   - **2.2.** *JAK2* mutation-enriched MPN
     - **2.2.1.** Polycythemia vera
     - **2.2.2.** Primary myelofibrosis
     - **2.2.3.** Essential thrombocythemia
   - **2.3.** Chronic neutrophilic leukemia (CNL)
   - **2.4.** Chronic eosinophilic leukemia, not otherwise specified (CEL-NOS)
   - **2.5.** Myeloproliferative neoplasm, unclassifiable (MPN-U)
3. **Myelodysplastic syndromes (MDS)**
   - **3.1.** MDS with single lineage dysplasia
   - **3.2.** MDS with ring sideroblasts (MDS-RS)
   - **3.3.** MDS with multilineage dysplasia
   - **3.4.** MDS with excess blasts
   - **3.5.** MDS with isolated del(5q)
   - **3.6.** MDS, unclassifiable (MDS-U)
   - **3.7.** *Provisional entity: Refractory cytopenia of childhood*
4. **MDS/MPN overlap**
   - **4.1.** Chronic myelomonocytic leukemia (CMML)
   - **4.2.** Atypical chronic myeloid leukemia (aCML), *BCR-ABL1* negative
   - **4.3.** Juvenile myelomonocytic leukemia (JMML)
   - **4.4.** MDS/MPN with ring sideroblasts and thrombocytosis (MDS/MPN-RS-T)
   - **4.5.** MDS/MPN, unclassifiable (MDS/MPN-U)
5. **Mastocytosis**
6. **Myeloid/lymphoid neoplasms with eosinophilia and rearrangement of *PDGFRA, PDGFRB*, *FGFR1* or with *PCM1-JAK2***
   - **6.1.** Myeloid/lymphoid neoplasms with *PDGFRA* rearrangement
   - **6.2.** Myeloid/lymphoid neoplasms with *PDGFRB* rearrangement
   - **6.3.** Myeloid/lymphoid neoplasms with *FGFR1* rearrangement
   - **6.4.** *Provisional entity: Myeloid/lymphoid neoplasms with PCM1-JAK2 translocation*

Myeloid neoplasms with germline predisposition

CHRONIC NEUTROPHILIC LEUKEMIA

CNL is a clonal proliferation of mature neutrophils with few or no circulating immature granulocytes. In 2013, CNL was associated with activating mutations of the gene (*CSF3R*) encoding for the receptor for granulocyte colony-stimulating factor (G-CSF), also known as colony stimulating factor 3 (CSF3). Patients with CNL might be asymptomatic at presentation but also display constitutional symptoms, splenomegaly, anemia, and thrombocytopenia. Median survival is ~2 years and causes of death include leukemic transformation, progressive disease associated with severe cytopenias, and marked treatment-refractory leukocytosis. CNL is rare with <200 reported cases. Median age at diagnosis is ~67 years, and the disease is equally prevalent in both genders.

Pathogenesis

CSF3 is the main growth factor for granulocyte proliferation and differentiation. Accordingly, recombinant CSF3 is used for the treatment of severe neutropenia, including severe congenital neutropenia (SCN). Some patients with SCN acquire *CSF3R* mutations and the frequency of such mutations is significantly higher (~ 80%) in those patients who experience leukemic transformation. SCN-associated *CSF3R* mutations occur in the region of the gene coding for the cytoplasmic domain of CSF3R, and result in truncation of the C-terminal-negative regulatory domain. In 2013, Maxson et al. described a different class of *CSF3R* mutations in ~ 90% of patients with CNL; these were mostly membrane proximal, the most frequent being a C-to-T substitution at nucleotide 1853 (T618I). In a subsequent confirmatory study, *CSF3R* mutations were found to be specific to WHO-defined CNL. About 40% of the T618I-mutated cases also harbored *SETBP1* mutations. *CSF3R*T618I has been shown to induce lethal myeloproliferative disorder in a mouse model and in vitro sensitivity to JAK inhibition.

Diagnosis

Diagnosis of CNL requires exclusion of the more common causes of neutrophilia including infections and inflammatory processes. In addition, one should be mindful of the association between some forms of metastatic cancer or plasma cell neoplasms with secondary neutrophilia. Neoplastic neutrophilia also occurs in other myeloid malignancies including aCML and CMML. Accordingly, the WHO diagnostic criteria for CNL are designed to exclude the possibilities of both secondary/reactive neutrophilia and leukocytosis associated with myeloid malignancies other than CNL (Table 106-2): leukocytosis (≥25 × 10^9/L), ≥80% segmented/band neutrophils, <10% immature myeloid cells, <1% circulating blasts and absence of dysgranulopoiesis or monocytosis (monocyte count <1 × 10^9/L). BM in CNL is hypercellular and displays increased number and percentage of neutrophils with very high myeloid to erythroid ratio and minimal left shift, myeloid dysplasia or reticulin fibrosis.

The recent discovery of *CSF3R* mutations (see above) and their almost invariable association with WHO-defined CNL has allowed its incorporation in the WHO diagnostic criteria **(Table 106-5)**. In practical terms, the presence of a membrane proximal *CSF3R* mutation in a patient with predominantly neutrophilic granulocytosis should be sufficient for the diagnosis of CNL, regardless of the degree of leukocytosis.

Unfortunately, several exclusionary criteria still need to be met for diagnosing CNL in the absence of *CSF3R* mutations (Table 106-5).

TABLE 106-5

2016 World Health Organization (WHO) Diagnostic Criteria for Chronic Neutrophilic Leukemia (CNL), Atypical Chronic Myeloid Leukemia, *BCR-ABL1*-negative (aCML), and Chronic Myelomonocytic Leukemia (CMML)

Variables	Chronic Neutrophilic Leukemia	Atypical Chronic Myeloid Leukemia	Chronic Myelomonocytic Leukemia
PB leukocyte count	≥25 × 10^9/L	Granulocytosis	
PB segmented neutrophils/bands	≥80%		
PB immature granulocytes[a]	<10%	≥10%	
PB blast count	<1%	<20%	<20%
PB monocyte count	<1 x 10(9)/L	No or minimal monocytosis	≥1 × 10^9/L Persistent and lasting for at least 3 months
Dysgranulopoiesis	No	Yes	
PB basophil percentage		<2%	
PB monocyte percentage		<10%	≥10%
Bone marrow	Hypercellular ↑Neutrophils, number and % <5% blasts Normal neutrophilic maturation	Hypercellular ↑Granulocyte proliferation Granulocytic dysplasia ±erythroid/megakaryocyte Dysplasia <20% blasts	Dysplasia in ≥1 myeloid lineages or clonal cytogenetic/molecular abnormality <20% blasts or promonocytes
BCR-ABL1	No	No	No

Variables	Chronic Neutrophilic Leukemia	Atypical Chronic Myeloid Leukemia	Chronic Myelomonocytic Leukemia
PDGFRA, *PDGFRB*, *FGFR1*, or *PCM1-JAK2* rearrangement	No	No	No
CSF3R T618I or other activating CSF3R mutation or persistent neutrophilia, splenomegaly, and no identifiable cause of reactive neutrophilia	Yes		
PB and BM blasts/promonocytes		<20%	<20%
Evidence for other MPN: CML, PV, ET, or PMF	No	No	No
Evidence for reactive Leukocytosis[b] or monocytosis	No		No

[a]Immature granulocytes include myeloblasts, promyelocytes, myelocytes, and metamyelocytes. [b]Causes of reactive neutrophilia include plasma cell neoplasms, solid tumor, infections, and inflammatory processes.

Abbreviations: BM, bone marrow; CML, chronic myeloid leukemia; ET, essential thrombocythemia; MPN, myeloproliferative neoplasms; PB, peripheral blood; PMF, primary myelofibrosis; PV, polycythemia vera.

Treatment

Current treatment in CNL is largely palliative and suboptimal in its efficacy. Several drugs alone or in combination have been tried and none have shown remarkable efficacy. As such, allogeneic stem cell transplant (ASCT) is reasonable to consider in the presence of symptomatic disease, especially in younger patients. Otherwise, cytoreductive therapy with hydroxyurea is probably as good as anything, and a more intensive combination chemotherapy may not have additional value. However, response to hydroxyurea therapy is often transient, and some have successfully used interferon α as an alternative drug. Response to treatment with ruxolitinib (a JAK1 and JAK2 inhibitor) has been reported in several case reports but, as is the case with hydroxyurea treatment, the response is often incomplete and temporary.

ATYPICAL CHRONIC MYELOID LEUKEMIA

"Atypical chronic myeloid leukemia, *BCR-ABL1* negative (aCML)" is formally classified under the MDS/MPN category of myeloid malignancies and is characterized by left shifted granulocytosis and dysgranulopoiesis. The differential diagnosis of aCML includes CML, which is distinguished by the presence of *BCR-ABL1*; CNL, which is distinguished by the absence of dysgranulopoiesis and presence of *CSF3R* mutations; and CMML, which is distinguished by the presence of monocytosis (absolute monocyte count ≥1 × 10^9/L). The WHO diagnostic criteria for aCML are listed in Table 106-5 and include granulocytosis, dysgranulopoiesis, ≥10% immature granulocytes, <20% PB or BM myeloblasts, <10% PB monocytes, <2% basophils, absence of otherwise specific mutations such as *BCR-ABL1, PDGFRA, PDGFRB, FGFR1, or PCM1-JAK2* and not meeting WHO criteria for CML, PMF, PV, or ET. The BM in aCML is hypercellular with granulocyte proliferation and dysplasia with or without erythroid or megakaryocytic dysplasia.

The molecular pathogenesis of aCML is incompletely understood; about a fourth of the patients express *SETBP1* mutations, which are, however, also found in several other myeloid malignancies, including CNL and CMML. *SETBP1* mutations in aCML were prognostically detrimental and mostly located between codons 858 and 871; similar mutations are seen with Schinzel-Giedion syndrome (a congenital disease with severe developmental delay and various physical stigmata including midface retraction, large forehead, and macroglossia). A somatic missense mutation in ethanolamine kinase 1 (*ETNK1*N244S) was described in 9% of patients with aCML but was also seen in 14% of patients with CMML, 6% of patients with mastocytosis (especially in association with eosinophilia), and rarely in other MPN.

In a series of 55 patients with WHO-defined aCML, median age at diagnosis was 62 years with female preponderance (57%), splenomegaly was reported in 54% of the patients, red cell transfusion requirement in 65%, abnormal karyotype in 20% (20q- and trisomy 8 being the most frequent) and leukemic transformation in 40%. Median survival was 25 months. Outcome was worse in patients with marked leukocytosis, transfusion requirement, and increased immature cells in the PB. Conventional chemotherapy is largely ineffective in the treatment of aCML. However, a favorable experience with ASCT was reported in nine patients; after a median follow-up of 55 months, the majority of the patients remained in complete remission.

CHRONIC MYELOMONOCYTIC LEUKEMIA

CMML is classified under the WHO category of MDS/MPN and is defined by an absolute monocyte count (AMC) of ≥1 × 10^9/L in the PB and accounting for ≥10% of the leukocyte count. Median age at diagnosis ranges between 65 and 75 years and there is a 2:1 male predominance. Clinical presentation is variable and depends on whether the disease presents with MDS-like or MPN-like phenotype; the former is associated with cytopenias and the latter with splenomegaly and features of myeloproliferation such as fatigue, night sweats, weight loss, and cachexia. About 20% of patients with CMML experience serositis involving the joints (arthritis), pericardium (pericarditis and pericardial effusion), pleura (pleural effusion), or peritoneum (ascites).

Pathogenesis

Clonal cytogenetic abnormalities are seen in about a third of patients with CMML and include trisomy 8 and abnormalities of chromosome 7. Almost all patients with CMML harbor somatic mutations involving epigenetic regulator genes (e.g., *ASXL1, TET2*), spliceosome pathway genes (e.g., *SRSF2)*, DNA damage response genes (e.g., *TP53*), and tyrosine kinases/transcription factors (e.g., *KRAS, NRAS, CBL*, and *RUNX1*). However, none of these mutations are specific to CMML, and their precise pathogenetic contribution is unclear.

Diagnosis

Reactive monocytosis is uncommon but has been reported in association with certain infections and inflammatory conditions. Clonal (i.e., neoplastic) monocytosis defines CMML but is also seen with JMML and AML with monocytic differentiation. The WHO diagnostic criteria for CMML are listed in Table 106-5 and include (1) persistent PB monocyte count of ≥1 × 10^9/L with monocyte percentage of ≥10%, (2) absence of *BCR-ABL1, PDGFRA, PDGFRB, FGFR1, or PCM1-JAK2 rearrangements*, (3) not meeting WHO criteria for CML, PV, ET, or PMF, (4) <20% blasts and promonocytes in the PB and BM, and (5) dysplasia involving one or more myeloid lineages or, in the absence of dysplasia, presence of an acquired clonal cytogenetic or molecular genetic abnormality or non-reactive monocytosis lasting for at least 3 months.

The BM in CMML is hypercellular with granulocytic and monocytic proliferation. Dysplasia is often present and may involve one, two, or all myeloid lineages. On immunophenotyping the abnormal cells often express myelomonocytic antigens such as CD13 and CD33, with variable expression of CD14, CD68, CD64, and CD163. Monocytic-derived cells are almost always positive for the cytochemical non-specific esterases (e.g., butyrate esterase), while normal granulocytic precursors are positive for lysozyme and chloroacetate esterase. In CMML, it is common to have a hybrid cytochemical staining pattern with cells expressing both chloroacetate and butyrate esterases simultaneously (dual esterase staining).

Prognosis

A recent meta-analysis showed median survival of 1.5 years in CMML. Numerous prognostic systems have attempted to better define and stratify the natural history of CMML. One of these, the Mayo prognostic model, assigns one point each to the following four independent prognostic variables: AMC >10 × 10^9/L, presence of circulating immature cells, hemoglobin <10 gm/dL and platelet count <100,000/mL. This model stratified patients into three risk groups: low (0 points), intermediate (1 point), and high (≥2 points), translating to median survival of 32, 18, and 10 months, respectively.

A French study incorporated *ASXL1* mutational status in 312 CMML patients. In a multivariable model, independent predictors of poor survival were WBC >15 × 10^9/L (3 points), *ASXL1* mutations (2 points), age >65 years (2 points), platelet count <100,000/mL (2 points), and hemoglobin <10 gm/dL in females and <11 gm/dL in males (2 points). This model stratified patients into three groups: low (0–4 points), intermediate (5–7 points), and high risk (8–12 points), with median survival of "not reached," 38.5 and 14.4 months, respectively.[27] *ASXL1* and *DNMT3A* mutations also have an adverse effect on CMML.

Treatment

Current treatment in CMML consists of hydroxyurea and supportive care, including red cell transfusions and use of erythropoiesis-stimulating agents (ESA). The value of hydroxyurea was reinforced by a randomized trial against oral etoposide. No other single or combination chemotherapy has been shown to be superior to hydroxyurea. ASCT is a viable treatment option for transplant-eligible patients with poor prognostic features. Given the MDS/MPN overlap phenotype and the presence of MDS-like genetic/methylation abnormalities in CMML, hypomethylating agents such as 5-azacitidine and decitabine have been used with limited efficacy; in a study using decitabine in CMML, overall response rate was 48% with 17% complete remissions and median survival of 17 months. The experience with 5-azacytidine was somewhat similar.

JUVENILE MYELOMONOCYTIC LEUKEMIA

JMML is primarily a disease of early childhood and is included, along with CMML, in the "MDS/MPN" WHO category. Both CMML and JMML feature leukocytosis, monocytosis, and hepatosplenomegaly. Additional characteristic features in JMML include thrombocytopenia and elevated fetal hemoglobin. Myeloid progenitors in JMML display GM-CSF hypersensitivity that has been attributed to dysregulated RAS/MAPK signaling. The latter is believed to result from mutually exclusive mutations involving *RAS*, *PTPN11*, and *NF1*. A third of patients with JMML that is not associated with Noonan syndrome carry *PTPN11* mutations while the incidence of *NF1* in patients without neurofibromatosis, type 1 and *RAS* mutations are ~15% each. In general, in about 85% of JMML cases, one of the classical RAS pathway mutations (*PTPN11*, *NRAS*, *KRAS*, *NF1*, and *CBL*) is present; in addition, a myrof other mutations, such as *ASXL1*, *RUNX1*, *SETBP1*, *JAK3*, *CUX1*, and others have been reported. Drug therapy is relatively ineffective in JMML, and the treatment of choice is ASCT, which results in a 5-year survival of ~50%.

The 2016 revised WHO diagnostic criteria for JMML requires the presence of PB monocyte count ≥1×10^9/L, <20% blasts in blood or BM, splenomegaly, and absence of *BCR-ABL1*. Diagnosis also requires the presence of one of the following: somatic mutation of *PTPN11, KRAS*, or *NRAS*; clinical diagnosis of NF1 or *NF1* mutation; germline mutation of *CBL* and loss of heterozygosity. Diagnosis of JMML can still be considered without the aforementioned genetic features, in the presence of monosomy 7 or any other cytogenetic abnormality or in the presence of two of the following: increased hemoglobin F, presence of myeloid or erythroid precursors in the PB, GM-CSF hypersensitivity in colony assay, and hyperphosphorylation of STAT5.

MDS/MPN, UNCLASSIFIABLE (MDS/MPN-U)

The WHO classifies patients with morphologic and laboratory features that resemble both MDS and MPN as "MDS/MPN overlap." This category includes CMML, aCML, and JMML, which have been described above. In addition, MDS/MPN includes a fourth category referred to as MDS/MPN, unclassifiable (MDS/MPN-U). Diagnosis of MDS/MPN-U requires the presence of both MDS and MPN features that are not adequate to classify patients as CMML, aCML, or JMML. MDS/MPN also includes the provisional category of refractory anemia with ring sideroblasts and thrombocytosis (RARS-T); the 2016 revision of the WHO classification document has changed the term RARS-T into "MDS/MPN-RS-T."

In a study of 85 patients with MDS/MPN-U, median age was 70 years and 72% were males. Splenomegaly at presentation was present in 33%, thrombocytosis in 13%, leukocytosis in 18%, *JAK2* mutations in 30%, and abnormal karyotype 51%; the most frequent cytogenetic abnormality was trisomy 8. Median survival was 12.4 months and favorably affected by thrombocytosis. Treatment with hypomethylating agents, immunomodulators, or ASCT did not appear to favorably affect survival.

MDS/MPN WITH RING SIDEROBLASTS AND THROMBOCYTOSIS (MDS/MPN-RS-T)

MDS/MPN-RS-T is classified in the MDS/MPN category because it shares dysplastic features with MDS-RS and myeloproliferative features with ET. The 2016 revised WHO diagnostic criteria for MDS/MPN-RS-T includes anemia associated with erythroid lineage dysplasia, presence of ≥15% ring sideroblasts, blast count of <5% in BM and <1% in the PB, platelet count of ≥450 × 10^9/L, and absence of *BCR-ABL1*, *PDGFRA, PDGFRB, FGFR1, PCM1-JAK2* mutations or t(3;3)(q21;q26), inv(3)(q21q26), or del(5q). These diagnostic criteria also require the absence of history of MPN, MDS, or other type of MDS/MPN and also either the presence of *SF3B1* mutation or absence of exposure to cytotoxic or other treatment that could be blamed for the morphologic abnormalities.

One hundred eleven patients with MDS/MPN-RS-T were compared with 33 patients with RARS. The frequency of *SF3B1* mutations in MDS/MPN-RS-T-T (87%) was similar to that in MDS-RS (85%). *JAK2 V617F* mutation was detected in 49% of MDS/MPN-RS-T patients (including 48% of those mutated for *SF3B1*), but none of those with MDS-RS. In MDS/MPN-RS-T, *SF3B1* mutations were more frequent in females (95%) than in males (77%), and mean ring sideroblast counts were higher in *SF3B1-mutated* patients. Median overall survival was 6.9 years in *SF3B1-mutated* vs 3.3 years in unmutated cases. Six-year survival was 67% in *JAK2*-mutated vs 32% in unmutated cases. Multivariable analysis identified younger age, *JAK2* and *SF3B1* mutations as favorable factors. Predictors of poor survival in MDS/MPN-RS-T include anemia, abnormal karyotype, and presence of *ASXL1* or *SETBP1* mutations. Interestingly, the presence of *SF3B1* mutations in MDS/MPN-RS-T is associated with increased risk of thrombosis. Several case reports have suggested that treatment with lenalidomide might induce red cell transfusion independence and complete remissions in MDS/MPN-RS-T.

MYELOPROLIFERATIVE NEOPLASM, UNCLASSIFIABLE (MPN-U)

The category of MPN-U includes MPN-like neoplasms that cannot be clearly classified as one of the other seven subcategories of MPN (Table 106-4). Examples include patients presenting with unusual thrombosis or unexplained organomegaly with normal blood counts but found to carry MPN-characteristic mutations such as *JAK2* and *CALR* or display BM morphology that is consistent with MPN. It is possible that some cases of MPN-U represent earlier disease stages in PV or ET, which however fail to meet the threshold hemoglobin levels or platelet counts that are required per WHO diagnostic criteria. Specific treatment interventions might not be necessary in asymptomatic patients with MPN-U, whereas patients with arterial thrombotic complications might require cytoreductive and aspirin therapy, and those with venous thrombosis might require systemic anticoagulation.

MYELOID NEOPLASMS WITH GERM LINE PREDISPOSITION

The 2016 WHO revision on the classification of myeloid neoplasms adds a section referred to as “myeloid neoplasms with germ line predisposition” and includes cases of AML, MDS, and MDS/MPN that arise in the setting of a germ line predisposition mutation, such as *CEBPA, DDX41, RUNX1, ANKRD26, ETV6*, or *GATA2*. This particular category of diseases also includes myeloid neoplasms that arise in the background of BM failure syndromes, Down syndrome, Noonan syndrome, neurofibromatosis, and telomeropathies.

TRANSIENT MYELOPROLIFERATIVE DISORDER

TMD, also referred to as transient abnormal myelopoiesis (TAM), constitutes an often but not always transient phenomenon of abnormal megakaryoblast proliferation, which occurs in ~10% of infants with Down syndrome. TMD is usually recognized at birth and either undergoes spontaneous regression (75% of the cases) or progress into acute megakaryoblastic leukemia (AMKL) (25% of the cases). Almost all patients with TMD and TMD-derived AMKL display somatic *GATA1* mutations. TMD-associated *GATA1* mutations constitute exon 2 insertions, deletions, or missense mutations, affecting the N-terminal transactivation domain of GATA-1 and result in loss of full-length (50-kD) GATA-1 and its replacement with a shorter isoform (40-kD) that retains friend of GATA-1 (FOG-1) binding. In contrast, inherited forms of exon 2 *GATA1* mutations produce a phenotype with anemia whereas exon 4 mutations that affect the N-terminal, FOG-1-interactive domain, produce familial dyserythropoietic anemia with thrombocytopenia[51] or X-linked macrothrombocytopenia.

PRIMARY EOSINOPHILIA

Eosinophilia refers to a PB absolute eosinophil count (AEC) that is above the upper normal limit of the reference range. The term “hypereosinophilia” is used when the AEC is above 1500×10^9/L. Eosinophilia is operationally classified into secondary (non-neoplastic proliferation of eosinophils) and primary (proliferation of eosinophils that is either neoplastic or otherwise unexplained). Secondary eosinophilia is by far the most frequent cause of eosinophilia and is often associated with infections, especially those related to tissue-invasive helminths, allergic/vasculitic diseases, drugs, and metastatic cancer. Primary eosinophilia is the focus of this chapter and is considered when a cause for secondary eosinophilia is not readily apparent.

Primary eosinophilia is classified as clonal or idiopathic. Diagnosis of clonal eosinophilia requires morphologic, cytogenetic, or molecular evidence of a myeloid neoplasm. Idiopathic eosinophilia is considered when both secondary and clonal eosinophilias have been ruled out as a possibility. HES is a subcategory of idiopathic eosinophilia with persistent AEC of $\geq 1.5 \times 10^9$/L and associated with eosinophil-mediated organ damage **(Table 106-6)**. An HES-like disorder that is associated with clonal or phenotypically abnormal T cells is referred to as “lymphocytic variant hypereosinophilia” (Table 106-6).

TABLE 106-6

Primary Eosinophilia Classification

Variables	Eosinophilia Associated with *PDGFRA, PDGFRB, FGFR1* or *PCM1-JAK2* Abnormality	Chronic Eosinophilia Not Otherwise Specified (CEL-NOS)	Lymphocytic Variant Hypereosinophilia	Hypereosinophilic Syndrome
Absolute eosinophil count	>600 × 10^9/L	>1500 × 10^9/L	>1500 × 10^9/L	>1500 × 10^9/L
Peripheral blood blast >2%	Yes or no	Yes or no	No	No
Bone marrow blast >5%	Yes or no	Yes or No	No	No
Abnormal karyotype	Yes or no	Yes or no	No	No
PDGFRA, PDGFRB, *FGFR1 or PCM1-JAK2* abnormality	Yes	No	No	No
BCR-ABL1	No	No	No	No
Abnormal T lymphocyte phenotype or clonal T cell clones	No	No	Yes	No
Eosinophil-mediated tissue damage	Yes or no	Yes or no	Yes or no	Yes

Clonal Eosinophilia

Examples of clonal eosinophilia include eosinophilia associated with AML, MDS, CML, mastocytosis, and MDS/MPN overlap. Myeloid neoplasm-associated eosinophilia also includes the WHO MPN subcategory of chronic eosinophilic leukemia, not otherwise specified (CEL-NOS) and the WHO myeloid malignancy subcategory referred to as "myeloid/lymphoid neoplasms with eosinophilia and rearrangement of platelet-derived growth factor receptor (*PDGFR*) α/β or fibroblast growth factor receptor 1 (*FGFR1*)" or with *PCM1-JAK2* (Table 106-4).

The diagnostic workup for clonal eosinophilia that is not associated with morphologically overt myeloid malignancy should start with PB mutation screening for *FIP1L1-PDGFRA* and *PDGFRB* mutations using fluorescence in situ hybridization (FISH) or reverse transcription-polymerase chain reaction. This is crucial since such eosinophilia is easily treated with imatinib. If mutation screening is negative, a BM examination with cytogenetic studies is indicated. In this regard, one must first pay attention to the presence or absence of 5q33, 4q12, 8p11.2, or t(8;9)(p22;p24.1) translocations, which, if present, would suggest *PDGFRB*, *PDGFRA*, or *FGFR1*-rearranged or *PCM1-JAK2*-associated clonal eosinophilia, respectively. The presence of 5q33 or 4q12 translocations predicts favorable response to treatment with imatinib mesylate and that of t(8;9)(p22;p24.1), a transient response to ruxolutinib while 8p11.2 translocations are associated with aggressive myeloid malignancies that are refractory to current drug therapy.

Chronic Eosinophilic Leukemia, Not Otherwise Specified (CEL-NOS)

CEL-NOS is a subset of clonal eosinophilia that is neither molecularly defined nor classified as an alternative clinicopathologically assigned myeloid malignancy. We prefer to use the term strictly in patients with an "HES" phenotype who also display either a clonal cytogenetic/molecular abnormality or excess blasts in the BM or PB. The WHO defines CEL-NOS in the presence of ≥1.5×10^9/L AEC that is accompanied by either the presence of myeloblast excess (either >2% in the PB or 5–19% in the BM) or evidence of myeloid clonality. Cytogenetic abnormalities in CEL, other than those that are associated with molecularly defined eosinophilic disorders, include trisomy 8 (the most frequent), t(10;11)(p14;q21), and t(7;12)(q11;p11). CEL-NOS does not respond to imatinib, and treatment strategies are often not different from those utilized in other similar MPNs; ASCT for transplant-eligible patients with poor risk factors and participation in experimental treatment protocols otherwise.

PDGFR Mutated Eosinophilia

Both platelet-derived growth factor receptors α (*PDGFRA* located on chromosome 4q12) and β (*PDGFRB* located on chromosome 5q31-q32) are involved in MPN-relevant activating mutations. Clinical phenotype in both instances includes prominent blood eosinophilia and excellent response to imatinib therapy. In regards to *PDGFRA* mutations, the most popular is *FIP1L1-PDGFRA*, a karyotypically occult del(4)(q12), that was described in 2003 as an imatinib-sensitive activating mutation. Functional studies have demonstrated transforming properties in cell lines and the induction of MPN in mice. Cloning of the *FIP1L1-PDGFRA* fusion gene identified a novel molecular mechanism for generating this constitutively active fusion tyrosine kinase, wherein a ~800kb interstitial deletion within 4q12 fuses the 5' portion of *FIP1L1* to the 3' portion of

PDGFRA.[57] *FIP1L1-PDGFRA* occurs in a very small subset of patients who present with the phenotypic features of either SM or HES, but the presence of the mutation reliably predicts complete hematologic and molecular response to imatinib therapy.

The association between eosinophilic myeloid malignancies and *PDGFRB* rearrangement was first characterized and published in 1994 where fusion of the tyrosine kinase encoding region of *PDGFRB* to the *ets*- like gene, *ETV6* (*ETV6-PDGFRB*, t(5;12)(q33;p13) was demonstrated. The fusion protein was transforming to cell lines and resulted in constitutive activation of PDGFRB signaling. Since then, several other *PDGFRB* fusion transcripts with similar disease phenotypes have been described, cell line transformation and MPD-induction in mice has been demonstrated, and imatinib therapy was effective when employed.

FGFR1 Mutated Eosinophilia

The 8p11 myeloproliferative syndrome (EMS) (also known as human stem cell leukemic/lymphoma syndrome) constitutes a clinical phenotype with features of both lymphoma and eosinophilic MPN and characterized by a fusion mutation that involves the gene for fibroblast growth factor receptor-1 (*FGFR1*), which is located on chromosome 8p11. In EMS, both myeloid and lymphoid lineage cells exhibit the 8p11 translocation, thus demonstrating the stem cell origin of the disease. The disease features several 8p11-linked chromosome translocations and some of the corresponding fusion *FGFR1* mutants have been shown to transform cell lines and induce EMS- or CML-like disease in mice depending on the specific *FGFR1* partner gene, *ZNF 198* or *BCR*, respectively. Consistent with this laboratory observation, some patients with *BCR-FGFR1* mutation manifest a more indolent CML-like disease. The mechanism of FGFR1 activation in EMS is similar to that seen with PDGFRB-associated MPD; the tyrosine kinase domain of FGFR1 is juxtaposed to a dimerization domain from the partner gene. EMS is aggressive and requires combination chemotherapy followed by ASCT.

PCM1-JAK2 Associated Myeloid/Lymphoid Neoplasm with Eosinophilia

The 2016 revised WHO document includes a provisional entity under myeloid/lymphoid neoplasms with eosinophilia referred to as “myeloid/lymphoid neoplasms with *PCM1-JAK2.*” The entity is characterized by the t(8;9)(p22;p24.1) cytogenetic abnormality and a phenotype that displays marked male predominance, organomegaly, eosinophilia, and hetereogenous morphologic features similar to MPN, MDS, or MDS/MPN. Current drug therapy for *PCM1-JAK2*-associated disease is suboptimal, although some affected patients have displayed transient responses to ruxolutinib therapy.

Hypereosinophilic Syndrome

Blood eosinophilia that is neither secondary nor clonal is operationally labeled as being “idiopathic.” HES is a sub-category of idiopathic eosinophilia with persistent increase of the AEC to ≥1.5×10^9/L and presence of eosinophil-mediated organ damage, including cardiomyopathy, gastroenteritis, cutaneous lesions, sinusitis, pneumonitis, neuritis, and vasculitis. In addition, some patients manifest thromboembolic complications, hepatosplenomegaly, and either cytopenia or cytosis.

BM histological and cytogenetic/molecular studies should be examined before a working diagnosis of HES is made. Additional blood studies that are currently recommended during the evaluation of "HES" include serum tryptase (an increased level suggests mastocytosis and warrants molecular studies to detect *FIP1L1-PDGFRA*), T-cell immunophenotyping, as well as T-cell receptor antigen gene rearrangement analysis (a positive test suggests an underlying clonal or phenotypically abnormal T-cell disorder). In addition, initial evaluation in HES should include echocardiogram and measurement of serum troponin levels to screen for myocardial involvement by the disease.

Initial evaluation of the patient with eosinophilia should include tests that facilitate assessment of target organ damage: complete blood count, chest x-ray, echocardiogram, and serum troponin level. Increased level of serum cardiac troponin has been shown to correlate with the presence of cardiomyopathy in HES. Typical echocardiographic findings in HES include ventricular apical thrombus, posterior mitral leaflet or tricuspid valve abnormality, endocardial thickening, dilated left ventricle, and pericardial effusion.

In a Mayo Clinic study of 98 consecutive patients with idiopathic eosinophilia, including HES, median age was 53 years (55% males), and overt organ involvement was seen >80% of the cases including 54% involving organs other than the skin. The frequencies of cardiac involvement, hepatosplenomegaly, increased serum tryptase and interleukin-5 levels were 8, 4, 24, and 31%, respectively. The study also revealed that 11% of the affected patients harbored pathogenetic mutations including *TET2*, *ASXL1*, and *KIT*; the presence of such mutations did not appear to influence phenotype and the number of informative cases was too small to assess prognostic relevance. Instead, the study identified anemia and presence of cardiac involvement or hepatosplenomegaly as risk factors for survival.

Glucocorticoids are the cornerstone of therapy in HES. Treatment with oral prednisone is usually started at 1mg/kg/d and continued for 1 to 2 weeks before the dose is tapered slowly over the ensuing 2 to 3 months. If symptoms recur at a prednisone dose level of >10 mg/d, either hydroxyurea or interferon α is used as steroid-sparing agent. In patients who do not respond to usual therapy as outlined above, mepolizumab or alemtuzumab might be considered. Mepolizumab targets IL-5, a well-recognized survival factor for eosinophils. Alemtuzumab targets the CD52 antigen, which has been shown to be expressed by eosinophils but not by neutrophils.

MASTOCYTOSIS

Mast cell disease (MCD) is defined as tissue infiltration by morphologically as well as immunophenotypically abnormal mast cells. MCD is classified into two broad categories; cutaneous and systemic mastocytosis (SM). MCD in adults is usually systemic and the clinical course can be either indolent or aggressive, depending on the respective absence or presence of impaired organ function. Symptoms and signs of MCD include urticaria pigmentosa, mast cell mediator release symptoms (e.g., headache, flushing, lightheadedness, syncope, anaphylaxis, pruritus, urticaria, angioedema, nausea, diarrhea, abdominal cramps), and organ damage (lytic bone lesions, osteoporosis, hepatosplenomegaly, cytopenia). Aggressive SM can be associated with another myeloid malignancy, including MPN, MDS, MDS/MPN overlap (e.g., CMML), or present as overt mast cell

leukemia (MCL). In general, life expectancy is near normal in indolent SM but significantly shortened in aggressive SM.

Diagnosis of SM is based on BM examination that shows clusters of morphologically abnormal, spindle-shaped mast cells that are best evaluated by the use of immunohistochemical stains that are specific to mast cells (tryptase, CD117). In addition, mast cell immunophenotyping reveals aberrant CD25 expression by neoplastic mast cells. Other laboratory findings in SM include increased levels of serum tryptase, histamine and urine histamine metabolites and prostaglandins. SM is associated with *KIT* mutations, usually *KIT*D816V, in the majority of patients. Accordingly, mutation screening for *KIT*D816V is diagnostically useful. However, the ability to detect *KIT*D816V depends on assay sensitivity and mast cell content of the test sample. The 2016 WHO classification of mastocytosis includes (1) cutaneous mastocytosis (CM), (2) SM, and (3) mast cell sarcoma (MCS). SM is further classified into (a) indolent SM (ISM), (b) smoldering SM (SSM), (c) SM with an associated hematological neoplasm (SM-AHN), (d) aggressive SM (ASM), and (e) MCL.

Both indolent and aggressive SM patients might experience mast cell mediator release symptoms, which are usually managed by both H-1 and H-2 histamine receptor blockers as well as cromolyn sodium. In addition, patients with propensity to vasodilatory shock should wear a medical alert bracelet as well as carry an Epi-Pen self-injector for self-administration of subcutaneous epinephrine. Urticaria pigmentosa shows variable response to both topical and systemic corticosteroid therapy. Cytoreductive therapy is not recommended for indolent SM. In aggressive SM, either interferon α or cladribine is considered first-line therapy and benefits the majority of patients. In contrast, imatinib is ineffective in the treatment of *PDGFR*-unmutated SM. A controlled study of patients with ISM or SSM demonstrated marginal value of masitinib (oral tyrosine kinase inhibitor that inhibits KIT and LYN) with reported cumulative symptomatic response rate of 18.7% vs 7.4% for placebo. Treatment responses were more impressive in another study that used the multikinase inhibitor midostaurin in patients with the more aggressive forms of SM, with 45% of the patients achieving major response.

DENDRITIC AND HISTIOCYTIC NEOPLASMS

DC and histiocyte/macrophage neoplasms are extremely rare. DCs are antigen-presenting cells, whereas histiocyte/macrophages are antigen-processing. BM myeloid stem cells (CD34+) give rise to monocyte (CD14+, CD68+, CD11c+, CD1a–) and DC (CD14-, CD11c+/–, CD1a+/c) precursors. Monocyte precursors, in turn, give rise to macrophages (CD14+, CD68+, CD11c+, CD163+, lysozyme+) and interstitial DCs (CD68+, CD1a–). DC precursors give rise to Langerhans cell DCs (Birbeck granules, CD1a+, S100+, langerin+) and plasmacytoid DCs (CD68+, CD123+). Follicular DCs (CD21+, CD23+, CD35+) originate from mesenchymal stem cells. Dendritic and histiocytic neoplasms are operationally classified into macrophage/histiocyte-related and DC-related. The former includes histiocytic sarcoma/malignant histiocytosis (MH) and the latter Langerhans cell (LC) histiocytosis, LC sarcoma, interdigitating DC sarcoma, and follicular DC sarcoma.

Histiocytic Sarcoma/Malignant Histiocytosis

Histiocytic sarcoma represents malignant proliferation of mature tissue histiocytes and is often localized. Median age at diagnosis is estimated at 46 years with slight male predilection. Some patients might have history of lymphoma, MDS, or germ cell tumors at time of disease presentation. The three typical disease sites are lymph nodes, skin, and the gastrointestinal system. Patients may or may not have systemic symptoms including fever and weight loss, and other symptoms include hepatosplenomegaly, lytic bone lesions, and pancytopenia. Immunophenotype includes presence of histiocytic markers (CD68, lysozyme, CD11c, CD14) and absence of myeloid or lymphoid markers. Prognosis is poor and treatment often ineffective. The term MH refers to a disseminated disease and systemic symptoms. Lymphoma-like treatment induces complete remissions in some patients and median survival is estimated at 2 years.

Langerhans Cell Histiocytosis

LCs are specialized DCs that reside in mucocutaneous tissue and upon activation become specialized for antigen presentation to T cells. LC histiocytosis (LCH; also known as histiocytosis X) represents neoplastic proliferation of LCs (S-100+, CD1a+, and Birbeck granules on electron microscopy). LCH incidence is estimated at 5 per million and the disease typically affects children with a male predilection. Presentation can be either uni- (eosinophilic granuloma) or multi-focal. The former usually affects bones and less frequently lymph nodes, skin, and lung, while the latter is more disseminated. Unifocal disease often affects older children and adults while multisystem disease affects infants. LCH of the lung in adults is characterized by bilateral nodules. Prognosis depends on organs involved. Only 10% of patients progress from unifocal to multiorgan disease. LCH of the lung might improve upon cessation of smoking. Approximately 55% of patients with LCH harbor *BRAF*V600E gain-of-function mutations, which indicates high-risk disease and resistance to first-line therapy, while responses to targeted therapy with vemurafenib have been reported.

Langerhans Cell Sarcoma

Langerhans cell sarcoma (LCS) also represents neoplastic proliferation of LCs with overtly malignant morphology. The disease can present de novo or progress from antecedent LCH. There is a female predilection and median age at diagnosis is estimated at 41 years. Immunophenotype is similar to that seen in LCH and liver, spleen, lung, and bone are the usual sites of disease. Prognosis is poor and treatment generally ineffective.

Interdigitating Dendritic Cell Sarcoma

Interdigitating DC sarcoma (IDCS), also known as reticulum cell sarcoma, represents neoplastic proliferation of IDCs. The disease is extremely rare and affects elderly adults with no sex predilection. Typical presentation is asymptomatic solitary lymphadenopathy. Immunophenotype includes S-100+ and negative for vimentin and CD1a. Prognosis ranges from benign local disease to widespread lethal disease.

Follicular Dendritic Cell Sarcoma

FDC reside in B-cell follicles and present antigen to B cells. FDC neoplasms (FDCN) are usually localized and often affect adults. FDCN might be associated with Castleman's disease in 10–20% of cases and increased incidence in schizophrenia has been reported. Cervical lymph nodes are the most frequent site of involvement in FDCN and other sites include maxillary, mediastinal, and retroperitoneal lymph nodes, oral cavity, the gastrointestinal system, skin, and breast. Sites of metastasis include lung and liver. Immunophenotype includes CD21, CD35, and CD23. Clinical course is typically indolent, and treatment includes surgical excision followed by regional radiotherapy and sometimes systemic chemotherapy.

Hemophagocytic Syndromes

HPS represents non-neoplastic proliferation and activation of macrophages that induces cytokine-mediated BM suppression and features of intense phagocytosis in BM and liver. HPS may result from genetic or acquired disorders of macrophages. The former entail genetically determined inability to regulate macrophage proliferation and activation. Acquired HPS is often precipitated by viral infections, most notably Epstein-Barr virus. HPS might also accompany certain malignancies such as T-cell lymphoma. It is characterized by pancytopenia and elevated ferritin levels. Interferon γ is thought to play a role. Clinical course is often fulminant and fatal.

FURTHER READING

Arber DA et al: The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 127:2391, 2016.

Haroche J et al: Dramatic efficacy of vemurafenib in both multisystemic and refractory Erdheim-Chester disease and Langerhans cell histiocytosis harboring the BRAF V600E mutation. Blood 121:1495, 2013.

Lortholary O et al: Masitinib for treatment of severely symptomatic indolent systemic mastocytosis: a randomised, placebo-controlled, phase 3 study. Lancet 389:612, 2017.

Maxson JE et al: Oncogenic CSF3R mutations in chronic neutrophilic leukemia and atypical CML. N Engl J Med 368:1781, 2013.

Pardanani A et al: Predictors of survival in WHO-defined hypereosinophilic syndrome and idiopathic hypereosinophilia and the role of next-generation sequencing. Leukemia 30:1924, 2016.

Patnaik MM et al: Mayo prognostic model for WHO-defined chronic myelomonocytic leukemia: ASXL1 and spliceosome component mutations and outcomes. Leukemia 27:1504, 2013.

Swerdlow SH et al: The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood 127(20):2375, 2016. [PubMed: 26980727]

Taylor J et al: Diagnosis and classification of hematologic malignancies on the basis of genetics. Blood 130:410, 2017.

Thompson PA, Ravandi F: How I manage hairy cell leukemia. Br J Haematol 177:543, 2017.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 107: Plasma Cell Disorders

Nikhil C. Munshi; Dan L. Longo; Kenneth C. Anderson

INTRODUCTION

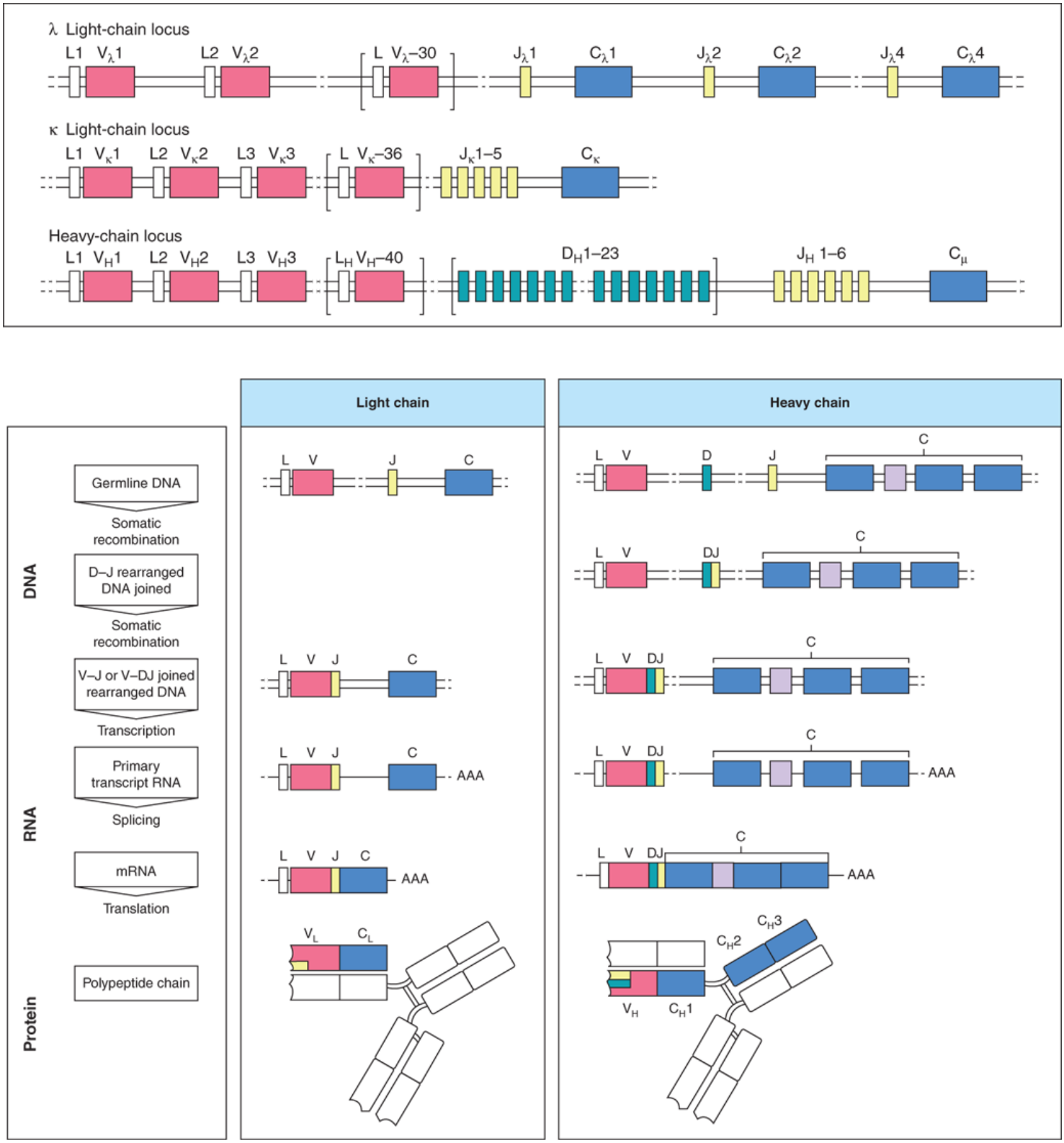
The *plasma cell disorders* are monoclonal neoplasms related to each other by virtue of their development from common progenitors in the late B-lymphocyte lineage. Multiple myeloma (MM), Waldenström's macroglobulinemia, primary amyloidosis **(Chap. 108)**, and the heavy chain diseases comprise this group and may be designated by a variety of synonyms such as *monoclonal gammopathies, paraproteinemias, plasma cell dyscrasias*, and *dysproteinemias*. Mature B lymphocytes destined to produce IgG bear surface immunoglobulin molecules of both μ and γ heavy chain isotypes with both isotypes having identical idiotypes (variable regions). Under normal circumstances, maturation to antibody-secreting plasma cells and their proliferation is stimulated by exposure to the antigen for which the surface immunoglobulin is specific; however, in the plasma cell disorders, the control over this process is lost. The clinical manifestations of all the plasma cell disorders relate to the expansion of the neoplastic cells, to the secretion of cell products (immunoglobulin molecules or subunits, lymphokines), and to some extent to the host's response to the tumor. **Normal development of B lymphocytes is discussed in Chap. 342 and depicted in Fig. 104-2.**

Three categories of structural variation are present among immunoglobulin molecules that form antigenic determinants, and these are used to classify immunoglobulins. *Isotypes* are those determinants that distinguish among the main classes of antibodies of a given species and are the same in all normal individuals of that species. Therefore, isotypic determinants are, by definition, recognized by antibodies from a distinct species (heterologous sera) but not by antibodies from the same species (homologous sera). There are five heavy chain isotypes (M, G, A, D, E) and two light chain isotypes (κ, λ). *Allotypes* are distinct determinants that reflect regular small differences between individuals of the same species in the amino acid sequences of otherwise similar immunoglobulins. These differences are determined by allelic genes; by definition, they are detected by antibodies made in the same species. *Idiotypes* are the third category of antigenic determinants. They are unique to the molecules produced by a given clone of antibody-producing cells. Idiotypes are formed by the unique structure of the antigen-binding portion of the molecule.

Antibody molecules **(Fig. 107-1)** are composed of two heavy chains (~50,000 mol wt) and two light chains (~25,000 mol wt). Each chain has a constant portion (limited amino acid sequence variability) and a variable region (extensive sequence variability). The light and heavy chains are linked by disulfide bonds and are aligned so that their variable regions are adjacent to one another. This variable region forms the antigen recognition site of the antibody molecule; its unique structural features form idiotypes that are reliable markers for a particular clone of cells because each antibody is formed and secreted by a single clone. Because of the mechanics of the gene rearrangements necessary to specify the immunoglobulin variable regions (VDJ joining for the heavy chain, VJ joining for the light chain), a particular clone rearranges only one of the two chromosomes to produce an immunoglobulin molecule of only one light chain isotype and only one allotype (allelic exclusion) (Fig. 107-1). After exposure to antigen, the variable region may become associated with a new heavy chain isotype (class switch). Each clone of cells performs these sequential gene arrangements in a unique way. This results in each clone producing a unique immunoglobulin molecule. In most plasma cells, light chains are synthesized in slight excess, secreted as free light chains, and cleared by the kidney, but <10 mg of such light chains is excreted per day.

FIGURE 107-1

Immunoglobulin genetics and the relationship of gene segments to the antibody protein. The *top* portion of the figure is a schematic of the organization of the immunoglobulin genes, λ on chromosome 22, κ on chromosome 2, and the heavy chain locus on chromosome 14. The heavy chain locus is >2 megabases, and some of the D region gene segments are only a few bases long, so the figure depicts the schematic relationship among the segments, not their actual size. The *bottom* portion of the figure outlines the steps in going from the noncontiguous germline gene segments to an intact antibody molecule. Two recombination events juxtapose the V-D-J (or V-J for light chains) segments. The rearranged gene is transcribed, and RNA splicing cuts out intervening sequences to produce an mRNA, which is then translated into an antibody light or heavy chain. The sites on the antibody that bind to antigen (the so-called CDR3 regions) are encoded by D and J segments for heavy chains and the J segments for light chains. *(From K Murphy: Janeway's Immunobiology, 8th ed. Garland Science, 2011.)*



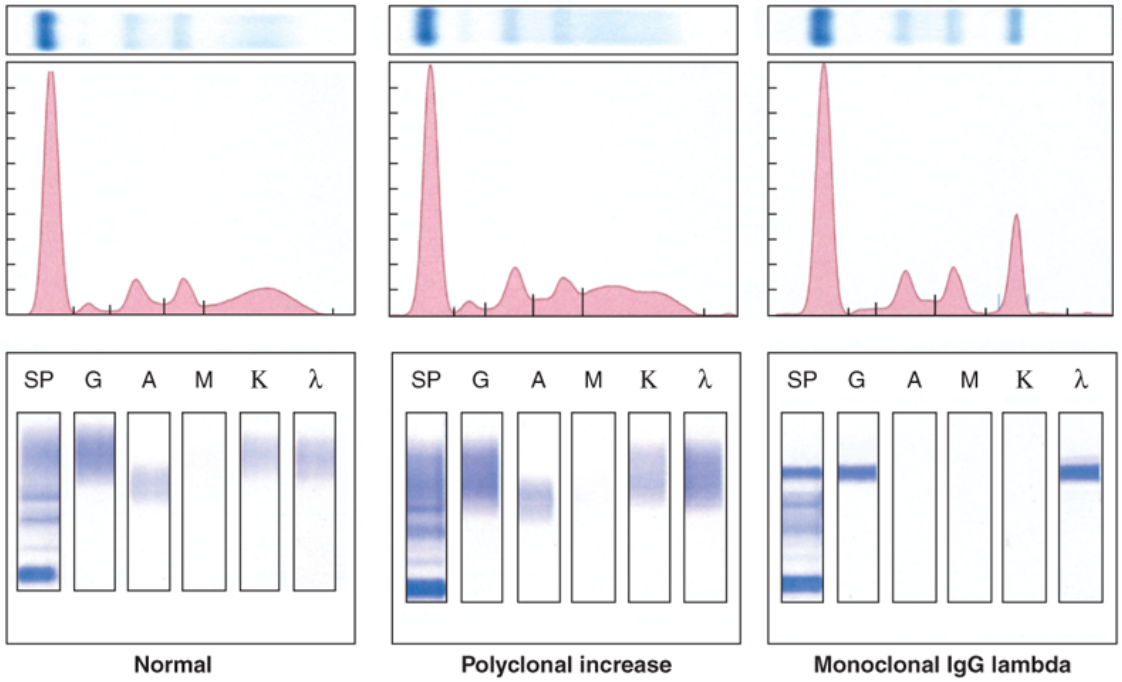
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Electrophoretic analysis permits separation of components of the serum proteins (Fig. 107-2). The immunoglobulins move heterogeneously in an electric field and form a broad peak in the gamma region, which is usually increased in the sera of patients with plasma cell tumors. There is a sharp spike in this region called an *M component* (M for monoclonal). Less commonly, the M component may appear in the β_2 or α_2 globulin region. The monoclonal antibody must be present at a concentration of at least 5 g/L (0.5 g/dL) to be accurately quantitated by this method. This corresponds to ~10^9 cells producing the antibody. Confirmation of the type of immunoglobulin and that it is truly monoclonal is determined by immunoelectrophoresis that reveals a single heavy and/or light chain type. Hence immunoelectrophoresis and electrophoresis provide qualitative and quantitative assessment of the M component, respectively. Once the presence of an M component has been confirmed, the amount of M component in the serum is a reliable measure of the tumor burden, making M component an excellent tumor marker to manage therapy, yet it is not specific enough to be used to screen asymptomatic

patients. In addition to the plasma cell disorders, M components may be detected in other lymphoid neoplasms such as chronic lymphocytic leukemia (CLL) and lymphomas of B- or T-cell origin; nonlymphoid neoplasms such as chronic myeloid leukemia, breast cancer, and colon cancer; a variety of nonneoplastic conditions such as cirrhosis, sarcoidosis, parasitic diseases, Gaucher's disease, and pyoderma gangrenosum; and a number of autoimmune conditions, including rheumatoid arthritis, myasthenia gravis, and cold agglutinin disease. Monoclonal proteins are also observed in immunosuppressed patients after organ transplant and, rarely, allogeneic transplant. At least two very rare skin diseases—lichen myxedematosus (also known as papular mucinosis) and necrobiotic xanthogranuloma—are associated with a monoclonal gammopathy. In papular mucinosis, highly cationic IgG is deposited in the dermis of patients. This organ specificity may reflect the specificity of the antibody for some antigenic component of the dermis. Necrobiotic xanthogranuloma is a histiocytic infiltration of the skin, usually of the face, that produces red or yellow nodules that can enlarge to plaques. Approximately 10% progress to myeloma. Five percent of patients with sensory motor neuropathy also have a monoclonal paraprotein.

FIGURE 107-2

Representative patterns of serum electrophoresis and immunofixation. The *upper panels* represent agarose gel, middle panels are the densitometric tracing of the gel, and *lower panels* are immunofixation patterns. Panel on the *left* illustrates the normal pattern of serum protein on electrophoresis. Because there are many different immunoglobulins in the serum, their differing mobilities in an electric field produce a broad peak. In conditions associated with increases in polyclonal immunoglobulin, the broad peak is more prominent (*middle panel*). In monoclonal gammopathies, the predominance of a product of a single cell produces a "church spire" sharp peak, usually in the γ globulin region (*right panel*). The immunofixation (*lower panel*) identifies the type of immunoglobulin. For example, normal and polyclonal increase in immunoglobulins produce no distinct bands; however, the *right panel* shows distinct bands in IgG and lambda protein lanes, confirming the presence of IgG lambda monoclonal protein. *(Courtesy of Dr. Neal I. Lindeman; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The nature of the M component is variable in plasma cell disorders. It may be an intact antibody molecule of any heavy chain subclass, or it may be an altered antibody or fragment. Isolated light or heavy chains may be produced. In some plasma cell tumors such as extramedullary or solitary bone plasmacytomas, less than one-third of patients will have an M component. In ~20% of myelomas, only light chains are produced and, in most cases, are secreted in the urine as Bence Jones proteins. The frequency of myelomas of a particular heavy chain class is roughly proportional to the serum concentration, and therefore, IgG myelomas are more common than IgA and IgD myelomas. In ~1% of patients with myeloma, biclonal or triclonal gammopathy is observed.

MULTIPLE MYELOMA

DEFINITION

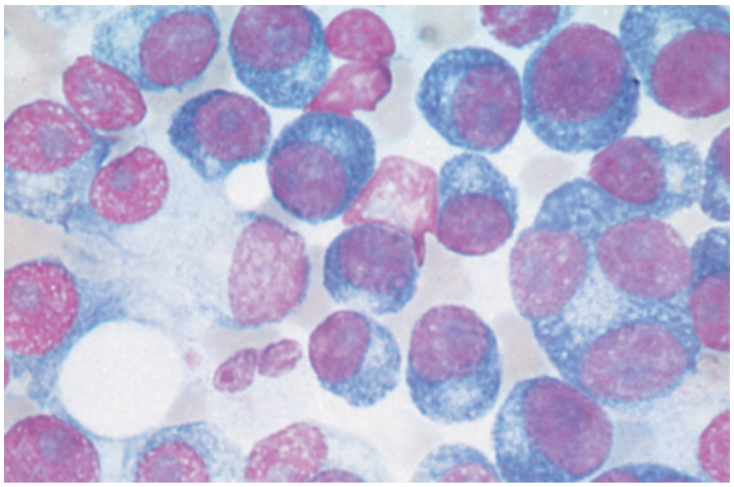
MM represents a malignant proliferation of plasma cells derived from a single clone. The tumor, its products, and the host response to it result in a number of organ dysfunctions and symptoms, including bone pain or fracture, renal failure, susceptibility to infection, anemia, hypercalcemia, and occasionally clotting abnormalities, neurologic symptoms, and manifestations of hyperviscosity.

ETIOLOGY

The cause of myeloma is not known. Myeloma occurred with increased frequency in those exposed to the radiation of nuclear warheads in World War II after a 20-year latency. Myeloma has been seen more commonly than expected among farmers, wood workers, leather workers, and those exposed to petroleum products. A variety of chromosomal alterations have been found in patients with myeloma: hyperdiploidy, 13q14 deletions, translocations t(11;14)(q13;q32), t(4;14)(p16;q32), and t(14;16), 1q amplification or 1p deletion, and 17p13 deletions. Evidence is strong that errors in switch recombination—the genetic mechanism to change antibody heavy chain isotype—participate in the early transformation process. However, no single common molecular pathogenetic pathway has yet emerged. Genome sequencing studies have failed to identify any recurrent mutation with frequency >20%; *N-ras, K-ras*, and *B-raf* mutations are most common and combined occur in >40% of patients. Evidence of complex clusters of subclonal variants is present at diagnosis, acquires additional mutations over time, indicative of genomic evolution that may drive disease progression. Interleukin (IL) 6 may play a role in driving myeloma cell proliferation. It remains difficult to distinguish benign from malignant plasma cells based on morphologic criteria in all but a few cases (**Fig. 107-3**).

FIGURE 107-3

Multiple myeloma (marrow). The cells bear characteristic morphologic features of plasma cells, round or oval cells with an eccentric nucleus composed of coarsely clumped chromatin, a densely basophilic cytoplasm, and a perinuclear clear zone containing the Golgi apparatus. Binucleate and multinucleate malignant plasma cells can be seen.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

INCIDENCE AND PREVALENCE

An estimated 30,280 new cases of myeloma were diagnosed in 2017, and 12,590 people died from the disease in the United States. Myeloma increases in incidence with age. The median age at diagnosis is 69 years; it is uncommon under age 40. Males are more commonly affected than females, and blacks have nearly twice the incidence of whites. Myeloma accounts for 1.3% of all malignancies in whites and 2% in blacks, and 13% of all hematologic cancers in whites and 33% in blacks.

GLOBAL CONSIDERATIONS

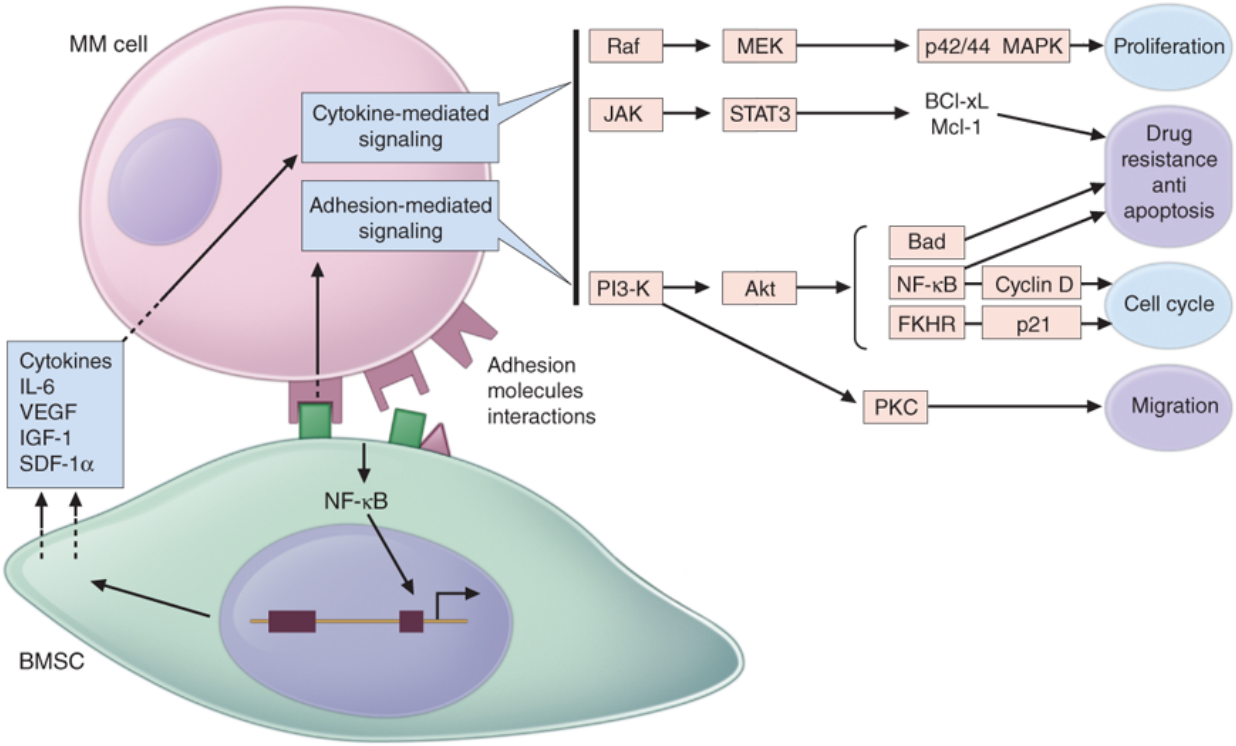
The incidence of myeloma is highest in African Americans and Pacific Islanders; intermediate in Europeans and North American whites; and lowest in people from developing countries including Asia. The higher incidence in more developed countries may result from the combination of a longer life expectancy and more frequent medical surveillance. Incidence of MM in other ethnic groups including native Hawaiians, female Hispanics, American Indians from New Mexico, and Alaskan natives is higher relative to U.S. whites in the same geographic area. Chinese and Japanese populations have a lower incidence than whites. Immunoproliferative small-intestinal disease (IPSID) with alpha heavy chain disease is most prevalent in the Mediterranean area. Despite these differences in prevalence, the characteristics, response to therapy, and prognosis of myeloma are similar worldwide.

PATHOGENESIS AND CLINICAL MANIFESTATIONS

MM cells bind via cell-surface adhesion molecules to bone marrow stromal cells (BMSCs) and extracellular matrix (ECM), which triggers MM cell growth, survival, drug resistance, and migration in the bone marrow milieu (**Fig. 107-4**). These effects are due both to direct MM cell–BMSC binding via adhesion molecules and to induction of various cytokines, including IL-6, insulin-like growth factor type I (IGF-I), vascular endothelial growth factor (VEGF), and stromal cell–derived growth factor (SDF)-1α. Growth, drug resistance, and migration are mediated via Ras/Raf/mitogen-activated protein kinase, PI3K/Akt, and protein kinase C signaling cascades, respectively.

FIGURE 107-4

Pathogenesis of multiple myeloma. Multiple myeloma (MM) cells interact with bone marrow stromal cells (BMSCs) and extracellular matrix proteins via adhesion molecules, triggering adhesion-mediated signaling as well as cytokine production. This triggers cytokine-mediated signaling that provides growth, survival, and antiapoptotic effects as well as development of drug resistance.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Bone pain is the most common symptom in myeloma, affecting nearly 70% of patients. Persistent localized pain usually signifies a pathologic fracture. The bone lesions of myeloma are caused by the proliferation of tumor cells, activation of osteoclasts that destroy bone, and suppression of osteoblasts that form new bone. The increased osteoclast activity is mediated by osteoclast activating factors (OAFs) produced by the myeloma cells (mediated by several cytokines, including IL-1, lymphotoxin, VEGF, receptor activator of NF-κB [RANK] ligand, macrophage inhibitory factor [MIP]-1α, and tumor necrosis factor [TNF]). The bone lesions are lytic in nature **(Fig. 107-5)** and are rarely associated with osteoblastic new bone formation due to their suppression by dickhoff-1 (DKK-1) produced by myeloma cells. Therefore, radioisotopic bone scanning is less useful in diagnosis than is plain radiography. The bony lysis results in substantial mobilization of calcium from bone, and serious acute and chronic complications of hypercalcemia may dominate the clinical picture (see below). Localized bone lesions may cause the collapse of vertebrae leading to spinal cord compression. The next most common clinical problem in patients with myeloma is susceptibility to bacterial infections. The most common infections are pneumonias and pyelonephritis, and the most frequent pathogens are *Streptococcus pneumoniae, Staphylococcus aureus*, and *Klebsiella pneumoniae* in the lungs and *Escherichia coli* and other gram-negative organisms in the urinary tract. In ~25% of patients, recurrent infections are the presenting features, and >75% of patients will have a serious infection at some time in their course. The susceptibility to infection has several contributing causes. First, patients with myeloma have diffuse hypogammaglobulinemia if the M component is excluded. The hypogammaglobulinemia is related to both decreased production and increased destruction of normal antibodies. The large M component results in fractional catabolic rates of 8–16% instead of the normal 2%. Moreover, some patients generate a population of circulating regulatory cells in response to their myeloma that can suppress normal antibody synthesis. These patients have very poor antibody responses, especially to polysaccharide antigens such as those on bacterial cell walls. Various abnormalities in T-cell function are also observed including decreased Th_1 response, increase in Th17 cells producing proinflammatory cytokines, and aberrant T_{reg} cell function. Granulocyte lysozyme content is low, and granulocyte migration is not as rapid as normal in patients with myeloma, probably the result of a tumor product. There are also a variety of abnormalities in complement functions in myeloma patients. All these factors contribute to the immune deficiency in these patients. Some commonly used therapeutic agents, e.g., dexamethasone, suppress immune responses and increase susceptibility to bacterial and fungal infection, and bortezomib predisposes to herpesvirus reactivation.

FIGURE 107-5

Bony lesions in multiple myeloma (MM). ***A.*** The skull demonstrates the typical "punched out" lesions characteristic of MM. The lesion represents a purely osteolytic lesion with little or no osteoblastic activity (above). ***B.*** PET/CT showing multiple fluorodeoxyglucose (FDG)-avid lesions in skeleton (left panel) with their resolution on achieving complete response (CR) (right panel). *(Part A courtesy of Dr. Geraldine Schechter; with permission. Part B courtesy of Dr. Sundar Jagannath; with permission.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Renal failure occurs in nearly 25% of myeloma patients, and some renal pathology is noted in >50%. Of many contributing factors, hypercalcemia is the most common cause of renal failure. Glomerular deposits of amyloid, hyperuricemia, recurrent infections, frequent use of nonsteroidal anti-inflammatory agents for pain control, use of iodinated contrast dye for imaging, bisphosphonate use, and occasional infiltration of the kidney by myeloma cells all may contribute to renal dysfunction. However, tubular damage associated with the excretion of light chains is almost always present. Normally, light chains are filtered, reabsorbed in the tubules, and catabolized. With the increase in the amount of light chains presented to the tubule, the tubular cells become overloaded with these proteins, and tubular damage results either directly from light chain toxic effects or indirectly from the release of intracellular lysosomal enzymes. The earliest manifestation of this tubular damage is the adult Fanconi's syndrome (a type 2 proximal renal tubular acidosis), with loss of glucose and amino acids, as well as defects in the ability of the kidney to acidify and concentrate the urine. The proteinuria is not accompanied by hypertension, and the protein is nearly all light chains. Generally, very little albumin is in the urine because glomerular function is usually normal. When the glomeruli are involved, nonselective proteinuria is also observed. Patients with myeloma also have a decreased anion gap [i.e., $Na^+ - (Cl^- + HCO_3^-)$]

because the M component is cationic, resulting in retention of chloride. This is often accompanied by hyponatremia that is felt to be artificial (pseudohyponatremia) because each volume of serum has less water as a result of the increased protein. Renal dysfunction due to light chain deposition disease, light chain cast nephropathy, and amyloidosis is partially reversible with effective therapy. Myeloma patients are susceptible to developing acute renal failure if they become dehydrated.

Normocytic and normochromic anemia occurs in ~80% of myeloma patients. It is usually related to the replacement of normal marrow by expanding tumor cells, to the inhibition of hematopoiesis by factors made by the tumor, to reduced production of erythropoietin by the kidney, and to the effects of long-term therapy. In addition, mild hemolysis may contribute to the anemia. A larger than expected fraction of patients may have megaloblastic anemia due to either folate or vitamin B_{12} deficiency. Granulocytopenia and thrombocytopenia are rare except when therapy-induced. Clotting abnormalities may be seen due to the failure of antibody-coated platelets to function properly; the interaction of the M component with clotting factors I, II, V, VII, or VIII; antibody to clotting factors; or amyloid damage of endothelium. Deep venous thrombosis is also observed with use of thalidomide, lenalidomide, or pomalidomide in combination with dexamethasone. Raynaud's phenomenon and impaired circulation may result if the M component forms cryoglobulins, and hyperviscosity syndromes may develop depending on the physical properties of the M component (most common with IgM, IgG3, and IgA paraproteins). Hyperviscosity is defined based on the relative viscosity of serum as compared with water. Normal relative serum viscosity is 1.8 (i.e., serum is normally almost twice as viscous as water). Symptoms of hyperviscosity occur at a level greater than 4 centipoise (cP), which is usually reached at paraprotein concentrations of ~40 g/L (4 g/dL) for IgM, 50 g/L (5 g/dL) for IgG3, and 70 g/L (7 g/dL) for IgA; however, depending on chemical and physical properties of the paraprotein molecule, it can occasionally be observed at lower levels.

Although neurologic symptoms occur in a minority of patients, they may have many causes. Hypercalcemia may produce lethargy, weakness, depression, and confusion. Hyperviscosity may lead to headache, fatigue, shortness of breath, exacerbation or precipitation of heart failure, visual disturbances, ataxia, vertigo, retinopathy, somnolence, and coma. Bony damage and collapse may lead to cord compression, radicular pain, and loss of bowel and bladder control. Infiltration of peripheral nerves by amyloid can be a cause of carpal tunnel syndrome and other sensorimotor mono- and polyneuropathies. Neuropathy associated with monoclonal gammopathy of undetermined significance (MGUS) and myeloma is more frequently sensory than motor neuropathy and is associated with IgM more than other isotypes. In >50% of patients with neuropathy, the IgM monoclonal protein is directed against myelin-associated globulin (MAG). Sensory neuropathy is also a side effect of therapy, specifically thalidomide and bortezomib.

Many of the clinical features of myeloma, e.g., cord compression, pathologic fractures, hyperviscosity, sepsis, and hypercalcemia, can present as medical emergencies. Despite the widespread distribution of plasma cells in the body, tumor expansion is dominantly within bone and bone marrow and, for reasons unknown, rarely causes enlargement of spleen, lymph nodes, or gut-associated lymphatic tissue.

DIAGNOSIS AND STAGING

The diagnosis of myeloma requires marrow plasmacytosis (>10%), a serum and/or urine M component, and at least one of the myeloma defining events detailed in **Table 107-1**. Bone marrow plasma cells are CD138+ and either monoclonal kappa or lambda light chain positive. The most important differential diagnosis in patients with myeloma involves their separation from individuals with MGUS or smoldering multiple myeloma (SMM). MGUS is vastly more common than myeloma, occurring in 1% of the population aged >50 years and in up to 10% of individuals aged >75 years. The diagnostic criteria for MGUS, SMM, and myeloma are described in Table 107-1. Although ~1% of patients per year with MGUS go on to develop myeloma, all cases of myeloma are preceded by MGUS. Non-IgG subtype, abnormal kappa/lambda free light chain ratio, and serum M protein >15 g/L (1.5 g/dL) are associated with higher incidence of progression of MGUS to myeloma. Absence of all three features predicts a 5% chance of progression, whereas higher risk MGUS with the presence of all three features predicts a 60% chance of progression >20 years. The features responsible for higher risk of progression from SMM to MM are bone marrow plasmacytosis >10%, abnormal kappa/lambda free light chain ratio, and serum M protein >30 g/L (3 g/dL). Patients with only one of these three features have a 25% chance of progression to MM in 5 years, whereas patients with high-risk SMM with all three features have a 76% chance of progression. There are two important variants of myeloma—solitary bone plasmacytoma and solitary extramedullary plasmacytoma. These lesions are associated with an M component in <30% of the cases, they may affect younger individuals, and both are associated with median survivals of ≥10 years. Solitary bone plasmacytoma is a single lytic bone lesion without marrow plasmacytosis. Extramedullary plasmacytomas usually involve the submucosal lymphoid tissue of the nasopharynx or paranasal sinuses without marrow plasmacytosis. Both tumors are highly responsive to local radiation therapy. If an M component is present, it should disappear after treatment. Solitary bone plasmacytomas may recur in other bony sites or evolve into myeloma. Extramedullary plasmacytomas rarely recur or progress.

TABLE 107-1

Diagnostic Criteria for Multiple Myeloma, Myeloma Variants, and Monoclonal Gammopathy of Undetermined Significance

Monoclonal Gammopathy of Undetermined Significance (MGUS)
Serum monoclonal protein (non-IgM type) <30 g/L Clonal bone marrow plasma cells <10%* Absence of myeloma defining events or amyloidosis that can be attributed to the plasma cell proliferative disorder
Smoldering Multiple Myeloma (Asymptomatic Myeloma)
Both criteria must be met: • Serum monoclonal protein (IgG or IgA) ≥30 g/L or urinary monoclonal protein ≥500 mg per 24 h and/or clonal bone marrow plasma cells 10–60% • Absence of myeloma defining events or amyloidosis
Symptomatic Multiple Myeloma
Clonal bone marrow plasma cells or biopsy-proven bony or extramedullary plasmacytoma[a] and any one or more of the following myeloma defining events: • Evidence of one or more end-organ damage that can be attributed to the underlying plasma cell proliferative disorder, specifically: ◦ Hypercalcemia: serum calcium >0.25 mmol/L (>1 mg/dL) higher than the upper limit of normal or >2.75 mmol/L (>11 mg/dL) ◦ Renal insufficiency: creatinine clearance <40 mL per min[b] or serum creatinine >177 µmol/L (>2 mg/dL) ◦ Anemia: hemoglobin value of >20 g/L below the lower limit of normal, or a hemoglobin value <100 g/L ◦ Bone lesions: one or more osteolytic lesions on skeletal radiography, CT, or PET-CT[c] ◦ Any one or more of the following biomarkers of malignancy: ▪ Clonal bone marrow plasma cell percentage[a] ≥60% ▪ Involved: uninvolved serum free light chain ratio[d] ≥100 ▪ >1 focal lesions on MRI studies[e]
Nonsecretory Myeloma
No M protein in serum and/or urine with immunofixation[f] Bone marrow clonal plasmacytosis ≥10% or plasmacytoma[a] Myeloma-related organ or tissue impairment (end-organ damage, as described above)
Solitary Plasmacytoma
Biopsy-proven solitary lesion of bone or soft tissue with evidence of clonal plasma cells Normal bone marrow with no evidence of clonal plasma cells[a] Normal skeletal survey and MRI (or CT) of spine and pelvis (except for the primary solitary lesion) Absence of end-organ damage such as hypercalcemia, renal insufficiency, anemia, or bone lesions (CRAB) that can be attributed to a lymphoplasma cell proliferative disorder
POEMS Syndrome
All of the following four criteria must be met: 1. Polyneuropathy 2. Monoclonal plasma cell proliferative disorder 3. Any one of the following: (a) sclerotic bone lesions; (b) Castleman's disease; (c) elevated levels of vascular endothelial growth factor (VEGF) 4. Any one of the following: (a) organomegaly (splenomegaly, hepatomegaly, or lymphadenopathy); (b) extravascular volume overload (edema, pleural effusion, or ascites); (c) endocrinopathy (adrenal, thyroid, pituitary, gonadal, parathyroid, and pancreatic); (d) skin changes (hyperpigmentation, hypertrichosis, glomeruloid hemangiomata, plethora, acrocyanosis, flushing, and white nails); (e) papilledema; (f) thrombocytosis/polycythemia[g]

PET-CT=^{18}F-fluorodeoxyglucose PET with CT. [a]Clonality should be established by showing κ/λ-light-chain restriction on flow cytometry, immunohistochemistry, or immunofluorescence. Bone marrow plasma cell percentage should preferably be estimated from a core biopsy specimen; in case of a disparity between the aspirate and core biopsy, the highest value should be used. [b]Measured or estimated by validated equations. [C]If bone marrow has less than 10% clonal plasma cells, more than one bone lesion is required to distinguish from solitary plasmacytoma with minimal marrow involvement. [d]These values are based on the serum Freelite assay (The Binding Site Group, Birmingham, UK). The involved free light chain must be ≥100 mg/L. [e]Each focal lesion must be 5 mm or more in size. [f]A small M component may sometimes be present. [g]These features should have no attributable other causes and have temporal relation with each other.

Abbreviation: POEMS, polyneuropathy, organomegaly, endocrinopathy, M-protein, and skin changes.

Serum protein electrophoresis and measurement of serum immunoglobulins and free light chains are useful for detecting and characterizing M spikes, supplemented by immunoelectrophoresis, which is especially sensitive for identifying low concentrations of M components not detectable by protein electrophoresis. A 24-h urine specimen is necessary to quantitate Bence Jones protein excretion. Serum alkaline phosphatase is usually normal even with extensive bone involvement because of the absence of osteoblastic activity. It is also important to quantitate serum β_2-microglobulin and albumin (see below).

The serum M component will be IgG in 53% of patients, IgA in 25%, and IgD in 1%; 20% of patients will have only light chains in serum and urine. Dipsticks for detecting proteinuria are not reliable at identifying light chains, and the heat test for detecting Bence Jones protein is falsely negative in ~50% of patients with light chain myeloma. Fewer than 1% of patients have no identifiable M component; these patients usually have light chain myeloma in which renal catabolism has made the light chains undetectable in the urine. In most of these patients, light chains can now be detected by serum free light chain assay. IgD myeloma may also present with light chain disease. About two-thirds of patients with serum M components also have urinary light chains. The light chain isotype may have an impact on disease behavior. Whether this is due to some genetically important determinant of cell proliferation or because lambda light chains are more likely to cause renal damage and form amyloid than are kappa light chains is unclear. The heavy chain isotype may have an impact on patient management as well. About half of patients with IgM paraproteins develop hyperviscosity compared with only 2–4% of patients with IgA and IgG M components. Among IgG myelomas, it is the IgG3 subclass that has the highest tendency to form both concentration- and temperature-dependent aggregates, leading to hyperviscosity and cold agglutination at lower serum concentrations. A standard workup directed at detecting monoclonal plasma cells and myeloma-defining events as well as prognosis is detailed in **Table 107-2**. A complete blood count with differential may reveal anemia. Erythrocyte sedimentation rate is elevated. Rare patients (~1%) may have plasma cell leukemia with >2000 plasma cells/μL. This may be seen in disproportionate frequency in IgD (12%) and IgE (25%) myelomas. Serum calcium, urea nitrogen, creatinine, and uric acid levels may be elevated.

TABLE 107-2

Standard Investigative Workup in MM

Investigations to Evaluate for Clonal Plasma Cells
Bone marrow aspirate and biopsy (fine needle aspiration of plasmacytoma if indicated) - Histology - Clonality by kappa/lambda immunostaining by flow cytometry or immunohistochemistry
Investigations to Evaluate Clonal Paraprotein
Serum protein electrophoresis and immunofixation Quantitative serum immunoglobulin levels (IgG, IgA, and IgM) 24-h Urine protein electrophoresis and immunofixation Serum free light chain and ratio Immunofixation for IgD or IgE in select cases
Investigation to Evaluate End-Organ Damage
Hemogram to assess for anemia Chemistry panel for renal function and calcium Skeletal survey to evaluate bone lesions - PET/CT or MRI if SMM or solitary plasmacytoma with no other MDE or extramedullary disease
Investigation for Risk Stratification
- β-2 microglobulin and serum albumin for ISS stage - Fluorescent in situ hybridization for hyperdipoidy, del17p, t(4;14); t(14;16), amp1q34, and del 13 on bone marrow sample - LDH
Specialized Investigation in Selected Cases
Abdominal fat pad for amyloid
Serum viscosity if IgM component or high IgA levels or serum M-component >7 g/dL

Abbreviations: ISS, International Staging System; LDH, lactate dehydrogenase; MRI, magnetic resonance imaging; PET/CT, positron emission tomography/computerized tomography.

The clinical evaluation of patients with myeloma includes a careful physical examination searching for tender bones and masses. Chest and bone radiographs may reveal lytic lesions or diffuse osteopenia. Magnetic resonance imaging (MRI) offers a sensitive means to document extent of bone marrow infiltration and cord or root compression in patients with pain syndromes. ^{18}F-fluorodeoxyglucose (^{18}F-FDG) PET/CT is a valuable tool to assess bone damage and detect extramedullary sites of the disease (Fig. 107-5). The use of ^{18}F-FDG PET/CT is recommended to distinguish between smoldering and active MM and to confirm a suspected diagnosis of solitary plasmacytoma. It is also a valuable tool to evaluate response in patients with oligo- or nonsecretory myeloma.

PROGNOSIS

Serum β_2-microglobulin is the single most powerful predictor of survival and can substitute for staging. β_2-Microglobulin is the light chain of the class I major histocompatibility antigens (HLA-A, -B, -C) on the surface of every cell. Combination of serum β_2-microglobulin and albumin levels forms the basis for a three-stage International Staging System (ISS) (**Table 107-3**) that predicts survival. With the use of high-dose therapy and the newer agents, the Durie-Salmon staging system is unable to predict outcome and thus is no longer used. High labeling index, circulating plasma cells, performance status, and high levels of lactate dehydrogenase are also associated with poor prognosis.

TABLE 107-3

Risk Stratification in Myeloma

Chromosomal Abnormalities		
Method	**Standard Risk (80%) (Expected Survival 6–7+ Years)**	**High Risk (20%) (Expected Survival 2–3 Years)**
Karyotype	No chromosomal aberration	Any abnormality on conventional karyotype
FISH	t(11;14)	Del(17p)
	t(6:14)	t(4:14)
	Del(13)	t(14:16)
		t(14;20)
		amp 1q34
International Staging System		
	Stage	**Median Survival, Months**
β_2M <3.5, alb ≥3.5	I (28%)[a]	62
β_2M <3.5, alb <3.5 *or* β_2M = 3.5–5.5	II (39%)	44
β_2M >5.5	III (33%)	29
Other features suggesting high-risk disease: De novo plasma cell leukemia Extramedullary disease Elevated lactate dehydrogenate (LDH) High-risk gene expression profile		

[a]Percentage of patients presenting at each stage.

Abbreviations: β_2M, serum β_2-microglobulin in mg/L; alb, serum albumin in g/dL; FISH, fluorescent in situ hybridization.

Other factors that may influence prognosis are the presence of cytogenetic abnormalities and hypodiploidy by karyotype, fluorescent in situ hybridization (FISH)–identified chromosome 17p deletion, and translocations t(4;14), (14;16), and t(14;20) and 1q34 amplification. Chromosome 13q deletion, previously thought to predict poor outcome, is not a predictor following the use of newer agents. The ISS system, along with cytogenetic changes, is the most widely used method for assessing prognosis (Table 107-3). Microarray profiling has formed the basis for RNA-based prognostic staging systems. Genome sequencing efforts have allowed for characterization of critical genes, pathways, and clonal heterogeneity in myeloma. The median number of mutations per transcribed genome in myeloma is around 58. A very heterogeneous mutational landscape with no unifying mutation has been observed. The most frequently mutated genes are *KRAS* and *NRAS* (about 20% each), followed by *TP53, DIS3, FAM46C*, and *BRAF*, all mutated in 5–10% of the patients. All other mutations were observed in <5% of the patients. These results are now being applied to develop new targeted personalized therapies in myeloma.

TREATMENT

TREATMENT

Multiple Myeloma

No specific intervention is indicated for patients with MGUS. Follow-up once a year or less frequently is adequate except in higher risk MGUS, where serum protein electrophoresis, complete blood count, creatinine, and calcium should be repeated every 6 months. A patient with MGUS and severe polyneuropathy is considered for therapeutic intervention if a causal relationship can be assumed, especially in the absence of any other potential causes for neuropathy. Therapy can include plasmapheresis and occasionally rituximab in patients with IgM MGUS or myeloma-like therapy in those with IgG or IgA disease. About 10% of patients have smoldering MM (SMM) and will have an indolent course demonstrating only slow progression of disease over many

years. For patients with SMM, no specific therapeutic intervention is indicated, although early intervention with lenalidomide and dexamethasone may prevent progression from high-risk SMM to active MM. At present, patients with SMM only require antitumor therapy when myeloma-defining events are identified. Patients with solitary bone plasmacytomas and extramedullary plasmacytomas may be expected to enjoy prolonged disease-free survival after local radiation therapy at a dose of around 40 Gy. Occult marrow involvement may occur at low incidence in patients with solitary bone plasmacytoma. Such patients are usually identified because their serum M component falls slowly or disappears initially after local therapy, only to return after a few months. These patients respond well to systemic therapy.

Patients with symptomatic and/or progressive myeloma require therapeutic intervention. In general, such therapy has two purposes: (1) systemic therapy to control myeloma; and (2) supportive care to control symptoms of the disease, its complications, and adverse effects of therapy. Therapy can significantly prolong survival and improve the quality of life for myeloma patients.

The therapy of myeloma includes an initial induction regimen followed by consolidation and/or maintenance therapy and, on subsequent progression, management of relapsed disease. A number of agents available for use at various stages of the therapy and their doses, schedules, and combinations are detailed in **Table 107-4**. Therapy is partly dictated by the patient's age and comorbidities, which may affect a patient's ability to undergo high-dose therapy and transplantation (**Fig. 107-6**).

Thalidomide, when combined with dexamethasone, achieved responses in two-thirds of newly diagnosed MM patients. Subsequently, lenalidomide, an immunomodulatory derivative of thalidomide, and bortezomib, a proteasome inhibitor, have each been combined with dexamethasone with high response rates (>80%) in newly diagnosed patients with MM. Importantly, their lower toxicity profile with improved efficacy has made them the preferred agents for induction therapy. Efforts to improve the depth of response and the fraction of patients responding have involved using three-drug regimens. The combination of lenalidomide, bortezomib, and dexamethasone achieves close to a 100% response rate and 30% complete response (CR) rate, making it one of the preferred induction regimens in transplant-eligible patients. Other similar three-drug combinations (bortezomib, thalidomide, and dexamethasone or bortezomib, cyclophosphamide, and dexamethasone) also achieve >90% response rate. Herpes zoster prophylaxis is indicated if bortezomib is used, and neuropathy attendant to bortezomib can be decreased both by its subcutaneous administration and administration on a weekly schedule. Lenalidomide use requires prophylaxis for deep-vein thrombosis (DVT) with either aspirin or if patients are at a greater risk of DVT, warfarin or low-molecular-weight heparin. In patients receiving lenalidomide, stem cells should be collected within 6 months, because the continued use of lenalidomide may compromise the ability to collect adequate numbers of stem cells. Initial therapy is continued until maximal cytoreduction. In patients who are transplant candidates, alkylating agents such as melphalan should be avoided because they damage stem cells, leading to decreased ability to collect stem cells for autologous transplant.

In patients who are not transplant candidates due to physiologic age >70 years, significant cardiopulmonary problems, or other comorbid illnesses, the same two- or three-drug combinations described above are considered standard of care as induction therapy. Previously, therapy consisting of intermittent pulses of melphalan, an alkylating agent, with prednisone (MP) was used. However, a number of studies have combined novel agents with MP and reported superior response and survival outcomes. In patients >65 years old, combining thalidomide with MP (MPT) obtains higher response rates and overall survival compared with MP alone. Similarly, significantly improved response (71 vs 35%) and overall survival (3-year survival 72 vs 59%) were observed with the combination of bortezomib and MP compared with MP alone. Lenalidomide added to MP followed by lenalidomide maintenance also prolonged progression-free survival compared with MP alone. These combinations of novel agents with MP also achieve high CR rates (MPT, ~15%; MP plus bortezomib, ~30%; MP plus lenalidomide, ~20%; and MP, ~2–4%). Continuous use of the lenalidomide and dexamethasone combination appears to be superior to the MPT regimen, making it a standard of care for older adults with myeloma.

High-dose therapy (HDT) and consolidation/maintenance are standard practice in the majority of eligible patients. Randomized studies comparing standard-dose therapy to high-dose melphalan therapy with hematopoietic stem cell support have shown that HDT can achieve high overall response rates, with up to 25–40% additional CRs and prolonged progression-free and overall survival; however, few, if any, patients are cured. Although two successive HDTs (tandem transplantations) are more effective than single HDT, the benefit is only observed in the subset of patients who do not achieve a complete or very good partial response to the first transplantation, which is rare. Moreover, a randomized study failed to show any significant difference in overall survival between early transplantation after induction therapy versus delayed transplantation at relapse. These data allow an option to delay transplantation, especially with the availability of more agents and combinations. Allogeneic transplantations may also produce high response rates, but with significant toxicities. Nonmyeloablative allogeneic transplantation can reduce toxicity but is recommended only under the auspices of a clinical trial to exploit an immune graft-versus-myeloma effect while avoiding attendant toxicity.

Maintenance therapy prolongs remissions following standard-dose regimens as well as HDT. Two phase 3 studies have demonstrated improved progression-free survival, and one study showed prolonged overall survival in patients receiving lenalidomide compared to placebo as maintenance therapy after HDT. In nontransplant candidates, two phase 3 studies showed prolonged progression-free survival with lenalidomide maintenance after MP plus lenalidomide or lenalidomide with dexamethasone induction therapy. Although concern arises regarding an increased incidence of second primary malignancies in patients receiving lenalidomide maintenance, its benefits in reducing the risk of progressive disease and death from myeloma far outweigh the small increased risk of second cancers. In patients with high-risk cytogenetics, lenalidomide and bortezomib have been combined and show promise as maintenance therapy after transplantation.

Relapsed myeloma can be treated with a number of agents including lenalidomide and/or bortezomib, if previously not used. These agents in combination with dexamethasone can achieve a partial response rate of up to 60% and a 10–15% CR rate in patients with relapsed disease. The combination of bortezomib and liposomal doxorubicin is active in relapsed myeloma. Thalidomide, if not used as initial therapy, can achieve responses in refractory cases. The second-generation proteasome inhibitor carfilzomib and immunomodulatory agent pomalidomide have shown efficacy in relapsed and refractory MM, even MM refractory to lenalidomide and bortezomib. An oral proteasome inhibitor, ixazomib has also been approved in combination with lenalidomide and dexamethasone as an all-oral regimen for relapsed MM. Two antibodies are approved for treatment of relapsed MM. Daratumumab targeting CD38 achieves high response rates and improved progression-free survival as a single agent with further improvement in response and survival when added to bortezomib and dexamethasone, or lenalidomide and dexamethasone. Elotuzumab targeting SLAMF7 has shown significant activity in combination with lenalidomide and dexamethasone in relapsed/refractory myeloma but not as a single agent. Panobinostat, a histone deacetylase inhibitor, in combination with bortezomib and dexamethasone has been approved for treatment of relapsed refractory myeloma based on superior response and progression-free survival compared to bortezomib and dexamethasone alone. Incorporation of the large number of active agents at various stages of therapies including in the newly diagnosed patients is improving survival as well as quality of life.

Improvement in the serum M component may lag behind the symptomatic improvement due to longer half-life (~3 weeks) of the immunoglobulin. The fall in M component depends on the rate of tumor kill and the fractional catabolic rate of immunoglobulin. Light chain excretion, with a functional half-life of ~6 h, may fall within the first week of treatment. Because urine light chain levels may relate to renal tubular function, they are not a reliable measure of tumor cell kill in patients with renal dysfunction; however, improvements in serum free light chain measurement are often seen sooner. Sequencing- and multicolor flow cytometry-based methods are now used to assess minimal residual disease (MRD) in bone marrow. Absence of MRD predicts longer survival. Although patients may not achieve complete remission, clinical responses may last for long periods of time.

The median overall survival of patients with myeloma is 8+ years, with subsets of younger patients surviving >10 years. The major causes of death are progressive myeloma, renal failure, sepsis, or therapy-related myelodysplasia. Nearly a quarter of patients die of myocardial infarction, chronic lung disease, diabetes, or stroke—all intercurrent illnesses related more to the age of the patient group than to the tumor.

Supportive care directed at the anticipated complications of the disease may be as important as primary antitumor therapy. Hypercalcemia generally responds well to bisphosphonates, glucocorticoid therapy, hydration, and natriuresis, and rarely requires calcitonin as well. Bisphosphonates (e.g., pamidronate 90 mg or zoledronate 4 mg initially once a month and later less frequently) reduce osteoclastic bone resorption and preserve performance status and quality of life, decrease bone-related complications, and may also have antitumor effects. Osteonecrosis of the jaw and renal dysfunction can occur in a minority of patients receiving bisphosphonate therapy. Treatments aimed at strengthening the skeleton such as fluorides, calcium, and vitamin D, with or without androgens, have been suggested, but are not of proven efficacy. Kyphoplasty or vertebroplasty should be considered in patients with painful collapsed vertebra. Iatrogenic worsening of renal function may be prevented by maintaining a high fluid intake to prevent dehydration and enhance excretion of light chains and calcium. In the event of acute renal failure, plasmapheresis is ~10 times more effective at clearing light chains than peritoneal dialysis; however, its role in reversing renal failure remains controversial. Importantly, reducing the protein load by effective antitumor therapy with agents such as bortezomib may result in improvement in renal function in over half of the patients. Use of lenalidomide in renal failure is possible but requires dose modification, as it is renally excreted. Urinary tract infections should be watched for and treated early. Plasmapheresis may be the treatment of choice for hyperviscosity syndromes. Although the pneumococcus is a dreaded pathogen in myeloma patients, pneumococcal polysaccharide vaccines may not elicit an antibody response. The pneumococcal conjugate vaccines may be more protective. Prophylactic administration of intravenous γ globulin preparations is used in the setting of recurrent serious infections. Chronic oral antibiotic prophylaxis is not warranted. Patients developing neurologic symptoms in the lower extremities, severe localized back pain, or problems with bowel and bladder control may need emergency MRI and local radiation therapy and glucocorticoids if cord compression is identified. In patients in whom neurologic deficit is increasing or substantial, emergent surgical decompression may be necessary. Most bone lesions respond to analgesics and systemic therapy, but certain painful lesions may respond most promptly to localized radiation. The anemia associated with myeloma may respond to erythropoietin along with hematinics (iron, folate, cobalamin). The pathogenesis of the anemia should be established and specific therapy instituted, whenever possible.

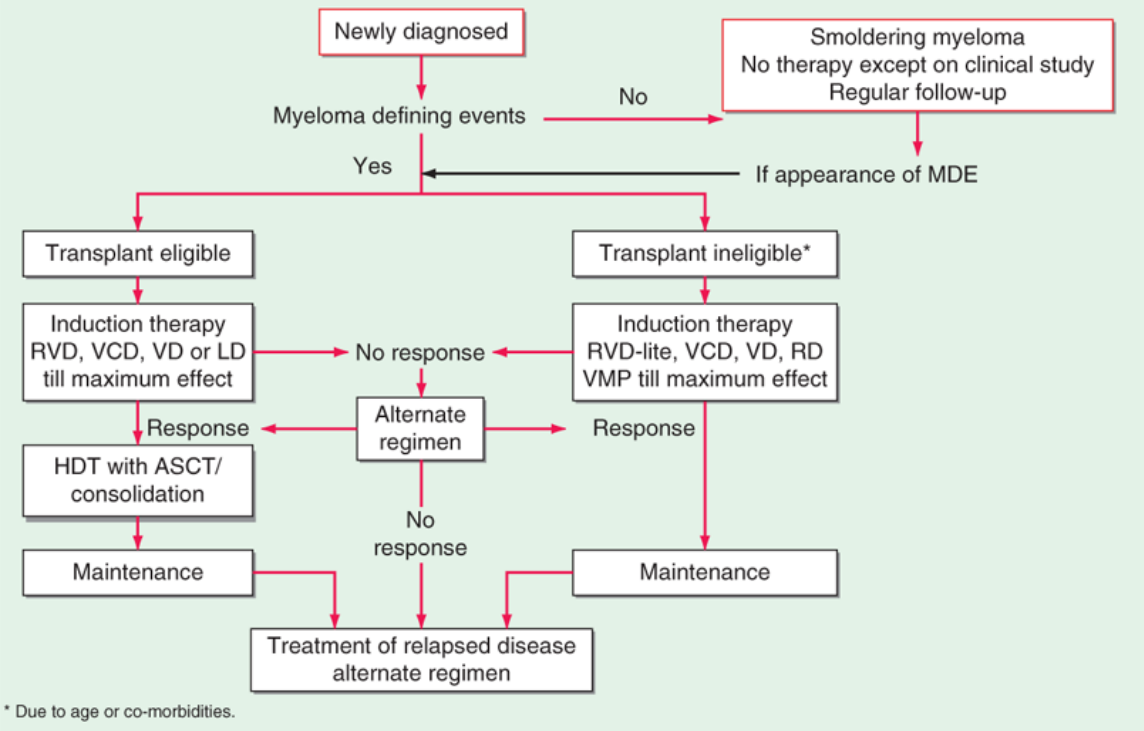
TABLE 107-4

Standard Therapeutic Agents in Myeloma

Class	Agent	Standard Dosage and Administration	Combination	Myeloma Indication
Immunomodulatory agents	Thalidomide (T)	Oral 50–200 mg qd	TD, VTD	Newly diagnosed and relapsed
	Lenalidomide (R)	Oral 5–25 mg daily × 21 days q 4 week	RD, RVD, DaRD, ERD, KRD, IRD	Relapsed
	Pomalidomide (P)	Oral 2–4 mg daily × 21 days q 4 week	PD	Relapsed
Proteasome inhibitor	Bortezomib (V)	IV or SC 1.3 mg/m^2 days 1, 4, 8, 11 OR 1, 8, 15	VD, VTD, VRD, DaVD, VCD	Newly diagnosed and relapsed
	Carfilzomib (K)	IV 20–56 mg/m^2 days 1, 2, 8, 9, 15, 16 q 4 weeks	KD, KRD	Relapsed
	Ixazomib (I)	Oral 4 mg days 1,8,15	IRD	Relapsed
Antibodies	Daratumumab (Da)	IV 16 mg/kg/week for 8 weeks then every 2 weeks for 16 weeks and then every 4 weeks thereafter	Dara, DaRD, DaVD	Relapsed
	Elotuzumab (E)	IV 10 mg/kg days 1, 8, 15, and 22 for first two cycles then on days 1 and 15. Along with RD	EloRD	Relapsed
Histone deacetylase inhibitor	Panobinostat (Pa)		PaVD	Relapsed
Alkylating agents	Melphalan (M)	Oral melphalan, 0.25 mg/kg per day for 4 days (with Pred) every 4–6 weeks	MP, MPT, MPR, MPV, high-dose M	Newly diagnosed and relapsed conditioning
	Cyclophosphamide (C)	IV—300–500 mg/m^2 weekly × 2 q 4 weeks Oral—50 mg qd × 21 days	VCD	Newly diagnosed and relapsed
	Bendamustine (B)	IV 70–90 mg days 1, 2 or days 1, 8 q 4 weeks	BD or BVD	Relapsed
Steroid	Dexamethasone (D) Prednisone (P)	Oral 10–40 mg q week Oral 1 mg/kg		All stages

FIGURE 107-6

Treatment algorithm for multiple myeloma. C, cyclophosphamide; D, dexamethasone; M, melphalan; P, prednisone; R, lenalidomide; RVD-lite, weekly regimen; V, bortezomib. Alternate regimen-combinations including daratumumab; elotuzumab; panobinostat; carfilzomib; ixazomib, pomalidomide or agents; ASCT, autologous stem cell transplantation; HDT, high-dose therapy; MDE, myeloma defining events.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

WALDENSTRÖM'S MACROGLOBULINEMIA

In 1948, Waldenström described a malignancy of lymphoplasmacytoid cells that secreted IgM. In contrast to myeloma, the disease was associated with lymphadenopathy and hepatosplenomegaly, but the major clinical manifestation was hyperviscosity syndrome. The disease resembles the related diseases CLL, myeloma, and lymphocytic lymphoma. It originates from a postgerminal center B cell that has undergone somatic mutations and antigenic selection in the lymphoid follicle and has the characteristics of an IgM-bearing memory B cell. Waldenström's macroglobulinemia (WM) and IgM myeloma follow a similar clinical course, but therapeutic options are different. The diagnosis of IgM myeloma is usually reserved for patients with lytic bone lesions and predominant infiltration with CD138+ plasma cells in the bone marrow. Such patients are at greater risk of pathologic fractures than patients with WM.

A familial occurrence is common in WM, but its molecular bases are yet unclear. A distinct *MYD88* L265P somatic mutation is present in >90% of patients with WM and the majority of IgM MGUS. Other commonly occurring mutations include *CXCR4* (30–40%), *ARID1A* (17%), and *CD79B* (8–15%). Presence of *MYD88* mutation status is now used as a diagnostic test to discriminate WM from marginal zone lymphomas (MZLs), IgM-secreting myeloma, and CLL with plasmacytic differentiation. This mutation also explains the molecular pathogenesis of the disease, with involvement of Toll-like receptor (TLR) and interleukin 1 receptor (IL-1R) signaling leading to activation of IL-1R–associated kinase (IRAK) 4 and IRAK1 followed by nuclear factor-κB (NF-κB) activation. *MYD88* mutation also triggers Burton's tyrosine kinase (BTK), hemopoietic cell kinase (HCK) growth, and survival signaling, which are now important therapeutic targets in WM. *CXCR4* mutations induce AKT and extracellular regulated kinase-1/2 (ERK 1/2) signaling. This pathway can lead to development of drug resistance in the presence of its ligand CXCL12.

The disease is similar to myeloma in being slightly more common in men and occurring with increased incidence with increasing age (median 64 years). The IgM in some patients with macroglobulinemia may have specificity for myelin-associated glycoprotein (MAG), a protein that has been associated with demyelinating disease of the peripheral nervous system and may be lost earlier and to a greater extent than the better known myelin basic protein in patients with multiple sclerosis. Sometimes patients with macroglobulinemia develop a peripheral neuropathy, and half of these patients are positive for anti-MAG antibody. The neuropathy may precede the appearance of the neoplasm. The whole process may begin with a viral infection that may elicit an antibody response that cross-reacts with a normal tissue component.

Like myeloma, the disease involves the bone marrow, but unlike myeloma, it does not cause bone lesions or hypercalcemia. Bone marrow shows >10% infiltration with lymphoplasmacytic cells (surface IgM+, CD19+, CD20+, and CD22+, rarely CD5+, but CD10− and CD23−) with an increase in number of mast cells. Like myeloma, an M component is present in the serum in excess of 30 g/L (3 g/dL), but unlike myeloma, the size of the IgM paraprotein results in little renal excretion, and only ~20% of patients excrete light chains. Therefore, renal disease is not common. The light chain isotype is kappa in 80% of the cases. Patients present with weakness, fatigue, and recurrent infections similar to myeloma patients, but epistaxis, visual disturbances, and neurologic symptoms such as peripheral neuropathy, dizziness, headache, and transient paresis are much more common in macroglobulinemia. Presence of *MYD88* and *CXCR4* mutations also affects disease presentation. Presence of *CXCR4* mutations is associated with higher bone marrow disease burden and higher incidence of hyperviscosity. Patients with wild-type *MYD88* show lower bone marrow disease burden.

Physical examination reveals adenopathy and hepatosplenomegaly, and ophthalmoscopic examination may reveal vascular segmentation and dilation of the retinal veins characteristic of hyperviscosity states. Patients may have a normocytic, normochromic anemia, but rouleaux formation and a positive Coombs' test are much more common than in myeloma. Malignant lymphocytes are usually present in the peripheral blood. About 10% of macroglobulins are cryoglobulins. These are pure M components and are not the mixed cryoglobulins seen in rheumatoid arthritis and other autoimmune diseases. Mixed cryoglobulins are composed of IgM or IgA complexed with IgG, for which they are specific. In both cases, Raynaud's phenomenon and serious vascular symptoms precipitated by the cold may occur, but mixed cryoglobulins are not commonly associated with malignancy. Patients suspected of having a cryoglobulin based on history and physical examination should have their blood drawn into a warm syringe and delivered to the laboratory in a container of warm water to avoid errors in quantitating the cryoglobulin.

TREATMENT

TREATMENT

Waldenström's Macroglobulinemia

Control of serious hyperviscosity symptoms such as an altered state of consciousness or paresis can be achieved acutely by plasmapheresis because 80% of the IgM paraprotein is intravascular. The median survival of affected individuals is ~50 months. However, many patients with WM have indolent disease that does not require therapy. Pretreatment parameters including older age, male sex, general symptoms, and cytopenias define a high-risk population. Treatment is usually not initiated unless the disease is symptomatic or increasing anemia, hyperviscosity, lymphadenopathy, or hepatosplenomegaly is present. Ibrutinib is approved by the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) for use in patients with symptomatic WM. It targets the constitutively activated BTK. In patients with one prior line of therapy, the overall response to ibrutinib was 91%. Best responses to ibrutinib are observed in patients with mutated *MYD88* and wild-type *CXCR4* status, while delayed and lower response rates to ibrutinib are observed in patients with mutated *CXCR4*. The other first line treatments include rituximab (anti-CD20) alone or combined with alkylators (bendamustine and cyclophosphamide), or proteasome inhibitors (bortezomib). Rituximab can produce IgM flare, so either plasmapheresis should be used before rituximab or its use should be initially withheld in patients with high IgM levels. Fludarabine (25 mg/m^2 per day for 5 days every 4 weeks) and cladribine (0.1 mg/kg per day for 7 days every 4 weeks) are also highly effective single agents. With identification of the *MYD88* mutation, inhibitors targeting IRAK1/4 and BCL2 are being evaluated. Although high-dose therapy plus autologous transplantation is an option, its use has declined due to the availability of other effective agents.

POEMS SYNDROME

The features of this syndrome are *p*olyneuropathy, *o*rganomegaly, *e*ndocrinopathy, *M*-protein, and *s*kin changes (POEMS). Diagnostic criteria are described in Table 107-1. Patients usually have a severe, progressive sensorimotor polyneuropathy associated with sclerotic bone lesions from myeloma. Polyneuropathy occurs in ~1.4% of myelomas, but the POEMS syndrome is only a rare subset of that group. Unlike typical myeloma, hepatomegaly and lymphadenopathy occur in about two-thirds of patients, and splenomegaly is seen in one-third. The lymphadenopathy frequently resembles Castleman's disease histologically, a condition that has been linked to IL-6 overproduction. The endocrine manifestations include amenorrhea in women and impotence and gynecomastia in men. Hyperprolactinemia due to loss of normal inhibitory control by the hypothalamus may be associated with other central nervous system manifestations such as papilledema and elevated cerebrospinal fluid pressure and protein. Type 2 diabetes mellitus occurs in about one-third of patients. Hypothyroidism and adrenal insufficiency are occasionally noted. Skin changes are diverse: hyperpigmentation, hypertrichosis, skin thickening, and digital clubbing. Other manifestations include peripheral edema, ascites, pleural effusions, fever, and thrombocytosis. Not all the components of POEMS syndrome may be present initially.

The pathogenesis of the disease is unclear, but high circulating levels of the proinflammatory cytokines IL-1, IL-6, VEGF, and TNF have been documented, and levels of the inhibitory cytokine transforming growth factor β are lower than expected. Treatment of the myeloma may result in an improvement in the other disease manifestations.

Patients are often treated similarly to those with myeloma. Plasmapheresis does not appear to be of benefit in POEMS syndrome. Patients presenting with isolated sclerotic lesions may have resolution of neuropathic symptoms after local therapy for plasmacytoma with radiotherapy. Similar to MM, novel agents and high-dose therapy with autologous stem cell transplantation have been pursued in selected patients and have been associated with prolonged progression-free survival.

HEAVY CHAIN DISEASES

The heavy chain diseases are rare lymphoplasmacytic malignancies. Their clinical manifestations vary with the heavy chain isotype. Patients have absence of light chain and secrete a defective heavy chain that usually has an intact Fc fragment and a deletion in the Fd region. Gamma, alpha, and mu heavy chain diseases have been described, but no reports of delta or epsilon heavy chain diseases have appeared. Molecular biologic analysis of these tumors has revealed structural genetic defects that may account for the aberrant chain secreted.

GAMMA HEAVY CHAIN DISEASE (FRANKLIN'S DISEASE)

This disease affects individuals of widely different age groups and countries of origin. It is characterized by lymphadenopathy, fever, anemia, malaise, hepatosplenomegaly, and weakness. It is frequently associated with autoimmune diseases, especially rheumatoid arthritis. Its most distinctive symptom is palatal edema, resulting from involvement of nodes in Waldeyer's ring, and this may progress to produce respiratory compromise. The diagnosis depends on the demonstration of an anomalous serum M component (often <20 g/L [<2 g/dL]) that reacts with anti-IgG but not antilight chain reagents. The M component is typically present in both serum and urine. Most of the paraproteins have been of the γ_1 subclass, but other subclasses have been seen. The patients may have thrombocytopenia, eosinophilia, and nondiagnostic bone marrow that may show increased numbers of lymphocytes or plasma cells that do not stain for light chain. Patients usually have a rapid downhill course and die of infection; however, some patients have survived 5 years with chemotherapy. Therapy is indicated when symptomatic and involves chemotherapeutic combinations used in low-grade lymphoma. Rituximab has also been reported to show efficacy.

ALPHA HEAVY CHAIN DISEASE (SELIGMANN'S DISEASE)

This is the most common of the heavy chain diseases. It is closely related to a malignancy known as Mediterranean lymphoma, a disease that affects young persons in parts of the world where intestinal parasites are common, such as the Mediterranean, Asia, and South America. The disease is characterized by an infiltration of the lamina propria of the small intestine with lymphoplasmacytoid cells that secrete truncated alpha chains. Demonstrating alpha heavy chains is difficult because the alpha chains tend to polymerize and appear as a smear instead of a sharp peak on electrophoretic profiles. Despite the polymerization, hyperviscosity is not a common problem in alpha heavy chain disease. Without J chain–facilitated dimerization, viscosity does not increase dramatically. Light chains are absent from serum and urine. The patients present with chronic diarrhea, weight loss, and malabsorption and have extensive mesenteric and paraaortic adenopathy. Respiratory tract involvement occurs rarely. Patients may vary widely in their clinical course. Some may develop diffuse aggressive histologies of malignant lymphoma. Chemotherapy may produce long-term remissions. Rare patients appear to have responded to antibiotic therapy, raising the question of the etiologic role of antigenic stimulation, perhaps by some chronic intestinal infection. Chemotherapy plus antibiotics may be more effective than chemotherapy alone. IPSID is recognized as an infectious pathogen–associated human lymphoma that has association with *Campylobacter jejuni*. It involves mainly the proximal small intestine resulting in malabsorption, diarrhea, and abdominal pain. IPSID is associated with excessive plasma cell differentiation and produces truncated alpha heavy chain proteins lacking the light chains as well as the first constant domain. Early-stage IPSID responds to antibiotics (30–70% complete remission). Most untreated IPSID patients progress to lymphoplasmacytic and immunoblastic lymphoma. Patients not responding to antibiotic therapy are considered for treatment with combination chemotherapy used to treat low-grade lymphoma.

MU HEAVY CHAIN DISEASE

The secretion of isolated mu heavy chains into the serum appears to occur in a very rare subset of patients with CLL. The only features that may distinguish patients with mu heavy chain disease are the presence of vacuoles in the malignant lymphocytes and the excretion of kappa light chains in the urine. The diagnosis requires ultracentrifugation or gel filtration to confirm the nonreactivity of the paraprotein with the light chain reagents because some intact macroglobulins fail to interact with these serums. The tumor cells seem to have a defect in the assembly of light and heavy chains because they appear to contain both in their cytoplasm. Such patients are not treated differently from other patients with CLL **(Chap. 104)**.

FURTHER READING

Attal M et al: IFM 2009 study. Lenalidomide, bortezomib, and dexamethasone with transplantation for myeloma. N Engl J Med 376:1311, 2017.

Avet-Loiseau H et al: Long-term analysis of the IFM 99 trials for myeloma: Cytogenetic abnormalities [t(4;14), del(17p), 1q gains] play a major role in defining long-term survival. J Clin Oncol 30:1949, 2012.

Chng WJ et al: International Myeloma Working Group. IMWG consensus on risk stratification in multiple myeloma. Leukemia 28:269, 2014.

Dimopoulos MA et al: Daratumumab, lenalidomide, and dexamethasone for multiple myeloma. N Engl J Med 375:1319, 2016.

Hunter ZR et al: Genomics, signaling, and treatment of Waldenstrom macroglobulinemia. J Clin Oncol 35:994, 2017.

Landgren O et al: Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: A prospective study. Blood 113:5412, 2009.

Palumbo A et al: Revised International Staging System for multiple myeloma: A report from International Myeloma Working Group. J Clin Oncol 33:2863, 2015.

Raje N, Roodman GD: Advances in the biology and treatment of bone disease in multiple myeloma. Clin Cancer Res 17:1278, 2011.

Rajkumar SV et al: International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. Lancet Oncol 15:e538, 2014. [PubMed: 25439696]

Robiou du Pont S et al: Genomics of multiple myeloma. J Clin Oncol 35:963, 2017.

Chapter 108: Amyloidosis

John L. Berk; Vaishali Sanchorawala

GENERAL PRINCIPLES

Amyloidosis is the term for a group of protein misfolding disorders characterized by the extracellular deposition of insoluble polymeric protein fibrils in tissues and organs. A robust cellular machinery exists to chaperone proteins during the process of synthesis and secretion, to ensure that they achieve correct tertiary conformation and function, and to eliminate proteins that misfold. However, genetic mutation, incorrect processing, and other factors may favor misfolding, with consequent loss of normal protein function and intracellular or extracellular aggregation. Many diseases, ranging from cystic fibrosis to Alzheimer's disease, are now known to involve protein misfolding. In the amyloidoses, the aggregates are typically extracellular, and the misfolded protein subunits assume a common antiparallel, β-pleated sheet–rich structural conformation that leads to the formation of higher-order oligomers and then fibrils with unique staining properties. The term *amyloid* was coined around 1854 by the pathologist Rudolf Virchow, who thought that these deposits resembled starch (Latin *amylum*) under the microscope.

Amyloid diseases, defined by the biochemical nature of the protein composing the fibril deposits, are classified according to whether they are systemic or localized, whether they are acquired or inherited, and their clinical patterns (**Table 108-1**). The standard nomenclature is *AX*, where *A* indicates amyloidosis and *X* represents the protein present in the fibril. This chapter focuses primarily on the systemic forms. *AL* refers to amyloid composed of immunoglobulin light chains (LCs); this disorder, formerly termed *primary systemic amyloidosis*, arises from a clonal B cell or plasma cell disorder and can be associated with myeloma or lymphoma. *ATTR* refers to the most prevalent of the *familial amyloidoses*, which are most commonly due to mutations in transthyretin (TTR), the transport protein for thyroid hormone and retinol-binding protein. *AA* amyloid is composed of the acute-phase reactant protein serum amyloid A (SAA) and occurs in the setting of chronic inflammatory or infectious diseases; for this reason, this type was formerly known as *secondary amyloidosis*. *A*β_2*M* amyloid results from misfolded β_2-microglobulin, occurring in individuals with long-standing renal disease who have undergone dialysis, typically for years. *Aβ*, the most common form of localized amyloidosis, is found in the brain of patients with Alzheimer's disease after abnormal proteolytic processing and aggregation of polypeptides derived from the amyloid precursor protein.

TABLE 108-1

Amyloid Precursor Proteins and Their Clinical Syndromes

Designation	Precursor	Clinical Syndrome	Clinical Involvement
Systemic Amyloidoses			
AL	Immunoglobulin light chain	Primary or myeloma-associated[a]	Any
AH	Immunoglobulin heavy chain	Rare variant of primary or myeloma-associated	Any
AA	Serum amyloid A protein	Secondary; reactive[b]	Renal, heart, other
$A\beta_2M$	β_2-Microglobulin	Hemodialysis-associated	Synovial tissue, bone
ATTR	Transthyretin	Familial (mutant) Age-related (wild type)	Cardiac, peripheral and autonomic nerves
AApoAI	Apolipoprotein AI	Familial	Hepatic, renal
AApoAII	Apolipoprotein AII	Familial	Renal
AGel	Gelsolin	Familial	Cornea, cranial nerves, skin, renal
AFib	Fibrinogen Aα	Familial	Renal
ALys	Lysozyme	Familial	Renal, hepatic
ALECT2	Leukocyte chemotactic factor 2	Undefined	Renal
Localized Amyloidoses			
Aβ	Amyloid β protein	Alzheimer's disease; Down's syndrome	Central nervous system

Designation	Precursor	Clinical Syndrome	Clinical Involvement
ACys	Cystatin C	Cerebral amyloid angiopathy	Central nervous system, vascular
APrP	Prion protein	Spongiform encephalopathies	Central nervous system
AIAPP	Islet amyloid polypeptide (amylin)	Diabetes-associated	Pancreas
ACal	Calcitonin	Medullary carcinoma of the thyroid	Thyroid
AANF	Atrial natriuretic factor	Atrial fibrillation	Cardiac atria
APro	Prolactin	Endocrinopathy	Pituitary
ASgl	Semenogelin I	Age-related; incidental autopsy or biopsy finding	Seminal vesicles

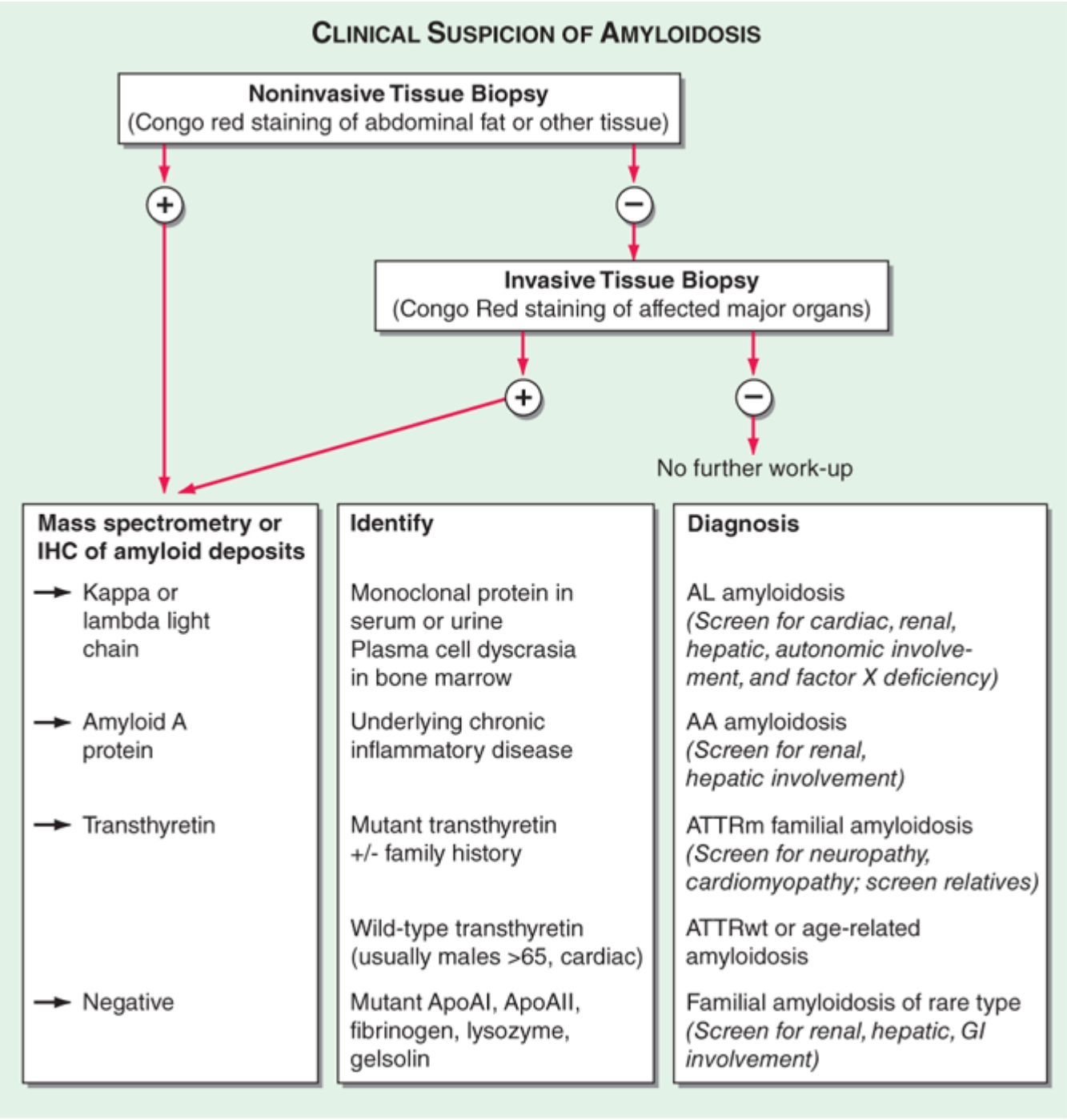
[a]Localized AL deposits can occur in skin, conjunctiva, urinary bladder, and the tracheobronchial tree. [b]Secondary to chronic inflammation or infection or to a hereditary periodic fever syndrome such as familial Mediterranean fever.

Diagnosis and treatment of the amyloidoses rest upon the histopathologic identification of amyloid deposits and immunohistochemical, biochemical, or genetic determination of amyloid type (**Fig. 108-1**). In the systemic amyloidoses, the clinically involved organs can be biopsied, but amyloid deposits may be found in any tissue of the body. Historically, blood vessels of the gingiva or rectal mucosa were often examined, but the most easily accessible tissue—positive in more than 80% of patients with systemic amyloidosis—is abdominal fat. After local anesthesia, fat is aspirated from the abdominal pannus with a 16-gauge needle. Fat globules expelled onto a glass slide can be stained, thus avoiding a surgical procedure. If this material is negative, more invasive biopsies of the kidney, heart, liver, or gastrointestinal tract can be considered in patients in whom amyloidosis is suspected. The regular β-sheet structure of amyloid deposits exhibits a unique “apple green” birefringence by polarized light microscopy when stained with Congo red dye; other regular protein structures (e.g., collagen) appear white under these conditions. The 10-nm-diameter fibrils can also be visualized by electron microscopy of paraformaldehyde-fixed tissue. Once amyloid is found, the precursor protein type must be determined by immunohistochemistry, immunoelectron microscopy, or extraction and biochemical analysis employing mass spectrometry; gene sequencing is used to identify mutants causing hereditary amyloidosis. The patient’s history, physical findings, and clinical presentation, including age and ethnic origin, organ system involvement, underlying diseases, and family history, may

provide helpful clues as to the type of amyloidosis. However, there can be considerable overlap in clinical presentations, and accurate typing is essential to guide appropriate therapy.

FIGURE 108-1

Algorithm for the diagnosis of amyloidosis and determination of type. Clinical suspicion: unexplained nephropathy, cardiomyopathy, neuropathy, enteropathy, arthropathy, and macroglossia. ApoAI, apolipoprotein AI; ApoAII, apolipoprotein AII; GI, gastrointestinal.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The mechanisms of fibril formation and tissue toxicity remain controversial. The “amyloid hypothesis,” as it is currently understood, proposes that precursor proteins undergo a process of reversible unfolding or

misfolding; misfolded proteins form oligomeric aggregates, higher-order polymers, and then fibrils that deposit in tissues. Accumulating evidence suggests that the oligomeric intermediates may constitute the most toxic species. Oligomers are more capable than large fibrils of interacting with cells and inducing formation of reactive oxygen species and stress signaling. Ultimately, the fibrillar tissue deposits are likely to interfere with normal organ function. A more sophisticated understanding of the mechanisms leading to amyloid formation and cell and tissue dysfunction will continue to provide new targets for therapies.

The clinical syndromes of the amyloidoses are associated with relatively nonspecific alterations in routine laboratory tests. Blood counts are usually normal, although the erythrocyte sedimentation rate is frequently elevated. Patients with glomerular kidney involvement generally have proteinuria, often in the nephrotic range, leading to hypoalbuminemia that may be severe; patients with serum albumin levels <2 g/dL generally have pedal edema or anasarca. Amyloid cardiomyopathy is characterized by concentric ventricular hypertrophy and diastolic dysfunction associated with elevation of brain natriuretic peptide or N-terminal pro–brain natriuretic peptide as well as troponin. These cardiac biomarkers can be used for disease staging, prognostication, and disease activity monitoring in patients with AL amyloidosis. Notably, renal insufficiency can falsely elevate levels of these biomarkers. Recently, biomarkers of cardiac remodeling—that is, matrix metalloproteinases and tissue inhibitors of metalloproteinases—have been found to be altered in the serum of patients with amyloid cardiomyopathy. Electrocardiographic and echocardiographic features of amyloid cardiomyopathy are described below. Patients with liver involvement, even when advanced, usually develop cholestasis with an elevated alkaline phosphatase concentration with minimal alteration of the aminotransferases and preservation of synthetic function. In AL amyloidosis, endocrine organs may be infiltrated with fibrils, and hypothyroidism, hypoadrenalism, or even hypopituitarism can occur. Although none of these findings is specific for amyloidosis, the presence of abnormalities in multiple organ systems should raise suspicion regarding this diagnosis.

AL AMYLOIDOSIS

Etiology and Incidence

AL amyloidosis is most frequently caused by a clonal expansion of bone-marrow plasma cells that secrete a monoclonal immunoglobulin LC depositing as amyloid fibrils in tissues. Whether the clonal plasma cells produce a LC that misfolds and leads to AL amyloidosis or an LC that folds properly, allowing the cells to inexorably expand over time and develop into multiple myeloma **(Chap. 107)**, may depend upon primary sequence of the clonal LC or other genetic or epigenetic factors. AL amyloidosis can occur with multiple myeloma or other B lymphoproliferative diseases, including non-Hodgkin's lymphoma **(Chap. 104)** and Waldenström's macroglobulinemia **(Chap. 107)**. AL amyloidosis is the most common type of systemic amyloidosis diagnosed in North America. Its incidence has been estimated at 4.5 cases/100,000 population; however, ascertainment continues to be inadequate, and the true incidence may be much higher. AL amyloidosis, like other plasma cell diseases, usually occurs after age 40 and is often rapidly progressive and fatal if untreated.

Pathology and Clinical Features

Amyloid deposits are usually widespread in AL amyloidosis and can be present in the interstitium of any organ outside the central nervous system. The amyloid fibril deposits are composed of full-length 23-kDa monoclonal immunoglobulin LCs as well as fragments. Accessory molecules co-deposited with LC fibrils (as well as with other amyloid fibrils) include serum amyloid P component, apolipoproteins e and AIV, glycosaminoglycans, and metal ions. Although all kappa and lambda LC subtypes have been identified in AL amyloid fibrils, lambda subtypes predominate. The lambda 6 subtype appears to have unique structural properties that predispose it to fibril formation, often in the kidney.

AL amyloidosis is often a rapidly progressive disease that presents as a pleiotropic set of clinical syndromes, recognition of which is key for initiation of the appropriate workup. Nonspecific symptoms of fatigue and weight loss are common; however, the diagnosis is rarely considered until symptoms referable to a specific organ develop. The kidneys are the most frequently involved organ and are affected in 70–80% of patients. Renal amyloidosis usually manifests as proteinuria, often in the nephrotic range and associated with hypoalbuminemia, secondary hypercholesterolemia and hypertriglyceridemia, and edema or anasarca. In some patients, interstitial rather than glomerular amyloid deposition can produce azotemia without proteinuria. The heart is the second most commonly affected organ (50–60% of patients), and cardiac involvement is the leading cause of death from AL amyloidosis. Early on, the electrocardiogram may show low voltage in the limb leads with a pseudo-infarct pattern. Echocardiographic features of disease include concentrically thickened ventricles and diastolic dysfunction with an abnormal global longitudinal strain pattern; a “sparkly” appearance has been described but is often not seen with modern high-resolution echocardiographic techniques. Poor atrial contractility occurs even in sinus rhythm, and patients with cardiac amyloidosis are at risk for development of atrial thrombi and stroke. Cardiac MRI can show increased wall thickness, and characteristic delayed enhancement of the subendocardium has been described following injection of gadolinium contrast. Nervous system symptoms include peripheral sensorimotor neuropathy and/or autonomic dysfunction manifesting as gastrointestinal motility disturbances (early satiety, diarrhea, constipation), dry eyes and mouth, impotence, orthostatic hypotension, and/or neurogenic bladder. Macroglossia (**Fig. 108-2A**), a pathognomonic sign of AL amyloidosis, is seen in only ~10% of patients. Liver involvement causes cholestasis and hepatomegaly. The spleen is frequently involved, and there may be functional hyposplenism in the absence of significant splenomegaly. Many patients experience “easy bruising” due to amyloid deposits in capillaries or deficiency of clotting factor X due to binding to amyloid fibrils; cutaneous ecchymoses appear, particularly around the eyes, producing another uncommon but pathognomonic finding, the “raccoon-eye” sign (**Fig. 108-2B**). Other findings include nail dystrophy (**Fig. 108-2C**), alopecia, and amyloid arthropathy with thickening of synovial membranes in the wrists and shoulders. The presence of a multisystemic illness or general fatigue along with any of these clinical syndromes should prompt a workup for amyloidosis.

FIGURE 108-2

Clinical signs of AL amyloidosis. ***A.*** Macroglossia. ***B.*** Periorbital ecchymoses. ***C.*** Fingernail dystrophy.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

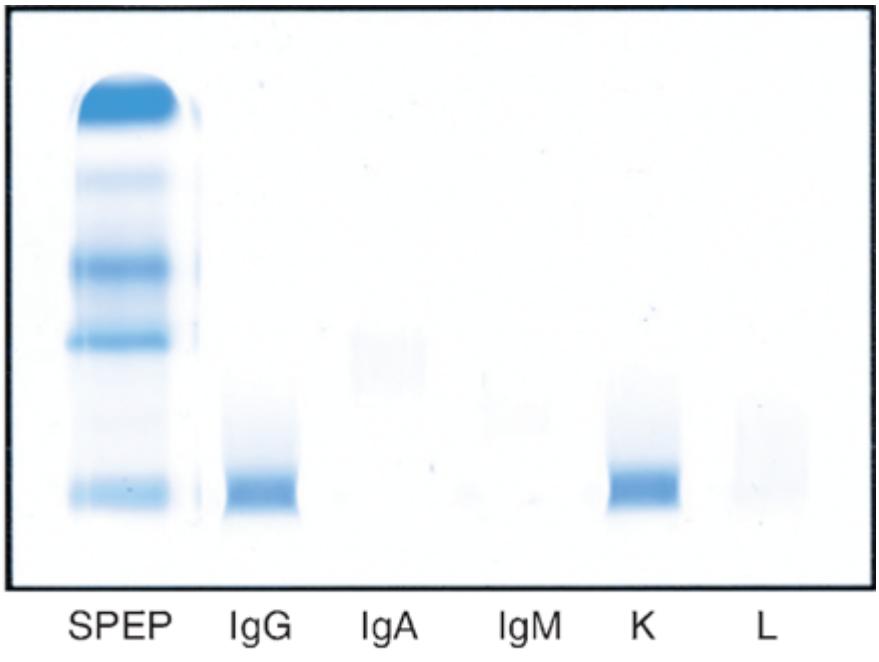
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Diagnosis

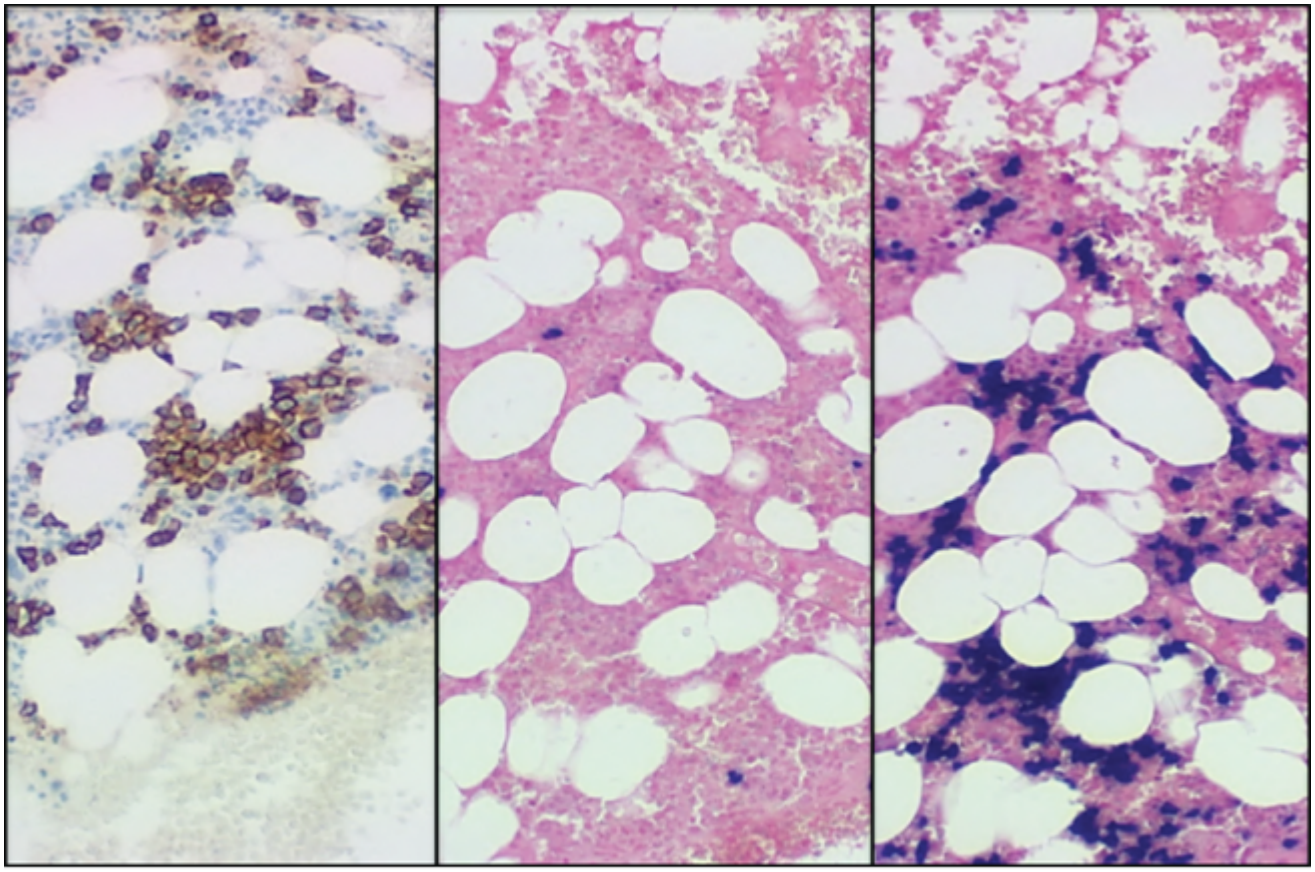
Identification of an underlying clonal plasma cell or B lymphoproliferative process and a clonal LC are key to the diagnosis of AL amyloidosis. Serum protein electrophoresis and urine protein electrophoresis, although of value in multiple myeloma, are *not* useful screening tests if AL amyloidosis is suspected because the clonal LC or whole immunoglobulin often is not present in sufficient amounts to produce a monoclonal "M-spike" in the serum or LC (Bence Jones) protein in the urine. However, more than 90% of patients with AL amyloidosis have serum or urine monoclonal LC or whole immunoglobulin detectable by immunofixation electrophoresis of serum (SIFE) or urine (UIFE) **(Fig. 108-3A)** or by nephelometric measurement of serum "free" LCs (i.e., LCs circulating in monomeric form rather than in an immunoglobulin tetramer with heavy chain). Examining the ratio as well as the absolute amount of serum-free LCs is essential, as renal insufficiency reduces LC clearance, nonspecifically elevating both isotypes. In addition, an increased percentage of plasma cells in the bone marrow—typically 5–30% of nucleated cells—is found in ~90% of patients. Kappa or lambda clonality should be demonstrated by flow cytometry, immunohistochemistry, or in situ hybridization for LC mRNA **(Fig. 108-3B)**.

FIGURE 108-3

Laboratory features of AL amyloidosis. ***A.*** Serum immunofixation electrophoresis reveals an IgGκ monoclonal protein in this example; serum protein electrophoresis is often normal. ***B.*** Bone-marrow biopsy sections stained by immunohistochemistry with antibody to CD138 (syndecan, highly expressed on plasma cells) (*left*) or by in situ hybridization with fluorescein-tagged probes (Ventana Medical Systems) binding to κ mRNA (*center*) and λ mRNA (*right*) in plasma cells. *(Photomicrograph courtesy of C. O'Hara; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A monoclonal serum protein by itself is not diagnostic of amyloidosis, since monoclonal gammopathy of uncertain significance is common in older patients (Chap. 107). However, when monoclonal gammopathy of uncertain significance is found in patients with biopsy-proven amyloidosis, the AL type should be ruled out. Similarly, patients thought to have "smoldering myeloma" because of a modest elevation of bone-marrow plasma cells should be screened for AL amyloidosis if they have signs or symptoms of renal, cardiac, or neurologic disease. Accurate tissue amyloid typing is essential for appropriate treatment. Immunohistochemical staining of the amyloid deposits is useful if they selectively bind one LC antibody in preference to the other; some AL deposits bind antibodies nonspecifically. Immunoelectron microscopy is more reliable, while mass spectrometry–based microsequencing of small amounts of protein extracted from fibril deposits has become the diagnostic standard. In ambiguous cases, other forms of amyloidosis should be thoroughly excluded with appropriate genetic and other testing.

Treatment

TREATMENT

AL Amyloidosis

Extensive multisystemic involvement typifies AL amyloidosis, and the median survival period without treatment is usually only ~1–2 years from the time of diagnosis. Current therapies target the clonal bone-

marrow plasma cells, using approaches employed for multiple myeloma. Treatment with oral melphalan and prednisone can decrease the plasma cell burden but rarely leads to complete hematologic remission, meaningful organ responses, or improved survival and is no longer widely used. The substitution of dexamethasone for prednisone produces a higher response rate and more durable remissions, although dexamethasone is not always well tolerated by patients with significant edema or cardiac disease. High-dose IV melphalan followed by autologous stem cell transplantation (HDM/SCT) produces complete hematologic responses in ~40% of treated patients, as determined by loss of clonal plasma cells in the bone marrow and disappearance of the amyloidogenic monoclonal LC, as determined by SIFE/UIFE and free LC quantitation. Six to 12 months after achieving a hematologic responses, improvements in organ function and quality of life may occur. Hematologic responses appear to be more durable after HDM/SCT than in multiple myeloma, with remissions continuing in some patients beyond 15 years without additional treatment. Unfortunately, only about 30–40% of all AL amyloidosis patients are suitable for aggressive treatment, and, even at specialized treatment centers, transplantation-related mortality rates are higher than those for other hematologic diseases because of impaired organ function at initial presentation. Amyloid cardiomyopathy, poor nutritional and performance status, and multiorgan disease contribute to excess morbidity and mortality. A bleeding diathesis resulting from adsorption of clotting factor X to amyloid fibrils also increases mortality rates; however, this syndrome occurs in only 5–10% of patients. A randomized multicenter trial conducted in France compared oral melphalan and dexamethasone with HDM/SCT and failed to show a benefit of dose-intensive treatment, although the transplantation-related mortality rate in this study was very high. It has become clear that careful selection of patients and expert peritransplantation management are essential in reducing transplantation-related mortality.

For patients with AL amyloidosis and impaired cardiac function or arrhythmias due to involvement of the myocardium, the median survival period is only ~6 months without treatment. In these patients, cardiac transplantation can be performed and followed by HDM/SCT to eliminate the noxious LC clone and prevent amyloid deposition in the transplanted heart or other organs.

Novel anti–plasma cell agents have been investigated for treatment of plasma cell diseases. The immunomodulators thalidomide, lenalidomide, and pomalidomide display activity; dosing may need to be adjusted compared to their usage for myeloma. The proteasome inhibitor bortezomib has also been found to be effective in single-center and multicenter trials. Anti-fibril small molecules and humanized monoclonal antibodies are also being tested. Clinical trials are essential in improving therapy for this rare disease.

Supportive care is important for patients with any type of amyloidosis. For nephrotic syndrome, diuretics and supportive stockings can ameliorate edema; angiotensin-converting enzyme inhibitors should be used with caution and have not been shown to slow renal disease progression. Effective diuresis can be facilitated with albumin infusions to raise intravascular oncotic pressure. Congestive heart failure due to amyloid cardiomyopathy is best treated with diuretics; it is important to note that digitalis, calcium channel blockers, and beta blockers are relatively contraindicated as they can interact with amyloid fibrils and produce heart block and worsening heart failure. Amiodarone has been used for atrial and ventricular arrhythmias. Automatic implantable defibrillators appear to have reduced effectiveness due to the thickened

myocardium, but they may benefit some patients. Atrial ablation is an effective approach for atrial fibrillation. For conduction abnormalities, ventricular pacing may be indicated. Atrial contractile dysfunction is common in amyloid cardiomyopathy and associated with increased thromboembolic complications, prompting considerations of anticoagulation even in the absence of atrial fibrillation. Autonomic neuropathy can be treated with α agonists such as midodrine to support postural blood pressure; gastrointestinal dysfunction may respond to motility or bulk agents. Nutritional supplementation, either oral or parenteral, is also important.

In localized AL amyloidosis, amyloid deposits can be produced by clonal plasma cells infiltrating local sites in the airways, bladder, skin, or lymph nodes (Table 108-1). These deposits may respond to surgical intervention or elimination of the responsible plasma cell clone by low-dose radiation therapy (typically only 20 Gy); systemic treatment generally is not appropriate. Patients should be referred to a center familiar with management of these rare manifestations of amyloidosis.

AA AMYLOIDOSIS

Etiology and Incidence

AA amyloidosis can occur in association with almost any chronic inflammatory state (e.g., rheumatoid arthritis, inflammatory bowel disease, familial Mediterranean fever **[Chap. 362]**, or other periodic fever syndromes) or chronic infections such as tuberculosis or subacute bacterial endocarditis. In the United States and Europe, AA amyloidosis has become less common, occurring in fewer than 2% of patients with these diseases, presumably because of advances in anti-inflammatory and antimicrobial therapies. It has also been described in association with Castleman's disease, lymphomas, renal cell carcinoma, emphasizing the diagnostic importance of CT scanning to look for such tumors as well as serologic and microbiologic studies. In up to 30% of patients, AA amyloidosis can also be seen without any identifiable underlying disease. AA is the most frequent systemic amyloidosis that occurs in children.

Pathology and Clinical Features

Organ involvement in AA amyloidosis usually begins in the kidneys. Hepatomegaly, splenomegaly, and autonomic neuropathy can also occur as the disease progresses; cardiomyopathy occurs in later disease. The symptoms and signs of AA disease cannot be reliably distinguished from those of AL amyloidosis. AA amyloid fibrils are usually composed of an 8-kDa, 76-amino-acid *N*-terminal portion of the 12-kDa precursor protein SAA. This acute-phase apoprotein is synthesized in the liver and transported by high-density lipoprotein (HDL3) in the plasma. Several years of an underlying inflammatory disease causing chronic elevation of SAA levels usually precede fibril formation, although infections can lead to AA deposition more rapidly.

Treatment

TREATMENT

AA Amyloidosis

Primary therapy for AA amyloidosis consists of treatment of the underlying inflammatory or infectious disease. Treatment that suppresses or eliminates the inflammatory state or infection decreases the SAA concentration and rate of amyloid fibril formation. For familial Mediterranean fever, colchicine at a dose of 1.2–1.8 mg/d is the standard treatment. However, colchicine has not been helpful for AA amyloidosis of other causes or for other amyloidoses. Tumor necrosis factor and interleukin 1 antagonists can be effective in syndromes related to cytokine elevation. Efforts to develop a fibril-specific agent (eprodisate) that interferes with the interaction of serum amyloid A protein and glycosaminoglycans to prevent or disrupt fibril formation failed in phase III trials.

ATTR AND AF AMYLOIDOSIS

 The familial amyloidoses are autosomal dominant diseases in which, beginning in midlife, a variant (FINE) plasma protein forms amyloid deposits. These diseases are rare, with an estimated incidence of <1 case/100,000 population in the United States, although founder effects in isolated areas of Portugal, Sweden, and Japan have led to a much higher incidence. The most common form of AF amyloidosis is ATTRm in the updated nomenclature, caused by mutation of the abundant plasma protein transthyretin (TTR, also known as *prealbumin*). More than 120 TTR mutations are known, and most are associated with ATTR amyloidosis. One variant, V122I, has a carrier frequency of nearly 4% in the African-American population and is associated with late-onset cardiac amyloidosis. The actual incidence and penetrance of disease in the African-American population is the subject of ongoing research, but ATTR amyloidosis warrants consideration in the differential diagnosis of African-American patients who present with concentric cardiac hypertrophy and evidence of diastolic heart failure, particularly in the absence of a history of hypertension. Other familial amyloidoses, caused by variant apolipoproteins AI or AII, gelsolin, fibrinogen Aα, or lysozyme, are reported in only a few families worldwide. New amyloidogenic serum proteins continue to be identified periodically, including recently the leukocyte chemotactic factor LECT2, a cause of renal amyloidosis in Hispanic and Pakistani populations. To date, no mutation in the coding sequence for the LECT2 gene has been identified, so the heritability of ALECT2 remains uncertain.

Amyloid deposits composed of unmutated precursor protein (wild type TTR) occur with aging. Due to a rapidly aging world population, wild-type ATTR (ATTRwt) is being diagnosed with increasing frequency in Caucasian men >65 years of age with amyloid cardiomyopathy. Formerly termed senile systemic amyloidosis, ATTRwt has been found at autopsy in 25% of hearts from men older than age 80 years. Why a wild type protein becomes amyloidogenic, and why patients bearing mutant TTR genes do not express disease until adulthood, remains a mystery.

Clinical Features and Diagnosis

AF amyloidosis has a presentation that is variable but is usually consistent within kindreds affected by the same mutant protein. A family history makes AF disease more likely, but many patients present with

previously unrecognized familial disease or sporadically with new mutations. ATTRm amyloidosis typically presents as a syndrome of familial amyloidotic polyneuropathy or familial amyloidotic cardiomyopathy. Peripheral neuropathy begins as a small-fiber length-dependent lower-extremity sensorimotor neuropathy and progresses to the upper extremities. Autonomic neuropathy manifests as diarrhea with weight loss and orthostatic hypotension. Patients with ATTR V30M, the most common mutation, often develop conduction defects late in the disease. Patients with ATTR T60A and several other mutations have myocardial thickening similar to that caused by AL amyloidosis, although heart failure is less common and long-term untreated survival rates are usually better. Vitreous opacities caused by amyloid deposits are pathognomonic for ATTR amyloidosis.

Typical syndromes associated with other forms of AF disease include renal amyloidosis with mutant fibrinogen, lysozyme, or apolipoproteins; hepatic amyloidosis with apolipoprotein AI; and amyloidosis of cranial nerves and cornea with gelsolin. Patients with AF amyloidosis can present with clinical syndromes that mimic those of patients with AL disease. Rarely, AF carriers can develop AL disease or AF patients may have monoclonal gammopathy without AL. Thus, it is important to screen both for plasma cell disorders and for mutations in patients with amyloidosis. Variant TTRs can usually be detected by isoelectric focusing, but DNA sequencing is now standard for diagnosis of ATTRm and other AF mutations.

Treatment

TREATMENT

ATTR Amyloidosis

Without intervention, the survival period after onset of ATTR disease is 5–15 years. Orthotopic liver transplantation replaces the major source of variant TTR production with one producing normal TTR. While liver transplantation can slow disease progression and improve chances of survival, it does not reverse sensorimotor neuropathy. Liver transplants are most successful in young patients with early peripheral neuropathy; older patients with familial amyloidotic cardiomyopathy or advanced polyneuropathy often experience end-organ disease progression despite successful liver transplantation. Post-transplant ATTR disease progression results from deposition of wild-type TTR onto pre-existing fibrillar deposits principally composed of mutant TTR.

The rate-limiting step in ATTR amyloidosis is dissociation of the TTR tetramer into monomeric protein followed by misfolding, oligomerization, and fibril aggregation. TTR tetramers can be stabilized by thyroxine binding or by thyroxine-mimetic small molecules such as the non-steroidal anti-inflammatory drug diflunisal or the rationally designed small-molecule therapeutic tafamidis. A placebo-controlled randomized trial of diflunisal demonstrated a significant reduction in the progression of polyneuropathy and maintenance of quality of life in patients with a wide variety of ATTR mutations who received the “repurposed” diflunisal. Tafamidis tested in a similar fashion in patients with the V30M ATTR mutation failed to meet its primary endpoints but was approved for marketing by the European Medicines Agency since most secondary endpoints favored the drug. These agents are now being investigated for effects on ATTR cardiomyopathy. In

vitro data and serendipitous observations in patients suggest that ATTRm disease can be ameliorated by "trans-suppression," in which a T119M TTR variant stabilizes tetramers that also contain amyloidogenic subunits. Interestingly, in a large population study in Denmark, 0.5% of participants were heterozygous for the T119M allele, and this small group had higher levels of TTR in their blood, a reduced incidence of cerebrovascular disease, and a 5- to 10-year survival advantage compared with participants lacking this allele. The newest approach to controlling ATTR disease, TTR gene silencing, involves nearly complete suppression of TTR production by the liver, effectively preventing amyloid fibril formation by eliminating the precursor protein. TTR gene silencers including RNA interference and anti-sense oligonucleotide agents directed against TTR RNA are in phase III clinical trials to determine their effect on progression of ATTR polyneuropathy and cardiomyopathy.

Aβ_2M AMYLOIDOSIS

Aβ_2M amyloid is composed of β_2-microglobulin, the invariant chain of class I human leukocyte antigens, and produces rheumatologic manifestations in patients undergoing long-term hemodialysis. β_2-Microglobulin is excreted by the kidney, and levels become elevated in end-stage renal disease. The molecular mass of β_2M is 11.8 kDa—above the cutoff of some dialysis membranes. The incidence of this disease appears to be declining with the use of newer membranes in high-flow dialysis techniques. Aβ_2M amyloidosis usually presents as carpal tunnel syndrome, persistent joint effusions, spondyloarthropathy, or cystic bone lesions. Carpal tunnel syndrome is often the first symptom. In the past, persistent joint effusions accompanied by mild discomfort were found in up to 50% of patients who had undergone dialysis for >12 years. Involvement is bilateral, and large joints (shoulders, knees, wrists, and hips) are most frequently affected. The synovial fluid is noninflammatory, and β_2M amyloid can be found if the sediment is stained with Congo red. Although less common, visceral β_2M amyloid deposits do occasionally occur in the gastrointestinal tract, heart, tendons, and subcutaneous tissues of the buttocks. There is no specific therapy for Aβ_2M amyloidosis, but cessation of dialysis after renal allografting may lead to symptomatic improvement.

THERAPEUTIC FRONTIERS

To date, treatment strategies have focused on limiting formation of amyloidogenic proteins. Disruption of existing amyloid by targeting ubiquitous components of the tissue deposits may offer means to improving major end organ function. Two antibody approaches—one involving an epitope exposed during immunoglobulin light chain misfolding (AL disease) and the other directed against serum amyloid P component—are undergoing evaluation in early phase III studies.

SUMMARY

A diagnosis of amyloidosis should be considered in patients with unexplained nephropathy, cardiomyopathy (particularly with diastolic dysfunction), neuropathy (either peripheral or autonomic), enteropathy, or the

pathognomonic soft tissue findings of macroglossia or periorbital ecchymoses. Pathologic identification of amyloid fibrils can be made with Congo red staining of aspirated abdominal fat or of an involved-organ biopsy specimen. Accurate typing by a combination of immunologic, biochemical, and genetic testing is essential in selecting appropriate therapy (Fig. 108-1). Systemic amyloidosis should not be considered an untreatable condition, as anti–plasma cell chemotherapy is highly effective in AL disease and targeted therapies are being developed for AA and ATTR disease. The combination of precursor and end organ amyloid therapeutics may permit not only disease control but also functional improvement to patients with amyloidosis. Tertiary referral centers can provide specialized diagnostic techniques and access to clinical trials for patients with these rare diseases.

ACKNOWLEDGMENT

This chapter represents a revised version of chapters co-authored by Drs. Martha Skinner, David C. Seldin, and John L. Berk in previous editions of Harrison's Principles of Internal Medicine.

FURTHER READING

Connors LH et al: Heart failure resulting from age-related cardiac amyloid disease associated with wild-type transthyretin: A prospective, observational cohort study. Circulation 133:282, 2016.

Hawkins PN et al: Evolving landscape in the management of transthyretin amyloidosis. Ann Med 47:625, 2015.

Richards DB et al: Therapeutic clearance of amyloid by antibodies to serum amyloid P component. N Engl J Med 373:1106, 2015.

Sanchorawala V et al: Long-term outcome of patients with AL amyloidosis treated with high-dose melphalan and stem cell transplantation: 20-year experience. Blood 126:2345, 2015.

Wechalekar AD et al: Systemic amyloidosis. Lancet 387:2641, 2016.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 109: Transfusion Biology and Therapy

Jeffery S. Dzieczkowski; Pierre Tiberghien; Kenneth C. Anderson

FIGURE 109-1

BLOOD GROUP ANTIGENS AND ANTIBODIES

Antigens systems important in transfusion medicine comprise red blood cell (RBC), platelet, neutrophil, and the widely distributed human leukocytes (HLA) antigens.

The study of RBC antigens and antibodies forms the foundation of transfusion medicine. Serologic studies initially characterized these antigens, but now the molecular composition and structure of many are known. Antigens, either carbohydrate or protein, are assigned to a blood group system based on the structure and similarity of the determinant epitopes. Other cellular blood elements such as platelets and plasma proteins are also antigenic and can result in alloimmunization, the production of antibodies directed against antigenic determinants of another individual. These antibodies, called alloantibodies, can comprise anti-RBC Abs, anti-human platelet antigens (HPA) Abs, as well as anti-human leukocytes antigens (HLA) Ab.

Antibodies directed against RBC antigens may result from “natural” exposure, particularly to carbohydrates that mimic some blood group antigens that are present in the environment, particularly saprophyte bacteria. Those antibodies that occur via natural stimuli are usually produced by a T cell–independent response (thus, generating no immune memory) and are mainly IgM isotype. *Autoantibodies* (antibodies against autologous blood group antigens) arise spontaneously or as the result of infectious sequelae (e.g., from *Mycoplasma pneumoniae*) and are also often IgM. These antibodies are often clinically insignificant due to their low affinity for antigen at body temperature. However, IgM antibodies can activate the complement cascade and result in hemolysis. Autoantibodies can also arise in an autoimmune setting with most often an IgG isotype. Antibodies that result from allogeneic exposure, such as transfusion or pregnancy, are usually IgG. IgG antibodies commonly bind to antigen at warmer temperatures and may hemolyze RBCs. Unlike IgM antibodies, IgG antibodies can cross the placenta and bind fetal erythrocytes bearing the corresponding antigen, resulting in hemolytic disease of the newborn, or *hydrops fetalis*. The same holds true for IgG directed against HPA antigens on platelets that can lead to fetal or neonatal immunization and result in intracranial hemorrhage.

Recipient alloimmunization to leukocytes, platelets, and plasma proteins may also result in transfusion complications such as fevers and urticaria as well as platelet transfusion refractoriness, but generally does not cause hemolysis. Such an alloimmunization in the blood donor may also result in a severe lung disorder called transfusion-related acute lung injury (TRALI). Assay for these non-hemolytic alloantibodies is not routinely performed; however, they may be detected using special assays.

ABO ANTIGENS AND ANTIBODIES

The first blood group antigen system, recognized in 1900, was ABO, the most important in transfusion medicine. The major blood groups of this system are A, B, AB, and O. O type RBCs lack A or B antigens. These antigens are carbohydrates attached to a precursor backbone, may be found on the cellular membrane either as glycosphingolipids or glycoproteins, and are secreted into plasma and body fluids as glycoproteins. H substance is the immediate precursor on which the A and B antigens are added. This H substance is formed by the addition of fucose to the glycolipid or glycoprotein backbone. The subsequent addition of *N*-acetylgalactosamine creates the A antigen, while the addition of galactose produces the B antigen.

The genes that determine the A and B phenotypes are found on chromosome 9p and are expressed in a Mendelian codominant manner. The gene products are glycosyl transferases, which confer the enzymatic capability of attaching the specific antigenic carbohydrate. Individuals who lack the “A” and “B” transferases are phenotypically type “O,” while those who inherit both transferases are type “AB.” Rare individuals lack the H gene, which codes for fucose transferase, and cannot form H substance. These individuals are homozygous for the silent h allele (hh) and have Bombay phenotype (O_h).

The ABO blood group system is important because essentially all individuals produce antibodies to the ABH carbohydrate antigen that they lack. The “naturally” occurring anti-A and anti-B antibodies are termed *isoagglutinins*. Thus, type A individuals produce anti-B, while type B individuals make anti-A. Neither isoagglutinin is found in type AB individuals, while type O individuals produce both anti-A and anti-B. Thus, persons with type AB are considered “universal recipients” with regard to RBC transfusion because they do not have antibodies against any ABO phenotype, while persons with type O blood can donate to essentially all recipients because their cells are not recognized by any ABO isoagglutinins. The rare individuals with Bombay phenotype produce antibodies to H substance (which is present on all red cells except those of *hh* phenotype) as well as to both A and B antigens and are therefore compatible only with other *hh* donors.

In most people (80%), A, B, and H antigens are secreted by the cells and are present in the circulation as well as in various secretions such as saliva (*Se* phenotype). Others, called “nonsecretors” do not secrete A, B, and H antigens (*se* phenotype). ABO and Se/se systems influence the susceptibility to a variety of diseases. For example, malaria is less severe in O than non O patients. Conversely, group O is associated with enhanced susceptibility to *Helicobacter pylori* (and gastric ulcer) as well as to cholera bacillus or to norovirus. Furthermore, group O individuals exhibit a lesser procoagulation phenotype when compared to non O individuals.

RH SYSTEM

The Rh system is the second most important blood group system in pretransfusion testing. The Rh antigens are found on a 30- to 32-kDa RBC membrane protein that has no defined function. Although >40 different antigens in the Rh system have been described, five determinants account for the vast majority of phenotypes. The two RH genes are located on chromosome 1. The RHD gene codes for the RhD protein. The RHCE gene codes for RhCE proteins expressing C and/or c, and E and/or e antigens. The presence of the D antigen confers Rh “positivity,” while persons who lack the D antigen are Rh negative. Two allelic antigen pairs, E/e and C/c, are also found on the Rh protein. These two Rh genes, RHD and RHCE, determine eight main haplotypes (DCe, DcE, Dce, DCE, dce, dCe, dcE,

and dCE) whose frequencies differ among different populations. The high diversity of the RH antigens includes weak or partial expression. Identifying individuals (especially young female of childbearing potential and multi-transfused patients) with a weak or RhD antigen is important to adequately select RhD positive or negative RBC. Molecular biology is now routinely applied to resolve such situations.

The D antigen is a potent alloantigen. About 15% of individuals lack this antigen. Exposure of these Rh-negative people to even small amounts of Rh-positive cells, by either transfusion or pregnancy, can result in the production of anti-D alloantibody.

OTHER BLOOD GROUP SYSTEMS AND ALLOANTIBODIES

Thirty-six RBC group systems, six collections (related set of antigens not sufficiently distinct from existing systems to qualify as systems) and two series (low frequency [“private”] and high frequency [“public”] individual antigens) are presently recognized, composed of more than 350 antigens. The presence or absence of certain antigens has been associated with various diseases and anomalies; antigens also act as receptors for infectious agents. Alloantibodies of importance in routine clinical practice are listed in Table 109-1.

TABLE 109-1

RBC Blood Group Systems and Alloantigens

Blood Group System	Antigen	Alloantibody	Clinical Significance
Rh (D, C/c, E/e)	RBC protein	IgG	HTR, HDN
Lewis (Le^a, Le^b)	Oligosaccharide	IgM/IgG	Rare HTR
Kell (K/k)	RBC protein	IgG	HTR, HDN
Duffy (Fy^a/Fy^b)	RBC protein	IgG	HTR, HDN
Kidd (Jk^a/Jk^b)	RBC protein	IgG	HTR (often delayed), HDN (mild)
I	Carbohydrate	IgM	None
MNS	RBC protein	IgM/IgG	Anti-M rare HDN, anti-S, -s, and -U HDN, HTR

Abbreviations: HDN, hemolytic disease of the newborn; HTR, hemolytic transfusion reaction; RBC, red blood cell.

Antibodies to *Lewis system* carbohydrate antigens are the most common cause of incompatibility during pretransfusion screening. The Lewis gene product is a fucosyl transferase and maps to chromosome 19. The antigen is not an integral membrane structure but is adsorbed to the RBC membrane from the plasma. Antibodies to Lewis antigens are usually IgM and cannot cross the placenta. Lewis antigens may be adsorbed onto tumor cells and may be targets of therapy.

I system antigens are also oligosaccharides related to H, A, B, and Le. I and i are not allelic pairs but are carbohydrate antigens that differ only in the extent of branching. The i antigen is an unbranched chain that is converted by the I gene product, a glycosyltransferase, into a branched chain. The branching process affects all the ABH antigens, which become progressively more branched in the first 2 years of life. Some patients with cold agglutinin disease or lymphomas can produce anti-I autoantibodies that cause RBC destruction. Occasional patients with mononucleosis or *Mycoplasma* pneumonia may develop cold agglutinins of either anti-I or anti-i specificity. Most adults lack i expression; thus, finding a donor for patients with anti-i is not difficult. Even though most adults express I antigen, binding is generally low at body temperature. Thus, administration of warm blood prevents isoagglutination.

The *P system* is another group of carbohydrate antigens controlled by specific glycosyltransferases. The very rare individuals lacking the P antigen produce an anti-P antibody. Finding a compatible donor for such individuals is difficult. P antigen is also the target of auto-antibodies in diseases such as syphilis and viral diseases in children, and can result in paroxysmal cold hemoglobinuria. In the latter case, an autoantibody to P binds to RBCs in the cold and fixes complement upon warming. Antibodies with these biphasic properties are called *Donath-Landsteiner antibodies*. The P antigen is the cellular receptor of parvovirus B19 and also may be a receptor for *Escherichia coli* binding to urothelial cells.

The *MNS system* is regulated by genes on chromosome 4. M and N are determinants on glycophorin A, an RBC membrane protein, and S and s are determinants on glycophorin B. Anti-S and anti-s IgG antibodies may develop after pregnancy or transfusion and lead to hemolysis. Glycophorin B expresses a public antigen named U. Anti-U antibodies can occur in the rare individuals lacking the U antigen. Such occurrence is problematic; virtually every donor is incompatible because nearly all persons express U.

The *Kell* protein is very large (720 amino acids), and its secondary structure contains many different antigenic epitopes. The immunogenicity of Kell is third behind the ABO and Rh systems. The Kell protein is linked to another blood group protein termed Kx. The rare absence of this protein (controlled by a gene on X) is associated with weak KEL antigen, acanthocytosis, shortened RBC survival, and a progressive form of muscular dystrophy that includes cardiac defects. This rare condition is called the *McLeod phenotype*. The K_x gene is linked to the 91-kDa component of the NADPH-oxidase on the X chromosome, deletion or mutation of which accounts for about 60% of cases of chronic granulomatous disease. Such individuals can produce an anti-Kx antibody that makes finding a compatible blood product difficult.

The *Duffy* antigens are codominant alleles, Fy^a and Fy^b, that also serve as receptors for *Plasmodium vivax*. More than 70% of persons in sub-Saharan Africa lack these antigens, probably from selective influences of malaria infection on the population. For unknown reasons, the lack of the Duffy antigen receptor for cytokines (DARC) is associated with mild neutropenia.

The *Kidd* antigens, Jk^a and Jk^b, may elicit antibodies transiently. A delayed hemolytic transfusion reaction (DHTR) that occurs with blood tested as compatible is often related to delayed appearance of anti-Jk^a.

PRETRANSFUSION TESTING

Pretransfusion testing of a potential recipient consists of the “type and screen.” The “forward type” determines the ABO and Rh phenotype of the recipient’s RBC by using antisera directed against the A, B, and D antigens. The “reverse type” detects isoagglutinins in the patient’s serum and should correlate with the ABO phenotype, or forward type. Additional typing for other main Rh antigens (CcEe), the K antigen, and more rarely Duffy, Kidd and Ss antigens, can be required depending on the clinical setting. Molecular typing is increasingly used to predict the RBC phenotype and facilitate the selection of a compatible blood component.

The alloantibody screen identifies antibodies directed against other RBC antigens. The alloantibody screen is performed by mixing patient serum with type O RBCs that contain the major antigens of most blood group systems and whose extended phenotype is known. The specificity of the alloantibody is identified by correlating the presence or absence of antigen with the results of the agglutination. Special attention should be given to patients receiving monoclonal antibody treatment that may bind to RBC (such as anti-CD38 treatment for myeloma) and therefore interfere with alloantibody screening.

Cross-matching is ordered when there is a high probability that the patient will require a packed RBC (PRBC) transfusion. In the setting of a systematic alloantibody screen, such a crossmatch can be restricted to alloimmunized patients as well as patients at high risk of alloimmunization (prior pregnancies, transfusions). Blood selected for cross-matching must be ABO compatible and lack antigens for which the patient has alloantibodies. Nonreactive cross-matching confirms the absence of any major incompatibility and reserves that unit for the patient.

In the case of Rh (D) -negative patients, every attempt must be made to provide Rh-negative blood components to prevent alloimmunization to the D antigen. In an emergency, Rh-positive blood can be safely transfused to an Rh-negative patient who lacks anti-D; however, the recipient is likely to become alloimmunized and produce anti-D. Rh-negative women of childbearing age who are transfused with products containing Rh-positive RBCs should receive passive immunization with anti-D (RhoGam or WinRho) to reduce or prevent sensitization.

BLOOD COMPONENTS

Blood products intended for transfusion are routinely collected as whole blood (450 mL) in various anticoagulants. Most donated blood is processed into components: PRBCs, platelets, and fresh-frozen plasma (FFP) or cryoprecipitate (**Table 109-2**). Whole blood can be first separated into PRBCs and platelet-rich plasma by slow centrifugation. The platelet-rich plasma is then centrifuged at high speed to yield one unit of random donor (RD) platelets (subsequently pooled) and one unit of FFP. Alternatively, whole blood can undergo high speed centrifugation to separate a PRBC, a FFP and a “buffy-coat” containing leukocytes and platelets. The buffy coat then undergoes pooling and is centrifuged at low speed to produce pooled platelets. The leukocyte level of blood products can be lowered by an additional filtration step after which they are referred to as leukoreduced (or leukodepleted) (<1 to 5×10^6 leukocytes per product). Cryoprecipitate is produced by thawing FFP to precipitate the plasma proteins, then separated by centrifugation.

TABLE 109-2

Characteristics of Selected Blood Components

Component	Volume, mL	Content	Clinical Response
PRBC	250–300 (including additive solution)	RBCs with variable leukocyte content and small amount of plasma	Increase hemoglobin 10 g/L and hematocrit 3%
Platelets (from whole blood)	50–70 /RD unit pool of 4 to 6 RD unit	5.5×10^{10}/RD unit	Increase platelet count 5000–10,000/μL/RD unit
Platelets (from apheresis)	200–400	$\geq 2 \times 10^{11}$/SDAP product	For pooled RD and SDAP: CCI ≥10 × 10^9/L within 1 h and ≥7.5 × 10^9/L within 24 h posttransfusion
FFP	200–250	Plasma proteins—coagulation factors, proteins C and S, antithrombin	Increases coagulation factors about 2%
Cryoprecipitate (may be used as topical fibrin glue as well)	10–15 /unit, pool of 4 to 5 units	Cold-insoluble plasma proteins, fibrinogen, factor VIII, factor XIII, vWF fibronectin	Increased plasma fibrinogen 0,3-1 g/L, increased factor VIII

Abbreviations: CCI, corrected count increment; FFP, fresh-frozen plasma; PRBC, packed red blood cells; RBC, red blood cell; RD, random donor; SDAP, single-donor apheresis platelets; vWF, von Willebrand factor.

Apheresis technology is used for the collection of multiple units of platelets from a single donor. These single-donor apheresis platelets (SDAP) contain the equivalent of at least five units of RD platelets and before eventual leukoreduced, have fewer contaminating leukocytes than pooled RD platelets.

Plasma as well as RBCs and granulocytes may also be collected by apheresis. Plasma-derived products such as albumin, intravenous immunoglobulin, antithrombin, and coagulation factor concentrates are prepared from pooled plasma from many donors. Plasma fractionation includes steps that eliminate infectious agents.

WHOLE BLOOD

Whole blood provides both oxygen-carrying capacity and volume expansion. It is the ideal component for patients who have sustained acute hemorrhage of ≥25% total blood volume loss. Whole blood is stored at 4°C to maintain erythrocyte viability, but platelet dysfunction and degradation of some coagulation factors occur. In addition, 2,3-

bisphosphoglycerate levels fall over time, leading to an increase in the oxygen affinity of the hemoglobin and a decreased capacity to deliver oxygen to the tissues, a problem with all red cell storage. Fresh whole blood avoids these problems, but it is typically used only in emergency settings (i.e., military). Whole blood is not readily available, since it is routinely processed into components.

PACKED RBCS

This product increases oxygen-carrying capacity in the anemic patient. PRBC are stored in additive solution up to 35–42 days at 4°C. Adequate oxygenation can be maintained with a hemoglobin content of 70 g/L in the normovolemic patient without cardiac disease; however, comorbid factors may necessitate transfusion at a higher threshold. The decision to transfuse should be guided by the clinical situation and not by an arbitrary laboratory value. In the critical care setting, liberal use of transfusions to maintain near-normal levels of hemoglobin has not proven advantageous. In most patients requiring transfusion, levels of hemoglobin of 80 g/L are sufficient to keep oxygen supply from being critically low.

PRBCs may be modified to prevent certain adverse reactions. The majority of cellular blood products are now leukocyte-reduced and universal prestorage leukocyte reduction has been recommended. Prestorage filtration appears superior to bedside filtration as smaller amounts of cytokines are generated in the stored product. These PRBC units contain <1 to 5.10^6 donor leukocytes, and their use lowers the incidence of posttransfusion fever and chills, cytomegalovirus (CMV) infections, and alloimmunization. Other theoretical benefits include less immunosuppression in the recipient and lower risk of infections with intracellular pathogens (in addition to CMV). Plasma, which may cause allergic reactions, can be removed from cellular blood components by washing. Patients with hemoglobinopathies such as sickle cell disease may undergo RBC exchange, an apheresis procedure by which sickled RBCs are replaced by donor RBCs.

PLATELETS

Thrombocytopenia is a risk factor for hemorrhage, and platelet transfusion reduces the incidence of bleeding. Platelets are stored in plasma or in additive solution up to 5–7 days at 20–24°C and under permanent motion. The threshold for prophylactic platelet transfusion is 10,000/µL. In patients without fever or infections, a threshold of 5000/µL may be sufficient to prevent spontaneous hemorrhage. For invasive procedures, 50,000/µL platelets is the usual target level.

Platelets are given either as pools of 4 to 6 prepared RDs or as SDAPs from a single donor. In an unsensitized patient without increased platelet consumption (splenomegaly, fever, disseminated intravascular coagulation [DIC]), two units of transfused RD per square-meter body surface area (BSA) is anticipated to increase the platelet count by ~10,000/uL. Patients who have received multiple transfusions may be alloimmunized to many HLA- and platelet-specific antigens and have little or no increase in their posttransfusion platelet counts. Patients who may require multiple transfusions are best served by receiving leukocyte-reduced components to lower the risk of alloimmunization.

Refractoriness to platelet transfusion may be evaluated using the corrected count increment (CCI):

$$\text{CCI} = \frac{\text{posttransfusion count}(/\mu\text{L}) - \text{pretransfusion count}(/\mu\text{L})}{\text{number of platelets transfused} \times 10^{-11}} \times \text{BSA}\ (\text{m}^2)$$

where BSA is body surface area measured in square meters. The platelet count performed 1 h after the transfusion is acceptable if the CCI is 10×10^9/mL, and after 18–24 h an increment of 7.5×10^9/mL is expected. Patients who have suboptimal responses are likely to have received multiple transfusions and have antibodies directed against class I HLA antigens. Refractoriness can be investigated by detecting anti-HLA antibodies in the recipient's serum. Patients who are sensitized will often react with 100% of the lymphocytes used for the HLA-antibody screen, and HLA-matched SDAPs should be considered for those patients who require transfusion. Although ABO-identical HLA-matched SDAPs provide the best chance for increasing the platelet count, locating these products is difficult. Platelet cross-matching is available in some centers. Additional clinical causes for a low platelet CCI include fever, bleeding, splenomegaly, DIC, or medications in the recipient.

FRESH-FROZEN PLASMA

FFP contains stable coagulation factors and plasma proteins: fibrinogen, antithrombin, albumin, as well as proteins C and S. Indications for FFP include correction of coagulopathies, including the rapid reversal of warfarin; supplying deficient plasma proteins; and treatment of auto-antibody-mediated thrombotic thrombocytopenic purpura (TTP). In the latter case, therapeutic plasma exchange allows both the removal of the autoantibody and the supplementation of the depleted enzyme (ADAMTS13). Other auto-immune diseases such as Guillain-Barré syndrome and myasthenia gravis may benefit from plasma exchange. FFP should not be routinely used to expand blood volume. FFP is an acellular component and does not transmit intracellular infections, e.g., CMV. In addition to FFP, pre-thawed or never frozen plasma as well as freeze-dried plasma are increasingly used to insure immediate availability when required. Patients who are IgA-deficient and require plasma support should receive FFP from IgA-deficient donors to prevent anaphylaxis (see below).

CRYOPRECIPITATE

Cryoprecipitate is a source of fibrinogen, factor VIII, and von Willebrand factor (vWF). It is ideal for supplying fibrinogen to the volume-sensitive patient. When factor VIII concentrates are not available, cryoprecipitate may be used since each unit contains ~80 units of factor VIII. Cryoprecipitate may also supply vWF to patients with dysfunctional (type II) or absent (type III) von Willebrand disease.

PLASMA DERIVATIVES

Plasma from thousands of donors may be pooled to derive specific protein concentrates, including albumin, intravenous immunoglobulin, antithrombin, and coagulation factors. In addition, donors who have high-titer antibodies to specific agents or antigens provide hyperimmune globulins, such as anti-D (RhoGam, WinRho), and antisera to hepatitis B virus (HBV), varicella-zoster virus, CMV, and other infectious agents.

ADVERSE REACTIONS TO BLOOD TRANSFUSION

Adverse reactions to transfused blood components occur despite multiple tests, inspections, and checks. Fortunately, the most common reactions are not life threatening, although serious reactions can present with mild symptoms and signs. Some reactions can be reduced or prevented by modified (filtered, washed, or irradiated) blood components. Blood product pathogen reduction is an option for platelets and plasma, and underway for whole blood and PRBC. When an adverse reaction is suspected, the transfusion should be stopped and reported to the blood bank for investigation.

Transfusion reactions may result from immune and nonimmune mechanisms. Immune-mediated reactions are often due to preformed donor or recipient antibody; however, cellular elements may also cause adverse effects. Nonimmune causes of reactions are due to the chemical and physical properties of the stored blood component and its additives.

Transfusion-transmitted viral infections are increasingly rare due to improved screening and testing. As the risk of viral infection is reduced, the relative risk of other reactions increases, such as hemolytic transfusion reactions and sepsis from bacterially contaminated components. However, one must remain concerned by novel or previously unidentified viral risks. Pretransfusion quality assurance improvements further increase the safety of transfusion therapy. Infections, like any adverse transfusion reaction, must be brought to the attention of the blood bank for appropriate actions (Table 109-3).

TABLE 109-3

Risks of Transfusion Complications

Frequency, Episodes: Unit	
Main Reactions	
Circulatory overload (TACO)	• 10.9:100,000
Febrile (FNHTR)	• 100–1000:100,000, frequently under reported
Allergic	• 100–400:100,000, product-dependent
TRALI	• 0,4–1:100,000, with mitigation, product-dependent
Delayed hemolytic	• 40:100,000
Acute hemolytic	• 2.5–7.9:100,000
Infections[a]	
Bacteria (septic transfusion reaction)	• 0,3–25:1,000,000 (product and detection- or pathogen-reduction-dependent
Hepatitis B	• 1:300,000 (<1:1,000,000*)
Hepatitis C*	• <0,1–1:1,000,000
HIV-1*, -2	• 0,1–1:1,000,000
HTLV-I and -II	• 1:3,000,000
Malaria	• 1:4,000,000
*: with nucleic acid testing (NAT) screening	
Other Complications	
RBC allosensitization	• 1:100
HLA allosensitization	• 1:10 (in the absence of leukodepletion)
Graft-versus-host disease	Extremely rare (with blood product irradiation in immunosuppressed patients)

[a]Other infectious agents associated with transfusion include arborvirus (West Nile virus, Dengue virus, Zika virus) hepatitis A and E virus, parvovirus B-19, *Babesia microti* and *Babesia duncani* (babesiosis), *Anaplasma phagocytophilum* (human granulocytic ehrlichiosis), *Trypanosoma cruzi* (Chagas disease), *Treponema pallidum*, and human herpesvirus-8. Frequency of infectious risks differ significantly world-wide.

Abbreviations: FNHTR, febrile nonhemolytic transfusion reaction; TRALI, transfusion-related acute lung injury; HTLV, human T lymphotropic virus; RBC, red blood cell; TACO, transfusion-associated circulatory overload.

IMMUNE-MEDIATED REACTIONS

Acute Hemolytic Transfusion Reactions

Immune-mediated hemolysis occurs when the recipient has preformed antibodies that lyse donor erythrocytes. The anti-A or anti-B antibodies are responsible for the majority of these reactions. However, alloantibodies directed against other RBC antigens, i.e., Rh, Kell, and Duffy, are responsible for fatal hemolytic transfusion reactions as well.

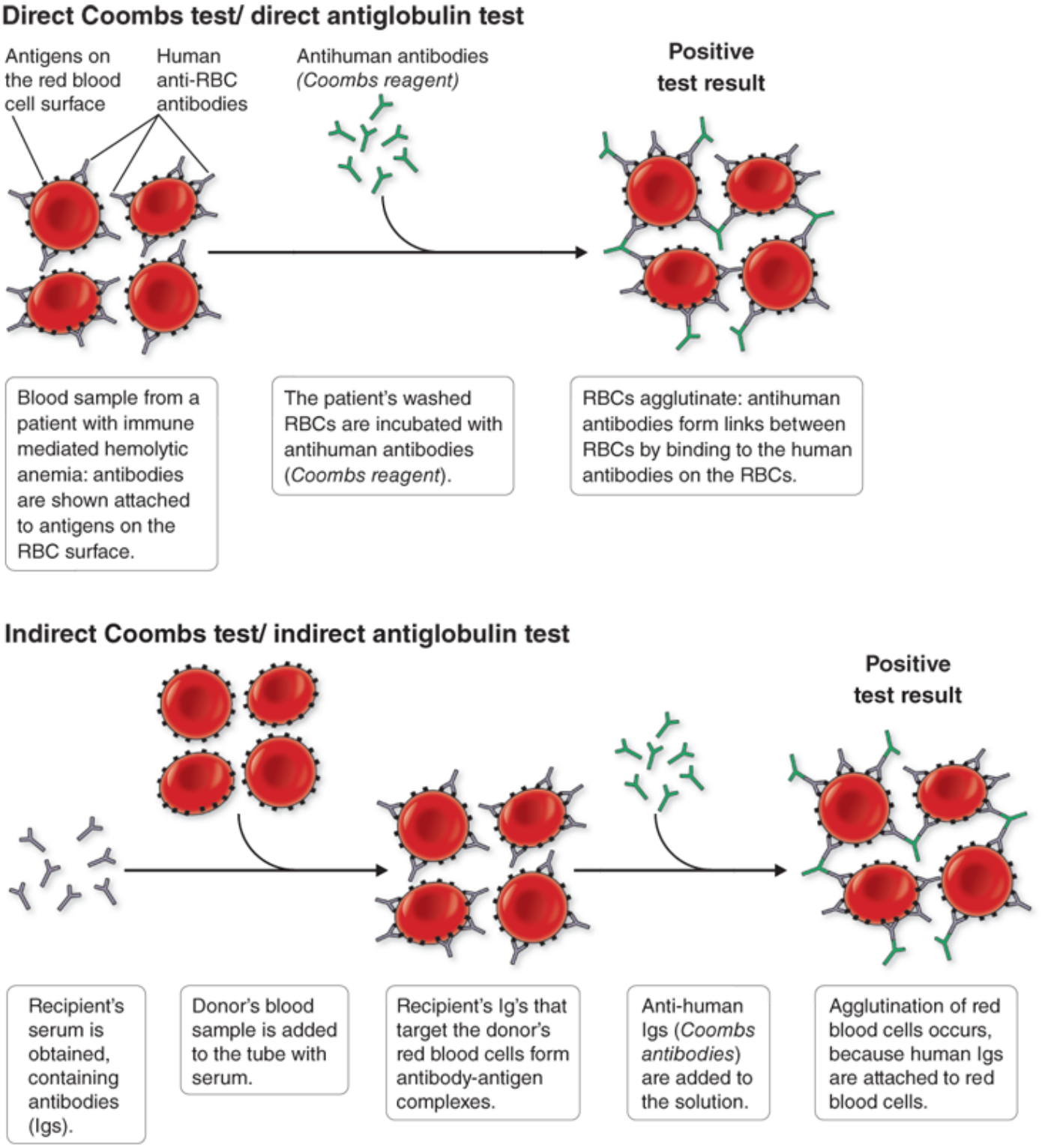
Acute hemolytic reactions may present with hypotension, tachypnea, tachycardia, fever, chills, hemoglobinemia, hemoglobinuria, chest and/or flank pain, and discomfort at the infusion site. Monitoring the patient's vital signs before and during the transfusion is important to identify reactions promptly. When acute hemolysis is suspected, the transfusion must be stopped immediately, intravenous access maintained, and the reaction reported to the blood bank. A correctly labeled posttransfusion blood sample and any untransfused blood should be sent to the blood bank for analysis. The laboratory evaluation for hemolysis includes the measurement of serum haptoglobin, lactate dehydrogenase (LDH), and indirect bilirubin levels.

The immune complexes that result in RBC lysis can cause renal dysfunction and failure. Diuresis should be induced with intravenous fluids and furosemide or mannitol. Tissue factor released from the lysed erythrocytes may initiate DIC. Coagulation studies including prothrombin time (PT), activated partial thromboplastin time (aPTT), fibrinogen, and platelet count should be monitored in patients with hemolytic reactions.

Errors at the patient's bedside, such as mislabeling the sample or transfusing the wrong patient, are responsible for the majority of these reactions. The blood bank investigation of these reactions includes examination of the pre- and posttransfusion samples for hemolysis and repeat typing of the patient samples; direct antiglobulin test (DAT), sometimes called the *direct Coombs test*, of the posttransfusion sample; repeating the cross-matching of the blood component; and checking all clerical records for errors. DAT detects the presence of antibody or complement bound to RBCs in vivo **(Fig. 109-1)**.

FIGURE 109-1

Direct and indirect Coombs test. The direct Coombs (antiglobulin) test detects the presence of antibodies (or complement) on the surface of erythrocytes. The indirect Coombs (antiglobulin) test detects antibodies in the serum that may bind to donor erythrocytes. RBC, red blood cell. *(Adapted from https://commons.wikimedia.org/wiki/File:Coombs_test_schematic.png.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Delayed Hemolytic and Serologic Transfusion Reactions

DHTRs are not completely preventable. These reactions occur in patients previously sensitized to RBC alloantigens who have a negative alloantibody screen due to low antibody levels. When the patient is transfused with antigen-positive blood, an anamnestic response results in the early production of alloantibody that binds donor RBCs. The alloantibody is detectable 1–2 weeks following the transfusion, and the posttransfusion DAT may become positive

due to circulating donor RBCs coated with antibody or complement. The transfused, alloantibody-coated erythrocytes are cleared by the reticuloendothelial system. These reactions are detected most commonly in the blood bank when a subsequent patient sample reveals a positive alloantibody screen or a new alloantibody in a recently transfused recipient.

No specific therapy is usually required, although additional RBC transfusions may be necessary. Delayed serologic transfusion reactions are similar to DHTR, as the DAT is positive and alloantibody is detected; however, RBC clearance is not increased.

Febrile Nonhemolytic Transfusion Reaction

The most frequent reaction associated with the transfusion of cellular blood components is a febrile nonhemolytic transfusion reaction (FNHTR). These reactions are characterized by chills and rigors and a ≥1°C rise in temperature. FNHTR is diagnosed when other causes of fever in the transfused patient are ruled out. Antibodies directed against donor leukocyte and HLA antigens may mediate these reactions; thus, multiply transfused patients and multiparous women are felt to be at increased risk. Although anti-HLA antibodies may be demonstrated in the recipient's serum, investigation is not routinely done because of the mild nature of most FNHTRs. The use of leukocyte-reduced blood products may prevent or delay sensitization to leukocyte antigens and thereby reduce the incidence of these febrile episodes. Cytokines released from leukocytes within stored blood components may mediate FNHTR; thus, leukoreduction before storage may prevent these reactions. Likewise, cytokines and chemokines released from platelets components, released during storage may also mediate FNHTR.

Allergic Reactions

Urticarial reactions are related to plasma proteins found in transfused components. Mild reactions may be treated symptomatically by temporarily stopping the transfusion and administering antihistamines (diphenhydramine, 50 mg orally or intramuscularly). The transfusion may be completed after the signs and/or symptoms resolve. Patients with a history of allergic transfusion reaction may be premedicated with an antihistamine. Cellular components can be washed to remove residual plasma for the extremely sensitized patient. Most of the allergic presentation may not depend on preformed antibodies and may be attributable to biological response modifiers triggering histamine and serotonin release from platelets and leukocytes.

Anaphylactic Reaction

This severe allergic reaction presents after transfusion of only a few milliliters of the blood component. Symptoms and signs include difficulty breathing, coughing, nausea and vomiting, hypotension, bronchospasm, loss of consciousness, respiratory arrest, and shock. Treatment includes stopping the transfusion, maintaining vascular access, and administering epinephrine (0.5–1 mL of 1:1000 dilution subcutaneously). Glucocorticoids may be required in severe cases.

Patients who are IgA-deficient, <1% of the population, may be sensitized to this Ig class and are at risk for anaphylactic reactions associated with plasma transfusion. Individuals with severe IgA deficiency should therefore receive only IgA-deficient plasma and washed cellular blood components. Patients who have anaphylactic or

repeated allergic reactions to blood components should be tested for IgA deficiency. Of note, the importance of the allergic risk associated with IgA deficiency may be overestimated and is currently debated.

Graft-Versus-Host Disease

Graft-versus-host disease (GVHD) is a frequent complication of allogeneic stem cell transplantation, in which lymphocytes from the donor attack and cannot be eliminated by an immunodeficient host. Transfusion-related GVHD is mediated by donor T lymphocytes that recognize host HLA antigens as foreign and mount an immune response, which is manifested clinically by the development of fever, a characteristic cutaneous eruption, diarrhea, and liver function abnormalities. GVHD can also occur when blood components that contain viable T lymphocytes are transfused to immunodeficient recipients or to immunocompetent recipients who share HLA antigens with the donor (e.g., a family donor). In addition to the aforementioned clinical features of GVHD, transfusion-associated GVHD (TA-GVHD) is characterized by marrow aplasia and pancytopenia. TA-GVHD is highly resistant to treatment with immunosuppressive therapies, including glucocorticoids, cyclosporine, antithymocyte globulin, and ablative therapy followed by allogeneic bone marrow transplantation. Clinical manifestations appear at 8–10 days, and death occurs at 3–4 weeks posttransfusion.

TA-GVHD can be prevented by irradiation of cellular components (minimum of 2500 cGy) before transfusion to patients at risk. Recently, pathogen inactivation technologies have shown to prevent TA-GVHD as well. Patients at risk for TA-GVHD include fetuses receiving intrauterine transfusions, selected immunocompetent (e.g., lymphoma patients) or immunocompromised recipients, recipients of donor units known to be from a blood relative, and recipients who have undergone marrow transplantation. Directed donations by family members should be discouraged (they are not less likely to transmit infection); lacking other options, the blood products from family members should always be irradiated.

Transfusion-Related Acute Lung Injury

TRALI is among the most common cause of transfusion related fatalities. The recipient develops symptoms of hypoxia (PaO2/FiO2 < 300 mmHg) and signs of noncardiogenic pulmonary edema, including bilateral interstitial infiltrates on chest x-ray, either during or within 6 h of transfusion. Treatment is supportive, and patients usually recover without sequelae. TRALI usually results from the transfusion of donor plasma that contains high-titer anti-HLA class II antibodies that bind recipient leukocytes. Anti-HLA class I and anti-human neutrophil antigen (HNA) antibodies can be involved as well. The leukocytes aggregate in the pulmonary vasculature and release mediators that increase capillary permeability. Testing the donor's plasma for anti-HLA antibodies can support this diagnosis. The implicated donors are frequently multiparous women. The transfusion of plasma and platelets from male and nulliparous women donors reduces the risk of TRALI. Recipient factors that are associated with increased risk of TRALI include smoking, chronic alcohol use, shock, liver surgery (transplantation), mechanical ventilation with >30 cm H_20 pressure support and positive fluid balance.

Posttransfusion Purpura

This reaction presents as thrombocytopenia 7–10 days after platelet transfusion and occurs predominantly in women. Platelet-specific antibodies are found in the recipient's serum, and the most frequently recognized antigen is HPA-1a found on the platelet glycoprotein IIIa receptor. The delayed thrombocytopenia is due to the production

of antibodies that react to both donor and recipient platelets. Additional platelet transfusions can worsen the thrombocytopenia and should be avoided. Treatment with intravenous immunoglobulin may neutralize the effector antibodies, or plasmapheresis can be used to remove the antibodies.

Alloimmunization

A recipient may become alloimmunized to a number of antigens on cellular blood elements and plasma proteins. Alloantibodies to RBC antigens are detected during pretransfusion testing, and their presence may delay finding antigen-negative cross-match-compatible products for transfusion. Women of childbearing age who are sensitized to certain RBC antigens (i.e., D, c, E, Kell, or Duffy) are at risk for bearing a fetus with hemolytic disease of the newborn. Matching for RBC antigen is the only pretransfusion selection test to prevent RBC alloimmunization.

Alloimmunization to antigens on leukocytes and platelets can result in refractoriness to platelet transfusions. Once alloimmunization has developed, HLA-compatible platelets from donors who share similar antigens with the recipient may be difficult to find. Hence, prudent transfusion practice is directed at preventing sensitization through the use of leukocyte-reduced cellular components, as well as limiting antigenic exposure by the judicious use of transfusions and use of SDAPs.

NONIMMUNOLOGIC REACTIONS

Fluid Overload

Blood components are excellent volume expanders, and transfusion may quickly lead to transfusion-associated circulatory overload (TACO). Dyspnea with O_2% <90 on room air, bilateral infiltrates on CXR with systolic hypertension are found with TACO. Brain natriuretic peptide (BNP) is often elevated (>1.5) that of pre-transfusion levels. Monitoring the rate and volume of the transfusion and using a diuretic can minimize this problem.

Hypothermia

Refrigerated (4°C) or frozen (−18°C or below) blood components can result in hypothermia when rapidly infused. Cardiac dysrhythmias can result from exposing the sinoatrial node to cold fluid. Use of an in-line warmer will prevent this complication.

Electrolyte Toxicity

RBC leakage during storage increases the concentration of potassium in the unit. Neonates and patients in renal failure are at risk for hyperkalemia. Preventive measures, such as using fresh or washed RBCs, are warranted for neonatal transfusions because this complication can be fatal.

Citrate, commonly used to anticoagulate blood components, chelates calcium and thereby inhibits the coagulation cascade. Hypocalcemia, manifested by circumoral numbness and/or tingling sensation of the fingers and toes, may result from multiple rapid transfusions. Because citrate is quickly metabolized to bicarbonate, calcium infusion is seldom required in this setting. If calcium or any other intravenous infusion is necessary, it must be given through a separate line.

Iron Overload

Each unit of RBCs contains 200–250 mg of iron. Symptoms and signs of iron overload affecting endocrine, hepatic, and cardiac function are common after 100 units of RBCs have been transfused (total-body iron load of 20 g). Preventing this complication by using alternative therapies (e.g., erythropoietin) and judicious transfusion is preferable and cost effective. Chelating agents, such as deferoxamine and deferasirox, are available, but the response though is often suboptimal.

Hypotensive Reactions

Transient hypotension may be noted among transfused patients who take angiotensin-converting enzyme (ACE) inhibitors. Since blood products contain bradykinin that is normally degraded by ACE, patients on ACE inhibitors may have increased bradykinin levels that cause hypotension in the recipient. The blood pressure typically returns to normal without intervention.

Immunomodulation

Transfusion of allogeneic blood is immunosuppressive. Multiply transfused renal transplant recipients are less likely to reject the graft, and transfusion may result in poorer outcomes in cancer patients and increase the risk of infections. Transfusion-related immunomodulation is thought to be mediated by transfused leukocytes. Leukocyte-depleted cellular products may cause less immunosuppression, though controlled data are unlikely to be obtained as the blood supply becomes universally leukocyte-depleted.

INFECTIOUS COMPLICATIONS

The blood supply is initially screened by selecting healthy donors without high-risk lifestyles, medical conditions, or exposure to transmissible pathogens, such as intravenous drug use or visiting malaria endemic areas. Multiple tests performed on donated blood to detect the presence of infectious agents using nucleic acid amplification testing (NAT) or evidence of prior infections by testing for antibodies to pathogens and sterility of platelet products further reduce the risk of transfusion-acquired infections. Pathogen reduction of platelets and plasma offer an additional mean to reduce such risk.

Viral Infections

HEPATITIS C VIRUS

Blood donations are tested for antibodies to HCV and HCV RNA. The risk of acquiring HCV through transfusion is now calculated to be 0,1 to 1 in 1,000,000 units. Infection with HCV may be asymptomatic or lead to chronic active hepatitis, cirrhosis, and liver failure.

HUMAN IMMUNODEFICIENCY VIRUS TYPE 1

Donated blood is tested for antibodies to HIV-1, HIV-1 p24 antigen, and HIV RNA using NAT. The risk of HIV-1 infection per transfusion episode is 0,1 to 1 in 1 million. Antibodies to HIV-2 are also measured in donated blood. No cases of HIV-2 infection in blood donors have been reported in the United States since 1992.

HEPATITIS B VIRUS

Donated blood is screened for HBV using assays for hepatitis B surface antigen (HbsAg) most combined with NAT testing. The risk of transfusion-associated HBV infection is several times greater than for HCV. Vaccination of individuals who require long-term transfusion therapy can prevent this complication.

OTHER HEPATITIS VIRUSES

Hepatitis A virus is rarely transmitted by transfusion; infection is typically asymptomatic and does not lead to chronic disease. Hepatitis E (HEV) can be transmitted by transfusion and may lead to chronic disease. Routine HEV RNA testing has been introduced in several European countries starting in 2015. West Nile virus (WNV) Transfusion-transmitted WNV infections were documented in 2002. This RNA virus can be detected using NAT; routine screening began in 2003. WNV infections range in severity from asymptomatic to fatal, with the older population at greater risk.

CYTOMEGALOVIRUS

This ubiquitous virus infects ≥50% of the general population and is transmitted by the infected “passenger” leukocytes found in transfused PRBCs or platelet components. Cellular components that are leukocyte-reduced have a decreased risk of transmitting CMV, regardless of the serologic status of the donor. Groups at risk for CMV infections include immunosuppressed patients, CMV-seronegative transplant recipients, and neonates; these patients should receive leukocyte-depleted components or CMV seronegative products.

HUMAN T LYMPHOTROPIC VIRUS (HTLV) TYPE I

Assays to detect HTLV-I and -II are used to screen all donated blood. HTLV-I is associated with adult T cell leukemia/lymphoma and tropical spastic paraparesis in a small percentage of infected persons **(Chap. 196)**. The risk of transfusion-mediated HTLV-1 infection is further mitigated by blood product leukoreduction. HTLV-II is not clearly associated with any disease.

PARVOVIRUS B-19

Blood components and pooled plasma products can transmit this virus, the etiologic agent of erythema infectiosum, or fifth disease, in children. Parvovirus B-19 shows tropism for erythroid precursors and inhibits both erythrocyte production and maturation. Pure red cell aplasia, presenting either as acute aplastic crisis or chronic anemia with shortened RBC survival, may occur in individuals with an underlying hematologic disease, such as sickle cell disease or thalassemia **(Chap. 94)**. The fetus of a seronegative woman is at risk for developing hydrops from this virus.

Bacterial Contamination

The relative risk of transfusion-transmitted bacterial infection has increased as the absolute risk of viral infections has dramatically decreased.

Most bacteria do not grow well at cold temperatures; thus, PRBCs and FFP are not common sources of bacterial contamination. However, some gram-negative bacteria can grow at 1° to 6°C. *Yersinia*, *Pseudomonas*, *Serratia*, *Acinetobacter*, and *Escherichia* species have all been implicated in infections related to PRBC transfusion. Platelet concentrates, which are stored at room temperature, are more likely to contain skin contaminants such as gram-positive organisms, including coagulase-negative staphylococci. It is estimated that 1 in 1000–2000 platelet components is contaminated with bacteria. The risk of death due to transfusion-associated sepsis is estimated to be in the order of 1 in 200,000–400,000 platelets products. Since 2004, blood banks have instituted methods to detect contaminated platelet components. Pathogen-reduced platelets have been available and offer an alternative to prevent transfusion-transmitted bacterial infection.

Recipients of transfusion contaminated with bacteria may develop fever and chills, which can progress to septic shock and DIC. These reactions may occur abruptly, within minutes of initiating the transfusion, or after several hours. The onset of symptoms and signs is often sudden and fulminant, which distinguishes bacterial contamination from an FNHTR. The reactions, particularly those related to gram-negative contaminants, are the result of infused endotoxins formed within the contaminated stored component.

When these reactions are suspected, the transfusion must be stopped immediately. Therapy is directed at reversing any signs of shock, and broad-spectrum antibiotics should be given. The blood bank should be notified to identify any clerical or serologic error. The blood component bag should be sent for culture and Gram stain.

Other Infectious Agents

Various parasites, including those causing malaria, babesiosis, and Chagas disease, can be transmitted by blood transfusion. Geographic migration and travel of donors shift the incidence of these infections. Other agents implicated in transfusion transmission include dengue, zika virus, variant Creutzfeldt-Jakob disease, *Anaplasma phagocytophilum*, and yellow fever vaccine virus and the list will grow. Tests for some pathogens are available, such as *Trypanosoma cruzi*, but not universally required while others are being developed (Babesia microti). These infections should be considered in the transfused patient in the appropriate clinical setting.

ALTERNATIVES TO TRANSFUSION

Alternatives to allogeneic blood transfusions that avoid homologous donor exposures with attendant immunologic and infectious risks remain attractive. Autologous blood remains an option when transfusion is anticipated. However, the cost-benefit ratio of autologous transfusion remains high. No transfusion is a zero-risk event; clerical errors and bacterial contamination remain potential complications even with autologous transfusions. Additional methods of autologous transfusion in the surgical patient include preoperative hemodilution, recovery of shed blood from sterile surgical sites, and postoperative drainage collection. Directed or designated donation from friends and family of the potential recipient has not been safer than volunteer donor component transfusions. Such directed donations may in fact place the recipient at higher risk for complications such as GVHD and alloimmunization.

Granulocyte and granulocyte-macrophage colony-stimulating factors are clinically useful to hasten leukocyte recovery in patients with leukopenia related to high-dose chemotherapy. Erythropoietin stimulates erythrocyte production in patients with anemia of chronic renal failure and other conditions, thus avoiding or reducing the

need for transfusion. This hormone can also stimulate erythropoiesis in the autologous donor to enable additional donation.

Gene therapy approaches in patients with sickle cell or major thalassemia offers the potential of dramatically reducing their transfusion needs. Stem cell-derived blood cells such as RBCs or platelets may in the future become a suitable alternative to very rare blood donors.

Lastly, optimizing the use of blood products through patient blood management programs can improve the therapeutic index of transfusion medicine.

ACKNOWLEDGMENT

We are pleased to acknowledge the assistance of Olivier Garaud, MD, PhD, and Jacques Chiaroni, MD, PhD, in the preparation of this chapter.

FURTHER READING

Carson JL et al: Indications for and adverse effects of red-cell transfusion. N Engl J Med 377:1261, 2017.

Delaney M et al: Biomedical Excellence for Safer Transfusion (BEST) Collaborative. Transfusion reactions: Prevention, diagnosis, and treatment. Lancet 388:2825, 2016.

Goodnough LT et al: Concepts of blood transfusion in adults. Lancet 381:1845, 2013.

Hendrickson JE, Tormey CA: Red blood cell antibodies in hematology/oncology patients: Interpretation of immunohematologic tests and clinical significance of detected antibodies. Hematol Oncol Clin North Am 30:635, 2016.

Zou S et al: Donor testing and risk: Current prevalence, incidence, and residual risk of transfusion-transmissible agents in US allogeneic donations. Transfus Med Rev 26:119, 2012.

Chapter 110: Hematopoietic Cell Transplantation

Frederick R. Appelbaum

FIGURE 110-1

INTRODUCTION

Bone marrow transplantation was the original term used to describe the collection and transplantation of hematopoietic stem cells, but with the demonstration that peripheral blood and umbilical cord blood are also useful sources of stem cells, *hematopoietic cell transplantation* (HCT) has become the preferred generic term for this process. HCT is used to treat patients with an abnormal but nonmalignant lymphohematopoietic system by replacing it with one from a normal donor. HCT is also used to treat malignancy by allowing the administration of higher doses of myelosuppressive therapy than would otherwise be possible, and in the setting of allogeneic HCT, by conferring an immunologic graft-versus-tumor effect. The use of HCT has been increasing, both because of its efficacy in selected diseases and because of increasing availability of donors.

The Center for International Blood and Marrow Transplant Research (*http://www.cibmtr.org*) estimates that worldwide about 70,000 transplants were performed in 2016. The frequency of transplantation varied widely from country to country with a close association of transplant rates with Gross National Income (GNI)/capita. However, even among countries with similar GNIs/capita, there are substantial differences between countries and regions concerning the frequency of transplantation, disease indications, and choice of donor type.

THE HEMATOPOIETIC STEM CELL

Several features of the hematopoietic stem cell make transplantation clinically feasible, including its remarkable regenerative capacity, its ability to home to the marrow space following intravenous injection, and the ability of the stem cell to be cryopreserved **(Chap. 92)**. Transplantation of a single stem cell can replace the entire lymphohematopoietic system of an adult mouse. In humans, transplantation of a small percentage of a donor's bone marrow volume regularly results in complete and sustained replacement of the recipient's entire lymphohematopoietic system, including all red cells, granulocytes, B and T lymphocytes, and platelets, as well as cells comprising the fixed macrophage population, including Kupffer cells of the liver, pulmonary alveolar macrophages, osteoclasts, Langerhans cells of the skin, and brain microglial cells. The ability of the hematopoietic stem cell to home to the marrow following intravenous injection is mediated, in part, by an interaction between CXCL12, also known as stromal cell–derived factor 1, produced by marrow stromal cells and the alpha-chemokine receptor CXCR4 found on stem cells. Homing is also

influenced by the interaction of cell-surface molecules, termed *selectins*, including E- and L-selectin, on bone marrow endothelial cells with ligands, termed *integrins*, such as VLA-4, on early hematopoietic cells. Human hematopoietic stem cells can survive freezing and thawing with little, if any, damage, making it possible to remove and store a portion of the patient's own bone marrow for later reinfusion following treatment of the patient with high-dose myelotoxic therapy.

CATEGORIES OF HCT

HCT can be described according to the relationship between the patient and the donor and by the anatomic source of stem cells. In ~1% of cases, patients have identical twins who can serve as donors. With the use of syngeneic donors, there is no risk of graft-versus-host disease (GVHD), and unlike the use of autologous marrow, there is no risk that the stem cells are contaminated with tumor cells.

Allogeneic transplantation involves a donor and a recipient who are not genetically identical. Following allogeneic transplantation, immune cells transplanted with the stem cells or developing from them can react against the patient, causing GVHD. Alternatively, if the immunosuppressive preparative regimen used to treat the patient before transplant is inadequate, immunocompetent cells of the patient can cause graft rejection. The risks of these complications are greatly influenced by the degree of matching between donor and recipient for human leukocyte antigen (HLA) molecules encoded by genes of the major histocompatibility complex.

HLA molecules are responsible for binding antigenic proteins and presenting them to T cells. The antigens presented by HLA molecules may derive from exogenous sources (e.g., during active infections) or may be endogenous proteins. If individuals are not HLA-matched, T cells from one individual will react strongly to the mismatched HLA, or "major antigens," of the second. Even if the individuals are HLA-matched, the T cells of the donor may react to differing endogenous or "minor antigens" presented by the HLA of the recipient. Reactions to minor antigens tend to be less vigorous. The genes of major relevance to transplantation include HLA-A, -B, -C, and -D; they are closely linked and therefore tend to be inherited as haplotypes, with only rare crossovers between them. Thus, the odds that any one full sibling will match a patient are one in four, and the probability that the patient has an HLA-identical sibling is $1 - (0.75)^n$, where n equals the number of siblings.

With standard techniques, the risk of graft rejection is 1–3%, and the risk of severe, life-threatening acute GVHD is ~15% following transplantation between HLA-identical siblings. The incidence of graft rejection and GVHD increases progressively with the use of family member donors mismatched for one, two, or three antigens. Although survival following a one-antigen mismatched transplant is not markedly altered, survival following two- or three-antigen mismatched transplants is reduced. Since the formation of the National Marrow Donor Program and other registries, it has become possible to identify HLA-matched unrelated donors for many patients. The genes encoding HLA antigens are highly polymorphic, and thus the odds of any two unrelated individuals being HLA identical are extremely low, somewhat less than 1 in 10,000. However, by identifying and typing >25 million volunteer donors, HLA-matched donors can now be found for

~60% of patients for whom a search is initiated, with higher rates among whites and lower rates among minorities and patients of mixed race. It takes, on average, 3–4 months to complete a search and schedule and initiate an unrelated donor transplant. With improvements in HLA typing and supportive care measures, survival following matched unrelated donor transplantation is essentially the same as that seen with HLA-matched siblings. Methods have recently been developed that enable the selection of "permissive" single-antigen mismatched unrelated donors that result in transplant outcomes similar to those seen with full matches.

Allogeneic HCT can be carried out across ABO blood barriers by removing isoagglutinins and/or incompatible red blood cells from the donor graft. However, depending on the direction of the mismatch, hemolysis of donor cells by persistent isoagglutinins in the host, or hemolysis of recipient red cells by isoagglutinins in the graft or developing from it may occur despite appropriate manipulation of the donor cell product.

Autologous transplantation involves the removal and storage of the patient's own stem cells with subsequent reinfusion after the patient receives high-dose myeloablative therapy. Unlike allogeneic transplantation, there is no risk of GVHD or graft rejection with autologous transplantation. On the other hand, autologous transplantation lacks a graft-versus-tumor (GVT) effect, and the autologous stem cell product can be contaminated with tumor cells, which could lead to relapse. A variety of techniques have been developed to "purge" autologous products of tumor cells, but no prospective randomized trials have yet shown that any approach results in a decrease in relapse rates or improvement in disease-free or overall survival.

Bone marrow aspirated from the posterior and anterior iliac crests initially was the source of hematopoietic stem cells for transplantation. Typically, anywhere from 1.5 to 5×10^8 nucleated marrow cells per kilogram are collected for allogeneic transplantation. Several studies have found improved survival in the settings of both matched sibling and unrelated transplantation by transplanting higher numbers of bone marrow cells.

Hematopoietic stem cells circulate in the peripheral blood but in very low concentrations. Following the administration of a myeloid growth factor such as granulocyte colony-stimulating factor (G-CSF) and during recovery from intensive chemotherapy, the concentration of hematopoietic progenitor cells in blood, as measured either by colony-forming units or expression of the CD34 antigen, increases markedly. This has made it possible to harvest adequate numbers of stem cells from the peripheral blood for transplantation. Donors are typically treated with 4 or 5 days of hematopoietic growth factor, following which stem cells are collected in one or two 4-h pheresis sessions. In the autologous setting, transplantation of $>2.5 \times 10^6$ CD34 cells per kilogram, a number that can be collected in most circumstances, leads to rapid and sustained engraftment in virtually all cases. In the 10–20% of patients who fail to mobilize sufficient CD34+ cells with growth factor alone, the addition of plerixafor, an antagonist of CXCR4, may be useful. When compared to the use of autologous marrow, use of peripheral blood stem cells results in more rapid hematopoietic recovery. Although this more rapid recovery diminishes the morbidity rate of transplantation, no studies show improved survival.

Clinical trials have shown that the use of growth factor–mobilized peripheral blood stem cells from HLA-matched family members leads to faster engraftment without an increase in acute GVHD. Chronic GVHD may

be increased with peripheral blood stem cells, but in trials conducted so far, this has been more than balanced by reductions in relapse rates and non-relapse mortality rates, with the use of peripheral blood stem cells resulting in improved overall survival. However, in the setting of matched unrelated donor transplantation, use of peripheral blood results in more chronic GVHD without a compensatory survival advantage, favoring the use of bone marrow in this setting.

Umbilical cord blood contains a high concentration of hematopoietic progenitor cells, allowing for its use as a source of stem cells for transplantation. Cord blood transplantation from family members has been explored in the setting where the immediate need for transplantation precludes waiting the 9 or so months generally required for the baby to mature to the point of donating marrow. Use of cord blood results in slower peripheral count recovery than seen with marrow but a lower incidence of GVHD, perhaps reflecting the low number of T cells in cord blood. Multiple cord blood banks have been developed to harvest and store cord blood for possible transplantation to unrelated patients from material that would otherwise be discarded. Currently more than 500,000 units are cryopreserved and available for use. The advantages of unrelated cord blood are rapid availability and decreased immune reactivity allowing for the use of partially matched units, which is of particular importance for those without matched unrelated donors. The risks of graft failure and transplant-related mortality are related to the dose of cord blood cells per kilogram, which previously limited the application of single cord blood transplantation to pediatric and smaller adult patients. Subsequent trials have found that for patients without suitable single cord units, the use of double cord transplants diminishes the risk of graft failure and early mortality even though only one of the donors ultimately engrafts. Survival rates are now similar with unrelated donor and cord blood transplantation with the result that a potential allogeneic donor can be found for almost every patient in need **(see Table 110-1)**.

TABLE 110-1

Probability of Identifying a Donor Based on Stem Cell Source and Patient Ethnicity

	Unrelated Adult %		**Unrelated Cord %**
Ethnicity	8/8[a]	7/8[a]	≥4/6[b]
Caucasian	75	90	>95
Hispanic	35	75	95
Black	18	70	90

[a]Matching for HLA-A, -B, -C, and DRB1. [b]Matching for HLA-A, -B, and DRB1.

THE TRANSPLANT PREPARATIVE REGIMEN

The treatment regimen administered to patients immediately preceding transplantation is designed to eradicate the patient's underlying disease and, in the setting of allogeneic transplantation, immunosuppress the patient adequately to prevent rejection of the transplanted marrow. The appropriate regimen therefore depends on the disease setting and source of marrow. For example, when transplantation is performed to treat severe combined immunodeficiency and the donor is a histocompatible sibling, no treatment is needed because no host cells require eradication and the patient is already too immunoincompetent to reject the transplanted marrow. For aplastic anemia, there is no large population of cells to eradicate, and high-dose cyclophosphamide plus antithymocyte globulin are sufficient to immunosuppress the patient adequately to accept the marrow graft. In the setting of thalassemia and sickle cell anemia, high-dose busulfan is frequently added to cyclophosphamide in order to eradicate hyperplastic host hematopoiesis. A variety of different regimens have been developed to treat malignant diseases. Most of these regimens include agents that have high activity against the tumor in question at conventional doses and have myelosuppression as their predominant dose-limiting toxicity. Therefore, these regimens commonly include busulfan, cyclophosphamide, melphalan, thiotepa, carmustine, etoposide, and total-body irradiation in various combinations.

Although high-dose treatment regimens were the initial approach to transplantation for malignancies, the understanding that much of the antitumor effect of transplantation derives from an immunologically mediated GVT response led investigators to ask if reduced-intensity conditioning regimens might be effective and more tolerable. Evidence for a GVT effect comes from studies showing that post-transplant relapse rates are lowest in patients who develop acute and chronic GVHD, higher in those without GVHD, and higher still in recipients of T cell–depleted allogeneic or syngeneic marrow. The demonstration that complete remissions can be obtained in many patients who have relapsed after transplant by simply administering viable lymphocytes from the original donor further strengthens the argument for a potent GVT effect. Accordingly, a variety of alternative regimens have been studied, ranging from non-myeloablative that are the very minimum required to achieve engraftment (e.g., fludarabine plus 200 cGy total-body irradiation) and would cause only transient myelosuppression if no transplant were performed, to so-called reduced intensity regimens, which would cause significant but not necessarily fatal myelosuppression in the absence of transplantation (e.g., fludarabine plus melphalan). Studies to date document that engraftment can be readily achieved with less toxicity than seen with conventional transplantation. Complete sustained responses have been documented in many patients, particularly those with more indolent hematologic malignancies. In general, relapse rates are higher following reduced-intensity conditioning, but transplant-related mortality is lower, favoring the use of reduced-intensity conditioning in patients with significant comorbidities. High-dose regimens are favored in all those felt able to tolerate the treatment.

THE TRANSPLANT PROCEDURE

Marrow is usually collected from the donor's posterior and sometimes anterior iliac crests, with the donor under general or spinal anesthesia. Typically, 10–15 mL/kg of marrow is aspirated, placed in heparinized media, and filtered through 0.3- and 0.2-mm screens to remove fat and bony spicules. The collected marrow may undergo further processing depending on the clinical situation, such as the removal of red cells to

prevent hemolysis in ABO-incompatible transplants, the removal of donor T cells to prevent GVHD, or attempts to remove possible contaminating tumor cells in autologous transplantation. Marrow donation is safe, with only very rare complications reported.

Peripheral blood stem cells are collected by leukapheresis after the donor has been treated with hematopoietic growth factors or, in the setting of autologous transplantation, sometimes after treatment with a combination of chemotherapy and growth factors. Stem cells for transplantation are infused through a large-bore central venous catheter. Such infusions are usually well tolerated, although occasionally patients develop fever, cough, or shortness of breath. These symptoms typically resolve with slowing of the infusion. When the stem cell product has been cryopreserved using dimethyl sulfoxide, patients more often experience short-lived nausea or vomiting due to the odor and taste of the cryoprotectant.

ENGRAFTMENT AND IMMUNE RECONSTITUTION

Peripheral blood counts usually reach their nadir several days to a week after transplant as a consequence of the preparative regimen; then cells produced by the transplanted stem cells begin to appear in the peripheral blood. The rate of recovery depends on the source of stem cells and the use of post-transplant growth factors. If marrow is the source of stem cells, recovery to 100 granulocytes/μL occurs on average by day 16 and to 500/μL by day 22. Use of G-CSF–mobilized peripheral blood stem cells speeds the rate of recovery by ~1 week when compared to marrow, whereas engraftment following cord blood transplantation is typically delayed by ~1 week compared to marrow. Use of a myeloid growth factor after transplant can accelerate recovery by 3–5 days. Platelet counts usually recover shortly after granulocytes.

While granulocytes and other components of innate immunity recover rapidly after HCT, adaptive immunity, which consists of cellular (T cell) and humoral (B cell) immunity, may take 1–2 years to fully recover. Survival and peripheral expansion of infused donor T cells is the dominant mechanism for T cell recovery in the first months after HCT and results in mostly CD8+ T cells with a limited repertoire. After several months, de novo generation of donor-derived CD4+ and CD8+ T cells becomes dominant providing a more diverse T cell repertoire. B cell counts recover by 6 months after autologous HCT and 9 months after allogeneic HCT. In general, immune recovery occurs more rapidly after autologous than allogeneic HCT and after receipt of unmodified grafts compared to the setting of in vivo or ex vivo T-cell depletion.

Following allogeneic transplantation, engraftment can be documented using fluorescence in situ hybridization of sex chromosomes if donor and recipient are sex-mismatched or by analysis of short tandem repeat polymorphisms after DNA amplification.

COMPLICATIONS FOLLOWING HEMATOPOIETIC CELL TRANSPLANT

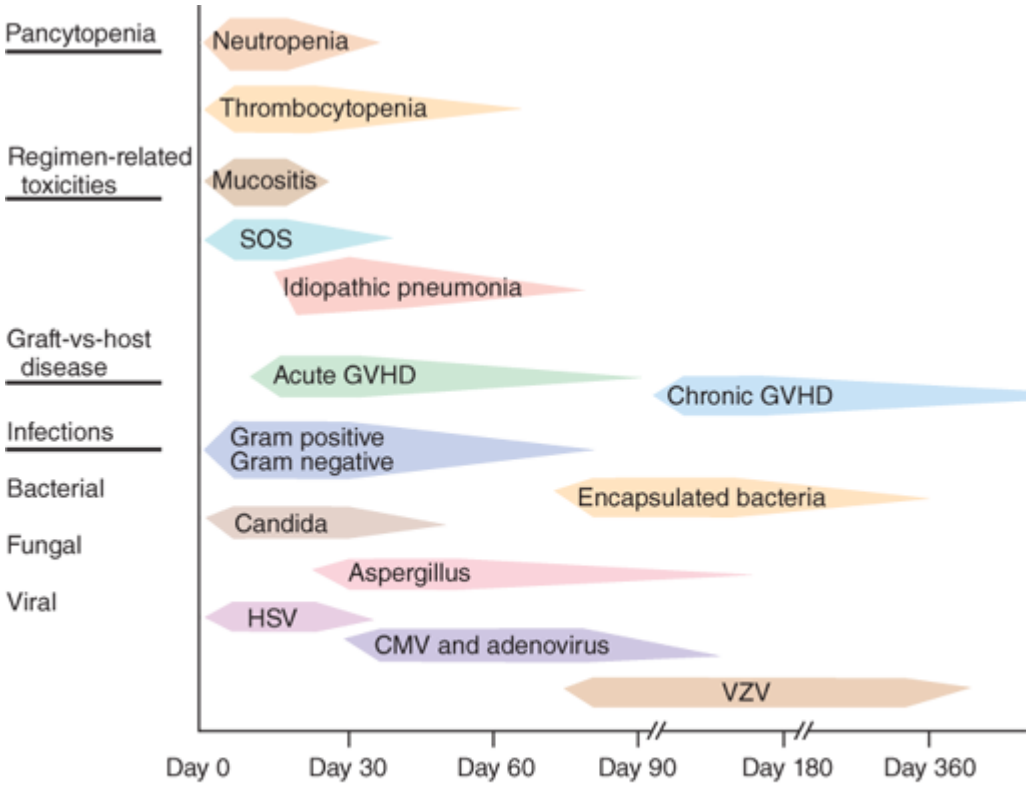
Early Direct Chemoradiotoxicities

The transplant preparative regimen may cause a spectrum of acute toxicities that vary according to intensity of the regimen and the specific agents used but frequently include nausea, vomiting, and mild skin erythema

(Fig. 110-1). High-dose cyclophosphamide can result in hemorrhagic cystitis, which can usually be prevented by bladder irrigation or with the sulfhydryl compound mercaptoethanesulfonate (MESNA). Most high-dose preparative regimens will result in oral mucositis, which typically develops 5–7 days after transplant and often requires narcotic analgesia. Use of a patient-controlled analgesic pump provides the greatest patient satisfaction and results in a lower cumulative dose of narcotic. Keratinocyte growth factor (palifermin) can shorten the duration of mucositis by several days following autologous transplantation. Patients begin losing their hair 5–6 days after transplant and by 1 week are usually profoundly pancytopenic.

FIGURE 110-1

Major syndromes complicating marrow transplantation. CMV, cytomegalovirus; GVHD, graft-versus-host disease; HSV, herpes simplex virus; SOS, sinusoidal obstructive syndrome (formerly venoocclusive disease); VZV, varicella-zoster virus. The size of the shaded area roughly reflects the period of risk of the complication.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Depending on the intensity of the conditioning regimen, 3–10% of patients will develop sinusoidal obstruction syndrome (SOS) of the liver (formerly called venoocclusive disease), a syndrome that results from direct cytotoxic injury to hepatic-venular and sinusoidal endothelium, with subsequent deposition of fibrin and the development of a local hypercoagulable state. This chain of events leads to the clinical symptoms of tender hepatomegaly, ascites, jaundice, and fluid retention. These symptoms can develop any time during the first month after transplant, with the peak incidence at day 16. Predisposing factors include prior exposure to intensive chemotherapy, pre-transplant hepatitis of any cause, and use of more intense conditioning regimens. The mortality rate of sinusoidal obstruction syndrome is ~30%, with progressive hepatic failure culminating in a terminal hepatorenal syndrome. Treatment of severe SOS with defibrotide, a polydeoxyribonucleotide, can reduce mortality.

Although most pneumonias developing early after transplant are caused by infectious agents, in a small percent of patients a diffuse interstitial pneumonia will develop that is a result of direct toxicity of high-dose preparative regimens. Bronchoalveolar lavage usually shows alveolar hemorrhage, and biopsies are typically characterized by diffuse alveolar damage, although some cases may have a more clearly interstitial pattern. High-dose glucocorticoids or antitumor necrosis factor therapies are sometimes used as treatment, although randomized trials proving their utility have not been reported.

Late Direct Chemoradiotoxicities

Two categories of chronic pulmonary disease occur in patients more than 3 months after HCT. Cryptogenic organizing pneumonia is a restrictive lung disease characterized by dry cough, shortness of breath, and chest imaging showing a diffuse, fluffy infiltrate. Biopsy shows granulation tissue within alveolar spaces and small airways and no infectious agents. The disease responds well to corticosteroids and is entirely reversible. Bronchiolitis obliterans is an obstructive disease presenting with cough, progressive dyspnea, and radiologic evidence of air trapping. Pathology shows collagen and granulation tissue in and around bronchial structures and eventually obliteration of small airways. The disease is usually associated with chronic GVHD and while it may respond to increasing immunosuppression, complete reversal is uncommon.

Other late complications of the preparative regimen include decreased growth velocity in children and delayed development of secondary sex characteristics. These complications can be partly ameliorated with the use of appropriate growth and sex hormone replacement. Most men become azoospermic, and most postpubertal women will develop ovarian failure, which should be treated. However, pregnancy is possible after transplantation, and patients should be counseled accordingly. Thyroid dysfunction, usually well compensated, is sometimes seen. Cataracts develop in 10–20% of patients and are most common in patients treated with total-body irradiation and those who receive glucocorticoid therapy after transplant for treatment of GVHD. Aseptic necrosis of the femoral head is seen in 10% of patients and is particularly frequent in those receiving chronic glucocorticoid therapy. Both acute and late chemoradiotoxicities (except those due to glucocorticoids and other agents used to treat GVHD) are less frequent in recipients of reduced-intensity compared to high-dose preparative regimens.

Graft Failure

Although complete and sustained engraftment is usually seen after transplant, occasionally marrow function either does not return or, after a brief period of engraftment, is lost. Graft failure after autologous transplantation can be the result of inadequate numbers of stem cells being transplanted, damage during ex vivo treatment or storage, or exposure of the patient to myelotoxic agents after transplant. Infections with cytomegalovirus (CMV) or human herpesvirus type 6 have also been associated with loss of marrow function. Graft failure after allogeneic transplantation can also be due to immunologic rejection of the graft by immunocompetent host cells. Such rejection is generally thought to be mostly T-cell mediated, but the presence of pre-HCT of donor-specific HLA antibodies in the patient is associated with poor engraftment, leading to the recommendation for screening for donor-directed anti-HLA antibodies in recipients prior to transplant. Immunologically based graft rejection is more common following use of less immunosuppressive

preparative regimens, in recipients of T cell–depleted stem cell products, and in patients receiving grafts from HLA-mismatched donors or cord blood.

Treatment of graft failure usually involves removing all potentially myelotoxic agents from the patient's regimen and attempting a short trial of a myeloid growth factor. Persistence of lymphocytes of host origin in allogeneic transplant recipients with graft failure indicates immunologic rejection. Reinfusion of donor stem cells in such patients is usually unsuccessful unless preceded by a second immunosuppressive preparative regimen. Standard high-dose preparative regimens are generally tolerated poorly if administered within 100 days of a first transplant because of cumulative toxicities. However, use of reduced-intensity conditioning regimens has been effective in some cases.

Graft-versus-Host Disease

Acute GVHD usually occurs within the first 3 months after allogeneic transplant with a peak onset around 4 weeks and is characterized by an erythematous maculopapular rash; by persistent anorexia or diarrhea, or both; and by liver disease with increased serum levels of bilirubin, alanine and aspartate aminotransferase, and alkaline phosphatase. Because many conditions can mimic acute GVHD, the diagnosis usually requires skin, liver, or endoscopic biopsy for confirmation. In all these organs, endothelial damage and lymphocytic infiltrates are seen. In skin, the epidermis and hair follicles are damaged; in liver, the small bile ducts show segmental disruption; and in intestines, destruction of the crypts and mucosal ulceration may be noted. A commonly used rating system for acute GVHD is shown in Table 110-2. Grade I acute GVHD is of little clinical significance, does not affect the likelihood of survival, and does not require treatment. In contrast, grades II to IV GVHD are associated with significant symptoms and a poorer probability of survival, and require aggressive therapy. The incidence of acute GVHD is higher in recipients of stem cells from mismatched or unrelated donors, in older patients, and in patients unable to receive full doses of drugs used to prevent the disease.

TABLE 110-2

Clinical Staging and Grading of Acute Graft-versus-Host Disease

Clinical Stage	Skin	Liver—Bilirubin, μmol/L (mg/dL)	Gut
1	Rash <25% body surface	34–51 (2–3)	Diarrhea 500–1000 mL/d
2	Rash 25–50% body surface	51–103 (3–6)	Diarrhea 1000–1500 mL/d
3	Generalized erythroderma	103–257 (6–15)	Diarrhea >1500 mL/d
4	Desquamation and bullae	>257 (>15)	Ileus
Overall Clinical Grade	**Skin Stage**	**Liver Stage**	**Gut Stage**
I	1–2	0	0
II	1–3	1	1
III	1–3	2–3	2–3
IV	2–4	2–4	2–4

Currently, the standard approach to GVHD prevention is the administration of a calcinuerin inhibitor (cyclosporine or tacrolimus) combined with an antimetabolite (methotrexate or mycophenolate mofetil) following transplantation. The addition of anti T-cell immune globulin (ATG) may further reduce the incidence of GVHD but has not been shown to improve survival. Other approaches being tested in phase III studies include the addition of sirolimus to the standard two-drug regimen, the removal of subsets or all T cells from the stem cell inoculum, and the use of cyclophosphamide administered several days after transplant in an effort to deplete activated allo-reactive T cells.

Despite prophylaxis, significant acute GVHD will develop in ~30% of recipients of stem cells from matched siblings. Factors associated with a greater risk of acute GVHD include HLA-mismatching between recipient and donor, patient and donor age, use of more intense preparative regimens and use of multiparous women

as donors. Biomarkers, including ST2, REG32, and TNF R1 have been identified that are able to predict the severity of acute GVHD. The disease is usually treated with prednisone at a daily dose of 1–2 mg/kg. Patients who fail to respond to prednisone sometimes respond to alternative immunosuppressants or ATG.

Chronic GVHD occurs most commonly between 3 months and 2 years after allogeneic transplant, developing in 20–50% of recipients. The disease is more common in older patients, in recipients of mismatched or unrelated stem cells, and in those with a preceding episode of acute GVHD. The disease resembles an autoimmune disorder with malar rash, sicca syndrome, arthritis, obliterative bronchiolitis, and bile duct degeneration and cholestasis. Mild chronic GVHD can sometimes be managed using local therapies (topical glucocorticoids to skin and cyclosporine eye drops). More severe disease requires systemic therapy usually with prednisone alone or in combination with cyclosporine. Mortality rates from chronic GVHD average around 15%, but range from 5 to 50% depending on severity. In most patients, chronic GVHD resolves, but it may require 1–3 years of immunosuppressive treatment before these agents can be withdrawn without the disease recurring. Because patients with chronic GVHD are susceptible to significant infection, they should receive prophylactic trimethoprim-sulfamethoxazole, and all suspected infections should be investigated and treated aggressively.

Although onset before or after 3 months after transplant is often used to discriminate between acute and chronic GVHD, occasional patients will develop signs and symptoms of acute GVHD after 3 months (late-onset acute GVHD), whereas others will exhibit signs and symptoms of both acute and chronic GVHD (overlap syndrome). There are as yet no data to suggest that these patients should be treated differently than those with classic acute or chronic GVHD.

From 3 to 5% of patients will develop an autoimmune disorder following allogeneic HCT, most commonly autoimmune hemolytic anemia or idiopathic thrombocytopenic purpura. Unrelated donor source and chronic GVHD are risk factors, but autoimmune disorders have been reported in patients with no obvious GVHD. Treatment is with prednisone, cyclosporine, or rituximab.

Infection

Post-transplant patients, particularly recipients of allogeneic transplantation, require unique approaches to the problem of infection. Early after transplantation, patients are profoundly neutropenic, and because the risk of bacterial infection is so great, most centers place patients on broad spectrum antibiotics once the granulocyte count falls to <500/μL. Prophylaxis against fungal infections reduces rates of infection and improves overall survival. Fluconazole is often used for patients with standard risk, while prophylaxis with mold active agents (voriconizole or posaconazole) should be considered for patients at higher risk, such as those with a prior fungal infection. Patients seropositive for herpes simplex should receive acyclovir prophylaxis. One approach to infection prophylaxis is shown in **Table 110-3**. Despite these prophylactic measures, most patients will develop fever and signs of infection after transplant. The management of patients who become febrile despite bacterial and fungal prophylaxis is a difficult challenge and is guided by individual aspects of the patient and by the institution's experience.

TABLE 110-3

Approach to Infection Prophylaxis in Allogeneic Transplant Recipients

Organism	Agent	Approach
Bacterial	Levofloxacin	750 mg PO or IV daily
Fungal	Fluconazole	400 mg PO qd to day 75 posttransplant
Pneumocystis carinii	Trimethoprim-sulfamethoxazole	1 double-strength tablet PO bid 2 days/week until day 180 or off immunosuppression
Viral		
Herpes simplex	Acyclovir	800 mg PO bid to day 30
Varicella-zoster	Acyclovir	800 mg PO bid to day 365
Cytomegalovirus	Ganciclovir	5 mg/kg IV bid for 7 days, then 5 (mg/kg)/d 5 days/week to day 100

The general problem of infection in the immunocompromised host is discussed in Chap. 138.

Once patients engraft, the incidence of bacterial infection diminishes; however, patients, particularly allogeneic transplant recipients, remain at significant risk of infection. During the period from engraftment until about 3 months after transplant, the most common causes of infection are gram-positive bacteria, fungi (particularly *Aspergillus*), and viruses including CMV. CMV infection, which in the past was frequently seen and often fatal, can be prevented in seronegative patients transplanted from seronegative donors by the use of either seronegative blood products or products from which the white blood cells have been removed. In seropositive patients or patients transplanted from seropositive donors, the use of ganciclovir, either as prophylaxis beginning at the time of engraftment or initiated when CMV first reactivates as evidenced by development of antigenemia or viremia, can significantly reduce the risk of CMV disease. Foscarnet is effective for some patients who develop CMV antigenemia or infection despite the use of ganciclovir or who cannot tolerate the drug, but can be associated with severe electrolyte wasting.

Pneumocystis jiroveci pneumonia, once seen in 5–10% of patients, can be prevented by treating patients with oral trimethoprim-sulfamethoxazole for 1 week before transplant and resuming the treatment once patients have engrafted.

Respiratory viruses that cause community-acquired infections, including respiratory syncytial virus (RSV), parainfluenza virus, influenza virus and, metapneumovirus can be life threatening or fatal in the post-

transplant patient. Protection of patients from infected visitors and staff by avoiding such contacts is critical. Neuraminidase inhibitors are effective for influenza infections. Inhaled ribavirin is sometimes used for RSV.

The risk of infection diminishes considerably beyond 3 months after transplant unless chronic GVHD develops, requiring continuous immunosuppression. Most transplant centers recommend continuing trimethoprim-sulfamethoxazole prophylaxis while patients are receiving any immunosuppressive drugs and also recommend careful monitoring for late CMV reactivation. In addition, many centers recommend prophylaxis against varicella-zoster, using acyclovir for 1 year after transplant. Patients should be revaccinated against tetanus, diphtheria, *Haemophilus influenzae*, polio, and pneumococcal pneumonia starting at 12 months after transplant and against measles, mumps, and rubella (MMR), varicella-zoster virus, and possibly pertussis at 24 months.

TREATMENT OF SPECIFIC DISEASES USING HEMATOPOIETIC CELL TRANSPLANTATION

TREATMENT

TREATMENT

Nonmalignant Diseases

IMMUNODEFICIENCY DISORDERS

By replacing abnormal stem cells with cells from a normal donor, HCT can cure patients of a variety of immunodeficiency disorders including severe combined immunodeficiency, Wiskott-Aldrich syndrome, and Chédiak-Higashi syndrome. The widest experience has been with severe combined immunodeficiency disease, where cure rates of 90% can be expected with HLA-identical donors and success rates of 50–70% have been reported using haplotype-mismatched parents as donors **(Table 110-4)**.

APLASTIC ANEMIA

Transplantation from matched siblings after a preparative regimen of high-dose cyclophosphamide and antithymocyte globulin can cure up to 90% of patients age <40 years with severe aplastic anemia. Results in older patients and in recipients of mismatched family member or unrelated marrow are less favorable; therefore, a trial of immunosuppressive therapy is generally recommended for such patients before considering transplantation. Transplantation is effective in all forms of aplastic anemia including, for example, the syndromes associated with paroxysmal nocturnal hemoglobinuria and Fanconi's anemia. Patients with Fanconi's anemia are abnormally sensitive to the toxic effects of alkylating agents, and so less intensive preparative regimens must be used in their treatment **(Chap. 98)**.

HEMOGLOBINOPATHIES

Marrow transplantation from an HLA-identical sibling following a preparative regimen of busulfan and cyclophosphamide can cure 80–90% of patients with thalassemia major. The best outcomes can be expected if patients are transplanted before they develop hepatomegaly or portal fibrosis and if they have been given

adequate iron chelation therapy. Among such patients, the probabilities of 5-year survival and disease-free survival are 95 and 90%, respectively. Although prolonged survival can be achieved with aggressive chelation therapy, transplantation is the only curative treatment for thalassemia. Transplantation is potentially curative for patients with sickle cell anemia. Two-year survival and disease-free survival rates of 95 and 85%, respectively, have been reported following matched sibling or cord blood transplantation. Decisions about patient selection and the timing of transplantation remain difficult, but transplantation represents a reasonable option for children and young adults who have suffered complications of sickle cell anemia including stroke, recurrent vasoocclusive pain, sickle cell lung disease, or sickle nephropathy **(Chap. 94)**.

OTHER NONMALIGNANT DISEASES

Theoretically, HCT should be able to cure any disease that results from an inborn error of the lymphohematopoietic system. Transplantation has been used successfully to treat congenital disorders of white blood cells such as Kostmann's syndrome, chronic granulomatous disease, and leukocyte adhesion deficiency. Congenital anemias such as Blackfan-Diamond anemia can also be cured with transplantation. Since the penetrance of some congenital marrow failure states is variable, potential family member donors should be carefully screened before use to assure they are not affected. Infantile malignant osteopetrosis is due to an inability of the osteoclast to resorb bone, and because osteoclasts derive from the marrow, transplantation can cure this rare inherited disorder.

HCT has been used as treatment for a number of storage diseases caused by enzymatic deficiencies, such as Gaucher's disease, Hurler's syndrome, Hunter's syndrome, and infantile metachromatic leukodystrophy. Transplantation for these diseases has not been uniformly successful, but treatment early in the course of these diseases, before irreversible damage to extramedullary organs has occurred, increases the chance for success.

Transplantation is being explored as a treatment for severe acquired autoimmune disorders. These trials are based on studies demonstrating that transplantation can reverse autoimmune disorders in animal models and on the observation that occasional patients with coexisting autoimmune disorders and hematologic malignancies have been cured of both with transplantation.

TABLE 110-4

Estimated 5-Year Survival Rates Following Transplantation[a]

Disease	Allogeneic, %	Autologous, %
Severe combined immunodeficiency	90	N/A
Aplastic anemia	90	N/A
Thalassemia	90	N/A
Acute myeloid leukemia		
First remission	55–60	50
Second remission	40	30
Acute lymphocytic leukemia		
First remission	50	40
Second remission	40	30
Chronic myeloid leukemia		
Chronic phase	70	ID
Accelerated phase	40	ID
Blast crisis	15	ID
Chronic lymphocytic leukemia	50	ID
Myelodysplasia	45	ID
Multiple myeloma—initial therapy	N/A	60
Non-Hodgkin’s lymphoma		
First relapse/second remission	40	40
Hodgkin’s disease		

Disease	Allogeneic, %	Autologous, %
First relapse/second remission	40	50

[a]These estimates are generally based on data reported by the International Bone Marrow Transplant Registry. The analysis has not been reviewed by their Advisory Committee.

Abbreviations: ID, insufficient data; N/A, not applicable.

TREATMENT

TREATMENT

Malignant Diseases

ACUTE LEUKEMIA

Allogeneic HCT cures 15–20% of patients who do not achieve complete response after induction chemotherapy for acute myeloid leukemia (AML) and is the only form of therapy that can cure such patients. Thus, all patients with AML who are possible transplant candidates should have their HLA type determined soon after diagnosis to enable HCT for those who fail to enter remission. Cure rates of 30–35% are seen when patients are transplanted in second remission or in first relapse. The best results with allogeneic transplantation are achieved when applied during first remission, with disease-free survival rates averaging 55–60%. Meta-analyses of studies comparing matched related donor transplantation to chemotherapy for adult AML patients age <60 years show a survival advantage with transplantation. This advantage is greatest for those with unfavorable-risk AML and is lost in those with favorable-risk disease. While HCT can be performed in patients up to age 75 and possibly beyond, prospective trials comparing HCT with chemotherapy are lacking for older patients. The role of autologous transplantation in the treatment of AML is less well defined. The rates of disease recurrence with autologous transplantation are higher than those seen after allogeneic transplantation, and cure rates are somewhat less.

Similar to patients with AML, adults with acute lymphocytic leukemia who do not achieve a complete response to induction chemotherapy can be cured in 15–20% of cases with immediate transplantation. Cure rates improve to 30–50% in second remission, and therefore transplantation can be recommended for adults who have persistent disease after induction chemotherapy or who have subsequently relapsed. Transplantation in first remission results in cure rates about 55%. Transplantation appears to offer a survival advantage over chemotherapy for patients with high-risk disease, such as those with Philadelphia chromosome–positive disease. Debate continues about whether adults with standard-risk disease should be transplanted in first remission or whether transplantation should be reserved until relapse. Autologous transplantation is associated with a higher relapse rate but a somewhat lower risk of non-relapse mortality when compared to allogeneic transplantation. There is no obvious role of autologous transplantation for

acute lymphocytic leukemia in first remission, and for second-remission patients, most experts recommend use of allogeneic stem cells if an appropriate donor is available.

CHRONIC LEUKEMIA

Allogeneic HCT is indicated for patients with chronic myeloid leukemia who are in chronic phase but have failed therapy with two or more tyrosine kinase inhibitors. In such patients, cure rates of 70% can be expected. HCT is also recommended for patients with CML who present or progress to accelerated phase or blast crisis, although lower cure rates are seen in such patients **(Chap. 101)**.

Although allogeneic transplantation can cure patients with chronic lymphocytic leukemia (CLL), it has not been extensively studied because of the chronic nature of the disease, the age profile of patients and, more recently, the availability of multiple effective therapies. In those cases where it was studied, complete remissions were achieved in the majority of patients, with disease-free survival rates of ~50% at 3 years, despite the advanced stage of the disease at the time of transplant.

MYELODYSPLASIA AND MYELOPROLIFERATIVE DISORDERS

Between 20 and 65% of patients with myelodysplasia appear to be cured with allogeneic transplantation. Results are better among younger patients and those with less advanced disease. However, patients with early-stage myelodysplasia can live for extended periods without intervention, and so transplantation is generally reserved for patients with an International Prognostic Scoring System (IPSS) score of Int-2 or higher, or for selected patients with an IPSS score of Int-1 who have other poor prognostic features **(Chap. 98)**. Allogeneic HCT can cure patients with primary myelofibrosis or myelofibrosis secondary to polycythemia vera or essential thrombocythemia, with 5-year progression-free survival rates in excess of 65% being reported. It may require many months for the fibrosis to resolve.

LYMPHOMA

Patients with disseminated intermediate- or high-grade non-Hodgkin's lymphoma who have not been cured by first-line chemotherapy and are transplanted in first relapse or second remission can still be cured in 40–50% of cases. This represents a clear advantage over results obtained with conventional-dose salvage chemotherapy. It is unsettled whether patients with high-risk disease benefit from transplantation in first remission. Most experts favor the use of autologous rather than allogeneic transplantation for patients with intermediate- or high-grade non-Hodgkin's lymphoma, because fewer complications occur with this approach and survival appears equivalent. Although autologous transplantation results in high response rates in patients with recurrent disseminated indolent non-Hodgkin's lymphoma, the availability of newer agents for this category of patient leaves the role of transplantation unsettled. Reduced-intensity conditioning regimens followed by allogeneic transplantation result in high rates of complete and enduring complete responses in patients with recurrent indolent lymphomas.

The role of transplantation in Hodgkin's disease is similar to that in intermediate- and high-grade non-Hodgkin's lymphoma. With transplantation, 5-year disease-free survival is 20–30% in patients who never achieve a first remission with standard chemotherapy and up to 70% for those transplanted in second remission. Transplantation has no defined role in first remission in Hodgkin's disease.

MYELOMA

Patients with myeloma who have progressed on first-line therapy can sometimes benefit from allogeneic or autologous transplantation. Prospective randomized studies demonstrate that the inclusion of autologous transplantation as part of the initial therapy of patients results in improved disease-free survival and overall survival. Further benefit is seen with the use of lenalidomide maintenance therapy following transplantation. The use of autologous transplantation followed by non-myeloablative allogeneic transplantation has yielded mixed results.

SOLID TUMORS

Patients with testicular cancer in whom first-line platinum-containing chemotherapy has failed can still be cured in ~50% of cases if treated with high-dose chemotherapy with autologous stem cell support, an outcome better than that seen with low-dose salvage chemotherapy. The use of high-dose chemotherapy with autologous stem cell support is being studied for several other solid tumors, including neuroblastoma and pediatric sarcomas. As in most other settings, the best results have been obtained in patients with limited amounts of disease and where the remaining tumor remains sensitive to conventional-dose chemotherapy. Few randomized trials of transplantation in these diseases have been completed.

POSTTRANSPLANT RELAPSE

Patients who relapse following autologous transplantation sometimes respond to further chemotherapy and may be candidates for possible allogeneic transplantation, particularly if the remission following the initial autologous transplant was long. Several options are available for patients who relapse following allogeneic transplantation. Treatment with infusions of unirradiated donor lymphocytes results in complete responses in as many as 75% of patients with chronic myeloid leukemia, 40% in myelodysplasia, 25% in AML, and 15% in myeloma have been reported. Major complications of donor lymphocyte infusions include transient myelosuppression and the development of GVHD. These complications depend on the number of donor lymphocytes given and the schedule of infusions, with less GVHD seen with lower dose, fractionated schedules.

FURTHER READING

Anasetti C et al: Peripheral-blood stem cells versus bone marrow from unrelated donors. N Engl J Med 367:1487, 2012.

Gooley TA et al: Reduced mortality after allogeneic hematopoietic-cell transplantation. N Engl J Med 363:2091, 2010.

Gragert L et al: HLA match likelihoods for hematopoietic stem-cell grafts in the U.S. registry. N Engl J Med 371:339, 2014.

Gratwohl A et al: One million haemopoietic stem-cell transplants: A retrospective observational study. Lancet Haematol 2:e91, 2015. [PubMed: 26687803]

Holtan SG et al: Acute graft-versus-host disease: A bench-to-bedside update (Review). Blood 124:363, 2014.

Kekre N, Antin JH: Hematopoietic stem cell transplantation donor sources in the 21st century: Choosing the ideal donor when a perfect match does not exist. Blood 124:334, 2014.

Koreth J et al: Allogeneic stem cell transplantation for acute myeloid leukemia in first complete remission: A systematic review and meta-analysis of prospective clinical trials. JAMA 301:2349, 2009.

Milano F et al: Cord-blood transplantation in patients with minimal residual disease. N Engl J Med 375:944, 2016.

Pai SY et al: Transplantation outcomes for severe combined immunodeficiency, 2000–2009. N Engl J Med 371:434, 2014.

Vander Lugt MT et al: ST2 as a marker for risk of therapy-resistant graft-versus-host disease and death. N Engl J Med 369:529, 2013.

Chapter 111: Disorders of Platelets and Vessel Wall

Barbara A. Konkle

INTRODUCTION

Hemostasis is a dynamic process in which the platelet and the blood vessel wall play key roles. Platelets become activated upon adhesion to von Willebrand factor (VWF) and collagen in the exposed subendothelium after injury. Platelet activation is also mediated through shear forces imposed by blood flow itself, particularly in areas where the vessel wall is diseased, and is also affected by the inflammatory state of the endothelium. The activated platelet surface provides the major physiologic site for coagulation factor activation, which results in further platelet activation and fibrin formation. Genetic and acquired influences on the platelet and vessel wall, as well as on the coagulation and fibrinolytic systems, determine whether normal hemostasis or bleeding or clotting symptoms will result.

THE PLATELET

Platelets are released from the megakaryocyte, likely under the influence of flow in the capillary sinuses. The normal blood platelet count is 150,000–450,000/μL. The major regulator of platelet production is the hormone thrombopoietin (TPO), which is synthesized in the liver. Synthesis is increased with inflammation and specifically by interleukin 6. TPO binds to its receptor on platelets and megakaryocytes, by which it is removed from the circulation. Thus a reduction in platelet and megakaryocyte mass increases the level of TPO, which then stimulates platelet production. Platelets circulate with an average life span of 7–10 days. Approximately one-third of the platelets reside in the spleen, and this number increases in proportion to splenic size, although the platelet count rarely decreases to <40,000/μL as the spleen enlarges. Platelets are physiologically very active, but are anucleate, and thus have limited capacity to synthesize new proteins.

Normal vascular endothelium contributes to preventing thrombosis by inhibiting platelet function **(Chap. 61)**. When vascular endothelium is injured, these inhibitory effects are overcome, and platelets adhere to the exposed intimal surface primarily through VWF, a large multimeric protein present in both plasma and in the extracellular matrix of the subendothelial vessel wall. Platelet adhesion results in the generation of intracellular signals that lead to activation of the platelet glycoprotein (Gp) IIb/IIIa ($\alpha_{IIb}\beta_3$) receptor and resultant platelet aggregation.

Activated platelets undergo release of their granule contents, which include nucleotides, adhesive proteins, growth factors, and procoagulants that serve to promote platelet aggregation and blood clot formation and

influence the environment of the forming clot. During platelet aggregation, additional platelets are recruited to the site of injury, leading to the formation of an occlusive platelet thrombus. The platelet plug is stabilized by the fibrin mesh that develops simultaneously as the product of the coagulation cascade.

THE VESSEL WALL

Endothelial cells line the surface of the entire circulatory tree, totaling 1–6 × 10^{13} cells, enough to cover a surface area equivalent to about six tennis courts. The endothelium is physiologically active, controlling vascular permeability, flow of biologically active molecules and nutrients, blood cell interactions with the vessel wall, the inflammatory response, and angiogenesis.

The endothelium normally presents an antithrombotic surface **(Chap. 61)** but rapidly becomes prothrombotic when stimulated, which promotes coagulation, inhibits fibrinolysis, and activates platelets. In many cases, endothelium-derived vasodilators are also platelet inhibitors (e.g., nitric oxide) and, conversely, endothelium-derived vasoconstrictors (e.g., endothelin) can also be platelet activators. The net effect of vasodilation and inhibition of platelet function is to promote blood fluidity, whereas the net effect of vasoconstriction and platelet activation is to promote thrombosis. Thus, blood fluidity and hemostasis are regulated by the balance of antithrombotic/prothrombotic and vasodilatory/vasoconstrictor properties of endothelial cells.

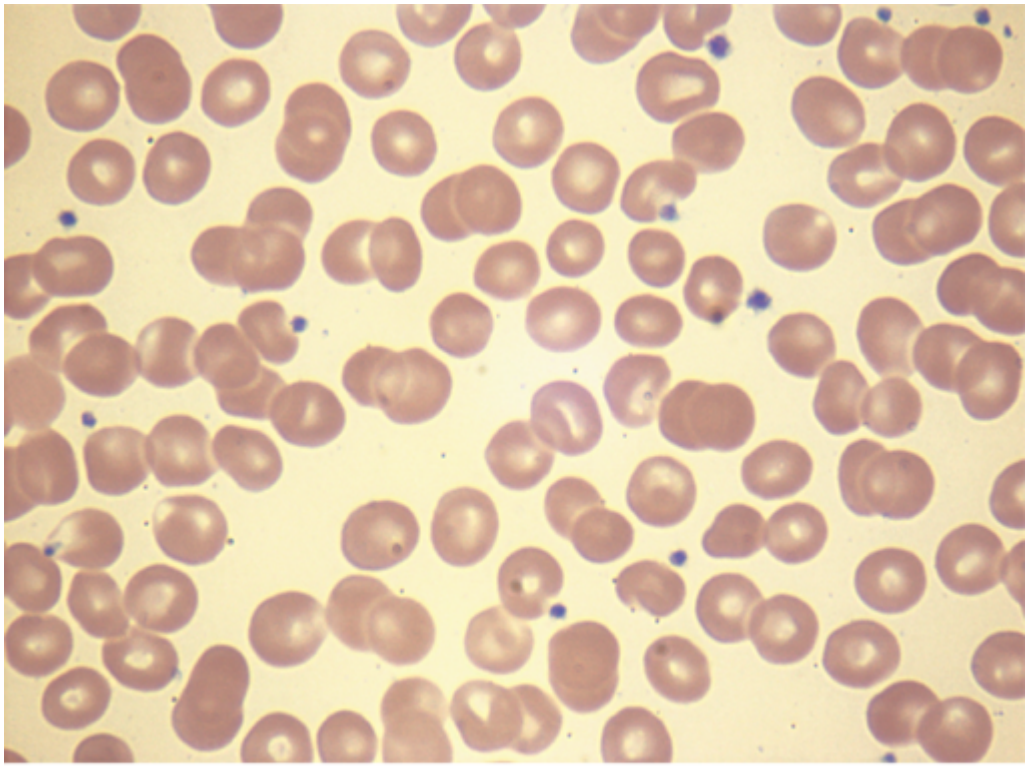
DISORDERS OF PLATELETS

THROMBOCYTOPENIA

Thrombocytopenia results from one or more of three processes: (1) decreased bone marrow production; (2) sequestration, usually in an enlarged spleen; and/or (3) increased platelet destruction. Disorders of production may be either inherited or acquired. In evaluating a patient with thrombocytopenia, a key step is to review the peripheral blood smear and to first rule out “pseudothrombocytopenia,” particularly in a patient without an apparent cause for the thrombocytopenia. Pseudothrombocytopenia **(Fig. 111-1B)** is an in vitro artifact resulting from platelet agglutination via antibodies (usually IgG, but also IgM and IgA) when the calcium content is decreased by blood collection in ethylenediamine tetraacetic (EDTA) (the anticoagulant present in tubes [purple top] used to collect blood for complete blood counts [CBCs]). If a low platelet count is obtained in EDTA-anticoagulated blood, a blood smear should be evaluated and a platelet count determined in blood collected into sodium citrate (blue top tube) or heparin (green top tube), or a smear of freshly obtained unanticoagulated blood, such as from a finger stick, can be examined.

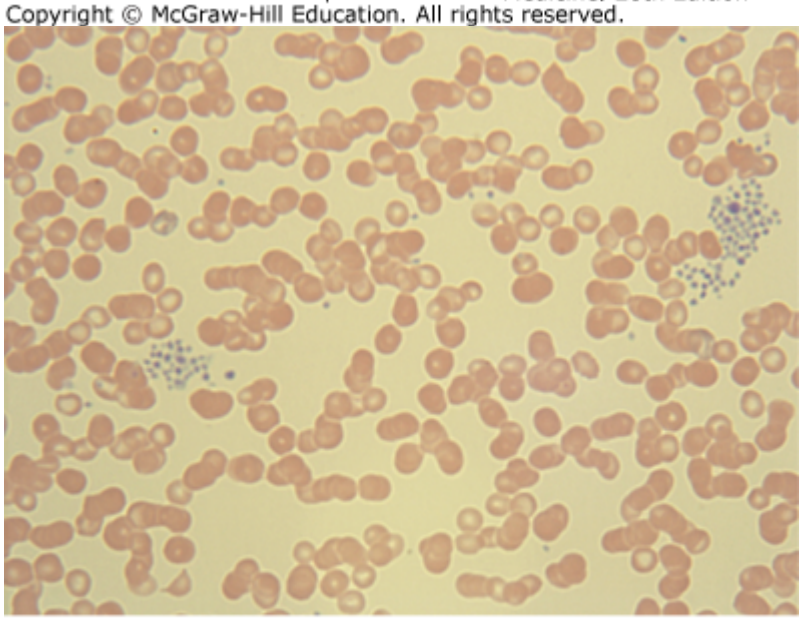
FIGURE 111-1

Photomicrographs of peripheral blood smears. ***A.*** Normal peripheral blood. ***B.*** Platelet clumping in pseudothrombocytopenia. ***C.*** Abnormal large platelet in autosomal dominant macrothrombocytopenia. ***D.*** Schistocytes and decreased platelets in microangiopathic hemolytic anemia.



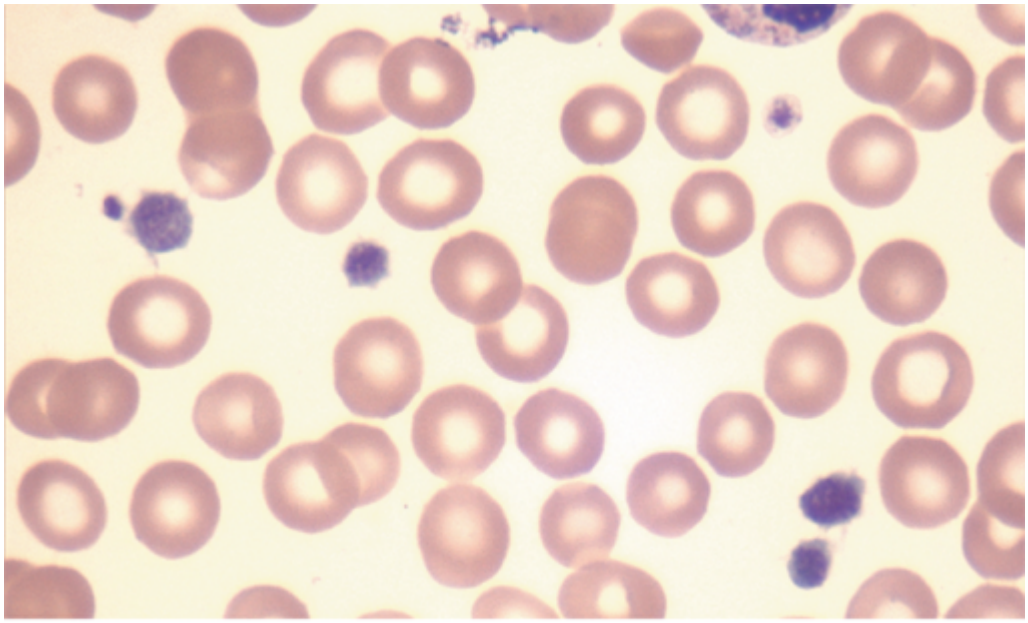
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



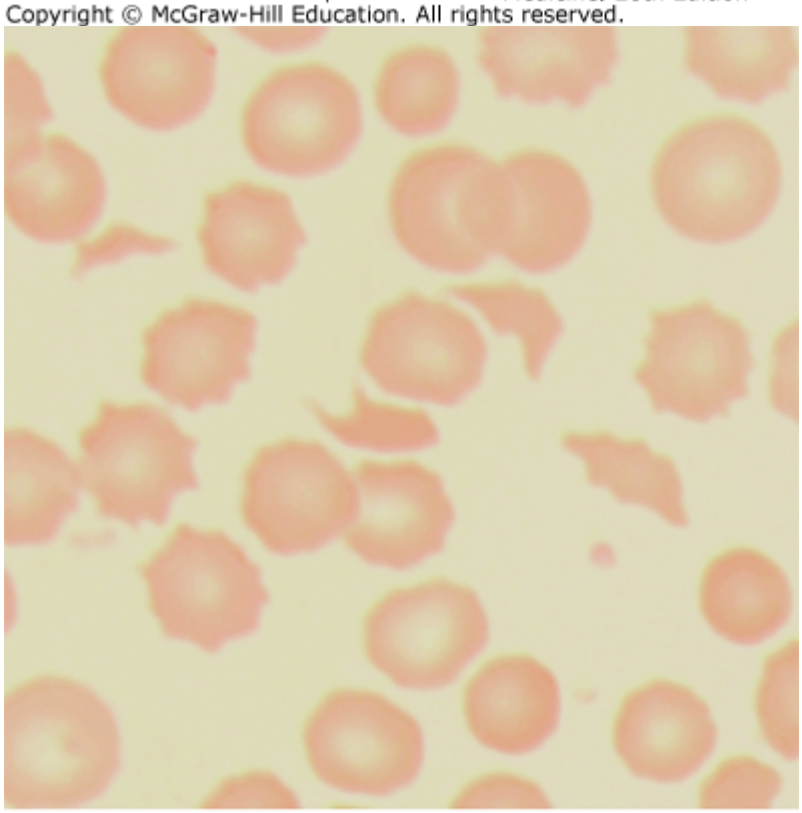
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

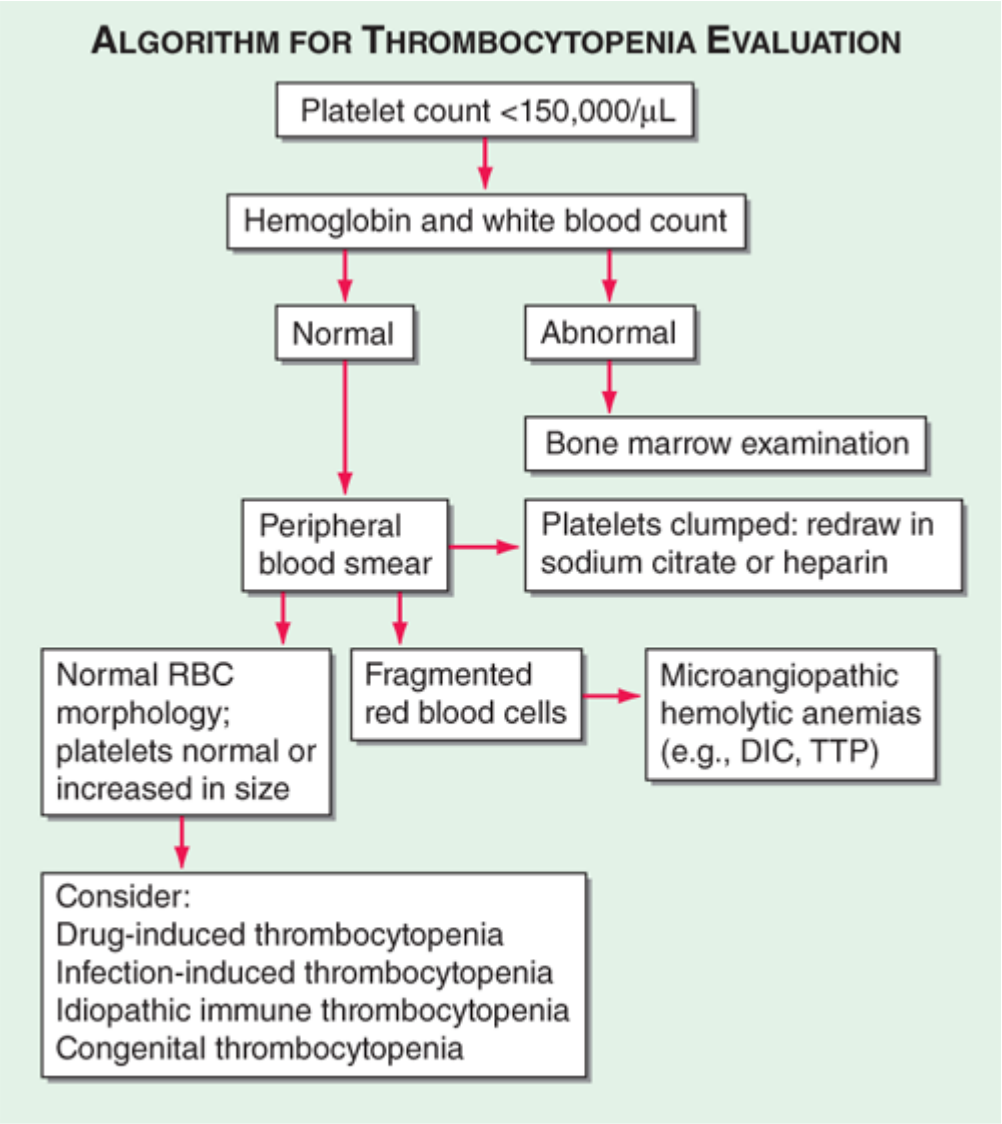
Thrombocytopenia

The history and physical examination, results of the CBC, and review of the peripheral blood smear are all critical components in the initial evaluation of thrombocytopenic patients **(Fig. 111-2)**. The overall health of the patient and whether he or she is receiving drug treatment will influence the differential diagnosis. A healthy young adult with thrombocytopenia will have a much more limited differential diagnosis than an ill hospitalized patient who is receiving multiple medications. Except in unusual inherited disorders, decreased platelet production usually results from bone marrow disorders that also affect red blood cell (RBC) and/or white blood cell (WBC) production. Because myelodysplasia can present with isolated thrombocytopenia, the bone marrow should be examined in patients presenting with isolated thrombocytopenia who are older than 60 years of age. While inherited thrombocytopenia is rare, any prior platelet counts should be retrieved and a family history regarding thrombocytopenia obtained. A careful history of drug ingestion should be obtained, including nonprescription and herbal remedies, because drugs are the most common cause of thrombocytopenia.

The physical examination can document an enlarged spleen, evidence of chronic liver disease, and other underlying disorders. Mild to moderate splenomegaly may be difficult to appreciate in many individuals due to body habitus and/or obesity but can be easily assessed by abdominal ultrasound. A platelet count of ~5000–10,000 is required to maintain vascular integrity in the microcirculation. When the count is markedly decreased, petechiae first appear in areas of increased venous pressure, the ankles and feet in an ambulatory patient. Petechiae are pinpoint, nonblanching hemorrhages and are usually a sign of a decreased platelet number and not platelet dysfunction. Wet purpura, blood blisters that form on the oral mucosa, are thought to denote an increased risk of life-threatening hemorrhage in the thrombocytopenic patient. Excessive bruising is seen in disorders of both platelet number and function.

FIGURE 111-2

Algorithm for evaluating the thrombocytopenic patient. DIC, disseminated intravascular coagulation; RBC, red blood cell; TTP, thrombotic thrombocytopenic purpura.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Infection-Induced Thrombocytopenia

Many viral and bacterial infections result in thrombocytopenia and are the most common noniatrogenic cause of thrombocytopenia. This may or may not be associated with laboratory evidence of disseminated intravascular coagulation (DIC), which is most commonly seen in patients with systemic infections with gram-negative bacteria. Infections can affect both platelet production and platelet survival. In addition, immune mechanisms can be at work, as in infectious mononucleosis and early HIV infection. Late in HIV infection, pancytopenia and decreased and dysplastic platelet production are more common. Immune-mediated thrombocytopenia in children usually follows a viral infection and almost always resolves spontaneously. This association of infection with immune thrombocytopenic purpura is less clear in adults.

Drug-Induced Thrombocytopenia

Many drugs have been associated with thrombocytopenia. A predictable decrease in platelet count occurs after treatment with many chemotherapeutic drugs due to bone marrow suppression (Chap. 69). Drugs that cause isolated thrombocytopenia and have been confirmed with positive laboratory testing are listed in Table 111-1, but all drugs should be suspect in a patient with thrombocytopenia without an apparent cause and should be stopped, or substituted, if possible. A helpful website, Platelets on the Internet (*http://www.ouhsc.edu/platelets/*), lists drugs and supplements reported to have caused thrombocytopenia and the level of evidence supporting the association. Although not as well studied, herbal and over-the-counter preparations may also result in thrombocytopenia and should be discontinued in patients who are thrombocytopenic.

TABLE 111-1

Drugs Reported as Definitely or Probably Causing Isolated Thrombocytopenia[a]

Abciximab	Mirtazapine
Acetaminophen	Naproxen
Amiodarone	Oxaliplatin
Amlodipine	Penicillin
Ampicillin	Phenytoin
Carbamazepine	Piperacillin
Ceftriaxone	Quinine
Cephamandole	Quinidine
Ciprofloxacin	Ranitidine
Diazepam	Rosiglitazone
Eptifibatide	Roxifiban
Furosemide	Sulfisoxazole
Gold	Suramin
Haloperidol	Tirofiban
Heparin	Tranilast
Ibuprofen	Trimethoprim/sulfamethoxazole
Lorazepam	Vancomycin

[a]Based on scoring requiring a compatible clinical picture and positive laboratory testing.

Source: Adapted from DM Arnold et al: J Thromb Hemost 11:169, 2013.

Classic drug-dependent antibodies are antibodies that react with specific platelet surface antigens and result in thrombocytopenia only when the drug is present. Many drugs are capable of inducing these antibodies, but for some reason, they are more common with quinine and sulfonamides. Drug-dependent antibody binding can be demonstrated by laboratory assays, showing antibody binding in the presence of, but not without, the drug present in the assay. The thrombocytopenia typically occurs after a period of initial

exposure (median length 21 days), or upon reexposure, and usually resolves in 7–10 days after drug withdrawal. The thrombocytopenia caused by the platelet Gp IIb/IIIa inhibitory drugs, such as abciximab, differs in that it may occur within 24 h of initial exposure. This appears to be due to the presence of naturally occurring antibodies that cross-react with the drug bound to the platelet.

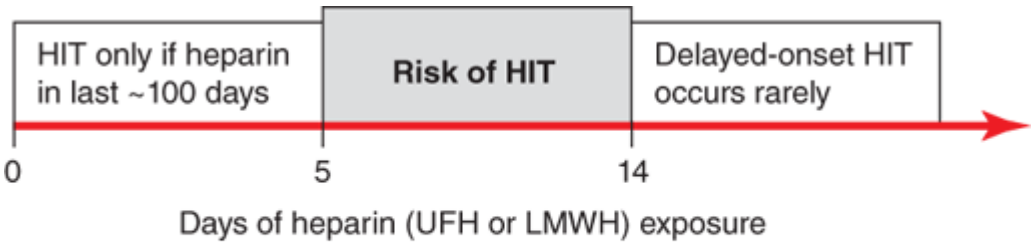
Heparin-Induced Thrombocytopenia

Drug-induced thrombocytopenia due to heparin differs from that seen with other drugs in two major ways. (1) The thrombocytopenia is not usually severe, with nadir counts rarely <20,000/μL. (2) Heparin-induced thrombocytopenia (HIT) is not associated with bleeding and, in fact, markedly increases the risk of thrombosis. HIT results from antibody formation to a complex of the platelet-specific protein platelet factor 4 (PF4) and heparin. The anti-heparin/PF4 antibody can activate platelets through the FcγRIIa receptor and also activate monocytes and endothelial cells. Many patients exposed to heparin develop antibodies to heparin/PF4, but do not appear to have adverse consequences. A fraction of those who develop antibodies will develop HIT, and a portion of those (up to 50%) will develop thrombosis (HITT).

HIT can occur after exposure to low-molecular-weight heparin (LMWH) as well as unfractionated heparin (UFH), although it is more common with the latter. Most patients develop HIT after exposure to heparin for 5–14 days **(Fig. 111-3)**. It occurs before 5 days in those who were exposed to heparin in the prior few weeks or months (<~100 days) and have circulating anti-heparin/PF4 antibodies. Rarely, thrombocytopenia and thrombosis begin several days after all heparin has been stopped (termed *delayed-onset HIT*). The “4T’s” have been recommended to be used in a diagnostic algorithm for HIT: *t*hrombocytopenia, *t*iming of platelet count drop, *thrombosis* and other sequelae such as localized skin reactions, and o*t*her causes of thrombocytopenia not evident. Application of the 4T scoring system is very useful in excluding a diagnosis of HIT but will result in over-diagnosis of HIT in situations where thrombocytopenia and thrombosis due to other etiologies are common, such as in the intensive care unit.

FIGURE 111-3

Time course of heparin-induced thrombocytopenia (HIT) development after heparin exposure. The timing of development after heparin exposure is a critical factor in determining the likelihood of HIT in a patient. HIT occurs early after heparin exposure in the presence of preexisting heparin/platelet factor 4 (PF4) antibodies, which disappear from circulation by ~100 days following a prior exposure. Rarely, HIT may occur later after heparin exposure (termed delayed-onset HIT). In this setting, heparin/PF4 antibody testing is usually markedly positive. HIT can occur after exposure to either unfractionated (UFH) or low-molecular-weight heparin (LMWH).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

LABORATORY TESTING FOR HIT

HIT (anti-heparin/PF4) antibodies can be detected using two types of assays. The most widely available is an enzyme-linked immunoassay (ELISA) with PF4/polyanion complex as the antigen. Because many patients develop antibodies but do not develop clinical HIT, the test has a low specificity for the diagnosis of HIT. This is especially true in patients who have undergone cardiopulmonary bypass surgery, where ~50% of patients develop these antibodies postoperatively. IgG-specific ELISAs increase specificity but may decrease sensitivity. The other assay is a platelet activation assay, most commonly the serotonin release assay, which measures the ability of the patient's serum to activate platelets in the presence of heparin in a concentration-dependent manner. This test has lower sensitivity but higher specificity than the ELISA. However, HIT remains a clinical diagnosis.

TREATMENT

TREATMENT

Heparin-Induced Thrombocytopenia

Early recognition is key in treatment of HIT, with prompt discontinuation of heparin and use of alternative anticoagulants if bleeding risk does not outweigh thrombotic risk. Thrombosis is a common complication of HIT, even after heparin discontinuation, and can occur in both the venous and arterial systems. Patients with higher anti-heparin/PF4 antibody titers may have a higher risk of thrombosis. In patients diagnosed with HIT, imaging studies to evaluate the patient for thrombosis (at least lower extremity duplex Doppler imaging) are recommended. Patients requiring anticoagulation should be switched from heparin to an alternative anticoagulant. The direct thrombin inhibitor (DTI) argatroban is effective in HITT. The DTI bivalirudin and the antithrombin-binding pentasaccharide fondaparinux are also effective but not approved by the U.S. Food and Drug Administration (FDA) for this indication. Danaparoid, a mixture of glycosaminoglycans with anti-Xa activity, has been used extensively for the treatment of HITT; it is no longer available in the United States but is available in other countries. HIT antibodies cross-react with LMWH, and these preparations should not be used in the treatment of HIT.

Because of the high rate of thrombosis in patients with HIT, anticoagulation should be considered, even in the absence of thrombosis. In patients with thrombosis, patients can be transitioned to warfarin, with treatment usually for 3–6 months. In patients without thrombosis, the duration of anticoagulation needed is undefined. An increased risk of thrombosis is present for at least 1 month after diagnosis; however, most thromboses occur early, and whether thrombosis occurs later if the patient is initially anticoagulated is unknown. Options include continuing anticoagulation until a few days after platelet recovery or for 1 month.

Introduction of warfarin alone in the setting of HIT or HITT may precipitate thrombosis, particularly venous gangrene, presumably due to clotting activation and severely reduced levels of proteins C and S. Warfarin therapy, if started, should be overlapped with a DTI or fondaparinux and started after resolution of the thrombocytopenia and lessening of the prothrombotic state. Evidence for use of an oral direct Xa inhibitor in this setting is growing, but more data are needed to establish efficacy.

Immune Thrombocytopenic Purpura

Immune thrombocytopenic purpura (ITP; also termed *idiopathic thrombocytopenic purpura*) is an acquired disorder in which there is immune-mediated destruction of platelets and possibly inhibition of platelet release from the megakaryocyte. In children, it is usually an acute disease, most commonly following an infection, and with a self-limited course. In adults, it is a more chronic disease, although in some adults, spontaneous remission occurs, usually within months of diagnosis. ITP is termed *secondary* if it is associated with an underlying disorder; autoimmune disorders, particularly systemic lupus erythematosus (SLE), and infections, such as HIV and hepatitis C, are common causes. The association of ITP with *Helicobacter pylori* infection is unclear but appears to have a geographic distribution.

ITP is characterized by mucocutaneous bleeding and a low, often very low, platelet count, with an otherwise normal peripheral blood cells and smear. Patients usually present either with ecchymoses and petechiae, or with thrombocytopenia incidentally found on a routine CBC. Mucocutaneous bleeding, such as oral mucosa, gastrointestinal, or heavy menstrual bleeding, may be present. Rarely, life-threatening, including central nervous system, bleeding can occur. Wet purpura (blood blisters in the mouth) and retinal hemorrhages may herald life-threatening bleeding.

LABORATORY TESTING IN ITP

Laboratory testing for antibodies (serologic testing) is usually not helpful due to the low sensitivity and specificity of the current tests. Bone marrow examination can be reserved for those who have other signs or laboratory abnormalities not explained by ITP or in patients who do not respond to initial therapy. The peripheral blood smear may show large platelets, with otherwise normal morphology. Depending on the bleeding history, iron-deficiency anemia may be present.

Laboratory testing is performed to evaluate for secondary causes of ITP and should include testing for HIV infection and hepatitis C (and other infections if indicated). Serologic testing for SLE, serum protein electrophoresis, immunoglobulin levels to potentially detect hypogammaglobulinemia, selective testing for IgA deficiency or monoclonal gammopathies, and testing for *H. pylori* infection should be considered, depending on the clinical circumstance. If anemia is present, direct antiglobulin testing (Coombs' test) should be performed to rule out combined autoimmune hemolytic anemia with ITP (Evans' syndrome).

TREATMENT

TREATMENT

Immune Thrombocytopenic Purpura

The treatment of ITP uses drugs that decrease reticuloendothelial uptake of the antibody-bound platelet, decrease antibody production, and/or increase platelet production. The diagnosis of ITP does not necessarily mean that treatment must be instituted. Patients with platelet counts >30,000/μL appear not to have increased mortality related to the thrombocytopenia.

Initial treatment in patients without significant bleeding symptoms, severe thrombocytopenia (<5000/μL), or signs of impending bleeding (such as retinal hemorrhage or large oral mucosal hemorrhages) can be instituted as an outpatient using single agents. Traditionally, this has been prednisone at 1 mg/kg, although Rh_0(D) immune globulin therapy (WinRho SDF), at 50–75 μg/kg, is also being used in this setting. Rh_0(D) immune globulin must be used only in Rh-positive patients because the mechanism of action is production of limited hemolysis, with antibody-coated cells “saturating” the Fc receptors, inhibiting Fc receptor function. Monitoring patients for 8 h after infusion is now advised by the FDA because of the rare complication of severe intravascular hemolysis. Intravenous gamma globulin (IVIgG), which is pooled, primarily IgG antibodies, also blocks the Fc receptor system, but appears to work primarily through different mechanism(s). IVIgG has more efficacy than anti-Rh_0(D) in postsplenectomized patients. IVIgG is dosed at 1–2 g/kg total, given over 1–5 days. Side effects are usually related to the volume of infusion and infrequently include aseptic meningitis and renal failure. All immunoglobulin preparations are derived from human plasma and undergo treatment for viral inactivation.

For patients with severe ITP and/or symptoms of bleeding, hospital admission and combined-modality therapy is given using high-dose glucocorticoids with IVIgG or anti-Rh_0(D) therapy and, as needed, additional immunosuppressive agents. Rituximab, an anti-CD20 (B cell) antibody, has shown efficacy in the treatment of refractory ITP, although long-lasting remission only occurs in ∼30% of patients.

Splenectomy has been used for treatment of patients who relapse after glucocorticoids are tapered. Splenectomy remains an important treatment option; however, more patients than previously thought will go into a remission over time. Observation, if the platelet count is high enough, or intermittent treatment with anti-Rh_0(D) or IVIgG, or initiation of treatment with a TPO receptor agonist (see below) may be a reasonable approach to see if the ITP will resolve. Vaccination against encapsulated organisms (especially pneumococcus, but also meningococcus and *Haemophilus influenzae*, depending on patient age and potential exposure) is recommended before splenectomy. Accessory spleen(s) are a very rare cause of relapse.

TPO receptor agonists are available for the treatment of ITP. This approach stems from the finding that many patients with ITP do not have increased TPO levels, as was previously hypothesized. TPO levels reflect megakaryocyte mass, which is usually normal in ITP. TPO levels are not increased in the setting of platelet destruction. Two agents, one administered subcutaneously (romiplostim) and another orally (eltrombopag), are effective in raising platelet counts in patients with ITP and are recommended for adults at risk of bleeding who relapse after splenectomy or who have been unresponsive to at least one other therapy, particularly in those who have a contraindication to splenectomy. However, with the recognition that ITP will resolve spontaneously in some adult patients, short-term treatment with a TPO agonist can be considered before

splenectomy in patients who need therapy. Eltrombopag is FDA approved for use in children over 1 year of age. Romiplostim is not yet FDA approved in children but a randomized trial supports efficacy.

Inherited Thrombocytopenia

Thrombocytopenia is rarely inherited, either as an isolated finding or as part of a syndrome, and may be inherited in an autosomal dominant, autosomal recessive, or X-linked pattern. Many forms of autosomal dominant thrombocytopenia are now known to be associated with mutations in the nonmuscle myosin heavy chain *MYH9* gene. Interestingly, these include the May-Hegglin anomaly, and Sebastian, Epstein's, and Fechtner syndromes, all of which have distinct distinguishing features. A common feature of these disorders is large platelets (Fig. 111-1C). Autosomal recessive disorders include congenital amegakaryocytic thrombocytopenia, thrombocytopenia with absent radii, and Bernard-Soulier syndrome. The latter is primarily a functional platelet disorder due to absence of Gp Ib-IX-V, the VWF adhesion receptor. X-linked disorders include Wiskott-Aldrich syndrome and a dyshematopoietic syndrome resulting from a mutation in *GATA-1*, an important transcriptional regulator of hematopoiesis.

THROMBOTIC THROMBOCYTOPENIC PURPURA AND HEMOLYTIC-UREMIC SYNDROME

Thrombotic thrombocytopenic microangiopathies are a group of disorders characterized by thrombocytopenia, a microangiopathic hemolytic anemia evident by fragmented RBCs (Fig. 111-1D) and laboratory evidence of hemolysis, and microvascular thrombosis. They include thrombotic thrombocytopenic purpura (TTP) and hemolytic-uremic syndrome (HUS), as well as syndromes complicating bone marrow transplantation, certain medications and infections, pregnancy, and vasculitis. In DIC, although thrombocytopenia and microangiopathy are seen, a coagulopathy predominates, with consumption of clotting factors and fibrinogen resulting in an elevated prothrombin time (PT) and often activated partial thromboplastin time (aPTT). The PT and aPTT are characteristically normal in TTP or HUS.

Thrombotic Thrombocytopenic Purpura

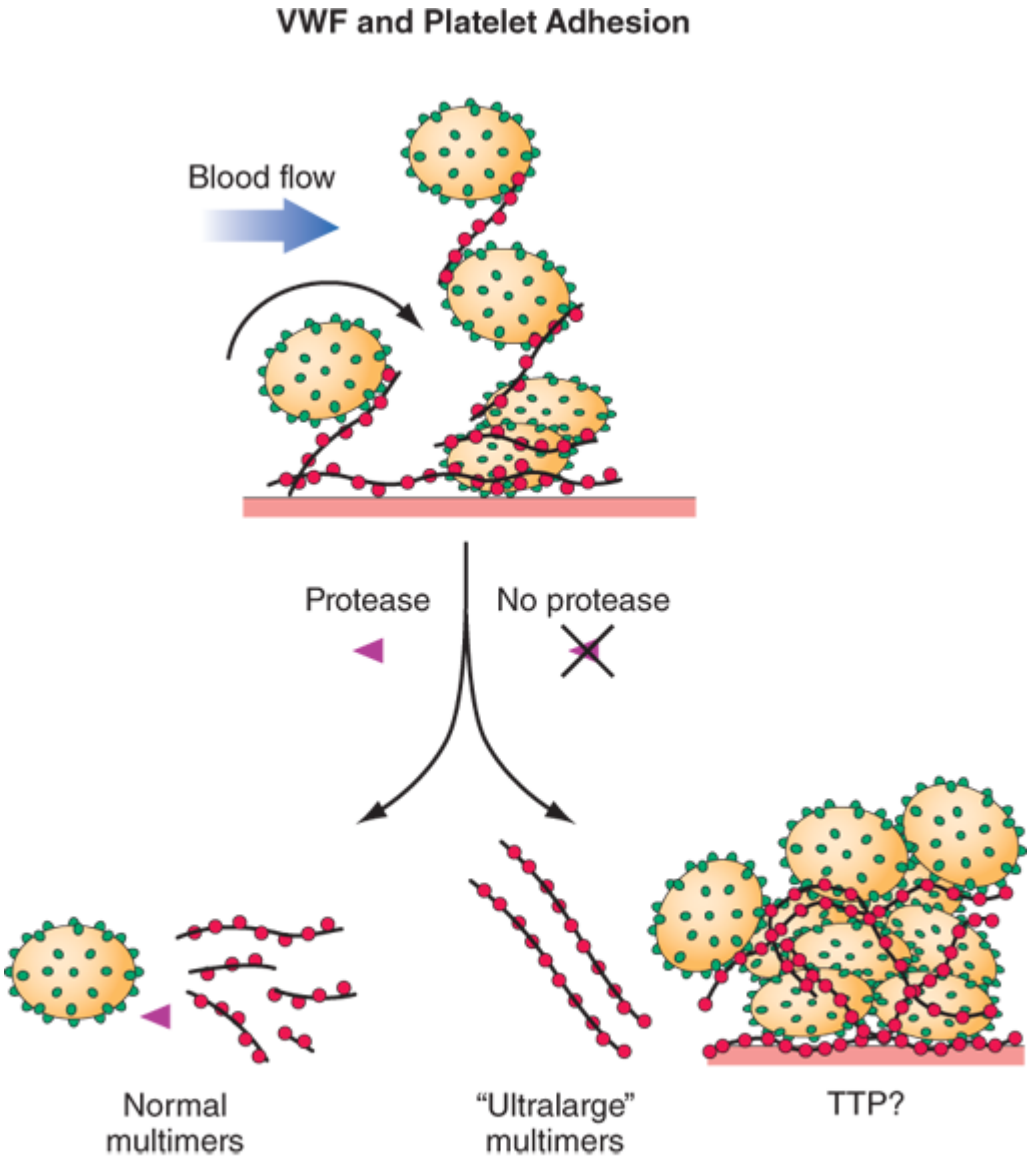
TTP and HUS were previously considered overlap syndromes. However, with better understanding of the pathophysiology of both TTP and HUS they are clearly separate entities. TTP was first described in 1924 by Eli Moschcowitz and characterized by a pentad of findings that include microangiopathic hemolytic anemia, thrombocytopenia, renal failure, neurologic findings, and fever. The full-blown syndrome is less commonly seen now, probably due to earlier diagnosis. The introduction of treatment with plasma exchange markedly improved the prognosis in patients, with a decrease in mortality from 85–100% to 10–30%.

The pathogenesis of inherited (Upshaw-Schulman syndrome) and idiopathic TTP is related to a deficiency of, or antibodies to, the metalloprotease ADAMTS13, which cleaves VWF. VWF is normally secreted as ultra-large multimers, which are then cleaved by ADAMTS13. The persistence of ultra-large VWF molecules is thought to contribute to pathogenic platelet adhesion and aggregation (**Fig. 111-4**). This defect alone, however, is not sufficient to result in TTP because individuals with a congenital absence of ADAMTS13 develop TTP only

episodically. The level of ADAMTS13 activity, as well as antibodies, can be detected by laboratory assays. ADAMTS13 activity levels of <10% are more clearly associated with idiopathic (antibody-mediated) TTP.

FIGURE 111-4

Pathogenesis of thrombotic thrombocytopenic purpura (TTP). Normally the ultra-high-molecular-weight multimers of von Willebrand factor (VWF) produced by the endothelial cells are processed into smaller multimers by a plasma metalloproteinase called ADAMTS13. In TTP, the activity of the protease is inhibited, and the ultra-high-molecular-weight multimers of VWF initiate platelet aggregation and thrombosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Idiopathic TTP appears to be more common in women than in men. No geographic or racial distribution has been defined. TTP is more common in patients with HIV infection and in pregnant women. Medication-related microangiopathic hemolytic anemia may be secondary to antibody formation (ticlopidine and possibly clopidogrel) or direct endothelial toxicity (cyclosporine, mitomycin C, tacrolimus, quinine), although this is not always so clear, and fear of withholding treatment, as well as lack of other treatment alternatives,

may result in initial application of plasma exchange. However, withdrawal, or reduction in dose, of endothelial toxic agents usually decreases the microangiopathy.

Treatment

TREATMENT

Thrombotic Thrombocytopenic Purpura

TTP is a devastating disease if not diagnosed and treated promptly. In patients presenting with new thrombocytopenia, with or without evidence of renal insufficiency and other elements of classic TTP, laboratory data should be obtained to rule out DIC and to evaluate for evidence of microangiopathic hemolytic anemia. Findings to support the TTP diagnosis include an increased lactate dehydrogenase and indirect bilirubin, decreased haptoglobin, and increased reticulocyte count, with a negative direct antiglobulin test. The peripheral smear should be examined for evidence of schistocytes (Fig. 111-1D). Polychromasia is usually also present due to the increased number of young red blood cells, and nucleated RBCs are often present, which is thought to be due to infarction in the microcirculatory system of the bone marrow.

Plasma exchange remains the mainstay of treatment of TTP. ADAMTS13 antibody-mediated TTP (idiopathic TTP) appears to respond best to plasma exchange. Plasma exchange is continued until the platelet count is normal and signs of hemolysis are resolved for at least 2 days. Although never evaluated in clinical trials, the use of glucocorticoids seems a reasonable approach, but should only be used as an adjunct to plasma exchange. Additionally, other immunomodulatory therapies have been reported to be successful in refractory or relapsing TTP, including rituximab, vincristine, cyclophosphamide, and splenectomy, with rituximab having the most evidence for efficacy. A significant relapse rate is noted; 25–45% of patients relapse within 30 days of initial “remission,” and 12–40% of patients have late relapses. Relapses are more frequent in patients with severe ADAMTS13 deficiency at presentation.

Hemolytic-Uremic Syndrome

HUS is a syndrome characterized by acute renal failure, microangiopathic hemolytic anemia, and thrombocytopenia. It is seen preceded by an episode of diarrhea, often hemorrhagic in nature, predominantly in children. *Escherichia coli* O157:H7 is the most frequent, although not only, etiologic serotype. HUS not associated with diarrhea is more heterogeneous in presentation and course. Atypical HUS (aHUS) is usually due to genetic defects that result in chronic complement activation or antibodies directed against complement regulatory proteins. Laboratory testing for DNA variants in complement regulatory genes is available, although assigning pathogenicity to variants remains challenging. Currently there is not a functional assay that is diagnostic of the disease.

Treatment

TREATMENT

Hemolytic-Uremic Syndrome

Treatment of HUS is primarily supportive. In HUS associated with diarrhea, many (~40%) children require at least some period of support with dialysis; however, the overall mortality is <5%. In HUS not associated with diarrhea, the mortality is higher, ~26%. Plasma infusion or plasma exchange has not been shown to alter the overall course. ADAMTS13 levels are generally reported to be normal in HUS, although occasionally they have been reported to be decreased. In patients with aHUS, eculizumab, a humanized monoclonal antibody against C5 that blocks terminal complement, has efficacy in resolution of HUS and improving or preserving renal function. Patients with aHUS may initially be treated with plasma exchange, until the ADAMTS13 level is returned and the diagnosis more clear, since aHUS remains a diagnosis of exclusion. However, plasma exchange has not been shown to affect clinical outcomes in aHUS.

THROMBOCYTOSIS

Thrombocytosis is almost always due to (1) iron deficiency; (2) inflammation, cancer, or infection (reactive thrombocytosis); or (3) an underlying myeloproliferative process (essential thrombocythemia or polycythemia vera) **(Chap. 99)** or, rarely, the 5q– myelodysplastic process **(Chap. 98)**. Patients presenting with an elevated platelet count should be evaluated for underlying inflammation or malignancy, and iron deficiency should be ruled out. Thrombocytosis in response to acute or chronic inflammation has not been clearly associated with an increased thrombotic risk. In fact, patients with markedly elevated platelet counts (>1.5 million), usually seen in the setting of a myeloproliferative disorder, have an increased risk of bleeding. This appears to be due, at least in part, to acquired von Willebrand disease (VWD) due to platelet-VWF binding and removal from the circulation.

QUALITATIVE DISORDERS OF PLATELET FUNCTION

Inherited Disorders of Platelet Function

Inherited platelet function disorders are thought to be relatively rare, although the prevalence of mild disorders of platelet function is unclear, in part because our testing for such disorders is suboptimal. Rare qualitative disorders include the autosomal recessive disorders Glanzmann's thrombasthenia (absence of the platelet Gp IIb/IIIa receptor) and Bernard-Soulier syndrome (absence of the platelet Gp Ib-IX-V receptor). Both are inherited in an autosomal recessive fashion and present with bleeding symptoms in childhood.

Platelet storage pool disorder (SPD) is the classic autosomal dominant qualitative platelet disorder. This results from abnormalities of platelet granule formation. It is also seen as a part of inherited disorders of granule formation, such as Hermansky-Pudlak syndrome. Bleeding symptoms in SPD are variable, but often are mild. The most common inherited disorders of platelet function prevent normal secretion of granule content and are termed *secretion defects*. Few of these abnormalities have been dissected at the molecular level, but they likely result from various DNA variants.

Treatment

TREATMENT

Inherited Disorders of Platelet Dysfunction

Bleeding symptoms or prevention of bleeding in patients with severe platelet dysfunction frequently requires platelet transfusion. Care is taken to limit the risk of alloimmunization by limiting exposure, using HLA-matched leuko-depleted platelet concentrates for transfusion. Platelet disorders associated with milder bleeding symptoms frequently respond to desmopressin (1-deamino-8-D-arginine vasopressin [DDAVP]). DDAVP increases plasma VWF and factor VIII levels; it may also have a direct effect on platelet function. Particularly for mucosal bleeding symptoms, antifibrinolytic therapy (ε-aminocaproic acid or tranexamic acid) is used alone or in conjunction with DDAVP or platelet therapy.

Acquired Disorders of Platelet Function

Acquired platelet dysfunction is common, usually due to medications, either intentionally as with antiplatelet therapy or unintentionally as with high-dose penicillins. Acquired platelet dysfunction occurs in uremia. This is likely multifactorial, but the resultant effect is defective adhesion and activation. The platelet defect is improved most by dialysis but may also be improved by increasing the hematocrit to 27–32%, giving DDAVP (0.3 μg/kg), or use of conjugated estrogens. Platelet dysfunction also occurs with cardiopulmonary bypass due to the effect of the artificial circuit on platelets, and bleeding symptoms respond to platelet transfusion. Platelet dysfunction seen with underlying hematologic disorders can result from nonspecific interference by circulating paraproteins or intrinsic platelet defects in myeloproliferative and myelodysplastic syndromes.

VON WILLEBRAND DISEASE

VWD is the most common inherited bleeding disorder. Estimates from laboratory data suggest a prevalence of ~1%, but data based on symptomatic individuals suggest that it is closer to 0.1% of the population. VWF serves two roles: (1) as the major adhesion molecule that tethers the platelet to the exposed subendothelium; and (2) as the binding protein for factor VIII (FVIII), resulting in significant prolongation of the FVIII half-life in circulation. The platelet-adhesive function of VWF is critically dependent on the presence of large VWF multimers, whereas FVIII binding is not. Most of the symptoms of VWD are “platelet-like” except in more severe VWD when the FVIII is low enough to produce symptoms similar to those found in FVIII deficiency (hemophilia A).

VWD has been classified into three major types, with four subtypes of type 2 **(Table 111-2; Fig. 111-5)**. By far the most common type of VWD is type 1 disease, with a parallel decrease in VWF protein, VWF function, and FVIII levels, accounting for at least 80% of cases. Patients have predominantly mucosal bleeding symptoms, although postoperative bleeding can also be seen. Bleeding symptoms are very uncommon in infancy and usually manifest later in childhood with excessive bruising and epistaxis. Because these symptoms occur commonly in childhood, the clinician should particularly note bruising at sites unlikely to be traumatized and/or prolonged epistaxis requiring medical attention. Menorrhagia is a common manifestation of VWD.

Menstrual bleeding resulting in anemia should warrant an evaluation for VWD and, if negative, functional platelet disorders. Frequently, mild type 1 VWD first manifests with dental extractions, particularly wisdom tooth extraction, or tonsillectomy.

TABLE 111-2

Laboratory Diagnosis of von Willebrand Disease (VWD)

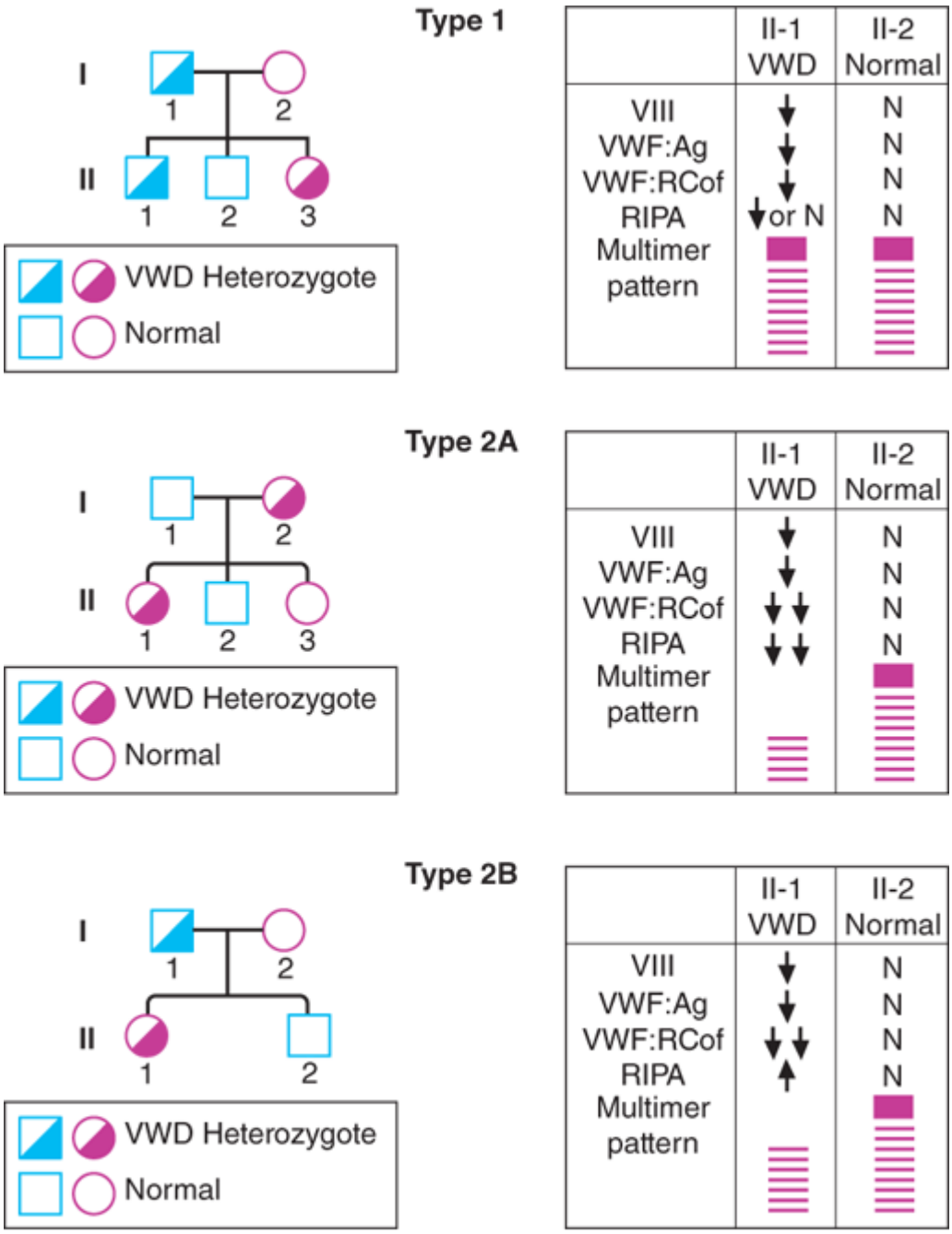
Type	aPTT	VWF Antigen	VWF Activity	FVIII Activity	Multimer
1	Nl or ↑	↓	↓	↓	Normal distribution, decreased in quantity
2A	Nl or ↑	↓	↓↓	↓	Loss of high- and intermediate-MW multimers
2B[a]	Nl or ↑	↓	↓↓	↓	Loss of high-MW multimers
2M	Nl or ↑	↓	↓↓	↓	Normal distribution, decreased in quantity
2N	↑↑	Nl or ↓[b]	Nl or ↓[b]	↓↓	Normal distribution
3	↑↑	↓↓	↓↓	↓↓	Absent

[a]Usually also decreased platelet count. [b]For type 2N, in the homozygous state, FVIII is very low; in the heterozygous state, it is only seen in conjunction with type 1 VWD.

Abbreviations: aPTT, activated partial thromboplastin time; F, factor; MW, molecular weight; Nl, normal; VWF, von Willebrand factor.

FIGURE 111-5

Pattern of inheritance and laboratory findings in von Willebrand disease (VWD). The assays of platelet function include a coagulation assay of factor VIII bound and carried by von Willebrand factor (VWF), abbreviated as VIII; immunoassay of total VWF protein (VWF:Ag); bioassay of the ability of patient plasma to support ristocetin-induced agglutination of normal platelets (VWF:RCoF), a measure of VWF activity; and ristocetin-induced aggregation of patient platelets, abbreviated RIPA. The multimer pattern illustrates the multimer bands present when plasma is electrophoresed in a polyacrylamide gel. The II-1 and II-2 columns refer to the phenotypes of the second-generation offspring.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Not all patients with low VWF levels have bleeding symptoms. Whether patients bleed or not will depend on the overall hemostatic balance they have inherited, along with environmental influences and the type of hemostatic challenges they experience. Although the inheritance of VWD is autosomal, many factors modulate both VWF levels and bleeding symptoms. These have not all been defined, but include blood type, thyroid hormone status, race, stress, exercise, hormonal (both endogenous and exogenous) influences, and modulators of VWF clearance. Patients with type O blood have VWF protein levels of approximately one-half that of patients with AB blood type; and, in fact, the normal range for patients with type O blood overlaps that which has been considered diagnostic for VWD. A mildly decreased VWF level should be viewed more as a risk factor for bleeding than as an actual disease.

Patients with type 2 VWD have functional defects; thus, the VWF antigen measurement is significantly higher than the test of function. For types 2A, 2B, and 2M VWD, platelet-binding and/or collagen-binding VWF activity is decreased. In type 2A VWD, the impaired function is due either to increased susceptibility to cleavage by ADAMTS13, resulting in loss of intermediate- and high-molecular-weight multimers, or to decreased production of these multimers by the cell. Type 2B VWD results from gain-of-function mutations that result in increased spontaneous binding of VWF to platelets in circulation, with increased cleavage by ADAMTS13 and clearance of this complex by the reticuloendothelial system. The resulting VWF in the patients' plasma lacks the highest molecular-weight multimers, and the platelet count is usually modestly reduced. Type 2M occurs as a consequence of a group of DNA variants that cause dysfunction but do not affect multimer structure.

Type 2N VWD is due to variants in the *VWF* gene that affect binding of FVIII. As FVIII is stabilized by binding to VWF, the FVIII in patients with type 2N VWD has a very short half-life, and the FVIII level is markedly decreased. This is sometimes termed *autosomal hemophilia*. Type 3 VWD, or severe VWD, describes patients with virtually no VWF protein and FVIII levels <10%. Patients experience mucosal and joint bleeding, surgery-related bleeding, and other bleeding symptoms. Some patients with type 3 VWD, particularly those with large VWF gene deletions, are at risk of developing antibodies to infused VWF.

Acquired VWD is a rare disorder, most commonly seen in patients with underlying lymphoproliferative disorders, including monoclonal gammopathies of underdetermined significance (MGUS), multiple myeloma, and Waldenström's macroglobulinemia. It is seen most commonly in the setting of MGUS and should be suspected in patients, particularly elderly patients, with a new onset of severe mucosal bleeding symptoms. Laboratory evidence of acquired VWD is found in some patients with aortic valvular disease. Heyde's syndrome (aortic stenosis with gastrointestinal bleeding) is attributed to the presence of angiodysplasia of the gastrointestinal tract in patients with aortic stenosis. The shear stress on blood passing through the stenotic aortic valve appears to unfold VWF, making it susceptible to proteolysis. Consequently, large multimer forms are lost, leading to an acquired type 2 VWD, but return when the stenotic valve is replaced.

TREATMENT

TREATMENT

von Willebrand Disease

The mainstay of treatment for type 1 VWD is DDAVP (desmopressin), which results in release of VWF and FVIII from endothelial stores. DDAVP can be given intravenously or by a high-concentration intranasal spray (1.5 mg/mL). The peak activity when given intravenously is ~30 min, whereas it is 2 h when given intra-nasally. The usual dose is 0.3 μg/kg intravenously or two squirts (one in each nostril) for patients >50 kg (one squirt for those <50 kg). It is recommended that patients with VWD be tested with DDAVP to assess their response before using it. In patients who respond well (increase in laboratory values of two- to fourfold), it can be used for procedures with minor to moderate risk of bleeding. Depending on the procedure, additional doses may

be needed; it is usually given every 12–24 h. Less frequent dosing may result in less tachyphylaxis, which occurs when synthesis cannot compensate for the released stores. The major side effect of DDAVP is hyponatremia due to decreased free water clearance. This occurs most commonly in the very young and the very old, but fluid restriction should be advised for all patients for the 24 h following each dose.

Some patients with types 2A and 2M VWD respond to DDAVP such that it can be used for minor procedures. For the other subtypes, for type 3 disease, and for major procedures requiring longer periods of normal hemostasis, VWF replacement can be given. Virally inactivated VWF-plasma-derived and recombinant factor concentrates are safer than cryoprecipitate as the replacement product.

Antifibrinolytic therapy using either ε-aminocaproic acid or tranexamic acid is an important therapy, either alone or in an adjunctive capacity, particularly for the prevention or treatment of mucosal bleeding. These agents are particularly useful in treatment of menorrhagia and post partum hemorrhage, prophylaxis for dental procedures, and with DDAVP or factor concentrate for dental extractions, tonsillectomies, and prostate procedures. Antifibrinolytic agents are contraindicated in the setting of upper urinary tract bleeding, due to the risk of ureteral obstruction.

DISORDERS OF THE VESSEL WALL

The vessel wall is an integral part of hemostasis, and separation of a fluid phase is artificial, particularly in disorders such as TTP or HIT that clearly involve the endothelium as well. Inflammation localized to the vessel wall, such as vasculitis, and inherited connective tissue disorders are abnormalities inherent to the vessel wall.

METABOLIC AND INFLAMMATORY DISORDERS

Acute febrile illnesses may result in vascular damage. This can result from immune complexes containing viral antigens or the viruses themselves. Certain pathogens, such as the rickettsiae causing Rocky Mountain spotted fever, replicate in endothelial cells and damage them. Vascular purpura may occur in patients with polyclonal gammopathies but more commonly in those with monoclonal gammopathies, including Waldenström's macroglobulinemia, multiple myeloma, and cryoglobulinemia. Patients with mixed cryoglobulinemia develop a more extensive maculopapular rash due to immune complex–mediated damage to the vessel wall.

Patients with scurvy (vitamin C deficiency) develop painful episodes of perifollicular skin bleeding as well as more systemic bleeding symptoms. Vitamin C is needed to synthesize hydroxyproline, an essential constituent of collagen. Patients with Cushing's syndrome or on chronic glucocorticoid therapy develop skin bleeding and easy bruising due to atrophy of supporting connective tissue. A similar phenomenon is seen with aging, where following minor trauma, blood spreads superficially under the epidermis. This has been termed *senile purpura*. It is most common on skin that has been previously damaged by sun exposure.

Henoch-Schönlein, or anaphylactoid, purpura is a distinct, self-limited type of vasculitis that occurs in children and young adults. Patients have an acute inflammatory reaction with IgA and complement components in capillaries, mesangial tissues, and small arterioles leading to increased vascular permeability

and localized hemorrhage. The syndrome is often preceded by an upper respiratory infection, commonly with streptococcal pharyngitis, or is triggered by drug or food allergies. Patients develop a purpuric rash on the extensor surfaces of the arms and legs, usually accompanied by polyarthralgias or arthritis, abdominal pain, and hematuria from focal glomerulonephritis. All coagulation tests are normal, but renal impairment may occur. Glucocorticoids can provide symptomatic relief but do not alter the course of the illness.

INHERITED DISORDERS OF THE VESSEL WALL

Patients with inherited disorders of the connective tissue matrix, such as Marfan's syndrome, Ehlers-Danlos syndrome, and pseudoxanthoma elasticum, frequently report easy bruising. Inherited vascular abnormalities can result in increased bleeding. This is notably seen in hereditary hemorrhagic telangiectasia (HHT, or Osler-Weber-Rendu disease), a disorder where abnormal telangiectatic capillaries result in frequent bleeding episodes, primarily from the nose and gastrointestinal tract. Arteriovenous malformation (AVM) in the lung, brain, and liver may also occur in HHT. The telangiectasia can often be visualized on the oral and nasal mucosa. Signs and symptoms develop over time. Epistaxis begins, on average, at the age of 12 and occurs in >95% of affected individuals by middle age. Approximately 25% have GI bleeding usually beginning after the age of 50. HHT is caused by pathogenic DNA variants in number of genes involved in the TGFβ/BMP signaling cascade.

FURTHER READING

Arnold DM: Bleeding complications in immune thrombocytopenia. Hematology 2015:237, 2015.

Aster RH et al: Drug-induced immune thrombocytopenia: Pathogenesis, diagnosis and management. J Thromb Haemost 7:911, 2009.

Boender J et al: A diagnostic approach to mild bleeding disorders. J Thromb Haemost 14:1507, 2016.

Brodsky RA: Complement in haemolytic anemia. Blood 126:2459, 2015.

Gresele P et al: Diagnosis of inherited platelet function disorders: Guidance from the SSC of the ISTH. J Thromb Haemost 13:314, 2014.

Hunt BJ: Bleeding and coagulopathies in critical care. N Engl J Med 370:847, 2014.

Kremer Hovinga JA et al: Survival and relapse in patients with thrombotic thrombocytopenic purpura. Blood 115:1500, 2010.

Leebeek FW, Eikenboom JC: von Willebrand's disease. N Engl J Med 375:2067, 2016.

Warkentin TE, Greinacher A: Management of heparin-induced thrombocytopenia. Curr Opin Hematol 23:462, 2016.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 112: Coagulation Disorders

Valder R. Arruda; Katherine A. High

INTRODUCTION

Deficiencies of coagulation factors have been recognized for centuries. Patients with genetic deficiencies of plasma coagulation factors exhibit lifelong recurrent bleeding episodes into joints, muscles, and closed spaces, either spontaneously or following an injury. The most common inherited factor deficiencies are the hemophilias, X-linked diseases caused by deficiency of factor (F) VIII (hemophilia A) or FIX (hemophilia B). Rare congenital bleeding disorders due to deficiencies of other factors, including FII (prothrombin), FV, FVII, FX, FXI, FXIII, and fibrinogen, are commonly inherited in an autosomal recessive manner (Table 112-1). Advances in characterization of the molecular bases of clotting factor deficiencies have contributed to better understanding of the disease phenotypes and may eventually allow more targeted therapeutic approaches through the development of small molecules, recombinant proteins, or cell- and gene-based therapies.

TABLE 112-1

Genetic and Laboratory Characteristics of Inherited Coagulation Disorders

Clotting Factor Deficiency	Inheritance	Prevalence in General Population	Laboratory Abnormality[a]			Minimum Hemostatic Levels	Treatment	Plasma Half-Life
			aPTT	PT	TT			
Fibrinogen	AR	1 in 1,000,000	+	+	+	100 mg/dL	Cryoprecipitate	2–4 d
Prothrombin	AR	1 in 2,000,000	+	+	−	20–30%	FFP/PCC	3–4 d
Factor V	AR	1 in 1,000,000	+/−	+/−	−	15–20%	FFP[c]	36 h
Factor VII	AR	1 in 500,000	−	+	−	15–20%	FFP/PCC	4–6 h
Factor VIII	X-linked	1 in 5,000	+	−	−	30%	FVIII concentrates	8–12 h
Factor IX	X-linked	1 in 30,000	+	−	−	30%	FIX concentrates	18–24 h
Factor X	AR	1 in 1,000,000	+/−	+/−	−	15–20%	FFP/PCC	40–60 h
Factor XI	AR	1 in 1,000,000	+	−	−	15–20%	FFP	40–70 h
Factor XII	AR	ND	+	−	−	[b]	[b]	60 h
HK	AR	ND	+	−	−	[b]	[b]	150 h
Prekallikrein	AR	ND	+	−	−	[b]	[b]	35 h
Factor XIII	AR	1 in 2,000,000	−	−	+/−	2–5%	Cryoprecipitate/FXIII concentrates	11–14 d

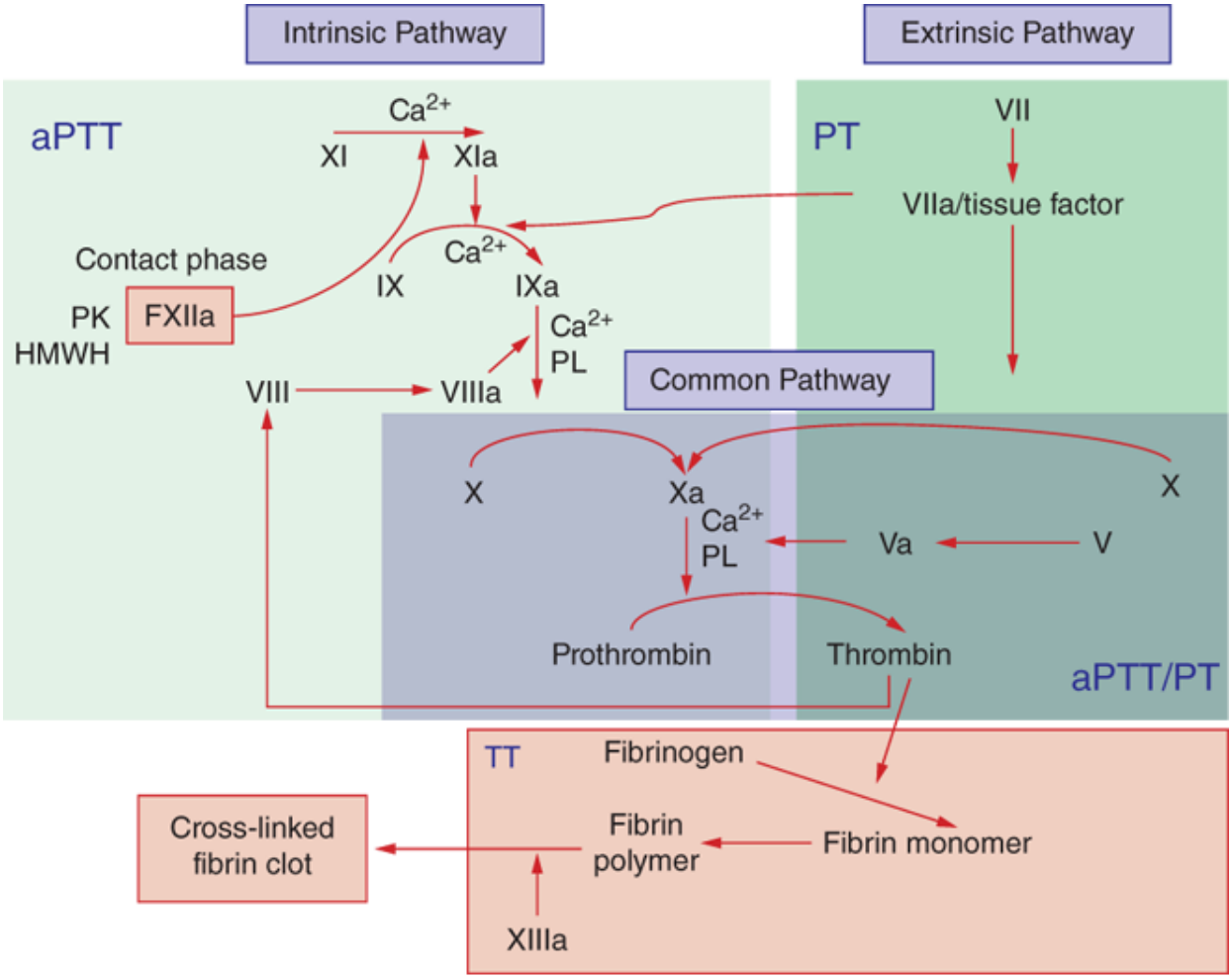
[a]Values within normal range (−) or prolonged (+). [b]No risk for bleeding; treatment is not indicated. [c]Since platelets contain FV, platelet transfusion can be used as therapy.

Abbreviations: aPTT, activated partial thromboplastin time; AR, autosomal recessive; FFP, fresh-frozen plasma; HK, high-molecular-weight kininogen; ND, not determined; PCC, prothrombin complex concentrates; PT, prothrombin time; TT, thrombin time.

Commonly used tests of hemostasis provide the initial screening for clotting factor activity (**Fig. 112-1**), and disease phenotype often correlates with the level of clotting activity. An isolated abnormal prothrombin time (PT) suggests FVII deficiency, whereas a prolonged activated partial thromboplastin time (aPTT) indicates most commonly hemophilia or FXI deficiency (Fig. 112-1). The prolongation of both PT and aPTT suggests deficiency of FV, FX, FII, or fibrinogen abnormalities. The addition of the missing factor at a range of doses to the subject's plasma will correct the abnormal clotting times; the result is expressed as a percentage of the activity observed in normal subjects.

FIGURE 112-1

Coagulation cascade and laboratory assessment of clotting factor deficiency by activated partial prothrombin time (aPTT), prothrombin time (PT), thrombin time (TT), and phospholipid (PL).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Acquired deficiencies of plasma coagulation factors are more frequent than congenital disorders; the most common disorders include hemorrhagic diathesis of liver disease, disseminated intravascular coagulation (DIC), and vitamin K deficiency. In these disorders, blood coagulation is hampered by the deficiency of more than one clotting factor, and the bleeding episodes are the result of perturbation of both primary (e.g., platelet and vessel wall interactions) and secondary (coagulation) hemostasis.

The development of alloantibodies to coagulation plasma proteins, clinically termed *inhibitors*, is a relatively rare disease that often affects hemophilia A or B and FXI-deficient patients on repetitive exposure to the missing protein to control bleeding episodes. Inhibitory autoantibodies also occur among subjects without genetic deficiency of clotting factors (e.g., in the postpartum setting as a manifestation of underlying autoimmune or neoplastic disease, or idiopathically). Rare cases of inhibitors to thrombin or FV have been reported in patients receiving topical bovine thrombin preparation as a local hemostatic agent in complex surgeries. The diagnosis of inhibitors is based on the same tests as those used to diagnose inherited plasma coagulation factor deficiencies. However, the addition of the missing protein to the plasma of a subject with an inhibitor does not correct the abnormal aPTT and/or PT tests (known as mixing tests). This is the major laboratory difference between deficiencies and inhibitors. Additional tests are required to measure the specificity of the inhibitor and its titer.

The treatment of these bleeding disorders often requires replacement of the deficient protein using recombinant or purified plasma-derived products or fresh-frozen plasma (FFP). Therefore, it is imperative to arrive at a proper diagnosis to optimize patient care without unnecessary exposure to suboptimal treatment and the risks of bloodborne disease.

HEMOPHILIA

PATHOGENESIS AND CLINICAL MANIFESTATIONS

Hemophilia is an X-linked recessive hemorrhagic disease due to mutations in the *F8* gene (hemophilia A or classic hemophilia) or *F9* gene (hemophilia B). The disease affects 1 in 10,000 males worldwide, in all ethnic groups; hemophilia A represents 80% of all cases. Male subjects are clinically affected; women, who carry a single mutated gene, are generally asymptomatic. Family history of the disease is absent in ~30% of cases, and in these cases, 80% of the mothers are carriers of the de novo mutated allele. More than 500 different mutations have been identified in the *F8* or *F9* genes of patients with hemophilia A or B, respectively. One of the most common hemophilia A mutations results from an inversion of the intron 22 sequence, and it is present in 40% of cases of severe hemophilia A. Advances in molecular diagnosis now permit precise identification of mutations, allowing accurate diagnosis of women carriers of the hemophilia gene in affected families.

Clinically, hemophilia A and hemophilia B are indistinguishable. The disease phenotype correlates with the residual activity of FVIII or FIX and can be classified as severe (<1%), moderate (1–5%), or mild (6–30%). In the severe and moderate forms, the disease is characterized by bleeding into the joints (hemarthrosis), soft tissues, and muscles after minor trauma or even spontaneously. Patients with mild disease experience infrequent bleeding that is usually secondary to trauma. Among those with residual FVIII or FIX activity >25% of normal, the disease is discovered only by bleeding after major trauma or during routine presurgery laboratory tests. Typically, the global tests of coagulation show only an isolated prolongation of the aPTT assay. Patients with hemophilia have normal bleeding times and platelet counts. The diagnosis is made after specific determination of FVIII or FIX clotting activity.

Early in life, bleeding may present after circumcision or rarely as intracranial hemorrhages. The disease is more evident when children begin to walk or crawl. In the severe form, the most common bleeding manifestations are the recurrent hemarthroses, which can affect every joint but mainly affect knees, elbows, ankles, shoulders, and hips. Acute hemarthroses are painful, and clinical signs are local swelling and erythema. To avoid pain, the patient may adopt a fixed position, which leads eventually to muscle contractures. Very young children unable to communicate

verbally show irritability and a lack of movement of the affected joint. Chronic hemarthroses are debilitating, with synovial thickening and synovitis in response to the intraarticular blood. After a joint has been damaged, recurrent bleeding episodes result in the clinically recognized “target joint,” which then establishes a vicious cycle of bleeding, resulting in progressive joint deformity that in critical cases requires surgery as the only therapeutic option. Hematomas into the muscle of distal parts of the limbs may lead to external compression of arteries, veins, or nerves that can evolve to a compartment syndrome.

Bleeding into the oropharyngeal spaces, central nervous system (CNS), or retroperitoneum is life threatening and requires immediate therapy. Retroperitoneal hemorrhages can accumulate large quantities of blood with formation of masses with calcification and inflammatory tissue reaction (pseudotumor syndrome) and also result in damage to the femoral nerve. Pseudotumors can also form in bones, especially long bones of the lower limbs. Hematuria is frequent among hemophilia patients, even in the absence of genitourinary pathology. It is often self-limited and may not require specific therapy.

TREATMENT

TREATMENT

Hemophilia

Without treatment, severe hemophilia may limit life expectancy. Advances in the blood fractionation industry during World War II resulted in the realization that plasma could be used to treat hemophilia, but the volumes required to achieve even modest elevation of circulating factor levels limit the utility of plasma infusion as an approach to disease management. The discovery in the 1960s that the cryoprecipitate fraction of plasma was enriched for FVIII, and the eventual purification of FVIII and FIX from plasma, led to the introduction of home infusion therapy with factor concentrates in the 1970s. The availability of factor concentrates resulted in a dramatic improvement in life expectancy and in quality of life for people with severe hemophilia. However, the contamination of the blood supply with hepatitis viruses and, subsequently, HIV resulted in widespread transmission of these bloodborne infections within the hemophilia population; complications of HIV and of hepatitis C are now the leading causes of death among U.S. adults with severe hemophilia. The introduction of viral inactivation steps in the preparation of plasma-derived products in the mid-1980s greatly reduced the risk of HIV and hepatitis, and the risks were further reduced by the successful production of recombinant FVIII and FIX proteins, both licensed in the 1990s. It is uncommon for hemophilic patients born after 1985 to have contracted either hepatitis or HIV, and for these individuals, life expectancy is ~65 years. In fact, since 1998, no evidence of new infections with viral hepatitis or HIV has been reported in patients using blood products. Factor replacement therapy for hemophilia can be provided either in response to a bleeding episode or as a prophylactic treatment. Primary prophylaxis is defined as a strategy for maintaining the missing clotting factor at levels ~1% or higher on a regular basis in order to prevent bleeds, especially the onset of hemarthroses. Hemophilic boys receiving regular infusions of FVIII (3 days/week) or FIX (2 days/week) can reach puberty without detectable joint abnormalities. Prophylaxis has become gradually more common in young patients. The Centers for Disease Control and Prevention reported that 51% of children with severe hemophilia who are aged <6 years receive prophylaxis, increasing considerably from 33% in 1995. Prophylaxis is the standard care for children; however, teenagers and young adults do not always maintain the treatment regularly. Although highly recommended, the high cost and difficulties in accessing peripheral veins in young patients and the potential infectious and thrombotic risks of long-term central vein catheters are important limiting factors for many young patients. Prophylaxis is also increasing among adults with severe hemophilia.

General considerations regarding the treatment of bleeds in hemophilia include the following: (1) Treatment should begin as soon as possible because symptoms often precede objective evidence of bleeding; because of the superior efficacy of early therapeutic intervention, classic symptoms of bleeding into the joint in a reliable patient, headaches, or automobile or other accidents require prompt replacement and further laboratory investigation. (2) Drugs that hamper platelet function, such as aspirin or aspirin-containing drugs, should be avoided; to control pain, drugs such as ibuprofen or propoxyphene are preferred. FVIII and FIX are dosed in units. One unit is defined as amount of FVIII (100 ng/mL) or FIX (5 μg/mL) in 1 mL of normal plasma. One unit of FVIII per kilogram of body weight increases the plasma FVIII level by 2%. One can calculate the dose needed to increase FVIII levels to 100% in a 70-kg severe hemophilia patient (<1%) using the simple formula below. Thus, 3500 units of FVIII will raise the circulating level to 100%.

$$\text{FVIII dose (IU)} = \text{Target FVIII levels} - \text{FVIII baseline levels} \times \text{body weight (kg)} \times 0.5 \text{ unit/kg}$$

The doses for FIX replacement are different from those for FVIII, because FIX recovery after infusion is usually only 50% of the predicted value. Therefore, the formula for FIX replacement is as follows:

$$\text{FIX dose (IU)} = \text{Target FIX levels} - \text{FIX baseline levels} \times \text{body weight (kg)} \times 1 \text{ unit/kg}$$

The FVIII half-life of 8–12 h requires injections twice a day to maintain therapeutic levels, whereas the FIX half-life is longer, ~24 h, so that once-a-day injection is sufficient. In specific situations such as after surgery, continuous infusion of factor may be desirable because of its safety in achieving sustained factor levels at a lower total cost.

Cryoprecipitate is enriched with FVIII protein (each bag contains ~80 IU of FVIII) and was commonly used for the treatment of hemophilia A decades ago; it is still in use in some developing countries, but because of the risk of bloodborne diseases, this product should be avoided in hemophilia patients when factor concentrates are available.

Mild bleeds such as uncomplicated hemarthroses or superficial hematomas require initial therapy with factor levels of 30–50%. Additional doses to maintain levels of 15–25% for 2 or 3 days are indicated for severe hemarthroses, especially when these episodes affect the “target joint.” Large hematomas, or bleeds into deep muscles, require factor levels of 50% or even higher if the clinical symptoms do not improve, and factor replacement may be required for a period of 1 week or longer. The control of serious bleeds including those that affect the oropharyngeal spaces, CNS, and the retroperitoneum require sustained protein levels of 50–100% for 7–10 days. Prophylactic replacement for surgery is aimed at achieving normal factor levels (100%) for a period of 7–10 days; replacement can then be tapered depending on the extent of the surgical wounds. Oral surgery is associated with extensive tissue damage that usually requires factor replacement for 1–3 days coupled with oral antifibrinolytic drugs.

NONTRANSFUSION THERAPY IN HEMOPHILIA

DDAVP (1-Amino-8-D-Arginine Vasopressin)

DDAVP is a synthetic vasopressin analog that causes a transient rise in FVIII and von Willebrand factor (VWF), but not FIX, through a mechanism involving release from endothelial cells. Patients with moderate or mild hemophilia A should be tested to determine if they respond to DDAVP before a therapeutic application. DDAVP at doses of 0.3 μg/kg body weight, over a 20-min period, is expected to raise FVIII levels by two- to threefold over baseline, peaking

between 30 and 60 min after infusion. DDAVP does not improve FVIII levels in severe hemophilia A patients because there are no stores to release. Repeated dosing of DDAVP results in tachyphylaxis because the mechanism is an increase in release rather than de novo synthesis of FVIII and VWF. More than three consecutive doses become ineffective, and if further therapy is indicated, FVIII replacement is required to achieve hemostasis.

Antifibrinolytic Drugs

Bleeding in the gums, gastrointestinal tract, and during oral surgery requires the use of oral antifibrinolytic drugs such as ε-amino caproic acid (EACA) or tranexamic acid to control local hemostasis. The duration of the treatment depending on the clinical indication is 1 week or longer. Tranexamic acid is given at doses of 25 mg/kg three to four times a day. EACA treatment requires a loading dose of 200 mg/kg (maximum of 10 g) followed by 100 mg/kg per dose (maximum 30 g/d) every 6 h. These drugs are not indicated to control hematuria because of the risk of formation of an occlusive clot in the lumen of genitourinary tract structures.

COMPLICATIONS

Inhibitor Formation

The formation of alloantibodies to FVIII or FIX is currently the major complication of hemophilia treatment. The prevalence of inhibitors to FVIII is estimated to be ~30% of severe hemophilia A patients and 10% among patients with non-severe hemophilia A. Inhibitors to FIX are detected in only 3–5% of all hemophilia B patients. The high-risk group for inhibitor formation includes severe deficiency (>80% of all cases of inhibitors), familial history of inhibitor, African descent, mutations in the FVIII or FIX gene resulting in deletion of large coding regions, or gross gene rearrangements. Inhibitors usually appear early in life, at a median of 2 years of age, and after 10 cumulative days of exposure. However, intensive replacement therapy such as for major surgery, intracranial bleeding, or trauma increases the risk of inhibitor formation for patients of all ages and degree of clinical severity, which requires close laboratory monitoring in the following weeks.

The clinical diagnosis of an inhibitor is suspected when patients do not respond to factor replacement at therapeutic doses. Inhibitors increase both morbidity and mortality in hemophilia. Because early detection of an inhibitor is critical to a successful correction of the bleeding or to eradication of the antibody, most hemophilia centers perform annual screening for inhibitors. The laboratory test required to confirm the presence of an inhibitor is an aPTT with a mix (with normal plasma). In most hemophilia patients, a 1:1 mix with normal plasma completely corrects the aPTT. In inhibitor patients, the aPTT on a 1:1 mix is abnormally prolonged because the inhibitor neutralizes the FVIII clotting activity of the normal plasma. The Bethesda assay uses a similar principle and defines the specificity of the inhibitor and its titer. The results are expressed in Bethesda units (BU), in which 1 BU is the amount of antibody that neutralizes 50% of the FVIII or FIX present in normal plasma after 2 h of incubation at 37°C. Clinically, inhibitor patients are classified as low responders or high responders, which provides guidelines for optimal therapy. Therapy for inhibitor patients has two goals: the control of acute bleeding episodes and the eradication of the inhibitor. For the control of bleeding episodes, low responders, those with titer <5 BU, respond well to high doses of human FVIII (50–100 U/kg), with minimal or no increase in the inhibitor titers. However, high-responder patients, those with initial inhibitor titer >5 BU or an anamnestic response in the antibody titer to >5 BU, even if low titer initially, do not respond to FVIII. The control of bleeding episodes in high-responder patients can be achieved by using concentrates enriched for prothrombin, FVII, FIX, FX (prothrombin complex concentrates [PCCs] or activated PCCs [aPCCs]), and recombinant activated factor VII (FVIIa) known as “bypass agents” (Fig. 112-1). For FIX inhibitor patients, high doses of FIX can be used (<5 BU); however, allergic or anaphylactic reactions are common in FIX inhibitors, thus the use of bypass products should be used to treat or prevent bleeding as well as for those cases of high titer inhibitors. For

eradication of the inhibitory antibody, immunosuppression alone is not effective. The most effective strategy is the immune tolerance induction (ITI) based on daily infusion of missing protein until the inhibitor disappears, typically requiring periods >1 year, with success rates of ~60%. The management of patients with severe hemophilia and inhibitors resistant to ITI is challenging. The use of anti-CD20 monoclonal antibody (rituximab) combined with ITI was thought to be effective. Although this therapy may reduce the inhibitor titers in some cases, sustained eradication is uncommon.

Novel Therapeutic Approaches in Development for Hemophilia

Clinical studies using long-acting clotting factors with enhanced half-lives are in the late phase of clinical testing, and these new-generation products (for FVIII and FIX) may facilitate prophylaxis by requiring fewer injections to maintain circulating levels >1%. In hemophilia A, the use of these products reduced the frequency of injections of FVIII from 3 to 2 days a week, and notably for hemophilia B, most patients will require once-a-week injections of long-acting FIX.

The use of recombinant interleukin 11 in patients with moderate or mild hemophilia A unresponsive to DDAVP has been tested in early-phase clinical trials and may be an alternate therapeutic strategy for clinical situations that require transient increases in FVIII levels.

Gene therapy trials for hemophilia A and B using adeno-associated viral vectors are ongoing **(Chap. 458)**.

INFECTIOUS DISEASES

Hepatitis C virus (HCV) infection is the major cause of morbidity and the second leading cause of death in hemophilia patients exposed to older clotting factor concentrates. The vast majority of young patients treated with plasma-derived products from 1970 to 1985 became infected with HCV. It has been estimated that >80% of patients aged >20 years are HCV antibody positive as of 2006. The comorbidity of the underlying liver disease in hemophilia patients is clear when these individuals require invasive procedures; correction of both genetic and acquired (secondary to liver disease) deficiencies may be needed. Infection with HIV also swept the population of patients using plasma-derived concentrates two decades ago. Co-infection of HCV and HIV, present in almost 50% of hemophilia patients, is an aggravating factor for the evolution of liver disease. The response to HCV antiviral therapy in hemophilia is restricted to <30% of patients and even poorer among those with both HCV and HIV infection. The development of effective direct-acting antivirals for the treatment of HCV may change this scenario. End-stage liver disease requiring organ transplantation may be curative for both the liver disease and for hemophilia

EMERGING CLINICAL PROBLEMS IN AGING HEMOPHILIA PATIENTS

There has been continuous improvement of the management of hemophilia since the increase in the population of adults living beyond middle age in the developing world. The life expectancy of a patient with severe hemophilia is only ~10 years shorter than the general male population. In patients with mild or moderate hemophilia, life expectancy is approaching that of the male population without coagulopathy. Elderly hemophilia patients have different problems compared to the younger generation; they have more severe arthropathy and chronic pain, due to suboptimal treatment, and high rates of HCV and/or HIV infections.

Early data indicate that mortality from coronary artery disease is lower in hemophilia patients than the general male population. The underlying hypocoagulability probably provides a protective effect against thrombus formation, but it does not prevent atherogenesis. Similar to the general population, these patients are exposed to cardiovascular risk factors such as age, obesity, and smoking. Moreover, physical inactivity, hypertension, and chronic renal disease are commonly observed in hemophilia patients. In HIV patients on combined antiretroviral therapy, there may be a

further increase in the risk of cardiovascular disease. Therefore, these patients should be carefully considered for preventive and therapeutic approaches to minimize the risk of cardiovascular disease.

Excessive replacement therapy should be avoided, and it is prudent to slowly infuse factor concentrates. Continuous infusion of clotting factor is preferable to bolus dosing in patients with cardiovascular risk factors undergoing invasive procedures. The management of an acute ischemic event and coronary revascularization should include the collaboration of hematologists and internists. The early assumption that hemophilia would protect against occlusive vascular disease may change in this aging population. Cancer is a common cause of mortality in aging hemophilia patients because they are at risk for HIV- and HCV-related malignancies. Hepatocellular carcinoma (HCC) is the most prevalent primary liver cancer and a common cause of death in HIV-negative patients. The recommendations for cancer screening for the general population should be the same for age-matched hemophilia patients. Among those with high-risk HCV, a semiannual or annual ultrasound and α fetoprotein are recommended for HCC. Screening for urogenital neoplasm in the presence of hematuria or hematochezia may be delayed due to the underlying bleeding disease, thus preventing early intervention. Multidisciplinary interaction should facilitate the attempts to ensure optimal cancer prevention and treatment recommendations for those with hemophilia.

MANAGEMENT OF CARRIERS OF HEMOPHILIA

Usually hemophilia carriers, with factor levels of ~50% of normal, have not been considered to be at risk for bleeding. However, a wide range of values (22–116%) have been reported due to random inactivation of the X chromosomes (*lyonization*). Therefore, it is important to measure the factor level of carriers to recognize those at risk of bleeding and to optimize preoperative and postoperative management. During pregnancy, both FVIII and FIX levels increase gradually until delivery. FVIII levels increase approximately two- to threefold compared to nonpregnant women, whereas an FIX increase is less pronounced. After delivery, there is a rapid fall in the pregnancy-induced rise of maternal clotting factor levels. This represents an imminent risk of bleeding that can be prevented by infusion of factor concentrate to levels of 50–70% for 3 days in the setting of vaginal delivery and up to 5 days for cesarean section. In mild cases, the use of DDAVP and/or antifibrinolytic drugs is recommended.

FACTOR XI DEFICIENCY

Factor XI is a zymogen of an active serine protease (FIXa) in the intrinsic pathway of blood coagulation that activates FIX (Fig. 112-1). There are two pathways for the formation of FXIa. In an aPTT-based assay, the protease is the result of activation by FXIIa in conjunction with high-molecular-weight kininogen and kallikrein. In vivo data suggest that thrombin is the physiologic activator of FXI. The generation of thrombin by the tissue factor/factor VIIa pathway activates FXI on the platelet surface that contributes to additional thrombin generation after the clot has formed and thus augments resistance to fibrinolysis through a thrombin-activated fibrinolytic inhibitor (TAFI).

Factor XI deficiency is a rare bleeding disorder that occurs in the general population at a frequency of one in a million. However, the disease is highly prevalent among Ashkenazi and Iraqi Jewish populations, reaching a frequency of 6% as heterozygotes and 0.1–0.3% as homozygotes. More than 65 mutations in the FXI gene have been reported, whereas fewer mutations (two to three) are found among affected Jewish populations.

Normal FXI clotting activity levels range from 70 to 150 U/dL. In heterozygous patients with moderate deficiency, FXI ranges from 20 to 70 U/dL, whereas in homozygous or double heterozygote patients, FXI levels are <1–20 U/dL. Patients with FXI levels <10% of normal have a high risk of bleeding, but the disease phenotype does not always correlate with residual FXI clotting activity. A family history is indicative of the risk of bleeding in the propositus.

Clinically, the presence of mucocutaneous hemorrhages such as bruises, gum bleeding, epistaxis, hematuria, and menorrhagia are common, especially following trauma. This hemorrhagic phenotype suggests that tissues rich in fibrinolytic activity are more susceptible to FXI deficiency. Postoperative bleeding is common but not always present, even among patients with very low FXI levels.

FXI replacement is indicated in patients with severe disease required to undergo a surgical procedure. A negative history of bleeding complications following invasive procedures does not exclude the possibility of an increased risk for hemorrhage.

TREATMENT

TREATMENT

Factor XI Deficiency

The treatment of FXI deficiency is based on the infusion of FFP at doses of 15–20 mL/kg to maintain trough levels ranging from 10 to 20%. Because FXI has a half-life of 40–70 h, the replacement therapy can be given on alternate days. The use of antifibrinolytic drugs is beneficial to control bleeds, with the exception of hematuria or bleeds in the bladder. The development of an FXI inhibitor was observed in 10% of severely FXI-deficient patients who received replacement therapy. Patients with severe FXI deficiency who develop inhibitors usually do not bleed spontaneously. However, bleeding following a surgical procedure or trauma can be severe. In these patients, FFP and FXI concentrates should be avoided. The use of PCC/aPCC or recombinant activated FVII has been effective.

RARE BLEEDING DISORDERS

Collectively, the inherited disorders resulting from deficiencies of clotting factors other than FVIII, FIX, and FXI (Table 112-1) represent a group of rare bleeding diseases. The bleeding symptoms in these patients vary from asymptomatic (dysfibrinogenemia or FVII deficiency) to life-threatening (FX or FXIII deficiency). There is no pathognomonic clinical manifestation that suggests one specific disease, but overall, in contrast to hemophilia, hemarthrosis is a rare event and bleeding in the mucosal tract or after umbilical cord clamping is common. Individuals heterozygous for plasma coagulation deficiencies are often asymptomatic. The laboratory assessment for the specific deficient factor following screening with general coagulation tests (Table 112-1) will define the diagnosis.

Replacement therapy using FFP or PCCs (containing prothrombin, FVII, FIX, and FX) provides adequate hemostasis in response to bleeds or as prophylactic treatment. The use of PCC should be carefully monitored and avoided in patients with underlying liver disease, or those at high risk for thrombosis because of the risk of DIC.

FAMILIAL MULTIPLE COAGULATION DEFICIENCIES

There are several bleeding disorders characterized by the inherited deficiency of more than one plasma coagulation factor. To date, the genetic defects in two of these diseases have been characterized, and they provide new insights into the regulation of hemostasis by gene-encoding proteins outside blood coagulation.

Combined Deficiency of FV and FVIII

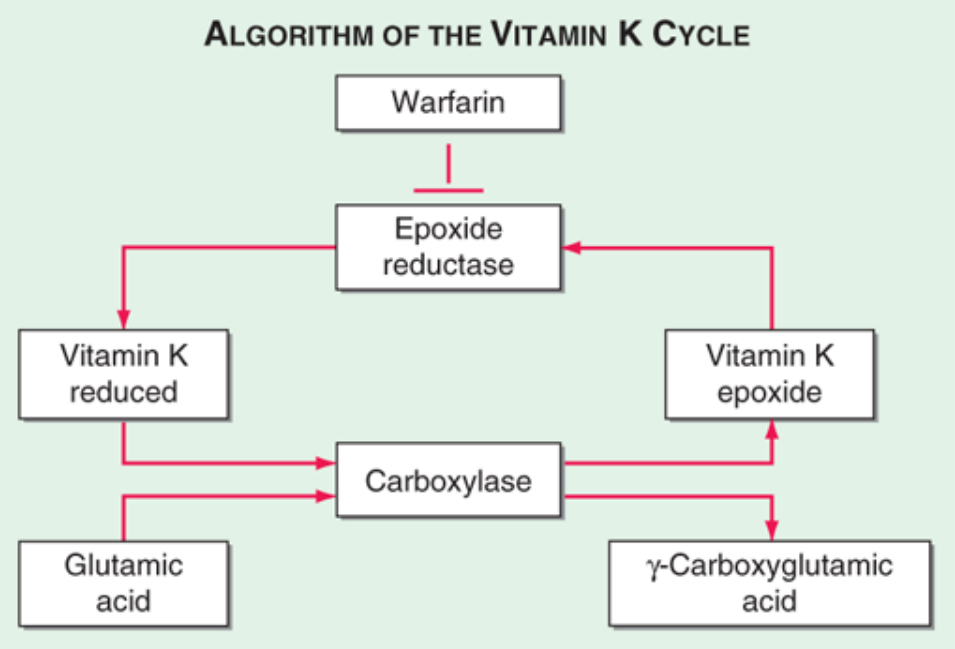
Patients with combined FV and FVIII deficiency exhibit ~5% of residual clotting activity of each factor. Interestingly, the disease phenotype is a mild bleeding tendency, often following trauma. An underlying mutation has been identified in the *lectin mannose binding 1* (*LMAN1*) gene, a mannose-binding protein localized in the Golgi apparatus that functions as a chaperone for both FV and FVIII. In other families, mutations in the multiple coagulation factor deficiency 2 (*MCFD2*) gene have been defined; this gene encodes a protein that forms a Ca^{2+} dependent complex with LMAN1 and provides cofactor activity in the intracellular mobilization of both FV and FVIII. The replacement therapy to control or prevent bleeding consisted of FFP to maintain FV levels and DDAVP or FVIII concentrate to achieve FVIII levels of 20–40%. Alternatively, since platelets contain FV, transfusion of platelets can also be used.

Multiple Deficiencies of Vitamin K–Dependent Coagulation Factors

Two enzymes involved in vitamin K metabolism have been associated with combined deficiency of all vitamin K–dependent proteins, including the procoagulant proteins prothrombin, VII, IX, and X and the anticoagulant proteins C and S. Vitamin K is a fat-soluble vitamin that is a cofactor for carboxylation of the gamma carbon of the glutamic acid residues in the vitamin K–dependent factors, a critical step for calcium and phospholipid binding of these proteins **(Fig. 112-2)**. The enzymes γ-glutamylcarboxylase and epoxide reductase are critical for the metabolism and regeneration of vitamin K. Mutations in the genes encoding the γ-carboxylase (GGCX) or vitamin K epoxide reductase complex 1 (VKORC1) result in defective enzymes and thus in vitamin K–dependent factors with reduced activity, varying from 1 to 30% of normal. The disease phenotype is characterized by mild to severe bleeding episodes present from birth. Some patients respond to oral administration of vitamin K1 (5–20 mg/d), or in poor responders parenteral vitamin K1 at doses of 5–20 mg/week could be considered. For severe bleeding, replacement therapy with FFP or PCC may be necessary to achieve full hemostatic control.

FIGURE 112-2

The vitamin K cycle. Vitamin K is a cofactor for the formation of γ-carboxyglutamic acid residues on coagulation proteins. Vitamin K–dependent γ-glutamylcarboxylase, the enzyme that catalyzes the vitamin K epoxide reductase, regenerates reduced vitamin K. Warfarin blocks the action of the reductase and competitively inhibits the effects of vitamin K.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DISSEMINATED INTRAVASCULAR COAGULATION

DIC is a clinicopathologic syndrome characterized by widespread intravascular fibrin formation in response to excessive blood protease activity that overcomes the natural anticoagulant mechanisms. There are several underlying pathologies associated with DIC (Table 112-2).

TABLE 112-2

Common Clinical Causes of Disseminated Intravascular Coagulation

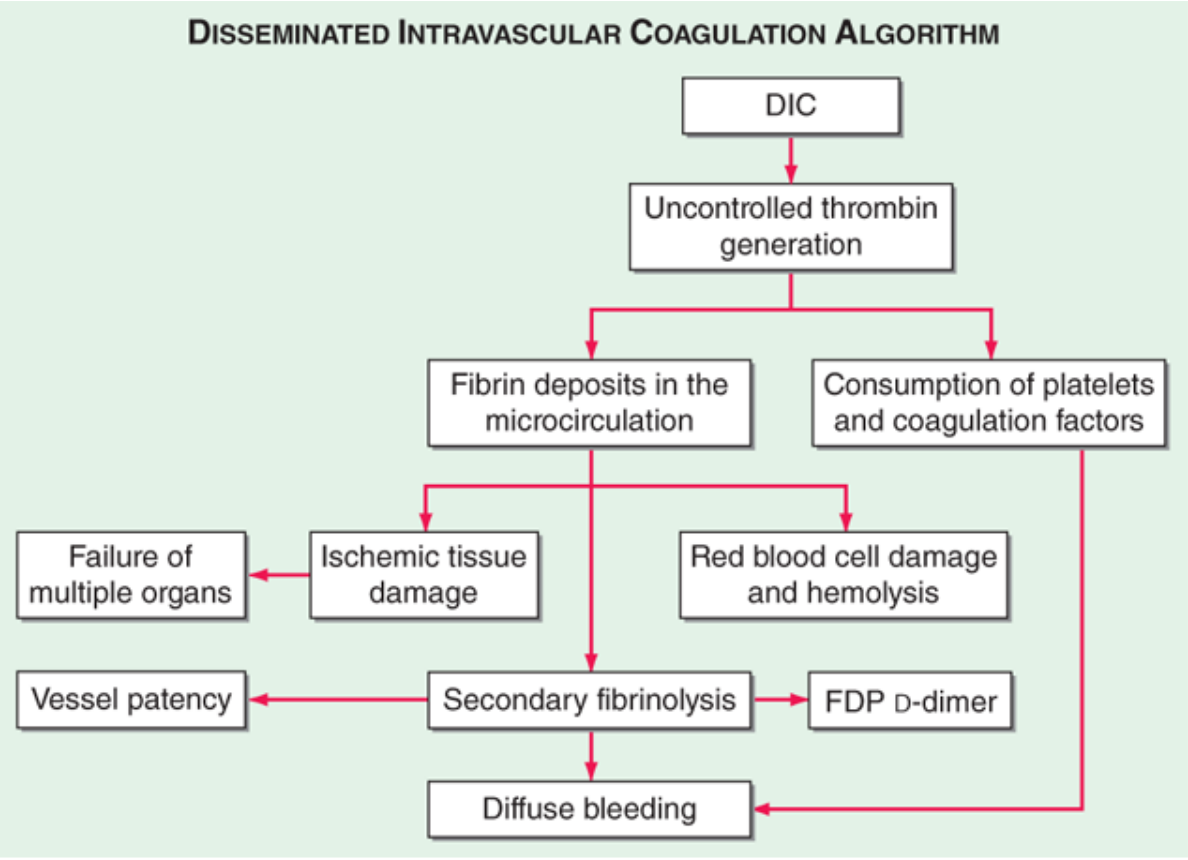
Sepsis	**Immunologic Disorders**
• Bacterial: Staphylococci, streptococci, pneumococci, meningococci, gram-negative bacilli • Viral • Mycotic • Parasitic • Rickettsial	• Acute hemolytic transfusion reaction • Organ or tissue transplant rejection • Immunotherapy • Graft-versus-host disease
Trauma and Tissue Injury	**Drugs**
• Brain injury (gunshot) • Extensive burns • Fat embolism • Rhabdomyolysis	• Fibrinolytic agents • Aprotinin • Warfarin (especially in neonates with protein C deficiency) • Prothrombin complex concentrates • Recreational drugs (amphetamines)
Vascular Disorders	**Envenomation**
• Giant hemangiomas (Kasabach-Merritt syndrome) • Large vessel aneurysms (e.g., aorta)	• Snake • Insects
Obstetrical Complications	**Liver Disease**
• Abruptio placentae • Amniotic fluid embolism • Dead fetus syndrome • Septic abortion	• Fulminant hepatic failure • Cirrhosis • Fatty liver of pregnancy
Cancer	**Miscellaneous**
• Adenocarcinoma (prostate, pancreas, etc.) • Hematologic malignancies (acute promyelocytic leukemia)	• Shock • Respiratory distress syndrome • Massive transfusion

The most common causes are bacterial sepsis, malignant disorders such as solid tumors or acute promyelocytic leukemia, and obstetric causes. DIC is diagnosed in almost one-half of pregnant women with abruptio placentae or with amniotic fluid embolism. Trauma, particularly to the brain, can also result in DIC. The exposure of blood to phospholipids from damaged tissue, hemolysis, and endothelial damage are all contributing factors to the development of DIC in this setting. Purpura fulminans is a severe form of DIC resulting from thrombosis of extensive areas of the skin; it affects predominantly young children following viral or bacterial infection, particularly those with inherited or acquired hypercoagulability due to deficiencies of the components of the protein C pathway. Neonates homozygous for protein C deficiency also present high risk for purpura fulminans with or without thrombosis of large vessels.

The central mechanism of DIC is the uncontrolled generation of thrombin by exposure of the blood to pathologic levels of tissue factor (**Fig. 112-3**). Simultaneous suppression of physiologic anticoagulant mechanisms and abnormal fibrinolysis further accelerate the process. Together, these abnormalities contribute to systemic fibrin deposition in small and midsize vessels. The duration and intensity of the fibrin deposition can compromise the blood supply of many organs, especially the lung, kidney, liver, and brain, with consequent organ failure. The sustained activation of coagulation results in consumption of clotting factors and platelets, which in turn leads to systemic bleeding. This is further aggravated by secondary hyperfibrinolysis. Studies in animals demonstrate that the fibrinolytic system is indeed suppressed at the time of maximal activation of coagulation. Interestingly, in patients with acute promyelocytic leukemia, a severe hyperfibrinolytic state often occurs in addition to the coagulation activation. The release of several proinflammatory cytokines such as interleukin 6 and tumor necrosis factor α plays a central role in mediating the coagulation defects in DIC and symptoms associated with systemic inflammatory response syndrome (SIRS).

FIGURE 112-3

The pathophysiology of disseminated intravascular coagulation (DIC). Interactions between coagulation and fibrinolytic pathways result in bleeding and thrombosis in the microcirculation in patients with DIC. FDP, fibrin degradation product.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical manifestations of DIC are related to the magnitude of the imbalance of hemostasis, to the underlying disease, or to both. The most common findings are bleeding ranging from oozing from venipuncture sites, petechiae, and ecchymoses to severe hemorrhage from the gastrointestinal tract, lung, or into the CNS. In chronic DIC, the bleeding symptoms are discrete and restricted to skin or mucosal surfaces. The hypercoagulability of DIC manifests as the occlusion of vessels in the microcirculation and resulting organ failure. Thrombosis of large vessels and cerebral embolism can also occur. Hemodynamic complications and shock are common among patients with acute DIC. The mortality ranges from 30 to >80% depending on the underlying disease, the severity of the DIC, and the age of the patient.

The diagnosis of clinically significant DIC is based on the presence of clinical and/or laboratory abnormalities of coagulation or thrombocytopenia. The laboratory diagnosis of DIC should prompt a search for the underlying disease if it is not already apparent. There is no single test that establishes the diagnosis of DIC. The laboratory investigation should include coagulation tests (aPTT, PT, thrombin time [TT]) and markers of fibrin degradation products (FDPs), in addition to platelet and red cell count and analysis of the blood smear. These tests should be repeated over a period of 6–8 h because an initially mild abnormality can change dramatically in patients with severe DIC.

Common findings include the prolongation of PT and/or aPTT; platelet counts µ100,000/µL, or a rapid decline in platelet numbers; the presence of schistocytes (fragmented red cells) in the blood smear; and elevated levels of FDP. The most sensitive test for DIC is the FDP level. DIC is an unlikely diagnosis in the presence of normal levels of FDP. The D-dimer test is more specific for detection of fibrin—but not fibrinogen—degradation products and indicates that the cross-linked fibrin has been digested by plasmin. Because fibrinogen has a prolonged half-life, plasma levels

diminish acutely only in severe cases of DIC. High-grade DIC is also associated with levels of antithrombin III or plasminogen activity <60% of normal.

Chronic DIC

Low-grade, compensated DIC can occur in clinical situations including giant hemangioma, metastatic carcinoma, or the dead fetus syndrome. Plasma levels of FDP or D-dimers are elevated. aPTT, PT, and fibrinogen values are within the normal range or high. Mild thrombocytopenia or normal platelet counts are also common findings. Red cell fragmentation is often detected but at a lower degree than in acute DIC.

Differential Diagnosis

The differential diagnosis between DIC and severe liver disease is challenging and requires serial measurements of the laboratory parameters of DIC. Patients with severe liver disease are at risk for bleeding and manifest laboratory features including thrombocytopenia (due to platelet sequestration, portal hypertension, or hypersplenism), decreased synthesis of coagulation factors and natural anticoagulants, and elevated levels of FDP due to reduced hepatic clearance. However, in contrast to DIC, these laboratory parameters in liver disease do not change rapidly. Other important differential findings include the presence of portal hypertension or other clinical or laboratory evidence of an underlying liver disease.

Microangiopathic disorders such as thrombotic thrombocytopenic purpura present an acute clinical onset of illness accompanied by thrombocytopenia, red cell fragmentation, and multiorgan failure. However, there is no consumption of clotting factors or hyperfibrinolysis.

Over the last few years, several clinical trials on immune therapies for neoplasias using monoclonal antibodies or gene-modified T cells targeting tumor-specific antigens showed unwanted inflammatory responses with increased cytokine release. These complications are sometimes associated with increased D-dimers and decreased fibrinogen levels, cytopenias, and liver dysfunction; thus, careful screening tests for DIC are indicated.

Treatment

TREATMENT

Disseminated Intravascular Coagulation

The morbidity and mortality associated with DIC are primarily related to the underlying disease rather than the complications of the DIC. The control or elimination of the underlying cause should therefore be the primary concern. Patients with severe DIC require control of hemodynamic parameters, respiratory support, and sometimes invasive surgical procedures. Attempts to treat DIC without accompanying treatment of the causative disease are likely to fail.

MANAGEMENT OF HEMORRHAGIC SYMPTOMS

Administration of FFP and/or platelet concentrates is indicated for patients with active bleeding or at high risk of bleeding, such as in preparation for invasive procedures or after chemotherapy. The control of bleeding in DIC patients with marked thrombocytopenia (platelet counts <10,000–20,000/μL) and low levels of coagulation factors will require replacement therapy. The PT (>1.5 times the normal) provides a good indicator of the severity of the

clotting factor consumption. Replacement with FFP is indicated (1 unit of FFP increases most coagulation factors by 30% in an adult without DIC). Low levels of fibrinogen (<100 mg/dL) or brisk hyperfibrinolysis will require infusion of cryoprecipitate (plasma fraction enriched for fibrinogen, FVIII, and VWF). The replacement of 10 U of cryoprecipitate for every 2–3 U of FFP is sufficient to correct the hemostasis. The transfusion scheme must be adjusted according to the patient's clinical and laboratory evolution. Platelet concentrates at a dose of 1–2 U/10 kg body weight are sufficient for most DIC patients with severe thrombocytopenia. Clotting factor concentrates are not recommended for control of bleeding in DIC because of the limited efficacy afforded by replacement of single factors (FVIII or FIX concentrates) and the high risk of products containing traces of aPCCs that further aggravate the disease.

REPLACEMENT OF COAGULATION OR FIBRINOLYSIS INHIBITORS

Drugs to control coagulation such as heparin, antithrombin III (ATIII) concentrates, or antifibrinolytic drugs have all been tried in the treatment of DIC. Low doses of continuous-infusion heparin (5–10 U/kg per h) may be effective in patients with low-grade DIC associated with solid tumor, acute promyelocytic leukemia, or in a setting with recognized thrombosis. Heparin is also indicated for the treatment of purpura fulminans during the surgical resection of giant hemangiomas and during removal of a dead fetus. In acute DIC, the use of heparin is likely to aggravate bleeding. To date, the use of heparin in patients with severe DIC has no proven survival benefit. The use of antifibrinolytic drugs, EACA, or tranexamic acid to prevent fibrin degradation by plasmin may reduce bleeding episodes in patients with DIC and confirmed hyperfibrinolysis. However, these drugs can increase the risk of thrombosis, and concomitant use of heparin is indicated. Patients with acute promyelocytic leukemia or those with chronic DIC associated with giant hemangiomas are among the few patients who may benefit from this therapy. The use of protein C concentrates to treat purpura fulminans associated with acquired protein C deficiency or meningococcemia has been proven efficacious. The results from the replacement of ATIII in early-phase studies are promising but require further study.

Guidance for diagnosis and treatment of DIC had been proposed by the International Society of Thrombosis and Haemostasis. This initiative will permit more detailed clinical data on diagnosis and treatment of DIC. The clinical utility of these scoring systems and therapeutic recommendations contained in these guidelines is not yet known.

VITAMIN K DEFICIENCY

Vitamin K–dependent proteins are a heterogenous group, including clotting factor proteins and also proteins found in bone, lung, kidney, and placenta. Vitamin K mediates posttranslational modification of glutamate residues to γ-carboxylglutamate, a critical step for the activity of vitamin K–dependent proteins for calcium binding and proper assembly to phospholipid membranes (Fig. 112-2). Inherited deficiency of the functional activity of the enzymes involved in vitamin K metabolism, notably the GGCX or VKORC1 (see above), results in bleeding disorders. The amount of vitamin K in the diet is often limiting for the carboxylation reaction; thus recycling of the vitamin K is essential to maintain normal levels of vitamin K–dependent proteins. In adults, low dietary intake alone is seldom reason for severe vitamin K deficiency but may become common in association with the use of broad-spectrum antibiotics. Disease or surgical interventions that affect the ability of the intestinal tract to absorb vitamin K, either through anatomic alterations or by changing the fat content of bile salts and pancreatic juices in the proximal small bowel, can result in significant reduction of vitamin K levels. Chronic liver diseases such as primary biliary cirrhosis also deplete vitamin K stores. Neonatal vitamin K deficiency and the resulting hemorrhagic disease of the newborn have been almost entirely eliminated by routine administration of vitamin K to all neonates. Prolongation of PT values is the most common and earliest finding in vitamin K–deficient patients due to reduction in prothrombin, FVII,

FIX, and FX levels. FVII has the shortest half-life among these factors that can prolong the PT before changes in the aPTT. Parenteral administration of vitamin K at a total dose of 10 mg is sufficient to restore normal levels of clotting factor within 8–10 h. In the presence of ongoing bleeding or a need for immediate correction before an invasive procedure, replacement with FFP or PCC is required. The latter should be avoided in patients with severe underlying liver disorders due to high risk of thrombosis. The reversal of excessive anticoagulant therapy with warfarin or warfarin-like drugs can be achieved by minimal doses of vitamin K (1 mg orally or by intravenous injection) for asymptomatic patients. This strategy can diminish the risk of bleeding while maintaining therapeutic anticoagulation for an underlying prothrombotic state.

In patients with life-threatening bleeds, the use of recombinant factor VIIa in nonhemophilia patients on anticoagulant therapy has been shown to be effective at restoring hemostasis rapidly, allowing emergency surgical intervention. However, patients with underlying vascular disease, vascular trauma and other comorbidities are at risk for thromboembolic complications that affect both arterial and venous systems. Thus, the use of factor VIIa in this setting is limited to administration of low doses given for only a limited number of injections. Close monitoring for vascular complications is highly indicated.

COAGULATION DISORDERS ASSOCIATED WITH LIVER FAILURE

The liver is central to hemostasis because it is the site of synthesis and clearance of most procoagulant and natural anticoagulant proteins and of essential components of the fibrinolytic system. Liver failure is associated with a high risk of bleeding due to deficient synthesis of procoagulant factors and enhanced fibrinolysis. Thrombocytopenia is common in patients with liver disease, and may be due to congestive splenomegaly (hypersplenism) or immune-mediated shortened platelet lifespan (primary biliary cirrhosis). In addition, several anatomic abnormalities secondary to underlying liver disease further promote the occurrence of hemorrhage (**Table 112-3**). Dysfibrinogenemia is a relatively common finding in patients with liver disease due to impaired fibrin polymerization. The development of DIC concomitant to chronic liver disease is not uncommon and may enhance the risk for bleeding. Laboratory evaluation is mandatory for an optimal therapeutic strategy, either to control ongoing bleeding or to prepare patients with liver disease for invasive procedures. Typically, these patients present with prolonged PT, aPTT, and TT depending on the degree of liver damage, thrombocytopenia, and normal or slight increase of FDP. Fibrinogen levels are diminished only in fulminant hepatitis, decompensated cirrhosis, or advanced liver disease, or in the presence of DIC. The presence of prolonged TT and normal fibrinogen and FDP levels suggest dysfibrinogenemia. FVIII levels are often normal or elevated in patients with liver failure, and decreased levels suggest superimposing DIC. Because FV is only synthesized in the hepatocyte and is not a vitamin K–dependent protein, reduced levels of FV may be an indicator of hepatocyte failure. Normal levels of FV and low levels of FVII suggest vitamin K deficiency. Vitamin K levels may be reduced in patients with liver failure due to compromised storage in hepatocellular disease, changes in bile acids, or cholestasis that can diminish the absorption of vitamin K. Replacement of vitamin K may be desirable (10 mg given by slow intravenous injection) to improve hemostasis.

TABLE 112-3

Coagulation Disorders and Hemostasis in Liver Disease

Bleeding
Portal hypertension
Esophageal varices
Thrombocytopenia
Splenomegaly
Chronic or acute DIC
Decreased synthesis of clotting factors
Hepatocyte failure
Vitamin K deficiency
Systemic fibrinolysis
DIC
Dysfibrinogenemia
Thrombosis
Decreased synthesis of coagulation inhibitors: protein C, protein S, antithrombin
Hepatocyte failure
Vitamin K deficiency (protein C, protein S)
Failure to clear activated coagulation proteins (DIC)
Dysfibrinogenemia
Iatrogenic: Transfusion of prothrombin complex concentrates
Antifibrinolytic agents: EACA, tranexamic acid

Abbreviations: DIC, disseminated intravascular coagulation; EACA, ε-aminocaproic acid.

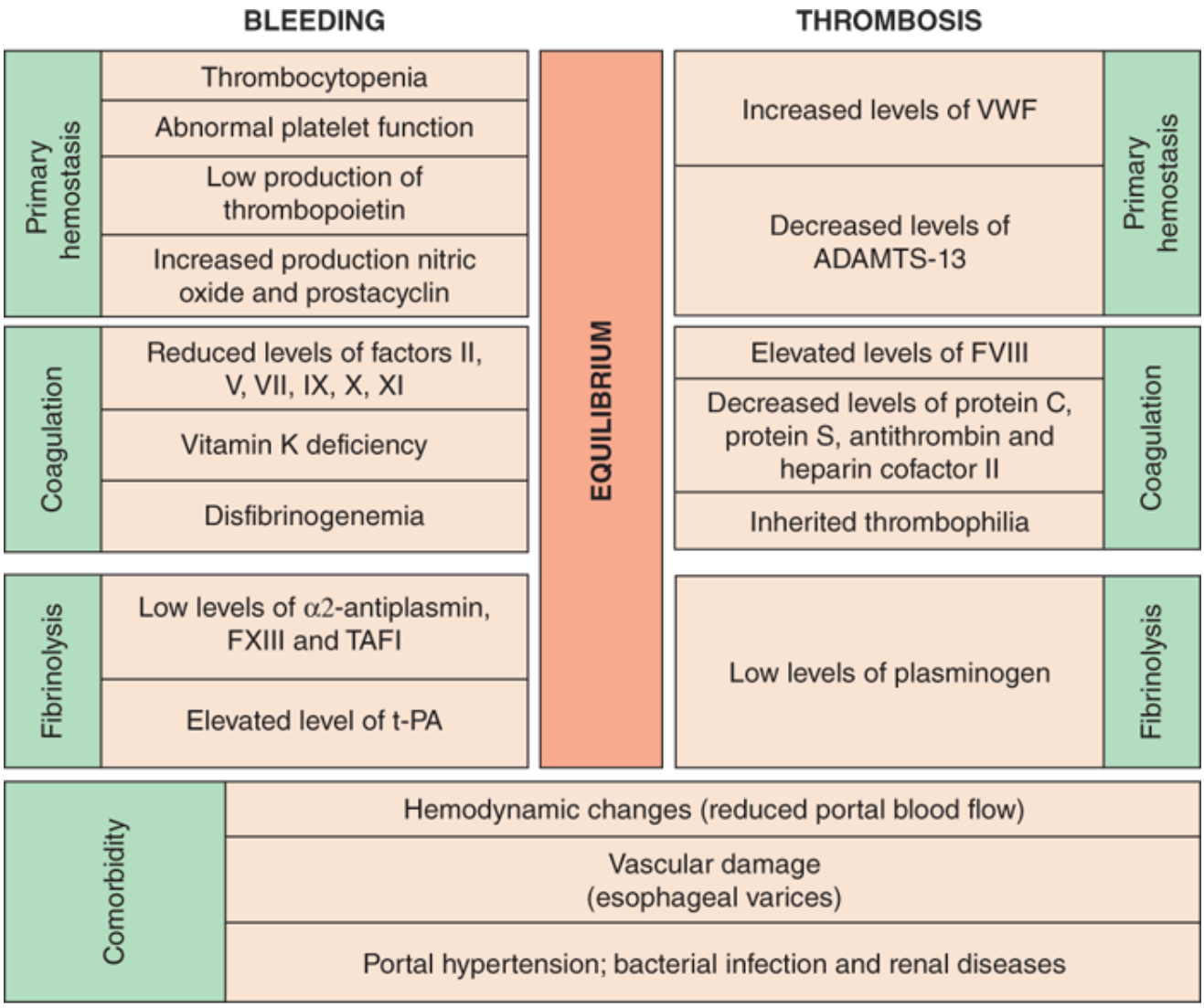
Treatment with FFP is the most effective to correct hemostasis in patients with liver failure. Infusion of FFP (5–10 mL/kg; each bag contains ~200 mL) is sufficient to ensure 10–20% of normal levels of clotting factors but not correction of PT or aPTT. Even high doses of FFP (20 mL/kg) do not correct the clotting times in all patients. Monitoring for clinical symptoms and clotting times will determine if repeated doses are required 8–12 h after the first infusion. Platelet concentrates are indicated when platelet counts are <10,000–20,000/μL to control an ongoing bleed or immediately before an invasive procedure if counts are <50,000/μL. Cryoprecipitate is indicated only when fibrinogen levels are less than 100 mg/mL; dosing is six bags for a 70-kg patient daily. PCC infusion in patients with liver failure should be avoided due to the high risk of thrombotic complications. The safety of the use of antifibrinolytic drugs to control bleeding in patients with liver failure is not yet well defined and should be avoided.

Liver Disease and Thromboembolism

The clinical bleeding phenotype of hemostasis in patients with stable liver disease is often mild or even asymptomatic. However, as the disease progresses, the hemostatic balance is less stable and more easily disturbed than in healthy individuals. Furthermore, the hemostatic balance is compromised by comorbid complications such as infections and renal failure **(Fig. 112-4)**. Based on the clinical bleeding complications in patients with cirrhosis and laboratory evidence of hypocoagulation such as a prolonged PT/aPTT, it has long been assumed that these patients are protected against thrombotic disease. Cumulative clinical experience, however, has demonstrated that these patients are at risk for thrombosis, especially those with advanced liver disease. Although hypercoagulability could explain the occurrence of venous thrombosis, according to Virchow's triad, hemodynamic changes and damaged vasculature may also be a contributing factor, and both processes may potentially also occur in patients with liver disease. Liver-related thrombosis, in particular, thrombosis of the portal and mesenteric veins, is common in patients with advanced cirrhosis. Hemodynamic changes, such as decreased portal flow, and evidence that inherited thrombophilia may enhance the risk for portal vein thrombosis in patients with cirrhosis suggest that hypercoagulability may play a role as well. Patients with liver disease develop deep-vein thrombosis and pulmonary embolism at appreciable rates (ranging from 0.5 to 1.9%). The implication of these findings is relevant to the erroneous exclusion of thrombosis in patients with advanced liver disease, even in the presence of prolongation of routine clotting times, and caution should be advised on overcorrection of these laboratory abnormalities.

FIGURE 112-4

Balance of hemostasis in liver disease. TAFI, thrombin-activated fibrinolytic inhibitor; t-PA, tissue plasminogen activator; VWF, von Willebrand factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Acquired Inhibitors of Coagulation Factors

An acquired inhibitor is an immune-mediated disease characterized by the presence of an autoantibody against a specific clotting factor. FVIII is the most common target of antibody formation, and is sometimes referred to as acquired hemophilia A, but inhibitors to prothrombin, FV, FIX, FX, and FXI are also reported. Acquired inhibitor to FVIII occurs predominantly in older adults (median age of 60 years), but occasionally in pregnant or postpartum women with no previous history of bleeding. In 50% of patients with inhibitors, no underlying disease is identified at the time of diagnosis. In the remaining patients, the causes are autoimmune diseases, malignancies (lymphomas, prostate cancer), dermatologic diseases, and pregnancy. Bleeding episodes occur commonly in soft tissues, the gastrointestinal or urinary tracts, and skin. In contrast to hemophilia, hemarthrosis is rare in these patients. Retroperitoneal hemorrhages and other life-threatening bleeding may appear suddenly. The overall mortality in untreated patients ranges from 8 to 22%, and most deaths occur within the first few weeks after presentation. The diagnosis is based on the prolonged aPTT with normal PT and TT. The aPTT remains prolonged after mixture of the test plasma with equal amounts of pooled normal plasma for 2 h at 37°C. The Bethesda assay using factor specific-deficient plasma as performed for inhibitor detection in hemophilia will confirm the diagnosis. Major bleeding is

treated with bypass products such as PCC/aPCC or recombinant FVIIa. Recombinant porcine FVIII is also available for the treatment of acquired hemophilia.

In contrast to hemophilia, inhibitors in nonhemophilic patients are typically responsive to immune suppression, and therapy should be initiated early for most cases. The first choice includes steroid or a combination of steroid with cytotoxic therapy (e.g., cyclophosphamide), with complete eradication of the inhibitors in more than 70% of patients. High-dose intravenous γ-globulin and anti-CD20 monoclonal antibody have been reported to be effective in patients with autoantibodies to FVIII; however, there is no firm evidence that these alternatives are superior to the first line of immunosuppressive drugs. Notably, relapse of the inhibitor to FVIII is relatively common (up to 20%) within the first 6 months following withdrawal of immunosuppression. Thus, after eradication, patients should be followed up regularly for early therapeutic intervention when indicated or prior to invasive procedure.

Topical plasma-derived bovine and human thrombin are commonly used in the United States and worldwide. These effective hemostatic sealants are used during major surgery such as for cardiovascular, thoracic, neurologic, pelvic, and trauma indications, as well as in the setting of extensive burns. The development of antibody formation to the xenoantigen or its contaminant (bovine clotting protein) has the potential to show cross-reactivity with human clotting factors that may hamper their function and induce bleeding.

Clinical features of these antibodies include bleeding from a primary hemostatic defect or coagulopathy that sometimes can be life threatening. The clinical diagnosis of these acquired coagulopathies is often complicated by the fact that the bleeding episodes may be detectable during or immediately following major surgery and could be assumed to be due to the procedure itself.

Notably, the risk of this complication is further increased by repeated exposure to topical thrombin preparations. Thus, a careful medical history of previous surgical interventions that may have occurred even decades earlier is critical to assessing risk.

The laboratory abnormalities are reflected by combined prolongation of the aPTT and PT that often fails to improve by transfusion of FFP and vitamin K. The abnormal laboratory tests cannot be corrected by mixing a test with equal parts of normal plasma that denotes the presence of inhibitory antibodies. The diagnosis of a specific antibody is obtained by the determination of the residual activity of human FV or other suspected human clotting factor. There are no commercially available assays specific for bovine thrombin coagulopathy.

There are no established treatment guidelines. Platelet transfusions have been used as a source of FV replacement for patients with FV inhibitors. Frequent injections of FFP and vitamin K supplementation may function as co-adjuvant rather than an effective treatment of the coagulopathy itself. Experience with recombinant FVIIa as a bypass agent is limited, and outcomes have been generally poor. Specific treatments to eradicate the antibodies based on immunosuppression with steroids, intravenous immunoglobulin, or serial plasmapheresis have been sporadically reported. Patients should be advised to avoid any topical thrombin sealant in the future.

Novel plasma-derived and recombinant human thrombin preparations for topical hemostasis have been approved by the U.S. Food and Drug Administration. These preparations have demonstrated hemostatic efficacy with reduced immunogenicity compared to the first generation of bovine thrombin products.

The presence of lupus anticoagulant can be associated with venous or arterial thrombotic disease. However, bleeding has also been reported in lupus anticoagulant; it is due to the presence of antibodies to prothrombin,

which results in hypoprothrombinemia. Both disorders show a prolonged PTT that does not correct on mixing. To distinguish acquired inhibitors from lupus anticoagulant, note that the dilute Russell's viper venom test and the hexagonal-phase phospholipids test will be negative in patients with an acquired inhibitor and positive in patients with lupus anticoagulants. Moreover, lupus anticoagulant interferes with the clotting activity of many factors (FVIII, FIX, FXII, FXI), whereas acquired inhibitors are specific to a single factor.

FURTHER READING

Gando S et al: Disseminated intravascular coagulation. Nat Rev Dis Primers 2:16037, 2017.

George LA et al: Hemophilia B gene therapy with a high-specific activity factor IX variant. N Engl J Med 377:2215, 2017.

Lee CA et al (eds): *Textbook of Hemophilia*, 3rd ed. Oxford, Wiley Blackwell, 2014.
CrossRef

Nienhuis AW et al: Gene therapy for hemophilia. Mol Ther 25:1163, 2017.

Peyvandi F, Menegatti M: Treatment of rare factor deficiencies in 2016. Hematology Am Soc Hematol Educ Program 1:663, 2016.

Rangarajan S et al: AAV5-factor Viii gene transfer in severe hemophilia A. N Engl J Med 377:2519, 2017.

Tripodi A et al: Changing concepts of cirrhotic coagulopathy. Am J Gastroenterol 112:274, 2017.

Chapter 113: Arterial and Venous Thrombosis

Jane E. Freedman; Joseph Loscalzo

OVERVIEW OF THROMBOSIS

GENERAL OVERVIEW

Thrombosis, the obstruction of blood flow due to the formation of clot, may result in tissue anoxia and damage, and it is a major cause of morbidity and mortality in a wide range of arterial and venous diseases and patient populations. In 2013 in the United States, cardiovascular disease accounted for 30.8% (800,937) of all 2,596,993 deaths, or about 1 of every 3 deaths. Approximately 735,000 people had a new or recurrent myocardial infarction, and ~690,000 people experienced a new or recurrent ischemic stroke. It is estimated that 300,000–600,000 people each year have a pulmonary embolism or deep-venous thrombotic event. In the nondiseased state, physiologic hemostasis reflects a delicate interplay between factors that promote and inhibit blood clotting, favoring the former. This response is crucial as it prevents uncontrolled hemorrhage and exsanguination following injury. In specific settings, the same processes that regulate normal hemostasis can cause pathologic thrombosis, leading to arterial or venous occlusion. Importantly, many commonly used therapeutic interventions may also alter the thrombotic–hemostatic balance adversely.

Hemostasis and thrombosis primarily involve the interplay among three factors: the vessel wall, coagulation and fibrinolytic proteins, and platelets. Many prevalent acute vascular diseases are due to thrombus formation within a vessel, including myocardial infarction, thrombotic cerebrovascular events, and venous thrombosis. Although the end result is vessel occlusion and tissue ischemia, the pathophysiologic processes governing these pathologies have similarities as well as distinct differences. While many of the pathways regulating thrombus formation are similar to those that regulate hemostasis, the processes triggering or perpetuating thrombosis may be distinct and can vary in different clinical and genetic settings. In venous thrombosis, primary hypercoagulable states reflecting defects in the proteins governing coagulation and/or fibrinolysis or secondary hypercoagulable states involving abnormalities of blood vessels and blood flow or stasis lead to thrombosis. By contrast, arterial thrombosis is highly dependent on the state of the vessel wall, the platelet, and factors related to blood flow.

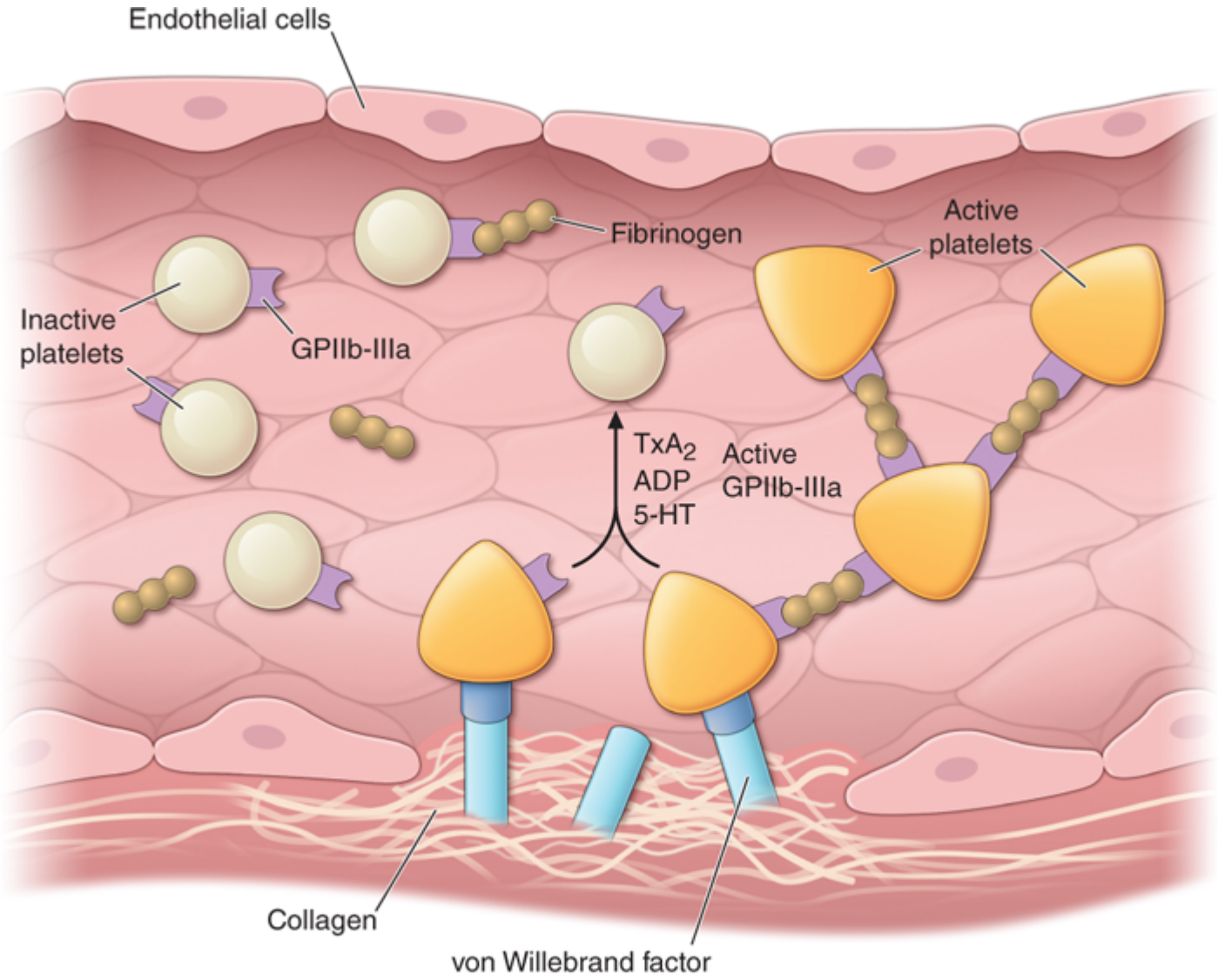
ARTERIAL THROMBOSIS

OVERVIEW OF ARTERIAL THROMBOSIS

In arterial thrombosis, platelets and abnormalities of the vessel wall typically play a key role in vessel occlusion. Arterial thrombus forms via a series of sequential steps in which platelets adhere to the vessel wall, additional platelets are recruited, and thrombin is activated (**Fig. 113-1**). The regulation of platelet adhesion, activation, aggregation, and recruitment will be described in detail below. In addition, while the primary function of platelets is regulation of hemostasis, our understanding of their role in other processes, such as immunity, metastasis, wound healing, and inflammation, continues to evolve.

FIGURE 113-1

Platelet activation and thrombosis. Platelets circulate in an inactive form in the vasculature. Damage to the endothelium and/or external stimuli activates platelets that adhere to the exposed subendothelial von Willebrand factor and collagen. This adhesion leads to activation of the platelet, shape change, and the synthesis and release of thromboxane (TxA_2), serotonin (5-HT), and adenosine diphosphate (ADP). Platelet stimuli cause conformational change in the platelet integrin glycoprotein (GP) IIb/IIIa receptor, leading to the high-affinity binding of fibrinogen and the formation of a stable platelet thrombus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ARTERIAL THROMBOSIS AND VASCULAR DISEASE

Arterial thrombosis is a major cause of morbidity and mortality both in the United States and, increasingly, worldwide. Although the rates have declined in the United States, the overall burden remains high. Overall, in 2013 coronary heart disease was estimated to cause about 1 of every 7 deaths in the United States. In addition to the 660,000 Americans who will have a new coronary event, an additional 160,000 silent first myocardial infarctions are projected to occur annually. Although the rate of strokes has fallen by a third, each year, about 690,000 people experience a new or recurrent ischemic stroke. It is estimated that 1 of every 20 deaths in the United States is due to stroke.

THE PLATELET

Many processes in platelets have parallels with other cell types, such as the presence of specific receptors and signaling pathways; however, unlike most cells, platelets lack a nucleus and are unable to adapt to changing biologic settings by altered gene transcription. Platelets sustain limited protein synthetic capacity from megakaryocyte-derived and intracellularly transported messenger RNA (mRNA) and microRNA (miRNA). Most of the molecules needed to respond to various stimuli, however, are maintained in storage granules and membrane compartments.

Platelets are disc-shaped, very small, anucleate cells (1–5 μm in diameter) that circulate in the blood at concentrations of 200–400,000/μL, with an average life span of 7–10 days. Platelets are derived from megakaryocytes, polyploidal hematopoietic cells found in the bone marrow. The primary regulator of platelet formation is thrombopoietin (TPO). The precise mechanism by which megakaryocytes produce and release fully formed platelets is unclear, but the process likely involves formation of proplatelets, pseudopod-like structures generated by the evagination of the cytoplasm from which platelets bud. After release into the circulation, (young, large) platelets may continue to divide. Platelet granules are synthesized in megakaryocytes before thrombopoiesis and contain an array of prothrombotic, proinflammatory, and antimicrobial mediators. The two major types of platelet granules, alpha and dense, are distinguished by their size, abundance, and content. Alpha-granules contain soluble coagulation proteins, adhesion molecules, growth factors, integrins, cytokines, and inflammatory modulators. Platelet dense-granules are smaller than alpha-granules and less abundant. Whereas alpha-granules contain proteins that may be more important in the inflammatory response, dense-granules contain high concentrations of small molecules, including adenosine diphosphate (ADP) and serotonin that influence platelet aggregation and other related vascular processes, such as vasomotor tone.

Platelet Adhesion

(See Fig. 113-1) The formation of a thrombus is initiated by the adherence of platelets to the damaged vessel wall. Damage exposes subendothelial components responsible for triggering platelet reactivity, including collagen, von Willebrand factor, fibronectin, and other adhesive proteins, such as vitronectin and thrombospondin. The hemostatic response may vary, depending on the extent of damage, the specific proteins exposed, and flow conditions. Certain proteins are expressed on the platelet surface that

subsequently regulate collagen-induced platelet adhesion, particularly under flow conditions, and include glycoprotein (GP) IV, GPVI, and the integrin $\alpha_2\beta_1$. The platelet GPIb-IX-V complex adhesive receptor is central both to platelet adhesion and to the initiation of platelet activation. Damage to the blood vessel wall exposes subendothelial von Willebrand factor and collagen to the circulating blood. The GPIb-IX-V complex binds to the exposed von Willebrand factor, causing platelets to adhere (Fig. 113-1). In addition, the engagement of the GPIb-IX-V complex with ligand induces signaling pathways that lead to platelet activation. von Willebrand factor–bound GPIb-IX-V promotes a calcium-dependent conformational change in the GPIIb/IIIa receptor, transforming it from an inactive low-affinity state to an active high-affinity receptor for fibrinogen.

Platelet Activation

The activation of platelets is controlled by a variety of surface receptors that regulate various functions in the activation process. Platelet receptors control many distinct processes and are stimulated by a wide variety of agonists and adhesive proteins that result in variable degrees of activation. In general terms, the stimulation of platelet receptors triggers two specific processes: (1) activation of internal signaling pathways that lead to further platelet activation and granule release, and (2) the capacity of the platelet to bind to other adhesive proteins/platelets. Both of these processes contribute to the formation of a thrombus. Stimulation of nonthrombotic receptors results in platelet adhesion or interaction with other vascular cells, including endothelial cells, neutrophils, and mononuclear cells.

Many families and subfamilies of receptors are found on platelets that regulate a variety of platelet functions. These include the seven transmembrane receptor family, which is the main agonist-stimulated receptor family. Several seven transmembrane receptors are found on platelets, including the ADP receptors, prostaglandin receptors, lipid receptors, and chemokine receptors. Receptors for thrombin comprise the major seven transmembrane receptors found on platelets. Among this last group, the first identified was the protease activation receptor 1 (PAR1). The PAR class of receptors has a distinct mechanism of activation that involves specific cleavage of the N-terminus by thrombin, which, in turn, acts as a ligand for the receptor. Other PAR receptors are present on platelets, including PAR2 (not activated by thrombin) and PAR4. Adenosine receptors are responsible for transduction of ADP-induced signaling events, which are initiated by the binding of ADP to purinergic receptors on the platelet surface. There are several distinct ADP receptors, classified as $P2X_1$, $P2Y_1$, and $P2Y_{12}$. The activation of both the $P2Y_{12}$ and $P2Y_1$ receptors is essential for ADP-induced platelet aggregation. The thienopyridine derivatives, clopidogrel and prasugrel, are clinically used inhibitors of ADP-induced platelet aggregation.

Platelet Aggregation

Activation of platelets results in a rapid series of signal transduction events, including tyrosine kinase, serine/threonine kinase, and lipid kinase activation. In unstimulated platelets, the major platelet integrin GPIIb/IIIa is maintained in an inactive conformation and functions as a low-affinity adhesion receptor for fibrinogen. This integrin is unique as it is only expressed on platelets. After stimulation, the interaction between fibrinogen and GPIIb/IIIa forms intercellular connections between platelets, leading to the

formation of a platelet aggregate (Fig. 113-1). A calcium-sensitive conformational change in the extracellular domain of GPIIb/IIIa enables the high-affinity binding of soluble plasma fibrinogen as a result of a complex network of inside-out signaling events. The GPIIb/IIIa receptor serves as a bidirectional conduit with GPIIb/IIIa-mediated signaling (outside-in) occurring immediately after the binding of fibrinogen. This leads to additional intracellular signaling that further stabilizes the platelet aggregate and transforms platelet aggregation from a reversible to an irreversible process (inside-out).

THE ROLE OF PLATELETS AND THROMBOSIS IN INFLAMMATION

Inflammation plays an important role during the acute thrombotic phase of acute coronary syndromes. In the setting of acute upper respiratory infections, people are at higher risk of myocardial infarction and thrombotic stroke. Patients with acute coronary syndromes have not only increased interactions between platelets (homotypic aggregates), but also increased interactions between platelets and leukocytes (heterotypic aggregates) detectable in circulating blood. These latter aggregates form when platelets are activated and adhere to circulating leukocytes. Platelets bind via P-selectin (CD62P) expressed on the surface of activated platelets to the leukocyte receptor, P-selectin glycoprotein ligand 1 (PSGL-1). This association leads to increased expression of CD11b/CD18 (Mac-1) on leukocytes, which itself supports interactions with platelets partially via bivalent fibrinogen linking this integrin with its platelet surface counterpart, GPIIb/IIIa. Platelet surface P-selectin also induces the expression of tissue factor on monocytes, which promotes fibrin formation.

In addition to platelet–monocyte aggregates, the immunomodulator, soluble CD40 ligand (CD40L or CD154), also reflects a link between thrombosis and inflammation. The CD40 ligand is a trimeric transmembrane protein of the tumor necrosis factor family and, with its receptor CD40, is an important contributor to the inflammatory process leading both to thrombosis and atherosclerosis. While many immunologic and vascular cells have been found to express CD40 and/or CD40 ligand, in platelets, CD40 ligand is rapidly translocated to the surface after stimulation and is upregulated in the newly formed thrombus. The surface-expressed CD40 ligand is cleaved from the platelet to generate a soluble fragment (soluble CD40 ligand).

Links have also been established among platelets, infection, immunity, and inflammation. Bacterial and viral infections are associated with a transient increase in the risk of acute thrombotic events, such as acute myocardial infarction and stroke. In addition, platelets contribute significantly to the pathophysiology and high mortality rates of sepsis. The expression, functionality, and signaling pathways of Toll-like receptors (TLRs) have been established in platelets. Stimulation of platelet TLR2, TLR3, and TLR4 directly and indirectly activates the platelet's thrombotic and inflammatory responses, and live bacteria induce a proinflammatory response in platelets in a TLR2-dependent manner, suggesting a mechanism by which specific bacteria and bacterial components can directly activate platelet-dependent thrombosis. Additionally, viruses, such as HIV, HCV, and Dengue, are also known to cause elevated levels of thrombosis, and, recently, platelets have been shown to regulate immune responses to viruses via receptors TLR7 and TLR8.

Risk Factors for Arterial Thrombosis

Various factors increase the risk of developing arterial thrombosis. Classically, the cardiovascular-dependent risk factors implicated in thrombosis have been hypertension, high levels of low-density lipoprotein-cholesterol, and smoking. However, diabetes, pregnancy, age, chemotherapeutics, and infectious burden may also contribute to arterial thrombosis. Stillbirth and loss of multiple pregnancies may increase the risk of ischemic stroke and MI as does hormonal replacement therapy. Systemic lupus erythematosus and rheumatoid arthritis are now well-recognized risks for thrombosis, and the former, in particular, may contribute in the pediatric population. The antiphospholipid syndrome is also another widely recognized autoimmune prothrombotic risk for arterial (and venous) thrombosis.

GENETICS OF ARTERIAL THROMBOSIS



Some studies have associated arterial thrombosis with genetic variants **(Table 113-1 *A*)**; however, the associations have been weak and not confirmed in larger series. Platelet count and mean platelet volume have been studied by genome-wide association studies (GWAS), and this approach identified signals located to noncoding regions. Of 15 quantitative trait loci associated with mean platelet volume and platelet count, one located at 12q24 is also a risk locus for coronary artery disease.

TABLE 113-1

Heritable Causes of Arterial and Venous Thrombosis

A. Arterial Thrombosis
Platelet Receptors
β3 and α2 integrins
P_l A2 polymorphism
Fc(gamma)RIIA
GPIV T13254C polymorphism
GPIb
Thrombin receptor PAR1-5061 → D
Redox Enzymes
Plasma glutathione peroxidase, GPx3, promoter haploytpe H2
H2 promoter haplotype
Endothelial nitric oxide synthase
–786T/C, –922A/G, –1468T/A
Paraoxonase
–107T allele, 192R allele
Homocysteine
Cystathionine β-synthase 833T → C
5,10-Methylene tetrahydrofolate reductase (MTHFR) 677C → T
B. Venous Thrombosis
Procoagulant Proteins

Fibrinogen
–455G/A, –854G/A
Prothrombin (20210G → A)
Protein C Anticoagulant Pathway
Factor V Leiden: 1691G → A (Arg506Gln)
Thrombomodulin 1481C → T (Ala455Val)
Fibrinolytic Proteins with Known Polymorphisms
Tissue plasminogen activator (tPA)
7351C/T, 20 099T/C in exon 6, 27 445T/A in intron 10
Plasminogen activator inhibitor (PAI-1)
4G/5G insertion/deletion polymorphism at position –675
Homocysteine
Cystathionine β-synthase 833T → C
5,10-MTHFR 677C → T

In the area of genetic variability and platelet function, studies have primarily dealt with pharmacogenetics, the field of pharmacology dealing with the interindividual variability in drug response based on genetic determinants (**Table 113-2**). This focus has been driven by the wide variability among individuals in terms of response to antithrombotic drugs and the lack of a common explanation for this variance. The best described is the issue of “aspirin resistance,” although heterogeneity for other antithrombotics (e.g., clopidogrel) has also been extensively examined. Primarily, platelet-dependent genetic determinants have been defined at the level of (1) drug effect, (2) drug compliance, and (3) drug metabolism. Many candidate platelet genes have been studied for their interaction with antiplatelet and antithrombotic agents.

TABLE 113-2

Genetic Variation and Pharmacogenetic Responses to Platelet Inhibitors

Potential Gene Altered	Target Therapeutic Class	Specific Drug
P2Y1 and *P2Y12 CYP2C19, CYP3A4, CYP3A5*	ADP receptor inhibitors	Clopidogrel, prasugrel
COX1, COX2	Cyclooxygenase inhibitors	Aspirin
PlA1/A2	Receptor inhibitors	Abciximab, eptifibatide, tirofiban
INTB3, GPIbA	Glycoprotein IIb-IIIa receptor inhibitors	

Many patients have an inadequate response to the inhibitory effects of aspirin. Heritable factors contribute to the variability; however, ex vivo tests of residual platelet responsiveness after aspirin administration have not provided firm evidence for a pharmacogenetic interaction between aspirin and *COX1* or other relevant platelet receptors. As such, currently, there is no clinical indication for genotyping to optimize aspirin's antiplatelet efficiency. For the platelet P2Y12 receptor inhibitor clopidogrel, additional data suggest that genetics may affect the drug's responsiveness and utility. The responsible genetic variant appears not to be the expected P2Y12 receptor but an enzyme responsible for drug metabolism. Clopidogrel is a prodrug, and liver metabolism by specific cytochrome P450 enzymes is required for activation. The genes encoding the CYP-dependent oxidative steps are polymorphic, and carriers of specific alleles of the CYP2C19 and CYP3A4 loci have increased platelet aggregability. Increased platelet activity has also been specifically associated with the CYP2C19*2 allele, which causes loss of platelet function in select patients. Because these are common genetic variants, this observation has been shown to be clinically relevant in large studies. In summary, although the loss-of-function polymorphisms in CYP2C19 is the strongest individual variable affecting pharmacokinetics and antiplatelet response to clopidogrel; it only accounts for 5–12% of the variability in ADP-induced platelet aggregation on clopidogrel. In addition, genetic variables do not appear to contribute significantly to the clinical outcomes of patients treated with the P2Y12 receptor antagonists prasugrel or ticagrelor.

VENOUS THROMBOSIS

OVERVIEW OF VENOUS THROMBOSIS

Coagulation is the process by which thrombin is activated and soluble plasma fibrinogen is converted into insoluble fibrin. These steps account for both normal hemostasis and the pathophysiologic processes

influencing the development of venous thrombosis. The primary forms of venous thrombosis are deep-vein thrombosis (DVT) in the extremities and the subsequent embolization to the lungs (pulmonary embolism), referred to together as venous thromboembolic disease (VTE). Venous thrombosis occurs due to heritable causes (Table 113-1 *B*) and acquired causes **(Table 113-3)**.

TABLE 113-3

Acquired Causes of Venous Thrombosis

Surgery
Neurosurgery
Major abdominal surgery
Malignancy
Antiphospholipid syndrome
Other
Trauma
Pregnancy
Long-distance travel
Obesity
Oral contraceptives/hormone replacement
Myeloproliferative disorders
Polycythemia vera

DEEP-VENOUS THROMBOSIS AND PULMONARY EMBOLISM

It is estimated that 300,000–600,000 new cases of venous thromboembolism occur each year with 60,000–80,000 deaths attributed to DVT or PE. Of new cases, up to 30% of patients die within 30 days and one-fifth suffer sudden death due to pulmonary embolism; 30% go on to develop recurrent venous thromboembolism within 10 years. Data from the Atherosclerosis Risk in Communities (ARIC) study reported a 9% 28-day fatality

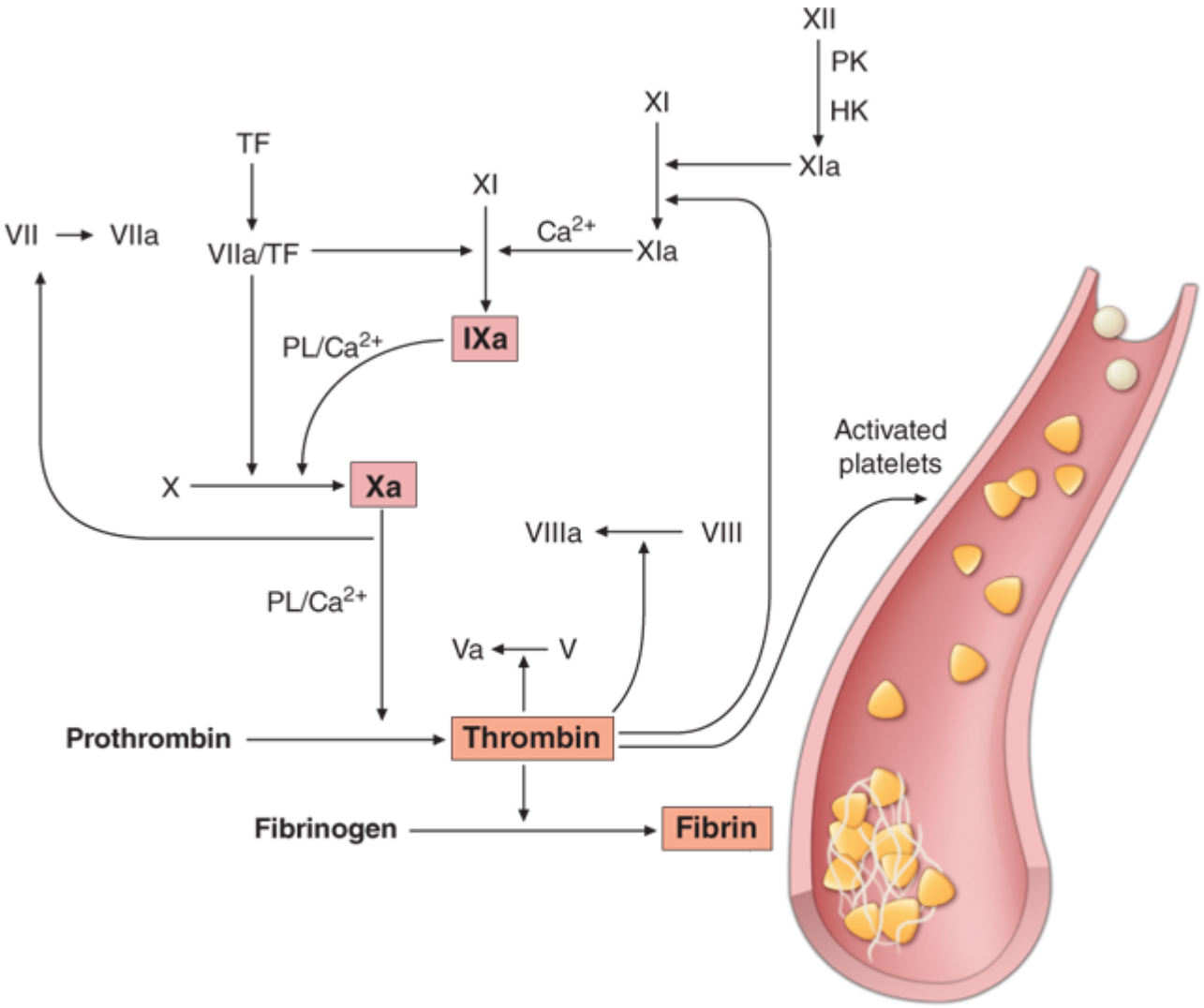
rate from DVT and a 15% fatality rate from pulmonary embolism. Pulmonary embolism in the setting of cancer has a 25% fatality rate. The mean incidence of first DVT in the general population is 5 per 10,000 person-years; the incidence is similar in males and females when adjusting for factors related to reproduction and birth control and increases dramatically with age from 2 to 3 per 10,000 person-years at 30–49 years of age to 20 per 10,000 person-years at 70–79 years of age.

OVERVIEW OF THE COAGULATION CASCADE AND ITS ROLE IN VENOUS THROMBOSIS

Coagulation is defined as the formation of fibrin by a series of linked enzymatic reactions in which each reaction product converts the subsequent inactive zymogen into an active serine protease **(Fig. 113-2)**. This coordinated sequence is called the coagulation cascade and is a key mechanism for regulating hemostasis. Central to the function of the coagulation cascade is the principle of amplification: due to a series of linked enzymatic reactions, a small stimulus can lead to much greater quantities of fibrin, the end product that prevents hemorrhage at the site of vascular injury. In addition to the known risk factors relevant to hypercoagulopathy, stasis, and vascular dysfunction, newer areas of research have identified contributions from procoagulant microparticles, inflammatory cells, microvesicles, and fibrin structure.

FIGURE 113-2

Summary of the coagulation pathways. Specific coagulation factors (“a” indicates activated form) are responsible for the conversion of soluble plasma fibrinogen into insoluble fibrin. This process occurs via a series of linked reactions in which the enzymatically active product subsequently converts the downstream inactive protein into an active serine protease. In addition, the activation of thrombin leads to stimulation of platelets. HK, high-molecular-weight kininogen; PK, prekallikrein; TF, tissue factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The coagulation cascade is primarily initiated by vascular injury exposing tissue factor to blood components (Fig. 113-2). Tissue factor may also be found in bloodborne cell-derived microparticles and, under pathophysiologic conditions, in leukocytes or platelets. Plasma factor VII (FVII) is the ligand for and is activated (FVIIa) by binding to tissue factor exposed at the site of vessel damage. The binding of FVII/VIIa to tissue factor activates the downstream conversion of factor X (FX) to active FX (FXa). In an alternative reaction, the FVII/FVIIa–tissue factor complex initially converts FIX to FIXa, which then activates FX in conjunction with its cofactor factor VIII (FVIIIa). Factor Xa with its cofactor FVa converts prothrombin to thrombin, which then converts soluble plasma fibrinogen to insoluble fibrin, leading to clot or thrombus formation. Thrombin also activates FXIII to FXIIIa, a transglutaminase that covalently cross-links and stabilizes the fibrin clot. Formation of thrombi is affected by mechanisms governing fibrin structure and stability, including specific fibrinogen variants and how they alter fibrin formation, strength, and structure.

Several antithrombotic factors also regulate coagulation; these include antithrombin, tissue factor pathway inhibitor (TFPI), heparin cofactor II, and protein C/protein S. Under normal conditions, these factors limit the production of thrombin to prevent the perpetuation of coagulation and thrombus formation. Typically, after the clot has caused occlusion at the damaged site and begins to expand toward adjacent uninjured vessel

segments, the anticoagulant reactions governed by the normal endothelium become pivotal in limiting the extent of this hemostatically protective clot.

RISK FACTORS FOR VENOUS THROMBOSIS

An array of different factors contributes to the risk of VTE, and it is notable that women and men of all ages, races, and ethnicities are at risk for venous thromboembolism. The risk factors for venous thrombosis are primarily related to hypercoagulability, which can be genetic (Table 113-1) or acquired, or due to immobilization and venous stasis. Independent predictors for recurrence include increasing age, obesity, malignant neoplasm, and acute extremity paresis. It is estimated that 5–8% of the U.S. population has a genetic risk factor known to predispose to venous thrombosis. Often, multiple risk factors are present in a single individual. Significant risk is incurred by major orthopedic, abdominal, or neurologic surgeries. Cancer patients have an approximately fourfold increased risk of VTE as compared with the general population, and cancer patients with VTE have reduced survival. Hospitalized patients have a greatly increased risk of venous thrombosis with risk factors (increased age, male, ethnicity) and comorbid conditions, including infection, renal disease, and weight loss. Moderate risk is promoted by prolonged bedrest; certain types of cancer, pregnancy, hormone replacement therapy, or oral contraceptive use; and other sedentary conditions such as long-distance plane travel. It has been reported that the risk of developing a venous thromboembolic event doubles after air travel lasting 4 h, although the absolute risk remains low (1 in 6000). The relative risk of VTE among pregnant or postpartum women is 4.3, and the overall incidence (absolute risk) is 199.7 per 100,000 woman-years.

GENETICS OF VENOUS THROMBOSIS

(See Table 113-2) Less common causes of venous thrombosis are those due to genetic variants. These abnormalities include loss-of-function mutations of endogenous anticoagulants as well as gain-of-function mutations of procoagulant proteins. Heterozygous antithrombin deficiency and homozygosity of the factor V Leiden mutation significantly increase the risk of venous thrombosis. While homozygous protein C or protein S deficiencies are rare and may lead to fatal purpura fulminans, heterozygous deficiencies are associated with a moderate risk of thrombosis. Activated protein C impairs coagulation by proteolytic degradation of FVa. Patients resistant to the activity of activated protein C may have a point mutation in the FV gene located on chromosome 1, a mutant denoted factor V Leiden. Mildly increased risk has been attributed to elevated levels of procoagulant factors, as well as low levels of tissue factor pathway inhibitor. Polymorphisms of methylene tetrahydrofolate reductase as well as hyperhomocysteinemia have been shown to be independent risk factors for venous thrombosis, as well as arterial vascular disease; however, many of the initial descriptions of genetic variants and their associations with thromboembolism are being questioned in larger, more contemporary studies.

FIBRINOLYSIS AND THROMBOSIS

Specific abnormalities in the fibrinolytic system have been associated with enhanced thrombosis. Factors such as elevated levels of tissue plasminogen activator (tPA) and plasminogen activator inhibitor type 1 (PAI-

1) have been associated with decreased fibrinolytic activity and an increased risk of arterial thrombotic disease. Specific genetic variants have been associated with decreased fibrinolytic activity, including the 4G/5G insertion/deletion polymorphism in the (plasminogen activator type 1) *PAI-1* gene. Additionally, the 311-bp Alu insertion/deletion in tPA's intron 8 has been associated with enhanced thrombosis; however, genetic abnormalities have not been associated consistently with altered function or tPA levels, raising questions about the relevant pathophysiologic mechanism. Thrombin-activatable fibrinolysis inhibitor (TAFI) is a carboxypeptidase that regulates fibrinolysis; elevated plasma TAFI levels have been associated with an increased risk of both DVT and cardiovascular disease.

The metabolic syndrome also is accompanied by altered fibrinolytic activity. This syndrome, which comprises abdominal fat (central obesity), altered glucose and insulin metabolism, dyslipidemia, and hypertension, has been associated with atherothrombosis. The mechanism for enhanced thrombosis appears to be due both to altered platelet function and to a procoagulant and hypofibrinolytic state. One of the most frequently documented prothrombotic abnormalities reported in this syndrome is an increase in plasma levels of PAI-1.

In addition to contributing to platelet function, inflammation plays a role in both coagulation-dependent thrombus formation and thrombus resolution. Both polymorphonuclear neutrophils and monocytes/macrophages contribute to multiple overlapping thrombotic functions, including fibrinolysis, chemokine and cytokine production, and phagocytosis.

THE DISTINCTION BETWEEN ARTERIAL AND VENOUS THROMBOSIS

Although there is overlap, venous thrombosis and arterial thrombosis are initiated differently, and clot formation progresses by somewhat distinct pathways. In the setting of stasis or states of hypercoagulability, venous thrombosis is activated with the initiation of the coagulation cascade primarily due to exposure of tissue factor; this leads to the formation of thrombin and the subsequent conversion of fibrinogen to fibrin. In the artery, thrombin formation also occurs, but thrombosis is primarily promoted by the adhesion of platelets to an injured vessel and stimulated by exposed extracellular matrix (Figs. 113-1 and 113-2). There is wide variation in individual responses to vascular injury, an important determinant of which is the predisposition an individual has to arterial or venous thrombosis. This concept has been supported indirectly in prothrombotic animal models in which there is poor correlation between the propensity to develop venous versus arterial thrombosis.

Despite considerable progress in understanding the role of hypercoagulable states in VTE, the contribution of hypercoagulability to arterial vascular disease is much less well understood. Although specific thrombophilic conditions, such as factor V Leiden and the prothrombin G20210A mutation, are risk factors for DVT, pulmonary embolism, and other venous thromboembolic events, their contribution to arterial thrombosis is less well defined. In fact, to the contrary, many of these thrombophilic factors have not been found to be clinically important risk factors for arterial thrombotic events, such as acute coronary syndromes.

Clinically, although the pathophysiology is distinct, arterial and venous thrombosis do share common risk factors, including age, obesity, cigarette smoking, diabetes mellitus, arterial hypertension, hyperlipidemia, and metabolic syndrome. Select genetic variants, including those of the glutathione peroxidase-3 (GPx3) gene, have also been associated with arterial and venous thrombo-occlusive disease. Importantly, arterial and venous thrombosis may both be triggered by pathophysiologic stimuli responsible for activating inflammatory and oxidative pathways.

The diagnosis and treatment of ischemic heart disease are discussed in Chap. 267. Stroke diagnosis and management are discussed in Chap. 301. The diagnosis and management of DVT and pulmonary embolus are discussed in Chap. 273.

FURTHER READING

Alexander KP, Peterson ED: Minimizing the risks of anticoagulants and platelet inhibitors. Circulation 121:1960, 2010.

Baron TH et al: Current concepts: Management of antithrombotic therapy in patients undergoing invasive procedures. N Engl J Med 368:2113, 2013.

Davì G, Patrono C: Platelet activation and atherothrombosis. N Engl J Med 357:2482, 2007.

Engelmann B, Massberg S: Thrombosis as an intravascular effector of innate immunity. Nat Rev Immunol 13:34, 2013.

Furie B, Furie BC: Mechanisms of thrombus formation. N Engl J Med 359:938, 2008.

Koupenova M et al: Thrombosis and platelets: An update. Eur Heart J 38:785, 2017.

Mackman N: New insights into the mechanisms of venous thrombosis. J Clin Invest 122:2331, 2012.

Mozaffarian D et al: Heart disease and stroke statistics—A Report from the American Heart Association 2016 Update. Circulation 133:447, 2016.

Tapson VF: Acute pulmonary embolism. N Engl J Med 358:1037, 2008.

Tichelaar YI et al: Infections and inflammatory diseases as risk factors for venous thrombosis. A systematic review. Thromb Haemost 107:827, 2012.

Chapter 114: Antiplatelet, Anticoagulant, and Fibrinolytic Drugs

Jeffrey I. Weitz

INTRODUCTION

Thromboembolic disorders are major causes of morbidity and mortality. Thrombosis can occur in arteries or veins. Arterial thrombosis is the most common cause of acute myocardial infarction (MI), ischemic stroke, and limb gangrene. Venous thromboembolism encompasses deep vein thrombosis (DVT), which can lead to postthrombotic syndrome, and pulmonary embolism (PE), which can be fatal or can result in chronic thromboembolic pulmonary hypertension.

Most arterial thrombi are superimposed on disrupted atherosclerotic plaque because plaque rupture exposes thrombogenic material in the core to the blood. This material then triggers platelet aggregation and fibrin formation, which results in the generation of a platelet-rich thrombus that can temporarily or permanently occlude blood flow. In contrast, venous thrombi rarely form at sites of obvious vascular disruption. Although they can develop after surgical trauma to veins or secondary to indwelling venous catheters, venous thrombi usually originate in the valve cusps of the deep veins of the calf or in the muscular sinuses. Sluggish blood flow reduces the oxygen supply to the avascular valve cusps. Endothelial cells lining these valve cusps become activated and express adhesion molecules on their surface. Tissue factor–bearing leukocytes and microparticles adhere to these activated cells and induce coagulation. DNA extruded from neutrophils forms neutrophil extracelluar traps (NETs) that provide a scaffold that binds platelets and promotes their activation and aggregation. Local thrombus formation is exacerbated by reduced clearance of activated clotting factors as a result of impaired blood flow. If the thrombi extend from the calf veins into the popliteal and more proximal veins of the leg, thrombus fragments can dislodge, travel to the lungs, and produce a PE.

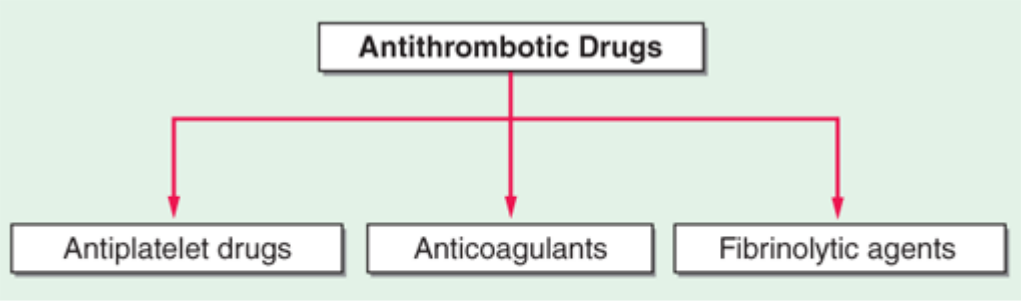
Arterial and venous thrombi are composed of platelets, fibrin, and trapped red blood cells, but the proportions differ. Arterial thrombi are rich in platelets because of the high shear in the injured arteries. In contrast, venous thrombi, which form under low shear conditions, contain relatively few platelets and are predominantly composed of fibrin and trapped red cells. Because of the predominance of platelets, arterial thrombi appear white, whereas venous thrombi are red in color, reflecting the trapped red cells.

Antithrombotic drugs are used for prevention and treatment of thrombosis. Targeting the components of thrombi, these agents include (1) antiplatelet drugs, (2) anticoagulants, and (3) fibrinolytic agents (**Fig. 114-**

1). With the predominance of platelets in arterial thrombi, strategies to attenuate arterial thrombosis focus mainly on antiplatelet agents, although, in the acute setting, may include anticoagulants and fibrinolytic agents. Anticoagulants are the mainstay of prevention and treatment of venous thromboembolism because fibrin is the predominant component of venous thrombi. Antiplatelet drugs are less effective than anticoagulants in this setting because of the limited platelet content of venous thrombi. Fibrinolytic therapy is used in selected patients with venous thromboembolism. For example, patients with massive PE can benefit from systemic or catheter-directed fibrinolytic therapy. Pharmaco-mechanical therapy also is used to restore blood flow in patients with extensive DVT involving the iliac and/or femoral veins.

FIGURE 114-1

Classification of antithrombotic drugs.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ANTIPLATELET DRUGS

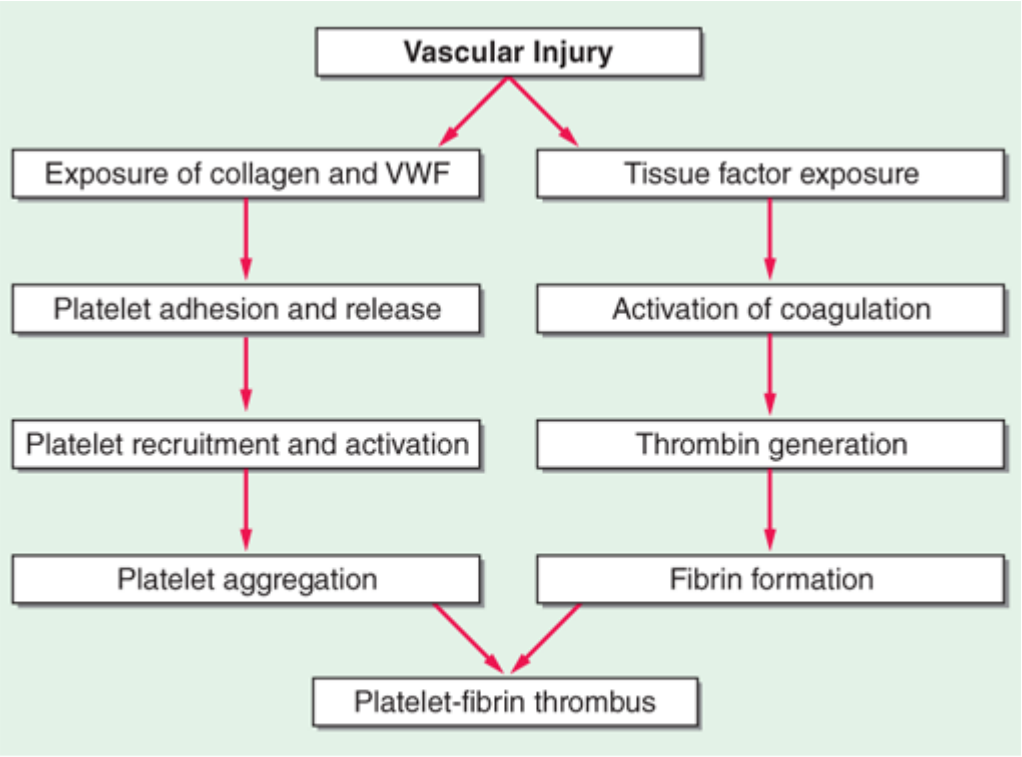
ROLE OF PLATELETS IN ARTERIAL THROMBOSIS

In healthy vasculature, circulating platelets are maintained in an inactive state by nitric oxide (NO) and prostacyclin released by endothelial cells lining the blood vessels. In addition, endothelial cells also express CD39 on their surface, a membrane-associated ecto-adenosine diphosphatase (ADPase) that degrades ADP released from activated platelets. When the vessel wall is damaged, release of these substances is impaired and subendothelial matrix is exposed. Platelets adhere to exposed collagen via $\alpha_2\beta_1$ and glycoprotein (Gp) V1 and to von Willebrand factor (VWF) via Gp Ibα and Gp IIb/IIIa ($\alpha_{IIb}\beta_3$)—receptors that are constitutively expressed on the platelet surface. Adherent platelets undergo a change in shape, secrete ADP from their dense granules, and synthesize and release thromboxane A_2. Released ADP and thromboxane A_2, which are platelet agonists, activate ambient platelets and recruit them to the site of vascular injury (**Fig. 114-2**).

FIGURE 114-2

Coordinated role of platelets and the coagulation system in thrombogenesis. Vascular injury simultaneously triggers platelet activation and aggregation and activation of the coagulation system. Platelet activation is initiated by exposure of subendothelial collagen and von Willebrand factor (VWF), onto which platelets adhere. Adherent platelets become activated and release ADP and thromboxane A_2, platelet agonists that activate ambient platelets and recruit them to the site of injury. When platelets are activated, glycoprotein

IIb/IIIa on their surface undergoes a conformational change that enables it to ligate fibrinogen and/or VWF and mediate platelet aggregation. Coagulation is triggered by tissue factor exposed at the site of injury. Tissue factor triggers thrombin generation. As a potent platelet agonist, thrombin amplifies platelet recruitment to the site of injury. Thrombin also converts fibrinogen to fibrin, and the fibrin strands then weave the platelet aggregates together to form a platelet/fibrin thrombus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Disruption of the vessel wall also exposes tissue factor–expressing cells to the blood. Tissue factor binds factor VIIa and initiates coagulation. Activated platelets potentiate coagulation by providing a surface that binds clotting factors and supports the assembly of activation complexes that enhance thrombin generation. In addition to converting fibrinogen to fibrin, thrombin serves as a potent platelet agonist and recruits more platelets to the site of vascular injury. Thrombin also amplifies its own generation by feedback activation of factors V, VIII, and XI and solidifies the fibrin network by activating factor XIII, which then cross-links the fibrin strands.

When platelets are activated, Gp IIb/IIIa, the most abundant receptor on the platelet surface, undergoes a conformational change that enables it to bind fibrinogen and, under high shear conditions, VWF. Divalent fibrinogen or multivalent VWF molecules bridge adjacent platelets together to form platelet aggregates. Fibrin strands, generated through the action of thrombin, then weave these aggregates together to form a platelet/fibrin mesh.

Antiplatelet drugs target various steps in this process. The commonly used drugs include aspirin, ADP receptor inhibitors, which include the thienopyridines (clopidogrel and prasugrel) as well as ticagrelor and cangrelor, dipyridamole, Gp IIb/IIIa antagonists, and vorapaxar.

ASPIRIN

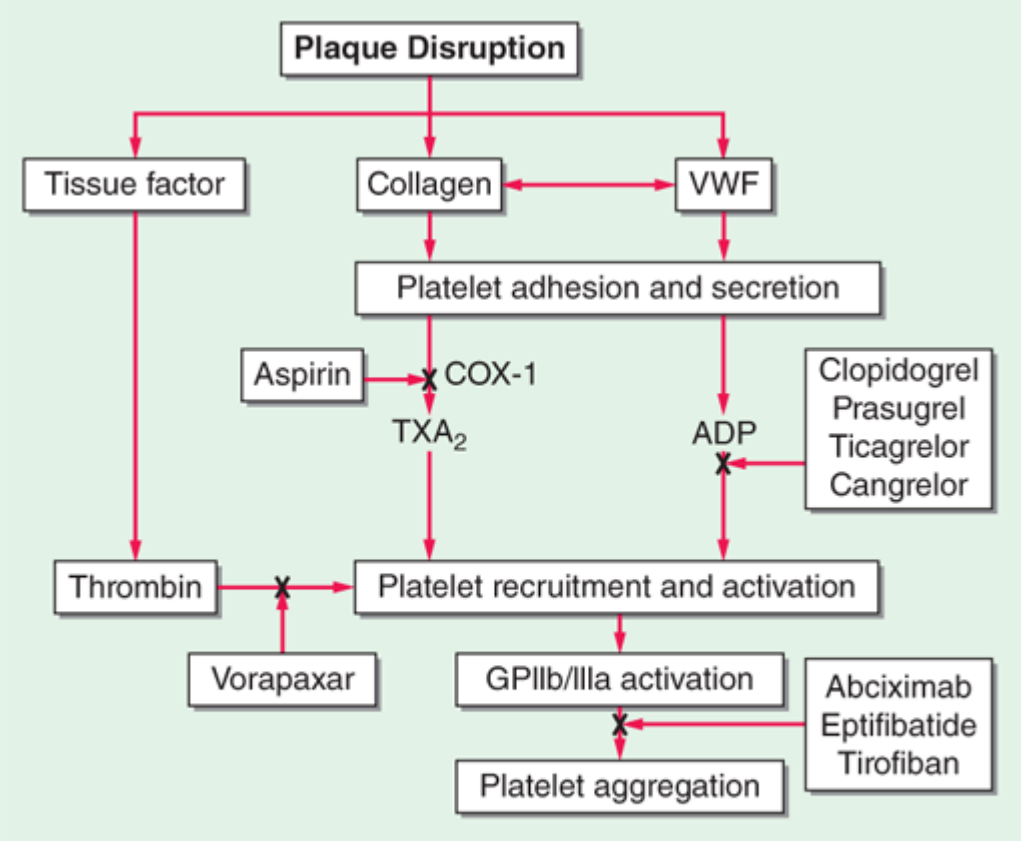
The most widely used antiplatelet agent worldwide is aspirin. As a cheap and effective antiplatelet drug, aspirin serves as the foundation of most antiplatelet strategies.

Mechanism of Action

Aspirin produces its antithrombotic effect by irreversibly acetylating and inhibiting platelet cyclooxygenase (COX)-1 **(Fig. 114-3)**, a critical enzyme in the biosynthesis of thromboxane A_2. At high doses (~1 g/d), aspirin also inhibits COX-2, an inducible COX isoform found in endothelial cells and inflammatory cells. In endothelial cells, COX-2 initiates the synthesis of prostacyclin, a potent vasodilator and inhibitor of platelet aggregation.

FIGURE 114-3

Site of action of antiplatelet drugs. Aspirin inhibits thromboxane A_2 (TXA_2) synthesis by irreversibly acetylating cyclooxygenase-1 (COX-1). Reduced TXA_2 release attenuates platelet activation and recruitment to the site of vascular injury. Clopidogrel and prasugrel irreversibly block $P2Y_{12}$, a key ADP receptor on the platelet surface; cangrelor and ticagrelor are reversible inhibitors of $P2Y_{12}$. Abciximab, eptifibatide, and tirofiban inhibit the final common pathway of platelet aggregation by blocking fibrinogen and von Willebrand factor binding to activated glycoprotein (Gp) IIb/IIIa. Vorapaxar inhibits thrombin-mediated platelet activation by targeting protease-activated receptor-1 (PAR-1), the major thrombin receptor on human platelets.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Indications

Aspirin is widely used for secondary prevention of cardiovascular events in patients with established coronary artery, cerebral artery, or peripheral artery disease. Compared with placebo in this setting, aspirin produces a 25% reduction in the risk of cardiovascular death, MI, or stroke. Use of aspirin for primary prevention is controversial. It is unclear whether the benefits of daily aspirin for primary cardiac protection outweigh its associated risks for gastrointestinal and intracerebral hemorrhage. Consequently, aspirin is no longer recommended for primary cardiac prevention unless the baseline cardiovascular risk is at least 1% per year and 10% at 10 years and patients are at low risk for bleeding.

Dosages

Aspirin is usually administered at doses of 75–325 mg once daily. Higher doses of aspirin are not more effective than lower aspirin doses, and some analyses suggest reduced efficacy with higher doses. Because the side effects of aspirin are dose-related, daily aspirin doses of 75–100 mg are recommended for most indications. When rapid platelet inhibition is required, an initial aspirin dose of at least 160 mg should be given.

Side Effects

The most common side effects are gastrointestinal and range from dyspepsia to erosive gastritis or peptic ulcers with bleeding and perforation. These side effects are dose-related. Use of enteric-coated or buffered aspirin in place of plain aspirin does not eliminate gastrointestinal side effects. The overall risk of major bleeding with aspirin is 1–3% per year. The risk of bleeding is increased two- to threefold when aspirin is given in conjunction with other antiplatelet drugs, such as clopidogrel, or with anticoagulants, such as warfarin. When dual or triple therapy is prescribed, low-dose aspirin should be given (75–100 mg daily). Eradication of *Helicobacter pylori* infection and administration of proton pump inhibitors may reduce the risk of aspirin-induced upper gastrointestinal bleeding in patients with peptic ulcer disease.

Aspirin should not be administered to patients with a history of aspirin allergy characterized by bronchospasm. This problem occurs in ~0.3% of the general population but is more common in those with chronic urticaria or asthma, particularly in individuals with nasal polyps or chronic rhinitis. Hepatic and renal toxicity are observed with aspirin overdose.

Aspirin Resistance

Clinical aspirin resistance is defined as the failure of aspirin to protect patients from ischemic vascular events. This is not a helpful definition because it is made after the event occurs. Furthermore, it is not realistic to expect aspirin, which only blocks thromboxane A_2–induced platelet activation, to prevent all vascular events.

Aspirin resistance has also been described biochemically as failure of the drug to produce its expected inhibitory effects on tests of platelet function, such as thromboxane A_2 synthesis or arachidonic acid-induced platelet aggregation. Potential causes of aspirin resistance include poor compliance, reduced absorption, drug-drug interaction with ibuprofen, and overexpression of COX-2. Unfortunately, the tests for aspirin resistance have not been well standardized, and there is little evidence that they identify patients at increased risk of recurrent vascular events, or that resistance can be reversed by giving higher doses of aspirin or by adding other antiplatelet drugs. Until such information is available, testing for aspirin resistance remains a research tool.

ADP RECEPTOR ANTAGONISTS

The ADP receptor antagonists include the thienopyridines (clopidogrel and prasugrel) as well as ticagrelor and cangrelor. All of these drugs target $P2Y_{12}$, the key ADP receptor on platelets.

Thienopyridines

MECHANISM OF ACTION

The thienopyridines are structurally related drugs that selectively inhibit ADP-induced platelet aggregation by irreversibly blocking $P2Y_{12}$ (Fig. 114-3). Clopidogrel and prasugrel are prodrugs that require metabolic activation by the hepatic cytochrome P450 (CYP) enzyme system. Prasugrel is about tenfold more potent

than clopidogrel and has a more rapid onset of action because of better absorption and more streamlined metabolic activation.

INDICATIONS

When compared with aspirin in patients with recent ischemic stroke, recent MI, or a history of peripheral arterial disease, clopidogrel reduced the risk of cardiovascular death, MI, and stroke by 8.7%. Therefore, clopidogrel is more effective than aspirin but is also more expensive. In some patients, clopidogrel and aspirin are combined to capitalize on their capacity to block complementary pathways of platelet activation. For example, the combination of aspirin plus clopidogrel is recommended for at least 4 weeks after implantation of a bare metal stent in a coronary artery and for at least 6 months in those with a drug-eluting stent. Concerns about late in-stent thrombosis with drug-eluting stents have led some experts to recommend longer-term use of clopidogrel plus aspirin for the latter indication in patients without bleeding complications.

The combination of clopidogrel and aspirin is also effective in patients with unstable angina. Thus, in 12,562 such patients, the risk of cardiovascular death, MI, or stroke was 9.3% in those randomized to the combination of clopidogrel and aspirin and 11.4% in those given aspirin alone. This 20% relative risk reduction with combination therapy was highly statistically significant. However, combining clopidogrel with aspirin increases the risk of major bleeding to about 2% per year. This bleeding risk persists even if the daily dose of aspirin is ≤100 mg. Therefore, the combination of clopidogrel and aspirin should only be used when there is a clear benefit. For example, this combination has not proven to be superior to clopidogrel alone in patients with acute ischemic stroke or to aspirin alone for primary prevention in those at risk for cardiovascular events.

Prasugrel was compared with clopidogrel in 13,608 patients with acute coronary syndromes who were scheduled to undergo percutaneous coronary intervention. The incidence of the primary efficacy endpoint, a composite of cardiovascular death, MI, or stroke, was significantly lower with prasugrel than with clopidogrel (9.9% and 12.1%, respectively), mainly reflecting a reduction in the incidence of nonfatal MI. The incidence of stent thrombosis also was significantly lower with prasugrel (1.1% and 2.4%, respectively). However, these advantages were at the expense of significantly higher rates of fatal bleeding (0.4% and 0.1%, respectively) and life-threatening bleeding (1.4% and 0.9%, respectively) with prasugrel. Because patients older than age 75 years and those with a history of prior stroke or transient ischemic attack have a particularly high risk of bleeding, prasugrel should generally be avoided in older patients, and the drug is contraindicated in those with a history of cerebrovascular disease. Caution is required if prasugrel is used in patients weighing <60 kg or in those with renal impairment.

When prasugrel was compared with clopidogrel in 7243 patients with unstable angina or MI without ST-segment elevation, prasugrel failed to reduce the rate of the primary efficacy endpoint, which was a composite of cardiovascular death, MI, and stroke. Because of the negative results of this study, prasugrel is reserved for patients undergoing percutaneous coronary intervention. In this setting, prasugrel is usually given in conjunction with aspirin. To reduce the risk of bleeding, the daily aspirin dose should be ≤100 mg.

DOSING

Clopidogrel is given once daily at a dose of 75 mg. Loading doses of clopidogrel are given when rapid ADP receptor blockade is desired. For example, patients undergoing coronary stenting are often given a loading dose of 300–600 mg, which produces inhibition of ADP-induced platelet aggregation in about 4–6 h. After a loading dose of 60 mg, prasugrel is given once daily at a dose of 10 mg. Patients older than age 75 years or weighing less than 60 kg should receive a lower daily prasugrel dose of 5 mg.

SIDE EFFECTS

The most common side effect of clopidogrel and prasugrel is bleeding. Because of its greater potency, bleeding is more common with prasugrel than clopidogrel. To reduce the risk of bleeding, clopidogrel and prasugrel should be stopped 5–7 days before major surgery. In patients taking clopidogrel or prasugrel who present with serious bleeding, platelet transfusion may be helpful.

Hematologic side effects, including neutropenia, thrombocytopenia, and thrombotic thrombocytopenic purpura, are rare.

THIENOPYRIDINE RESISTANCE

The capacity of clopidogrel to inhibit ADP-induced platelet aggregation varies among subjects. This variability reflects, at least in part, genetic polymorphisms in the CYP isoenzymes involved in the metabolic activation of clopidogrel. Most important of these is CYP2C19. Clopidogrel-treated patients with the loss-of-function *CYP2C19* 2* allele exhibit reduced platelet inhibition compared with those with the wild-type *CYP2C19* 1* allele and experience a higher rate of cardiovascular events. This is important because estimates suggest that up to 25% of whites, 30% of African Americans, and 50% of Asians carry the loss-of-function allele, which would render them resistant to clopidogrel. Even patients with the reduced function *CYP2C19* 3*, ** 4*, or ** 5* alleles may derive less benefit from clopidogrel than those with the full-function *CYP2C19* 1* allele. Concomitant administration of clopidogrel with proton pump inhibitors, which are inhibitors of *CYP2C19*, produces a small reduction in the inhibitory effects of clopidogrel on ADP-induced platelet aggregation. The extent to which this interaction increases the risk of cardiovascular events remains controversial.

In contrast to their effect on the metabolic activation of clopidogrel, *CYP2C19* polymorphisms appear to be less important determinants of the activation of prasugrel. Thus, no association was detected between the loss-of-function allele and decreased platelet inhibition or increased rate of cardiovascular events with prasugrel. The observation that genetic polymorphisms affecting clopidogrel absorption or metabolism influence clinical outcomes raises the possibilities that pharmacogenetic profiling may be useful to identify clopidogrel-resistant patients and that point-of-care assessment of the extent of clopidogrel-induced platelet inhibition may help detect patients at higher risk for subsequent cardiovascular events. Clinical trials designed to evaluate these possibilities have thus far been negative. Although administration of higher doses of clopidogrel can overcome a reduced response to clopidogrel, the clinical benefit of this approach is uncertain. Instead, prasugrel or ticagrelor may be better choices for these patients.

Ticagrelor

As an orally active inhibitor of $P2Y_{12}$, ticagrelor differs from the thienopyridines in that ticagrelor does not require metabolic activation and it produces reversible inhibition of the ADP receptor.

MECHANISM OF ACTION

Like the thienopyridines, ticagrelor inhibits $P2Y_{12}$. Because it does not require metabolic activation, ticagrelor has a more rapid onset and offset of action than clopidogrel, and it produces greater and more predictable inhibition of ADP-induced platelet aggregation than clopidogrel.

Indications

When compared with clopidogrel in patients with acute coronary syndromes, ticagrelor produced a greater reduction in the primary efficacy endpoint—a composite of cardiovascular death, MI, and stroke at 1 year—than clopidogrel (9.8% and 11.7%, respectively; $p = 0.001$). This difference reflected a significant reduction in both cardiovascular death (4.0% and 5.1%, respectively; $p = 0.001$) and MI (5.8% and 6.9%, respectively; $p = 0.005$) with ticagrelor compared with clopidogrel. Rates of stroke were similar with ticagrelor and clopidogrel (1.5% and 1.3%, respectively), and no difference in rates of major bleeding was noted. When minor bleeding was added to the major bleeding results, however, ticagrelor showed an increase relative to clopidogrel (16.1% and 14.6%, respectively; $p = 0.008$). Ticagrelor also was superior to clopidogrel in patients with acute coronary syndrome who underwent percutaneous coronary intervention or cardiac surgery. Based on these observations, guidelines give preference to ticagrelor over clopidogrel, particularly in higher risk patients.

DOSING

Ticagrelor is initiated with an oral loading dose of 180 mg followed by 90 mg twice daily. The dose does not require adjustment in patients with renal impairment, but the drug should be used with caution in patients with hepatic disease and in those receiving potent inhibitors or inducers of CYP3A4 because ticagrelor is metabolized in the liver via CYP3A4. Ticagrelor is usually administered in conjunction with aspirin; the daily aspirin dose should not exceed 100 mg.

SIDE EFFECTS

In addition to bleeding, the most common side effects of ticagrelor are dyspnea, which can occur in up to 15% of patients, and asymptomatic ventricular pauses. The dyspnea, which tends to occur soon after initiating ticagrelor, is usually self-limiting and mild in intensity. The mechanism responsible for this side effect is unknown.

To reduce the risk of bleeding, ticagrelor should be stopped 5–7 days prior to major surgery. Platelet transfusions are unlikely to be of benefit in patients with ticagrelor-related bleeding because the drug will bind to $P2Y_{12}$ on the transfused platelets.

Cangrelor

Cangrelor is a rapidly acting reversible inhibitor of P2Y12 that is administered intravenously. It has an immediate onset of action, a half-life of 3–5 min, and an offset of action within an hour. Cangrelor is licensed

for use in patients undergoing percutaneous coronary intervention and produces rapid ADP receptor blockade in those who have not received pre-treatment with clopidogrel, prasugrel, or ticagrelor.

DIPYRIDAMOLE

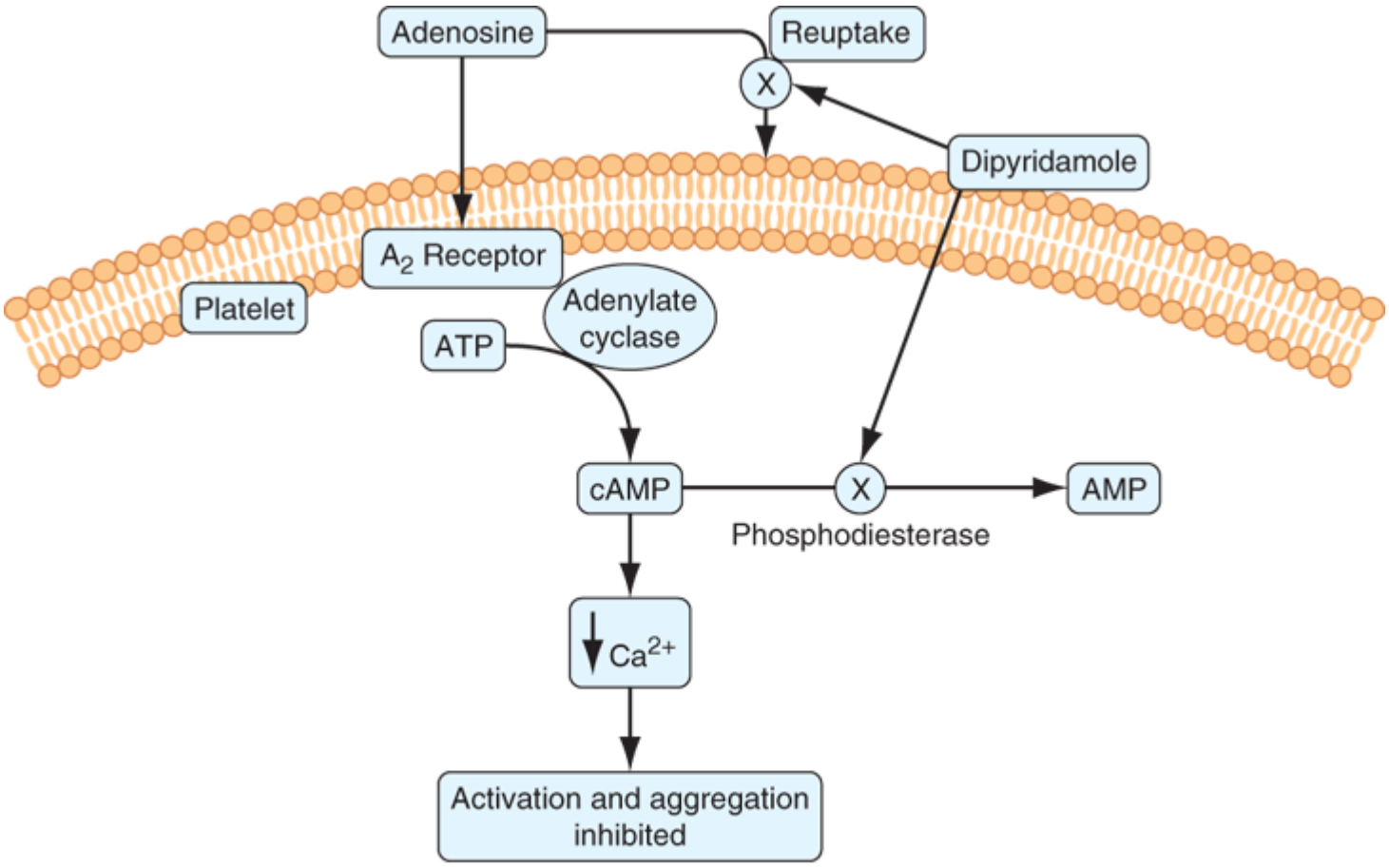
Dipyridamole is a relatively weak antiplatelet agent on its own, but an extended-release formulation of dipyridamole combined with low-dose aspirin, a preparation known as *Aggrenox*, is used for prevention of stroke in patients with transient ischemic attacks.

Mechanism of Action

By inhibiting phosphodiesterase, dipyridamole blocks the breakdown of cyclic adenosine monophosphate (AMP). Increased levels of cyclic AMP reduce intracellular calcium and inhibit platelet activation. Dipyridamole also blocks the uptake of adenosine by platelets and other cells. This produces a further increase in local cyclic AMP levels because the platelet adenosine A_2 receptor is coupled to adenylate cyclase (Fig. 114-4).

FIGURE 114-4

Mechanism of action of dipyridamole. Dipyridamole increases levels of cyclic AMP (cAMP) in platelets by (1) blocking the reuptake of adenosine and (2) inhibiting phosphodiesterase-mediated cyclic AMP degradation. By promoting calcium uptake, cyclic AMP reduces intracellular levels of calcium. This, in turn, inhibits platelet activation and aggregation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Indications

Dipyridamole plus aspirin was compared with aspirin or dipyridamole alone, or with placebo, in patients with an ischemic stroke or transient ischemic attack. The combination reduced the risk of stroke by 22.1% compared with aspirin and by 24.4% compared with dipyridamole. A second trial compared dipyridamole plus aspirin with aspirin alone for secondary prevention in patients with ischemic stroke. Vascular death, stroke, or MI occurred in 13% of patients given combination therapy and in 16% of those treated with aspirin alone. Another trial randomized 20,332 patients with noncardioembolic ischemic stroke to either Aggrenox or clopidogrel. The primary efficacy endpoint of recurrent stroke occurred in 9.0% of those given Aggrenox and in 8.8% of patients treated with clopidogrel. Although this difference was not statistically significant, the study failed to meet the prespecified margin to claim noninferiority of Aggrenox relative to clopidogrel. These results have dampened enthusiasm for the use of Aggrenox.

Because of its vasodilatory effects and the paucity of data supporting the use of dipyridamole in patients with symptomatic coronary artery disease, Aggrenox should not be used for stroke prevention in such patients. Clopidogrel is a better choice in this setting.

Dosing

Aggrenox is given twice daily. Each capsule contains 200 mg of extended-release dipyridamole and 25 mg of aspirin.

Side Effects

Because dipyridamole has vasodilatory effects, it must be used with caution in patients with coronary artery disease. Gastrointestinal complaints, headache, facial flushing, dizziness, and hypotension can also occur. These symptoms often subside with continued use of the drug.

GP IIB/IIIA RECEPTOR ANTAGONISTS

As a class, parenteral Gp IIb/IIIa receptor antagonists have an established niche in patients with acute coronary syndromes. The three agents in this class are abciximab, eptifibatide, and tirofiban.

Mechanism of Action

A member of the integrin family of adhesion receptors, Gp IIb/IIIa is found on the surface of platelets and megakaryocytes. With about 80,000 copies per platelet, Gp IIb/IIIa is the most abundant receptor. Consisting of a noncovalently linked heterodimer, Gp IIb/IIIa is inactive on resting platelets. When platelets are activated, inside-outside signal transduction pathways trigger a conformational activation of the receptor. Once activated, Gp IIb/IIIa binds adhesive molecules, such as fibrinogen and, under high shear conditions, VWF. Binding is mediated by the Arg-Gly-Asp (RGD) sequence found on the α chains of fibrinogen and on VWF, and by the Lys-Gly-Asp (KGD) sequence located within a unique dodecapeptide domain on the γ chains of fibrinogen. Once bound, fibrinogen and/or VWF bridge adjacent platelets together to induce platelet aggregation.

Although abciximab, eptifibatide, and tirofiban all target the Gp IIb/IIIa receptor, they are structurally and pharmacologically distinct (**Table 114-1**). Abciximab is a Fab fragment of a humanized murine monoclonal antibody directed against the activated form of Gp IIb/IIIa. Abciximab binds to the activated receptor with high affinity and blocks the binding of adhesive molecules. In contrast, eptifibatide and tirofiban are synthetic small molecules. Eptifibatide is a cyclic heptapeptide that binds Gp IIb/IIIa because it incorporates the KGD motif, whereas tirofiban is a nonpeptidic tyrosine derivative that acts as an RGD mimetic. Abciximab has a long half-life and can be detected on the surface of platelets for up to 2 weeks; eptifibatide and tirofiban have short half-lives.

TABLE 114-1

Features of GpIIb/IIIa Antagonists

Feature	Abciximab	Eptifibatide	Tirofiban
Description	Fab fragment of humanized mouse monoclonal antibody	Cyclical KGD-containing heptapeptide	Nonpeptidic RGD mimetic
Specific for Gp IIb/IIIa	No	Yes	Yes
Plasma half-life	Short (min)	Long (2.5 h)	Long (2.0 h)
Platelet-bound half-life	Long (days)	Short (s)	Short (s)
Renal clearance	No	Yes	Yes

Indications

Abciximab and eptifibatide are used in patients undergoing percutaneous coronary interventions, particularly those who have not been pretreated with an ADP receptor antagonist. Tirofiban is used in high-risk patients with unstable angina. Eptifibatide also can be used for this indication.

Dosing

All of the Gp IIb/IIIa antagonists are given as an IV bolus followed by an infusion. The recommended dose of abciximab is a bolus of 0.25 mg/kg followed by an infusion of 0.125 µg/kg per minute to a maximum of 10 µg/kg for 12 h. Eptifibatide is given as two 180 µg/kg boluses given 10 min apart, followed by an infusion of 2.0 µg/kg per minute for at least 12 h. Tirofiban is given as a bolus of 25 µg/kg; the drug is then continued at a rate of 0.15 µg/kg per minute for up to 18 h. Because these agents are cleared by the kidneys, the doses of eptifibatide and tirofiban must be reduced in patients with renal insufficiency. Thus, the eptifibatide infusion is reduced to 1 µg/kg per minute in patients with a creatinine clearance below 50 mL/min, whereas the post loading dose of tirofiban is cut in half for patients with a creatinine clearance below 60 mL/min.

Side Effects

In addition to bleeding, thrombocytopenia is the most serious complication. Thrombocytopenia is immune-mediated and is caused by antibodies directed against neoantigens on Gp IIb/IIIa that are exposed upon antagonist binding. With abciximab, thrombocytopenia occurs in up to 5% of patients. Thrombocytopenia is severe in ~1% of these individuals. Thrombocytopenia is less common with the other two agents, occurring in ~1% of patients.

VORAPAXAR

An orally active PAR-1 antagonist, vorapaxar is slowly eliminated with a half-life of about 200 h. When compared with placebo in 12,944 patients with acute coronary syndrome without ST-segment elevation, vorapaxar failed to significantly reduce the primary efficacy endpoint, a composite of cardiovascular death, MI, stroke, recurrent ischemia requiring rehospitalization, and urgent coronary revascularization. Moreover, vorapaxar was associated with increased rates of bleeding, including intracranial bleeding.

In a second trial, vorapaxar was compared with placebo for secondary prevention in 26,449 patients with prior MI, ischemic stroke, or peripheral arterial disease. Overall, vorapaxar reduced the risk for cardiovascular death, MI, or stroke by 13%, but doubled the risk of intracranial bleeding. In the prespecified subgroup of 17,779 patients with prior MI, however, vorapaxar reduced the risk for cardiovascular death, MI, or stroke by 20% compared with placebo (from 9.7% to 8.1%, respectively). The rate of intracranial hemorrhage was higher with vorapaxar than with placebo (0.6% and 0.4%, respectively; $p = 0.076$) as was the rate of moderate or severe bleeding (3.4% and 2.1%, respectively; $P < 0.0001$). Based on these data, vorapaxar is licensed for patients younger than 75 years with MI who have no history of stroke, transient ischemic attack, or history of intracranial bleeding and weigh more than 60 kg.

ANTICOAGULANTS

There are both parenteral and oral anticoagulants. The parenteral anticoagulants include heparin, low-molecular-weight heparin (LMWH), fondaparinux (a synthetic pentasaccharide), lepirudin, desirudin, bivalirudin, and argatroban. Currently available oral anticoagulants include warfarin; dabigatran etexilate, an oral thrombin inhibitor; and rivaroxaban, apixaban, and edoxaban, which are oral factor Xa inhibitors.

PARENTERAL ANTICOAGULANTS

Heparin

A sulfated polysaccharide, heparin is isolated from mammalian tissues rich in mast cells. Most commercial heparin is derived from porcine intestinal mucosa and is a polymer of alternating D-glucuronic acid and *N*-acetyl-D-glucosamine residues.

MECHANISM OF ACTION

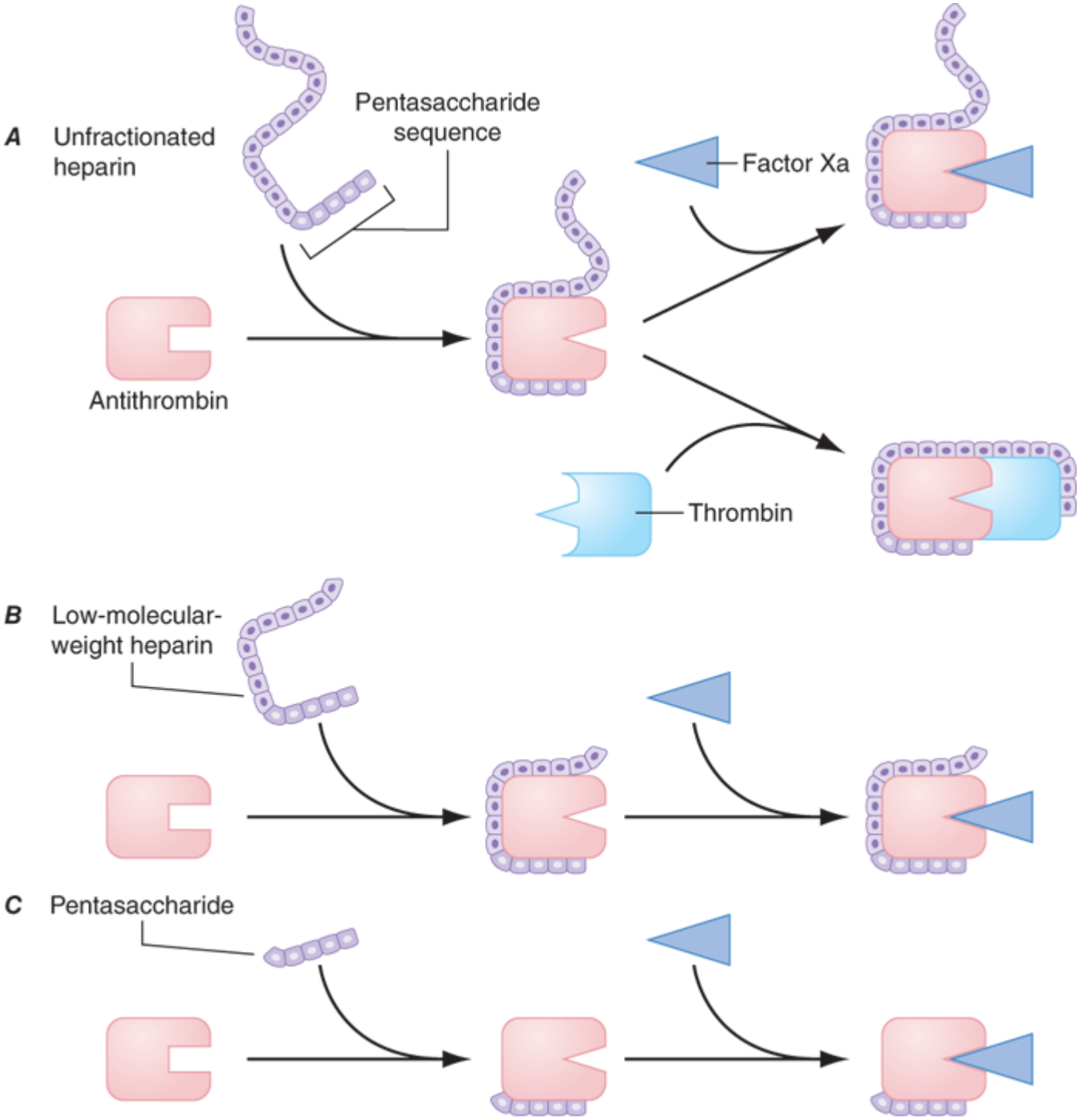
Heparin acts as an anticoagulant by activating antithrombin (previously known as antithrombin III) and accelerating the rate at which antithrombin inhibits clotting enzymes, particularly thrombin and factor Xa.

Antithrombin, the obligatory plasma cofactor for heparin, is a member of the serine protease inhibitor (serpin) superfamily. Synthesized in the liver and circulating in plasma at a concentration of 2.6 ± 0.4 μM, antithrombin acts as a suicide substrate for its target enzymes.

To activate antithrombin, heparin binds to the serpin via a unique pentasaccharide sequence that is found on one-third of the chains of commercial heparin (**Fig. 114-5**). Heparin chains without this pentasaccharide sequence have little or no anticoagulant activity. Once bound to antithrombin, heparin induces a conformational change in the reactive center loop of antithrombin that renders it more readily accessible to its target proteases. This conformational change enhances the rate at which antithrombin inhibits factor Xa by at least two orders of magnitude but has little effect on the rate of thrombin inhibition. To catalyze thrombin inhibition, heparin serves as a template that binds antithrombin and thrombin simultaneously. Formation of this ternary complex brings the enzyme in close apposition to the inhibitor, thereby promoting the formation of a stable covalent thrombin-antithrombin complex.

FIGURE 114-5

Mechanism of action of heparin, low-molecular-weight heparin (LMWH), and fondaparinux, a synthetic pentasaccharide. ***A.*** Heparin binds to antithrombin via its pentasaccharide sequence. This induces a conformational change in the reactive center loop of antithrombin that accelerates its interaction with factor Xa. To potentiate thrombin inhibition, heparin must simultaneously bind to antithrombin and thrombin. Only heparin chains composed of at least 18 saccharide units, which correspond to a molecular weight of 5400, are of sufficient length to perform this bridging function. With a mean molecular weight of 15,000, all of the heparin chains are long enough to do this. ***B***. LMWH has greater capacity to potentiate factor Xa inhibition by antithrombin than thrombin because, with a mean molecular weight of 4500–5000, at least half of the LMWH chains are too short to bridge antithrombin to thrombin. ***C.*** The pentasaccharide only accelerates factor Xa inhibition by antithrombin because the pentasaccharide is too short to bridge antithrombin to thrombin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Only pentasaccharide-containing heparin chains composed of at least 18 saccharide units (which correspond to a molecular weight of 5400) are of sufficient length to bridge thrombin and antithrombin together. With a mean molecular weight of 15,000, and a range of 5000–30,000, almost all of the chains of unfractionated heparin are long enough to do so. Consequently, by definition, heparin has equal capacity to promote the inhibition of thrombin and factor Xa by antithrombin and is assigned an anti-factor Xa to anti-factor IIa (thrombin) ratio of 1:1.

Heparin causes the release of tissue factor pathway inhibitor (TFPI) from the endothelium. A factor Xa–dependent inhibitor of tissue factor–bound factor VIIa, TFPI may contribute to the antithrombotic activity of heparin. Longer heparin chains induce the release of more TFPI than shorter ones.

PHARMACOLOGY

Heparin must be given parenterally. It is usually administered SC or by continuous IV infusion. When used for therapeutic purposes, the IV route is most often employed. If heparin is given SC for treatment of thrombosis, the dose of heparin must be high enough to overcome the limited bioavailability associated with this method of delivery.

In the circulation, heparin binds to the endothelium and to plasma proteins other than antithrombin. Heparin binding to endothelial cells explains its dose-dependent clearance. At low doses, the half-life of heparin is short because it binds rapidly to the endothelium. With higher doses of heparin, the half-life is longer because heparin is cleared more slowly once the endothelium is saturated. Clearance is mainly extra renal; heparin binds to macrophages, which internalize and depolymerize the long heparin chains and secrete shorter chains back into the circulation. Because of its dose-dependent clearance mechanism, the plasma half-life of heparin ranges from 30 to 60 min with bolus IV doses of 25 and 100 units/kg, respectively.

Once heparin enters the circulation, it binds to plasma proteins other than antithrombin, a phenomenon that reduces its anticoagulant activity. Some of the heparin-binding proteins found in plasma are acute-phase reactants whose levels are elevated in ill patients. Others, such as high-molecular-weight multimers of VWF, are released from activated platelets or endothelial cells. Activated platelets also release platelet factor 4 (PF4), a highly cationic protein that binds heparin with high affinity. The large amounts of PF4 found in the vicinity of platelet-rich arterial thrombi can neutralize the anticoagulant activity of heparin. This phenomenon may attenuate heparin’s capacity to suppress thrombus growth.

Because the levels of heparin-binding proteins in plasma vary from person to person, the anticoagulant response to fixed or weight-adjusted doses of heparin is unpredictable. Consequently, coagulation monitoring is essential to ensure that a therapeutic response is obtained. This is particularly important when heparin is administered for treatment of established thrombosis because a subtherapeutic anticoagulant response may render patients at risk for recurrent thrombosis, whereas excessive anticoagulation increases the risk of bleeding.

MONITORING THE ANTICOAGULANT EFFECT

Heparin therapy can be monitored using the activated partial thromboplastin time (aPTT) or anti-factor Xa level. Although the aPTT is the test most often used for this purpose, there are problems with this assay. aPTT reagents vary in their sensitivity to heparin, and the type of coagulometer used for testing can influence the results. Consequently, laboratories must establish a therapeutic aPTT range with each reagent-coagulometer combination by measuring the aPTT and anti-factor Xa level in plasma samples collected from heparin-treated patients. For most of the aPTT reagents and coagulometers in current use, therapeutic heparin levels are achieved with a two- to threefold prolongation of the aPTT. Anti-factor Xa levels also can be used to monitor heparin therapy. With this test, therapeutic heparin levels range from 0.3 to 0.7 units/mL.

Up to 25% of heparin-treated patients with venous thromboembolism require >35,000 units/d to achieve a therapeutic aPTT. These patients are considered heparin-resistant. It is useful to measure anti-factor Xa levels in heparin-resistant patients because many will have a therapeutic anti-factor Xa level despite a subtherapeutic aPTT. This dissociation in test results occurs because elevated plasma levels of fibrinogen and factor VIII, both of which are acute-phase proteins, shorten the aPTT but have no effect on anti-factor Xa levels. Heparin therapy in patients who exhibit this phenomenon is best monitored using anti-factor Xa levels instead of the aPTT. Patients with congenital or acquired antithrombin deficiency and those with elevated levels of heparin-binding proteins may also need high doses of heparin to achieve a therapeutic aPTT or anti-factor Xa level. If there is good correlation between the aPTT and the anti-factor Xa levels, either test can be used to monitor heparin therapy.

DOSING

For prophylaxis, heparin is usually given in fixed doses of 5000 units SC two or three times daily. With these low doses, coagulation monitoring is unnecessary. In contrast, monitoring is essential when the drug is given in therapeutic doses. Fixed-dose or weight-based heparin nomograms are used to standardize heparin dosing and to shorten the time required to achieve a therapeutic anticoagulant response. At least two heparin nomograms have been validated in patients with venous thromboembolism and reduce the time required to achieve a therapeutic aPTT. Weight-adjusted heparin nomograms have also been evaluated in patients with acute coronary syndromes. After an IV heparin bolus of 5000 units or 70 units/kg, a heparin infusion rate of 12–15 units/kg per hour is usually administered. In contrast, weight-adjusted heparin nomograms for patients with venous thromboembolism use an initial bolus of 5000 units or 80 units/kg, followed by an infusion of 18 units/kg per hour. Thus, patients with venous thromboembolism appear to require higher doses of heparin to achieve a therapeutic aPTT than do patients with acute coronary syndromes. This may reflect differences in the thrombus burden. Heparin binds to fibrin, and the amount of fibrin in patients with extensive DVT is greater than that in those with coronary thrombosis.

Heparin manufacturers in North America have traditionally measured heparin potency in USP units, with a unit defined as the concentration of heparin that prevents 1 mL of citrated sheep plasma from clotting for 1 h after calcium addition. In contrast, manufacturers in Europe measured heparin potency with anti-Xa assays using an international heparin standard for comparison. Because of problems with heparin contamination with oversulfated chondroitin sulfate, which the USP assay system does not detect, North American heparin manufacturers now use the anti-Xa assay to assess heparin potency. The use of international units in place of USP units results in a 10% reduction in heparin doses, which is a difference unlikely to affect patient care because monitoring ensures that a therapeutic anticoagulant response has been achieved.

LIMITATIONS

Heparin has pharmacokinetic and biophysical limitations (**Table 114-2**). The pharmacokinetic limitations reflect heparin's propensity to bind in a pentasaccharide-independent fashion to cells and plasma proteins. Heparin binding to endothelial cells explains its dose-dependent clearance, whereas binding to plasma proteins results in a variable anticoagulant response and can lead to heparin resistance.

TABLE 114-2

Pharmacokinetic and Biophysical Limitations of Heparin

Limitations	Mechanism
Poor bioavailability at low doses	Binds to endothelial cells and macrophages
Dose-dependent clearance	Binds to macrophages
Variable anticoagulant response	Binds to plasma proteins whose levels vary from patient to patient
Reduced activity in the vicinity of platelet-rich thrombi	Neutralized by platelet factor 4 released from activated platelets
Limited activity against factor Xa incorporated in the prothrombinase complex and thrombin bound to fibrin	Reduced capacity of heparin-antithrombin complex to inhibit factor Xa bound to activated platelets and thrombin bound to fibrin

The biophysical limitations of heparin reflect the inability of the heparin-antithrombin complex to inhibit factor Xa when it is incorporated into the prothrombinase complex, the complex that converts prothrombin to thrombin, and to inhibit thrombin bound to fibrin. Consequently, factor Xa bound to activated platelets within platelet-rich thrombi has the potential to generate thrombin, even in the face of heparin. Once this thrombin binds to fibrin, it too is protected from inhibition by the heparin-antithrombin complex. Clot-associated thrombin can then trigger thrombus growth by locally activating platelets and amplifying its own generation through feedback activation of factors V, VIII, and XI. Further compounding the problem is the potential for heparin neutralization by the high concentrations of PF4 released from activated platelets within the platelet-rich thrombus.

SIDE EFFECTS

The most common side effect of heparin is bleeding. Other complications include thrombocytopenia, osteoporosis, and elevated levels of transaminases.

Bleeding

The risk of bleeding rises as the dose of heparin is increased. Concomitant administration of drugs that affect hemostasis, such as antiplatelet or fibrinolytic agents, increases the risk of bleeding, as does recent surgery or trauma. Heparin-treated patients with serious bleeding can be given protamine sulfate to neutralize the heparin. Protamine sulfate, a mixture of basic polypeptides isolated from salmon sperm, binds heparin with high affinity, and the resultant protamine-heparin complexes are then cleared. Typically, 1 mg of protamine sulfate neutralizes 100 units of heparin. Protamine sulfate is given IV. Anaphylactoid reactions to protamine sulfate can occur, and drug administration by slow IV infusion is recommended to reduce the risk.

THROMBOCYTOPENIA

Heparin can cause thrombocytopenia. Heparin-induced thrombocytopenia (HIT) is an antibody-mediated process that is triggered by antibodies directed against neoantigens on PF4 that are exposed when heparin binds to this protein. These antibodies, which are usually of the IgG isotype, bind simultaneously to the heparin-PF4 complex and to platelet Fc receptors. Such binding activates the platelets and generates platelet microparticles. Circulating microparticles are prothrombotic because they express anionic phospholipids on their surface and can bind clotting factors and promote thrombin generation.

The clinical features of HIT are illustrated in **Table 114-3**. Typically, HIT occurs 5–10 days after initiation of heparin therapy, but it can manifest earlier if the patient has received heparin within the past 3 months. A platelet count <100,000/μL or a 50% decrease in the platelet count from the pretreatment value should raise the suspicion of HIT in those receiving heparin. HIT is more common in surgical patients than in medical patients and, like many autoimmune disorders, occurs more frequently in females than in males.

TABLE 114-3

Features of Heparin-Induced Thrombocytopenia

Features	Details
Thrombocytopenia	Platelet count of ≤100,000/μL or a decrease in platelet count of ≥50%
Timing	Platelet count falls 5–10 days after starting heparin
Type of heparin	More common with unfractionated heparin than low-molecular-weight heparin
Type of patient	More common in surgical patients and patients with cancer than general medical patients; more common in women than in men
Thrombosis	Venous thrombosis more common than arterial thrombosis

HIT can be associated with thrombosis, either arterial or venous. Venous thrombosis, which manifests as DVT and/or PE, is more common than arterial thrombosis. Arterial thrombosis can manifest as ischemic stroke or acute MI. Rarely, platelet-rich thrombi in the distal aorta or iliac arteries can cause critical limb ischemia.

The diagnosis of HIT is established using enzyme-linked immuno- assays to detect antibodies against heparin-PF4 complexes or with platelet activation assays. Enzyme-linked assays are sensitive but can be positive in the absence of any clinical evidence of HIT. However, because of its sensitivity, negative enzyme-linked immunoassay excludes the diagnosis of HIT. The most specific diagnostic test for HIT is the serotonin release assay. This test is performed by quantifying serotonin release when washed platelets loaded with labeled serotonin are exposed to patient serum in the absence or presence of varying concentrations of

heparin. If the patient serum contains the HIT antibody, heparin addition induces platelet activation and serotonin release.

Management of HIT is outlined in **Table 114-4**. Heparin should be stopped in patients with suspected or documented HIT, and an alternative anticoagulant should be administered to prevent or treat thrombosis. The agents most often used for this indication are parenteral direct thrombin inhibitors, such as lepirudin, argatroban, or bivalirudin, or factor Xa inhibitors, such as fondaparinux or rivaroxaban.

TABLE 114-4

Management of Heparin-Induced Thrombocytopenia

Stop all heparin.
Give an alternative anticoagulant, such as lepirudin, argatroban, bivalirudin, fondaparinux or rivaroxaban.
Do not give platelet transfusions.
Do not give warfarin until the platelet count returns to its baseline level. If warfarin was administered, give vitamin K to restore the INR to normal.
Evaluate for thrombosis, particularly deep vein thrombosis.

Abbreviation: INR, international normalized ratio.

Patients with HIT, particularly those with associated thrombosis, often have evidence of increased thrombin generation that can lead to consumption of protein C. If these patients are given warfarin without a concomitant parenteral anticoagulant to inhibit thrombin or thrombin generation, the further decrease in protein C levels induced by the vitamin K antagonist can trigger skin necrosis. To avoid this problem, patients with HIT should be treated with a direct thrombin inhibitor or with fondaparinux or rivaroxaban until the platelet count returns to normal levels. At this point, low-dose warfarin therapy can be introduced, and the thrombin inhibitor can be discontinued when the anticoagulant response to warfarin has been therapeutic for at least 2 days. Alternatively, the patient can be maintained or switched to rivaroxaban.

Osteoporosis

Treatment with therapeutic doses of heparin for >1 month can cause a reduction in bone density. This complication has been reported in up to 30% of patients given long-term heparin therapy, and symptomatic vertebral fractures occur in 2–3% of these individuals.

Heparin causes bone loss both by decreasing bone formation and by enhancing bone resorption. Thus, heparin affects the activity of both osteoblasts and osteoclasts.

Elevated Levels of Transaminases

Therapeutic doses of heparin are frequently associated with modest elevations in the serum levels of hepatic transaminases without a concomitant increase in the level of bilirubin. The levels of transaminases rapidly return to normal when the drug is stopped. The mechanism responsible for this phenomenon is unknown.

Low-Molecular-Weight Heparin

Consisting of smaller fragments of heparin, LMWH is prepared from unfractionated heparin by controlled enzymatic or chemical depolymerization. The mean molecular weight of LMWH is about 5000, one-third the mean molecular weight of unfractionated heparin. LMWH has advantages over heparin (**Table 114-5**) and has replaced heparin for most indications.

TABLE 114-5

Advantages of LMWH over Heparin

Advantage	Consequence
Better bioavailability and longer half-life after subcutaneous injection	Can be given subcutaneously once or twice daily for both prophylaxis and treatment
Dose-independent clearance	Simplified dosing
Predictable anticoagulant response	Coagulation monitoring is unnecessary in most patients
Lower risk of heparin-induced thrombocytopenia	Safer than heparin for short- or long-term administration
Lower risk of osteoporosis	Safer than heparin for extended administration

Abbreviation: LMWH, low-molecular-weight heparin.

MECHANISM OF ACTION

Like heparin, LMWH exerts its anticoagulant activity by activating antithrombin. With a mean molecular weight of 5000, which corresponds to about 17 saccharide units, at least half of the pentasaccharide-containing chains of LMWH are too short to bridge thrombin to antithrombin (Fig. 114-5). However, these chains retain the capacity to accelerate factor Xa inhibition by antithrombin because this activity is largely the result of the conformational changes in antithrombin evoked by pentasaccharide binding. Consequently, LMWH catalyzes factor Xa inhibition by antithrombin more than thrombin inhibition. Depending on their unique molecular weight distributions, LMWH preparations have anti-factor Xa to anti-factor IIa ratios ranging from 2:1 to 4:1.

PHARMACOLOGY

Although usually given SC, LMWH also can be administered IV if a rapid anticoagulant response is needed. LMWH has pharmacokinetic advantages over heparin. These advantages reflect the fact that shorter heparin chains bind less avidly to endothelial cells, macrophages, and heparin-binding plasma proteins. Reduced binding to endothelial cells and macrophages eliminates the rapid, dose-dependent, and saturable mechanism of clearance that is a characteristic of unfractionated heparin. Instead, the clearance of LMWH is dose-independent and its plasma half-life is longer. Based on measurement of anti-factor Xa levels, LMWH has a plasma half-life of ~4 h. LMWH is cleared almost exclusively by the kidneys, and the drug can accumulate in patients with renal insufficiency.

LMWH exhibits about 90% bioavailability after SC injection. Because LMWH binds less avidly to heparin-binding proteins in plasma than heparin, LMWH produces a more predictable dose response, and resistance to LMWH is rare. With a longer half-life and more predictable anticoagulant response, LMWH can be given SC once or twice daily without coagulation monitoring, even when the drug is given in treatment doses. These properties render LMWH more convenient than unfractionated heparin. Capitalizing on this feature, studies in patients with venous thromboembolism have shown that home treatment with LMWH is as effective and safe as in-hospital treatment with continuous IV infusions of heparin. Outpatient treatment with LMWH streamlines care, reduces health care costs, and increases patient satisfaction.

MONITORING

In the majority of patients, LMWH does not require coagulation monitoring. If monitoring is necessary, anti-factor Xa levels must be measured because most LMWH preparations have little effect on the aPTT. Therapeutic anti-factor Xa levels with LMWH range from 0.5 to 1.2 units/mL when measured 3–4 h after drug administration. When LMWH is given in prophylactic doses, peak anti-factor Xa levels of 0.2–0.5 units/mL are desirable.

Indications for LMWH monitoring include renal insufficiency and obesity. LMWH monitoring in patients with a creatinine clearance of ≤50 mL/min is advisable to ensure that there is no drug accumulation. Although weight-adjusted LMWH dosing appears to produce therapeutic anti-factor Xa levels in patients who are overweight, this approach has not been extensively evaluated in those with morbid obesity. It may also be advisable to monitor the anticoagulant activity of LMWH during pregnancy because dose requirements can change, particularly in the third trimester. Monitoring should also be considered in high-risk settings, such as in patients with mechanical heart valves who are given LMWH for prevention of valve thrombosis, and when LMWH is used in treatment doses in infants or children.

DOSING

The doses of LMWH recommended for prophylaxis or treatment vary depending on the LMWH preparation. For prophylaxis, once-daily SC doses of 4000–5000 units are often used, whereas doses of 2500–3000 units are given when the drug is administered twice daily. For treatment of venous thromboembolism, a dose of 150–200 units/kg is given if the drug is administered once daily. If a twice-daily regimen is used, a dose of 100 units/kg is given. In patients with unstable angina, LMWH is given SC on a twice-daily basis at a dose of 100–120 units/kg.

SIDE EFFECTS

The major complication of LMWH is bleeding. Meta-analyses suggest that the risk of major bleeding is lower with LMWH than with unfractionated heparin. HIT and osteoporosis are less common with LMWH than with unfractionated heparin.

Bleeding
Like the situation with heparin, bleeding with LMWH is more common in patients receiving concomitant therapy with antiplatelet or fibrinolytic drugs. Recent surgery, trauma, or underlying hemostatic defects also increase the risk of bleeding with LMWH.

Although protamine sulfate can be used as an antidote for LMWH, protamine sulfate incompletely neutralizes the anticoagulant activity of LMWH because it only binds the longer chains of LMWH. Because longer chains are responsible for catalysis of thrombin inhibition by antithrombin, protamine sulfate completely reverses the anti-factor IIa activity of LMWH. In contrast, protamine sulfate only partially reverses the anti-factor Xa activity of LMWH because the shorter pentasaccharide-containing chains of LMWH do not bind to protamine sulfate. Consequently, patients at high risk for bleeding may be more safely treated with continuous IV unfractionated heparin than with SC LMWH.

Thrombocytopenia
The risk of HIT is about fivefold lower with LMWH than with heparin. LMWH binds less avidly to platelets and causes less PF4 release. Furthermore, with lower affinity for PF4 than heparin, LMWH is less likely to induce the conformational changes in PF4 that trigger the formation of HIT antibodies.

LMWH should not be used to treat HIT patients because most HIT antibodies exhibit cross-reactivity with LMWH. This in vitro cross-reactivity is not simply a laboratory phenomenon because there are case reports of thrombosis when HIT patients were switched from heparin to LMWH.

Osteoporosis

Because the risk of osteoporosis is lower with LMWH than with heparin, LMWH is a better choice for extended treatment.

Fondaparinux

A synthetic analogue of the antithrombin-binding pentasaccharide sequence, fondaparinux differs from LMWH in several ways (**Table 114-6**). Fondaparinux is licensed for thromboprophylaxis in general medical or surgical patients and in high-risk orthopedic patients and as an alternative to heparin or LMWH for initial treatment of patients with established venous thromboembolism. Although fondaparinux is used in Europe as an alternative to heparin or LMWH in patients with acute coronary syndrome, the drug is not licensed for this indication in the United States.

TABLE 114-6

Comparison of LMWH and Fondaparinux

Features	LMWH	Fondaparinux
Number of saccharide units	15–17	5
Catalysis of factor Xa inhibition	Yes	Yes
Catalysis of thrombin inhibition	Yes	No
Bioavailability after subcutaneous administration (%)	90	100
Plasma half-life (h)	4	17
Renal excretion	Yes	Yes
Induces release of tissue factor pathway inhibitor	Yes	No
Neutralized by protamine sulfate	Partially	No

MECHANISM OF ACTION

As a synthetic analogue of the antithrombin-binding pentasaccharide sequence found in heparin and LMWH, fondaparinux has a molecular weight of 1728. Fondaparinux binds only to antithrombin (Fig. 114-5) and is too short to bridge thrombin to antithrombin. Consequently, fondaparinux catalyzes factor Xa inhibition by antithrombin and does not enhance the rate of thrombin inhibition.

PHARMACOLOGY

Fondaparinux exhibits complete bioavailability after SC injection. With no binding to endothelial cells or plasma proteins, the clearance of fondaparinux is dose-independent and its plasma half-life is 17 h. The drug is given SC once daily. Because fondaparinux is cleared unchanged via the kidneys, it is contraindicated in patients with a creatinine clearance <30 mL/min and should be used with caution in those with a creatinine clearance <50 mL/min.

Fondaparinux produces a predictable anticoagulant response after administration in fixed doses because it does not bind to plasma proteins. The drug is given at a dose of 2.5 mg once daily for prevention of venous thromboembolism. For initial treatment of established venous thromboembolism, fondaparinux is given at a dose of 7.5 mg once daily. The dose can be reduced to 5 mg once daily for those weighing <50 kg and increased to 10 mg for those >100 kg. When given in these doses, fondaparinux is as effective as heparin or LMWH for initial treatment of patients with DVT or PE and produces similar rates of bleeding.

Fondaparinux is used at a dose of 2.5 mg once daily in patients with acute coronary syndromes. When this prophylactic dose of fondaparinux was compared with treatment doses of enoxaparin in patients with non-ST-segment elevation acute coronary syndrome, there was no difference in the rate of cardiovascular death, MI, or stroke at 9 days. However, the rate of major bleeding was 50% lower with fondaparinux than with enoxaparin, a difference that likely reflects the fact that the dose of fondaparinux was lower than that of enoxaparin. In acute coronary syndrome patients who require percutaneous coronary intervention, there is a risk of catheter thrombosis with fondaparinux unless adjunctive heparin is given.

SIDE EFFECTS

Fondaparinux does not cause HIT because it does not bind to PF4. In contrast to LMWH, there is no cross-reactivity of fondaparinux with HIT antibodies. Consequently, fondaparinux appears to be effective for treatment of HIT patients, although large clinical trials supporting its use are lacking.

The major side effect of fondaparinux is bleeding. There is no antidote for fondaparinux. Protamine sulfate has no effect on the anticoagulant activity of fondaparinux because it fails to bind to the drug. Recombinant activated factor VII reverses the anticoagulant effects of fondaparinux in volunteers, but it is unknown whether this agent controls fondaparinux-induced bleeding.

Parenteral Direct Thrombin Inhibitors

Direct thrombin inhibitors bind directly to thrombin and block its interaction with its substrates. Approved parenteral direct thrombin inhibitors include recombinant hirudins (lepirudin and desirudin), argatroban, and bivalirudin **(Table 114-7)**. Lepirudin and argatroban are licensed for treatment of patients with HIT, desirudin is licensed for thromboprophylaxis after elective hip arthroplasty, and bivalirudin is approved as an alternative to heparin in patients undergoing percutaneous coronary intervention, including those with HIT.

TABLE 114-7

Comparison of the Properties of Lepirudin, Bivalirudin, and Argatroban

	Lepirudin/Desirudin	Bivalirudin	Argatroban
Molecular mass	7000	1980	527
Site(s) of interaction with thrombin	Active site and exosite 1	Active site and exosite 1	Active site
Renal clearance	Yes	No	No
Hepatic metabolism	No	No	Yes
Plasma half-life (min)	60 (IV) 120–180 (SC)	25	45

LEPIRUDIN AND DESIRUDIN

Recombinant forms of hirudin, lepirudin and desirudin are bivalent direct thrombin inhibitors that interact with the active site and exosite 1, the substrate-binding site on thrombin. For rapid anticoagulation, lepirudin is given by continuous IV infusion, but the drug can be given SC. Lepirudin has a plasma half-life of 60 min after IV infusion and is cleared by the kidneys. Consequently, lepirudin accumulates in patients with renal insufficiency. For thromboprophylaxis, desirudin is given SC twice daily in fixed doses; the half-life of desirudin is 2–3 h after SC injection.

A high proportion of lepirudin-treated patients develop antibodies against the drug; antibody formation is rare with SC desirudin. Although lepirudin-directed antibodies rarely cause problems, in a small subset of patients, they can delay lepirudin clearance and enhance its anticoagulant activity. Serious bleeding has been reported in some of these patients.

Lepirudin is usually monitored using the aPTT, and the dose is adjusted to maintain an aPTT that is 1.5–2.5 times the control. The aPTT is not an ideal test for monitoring lepirudin therapy because the clotting time plateaus with higher drug concentrations. Although the clotting time with ecarin, a snake venom that converts prothrombin to meizothrombin, provides a better index of lepirudin dose than the aPTT, the ecarin clotting time is not widely available. When used for thromboprophylaxis, desirudin does not require monitoring.

ARGATROBAN

A univalent inhibitor that targets the active site of thrombin, argatroban is metabolized in the liver. Consequently, this drug must be used with caution in patients with hepatic insufficiency. Argatroban is not cleared via the kidneys, so this drug is safer than lepirudin for HIT patients with renal insufficiency.

Argatroban is administered by continuous IV infusion and has a plasma half-life of ~45 min. The aPTT is used to monitor its anticoagulant effect, and the dose is adjusted to achieve an aPTT 1.5–3 times the baseline value, but not to exceed 100 s. Argatroban also prolongs the international normalized ratio (INR), a feature that can complicate the transitioning of patients to warfarin. This problem can be circumvented by using the levels of factor X to monitor warfarin in place of the INR. Alternatively, argatroban can be stopped for 2–3 h before INR determination.

BIVALIRUDIN

A synthetic 20-amino-acid analogue of hirudin, bivalirudin is a divalent thrombin inhibitor. Thus, the N-terminus of bivalirudin interacts with the active site of thrombin, whereas its C-terminus binds to exosite 1. Bivalirudin has a plasma half-life of 25 min, the shortest half-life of all the parenteral direct thrombin inhibitors. Bivalirudin is degraded by peptidases and is partially excreted via the kidneys. When given in high doses in the cardiac catheterization laboratory, the anticoagulant activity of bivalirudin is monitored using the activated clotting time. With lower doses, its activity can be assessed using the aPTT.

Bivalirudin is licensed as an alternative to heparin in patients undergoing percutaneous coronary intervention. Bivalirudin also has been used successfully in HIT patients who require percutaneous coronary intervention or cardiac bypass surgery.

ORAL ANTICOAGULANTS

For many years, vitamin K antagonists such as warfarin were the only available oral anticoagulants. This situation changed with the introduction of the direct oral anticoagulants, which include dabigatran, rivaroxaban, apixaban, and edoxaban.

Warfarin

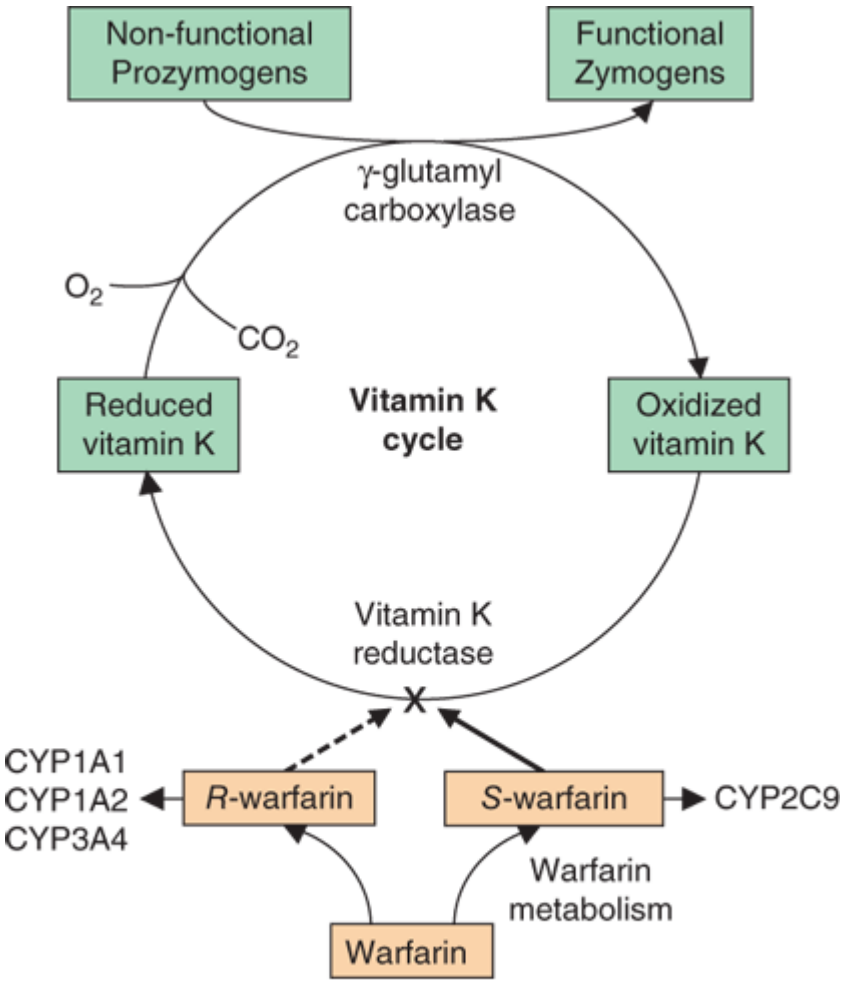
A water-soluble vitamin K antagonist initially developed as a rodenticide, warfarin is the coumarin derivative most often prescribed in North America. Like other vitamin K antagonists, warfarin interferes with the synthesis of the vitamin K–dependent clotting proteins, which include prothrombin (factor II) and factors VII, IX, and X. The synthesis of the vitamin K–dependent anticoagulant proteins, proteins C and S, is also reduced by vitamin K antagonists.

MECHANISM OF ACTION

All of the vitamin K–dependent clotting factors possess glutamic acid residues at their N termini. A posttranslational modification adds a carboxyl group to the γ-carbon of these residues to generate γ-carboxyglutamic acid. This modification is essential for expression of the activity of these clotting factors because it permits their calcium-dependent binding to negatively charged phospholipid surfaces. The γ-carboxylation process is catalyzed by a vitamin K–dependent carboxylase. Thus, vitamin K from the diet is reduced to vitamin K hydroquinone by vitamin K reductase (**Fig. 114-6**). Vitamin K hydroquinone serves as a cofactor for the carboxylase enzyme, which in the presence of carbon dioxide replaces the hydrogen on the γ-carbon of glutamic acid residues with a carboxyl group. During this process, vitamin K hydroquinone is oxidized to vitamin K epoxide, which is then reduced to vitamin K by vitamin K epoxide reductase.

FIGURE 114-6

Mechanism of action of warfarin. A racemic mixture of *S*- and *R*-enantiomers, *S*-warfarin is most active. By blocking vitamin K epoxide reductase, warfarin inhibits the conversion of oxidized vitamin K into its reduced form. This inhibits vitamin K–dependent γ-carboxylation of factors II, VII, IX, and X because reduced vitamin K serves as a cofactor for γ-glutamyl carboxylase, which catalyzes the γ-carboxylation process, thereby converting prozymogens to zymogens capable of binding calcium and interacting with anionic phospholipid surfaces. *S*-warfarin is metabolized by CYP2C9. Common genetic polymorphisms in this enzyme can influence warfarin metabolism. Polymorphisms in the C1 subunit of vitamin K epoxide reductase (VKORC1) also can affect the susceptibility of the enzyme to warfarin-induced inhibition, thereby influencing warfarin dosage requirements.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Warfarin inhibits vitamin K epoxide reductase (VKOR), thereby blocking the γ-carboxylation process. This results in the synthesis of vitamin K–dependent clotting proteins that are only partially γ-carboxylated. Warfarin acts as an anticoagulant because these partially γ-carboxylated proteins have reduced or absent biologic activity. The onset of action of warfarin is delayed until the newly synthesized clotting factors with reduced activity gradually replace their fully active counterparts.

The antithrombotic effect of warfarin depends on a reduction in the functional levels of factor X and prothrombin, clotting factors that have half-lives of 24 and 72 h, respectively. Because the antithrombotic

effect of warfarin is delayed, patients with established thrombosis or at high risk for thrombosis require concomitant treatment with a rapidly acting parenteral anticoagulant, such as heparin, LMWH, or fondaparinux, for at least 5 days.

PHARMACOLOGY

Warfarin is a racemic mixture of R and S isomers. Warfarin is rapidly and almost completely absorbed from the gastrointestinal tract. Levels of warfarin in the blood peak about 90 min after drug administration. Racemic warfarin has a plasma half-life of 36–42 h, and >97% of circulating warfarin is bound to albumin. Only the small fraction of unbound warfarin is biologically active.

Warfarin accumulates in the liver where the two isomers are metabolized via distinct pathways. *CYP2C9* mediates oxidative metabolism of the more active S isomer (**Fig. 114-6**). Two relatively common variants, *CYP2C9** *2* and *CYP2C9** *3*, encode an enzyme with reduced activity. Patients with these variants require lower maintenance doses of warfarin. Approximately 25% of Caucasians have at least one variant allele of *CYP2C9** *2* or *CYP2C9** *3*, whereas those variant alleles are less common in African Americans and Asians (**Table 114-8**). Heterozygosity for *CYP2C9** *2* or *CYP2C9** *3* decreases the warfarin dose requirement by 20–30% relative to that required in subjects with the wild-type *CYP2C9** *1*/** *1* alleles, whereas homozygosity for the *CYP2C9** *2* or *CYP2C9** *3* alleles reduces the warfarin dose requirement by 50–70%.

TABLE 114-8

Frequencies of *CYP2C9* Genotypes and *VKORC1* Haplotypes in Different Populations and Their Effect on Warfarin Dose Requirements

	Frequency, %			
Genotype/Haplotype	**Caucasians**	**African Americans (A/A)**	**Asians (A)**	**Dose Reduction Compared with Wild-Type**
CYP2C9				
*1/*1	70	90	95	–
*1/*2	17	2	0	22
*1/*3	9	3	4	34
*2/*2	2	0	0	43
*2/*3	1	0	0	53
*3/*3	0	0	1	76
VKORC1				
Non-A/non-A	37	82	7	–
Non-A/A	45	12	30	26
A/A	18	6	63	50

Consistent with their decreased warfarin dose requirement, subjects with at least one *CYP2C9* variant allele are at increased risk for bleeding. Compared with individuals with no variant alleles, the relative risks for warfarin-associated bleeding in *CYP2C9** 2* or *CYP2C9** 3* carriers are 1.9 and 1.8, respectively.

Polymorphisms in *VKORC1* also can influence the anticoagulant response to warfarin. Several genetic variations of *VKORC1* are in strong linkage disequilibrium and have been designated as non-A haplotypes. *VKORC1* variants are more prevalent than variants of *CYP2C9*. Asians have the highest prevalence of *VKORC1* variants, followed by Caucasians and African Americans (Table 114-8). Polymorphisms in *VKORC1* likely explain 30% of the variability in warfarin dose requirements. Compared with *VKORC1* non-A/non-A

homozygotes, the warfarin dose requirement decreases by 25 and 50% in A haplotype heterozygotes and homozygotes, respectively. These findings prompted the Food and Drug Administration to amend the prescribing information for warfarin to indicate that lower initiation doses should be considered for patients with *CYP2C9* and *VKORC1* genetic variants. In addition to genotype data, other pertinent patient information has been incorporated into warfarin dosing algorithms. Although such algorithms help predict suitable warfarin doses, it remains unclear whether better dose identification improves patient outcome in terms of reducing hemorrhagic complications or recurrent thrombotic events.

In addition to genetic factors, the anticoagulant effect of warfarin is influenced by diet, drugs, and various disease states. Fluctuations in dietary vitamin K intake affect the activity of warfarin. A wide variety of drugs can alter absorption, clearance, or metabolism of warfarin. Because of the variability in the anticoagulant response to warfarin, coagulation monitoring is essential to ensure that a therapeutic response is obtained.

MONITORING

Warfarin therapy is most often monitored using the prothrombin time, a test that is sensitive to reductions in the levels of prothrombin, factor VII, and factor X. The test is performed by adding thromboplastin, a reagent that contains tissue factor, phospholipid, and calcium, to citrated plasma and determining the time to clot formation. Thromboplastins vary in their sensitivity to reductions in the levels of the vitamin K–dependent clotting factors. Thus, less sensitive thromboplastins will trigger the administration of higher doses of warfarin to achieve a target prothrombin time. This is problematic because higher doses of warfarin increase the risk of bleeding.

The INR was developed to circumvent many of the problems associated with the prothrombin time. To calculate the INR, the patient's prothrombin time is divided by the mean normal prothrombin time, and this ratio is then multiplied by the international sensitivity index (ISI), which is an index of the sensitivity of the thromboplastin used for prothrombin time determination to reductions in the levels of the vitamin K–dependent clotting factors. Highly sensitive thromboplastins have an ISI of 1.0. Most current thromboplastins have ISI values that range from 0.9 to 1.4.

Although the INR has helped to standardize anticoagulant practice, problems persist. The precision of INR determination varies depending on reagent-coagulometer combinations. This leads to variability in the INR results. Also complicating INR determination is unreliable reporting of the ISI by thromboplastin manufacturers. Furthermore, every laboratory must establish the mean normal prothrombin time with each new batch of thromboplastin reagent. To accomplish this, the prothrombin time must be measured in fresh plasma samples from at least 20 healthy volunteers using the same coagulometer that is used for patient samples.

For most indications, warfarin is administered in doses that produce a target INR of 2.0–3.0. An exception is patients with mechanical heart valves, particularly those in the mitral position or older ball and cage valves in the aortic position, where a target INR of 2.5–3.5 is recommended. Studies in atrial fibrillation demonstrate an increased risk of cardioembolic stroke when the INR falls to <1.7 and an increase in intracranial bleeding with INR values >4.5. These findings highlight the fact that vitamin K antagonists have a narrow therapeutic

window. In support of this concept, a study in patients receiving long-term warfarin therapy for unprovoked venous thromboembolism demonstrated a higher rate of recurrent venous thromboembolism with a target INR of 1.5–1.9 compared with a target INR of 2.0–3.0.

DOSING

Warfarin is usually started at a dose of 5–10 mg. Lower doses are used for patients with *CYP2C9* or *VKORC1* polymorphisms, which affect the pharmacodynamics or pharmacokinetics of warfarin and render patients more sensitive to the drug. The dose is then titrated to achieve the desired target INR. Because of its delayed onset of action, patients with established thrombosis or those at high risk for thrombosis are given concomitant initial treatment with a rapidly acting parenteral anticoagulant, such as heparin, LMWH, or fondaparinux. Early prolongation of the INR reflects reduction in the functional levels of factor VII. Consequently, concomitant treatment with the parenteral anticoagulant should be continued until the INR has been therapeutic for at least 2 consecutive days. A minimum 5-day course of parenteral anticoagulation is recommended to ensure that the levels of factor Xa and prothrombin have been reduced into the therapeutic range with warfarin.

Because warfarin has a narrow therapeutic window, frequent coagulation monitoring is essential to ensure that a therapeutic anticoagulant response is maintained. Even patients with stable warfarin dose requirements should have their INR determined every 3–4 weeks. More frequent monitoring is necessary when new medications are introduced because so many drugs enhance or reduce the anticoagulant effects of warfarin.

SIDE EFFECTS

Like all anticoagulants, the major side effect of warfarin is bleeding. A rare complication is skin necrosis. Warfarin crosses the placenta and can cause fetal abnormalities. Consequently, warfarin should not be used during pregnancy.

Bleeding

At least half of the bleeding complications with warfarin occur when the INR exceeds the therapeutic range. Bleeding complications may be mild, such as epistaxis or hematuria, or more severe, such as retroperitoneal or gastrointestinal bleeding. Life-threatening intracranial bleeding can also occur.

To minimize the risk of bleeding, the INR should be maintained in the therapeutic range. In asymptomatic patients whose INR is between 3.5 and 10, warfarin should be withheld until the INR returns to the therapeutic range. If the INR is over 10, oral vitamin K should be administered, at a dose of 2.5–5 mg, although there is no evidence that doing so reduces the bleeding risk. Higher doses of oral vitamin K (5–10 mg) produce more rapid reversal of the INR but may render patients temporarily resistant to warfarin when the drug is restarted.

Patients with serious bleeding need more aggressive treatment. These patients should be given 5–10 mg of vitamin K by slow IV infusion. Additional vitamin K should be given until the INR is in the normal range. Treatment with vitamin K should be supplemented with four factor prothrombin complex concentrate, which

contains all four vitamin K–dependent clotting proteins. Prothrombin complex concentrate normalizes the INR more rapidly than transfusion of fresh frozen plasma.

Warfarin-treated patients who experience bleeding when their INR is in the therapeutic range require investigation into the cause of the bleeding. Those with gastrointestinal or genitourinary bleeding often have an underlying lesion.

Skin Necrosis

A rare complication of warfarin, skin necrosis usually is seen 2–5 days after initiation of therapy. Well-demarcated erythematous lesions form on the thighs, buttocks, breasts, or toes. Typically, the center of the lesion becomes progressively necrotic. Examination of skin biopsies taken from the border of these lesions reveals thrombi in the microvasculature.

Warfarin-induced skin necrosis is seen in patients with congenital or acquired deficiencies of protein C or protein S. Initiation of warfarin therapy in these patients produces a precipitous fall in plasma levels of proteins C or S, thereby eliminating this important anticoagulant pathway before warfarin exerts an antithrombotic effect through lowering of the functional levels of factor X and prothrombin. The resultant procoagulant state triggers thrombosis. Why the thrombosis is localized to the microvasculature of fatty tissues is unclear.

Treatment involves discontinuation of warfarin and reversal with vitamin K, if needed. An alternative anticoagulant, such as heparin or LMWH, should be given in patients with thrombosis. Protein C concentrate can be given to protein C–deficient patients to accelerate healing of the skin lesions; fresh-frozen plasma may be of value if protein C concentrate is unavailable and for those with protein S deficiency. Occasionally, skin grafting is necessary when there is extensive skin loss.

Because of the potential for skin necrosis, patients with known protein C or protein S deficiency require overlapping treatment with a parenteral anticoagulant when initiating warfarin therapy. Warfarin should be started in low doses in these patients, and the parenteral anticoagulant should be continued until the INR is therapeutic for at least 2–3 consecutive days. Alternatively, treatment with rivaroxaban or apixaban could be given although there is limited information about their efficacy and safety in patients with severe protein C or S deficiency.

Pregnancy

Warfarin crosses the placenta and can cause fetal abnormalities or bleeding. The fetal abnormalities include a characteristic embryopathy, which consists of nasal hypoplasia and stippled epiphyses. The risk of embryopathy is highest if warfarin is given in the first trimester of pregnancy. Central nervous system abnormalities can also occur with exposure to warfarin at any time during pregnancy. Finally, maternal administration of warfarin produces an anticoagulant effect in the fetus that can cause bleeding. This is of particular concern at delivery when trauma to the head during passage through the birth canal can lead to intracranial bleeding. Because of these potential problems, warfarin is contraindicated in pregnancy, particularly in the first and third trimesters. Instead, heparin, LMWH, or fondaparinux can be given during pregnancy for prevention or treatment of thrombosis.

Warfarin does not pass into the breast milk. Consequently, warfarin can safely be given to nursing mothers.

Special Problems

Patients with a lupus anticoagulant and those who need urgent or elective surgery present special challenges. Although observational studies suggested that patients with thrombosis complicating the antiphospholipid syndrome required higher intensity warfarin regimens to prevent recurrent thromboembolic events, two randomized trials showed that targeting an INR of 2.0–3.0 is as effective as higher intensity treatment and produces less bleeding. Monitoring warfarin therapy can be problematic in patients with antiphospholipid syndrome if the lupus anticoagulant prolongs the baseline INR; chromogenic factor X levels can be used instead of the INR in such patients.

There is no need to stop warfarin before procedures associated with a low risk of bleeding; these include dental cleaning, simple dental extraction, cataract surgery, or skin biopsy. For procedures associated with a moderate or high risk of bleeding, warfarin should be stopped 5 days before the procedure to allow the INR to return to normal levels. Patients at high risk for thrombosis, such as those with mechanical heart valves, can be bridged with once- or twice-daily SC injections of LMWH when the INR falls to <2.0. The last dose of LMWH should be given 12–24 h before the procedure, depending on whether LMWH is administered twice or once daily. After the procedure, treatment with warfarin can be restarted.

Direct Oral Anticoagulants

Direct oral anticoagulants are available as alternatives to warfarin. These agents include dabigatran, which inhibits thrombin, and rivaroxaban, apixaban, and edoxaban, which inhibit factor Xa. All of these drugs have a rapid onset and offset of action and have half-lives that permit once- or twice-daily administration. Designed to produce a predictable level of anticoagulation, the direct oral agents are more convenient to administer than warfarin because they are given in fixed doses without routine coagulation monitoring.

MECHANISM OF ACTION

The direct oral anticoagulants are small molecules that bind reversibly to the active site of their target enzyme. Table 114-9 summarizes the distinct pharmacological properties of these agents.

TABLE 114-9

Comparison of the Pharmacologic Properties of the New Oral Anticoagulants

Characteristic	Rivaroxaban	Apixaban	Edoxaban	Dabigatran
Target	Factor Xa	Factor Xa	Factor Xa	Thrombin
Prodrug	No	No	No	Yes
Bioavailability	80%	60%	50%	6%
Dosing	qd (bid)	bid	qd	bid (qd)
Half-life	7–11 h	12 h	9–11 h	12–17 h
Renal excretion	33% (66%)	25%	35%	80%
Interactions	CYP 3A4/P-gp	CYP 3A4/P-gp	P-gp	P-gp

Abbreviations: bid, twice a day; CYP, cytochrome P450; P-gp, P-glycoprotein; qd, once a day.

INDICATIONS

The direct oral anticoagulants have been compared with warfarin for stroke prevention in patients with nonvalvular atrial fibrillation in four randomized trials that enrolled 71,683 patients. A meta-analysis of these data demonstrates that compared with warfarin, the higher doses of the direct oral anticoagulants significantly reduce stroke or systemic embolism by 19% ($p = 0.001$), primarily driven by a 51% reduction in hemorrhagic stroke ($p < 0.0001$), and are associated with a 10% reduction in mortality ($p < 0.0001$). The direct oral anticoagulants reduce intracranial hemorrhage by 52% compared with warfarin ($p < 0.0001$), but increase gastrointestinal bleeding by about 24% ($p = 0.04$). Overall, the direct oral anticoagulants demonstrate a favorable benefit-to-risk profile compared with warfarin, and their relative efficacy and safety are maintained across a wide spectrum of atrial fibrillation patients, including those over the age of 75 years and those with a prior history of stroke. Based on these findings, dabigatran, rivaroxaban, apixaban, and edoxaban are licensed as alternatives to warfarin for stroke prevention in nonvalvular atrial fibrillation, which is defined as atrial fibrillation occurring in patients without mechanical heart valves or severe rheumatic valvular disease, particularly mitral stenosis and/or regurgitation.

The direct oral anticoagulants were compared with conventional anticoagulation therapy in 27,023 patients with acute venous thromboembolism. The primary efficacy endpoint in these trials was recurrent venous thromboembolism, while the primary safety outcome was either major bleeding or the composite of major

and clinically relevant non-major bleeding. In a pooled analysis of these trials, recurrent fatal and non-fatal venous thromboembolism occurred in 2.0% of those given direct oral anticoagulants compared with 2.2% of those given a vitamin K antagonist (relative risk [RR] 0.90, 95% confidence interval [CI]: 0.77–1.06). Compared with vitamin K antagonists, direct oral anticoagulants were associated with a 39% reduction in the risk of major bleeding (RR 0.61, 95% CI: 0.45–0.83), a 63% reduction in intracranial bleeding (RR 0.37, 95% CI: 0.21–0.68), and a 64% reduction in fatal bleeding (RR 0.36, 95% CI: 0.15–0.84). In addition, clinically relevant non-major bleeding was reduced by 27% with the direct oral anticoagulants compared with vitamin K antagonists (RR 0.73, 95% CI: 0.58–0.93). Therefore, the direct oral anticoagulants are non-inferior to well-managed vitamin K antagonist therapy for treatment of venous thromboembolism, but are associated with significantly less bleeding.

Whereas dabigatran and edoxaban were started after a minimum of a 5-day course of parenteral anticoagulant therapy, rivaroxaban and apixaban were administered in all-oral regimens starting with a higher dose for 21 days and 7 days, respectively. When used in this all-oral fashion, both agents were non-inferior to conventional therapy and were associated with significantly less major bleeding. Therefore, rivaroxaban and apixaban simplify treatment and facilitate out-of-hospital management of most patients with DVT and many with PE, thereby reducing healthcare costs. With these advantages, clinical guidelines now endorse the direct oral anticoagulants for first-line treatment of venous thromboembolism in patients without active cancer. For those with active cancer, LMWH remains the preferred therapy.

Dabigatran, rivaroxaban, and apixaban have also been compared with enoxaparin for thromboprophylaxis after elective hip or knee arthroplasty and although all three are licensed for this indication, dabigatran is rarely used.

DOSING

For stroke prevention in patients with nonvalvular atrial fibrillation, dabigatran is given at a dose of 150 mg twice daily with a dose reduction to 75 mg twice daily in those with a creatinine clearance of 15–30 mL/min; rivaroxaban is given at a dose of 20 mg once daily with a dose reduction to 15 mg once daily in patients with a creatinine clearance of 15–49 mL/min; apixaban is given at a dose of 5 mg twice daily with a dose reduction to 2.5 mg twice daily for patients with at least two of the following criteria: age of 80 years or more, body weight of 60 kg or less, and serum creatinine >1.5 g/dL; and edoxaban is given at a dose of 60 mg once daily with a dose reduction to 30 mg once daily for patients with a creatinine clearance of 15–50 mL/min, body weight of 60 kg or less, or receiving potent P-glycoprotein inhibitors such as dronedarone or verapamil.

For treatment of venous thromboembolism, dabigatran is given at a dose of 150 mg twice daily after a minimum of a 5-day course of heparin or LMWH; rivaroxaban is started at a dose of 15 mg twice daily for 21 days, and the dose is then reduced to 20 mg once daily thereafter; apixaban is started at a dose of 10 mg twice daily for 7 days, and the dose is then reduced to 5 mg twice daily for the following 6 months after which the dose can be further reduced to 2.5 mg twice daily; and edoxaban is given at a dose of 60 mg once daily after a minimum of a 5-day course of heparin or LMWM with the dose reduced to 30 mg once daily for patients with a creatinine clearance of 15–50 mL/min, body weight of 60 kg or less, or receiving potent P-glycoprotein inhibitors.

For thromboprophylaxis after elective hip or knee replacement surgery, rivaroxaban is given at a dose of 10 mg once daily, whereas apixaban is given at a dose of 2.5 mg twice daily. The drugs are usually given for 14 days after knee arthroplasty and for 35 days after hip arthroplasty.

MONITORING

Although designed to be administered without routine monitoring, there are situations where determination of the anticoagulant activity of the new oral anticoagulants can be helpful. These include assessment of adherence, detection of accumulation or overdose, identification of bleeding mechanisms, and determination of activity prior to surgery or intervention. For qualitative assessment of anticoagulant activity, the prothrombin time can be used for factor Xa inhibitors and the aPTT for dabigatran. Rivaroxaban and edoxaban prolong the prothrombin time more than apixaban. In fact, because apixaban has such a limited effect on the prothrombin time, anti-factor Xa assays are needed to assess its activity. The effect of the drugs on tests of coagulation varies depending on the time that the blood is drawn relative to the timing of the last dose of the drug and the reagents used to perform the tests. Chromogenic anti-factor Xa assays and a diluted thrombin clotting time with appropriate calibrators provide quantitative assays to measure the plasma levels of the factor Xa inhibitors and dabigatran, respectively.

SIDE EFFECTS

Like all anticoagulants, bleeding is the most common side effect of the direct oral anticoagulants. The direct oral anticoagulants are associated with less intracranial bleeding than warfarin. The increased risk of intracranial bleeding with warfarin likely reflects the reduction in functional levels of factor VII, which precludes efficient thrombin generation at sites of microvascular bleeding in the brain. Because the direct oral anticoagulants target downstream coagulation enzymes, they produce less impairment of hemostatic plug formation at sites of vascular injury.

A downside of the new oral anticoagulants is the increased risk of gastrointestinal bleeding. This likely occurs because unabsorbed active drug in the gut exacerbates bleeding from lesions. Although dabigatran etexilate is a prodrug, only 7% is absorbed and the remainder passes through the gut where at least two-thirds is metabolically activated to dabigatran by gut esterases.

Dyspepsia occurs in up to 10% of patients treated with dabigatran; this problem improves with time and can be minimized by administering the drug with food. Dyspepsia is rare with rivaroxaban, apixaban, and edoxaban.

PERIPROCEDURAL MANAGEMENT

Like warfarin, the new oral anticoagulants must be stopped before procedures associated with a moderate or high risk of bleeding. The drugs should be held for 1–2 days, or longer if renal function is impaired. Assessment of residual anticoagulant activity before procedures associated with a high bleeding risk is prudent.

MANAGEMENT OF BLEEDING

With minor bleeding, holding one or two doses of drug is usually sufficient. The approach to serious bleeding is similar to that with warfarin except that vitamin K administration is of no benefit. Thus, the anticoagulant and antiplatelet drugs should be held, the patient should be resuscitated with fluids and blood products as

necessary, and, if possible, the bleeding site should be identified and managed. Coagulation testing will determine the extent of anticoagulation and renal function should be assessed so that the half-life of the drug can be calculated. Timing of the last dose of anticoagulant is important; administration of oral activated charcoal may help to prevent absorption of drug administered in the past 2–4 h.

Anticoagulant reversal should be considered in patients with life-threatening bleeding, such as intracranial bleeding, if bleeding continues despite supportive measures or if patients require urgent surgery. Idarucizumab is a specific reversal agent for dabigatran. A monoclonal antibody fragment, idarucizumab binds dabigatran with high affinity to form a 1:1 complex that is then cleared by the kidneys. Idarucizumab is given as an intravenous bolus of 5 g.

Andexanet alfa and ciraparantag are under development for reversal of rivaroxaban, apixaban, and edoxaban but neither is approved. Until these reversal agents are available, prothrombin complex concentrate should be considered for reversal of the oral factor Xa inhibitors in patients with life-threatening or ongoing bleeding.

PREGNANCY

As small molecules, the direct oral anticoagulants can all pass through the placenta. Consequently, these agents are contraindicated in pregnancy, and when used by women of childbearing potential, appropriate contraception is important. The direct oral anticoagulants should be avoided in nursing mothers and their safety in children has yet to be established.

FIBRINOLYTIC DRUGS

ROLE OF FIBRINOLYTIC THERAPY

Fibrinolytic drugs can be used to degrade thrombi and are administered systemically or can be delivered via catheters directly into the substance of the thrombus. Systemic delivery is used for treatment of acute MI, acute ischemic stroke, and most cases of massive PE. The goal of therapy is to produce rapid thrombus dissolution, thereby restoring blood flow. In the coronary circulation, restoration of blood flow reduces morbidity and mortality rates by limiting myocardial damage, whereas in the cerebral circulation, rapid thrombus dissolution decreases the neuronal death and brain infarction that produce irreversible brain injury. For patients with massive PE, the goal of thrombolytic therapy is to restore pulmonary artery perfusion.

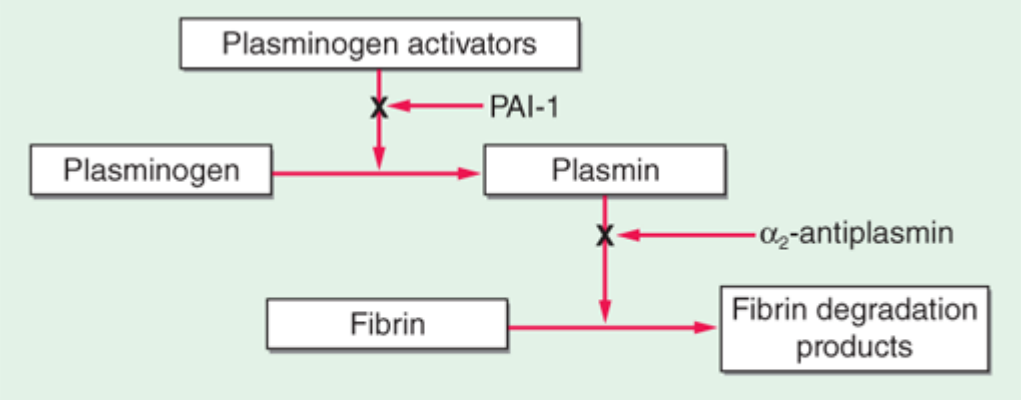
Peripheral arterial thrombi and thrombi in the proximal deep veins of the leg are most often treated using catheter-directed thrombolytic therapy. Catheters with multiple side holes can be used to enhance drug delivery. In some cases, intravascular devices that fragment and extract the thrombus are used to hasten treatment. These devices can be used alone or in conjunction with fibrinolytic drugs.

MECHANISM OF ACTION

Currently approved fibrinolytic agents include streptokinase; acylated plasminogen streptokinase activator complex (anistreplase); urokinase; recombinant tissue-type plasminogen activator (rtPA), which is also known as alteplase or activase; and two recombinant derivatives of rtPA, tenecteplase and reteplase. All of these agents act by converting plasminogen, the zymogen, to plasmin, the active enzyme **(Fig. 114-7)**. Plasmin then degrades the fibrin matrix of thrombi and produces soluble fibrin degradation products.

FIGURE 114-7

The fibrinolytic system and its regulation. Plasminogen activators convert plasminogen to plasmin. Plasmin then degrades fibrin into soluble fibrin degradation products. The system is regulated at two levels. Type 1 plasminogen activator inhibitor (PAI-1) regulates the plasminogen activators, whereas α_2-antiplasmin serves as the major inhibitor of plasmin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Endogenous fibrinolysis is regulated at two levels. Plasminogen activator inhibitors, particularly the type 1 form (PAI-1), prevent excessive plasminogen activation by regulating the activity of tPA and urokinase-type plasminogen activator (uPA). Once plasmin is generated, it is regulated by plasmin inhibitors, the most important of which is α_2-antiplasmin. The plasma concentration of plasminogen is twofold higher than that of α_2-antiplasmin. Consequently, with pharmacologic doses of plasminogen activators, the concentration of plasmin that is generated can exceed that of α_2-antiplasmin. In addition to degrading fibrin, unregulated plasmin can also degrade fibrinogen and other clotting factors. This process, which is known as the *systemic lytic state*, reduces the hemostatic potential of the blood and increases the risk of bleeding.

The endogenous fibrinolytic system is geared to localize plasmin generation to the fibrin surface. Both plasminogen and tPA bind to fibrin to form a ternary complex that promotes efficient plasminogen activation. In contrast to free plasmin, plasmin generated on the fibrin surface is relatively protected from inactivation by α_2-antiplasmin, a feature that promotes fibrin dissolution. Furthermore, C-terminal lysine residues, exposed as plasmin degrades fibrin, serve as binding sites for additional plasminogen and tPA molecules. This creates a positive feedback that enhances plasmin generation. When used pharmacologically, the various plasminogen activators capitalize on these mechanisms to a lesser or greater extent.

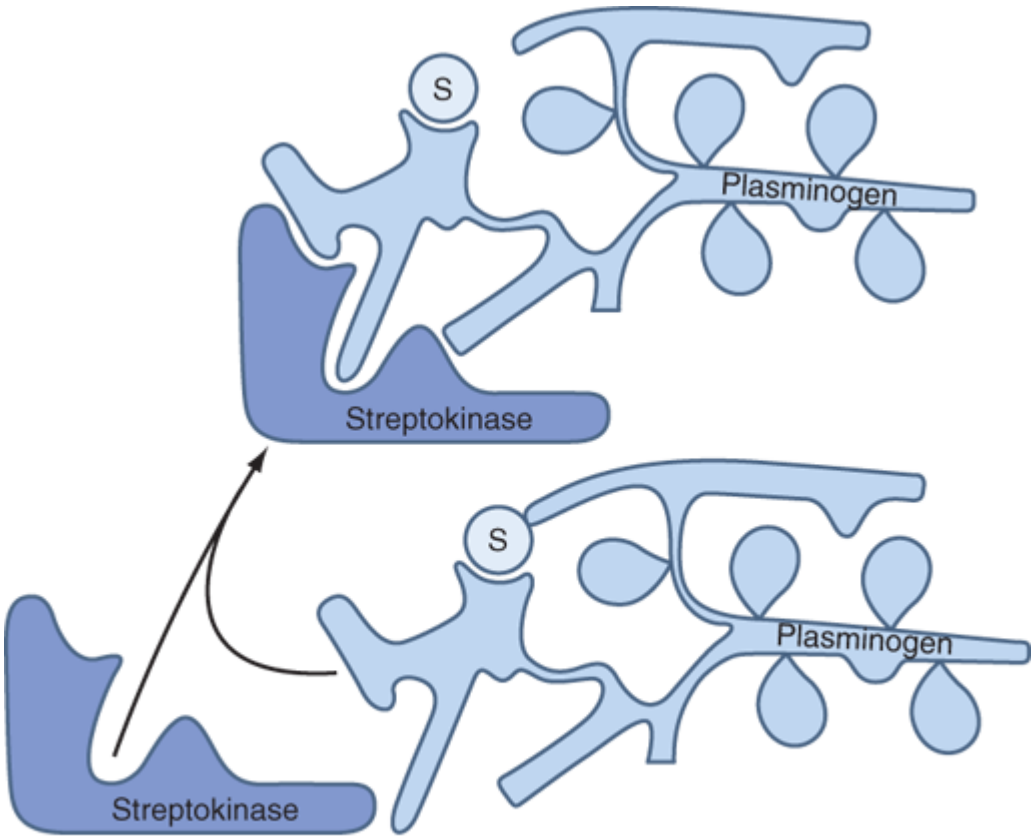
Plasminogen activators that preferentially activate fibrin-bound plasminogen are considered fibrin-specific. In contrast, nonspecific plasminogen activators do not discriminate between fibrin-bound and circulating plasminogen. Activation of circulating plasminogen results in the generation of unopposed plasmin that can trigger the systemic lytic state. Alteplase and its derivatives are fibrin-specific plasminogen activators, whereas streptokinase, anistreplase, and urokinase are nonspecific agents.

STREPTOKINASE

Unlike other plasminogen activators, streptokinase is not an enzyme and does not directly convert plasminogen to plasmin. Instead, streptokinase forms a 1:1 stoichiometric complex with plasminogen. Formation of this complex induces a conformational change in plasminogen that exposes its active site (**Fig. 114-8**). The streptokinase-plasminogen complex then converts additional plasminogen to plasmin.

FIGURE 114-8

Mechanism of action of streptokinase. Streptokinase binds to plasminogen and induces a conformational change in plasminogen that exposes its active site. The streptokinase/plasmin(ogen) complex then serves as the activator of additional plasminogen.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Streptokinase has no affinity for fibrin, and the streptokinase-plasminogen complex activates both free and fibrin-bound plasminogen. Activation of circulating plasminogen generates sufficient amounts of plasmin to overwhelm α_2-antiplasmin. Unopposed plasmin not only degrades fibrin in the occlusive thrombus but also induces a systemic lytic state.

When given systemically to patients with acute MI, streptokinase reduces mortality. For this indication, the drug is usually given as an IV infusion of 1.5 million units over 30–60 min. Patients who receive streptokinase can develop antibodies against the drug, as can patients with prior streptococcal infection. These antibodies can reduce the effectiveness of streptokinase.

Allergic reactions occur in ~5% of patients treated with streptokinase. These may manifest as a rash, fever, chills, and rigors. Although anaphylactic reactions can occur, these are rare. Transient hypotension is common with streptokinase and has been attributed to plasmin-mediated release of bradykinin from kininogen. The hypotension usually responds to leg elevation and administration of IV fluids and low doses of vasopressors, such as dopamine or norepinephrine.

ANISTREPLASE

To generate this drug, streptokinase is combined with equimolar amounts of Lys-plasminogen, a plasmin-cleaved form of plasminogen with a Lys residue at its N terminal. The active site of Lys-plasminogen that is exposed upon combination with streptokinase is then masked with an anisoyl group. After IV infusion, the anisoyl group is slowly removed by deacylation, giving the complex a half-life of ~100 min. This allows drug administration via a single bolus infusion.

Although it is more convenient to administer, anistreplase offers few mechanistic advantages over streptokinase. Like streptokinase, anistreplase does not distinguish between fibrin-bound and circulating plasminogen. Consequently, it too produces a systemic lytic state. Likewise, allergic reactions and hypotension are just as frequent with anistreplase as they are with streptokinase.

When anistreplase was compared with alteplase in patients with acute MI, reperfusion was obtained more rapidly with alteplase than with anistreplase. Improved reperfusion was associated with a trend toward better clinical outcomes and reduced mortality rate with alteplase. These results and the high cost of anistreplase have dampened the enthusiasm for its use.

UROKINASE

Urokinase is a two-chain serine protease derived from cultured fetal kidney cells with a molecular weight of 34,000. Urokinase converts plasminogen to plasmin directly by cleaving the Arg560-Val561 bond. Unlike streptokinase, urokinase is not immunogenic and allergic reactions are rare. Urokinase produces a systemic lytic state because it does not discriminate between fibrin-bound and circulating plasminogen.

Despite many years of use, urokinase has never been systemically evaluated for coronary thrombolysis. Instead, urokinase is often employed for catheter-directed lysis of thrombi in the deep veins or the peripheral arteries. Because of production problems, the availability of urokinase is limited.

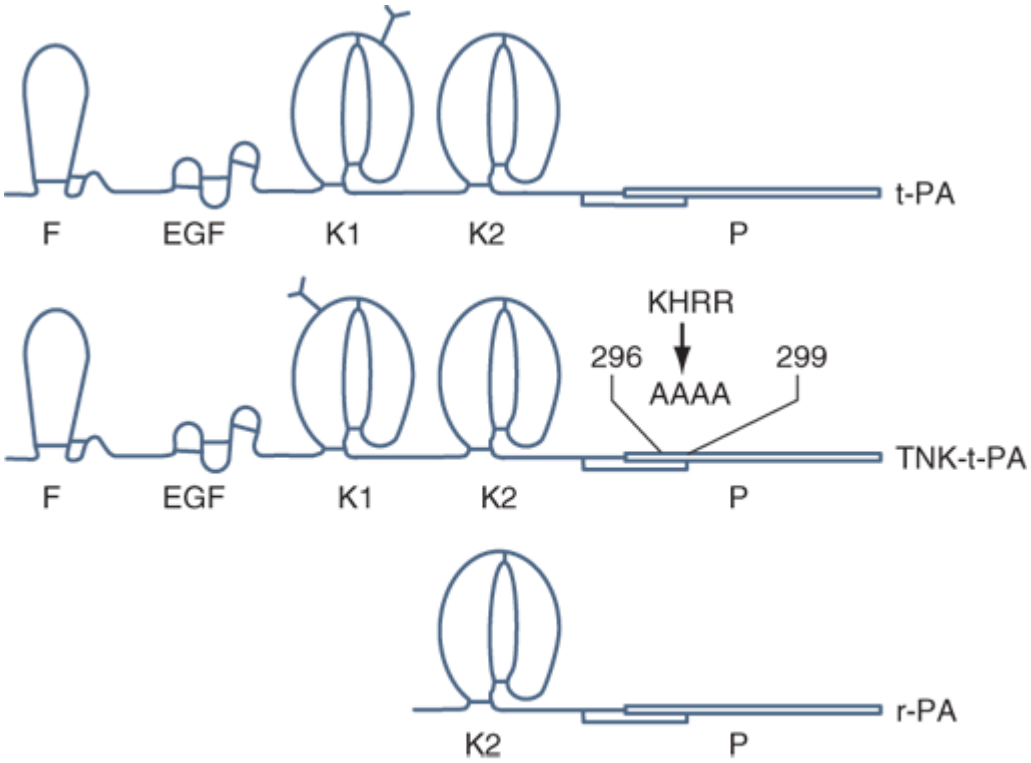
ALTEPLASE

A recombinant form of single-chain tPA, alteplase has a molecular weight of 68,000. Alteplase is rapidly converted into its two-chain form by plasmin. Although single- and two-chain forms of tPA have equivalent activity in the presence of fibrin, in its absence, single-chain tPA has tenfold lower activity.

Alteplase consists of five discrete domains **(Fig. 114-9)**; the N-terminal A chain of two-chain alteplase contains four of these domains. Residues 4 through 50 make up the finger domain, a region that resembles the finger domain of fibronectin; residues 50 through 87 are homologous with epidermal growth factor, whereas residues 92 through 173 and 180 through 261, which have homology to the kringle domains of plasminogen, are designated as the first and second kringle, respectively. The fifth alteplase domain is the protease domain; it is located on the C-terminal B chain of two-chain alteplase.

FIGURE 114-9

Domain structures of alteplase (tPA), tenecteplase (TNK-tPA), and reteplase (r-PA). The finger (F), epidermal growth factor (EGF), first and second kringles (K1 and K2, respectively), and protease (P) domains are illustrated. The glycosylation site (Y) on K1 has been repositioned in tenecteplase to endow it with a longer half-life. In addition, a tetra-alanine substitution in the protease domain renders tenecteplase resistant to type 1 plasminogen activator inhibitor (PAI-1) inhibition. Reteplase is a truncated variant that lacks the F, EGF, and K1 domains.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The interaction of alteplase with fibrin is mediated by the finger domain and, to a lesser extent, by the second kringle domain. The affinity of alteplase for fibrin is considerably higher than that for fibrinogen. Consequently, the catalytic efficiency of plasminogen activation by alteplase is two to three orders of magnitude higher in the presence of fibrin than in the presence of fibrinogen. This phenomenon helps to localize plasmin generation to the fibrin surface.

Although alteplase preferentially activates plasminogen in the presence of fibrin, alteplase is not as fibrin-selective as was first predicted. Its fibrin specificity is limited because like fibrin, (DD)E, which is the complex of D-dimer non-covalently linked to fragment E is the major soluble degradation product of cross-linked fibrin. (DD)E binds alteplase and plasminogen with high affinity. Consequently, (DD)E is as potent as fibrin as a stimulator of plasminogen activation by alteplase. Whereas plasmin generated on the fibrin surface results in thrombolysis, plasmin generated on the surface of circulating (DD)E degrades fibrinogen. Fibrinogen degradation results in the accumulation of fragment X, a high-molecular-weight clottable fibrinogen degradation product. Incorporation of fragment X into hemostatic plugs formed at sites of vascular injury renders them susceptible to lysis. This phenomenon may contribute to alteplase-induced bleeding.

A trial comparing alteplase with streptokinase for treatment of patients with acute MI demonstrated significantly lower mortality with alteplase than with streptokinase, although the absolute difference was small. The greatest benefit was seen in patients age <75 years with anterior MI who presented <6 h after symptom onset.

For treatment of acute MI or acute ischemic stroke, alteplase is given as an IV infusion over 60–90 min. The total dose of alteplase usually ranges from 90 to 100 mg. Allergic reactions and hypotension are rare, and alteplase is not immunogenic.

TENECTEPLASE

Tenecteplase is a genetically engineered variant of tPA and was designed to have a longer half-life than tPA and to be resistant to inactivation by PAI-1. To prolong its half-life, a new glycosylation site was added to the first kringle domain (Fig. 114-9). Because addition of this extra carbohydrate side chain reduced fibrin affinity, the existing glycosylation site on the first kringle domain was removed. To render the molecule resistant to inhibition by PAI-1, a tetra-alanine substitution was introduced at residues 296–299 in the protease domain, the region responsible for the interaction of tPA with PAI-1.

Tenecteplase is more fibrin-specific than tPA. Although both agents bind to fibrin with similar affinity, the affinity of tenecteplase for (DD)E is significantly lower than that of tPA. Consequently, (DD)E does not stimulate systemic plasminogen activation by tenecteplase to the same extent as tPA. As a result, tenecteplase produces less fibrinogen degradation than tPA.

For coronary thrombolysis, tenecteplase is given as a single IV bolus. In a large phase III trial that enrolled >16,000 patients, the 30-day mortality rate with single-bolus tenecteplase was similar to that with accelerated-dose tPA. Although rates of intracranial hemorrhage were also similar with both treatments, patients given tenecteplase had fewer noncerebral bleeds and a reduced need for blood transfusions than those treated with tPA. The improved safety profile of tenecteplase likely reflects its enhanced fibrin specificity.

RETEPLASE

Reteplase is a single-chain, recombinant tPA derivative that lacks the finger, epidermal growth factor, and first kringle domains (Fig. 114-9). This truncated derivative has a molecular weight of 39,000. Reteplase binds fibrin more weakly than tPA because it lacks the finger domain. Because it is produced in *Escherichia coli*, reteplase is not glycosylated. This endows it with a plasma half-life longer than that of tPA. Consequently, reteplase is given as two IV boluses, which are separated by 30 min. Clinical trials have demonstrated that reteplase is at least as effective as streptokinase for treatment of acute MI, but the agent is not superior to tPA.

CONCLUSIONS AND FUTURE DIRECTIONS

Thrombosis involves a complex interplay among the vessel wall, platelets, the coagulation system, and the fibrinolytic pathways. Activation of coagulation also triggers inflammatory pathways that may exacerbate thrombosis. A better understanding of the biochemistry of blood coagulation and advances in structure-based drug design have identified new targets and resulted in the development of novel antithrombotic drugs. Well-designed clinical trials have provided detailed information on which drugs to use and when to use them. Despite these advances, however, thromboembolic disorders remain a major cause of morbidity and mortality. Therefore, the search for better and safer targets continues.

FURTHER READING

Badimon L, Vilahur G: Thrombosis formation on atherosclerotic lesions and plaque rupture. J Int Med 276:618, 2014.

Chapin JC, Hajjar KA: Fibrinolysis and the control of blood coagulation. Blood Rev 29:17, 2015.

Eikelboom JW et al: Idarucizumab: The antidote for reversal of dabigatran. Circulation 132:2412, 2015.

Garcia D et al: Laboratory assessment of the anticoagulant effects of the next generation of oral anticoagulants. J Thromb Haemost 11:245, 2013.

Mackman N: New insights into the mechanisms of venous thrombosis. J Clin Invest 122:2331, 2012.

Mora S et al: Low-dose aspirin in the primary prevention of cardiovascular disease: Shared decision making in clinical practice. JAMA 316:709, 2016.

Mulloy B et al: Pharmacology of heparin and related drugs. Pharmacol Rev 68:76, 2016.

Tantry US et al: Vorapaxar in the secondary prevention of atherothrombosis. Expert Rev Cardiovasc Ther 13:1293, 2015.

Warkentin TE: Heparin-induced thrombocytopenia. Curr Opin Crit Care 21:576, 2015.

Weitz JI: Anticoagulation therapy in 2015: Where we are and where we are going. J Thromb Thrombolysis 39:264, 2015.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 115: Approach to the Patient with an Infectious Disease

Neeraj K. Surana; Dennis L. Kasper

FIGURE 115-1

HISTORICAL PERSPECTIVE

The origins of the field of infectious diseases are humble. The notion that communicable diseases were due to a *miasma* ("bad air") can be traced back to at least the mid-sixteenth century. Not until the work of Louis Pasteur and Robert Koch in the late nineteenth century was there credible evidence supporting the germ theory of disease—i.e., that microorganisms are the direct cause of infections. In contrast to this relatively slow start, the twentieth century saw remarkable advances in the field of infectious diseases, and the etiologic agents of numerous infectious diseases were soon identified. Furthermore, the discovery of antibiotics and the advent of vaccines against some of the most deadly and debilitating infections greatly altered the landscape of human health. Indeed, the twentieth century saw the elimination of smallpox, one of the great scourges in the history of humanity. These remarkable successes prompted Sir Frank MacFarlane Burnet, a noted immunologist and Nobel laureate, to write in a 1962 publication entitled *Natural History of Infectious Diseases*: "In many ways one can think of the middle of the twentieth century as the end of one of the most important social revolutions in history, the virtual elimination of infectious disease." Professor Burnet was not alone in this view. Robert Petersdorf, a renowned infectious disease expert and former editor of this textbook, wrote in 1978 that "even with my great personal loyalties to infectious diseases, I cannot conceive a need for 309 more [graduating trainees in infectious diseases] unless they spend their time culturing each other." Given the enormous growth of interest in the microbiome in the past 10 years, Dr. Petersdorf's statement might have been ironically clairvoyant, although he could have had no idea what was in store for humanity, with an onslaught of new, emerging, and re-emerging infectious diseases.

Clearly, even with all the advances of the twentieth century, infectious diseases continue to represent a formidable challenge for patients and physicians alike. Furthermore, during the latter half of the century, several chronic diseases were demonstrated to be directly or indirectly caused by infectious microbes; perhaps the most notable examples are the associations of *Helicobacter pylori* with peptic ulcer disease and gastric carcinoma, human papillomavirus with cervical cancer, and hepatitis B and C viruses with liver cancer. In fact, ~16% of all malignancies are now known to be associated with an infectious cause. In addition, numerous emerging and re-emerging infectious diseases continue to have a dire impact on global health: HIV/AIDS, pandemic influenza, Ebola, and Zika are but a few examples. The fear of weaponizing pathogens for bioterrorism is ever present and poses a potentially enormous threat to public health. Moreover, escalating antimicrobial resistance in clinically relevant microbes (e.g., *Mycobacterium tuberculosis*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Plasmodium* species, and HIV) signifies that the administration of antimicrobial agents—once thought to be a panacea—requires appropriate stewardship. For all these reasons, infectious diseases continue to exert grim effects on individual patients as well as on international public health. Even with all the successes of the past century, physicians must be as thoughtful about infectious diseases now as they were at the beginning of the twentieth century.

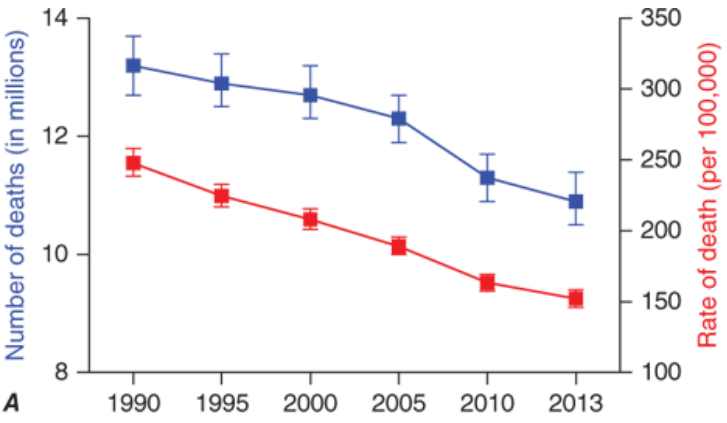
GLOBAL CONSIDERATIONS

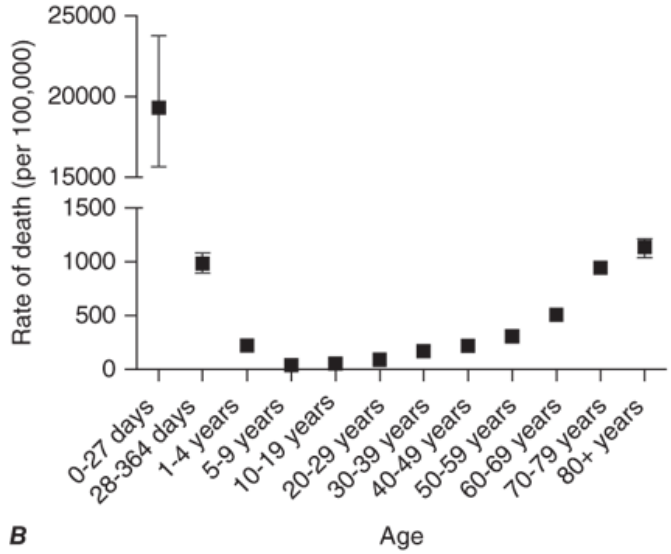
Infectious diseases remain the second leading cause of death worldwide. Although the rate of infectious disease–related deaths has decreased dramatically over the past 25 years, there were still 10.9 million such deaths in 2013 **(Fig. 115-1A)**. These deaths disproportionately affect children <1 year of age, adults older than 70 years, and persons living in low- and middle-income countries **(Fig. 115-1B and 115-1C; Chap. 462)**; in 2013, 20% of all deaths worldwide were related to infectious diseases, with rates >50% in most sub-Saharan African countries.

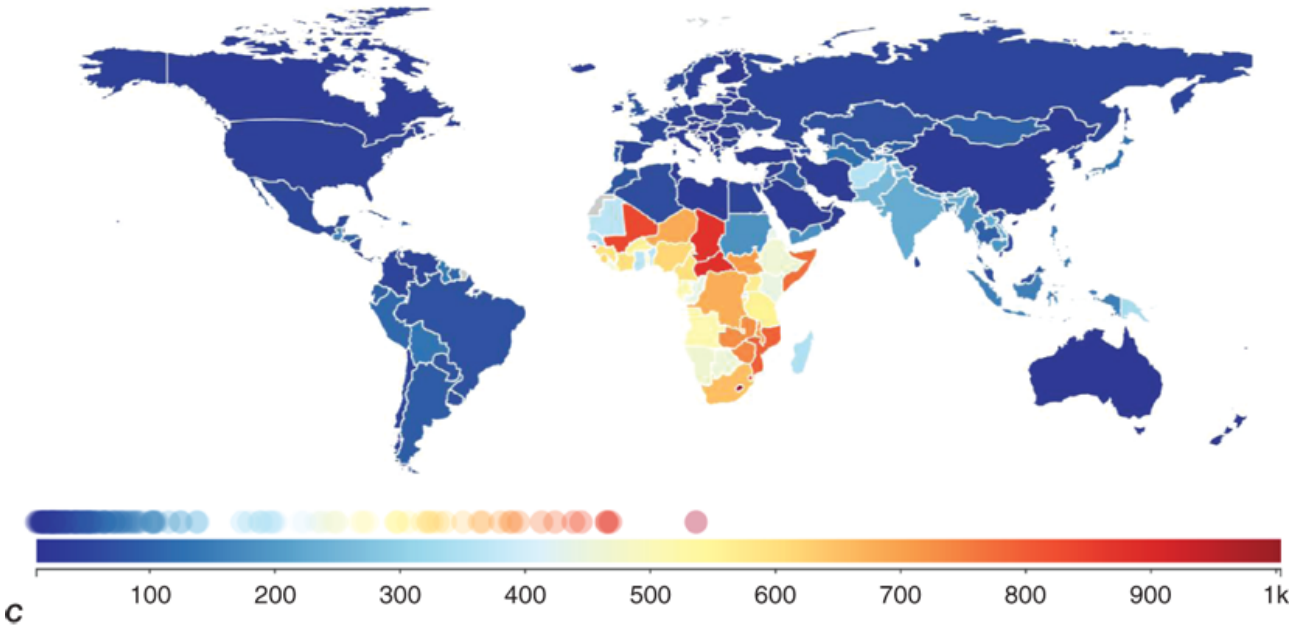
FIGURE 115-1

Magnitude of infectious disease–related deaths globally. ***A.*** The absolute number (*blue line; left axis*) and rate (*red line; right axis*) of infectious disease–related deaths throughout the world since 1990. ***B.*** Age-specific rates of infectious disease–related deaths in 2013. In both ***A*** and ***B***, the charts depict the mean estimate and 95% uncertainty intervals. ***C.*** A map depicting country-specific data for the percentages of

total deaths that were attributable to communicable, maternal, neonatal, and nutritional disorders in 2013. *(Source: Global Burden of Disease Study, Institute for Health Metrics and Evaluation.)*







Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Given that infectious diseases are still a major cause of global mortality, understanding the local epidemiology of disease is critically important in evaluating patients. Diseases such as HIV/AIDS have decimated sub-Saharan Africa, with HIV-infected adults representing 19–

29% of the total population in countries like South Africa, Botswana, and Swaziland. Moreover, drug-resistant tuberculosis is rampant throughout the former Soviet-bloc countries, India, China, and South Africa. The ready availability of this type of information allows physicians to develop appropriate differential diagnoses and treatment plans for individual patients. Programs such as the Global Burden of Disease seek to quantify human losses (e.g., deaths, disability-adjusted life years) due to diseases by age, sex, and country over time; these data not only help inform local, national, and international health policy but can also help guide local medical decision-making.

Even though some diseases (e.g., pandemic influenza, Middle East respiratory syndrome) are seemingly geographically restricted, the increasing ease of rapid worldwide travel has raised concern about their swift spread around the globe. Indeed, human migration has historically been the source of epidemics: *Yersinia pestis* spread along trade routes in the fourteenth century, Native American populations were devastated by diseases such as smallpox and measles that were imported by European explorers in the fifteenth and sixteenth centuries, military maneuvers helped facilitate the spread of the 1918 influenza pandemic, and religious pilgrimages (e.g., the Hajj) provide the means for worldwide dissemination of diseases. The introduction of cholera into Haiti, the transmission of Ebola within the United States, and the emerging outbreak of Zika virus infection are recent examples that highlight the continued effects of global travel on the spread of infectious diseases. Not only can travelers carry person-to-person transmitted infections (e.g., influenza, HIV) anywhere in the world, but they can also introduce vector-borne infections to new geographic areas (e.g., chikungunya and Zika viruses) and contribute to the worldwide spread of multidrug-resistant organisms. The world's increasing interconnectedness has profound implications not only for the global economy but also for medicine and the spread of infectious diseases.

UNDERSTANDING THE MICROBIOTA

Normal, healthy humans are colonized with ~50 trillion bacteria as well as countless viruses, fungi, and archaea; taken together, these microorganisms outnumber human cells by ~10 times in the human body **(Chap. 459)**. The major reservoir of these microbes is the gastrointestinal tract, but substantial numbers of microbes live in the female genital tract, the oral cavity, and the nasopharynx. There is increasing interest in the skin and lungs as sites where microbial colonization might be highly relevant to the biology and disease susceptibility of the host. These commensal organisms provide the host with myriad benefits, from aiding in metabolism to shaping the immune system. With regard to infectious diseases, the vast majority of infections are caused by organisms that are part of the normal microbiota (e.g., *S. aureus*, *S. pneumoniae*, *Pseudomonas aeruginosa*), with relatively few infections due to organisms that are strictly pathogens (e.g., *Neisseria gonorrhoeae*, rabies virus). Perhaps it is not surprising that a general understanding of the microbiota is essential in the evaluation of infectious diseases. Individuals' microbiotas likely have a major impact on their susceptibility to infectious diseases and even their responses to vaccines. Site-specific knowledge of the indigenous microbiota may facilitate appropriate interpretation of culture results, aid in selection of empirical antimicrobial therapy based on the likely causative agents, and provide additional impetus for rational antibiotic use to minimize the untoward effects of these drugs on the "beneficial" microbes that inhabit the body.

WHEN TO CONSIDER AN INFECTIOUS ETIOLOGY

The title of this chapter may appear to presuppose that the physician knows when a patient has an infectious disease. In reality, this chapter can serve only as a guide to the evaluation of a patient in whom an infectious disease is a possibility. Once a specific diagnosis is made, the reader should consult the subsequent chapters that deal with specific microorganisms in detail. The challenge for the physician is to recognize which patients may have an infectious disease as opposed to some other underlying disorder. This task is greatly complicated by the fact that infections have an infinite range of presentations, from acute life-threatening conditions (e.g., meningococcemia) to chronic diseases of varying severity (e.g., *H. pylori*–associated peptic ulcer disease) to no symptoms at all (e.g., latent *M. tuberculosis* infection). While it is impossible to generalize about a presentation that encompasses all infections, common findings in the history, physical examination, and basic laboratory testing often suggest that the patient either has an infectious disease or should be more closely evaluated for one. This chapter focuses on these common findings and how they may direct the ongoing evaluation of the patient.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Infectious Disease

See also Chap. 117.

HISTORY

As in all of medicine, a complete and thorough history is paramount in the evaluation of a patient with a possible infectious disease. The history is critical for developing a focused differential diagnosis and for guiding the physical exam and initial diagnostic testing. Although a detailing of all the elements of a history is beyond the scope of this chapter, specific components relevant to infectious diseases require particular attention. In general, these aspects focus on two areas: (1) an exposure history that may identify microorganisms with which the patient may have come into contact and (2) host-specific factors that may predispose to the development of an infection.

Exposure History

History of infections or exposure to drug-resistant microbes

Information about a patient's previous infections, with the associated microbial susceptibility profiles, is very helpful in determining possible etiologic agents. Specifically, knowing whether a patient has a history of infection with drug-resistant organisms (e.g., methicillin-resistant *S. aureus,* vancomycin-resistant *Enterococcus* species, enteric organisms that produce an extended-spectrum β-lactamase or carbapenemase) or may have been exposed to drug-resistant microbes (e.g., during a recent stay in a hospital, nursing home, or long-term acute-care facility) may alter the choice of empirical antibiotics. For example, a patient presenting with sepsis who is known to have a history of invasive infection with a multidrug-resistant isolate of *P. aeruginosa* should be treated empirically with an antimicrobial regimen that will cover this strain.

Social history

Although the social history taken by physicians is often limited to inquiries about a patient's alcohol and tobacco use, a complete social history can offer a number of clues to the underlying diagnosis. Knowing whether the patient has any high-risk behaviors (e.g., unsafe sexual behaviors, IV drug use), potential hobby-associated exposures (e.g., avid gardening, with possible *Sporothrix schenckii* exposure), or occupational exposures (e.g., increased risk for *M. tuberculosis* exposure in funeral service workers) can facilitate diagnosis. The importance of the social history is exemplified by a case in 2009 in which a laboratory researcher died of a *Y. pestis* infection acquired during his work; although this patient had visited both an outpatient clinic and an emergency department, his records at both sites failed to include his occupation—information that potentially could have led quickly to appropriate treatment and infection control measures.

Dietary habits

As certain pathogens are associated with specific dietary habits, inquiring about a patient's diet can provide insight into possible exposures. For example, Shiga toxin–producing strains of *Escherichia coli* and *Toxoplasma gondii* are associated with the consumption of raw or undercooked meat; *Salmonella typhimurium*, *Listeria monocytogenes*, and *Mycobacterium bovis* with unpasteurized milk; *Leptospira* species, parasites, and enteric bacteria with unpurified water; and *Vibrio* species, norovirus, helminths, and protozoa with raw seafood.

Animal exposures

Because animals are often important vectors of infectious diseases, patients should be asked about exposures to any animals, including contact with their own pets, visits to petting zoos, or random encounters (e.g., home rodent infestation). For example, dogs can carry ticks that serve as agents for the transmission of several infectious diseases, including Lyme disease, Rocky Mountain spotted fever, and ehrlichiosis. Cats are associated with *Bartonella henselae* infection, reptiles with *Salmonella* infection, rodents with leptospirosis, and rabbits with tularemia **(Chap. 136)**.

Travel history

Attention should be paid to both international and domestic travel. Fever in a patient who has recently returned from abroad significantly broadens the differential diagnosis **(Chap. 119)**; even a remote history of international travel may reflect patients' exposure to infections with pathogens such as *M. tuberculosis* or *Strongyloides stercoralis*. Similarly, domestic travel may have exposed patients to pathogens that are not normally found in their local environment and therefore may not routinely be considered in the differential diagnosis. For example, a patient who has recently visited California or Martha's Vineyard may have been exposed to *Coccidioides immitis* or *Francisella tularensis*, respectively. Beyond simply identifying locations that a patient may have visited, the physician needs to delve deeper to learn what kinds of activities and behaviors the patient engaged in during travel (e.g., the types of food and sources of water consumed, freshwater swimming, animal exposures) and whether the patient had the necessary immunizations and/or took the necessary prophylactic medications prior to travel; these additional exposures, which the patient may not think to report without specific prompting, are as important as exposures during a patient's routine daily living.

Host-Specific Factors

Because many opportunistic infections (e.g., with *Pneumocystis jirovecii*, *Aspergillus* species, or JC virus) affect primarily immunocompromised patients, it is of vital importance to determine the immune status of the patient. Defects in the immune system may be due to an underlying disease (e.g., malignancy, HIV infection, malnutrition), a medication (e.g., chemotherapy, glucocorticoids, monoclonal

antibodies to components of the immune system), a treatment modality (e.g., total body irradiation, splenectomy), or a primary immunodeficiency. The type of infection for which the patient is at increased risk varies with the specific type of immune defect. In concert with determining whether a patient is immunocompromised for any reason, the physician should review the immunization record to ensure that the patient is adequately protected against vaccine-preventable diseases (Chap. 118).

PHYSICAL EXAMINATION

Like the history, a thorough physical examination is crucial in evaluating patients with an infectious disease. Some elements of the physical exam (e.g., skin, lymphatics) that are often performed in a cursory manner as a result of the ever-increasing pace of medical practice may help identify the underlying diagnosis. Moreover, serial exams are critical since new findings may appear as the illness progresses. A description of all the elements of a physical exam is beyond the scope of this chapter, but the following components have particular relevance to infectious diseases.

Vital Signs

Given that elevations in temperature are often a hallmark of infection, paying close attention to the temperature may be of value in diagnosing an infectious disease. The idea that 37°C (98.6°F) is the normal human body temperature dates back to the nineteenth century and was initially based on axillary measurements. Rectal temperatures more accurately reflect the core body temperature and are 0.4°C (0.7°F) and 0.8°C (1.4°F) higher than oral and axillary temperatures, respectively. Although the definition of fever varies greatly throughout the medical literature, the most common definition, which is based on studies defining fever of unknown origin (Chap. 17), uses a core temperature ≥38.3°C (≥101°F). Although fever is very commonly associated with infection, it is also documented in many other diseases (Chap. 15). For every 1°C (1.8°F) increase in core temperature, the heart rate typically rises by 15–20 beats/min. Table 115-1 lists infections that are associated with relative bradycardia (*Faget's sign*), where patients have a lower heart rate than might be expected for a given body temperature. Although this pulse–temperature dissociation is not highly sensitive or specific for establishing a diagnosis, it is potentially useful in low-resource settings given its ready availability and simplicity.

Lymphatics

There are ~600 lymph nodes throughout the body, and infections are an important cause of lymphadenopathy. A physical examination should include evaluation of lymph nodes in multiple regions (e.g., popliteal, inguinal, epitrochlear, axillary, multiple cervical regions), with notation of the location, size (normal, <1 cm), presence or absence of tenderness, and consistency (soft, firm, or rubbery) and of whether the nodes are matted (i.e., connected and moving together). Nodes that are small and firm can also be described as "shotty," referring to the size and consistency of buckshot pellets. Of note, palpable epitrochlear nodes are always pathologic. Of patients presenting with lymphadenopathy, 75% have localized findings, and the remaining 25% have generalized lymphadenopathy (i.e., that involving more than one anatomic region). Localized lymphadenopathy in the head and neck region is found in 55% of patients, inguinal lymphadenopathy in 14%, and axillary lymphadenopathy in 5%. Determining whether the patient has generalized versus localized lymphadenopathy can help narrow the differential diagnosis, as various infections present differently.

Skin

The fact that many infections have cutaneous manifestations gives the skin examination particular importance in the evaluation of patients (Chaps. 16, 54, 124, and A1). It is important to perform a complete skin exam, with attention to both front and back. Specific rashes are often extremely helpful in narrowing the differential diagnosis of an infection (Chaps. 16 and A1). In numerous anecdotal instances, patients in the intensive care unit have had "fever of unknown origin" that was actually due to unrecognized pressure ulcers. Moreover, close examination of the distal extremities for splinter hemorrhages, Janeway lesions, or Osler's nodes may yield evidence of endocarditis or other causes of septic emboli.

Foreign Bodies

As previously mentioned, many infections are caused by members of the indigenous microbiota. These infections typically occur when these microbes escape their normal habitat and enter a new one. Thus, maintenance of epithelial barriers is one of the most important mechanisms in protection against infection. However, hospitalization of patients is often associated with breaches of these barriers—e.g., due to placement of IV lines, surgical drains, or tubes (such as endotracheal tubes and Foley catheters) that allow microorganisms to localize in sites to which they normally would not have access (Chap. 137). Accordingly, knowing what lines, tubes, and drains are in place is helpful in ascertaining what body sites might be infected.

DIAGNOSTIC TESTING

Laboratory and radiologic testing has advanced greatly over the past few decades and has become an important component in the evaluation of patients. The dramatic increase in the number of serologic diagnostics, antigen tests, and molecular diagnostics available to

the physician has, in fact, revolutionized medical care. However, all of these tests should be viewed as adjuncts to the history and physical examination—not a replacement for them. The selection of initial tests should be based directly on the patient's history and physical exam findings. Moreover, diagnostic testing should generally be limited to those conditions that are reasonably likely and treatable, important in terms of public health considerations, and/or capable of providing a definitive diagnosis that will consequently limit other testing.

White Blood Cell (WBC) Count

Elevations in the WBC count are often associated with infection, though many viral infections are associated with leukopenia. It is important to assess the WBC differential, given that different classes of microbes are associated with various leukocyte types. For example, bacteria are associated with an increase in polymorphonuclear neutrophils, often with elevated levels of earlier developmental forms such as bands; viruses are associated with an increase in lymphocytes; and certain parasites are associated with an increase in eosinophils. Table 115-2 lists the major infectious causes of eosinophilia.

Inflammatory Markers

The erythrocyte sedimentation rate (ESR) and the C-reactive protein (CRP) level are indirect and direct measures of the acute-phase response, respectively, that can be used to assess a patient's general level of inflammation. Moreover, these markers can be followed serially over time to monitor disease progress/resolution. It is noteworthy that the ESR changes relatively slowly, and its measurement more often than weekly usually is not useful; in contrast, CRP concentrations change rapidly, and daily measurements can be useful in the appropriate context. Although these markers are sensitive indicators of inflammation, neither is very specific. An extremely elevated ESR (>100 mm/h) has a 90% predictive value for a serious underlying disease (Table 115-3). Work is ongoing to identify other potentially useful inflammatory markers (e.g., procalcitonin, serum amyloid A protein); however, their clinical utility requires further validation.

Analysis of Cerebrospinal Fluid (CSF)

Assessment of CSF is critical for patients with suspected meningitis or encephalitis. An opening pressure should always be recorded, and fluid should routinely be sent for cell counts, Gram's stain and culture, and determination of glucose and protein levels. A CSF Gram's stain typically requires >10^5 bacteria/mL for reliable positivity; its specificity approaches 100%. Table 115-4 lists the typical CSF profiles for various infections. In general, CSF with lymphocytic pleocytosis and a low glucose concentration suggests either infection (e.g., with *Listeria*, *M. tuberculosis*, or a fungus) or a noninfectious disorder (e.g., neoplastic meningitis, sarcoidosis). Bacterial antigen tests of CSF (e.g., latex agglutination tests for *Haemophilus influenzae* type b, group B *Streptococcus*, *S. pneumoniae*, and *Neisseria meningitidis*) are not recommended for screening, given that these tests are no more sensitive than Gram's stain; however, these assays can be helpful in presumptively identifying organisms seen on Gram's stain. In contrast, other antigen tests (e.g., for *Cryptococcus*) and some CSF serologic testing (e.g., for *Treponema pallidum*, *Coccidioides*) are highly sensitive and are useful for select patients. In addition, polymerase chain reaction (PCR) analysis of CSF is increasingly being used for the diagnosis of bacterial (e.g., *N. meningitidis*, *S. pneumoniae*, mycobacteria) and viral (e.g., herpes simplex virus, enterovirus) infections; while these molecular tests permit rapid diagnosis with a high degree of sensitivity and specificity, they often do not allow determination of antimicrobial resistance profiles.

Cultures

The mainstays of infectious disease diagnosis include the culture of infected tissue (e.g., surgical specimens) or fluid (e.g., blood, urine, sputum, pus from a wound). Samples can be sent for culture of bacteria (aerobic or anaerobic), fungi, or viruses. Ideally, specimens are collected before the administration of antimicrobial therapy; in instances where this order of events is not clinically feasible, microscopic examination of the specimen (e.g., Gram-stained or potassium hydroxide [KOH]–treated preparations) is particularly important. Culture of the organism(s) allows identification of the etiologic agent(s), determination of the antimicrobial susceptibility profile, and—when there is concern about an outbreak—isolate typing. While cultures are extremely useful in the evaluation of patients, determining whether culture results are clinically meaningful or represent contamination (e.g., a non-*aureus*, non-*lugdunensis* staphylococcal species growing in a blood culture) can sometimes be challenging and requires an understanding of the patient's immune status, exposure history, and microbiota. In some cases, serial cultures to demonstrate clearance of the organism may be helpful.

Pathogen-Specific Testing

Numerous pathogen-specific tests (e.g., serology, antigen testing, PCR testing) are commercially available, and many hospitals now offer some of these tests in-house to facilitate rapid turnaround that ultimately enhances patient care. The reader is directed to relevant chapters on the pathogens of interest for specific details. Some of these tests (e.g., universal PCRs) identify organisms that currently are not cultivable and have unclear relationships to disease, thereby complicating diagnosis. As these tests become more commonplace and the work of the Human Microbiome Project progresses, the relevance of some of these previously unrecognized bacteria to human health will likely become more apparent.

Radiology

Imaging provides an important adjunct to the physical examination, allowing evaluation for lymphadenopathy in regions that are not externally accessible (e.g., mediastinum, intraabdominal sites), assessment of internal organs for evidence of infection, and facilitation of image-guided percutaneous sampling of deep spaces. The choice of imaging modality (e.g., CT, MRI, ultrasound, nuclear medicine, use of contrast) is best made in consultation with a radiologist to ensure that the results will address the physician's specific concerns.

TREATMENT

Physicians often must balance the need for empirical antibiotic treatment with the patient's clinical condition. When clinically feasible, it is best to obtain relevant samples (e.g., blood, CSF, tissue, purulent exudate) for culture prior to the administration of antibiotics, as antibiotic treatment often makes subsequent diagnosis more difficult. Although a general maxim for antibiotic treatment is to use a regimen with as narrow a spectrum as possible **(Chap. 139)**, empirical regimens are necessarily somewhat broad, given that a specific diagnosis has not yet been made. **Table 115-5** lists empirical antibiotic treatment regimens for commonly encountered infectious presentations. These regimens should be narrowed as appropriate once a specific diagnosis is made. In addition to antibiotics, there is sometimes a role for adjunctive therapies, such as intravenous immunoglobulin G (IVIG) pooled from healthy adults or hyperimmune globulin prepared from the blood of individuals with high titers of specific antibodies to select pathogens (e.g., cytomegalovirus, hepatitis B virus, rabies virus, vaccinia virus, *Clostridium tetani*, varicella-zoster virus, *Clostridium botulinum* toxin). Although the data suggesting efficacy are limited, IVIG is sometimes used for patients with suspected staphylococcal or streptococcal toxic shock syndrome.

INFECTION CONTROL

When evaluating a patient with a suspected infectious disease, the physician must consider what infection control methods are necessary to prevent transmission of any possible infection to other people. In 2007, the U.S. Centers for Disease Control and Prevention published guidelines for isolation precautions that are available for download at *www.cdc.gov/hicpac/2007IP/2007isolationPrecautions.html*. Persons exposed to certain pathogens (e.g., *N. meningitidis*, HIV, *Bacillus anthracis*) should receive postexposure prophylaxis to prevent disease acquisition. (See relevant chapters for details on specific pathogens.)

WHEN TO OBTAIN AN INFECTIOUS DISEASE CONSULT

At times, primary physicians need assistance with patient management from a diagnostic and/or therapeutic perspective. Multiple studies have demonstrated that an infectious disease consult is associated with improved outcomes, shorter length of hospital stay, and decreased costs for patients with various diseases. For example, in a prospective cohort study of patients with *S. aureus* bacteremia, infectious disease consultation was independently associated with a 56% reduction in 28-day mortality. In addition, infectious disease specialists provide other services (e.g., infection control, antimicrobial stewardship, management of outpatient antibiotic therapy, occupational exposure programs) that have been shown to benefit patients. Whenever such assistance would be advantageous to a patient with a possible infection, the primary physician should opt for an infectious disease consult. Specific situations that might prompt a consult include (1) difficult-to-diagnose patients with presumed infections, (2) patients who are not responding to treatment as expected, (3) patients with a complicated medical history (e.g., organ transplant recipients, patients immunosuppressed due to autoimmune or inflammatory conditions), and (4) patients with "exotic" diseases (i.e., diseases that are not typically seen within the region).

TABLE 115-1

Causes of Relative Bradycardia

Infectious Causes	
Intracellular organisms	
Gram-negative bacteria	*Salmonella typhi* *Francisella tularensis* *Brucella* spp. *Coxiella burnetii* (Q fever) *Leptospira interrogans* *Legionella pneumophila* *Mycoplasma pneumoniae*
Tick-borne organisms	*Rickettsia* spp. *Orientia tsutsugamushi* (scrub typhus) *Babesia* spp.
Other	*Corynebacterium diphtheriae* *Plasmodium* spp. (malaria)
Viruses/viral infections	Yellow fever virus Dengue virus Viral hemorrhagic fevers[a] Viral myocarditis
Noninfectious Causes	
	Drug fever Beta blocker use Central nervous system lesions Malignant lymphoma Factitious fever

[a]Primarily early in the course of infection with Marburg or Ebola virus.

TABLE 115-2

Major Infectious Causes of Eosinophilia[a]

Organ Involved	Organism	Exposure	Geographic Distribution	Degree of Eosinophilia[b]
Central nervous system	*Angiostrongylus* *Gnathostoma*	Raw seafood Raw poultry and seafood	Asia Asia	Mild Moderate to extreme
Eye	*Loa loa* *Onchocerca*	Insect bite Insect bite	Africa Africa	Moderate (expatriates), mild (patients living in endemic areas) Mild (expatriates), moderate (patients living in endemic areas)
Lung	*Chlamydia trachomatis* *Strongyloides* *Toxocara canis/Toxocara cati*[c] *Paragonimus* *Coccidioides immitis* *Brugia malayi* *Pneumocystis jirovecii*	Sexual transmission Soil Dogs, soil Crabs and crayfish Soil Insect bite Air	Worldwide Tropical Worldwide Asia Southwestern United States Asia Worldwide	Mild Moderate (acute), mild (chronic) Moderate to extreme Moderate (acute), mild (chronic) Mild (acute), extreme (disseminated) Mild to moderate Mild
Liver	*Schistosoma japonicum* *Schistosoma mansoni* *Fasciola* *Clonorchis* *Opisthorchis*	Freshwater swimming Freshwater swimming Watercress Raw seafood Raw seafood	Asia Africa, Middle East, Latin America Worldwide Asia Asia	Moderate (acute), mild (chronic) Moderate (acute), mild (chronic) Moderate Mild to moderate Mild to moderate
Intestines	*Ascaris*[d] Hookworm *Trichuris* *Cystoisospora belli* *Dientamoeba fragilis* *Capillaria* *Heterophyes* *Anisakis* *Baylisascaris procyonis*[e] *Hymenolepis nana*	Raw fruits and vegetables, contaminated water Soil Raw fruits and vegetables, contaminated water Contaminated water and food Unclear; spread via fecal–oral route Raw seafood Raw seafood Raw seafood Soil Contaminated water, soil	Worldwide Worldwide Tropical Worldwide Worldwide Asia Asia, Middle East Worldwide North America Worldwide	Mild to extreme Mild to moderate Mild Mild Mild Extreme Mild Mild Moderate to extreme Mild
Bladder	*Schistosoma haematobium*	Freshwater swimming	Africa, Middle East	Moderate (acute), mild (chronic)
Muscle	*Trichinella*	Pork	Worldwide	Moderate to extreme

Organ Involved	Organism	Exposure	Geographic Distribution	Degree of Eosinophilia[b]
Lymphatics	*Wuchereria bancrofti*[d] *Bartonella henselae*	Insect bite Cats	Tropical Worldwide	Moderate to extreme[f] Mild
Other	Recovery from bacterial or viral infections HIV *Cryptococcus neoformans*	— Contaminated bodily fluid Soil	— Worldwide Worldwide	Mild Mild Moderate to extreme (disseminated)

[a]There are numerous noninfectious causes of eosinophilia, such as atopic disease, DRESS (drug reaction with eosinophilia and systemic symptoms) syndrome, and pernicious anemia, which can cause mild eosinophilia; drug hypersensitivity and serum sickness, which can cause mild to moderate eosinophilia; collagen vascular disease, which can cause moderate eosinophilia; and malignancy, Churg-Strauss syndrome, and hyper-IgE syndromes, which can cause moderate to extreme eosinophilia. [b]Mild: 500–1500 cells/μL; moderate: 1500–5000 cells/μL; extreme: >5000 cells/μL. [c]Can also affect the liver and the eyes. [d]Can also affect the lungs. [e]Can also affect the eyes and the central nervous system. [f]Levels are typically higher with pulmonary infections.

TABLE 115-3

Causes of an Extremely Elevated Erythrocyte Sedimentation Rate (>100 mm/h)

Etiologic Category (% of Cases)	Specific Causes
Infectious diseases (35–40)	Subacute bacterial endocarditis Abscesses Osteomyelitis Tuberculosis Urinary tract infection
Inflammatory diseases (15–20)	Giant cell arteritis Rheumatoid arthritis Systemic lupus erythematosus
Malignancies (15–20)	Multiple myeloma Leukemias Lymphomas Carcinomas
Other (20–35)	Drug hypersensitivity reactions (drug fever) Ischemic tissue injury/trauma Renal diseases

TABLE 115-4

Typical Cerebrospinal Fluid Profiles for Meningitis and Encephalitis[a]

	Normal	Bacterial Meningitis	Viral Meningitis	Fungal Meningitis[b]	Parasitic Meningitis	Tuberculous Meningitis	Encephalitis
WBC count (per µL)	<5	>1000	25–500	40–600	150–2000	25–100	50–500
Differential of WBC	60–70% lymphocytes, ≤30% monocytes/macrophages	↑↑PMNs (≥80%)	Predominantly lymphocytes[c]	Lymphocytes or PMNs, depending on specific organism	↑↑ Eosinophils (≥50%)[d]	Predominantly lymphocytes[c]	Predominantly lymphocytes[c]
Gram's stain	Negative	Positive (in >60% of cases)	Negative	Rarely positive	Negative	Occasionally positive[e]	Negative
Glucose (mg/dL)	40–85	<40	Normal	↓ to normal	Normal	<50 in 75% of cases	Normal
Protein (mg/dL)	15–45	>100	20–80	150–300	50–200	100–200	50–100
Opening pressure (mmH_2O)	50–180	>300	100–350	160–340	Normal	150–280	Normal to ↑
Common causes	—	*Streptococcus pneumoniae*, *Neisseria meningitidis*	Enteroviruses	*Candida*, *Cryptococcus*, and *Aspergillus* spp.	*Angiostrongylus cantonensis*, *Gnathostoma spinigerum*, *Baylisascaris procyonis*	*Mycobacterium tuberculosis*	Herpesviruses, enteroviruses, influenza virus, rabies virus

[a]Numbers indicate typical results, but actual results may vary. [b]Cerebrospinal fluid characteristics depend greatly on the specific organism. [c]Neutrophils may predominate early in the disease course. [d]Patients typically have striking eosinophilia as well. [e]Sensitivity can be increased by examination of a smear of protein coagulum (pellicle) and the use of acid-fast stains.

Abbreviations: PMNs, polymorphonuclear neutrophils; WBC, white blood cell.

TABLE 115-5

Initial Empirical Antibiotic Therapy for Common Infectious Disease Presentations[a]

Clinical Syndrome	Common Etiologies	Antibiotic(s)	Comments	See Chapter(s)
Septic shock	*Staphylococcus aureus, Streptococcus pneumoniae,* enteric gram-negative bacilli	Vancomycin, 15 mg/kg q12h[b] ***plus*** A broad-spectrum antipseudomonal β-lactam (piperacillin-tazobactam, 4.5 g q6h; imipenem, 1 g q8h; meropenem, 1 g q8h; or cefepime, 1–2 g q8–12h)	—	297
Meningitis	*S. pneumoniae*, *Neisseria meningitidis*	Vancomycin, 15 mg/kg q12h[b] ***plus*** Ceftriaxone, 2 g q12h	Dexamethasone (0.15 mg/kg IV q6h for 2–4 d) should be added for patients with suspected or proven pneumococcal meningitis, with the first dose administered 10–20 min before the first dose of antibiotics.	133 and pathogen-specific chapters
CNS abscess	*Streptococcus* spp., *Staphylococcus* spp., anaerobes, gram-negative bacilli	Vancomycin, 15 mg/kg q12h[b] ***plus*** Ceftriaxone, 2 g q12h ***plus*** Metronidazole, 500 mg q8h	—	133
Acute endocarditis (native valve)	*S. aureus, Streptococcus* spp., coagulase-negative staphylococci	Vancomycin, 15 mg/kg q12h[b] ***plus*** Cefepime, 2 g q8h	—	123

Clinical Syndrome	Common Etiologies	Antibiotic(s)	Comments	See Chapter(s)
Pneumonia Community-acquired, outpatient	*S. pneumoniae, Mycoplasma pneumoniae, Haemophilus influenzae, Chlamydia pneumoniae*	Azithromycin, 500 mg PO × 1, then 250 mg PO qd × 4 days A respiratory fluoroquinolone (moxifloxacin, 400 mg IV/PO qd; gemifloxacin, 320 mg PO qd; or levofloxacin, 750 mg IV/PO qd)	If MRSA is a consideration, add vancomycin (15 mg/kg q12h[b]) or linezolid (600 mg q12h); daptomycin should not be used in patients with pneumonia.	121 and pathogen-specific chapters
Inpatient, non-ICU		***or*** A β-lactam (cefotaxime, ceftriaxone, or ampicillin-sulbactam) *plus* azithromycin		
Inpatient, ICU	Above plus *Legionella* spp.	A β-lactam ***plus*** Azithromycin *or* a respiratory fluoroquinolone		
Hospital-acquired pneumonia[d]	*S. pneumoniae, H. influenzae, S. aureus*, gram-negative bacilli (e.g., *Pseudomonas aeruginosa, Klebsiella pneumoniae, Acinetobacter* spp.)	An antipseudomonal β-lactam (cefepime, 1–2 g q8–12h; ceftazidime, 2 g q8h; imipenem, 1 g q8h; meropenem, 1 g q8h; or piperacillin-tazobactam, 4.5 g q6h) ***plus*** An antipseudomonal fluoroquinolone (levofloxacin or ciprofloxacin, 400 mg q8h) *or* an aminoglycoside (amikacin, 20 mg/kg q24h[c]; gentamicin, 7 mg/kg q24h[e]; or tobramycin, 7 mg/kg q24h[e])		
Complicated intraabdominal infection Mild to moderate severity	Anaerobes (*Bacteroides* spp., *Clostridium* spp.), gram-negative bacilli (*Escherichia coli*), *Streptococcus* spp.	Cefoxitin, 2 g q6h ***or*** A combination of metronidazole (500 mg q8–12h) *plus one of the following*: cefazolin (1–2 g q8h), cefuroxime (1.5 g q8h), ceftriaxone (1–2 g q12–24h), cefotaxime (1–2 g q6–8h)	If MRSA is a consideration, add vancomycin (15 mg/kg q12h[b]).	127, 172, and pathogen-specific chapters
High-risk patient or high degree of severity	Same as above	A carbapenem (imipenem, 1 g q8h; meropenem, 1 g q8h; doripenem, 500 mg q8h) ***or*** Piperacillin-tazobactam, 3.375 g q6h[f] ***or*** A combination of metronidazole (500 mg q8–12h) ***plus either*** an antipseudomonal cephalosporin (cefepime, 2 g q8–12h; ceftazidime, 2 g q8h) *or* an antipseudomonal fluoroquinolone (ciprofloxacin, 400 mg q12h; levofloxacin, 750 mg q24h)		

Clinical Syndrome	Common Etiologies	Antibiotic(s)	Comments	See Chapter(s)
Skin and soft tissue infection	*S. aureus*, *Streptococcus pyogenes*	Dicloxacillin, 250–500 mg PO qid ***or*** Cephalexin, 250–500 mg PO qid ***or*** Clindamycin, 300–450 mg PO tid ***or*** Nafcillin/oxacillin, 1–2 g q4h	If MRSA is a consideration, clindamycin, vancomycin (15 mg/kg q12h[b]), linezolid (600 mg IV/PO q12h), or TMP-SMX (1–2 double-strength tablets PO bid[g]) can be used.	**124** and pathogen-specific chapters

[a]This table refers to immunocompetent adults with normal renal and hepatic function. All doses listed are for parenteral administration unless indicated otherwise. Local antimicrobial susceptibility profiles may influence the choice of antibiotic. Therapy should be tailored once a specific etiologic agent and its susceptibilities are identified. [b]Trough levels for vancomycin should be 15–20 µg/mL. [c]Trough levels for amikacin should be <4 µg/mL. [d]In patients with late onset (i.e., after ≥5 days of hospitalization) or risk factors for multidrug-resistant organisms. [e]Trough levels for gentamicin and tobramycin should be <1 µg/mL. [f]If *P. aeruginosa* is a concern, the dosage may be increased to 3.375 g IV q4h or 4.5 g IV q6h. [g]Data on the efficacy of TMP-SMX in skin and soft tissue infections are limited.

Abbreviations: CNS, central nervous system; ICU, intensive care unit; MRSA, methicillin-resistant *S. aureus*; TMP-SMX, trimethoprim-sulfamethoxazole.

PERSPECTIVE

The study of infectious diseases is really a study of host–bacterial interactions and represents evolution by both the host and the bacteria—an endless struggle in which microbes have generally been more creative and adaptive. Given that one-fifth of deaths worldwide are still related to infectious diseases, it is clear that the war against infectious diseases has not been won. For example, a cure for HIV infection is still lacking, there have been only marginal improvements in the methods for detection and treatment of tuberculosis after more than a half century of research, new infectious disease outbreaks (e.g., pandemic influenza, viral hemorrhagic fevers, Zika) continue to emerge, and the threat of microbial bioterrorism remains high. The subsequent chapters in Part 5 detail—on both a syndrome and a microbe-by-microbe basis—the current state of medical knowledge about infectious diseases. At their core, all of these chapters carry a similar message: Despite numerous advances in the diagnosis, treatment, and prevention of infectious diseases, much work and research are required before anyone can confidently claim we have achieved “the virtual elimination of infectious disease.” In reality, this goal will never be attained, given the rapid adaptability of microbes.

FURTHER READING

Bartlett JG: Why infectious diseases. Clin Infect Dis 59(Suppl 2):S85, 2014. [PubMed: 25151484]

Khabbaz RF et al: Challenges of infectious diseases in the USA. Lancet 384:53, 2014.

Marston HD et al: Antimicrobial resistance. JAMA 316:1193, 2016.

Schmitt S et al: Infectious diseases specialty intervention is associated with decreased mortality and lower healthcare costs. Clin Infect Dis 58:22, 2014.

Verghese A et al: Inadequacies of physical examination as a cause of medical errors and adverse events: A collection of vignettes. Am J Med 128:1322, 2015.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.

Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 116: Molecular Mechanisms of Microbial Pathogenesis

Gerald B. Pier

INTRODUCTION

Over the past five decades, molecular studies of the pathogenesis of microorganisms have yielded a torrent of information about the various microbial and host molecules that contribute to the processes of infection and disease. These processes can be classified into several stages: microbial encounter with and entry into the host; microbial growth after entry; avoidance of innate host defenses; tissue invasion and tropism; tissue damage; and transmission to new hosts. *Virulence* is the measure of an organism's capacity to cause disease and is a function of the pathogenic factors elaborated by microbes. These factors promote *colonization* (the simple presence of potentially pathogenic microbes in or on a host), *infection* (attachment and growth of pathogens and avoidance of host defenses), and *disease* (often, but not always, the result of activities of secreted toxins or toxic metabolites). In addition, the host's inflammatory response to infection greatly contributes to disease and its attendant clinical signs and symptoms. A recent explosion of interest in the *microbiome* (the collection of microbial genomes present in or on mammalian organisms) and the *microbiota* (the collection of microbes residing in and on mammalian organisms) and their impact on physiology of, susceptibility to, and response to infection and on immune system development has greatly expanded our understanding of host–pathogen interactions. Furthermore, investigations in this field have documented effects of the microbiome on all aspects of animal—and even plant—physiology, greatly increasing our knowledge of the everyday influence of host–microbe interactions on life.

MICROBIAL ENTRY AND ADHERENCE

The Microbiome

We now know that the indigenous microbial organisms living in close association with almost all animals and plants are organized into complex communities that strongly modulate overall host physiology, including the ability of pathogenic microbes to establish themselves in or on host surfaces. The sheer numbers of these microbes and their genomic variability often exceed the numbers of host cells and the variability of host genes in a typical animal. Changes and differences in microbiomes within and between individuals, currently characterized by high-throughput DNA sequencing techniques and bioinformatic analysis, impact such diverse conditions as obesity; type 1 diabetes; cognition; neurologic states; autoimmune diseases; skin, gastrointestinal, respiratory, and vaginal infectious diseases; and development and control of the immune system. It has been difficult to directly associate specific types of microbiomes with pathophysiologic states, and our understanding of the degree to which microbial species are conserved or variable within human and other animal microbiotas is evolving. Experimental studies in laboratory animals, particularly in germ-free mammals, show the potent ability of changes in the microbiota to manipulate health status and outcomes. One of the clearest functions of the microbiota is to influence and mature the cells of the immune system, thereby exerting a major effect on susceptibility and resistance to microbial infection. The degree to which studies of the microbiome will translate into strategies for the management of human health and disease (e.g., the use of fecal transplants to treat and prevent recurrences of serious *Clostridium difficile* infection) is still an open question. For the moment, defining clusters of organisms associated with diseases may be more feasible than identifying single organisms or microbial molecules. Results from the Human Microbiome Project suggest a high level of variability among individuals in microbiome components, although many individuals appear to maintain a fairly conserved microbiome throughout their lives. In the context of infectious diseases, changes and disruptions of the indigenous microbiome—i.e., alterations of the normal flora due to antibiotic and immunosuppressive drug use, environmental changes, and the effects of microbial virulence factors used to displace the indigenous microbial flora and thus to facilitate pathogen colonization—have a strong and often fundamental impact on the progression of infection. While the technology for defining and understanding the microbiome is still quite young, there is little doubt that the resulting data will markedly affect our concepts of and approaches to microbial pathogenesis and infectious disease treatment.

Entry Sites

A microbial pathogen can potentially enter any part of a host organism. In general, the type of disease produced by a particular microbe is often a direct consequence of its route of entry into the body. The most common sites of entry are mucosal surfaces (the respiratory, alimentary, and urogenital tracts) and the skin. Ingestion, inhalation, and sexual contact are typical routes of microbial entry. Other portals of entry include sites of skin injury (cuts, bites, burns, trauma) along with injection via natural (e.g., vector-borne) or artificial (e.g., needlestick injury) routes. A few pathogens, such as *Schistosoma* species, can penetrate unbroken skin. The conjunctiva can serve as an entry point for pathogens of the eye, which occasionally spread systemically from that site.

Microbial entry usually relies on the presence of specific factors needed for persistence and growth in a tissue. Fecal–oral spread via the alimentary tract requires a biologic profile consistent with survival in the varied environments of the gastrointestinal tract (including the low pH of the stomach and the high bile content of the intestine) as well as in contaminated food or water outside the host. Organisms that gain entry via the respiratory tract survive well in small moist droplets produced during sneezing and coughing. Pathogens that enter by venereal routes often survive best in the warm moist environment of the urogenital mucosa and have restricted host ranges (e.g., *Neisseria gonorrhoeae*, *Treponema pallidum*, and HIV).

The biology of microbes entering through the skin is highly varied. Some of these organisms can survive under a broad range of environmental conditions, such as those in the salivary glands or alimentary tracts of arthropod vectors, the mouths of larger animals, soil, and water. A complex biology allows protozoan parasites such as *Plasmodium*, *Leishmania*, and *Trypanosoma* species to undergo morphogenic changes that permit transmission of the organism to mammalian hosts during insect feeding for blood meals. Plasmodia are injected as infective sporozoites from the salivary glands during mosquito feeding. *Leishmania* parasites are regurgitated as promastigotes from the alimentary tract of sandflies and injected by bite into a susceptible host. Trypanosomes are first ingested from infected hosts by reduviid bugs; the pathogens then multiply in the gastrointestinal tract of the insects and are released in feces onto the host's skin during subsequent feedings. Most microbes that land directly on intact skin are destined to die, as survival on the skin or in hair follicles requires resistance to fatty acids, low pH, and other antimicrobial factors on the skin. Once it is damaged (and particularly if it becomes necrotic), the skin can be a major portal of entry and growth for pathogens and elaboration of their toxic products. Burn wound infections and tetanus are clear examples. After animal bites, pathogens resident in the animal's saliva gain access to the victim's tissues through the damaged skin. Rabies is the paradigm for this pathogenic process; rabies virus grows in striated muscle cells at the site of inoculation.

Microbial Adherence

Once in or on a host, many microbes must situate themselves favorably to avoid clearance mechanisms, in part by microbial anchoring to a tissue or tissue factor. (One possible exception is an organism that directly enters the bloodstream and multiplies there.) Because most host cells—responding to activation of innate immunity (see "Avoidance of Innate Host Defenses," below)—express multiple surface and cytoplasmic molecules that detect pathogens and pathogen factors, a complex interplay ensues and determines whether the microbe will avoid host clearance and remain in a tissue. Viruses and intracellular pathogens like *Mycobacterium tuberculosis* must bind to cells and enter them, whereas common extracellular bacterial pathogens of the human respiratory tract survive better if they avoid binding to pulmonary epithelial cells.

Specific ligands or adhesins for host receptors constitute a major area of study in microbial pathogenesis. Adhesins comprise a wide range of surface structures, anchoring the microbe to a tissue and promoting cellular entry as well as eliciting host responses critical to innate immunity (**Table 116-1**). Most microbes produce multiple adhesins specific for multiple host receptors that often are redundant, are serologically variable, and act additively or synergistically with other microbial factors to promote sticking to host tissues. In addition, some microbes adsorb host proteins onto their surface and use the natural host protein receptor for binding and entry into cells. While it is clear that, for some pathogenic organisms, blocking adherence can be a means to prevent infection, for others it could have unintended consequences, decreasing the innate host response that facilitates elimination of the infecting microbe.

TABLE 116-1

Examples of Microbial Ligand–Receptor Interactions

MICROORGANISM	TYPE OF MICROBIAL LIGAND	HOST RECEPTOR
Viral Pathogens		
Influenza virus	Hemagglutinin	Sialic acid
Measles virus		
Vaccine strain	Hemagglutinin	CD46/moesin/signaling lymphocytic activation molecule (SLAM)/nectin-4
Wild-type strains	Hemagglutinin	
Human herpesvirus type 6A	Glycoprotein complex gH/gL/gQ1/gQ2	CD46
Herpes simplex virus	Glycoprotein C	Heparan sulfate
HIV	Surface glycoprotein	CD4 and chemokine receptors (CCR5 and CXCR4)
Epstein-Barr virus	Envelope protein	CD21 (CR2)
Adenovirus	Fiber protein	Coxsackie-adenovirus receptor (CAR)
Coxsackievirus	Viral coat proteins	CAR and major histocompatibility class I antigens
Bacterial Pathogens		
Neisseria spp.	Pili	Membrane cofactor protein (CD46)
Pseudomonas aeruginosa	Pili and flagella Lipopolysaccharide	Asialo-GM1 Cystic fibrosis transmembrane conductance regulator (CFTR)
Escherichia coli	Pili	Ceramides/mannose and digalactosyl residues
Streptococcus pyogenes	Hyaluronic acid capsule	CD44
Yersinia spp.	Invasin/accessory invasin locus	β_1 Integrins
Bordetella pertussis	Filamentous hemagglutinin	CR3
Legionella pneumophila	Adsorbed C3bi	CR3
Mycobacterium tuberculosis	Adsorbed C3bi	CR3; DC-SIGN
Fungal Pathogens		
Blastomyces dermatitidis	WI-1	Possibly matrix proteins and integrins
Candida albicans	Int1p	Extracellular matrix proteins
Protozoal Pathogens		
Plasmodium vivax	Merozoite form	Duffy Fy antigen
Plasmodium falciparum	Erythrocyte-binding protein 175 (EBA-175)	Glycophorin A
Entamoeba histolytica	Surface lectin	*N*-Acetylglucosamine

Viral Adhesins

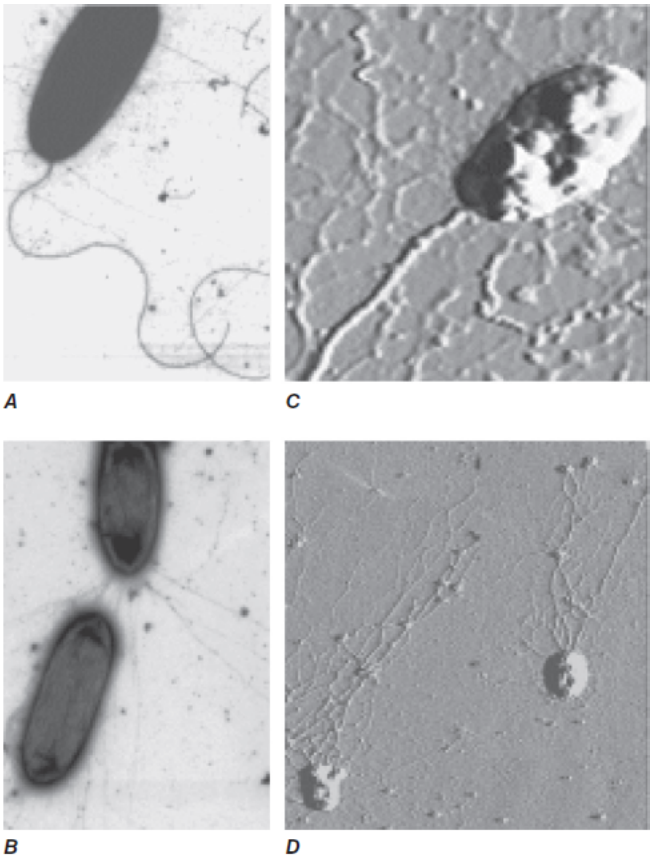
All viral pathogens must bind to host cells, enter them, and replicate within them. Viral coat proteins serve as ligands for cellular entry, and more than one ligand–receptor interaction may be needed. In some types of viruses, such as lipid bilayer–encapsulated Retroviridae or Rhabdoviridae, a single protein mediates both viral binding and entry via fusion with the host cell membrane. In other cases, a second viral fusion protein is needed to complete viral entry. HIV uses its envelope glycoprotein (gp) 120 to enter host cells by binding to both CD4 and one of two receptors for chemokines (CCR5 or CXCR4). Measles virus requires two proteins for cellular entry: the hemagglutinin (H) glycoprotein of wild-type measles virus binds to the signaling lymphocytic activation molecule (SLAM or CD150) on macrophages and dendritic cells, where the virus initially replicates, and also to nectin-4 on respiratory epithelial cells, where later replication occurs. The vaccine strain of measles virus binds to both CD46 and SLAM. For full cellular entry, however, measles virus requires a second fusion (F) protein. The gB and gC proteins on herpes simplex virus bind to heparan sulfate, although this adherence is not essential for entry but rather serves to concentrate virions close to the cell surface; this step is followed by attachment to mammalian cells mediated by the viral gD protein, with subsequent formation of a homotrimer of viral gB protein or a heterodimer of viral gH and gL proteins that permits fusion of the viral envelope with the host cell membrane. Herpes simplex virus can use a number of eukaryotic cell-surface receptors for entry, including the herpesvirus entry mediator, members of the immunoglobulin superfamily, the proteins nectin-1 and nectin-2, and modified heparan sulfate.

Bacterial Adhesins

Among the adhesins studied in greatest detail are bacterial pili and flagella **(Fig. 116-1)**. *Pili* or *fimbriae* are commonly used by gram-negative bacteria for attachment to host cells and tissues; similar factors are produced by gram-positive organisms such as group B streptococci. In electron micrographs, these hairlike projections (up to several hundred per cell) may be confined to one end of the organism (*polar pili*) or distributed more evenly over the surface. An individual cell may have pili with a variety of functions. Most pili are made up of a major pilin protein subunit (17,000–30,000 Da) that polymerizes to form the pilus. Many strains of *Escherichia coli* isolated from urinary tract infections express a mannose-binding type 1 pilus that attaches to the uroplakins coating the cells in the bladder epithelium. Other strains produce the Pap (pyelonephritis-associated) or P pilus adhesin that mediates binding to digalactose (gal-gal) residues on globosides of the human P blood groups. Both of these pili have proteins located at the tips of the main pilus unit that are critical to the binding specificity of the whole pilus unit. *E. coli* cells causing diarrheal disease express pilus-like receptors for enterocytes on the small bowel, along with other receptors termed *colonization factors*.

FIGURE 116-1

Bacterial surface structures. ***A*** and ***B.*** Traditional electron micrographic images of fixed cells of *Pseudomonas aeruginosa*. Flagella (***A***) and pili (***B***) project out from the bacterial poles. ***C*** and ***D.*** Atomic force microscopic image of live *P. aeruginosa* freshly planted onto a smooth mica surface. This technology reveals the fine, three-dimensional detail of the bacterial surface structures.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The type IV pili found in *Neisseria* species, *Moraxella* species, *Vibrio cholerae*, *Legionella pneumophila, Salmonella enterica* serovar Typhi, enteropathogenic *E. coli*, and *Pseudomonas aeruginosa* often mediate adherence of organisms to target surfaces. Type IV pili tend to have a relatively conserved amino-terminal region and a more variable carboxyl-terminal region. For some species (e.g., *N. gonorrhoeae, Neisseria meningitidis*, and enteropathogenic *E. coli*), the pili are critical for attachment to mucosal epithelial cells. For others, such as *P. aeruginosa*, the pili may inhibit colonization; recent studies of *P. aeruginosa* colonization showed that,

in a bank of mutants in which all nonessential genes were interrupted, those unable to produce type IVa pili were actually better able to colonize the gastrointestinal and lung mucosa of mice. *V. cholerae* cells appear to use two different types of pili for intestinal colonization. Whereas interference with this stage of colonization would appear to be an effective antibacterial strategy, attempts to develop pilus-based vaccines against human diseases have not been highly successful to date.

Flagella are long appendages attached at one or both ends of the bacterial cell (*polar flagella*) or distributed over the entire cell surface (*peritrichous flagella*). Flagella, like pili, are composed of a polymerized or aggregated basic protein. In flagella, the protein subunits form a tight helical structure and vary serologically with the species. Spirochetes such as *T. pallidum* and *Borrelia burgdorferi* have axial filaments similar to flagella running down the long axis of the center of the cell, and they "swim" by rotation around these filaments. Some bacteria can glide over a surface in the absence of obvious motility structures.

Other bacterial structures involved in adherence to host tissues include staphylococcal and streptococcal proteins that bind to human extracellular matrix proteins such as fibrin, fibronectin, fibrinogen, laminin, and collagen. Fibronectin is a commonly used receptor for various pathogens; a particular amino acid sequence in fibronectin, Arg-Gly-Asp or RGD, is a conserved target used by bacteria to bind to host tissues. Binding of the *Staphylococcus aureus* surface protein clumping factor A (ClfA) to fibrinogen has been implicated in many aspects of pathogenesis. The conserved outer-core portion of the lipopolysaccharide (LPS) of *P. aeruginosa* mediates binding to the cystic fibrosis transmembrane conductance regulator (CFTR) on airway epithelial cells—an event that appears to be critical for normal host resistance to infection, initiating recruitment of polymorphonuclear neutrophils (PMNs) to the lung mucosa to kill the cells via opsonophagocytosis. A large number of microbial pathogens encompassing major gram-positive bacteria (staphylococci and streptococci), gram-negative bacteria (major enteric species and coccobacilli), fungi (*Candida*, *Fusobacterium*, *Aspergillus*), and even eukaryotic pathogens (*Trichomonas vaginalis* and *Plasmodium falciparum*) express a surface polysaccharide composed of β-1-6-linked-poly-*N*-acetyl-d-glucosamine (PNAG). One of its functions is to promote binding to synthetic materials used in catheters and other types of implanted devices. This polysaccharide may be a critical factor in the establishment of device-related infections by pathogens such as staphylococci and *E. coli*. High-powered imaging techniques (e.g., atomic force microscopy) have revealed that bacterial cells have a nonhomogeneous surface that is probably attributable to different concentrations of cell surface molecules, including microbial adhesins, at specific locations (Fig. 116-1, panels C and D).

Fungal Adhesins

Fungi produce adhesins that mediate colonization of epithelial surfaces, adhering particularly to structures like fibronectin, laminin, and collagen. The *Candida albicans* INT1 protein bears similarity to mammalian integrins that bind to extracellular matrix proteins. The agglutinin-like sequence (ALS) adhesins are large cell-surface glycoproteins mediating adherence of pathogenic *Candida* to host tissues. These adhesins possess a conserved three-domain structure composed of an N-terminus that mediates adherence to host tissue receptors, a central motif consisting of a number of repeats of a conserved sequence of 36 amino acids, and a C-terminal domain that varies in length and sequence and contains a glycosylphosphatidylinositol (GPI) anchor addition that allows the adhesins to bind to the fungal cell wall. Variability in the number of central domains characterizes different ALS proteins with specificity for different host receptors. The ALS adhesins are expressed under certain environmental conditions and are crucial for pathogenesis of fungal infections.

For several respiratory fungal pathogens, the inoculum is ingested by alveolar macrophages in which the fungal cells transform to pathogenic phenotypes. Like *C. albicans*, *Blastomyces dermatitidis* produces a 120-kDa surface protein, designated WI-1, that binds to CD11b/CD18 integrins as well as to CD14 on macrophages. An unidentified factor on *Histoplasma capsulatum* also mediates binding to the integrin surface proteins.

Eukaryotic Pathogen Adhesins

Eukaryotic parasites use complicated surface glycoproteins as adhesins, some of which are lectins (proteins that bind to specific carbohydrates on host cells). *Plasmodium vivax*, one of six *Plasmodium* species causing malaria, binds (via Duffy-binding protein) to the Duffy blood group carbohydrate antigen Fy on erythrocytes. *Entamoeba histolytica*, the third leading cause of death from parasitic diseases, expresses two proteins that bind to the disaccharide galactose/*N*-acetyl galactosamine. Children with mucosal IgA antibody to one of these lectins are resistant to reinfection with virulent *E. histolytica*. A major surface glycoprotein (gp63) of *Leishmania* promastigotes is needed for these parasites to enter human macrophages—the principal target cell of infection. This glycoprotein promotes complement binding but inhibits complement lytic activity, allowing the parasite to use complement receptors for entry into macrophages; gp63 also binds to fibronectin receptors on macrophages. As part of hepatic granuloma formation, *Schistosoma mansoni* expresses a carbohydrate epitope related to the Lewis X blood group antigen that promotes adherence of helminthic eggs to vascular endothelial cells under inflammatory conditions.

Host Receptors

Host receptors are found both on target cells (such as epithelial cells lining mucosal surfaces) and within the mucus layer covering these cells. Microbial pathogens bind to a wide range of host receptors to establish infection (Table 116-1). Selective loss of host receptors for a pathogen may confer natural resistance to an otherwise susceptible population. For example, 70% of individuals in western Africa lack Fy antigens and are resistant to *P. vivax* infection. *S. enterica* serovar Typhi, the etiologic agent of typhoid fever, produces a pilus protein that binds to CFTR to enter the gastrointestinal submucosa after being ingested by enterocytes. As homozygous mutations in *CFTR* are the cause of the life-shortening disease cystic fibrosis, heterozygote carriers (e.g., 4–5% of individuals of European ancestry) may have had a selective advantage due to decreased susceptibility to typhoid fever.

Numerous virus–target cell interactions have been described, and it is now clear that different viruses can use similar host cell receptors for entry. The list of certain and likely host receptors for viral pathogens is long. Among the host membrane components that can serve as receptors for viruses are sialic acids, gangliosides, glycosaminoglycans, integrins and other members of the immunoglobulin superfamily, histocompatibility antigens, and regulators and receptors for complement components. An example of the effect of host receptors on the pathogenesis of infection has emerged from studies comparing the binding of avian influenza A virus subtype H5N1 with that of influenza A strains expressing the H1 hemagglutinin subtype. These subtypes are highly pathogenic and transmissible from human to human, and they bind to a receptor composed of two sugar molecules: sialic acid linked α-2-6 to galactose. This receptor is expressed at high levels in the human airway epithelium; when virus is shed from this surface, its transmission via coughing and aerosol droplets is facilitated. In contrast, the H5N1 avian influenza virus

binds to sialic acid linked α-2-3 to galactose, and this receptor is expressed at high levels on cells in the terminal bronchioles, including type II pneumocytes, alveolar macrophages, and nonciliated cuboidal epithelial cells. Infection at these sites is thought to underlie the high mortality rate associated with avian influenza but also the low interhuman transmissibility of this strain, which is not readily transported to the airways from which it can be expelled by coughing. Nonetheless, it has been shown that H5 hemagglutinins can acquire mutations leading to binding to α-2-6-linked sialic acids that increase their human transmissibility but retain their high level of lethality.

MICROBIAL GROWTH AFTER ENTRY

Once established on a mucosal or skin site, pathogenic microbes must replicate before causing full-blown infection and disease. Within cells, viral particles release their nucleic acids, which may be directly translated into viral proteins (positive-strand RNA viruses), transcribed from a negative strand of RNA into a complementary mRNA (negative-strand RNA viruses), or transcribed into a complementary strand of DNA (retroviruses). For DNA viruses, mRNA may be transcribed directly from viral DNA, either in the cell nucleus or in the cytoplasm. To grow, bacteria must acquire specific nutrients or synthesize them from precursors in host tissues. Many infectious processes are most often found in specific sites—e.g., H1 influenza in the respiratory mucosa, gonorrhea in the urogenital epithelium, and shigellosis in the gastrointestinal epithelium. While there are multiple reasons for this specificity, one important consideration is the ability of these pathogens to obtain the nutrients needed for growth and survival.

Temperature restrictions also play a role in limiting certain pathogens to specific tissues. Rhinoviruses, a cause of the common cold, grow best at 33°C and replicate in cooler nasal tissues but not in the lung. Leprosy lesions due to *Mycobacterium leprae* are found in and on relatively cool body sites. Fungal pathogens that infect the skin, hair follicles, and nails (dermatophyte infections) remain confined to the cooler, exterior, keratinous layer of the epithelium.

A topic of major interest is the ability of many bacterial, fungal, and protozoal species to grow in multicellular masses referred to as *biofilms*. These masses are biochemically and morphologically quite distinct from the free-living individual cells referred to as *planktonic cells*. Growth in biofilms leads to altered microbial metabolism, production of extracellular virulence factors, and decreased susceptibility to biocides, antimicrobial agents, and host defense molecules and cells. *P. aeruginosa* growing on the bronchial mucosa during chronic infection, staphylococci and other pathogens growing on implanted medical devices, and dental pathogens growing on tooth surfaces to form plaques represent several examples of microbial biofilm growth associated with human disease. Many other pathogens can form biofilms during in vitro growth. This mode of growth contributes to microbial virulence and induction of disease and can also be an important factor in microbial survival outside the host, promoting transmission to additional susceptible individuals.

AVOIDANCE OF INNATE HOST DEFENSES

Microbes have interacted with mucosal/epithelial surfaces since the emergence of multicellular organisms. Thus it is not surprising that multicellular hosts have a variety of innate surface defense mechanisms that can sense when pathogens are present and contribute to their elimination. The skin is acidic and bathed with fatty acids toxic to many microbes. Skin pathogens such as staphylococci must tolerate these adverse conditions. Mucosal surfaces are covered by a barrier composed of a thick mucus layer that entraps microbes and facilitates their transport out of the body by mucociliary clearance, coughing, and urination. Mucous secretions, saliva, and tears contain antibacterial factors such as lysozyme and antimicrobial peptides as well as antiviral factors such as interferons (IFNs). Gastric acidity and bile salts are inimical to the survival of many ingested organisms, and most mucosal surfaces—particularly the nasopharynx, vaginal tract, and gastrointestinal tract—contain a resident flora of commensal microbes that interfere with the ability of pathogens to colonize and infect a host. Major advances in the use of nucleic acid sequencing now allow extensive identification and characterization of the vast array of commensal organisms that have come to be referred to as the *microbiota*. In addition to its role in providing competition for mucosal colonization, acquisition of a normal microbiota is critical for proper development of the immune system, impacting maturation and differentiation of components of both the innate and acquired immune systems.

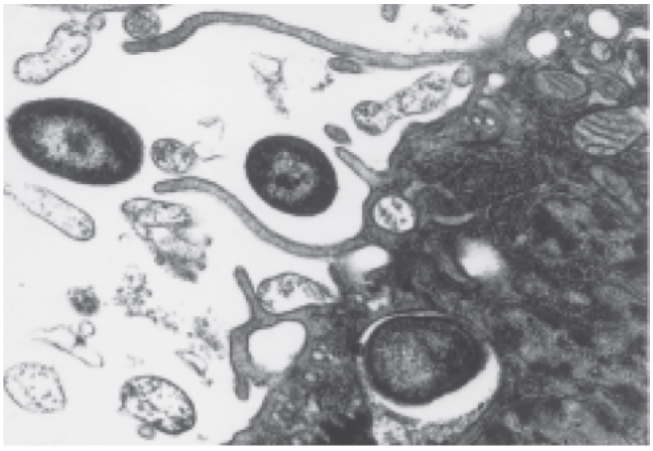
Pathogens that survive local antimicrobial factors must still contend with host endocytic, phagocytic, and inflammatory responses as well as with host genetic factors that determine the degree to which a pathogen can survive and grow. The growth of viral pathogens entering skin or mucosal epithelial cells can be limited by a variety of host genetic factors, including production of IFNs, modulation of receptors for viral entry, and age- and hormone-related susceptibility factors; by nutritional status; and even by personal habits such as smoking and exercise. The list of genes whose variants can affect host susceptibility and resistance to infection is rapidly expanding. A classic example is a 32-bp deletion in the gene for the HIV-1 co-receptor known as chemokine receptor 5 (CCR5), which, when present in the homozygous state, confers high-level resistance to HIV-1 infection. A now-famous case is that of the “Berlin Patient,” a man infected with HIV who received a hematopoietic stem-cell transplant from a donor homozygous for the 32-bp CCR5 deletion to treat acute myeloid leukemia. The apparent sterilizing cure of this patient’s HIV infection is likely due to his having only HIV-resistant T cells after the stem cell transplantation.

Encounters with Epithelial Cells

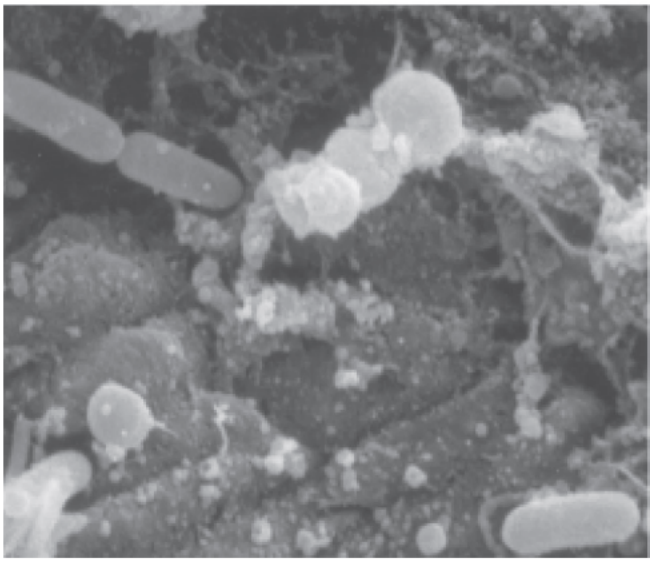
Over the past two decades, many pathogens have been shown to enter epithelial cells **(Fig. 116-2)** by using specialized surface structures that bind to receptors, with consequent internalization. However, the exact role and the importance of this process in infection and disease are not well defined for most of these pathogens. Microbial entry into host epithelial cells is seen as a path for translocation to adjacent or deeper tissues or as a route to a sanctuary site to avoid killing by professional phagocytes. Epithelial cell entry is a critical aspect of dysentery induction by *Shigella*.

FIGURE 116-2

Entry of bacteria into epithelial cells. ***A.*** Internalization of *Pseudomonas aeruginosa* by cultured airway epithelial cells expressing wild-type cystic fibrosis transmembrane conductance regulator, the cell receptor for bacterial ingestion. ***B.*** Entry of *P. aeruginosa* into murine tracheal epithelial cells after infection of mice by the intranasal route.



A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Curiously, less virulent strains of many bacterial pathogens are more adept at entering epithelial cells than are more virulent strains; examples include pathogens that lack the surface polysaccharide capsule needed to cause serious disease. Thus, for *Haemophilus influenzae*, *Streptococcus pneumoniae*, *Streptococcus agalactiae* (group B *Streptococcus*), *N. meningitidis*, and *S. pyogenes*, isogenic mutants or variants lacking capsules enter epithelial cells more easily than the wild-type, encapsulated parental forms that cause disseminated disease. These observations have led to the proposal that epithelial cell entry may be primarily a manifestation of host defense, resulting in bacterial clearance by both shedding of epithelial cells containing internalized bacteria and initiation of a protective and nonpathogenic inflammatory response. However, a possible consequence of this process could be the opening of a hole in the epithelium, potentially allowing uningested organisms to enter the submucosa. This scenario has been documented in murine *S. enterica* serovar Typhimurium infections and in experimental bladder infections caused by uropathogenic *E. coli*. In the latter system, bacterial pilus-mediated attachment to *uroplakins* induces exfoliation of the cells with attached bacteria. Subsequently, infection is produced by residual bacterial cells that invade the superficial bladder epithelium, where they can grow intracellularly into biofilm-like masses encased in an extracellular polysaccharide-rich matrix and surrounded by uroplakin. It is likely that at low bacterial inocula epithelial cell ingestion and subclinical inflammation are efficient means to eliminate pathogens, whereas at higher inocula a proportion of surviving bacterial cells enter host tissue through the damaged mucosal surface and multiply, producing disease. Alternatively, failure of the appropriate epithelial cell response to a pathogen may allow the organism to survive on a mucosal surface where, if it avoids other host defenses, it can grow and cause a local infection. Along these lines, as noted above, *P. aeruginosa* is taken into epithelial cells by CFTR, a protein missing or nonfunctional in most patients with severe cases of cystic fibrosis. The major clinical consequence of this disease is chronic airway-surface infection with *P. aeruginosa* in 80–90% of patients. The failure of airway epithelial cells to ingest and promote the removal of *P. aeruginosa* via a properly regulated inflammatory response has been proposed as a key component of the hypersusceptibility of these patients to chronic airway infection with this organism.

Encounters with Phagocytes

Phagocytosis and Inflammation

Phagocytosis of microbes is a major innate host defense that limits the growth and spread of pathogens. Phagocytes appear rapidly at sites of infection in conjunction with the initiation of inflammation. Ingestion of microbes by both tissue-fixed macrophages and migrating phagocytes probably accounts for the limited ability of most microbial agents to cause disease. A family of related molecules called *collectins*, *soluble defense collagens*, or *pattern-recognition molecules* are found in blood (mannose-binding lectins), lung (surfactant proteins A and D), and most likely other tissues and bind to carbohydrates on microbial surfaces to

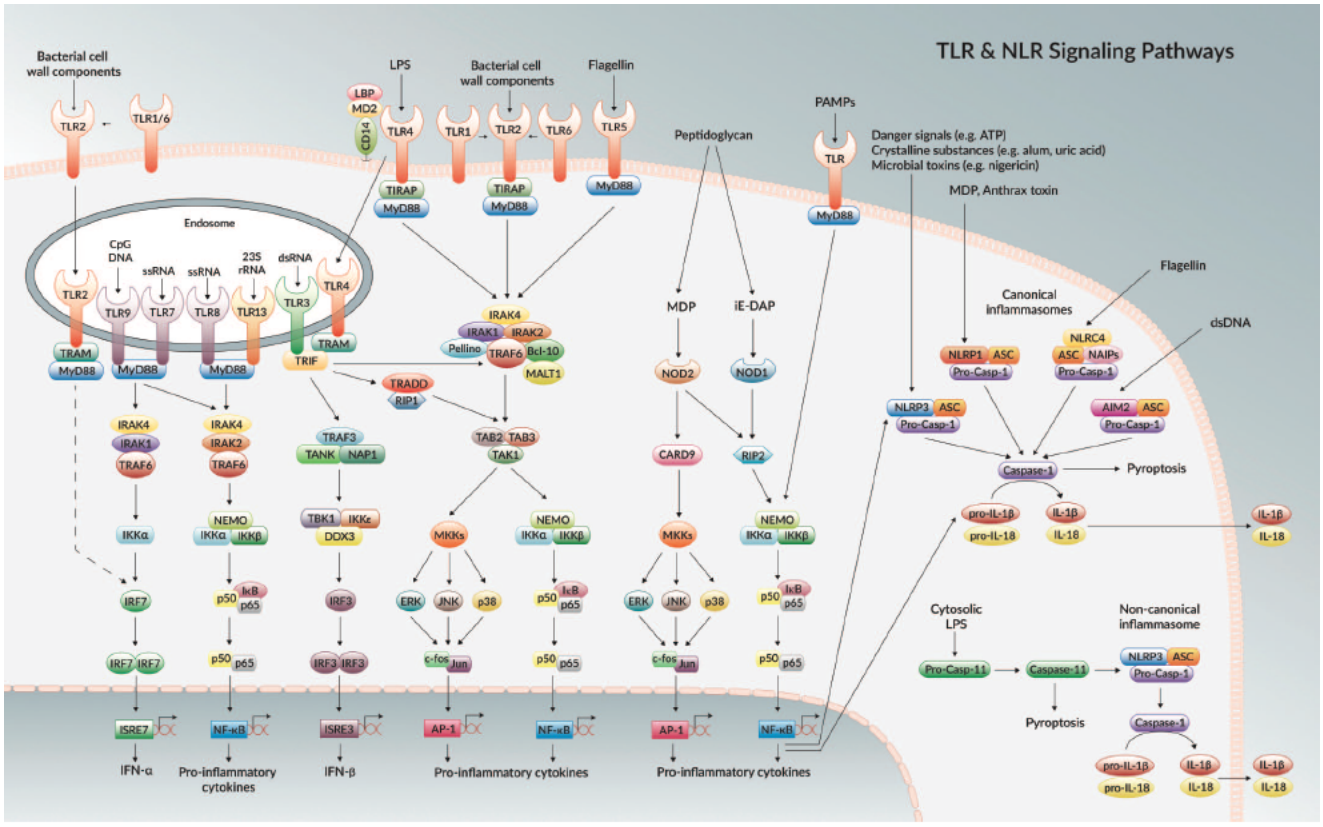
promote phagocyte clearance. Bacterial pathogens are ingested principally by PMNs, while eosinophils are frequently found at sites of infection by protozoan or multicellular parasites. Successful pathogens, by definition, must avoid being cleared by professional phagocytes. One of several antiphagocytic strategies employed by bacteria and by the fungal pathogen *Cryptococcus neoformans* is to elaborate large-molecular-weight surface polysaccharide antigens, often in the form of a capsule that coats the cell surface. Most pathogenic bacteria produce such antiphagocytic capsules. On occasion, proteins or polypeptides form capsule-like coatings for organisms such as group A streptococci and *Bacillus anthracis*.

An area of both intense interest and controversy is the role of the release of neutrophil extracellular traps (NETs) in protection against infection. NETs are composed of DNA and other intracellular components with antimicrobial properties, such as histones, myeloperoxidase, and elastase. NET release has been described as both a "suicidal" event, wherein, in response to stimuli, PMNs lyse and release NET components, and a "vital" event, wherein intracellular NET components are released but neutrophils remain viable and functional. Microbial particle size might regulate release of NETs, as has been reported for larger microbial structures like *C. albicans* hyphae or cellular aggregates. NET formation can also be pathologic as these networks are associated with damage to cells, thrombosis, inhibition of wound healing, and autoimmunity.

As activation of local phagocytes in tissues is a key step in initiating inflammation and migration of additional phagocytes into infected sites, much attention has been paid to microbial factors that initiate inflammation. These are usually conserved factors critical to the microbes' survival and are referred to as *pathogen-associated molecular patterns* (PAMPs). Cellular responses to microbial encounters with phagocytes are governed largely by the structure of the microbial PAMPs that elicit inflammation, and detailed knowledge of these structures of bacterial pathogens has contributed greatly to our understanding of molecular mechanisms of microbial pathogenesis mediated by activation of host cell molecules such as Toll-like receptors (TLRs; **Fig. 116-3**). One of the best-studied systems involves the interaction of LPS from gram-negative bacteria with the GPI-anchored membrane protein CD14 found on the surface of professional phagocytes, including migrating and tissue-fixed macrophages and PMNs. A soluble form of CD14 is also found in plasma and on mucosal surfaces. A plasma protein, LPS-binding protein, transfers LPS to membrane-bound CD14 on myeloid cells and promotes binding of LPS to soluble CD14. Soluble CD14/LPS/LPS-binding protein complexes bind to many cell types and may be internalized to initiate cellular responses to microbial pathogens. It has been shown that peptidoglycan and lipoteichoic acid from gram-positive bacteria and cell surface products of mycobacteria and spirochetes can interact with CD14 (Fig. 116-3). Additional molecules, such as MD-2, also participate in the recognition of bacterial activators of inflammation.

FIGURE 116-3

Toll-like receptor (TLR) and NOD-like receptor (NLR) signaling pathways. Microbial cell-surface constituents interact with TLRs, in some cases requiring additional factors such as MD2, which facilitates the response to lipopolysaccharide (LPS) via TLR4. Although microbial cell-surface constituents are depicted as interacting with the TLRs on the cell surface, TLRs contain extracellular leucine-rich domains that become localized to the lumen of the phagosome upon uptake of bacterial cells. The internalized TLRs can bind to microbial products. The TLRs are oligomerized, usually forming homodimers, and then bind to the general adapter protein MyD88 via the C-terminal Toll/interleukin 1 receptor (IL-1R) (TIR) domains, which also bind to TIRAP (TIR domain–containing adapter protein), a molecule that participates in the transduction of signals from TLRs 1, 2, 4, and 6. The MyD88/TIRAP complex activates signal-transducing molecules such as IRAK4 (IL-1R-associated kinase 4), which in turn activates IRAK1. This activation can be blocked by IRAKM and Toll interacting protein (TOLLIP). IRAK1 activates TRAF6 (tumor necrosis factor receptor–associated factor 6), TAK1 (transforming growth factor β–activating kinase 1), and TAB1/2 (TAK1-binding protein 1/2). This signaling complex associates with the ubiquitin-conjugating enzyme Ubc13 and the Ubc-like protein UEV1A to catalyze the formation of a polyubiquitin chain on TRAF6. Polyubiquitination of TRAF6 activates TAK1, which, along with TAB1/2 (a protein that binds to lysine residue 63 in polyubiquitin chains via a conserved zinc-finger domain), phosphorylates the inducible kinase complex: IKKα, -β, and -γ. IKKγ is also called NEMO [nuclear factor κB (NF-κB) essential modulator]. This large complex phosphorylates the inhibitory component of NF-κB, IκBα, resulting in release of IκBα from NF-κB. Phosphorylated (PP) IκB is then ubiquitinated (ub) and degraded, and the two components of NF-κB—p50 (or Rel) and p65—translocate to the nucleus, where they bind to regulatory transcriptional sites on target genes, many of which encode inflammatory proteins. In addition to inducing NF-κB nuclear translocation, the TAK1/TAB1/2 complex activates MAP kinase transducers such as MKK 4/7 and MKK 3/6, an event that can lead to nuclear translocation of transcription factors such as AP1. TLR4 can also activate NF-κB nuclear translocation via the MyD88-independent TRIF [TIR domain–containing, adapter-inducing interferon β (IFN-β)] and TRAM (TRIF-related adapter molecule) cofactors. Intracellular TLRs 3, 7, 8, and 9 also use MyD88 and TRIF to activate IFN response factors 3 and 7 (IRF3 and IRF7), which also function as transcriptional factors in the nucleus. The nucleotide-binding oligomerization domain–like receptor (NLR) family of proteins is involved in the regulation of innate immune responses. These proteins sense pathogen-associated molecular patterns (PAMPs) in the cytosol as well as the host-derived signals known as *damage-associated molecular patterns* (DAMPs). Certain NLRs induce the assembly of large caspase 1–activating complexes called *inflammasomes*. Activation of caspase 1 through autoproteolytic maturation leads to the processing and secretion of the proinflammatory cytokines interleukin 1β (IL-1β) and IL-18. So far, four inflammasomes have been identified and defined by the NLR protein they contain: the NLRP1/NALP1b inflammasome, the NLRC4/IPAF inflammasome, the NLRP3/NALP3 inflammasome, and the AIM2 (absent in melanoma 2)–containing inflammasome. (*Pathway diagram reproduced with permission from InvivoGen; www.invivogen.com/review-inflammasome.*)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GPI-anchored receptors do not have intracellular signaling domains; therefore, it is the TLRs that transduce signals for cellular activation due to LPS binding. Binding of microbial factors to TLRs to activate signal transduction occurs in the phagosome—and not on the surface—of dendritic cells that have internalized the microbe. This binding is probably due to the release of the microbial surface factor from the cell in the environment of the phagosome, where the liberated factor can bind to its cognate TLRs. TLRs initiate cellular activation through a series of signal-transducing molecules (Fig. 116-3) that lead to nuclear translocation of the transcription factor nuclear factor κB (NF-κB), a master-switch for production of important inflammatory cytokines such as tumor necrosis factor α (TNF-α) and interleukin 1 (IL-1).

The initiation of inflammation can also occur with viral particles and other microbial products such as polysaccharides, enzymes, and toxins. Bacterial flagella activate inflammation by binding of a conserved sequence to TLR5. Some pathogens (e.g., *Campylobacter jejuni, Helicobacter pylori,* and *Bartonella bacilliformis*) make flagella that lack this sequence and do not bind to TLR5; thus efficient host responses to infection are prevented. Bacteria also produce a high proportion of DNA molecules with unmethylated CpG residues that activate inflammation through TLR9. TLR3 recognizes double-strand RNA, a pattern-recognition molecule produced by many viruses during their replicative cycle. TLR1 and TLR6 associate with TLR2 to promote recognition of acylated microbial proteins and peptides.

The myeloid differentiation factor 88 (MyD88) molecule and the Toll/IL-1R (TIR) domain–containing adapter protein (TIRAP) bind to the cytoplasmic domains of TLRs and also to receptors that are part of the IL-1 receptor families. Numerous studies have shown that MyD88/TIRAP-mediated transduction of signals from TLRs and other receptors is critical for innate resistance to infection, activating MAP kinases and NF-κB and thereby leading to production of cytokines/chemokines. Mice lacking MyD88 are more susceptible than normal mice to infections with a broad range of pathogens. In one study, nine children homozygous for defective MyD88 genes had recurrent infections with *S. pneumoniae*, *S. aureus*, and *P. aeruginosa*—three bacterial species showing increased virulence in MyD88-deficient mice. The MyD88-deficient children seemed to have no greater susceptibility to other bacteria, viruses, fungi, or parasites. Another component of the MyD88-dependent signaling pathway is a molecule known as IL-1 receptor–associated kinase 4 (IRAK4). Individuals with a homozygous deficiency in genes encoding this protein are at increased risk for *S. pneumoniae* and *S. aureus* infections and, to some degree, *P. aeruginosa* infections as well.

Some TLRs (e.g., TLR3 and TLR4) can also activate signal transduction via a MyD88-independent pathway involving TIR domain–containing, adapter-inducing IFN-β (TRIF) and the TRIF-related adapter molecule (TRAM). Signaling through TRIF and TRAM activates the production of both NF-κB-dependent cytokines/chemokines and type 1 IFNs. The type 1 IFNs bind to the IFN-α receptor composed of two protein chains, IFNAR1 and IFNAR2. Humans produce three type 1 IFNs: IFN-α, IFN-β, and IFN-γ. These molecules activate another class of proteins known as signal transducer and activator of transcription (STAT) complexes. The STAT factors are important in regulating immune system genes and thus play a critical role in responses to microbial infections.

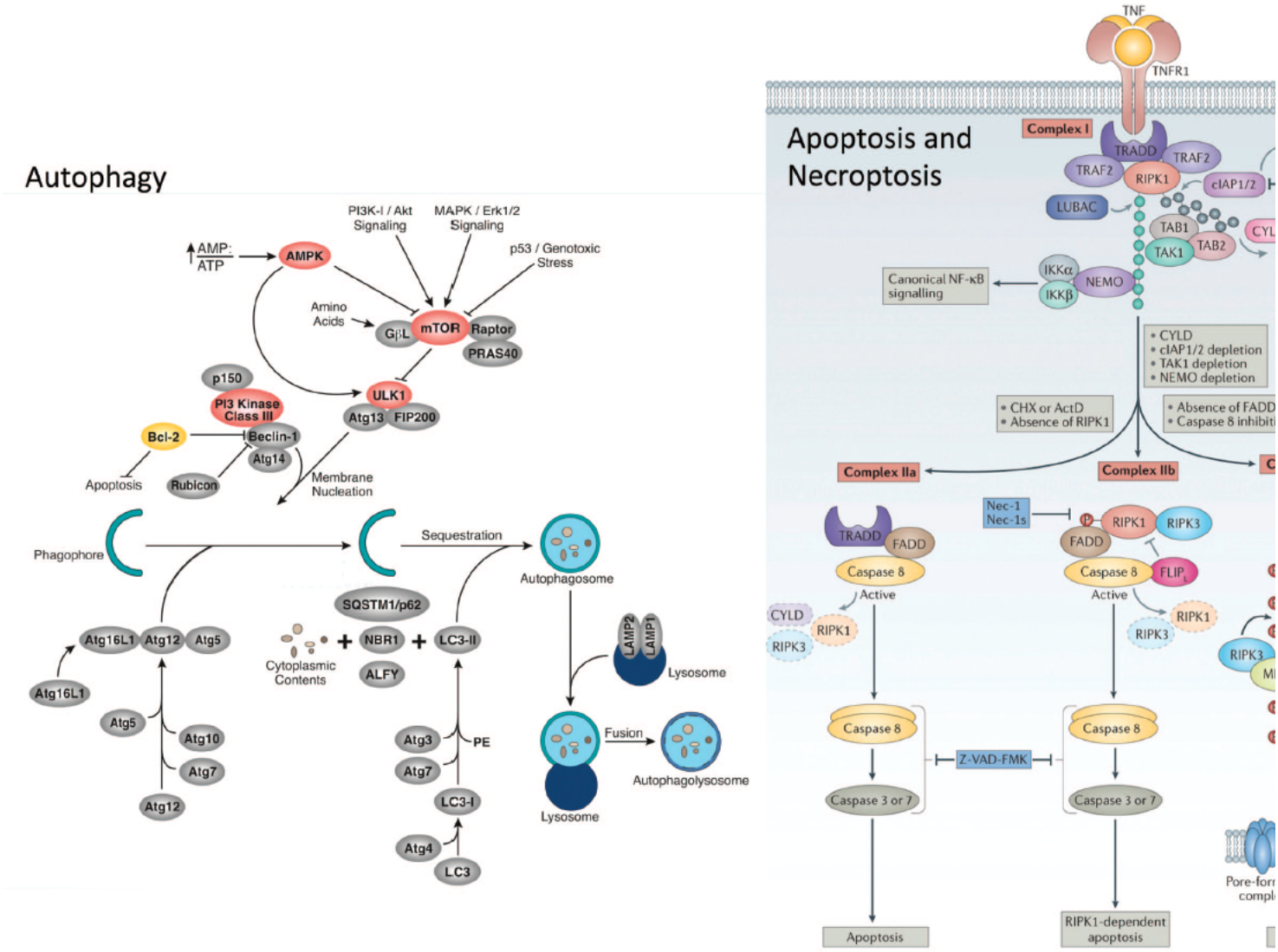
Another intracellular complex of proteins found to be a major factor in the host cell response to infection is the *inflammasome* (Fig. 116-3), where inflammatory cytokines IL-1 and IL-18 are changed from their precursors to active forms by the cysteine protease caspase 1 and then secreted. The inflammasome is composed of proteins that are members of the nucleotide binding and oligomerization domain (NOD)–like receptor (NLR) family. Like the TLRs, NOD proteins within inflammasomes sense the presence of the conserved microbial factors either internalized from outside the cell to form canonical inflammasomes or released inside a cell (probably after microbial uptake) to directly activate pro-caspase-2. Recognition of these PAMPs by NLRs leads to caspase 1 or caspase 2 activation and to secretion of active IL-1 and IL-18. Studies of mice indicate that as many as four canonical inflammasomes with different components can be formed (Fig. 116-3). The

components depend on the type of stimulus driving inflammasome formation and activation. A fifth, noncanonical inflammasome responding to intracellular LPS also has been described (Fig. 116-3).

Some recent additions to the identified intracellular components responding to microbial infection are autophagy **(Fig. 116-4)**, initially described as an intracellular process for degradation and recycling of cellular components for reuse, and several pathways leading to cell death, including apoptosis, RIPK1-dependent apoptosis, and necroptosis (Fig. 116-4). The latter three pathways are means by which cells undergo a death program in response to infection (notably, viral infection) and inflammation. Autophagy is an early defense mechanism mediated by caspases in response to pathogens wherein, after ingestion, microbes in either vacuoles or the cytoplasm are delivered to lysosomal compartments for degradation. Avoidance of this process is critical if pathogens are to cause disease. Pathogens can avoid autophagy by multiple mechanisms; examples include the inhibition of proteins within the autophagic vacuole by *Shigella*, the recruitment of host proteins to prevent autophagy of *Listeria monocytogenes*, and the inhibition of vacuole formation by *L. pneumophila*. In the death pathways, cell death to inhibit viral replication (Fig. 116-4) is mediated by a series of reactions commencing with TNF-α production and binding of this molecule to its receptor, TNFR1. In the two apoptotic pathways, the final steps are activation of effector caspases 3 and 7 and apoptotic cell death. In necroptosis, oligomers of the mixed-lineage kinase domain–like protein (MLKL) form and insert into the cell's plasma membrane; their insertion leads to lysis and release of damage-associated molecular patterns (DAMPs), resulting in protective innate immune responses. Finally, additional pathways of cell death are being described, including ferroptosis, oxytosis, parthanatos, pyronecrosis, and pyroptosis. The impact of these pathways on host–pathogen interactions is only beginning to be investigated.

FIGURE 116-4

Autophagy, apoptosis, and necroptosis. ***A.*** *Autophagy* is a catabolic process that results in the autophagosomic–lysosomal degradation of bulk cytoplasmic contents, abnormal protein aggregates, and excess or damaged organelles. Autophagy is generally activated by conditions of nutrient deprivation but has also been associated with physiologic as well as pathologic processes such as development, differentiation, neurodegenerative disease, stress, infection, and cancer. The kinase mTOR is a critical regulator of autophagy induction, with activated mTOR (Akt and MAPK signaling) suppressing autophagy and negative regulation of mTOR (AMPK and p53 signaling) promoting it. Three related serine/threonine kinases—UNC-51-like kinases 1, 2, and 3 (ULK1, ULK2, ULK3), which play a role similar to that played by the yeast Atg1—act downstream of the mTOR complex. ULK1 and ULK2 form a large complex with the mammalian homolog of an autophagy-related (Atg) gene product (mAtg13) and the scaffold protein FIP200 (an ortholog of yeast Atg17). The class III PI3K complex, containing hVps34, beclin-1 (a mammalian homolog of yeast Atg6), p150 (a mammalian homolog of yeast Vps15), and Atg14-like protein (Atg14L or Barkor) or ultraviolet irradiation resistance–associated gene (UVRAG), is required for the induction of autophagy. The Atg genes control autophagosome formation through Atg12-Atg5 and LC3-II (Atg8-II) complexes. Atg12 is conjugated to Atg5 in a ubiquitin-like reaction that requires Atg7 and Atg10 (E1- and E2-like enzymes, respectively). The Atg12-Atg5 conjugate then interacts noncovalently with Atg16 to form a large complex. LC3/Atg8 is cleaved at its C-terminus by Atg4 protease to generate the cytosolic LC3-I. LC3-I is conjugated to phosphatidylethanolamine (PE), also in a ubiquitin-like reaction that requires Atg7 and Atg3 (E1- and E2-like enzymes, respectively). The lipidated form of LC3, known as LC3-II, is attached to the autophagosome membrane. Autophagy and apoptosis are connected both positively and negatively, and extensive crosstalk exists between the two processes. During nutrient deficiency, autophagy functions as a pro-survival mechanism; however, excessive autophagy may lead to cell death, a process morphologically distinct from apoptosis. Several pro-apoptotic signals, such as tumor necrosis factor (TNF), TNF-related apoptosis-inducing ligand (TRAIL), and FAS-associated death domain (FADD), also induce autophagy. Moreover, Bcl-2 inhibits beclin-1-dependent autophagy, thereby functioning as both a pro-survival and an anti-autophagic regulator. *Mitophagy* is a selective autophagic process specifically designed for the removal of damaged or unneeded mitochondria from a cell. Upon mitochondrial damage, the protein PINK, which is continually degraded in the healthy state through the action of PARL, is stabilized and recruits the E3 ligase Parkin to initiate mitophagy. Polyubiquitination of mitochondrial membrane proteins by Parkin results in the recruitment of autophagy adaptor proteins SQSTM1/p62, NBR1, and Ambra1, which bind to LC3 via their LC3-interacting region (LIR). In addition, BNIP3 and BNIP3L/NIX, which also contain LIRs, directly recruit autophagic machinery by a ubiquitin-independent mechanism to induce autophagosome formation in certain cell types. *[Reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com).]* ***B.*** Apoptosis and necroptosis are initiated by the binding of TNF to its cognate receptor TNFR1, triggering the assembly of complex I, which comprises TNFR1, TNFR1-associated death domain (TRADD), receptor-interacting serine/threonine-protein kinase 1 (RIPK1), TNFR-associated factor 2 (TRAF2), cellular inhibitor of apoptosis protein 1/2 (cIAP1/2), and linear ubiquitin chain assembly complex (LUBAC). Complex I provides the platform for Lys63-linked ubiquitylation (gray circles) or linear ubiquitylation (green circles) of RIPK1 by cIAP1/2 and LUBAC, respectively. This ubiquitylation is implicated in the decision between NF-κB or survival signaling and cell death signaling. Ubiquitylation leads to the recruitment of other factors, such as transforming growth factor β–activated kinase (TAK1), TAK1-binding protein 1 (TAB1), TAB2, NF-κB essential modulator (NEMO), and the inhibitor of the NF-κB kinase-α (IKKα)–IKKβ complex; this recruitment usually triggers canonical NF-κB signaling. In the presence of the translational inhibitor cycloheximide (CHX), TNF stimulation leads to the formation of cytosolic complex IIa, in which RIPK1 disappears, whereas interaction of TRADD and FADD leads to the activation of caspase 8 and effector caspases (e.g., caspase 3 and caspase 7) and to apoptotic cell death. With cIAP1/2 inhibitors [second mitochondria-derived activator of caspase (SMAC) mimetics], knockdown of cIAPs, or inhibition or depletion of TAK1 or NEMO, complex IIb is formed. Complex IIb consists of RIPK1, RIPK3, FADD, and caspase 8 and favors RIPK1-kinase activity–dependent apoptosis. Heteromeric pro-caspase 8–FLICE-like inhibitory protein long isoform ($FLIP_L$) inhibits necroptosis, probably by cleaving RIPK1, RIPK3, and cylindromatosis (CYLD). With sufficient expression of RIPK3 and concomitant inhibition or reduced expression of pro-caspase 8 and $FLIP_L$, complex IIc (also known as the necrosome) is formed. The formation of complex IIc entails the association of RIPK1 and RIPK3 followed by a series of auto- and transphosphorylation events of RIPK1 and RIPK3. Activated RIPK3 phosphorylates and recruits mixed-lineage kinase domain–like protein (MLKL), eventually leading to the formation of a supramolecular protein complex at the plasma membrane and necroptosis. SMAC mimetics are being developed to impair survival signaling and to sensitize cells to the triggering of cell death in the context of tumor treatment. Z-VAD-FMK is a bona fide pan-caspase inhibitor. Necrostatin-1 (Nec-1) and the more specific Nec-1s, Cpd27, and (more recently) PN10—a hybrid molecule consisting of Nec-1s and ponatinib—all inhibit the kinase activity of RIPK1 and thus inhibit necroptotic signaling. Additional necroptosis inhibitors include the RIPK3 inhibitors GSK'840, GSK'843 and GSK'872 as well as the MLKL inhibitors necrosulfonamide (NSA) and compound 1. However, the specificity of these MLKL inhibitors is not restricted to inhibition of MLKL, and their efficacy is probably also due to effects on other steps in the necroptosis pathway. ActD, actinomycin D. *(Reprinted by permission from Macmillan Publishers Ltd: M Conrad et al: Regulated necrosis: Disease relevance and therapeutic opportunities. Nature Reviews Drug Discovery 15:348, copyright 2016.)*



A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Additional Interactions of Microbial Pathogens and Phagocytes

Other ways that microbial pathogens avoid destruction by phagocytes include production of factors that are toxic to these cells or that interfere with their chemotactic and ingestion function. Hemolysins, leukocidins, and the like are microbial proteins that can kill phagocytes. *S. aureus* elaborates a family of bi-component leukocidins that bind to host receptors such as the HIV co-receptor CCR5, which is also a receptor for the LukE/D toxin, and the receptor of the C5a component of activated complement used by LukF/S, also known as the *Panton-Valentine leukocidin*. The cytolytic staphylococcal α hemolysin binds to the disintegrin and metalloprotease 10 (ADAM-10) protein expressed on a variety of cells and also activates the NLRP3 inflammasome in monocytic cells, with consequent production of inflammatory cytokines as well as cell death. Streptolysin O made by *S. pyogenes* binds to cholesterol in phagocyte membranes and initiates a process of internal degranulation, with the release of normally granule-sequestered toxic components into the phagocyte's cytoplasm. *E. histolytica*, an intestinal protozoan that causes amebic dysentery, can disrupt phagocyte membranes after direct contact via the release of protozoal phospholipase A and pore-forming peptides.

Microbial Survival Inside Phagocytes

Many important microbial pathogens use a variety of strategies to survive inside phagocytes (particularly macrophages) after ingestion. Inhibition of fusion of the phagocytic vacuole (the phagosome) containing the initially ingested microbe with the lysosomal granules containing antimicrobial substances (the lysosome) allows *M. tuberculosis*, *S. enterica* serovar Typhi, and *Toxoplasma gondii* to survive inside macrophages. Some organisms, such as *L. monocytogenes*, escape into the phagocyte's cytoplasm to grow and eventually spread to other cells. Resistance to killing within the macrophage and subsequent growth are critical to successful infection by herpes-type viruses, measles virus, poxviruses, *Salmonella*, *Yersinia*, *Legionella*, *Mycobacterium*, *Trypanosoma*, *Nocardia*, *Histoplasma*, *Toxoplasma*, and *Rickettsia*. *Salmonella* species use a master regulatory system—in which the *PhoP*/*PhoQ* genes control other genes—to enter and survive within cells, with intracellular survival entailing structural changes in the cell envelope LPS.

TISSUE INVASION AND TISSUE TROPISM

Tissue Invasion

Most viral pathogens cause disease by growth at skin or mucosal entry sites, but some pathogens spread from the initial site to deeper tissues. Virus can spread via the nerves (rabies virus) or plasma (picornaviruses) or within migratory blood cells (poliovirus, Epstein-Barr virus, and many others). Specific viral genes determine where and how individual viral strains can spread.

Bacteria may invade deeper layers of mucosal tissue via intracellular uptake by epithelial cells, traversal of epithelial cell junctions, or penetration through denuded epithelial surfaces. Among virulent *Shigella* strains and invasive *E. coli*, outer-membrane proteins are critical to epithelial cell invasion and bacterial multiplication. *Neisseria* and *Haemophilus* species penetrate mucosal cells by poorly understood mechanisms before dissemination into the bloodstream. Staphylococci and streptococci elaborate a variety of extracellular enzymes, such as hyaluronidase, lipases, nucleases, and hemolysins, that are probably important in breaking down cellular and matrix structures and allowing the bacteria access to deeper tissues and blood. Staphylococcal α hemolysin binding to ADAM-10 leads to endothelial cell damage and disruption of vascular barrier function, events that are probably critical for systemic spread of *S. aureus* from an initial infectious site. Organisms that colonize the gastrointestinal tract can often translocate through the mucosa into the blood and, under circumstances in which host defenses are inadequate, cause bacteremia. *Yersinia enterocolitica* can invade the mucosa through the activity of the invasin protein. The complex milieu of basement membrane–containing structures such as laminin and collagen that anchor epithelial cells to mucosal surfaces must often be breached. A family of microbial surface components recognizing adhesive matrix molecules (MSCRAMMs) can attach bacteria to the extracellular matrix and, along with proteases that degrade the basement proteins as well as surface-bound plasminogen and matrix metalloproteinases recruited from the host, permit breaching of this structure. Some bacteria (e.g., *Brucella*) can be carried from a mucosal site to a distant site by phagocytic cells that ingest but fail to kill the bacteria.

Fungal pathogens almost always take advantage of host immunocompromise to spread hematogenously to deeper tissues. The AIDS epidemic has resoundingly illustrated this principle: the immunodeficiency of many HIV-infected patients permits the development of life-threatening fungal infections of the lung, blood, and brain. Other than the capsule of *C. neoformans*, specific fungal antigens involved in tissue invasion are not well characterized. Both fungal pathogens and protozoal pathogens (e.g., *Plasmodium* species and *E. histolytica*) undergo morphologic changes to spread within a host. *C. albicans* undertakes a yeast–hyphal transformation wherein the hyphal forms are found where the fungus is infiltrating the mucosal barrier of tissues, while the yeast form grows on epithelial cell surfaces as well as the tips of hyphae that have infiltrated tissues. Malarial parasites grow in liver cells as merozoites and are released into the blood to invade erythrocytes and become trophozoites. *E. histolytica* is found as both a cyst and a trophozoite in the intestinal lumen, through which this pathogen enters the host, but only the trophozoite form can spread systemically to cause amebic liver abscesses. Other protozoal pathogens, such as *T. gondii*, *Giardia lamblia*, and *Cryptosporidium*, also undergo extensive morphologic changes after initial infection to spread to other tissues.

Tissue Tropism

While it is well known that certain microbes cause disease by infecting specific tissues, the molecular basis for tissue tropism is understood somewhat better for viral pathogens than for other infectious agents. Specific receptor–ligand interactions clearly underlie the ability of certain viruses to enter cells within tissues and disrupt normal tissue function, but the mere presence of a receptor for a virus on a target tissue is not sufficient for tissue tropism. Factors in the cell, route of viral entry, viral capacity to penetrate into cells, viral genetic elements that regulate gene expression, and pathways of viral spread in a tissue all affect tissue tropism. Some viral genes are best transcribed in specific target cells, such as hepatitis B genes in liver cells and Epstein-Barr virus genes in B lymphocytes. The route of inoculation of poliovirus determines its neurotropism, although the molecular basis for this association is not understood.

Compared with viral tissue tropism, the tissue tropism of bacterial and parasitic infections has not been as clearly elucidated, but studies of *Neisseria* species have provided insights. Both *N. gonorrhoeae,* which colonizes and infects the human genital tract, and *N. meningitidis*, which principally colonizes the human oropharynx but can spread to the brain, produce type IV pili (Tfp) that mediate adherence to host tissues. In the case of *N. gonorrhoeae*, the Tfp bind to a glucosamine-galactose-containing adhesin on the surface of cervical and urethral cells; in the case of *N. meningitidis*, the Tfp bind to cells in the human meninges in order to cross the blood-brain barrier. *N. gonorrhoeae* can use cytidine monophosphate *N*-acetylneuraminic acid from host tissues to add *N*-acetylneuraminic acid (sialic acid) to its lipooligosaccharide O side chain, and this alteration makes the organism resistant to host defenses. Lactate, present at high levels on genital mucosal surfaces, stimulates sialylation of gonococcal lipooligosaccharide. Bacteria with sialic acid sugars in their capsules, such as *N. meningitidis*, *E. coli* K1, and group B streptococci, have a propensity to cause meningitis, but this generalization has many exceptions. For example, all recognized serotypes of group B streptococci contain sialic acid in their capsules, but only one serotype (III) is responsible for most cases of group B streptococcal meningitis. Moreover, both *H. influenzae* and *S. pneumoniae* can readily cause meningitis, but these organisms do not have sialic acid in their capsules.

TISSUE DAMAGE AND DISEASE

Disease is a complex phenomenon resulting from tissue invasion and destruction, toxin elaboration, and host response. Viruses cause much of their damage by exerting a cytopathic effect on host cells and inhibiting host defenses. The growth of bacterial, fungal, and protozoal parasites in tissue, which may or may not be accompanied by toxin elaboration, can also compromise tissue function and lead to disease. For some bacterial and possibly some fungal pathogens, toxin production is one of the best-characterized molecular mechanisms of pathogenesis, while host factors such as IL-1, TNF-α, kinins, inflammatory proteins, products of complement activation, and mediators derived from arachidonic acid metabolites (leukotrienes) and cellular degranulation (histamines) readily contribute to the severity of disease.

Viral Disease

Viral pathogens inhibit host immune responses by a variety of mechanisms—e.g., by decreasing production of major histocompatibility complex molecules (adenovirus E3 protein), diminishing cytotoxic T cell recognition of virus-infected cells (Epstein-Barr virus nuclear antigen 1 and cytomegalovirus intermediate–early protein), producing virus-encoded complement receptor proteins that protect infected cells from complement-mediated lysis (herpesvirus and vaccinia virus),

making proteins that interfere with the action of IFN (influenza virus and poxvirus), and elaborating superantigen-like proteins (mouse mammary tumor virus and related retroviruses and the rabies nucleocapsid). Superantigens activate large populations of T cells that express particular subsets of the T cell receptor β protein, causing massive cytokine release and subsequent host reactions. Dengue virus is the most common insect-transmitted virus in the world, causing symptoms ranging from none to serious systemic illness or "break-bone" fever (severe fever and pain). Along with the recent epidemic of the related flavivirus Zika virus, disruptions to host innate immunity and inhibition of programmed cell death that allows continued viral replication underlie the ability of these viruses to cause infections. Infections of pregnant women with Zika virus can result in viral crossing of the placenta; viral entry into and growth in fetal brain tissues result in the birth of neonates with microcephaly. Viruses also produce peptide growth factors for host cells, which disrupt normal cellular growth, proliferation, and differentiation. In addition, viral factors can bind to and interfere with the function of host receptors for signaling molecules. Modulation of cytokine production during viral infection can stimulate viral growth inside cells with receptors for the cytokine, and virus-encoded cytokine homologues (e.g., the Epstein-Barr virus BCRF1 protein, which is highly homologous to the immunoinhibitory IL-10 molecule) can prevent immune-mediated clearance of viral particles. Viruses cause disease in neural cells by interfering with levels of neurotransmitters without necessarily destroying the cells, or they may induce either programmed cell death (apoptosis) to destroy tissues or inhibitors of apoptosis to allow prolonged viral infection of cells. For infection to spread, many viruses must be released from cells. The HIV protein U (Vpu) facilitates virus release, a process that is specific to certain cells. Mammalian cells produce a restriction factor that inhibits release of some viruses; for HIV, this factor is designated BST-2 (bone marrow stromal antigen 2)/HM1.24/CD317, or tetherin. Vpu of HIV interacts with tetherin, allowing release of infectious virus. Overall, virus-induced disruption of normal cellular and tissue function promotes clinical disease.

Bacterial Toxins

Among the first infectious diseases to be understood were those due to toxin-elaborating bacteria. Diphtheria, botulism, and tetanus toxins are responsible for the diseases associated with local infections due to *Corynebacterium diphtheriae*, *Clostridium botulinum*, and *Clostridium tetani*, respectively. *C. difficile* is an anaerobic gram-positive organism that elaborates two toxins, A and B, responsible for disruption of the intestinal mucosa when its numbers expand in the intestine, leading to antibiotic-associated diarrhea and potentially pseudomembranous colitis. A clinical trial evaluating prevention of recurrence of *C. difficile* infection by monoclonal antibodies to toxins A and B showed that antibody to toxin B, but not toxin A, had a significant impact. Enterotoxins produced by *E. coli*, *Salmonella, Shigella*, staphylococci, and *V. cholerae* contribute to diarrheal disease caused by these organisms. Staphylococci, streptococci, *P. aeruginosa*, and *Bordetella* elaborate various toxins that cause or contribute to disease, including toxic shock syndrome toxin 1; erythrogenic toxin; exotoxins A, S, T, and U; and pertussis toxin. A number of these toxins (e.g., cholera toxin, diphtheria toxin, pertussis toxin, *E. coli* heat-labile toxin, and *P. aeruginosa* exotoxin) have adenosine diphosphate ribosyl transferase activity; i.e., the toxins enzymatically catalyze the transfer of the adenosine diphosphate ribosyl portion of nicotinamide adenine diphosphate to target proteins and inactivate them. The staphylococcal enterotoxins, toxic shock syndrome toxin 1, and the streptococcal pyogenic exotoxins behave as superantigens, stimulating certain T cells to proliferate without processing of the protein toxin by antigen-presenting cells. Part of this process involves stimulation of the antigen-presenting cells to produce IL-1 and TNF-α, which have been implicated in many clinical features of diseases like toxic shock syndrome and scarlet fever. A number of gram-negative pathogens (*Salmonella*, *Yersinia*, and *P. aeruginosa*) can inject toxins directly into host target cells by means of a complex set of proteins referred to as the *type III secretion system*. Loss or inactivation of this virulence system usually greatly reduces the capacity of a bacterial pathogen to cause disease.

Endotoxin

The lipid A portion of LPS in some gram-negative bacteria has potent biologic activities that cause many of the clinical manifestations of gram-negative bacterial sepsis, including fever, muscle proteolysis, uncontrolled intravascular coagulation, and shock. The effects of lipid A appear to be mediated by the production of potent cytokines due to LPS binding to CD14 and signal transduction via TLRs, particularly TLR4. Cytokines exhibit potent hypothermic activity through effects on the hypothalamus; they also increase vascular permeability, alter the activity of endothelial cells, and induce endothelial-cell procoagulant activity. Numerous therapeutic strategies aimed at neutralizing the effects of endotoxin are under investigation, and while studies with laboratory animals have been promising, they have not yet translated into positive results for human septic shock.

Invasion

Many diseases are caused primarily by pathogens growing to high levels in tissues. Pneumococcal pneumonia is mostly attributable to the growth of *S. pneumoniae* in the lung and the attendant host inflammatory response, although specific factors that enhance this process (e.g., pneumolysin) may be responsible for some of the pathogenic properties of the pneumococcus. Disease that follows bacteremia and invasion of the meninges by meningitis-producing bacteria such as *N. meningitidis*, *H. influenzae*, *E. coli* K1, and group B streptococci appears to be due solely to the ability of these organisms to gain access to these tissues, multiply in them, and provoke cytokine production leading to tissue-damaging host inflammation.

Specific molecular mechanisms accounting for tissue invasion by fungal and protozoal pathogens are less well described. Except for studies pointing to factors like capsule and melanin production by *C. neoformans,* and possibly levels of cell wall glucans in some pathogenic fungi, the molecular basis for fungal invasiveness is not well defined. Melanism has been shown to protect the fungal cell against death caused by phagocyte factors such as nitric oxide, superoxide, and hypochlorite. Morphogenic variation and production of proteases (e.g., the *Candida* aspartyl proteinase) have been implicated in fungal invasion of host tissues.

If pathogens are to invade host tissues (particularly the blood), they must avoid the major host defenses represented by complement and phagocytic cells. Bacteria most often avoid these defenses through their surface polysaccharides—either capsular polysaccharides or long O-side-chain antigens characteristic of the smooth LPS of gram-negative bacteria. These molecules can prevent the activation and/or deposition of complement opsonins or can limit the access of phagocytic cells with receptors for complement opsonins to these molecules when they are deposited on the bacterial surface below the capsular layer. Another potential mechanism of microbial virulence is the ability of some organisms to present the capsule as an apparent self antigen through molecular mimicry. For example, the polysialic acid capsule of group B *N. meningitidis* is chemically identical to an oligosaccharide found on human brain cells.

Immunochemical studies of capsular polysaccharides have led to an appreciation of the tremendous chemical diversity that can result from the linking of a few monosaccharides. For example, three hexoses can link up in more than 300 different, potentially serologically distinct ways, while three amino acids have only six possible peptide combinations. Capsular polysaccharides have been used as effective vaccines against meningococcal meningitis as well as against pneumococcal and *H. influenzae* infections and may prove to be of value as vaccines against any organism that expresses a nontoxic, immunogenic capsular polysaccharide. In addition, most encapsulated pathogens become virtually avirulent when capsule production is interrupted by genetic manipulation; this observation emphasizes the importance of this structure in pathogenesis. A capsule-like surface polysaccharide, PNAG, has been found as a conserved structure shared by many microbes but generally is a poor target for antibody-mediated immunity because of the propensity of most humans and animals, all of which are colonized by PNAG-producing microbes, to produce a nonprotective type of antibody. Altering the structure of PNAG by removing the acetate substituents on the *N*-acetyl glucosamine monomers leads to an immunogenic form, deacetylated PNAG, that has been reported to induce antibodies that are protective in animals against diverse microbial pathogens.

Host Response

The inflammatory response of the host is critical for interruption and resolution of the infectious process but also is often responsible for the signs and symptoms of disease. Infection promotes a complex series of host responses involving the complement, kinin, and coagulation pathways. The production of cytokines such as IL-1, IL-17, IL-18, TNF-α, IFN-γ, and other factors regulated in part by the NF-κB transcription factor leads to fever, muscle proteolysis, and other effects. An inability to kill or contain the microbe usually results in further damage due to the progression of inflammation and infection. For example, in many chronic infections, degranulation of host inflammatory cells leads to release of host proteases, elastases, histamines, and other toxic substances that can degrade host tissues. Chronic inflammation in any tissue can lead to the destruction of that tissue and to clinical disease associated with loss of organ function, such as sterility from pelvic inflammatory disease caused by chronic infection with *N. gonorrhoeae*.

The nature of the host response elicited by the pathogen often determines the pathology of a particular infection. Local inflammation produces local tissue damage, while systemic inflammation, such as that seen during sepsis, can result in the signs and symptoms of septic shock. The severity of septic shock is associated with the degree of production of host effectors. Disease due to intracellular parasitism results from the formation of granulomas wherein the host attempts to wall off the parasite inside a fibrotic lesion surrounded by fused epithelial cells that make up so-called multinucleated giant cells. A number of pathogens, particularly anaerobic bacteria, staphylococci, and streptococci, provoke the formation of an abscess, probably because of the presence of zwitterionic surface polysaccharides such as the capsular polysaccharide of *Bacteroides fragilis*. The outcome of an infection depends on the balance between an effective host response that eliminates a pathogen and an excessive inflammatory response that is associated with an inability to eliminate a pathogen and with the resultant tissue damage that leads to disease.

TRANSMISSION TO NEW HOSTS

As part of the pathogenic process, most microbes are shed from the host, often in a form infectious for susceptible individuals. However, the rate of transmissibility may not necessarily be high, even if the disease is severe in the infected individual, as these traits are not linked. Most pathogens exit via the same route by which they entered: respiratory pathogens by aerosols from sneezing or coughing or through salivary spread, gastrointestinal pathogens by fecal–oral spread, sexually transmitted diseases by venereal spread, and vector-borne organisms by either direct contact with the vector through a blood meal or indirect contact with organisms shed into environmental sources such as water. Microbial factors that specifically promote transmission are not well characterized. Respiratory shedding is facilitated by overproduction of mucous secretions, with consequently enhanced sneezing and coughing. Diarrheal toxins such as cholera toxin, *E. coli* heat-labile toxins, and *Shigella* toxins probably facilitate fecal–oral spread of microbial cells in the high volumes of diarrheal fluid produced during infection. The ability to produce phenotypic variants that resist hostile environmental factors (e.g., the highly resistant cysts of *E. histolytica* shed in feces) represents another mechanism of pathogenesis relevant to transmission. Blood parasites such as *Plasmodium* species change phenotype after ingestion by a mosquito—a prerequisite for the continued transmission of this pathogen. Venereally transmitted pathogens may undergo phenotypic variation due to the production of specific factors to facilitate transmission, but shedding of these pathogens into the environment does not result in the formation of infectious foci.

SUMMARY

In summary, the molecular mechanisms used by pathogens to colonize, invade, infect, and disrupt the host are numerous and diverse. Each phase of the infectious process involves a variety of microbial and host factors interacting in a manner that can result in disease. Recognition of the coordinated genetic regulation of virulence factor elaboration when organisms move from their natural environment into the mammalian host emphasizes the complex nature of the host–parasite interaction. Fortunately, the need for diverse factors in successful infection and disease implies that a variety of therapeutic strategies may be developed to interrupt this process and thereby prevent and treat microbial infections.

FURTHER READING

Bäumler AJ, Sperandio V: Interactions between the microbiota and pathogenic bacteria in the gut. Nature 535:85, 2016. [PubMed: 27383983]

Brunke S et al: Virulence factors in fungal pathogens of man. Curr Opin Microbiol 32:89, 2016. [PubMed: 27257746]

Fuchs Y, Steller H: Live to die another way: Modes of programmed cell death and the signals emanating from dying cells. Nat Rev Mol Cell Bio l16:329, 2016.

Honda K, Littman DR: The microbiota in adaptive immune homeostasis and disease. Nature 535:75, 2016. [PubMed: 27383982]

Olagnier D et al: Dengue virus immunopathogenesis: Lessons applicable to the emergence of Zika virus. J Mol Biol 428:3429, 2016. [PubMed: 27130436]

Chapter 117: Approach to the Acutely Ill Infected Febrile Patient

Tamar F. Barlam; Dennis L. Kasper

INTRODUCTION

The physician treating the acutely ill febrile patient must be able to recognize infections that require emergent attention. If such infections are not adequately evaluated and treated at initial presentation, the opportunity to alter an adverse outcome may be lost. In this chapter, the clinical presentations of and approach to patients with relatively common infectious disease emergencies are discussed. These infectious processes and their treatments are discussed in detail in other chapters.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Acute Febrile Illness

Before the history is elicited and a physical examination is performed, an immediate assessment of the patient's general appearance can yield valuable information. The perceptive physician's subjective sense that a patient is septic or toxic often proves accurate. Visible agitation or anxiety in a febrile patient can be a harbinger of critical illness.

HISTORY

Presenting symptoms are frequently nonspecific. Detailed questions should be asked about the onset and duration of symptoms and about changes in severity or rate of progression over time. Host factors and comorbid conditions may increase the risk of infection with certain organisms or of a more fulminant course than is usually seen. Lack of splenic function, alcoholism with significant liver disease, IV drug use, HIV infection, diabetes, malignancy, organ transplantation, and chemotherapy all predispose to specific infections and frequently to increased severity. The patient should be questioned about factors that might help identify a nidus for invasive infection, such as recent upper respiratory tract infections, influenza, or varicella; prior trauma; disruption of cutaneous barriers due to lacerations, burns, surgery, body piercing, or decubiti; and the presence of foreign bodies, such as nasal packing after rhinoplasty, tampons, or prosthetic joints. Travel, contact with pets or other animals, or activities that might result in tick or mosquito exposure can lead to diagnoses that would not otherwise be considered. Recent dietary intake, medication use, social

or occupational contact with ill individuals, vaccination history, recent sexual contacts, and menstrual history may be relevant. Pregnancy might increase the risk and severity of some illnesses, such as influenza, or increase the risk of significant morbidity for the fetus, as in *Listeria* or Zika virus infection. A review of systems should focus on any neurologic signs or sensorium alterations, rashes or skin lesions, and focal pain or tenderness and should also include a general review of respiratory, gastrointestinal, or genitourinary symptoms.

PHYSICAL EXAMINATION

A complete physical examination should be performed, with special attention to several areas that are sometimes given short shrift in routine examinations. Assessment of the patient's general appearance and vital signs, skin and soft tissue examination, and the neurologic evaluation are of particular importance.

The patient may appear either anxious and agitated or lethargic and apathetic. Fever is usually present, although elderly patients and compromised hosts (e.g., patients who are uremic or cirrhotic and those who are taking glucocorticoids or nonsteroidal anti-inflammatory drugs) may be afebrile despite serious underlying infection. Critically ill patients may be hypothermic, with a high risk of organ failure and mortality. Measurement of blood pressure, heart rate, and respiratory rate helps determine the degree of hemodynamic and metabolic compromise. The patient's airway must be evaluated to rule out the risk of obstruction from an invasive oropharyngeal infection.

The etiologic diagnosis may become evident in the context of a thorough skin examination **(Chap. 16)**. Petechial rashes are typically seen with meningococcemia or Rocky Mountain spotted fever (RMSF; **see Fig. A1-16**); erythroderma is associated with toxic shock syndrome (TSS) and drug fever. The soft tissue and muscle examination is critical. Areas of erythema or duskiness, edema, and tenderness may indicate underlying necrotizing fasciitis, myositis, or myonecrosis. The neurologic examination must include a careful assessment of mental status for signs of early encephalopathy. Evidence of nuchal rigidity or focal neurologic findings should be sought.

DIAGNOSTIC WORKUP

After a quick clinical assessment, diagnostic material should be obtained rapidly and antibiotic and supportive treatment begun. Blood (for cultures; baseline complete blood count with differential; measurement of serum electrolytes, blood urea nitrogen, serum creatinine, and serum glucose; and liver function tests) can be obtained at the time an IV line is placed and before antibiotics are administered. The blood lactate concentration also should be measured. Three sets of blood cultures should be performed for patients with possible acute endocarditis. Blood smears from patients at risk for severe parasitic disease, such as malaria or babesiosis **(Chaps. 219, 220, and A6)**, must be examined for the diagnosis and quantitation of parasitemia. Blood smears may also be diagnostic in ehrlichiosis and anaplasmosis.

Patients with possible meningitis should have cerebrospinal fluid (CSF) drawn before the initiation of antibiotic therapy. Focal findings, depressed mental status, or papilledema should be evaluated by brain imaging prior to lumbar puncture, which, in this setting, could initiate herniation. *Antibiotics should be administered before imaging but after blood for cultures has been drawn*. If CSF cultures are negative, blood

cultures will provide the diagnosis in 50–70% of cases. Molecular diagnostic techniques (e.g., broad-range 16S rRNA gene polymerase chain reaction testing for bacterial meningitis pathogens) are of increasing importance in the rapid diagnosis of life-threatening infections.

Focal abscesses necessitate immediate CT or MRI as part of an evaluation for surgical intervention. Other diagnostic procedures, such as wound cultures, should not delay the initiation of treatment for more than minutes. Once emergent evaluation, diagnostic procedures, and (if appropriate) surgical consultation (see below) have been completed, other laboratory tests can be conducted. Appropriate radiography, computed axial tomography, MRI, urinalysis, measurement of the erythrocyte sedimentation rate and/or C-reactive protein level, procalcitonin monitoring, and transthoracic or transesophageal echocardiography all may prove important.

TREATMENT

TREATMENT

The Acutely Ill Patient

In the acutely ill patient, empirical antibiotic therapy is critical and should be administered without undue delay in addition to fluid resuscitation and vasopressor support as needed. Increased prevalence of antibiotic resistance in community-acquired bacteria must be considered when antibiotics are selected. Table 117-1 lists first-line empirical regimens for infections considered in this chapter. In addition to the rapid initiation of antibiotic therapy, several of these infections require urgent surgical attention. Neurosurgical evaluation for subdural empyema, otolaryngologic surgery for possible mucormycosis, and cardiothoracic surgery for critically ill patients with acute endocarditis are as important as antibiotic therapy. For infections such as necrotizing fasciitis and clostridial myonecrosis, rapid surgical intervention supersedes other diagnostic or therapeutic maneuvers.

Adjunctive treatments may reduce morbidity and mortality rates and include dexamethasone for bacterial meningitis or IV immunoglobulin for TSS and necrotizing fasciitis caused by group A *Streptococcus*. Adjunctive therapies should usually be initiated within the first hours of treatment; however, dexamethasone for bacterial meningitis must be given before or at the time of the first dose of antibiotic. Glucocorticoids can also be harmful, sometimes resulting in worse outcomes—e.g., when given in the setting of cerebral malaria or viral hepatitis.

TABLE 117-1

Empirical Treatment for Common Infectious Disease Emergencies[a]

CLINICAL SYNDROME	POSSIBLE ETIOLOGIES	TREATMENT	COMMENTS	SEE CHAP(S).
Sepsis without a Clear Focus				
Septic shock	*Pseudomonas* spp., gram-negative enteric bacilli, *Staphylococcus* spp., *Streptococcus* spp.	Vancomycin (15 mg/kg q12h)[b] *plus* gentamicin (5 mg/kg per day) ***plus either*** Piperacillin/tazobactam (3.375–4.5 g q6h) *or* cefepime (2 g q8h)[c]	Empirical therapy should be tailored to local resistance patterns. Adjust treatment when culture data become available.	142, 143, 156, 159, 297
Overwhelming post-splenectomy sepsis	*Streptococcus pneumoniae*, *Haemophilus influenzae*, *Neisseria meningitidis*	Ceftriaxone (2 g q12h) *plus* vancomycin (15 mg/kg q12h)[b]	If a β-lactam-sensitive strain is identified, vancomycin can be discontinued.	297
Babesiosis	*Babesia microti* (U.S.), *B. divergens* (Europe)	Clindamycin (600 mg q8h) *plus* quinine (650 mg q8h)	Atovaquone and azithromycin can be used in less severe disease and are associated with fewer side effects. Treatment with doxycycline (100 mg bid) for potential co-infection with *Borrelia burgdorferi* or *Anaplasma* spp. may be prudent.	217, 220
Sepsis with Skin Findings				

CLINICAL SYNDROME	POSSIBLE ETIOLOGIES	TREATMENT	COMMENTS	SEE CHAP(S).
Meningococcemia	*N. meningitidis*	Penicillin (4 mU q4h) *or* ceftriaxone (2 g q12h)	Ceftriaxone eradicates nasopharyngeal carriage of the organism. Close contacts require chemoprophylaxis with rifampin (600 mg q12h for 2 days) or ciprofloxacin (a single dose, 500 mg).	150
Rocky Mountain spotted fever (RMSF)	*Rickettsia rickettsii*	Doxycycline (100 mg bid)	If both meningococcemia and RMSF are being considered, use ceftriaxone (2 g q12h) *plus* doxycycline (100 mg bid). If RMSF is diagnosed, doxycycline is the proven superior agent.	182
Purpura fulminans	*S. pneumoniae, H. influenzae*, *N. meningitidis*	Ceftriaxone (2 g q12h) *plus* vancomycin (15 mg/kg q12h)[b]	If a β-lactam-sensitive strain is identified, vancomycin can be discontinued.	141, 150, 152, 297
Erythroderma: toxic shock syndrome	Group A *Streptococcus, Staphylococcus aureus*	Vancomycin (15 mg/kg q12h)[b] *plus* clindamycin (600 mg q8h	If a penicillin- or oxacillin-sensitive strain is isolated, these agents are superior to vancomycin (penicillin, 2 mU q4h; or oxacillin, 2 g IV q4h). The site of toxigenic bacteria should be debrided; IV immunoglobulin can be used in severe cases.[d]	142, 143

CLINICAL SYNDROME	POSSIBLE ETIOLOGIES	TREATMENT	COMMENTS	SEE CHAP(S).
Sepsis with Soft Tissue Findings				
Necrotizing fasciitis	Group A *Streptococcus,* mixed aerobic/anaerobic flora, CA-MRSA[e]	Vancomycin (15 mg/kg q12h)[b] *plus* clindamycin (600 mg q8h) *plus* gentamicin (5 mg/kg per day)	Urgent surgical evaluation is critical. Adjust treatment when culture data become available.	124, 142, 143
Clostridial myonecrosis	*Clostridium perfringens*	Penicillin (2 mU q4h) *plus* clindamycin (600 mg q8h)	Urgent surgical evaluation is critical.	149
Neurologic Infections				
Bacterial meningitis	*S. pneumoniae*, *N. meningitidis*	Ceftriaxone (2 g q12h) *plus* vancomycin (15 mg/kg q12h)[b]	If a β-lactam–sensitive strain is identified, vancomycin can be discontinued. If the patient is >50 years old or has comorbid disease, add ampicillin (2 g q4h) for *Listeria* coverage. Dexamethasone (10 mg q6h for 4 days) improves outcome in adults with meningitis (especially pneumococcal).	133
Brain abscess, suppurative intracranial infections	*Streptococcus* spp., *Staphylococcus* spp., anaerobes, gram-negative bacilli	Vancomycin (15 mg/kg q12h)[b] *plus* metronidazole (500 mg q8h) *plus* ceftriaxone (2 g q12h)	Urgent surgical evaluation is critical. If a penicillin- or oxacillin-sensitive strain is isolated, these agents are superior to vancomycin (penicillin, 4 mU q4h; *or* oxacillin, 2 g q4h).	133

CLINICAL SYNDROME	POSSIBLE ETIOLOGIES	TREATMENT	COMMENTS	SEE CHAP(S).
Cerebral malaria	*Plasmodium falciparum*	Artesunate (2.4 mg/kg IV at 0, 12, and 24 h; then once daily)[f] *or* quinine (IV loading dose of 20 mg salt/kg; then 10 mg/kg q8h)	Do not use glucocorticoids. Use IV quinidine if IV quinine is not available. During IV quinidine treatment, blood pressure and cardiac function should be monitored continuously and blood glucose periodically.	217, 219
Spinal epidural abscess	*Staphylococcus* spp., gram-negative bacilli	Vancomycin (15 mg/kg q12h)[b] ***plus either*** Piperacillin/tazobactam (3.375–4.5 g q6h) *or* cefepime (2 g q8h)[c]	Surgical evaluation is essential. If a penicillin- or oxacillin-sensitive strain is isolated, these agents are superior to vancomycin (penicillin, 4 mU q4h; *or* oxacillin, 2 g q4h).	434
Focal Infections				
Acute bacterial endocarditis	*S. aureus,* β-hemolytic streptococci, HACEK group,[g] *Neisseria* spp., *S. pneumoniae*	Ceftriaxone (2 g q12h) *plus* vancomycin (15 mg/kg q12h)[b]	Adjust treatment when culture data become available. Surgical evaluation is essential.	123

[a]These empirical regimens include coverage for gram-positive pathogens that are resistant to β-lactam antibiotics. Local resistance patterns should be considered and may alter the need for empirical vancomycin. [b]A vancomycin loading dose of 20–25 mg/kg can be considered in critically ill patients. [c]β-Lactam antibiotics may exhibit unpredictable pharmacodynamics in sepsis. Higher dosing or prolonged or continuous infusions can be considered. [d]The optimal dose of IV immunoglobulin has not been determined, but the median dose in observational studies is 2 g/kg (total dose administered for 1–5 days). [e]Community-acquired methicillin-resistant *S. aureus*. [f]In the United States, artesunate must be obtained through the Centers for Disease Control and Prevention. For patients diagnosed with

severe malaria, full doses of parenteral antimalarial treatment should be started with whichever recommended antimalarial agent is first available. [g] *Haemophilus* spp., *Aggregatibacter* spp., *Cardiobacterium hominis, Eikenella corrodens,* and *Kingella kingae*.

SPECIFIC PRESENTATIONS

The infections considered below according to common clinical presentation can have rapidly catastrophic outcomes, and their immediate recognition and treatment can be life-saving. Recommended empirical therapeutic regimens are presented in Table 117-1.

SEPSIS WITHOUT AN OBVIOUS FOCUS OF PRIMARY INFECTION

Patients initially have a brief prodrome of nonspecific symptoms and signs that progresses quickly to hemodynamic instability with hypotension, tachycardia, tachypnea, respiratory distress, and altered mental status. Disseminated intravascular coagulation (DIC) with clinical evidence of a hemorrhagic diathesis is a poor prognostic sign.

Septic Shock

Patients with bacteremia leading to septic shock may have a primary site of infection (e.g., pneumonia, pyelonephritis, or cholangitis) that is not evident initially (See also Chap. 297). Elderly patients with comorbid conditions, hosts compromised by malignancy and neutropenia, and patients who have recently undergone a surgical procedure or hospitalization are at increased risk for an adverse outcome. Gram-negative bacteremia with organisms such as *Pseudomonas aeruginosa* or *Escherichia coli* and gram-positive infection with organisms such as *Staphylococcus aureus* (including methicillin-resistant *S. aureus* [MRSA]) or group A streptococci can present as intractable hypotension and multiorgan failure. Treatment can usually be initiated empirically on the basis of the presentation, host factors **(Chap. 297)**, and local patterns of bacterial resistance. Outcomes are worse when antimicrobial treatment is delayed or when the responsible pathogen ultimately proves not to be susceptible to the initial regimen. Active empirical antimicrobial coverage administered before admission to the intensive care unit is strongly associated with improved survival. Broad-spectrum antimicrobial agents are therefore recommended and should be instituted rapidly, preferably within the first hours after presentation. Risk factors for fungal infection should be assessed, as the incidence of fungal septic shock is increasing. Biomarkers such as C-reactive protein and procalcitonin have not proved reliable diagnostically but, when measured over time, can facilitate appropriate de-escalation of therapy and predict outcome. Glucocorticoids are often considered for patients with severe sepsis who do not respond to fluid resuscitation and vasopressor therapy. However, conclusive evidence for the efficacy of glucocorticoids in this setting is lacking.

Overwhelming Infection in Asplenic Patients

Patients without splenic function are at risk for overwhelming bacterial sepsis (See also Chap. 297). Asplenic adult patients succumb to sepsis at 58 times the rate of the general population. Most infections are thought to occur within the first l or 2 years, but the increased risk persists throughout life. The median interval between splenectomy and sepsis is 5.75 years, with a range of 1–19 years. In asplenia, encapsulated bacteria cause the majority of infections. Adults, who are more likely to have antibody to these organisms, are at lower risk than children. *Streptococcus pneumoniae* is the most common isolate, causing 40–70% of cases. The risk of infection with *Haemophilus influenzae* or *Neisseria meningitidis* is also greater in patients without splenic function, but reported cases are declining. Severe clinical manifestations of infections due to *E. coli*, *S. aureus*, group B streptococci, *P. aeruginosa*, *Bordetella holmesii*, and *Capnocytophaga*, *Babesia*, and *Plasmodium* species have been described.

Babesiosis

A history of recent travel to endemic areas raises the possibility of infection with *Babesia* (See also Chap. 220). Between 1 and 4 weeks after a tick bite, the patient experiences chills, fatigue, anorexia, myalgia, arthralgia, shortness of breath, nausea, and headache; ecchymosis and/or petechiae are occasionally seen. The tick that most commonly transmits *Babesia*, *Ixodes scapularis*, also transmits *Borrelia burgdorferi* (the agent of Lyme disease) and *Anaplasma*; co-infection can occur, resulting in more severe disease. Infection with the European species *Babesia divergens* is more frequently fulminant than that due to the U.S. species *Babesia microti. B. divergens* causes a febrile syndrome with hemolysis, jaundice, hemoglobinemia, and renal failure and is associated with a mortality rate of >40%. Severe babesiosis is especially common in asplenic hosts but does occur in hosts with normal splenic function, particularly those >60 years of age and those with underlying immunosuppressive conditions such as HIV infection or malignancy. Complications include renal failure, acute respiratory failure, and DIC.

Other Sepsis Syndromes

Tularemia **(Chap. 165)** is seen throughout the United States, but most cases recorded in 2015 occurred in South Dakota, Nebraska, Colorado, and Wyoming. This disease is associated with wild rabbit, tick, and tabanid fly contact. It can be transmitted by arthropod bite, handling of infected animal carcasses, consumption of contaminated food and water, or inhalation. The typhoidal form can be associated with gram-negative septic shock and a mortality rate of >30%, especially in patients with underlying comorbid or immunosuppressive conditions. Plague occurs infrequently in the United States **(Chap. 166)**, primarily after contact with ground squirrels, prairie dogs, or chipmunks, but is endemic in other parts of the world, with >90% of all cases occurring in Africa. The septic form is particularly rare and is associated with shock, multiorgan failure, and a 30% mortality rate. These infections should be considered in the appropriate epidemiologic setting. The Centers for Disease Control and Prevention lists *Francisella tularensis* and *Yersinia pestis* (the agents of tularemia and plague, respectively) along with *Bacillus anthracis* (the agent of anthrax) as important organisms that might be used for bioterrorism (Chap. S2).

SEPSIS WITH SKIN MANIFESTATIONS

Maculopapular rashes may reflect early meningococcal or rickettsial disease but are usually associated with nonemergent infections (See also Chap. 16). Exanthems are usually viral. Primary HIV infection commonly presents with a rash that is typically maculopapular and involves the upper part of the body but can spread to the palms and soles. The patient is usually febrile and can have lymphadenopathy, severe headache, dysphagia, diarrhea, myalgias, and arthralgias. Recognition of this syndrome provides an opportunity to prevent transmission and to institute treatment and monitoring early on.

Petechial rashes caused by viruses are seldom associated with hypotension or a toxic appearance, although there can be exceptions (e.g., severe measles or arboviral infection). Petechial rashes limited to the distribution of the superior vena cava are rarely associated with severe disease. In other settings, petechial rashes require more urgent attention.

Meningococcemia


Almost three-quarters of patients with *N. meningitidis* bacteremia have a rash (See also Chap. 150). Meningococcemia most often affects young children (i.e., those 6 months to 5 years old). In sub-Saharan Africa, the high prevalence of serogroup A meningococcal disease has been a threat to public health for more than a century. Thousands of deaths occur annually in this area, which is known as the “meningitis belt,” and large epidemic waves occur approximately every 8–12 years. Serogroups W135 and X are also important emerging pathogens in Africa. In the United States, sporadic cases and outbreaks occur in day-care centers, schools (grade school through college, particularly among college freshmen living in residential halls), and army barracks. Household contacts of index cases are at 400–800 times greater risk of disease than the general population. Patients may have fever, headache, nausea, vomiting, myalgias, changes in mental status, and meningismus. However, the rapidly progressive form of disease is not usually associated with meningitis. The rash is initially pink, blanching, and maculopapular, appearing on the trunk and extremities, but then becomes hemorrhagic, forming petechiae. Petechiae are first seen at the ankles, wrists, axillae, mucosal surfaces, and palpebral and bulbar conjunctiva, with subsequent spread on the lower extremities and to the trunk. A cluster of petechiae may be seen at pressure points—e.g., where a blood pressure cuff has been inflated. In rapidly progressive meningococcemia (10–20% of cases), the petechial rash quickly becomes purpuric **(see Fig. A1-41)**, and patients develop DIC, multiorgan failure, and shock; 50–60% of these patients die, and survivors often require extensive debridement or amputation of gangrenous extremities. Hypotension with petechiae for <12 h is associated with significant mortality. Cyanosis, coma, oliguria, metabolic acidosis, and elevated partial thromboplastin time also are associated with a fatal outcome. Antibiotics given in the office by the primary care provider before hospital evaluation and admission may improve prognosis; this observation suggests that early initiation of treatment may be life-saving. Meningococcal conjugate vaccines are protective against serogroups A, C, Y and W135 and are recommended for children 11–18 years of age and for other high-risk patients. Vaccines active against serogroup B are available and are recommended for high-risk individuals >10 years of age.

Rocky Mountain Spotted Fever and Other Rickettsial Diseases

RMSF is a tickborne disease caused by *Rickettsia rickettsii* that occurs throughout North and South America (See also Chap. 182). Other rickettsiae (e.g., *R. parkeri*, *R. akari*) can also cause spotted fever. Up to 40% of patients do not report a history of a tick bite, but a history of travel or outdoor activity (e.g., camping in tick-infested areas) can often be ascertained. For the first 3 days, headache, fever, malaise, myalgias, nausea, vomiting, and anorexia are documented. By day 3, half of patients have skin findings. Blanching macules develop initially on the wrists and ankles and then spread over the legs and trunk. The lesions become hemorrhagic and are frequently petechial. The rash spreads to palms and soles later in the course. The centripetal spread is a classic feature of RMSF but occurs in a minority of patients. Moreover, 10–15% of patients with RMSF never develop a rash. The patient can be hypotensive and develop noncardiogenic pulmonary edema, confusion, lethargy, and encephalitis progressing to coma. The CSF contains 10–100 cells/μL, usually with a predominance of mononuclear cells. The CSF glucose level is often normal; the protein concentration may be slightly elevated. Renal and hepatic injury as well as bleeding secondary to vascular damage are noted. For untreated infections, mortality rates are 20–30%. Delayed recognition and treatment are associated with a greater risk of death; Native Americans, children 5–9 years of age, adults >70 years old, and persons with underlying immunosuppression are at a 3- to 5-fold increased risk of death.


Other rickettsial diseases cause significant morbidity and mortality worldwide. *Mediterranean spotted fever* caused by *Rickettsia conorii* is found in Africa, southwestern and south-central Asia, and southern Europe. Patients have fever, flu-like symptoms, and an inoculation eschar at the site of the tick bite. A maculopapular rash develops within 1–7 days, involving the palms and soles but sparing the face. Elderly patients or those with diabetes, alcoholism, uremia, or congestive heart failure are at risk for severe disease characterized by neurologic involvement, respiratory distress, and gangrene of the digits or purpura fulminans. Mortality rates associated with this severe form of disease approach 50%. *Epidemic typhus*, caused by *Rickettsia prowazekii*, is transmitted in louse-infested environments and emerges in conditions of extreme poverty, war, and natural disaster. Patients experience a sudden onset of high fevers, severe headache, cough, myalgias, and abdominal pain. A maculopapular rash develops (primarily on the trunk) in more than half of patients and can progress to petechiae and purpura. Serious signs include delirium, coma, seizures, noncardiogenic pulmonary edema, skin necrosis, and peripheral gangrene. Mortality rates approached 60% in the preantibiotic era and continue to exceed 10–15% in contemporary outbreaks. *Scrub typhus*, caused by *Orientia tsutsugamushi* (a separate genus in the family Rickettsiaceae), is transmitted by larval mites or chiggers and is one of the most common infections in southeastern Asia and the western Pacific. The organism is found in areas of heavy scrub vegetation (e.g., along riverbanks). Patients may have an inoculation eschar and may develop a maculopapular rash, lymphadenopathy, and dyspnea. Severe cases progress to pneumonia, meningoencephalitis, myocarditis, DIC, and renal failure. Mortality rates range from 1% to 70% and vary by location, increasing age, myocarditis, delirium, pneumonitis, or signs of hemorrhage.

If recognized in a timely fashion, rickettsial disease is very responsive to treatment. Doxycycline (100 mg twice daily for 3–14 days) is the treatment of choice for both adults and children. The newer macrolides may be a suitable alternative, but mortality rates are higher when tetracycline-based treatment is not given.

Purpura Fulminans

Purpura fulminans is the cutaneous manifestation of DIC and presents as large ecchymotic areas and hemorrhagic bullae (See also Chaps. 150 and 297). Progression of petechiae to purpura, ecchymoses, and gangrene is associated with congestive heart failure, septic shock, acute renal failure, acidosis, hypoxia, hypotension, and death. Purpura fulminans has been associated primarily with *N. meningitidis* but, in splenectomized patients, may be associated with *S. pneumoniae*, *H. influenzae*, and *S. aureus*.

Ecthyma Gangrenosum

Septic shock caused by *P. aeruginosa* or *Aeromonas hydrophila* can be associated with ecthyma gangrenosum **(see Figs. 159-1 and A1-34)**: hemorrhagic vesicles surrounded by a rim of erythema with central necrosis and ulceration. These gram-negative bacteremias are most common among patients with neutropenia, extensive burns, and hypogammaglobulinemia.

Other Infections Associated with Rash

Vibrio vulnificus and other noncholera *Vibrio* bacteremic infections **(Chap. 163)** can cause focal skin lesions and overwhelming sepsis in hosts with chronic liver disease, heavy alcohol consumption, iron storage disorders, diabetes, renal insufficiency, hematologic disease, or malignancy or other immunocompromising conditions. After ingestion of contaminated raw shellfish (typically oysters from the Gulf Coast in U.S. cases), there is a sudden onset of malaise, chills, fever, and hypotension. The patient develops bullous or hemorrhagic skin lesions, usually on the lower extremities, and 75% of patients have leg pain. The mortality rate can be as high as 50–60%, particularly when the patient presents with hypotension. Outcomes are improved when patients are treated with fluoroquinolones with or without cephalosporins or with tetracycline-containing regimens. Other infections, caused by agents such as *Aeromonas*, *Klebsiella*, and *E. coli*, can cause hemorrhagic bullae and death due to overwhelming sepsis in cirrhotic patients. *Capnocytophaga canimorsus* can cause septic shock in asplenic patients. Infection typically follows a dog bite. Patients present with fever, chills, myalgia, vomiting, diarrhea, dyspnea, confusion, and headache. Findings can include an exanthem or erythema multiforme **(see Figs. 52-9 and A1-24)**, cyanotic mottling or peripheral cyanosis, petechiae, and ecchymosis. About 30% of patients with this fulminant form die of overwhelming sepsis and DIC, and survivors may require amputation because of gangrene.

Erythroderma

TSS **(Chaps. 142 and 143)** is usually associated with erythroderma. The patient presents with fever, malaise, myalgias, nausea, vomiting, diarrhea, and confusion. There is a sunburn-type rash that may be subtle and patchy but is usually diffuse and is found on the face, trunk, and extremities. Erythroderma, which desquamates after 1–2 weeks, is more common in *Staphylococcus*-associated than in *Streptococcus*-associated TSS. Hypotension develops rapidly—often within hours—after the onset of symptoms. Multiorgan failure occurs. Early renal failure may precede hypotension and distinguishes this syndrome from other septic shock syndromes. There may be no indication of a primary focal infection, although possible cutaneous or mucosal portals of entry for the organism can be ascertained when a careful history is taken. Colonization rather than overt infection of the vagina or a postoperative wound, for example, is typical with

staphylococcal TSS, and the mucosal areas appear hyperemic but not infected. Streptococcal TSS is more often associated with skin or soft tissue infection (including necrotizing fasciitis), and patients are more likely to be bacteremic. TSS caused by *Clostridium sordellii* is associated with childbirth or with skin injection of black-tar heroin. The diagnosis of TSS is defined by the clinical criteria of fever, rash, hypotension, and multiorgan involvement. (Of note, fever is typically absent when TSS is caused by *C. sordellii*.) The mortality rate is 5% for menstruation-associated TSS, 10–15% for nonmenstrual TSS, 30–70% for streptococcal TSS, and up to 90% for obstetric *C. sordellii* TSS. Clindamycin improves outcomes when included in the treatment regimen. Some studies have shown that use of IV immunoglobulin is associated with improved survival as well.

Viral Hemorrhagic Fevers


Viral hemorrhagic fevers **(Chaps. 204 and 205)** are zoonotic illnesses caused by viruses that reside in either animal reservoirs or arthropod vectors. These diseases occur worldwide and are restricted to areas where the host species live. They are caused by four major groups of viruses: Arenaviridae (e.g., Lassa fever in Africa), Bunyaviridae (e.g., Rift Valley fever in Africa; hantavirus hemorrhagic fever with renal syndrome in Asia; and Crimean-Congo hemorrhagic fever, which has an extensive geographic distribution), Filoviridae (e.g., Ebola and Marburg virus infections in Africa), and Flaviviridae (e.g., yellow fever in Africa and South America and dengue in Asia, Africa, and the Americas). Lassa fever and Ebola and Marburg virus infections are also transmitted from person to person. The vectors for most viral fevers are found in rural areas; dengue and yellow fever are important exceptions. After a prodrome of fever, myalgias, and malaise, patients develop evidence of vascular damage, petechiae, and local hemorrhage. Shock, multifocal hemorrhaging, and neurologic signs (e.g., seizures or coma) predict a poor prognosis. Dengue **(Chap. 204)** is the most common arboviral disease worldwide. More than half a million cases of dengue hemorrhagic fever occur each year, with at least 12,000 deaths. Patients have a triad of symptoms: hemorrhagic manifestations, evidence of plasma leakage, and platelet counts of <100,000/μL. Mortality rates are 10–20%. If dengue shock syndrome develops, mortality rates can reach 40%. Ebola infection has been associated with outbreaks with high mortality rates. The 2014 outbreak in West Africa had a mortality rate of >50%. Symptoms can appear 2–21 days after exposure, but most patients become ill within 9 days. The patient first presents with fatigue, fever, headache, and muscle pains, and the illness can progress to multiorgan failure and hemorrhaging. Careful volume-replacement therapy to maintain blood pressure and intravascular volume is key to survival in these infections. Ribavirin also may be useful against Arenaviridae and Bunyaviridae.

Other viral illnesses with rash, such as measles, can be associated with significant mortality rates. Steroids may sometimes be useful in severe disease in malnourished populations, especially if neurologic complications are present.

SEPSIS WITH A SOFT TISSUE/MUSCLE PRIMARY FOCUS

See also Chap. 124.

Necrotizing Fasciitis

This infection is characterized by extensive necrosis of the subcutaneous tissue and fascia. It may arise at a site of minimal trauma or surgical incision and may also be associated with recent varicella, childbirth, or muscle strain. The most common causes of necrotizing fasciitis are group A streptococci alone **(Chap. 143)** and a mixed facultative and anaerobic flora **(Chap. 124)**; the incidence of group A streptococcal necrotizing fasciitis has been increasing for the past quarter-century. Diabetes mellitus, IV drug use, chronic liver or renal disease, and malignancy are associated risk factors. Physical findings are initially minimal compared with the severity of pain and the degree of fever. The examination is often unremarkable except for soft tissue edema and erythema. The infected area is red, hot, shiny, swollen, and exquisitely tender. In untreated infection, the overlying skin develops blue-gray patches after 36 h, and cutaneous bullae and necrosis develop after 3–5 days. Necrotizing fasciitis due to a mixed flora, but not that due to group A streptococci, can be associated with gas production. Without treatment, pain decreases because of thrombosis of the small blood vessels and destruction of the peripheral nerves—an ominous sign. The mortality rate is 15–34% overall, >70% in association with TSS, and nearly 100% without surgical intervention. Necrotizing fasciitis may also be due to *Clostridium perfringens* **(Chap. 149)**; in this condition, the patient is extremely toxic and the mortality rate is high. Within 48 h, rapid tissue invasion and systemic toxicity associated with hemolysis and death ensue. The distinction between this entity and clostridial myonecrosis is made by muscle biopsy. Necrotizing fasciitis caused by community-acquired MRSA also has been reported.

Clostridial Myonecrosis

Myonecrosis is often associated with trauma or surgery but can develop spontaneously (See also Chap. 149). The incubation period is usually 12–24 h long, and massive necrotizing gangrene develops within hours of onset. Systemic toxicity, shock, and death can occur within 12 h. The patient's pain and toxic appearance are out of proportion to physical findings. On examination, the patient is febrile, apathetic, tachycardic, and tachypneic and may express a feeling of impending doom. Hypotension and renal failure develop later, and hyperalertness is evident preterminally. The skin over the affected area is bronze-brown, mottled, and edematous. Bullous lesions with serosanguineous drainage and a mousy or sweet odor can develop. Crepitus can occur secondary to gas production in muscle tissue. The mortality rate is >65% for spontaneous myonecrosis, which is often associated with *Clostridium septicum* or *C. tertium* and underlying malignancy. The mortality rates associated with trunk and limb infection are 63% and 12%, respectively, and any delay in surgical treatment increases the risk of death.

NEUROLOGIC INFECTIONS WITH OR WITHOUT SEPTIC SHOCK

Bacterial Meningitis

Bacterial meningitis is one of the most common infectious disease emergencies involving the central nervous system (See also Chap. 133). Although hosts with cell-mediated immune deficiency (including transplant recipients, diabetic patients, elderly patients, and cancer patients receiving certain chemotherapeutic agents) are at particular risk for *Listeria monocytogenes* meningitis, most cases in adults are due to *S. pneumoniae* (30–60%) and *N. meningitidis* (10–35%). The classic presentation of fever, meningismus, and altered mental status is seen in only one-half to two-thirds of patients. The elderly can present without fever

or meningeal signs. Cerebral dysfunction is evidenced by confusion, delirium, and lethargy that can progress to coma. In some cases, the presentation is fulminant, with sepsis and brain edema; papilledema at presentation is unusual and suggests another diagnosis (e.g., an intracranial lesion). Focal signs, including cranial nerve palsies (IV, VI, VII), can be seen in 10–20% of cases; 50–70% of patients have bacteremia. A poor outcome is associated with coma, seizures, hypotension, a pneumococcal etiology, respiratory distress, a CSF glucose level of <0.6 mmol/L (<10 mg/dL), a CSF protein level of >2.5 g/L, a peripheral white blood cell count of <5000/μL, and a serum sodium level of <135 mmol/L. Rapid initiation of treatment is essential; the odds of an unfavorable outcome may increase by 30% for each hour that treatment is delayed. Dexamethasone is an adjunctive treatment for meningitis in adults, especially for infections caused by *S. pneumoniae*. It must be given before or with the first dose of antibiotics; otherwise, it is unlikely to improve outcomes.

Suppurative Intracranial Infections

In suppurative intracranial infections, rare intracranial lesions present along with sepsis and hemodynamic instability (See also Chap. 135). Rapid recognition of the toxic patient with central neurologic signs is crucial to improvement of the dismal prognosis of these entities. Patients with diabetes or hematologic disease may be at increased risk for these infections. *Subdural empyema* arises from the paranasal sinus in 60–70% of cases. Microaerophilic streptococci and staphylococci are the predominant etiologic organisms. The patient is toxic, with fever, headache, and nuchal rigidity. Of all patients, 75% have focal signs and 6–20% die. Despite improved survival rates, 15–44% of patients are left with permanent neurologic deficits. *Septic cavernous sinus thrombosis* follows a facial or sphenoid sinus infection; 70% of cases are due to staphylococci (including MRSA), and the remainder are due primarily to aerobic or anaerobic streptococci. Fungi have been common in some series. A unilateral or retro-orbital headache progresses to a toxic appearance and fever within days. Three-quarters of patients have unilateral periorbital edema that becomes bilateral and then progresses to ptosis, proptosis, ophthalmoplegia, and papilledema. The mortality rate is as high as 30%. *Septic thrombosis of the superior sagittal sinus* spreads from the ethmoid or maxillary sinuses and is caused by *S. pneumoniae*, other streptococci, and staphylococci. The fulminant course is characterized by headache, nausea, vomiting, rapid progression to confusion and coma, nuchal rigidity, and brainstem signs. If the sinus is totally thrombosed, the mortality rate exceeds 80%. Broad-spectrum antibiotics and early surgical intervention at the primary site of infection may improve outcomes. Anticoagulation or steroids are of uncertain benefit.

Brain Abscess

Brain abscess often occurs without systemic signs (See also Chap. 135). Almost half of patients are afebrile, and presentations are more consistent with a space-occupying lesion in the brain; 70% of patients have headache and/or altered mental status, 50% have focal neurologic signs, and 25% have papilledema. Abscesses can present as single or multiple lesions resulting from contiguous foci or hematogenous infection, such as endocarditis, or after surgery or trauma. The infection progresses over several days from cerebritis to an abscess with a mature capsule. More than half of infections are polymicrobial, with an

etiology consisting of aerobic bacteria (primarily streptococcal species) and anaerobes. Abscesses arising hematogenously are especially apt to rupture into the ventricular space, causing a sudden and severe deterioration in clinical status and a high mortality rate. Otherwise, mortality is low (<20%) but morbidity is high (30–55%). Patients presenting with stroke and a parameningeal infectious focus, such as sinusitis or otitis, may have a brain abscess, and physicians must maintain a high level of suspicion. Prognosis worsens in patients with a fulminant course, delayed diagnosis, abscess rupture into the ventricles, multiple abscesses, or abnormal neurologic status at presentation. In one study, mortality at 1 year was 19%.

Cerebral Malaria

This entity should be urgently considered if patients who have recently traveled to areas endemic for malaria present with a febrile illness and lethargy or other neurologic signs (See also Chap. 219). Fulminant malaria is caused by *Plasmodium falciparum* and is associated with temperatures of >40°C (>104°F), hypotension, jaundice, acute respiratory distress syndrome, and bleeding. By definition, any patient with a change in mental status or repeated seizure in the setting of fulminant malaria has cerebral malaria. In adults, this nonspecific febrile illness progresses to coma over several days; occasionally, coma occurs within hours and death within 24 h. Nuchal rigidity and photophobia are rare. On physical examination, symmetric encephalopathy is typical, and upper motor neuron dysfunction with decorticate and decerebrate posturing can be seen in advanced disease. Unrecognized infection results in a 20–30% mortality rate.

Intracranial and Spinal Epidural Abscesses

Spinal and intracranial epidural abscesses (SEAs and ICEAs) can result in permanent neurologic deficits, sepsis, and death (See also Chap. 434). At-risk patients include those with diabetes mellitus; IV drug use; chronic alcohol abuse; recent spinal trauma, surgery, or epidural anesthesia; and other comorbid conditions, such as HIV infection. Fungal epidural abscess and meningitis can follow epidural or paraspinal glucocorticoid injections. In the United States and Canada, where early treatment of otitis and sinusitis is typical, ICEA is rare but the number of cases of SEA is on the rise. In Africa and areas with limited access to health care, SEAs and ICEAs cause significant morbidity and mortality. ICEAs typically present as fever, mental status changes, and neck pain, while SEAs often present as fever, localized spinal tenderness, and back pain. ICEAs are typically polymicrobial, whereas SEAs are most often due to hematogenous seeding, with staphylococci the most common etiologic agent. Early diagnosis and treatment, which may include surgical drainage, minimize rates of mortality and permanent neurologic sequelae. Outcomes are worse for SEA due to MRSA, for infection at a higher vertebral-body level, for impaired neurologic status on presentation, and for dorsal rather than ventral location of the abscess. Elderly patients and persons with renal failure, malignancy, and other comorbidities also have less favorable outcomes.

OTHER FOCAL SYNDROMES WITH A FULMINANT COURSE

Infection at virtually any primary focus (e.g., osteomyelitis, pneumonia, pyelonephritis, or cholangitis) can result in bacteremia and sepsis. Lemierre's syndrome—jugular septic thrombophlebitis caused by *Fusobacterium necrophorum*—is associated with metastatic infectious emboli (primarily to the lung but

sometimes to the liver or other organs) and sepsis, with mortality rates of >15%. TSS has been associated with focal infections such as septic arthritis, peritonitis, sinusitis, and wound infection. Rapid clinical deterioration and death can be associated with destruction of the primary site of infection, as is seen in endocarditis and in infections of the oropharynx (e.g., Ludwig's angina or epiglottitis, in which edema suddenly compromises the airway).

Rhinocerebral Mucormycosis

Individuals with diabetes or immunocompromising conditions such as solid organ transplants or hematologic malignancies are at risk for invasive rhinocerebral mucormycosis (See also Chap. 213). Patients present with low-grade fever, dull sinus pain, diplopia, decreased mental status, decreased ocular motion, chemosis, proptosis, dusky or necrotic nasal turbinates, and necrotic hard-palate lesions that respect the midline. Without rapid recognition and intervention, the process continues on an inexorable invasive course, with mortality rates of 50–85% or greater. Uncontrolled diabetes and increasing age are negative prognostic factors.

Acute Bacterial Endocarditis

This entity presents with a much more aggressive course than subacute endocarditis (See also Chap. 123). Bacteria such as *S. aureus*, *S. pneumoniae*, *L. monocytogenes*, *Haemophilus* species, and streptococci of groups A, B, and G attack native valves. Native-valve endocarditis caused by *S. aureus* (including MRSA strains) is increasing, particularly in health care settings. Mortality rates range from 10% to 40%. The host may have comorbid conditions such as underlying malignancy, diabetes mellitus, IV drug use, or alcoholism. The patient presents with fever, fatigue, and malaise <2 weeks after onset of infection. On physical examination, a changing murmur and congestive heart failure may be noted. Hemorrhagic macules on palms or soles (*Janeway lesions*) sometimes develop. Petechiae, Roth's spots, splinter hemorrhages, and splenomegaly are unusual. Rapid valvular destruction, particularly of the aortic valve, results in pulmonary edema and hypotension. Myocardial abscesses can form, eroding through the septum or into the conduction system and causing life-threatening arrhythmias or high-degree conduction block. Large friable vegetations can result in major arterial emboli, metastatic infection, or tissue infarction. Older patients with *S. aureus* endocarditis are especially likely to present with nonspecific symptoms—a circumstance that delays diagnosis and worsens prognosis. Rapid intervention is crucial for a successful outcome.

Inhalational Anthrax

Inhalational anthrax, the most severe form of disease caused by *B. anthracis*, had not been reported in the United States for more than 25 years until the use of this organism as an agent of bioterrorism in 2001 (See also Chap. S2). Patients presented with malaise, fever, cough, nausea, drenching sweats, shortness of breath, and headache. Rhinorrhea was unusual. All patients had abnormal chest roentgenograms at presentation. Pulmonary infiltrates, mediastinal widening, and pleural effusions were the most common findings. Hemorrhagic meningitis was documented in 38% of these patients. Survival was more likely when antibiotics were given during the prodromal period and when multidrug regimens were used. In the absence of urgent

intervention with antimicrobial agents and supportive care, inhalational anthrax progresses rapidly to hypotension, cyanosis, and death.

Viral Respiratory Tract Illness

Viral respiratory tract illnesses can cause severe disease; several new syndromes have been described in the past decade. For patients who present with a respiratory illness and a relevant exposure and travel history, these viral illnesses must be considered and appropriate infection control measures instituted in addition to supportive care.

Avian and Swine Influenza

Human cases of avian influenza have occurred primarily in Southeast Asia, particularly Vietnam (H5N1) and China (H7N9) (See also Chap. 195). Avian influenza should be considered in patients with severe respiratory tract illness, particularly if they have been exposed to poultry. Patients present with high fever, an influenza-like illness, and lower respiratory tract symptoms; this illness can progress rapidly to bilateral pneumonia, acute respiratory distress syndrome, multiorgan failure, and death. Younger age appears to be associated with a lower risk of complications. Early antiviral treatment with neuraminidase inhibitors should be initiated along with aggressive supportive measures. Unlike avian influenza, whose human-to-human transmission has so far been rare and has not been sustained, influenza caused by a novel swine-associated A/H1N1 virus has spread rapidly throughout the world; by 2012, 214 countries had diagnosed cases of influenza A/H1N1, with 18,449 deaths. Patients most at risk of severe disease are children <5 years of age, elderly persons, patients with underlying chronic conditions, and pregnant women. Obesity also has been identified as a risk factor for severe illness. Immunosuppression and co-infection with *S. aureus* at presentation are independent risk factors for increased mortality.

SARS and MERS

Severe acute respiratory syndrome (SARS) was identified in 2002 in China but has been diagnosed in several countries, primarily in Asia. Possible animal reservoirs include bats and civets. SARS is caused by a coronavirus and is characterized by efficient human transmission but relatively low mortality. It spreads from person to person via droplets; “super-spreader” airborne events have occurred. The potential pandemic with SARS was controlled through identification and isolation of infected patients. A 3- to 7-day prodrome characterized by fever, malaise, headache, and myalgia can progress to nonproductive cough, dyspnea, and respiratory failure. The risk of contagion is low during the prodrome. Older patients and those with diabetes mellitus, chronic hepatitis B, and other comorbidities can have less favorable outcomes.

Middle East respiratory syndrome (MERS) is caused by a novel betacoronavirus and was first recognized in 2012 in Saudi Arabia. Human cases have been associated with direct and indirect contact with dromedary camels. Unlike SARS, MERS exhibits inefficient human transmission but carries a high mortality rate. As of 2015, 1180 cases had been confirmed, with 40% mortality. MERS ranges from asymptomatic infection to acute respiratory distress syndrome, multiorgan failure, and death. Elderly men with comorbidities appear to

be at highest risk for poor outcomes. Despite little documented human-to-human transmission in the community, nosocomial infection must be prevented by adherence to strict infection control practices. MERS is currently a low-level public health threat and is likely to remain so unless the virus mutates and its transmissibility increases.

Hantavirus Pulmonary Syndrome

Hantavirus pulmonary syndrome has been documented in the United States since 1993 (primarily the southwestern states, west of the Mississippi River), Canada, and South America (See also Chap. 204). Most cases occur in rural areas and are associated with exposure to rodents. Patients present with a nonspecific viral prodrome of fever, malaise, myalgias, nausea, vomiting, and dizziness that may progress to pulmonary edema, respiratory failure, and death. Hantavirus pulmonary syndrome causes myocardial depression and increased pulmonary vascular permeability; therefore, careful fluid resuscitation and use of pressor agents are crucial. Aggressive cardiopulmonary support during the first few hours of illness can be life-saving in this high-mortality syndrome. The early onset of thrombocytopenia may help distinguish this syndrome from other febrile illnesses in an appropriate epidemiologic setting.

Clostridium difficile Infection

C. difficile infection (CDI) is a toxin-mediated diarrheal syndrome that is strongly associated with prior antibiotic use. Proton-pump inhibitors have also been identified as a potential risk factor for the disease. Although most cases of CDI have occurred in the health care setting, community-onset CDI is increasing. Overall, community-onset cases occur in younger patients than nosocomial cases. Patients with community-onset CDI are less likely to have a history of antibiotic or protein-pump inhibitor use. CDI is associated with significant morbidity and mortality, particularly among older patients. The Centers for Disease Control and Prevention has reported that *C. difficile* infection is one of the top three health threats associated with antibiotic use.

SUMMARY

Acutely ill febrile patients with the syndromes discussed in this chapter require close observation, aggressive supportive measures, and—in most cases—admission to intensive care units. The most important task of the physician is to distinguish these patients from other infected febrile patients whose illness will not progress to fulminant disease. The alert physician must recognize the acute infectious disease emergency and then proceed with appropriate urgency.

FURTHER READING

Nicolasora N, Kaul DR: Infectious disease emergencies. Med Clin N Am 92:427, 2008. [PubMed: 18298987]

Seymour CW et al: Assessment of clinical criteria for sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 315:762, 2016. [PubMed: 26903335]

Simmons CP et al: Dengue. N Engl J Med 366:15, 2012.

Theilacker C et al: Overwhelming postsplenectomy infection: A prospective multicenter cohort study. Clin Infect Dis 62:871, 2016. [PubMed: 26703862]

Thigpen MC et al: Bacterial meningitis in the United States, 1998–2007. N Engl J Med 364:2016, 2011.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 118: Immunization Principles and Vaccine Use

Nancy Messonnier; Anne Schuchat; Lisa A. Jackson

INTRODUCTION

Few medical interventions of the past century can rival the effect that immunization has had on longevity, economic savings, and quality of life. Seventeen diseases are now preventable through vaccines routinely administered to children and adults in the United States (Table 118-1), and most vaccine-preventable diseases of childhood are at historically low levels (Table 118-2). Health care providers deliver the vast majority of vaccines in the United States in the course of providing routine health services and therefore play an integral role in the nation's public health system.

TABLE 118-1

Diseases Preventable with Vaccines Routinely Administered in the United States to Children and/or Adults

CONDITION	TARGET POPULATION(S) FOR ROUTINE USE
Pertussis	Children, adolescents, adults
Diphtheria	Children, adolescents, adults
Tetanus	Children, adolescents, adults
Poliomyelitis	Children
Measles	Children
Mumps	Children
Rubella, congenital rubella syndrome	Children
Hepatitis B	Children and high-risk adults
Haemophilus influenzae type b infection	Children
Hepatitis A	Children
Influenza	Children, adolescents, adults
Varicella	Children
Pneumococcal disease	Children, older adults
Meningococcal disease	Adolescents and high-risk adults
Rotavirus infection	Infants
Human papillomavirus infection, cervical and anogenital cancers	Adolescents and young adults
Zoster	Older adults

TABLE 118-2

Decline in Vaccine-Preventable Diseases in the United States Following Widespread Implementation of National Vaccine Recommendations

CONDITION	ANNUAL NO. OF PREVACCINE CASES (AVERAGE)	NO. OF CASES REPORTED IN 2016[a]	REDUCTION (%) IN CASES AFTER WIDESPREAD VACCINATION
Smallpox	29,005	0	100
Diphtheria	21,053	0	100
Measles	530,217	69	>99
Mumps	162,344	5311	97
Pertussis	200,752	15,737	92
Polio (paralytic)	16,316	0	100
Rubella	47,745	5	>99
Congenital rubella syndrome	152	1	99
Tetanus	580	33	94
Haemophilus influenzae type b infection	20,000	22[b]	>99
Hepatitis A	117,333	2500[c]	98
Hepatitis B (acute)	66,232	19,200[c]	71
Invasive pneumococcal infection: all ages	63,067	29,000[d]	54
Varicella	4,085,120	126,639[e]	97

[a]2016 reported cases unless otherwise specified. [b]An additional 11 type b infections are estimated to have occurred among 222 reports of *H. influenzae* infection caused by unknown types among children <5 years of age. [c]Data are from the CDC's Viral Hepatitis Surveillance, 2014. [d]Data are from the CDC's Active Bacterial Core Surveillance 2015 Provisional Report. [e]Data are from Morb Mortal Wkly Rep 65:1306, 2016 (2015 final data).

Source: Adapted from SW Roush et al: JAMA 298:2155, 2007 and Morb Mortal Wkly Rep 65:924, 2017.

VACCINE IMPACT

Direct and Indirect Effects

Immunizations against specific infectious diseases protect individuals against infection and thereby prevent symptomatic illnesses. Specific vaccines may blunt the severity of clinical illness (e.g., rotavirus vaccines and severe gastroenteritis) or reduce complications (e.g., zoster vaccines and postherpetic neuralgia). Some immunizations also reduce transmission of infectious disease agents from immunized people to others, thereby reducing the impact of infection spread. This indirect impact is known as *herd immunity*. The level of immunization in a population that is required to achieve indirect protection of unimmunized people varies substantially with the specific vaccine and disease.

Since childhood vaccines have become widely available in the United States, major declines in rates of vaccine-preventable diseases among both children and adults have become evident (Table 118-2). For example, vaccination of children <5 years of age against seven types of *Streptococcus*

pneumoniae led to a >90% overall reduction in invasive disease caused by those types. Among children born during 1994–2013, a series of childhood vaccines targeting 13 vaccine-preventable diseases will prevent 322 million illnesses and 732,000 deaths over the course of their lifetimes and save $1.38 trillion (U.S.).

Control, Elimination, and Eradication of Vaccine-Preventable Diseases

Immunization programs are associated with the goals of controlling, eliminating, or eradicating a disease. *Control* of a vaccine-preventable disease reduces poor illness outcomes and often limits the disruptive impacts associated with outbreaks of disease in communities, schools, and institutions. Control programs can also reduce absences from work for ill persons and for parents caring for sick children, decrease absences from school, and limit health care utilization associated with treatment visits.

Elimination of a disease is a more demanding goal than control, usually requiring the reduction to zero of cases in a defined geographic area but sometimes defined as reduction in the indigenous sustained transmission of an infection in a geographic area. As of 2016, the United States had eliminated indigenous transmission of measles, rubella, poliomyelitis, and diphtheria. Importation of pathogens from other parts of the world continues to be important, and public health efforts are intended to respond promptly to such cases in order to limit forward spread of the infectious agent.

 *Eradication* of a disease is achieved when its elimination can be sustained without the need to continue interventions. The only vaccine-preventable disease of humans that has been globally eradicated thus far is smallpox. Although smallpox vaccine is no longer given routinely, the disease has not reemerged naturally because all chains of human transmission were interrupted through earlier vaccination efforts and humans were the only natural reservoir of the virus. Currently, a major health initiative is targeting the global eradication of polio. Sustained transmission of polio has been eliminated from most nations but has not yet been interrupted in Afghanistan and Pakistan. In 2016, after Nigeria completed 2 years without a wild polio case detected, three cases were identified in a region where vaccinators have been unable to reach hundreds of thousands of children because of insurgency and the virus is likely to have been circulating undetected. Detection of a case of disease that has been targeted for eradication or elimination is considered a sentinel event that could permit the infectious agent to become reestablished in the community or region. Therefore, such episodes must be promptly reported to public health authorities.

Outbreak Detection and Control

Clusters of cases of a vaccine-preventable disease detected in an institution, a medical practice, or a community may signal important changes in the pathogen, vaccine, or environment. Several factors can give rise to increases in vaccine-preventable disease, including (1) low rates of immunization that result in an accumulation of susceptible people (e.g., measles resurgence among vaccination abstainers); (2) changes in the infectious agent that permit it to escape vaccine-induced protection (e.g., non-vaccine-type pneumococci); (3) waning of vaccine-induced immunity (e.g., pertussis among adolescents and adults vaccinated in early childhood); and (4) point-source introductions of large inocula (e.g., food-borne exposure to hepatitis A virus). Reporting episodes of outbreak-prone diseases to public health authorities can facilitate recognition of clusters that require further interventions.

Public Health Reporting

Recognition of suspected cases of diseases targeted for elimination or eradication—along with other diseases that require urgent public health interventions, such as contact tracing, administration of chemo- or immunoprophylaxis, or epidemiologic investigation for common-source exposure—is typically associated with special reporting requirements. Many diseases against which vaccines are routinely used, including measles, pertussis, *Haemophilus influenzae* type b invasive disease, and varicella, are nationally notifiable. Clinicians and laboratory staff have a responsibility to report some vaccine-preventable disease occurrences to local or state public health authorities according to specific case-definition criteria. All providers should be aware of state or city disease-reporting requirements and the best ways to contact public health authorities. A prompt response to vaccine-preventable disease outbreaks can greatly enhance the effectiveness of control measures.

Global Considerations

 Several international health initiatives currently focus on reducing vaccine-preventable diseases in regions throughout the world. The American Red Cross, the World Health Organization (WHO), the United Nations Foundation, the United Nations Children's Fund (UNICEF), and the Centers for Disease Control and Prevention (CDC) are partners in the Measles & Rubella Initiative, which targets reduction of worldwide measles deaths. During 2000–2014, global measles mortality rates declined by 78%—i.e., from an estimated 535,300 deaths in 2000 to 114,900 deaths in 2014. In 2015, the Americas became the first WHO region to be declared free of endemic transmission of rubella. Rotary International, UNICEF, the CDC, and the WHO are leading partners in the global eradication of polio, an endeavor that reduced the annual number of paralytic polio cases from

350,000 in 1988 to 74 in 2015. The GAVI Alliance and the Bill and Melinda Gates Foundation have brought substantial momentum to global efforts to reduce vaccine-preventable diseases, expanding on earlier efforts by the WHO, UNICEF, and governments in developed and developing countries.

Enhancing Immunization in Adults

Although immunization has become a centerpiece of routine pediatric medical visits, it has not been as well integrated into routine health care visits for adults. This chapter focuses on immunization principles and vaccine use in adults. Accumulating evidence suggests that immunization coverage can be increased through efforts directed at consumer-, provider-, institution-, and system-level factors. The literature suggests that the application of multiple strategies is more effective at raising coverage rates than is the use of any single strategy.

Recommendations for Adult Immunizations

The CDC's Advisory Committee on Immunization Practices (ACIP) is the main source of recommendations for administration of vaccines approved by the U.S. Food and Drug Administration (FDA) for use in children and adults in the U.S. civilian population. The ACIP is a federal advisory committee that consists of 15 voting members (experts in fields associated with immunization) appointed by the Secretary of the U.S. Department of Health and Human Services; 8 ex officio members representing federal agencies; and 30 nonvoting representatives of various liaison organizations, including major medical societies and managed-care organizations. The ACIP recommendations, which are available at *www.cdc.gov/vaccines/hcp/acip-recs/*, are harmonized to the greatest extent possible with vaccine recommendations made by other organizations, including the American College of Obstetricians and Gynecologists, the American Academy of Family Physicians, and the American College of Physicians.

Adult Immunization Schedules

Immunization schedules for adults in the United States are updated annually and can be found online (*www.cdc.gov/vaccines/schedules/hcp/adult.html*). In February, the schedules are published in *American Family Physician*, *Annals of Internal Medicine*, and *Morbidity and Mortality Weekly Report* (*www.cdc.gov/mmwr*). The adult immunization schedules for 2016 are summarized in **Fig. 118-1**. Additional information and specifications are contained in the footnotes to these schedules. In the time between annual publications, additions and changes to schedules are published in *Morbidity and Mortality Weekly Report*.

FIGURE 118-1

Recommended adult immunization schedules, United States, 2018. Additional information, including footnotes for each vaccine, contraindications, and precautions, can be found at *https://www.cdc.gov/vaccines/schedules/hcp/imz/adult.html*. The recommendations in this schedule were approved by the Centers for Disease Control and Prevention (CDC) Advisory Committee on Immunization Practices (ACIP), the American Academy of Family Physicians (AAFP), the American College of Physicians (ACP), the American College of Obstetricians and Gynecologists (ACOG), and the American College of Nurse-Midwives (ACNM). For complete statements by the ACIP, visit *www.cdc.gov/vaccines/hcp/acip-recs/*.

Recommended Adult Immunization Schedule—United States - 2016

Note: These recommendations must be read with the footnotes that follow containing number of doses, intervals between doses, and other important information.

Figure 1. Recommended immunization schedule for adults aged 19 years or older, by vaccine and age group[1]

VACCINE ▼ / AGE GROUP ►	19-21 years	22-26 years	27-49 years	50-59 years	60-64 years	≥ 65 years
Influenza*,[2]	1 dose annually					
Tetanus, diphtheria, pertussis (Td/Tdap)*,[3]	Substitute Tdap for Td once, then Td booster every 10 yrs					
Varicella*,[4]	2 doses					
Human papillomavirus (HPV) Female*,[5]	3 doses					
Human papillomavirus (HPV) Male*,[5]	3 doses					
Zoster[6]					1 dose	
Measles, mumps, rubella (MMR)*,[7]	1 or 2 doses depending on indication					
Pneumococcal 13-valent conjugate (PCV13)*,[8]	1 dose					
Pneumococcal 23-valent polysaccharide (PPSV23)[8]	1 or 2 doses depending on indication					1 dose
Hepatitis A*,[9]	2 or 3 doses depending on vaccine					
Hepatitis B*,[10]	3 doses					
Meningococcal 4-valent conjugate (MenACWY) or polysaccharide (MPSV4)*,[11]	1 or more doses depending on indication					
Meningococcal B (MenB)[11]	2 or 3 doses depending on vaccine					
Haemophilus influenzae type b (Hib)*,[12]	1 or 3 doses depending on indication					

*Covered by the Vaccine Injury Compensation Program

Recommended for all persons who

Report all clinically significant postvaccination reactions to the Vaccine Adverse Event Reporting System (VAERS). Reporting forms and instructions on filing a VAERS report are available at www.vaers.hhs.gov or by telephone, 800-822-7967.

meet the age requirement, lack documentation of vaccination, or lack evidence of past infection; zoster vaccine is recommended regardless of past episode of zoster

Recommended for persons with a risk factor (medical, occupational, lifestyle, or other indication)

No recommendation

Information on how to file a Vaccine Injury Compensation Program claim is available at www.hrsa.gov/vaccinecompensation or by telephone, 800-338-2382. To file a claim for vaccine injury, contact the U.S. Court of Federal Claims, 717 Madison Place, N.W., Washington, D.C. 20005; telephone, 202-357-6400.

Additional information about the vaccines in this schedule, extent of available data, and contraindications for vaccination is also available at www.cdc.gov/vaccines or from the CDC-INFO Contact Center at 800-CDC-INFO (800-232-4636) in English and Spanish, 8:00 a.m. - 8:00 p.m. Eastern Time, Monday - Friday, excluding holidays.

Use of trade names and commercial sources is for identification only and does not imply endorsement by the U.S. Department of Health and Human Services.

The recommendations in this schedule were approved by the Centers for Disease Control and Prevention's (CDC) Advisory Committee on Immunization Practices (ACIP), the American Academy of Family Physicians (AAFP), the America College of Physicians (ACP), the American College of Obstetricians and Gynecologists (ACOG) and the American College of Nurse-Midwives (ACNM).

Figure 2. Vaccines that might be indicated for adults aged 19 years or older based on medical and other indications[1]

VACCINE ▼ INDICATION ►	Pregnancy	Immuno-compromising conditions (excluding HIV infection)[4,6,7,8,13]	HIV infection CD4+ count (cells/μL)[4,6,7,8,13] < 200	HIV infection CD4+ count (cells/μL)[4,6,7,8,13] ≥ 200	Men who have sex with men (MSM)	Kidney failure, end-stage renal disease, on hemodialysis	Heart disease, chronic lung disease, chronic alcoholism	Asplenia and persistent complement component deficiencies[8,11,12]	Chronic liver disease	Diabetes	Healthcare personnel
Influenza*[,2]	1 dose annually										
Tetanus, diphtheria, pertussis (Td/Tdap)*[,3]	1 dose Tdap each pregnancy	Substitute Tdap for Td once, then Td booster every 10 yrs									
Varicella*[,4]	Contraindicated			2 doses							
Human papillomavirus (HPV) Female*[,5]		3 doses through age 26 yrs				3 doses through age 26 yrs					
Human papillomavirus (HPV) Male*[,5]		3 doses through age 26 yrs				3 doses through age 21 yrs					
Zoster[6]	Contraindicated			1 dose							
Measles, mumps, rubella (MMR)*[,7]	Contraindicated			1 or 2 doses depending on indication							
Pneumococcal 13-valent conjugate (PCV13)*[,8]		1 dose									
Pneumococcal polysaccharide (PPSV23)[8]		1, 2, or 3 doses depending on indication									
Hepatitis A*[,9]		2 or 3 doses depending on vaccine									
Hepatitis B*[,10]		3 doses									
Meningococcal 4-valent conjugate (MenACWY) or polysaccharide (MPSV4)*[,11]		1 or more doses depending on indication									
Meningococcal B (MenB)[11]		2 or 3 doses depending on vaccine									
Haemophilus influenzae type b (Hib)*[,12]		3 doses post-HSCT recipients only		1 dose							

*Covered by the Vaccine Injury Compensation Program

Recommended for all persons who meet the age requirement, lack documentation of vaccination, or lack evidence of past infection; zoster vaccine is recommended regardless of past episode of zoster

Recommended for persons with a risk factor (medical, occupational, lifestyle, or other indication)

No recommendation

Contraindicated

U.S. Department of Health and Human Services
Centers for Disease Control and Prevention

These schedules indicate the recommended age groups and medical indications for which administration of currently licensed vaccines is commonly recommended for adults aged ≥19 years, as of February 2016. For all vaccines being recommended on the Adult Immunization Schedule: a vaccine series does not need to be restarted, regardless of the time that has elapsed between doses. Licensed combination vaccines may be used whenever any components of the combination are indicated and when the vaccine's other components are not contraindicated. For detailed recommendations on all vaccines, including those used primarily for travelers or that are issued during the year, consult the manufacturers' package inserts and the complete statements from the Advisory Committee on Immunization Practices (www.cdc.gov/vaccines/hcp/acip-recs/index.html). Use of trade names and commercial sources is for identification only and does not imply endorsement by the U.S. Department of Health and Human Services.

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IMMUNIZATION PRACTICE STANDARDS

Administering immunizations to adults involves a number of processes, such as deciding whom to vaccinate, assessing vaccine contraindications and precautions, providing vaccine information statements (VISs), ensuring appropriate storage and handling of vaccines, administering vaccines, and maintaining vaccine records. In addition, provider reporting of adverse events that follow vaccination is an essential component of the vaccine safety monitoring system. In 2014, the standards for adult immunization were revised to focus on vaccinating adults at every opportunity.

Deciding Whom to Vaccinate

Every effort should be made to ensure that adults receive all indicated vaccines as expeditiously as possible. When adults present for care, their immunization history should be assessed and recorded, and this information should be used to identify needed vaccinations according to the most current version of the adult immunization schedule. Decision-support tools incorporated into electronic health records can provide prompts for needed vaccinations. Standing orders, which are often used for routinely indicated vaccines (e.g., influenza and pneumococcal vaccines), permit a nurse or another approved licensed practitioner to administer vaccines without a specific physician order, thus lowering barriers to adult immunization.

Assessing Contraindications and Precautions

Before vaccination, all patients should be screened for contraindications and precautions. A *contraindication* is a condition that increases the risk of a serious adverse reaction to vaccination. A vaccine should not be administered when a contraindication is documented. For example, a history of an anaphylactic reaction to a dose of vaccine or to a vaccine component is a contraindication for further doses. A *precaution* is a condition that may increase the risk of an adverse event or that may compromise the ability of the vaccine to evoke immunity (e.g., administering measles vaccine to a person who has recently received a blood transfusion and may consequently have transient passive immunity to measles virus). Normally, a vaccine is not administered when a precaution is noted. However, situations may arise when the benefits of vaccination outweigh the estimated risk of an adverse event, and the provider may decide to vaccinate the patient despite the precaution.

In some cases, contraindications and precautions are temporary and may lead to mere deferral of vaccination until a later time. For example, moderate or severe acute illness with or without fever is generally considered a transient precaution to vaccination and results in postponement of vaccine administration until the acute phase has resolved; thus the superimposition of adverse effects of vaccination on the underlying illness and the mistaken attribution of a manifestation of the underlying illness to the vaccine are avoided. Contraindications and precautions to vaccines licensed in the United States for use in civilian adults are summarized in **Table 118-3**. It is important to recognize conditions that are *not* contraindications in order not to miss opportunities for vaccination. For example, in most cases, mild acute illness (with or without fever), a history of a mild to moderate local reaction to a previous dose of the vaccine, and breast-feeding are not contraindications to vaccination.

TABLE 118-3

Contraindications and Precautions for Commonly Used Vaccines in Adults

VACCINE FORMULATION	CONTRAINDICATIONS AND PRECAUTIONS
All vaccines	**Contraindication** Severe allergic reaction (e.g., anaphylaxis) after a previous vaccine dose or to a vaccine component **Precaution** Moderate or severe acute illness with or without fever. Defer vaccination until illness resolves.
Td	**Precautions** GBS within 6 weeks after a previous dose of TT-containing vaccine History of Arthus-type hypersensitivity reactions after a previous dose of TD- or DT-containing vaccines (including MCV4). Defer vaccination until at least 10 years have elapsed since the last dose. History of severe allergic reaction to dry natural rubber (latex) (certain formulations; syringe; see text)
Tdap	**Contraindication** History of encephalopathy (e.g., coma or prolonged seizures) not attributable to another identifiable cause within 7 days of administration of a vaccine with pertussis components, such as DTaP or Tdap **Precautions** GBS within 6 weeks after a previous dose of TT-containing vaccine Progressive or unstable neurologic disorder, uncontrolled seizures, or progressive encephalopathy. Defer vaccination until a treatment regimen has been established and the condition has stabilized. History of Arthus-type hypersensitivity reactions after a previous dose of TT- or DT-containing vaccines (including MCV4). Defer vaccination until at least 10 years have elapsed since the last dose. History of severe allergic reaction to dry natural rubber (latex) (syringe; see text)
HPV	**Contraindications** History of immediate hypersensitivity to yeast (for Gardasil) History of severe allergic reaction to dry natural rubber (latex) (certain formulations; see text) **Precaution** Pregnancy (If a woman is found to be pregnant after initiation of the vaccination series, the remainder of the 3-dose regimen should be delayed until after completion of the pregnancy. If a vaccine dose has been administered during pregnancy, no intervention is needed. Exposure to Gardasil during pregnancy should be reported to Merck at 800-986-8999; exposure to Cervarix during pregnancy should be reported to GlaxoSmithKline at 888-452-9622.)
MMR	**Contraindications** History of immediate hypersensitivity reaction to gelatin[a] or neomycin Pregnancy Known severe immunodeficiency (e.g., hematologic and solid tumors; chemotherapy; congenital immunodeficiency; long-term immunosuppressive therapy; severe immunocompromise due to HIV infection) **Precautions** Recent receipt (within 11 months) of antibody-containing blood product History of thrombocytopenia or thrombocytopenic purpura
Varicella	**Contraindications** Pregnancy Known severe immunodeficiency History of immediate hypersensitivity reaction to gelatin[a] or neomycin **Precaution** Recent receipt (within 11 months) of antibody-containing blood product

VACCINE FORMULATION	CONTRAINDICATIONS AND PRECAUTIONS
Influenza, inactivated, injectable	**Contraindication** History of severe allergic reaction to dry natural rubber (latex) (certain formulations; see text) **Precaution**[b] History of GBS within 6 weeks after a previous influenza vaccine dose
Influenza, live attenuated nasal spray	**Contraindications**[b] Age ≥50 years Pregnancy Immunosuppression, including that caused by medications or by HIV infection; known severe immunodeficiency (e.g., hematologic and solid tumors; chemotherapy; congenital immunodeficiency; long-term immunosuppressive therapy; severe immunocompromise due to HIV infection) Certain chronic medical conditions, such as diabetes mellitus; chronic pulmonary disease (including asthma); chronic cardiovascular disease (except hypertension); renal, hepatic, neurologic/neuromuscular, hematologic, or metabolic disorders Close contact with severely immunosuppressed persons who require a protected environment, such as isolation in a bone marrow transplantation unit Close contact with persons with lesser degrees of immunosuppression (e.g., persons receiving chemotherapy or radiation therapy who are not being cared for in a protective environment; persons with HIV infection) is *not* a contraindication or a precaution. Health care personnel in neonatal intensive care units or oncology clinics may receive live attenuated influenza vaccine. **Precautions** History of GBS within 6 weeks of a previous influenza vaccine dose Receipt of specific antiviral agents (i.e., amantadine, rimantadine, zanamivir, or oseltamivir) within 48 h before vaccination
Pneumococcal polysaccharide	None, other than those listed for all vaccines
Pneumococcal conjugate	None, other than those listed for all vaccines
Hepatitis A	**Contraindication** History of severe allergic reaction to dry natural rubber (latex) (syringe; see text)
Hepatitis B	**Contraindications** History of immediate hypersensitivity to yeast History of severe allergic reaction to dry natural rubber (latex) (syringe; see text)
Meningococcal conjugate	None, other than those listed for all vaccines
Meningococcal polysaccharide	**Contraindication** History of severe allergic reaction to dry natural rubber (latex)
Serogroup B meningococcal	**Contraindication** History of severe allergic reaction to dry natural rubber (latex) (certain formulations; syringe; see text)
Zoster	**Contraindications** Pregnancy Known severe immunodeficiency History of immediate hypersensitivity reaction to gelatin[a] or neomycin **Precaution** Receipt of specific antiviral agents (i.e., acyclovir, famciclovir, or valacyclovir) within 24 h before vaccination

[a]Extreme caution must be exercised in administering MMR, varicella, or zoster vaccine to persons with a history of anaphylactic reaction to gelatin or gelatin-containing products. Before administration, skin testing for sensitivity to gelatin can be considered. However, no specific protocols for this purpose have been published. [b]History of severe allergic reaction (e.g., anaphylaxis) to egg is a labeled contraindication to the use of inactivated influenza vaccine and live attenuated influenza vaccine. However, CDC's Advisory Committee on Immunization Practices recommends that any licensed, recommended, and appropriate inactivated influenza vaccine or recombinant influenza vaccine may be administered to persons with egg allergy of any severity *(www.cdc.gov/vaccines/hcp/acip-recs/vacc-specific/flu.html)*.

Abbreviations: DT, diphtheria toxoid; DTaP, diphtheria, tetanus, and pertussis; GBS, Guillain-Barré syndrome; HPV, human papillomavirus; MCV4, quadrivalent meningococcal conjugate vaccine; MMR, measles, mumps, and rubella; Td, tetanus and diphtheria toxoids; Tdap, tetanus and diphtheria toxoids and acellular pertussis; TT, tetanus toxoid.

History of Immediate Hypersensitivity to a Vaccine Component

A severe allergic reaction (e.g., anaphylaxis) to a previous dose of a vaccine or to one of its components is a contraindication to vaccination. While most vaccines have many components, substances to which individuals are most likely to have had a severe allergic reaction include egg protein, gelatin, and yeast. In addition, although natural rubber (latex) is not a vaccine component, some vaccines are supplied in vials or syringes that contain natural rubber latex. These vaccines can be identified by the product insert and should not be administered to persons who report a severe (anaphylactic) allergy to latex unless the benefit of vaccination clearly outweighs the risk for a potential allergic reaction. The much more common local or contact hypersensitivity to latex, such as to medical gloves (which contain synthetic latex that is not linked to allergic reactions), is *not* a contraindication to administration of a vaccine supplied in a vial or syringe that contains natural rubber latex. Vaccines routinely indicated for adults that, as of February 2015, were sometimes supplied in a vial or syringe containing natural rubber include Havrix hepatitis A vaccine (syringe); Vaqta hepatitis A vaccine (vial and syringe); Engerix-B hepatitis B vaccine (syringe); Recombivax HB hepatitis B vaccine (vial); Cervarix HPV vaccine (syringe); Fluvirin, Agriflu (syringe), and Flucelvax (syringe) influenza vaccines; Adacel and Boostrix Tdap (tetanus and diphtheria toxoids and acellular pertussis) vaccines (syringe); Td (tetanus and diphtheria toxoids) vaccines (syringe); Twinrix hepatitis A and B vaccine (syringe); Menomune meningococcal polysaccharide vaccine (vial); and Bexsero meningococcal serogroup B vaccine (syringe).

Pregnancy

Live-virus vaccines are contraindicated during pregnancy because of the hypothetical risk that vaccine virus replication will cause congenital infection or have other adverse effects on the fetus. Most live-virus vaccines, including varicella vaccine, are not secreted in breast milk; therefore, breast-feeding is not a contraindication for live-virus or other vaccines. Pregnancy is not a contraindication to administration of inactivated vaccines, but most are avoided during pregnancy because relevant safety data are limited. Two inactivated vaccines, Tdap vaccine and inactivated influenza vaccine, are routinely recommended for pregnant women in the United States. Tdap vaccine is recommended during each pregnancy, regardless of prior vaccination status, in order to prevent pertussis in neonates. Annual influenza vaccination is recommended for all persons 6 months of age and older, regardless of pregnancy status. Some other inactivated vaccines, such as meningococcal vaccines, may be given to pregnant women in certain circumstances.

Immunosuppression

Live-virus vaccines elicit an immune response due to replication of the attenuated (weakened) vaccine virus that is contained by the recipient's immune system. In persons with compromised immune function, enhanced replication of vaccine viruses is possible and could lead to disseminated infection with the vaccine virus. For this reason, live-virus vaccines are contraindicated for persons with severe immunosuppression, the definition of which may vary with the vaccine. Severe immunosuppression may be caused by many disease conditions, including HIV infection and hematologic or generalized malignancy. In some of these conditions, all affected persons are severely immunocompromised. In others (e.g., HIV infection), the degree to which the immune system is compromised depends on the severity of the condition, which in turn depends on the stage of disease or treatment. For example, measles-mumps-rubella (MMR) vaccine may be given to HIV-infected persons who are not severely immunocompromised. Severe immunosuppression may also be due to therapy with immunosuppressive agents, including high-dose glucocorticoids. In this situation, the dose, duration, and route of administration may influence the degree of immunosuppression.

VACCINE INFORMATION STATEMENTS

A VIS is a one-page (two-sided) information sheet produced by the CDC that informs vaccine recipients (or their parents or legal representatives) about the benefits and risks of a vaccine. VISs are mandated by the National Childhood Vaccine Injury Act (NCVIA) of 1986 and—whether the vaccine recipient is a child or an adult—must be provided for any vaccine covered by the Vaccine Injury Compensation Program. As of July 2016, vaccines that are covered by the NCVIA and that are licensed for use in adults include Td, Tdap, hepatitis A, hepatitis B, human papillomavirus, inactivated

influenza, live intranasal influenza, MMR, pneumococcal conjugate, meningococcal conjugate, serogroup B meningococcal, polio, and varicella vaccines. When combination vaccines for which no separate VIS exists are given (e.g., hepatitis A and B combination vaccine), all relevant VISs should be provided. VISs also exist for some vaccines not covered by the NCVIA, such as pneumococcal polysaccharide, Japanese encephalitis, rabies, herpes zoster, typhoid, anthrax, and yellow fever vaccines. The use of these VISs is encouraged but is not mandated.

All current VISs are available on the Internet at two websites: the CDC's Vaccines & Immunizations site *(www.cdc.gov/vaccines/hcp/vis/)* and the Immunization Action Coalition's site *(www.immunize.org/vis/)*. (The latter site also includes translations of the VISs.) VISs from these sites can be downloaded and printed.

STORAGE AND HANDLING

Injectable vaccines are packaged in multidose vials, single-dose vials, or manufacturer-filled single-dose syringes. The live attenuated nasal-spray influenza vaccine is packaged in single-dose sprayers. Oral typhoid vaccine is packaged in capsules. Some vaccines, such as MMR, varicella, zoster, and meningococcal polysaccharide vaccines, come as lyophilized (freeze-dried) powders that must be reconstituted (i.e., mixed with a liquid diluent) before use. The lyophilized powder and the diluent come in separate vials. Diluents are not interchangeable but rather are specifically formulated for each type of vaccine; only the specific diluent provided by the manufacturer for each type of vaccine should be used. Once lyophilized vaccines have been reconstituted, their shelf-life is limited and they must be stored under appropriate temperature and light conditions. For example, varicella and zoster vaccines must be protected from light and administered within 30 min of reconstitution; MMR vaccine likewise must be protected from light but can be used up to 8 h after reconstitution. Single-dose vials of meningococcal polysaccharide vaccine must be used within 30 min of reconstitution, while multidose vials must be used within 35 days.

Vaccines are stored either at refrigerator temperature (2–8°C) or at freezer temperature (–15°C or colder). In general, inactivated vaccines (e.g., inactivated influenza, pneumococcal polysaccharide, and meningococcal conjugate vaccines) are stored at refrigerator temperature, while vials of lyophilized-powder live-virus vaccines (e.g., varicella, zoster, and MMR vaccines) are stored at freezer temperature. Diluents for lyophilized vaccines may be stored at refrigerator or room temperature. Live attenuated influenza vaccine—a live-virus liquid formulation administered by nasal spray—is stored at refrigerator temperature.

Vaccine storage and handling errors can result in the loss of vaccines worth millions of dollars, and administration of improperly stored vaccines may elicit inadequate immune responses in patients. To improve the standard of vaccine storage and handling practices, the CDC has published detailed guidance (available at *www.cdc.gov/vaccines/hcp/admin/storage/toolkit/storage-handling-toolkit.pdf*). For vaccine storage, the CDC recommends stand-alone units—i.e., self-contained units that either refrigerate or freeze but do not do both—as these units maintain the required temperatures better than combination refrigerator/freezer units. Dormitory-style combined refrigerator/freezer units should never be used for vaccine storage.

The temperature of refrigerators and freezers used for vaccine storage must be monitored and recorded at least twice each workday. Ideally, continuous thermometers that measure and record temperature all day and all night are used, and minimal and maximal temperatures are read and documented each workday. The CDC recommends the use of calibrated digital thermometers with a probe in thermal-buffered material; more detailed information on specifications of storage units and temperature-monitoring devices is provided at the link given above.

ADMINISTRATION OF VACCINES

Most parenteral vaccines recommended for routine administration to adults in the United States are given by either the IM or the SC route; one influenza vaccine formulation approved for use in adults 18–64 years of age is given intradermally. Live-virus vaccines such as varicella, zoster, and MMR are given SC. Most inactivated vaccines are given IM. The 23-valent pneumococcal polysaccharide vaccine may be given either IM or SC, but IM administration is preferred because it is associated with a lower risk of injection-site reactions.

Vaccines given to adults by the SC route are administered with a 5/8-inch needle into the upper outer-triceps area. Vaccines administered to adults by the IM route are injected into the deltoid muscle **(Fig. 118-2)** with a needle whose length should be selected on the basis of the recipient's sex and weight to ensure adequate penetration into the muscle. Current guidelines indicate that, for men and women weighing <152 lbs (<70 kg), a 1-inch needle is sufficient; for women weighing 152–200 lbs (70–90 kg) and men weighing 152–260 lbs (70–118 kg), a 1- to 1.5-inch needle is needed; and for women weighing >200 lbs (>90 kg) and men weighing >260 lbs (>118 kg), a 1.5-inch needle is required. Additional illustrations of vaccine injection locations and techniques may be found at *www.immunize.org/catg.d/p2020a.pdf*.

FIGURE 118-2

Technique for IM administration of vaccine. *(Photo credit: James Gathany, Centers for Disease Control and Prevention; accessible at Public Health Image Library, www.cdc.gov. PHIL ID#9420.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Aspiration, the process of pulling back on the plunger of the syringe after skin penetration but prior to injection, is not necessary because no large blood vessels are present at the recommended vaccine injection sites.

Multiple vaccines can be administered at the same visit; indeed, administration of all needed vaccines at one visit is encouraged. Studies have shown that vaccines are as effective when administered simultaneously as they are individually, and simultaneous administration of multiple vaccines is not associated with an increased risk of adverse effects. If more than one vaccine must be administered in the same limb, the injection sites should be separated by 1–2 inches so that any local reactions can be differentiated. If a vaccine and an immune globulin preparation are administered simultaneously (e.g., Td vaccine and tetanus immune globulin), a separate anatomic site should be used for each injection.

For certain vaccines (e.g., HPV vaccine and hepatitis B vaccine), multiple doses are required for an adequate and persistent antibody response. The recommended vaccination schedule specifies the interval between doses. Many adults who receive the first dose in a multiple-dose vaccine series do not complete the series or do not receive subsequent doses within the recommended interval; this lack of adherence to protocol compromises vaccine efficacy and/or the duration of protection. Providers should implement recall systems that will prompt patients to return for subsequent doses in a vaccination series at the appropriate intervals. With the exception of oral typhoid vaccination, an interruption in the schedule does not require restarting of the entire series or the addition of extra doses.

Syncope may follow vaccination, especially in adolescents and young adults. Serious injuries, including skull fracture and cerebral hemorrhage, have occurred. Adolescents and adults should be seated or lying down during vaccination. The majority of reported syncope episodes after vaccination occur within 15 min. The ACIP recommends that vaccine providers strongly consider observing patients, particularly adolescents, with patients seated or lying down for 15 min after vaccination. If syncope develops, patients should be observed until the symptoms resolve.

Anaphylaxis is a rare complication of vaccination. All facilities providing immunizations should have an emergency kit containing aqueous epinephrine for administration in the event of a systemic anaphylactic reaction.

MAINTENANCE OF VACCINE RECORDS

All vaccines administered should be fully documented in the patient's permanent medical record. Documentation should include the date of administration, the name or common abbreviation of the vaccine, the vaccine lot number and manufacturer, the administration site, the VIS edition, the date the VIS was provided, and the name, address, and title of the person who administered the vaccine. Increasing use of two-dimensional bar codes on vaccine vials and syringes that can be scanned for data entry into compatible electronic medical records and immunization information systems may facilitate more complete and accurate recording of required information.

VACCINE SAFETY MONITORING AND ADVERSE EVENT REPORTING

Prelicensure Evaluations of Vaccine Safety

Before vaccines are licensed by the FDA, they are evaluated in clinical trials with volunteers. These trials are conducted in three progressive phases. Phase 1 trials are small, usually involving fewer than 100 volunteers. Their purposes are to provide a basic evaluation of safety and to identify common adverse events. Phase 2 trials, which are larger and may involve several hundred participants, collect additional information on safety and are usually designed to evaluate immunogenicity as well. Data gained from phase 2 trials can be used to determine the composition of the vaccine, the number of doses required, and a profile of common adverse events. Vaccines that appear promising are evaluated in phase 3 trials, which

typically involve several hundred to several thousand volunteers and are generally designed to demonstrate vaccine efficacy and provide additional information on vaccine safety.

Postlicensure Monitoring of Vaccine Safety

After licensure, a vaccine's safety is assessed by several mechanisms. The NCVIA of 1986 requires health care providers to report certain adverse events that follow vaccination. As a mechanism for that reporting, the Vaccine Adverse Event Reporting System (VAERS) was established in 1990 and is jointly managed by the CDC and the FDA. This safety surveillance system collects reports of adverse events associated with vaccines currently licensed in the United States. *Adverse events* are defined as untoward events that occur after immunization and that might be caused by the vaccine product or vaccination process. While the VAERS was established in response to the NCVIA, any adverse event following vaccination—whether in a child or an adult, and whether or not it is believed to have actually been caused by vaccination—may be reported through the VAERS. The adverse events that health care providers are required to report are listed in the reportable-events table on the VAERS website at *vaers.hhs.gov/reportable.htm*. During 2011–2014, approximately 30,000 VAERS reports were filed annually, with ~7% reporting serious events resulting in hospitalization, life-threatening illness, disability, or death.

Anyone can file a VAERS report, including health care providers, manufacturers, and vaccine recipients or their parents or guardians. VAERS reports may be submitted online (*https://vaers.hhs.gov/reportevent.html*) or by completing a paper form requested by email (info@vaers.org) or phone (800-822-7967). The VAERS form asks for the following information: the type of vaccine received; the timing of vaccination; the time of onset of the adverse event; and the recipient's current illnesses or medications, history of adverse events following vaccination, and demographic characteristics (e.g., age and sex). This information is entered into a database. The individual who reported the adverse event then receives a confirmation letter by mail with a VAERS identification number that can be used if additional information is submitted later. In selected cases of serious adverse reaction, the patient's recovery status may be followed up at 60 days and 1 year after vaccination. The FDA and the CDC have access to VAERS data and use this information to monitor vaccine safety and conduct research studies. VAERS data (minus personal information) are also available to the public.

While the VAERS provides useful information on vaccine safety, this passive reporting system has important limitations. One is that events following vaccination are merely reported; the system cannot assess whether a given type of event occurs more often than expected after vaccination. A second is that event reporting is incomplete and is biased toward events that are believed to be more likely to be due to vaccination and that occur relatively soon after vaccination. To obtain more systematic information on adverse events occurring in both vaccinated and unvaccinated persons, the Vaccine Safety Datalink project was initiated in 1991. Directed by the CDC, this project includes nine managed-care organizations in the United States; member databases include information on immunizations, medical conditions, demographics, laboratory results, and medication prescriptions. The Department of Defense oversees a similar system monitoring the safety of immunizations among active-duty military personnel. In addition, postlicensure evaluations of vaccine safety may be conducted by the vaccine manufacturer. In fact, such evaluations are often required by the FDA as a condition of vaccine licensure.

CONSUMER ACCESS TO AND DEMAND FOR IMMUNIZATION

By removing barriers to the consumer or patient, providers and health care institutions can improve vaccine use. Financial barriers have traditionally been important constraints, particularly among uninsured adults. Even for insured adults, out-of-pocket costs associated with newer, more expensive adult vaccines (e.g., zoster vaccine) are an obstacle to be overcome. After influenza vaccine was included by Medicare for all beneficiaries in 1993, coverage among persons ≥65 years of age doubled (from ~30% in 1989 to >60% in 1997). Other strategies that enhance patients' access to vaccination include extended office hours (e.g., evening and weekend hours) and scheduled vaccination-only clinics where waiting times are reduced. Provision of vaccines outside the "medical home" (e.g., through occupational clinics, universities, pharmacies, and retail settings) can expand access for adults who do not make medical visits frequently. Increasing proportions of adults are being vaccinated in these settings.

Health promotion efforts aimed at increasing the demand for immunization are common. Direct-to-consumer advertising by pharmaceutical companies has been used for some newer adolescent and adult vaccines. Efforts to raise consumer demand for vaccines have not increased immunization rates unless implemented in conjunction with other strategies that target strengthening of provider practices or reduction of consumer barriers. Attitudes and beliefs related to vaccination can be considerable impediments to consumer demand. Many adults view vaccines as important for children but are less familiar with vaccinations targeting disease prevention in adults. Several vaccines are recommended for adults with certain medical risk factors, but self-identification as a high-risk individual is relatively rare. Communication research suggests that adults are motivated to get vaccines to protect their own health and many would get vaccinated to protect loved ones. Adults with chronic conditions are more likely to be aware that they need to protect their own health. Some vaccines are explicitly recommended for persons at relatively low risk of serious complications, with the goal of reducing the risk of transmission to higher-risk contacts. For example, for protection of newborns, vaccinations against influenza and pertussis are recommended for pregnant women.

STRATEGIES FOR PROVIDERS AND HEALTH CARE FACILITIES

Recommendation from the Provider

Health care providers can have great influence on patients with regard to immunization. A recommendation from a doctor or nurse carries more weight than do recommendations from professional societies or endorsements by celebrities. Providers should be well informed about vaccine risks and benefits so that they can address patients' common concerns. The CDC, the American College of Physicians, and the American Academy of Family Physicians review and update the schedule for adult immunization on an annual basis and have developed educational materials to facilitate provider–patient discussions about vaccination (*www.cdc.gov/vaccines/hcp.htm*).

System Supports

Medical offices can incorporate a variety of methods to ensure that providers consistently offer specific immunizations to patients with indications for specific vaccines. Decision-support tools have been incorporated into some electronic health records to alert the provider when specific vaccines are indicated. Manual or automated reminders and standing orders have been discussed (see "Deciding Whom to Vaccinate," above) and have consistently improved vaccination coverage in both office and hospital settings. Most clinicians' estimates of their own performance diverge from objective measurements of their patients' immunization coverage; quantitative assessment and feedback have been shown in pediatric and adolescent practices to increase immunization performance significantly. Some health plans have instituted incentives for providers with high rates of immunization coverage. Specialty providers, including obstetrician–gynecologists, may be the only providers serving some high-risk patients with indications for selected vaccines (e.g., Tdap, influenza, or pneumococcal polysaccharide vaccine).

Immunization Requirements

Vaccination against selected communicable diseases is required for attendance at many universities and colleges as well as for service in the U.S. military or in some occupational settings (e.g., child care, laboratory, veterinary, and health care). Immunizations are recommended and sometimes required for travel to certain countries **(Chap. 119)**.

Vaccination of Health Care Staff

A particular area of focus for medical settings is vaccination of health care workers, including those with and without direct patient-care responsibilities. The Joint Commission (which accredits health care organizations), the CDC's Healthcare Infection Control Practices Advisory Committee, and the ACIP all recommend influenza vaccination of all health care personnel; recommendations also focus on requiring documentation of declination for providers who do not accept annual influenza vaccination. As part of their participation in the Centers for Medicare and Medicaid Services' Hospital Inpatient Quality Reporting program, acute-care hospitals are required to report the proportion of their health care personnel who have received seasonal influenza vaccine. Some institutions and jurisdictions have added mandates on influenza vaccination of health care workers and have expanded on earlier requirements related to vaccination or proof of immunity for hepatitis B, measles, mumps, rubella, and varicella.

VACCINATION IN NONMEDICAL SETTINGS

Receipt of vaccination in medical offices is most frequent among young children and adults ≥65 years of age. Patients in these age groups make more office visits and are more likely to receive care in a consistent "medical home" than are older children, adolescents, and nonelderly adults. Vaccination outside the medical home can expand access to those whose health care visits are limited and reduce the burden on busy clinical practices. In some locations, financial constraints related to inventory and storage requirements have led providers to stock few or no vaccines. Outside private office and hospital settings, vaccination may also occur at health department venues, workplaces, retail sites (including pharmacies and supermarkets), and schools or colleges.

When vaccines are given in nonmedical settings, it remains important for standards of immunization practice to be followed. Consumers should be provided with information on how to report adverse events (e.g., via provision of a VIS), and procedures should ensure that documentation of vaccine administration is forwarded to the primary care provider and the state or city public health immunization registry. Detailed documentation may be required for employment, school attendance, and travel. Personalized health records can help consumers keep track of their immunizations, and some occupational health clinics have incorporated automated immunization reports that help employees stay up-to-date with recommended vaccinations. Some pharmacy chain establishments are using automated systems to report immunization information to the state or local immunization information system.

PERFORMANCE MONITORING

Tracking of immunization coverage at national, state, institution, and practice levels can yield feedback to practitioners and programs and facilitate quality improvement. Healthcare Effectiveness Data and Information Set (HEDIS) measures related to adult immunization facilitate comparison of health plans. The CDC's National Immunization Survey and National Health Interview Survey provide selected information on immunization coverage among adults and track progress toward achievement of Healthy People 2020 targets for immunization coverage. Influenza and pneumococcal vaccine coverage rates have been higher among persons ≥65 years of age (60–70%) than among high-risk 18- to 64-year-olds. Figures on state-specific immunization coverage with pneumococcal polysaccharide and influenza vaccines (as measured through the CDC's Behavioral Risk Factor Surveillance System) reveal substantial geographic variation in coverage. There are persistent disparities in adult immunization coverage rates between whites and racial and ethnic minorities. In contrast, racial and economic disparities in immunization of young children have been dramatically reduced during the past 20 years. Much of this progress is attributed to the Vaccines for Children Program, which since 1994 has entitled uninsured children to receive free vaccines.

FUTURE TRENDS

Although most vaccines developed in the twentieth century targeted common acute infectious diseases of childhood, more recently developed vaccines prevent chronic conditions prevalent among adults. Hepatitis B vaccine prevents hepatitis B–related cirrhosis and hepatocellular carcinoma, and HPV vaccine prevents some types of cervical cancer, genital warts, and anogenital cancers and may also prevent some oropharyngeal cancers. A new herpes zoster subunit vaccine that was licensed in 2017 should substantially improve protection against zoster and postherpetic neuralgia. New targets of vaccine development and research may further broaden the definition of vaccine-preventable disease. Research is ongoing on vaccines to prevent insulin-dependent diabetes mellitus, nicotine addiction, and Alzheimer's disease. Expanding strategies for vaccine development are incorporating molecular approaches such as DNA, vector, and peptide vaccines. New technologies, such as the use of transdermal and other needle-less routes of administration, are being applied to vaccine delivery.

FURTHER READING

Centers for Disease Control and Prevention: *Epidemiology and Prevention of Vaccine-Preventable Diseases*, 13th ed. J Hamborsky et al (eds). Washington DC, Public Health Foundation, 2015.

McNeil MM et al: The vaccine safety datalink: Successes and challenges monitoring vaccine safety. Vaccine 32:5390, 2014. [PubMed: 25108215]

NATIONAL VACCINE ADVISORY COMMITTEE: Recommendations from the National Vaccine Advisory Committee: Standards for Adult Immunization Practice. Public Health Rep 129:115, 2014.

Plotkin SA et al (eds): *Plotkin's Vaccines*, 7th ed. Philadelphia, Elsevier, 2017.

Schuchat A, Bell BP: Monitoring the impact of vaccines post-licensure: New challenges, new opportunities. Expert Rev Vaccines 7:437, 2008. [PubMed: 18444891]

Whitney CW et al: Benefits from immunization during the Vaccines for Children Program era—United States, 1994–2013. MMWR 63:352, 2014. [PubMed: 24759657]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 119: Health Recommendations for International Travel

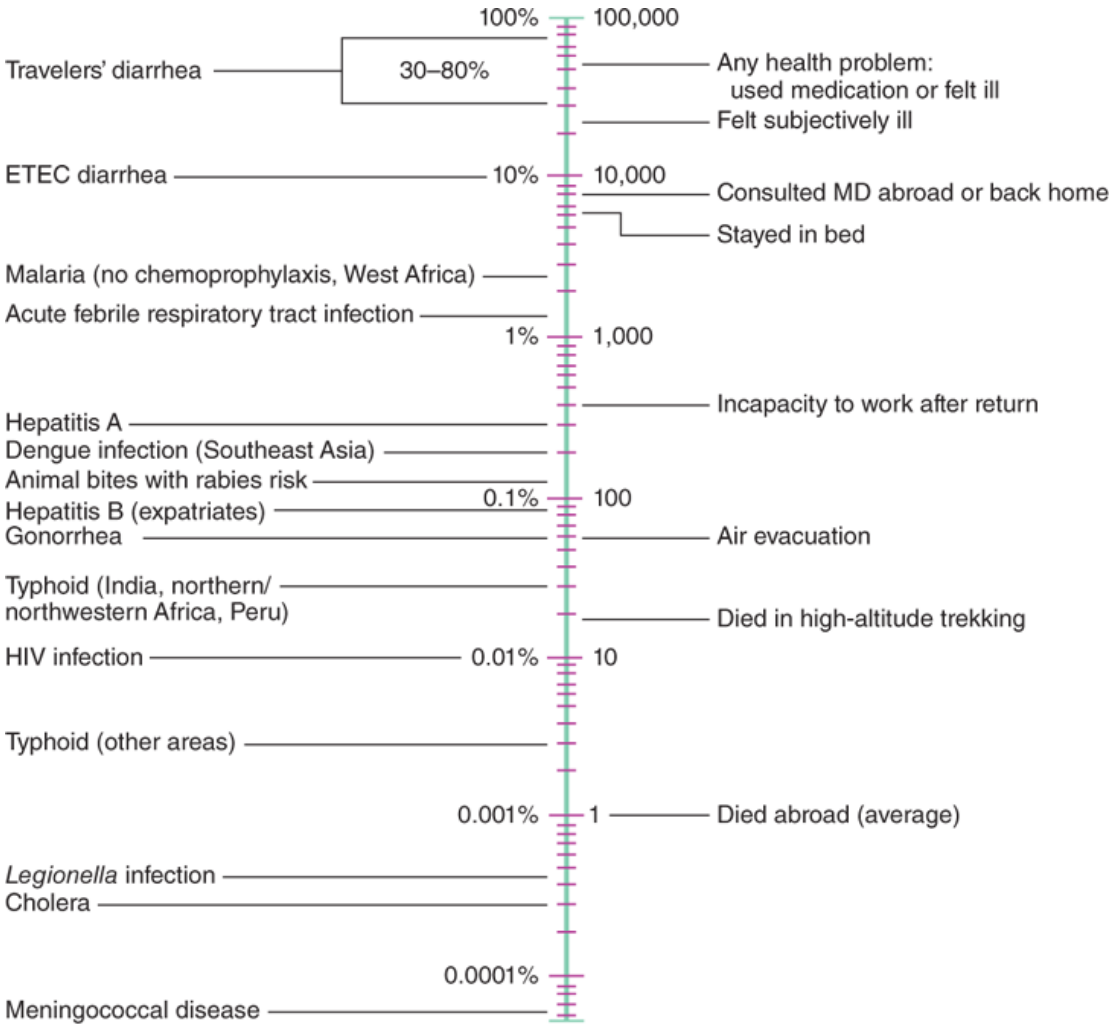
Jay S. Keystone; Phyllis E. Kozarsky

INTRODUCTION

According to the United Nations World Tourism Organization, international tourist arrivals grew dramatically from 25 million in 1950 to >1 billion in 2016; limited data suggest that the only areas without the growth of tourism were in Africa. Not only are more people traveling; travelers are seeking more exotic and remote destinations. In addition to tourism travel, travel across borders has increased in other sectors as well—e.g., for visits with friends and relatives (VFRs) in travelers' places of birth, for immigration, for business, and for missionary and volunteer work. Travel from industrialized to developing regions has been increasing, with Asia and the Pacific and the Middle East the emerging destinations. **Figure 119-1** summarizes the monthly incidence of health problems during travel in developing countries. Studies continue to show that 50–75% of short-term travelers to the tropics or subtropics report some health impairment. Most of these health problems are minor: only 5% require medical attention, and <1% require hospitalization. Although infectious agents contribute substantially to morbidity among travelers, these pathogens account for only ~1% of deaths in this population. Cardiovascular disease and injuries are the most frequent causes of death among travelers from the United States, accounting for 49 and 22% of deaths, respectively. Age-specific rates of death due to cardiovascular disease are similar among travelers and non-travelers. In contrast, rates of death due to injury (the majority from motor vehicle, drowning, or aircraft accidents) are several times higher among travelers. Motor vehicle accidents account for >40% of travelers' deaths that are not due to cardiovascular disease or preexisting illness.

FIGURE 119-1

Monthly incidence rates of health problems during stays in developing countries. ETEC, enterotoxigenic *Escherichia coli. (From R Steffen et al: Int J Antimicrob Agents 21:89, 2003. Reprinted with permission from Elsevier Science B.V. and International Society of Chemotherapy. https://www-sciencedirect-com.ezproxy1.library.usyd.edu.au/journal/international-journal-of-antimicrobial-agents.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GENERAL ADVICE

Health maintenance recommendations are based not only on the traveler's destination but also on assessment of risk, which is determined by such variables as health status, specific itinerary, purpose of travel, season, and lifestyle during travel. Detailed information regarding country-specific risks and recommendations may be obtained from the Centers for Disease Control and Prevention (CDC) publication *Health Information for International Travel* (available at *www.cdc.gov/travel*).

Fitness for travel is an issue of growing concern in view of the increased numbers of elderly and chronically ill individuals journeying to exotic destinations (see "Travel and Special Hosts," below). Since most commercial aircraft are pressurized to 2500 m (8000 ft) above sea level (corresponding to a Pao_2 of ~55 mmHg), individuals with serious cardiopulmonary problems or anemia should be evaluated before travel. In addition, those who have recently had surgery, a myocardial infarction, a cerebrovascular accident, or a deep-vein thrombosis may be at high risk for adverse events during flight. A summary of current recommendations regarding fitness to fly has been published by the Aerospace Medical Association Air Transport Medicine Committee (*www.asma.org/publications/medical-publications-for-airline-travel*). A pre-travel health assessment is advisable for individuals considering particularly adventurous recreational activities, such as mountain climbing and scuba diving.

IMMUNIZATIONS FOR TRAVEL

Immunizations for travel fall into three broad categories: *routine* (childhood/adult immunizations and boosters that are important regardless of travel), *required* (immunizations that are mandated by international regulations for entry into certain areas or for border crossings), and *recommended* (immunizations that are desirable because of travel-related risks). Required and recommended vaccines commonly given to travelers are listed in **Table 119-1**.

TABLE 119-1

Vaccines Commonly Used for Travel in Adults

Vaccine	Primary Series	Booster Interval
Cholera: Dukoral[a] (inactivated whole-cell recombinant subunit; available in Canada and Europe), Vaxchora (live attenuated; available in U.S.)	1 dose	2 years for Dukoral; unknown for Vaxchora
Hepatitis A (Havrix), 1440 enzyme immunoassay U/mL	2 doses, 6–12 months apart, IM	None required
Hepatitis A (VAQTA, AVAXIM, EPAXAL)	2 doses, 6–18 months apart, IM	None required
Hepatitis A/B combined (Twinrix)	3 doses at 0, 1, and 6 months or 0, 7, and 21–30 days plus booster at 1 year, IM	None required *except* 12 months (once only, for accelerated schedule)
Hepatitis B (Engerix B): accelerated schedule	3 doses at 0, 1, and 2 months or 0, 7, and 21 days plus booster at 1 year, IM	12 months, once only
Hepatitis B (Engerix B or Recombivax): standard schedule	3 doses at 0, 1, and 6 months, IM	None required
Japanese encephalitis (Ixiaro)	2 doses at 0 and 28 days, IM	>1 year after primary series (optimal booster schedule not yet determined)
Meningococcus, quadrivalent (Menomune [polysaccharide], Menactra, Menveo [conjugate])	1 dose (Menactra/Menveo, IM; Menomune, SC)	>3 years (optimal booster schedule not yet determined)
Rabies human diploid cell vaccine (Imovax), rabies vaccine absorbed (RVA), or purified chick embryo cell vaccine (RabAvert)	3 doses at 0, 7, and 21 or 28 days, IM	None required except with exposure
Typhoid Ty21a, oral live attenuated (Vivotif)	1 capsule every other day × 4 doses	5 years
Typhoid Vi capsular polysaccharide, injectable (Typhim Vi)	1 dose IM	2 years
Yellow fever	1 dose SC	1 lifetime dose

[a]Cross-protects against enterotoxigenic *Escherichia coli* and provides 30–50% protection against travelers' diarrhea.

Routine Immunizations

DIPHTHERIA, TETANUS, AND POLIO

Diphtheria **(Chap. 145)** continues to be a problem worldwide. Large outbreaks have occurred in countries that do not have rigorous vaccination programs or that have reduced their public vaccination programs. Serologic surveys show that tetanus **(Chap. 147)** antibodies are lacking in many North Americans, especially in women aged >50. With the recent increase in pertussis among adults, the diphtheria–tetanus–acellular pertussis (Tdap) combination is now recommended for adults as a once-only replacement for the 10-year tetanus–diphtheria (Td) booster.

The risk of polio **(Chap. 199)** to the international traveler is extremely low despite challenges faced by eradication programs. Wild-type poliovirus has been eradicated from most areas of the world; Nigeria, Afghanistan, and Pakistan are the only countries where polio continues to be endemic. Some countries may actually require travelers who have been in country for >4 weeks to show proof on exiting that they have received polio vaccine within the previous year. (Because this list of countries changes, providers should check the CDC travelers' health website at *www.cdc.gov/travel*.) Studies in the United States suggest that 12% of adult travelers are unprotected against at least one poliovirus serogroup. Foreign travel offers an ideal opportunity to have polio immunization updated.

MEASLES

Measles (rubeola) continues to be a major cause of morbidity and death in the developing world (Chap. 200). Several outbreaks of measles in the United States and Canada have been linked to imported cases, especially from Europe, where large outbreaks have occurred. The group at highest risk consists of persons born after 1956 and vaccinated before 1980, in many of whom primary vaccination failed. The measles–mumps–rubella (MMR) vaccine is typically used; its coverage of rubella also addresses a growing concern in some areas of Eastern Europe and Asia.

INFLUENZA

Influenza (Chap. 195)—possibly the most common vaccine-preventable infection in travelers—occurs year-round in the tropics and during the summer months in the Southern Hemisphere (coinciding with the winter months in the Northern Hemisphere). One prospective study showed that influenza developed in 1% of travelers to Southeast Asia per month of stay. Annual vaccination should be considered for all travelers who do not have a contraindication. The speed of global spread of the pandemic H1N1 virus in 2009 illustrated why influenza immunization is so important for travelers.

PNEUMOCOCCAL INFECTION

Regardless of travel, pneumococcal vaccine (Chap. 141) should be administered routinely to all persons aged >65 and to persons between the ages of 2 and 64 who are at high risk of serious infection, including those with diabetes mellitus; those with chronic heart, lung, or kidney disease; those who have been splenectomized or are immunocompromised; and those who have sickle cell disease.

Required Immunizations

YELLOW FEVER

Documentation of vaccination against yellow fever (Chap. 204) may be required or recommended as a condition for entry into or passage through countries of sub-Saharan Africa and equatorial South America, where the disease is endemic or epidemic, or (according to the International Health Regulations [IHR]) for entry into countries at risk of having the infection introduced. In 2014, the World Health Organization (WHO) adopted a recommendation to remove the 10-year booster-dose requirement from the IHR as of June 2016. Thus one dose of yellow fever vaccine and a completed International Certificate of Vaccination or Prophylaxis should be valid for the lifetime of the vaccinee. Some countries have already adopted this change, as noted on the CDC's website under the yellow fever vaccine requirements on each country's destination page. However, it is uncertain when and whether all countries with yellow fever vaccination requirements will adopt this change. Some countries may still require a booster after 10 years, and a booster may be recommended for other travelers as well. This vaccine is given only by state-authorized yellow fever centers, and its administration must be documented on an official International Certificate of Vaccination or Prophylaxis. A registry of U.S. clinics that provide the vaccine is available from the CDC (*www.cdc.gov/travel*). Data suggest that fewer than 50% of travelers entering areas endemic for yellow fever are immunized, and lack of coverage is a serious problem. Severe adverse events associated with this vaccine have increased in incidence. First-time vaccine recipients may present with a syndrome characterized as either neurotropic (1 case per 125,000 doses) or viscerotropic (overall, 1 case per 250,000 doses; among persons 60–69 years of age, 1 case per 100,000 doses; and among persons ≥70 years of age, 1 case per 40,000 doses). Immunosuppression and thymic disease increase the risk of these adverse events (*https://www.cdc.gov/vaccines/hcp/vis/vis-statements/yf.pdf*).

MENINGOCOCCAL MENINGITIS

Protection against meningitis is required for entry into Saudi Arabia during the Hajj (Chap. 150). Hajj visas cannot be issued without proof of meningococcal vaccination. All adults and children >2 years of age must have received a single dose of quadrivalent A/C/Y/W-135 vaccine and must show proof of vaccination on a valid International Certificate of Vaccination or Prophylaxis.

Recommended Immunizations

HEPATITIS A AND B

Hepatitis A (Chap. 332) is one of the most common vaccine-preventable infections of travelers. Older data demonstrated a risk six times greater for travelers who stray from the usual tourist routes. The mortality rate for hepatitis A increases with age, reaching almost 2% among individuals aged >50. Of the four hepatitis A vaccines currently available in North America (two in the United States), all are interchangeable and have an efficacy of >95%. Hepatitis A vaccine is currently given to all children in the United States. Since the most frequently identified risk factor for hepatitis A in the United States is international travel, and since morbidity and mortality risk increase with age, it seems appropriate that all adults be immune prior to travel.

Long-stay overseas workers appear to be at considerable risk for hepatitis B infection (Chap. 332), although even short-term travelers can acquire this infection if they indulge in behaviors that place them at risk. The recommendation that all travelers be immunized against hepatitis B before departure is supported by two studies showing that 17% of the assessed travelers who received health care abroad had some type of injection; according to the WHO, nonsterile equipment is used for up to 75% of all injections given in parts of the developing world. A 3-week accelerated schedule of the combined hepatitis A and B vaccine has been approved in the United States. Although no data are available on the

specific risk of infection with hepatitis B virus among U.S. travelers, ~240 million people in the world have chronic infection. All children and adolescents in the United States are immunized against this illness. Hepatitis B vaccination should be considered for all travelers.

TYPHOID AND PARATYPHOID FEVER

Most cases of typhoid fever in North America are due to travel, with ~300 cases seen per year in the United States. The attack rate for typhoid fever (Chap. 160) is 1 case per 30,000 travelers per month of travel to the developing world. In the United States, >80% of reports of typhoid fever and >90% of reports of paratyphoid fever caused by *Salmonella* Paratyphi A are in travelers to southern Asia. One group at particular risk are immigrants and their families who have returned to their homelands for VFRs. Between 1999 and 2006 in the United States, 66% of imported cases of *S. typhi* infection involved the latter group. Unfortunately, data show that both *S. typhi* and *S. paratyphi* A have become increasingly resistant to fluoroquinolone antibiotics (especially strains acquired on the Indian subcontinent). Both of the available vaccines—one oral (live) and the other injectable (polysaccharide)—have efficacy rates of ~70% but are not protective against Paratyphi disease. In some countries, a combined hepatitis A/typhoid vaccine is available.

MENINGOCOCCAL MENINGITIS

Although the risk of meningococcal disease among travelers has not been quantified, it is likely to be higher among travelers who live with poor indigenous populations in overcrowded conditions (Chap. 150). Because of its enhanced ability to prevent nasal carriage (compared with the older polysaccharide vaccine), a quadrivalent conjugate vaccine is the product of choice (regardless of age) for immunization of persons traveling to sub-Saharan Africa during the dry season or to areas of the world where there are epidemics. The vaccine, which protects against serogroups A, C, Y, and W-135, has an efficacy rate of >90%. Except in rare outbreak situations, there is no role for meningococcal serogroup B immunization of travelers.

JAPANESE ENCEPHALITIS

The risk of Japanese encephalitis (Chap. 204), an infection transmitted by night-biting mosquitoes in rural Asia and Southeast Asia, can be as high as ~1 case per 5000 travelers per month of stay in an endemic area. Most infections are asymptomatic; however, among the very small proportion of infected persons who become ill, death and severe neurologic sequelae are common. Most symptomatic infections in recent years have occurred in tourists to Southeast Asia, of whom one-third had traveled for <1 month in an endemic area. The vaccine efficacy rate is >90%, and vaccination is recommended by the CDC for persons staying >1 month in rural endemic areas or for shorter periods if their activities in these areas (e.g., camping, bicycling, hiking) will increase exposure risk. Recent studies suggest that the vaccine schedule (off-label) may be accelerated to 2 doses within 1 week for last-minute travelers.

CHOLERA

The risk of cholera (Chap. 163) is extremely low, with ~1 case per 500,000 journeys to endemic areas. A live oral cholera vaccine was recently approved in the United States by the U.S. Food and Drug Administration (FDA). Its use should be considered for aid and health care workers in refugee camps or in disaster-stricken/war-torn areas.

RABIES

Domestic animals, primarily dogs, are the major transmitters of rabies in developing countries (Chap. 203). Several studies have shown that the risk of rabies posed by a dog bite in an endemic area translates into 1–3.6 cases per 1000 travelers per month of stay. Countries where canine rabies is highly endemic include Mexico, the Philippines, Sri Lanka, India, Thailand, China, and Vietnam. The two vaccines available in the United States provide >90% protection. Rabies vaccine is recommended for long-stay travelers, particularly children (who tend to play with animals and may not report bites), and for persons who may be occupationally exposed to rabies in endemic areas; however, in a large-scale study, almost 50% of potential exposures occurred within the first month of travel. Even after receipt of a preexposure rabies vaccine series, two postexposure doses are required; rabies immune globulin (which is often unavailable in developing countries) is not necessary.

PREVENTION OF MALARIA AND OTHER INSECT-BORNE DISEASES

It is estimated that >30,000 American and European travelers develop malaria each year (Chap. 219). The risk to travelers is highest in Oceania and sub-Saharan Africa (estimated at 1:5 and 1:50 per month of stay, respectively, among persons not using chemoprophylaxis); intermediate in malarious areas on the Indian subcontinent and in Southeast Asia (1:250–1:1000 per month); and low in South and Central America (1:2500–1:10,000 per month). Malaria surveillance in the United States from 2012 to 2013 showed a 2% increase in cases, and annual increases have been reported since the early 1970s. Of the more than 1700 cases reported in 2014, 66% were due to *Plasmodium falciparum*; of cases in which a purpose of travel was reported, 58% were associated with VFRs. Patients traveling for VFRs are at the highest risk of acquiring malaria and may die of the disease if their immunity has waned after living outside an endemic area. There were five deaths due to malaria in the United States in 2014. Only 8% of those infected had adhered to CDC guidelines for chemoprophylaxis. With the worldwide increase in chloroquine- and multidrug-resistant falciparum malaria, decisions about chemoprophylaxis have become more difficult. Table 119-2 lists the currently recommended drugs of choice for prophylaxis of malaria, by destination.

TABLE 119-2

Malaria Chemosuppressive Regimens, According to Geographic Area[a]

Geographic Area	Drug of Choice[b]	Alternatives
Central America (north of Panama), Iraq, Turkey, northern Argentina, and Paraguay	Chloroquine	Atovaquone-proguanil[c] Doxycycline Mefloquine Primaquine
South America (but not northern Argentina or Paraguay, where chloroquine may be used); Central America (only Panama east of the Canal); Asia (including Southeast Asia); Africa; and Oceania	Doxycycline Atovaquone-proguanil[c] Mefloquine	
Thai–Myanmar and Thai–Cambodian borders, central Vietnam	Atovaquone-proguanil[c] Doxycycline	

[a]See the CDC's *Health Information for International Travel 2018* (*www.cdc.gov/travel*). [b]In all areas where chloroquine can still be used, the other drugs listed may be used as alternatives. [c]Malarone.

Note: See also **Chap. 219**.

Several studies indicate that fewer than 50% of travelers adhere to basic recommendations for malaria prevention. Keys to the prevention of malaria include both personal protection measures against mosquito bites (especially between dusk and dawn) and malaria chemoprophylaxis. The former measures entail the use of DEET or picaridin-containing insect repellents, permethrin-impregnated bed nets and clothing, screened sleeping accommodations, and protective clothing. Thus, in regions where infections such as malaria are transmitted, protective products are recommended, even for children and infants. In general, higher concentrations of any active ingredient provide a longer duration of protection. However, studies suggest that concentrations of DEET above ~50% do not offer a marked increase in protection time against mosquitoes. The CDC also recommends oil of lemon eucalyptus (PMD, para-menthane-3,8-diol) and IR3535 (3-[*N*-butyl-*N*-acetyl]-aminopropionic acid, ethyl ester). Personal protection measures also help prevent other insect-transmitted illnesses, such as dengue, chikungunya, and Zika **(Chap. 204)**.

Over the past decade, the incidence of dengue has increased considerably, particularly in the Caribbean region, Latin America, Southeast Asia, and Africa. Chikungunya, another mosquito-borne infection that clinically resembles dengue but primarily causes symptoms and signs of arthralgia and arthritis (at times chronic and destructive) has particularly affected the Caribbean in the last few years. Zika virus has also emerged in the past 2 years. Although only 20% of those who are infected have symptoms, Zika virus has been associated with severe complications (microcephaly and other neurologic and organ-system problems) in newborns of women who become infected during pregnancy. In addition, Guillain-Barré syndrome has been associated with Zika virus. Many questions linger with respect to this illness, its complications, and its transmission, especially its sexual transmission.

The CDC travelers' health website must be checked prior to travel in order to assess the risk of all these mosquito-borne diseases at specific destinations. Pregnant women should not travel to Zika-affected areas. Dengue, chikungunya, and Zika viruses are transmitted by an urban-dwelling mosquito that may be found indoors and that bites during daylight, primarily at dawn and dusk. Mosquito avoidance measures are crucial for all travelers to regions where these vector-borne diseases are transmitted.

PREVENTION OF GASTROINTESTINAL ILLNESS

Diarrhea, the leading cause of illness in travelers **(Chap. 128)**, is usually a short-lived, self-limited condition. However, 40% of affected individuals need to alter their scheduled activities, and another 20% are confined to bed. The most important determinant of risk is the destination. Incidence rates per 2-week stay have been reported to be 10–40%, with the highest rates in parts of Africa and southern Asia. Infants and young

adults are at particularly high risk for gastrointestinal illness and for complications such as dehydration. Recent reviews suggest that there is little correlation between dietary indiscretions and the occurrence of travelers' diarrhea (TD). Earlier studies of U.S. students in Mexico showed that eating meals in restaurants and cafeterias or consuming food from street vendors was associated with increased risk. For further discussion, see "Precautions," below.

Etiology

The most frequently identified pathogens causing TD are enterotoxigenic *Escherichia coli* (ETEC) and enteroaggregative *E. coli* (EAEC) **(Chap. 156)**, although in some parts of the world (notably northern Africa and Southeast Asia) *Campylobacter* infections **(Chap. 162)** appear to predominate **(See also Table 128-3)**. Other common causative organisms include *Salmonella* **(Chap. 160)**, *Shigella* **(Chap. 161)**, rotavirus **(Chap. 198)**, and norovirus **(Chap. 198)**. The latter virus has caused numerous outbreaks on cruise ships and is an increasingly recognized cause of TD, causing up to 30% of such cases in some studies. Except for giardiasis **(Chap. 224)**, parasitic infections are uncommon causes of TD in short-term travelers. A growing problem for travelers is the development of antibiotic resistance among many bacterial pathogens and the movement of such pathogens worldwide. In centers that have molecular diagnostics capacity, other organisms are being identified in stools of patients with acute and chronic TD, although difficulties are encountered in their significance. The greater availability of these new modes for detection of pathogens in stool samples will reveal more about other and perhaps new pathogens responsible for TD.

Precautions

Some experts think that it is not only *what* travelers eat but also *where* they eat that puts them at risk of illness. Food sold by street vendors can carry a high risk, and restaurant hygiene can be a major problem over which the traveler has no control. In addition to discretion in choosing the source of food and water, general precautions include eating foods piping hot; avoiding foods that are raw or poorly cooked; and drinking only boiled or commercially bottled beverages, particularly those that are carbonated. Heating kills diarrhea-causing organisms, whereas freezing does not; therefore, ice cubes made from unpurified water should be avoided. In spite of these recommendations, the literature has repeatedly documented dietary indiscretions by 98% of travelers within the first 72 h after arrival at their destination. The maxim "Boil it, cook it, peel it, or forget it!" is easy to remember but apparently difficult to follow. Using hand sanitizer regularly has been shown to reduce TD.

Self-Treatment

As TD often occurs despite rigorous food and water precautions, travelers may want to carry medications for self-treatment **(See also Table 128-5)**. An antibiotic is useful in reducing the frequency of bowel movements and the duration of illness in moderate to severe diarrhea. The standard regimen is a 3-day course of a quinolone taken twice daily (or, in the case of some formulations, once daily) or, alternatively, a short regimen of azithromycin. However, studies have shown that one double dose of a quinolone or one dose of azithromycin may be equally effective. For diarrhea acquired in areas such as southern and Southeast Asia, where *Campylobacter* and other infections may be quinolone-resistant, azithromycin is the antibiotic of choice. Rifaximin, a poorly absorbed rifampin derivative, is highly effective against noninvasive bacterial pathogens such as ETEC and EAEC. The current approach to self-treatment of moderate to severe TD for the typical short-term traveler is to carry three once-daily doses of an antibiotic and to use as many doses as necessary to resolve the illness. If neither high fever nor blood in the stool accompanies diarrhea, loperamide may be taken alone or in combination with an antibiotic; studies have shown that the combination is better than the antibiotic alone and does not prolong illness.

Because of the growing problem of antimicrobial resistance worldwide, there is an effort to remind travelers that mild to moderate diarrhea may be managed with loperamide alone. For diarrhea that interferes with activity, adding an antibiotic is reasonable. However, the downside of antibiotic use is the change in the gut microbiota leading to carriage with extended-spectrum β-lactamase (ESBL)–producing Enterobacteriaceae that can persist for many months after return (50% for 1 month and 10% for 1 year). In one study, 28–80% of returned travelers carried these organisms after antibiotic self-treatment.

Prophylaxis

Bismuth subsalicylate remains a good option for the prophylaxis of TD but is only ~60% effective. For certain high-risk individuals (e.g., athletes, persons with a repeated history of TD, and persons with chronic diseases), a daily dose of a quinolone, azithromycin, or rifaximin during travel of <1 month's duration is 75–90% efficacious in preventing TD. A recommendation for the use of probiotics or prebiotics is premature; not enough information is available about the efficacy of using agents containing different organisms, the ideal number of organisms per dose, and the lack of product standardization. Further research on the gut microbiome will further elucidate this area. In Europe and Canada, an oral subunit cholera vaccine (Dukoral) that cross-protects against ETEC has been shown to provide 30–50% protection against TD. However, given the epidemiology of ETEC-induced TD, it is expected that only ~10% of travelers will benefit from this vaccine.

Illness after Return

Although extremely common, acute TD is usually self-limited or amenable to antibiotic therapy. Persistent bowel problems after the traveler returns home have a less well-defined etiology and usually require medical attention from a specialist. Infectious agents (e.g., *Giardia lamblia, Cyclospora cayetanensis, Entamoeba histolytica*) appear to be responsible for only a small proportion of cases with persistent bowel symptoms. One of the most common diagnoses in persistent diarrhea after travel is postinfectious irritable bowel syndrome. In as many as 4–13% of cases, symptoms may last months or years. When no infectious etiology can be identified, a trial of metronidazole therapy for presumed giardiasis (or small-intestinal bowel overgrowth), a strict lactose-free diet for a short period, or a trial of high-dose hydrophilic mucilloid may relieve symptoms. Because management of postinfectious irritable bowel syndrome can be quite complicated, these patients should be referred to a specialist.

PREVENTION OF OTHER TRAVEL-RELATED PROBLEMS

Travelers are at high risk for sexually transmitted diseases **(Chap. 131)**. Surveys have shown that large numbers of travelers engage in casual sex, and condoms are not used consistently. Increasing numbers of travelers are being diagnosed with illnesses such as schistosomiasis **(Chap. 229)**, Zika **(Chap. 204)**, dengue **(Chap. 204)**, chikungunya **(Chap. 204)**, and tick-borne rickettsial disease **(Chap. 182)**. Travelers are cautioned to avoid bathing, swimming, or wading in freshwater lakes, streams, or rivers in parts of northeastern South America, parts of the Caribbean, Africa, and Southeast Asia. Travelers should avoid walking barefoot because of the risk of hookworm and *Strongyloides* infections **(Chap. 227)** and snakebites **(Chap. 451)**. Insect repellents are important for prevention not only of malaria but also of other vector-borne diseases.

Prevention of travel-associated injury depends mostly on common-sense precautions. Riding on motorcycles (especially without helmets) and in overcrowded public vehicles is not recommended; in developing countries, individuals should *never* travel by road in rural areas after dark. Of persons who die during travel, fewer than 1% die of infection, whereas 40% die in motor vehicle accidents. Excessive alcohol use has been a significant factor in motor vehicle accidents, drownings, assaults, and injuries. During travel, situational awareness is important; wearing culturally appropriate clothing, avoiding flashy jewelry, and protecting one's money and passport are safety essentials.

THE TRAVELER'S MEDICAL KIT

A traveler's medical kit is strongly advisable. The contents may vary widely, depending on the itinerary, duration of stay, style of travel, and local medical facilities. While many medications are available abroad (often over the counter), directions for their use may be nonexistent or in a foreign language, or a product may be outdated or counterfeit. Recent studies of antimalarial products in sub-Saharan Africa and Southeast Asia showed that 30–50% were counterfeit or contained inadequate amounts of active drug. The sale and marketing of such medications are a growing industry that is expected to expand. In the medical kit, the short-term traveler should consider carrying an analgesic; an antidiarrheal agent and an antibiotic for self-treatment of TD; antihistamines; a laxative; oral rehydration salts; a sunscreen with broad-spectrum protection (UVA and UVB, with the latter at a level of at least 30 SPF); a DEET- or picaridin-containing insect repellent; an insecticide for clothing (permethrin); and, if necessary, an antimalarial drug. To these medications, the long-stay traveler might add a broad-spectrum general-purpose antibiotic (levofloxacin or azithromycin), an antibacterial eye and skin ointment, and a topical antifungal cream. The appropriate use of all antimicrobial agents should be reviewed prior to travel. Regardless of the duration of travel, a first-aid kit containing such items as scissors, tweezers, and bandages should be considered.

TRAVEL AND SPECIAL HOSTS

PREGNANCY AND TRAVEL

A woman's medical history and itinerary, the quality of medical care at her destinations, and her degree of flexibility determine whether travel is wise during pregnancy **(See also Chap. 466)**. According to the American College of Obstetrics and Gynecology, the safest part of pregnancy in which to travel is between 18 and 24 weeks, when there is the least danger of spontaneous abortion or premature labor. Some obstetricians prefer that women stay within a few hundred miles of home after the 28th week of pregnancy in case problems arise. In general, however, healthy women may be advised that it is acceptable to travel.

Relative contraindications to international travel during pregnancy include a history of miscarriage, premature labor, incompetent cervix, or toxemia. General medical problems such as diabetes, heart failure, severe anemia, or a history of thromboembolic disease also should prompt the pregnant woman to postpone her travels. Finally, destinations in which the pregnant woman and her fetus may be at excessive risk (e.g., those in Zika-affected areas, those at high altitudes, those where live-virus vaccines are required, and those where multidrug-resistant malaria is endemic) should be avoided.

Malaria

Malaria during pregnancy carries a significant risk of morbidity and mortality. Levels of parasitemia are highest and failure to clear the parasites after treatment is most frequent among primigravidae. Severe disease, with complications such as cerebral malaria, massive hemolysis, and renal failure, is especially likely in pregnancy. Fetal sequelae include spontaneous abortion, stillbirth, preterm delivery, and congenital infection. Chloroquine and mefloquine are considered to be safe in all trimesters.

Enteric Infections

Pregnant travelers must be extremely cautious regarding their food and beverage intake. Dehydration due to TD can lead to inadequate placental blood flow. Infections such as toxoplasmosis, hepatitis E, and listeriosis also can cause serious sequelae in pregnancy.

The mainstay of therapy for TD during pregnancy is rehydration. Loperamide may be used if necessary. For self-treatment, azithromycin may be the best option. Although quinolones are increasingly being used safely during pregnancy and rifaximin is poorly absorbed from the gastrointestinal tract, these drugs are not approved for this indication.

Because of the serious problems encountered when infants are given local foods and beverages, women are strongly encouraged to breast-feed when traveling with a neonate. A nursing mother with TD should not stop breast-feeding but should increase her fluid intake.

Air Travel and High-Altitude Destinations

Commercial air travel is not a risk to the healthy pregnant woman or to the fetus. The higher radiation levels reported at altitudes of >10,500 m (>35,000 ft) should pose no problem for the healthy pregnant traveler. Since each airline has a policy regarding pregnancy and flying, it is best to check with the specific carrier when booking reservations. Domestic air travel is usually permitted until the 36th week, whereas international air travel is generally curtailed after the 32nd week.

There are no known risks for pregnant women who travel to high-altitude destinations and stay for short periods. However, there are likewise no data on the safety of pregnant women at altitudes of >4500 m (15,000 ft). A prominent concern is that most of these destinations are remote.

THE HIV-INFECTED TRAVELER

The HIV-infected traveler is at special risk of serious infections due to a number of pathogens that may be more prevalent at travel destinations than at home **(See also Chap. 197)**. However, the degree of risk depends primarily on the state of the immune system at the time of travel. For persons whose CD4+ T cell counts are normal or >500/μL, data suggest no greater risk during travel than for persons without HIV infection. Individuals with AIDS (CD4+ T cell counts of <200/μL) and others who are symptomatic need special counseling and should visit a travel medicine practitioner before departure, especially when traveling to developing countries.

Several countries deny entry to HIV-positive individuals for prolonged stay, even though these restrictions do not appear to decrease rates of transmission of the virus. In general, HIV testing is required for individuals who wish to stay abroad for >3 months or who intend to work or study abroad. Some countries will accept an HIV serologic test done within 6 months of departure, whereas others will not accept a blood test done at any time in the traveler's home country. Border officials often have the authority to make inquiries of individuals entering a country and to check the medications they are carrying. If antiretroviral drugs are identified, the person may be barred from entering the country. Information on testing requirements for specific countries is available from consular offices but is subject to frequent change.

Immunizations

All of the HIV-infected traveler's routine immunizations should be up to date **(Chap. 118)**. The response to immunization may be impaired at CD4+ T cell counts of <200/μL and in some cases at even higher counts. Thus HIV-infected persons should be vaccinated as early as possible to ensure adequate immune responses. For patients receiving antiretroviral therapy, at least 3 months must elapse before regenerated CD4+ T cells can be considered fully functional; therefore, vaccination of these patients should be delayed. However, when the risk of illness is high or the sequelae of illness are serious, immunization is recommended. In certain circumstances, it may be prudent to check the adequacy of the serum antibody response before departure.

Because of the increased risk of infections due to *Streptococcus pneumoniae* and other bacterial pathogens that cause pneumonia after influenza, the conjugate pneumococcal vaccine (Prevnar 13) followed by the 23-valent polysaccharide vaccine (Pneumovax) as well as influenza vaccine should be administered. The estimated rates of response to influenza vaccine are >80% among persons with asymptomatic HIV infection and <50% among those with AIDS.

In general, live attenuated vaccines are contraindicated for persons with immune dysfunction. Because measles (rubeola) can be a severe or lethal infection in HIV-positive patients, these patients should receive the measles vaccine (or the combination MMR vaccine) unless the CD4+ T cell count is <200/μL. Between 18 and 58% of symptomatic HIV-infected vaccinees develop adequate measles antibody titers, and 50–100% of asymptomatic HIV-infected persons seroconvert.

In asymptomatic HIV-infected individuals, CD4+ T cell counts of 200–499/μL (moderate immune suppression) are considered a precaution for yellow fever vaccination; since there is no contraindication, yellow fever vaccine may be considered when these persons travel to endemic areas. Studies of these individuals, along with asymptomatic persons whose CD4+ T cell count is >500/μL, found no serious adverse events, although their immunologic response may be decreased. If the CD4+ T cell count is <200/μL, an alternative itinerary that poses no risk of exposure to yellow fever is recommended. If the traveler is passing through or traveling to an area where the vaccine is required but the disease risk is low, a physician's waiver should be issued. Although the WHO indicates that a single dose of yellow fever vaccine provides lifetime protection, the CDC recommends boosters every 10 years for HIV-infected individuals.

A transient increase in HIV viremia (lasting days to weeks) has been demonstrated in HIV-infected individuals after immunization against influenza, pneumococcal infection, and tetanus **(Chap. 197)**. At this point, however, no evidence indicates that this transient increase is detrimental.

Gastrointestinal Illness

Decreased levels of gastric acid, abnormal gastrointestinal mucosal immunity, other complications of HIV infection, and medications taken by HIV-infected patients make TD especially problematic in these individuals. TD is likely to occur more frequently, to be more severe, to be accompanied by bacteremia, and to be more difficult to treat. *Cryptosporidium*, *Isospora belli*, and *Microsporidium* infections, although uncommon, are associated with increased morbidity and mortality rates in patients with AIDS.

The HIV-infected traveler must be careful to consume only appropriately prepared foods and beverages, and some may benefit from antibiotic prophylaxis for TD. Sulfonamides (as used to prevent pneumocystosis) are ineffective because of widespread resistance.

Other Travel-Related Infections

Data are lacking on the severity of many vector-borne diseases in HIV-infected individuals. Malaria is especially severe in asplenic persons and in those with AIDS. The HIV load doubles during malaria, with subsidence in ~8–9 weeks; the significance of this increase in viral load is unknown.

Visceral leishmaniasis **(Chap. 221)** has been reported in numerous HIV-infected travelers. Diagnosis may be difficult, given that splenomegaly and hyperglobulinemia are often lacking and serologic results are frequently negative. Sandfly bites may be prevented by evening use of insect repellents.

Certain respiratory illnesses, such as histoplasmosis and coccidioidomycosis, cause greater morbidity and mortality among patients with AIDS. Although tuberculosis is common among HIV-infected persons (especially in developing countries), its acquisition by the short-term HIV-infected traveler has not been reported as a major problem. From a prospective study, it is estimated that, for travelers not engaged in health care, the risk of tuberculosis infection is ~3% per year of travel.

Medications

Adverse events due to medications and drug interactions are common and raise complex issues for HIV-infected persons. Rates of cutaneous reaction (e.g., increased cutaneous sensitivity to sulfonamides) are unusually high among patients with AIDS. Doxycycline appears to have no clinically significant interactions with either the protease inhibitors or the non-nucleoside reverse transcriptase inhibitors (NNRTIs). Zidovudine levels may be increased with atovaquone. The drug combination atovaquone/proguanil may interact with antiretroviral protease inhibitors such as ritonavir, darunavir, atazanavir, indinavir, and lopinavir as well as the NNRTIs nevirapine, etravirine, and efavirenz. In spite of potential interactions, atovaquone/proguanil is well tolerated and remains the choice for most HIV-infected travelers.

Concomitant administration of the antimalarial drug mefloquine and antiretroviral protease inhibitors such as ritonavir, lopinavir, darunavir, and atazanavir may result in increased levels of mefloquine, with an increased risk of QT prolongation. Similarly, ritonavir may increase chloroquine levels. On the other hand, serum levels of mefloquine may be lowered with the use of efavirenz, nevirapine, or etravirine. Because of the increase in antiretroviral agents and the lack of accumulated data on their interactions with antimalarial agents, decisions about malaria chemoprophylaxis continue to be difficult; with a short duration of travel, an interaction may be inconsequential. With regard to malaria treatment, a hypothetical concern is that the antimalarial drugs lumefantrine (combined with artemether in Coartem) and halofantrine (no

longer recommended due to toxicity) may interact with HIV protease inhibitors and NNRTIs since drugs in the latter two categories are known to be potent inhibitors of cytochrome P450. In keeping current with antiretroviral drug interactions, a website from the University of Liverpool (*www.hiv-druginteractions.org*) is helpful.

CHRONIC ILLNESS, DISABILITY, AND TRAVEL

Chronic health problems need not prevent travel, but special measures can make the journey safer and more comfortable.

Heart Disease

Cardiovascular events are the main cause of deaths among travelers and of in-flight emergencies on commercial aircraft. Extra supplies of all medications should be kept in carry-on luggage, along with a copy of a recent electrocardiogram and the name and telephone number of the traveler's physician at home. Pacemakers are not affected by airport security devices, although electronic telephone checks of pacemaker function cannot be transmitted by international satellites. Travelers with electronic defibrillators should carry a note to that effect and ask for hand screening. A traveler may benefit from supplemental oxygen; since oxygen delivery systems are not standard, supplementary oxygen should be ordered by the traveler's physician well before flight time. Travelers may benefit from aisle seating and should walk, perform stretching and flexing exercises, consider wearing support hose, and remain hydrated during the flight to prevent venous thrombosis and pulmonary embolism.

Chronic Lung Disease

Chronic obstructive pulmonary disease is one of the most common diagnoses in patients who require emergency-department evaluation for symptoms occurring during airline flights. The best predictor of the development of in-flight problems is the sea-level Pa_{O_2}. A Pa_{O_2} of at least 72 mmHg corresponds to an in-flight arterial Pa_{O_2} of ~55 mmHg when the cabin is pressurized to 2500 m (8000 ft). If the traveler's baseline Pa_{O_2} is <72 mmHg, the provision of supplemental oxygen should be considered. Contraindications to flight include active bronchospasm, lower respiratory infection, lower-limb deep-vein phlebitis, pulmonary hypertension, and recent thoracic surgery (within the preceding 3 weeks) or pneumothorax. Decreased outdoor activity at the destination should be considered if air pollution is excessive.

Diabetes Mellitus

Alterations in glucose control and changes in insulin requirements are common problems among patients with diabetes who travel. Changes in time zones, in the amount and timing of food intake, and in physical activity demand vigilant assessment of metabolic control. Because of the risk of foot ulcers, travelers should wear closed footwear that has been proven to be comfortable. The traveler with diabetes should pack medication (including a bottle of regular insulin for emergencies), insulin syringes and needles, equipment and supplies for glucose monitoring, and snacks in carry-on luggage. Insulin is stable for ~3 months at room temperature but should be kept as cool as possible. The name and telephone number of the home physician and a card and bracelet listing the patient's medical problems and the type and dose of insulin used should accompany the traveler. In order to facilitate international border crossings, travelers should carry a physician's letter authorizing the carriage of needles and syringes. In traveling eastward (e.g., from the United States to Europe), the patient may need to decrease the morning insulin dose on arrival. The blood glucose can then be checked during the day to determine whether additional insulin is required. For flights westward, with lengthening of the day, an additional dose of regular insulin may be required.

Other Special Groups

Other groups for whom special travel measures are encouraged include patients undergoing dialysis, those with transplants, and those with other disabilities. Up to 13% of travelers have some disability, but few advocacy groups and tour companies dedicate themselves to this growing population. Medication interactions are a source of serious concern for these travelers, and appropriate medical information should be carried, along with the home physician's name and telephone number. Some travelers taking glucocorticoids carry stress doses in case they become ill. Immunization of these immunocompromised travelers may result in less than adequate protection. Thus the traveler and the physician must carefully consider which destinations are appropriate.

TRAVEL HEALTH INSURANCE

Today, more elderly or chronically ill individuals travel, and more of these individuals journey to remote locations and enjoy adventurous activities. Illness or injury abroad is not uncommon and is best considered before the journey. Persons who develop health problems abroad may incur enormous out-of-pocket expenses. Thus prospective travelers should consider purchasing supplemental travel health insurance and

should check with their health insurance company regarding whether they have coverage for illness or injury overseas. Unfortunately, many insurance companies will not cover preexisting illness if it is the reason for trip cancellation or illness abroad. Most countries do not accept routine health insurance from other countries unless there is a special traveler supplement. In most circumstances, travelers are asked to pay in cash for services rendered on an emergency basis, whether in a physician's office, in an emergency or urgent care center, or even in a hospital. There are several types of travel insurance. It is wise to purchase *trip cancellation insurance*, especially, for example, if the traveler has an underlying chronic illness and may need to cancel a trip due to an exacerbation of disease. *Travel health insurance* will cover expenses in the event that medical care abroad is needed. *Evacuation insurance* will cover medical evacuation, usually to a medical center in another location where it is deemed that the care is similar to that available in the traveler's home country. The cost of medical evacuation can easily exceed $100,000 U.S. There are a number of travel insurance providers, and it is very important to read the fine print carefully and to determine exactly what each company provides, thereby ensuring an appropriate fit for the individual's particular circumstances. The U.S. Department of State website lists travel health insurance companies and provides information about whom to contact in an emergency through STEP (the Smart Traveler Enrollment Program: *https://travel.state.gov/content/passports/en/go/health/doctors.html*).

MEDICAL TOURISM

Travel for the purpose of obtaining health care abroad has received a great deal of attention in the medical literature and the media. Although the data are difficult to confirm, it has been estimated that at least 750,000 Americans travel each year for medical purposes. According to the Department of Commerce website, the numbers of such travelers have doubled recently, and some experts predict massive increases. Lower cost is usually cited as the motivation for this type of tourism, and an entire industry has flourished as a result of this phenomenon. However, the quality of facilities, assistance services, and care is neither uniform nor regulated; thus, in most instances, responsibility for assessing the suitability of an individual program or facility lies solely with the traveler. Persons considering this option must recognize that they are almost always at a disadvantage when being treated in a foreign country, particularly if there are complications. Concerns to be addressed include the quality of the health care facility and its staff; language and cultural differences that may impede accurate interpretation of both verbal and nonverbal communication; religious and ethical differences that may be encountered over issues such as efforts to preserve life and limb or the provision of care for the terminally ill; lack of familiarity with the local medical system; limited access of the care provider to the patient's medical history; the use of unfamiliar drugs and medicines; the relative difficulty of arranging follow-up care back in the United States; and the possibility that such follow-up care may be fraught with problems should there be complications. If serious issues arise, legal recourse may be difficult or impossible. Patients planning to travel abroad to obtain health care, particularly when surgery is involved, should be immunized for hepatitis B and should consider having baseline hepatitis C and HIV tests preoperatively. Prevalence rates of hepatitis B and C and HIV infection vary considerably around the world and are generally higher in developing regions than in the United States and Western Europe. The latest information available on the safety of the blood supply outside the United States is the WHO's Global Database on Blood Safety based on data from 2011 (*www.who.int/topics/blood_safety/en*). Persons researching the accreditation status of overseas facilities should note that, although these facilities may be part of a chain, they are surveyed and accredited individually. Accreditation resources include (1) the Joint Commission International (*www.jointcommissioninternational.org*), (2) the Australian Council for Healthcare Standards International (*www.achs.org.au/achs-international/*), and (3) Accreditation Canada International (*www.internationalaccreditation.ca*). The American Medical Association also offers guidelines for medical tourism (*www.medretreat.com/templates/UserFiles/Documents/Whitepapers/AMAGuidelines.pdf*).

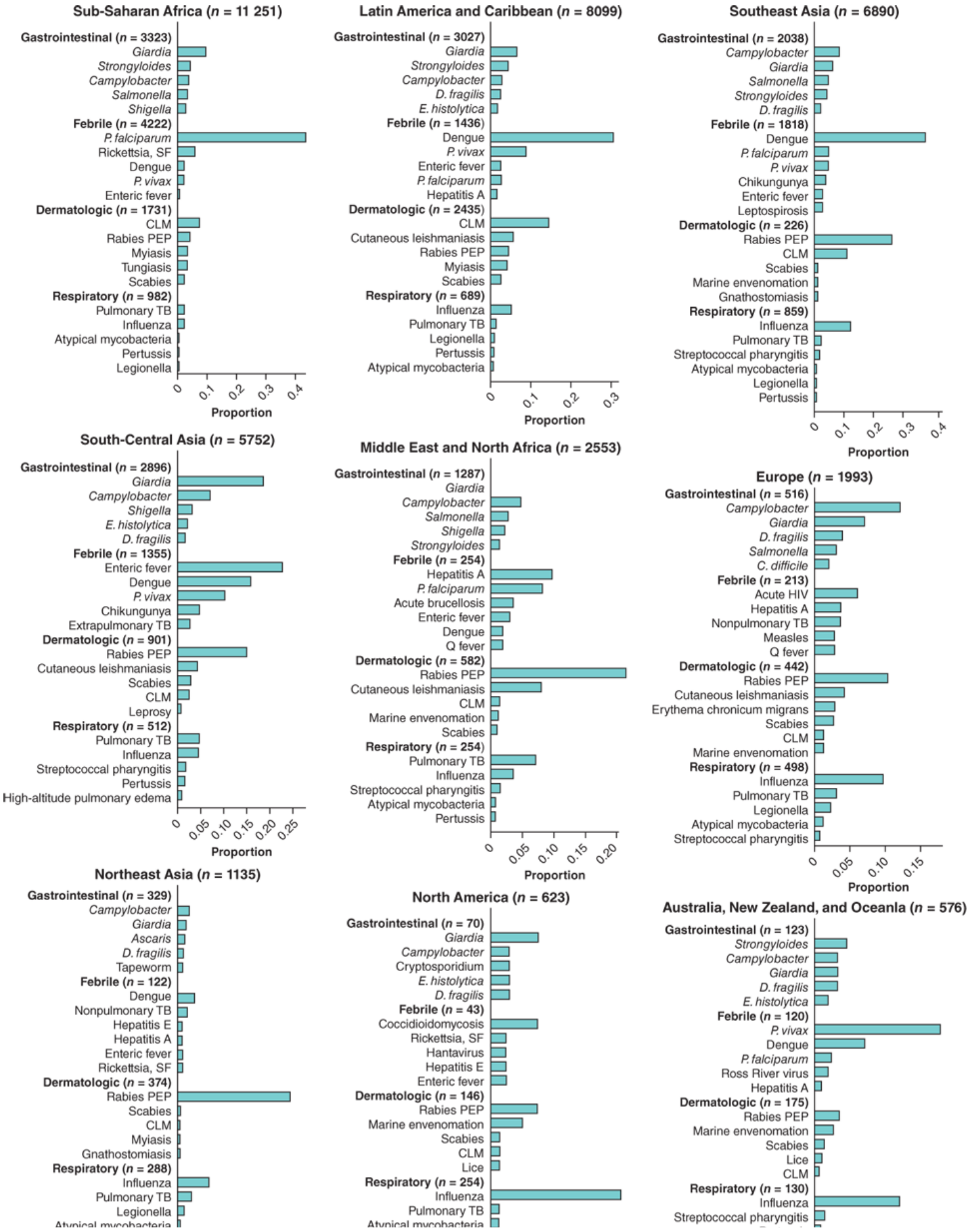
PROBLEMS AFTER RETURN

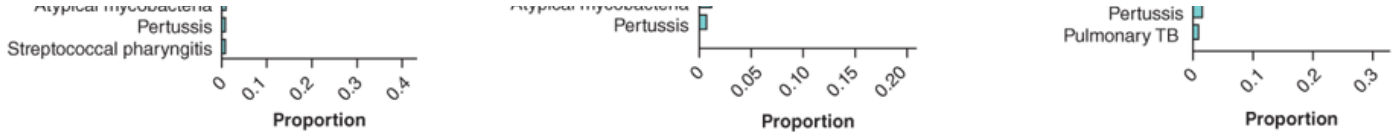
The most common medical problems encountered by travelers after their return home are diarrhea, fever, respiratory illnesses, and skin diseases **(Fig. 119-2)**. Frequently ignored problems are fatigue and emotional stress, especially in long-stay travelers. The approach to diagnosis requires some knowledge of geographic medicine, in particular the epidemiology and clinical presentation of infectious disorders. A geographic history should focus on the traveler's exact itinerary, including dates of arrival and departure; exposure history (food indiscretions, drinking-water sources, freshwater contact, sexual activity, animal contact, insect bites); location and style of travel (urban vs rural, first-class hotel accommodation vs. camping); immunization history; and use of antimalarial chemosuppression. Recently, some travelers who have been hospitalized abroad have been shown on return to be colonized with multidrug-resistant bacteria such as Enterobacteriaceae producing ESBLs and bacteria producing NDM-1 (New Delhi metallo-β-lactamase 1); these bacteria may be transmitted to family members and other contacts, especially within the health care system.

FIGURE 119-2

Top identified causes of gastrointestinal, febrile, dermatologic, and respiratory illnesses, by region, among ill returned travelers. More than five diagnoses are shown if more than one cause had equal numbers of cases. These graphs represent proportions, and there is variability in the number of ill travelers represented from panel to panel (shown from largest to smallest traveler numbers). CLM, cutaneous larva migrans; *D.*

fragilis, *Dientamoeba fragilis*; *E. histolytica*, *Entamoeba histolytica*; *P. falciparum*, *Plasmodium falciparum*; *P. vivax*, *Plasmodium vivax*; PEP, postexposure prophylaxis; SF, spotted fever; TB, tuberculosis. *(Reprinted with permission from K Leder et al: Ann Intern Med 158:456, 2013.)*





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIARRHEA

See "Prevention of Gastrointestinal Illness," above.

FEVER

Fever in a traveler who has returned from a malarious area should be considered a medical emergency because death from *Plasmodium falciparum* malaria can follow an illness of only several days' duration. Although "fever from the tropics" does not always have a tropical cause, malaria should be the first diagnosis considered. The risk of *P. falciparum* malaria is highest among travelers returning from Africa or Oceania and among those who become symptomatic within the first 2 months after return. Other important causes of fever after travel include dengue, chikungunya, Zika, viral hepatitis (A and E), typhoid and paratyphoid fever, bacterial enteritis, rickettsial infections (including tick typhus, scrub typhus, and Q fever), and—in rare instances—leptospirosis, acute HIV infection, and amebic liver abscess. Studies published by GeoSentinel (an emerging infectious disease surveillance group established by the CDC and the International Society of Travel Medicine) have informed providers about specific illness and syndrome risks in ill returned travelers (Fig. 119-2). Outbreaks of dengue, previously considered to be very rare in Africa, have been documented recently across central Africa. However, in at least 25% of cases, no etiology of fever in a returning traveler can be found, and the fever resolves spontaneously. Clinicians should keep in mind that no present-day antimalarial agent guarantees protection from malaria and that some immunizations (notably, that against typhoid fever) are only partially protective.

When no specific diagnosis is forthcoming, the following investigations, where applicable, are suggested: complete blood count, liver function tests, thick/thin blood films or rapid diagnostic testing for malaria (repeated several times if necessary), urinalysis, urine and blood cultures (repeated once), rapid influenza diagnostic testing, chest x-ray, and collection of an acute-phase serum sample to be held for subsequent examination along with a paired convalescent-phase serum sample.

SKIN DISEASES

Pyodermas, sunburn, insect bites, skin ulcers, and cutaneous larva migrans are the most common skin conditions affecting travelers after their return home. In individuals with persistent skin ulcers, a diagnosis of cutaneous leishmaniasis, mycobacterial infection, or fungal infection should be considered. Careful, complete inspection of the skin is important in detecting the rickettsial eschar in a febrile patient or the central breathing hole in a "boil" due to myiasis.

EMERGING INFECTIOUS DISEASES

In recent years, travel and commerce have fostered the worldwide spread of chikungunya, Zika, influenza caused by novel serotypes, severe acute respiratory syndrome (SARS), and (earlier) HIV infection and the reemergence of cholera as a global health threat. For travelers, some of these issues, along with others, remain a concern. Dengue continues to pose serious health threats in Latin America and Southeast Asia; schistosomiasis is being described in previously unaffected lakes in Africa; and antibiotic-resistant strains of sexually transmitted and enteric pathogens are emerging at an alarming rate.

CONCLUSIONS

The growth of global travel and migration now demand that the clinician become as familiar as possible with travel medicine. Practitioners may choose either to refer their patients to a travel clinic before departure or to acquire knowledge that enables them to provide pre-travel counseling and to prescribe appropriate vaccinations and chemoprophylaxis. It is equally important for physicians seeing ill returned travelers to be familiar with common post-travel syndromes and diseases, particularly those that may have been acquired in the developing world, and to identify other physicians who can assist with complex post-travel illnesses. The CDC publishes a biennial text, *Health Information for International Travel* (accessed through their website at *www.cdc.gov/travel*), that provides pre-travel health recommendations. The International Society of Travel Medicine (*www.istm.org*) publishes a list of travel clinics, and the American Society of Tropical Medicine and Hygiene (*www.astmh.org*) publishes a list of clinical tropical medicine specialists.

As Nobel Laureate Dr. Joshua Lederberg pointed out, “The microbe that felled one child in a distant continent yesterday can reach yours today and seed a global pandemic tomorrow.” The vigilant clinician understands that the importance of a thorough travel history cannot be overemphasized.

FURTHER READING

Aung AK et al: Travel risk assessment, advice and vaccinations in immune compromised travelers. Travel Med Infect Dis 13:31, 2015.

Centers for Disease Control and Prevention: CDC health information for international travel 2018. New York, Oxford University Press, 2017.

Centers for Disease Control and Prevention: Malaria surveillance—United States, 2014. Morb Mortal Wkly Rep 66(12):1, 2017.

Connor BA, Riddle MS: Post-infectious sequelae of travelers’ diarrhea. J Travel Med 20:303, 2013.

Freedman DO et al: Medical considerations before international travel. N Engl J Med 375:247, 2016.

Keystone JS et al (eds): *Travel Medicine*, 4th ed. London, Elsevier, 2018 (in press).

Kogelman L et al: Knowledge, attitudes, and practices of US practitioners who provide pre-travel advice. J Travel Med 21:104, 2014.

Leder K et al: GeoSentinel surveillance of illness in returned travelers, 2007–2011. Ann Intern Med 158:456, 2013.

Plotkin SA et al (eds): *Vaccines*, 7th ed. London, Elsevier, 2017.

Riddle MS et al: Guidelines for the prevention and treatment of travelers’ diarrhea: A graded expert panel report. J Travel Med 24(Suppl 1):S2, 2017. [PubMed: 28520996]

Sanford CA, Jong EC: Immunizations. Med Clin North Am 100:247, 2016.

Steffen R et al: Traveler’s diarrhea: A clinical review. JAMA 313:71, 2015.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 120: Climate Change and Infectious Disease

Aaron S. Bernstein

INTRODUCTION

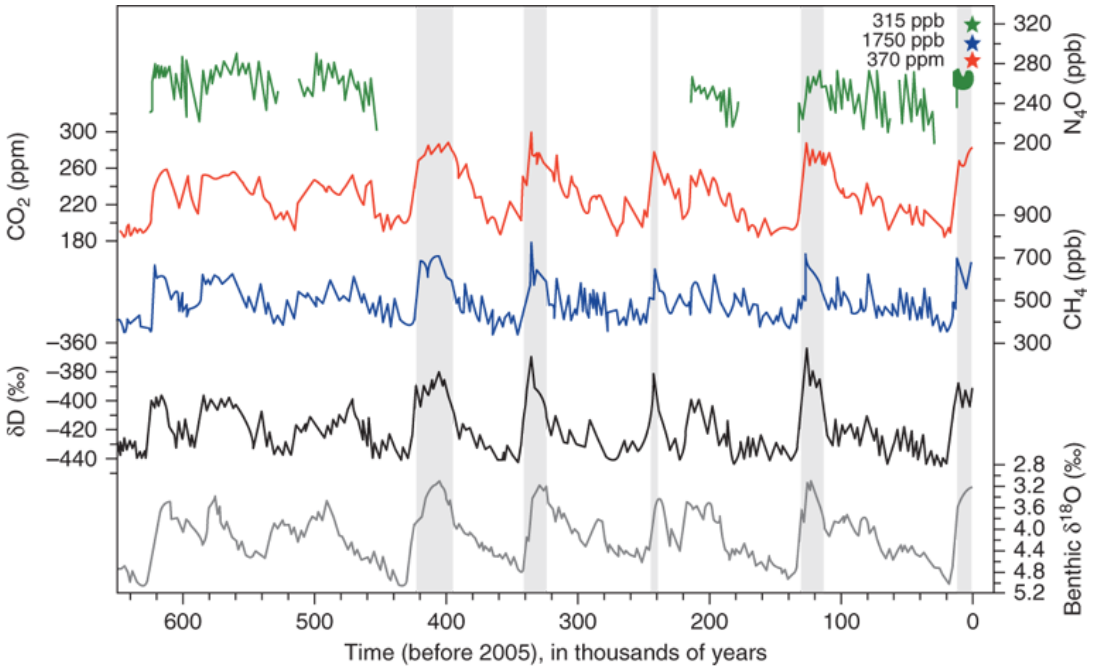
The release of greenhouse gases—principally carbon dioxide—into Earth's atmosphere since the late nineteenth century has contributed to a climate unfamiliar to our species, *Homo sapiens*. This new climate has already altered the epidemiology of some infectious diseases. Continued accumulation of greenhouse gases in the atmosphere will further alter the planet's climate. In some cases climate change may establish conditions favoring the emergence of infectious diseases, while in others it may render areas that are presently suitable for certain diseases unsuitable. This chapter presents the current state of knowledge regarding the known and prospective infectious-disease consequences of climate change.

OVERVIEW

The term *climate change* refers to long-term alterations in temperature, precipitation, wind, humidity, and other components of weather. Over the past 2.5 million years, the earth has warmed and cooled, cycling between glacial and interglacial periods during which average global temperatures moved up and down by 4–7°C. During the last glacial period, which ended roughly 12,000 years ago, global temperatures were, on average, 5°C cooler than in the mid-twentieth century **(Fig. 120-1)**.

FIGURE 120-1

Overview of the earth's temperature and primary greenhouse gases over the last 600,000 years. Variations of deuterium (δD; *black*) serve as a proxy for temperature. Atmospheric concentrations of greenhouse gases—CO_2 (*red*), CH_4 (*blue*), and nitrous oxide (N_2O; *green*)—were derived from air trapped within Antarctic ice cores and from recent atmospheric measurements. *Shaded areas* indicate interglacial periods. Benthic $\delta^{18}O$ marine records (*dark gray*) are a proxy for global ice-volume fluctuations and can be compared to the ice core data. Downward trends in the benthic $\delta^{18}O$ curve reflect increasing ice volumes on land. The *stars* and *labels* indicate atmospheric concentrations as of the year 2000. CO_2 levels surpassed 400 ppm as of 2013 and are rising at a rate of 2–2.5 ppm per year. *(From Intergovernmental Panel on Climate Change Fourth Assessment Report. Working Group I, Chapter 6, Figure 6.3. Cambridge University Press, 2007.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The present climate period, known as the Holocene, is remarkable for its stability: temperatures have largely remained within a range of 2–3°C. This stability has enabled the successful population and cultivation of much of the earth's landmass by humanity. Current climate change differs from that in the past not only because its primary cause is human activities but also because its pace is faster. The current rate of warming on Earth is unprecedented in the

last 50 million years. The 5°C of warming that occurred at the end of the last ice age about 12,000 years ago took roughly 5000 years, whereas such a temperature increment may occur within the next 150 years unless the release of greenhouse gases is substantially reduced in coming decades. Climate science, although still a relatively new discipline, has provided an ever-clearer picture of how the changing chemistry of the atmosphere has influenced, and will continue to influence, the global climate.

GREENHOUSE GASES

Greenhouse gases (**Table 120-1 and Fig. 120-2**) are a group of gases in Earth's atmosphere that absorb infrared radiation and thus retain heat inside the atmosphere. In the absence of these gases, the earth's average temperature would be about 33°C colder. Carbon dioxide, released into the atmosphere primarily from fossil fuel combustion and deforestation, has had the greatest effect on climate since the Industrial Revolution. Of note, the Swedish scientist Svante Arrhenius first suggested in the late nineteenth century that the addition of carbon dioxide to the Earth's atmosphere would increase the planet's surface temperature. Water vapor is the most abundant and a highly potent greenhouse gas but, given its short atmospheric life span and sensitivity to temperature, is not a major factor in recently observed climate change.

TABLE 120-1

Greenhouse Gases: Sources, Sinks, and Forcings

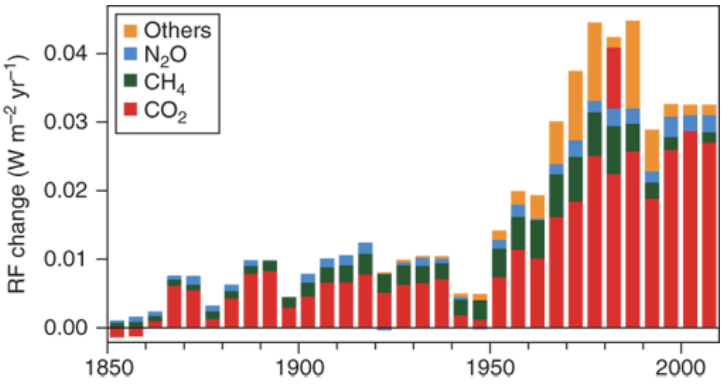
Gas	Human Sources	Sink[a]	Radiative Forcing[b] (95% Confidence Interval)
Carbon dioxide (CO_2)	Fossil fuel combustion, deforestation	Uptake by oceans (~30%), plants	1.68 (1.33–2.03)
Methane (CH_4)	Fossil fuel production, ruminant animals, decomposition in landfills	Hydroxyl radicals in the troposphere	0.97 (0.74–1.20)
Nitrous oxide (N_2O)	Fertilizer, fossil fuel combustion, biomass burning, livestock manure	Photolysis in the stratosphere	0.17 (0.14–0.23)
Halocarbons	Refrigerants, electrical insulation, aluminum production	Hydroxyl radicals in the troposphere, sunlight in the stratosphere	0.18 (0.01–0.35)

[a]In this table, a *sink* refers to the place where greenhouse gases are naturally stored or the mechanism through which they are destroyed. [b]*Radiative forcing*, measured in watts per meter squared, refers to how much an entity can alter the balance of incoming and outgoing radiation to and from Earth's atmosphere. It is measured relative to a preindustrial (i.e., 1750) baseline. Greenhouse gases have a positive "forcing"; that is, on balance, they increase the amount of radiation (and specifically infrared radiation) that is retained in Earth's atmosphere.

Sources: Intergovernmental Panel on Climate Change Fifth Assessment Report, Working Group 1, Chapter 8; American Chemical Society "Greenhouse gas sources and sinks," available at *www.acs.org/content/acs/en/climatescience/greenhousegases/sourcesandsinks.html*.

FIGURE 120-2

Acceleration of radiative forcing (RF) from release of major greenhouse gases, 1850–2011. For definition of *radiative forcing*, see footnote *b* to Table 120-1. *(From Intergovernmental Panel on Climate Change Fifth Assessment Report, Working Group 1, Chapter 8, Figure 8.6, p. 677.)*

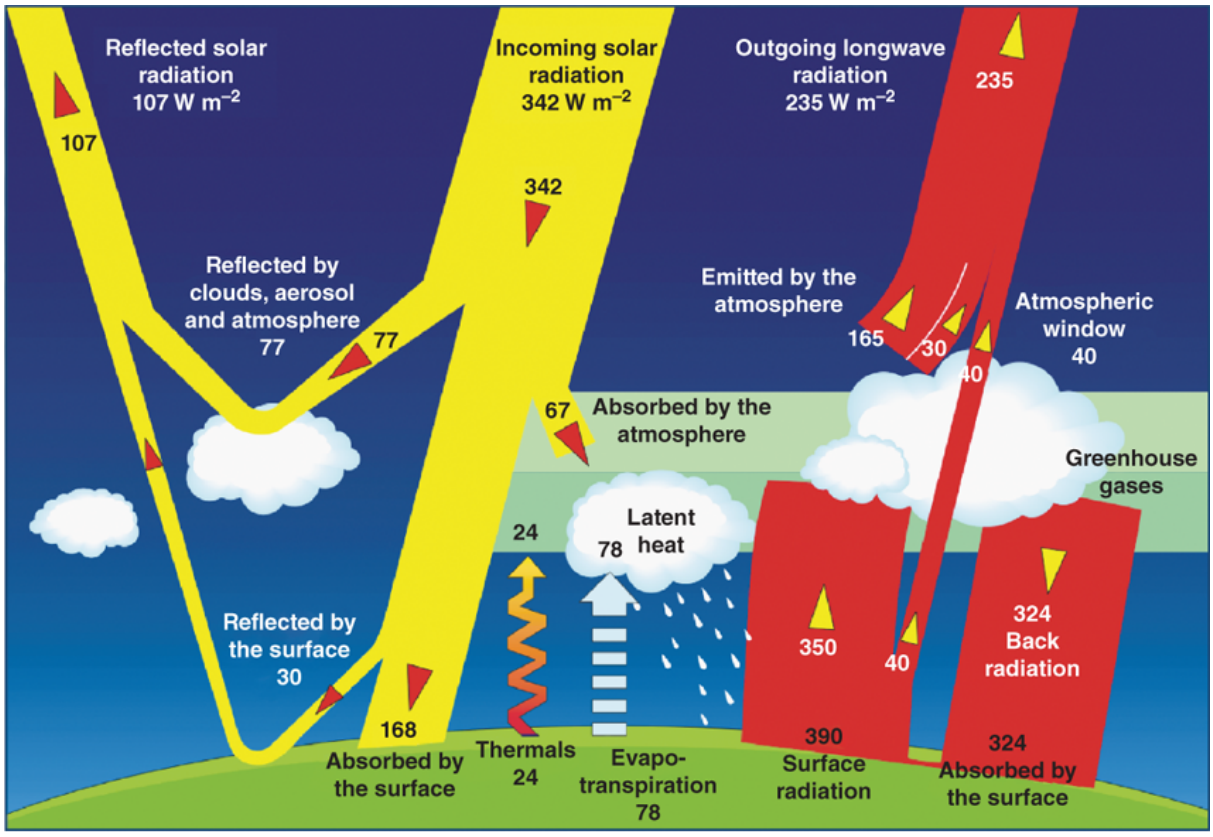


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The atmosphere, some of the aerosols suspended in it, and clouds reflect a portion of incoming solar radiation back toward space. The remainder reaches Earth's surface, where it is absorbed and some is then emitted back at the atmosphere. The earth emits energy absorbed from the sun at longer wavelengths, primarily infrared, that greenhouse gases are able to absorb. The change in wavelength that occurs as solar radiation is absorbed and re-emitted from the earth's surface is fundamental to the greenhouse effect **(Fig. 120-3)**.

FIGURE 120-3

Earth's energy balance. *(From JT Kiehl: Earth's annual global mean energy budget. Bull Am Meteor Soc 78: 197, 1997, Fig 7.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TEMPERATURE

Climate change has become nearly synonymous with global warming, as a clear signal from rising greenhouse gas concentrations has been an increase in the mean global surface temperature of ~0.85°C since 1880. However, this mean warming belies warming that is occurring much faster in certain regions. The Arctic has warmed twice as fast overall, and winters are warming faster than summers. Nighttime minimum temperatures are also rising faster than daytime high temperatures. Each of these nuances bears upon the incidence of infectious diseases in general and vector-borne disease specifically.

A moderate projection based on the best available scientific evidence suggests that average global temperatures will warm an additional 1.4–3.1°C by 2100 as compared to the period 1986–2005. Because of climate change, extreme heat waves have already become more common and are expected to be even more frequent later in this century. Besides contributing directly to morbidity and mortality in human populations, heat waves wilt crops and are expected to contribute substantially to predicted agricultural losses. For example, the 2010 heat wave in Russia, which was unprecedented in its severity, contributed to hundreds of forest fires that generated enough air pollution to kill an estimated 56,000 people and that burned 300,000 acres of crops, including roughly 25% of the nation's wheat fields. Nutritional deficiencies underlie a substantial portion of the global burden of many infectious diseases.

PRECIPITATION

In addition to changing temperature, the emission of greenhouse gases and the consequent increase in energy in Earth's atmosphere have influenced the planet's water cycle. Since 1950, substantial increases in the heaviest precipitation events (i.e., those above the 95th percentile) have been observed in Europe and North America. While trends over that same interval are less clear in other regions because of limited data, regions of Southeast Asia and southern South America have likely experienced increases in heavy precipitation as well. Other areas have seen greater drought, notably southern Australia and the southwestern United States.

A warmer atmosphere holds more water vapor; specifically, there is 6–7.5% more water vapor per degree (Celsius) of warming in the lower atmosphere. For areas that have traditionally had more precipitation on average, warming tends to promote heavier precipitation events. In contrast, in regions prone to

drought, warming tends to result in greater periods between rainfalls and in the risk of drought.

HURRICANES

The world's oceans have absorbed 90% of the excess heat that greenhouse gases have kept in Earth's atmosphere since the 1960s. Ocean heat provides energy for hurricanes, and warmer years tend to have greater hurricane activity. Atlantic hurricanes are the best studied and have the most data available. An analysis of satellite observations from 1983 to 2005 has shown a trend toward increasing severity—albeit decreasing frequency—of Atlantic hurricanes. Modeling of future tropical cyclones suggests that their intensity may increase 2–11% by 2100 and that the average storm will bring 20% more rainfall.

SEA LEVEL RISE

Between 1901 and 2010, the global sea level rose ~200 mm, or ~1.7 mm per year on average. From 1993 to 2010, the rate of rise nearly doubled—i.e., to 3.2 mm annually. Most of this sea level rise has resulted from the thermal expansion of water. Glacial ice melt is the second greatest factor, and its contribution is accelerating. By 2100, global sea level may rise by 0.8–2 m, with an annual rate of rise of 8–16 mm at the century's end. A large section of the West Antarctic ice sheet has begun to fall apart, and its melting alone may cause sea level to rise by ≥3 m in coming centuries.

Sea level rise is not uniform. The rate of rise on the eastern seaboard of North America has been roughly double the global rate. Compounding sea level rise is the subsidence of coastal areas due to human settlement. In the absence of levee upgrades, an estimated 170 million people living near coasts worldwide will be at risk of flooding in 2100 because of the combined effects of subsidence, erosion, and sea level rise.

Along with extreme storms and overuse of coastal aquifers, rising seas also contribute to salinization of coastal groundwater. About 1 billion people rely on coastal aquifers for potable water.

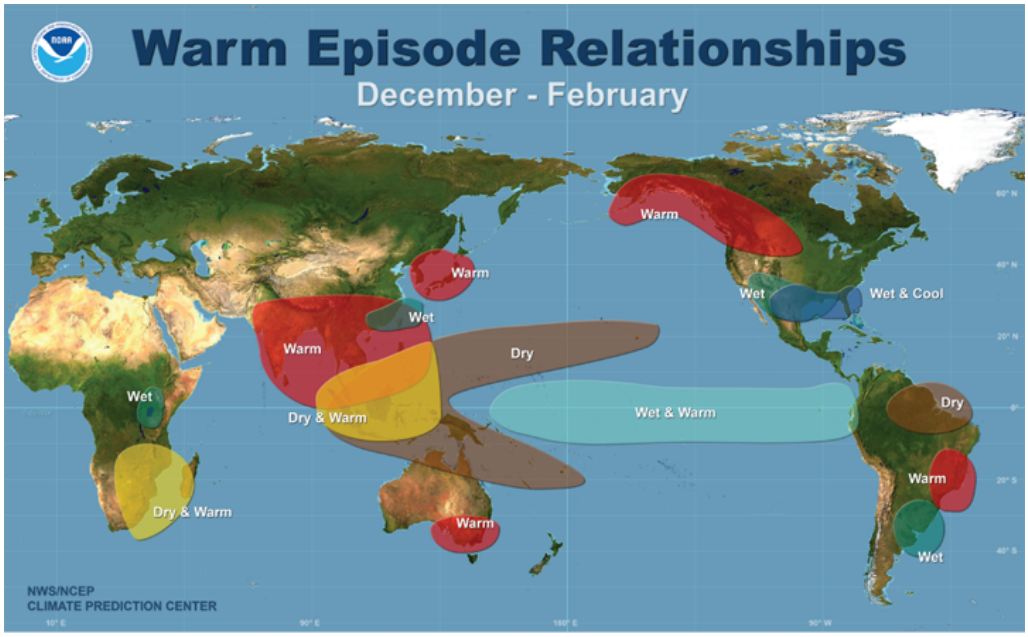
EL NIÑO SOUTHERN OSCILLATION

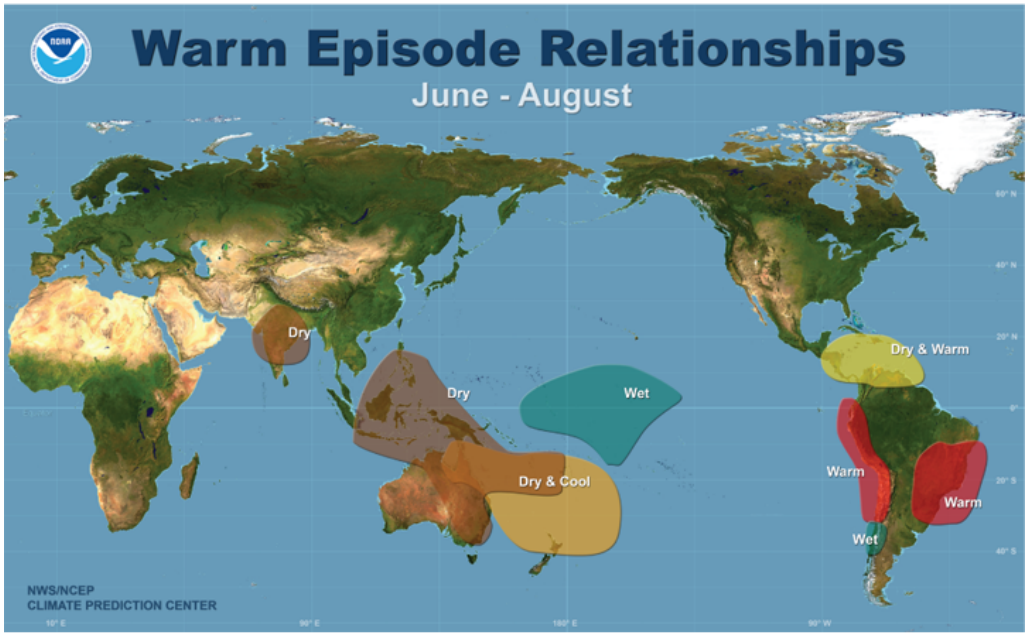
The *El Niño Southern Oscillation* (ENSO) refers to periodic changes in water temperature in the eastern Pacific Ocean that occur roughly every 5 years. ENSO cycles have dramatic effects on weather around the globe. Warmer-than-average water temperatures in the eastern Pacific define *El Niño events* (see below), whereas cooler-than-average water temperatures define La Niña periods. Evidence is accruing that climate change may be increasing the frequency and severity of El Niño events.

El Niño events drive alterations in weather worldwide **(Fig. 120-4)** and are associated with extreme events and consequently higher rates of morbidity and mortality. Hurricane Mitch, one of the most powerful hurricanes ever observed, with winds reaching 290 km/h, dropped 1–1.8 m (3–6 feet) of rain over 72 h on parts of Honduras and Nicaragua. As a result of this storm, 11,000 people died and 2.7 million were displaced. Outbreaks of cholera, leptospirosis, and dengue occurred in the storm's aftermath.

FIGURE 120-4

Characteristic weather anomalies, by season, during El Niño events. *(Source: www.cpc.ncep.noaa.gov/products/precip/CWlink/ENSO/ENSO-Global-Impacts/High-Resolution/.)*





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

POPULATION MIGRATION AND CONFLICT

The final common outcome of all climate-change effects is often population migration. Sea level rise, extreme heat and precipitation, droughts, and salinization of water supplies all conspire to make regions (including some inhabited by humans for millennia) uninhabitable. Among climate-change migrants in the near future may be the inhabitants of low-lying South Pacific islands that are vulnerable to sea level rise and residents of the Alaskan archipelago, where melting of permafrost has rendered traditional means of cold food storage difficult.

Climate change may also be contributing to humanitarian crises and conflicts. A severe 2011 drought in East Africa may have incited the Somali famine that resulted in 1 million refugees; mortality rates reached 7.4/10,000 in some refugee camps. Crop losses associated with the 2010 Russian heat wave led Russia to halt grain exports, causing higher grain prices on the world market and food riots in developing nations.

EFFECTS OF CLIMATE CHANGE ON INFECTIOUS DISEASE

The incidence of most, if not all, infectious diseases depends on climate. For any given infection, however, climate change is but one of many factors that determine disease epidemiology, and often it is not the most influential factor. Even in instances in which climate change creates conditions favorable to the spread of infections, diseases may be kept in check through interventions such as vector control or antibiotic treatment.

Detecting climate-change influence on an emerging human disease can be challenging. Research with animal pathogens, which in most instances are less well monitored and intervened upon than that with their human counterparts, has suggested how climate change may influence disease spread. For example, the life cycle of nematode parasites of caribou and musk oxen shortens as temperatures rise. As the Arctic has warmed, higher nematode burdens and consequently higher rates of morbidity and mortality have been observed. Other examples from animals, such as the spread of the protozoan parasite *Perkinsus marinus* in oysters, demonstrate how warming can enable range expansion of pathogens previously held in check by colder temperatures.

As these and other examples from studies of animals make clear, the influence of climate change on infectious diseases can be pronounced. The following sections deal with the infectious diseases for which research has explored the influence of climate change.

VECTOR-BORNE DISEASE

Because insects are cold-blooded, ambient temperature dictates their geographic distribution. With increases in temperatures (in particular, nighttime minimum temperatures), insects are freed to move poleward and up mountainsides. At the same time, as new areas become climatically suitable, current mosquito habitats may become unsuitable as a result of heat extremes.

In addition, insects tend to be sensitive to water availability. Mosquitoes that transmit malaria, dengue, and other infections may breed in pools of water created by heavy downpours. As has been observed in the Amazon, breeding pools can also appear during periods of drought when rivers recede and leave behind stagnant pools of water for *Anopheles* mosquitoes. These circumstances have raised interest in the potentially favorable impact of water-cycle intensification on the spread of mosquito-borne disease.

Malaria

TEMPERATURE

Higher temperatures promote higher mosquito-biting rates, shorter parasite reproductive cycles, and the potential for the survival of mosquito vectors of *Plasmodium* infection in locations previously too cold to sustain them. Modeling experiments have identified highland areas of East Africa and South America as perhaps most vulnerable to increased malarial incidence as a result of rising temperatures. In addition, an analysis of interannual malaria in Ecuador and Colombia has documented a greater incidence of malaria at higher altitudes in warmer years. Highland populations may be more vulnerable to malaria epidemics because they lack immunity.

Although rising temperature has the potential to expand the viable range of disease, malaria incidence is not associated with temperature in a strictly linear fashion. While mosquitoes and parasites may adapt to a warming climate, the present optimal temperature for malaria transmission is ~25°C, with a range of transmission temperatures between 16°C and 34°C. Rising temperatures also can have differential effects on parasite development during external incubation and on the mosquitoes' gonotrophic cycle. Asynchrony between these two temperature-sensitive processes has been shown to decrease the vectorial capacity of mosquitoes.[1]

[1] rVc is the vectorial capacity relative to the vector-to-human population ratio and is defined by the equation $rVc = a^2 b_h b_m e^{-\mu m n}/\mu_m$ where a is the vector biting rate; b_h is the probability of vector-to-human transmission per bite; b_m is the probability of human-to-vector infection per bite; n is the duration of the extrinsic incubation period; and μ_m is the vector mortality rate.

PRECIPITATION

The abundance of *Anopheles* mosquitoes is strongly correlated with the availability of surface-water pools for mosquito breeding, and biting rates have been linked to soil moisture (a surrogate for breeding pools). Research in the East African highlands has documented that increased variance in rainfall over time has strengthened the association between precipitation and disease incidence. These disease-promoting effects of precipitation may be countered by the potential for extreme rainfall to flush mosquito larvae from breeding sites.

PROJECTIONS

Climate models have begun to deliver output on regional scales, permitting projections of climate-suitable regions to assist national and local health authorities. Climate models speak to the temperature and precipitation ranges necessary for malaria transmission but do not—and cannot—account for the capacity of malaria control programs to halt the spread of disease. The global reduction in malaria distribution over the past century makes it clear that, even with climate change, malaria occurs in far fewer places today because of public health interventions.

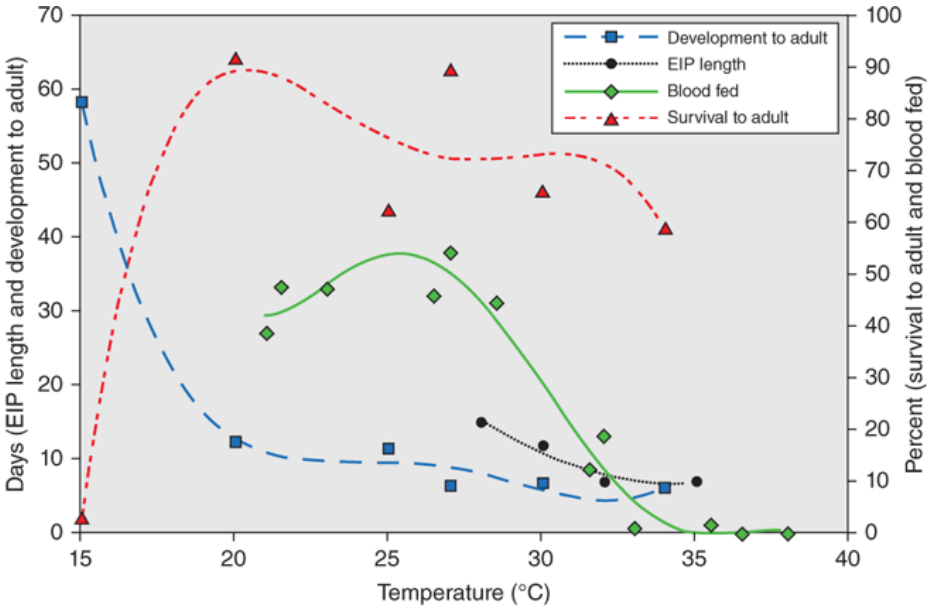
Despite intensive efforts, malaria remains the single greatest vector-borne disease cause of morbidity and death in the world. Particularly in regions that are most affected by malaria and where the public health infrastructure is inadequate to contain it, climate modeling may provide a useful tool in determining where the disease may spread. Modeling studies in sub-Saharan Africa have suggested that, although East African nations may encompass regions that will become more climatically suitable for malaria over this century, West African nations may not. By 2100, temperatures in West Africa may largely exceed those optimal for malaria transmission, and the climate may become drier; in contrast, higher temperatures and changes in precipitation may allow malaria to move up the mountainsides of East African countries. Climate change may create conditions favorable to malaria in subtropical and temperate regions of the Americas, Europe, and Asia as well.

Dengue

Like malaria epidemics, dengue fever epidemics depend on temperature **(Fig. 120-5)**. Higher temperatures increase the rate of larval development and accelerate the emergence of adult *Aedes* mosquitoes. The daily temperature range may also influence dengue virus transmission, with a smaller range corresponding to a higher transmission potential. Temperatures <15°C or >36°C substantially reduce mosquito feeding. In a *Rhesus* model of dengue, viral replication can occur in as little as 7 days with temperatures of >32–35°C; at 30°C, replication takes ≥12 days; and replication does not reliably occur at 26°C. Research on dengue in New Caledonia has shown peak transmission at ~32°C, reflecting combined effects of a shorter extrinsic incubation period, a higher feeding frequency, and more rapid development of mosquitoes. Along with temperature, peak relative humidity is a strong predictor of dengue outbreaks.

FIGURE 120-5

Effects of temperature on variables associated with dengue transmission. Shown are the number of days required for development of immature *Aedes aegypti* mosquitoes to adults, the length of the dengue virus type 2 extrinsic incubation period (EIP), the percentage of *A. aegypti* mosquitoes that complete a blood meal within 30 min after a blood source is made available, and the percentage of hatched *A. aegypti* larvae surviving to adulthood. *(Reproduced from CW Morin et al: Climate and dengue transmission: Evidence and implications. Environ Health Perspect 121:1264, 2013.)*



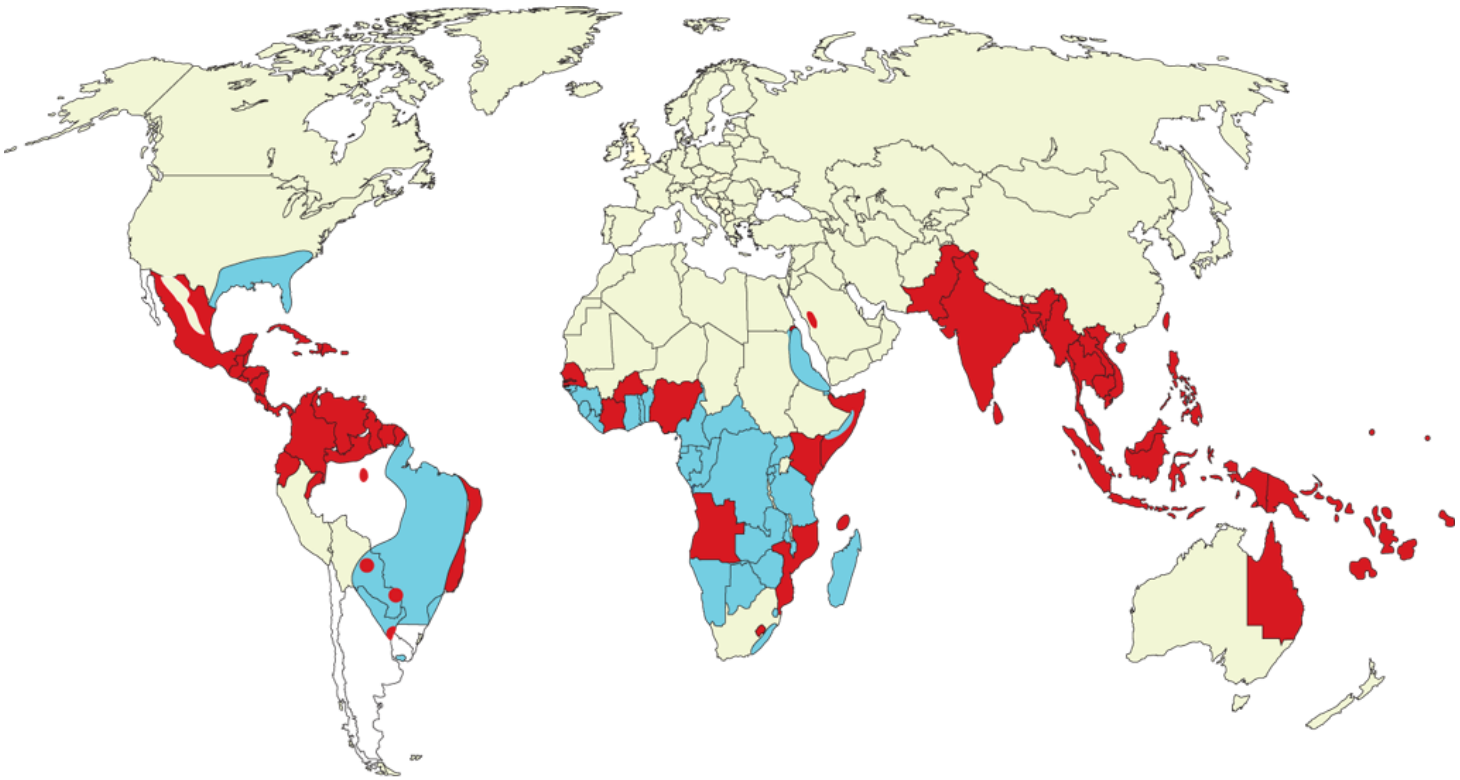
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The association between dengue epidemics and precipitation is less consistent in the peer-reviewed literature, possibly because of the mosquito vector's greater reliance on domestic breeding sites than on natural pools of water. For instance, in some studies, increased access to a piped water supply has been linked to dengue epidemics, presumably because of associated increased domestic water storage. Nonetheless, several studies have established rainfall as a predictor of the seasonal timing of dengue epidemics.

The current global distribution of dengue largely overlaps the geographic spread of *Aedes* mosquitoes **(Fig. 120-6)**. The presence of *Aedes* without dengue endemicity in large regions of North and South America and Africa illustrates the relevance of variables other than climate to disease incidence. Nevertheless, coupled climatic–epidemiologic modeling suggests dramatic shifts in the relative vectorial capacity for dengue by the end of this century should little or no mitigation of greenhouse gas emissions occur **(Fig. 120-7)**. Given the joint effects of climate change and population growth, the number of people exposed to *A. aegypti* globally may nearly double by 2100 from roughly 4 billion to 8 billion or more.

FIGURE 120-6

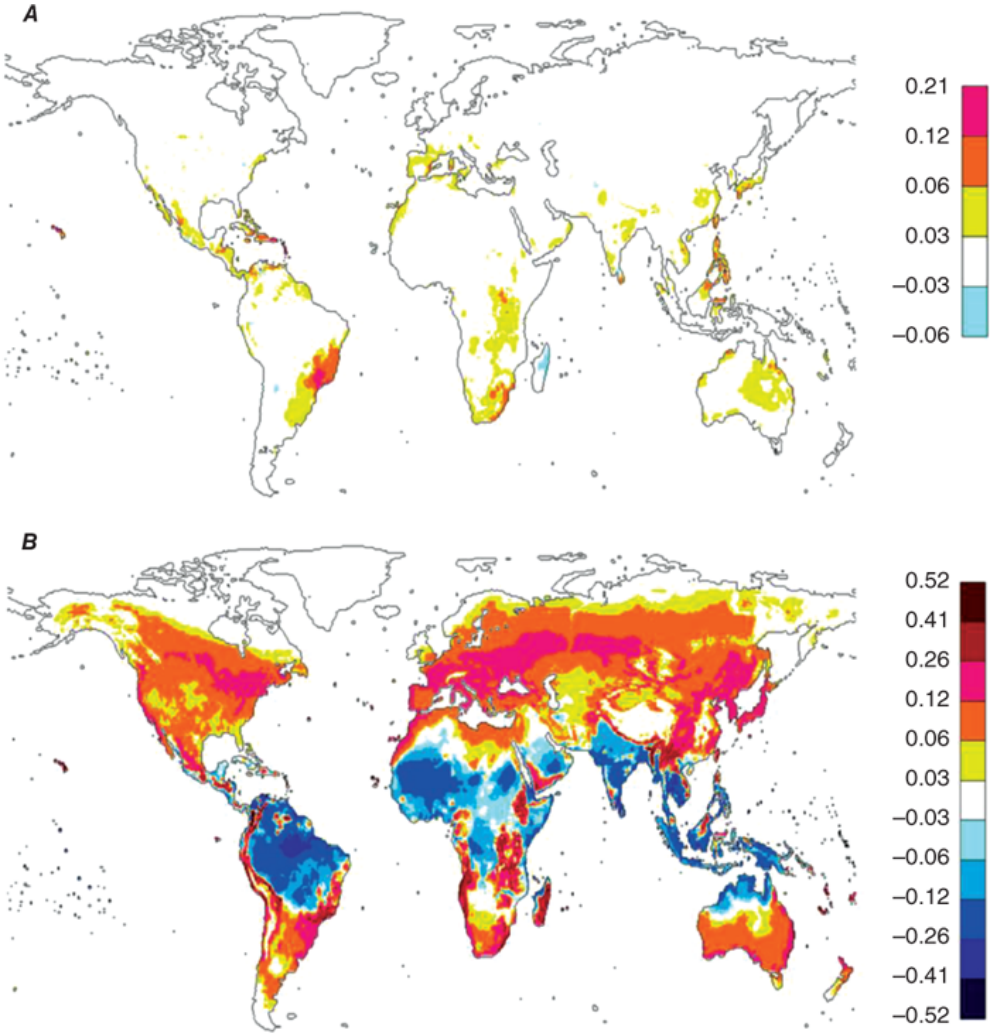
Distribution of *Aedes aegypti* mosquitoes (*turquoise*) and dengue fever epidemics (*red*). *(Map produced by the Agricultural Research Service of the U.S. Department of Agriculture.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 120-7

Trend of annually averaged global dengue epidemic potential (*rVc*). Differences in rVc are based on 30-year averages of temperature and daily temperature range. ***A.*** Differences between 1980–2009 and 1901–1930. ***B.*** Differences between 2070–2099 and 1980–2009. The mean value of rVc was averaged from five global climate models under RCP8.5, a scenario of high greenhouse-gas emission. The color bar describes the values of the rVc. *(From J Liu-Helmersson et al: Vectorial capacity of Aedes aegypti: Effects of temperature and implications for global dengue epidemic potential. PLoS ONE 9:e89783, 2014 [doi:10.1371/journal.pone.0089783].)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Other Arbovirus Infections

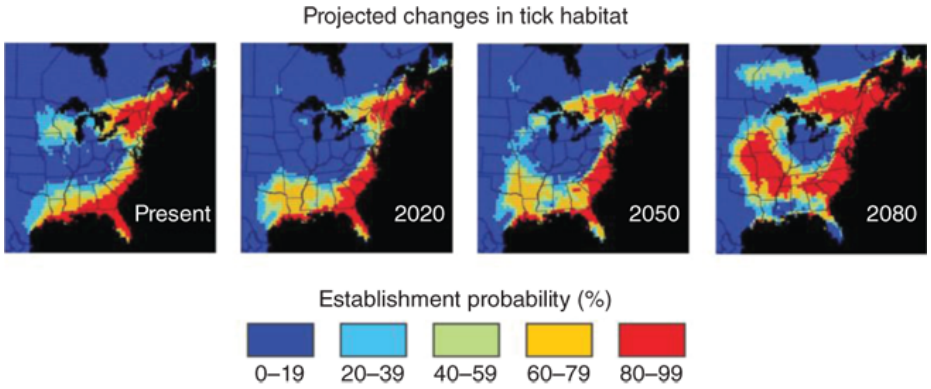
Climate change may favor increased geographic spread of other arboviral diseases, including Zika virus disease, chikungunya virus disease, West Nile virus disease, and eastern equine encephalitis. Zika virus moved to the Western Hemisphere from French Polynesia around 2013 and rapidly spread in Brazil in 2016. Although air travel was essential for the delivery of the virus to the Americas, the available evidence suggests that the 2015 El Niño event provided an optimal climate for the infection to take root and spread. *A. aegypti* is the primary vector for Zika virus. Chikungunya virus disease emerged in Italy in 2007, having previously been mostly a disease of African nations. Climate models predict that, should competent vectors be present, conditions will be suitable for chikungunya virus to gain a foothold in Western Europe, especially France, in the first half of the twenty-first century. In North America, areas favorable to West Nile virus outbreaks are expected to shift northward in this century. Current hotspots in North America are the California Central Valley, southwestern Arizona, southern Texas, and Louisiana, which have both compatible climates and avian reservoirs for the disease. By mid-century, the upper Midwest and New England will be more climatically suited to West Nile virus; by the end of the century, the area of risk may shift even further north to southern Canada. Whether the disease will ultimately move northward will depend on reservoir availability and mosquito control programs, among other factors.

Lyme Disease

In the past few decades, *Ixodes scapularis*, the primary tick vector for Lyme disease as well as for anaplasmosis and babesiosis in New England, has become established in Canada because of warming temperatures. With climate change, the range of the tick is expected to expand further **(Fig. 120-8)**.

FIGURE 120-8

Present and projected probability of establishment of *Ixodes scapularis.* *(From U.S. National Climate Assessment 2014, adapted from JS Brownstein et al: Effect of climate change on Lyme disease risk in North America. Ecohealth 2:38, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

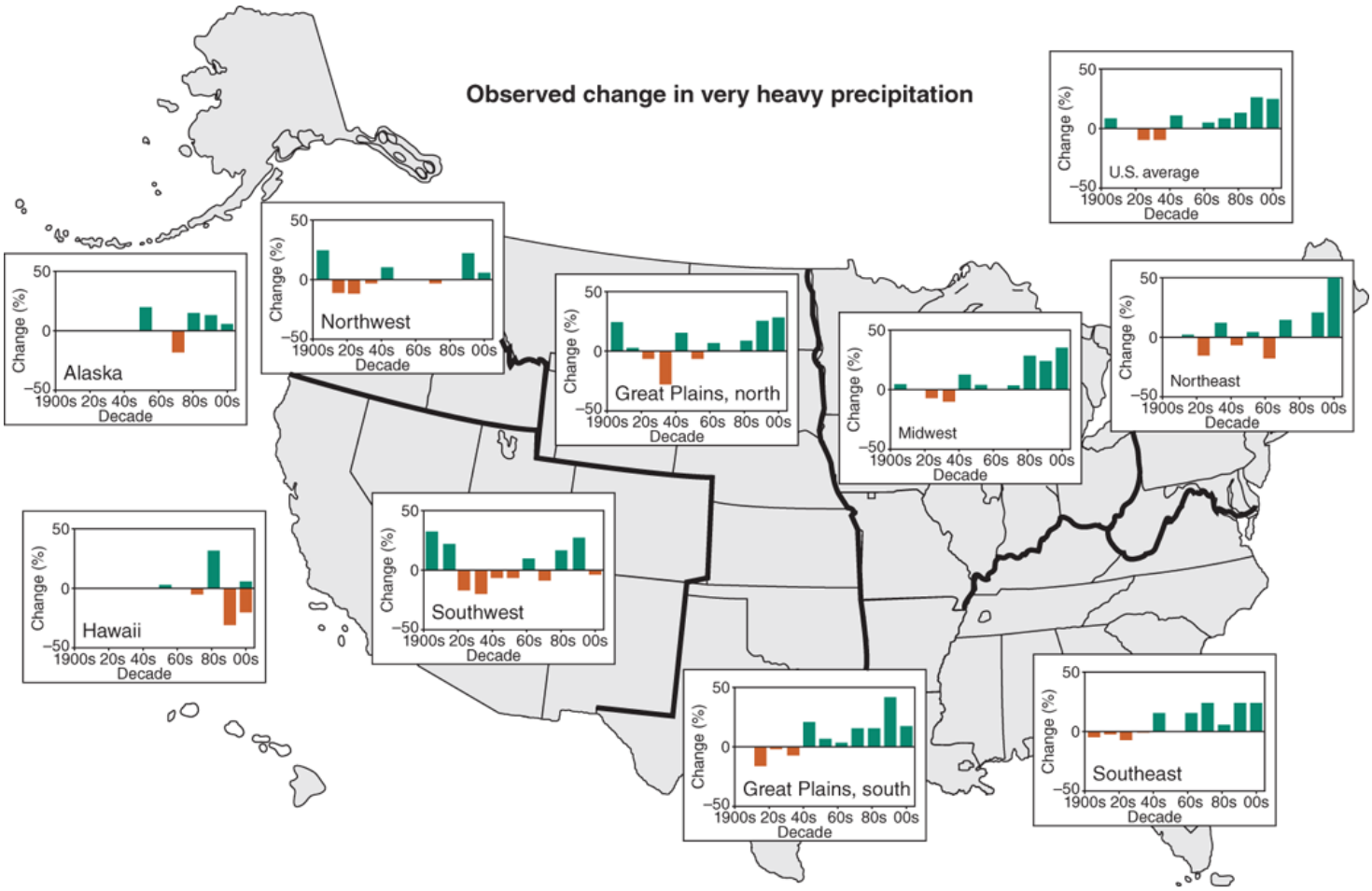
Lyme disease, caused by the spirochete *Borrelia burgdorferi*, is the most commonly reported vector-borne disease in North America, with ~30,000 cases per year. The model used in Fig. 120-8 showed 95% accuracy in predicting current *I. scapularis* distribution and suggests substantial expansion of tick habitat and consequently of populations at risk for the diseases this tick transmits, particularly in Quebec, Iowa, and Arkansas, by 2080. Of note, some areas on the Gulf Coast may become less suitable for ticks by the end of the century.

WATERBORNE DISEASE

Outbreaks of waterborne disease are associated with heavy rainfall events. A review of 548 waterborne disease outbreaks in the United States found that 51% were preceded by precipitation levels above the 90th percentile. Since 1900, most regions of the United States except the Southwest and Hawaii have experienced an increase in heavy downpours **(Fig. 120-9)**, with the greatest intensification of the water cycle in New England and Alaska. Climate models suggest that by 2100 daily heavy-precipitation events, which are defined as a cumulative daily amount that now occurs once every 20 years, will increase nationwide **(Fig. 120-10)**. This scenario may be from two to as much as five times more likely, depending on the extent of greenhouse gas emission reductions achieved early in the twenty-first century.

FIGURE 120-9

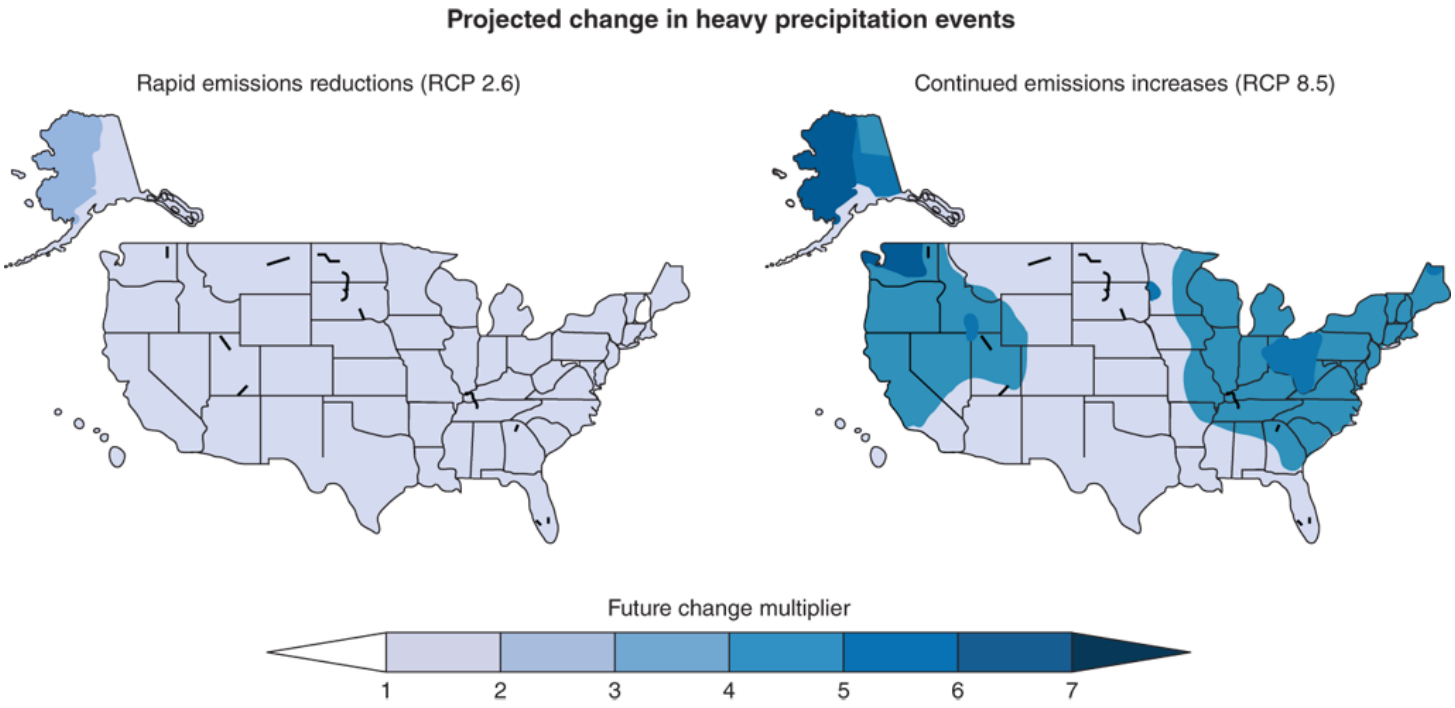
Percentage changes in the annual amount of precipitation falling in very heavy events, defined as the heaviest 1% of all daily events from 1901 to 2012 for each region. Changes are relative to a 1901–1960 average for all regions except values for Alaska and Hawaii, which are relative to the 1951–1980 average. *(From U.S. National Climate Assessment 2014, NOAA National Climate Data Center/Cooperative Institute for Climate and Satellites, North Carolina.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 120-10

Increased frequency of extreme daily precipitation events (defined as a daily amount that now occurs once in 20 years) by the latter part of the twenty-first century (2081–2100) compared to the frequency in the latter part of the twentieth century (1981–2000). A representative concentration pathway (RCP) describes a plausible climate future based on a net radiative forcing (e.g., 2.6 or 8.5) in 2100. *(From U.S. National Climate Assessment 2014, NOAA National Climate Data Center/Cooperative Institute for Climate and Satellites, North Carolina.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Most disease outbreaks after heavy precipitation occur through contamination of drinking-water supplies. While outbreaks related to surface-water contamination generally occur within a month of the precipitation event, disease outbreaks from groundwater contamination tend to occur ≥2 months later. According to a review of published reports of waterborne disease outbreaks, *Vibrio* and *Leptospira* species are the pathogens most commonly involved in the wake of heavy precipitation.

Combined Sewer Systems

Roughly 40 million people in the United States and millions more around the world rely on combined sewer systems in which storm water and sanitary wastewater are conveyed in the same pipe to treatment facilities. These systems were designed on the basis of the nineteenth-century climate, in which heavy downpours were less frequent than they are today. The frequency of combined sewer overflows resulting in untreated sewage discharge, usually into freshwater bodies, has been increasing in cities worldwide. Overflows are associated with discharges of heavy metals and other chemical pollutants as well as a variety of pathogens. Outbreaks of hepatitis A, *Escherichia coli* O157:H7 infection, and cryptosporidial disease have been associated with sewer overflows in the United States.

Rising Temperatures and *Vibrio* Species

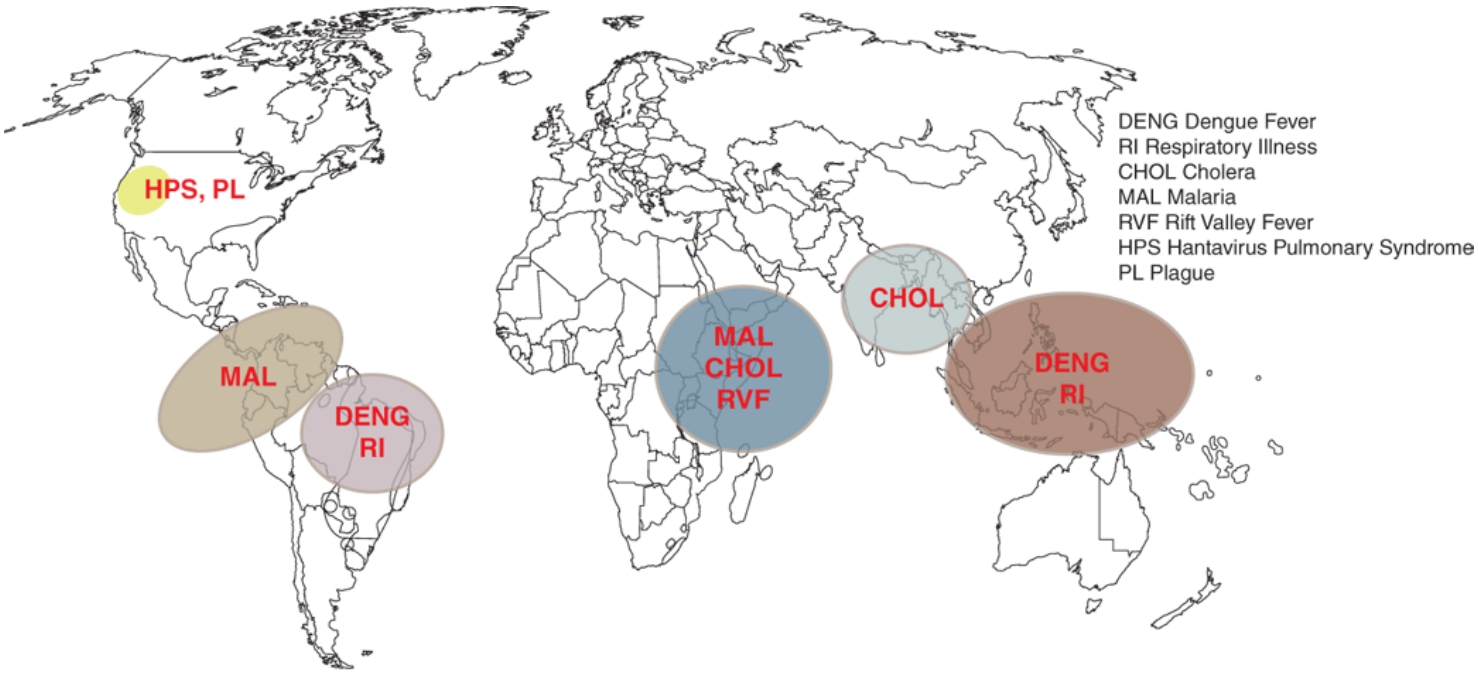
Warmer temperatures favor proliferation of *Vibrio* species and disease outbreaks, as has been demonstrated in countries surrounding the Baltic Sea, Chile, Israel, northwestern Spain, and the U.S. Pacific Northwest. Around the Baltic Sea, outbreaks of *Vibrio* infection may be particularly likely because of faster warming near the poles and the sea's relatively low salt content. In 2004, a *Vibrio parahaemolyticus* outbreak arising from consumption of Alaskan oysters occurred. This pathogen was unknown in Alaskan oysters prior to this event and extended the known geographic range of the disease 1000 km northward.

ENSO-Related Outbreaks

In the past, El Niño events were used as a model to investigate the potential for extreme weather–related infectious disease epidemics occurring in association with climate change. Recent evidence indicates that climate change itself may be strengthening El Niño events. These events tend to promote epidemic infections in certain regions **(Fig. 120-11)**.

FIGURE 120-11

Characteristic patterns of disease outbreaks associated with El Niño events, determined on the basis of 2006–2007 conditions. *(From A Anyamba et al: Developing global climate anomalies suggest potential disease risks for 2006–2007. Int J Health Geogr 5:60, 2006.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Associations of El Niño with outbreaks of Rift Valley fever in eastern and southern Africa have been known since the 1950s. El Niño favors wet conditions suitable for the insect vectors of the disease in these regions. Given the strong association between El Niño conditions and disease incidence, models have successfully predicted Rift Valley fever epidemics in humans and animals. In the 2006–2007 El Niño season, for example, outbreaks of Rift Valley fever were accurately predicted 2–6 weeks prior to epidemics in Somalia, Kenya, and Tanzania.

El Niño has had inconsistent associations with malaria incidence in African countries. Some of the strongest associations between El Niño and malaria have been identified in South Africa and Swaziland, where available data on incidence are relatively robust; however, even in these instances, the observed increased risk did not reach statistical significance. A stronger link to El Niño has been found in several studies done in South America. Research on malaria incidence in Colombia between 1960 and 2006 found that a 1°C temperature rise contributed to a 20% increase in incidence.

El Niño years are often associated with an increased incidence of dengue. Research on dengue outbreaks in Thailand from 1996 to 2005 revealed that 15–22% of the variance in monthly dengue disease incidence was attributable to El Niño. In South America, data on dengue outbreaks between 1995 and 2010 showed an increased incidence during the El Niño events of 1997–1998 and 2006–2007.

CLIMATE CHANGE, POPULATION DISPLACEMENT, AND INFECTIOUS DISEASE EPIDEMICS

For many reasons, including freshwater shortages, flooding, food shortages, and climate change–driven conflicts, climate change has and will continue to put pressure on human populations to move. Human migrations have long been associated with epidemic disease in the migrating populations themselves and in the communities in which they settle. The specific pathogens and patterns of disease that may appear after population migration relate to endemic diseases present in the migrant populations.

Large-scale migrations are common after extreme precipitation events. Hurricane Katrina, for instance, displaced about 1 million people from the U.S. Gulf Coast. Among Katrina refugees, outbreaks of respiratory, diarrheal, and skin diseases were most common. While attribution of a single weather event to increased greenhouse-gas emissions is difficult, research can provide information on the likelihood of such events. It is expected, for example, that warming by 1°C increases the odds of a storm as strong as or stronger than Katrina two- to sevenfold.

In the developing world, infectious disease outbreaks associated with population displacement due to extreme weather events may be especially hard to detect and respond to. Mitigation of disease risk requires overlaying of climate-related migration risk with foci of disease epidemics.

A BROADER VIEW OF CLIMATE CHANGE AND HEALTH

Climate change has far-reaching implications for the distribution and spread of infectious diseases worldwide. However, the greatest disease burdens related to climate change may not be due to infections. Because climate change disrupts the foundations of health, such as access to safe drinking water and food, it has the potential to undermine progress against major existing health problems such as malnutrition. In addition, resource scarcity and climate instability are increasingly associated with conflicts. Scholars have argued that events related to climate change were a factor in the revolutions of the Arab Spring and the Syrian civil war.

The public health response to climate change entails both mitigation and adaptation measures. *Mitigation* represents primary prevention and involves the reduction of greenhouse gas emissions into the atmosphere. Although no clear safety threshold of greenhouse gas emissions has been agreed upon, national governments from the major industrialized countries have agreed to set a warming target of <2°C above preindustrial levels by 2050; the attainment of this goal will require reducing greenhouse gas emissions by 40–70% below 2010 levels. The 2016 Paris Agreement on climate change provides a framework for the establishment of a global carbon market that may speed reductions in greenhouse gas emissions, along with several other seminal provisions. Mitigation also confers co-benefits, including better air quality that results when fewer bio- or fossil fuels are burned. Biofuel-burning cookstoves, for instance, used by some 3 billion people around the world, release air pollution that constitutes about one-fourth of the global black-carbon emissions that warm the planet and kill roughly 4 million people a year. Clean cookstoves simultaneously mitigate climate change and indoor air pollution–related mortality.

Adaptation represents secondary prevention and is aimed at reducing the harms associated with sea level rise, heat waves, floods, droughts, wildfires, and other greenhouse gas–driven events. The efficacy of adaptation is constrained by the challenges inherent in predicting the precise location, duration, and severity of extreme weather events and flooding related to sea level rise, among other considerations.

FURTHER READING

Anyamba A et al: Prediction of a Rift Valley fever outbreak. Proc Natl Acad Sci USA 106:955, 2009.

Cai W et al: Increasing frequency of extreme El Niño events due to greenhouse warming. Nat Clim Change 4:111, 2014.

Caminade C et al: Impact of climate change on global malaria distribution. Proc Natl Acad Sci USA 111:3286, 2014.

Colón-González FJ et al: The effects of weather and climate change on dengue. PLoS Negl Trop Dis 7:e2503, 2013. [PubMed: 24244765]

Gething PW et al: Climate change and the global malaria recession. Nature 465:342, 2010.

National Climatic Data Center: Mitch: The deadliest Atlantic hurricane since 1780. Available from *www.ncdc.noaa.gov/oa/reports/mitch/mitch.html*. Accessed January 13, 2017.

Ogden NH et al: Climate change and the potential for range expansion of the Lyme disease vector *Ixodes scapularis* in Canada. Int J Parasitol 36:63, 2006.

Paaijmans KP et al: Temperature-dependent pre-bloodmeal period and temperature-driven asynchrony between parasite development and mosquito biting rate reduce malaria transmission intensity. PLoS One 8:e55777, 2013. [PubMed: 23383280]

Watts DM et al: Effect of temperature on the vector efficiency of *Aedes aegypti* for dengue 2 virus. Am J Trop Med Hyg 36:143, 1987.

Zhou G et al: Association between climate variability and malaria epidemics in the East African highlands. Proc Natl Acad Sci USA 101:2375, 2004.

Chapter 121: Pneumonia

Lionel A. Mandell; Richard Wunderink

DEFINITION

Pneumonia is an infection of the pulmonary parenchyma. Despite being the cause of significant morbidity and mortality, it is often misdiagnosed, mistreated, and underestimated. Pneumonia historically was typically classified as community-acquired (CAP), hospital-acquired (HAP), or ventilator-associated (VAP). A fourth category, health care–associated pneumonia (HCAP), was introduced recently. This category was meant to encompass those cases of CAP that were caused by multidrug-resistant (MDR) pathogens typically associated with HAP. Unfortunately, the original definitions appear to have been overly sensitive, resulting in the treatment of a high proportion of patients who had community-onset pneumonia with broad-spectrum antibiotics consistent with HAP treatment. Retrospective studies have actually suggested a worse outcome when broad-spectrum antibiotics were used in these cases.

Rather than relying on a predefined subset or category of pneumonia cases, it is likely to be of greater value to assess each case individually on the basis of risk factors for infection with an MDR organism. Rather than originating in primary pneumonia research, the original HCAP definition was modified from a study of health care–associated bacteremia. Recent studies have more closely identified patients at risk for pathogens resistant to the antibiotics usually used; have defined risk factors for infection with methicillin-resistant *Staphylococcus aureus* (MRSA) independent of other MDR pathogens; and have found that at least two, if not three, risk factors are required before the probability of drug-resistant pathogens is sufficient to influence initial empirical broad-spectrum antibiotic therapy. These risk factors are listed in **Table 121-1**.

TABLE 121-1

Risk Factors for Pathogens Resistant to Usual Therapy for Community-Acquired Pneumonia[a]

MULTIDRUG-RESISTANT GRAM-NEGATIVE BACTERIA AND MRSA	NOSOCOMIAL MRSA	COMMUNITY-ACQUIRED MRSA
Hospitalization ≥2 days in previous 90 days Use of antibiotics in previous 90 days Immunosuppression Nonambulatory status Tube feedings Gastric acid suppression Severe COPD or bronchiectasis[b]	Hospitalization ≥ 2 days in previous 90 days Use of antibiotics in previous 90 days Chronic hemodialysis in previous 30 days Documented prior MRSA colonization Congestive heart failure Gastric acid suppression	Cavitary infiltrate or necrosis Gross hemoptysis Neutropenia Erythematous rash Concurrent influenza Young, previously healthy status Summer-month onset

[a]Cephalosporin/macrolide or respiratory fluoroquinolone. [b]Risk for Pseudomonas aeruginosa infection.

Abbreviations: COPD, chronic obstructive pulmonary disease; MRSA, methicillin-resistant *Staphylococcus aureus*.

This chapter deals with pneumonia in patients who are not considered to be immunocompromised. **Pneumonia in severely immunocompromised patients, some of whom overlap with the groups of patients considered in this chapter, warrants separate discussion (see Chaps. 70, 138, and 197).**

PATHOPHYSIOLOGY

Pneumonia results from the proliferation of microbial pathogens at the alveolar level and the host's response to those pathogens. Microorganisms gain access to the lower respiratory tract in several ways. The most common is by aspiration from the oropharynx. Small-volume aspiration occurs frequently during sleep (especially in the elderly) and in patients with decreased levels of consciousness. Rarely, pneumonia occurs via hematogenous spread (e.g., from tricuspid endocarditis) or by contiguous extension from an infected pleural or mediastinal space.

Mechanical factors are critically important in host defense. The hairs and turbinates of the nares capture larger inhaled particles before they reach the lower respiratory tract. The branching architecture of the tracheobronchial tree traps microbes on the airway lining, where mucociliary clearance and local antibacterial factors either clear or kill the potential pathogen. The gag and cough reflexes offer critical protection from aspiration. In addition, the normal flora adhering to mucosal cells of the oropharynx, whose

components are remarkably constant, prevents pathogenic bacteria from binding and thereby decreases the risk of pneumonia.

When these barriers are overcome or when microorganisms are small enough to be inhaled to the alveolar level, resident alveolar macrophages are extremely efficient at clearing and killing pathogens. Macrophages are assisted by proteins that are produced by the alveolar epithelial cells (e.g., surfactant proteins A and D) and that have intrinsic opsonizing properties or antibacterial or antiviral activity. Once engulfed by the macrophage, the pathogens—even if they are not killed—are eliminated via either the mucociliary elevator or the lymphatics and no longer represent an infectious challenge. Only when the capacity of the alveolar macrophages to ingest or kill the microorganisms is exceeded does clinical pneumonia become manifest. In that situation, the alveolar macrophages initiate the inflammatory response to bolster lower respiratory tract defenses. The host inflammatory response, rather than proliferation of microorganisms, triggers the clinical syndrome of pneumonia. The release of inflammatory mediators, such as interleukin 1 and tumor necrosis factor, results in fever. Chemokines, such as interleukin 8 and granulocyte colony-stimulating factor, stimulate the release of neutrophils and their attraction to the lung, producing both peripheral leukocytosis and increased purulent secretions. Inflammatory mediators released by macrophages and the newly recruited neutrophils create an alveolar capillary leak equivalent to that seen in acute respiratory distress syndrome, although in pneumonia this leak is localized (at least initially). Even erythrocytes can cross the alveolar–capillary membrane, with consequent hemoptysis. The capillary leak results in a radiographic infiltrate and rales detectable on auscultation, and hypoxemia results from alveolar filling. Moreover, some bacterial pathogens appear to interfere with the hypoxemic vasoconstriction that would normally occur with fluid-filled alveoli, and this interference can result in severe hypoxemia. Increased respiratory drive in the systemic inflammatory response syndrome (Chap. 297) leads to respiratory alkalosis. Decreased compliance due to capillary leak, hypoxemia, increased respiratory drive, increased secretions, and occasionally infection-related bronchospasm all lead to dyspnea. If severe enough, the changes in lung mechanics secondary to reductions in lung volume and compliance and the intrapulmonary shunting of blood may cause respiratory failure and death.

The presence of a normal alveolar microbiota raises the possibility of an alternative pathway for development of pneumonia. This microbiota is similar to the oropharyngeal microbiota; both are predominantly gram-positive in contrast to the gram-negative milieu of the normal gastrointestinal microbiota. Rather than invasion of a sterile lower respiratory tract by pathogens to cause pneumonia, alterations in host defense may allow overgrowth of one or more components of the normal bacterial flora. The fact that many CAP pathogens are components of the normal alveolar microbiota supports this alternative-pathogenesis model. The two most likely sources of an altered alveolar microbiota are viral upper respiratory tract infections for CAP and antibiotic therapy for HAP/VAP.

PATHOLOGY

Classic pneumonia evolves through a series of pathologic changes. The initial phase is one of *edema*, with the presence of a proteinaceous exudate—and often of bacteria—in the alveoli. This phase is rarely evident in

clinical or autopsy specimens because of the rapid transition to the *red hepatization* phase. The presence of erythrocytes in the cellular intra-alveolar exudate gives this second stage its name, but neutrophil influx is more important with regard to host defense. Bacteria are occasionally seen in pathologic specimens collected during this phase. In the third phase, *gray hepatization*, no new erythrocytes are extravasating, and those already present have been lysed and degraded. The neutrophil is the predominant cell, fibrin deposition is abundant, and bacteria have disappeared. This phase corresponds with successful containment of the infection and improvement in gas exchange. In the final phase, *resolution*, the macrophage reappears as the dominant cell type in the alveolar space, and the debris of neutrophils, bacteria, and fibrin has been cleared, as has the inflammatory response.

This pattern has been described best for lobar pneumococcal pneumonia and may not apply to pneumonia of all etiologies, especially viral or *Pneumocystis* pneumonia. In VAP, respiratory bronchiolitis may precede the development of a radiologically apparent infiltrate. Because of the microaspiration mechanism, a bronchopneumonia pattern is most common in nosocomial pneumonias, whereas a lobar pattern is more common in bacterial CAP. Despite the radiographic appearance, viral and *Pneumocystis* pneumonias represent alveolar rather than interstitial processes.

COMMUNITY-ACQUIRED PNEUMONIA

ETIOLOGY

The extensive list of potential etiologic agents in CAP includes bacteria, fungi, viruses, and protozoa. Newly identified pathogens include metapneumoviruses, the coronaviruses responsible for severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS), and community-acquired strains of MRSA. Most cases of CAP, however, are caused by relatively few pathogens **(Table 121-2)**. Although *Streptococcus pneumoniae* is most common, other organisms must also be considered in light of the patient's risk factors and severity of illness. Separation of potential agents into "typical" bacterial pathogens or "atypical" organisms may be helpful. The former category includes *S. pneumoniae*, *Haemophilus influenzae*, and (in selected patients) *S. aureus* and gram-negative bacilli such as *Klebsiella pneumoniae* and *Pseudomonas aeruginosa.* The "atypical" organisms include *Mycoplasma pneumoniae*, *Chlamydia pneumoniae*, and *Legionella* species as well as respiratory viruses such as influenza viruses, adenoviruses, human metapneumovirus, and respiratory syncytial viruses. Overall, with the increasing use of pneumococcal vaccine, the incidence of pneumococcal pneumonia appears to be decreasing. Cases due to *M. pneumoniae* and *C. pneumoniae*, however, appear to be increasing in incidence, especially among young adults. Viruses may be responsible for a large proportion of CAP cases that require hospital admission, even in adults. Polymerase chain reaction (PCR)–based testing shows that viruses may be present in 20–30% of healthy adults and in the same percentage of patients with pneumonia, including those who are severely ill. The most common of these viruses are influenza, parainfluenza, and respiratory syncytial viruses. Whether they are etiologic pathogens, co-pathogens, or simply colonizers cannot always be determined. Atypical organisms cannot be cultured on standard media or seen on Gram's stain. The frequency and importance of atypical pathogens have significant implications for therapy. They are intrinsically resistant to all β-lactam

agents and must be treated with a macrolide, a fluoroquinolone, or a tetracycline. In the ~10–15% of CAP cases that are polymicrobial, the etiology usually includes a combination of typical and atypical pathogens.

TABLE 121-2

Microbial Causes of Community-Acquired Pneumonia, by Site of Care

OUTPATIENTS	HOSPITALIZED PATIENTS	
	NON-ICU	ICU
Streptococcus pneumoniae *Mycoplasma pneumoniae* *Haemophilus influenzae* *C. pneumoniae* Respiratory viruses[a]	*S. pneumoniae* *M. pneumoniae* *Chlamydia pneumoniae* *H. influenzae* *Legionella* spp. Respiratory viruses[a]	*S. pneumoniae* *Staphylococcus aureus* *Legionella* spp. Gram-negative bacilli *H. influenzae* Respiratory viruses

[a]Influenza A and B viruses, human metapneumovirus, adenoviruses, respiratory syncytial viruses, parainfluenza viruses.

Abbreviation: ICU, intensive care unit.

Anaerobes play a significant role only when an episode of aspiration has occurred days to weeks before presentation for pneumonia. The combination of an unprotected airway (e.g., in patients with alcohol or drug overdose or a seizure disorder) and significant gingivitis constitutes the major risk factor. Anaerobic pneumonias are often complicated by abscess formation and by significant empyemas or parapneumonic effusions.

S. aureus pneumonia is well known to complicate influenza infection. However, MRSA has been reported as a primary etiologic agent of CAP. While this entity is still relatively uncommon, clinicians must be aware of its potentially serious consequences, such as necrotizing pneumonia. Two important developments have led to this problem: the spread of MRSA from the hospital setting to the community and the emergence of genetically distinct strains of MRSA in the community. The community-acquired MRSA (CA-MRSA) strains may infect healthy individuals with no association with health care.

Unfortunately, despite a careful history and physical examination as well as routine radiographic studies, the causative pathogen in a case of CAP is difficult to predict with any degree of certainty; in more than one-half of cases, a specific etiology is never determined. Nevertheless, epidemiologic and risk factors may suggest the involvement of certain pathogens **(Table 121-3)**.

TABLE 121-3

Epidemiologic Factors Suggesting Possible Causes of Community-Acquired Pneumonia

FACTOR	POSSIBLE PATHOGEN(S)
Alcoholism	*Streptococcus pneumoniae*, oral anaerobes, *Klebsiella pneumoniae*, *Acinetobacter* spp., *Mycobacterium tuberculosis*
COPD and/or smoking	*Haemophilus influenzae*, *Pseudomonas aeruginosa*, *Legionella* spp., *S. pneumoniae*, *Moraxella catarrhalis*, *Chlamydia pneumoniae*
Structural lung disease (e.g., bronchiectasis)	*P. aeruginosa*, *Burkholderia cepacia*, *Staphylococcus aureus*
Dementia, stroke, decreased level of consciousness	Oral anaerobes, gram-negative enteric bacteria
Lung abscess	CA-MRSA, oral anaerobes, endemic fungi, *M. tuberculosis*, atypical mycobacteria
Travel to Ohio or St. Lawrence river valley	*Histoplasma capsulatum*
Travel to southwestern United States	Hantavirus, *Coccidioides* spp.
Travel to Southeast Asia	*Burkholderia pseudomallei*, avian influenza virus
Stay in hotel or on cruise ship in previous 2 weeks	*Legionella* spp.
Local influenza activity	Influenza virus, *S. pneumoniae*, *S. aureus*
Exposure to bats or birds	*H. capsulatum*
Exposure to birds	*Chlamydia psittaci*
Exposure to rabbits	*Francisella tularensis*
Exposure to sheep, goats, parturient cats	*Coxiella burnetii*

Abbreviations: CA-MRSA, community-acquired methicillin-resistant *Staphylococcus aureus;* COPD, chronic obstructive pulmonary disease.

EPIDEMIOLOGY

More than 5 million CAP cases occur annually in the United States. Along with influenza, CAP is the eighth leading cause of death in this country. Usually, 80% of the affected patients are treated as outpatients and 20% as inpatients. The mortality rate among outpatients is usually <5%, whereas among hospitalized patients the rate can range from ~12% to 40%, depending on whether treatment is provided in or outside of the intensive care unit (ICU). In the United States, CAP is the number one cause of death from infection among patients >65 years of age. Further compounding its impact is the fact that 18% of hospitalized CAP patients are readmitted within 1 month of discharge. CAP results in more than 1.2 million hospitalizations and more than 55,000 deaths annually. The overall yearly cost associated with CAP is estimated at $17 billion. The incidence rates are highest at the extremes of age. The overall annual rate in the United States is 12 cases/1000 persons, but the figure increases to 12–18/1000 among children <4 years of age and to 20/1000 among persons >60 years of age.

The risk factors for CAP in general and for pneumococcal pneumonia in particular have implications for treatment regimens. Risk factors for CAP include alcoholism, asthma, immunosuppression, institutionalization, and an age of ≥70 years. In the elderly, factors such as decreased cough and gag reflexes as well as reduced antibody and Toll-like receptor responses increase the likelihood of pneumonia. Risk factors for pneumococcal pneumonia include dementia, seizure disorders, heart failure, cerebrovascular disease, alcoholism, tobacco smoking, chronic obstructive pulmonary disease (COPD), and HIV infection. CA-MRSA pneumonia is more likely in patients with skin colonization or infection with CA-MRSA. Enterobacteriaceae tend to infect patients who have recently been hospitalized and/or received antibiotic therapy or who have comorbidities such as alcoholism, heart failure, or renal failure. *P. aeruginosa* is a particular problem in patients with severe structural lung disease, such as bronchiectasis, cystic fibrosis, or severe COPD. Risk factors for *Legionella* infection include diabetes, hematologic malignancy, cancer, severe renal disease, HIV infection, smoking, male gender, and a recent hotel stay or ship cruise.

CLINICAL MANIFESTATIONS

CAP can vary from indolent to fulminant in presentation and from mild to fatal in severity. Manifestations of progression and severity include both constitutional findings and those limited to the lung and associated structures.

The patient is frequently febrile with tachycardia or may have a history of chills and/or sweats. Cough may be either nonproductive or productive of mucoid, purulent, or blood-tinged sputum. Gross hemoptysis is suggestive of CA-MRSA pneumonia. Depending on severity, the patient may be able to speak in full sentences or may be very short of breath. If the pleura is involved, the patient may experience pleuritic chest pain. Up to

20% of patients may have gastrointestinal symptoms such as nausea, vomiting, and/or diarrhea. Other symptoms may include fatigue, headache, myalgias, and arthralgias.

Findings on physical examination vary with the degree of pulmonary consolidation and the presence or absence of a significant pleural effusion. An increased respiratory rate and use of accessory muscles of respiration are common. Palpation may reveal increased or decreased tactile fremitus, and the percussion note can vary from dull to flat, reflecting underlying consolidated lung and pleural fluid, respectively. Crackles, bronchial breath sounds, and possibly a pleural friction rub may be heard on auscultation. The clinical presentation may not be so obvious in the elderly, who may initially display new-onset or worsening confusion and few other manifestations. Severely ill patients may have septic shock and evidence of organ failure.

The risk of cardiac complications secondary to enhanced inflammation and procoagulant activity is increased. These complications include myocardial infarction, congestive heart failure, and arrhythmias, particularly in the elderly. In pneumococcal CAP, the increased risk of acute coronary events may be partially driven by pneumolysis, which increases platelet activation. Up to 90% of acute coronary syndromes occur in the first week after onset of CAP, and the risk of new-onset congestive heart failure in elderly hospitalized CAP patients can extend up to 1 year.

DIAGNOSIS

When confronted with possible CAP, the physician must ask two questions: Is this pneumonia, and, if so, what is the likely etiology? The former question is typically answered by clinical and radiographic methods, whereas the latter requires the aid of laboratory techniques.

Clinical Diagnosis

The differential diagnosis includes both infectious and noninfectious entities such as acute bronchitis, acute exacerbations of chronic bronchitis, heart failure, pulmonary embolism, hypersensitivity pneumonitis, and radiation pneumonitis. The importance of a careful history cannot be overemphasized. For example, known cardiac disease may suggest worsening pulmonary edema, while underlying carcinoma may suggest lung injury secondary to irradiation.

Unfortunately, the sensitivity and specificity of the findings on physical examination are less than ideal, averaging 58% and 67%, respectively. As mentioned earlier, the elderly may initially present with confusion alone. Therefore, chest radiography is often necessary to differentiate CAP from other conditions. Radiographic findings may include risk factors for increased severity (e.g., cavitation or multilobar involvement). Occasionally, radiographic results suggest an etiologic diagnosis. For example, pneumatoceles suggest infection with *S. aureus*, and an upper-lobe cavitating lesion suggests tuberculosis. CT may be of value in a patient with suspected postobstructive pneumonia caused by a tumor or foreign body or suspected cavitary disease. For outpatients, the clinical and radiologic assessments are usually all that is done before treatment for CAP is started since most laboratory results are not available soon enough to

influence initial management significantly. In certain cases, the availability of rapid point-of-care outpatient diagnostic tests can be very important; for example, rapid diagnosis of influenza virus infection can prompt specific anti-influenza drug treatment and secondary prevention.

Etiologic Diagnosis

The etiology of pneumonia usually cannot be determined solely on the basis of clinical presentation. Except for CAP patients admitted to the ICU, no data exist to show that treatment directed at a specific pathogen is statistically superior to empirical therapy. The benefit of establishing a microbial etiology can therefore be questioned, particularly in light of the cost of diagnostic testing. However, a number of reasons can be advanced for attempting an etiologic diagnosis. Identification of an unexpected pathogen allows narrowing of the initial empirical regimen, thereby decreasing antibiotic selection pressure and lessening the risk of resistance. Pathogens with important public safety implications, such as *Mycobacterium tuberculosis* and influenza virus, may be found in some cases. Finally, without culture and susceptibility data, trends in resistance cannot be followed accurately, and appropriate empirical therapeutic regimens are harder to devise.

Gram's Stain and Culture of Sputum

The main purpose of the sputum Gram's stain is to ensure that a sample is suitable for culture. However, Gram's staining may also identify certain pathogens (e.g., *S. pneumoniae*, *S. aureus*, and gram-negative bacteria) by their characteristic appearance. To be adequate for culture, a sputum sample must have >25 neutrophils and <10 squamous epithelial cells per low-power field. The sensitivity and specificity of the sputum Gram's stain and culture are highly variable. Even in cases of proven bacteremic pneumococcal pneumonia, the yield of positive cultures from sputum samples is ≤50%.

Many patients, particularly elderly individuals, may not be able to produce an appropriate expectorated sputum sample. Others may already have started a course of antibiotics that can interfere with culture results at the time a sample is obtained. Inability to produce sputum can result from dehydration, and its correction may result in increased sputum production and a more obvious infiltrate on chest radiography. For patients admitted to the ICU and intubated, a deep-suction aspirate or bronchoalveolar lavage sample (obtained either via bronchoscopy or non-bronchoscopically) has a high yield on culture when sent to the microbiology laboratory as soon as possible. Since the etiologies in severe CAP are somewhat different from those in milder disease (Table 121-2), the greatest benefit of staining and culturing respiratory secretions is to alert the physician of unsuspected and/or resistant pathogens and to permit appropriate modification of therapy. Other stains and cultures (e.g., specific stains for *M. tuberculosis* or fungi) may be useful as well.

Blood Cultures

The yield from blood cultures, even when samples are collected before antibiotic therapy, is disappointingly low. Only 5–14% of cultures of blood from patients hospitalized with CAP are positive, and the most frequently isolated pathogen is *S. pneumoniae.* Since recommended empirical regimens all provide pneumococcal coverage, a blood culture positive for this pathogen has little, if any, effect on clinical outcome. However, susceptibility data may allow narrowing of antibiotic therapy in appropriate cases. Because of the low yield and the lack of significant impact on outcome, blood cultures are no longer

considered *de rigueur* for all hospitalized CAP patients. Certain high-risk patients—including those with neutropenia secondary to pneumonia, asplenia, complement deficiencies, chronic liver disease, or severe CAP—should have blood cultured.

Urinary Antigen Tests

Two commercially available tests detect pneumococcal and *Legionella* antigen in urine. The test for *Legionella pneumophila* detects only serogroup 1, but this serogroup accounts for most community-acquired cases of Legionnaires' disease in the United States. The sensitivity and specificity of the *Legionella* urine antigen test are as high as 70% and 99%, respectively. The pneumococcal urine antigen test is also quite sensitive and specific (70% and >90%, respectively). Although false-positive results can be obtained with samples from pneumococcus-colonized children, the test is generally reliable. Both tests can detect antigen even after the initiation of appropriate antibiotic therapy.

Polymerase Chain Reaction

PCR tests, which amplify a microorganism's DNA or RNA, are available for a number of pathogens. PCR of nasopharyngeal swabs, for example, has become the standard for diagnosis of respiratory viral infection. In addition, PCR can detect the nucleic acid of *Legionella* species, *M. pneumoniae*, *C. pneumoniae*, and mycobacteria. The cost-effectiveness of PCR testing, however, has not been definitively established. In patients with pneumococcal pneumonia, an increased bacterial load documented in whole blood by PCR is associated with an increased risk of septic shock, the need for mechanical ventilation, and death. Clinical availability of such a test could conceivably help identify patients suitable for ICU admission.

Serology

A fourfold rise in specific IgM antibody titer between acute- and convalescent-phase serum samples is generally considered diagnostic of infection with the pathogen in question. In the past, serologic tests were used to help identify atypical pathogens as well as selected unusual organisms such as *Coxiella burnetii.* Recently, however, they have fallen out of favor because of the time required to obtain a final result for the convalescent-phase sample and the difficulty of interpretation.

Biomarkers

A number of substances can serve as markers of severe inflammation. The two most commonly in use are C-reactive protein (CRP) and procalcitonin (PCT). Levels of these acute-phase reactants increase in the presence of an inflammatory response, particularly to bacterial pathogens. CRP may be of use in the identification of worsening disease or treatment failure, and PCT may play a role in distinguishing bacterial from viral infection, determining the need for antibacterial therapy, or deciding when to discontinue treatment. PCT testing can result in less antibiotic use in CAP with no concomitant increase in treatment failure or mortality risk. These tests should not be used on their own, but, when interpreted in conjunction with other findings from the history, physical examination, radiology, and laboratory tests, may help with antibiotic stewardship and appropriate management of seriously ill patients with CAP.

TREATMENT

TREATMENT

SITE OF CARE

The cost of inpatient management exceeds that of outpatient treatment by a factor of 20, and hospitalization accounts for most CAP-related expenditures. Thus the decision to hospitalize a patient with CAP has considerable implications, and late admission to the ICU is associated with increased mortality risk. Certain patients can be managed at home, and others clearly require treatment in the hospital, but the choice is sometimes difficult. Tools that objectively assess the risk of adverse outcomes, including severe illness and death, can minimize unnecessary hospital admissions. Although a number of prediction rules exist, the two most frequently used are the Pneumonia Severity Index (PSI), a prognostic model used to identify patients at low risk of dying, and the CURB-65 criteria, a severity-of-illness score.

To determine the PSI, points are given for 20 variables, including age, coexisting illness, and abnormal physical and laboratory findings. On the basis of the resulting score, patients are assigned to one of five classes with the following mortality rates: class 1, 0.1%; class 2, 0.6%; class 3, 2.8%; class 4, 8.2%; and class 5, 29.2%. Determination of the PSI is often impractical in a busy emergency-department setting because of the number of variables. However, clinical trials demonstrate that routine use of the PSI results in lower admission rates for class 1 and class 2 patients. Patients in class 3 could ideally be admitted to an observation unit until a further decision can be made.

The CURB-65 criteria include five variables: confusion (C); urea >7 mmol/L (U); respiratory rate ≥30/min (R); blood pressure, systolic ≤90 mmHg or diastolic ≤60 mmHg (B); and age ≥65 years. Patients with a score of 0, among whom the 30-day mortality rate is 1.5%, can be treated outside the hospital. With a score of 1 or 2, the patient should be hospitalized unless the score is entirely or in part attributable to an age of ≥65 years. In such cases, hospitalization may not be necessary. Among patients with scores of ≥3, mortality rates are 22% overall; these patients may require ICU admission.

It is not clear which assessment tool is superior. Whichever system is used, these objective criteria must always be tempered by careful consideration of factors relevant to individual patients, including the ability to comply reliably with an oral antibiotic regimen and the resources available to the patient outside the hospital.

Neither PSI nor CURB-65 is accurate in determining the need for ICU admission. Septic shock or respiratory failure in the emergency department is an obvious indication for ICU care. However, mortality rates are higher among less ill patients who are admitted to the floor and then deteriorate than among equally ill patients monitored in the ICU. A variety of scores have been proposed to identify patients most likely to have early deterioration (**Table 121-4**). Most factors in these scores are similar to the minor severity criteria proposed by the Infectious Diseases Society of America (IDSA) and the American Thoracic Society (ATS) in their guidelines for the management of CAP. Recent data suggest that thrombocytopenia, leukopenia, and hypothermia can be removed from the list of minor criteria.

ANTIBIOTIC RESISTANCE

Antimicrobial resistance is a significant problem that threatens to diminish our therapeutic armamentarium. Misuse of antibiotics results in increased antibiotic selection pressure that can affect resistance locally and globally by clonal dissemination. For CAP, the main resistance issues currently involve *S. pneumoniae* and CA-MRSA.

S. pneumoniae

In general, pneumococcal resistance is acquired by direct DNA incorporation and remodeling resulting from contact with closely related oral commensal bacteria, by the process of natural transformation, or by mutation of certain genes.

The minimal inhibitory concentration (MIC) cutoffs for penicillin in pneumonia are ≤2 µg/mL for susceptible, >2–4 µg/mL for intermediate, and ≥8 µg/mL for resistant. A change in susceptibility thresholds resulted in a dramatic decrease in the proportion of pneumococcal isolates considered nonsusceptible. For meningitis, MIC thresholds remain at the former higher levels. Fortunately, resistance to penicillin appeared to plateau even before the change in MIC thresholds. Pneumococcal resistance to β-lactam drugs is due solely to low-affinity penicillin-binding proteins. Risk factors for penicillin-resistant pneumococcal infection include recent antimicrobial therapy, an age of <2 years or >65 years, attendance at day-care centers, recent hospitalization, and HIV infection.

 In contrast to penicillin resistance, resistance to macrolides is increasing through several mechanisms. *Target-site modification* caused by ribosomal methylation in 23S rRNA encoded by the *ermB* gene results in high-level resistance (MICs, ≥64 µg/mL) to macrolides, lincosamides, and streptogramin B–type antibiotics. The *efflux mechanism* encoded by the *mef* gene (*M phenotype*) is usually associated with low-level resistance (MICs, 1–32 µg/mL). These two mechanisms account for ~45% and ~65%, respectively, of resistant pneumococcal isolates in the United States. High-level resistance to macrolides is more common in Europe, whereas lower-level resistance predominates in North America. In some countries, including the United States, the prevalence of macrolide-resistant *S. pneumoniae* exceeds 25%. In such situations, a macrolide should not be used as empirical monotherapy.

Pneumococcal resistance to fluoroquinolones (e.g., ciprofloxacin and levofloxacin) has been reported. Changes can occur in one or both target sites (topoisomerases II and IV) from mutations in the *gyrA* and *parC* genes, respectively. In addition, an efflux pump may play a role in pneumococcal resistance to fluoroquinolones.

Isolates resistant to drugs from three or more antimicrobial classes with different mechanisms of action are considered MDR strains. The propensity for an association of pneumococcal resistance to penicillin with reduced susceptibility to other drugs, such as macrolides, tetracyclines, and trimethoprim-sulfamethoxazole, is also of concern. In the United States, 58.9% of penicillin-resistant pneumococcal isolates from blood are also resistant to macrolides.

The most important risk factor for antibiotic-resistant pneumococcal infection is use of a specific antibiotic within the previous 3 months. Therefore, a patient's history of prior antibiotic treatment is a critical factor in avoiding the use of an inappropriate antibiotic.

M. pneumoniae

Macrolide-resistant *M. pneumoniae* has been reported in a number of countries, including Germany (3%), Japan (30%), China (95%), and France and the United States (5–13%). *Mycoplasma* resistance to macrolides is on the rise as a result of binding-site mutation in domain V of 23S rRNA.

CA-MRSA

CAP due to MRSA may be caused by the classic hospital-acquired strains or by genotypically and phenotypically distinct community-acquired strains. Most infections with the former strains have been acquired either directly or indirectly by contact with the health care environment (Table 121-1). However, in some hospitals, CA-MRSA strains are displacing the classic hospital-acquired strains—a trend suggesting that the newer strains may be more robust and blurring this distinction.



Methicillin resistance in *S. aureus* is determined by the *mecA* gene, which encodes for resistance to all β-lactam drugs. At least five *staphylococcal chromosomal cassette mec* (*SCCmec*) types have been described. The typical hospital-acquired strain usually has type II or III, whereas CA-MRSA has a type IV SCC*mec* element. CA-MRSA isolates tend to be less resistant than the older hospital-acquired strains and are often susceptible to trimethoprim-sulfamethoxazole, clindamycin, and tetracycline in addition to vancomycin and linezolid. However, the most important distinction is that CA-MRSA strains also carry genes for superantigens, such as enterotoxins B and C and Panton-Valentine leukocidin, a membrane-tropic toxin that can create cytolytic pores in polymorphonuclear neutrophils, monocytes, and macrophages.

Gram-Negative Bacilli

A detailed discussion of resistance among gram-negative bacilli is beyond the scope of this chapter **(see Chap. 156)**. Fluoroquinolone resistance among isolates of *Escherichia coli* from the community appears to be increasing. *Enterobacter* species are typically resistant to cephalosporins; the drugs of choice for use against these bacteria are usually fluoroquinolones or carbapenems. Similarly, when infections due to bacteria producing extended-spectrum β-lactamases are documented or suspected, a fluoroquinolone or a carbapenem should be used.

INITIAL ANTIBIOTIC MANAGEMENT

Since the etiology of CAP is rarely known at the outset of treatment, initial therapy is usually empirical, designed to cover the most likely pathogens (Table 121-2). In all cases, antibiotic treatment should be initiated as expeditiously as possible. The CAP treatment guidelines in the United States (summarized in Table 121-5) represent joint statements from the IDSA and the ATS; the Canadian guidelines come from the Canadian Infectious Disease Society and the Canadian Thoracic Society. In these guidelines, coverage is always provided for the pneumococcus and atypical pathogens. In contrast, guidelines from some European countries do not always include atypical coverage based on local epidemiologic data. The U.S./Canadian

approach is supported by retrospective data derived from administrative databases including thousands of patients. Atypical pathogen coverage provided by the addition of a macrolide to a β-lactam or by the use of a fluoroquinolone alone has been consistently associated with a significant reduction in mortality rates compared with those for β-lactam coverage alone.

For the treatment of severe CAP, accumulating data continue to demonstrate the benefits of including a macrolide, such as reduced mortality. However, two recent studies of patients hospitalized with moderate CAP yielded differing results. One study demonstrated a more rapid return to clinical stability and fewer adverse events with a β-lactam–macrolide combination than with a β-lactam alone. Using cluster randomization, the second study reported no difference among three regimens—a β-lactam alone, a β-lactam–macrolide combination, and a fluoroquinolone—but had significant design flaws, including a lack of chest radiographic confirmation in 24% of cases and significant rates of noncompliance with the assigned regimen.

Empirical treatment regimens for CAP are listed in Table 121-5. In general, the recommendations in the IDSA/ATS guidelines published in 2007 continue to apply but with a possible exception for treatment of outpatients who have previously been well and have not received an antibiotic within 3 months. Given the rise of macrolide resistance among pneumococci, consideration of local epidemiologic and susceptibility data as well as the patient's recent use of any antibiotics is imperative before selection of a regimen, particularly as regards macrolide monotherapy. If concern exists about macrolide resistance, the patient is otherwise well and has not recently received antibiotics, and the local doxycycline resistance rate among pneumococcal isolates is <25%, doxycycline may be used instead of macrolide monotherapy. Otherwise, a fluoroquinolone or a β-lactam plus a macrolide should be used.

A meta-analysis found ceftaroline to be superior to ceftriaxone as the β-lactam component of IV empirical treatment of CAP in hospitalized patients in PORT risk class III or IV who have not received prior antibiotics. Clinical response rates for patients infected with *S. pneumoniae* or *S. aureus* also favored ceftaroline. Patients who had documented or suspected infection due to *P. aeruginosa* were excluded.

Once the etiologic agent(s) and their susceptibilities are known, therapy may be altered to target the specific pathogen(s). However, this decision is not always straightforward. If blood cultures yield *S. pneumoniae* sensitive to penicillin after 2 days of treatment with a macrolide plus a β-lactam or with a fluoroquinolone alone, should therapy be switched to penicillin alone? The concern here is that a β-lactam alone would not be effective in the potential 15% of cases with atypical co-infection. No standard approach exists. Some experts think that 3 days of macrolide therapy is adequate for *Mycoplasma* infection and that, unless the test for *Legionella* urinary antigen is positive, treatment can be continued with a β-lactam alone. In all cases, the individual patient and the various risk factors must be considered.

Management of bacteremic pneumococcal pneumonia is also controversial. Data from nonrandomized studies suggest that combination therapy (especially with a β-lactam–macrolide combination) is associated with a lower mortality rate than monotherapy, particularly in severely ill patients. The exact reason is

unknown, but possible explanations include an additive or synergistic antibacterial effect, antimicrobial tolerance, atypical co-infection, or the immunomodulatory effects of the macrolides.

For CAP patients admitted to the ICU, the risk of infection with *P. aeruginosa* or CA-MRSA is increased. Empirical coverage should be considered when a patient has risk factors or a Gram's stain suggestive of these pathogens (Table 121-5). If CA-MRSA is suspected, either linezolid or vancomycin—with or without clindamycin to inhibit toxin production—can be added to the initial empirical regimen. There is increasing concern about vancomycin's loss of potency against MRSA, poor penetration into epithelial lining fluid, and lack of effect on toxin production relative to linezolid.

Although hospitalized patients have traditionally received initial therapy by the IV route, some drugs—particularly the fluoroquinolones—are very well absorbed and can be given orally from the outset to certain patients. For patients initially treated IV, a switch to oral treatment is appropriate as long as the patient can ingest and absorb the drugs, is hemodynamically stable, and is showing clinical improvement.

The duration of treatment for CAP has generated considerable interest. Studies with fluoroquinolones and telithromycin suggest that a 5-day course is sufficient for otherwise uncomplicated CAP but a longer course may be required for patients with bacteremia, metastatic infection, or infection with a virulent pathogen such as *P. aeruginosa* or CA-MRSA.

ADJUNCTIVE MEASURES

In addition to appropriate antimicrobial therapy, certain adjunctive measures should be used. Adequate hydration, oxygen therapy for hypoxemia, vasopressors, and assisted ventilation when necessary are critical to successful treatment. Randomized placebo-controlled trials have shown a benefit in treatment of hospitalized patients and patients who have severe CAP with prednisone and methylprednisolone, respectively. The value of adjunctive therapy with agents such as statins and angiotensin-converting enzyme inhibitors remains unproven in the management of CAP.

FAILURE TO IMPROVE

Patients slow to respond to therapy should be reevaluated at about day 3 (sooner if their condition is worsening rather than simply not improving), and several possible scenarios should be considered. A number of noninfectious conditions mimic pneumonia, including pulmonary edema, pulmonary embolism, lung carcinoma, radiation and hypersensitivity pneumonitis, and connective tissue disease involving the lungs. If the patient truly has CAP and empirical treatment is aimed at the correct pathogen, lack of response may be explained in a number of ways. The pathogen may be resistant to the drug selected, or a sequestered focus (e.g., lung abscess or empyema) may be blocking access of the antibiotic(s) to the pathogen. The patient may be getting either the wrong drug or the correct drug at the wrong dose or frequency of administration. Another possibility is that CAP is the correct diagnosis but an unsuspected pathogen (e.g., CA-MRSA, *M. tuberculosis*, or a fungus) is the cause. Nosocomial superinfections—both pulmonary and extrapulmonary—are other possible explanations for a hospitalized patient's failure to improve or deterioration. In all cases of delayed response or worsening condition, the patient must be carefully reassessed and appropriate studies initiated, possibly including procedures such as CT or bronchoscopy.

COMPLICATIONS

Complications of severe CAP include respiratory failure, shock and multiorgan failure, coagulopathy, and exacerbation of comorbid illnesses. Three particularly noteworthy conditions are metastatic infection, lung abscess, and complicated pleural effusion. Metastatic infection (e.g., brain abscess or endocarditis) is very unusual and will require a high degree of suspicion and a detailed workup for proper treatment. Lung abscess may occur in association with aspiration or with infection caused by a single CAP pathogen, such as CA-MRSA, *P. aeruginosa*, or (rarely) *S. pneumoniae.* Aspiration pneumonia is typically a polymicrobial infection involving both aerobes and anaerobes. A significant pleural effusion should be tapped for both diagnostic and therapeutic purposes. If the fluid has a pH of <7, a glucose level of <2.2 mmol/L, and a lactate dehydrogenase concentration of >1000 U/L or if bacteria are seen or cultured, it should be completely drained; a chest tube is often required, and video-assisted thoracoscopy may be needed for late treatment or difficult cases.

FOLLOW-UP

Fever and leukocytosis usually resolve within 2–4 days in otherwise healthy patients with CAP, but physical findings may persist longer. Chest radiographic abnormalities are slowest to resolve (4–12 weeks), with the speed of clearance depending on the patient's age and underlying lung disease. Patients may be discharged from the hospital once their clinical conditions, including comorbidities, are stable. The site of residence after discharge (nursing home, home with family, home alone) is an important discharge consideration, particularly for elderly patients. For a hospitalized patient, a follow-up radiograph ~4–6 weeks later is recommended. If relapse or recurrence is documented, particularly in the same lung segment, the possibility of an underlying neoplasm must be considered.

TABLE 121-4

Risk Factors for Early Deterioration in Community-Acquired Pneumonia

Multilobar infiltrates	Hypoalbuminemia
Severe hypoxemia (arterial saturation <90%)	Neutropenia
Severe acidosis (pH <7.30)	Thrombocytopenia
Mental confusion	Hyponatremia
Severe tachypnea (>30 breaths/min)	Hypoglycemia

TABLE 121-5

Empirical Antibiotic Treatment of Community-Acquired Pneumonia

Outpatients

1. Previously healthy and no antibiotics in past 3 months
   - A macrolide [clarithromycin (500 mg PO bid) or azithromycin (500 mg PO once, then 250 mg qd)] *or*
   - Doxycycline (100 mg PO bid)
2. Comorbidities or antibiotics in past 3 months: select an alternative from a different class
   - A respiratory fluoroquinolone [moxifloxacin (400 mg PO qd), gemifloxacin (320 mg PO qd), levofloxacin (750 mg PO qd)] *or*
   - A β-lactam [preferred: high-dose amoxicillin (1 g tid) or amoxicillin/clavulanate (2 g bid); alternatives: ceftriaxone (1–2 g IV qd), cefpodoxime (200 mg PO bid), or cefuroxime (500 mg PO bid)] *plus* a macrolide[a]
3. In regions with a high rate of "high-level" pneumococcal macrolide resistance,[b] consider alternatives listed above for patients with comorbidities.

Inpatients, Non-ICU

- A respiratory fluoroquinolone [e.g., moxifloxacin (400 mg PO or IV qd) or levofloxacin (750 mg PO or IV qd)]
- A β-lactam[c] [e.g., ceftriaxone (1–2 g IV qd), ampicillin (1–2 g IV q4–6h), cefotaxime (1–2 g IV q8h), ertapenem (1 g IV qd)] *plus* a macrolide[d] [e.g., oral clarithromycin or azithromycin as listed above or IV azithromycin (1 g once, then 500 mg qd)]

Inpatients, ICU

- A β-lactam[e] [e.g., ceftriaxone (2 g IV qd), ampicillin-sulbactam (2 g IV q8h), or cefotaxime (1–2 g IV q8h)] *plus* either azithromycin or a fluoroquinolone (as listed above for inpatients, non-ICU)

Special Concerns

If Pseudomonas is a consideration:

- An antipseudomonal β-lactam [e.g., piperacillin/tazobactam (4.5 g IV q6h), cefepime (1–2 g IV q12h), imipenem (500 mg IV q6h), meropenem (1 g IV q8h)] *plus* either ciprofloxacin (400 mg IV q12h) or levofloxacin (750 mg IV qd)
- The above β-lactams *plus* an aminoglycoside [amikacin (15 mg/kg qd) or tobramycin (1.7 mg/kg qd)] *plus* azithromycin
- The above β-lactams[f] *plus* an aminoglycoside *plus* an antipneumococcal fluoroquinolone

If CA-MRSA is a consideration:

- Add linezolid (600 mg IV q12h) or vancomycin (15 mg/kg q12h initially, with adjusted doses) *plus* clindamycin (300 mg q6h)

[a]Doxycycline (100 mg PO bid) is an alternative to the macrolide. [b]MICs >16 µg/mL in 25% of isolates. [c]A respiratory fluoroquinolone should be used for penicillin-allergic patients. [d]Doxycycline (100 mg IV q12h) is an alternative to the macrolide. [e]For penicillin-allergic patients, use a respiratory fluoroquinolone and aztreonam (2 g IV q8h). [f]For penicillin-allergic patients, substitute aztreonam.

Abbreviations: CA-MRSA, community-acquired methicillin-resistant *Staphylococcus aureus;* ICU, intensive care unit.

PROGNOSIS

The prognosis of CAP depends on the patient's age, comorbidities, and site of treatment (inpatient or outpatient). Young patients without comorbidity do well and usually recover fully after ~2 weeks. Older patients and those with comorbid conditions can take several weeks longer to recover fully. The overall mortality rate for the outpatient group is <5%. For patients requiring hospitalization, the overall mortality rate ranges from 2 to 40%, depending on the category of patient and the processes of care, particularly the administration of appropriate antibiotics as soon as possible.

PREVENTION

The main preventive measure is vaccination **(Chap. 118)**. Recommendations of the Advisory Committee on Immunization Practices should be followed for influenza and pneumococcal vaccines.

A pneumococcal polysaccharide vaccine (PPSV23) and a protein conjugate pneumococcal vaccine (PCV13) are available in the United States **(Chap. 141)**. The former product contains capsular material from 23 pneumococcal serotypes; in the latter, capsular polysaccharide from 13 of the most common pneumococcal pathogens affecting children is linked to an immunogenic protein. PCV13 produces T cell–dependent antigens that result in long-term immunologic memory. Administration of this vaccine to children has led to an overall decrease in the prevalence of antimicrobial-resistant pneumococci and in the incidence of invasive pneumococcal disease among both children and adults. However, vaccination can be followed by the replacement of vaccine serotypes with nonvaccine serotypes, as was seen with serotypes 19A and 35B after introduction of the original 7-valent conjugate vaccine. PCV13 is also recommended for the elderly and for younger immunocompromised patients. Because of an increased risk of pneumococcal infection, even among patients without obstructive lung disease, smokers should be strongly encouraged to stop smoking.

The influenza vaccine is available in an inactivated or recombinant form. The live attenuated influenza vaccine or "nasal spray" vaccine is no longer recommended. In the event of an influenza outbreak, unprotected patients at risk from complications should be vaccinated immediately and given

chemoprophylaxis with either oseltamivir or zanamivir for 2 weeks—i.e., until vaccine-induced antibody levels are sufficiently high.

VENTILATOR-ASSOCIATED PNEUMONIA

Most research on hospital-acquired pneumonia has focused on VAP. However, the information and principles based on this research can be applied to non-ICU HAP as well. The greatest difference between VAP and HAP studies is the dependence on expectorated sputum for a microbiologic diagnosis of HAP (as for that of CAP), which is further complicated by frequent colonization by pathogens in patients with HAP. Therefore, most of the literature has focused on HAP resulting in intubation, where, once again, access to the lower respiratory tract facilitates an etiologic diagnosis.

ETIOLOGY

Potential etiologic agents of VAP include both MDR and non-MDR bacterial pathogens (**Table 121-6**). The non-MDR group is nearly identical to the pathogens found in severe CAP (Table 121-2); it is not surprising that such pathogens predominate if VAP develops in the first 5–7 days of the hospital stay. However, if patients have other risk factors, MDR pathogens are a consideration, even early in the hospital course. The relative frequency of individual MDR pathogens can vary significantly from hospital to hospital and even between different critical care units within the same institution. Most hospitals have problems with *P. aeruginosa* and MRSA, but other MDR pathogens are often institution-specific. Less commonly, fungal and viral pathogens cause VAP, usually affecting severely immunocompromised patients. Rarely, community-associated viruses cause mini-epidemics, usually when introduced by ill health care workers.

TABLE 121-6

Microbiologic Causes of Ventilator-Associated Pneumonia

NON-MDR PATHOGENS	MDR PATHOGENS
Streptococcus pneumoniae	*Pseudomonas aeruginosa*
Other *Streptococcus* spp.	Methicillin-resistant *S. aureus*
Haemophilus influenzae	*Acinetobacter* spp.
Methicillin-sensitive *Staphylococcus aureus*	Antibiotic-resistant Enterobacteriaceae
Antibiotic-sensitive Enterobacteriaceae	ESBL-positive strains
Escherichia coli	Carbapenem-resistant strains
Klebsiella pneumoniae	*Legionella pneumophila*
Proteus spp.	*Burkholderia cepacia*
Enterobacter spp.	*Aspergillus* spp.
Serratia marcescens	

Abbreviations: ESBL, extended-spectrum β-lactamase; MDR, multidrug-resistant.

EPIDEMIOLOGY

Pneumonia is a common complication among patients requiring mechanical ventilation. Prevalence estimates vary between 6 and 52 cases per 100 patients, depending on the population studied. On any given day in the ICU, an average of 10% of patients will have pneumonia—VAP in the overwhelming majority of cases. The frequency of diagnosis is not static but changes with the duration of mechanical ventilation, with the highest hazard ratio in the first 5 days and a plateau in additional cases (1% per day) after ~2 weeks. However, the cumulative rate among patients who remain ventilated for as long as 30 days is as high as 70%. These rates often do not reflect the recurrence of VAP in the same patient. Once a ventilated patient is transferred to a chronic-care facility or to home, the incidence of pneumonia drops significantly, especially in the absence of other risk factors for pneumonia. However, in chronic ventilator units, purulent tracheobronchitis becomes a significant issue, often interfering with efforts to wean patients off mechanical ventilation (Chap. 295).

Three factors are critical in the pathogenesis of VAP: colonization of the oropharynx with pathogenic microorganisms, aspiration of these organisms from the oropharynx into the lower respiratory tract, and compromise of the normal host defense mechanisms. Most risk factors and their corresponding prevention strategies pertain to one of these three factors (Table 121-7).

TABLE 121-7

Pathogenic Mechanisms and Corresponding Prevention Strategies for Ventilator-Associated Pneumonia

PATHOGENIC MECHANISM	PREVENTION STRATEGY
Oropharyngeal colonization with pathogenic bacteria	
Elimination of normal flora	Avoidance of prolonged antibiotic courses
Large-volume oropharyngeal aspiration around time of intubation	Short course of prophylactic antibiotics for comatose patients[a]
Gastroesophageal reflux	Postpyloric enteral feeding[b]; avoidance of high gastric residuals, prokinetic agents
Bacterial overgrowth of stomach	Avoidance of prophylactic agents that raise gastric pH[b]; selective decontamination of digestive tract with nonabsorbable antibiotics[b]
Cross-infection from other colonized patients	Hand washing, especially with alcohol-based hand rub; intensive infection control education[a]; isolation; proper cleaning of reusable equipment
Large-volume aspiration	Endotracheal intubation; rapid-sequence intubation technique; avoidance of sedation; decompression of small-bowel obstruction
Microaspiration around endotracheal tube	
Endotracheal intubation	Noninvasive ventilation[a]
Prolonged duration of ventilation	Daily awakening from sedation,[a] weaning protocols[a]
Abnormal swallowing function	Early percutaneous tracheostomy[a]

PATHOGENIC MECHANISM	PREVENTION STRATEGY
Secretions pooled above endotracheal tube	Head of bed elevated[a]; continuous aspiration of subglottic secretions with specialized endotracheal tube[a]; avoidance of reintubation; minimization of sedation and patient transport
Altered lower respiratory host defenses	Tight glycemic control[b]; lowering of hemoglobin transfusion threshold

[a]Strategies demonstrated to be effective in at least one randomized controlled trial. [b]Strategies with negative randomized trials or conflicting results.

The most obvious risk factor is the endotracheal tube, which bypasses the normal mechanical factors preventing aspiration. While the presence of an endotracheal tube may prevent large-volume aspiration, microaspiration is actually exacerbated by secretions pooling above the cuff. The endotracheal tube and the concomitant need for suctioning can damage the tracheal mucosa, thereby facilitating tracheal colonization. In addition, pathogenic bacteria can form a glycocalyx biofilm on the tube's surface that protects them from both antibiotics and host defenses. The bacteria can also be dislodged during suctioning and can reinoculate the trachea, or tiny fragments of glycocalyx can embolize to distal airways, carrying bacteria with them.

In a high percentage of critically ill patients, the normal oropharyngeal flora is replaced by pathogenic microorganisms. The most important risk factors are antibiotic selection pressure, cross-infection from other infected/colonized patients or contaminated equipment, and malnutrition. Of these factors, antibiotic exposure poses the greatest risk by far. Pathogens such as *P. aeruginosa* almost never cause infection in patients without prior exposure to antibiotics. The recent emphasis on hand hygiene has lowered the cross-infection rate.

How the lower respiratory tract defenses become overwhelmed remains poorly understood. Almost all intubated patients experience microaspiration and are at least transiently colonized with pathogenic bacteria. However, only around one-third of colonized patients develop VAP. Colony counts increase to high levels, sometimes days before the development of clinical pneumonia; these increases suggest that the final step in VAP development, independent of aspiration and oropharyngeal colonization, is the overwhelming of host defenses. Severely ill patients with sepsis and trauma appear to enter a state of immunoparalysis several days after admission to the ICU—a time that corresponds to the greatest risk of developing VAP. The mechanism of this immunosuppression is not clear, although several factors have been suggested. Hyperglycemia and more frequent transfusions adversely affect the immune response.

CLINICAL MANIFESTATIONS

The clinical manifestations are generally the same in VAP as in all other forms of pneumonia: fever, leukocytosis, increase in respiratory secretions, and pulmonary consolidation on physical examination, along with a new or changing radiographic infiltrate. The frequency of abnormal chest radiographs before the onset of pneumonia in intubated patients and the limitations of portable radiographic technique make interpretation of radiographs more difficult than in patients who are not intubated. Other clinical features may include tachypnea, tachycardia, worsening oxygenation, and increased minute ventilation.

DIAGNOSIS

No single set of criteria is reliably diagnostic of pneumonia in a ventilated patient. The inability to accurately identify such patients compromises efforts to prevent and treat VAP and even calls into question estimates of the impact of VAP on mortality rates.

Application of the clinical criteria typical for CAP consistently results in overdiagnosis of VAP, largely because of three common findings in at-risk patients: (1) frequent tracheal colonization with pathogenic bacteria in patients with endotracheal tubes, (2) multiple alternative causes of radiographic infiltrates in mechanically ventilated patients, and (3) the high frequency of other sources of fever in critically ill patients. The differential diagnosis of VAP includes a number of entities such as atypical pulmonary edema, pulmonary contusion, alveolar hemorrhage, hypersensitivity pneumonitis, acute respiratory distress syndrome, and pulmonary embolism. Clinical findings in ventilated patients with fever and/or leukocytosis may have alternative causes, including antibiotic-associated diarrhea, central line–associated infection, sinusitis, urinary tract infection, pancreatitis, and drug fever. Conditions mimicking pneumonia are often documented in patients in whom VAP has been ruled out by accurate diagnostic techniques. Most of these alternative diagnoses do not require antibiotic treatment; require antibiotics different from those used to treat VAP; or require some additional intervention, such as surgical drainage or catheter removal, for optimal management.

This diagnostic dilemma has led to debate and controversy. The major question is whether a quantitative-culture approach as a means of eliminating false-positive clinical diagnoses is superior to the clinical approach enhanced by principles learned from quantitative-culture studies. The most recent IDSA/ATS guidelines for HAP/VAP gave a weak recommendation for the clinical approach based on availability of resources, cost, and availability of expertise. The guidelines did acknowledge that the use of a quantitative approach may result in less antibiotic use, which may be critical for antibiotic stewardship in the ICU. Therefore, the approach at each institution, or potentially for each patient, should balance the frequency of complex illnesses that are associated with (1) greater frequency of alternative causes of the clinical manifestations, (2) higher colonization rates, and (3) more frequent prior antibiotic therapy versus availability and expertise of invasive techniques with quantitative cultures.

Quantitative-Culture Approach

The essence of the quantitative-culture approach is discrimination between colonization and true infection through determination of the bacterial burden. The more distal in the respiratory tree the diagnostic

sampling, the more specific the results and therefore the lower the threshold of growth necessary to diagnose pneumonia and exclude colonization. For example, a quantitative endotracheal aspirate yields proximate samples, and the diagnostic threshold is 10^6 cfu/mL. The protected specimen brush method, in contrast, obtains distal samples and has a threshold of 10^3 cfu/mL. Conversely, sensitivity declines as more distal secretions are obtained, especially when they are collected blindly (i.e., by a technique other than bronchoscopy). Additional tests that may increase the diagnostic yield include Gram's staining, differential cell counts, staining for intracellular organisms, and detection of local protein levels elevated in response to infection.

The key piece of a quantitative-culture approach is to base subsequent antibiotic therapy on the results of the quantitative cultures. In a study comparing the quantitative with the clinical approach, the use of bronchoscopic quantitative cultures resulted in significantly less antibiotic use at 14 days after study entry, a lower 14-day mortality rate, and a lower 28-day severity-adjusted mortality rate. In addition, more alternative sites of infection were found in patients randomized to the quantitative-culture strategy. A critical aspect of this study was that antibiotic treatment was initiated only in patients whose gram-stained respiratory sample was positive or who displayed signs of hemodynamic instability. Fewer than half as many patients were treated for pneumonia in the bronchoscopy group, and only one-third as many microorganisms were cultured. Other randomized trials of the quantitative-culture approach did not closely link antibiotic management with the results of cultures; thus the validity of their results was compromised.

The Achilles heel of the quantitative approach is the effect of antibiotic therapy. With sensitive microorganisms, a single antibiotic dose can reduce colony counts below the diagnostic threshold. Recent changes in antibiotic therapy are the most significant. After 3 days, the operating characteristics of the tests improve to the point at which they are equivalent to results when no prior antibiotic therapy has been given. Conversely, colony counts above the diagnostic threshold during antibiotic therapy suggest that the current antibiotics are ineffective. Even the normal host response may be sufficient to reduce quantitative-culture counts below the diagnostic threshold if sampling is delayed. In short, expertise in quantitative-culture techniques is critical, with a specimen obtained as soon as pneumonia is suspected and before antibiotic therapy is initiated or changed.

Clinical Approach

General knowledge of the lack of specificity of a clinical diagnosis of VAP and results from invasive quantitative-culture studies have actually improved the clinical approach to the diagnosis of VAP. Tracheal aspirates generally yield at least twice as many potential pathogens as quantitative cultures. However, the causative pathogen is almost always present. The absence of bacteria in gram-stained endotracheal aspirates makes pneumonia an unlikely cause of fever or pulmonary infiltrates. These findings, coupled with a heightened awareness of the alternative diagnoses possible in patients with suspected VAP, can prevent inappropriate overtreatment for pneumonia. Furthermore, the absence of an MDR pathogen in tracheal aspirate cultures eliminates the need for MDR coverage, allowing empirical antibiotic therapy to be de-escalated. Since the main benefits of bronchoscopic quantitative cultures are decreased antibiotic selection

pressure (which reduces the risk of subsequent infection with MDR pathogens) and the identification of alternative sources of infection, a clinical diagnostic approach that incorporates such principles may result in similar outcomes.

TREATMENT

TREATMENT

Ventilator-Associated Pneumonia

Many studies have demonstrated higher mortality rates with initially inappropriate empirical antibiotic therapy. The key to appropriate antibiotic management of VAP is an appreciation of the resistance patterns of the most likely pathogens in a given patient.

ANTIBIOTIC RESISTANCE

If not for the higher risk of infection with MDR pathogens (Table 121-6), VAP could be treated with the same antibiotics used for severe CAP. However, antibiotic selection pressure leads to the frequent involvement of MDR pathogens by selecting either for drug-resistant isolates of common pathogens (MRSA and Enterobacteriaceae producing extended-spectrum β-lactamases or carbapenemases) or for intrinsically resistant pathogens (*P. aeruginosa* and *Acinetobacter* species). Frequent use of β-lactam drugs, especially cephalosporins, appears to be the major risk factor for infection with MRSA and extended-spectrum β-lactamase–positive strains.

P. aeruginosa has demonstrated the ability to develop resistance to all routinely used antibiotics. Unfortunately, even if initially sensitive, *P. aeruginosa* isolates also have a propensity to develop resistance during treatment. Either de-repression of resistance genes or selection of resistant clones within the large bacterial inoculum associated with most pneumonias may be the cause. *Acinetobacter* species, *Stenotrophomonas maltophilia*, and *Burkholderia cepacia* are intrinsically resistant to many of the empirical antibiotic regimens employed (see below). VAP caused by these pathogens emerges during treatment of other infections, and resistance is always evident at initial diagnosis.

EMPIRICAL THERAPY

Recommended options for empirical therapy are listed in **Table 121-8**. Treatment should be started once diagnostic specimens have been obtained. The major factor in the selection of agents is the presence of risk factors for MDR pathogens. Choices among the various options listed depend on local patterns of resistance and—a very important factor—the patient's prior antibiotic exposure. Knowledge of the local hospital's—and even the specific ICU's—antibiogram and the local incidence of specific MDR pathogens (e.g., MRSA) is critical in selecting appropriate empirical therapy.

The majority of patients *without* risk factors for MDR infection can be treated with a single agent. Unfortunately, the proportion of patients with no MDR risk factors is <10% in some ICUs and is unknown for HAP patients. The major difference from CAP is the markedly lower incidence of atypical pathogens in VAP; the exception is *Legionella*, which can be a nosocomial pathogen, especially with breakdowns in the

treatment of potable water in the hospital. The standard recommendation for patients *with* risk factors for MDR infection is for three antibiotics: two directed at *P. aeruginosa* and one at MRSA. A β-lactam agent provides the greatest coverage, yet even the broadest-spectrum agent—a carbapenem—still provides inappropriate initial therapy in up to 10–15% of cases at some centers. The emergence of carbapenem resistance at some institutions requires the addition of polymyxins to the combination-therapy options.

SPECIFIC TREATMENT

Once an etiologic diagnosis is made, broad-spectrum empirical therapy can be modified to specifically address the known pathogen. For patients with MDR risk factors, antibiotic regimens can be reduced to a single agent in most cases. Only a minority of cases require a complete course with two or three drugs. A negative tracheal-aspirate culture or growth below the threshold for quantitative cultures of samples obtained before any antibiotic change strongly suggests that antibiotics should be discontinued or that a search for an alternative diagnosis should be pursued. Identification of other confirmed or suspected sites of infection may require ongoing antibiotic therapy, but the spectrum of pathogens (and the corresponding antibiotic choices) may be different from those for VAP. A 7- or 8-day course of therapy is just as effective as a 2-week course and is associated with less frequent emergence of antibiotic-resistant strains.

The major controversy regarding specific therapy for VAP concerns the need for ongoing combination treatment of *Pseudomonas* pneumonia. No randomized controlled trials have demonstrated a benefit of combination therapy with a β-lactam and an aminoglycoside, nor have subgroup analyses in other trials found a survival benefit with such a regimen. The unacceptably high rates of clinical failure and death for VAP caused by *P. aeruginosa* despite combination therapy (see “Failure to Improve,” below) indicate that better regimens are needed, perhaps including aerosolized antibiotics. Current guidelines recommend against continued combination therapy for most cases of *Pseudomonas* pneumonia.

FAILURE TO IMPROVE

Treatment failure is not uncommon in VAP, especially that caused by MDR pathogens. VAP caused by MRSA is associated with a 40% clinical failure rate when treated with standard-dose vancomycin. One proposed but unproven solution is the use of high-dose individualized treatment, although the risk of renal toxicity increases with this strategy. In addition, the MIC of vancomycin has been increasing, and a high percentage of clinical failures occur when the MIC is in the upper range of sensitivity (i.e., 1.5–2 μg/mL). Linezolid appears to be 15% more efficacious than even adjusted-dose vancomycin and is clearly preferred in patients with renal insufficiency and those infected with high-MIC isolates of MRSA. VAP due to *Pseudomonas* has a 40–50% failure rate, no matter what the regimen. Causes of clinical failure vary with the pathogen(s) and the antibiotic(s). Inappropriate initial therapy can usually be minimized by use of the recommended combination regimen (Table 121-8). However, the emergence of β-lactam resistance during therapy is an important problem, especially in infection with *Pseudomonas* and *Enterobacter* species. Recurrent VAP caused by the same pathogen is possible because the biofilm on endotracheal tubes allows reintroduction of the microorganism. Studies of VAP caused by *Pseudomonas* show that approximately half of recurrent cases are caused by a new strain.

Treatment failure is very difficult to diagnose early in the therapeutic course, and discrimination among the various potential causes is a challenge. Pneumonia due to a new superinfection, the presence of extrapulmonary infection, and drug toxicity must be considered in the differential diagnosis of treatment failure. Serial measurements of procalcitonin levels appear to track the clinical response accurately, while repeat quantitative cultures may clarify the microbiologic response.

COMPLICATIONS

Apart from death, the major complication of VAP is prolongation of mechanical ventilation, with corresponding increases in the duration of ICU stay and hospitalization. In most studies, an additional week of mechanical ventilation resulting from VAP is common. The additional expense of this complication often warrants costly and aggressive efforts at prevention.

In rare cases, necrotizing pneumonia (e.g., that due to *P. aeruginosa*) can cause significant pulmonary hemorrhage. More commonly, necrotizing infections result in the long-term complications of bronchiectasis and parenchymal scarring leading to recurrent pneumonia. Other long-term complications of pneumonia are underappreciated. Pneumonia results in a catabolic state in a patient already nutritionally at risk. The muscle loss and general debilitation from an episode of VAP often require prolonged rehabilitation and, in the elderly, often result in an inability to return to independent function and the need for nursing home placement.

FOLLOW-UP

Clinical improvement, if it occurs, is usually evident within 48–72 h of the initiation of antimicrobial treatment. Because findings on chest radiography often worsen initially during treatment, they are less helpful than clinical criteria as an indicator of clinical response in severe pneumonia.

TABLE 121-8

Empirical Antibiotic Treatment of Hospital-Acquired and Ventilator-Associated Pneumonia

NO RISK FACTORS FOR RESISTANT GRAM-NEGATIVE PATHOGEN	RISK FACTORS FOR RESISTANT GRAM-NEGATIVE PATHOGEN[a] *(CHOOSE ONE FROM EACH COLUMN)*	
Piperacillin-tazobactam (4.5 g IV q6h[b]) Cefepime (2 g IV q8h) Levofloxacin (750 mg IV q24h)	Piperacillin-tazobactam (4.5 g IV q6h[b]) Cefepime (2 g IV q8h) Ceftazidime (2 g IV q8h) Imipenem (500 mg IV q6h[b]) Meropenem (1 g IV q8h)	Amikacin (15–20 mg/kg IV q24h) Gentamicin (5–7 mg/kg IV q24h) Tobramycin (5–7 mg/kg IV q24h) Ciprofloxacin (400 mg IV q8h) Levofloxacin (750 mg IV q24h) Colistin (loading dose of 5 mg/kg IV followed by maintenance doses of 2.5 mg × [1.5 × CrCl + 30] IV q12h) Polymyxin B (2.5–3.0 mg/kg per day IV in 2 divided doses)
Risk Factors for MRSA[b] *(Add to above)*		
Linezolid (600 mg IV q12h) *or* Adjusted-dose vancomycin (trough level, 15–20 mg/dL)		

[a]Prior antibiotic therapy, prior hospitalization, local antibiogram. [b]Prior antibiotic therapy, prior hospitalization, known MRSA colonization, chronic hemodialysis, local documented MRSA pneumonia rate >10% (or local rate unknown).

Abbreviations: CrCl, creatinine clearance rate; MRSA, methicillin-resistant *Staphylococcus aureus*.

PROGNOSIS

VAP is associated with crude mortality rates as high as 50–70%, but the real issue is attributable mortality. Many patients with VAP have underlying diseases that would result in death even if VAP did not occur. Attributable mortality exceeded 25% in one matched-cohort study, while more recent studies have suggested

much lower rates. Some variability in VAP mortality rates is clearly related to the type of patient and ICU studied. VAP in trauma patients is not associated with attributable mortality, possibly because many of the patients were otherwise healthy before being injured. The causative pathogen also plays a major role. Generally, MDR pathogens are associated with significantly greater attributable mortality than non-MDR pathogens. Pneumonia caused by some pathogens (e.g., *S. maltophilia*) is simply a marker for a patient whose immune system is so compromised that death is almost inevitable.

PREVENTION

Because of the significance of endotracheal intubation as a risk factor for VAP, the most important preventive intervention is to avoid intubation or minimize its duration (Table 121-7). Successful noninvasive ventilation avoids many of the problems associated with endotracheal tubes. Strategies that minimize the duration of ventilation through daily holding of sedation and formal weaning protocols have also been highly effective in preventing VAP.

Unfortunately, a tradeoff in risks is sometimes necessary. Aggressive attempts to extubate early may result in reintubation(s) and increase aspiration, posing a risk of VAP. Heavy continuous sedation increases VAP risk, but self-extubation because of insufficient sedation is also a risk. The tradeoffs also apply to antibiotic therapy. Short-course antibiotic prophylaxis can decrease the risk of VAP in comatose patients requiring intubation, and data suggest that antibiotics decrease VAP rates overall. However, the major benefit appears to be a decrease in the incidence of early-onset VAP, which is usually caused by the less pathogenic non-MDR microorganisms. Conversely, prolonged courses of antibiotics consistently increase the risk of VAP caused by more lethal MDR pathogens. Despite its virulence and associated mortality, VAP caused by *Pseudomonas* is rare among patients who have not recently received antibiotics.

Minimizing microaspiration around the endotracheal tube cuff is also a strategy for avoidance of VAP. Simply elevating the head of the bed (at least 30° above horizontal but preferably 45°) decreases VAP rates. Specially modified endotracheal tubes that allow removal of the secretions pooled above the cuff may also prevent VAP. The risk-to-benefit ratio of transporting the patient outside the ICU for diagnostic tests or procedures should be carefully considered, since VAP rates are increased among transported patients.

The role played by overgrowth of the normal bowel flora in the stomach in the pathogenesis of VAP is questionable. MRSA and the nonfermenters *P. aeruginosa* and *Acinetobacter* species are not normally part of the bowel flora but reside primarily in the nose and on the skin, respectively. Therefore, emphasis on controlling overgrowth of the bowel flora by avoidance of agents that raise gastric pH may be relevant only in certain populations, such as liver transplant recipients and patients who have undergone other major intraabdominal procedures or who have bowel obstruction.

In outbreaks of VAP due to specific pathogens, the possibility of a breakdown in infection control measures (particularly contamination of reusable equipment) should be investigated. Even high rates of pathogens that are already common in a particular ICU may result from cross-infection. Education and reminders of the need for consistent hand washing and other infection-control practices can minimize this risk.

HOSPITAL-ACQUIRED PNEUMONIA

While significantly less well studied than VAP, HAP in non-intubated patients—both inside and outside the ICU—is similar to VAP. The main differences are the higher frequency of non-MDR pathogens and the generally better underlying host immunity in non-intubated patients. The lower frequency of MDR pathogens allows monotherapy in a larger proportion of cases of HAP than of VAP.

The only pathogens that may be more common in the non-VAP population are anaerobes. The greater risk of macroaspiration by non-intubated patients and the lower oxygen tensions in the lower respiratory tract of these patients increase the likelihood of a role for anaerobes. While more common in patients with HAP, anaerobes usually contribute only to polymicrobial pneumonias. As in the management of CAP, specific therapy targeting anaerobes probably is not needed since many of the recommended antibiotics are active against anaerobes.

Diagnosis is even more difficult for HAP in the non-intubated patient than for VAP. Lower respiratory tract samples appropriate for culture are considerably more difficult to obtain from non-intubated patients. Many of the underlying diseases that predispose a patient to HAP are also associated with an inability to cough adequately. Since blood cultures are infrequently positive (<15% of cases), the majority of patients with HAP do not have culture data on which antibiotic modifications can be based. Therefore, de-escalation of therapy is less likely in patients with risk factors for MDR pathogens. Despite these difficulties, the better host defenses in non-ICU patients result in lower mortality rates than are documented for VAP. In addition, the risk of antibiotic failure is lower in HAP.

GLOBAL IMPACT

From the available data, it is virtually impossible to accurately assess the impact of pneumonia from a global perspective. Any differences in incidence, disease burden, and costs across different age, ethnic, and racial groups are compounded by differences among countries in terms of etiologic pathogens, resistance rates, access to health-care and diagnostic facilities, and vaccine availability and usage.

A standard approach with clearly defined outcome measures is needed before the impact of pneumonia can be accurately evaluated. However, simple extrapolation from U.S. data for CAP and HAP/VAP shows that pneumonia has a significant impact on quality of life, morbidity, health costs, and mortality rates and that this impact has implications for patients and for society as a whole.

FURTHER READING

Chastre J et al: Comparison of 8 vs 15 days of antibiotic therapy for ventilator-associated pneumonia in adults: a randomized trial. JAMA 290:2588, 2003. [PubMed: 14625336]

Fagon JY et al: Invasive and noninvasive strategies for management of suspected ventilator-associated pneumonia. A randomized trial. Ann Intern Med 132:621, 2000. [PubMed: 10766680]

Jain S et al: Community-acquired pneumonia requiring hospitalization among U.S. adults. N Engl J Med 373:415, 2015. [PubMed: 26172429]

Kalil AC et al: Management of adults with hospital-acquired and ventilator-associated pneumonia: 2016 clinical practice guidelines by the Infectious Diseases Society of America and the American Thoracic Society. Clin Infect Dis 63:e61, 2016. [PubMed: 27418577]

Mandell LA et al: Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. Clin Infect Dis 44(Suppl 2):S27, 2007. [PubMed: 17278083]

Musher DM, Thorner AR: Community-acquired pneumonia. N Engl J Med 371:1619, 2014. [PubMed: 25337751]

Shindo Y et al: Risk factors for drug-resistant pathogens in community-acquired and healthcare-associated pneumonia. Am J Respir Crit Care Med 188:985, 2013. [PubMed: 23855620]

Wunderink RG, Waterer GW: Community-acquired pneumonia. N Engl J Med 370:541, 2014.

Chapter 122: Lung Abscess

Rebecca M. Baron; Miriam Baron Barshak

FIGURE 122-1

INTRODUCTION

Lung abscess represents necrosis and cavitation of the lung following microbial infection. Lung abscesses can be single or multiple but usually are marked by a single dominant cavity >2 cm in diameter.

ETIOLOGY

The low prevalence of lung abscesses makes them difficult to study in randomized controlled trials. Although the incidence of lung abscesses has decreased in the antibiotic era, they are still a source of significant morbidity and mortality.

Lung abscesses are usually characterized as either primary (~80% of cases) or secondary. *Primary* lung abscesses usually arise from aspiration, are often caused principally by anaerobic bacteria, and occur in the absence of an underlying pulmonary or systemic condition. *Secondary* lung abscesses arise in the setting of an underlying condition, such as a postobstructive process (e.g., a bronchial foreign body or tumor) or a systemic process (e.g., HIV infection or another immunocompromising condition). Lung abscesses can also be characterized as acute (<4–6 weeks in duration) or chronic (~40% of cases).

EPIDEMIOLOGY

The majority of the existing epidemiologic information involves primary lung abscesses. In general, middle-aged men are more commonly affected than middle-aged women. The major risk factor for primary lung abscesses is aspiration. Patients at particular risk for aspiration, such as those with altered mental status, alcoholism, drug overdose, seizures, bulbar dysfunction, prior cerebrovascular or cardiovascular events, or neuromuscular disease, are most commonly affected. In addition, patients with esophageal dysmotility or esophageal lesions (strictures or tumors) and those with gastric distention and/or gastroesophageal reflux, especially those who spend substantial time in the recumbent position, are at risk for aspiration.

It is widely thought that colonization of the gingival crevices by anaerobic bacteria or microaerophilic streptococci (especially in patients with gingivitis and periodontal disease), combined with a risk of

aspiration, is important in the development of lung abscesses. In fact, many physicians consider it extremely rare for lung abscesses to develop in the absence of teeth as a nidus for bacterial colonization.

The importance of these risk factors in the development of lung abscesses is highlighted by a significant reduction in abscess incidence in the late 1940s that coincided with a change in oral surgical technique: beginning at that time, these operations were no longer performed with the patient in the seated position without a cuffed endotracheal tube, and the frequency of perioperative aspiration events was thus decreased. In addition, the introduction of penicillin around the same time significantly reduced the incidence of and mortality rate from lung abscess.

PATHOGENESIS

Primary Lung Abscesses

The development of primary lung abscesses is thought to originate when chiefly anaerobic bacteria (as well as microaerophilic streptococci) in the gingival crevices are aspirated into the lung parenchyma in a susceptible host **(Table 122-1)**. Patients who develop primary lung abscesses usually carry an overwhelming burden of aspirated material or are unable to clear the bacterial load. Pneumonitis develops initially (exacerbated in part by tissue damage caused by gastric acid); then, over a period of 7–14 days, the anaerobic bacteria produce parenchymal necrosis and cavitation whose extent depends on host–pathogen interaction **(Fig. 122-1)**. Anaerobes are thought to produce more extensive tissue necrosis in polymicrobial infections in which virulence factors of the various bacteria can act synergistically to cause more significant tissue destruction.

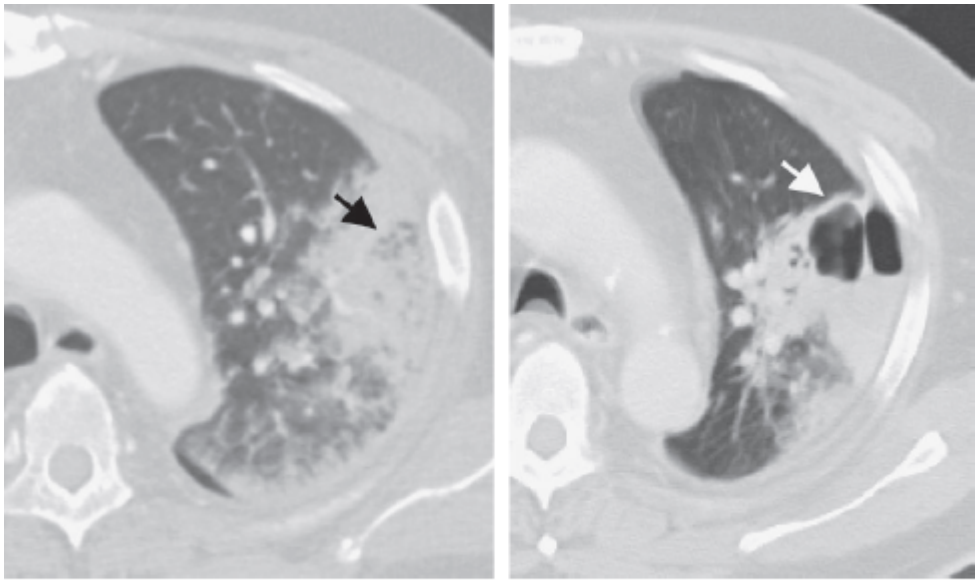
TABLE 122-1

Examples of Microbial Pathogens That Can Cause Lung Abscesses

CLINICAL CONDITION	PATHOGENS
Primary lung abscess (usually with risk factors for aspiration)	Anaerobes (e.g., *Peptostreptococcus* spp., *Prevotella* spp., *Bacteroides* spp., *Streptococcus milleri*), microaerophilic streptococci
Secondary lung abscess (often with underlying immunocompromise)	*Staphylococcus aureus*, gram-negative rods (e.g., *Pseudomonas aeruginosa*, Enterobacteriaceae), *Nocardia* spp., *Aspergillus* spp., Mucorales, *Cryptococcus* spp., *Legionella* spp., *Rhodococcus equi*, *Pneumocystis jirovecii*
Embolic lesions	*Staphylococcus aureus* (often from endocarditis), *Fusobacterium necrophorum* (Lemierre's syndrome; see text for details)
Endemic infections (with or without underlying immunocompromise)	*Mycobacterium tuberculosis* (as well as *Mycobacterium avium* and *Mycobacterium kansasii*), *Coccidioides* spp., *Histoplasma capsulatum*, *Blastomyces* spp., parasites (e.g., *Entamoeba histolytica*, *Paragonimus westermani*, *Strongyloides stercoralis*)
Miscellaneous conditions	Bacterial pathogen (often *S. aureus*) after influenza or another viral infection, *Actinomyces* spp.

FIGURE 122-1

Representative chest CT scans demonstrating development of lung abscesses. This patient was immunocompromised by underlying lymphoma and developed severe *Pseudomonas aeruginosa* pneumonia, as represented by a left lung infiltrate with concern for central regions of necrosis (*panel **A**, black arrow*). Two weeks later, areas of cavitation with air-fluid levels were visible in this region and were consistent with the development of lung abscesses (*panel **B**, white arrow*). *(Images provided by Dr. Ritu Gill, Division of Chest Radiology, Brigham and Women's Hospital, Boston; with permission.)*



A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Secondary Lung Abscesses

The pathogenesis of secondary abscesses depends on the predisposing factor. For example, in cases of bronchial obstruction from malignancy or a foreign body, the obstructing lesion prevents clearance of oropharyngeal secretions, leading to abscess development. With underlying systemic conditions (e.g., immunosuppression after bone marrow or solid organ transplantation), impaired host defense mechanisms lead to increased susceptibility to the development of lung abscesses caused by a broad range of pathogens, including opportunistic organisms (Table 122-1).

Lung abscesses also arise from septic emboli, either in tricuspid valve endocarditis (often involving *Staphylococcus aureus*) or in Lemierre's syndrome, in which an infection begins in the pharynx (classically involving *Fusobacterium necrophorum*) and then spreads to the neck and the carotid sheath (which contains the jugular vein) to cause septic thrombophlebitis.

PATHOLOGY AND MICROBIOLOGY

Primary Lung Abscesses

The dependent segments (posterior upper lobes and superior lower lobes) are the most common locations of primary lung abscesses, given the predisposition of aspirated materials to be deposited in these areas. Generally, the right lung is affected more commonly than the left because the right mainstem bronchus is less angulated.

The microbiology of primary lung abscesses is often polymicrobial, primarily including anaerobic organisms as well as microaerophilic streptococci (Table 122-1). The retrieval and culture of anaerobes can be complicated by the contamination of samples with microbes from the oral cavity, the need for expeditious

transport of the cultures to the laboratory, the need for early plating with special culture techniques, the prolonged time required for culture growth, and the need for collection of specimens prior to administration of antibiotics. When attention is paid to these factors, rates of recovery of specific isolates are reportedly as high as 78%.

Because it is not clear that knowing the identity of the causative anaerobic isolate alters the response to treatment of a primary lung abscess, practice has shifted away from the use of specialized techniques to obtain material for culture, such as transtracheal aspiration and bronchoalveolar lavage with protected brush specimens that allow recovery of culture material while avoiding contamination from the oral cavity. When no pathogen is isolated from a primary lung abscess (which is the case as often as 40% of the time), the abscess is termed a *nonspecific lung abscess*, and the presence of anaerobes is often presumed. A *putrid lung abscess* refers to cases with foul-smelling breath, sputum, or empyema; these manifestations are essentially diagnostic of an anaerobic lung abscess.

Secondary Lung Abscesses

The location of secondary abscesses may vary with the underlying cause. The microbiology of secondary lung abscesses can encompass quite a broad bacterial spectrum, with infection by *Pseudomonas aeruginosa* and other gram-negative rods most common. In addition, a broad array of pathogens can be identified in patients from certain endemic areas and in specific clinical scenarios (e.g., a significant incidence of fungal infections among immunosuppressed patients following bone marrow or solid organ transplantation). Because immunocompromised hosts and patients without the classic presentation of a primary lung abscess can be infected with a wide array of unusual organisms (Table 122-1), it is of special importance to obtain culture material in order to target therapy.

CLINICAL MANIFESTATIONS

Clinical manifestations may initially be similar to those of pneumonia, with fevers, cough, sputum production, and chest pain; a more chronic and indolent presentation that includes night sweats, fatigue, and anemia is often observed with anaerobic lung abscesses. A subset of patients with putrid lung abscesses may report discolored phlegm and foul-tasting or foul-smelling sputum. Patients with lung abscesses due to non-anaerobic organisms, such as *S. aureus*, may present with a more fulminant course characterized by high fevers and rapid progression.

Findings on physical examination may include fevers, poor dentition, and/or gingival disease as well as amphoric and/or cavernous breath sounds on lung auscultation. Additional findings may include digital clubbing and the absence of a gag reflex.

DIFFERENTIAL DIAGNOSIS

The differential diagnosis of lung abscesses is broad and includes other noninfectious processes that result in cavitary lung lesions, including lung infarction, malignancy, sequestration, cryptogenic organizing pneumonia, sarcoidosis, vasculitides and other autoimmune diseases (e.g., granulomatosis with polyangiitis), lung cysts or bullae containing fluid, and septic emboli (e.g., from tricuspid valve endocarditis). Other less common entities can include pulmonary manifestations of diseases that usually present at locations other than the chest (e.g., inflammatory bowel disease, pyoderma gangrenosum).

DIAGNOSIS

Lung abscesses are documented by chest imaging. Although a chest radiograph usually detects a thick-walled cavity with an air-fluid level, CT permits better definition and may provide earlier evidence of cavitation. CT may also yield additional information regarding a possible underlying cause of lung abscess, such as malignancy, and may help distinguish a peripheral lung abscess from a pleural infection. This distinction has important implications for treatment because a pleural space infection, such as an empyema, may require urgent drainage.

As described earlier (see "Pathology and Microbiology," above), more invasive diagnostics (such as transtracheal aspiration) were traditionally undertaken for primary lung abscesses, whereas empirical therapy that includes drugs targeting anaerobic organisms currently is used more often. While sputum can be collected noninvasively for Gram's stain and culture, which may yield a pathogen, the infection is likely to be polymicrobial, and culture results may not reflect the presence of anaerobic organisms. As stated above, many physicians consider putrid-smelling sputum to be virtually diagnostic of an anaerobic infection.

When a secondary lung abscess is present or empirical therapy fails to elicit a response, sputum and blood cultures are advised in addition to serologic studies for opportunistic pathogens (e.g., viruses and fungi causing infections in immunocompromised hosts). Additional diagnostics, such as bronchoscopy with bronchoalveolar lavage or protected brush specimen collection and CT-guided percutaneous needle aspiration, can be undertaken. Risks posed by these more invasive diagnostics include spillage of abscess contents into the other lung (with bronchoscopy) and pneumothorax and bronchopleural fistula development (with CT-guided needle aspiration). However, early diagnostics in secondary abscesses, especially in immunocompromised hosts, are particularly important, because the patients involved may be especially fragile, at risk for infection with a broad array of pathogens, and therefore less likely than other patients to respond to empirical therapy.

TREATMENT

TREATMENT

Lung Abscess

The availability of antibiotics in the 1940s and 1950s established therapy with this drug class as the primary approach to the treatment of lung abscess. Previously, surgery had been relied upon much more frequently.

For many decades, penicillin was the antibiotic of choice for primary lung abscesses in light of its anaerobic coverage; however, because oral anaerobes can produce β-lactamases, clindamycin has proved superior to penicillin in clinical trials. For primary lung abscesses, the recommended regimens are (1) clindamycin (600 mg IV three times daily; then, with the disappearance of fever and clinical improvement, 300 mg PO four times daily) or (2) an IV-administered β-lactam/β-lactamase combination, followed—once the patient's condition is stable—by orally administered amoxicillin-clavulanate. This therapy should be continued until imaging demonstrates that the lung abscess has cleared or regressed to a small scar. Treatment duration may range from 3–4 weeks to as long as 14 weeks. One small study suggested that moxifloxacin (400 mg/d PO) is as effective and well tolerated as ampicillin-sulbactam. Notably, metronidazole is not effective as a single agent: it covers anaerobic organisms but not the microaerophilic streptococci that are often components of the mixed flora of primary lung abscesses.

In secondary lung abscesses, antibiotic coverage should be directed at the identified pathogen, and a prolonged course (until resolution of the abscess is documented) is often required. Treatment regimens and courses vary widely, depending on the immune state of the host and the identified pathogen. Other interventions may be necessary as well, such as relief of an obstructing lesion or treatment directed at the underlying condition predisposing the patient to lung abscess. Similarly, if the condition of patients with presumed primary lung abscess fails to improve, additional studies to rule out an underlying predisposing cause for a secondary lung abscess are indicated.

Although it can take as long as 7 days for patients receiving appropriate therapy to defervesce, as many as 10–20% of patients may not respond at all, with continued fevers and progression of the abscess cavity on imaging. An abscess >6–8 cm in diameter is less likely to respond to antibiotic therapy without additional interventions. Options for patients who do not respond to antibiotics and whose additional diagnostic studies fail to identify an additional pathogen that can be treated include surgical resection and percutaneous drainage of the abscess, especially when the patient is a poor surgical candidate. Timing of surgical intervention can be challenging; the goal is to balance the morbidity/mortality risk of a procedure with the need for definitively clearing the abscess in the setting of persistent infection that is not responsive to nonsurgical approaches. Possible complications of percutaneous drainage include bacterial contamination of the pleural space as well as pneumothorax and hemothorax.

COMPLICATIONS

Larger cavity size on presentation may correlate with the development of persistent cystic changes (pneumatoceles) or bronchiectasis. Additional possible complications include recurrence of abscesses despite appropriate therapy, extension to the pleural space with development of empyema, life-threatening hemoptysis, and massive aspiration of lung abscess contents.

PROGNOSIS AND PREVENTION

Reported mortality rates for primary abscesses have been as low as 2%, while rates for secondary abscesses are generally higher—as high as 75% in some case series. Other poor prognostic factors include an age of >60, the presence of aerobic bacteria, sepsis at presentation, symptom duration of >8 weeks, and abscess size of >6 cm.

Mitigation of underlying risk factors may be the best approach to prevention of lung abscesses, with attention directed toward airway protection, oral hygiene, and minimized sedation with elevation of the head of the bed for patients at risk for aspiration. Prophylaxis against certain pathogens in at-risk patients (e.g., recipients of bone marrow or solid organ transplants or patients whose immune systems are significantly compromised by HIV infection) may be undertaken.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Lung Abscess

For patients with a lung abscess and a low likelihood of malignancy (e.g., smokers <45 years old) and with risk factors for aspiration, it is reasonable to administer empirical treatment and then to pursue further evaluation if therapy does not elicit a response. However, some clinicians may opt for up-front cultures, even in primary lung abscesses. In patients with risk factors for malignancy or other underlying conditions (especially immunocompromised hosts) or with an atypical presentation, earlier diagnostics should be considered, such as bronchoscopy with biopsy or CT-guided needle aspiration. Bronchoscopy should be performed early in patients whose history, symptoms, or imaging findings are consistent with possible bronchial obstruction. In patients from areas endemic for tuberculosis or patients with other risk factors for tuberculosis (e.g., underlying HIV infection), induced sputum samples should be examined early in the workup to rule out this disease.

FURTHER READING

Bartlett JG: How important are anaerobic bacteria in aspiration pneumonia: When should they be treated and what is optimal therapy. Infect Dis Clin North Am 27:149, 2013. [PubMed: 23398871]

Davis B, Systrom DM: Lung abscess: Pathogenesis, diagnosis, and treatment. Curr Clin Top Infect Dis 18:252, 1998. [PubMed: 9779359]

Desai H, Agrawal A: Pulmonary emergencies: Pneumonia, acute respiratory distress syndrome, lung abscess, and empyema. Med Clin North Am 96:1127, 2012. [PubMed: 23102481]

Ott SR et al: Moxifloxacin vs ampicillin/sulbactam in aspiration pneumonia and primary lung abscess. Infection 36:23, 2008. [PubMed: 18231720]

Raymond D: Surgical intervention for thoracic infections. Surg Clin North Am 94:1283, 2014. [PubMed: 25440124]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 123: Infective Endocarditis

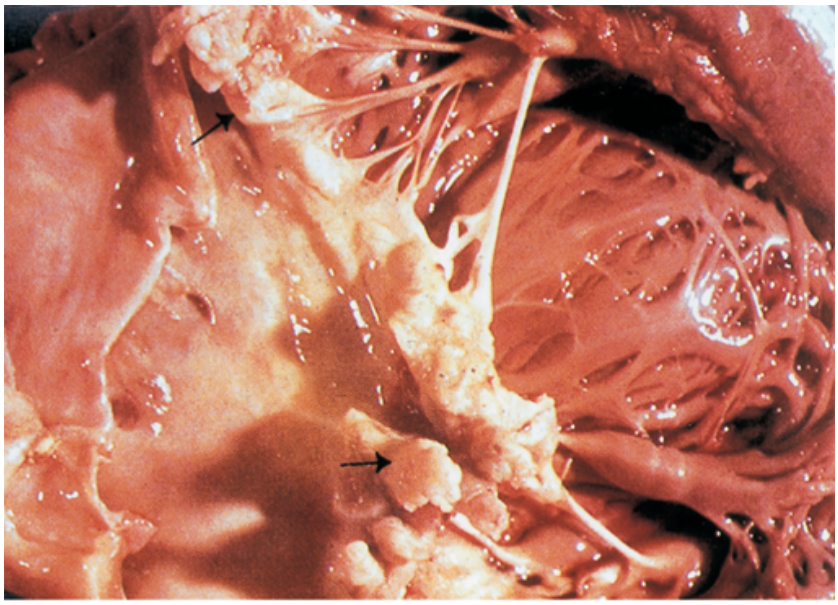
Adolf W. Karchmer

INTRODUCTION

The prototypic lesion of infective endocarditis, the *vegetation* (Fig. 123-1), is a mass of platelets, fibrin, microorganisms, and scant inflammatory cells. Infection most commonly involves heart valves but may also occur on the low-pressure side of a ventricular septal defect, on mural endocardium damaged by aberrant jets of blood or foreign bodies, or on intracardiac devices themselves. The analogous process involving arteriovenous shunts, arterio-arterial shunts (patent ductus arteriosus), or a coarctation of the aorta is called *infective endarteritis*.

FIGURE 123-1

Vegetations (arrows) due to viridans streptococcal endocarditis involving the mitral valve.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Endocarditis can be classified according to the temporal evolution of disease, the site of infection, the cause of infection, or the predisposing risk factor (e.g., injection drug use, association with health care). *Acute endocarditis* is a hectically febrile illness that rapidly damages cardiac structures, seeds extracardiac sites, and, if untreated, progresses to death within weeks. *Subacute endocarditis* follows an indolent course; causes structural cardiac damage only slowly, if at all; rarely metastasizes; and is gradually progressive unless complicated by a major embolic event or a ruptured mycotic aneurysm.

In developed countries, the incidence of endocarditis ranges from 4 to 7 cases per 100,000 population per year and has remained relatively stable during recent decades. While congenital heart diseases remain a constant predisposition, predisposing conditions in developed countries have shifted from chronic rheumatic heart disease (still a common predisposition in developing countries) to illicit IV drug use, degenerative valve disease, and intracardiac devices. The incidence of endocarditis is notably increased among the elderly. In developed countries, 25–35% of cases of native-valve endocarditis (NVE) are associated with health care, and 16–30% of all cases are prosthetic-valve infections (PVE). The risk of PVE is greatest during the first 6–12 months after valve replacement; gradually declines to a low, stable rate thereafter; and is similar for mechanical and bioprosthetic devices. The incidence of infection involving

transcatheter implanted aortic valves (a new form of PVE that is beyond the scope of this chapter) is 0.97% in the initial year of follow-up. Endocarditis involving cardiovascular implantable electronic devices (CIEDs)—primarily permanent pacemakers and implantable cardioverter-defibrillators—occurs in 0.5–1.14 cases per 1000 device recipients, with higher rates among patients with an implantable cardioverter-defibrillator than among those with a permanent pacemaker.

ETIOLOGY

Although many species of bacteria and fungi cause sporadic episodes of endocarditis, a few bacterial species cause the majority of cases **(Table 123-1)**. The oral cavity, skin, and upper respiratory tract are the respective primary portals for viridans streptococci, staphylococci, and HACEK organisms (*Haemophilus* species, *Aggregatibacter* species, *Cardiobacterium hominis*, *Eikenella corrodens*, and *Kingella kingae*). *Streptococcus gallolyticus* subspecies *gallolyticus* (formerly *S. bovis* biotype 1) originates from the gastrointestinal tract, where it is associated with polyps and colonic tumors, and enterococci enter the bloodstream primarily from the genitourinary tract. Health care–associated NVE, most commonly caused by *Staphylococcus aureus*, coagulase-negative staphylococci (CoNS), and enterococci, may have either a nosocomial onset (55%) or a community onset (45%) in patients who have had extensive contact with the health care system. Endocarditis complicates 6–25% of episodes of catheter-associated *S. aureus* bacteremia; the higher rates are detected in high-risk patients studied by transesophageal echocardiography (TEE) (see "Cardiac Imaging" below).

TABLE 123-1

Organisms Causing Major Clinical Forms of Endocarditis

	Percentage of Cases								
	Native-Valve Endocarditis		Prosthetic-Valve Endocarditis at Indicated Time of Onset (Months) after Valve Surgery			Endocarditis in Injection Drug Users			
Organism(s)	**Community-Acquired (*n* = 1718)**	**Health Care–Associated (*n* = 1110)**	**<2 (*n* = 144)**	**2–12 (*n* = 31)**	**>12 (*n* = 194)**	**Right-Sided (*n* = 346)**	**Left-Sided (*n* = 204)**	**Total (*n* = 675)[a]**	**CIED (*n* = 337)**
Streptococci[b]	40	13	1	9	31	5	15	12	2
Pneumococci	2	—	—	—	—	—	—	—	—
Enterococci[c]	9	16	8	12	11	2	24	9	4
Staphylococcus aureus	28	52[d]	22	12	18	77	23	57	36
Coagulase-negative staphylococci	5	11	33	32	11	—	—	—	41
Fastidious gram-negative coccobacilli (HACEK group)[e]	3	—	—	—	6	—	—	—	—
Gram-negative bacilli	1	1	13	3	6	5	13	7	6
Candida spp.	<1	1	8	12	1	—	12	4	2
Polymicrobial/miscellaneous	3	3	3	6	5	8	10	7	2
Diphtheroids	—	<1	6	—	3	—	—	0.1	1
Culture-negative	9	3	5	6	8	3	3	3	6

[a]The total number of cases is larger than the sum of right- and left-sided cases because the location of infection was not specified in some cases. [b]Includes viridans streptococci; *Streptococcus gallolyticus*; other non–group A, groupable streptococci; and *Abiotrophia* and *Granulicatella* spp. (nutritionally variant, pyridoxal-requiring streptococci). [c]Primarily *E. faecalis* or nonspeciated isolates; occasionally *E. faecium* or other, less likely species. [d]Methicillin resistance is common among these *S. aureus* strains. [e]Includes *Haemophilus* spp., *Aggregatibacter* spp., *Cardiobacterium hominis*, *Eikenella corrodens*, and *Kingella kingae*.

Abbreviation: CIED, cardiac implantable electronic device.

Note: Data are compiled from multiple studies.

PVE arising within 2 months of valve surgery—i.e., early PVE—is generally nosocomial and is the result of intraoperative contamination of the prosthesis or a bacteremic postoperative complication. This nosocomial origin is reflected in the primary microbial causes: *S.*

aureus, CoNS, facultative gram-negative bacilli, diphtheroids, and fungi. The portals of entry and organisms causing cases beginning >12 months after surgery—i.e., late PVE—are similar to those in community-acquired NVE. PVE due to CoNS that presents 2–12 months after surgery often represents delayed-onset nosocomial infection. Regardless of the time of onset after surgery, at least 68–85% of CoNS strains that cause PVE are resistant to methicillin.

CIED endocarditis involves the device or the endothelium at points of device contact. Occasionally, there is concurrent aortic or mitral valve infection. One-third of cases of CIED endocarditis present within 3 months after device implantation or manipulation, one-third present at 4–12 months, and one-third present at >1 year. *S. aureus* and CoNS, both of which are often resistant to methicillin, cause the majority of cases.

Injection drug use–associated endocarditis, especially that involving the tricuspid valve, is commonly caused by *S. aureus*, which in many cases is resistant to methicillin. Left-sided valve infections in addicts have a more varied etiology. In addition to the usual causes of endocarditis, these cases can be due to *Pseudomonas aeruginosa* and *Candida* species, and sporadic cases can be caused by unusual organisms such as *Bacillus*, *Lactobacillus*, and *Corynebacterium* species. Polymicrobial endocarditis occurs among injection drug users. HIV infection in drug users does not significantly influence the causes of endocarditis.

From 5 to 15% of patients with endocarditis have negative blood cultures; in one-third to one-half of these cases, cultures are negative because of prior antibiotic exposure. The remainder of these patients are infected by fastidious organisms, such as some streptococci (nutritionally variant bacteria now designated *Granulicatella* and *Abiotrophia* species), HACEK organisms, *Coxiella burnetii*, and *Bartonella* species. Some fastidious organisms occur in characteristic geographic settings (e.g., *C. burnetii* and *Bartonella* species in Europe, *Brucella* species in the Middle East). *Tropheryma whipplei* causes an indolent, culture-negative, afebrile form of endocarditis. *C. burnetii* has a predilection for prosthetic valves. *Corynebacterium* species and *Propionibacterium acnes* may involve intracardiac devices and be slow to grow in blood cultures. Lastly, atrial myxoma, marantic endocarditis, and the antiphospholipid antibody syndrome may mimic culture-negative infectious endocarditis.

PATHOGENESIS

The undamaged endothelium is resistant to infection by most bacteria and to thrombus formation. Endothelial injury (e.g., at the site of impact of high-velocity blood jets or on the low-pressure side of a cardiac structural lesion) allows either direct infection by virulent organisms or the development of a platelet–fibrin thrombus—a condition called *nonbacterial thrombotic endocarditis* (NBTE). This thrombus serves as a site of bacterial attachment during transient bacteremia. The cardiac conditions most commonly resulting in NBTE are mitral regurgitation, aortic stenosis, aortic regurgitation, ventricular septal defects, and complex congenital heart disease. NBTE also arises as a result of a hypercoagulable state; this phenomenon gives rise to *marantic endocarditis* (uninfected vegetations seen in patients with malignancy and chronic diseases) and to bland vegetations complicating systemic lupus erythematosus and antiphospholipid antibody syndrome.

Organisms that cause endocarditis enter the bloodstream from mucosal surfaces, the skin, or sites of focal infection. Except for more virulent bacteria (e.g., *S. aureus*) that can adhere directly to intact endothelium or exposed subendothelial tissue, microorganisms in the blood adhere at sites of NBTE. The organisms that commonly cause endocarditis have surface adhesin molecules, collectively called microbial surface components recognizing adhesin matrix molecules (MSCRAMMs), that mediate adherence to NBTE sites or injured endothelium. Adherence is facilitated by fibronectin-binding proteins present on many gram-positive bacteria; by clumping factor (a fibrinogen- and fibrin-binding surface protein) on *S. aureus*; by fibrinogen-binding surface proteins (Fss2), collagen-binding surface protein (Ace), and Ebp pili (the latter mediating platelet adherence) on *Enterococcus faecalis*; and by glucans or FimA (a member of the family of oral mucosal adhesins) on streptococci. Fibronectin-binding proteins are required for *S. aureus* invasion of intact endothelium; thus these surface proteins may facilitate infection of previously normal valves. If resistant to the bactericidal activity of serum and the microbicidal peptides released locally by platelets, adherent organisms proliferate to form dense microcolonies. Microorganisms also induce platelet deposition and a localized procoagulant state by eliciting tissue factor from the endothelium and, in the case of *S. aureus,* from monocytes as well. Fibrin deposition combines with platelet aggregation and microorganism proliferation to generate an infected vegetation. Organisms deep in vegetations are metabolically inactive (nongrowing) and relatively resistant to killing by antimicrobial agents. Proliferating surface organisms are shed into the bloodstream continuously.

The clinical manifestations of endocarditis—other than constitutional symptoms, which probably result from cytokine production—arise from damage to intracardiac structures; embolization of vegetation fragments, leading to infection or infarction of remote tissues; hematogenous infection of sites during bacteremia; and tissue injury due to the deposition of circulating immune complexes or immune responses to deposited bacterial antigens.

CLINICAL MANIFESTATIONS

The clinical endocarditis syndrome is highly variable and spans a continuum between acute and subacute presentations. NVE, PVE, and endocarditis due to injection drug use share clinical and laboratory manifestations (Table 123-2). The causative microorganism is primarily responsible for the temporal course of endocarditis. β-Hemolytic streptococci, *S. aureus*, and pneumococci typically result in an acute course, although *S. aureus* occasionally causes subacute disease. Endocarditis caused by *Staphylococcus lugdunensis* (a coagulase-negative species) or by enterococci may present acutely. Subacute endocarditis is typically caused by viridans streptococci, enterococci, CoNS, and the HACEK group. Endocarditis caused by *Bartonella* species, *T. whipplei*, or *C. burnetii* is exceptionally indolent.

TABLE 123-2

Clinical and Laboratory Features of Infective Endocarditis

Feature	Frequency, %
Fever	80–90
Chills and sweats	40–75
Anorexia, weight loss, malaise	25–50
Myalgias, arthralgias	15–30
Back pain	7–15
Heart murmur	80–85
New/worsened regurgitant murmur	20–50
Arterial emboli	20–50
Splenomegaly	15–50
Clubbing	10–20
Neurologic manifestations	20–40
Peripheral manifestations (Osler's nodes, subungual hemorrhages, Janeway lesions, Roth's spots)	2–15
Petechiae	10–40
Laboratory manifestations	
Anemia	70–90
Leukocytosis	20–30
Microscopic hematuria	30–50
Elevated erythrocyte sedimentation rate	60–90
Elevated C-reactive protein level	>90
Rheumatoid factor	50
Circulating immune complexes	65–100
Decreased serum complement	5–40

In patients with subacute presentations, fever is typically low-grade and rarely exceeds 39.4°C (103°F); in contrast, temperatures of 39.4°–40°C (103°–104°F) are often noted in acute endocarditis. Fever may be blunted in patients who are elderly, are severely

debilitated, or have renal failure.

Cardiac Manifestations

Although heart murmurs are usually indicative of the predisposing cardiac pathology rather than of endocarditis, valvular damage and ruptured chordae may result in new regurgitant murmurs. In acute endocarditis involving a normal valve, murmurs may be absent initially but ultimately are detected in 85% of cases. Congestive heart failure (CHF) develops in 30–40% of patients as a consequence of valve dysfunction or, occasionally, intracardiac fistulae. Heart failure due to aortic valve dysfunction progresses more rapidly than does that due to mitral valve dysfunction. Extension of infection beyond valve leaflets into adjacent annular or myocardial tissue results in perivalvular abscesses, which in turn may cause intracardiac fistulae with new murmurs. Abscesses may burrow from the aortic valve annulus into the upper ventricular septum, where they may interrupt the conduction system, leading to varying degrees of heart block. Mitral perivalvular abscesses, which are usually more distant from the conduction system, only rarely cause conduction abnormalities. Emboli to a coronary artery occur in 2% of patients and may result in myocardial infarction.

Noncardiac Manifestations

The classic nonsuppurative peripheral manifestations of subacute endocarditis (e.g., Janeway lesions; **Fig. 123-2A**) are related to prolonged infection; with early diagnosis and treatment, these have become infrequent. In contrast, septic embolization mimicking some of these lesions (subungual hemorrhage, Osler's nodes) is common in patients with acute *S. aureus* endocarditis **(Fig. 123-2B)**. Musculoskeletal pain usually remits promptly with treatment but must be distinguished from focal metastatic infections (e.g., spondylodiscitis), which may complicate 10–15% of cases. Hematogenously seeded focal infection occurs most often in the skin, spleen, kidneys, skeletal system, and meninges. Arterial emboli, one-half of which precede the diagnosis of endocarditis, are clinically apparent in up to 50% of patients. Endocarditis caused by *S. aureus,* mobile vegetations >10 mm in diameter, and infection involving the mitral valve, especially the anterior leaflet, are independently associated with an increased risk of embolization. Embolic arterial occlusion causes regional pain or ischemia-induced organ dysfunction (e.g., of the kidney, spleen, bowel, extremity). Cerebrovascular emboli presenting as strokes or occasionally as encephalopathy complicate 15–35% of cases of endocarditis; however, evidence of clinically asymptomatic emboli is found on MRI in 30–65% of patients with left-sided endocarditis. The frequency of stroke is 8 per 1000 patient-days during the week prior to diagnosis; the subsequent fall in frequency—to 4.8 and 1.7 per 1000 patient-days during the first and second weeks of effective antimicrobial therapy, respectively—is unrelated to change in vegetation size. Only 3% of strokes occur after 1 week of effective therapy. Emboli occurring late during or after effective therapy do not in themselves constitute evidence of failed antimicrobial treatment.

FIGURE 123-2

A. Janeway lesions on the toe (*left*) and plantar surface (*right*) of the foot in subacute *Neisseria mucosa* endocarditis. *(Images courtesy of Rachel Baden, MD.)* ***B.*** Septic emboli with hemorrhage and infarction due to acute *Staphylococcus aureus* endocarditis.



A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Other neurologic complications include aseptic or purulent meningitis, intracranial hemorrhage due to hemorrhagic infarcts or ruptured mycotic aneurysms, and seizures. (*Mycotic aneurysms* are focal dilations of arteries occurring at points in the artery wall that have been weakened by infection in the vasa vasorum or where septic emboli have lodged.) Microabscesses in brain and meninges occur commonly in *S. aureus* endocarditis; surgically drainable intracerebral abscesses are infrequent.

Immune complex deposition on the glomerular basement membrane causes diffuse hypocomplementemic glomerulonephritis and renal dysfunction, which typically improve with effective antimicrobial therapy. Embolic renal infarcts cause flank pain and hematuria but rarely cause renal dysfunction.

Manifestations of Specific Predisposing Conditions

Almost 50% of endocarditis associated with injection drug use is limited to the tricuspid valve and presents with fever but with faint or no murmur and no peripheral manifestations. Septic pulmonary emboli, which are common with tricuspid endocarditis, cause cough, pleuritic chest pain, nodular pulmonary infiltrates, and occasionally empyema or pyopneumothorax. Infection of the aortic or mitral valves presents with the typical clinical features of endocarditis, including peripheral manifestations.

If not associated with an intracardiac device or masked by the symptoms of concurrent comorbid illness, health care–associated endocarditis has typical manifestations. CIED endocarditis may be associated with obvious (especially within 6 months of device

manipulation) or cryptic generator pocket infection and results in fever, minimal murmur, and pulmonary symptoms due to septic emboli. Late-onset PVE presents with typical clinical features. In early PVE, typical symptoms may be obscured by comorbidity associated with recent surgery. In both early and late PVE, perivalvular infection is common and often results in partial valve dehiscence, regurgitant murmurs, CHF, or disruption of the conduction system.

DIAGNOSIS

Careful clinical, microbiologic, and echocardiographic evaluations should be pursued when febrile patients have endocarditis predispositions, cardiac or noncardiac (e.g., stroke or splenic infarct) features of endocarditis, or blood cultures yielding an endocarditis-associated organism.

Duke Criteria

The diagnosis of infective endocarditis is established with certainty only when vegetations are examined histologically and microbiologically. Nevertheless, a highly sensitive and specific diagnostic schema—known as the *modified Duke criteria*—is based on clinical, laboratory, and echocardiographic findings commonly encountered in patients with endocarditis **(Table 123-3)**. Although developed as a research tool, the criteria can help with diagnosis if the appropriate data are collected. Nevertheless, clinical judgment must be exercised in order to use the criteria effectively. Documentation of two major criteria, of one major criterion and three minor criteria, or of five minor criteria allows a clinical diagnosis of definite endocarditis. The diagnosis of endocarditis is rejected if an alternative diagnosis is established, if symptoms resolve and do not recur with ≤4 days of antibiotic therapy, or if surgery or autopsy after ≤4 days of antimicrobial therapy yields no histologic evidence of endocarditis. Illnesses not classified as definite endocarditis or rejected as such are considered cases of possible infective endocarditis when either one major and one minor criterion or three minor criteria are fulfilled. Requiring some clinical features of endocarditis for classification as possible infective endocarditis increases the specificity of the schema without significantly reducing its sensitivity. Unless there are extenuating circumstances, patients with definite or possible endocarditis are treated as having endocarditis.

TABLE 123-3

The Modified Duke Criteria for the Clinical Diagnosis of Infective Endocarditis[a]

Major Criteria
1. Positive blood culture
Typical microorganism for infective endocarditis from two separate blood cultures
Viridans streptococci, *Streptococcus gallolyticus*, HACEK group organisms, *Staphylococcus aureus*, *or*
Community-acquired enterococci in the absence of a primary focus,
or
Persistently positive blood culture, defined as recovery of a microorganism consistent with infective endocarditis from:
Blood cultures drawn >12 h apart; *or*
All of 3 or a majority of ≥4 separate blood cultures, with first and last drawn at least 1 h apart
or
Single positive blood culture for *Coxiella burnetii* or phase I IgG antibody titer of >1:800
2. Evidence of endocardial involvement
Positive echocardiogram[b]
Oscillating intracardiac mass on valve or supporting structures or in the path of regurgitant jets or in implanted material, in the absence of an alternative anatomic explanation, *or*
Abscess, *or*
New partial dehiscence of prosthetic valve,
or
New valvular regurgitation (increase or change in preexisting murmur not sufficient)
Minor Criteria
1. Predisposition: predisposing heart conditions[c] or injection drug use
2. Fever ≥38.0°C (≥100.4°F)
3. Vascular phenomena: major arterial emboli, septic pulmonary infarcts, mycotic aneurysm, intracranial hemorrhage, conjunctival hemorrhages, Janeway lesions
4. Immunologic phenomena: glomerulonephritis, Osler's nodes, Roth's spots, rheumatoid factor

5. Microbiologic evidence: positive blood culture but not meeting major criterion, as noted previously,[d] *or* serologic evidence of active infection with an organism consistent with infective endocarditis

[a]Definite endocarditis is defined by documentation of two major criteria, of one major criterion and three minor criteria, or of five minor criteria. See text for further details. [b]Transesophageal echocardiography is required for optimal assessment of possible prosthetic valve endocarditis or complicated endocarditis. [c]Valvular disease with stenosis or regurgitation, presence of a prosthetic valve, congenital heart disease including corrected or partially corrected conditions (except isolated atrial septal defect, repaired ventricular septal defect, or closed patent ductus arteriosus), prior endocarditis, or hypertrophic cardiomyopathy. [d]Excluding single positive cultures for coagulase-negative staphylococci and diphtheroids, which are common culture contaminants, or for organisms that do not cause endocarditis frequently, such as gram-negative bacilli.

Source: Adapted from JS Li et al: Clin Infect Dis 30:633, 2000. With permission from Oxford University Press.

The modified Duke criteria emphasize bacteremia and echocardiographic findings typical of endocarditis. The requirement for multiple positive blood cultures over time is consistent with the continuous low-density bacteremia characteristic of endocarditis. The diagnostic criteria attach significance to the species of organism isolated from blood cultures. To fulfill a major criterion, an organism that causes both endocarditis and non-endocarditis-related bacteremia (e.g., *S. aureus*, enterococci) must be recovered in multiple blood cultures (i.e., persistent bacteremia) and in the absence of an extracardiac focus of infection. Organisms that rarely cause endocarditis but commonly contaminate blood cultures (e.g., diphtheroids, CoNS) must be found in repeated blood cultures if they are to satisfy a major criterion.

Blood Cultures

Isolation of the causative microorganism from blood cultures is critical for diagnosis and for treatment planning. In patients with suspected NVE, PVE, or CIED endocarditis who have not received antibiotics during the prior 2 weeks, three 2-bottle blood culture sets, separated from one another by at least 2 h, should be obtained from different venipuncture sites over 24 h. If the cultures remain negative after 48–72 h, two or three additional blood culture sets should be obtained, and the laboratory should be consulted for advice regarding optimal culture techniques. Pending culture results, empirical antimicrobial therapy should be withheld initially from hemodynamically stable patients with suspected subacute endocarditis, especially those who have received antibiotics within the preceding 2 weeks. The delay allows blood for additional cultures to be obtained without the confounding effect of empirical treatment. Patients with acute endocarditis or with deteriorating hemodynamics who may require urgent surgery should receive empirical treatment immediately after three sets of blood cultures are obtained over several hours.

Non-Blood-Culture Tests

Serologic tests can be used to implicate organisms that are difficult to recover by blood culture: *Brucella*, *Bartonella*, *Legionella*, *Chlamydia psittaci*, and *C. burnetii*. In vegetations recovered at surgery or by embolectomy, pathogens can also be identified by culture; by microscopic examination with special stains; and by polymerase chain reaction (PCR) recovery of microbial DNA or DNA encoding the 16S or 28S ribosomal unit (16S rRNA or 28S rRNA), which when sequenced allows identification of bacteria and fungi, respectively.

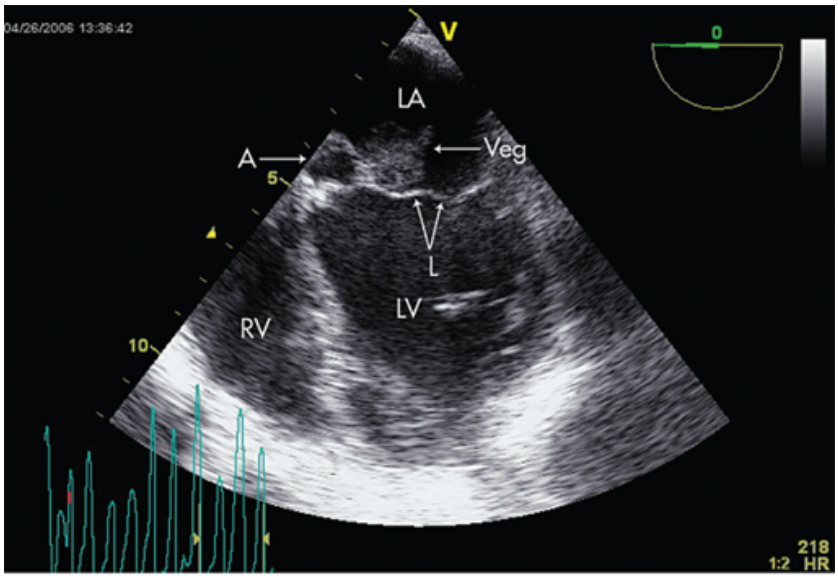
Cardiac Imaging

Echocardiography anatomically confirms and measures vegetations, detects intracardiac complications, and assesses cardiac function (Fig. 123-3). Transthoracic echocardiography (TTE) is noninvasive and exceptionally specific; however, it cannot image vegetations <2 mm in diameter, and in 20% of patients the images are inadequate. TTE detects vegetations in 65–80% of patients with definite clinical endocarditis but is not optimal for evaluating prosthetic valves or detecting intracardiac complications. TEE is safe and detects vegetations in >90% of patients with definite endocarditis; nevertheless, initial studies may yield false-negative results in 6–18% of endocarditis patients. When endocarditis is likely, a negative TEE result does not exclude the diagnosis but rather warrants repeating the study in 7–10 days. TEE is the optimal method for the diagnosis of PVE and CIED endocarditis as well as for the detection of myocardial abscess, valve perforations, or intracardiac fistulae. In patients with a CIED and a low likelihood of intracardiac infection, a

mass adherent to the lead may be a bland thrombosis rather than an infected vegetation. Three-dimensional TEE may occasionally augment the findings of standard TEE. In addition, 18-fluorodeoxyglucose positron emission tomography (FDG-PET)/CT, a technique still under evaluation, may identify perivalvular or perigraft infection not seen on TEE in patients with PVE or prosthesis–aorta graft infection (Bental procedure).

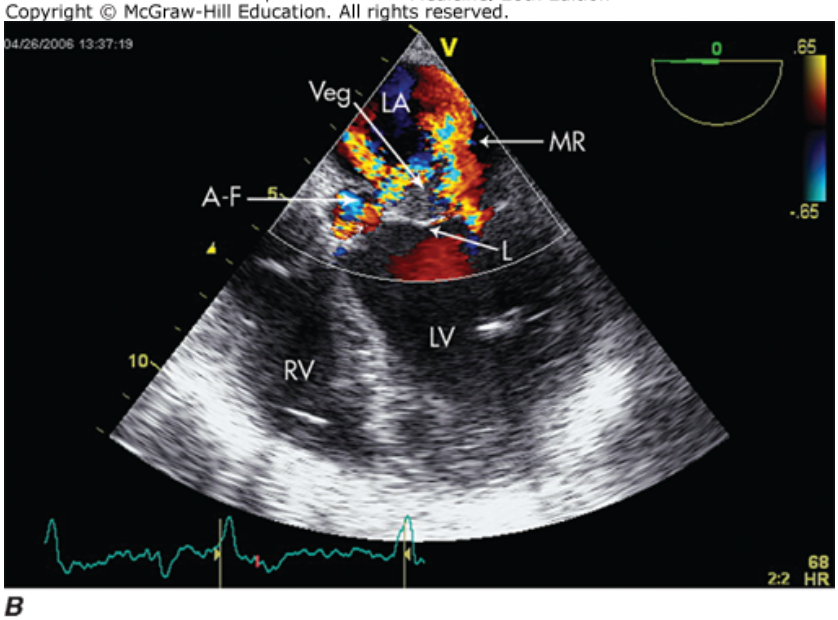
FIGURE 123-3

Imaging of a mitral valve infected with *Staphylococcus aureus* by low-esophageal, four-chamber view, transesophageal echocardiography (TEE). *A.* Two-dimensional echocardiogram showing a large vegetation with an adjacent echolucent abscess cavity. ***B.*** Color-flow Doppler image showing severe mitral regurgitation through both the abscess-fistula and the central valve orifice. A, abscess; A-F, abscess-fistula; L, valve leaflets; LA, left atrium; LV, left ventricle; MR, mitral central valve regurgitation; RV, right ventricle; veg, vegetation. *(With permission of Andrew Burger, MD.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

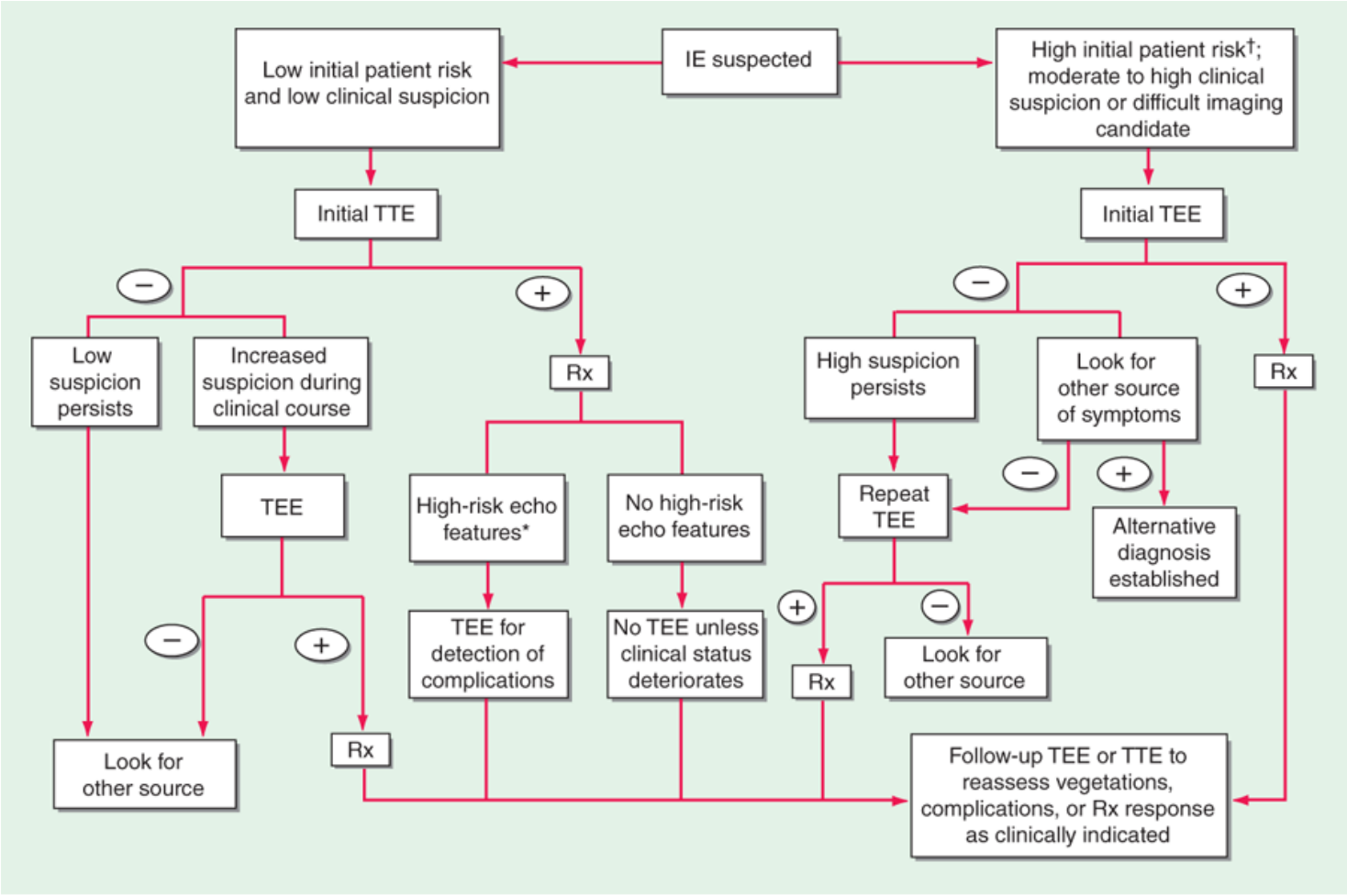
Because *S. aureus* bacteremia is associated with a high prevalence of endocarditis, routine echocardiographic evaluation (TTE or preferably TEE) is recommended in these patients. Patients with nosocomial *S. aureus* bacteremia are at increased risk of endocarditis if one or more of the following are present: positive blood cultures for 2–4 days, hemodialysis dependency, a permanent intracardiac

device, spine infection, nonvertebral osteomyelitis, or an endocarditis-predisposing valve abnormality. Ideally, these patients should be evaluated with TEE. In patients with none of these findings, the risk of endocarditis is low and evaluation with TTE may suffice.

Experts favor echocardiographic evaluation of all patients with a clinical diagnosis of endocarditis; however, the test should not be used to screen patients with a low probability of endocarditis. An approach to echocardiographic evaluation of patients with suspected endocarditis is illustrated in **Fig. 123-4**.

FIGURE 123-4

The diagnostic use of transesophageal and transtracheal echocardiography (TEE and TTE, respectively). †High initial patient risk for infective endocarditis (IE), as listed in Table 123-8, or evidence of intracardiac complications (new regurgitant murmur, new electrocardiographic conduction changes, or congestive heart failure). *High-risk echocardiographic features include large vegetations, valve insufficiency, paravalvular infection, or ventricular dysfunction. Rx indicates initiation of antibiotic therapy. *(Reproduced with permission from Diagnosis and Management of Infective Endocarditis and Its Complications. Circulation 98:2936, 1998. © 1998 American Heart Association.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Other Studies

Many studies that are not diagnostic—i.e., complete blood count, creatinine determination, liver function tests, chest radiography, and electrocardiography—are important in the management of patients with endocarditis. The erythrocyte sedimentation rate, C-reactive protein level, rheumatoid factor, and circulating immune complex titer are commonly increased in endocarditis (Table 123-2). Cardiac catheterization is used to assess coronary artery patency in older individuals who are to undergo surgery for endocarditis.

TREATMENT

TREATMENT

Infective Endocarditis

ANTIMICROBIAL THERAPY

To cure endocarditis, all bacteria in the vegetation must be killed. However, it is difficult to eradicate these bacteria because local host defenses are deficient and because the bacteria are largely nongrowing and metabolically inactive and thus are less easily killed by antibiotics. Accordingly, therapy must be bactericidal and prolonged. Antibiotics are generally given parenterally to achieve serum concentrations that, through passive diffusion, result in effective concentrations in the depths of the vegetation. To select effective therapy requires knowledge of the susceptibility of the causative microorganisms. The decision to initiate treatment empirically must balance the need to establish a microbiologic diagnosis against the potential progression of disease or the need for urgent surgery (see "Blood Cultures," above). Simultaneous infection at other sites (such as the meninges), allergies, end-organ dysfunction, interactions with concomitantly administered medications, and risks of adverse events must be considered in the selection of therapy.

Although given for several weeks longer, the regimens recommended for the treatment of PVE (except that caused by staphylococci) are similar to those used to treat NVE **(Table 123-4)**. Recommended doses and durations of therapy should be followed unless alterations are required by end-organ dysfunction or adverse events. The duration of therapy is measured from the time blood cultures become negative.

Organism-Specific Therapies

Streptococci

The recommended therapies for streptococcal endocarditis are based on the minimal inhibitory concentration (MIC) of penicillin for the causative isolate (Table 123-4). The 2-week penicillin/gentamicin and ceftriaxone/gentamicin regimens should not be used to treat PVE or NVE complicated by cardiac or extracardiac abscess. Caution should be exercised in considering aminoglycoside-containing regimens for the treatment of patients at increased risk for aminoglycoside toxicity (renal or eighth cranial nerve). The regimens recommended for relatively penicillin-resistant streptococci are advocated for treatment of group B, C, or G streptococcal endocarditis. *Granulicatella* or *Abiotrophia* species (nutritionally variant streptococci) and *Gemella* species are treated with the regimens for moderately penicillin-resistant streptococci, as is PVE caused by these organisms or by streptococci with a penicillin MIC of >0.1 μg/mL (Table 123-4).

Enterococci

Enterococci are resistant to oxacillin, nafcillin, and the cephalosporins and are only inhibited—not killed—by penicillin, ampicillin, teicoplanin (not available in the United States), and vancomycin. Enterococci are killed by the synergistic interaction of a cell wall–active antibiotic that is effective at achievable serum concentrations (penicillin, ampicillin, vancomycin, or teicoplanin) combined with an aminoglycoside (gentamicin or streptomycin) to which the isolate does not exhibit high-level resistance. An isolate's resistance to cell wall–active agents or its ability to replicate in the presence of gentamicin at ≥500 μg/mL or streptomycin at 1000–2000 μg/mL—a phenomenon called *high-level aminoglycoside resistance*—indicates that the ineffective antimicrobial agent cannot participate in the interaction to produce killing. High-level resistance to gentamicin predicts that tobramycin, netilmicin, amikacin, and kanamycin will be ineffective also. In fact, even when enterococci are not highly resistant to gentamicin, it is difficult to predict that aminoglycosides other than gentamicin and streptomycin will participate in synergistic killing; consequently, only these two aminoglycosides should be considered for synergy in treating enterococcal endocarditis. High concentrations of ampicillin plus ceftriaxone or cefotaxime, by expanded binding of penicillin-binding proteins, also kill *E. faecalis* in vitro and in animal models of endocarditis.

Enterococci must be tested for high-level resistance to streptomycin and gentamicin, β-lactamase production, and susceptibility to penicillin and ampicillin (MIC, <8 μg/mL) and to vancomycin (MIC, ≤4 μg/mL) and teicoplanin (MIC, ≤2 μg/ml). If the isolate produces β-lactamase, ampicillin/sulbactam or vancomycin can be used as the cell wall–active component; if the penicillin/ampicillin MIC is ≥8 μg/mL, vancomycin can be considered; and if the vancomycin MIC is ≥8 μg/mL, penicillin or ampicillin can be considered. In the absence of high-level resistance, gentamicin or streptomycin should be used as the aminoglycoside (Table 123-4). Although the dose of gentamicin used to achieve bactericidal synergy in treating enterococcal endocarditis is smaller than that used in standard therapy, nephrotoxicity (or vestibular toxicity with streptomycin) is not uncommon during treatment lasting 4–6 weeks. Regimens in which the

gentamicin component is given for only 2–3 weeks have been curative and associated with less nephrotoxicity than those using longer courses. Thus some experts prefer regimens wherein gentamicin is administered for only 2–3 weeks.

If there is high-level resistance to both gentamicin and streptomycin, a synergistic bactericidal effect cannot be achieved with an aminoglycoside; thus an aminoglycoside should not be given. Instead, an 8- to 12-week course of a single cell wall–active agent can be considered; however, high doses of ampicillin combined with ceftriaxone or cefotaxime have been suggested for *E. faecalis* endocarditis (Table 123-4). Nonrandomized comparative studies suggest that ampicillin-ceftriaxone may be as effective as (and less nephrotoxic than) penicillin or ampicillin plus an aminoglycoside in the treatment of *E. faecalis* endocarditis and may provide effective treatment when strains possess high-level resistance to gentamicin and streptomycin. This regimen may also be preferred in patients who are at increased risk for aminoglycoside nephrotoxicity or in lieu of streptomycin.

If the enterococcal isolate is resistant to all of the commonly used agents, suppression of bacteremia followed by surgical treatment should be considered. The role of agents potentially active against multidrug-resistant enterococci (quinupristin/dalfopristin [*E. faecium* only], linezolid, and daptomycin) in the treatment of endocarditis has not been established.

Staphylococci

The regimens used to treat staphylococcal endocarditis (Table 123-4) are based not on coagulase production but rather on the presence or absence of a prosthetic valve or foreign device, the native valve(s) involved (right vs left side), and the susceptibility of the isolate to penicillin, methicillin, and vancomycin. All staphylococci are considered potentially penicillin resistant and, except in specific countries, methicillin resistant. Thus empirical therapy for possible staphylococcal NVE should use a regimen that covers methicillin-resistant organisms. Therapy should be revised to a β-lactam agent if the isolate is susceptible to methicillin. The addition of 3–5 days of gentamicin to a β-lactam antibiotic or vancomycin to enhance therapy for left-sided NVE has not improved survival rates and is associated with nephrotoxicity. Most guidelines do not recommend the routine addition of gentamicin, fusidic acid, or rifampin to regimens for *S. aureus* NVE.

For treatment of NVE due to *methicillin-resistant S. aureus* (MRSA), vancomycin, dosed to achieve trough concentrations of 15–20 μg/mL, is recommended, with the caveat that this regimen may be associated with nephrotoxicity. Although resistance to vancomycin among staphylococci is rare, reduced vancomycin susceptibility among MRSA strains is increasingly encountered. Isolates with a vancomycin MIC of 4–16 μg/mL have intermediate susceptibility and are referred to as *vancomycin-intermediate S. aureus* (VISA). Isolates with a vancomycin MIC of 2 μg/mL may harbor subpopulations with higher MICs. Isolates with these subpopulations, called *heteroresistant VISA* (hVISA), are not detectable by routine susceptibility testing and yet may impair vancomycin efficacy. Because of the pharmacokinetics/pharmacodynamics of vancomycin, killing of MRSA with a vancomycin MIC of >1.0 μg/mL is unpredictable, even with aggressive vancomycin dosing. As an alternative to vancomycin, daptomycin (8–10 mg/kg IV once daily) has provided effective treatment for left-sided NVE caused by documented daptomycin-susceptible VISA, hVISA, or isolates with a vancomycin MIC of >1.0 μg/mL (not approved by the U.S. Food and Drug Administration for this indication). Daptomycin activity against MRSA—even against some isolates with reduced daptomycin susceptibility—is enhanced in combination with nafcillin or ceftaroline. Case series suggest that either high-dose daptomycin combined with nafcillin or ceftaroline or ceftaroline alone (600 mg IV q8h) may be effective treatment for vancomycin-unresponsive MRSA endocarditis. Infectious disease consultation is recommended for treatment of MRSA endocarditis when bacteremia persists despite therapy. The efficacy of linezolid or telavancin for left-sided MRSA endocarditis has not been established. Although it is not advocated by other groups, the British Society for Antimicrobial Chemotherapy recommends the addition of a second drug to vancomycin (rifampin) or to daptomycin (rifampin, gentamicin, or linezolid) for the treatment of MRSA NVE.

Methicillin-susceptible *S. aureus* endocarditis that is uncomplicated and limited to the tricuspid or pulmonic valve can often be treated with a 2-week course that combines oxacillin or nafcillin (but not vancomycin) with gentamicin. However, patients with prolonged fever (≥5 days) during therapy or multiple septic pulmonary emboli should receive standard-duration therapy. Vancomycin plus gentamicin for 2 weeks for right-sided endocarditis caused by MRSA yields suboptimal results; thus this entity is treated for at least 4 weeks with vancomycin or daptomycin (6 mg/kg as a single daily dose).

Staphylococcal PVE is treated for 6–8 weeks with a multidrug regimen (Table 123-4). Rifampin is an essential component because it kills staphylococci that are adherent to foreign material in a biofilm. Two other agents (selected on the basis of susceptibility testing) are combined with rifampin to prevent in vivo emergence of rifampin resistance. Because many staphylococci (particularly MRSA and

Staphylococcus epidermidis causing PVE) are resistant to gentamicin, the isolate's susceptibility to gentamicin or an alternative agent should be established before rifampin treatment is begun. Possible alternatives for gentamicin include another aminoglycoside, a fluoroquinolone (chosen on the basis of susceptibility), ceftaroline, or another active agent.

Other Organisms

In the absence of meningitis, endocarditis caused by *Streptococcus pneumoniae* isolates with a penicillin MIC of ≤4 μg/mL can be treated with IV penicillin (4 million units every 4 h), ceftriaxone (2 g/d as a single dose), cefotaxime (at a comparable dose), or vancomycin. Ceftriaxone or vancomycin is preferred for pneumococcal strains with a penicillin MIC of ≥2 μg/mL. If meningitis is suspected, treatment with vancomycin plus ceftriaxone—at the doses advised for meningitis—should be initiated until susceptibility results are known. Definitive therapy should then be selected on the basis of meningitis breakpoints (penicillin MIC, 0.06 μg/mL; or ceftriaxone MIC, 0.5 μg/mL). Pneumococcal NVE is treated for 4 weeks and pneumococcal PVE for 6 weeks. *P. aeruginosa* endocarditis is treated with an antipseudomonal β-lactam (piperacillin or a cephalosporin) and high doses of tobramycin (8 mg/kg per day in three divided doses). Endocarditis caused by Enterobacteriaceae is treated with a potent β-lactam antibiotic plus an aminoglycoside. Corynebacterial endocarditis is treated with penicillin plus an aminoglycoside (if the organism is susceptible to the aminoglycoside) or with vancomycin, which is highly bactericidal for most strains. Therapy for *Candida* endocarditis consists of a lipid formulation of amphotericin B (3–5 mg/kg IV qd) plus flucytosine (25 mg/kg PO q6h) or a high-dose echinocandin (caspofungin or micafungin, 150 mg IV qd; or anidulafungin, 200 mg IV qd). Early surgery is advised, as is long-term (if not indefinite) suppression with an oral azole.

Empirical Therapy and Treatment for Culture-Negative Endocarditis

In designing therapy to be administered before culture results are known or when cultures are truly negative, clinical clues to etiology (e.g., acute vs. subacute presentation, NVE, early or late PVE, the patient's predispositions) as well as epidemiologic clues (region of residence, animal exposure) must be considered. Thus empirical therapy for acute endocarditis in an injection drug user or for health care–associated NVE should cover MRSA and potentially antibiotic-resistant gram-negative bacilli. Treatment with vancomycin plus gentamicin or cefepime, initiated immediately after blood cultures are obtained, covers these organisms as well as many other potential causes. For empirical treatment of NVE with a subacute presentation, vancomycin plus ceftriaxone is reasonable. For blood culture–pending PVE, vancomycin, gentamicin, and cefepime should be used if the prosthetic valve has been in place for ≤1 year. Empirical therapy for infected prosthetic valves in place for >1 year is similar to that for culture-negative NVE. Therapy is revised once a pathogen has been identified.

In the treatment of blood culture–negative episodes, marantic endocarditis and the antiphospholipid antibody syndrome must be considered. Fastidious organisms should be investigated by serologic testing. In the absence of prior antibiotic therapy, it is unlikely that infection due to *S. aureus*, CoNS, enterococci, or Enterobacteriaceae will present with negative blood cultures; thus, in this situation, recommended empirical therapy targets not these organisms but rather fastidious streptococci, nutritionally variant organisms, the HACEK group, and *Bartonella* species. Pending the availability of diagnostic data, blood culture–negative subacute NVE is treated with vancomycin plus ampicillin-sulbactam (12 g every 24 h) or ceftriaxone; doxycycline (100 mg twice daily) is added for enhanced *Bartonella* coverage. If cultures are negative because of prior antibiotic administration, pathogens that are likely to be inhibited by the specific prior therapy should be considered.

CIED Endocarditis

Antimicrobial therapy for CIED endocarditis (as well as for generator pocket and lead infection) is adjunctive to complete removal of the device. The antimicrobial selected is based on the causative organism and should be used as recommended for NVE (Table 123-4). Bacteremic CIED infection may be complicated by coincident left-sided NVE, PVE, or remote-site infection (e.g., osteomyelitis) and may require modification of antimicrobial therapy. A 4- to 6-week course of endocarditis-targeted therapy is recommended for patients with CIED endocarditis and for those with bacteremia that continues during ongoing antimicrobial therapy after device removal. Generator pocket infection without bacteremia is treated with a 10- to 14-day course, some of which can be given orally. In the absence of another source, *S. aureus* bacteremia (and persistent CoNS bacteremia) in patients with a CIED is likely to be indicative of CIED or valvular endocarditis and should be managed accordingly. However, not all bloodstream infections in these patients indicate endocarditis. If evidence suggesting endocarditis is lacking, bloodstream infection due to gram-negative bacilli, streptococci, enterococci, and *Candida* species may not indicate endocarditis and can be treated with an abbreviated course of antimicrobial therapy. However, in the absence of another source, bacteremia relapse after antimicrobial therapy increases the likelihood of CIED

endocarditis and warrants treatment as such. Attempted salvage of an infected CIED with antibiotics alone is usually unsuccessful and should be reserved for patients whose devices cannot be removed or who refuse removal.

Outpatient Antimicrobial Therapy

Fully compliant, clinically stable patients who are no longer bacteremic, are not febrile, and have no clinical or echocardiographic findings that suggest an impending complication may complete therapy as outpatients. Careful follow-up and a stable home setting are necessary, as are predictable IV access and use of antimicrobial agents that are stable in solution. Recommended regimens should not be compromised to accommodate outpatient therapy.

Monitoring Antimicrobial Therapy

Antibiotic toxicity, including allergic reactions, occurs in 25–40% of endocarditis patients and commonly arises after several weeks of therapy. Blood tests to detect renal, hepatic, and hematologic toxicity should be performed periodically. Serum concentrations of aminoglycosides and vancomycin should be monitored periodically and doses adjusted to optimize treatment and to avoid or address toxicity.

Blood cultures should be repeated daily until sterile in patients with endocarditis due to *S. aureus* or difficult-to-treat organisms, rechecked if there is recrudescent fever, and performed again 4–6 weeks after therapy to document cure. Blood cultures become sterile within 2 days after the start of appropriate therapy when infection is caused by viridans streptococci, enterococci, or HACEK organisms. In *S. aureus* endocarditis, β-lactam therapy results in sterile cultures in 3–5 days, whereas in MRSA endocarditis, positive cultures may persist for 7–9 days with vancomycin or daptomycin treatment. MRSA bacteremia persisting despite an adequate dosage of vancomycin or daptomycin may indicate emergence of reduced susceptibility in the infecting strain and point to a need for alternative therapy. When fever persists for 7 days despite appropriate antibiotic therapy, patients should be evaluated for paravalvular abscess, extracardiac abscesses (spleen, kidney), or complications (embolic events). Recrudescent fever raises the possibility of these complications but also of drug reactions or complications of hospitalization. Vegetations become smaller with effective therapy; however, 3 months after cure, 50% are unchanged and 25% are slightly larger or smaller.

Antithrombotic Therapy

A decision to initiate antithrombotic (anticoagulant or antiplatelet) therapy in patients with infective endocarditis requires careful consideration of the risks and benefits, including temporal considerations of each. Patients with infective endocarditis are at risk for emboli, for hemorrhagic transformation of embolic strokes, and for intracerebral hemorrhage from septic arteritis or ruptured mycotic aneurysms. Antithrombotic therapy can render this bleeding catastrophic. Neither anticoagulant nor antiplatelet therapy reduces the risk of emboli in patients with NVE, and thus such treatment is not indicated for that purpose. However, patients with infective endocarditis may have coexisting conditions wherein anticoagulation is indicated. Thus, in the absence of a contraindication (i.e., no clinical or imaging evidence of a recent large embolic stroke, intracerebral hemorrhage, or mycotic aneurysm), anticoagulant therapy is given to patients who have a mechanical prosthetic valve, atrial fibrillation with either mitral stenosis or a CHADS2 score ≥2, or deep-vein thrombophlebitis. Most experts prefer to use unfractionated or low-molecular-weight heparin for ease of reversal. Anticoagulant therapy should be reversed, at least temporarily, in most patients who have had an acute ischemic stroke or an intracerebral hemorrhage.

SURGICAL TREATMENT

Intracardiac and central nervous system complications of endocarditis are important causes of morbidity and death. In some cases, effective treatment for these complications requires surgery. The indications for cardiac surgical treatment of endocarditis (Table 123-5) have been derived from observational studies and expert opinion. The strength of individual indications varies; thus the risks and benefits as well as the timing of surgery must be individualized (Table 123-6). These complex considerations are best weighed by a team that includes cardiologists, cardiac surgeons, infectious disease physicians, and neurologists if there have been neurologic complications. From 25 to 40% of patients with left-sided endocarditis undergo cardiac surgery during active infection, with slightly higher surgery rates for PVE than NVE. Intracardiac complications and CHF are the most commonly cited indications for surgery. The benefit of surgery has been assessed primarily in studies comparing populations of medically and surgically treated patients matched for the necessity of surgery, with adjustments for predictors of death (comorbidities) and the timing of surgical intervention (a correction for survival bias). Although study results vary, surgery for NVE based on current indications appears to convey a significant survival benefit (27–55%) that becomes increasingly apparent among patients with the most pressing indications and with follow-up

for ≥6 months. The effect of surgery for PVE is more nuanced, with survival benefits accruing largely to those with intracardiac complications. Of note, surgery itself carries mortality risks that may offset survival benefits in patients with lesser indications.

Indications

Congestive Heart Failure

Moderate to severe refractory CHF caused by new or worsening valve dysfunction or intracardiac fistulae is the major indication for cardiac surgery. Surgery can relieve functional stenosis due to large vegetations or restore competence to damaged regurgitant valves by repair or replacement. At 6–12 months of follow-up, the mortality rate is 50% among patients with left-sided NVE or PVE and moderate to severe heart failure due to valve dysfunction who are treated medically, while that among matched patients treated surgically is 15%. The survival benefit with surgery is inversely related to the severity of preoperative CHF; thus surgery should not be delayed in the face of deteriorating hemodynamics.

Perivalvular Infection

This complication, which is most common with aortic valve infection, occurs in 10–15% of patients with NVE and in 45–60% of those with PVE. It is suggested clinically by persistent unexplained fever during appropriate therapy, new electrocardiographic conduction disturbances, or pericarditis. TEE with color Doppler is the test of choice to detect perivalvular abscesses (sensitivity, ≥85%). Occasionally, three-dimensional TEE and FDG-PET/CT demonstrate perivalvular infection not detected by TEE. For optimal outcome, surgery is required, especially when fever persists, fistulae develop, prostheses are dehisced and unstable, or infection relapses after appropriate treatment. Cardiac rhythm must be monitored since high-grade heart block may require insertion of a pacemaker.

Uncontrolled Infection

Continued positive blood cultures or otherwise unexplained persistent fevers despite optimal antibiotic therapy may reflect uncontrolled infection that warrants surgery. Surgical treatment is also advised for endocarditis caused by organisms against which effective antimicrobial therapy is lacking (e.g., yeasts, fungi, *P. aeruginosa*, other highly antibiotic-resistant bacteria, *Brucella* species).

S. aureus Endocarditis

The mortality rate for *S. aureus* PVE exceeds 50% with medical treatment but is reduced to 25% with surgical treatment. When patients have intracardiac complications, surgical treatment reduces the mortality rate twentyfold. However, surgery is not routinely advised for uncomplicated *S. aureus* PVE. Surgical treatment should be considered for patients with MRSA left-sided NVE who remain septic and unresponsive to alternative antibiotics. Isolated tricuspid-valve *S. aureus* endocarditis, even with persistent fever, rarely requires surgery.

Prevention of Systemic Emboli

Persisting morbidity and/or death may result from cerebral or coronary artery emboli. Antithrombotic therapy does not prevent systemic emboli in NVE. The frequency of embolization decreases rapidly with effective antimicrobial therapy. Thus, to further reduce emboli through cardiac surgery, the surgery must be performed very early. Predicting a high risk of systemic embolization by echocardiographic determination of vegetation size and anatomy does not identify those patients in whom surgery to prevent emboli will result in increased survival. In a small randomized trial in patients who were at low risk of surgery-related mortality and had large vegetations (>10 mm) and significant valve dysfunction, emboli were prevented by early surgery (≤48 h after diagnosis), but there was no survival benefit. Rarely is the indication for surgery solely to prevent systemic emboli; however, this goal may be an additional benefit of early surgery for other indications. Valve repair, with the consequent avoidance of a prosthesis, improves the benefit-to-risk ratio of surgery performed to eliminate vegetations.

Cied Endocarditis

Removal of all hardware is recommended for patients with established CIED endocarditis as well as for pocket or intracardiac lead infection. Percutaneous lead extraction is preferred; with retained hardware after attempted percutaneous extraction, surgical removal should be considered. With lead vegetations >2 cm, there is a risk of a pulmonary embolism; nevertheless, the need for surgical removal of the CIED is unclear. Removal of the infected CIED during the initial hospitalization is associated with increased 30-day and 1-year survival rates over those attained with antibiotic therapy and attempted device retention. The CIED, if needed, can be reimplanted at a new site after at least 10–14 days of effective antimicrobial therapy. CIEDs should be replaced when patients undergo surgery for endocarditis.

Timing of Cardiac Surgery

With life-threatening indications for surgery (valve dysfunction and severe CHF, perivalvular abscess, major prosthesis dehiscence), surgery during the initial days of therapy is associated with greater survival than later surgery. With less compelling indications, surgery may reasonably be delayed to allow further treatment as well as improvement in overall health (Table 123-6). Recrudescent endocarditis on a newly implanted prosthetic valve follows surgery for active NVE and PVE in 2% and 6–15% of patients, respectively. These frequencies do not justify the increased mortality risk associated with delaying surgery in patients with severe heart failure, valve dysfunction, and uncontrolled infections. Delay is justified when infection is controlled and CHF is resolved with medical therapy.

Neurologic complications of endocarditis may be exacerbated during cardiac surgery. The risk of neurologic deterioration is related to the type and severity of the preoperative neurologic complication and the interval between the complication and surgery. When the surgical indication is not urgent, cardiac surgery should be delayed for 2–3 weeks after a large nonhemorrhagic embolic infarction and for 4 weeks after a cerebral hemorrhage. A ruptured mycotic aneurysm should be treated before cardiac surgery. In a non-obtunded patient with an ischemic stroke and hemorrhage excluded by imaging, cardiac surgery, if urgent, should be performed early.

Antibiotic Therapy after Cardiac Surgery

Organisms have been detected on Gram's stain—or their DNA has been detected by PCR—in excised valves from 45% of patients who have successfully completed the recommended therapy for endocarditis. In only 7% of these patients were the organisms—most of which are unusual and antibiotic resistant—cultured from the valve. Detection of organisms or their DNA does not necessarily indicate antibiotic failure; in fact, relapse after surgery for active endocarditis is uncommon. Thus, when valve cultures are negative in uncomplicated NVE caused by susceptible organisms, the duration of preoperative plus postoperative treatment should equal the total duration of recommended therapy. For endocarditis complicated by perivalvular abscess, partially treated PVE, or culture-positive valves, a full course of therapy should be given postoperatively.

Extracardiac Complications

Splenic abscess develops in 3–5% of patients with endocarditis. Effective therapy requires either image-guided percutaneous drainage or splenectomy. Mycotic aneurysms occur in 2–15% of endocarditis patients; one-half of these cases involve the cerebral arteries and present as headaches, focal neurologic symptoms, or hemorrhage. Cerebral aneurysms should be monitored by angiography. Some will resolve with effective antimicrobial therapy, but those that persist, enlarge, or leak should be treated surgically if possible. Extracerebral aneurysms present as local pain, a mass, local ischemia, or bleeding; these aneurysms are treated surgically.

TABLE 123-4

Antibiotic Treatment for Infective Endocarditis Caused by Common Organisms[a]

Organism(s)	Drug (Dose, Duration)	Comments
Streptococci		
Penicillin-susceptible streptococci, *S. gallolyticus* (MIC ≤0.12 µg/mL[b])	• Penicillin G (2–3 mU IV q4h for 4 weeks)	Can use ampicillin or amoxicillin (2 g IV q4h) if penicillin is unavailable.
	• Ceftriaxone (2 g/d IV as a single dose for 4 weeks)	Can use ceftriaxone in patients with non-immediate penicillin allergy.
	• Vancomycin[c] (15 mg/kg IV q12h for 4 weeks)	Use vancomycin in patients with severe or immediate β-lactam allergy.
	• Penicillin G (2–3 mU IV q4h) *or* ceftriaxone (2 g IV qd) for 2 weeks ***plus*** Gentamicin[d] (3 mg/kg qd IV or IM, as a single dose[e] or divided into equal doses q8h for 2 weeks)	Avoid 2-week regimen when risk of aminoglycoside toxicity is increased and in prosthetic–valve or complicated endocarditis.
Relatively penicillin-resistant streptococci, *S. gallolyticus* (MIC >0.12 µg/mL and <0.5 µg/mL[f])	• Penicillin G (4 mU IV q4h) *or* ceftriaxone (2 g IV qd) for 4 weeks ***plus*** Gentamicin[d] (3 mg/kg qd IV or IM, as a single dose[e] or divided into equal doses q8h for 2 weeks)	Can use ampicillin or amoxicillin (2 g IV q4h) if penicillin is unavailable. Penicillin alone at this dose for 6 weeks or with gentamicin during the initial 2 weeks is preferred for prosthetic–valve endocarditis caused by streptococci with penicillin MICs of ≤0.1 µg/mL.
	• Vancomycin[c] as noted above for 4 weeks	Use vancomycin if unable to tolerate penicillins. Ceftriaxone alone or with gentamicin can be used in patients with non-immediate β-lactam allergy.

Organism(s)	Drug (Dose, Duration)	Comments
Moderately penicillin-resistant streptococci (MIC, ≥0.5 µg/mL and <8 µg/mL[g]); *Granulicatella*, *Abiotrophia*, or *Gemella* spp.	• Penicillin G (4–5 mU IV q4h) *or* ceftriaxone (2 g IV qd) for 6 weeks ***plus*** Gentamicin[d] (3 mg/kg qd IV or IM as a single dose[e] *or* divided into equal doses q8h for 6 weeks)	Preferred for PVE caused by streptococci with penicillin MICs of >0.1 µg/mL.
	• Vancomycin[c] as noted above for 4 weeks	Regimen is preferred by some.
Enterococci[h]		
	• Penicillin G (4–5 mU IV q4h) *plus* gentamicin[d] (1 mg/kg IV q8h), both for 4–6 weeks	Can treat NVE for 4 weeks if symptoms last <3 months. Treat PVE and NVE with >3 months of symptoms for 6 weeks. Can abbreviate gentamicin course in some patients (see text). Can use streptomycin (7.5 mg/kg q12h) in lieu of gentamicin if there is not high-level resistance to streptomycin
	• Ampicillin (2 g IV q4h) *plus* gentamicin[d] (1 mg/kg IV q8h), both for 4–6 weeks	Can use amoxicillin in lieu of ampicillin (same dose)
	• Vancomycin[c] (15 mg/kg IV q12h) *plus* gentamicin[d] (1 mg/kg IV q8h), both for 6 weeks	Use vancomycin plus gentamicin only for penicillin-allergic patients (preferable to desensitize to penicillin) and for isolates resistant to penicillin/ampicillin.
	• Ampicillin (2 g IV q4h) *plus* ceftriaxone (2 g IV q12h), both for 6 weeks	Use for *E. faecalis* isolates with or without high-level resistance to gentamicin and streptomycin or for patients at high risk for aminoglycoside nephrotoxicity (creatinine clearance rate <50 mL/min; see text).
Staphylococci (*S. aureus* and coagulase-negative)		
MSSA infecting native valves (no foreign devices)	• Nafcillin, oxacillin, *or* flucloxacillin (2 g IV q4h for 4–6 weeks)	Can use penicillin (4 mU q4h) if isolate is penicillin susceptible (i.e., does not produce β-lactamase); 6-week course preferred
	• Cefazolin (2 g IV q8h for 4–6 weeks)	Can use cefazolin regimen for patients with non-immediate penicillin allergy; 6-week course preferred

Organism(s)	Drug (Dose, Duration)	Comments
	• Vancomycin[c] (15 mg/kg IV q12h for 4–6 weeks)	Only use vancomycin for patients with immediate (urticarial) or severe penicillin allergy; see text regarding addition of gentamicin, fusidic acid, or rifampin. A 6-week course is preferred.
MRSA infecting native valves (no foreign devices)	• Vancomycin[c] (15 mg/kg IV q8–12h for 4–6 weeks)	No role for routine use of rifampin (see text). Consider high-dose daptomycin treatment (see text) for MRSA with vancomycin MIC >1.0 or persistent bacteremia during vancomycin therapy.
MSSA infecting prosthetic valves	• Nafcillin, oxacillin, *or* flucloxacillin (2 g IV q4h for 6–8 weeks) ***plus*** Gentamicin[d] (1 mg/kg IM or IV q8h for 2 weeks) ***plus*** • Rifampin[i] (300 mg PO q8h for 6–8 weeks)	Use gentamicin during initial 2 weeks; determine susceptibility to gentamicin before initiating rifampin (see text); if patient is highly allergic to penicillin, use regimen for MRSA; if β-lactam allergy is of the minor non-immediate type, cefazolin can be substituted for oxacillin, nafcillin, or flucloxacillin.
MRSA infecting prosthetic valves	• Vancomycin[c] (15 mg/kg IV q12h for 6–8 weeks) ***plus*** Gentamicin[d] (1 mg/kg IM or IV q8h for 2 weeks) ***plus*** Rifampin[i] (300 mg PO q8h for 6–8 weeks)	Use gentamicin during initial 2 weeks; determine gentamicin susceptibility before initiating rifampin (see text).
HACEK Organisms		
	• Ceftriaxone (2 g/d IV as a single dose for 4 weeks)	Can use another third-generation cephalosporin at comparable dose.
	• Ampicillin/sulbactam (3 g IV q6h for 4 weeks)	If the isolate is susceptible, ciprofloxacin (400 mg IV q12h) can be used.
Coxiella burnetii		
	• Doxycycline (100 mg PO q12h) *plus* hydroxychloroquine (200 mg PO q8h), both for at least 18 (native valve) or 24 (prosthetic valve) months	Follow serology to monitor response during treatment (antiphase I IgG and IgA decreased 4-fold and IgM antiphase II negative) and thereafter for relapse.
***Bartonella* spp.**		

Organism(s)	Drug (Dose, Duration)	Comments
	• Doxycycline (100 mg q12h PO) for 6 weeks ***plus*** Gentamicin (1 mg/kg IV q8h for 2 weeks)	If doxycycline is not tolerated, use azithromycin (500 mg PO qd). Some experts recommend that doxycycline be continued for 3–6 months unless all infection is resected surgically.

[a]Regimens adapted from the guidelines of the American Heart Association, the European Society of Cardiology (ESC), and to a lesser extent the British Society for Antimicrobial Chemotherapy (BSAC). Doses of gentamicin, streptomycin, and vancomycin must be adjusted for reduced renal function. Ideal body weight is used to calculate doses of gentamicin and streptomycin per kilogram (men = 50 kg + 2.3 kg per inch over 5 feet; women = 45.5 kg + 2.3 kg per inch over 5 feet). [b]MIC ≤0.125 μg/mL per ESC and BSAC. [c]Vancomycin dose is based on actual body weight. Adjust for trough level of 10–15 μg/mL for streptococcal and enterococcal infections and 15–20 μg/mL for staphylococcal infections. [d]Aminoglycosides should not be administered as single daily doses for enterococcal endocarditis and should be introduced as part of the initial treatment. Target peak and trough serum concentrations of divided-dose gentamicin 1 h after a 20- to 30-min infusion or IM injection are ~3.5 μg/mL and ≤1 μg/mL, respectively; target peak and trough serum concentrations of streptomycin (timing as with gentamicin) are 20–35 μg/mL and <10 μg/mL, respectively. [e]Netilmicin (4 mg/kg qd, as a single dose) can be used in lieu of gentamicin. [f]MIC >0.125 μg/mL and ≤2.0 μg/mL per ESC; MIC >0.125 μg/mL and ≤0.5 μg/mL per BSAC. [g]MIC >2.0 μg/mL per ESC; treat with regimen for enterococci (BSAC). [h]Antimicrobial susceptibility must be evaluated; see text. [i]Rifampin increases warfarin and dicumarol requirements for anticoagulation.

Abbreviations: MIC, minimal inhibitory concentration; MRSA, methicillin-resistant *S. aureus*; MSSA, methicillin-sensitive *S. aureus*; NVE, native-valve endocarditis; PVE, prosthetic-valve endocarditis.

TABLE 123-5

Indications for Cardiac Surgical Treatment in Patients with Endocarditis

Surgery Required for Optimal Outcome
Native-valve or prosthetic-valve endocarditis
Moderate or severe congestive heart failure or shock due to valve dysfunction
Perivalvular extension of infection with abscess, fistula, or heart block
Persistent bacteremia without an extracardiac cause despite 7–10 days of optimal antimicrobial therapy
Lack of effective antimicrobial therapy (e.g., fungal, *Brucella*, multidrug-resistant gram-negative bacillary endocarditis)
Prosthetic-valve endocarditis
Partially dehisced unstable prosthetic valve
Surgery to Be Strongly Considered for Improved Outcome[a]
Prosthetic-valve endocarditis
S. aureus infection with intracardiac complications
Relapse after optimal antimicrobial therapy
Native-valve endocarditis
Large (>10-mm) hypermobile vegetation, particularly with prior systemic embolus and significant valve dysfunction[b]
Very large (>30-mm) vegetation
Persistent unexplained fever (≥10 days) in blood culture–negative endocarditis
Poorly responsive or relapsed endocarditis due to highly antibiotic-resistant enterococci or gram-negative bacilli

[a]Carefully consider surgery. Multiple findings are often combined to justify surgery. [b]In the group with an estimated low cardiac-surgery mortality risk.

TABLE 123-6

Timing of Cardiac Surgical Intervention in Patients with Endocarditis

	Indication for Surgical Intervention	
Timing	**Strong Supporting Evidence**	**Conflicting Evidence, but Majority of Opinions Favor Surgery**
Emergent (same day)	Valve dysfunction with pulmonary edema or cardiogenic shock Acute aortic regurgitation plus preclosure of mitral valve Sinus of Valsalva abscess ruptured into right heart Rupture into pericardial sac	
Urgent (within 1–2 days)	Valve obstruction by vegetation Unstable (dehisced) prosthesis Acute aortic or mitral regurgitation with heart failure (New York Heart Association class III or IV) Septal perforation Perivalvular extension of infection with or without new electrocardiographic conduction system changes Lack of effective antibiotic therapy	Vegetation diameter >10 mm plus severe but not urgent aortic or mitral valve dysfunction[a] Major embolus plus persisting large vegetation (>10 mm) Mobile vegetation >30 mm
Elective (earlier usually preferred)	Progressive paravalvular prosthetic regurgitation Valve dysfunction plus persisting infection after ≥7–10 days of antimicrobial therapy Fungal (mold) endocarditis	Staphylococcal prosthetic-valve endocarditis with intracardiac complications Early prosthetic–valve endocarditis (≤2 months after valve surgery) *Candida* spp. endocarditis Antibiotic-resistant organisms

[a]Supported by a single-institution randomized trial showing benefit from early surgery. Implementation requires clinical judgment. If surgery is elected, it must be done early (see text).

Source: Adapted from L Olaison, G Pettersson: Infect Dis Clin North Am 16:453, 2002.

OUTCOME

Endocarditis is a heterogeneous disease that occurs in extremely heterogeneous patient populations. Many factors can adversely affect outcome; these include older age, severe comorbid conditions and diabetes, delayed diagnosis, involvement of prosthetic valves or the aortic valve, an invasive (*S. aureus*) or antibiotic-resistant (*P. aeruginosa*, yeast) pathogen, intracardiac and major neurologic complications, and an association of infection with health care. Death or poor outcome often is related not to failure of antibiotic therapy but rather to the interactions of comorbidities and endocarditis-related end-organ complications. In developed countries, overall survival rates are 80–85%; however, rates vary considerably among subpopulations of endocarditis patients. Thus predicting the outcome for a given patient must focus on that individual's infection, the complexity of required therapy, and preexisting comorbidities. Survival rates for patients with NVE caused by viridans streptococci, HACEK organisms, or enterococci (susceptible to synergistic therapy) are 85–90%. For *S. aureus* NVE in patients who do not inject drugs, survival rates are 55–70%, whereas 85–90% of injection drug users survive their initial episode of *S. aureus* endocarditis. However, if addiction is not successfully addressed in the latter group, the longer-term prognosis is guarded. PVE beginning within 2 months after valve replacement results in

mortality rates of 40–50%, whereas rates are only 10–20% in later-onset cases. Overall survival rates 1 year after successful endocarditis treatment are 80–90%.

PREVENTION

Prevention of endocarditis has long been a goal in clinical practice. However, expert committees have concluded that the evidence favoring antibiotic prophylaxis for endocarditis is insufficient to recommend this treatment as a widespread standard of care. Weighing the potential benefits, potential adverse events, and costs associated with antibiotic prophylaxis, the American Heart Association and the European Society of Cardiology now recommend prophylactic antibiotics **(Table 123-7)** only for those patients at highest risk for severe morbidity or death from endocarditis **(Table 123-8)**. Maintaining good dental hygiene in at-risk patients is essential and a recommended goal. Prophylaxis is recommended only when there is manipulation of gingival tissue or the periapical region of the teeth or perforation of the oral mucosa (including surgery on the respiratory tract). Prophylaxis is not advised for patients undergoing gastrointestinal or genitourinary tract procedures. High-risk patients should be treated before or when they undergo procedures on an infected genitourinary tract or infected skin. The National Institute for Health and Clinical Excellence in the United Kingdom has advised discontinuation of all antibiotic prophylaxis for endocarditis. Surveillance studies have attempted to assess the impact of these new prophylaxis guidelines but have been inconclusive because of inadequate data.

TABLE 123-7

Antibiotic Regimens for Prophylaxis of Endocarditis in Adults with High-Risk Cardiac Lesions[a,b]

A. Standard oral regimen
Amoxicillin: 2 g PO 1 h before procedure
B. Inability to take oral medication
Ampicillin: 2 g IV or IM within 1 h before procedure
C. Penicillin allergy
1. Clarithromycin or azithromycin: 500 mg PO 1 h before procedure
2. Cephalexin[c]: 2 g PO 1 h before procedure
3. Clindamycin: 600 mg PO 1 h before procedure
D. Penicillin allergy, inability to take oral medication
1. Cefazolin[c] or ceftriaxone[c]: 1 g IV or IM 30 min before procedure
2. Clindamycin: 600 mg IV or IM 1 h before procedure

[a]Dosing for children: for amoxicillin, ampicillin, cephalexin, or cefadroxil, use 50 mg/kg PO; cefazolin, 25 mg/kg IV; clindamycin, 20 mg/kg PO or 25 mg/kg IV; clarithromycin, 15 mg/kg PO; and vancomycin, 20 mg/kg IV. [b]For high-risk lesions, see Table 123-8. Prophylaxis is not advised for other lesions. [c]Do not use cephalosporins in patients with immediate hypersensitivity (urticaria, angioedema, anaphylaxis) to penicillin.

Source: Table created using the guidelines published by the American Heart Association and the European Society of Cardiology (W Wilson et al: Circulation 116:1736, 2007; and G Habib et al: Eur Heart J 30:2369, 2009).

TABLE 123-8

High-Risk Cardiac Lesions for Which Endocarditis Prophylaxis Is Advised Before Dental Procedures

Prosthetic heart valves
Prior endocarditis
Unrepaired cyanotic congenital heart disease, including palliative shunts or conduits
Completely repaired congenital heart defects during the 6 months after repair
Incompletely repaired congenital heart disease with residual defects adjacent to prosthetic material
Valvulopathy developing after cardiac transplantation[a]

[a]Not a target population for prophylaxis according to recommendations of the European Society for Cardiology.

Source: Table created using the guidelines published by the American Heart Association and the European Society of Cardiology (W Wilson et al: Circulation 116:1736, 2007; and G Habib et al: Eur Heart J 30:2369, 2009).

In patients with aortic or mitral valve regurgitation or a prosthetic valve, treatment of acute Q fever for 12 months with doxycycline plus hydroxychloroquine (see Table 123-4) is highly effective in preventing *C. burnetii* endocarditis.

FURTHER READING

Amat-Santos IJ et al: Infective endocarditis after transcatheter aortic valve implantation: Results from a large multicenter registry. Circulation 131:1566, 2015.

Baddour LM et al: Update on cardiovascular implantable electronic device infections and their management: A scientific statement from the American Heart Association. Circulation 121:458, 2010.

Baddour LM et al: Infective endocarditis in adults: Diagnosis, antimicrobial therapy, and management of complications: A scientific statement for healthcare professionals from the American Heart Association. Circulation 132:1435, 2015.

Chirouze C et al: Impact of early valve surgery on outcome of *Staphylococcus aureus* prosthetic valve infective endocarditis: Analysis in the International Collaboration of Endocarditis–Prospective Cohort Study. Clin Infect Dis 60:741, 2015.

Gould FK et al: Guidelines for the diagnosis and antibiotic treatment of endocarditis in adults: A report of the Working Party of the British Society for Antimicrobial Chemotherapy. J Antimicrob Chemother 67:269, 2011.

Habib G et al: 2015 ESC guidelines for the management of infective endocarditis. Eur Heart J 36:3075, 2015.

Lalani T et al: Analysis of the impact of early surgery on in-hospital mortality of native valve endocarditis: Use of propensity score and instrumental variable methods to adjust for treatment-selection bias. Circulation 121:1005, 2010.

Lalani T et al: In-hospital and 1-year mortality in patients undergoing early surgery for prosthetic valve endocarditis. JAMA Intern Med 173:1495, 2013.

Sandoe JA et al: Guidelines for the diagnosis, prevention and management of implantable cardiac electronic device infection. Report of a joint Working Party project on behalf of the British Society for Antimicrobial Chemotherapy (BSAC, host organization), British

Heart Rhythm Society (BHRS), British Cardiovascular Society (BCS), British Heart Valve Society (BHVS) and British Society for Echocardiography (BSE). J Antimicrob Chemother 70:325, 2015.

Chapter 124: Infections of the Skin, Muscles, and Soft Tissues

Dennis L. Stevens

INTRODUCTION

Skin and soft tissue infections occur in all races, all ethnic groups, and all geographic locations, although some have unique geographic niches. In modern times, the frequency and severity of some skin and soft tissue infections have increased for several reasons. First, microbes are rapidly disseminated throughout the world via efficient air travel, acquiring genes for virulence factors and antibiotic resistance. Second, natural disasters, such as earthquakes, tsunamis, tornadoes, and hurricanes, appear to be increasing in frequency, and the injuries sustained during these events commonly cause major skin and soft-tissue damage that predisposes to infection. Third, trauma and casualties resulting from combat and terrorist activities can markedly damage or destroy tissues and provide both endogenous and exogenous pathogens with ready access to deeper structures. Unfortunately, because the marvels of modern medicine may not be available during human-instigated and natural disasters, primary treatment may be delayed and the likelihood of severe infection and death increased.

ANATOMIC RELATIONSHIPS: CLUES TO THE DIAGNOSIS OF SOFT TISSUE INFECTIONS

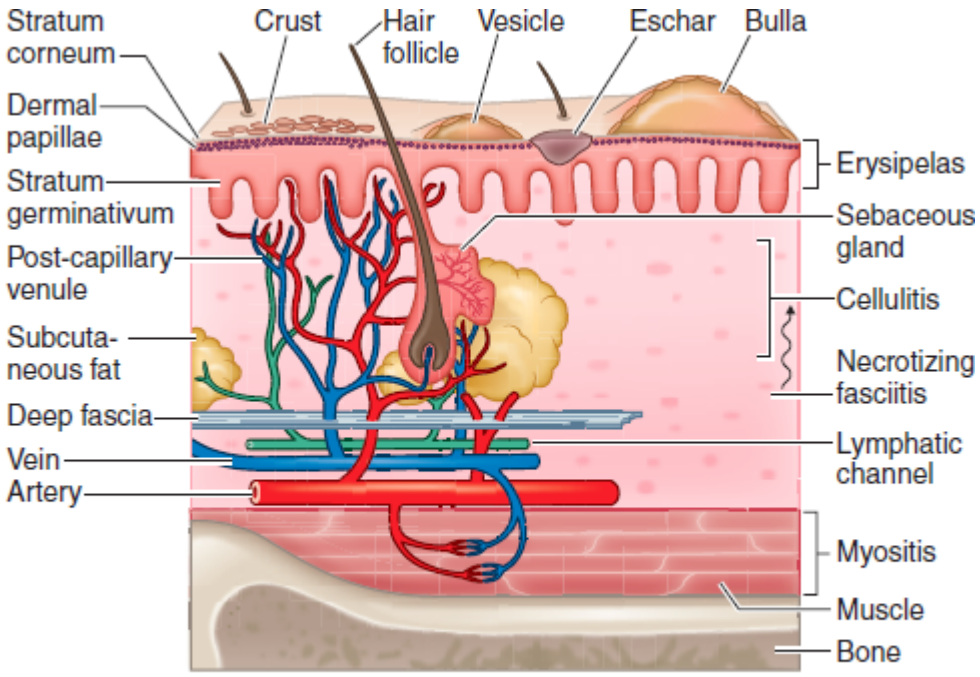
Skin and soft tissue infections have been common human afflictions for centuries. However, between 2000 and 2004, hospital admissions for these infections rose by 27%, a remarkable increase that was attributable largely to the emergence of the USA300 clone of methicillin-resistant *Staphylococcus aureus* (MRSA). This chapter provides an anatomic approach to understanding the types of soft tissue infections and the diverse microbes responsible.

Protection against infection of the epidermis depends on the mechanical barrier afforded by the stratum corneum, since the epidermis itself is devoid of blood vessels **(Fig. 124-1)**. Disruption of this layer by burns or bites, abrasions, foreign bodies, primary dermatologic disorders (e.g., herpes simplex, varicella, ecthyma gangrenosum), surgery, or vascular or pressure ulcer allows penetration of bacteria to the deeper structures. Similarly, the hair follicle can serve as a portal either for components of the normal flora (e.g., *Staphylococcus*) or for extrinsic bacteria (e.g., *Pseudomonas* in hot-tub folliculitis). Intracellular infection of the squamous epithelium with vesicle formation may arise from cutaneous inoculation, as in infection with

herpes simplex virus (HSV) type 1; from the dermal capillary plexus, as in varicella and infections due to other viruses associated with viremia; or from cutaneous nerve roots, as in herpes zoster. Bacteria infecting the epidermis, such as *Streptococcus pyogenes*, may be translocated laterally to deeper structures via lymphatics, an event that results in the rapid superficial spread of erysipelas. Later, engorgement or obstruction of lymphatics causes flaccid edema of the epidermis, another characteristic of erysipelas.

FIGURE 124-1

Structural components of the skin and soft tissues, superficial infections, and infections of the deeper structures. The rich capillary network beneath the dermal papillae plays a key role in the localization of infection and in the development of the acute inflammatory reaction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The rich plexus of capillaries beneath the dermal papillae provides nutrition to the stratum germinativum, and physiologic responses of this plexus produce important clinical signs and symptoms. For example, infective vasculitis of the plexus results in petechiae, Osler's nodes, Janeway lesions, and palpable purpura, which, if present, are important clues to the existence of endocarditis (**Chap. 123**). In addition, metastatic infection within this plexus can result in cutaneous manifestations of disseminated fungal infection (**Chap. 211**), gonococcal infection (**Chap. 151**), *Salmonella* infection (**Chap. 160**), *Pseudomonas* infection (i.e., ecthyma gangrenosum; **Chap. 159**), meningococcemia (**Chap. 150**), and staphylococcal infection (**Chap. 142**). The plexus also provides bacteria with access to the circulation, thereby facilitating local spread or bacteremia. The postcapillary venules of this plexus are a prominent site of polymorphonuclear leukocyte sequestration, diapedesis, and chemotaxis to the site of cutaneous infection.

Amplification of these physiologic mechanisms by excessive levels of cytokines or bacterial toxins causes leukostasis, venous occlusion, and pitting edema. Edema with purple bullae, ecchymosis, and cutaneous anesthesia suggests loss of vascular integrity and necessitates exploration of the deeper structures for evidence of necrotizing fasciitis or myonecrosis. An early diagnosis requires a high level of suspicion in

instances of unexplained fever and of pain and tenderness in the soft tissue, even in the absence of acute cutaneous inflammation.

Table 124-1 indicates the chapters in which the infections described below are discussed in greater detail. Many of these infections are illustrated in the chapters cited or in Chap. A1.

TABLE 124-1

Skin and Soft Tissue Infections

LESION, CLINICAL SYNDROME	INFECTIOUS AGENT(S)	SEE ALSO CHAP(S).
Vesicles		
Smallpox	Variola virus	S2
Chickenpox	Varicella-zoster virus	188
Shingles (herpes zoster)	Varicella-zoster virus	188
Cold sores, herpetic whitlow, herpes gladiatorum	Herpes simplex virus	187
Hand-foot-and-mouth disease	Coxsackievirus A16	199
Orf	Parapoxvirus	191
Molluscum contagiosum	Molluscum contagiosum poxvirus	191
Rickettsialpox	*Rickettsia akari*	182
Blistering distal dactylitis	*Staphylococcus aureus* or *Streptococcus pyogenes*	142, 143
Bullae		
Staphylococcal scalded-skin syndrome	*S. aureus*	142
Necrotizing fasciitis	*S. pyogenes*, *Clostridium* spp., mixed aerobes and anaerobes	143, 149, 172
Gas gangrene	*Clostridium* spp.	149
Halophilic vibrio	*Vibrio vulnificus*	163
Crusted lesions		

LESION, CLINICAL SYNDROME	INFECTIOUS AGENT(S)	SEE ALSO CHAP(S).
Bullous impetigo/ecthyma	*S. aureus*	142
Impetigo contagiosa	*S. pyogenes*	143
Ringworm	Superficial dermatophyte fungi	214
Sporotrichosis	*Sporothrix schenckii*	214
Histoplasmosis	*Histoplasma capsulatum*	207
Coccidioidomycosis	*Coccidioides immitis*	208
Blastomycosis	*Blastomyces dermatitidis*	209
Cutaneous leishmaniasis	*Leishmania* spp.	221
Cutaneous tuberculosis	*Mycobacterium tuberculosis*	173
Nocardiosis	*Nocardia asteroides*	169
Folliculitis		
Furunculosis	*S. aureus*	142
Hot-tub folliculitis	*Pseudomonas aeruginosa*	159
Swimmer’s itch	*Schistosoma* spp.	229
Acne vulgaris	*Propionibacterium acnes*	53
Papular and nodular lesions		
Fish-tank or swimming-pool granuloma	*Mycobacterium marinum*	175
Creeping eruption (cutaneous larva migrans)	*Ancylostoma braziliense*	226
Dracunculiasis	*Dracunculus medinensis*	228
Cercarial dermatitis	*Schistosoma mansoni*	229

LESION, CLINICAL SYNDROME	INFECTIOUS AGENT(S)	SEE ALSO CHAP(S).
Verruca vulgaris	Human papillomaviruses 1, 2, 4	193
Condylomata acuminata (anogenital warts)	Human papillomaviruses 6, 11, 16, 18	193
Onchocerciasis nodule	*Onchocerca volvulus*	228
Cutaneous myiasis	*Dermatobia hominis*	452
Verruca peruana	*Bartonella bacilliformis*	167
Cat-scratch disease	*Bartonella henselae*	167
Lepromatous leprosy	*Mycobacterium leprae*	174
Secondary syphilis (papulosquamous and nodular lesions, condylomata lata)	*Treponema pallidum*	177
Tertiary syphilis (nodular gummatous lesions)	*T. pallidum*	177
Ulcers with or without eschars		
Anthrax	*Bacillus anthracis*	S2
Ulceroglandular tularemia	*Francisella tularensis*	165, S2
Bubonic plague	*Yersinia pestis*	166, S2
Buruli ulcer	*Mycobacterium ulcerans*	175
Leprosy	*M. leprae*	174
Cutaneous tuberculosis	*M. tuberculosis*	173
Chancroid	*Haemophilus ducreyi*	152
Primary syphilis	*T. pallidum*	177
Erysipelas	*S. pyogenes*	143

LESION, CLINICAL SYNDROME	INFECTIOUS AGENT(S)	SEE ALSO CHAP(S).
Cellulitis	*Staphylococcus* spp., *Streptococcus* spp., various other bacteria	Various
Necrotizing fasciitis		
Streptococcal gangrene	*S. pyogenes*	143
Fournier's gangrene	Mixed aerobic and anaerobic bacteria	172
Staphylococcal necrotizing fasciitis	Methicillin-resistant *S. aureus*	142
Myositis and myonecrosis		
Pyomyositis	*S. aureus*	142
Streptococcal necrotizing myositis	*S. pyogenes*	143
Gas gangrene	*Clostridium* spp.	149
Nonclostridial (crepitant) myositis	Mixed aerobic and anaerobic bacteria	172
Synergistic nonclostridial anaerobic myonecrosis	Mixed aerobic and anaerobic bacteria	172

INFECTIONS ASSOCIATED WITH VESICLES

(Table 124-1) Vesicle formation due to infection is caused by viral proliferation within the epidermis. In varicella and variola, viremia precedes the onset of a diffuse centripetal rash that progresses from macules to vesicles, then to pustules, and finally to scabs over the course of 1–2 weeks. Vesicles of varicella have a "dewdrop" appearance and develop in crops randomly about the trunk, extremities, and face over 3–4 days. Herpes zoster occurs in a single dermatome; the appearance of vesicles is preceded by pain for several days. Zoster may occur in persons of any age but is most common among immunosuppressed individuals and elderly patients, whereas most cases of varicella occur in young children. Vesicles due to HSV are found on or around the lips (HSV-1) or genitals (HSV-2) but also may appear on the head and neck of young wrestlers (herpes gladiatorum) or on the digits of health care workers (herpetic whitlow). Recurrent herpes labialis (HSV-1) and herpes genitalis commonly follow primary infection. Coxsackievirus A16 characteristically causes vesicles on the hands, feet, and mouth of children. Orf is caused by a DNA virus related to smallpox virus and

infects the fingers of individuals who work around goats and sheep. Molluscum contagiosum virus induces flaccid vesicles on the skin of healthy and immunocompromised individuals. Although variola (smallpox) in nature was eradicated as of 1977, postmillennial terrorist events have renewed interest in this devastating infection **(Chap. S2)**. Viremia beginning after an incubation period of 12 days is followed by a diffuse maculopapular rash, with rapid evolution to vesicles, pustules, and then scabs. Secondary cases can occur among close contacts.

Rickettsialpox begins after mite-bite inoculation of *Rickettsia akari* into the skin. A papule with a central vesicle evolves to form a 1- to 2.5-cm painless crusted black eschar with an erythematous halo and proximal adenopathy. While more common in the northeastern United States and Ukraine in 1940–1950, rickettsialpox has recently been described in Ohio, Arizona, and Utah. Blistering dactylitis is a painful, vesicular, localized *S. aureus* or group A streptococcal infection of the pulps of the distal digits of the hands.

INFECTIONS ASSOCIATED WITH BULLAE

(Table 124-1) Staphylococcal scalded-skin syndrome (SSSS) in neonates is caused by a toxin (exfoliatin) from phage group II *S. aureus*. SSSS must be distinguished from toxic epidermal necrolysis (TEN), which occurs primarily in adults, is drug-induced, and is associated with a higher mortality rate. Punch biopsy with frozen section is useful in making this distinction since the cleavage plane is the stratum corneum in SSSS and the stratum germinativum in TEN (Fig. 124-1). Intravenous γ-globulin is a promising treatment for TEN. Necrotizing fasciitis and gas gangrene also induce bulla formation (see "Necrotizing Fasciitis," below). Halophilic vibrio infection can be as aggressive and fulminant as necrotizing fasciitis; a helpful clue in its diagnosis is a history of exposure to waters of the Gulf of Mexico or the Atlantic seaboard or (in a patient with cirrhosis) the ingestion of raw seafood. The etiologic organism (*Vibrio vulnificus*) is highly susceptible to tetracycline.

INFECTIONS ASSOCIATED WITH CRUSTED LESIONS

(Table 124-1) Impetigo contagiosa is caused by *S. pyogenes*, and bullous impetigo is due to *S. aureus*. Both skin lesions may have an early bullous stage but then appear as thick crusts with a golden-brown color. Epidemics of impetigo caused by MRSA have been reported. Streptococcal lesions are most common among children 2–5 years of age, and epidemics may occur in settings of poor hygiene, particularly among children in lower socioeconomic settings in tropical climates. It is important to recognize impetigo contagiosa because of its relationship to poststreptococcal glomerulonephritis. Rheumatic fever is not a complication of skin infection caused by *S. pyogenes*. Superficial dermatophyte infection (ringworm) can occur on any skin surface, and skin scrapings with KOH staining are diagnostic. Primary infections with dimorphic fungi such as *Blastomyces dermatitidis* and *Sporothrix schenckii* can initially present as crusted skin lesions resembling ringworm. Disseminated infection with *Coccidioides immitis* can also involve the skin, and biopsy and culture should be performed on crusted lesions when the patient is from an endemic area. Crusted nodular lesions caused by *Mycobacterium chelonei* have been described in HIV-seropositive patients. Treatment with clarithromycin looks promising.

FOLLICULITIS

(Table 124-1) Hair follicles serve as portals for a number of bacteria, although *S. aureus* is the most common cause of localized folliculitis. Sebaceous glands empty into hair follicles and ducts and, if these portals are blocked, form sebaceous cysts that may resemble staphylococcal abscesses or may become secondarily infected. Inflammation of sweat glands (hidradenitis suppurativa) also can mimic infection of hair follicles, particularly in the axillae, but new treatments with potent anti-inflammatory agents hold promise. Chronic folliculitis is uncommon except in acne vulgaris, where constituents of the normal flora (e.g., *Propionibacterium acnes*) may play a role.

Diffuse folliculitis occurs in two settings. *Hot-tub folliculitis* is caused by *Pseudomonas aeruginosa* in waters that are insufficiently chlorinated and maintained at temperatures of 37–40°C. Infection is usually self-limited, although bacteremia and shock have been reported. *Swimmer's itch* occurs when a skin surface is exposed to water infested with freshwater avian schistosomes. Warm water temperatures and alkaline pH are suitable for mollusks that serve as intermediate hosts between birds and humans. Free-swimming schistosomal cercariae readily penetrate human hair follicles or pores but quickly die and elicit a brisk allergic reaction, causing intense itching and erythema.

PAPULAR AND NODULAR LESIONS

(Table 124-1) Raised lesions of the skin occur in many different forms. *Mycobacterium marinum* infections of the skin may present as cellulitis or as raised erythematous nodules. Similar lesions caused by *Mycobacterium abscessus* and *M. chelonei* have been described among patients undergoing cosmetic laser surgery and tattooing, respectively. Erythematous papules are early manifestations of cat-scratch disease (with lesions developing at the primary site of inoculation of *Bartonella henselae*) and bacillary angiomatosis (also caused by *B. henselae*). Raised serpiginous or linear eruptions are characteristic of cutaneous larva migrans, which is caused by burrowing larvae of dog or cat hookworms (*Ancylostoma braziliense*) and which humans acquire through contact with soil that has been contaminated with dog or cat feces. Similar burrowing raised lesions are present in dracunculiasis caused by migration of the adult female nematode *Dracunculus medinensis*. Nodules caused by *Onchocerca volvulus* measure 1–10 cm in diameter and occur mostly in persons bitten by *Simulium* flies in Africa. The nodules contain the adult worm encased in fibrous tissue. Migration of microfilariae into the eyes may result in blindness. Verruga peruana is caused by *Bartonella bacilliformis*, which is transmitted to humans by the sandfly *Phlebotomus*. This condition can take the form of single gigantic lesions (several centimeters in diameter) or multiple small lesions (several millimeters in diameter). Numerous subcutaneous nodules may also be present in cysticercosis caused by larvae of *Taenia solium*. Multiple erythematous papules develop in schistosomiasis; each represents a cercarial invasion site. Skin nodules as well as thickened subcutaneous tissue are prominent features of lepromatous leprosy. Large nodules or gummas are features of tertiary syphilis, whereas flat papulosquamous lesions are characteristic of secondary syphilis. Human papillomavirus may cause singular warts (verruca vulgaris) or multiple warts in the anogenital area (condylomata acuminata). The latter are major problems in HIV-infected individuals.

ULCERS WITH OR WITHOUT ESCHARS

(Table 124-1) Cutaneous anthrax begins as a pruritic papule, which develops within days into an ulcer with surrounding vesicles and edema and then into an enlarging ulcer with a black eschar. Cutaneous anthrax may cause chronic nonhealing ulcers with an overlying dirty-gray membrane, although lesions may also mimic psoriasis, eczema, or impetigo. Ulceroglandular tularemia may have associated ulcerated skin lesions with painful regional adenopathy. Although buboes are the major cutaneous manifestation of plague, ulcers with eschars, papules, or pustules are also present in 25% of cases.

Mycobacterium ulcerans typically causes chronic skin ulcers on the extremities of individuals living in the tropics. *Mycobacterium leprae* may be associated with cutaneous ulcerations in patients with lepromatous leprosy related to Lucio's phenomenon, in which immune-mediated destruction of tissue bearing high concentrations of *M. leprae* bacilli occurs, usually several months after initiation of effective therapy. *Mycobacterium tuberculosis* also may cause ulcerations, papules, or erythematous macular lesions of the skin in both immunocompetent and immunocompromised patients.

Decubitus ulcers are due to tissue hypoxemia secondary to pressure-induced vascular insufficiency and may become secondarily infected with components of the skin and gastrointestinal flora, including anaerobes. Ulcerative lesions on the anterior shins may be due to pyoderma gangrenosum, which must be distinguished from similar lesions of infectious etiology by histologic evaluation of biopsy sites. Ulcerated lesions on the genitals may be either painful (chancroid) or painless (primary syphilis).

ERYSIPELAS

(Table 124-1) Erysipelas is due to *S. pyogenes* and is characterized by an abrupt onset of fiery-red swelling of the face or extremities. The distinctive features of erysipelas are well-defined indurated margins, particularly along the nasolabial fold; rapid progression; and intense pain. Flaccid bullae may develop during the second or third day of illness, but extension to deeper soft tissues is rare. Treatment with penicillin is effective; swelling may progress despite appropriate treatment, although fever, pain, and the intense red color diminish. Desquamation of the involved skin occurs 5–10 days into the illness. Infants and elderly adults are most commonly afflicted, and the severity of systemic toxicity varies.

CELLULITIS

(Table 124-1) Cellulitis is an acute inflammatory condition of the skin that is characterized by localized pain, erythema, swelling, and heat. It may be caused by indigenous flora colonizing the skin and appendages (e.g., *S. aureus* and *S. pyogenes*) or by a wide variety of exogenous bacteria. Because the exogenous bacteria involved in cellulitis occupy unique niches in nature, a thorough history (including epidemiologic data) offers important clues to etiology. When there is drainage, an open wound, or an obvious portal of entry, Gram's stain and culture provide a definitive diagnosis. In the absence of these findings, the bacterial etiology of cellulitis is difficult to establish, and in some cases staphylococcal and streptococcal cellulitis may have

similar features. Even with needle aspiration of the leading edge or a punch biopsy of the cellulitis tissue itself, cultures are positive in only 20% of cases. This observation suggests that relatively low numbers of bacteria may cause cellulitis and that the expanding area of erythema within the skin may be a direct effect of extracellular toxins or of the soluble mediators of inflammation elicited by the host.

Bacteria may gain access to the epidermis through cracks in the skin, abrasions, cuts, burns, insect bites, surgical incisions, and IV catheters. Cellulitis caused by *S. aureus* spreads from a central localized infection, such as an abscess, folliculitis, or an infected foreign body (e.g., a splinter, a prosthetic device, or an IV catheter). MRSA is rapidly replacing methicillin-sensitive *S. aureus* (MSSA) as a cause of cellulitis in both inpatient and outpatient settings. Cellulitis caused by MSSA or MRSA is usually associated with a focal infection, such as a furuncle, a carbuncle, a surgical wound, or an abscess; the U.S. Food and Drug Administration preferentially refers to these types of infection as *purulent cellulitis*. In contrast, cellulitis due to *S. pyogenes* is a more rapidly spreading, diffuse process that is frequently associated with lymphangitis and fever and should be referred to as *nonpurulent cellulitis.* Recurrent streptococcal cellulitis of the lower extremities may be caused by organisms of group A, C, or G in association with chronic venous stasis or with saphenous venectomy for coronary artery bypass surgery. Streptococci also cause recurrent cellulitis among patients with chronic lymphedema resulting from elephantiasis, lymph node dissection, or Milroy disease. Recurrent staphylococcal cutaneous infections are more common among individuals who have eosinophilia and elevated serum levels of IgE (Job syndrome) and among nasal carriers of staphylococci. Cellulitis caused by *Streptococcus agalactiae* (group B *Streptococcus*) occurs primarily in elderly patients and those with diabetes mellitus or peripheral vascular disease. *Haemophilus influenzae* typically causes periorbital cellulitis in children in association with sinusitis, otitis media, or epiglottitis. It is unclear whether this form of cellulitis will (like meningitis) become less common as a result of the impressive efficacy of the *H. influenzae* type b vaccine.

Many other bacteria also cause cellulitis. It is fortunate that these organisms occur in such characteristic settings that a good history provides useful clues to the diagnosis. Cellulitis associated with cat bites and, to a lesser degree, with dog bites is commonly caused by *Pasteurella multocida*, although in the latter case *Staphylococcus intermedius* and *Capnocytophaga canimorsus* also must be considered. Sites of cellulitis and abscesses associated with dog bites and human bites also contain a variety of anaerobic organisms, including *Fusobacterium*, *Bacteroides*, aerobic and anaerobic streptococci, and *Eikenella corrodens*. *Pasteurella* is notoriously resistant to dicloxacillin and nafcillin but is sensitive to all other β-lactam antimicrobial agents as well as to quinolones, tetracycline, and erythromycin. Amoxicillin-clavulanate, ampicillin-sulbactam, and cefoxitin are good choices for the treatment of animal or human bite infections. *Aeromonas hydrophila* causes aggressive cellulitis and occasionally necrotizing fasciitis in tissues surrounding lacerations sustained in freshwater (lakes, rivers, and streams). This organism remains sensitive to aminoglycosides, fluoroquinolones, chloramphenicol, trimethoprim-sulfamethoxazole, and third-generation cephalosporins; it is resistant to ampicillin, however. *P. aeruginosa* causes three types of soft tissue infection: ecthyma gangrenosum in neutropenic patients, hot-tub folliculitis, and cellulitis following penetrating injury. Most commonly, *P. aeruginosa* is introduced into the deep tissues when a person steps on a nail. Treatment includes surgical inspection and drainage, particularly if the injury also involves bone or

joint capsule. Choices for empirical treatment while antimicrobial susceptibility data are awaited include an aminoglycoside, a third-generation cephalosporin (ceftazidime, cefoperazone, or cefotaxime), a semisynthetic penicillin (ticarcillin, mezlocillin, or piperacillin), or a fluoroquinolone (although drugs of the last class are not indicated for the treatment of children <13 years old).

Gram-negative bacillary cellulitis, including that due to *P. aeruginosa*, is most common among hospitalized, immunocompromised hosts. Cultures and sensitivity tests are critically important in this setting because of multidrug resistance **(Chap. 159)**.

The gram-positive aerobic rod *Erysipelothrix rhusiopathiae* is most often associated with fish and domestic swine and causes cellulitis primarily in bone renderers and fishmongers. *E. rhusiopathiae* remains susceptible to most β-lactam antibiotics (including penicillin), erythromycin, clindamycin, tetracycline, and cephalosporins but is resistant to sulfonamides, chloramphenicol, and vancomycin. Its resistance to vancomycin, which is unusual among gram-positive bacteria, is of potential clinical significance since this agent is sometimes used in empirical therapy for skin infection. Fish food containing the water flea *Daphnia* is sometimes contaminated with *M. marinum*, which can cause cellulitis or granulomas on skin surfaces exposed to the water in aquariums or injured in swimming pools. Rifampin plus ethambutol has been an effective therapeutic combination in some cases, although no comprehensive studies have been undertaken. In addition, some strains of *M. marinum* are susceptible to tetracycline or to trimethoprim-sulfamethoxazole.

NECROTIZING FASCIITIS

(Table 124-1) Necrotizing fasciitis, formerly called streptococcal gangrene, may be associated with group A *Streptococcus* or mixed aerobic–anaerobic bacteria or may occur as a component of gas gangrene caused by *Clostridium perfringens*. Strains of MRSA that produce the Panton-Valentine leukocidin (PVL) toxin have been reported to cause necrotizing fasciitis. Early diagnosis may be difficult when pain or unexplained fever is the only presenting manifestation. Swelling then develops and is followed by brawny edema and tenderness. With progression, dark-red induration of the epidermis appears, along with bullae filled with blue or purple fluid. Later the skin becomes friable and takes on a bluish, maroon, or black color. By this stage, thrombosis of blood vessels in the dermal papillae (Fig. 124-1) is extensive. Extension of infection to the level of the deep fascia causes this tissue to take on a brownish-gray appearance. Rapid spread occurs along fascial planes, through venous channels and lymphatics. Patients in the later stages are toxic and frequently manifest shock and multiorgan failure.

Necrotizing fasciitis caused by mixed aerobic–anaerobic bacteria begins with a breach in the integrity of a mucous membrane barrier, such as the mucosa of the gastrointestinal or genitourinary tract. The portal can be a malignancy, a diverticulum, a hemorrhoid, an anal fissure, or a urethral tear. Other predisposing factors include peripheral vascular disease, diabetes mellitus, surgery, and penetrating injury to the abdomen. Leakage into the perineal area results in a syndrome called *Fournier's gangrene*, characterized by massive swelling of the scrotum and penis with extension into the perineum or the abdominal wall and the legs.

Necrotizing fasciitis caused by *S. pyogenes* has increased in frequency and severity since 1985. There are two distinct clinical presentations: those with no portal of entry and those with a defined portal of entry. Infections in the first category often begin deep at the site of a nonpenetrating minor trauma, such as a bruise or a muscle strain. Seeding of the site via transient bacteremia is likely, although most patients deny antecedent streptococcal infection. The affected patients present with only severe pain and fever. Late in the course, the classic signs of necrotizing fasciitis, such as purple (violaceous) bullae, skin sloughing, and progressive toxicity, develop. In infections of the second type, *S. pyogenes* may reach the deep fascia from a site of cutaneous infection or penetrating trauma. These patients have early signs of superficial skin infection with progression to necrotizing fasciitis. In either case, toxicity is severe, and renal impairment may precede the development of shock. In 20–40% of cases, myositis occurs concomitantly, and, as in gas gangrene (see below), serum creatine phosphokinase levels may be markedly elevated. Necrotizing fasciitis due to mixed aerobic–anaerobic bacteria may be associated with gas in deep tissue, but gas usually is not present when the cause is *S. pyogenes* or MRSA. Prompt surgical exploration down to the deep fascia and muscle is essential. Necrotic tissue must be surgically removed, and Gram's staining and culture of excised tissue are useful in establishing whether group A streptococci, mixed aerobic–anaerobic bacteria, MRSA, or *Clostridium* species are present (see "Treatment," below).

MYOSITIS AND MYONECROSIS

(Table 124-1) Muscle involvement can occur with viral infection (e.g., influenza, dengue, or coxsackievirus B infection) or parasitic invasion (e.g., trichinellosis, cysticercosis, or toxoplasmosis). Although myalgia develops in most of these infections, severe muscle pain is the hallmark of pleurodynia (coxsackievirus B), trichinellosis, and bacterial infection. Acute rhabdomyolysis predictably occurs with clostridial and streptococcal myositis but may also be associated with influenza virus, echovirus, coxsackievirus, Epstein-Barr virus, and *Legionella* infections.

Pyomyositis is usually due to *S. aureus*, is common in tropical areas, and generally has no known portal of entry. Cases of pyomyositis caused by MRSA producing the PVL toxin have been described among children in the United States. Muscle infection begins at the exact site of blunt trauma or muscle strain. Infection remains localized, and shock does not develop unless organisms produce toxic shock syndrome toxin 1 or certain enterotoxins and the patient lacks antibodies to the toxin produced by the infecting organisms. In contrast, *S. pyogenes* may induce primary myositis (referred to as *streptococcal necrotizing myositis*) in association with severe systemic toxicity. Myonecrosis occurs concomitantly with necrotizing fasciitis in ~50% of cases. Both are part of the streptococcal toxic shock syndrome.

Gas gangrene usually follows severe penetrating injuries that result in interruption of the blood supply and introduction of soil into wounds. Such cases of traumatic gangrene are usually caused by the clostridial species *C. perfringens*, *C. septicum*, and *C. histolyticum*. Rarely, latent or recurrent gangrene can occur years after penetrating trauma; dormant spores that reside at the site of previous injury are most likely responsible. Spontaneous nontraumatic gangrene among patients with neutropenia, gastrointestinal malignancy, diverticulosis, or recent radiation therapy to the abdomen is caused by several clostridial

species, of which *C. septicum* is the most commonly involved. The tolerance of this anaerobe to oxygen probably explains why it can initiate infection spontaneously in normal tissue anywhere in the body.

Gas gangrene of the uterus, especially that due to *Clostridium sordellii,* historically occurred as a consequence of illegal or self-induced abortion and nowadays also follows spontaneous abortion, vaginal delivery, and cesarean section. *C. sordellii* has also been implicated in medically induced abortion. Postpartum *C. sordellii* infections in young, previously healthy women present as a unique clinical picture: little or no fever, lack of a purulent discharge, refractory hypotension, extensive peripheral edema and effusions, hemoconcentration, and a markedly elevated white blood cell count. The infection is almost uniformly fatal, with death ensuing rapidly. *C. sordellii* and *C. novyi* have also been associated with cutaneous injection of black tar heroin; mortality rates are lower among the affected individuals, probably because their aggressive injection-site infections are readily apparent and diagnosis is therefore prompt.

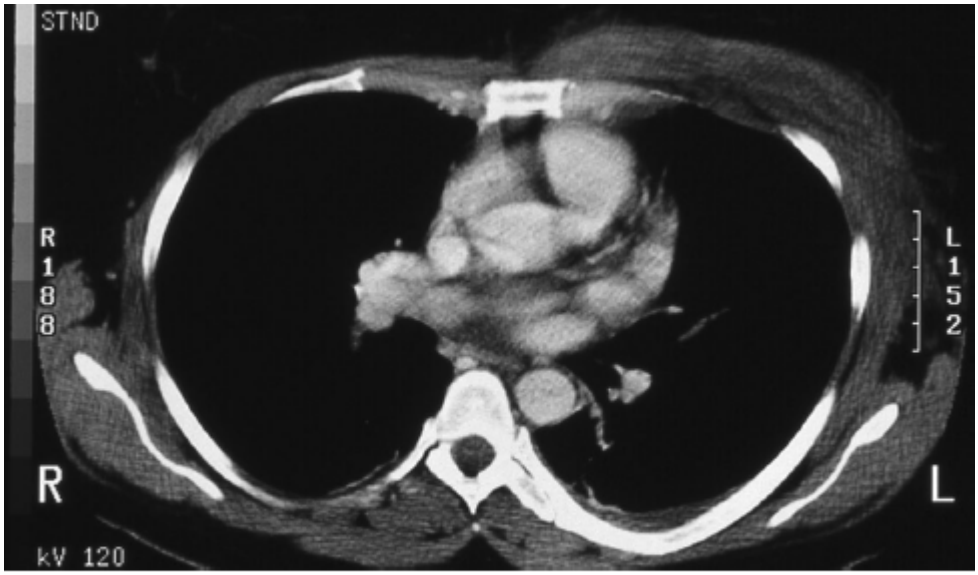
Synergistic nonclostridial anaerobic myonecrosis, also known as necrotizing cutaneous myositis and synergistic necrotizing cellulitis, is a variant of necrotizing fasciitis caused by mixed aerobic and anaerobic bacteria with the exclusion of clostridial organisms (see "Necrotizing Fasciitis," above).

DIAGNOSIS

This chapter emphasizes the physical appearance and location of lesions within the soft tissues as important diagnostic clues. Other crucial considerations in narrowing the differential diagnosis are the temporal progression of the lesions as well as the patient's travel history, animal exposure or bite history, age, underlying disease status, and lifestyle. However, even the astute clinician may find it challenging to diagnose all infections of the soft tissues by history and inspection alone. Soft tissue radiography, CT **(Fig. 124-2)**, and MRI may be useful in determining the depth of infection and should be performed when the patient has rapidly progressing lesions or evidence of a systemic inflammatory response syndrome. These tests are particularly valuable for defining a localized abscess or detecting gas in tissue. Unfortunately, they may reveal only soft tissue swelling and thus are not specific for fulminant infections such as necrotizing fasciitis or myonecrosis caused by group A *Streptococcus* (Fig. 124-2), where gas is not found in lesions.

FIGURE 124-2

CT showing edema and inflammation of the left chest wall in a patient with necrotizing fasciitis and myonecrosis caused by group A *Streptococcus*.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Aspiration of the leading edge or punch biopsy with frozen section may be helpful if the results of imaging tests are positive, but false-negative results occur in ~80% of cases. There is some evidence that aspiration alone may be superior to injection and aspiration with normal saline. Frozen sections are especially useful in distinguishing SSSS from TEN and are quite valuable in cases of necrotizing fasciitis. Open surgical inspection, with debridement as indicated, is clearly the best way to determine the extent and severity of infection and to obtain material for Gram's staining and culture. Such an aggressive approach is important and may be lifesaving if undertaken early in the course of fulminant infections where there is evidence of systemic toxicity.

TREATMENT

TREATMENT

Infections of the Skin, Muscles, and Soft Tissues

A full description of the treatment of all the clinical entities described herein is beyond the scope of this chapter. As a guide to the clinician in selecting appropriate treatment, the antimicrobial agents useful in the most common and the most fulminant cutaneous infections are listed in Table 124-2.

Furuncles, carbuncles, and abscesses caused by MRSA and MSSA are common, and their treatment depends upon the size of the lesion. Furuncles <2.5 cm in diameter are usually treated with moist heat. Those that are larger (4.5 cm of erythema and induration) require surgical drainage, and the occurrence of these larger lesions in association with fever, chills, or leukocytosis requires both drainage and antibiotic treatment. Previous studies in children demonstrated that surgical drainage of abscesses (mean diameter, 3.8 cm) was as effective when used alone as when combined with trimethoprim-sulfamethoxazole treatment. However, the rate of recurrence of new lesions was lower in the group undergoing both drainage and antibiotic treatment. Recent studies in adults with predominantly MRSA localized abscesses suggested that a 7- to 10-

day course of treatment with trimethoprim-sulfamethoxazole was associated with higher cure rates and fewer recurrences. In children, a 3-day course was not as effective as a 7-day course.

Early and aggressive surgical exploration is essential in cases of suspected necrotizing fasciitis, myositis, or gangrene in order to (1) visualize the deep structures, (2) remove necrotic tissue, (3) reduce compartment pressure, and (4) obtain suitable material for Gram's staining and for aerobic and anaerobic cultures. Appropriate empirical antibiotic treatment for mixed aerobic–anaerobic infections could consist of ampicillin-sulbactam, cefoxitin, or the following combination: (1) clindamycin (600–900 mg IV every 8 h) or metronidazole (500 mg every 6 h) plus (2) ampicillin or ampicillin-sulbactam (1.5–3 g IV every 6 h) plus (3) gentamicin (1–1.5 mg/kg every 8 h). Group A streptococcal and clostridial infection of the fascia and/or muscle carries a mortality rate of 20–50% with penicillin treatment. In experimental models of streptococcal and clostridial necrotizing fasciitis/myositis, clindamycin has exhibited markedly superior efficacy, but no comparative clinical trials have been performed. A retrospective study of children with invasive group A streptococcal infection demonstrated higher survival rates with clindamycin treatment than with β-lactam antibiotic therapy. Hyperbaric oxygen treatment also may be useful in gas gangrene due to clostridial species. Antibiotic treatment should be continued until all signs of systemic toxicity have resolved, all devitalized tissue has been removed, and granulation tissue has developed **(Chaps. 143, 149, and 172)**.

In summary, infections of the skin and soft tissues are diverse in presentation and severity and offer a great challenge to the clinician. This chapter provides an approach to diagnosis and understanding of the pathophysiologic mechanisms involved in these infections. More in-depth information is found in chapters on specific infections.

TABLE 124-2

Treatment of Common Infections of the Skin

DIAGNOSIS/CONDITION	PRIMARY TREATMENT	ALTERNATIVE TREATMENT	SEE ALSO CHAP(S).
Animal bite (prophylaxis or early infection)[a]	Amoxicillin–clavulanate (875/125 mg PO bid)	Doxycycline (100 mg PO bid)	136
Animal bite[a] (established infection)	Ampicillin–sulbactam (1.5–3 g IV q6h)	Clindamycin (600–900 mg IV q8h) ***plus*** Ciprofloxacin (400 mg IV q12h) *or* cefoxitin (2 g IV q6h)	136
Bacillary angiomatosis	Erythromycin (500 mg PO qid)	Doxycycline (100 mg PO bid)	167
Herpes simplex (primary genital)	Acyclovir (400 mg PO tid for 10 days)	Famciclovir (250 mg PO tid for 5–10 days) *or* valacyclovir (1000 mg PO bid for 10 days)	187
Herpes zoster (immunocompetent host >50 years of age)	Acyclovir (800 mg PO 5 times daily for 7–10 days)	Famciclovir (500 mg PO tid for 7–10 days) *or* valacyclovir (1000 mg PO tid for 7 days)	188
Cellulitis (staphylococcal or streptococcal[b,c])	Nafcillin or oxacillin (2 g IV q4–6h)	Cefazolin (1–2 g q8h) *or* ampicillin/sulbactam (1.5–3 g IV q6h) *or* erythromycin (0.5–1 g IV q6h) *or* clindamycin (600–900 mg IV q8h)	142, 143
MRSA skin infection[d]	Vancomycin (1 g IV q12h)	Linezolid (600 mg IV q12h)	142
Necrotizing fasciitis (group A streptococcal[b])	Clindamycin (600–900 mg IV q6–8h) *plus* penicillin G (4 million units IV q4h)	Clindamycin (600–900 mg IV q6–8h) *plus* a cephalosporin (first- or second-generation)	143

DIAGNOSIS/CONDITION	PRIMARY TREATMENT	ALTERNATIVE TREATMENT	SEE ALSO CHAP(S).
Necrotizing fasciitis (mixed aerobes and anaerobes)	Ampicillin (2 g IV q4h) *plus* clindamycin (600–900 mg IV q6–8h) *plus* ciprofloxacin (400 mg IV q6–8h)	Vancomycin (1 g IV q6h) *plus* metronidazole (500 mg IV q6h) *plus* ciprofloxacin (400 mg IV q6–8h)	117, 172
Gas gangrene	Clindamycin (600–900 mg IV q6–8h) *plus* penicillin G (4 million units IV q4–6h)	Clindamycin (600–900 mg IV q6–8h) *plus* cefoxitin (2 g IV q6h)	149

[a]*Pasteurella multocida,* a species commonly associated with both dog and cat bites, is resistant to cephalexin, dicloxacillin, clindamycin, and erythromycin. *Eikenella corrodens*, a bacterium commonly associated with human bites, is resistant to clindamycin, penicillinase-resistant penicillins, and metronidazole but is sensitive to trimethoprim-sulfamethoxazole and fluoroquinolones. [b]The frequency of erythromycin resistance in group A *Streptococcus* is currently ~5% in the United States but has reached 70–100% in some other countries. Most, but not all, erythromycin-resistant group A streptococci are susceptible to clindamycin. Approximately 90% of *Staphylococcus aureus* strains are sensitive to clindamycin, but resistance—both intrinsic and inducible—is increasing. [c]Severe hospital-acquired *S. aureus* infections or community-acquired *S. aureus* infections that are not responding to the β-lactam antibiotics recommended in this table may be caused by methicillin-resistant strains, requiring a switch to vancomycin, daptomycin, or linezolid. [d]Some strains of methicillin-resistant *S. aureus* (MRSA) remain sensitive to tetracycline and trimethoprim-sulfamethoxazole. Daptomycin (4 mg/kg IV q24h) or tigecycline (100-mg loading dose followed by 50 mg IV q12h) is an alternative treatment for MRSA.

FURTHER READING

Aldape MJ et al: *Clostridium sordellii* infection: Epidemiology, clinical findings, and current perspectives on diagnosis and treatment. Clin Infect Dis 43:1436, 2006. [PubMed: 17083018]

Stevens DL, Bryant AE: Life threatening skin and soft tissue infections, in *Netter's Infectious Diseases*, EC Jong, and DL Stevens (eds). Philadelphia, Elsevier, 2011, pp 94–101.

Stevens DL et al: Practice guidelines for the diagnosis and management of skin and soft tissue infections: 2014 update by the Infectious Diseases Society of America. Clin Infect Dis 59:e10, 2014 (Erratum: Clin Infect Dis 60:1448, 2015). [PubMed: 24973422]

Stevens DL et al: Trimethoprim-sulfamethoxazole vs placebo for the treatment of uncomplicated skin abscesses. N Engl J Med 374:824, 2016.

Talan DA et al: Bacteriologic analysis of infected dog and cat bites. Emergency Medicine Animal Bite Infection Study Group. N Engl J Med 340:85, 1999. [PubMed: 9887159]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 125: Infectious Arthritis

Lawrence C. Madoff

INTRODUCTION

Although *Staphylococcus aureus*, *Neisseria gonorrhoeae*, and other bacteria are the most common causes of infectious arthritis, various mycobacteria, spirochetes, fungi, and viruses also infect joints (Table 125-1). Since acute bacterial infection can destroy articular cartilage rapidly, all inflamed joints must be evaluated without delay to exclude noninfectious processes and determine appropriate antimicrobial therapy and drainage procedures. For more detailed information on infectious arthritis caused by specific organisms, the reader is referred to the chapters on those organisms.

TABLE 125-1

Differential Diagnosis of Arthritis Syndromes

ACUTE MONARTICULAR ARTHRITIS	CHRONIC MONARTICULAR ARTHRITIS	POLYARTICULAR ARTHRITIS
Staphylococcus aureus	*Mycobacterium tuberculosis*	*Neisseria meningitidis*
Streptococcus pneumoniae	Nontuberculous mycobacteria	*N. gonorrhoeae*
β-Hemolytic streptococci	*Borrelia burgdorferi*	Nongonococcal bacterial arthritis
Gram-negative bacilli	*Treponema pallidum*	Bacterial endocarditis
Neisseria gonorrhoeae	*Candida* spp.	*Candida* spp.
Candida spp.	*Sporothrix schenckii*	Poncet's disease (tuberculous rheumatism)
Crystal-induced arthritis	*Coccidioides immitis*	Hepatitis B virus
Fracture	*Blastomyces dermatitidis*	Parvovirus B19
Hemarthrosis	*Aspergillus* spp.	HIV
Foreign body	*Cryptococcus neoformans*	Human T-lymphotropic virus type 1
Osteoarthritis	*Nocardia* spp.	Rubella virus
Ischemic necrosis	*Brucella* spp.	Arthropod-borne viruses
Monarticular rheumatoid arthritis	Legg-Calvé-Perthes disease	Sickle cell disease flare
	Osteoarthritis	Reactive arthritis
		Serum sickness
		Acute rheumatic fever
		Inflammatory bowel disease
		Systemic lupus erythematosus
		Rheumatoid arthritis/Still's disease
		Other vasculitides
		Sarcoidosis

Acute bacterial infection typically involves a single joint or a few joints. Subacute or chronic monarthritis or oligoarthritis suggests mycobacterial or fungal infection; episodic inflammation is seen in syphilis, Lyme disease, and the reactive arthritis that follows enteric infections and chlamydial urethritis. Acute polyarticular inflammation occurs as an immunologic reaction during the course of endocarditis, rheumatic fever, disseminated neisserial infection, and acute hepatitis B. Bacteria and viruses occasionally infect multiple joints, the former most commonly in persons with rheumatoid arthritis.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Infectious Arthritis

Aspiration of synovial fluid—an essential element in the evaluation of potentially infected joints—can be performed without difficulty in most cases by the insertion of a large-bore needle into the site of maximal fluctuance or tenderness or by the route of easiest access. Ultrasonography or fluoroscopy may be used to guide aspiration of difficult-to-localize effusions of the hip and, occasionally, the shoulder and other joints. Normal synovial fluid contains <180 cells (predominantly mononuclear cells) per microliter. Synovial cell counts averaging 100,000/μL (range, 25,000–250,000/μL), with >90% neutrophils, are characteristic of acute bacterial infections. Crystal-induced, rheumatoid, and other noninfectious inflammatory arthritides usually are associated with <30,000–50,000 cells/μL; cell counts of 10,000–30,000/μL, with 50–70% neutrophils and the remainder lymphocytes, are common in mycobacterial and fungal infections. Definitive diagnosis of an infectious process relies on identification of the pathogen in stained smears of synovial fluid, isolation of the pathogen from cultures of synovial fluid and blood, or detection of microbial nucleic acids and proteins by nucleic acid amplification (NAA)–based assays and immunologic techniques.

ACUTE BACTERIAL ARTHRITIS

PATHOGENESIS

Bacteria enter the joint from the bloodstream; from a contiguous site of infection in bone or soft tissue; or by direct inoculation during surgery, injection, animal or human bite, or trauma. In hematogenous infection, bacteria escape from synovial capillaries, which have no limiting basement membrane, and within hours provoke neutrophilic infiltration of the synovium. Neutrophils and bacteria enter the joint space; later, bacteria adhere to articular cartilage. Degradation of cartilage begins within 48 h as a result of increased intraarticular pressure, release of proteases and cytokines from chondrocytes and synovial macrophages, and invasion of the cartilage by bacteria and inflammatory cells. Histologic studies reveal bacteria lining the synovium and cartilage as well as abscesses extending into the synovium, cartilage, and—in severe cases—subchondral bone. Synovial proliferation results in the formation of a pannus over the cartilage, and thrombosis of inflamed synovial vessels develops. Bacterial factors that appear important in the pathogenesis of infective arthritis include various surface-associated adhesins in *S. aureus* that permit adherence to cartilage and endotoxins that promote chondrocyte-mediated breakdown of cartilage.

MICROBIOLOGY

The hematogenous route of infection is the most common route in all age groups, and nearly every bacterial pathogen is capable of causing septic arthritis. In infants, group B streptococci, gram-negative enteric bacilli, and *S. aureus* are the most common pathogens. Since the advent of the *Haemophilus influenzae* vaccine, the predominant causes among children <5 years of age have been *S. aureus*, *Streptococcus pyogenes* (group A *Streptococcus*), and (in some centers) *Kingella kingae*. Among young adults and adolescents, *N. gonorrhoeae* is the most commonly implicated organism. *S. aureus* accounts for most nongonococcal isolates in adults of all ages; gram-negative bacilli, pneumococci, and β-hemolytic streptococci—particularly groups A and B but also groups C, G, and F—are involved in up to one-third of cases in older adults, especially those with underlying comorbid illnesses.

Infections after surgical procedures or penetrating injuries are due most often to *S. aureus* and occasionally to other gram-positive bacteria or gram-negative bacilli. Infections with coagulase-negative staphylococci are unusual except after the implantation of prosthetic joints or arthroscopy. Anaerobic organisms, often in association with aerobic or facultative bacteria, are found after human bites and when decubitus ulcers or intraabdominal abscesses spread into adjacent joints. Polymicrobial infections complicate traumatic injuries with extensive contamination. Bites and scratches from cats and other animals may introduce *Pasteurella multocida* or *Bartonella henselae* into joints either directly or hematogenously, and bites from humans may introduce *Eikenella corrodens* or other components of the oral flora. Penetration of a sharp object through a shoe is associated with *Pseudomonas aeruginosa* arthritis in the foot.

NONGONOCOCCAL BACTERIAL ARTHRITIS

Epidemiology

Although hematogenous infections with virulent organisms such as *S. aureus*, *H. influenzae*, and pyogenic streptococci occur in healthy persons, there is an underlying host predisposition in many cases of septic arthritis. Patients with rheumatoid arthritis have the highest incidence of infective arthritis (most often secondary to *S. aureus*) because of chronically inflamed joints; glucocorticoid therapy; and frequent breakdown of rheumatoid nodules, vasculitic ulcers, and skin overlying deformed joints. Diabetes mellitus, glucocorticoid therapy, hemodialysis, and malignancy all carry an increased risk of infection with *S. aureus* and gram-negative bacilli. Tumor necrosis factor inhibitors (e.g., etanercept, infliximab), which increasingly are used for the treatment of rheumatoid arthritis, predispose to mycobacterial infections and

possibly to other pyogenic bacterial infections and could be associated with septic arthritis in this population. Pneumococcal infections complicate alcoholism, deficiencies of humoral immunity, and hemoglobinopathies. Pneumococci, *Salmonella* species, and *H. influenzae* cause septic arthritis in persons infected with HIV. Persons with primary immunoglobulin deficiency are at risk for mycoplasmal arthritis, which results in permanent joint damage if tetracycline and replacement therapy with IV immunoglobulin are not administered promptly. IV drug users acquire staphylococcal and streptococcal infections from their own flora and acquire pseudomonal and other gram-negative infections from drugs and injection paraphernalia.

Clinical Manifestations

Some 90% of patients present with involvement of a single joint—most commonly the knee; less frequently the hip; and still less often the shoulder, wrist, or elbow. Small joints of the hands and feet are more likely to be affected after direct inoculation or a bite. Among IV drug users, infections of the spine, sacroiliac joints, and sternoclavicular joints (**Fig. 125-1**) are more common than infections of the appendicular skeleton. Polyarticular infection is most common among patients with rheumatoid arthritis and may resemble a flare of the underlying disease.

FIGURE 125-1

Acute septic arthritis of the sternoclavicular joint. A man in his forties with a history of cirrhosis presented with a new onset of fever and lower neck pain. He had no history of IV drug use or previous catheter placement. Jaundice and a painful swollen area over his left sternoclavicular joint were evident on physical examination. Cultures of blood drawn at admission grew group B *Streptococcus*. The patient recovered after treatment with IV penicillin. *(Courtesy of Francisco M. Marty, MD, Brigham and Women's Hospital, Boston; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The usual presentation consists of moderate to severe pain that is uniform around the joint, effusion, muscle spasm, and decreased range of motion. Fever in the range of 38.3–38.9°C (101–102°F) and sometimes higher is common but may not be present, especially in persons with rheumatoid arthritis, renal or hepatic insufficiency, or conditions requiring immunosuppressive therapy. The inflamed, swollen joint is usually evident on examination except in the case of a deeply situated joint such as the hip, shoulder, or sacroiliac joint. Cellulitis, bursitis, and acute osteomyelitis, which may produce a similar clinical picture, should be distinguished from septic arthritis by their greater range of motion and less-than-circumferential swelling. A focus of extraarticular infection, such as a boil or pneumonia, should be sought. Peripheral-blood leukocytosis with a left shift and elevation of the erythrocyte sedimentation rate or C-reactive protein level are common.

Plain radiographs show evidence of soft-tissue swelling, joint-space widening, and displacement of tissue planes by the distended capsule. Narrowing of the joint space and bony erosions indicate advanced infection and a poor prognosis. Ultrasound is useful for detecting effusions in the hip, and CT or MRI can demonstrate infections of the sacroiliac joint, the sternoclavicular joint, and the spine very well.

Laboratory Findings

Specimens of peripheral blood and synovial fluid should be obtained before antibiotics are administered. Blood cultures are positive in up to 50–70% of *S. aureus* infections but are less frequently positive in infections due to other organisms. The synovial fluid is turbid, serosanguineous, or frankly purulent. Gram-stained smears confirm the presence of large numbers of neutrophils. Levels of total protein and lactate dehydrogenase in synovial fluid are elevated, and the glucose level is depressed; however, these findings are not specific for infection, and measurement of these levels is not necessary for diagnosis. The synovial fluid should be examined for crystals because gout and

pseudogout can resemble septic arthritis clinically and infection and crystal-induced disease occasionally occur together. Organisms are seen on synovial fluid smears in nearly three-quarters of infections with *S. aureus* and streptococci and in 30–50% of infections due to gram-negative and other bacteria. Cultures of synovial fluid are positive in >90% of cases. Inoculation of synovial fluid into bottles containing liquid media for blood cultures increases the yield of a culture, especially if the pathogen is a fastidious organism or the patient is taking an antibiotic. NAA-based assays for bacterial DNA, when available, can be useful for the diagnosis of partially treated or culture-negative bacterial arthritis.

TREATMENT

TREATMENT

Nongonococcal Bacterial Arthritis

Prompt administration of systemic antibiotics and drainage of the involved joint can prevent destruction of cartilage, postinfectious degenerative arthritis, joint instability, or deformity. Once samples of blood and synovial fluid have been obtained for culture, empirical antibiotics should be directed against the bacteria visualized on smears or the pathogens that are likely in light of the patient's age and risk factors. Initial therapy should consist of IV-administered bactericidal agents; direct instillation of antibiotics into the joint is not necessary to achieve adequate levels in synovial fluid and tissue. An IV third-generation cephalosporin such as cefotaxime (1 g every 8 h) or ceftriaxone (1–2 g every 24 h) provides adequate empirical coverage for most community-acquired infections in adults when smears show no organisms. IV vancomycin (1 g every 12 h) is used if there are gram-positive cocci on the smear. If methicillin-resistant *S. aureus* is an unlikely pathogen (e.g., when it is not widespread in the community), cefazolin (2 g every 8 h), oxacillin (2 g every 4 h), or nafcillin (2 g every 4 h) should be given. In addition, an aminoglycoside or third-generation cephalosporin should be given to IV drug users and to other patients in whom *P. aeruginosa* may be the responsible agent.

Definitive therapy is based on the identity and antibiotic susceptibility of the bacteria isolated in culture. Infections due to staphylococci are treated with cefazolin, oxacillin, nafcillin, or vancomycin for 4 weeks. Pneumococcal and streptococcal infections due to penicillin-susceptible organisms respond to 2 weeks of therapy with penicillin G (2 million units IV every 4 h); infections caused by *H. influenzae* and by strains of *Streptococcus pneumoniae* that are resistant to penicillin are treated with cefotaxime or ceftriaxone for 2 weeks. Most enteric gram-negative infections can be cured in 3–4 weeks by a second- or third-generation cephalosporin given IV or by a fluoroquinolone such as levofloxacin (500 mg IV or PO every 24 h). *P. aeruginosa* infection should be treated for at least 2 weeks with a combination regimen composed of an aminoglycoside plus either an extended-spectrum penicillin such as mezlocillin (3 g IV every 4 h) or an antipseudomonal cephalosporin such as ceftazidime (1 g IV every 8 h). If tolerated, this regimen is continued for an additional 2 weeks; alternatively, a fluoroquinolone such as ciprofloxacin (750 mg PO twice daily) is given by itself or with the penicillin or cephalosporin in place of the aminoglycoside.

Timely drainage of pus and necrotic debris from the infected joint is required for a favorable outcome. Needle aspiration of readily accessible joints such as the knee may be adequate if loculations or particulate matter in the joint does not prevent its thorough decompression. Arthroscopic drainage and lavage may be employed initially or within several days if repeated needle aspiration fails to relieve symptoms, decrease the volume of the effusion and the synovial white cell count, and clear bacteria from smears and cultures. In some cases, arthrotomy is necessary to remove loculations and debride infected synovium, cartilage, or bone. Septic arthritis of the hip is best managed with arthrotomy, particularly in young children, in whom infection threatens the viability of the femoral head. Septic joints do not require immobilization except for pain control before symptoms are alleviated by treatment. Weight bearing should be avoided until signs of inflammation have subsided, but frequent passive motion of the joint is indicated to maintain full mobility. Although addition of glucocorticoids to antibiotic treatment improves the outcome of *S. aureus* arthritis in experimental animals, no clinical trials have evaluated this approach in humans.

GONOCOCCAL ARTHRITIS

Epidemiology

Although its incidence has declined in recent years, gonococcal arthritis **(Chap. 151)** has accounted for up to 70% of episodes of infectious arthritis in persons <40 years of age in the United States. Arthritis due to *N. gonorrhoeae* is a consequence of bacteremia arising from gonococcal infection or, more frequently, from asymptomatic gonococcal mucosal colonization of the urethra, cervix, or pharynx. Women are at greatest risk during menses and during pregnancy and overall are two to three times more likely than men to develop disseminated gonococcal infection (DGI) and arthritis. Persons with complement deficiencies, especially of the terminal components, are prone to recurrent episodes of gonococcemia. Strains of gonococci that are most likely to cause DGI include those which produce transparent colonies in culture, have the type IA outer-membrane protein, or are of the AUH-auxotroph type.

Clinical Manifestations and Laboratory Findings

The most common manifestation of DGI is a syndrome of fever, chills, rash, and articular symptoms. Small numbers of papules that progress to hemorrhagic pustules develop on the trunk and the extensor surfaces of the distal extremities. Migratory arthritis and tenosynovitis of the knees, hands, wrists, feet, and ankles are prominent. The cutaneous lesions and articular findings are believed to be the consequence of an immune reaction to circulating gonococci and immune-complex deposition in tissues. Thus, cultures of synovial fluid are consistently negative, and blood cultures are positive in fewer than 45% of patients. Synovial fluid may be difficult to obtain from inflamed joints and usually contains only 10,000–20,000 leukocytes/μL.

True gonococcal septic arthritis is less common than the DGI syndrome and always follows DGI, which is unrecognized in one-third of patients. A single joint such as the hip, knee, ankle, or wrist is usually involved. Synovial fluid, which contains >50,000 leukocytes/μL, can be obtained with ease; the gonococcus is evident only occasionally in Gram-stained smears, and cultures of synovial fluid are positive in fewer than 40% of cases. Blood cultures are almost always negative.

Because it is difficult to isolate gonococci from synovial fluid and blood, specimens for culture should be obtained from potentially infected mucosal sites. NAA-based urine tests also may be positive. Cultures and Gram-stained smears of skin lesions are occasionally positive. All specimens for culture should be plated onto Thayer-Martin agar directly or in special transport media at the bedside and transferred promptly to the microbiology laboratory in an atmosphere of 5% CO_2, as generated in a candle jar. NAA-based assays are extremely sensitive in detecting gonococcal DNA in synovial fluid. A dramatic alleviation of symptoms within 12–24 h after the initiation of appropriate antibiotic therapy supports a clinical diagnosis of the DGI syndrome if cultures are negative.

TREATMENT

TREATMENT

Gonococcal Arthritis

Initial treatment consists of ceftriaxone (1 g IV or IM every 24 h) to cover possible penicillin-resistant organisms. Once local and systemic signs are clearly resolving, the 7-day course of therapy can be completed with an oral fluoroquinolone such as ciprofloxacin (500 mg twice daily) if the organism is known to be susceptible. If penicillin-susceptible organisms are isolated, amoxicillin (500 mg three times daily) may be used. Suppurative arthritis usually responds to needle aspiration of involved joints and 7–14 days of antibiotic treatment. Arthroscopic lavage or arthrotomy is rarely required. Patients with DGI should be treated for *Chlamydia trachomatis* infection unless this infection is ruled out by appropriate testing. Addition of azithromycin (1 g orally as a single dose) is recommended to treat chlamydial co-infection, which is common. Sexual partners should be offered testing and presumptive treatment for gonorrhea and chlamydial infection.

It is noteworthy that arthritis symptoms similar to those seen in DGI occur in meningococcemia. A dermatitis–arthritis syndrome, purulent monarthritis, and reactive polyarthritis have been described. All respond to treatment with IV penicillin.

SPIROCHETAL ARTHRITIS

LYME DISEASE

Lyme disease **(Chap. 181)** due to infection with the spirochete *Borrelia burgdorferi* causes arthritis in up to 60% of persons who are not treated. Intermittent arthralgias and myalgias—but not arthritis—occur within days or weeks of inoculation of the spirochete by the *Ixodes* tick. Later, there are three patterns of joint disease: (1) Fifty percent of untreated persons experience intermittent episodes of monarthritis or oligoarthritis involving the knee and/or other large joints. The symptoms wax and wane without treatment over months, and each year 10–20% of patients report loss of joint symptoms. (2) Twenty percent of untreated persons develop a pattern of waxing and waning arthralgias. (3) Ten percent of untreated patients develop chronic inflammatory synovitis that results in erosive lesions and destruction of the joint. Serologic tests for IgG antibodies to *B. burgdorferi* are positive in more than 90% of persons with Lyme arthritis, and an NAA-based assay detects *Borrelia* DNA in 85%.

TREATMENT

TREATMENT

Lyme Arthritis

Lyme arthritis generally responds well to therapy. A regimen of oral doxycycline (100 mg twice daily for 28 days), oral amoxicillin (500 mg three times daily for 28 days), or parenteral ceftriaxone (2 g/d for 2–4 weeks) is recommended. Patients who do not respond to a total of 2 months of oral therapy or 1 month of parenteral therapy are unlikely to benefit from additional antibiotic therapy and are treated with anti-inflammatory

agents or synovectomy. Failure of therapy is associated with host features such as the human leukocyte antigen DR4 (HLA-DR4) genotype, persistent reactivity to OspA (outer-surface protein A), and the presence of hLFA-1 (human leukocyte function–associated antigen 1), which cross-reacts with OspA.

SYPHILITIC ARTHRITIS

Articular manifestations occur in different stages of syphilis **(Chap. 177)**. In early congenital syphilis, periarticular swelling and immobilization of the involved limbs (*Parrot's pseudoparalysis*) complicate osteochondritis of long bones. *Clutton's joint*, a late manifestation of congenital syphilis that typically develops between ages 8 and 15 years, is caused by chronic painless synovitis with effusions of large joints, particularly the knees and elbows. Secondary syphilis may be associated with arthralgias, with symmetric arthritis of the knees and ankles and occasionally of the shoulders and wrists, and with sacroiliitis. The arthritis follows a subacute to chronic course with a mixed mononuclear and neutrophilic synovial-fluid pleocytosis (typical cell counts, 5000–15,000/μL). Immunologic mechanisms may contribute to the arthritis, and symptoms usually improve rapidly with penicillin therapy. In tertiary syphilis, Charcot joint results from sensory loss due to tabes dorsalis. Penicillin is not helpful in this setting.

MYCOBACTERIAL ARTHRITIS

Tuberculous arthritis **(Chap. 173)** accounts for ~1% of all cases of tuberculosis and 10% of extrapulmonary cases. The most common presentation is chronic granulomatous monarthritis. An unusual syndrome, *Poncet's disease*, is a reactive symmetric form of polyarthritis that affects persons with visceral or disseminated tuberculosis. No mycobacteria are found in the joints, and symptoms resolve with antituberculous therapy.

Unlike tuberculous osteomyelitis **(Chap. 126)**, which typically involves the thoracic and lumbar spine (50% of cases), tuberculous arthritis primarily involves the large weight-bearing joints, in particular the hips, knees, and ankles, and only occasionally involves smaller non-weight-bearing joints. Progressive monarticular swelling and pain develop over months or years, and systemic symptoms are seen in only half of all cases. Tuberculous arthritis occurs as part of a disseminated primary infection or through late reactivation, often in persons with HIV infection or other immunocompromised hosts. Coexistent active pulmonary tuberculosis is unusual.

Aspiration of the involved joint yields fluid with an average cell count of 20,000/μL, with ~50% neutrophils. Acid-fast staining of the fluid yields positive results in fewer than one-third of cases, and cultures are positive in 80%. Culture of synovial tissue taken at biopsy is positive in ~90% of cases and shows granulomatous inflammation in most. NAA methods can shorten the time to diagnosis to 1 or 2 days. Radiographs reveal peripheral erosions at the points of synovial attachment, periarticular osteopenia, and eventually joint-space narrowing. Therapy for tuberculous arthritis is the same as that for tuberculous pulmonary disease, requiring the administration of multiple agents for 6–9 months. Therapy is more prolonged in immunosuppressed individuals, such as those infected with HIV.

Various atypical mycobacteria **(Chap. 175)** found in water and soil may cause chronic indolent arthritis. Such disease results from trauma and direct inoculation associated with farming, gardening, or aquatic activities. Smaller joints, such as the digits, wrists, and knees, are usually involved. Involvement of tendon sheaths and bursae is typical. The mycobacterial species involved include *Mycobacterium marinum*, *M. avium-intracellulare*, *M. terrae*, *M. kansasii*, *M. fortuitum*, and *M. chelonae*. In persons who have HIV infection or are receiving immunosuppressive therapy, hematogenous spread to the joints has been reported for *M. kansasii*, *M. avium* complex, and *M. haemophilum*. Diagnosis usually requires biopsy and culture, and therapy is based on antimicrobial susceptibility patterns.

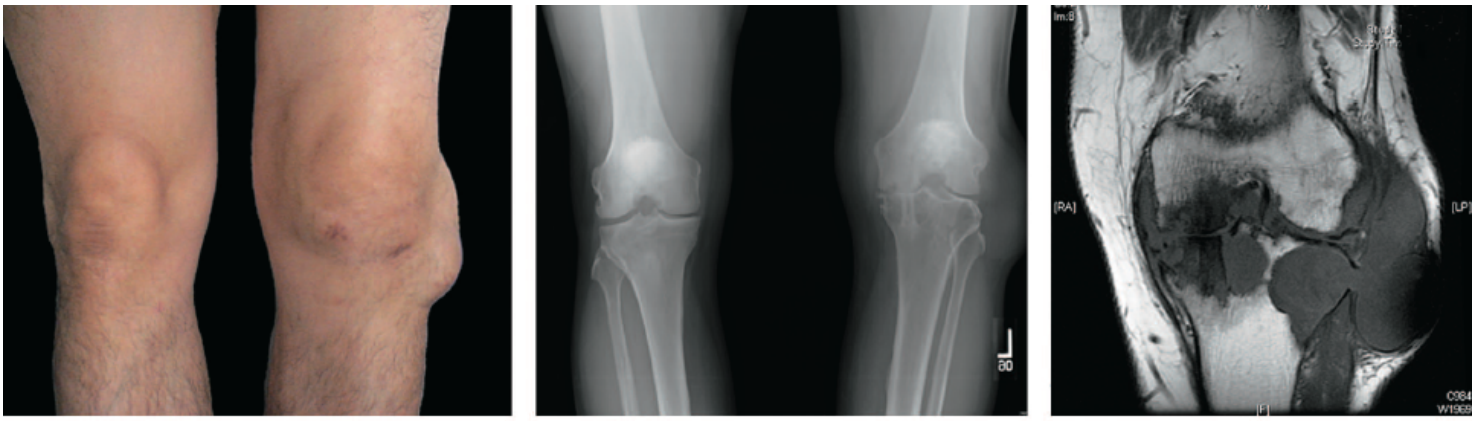
FUNGAL ARTHRITIS

Fungi are an unusual cause of chronic monarticular arthritis. Granulomatous articular infection with the endemic dimorphic fungi *Coccidioides immitis*, *Blastomyces dermatitidis*, and (less commonly) *Histoplasma capsulatum* **(Fig. 125-2)** results from hematogenous seeding or direct extension from bony lesions in persons with disseminated disease. Joint involvement is an unusual complication of sporotrichosis (infection with *Sporothrix schenckii*) among gardeners and other persons who work with soil or sphagnum moss. Articular sporotrichosis is six times more common among men than among women, and alcoholics and other debilitated hosts are at risk for polyarticular infection.

FIGURE 125-2

Chronic arthritis caused by *Histoplasma capsulatum* in the left knee. *A.* A man in his sixties from El Salvador presented with a history of progressive knee pain and difficulty walking for several years. He had undergone arthroscopy for a meniscal tear 7 years before presentation (without relief) and had received several intraarticular glucocorticoid injections. The patient developed significant deformity of the knee over time, including a large effusion in the lateral aspect. ***B.*** An x-ray of the knee showed multiple abnormalities, including severe medial

femorotibial joint-space narrowing, several large subchondral cysts within the tibia and the patellofemoral compartment, a large suprapatellar joint effusion, and a large soft-tissue mass projecting laterally over the knee. ***C.*** MRI further defined these abnormalities and demonstrated the cystic nature of the lateral knee abnormality. Synovial biopsies demonstrated chronic inflammation with giant cells, and cultures grew *H. capsulatum* after 3 weeks of incubation. All clinical cystic lesions and the effusion resolved after 1 year of treatment with itraconazole. The patient underwent a left total-knee replacement for definitive treatment. *(Courtesy of Francisco M. Marty, MD, Brigham and Women's Hospital, Boston; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Candida infection involving a single joint—usually the knee, hip, or shoulder—results from surgical procedures, intraarticular injections, or (among critically ill patients with debilitating illnesses such as diabetes mellitus or hepatic or renal insufficiency and patients receiving immunosuppressive therapy) hematogenous spread. *Candida* infections in IV drug users typically involve the spine, sacroiliac joints, or other fibrocartilaginous joints. Unusual cases of arthritis due to *Aspergillus* species, *Cryptococcus neoformans*, *Pseudallescheria boydii*, and the dematiaceous fungi also have resulted from direct inoculation or disseminated hematogenous infection in immunocompromised persons. In the United States, a 2012 national outbreak of fungal arthritis (and meningitis) caused by *Exserohilum rostratum* was linked to intraspinal and intraarticular injection of a contaminated preparation of methylprednisolone acetate.

The synovial fluid in fungal arthritis usually contains 10,000–40,000 cells/μL, with ~70% neutrophils. Stained specimens and cultures of synovial tissue often confirm the diagnosis of fungal arthritis when studies of synovial fluid give negative results. Treatment consists of drainage and lavage of the joint and systemic administration of an antifungal agent directed at a specific pathogen. The doses and duration of therapy are the same as for disseminated disease **(see Part 5, Section 16)**. Intraarticular instillation of amphotericin B has been used in addition to IV therapy.

VIRAL ARTHRITIS

Viruses produce arthritis by infecting synovial tissue during systemic infection or by provoking an immunologic reaction that involves joints. As many as 50% of women report persistent arthralgias and 10% report frank arthritis within 3 days of the rash that follows natural infection with rubella virus and within 2–6 weeks after receipt of live-virus vaccine. Episodes of symmetric inflammation of fingers, wrists, and knees uncommonly recur for >1 year, but a syndrome of chronic fatigue, low-grade fever, headaches, and myalgias can persist for months or years. IV immunoglobulin has been helpful in selected cases. Self-limited monarticular or migratory polyarthritis may develop within 2 weeks of the parotitis of mumps; this sequela is more common among men than among women. Approximately 10% of children and 60% of women develop arthritis after infection with parvovirus B19. In adults, arthropathy sometimes occurs without fever or rash. Pain and stiffness, with less prominent swelling (primarily of the hands but also of the knees, wrists, and ankles), usually resolve within weeks, although a small proportion of patients develop chronic arthropathy.

About 2 weeks before the onset of jaundice, up to 10% of persons with acute hepatitis B develop an immune complex–mediated, serum sickness–like reaction with maculopapular rash, urticaria, fever, and arthralgias. Less common developments include symmetric arthritis involving the hands, wrists, elbows, or ankles and morning stiffness that resembles a flare of rheumatoid arthritis. Symptoms resolve at the time jaundice develops. Many persons with chronic hepatitis C infection report persistent arthralgia or arthritis, both in the presence and in the absence of cryoglobulinemia.

Painful arthritis involving larger joints often accompanies the fever and rash of several arthropod-borne viral infections, including those caused by Zika, chikungunya, O'nyong-nyong, Ross River, Mayaro, and Barmah Forest viruses **(Chap. 204)**. Symmetric arthritis involving the

hands and wrists may occur during the convalescent phase of infection with lymphocytic choriomeningitis virus. Patients infected with an enterovirus frequently report arthralgias, and echovirus has been isolated from patients with acute polyarthritis.

Several arthritis syndromes are associated with HIV infection. Reactive arthritis with painful lower-extremity oligoarthritis often follows an episode of urethritis in HIV-infected persons. HIV-associated reactive arthritis appears to be extremely common among persons with the HLA-B27 haplotype, but sacroiliac joint disease is unusual and is seen mostly in the absence of HLA-B27. Up to one-third of HIV-infected persons with psoriasis develop psoriatic arthritis. Painless monarthropathy and persistent symmetric polyarthropathy occasionally complicate HIV infection. Chronic persistent oligoarthritis of the shoulders, wrists, hands, and knees occurs in women infected with human T-lymphotropic virus type 1. Synovial thickening, destruction of articular cartilage, and leukemic-appearing atypical lymphocytes in synovial fluid are characteristic, but progression to T cell leukemia is unusual.

PARASITIC ARTHRITIS

Arthritis due to parasitic infection is rare. The guinea worm *Dracunculus medinensis* may cause destructive joint lesions in the lower extremities as migrating gravid female worms invade joints or cause ulcers in adjacent soft tissues that become secondarily infected. Hydatid cysts infect bones in 1–2% of cases of infection with *Echinococcus granulosus*. The expanding destructive cystic lesions may spread to and destroy adjacent joints, particularly the hip and pelvis. In rare cases, chronic synovitis has been associated with the presence of schistosomal eggs in synovial biopsies. Monarticular arthritis in children with lymphatic filariasis appears to respond to therapy with diethylcarbamazine even in the absence of microfilariae in synovial fluid. Reactive arthritis has been attributed to hookworm, *Strongyloides*, *Cryptosporidium*, and *Giardia* infection in case reports, but confirmation is required.

POSTINFECTIOUS OR REACTIVE ARTHRITIS

Reactive polyarthritis develops several weeks after ~1% of cases of nongonococcal urethritis and 2% of enteric infections, particularly those due to *Yersinia enterocolitica*, *Shigella flexneri*, *Campylobacter jejuni*, and *Salmonella* species. Only a minority of these patients have the other findings of classic reactive arthritis, including urethritis, conjunctivitis, uveitis, oral ulcers, and rash. Studies have identified microbial DNA or antigen in synovial fluid or blood, but the pathogenesis of this condition is poorly understood.

Reactive arthritis is most common among young men (except after *Yersinia* infection) and has been linked to the HLA-B27 locus as a potential genetic predisposing factor. Patients report painful, asymmetric oligoarthritis that affects mainly the knees, ankles, and feet. Low-back pain is common, and radiographic evidence of sacroiliitis is found in patients with long-standing disease. Most patients recover within 6 months, but prolonged recurrent disease is more common in cases that follow chlamydial urethritis. Anti-inflammatory agents help relieve symptoms, but the role of prolonged antibiotic therapy in eliminating microbial antigen from the synovium is controversial.

Migratory polyarthritis and fever constitute the usual presentation of acute rheumatic fever in adults (**Chap. 352**). This presentation is distinct from that of poststreptococcal reactive arthritis, which also follows infections with group A *Streptococcus* but is not migratory, lasts beyond the typical 3-week maximum of acute rheumatic fever, and responds poorly to aspirin.

INFECTIONS IN PROSTHETIC JOINTS

Infection complicates 1–4% of total joint replacements. The majority of infections are acquired intraoperatively or immediately postoperatively as a result of wound breakdown or infection; less commonly, these joint infections develop later after joint replacement and are the result of hematogenous spread or direct inoculation. The presentation may be acute, with fever, pain, and local signs of inflammation, especially in infections due to *S. aureus*, pyogenic streptococci, and enteric bacilli. Alternatively, infection may persist for months or years without causing constitutional symptoms when less virulent organisms, such as coagulase-negative staphylococci or diphtheroids, are involved. Such indolent infections usually are acquired during joint implantation and are discovered during evaluation of chronic unexplained pain or after a radiograph shows loosening of the prosthesis; the erythrocyte sedimentation rate and C-reactive protein level are usually elevated in such cases.

The diagnosis is best made by needle aspiration of the joint; accidental introduction of organisms during aspiration must be avoided meticulously. Synovial fluid pleocytosis with a predominance of polymorphonuclear leukocytes is highly suggestive of infection, since other inflammatory processes uncommonly affect prosthetic joints. Culture and Gram's stain usually yield the responsible pathogen. Sonication of explanted prosthetic material can improve the yield of culture, presumably by breaking up bacterial biofilms on the surfaces of prostheses. Use of special media for unusual pathogens such as fungi, atypical mycobacteria, and *Mycoplasma* may be necessary if routine and anaerobic cultures are negative.

TREATMENT

TREATMENT

Prosthetic Joint Infections

Treatment includes surgery and high doses of parenteral antibiotics, which are given for 4–6 weeks because bone is usually involved. In most cases, the prosthesis must be replaced to cure the infection. Implantation of a new prosthesis is best delayed for several weeks or months because relapses of infection occur most commonly within this time frame. In some cases, reimplantation is not possible, and the patient must manage without a joint, with a fused joint, or even with amputation. Cure of infection without removal of the prosthesis is occasionally possible in cases that are due to streptococci or pneumococci and that lack radiologic evidence of loosening of the prosthesis. In these cases, antibiotic therapy must be initiated within several days of the onset of infection, and the joint should be drained vigorously by open arthrotomy or arthroscopically. In selected patients who prefer to avoid the high morbidity rate associated with joint removal and reimplantation, suppression of the infection with antibiotics may be a reasonable goal. A high cure rate with retention of the prosthesis has been reported when the combination of oral rifampin and another antibiotic (e.g., a quinolone, an antistaphylococcal penicillin, or vancomycin) is given for 3–6 months to persons with staphylococcal prosthetic joint infection of short duration. This approach, which is based on the ability of rifampin to kill organisms adherent to foreign material and in the stationary growth phase, requires confirmation in prospective trials.

PREVENTION

To avoid the disastrous consequences of infection, candidates for joint replacement should be selected with care. Rates of infection are particularly high among patients with rheumatoid arthritis, persons who have undergone previous surgery on the joint, and persons with medical conditions requiring immunosuppressive therapy. Perioperative antibiotic prophylaxis, usually with cefazolin, and measures to decrease intraoperative contamination, such as laminar flow, have lowered the rates of perioperative infection to <1% in many centers. After implantation, measures should be taken to prevent or rapidly treat extra-articular infections that might give rise to hematogenous spread to the prosthesis. The effectiveness of prophylactic antibiotics for the prevention of hematogenous infection after dental procedures has not been demonstrated; in fact, viridans streptococci and other components of the oral flora are extremely unusual causes of prosthetic joint infection. Accordingly, the American Dental Association and the American Academy of Orthopaedic Surgeons do not recommend antibiotic prophylaxis for most dental patients with total joint replacements and have stated that there is no convincing evidence to support its use. Similarly, guidelines issued by the American Urological Association and the American Academy of Orthopaedic Surgeons do not recommend the use of prophylactic antibiotics for most patients with prosthetic joints who are undergoing urologic procedures but state that prophylaxis should be considered in certain situations—e.g., for patients (especially immunocompromised patients) who are undergoing a procedure posing a relatively high risk of bacteremia (such as lithotripsy or surgery involving bowel segments).

ACKNOWLEDGMENTS

The contributions of James H. Maguire and the late Scott J. Thaler to this chapter in earlier editions are gratefully acknowledged.

FURTHER READING

Bardin T: Gonococcal arthritis. Best Pract Res Clin Rheumatol 17:201, 2003. [PubMed: 12787521]

Borzio R et al: Predictors of septic arthritis in the adult population. Orthopedics 39:e657, 2016. [PubMed: 27286047]

Franssila R, Hedman K: Infection and musculoskeletal conditions: Viral causes of arthritis. Best Pract Res Clin Rheumatol 20:1139, 2006. [PubMed: 17127201]

Harrington JT: Mycobacterial and fungal arthritis. Curr Opin Rheumatol 10:335, 1998. [PubMed: 9725095]

Meehan AM et al: Outcome of penicillin-susceptible streptococcal prosthetic joint infection treated with debridement and retention of the prosthesis. Clin Infect Dis 36:845, 2003. [PubMed: 12652384]

Zeller V et al: One-stage exchange arthroplasty for chronic periprosthetic hip infection: Results of a large prospective cohort study. J Bone Joint Surg Am 96:e1, 2014. [PubMed: 24382729]

Zimmerli W et al: Prosthetic-joint infections. N Engl J Med 351:145, 2004. [PubMed: 15247354]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 126: Osteomyelitis

Werner Zimmerli

INTRODUCTION

Osteomyelitis, an infection of bone, can be caused by various microorganisms that arrive at bone through different routes. Spontaneous hematogenous osteomyelitis may occur in otherwise healthy individuals, whereas local microbial spread mainly affects either individuals who have underlying disease (e.g., vascular insufficiency) or patients who have compromised skin or other tissue barriers, with consequent exposure of bone. The latter situation typically follows surgery involving bone, such as sternotomy or orthopedic repair.

The manifestations of osteomyelitis are different in children and adults. In children circulating microorganisms seed mainly long bones, whereas in adults the vertebral column is the most commonly affected site.

Management of osteomyelitis differs greatly depending on whether an implant is involved. The most important aim of the management of either type of osteomyelitis is to prevent progression to chronic osteomyelitis by rapid diagnosis and prompt treatment. Device-related bone and joint infection necessitates a multidisciplinary approach requiring antibiotic therapy and, in many cases, surgical removal of the device. For most types of osteomyelitis, the optimal duration of antibiotic treatment has not been established in clinical trials. Therefore, the recommendations for therapy in this chapter reflect mainly expert opinions.

CLASSIFICATION

There is no generally accepted, comprehensive system for classification of osteomyelitis, primarily because of the multifaceted presentation of this infection. Different specialists are confronted with different facets of bone disease. Most often, however, general practitioners or internists are the first to encounter patients with the initial signs and symptoms of osteomyelitis. These primary care physicians should be able to recognize this disease in any of its forms. Osteomyelitis cases can be classified by various criteria, including pathogenesis, duration of infection, location of infection, and presence or absence of foreign material. The widely used Cierny-Mader staging system classifies osteomyelitis according to anatomic site, comorbidity, and radiographic findings, with stratification of long-bone osteomyelitis to optimize surgical management; this system encompasses both systemic and local factors affecting immune status, metabolism, and local vascularity.

Any of three mechanisms can underlie osteomyelitis: (1) hematogenous spread; (2) spread from a contiguous site following surgery; and (3) secondary infection in the setting of vascular insufficiency or concomitant neuropathy. Hematogenous osteomyelitis in adults typically involves the vertebral column. In only about half of patients a primary focus can be detected. The most common primary foci of infection are the urinary tract, skin/soft tissue, intravascular catheterization sites, and the endocardium. Spread from a contiguous source follows either bone trauma or surgical intervention. Wound infection leading to osteomyelitis typically occurs after cardiovascular intervention involving the sternum, orthopedic repair after open fracture, or prosthetic joint insertion. Osteomyelitis secondary to

vascular insufficiency or peripheral neuropathy most often follows chronic, progressively deep skin and soft tissue infection of the foot. The most common underlying condition is diabetes. In diabetes that is poorly controlled, the *diabetic foot syndrome* is caused by skin, soft tissue, and bone ischemia combined with motor, sensory, and autonomic neuropathy.

Classification of osteomyelitis according to the duration of infection, although ill defined, is useful because the management of acute and chronic osteomyelitis differs. Whereas acute osteomyelitis can generally be treated with antibiotics alone, antibiotic treatment for chronic osteomyelitis should be combined with debridement surgery. Acute hematogenous or contiguous osteomyelitis evolves over a short period—i.e., a few days or weeks. In contrast, subacute or chronic osteomyelitis lasts for weeks or months before treatment is started. Typical examples of a subacute course are vertebral osteomyelitis due to tuberculosis or brucellosis and delayed implant-associated infections caused mainly by low-virulence microorganisms (coagulase-negative staphylococci, *Propionibacterium acnes*). Chronic osteomyelitis develops when insufficient therapy leads to persistence or recurrence, most often after sternal, mandibular, or foot infection.

Classification by location distinguishes among cases in the long bones, the vertebral column, and the periarticular bones. Long bones are generally involved after hematogenous seeding in children or contiguous spread following trauma or surgery. The risk of vertebral osteomyelitis in adults increases with age. Periarticular osteomyelitis, which complicates septic arthritis that has not been adequately treated, is especially common in periprosthetic joint infection.

Osteomyelitis involving a foreign device requires surgical management for cure. Even acute implant-associated infection calls for prolonged antimicrobial therapy. Therefore, identification of this type of disease is of practical importance.

VERTEBRAL OSTEOMYELITIS

PATHOGENESIS

Vertebral osteomyelitis, also referred to as disk-space infection, septic diskitis, spondylodiskitis, or spinal osteomyelitis, is the most common manifestation of hematogenous bone infection in adults. This designation reflects a pathogenic process leading to involvement of the adjacent vertebrae and the corresponding intervertebral disk. In adults, the disk is avascular. Microorganisms invade via the segmental arterial circulation in adjacent endplates and then spread into the disk. Alternative routes of infection are retrograde seeding through the prevertebral venous plexus and direct inoculation during spinal surgery, epidural infiltration, or trauma. In the setting of implant surgery, microorganisms are inoculated either during the procedure or, if wound healing is impaired, in the early postoperative period.

EPIDEMIOLOGY

Vertebral osteomyelitis occurs more often in male than in female patients (ratio, 1.5:1). Between 1995 and 2008, the incidence rate increased from 2.2 to 5.8 cases/100,000 person-years. There is a clear age-dependent increase from 0.3 case/100,000 at ages <20 years to 6.5 cases/100,000 at ages >70 years. The observed increase in reported cases during the past two decades may reflect improvements in diagnosis resulting from the broad availability of MRI technology. In addition, the fraction of cases of vertebral osteomyelitis acquired in association with health care is increasing as a consequence of comorbidity and the rising number of invasive interventions.

MICROBIOLOGY

Vertebral osteomyelitis is typically classified as pyogenic or nonpyogenic. However, this distinction is arbitrary: in “nonpyogenic” cases (tuberculous, brucellar), macroscopic pus formation (caseous necrosis, abscess) is quite common. A more accurate scheme is to classify cases as acute or subacute/chronic. Whereas the microbiologic spectrum of acute cases is similar in different parts of the world, the spectrum of subacute/chronic cases varies according to the geographic region. The great majority of cases are monomicrobial in etiology. Of episodes of acute vertebral osteomyelitis, 40–50% are caused by *Staphylococcus aureus*, 12% by streptococci, and 20% by gram-negative bacilli—mainly *Escherichia coli* (9%) and *Pseudomonas aeruginosa* (6%). Subacute vertebral osteomyelitis is typically caused by *Mycobacterium tuberculosis* or *Brucella* species in regions where these microorganisms are endemic. Osteomyelitis due to viridans streptococci also has a subacute presentation; these infections most often occur as secondary foci in patients with endocarditis. In vertebral osteomyelitis due to *Candida* species, the diagnosis is often delayed by several weeks; this etiology should be suspected in IV drug users who do not use sterile paraphernalia. In implant-associated spinal osteomyelitis, coagulase-negative staphylococci and *P. acnes*—which, in the absence of an implant, are generally considered contaminants—typically cause low-grade (chronic) infections. As an exception, coagulase-negative staphylococci can cause native spinal osteomyelitis in cases of prolonged bacteremia (e.g., in patients with infected pacemaker electrodes or implanted vascular catheters that are not promptly removed).

CLINICAL MANIFESTATIONS

The signs and symptoms of vertebral osteomyelitis are nonspecific. Only about half of patients develop fever >38°C (>100.4°F), perhaps because patients frequently use analgesic drugs. Back pain is the leading initial symptom (>85% of cases). The location of the pain corresponds to the site of infection: the cervical spine in ~10% of cases, the thoracic spine in 30%, and the lumbar spine in 60%. One exception is involvement at the thoracic level in two-thirds of cases of tuberculous osteomyelitis and at the lumbar level in only one-third. This difference is due to direct mycobacterial spread via pleural or mediastinal lymph nodes in pulmonary tuberculosis.

Neurologic deficits, such as radiculopathy, weakness, or sensory loss, are observed in about one-third of cases of vertebral osteomyelitis. In brucellar vertebral osteomyelitis, neurologic impairment is less common; in tuberculous osteomyelitis, it is about twice as common as in cases of other etiologies. Neurologic signs and symptoms are caused mostly by spinal epidural abscess. This complication starts with severe localized back pain and progresses to radicular pain, reflex changes, sensory abnormalities, motor weakness, bowel and bladder dysfunction, and paralysis.

A primary focus should always be sought but is found in only half of cases. Overall, endocarditis is identified in ~10% of patients. In osteomyelitis caused by viridans streptococci, endocarditis is the source in about half of patients.

Implant-associated spinal osteomyelitis can present as either early- or late-onset infection. Early-onset infection is diagnosed within 30 days after implant placement. *S. aureus* is the most common pathogen. Wound healing impairment and fever are the leading findings. Late-onset infection is diagnosed beyond 30 days after surgery, with low-virulence organisms such as coagulase-negative staphylococci or *P. acnes* as typical infecting agents. Fever is rare. One-quarter of patients have a sinus tract. Because of the delayed course and the lack of classic signs of infection, rapid diagnosis requires a high degree of suspicion.

DIAGNOSIS

Leukocytosis and neutrophilia have low levels of diagnostic sensitivity (only 65 and 40%, respectively). In contrast, an increased erythrocyte sedimentation rate or C-reactive protein (CRP) level has been reported in 98 and 100% of cases, respectively; thus, these tests are helpful in excluding vertebral

osteomyelitis. The fraction of blood cultures that yield positive results depends heavily on whether the patient has been pretreated with antibiotics; across studies, the range is from 30 to 78%. In view of this low rate of positive blood culture after antibiotic treatment, such therapy should be withheld until microbial growth is proven unless the patient has sepsis syndrome. In patients with negative blood cultures, CT-guided or open biopsy is needed. Whether a CT-guided biopsy with a negative result is repeated or followed by open biopsy depends on the experience of personnel at the specific center. Bone samples should be cultured for aerobic, anaerobic, and fungal agents, with a portion of the sample sent for histopathologic study. In cases with a subacute/chronic presentation, a suggestive history, or a granuloma detected during histopathologic analysis, mycobacteria and brucellae also should be sought. When blood and tissue cultures are negative despite suggestive histopathology, broad-range polymerase chain reaction analysis of biopsy specimens or aspirated pus should be considered. This technique allows detection of unusual pathogens such as *Tropheryma whipplei*.

Given that signs and symptoms of osteomyelitis are nonspecific, the clinical differential diagnosis of febrile back pain is broad, including pyelonephritis, pancreatitis, and viral syndromes. In addition, multiple noninfectious pathologies of the vertebral column, such as osteoporotic fracture, seronegative spondylitis (ankylosing spondylitis, psoriasis, reactive arthritis, enteropathic arthritis), and spinal stenosis must be considered.

Imaging procedures are the most important tools not only for the diagnosis of vertebral osteomyelitis but also for the detection of pyogenic complications and alternative conditions (e.g., bone metastases or osteoporotic fractures). Plain radiography is a reasonable first step in evaluating patients without neurologic symptoms and may reveal an alternative diagnosis. Because of its low sensitivity, plain radiography generally is not helpful in acute osteomyelitis, but it can be useful in subacute or chronic cases. The gold standard is MRI, which should be performed expeditiously in patients with neurologic impairment in order to rule out a herniated disk or to detect pyogenic complications in a timely manner. Even if the pathologic findings on MRI suggest vertebral osteomyelitis, alternative diagnoses should be considered, especially when blood cultures are negative. The most common alternative diagnosis is erosive osteochondrosis. Septic bone necrosis, gouty spondylodiskitis, and erosive diskovertebral lesions (Andersson lesions) in ankylosing spondylitis may likewise mimic vertebral osteomyelitis. CT is less sensitive than MRI but may be helpful in guiding a percutaneous biopsy. In the future, positron-emission tomography (PET) with ^{18}F-fluorodeoxyglucose, which has a high degree of diagnostic accuracy, may be an alternative imaging procedure when MRI is contraindicated. ^{18}F-fluorodeoxyglucose PET should be considered for patients with implants and patients in whom several foci are suspected.

TREATMENT

TREATMENT Vertebral Osteomyelitis

The aims of therapy for vertebral osteomyelitis are (1) elimination of the pathogen(s), (2) protection from further bone loss, (3) relief of back pain, (4) prevention of complications, and (5) stabilization, if needed.

Table 126-1 summarizes suggested antimicrobial regimens for infections attributable to the most common etiologic agents. For optimal antimicrobial therapy, identification of the infecting agent is required. Therefore, in patients without sepsis syndrome, antibiotics should not be administered until the pathogen is identified in a blood culture, a bone biopsy, or an aspirated pus collection. Traditionally, bone infections are at least initially treated by the IV route. Unfortunately, relevant controlled trials are lacking, and the preference for the IV route is not evidence based. There are no good arguments for the assumption that IV therapy is superior to oral administration if the following requirements are met: (1) optimal antibiotic spectrum, (2) excellent bioavailability of the oral drug, (3) clinical studies confirming efficacy of the oral drug, (4) normal intestinal function, and (5) no vomiting. However, a short initial course of parenteral therapy with a β-lactam antibiotic may lower the risk of

emergence of fluoroquinolone resistance, especially if *P. aeruginosa* infection is treated with ciprofloxacin or staphylococcal infection with the combination of a fluoroquinolone plus rifampin. These suggestions are based on observational studies and expert opinion. A recent randomized, controlled trial showed that 6 weeks of antibiotic treatment is not inferior to a 12-week course in patients with pyogenic vertebral osteomyelitis. The cure rate was 90.9% in both groups 1 year after therapy. Thus, prolonged antibiotic therapy is required only for patients with undrained abscesses and for patients with spinal implants. Treatment efficacy should be regularly monitored through inquiries about signs and symptoms (fever, pain) and assessment for signs of inflammation (elevated CRP concentrations). Follow-up MRI is appropriate only for patients with pyogenic complications, since the correlation between clinical healing and improvement on MRI is very poor.

Surgical treatment generally is not needed in acute hematogenous vertebral osteomyelitis. However, it is always necessary in implant-associated spinal infection. Early infections (those occurring up to 30 days after internal stabilization) can be cured with debridement, implant retention, and a 3-month course of antibiotics (**Table 126-2**). In contrast, in late infection with a duration of >30 days, implant removal and a 6-week course of antibiotics (Table 126-1) are required for complete elimination of the infection. If implants cannot be removed, oral suppressive long-term treatment should follow the initial course of IV antibiotics. The optimal duration of suppressive therapy is unknown. However, if antibiotic therapy is discontinued after, for example, 1 year, close clinical and laboratory (CRP) follow-up is needed.

TABLE 126-1 Antibiotic Therapy for Osteomyelitis in Adults Without Implants[a]

MICROORGANISM	ANTIMICROBIAL AGENT (DOSE,[b] ROUTE)
Staphylococcus spp.	
Methicillin-susceptible	Nafcillin or oxacillin[c] (2 g IV q6h)
	followed by
	Rifampin (300–450 mg PO q12h) plus levofloxacin (750 mg PO q24h or 500 mg PO q12h)
Methicillin-resistant	Vancomycin[d] (15 mg/kg IV q12h) or daptomycin (>6–8 mg/kg IV q24h)
	followed by
	Rifampin (300–450 mg PO q12h)
	plus
	Levofloxacin (750 mg PO q24h or 500 mg PO q12h) or TMP-SMX[e] (1 double-strength tablet PO q8h) or fusidic acid (500 mg PO q8h)
Streptococcus spp.	Penicillin G[c] (5 million units IV q6h) or ceftriaxone (2 g IV q24h)
Enterobacteriaceae	
Quinolone-susceptible	Ciprofloxacin (750 mg PO q24h)
Quinolone-resistant[f]	Imipenem (500 mg IV q6h)
Pseudomonas aeruginosa	Cefepime or ceftazidime (2 g IV q8h) *plus* an aminoglycoside[g]
	or
	Piperacillin-tazobactam (4.5 g IV q8h) plus an aminoglycoside[g] for 2–4 weeks
	followed by
	Ciprofloxacin[h] (750 mg PO q12h)
Anaerobes	Clindamycin (600 mg IV q6–8h) for 2–4 weeks
	followed by
	Clindamycin[i] (300 mg PO q6h)

[a]Unless otherwise indicated, the total duration of antimicrobial treatment is generally 6 weeks. [b]All dosages are for adults with normal renal function. [c]When the patient has delayed-type penicillin hypersensitivity, cefuroxime (1.5 g IV q6–8h) can be administered. When the patient has immediate-type penicillin hypersensitivity, the penicillin should be replaced by vancomycin (1 g IV q12h). [d]Target vancomycin trough level: 15–20 μg/mL. [e]Trimethoprim-sulfamethoxazole. A double-strength tablet contains 160 mg of trimethoprim and 800 mg of sulfamethoxazole. [f]Including isolates producing extended-spectrum β-lactamase. [g]The need for addition of an aminoglycoside has not yet been proven. However, this addition may decrease the risk of emergence of resistance to the β-lactam. [h]The rationale for starting ciprofloxacin treatment only after pretreatment with a β-lactam is the increased risk of emergence of quinolone resistance in the presence of a heavy bacterial load. [i]Alternatively, penicillin G (5 million units IV q6h) or ceftriaxone (2 g IV q24h) can be used against gram-positive anaerobes (e.g., *Propionibacterium acnes*), and metronidazole (500 mg IV/PO q8h) can be used against gram-negative anaerobes (e.g., *Bacteroides* spp.).

Source: From W Zimmerli: N Engl J Med 362:1022, 2010. © Massachusetts Medical Society. Reprinted with permission.

TABLE 126-2 Antibiotic Therapy for Osteomyelitis Associated with Orthopedic Devices

MICROORGANISM	ANTIMICROBIAL AGENT[a] (DOSE, ROUTE)
Staphylococcus spp.	*Recommendation for initial treatment phase (2 weeks with implant)*
Methicillin-susceptible	Rifampin (450 mg PO/IV q12h[b]) ***plus*** Nafcillin or oxacillin[c] (2 g IV q6h)
Methicillin-resistant	Rifampin (450 mg PO/IV q12h[b]) ***plus*** Vancomycin (15 mg/kg IV q12h) or daptomycin (6–10 mg/kg IV q24h)
Staphylococcus spp.	*Recommendation after completion of initial treatment phase* Rifampin (450 mg PO q12h[b]) ***plus*** Levofloxacin (750 mg PO q24h or 500 mg PO q12h) or ciprofloxacin (750 mg PO q12h) or fusidic acid (500 mg PO q8h) or TMP-SMX[d] (1 double-strength tablet PO q8h) or minocycline (100 mg PO q12h) or linezolid (600 mg PO q12h) or clindamycin (1200–1350 mg/d PO in 3 or 4 divided doses)
Streptococcus spp.[e]	Penicillin G[c] (18–24 million units/d IV in 6 divided doses) or ceftriaxone (2 g IV q24h) for 4 weeks ***followed by*** Amoxicillin (750–1000 mg PO q6–8h) or clindamycin (1200–1350 mg/d PO in 3 or 4 divided doses)
Enterococcus spp.[f]	
Penicillin-susceptible	Penicillin G[c] (24 million units/d IV in 6 divided doses) *or* ampicillin or amoxicillin[g] (2 g IV q4–6h)
Penicillin-resistant	Vancomycin (15 mg/kg IV q12h) or daptomycin (6–10 mg/kg IV q24h) or linezolid (600 mg IV/PO q12h)
Enterobacteriaceae	A β-lactam selected in light of in vitro susceptibility profile for 2 weeks[h] ***followed by*** Ciprofloxacin (750 mg PO q12h)
Enterobacter spp.[i] and nonfermenters[j] (e.g., *Pseudomonas aeruginosa*)	Cefepime or ceftazidime (2 g IV q8h) *or* meropenem (1 g IV q8h[k]) for 2–4 weeks ***followed by*** Ciprofloxacin (750 mg PO q12h)
Propionibacterium spp.	Penicillin G[c] (18–24 million units/d IV in 6 divided doses) or clindamycin (600–900 mg IV q8h) for 2–4 weeks ***followed by*** Amoxicillin (750–1000 mg PO q6–8h) or clindamycin (1200–1350 mg/d PO in 3 or 4 divided doses)

MICROORGANISM	ANTIMICROBIAL AGENT[a] (DOSE, ROUTE)
Gram-negative anaerobes (e.g., *Bacteroides* spp.)	Metronidazole (500 mg IV/PO q8h)
Mixed bacteria (without methicillin-resistant staphylococci)	Ampicillin-sulbactam (3 g IV q6h) or amoxicillin-clavulanate[l] (2.2 g IV q6h) or piperacillin-tazobactam (4.5 g IV q8h) or imipenem (500 mg IV q6h) or meropenem (1 g IV q8h[k]) for 2–4 weeks ***followed by*** Individualized oral regimens chosen in light of antimicrobial susceptibility

[a]Antimicrobial agents should be chosen in light of the isolate's in vitro susceptibility, the patient's drug allergies and intolerances, potential drug interactions, and contraindications to specific drugs. All dosages recommended are for adults with normal renal and hepatic function. See text for total durations of antibiotic treatment. [b]Other dosages and intervals of administration with equivalent success rates have been reported. [c]When the patient has delayed-type penicillin hypersensitivity, cefazolin (2 g IV q8h) can be administered. When the patient has immediate-type penicillin hypersensitivity, the penicillin should be replaced by vancomycin (1 g IV q12h). [d]Trimethoprim-sulfamethoxazole. A double-strength tablet contains 160 mg of trimethoprim and 800 mg of sulfamethoxazole. [e]Determination of the minimal inhibitory concentration (MIC) of penicillin is advisable. [f]Combination therapy with an aminoglycoside is optional since its superiority to monotherapy for prosthetic joint infection is unproved. When using combination therapy, monitor for signs of aminoglycoside ototoxicity and nephrotoxicity; the latter is potentiated by other nephrotoxic agents (e.g., vancomycin). [g]For patients with hypersensitivity to penicillin, see treatment options for penicillin-resistant enterococci. [h]Ciprofloxacin (PO or IV) can be administered to patients with hypersensitivity to β-lactams. [i]Ceftriaxone and ceftazidime should not be administered for treatment targeting *Enterobacter* species, even strains that test susceptible in the laboratory, but can be used against nonfermenters. Strains producing extended-spectrum β-lactamases should not be treated with any cephalosporin, including cefepime. *Enterobacter* infections can also be treated with ertapenem (1 g IV q24h); however, ertapenem is not effective against *Pseudomonas* spp. and other nonfermenters. [j]Addition of an aminoglycoside is optional. Use of two active drugs can be considered in light of the patient's clinical condition. [k]The recommended dosage is in line with the guidelines of the Infectious Diseases Society of America. In Europe, 2 g IV q8h is suggested for *P. aeruginosa* infections. [l]Not available as an IV formulation in the United States.

Source: Modified from W Zimmerli et al: N Engl J Med 351:1645, 2004. © Massachusetts Medical Society. Reprinted with permission.

COMPLICATIONS

Complications should be suspected when there is persistent pain, a persistently increased CRP level, and new-onset or persistent neurologic impairment. In cases of persistent pain with or without signs of inflammation, paravertebral, epidural, or psoas abscesses **(Fig. 126-1)** must be sought. Epidural abscesses occur in 15–20% of cases. This complication is more common in the cervical column (30%) than in the lumbar spine (12%). Persistent pain despite normalization of CRP values indicates mechanical complications such as severe osteonecrosis or spinal instability. These patients require a consult with an experienced orthopedic surgeon.

FIGURE 126-1

CT scan of acute vertebral osteomyelitis (L1/L2) due to *Staphylococcus aureus* in a 64-year-old man. Low-grade fever persisted despite appropriate IV antibiotic therapy. The scan revealed a psoas abscess on the right side.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GLOBAL CONSIDERATIONS

The incidence rate of acute vertebral osteomyelitis is similar in different regions of the world. In contrast, subacute/chronic vertebral osteomyelitis predominates in defined regions. Cases attributable to brucellosis predominate in endemic areas such as the Middle East, Africa, Central and South America, and the Indian subcontinent. Tuberculosis is an especially frequent cause in Africa and Asia (India, Indonesia, China), where more than two-thirds of the global tuberculosis burden is reported. Thus, specific diagnostic tests are needed in patients either living in or having traveled to these regions.

OSTEOMYELITIS IN LONG BONES

PATHOGENESIS

Osteomyelitis in long bones is a consequence of hematogenous seeding, exogenous contamination during trauma (open fracture), or perioperative contamination during orthopedic repairs. Its presentation is either acute (with a duration of days to a few weeks) or chronic. Hematogenous infection in long bones typically occurs in children. Ineffectively treated hematogenous osteomyelitis

during childhood can progress to chronic disease. In adults, the leading pathogenic source is exogenous infection, mainly associated with internal fixation devices. Chronic osteomyelitis can recur after a symptom-free interval of >70 years. Such recurrences are most common among elderly patients who developed osteomyelitis in the preantibiotic era.

EPIDEMIOLOGY

In adults, most cases of long-bone osteomyelitis are posttraumatic or postsurgical; less frequently, late recurrence arises from hematogenous infections during childhood. The risk of infection depends on the type of fracture. After closed fracture, implant-associated infection occurs in fewer than 1% of patients. In contrast, after open fracture, the risk of osteomyelitis ranges from ~2 up to 30%, with the precise figure depending on the degree of tissue damage during trauma and the time between injury and admission to a specialized center.

MICROBIOLOGY

The spectrum of microorganisms causing hematogenous long-bone osteomyelitis does not differ from that in vertebral osteomyelitis. *S. aureus* is most commonly isolated in each type of osteomyelitis. In rare cases, mycobacteria or fungal agents such as *Cryptococcus* species, *Sporothrix schenckii*, *Blastomyces dermatitidis*, or *Coccidioides* species are found in patients who live or have traveled in endemic regions. Impaired cellular immunity (e.g., in HIV infection or after transplantation) predisposes to these etiologies. Coagulase-negative staphylococci are the second most common etiologic agents (after *S. aureus*) in implant-associated osteomyelitis. After open fracture, contiguous long-bone osteomyelitis is typically caused by gram-negative bacilli or a polymicrobial mixture of organisms.

CLINICAL MANIFESTATIONS

The leading symptoms in adults with primary or recurrent hematogenous long-bone osteomyelitis are pain and low-grade fever. Infection occasionally manifests as clinical sepsis and local signs of inflammation (erythema and swelling). After internal fixation, osteomyelitis can be classified as early (acute; <3 weeks), delayed (3–10 weeks), or late (chronic) infection. Early/acute long-bone osteomyelitis manifests as signs of surgical site infection, such as erythema and impaired wound healing. Acute implant-associated infection may also follow hematogenous seeding at any time after implantation of a device. Typical symptoms are new-onset pain and signs of sepsis. Delayed or late (chronic) infections are usually caused by low-virulence microorganisms or occur after ineffective treatment of early-onset infection. Patients may present with persisting pain, subtle local signs of inflammation, intermittent discharge of pus, or fluctuating erythema over the scar **(Fig. 126-2)**.

FIGURE 126-2

A 42-year-old man who had had a malleolar fracture 6 weeks previously had persistent pain and slight inflammation after orthopedic repair. His infection was treated with oral antibiotics without debridement surgery. This insufficient management of an implant-associated *Staphylococcus aureus* infection was complicated by a sinus tract.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS

The diagnostic workup for acute hematogenous long-bone osteomyelitis is similar to that for vertebral osteomyelitis. Bone remodeling and thus marker uptake are increased for at least 1 year after surgery. Therefore, the three-phase bone scan is not useful during this interval. However, in late recurrences it allows rapid diagnosis at low cost. If the results are positive, CT is required in order to estimate the extent of inflamed tissue and to detect bone necrosis (sequesters). Implant-associated infection should be suspected if CRP values do not return to the normal range or rise after an initial decrease. Clinical and laboratory suspicion should prompt surgical exploration and sampling.

In chronic osteomyelitis of >1 year's duration, single-photon emission CT plus conventional CT (SPECT/CT) is a good option, either with ^{99m}Tc methylene diphosphonate (^{99m}Tc-MDP)–labeled leukocytes or with labeled monoclonal antibodies to granulocytes. Surgical debridement is needed for diagnostic (biopsy culture, histology) and therapeutic reasons.

TREATMENT

TREATMENT Osteomyelitis in Long Bones

Treatment for acute hematogenous infection in long bones is identical to that for acute vertebral osteomyelitis (Table 126-1). The suggested duration of antibiotic therapy is 4–6 weeks. In contrast to chronic or implant-associated osteomyelitis, acute hematogenous infection does not generally require surgical intervention. Initial IV administration of antimicrobial agents is followed by long-term oral treatment. The duration of the initial IV phase of therapy has not been defined. The IV course can be as short as 2 days for a drug with excellent bioavailability. In recurrences of chronic osteomyelitis as well as in each type of exogenous osteomyelitis (acute, chronic, with or without an implant), a combination of surgical debridement, obliteration of dead space, and long-term antibiotic therapy is needed.

The therapeutic aims in patients whose infections are associated with internal fixation devices are consolidation of the fracture and prevention of chronic osteomyelitis. Stable implants can be maintained except in patients with uncontrolled sepsis. Appropriate antimicrobial therapies are listed in Table 126-2. The cure rate for early staphylococcal implant-associated infections treated with a fluoroquinolone plus rifampin is >90%. Rifampin is efficacious against staphylococcal biofilms of ≤3 weeks' duration. Similarly, fluoroquinolones are active against biofilms formed by gram-negative bacilli. In these cases, an initial 2-week course of IV therapy with a β-lactam is suggested in order to minimize the risk of emergence of resistance to the oral drugs. The total duration of treatment is 3 months, and the device can be retained even after antibiotics have been discontinued. In contrast, in cases caused by rifampin-resistant staphylococci or fluoroquinolone-resistant gram-negative bacilli, the hardware should be removed after consolidation of the fracture and before discontinuation of antibiotics. These patients are treated with an oral antibiotic (suppressive therapy) as long as retention of the hardware is necessary.

COMPLICATIONS

The main complication of long-bone osteomyelitis is the persistence of infection with progression to chronic osteomyelitis. This risk is especially high after internal fixation of an open fracture and among patients with implant-associated osteomyelitis that is treated without surgical debridement. In chronic osteomyelitis, recurrent sinus tracts result in severe damage to skin and soft tissue (Fig. 126-2). Patients who have chronic open wounds need a therapeutic approach combining orthopedic repair and plastic reconstructive surgery.

GLOBAL CONSIDERATIONS

 In North American and Western European countries, tuberculous osteomyelitis is extremely rare, occurring mainly in very old people, in HIV-infected patients, and in immigrants from endemic countries. In contrast, in countries where the prevalence of tuberculosis is high (India, Indonesia, China), tuberculous osteomyelitis must routinely be considered.

PERIPROSTHETIC JOINT INFECTION

PATHOGENESIS

Implanted foreign material is highly susceptible to local infection due to local immunodeficiency around the device. Infection occurs by either the exogenous or the hematogenous route. More rarely, contiguous spread from adjacent sites of osteomyelitis or deep soft-tissue infection may cause periprosthetic joint infection (PJI). The fact that foreign devices are covered with host proteins such as fibronectin favors the adherence of staphylococci and the formation of a biofilm that resists phagocytosis.

EPIDEMIOLOGY

The risk of infection manifesting during the first 2 postoperative years varies according to the joint. It is lowest after hip and knee arthroplasty (0.3–1.5%) and highest after ankle and elbow replacement (4–10%). The risk of hematogenous PJI is highest in the early postoperative period. However, hematogenous seeding occurs throughout life, and most cases therefore develop >2 years after implantation. The rate of risk for secondary PJI during *S. aureus* bacteremia is 30–40%.

MICROBIOLOGY

About 50–70% of cases of PJI are caused by staphylococci (*S. aureus* and coagulase-negative staphylococci), 6–10% by streptococci, 4–10% by gram-negative bacilli, and the rest by various other microorganisms. In some centers, the fraction of PJI cases caused by gram-negative bacilli is much higher for unknown reasons. All microorganisms can cause PJI, including fungi and mycobacteria. *P. acnes* causes up to one-third of episodes of periprosthetic shoulder infection.

CLASSIFICATION AND CLINICAL MANIFESTATIONS

PJI is traditionally classified as early (<3 months after implantation), delayed (3–24 months after surgery), or late (>2 years after implantation). For therapeutic decision-making (see below), it is more useful to classify PJI as (1) acute hematogenous PJI with <3 weeks of symptoms, (2) early postinterventional PJI manifesting within 1 month after surgery, or (3) chronic PJI with symptom duration of >3 weeks.

Acute exogenous PJI typically presents with local signs of infection **(Fig. 126-3)**. In contrast, acute hematogenous PJI, most often caused by *S. aureus*, is characterized by new-onset pain that initially is not accompanied by prominent local inflammatory signs. In most cases, an ongoing sepsis syndrome dominates the clinical picture. Key findings in chronic PJI are joint effusion, local pain, implant loosening, and occasionally a sinus tract. Chronic PJI is most commonly caused by low-virulence microorganisms such as coagulase-negative staphylococci or *P. acnes*. These infections are characterized by nonspecific symptoms, such as chronic pain caused by low-grade inflammation or early loosening.

FIGURE 126-3

Acute postoperative periprosthetic joint infection of the left hip caused by group B streptococci in a 68-year-old woman.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS

Blood tests such as the measurement of CRP (elevated levels, ≥10 mg/L) and erythrocyte sedimentation rate (elevated rates, ≥30 mm/h) are sensitive (91–97%) but not specific (70–78%).

Synovial fluid cell counts are ~90% sensitive and specific, with threshold values of 1700 leukocytes/μL in periprosthetic knee infection and 4200 leukocytes/μL in periprosthetic hip infection. In the future, α-defensin, a biomarker that can be tested in synovial fluid, may replace cell counts. This test is expensive but accurate and easy to perform. During debridement surgery, at least three but optimally six tissue samples should be obtained for culture and histopathology. If implant material (modular parts, screws, or the prosthesis) is removed, sonication of this material followed by culture and/or use of molecular methods to examine the sonicate fluid allows the detection of microorganisms in biofilms.

The three-phase bone scan is very sensitive for detecting PJI but is not specific. As mentioned above, this test does not differentiate bone remodeling from infection and therefore is not useful during at least the first year after implantation. CT and MRI detect soft tissue infection, prosthetic loosening, and bone erosion, but imaging artifacts caused by metal implants limit their use. ^{18}F-fluorodeoxyglucose PET (^{18}F-FD-PET) is an alternative method with fair sensitivity and specificity for the detection of PJI. However, ^{18}F-FDG-PET/CT still is not an established technique for the diagnosis of PJI because of controversial published results.

TREATMENT

TREATMENT Periprosthetic Joint Infection

Treatment of PJI requires a multidisciplinary approach involving an experienced orthopedic surgeon, an infectious disease specialist, a plastic reconstructive surgeon, and a microbiologist. Therefore, most patients are referred to a specialized center. In general, the goal of treatment is cure—i.e., a pain-free functional joint with complete eradication of the infecting pathogen(s). However, for patients with severe comorbidity, lifelong suppressive antimicrobial therapy may be preferred. As a rule, antimicrobial therapy without surgical intervention is not curative but merely suppressive. There are four curative surgical options: debridement and implant retention, one-stage implant exchange, two-stage implant exchange, and implant removal without replacement. Implant retention offers a good chance of infection-free survival (>80%) only if the following conditions are fulfilled: (1) acute infection, (2) stable implant, (3) pathogen susceptible to a biofilm-active antimicrobial agent (see below), and (4) skin and soft tissue in good condition.

Table 126-2 summarizes pathogen-specific antimicrobial therapy for PJI. Initial IV therapy is followed by long-term oral antibiotics. Efficacious treatment is best defined in staphylococcal implant-associated infections. Rifampin exhibits excellent activity against biofilms composed of susceptible staphylococci. Because of the risk of rapid emergence of resistance, rifampin must always be combined with another effective antibiotic. If gram-negative infections are treated with implant retention, fluoroquinolones should be used because of their activity against gram-negative biofilms.

PREVENTION OF HEMATOGENOUS INFECTION

As mentioned above, hematogenous seeding may occur throughout life. This risk is highest during *S. aureus* bacteremia from a distant focus. Therefore, documented bacterial infections should be promptly treated in patients with prosthetic joints. However, according to a large prospective case–control study, the risk of prosthetic hip or knee infection is not increased following dental procedures. Therefore, antibiotic prophylaxis is not needed during dental work.

GLOBAL CONSIDERATIONS

Rifampin and fluoroquinolones are still the only antimicrobial agents with good activity against staphylococcal and gram-negative biofilms, respectively. Thus, in countries with high rates of rifampin

resistance in staphylococci and/or high rates of fluoroquinolone resistance in gram-negative bacilli, debridement with implant retention generally does not yield a good cure rate.

STERNAL OSTEOMYELITIS

PATHOGENESIS

Sternal osteomyelitis occurs primarily after sternal surgery (with the entry of exogenous organisms) and more rarely by hematogenous seeding or contiguous extension from adjacent sites of sternocostal arthritis. Exogenous sternal osteomyelitis after open sternal surgery is also called *deep sternal-wound infection.* Exogenous infection may follow minor sternal trauma, sternal fracture, and manubriosternal septic arthritis. Tuberculous sternal osteomyelitis typically manifests during hematogenous seeding in children or as reactivated infection in adults. Reactivation is sometimes preceded by blunt trauma. In rare cases, tuberculous sternal osteomyelitis is caused by continuous infection from an infected internal mammary lymph node.

EPIDEMIOLOGY

The incidence of poststernotomy wound infection varies from 0.5 to 2%, but figures are even higher among patients with risk factors such as diabetes, obesity, chronic renal failure, emergency surgery, use of bilateral internal mammary arteries, and re-exploration for bleeding. Rapid diagnosis and correct management of superficial sternal wound infection prevent its progression to sternal osteomyelitis. Primary (hematogenous) sternal osteomyelitis accounts for only 0.3% of all cases of osteomyelitis. Risk factors are IV drug use, HIV infection, radiotherapy, blunt trauma, cardiopulmonary resuscitation, alcohol abuse, liver cirrhosis, and hemoglobinopathy.

MICROBIOLOGY

Poststernotomy osteomyelitis is generally caused by *S. aureus* (10–20% of cases), coagulase-negative staphylococci (40–60%), gram-negative bacilli (15–25%), or *P. acnes* (2–10%). Fungal infections caused by *Candida* species also play a role. The fact that ~20% of cases are polymicrobial is indicative of exogenous superinfection during therapy. Hematogenous sternal osteomyelitis is caused most commonly by *S. aureus*. Other microorganisms play a role in special populations—e.g., *P. aeruginosa* in IV drug users, *Salmonella* species in individuals with sickle cell anemia, and *M. tuberculosis* in patients from endemic areas who have previously had tuberculosis.

CLINICAL MANIFESTATIONS

Exogenous sternal osteomyelitis manifests as fever, increased local pain, erythema, wound discharge, and sternal instability **(Fig. 126-4)**. Contiguous mediastinitis is a feared complication, occurring in ~10–30% of patients with sternal osteomyelitis. Hematogenous sternal osteomyelitis is characterized by sternal pain, swelling, and erythema. In addition, most patients have systemic signs and symptoms of sepsis.

FIGURE 126-4

Sternal osteomyelitis caused by *Staphylococcus epidermidis* 5 weeks after sternotomy for aortocoronary bypass in a 72-year-old man.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The differential diagnosis of hematogenous sternal osteomyelitis includes immunologic processes typically presenting as systemic or multifocal inflammation of the sternum or of the sternoclavicular or sternocostal joints (e.g., SAPHO [synovitis, acne, pustulosis, hyperostosis, osteitis], vasculitis, and chronic multifocal relapsing osteomyelitis).

DIAGNOSIS

In primary sternal osteomyelitis, the diagnostic workup does not differ from that in other types of hematogenous osteomyelitis (see above). When a patient has grown up in regions where tuberculosis is endemic, a specific workup for mycobacterial infection should be performed, especially if osteomyelitis had its onset after a blunt sternal trauma. In secondary sternal osteomyelitis, leukocyte counts may be normal, but the CRP level is >100 mg/L in most cases. Tissue sampling for microbiologic studies is crucial. In osteomyelitis associated with sternal wires, low-virulence microorganisms, such as coagulase-negative staphylococci, play an important role. In order to differentiate between colonization and infection, samples from at least three deep biopsies should be subjected to microbiologic examination. Superficial swab cultures are not diagnostic and may be misleading. No studies have compared the value of the various imaging modalities in suspected primary sternal osteomyelitis. However, MRI is the current gold standard for detection of each type of osteomyelitis.

TREATMENT

TREATMENT Sternal Osteomyelitis

In cases of deep sternal-wound infection, antibiotic therapy should be started immediately after samples have been obtained for microbiologic analyses in order to control clinical sepsis. To protect a newly inserted heart valve, initial treatment should be directed against staphylococci, with consideration of the local susceptibility pattern. In centers with a high prevalence of methicillin-resistant *S. aureus*, vancomycin or daptomycin should be added to a broad-spectrum β-lactam drug. As soon as cultures of blood and/or deep wound biopsies have confirmed the pathogen's identity and susceptibility pattern, treatment should be optimized and narrowed accordingly. Tables 126-1 and 126-2 show appropriate therapeutic choices for the most frequently identified microorganisms causing

sternal osteomyelitis in the absence and presence, respectively, of an implanted device. In a recent observational study of patients with staphylococcal deep sternal-wound infection, the use of a rifampin-containing regimen was predictive of success. The optimal duration of antibiotic therapy has not been established. In acute sternal osteomyelitis without hardware, a 6-week course is the rule. In patients with remaining sternal wires, treatment duration is generally prolonged to 3 months (Table 126-2). Like other types of tuberculous bone infection, tuberculous sternal osteomyelitis is treated for 6–12 months.

Primary sternal osteomyelitis can generally be treated without surgery. In contrast, in secondary sternal osteomyelitis, debridement is always required. This procedure should be performed by a team of experienced surgeons, since mediastinitis, bone infection, and skin and soft tissue damage may need to be treated during the same intervention.

PROGNOSIS

Primary sternal osteomyelitis poses a minimal mortality risk. In contrast, the in-hospital mortality rates from secondary sternal osteomyelitis are 15–30% after sternal surgery.

GLOBAL CONSIDERATIONS

In endemic areas, microorganisms such as *M. tuberculosis*, *Salmonella* species, and *Brucella* species should be considered during sampling for microbiologic diagnosis.

FOOT OSTEOMYELITIS

PATHOGENESIS

Osteomyelitis of the foot usually occurs in patients with diabetes, peripheral arterial insufficiency, or peripheral neuropathy and after foot surgery. These entities are often linked to each other, especially in diabetic patients with late complications. However, foot osteomyelitis is also seen in patients with isolated peripheral neuropathy and can manifest as implant-associated osteomyelitis in patients without comorbidity due to a deep wound infection after foot surgery (hallux valgus surgery, arthrodesis, total ankle arthroplasty). Foot osteomyelitis is acquired almost exclusively by the exogenous route. It is a complication of deep pressure ulcers and of impaired wound healing after surgery.

EPIDEMIOLOGY

The incidence of diabetic foot infection is 30–40 cases/1000 persons with diabetes per year. The condition starts with skin and soft tissue lesions and progresses to osteomyelitis, especially in patients with risk factors. About 20–60% of patients with diabetic foot infection have confirmed osteomyelitis. Diabetic foot osteomyelitis increases the risk of amputation. With adequate management of the early stage of diabetic foot infections, the rate of amputation can be lowered.

RISK FACTORS

Risk factors for diabetic foot infection are (1) peripheral motor, sensory, and autonomic neuropathy; (2) neuro-osteoarthropathic deformities (Charcot foot; **Fig. 126-5**); (3) arterial insufficiency; (4) uncontrolled hyperglycemia; (5) disabilities such as reduced vision; and (6) maladaptive behavior.

FIGURE 126-5

Neuropathic joint disease (Charcot foot) complicated by chronic foot osteomyelitis in a 78-year- old woman with diabetes mellitus complicated by severe neuropathy.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MICROBIOLOGY

The correlation between cultures from bone biopsy and those from wound swabs or even deep soft-tissue punctures is poor. In a study of 31 patients with simultaneous sampling, the correlation between needle biopsy and bone biopsy cultures was only 24%. The correlation is better when *S. aureus* is isolated (40–50%) than when anaerobes (20–35%), gram-negative bacilli (20–30%), or coagulase-negative staphylococci (0–20%) are identified. When only bone-biopsy samples are considered, the leading pathogens are *S. aureus* (25–40%), anaerobes (5–20%), and various gram-negative bacilli (18–40%). The precise distribution depends on whether the patient has already been treated with antibiotics. Anaerobes are especially prevalent in chronic wounds. Pretreatment typically selects for *P. aeruginosa*, methicillin-resistant *S. aureus*, or enterococci.

DIAGNOSIS

In many cases, foot osteomyelitis can be diagnosed clinically, without imaging procedures. Most clinicians rely on the “probe-to-bone” test, which has a positive predictive value of ~90% in populations with a high pretest probability. Thus, in a patient with diabetes who is hospitalized for a chronic deep foot ulcer, the diagnosis of foot osteomyelitis is highly probable if bone can be directly touched with a metal instrument. In a patient with a lower pretest probability, MRI should be performed because of its high degree of sensitivity (80–100%) and specificity (80–90%). Plain radiography has a sensitivity of only 30–90% and a specificity of only 50–90%; it may be considered for follow-up of patients with confirmed diabetic foot osteomyelitis.

TREATMENT

TREATMENT Foot Osteomyelitis

As mentioned above, correlation between cultures of bone and those of wound swabs or wound punctures is poor. Antibiotic treatment should be based on bone culture. If no bone biopsy is performed, empirical therapy chosen in light of the most common infecting agents and the type of clinical syndrome should be given. Wound debridement combined with a 4- to 6-week course of antibiotics renders amputation unnecessary in about two-thirds of patients. According to the 2012 Infectious Diseases Society of America's clinical practice guideline for the diagnosis and treatment of diabetic foot infections, the following management strategies should be considered. If a foot ulcer is clinically infected, prompt empirical antimicrobial therapy may prevent progression to osteomyelitis. When the risk of methicillin-resistant *S. aureus* is considered high, an agent active against these strains (e.g., vancomycin) should be chosen. If the patient has not recently received antibiotics, the spectrum of the selected antibiotic must include gram-positive cocci (e.g., clindamycin, ampicillin-sulbactam). If the patient has received antibiotics within the past month, the spectrum of empirical antibiotics should include gram-negative bacilli (e.g., clindamycin plus a fluoroquinolone). If the patient has risk factors for *Pseudomonas* infection (previous colonization, residence in a warm climate, frequent exposure of the foot to water), an empirical antipseudomonal agent (e.g., piperacillin-tazobactam, cefepime) is indicated. If osteomyelitis is suspected either on clinical grounds (probe to bone) or on the basis of imaging procedures (MRI), bone biopsy should be performed. If infected bone is not entirely removed by surgery, the patient should be treated for 4–6 weeks in line with the identified pathogen(s) and their susceptibility. Treatment should initially be given by the IV route. Whether therapy can later be administered by the oral route depends on the bioavailability of oral drugs that cover the infecting agents. If dead bone cannot be removed, long-term therapy (at least 3 months) should be considered. In such cases, cure of osteomyelitis is usually the exception, and repetitive suppressive treatment may be needed.

GLOBAL CONSIDERATIONS

 The number of multiresistant microorganisms causing diabetic foot infection is increasing. The prevalence of methicillin-resistant *S. aureus* is 5–43% in various countries. In a study of 102 patients with diabetic foot infection from India, 69% of aerobic gram-negative bacilli produced extended-spectrum β-lactamase and 43% of *S. aureus* isolates were methicillin resistant. Risk factors for multidrug-resistant microorganisms are poor glycemic control, prolonged duration of infection, and large ulcer size.

FURTHER READING

Chan M et al: A retrospective study of deep sternal wound infections: Clinical and microbiological characteristics, treatment, and risk factors for complications. Diag Microbiol Infect Dis 84:261, 2016.
Lew DP, Waldvogel FA: Osteomyelitis. Lancet 364:369, 2004. [PubMed: 15276398]
Lipsky BA et al: 2012 Infectious Diseases Society of America clinical practice guideline for the diagnosis and treatment of diabetic foot infections. Clin Infect Dis 54:e132, 2012. [PubMed: 22619242]
Osmon DR et al: Diagnosis and management of prosthetic joint infection: Clinical practice guidelines by the Infectious Diseases Society of America. Clin Infect Dis 56:e1, 2013. [PubMed: 23223583]
Zimmerli W: Vertebral osteomyelitis. N Engl J Med 362:1022, 2010. [PubMed: 20237348]

Chapter 127: Intraabdominal Infections and Abscesses

Miriam Baron Barshak; Dennis L. Kasper

INTRODUCTION

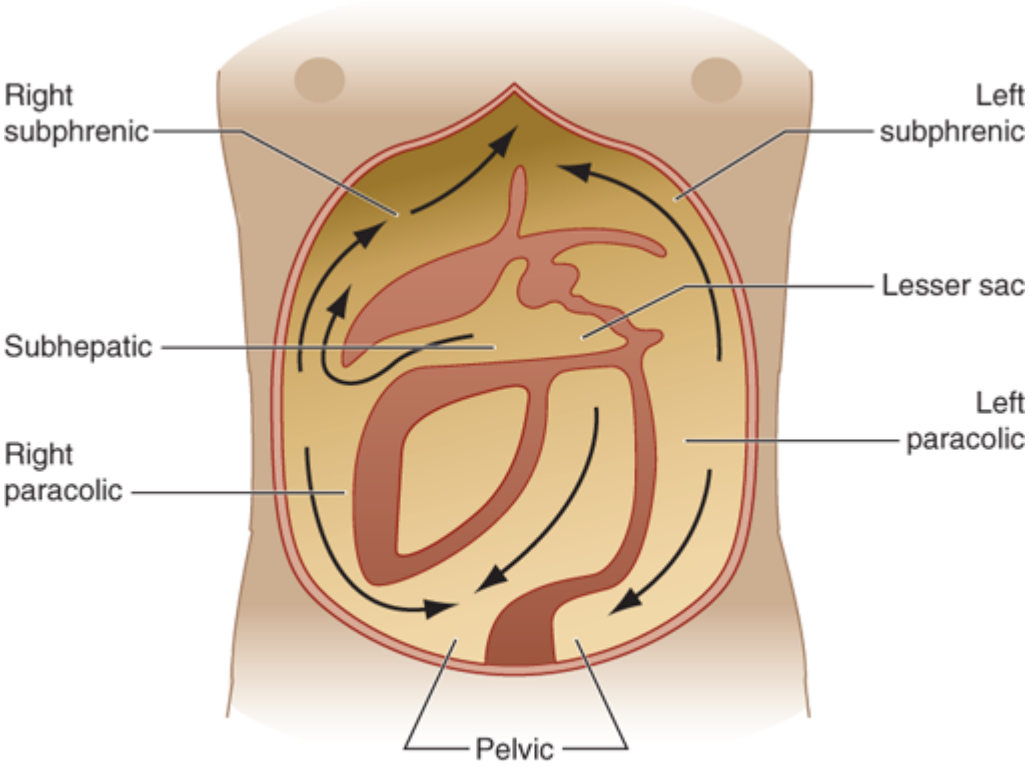
Intraperitoneal infections generally arise because a normal anatomic barrier is disrupted. This disruption may result from a variety of causes—e.g., when the appendix, a diverticulum, or an ulcer ruptures; when the bowel wall is weakened by ischemia, tumor, or inflammation (e.g., in inflammatory bowel disease); or with adjacent inflammatory processes, such as pancreatitis or pelvic inflammatory disease, in which enzymes (in the former case) or organisms (in the latter) may leak into the peritoneal cavity. Whatever the inciting event, once inflammation develops and organisms usually contained within the bowel or another organ enter the normally sterile peritoneal space, a knowable series of events takes place. Intraabdominal infections occur in two stages: peritonitis and—if the patient survives this stage and goes untreated—abscess formation. The types of microorganisms predominating in each stage of infection are responsible for the pathogenesis of disease.

PERITONITIS

Peritonitis is a life-threatening event that is often accompanied by bacteremia and sepsis syndrome (Chap. 297). The peritoneal cavity is large but is divided into compartments. The upper and lower peritoneal cavities are divided by the transverse mesocolon; the greater omentum extends from the transverse mesocolon and from the lower pole of the stomach to line the lower peritoneal cavity. The pancreas, duodenum, and ascending and descending colon are located in the anterior retroperitoneal space; the kidneys, ureters, and adrenals are found in the posterior retroperitoneal space. The other organs, including the liver, stomach, gallbladder, spleen, jejunum, ileum, transverse and sigmoid colon, cecum, and appendix, are within the peritoneal cavity. The cavity is lined with a serous membrane that can serve as a conduit for fluids—a property exploited in peritoneal dialysis (Fig. 127-1). A small amount of serous fluid is normally present in the peritoneal space, with a protein content (consisting mainly of albumin) of <30 g/L and <300 white blood cells (WBCs, generally mononuclear cells) per microliter. In bacterial infections, leukocyte recruitment into the infected peritoneal cavity consists of an early influx of polymorphonuclear leukocytes (PMNs) and a prolonged subsequent phase of mononuclear cell migration. The phenotype of the infiltrating leukocytes during the course of inflammation is regulated primarily by resident-cell chemokine synthesis.

FIGURE 127-1

Diagram of the intraperitoneal spaces, showing the circulation of fluid and potential areas for abscess formation. Some compartments collect fluid or pus more often than others. These compartments include the pelvis (the lowest portion), the subphrenic spaces on the right and left sides, and Morrison's pouch, which is a posterosuperior extension of the subhepatic spaces and is the lowest part of the paravertebral groove when a patient is recumbent. The falciform ligament separating the right and left subphrenic spaces appears to act as a barrier to the spread of infection; consequently, it is unusual to find bilateral subphrenic collections. *(Reprinted with permission from B Lorber [ed]: Atlas of Infectious Diseases, vol VII: Intra-abdominal Infections, Hepatitis, and Gastroenteritis. Philadelphia, Current Medicine, 1996, p 1.13.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PRIMARY (SPONTANEOUS) BACTERIAL PERITONITIS

Peritonitis is either primary (without an apparent source of contamination) or secondary. The types of organisms found and the clinical presentations of these two processes are different. In adults, primary bacterial peritonitis (PBP) occurs most commonly in conjunction with cirrhosis of the liver (frequently the result of alcoholism). However, the disease has been reported in adults with metastatic malignant disease, postnecrotic cirrhosis, chronic active hepatitis, acute viral hepatitis, congestive heart failure, systemic lupus erythematosus, and lymphedema as well as in patients with no underlying disease. Although PBP virtually always develops in patients with preexisting ascites, it is, in general, an uncommon event, occurring in ≤10% of cirrhotic patients. The cause of PBP has not been established definitively but is believed to involve hematogenous spread of organisms in a patient in whom a diseased liver and altered portal circulation result in a defect in the usual filtration function. Organisms multiply in ascites, a good medium for growth. Proteins of the complement cascade are found in peritoneal fluid, with lower levels in cirrhotic patients than in

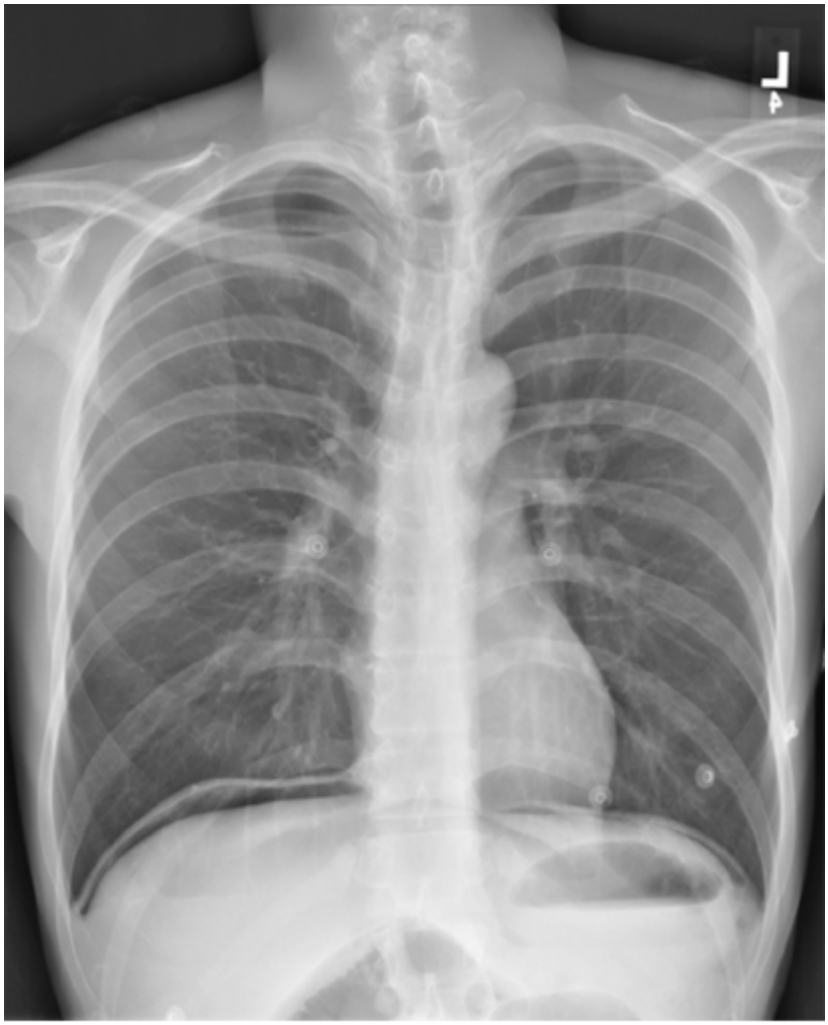
patients with ascites of other etiologies. The opsonic and phagocytic properties of PMNs are diminished in patients with advanced liver disease. Cirrhosis is associated with alterations in the gut microbiota, including an increased prevalence of potentially pathogenic bacteria such as Enterobacteriaceae. Small-intestinal bacterial overgrowth is frequently present in advanced stages of liver cirrhosis and has been linked with pathologic bacterial translocation and PBP. Factors promoting these changes in cirrhosis may include deficiencies in Paneth cell defensins, reduced intestinal motility, decreased pancreatobiliary secretions, and portal-hypertensive enteropathy.

The presentation of PBP differs from that of secondary peritonitis. The most common manifestation is fever, which is reported in up to 80% of patients. Ascites is found but virtually always predates infection. Abdominal pain, an acute onset of symptoms, and peritoneal irritation during physical examination can be helpful diagnostically, but the absence of any of these findings does not exclude this often-subtle diagnosis. Nonlocalizing symptoms (such as malaise, fatigue, or encephalopathy) without another clear etiology should also prompt consideration of PBP in a susceptible patient. It is vital to sample the peritoneal fluid of any cirrhotic patient with ascites and fever. The finding of >250 PMNs/μL is diagnostic for PBP, according to Conn. This criterion does not apply to secondary peritonitis (see below). The microbiology of PBP is also distinctive. While enteric gram-negative bacilli such as *Escherichia coli* are most commonly encountered, gram-positive organisms such as streptococci, enterococci, or even pneumococci are sometimes found. In an important development, widespread use of quinolones to prevent PBP in high-risk subgroups of patients, frequent hospitalizations, and exposure to broad-spectrum antibiotics have led to a change in the etiology of infections in patients with cirrhosis, with more gram-positive bacteria and extended-spectrum β-lactamase (ESBL)–producing Enterobacteriaceae in recent years. Risk factors for multidrug-resistant infections include nosocomial origin of infection, long-term norfloxacin prophylaxis, recent infection with multiresistant bacteria, and recent use of β-lactam antibiotics. In PBP, a single organism is typically isolated; anaerobes are found less frequently in PBP than in secondary peritonitis, in which a mixed flora including anaerobes is the rule. In fact, if PBP is suspected and multiple organisms including anaerobes are recovered from the peritoneal fluid, the diagnosis must be reconsidered and a source of secondary peritonitis sought.

The diagnosis of PBP is not easy. It depends on the exclusion of a primary intraabdominal source of infection. Contrast-enhanced CT is useful in identifying an intraabdominal source for infection. It may be difficult to recover organisms from cultures of peritoneal fluid, presumably because the burden of organisms is low. However, the yield can be improved if 10 mL of peritoneal fluid is placed directly into a blood culture bottle. Because bacteremia frequently accompanies PBP, blood should be cultured simultaneously. To maximize the yield, culture samples should be collected prior to administration of antibiotics. No specific radiographic studies are helpful in the diagnosis of PBP. A plain film of the abdomen would be expected to show ascites. Chest and abdominal radiography should be performed when patients have abdominal pain to exclude free air, which signals a perforation **(Fig. 127-2)**.

FIGURE 127-2

Pneumoperitoneum. Free air under the diaphragm on an upright chest film suggests the presence of a bowel perforation and associated peritonitis. *(Image courtesy of Dr. John Braver; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Primary Bacterial Peritonitis

Treatment for PBP is directed at the isolate from blood or peritoneal fluid. Gram's staining of peritoneal fluid often gives negative results in PBP. Therefore, until culture results become available, therapy should cover gram-negative aerobic bacilli and gram-positive cocci. Third-generation cephalosporins such as cefotaxime (2 g q8h, administered IV) provide reasonable initial coverage in moderately ill patients. Broad-spectrum antibiotics, such as penicillin/β-lactamase inhibitor combinations (e.g., piperacillin/tazobactam, 3.375 g q6h IV for adults with normal renal function) or ceftriaxone (2 g q24h IV), are also options. Broader empirical coverage aimed at resistant hospital-acquired gram-negative bacteria (e.g., treatment with a carbapenem) may be appropriate for nosocomially acquired PBP until culture results become available. Empirical coverage for anaerobes is not necessary. A mortality benefit from albumin (1.5 g/kg of body weight within 6 h of detection and 1.0 g/kg on day 3) has been demonstrated for patients who present with serum creatinine levels ≥1 mg/dL, blood urea nitrogen levels ≥30 mg/dL, or total bilirubin levels ≥4 mg/dL but not for patients

who do not meet these criteria. After the infecting organism is identified, therapy should be narrowed to target the specific pathogen. Patients with PBP usually respond within 72 h to appropriate antibiotic therapy. Antimicrobial treatment can be administered for as little as 5 days if rapid improvement occurs and blood cultures are negative, but a course of up to 2 weeks may be required for patients with bacteremia and for those whose improvement is slow. Persistence of WBCs in the ascitic fluid after therapy should prompt a search for additional diagnoses.

Prevention

Primary Prevention

Several observational studies and a meta-analysis raise the concern that gastric acid suppression may increase the risk of PBP. No prospective studies have yet addressed whether avoidance of such therapy may prevent PBP. Nonselective beta blockers may prevent secondary bacterial peritonitis. A 2012 guideline from the American Association for the Study of Liver Diseases recommends chronic antibiotic prophylaxis with a regimen described in the next section for patients who are at highest risk for PBP—that is, those with an ascitic-fluid total protein level <1.5 g/dL along with impaired renal function (creatinine, ≥1.2 mg/dL; blood urea nitrogen, ≥25 mg/dL; or serum sodium, ≤130 mg/dL) and/or liver failure (Child-Pugh score, ≥9; and bilirubin, ≥3 mg/dL). A 7-day course of antibiotic prophylaxis is recommended for patients with cirrhosis and gastrointestinal bleeding.

Secondary Prevention

PBP has a high rate of recurrence. Up to 70% of patients experience a recurrence within 1 year. Antibiotic prophylaxis is recommended for patients with a history of PBP to reduce this rate to <20% and improve short-term survival rates. Prophylactic regimens for adults with normal renal function include fluoroquinolones (ciprofloxacin, 500 mg weekly; or norfloxacin [not available in the United States], 400 mg/d) or trimethoprim-sulfamethoxazole (one double-strength tablet daily). However, long-term administration of broad-spectrum antibiotics in this setting has been shown to increase the risk of severe staphylococcal infections.

SECONDARY PERITONITIS

Secondary peritonitis develops when bacteria contaminate the peritoneum as a result of spillage from an intraabdominal viscus. The organisms found almost always constitute a mixed flora in which facultative gram-negative bacilli and anaerobes predominate, especially when the contaminating source is colonic. Early in the course of infection, when the host response is directed toward containment, exudate containing fibrin and PMNs is found. Early death in this setting is attributable to gram-negative bacillary sepsis and to potent endotoxins circulating in the bloodstream **(Chap. 297)**. Gram-negative bacilli, particularly *E. coli*, are common bloodstream isolates, but *Bacteroides fragilis* bacteremia also occurs. The severity of abdominal pain and the clinical course depend on the inciting process. The organisms isolated from the peritoneum also vary with the source of the initial process and the normal flora at that site. Secondary peritonitis can result primarily from chemical irritation and/or bacterial contamination. For example, as long as the patient is not achlorhydric, a ruptured gastric ulcer will release low-pH gastric contents that will serve as a chemical

irritant. The normal flora of the stomach comprises the same organisms found in the oropharynx but in lower numbers. Thus, the bacterial burden in a ruptured ulcer is negligible compared with that in a ruptured appendix. The normal flora of the colon below the ligament of Treitz contains ~10^{11} anaerobic organisms/g of feces but only 10^{8} aerobes/g; therefore, anaerobic species account for 99.9% of the bacteria **(Chap. 459)**. Leakage of colonic contents (pH 7–8) does not cause significant chemical peritonitis, but infection is intense because of the heavy bacterial load.

Depending on the inciting event, local symptoms may occur in secondary peritonitis—for example, epigastric pain from a ruptured gastric ulcer. In appendicitis **(Chap. 324)**, the initial presenting symptoms are often vague, with periumbilical discomfort and nausea followed in a number of hours by pain more localized to the right lower quadrant. Unusual locations of the appendix (including a retrocecal position) can complicate this presentation further. Once infection has spread to the peritoneal cavity, pain increases, particularly with infection involving the parietal peritoneum, which is innervated extensively. Patients usually lie motionless, often with knees drawn up to avoid stretching the nerve fibers of the peritoneal cavity. Coughing and sneezing, which increase pressure within the peritoneal cavity, are associated with sharp pain. There may or may not be pain localized to the infected or diseased organ from which secondary peritonitis has arisen. Patients with secondary peritonitis generally have abnormal findings on abdominal examination, with marked voluntary and involuntary guarding of the anterior abdominal musculature. Later findings include tenderness, especially rebound tenderness. In addition, there may be localized findings in the area of the inciting event. In general, patients are febrile, with marked leukocytosis and a left shift of the WBCs to band forms.

While recovery of organisms from peritoneal fluid is easier in secondary than in primary peritonitis, a tap of the abdomen is rarely the procedure of choice in secondary peritonitis. An exception is in cases involving trauma, where the possibility of a hemoperitoneum may need to be excluded early. Emergent studies (such as abdominal CT) to find the source of peritoneal contamination should be undertaken if the patient is hemodynamically stable; unstable patients may require surgical intervention without prior imaging. Results of cultures from drain sites are not reliable for defining the etiology of infections.

TREATMENT

TREATMENT

Secondary Peritonitis

Treatment for secondary peritonitis includes early administration of antibiotics aimed particularly at aerobic gram-negative bacilli and anaerobes (see below). The most appropriate regimen depends on the anticipated flora and the degree of illness. Community-acquired infections associated with mild to moderate disease can be treated with many drugs covering these organisms, including broad-spectrum penicillin/β-lactamase inhibitor combinations (e.g., ticarcillin/clavulanate, 3.1 g q4–6h IV; or piperacillin/tazobactam, 3.375 g q6h IV) or a combination of either a fluoroquinolone (e.g., levofloxacin, 750 mg q24h IV) or a third-generation cephalosporin (e.g., ceftriaxone, 2 g q24h IV) plus metronidazole (500 mg q8h IV). Patients in intensive care

units and/or those with health care–associated infections should receive antibiotics targeting more resistant gram-negative organisms such as *Pseudomonas aeruginosa*—e.g., imipenem (500 mg q6h IV), meropenem (1 g q8h IV), higher-dose piperacillin/tazobactam (4.5 g IV q6h), or drug combinations such as cefepime (2 g IV q8h) or ceftazidime (2 g IV q8h) plus metronidazole. The role of enterococci and *Candida* species in mixed infections is controversial; however, because cephalosporin-based regimens lack activity against enterococci, ampicillin or vancomycin can be added to these regimens for enterococcal coverage in very ill patients until culture results are available. For patients known to be colonized with ampicillin-resistant, vancomycin-resistant enterococci (VRE), a VRE-active agent, such as linezolid or daptomycin, should be included. Antifungal coverage is warranted if there is growth of *Candida* species from a sterile site. Patients who are known to be colonized with highly resistant gram-negative organisms may require treatment with a newer agent such as ceftazidime/avibactam or ceftolozane/tazobactam. Secondary peritonitis usually requires both surgical intervention to address the inciting process and antibiotics to treat early bacteremia, to decrease the incidence of abscess formation and wound infection, and to prevent distant spread of infection. Although surgery is rarely indicated in PBP in adults, it may be life-saving in secondary peritonitis. Recombinant human activated protein C (APC) was considered at one time for treatment of severe sepsis from causes including secondary peritonitis but was withdrawn from the market in 2011 after it was determined that the drug was associated with an increased risk of bleeding and that evidence for its beneficial effects was inadequate. Thus APC should not be used for sepsis or septic shock outside randomized clinical trials.

Peritonitis may develop as a complication of abdominal surgeries. These infections may be accompanied by localizing pain and/or nonlocalizing signs or symptoms such as fever, malaise, anorexia, and toxicity. As a nosocomial infection, postoperative peritonitis may be associated with organisms such as staphylococci, components of the gram-negative hospital microflora, and the microbes that cause PBP and secondary peritonitis, as described above.

PERITONITIS IN PATIENTS UNDERGOING CONTINUOUS AMBULATORY PERITONEAL DIALYSIS

A third type of peritonitis arises in patients who are undergoing continuous ambulatory peritoneal dialysis (CAPD). Unlike PBP and secondary peritonitis, which are caused by endogenous bacteria, CAPD-associated peritonitis usually involves skin organisms. The pathogenesis of infection is similar to that of intravascular device–related infection, in which skin organisms migrate along the catheter, which both serves as an entry point and exerts the effects of a foreign body. Exit-site or tunnel infection may or may not accompany CAPD-associated peritonitis. Like PBP, CAPD-associated peritonitis is usually caused by a single organism. Peritonitis is, in fact, the most common reason for discontinuation of CAPD. Improvements in equipment design, especially the Y-set connector, have resulted in a decrease from one case of peritonitis per 9 months of CAPD to one case per 24 months.

The clinical presentation of CAPD peritonitis resembles that of secondary peritonitis in that diffuse pain and peritoneal signs are common. The dialysate is usually cloudy and contains >100 WBCs/μL, >50% of which are neutrophils. However, the number of cells depends in part on dwell time. According to a guideline from the

International Society for Peritoneal Dialysis (2016), for patients undergoing automated peritoneal dialysis who present during their nighttime treatment and whose dwell time is much shorter than with CAPD, the clinician should use the percentage of PMNs rather than the absolute number of WBCs to diagnose peritonitis. As the normal peritoneum has very few PMNs, a proportion above 50% is strong evidence of peritonitis even if the absolute WBC count does not reach 100/μL. Meanwhile, patients undergoing automated peritoneal dialysis without a daytime exchange who present with abdominal pain may have no fluid to withdraw, in which case 1 L of dialysate should be infused and permitted to dwell a minimum of 1–2 h, then drained, examined for turbidity, and sent for cell count with differential and culture. The differential (with a shortened dwell time) may be more useful than the absolute WBC count. In equivocal cases or in patients with systemic or abdominal symptoms in whom the effluent appears clear, a second exchange is performed, with a dwell time of at least 2 h. Clinical judgment should guide initiation of therapy.

The most common organisms are *Staphylococcus* species, which accounted for ~45% of cases in one series. Historically, coagulase-negative staphylococcal species were identified most commonly in these infections, but these isolates have more recently been decreasing in frequency. *Staphylococcus aureus* is more often involved among patients who are nasal carriers of the organism than among those who are not, and this organism is the most common pathogen in overt exit-site infections. Gram-negative bacilli and fungi such as *Candida* species are also found. Vancomycin-resistant enterococci and vancomycin-intermediate *S. aureus* have been reported to produce peritonitis in CAPD patients. The finding of more than one organism in dialysate culture should prompt evaluation for secondary peritonitis. As with PBP, culture of dialysate fluid in blood culture bottles improves the yield. To facilitate diagnosis, several hundred milliliters of removed dialysis fluid should be concentrated by centrifugation before culture.

TREATMENT

TREATMENT

CAPD Peritonitis

Empirical therapy for CAPD peritonitis should be directed at *S. aureus*, coagulase-negative *Staphylococcus*, and gram-negative bacilli until the results of cultures become available. Guidelines suggest that agents should be chosen on the basis of local experience with resistant organisms. In some centers, a first-generation cephalosporin such as cefazolin (for gram-positive bacteria) and a fluoroquinolone or a third-generation cephalosporin such as ceftazidime (for gram-negative bacteria) may be reasonable; in areas with high rates of infection with methicillin-resistant *S. aureus*, vancomycin should be used instead of cefazolin, and gram-negative coverage may need to be broadened—e.g., with an aminoglycoside, ceftazidime, cefepime, or a carbapenem. Broad coverage including vancomycin should be particularly considered for patients with septic physiology or exit-site infections. Vancomycin should also be included in the regimen if the patient has a history of colonization or infection with methicillin-resistant *S. aureus* or has a history of severe allergy to penicillins and cephalosporins. Loading doses are administered intraperitoneally; doses depend on the dialysis method and the patient's renal function. Antibiotics are given either continuously (i.e., with each exchange) or intermittently (i.e., once daily, with the dose allowed to remain in the peritoneal

cavity for at least 6 h). If the patient is severely ill, IV antibiotics should be added at doses appropriate for the patient's degree of renal failure. The clinical response to an empirical treatment regimen should be rapid; if the patient has not responded after 48–96 h of treatment, new samples should be collected for cell counts and cultures, and catheter removal should be considered. For patients who lack exit-site or tunnel infection, the typical duration of antibiotic treatment is 14 days. For patients with exit-site or tunnel infection, catheter removal should be considered, and a longer duration of antibiotic therapy (up to 21 days) may be appropriate. In fungal infections, the catheter should be removed immediately.

TUBERCULOUS PERITONITIS

See Chap. 173.

INTRAABDOMINAL ABSCESSES

INTRAPERITONEAL ABSCESSES

Abscess formation is common in untreated peritonitis if overt gram-negative sepsis either does not develop or develops but is not fatal. In experimental models of abscess formation, mixed aerobic and anaerobic organisms have been implanted intraperitoneally. Without therapy directed at anaerobes, animals develop intraabdominal abscesses. As in humans, these experimental abscesses may stud the peritoneal cavity, lie within the omentum or mesentery, or even develop on the surface of or within viscera such as the liver.

Pathogenesis and Immunity

There is often disagreement about whether an abscess represents a disease state or a host response. In a sense, it represents both: while an abscess is an infection in which viable infecting organisms and PMNs are contained in a fibrous capsule, it is also a process by which the host confines microbes to a limited space, thereby preventing further spread of infection. In any event, abscesses do cause significant symptoms, and patients with abscesses can be quite ill. Experimental work has helped to define both the host cells and the bacterial virulence factors responsible—most notably in the case of *B. fragilis*. This organism, although accounting for only 0.5% of the normal colonic flora, is the anaerobe most frequently isolated from intraabdominal infections, is especially prominent in abscesses, and is the most common anaerobic bloodstream isolate. On clinical grounds, therefore, *B. fragilis* appears to be uniquely virulent. Moreover, *B. fragilis* acts alone to cause abscesses in animal models of intraabdominal infection, whereas most other *Bacteroides* species must act synergistically with a facultative organism to induce abscess formation.

Of the several virulence factors identified in *B. fragilis*, one is critical: the capsular polysaccharide complex found on the bacterial surface. This complex comprises at least eight distinct surface polysaccharides. Structural analysis of these polysaccharides has shown an unusual motif of oppositely charged sugars. Polysaccharides having these *zwitterionic* characteristics, such as polysaccharide A, evoke a host response in the peritoneal cavity that localizes bacteria into abscesses. *B. fragilis* and polysaccharide A have been found to adhere to primary mesothelial cells in vitro; this adherence, in turn, stimulates the production of tumor

necrosis factor α and intercellular adhesion molecule 1 by peritoneal macrophages. Although abscesses characteristically contain PMNs, the process of abscess induction depends on the stimulation of T lymphocytes by these unique zwitterionic polysaccharides. The stimulated CD4+ T lymphocytes secrete leukoattractant cytokines and chemokines. The alternative pathway of complement and fibrinogen also participate in abscess formation.

While antibodies to the capsular polysaccharide complex enhance bloodstream clearance of *B. fragilis*, CD4+ T cells are critical in immunity to abscesses. When administered experimentally, *B. fragilis* polysaccharide A has immunomodulatory characteristics and stimulates CD4+ T regulatory cells via an interleukin 2–dependent mechanism to produce interleukin 10. Interleukin 10 downregulates the inflammatory response, thereby preventing abscess formation.

Clinical Presentation

Of all intraabdominal abscesses, 74% are intraperitoneal or retroperitoneal and are not visceral. Most intraperitoneal abscesses result from fecal spillage from a colonic source, such as an inflamed appendix. Abscesses can also arise from other processes. They usually form within weeks of the development of peritonitis and may be found in a variety of locations from omentum to mesentery, pelvis to psoas muscles, and subphrenic space to a visceral organ such as the liver, where they may develop either on the surface of the organ or within it. Periappendiceal and diverticular abscesses occur commonly. Diverticular abscesses are least likely to rupture. Infections of the female genital tract and pancreatitis are also among the more common causative events. When abscesses occur in the female genital tract—either as a primary infection (e.g., tuboovarian abscess) or as an infection extending into the pelvic cavity or peritoneum—*B. fragilis* figures prominently among the organisms isolated. *B. fragilis* is not found in large numbers in the normal vaginal flora. For example, it is encountered less commonly in pelvic inflammatory disease and endometritis without an associated abscess. In pancreatitis with leakage of damaging pancreatic enzymes, inflammation is prominent. Therefore, clinical findings such as fever, leukocytosis, and even abdominal pain do not distinguish pancreatitis itself from complications such as pancreatic pseudocyst, pancreatic abscess (**Chap. 341**), or intraabdominal collections of pus. Especially in cases of necrotizing pancreatitis, in which the incidence of local pancreatic infection may be as high as 30%, needle aspiration under CT guidance is performed to sample fluid for culture. Traditionally, many centers have prescribed preemptive antibiotics for patients with necrotizing pancreatitis. Imipenem is frequently used for this purpose because it reaches high tissue levels in the pancreas (although it is not unique in this regard). Randomized controlled studies have not demonstrated a benefit from this practice, and many guidelines no longer recommend preemptive antibiotics for patients with acute pancreatitis. If needle aspiration yields infected fluid in the setting of acute necrotizing pancreatitis, antibiotic treatment is appropriate in conjunction with surgical and/or percutaneous drainage of infected material. Infected pseudocysts that occur remotely from acute pancreatitis are unlikely to be associated with significant amounts of necrotic tissue and may be treated with either surgical or percutaneous catheter drainage in conjunction with appropriate antibiotic therapy.

Diagnosis

Scanning procedures have considerably facilitated the diagnosis of intraabdominal abscesses. Abdominal CT probably has the highest yield, although ultrasonography is particularly useful for the right upper quadrant, kidneys, and pelvis. Both indium-labeled WBCs and gallium tend to localize in abscesses and may be useful in finding a collection. Because gallium is taken up in the bowel, indium-labeled WBCs may have a slightly greater yield for abscesses near the bowel. Neither indium-labeled WBC scans nor gallium scans serve as a basis for a definitive diagnosis, however; both need to be followed by other, more specific studies, such as CT, if an area of possible abnormality is identified. Abscesses contiguous with or contained within diverticula are particularly difficult to diagnose with scanning procedures. Although barium should not be injected if a perforation is suspected, a barium enema occasionally may detect a diverticular abscess not diagnosed by other procedures. If one study is negative, a second study sometimes reveals a collection. Although exploratory laparotomy has been less commonly used since the advent of CT, this procedure still must be undertaken on occasion if an abscess is strongly suspected on clinical grounds.

TREATMENT

TREATMENT

Intraperitoneal Abscesses

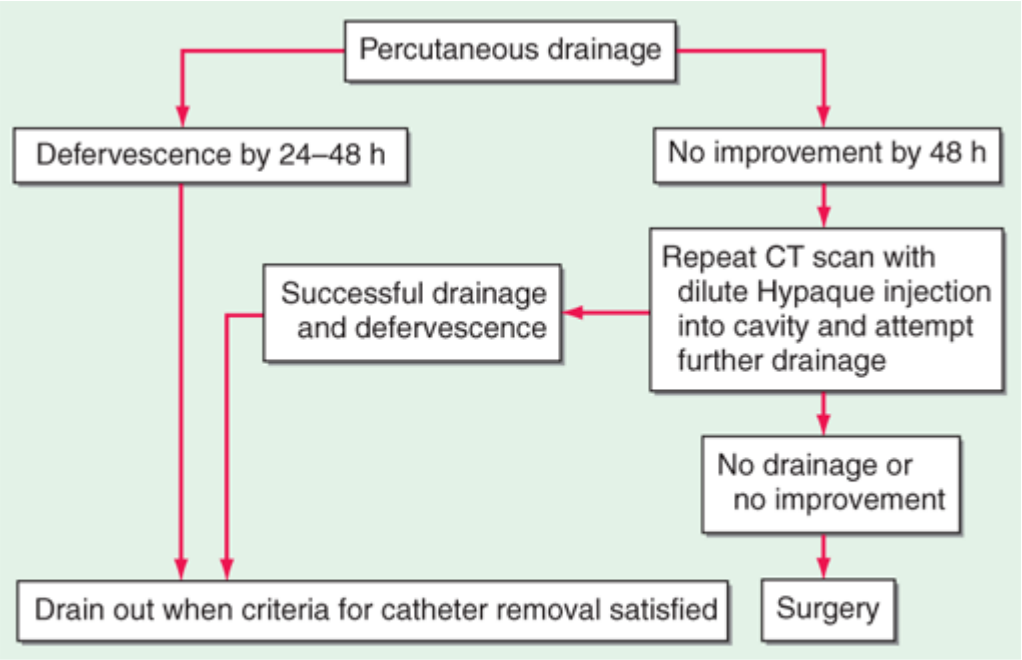
An algorithm for the management of patients with intraabdominal (including intraperitoneal) abscesses by percutaneous drainage is presented in Fig. 127-3. Treatment of intraabdominal infections involves determination of the initial focus of infection, administration of broad-spectrum antibiotics targeting the organisms involved, and performance of a drainage procedure if one or more definitive abscesses have formed. Antimicrobial therapy, in general, is adjunctive to drainage and/or surgical correction of an underlying lesion or process in intraabdominal abscesses. Results of cultures from drain sites are not reliable for defining the etiology of infections. Unlike the intraabdominal abscesses resulting from most causes, for which drainage of some kind is generally required, abscesses associated with diverticulitis usually wall off locally after rupture of a diverticulum, so that surgical intervention is not routinely required.

A number of agents exhibit excellent activity against aerobic gram-negative bacilli. Because death in intraabdominal sepsis is linked to gram-negative bacteremia, empirical therapy for intra-abdominal infection always needs to include adequate coverage of gram-negative aerobic, facultative, and anaerobic organisms. Even if anaerobes are not cultured from clinical specimens, they still must be covered by the therapeutic regimen. Empirical antibiotic therapy should be the same as that discussed above for secondary peritonitis. Most clinical treatment failures are due to failure to drain the abscess and thereby achieve source control. The appropriate duration of antibiotic treatment for abdominal abscesses depends on whether the presumptive source of the intraabdominal infection has been controlled. With adequate source control, antibiotic treatment may be limited to 4 or 5 days.

FIGURE 127-3

Algorithm for the management of patients with intraabdominal abscesses by percutaneous drainage. Antimicrobial therapy should be administered concomitantly. *(Reprinted with permission from B Lorber [ed]: Atlas of Infectious Diseases, vol VII: Intra-abdominal Infections, Hepatitis, and Gastroenteritis. Philadelphia,*

Current Medicine, 1996, p 1.30, as adapted from OD Rotstein, RL Simmons, in SL Gorbach et al [eds]: Infectious Diseases. Philadelphia, Saunders, 1992, p 668.)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

VISCERAL ABSCESSES

Liver Abscesses

The liver is the organ most subject to the development of abscesses. In one study of 540 intraabdominal abscesses, 26% were visceral. Liver abscesses made up 13% of the total number, or 48% of all visceral abscesses. Liver abscesses may be solitary or multiple; they may arise from hematogenous spread of bacteria or from local spread from contiguous sites of infection within the peritoneal cavity. In the past, appendicitis with rupture and subsequent spread of infection was the most common source for a liver abscess. Currently, associated disease of the biliary tract is most common. Pylephlebitis (suppurative thrombosis of the portal vein), usually arising from infection in the pelvis but sometimes from infection elsewhere in the peritoneal cavity, is another common source for bacterial seeding of the liver.

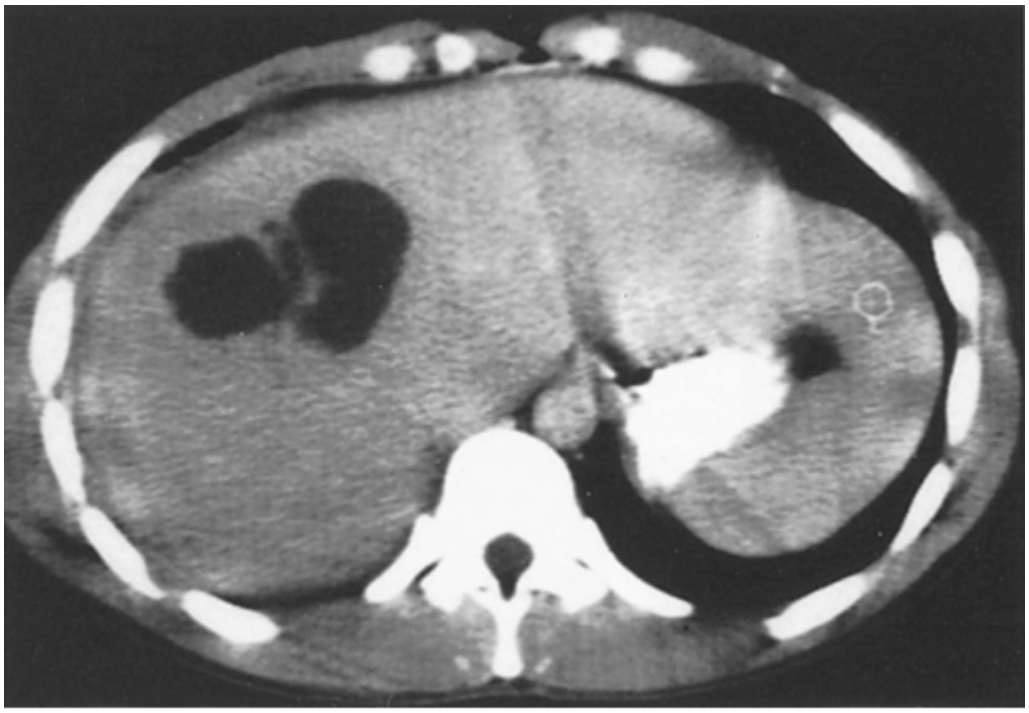
Fever is the most common presenting sign of liver abscess. Some patients, particularly those with associated disease of the biliary tract, have symptoms and signs localized to the right upper quadrant, including pain, guarding, punch tenderness, and even rebound tenderness. Nonspecific symptoms, such as chills, anorexia, weight loss, nausea, and vomiting, may also develop. Only 50% of patients with liver abscesses, however, have hepatomegaly, right-upper-quadrant tenderness, or jaundice; thus, one-half of patients have no symptoms or signs to direct attention to the liver. Fever of unknown origin may be the only manifestation of liver abscess, especially in the elderly. Diagnostic studies of the abdomen, especially the right upper quadrant, should be a part of any workup for fever of unknown origin. The single most reliable laboratory finding is an elevated serum concentration of alkaline phosphatase, which is documented in 70% of patients

with liver abscesses. Other tests of liver function may yield normal results, but 50% of patients have elevated serum levels of bilirubin, and 48% have elevated concentrations of aspartate aminotransferase. Other laboratory findings include leukocytosis in 77% of patients, anemia (usually normochromic, normocytic) in 50%, and hypoalbuminemia in 33%. Concomitant bacteremia is found in one-third to one-half of patients. A liver abscess is sometimes suggested by chest radiography, especially if a new elevation of the right hemidiaphragm is seen; other suggestive findings include a right basilar infiltrate and a right pleural effusion.

Imaging studies are the most reliable methods for diagnosing liver abscesses. These studies include ultrasonography, CT (**Fig. 127-4**), indium-labeled WBC or gallium scan, and MRI. More than one such study may be required.

FIGURE 127-4

Multilocular liver abscess on CT scan. Multiple or multilocular abscesses are more common than solitary abscesses. *(Reprinted with permission from B Lorber [ed]: Atlas of Infectious Diseases, vol VII: Intra-abdominal Infections, Hepatitis, and Gastroenteritis. Philadelphia, Current Medicine, 1996, Fig. 1.22.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Organisms recovered from liver abscesses vary with the source. In liver infection arising from the biliary tree, enteric gram-negative aerobic bacilli and enterococci are common isolates. *Klebsiella pneumoniae* liver abscess has been well described in Southeast Asia for more than 20 years and has become an emerging syndrome in North America and elsewhere. These community-acquired infections have been linked to a virulent hypermucoviscous *K. pneumoniae* phenotype and to a specific genotype. The typical syndrome includes liver abscess, bacteremia, and metastatic infection. Ampicillin/amoxicillin therapy started within the previous 30 days has been associated with increased risk for this syndrome, presumably because of selection for the causative strain. Unless previous surgery has been performed, anaerobes are not generally involved in

liver abscesses arising from biliary infections. In contrast, in liver abscesses arising from pelvic and other intraperitoneal sources, a mixed flora including both aerobic and anaerobic species is common; *B. fragilis* is the species most frequently isolated. With hematogenous spread of infection, usually only a single organism is encountered; this species may be *S. aureus* or a streptococcal species such as one in the *Streptococcus milleri* group. Liver abscesses may also be caused by *Candida* species; such abscesses usually follow fungemia in patients receiving chemotherapy for cancer and often present when PMNs return after a period of neutropenia. Amebic liver abscesses are not an uncommon problem **(Chap. 218)**. Amebic serologic testing gives positive results in >95% of cases. In addition, polymerase chain reaction (PCR) testing has been used in recent years. Negative results from these studies help to exclude this diagnosis.

TREATMENT

TREATMENT

Liver Abscesses

(Fig. 127-3) Drainage is the mainstay of therapy for intraabdominal abscesses, including liver abscesses; the approach can be either percutaneous (with a pigtail catheter kept in place or possibly with a device that can perform pulse lavage to fragment and evacuate the semisolid contents of a liver abscess) or surgical. However, there is growing interest in medical management alone for pyogenic liver abscesses. The drugs used for empirical therapy include the same ones used in intraabdominal sepsis and secondary bacterial peritonitis. Usually, blood cultures and a diagnostic aspirate of abscess contents should be obtained before the initiation of empirical therapy, with antibiotic choices adjusted when the results of Gram's staining and culture become available. Cases treated without definitive drainage generally require longer courses of antibiotic therapy. When percutaneous drainage was compared with open surgical drainage, the average length of hospital stay for the former was almost twice that for the latter, although both the time required for fever to resolve and the mortality rate were the same for the two procedures. The mortality rate was appreciable despite treatment, averaging 15%. Several factors predict the failure of percutaneous drainage and therefore may favor primary surgical intervention. These factors include the presence of multiple, sizable abscesses; viscous abscess contents that tend to plug the catheter; associated disease (e.g., disease of the biliary tract) requiring surgery; the presence of yeast; communication with an untreated obstructed biliary tree; or the lack of a clinical response to percutaneous drainage in 4–7 days.

Treatment of candidal liver abscesses often entails initial administration of liposomal amphotericin B (3–5 mg/kg IV daily) or an echinocandin, with subsequent fluconazole therapy **(Chap. 211)**. In some cases, therapy with fluconazole alone (6 mg/kg daily) may be used—e.g., in clinically stable patients whose infecting isolate is susceptible to this drug.

Splenic Abscesses

Splenic abscesses are much less common than liver abscesses. The incidence of splenic abscesses has ranged from 0.14 to 0.7% in various autopsy series. The clinical setting and the organisms isolated usually differ from those for liver abscesses. The degree of clinical suspicion for splenic abscess needs to be high

because this condition is frequently fatal if left untreated. Even in the most recently published series, diagnosis was made only at autopsy in 37% of cases. Although splenic abscesses may arise occasionally from contiguous spread of infection or from direct trauma to the spleen, hematogenous spread of infection is more common. Bacterial endocarditis is the most common associated infection **(Chap. 123)**. Splenic abscesses can develop in patients who have received extensive immunosuppressive therapy (particularly those with malignancy involving the spleen) and in patients with hemoglobinopathies or other hematologic disorders (especially sickle cell anemia).

Although ~50% of patients with splenic abscesses have abdominal pain, the pain is localized to the left upper quadrant in only one-half of these cases. Splenomegaly is found in ~50% of cases. Fever and leukocytosis are generally present; the development of fever preceded diagnosis by an average of 20 days in one series. Left-sided chest findings may include abnormalities to auscultation, and chest radiographic findings may include an infiltrate or a left-sided pleural effusion. CT scan of the abdomen has been the most sensitive diagnostic tool. Ultrasonography can yield the diagnosis but is less sensitive. Liver–spleen scan or gallium scan may also be useful. Streptococcal species are the most common bacterial isolates from splenic abscesses, followed by *S. aureus*—presumably reflecting the associated endocarditis. An increase in the prevalence of gram-negative aerobic isolates from splenic abscesses has been reported; these organisms often derive from a urinary tract focus, with associated bacteremia, or from another intraabdominal source. *Salmonella* species are seen fairly commonly, especially in patients with sickle cell hemoglobinopathy. Anaerobic species accounted for only 5% of isolates in the largest collected series, but the reporting of a number of “sterile abscesses” may indicate that optimal techniques for the isolation of anaerobes were not used.

TREATMENT

TREATMENT

Splenic Abscesses

Because of the high mortality figures reported for splenic abscesses, splenectomy with adjunctive antibiotics has traditionally been considered standard treatment and remains the best approach for complex, multilocular abscesses or multiple abscesses. However, percutaneous drainage has worked well for single, small (<3-cm) abscesses in some studies and may also be useful for patients with high surgical risk. Patients undergoing splenectomy should be vaccinated against encapsulated organisms (*Streptococcus pneumoniae*, *Haemophilus influenzae*, *Neisseria meningitidis*). The most important factor in successful treatment of splenic abscesses is early diagnosis.

Perinephric and Renal Abscesses

Perinephric and renal abscesses are not common. The former accounted for only ~0.02% of hospital admissions and the latter for ~0.2% in Altemeier’s series of 540 intraabdominal abscesses. Before antibiotics became available, most renal and perinephric abscesses were hematogenous in origin, usually complicating prolonged bacteremia, with *S. aureus* most commonly recovered. Now, in contrast, >75% of perinephric and renal abscesses arise from a urinary tract infection. Infection ascends from the bladder to the kidney, with

pyelonephritis preceding abscess development. Bacteria may directly invade the renal parenchyma from medulla to cortex. Local vascular channels within the kidney may also facilitate the transport of organisms. Areas of abscess developing within the parenchyma may rupture into the perinephric space. The kidneys and adrenal glands are surrounded by a layer of perirenal fat that, in turn, is surrounded by Gerota's fascia, which extends superiorly to the diaphragm and inferiorly to the pelvic fat. Abscesses extending into the perinephric space may track through Gerota's fascia into the psoas or transversalis muscles, into the anterior peritoneal cavity, superiorly to the subdiaphragmatic space, or inferiorly to the pelvis. Of the risk factors that have been associated with the development of perinephric abscesses, the most important is concomitant nephrolithiasis obstructing urinary flow. Of patients with perinephric abscess, 20–60% have renal stones. Other structural abnormalities of the urinary tract, prior urologic surgery, trauma, and diabetes mellitus have also been identified as risk factors.

The organisms most frequently encountered in perinephric and renal abscesses are *E. coli*, *Proteus* species, and *Klebsiella* species. *E. coli*, the aerobic species most commonly found in the colonic flora, seems to have unique virulence properties in the urinary tract, including factors promoting adherence to uroepithelial cells. The urease of *Proteus* species splits urea, thereby creating a more alkaline and more hospitable environment for bacterial proliferation. *Proteus* species are frequently found in association with large struvite stones caused by the precipitation of magnesium ammonium sulfate in an alkaline environment. These stones serve as a nidus for recurrent urinary tract infection. Although a single bacterial species is usually recovered from a perinephric or renal abscess, multiple species may also be found. If a urine culture is not contaminated with periurethral flora and is found to contain more than one organism, a perinephric or renal abscess should be considered in the differential diagnosis. Urine cultures may also be polymicrobial in cases of bladder diverticulum.

Candida species can cause renal abscesses. Fungi of this genus may spread to the kidney hematogenously or by ascension from the bladder. The hallmark of the latter route of infection is ureteral obstruction with large fungal balls.

The presentation of perinephric and renal abscesses is quite nonspecific. Flank pain and abdominal pain are common. At least 50% of patients are febrile. Pain may be referred to the groin or leg, particularly with extension of infection. The diagnosis of perinephric abscess, like that of splenic abscess, is frequently delayed, and the mortality rate in some series is appreciable, although lower than in the past. Perinephric or renal abscess should be most seriously considered when a patient presents with symptoms and signs of pyelonephritis and remains febrile after 4 or 5 days of treatment. Moreover, when a urine culture yields a polymicrobial flora, when a patient is known to have renal stones, or when fever and pyuria coexist with a sterile urine culture, these diagnoses should be entertained.

Renal ultrasonography and abdominal CT are the most useful diagnostic modalities. If a renal or perinephric abscess is diagnosed, nephrolithiasis should be excluded, especially when a high urinary pH suggests the presence of a urea-splitting organism.

TREATMENT

TREATMENT

Perinephric and Renal Abscesses

Treatment for perinephric and renal abscesses, like that for other intraabdominal abscesses, includes drainage of pus and antibiotic therapy directed at the organism(s) recovered. For perinephric abscesses, percutaneous drainage is usually successful.

Psoas Abscesses

The psoas muscle is another location in which abscesses are encountered. Psoas abscesses may arise from a hematogenous source, by contiguous spread from an intraabdominal or pelvic process, or by contiguous spread from nearby bony structures (e.g., vertebral bodies). Associated osteomyelitis due to spread from bone to muscle or from muscle to bone is common in psoas abscesses. When Pott's disease was common, *Mycobacterium tuberculosis* was a frequent cause of psoas abscess. Currently, either *S. aureus* or a mixture of enteric organisms including aerobic and anaerobic gram-negative bacilli is usually isolated from psoas abscesses in the United States. *S. aureus* is most likely to be isolated when a psoas abscess arises from hematogenous spread or a contiguous focus of osteomyelitis; a mixed enteric flora is the most likely etiology when the abscess has an intraabdominal or pelvic source. Patients with psoas abscesses frequently present with fever, lower abdominal or back pain, or pain referred to the hip or knee. CT is the most useful diagnostic technique.

TREATMENT

TREATMENT

Psoas Abscesses

Treatment includes surgical drainage and the administration of an antibiotic regimen directed at the inciting organism(s).

Pancreatic Abscesses

See Chap. 341.

ACKNOWLEDGMENT

The substantial contributions of Dori F. Zaleznik, MD, to this chapter in previous editions are gratefully acknowledged.

FURTHER READING

CONN HO et al (eds): *Spontaneous Bacterial Peritonitis: The Disease, Pathogenesis and Treatment*. New York, Marcel Dekker, 2000.

GOLAN Y: Empiric therapy for hospital-acquired, Gram-negative complicated intra-abdominal infection and complicated urinary tract infections: A systematic literature review of current and emerging treatment options. BMC Infect Dis 15:313, 2015. [PubMed: 26243291]

LI PK et al: ISPD peritonitis recommendations: 2016 update on prevention and treatment. Perit Dial Int 36:481, 2016. [PubMed: 27282851]

SAWYER RG et al: Trial of short-course antimicrobial therapy for intra-abdominal infection. N Engl J Med 372:1996, 2015. [PubMed: 25992746]

SOLOMKIN JS et al: Diagnosis and management of complicated intra-abdominal infection in adults and children: Guidelines by the Surgical Infection Society and the Infectious Diseases Society of America. Clin Infect Dis 50:133, 2010. [PubMed: 20034345]

TENNER S et al: American College of Gastroenterology guideline: Management of acute pancreatitis. Am J Gastroenterol 108:1400, 2013. [PubMed: 23896955]

Chapter 128: Acute Infectious Diarrheal Diseases and Bacterial Food Poisoning

Richelle C. Charles; Stephen B. Calderwood; Regina C. LaRocque

FIGURE 128-1

INTRODUCTION

Acute diarrheal disease is a leading cause of illness globally and is associated with an estimated 1.7 million deaths per year. Among children <5 years of age, diarrheal disease is second only to lower respiratory infection as the most common infectious cause of death. The morbidity from diarrhea also is significant. Recurrent intestinal infections are associated with physical and mental stunting, wasting, micronutrient deficiencies, and malnutrition. In short, diarrheal disease is a driving factor in global morbidity and mortality.

The wide range of clinical manifestations of acute gastrointestinal illnesses is matched by the wide variety of infectious agents involved, including viruses, bacteria, and parasites (Table 128-1). This chapter discusses factors that enable gastrointestinal pathogens to cause disease, reviews host defense mechanisms, and delineates an approach to the evaluation and treatment of patients presenting with acute diarrhea. Individual organisms causing acute gastrointestinal illnesses are discussed in detail in subsequent chapters.

TABLE 128-1

Gastrointestinal Pathogens Causing Acute Diarrhea

Mechanism	Location	Illness	Stool Findings	Examples of Pathogens Involved
Noninflammatory (enterotoxin)	Proximal small bowel	Watery diarrhea	No fecal leukocytes; mild or no increase in fecal lactoferrin	*Vibrio cholerae*, enterotoxigenic *Escherichia coli* (LT and/or ST), enteroaggregative *E. coli*, *Clostridium perfringens*, *Bacillus cereus*, *Staphylococcus aureus*, *Aeromonas hydrophila*, *Plesiomonas shigelloides*, rotavirus, norovirus, enteric adenoviruses, *Giardia lamblia*, *Cryptosporidium* spp., *Cyclospora* spp., microsporidia
Inflammatory (invasion or cytotoxin)	Colon or distal small bowel	Dysentery or inflammatory diarrhea	Fecal polymorphonuclear leukocytes; substantial increase in fecal lactoferrin	*Shigella* spp., *Salmonella* spp., *Campylobacter jejuni*, enterohemorrhagic *E. coli*, enteroinvasive *E. coli*, *Yersinia enterocolitica*, *Listeria monocytogenes*, *Vibrio parahaemolyticus*, *Clostridium difficile*, *A. hydrophila*, *P. shigelloides*, *Entamoeba histolytica*, *Klebsiella oxytoca*
Penetrating	Distal small bowel	Enteric fever	Fecal mononuclear leukocytes	*Salmonella* Typhi, *Y. enterocolitica*

Abbreviations: LT, heat-labile enterotoxin; ST, heat-stable enterotoxin.

PATHOGENIC MECHANISMS

Enteric pathogens have developed a variety of tactics to overcome host defenses. Understanding the virulence factors employed by these organisms is important in the diagnosis and treatment of clinical disease.

INOCULUM SIZE

The number of microorganisms that must be ingested to cause disease varies considerably from species to species. For *Shigella*, enterohemorrhagic *Escherichia coli*, *Giardia lamblia*, or *Entamoeba*, as few as 10–100 bacteria or cysts can produce infection, while 10^5–10^8 *Vibrio cholerae* organisms must be ingested to cause disease. The infective dose of *Salmonella* varies widely, depending on the species, host, and food vehicle. The ability of organisms to overcome host defenses has important implications for transmission; *Shigella*, enterohemorrhagic *E. coli*, *Entamoeba*, and *Giardia* can spread by person-to-person contact, whereas under some circumstances *Salmonella* may need to grow in food for several hours before reaching an effective infectious dose.

ADHERENCE

Many organisms must adhere to the gastrointestinal mucosa as an initial step in the pathogenic process; thus, organisms that can compete with the normal bowel flora and colonize the mucosa have an important advantage in causing disease. Specific cell-surface proteins involved in attachment of bacteria to intestinal cells are important virulence determinants. *V. cholerae*, for example, adheres to the brush border of small-intestinal enterocytes via specific surface adhesins, including the toxin-coregulated pilus and other accessory colonization factors. Enterotoxigenic *E. coli*, which causes watery diarrhea, produces an adherence protein called *colonization factor antigen* that is necessary for colonization of the upper small intestine by the organism prior to the production of enterotoxin. Enteropathogenic *E. coli*, an agent of diarrhea in young children, and enterohemorrhagic *E. coli*, which causes hemorrhagic colitis and the hemolytic-uremic syndrome, produce virulence determinants that allow these organisms to attach to and efface the brush border of the intestinal epithelium.

TOXIN PRODUCTION

The production of one or more exotoxins is important in the pathogenesis of numerous enteric organisms. Such toxins include *enterotoxins*, which cause watery diarrhea by acting directly on secretory mechanisms in the intestinal mucosa; *cytotoxins*, which cause destruction of mucosal cells and associated inflammatory diarrhea; and *neurotoxins*, which act directly on the central or peripheral nervous system.

The prototypical enterotoxin is cholera toxin, a heterodimeric protein composed of one A and five B subunits. The A subunit contains the enzymatic activity of the toxin, while the B subunit pentamer binds holotoxin to the enterocyte surface receptor, the ganglioside GM1. After the binding of holotoxin, a fragment of the A subunit is translocated across the eukaryotic cell membrane into the cytoplasm, where it catalyzes the adenosine diphosphate ribosylation of a guanosine triphosphate–binding protein and causes persistent activation of adenylate cyclase. The end result is an increase of cyclic adenosine monophosphate in the intestinal cell, which increases Cl^- secretion and decreases Na^+ absorption, leading to a loss of fluid and the production of diarrhea.

Enterotoxigenic strains of *E. coli* may produce a protein called *heat-labile enterotoxin* (LT) that is similar to cholera toxin and causes secretory diarrhea by the same mechanism. Alternatively, enterotoxigenic strains of *E. coli* may produce *heat-stable enterotoxin* (ST), one form of which causes diarrhea by activation of guanylate cyclase and elevation of intracellular cyclic guanosine monophosphate. Some enterotoxigenic strains of *E. coli* produce both LT and ST.

Bacterial cytotoxins, in contrast, destroy intestinal mucosal cells and produce the syndrome of dysentery, with bloody stools containing inflammatory cells. Enteric pathogens that produce such cytotoxins include *Shigella dysenteriae* type 1, *Vibrio parahaemolyticus*, and *Clostridium difficile*. *S. dysenteriae* type 1 and Shiga toxin–producing strains of *E. coli* produce potent cytotoxins and have been associated with outbreaks of hemorrhagic colitis and hemolytic-uremic syndrome.

Neurotoxins are usually produced by bacteria outside the host and therefore cause symptoms soon after ingestion. Included are the staphylococcal and *Bacillus cereus* toxins, which act on the central nervous system to produce vomiting.

INVASION

Dysentery may result not only from the production of cytotoxins but also from bacterial invasion and destruction of intestinal mucosal cells. Infections due to *Shigella* and enteroinvasive *E. coli* are characterized by the organisms' invasion of mucosal epithelial cells, intraepithelial multiplication, and subsequent spread to adjacent cells. *Salmonella* causes inflammatory diarrhea by invasion of the bowel mucosa, but generally is not associated with the destruction of enterocytes or the full clinical syndrome of dysentery. *Salmonella* Typhi and *Yersinia enterocolitica* can penetrate intact intestinal mucosa, multiply intracellularly in Peyer's patches and intestinal lymph nodes, and then disseminate through the bloodstream to cause enteric fever—a syndrome characterized by fever, headache, relative bradycardia, abdominal pain, splenomegaly, and leukopenia.

HOST DEFENSES

Given the enormous number of microorganisms ingested with every meal, the normal host must combat a constant influx of potential enteric pathogens. Studies of infections in patients with alterations in defense mechanisms have led to a greater understanding of the variety of ways in which the normal host can protect itself against disease.

INTESTINAL MICROBIOTA

The large numbers of bacteria that normally inhabit the intestine (*the intestinal microbiota*) act as an important host defense mechanism, preventing colonization by potential enteric pathogens. Persons with fewer intestinal bacteria, such as infants who have not yet developed normal enteric colonization or patients receiving antibiotics, are at significantly greater risk of developing infections with enteric pathogens. The composition of the intestinal microbiota is as important as the number of organisms present. More than 99%

of the normal colonic microbiota is made up of anaerobic bacteria, and the acidic pH and volatile fatty acids produced by these organisms appear to be critical elements in resistance to colonization.

GASTRIC ACID

The acidic pH of the stomach is an important barrier to enteric pathogens, and an increased frequency of infections due to *Salmonella*, *G. lamblia*, and a variety of helminths has been reported among patients who have undergone gastric surgery or are achlorhydric for some other reason. Neutralization of gastric acid with antacids, proton pump inhibitors, or H_2 blockers—a common practice in the management of hospitalized patients—similarly increases the risk of enteric colonization. In addition, some microorganisms can survive the extreme acidity of the gastric environment; rotavirus, for example, is highly stable to acidity.

INTESTINAL MOTILITY

Normal peristalsis is the major mechanism for clearance of bacteria from the proximal small intestine. When intestinal motility is impaired (e.g., by treatment with opiates or other antimotility drugs, anatomic abnormalities, or hypomotility states), the frequency of bacterial overgrowth and infection of the small bowel with enteric pathogens is increased. Some patients whose treatment for *Shigella* infection consists of diphenoxylate hydrochloride with atropine (Lomotil) experience prolonged fever and shedding of organisms, while patients treated with opiates for mild *Salmonella* gastroenteritis have a higher frequency of bacteremia than those not treated with opiates.

INTESTINAL MUCIN

A complex layer of mucus, produced by specialized secretory cells, covers the stomach, small intestine, and large intestine and separates the commensal microbiota from the epithelium. The thickness and constituents of this mucus barrier vary throughout the gastrointestinal tract. The mucus barrier turns over rapidly and comprises glycoproteins and a range of antimicrobial molecules and secreted immunoglobulins directed against specific microbial antigens. Enteric pathogens have evolved a wide range of strategies to overcome this barrier and thus to reach the underlying epithelium and cause disease. For example, pathogens can penetrate the mucus layer by secreting enzymes to degrade the mucus or through flagella-mediated motility. Some organisms, such as *Shigella*, secrete toxins that can diffuse through the mucus layer and disrupt the underlying epithelium. The resulting reduction of mucus production allows the pathogen to reach the cell surface.

IMMUNITY

Both cellular immune responses and antibody production play important roles in protection from enteric infections. Humoral immunity to enteric pathogens consists of systemic IgG and IgM as well as secretory IgA. The mucosal immune system may be the first line of defense against many gastrointestinal pathogens. The binding of bacterial antigens to the luminal surface of M cells in the distal small bowel and the subsequent presentation of antigens to subepithelial lymphoid tissue lead to the proliferation of sensitized lymphocytes.

These lymphocytes circulate and populate all of the mucosal tissues of the body as IgA-secreting plasma cells.

GENETIC DETERMINANTS



Host genetic variation influences susceptibility to diarrheal diseases. People with blood group O show increased susceptibility to disease due to *V. cholerae*, *Shigella*, *E. coli* O157, and norovirus. Polymorphisms in genes encoding inflammatory mediators have been associated with the outcome of infection with enteroaggregative *E. coli*, enterotoxin-producing *E. coli*, *Salmonella*, *C. difficile*, and *V. cholerae*.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Infectious Diarrhea or Bacterial Food Poisoning

The approach to the patient with possible infectious diarrhea or bacterial food poisoning is shown in **Fig. 128-1**.

HISTORY

The answers to questions with high discriminating value can quickly narrow the range of potential causes of diarrhea and help determine whether treatment is needed. Important elements of the narrative history are detailed in Fig. 128-1.

PHYSICAL EXAMINATION

The examination of patients for signs of dehydration provides essential information about the severity of the diarrheal illness and the need for rapid therapy. Mild dehydration is indicated by thirst, dry mouth, decreased axillary sweat, decreased urine output, and slight weight loss. Signs of moderate dehydration include an orthostatic fall in blood pressure, skin tenting, and sunken eyes (or, in infants, a sunken fontanelle). Signs of severe dehydration include lethargy, obtundation, feeble pulse, hypotension, and frank shock.

DIAGNOSTIC APPROACH

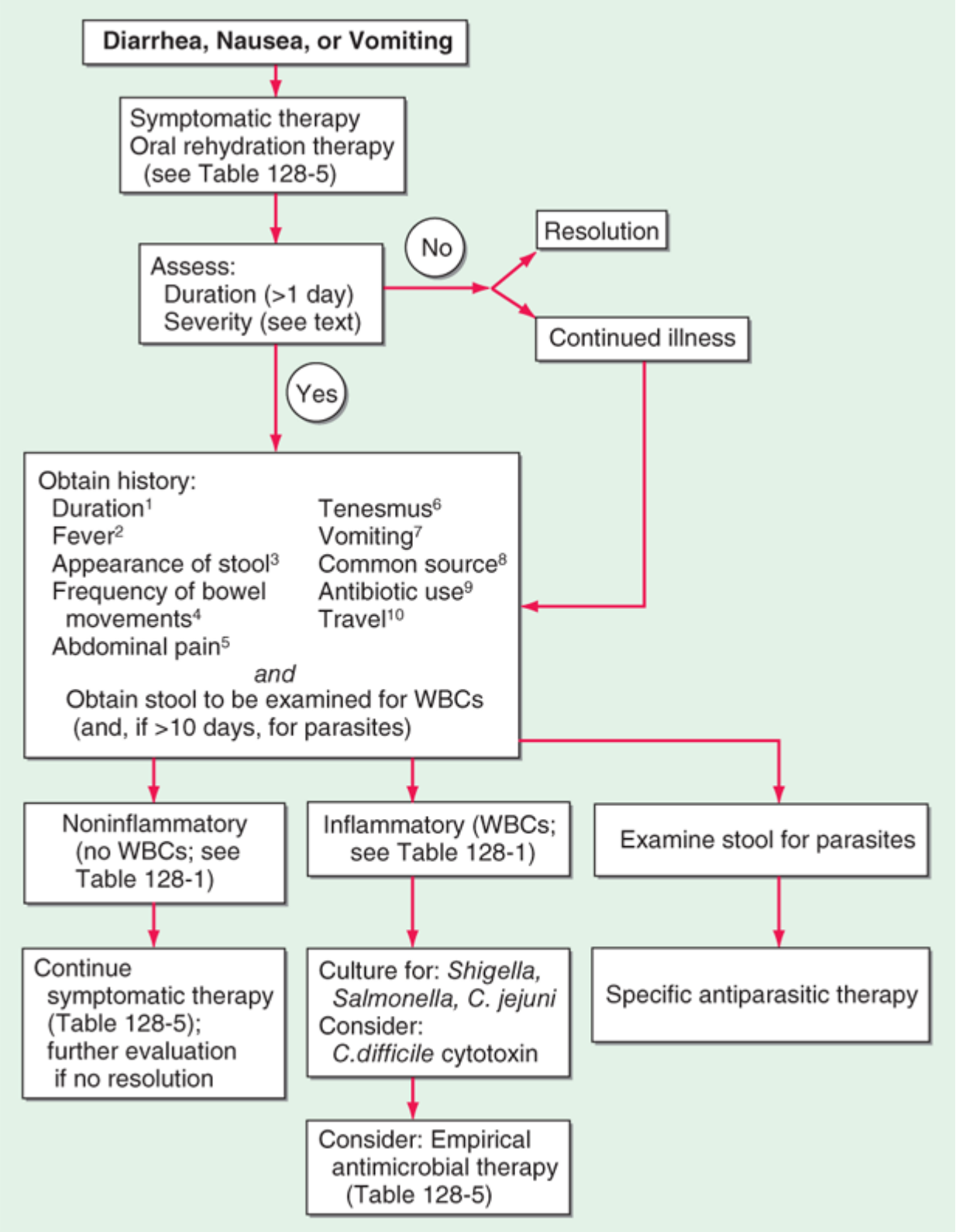
After the severity of illness is assessed, the clinician must distinguish between *inflammatory* and *noninflammatory* disease. Using the history and epidemiologic features of the case as guides, the clinician can then rapidly evaluate the need for further efforts to define a specific etiology and for therapeutic intervention. Examination of a stool sample may supplement the narrative history. Grossly bloody or mucoid stool suggests an inflammatory process. A test for fecal leukocytes (preparation of a thin smear of stool on a glass slide, addition of a drop of methylene blue, and examination of the wet mount) can suggest inflammatory disease in patients with diarrhea, although the predictive value of this test is still debated. A test for fecal lactoferrin, which is a marker of fecal leukocytes, is more sensitive and is available in latex agglutination and enzyme-linked immunosorbent assay formats. Causes of acute infectious diarrhea, categorized as inflammatory and noninflammatory, are listed in Table 128-1.

POST-DIARRHEA COMPLICATIONS

Chronic complications may follow the resolution of an acute diarrheal episode. The clinician should inquire about prior diarrheal illness if the conditions listed in **Table 128-2** are observed.

FIGURE 128-1

Clinical algorithm for the approach to patients with community-acquired infectious diarrhea or bacterial food poisoning. Key to superscripts: 1. Diarrhea lasting >2 weeks is generally defined as chronic; in such cases, many of the causes of acute diarrhea are much less likely, and a new spectrum of causes needs to be considered. 2. Fever often implies invasive disease, although fever and diarrhea may also result from infection outside the gastrointestinal tract, as in malaria. 3. Stools that contain blood or mucus indicate ulceration of the large bowel. Bloody stools without fecal leukocytes should alert the laboratory to the possibility of infection with Shiga toxin–producing enterohemorrhagic *Escherichia coli*. Bulky white stools suggest a small-intestinal process that is causing malabsorption. Profuse "rice-water" stools suggest cholera or a similar toxigenic process. 4. Frequent stools over a given period can provide the first warning of impending dehydration. 5. Abdominal pain may be most severe in inflammatory processes like those due to *Shigella*, *Campylobacter*, and necrotizing toxins. Painful abdominal muscle cramps, caused by electrolyte loss, can develop in severe cases of cholera. Bloating is common in giardiasis. An appendicitis-like syndrome should prompt a culture for *Yersinia enterocolitica* with cold enrichment. 6. Tenesmus (painful rectal spasms with a strong urge to defecate but little passage of stool) may be a feature of cases with proctitis, as in shigellosis or amebiasis. 7. Vomiting implies an acute infection (e.g., a toxin-mediated illness or food poisoning) but can also be prominent in a variety of systemic illnesses (e.g., malaria) and in intestinal obstruction. 8. Asking patients whether anyone else they know is sick is a more efficient means of identifying a common source than is constructing a list of recently eaten foods. If a common source seems likely, specific foods can be investigated. See text for a discussion of bacterial food poisoning. 9. Current antibiotic therapy or a recent history of treatment suggests *Clostridium difficile* diarrhea **(Chap. 129)**. Stop antibiotic treatment if possible and consider tests for *C. difficile* toxins. Antibiotic use may increase the risk of chronic intestinal carriage following salmonellosis. 10. See text (and **Chap. 119**) for a discussion of traveler's diarrhea. *(After TS Steiner, RL Guerrant: Principles and syndromes of enteric infection, in Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases, 7th ed, GL Mandell et al [eds]. Philadelphia, Churchill Livingstone, 2010, pp 1335–1351; RL Guerrant, DA Bobak: N Engl J Med 325:327, 1991; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 128-2

Post-Diarrhea Complications of Acute Infectious Diarrheal Illness

Complication	Comments
Chronic diarrhea (diarrhea lasting >4 weeks) • Lactase deficiency • Small-bowel bacterial overgrowth • Malabsorption syndromes (tropical and celiac sprue)	Occurs in ~1% of travelers with acute diarrhea • Protozoa account for ~1/3 of cases
Initial presentation or exacerbation of inflammatory bowel disease	May be precipitated by traveler's diarrhea
Irritable bowel syndrome	Occurs in ~10% of travelers with traveler's diarrhea
Reactive arthritis	Particularly likely after infection with invasive organisms (*Shigella*, *Salmonella*, *Campylobacter*, *Yersinia*)
Hemolytic-uremic syndrome (hemolytic anemia, thrombocytopenia, and renal failure)	Follows infection with Shiga toxin–producing bacteria (*Shigella dysenteriae* type 1 and enterohemorrhagic *Escherichia coli*)
Guillain-Barré syndrome	Particularly likely after *Campylobacter* infection

EPIDEMIOLOGY

TRAVEL HISTORY

Of the several million people who travel from temperate industrialized countries to tropical regions of Asia, Africa, and Central and South America each year, 20–50% experience a sudden onset of abdominal cramps, anorexia, and watery diarrhea; thus *traveler's diarrhea* is the most common travel-related infectious illness (Chap. 119). The time of onset is usually 3 days to 2 weeks after the traveler's arrival in a resource-poor area; most cases begin within the first 3–5 days. The illness is generally self-limited, lasting 1–5 days. The high rate of diarrhea among travelers to underdeveloped areas is related to the ingestion of contaminated food or water.

The organisms that cause traveler's diarrhea vary considerably with location **(Table 128-3)**, as does the pattern of antimicrobial resistance. In all areas, enterotoxigenic and enteroaggregative strains of *E. coli* are the most common isolates from persons with the classic secretory traveler's diarrhea syndrome. Infection with *Campylobacter jejuni* is especially common in areas of Asia.

TABLE 128-3

Causes of Traveler's Diarrhea

Etiologic Agent	Approximate Percentage of Cases	Comments
Bacteria	**50–75**	
Enterotoxigenic *Escherichia coli*	10–45	Single most important agent
Enteroaggregative *E. coli*	5–35	Emerging enteric pathogen with worldwide distribution
Campylobacter jejuni	5–25	More common in Asia
Shigella	0–15	Major cause of dysentery
Salmonella	0–15	—
Others	0–5	Including *Aeromonas*, *Plesiomonas*, and *Vibrio cholerae*
Viruses	**0–20**	
Norovirus	0–10	Associated with cruise ships
Rotavirus	0–5	Particularly common among children
Parasites	**0–10**	
Giardia lamblia	0–5	Affects hikers and campers who drink from freshwater streams
Cryptosporidium	0–5	Resistant to chlorine treatment of water sources
Entamoeba histolytica	<1	—
Cyclospora	<1	—
Other	**0–10**	

Etiologic Agent	Approximate Percentage of Cases	Comments
Acute food poisoning[a]	0–5	—
No pathogen identified	10–50	—

[a]For etiologic agents, see Table 128-4.

Source: After DR Hill et al: The practice of travel medicine: Guidelines by the Infectious Diseases Society of America. Clin Infect Dis 43:1499, 2006.

LOCATION

Closed and semi-closed communities, including day-care centers, schools, residential facilities, and cruise ships, are important settings for outbreaks of enteric infections. Norovirus, which is highly contagious and robust in surviving on surfaces, is the most common etiologic agent associated with outbreaks of acute gastroenteritis. Other common organisms, often spread by fecal–oral contact in such communities, are *Shigella*, *C. jejuni*, and *Cryptosporidium*. Rotavirus is rarely a cause of pediatric diarrheal outbreaks in the United States since rotavirus vaccination was broadly recommended in 2006. Similarly, hospitals are sites in which enteric infections are concentrated. Diarrhea is one of the most common manifestations of nosocomial infections. *C. difficile* is the predominant cause of nosocomial diarrhea among adults in the United States, and outbreaks of norovirus infection are common in health care settings. *Klebsiella oxytoca* has been identified as a cause of antibiotic-associated hemorrhagic colitis. Enteropathogenic *E. coli* has been associated with outbreaks of diarrhea in nurseries for newborns. One-third of elderly patients in chronic-care institutions develop a significant diarrheal illness each year; more than one-half of these cases are caused by cytotoxin-producing *C. difficile*. Antimicrobial therapy can predispose to pseudomembranous colitis by altering the normal colonic flora and allowing the multiplication of *C. difficile* (Chap. 129).

AGE

Globally, most morbidity and mortality from enteric pathogens involves children <5 years of age. Breast-fed infants are protected from pathogens in contaminated food and water and derive some protection from maternal antibodies, but their risk of infection rises dramatically when they begin to eat solid foods. Exposure to rotavirus is universal, with most children experiencing their first infection in the first or second year of life if not vaccinated. Older children and adults are more commonly infected with norovirus. Other organisms with higher attack rates among children than among adults include enterotoxigenic, enteropathogenic, and enterohemorrhagic *E. coli*; *Shigella*; *C. jejuni*; and *G. lamblia*.

HOST IMMUNE STATUS

Immunocompromised hosts are at elevated risk of acute and chronic infectious diarrhea. Individuals with defects in cell-mediated immunity (including those with AIDS) are at particularly high risk of invasive enteropathies, including salmonellosis, listeriosis, and cryptosporidiosis. Individuals with hypogammaglobulinemia are at particular risk of *C. difficile* colitis and giardiasis. Patients with cancer are more likely to develop *C. difficile* infection as a result of chemotherapy and frequent hospitalizations. Infectious diarrhea can be life-threatening in immunocompromised hosts, with complications including bacteremia and metastatic seeding of infection. Furthermore, dehydration may compromise renal function and increase the toxicity of immunosuppressive drugs.

BACTERIAL FOOD POISONING

If the history and the stool examination indicate a noninflammatory etiology of diarrhea and there is evidence of a common-source outbreak, questions concerning the ingestion of specific foods and the time of onset of diarrhea after a meal can provide clues to the bacterial cause of the illness. Potential causes of bacterial food poisoning are shown in **Table 128-4**.

TABLE 128-4

Bacterial Food Poisoning

Incubation Period, Organism	Symptoms	Common Food Sources
1–6 h		
Staphylococcus aureus	Nausea, vomiting, diarrhea	Ham, poultry, potato or egg salad, mayonnaise, cream pastries
Bacillus cereus	Nausea, vomiting, diarrhea	Fried rice
8–16 h		
Clostridium perfringens	Abdominal cramps, diarrhea (vomiting rare)	Beef, poultry, legumes, gravies
B. cereus	Abdominal cramps, diarrhea (vomiting rare)	Meats, vegetables, dried beans, cereals
>16 h		
Vibrio cholerae	Watery diarrhea	Shellfish, water
Enterotoxigenic *Escherichia coli*	Watery diarrhea	Salads, cheese, meats, water
Enterohemorrhagic *E. coli*	Bloody diarrhea	Ground beef, roast beef, salami, raw milk, raw vegetables, apple juice
Salmonella spp.	Inflammatory diarrhea	Beef, poultry, eggs, dairy products
Campylobacter jejuni	Inflammatory diarrhea	Poultry, raw milk
Shigella spp.	Dysentery	Potato or egg salad, lettuce, raw vegetables
Vibrio parahaemolyticus	Dysentery	Mollusks, crustaceans

Bacterial disease caused by an enterotoxin elaborated outside the host, such as that due to *Staphylococcus aureus* or *B. cereus*, has the shortest incubation period (1–6 h) and generally lasts <12 h. Most cases of staphylococcal food poisoning are caused by contamination from infected human carriers. Staphylococci can multiply at a wide range of temperatures; thus, if food is left to cool slowly and remains at room temperature after cooking, the organisms will have the opportunity to form enterotoxin. Outbreaks following picnics where potato salad, mayonnaise, and cream pastries have been served offer classic examples of staphylococcal food poisoning. Diarrhea, nausea, vomiting, and abdominal cramping are common, while fever is less so.

B. cereus can produce either a syndrome with a short incubation period—the *emetic* form, mediated by a staphylococcal type of enterotoxin—or one with a longer incubation period (8–16 h)—the *diarrheal* form, caused by an enterotoxin resembling *E. coli* LT, in which diarrhea and abdominal cramps are characteristic but vomiting is uncommon. The emetic form of *B. cereus* food poisoning is associated with contaminated fried rice; the organism is common in uncooked rice, and its heat-resistant spores survive boiling. If cooked rice is not refrigerated, the spores can germinate and produce toxin. Frying before serving may not destroy the preformed, heat-stable toxin.

Food poisoning due to *Clostridium perfringens* also has a slightly longer incubation period (8–14 h) and results from the survival of heat-resistant spores in inadequately cooked meat, poultry, or legumes. After ingestion, toxin is produced in the intestinal tract, causing moderately severe abdominal cramps and diarrhea; vomiting is rare, as is fever. The illness is self-limited, rarely lasting >24 h.

Not all food poisoning has a bacterial cause. Nonbacterial agents of short-incubation food poisoning include capsaicin, which is found in hot peppers, and a variety of toxins found in fish and shellfish (Chap. 451).

LABORATORY EVALUATION

Many cases of noninflammatory diarrhea are self-limited or can be treated empirically, and in these instances, the clinician may not need to determine a specific etiology. Potentially pathogenic *E. coli* cannot be distinguished from normal fecal flora by routine culture, and tests to detect enterotoxins are not available in most clinical laboratories. In situations in which cholera is a concern, stool should be cultured on selective media such as thiosulfate–citrate–bile salts–sucrose (TCBS) or tellurite–taurocholate–gelatin (TTG) agar. A latex agglutination test has made the rapid detection of rotavirus in stool practical for many laboratories, while reverse-transcriptase polymerase chain reaction (PCR) and specific antigen enzyme immunoassays have been developed for the identification of norovirus. Stool specimens should be examined by immunofluorescence-based rapid assays or (less sensitive) standard microscopy for *Giardia* cysts or *Cryptosporidium* if the level of clinical suspicion regarding the involvement of these organisms is high.

All patients with fever and evidence of inflammatory disease acquired outside the hospital should have stool evaluated for *Salmonella*, *Shigella*, and *Campylobacter*. *Salmonella* and *Shigella* can be selected on MacConkey agar as non-lactose-fermenting (colorless) colonies or can be grown on *Salmonella–Shigella* agar or in selenite enrichment broth, both of which inhibit most organisms except these pathogens. Evaluation of

nosocomial diarrhea should initially focus on *C. difficile*; stool culture for other pathogens in this setting has an extremely low yield and is not cost-effective. Toxins A and B produced by pathogenic strains of *C. difficile* can be detected by rapid enzyme immunoassays, latex agglutination tests, or PCR (Chap. 129). Isolation of *C. jejuni* requires inoculation of fresh stool onto selective growth medium and incubation at 42°C in a microaerophilic atmosphere. In many laboratories in the United States, *E. coli* O157:H7 is among the most common pathogens isolated from visibly bloody stools. Strains of this enterohemorrhagic serotype can be identified in specialized laboratories by serotyping but also can be identified presumptively in hospital laboratories as lactose-fermenting, indole-positive colonies of sorbitol nonfermenters (white colonies) on sorbitol MacConkey plates. If the clinical presentation suggests the possibility of intestinal amebiasis, stool should be examined by a rapid antigen detection assay or by (less sensitive and less specific) microscopy. Multiplex nucleic acid amplification methods for detection of many stool pathogens (viral, bacterial, and parasitic) are increasingly being used in clinical microbiology laboratories to decrease the time to detection of a pathogen. Although these tests may be more sensitive and rapid than standard culture methods, the lack of a microbial isolate prevents determination of antimicrobial susceptibility and typing of strains by public health authorities in order to detect and respond to common-source outbreaks. For this reason, the Centers for Disease Control and Prevention suggests that diagnosis of an enteric bacterial infection by a nucleic acid amplification method should be followed by attempted isolation of the pathogen by culture.

TREATMENT

TREATMENT

Infectious Diarrhea or Bacterial Food Poisoning

In many cases, a specific diagnosis is not necessary or not available to guide treatment. The clinician can proceed with the information obtained from the history, stool examination, and evaluation of dehydration severity. Empirical regimens for the treatment of traveler's diarrhea are listed in Table 128-5.

The mainstay of treatment is adequate rehydration. The treatment of cholera and other dehydrating diarrheal diseases was revolutionized by the promotion of oral rehydration solution (ORS), the efficacy of which depends on the fact that glucose-facilitated absorption of sodium and water in the small intestine remains intact in the presence of cholera toxin. The use of ORS has reduced cholera mortality rates from >50% (in untreated cases) to <1%. A number of ORS formulas have been used. Initial preparations were based on the treatment of patients with cholera and included a solution containing 3.5 g of sodium chloride, 2.5 g of sodium bicarbonate (or 2.9 g of sodium citrate), 1.5 g of potassium chloride, and 20 g of glucose (or 40 g of sucrose) per liter of water. Such a preparation can still be used for the treatment of severe cholera. Many causes of secretory diarrhea, however, are associated with less electrolyte loss than occurs in cholera. Beginning in 2002, the World Health Organization recommended a "reduced-osmolarity/reduced-salt" ORS that is better tolerated and more effective than classic ORS. This preparation contains 2.6 g of sodium chloride, 2.9 g of trisodium citrate, 1.5 g of potassium chloride, and 13.5 g of glucose (or 27 g of sucrose) per liter of water. ORS formulations containing rice or cereal as the carbohydrate source may be even more

effective than glucose-based solutions. Patients who are severely dehydrated or in whom vomiting precludes oral therapy should receive IV solutions such as Ringer's lactate.

Most secretory forms of traveler's diarrhea (usually due to enterotoxigenic or enteroaggregative *E. coli* or to *Campylobacter*) can be treated effectively with rehydration, bismuth subsalicylate, or antiperistaltic agents. Antimicrobial agents can shorten the duration of illness from 3–4 days to 24–36 h but may be associated with the acquisition of multidrug-resistant organisms. Changes in diet have not been shown to have an impact on the duration of illness, while the efficacy of probiotics continues to be debated. Most individuals who present with dysentery (bloody diarrhea and fever) should be treated empirically with an antimicrobial agent (e.g., a fluoroquinolone or a macrolide) pending microbiologic analysis of stool. Individuals with shigellosis should receive a 3- to 7-day course. Individuals with more severe or prolonged *Campylobacter* infection often benefit from antimicrobial treatment as well. Because of widespread resistance of *Campylobacter* to fluoroquinolones, especially in parts of Asia, a macrolide antibiotic such as erythromycin or azithromycin may be preferred for this infection.

Treatment of salmonellosis must be tailored to the individual patient. Since administration of antimicrobial agents often prolongs intestinal colonization with *Salmonella*, these drugs are usually reserved for individuals at high risk of complications from disseminated salmonellosis, such as infants, patients with prosthetic devices, patients over age 50, and immunocompromised persons. Antimicrobial agents should not be administered to individuals (especially children) in whom enterohemorrhagic *E. coli* infection is suspected. Laboratory studies of enterohemorrhagic *E. coli* strains have demonstrated that a number of antibiotics induce replication of Shiga toxin–producing lambdoid bacteriophages, thereby significantly increasing toxin production by these strains. Clinical studies have supported these laboratory results, and antibiotics may increase by twentyfold the risk of hemolytic-uremic syndrome and renal failure during enterohemorrhagic *E. coli* infection. A clinical clue in the diagnosis of the latter infection is bloody diarrhea with low fever or none at all.

TABLE 128-5

Treatment of Traveler's Diarrhea on the Basis of Clinical Features[a]

Clinical Syndrome	Suggested Therapy
Watery diarrhea (no blood in stool, no fever), 1 or 2 unformed stools per day without distressing enteric symptoms	Oral fluids (oral rehydration solution, Pedialyte, Lytren, or flavored mineral water) and saltine crackers
Watery diarrhea (no blood in stool, no fever), 1 or 2 unformed stools per day with distressing enteric symptoms	Bismuth subsalicylate (for adults): 30 mL or 2 tablets (262 mg/tablet) every 30 min for 8 doses; or loperamide[b]: 4 mg initially followed by 2 mg after passage of each unformed stool, not to exceed 8 tablets (16 mg) per day (prescription dose) or 4 caplets (8 mg) per day (over-the-counter dose); drugs can be taken for 2 days. Antibacterial drug[c] can be considered in selected circumstances.
Dysentery (passage of bloody stools) or fever (>37.8°C)	Antibacterial drug[c]
Vomiting, minimal diarrhea	Bismuth subsalicylate (for adults; see dose above)
Diarrhea in infants (<2 years old)	Fluids and electrolytes (oral rehydration solution, Pedialyte, Lytren); continue feeding, especially with breast milk; seek medical attention for moderate dehydration, fever lasting >24 h, bloody stools, or diarrhea lasting more than several days

[a]All patients should take oral fluids (Pedialyte, Lytren, or flavored mineral water) plus saltine crackers. If diarrhea becomes moderate or severe, if fever persists, or if bloody stools or dehydration develops, the patient should seek medical attention. [b]Loperamide should not be used by patients with fever or dysentery; its use may prolong diarrhea in patients with infection due to *Shigella* or other invasive organisms. [c]The recommended antibacterial drugs are as follows:

If the level of suspicion is low for fluoroquinolone-resistant *Campylobacter*:

Adults: (1) A fluoroquinolone such as ciprofloxacin, 750 mg as a single dose or 500 mg bid for 3 days; levofloxacin, 500 mg as a single dose or 500 mg qd for 3 days; or norfloxacin, 800 mg as a single dose or 400 mg bid for 3 days. (2) Azithromycin, 1000 mg as a single dose or 500 mg qd for 3 days. (3) Rifaximin, 200 mg tid or 400 mg bid for 3 days (not recommended for use in dysentery).

Children: Azithromycin, 10 mg/kg on day 1, 5 mg/kg on days 2 and 3 if diarrhea persists.

If fluoroquinolone-resistant *Campylobacter* is suspected (for example, following travel to Southeast Asia):

Adults: Azithromycin (at above dose for adults). *Children:* Same as for children traveling to other areas (see above).

Source: After DR Hill et al: The practice of travel medicine: Guidelines by the Infectious Diseases Society of America. Clin Infect Dis 43:1499, 2006.

PROPHYLAXIS

Improvements in hygiene to limit fecal–oral spread of enteric pathogens will be necessary if the prevalence of diarrheal diseases is to be significantly reduced in developing countries. Travelers can reduce their risk of diarrhea by eating only hot, freshly cooked food; by avoiding raw vegetables, salads, and unpeeled fruit; and by drinking only boiled or treated water and avoiding ice. Historically, few travelers to tourist destinations adhere to these dietary restrictions. Bismuth subsalicylate is an inexpensive agent for the prophylaxis of traveler's diarrhea; it is taken at a dosage of 2 tablets (525 mg) four times a day. Treatment appears to be effective and safe for up to 3 weeks, but adverse events such as temporary darkening of the tongue and tinnitus can occur. A meta-analysis suggests that probiotics may lessen the likelihood of traveler's diarrhea by ~15%. Prophylactic antimicrobial agents, although effective, are not generally recommended for the prevention of traveler's diarrhea except when travelers are immunosuppressed or have other underlying illnesses that place them at high risk for morbidity from gastrointestinal infection. If prophylaxis is indicated, the nonabsorbed antibiotic rifaximin can be considered.

The possibility of exerting a major impact on the worldwide morbidity and mortality associated with diarrheal diseases has led to intensive efforts to develop effective vaccines against the common bacterial and viral enteric pathogens. An effective rotavirus vaccine is available. Vaccines against *S.* Typhi and *V. cholerae* also are available, although the protection they offer is incomplete and/or short lived. At present, there are no effective commercially available vaccines against *Shigella*, *Campylobacter*, nontyphoidal *Salmonella*, norovirus, or intestinal parasites.

ACKNOWLEDGMENT

The authors thank Edward T. Ryan, MD, for his significant contributions to this chapter in previous editions.

FURTHER READING

Goldenberg JZ et al: Probiotics for the prevention of pediatric antibiotic-associated diarrhea. Cochrane Database Syst Rev 12:CD004827, 2015.

Guerrant RL et al: Practice guidelines for the management of infectious diarrhea. Clin Infect Dis 32:331, 2001.

Kotloff KL et al: Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): A prospective, case-control study. Lancet 382:209, 2013.

Lanata CF et al: Global causes of diarrheal disease mortality in children <5 years of age: A systematic review. PLoS One 8:e72788, 2013. [PubMed: 24023773]

Steffen R et al: Traveler's diarrhea: A clinical review. JAMA 313:71, 2015.

Chapter 129: *Clostridium difficile* Infection, Including Pseudomembranous Colitis

Dale N. Gerding; Stuart Johnson

DEFINITION

Clostridium difficile infection (CDI) is a unique colonic disease that is acquired most commonly in association with antimicrobial use and the consequent disruption of the normal colonic microbiota. The most commonly diagnosed diarrheal illness acquired in the hospital, CDI results from the ingestion of spores of *C. difficile* that vegetate, multiply, and secrete toxins, causing diarrhea and, in the most severe cases, pseudomembranous colitis (PMC).

ETIOLOGY AND EPIDEMIOLOGY

C. difficile is an obligately anaerobic, gram-positive, spore-forming bacillus whose spores are found widely in nature, particularly in the environment of hospitals and chronic-care facilities. CDI occurs frequently in hospitals and nursing homes (or shortly after discharge from these facilities) where the level of antimicrobial use is high and the environment is contaminated by *C. difficile* spores.

Clindamycin, ampicillin, and cephalosporins were the first antibiotics associated with CDI. The second- and third-generation cephalosporins, particularly cefotaxime, ceftriaxone, cefuroxime, and ceftazidime, are agents frequently responsible for this condition, and the fluoroquino-lones (ciprofloxacin, levofloxacin, and moxifloxacin) are the most recent drug class to be implicated in hospital outbreaks. Penicillin/β-lactamase-inhibitor combinations such as ticarcillin/clavulanate and piperacillin/tazobactam pose significantly less risk. However, all antibiotics, including vancomycin and metronidazole (the agents most commonly used to treat CDI), have been found to carry a risk of subsequent CDI. A few cases, especially in the community, are reported in patients without documentation of prior antibiotic exposure.

C. difficile is acquired exogenously—most often in the hospital or nursing home, but also in the outpatient setting—and is carried in the stool of both symptomatic and asymptomatic patients. The rate of fecal colonization increases in proportion to length of hospital stay and is often ≥20% among adult patients hospitalized for >2 weeks; in contrast, the rate is 1–3% among community residents. CDI is now the most common health care–associated infection in the United States, with an estimated 453,000 cases annually. The incidence is higher among female patients, Caucasians, and persons ≥65 years of age. The estimated

number of first recurrences of CDI is 83,000, and the estimated number of CDI-associated deaths is 29,300. Community-onset CDI without recent hospitalization, nursing home residence, or outpatient health-care contact probably accounts for ≤10% of all cases.

Asymptomatic fecal carriage of *C. difficile* in healthy neonates is very common, with repeated colonization by multiple strains in infants <1–2 years of age, but associated disease in these infants is extremely rare if it occurs at all. Spores of *C. difficile* are found on environmental surfaces (where the organism can persist for months) and on the hands of hospital personnel who fail to practice good hand hygiene. Hospital epidemics of CDI have been attributed to a single *C. difficile* strain and to multiple strains present simultaneously. Other identified risk factors for CDI include older age, greater severity of underlying illness, gastrointestinal surgery, use of electronic rectal thermometers, enteral tube feeding, and antacid treatment. Use of proton pump inhibitors may be a risk factor, but this risk is probably modest, and no firm data have implicated these agents in patients who are not already receiving antibiotics.

PATHOLOGY AND PATHOGENESIS

Spores of toxigenic *C. difficile* are ingested, survive gastric acidity, germinate in the small bowel, and colonize the lower intestinal tract, where they elaborate two large toxins: toxin A (an enterotoxin) and toxin B (a cytotoxin). These toxins initiate processes resulting in the disruption of epithelial-cell barrier function, diarrhea, and pseudomembrane formation. Toxin A is a potent neutrophil chemoattractant, and both toxins glucosylate the GTP-binding proteins of the Rho subfamily that regulate the actin cell cytoskeleton. Data from studies using molecular disruption of toxin genes in isogenic mutants suggest that toxin B may be the more important virulence factor; this possibility, if confirmed, might account for the occurrence of clinical disease caused by toxin A–negative strains but not by toxin B-negative strains. Disruption of the cytoskeleton results in loss of cell shape, adherence, and tight junctions, with consequent fluid leakage. A third toxin, binary toxin CDT, was previously found in only ~6% of strains but is present in all isolates of the widely recognized epidemic NAP1/BI/027 strain (see "Global Considerations," below); this toxin is related to *C. perfringens* iota toxin. Its role in the pathogenesis of CDI has not yet been defined.

The pseudomembranes of PMC are confined to the colonic mucosa and initially appear as 1- to 2-mm whitish-yellow plaques. The intervening mucosa appears unremarkable, but, as the disease progresses, the pseudomembranes coalesce to form larger plaques and become confluent over the entire colon wall **(Fig. 129-1)**. The whole colon is usually involved, but 10% of patients have rectal sparing. Viewed microscopically, the pseudomembranes have a mucosal attachment point and contain necrotic leukocytes, fibrin, mucus, and cellular debris. The epithelium is eroded and necrotic in focal areas, with neutrophil infiltration of the mucosa.

Figure 129-1

Autopsy specimen showing confluent pseudomembranes covering the cecum of a patient with pseudomembranous colitis. Note the sparing of the terminal ileum (*arrow*).



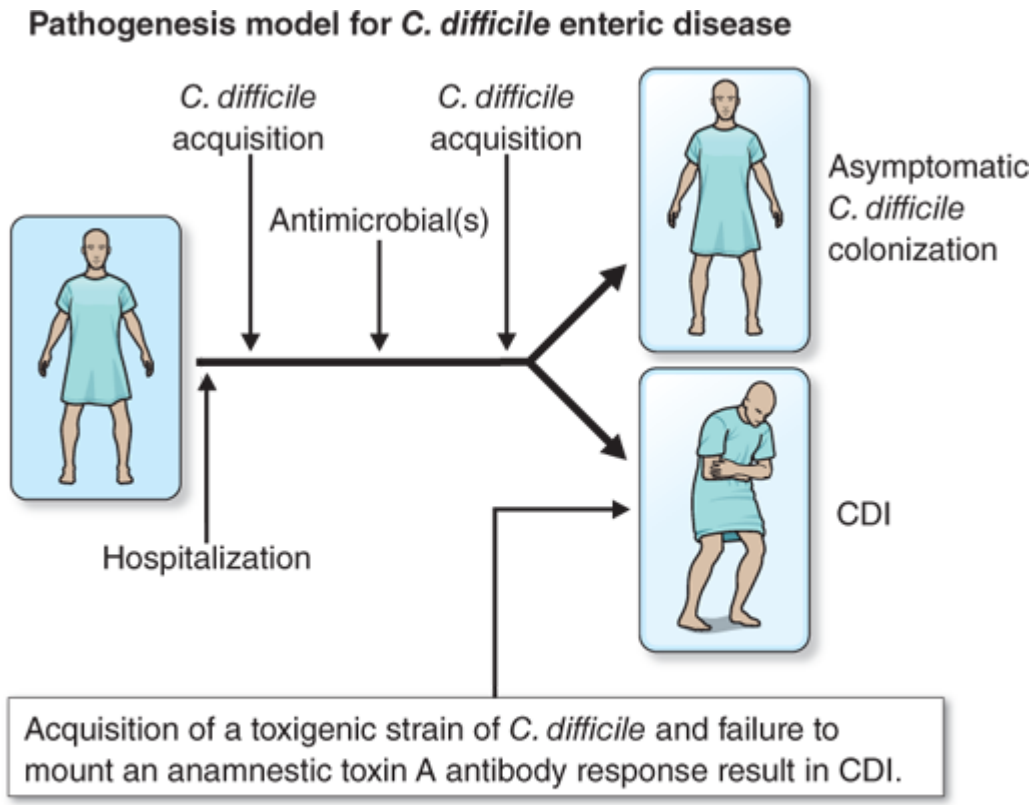
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Patients colonized with *C. difficile* were initially thought to be at high risk for CDI. However, four prospective studies have shown that colonized patients who have not previously had CDI actually have a decreased risk of CDI, possibly because many of these patients are colonized by nontoxigenic strains. At least three events are proposed as essential for the development of CDI (**Fig. 129-2**). Exposure to antimicrobial agents is the first event and establishes susceptibility to CDI, most likely through disruption of the normal gastrointestinal microbiota. The second event is exposure to toxigenic *C. difficile*. Given that the majority of patients do not develop CDI after the first two events, a third event is clearly essential for its occurrence. Candidate third events include exposure to a *C. difficile* strain of particular virulence, exposure to antimicrobial agents especially likely to cause CDI, and an inadequate host immune response. The host anamnestic serum IgG antibody response to toxin A of *C. difficile* is the most likely third event that determines which patients develop diarrhea and which patients remain asymptomatic. The majority of humans probably first develop antibody to *C. difficile* toxins when colonized asymptomatically during the first year of life or after CDI in childhood. Infants are thought not to develop symptomatic CDI because they lack suitable mucosal toxin receptors that develop later in life. In adulthood, serum levels of IgG antibody to toxin A increase more in response to infection in individuals who become asymptomatic carriers than in those who develop CDI. For persons who develop CDI, increasing levels of antitoxin A during treatment correlate with a lower risk of

recurrence. Two large clinical trials in which intravenous monoclonal antibodies to toxin A and toxin B were used together and as single agents in addition to standard antibiotic therapy showed that rates of recurrent CDI were significantly lower with the combination of antibodies and with the toxin B antibody alone than with placebo plus standard therapy. Antibody to toxin A alone was ineffective.

Figure 129-2

Pathogenesis model for hospital-acquired *Clostridium difficile* infection (CDI). At least three events are integral to *C. difficile* pathogenesis. Exposure to antibiotics establishes susceptibility to infection. Once susceptible, the patient may acquire nontoxigenic (nonpathogenic) or toxigenic strains of *C. difficile* as a second event. Acquisition of toxigenic *C. difficile* may be followed by asymptomatic colonization or CDI, depending on one or more additional events, including an inadequate host anamnestic IgG response to *C. difficile* toxin A.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GLOBAL CONSIDERATIONS

Rates and severity of CDI in the United States, Canada, and Europe have increased markedly since the year 2000. Rates in U.S. hospitals tripled between 2000 and 2005. Hospitals in Montreal, Quebec, reported rates in 2005 that were four times higher than the 1997 baseline, with directly attributable mortality of 6.9% (increased from 1.5%). An epidemic strain, variously known as toxinotype III, REA type BI, PCR ribotype 027, and pulsed-field type NAP1 (collectively designated NAP1/BI/027), is thought to account for much of the increase in incidence and has been found in North America, Europe, and Asia. It is now recognized that two

clones of NAP1/BI/027 originated in the United States and Canada and spread to the United Kingdom, Europe, and Asia. The epidemic organism is characterized by (1) an ability to produce 16–23 times as much toxin A and toxin B as control strains in vitro, (2) the presence of binary toxin CDT, and (3) high-level resistance to all fluoroquinolones. New strains have been and will probably continue to be implicated in outbreaks, including a strain commonly found in food animals that also carries binary toxin and has been associated with high mortality rates in human infections (toxinotype V, ribotype 078). In the last 5 years, the rates of CDI in the United Kingdom have markedly decreased, and the frequency of the NAP1/BI/027 strain has decreased in the European Union. The rate of CDI caused by NAP1/BI/027 is similarly decreasing in the United States, where recent epidemiologic data from the Centers for Disease Control and Prevention (CDC) indicate that this strain has been replaced as the most frequently isolated community-associated strain by ribotype 106/REA group DH, which was previously found to be epidemic in the United Kingdom.

CLINICAL MANIFESTATIONS

Diarrhea is the most common manifestation caused by *C. difficile*. Stools are almost never grossly bloody and range from soft and unformed to watery or mucoid in consistency, with a characteristic odor. Clinical and laboratory findings include fever in 28% of cases, abdominal pain in 22%, and leukocytosis in 50%. When adynamic ileus (which is seen on x-ray in ~20% of cases) results in cessation of stool passage, the diagnosis of CDI is frequently overlooked. A clue to the presence of unsuspected CDI in these patients is unexplained leukocytosis, with ≥15,000 white blood cells (WBCs)/μL. Such patients are at high risk for complications of *C. difficile* infection, particularly toxic megacolon and sepsis.

C. difficile diarrhea recurs after treatment in ~15–30% of cases; this figure may have increased as a result of NAP1/BI/027. Recurrences may represent either relapses due to the same strain or reinfections with a new strain. Susceptibility to recurrence of clinical CDI is likely a result of continued disruption of the normal fecal microbiota caused by the antibiotic used to treat CDI.

DIAGNOSIS

The diagnosis of CDI is based on a combination of clinical criteria: (1) diarrhea (≥3 unformed stools per 24 h for ≥2 days) with no other recognized cause plus (2) detection of toxin A or B in the stool, detection of toxin-producing *C. difficile* in the stool by nucleic acid amplification testing (NAAT; e.g., polymerase chain reaction [PCR]) or by culture, or visualization of pseudomembranes in the colon. PMC is a more advanced form of CDI and is visualized at endoscopy in only ~50% of patients with diarrhea who have a positive stool culture and toxin assay for *C. difficile* (**Table 129-1**). Endoscopy is a rapid diagnostic tool in seriously ill patients with suspected PMC and an acute abdomen, but a negative result in this examination does not rule out CDI.

TABLE 129-1

Relative Sensitivity and Specificity of Diagnostic Tests for *Clostridium difficile* Infection (CDI)

Type of Test	Relative Sensitivity[a]	Relative Specificity[a]	Comment
Stool culture for *C. difficile*	++++	+++	Most sensitive test; specificity of ++++ if the *C. difficile* isolate tests positive for toxin; turnaround time too slow for practical use
Cell culture cytotoxin test on stool	+++	++++	With clinical data, is diagnostic of CDI; highly specific but not as sensitive as stool culture; slow turnaround time
Enzyme immunoassay for toxins A and B in stool	++ to +++	+++	With clinical data, is diagnostic of CDI; rapid results, but not as sensitive as stool culture or cell culture cytotoxin test
Enzyme immunoassay for *C. difficile* common antigen in stool	+++ to ++++	+++	Detects glutamate dehydrogenase found in toxigenic and nontoxigenic strains of *C. difficile* and other stool organisms; more sensitive and less specific than enzyme immunoassay for toxins; requires confirmation with a toxin test; rapid results
Nucleic acid amplification tests for *C. difficile* toxin A or B gene in stool	++++	+++	Detects toxigenic *C. difficile* in stool; widely used in U.S. for clinical testing; more sensitive than enzyme immunoassay toxin testing; marked increase in CDI diagnoses when implemented
Colonoscopy or sigmoidoscopy	+	++++	Highly specific if pseudomembranes are seen; insensitive compared with other tests

[a]According to both clinical and test-based criteria.

Note: ++++, >90%; +++, 71–90%; ++, 51–70%; +, ~50%.

Despite the array of tests available for *C. difficile* and its toxins (Table 129-1), no single test has high sensitivity, high specificity, and rapid turnaround. Most laboratory tests for toxins, including enzyme immunoassays (EIAs), lack sensitivity. However, testing of multiple additional stool specimens is not

recommended. NAATs (including PCR) are widely used diagnostically and are both rapid and sensitive; however, concern has been raised that PCR may detect colonization with toxigenic *C. difficile* in patients who have diarrhea for a reason other than CDI. Confirmation of the presence of toxin in the stool in addition to PCR or glutamate dehydrogenase (GDH) positivity is recommended in the European CDI guidelines for diagnosis of CDI. Empirical treatment is appropriate if CDI is strongly suspected on clinical grounds and stool testing is delayed. Testing of asymptomatic patients is not recommended except for epidemiologic study purposes. In particular, so-called tests of cure following treatment are not recommended because more than 50% of patients continue to harbor the organism and its toxin after diarrhea has ceased and test results do not always predict the recurrence of CDI. The results of such tests should not be used to restrict placement of patients in long-term care or nursing home facilities.

TREATMENT

TREATMENT

Clostridium difficile Infection

PRIMARY CDI

When possible, discontinuation of any ongoing antimicrobial administration is recommended as the first step in treatment of CDI. Earlier studies indicated that 15–23% of patients respond to this simple measure. However, with the advent of the NAP1/BI/027 epidemic strain and the associated rapid clinical deterioration of some patients, prompt initiation of specific CDI treatment has become the standard. General treatment guidelines include hydration and the avoidance of antiperistaltic agents and opiates, which may mask symptoms and possibly worsen disease. Nevertheless, antiperistaltic agents have been used safely with vancomycin or metronidazole treatment for mild to moderate CDI.

Oral administration of vancomycin, fidaxomicin, or metronidazole has been recommended for CDI treatment. IV vancomycin is ineffective for CDI. Fidaxomicin is available only for oral administration. Two large clinical trials comparing vancomycin and fidaxomicin indicated comparable clinical resolution of diarrhea in ~90% of patients, and the rate of recurrent CDI was significantly lower with fidaxomicin. When IV metronidazole is administered, fecal bactericidal drug concentrations are achieved during acute diarrhea; however, in the presence of adynamic ileus, IV metronidazole treatment of CDI has failed. In previous randomized trials, diarrhea response rates to oral therapy with vancomycin or metronidazole were ≥94%, but four observational studies found that response rates for metronidazole had declined to 62–78%. In addition to observational reports of increases in metronidazole failures, a prospective, randomized, double-blind, placebo-controlled study demonstrated the superiority of vancomycin over metronidazole for treatment of severe CDI. Furthermore, the largest randomized controlled trial of vancomycin vs metronidazole showed that the vancomycin cure rate was superior to the metronidazole cure rate (81% vs 73%; $p = 0.034$) for all patients with CDI, regardless of severity. Although the mean time to resolution of diarrhea is 2–4 days, the response to metronidazole may be much slower. Treatment should not be deemed a failure until a drug has been given for at least 6 days. On the basis of data for shorter courses of vancomycin and the results of four

large clinical trials, it is recommended that vancomycin, fidaxomicin, or metronidazole be given for at least 10 days. Metronidazole is not approved for CDI by the U.S. Food and Drug Administration (FDA), and, despite its low cost, its use for CDI treatment is likely to decline once the results of recent randomized trials are incorporated into CDI treatment guidelines. It is important to initiate treatment with oral vancomycin for patients who appear seriously ill, particularly if they have a high WBC count (>15,000/μL) or a creatinine level that is ≥1.5 times higher than the premorbid value **(Table 129-2)**. Small randomized trials of nitazoxanide, bacitracin, rifaximin, and fusidic acid for treatment of CDI have been conducted. These drugs have not been extensively studied, shown to be superior, or approved by the FDA for CDI, but they provide potential alternatives to vancomycin and fidaxomicin.

RECURRENT CDI

Overall, ~15–30% of successfully treated patients experience recurrences of CDI, either as relapses caused by the original organism or as reinfections following treatment. CDI recurrence is significantly lower in patients treated with fidaxomicin than in those treated with vancomycin. Vancomycin and metronidazole have comparable recurrence rates. Recurrence rates are higher among patients ≥65 years old, those who continue to take antibiotics while being treated for CDI, and those who remain in the hospital after the initial episode of CDI. Patients who have a first recurrence of CDI have a high rate of second recurrence (38%). Fidaxomicin is superior to vancomycin in reducing further recurrences in patients who have had one CDI recurrence (Table 129-2). Recurrent disease, once thought to be relatively mild, has now been documented to pose a significant (11%) risk of serious complications (shock, megacolon, perforation, colectomy, or death within 30 days). There is no standard treatment for multiple recurrences, but long or repeated metronidazole courses should be avoided because of potential neurotoxicity. The use of vancomycin in tapering doses or with pulsed dosing every other day for 2–8 weeks may be the most practical approach to treating patients with multiple recurrences. Other experimental approaches include (1) administration of vancomycin followed by fecal microbiota transplantation (FMT) via nasoduodenal tube, colonoscope, or enema and (2) intentional colonization of the patient with a nontoxigenic strain of *C. difficile*. There is currently much interest in the use of FMT in patients with multiple recurrences of CDI, for which it appears to be effective. However, neither of these biotherapeutic approaches, including FMT, has been approved by the FDA for use in the United States. The results of randomized controlled trials of FMT continue to be reported, and, as would be expected, the results are not as impressive as in observational trials. Other FDA-unapproved antibiotic strategies include (1) sequential treatment with vancomycin (125 mg four times daily for 10–14 days) followed by rifaximin (400 mg twice daily for 14 days), (2) treatment with nitazoxanide (500 mg twice daily for 10 days), and (3) vancomycin (125 mg 4 times daily for 10 days) followed by fidaxomicin (200 mg daily for 7 days, then every other day for 13 doses).

SEVERE COMPLICATED OR FULMINANT CDI

Fulminant (rapidly progressive and severe) CDI presents the most difficult treatment challenge. Patients with fulminant disease often do not have diarrhea, and their illness mimics an acute surgical abdomen. Sepsis (hypotension, fever, tachycardia, leukocytosis) may result from fulminant CDI. An acute abdomen (with or without toxic megacolon) may include signs of obstruction, ileus, colon-wall thickening and ascites on abdominal CT, and peripheral-blood leukocytosis (≥20,000 WBCs/μL). With or without diarrhea, the

differential diagnosis of an acute abdomen, sepsis, or toxic megacolon should include CDI if the patient has received antibiotics in the past 2 months. Cautious sigmoidoscopy or colonoscopy to visualize PMC and an abdominal CT examination are the best diagnostic tests in patients without diarrhea.

Medical management of fulminant CDI is suboptimal because of the difficulty of delivering oral fidaxomicin, metronidazole, or vancomycin to the colon in the presence of ileus (Table 129-2). The combination of vancomycin (given orally or via nasogastric tube and by retention enema) plus IV metronidazole has been used with some success in uncontrolled studies, as has IV tigecycline in small-scale uncontrolled studies. Surgical colectomy may be life-saving if there is no response to medical management. If possible, colectomy should be performed before the serum lactate level reaches 5 mmol/L. The incidence of fulminant CDI requiring colectomy appears to be increasing in the evolving epidemic. However, mortality and morbidity associated with colectomy may be reduced by performing instead a laparoscopic ileostomy followed by colon lavage with polyethylene glycol and vancomycin infusion into the colon via the ileostomy.

TABLE 129-2

Recommendations for the Treatment of *Clostridium difficile* Infection (CDI)

Clinical Setting	Treatment(s)	Comments
Initial episode, mild to moderate	Oral vancomycin (125 mg qid × 10 d) ***or*** Fidaxomicin (200 mg bid × 10 d) ***or*** Oral metronidazole (500 mg tid × 10–14 d)	Oral metronidazole is less effective than the other options and may necessitate a longer treatment course for response. Metronidazole is recommended only if vancomycin or fidaxomicin is not readily accessible.
Initial episode, severe	Oral vancomycin (125 mg qid × 10 d) ***or*** Fidaxomicin (200 mg bid × 10 d)	Indicators of severe disease may include leukocytosis (≥15,000 white blood cells/μL) and a creatinine level ≥1.5 times the premorbid value.

Clinical Setting	Treatment(s)	Comments
Initial episode, fulminant	Vancomycin (500 mg PO or via nasogastric tube) plus metronidazole (500 mg IV q8h) ***plus consider*** Rectal instillation of vancomycin (500 mg in 100 mL of normal saline as a retention enema q6–8h)	Fulminant CDI is defined as severe CDI with the addition of hypotension, shock, ileus, or toxic megacolon. The duration of treatment may need to be >2 weeks and is dictated by response.
First recurrence	Oral vancomycin (125 mg qid × 10 d) ***or*** Oral vancomycin followed by a taper-and-pulse regimen,[a] ***or*** Fidaxomicin (200 mg bid × 10 d)	Treatment for the initial episode should be considered when choosing treatment for the first recurrence.

Clinical Setting	Treatment(s)	Comments
Multiple recurrences	Oral vancomycin treatment followed by a taper-and-pulse regimen ***or*** Vancomycin (125 mg qid × 10 d), then stop vancomycin and start rifaximin (400 mg bid × 2 weeks); ***or*** Fidaxomicin (200 mg bid × 10 d) ***or*** Fecal microbiota transplantation (FMT)	FMT has been compared to a treatment course of vancomycin and vancomycin followed by a taper-and-pulse vancomycin regimen. The true efficacy of a single FMT may be only 50–60%. It is recommended that FMT given by enema be considered only after appropriate antibiotic treatment for ≥2 recurrent CDI episodes. Other options for multiple CDI recurrences are lacking good comparative data but include: Nitazoxanide (500 mg bid × 10 d) ***or*** Vancomycin (125 mg qid × 10 d) followed by fidaxomicin (200 mg daily × 7 doses, then every other day × 13 doses)

[a]A typical taper-and-pulse vancomycin regimen following a 10-day treatment course includes: 125 mg bid × 1 week, then daily × 1 week, then q2–3d for 2–8 weeks.

PROGNOSIS

The mortality rate attributed to CDI, previously found to be 0.6–3.5%, has reached 6.9% in recent outbreaks and is progressively higher with increasing age. Most patients recover, but recurrences are common.

PREVENTION AND CONTROL

Strategies for the prevention of CDI are of two types: those aimed at preventing transmission of the organism to the patient and those aimed at reducing the risk of CDI if the organism is transmitted. Transmission of *C. difficile* in clinical practice has been prevented by gloving of personnel, elimination of the use of

contaminated electronic thermometers, and use of hypochlorite (bleach) solution for environmental decontamination of patients' rooms. Hand hygiene is critical; hand washing is recommended in CDI outbreaks because alcohol hand gels are not sporicidal. CDI outbreaks have been best controlled by restricting the use of specific antibiotics, such as clindamycin, second- and third-generation cephalosporins, and fluoroquinolones. Outbreaks of CDI due to clindamycin-resistant strains have resolved promptly when clindamycin use is restricted. Future prevention strategies include use of monoclonal antibodies, vaccines, and biotherapeutics with live organisms that restore protection from colonization.

FURTHER READING

HE M et al: Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. Nat Genet 45:109, 2013. [PubMed: 23222960]

HOTA SS et al: Oral vancomycin followed by fecal transplantation versus tapering oral vancomycin treatment for recurrent *Clostridium difficile* infection: An open-label, randomized controlled trial. Clin Infect Dis 64:265, 2016. [PubMed: 28011612]

JOHNSON S et al: Vancomycin, metronidazole, or tolevamer for *Clostridium difficile* infection: Results from two multinational, randomized, controlled trials. Clin Infect Dis 59:345, 2014. [PubMed: 24799326]

LESSA FC et al: Burden of *Clostridium difficile* infection in the United States. N Engl J Med 372:825, 2015. [PubMed: 25714160]

LOUIE TJ et al: Fidaxomicin versus vancomycin for *Clostridium difficile* infection. N Engl J Med. 364:422, 2011. [PubMed: 21288078]

MAGILL SS et al: Multistate point–prevalence survey of healthcare-associated infections. N Engl J Med 370:1198, 2014. [PubMed: 24670166]

MCDONALD LC et al: Clinical practice guidelines for *Clostridium difficile* infection in adults and children: 2017 update by the Infectious Diseases Society of America (IDSA) and the Society for Healthcare Epidemiology of America (SHEA). Clin Infect Dis 2018 (in press).

POLAGE CR et al: Overdiagnosis of *Clostridium difficile* infection in the molecular test era. JAMA Intern Med 175:1792, 2015. [PubMed: 26348734]

SEE I et al: NAP1 strain type predicts outcomes from *Clostridium difficile* infection. Clin Infect Dis 58:1394, 2014. [PubMed: 24604900]

WILCOX MH et al: Bezlotoxumab for prevention of *Clostridium difficile* infection recurrence. N Engl J Med 376:305, 2017. [PubMed: 28121498]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 130: Urinary Tract Infections, Pyelonephritis, and Prostatitis

Kalpana Gupta; Barbara W. Trautner

INTRODUCTION

Urinary tract infection (UTI) is a common and painful human illness that, fortunately, is rapidly responsive to modern antibiotic therapy. In the preantibiotic era, UTI caused significant morbidity. Hippocrates, writing about a disease that appears to have been acute cystitis, said that the illness could last for a year before either resolving or worsening to involve the kidneys. When chemotherapeutic agents used to treat UTI were introduced in the early twentieth century, they were relatively ineffective, and persistence of infection after 3 weeks of therapy was common. Nitrofurantoin, which became available in the 1950s, was the first tolerable and effective agent for the treatment of UTI.

Since the most common manifestation of UTI is acute cystitis and since acute cystitis is far more prevalent among women than among men, most clinical research on UTI has involved women. Many studies have enrolled women from college campuses or large health maintenance organizations in the United States. Therefore, when reviewing the literature and recommendations concerning UTI, clinicians must consider whether the findings are applicable to their patient populations.

DEFINITIONS

UTI may be asymptomatic (subclinical infection) or symptomatic (disease). Thus, the term *urinary tract infection* encompasses a variety of clinical entities, including asymptomatic bacteriuria (ASB), cystitis, prostatitis, and pyelonephritis. The distinction between symptomatic UTI and ASB has major clinical implications. Both UTI and ASB connote the presence of bacteria in the urinary tract, usually accompanied by white blood cells and inflammatory cytokines in the urine. However, ASB occurs in the absence of symptoms attributable to the bacteria in the urinary tract and usually does not require treatment, while UTI has more typically been assumed to imply symptomatic disease that warrants antimicrobial therapy. Much of the literature concerning UTI, particularly catheter-associated infection, does not differentiate between UTI and ASB. In this chapter, the term *urinary tract infection* denotes symptomatic disease; *cystitis*, symptomatic infection of the bladder; and *pyelonephritis*, symptomatic infection of the kidneys. *Uncomplicated urinary tract infection* refers to acute cystitis or pyelonephritis in nonpregnant outpatient women without anatomic abnormalities or instrumentation of the urinary tract; the term *complicated urinary tract infection* encompasses all other types of UTI. *Recurrent urinary tract infection* is not necessarily complicated; individual episodes can be uncomplicated and treated as such. *Catheter-associated bacteriuria* can be either symptomatic (CAUTI) or asymptomatic.

EPIDEMIOLOGY AND RISK FACTORS

Except among infants and the elderly, UTI occurs far more commonly in females than in males. During the neonatal period, the incidence of UTI is slightly higher among males than among females because male infants more commonly have congenital urinary tract anomalies. After 50 years of age, obstruction from prostatic hypertrophy becomes common in men, and the incidence of UTI is almost as high among men as among women. Between 1 year and ~50 years of age, UTI and recurrent UTI are predominantly diseases of females. The prevalence of ASB is ~5% among women between ages 20 and 40 and may be as high as 40–50% among elderly women and men.

As many as 50–80% of women in the general population acquire at least one UTI during their lifetime—uncomplicated cystitis in most cases. Recent use of a diaphragm with spermicide, frequent sexual intercourse, and a history of UTI are independent risk factors for

acute cystitis. Cystitis is temporally related to recent sexual intercourse in a dose–response manner, with an increased relative risk ranging from 1.4 with one episode of intercourse in the preceding week to 4.8 with five episodes. In healthy postmenopausal women, sexual activity, diabetes mellitus, and incontinence are risk factors for UTI.

Many factors predisposing women to cystitis also increase the risk of pyelonephritis. Factors independently associated with pyelonephritis in young healthy women include frequent sexual intercourse, a new sexual partner, a UTI in the previous 12 months, a maternal history of UTI, diabetes, and incontinence. The shared risk factors for cystitis and pyelonephritis are not surprising given that pyelonephritis typically arises through the ascent of bacteria from the bladder to the upper urinary tract. However, pyelonephritis can occur without symptomatic antecedent cystitis.

About 20–30% of women who have had one episode of UTI will have recurrent episodes. Early recurrence (within 2 weeks) is usually regarded as relapse rather than reinfection and may indicate the need to evaluate the patient for a sequestered focus. Intracellular pods of infecting organisms within the bladder epithelium have been demonstrated in animal models of UTI, but the clinical impact of this phenomenon in humans is not yet clear. The rate of recurrence ranges from 0.3 to 7.6 infections per patient per year, with an average of 2.6 infections per year. It is not uncommon for multiple recurrences to follow an initial infection, resulting in clustering of episodes. Clustering may be related temporally to the presence of a new risk factor, to the sloughing of the protective outer bladder epithelial layer in response to bacterial attachment during acute cystitis, or possibly to antibiotic-related alteration of the normal flora. The likelihood of a recurrence decreases with increasing time since the last infection. A case–control study of predominantly white premenopausal women with recurrent UTI identified frequent sexual intercourse, use of spermicide, a new sexual partner, a first UTI before 15 years of age, and a maternal history of UTI as independent risk factors for recurrent UTI. The only consistently documented behavioral risk factors for recurrent UTI include frequent sexual intercourse and spermicide use. In postmenopausal women, major risk factors for recurrent UTI include a history of premenopausal UTI and anatomic factors affecting bladder emptying, such as cystoceles, urinary incontinence, and residual urine.

In pregnant women, ASB has clinical consequences, and both screening for and treatment of this condition are indicated. Specifically, ASB during pregnancy is associated with maternal pyelonephritis, which in turn is associated with preterm delivery. Antibiotic treatment of ASB in pregnant women can reduce the risk of pyelonephritis, preterm delivery, and low-birth-weight babies.

The majority of men with UTI have a functional or anatomic abnormality of the urinary tract, most commonly urinary obstruction secondary to prostatic hypertrophy. That said, not all men with UTI have detectable urinary abnormalities; this point is particularly relevant for men ≤45 years of age. Lack of circumcision is associated with an increased risk of UTI because *Escherichia coli* is more likely to colonize the glans and prepuce and subsequently migrate into the urinary tract of uncircumcised men.

Women with diabetes have a two- to threefold higher rate of ASB and UTI than women without diabetes; there is insufficient evidence on which to base a corresponding statement about men. Increased duration of diabetes and the use of insulin rather than oral medication are associated with an elevated risk of UTI among women with diabetes. Poor bladder function, obstruction in urinary flow, and incomplete voiding are additional factors commonly found in patients with diabetes that increase the risk of UTI. Impaired cytokine secretion may contribute to ASB in diabetic women. The sodium–glucose co-transporter 2 (SGLT2) inhibitors used for treatment of diabetes result in glycosuria and may be associated with small increases in the risk of UTI.

ETIOLOGY

The uropathogens causing UTI vary by clinical syndrome but are usually enteric gram-negative rods that have migrated to the urinary tract. The susceptibility patterns of these organisms vary by clinical syndrome and by geography. In acute uncomplicated cystitis in the United States, the etiologic agents are highly predictable: *E. coli* accounts for 75–90% of isolates; *Staphylococcus saprophyticus* for 5–15% (with particularly frequent isolation from younger women); and *Klebsiella, Proteus, Enterococcus*, and *Citrobacter* species, along with other organisms, for 5–10%. Similar etiologic agents are found in Europe and Brazil. The spectrum of agents causing uncomplicated pyelonephritis is similar, with *E. coli* predominating. In complicated UTI (e.g., CAUTI), *E. coli* remains the predominant organism, but other aerobic gram-negative rods, such as *Pseudomonas aeruginosa* and *Klebsiella*, *Proteus*, *Citrobacter*, *Acinetobacter*, and *Morganella* species, also are frequently isolated. Gram-positive bacteria (e.g., enterococci and *Staphylococcus aureus*) and yeasts also are important pathogens in complicated UTI. Data on etiology and resistance are generally

obtained from laboratory surveys and should be understood in the context that organisms are identified only in cases in which urine is sent for culture—typically, when complicated UTI or pyelonephritis is suspected. Genetic sequencing of the bladder microbiome or of all the bacteria that can be identified in the bladder has consistently demonstrated that more bacterial species are present than can be identified by routine culture methods, in both symptomatic and asymptomatic states. The clinical significance of these non-cultivatable organisms is unknown but has challenged the assumption that the bladder is normally a sterile site.

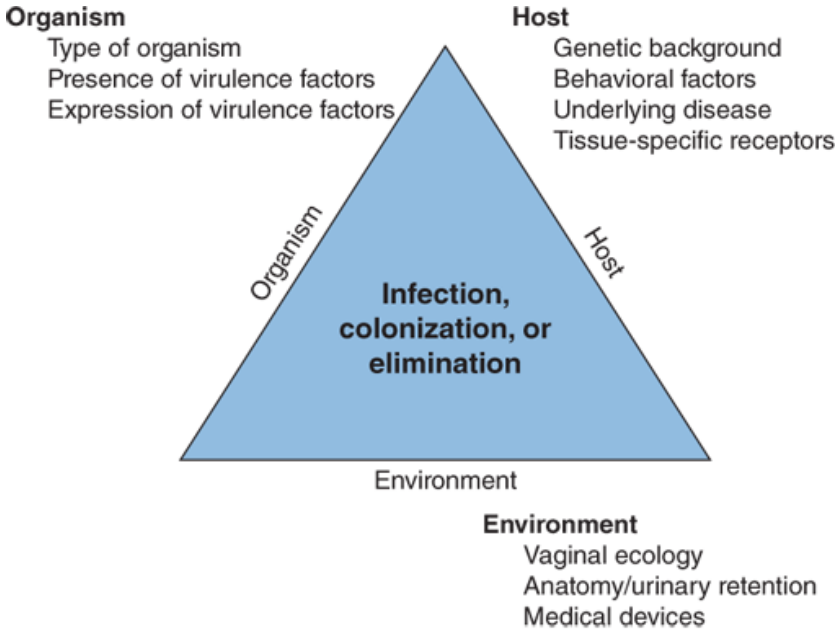
The available data demonstrate a worldwide increase in the resistance of *E. coli* to antibiotics commonly used to treat UTI. North American and European surveys from women with acute cystitis have documented resistance rates of >20% to trimethoprim-sulfamethoxazole (TMP-SMX) in many regions and >10% to ciprofloxacin in some regions. In community-acquired infections, the increased prevalence of multidrug-resistant uropathogens has left few oral options for therapy in some cases. Since resistance rates vary by local geographic region, with individual patient characteristics, and over time, it is important to use current and local data when choosing a treatment regimen.

PATHOGENESIS

The urinary tract can be viewed as an anatomic unit linked by a continuous column of urine extending from the urethra to the kidneys. In the majority of UTIs, bacteria establish infection by ascending from the urethra to the bladder. Continuing ascent up the ureter to the kidney is the pathway for most renal parenchymal infections. However, introduction of bacteria into the bladder does not inevitably lead to sustained and symptomatic infection. The interplay of host, pathogen, and environmental factors determines whether tissue invasion and symptomatic infection will ensue (**Fig. 130-1**). For example, bacteria often enter the bladder after sexual intercourse, but normal voiding and innate host defense mechanisms in the bladder eliminate these organisms. Any foreign body in the urinary tract, such as a urinary catheter or stone, provides an inert surface for bacterial colonization. Abnormal micturition and/or significant residual urine volume promotes infection. In the simplest of terms, anything that increases the likelihood of bacteria entering the bladder and staying there increases the risk of UTI.

FIGURE 130-1

Pathogenesis of urinary tract infection. The relationship among specific host, pathogen, and environmental factors determines the clinical outcome.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Bacteria can gain access to the urinary tract through the bloodstream. However, hematogenous spread accounts for <2% of documented UTIs and usually results from bacteremia caused by relatively virulent organisms, such as *Salmonella* and *S. aureus*. Indeed, the isolation of either of these pathogens from a patient without a catheter or other instrumentation warrants a search for a bloodstream source. Hematogenous infections may produce focal abscesses or areas of pyelonephritis within a kidney and result in

positive urine cultures. The pathogenesis of candiduria is distinct in that the hematogenous route is common. The presence of *Candida* in the urine of a non-instrumented immunocompetent patient implies either genital contamination or potentially widespread visceral dissemination.

Environmental Factors

Vaginal Ecology

Vaginal ecology is an important environmental factor affecting the risk of UTI in women. Colonization of the vaginal introitus and periurethral area with organisms from the intestinal flora (usually *E. coli*) is the critical initial step in the pathogenesis of UTI. Sexual intercourse is associated with an increased risk of vaginal colonization with *E. coli* and thereby increases the risk of UTI. Nonoxynol-9 in spermicide is toxic to the normal vaginal lactobacilli and thus is likewise associated with an increased risk of *E. coli* vaginal colonization and bacteriuria. In postmenopausal women, the previously predominant vaginal lactobacilli are replaced with colonizing gram-negative bacteria. The use of topical estrogens to prevent UTI in postmenopausal women is controversial; given the side effects of systemic hormone replacement, oral estrogens should not be used to prevent UTI.

Anatomic and Functional Abnormalities

Any condition that permits urinary stasis or obstruction predisposes the individual to UTI. Foreign bodies such as stones or urinary catheters provide an inert surface for bacterial colonization and formation of a persistent biofilm. Thus, vesicoureteral reflux, ureteral obstruction secondary to prostatic hypertrophy, neurogenic bladder, and urinary diversion surgery create an environment favorable to UTI. In persons with such conditions, *E. coli* strains lacking typical urinary virulence factors are often the cause of infection. Inhibition of ureteral peristalsis and decreased ureteral tone leading to vesicoureteral reflux are important in the pathogenesis of pyelonephritis in pregnant women. Anatomic factors—specifically, the distance of the urethra from the anus—are considered to be the primary reason why UTI is predominantly an illness of young women rather than of young men.

Host Factors

 The genetic background of the host influences the individual's susceptibility to recurrent UTI, at least among women. A familial disposition to UTI and to pyelonephritis is well documented. Women with recurrent UTI are more likely to have had their first UTI before the age of 15 years and to have a maternal history of UTI. A component of the underlying pathogenesis of this familial predisposition to recurrent UTI may be persistent vaginal colonization with *E. coli*, even during asymptomatic periods. Vaginal and periurethral mucosal cells from women with recurrent UTI bind threefold more uropathogenic bacteria than do mucosal cells from women without recurrent infection. Epithelial cells from women who are non-secretors of certain blood group antigens may possess specific types of receptors to which *E. coli* can bind, thereby facilitating colonization and invasion. Mutations in host innate immune response genes (e.g., those coding for Toll-like receptors and the interleukin 8 receptor) also have been linked to recurrent UTI and pyelonephritis. The genetic patterns that predispose to cystitis and pyelonephritis appear to be distinct.

Microbial Factors

 An anatomically normal urinary tract presents a stronger barrier to infection than a compromised urinary tract. Thus, strains of *E. coli* that cause invasive symptomatic infection of the urinary tract in otherwise normal hosts often possess and express genetic virulence factors, including surface adhesins that mediate binding to specific receptors on the surface of uroepithelial cells. The best-studied adhesins are the P fimbriae, hair-like protein structures that interact with a specific receptor on renal epithelial cells. (The letter *P* denotes the ability of these fimbriae to bind to blood group antigen P, which contains a D-galactose-D-galactose residue.) P fimbriae are important in the pathogenesis of pyelonephritis and subsequent bloodstream invasion from the kidney.

Another adhesin is the type 1 pilus (fimbria), which all *E. coli* strains possess but not all *E. coli* strains express. Type 1 pili are thought to play a key role in initiating *E. coli* bladder infection; they mediate binding to mannose on the luminal surface of bladder uroepithelial

cells. Toxins, metal (iron) acquisition systems, biofilm formation, and capsules can also contribute to the ability of pathogenic *E. coli* to thrive in the bladder.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Clinical Syndromes

The most important issue to be addressed when a UTI is suspected is the characterization of the clinical syndrome as ASB, uncomplicated cystitis, pyelonephritis, prostatitis, or complicated UTI. This information will shape the diagnostic and therapeutic approach.

ASYMPTOMATIC BACTERIURIA

A diagnosis of ASB can be considered only when the patient does not have local or systemic symptoms referable to the urinary tract. The clinical presentation is usually bacteriuria detected incidentally when a patient undergoes a screening urine culture for a reason unrelated to the genitourinary tract. Systemic signs or symptoms such as fever, altered mental status, and leukocytosis in the setting of a positive urine culture are nonspecific and do not merit a diagnosis of symptomatic UTI unless other potential etiologies have been considered.

CYSTITIS

The typical symptoms of cystitis are dysuria, urinary frequency, and urgency. Nocturia, hesitancy, suprapubic discomfort, and gross hematuria are often noted as well. Unilateral back or flank pain is generally an indication that the upper urinary tract is involved. Fever also is an indication of invasive infection of either the kidney or the prostate.

PYELONEPHRITIS

Mild pyelonephritis can present as low-grade fever with or without lower-back or costovertebral-angle pain, whereas severe pyelonephritis can manifest as high fever, rigors, nausea, vomiting, and flank and/or loin pain. Symptoms are generally acute in onset, and symptoms of cystitis may not be present. Fever is the main feature distinguishing cystitis from pyelonephritis. The fever of pyelonephritis typically exhibits a high spiking “picket-fence” pattern and resolves over 72 h of therapy. Bacteremia develops in 20–30% of cases of pyelonephritis. Patients with diabetes may present with obstructive uropathy associated with acute papillary necrosis when the sloughed papillae obstruct the ureter. Papillary necrosis may also be evident in some cases of pyelonephritis complicated by obstruction, sickle cell disease, analgesic nephropathy, or combinations of these conditions. In the rare cases of bilateral papillary necrosis, a rapid rise in the serum creatinine level may be the first indication of the condition. *Emphysematous* pyelonephritis is a particularly severe form of the disease that is associated with the production of gas in renal and perinephric tissues and occurs almost exclusively in diabetic patients **(Fig. 130-2)**. *Xanthogranulomatous* pyelonephritis occurs when chronic urinary obstruction (often by staghorn calculi), together with chronic infection, leads to suppurative destruction of renal tissue **(Fig. 130-3)**. On pathologic examination, the residual renal tissue frequently has a yellow coloration, with infiltration by lipid-laden macrophages. Pyelonephritis can also be complicated by intraparenchymal abscess formation; this development should be suspected when a patient has continued fever and/or bacteremia despite antibacterial therapy.

PROSTATITIS

Prostatitis includes both infectious and noninfectious abnormalities of the prostate gland. Infections can be acute or chronic, are almost always bacterial in nature, and are far less common than the noninfectious entity *chronic pelvic pain syndrome* (formerly known as chronic prostatitis). Acute bacterial prostatitis presents as dysuria, frequency, and pain in the prostatic pelvic or perineal area. Fever and chills are usually present, and symptoms of bladder outlet obstruction are common. Chronic bacterial prostatitis presents more insidiously as recurrent episodes of cystitis, sometimes with associated pelvic and perineal pain. Men who present with recurrent cystitis should be evaluated for a prostatic focus as well as urinary retention.

COMPLICATED UTI

Complicated UTI presents as a symptomatic episode of cystitis or pyelonephritis in a man or woman with an anatomic predisposition to infection, with a foreign body in the urinary tract, or with factors predisposing to a delayed response to therapy.

FIGURE 130-2

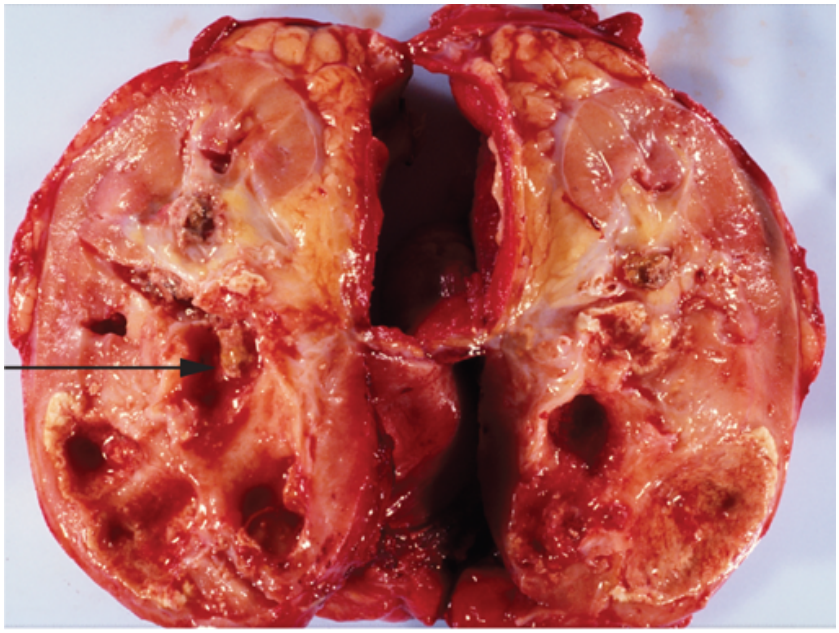
Emphysematous pyelonephritis. Infection of the right kidney of a diabetic man by *Escherichia coli*, a gas-forming, facultative anaerobic uropathogen, has led to destruction of the renal parenchyma (*arrow*) and tracking of gas through the retroperitoneal space (*arrowhead*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

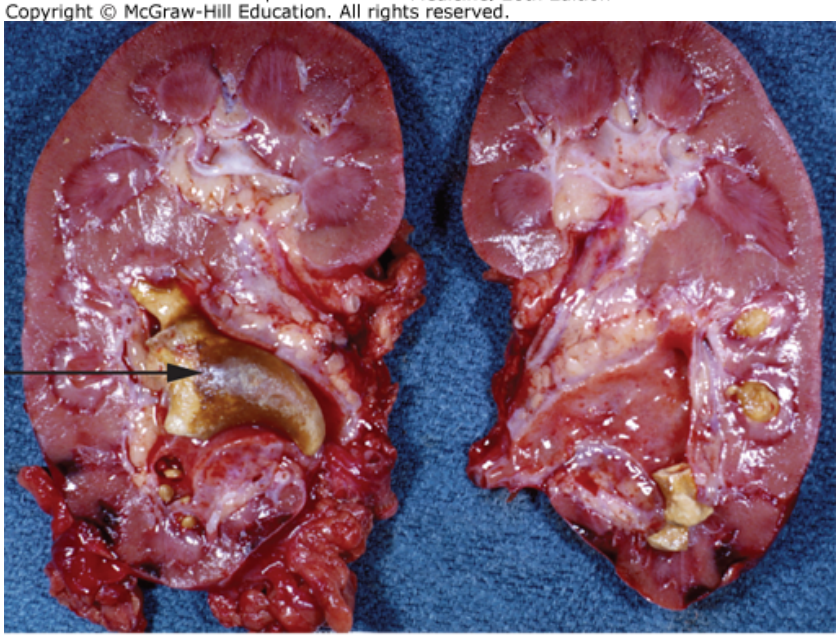
FIGURE 130-3

Xanthogranulomatous pyelonephritis. *A.* This photograph shows extensive destruction of renal parenchyma due to long-standing suppurative inflammation. The precipitating factor was obstruction by a staghorn calculus, which has been removed, leaving a depression (*arrow*). The mass effect of xanthogranulomatous pyelonephritis can mimic renal malignancy. ***B.*** A large staghorn calculus (*arrow*) is seen obstructing the renal pelvis and calyceal system. The lower pole of the kidney shows areas of hemorrhage and necrosis with collapse of cortical areas. *(Images courtesy of Dharam M. Ramnani, MD, Virginia Urology Pathology Laboratory, Richmond, VA.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSTIC TOOLS

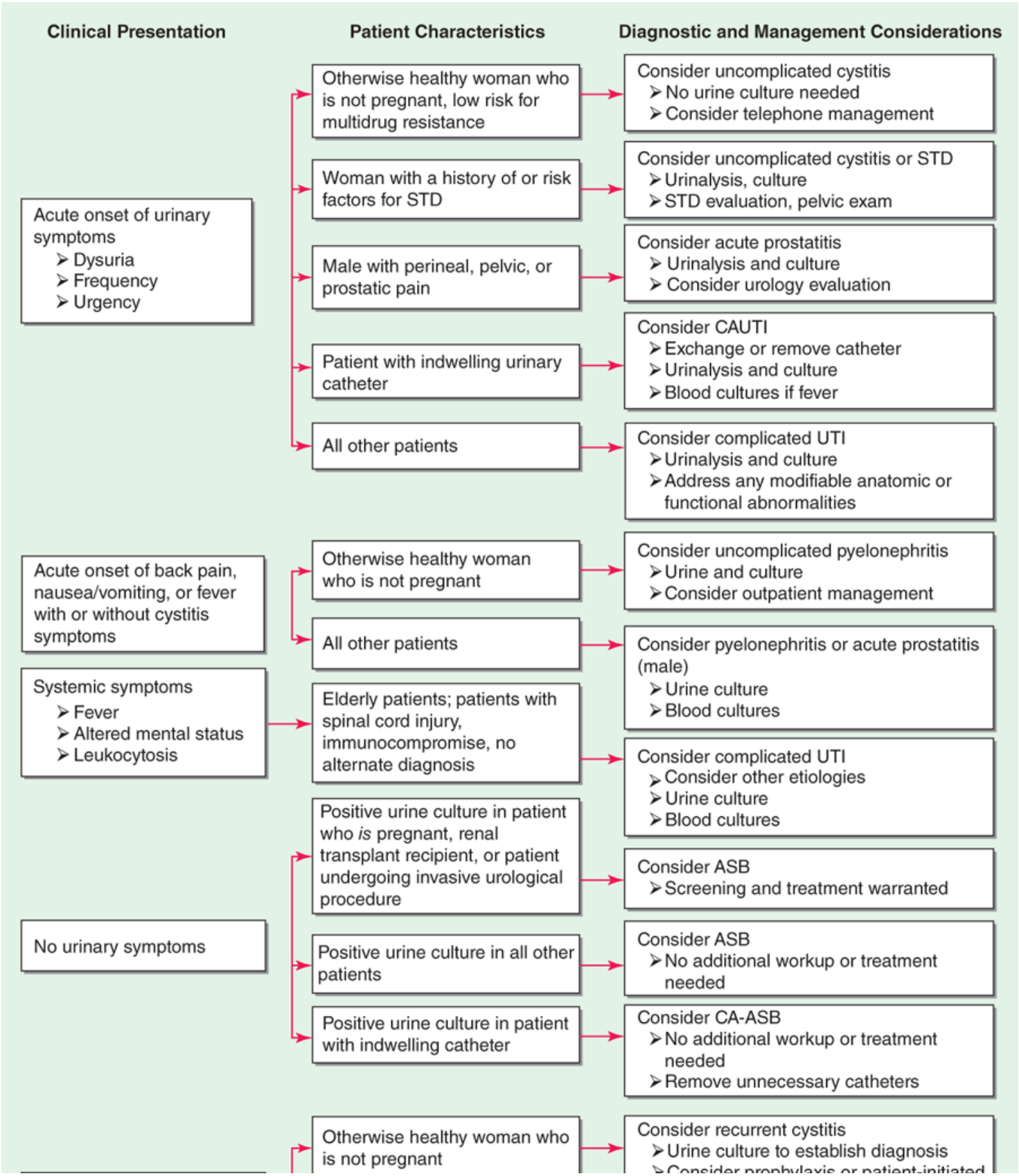
History

The diagnosis of any of the UTI syndromes or ASB begins with a detailed history (**Fig. 130-4**). The history given by the patient has a high predictive value in uncomplicated cystitis. A meta-analysis evaluating the probability of acute UTI on the basis of history and physical findings concluded that, in women presenting with at least one symptom of UTI (dysuria, frequency, hematuria, or back pain) and without complicating factors, the probability of acute cystitis or pyelonephritis is 50%. The even higher rates of accuracy of self-diagnosis among women with recurrent UTI probably account for the success of patient-initiated treatment of recurrent cystitis. If vaginal discharge and complicating factors are absent and risk factors for UTI are present, then the probability of UTI is close to 90%,

and no laboratory evaluation is needed. A combination of dysuria and urinary frequency in the absence of vaginal discharge increases the probability of UTI to 96%. Further laboratory evaluation with dipstick testing or urine culture is not necessary in such patients before the initiation of definitive therapy.

FIGURE 130-4

Diagnostic approach to urinary tract infection (UTI). ASB, asymptomatic bacteriuria; CA-ASB, catheter-associated ASB; CAUTI, catheter-associated UTI; STD, sexually transmitted disease.





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In applying the patient's history as a diagnostic tool, the physician must remember that the studies included in the meta-analysis cited above did not enroll children, adolescents, pregnant women, men, or patients with complicated UTI. One significant concern is that sexually transmitted disease—that caused by *Chlamydia trachomatis* in particular—may be inappropriately treated as UTI. This concern is particularly relevant for female patients under the age of 25. The differential diagnosis to be considered when women present with dysuria includes cervicitis (*C. trachomatis, Neisseria gonorrhoeae*), vaginitis (*Candida albicans, Trichomonas vaginalis*), herpetic urethritis, interstitial cystitis, and noninfectious vaginal or vulvar irritation. Women with more than one sexual partner and inconsistent use of condoms are at high risk for both UTI and sexually transmitted disease, and symptoms alone do not always distinguish between these conditions.

Urine Dipstick Test, Urinalysis, and Urine Culture

Useful diagnostic tools include the urine dipstick test and urinalysis, both of which provide point-of-care information, and the urine culture, which can retrospectively confirm a prior diagnosis. Understanding the parameters of the dipstick test is important in interpreting its results. Only members of the family Enterobacteriaceae convert nitrate to nitrite, and enough nitrite must accumulate in the urine to reach the threshold of detection. If a woman with acute cystitis is forcing fluids and voiding frequently, the dipstick test for nitrite is less likely to be positive, even when *E. coli* is present. The leukocyte esterase test detects this enzyme in polymorphonuclear leukocytes in the host's urine, whether the cells are intact or lysed. Many reviews have attempted to describe the diagnostic accuracy of dipstick testing. The bottom line for clinicians is that a urine dipstick test can confirm the diagnosis of uncomplicated cystitis in a patient with a reasonably high pretest probability of this disease; either nitrite or leukocyte esterase positivity can be interpreted as a positive result. Blood in the urine also may suggest a diagnosis of UTI. A dipstick test negative for both nitrite and leukocyte esterase in this type of patient should prompt consideration of other explanations for the patient's symptoms and collection of urine for culture. A negative dipstick test is not sufficiently sensitive to rule out bacteriuria in pregnant women, in whom it is important to detect all episodes of bacteriuria.

Urine microscopy reveals pyuria in nearly all cases of cystitis and hematuria in ~30% of cases. In current practice, most hospital laboratories use an automated system rather than manual examination for urine microscopy. A machine aspirates a sample of the urine and then classifies the particles in the urine by size, shape, contrast, light scatter, volume, and other properties. These automated systems can be overwhelmed by high numbers of dysmorphic red blood cells, white blood cells, or crystals; in general, counts of bacteria are less accurate than are counts of red and white blood cells. The authors' clinical recommendation is that the patient's symptoms and presentation should outweigh an incongruent result on automated urinalysis.

The detection of bacteria in a urine culture is the diagnostic gold standard for UTI; unfortunately, however, culture results do not become available until 24 h after the patient's presentation. Identifying specific organism(s) can require an additional 24 h. Studies of women with symptoms of cystitis have found that a colony count threshold of $\geq 10^2$ bacteria/mL is more sensitive (95%) and specific (85%) than a threshold of 10^5/mL for the diagnosis of acute cystitis in women. In men, the minimal level indicating infection appears to be 10^3/mL. Urine specimens frequently become contaminated with the normal microbial flora of the distal urethra, vagina, or skin. These contaminants can grow to high numbers if the collected urine is allowed to stand at room temperature. In most instances, a culture that yields mixed bacterial species is contaminated except in settings of long-term catheterization, chronic urinary retention, or the presence of a fistula between the urinary tract and the gastrointestinal or genital tract.

DIAGNOSTIC APPROACH

The approach to diagnosis is influenced by which of the clinical UTI syndromes is suspected (Fig. 130-4).

Uncomplicated Cystitis in Women

Uncomplicated cystitis in women can be treated on the basis of history alone. However, if the symptoms are not specific or if a reliable history cannot be obtained, then a urine dipstick test should be performed. A positive nitrite or leukocyte esterase result in a woman with one symptom of UTI increases the probability of UTI from 50% to ~80%, and empirical treatment can be considered without further testing. In this setting, a negative dipstick result does not rule out UTI, and a urine culture, close clinical follow-up, and possibly a pelvic examination are recommended. In women with complicated UTI (e.g., due to pregnancy, suspected bacterial resistance, or recent UTI), a urine culture is warranted to guide appropriate therapy.

Cystitis in Men

The signs and symptoms of cystitis in men are similar to those in women, but this disease differs in several important ways in the male population. Collection of urine for culture is strongly recommended when a man has symptoms of UTI, as the documentation of bacteriuria can differentiate the less common syndromes of acute and chronic bacterial prostatitis from the very common entity of chronic pelvic pain syndrome, which is not associated with bacteriuria and thus is not usually responsive to antibacterial therapy. Men with febrile UTI often have an elevated serum level of prostate-specific antigen as well as an enlarged prostate and enlarged seminal vesicles on ultrasound—findings indicative of prostate involvement. In a study of 85 men with febrile UTI, symptoms of urinary retention, early recurrence of UTI, hematuria at follow-up, and voiding difficulties were predictive of surgically correctable disorders. Men with none of these symptoms had normal upper and lower urinary tracts on urologic workup. In general, men with a first febrile UTI should have imaging performed (CT or ultrasound); if the diagnosis is unclear or if UTI is recurrent, referral for urologic consultation and further evaluation—including potential localization cultures using the two- or four-glass Meares-Stamey test (urine collection after prostate massage)—is appropriate.

Asymptomatic Bacteriuria

The diagnosis of ASB involves both microbiologic and clinical criteria. The microbiologic criterion (including in urinary catheter–associated asymptomatic bacteriuria) is $\geq 10^5$ bacterial CFU/mL of urine. The clinical criterion is an absence of signs or symptoms referable to UTI.

TREATMENT

TREATMENT

Urinary Tract Infections

Treatment of UTI accounts for a major proportion of antimicrobial use in ambulatory care, inpatient care, and long-term-care settings. Responsible use of antibiotics for this common infection has broad implications for preserving antibiotic effectiveness into the future. That said, antimicrobial therapy is warranted for any UTI that is truly symptomatic. The choice of antimicrobial agent, the dose, and the duration of therapy depend on the site of infection and the presence or absence of complicating conditions. Each category of UTI warrants a different approach based on the particular clinical syndrome.

Antimicrobial resistance among uropathogens varies from region to region and impacts the approach to empirical treatment of UTI. *E. coli* ST131 is the predominant multilocus sequence type found worldwide as the cause of multidrug-resistant UTI. Recommendations for treatment must be considered in the context of local resistance patterns and national differences in some agents' availability. For example, fosfomycin and pivmecillinam are not available in all countries but are considered first-line options where they are available because they retain activity against a majority of uropathogens that produce extended-spectrum β-lactamases. Thus, therapeutic choices should depend on local resistance, drug availability, and individual patient factors such as recent travel and antimicrobial use.

UNCOMPLICATED CYSTITIS IN WOMEN

Since the species and antimicrobial susceptibilities of the bacteria that cause acute uncomplicated cystitis are highly predictable, many episodes of uncomplicated cystitis can be managed over the telephone (Fig. 130-4). Most patients with other UTI syndromes

require further diagnostic evaluation. Although the risk of serious complications with telephone management appears to be low, studies of telephone management algorithms generally have involved otherwise healthy women who are at low risk of complications of UTI.

In 1999, TMP-SMX was recommended as the first-line agent for treatment of uncomplicated UTI in the published guidelines of the Infectious Diseases Society of America. Since then, antibiotic resistance among uropathogens causing uncomplicated cystitis has increased, appreciation of the importance of collateral damage (as defined below) has increased, and newer agents have been studied. Unfortunately, there is no longer a single best agent for acute uncomplicated cystitis.

Collateral damage refers to the adverse ecologic effects of antimicrobial therapy, including killing of the normal flora and selection of drug-resistant organisms. The implication of collateral damage for UTI management is that a drug that is highly efficacious for the treatment of UTI is not necessarily the optimal first-line agent if it also has pronounced secondary effects on the normal flora or is likely to adversely affect resistance patterns. Drugs used for UTI that have a minimal effect on fecal flora include pivmecillinam, fosfomycin, and nitrofurantoin. In contrast, trimethoprim, TMP-SMX, quinolones, and ampicillin affect the fecal flora more significantly; these drugs are notably the agents for which rising resistance levels have been documented.


Choosing judiciously whether to initiate antibiotic therapy and then selecting the most urinary-focused agent for the shortest appropriate duration are important factors in global efforts to stem the rise of antimicrobial-resistant organisms. Several effective therapeutic regimens are available for acute uncomplicated cystitis in women (**Table 130-1**). Well-studied first-line agents include TMP-SMX and nitrofurantoin. Second-line agents include β-lactams. There is increasing experience with the use of fosfomycin for UTIs (including complicated infections), particularly for infections caused by multidrug-resistant *E. coli*. According to an advisory from the U.S. Food and Drug Administration (FDA), fluoroquinolones should not be used for uncomplicated cystitis unless no alternatives are available. Pivmecillinam is not currently available in the United States or Canada but is a popular agent in some European countries. The pros and cons of specific agents are discussed briefly below.

Traditionally, TMP-SMX has been recommended as first-line treatment for acute cystitis, and it remains appropriate to consider the use of this drug in regions with resistance rates not exceeding 20%. In women with recurrent UTI, prior cultures can be used as a guide to TMP-SMX susceptibility, although interim acquisition of resistant bacteria can occur. TMP-SMX resistance has clinical significance: in TMP-SMX-treated patients with resistant isolates, the time to symptom resolution is longer and rates of both clinical and microbiologic failure are higher. Individual host factors associated with an elevated risk of UTI caused by a strain of *E. coli* resistant to TMP-SMX include recent use of TMP-SMX or another antimicrobial agent and recent travel to an area with high rates of TMP-SMX resistance. The optimal setting for empirical use of TMP-SMX is uncomplicated UTI in a female patient who has an established relationship with the practitioner and who can thus seek further care if her symptoms do not respond promptly.

Resistance to nitrofurantoin remains low despite >60 years of use, as several mutational steps are required for the development of bacterial resistance to this drug. Nitrofurantoin remains highly active against *E. coli* and most non–*E. coli* isolates. *Proteus*, *Pseudomonas*, *Serratia*, *Enterobacter*, and yeasts are all intrinsically resistant to this drug. Although nitrofurantoin has traditionally been prescribed as a 7-day regimen, guidelines now recommend a 5-day course, which is as effective as a 3-day course of TMP-SMX for treatment of acute cystitis; 3-day courses of nitrofurantoin are not recommended for acute cystitis. Nitrofurantoin does not reach significant levels in tissue and cannot be used to treat pyelonephritis.

Most fluoroquinolones are highly effective as short-course therapy for cystitis; the exception is moxifloxacin, which may not reach adequate urinary levels. The fluoroquinolones commonly used for UTI include ciprofloxacin and levofloxacin. The two main concerns about fluoroquinolone use for acute cystitis are the propagation of fluoroquinolone resistance, not only among uropathogens but also among other organisms causing more serious and difficult-to-treat infections at other sites, and their rare but potentially serious adverse effects. For example, quinolone use in certain populations, including adults >60 years of age, has been associated with an increased risk of Achilles tendon rupture. Other potential side effects include irreversible neuropathy. In light of these detrimental effects, the FDA issued an advisory against using fluoroquinolones to treat acute cystitis in patients who have other therapeutic options.

β-Lactam agents generally have not performed as well as TMP-SMX or fluoroquinolones in acute cystitis. Rates of pathogen eradication are lower and relapse rates are higher with β-lactam drugs. The generally accepted explanation is that β-lactams fail to eradicate

uropathogens from the vaginal reservoir. Many strains of *E. coli* that are resistant to TMP-SMX are also resistant to amoxicillin and cephalexin; thus, these drugs should be used only for patients infected with susceptible strains.

Urinary analgesics are appropriate in certain situations to speed resolution of bladder discomfort. The urinary tract analgesic phenazopyridine is widely used but can cause significant nausea. Combination analgesics containing urinary antiseptics (methenamine, methylene blue), a urine-acidifying agent (sodium phosphate), and an antispasmodic agent (hyoscyamine) also are available.

Interest in the responsible use of antibiotics has led to exploration of antibiotic-sparing approaches to the treatment of acute uncomplicated cystitis. Both placebo and analgesics alone have proved inferior to antibiotics for resolution of symptoms and prevention of pyelonephritis. Delayed therapy, in which a woman receives a prescription for antibiotics but fills it only if symptoms fail to resolve in a day or two, has the potential advantage of avoiding antibiotic use in those who either do not have cystitis to begin with or have a mild case that resolves spontaneously. The downside is that women who really do have cystitis endure discomfort for a longer period and may meanwhile progress to pyelonephritis. However, one certain measure for more responsible use of antibiotics in cystitis is to treat for the correct duration; in practice, many episodes of acute cystitis are treated longer than is recommended by evidence-based guidelines.

PYELONEPHRITIS

Since patients with pyelonephritis have tissue-invasive disease, the treatment regimen chosen should have a very high likelihood of eradicating the causative organism and should reach therapeutic blood levels quickly. High rates of TMP-SMX-resistant *E. coli* in patients with pyelonephritis have made fluoroquinolones the first-line therapy for acute uncomplicated pyelonephritis. Whether the fluoroquinolones are given orally or parenterally depends on the patient's tolerance for oral intake. A randomized clinical trial demonstrated that a 7-day course of therapy with oral ciprofloxacin (500 mg twice daily, with or without an initial IV 400-mg dose) was highly effective for the initial management of pyelonephritis in the outpatient setting. Oral TMP-SMX (one double-strength tablet twice daily for 14 days) also is effective for treatment of acute uncomplicated pyelonephritis if the uropathogen is known to be susceptible. If the pathogen's susceptibility is not known and TMP-SMX is used, an initial IV 1-g dose of ceftriaxone is recommended. Oral β-lactam agents are less effective than the fluoroquinolones and should be used with caution and close follow-up. Options for parenteral therapy for uncomplicated pyelonephritis include fluoroquinolones, an extended-spectrum cephalosporin with or without an aminoglycoside, or a carbapenem. Combinations of a β-lactam and a β-lactamase inhibitor (e.g., ampicillin-sulbactam, ticarcillin-clavulanate, piperacillin-tazobactam) or a carbapenem (imipenem-cilastatin, ertapenem, meropenem) can be used in patients with more complicated histories, previous episodes of pyelonephritis, anticipated antimicrobial resistance, or recent urinary tract manipulations; in general, the treatment of such patients should be guided by urine culture results. Once the patient has responded clinically, oral therapy should be substituted for parenteral therapy.

UTI IN PREGNANT WOMEN

Nitrofurantoin, ampicillin, and the cephalosporins are considered relatively safe in early pregnancy. One retrospective case-control study suggesting an association between nitrofurantoin and birth defects has not been confirmed. Sulfonamides should clearly be avoided both in the first trimester (because of possible teratogenic effects) and near term (because of a possible role in the development of kernicterus). Fluoroquinolones are avoided because of possible adverse effects on fetal cartilage development. Ampicillin and the cephalosporins have been used extensively in pregnancy and are the drugs of choice for the treatment of asymptomatic or symptomatic UTI in this group of patients. Generally, pregnant women with ASB are treated for 4–7 days in the absence of evidence to support single-dose therapy. For pregnant women with overt pyelonephritis, parenteral β-lactam therapy with or without aminoglycosides is the standard of care.

UTI IN MEN

Since the prostate is involved in the majority of cases of febrile UTI in men, the goal in these patients is to eradicate the prostatic infection as well as the bladder infection. A 7- to 14-day course of a fluoroquinolone or TMP-SMX is recommended if the uropathogen is susceptible. If acute bacterial prostatitis is suspected, antimicrobial therapy should be initiated after urine and blood are obtained for cultures. Therapy can be tailored to urine culture results and should be continued for 2–4 weeks. For documented chronic bacterial prostatitis, a 4- to 6-week course of antibiotics is often necessary. Recurrences, which are not uncommon in chronic prostatitis, often warrant a 12-week course of treatment.

COMPLICATED UTI

Complicated UTI (other than that discussed above) occurs in a heterogeneous group of patients with a wide variety of structural and functional abnormalities of the urinary tract and kidneys. The range of species and their susceptibility to antimicrobial agents are likewise heterogeneous. As a consequence, therapy for complicated UTI must be individualized and guided by urine culture results. Frequently, a patient with complicated UTI will have prior urine-culture data that can be used to guide empirical therapy while current culture results are pending. Xanthogranulomatous pyelonephritis is treated with nephrectomy. Percutaneous drainage can be used as the initial therapy in emphysematous pyelonephritis and can be followed by elective nephrectomy as needed. Papillary necrosis with obstruction requires intervention to relieve the obstruction and to preserve renal function.

ASYMPTOMATIC BACTERIURIA

Treatment of ASB does not decrease the frequency of symptomatic infections or complications except in pregnant women, persons undergoing urologic surgery, and perhaps neutropenic patients and renal transplant recipients. Treatment of ASB in pregnant women and patients undergoing urologic procedures should be directed by urine culture results. In all other populations, screening for and treatment of ASB are discouraged. The majority of cases of catheter-associated bacteriuria are asymptomatic and do not warrant antimicrobial therapy.

CATHETER-ASSOCIATED UTI

Multiple institutions have released guidelines for the treatment of CAUTI, which is defined by bacteriuria and symptoms in a catheterized patient. The signs and symptoms either are localized to the urinary tract or can include otherwise unexplained systemic manifestations, such as fever. The accepted threshold for bacteriuria to meet the definition of CAUTI is $\geq 10^3$ CFU/mL of urine, while the threshold for bacteriuria to meet the definition of ASB is $\geq 10^5$ CFU/mL.

As catheters provide a conduit for bacteria to enter the bladder, bacteriuria is inevitable with long-term catheter use. The typical signs and symptoms of UTI, including pain, urgency, dysuria, fever, peripheral leukocytosis, and pyuria, have less predictive value for the diagnosis of infection in catheterized patients. Furthermore, the presence of bacteria in the urine of a patient who is febrile and catheterized does not necessarily mean that the patient has CAUTI, and other explanations for the fever should be considered.

The etiology of CAUTI is diverse, and urine culture results are essential to guide treatment. Fairly good evidence supports the practice of catheter change during treatment for CAUTI. The goal is to remove biofilm-associated organisms that could serve as a nidus for reinfection. Pathology studies reveal that many patients with long-term catheters have occult pyelonephritis. A randomized trial in persons with spinal cord injury who were undergoing intermittent catheterization found that relapse was more common after 3 days of therapy than after 14 days. In general, a 7- to 14-day course of antibiotics is recommended, but further studies on the optimal duration of therapy are needed.

The best strategy for prevention of CAUTI is to avoid insertion of unnecessary catheters and to remove catheters once they are no longer necessary. Quality-improvement collaboratives that have addressed technical aspects of CAUTI prevention (such as avoidance of inappropriate catheterization) as well as team communication strategies have shown the benefit of this approach in decreasing CAUTI in both acute- and long-term-care settings. Antimicrobial catheters impregnated with silver or nitrofurazone have not been shown to provide significant clinical benefit in terms of reducing rates of symptomatic UTI. Evidence is insufficient to recommend suprapubic catheters and condom catheters as alternatives to indwelling urinary catheters as a means to prevent bacteriuria. However, intermittent catheterization may be preferable to long-term indwelling urethral catheterization in certain populations (e.g., spinal cord–injured persons) to prevent both infectious and anatomic complications.

CANDIDURIA

The appearance of *Candida* in the urine is an increasingly common complication of indwelling catheterization, particularly for patients in the intensive care unit, those taking broad-spectrum antimicrobial drugs, and those with underlying diabetes mellitus. In many studies, >50% of urinary *Candida* isolates have been found to be non-*albicans* species. The clinical presentation varies from a laboratory finding without symptoms to pyelonephritis and even sepsis. Removal of the urethral catheter results in resolution of candiduria in more than one-third of asymptomatic cases. Treatment of asymptomatic patients does not appear to decrease the frequency of recurrence of candiduria. Therapy is recommended for patients who have symptomatic cystitis or pyelonephritis and for those who are at high risk for disseminated disease. High-risk patients include those with neutropenia, those who are undergoing

urologic manipulation, those who are clinically unstable, and low-birth-weight infants. Fluconazole (200–400 mg/d for 7–14 days) reaches high levels in urine and is the first-line regimen for *Candida* infections of the urinary tract. Although instances of successful eradication of candiduria by some of the newer azoles and echinocandins have been reported, these agents are characterized by only low-level urinary excretion and thus are not recommended. For *Candida* isolates with high levels of resistance to fluconazole, oral flucytosine and/or parenteral amphotericin B are options. Bladder irrigation with amphotericin B generally is not recommended.

TABLE 130-1

Treatment Strategies for Acute Uncomplicated Cystitis

Drug and Dose	Estimated Clinical Efficacy, %	Estimated Bacterial Efficacy,[a] %	Common Side Effects
Nitrofurantoin, 100 mg bid × 5–7 d	87–95	82–92	Nausea, headache
TMP-SMX, 1 DS tablet bid × 3 d	86–100	85–100	Rash, urticaria, nausea, vomiting, hematologic abnormalities
Fosfomycin, 3-g single-dose sachet	83–95	78–98	Diarrhea, nausea, headache
Pivmecillinam, 400 mg bid × 3–7 d	55–82	74–84	Nausea, vomiting, diarrhea
Fluoroquinolones, dose varies by agent; 3-d regimen	81–98	78–96	Nausea, vomiting, diarrhea, headache, drowsiness, insomnia
β-Lactams, dose varies by agent; 5- to 7-d regimen	79–98	74–98	Diarrhea, nausea, vomiting, rash, urticaria

[a]Microbial response as measured by reduction of bacterial counts in the urine.

Note: Efficacy rates are averages or ranges calculated from the data and studies included in the 2010 Infectious Diseases Society of America/European Society of Clinical Microbiology and Infectious Diseases guideline for treatment of uncomplicated UTI and the 2014 *JAMA* systematic review on UTI in the outpatient setting. Ranges are estimates from published studies and may vary by specific agent and by rate of resistance.

Abbreviations: DS, double-strength; TMP-SMX, trimethoprim-sulfamethoxazole.

PREVENTION OF RECURRENT UTI IN WOMEN

Recurrence of uncomplicated cystitis in reproductive-age women is common, and a preventive strategy is indicated if recurrent UTIs are interfering with a patient's lifestyle. The threshold of two or more symptomatic episodes per year is not absolute; decisions about interventions should take the patient's preferences into account.

Three prophylactic strategies are available: continuous, postcoital, and patient-initiated therapy. Continuous prophylaxis and postcoital prophylaxis usually entail low doses of TMP-SMX, a fluoroquinolone, or nitrofurantoin. These regimens are all highly effective during the period of active antibiotic intake. Typically, a prophylactic regimen is prescribed for 6 months and then discontinued, at which point the rate of recurrent UTI often returns to baseline. If bothersome infections recur, the prophylactic program can be reinstituted for a longer period. Selection of resistant strains in the fecal flora has been documented in studies of women taking prophylactic antibiotics for 12 months.

Patient-initiated therapy involves supplying the patient with materials for urine culture and with a course of antibiotics for self-medication at the first symptoms of infection. The urine culture is refrigerated and delivered to the physician's office for confirmation of the diagnosis. When an established and reliable patient–provider relationship exists, the urine culture can be omitted as long as the symptomatic episodes respond completely to short-course therapy and are not followed by relapse.

Non-antimicrobial prevention is increasingly being studied. Lactobacillus probiotics are one appealing approach to UTI prevention, but there is a paucity of data to support this strategy. Similarly, studies of cranberry products for UTI prevention have produced mixed results. Varied dosing and product composition between studies remains an issue for providing clinical guidance.

PROGNOSIS

Cystitis is a risk factor for recurrent cystitis and pyelonephritis. ASB is common among elderly and catheterized patients but does not in itself increase the risk of death. The relationships among recurrent UTI, chronic pyelonephritis, and renal insufficiency have been widely studied. In the absence of anatomic abnormalities such as reflux, recurrent infection in children and adults does not lead to chronic pyelonephritis or to renal failure. Moreover, infection does not play a primary role in chronic interstitial nephritis; the primary etiologic factors in this condition are analgesic abuse, obstruction, reflux, and toxin exposure. In the presence of underlying renal abnormalities (particularly obstructing stones), infection as a secondary factor can accelerate renal parenchymal damage. In spinal cord–injured patients, use of a long-term indwelling bladder catheter is a well-documented risk factor for bladder cancer. Chronic bacteriuria resulting in chronic inflammation is one possible explanation for this observation.

FURTHER READING

BENT S et al: Does this woman have acute uncomplicated urinary tract infection? JAMA 287:2701, 2002. [PubMed: 12020306]

GRIGORYAN L et al: Urinary tract infections in young adults. JAMA 312:1677, 2014. [PubMed: 25335150]

GUPTA K et al: International clinical practice guidelines for the treatment of acute uncomplicated cystitis and pyelonephritis in women: A 2010 update by the Infectious Diseases Society of America and the European Society for Microbiology and Infectious Diseases. Clin Infect Dis 52:e103, 2011. [PubMed: 21292654]

HOOTON TM et al: Diagnosis, prevention, and treatment of catheter-associated urinary tract infection in adults: 2009 international clinical practice guidelines from the Infectious Diseases Society of America. Clin Infect Dis 50:625, 2010. [PubMed: 20175247]

HOOTON TM et al: Voided midstream urine culture and acute cystitis in premenopausal women. N Engl J Med 369; 1883, 2013. [PubMed: 24224622]

NICOLLE LE et al: Infectious Diseases Society of America guidelines for the diagnosis and treatment of asymptomatic bacteriuria in adults. Clin Infect Dis 40:643, 2005. [PubMed: 15714408]

Chapter 131: Sexually Transmitted Infections: Overview and Clinical Approach

Jeanne M. Marrazzo; King K. Holmes

CLASSIFICATION AND EPIDEMIOLOGY

Worldwide, most adults acquire at least one sexually transmitted infection (STI), and many remain at risk for complications. Each day, for example, more than 1 million STIs are acquired worldwide, placing many affected persons at risk for adverse reproductive health outcomes and neoplasia. Certain STIs, such as syphilis, gonorrhea, HIV infection, hepatitis B, and chancroid, often occur in highly interconnected sexual networks characterized by high rates of partner change or multiple concurrent partners. Such networks, for example, often include persons who engage in transactional sex, men who have sex with men (MSM), and persons involved in the use of illicit drugs, particularly methamphetamine. Other STIs are distributed more evenly throughout populations. For example, chlamydial infections, genital human papillomavirus (HPV) infections, and genital herpes can spread efficiently even in relatively low-risk populations. Finally, modern technologies based on detection of nucleic acid have accelerated elucidation of the role of sexual transmission in the spread of some viruses, including Ebola virus and Zika virus, and have provided new evidence of apparent sexual transmission of several bacteria, including group C *Neisseria meningitidis* and anaerobes associated with bacterial vaginosis (BV).

In general, the product of three factors determines the initial rate of spread of any STI within a population: rate of sexual exposure of susceptible to infectious people, efficiency of transmission per exposure, and duration of infectivity of those infected. Accordingly, efforts to prevent and control STIs aim to decrease the rate of sexual exposure of susceptible to infected persons (e.g., through education and efforts to change sexual behavior norms and through control efforts aimed at reducing the proportion of the population infected); to decrease the duration of infectivity (through early diagnosis and curative or suppressive treatment); and to decrease the efficiency of transmission (through promotion of condom use and safer sexual practices, use of effective vaccines, and male medical circumcision).

In all societies, STIs rank among the most common of all infectious diseases, with at least 40 microorganisms now classified as predominantly sexually transmitted or as frequently sexually transmissible (Table 131-1). In developing countries, with three-quarters of the world's population and 90% of the world's STIs, factors such as population growth (especially in adolescent and young-adult age groups), rural-to-urban migration, wars, limited or no provision of reproductive health services for women, and poverty create

exceptional vulnerability to disease resulting from unprotected sex. During the 1990s in China, Russia, the other states of the former Soviet Union, and South Africa, internal social structures changed rapidly as borders opened to the West, unleashing enormous new epidemics of HIV infection and other STIs. Despite advances in the provision of highly effective antiretroviral therapy worldwide, HIV remains the leading cause of death in some developing countries, and HPV and hepatitis B virus (HBV) remain important causes of cervical and hepatocellular carcinoma, respectively—two of the most common (and preventable) malignancies in the developing world. Sexually transmitted herpes simplex virus (HSV) infection causes most genital ulcer disease throughout the world, and an increasing proportion of cases of genital herpes occur in developing countries with generalized HIV epidemics, where the positive-feedback loop between HSV and HIV transmission remains intractable. Despite this consistent link, randomized trials evaluating the efficacy of antiviral therapy in suppressing HSV in both HIV-uninfected and HIV-infected persons have demonstrated no protective effect against acquisition or transmission of HIV. The World Health Organization estimated that 357 million new cases of four curable STIs—gonorrhea, chlamydial infection, syphilis, and trichomoniasis—occurred annually in recent years. Up to 50% of women of reproductive age in developing countries have BV (arguably acquired sexually). All of these curable STIs have been associated with increased risk of HIV transmission or acquisition.

TABLE 131-1

Sexually Transmitted and Sexually Transmissible Microorganisms

Bacteria	Viruses	Other[a]
Transmitted in Adults Predominantly by Sexual Intercourse		
Neisseria gonorrhoeae *Chlamydia trachomatis* *Treponema pallidum* *Haemophilus ducreyi* *Klebsiella (Calymmatobacterium) granulomatis* *Ureaplasma urealyticum* *Mycoplasma genitalium*	HIV (types 1 and 2) Human T-cell lymphotropic virus type 1 Herpes simplex virus type 2 Human papillomavirus (multiple genital genotypes) Hepatitis B virus[b] Molluscum contagiosum virus	*Trichomonas vaginalis* *Pthirus pubis*
Sexual Transmission Repeatedly Described but Not Well Defined or Not the Predominant Mode		
Mycoplasma hominis *Gardnerella vaginalis* and other vaginal bacteria Group B *Streptococcus* *Mobiluncus* spp. *Helicobacter cinaedi* *Helicobacter fennelliae* Anaerobes associated with bacterial vaginosis *Leptotrichia/Sneathia* Group C *Neisseria meningitidis*	Cytomegalovirus Human T-cell lymphotropic virus type 2 Hepatitis C virus (?) Hepatitis D virus Herpes simplex virus type 1 Zika virus Ebola virus (?) Epstein-Barr virus Human herpesvirus type 8	*Candida albicans* *Sarcoptes scabiei*
Transmitted by Sexual Contact Involving Oral–Fecal Exposure; of Declining Importance in Men Who Have Sex with Men		
Shigella spp. *Campylobacter* spp.	Hepatitis A virus	*Giardia lamblia* *Entamoeba histolytica*

[a]Includes protozoa, ectoparasites, and fungi. [b]Among U.S. patients for whom a risk factor can be ascertained, most hepatitis B virus infections are transmitted sexually.

In the United States, the prevalence of antibody to HSV-2 began to fall in the late 1990s, especially among adolescents and young adults; the decline was presumably due to delayed sexual debut, increased condom use, and lower rates of multiple (four or more) sex partners—all well documented by the U.S. Youth Risk Behavior Surveillance System. The estimated annual incidence of HBV infection has also declined dramatically since the mid-1980s; this decrease is probably attributable to now-widespread administration of hepatitis B vaccine in infancy. Genital HPV remains the most common sexually transmitted pathogen in the United States, infecting 60% of a cohort of initially HPV-negative, sexually active Washington state college women within 5 years in a study conducted from 1990 to 2000—i.e., during the pre–HPV immunization era. The scale-up of HPV vaccine coverage among young women has already shown promise in reducing the incidence of infection with the HPV types included in the vaccines and of conditions associated with these viruses.

In industrialized countries, fear of HIV infection in the mid-1980s and through the mid-2000s, coupled with widespread behavioral interventions and better-organized systems of care for the curable STIs, initially helped curb the transmission of several STDs. However, with well-tolerated and highly effective antiretroviral therapy now available, HIV has become for many a chronic disease associated with a normal life span and high quality of life. Rates of gonorrhea and syphilis remain higher in the United States than in any other Western industrialized country.

In the United States, the Centers for Disease Control and Prevention (CDC) has compiled reported rates of STIs since 1941. The incidence of reported gonorrhea peaked at 468 cases per 100,000 population in the mid-1970s and fell to a low of 98 cases per 100,000 in 2012. With increased testing and more sensitive tests, the incidence of reported *Chlamydia trachomatis* infection has been increasing steadily since reporting began in 1984, reaching an all-time peak of 457.6 cases per 100,000 in 2011. The incidence of primary and secondary syphilis per 100,000 peaked at 71 cases in 1946, fell rapidly to 3.9 cases in 1956, ranged from ~10 to 15 cases through 1987 (with markedly increased rates among MSM and African Americans), and then fell to a nadir of 2.1 cases in 2000–2001 (with rates falling most rapidly among heterosexual African Americans). However, since 1996, with the introduction of highly active antiretroviral therapy, gonorrhea, syphilis, and chlamydial infection have had a remarkable resurgence among MSM in North America and Europe, where outbreaks of a rare type of chlamydial infection (lymphogranuloma venereum [LGV]) that had virtually disappeared during the AIDS era have occurred. In 2014, ~75% of primary and secondary syphilis cases reported to the CDC were in MSM. Moreover, the uptake of daily oral emtricitabine/tenofovir as oral pre-exposure prophylaxis for HIV-1 acquisition has increased among MSM since its initial approval for this purpose in 2012 and has been associated with reports of reduced condom-use frequency and concomitantly increased STI acquisition. These developments have resulted in a soaring incidence of STIs, with increasing co-infection with HIV and other sexually transmitted pathogens (particularly *Treponema pallidum*, the cause of syphilis; and *Neisseria gonorrhoeae*, the cause of gonorrhea), primarily among MSM.

MANAGEMENT OF COMMON SEXUALLY TRANSMITTED DISEASE (STD) SYNDROMES

Although other chapters discuss management of specific STIs, most patients are managed (at least initially) on the basis of presenting symptoms and signs and associated risk factors, even in industrialized countries. Table 131-2 lists some of the most common clinical STD syndromes and their microbial etiologies. Strategies for their management are outlined below. **Chapters 196 and 197 address the management of infections with human retroviruses.**

TABLE 131-2

Major Sexually Transmitted Disease Syndromes and Sexually Transmitted Microbial Etiologies

Syndrome	Sexually Transmitted Microbial Etiologies
AIDS	HIV types 1 and 2
Urethritis: males	*Neisseria gonorrhoeae, Chlamydia trachomatis, Mycoplasma genitalium, Ureaplasma urealyticum* (subspecies *urealyticum*), *Trichomonas vaginalis,* HSV, some anaerobic bacteria, *Leptotrichia/Sneathia*
Epididymitis	*C. trachomatis, N. gonorrhoeae*, and (in older men or men who have sex with men) coliform bacteria
Lower genital tract infections: females	
Cystitis/urethritis	*C. trachomatis, N. gonorrhoeae,* HSV
Mucopurulent cervicitis	*C. trachomatis, N. gonorrhoeae, M. genitalium*
Vulvitis	*Candida albicans,* HSV
Bartholinitis	*C. albicans, T. vaginalis*
Vulvovaginitis	*C. albicans, T. vaginalis*
BV	BV-associated bacteria (see text)
Acute pelvic inflammatory disease	*N. gonorrhoeae, C. trachomatis,* BV-associated bacteria, *M. genitalium,* group B streptococci
Infertility	*N. gonorrhoeae, C. trachomatis,* BV-associated bacteria
Ulcerative lesions of the genitalia	HSV-1, HSV-2, *Treponema pallidum, Haemophilus ducreyi, C. trachomatis* (LGV strains), *Klebsiella (Calymmatobacterium) granulomatis*
Complications of pregnancy/puerperium	Several pathogens implicated
Intestinal infections	

Syndrome	Sexually Transmitted Microbial Etiologies
Proctitis	*C. trachomatis, N. gonorrhoeae,* HSV, *T. pallidum*
Proctocolitis or enterocolitis	*Campylobacter* spp., *Shigella* spp., *Entamoeba histolytica, Helicobacter* spp., other enteric pathogens
Enteritis	*Giardia lamblia*
Acute arthritis with urogenital infection or viremia	*N. gonorrhoeae* (e.g., DGI), *C. trachomatis* (e.g., reactive arthritis), HBV
Genital and anal warts	HPV (30 genital types)
Mononucleosis syndrome	CMV, HIV, EBV
Hepatitis	Hepatitis viruses, *T. pallidum,* CMV, EBV
Neoplasias	
Squamous cell dysplasias and cancers of the cervix, anus, vulva, vagina, or penis	HPV (especially types 16, 18, 31, 45)
Kaposi's sarcoma, body-cavity lymphomas	HHV-8
T cell leukemia	HTLV-1
Hepatocellular carcinoma	HBV
Tropical spastic paraparesis	HTLV-1
Scabies	*Sarcoptes scabiei*
Pubic lice	*Pthirus pubis*

Abbreviations: BV, bacterial vaginosis; CMV, cytomegalovirus; DGI, disseminated gonococcal infection; EBV, Epstein-Barr virus; HBV, hepatitis B virus; HHV-8, human herpesvirus type 8; HPV, human papillomavirus; HSV, herpes simplex virus; HTLV, human T-cell lymphotropic virus; LGV, lymphogranuloma venereum.

STD care and management begin with risk assessment and proceed to clinical assessment, diagnostic testing or screening, treatment, and prevention. Risk assessment guides detection and interpretation of symptoms that could denote an STD; decisions on screening or prophylactic/preventive treatment; risk reduction counseling and intervention (e.g., hepatitis B vaccination); treatment of partners of patients with known infections; and behavioral risk reduction by the patient. Consideration of routine demographic data (e.g., gender, age, area of residence) is a simple first step in this risk assessment. For example, national guidelines strongly recommend routine screening of sexually active females ≤25 years of age for *C. trachomatis* infection. **Table 131-3** provides a set of 11 STD/HIV risk-assessment questions that clinicians can pose verbally or that health care systems can adapt (with yes/no responses) into a routine self-administered questionnaire. The initial framing statement gives permission to discuss topics that may be difficult for the patient to disclose.

TABLE 131-3

Eleven-Question STD/HIV Risk Assessment

Framing Statement	
In order to provide the best care for you today and to understand your risk for certain infections, it is necessary for us to talk about your sexual behavior.	
Screening Questions	
(1) Do you have any reason to think you might have a sexually transmitted infection? If so, what reason? (2) For all adolescents <18 years old: Have you begun having any kind of sex yet?	
STD History	
(3) Have you ever had any sexually transmitted infections or any genital infections? If so, which ones?	
Sexual Preference	
(4) Have you had sex with men, women, or both?	
Injection Drug Use	
(5) Have you ever injected yourself (“shot up”) with drugs? (If yes, have you ever shared needles or injection equipment?) (6) Have you ever had sex with a gay or bisexual man or with anyone who had ever injected drugs?	
Characteristics of Partner(s)	
(7) Has your sex partner had any sexually transmitted infections? If so, which ones? (8) Has your sex partner had other sex partners during the time you’ve been together?	
STD Symptoms Checklist	
(9) Have you recently developed any of these symptoms?	
For Men	**For Women**
(a) Discharge of pus (drip) from the penis (b) Genital sores (ulcers) or rash	(a) Abnormal vaginal discharge (increased amount, abnormal odor, abnormal yellow color) (b) Genital sores (ulcers), rash, or itching

Sexual Practices, Past 2 Months (for patients answering yes to any of the above questions, to guide examination and testing)
(10) Now I'd like to ask what parts of your body may have been sexually exposed to an STD (e.g., your penis, mouth, vagina, anus).
Query about Interest in STD Screening Tests (for patients answering no to all of the above questions)
(11) Would you like to be tested for HIV or any other STDs today? (If yes, clinician can explore which STD and why.)

Source: Adapted from JR Curtis, KK Holmes, in KK Holmes et al (eds): *Sexually Transmitted Diseases,* 4th ed. New York, McGraw-Hill, 2008.

Risk assessment is followed by clinical assessment (elicitation of information on specific current symptoms and signs of STDs). Confirmatory diagnostic tests (for persons with symptoms or signs) or screening tests (for those without symptoms or signs) may involve microscopic examination, culture, nucleic acid amplification tests (NAATs), or serology. Initial syndrome-based treatment should cover the most likely causes. For certain syndromes, results of rapid tests can narrow the spectrum of this initial therapy (e.g., pH of vaginal fluid for women with vaginal discharge, Gram's stain of urethral discharge for men with urethral discharge, rapid plasma reagin test for genital ulcer to assess the probability of syphilis). After the institution of treatment, STD management proceeds to the "4 Cs" of prevention and control: *c*ontact tracing (see "Prevention and Control of STIs," below), ensuring *c*ompliance with therapy, and *c*ounseling on risk reduction, including *c*ondom promotion and provision as well as motivational interviewing for risk reduction.

Consistent with current guidelines, all adults should be screened for infection with HIV-1 at least once, and more frequently if they are at elevated risk for acquisition of this infection.

URETHRITIS IN MEN

Urethritis in men produces urethral discharge, dysuria, or both, usually without frequency of urination. Causes include *N. gonorrhoeae, C. trachomatis, Mycoplasma genitalium, Ureaplasma urealyticum, Trichomonas vaginalis*, HSV, and (rarely) adenovirus.

Until recently, *C. trachomatis* caused ~30–40% of cases of nongonococcal urethritis (NGU), particularly in heterosexual men; however, the proportion of cases due to this organism has probably declined in some populations served by effective chlamydial control programs, and older men with urethritis appear less likely to have chlamydial infection. HSV and *T. vaginalis* each cause a small proportion of NGU cases in the United States. Recently, multiple studies have consistently implicated *M. genitalium* as a probable cause of many *Chlamydia*-negative cases. Fewer studies than in the past have implicated *Ureaplasma*; the ureaplasmas have been differentiated into *U. urealyticum* and *Ureaplasma parvum*, and a few studies suggest that *U.*

urealyticum—but not *U. parvum*—is associated with NGU. Coliform bacteria can cause urethritis in men who practice insertive anal intercourse. More recently, anaerobic bacteria that are characteristically involved in BV, especially *Leptotrichia/Sneathia* species, have occasionally been associated with urethritis in heterosexual men. Recommendations for the initial diagnosis of urethritis in men currently include specific tests only for *N. gonorrhoeae* and *C. trachomatis*; they do not yet include testing for *M. genitalium,* although a NAAT is now commercially available for the latter.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Urethritis in Men

The following summarizes the approach to the male patient with suspected urethritis:

1. *Establish the presence of urethritis*. If proximal-to-distal "milking" of the urethra does not express a purulent or mucopurulent discharge, even after the patient has not voided for several hours (or preferably overnight), a Gram's-stained smear of an anterior urethral specimen obtained by passage of a small urethrogenital swab 2–3 cm into the urethra usually reveals ≥2 neutrophils per 1000× field when urethritis is present; in gonococcal infection, such a smear usually reveals gram-negative intracellular diplococci as well. Alternatively, the centrifuged sediment of the first 20–30 mL of voided urine—ideally collected as the first morning specimen—can be examined for inflammatory cells, either by microscopy showing ≥10 leukocytes per high-power field or by the leukocyte esterase test. Patients with symptoms who lack objective evidence of urethritis generally do not benefit from repeated courses of antibiotics, and other etiologies of such symptoms may be considered.

2. *Evaluate for complications or alternative diagnoses*. A brief history and examination can exclude epididymitis and systemic complications, such as disseminated gonococcal infection (DGI) and reactive arthritis. Although digital examination of the prostate gland seldom contributes to the evaluation of sexually active young men with urethritis, men with dysuria who lack evidence of urethritis as well as sexually inactive men with urethritis should undergo prostate palpation, urinalysis, and urine culture to exclude bacterial prostatitis and cystitis.

3. *Evaluate for gonococcal and chlamydial infection*. An absence of typical gram-negative diplococci on Gram's-stained smear of urethral exudate containing inflammatory cells warrants a preliminary diagnosis of NGU, as this test is 98% sensitive for the diagnosis of gonococcal urethral infection. However, an increasing proportion of men with symptoms and/or signs of urethritis are simultaneously assessed for infection with *N. gonorrhoeae* and *C. trachomatis* by NAATs of first-catch urine. The urine specimen tested should consist of the first 10–15 mL of the stream, and, if possible, patients should not have voided for the prior 2 h. Culture or NAAT for *N. gonorrhoeae* may yield positive results even when Gram's staining is negative; certain strains of *N. gonorrhoeae* can result in negative urethral Gram's stains in up to 30% of cases of urethral infection. Results of tests for gonococcal and chlamydial infection predict the patient's prognosis (with greater risk for recurrent NGU if neither chlamydiae nor gonococci are found than if either

is detected) and can guide both the counseling given to the patient and the management of the patient's sexual partner(s).

4. *Treat urethritis promptly while test results are pending.*

TREATMENT

TREATMENT

Urethritis in Men

Table 131-4 summarizes the steps in management of urethral discharge and/or dysuria in sexually active men.

In practice, if Gram's stain does not reveal gonococci, urethritis is treated with a regimen effective for NGU, such as azithromycin or doxycycline. Both are effective. Although azithromycin has been more effective than doxycycline for *M. genitalium* infection, the efficacy of azithromycin for treatment of *M. genitalium* is rapidly declining. Alternatives include moxifloxacin and pristinamycin, a stretogramin antibiotic available in some countries. If gonococci are demonstrated by Gram's stain or if no diagnostic tests are performed to exclude gonorrhea definitively, treatment should include parenteral cephalosporin therapy for gonorrhea **(Chap. 151)** plus oral azithromycin, primarily for additive activity against *N. gonorrhoeae* given concerns about evolving cephalosporin resistance. Azithromycin is effective for treating *C. trachomatis* infection, which can cause urethral co-infection in men with gonococcal urethritis. Sexual partners should also be tested for gonorrhea and chlamydial infection. Regardless of whether they are tested for these infections, however, they should receive the same regimen given to the male index case. Patients with confirmed persistence or recurrence of urethritis after treatment should be re-treated with the initial regimen if they did not comply with the original treatment or were reexposed to an untreated partner. Most persistent urethritis is due to *M. genitalium*, and prompt diagnostic testing and/or treatment for *M. genitalium* is recommended.

National and international guidelines do exist for treatment of gonococcal urethritis, typically with ceftriaxone plus azithromycin. However, consensus is still lacking on treatment of urethritis that persists after treatment and cure of gonorrhea. Ideally, the approach would involve testing for potential causes of the persistent urethritis (e.g., *M. genitalium*) and antimicrobial susceptibility testing in settings and populations where antimicrobial resistance is emerging. Currently, assays are available that can detect *M. genitalium*, and some experts believe it is time to integrate such testing into STD care. If *M. genitalium* is detected, the persistent urethritis can be treated with azithromycin or moxifloxacin in light of local patterns of antimicrobial susceptibility.

In heterosexual men with a high likelihood of exposure to trichomoniasis, an intraurethral swab specimen and a first-voided urine sample should be tested for *T. vaginalis* (often by culture, although NAATs are more sensitive and are approved for the diagnosis of trichomoniasis in women), and presumptive treatment with metronidazole or tinidazole (2 g by mouth in a single dose) should be given. For MSM, trichomoniasis is unlikely, and consideration of a course of moxifloxacin is warranted. Because MSM also have the highest

prevalence rates of antimicrobial-resistant *N. gonorrhoeae*, this possibility, even if apparently ruled out at the initial presentation, should be kept in mind.

TABLE 131-4

Management of Urethral Discharge in Men

Usual Causes	**Usual Initial Evaluation**
Chlamydia trachomatis *Neisseria gonorrhoeae* *Mycoplasma genitalium* *Ureaplasma urealyticum* *Trichomonas vaginalis* Herpes simplex virus	Demonstration of urethral discharge or pyuria Exclusion of local or systemic complications Urethral Gram's stain to confirm urethritis, detect gram-negative diplococci Test for *N. gonorrhoeae, C. trachomatis*
Initial Treatment for Patient and Partners	
Treat gonorrhea (unless excluded): Ceftriaxone (250 mg IM[a)] plus azithromycin (1 g PO)	
Management of Recurrence	
Confirm objective evidence of urethritis. If patient was reexposed to untreated or new partner, repeat treatment of patient and partner. If patient was not reexposed, consider infection with *T. vaginalis*[b] or antibiotic-resistant *M. genitalium*[c] or *Ureaplasma*, and consider treatment with metronidazole, azithromycin, or both.	

[a]Neither oral cephalosporins nor fluoroquinolones are recommended for treatment of gonorrhea in the United States because of the emergence of increasing fluoroquinolone resistance in *N. gonorrhoeae,* especially (but not only) among men who have sex with men, and the decreasing susceptibility of a still-small proportion of gonococci to ceftriaxone (Fig. 131-1). Updates on the emergence of antimicrobial resistance in *N. gonorrhoeae* can be obtained from the Centers for Disease Control and Prevention at *http://www.cdc.gov/std.* [b]In men, the diagnosis of *T. vaginalis* infection requires culture, DNA testing, or nucleic acid amplification testing (where available) of early-morning first-voided urine sediment or of a urethral swab specimen obtained before voiding. [c]*M. genitalium* is often resistant to doxycycline and azithromycin but is usually susceptible to the fluoroquinolone moxifloxacin. Moxifloxacin can be considered for treatment of refractory nongonococcal, nonchlamydial urethritis.

EPIDIDYMITIS

Acute epididymitis, almost always unilateral, produces pain, swelling, and tenderness of the epididymis, with or without symptoms or signs of urethritis. This condition must be differentiated from testicular torsion, tumor, and trauma. Torsion, a surgical emergency, usually occurs in the second or third decade of life and produces a sudden onset of pain, elevation of the testicle within the scrotal sac, rotation of the epididymis from a posterior to an anterior position, and absence of blood flow on Doppler ultrasound. Persistence of symptoms after a course of therapy for epididymitis suggests the possibility of testicular tumor or of a chronic granulomatous disease, such as tuberculosis. In sexually active men under age 35, acute epididymitis is caused most frequently by *C. trachomatis* and less commonly by *N. gonorrhoeae* and is usually associated with overt or subclinical urethritis. Acute epididymitis occurring in older men or following urinary tract instrumentation is usually caused by urinary pathogens. These older men usually have no urethritis but do have bacteriuria. Similarly, epididymitis in MSM who have practiced insertive rectal intercourse is often caused by Enterobacteriaceae.

TREATMENT

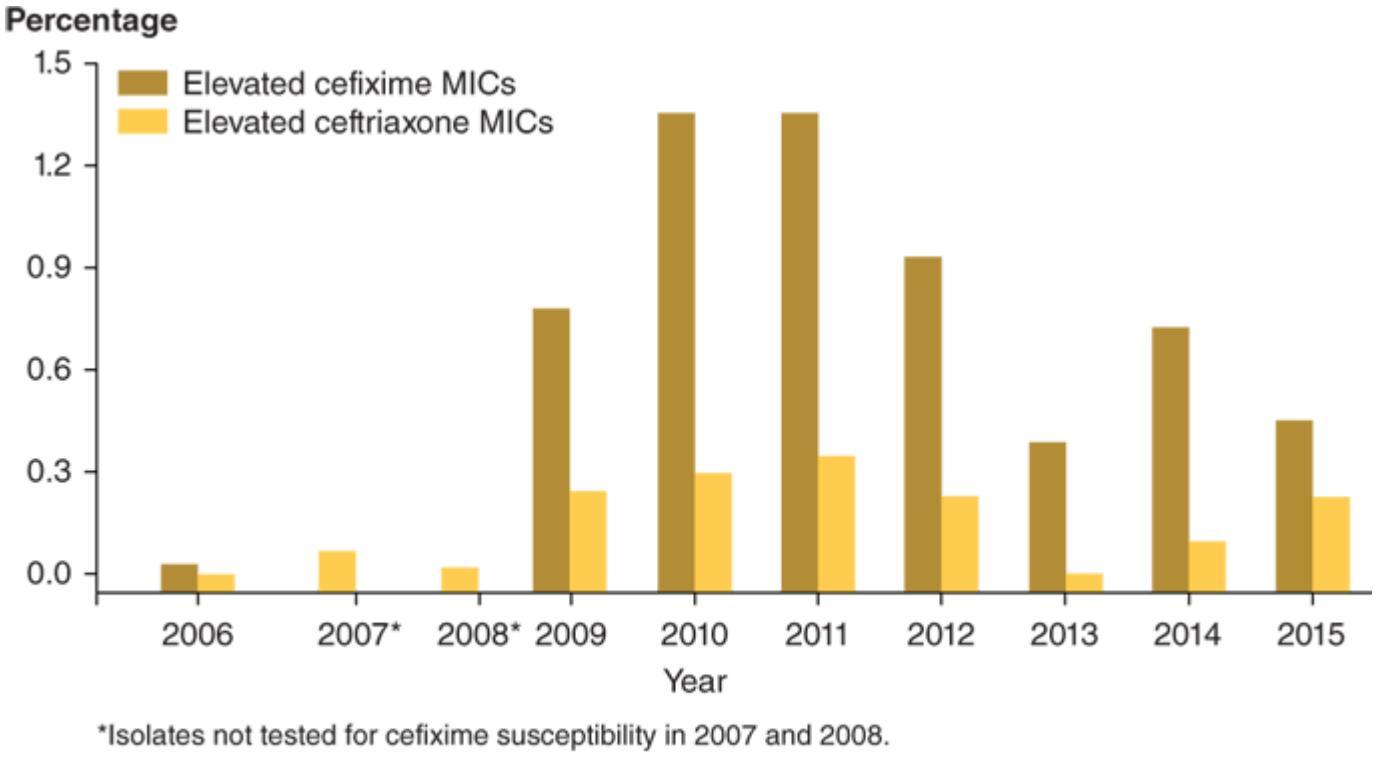
TREATMENT

Epididymitis

Ceftriaxone (250 mg as a single dose IM) followed by doxycycline (100 mg by mouth twice daily for 10 days) constitutes effective treatment for epididymitis caused by *N. gonorrhoeae* or *C. trachomatis*. Neither oral cephalosporins nor fluoroquinolones are recommended for treatment of gonorrhea in the United States because of resistance in *N. gonorrhoeae,* especially (but not only) among MSM **(Fig. 131-1)**. When infection with Enterobacteriaceae is suspected, oral levofloxacin (500 mg once daily for 10 days) or ofloxacin (300 mg twice daily for 10 days) is effective for syndrome-based initial treatment of epididymitis; however, because this regimen is not effective against gonococcal or chlamydial infection, it should be combined with effective therapy for possible gonococcal or chlamydial infection of the epididymis unless bacteriuria with Enterobacteriaceae is confirmed.

FIGURE 131-1

Proportion of *Neisseria gonorrhoeae* isolates with elevated minimal inhibitory concentrations (MICs) of ceftriaxone (≥0.125 µg/mL) and cefixime (≥0.25 µg/mL), United States, 2006–2015. *(From the Centers for Disease Control and Prevention: Gonococcal Isolate Surveillance Project [GISP], 2016.)*



*Isolates not tested for cefixime susceptibility in 2007 and 2008.

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

URETHRITIS AND THE URETHRAL SYNDROME IN WOMEN

C. trachomatis, N. gonorrhoeae, and occasionally HSV cause symptomatic urethritis—known as *the urethral syndrome* in women—that is characterized by "internal" dysuria (usually without urinary urgency or frequency), pyuria, and an absence of *Escherichia coli* and other uropathogens at counts of $\geq 10^2$/mL in urine. In contrast, the dysuria associated with vulvar herpes or vulvovaginal candidiasis (and perhaps with trichomoniasis) is often described as "external," being caused by painful contact of urine with the inflamed or ulcerated labia or introitus. Acute onset, association with urinary urgency or frequency, hematuria, or suprapubic bladder tenderness suggests bacterial cystitis. Among women with symptoms of acute bacterial cystitis, costovertebral pain and tenderness or fever suggest acute pyelonephritis. **The management of bacterial urinary tract infection (UTI) is discussed in Chap. 130.**

Signs of vulvovaginitis, coupled with symptoms of external dysuria, suggest vulvar infection (e.g., with HSV or *Candida albicans*). Among dysuric women without signs of vulvovaginitis, bacterial UTI must be differentiated from the urethral syndrome by assessment of risk, evaluation of the pattern of symptoms and signs, and specific microbiologic testing. An STI etiology of the urethral syndrome is suggested by young age, more than one current sexual partner, a new partner within the past month, a partner with urethritis, or coexisting mucopurulent cervicitis (see below). The finding of a single urinary pathogen, such as *E. coli* or *Staphylococcus saprophyticus*, at a concentration of $\geq 10^2$/mL in a properly collected specimen of midstream urine from a dysuric woman with pyuria indicates probable bacterial UTI, whereas pyuria with $<10^2$ conventional uropathogens per milliliter of urine ("sterile" pyuria) suggests acute urethral syndrome due to *C. trachomatis* or *N. gonorrhoeae*. Gonorrhea and chlamydial infection should be sought by specific tests

(e.g., NAATs of vaginal secretions collected with a swab). Among dysuric women with sterile pyuria caused by infection with *N. gonorrhoeae* or *C. trachomatis*, appropriate treatment alleviates dysuria. The role of *M. genitalium* in the urethral syndrome in women remains undefined.

VULVOVAGINAL INFECTIONS

Abnormal Vaginal Discharge

If directly questioned about vaginal discharge during routine health checkups, many women acknowledge having nonspecific symptoms of vaginal discharge that do not correlate with objective signs of inflammation or with actual infection. However, unsolicited reporting of abnormal vaginal discharge often denotes BV or trichomoniasis. Specifically, an abnormally increased amount or an abnormal odor of the discharge is associated with one or both of these conditions. Cervical infection with *N. gonorrhoeae* or *C. trachomatis* does not often cause an increased amount or abnormal odor of discharge; however, when these pathogens cause cervicitis, they—like *T. vaginalis*—often result in an increased number of neutrophils in vaginal fluid, which thus takes on a yellow color. Vulvar conditions such as genital herpes or vulvovaginal candidiasis can cause vulvar pruritus, burning, irritation, or lesions as well as external dysuria (as urine passes over the inflamed vulva or areas of epithelial disruption) or vulvar dyspareunia.

Certain vulvovaginal infections may have serious sequelae. Trichomoniasis, BV, and vulvovaginal candidiasis have all been associated with increased risk of acquisition of HIV infection; BV promotes HIV transmission from HIV-infected women to their male sex partners. Vaginal trichomoniasis and BV early in pregnancy independently predict premature onset of labor. BV can also lead to anaerobic bacterial infection of the endometrium and salpinges. Vaginitis may be an early and prominent feature of toxic shock syndrome, and recurrent or chronic vulvovaginal candidiasis develops with increased frequency among women who have systemic illnesses, such as diabetes mellitus or HIV-related immunosuppression (although only a very small proportion of women with recurrent vulvovaginal candidiasis in industrialized countries actually have a serious predisposing illness).

Thus vulvovaginal symptoms or signs warrant careful evaluation, including speculum and pelvic examination, diagnostic testing, and appropriate therapy specific for the infection identified. Unfortunately, clinicians do not always perform the tests required to establish the cause of such symptoms. Further, self-diagnosis of a specific type of infection—including vulvovaginal candidiasis—is often incorrect. The diagnosis and treatment of the three most common types of vaginal infection are summarized in **Table 131-5**.

TABLE 131-5

Diagnostic Features and Management of Vaginal Infection

Feature	Normal Vaginal Examination	Vulvovaginal Candidiasis	Trichomonal Vaginitis	Bacterial Vaginosis
Etiology	Uninfected; lactobacilli predominant	*Candida albicans*	*Trichomonas vaginalis*	Associated with *Gardnerella vaginalis*, various anaerobic and/or noncultured bacteria, and mycoplasmas
Typical symptoms	None	Vulvar itching and/or irritation	Profuse discharge; vulvar itching	Malodorous, slightly increased discharge
Discharge				
Amount	Variable; usually scant	Scant	Often profuse	Moderate
Color[a]	Clear or translucent	White	White or yellow	White or gray
Consistency	Nonhomogeneous, flocculent	Clumped; adherent plaques	Homogeneous	Homogeneous, low viscosity; uniformly coats vaginal walls
Inflammation of vulvar or vaginal epithelium	None	Erythema of vaginal epithelium, introitus; vulvar dermatitis, fissures common	Erythema of vaginal and vulvar epithelium; colpitis macularis	None
pH of vaginal fluid[b]	Usually ≤4.5	Usually ≤4.5	Usually ≥5	Usually >4.5

Feature	Normal Vaginal Examination	Vulvovaginal Candidiasis	Trichomonal Vaginitis	Bacterial Vaginosis
Amine ("fishy") odor with 10% KOH	None	None	May be present	Present
Microscopy[c]	Normal epithelial cells; lactobacilli predominant	Leukocytes, epithelial cells; mycelia or pseudomycelia in up to 80% of *C. albicans* culture–positive persons with typical symptoms	Leukocytes; motile trichomonads seen in 80–90% of symptomatic patients, less often in the absence of symptoms	Clue cells; few leukocytes; no lactobacilli or only a few outnumbered by profuse mixed microbiota, nearly always including *G. vaginalis* plus anaerobic species on Gram's stain (Nugent's score ≥7)
Other laboratory findings		Isolation of *Candida* spp.	Isolation of *T. vaginalis* or positive NAAT[d]	
Usual treatment	None	Azole cream, tablet, or suppository—e.g., miconazole (100-mg vaginal suppository) or clotrimazole (100-mg vaginal tablet) once daily for 7 days Fluconazole, 150 mg orally (single dose)	Metronidazole or tinidazole, 2 g orally (single dose) Metronidazole, 500 mg PO bid for 7 days	Metronidazole, 500 mg PO bid for 7 days Metronidazole gel, 0.75%, one applicator (5 g) intravaginally once daily for 5 days Clindamycin, 2% cream, one full applicator vaginally each night for 7 days

Feature	Normal Vaginal Examination	Vulvovaginal Candidiasis	Trichomonal Vaginitis	Bacterial Vaginosis
Usual management of sexual partner	None	None; topical treatment if candidal dermatitis of penis is detected	Examination for sexually transmitted infection; treatment with metronidazole, 2 g PO (single dose)	None

[a]Color of discharge is best determined by examination against the white background of a swab. [b]A pH determination is not useful if blood is present or if the test is performed on endocervical secretions. [c]To detect fungal elements, vaginal fluid is digested with 10% KOH prior to microscopic examination; to examine for other features, fluid is mixed (1:1) with physiologic saline. Gram's stain is also excellent for detecting yeasts (less predictive of vulvovaginitis) and pseudomycelia or mycelia (strongly predictive of vulvovaginitis) and for distinguishing normal flora from the mixed flora seen in bacterial vaginosis, but it is less sensitive than the saline preparation for detection of *T. vaginalis.* [d]NAAT, nucleic acid amplification test (where available).

Inspection of the vulva and perineum may reveal tender genital ulcerations or fissures (typically due to HSV infection or vulvovaginal candidiasis) or discharge visible at the introitus before insertion of a speculum (suggestive of BV or trichomoniasis). Speculum examination permits the clinician to discern whether the discharge appears abnormal and whether it emanates from the cervical os (mucoid and, if abnormal, yellow) or from the vagina (not mucoid, since the vaginal epithelium does not produce mucus). Symptoms or signs of abnormal vaginal discharge should prompt testing of vaginal fluid for pH, for a fishy odor when mixed with 10% KOH, and for certain microscopic features when mixed with saline (motile trichomonads and/or "clue cells") and with 10% KOH (pseudohyphae or hyphae indicative of vulvovaginal candidiasis). Additional objective laboratory tests, described below, are useful for establishing the cause of abnormal vaginal discharge. Gram's staining of vaginal fluid can be used to characterize the vaginal bacteria using the Nugent score but is used primarily for research purposes and requires familiarity with the morphotypes and scale involved.

TREATMENT

TREATMENT

Vaginal Discharge

Patterns of treatment for abnormal vaginal discharge vary widely. In developing countries, where clinics or pharmacies often dispense treatment based on symptoms alone without examination or testing, oral

treatment with metronidazole—particularly with a 7-day regimen—provides reasonable coverage against both trichomoniasis and BV, the usual causes of symptoms of vaginal discharge. Metronidazole treatment of sex partners prevents reinfection of women with *T. vaginalis*, although it does not help prevent the recurrence of BV. Guidelines for syndromic management promulgated by the World Health Organization suggest consideration of treatment for cervical infection and for trichomoniasis, BV, and vulvovaginal candidiasis in women with symptoms of abnormal vaginal discharge. However, it is important to note that the majority of chlamydial and gonococcal cervical infections produce no symptoms.

In industrialized countries, clinicians treating symptoms and signs of abnormal vaginal discharge should, at a minimum, differentiate between BV and trichomoniasis, because optimal management of patients and partners differs for these two conditions.

Vaginal Trichomoniasis (See also Chap. 224)

Symptomatic trichomoniasis characteristically produces a profuse, yellow, purulent, homogeneous vaginal discharge and vulvar irritation, sometimes with visible inflammation of the vaginal and vulvar epithelium and petechial lesions on the cervix (the so-called strawberry cervix, best visualized by colposcopy). The pH of vaginal fluid—normally <4.7—usually rises to ≥5. Microscopic examination of vaginal discharge mixed with saline reveals motile trichomonads in most culture-positive cases. However, saline microscopy probably detects only one-half of all cases, and, especially in the absence of symptoms or signs, culture or NAAT is usually required for detection of the organism. NAAT for *T. vaginalis* is more sensitive than culture. Treatment of asymptomatic as well as symptomatic cases reduces rates of transmission and prevents later development of symptoms.

TREATMENT

TREATMENT

Vaginal Trichomoniasis

Only nitroimidazoles (e.g., metronidazole and tinidazole) consistently cure trichomoniasis. A single 2-g oral dose of metronidazole is effective and less expensive than the alternatives. Tinidazole has a longer half-life than metronidazole, causes fewer gastrointestinal symptoms, and may be useful in treating trichomoniasis that fails to respond to metronidazole. Treatment of sexual partners—facilitated by dispensing metronidazole to the female patient to give to her partner(s), with a warning about avoiding the concurrent use of alcohol—significantly reduces both the risk of reinfection and the reservoir of infection; treating partners is the standard of care. Intravaginal treatment with 0.75% metronidazole gel is not reliable for vaginal trichomoniasis. Thus, systemic use of metronidazole is still recommended throughout pregnancy for treatment of trichomoniasis. In a large randomized trial, metronidazole treatment of trichomoniasis during pregnancy was associated with an increased frequency of perinatal morbidity. However, most studies, including randomized controlled trials, have shown no adverse effects of metronidazole use during pregnancy on preterm birth or birth defects.

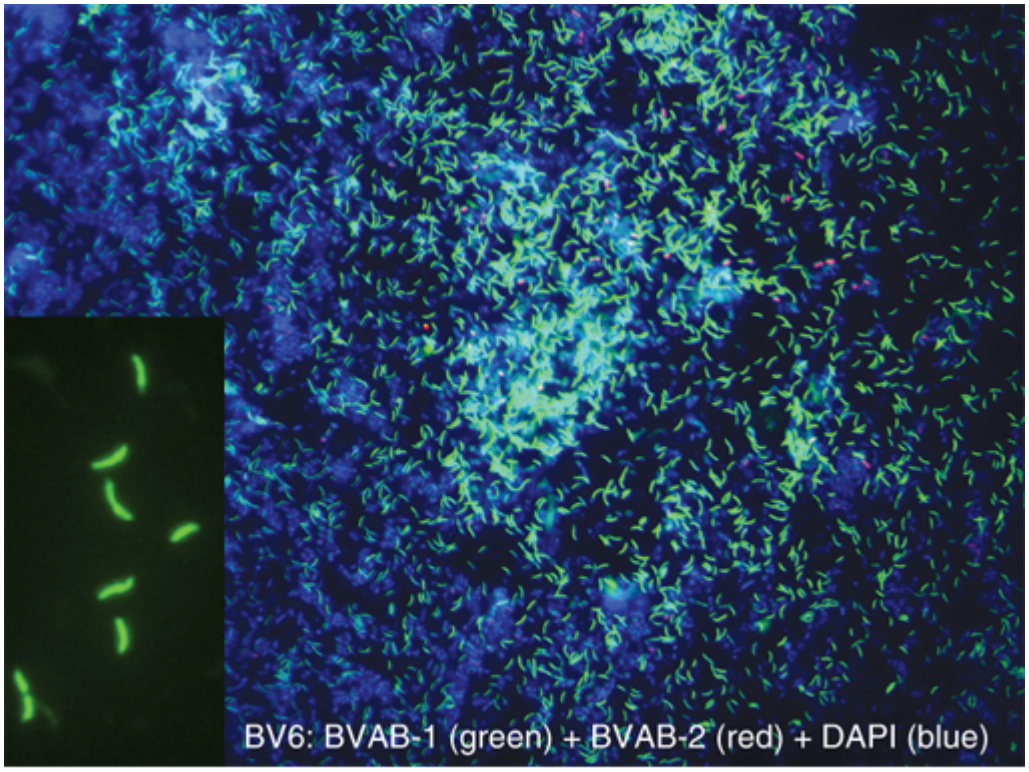
Bacterial Vaginosis

BV is a syndrome characterized by symptoms of vaginal malodor and increased white-gray discharge, which appears homogeneous, is low in viscosity, and uniformly covers the vaginal mucosa. BV has been associated with an increased risk of acquiring several other genital infections, including those caused by HIV, *C. trachomatis*, and *N. gonorrhoeae*. Other possible risk factors include recent unprotected vaginal intercourse, having a female sex partner, and vaginal douching. Although bacteria associated with BV have been detected under the foreskin of uncircumcised men and have been associated with urethritis, metronidazole treatment of male partners has not reduced the rate of recurrence of BV among affected women.

Among women with BV, culture of vaginal fluid has shown markedly increased prevalences and concentrations of *Gardnerella vaginalis, Mycoplasma hominis*, and several anaerobic bacteria (e.g., *Mobiluncus, Prevotella* [formerly *Bacteroides*], and some *Peptostreptococcus* species) as well as an absence of hydrogen peroxide–producing *Lactobacillus* species that constitute most of the normal vaginal microbiota and help protect against cervical and vaginal infections. Broad-range polymerase chain reaction (PCR) amplification of 16S rDNA in vaginal fluid, with subsequent identification of specific bacterial species by various methods, has documented even greater bacterial diversity, including several unique species not previously identified in culture **(Fig. 131-2)** and *Atopobium vaginae*, an organism that is strongly associated with BV and is resistant to metronidazole. Other genera newly implicated in BV include *Megasphaera, Leptotrichia, Eggerthella*, and *Dialister*.

FIGURE 131-2

Broad-range polymerase chain reaction amplification of 16S rDNA in vaginal fluid from a woman with bacterial vaginosis (BV) shows a field of bacteria hybridizing with probes for BV-associated bacterium 1 (BVAB-1, visible as a thin, curved green rod) and for BVAB-2 (red). The *inset* shows that BVAB-1 has a morphology similar to that of *Mobiluncus* (curved rod). *(Reprinted with permission from DN Fredricks et al: N Engl J Med 353:1899, 2005.)*

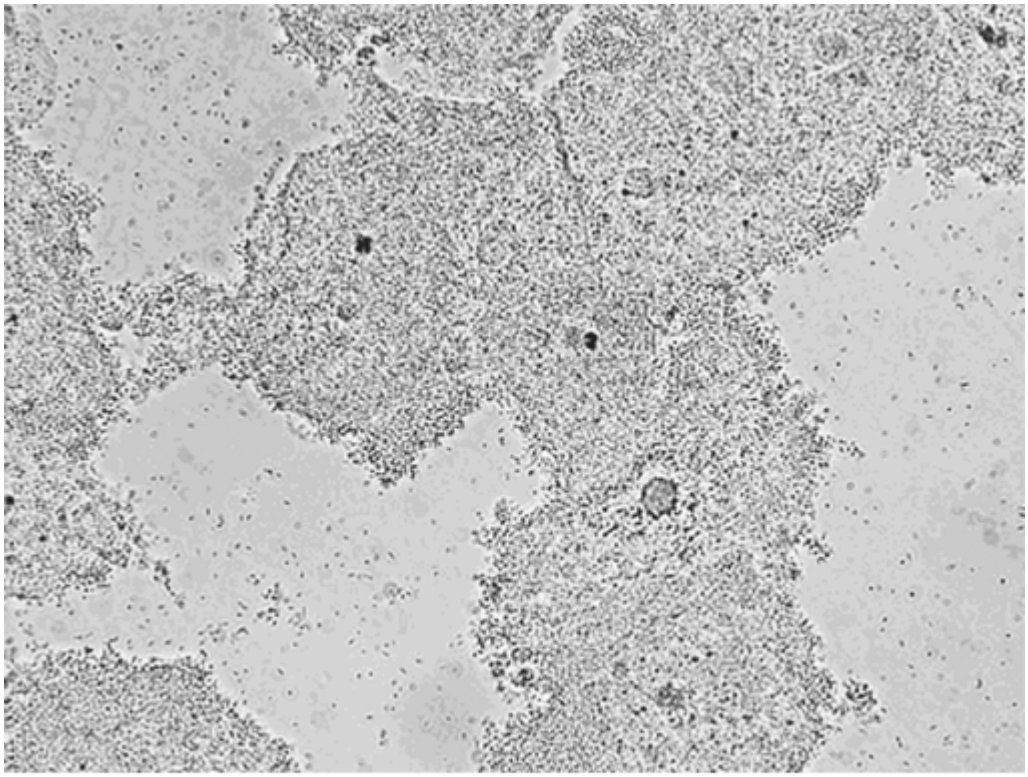


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

BV is conventionally diagnosed clinically with the Amsel criteria, which include any three of the following four clinical abnormalities: (1) objective signs of increased white homogeneous vaginal discharge; (2) a vaginal discharge pH of >4.5; (3) liberation of a distinct fishy odor (attributable to volatile amines such as trimethylamine) immediately after vaginal secretions are mixed with a 10% solution of KOH; and (4) microscopic demonstration of "clue cells" (vaginal epithelial cells coated with coccobacillary organisms, which have a granular appearance and indistinct borders; **Fig. 131-3**) on a wet mount prepared by mixing vaginal secretions with normal saline in a ratio of ~1:1.

FIGURE 131-3

Wet mount of vaginal fluid showing typical clue cells from a woman with bacterial vaginosis. Note the obscured epithelial cell margins and the granular appearance attributable to many adherent bacteria (×400). *(Photograph provided by Lorna K. Rabe, reprinted with permission from S Hillier et al, in KK Holmes et al [eds]: Sexually Transmitted Diseases, 4th ed. New York, McGraw-Hill, 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Bacterial Vaginosis

The standard dosage of oral metronidazole for the treatment of BV is 500 mg twice daily for 7 days. The single 2-g oral dose of metronidazole recommended for trichomoniasis produces significantly lower short-term cure rates and should not be used. Intravaginal treatment with 2% clindamycin cream (one full applicator [5 g containing 100 mg of clindamycin phosphate] each night for 7 nights) or with 0.75% metronidazole gel (one full applicator [5 g containing 37.5 mg of metronidazole] twice daily for 5 days) is also approved for use in the United States and does not elicit systemic adverse reactions; the response to both of these treatments is similar to the response to oral metronidazole. Other alternatives include oral clindamycin (300 mg twice daily for 7 days), clindamycin ovules (100 g intravaginally once at bedtime for 3 days), and oral tinidazole (1 g daily for 5 days or 2 g daily for 3 days). Unfortunately, recurrence over the long term (i.e., several months later) is distressingly common after either oral or intravaginal treatment. A randomized trial comparing intravaginal gel containing 37.5 mg of metronidazole with a suppository containing 500 mg of metronidazole plus nystatin (the latter not marketed in the United States) showed significantly higher rates of recurrence with the 37.5-mg regimen; this result suggests that higher metronidazole dosages may be important in topical intravaginal therapy. Recurrences can be significantly lessened with the twice-weekly use of suppressive intravaginal metronidazole gel.

Efforts to replenish numbers of vaginal lactobacilli that produce hydrogen peroxide and probably sustain vaginal health have been unsuccessful. While one randomized trial of orally ingested lactobacilli found reduced rates of recurrent BV, this result has not yet been either confirmed or refuted, and a randomized multicenter trial in the United States found no benefit of repeated intravaginal inoculation of a vaginal peroxide-producing *Lactobacillus* species following treatment of BV with metronidazole. A meta-analysis of 18 studies concluded that BV during pregnancy substantially increased the risk of preterm delivery and of spontaneous abortion. However, in most studies, topical intravaginal treatment of BV with clindamycin during pregnancy has not reduced adverse pregnancy outcomes. Numerous trials of oral metronidazole treatment during pregnancy have given inconsistent results, and recent reviews have concluded that antenatal treatment of women with BV—including those with previous preterm delivery—did not reduce the risk of preterm delivery. The U.S. Preventive Services Task Force thus recommends against routine screening of pregnant women for BV.

Vulvovaginal Pruritus, Burning, or Irritation

Vulvovaginal candidiasis produces vulvar pruritus, burning, or irritation, generally without symptoms of increased vaginal discharge or malodor. Genital herpes can produce similar symptoms, with lesions sometimes difficult to distinguish from the fissures and inflammation caused by candidiasis. Signs of vulvovaginal candidiasis include vulvar erythema, edema, fissures, and tenderness. With candidiasis, a white scanty vaginal discharge sometimes takes the form of white thrush-like plaques or cottage cheese-like curds adhering loosely to the vaginal epithelium. *C. albicans* accounts for nearly all cases of symptomatic vulvovaginal candidiasis, which probably arise from endogenous strains of *C. albicans* that have colonized the vagina or the intestinal tract. Complicated vulvovaginal candidiasis includes cases that recur four or more times per year; are unusually severe; are caused by non-*albicans Candida* species; or occur in women with uncontrolled diabetes, debilitation, immunosuppression, or pregnancy.

In addition to compatible clinical symptoms, the diagnosis of vulvovaginal candidiasis involves the demonstration of pseudohyphae or hyphae by microscopic examination of vaginal fluid mixed with saline or 10% KOH or subjected to Gram's staining. Microscopic examination is less sensitive than culture but correlates better with symptoms. Culture is typically reserved for cases that do not respond to standard first-line antimycotic agents and is undertaken to rule out imidazole or azole resistance (often associated with *Candida glabrata*) or before the initiation of suppressive antifungal therapy for recurrent disease.

TREATMENT

TREATMENT

Vulvovaginal Pruritus, Burning, or Irritation

Symptoms and signs of vulvovaginal candidiasis warrant treatment, usually intravaginal administration of any of several imidazole antibiotics (e.g., miconazole or clotrimazole) for 3–7 days or of a single dose of oral fluconazole (Table 131-5). Over-the-counter marketing of such preparations has reduced the cost of care and made treatment more convenient for many women with recurrent yeast vulvovaginitis. However, most

women who purchase these preparations do not have vulvovaginal candidiasis, whereas many have other vaginal infections that require different treatment. Therefore, only women with classic symptoms of vulvar pruritus and a history of previous episodes of yeast vulvovaginitis documented by an experienced clinician should self-treat. Short-course topical intravaginal azole drugs are effective for the treatment of uncomplicated vulvovaginal candidiasis (e.g., clotrimazole, two 100-mg vaginal tablets daily for 3 days; or miconazole, a 1200-mg vaginal suppository as a single dose). Single-dose oral treatment with fluconazole (150 mg) is also effective and is preferred by many patients. Management of complicated cases (see above) and those that do not respond to the usual intravaginal or single-dose oral therapy often involves prolonged or periodic oral therapy; this situation is discussed extensively in the 2015 CDC STD treatment guidelines (*http://www.cdc.gov/std/treatment*). Treatment of sexual partners is not routinely indicated.

Other Causes of Vaginal Discharge or Vaginitis

In the ulcerative vaginitis associated with staphylococcal toxic shock syndrome, *Staphylococcus aureus* should be promptly identified in vaginal fluid by Gram's stain and by culture. In desquamative inflammatory vaginitis, smears of vaginal fluid reveal neutrophils, massive vaginal epithelial-cell exfoliation with increased numbers of parabasal cells, and gram-positive cocci; this syndrome may respond to treatment with 2% clindamycin cream, often given in combination with topical steroid preparations for several weeks. Additional causes of vaginitis and vulvovaginal symptoms include retained foreign bodies (e.g., tampons), cervical caps, vaginal spermicides, vaginal antiseptic preparations or douches, vaginal epithelial atrophy (in postmenopausal women or during prolonged breast-feeding in the postpartum period), allergic reactions to latex condoms, vaginal aphthae associated with HIV infection or Behçet's syndrome, and vestibulitis.

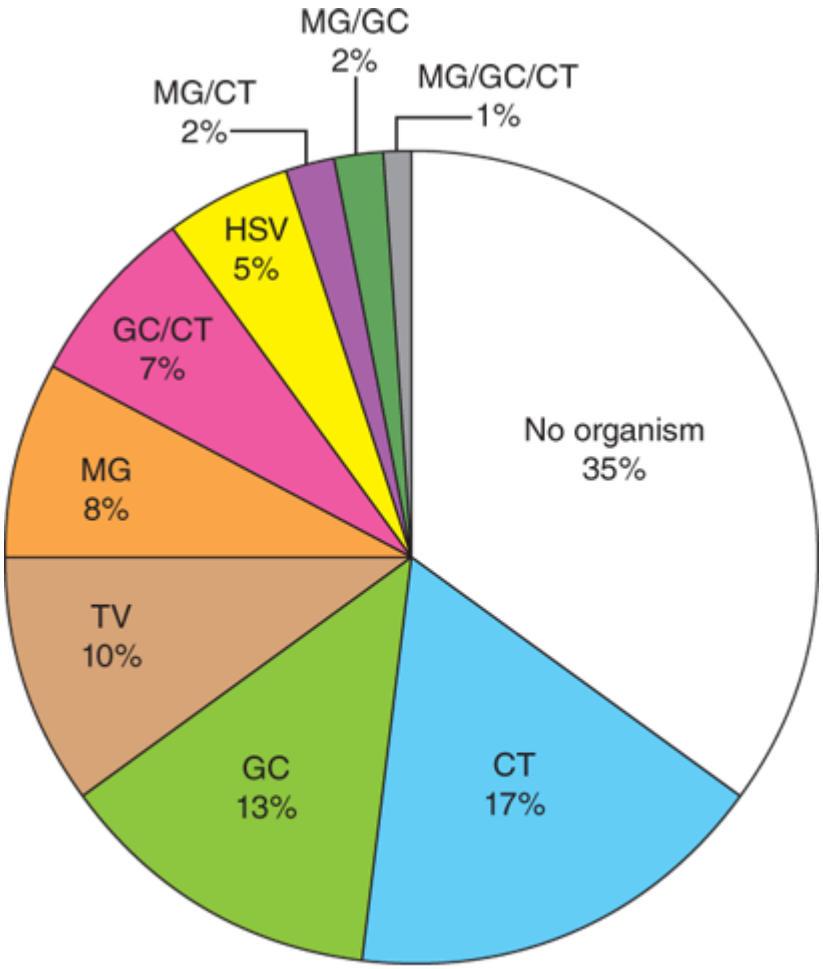
MUCOPURULENT CERVICITIS

Mucopurulent cervicitis (MPC) refers to inflammation of the columnar epithelium and subepithelium of the endocervix and of any contiguous columnar epithelium that lies exposed in an ectopic position on the ectocervix. MPC in women represents the "silent partner" of urethritis in men, being equally common and often caused by the same agents (*N. gonorrhoeae*, *C. trachomatis*, *M. genitalium*); however, MPC is more difficult than urethritis to recognize, given the nonspecific nature of symptoms (e.g., abnormal vaginal discharge) and the need for visualization by pelvic examination. As the most common manifestation of these serious bacterial infections in women, MPC can be a harbinger or sign of upper genital tract infection, also known as *pelvic inflammatory disease* (PID; see below). In pregnant women, MPC can lead to obstetric complications. In the pre-NAAT era, more than one-third of cervicovaginal specimens tested for *C. trachomatis, N. gonorrhoeae, M. genitalium*, HSV, and *T. vaginalis* revealed no identifiable etiology for MPC (Fig. 131-4). More recent studies employing NAATs for these pathogens have still failed to identify a microbiologic etiology in nearly one-half of women with MPC. Individual bacteria associated with BV may also elicit an inflammatory reaction at the cervix; thus, BV may be a cause of MPC.

FIGURE 131-4

Organisms detected among female sexually transmitted disease clinic patients with mucopurulent cervicitis (*n* = 167). CT, *Chlamydia trachomatis*; GC, gonococcus; MG, *Mycoplasma genitalium*; TV, *Trichomonas*

vaginalis; HSV, herpes simplex virus. *(Courtesy of Dr. Lisa Manhart; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The diagnosis of MPC rests on the detection of cardinal signs at the cervix, including yellow mucopurulent discharge from the cervical os, endocervical bleeding upon gentle swabbing, and edematous cervical ectopy (see below); the latter two findings are somewhat more common with MPC due to chlamydial infection, but signs alone do not allow a distinction among the causative pathogens. Unlike the endocervicitis produced by gonococcal or chlamydial infection, cervicitis caused by HSV produces ulcerative lesions on the stratified squamous epithelium of the ectocervix as well as on the columnar epithelium. Yellow cervical mucus on a white swab removed from the endocervix indicates the presence of polymorphonuclear leukocytes (PMNs). Gram's staining may confirm their presence, although it adds relatively little to the diagnostic value of assessment for cervical signs. The presence of ≥20 PMNs per 1000× microscopic field within strands of cervical mucus not contaminated by vaginal squamous epithelial cells or vaginal bacteria indicates endocervicitis. Detection of intracellular gram-negative diplococci in carefully collected endocervical mucus is quite specific but ≤50% sensitive for gonorrhea. Therefore, NAATs for *N. gonorrhoeae* and *C. trachomatis* are always indicated in the evaluation of MPC, as is a careful evaluation of vaginal discharge for the causes of vaginitis discussed above.

TREATMENT

TREATMENT

Mucopurulent Cervicitis

Although the above criteria for MPC are neither highly specific nor highly predictive of gonococcal or chlamydial infection in some settings, the 2015 CDC STD guidelines call for consideration of empirical treatment for MPC, pending test results, in most cases. Presumptive treatment with antibiotics active against *C. trachomatis* should be provided for women at increased risk for this common STI (risk factors: age <25 years, new or multiple sex partners, and unprotected sex), especially if follow-up cannot be ensured. Concurrent therapy for gonorrhea is indicated if the prevalence of this infection is substantial in the relevant patient population (e.g., young adults, a clinic with documented high prevalence). In this situation, therapy should include a single-dose regimen effective for gonorrhea plus treatment for chlamydial infection, as outlined in Table 131-4 for the treatment of urethritis. In settings where gonorrhea is much less common than chlamydial infection, initial therapy for chlamydial infection alone suffices, pending test results for gonorrhea. The etiology and potential benefit of treatment for endocervicitis not associated with gonorrhea or chlamydial infection have not been established. Although the antimicrobial susceptibility of *M. genitalium* is not yet well defined, the organism frequently persists after doxycycline therapy, and it currently seems reasonable to use azithromycin to treat possible *M. genitalium* infection in such cases. With resistance of *M. genitalium* to azithromycin now recognized, moxifloxacin may be a reasonable alternative. The sexual partner(s) of a woman with MPC should be examined and given a regimen similar to that chosen for the woman unless results of tests for gonorrhea or chlamydial infection in either partner warrant different therapy or no therapy.

CERVICAL ECTOPY

Cervical ectopy, often mislabeled “cervical erosion,” is easily confused with infectious endocervicitis. Ectopy represents the presence of the one-cell-thick columnar epithelium extending from the endocervix out onto the visible ectocervix. In ectopy, the cervical os may contain clear or slightly cloudy mucus but usually not yellow mucopus. Colposcopy shows intact epithelium. Normally found during adolescence and early adulthood, ectopy gradually recedes through the second and third decades of life, as squamous metaplasia replaces the ectopic columnar epithelium. Oral contraceptive use favors the persistence or reappearance of ectopy, while smoking apparently accelerates squamous metaplasia. Cauterization of ectopy is not warranted. Ectopy may render the cervix more susceptible to infection with *N. gonorrhoeae, C. trachomatis*, or HIV.

PELVIC INFLAMMATORY DISEASE

The term *pelvic inflammatory disease* usually refers to infection that ascends from the cervix or vagina to involve the endometrium and/or fallopian tubes. Infection can extend beyond the reproductive tract to cause pelvic peritonitis, generalized peritonitis, perihepatitis, perisplenitis, or pelvic abscess. Rarely, infection not related to specific sexually transmitted pathogens extends secondarily to the pelvic organs (1) from adjacent foci of inflammation (e.g., appendicitis, regional ileitis, or diverticulitis) or BV, (2) as a result of hematogenous

dissemination (e.g., of tuberculosis or staphylococcal bacteremia), or (3) as a complication of certain tropical diseases (e.g., schistosomiasis). Intrauterine infection can be primary (spontaneously occurring and usually sexually transmitted) or secondary to invasive intrauterine surgical procedures (e.g., dilation and curettage, termination of pregnancy, insertion of an intrauterine device [IUD], or hysterosalpingography) or to parturition.

Etiology

The agents most often implicated in acute PID include the primary causes of endocervicitis (*N. gonorrhoeae, C. trachomatis*, and *M. genitalium*) and anaerobes associated with BV. In general, PID is most often caused by *N. gonorrhoeae* in settings where there is a high incidence of gonorrhea. *M. genitalium* has also been significantly associated with histopathologic diagnoses of endometritis and with salpingitis.

Anaerobic and facultative organisms (especially *Prevotella* species, peptostreptococci, *E. coli, Haemophilus influenzae*, and group B streptococci) as well as genital mycoplasmas have been isolated from the peritoneal fluid or fallopian tubes in a varying proportion (typically one-fourth to one-third) of women with PID studied in the United States. The difficulty of determining the exact microbial etiology of an individual case of PID—short of using invasive procedures for specimen collection—has implications for the approach to empirical antimicrobial treatment of this infection.

Epidemiology

In the United States, the estimated annual number of initial visits to physicians' offices for PID by women 15–44 years of age fell from an average of 400,000 during the 1980s to 250,000 in 1999 and then to 51,000 in 2014. Hospitalizations for acute PID in the United States also declined steadily throughout the 1980s and early 1990s but have remained fairly constant at 70,000–100,000 per year since 1995. Important risk factors for acute PID include the presence of endocervical infection or BV, a history of salpingitis or of recent vaginal douching, and recent insertion of an IUD. Certain other iatrogenic factors, such as dilation and curettage or cesarean section, can increase the risk of PID, especially among women with endocervical gonococcal or chlamydial infection or BV. Symptoms of *N. gonorrhoeae*–associated and *C. trachomatis*–associated PID often begin during or soon after the menstrual period; this timing suggests that menstruation is a risk factor for ascending infection from the cervix and vagina. Experimental inoculation of the fallopian tubes of nonhuman primates has shown that repeated exposure to *C. trachomatis* leads to the greatest degree of tissue inflammation and damage; thus, immunopathology probably contributes to the pathogenesis of chlamydial salpingitis. Women using oral contraceptives appear to be at decreased risk of symptomatic PID, and tubal sterilization reduces the risk of salpingitis by preventing intraluminal spread of infection into the tubes.

Clinical Manifestations

ENDOMETRITIS: A CLINICAL PATHOLOGIC SYNDROME

A study of women with clinically suspected PID who were undergoing both endometrial biopsy and laparoscopy showed that those with endometritis alone differed from those who also had salpingitis in significantly less often having lower-quadrant, adnexal, or cervical motion or abdominal rebound tenderness; fever; or elevated C-reactive protein levels. In addition, women with endometritis alone differed from those with neither endometritis nor salpingitis in more often having gonorrhea, chlamydial infection, and risk factors such as douching or IUD use. Thus, women with endometritis alone were intermediate between those with neither endometritis nor salpingitis and those with salpingitis with respect to risk factors, clinical manifestations, cervical infection prevalence, and elevated C-reactive protein level. Women with endometritis alone are at lower risk of subsequent tubal occlusion and resulting infertility than are those with salpingitis.

SALPINGITIS

Symptoms of nontuberculous salpingitis classically evolve from a yellow or malodorous vaginal discharge caused by MPC and/or BV to midline abdominal pain and abnormal vaginal bleeding caused by endometritis and then to bilateral lower abdominal and pelvic pain caused by salpingitis, with nausea, vomiting, and increased abdominal tenderness if peritonitis develops.

The abdominal pain in nontuberculous salpingitis is usually described as dull or aching. In some cases, pain is lacking or atypical, but active inflammatory changes are found in the course of an unrelated evaluation or procedure, such as a laparoscopic evaluation for infertility. Abnormal uterine bleeding precedes or coincides with the onset of pain in ~40% of women with PID, symptoms of urethritis (dysuria) occur in 20%, and symptoms of proctitis (anorectal pain, tenesmus, and rectal discharge or bleeding) are occasionally seen in women with gonococcal or chlamydial infection.

Speculum examination shows evidence of MPC (yellow endocervical discharge, easily induced endocervical bleeding) in the majority of women with gonococcal or chlamydial PID. Cervical motion tenderness is produced by stretching of the adnexal attachments on the side toward which the cervix is pushed. Bimanual examination reveals uterine fundal tenderness due to endometritis and abnormal adnexal tenderness due to salpingitis that is usually, but not necessarily, bilateral. Adnexal swelling is palpable in about one-half of women with acute salpingitis, but evaluation of the adnexae in a patient with marked tenderness is not reliable. The initial temperature is >38°C in only about one-third of patients with acute salpingitis. Laboratory findings include elevation of the erythrocyte sedimentation rate (ESR) in 75% of patients with acute salpingitis and elevation of the peripheral white blood cell count in up to 60%.

Unlike nontuberculous salpingitis, genital tuberculosis often occurs in older women, many of whom are postmenopausal. Presenting symptoms include abnormal vaginal bleeding, pain (including dysmenorrhea), and infertility. About one-quarter of these women have had adnexal masses. Endometrial biopsy shows tuberculous granulomas and provides optimal specimens for culture.

PERIHEPATITIS AND PERIAPPENDICITIS

Pleuritic upper-abdominal pain and tenderness, usually localized to the right upper quadrant (RUQ), develop in 3–10% of women with acute PID. Symptoms of perihepatitis arise during or after the onset of symptoms of PID and may overshadow lower-abdominal symptoms, thereby leading to a mistaken diagnosis of

cholecystitis. In perhaps 5% of cases of acute salpingitis, early laparoscopy reveals perihepatic inflammation ranging from edema and erythema of the liver capsule to exudate with fibrinous adhesions between the visceral and parietal peritoneum. When treatment is delayed and laparoscopy is performed late, dense "violin-string" adhesions can be seen over the liver; chronic exertional or positional RUQ pain ensues when traction is placed on the adhesions. Although perihepatitis, also known as the *Fitz-Hugh–Curtis syndrome*, was for many years specifically attributed to gonococcal salpingitis, most cases are now attributed to chlamydial salpingitis. In patients with chlamydial salpingitis, serum titers of microimmunofluorescent antibody to *C. trachomatis* are typically much higher when perihepatitis is present than when it is absent.

Physical findings include RUQ tenderness and usually include adnexal tenderness and cervicitis, even in patients whose symptoms do not suggest salpingitis. Results of liver function tests and RUQ ultrasonography are nearly always normal. The presence of MPC and pelvic tenderness in a young woman with subacute pleuritic RUQ pain and normal ultrasonography of the gallbladder points to a diagnosis of perihepatitis.

Periappendicitis (appendiceal serositis without involvement of the intestinal mucosa) has been found in ~5% of patients undergoing appendectomy for suspected appendicitis and can occur as a complication of gonococcal or chlamydial salpingitis.

Among women with salpingitis, HIV infection is associated with increased severity of salpingitis and with tuboovarian abscess requiring hospitalization and surgical drainage. Nonetheless, among women with HIV infection and salpingitis, the clinical response to conventional antimicrobial therapy (coupled with drainage of tuboovarian abscess, when found) has usually been satisfactory.

Diagnosis

Treatment appropriate for PID must not be withheld from patients who have an equivocal diagnosis; it is better to err on the side of overdiagnosis and overtreatment. On the other hand, it is essential to differentiate between salpingitis and other pelvic pathology, particularly surgical emergencies such as appendicitis and ectopic pregnancy.

Nothing short of laparoscopy definitively identifies salpingitis, but routine laparoscopy to confirm suspected salpingitis is generally impractical. Most patients with acute PID have lower abdominal pain of <3 weeks' duration, pelvic tenderness on bimanual pelvic examination, and evidence of lower genital tract infection (e.g., MPC). Approximately 60% of such patients have salpingitis at laparoscopy, and perhaps 10–20% have endometritis alone. Among the patients with these findings, a rectal temperature >38°C, a palpable adnexal mass, and elevation of the ESR to >15 mm/h also raise the probability of salpingitis, which has been found at laparoscopy in 68% of patients with one of these additional findings, 90% of patients with two, and 96% of patients with three. However, only 17% of all patients with laparoscopy-confirmed salpingitis have had all three additional findings.

In a woman with pelvic pain and tenderness, increased numbers of PMNs (30 per 1000× microscopic field in strands of cervical mucus) or leukocytes outnumbering epithelial cells in vaginal fluid (in the absence of

trichomonal vaginitis, which also produces PMNs in vaginal discharge) increase the predictive value of a clinical diagnosis of acute PID, as do onset with menses, history of recent abnormal menstrual bleeding, presence of an IUD, history of salpingitis, and sexual exposure to a male with urethritis. Appendicitis or another disorder of the gut is favored by the early onset of anorexia, nausea, or vomiting; the onset of pain later than day 14 of the menstrual cycle; or unilateral pain limited to the right or left lower quadrant. Whenever the diagnosis of PID is being considered, serum assays for human β-chorionic gonadotropin should be performed; these tests are usually positive with ectopic pregnancy. Ultrasonography and MRI can be useful for the identification of tuboovarian or pelvic abscess. MRI of the tubes can also show increased tubal diameter, intratubal fluid, or tubal wall thickening in cases of salpingitis.

The primary value of laparoscopy in women with lower abdominal pain is for the exclusion of other surgical problems that cannot be resolved with non-invasive imaging. Some of the most common or serious problems that may be confused with salpingitis (e.g., acute appendicitis, ectopic pregnancy, corpus luteum bleeding, ovarian tumor) are unilateral. Unilateral pain or pelvic mass, although not incompatible with PID, is a strong indication for laparoscopy unless the clinical picture warrants laparotomy instead. Atypical clinical findings such as the absence of lower genital tract infection, a missed menstrual period, a positive pregnancy test, or failure to respond to appropriate therapy are other common indications for laparoscopy. Endometrial biopsy is relatively sensitive and specific for the diagnosis of endometritis, which correlates well with the presence of salpingitis.

Vaginal or endocervical swab specimens should be obtained for NAATs for *N. gonorrhoeae* and *C. trachomatis*. At a minimum, vaginal fluid should be evaluated for the presence of PMNs, and endocervical secretions ideally should be assessed by Gram's staining for PMNs and gram-negative diplococci, which indicate gonococcal infection. The clinical diagnosis of PID made by expert gynecologists is confirmed by laparoscopy or endometrial biopsy in ~90% of women who also have cultures positive for *N. gonorrhoeae* or *C. trachomatis*. Even among women with no symptoms suggestive of acute PID who were attending an STD clinic or a gynecology clinic in Pittsburgh, endometritis was significantly associated with endocervical gonorrhea or chlamydial infection or with BV, being detected in 26%, 27%, and 15% of women with these conditions, respectively.

TREATMENT

TREATMENT

Pelvic Inflammatory Disease

Recommended combination regimens for ambulatory or parenteral management of PID are presented in Table 131-6. Women managed as outpatients should receive a combined regimen with broad activity, such as ceftriaxone (to cover possible gonococcal infection) followed by doxycycline (to cover possible chlamydial infection). Metronidazole can be added to enhance activity against anaerobes; this addition should be strongly considered if BV is documented. Although few methodologically sound clinical trials (especially with

prolonged follow-up) have been conducted, one meta-analysis suggested a benefit of providing good coverage against anaerobes.

The CDC STD treatment guidelines recommend initiation of empirical treatment for PID in sexually active young women and other women at risk for PID if they are experiencing pelvic or lower abdominal pain, if no other cause for the pain can be identified, and if pelvic examination reveals one or more of the following criteria for PID: cervical motion tenderness, uterine tenderness, or adnexal tenderness. Women with suspected PID can be treated as either outpatients or inpatients. In the multicenter Pelvic Inflammatory Disease Evaluation and Clinical Health (PEACH) trial, 831 women with mild to moderately severe symptoms and signs of PID were randomized to receive either inpatient treatment with IV cefoxitin and doxycycline or outpatient treatment with a single IM dose of cefoxitin plus oral doxycycline. Short-term clinical and microbiologic outcomes and long-term outcomes were equivalent in the two groups. Nonetheless, hospitalization should be considered when (1) the diagnosis is uncertain and surgical emergencies such as appendicitis and ectopic pregnancy cannot be excluded, (2) the patient is pregnant, (3) pelvic abscess is suspected, (4) severe illness or nausea and vomiting preclude outpatient management, (5) the patient has HIV infection, (6) the patient is assessed as unable to follow or tolerate an outpatient regimen, or (7) the patient has failed to respond to outpatient therapy. Some experts also prefer to hospitalize adolescents with PID for initial therapy, although younger women do as well as older women on outpatient therapy.

Currently, no agents other than parenteral cephalosporins provide reliable coverage for gonococcal infection. Thus, adequate oral treatment of women with serious intolerance to cephalosporins is a challenge. If penicillins are an option, amoxicillin/clavulanic acid combined with doxycycline has elicited a short-term clinical response in one trial. Clinical trials performed outside the United States support the effectiveness of oral moxifloxacin. In this case, it is imperative to perform a sensitive diagnostic test for gonorrhea (ideally, a culture to test for antimicrobial susceptibility) before initiation of therapy. For women whose PID involves quinolone-resistant *N. gonorrhoeae*, treatment is uncertain but could include parenteral gentamicin or oral azithromycin, although the latter agent has not been studied for this purpose.

For hospitalized patients, the following two parenteral regimens (Table 131-6) have given nearly identical results in a multicenter randomized trial:

1. Doxycycline plus either cefotetan or cefoxitin: Administration of these drugs should be continued by the IV route for at least 48 h after the patient's condition improves and then followed with oral doxycycline (100 mg twice daily) to complete 14 days of therapy.

2. Clindamycin plus gentamicin in patients with normal renal function: Once-daily administration of gentamicin (with combination of the total daily dose into a single daily dose) has not been evaluated in PID but has been efficacious in other serious infections and could be substituted. Treatment with these drugs should be continued for at least 48 h after the patient's condition improves and then followed with oral doxycycline (100 mg twice daily) or clindamycin (450 mg four times daily) to complete 14 days of therapy. In cases with tuboovarian abscess, clindamycin rather than doxycycline for continued therapy provides better coverage for anaerobic infection.

FOLLOW-UP

Hospitalized patients should show substantial clinical improvement within 3–5 days. Women treated as outpatients should be clinically reevaluated within 72 h. A follow-up telephone survey of women seen in an emergency department and given a prescription for 10 days of oral doxycycline for PID found that 28% never filled the prescription and 41% stopped taking the medication early (after an average of 4.1 days), often because of persistent symptoms, lack of symptoms, or side effects. Women not responding favorably to ambulatory therapy should be hospitalized for parenteral therapy and further diagnostic evaluations, including a consideration of laparoscopy. Male sex partners should be evaluated and treated empirically for gonorrhea and chlamydial infection. After completion of treatment, tests for persistent or recurrent infection with *N. gonorrhoeae* or *C. trachomatis* should be performed if symptoms persist or recur or if the patient has not complied with therapy or has been reexposed to an untreated sex partner.

SURGERY

Surgery is necessary for the treatment of salpingitis only in the face of life-threatening infection (such as rupture or threatened rupture of a tuboovarian abscess) or for drainage of an abscess. Conservative surgical procedures are usually sufficient. Pelvic abscesses can often be drained by posterior colpotomy, and peritoneal lavage can be used for generalized peritonitis.

TABLE 131-6

Combination Antimicrobial Regimens Recommended for Outpatient Treatment or for Parenteral Treatment of Pelvic Inflammatory Disease

Outpatient Regimens[a]	Parenteral Regimens
Ceftriaxone (250 mg IM once) ***plus*** Doxycycline (100 mg PO bid for 14 days) ***plus***[b] Metronidazole (500 mg PO bid for 14 days)	Initiate parenteral therapy with either of the following regimens; continue parenteral therapy until 48 h after clinical improvement; then change to outpatient therapy, as described in the text **Regimen A** Cefotetan (2 g IV q12h) *or* cefoxitin (2 g IV q6h) ***plus*** Doxycycline (100 mg IV or PO q12h) **Regimen B** Clindamycin (900 mg IV q8h) ***plus*** Gentamicin (loading dose of 2 mg/kg IV or IM, then maintenance dose of 1.5 mg/kg q8h)

[a]See text for discussion of options in the patient who is intolerant of cephalosporins. [b]The addition of metronidazole is recommended by some experts, particularly if bacterial vaginosis is present.

Source: Adapted from Centers for Disease Control and Prevention: MMWR Recomm Rep 59(RR-12):1, 2010.

Prognosis

Late sequelae include infertility due to bilateral tubal occlusion, ectopic pregnancy due to tubal scarring without occlusion, chronic pelvic pain, and recurrent salpingitis. The overall post-salpingitis risk of infertility due to tubal occlusion in a large study in Sweden was 11% after one episode of salpingitis, 23% after two episodes, and 54% after three or more episodes. A University of Washington study found a sevenfold increase in the risk of ectopic pregnancy and an eightfold increase in the rate of hysterectomy after PID.

Prevention

A randomized controlled trial designed to determine whether selective screening for chlamydial infection reduces the risk of subsequent PID showed that women randomized to undergo screening had a 56% lower rate of PID over the following year than did women receiving the usual care without screening. This report helped prompt U.S. national guidelines for risk-based chlamydial screening of young women to reduce the incidence of PID and the prevalence of post-PID sequelae, while also reducing sexual transmission of *C. trachomatis*. The CDC and the U.S. Preventive Services Task Force recommend that sexually active women

≤25 years of age be screened annually for genital chlamydial infection. Despite this recommendation, screening coverage in many primary care settings remains low.

ULCERATIVE GENITAL OR PERIANAL LESIONS

Genital ulceration reflects a set of important STIs, most of which sharply increase the risk of sexual acquisition and shedding of HIV. In a 1996 study of genital ulcers in 10 of the U.S. cities with the highest rates of primary syphilis, PCR testing of ulcer specimens demonstrated HSV in 62% of patients, *T. pallidum* in 13%, and *Haemophilus ducreyi* (the cause of chancroid) in 12–20%. Today, genital herpes represents an even higher proportion of genital ulcers in the United States and other industrialized countries.

In Asia and Africa, chancroid **(Fig. 131-5)** was once considered the most common type of genital ulcer, followed in frequency by primary syphilis and then genital herpes **(Fig. 131-6)**. With increased efforts to control chancroid and syphilis and widespread use of broad-spectrum antibiotics to treat STI-related syndromes, together with more frequent recurrences or persistence of genital herpes attributable to HIV infection, PCR testing of genital ulcers now clearly implicates genital herpes as by far the most common cause of genital ulceration in most developing countries. LGV due to *C. trachomatis* **(Fig. 131-7)** and donovanosis (granuloma inguinale, due to *Klebsiella granulomatis*; **see Fig. 168-1**) continue to cause genital ulceration in some developing countries. LGV virtually disappeared in industrialized countries during the first 20 years of the HIV pandemic, but outbreaks are again occurring in Europe (including the United Kingdom), in North America, and in Australia. In these outbreaks, LGV typically presents as proctitis, with or without anal lesions, in men who report unprotected receptive anal intercourse, very often in association with HIV and/or hepatitis C virus infection; the latter may be an acute infection acquired through the same exposure. Other causes of genital ulcers include (1) candidiasis and traumatized genital warts—both readily recognized; (2) lesions due to genital involvement by more widespread dermatoses; (3) cutaneous manifestations of systemic diseases such as genital mucosal ulceration in Stevens-Johnson syndrome or Behçet's disease; (4) superinfections of lesions that may originally have been sexually acquired (for example, methicillin-resistant *S. aureus* complicating a genital ulcer due to HSV-2); and (5) localized drug reactions, such as the ulcers occasionally seen with topical paromomycin cream or boric acid preparations.

FIGURE 131-5

Chancroid: multiple, painful, punched-out ulcers with undermined borders on the labia occurring after autoinoculation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 131-6

Genital herpes. A relatively mild, superficial ulcer is typically seen in episodic outbreaks. *(Courtesy of Michael Remington, University of Washington Virology Research Clinic.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 131-7

Lymphogranuloma venereum (LGV): striking tender lymphadenopathy occurring at the femoral and inguinal lymph nodes, separated by a groove made by Poupart's ligament. This "sign-of-the-groove" is not considered specific for LGV; for example, lymphomas may present with this sign.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Diagnosis

Although most genital ulcerations cannot be diagnosed confidently on clinical grounds alone, clinical findings **(Table 131-7)** and epidemiologic considerations can usually guide initial management **(Table 131-8)** pending results of specific tests. Clinicians should order a rapid serologic test for syphilis in all cases of genital ulcer. To evaluate lesions except those highly characteristic of infection with HSV (i.e., those with herpetic vesicles), dark-field microscopy, direct immunofluorescence, and a NAAT for *T. pallidum* can be useful but are rarely available. It is important to note that 30% of syphilitic chancres—the primary ulcer of syphilis—are associated with an initially nonreactive syphilis serology. All patients presenting with genital ulceration should be counseled and tested for HIV infection.

TABLE 131-7

Clinical Features of Genital Ulcers

Feature	Syphilis	Herpes	Chancroid	Lymphogranuloma Venereum	Donovanosis
Incubation period	9–90 days	2–7 days	1–14 days	3 days–6 weeks	1–4 weeks (up to 6 months)
Early primary lesions	Papule	Vesicle	Pustule	Papule, pustule, or vesicle	Papule
No. of lesions	Usually one	Multiple	Usually multiple, may coalesce	Usually one; often not detected, despite lymphadenopathy	Variable
Diameter	5–15 mm	1–2 mm	Variable	2–10 mm	Variable
Edges	Sharply demarcated, elevated, round, or oval	Erythematous	Undermined, ragged, irregular	Elevated, round, or oval	Elevated, irregular
Depth	Superficial or deep	Superficial	Excavated	Superficial or deep	Elevated
Base	Smooth, nonpurulent, relatively nonvascular	Serous, erythematous, nonvascular	Purulent, bleeds easily	Variable, nonvascular	Red and velvety, bleeds readily
Induration	Firm	None	Soft	Occasionally firm	Firm
Pain	Uncommon	Frequently tender	Usually very tender	Variable	Uncommon

Feature	Syphilis	Herpes	Chancroid	Lymphogranuloma Venereum	Donovanosis
Lymphadenopathy	Firm, nontender, bilateral	Firm, tender, often bilateral with initial episode	Tender, may suppurate, loculated, usually unilateral	Tender, may suppurate, loculated, usually unilateral	None; pseudobuboes

Source: From RM Ballard, in KK Holmes et al (eds): *Sexually Transmitted Diseases*, 4th ed. New York, McGraw-Hill, 2008.

TABLE 131-8

Initial Management of Genital or Perianal Ulcer

Causative Pathogens
HSV *Treponema pallidum* (primary syphilis) *Haemophilus ducreyi* (chancroid)
Usual Initial Laboratory Evaluation
Dark-field examination (if available), direct FA, or PCR for *T. pallidum* RPR, VDRL, or EIA serologic test for syphilis[a] Culture, direct FA, ELISA, or PCR for HSV HSV-2-specific serology (consider) In chancroid-endemic area: PCR or culture for *H. ducreyi*
Initial Treatment
Herpes confirmed or suspected (history or sign of vesicles): Treat for genital herpes with acyclovir, valacyclovir, or famciclovir. **Syphilis confirmed** (dark-field, FA, or PCR showing *T. pallidum*, or RPR reactive): Benzathine penicillin (2.4 million units IM once to patient, to recent [e.g., within 3 months] seronegative partner[s], and to all seropositive partners)[b] **Chancroid confirmed or suspected** (diagnostic test positive, or HSV and syphilis excluded, and persistent lesion): Ciprofloxacin (500 mg PO as single dose) *or* Ceftriaxone (250 mg IM as single dose) *or* Azithromycin (1 g PO as single dose)

[a]If results are negative but primary syphilis is suspected, treat presumptively when indicated by epidemiologic and sexual risk assessment; repeat in 1 week. [b]The same treatment regimen is also effective in HIV-infected persons with early syphilis.

Abbreviations: EIA, enzyme immunoassay; ELISA, enzyme-linked immunosorbent assay; FA, fluorescent antibody; HSV, herpes simplex virus; PCR, polymerase chain reaction; RPR, rapid plasma reagin; VDRL, Venereal Disease Research Laboratory.

Typical vesicles or pustules or a cluster of painful ulcers preceded by vesiculopustular lesions suggest genital herpes. These typical clinical manifestations make detection of the virus optional; however, many patients

want confirmation of the diagnosis, and differentiation of HSV-1 from HSV-2 has prognostic implications, because the latter causes more frequent genital recurrences and is more infectious to vulnerable sex partners.

Painless, nontender, indurated ulcers with firm, nontender inguinal adenopathy suggest primary syphilis. If results of dark-field examination and a rapid serologic test for syphilis are initially negative, or if these tests are not available, presumptive therapy should be provided on the basis of the individual's risk. With historically high rates of syphilis among MSM in the United States, therapy for this infection should not be withheld pending watchful waiting and/or subsequent detection of seroconversion. Repeated serologic testing for syphilis 1 or 2 weeks after treatment of seronegative primary syphilis usually demonstrates seroconversion.

"Atypical" or clinically trivial ulcers may be more common manifestations of genital herpes than classic vesiculopustular lesions. Specific tests for HSV in such lesions are therefore indicated **(Chap. 187)**. Commercially available type-specific serologic tests for serum antibody to HSV-2 may give negative results, especially when patients present early with the initial episode of genital herpes or when HSV-1 is the cause of genital herpes (as is often the case today). Furthermore, a positive test for antibody to HSV-2 does not prove that the current lesions are herpetic, because nearly one-fifth of the general population of the United States (and no doubt a higher proportion of those at risk for other STIs) becomes seropositive for HSV-2 during early adulthood. Although even "type-specific" tests for HSV-2 that are commercially available in the United States are not 100% specific, a positive HSV-2 serology does enable the clinician to tell the patient that he or she has probably had genital herpes, should learn to recognize symptoms, and should avoid sex during recurrences. In addition, because genital shedding and sexual transmission of HSV-2 often occur in the absence of symptoms and signs of recurrent herpetic lesions, persons who have a history of genital herpes or who are seropositive for HSV-2 should consider the use of condoms or suppressive antiviral therapy, both of which can reduce the risk of HSV-2 transmission to a sexual partner.

Demonstration of *H. ducreyi* by culture (or by PCR, where available) is most useful when ulcers are painful and purulent, especially if inguinal lymphadenopathy with fluctuance or overlying erythema is noted; if chancroid is prevalent in the community; or if the patient has recently had a sexual exposure elsewhere in a chancroid-endemic area (e.g., a developing country). Enlarged, fluctuant lymph nodes should be aspirated for culture or PCR to detect *H. ducreyi* as well as for Gram's staining and culture to rule out the presence of other pyogenic bacteria.

When genital ulcers persist beyond the natural history of initial episodes of herpes (2–3 weeks) or of chancroid or syphilis (up to 6 weeks) and do not resolve with syndrome-based antimicrobial therapy, then—in addition to the usual tests for herpes, syphilis, and chancroid—biopsy is indicated to exclude donovanosis as well as carcinoma and other nonvenereal dermatoses.

TREATMENT

TREATMENT

Ulcerative Genital or Perianal Lesions

Immediate syndrome-based treatment for acute genital ulcer (after collection of all necessary diagnostic specimens at the first visit) is often appropriate before all test results become available because patients with typical initial or recurrent episodes of genital or anorectal herpes can benefit from prompt oral antiviral therapy (Chap. 187); because early treatment of sexually transmitted causes of genital ulcers decreases further transmission; and because some patients do not return for test results and treatment. A thorough assessment of the patient's sexual-risk profile and medical history is critical in determining the course of initial management. The patient who has risk factors consistent with exposure to syphilis (e.g., a male patient who reports sex with other men or who has HIV infection) should generally receive initial treatment for syphilis. Empirical therapy for chancroid should be considered if there has been an exposure in an area of the world where chancroid occurs or if regional lymph node suppuration is evident. Finally, empirical antimicrobial therapy may be indicated if ulcers persist and the diagnosis remains unclear after a week of observation despite attempts to diagnose herpes, syphilis, and chancroid.

PROCTITIS, PROCTOCOLITIS, ENTEROCOLITIS, AND ENTERITIS

Sexually acquired *proctitis*, with inflammation limited to the rectal mucosa (the distal 10–12 cm), results from direct rectal inoculation of typical STD pathogens. In contrast, inflammation extending from the rectum to the colon (*proctocolitis*), involving both the small and the large bowel (*enterocolitis*), or involving the small bowel alone (*enteritis*) can result from ingestion of typical intestinal pathogens through oral–anal exposure during sexual contact. Anorectal pain and mucopurulent, bloody rectal discharge suggest proctitis or protocolitis. Proctitis commonly produces tenesmus (causing frequent attempts to defecate, but not true diarrhea) and constipation, whereas proctocolitis and enterocolitis more often cause true diarrhea. In all three conditions, anoscopy usually shows mucosal exudate and easily induced mucosal bleeding (i.e., a positive "wipe test"), sometimes with petechiae or mucosal ulcers. Exudate should be sampled for Gram's staining and other microbiologic studies. Sigmoidoscopy or colonoscopy shows inflammation limited to the rectum in proctitis or disease extending at least up into the sigmoid colon in proctocolitis.

The AIDS era brought an extraordinary shift in the clinical and etiologic spectrum of intestinal infections among MSM. The number of cases of the acute intestinal STIs described above fell as high-risk sexual behaviors became less common in this group. At the same time, the number of AIDS-related opportunistic intestinal infections increased rapidly, many associated with chronic or recurrent symptoms. The incidence of these opportunistic infections has since fallen with increasingly widespread coverage of HIV-infected persons with effective antiretroviral therapy. Two species initially isolated in association with intestinal symptoms in MSM—now known as *Helicobacter cinaedi* and *Helicobacter fennelliae*—have both been isolated from the blood of HIV-infected men and other immunosuppressed persons, often in association with a syndrome of multifocal dermatitis and arthritis.

Acquisition of HSV, *N. gonorrhoeae*, or *C. trachomatis* (including LGV strains of *C. trachomatis*) during receptive anorectal intercourse causes most cases of infectious proctitis in women and MSM. Primary and secondary syphilis can also produce anal or anorectal lesions, with or without symptoms. Gonococcal or

chlamydial proctitis typically involves the most distal rectal mucosa and the anal crypts and is clinically mild, without systemic manifestations. In contrast, primary proctitis due to HSV and proctocolitis due to the strains of *C. trachomatis* that cause LGV usually produce severe anorectal pain and often cause fever. Perianal ulcers and inguinal lymphadenopathy, most commonly due to HSV, can also occur with LGV or syphilis. Sacral nerve root radiculopathies, usually presenting as urinary retention, laxity of the anal sphincter, or constipation, may complicate primary herpetic proctitis. In LGV, rectal biopsy typically shows crypt abscesses, granulomas, and giant cells—findings resembling those in Crohn's disease; such findings should always prompt rectal culture and serology for LGV, which is a curable infection. Syphilis can also produce rectal granulomas, usually in association with infiltration by plasma cells or other mononuclear cells. Syphilis, LGV, and HSV infection involving the rectum can produce perirectal adenopathy that is sometimes mistaken for malignancy; syphilis, LGV, HSV infection, and chancroid involving the anus can produce inguinal adenopathy because anal lymphatics drain to inguinal lymph nodes.

Diarrhea and abdominal bloating or cramping pain without anorectal symptoms and with normal findings on anoscopy and sigmoidoscopy occur with inflammation of the small intestine (enteritis) or with proximal colitis. In MSM without HIV infection, enteritis is often attributable to *Giardia lamblia*. Sexually acquired proctocolitis is most often due to *Campylobacter* or *Shigella* species.

TREATMENT

TREATMENT

Proctitis, Proctocolitis, Enterocolitis, and Enteritis

Acute proctitis in persons who have practiced receptive anorectal intercourse is usually sexually acquired. Such patients should undergo anoscopy to detect rectal ulcers or vesicles and petechiae after swabbing of the rectal mucosa; to examine rectal exudates for PMNs and gram-negative diplococci; and to obtain rectal swab specimens for testing for rectal gonorrhea, chlamydial infection, herpes, and syphilis. Pending test results, patients with proctitis should receive empirical syndromic treatment—e.g., with ceftriaxone (a single IM dose of 250 mg for gonorrhea) plus doxycycline (100 mg by mouth twice daily for 7 days for possible chlamydial infection) plus treatment for herpes or syphilis if indicated. If LGV proctitis is proven or suspected, the recommended treatment is doxycycline (100 mg by mouth twice daily for 21 days); alternatively, 1 g of azithromycin once a week for 3 weeks is likely to be effective but is little studied.

PREVENTION AND CONTROL OF STIs

Prevention and control of STIs require the following:

1. Reduction of the average rate of sexual exposure to STIs through alteration of sexual risk behaviors and behavioral norms among both susceptible and infected persons in all population groups. The necessary changes include reduction in the total number of sexual partners and the number of concurrent sexual partners. The U.S. Preventive Services Task Force recommends intensive behavioral counseling for all

sexually active adolescents and adults who are at increased risk for STIs (grade B recommendation). Motivational interviewing is one approach that has elicited behavioral changes, including safer sex practices and more consistent contraception, that contribute to these goals.

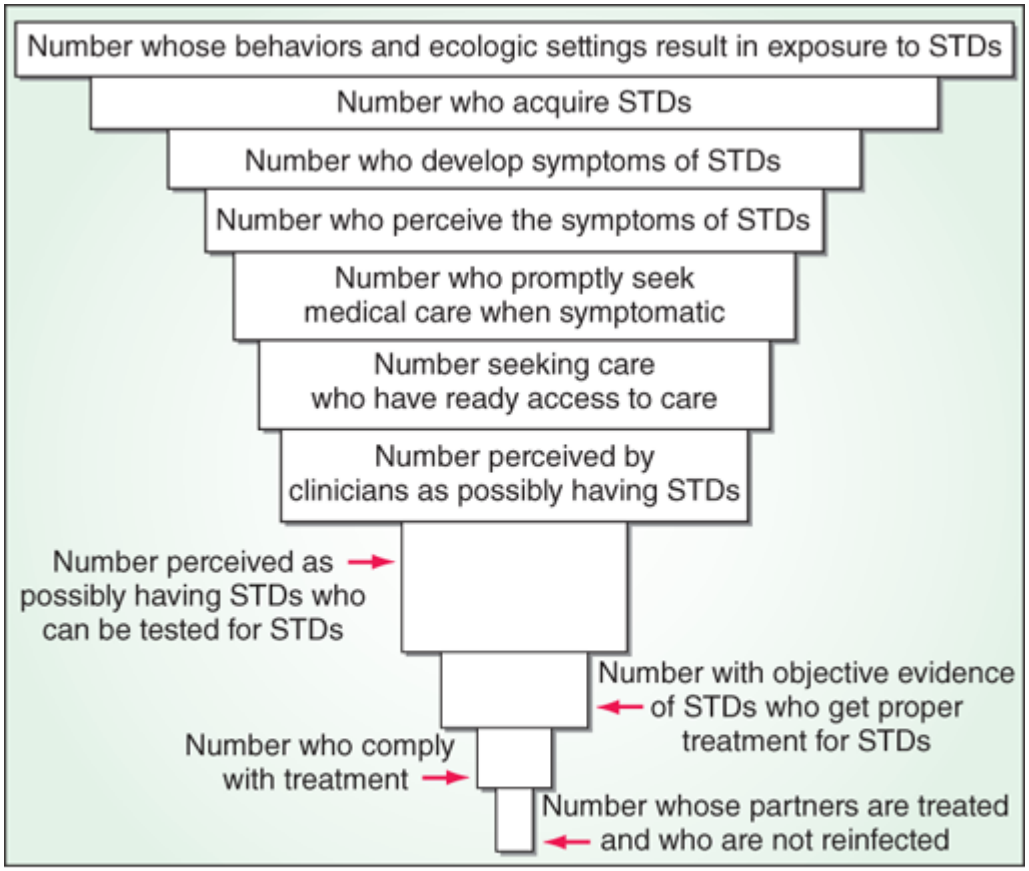
2. Reduction of the efficiency of transmission through the promotion of safer sexual practices, the use of condoms during casual or commercial sex, vaccination against HBV and HPV infection, male circumcision (which reduces risk of acquisition of HIV infection, chancroid, and perhaps other STIs), and a growing number of other approaches (e.g., early detection and treatment of other STIs to reduce the efficiency of sexual transmission of HIV). Longitudinal studies have shown that consistent condom use is associated with significant protection of both males and females against all STIs that have been examined, including HIV, HPV, and HSV infections as well as gonorrhea and chlamydial infection. The only exceptions are probably sexually transmitted *Pthirus pubis* and *Sarcoptes scabiei* infestations.

3. Shortening of the duration of infectivity of STIs through early detection and curative or suppressive treatment of patients and their sexual partners. The availability of curative therapy for hepatitis C virus infection and suppressive therapy for HBV infection exemplifies new opportunities for shortening infectivity in major STIs.

Financial and time constraints imposed by many clinical practices, along with the reluctance of some clinicians to ask questions about stigmatized sexual behaviors, often curtail screening and prevention services. As outlined in **Fig. 131-8**, the success of clinicians' efforts to detect and treat STIs depends in part on societal efforts to teach young people how to recognize symptoms of STIs; to motivate individuals with symptoms to seek care promptly; to educate persons who are at risk but have no symptoms about what tests they should undergo routinely; and to make high-quality, appropriate care accessible, affordable, and acceptable, especially to the young indigent patients most likely to acquire an STI.

FIGURE 131-8

Critical control points for preventive and clinical interventions against sexually transmitted diseases (STDs). *(Adapted from HT Waller and MA Piot: Bull World Health Organ 41:75, 1969 and 43:1, 1970; and from "Resource allocation model for public health planning—a case study of tuberculosis control," Bull World Health Organ 48[Suppl], 1973.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

STI RISK ASSESSMENT

Because many infected individuals develop no symptoms or fail to recognize and report symptoms, clinicians should routinely perform an STI risk assessment for teenagers and young adults as a guide to selective screening. As stated earlier, the U.S. Preventive Services Task Force recommends screening sexually active female patients ≤25 years of age for *C. trachomatis* whenever they present for health care (at least once a year); older women should be tested if they have more than one sexual partner, have begun a new sexual relationship since the previous test, or have another STI diagnosed. In women 25–29 years of age, chlamydial infection is uncommon but still may reach a prevalence of 3–5% in some settings; information provided by women in this age group on a sex partner's concurrency (whether a male partner has had another sex partner during the time they have been together) is helpful in identifying women at increased risk. In some regions of the United States, widespread selective screening and treatment of young women for cervical *C. trachomatis* infection have been associated with a 50–60% drop in prevalence. Such screening and treatment also protect the individual woman from PID. Sensitive urine-based genetic amplification tests permit expansion of screening to men, teenage boys, and girls in settings where examination is not planned or is impractical (e.g., during pre-participation sports examinations or during initial medical evaluation of adolescent girls). Vaginal swabs—collected either by the health care provider at a pelvic examination or by the woman herself—are highly sensitive and specific for the diagnosis of chlamydial and gonococcal infection; they are now the preferred type of specimen for screening and diagnosis of these infections.

Although gonorrhea is now substantially less common than chlamydial infection in industrialized countries, screening tests for *N. gonorrhoeae* are still appropriate for women and teenage girls attending STD clinics and for sexually active teens and young women from areas of high gonorrhea prevalence. Multiplex NAATs that combine screening for *N. gonorrhoeae* and *C. trachomatis*—and, more recently, for *T. vaginalis*—in a single low-cost assay now facilitate the prevention and control of these infections for populations at high risk.

All patients who have newly detected STIs or are at high risk for STIs according to routine risk assessment as well as all pregnant women should be encouraged to undergo serologic testing for syphilis and HIV infection, with appropriate HIV counseling before and after testing. Randomized trials have shown that risk-reduction counseling of patients with STIs significantly lowers subsequent risk of acquiring an STI; such counseling should now be considered a standard component of STI management. Preimmunization serologic testing for antibody to HBV is indicated for unvaccinated persons who are known to be at high risk, such as MSM and people who use injection drugs. In most young persons, however, it is more cost-effective to vaccinate against HBV without serologic screening. It is important to recognize that, while immunization against HBV has contributed to marked reductions in the incidence of infection with this virus, the majority of new cases that do occur are acquired through sex. In 2006, the CDC's Advisory Committee on Immunization Practices (ACIP) recommended the following: (1) Universal hepatitis B vaccination should be implemented for all unvaccinated adults in settings in which a high proportion of adults have risk factors for HBV infection (e.g., STD clinics, HIV testing and treatment facilities, drug-abuse treatment and prevention settings, health care settings targeting services to injection drug users or MSM, and correctional facilities). (2) In other primary care and specialty medical settings that provide care to adults at risk for HBV infection, health care providers should inform all patients about the health benefits of vaccination, the risk factors for HBV infection, and the persons for whom vaccination is recommended; they should vaccinate adults who report risk factors for HBV infection as well as any adult who requests protection from HBV infection. To promote vaccination in all settings, health care providers should implement standing orders to identify adults recommended for hepatitis B vaccination, should administer hepatitis B vaccine as part of routine clinical services, should not require acknowledgment of an HBV infection risk factor for adult vaccination, and should use available reimbursement mechanisms to remove financial barriers to hepatitis B vaccination.

In 2007, the ACIP made its first recommendation for routine immunization of 9- to 26-year-old girls and women with the quadrivalent HPV vaccine (against HPV types 6, 11, 16, and 18). In 2011, the ACIP recommended routine administration of quadrivalent HPV vaccine to boys at 11 or 12 years of age and to males 13–21 years of age who have not yet been vaccinated or who have not completed the three-dose vaccine series; HBV vaccination of men 22–26 years of age has also been recommended. Since that time, a nonavalent HPV vaccine has become available and has largely replaced the earlier vaccines. The optimal age for recommended vaccination is 11–12 years because of the very high risk of HPV infection after sexual debut.

Partner notification is the process of identifying and informing partners of infected patients about possible exposure to an STI and of examining, testing, vaccinating, and treating partners as appropriate. In a series of 22 reports concerning partner notification during the 1990s, index patients with gonorrhea or chlamydial

infection named a mean of 0.75–1.6 partners, of whom one-fourth to one-third were infected; those with syphilis named 1.8–6.3 partners, with one-third to one-half infected; and those with HIV infection named 0.76–5.31 partners, with up to one-fourth infected. Persons who transmit infection or who have recently been infected and are still in the incubation period usually have no symptoms or only mild symptoms and seek medical attention only when notified of their exposure. Therefore, the clinician must encourage patients to participate in partner notification, must ensure that exposed persons are notified and treated, and must guarantee confidentiality to all involved. In the United States, local health departments often offer assistance in partner notification, treatment, and/or counseling. It seems both feasible and most useful to notify those partners exposed within the patient's likely period of infectiousness, which is often considered the preceding 1 month for gonorrhea, 1–2 months for chlamydial infection, and up to 3 months for early syphilis.

Persons with a new-onset STI always have a *source* contact who gave them the infection; in addition, they may have a *secondary* (*spread* or *exposed*) contact with whom they had sex after becoming infected. The identification and treatment of these two types of contacts have different objectives. Treatment of the source contact (often a casual contact) benefits the community by preventing further transmission and benefits the source contact; treatment of the recently exposed secondary contact (typically a spouse or another steady sexual partner) prevents the development of serious complications (such as PID) in the partner, reinfection of the index patient, and further spread of infection. A survey of a random sample of U.S. physicians found that most instructed patients to abstain from sex during treatment, to use condoms, and to inform their sex partners after being diagnosed with gonorrhea, chlamydial infection, or syphilis; physicians sometimes gave the patients drugs for their partners. However, follow-up of the partners by physicians was infrequent. A randomized trial compared patients' delivery of therapy to partners exposed to gonorrhea or chlamydial infection with conventional notification and advice to partners to seek evaluation for STD; patients' delivery of partners' therapy, also known as *expedited partner therapy* (EPT), significantly reduced combined rates of reinfection of the index patient with *N. gonorrhoeae* or *C. trachomatis*. State-by-state variations in regulations governing this approach have not been well defined, but the 2015 CDC STD treatment guidelines describe its potential use. EPT, which is now commonly used by many practicing physicians, is currently permissible in 39 states and potentially allowable in another 8. (Updated information on the legal status of EPT is available at *http://www.cdc.gov/std/ept*.)

In summary, clinicians and public health agencies share responsibility for the prevention and control of STIs. In the current health care environment, the role of primary care clinicians has become increasingly important in STI prevention as well as in diagnosis and treatment, and the resurgence of bacterial STIs like syphilis and LGV among MSM—particularly those co-infected with HIV—emphasizes the need for risk assessment and routine screening.

FURTHER READING

Clement ME et al: Treatment of syphilis: A systematic review. JAMA 312:1905, 2014.

Gottlieb SL et al: The global roadmap for advancing development of vaccines against sexually transmitted infections: Update and next steps. Vaccine 34:2939, 2016.

Johnston C, Corey L: Current concepts for genital herpes simplex virus infection: Diagnostics and pathogenesis of genital tract shedding. Clin Microbiol Rev 29:149, 2016.

Kirkcaldy RD et al: *Neisseria gonorrhoeae* antimicrobial resistance among men who have sex with men and men who have sex exclusively with women: The Gonococcal Isolate Surveillance Project, 2005–2010. Ann Intern Med 158:321, 2013.

Mlisana K et al: Symptomatic vaginal discharge is a poor predictor of sexually transmitted infections and genital tract inflammation in high-risk women in South Africa. J Infect Dis 206:6, 2012.

Newman L et al: Global estimates of the prevalence and incidence of four curable sexually transmitted infections in 2012 based on systematic review and global reporting. PLoS One 10:e0143304, 2015. [PubMed: 26646541]

Price MJ et al: Risk of pelvic inflammatory disease following *Chlamydia trachomatis* infection: Analysis of prospective studies with a multistate model. Am J Epidemiol 178:484, 2013.

Wandeler G et al: Hepatitis C virus infections in the Swiss HIV Cohort Study: A rapidly evolving epidemic. Clin Infect Dis 55:1408, 2012.

Workowski KA, Bolan GA: Sexually transmitted disease treatment guidelines, 2015. MMWR Recomm Rep 64(RR-03):1, 2015. [PubMed: 26042815]

Zakher B et al: Screening for gonorrhea and *Chlamydia*: A systematic review for the US Preventive Services Task Force. Ann Intern Med 161:884, 2014.

Chapter 132: Encephalitis

Karen L. Roos; Kenneth L. Tyler

FIGURE 132-1

DEFINITION

Encephalitis is defined as an inflammation of the brain caused either by infection, usually with a virus, or from a primary autoimmune process. This chapter will focus on infectious causes of encephalitis; non-infectious etiologies are considered elsewhere (**Chaps. 90**, **436**, **and 437**). In contrast to meningitis **(Chaps. 133 and 134)**, in which the infectious process and associated inflammatory response are limited largely to the meninges, in encephalitis the brain parenchyma is also involved. Many patients with encephalitis also have evidence of associated meningitis (meningoencephalitis) and, in some cases, involvement of the spinal cord or nerve roots (encephalomyelitis, encephalomyeloradiculitis).

CLINICAL MANIFESTATIONS

In addition to the acute febrile illness with evidence of meningeal involvement characteristic of meningitis, the patient with encephalitis commonly has an altered level of consciousness (confusion, behavioral abnormalities), or a depressed level of consciousness ranging from mild lethargy to coma, and evidence of either focal or diffuse neurologic signs and symptoms. Patients with encephalitis may have hallucinations, agitation, personality change, behavioral disorders, and, at times, a frankly psychotic state. Focal or generalized seizures occur in many patients with encephalitis. Virtually every possible type of focal neurologic disturbance has been reported in viral encephalitis; the signs and symptoms reflect the sites of infection and inflammation. The most commonly encountered focal findings are aphasia, ataxia, upper or lower motor neuron patterns of weakness, involuntary movements (e.g., myoclonic jerks, tremor), and cranial nerve deficits (e.g., ocular palsies, facial weakness). Involvement of the hypothalamic-pituitary axis may result in temperature dysregulation, diabetes insipidus, or the development of the syndrome of inappropriate secretion of antidiuretic hormone (SIADH). Even though neurotropic viruses typically cause pathologic injury in distinct regions of the central nervous system (CNS), variations in clinical presentations make it impossible to reliably establish the etiology of a specific case of encephalitis on clinical grounds alone (see “Differential Diagnosis,” below).

ETIOLOGY

In the United States, there are an estimated ~20,000 cases of encephalitis per year, although the actual number of cases is likely to be significantly larger. Despite comprehensive diagnostic efforts, the majority of cases of acute encephalitis of suspected viral etiology remain of unknown cause. Hundreds of viruses are capable of causing encephalitis, although only a limited subset is responsible for most cases in which a specific cause is identified **(Table 132-1)**. The most commonly identified viruses causing sporadic cases of acute encephalitis in immunocompetent adults are herpesviruses (herpes simplex virus [HSV], varicella-zoster virus [VZV], Epstein-Barr virus [EBV]). Epidemics of encephalitis are caused by arboviruses, which belong to several different viral taxonomic groups including *Alphaviruses* (e.g., eastern equine encephalitis [EEE] virus), *Flaviviruses* (e.g., West Nile virus [WNV], St. Louis encephalitis virus, Japanese encephalitis virus, Powassan virus), and *Bunyaviruses* (e.g., California encephalitis virus serogroup, La Crosse virus). Historically, the largest number of cases of arbovirus encephalitis in the United States has been due to St. Louis encephalitis virus and the California encephalitis virus serogroup. However, since 2002, WNV has been responsible for the majority of arbovirus meningitis and encephalitis cases in the United States. WNV caused 21,405 confirmed cases of neuroinvasive disease (encephalitis, meningitis, or myelitis) in the years 1995–2016 with 1877 deaths. It is important to recognize that WNV epidemics are unpredictable and that cases have occurred in every state in the continental United States. New causes of viral CNS infections are constantly appearing, as evidenced by the outbreak of cases of encephalitis in Southeast Asia caused by Nipah virus, a member of the Paramyxoviridae family; meningitis in Europe caused by Toscana virus, an arbovirus belonging to the Bunyavirus family; neurological disorders associated with Zika virus, a flavivirus, in South America; and neurologic disorders associated with major epidemics of Chikungunya virus, a togavirus in Africa, India, and Southeast Asia. Parechoviruses including human parechovirus 3 (HPeV3), members of the Picornavirus family, have been reported as causes of fever, sepsis, and meningitis in infants (age <3 months) in the United States and abroad.

TABLE 132-1

Viruses Causing Acute Encephalitis in North America

Common	Less Common
Herpesviruses	Rabies
Cytomegalovirus[a]	Eastern equine encephalitis virus
Herpes simplex virus 1[b]	Powassan virus
Herpes simplex virus 2	Cytomegalovirus[a]
Human herpesvirus 6	Colorado tick fever virus
Varicella-zoster virus	Mumps
Epstein-Barr virus	
Arthropod-borne viruses	
La Crosse virus	
West Nile virus[c]	
St. Louis encephalitis virus	
Zika	
Enteroviruses	

[a]Immunocompromised host. [b]The most common cause of sporadic encephalitis. [c]The most common cause of epidemic encephalitis.

LABORATORY DIAGNOSIS

CSF Examination

Cerebrospinal fluid (CSF) examination should be performed in all patients with suspected viral encephalitis unless contraindicated by the presence of severely increased intracranial pressure (ICP). Ideally at least 20 mL should be collected with 5–10 mL stored frozen for later studies as needed. The characteristic CSF profile is indistinguishable from that of viral meningitis and typically consists of a lymphocytic pleocytosis, a mildly elevated protein concentration, and a normal glucose concentration. A CSF pleocytosis (>5 cells/μL) occurs in >95% of immunocompetent patients with documented viral encephalitis. In rare cases, a pleocytosis may be absent on the initial lumbar puncture (LP) but present on subsequent LPs. Patients who are severely immunocompromised by HIV infection, glucocorticoid or other immunosuppressant drugs, chemotherapy, or lymphoreticular malignancies may fail to mount a CSF inflammatory response. CSF cell counts exceed 500/μL in only about 10% of patients with encephalitis. Infections with certain arboviruses (e.g., EEE virus or California encephalitis virus), mumps, and lymphocytic choriomeningitis virus (LCMV) may occasionally result in cell counts >1000/μL, but this degree of pleocytosis should suggest the possibility of nonviral

infections or other inflammatory processes. Atypical lymphocytes in the CSF may be seen in EBV infection and less commonly with other viruses, including cytomegalovirus (CMV), HSV, and enteroviruses. Increased numbers of plasmacytoid or Mollaret-like large mononuclear cells have been reported in WNV encephalitis. Polymorphonuclear pleocytosis occurs in ~45% of patients with WNV encephalitis and is also a common feature in CMV myeloradiculitis in immunocompromised patients. Large numbers of CSF polymorphonuclear leukocytes may be present in patients with encephalitis due to EEE virus, echovirus 9, and, more rarely, other enteroviruses. However, persisting CSF neutrophilia should prompt consideration of bacterial infection, leptospirosis, amebic infection, and noninfectious processes such as acute hemorrhagic leukoencephalitis. About 20% of patients with encephalitis will have a significant number of red blood cells (>500/μL) in the CSF in a nontraumatic tap. The pathologic correlate of this finding may be a hemorrhagic encephalitis of the type seen with HSV; however, CSF red blood cells occur with similar frequency and in similar numbers in patients with nonherpetic focal encephalitides. A decreased CSF glucose concentration is distinctly unusual in viral encephalitis and should suggest the possibility of bacterial, fungal, tuberculous, parasitic, leptospiral, syphilitic, sarcoid, or neoplastic meningitis. Rare patients with mumps, LCMV, or advanced HSV encephalitis and many patients with CMV myeloradiculitis have low CSF glucose concentrations.

CSF POLYMERASE CHAIN REACTION

CSF PCR has become the primary diagnostic test for CNS infections caused by CMV, EBV, HHV-6, and enteroviruses. In the case of VZV CNS infection, CSF PCR and detection of virus-specific IgM or intrathecal antibody synthesis both provide important aids to diagnosis. The sensitivity and specificity of CSF PCRs vary with the virus being tested. The sensitivity (~96%) and specificity (~99%) of HSV CSF PCR are equivalent to or exceed those of brain biopsy. It is important to recognize that HSV CSF PCR results need to be interpreted after considering the likelihood of disease in the patient being tested, the timing of the test in relationship to onset of symptoms, and the prior use of antiviral therapy. A negative HSV CSF PCR test performed by a qualified laboratory at the appropriate time during illness in a patient with a high likelihood of HSV encephalitis based on clinical and laboratory abnormalities significantly reduces the likelihood of HSV encephalitis but does not exclude it. For example, in a patient with a pretest probability of 35% of having HSV encephalitis, a negative HSV CSF PCR reduces the posttest probability to ~2%, and for a patient with a pretest probability of 60%, a negative test reduces the posttest probability to ~6%. In both situations, a positive test makes the diagnosis almost certain (98–99%). There have been reports of initially negative HSV CSF PCR tests that were obtained early (≤72 h) following symptom onset and that became positive when repeated 1–3 days later. The frequency of positive HSV CSF PCRs in patients with herpes encephalitis also decreases as a function of the duration of illness, with only ~20% of cases remaining positive after ≥14 days. PCR results are generally not affected by ≤1 week of antiviral therapy. In one study, 98% of CSF specimens remained PCR-positive during the first week of antiviral therapy, but the numbers fell to ~50% by 8–14 days and to ~21% by >15 days after initiation of antiviral therapy.

The sensitivity and specificity of CSF PCR tests for viruses other than HSV have not been definitively characterized. Enteroviral (EV) CSF PCR appears to have a sensitivity and specificity of >95%. EV PCR sensitivity for EV71 may be considerably lower (~30% in some reports). Parechoviruses are also not detected

by standard EV RT-PCRs. The specificity of EBV CSF PCR has not been established. Positive EBV CSF PCRs associated with positive tests for other pathogens have been reported and may reflect reactivation of EBV latent in lymphocytes that enter the CNS as a result of an unrelated infectious or inflammatory process. In patients with CNS infection due to VZV, CSF antibody and PCR studies should be considered complementary because patients may have evidence of intrathecal synthesis of VZV-specific antibodies and negative CSF PCRs. In the case of WNV infection, CSF PCR appears to be less sensitive than detection of WNV-specific CSF IgM, although PCR testing remains useful in immunocompromised patients who may not mount an effective anti-WNV antibody response.

Unbiased rapid parallel sequencing technologies capable of identifying infectious genomes in CSF, brain, and other tissues have recently shown great promise for rapid diagnosis of obscure cases of encephalitis and other brain infections.

CSF Culture

CSF culture is generally of limited utility in the diagnosis of acute viral encephalitis. Culture may be insensitive (e.g., >95% of patients with HSV encephalitis have negative CSF cultures as do virtually all patients with EBV-associated CNS disease) and often takes too long to significantly effect immediate therapy.

Serologic Studies and Antigen Detection

For many arboviruses including WNV, serologic studies remain important diagnostic tools. Serum antibody determination is less useful for viruses with high seroprevalence rates in the general population such as HSV, VZV, CMV, and EBV. For viruses with low seroprevalence rates, diagnosis of acute viral infection can be made by documenting seroconversion between acute-phase and convalescent sera (typically obtained after 2–4 weeks) or by demonstrating the presence of virus-specific IgM antibodies. For viruses with high seroprevalence such as VZV and HSV, demonstration of synthesis of virus-specific antibodies in CSF, as shown by an increased IgG index or the presence of CSF IgM antibodies, may be useful and can provide presumptive evidence of CNS infection. Unfortunately, the delay between onset of infection and the host's generation of a virus-specific antibody response often means that serologic data are useful mainly for the retrospective establishment of a specific diagnosis, rather than in aiding acute diagnosis or management.

In patients with HSV encephalitis, antibodies to HSV-1 glycoproteins and HSV glycoprotein antigens have been detected in the CSF. Optimal detection of both HSV antibodies and antigen typically occurs after the first week of illness, limiting the utility of these tests in acute diagnosis. Nonetheless, HSV CSF antibody testing is of value in selected patients whose illness is >1 week in duration and who are CSF PCR–negative for HSV. In the case of VZV infection, CSF antibody tests may be positive when PCR fails to detect viral DNA, and both tests should be considered complementary rather than mutually exclusive.

Demonstration of WNV IgM antibodies is diagnostic of WNV encephalitis because IgM antibodies do not cross the blood-brain barrier, and their presence in CSF is therefore indicative of intrathecal synthesis. Timing of antibody collection may be important because the rate of CSF WNV IgM seropositivity increases by ~10% per

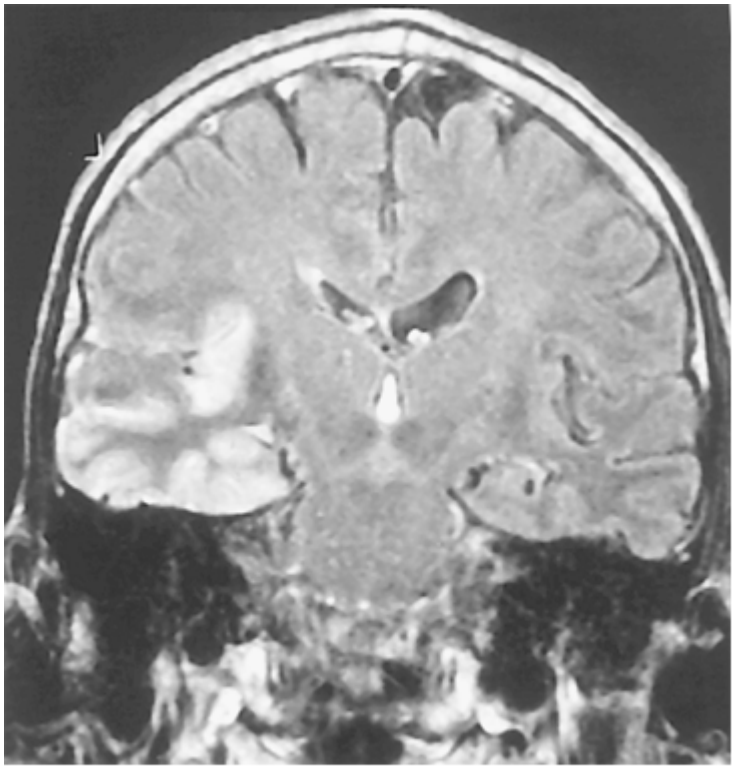
day during the first week after illness onset, reaching 80% or higher on day 7 after symptom onset. Although serum and CSF IgM antibodies generally persist for only a few months after acute infection, there are exceptions to this rule, and WNV serum IgM has been shown to persist in some patients for >1 year following acute infection.

MRI, CT, and EEG

Patients with suspected encephalitis almost invariably undergo neuroimaging studies and often electroencephalogram (EEG). These tests help identify or exclude alternative diagnoses and assist in the differentiation between a focal, as opposed to diffuse, encephalitic process. Focal findings in a patient with encephalitis should always raise the possibility of HSV encephalitis. Examples of focal findings include: (1) areas of increased signal intensity in the frontotemporal, cingulate, or insular regions of the brain on T2-weighted, fluid-attenuated inversion recovery (FLAIR), or diffusion-weighted MRI **(Fig. 132-1)**; (2) focal areas of low absorption, mass effect, and contrast enhancement on CT; or (3) periodic focal temporal lobe spikes on a background of slow or low-amplitude ("flattened") activity on EEG. Approximately 10% of patients with PCR-documented HSV encephalitis will have a normal MRI, although nearly 80% will have abnormalities in the temporal lobe, and an additional 10% in extratemporal regions. The lesions are typically hyperintense on T2-weighted images. The addition of FLAIR and diffusion-weighted images to the standard MRI sequences enhances sensitivity. Children with HSV encephalitis may have atypical patterns of MRI lesions and often show involvement of brain regions outside the frontotemporal areas. CT is less sensitive than MRI and is normal in up to 20–35% of patients. EEG abnormalities occur in >75% of PCR-documented cases of HSV encephalitis; they typically involve the temporal lobes but are often nonspecific. Some patients with HSV encephalitis have a distinctive EEG pattern consisting of periodic, stereotyped, sharp-and-slow complexes originating in one or both temporal lobes and repeating at regular intervals of 2–3 s. The periodic complexes are typically noted between days 2 and 15 of the illness and are present in two-thirds of pathologically proven cases of HSV encephalitis.

FIGURE 132-1

Coronal fluid-attenuated inversion recovery (FLAIR) magnetic resonance image from a patient with herpes simplex encephalitis. Note the area of increased signal in the right temporal lobe (*left side of image*) confined predominantly to the gray matter. This patient had predominantly unilateral disease; bilateral lesions are more common but may be quite asymmetric in their intensity.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Significant MRI abnormalities are found in only approximately two-thirds of patients with WNV encephalitis, a frequency less than that with HSV encephalitis. When present, abnormalities often involve deep brain structures, including the thalamus, basal ganglia, and brainstem, rather than the cortex and may only be apparent on FLAIR images. EEGs in patients with WNV encephalitis typically show generalized slowing that may be more anteriorly prominent rather than the temporally predominant pattern of sharp or periodic discharges more characteristic of HSV encephalitis. Patients with VZV encephalitis may show multifocal areas of hemorrhagic and ischemic infarction, reflecting the tendency of this virus to produce a CNS vasculopathy rather than a true encephalitis. Immunocompromised adult patients with CMV often have enlarged ventricles with areas of increased T2 signal on MRI outlining the ventricles and subependymal enhancement on T1-weighted postcontrast images. **Table 132-2** highlights specific diagnostic test results in encephalitis that can be useful in clinical decision-making.

TABLE 132-2

Use of Diagnostic Tests in Encephalitis

The best test for WNV encephalitis is the *CSF IgM antibody test.* The prevalence of positive CSF IgM tests increases by about 10% per day after illness onset and reaches 70–80% by the end of the first week. Serum WNV IgM can provide evidence for recent WNV infection, but in the absence of other findings does not establish the diagnosis of neuroinvasive disease (meningitis, encephalitis, acute flaccid paralysis). Approximately 80% of patients with proven HSV encephalitis have *MRI* abnormalities involving the temporal lobes. This percentage likely increases to >90% when FLAIR and diffusion-weighted MRI sequences are also used. The absence of temporal lobe lesions on MRI reduces the likelihood of HSV encephalitis and should prompt consideration of other diagnostic possibilities. The *CSF HSV PCR* test may be negative in the first 72 h of symptoms of HSV encephalitis. A repeat study should be considered in patients with an initial early negative PCR in whom diagnostic suspicion of HSV encephalitis remains high and no alternative diagnosis has yet been established. Detection of *intrathecal synthesis* (increased CSF/serum HSV antibody ratio corrected for breakdown of the blood-brain barrier) of *HSV-specific antibody* may be useful in diagnosis of HSV encephalitis in patients in whom only late (>1 week after onset) CSF specimens are available and PCR studies are negative. Serum serology alone is of no value in diagnosis of HSV encephalitis due to the high seroprevalence rate in the general population. Negative *CSF viral cultures* are of no value in excluding the diagnosis of HSV or EBV encephalitis. *VZV CSF IgM antibodies* may be present in patients with a negative VZV CSF PCR. Both tests should be performed in patients with suspected VZV CNS disease. The specificity of *EBV CSF PCR* for diagnosis of CNS infection is unknown. Positive tests may occur in patients with a CSF pleocytosis due to other causes. Detection of EBV CSF IgM or intrathecal synthesis of antibody to VCA supports the diagnosis of EBV encephalitis. Serologic studies consistent with acute EBV infection (e.g., IgM VCA, presence of antibodies against EA but not against EBNA) can help support the diagnosis.

Abbreviations: CNS, central nervous system; CSF, cerebrospinal fluid; DWI, diffusion-weighted imaging; EA, early antigen; EBNA, EBV-associated nuclear antigen; EBV, Epstein-Barr virus; FLAIR, fluid-attenuated inversion recovery; HSV, herpes simplex virus; IgM, immunoglobulin M; MRI, magnetic resonance imaging; PCR, polymerase chain reaction; VCA, viral capsid antibody; VZV, varicella-zoster virus; WNV, West Nile virus.

Brain Biopsy

Brain biopsy is now generally reserved for patients in whom CSF PCR studies fail to lead to a specific diagnosis, who have focal abnormalities on MRI, no serologic evidence of autoimmune disease, and who continue to show progressive clinical deterioration despite treatment with acyclovir and supportive therapy.

DIFFERENTIAL DIAGNOSIS

Infection by a variety of other organisms can mimic viral encephalitis. In studies of biopsy-proven HSV encephalitis, common infectious mimics of focal viral encephalitis included mycobacteria, fungi, rickettsiae, *Listeria, Mycoplasma*, and other bacteria (including *Bartonella* sp.). There are an increasing number of antibodies reported that cause autoimmune encephalitis, including those associated with antibodies against *N*-methyl-D-aspartate (NMDA) receptor, voltage-gated potassium channels/leucine-rich glioma inactivated protein-1 (VGKC/LGI-I), α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA), γ-aminobutyric acid (GABA) receptors, and glutamic acid decarboxylase (GAD 65), that can mimic that caused by viral infection. In most cases, diagnosis is made by detection of the specific autoantibodies in serum and/or CSF. NMDA receptor antibodies have been reported in some patients with HSE encephalitis, and their presence should not exclude appropriate testing and treatment for HSV encephalitis. It has been suggested that development of NMDAR antibodies in patients with HSE may contribute to delayed recovery or clinical relapse. Autoimmune encephalitis may also be associated with specific cancers (paraneoplastic) and onconeuronal antibodies (e.g., anti-Hu, Yo, Ma2, amphiphysin, CRMP5, CV2) **(Chap. 90)**. Subacute or chronic forms of encephalitis may occur in association with autoantibodies against thyroglobulin and thyroperoxidase (Hashimoto's encephalopathy) and with prion diseases.

Infection caused by the ameba *Naegleria fowleri* can also cause acute meningoencephalitis (primary amebic meningoencephalitis), whereas that caused by *Acanthamoeba* and *Balamuthia* more typically produces subacute or chronic granulomatous amebic meningoencephalitis. *Naegleria* thrive in warm, iron-rich pools of water, including those found in drains, canals, and both natural and human-made outdoor pools. Infection has typically occurred in immunocompetent children with a history of swimming in potentially infected water. The CSF, in contrast to the typical profile seen in viral encephalitis, often resembles that of bacterial meningitis with a neutrophilic pleocytosis and hypoglycorrhachia. Motile trophozoites can be seen in a wet mount of warm, fresh CSF. There have been an increasing number of cases of *Balamuthia mandrillaris* amebic encephalitis mimicking acute viral encephalitis in children and immunocompetent adults. This organism has also been associated with encephalitis in recipients of transplanted organs from a donor with unrecognized infection. No effective treatment has been identified, and mortality approaches 100%.

Encephalitis can be caused by the raccoon pinworm *Baylisascaris procyonis*. Clues to the diagnosis include a history of raccoon exposure, especially of playing in or eating dirt potentially contaminated with raccoon feces. Most patients are children, and many have an associated eosinophilia.

Once nonviral causes of encephalitis have been excluded, the major diagnostic challenge is to distinguish HSV from other viruses that cause encephalitis. This distinction is particularly important because in virtually every other instance the therapy is supportive, whereas specific and effective antiviral therapy is available for HSV, and its efficacy is enhanced when it is instituted early in the course of infection. HSV encephalitis should be considered when clinical features suggest involvement of the inferomedial frontotemporal regions of the brain, including prominent olfactory or gustatory hallucinations, anosmia, unusual or bizarre behavior or personality alterations, or memory disturbance. HSV encephalitis should always be suspected in patients with signs and symptoms consistent with acute encephalitis with focal findings on clinical examination, neuroimaging studies, or EEG. The diagnostic procedure of choice in these patients is CSF PCR analysis for

HSV. A positive CSF PCR establishes the diagnosis, and a negative test dramatically reduces the likelihood of HSV encephalitis (see above).

The anatomic distribution of lesions may provide an additional clue to diagnosis. Patients with rapidly progressive encephalitis and prominent brainstem signs, symptoms, or neuroimaging abnormalities may be infected by flaviviruses (WNV, St. Louis encephalitis virus, Japanese encephalitis virus), HSV, enterovirus 71 (EV71), rabies, or *Listeria monocytogenes*. Significant involvement of deep gray matter structures, including the basal ganglia and thalamus, should also suggest possible flavivirus infection. These patients may present clinically with prominent movement disorders (tremor, myoclonus) or parkinsonian features. Patients with WNV infection can also present with a poliomyelitis-like acute flaccid paralysis, as can patients infected with EV71, Enterovirus D-68 (EV-D68) and less commonly, other enteroviruses. Acute flaccid paralysis is characterized by the acute onset of a lower motor neuron type of weakness with flaccid tone, reduced or absent reflexes, and relatively preserved sensation. The complete eradication of polio remains an ongoing challenge despite a continuing World Health Organization poliovirus elimination campaign. There have been small outbreaks of poliomyelitis associated with vaccine strains of virus that have reverted to virulence through mutation or recombination with circulating wild-type enteroviruses in Hispaniola, China, the Philippines, Indonesia, Nigeria, and Madagascar.

Epidemiologic factors may provide important clues to the diagnosis of viral encephalitis. Particular attention should be paid to the season of the year; the geographic location and travel history; and possible exposure to animal bites or scratches, rodents, and ticks. Although transmission from the bite of an infected dog remains the most common cause of rabies worldwide, in the United States very few cases of dog rabies occur, and the most common risk factor is exposure to bats—although a clear history of a bite or scratch is often lacking. The classic clinical presentation of encephalitic (furious) rabies is fever, fluctuating consciousness, and autonomic hyperactivity. Phobic spasms of the larynx, pharynx, neck muscles, and diaphragm can be triggered by attempts to swallow water (*hydrophobia*) or by inspiration (*aerophobia*). Patients may also present with paralytic (dumb) rabies characterized by acute ascending paralysis. Rabies due to the bite of a bat has a different clinical presentation than classic rabies due to a dog or wolf bite. Patients present with focal neurologic deficits, myoclonus, seizures, and hallucinations; phobic spasms are not a typical feature. Patients with rabies have a CSF lymphocytic pleocytosis and may show areas of increased T2 signal abnormality in the brainstem, hippocampus, and hypothalamus. Diagnosis can be made by finding rabies virus antigen in brain tissue or in the neural innervation of hair follicles at the nape of the neck. PCR amplification of viral nucleic acid from CSF and saliva or tears may also enable diagnosis. Serology is frequently negative in both serum and CSF in the first week after onset of infection, which limits its acute diagnostic utility. No specific therapy is available, and cases are almost invariably fatal, with isolated survivors having devastating neurologic sequelae.

State public health authorities provide a valuable resource concerning isolation of particular agents in individual regions. Regular updates concerning the number, type, and distribution of cases of arboviral encephalitis can be found on the CDC and U.S. Geological Survey (USGS) websites (*http://www.cdc.gov* and *http://diseasemaps.usgs.gov*).

TREATMENT

TREATMENT

Viral Encephalitis

Specific antiviral therapy should be initiated when appropriate. Vital functions, including respiration and blood pressure, should be monitored continuously and supported as required. In the initial stages of encephalitis, many patients will require care in an intensive care unit. Basic management and supportive therapy should include careful monitoring of ICP, fluid restriction, avoidance of hypotonic intravenous solutions, and suppression of fever. Seizures should be treated with standard anticonvulsant regimens, and prophylactic therapy should be considered in view of the high frequency of seizures in severe cases of encephalitis. As with all seriously ill, immobilized patients with altered levels of consciousness, encephalitis patients are at risk for aspiration pneumonia, stasis ulcers and decubiti, contractures, deep-venous thrombosis and its complications, and infections of indwelling lines and catheters.

Acyclovir is of benefit in the treatment of HSV and should be started empirically in patients with suspected viral encephalitis, especially if focal features are present, while awaiting viral diagnostic studies. Treatment should be discontinued in patients found not to have HSV encephalitis, with the possible exception of patients with severe encephalitis due to VZV or EBV. HSV, VZV, and EBV all encode an enzyme deoxypyrimidine (thymidine) kinase that phosphorylates acyclovir to produce acyclovir-5′-monophosphate. Host cell enzymes then phosphorylate this compound to form a triphosphate derivative. It is the triphosphate that acts as an antiviral agent by inhibiting viral DNA polymerase and by causing premature termination of nascent viral DNA chains. The specificity of action depends on the fact that uninfected cells do not phosphorylate significant amounts of acyclovir to acyclovir-5′-monophosphate. A second level of specificity is provided by the fact that the acyclovir triphosphate is a more potent inhibitor of viral DNA polymerase than of the analogous host cell enzymes.

Adults should receive a dose of 10 mg/kg of acyclovir intravenously every 8 h (30 mg/kg per day total dose) for 21 days. Neonatal HSV CNS infection is less responsive to acyclovir therapy than HSV encephalitis in adults; it is recommended that neonates with HSV encephalitis receive 20 mg/kg of acyclovir every 8 h (60 mg/kg per day total dose) for a minimum of 21 days.

Prior to intravenous administration, acyclovir should be diluted to a concentration ≤7 mg/mL. (A 70-kg person would receive a dose of 700 mg, which would be diluted in a volume of 100 mL.) Each dose should be infused slowly over 1 h, rather than by rapid or bolus infusion, to minimize the risk of renal dysfunction. Care should be taken to avoid extravasation or intramuscular or subcutaneous administration. The alkaline pH of acyclovir can cause local inflammation and phlebitis (9%). Dose adjustment is required in patients with impaired renal glomerular filtration. Penetration into CSF is excellent, with average drug levels ~50% of serum levels. Complications of therapy include elevations in blood urea nitrogen and creatinine levels (5%), thrombocytopenia (6%), gastrointestinal toxicity (nausea, vomiting, diarrhea) (7%), and neurotoxicity (lethargy or obtundation, disorientation, confusion, agitation, hallucinations, tremors, seizures) (1%).

Acyclovir resistance may be mediated by changes in either the viral deoxypyrimidine kinase or DNA polymerase. To date, acyclovir-resistant isolates have not been a significant clinical problem in immunocompetent individuals. However, there have been reports of clinically virulent acyclovir-resistant HSV isolates from sites outside the CNS in immunocompromised individuals, including those with AIDS.

Oral antiviral drugs with efficacy against HSV, VZV, and EBV, including acyclovir, famciclovir, and valacyclovir, have not been evaluated in the treatment of encephalitis either as primary therapy or as supplemental therapy following completion of a course of parenteral acyclovir. The National Institute of Allergy and Infectious Diseases (NIAID)/National Institute of Neurological Disorders and Stroke–sponsored phase III trial of supplemental oral valacyclovir therapy (2 g tid for 3 months) following the initial 14- to 21-day course of therapy with parenteral acyclovir (*www.clinicaltrials.gov*, identifier NCT00031486) was terminated early due to low enrollment. Although analysis was compromised due to low numbers, no differences were seen in the 12-month endpoints including dementia rating scale, mini-mental state examination, and Glasgow coma score in patients receiving valacyclovir versus placebo. The role of adjunctive intravenous glucocorticoids in treatment of HSV and VZV infection remains unclear. Experimental models and case reports of HSV encephalitis suggest that glucocorticoids may be efficacious and a randomized clinical trial comparing acyclovir alone and acyclovir plus dexamethasone (10 mg every 6 h intravenously for 4 days) in patients with HSV encephalitis is underway in Europe (NCT03084783).

Ganciclovir and foscarnet, either alone or in combination, are often used in the treatment of CMV-related CNS infections, although their efficacy remains unproven. Cidofovir (see below) may provide an alternative in patients who fail to respond to ganciclovir and foscarnet, although data concerning its use in CMV CNS infections are extremely limited.

Ganciclovir is a synthetic nucleoside analogue of 2′-deoxyguanosine. The drug is preferentially phosphorylated by virus-induced cellular kinases. Ganciclovir triphosphate acts as a competitive inhibitor of the CMV DNA polymerase, and its incorporation into nascent viral DNA results in premature chain termination. Following intravenous administration, CSF concentrations of ganciclovir are 25–70% of coincident plasma levels. The usual dose for treatment of severe neurologic illnesses is 5 mg/kg every 12 h given intravenously at a constant rate over 1 h. Induction therapy is followed by maintenance therapy of 5 mg/kg every day for an indefinite period. Induction therapy should be continued until patients show a decline in CSF pleocytosis and a reduction in CSF CMV DNA copy number on quantitative PCR testing (where available). Doses should be adjusted in patients with renal insufficiency. Treatment is often limited by the development of granulocytopenia and thrombocytopenia (20–25%), which may require reduction in or discontinuation of therapy. Gastrointestinal side effects, including nausea, vomiting, diarrhea, and abdominal pain, occur in ~20% of patients. Some patients treated with ganciclovir for CMV retinitis have developed retinal detachment, but the causal relationship to ganciclovir treatment is unclear. Valganciclovir is an orally bioavailable prodrug that can generate high serum levels of ganciclovir, although studies of its efficacy in treating CMV CNS infections are limited.

Foscarnet is a pyrophosphate analogue that inhibits viral DNA polymerases by binding to the pyrophosphate-binding site. Following intravenous infusion, CSF concentrations range from 15 to 100% of coincident plasma

levels. The usual dose for serious CMV-related neurologic illness is 60 mg/kg every 8 h administered by constant infusion over 1 h. Induction therapy for 14–21 days is followed by maintenance therapy (60–120 mg/kg per day). Induction therapy may need to be extended in patients who fail to show a decline in CSF pleocytosis and a reduction in CSF CMV DNA copy number on quantitative PCR tests (where available). Approximately one-third of patients develop renal impairment during treatment, which is reversible following discontinuation of therapy in most, but not all, cases. This is often associated with elevations in serum creatinine and proteinuria and is less frequent in patients who are adequately hydrated. Many patients experience fatigue and nausea. Reductions in serum calcium, magnesium, and potassium occur in ~15% of patients and may be associated with tetany, cardiac rhythm disturbances, or seizures.

Cidofovir is a nucleotide analogue that is effective in treating CMV retinitis and equivalent to or better than ganciclovir in some experimental models of murine CMV encephalitis, although data concerning its efficacy in human CMV CNS disease are limited. The usual dose is 5 mg/kg intravenously once weekly for 2 weeks, then biweekly for two or more additional doses, depending on clinical response. Patients must be prehydrated with normal saline (e.g., 1 L over 1–2 h) prior to each dose and treated with probenecid (e.g., 1 g 3 h before cidofovir and 1 g 2 and 8 h after cidofovir). Nephrotoxicity is common; the dose should be reduced if renal function deteriorates.

Intravenous ribavirin (15–25 mg/kg per day in divided doses given every 8 h) has been reported to be of benefit in isolated cases of severe encephalitis due to California encephalitis (La Crosse) virus. Ribavirin might be of benefit for the rare patients, typically infants or young children, with severe adenovirus or rotavirus encephalitis and in patients with encephalitis due to LCMV or other arenaviruses. However, clinical trials are lacking. Hemolysis, with resulting anemia, has been the major side effect limiting therapy.

No specific antiviral therapy of proven efficacy is currently available for treatment of WNV encephalitis. Patients have been treated with interferon-α, ribavirin, an Israeli IVIg preparation that contains high-titer anti-WNV antibody (Omr-IgG-am) (*www.clinicaltrials.gov*, identifier NCT00069316 and 0068055), and humanized monoclonal antibodies directed against the viral envelope glycoprotein (*www.clinicaltrials.gov*, identifier NCT00927953 and 00515385). WNV chimeric vaccines, in which WNV envelope and premembrane proteins are inserted into the background of another flavivirus, DNA plasmid vaccines and inactivated virus vaccines have all been tested in phase I clinical trials and have been found to be both safe and immunogenic in healthy adults but have not yet been tested for disease prevention in humans (see *www.clinicaltrials.gov).* Both chimeric and killed inactivated WNV vaccines have been found to be safe and effective in preventing equine WNV infection, and effective vaccines are already in human use for prevention of other flavivirus infections including Japanese encephalitis, and yellow fever, suggesting that efficacy testing and commercial considerations rather than scientific issues will be the major impediment to creating a WNV vaccine.

SEQUELAE

There is considerable variation in the incidence and severity of sequelae in patients surviving viral encephalitis. In the case of EEE virus infection, nearly 80% of survivors have severe neurologic sequelae. At

the other extreme are infections due to EBV, California encephalitis virus, and Venezuelan equine encephalitis virus, where severe sequelae are unusual. For example, ~5–15% of children infected with La Crosse virus have a residual seizure disorder, and 1% have persistent hemiparesis. Detailed information about sequelae in patients with HSV encephalitis treated with acyclovir is available from the NIAID-Collaborative Antiviral Study Group (CASG) trials. Of 32 acyclovir-treated patients, 26 survived (81%). Of the 26 survivors, 12 (46%) had no or only minor sequelae, 3 (12%) were moderately impaired (gainfully employed but not functioning at their previous level), and 11 (42%) were severely impaired (requiring continuous supportive care). The incidence and severity of sequelae were directly related to the age of the patient and the level of consciousness at the time of initiation of therapy. Patients with severe neurologic impairment (Glasgow Coma Score 6) at initiation of therapy either died or survived with severe sequelae. Young patients (<30 years) with good neurologic function at initiation of therapy did substantially better (100% survival, 62% with no or mild sequelae) compared with their older counterparts (>30 years; 64% survival, 57% no or mild sequelae). Many patients with WNV infection have sequelae, including cognitive impairment; weakness; and hyper- or hypokinetic movement disorders, including tremor, myoclonus, and parkinsonism. In a large longitudinal study of prognosis in 156 patients with WNV infection, the mean time to achieve recovery (defined as 95% of maximal predicted score on specific validated tests) was 112–148 days for fatigue, 121–175 days for physical function, 131–139 days for mood, and 302–455 days for mental function (the longer interval in each case representing patients with invasive CNS disease).

CHRONIC ENCEPHALITIS

PROGRESSIVE MULTIFOCAL LEUKOENCEPHALOPATHY

Clinical Features and Pathology

Progressive multifocal leukoencephalopathy (PML) is characterized pathologically by multifocal areas of demyelination of varying size distributed throughout the brain but sparing the spinal cord and optic nerves. In addition to demyelination, there are characteristic cytologic alterations in both astrocytes and oligodendrocytes. Astrocytes are enlarged and contain hyperchromatic, deformed, and bizarre nuclei and frequent mitotic figures. Oligodendrocytes have enlarged, densely staining nuclei that contain viral inclusions formed by crystalline arrays of JC virus (JCV) particles. Patients often present with visual deficits (45%), typically a homonymous hemianopia; mental impairment (38%) (dementia, confusion, personality change); weakness, including hemi- or monoparesis; and ataxia. Seizures occur in ~20% of patients, predominantly in those with lesions abutting the cortex.

Almost all patients have an underlying immunosuppressive disorder or are receiving immunomodulatory therapy. In recent series, the most common associated conditions were AIDS (80%), hematologic malignancies (13%), transplant recipients (5%), and chronic inflammatory diseases (2%). It has been estimated that up to 5% of AIDS patients will develop PML. There have been >700 reported cases of PML occurring in patients being treated for multiple sclerosis and inflammatory bowel disease with natalizumab, a humanized monoclonal antibody that inhibits lymphocyte trafficking into CNS and bowel mucosa by

binding to α_4 integrins. Overall risk in these patients has been estimated at ~4.2 PML cases per 1000 treated patients, but the risk depends on a variety of factors including anti-JCV antibody serostatus and the magnitude of the JCV antibody response, prior immunosuppressive therapy use, and duration of natalizumab therapy. Patients who lack detectable JCV antibody have a risk of developing PML of <0.1 case/1000 patients, whereas those who are JCV seropositive and have been exposed to prior immunosuppressive therapy and have received >24 months of natalizumab therapy have a risk of >1.3 case/100 treated patients. PML cases have also been reported in patients receiving other humanized monoclonal antibodies with immunomodulatory activity including efalizumab and rituximab, although the relative risks have not been clearly established. The basic clinical and diagnostic features appear to be similar in HIV-associated PML and PML associated with immunomodulatory drugs with the exception of an increased likelihood of MRI enhancement of PML lesions in immunomodulatory cases. In natalizumab-associated PML, patients will also almost invariably develop clinical and radiographic worsening of lesions with discontinuation of therapy, attributed to development of immune reconstitution inflammatory syndrome (IRIS).

Diagnostic Studies

The diagnosis of PML is frequently suggested by MRI. MRI reveals multifocal asymmetric, coalescing white matter lesions located periventricularly, in the centrum semiovale, in the parietal-occipital region, and in the cerebellum. These lesions have increased signal on T2 and FLAIR images and decreased signal on T1-weighted images. HIV-PML lesions are classically nonenhancing (90%), but patients with immunomodulatory drug-associated PML may have peripheral ring enhancement. PML lesions are not typically associated with edema or mass effect. CT scans, which are less sensitive than MRI for the diagnosis of PML, often show hypodense nonenhancing white matter lesions.

The CSF is typically normal, although mild elevation in protein and/or IgG may be found. Pleocytosis occurs in <25% of cases, is predominantly mononuclear, and rarely exceeds 25 cells/µL. PCR amplification of JCV DNA from CSF has become an important diagnostic tool. The presence of a positive CSF PCR for JCV DNA in association with typical MRI lesions in the appropriate clinical setting is diagnostic of PML, reflecting the assay's relatively high specificity (92–100%); however, sensitivity is variable, and a negative CSF PCR does not exclude the diagnosis. In HIV-negative patients and HIV-positive patients not receiving highly active antiviral therapy (HAART), sensitivity is likely 70–90%. In HAART-treated patients, sensitivity may be closer to 60%, reflecting the lower JCV CSF viral load in this relatively more immunocompetent group. Patients with natalizumab-associated PML have highly variable amounts of JCV DNA in CSF. Some patients may have negative CSF PCRs performed in commercial laboratories where assay detection thresholds are typically >100 JCV DNA copies/µL, but positive results in reference laboratories using supersensitive techniques (detection of 10 JCV copies/µL or less). CSF studies with quantitative JCV PCR indicate that patients with low JCV loads (<100 copies/µL) have a generally better prognosis than those with higher viral loads. Patients with negative CSF PCR studies may require brain biopsy for definitive diagnosis. In biopsy or necropsy specimens of brain, JCV antigen and nucleic acid can be detected by immunocytochemistry, in situ hybridization, or PCR amplification.

Serologic studies are of no utility in diagnosis due to high basal seroprevalence level, but may contribute to risk stratification in patients contemplating therapy with immunomodulatory drugs such as natalizumab.

TREATMENT

TREATMENT

Progressive Multifocal Leukoencephalopathy

No effective therapy for PML is available. There are case reports of potential beneficial effects of the 5-HT_{2a} receptor antagonist mirtazapine, which may inhibit binding of JCV to its receptor on oligodendrocytes. Retrospective non-controlled studies have also suggested a possible beneficial effect of treatment with interferon-α. Neither of these agents has been tested in randomized controlled clinical trials. A prospective multicenter clinical trial to evaluate the efficacy of the antimalarial drug mefloquine failed to show benefit. Intravenous and/or intrathecal cytarabine were not shown to be of benefit in a randomized controlled trial in HIV-associated PML, although some experts suggest that cytarabine may have therapeutic efficacy in situations where breakdown of the blood-brain barrier allows sufficient CSF penetration. A randomized controlled trial of cidofovir in HIV-associated PML also failed to show significant benefit. Because PML almost invariably occurs in immunocompromised individuals, any therapeutic interventions designed to enhance or restore immunocompetence should be considered. Perhaps the most dramatic demonstration of this is disease stabilization and, in rare cases, improvement associated with the improvement in the immune status of HIV-positive patients with AIDS following institution of HAART. In HIV-positive PML patients treated with HAART, 1-year survival is ~50%, although up to 80% of survivors may have significant neurologic sequelae. HIV-positive PML patients with higher CD4 counts (>300/μL) and low or nondetectable HIV viral loads have a better prognosis than those with lower CD4 counts and higher viral loads. Although institution of HAART enhances survival in HIV-positive PML patients, the associated immune reconstitution in patients with an underlying opportunistic infection such as PML may also result in a severe CNS inflammatory syndrome (IRIS) associated with clinical worsening, CSF pleocytosis, and the appearance of new enhancing MRI lesions. Patients receiving natalizumab or other immunomodulatory antibodies who are suspected of having PML should have therapy immediately halted. Removal of drugs with long pharmacokinetic or biological half-lives, such as natalizumab, with plasma exchange or immunoadsorption is frequently utilized, although whether this improves outcomes has not been definitively established. Patients should be closely monitored for development of IRIS, which is generally treated with intravenous glucocorticoids, although controlled clinical trials of efficacy remain lacking.

SUBACUTE SCLEROSING PANENCEPHALITIS (SSPE)

SSPE is a rare, chronic, progressive demyelinating disease of the CNS associated with a chronic nonpermissive infection of brain tissue with measles virus. The frequency has been estimated at 1 in 100,000–500,000 measles cases. An average of five cases per year is reported in the United States. The incidence has declined dramatically since the introduction of a measles vaccine. Most patients give a history of primary measles infection at an early age (2 years), which is followed after a latent interval of 6–8 years by

the development of a progressive neurologic disorder. Some 85% of patients are between 5 and 15 years old at diagnosis. Initial manifestations include poor school performance and mood and personality changes. Typical signs of a CNS viral infection, including fever and headache, do not occur. As the disease progresses, patients develop progressive intellectual deterioration, focal and/or generalized seizures, myoclonus, ataxia, and visual disturbances. In the late stage of the illness, patients are unresponsive, quadriparetic, and spastic, with hyperactive tendon reflexes and extensor plantar responses.

Diagnostic Studies

MRI is often normal early, although areas of increased T2 signal develop in the white matter of the brain and brainstem as disease progresses. The EEG may initially show only nonspecific slowing, but with disease progression, patients develop a characteristic periodic pattern with bursts of high-voltage, sharp, slow waves every 3–8 s, followed by periods of attenuated (“flat”) background. The CSF is acellular with a normal or mildly elevated protein concentration and a markedly elevated gamma globulin level (>20% of total CSF protein). CSF antimeasles antibody levels are invariably elevated, and oligoclonal antimeasles antibodies are often present. Measles virus can be cultured from brain tissue using special cocultivation techniques. Viral antigen can be identified immunocytochemically, and viral genome can be detected by in situ hybridization or PCR amplification.

TREATMENT

TREATMENT

Subacute Sclerosing Panencephalitis

No definitive therapy for SSPE is available. Treatment with isoprinosine (Inosiplex, 100 mg/kg per day), alone or in combination with intrathecal or intraventricular interferon-α, has been reported to prolong survival and produce clinical improvement in some patients but has never been subjected to a controlled clinical trial.

PROGRESSIVE RUBELLA PANENCEPHALITIS

This is an extremely rare disorder that primarily affects males with congenital rubella syndrome, although isolated cases have been reported following childhood rubella. After a latent period of 8–19 years, patients develop progressive neurologic deterioration. The manifestations are similar to those seen in SSPE. CSF shows a mild lymphocytic pleocytosis, slightly elevated protein concentration, markedly increased gamma globulin, and rubella virus–specific oligoclonal bands. No therapy is available. Universal prevention of both congenital and childhood rubella through the use of the available live attenuated rubella vaccine would be expected to eliminate the disease.

FURTHER READING

Beckham JD, Tyler KL: Encephalitis, in *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Disease*, 8th ed, JE Bennett, R Dolin, MJ Blaser (eds). Philadelphia, Elsevier, 2015, pp 1144–1176.

Ramos-Estebanez et al: A systematic review on the role of adjunctive corticosteroids in herpes simplex virus encephalitis: Is timing critical for safety and efficacy? Antivir Ther 19:133, 2014.

Tunkel AR et al: The management of encephalitis: Clinical practice guidelines by the Infectious Diseases Society of America. Clin Infect Dis 47:303, 2008.

Venkatesan A et al: International Encephalitis Consortium. Case definitions, diagnostic algorithms, and priorities in encephalitis: Consensus statement of the International Encephalitis Consortium. Clin Infect Dis 57:1114, 2013.

Wilson MR et al: Acute West Nile virus meningoencephalitis diagnosed via metagenomic deep sequencing of cerebrospinal fluid in a renal transplant patient. Am J Transplant 17:803, 2017.

Chapter 133: Acute Meningitis

Karen L. Roos; Kenneth L. Tyler

FIGURE 133-1

BACTERIAL MENINGITIS

DEFINITION

Bacterial meningitis is an acute purulent infection within the subarachnoid space (SAS). It is associated with a CNS inflammatory reaction that may result in decreased consciousness, seizures, raised intracranial pressure (ICP), and stroke. The meninges, SAS, and brain parenchyma are all frequently involved in the inflammatory reaction (*meningoencephalitis*).

EPIDEMIOLOGY

Bacterial meningitis is the most common form of suppurative CNS infection, with an annual incidence in the United States of >2.5 cases/100,000 population. The organisms most often responsible for community-acquired bacterial meningitis are *Streptococcus pneumoniae* (~50%), *Neisseria meningitidis* (~25%), group B streptococci (~15%), and *Listeria monocytogenes* (~10%). *Haemophilus influenzae* type b accounts for <10% of cases of bacterial meningitis in most series. *N. meningitidis* is the causative organism of recurring epidemics of meningitis every 8–12 years.

ETIOLOGY

S. pneumoniae (Chap. 143) is the most common cause of meningitis in adults >20 years of age, accounting for nearly half the reported cases (1.1 per 100,000 persons per year). There are a number of predisposing conditions that increase the risk of pneumococcal meningitis, the most important of which is pneumococcal pneumonia. Additional risk factors include coexisting acute or chronic pneumococcal sinusitis or otitis media, alcoholism, diabetes, splenectomy, hypogammaglobulinemia, complement deficiency, and head trauma with basilar skull fracture and CSF rhinorrhea. The mortality rate remains ~20% despite antibiotic therapy.

The incidence of meningitis due to *N. meningitidis* (Chap. 150) has decreased with the routine immunization of 11- to 18-year-olds with the quadrivalent (serogroups A, C, W-135, and Y) meningococcal glycoconjugate vaccine. The vaccine does not contain serogroup B, which is responsible for one-third of cases of meningococcal disease. The Advisory Committee on Immunization Practices (ACIP) recommends that

adolescents and young adults aged 16–23 years may be vaccinated with the serogroup B meningococcal (MenB) vaccine. The presence of petechial or purpuric skin lesions can provide an important clue to the diagnosis of meningococcal infection. In some patients the disease is fulminant, progressing to death within hours of symptom onset. Infection may be initiated by nasopharyngeal colonization, which can result in either an asymptomatic carrier state or invasive meningococcal disease. The risk of invasive disease following nasopharyngeal colonization depends on both bacterial virulence factors and host immune defense mechanisms, including the host's capacity to produce antimeningococcal antibodies and to lyse meningococci by both classic and alternative complement pathways. Individuals with deficiencies of any of the complement components, including properdin, are highly susceptible to meningococcal infections.

Gram-negative bacilli cause meningitis in individuals with chronic and debilitating diseases such as diabetes, cirrhosis, or alcoholism and in those with chronic urinary tract infections. Gram-negative meningitis can also complicate neurosurgical procedures, particularly craniotomy, and head trauma associated with CSF rhinorrhea or otorrhea.

Otitis, mastoiditis, and sinusitis are predisposing and associated conditions for meningitis due to *Streptococci* sp., gram-negative anaerobes, *Staphylococcus aureus, Haemophilus* sp., and Enterobacteriaceae. Meningitis complicating endocarditis may be due to viridans streptococci, *S. aureus, Streptococcus bovis,* the HACEK group (*Haemophilus* sp., *Actinobacillus actinomycetemcomitans, Cardiobacterium hominis, Eikenella corrodens, Kingella kingae*), or enterococci.

Group B *Streptococcus*, or *Streptococcus agalactiae*, was previously responsible for meningitis predominantly in neonates, but it has been reported with increasing frequency in individuals aged >50 years, particularly those with underlying diseases.

L. monocytogenes (Chap. 146) is an increasingly important cause of meningitis in neonates (<1 month of age), pregnant women, individuals >60 years, and immunocompromised individuals of all ages. Infection is acquired by ingesting foods contaminated by *Listeria*. Foodborne human listerial infection has been reported from contaminated coleslaw, milk, soft cheeses, and several types of "ready-to-eat" foods, including delicatessen meat and uncooked hotdogs.

The frequency of *H. influenzae* type b (Hib) meningitis in children has declined dramatically since the introduction of the Hib conjugate vaccine, although rare cases of Hib meningitis in vaccinated children have been reported. More frequently, *H. influenzae* causes meningitis in unvaccinated children and older adults, and non-b *H. influenzae* is an emerging pathogen.

S. aureus and coagulase-negative staphylococci (Chap. 142) are important causes of meningitis that occurs following invasive neurosurgical procedures, particularly shunting procedures for hydrocephalus, or as a complication of the use of subcutaneous Ommaya reservoirs for administration of intrathecal chemotherapy.

PATHOPHYSIOLOGY

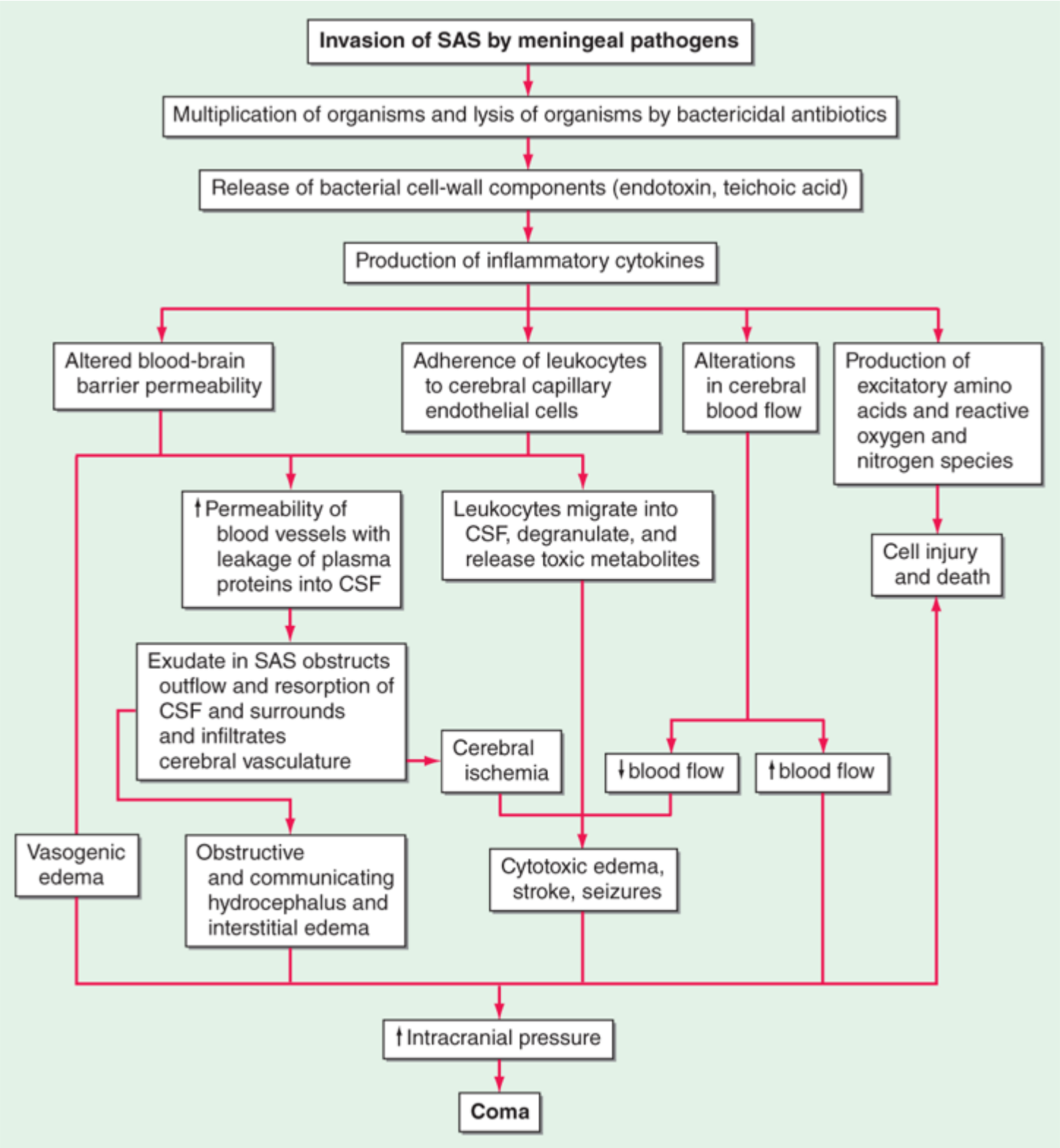
The most common bacteria that cause meningitis, *S. pneumoniae* and *N. meningitidis*, initially colonize the nasopharynx by attaching to nasopharyngeal epithelial cells. Bacteria are transported across epithelial cells in membrane-bound vacuoles to the intravascular space or invade the intravascular space by creating separations in the apical tight junctions of columnar epithelial cells. Once in the bloodstream, bacteria are able to avoid phagocytosis by neutrophils and classic complement-mediated bactericidal activity because of the presence of a polysaccharide capsule. Bloodborne bacteria can reach the intraventricular choroid plexus, directly infect choroid plexus epithelial cells, and gain access to the cerebrospinal fluid (CSF). Some bacteria, such as *S. pneumoniae*, can adhere to cerebral capillary endothelial cells and subsequently migrate through or between these cells to reach the CSF. Bacteria are able to multiply rapidly within CSF because of the absence of effective host immune defenses. Normal CSF contains few white blood cells (WBCs) and relatively small amounts of complement proteins and immunoglobulins. The paucity of the latter two prevents effective opsonization of bacteria, an essential prerequisite for bacterial phagocytosis by neutrophils. Phagocytosis of bacteria is further impaired by the fluid nature of CSF, which is less conducive to phagocytosis than a solid tissue substrate.

A critical event in the pathogenesis of bacterial meningitis is the inflammatory reaction induced by the invading bacteria. Many of the neurologic manifestations and complications of bacterial meningitis result from the immune response to the invading pathogen rather than from direct bacteria-induced tissue injury. As a result, neurologic injury can progress even after the CSF has been sterilized by antibiotic therapy.

The lysis of bacteria with the subsequent release of cell-wall components into the SAS is the initial step in the induction of the inflammatory response and the formation of a purulent exudate in the SAS **(Fig. 133-1)**. Bacterial cell-wall components, such as the lipopolysaccharide (LPS) molecules of gram-negative bacteria and teichoic acid and peptidoglycans of *S. pneumoniae*, induce meningeal inflammation by stimulating the production of inflammatory cytokines and chemokines by microglia, astrocytes, monocytes, microvascular endothelial cells, and CSF leukocytes. In experimental models of meningitis, cytokines including tumor necrosis factor alpha (TNF-α) and interleukin 1β (IL-1β) are present in CSF within 1–2 h of intracisternal inoculation of LPS. This cytokine response is quickly followed by an increase in CSF protein concentration and leukocytosis. Chemokines (cytokines that induce chemotactic migration in leukocytes) and a variety of other proinflammatory cytokines are also produced and secreted by leukocytes and tissue cells that are stimulated by IL-1β and TNF-α. In addition, bacteremia and the inflammatory cytokines induce the production of excitatory amino acids, reactive oxygen and nitrogen species (free oxygen radicals, nitric oxide, and peroxynitrite), and other mediators that can induce death of brain cells, especially in the dentate gyrus of the hippocampus.

FIGURE 133-1

The pathophysiology of the neurologic complications of bacterial meningitis. CSF, cerebrospinal fluid; SAS, subarachnoid space.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Much of the pathophysiology of bacterial meningitis is a direct consequence of elevated levels of CSF cytokines and chemokines. TNF-α and IL-1β act synergistically to increase the permeability of the blood-brain barrier, resulting in induction of vasogenic edema and the leakage of serum proteins into the SAS (Fig. 133-1). The subarachnoid exudate of proteinaceous material and leukocytes obstructs the flow of CSF

through the ventricular system and diminishes the resorptive capacity of the arachnoid granulations in the dural sinuses, leading to obstructive and communicating hydrocephalus and concomitant interstitial edema.

Inflammatory cytokines upregulate the expression of selectins on cerebral capillary endothelial cells and leukocytes, promoting leukocyte adherence to vascular endothelial cells and subsequent migration into the CSF. The adherence of leukocytes to capillary endothelial cells increases the permeability of blood vessels, allowing for the leakage of plasma proteins into the CSF, which adds to the inflammatory exudate. Neutrophil degranulation results in the release of toxic metabolites that contribute to cytotoxic edema, cell injury, and death. Contrary to previous beliefs, CSF leukocytes probably do little to contribute to the clearance of CSF bacterial infection.

During the very early stages of meningitis, there is an increase in cerebral blood flow, soon followed by a decrease in cerebral blood flow and a loss of cerebrovascular autoregulation **(Chap. 301)**. Narrowing of the large arteries at the base of the brain due to encroachment by the purulent exudate in the SAS and infiltration of the arterial wall by inflammatory cells with intimal thickening (*vasculitis*) also occur and may result in ischemia and infarction, obstruction of branches of the middle cerebral artery by thrombosis, thrombosis of the major cerebral venous sinuses, and thrombophlebitis of the cerebral cortical veins. The combination of interstitial, vasogenic, and cytotoxic edema leads to raised ICP and coma. Cerebral herniation usually results from the effects of cerebral edema, either focal or generalized; hydrocephalus and dural sinus or cortical vein thrombosis may also play a role.

CLINICAL PRESENTATION

Meningitis can present as either an acute fulminant illness that progresses rapidly in a few hours or as a subacute infection that progressively worsens over several days. The classic clinical triad of meningitis is fever, headache, and nuchal rigidity, but the classic triad may not be present. A decreased level of consciousness occurs in >75% of patients and can vary from lethargy to coma. Fever and either headache, stiff neck, or an altered level of consciousness will be present in nearly every patient with bacterial meningitis. Nausea, vomiting, and photophobia are also common complaints.

Nuchal rigidity (“stiff neck”) is the pathognomonic sign of meningeal irritation and is present when the neck resists passive flexion. Kernig’s and Brudzinski’s signs are also classic signs of meningeal irritation. *Kernig’s sign* is elicited with the patient in the supine position. The thigh is flexed on the abdomen, with the knee flexed; attempts to passively extend the knee elicit pain when meningeal irritation is present. *Brudzinski’s sign* is elicited with the patient in the supine position and is positive when passive flexion of the neck results in spontaneous flexion of the hips and knees. Although commonly tested on physical examinations, the sensitivity and specificity of Kernig’s and Brudzinski’s signs are uncertain. Both may be absent or reduced in very young or elderly patients, immunocompromised individuals, or patients with a severely depressed mental status. The high prevalence of cervical spine disease in older individuals may result in false-positive tests for nuchal rigidity.

Seizures occur as part of the initial presentation of bacterial meningitis or during the course of the illness in 20–40% of patients. Focal seizures are usually due to focal arterial ischemia or infarction, cortical venous thrombosis with hemorrhage, or focal edema. Generalized seizure activity and status epilepticus may be due to hyponatremia, cerebral anoxia, or, less commonly, the toxic effects of antimicrobial agents.

Raised ICP is an expected complication of bacterial meningitis and the major cause of obtundation and coma in this disease. More than 90% of patients will have a CSF opening pressure >180 mmH_2O, and 20% have opening pressures >400 mmH_2O. Signs of increased ICP include a deteriorating or reduced level of consciousness, papilledema, dilated poorly reactive pupils, sixth nerve palsies, decerebrate posturing, and the Cushing reflex (bradycardia, hypertension, and irregular respirations). The most disastrous complication of increased ICP is cerebral herniation. The incidence of herniation in patients with bacterial meningitis has been reported to occur in as few as 1% to as many as 8% of cases.

Specific clinical features may provide clues to the diagnosis of individual organisms and are discussed in more detail in specific chapters devoted to individual pathogens. The most important of these clues is the rash of meningococcemia, which begins as a diffuse erythematous maculopapular rash resembling a viral exanthem; however, the skin lesions of meningococcemia rapidly become petechial. Petechiae are found on the trunk and lower extremities, in the mucous membranes and conjunctiva, and occasionally on the palms and soles.

DIAGNOSIS

When bacterial meningitis is suspected, blood cultures should be immediately obtained and empirical antimicrobial and adjunctive dexamethasone therapy initiated without delay **(Table 133-1)**. The diagnosis of bacterial meningitis is made by examination of the CSF **(Table 133-2)**. The need to obtain neuroimaging studies (CT or MRI) prior to lumbar puncture (LP) requires clinical judgment. In an immunocompetent patient with no known history of recent head trauma, a normal level of consciousness, and no evidence of papilledema or focal neurologic deficits, it is considered safe to perform LP without prior neuroimaging studies. If LP is delayed in order to obtain neuroimaging studies, empirical antibiotic therapy should be initiated after blood cultures are obtained. Antibiotic therapy initiated a few hours prior to LP will not significantly alter the CSF WBC count or glucose concentration, nor is it likely to prevent visualization of organisms by Gram's stain or detection of bacterial nucleic acid by polymerase chain reaction (PCR) assay.

TABLE 133-1

Antibiotics Used in Empirical Therapy of Bacterial Meningitis and Focal Central Nervous System Infections[a]

Indication	**Antibiotic**	
Preterm infants to infants <1 month	Ampicillin + cefotaxime	
Infants 1–3 months	Ampicillin + cefotaxime or ceftriaxone	
Immunocompetent children >3 months and adults <55	Cefotaxime, ceftriaxone, or cefepime + vancomycin	
Adults >55 and adults of any age with alcoholism or other debilitating illnesses	Ampicillin + cefotaxime, ceftriaxone or cefepime + vancomycin	
Hospital-acquired meningitis, posttraumatic or postneurosurgery meningitis, neutropenic patients, or patients with impaired cell-mediated immunity	Ampicillin + ceftazidime or meropenem + vancomycin	
	Total Daily Dose and Dosing Interval	
Antimicrobial Agent	**Child (>1 month)**	**Adult**
Ampicillin	300 (mg/kg)/d, q6h	12 g/d, q4h
Cefepime	150 (mg/kg)/d, q8h	6 g/d, q8h
Cefotaxime	225–300 (mg/kg)/d, q6h	12 g/d, q4h
Ceftriaxone	100 (mg/kg)/d,	4 g/d, q12h

	q12h	
Ceftazidime	150 (mg/kg)/d, q8h	6 g/d, q8h
Gentamicin	7.5 (mg/kg)/d, q8h[b]	7.5 (mg/kg)/d, q8h
Meropenem	120 (mg/kg)/d, q8h	6 g/d, q8h
Metronidazole	30 (mg/kg)/d, q6h	1500–2000 mg/d, q6h
Nafcillin	100–200 (mg/kg)/d, q6h	9–12 g/d, q4h
Penicillin G	400,000 (U/kg)/d, q4h	20–24 million U/d, q4h
Vancomycin	45–60 (mg/kg)/d, q6h	45–60 (mg/kg)d, q6–12h[b]

[a]All antibiotics are administered intravenously; doses indicated assume normal renal and hepatic function. [b]Doses should be adjusted based on serum peak and trough levels: gentamicin therapeutic level: peak: 5–8 µg/mL; trough: <2 µg/mL; vancomycin therapeutic level: peak: 25–40 µg/mL; trough: 5–15 µg/mL.

TABLE 133-2

Cerebrospinal Fluid (CSF) Abnormalities in Bacterial Meningitis

Opening pressure	>180 mmH_2O
White blood cells	10/μL to 10,000/μL; neutrophils predominate
Red blood cells	Absent in nontraumatic tap
Glucose	<2.2 mmol/L (<40 mg/dL)
CSF/serum glucose	<0.4
Protein	>0.45 g/L (>45 mg/dL)
Gram's stain	Positive in >60%
Culture	Positive in >80%
PCR	Detects bacterial DNA
Latex agglutination	May be positive in patients with meningitis due to *Streptococcus pneumoniae, Neisseria meningitidis, Haemophilus influenzae* type b, *Escherichia coli*, group B streptococci
Limulus lysate	Positive in cases of gram-negative meningitis

Abbreviation: PCR, polymerase chain reaction.

The classic CSF abnormalities in bacterial meningitis (Table 133-2) are (1) polymorphonuclear (PMN) leukocytosis (>100 cells/μL in 90%), (2) decreased glucose concentration (<2.2 mmol/L [<40 mg/dL] and/or CSF/serum glucose ratio of <0.4 in ~60%), (3) increased protein concentration (>0.45 g/L [>45 mg/dL] in 90%), and (4) increased opening pressure (>180 mmH_2O in 90%). CSF bacterial cultures are positive in >80% of patients, and CSF Gram's stain demonstrates organisms in >60%.

CSF glucose concentrations <2.2 mmol/L (<40 mg/dL) are abnormal, and a CSF glucose concentration of zero can be seen in bacterial meningitis. Use of the CSF/serum glucose ratio corrects for hyperglycemia that may

mask a relative decrease in the CSF glucose concentration. The CSF glucose concentration is low when the CSF/serum glucose ratio is <0.6. A CSF/serum glucose ratio <0.4 is highly suggestive of bacterial meningitis but may also be seen in other conditions, including fungal, tuberculous, and carcinomatous meningitis. It takes from 30 min to several hours for the concentration of CSF glucose to reach equilibrium with blood glucose levels; therefore, administration of 50 mL of 50% glucose (D50) prior to LP, as commonly occurs in emergency room settings, is unlikely to alter CSF glucose concentration significantly unless more than a few hours have elapsed between glucose administration and LP.

There are a number of "CSF pathogen panels" available that use specific bacterial primers to detect the nucleic acid of *S. pneumoniae, N. meningitidis, Escherichia coli, L. monocytogenes, H. influenzae*, and *S. agalactiae* (Group B streptococci). The latex agglutination (LA) test for the detection of bacterial antigens of *S. pneumoniae, N. meningitidis, H. influenzae* type b, group B *Streptococcus*, and *E. coli* K1 strains in the CSF has been useful for making a diagnosis of bacterial meningitis but is being replaced by the CSF bacterial PCR assay. The CSF LA test has a *specificity* of 95–100% for *S. pneumoniae* and *N. meningitidis*, so a positive test is virtually diagnostic of bacterial meningitis caused by these organisms. However, the *sensitivity* of the CSF LA test is only 70–100% for detection of *S. pneumoniae* and 33–70% for detection of *N. meningitidis* antigens, so a negative test does not exclude infection by these organisms. The Limulus amebocyte lysate assay is a rapid diagnostic test for the detection of gram-negative endotoxin in CSF and thus for making a diagnosis of gram-negative bacterial meningitis. The test has a specificity of 85–100% and a sensitivity approaching 100%. Thus, a positive Limulus amebocyte lysate assay occurs in virtually all patients with gram-negative bacterial meningitis, but false positives may occur.

Almost all patients with bacterial meningitis will have neuroimaging studies performed during the course of their illness. MRI is preferred over CT because of its superiority in demonstrating areas of cerebral edema and ischemia. In patients with bacterial meningitis, diffuse meningeal enhancement is often seen after the administration of gadolinium. Meningeal enhancement is not diagnostic of meningitis but occurs in any CNS disease associated with increased blood-brain barrier permeability.

Petechial skin lesions, if present, should be biopsied. The rash of meningococcemia results from the dermal seeding of organisms with vascular endothelial damage, and biopsy may reveal the organism on Gram's stain.

DIFFERENTIAL DIAGNOSIS

Viral meningoencephalitis, and particularly herpes simplex virus (HSV) encephalitis, can mimic the clinical presentation of bacterial meningitis (encephalitis). HSV encephalitis typically presents with headache, fever, altered consciousness, focal neurologic deficits (e.g., dysphasia, hemiparesis), and focal or generalized seizures. The findings on CSF studies, neuroimaging, and electroencephalogram (EEG) distinguish HSV encephalitis from bacterial meningitis. The typical CSF profile with viral CNS infections is a lymphocytic pleocytosis with a normal glucose concentration, in contrast to the PMN pleocytosis and hypoglycorrhachia characteristic of bacterial meningitis. The CSF HSV PCR has a 96% sensitivity and a 99% specificity when CSF is examined 72 h following symptom onset, and in the first week of antiviral therapy. MRI abnormalities

(other than meningeal enhancement) are not seen in uncomplicated bacterial meningitis. By contrast, in HSV encephalitis, on T2-weighted, fluid-attenuated inversion recovery (FLAIR) and diffusion-weighted MRI images, high signal intensity lesions are seen in the orbitofrontal, anterior, and medial temporal lobes in the majority of patients within 48 h of symptom onset. Some patients with HSV encephalitis have a distinctive periodic pattern on EEG.

Rickettsial disease can resemble bacterial meningitis **(Chap. 182)**. Rocky Mountain spotted fever (RMSF) is transmitted by a tick bite and caused by the bacteria *Rickettsia rickettsii*. The disease may present acutely with high fever, prostration, myalgia, headache, nausea, and vomiting. Most patients develop a characteristic rash within 96 h of the onset of symptoms. The rash is initially a diffuse erythematous maculopapular rash that may be difficult to distinguish from that of meningococcemia. It progresses to a petechial rash, then to a purpuric rash, and if untreated, to skin necrosis or gangrene. The color of the lesions changes from bright red to very dark red, then yellowish-green to black. The rash typically begins in the wrist and ankles and then spreads distally and proximally within a matter of a few hours, involving the palms and soles. Diagnosis is made by immunofluorescent staining of skin biopsy specimens. Ehrlichioses are also transmitted by a tick bite. These are small gram-negative coccobacilli of which two species cause human disease. *Anaplasma phagocytophilum* causes human granulocytic ehrlichiosis (anaplasmosis), and *Ehrlichia chaffeensis* causes human monocytic ehrlichiosis. The clinical and laboratory manifestations of the infections are similar. Patients present with fever, headache, confusion, nausea, and vomiting. Twenty percent of patients have a maculopapular or petechial rash. There is laboratory evidence of leukopenia, thrombocytopenia, and anemia, and mild to moderate elevations in alanine aminotransferases, alkaline phosphatase, and lactate dehydrogenase. Patients with RMSF and those with ehrlichial infections may have an altered level of consciousness ranging from mild lethargy to coma, confusion, focal neurologic signs, cranial nerve palsies, hyperreflexia, and seizures.

Focal suppurative CNS infections, including subdural and epidural empyema and brain abscess, should also be considered, especially when focal neurologic findings are present. MRI should be performed promptly in all patients with suspected meningitis who have focal features, both to detect the intracranial infection and to search for associated areas of infection in the sinuses or mastoid bones.

A number of noninfectious CNS disorders can mimic bacterial meningitis. Subarachnoid hemorrhage (SAH; **Chap. 301**) is generally the major consideration. Other possibilities include medication-induced hypersensitivity meningitis; chemical meningitis due to rupture of tumor contents into the CSF (e.g., from a cystic glioma or craniopharyngioma epidermoid or dermoid cyst); carcinomatous or lymphomatous meningitis; meningitis associated with inflammatory disorders such as sarcoid, systemic lupus erythematosus (SLE), and Behçet's syndrome; pituitary apoplexy; and uveomeningitic syndromes (Vogt-Koyanagi-Harada syndrome).

On occasion, subacutely evolving meningitis **(Chap. 134)** may be considered in the differential diagnosis of acute meningitis. The principal causes include *Mycobacterium tuberculosis* **(Chap. 173)**, *Cryptococcus neoformans* **(Chap. 210)**, *Histoplasma capsulatum* **(Chap. 207)**, *Coccidioides immitis* **(Chap. 208)**, and *Treponema pallidum* **(Chap. 177)**.

TREATMENT

TREATMENT

Acute Bacterial Meningitis

EMPIRICAL ANTIMICROBIAL THERAPY

(Table 133-1) Bacterial meningitis is a medical emergency. The goal is to begin antibiotic therapy within 60 min of a patient's arrival in the emergency room. Empirical antimicrobial therapy is initiated in patients with suspected bacterial meningitis before the results of CSF Gram's stain and culture are known. *S. pneumoniae* **(Chap. 141)** and *N. meningitidis* **(Chap. 150)** are the most common etiologic organisms of community-acquired bacterial meningitis. Due to the emergence of penicillin- and cephalosporin-resistant *S. pneumoniae*, empirical therapy of community-acquired suspected bacterial meningitis in children and adults should include a combination of dexamethasone, a third- or fourth-generation cephalosporin (e.g., ceftriaxone, cefotaxime, or cefepime), and vancomycin, plus acyclovir, as HSV encephalitis is the leading disease in the differential diagnosis, and doxycycline during tick season to treat tick-borne bacterial infections. Ceftriaxone or cefotaxime provides good coverage for susceptible *S. pneumoniae*, group B streptococci, and *H. influenzae* and adequate coverage for *N. meningitidis*. Cefepime is a broad-spectrum fourth-generation cephalosporin with in vitro activity similar to that of cefotaxime or ceftriaxone against *S. pneumoniae* and *N. meningitidis* and greater activity against *Enterobacter* species and *Pseudomonas aeruginosa*. In clinical trials, cefepime has been demonstrated to be equivalent to cefotaxime in the treatment of penicillin-sensitive pneumococcal and meningococcal meningitis, and it has been used successfully in some patients with meningitis due to *Enterobacter* species and *P. aeruginosa*. Ampicillin should be added to the empirical regimen for coverage of *L. monocytogenes* in individuals <3 months of age, those >55, or those with suspected impaired cell-mediated immunity because of chronic illness, organ transplantation, pregnancy, malignancy, or immunosuppressive therapy. Metronidazole is added to the empirical regimen to cover gram-negative anaerobes in patients with otitis, sinusitis, or mastoiditis. In hospital-acquired meningitis, and particularly meningitis following neurosurgical procedures, staphylococci and gram-negative organisms including *P. aeruginosa* are the most common etiologic organisms. In these patients, empirical therapy should include a combination of vancomycin and ceftazidime or meropenem. Ceftazidime or meropenem should be substituted for ceftriaxone or cefotaxime in neurosurgical patients and in neutropenic patients, because ceftriaxone and cefotaxime do not provide adequate activity against CNS infection with *P. aeruginosa*. Meropenem is a carbapenem antibiotic that is highly active in vitro against *L. monocytogenes*, has been demonstrated to be effective in cases of meningitis caused by *P. aeruginosa*, and shows good activity against penicillin-resistant pneumococci. In experimental pneumococcal meningitis, meropenem was comparable to ceftriaxone and inferior to vancomycin in sterilizing CSF cultures. The number of patients with bacterial meningitis enrolled in clinical trials of meropenem has not been sufficient to definitively assess the efficacy of this antibiotic.

SPECIFIC ANTIMICROBIAL THERAPY

Meningococcal Meningitis (Table 133-3)

Although ceftriaxone and cefotaxime provide adequate empirical coverage for *N. meningitidis*, penicillin G remains the antibiotic of choice for meningococcal meningitis caused by susceptible strains. Isolates of *N. meningitidis* with moderate resistance to penicillin have been identified and are increasing in incidence worldwide. CSF isolates of *N. meningitidis* should be tested for penicillin and ampicillin susceptibility, and if resistance is found, cefotaxime or ceftriaxone should be substituted for penicillin. A 7-day course of intravenous antibiotic therapy is adequate for uncomplicated meningococcal meningitis. The index case and all close contacts should receive chemoprophylaxis with a 2-day regimen of rifampin (600 mg every 12 h for 2 days in adults and 10 mg/kg every 12 h for 2 days in children >1 year). Rifampin is not recommended in pregnant women. Alternatively, adults can be treated with one dose of azithromycin (500 mg) or one intramuscular dose of ceftriaxone (250 mg). Close contacts are defined as those individuals who have had contact with oropharyngeal secretions, either through kissing or by sharing toys, beverages, or cigarettes.

TABLE 133-3

Antimicrobial Therapy of Central Nervous System Bacterial Infections Based on Pathogen[a]

Organism	Antibiotic
Neisseria meningitides	
Penicillin-sensitive	Penicillin G or ampicillin
Penicillin-resistant	Ceftriaxone or cefotaxime
Streptococcus pneumoniae	
Penicillin-sensitive	Penicillin G
Penicillin-intermediate	Ceftriaxone or cefotaxime or cefepime
Penicillin-resistant	Ceftriaxone (or cefotaxime or cefepime) + vancomycin
Gram-negative bacilli (except *Pseudomonas* spp.)	Ceftriaxone or cefotaxime
Pseudomonas aeruginosa	Ceftazidime or meropenem
Staphylococci spp.	
Methicillin-sensitive	Nafcillin
Methicillin-resistant	Vancomycin
Listeria monocytogenes	Ampicillin + gentamicin
Haemophilus influenzae	Ceftriaxone or cefotaxime or cefepime
Streptococcus agalactiae	Penicillin G or ampicillin
Bacteroides fragilis	Metronidazole
Fusobacterium spp.	Metronidazole

[a]Doses are as indicated in Table 133-1.

Pneumococcal Meningitis

Antimicrobial therapy of pneumococcal meningitis is initiated with a cephalosporin (ceftriaxone, cefotaxime, or cefepime) and vancomycin. All CSF isolates of *S. pneumoniae* should be tested for sensitivity to penicillin and the cephalosporins. Once the results of antimicrobial susceptibility tests are known, therapy can be modified accordingly (Table 133-3). For *S. pneumoniae* meningitis, an isolate of *S. pneumoniae* is considered to be susceptible to penicillin with a minimal inhibitory concentration (MIC) <0.06 μg/mL and to be resistant when the MIC is >0.12 μg/mL. Isolates of *S. pneumoniae* that have cephalosporin MICs ≤0.5 μg/mL are considered sensitive to the cephalosporins (cefotaxime, ceftriaxone, cefepime). Those with MICs of 1 μg/mL are considered to have intermediate resistance, and those with MICs ≥2 μg/mL are considered resistant. For meningitis due to pneumococci, with cefotaxime or ceftriaxone MICs ≤0.5 μg/mL, treatment with cefotaxime or ceftriaxone is usually adequate. For MIC >1 μg/mL, vancomycin is the antibiotic of choice. Rifampin can be added to vancomycin for its synergistic effect but is inadequate as monotherapy because resistance develops rapidly when it is used alone.

A 2-week course of intravenous antimicrobial therapy is recommended for pneumococcal meningitis.

Patients with *S. pneumoniae* meningitis should have a repeat LP performed 24–36 h after the initiation of antimicrobial therapy to document sterilization of the CSF. Failure to sterilize the CSF after 24–36 h of antibiotic therapy should be considered presumptive evidence of antibiotic resistance. Patients with penicillin- and cephalosporin-resistant strains of *S. pneumoniae* who do not respond to intravenous vancomycin alone may benefit from the addition of intraventricular vancomycin. The intraventricular route of administration is preferred over the intrathecal route because adequate concentrations of vancomycin in the cerebral ventricles are not always achieved with intrathecal administration.

Listeria Meningitis

Meningitis due to *L. monocytogenes* is treated with ampicillin for at least 3 weeks (Table 133-3). Gentamicin is added in critically ill patients (2 mg/kg loading dose, then 7.5 mg/kg per day given every 8 h and adjusted for serum levels and renal function). The combination of trimethoprim (10–20 mg/kg per day) and sulfamethoxazole (50–100 mg/kg per day) given every 6 h may provide an alternative in penicillin-allergic patients.

Staphylococcal Meningitis

Meningitis due to susceptible strains of *S. aureus* or coagulase-negative staphylococci is treated with nafcillin (Table 133-3). Vancomycin is the drug of choice for methicillin-resistant staphylococci and for patients allergic to penicillin. In these patients, the CSF should be monitored during therapy. If the CSF is not sterilized after 48 h of intravenous vancomycin therapy, then either intraventricular or intrathecal vancomycin, 20 mg once daily, can be added.

Gram-Negative Bacillary Meningitis

The third-generation cephalosporins—cefotaxime, ceftriaxone, and ceftazidime—are equally efficacious for the treatment of gram-negative bacillary meningitis, with the exception of meningitis due to *P. aeruginosa*, which should be treated with ceftazidime or meropenem (Table 133-3). A 3-week course of intravenous antibiotic therapy is recommended for meningitis due to gram-negative bacilli.

ADJUNCTIVE THERAPY

The release of bacterial cell-wall components by bactericidal antibiotics leads to the production of the inflammatory cytokines IL-1β and TNF-α in the SAS. Dexamethasone exerts its beneficial effect by inhibiting the synthesis of IL-1β and TNF-α at the level of mRNA, decreasing CSF outflow resistance, and stabilizing the blood-brain barrier. The rationale for giving dexamethasone 20 min before antibiotic therapy is that dexamethasone inhibits the production of TNF-α by macrophages and microglia only if it is administered before these cells are activated by endotoxin. Dexamethasone does not alter TNF-α production once it has been induced. The results of clinical trials of dexamethasone therapy in meningitis due to *H. influenzae, S. pneumoniae*, and *N. meningitidis* have demonstrated its efficacy in decreasing meningeal inflammation and neurologic sequelae such as the incidence of sensorineural hearing loss.

A prospective European trial of adjunctive therapy for acute bacterial meningitis in 301 adults found that dexamethasone reduced the number of unfavorable outcomes (15 vs 25%, $p = .03$) including death (7 vs 15%, $p = .04$). The benefits were most striking in patients with pneumococcal meningitis. Dexamethasone (10 mg intravenously) was administered 15–20 min before the first dose of an antimicrobial agent, and the same dose was repeated every 6 h for 4 days. These results were confirmed in a second trial of dexamethasone in adults with pneumococcal meningitis. Therapy with dexamethasone should ideally be started 20 min before, or not later than concurrent with, the first dose of antibiotics. It is unlikely to be of significant benefit if started >6 h after antimicrobial therapy has been initiated. Dexamethasone may decrease the penetration of vancomycin into CSF, and it delays the sterilization of CSF in experimental models of *S. pneumoniae* meningitis. As a result, to assure reliable penetration of vancomycin into the CSF, children and adults are treated with vancomycin in a dose of 45–60 mg/kg per day. Alternatively, vancomycin can be administered by the intraventricular route. In clinical trials, dexamethasone has also been shown to reduce rates of death and hearing loss with no adverse effects in patients with meningococcal meningitis.

One of the concerns for using dexamethasone in adults with bacterial meningitis is that in experimental models of meningitis, dexamethasone therapy increased hippocampal cell injury and reduced learning capacity. This has not been the case in clinical series. The efficacy of dexamethasone therapy in preventing neurologic sequelae is different between high- and low-income countries. Three large randomized trials in low-income countries (sub-Saharan Africa, Southeast Asia) failed to show benefit in subgroups of patients. The lack of efficacy of dexamethasone in these trials has been attributed to late presentation to the hospital with more advanced disease, antibiotic pretreatment, malnutrition, infection with HIV, and treatment of patients with probable, but not microbiologically proven, bacterial meningitis. The results of these clinical trials suggest that patients in sub-Saharan Africa and those in low-income countries with negative CSF Gram's stain and culture should not be treated with dexamethasone.

INCREASED INTRACRANIAL PRESSURE

Emergency treatment of increased ICP includes elevation of the patient's head to 30–45°, intubation and hyperventilation ($Paco_2$ 25–30 mmHg), and mannitol. Patients with increased ICP should be managed in an intensive care unit; accurate ICP measurements are best obtained with an ICP monitoring device.

Treatment of increased ICP is discussed in detail in Chap. 301.

PROGNOSIS

Mortality rate is 3–7% for meningitis caused by *H. influenzae, N. meningitidis*, or group B streptococci; 15% for that due to *L. monocytogenes*; and 20% for *S. pneumoniae*. In general, the risk of death from bacterial meningitis increases with (1) decreased level of consciousness on admission, (2) onset of seizures within 24 h of admission, (3) signs of increased ICP, (4) young age (infancy) and age >50, (5) the presence of comorbid conditions including shock and/or the need for mechanical ventilation, and (6) delay in the initiation of treatment. Decreased CSF glucose concentration (<2.2 mmol/L [<40 mg/dL]) and markedly increased CSF protein concentration (>3 g/L [> 300 mg/dL]) have been predictive of increased mortality and poorer outcomes in some series. Moderate or severe sequelae occur in ~25% of survivors, although the exact incidence varies with the infecting organism. Common sequelae include decreased intellectual function, memory impairment, seizures, hearing loss and dizziness, and gait disturbances.

VIRAL MENINGITIS

CLINICAL MANIFESTATIONS

Immunocompetent adult patients with viral meningitis usually present with headache, fever, and signs of meningeal irritation coupled with an inflammatory CSF profile (see below). Headache is almost invariably present and often characterized as frontal or retroorbital and frequently associated with photophobia and pain on moving the eyes. Nuchal rigidity is present in most cases but may be mild and present only near the limit of neck anteflexion. Constitutional signs can include malaise, myalgia, anorexia, nausea and vomiting, abdominal pain, and/or diarrhea. Patients often have mild lethargy or drowsiness; however, profound alterations in consciousness, such as stupor, coma, or marked confusion, do not occur in viral meningitis and suggest the presence of encephalitis or other alternative diagnoses. Similarly, seizures or focal neurologic signs or symptoms or neuroimaging abnormalities indicative of brain parenchymal involvement are not typical of viral meningitis and suggest the presence of encephalitis or another CNS infectious or inflammatory process.

ETIOLOGY

Using a variety of diagnostic techniques, including CSF PCR, culture, and serology, a specific viral cause can be found in 60–90% of cases of viral meningitis. The most important agents are enteroviruses (including echoviruses and coxsackieviruses in addition to numbered enteroviruses), varicella-zoster virus (VZV), HSV

(HSV-2 > HSV-1), HIV, and arboviruses **(Table 133-4)**. CSF cultures are positive in 30–70% of patients, with the frequency of isolation depending on the specific viral agent. Approximately two-thirds of culture-negative cases of "aseptic" meningitis have a specific viral etiology identified by CSF PCR testing (see below).

TABLE 133-4

Viruses Causing Acute Meningitis in North America

Common	Less Common
Enteroviruses (coxsackieviruses, echoviruses, and human enteroviruses 68–71) Varicella-zoster virus Herpes simplex virus 2 Epstein-Barr virus Arthropod-borne viruses HIV	Herpes simplex virus 1 Human herpesvirus 6 Cytomegalovirus Lymphocytic choriomeningitis virus Mumps Zika

EPIDEMIOLOGY

Viral meningitis is not a nationally reportable disease; however, it has been estimated that the incidence is ~60,000–75,000 cases per year. In temperate climates, there is a substantial increase in cases during the nonwinter months, reflecting the seasonal predominance of enterovirus and arthropod-borne virus (arbovirus) infections in the summer and fall, with a peak monthly incidence of about 1 reported case per 100,000 population.

LABORATORY DIAGNOSIS

CSF Examination

The most important laboratory test in the diagnosis of viral meningitis is examination of the CSF. The typical profile is a pleocytosis, a normal or slightly elevated protein concentration (0.2–0.8 g/L [20–80 mg/dL]), a normal glucose concentration, and a normal or mildly elevated opening pressure (100–350 mmH_2O).

Organisms are *not* seen on Gram's stain of CSF. The total CSF cell count in viral meningitis is typically 25–500/μL, although cell counts of several thousand/μL are occasionally seen, especially with infections due to lymphocytic choriomeningitis virus (LCMV) and mumps virus. Lymphocytes are typically the predominant cell. Rarely, PMNs may predominate in the first 48 h of illness, especially with infections due to echovirus 9, West Nile virus (WNV), eastern equine encephalitis (EEE) virus, or mumps. A PMN pleocytosis occurs in 45% of patients with WNV meningitis and can persist for a week or longer before shifting to a lymphocytic pleocytosis. PMN pleocytosis with low glucose may also be a feature of cytomegalovirus (CMV) infections in

immunocompromised hosts. Despite these exceptions, the presence of a CSF PMN pleocytosis in a patient with suspected viral meningitis in whom a specific diagnosis has not been established should prompt consideration of alternative diagnoses, including bacterial meningitis or parameningeal infections. The CSF glucose concentration is typically normal in viral infections, although it may be decreased in 10–30% of cases due to mumps or LCMV. Rare instances of decreased CSF glucose concentration occur in cases of meningitis due to echoviruses and other enteroviruses, HSV-2, and VZV. As a rule, a lymphocytic pleocytosis with a low glucose concentration should suggest fungal or tuberculous meningitis, *Listeria* meningoencephalitis, or noninfectious disorders (e.g., sarcoid, neoplastic meningitis).

A number of tests measuring levels of various CSF proteins, enzymes, and mediators—including C-reactive protein, lactic acid, lactate dehydrogenase, neopterin, quinolinate, IL-1β, IL-6, soluble IL-2 receptor, β_2-microglobulin, and TNF—have been proposed as potential discriminators between viral and bacterial meningitis or as markers of specific types of viral infection (e.g., infection with HIV), but they remain of uncertain sensitivity and specificity and are not widely used for diagnostic purposes.

Polymerase Chain Reaction Amplification of Viral Nucleic Acid

Amplification of viral-specific DNA or RNA from CSF using PCR amplification has become the single most important method for diagnosing CNS viral infections. In both enteroviral and HSV infections of the CNS, CSF PCR has become the diagnostic procedure of choice and is substantially more sensitive than viral cultures. HSV CSF PCR is also an important diagnostic test in patients with recurrent episodes of “aseptic” meningitis, many of whom have amplifiable HSV DNA in CSF despite negative viral cultures. CSF PCR is also used routinely to diagnose CNS viral infections caused by CMV, Epstein-Barr virus (EBV), VZV, and human herpesvirus 6 (HHV-6). CSF PCR tests are available for WNV but are not as sensitive as detection of WNV-specific CSF IgM. PCR is also useful in the diagnosis of CNS infection caused by *Mycoplasma pneumoniae,* which can mimic viral meningitis and encephalitis. PCR of throat washings may assist in diagnosis of enteroviral and mycoplasmal CNS infections. PCR of stool specimens may also assist in diagnosis of enteroviral infections (see below).

Viral Culture

The sensitivity of CSF cultures for the diagnosis of viral meningitis and encephalitis, in contrast to its utility in bacterial infections, is generally poor. In addition to CSF, specific viruses may also be isolated from throat swabs, stool, blood, and urine. Enteroviruses and adenoviruses may be found in feces; arboviruses, some enteroviruses, and LCMV in blood; mumps and CMV in urine; and enteroviruses, mumps, and adenoviruses in throat washings. During enteroviral infections, viral shedding in stool may persist for several weeks. The presence of enterovirus in stool is not diagnostic and may result from residual shedding from a previous enteroviral infection; it also occurs in some asymptomatic individuals during enteroviral epidemics.

Serologic Studies

The basic approach to the serodiagnosis of viral meningitis is identical to that discussed earlier for viral encephalitis **(see Chap. 132)**. Serologic studies are important for the diagnosis of arboviruses such as WNV, however these tests are less useful for viruses such as HSV, VZV, CMV, and EBV that have a high seroprevalence in the general population.

CSF oligoclonal gamma globulin bands occur in association with a number of viral infections. The associated antibodies are often directed against viral proteins. Oligoclonal bands also occur commonly in certain noninfectious neurologic diseases (e.g., multiple sclerosis) and may be found in nonviral infections (e.g., neurosyphilis, Lyme neuroborreliosis).

Other Laboratory Studies

All patients with suspected viral meningitis should have a complete blood count and differential, liver and renal function tests, erythrocyte sedimentation rate (ESR), and C-reactive protein, electrolytes, glucose, creatine kinase, aldolase, amylase, and lipase. Neuroimaging studies (MRI preferable to CT) are not absolutely necessary in patients with uncomplicated viral meningitis but should be performed in patients with altered consciousness, seizures, focal neurologic signs or symptoms (see "Differential Diagnosis" below), atypical CSF profiles, or underlying immunocompromising treatments or conditions.

DIFFERENTIAL DIAGNOSIS

The most important issue in the differential diagnosis of viral meningitis is to consider diseases that can mimic viral meningitis, including (1) untreated or partially treated bacterial meningitis; (2) early stages of meningitis caused by fungi, mycobacteria, or *Treponema pallidum* (neurosyphilis), in which a lymphocytic pleocytosis is common, cultures may be slow growing or negative, and hypoglycorrhachia may not be present early; (3) meningitis caused by agents such as *Mycoplasma, Listeria* spp., *Brucella* spp., *Coxiella* spp., *Leptospira* spp., and *Rickettsia* spp.; (4) parameningeal infections; (5) neoplastic meningitis; and (6) meningitis secondary to noninfectious inflammatory diseases, including medication-induced hypersensitivity meningitis, SLE and other rheumatologic diseases, sarcoidosis, Behçet's syndrome, and the uveomeningitic syndromes. Studies in children >28 days of age suggest that the presence of CSF protein >0.5 g/L (sensitivity 89%, specificity 78%) and elevated serum procalcitonin levels >0.5 ng/mL (sensitivity 89%, specificity 89%) were clues to the presence of bacterial as opposed to "aseptic" meningitis. A variety of clinical algorithms for differentiating bacterial from aseptic meningitis have been developed. One such prospectively validated system, the *bacterial meningitis score*, suggests that the probability of bacterial meningitis is 0.3% or less (negative predictive value 99.7%, 95% confidence interval 99.6–100%) in children with CSF pleocytosis who have (1) a negative CSF Gram's stain, (2) CSF neutrophil count <1000 cells/μL, (3) CSF protein <80 mg/dL, (4) peripheral absolute neutrophil count of <10,000 cells/μL, and (5) no prior history or current presence of seizures.

SPECIFIC VIRAL ETIOLOGIES

Enteroviruses (EV) (Chap. 199) are the most common cause of viral meningitis, accounting for >85% of cases in which a specific etiology can be identified. Cases may either be sporadic or occur in clusters. EV71 has produced large epidemics of neurologic disease outside the United States, especially in Southeast Asia, but most recently reported cases in the United States have been sporadic. Enteroviruses are the most likely cause of viral meningitis in the summer and fall months, especially in children (<15 years), although cases occur at reduced frequency year round. Although the incidence of enteroviral meningitis declines with increasing age, some outbreaks have preferentially affected older children and adults. Meningitis outside the neonatal period is usually benign. Patients present with sudden onset of fever; headache; nuchal rigidity; and often constitutional signs, including vomiting, anorexia, diarrhea, cough, pharyngitis, and myalgias. The physical examination should include a careful search for stigmata of enterovirus infection, including exanthems, hand-foot-mouth disease, herpangina, pleurodynia, myopericarditis, and hemorrhagic conjunctivitis. The CSF profile is typically a lymphocytic pleocytosis (100–1000 cells/μL) with normal glucose and normal or mildly elevated protein concentration. However, up to 15% of patients, most commonly young infants rather than older children or adults, have a normal CSF leukocyte count. In rare cases, PMNs may predominate during the first 48 h of illness. CSF reverse transcriptase PCR (RT-PCR) is the diagnostic procedure of choice and is both sensitive (>95%) and specific (>100%). CSF PCR has the highest sensitivity if performed within 48 h of symptom onset, with sensitivity declining rapidly after day 5 of symptoms. PCR of throat washings or stool specimens may be positive for several weeks, and positive results can help support the diagnosis of an acute enteroviral infection. The sensitivity of routine enteroviral PCRs for detecting EV71 is low, and specific testing may be required. Treatment is supportive, and patients usually recover without sequelae. Chronic and severe infections can occur in neonates and in individuals with hypo- or agammaglobulinemia.

Arbovirus infections (Chap. 204) occur predominantly in the summer and early fall. Arboviral meningitis should be considered when clusters of meningitis and encephalitis cases occur in a restricted geographic region during the summer or early fall. In the United States, the most important causes of arboviral meningitis and encephalitis are WNV, St. Louis encephalitis virus, and the California encephalitis group of viruses. In WNV epidemics, avian deaths may serve as sentinel infections for subsequent human disease. A history of tick exposure or travel or residence in the appropriate geographic area should suggest the possibility of Colorado tick fever virus or Powassan virus infection, although nonviral tick-borne diseases, including RMSF and Lyme neuroborreliosis, may present similarly. Arbovirus meningitis is typically associated with a CSF lymphocytic pleocytosis, normal glucose concentration, and normal or mildly elevated protein concentration. However, ~45% of patients with WNV meningitis have CSF neutrophilia, which can persist for a week or more. The rarity of hypoglycorrhachia in WNV infection, the absence of positive Gram's stains, and the negative cultures help distinguish these patients from those with bacterial meningitis. Definitive diagnosis of arboviral meningitis is based on demonstration of viral-specific IgM in CSF or seroconversion. The prevalence of CSF IgM increases progressively during the first week after infection, peaking at >80% in patients with neuroinvasive disease; as a result, repeat studies may be needed when disease suspicion is high and an early study is negative. CSF PCR tests are available for some viruses in selected diagnostic laboratories and at the Centers for Disease Control and Prevention (CDC), but in the case of WNV, sensitivity (~70%) of CSF PCR is less than that of CSF serology. WNV CSF PCR may be useful in immunocompromised patients who may have absent or reduced antibody responses.

HSV meningitis (Chap. 187) has been increasingly recognized as a major cause of viral meningitis in adults, and overall, it is probably second in importance to enteroviruses as a cause of viral meningitis, accounting for 5% of total cases overall and undoubtedly a higher frequency of those cases occurring in adults and/or outside of the summer-fall period when enterovirus infections are increasingly common. In adults, the majority of cases of uncomplicated meningitis are due to HSV-2, whereas HSV-1 is responsible for 90% of cases of HSV encephalitis. HSV meningitis occurs in ~25–35% of women and ~10–15% of men at the time of an initial (primary) episode of genital herpes. Of these patients, 20% go on to have recurrent attacks of meningitis. Diagnosis of HSV meningitis is usually by HSV CSF PCR because cultures may be negative, especially in patients with recurrent meningitis. Demonstration of intrathecal synthesis of HSV-specific antibody may also be useful in diagnosis, although antibody tests are less sensitive and less specific than PCR and may not become positive until after the first week of infection. Although a history of or the presence of HSV genital lesions is an important diagnostic clue, many patients with HSV meningitis give no history and have no evidence of active genital herpes at the time of presentation. Most cases of recurrent viral or "aseptic" meningitis, including cases previously diagnosed as Mollaret's meningitis, are due to HSV.

VZV meningitis should be suspected in the presence of concurrent chickenpox or shingles. However, it is important to recognize that VZV is being increasingly identified as an important cause of both meningitis and encephalitis in patients without rash. The frequency of VZV as a cause of meningitis is extremely variable, ranging from as low as 3% to as high as 20% in different series. Diagnosis is usually based on CSF PCR, although the sensitivity of this test is not as high as for the other herpesviruses. VZV serologic studies complement PCR testing, and the diagnosis of VZV CNS infection can be made by the demonstration of VZV-specific intrathecal antibody synthesis and/or the presence of VZV CSF IgM antibodies, or by positive CSF cultures.

EBV infections may also produce aseptic meningitis, with or without associated infectious mononucleosis. The presence of atypical lymphocytes in the CSF or peripheral blood is suggestive of EBV infection but may occasionally be seen with other viral infections. EBV is almost never cultured from CSF. Serum and CSF serology help establish the presence of acute infection, which is characterized by IgM viral capsid antibodies (VCAs), antibodies to early antigens (EAs), and the absence of antibodies to EBV-associated nuclear antigen (EBNA). CSF PCR is another important diagnostic test, although false-positive results may reflect viral reactivation associated with other infectious or inflammatory processes or the presence of latent viral DNA in lymphocytes recruited due to other inflammatory conditions.

HIV meningitis should be suspected in any patient presenting with a viral meningitis with known or suspected risk factors for HIV infection. Meningitis may occur following primary infection with HIV in 5–10% of cases and less commonly at later stages of illness. Cranial nerve palsies, most commonly involving cranial nerves V, VII, or VIII, are more common in HIV meningitis than in other viral infections. Diagnosis can be confirmed by detection of HIV genome in blood or CSF. Seroconversion may be delayed, and patients with negative HIV serologies who are suspected of having HIV meningitis should be monitored for delayed seroconversion. **For further discussion of HIV infection, see Chap. 197.**

Mumps (Chap. 202) should be considered when meningitis occurs in the late winter or early spring, especially in males (male-to-female ratio 3:1). With the widespread use of the live attenuated mumps vaccine in the United States since 1967, the incidence of mumps meningitis has fallen by >95%; however, mumps remains a potential source of infection in nonimmunized individuals and populations. Rare cases (10–100:100,000 vaccinated individuals) of vaccine-associated mumps meningitis have been described, with onset typically 2–4 weeks after vaccination. The presence of parotitis, orchitis, oophoritis, pancreatitis, or elevations in serum lipase and amylase is suggestive of mumps meningitis; however, their absence does not exclude the diagnosis. Clinical meningitis was previously estimated to occur in 10–30% of patients with mumps parotitis; however, in a recent U.S. outbreak of nearly 2600 cases of mumps, only 11 cases of meningitis were identified, suggesting the incidence may be lower than previously suspected. Mumps infection confers lifelong immunity, so a documented history of previous infection excludes this diagnosis. Patients with meningitis have a CSF pleocytosis that can exceed 1000 cells/μL in 25%. Lymphocytes predominate in 75%, although CSF neutrophilia occurs in 25%. Hypoglycorrhachia occurs in 10–30% of patients and may be a clue to the diagnosis when present. Diagnosis is typically made by culture of virus from CSF or by detecting IgM antibodies or seroconversion. CSF PCR is available in some diagnostic and research laboratories.

LCMV infection (Chap. 204) should be considered when aseptic meningitis occurs in the late fall or winter and in individuals with a history of exposure to house mice (*Mus musculus*), pet or laboratory rodents (e.g., hamsters, rats, mice), or their excreta. Some patients have an associated rash, pulmonary infiltrates, alopecia, parotitis, orchitis, or myopericarditis. Laboratory clues to the diagnosis of LCMV, in addition to the clinical findings noted above, may include the presence of leukopenia, thrombocytopenia, or abnormal liver function tests. Some cases present with a marked CSF pleocytosis (>1000 cells/μL) and hypoglycorrhachia (<30%). Diagnosis is based on serology and/or culture of virus from CSF.

TREATMENT

TREATMENT

Acute Viral Meningitis

Treatment of almost all cases of viral meningitis is primarily symptomatic and includes use of analgesics, antipyretics, and antiemetics. Fluid and electrolyte status should be monitored. Patients with suspected bacterial meningitis should receive appropriate empirical therapy pending culture results (see above). Hospitalization may not be required in immunocompetent patients with presumed viral meningitis and no focal signs or symptoms, no significant alteration in consciousness, and a classic CSF profile (lymphocytic pleocytosis, normal glucose, negative Gram's stain) if adequate provision for monitoring at home and medical follow-up can be ensured. Immunocompromised patients; patients with significant alteration in consciousness, seizures, or the presence of focal signs and symptoms suggesting the possibility of encephalitis or parenchymal brain involvement; and patients who have an atypical CSF profile should be hospitalized. Oral or intravenous acyclovir may be of benefit in patients with meningitis caused by HSV-1 or 2 and in cases of severe EBV or VZV infection. Data concerning treatment of HSV, EBV, and VZV meningitis are extremely limited. Seriously ill patients should probably receive intravenous acyclovir (15–30 mg/kg per day

in three divided doses), which can be followed by an oral drug such as acyclovir (800 mg five times daily), famciclovir (500 mg tid), or valacyclovir (1000 mg tid) for a total course of 7–14 days. Patients who are less ill can be treated with oral drugs alone. Patients with HIV meningitis should receive highly active antiretroviral therapy **(Chap. 197)**. There is no specific therapy of proven benefit for patients with arboviral encephalitis, including that caused by WNV.

Patients with viral meningitis who are known to have deficient humoral immunity (e.g., X-linked agammaglobulinemia) and who are not already receiving either intramuscular gamma globulin or intravenous immunoglobulin (IVIg) should be treated with these agents. Intraventricular administration of immunoglobulin through an Ommaya reservoir has been tried in some patients with chronic enteroviral meningitis who have not responded to intramuscular or intravenous immunoglobulin.

Vaccination is an effective method of preventing the development of meningitis and other neurologic complications associated with poliovirus, mumps, measles, rubella, and varicella infection. A live attenuated VZV vaccine (Varivax) is available in the United States. Clinical studies indicate an effectiveness rate of 70–90% for this vaccine, but a booster may be required after ~10 years to maintain immunity. A live attenuated vaccine (Zostavax) is recommended for prevention of herpes zoster (shingles) in adults aged >60 at the present time. A new vaccine should be available soon. The herpes zoster subunit vaccine (HZ/su) containing recombinant varicella-zoster virus glycoprotein E and an adjuvant system has greater efficacy in preventing zoster in adults aged ≥70 years than the live attenuated vaccine. An inactivated varicella vaccine is available for transplant recipients and others for whom live viral vaccines are contraindicated.

PROGNOSIS

In adults, the prognosis for full recovery from viral meningitis is excellent. Rare patients complain of persisting headache, mild mental impairment, incoordination, or generalized asthenia for weeks to months. The outcome in infants and neonates (<1 year) is less certain; intellectual impairment, learning disabilities, hearing loss, and other lasting sequelae have been reported in some studies.

SUBACUTE MENINGITIS

CLINICAL MANIFESTATIONS

Patients with subacute meningitis typically have an unrelenting headache, stiff neck, low-grade fever, and lethargy for days to several weeks before they present for evaluation. Cranial nerve abnormalities and night sweats may be present. **This syndrome overlaps that of chronic meningitis, discussed in detail in Chap. 134, but is included here as the meningeal pathogens of subacute meningitis can also present as an acute meningitis.**

ETIOLOGY

Common causative organisms include *M. tuberculosis, C. neoformans, H. capsulatum, C. immitis*, and *T. pallidum*. Initial infection with *M. tuberculosis* is acquired by inhalation of aerosolized droplet nuclei. Tuberculous meningitis in adults does not develop acutely from hematogenous spread of tubercle bacilli to the meninges. Rather, millet seed–sized (miliary) tubercles form in the parenchyma of the brain during hematogenous dissemination of tubercle bacilli in the course of primary infection. These tubercles enlarge and are usually caseating. The propensity for a caseous lesion to produce meningitis is determined by its proximity to the SAS and the rate at which fibrous encapsulation develops. Subependymal caseous foci cause meningitis via discharge of bacilli and tuberculous antigens into the SAS. Mycobacterial antigens produce an intense inflammatory reaction that leads to the production of a thick exudate that fills the basilar cisterns and surrounds the cranial nerves and major blood vessels at the base of the brain.

 Fungal infections are typically acquired by the inhalation of airborne fungal spores. The initial pulmonary infection may be asymptomatic or present with fever, cough, sputum production, and chest pain. The pulmonary infection is often self-limited. A localized pulmonary fungal infection can then remain dormant in the lungs until there is an abnormality in cell-mediated immunity that allows the fungus to reactivate and disseminate to the CNS. The most common pathogen causing fungal meningitis is *C. neoformans*. This fungus is found worldwide in soil and bird excreta. *H. capsulatum* is endemic to the Ohio and Mississippi River valleys of the central United States and to parts of Central and South America. *C. immitis* is endemic to the desert areas of the southwest United States, northern Mexico, and Argentina.

Syphilis is a sexually transmitted disease that is manifest by the appearance of a painless chancre at the site of inoculation. *T. pallidum* invades the CNS early in the course of syphilis. Cranial nerves VII and VIII are most frequently involved.

LABORATORY DIAGNOSIS

The classic CSF abnormalities in tuberculous meningitis are as follows: (1) elevated opening pressure, (2) lymphocytic pleocytosis (10–500 cells/μL), (3) elevated protein concentration in the range of 1–5 g/L, and (4) decreased glucose concentration in the range of 1.1–2.2 mmol/L (20–40 mg/dL). *The combination of unrelenting headache, stiff neck, fatigue, night sweats, and fever with a CSF lymphocytic pleocytosis and a mildly decreased glucose concentration is highly suspicious for tuberculous meningitis.* The last tube of fluid collected at LP is the best tube to send for a smear for acid-fast bacilli (AFB). If there is a pellicle in the CSF or a cobweb-like clot on the surface of the fluid, AFB can best be demonstrated in a smear of the clot or pellicle. Positive smears are typically reported in only 10–40% of cases of tuberculous meningitis in adults. Cultures of CSF take 4–8 weeks to identify the organism and are positive in ~50% of adults. Culture remains the gold standard to make the diagnosis of tuberculous meningitis. PCR for the detection of *M. tuberculosis* DNA should be sent on CSF if available, but the sensitivity and specificity on CSF have not been defined. The CDC recommends the use of nucleic acid amplification tests for the diagnosis of pulmonary tuberculosis.

The characteristic CSF abnormalities in fungal meningitis are a mononuclear or lymphocytic pleocytosis, an increased protein concentration, and a decreased glucose concentration. There may be eosinophils in the

CSF in *C. immitis* meningitis. Large volumes of CSF are often required to demonstrate the organism on India ink smear or grow the organism in culture. If spinal fluid examined by LP on two separate occasions fails to yield an organism, CSF should be obtained by high-cervical or cisternal puncture.

The cryptococcal polysaccharide antigen test is a highly sensitive and specific test for cryptococcal meningitis. A reactive CSF cryptococcal antigen test establishes the diagnosis. The detection of the *Histoplasma* polysaccharide antigen in CSF establishes the diagnosis of a fungal meningitis but is not specific for meningitis due to *H. capsulatum*. It may be falsely positive in coccidioidal meningitis. The CSF complement fixation antibody test is reported to have a specificity of 100% and a sensitivity of 75% for coccidioidal meningitis.

The diagnosis of syphilitic meningitis is made when a reactive serum treponemal test (fluorescent treponemal antibody absorption test [FTA-ABS] or microhemagglutination assay–*T. pallidum* [MHA-TP]) is associated with a CSF lymphocytic or mononuclear pleocytosis and an elevated protein concentration, or when the CSF Venereal Disease Research Laboratory (VDRL) test is positive. A reactive CSF FTA-ABS is not definitive evidence of neurosyphilis. The CSF FTA-ABS can be falsely positive from blood contamination. A negative CSF VDRL does not rule out neurosyphilis. A negative CSF FTA-ABS or MHA-TP rules out neurosyphilis.

TREATMENT

TREATMENT

Subacute Meningitis

Empirical therapy of tuberculous meningitis is often initiated on the basis of a high index of suspicion without adequate laboratory support. Initial therapy is a combination of isoniazid (300 mg/d), rifampin (10 mg/kg per day), pyrazinamide (30 mg/kg per day in divided doses), ethambutol (15–25 mg/kg per day in divided doses), and pyridoxine (50 mg/d). When the antimicrobial sensitivity of the *M. tuberculosis* isolate is known, ethambutol can be discontinued. If the clinical response is good, pyrazinamide can be discontinued after 8 weeks and isoniazid and rifampin continued alone for the next 6–12 months. A 6-month course of therapy is acceptable, but therapy should be prolonged for 9–12 months in patients who have an inadequate resolution of symptoms of meningitis or who have positive mycobacterial cultures of CSF during the course of therapy. Dexamethasone therapy is recommended for HIV-negative patients with tuberculous meningitis. The dose is 12–16 mg/d for 3 weeks, and then tapered over 3 weeks.

Meningitis due to *C. neoformans* in non-HIV, nontransplant patients is treated with induction therapy with amphotericin B (AmB) (0.7 mg/kg IV per day) plus flucytosine (100 mg/kg per day in four divided doses) for at least 4 weeks if CSF culture results are negative after 2 weeks of treatment. Therapy should be extended for a total of 6 weeks in the patient with neurologic complications. Induction therapy is followed by consolidation therapy with fluconazole 400 mg/d for 8 weeks. Organ transplant recipients are treated with liposomal AmB (3–4 mg/kg per day) or AmB lipid complex (ABLC) 5 mg/kg per day plus flucytosine (100 mg/kg per day in four divided doses) for at least 2 weeks or until CSF culture is sterile. Follow CSF yeast cultures for sterilization

rather than the cryptococcal antigen titer. This treatment is followed by an 8- to 10-week course of fluconazole (400–800 mg/d [6–12 mg/kg] PO). If the CSF culture is sterile after 10 weeks of acute therapy, the dose of fluconazole is decreased to 200 mg/d for 6 months to a year. Patients with HIV infection are treated with AmB or a lipid formulation plus flucytosine for at least 2 weeks, followed by fluconazole for a minimum of 8 weeks. HIV-infected patients may require indefinite maintenance therapy with fluconazole 200 mg/d. Meningitis due to *H. capsulatum* is treated with AmB (0.7–1.0 mg/kg per day) for 4–12 weeks. A total dose of 30 mg/kg is recommended. Therapy with AmB is not discontinued until fungal cultures are sterile. After completing a course of AmB, maintenance therapy with itraconazole 200 mg two or three times daily is initiated and continued for at least 9 months to a year. *C. immitis* meningitis is treated with either high-dose fluconazole (1000 mg daily) as monotherapy or intravenous AmB (0.5–0.7 mg/kg per day) for >4 weeks. Intrathecal AmB (0.25–0.75 mg/d three times weekly) may be required to eradicate the infection. Lifelong therapy with fluconazole (200–400 mg daily) is recommended to prevent relapse. AmBisome (5 mg/kg per day) or AmB lipid complex (5 mg/kg per day) can be substituted for AmB in patients who have or who develop significant renal dysfunction. The most common complication of fungal meningitis is hydrocephalus. Patients who develop hydrocephalus should receive a CSF diversion device. A ventriculostomy can be used until CSF fungal cultures are sterile, at which time the ventriculostomy is replaced by a ventriculoperitoneal shunt.

Syphilitic meningitis is treated with aqueous penicillin G in a dose of 3–4 million units intravenously every 4 h for 10–14 days. An alternative regimen is 2.4 million units of procaine penicillin G intramuscularly daily with 500 mg of oral probenecid four times daily for 10–14 days. Either regimen is followed with 2.4 million units of benzathine penicillin G intramuscularly once a week for 3 weeks. The standard criterion for treatment success is reexamination of the CSF. The CSF should be reexamined at 6-month intervals for 2 years. The cell count is expected to normalize within 12 months, and the VDRL titer to decrease by two dilutions or revert to nonreactive within 2 years of completion of therapy. Failure of the CSF pleocytosis to resolve or an increase in the CSF VDRL titer by two or more dilutions requires retreatment.

FURTHER READING

Cunningham Al et al: Efficacy of the herpes zoster subunit vaccine in adults 70 years of age or older. N Engl J Med 375:1019, 2016.

Heckenberg SG et al: Adjunctive dexamethasone in adults with meningococcal meningitis. Neurology 79:1563, 2012.

Roos KL et al: Acute bacterial meningitis, in *Infections of the Central Nervous System,* 4th ed. Scheld WM, Whitley RJ, Marra (eds). Philadelphia, Wolters Kluwer Health, 2014, pp 365–419.

Chapter 134: Chronic and Recurrent Meningitis

Walter J. Koroshetz; Michael R. Wilson; Avindra Nath

FIGURE 134-1

INTRODUCTION

Chronic inflammation of the meninges (pia, arachnoid, and dura) can produce profound neurologic disability and may be fatal if not successfully treated. Chronic meningitis is diagnosed when a characteristic neurologic syndrome exists for >4 weeks and is associated with a persistent inflammatory response in the cerebrospinal fluid (CSF) (white blood cell count >5/μL). The causes are varied, and appropriate treatment depends on identification of the etiology. Five categories of disease account for most cases of chronic meningitis: (1) meningeal infections, (2) malignancy, (3) autoimmune inflammatory disorders, (4) chemical meningitis, and (5) parameningeal infections.

CLINICAL PATHOPHYSIOLOGY

Neurologic manifestations of chronic meningitis (Table 134-1) are determined by the anatomic location of the inflammation and its consequences. Persistent headache, hydrocephalus, cranial neuropathies, radiculopathies, and cognitive or personality changes are the cardinal features. These can occur alone or in combination. When they appear in combination, widespread dissemination of the inflammatory process along CSF pathways has occurred. In some cases, the presence of an underlying systemic illness points to a specific agent or class of agents as the probable cause. The diagnosis of chronic meningitis is usually made when the clinical presentation prompts the physician to examine the CSF for signs of inflammation. CSF is produced by the choroid plexus of the cerebral ventricles, exits through narrow foramina into the subarachnoid space surrounding the brain and spinal cord, circulates around the base of the brain and over the cerebral hemispheres, and is resorbed by arachnoid villi projecting into the superior sagittal sinus. CSF flow provides a pathway for rapid spread of infectious and other infiltrative processes over the brain, spinal cord, cranial, and spinal nerve roots. Spread from the subarachnoid space into brain parenchyma may occur via the arachnoid cuffs that surround blood vessels that penetrate brain tissue (Virchow-Robin spaces).

TABLE 134-1

Symptoms and Signs of Chronic Meningitis

Symptom	Sign
Chronic headache	± Papilledema
Neck or back pain/stiffness	Brudzinski's or Kernig's sign of meningeal irritation
Change in personality	Altered mental status—drowsiness, inattention, disorientation, memory loss, frontal release signs (grasp, suck, snout), perseveration
Facial weakness	Peripheral seventh CN paresis
Double vision	Paresis of CNs III, IV, VI
Diminished vision	Papilledema, optic atrophy
Hearing loss	Eighth CN paresis
Arm or leg weakness	Myelopathy or radiculopathy
Numbness in arms or legs	Myelopathy or radiculopathy
Urinary retention/incontinence	Myelopathy or radiculopathy Frontal lobe dysfunction (hydrocephalus)
Clumsiness	Ataxia

Abbreviation: CN, cranial nerve.

Intracranial Meningitis

Nociceptive nerve fibers of the meninges are stimulated by the inflammatory process, resulting in headache, neck pain, or back pain. Obstruction of CSF pathways at the foramina or arachnoid villi may produce hydrocephalus and signs and symptoms of raised intracranial pressure (ICP), including headache, vomiting, apathy or drowsiness, gait instability, papilledema, visual loss, impaired upgaze, or palsy of the sixth cranial nerve (CN) **(Chap. 433)**. Cognitive and behavioral changes during the course of chronic meningitis may also

result from vascular damage due to inflammation around the blood vessels that course in the subarachnoid space, causing infarction. Inflammatory deposits seeded via the CSF circulation are often prominent around the brainstem and cranial nerves and along the undersurface of the frontal and temporal lobes. Such cases, termed basal meningitis, often present as multiple cranial neuropathies, with decreased vision (CN II), facial weakness (CN VII), decreased hearing (CN VIII), diplopia (CNs III, IV, and VI), sensory or motor abnormalities of the oropharynx (CNs IX, X, and XII), decreased olfaction (CN I), or decreased facial sensation and masseter weakness (CN V).

Spinal Meningitis

Injury may occur to motor and sensory nerve roots as they traverse the subarachnoid space and penetrate the meninges. These cases present as multiple radiculopathies with combinations of radicular pain, sensory loss, motor weakness, and urinary or fecal incontinence. In some cases chronic inflammation causes clumping of the lower nerve roots and thickening of the meninges, so called pachymeningitis. Meningeal inflammation can encircle and damage the cord, resulting in a myelopathy. Patients with slowly progressive involvement of multiple cranial nerves and/or spinal nerve roots are likely to have chronic meningitis. Electrophysiologic testing (electromyography, nerve conduction studies, and evoked response testing) may be helpful in determining whether there is involvement of cranial and spinal nerve roots.

Systemic Manifestations

In some patients, evidence of systemic disease provides clues to the underlying cause of chronic meningitis. A complete history of travel, sexual practice, and exposure to infectious agents should be sought. Infectious causes are often associated with fever, malaise, anorexia, and signs of localized or disseminated infection outside the nervous system. Infectious causes are of major concern in the immunosuppressed patient, especially in patients with HIV infection, in whom chronic meningitis may present without headache or fever. Noninfectious inflammatory disorders most often produce systemic manifestations first, but meningitis may be the initial manifestation. Carcinomatous meningitis may or may not be accompanied by clinical evidence of the primary neoplasm.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Chronic Meningitis

The occurrence of chronic headache, hydrocephalus, cranial neuropathy, radiculopathy, and/or cognitive decline in a patient should prompt consideration of a lumbar puncture for evidence of meningeal inflammation. On occasion, the diagnosis is made when a contrast-enhanced imaging study (CT or MRI) shows leakage of contrast agent into the meninges. Meningeal enhancement is always concerning with the exception of dural enhancement after lumbar puncture, neurosurgical procedures, concussion, or spontaneous CSF leakage. Once chronic meningitis is confirmed by CSF examination, effort is focused on identifying the cause (**Tables 134-2 and 134-3**) by (1) further analysis of the CSF, (2) diagnosis of an underlying

systemic infection or noninfectious inflammatory condition, or (3) pathologic examination of meningeal biopsy specimens.

Two clinical forms of chronic meningitis exist. In the first, the symptoms are chronic and persistent, whereas in the second there are recurrent, discrete episodes of illness. In the latter group, all symptoms, signs, and CSF parameters of meningeal inflammation resolve completely between episodes either spontaneously or in response to a specific therapy. In such patients, the likely etiologies include Mollaret's meningitis which is most often due to herpes simplex virus (HSV) type 2; chemical meningitis due to episodic leakage from an epidermoid tumor, craniopharyngioma, or cholesteatoma into CSF; primary autoimmune inflammatory conditions, including Vogt-Koyanagi-Harada syndrome, Behçet's syndrome, systemic lupus erythematosus (SLE), rheumatoid arthritis, IgG4-related disease and drug hypersensitivity with repeated administration of the offending agent.

The epidemiologic history is of considerable importance in diagnosis of chronic meningitis and may provide direction for selection of laboratory studies. Pertinent features include a history of tuberculosis or exposure; recent epidural injection that led to epidemic of fungal meningitis by *Exserohilum rostratum*; past travel to areas endemic for fungal infections (the San Joaquin Valley in California and southwestern states for coccidioidomycosis, midwestern states for histoplasmosis, southeastern states for blastomycosis); travel to the Mediterranean region or ingestion of imported unpasteurized dairy products (*Brucella*); time spent in wooded areas endemic for Lyme disease; exposure to sexually transmitted disease (syphilis); exposure of an immunocompromised host to pigeons and their droppings (*Cryptococcus neoformans*); exposure to decaying wood in Vancouver Island, South Africa, and Australia (*Cryptococcus neoformans var. gatti*); gardening (*Sporothrix schenckii*); ingestion of poorly cooked meat or contact with a household cat (*Toxoplasma gondii*); residence in Thailand or Japan (*Gnathostoma spinigerum*), Latin America (*Paracoccidioides brasiliensis*), or the South Pacific (*Angiostrongylus cantonensis*); rural residence and raccoon exposure (*Baylisascaris procyonis*); and residence in Latin America, the Philippines, sub-Saharan Africa, or Southeast Asia (*Taenia solium/cysticercosis*).

The presence of focal cerebral signs in a patient with chronic meningitis suggests the possibility of a brain abscess, parameningeal infection or infarct; identification of a potential source of infection (chronic draining ear, sinusitis, right-to-left cardiac or pulmonary shunt, chronic pleuropulmonary infection) supports this diagnosis. In some cases, diagnosis may be established by recognition and biopsy of unusual skin lesions (Behçet's syndrome, SLE, cryptococcosis, blastomycosis, Lyme disease, sporotrichosis, trypanosomiasis, IV drug use) or enlarged lymph nodes (lymphoma, sarcoid, tuberculosis, HIV, secondary syphilis, or Whipple's disease). A careful ophthalmologic examination may reveal uveitis (Vogt-Koyanagi-Harada syndrome, sarcoid, or central nervous system [CNS] lymphoma), keratoconjunctivitis sicca (Sjögren's syndrome), or iridocyclitis (Behçet's syndrome) and is essential to assess visual loss from papilledema. Aphthous oral lesions, genital ulcers, and hypopyon suggest Behçet's syndrome. Hepatosplenomegaly suggests lymphoma, sarcoid, tuberculosis, or brucellosis. Herpetic lesions in the genital area or on the thighs suggest HSV-2 infection. A breast nodule, a suspicious hyperpigmented skin lesion, focal bone pain, hard, fixed lymph nodes, or an abdominal mass directs attention to possible carcinomatous meningitis.

IMAGING

Once the clinical syndrome is recognized as a potential manifestation of chronic meningitis, proper analysis of the CSF is essential. However, if the possibility of raised ICP exists, a brain imaging study should be performed before lumbar puncture. If ICP is elevated because of a mass lesion, brain swelling, or a block in ventricular CSF outflow (obstructive hydrocephalus), then lumbar puncture carries the potential risk of brain herniation. Obstructive hydrocephalus usually requires direct ventricular drainage. In patients with open CSF flow pathways, elevated ICP can still occur due to impaired resorption of CSF by arachnoid villi. In such patients, lumbar puncture is usually safe and may be therapeutic, but repetitive or continuous lumbar drainage may be necessary to prevent abrupt deterioration and death from raised ICP. In some patients, especially those with cryptococcal meningitis, fatal levels of raised ICP can occur without enlarged ventricles.

Contrast-enhanced MRI or CT studies of the brain and spinal cord can identify meningeal enhancement, parameningeal infections (including brain abscess), encasement of the spinal cord (malignancy, inflammation or infection), or nodular deposits on the meninges or nerve roots (malignancy or sarcoidosis) (Fig. 134-1). Imaging studies are also useful to guide biopsy of affected meninges.

Angiographic studies can identify evidence of cerebral arteritis in patients with chronic meningitis and stroke.

CEREBROSPINAL FLUID ANALYSIS

The CSF pressure should be measured and samples sent for bacterial, fungal, and tuberculous culture; Venereal Disease Research Laboratories (VDRL) test; cell count and differential; Gram's stain; and measurement of glucose and protein. Wet mount for fungus and parasites, india ink preparation, culture for fastidious bacteria and fungi, assays for cryptococcal antigen and oligoclonal immunoglobulin bands, and cytology should be performed. Other specific CSF tests (Tables 134-2 and 134-3) or blood tests and cultures should be ordered as indicated on the basis of the history, physical examination, or preliminary CSF results (i.e., eosinophilic, mononuclear, or polymorphonuclear meningitis). Rapid diagnosis may be facilitated by serologic tests and polymerase chain reaction (PCR) testing to identify DNA sequences in the CSF that are specific for the suspected pathogen. 16s ribosomal RNA (rRNA) PCR can be used to detect a broad range of bacterial causes of meningitis and can be particularly useful in partially treated meningitis when the yield of culture is low. 18s and 28s rRNA can similarly be useful for detecting a broad range of fungal species. In patients with suspected fungal infections, when other tests are negative, assays for beta-glucans may be a useful adjunct in establishing the diagnosis. Building on progress in parallel deep sequencing and informatics, unbiased metagenomic next-generation sequencing is becoming generally available, representing an efficient and powerful method for diagnosis of challenging diagnostic cases.

In most categories of chronic (not recurrent) meningitis, mononuclear cells predominate in the CSF. When neutrophils predominate after 3 weeks of illness, the principal etiologic considerations are *Nocardia asteroides*, *Actinomyces israelii*, *Brucella*, *Mycobacterium tuberculosis* (5–10% of early cases only), various fungi (*Blastomyces dermatitidis*, *Candida albicans*, *Histoplasma capsulatum*, *Aspergillus* spp., *Pseudallescheria boydii*, *Cladophialophora bantiana*), and noninfectious causes (SLE, exogenous chemical meningitis). When eosinophils predominate or are present in limited numbers in a primarily mononuclear

cell response in the CSF, the differential diagnosis includes parasitic diseases (*A. cantonensis*, *G. spinigerum*, *B. procyonis*, or *Toxocara canis* infection, cysticercosis, schistosomiasis, echinococcal disease, *T. gondii* infection), fungal infections (6–20% eosinophils along with a predominantly lymphocyte pleocytosis, particularly with coccidioidal meningitis), neoplastic disease (lymphoma, leukemia, metastatic carcinoma), or other inflammatory processes (sarcoidosis, hypereosinophilic syndrome).

It is often necessary to broaden the number of diagnostic tests if the initial workup does not reveal the cause. In addition, repeated samples (three or more) of large volumes of lumbar CSF may be required to diagnose certain infectious and malignant causes of chronic meningitis. Lymphomatous or carcinomatous meningitis may be diagnosed by examination of sections cut from a cell block formed by spinning down the sediment from a large volume of CSF. Flow cytometry for malignant cells may also be useful in patients with suspected carcinomatous meningitis. The diagnosis of fungal meningitis may also require large volumes of CSF for culture of sediment. If standard lumbar puncture is unrewarding, a cervical cisternal tap to sample CSF near to the basal meninges may be fruitful. Ventricular fluid may appear sterile in cases with active infection in the lower lumbar space.

LABORATORY INVESTIGATION

In addition to the CSF examination, an attempt should be made to uncover pertinent underlying illnesses. Tuberculin skin test, chest radiograph, urine analysis and culture, blood count and differential, renal and liver function tests, alkaline phosphatase, sedimentation rate, antinuclear antibody, anti-Ro antibody, anti-La antibody, rheumatoid factor, and IgG4 level are often indicated. In some cases, a thorough search for a systemic site of infection is indicated. Pulmonary foci of infection may be present, particularly with fungal or tuberculous disease. Hence a CT or MRI of the chest and a sputum examination may be helpful. Abnormalities can be pursued by bronchoscopy or transthoracic needle biopsy. A tuberculin skin test is often placed, although the test has limited specificity and sensitivity for diagnosis of active disease. Where available gamma interferon release assays may be used to diagnose latent tuberculosis. Liver, bone marrow, or lymph node biopsy may be diagnostic in some cases of miliary tuberculosis, disseminated fungal infection, sarcoidosis, or metastatic malignancy. Positron emission tomography with fluorodeoxyglucose may be useful in identifying a systemic site for biopsy in patients with suspected carcinomatous meningitis or sarcoidosis when other tests are unrevealing. Genetic testing can identify mutations that cause rare monogenic autoinflammatory disorders.

MENINGEAL BIOPSY

If CSF is not diagnostic then a meningeal biopsy should be strongly considered in patients who are severely disabled, who need chronic ventricular decompression, or whose illness is progressing rapidly. The activities of the surgeon, pathologist, microbiologist, and cytologist should be coordinated so that a large enough sample is obtained and the appropriate cultures and histologic and molecular studies, including electron-microscopic and PCR studies, are performed. The diagnostic yield of meningeal biopsy can be increased by targeting regions that enhance with contrast on MRI or CT. With current microsurgical techniques, most areas of the basal meninges can be accessed for biopsy via a limited craniotomy. In one series, MRI demonstrated meningeal enhancement in 47% of patients undergoing meningeal biopsy; biopsy of an enhancing region

was diagnostic in 80% of cases, biopsy of nonenhancing regions was diagnostic in only 9%, and sarcoid (31%) and metastatic adenocarcinoma (25%) were the most common conditions identified. Tuberculosis is the most common condition identified in many reports from outside the United States.

APPROACH TO THE ENIGMATIC CASE

In approximately one-third of cases, the diagnosis is not known despite careful evaluation of CSF and potential extraneural sites of disease. A number of the organisms that cause chronic meningitis may take weeks to be identified by cultures. In enigmatic cases, several options are available, determined by the extent of the clinical deficits and rate of progression. It is prudent to wait until cultures are finalized if the patient is asymptomatic or symptoms are mild and not progressive. Unfortunately, in many cases progressive neurologic deterioration occurs, and rapid treatment is required. Ventricular-peritoneal shunts may be placed to relieve hydrocephalus, but the risk of disseminating the undiagnosed inflammatory process into the abdomen must be considered.

Empirical Treatment

Diagnosis of the causative agent is essential because effective therapies exist for many etiologies of chronic meningitis, but if the condition is left untreated, progressive damage to the CNS and cranial nerves and roots is likely to occur. Occasionally, empirical therapy must be initiated when all attempts at diagnosis fail. In general, empirical therapy in the United States consists of antimycobacterial agents, amphotericin for fungal infection, and/or glucocorticoids for noninfectious inflammatory causes. It is important to direct empirical therapy of lymphocytic meningitis at tuberculosis, particularly if the condition is associated with low CSF glucose, since untreated disease can be devastating within weeks. Prolonged anti-tumor necrosis factor therapy and anti-programmed death-1 (PD-1) inhibitors can cause reactivation of TB, and such patients who develop chronic meningitis should be treated empirically with antituberculous therapy if the etiology is uncertain. In the Mayo Clinic series, the most useful empirical therapy was administration of glucocorticoids rather than antituberculous therapy. When proceeding with empiric glucocorticoids, caution should be maintained whenever a transient response to treatment is noted, as some infectious (e.g., tuberculosis and cysticercosis) and non-infectious (e.g., lymphoma) etiologies may temporarily respond to glucocorticoid monotherapy. Carcinomatous or lymphomatous meningitis may be difficult to diagnose initially, but the diagnosis becomes evident with time.

TABLE 134-2

Infectious Causes of Chronic Meningitis

Causative Agent	CSF Formula	Helpful Diagnostic Tests	Risk Factors and Systemic Manifestations
Common Bacterial Causes			
Partially treated suppurative meningitis	Mononuclear or mixed mononuclear-polymorphonuclear cells	CSF culture and Gram's stain; CSF 16s rRNA PCR	History consistent with acute bacterial meningitis and incomplete treatment
Parameningeal infection	Mononuclear or mixed mononuclear-polymorphonuclear cells	Contrast-enhanced CT or MRI to detect parenchymal, subdural, epidural, or sinus infection	Otitis media, pleuropulmonary infection, right-to-left cardiopulmonary shunt for brain abscess; focal neurologic signs; neck, back, ear, or sinus tenderness
Mycobacterium tuberculosis	Mononuclear cells except polymorphonuclear cells in early infection (commonly <500 WBC/μL); low CSF glucose; high protein	Tuberculin skin test may be negative; interferon gamma release assay; PCR and AFB culture of CSF (sputum, urine, gastric contents if indicated); identify tubercle bacillus on acid-fast stain of CSF or protein pellicle	Exposure history; previous tuberculous illness; immunosuppressed, anti-TNF therapy or AIDS; young children; fever, meningismus, night sweats, miliary TB on x-ray or liver biopsy; stroke due to arteritis
Lyme disease (Bannwarth's syndrome) *Borrelia burgdorferi*	Mononuclear cells; elevated protein	Serum Lyme antibody titer; western blot confirmation; (patients with syphilis may have false-positive Lyme titer)	History of tick bite or appropriate exposure history; erythema chronicum migrans skin rash; arthritis, radiculopathy, Bell's palsy, meningoencephalitis–multiple sclerosis-like syndrome

Causative Agent	CSF Formula	Helpful Diagnostic Tests	Risk Factors and Systemic Manifestations
Syphilis (secondary, tertiary) *Treponema pallidum*	Mononuclear cells; elevated protein	CSF VDRL; serum VDRL (or RPR); fluorescent treponemal antibody-absorbed (FTA) or MHA-TP; serum VDRL may be negative in tertiary syphilis	Appropriate exposure history; HIV-seropositive individuals at increased risk of aggressive infection; "dementia"; cerebral infarction due to endarteritis
Uncommon Bacterial Causes			
Actinomyces	Polymorphonuclear cells	Anaerobic culture	Parameningeal abscess or sinus tract (oral or dental focus); pneumonitis
Nocardia	Polymorphonuclear; occasionally mononuclear cells; often low glucose	Isolation may require weeks; weakly acid fast	Associated brain abscess may be present
Brucella	Mononuclear cells (rarely polymorphonuclear); elevated protein; often low glucose	CSF antibody detection; serum antibody detection	Intake of unpasteurized dairy products; exposure to goats, sheep, cows; fever, arthralgia, myalgia, vertebral osteomyelitis
Whipple's disease *Tropheryma whipplei*	Mononuclear cells	Biopsy of small bowel or lymph node; CSF PCR for *T. whipplei;* brain and meningeal biopsy (with PAS stain and EM examination)	Diarrhea, weight loss, arthralgias, fever; dementia, ataxia, paresis, ophthalmoplegia, oculomasticatory myoclonus
Rare Bacterial Causes			
Leptospirosis (occasionally if left untreated may last 3–4 weeks)			
Fungal Causes			

Causative Agent	CSF Formula	Helpful Diagnostic Tests	Risk Factors and Systemic Manifestations
Cryptococcus neoformans and var. gatti	Mononuclear cells; count not elevated in some patients with AIDS	India ink or fungal wet mount of CSF (budding yeast); blood and urine cultures; antigen detection in CSF	AIDS and immune suppression; pigeon exposure for neoformans, decaying wood exposure for var. gatti; skin and other organ involvement due to disseminated infection
Coccidioides immitis	Mononuclear cells (sometimes 10–20% eosinophils); often low glucose	Antibody detection in CSF and serum, antigen detection in CSF	Exposure history—southwestern United States; increased virulence in dark-skinned races
Candida sp.	Polymorphonuclear or mononuclear	Fungal stain and culture of CSF	IV drug abuse; postsurgery; prolonged IV therapy; disseminated candidiasis, recent epidural injection
Histoplasma capsulatum	Mononuclear cells; low glucose	Fungal stain and culture of large volumes of CSF; antigen detection in CSF, serum, and urine; antibody detection in serum, CSF	Exposure history—Ohio and central Mississippi River Valley; AIDS; mucosal lesions
Blastomyces dermatitidis	Mononuclear cells	Fungal stain and culture of CSF; biopsy and culture of skin, lung lesions; antibody detection in serum	Midwestern and southeastern United States; usually systemic infection; abscesses, draining sinus, ulcers
Aspergillus sp.	Mononuclear or polymorphonuclear	CSF culture	Sinusitis; granulocytopenia or immunosuppression
Sporothrix schenckii	Mononuclear cells	Antibody detection in CSF and serum; CSF culture	Traumatic inoculation; IV drug use; ulcerated skin lesion
Rare Fungal Causes			

Causative Agent	CSF Formula	Helpful Diagnostic Tests	Risk Factors and Systemic Manifestations
Xylohypha (formerly *Cladosporium*) *trichoides* and other dark-walled (dematiaceous) fungi such as *Curvularia; Drechslera; Mucor*; and, after water aspiration, *Pseudallescheria boydii;* iatrogenic *Exserohilum rostratum* infection following spinal blocks			
Protozoal Causes			
Toxoplasma gondii	Mononuclear cells	Biopsy or response to empirical therapy in clinically appropriate context (including presence of antibody in serum)	Usually with intracerebral abscesses; common in HIV-seropositive patients
Trypanosomiasis *Trypanosoma gambiense*, *T. rhodesiense*	Mononuclear cells; elevated protein	Elevated CSF IgM; identification of trypanosomes in CSF and blood smear	Endemic in Africa; chancre, lymphadenopathy; prominent sleep disorder
Rare Protozoal Causes			
Acanthamoeba sp. causing granulomatous amebic encephalitis and meningoencephalitis in immunocompromised and debilitated individuals. *Balamuthia mandrillaris* causing chronic meningoencephalitis in immunocompetent hosts.			
Helminthic Causes			
Cysticercosis (infection with cysts of *Taenia solium*)	Mononuclear cells; may have eosinophils; glucose level may be low	Indirect hemagglutination assay in CSF; ELISA immunoblotting in serum	Usually with multiple cysts in basal meninges and hydrocephalus; cerebral cysts, muscle calcification
Gnathostoma spinigerum	Eosinophils, mononuclear cells	Peripheral eosinophilia	History of eating raw fish; common in Thailand and Japan; subarachnoid hemorrhage; painful radiculopathy

Causative Agent	CSF Formula	Helpful Diagnostic Tests	Risk Factors and Systemic Manifestations
Angiostrongylus cantonensis	Eosinophils, mononuclear cells	Recovery of worms from CSF	History of eating raw shellfish; common in tropical Pacific regions; often benign
Baylisascaris procyonis (raccoon ascarid)	Eosinophils, mononuclear cells		Infection follows accidental ingestion of *B. procyonis* eggs from raccoon feces; fatal meningoencephalitis
Rare Helminthic Causes			
Trichinella spiralis (trichinosis); *Fasciola hepatica* (liver fluke), *Echinococcus* cysts; *Schistosoma* sp. The former may produce a lymphocytic pleocytosis whereas the latter two may produce an eosinophilic response in CSF associated with cerebral cysts (*Echinococcus*) or granulomatous lesions of brain or spinal cord			
Viral Causes			
Mumps	Mononuclear cells	Antibody in serum	No prior mumps or immunization; may produce meningoencephalitis; may persist for 3–4 weeks
Lymphocytic choriomeningitis	Mononuclear cells; may have low glucose	Antibody in serum; PCR for LCMV in CSF	Contact with rodents or their excreta; may persist for 3–4 weeks
Echovirus	Mononuclear cells; may have low glucose	Virus isolation from CSF	Congenital hypogammaglobulinemia; history of recurrent meningitis

Causative Agent	CSF Formula	Helpful Diagnostic Tests	Risk Factors and Systemic Manifestations
HIV (acute retroviral syndrome)	Mononuclear cells	PCR for HIV in blood and CSF	HIV risk factors; rash, fever, lymphadenopathy; lymphopenia in peripheral blood; syndrome may persist long enough to be considered as “chronic meningitis”; or chronic meningitis may develop in later stages (AIDS) due to HIV
Human herpes viruses	Mononuclear cells	PCR for HSV, CMV DNA; CSF antibody for HSV, EBV	Recurrent meningitis due to HSV-2 (rarely HSV-1) often associated with genital recurrences; EBV associated with myeloradiculopathy, CMV with polyradiculopathy

Abbreviations: AFB, acid-fast bacillus; CMV, cytomegalovirus; CSF, cerebrospinal fluid; CT, computed tomography; EBV, Epstein-Barr virus; ELISA, enzyme-linked immunosorbent assay; EM, electron microscopy; FTA, fluorescent treponemal antibody absorption test; HSV, herpes simplex virus; MHA-TP, microhemagglutination assay–*T. pallidum*; MRI, magnetic resonance imaging; PAS, periodic acid–Schiff; PCR, polymerase chain reaction; RPR, rapid plasma reagin test; TB, tuberculosis; VDRL, Venereal Disease Research Laboratories test.

TABLE 134-3

Noninfectious Causes of Chronic Meningitis

Causative Agents	CSF Formula	Helpful Diagnostic Tests	Risk Factors and Systemic Manifestations
Malignancy	Mononuclear cells; elevated protein; low glucose	Repeated cytologic examination of large volumes of CSF; CSF exam by polarizing microscopy; clonal lymphocyte markers; deposits on nerve roots or meninges seen on myelogram or contrast-enhanced MRI; meningeal biopsy	Metastatic cancer of breast, lung, stomach, or pancreas; melanoma, lymphoma, leukemia; meningeal gliomatosis; sarcoma; cerebral dysgerminoma
Chemical compounds (may cause recurrent meningitis)	Mononuclear or PMNs; low glucose, elevated protein; xanthochromia from subarachnoid hemorrhage in week prior to presentation with “meningitis”	Contrast-enhanced CT scan or MRI; cerebral angiogram to detect aneurysm. Enhancement and clumping of nerve roots of the cauda equina in arachnoiditis/pachymeningitis	History of recent injection into the subarachnoid space; history of sudden onset of headache; recent resection of acoustic neuroma or craniopharyngioma; epidermoid tumor of brain or spine, sometimes with dermoid sinus tract; pituitary apoplexy
Primary Inflammation			
CNS sarcoidosis	Mononuclear cells; elevated protein; often low glucose	Serum and CSF angiotensin-converting enzyme levels; biopsy of extraneural affected tissues or brain lesion/meningeal biopsy	CN palsy, especially CN VII; hypothalamic dysfunction, especially diabetes insipidus; abnormal chest radiograph; peripheral neuropathy or myopathy

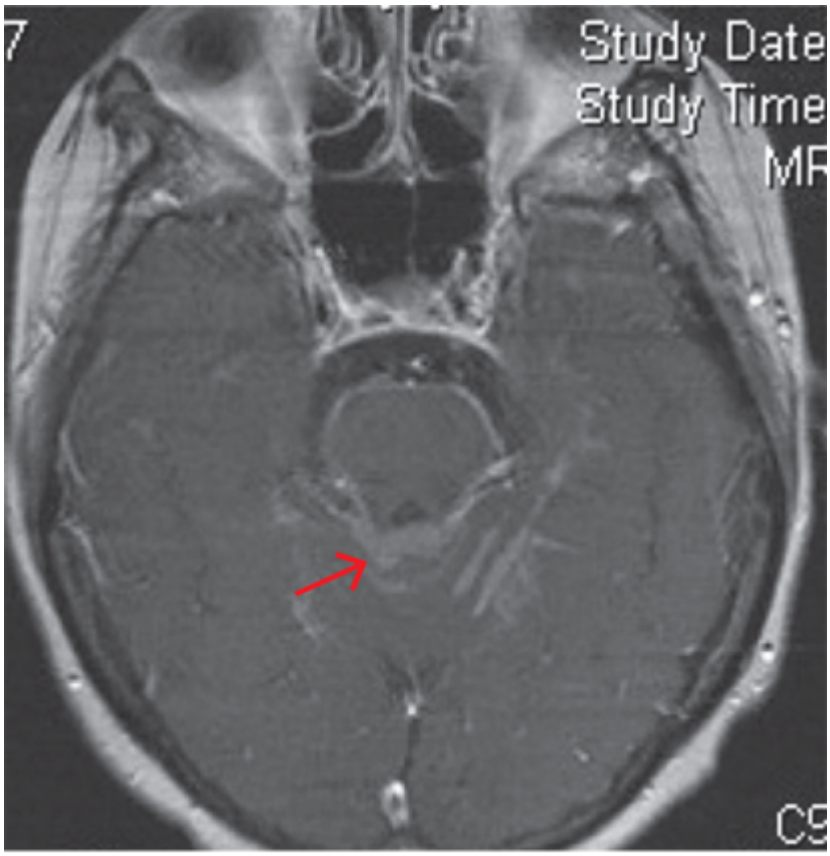
Causative Agents	CSF Formula	Helpful Diagnostic Tests	Risk Factors and Systemic Manifestations
Vogt-Koyanagi-Harada syndrome (recurrent meningitis)	Mononuclear cells		Recurrent meningoencephalitis with uveitis, retinal detachment, alopecia, lightening of eyebrows and lashes, dysacousia, cataracts, glaucoma
Isolated granulomatous angiitis of the nervous system	Mononuclear cells; elevated protein	Angiography; meningeal biopsy may be necessary if confined to small vessels. VZV PCR in blood and biopsy tissue	Subacute dementia; multiple cerebral infarctions; recent zoster ophthalmicus
Systemic lupus erythematosus	Mononuclear or PMNs	Anti-DNA antibody, antinuclear antibodies	Encephalopathy; seizures; stroke; transverse myelopathy; rash; arthritis
Behçet's syndrome (recurrent meningitis)	Mononuclear or PMNs; elevated protein		Oral and genital aphthous ulcers; iridocyclitis; retinal hemorrhages; pathergic lesions at site of skin puncture
Chronic benign lymphocytic meningitis	Mononuclear cells		Recovery in 2–6 months, diagnosis by exclusion
Mollaret's meningitis (recurrent meningitis)	Large endothelial cells and PMNs in first hours, followed by mononuclear cells	PCR for herpes; MRI/CT to rule out epidermoid tumor or dural cyst	Recurrent meningitis; exclude HSV-2; rare cases due to HSV-1; occasional case associated with dural cyst

Causative Agents	CSF Formula	Helpful Diagnostic Tests	Risk Factors and Systemic Manifestations
Drug hypersensitivity	PMNs; occasionally mononuclear cells or eosinophils	Complete blood count (eosinophilia)	Exposure to nonsteroidal anti-inflammatory agents, sulfonamides, isoniazid, tolmetin, ciprofloxacin, penicillin, carbamazepine, lamotrigine, IV immunoglobulin, OKT3 antibodies, phenazopyridine; improvement after discontinuation of drug; recurrence with repeat exposure
Granulomatosis with polyangiitis (Wegener's)	Mononuclear cells	Chest and sinus radiographs; urinalysis; ANCA antibodies in serum	Associated sinus, pulmonary, or renal lesions; CN palsies; skin lesions; peripheral neuropathy
Neonatal-Onset Multisystem Inflammatory Disorder	Mononuclear and PMNs	Gain of function mutation in NLRP3 gene leading to elevated IL-1β	Recurrent fever, urticaria, arthralgia, sensorineural hearing loss, papilledema, increased ICP
IgG4-Related Hypertrophic Pachymeningitis	Mild lymphocytic pleocytosis in some cases; normal to mildly increased protein; normal glucose	Serum IgG4 levels frequently elevated; ESR and C-reactive protein; meningeal biopsy shows swirling "storiform" fibrosis with lymphocytic infiltrates, obliterative phlebitis and IgG4+ plasma cells	Headache; seizures; focal symptoms from dural involvement in spinal cord/nerve roots, clivus, periorbital, vestibular and brainstem structures. Systemic IgG4-related disease can involve many tissues including pancreas, thyroid, lungs, and retroperitoneum
Other: multiple sclerosis, Sjögren's syndrome, and rarer forms of vasculitis (e.g., Cogan's syndrome)			

Abbreviations: ANCA, antineutrophil cytoplasmic antibodies; CN, cranial nerve; CSF, cerebrospinal fluid; CT, computed tomography; HSV, herpes simplex virus; MRI, magnetic resonance imaging; PCR, polymerase chain reaction; PMNs, polymorphonuclear cells.

FIGURE 134-1

Chronic meningitis illustrating meningeal enhancement on contrast MRI scan. ***A*** and ***B*** are images from a patient with chronic meningitis due to carcinoma. ***C*** and ***D*** are from a patient with chronic meningitis due to Cryptococcus infection. *Arrows* point to the most prominent areas of meningeal inflammation around the brainstem and cerebellar folia (***A***), cerebellum (***C***), along the dorsal spinal cord (***B***), and clumping of roots in the cauda equina (***D***).



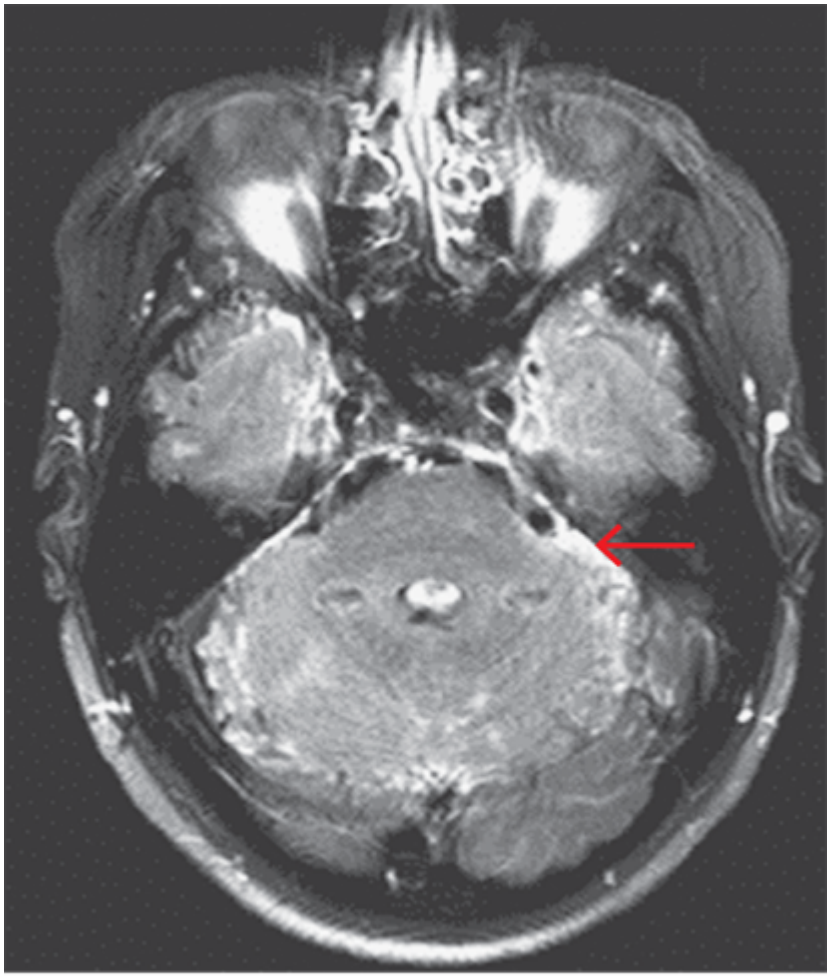
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE IMMUNOSUPPRESSED PATIENT

Chronic meningitis is not uncommon in the course of HIV infection. Pleocytosis and mild meningeal signs often occur at the onset of HIV infection, and occasionally low-grade meningitis persists. Toxoplasmosis commonly presents as intracranial abscesses and also may be associated with meningitis. Other important causes of chronic meningitis in AIDS include infection with *Cryptococcus*, *Nocardia*, *Candida*, or other fungi; syphilis; and lymphoma (Fig. 134-1). Toxoplasmosis, cryptococcosis, nocardiosis, and other fungal infections are important etiologic considerations in individuals with immunodeficiency states other than AIDS, including those due to immunosuppressive medications. Because of the increased risk of chronic meningitis and the attenuation of clinical signs of meningeal irritation in immunosuppressed individuals, CSF examination should be performed for any persistent headache or unexplained change in mental state.

FURTHER READING

Baldwin K, Whiting C: Chronic meningitis: Simplifying a diagnostic challenge. Curr Neurol Neurosci Rep 16:30, 2016.

Cheng TM et al: Chronic meningitis: The role of meningeal or cortical biopsy. Neurosurgery 34:590, 1994.

Le LT, Spudich SS: HIV-associated neurologic disorders and central nervous system opportunistic infections in HIV. Semin Neurol 36:373, 2016.

Lu LX et al: IgG4-related hypertrophic pachymeningitis: Clinical features, diagnostic criteria and treatment. JAMA Neurol 71:785, 2014.

Torok ME: Tuberculous meningitis: Advances in diagnosis and treatment. Br Med Bull 113:117, 2015.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 135: Brain Abscess and Empyema

Karen L. Roos; Kenneth L. Tyler

BRAIN ABSCESS

DEFINITION

A brain abscess is a focal, suppurative infection within the brain parenchyma, typically surrounded by a vascularized capsule. The term *cerebritis* is often employed to describe a nonencapsulated brain abscess.

EPIDEMIOLOGY


A bacterial brain abscess is a relatively uncommon intracranial infection, with an incidence of ~0.3–1.3:100,000 persons per year. Predisposing conditions include otitis media and mastoiditis, paranasal sinusitis, pyogenic infections in the chest or other body sites, penetrating head trauma or neurosurgical procedures, and dental infections. In immunocompetent individuals the most important pathogens are *Streptococcus* spp. (anaerobic, aerobic, and viridans [40%]), Enterobacteriaceae (*Proteus* spp., *E. coli* sp., *Klebsiella* spp. [25%]), anaerobes (e.g., *Bacteroides* spp., *Fusobacterium* spp. [30%]), and staphylococci (10%). In immunocompromised hosts with underlying HIV infection, organ transplantation, cancer, or immunosuppressive therapy, most brain abscesses are caused by *Nocardia* spp., *Toxoplasma gondii, Aspergillus* spp., *Candida* spp., and *C. neoformans*. In Latin America and in immigrants from Latin America, the most common cause of brain abscess is *Taenia solium* (neurocysticercosis). In India and East Asia, mycobacterial infection (tuberculoma) remains a major cause of focal CNS mass lesions.

ETIOLOGY

A brain abscess may develop (1) by direct spread from a contiguous cranial site of infection, such as paranasal sinusitis, otitis media, mastoiditis, or dental infection; (2) following head trauma or a neurosurgical procedure; or (3) as a result of hematogenous spread from a remote site of infection. In up to 25% of cases, no obvious primary source of infection is apparent (cryptogenic brain abscess).

Approximately one-third of brain abscesses are associated with otitis media and mastoiditis, often with an associated cholesteatoma. Otogenic abscesses occur predominantly in the temporal lobe (55–75%) and cerebellum (20–30%). In some series, up to 90% of cerebellar abscesses are otogenic. Common organisms include streptococci, *Bacteroides* spp., *Pseudomonas* spp., *Haemophilus* spp., and Enterobacteriaceae. Abscesses that develop as a result of direct spread of infection from the frontal, ethmoidal, or sphenoidal sinuses and those that occur due to dental infections are usually located in the frontal lobes. Approximately 10% of brain abscesses are associated with paranasal sinusitis, and this association is particularly strong in young males in their second and third decades of life. The most common pathogens in brain abscesses associated with paranasal sinusitis are streptococci (especially *Streptococcus milleri*), *Haemophilus* spp., *Bacteroides* spp., *Pseudomonas* spp., and *Staphylococcus aureus*. Dental infections are associated with ~2% of brain abscesses, although it is often suggested that many "cryptogenic" abscesses are in fact due to dental infections. The most common pathogens in this setting are streptococci, staphylococci, *Bacteroides* spp., and *Fusobacterium* spp.

Hematogenous abscesses account for ~25% of brain abscesses. Hematogenous abscesses are often multiple, and multiple abscesses often (50%) have a hematogenous origin. These abscesses show a predilection for the territory of the middle cerebral artery (i.e., posterior frontal or parietal lobes). Hematogenous abscesses are often located at the junction of the gray and white matter and are often poorly encapsulated. The microbiology of hematogenous abscesses is dependent on the primary source of infection. For example, brain abscesses that develop as a complication of infective endocarditis are often due to viridans streptococci or *S. aureus*. Abscesses associated with pyogenic lung infections such as lung abscess or bronchiectasis are often due to streptococci, staphylococci, *Bacteroides* spp., *Fusobacterium* spp., or Enterobacteriaceae. Abscesses that follow penetrating head trauma or neurosurgical procedures are frequently due to methicillin-resistant *S. aureus* (MRSA), *S. epidermidis*, Enterobacteriaceae, *Pseudomonas* spp., and *Clostridium* spp. Enterobacteriaceae and *P. aeruginosa* are important causes of abscesses associated with urinary sepsis. Congenital cardiac malformations that produce a right-to-left shunt, such as tetralogy of Fallot, patent ductus arteriosus, and atrial and ventricular septal defects, allow bloodborne bacteria to bypass the pulmonary capillary bed and reach the brain. Similar phenomena can occur with pulmonary arteriovenous malformations. The decreased arterial

oxygenation and saturation from the right-to-left shunt and polycythemia may cause focal areas of cerebral ischemia, thus providing a nidus for microorganisms that bypassed the pulmonary circulation to multiply and form an abscess. Streptococci are the most common pathogens in this setting.

PATHOGENESIS AND HISTOPATHOLOGY

Results of experimental models of brain abscess formation suggest that for bacterial invasion of brain parenchyma to occur, there must be preexisting or concomitant areas of ischemia, necrosis, or hypoxemia in brain tissue. The intact brain parenchyma is relatively resistant to infection. Once bacteria have established infection, brain abscess frequently evolves through a series of stages, influenced by the nature of the infecting organism and by the immunocompetence of the host. The early cerebritis stage (days 1–3) is characterized by a perivascular infiltration of inflammatory cells, which surround a central core of coagulative necrosis. Marked edema surrounds the lesion at this stage. In the late cerebritis stage (days 4–9), pus formation leads to enlargement of the necrotic center, which is surrounded at its border by an inflammatory infiltrate of macrophages and fibroblasts. A thin capsule of fibroblasts and reticular fibers gradually develops, and the surrounding area of cerebral edema becomes more distinct than in the previous stage. The third stage, early capsule formation (days 10–13), is characterized by the formation of a capsule that is better developed on the cortical than on the ventricular side of the lesion. This stage correlates with the appearance of a ring-enhancing capsule on neuroimaging studies. The final stage, late capsule formation (day 14 and beyond), is defined by a well-formed necrotic center surrounded by a dense collagenous capsule. The surrounding area of cerebral edema has regressed, but marked gliosis with large numbers of reactive astrocytes has developed outside the capsule. This gliotic process may contribute to the development of seizures as a sequela of brain abscess.

CLINICAL PRESENTATION

A brain abscess typically presents as an expanding intracranial mass lesion rather than as an infectious process. Although the evolution of signs and symptoms is extremely variable, ranging from hours to weeks or even months, most patients present to the hospital 11–12 days following onset of symptoms. The classic clinical triad of headache, fever, and a focal neurologic deficit is present in <50% of cases. The most common symptom in patients with a brain abscess is headache, occurring in >75% of patients. The headache is often characterized as a constant, dull, aching sensation, either hemicranial or generalized, and it becomes progressively more severe and refractory to therapy. Fever is present in only 50% of patients at the time of diagnosis, and its absence should not exclude the diagnosis. The new onset of focal or generalized seizure activity is a presenting sign in 15–35% of patients. Focal neurologic deficits including hemiparesis, aphasia, or visual field defects are part of the initial presentation in >60% of patients.

The clinical presentation of a brain abscess depends on its location, the nature of the primary infection if present, and the level of the ICP. Hemiparesis is the most common localizing sign of a frontal lobe abscess. A temporal lobe abscess may present with a disturbance of language (dysphasia) or an upper homonymous quadrantanopia. Nystagmus and ataxia are signs of a cerebellar abscess. Signs of raised ICP—papilledema, nausea and vomiting, and drowsiness or confusion—can be the dominant presentation of some abscesses, particularly those in the cerebellum. Meningismus is not present unless the abscess has ruptured into the ventricle or the infection has spread to the subarachnoid space.

DIAGNOSIS

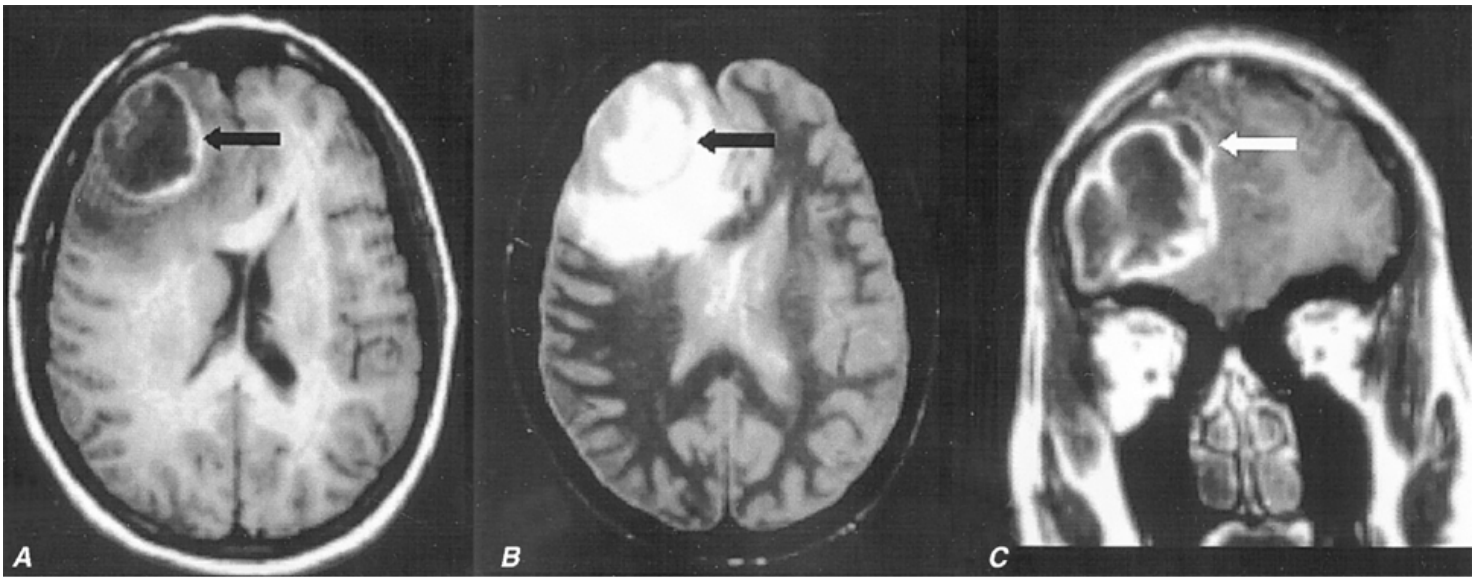
Diagnosis is made by neuroimaging studies. MRI **(Fig. 135-1)** is better than CT for demonstrating abscesses in the early (cerebritis) stages and is superior to CT for identifying abscesses in the posterior fossa. Cerebritis appears on MRI as an area of low-signal intensity on T1-weighted images with irregular postgadolinium enhancement and as an area of increased signal intensity on T2-weighted images. Cerebritis is often not visualized by CT scan, but when present, appears as an area of hypodensity. On a contrast-enhanced CT scan, a mature brain abscess appears as a focal area of hypodensity surrounded by ring enhancement with surrounding edema (hypodensity). On contrast-enhanced T1-weighted MRI, a mature brain abscess has a capsule that enhances surrounding a hypodense center and surrounded by a hypodense area of edema. On T2-weighted MRI, there is a hyperintense central area of pus surrounded by a well-defined hypointense capsule and a hyperintense surrounding area of edema. It is important to recognize that the CT and MRI appearance, particularly of the capsule, may be altered by treatment with glucocorticoids. The distinction between a brain abscess and other focal CNS lesions such as primary or metastatic tumors may be facilitated by the use of diffusion-weighted imaging sequences on which a brain abscess typically shows increased signal due to restricted diffusion of the abscess cavity with corresponding low signal on apparent diffusion coefficient images.

Microbiologic diagnosis of the etiologic agent is most accurately determined by Gram's stain and culture of abscess material obtained by CT-guided stereotactic needle aspiration. Aerobic and anaerobic bacterial cultures and mycobacterial and fungal cultures should be obtained. Up to

10% of patients will also have positive blood cultures. LP should not be performed in patients with known or suspected focal intracranial infections such as abscess or empyema; CSF analysis contributes nothing to diagnosis or therapy, and LP increases the risk of herniation.

FIGURE 135-1

Pneumococcal brain abscess. Note that the abscess wall has hyperintense signal on the axial T1-weighted magnetic resonance imaging (MRI) (***A***, *black arrow*), hypointense signal on the axial proton density images (***B***, *black arrow*), and enhances prominently after gadolinium administration on the coronal T1-weighted image (***C***). The abscess is surrounded by a large amount of vasogenic edema and has a small "daughter" abscess (***C***, *white arrow*). (*Courtesy of Joseph Lurito, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Additional laboratory studies may provide clues to the diagnosis of brain abscess in patients with a CNS mass lesion. About 50% of patients have a peripheral leukocytosis, 60% an elevated ESR, and 80% an elevated C-reactive protein. Blood cultures are positive in ~10% of cases overall but may be positive in >85% of patients with abscesses due to *Listeria*.

DIFFERENTIAL DIAGNOSIS

Conditions that can cause headache, fever, focal neurologic signs, and seizure activity include brain abscess, subdural empyema, bacterial meningitis, viral meningoencephalitis, superior sagittal sinus thrombosis, and acute disseminated encephalomyelitis. When fever is absent, primary and metastatic brain tumors become the major differential diagnosis. Less commonly, cerebral infarction or hematoma can have an MRI or CT appearance resembling brain abscess.

TREATMENT

TREATMENT

Brain Abscess

Optimal therapy of brain abscesses involves a combination of high-dose parenteral antibiotics and neurosurgical drainage. Empirical therapy of community-acquired brain abscess in an immunocompetent patient typically includes a third- or fourth-generation cephalosporin (e.g., cefotaxime, ceftriaxone, or cefepime) and metronidazole (**see Table 133-1 for antibiotic dosages**). In patients with penetrating head trauma or recent neurosurgical procedures, treatment should include ceftazidime as the third-generation cephalosporin to enhance coverage of *Pseudomonas* spp. and vancomycin for coverage of staphylococci. Meropenem plus vancomycin also provides good coverage in this setting.

Aspiration and drainage of the abscess under stereotactic guidance are beneficial for both diagnosis and therapy. Empirical antibiotic coverage should be modified based on the results of Gram's stain and culture of the abscess contents. Complete excision of a bacterial abscess via craniotomy or craniectomy is generally reserved for multiloculated abscesses or those in which stereotactic aspiration is unsuccessful.

Medical therapy alone is not optimal for treatment of brain abscess and should be reserved for patients whose abscesses are neurosurgically inaccessible, for patients with small (<2–3 cm) or nonencapsulated abscesses (cerebritis), and for patients whose condition is too tenuous to

allow performance of a neurosurgical procedure. All patients should receive a minimum of 6–8 weeks of parenteral antibiotic therapy. The role, if any, of supplemental oral antibiotic therapy following completion of a standard course of parenteral therapy has never been adequately studied.

In addition to surgical drainage and antibiotic therapy, patients should receive prophylactic anticonvulsant therapy because of the high risk (~35%) of focal or generalized seizures. Anticonvulsant therapy is continued for at least 3 months after resolution of the abscess, and decisions regarding withdrawal are then based on the EEG. If the EEG is abnormal, anticonvulsant therapy should be continued. If the EEG is normal, anticonvulsant therapy can be slowly withdrawn, with close follow-up and repeat EEG after the medication has been discontinued.

Glucocorticoids should not be given routinely to patients with brain abscesses. Intravenous dexamethasone therapy (10 mg every 6 h) is usually reserved for patients with substantial periabscess edema and associated mass effect and increased ICP. Dexamethasone should be tapered as rapidly as possible to avoid delaying the natural process of encapsulation of the abscess.

Serial MRI or CT scans should be obtained on a monthly or twice-monthly basis to document resolution of the abscess. More frequent studies (e.g., weekly) are probably warranted in the subset of patients who are receiving antibiotic therapy alone. A small amount of enhancement may remain for months after the abscess has been successfully treated.

PROGNOSIS

The mortality rate of brain abscess has declined in parallel with the development of enhanced neuroimaging techniques, improved neurosurgical procedures for stereotactic aspiration, and improved antibiotics. In modern series, the mortality rate is typically <15%. Significant sequelae, including seizures, persisting weakness, aphasia, or mental impairment, occur in ≥20% of survivors.

NONBACTERIAL CAUSES OF INFECTIOUS FOCAL CNS LESIONS

ETIOLOGY

Neurocysticercosis is the most common parasitic disease of the CNS worldwide. Humans acquire cysticercosis by the ingestion of food contaminated with the eggs of the parasite *T. solium*. Toxoplasmosis is a parasitic disease caused by *T. gondii* and acquired from the ingestion of undercooked meat and from handling cat feces.

CLINICAL PRESENTATION

The most common manifestation of neurocysticercosis is new-onset partial seizures with or without secondary generalization. Cysticerci may develop in the brain parenchyma and cause seizures or focal neurologic deficits. When present in the subarachnoid or ventricular spaces, cysticerci can produce increased ICP by interference with CSF flow. Spinal cysticerci can mimic the presentation of intraspinal tumors. When the cysticerci first lodge in the brain, they frequently cause little in the way of an inflammatory response. As the cysticercal cyst degenerates, it elicits an inflammatory response that may present clinically as a seizure. Eventually the cyst dies, a process that may take several years and is typically associated with resolution of the inflammatory response and, often, abatement of seizures.

Primary *Toxoplasma* infection is often asymptomatic. However, during this phase parasites may spread to the CNS, where they become latent. Reactivation of CNS infection is almost exclusively associated with immunocompromised hosts, particularly those with HIV infection. During this phase patients present with headache, fever, seizures, and focal neurologic deficits.

DIAGNOSIS

The lesions of neurocysticercosis are readily visualized by MRI or CT scans. Lesions with viable parasites appear as cystic lesions. The scolex can often be visualized on MRI. Lesions may appear as contrast-enhancing lesions surrounded by edema. A very early sign of cyst death is hypointensity of the vesicular fluid on T2-weighted images when compared with CSF. Parenchymal brain calcifications are the most common finding and evidence that the parasite is no longer viable. MRI findings of toxoplasmosis consist of multiple lesions in the deep white matter, the thalamus, and basal ganglia and at the gray-white junction in the cerebral hemispheres. With contrast administration, the majority of the lesions enhance in a ringed, nodular, or homogeneous pattern and are surrounded by edema. In the presence of the characteristic neuroimaging abnormalities of *T. gondii* infection, serum IgG antibody to *T. gondii* should be obtained and, when positive, the patient should be treated.

TREATMENT

TREATMENT

Anticonvulsant therapy is initiated when the patient with neurocysticercosis presents with a seizure. There is controversy about whether or not anthelmintic therapy should be given to all patients, and recommendations are based on the stage of the lesion. Cysticerci appearing as cystic lesions in the brain parenchyma with or without pericystic edema or in the subarachnoid space at the convexity of the cerebral hemispheres should be treated with anticysticidal therapy. Cysticidal drugs accelerate the destruction of the parasites, resulting in a faster resolution of the infection. Albendazole and praziquantel are used in the treatment of neurocysticercosis. Approximately 85% of parenchymal cysts are destroyed by a single course of albendazole, and ~75% are destroyed by a single course of praziquantel. The dose of albendazole is 15 mg/kg per day in two doses for 8 days. The dose of praziquantel is 50 mg/kg per day for 15 days, although a number of other dosage regimens are also frequently cited. Prednisone or dexamethasone is given with anticysticidal therapy to reduce the host inflammatory response to degenerating parasites. Only cysts in the vesicular stage, where the cyst contains living larva (scolex seen on CT or MRI), and cysts in the colloidal stage as the larva degnerates (edema surrounds the lesion), are treated with anticysticidal therapy. Some, but not all, experts recommend anticysticidal therapy for lesions that are in the “granulo-nodular” stage (surrounded by a contrast-enhancing ring). There is universal agreement that calcified lesions do not need to be treated with anticysticidal therapy. Antiepileptic therapy can be stopped once the follow-up CT scan shows resolution of the lesion. Long-term antiepileptic therapy is recommended when seizures occur after resolution of edema and resorption or calcification of the degenerating cyst.

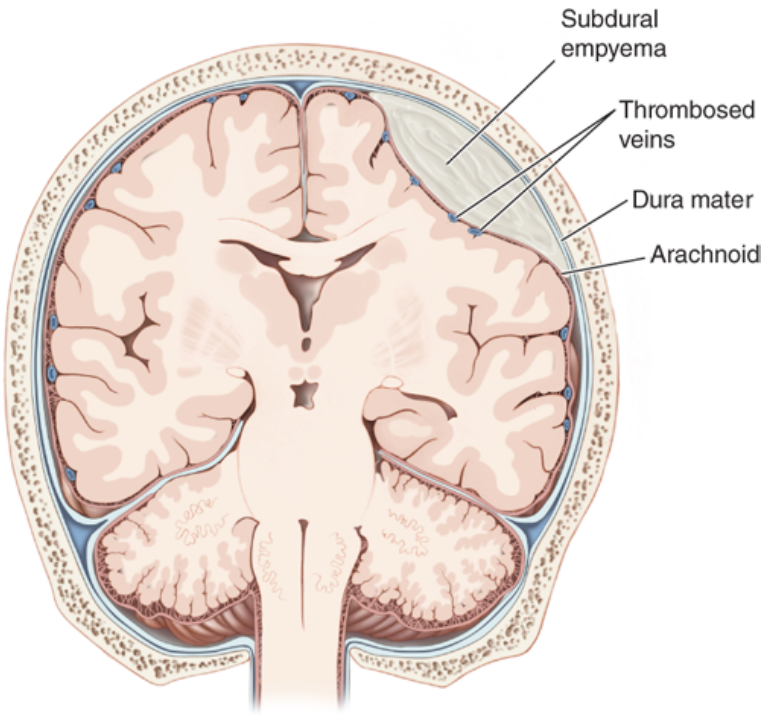
CNS toxoplasmosis is treated with a combination of sulfadiazine, 1.5–2.0 g orally qid, plus pyrimethamine, 100 mg orally to load, then 75–100 mg orally qd, plus folinic acid, 10–15 mg orally qd. Folinic acid is added to the regimen to prevent megaloblastic anemia. Therapy is continued until there is no evidence of active disease on neuroimaging studies, which typically takes at least 6 weeks, and then the dose of sulfadiazine is reduced to 2–4 g/d and pyrimethamine to 50 mg/d. Clindamycin plus pyrimethamine is an alternative therapy for patients who cannot tolerate sulfadiazine, but the combination of pyrimethamine and sulfadiazine is more effective.

SUBDURAL EMPYEMA

A subdural empyema (SDE) is a collection of pus between the dura and arachnoid membranes **(Fig. 135-2)**.

FIGURE 135-2

Subdural empyema.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPIDEMIOLOGY

SDE is a rare disorder that accounts for 15–25% of focal suppurative CNS infections. Sinusitis is the most common predisposing condition and typically involves the frontal sinuses, either alone or in combination with the ethmoid and maxillary sinuses. Sinusitis-associated empyema has a striking predilection for young males, possibly reflecting sex-related differences in sinus anatomy and development. It has been suggested that SDE may complicate 1–2% of cases of frontal sinusitis severe enough to require hospitalization. As a consequence of this epidemiology, SDE shows an ~3:1 male/female predominance, with 70% of cases occurring in the second and third decades of life. SDE may also develop as a complication of head trauma or neurosurgery. Secondary infection of a subdural effusion may also result in empyema, although secondary infection of hematomas, in the absence of a prior neurosurgical procedure, is rare.

ETIOLOGY

Aerobic and anaerobic streptococci, staphylococci, Enterobacteriaceae, and anaerobic bacteria are the most common causative organisms of sinusitis-associated SDE. Staphylococci and gram-negative bacilli are often the etiologic organisms when SDE follows neurosurgical procedures or head trauma. Up to one-third of cases are culture-negative, possibly reflecting difficulty in obtaining adequate anaerobic cultures.

PATHOPHYSIOLOGY

Sinusitis-associated SDE develops as a result of either retrograde spread of infection from septic thrombophlebitis of the mucosal veins draining the sinuses or contiguous spread of infection to the brain from osteomyelitis in the posterior wall of the frontal or other sinuses. SDE may also develop from direct introduction of bacteria into the subdural space as a complication of a neurosurgical procedure. The evolution of SDE can be extremely rapid because the subdural space is a large compartment that offers few mechanical barriers to the spread of infection. In patients with sinusitis-associated SDE, suppuration typically begins in the upper and anterior portions of one cerebral hemisphere and then extends posteriorly. SDE is often associated with other intracranial infections, including epidural empyema (40%), cortical thrombophlebitis (35%), and intracranial abscess or cerebritis (>25%). Cortical venous infarction produces necrosis of underlying cerebral cortex and subcortical white matter, with focal neurologic deficits and seizures (see below).

CLINICAL PRESENTATION

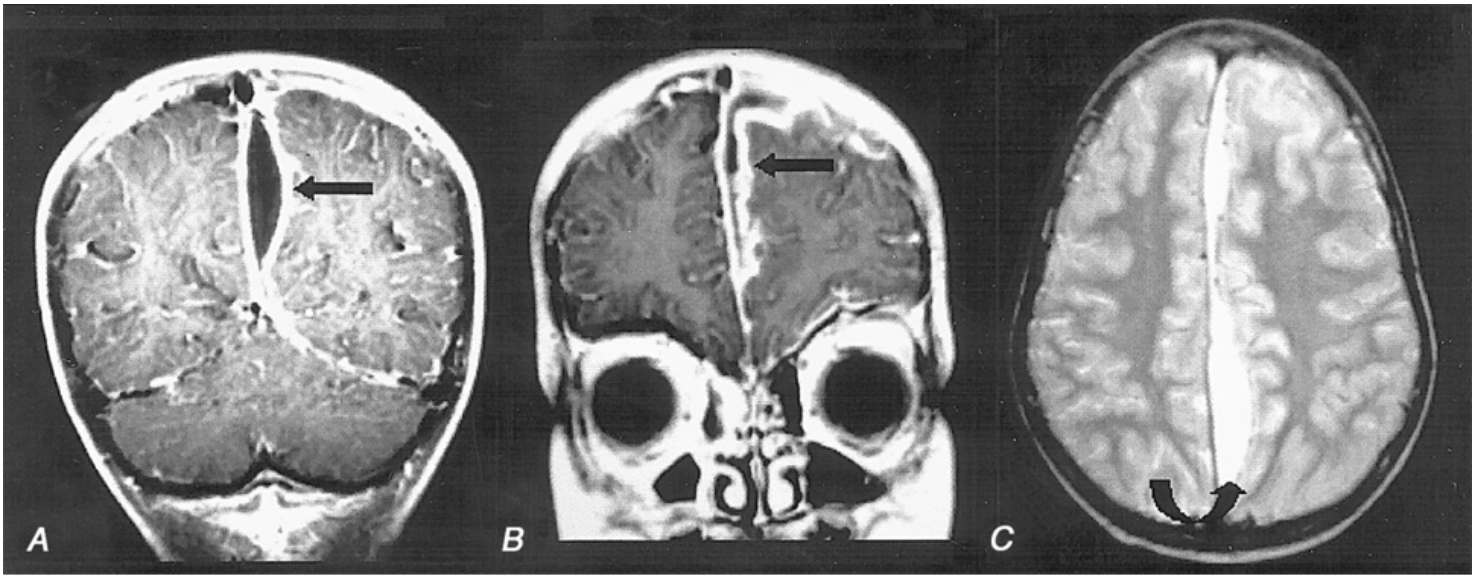
A patient with SDE typically presents with fever and a progressively worsening headache. The diagnosis of SDE should always be suspected in a patient with known sinusitis who presents with new CNS signs or symptoms. Patients with underlying sinusitis frequently have symptoms related to this infection. As the infection progresses, focal neurologic deficits, seizures, nuchal rigidity, and signs of increased ICP commonly occur. Headache is the most common complaint at the time of presentation; initially it is localized to the side of the subdural infection, but then it becomes more severe and generalized. Contralateral hemiparesis or hemiplegia is the most common focal neurologic deficit and can occur from the direct effects of the SDE on the cortex or as a consequence of venous infarction. Seizures begin as partial motor seizures that then become secondarily generalized. Seizures may be due to the direct irritative effect of the SDE on the underlying cortex or result from cortical venous infarction (see above). In untreated SDE, the increasing mass effect and increase in ICP cause progressive deterioration in consciousness, leading ultimately to coma.

DIAGNOSIS

MRI (Fig. 135-3) is superior to CT in identifying SDE and any associated intracranial infections. The administration of gadolinium greatly improves diagnosis by enhancing the rim of the empyema and allowing the empyema to be clearly delineated from the underlying brain parenchyma. Cranial MRI is also extremely valuable in identifying sinusitis, other focal CNS infections, cortical venous infarction, cerebral edema, and cerebritis. CT may show a crescent-shaped hypodense lesion over one or both hemispheres or in the interhemispheric fissure. Frequently the degree of mass effect, exemplified by midline shift, ventricular compression, and sulcal effacement, is far out of proportion to the mass of the SDE.

FIGURE 135-3

Subdural empyema. There is marked enhancement of the dura and leptomeninges (***A***, ***B***, *straight arrows*) along the left medial hemisphere. The pus is hypointense on T1-weighted images (***A***, ***B***) but markedly hyperintense on the proton density–weighted (***C***, *curved arrow*) image. *(Courtesy of Joseph Lurito, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CSF examination should be avoided in patients with known or suspected SDE because it adds no useful information and is associated with the risk of cerebral herniation.

DIFFERENTIAL DIAGNOSIS

The differential diagnosis of the combination of headache, fever, focal neurologic signs, and seizure activity that progresses rapidly to an altered level of consciousness includes subdural hematoma, bacterial meningitis, viral encephalitis, brain abscess, superior sagittal sinus thrombosis, and acute disseminated encephalomyelitis. The presence of nuchal rigidity is unusual with brain abscess or epidural empyema and should suggest the possibility of SDE when associated with significant focal neurologic signs and fever. Patients with bacterial meningitis also have nuchal rigidity but do not typically have focal deficits of the severity seen with SDE.

TREATMENT

TREATMENT

Subdural Empyema

SDE is a medical emergency. Emergent neurosurgical evacuation of the empyema, either through craniotomy, craniectomy, or burr-hole drainage, is the definitive step in the management of this infection. Empirical antimicrobial therapy for community-acquired SDE should include a combination of a third-generation cephalosporin (e.g., cefotaxime or ceftriaxone), vancomycin, and metronidazole (**see Table 133-1 for dosages**). Patients with hospital-acquired SDE may have infections due to *Pseudomonas* spp. or MRSA and should receive coverage with a carbapenem (e.g., meropenem) and vancomycin. Metronidazole is not necessary for antianaerobic therapy when meropenem is being used. Parenteral antibiotic therapy should be continued for a minimum of 3–4 weeks after SDE drainage. Patients with associated cranial osteomyelitis may require longer therapy. Specific diagnosis of the etiologic organisms is made based on Gram's stain and culture of fluid obtained via either burr holes or craniotomy; the initial empirical antibiotic coverage can be modified accordingly.

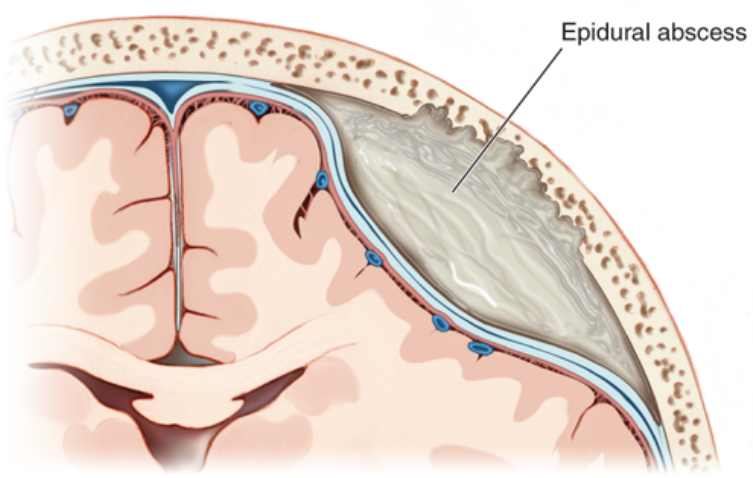
PROGNOSIS

Prognosis is influenced by the level of consciousness of the patient at the time of hospital presentation, the size of the empyema, and the speed with which therapy is instituted. Long-term neurologic sequelae, which include seizures and hemiparesis, occur in up to 50% of cases.

CRANIAL EPIDURAL ABSCESS

Cranial epidural abscess is a suppurative infection occurring in the potential space between the inner skull table and dura **(Fig. 135-4)**.

FIGURE 135-4
Cranial epidural abscess is a collection of pus between the dura and the inner table of the skull.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ETIOLOGY AND PATHOPHYSIOLOGY

Cranial epidural abscess is less common than either brain abscess or SDE and accounts for <2% of focal suppurative CNS infections. A cranial epidural abscess develops as a complication of a craniotomy or compound skull fracture or as a result of spread of infection from the frontal sinuses, middle ear, mastoid, or orbit. An epidural abscess may develop contiguous to an area of osteomyelitis, when craniotomy is complicated by infection of the wound or bone flap, or as a result of direct infection of the epidural space. Infection in the frontal sinus, middle ear, mastoid, or orbit can reach the epidural space through retrograde spread of infection from septic thrombophlebitis in the emissary veins that drain these areas or by way of direct spread of infection through areas of osteomyelitis. Unlike the subdural space, the epidural space is really a potential rather than an actual compartment. The dura is normally tightly adherent to the inner skull table, and infection must dissect the dura away from the skull table as it spreads. As a result, epidural abscesses are often smaller than SDEs. Cranial epidural abscesses, unlike brain abscesses, only rarely result from hematogenous spread of infection from extracranial primary sites. The bacteriology of a cranial epidural abscess is similar to that of SDE (see above). The etiologic organisms of an epidural abscess that arises from frontal sinusitis, middle-ear infections, or mastoiditis are usually streptococci or anaerobic organisms. Staphylococci or gram-negative organisms are the usual cause of an epidural abscess that develops as a complication of craniotomy or compound skull fracture.

CLINICAL PRESENTATION

Patients present with fever (60%), headache (40%), nuchal rigidity (35%), seizures (10%), and focal deficits (5%). Development of symptoms may be insidious, as the empyema usually enlarges slowly in the confined anatomic space between the dura and the inner table of the skull. Periorbital edema and Pott's puffy tumor, reflecting underlying associated frontal bone osteomyelitis, are present in ~40%. In patients with a recent neurosurgical procedure, wound infection is invariably present, but other symptoms may be subtle and can include altered mental status (45%), fever (35%), and headache (20%). The diagnosis should be considered when fever and headache follow recent head trauma or occur in the setting of frontal sinusitis, mastoiditis, or otitis media.

DIAGNOSIS

Cranial MRI with gadolinium enhancement is the procedure of choice to demonstrate a cranial epidural abscess. The sensitivity of CT is limited by the presence of signal artifacts arising from the bone of the inner skull table. The CT appearance of an epidural empyema is that of a lens or crescent-shaped hypodense extraaxial lesion. On MRI, an epidural empyema appears as a lentiform or crescent-shaped fluid collection that is hyperintense compared to CSF on T2-weighted images. On T1-weighted images, the fluid collection may be either isointense or hypointense compared to brain. Following the administration of gadolinium, there is linear enhancement of the dura on T1-weighted images. In distinction to subdural empyema, signs of mass effect or other parenchymal abnormalities are uncommon.

TREATMENT

TREATMENT

Epidural Abscess

Immediate neurosurgical drainage is indicated. Empirical antimicrobial therapy, pending the results of Gram's stain and culture of the purulent material obtained at surgery, should include a combination of a third-generation cephalosporin, vancomycin, and metronidazole **(see Table 133-1)**. Ceftazidime or meropenem should be substituted for ceftriaxone or cefotaxime in neurosurgical patients. Metronidazole is not necessary for antianaerobic coverage in patients receiving meropenem. When the organism has been identified, antimicrobial therapy can be modified accordingly. Antibiotics should be continued for 3–6 weeks after surgical drainage. Patients with associated osteomyelitis may require additional therapy.

PROGNOSIS

The mortality rate is <5% in modern series, and full recovery is the rule in most survivors.

SUPPURATIVE THROMBOPHLEBITIS

DEFINITION

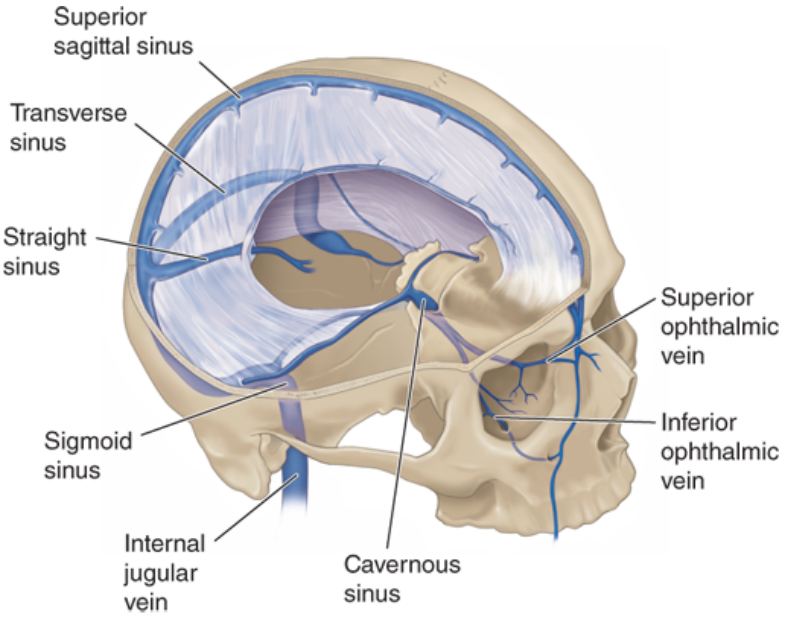
Suppurative intracranial thrombophlebitis is septic venous thrombosis of cortical veins and sinuses. This may occur as a complication of bacterial meningitis; SDE; epidural abscess; or infection in the skin of the face, paranasal sinuses, middle ear, or mastoid.

ANATOMY AND PATHOPHYSIOLOGY

The cerebral veins and venous sinuses have no valves; therefore, blood within them can flow in either direction. The superior sagittal sinus is the largest of the venous sinuses **(Fig. 135-5)**. It receives blood from the frontal, parietal, and occipital superior cerebral veins and the diploic veins, which communicate with the meningeal veins. Bacterial meningitis is a common predisposing condition for septic thrombosis of the superior sagittal sinus. The diploic veins, which drain into the superior sagittal sinus, provide a route for the spread of infection from the meninges, especially in cases where there is purulent exudate near areas of the superior sagittal sinus. Infection can also spread to the superior sagittal sinus from nearby SDE or epidural abscess. Dehydration from vomiting, hypercoagulable states, and immunologic abnormalities, including the presence of circulating antiphospholipid antibodies, also contribute to cerebral venous sinus thrombosis. Thrombosis may extend from one sinus to another, and at autopsy, thrombi of different histologic ages can often be detected in several sinuses. Thrombosis of the superior sagittal sinus is often associated with thrombosis of superior cortical veins and small parenchymal hemorrhages.

FIGURE 135-5

Anatomy of the cerebral venous sinuses.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The superior sagittal sinus drains into the transverse sinuses (Fig. 135-5). The transverse sinuses also receive venous drainage from small veins from both the middle ear and mastoid cells. The transverse sinus becomes the sigmoid sinus before draining into the internal jugular vein. Septic transverse/sigmoid sinus thrombosis can be a complication of acute and chronic otitis media or mastoiditis. Infection spreads from the

mastoid air cells to the transverse sinus via the emissary veins or by direct invasion. The cavernous sinuses are inferior to the superior sagittal sinus at the base of the skull. The cavernous sinuses receive blood from the facial veins via the superior and inferior ophthalmic veins. Bacteria in the facial veins enter the cavernous sinus via these veins. Bacteria in the sphenoid and ethmoid sinuses can spread to the cavernous sinuses via the small emissary veins. The sphenoid and ethmoid sinuses are the most common sites of primary infection resulting in septic cavernous sinus thrombosis.

CLINICAL MANIFESTATIONS

Septic thrombosis of the superior sagittal sinus presents with headache, fever, nausea and vomiting, confusion, and focal or generalized seizures. There may be a rapid development of stupor and coma. Weakness of the lower extremities with bilateral Babinski's signs or hemiparesis is often present. When superior sagittal sinus thrombosis occurs as a complication of bacterial meningitis, nuchal rigidity and Kernig's and Brudzinski's signs may be present.

The oculomotor nerve, the trochlear nerve, the abducens nerve, the ophthalmic and maxillary branches of the trigeminal nerve, and the internal carotid artery all pass through the cavernous sinus (**see Fig. 433-4**). The symptoms of *septic cavernous sinus thrombosis* are fever, headache, frontal and retroorbital pain, and diplopia. The classic signs are ptosis, proptosis, chemosis, and extraocular dysmotility due to deficits of cranial nerves III, IV, and VI; hyperesthesia of the ophthalmic and maxillary divisions of the fifth cranial nerve and a decreased corneal reflex may be detected. There may be evidence of dilated, tortuous retinal veins and papilledema.

Headache and earache are the most frequent symptoms of *transverse sinus thrombosis*. A transverse sinus thrombosis may also present with otitis media, sixth nerve palsy, and retroorbital or facial pain (*Gradenigo's syndrome*). Sigmoid sinus and internal jugular vein thrombosis may present with neck pain.

DIAGNOSIS

The diagnosis of septic venous sinus thrombosis is suggested by an absent flow void within the affected venous sinus on MRI and confirmed by magnetic resonance venography, CT angiogram, or the venous phase of cerebral angiography. The diagnosis of thrombophlebitis of intracerebral and meningeal veins is suggested by the presence of intracerebral hemorrhage but requires cerebral angiography for definitive diagnosis.

TREATMENT

TREATMENT

Suppurative Thrombophlebitis

Septic venous sinus thrombosis is treated with antibiotics, hydration, and removal of infected tissue and thrombus in septic lateral or cavernous sinus thrombosis. The choice of antimicrobial therapy is based on the bacteria responsible for the predisposing or associated condition. Optimal duration of therapy is unknown, but antibiotics are usually continued for 6 weeks or until there is radiographic evidence of resolution of thrombosis. Anticoagulation with dose-adjusted intravenous heparin is recommended for aseptic venous sinus thrombosis and in the treatment of septic venous sinus thrombosis complicating bacterial meningitis in patients who have progressive neurologic deterioration despite antimicrobial therapy and intravenous fluids. The presence of a small intracerebral hemorrhage from septic thrombophlebitis is not an absolute contraindication to heparin therapy. Successful management of aseptic venous sinus thrombosis has been reported with surgical thrombectomy, catheter-directed urokinase therapy, and a combination of intrathrombus recombinant tissue plasminogen activator (rtPA) and intravenous heparin, but there are not enough data to recommend these therapies in septic venous sinus thrombosis.

FURTHER READING

Baird A et al: Evidence-based guideline: Treatment of parenchymal neurocysticerosis. Report of the Guideline Development Subcommittee of the American Academy of Neurology. Neurology 80:1424, 2013.

Garcia HH et al: Clinical symptoms, diagnosis and treatment of neurocysticercosis. Lancet Neurol 13:1202, 2014.

Chapter 136: Infectious Complications of Bites

Sandeep S. Jubbal; Florencia Pereyra; Lawrence C. Madoff

INTRODUCTION

The skin is an essential component of nonspecific immunity, protecting the host from potential pathogens in the environment. Breaches in this protective barrier thus represent a form of immunocompromise that predisposes the patient to infection. Bites and scratches from animals and humans allow the inoculation of microorganisms past the skin's protective barrier into deeper, susceptible host tissues.

Each year in the United States, millions of animal-bite wounds are sustained. The vast majority are inflicted by pet dogs and cats, which number >100 million; the annual incidence of dog and cat bites has been reported as 300 bites per 100,000 population. Other bite wounds are a consequence of encounters with animals in the wild or in occupational settings. While many of these wounds require minimal or no therapy, a significant number result in infection, which may be life-threatening. The microbiology of bite-wound infections in general reflects the oropharyngeal flora of the biting animal, although organisms from the soil, the skin of the animal and the victim, and the animal's feces may also be involved.

DOG BITES

In the United States, dogs bite >4.7 million people each year and are responsible for 80% of all animal-bite wounds, an estimated 15–20% of which become infected. Each year, 800,000 Americans seek medical attention for dog bites; of those injured, 386,000 require treatment in an emergency department, with >1000 emergency department visits each day and ~30 deaths per year. Most dog bites are provoked and are inflicted by the victim's pet or by a dog known to the victim. These bites are frequently sustained during efforts to break up a dogfight. Children are more likely than adults to sustain canine bites, with the highest incidence of 6 bites per 1000 population among boys 5–9 years old. Victims are more often male than female, and bites most often involve an upper extremity. Among children <4 years old, two-thirds of all these injuries involve the head or neck. Infection typically manifests 8–24 h after the bite as pain at the site of injury with cellulitis accompanied by purulent, sometimes foul-smelling discharge. Septic arthritis and osteomyelitis may develop if a canine tooth penetrates synovium or bone. Systemic manifestations (e.g., fever, lymphadenopathy, and lymphangitis) also may occur. The microbiology of dog-bite wound infections is usually mixed and includes *Pasteurella* species, β-hemolytic streptococci, *Staphylococcus* species (including methicillin-resistant *Staphylococcus aureus* [MRSA] and *Staphylococcus intermedius*), *Neisseria* species (commonly *Neisseria weaveri*, formerly known as CDC group M-5), *Eikenella corrodens*, and *Capnocytophaga*

canimorsus. Many wounds also include anaerobic bacteria such as *Actinomyces, Fusobacterium, Prevotella*, and *Porphyromonas* species.

While most infections resulting from dog-bite injuries are localized to the area of injury, many of the microorganisms involved are capable of causing systemic infection, including bacteremia, meningitis, brain abscess, endocarditis, and chorioamnionitis. These infections are particularly likely in hosts with edema or compromised lymphatic drainage in the involved extremity (e.g., after a bite on the arm in a woman who has undergone mastectomy) and in patients who are immunocompromised by medication or disease (e.g., glucocorticoid use, systemic lupus erythematosus, acute leukemia, or hepatic cirrhosis). In addition, dog bites and scratches may result in systemic illnesses such as rabies **(Chap. 203)** and tetanus **(Chap. 147)**.

Infection with *C. canimorsus* following dog-bite wounds (or licking of preexisting wounds) may result in fulminant sepsis, disseminated intravascular coagulation, and renal failure, particularly in hosts who have impaired hepatic function, who have undergone splenectomy, or who are immunosuppressed. This thin gram-negative rod is difficult to culture on most solid media but grows in a variety of liquid media. It may require up to 14 days of incubation to grow on blood cultures. The bacteria are occasionally seen within polymorphonuclear leukocytes on Wright-stained smears of peripheral blood from septic patients. Tularemia **(Chap. 165)** also has been reported to follow dog bites.

CAT BITES

Although less common than dog bites, cat bites and scratches result in infection in more than half of all cases. Because the cat's narrow, sharp canine teeth penetrate deeply into tissue, cat bites are more likely than dog bites to cause septic arthritis and osteomyelitis; the development of these conditions is particularly likely when punctures are located over or near a joint, especially in the hand. Women sustain cat bites more frequently than do men. These bites most often involve the hands and arms. Both bites and scratches from cats are prone to infection from organisms in the cat's oropharynx. *Pasteurella multocida*, a normal component of the feline oral flora, is a small gram-negative coccobacillus implicated in the majority of cat-bite wound infections. Like that of dog-bite wound infections, however, the microflora of cat-bite wound infections is usually mixed. Other microorganisms causing infection after cat bites are similar to those causing dog-bite wound infections.

The same risk factors for systemic infection following dog-bite wounds apply to cat-bite wounds. *Pasteurella* infections tend to advance rapidly, often within hours, causing severe inflammation accompanied by purulent drainage with adenitis; *Pasteurella* may also be spread by respiratory droplets from animals, resulting in pneumonia or bacteremia. Like dog-bite wounds, cat-bite wounds may result in the transmission of rabies or in the development of tetanus. Infection with *Bartonella henselae* causes cat-scratch disease **(Chap. 167)** and is an important late consequence of cat bites and scratches. Tularemia **(Chap. 165)** also has been reported to follow cat bites. Occasionally, sporotrichosis **(Chap. 214)** has been associated with scratches or bites by animals, especially domestic cats.

OTHER ANIMAL BITES

Infections have been attributed to bites from many animal species. Often these bites are sustained as a consequence of occupational exposure (farmers, laboratory workers, veterinarians) or recreational exposure (hunters and trappers, wilderness campers, owners of exotic pets). Generally, the microflora of bite wounds reflects the oral flora of the biting animal. Most members of the cat family, including feral cats, harbor *P. multocida*. Bite wounds from aquatic animals such as alligators or piranhas may contain *Aeromonas hydrophila*. Shark, moray eel, and barracuda bites, like other injuries sustained in saltwater, are often associated with infections with marine *Vibrio* species. Venomous snakebites (Chap. 451) result in severe inflammatory responses and tissue necrosis—conditions that render these injuries prone to infection. The snake's oral flora includes many species of aerobes and anaerobes, such as *Pseudomonas aeruginosa, Serratia marcescens, Proteus* species, *Staphylococcus epidermidis, Bacteroides fragilis*, and *Clostridium* species. Bites from nonhuman primates are highly susceptible to infection with pathogens similar to those isolated from human bites (see below). Bites from Old World monkeys (*Macaca*) may also result in the transmission of B virus (*Macacine herpesvirus 1, Herpesvirus simiae, Cercopithecine herpesvirus*), a cause of serious infection of the human central nervous system. *Actinobacillus lignieresii* has often been reported in infected wounds of humans bitten by horses, pigs, and sheep. Bites of seals, walruses, and polar bears may cause a chronic suppurative infection known as *seal finger*, which is probably due to one or more species of *Mycoplasma* colonizing these animals.


Small rodents, including rats, mice, and gerbils, as well as animals that prey on rodents may transmit *Streptobacillus moniliformis* (a microaerophilic, pleomorphic gram-negative rod) or *Spirillum minor* (a spirochete); these organisms cause a clinical illness known as *rat-bite fever*. The vast majority of cases in the United States are streptobacillary, whereas *Spirillum* infection occurs mainly in Asia.

In the United States, the risk of rodent bites is usually greatest among laboratory workers or inhabitants of rodent-infested dwellings (particularly children). Rat-bite fever is distinguished from acute bite-wound infection by its typical manifestation after the initial wound has healed. Streptobacillary disease follows an incubation period of 3–10 days. Fever, chills, myalgias, headache, and severe migratory arthralgias are usually followed by a maculopapular rash, which characteristically involves the palms and soles and may become confluent or purpuric. Complications include endocarditis, myocarditis, meningitis, pneumonia, and abscesses in many organs. *Haverhill fever* is an *S. moniliformis* infection acquired from contaminated milk or drinking water and has similar manifestations. Streptobacillary rat-bite fever was frequently fatal in the preantibiotic era. The differential diagnosis includes Rocky Mountain spotted fever, Lyme disease, leptospirosis, and secondary syphilis. The diagnosis is made by direct observation of the causative organisms in tissue or blood, by culture of the organisms on enriched media, or by serologic testing with specific agglutinins.

Spirillum infection (referred to in Japan as *sodoku*) causes pain and purple swelling at the site of the initial bite, with associated lymphangitis and regional lymphadenopathy, after an incubation period of 1–4 weeks. The systemic illness includes fever, chills, and headache. The original lesion may eventually progress to an

eschar. The infection is diagnosed by direct visualization of the spirochetes in blood or tissue or by animal inoculation.

Finally, NO-1 (CDC nonoxidizer group 1) is a bacterium associated with dog- and cat-bite wounds. Infections in which NO-1 has been isolated have tended to manifest locally (i.e., as abscess and cellulitis). These infections have occurred in healthy persons with no underlying illness and in some instances have progressed from localized to systemic illnesses. The phenotypic characteristics of NO-1 are similar to those of asaccharolytic *Acinetobacter* species; i.e., NO-1 is oxidase-, indole-, and urease-negative. To date, all strains identified have been shown to be susceptible to aminoglycosides, β-lactam antibiotics, tetracyclines, quinolones, and sulfonamides.

HUMAN BITES

Human bites may be self-inflicted; may be sustained by medical personnel caring for patients; or may take place during fights, domestic abuse, or sexual activity. Human-bite wounds become infected more frequently (~10–15% of the time) than do bites inflicted by other animals. These infections reflect the diverse oral microflora of humans, which includes multiple species of aerobic and anaerobic bacteria. Common aerobic isolates include viridans streptococci, *S. aureus, E. corrodens* (which is particularly common in clenched-fist injury; see below), and *Haemophilus influenzae*. Anaerobic species, including *Fusobacterium nucleatum* and *Prevotella, Porphyromonas*, and *Peptostreptococcus* species, are isolated from 50% of wound infections due to human bites; many of these isolates produce β-lactamases. The oral flora of hospitalized and debilitated patients often includes Enterobacteriaceae in addition to the usual organisms. Hepatitis B, hepatitis C, herpes simplex virus infection, syphilis, tuberculosis, actinomycosis, and tetanus have been reported to be transmitted by human bites; it is biologically possible to transmit HIV through human bites, although this event is quite unlikely.

Human bites are categorized as either *occlusional* injuries, which are inflicted by actual biting, or *clenched-fist* injuries, which are sustained when the fist of one individual strikes the teeth of another, causing traumatic laceration of the hand. For several reasons, clenched-fist injuries, which are sometimes referred to as “fight bite” and which are more common than occlusional injuries, result in particularly serious infections. The deep spaces of the hand, including the bones, joints, and tendons, are frequently inoculated with organisms in the course of such injuries. The clenched position of the fist during injury, followed by extension of the hand, may further promote the introduction of bacteria as contaminated tendons retract beneath the skin’s surface. Moreover, medical attention is often sought only after frank infection develops.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Animal or Human Bites

A careful history should be elicited, including the type of biting animal, the type of attack (provoked or unprovoked), and the amount of time elapsed since injury. Local and regional public-health authorities should be contacted to determine whether an individual species could be rabid and/or to locate and observe the biting animal when rabies prophylaxis may be indicated (Chap. 203). Suspicious human-bite wounds should provoke careful questioning regarding domestic or child abuse. Details on antibiotic allergies, immunosuppression, splenectomy, liver disease, mastectomy, and immunization history should be obtained. The wound should be inspected carefully for evidence of infection, including redness, exudate, and foul odor. The type of wound (puncture, laceration, or scratch); the depth of penetration; and the possible involvement of joints, tendons, nerves, and bones should be assessed. It is often useful to include a diagram or photograph of the wound in the medical record. In addition, a general physical examination should be conducted and should include an assessment of vital signs as well as an evaluation for evidence of lymphangitis, lymphadenopathy, dermatologic lesions, and functional limitations. Injuries to the hand warrant consultation with a hand surgeon for the assessment of tendon, nerve, and muscular damage. Radiographs should be obtained when bone may have been penetrated or a tooth fragment may be present. Culture and Gram's staining of all infected wounds are essential; anaerobic cultures should be undertaken if abscesses, devitalized tissue, or foul-smelling exudate is present. A small-tipped swab may be used to culture deep punctures or small lacerations. It is also reasonable to culture samples from apparently uninfected wounds due to bites inflicted by animals other than dogs and cats, since the microorganisms causing disease are less predictable in these cases. The white blood cell count should be determined and the blood cultured if systemic infection is suspected.

TREATMENT

TREATMENT

Bite-Wound Infections

WOUND MANAGEMENT

Wound closure is controversial in bite injuries. Many authorities prefer not to attempt primary closure of wounds that are or may become infected, choosing instead to irrigate these wounds copiously, debride devitalized tissue, remove foreign bodies, and approximate the wound edges. Delayed primary closure may be undertaken after the risk of infection is over. Small uninfected wounds may be allowed to close by secondary intention. Puncture wounds due to cat bites should be left unsutured because of the high rate at which they become infected. Facial wounds are usually sutured after thorough cleaning and irrigation because of the importance of a good cosmetic result in this area and because anatomic factors such as an excellent blood supply and the absence of dependent edema lessen the risk of infection. In general, wounds >12 h old (for bites to the arm or leg) or >24 h old (for bites to the face) should not be closed primarily and may require prophylactic antibiotics (see below).

ANTIBIOTIC THERAPY

Established Infection

Antibiotics should be administered for all established bite-wound infections and should be chosen in light of the most likely potential pathogens, as indicated by the biting species and by Gram's stain and culture results **(Table 136-1)**. For dog and cat bites, antibiotics should be effective against *S. aureus, Pasteurella* species, *C. canimorsus*, streptococci, and oral anaerobes. For human bites, agents with activity against *S. aureus, H. influenzae*, and β-lactamase-positive oral anaerobes should be used. The combination of an extended-spectrum penicillin with a β-lactamase inhibitor (amoxicillin/clavulanic acid, ticarcillin/clavulanic acid, ampicillin/sulbactam) appears to offer the most reliable coverage for these pathogens. Second- and third-generation cephalosporins (cefuroxime, cefoxitin, cefpodoxime) also offer substantial coverage when given in conjunction with a drug that provides anaerobic coverage (clindamycin or metronidazole). The choice of antibiotics for penicillin-allergic patients (particularly those in whom immediate-type hypersensitivity makes the use of cephalosporins hazardous) is more difficult and is based primarily on in vitro sensitivity since data on clinical efficacy are inadequate. The combination of an antibiotic active against gram-positive cocci and anaerobes (such as clindamycin) with trimethoprim-sulfamethoxazole or a fluoroquinolone, which is active against many of the other potential pathogens, would appear reasonable. In vitro data suggest that azithromycin alone provides coverage against most commonly isolated bite-wound pathogens; however, this agent has variable activity against *P. multocida*, *E. corrodens,* and fusobacteria and thus should be avoided unless no alternative agent is available. As MRSA becomes more common in the community and evidence of its transmission between humans and their animal contacts increases, empirical use of agents active against MRSA should be considered in high-risk situations while culture results are awaited.

Antibiotics are generally given for 10–14 days, but the response to therapy must be carefully monitored. Failure to respond should prompt a consideration of diagnostic alternatives and surgical evaluation for possible drainage or debridement. Complications such as osteomyelitis or septic arthritis mandate a longer duration of therapy.

Management of *C. canimorsus* sepsis requires a 2-week course of IV penicillin G (2 million units IV every 4 h) and supportive measures. Alternative agents for the treatment of *C. canimorsus* infection include cephalosporins and fluoroquinolones. Serious infection with *P. multocida* (e.g., pneumonia, sepsis, or meningitis) also should be treated with IV penicillin G. Alternative agents include a second- or third-generation cephalosporin or ciprofloxacin.

Bites by venomous snakes **(Chap. 451)** may not require antibiotic treatment. Because it is often difficult to distinguish signs of infection from tissue damage caused by the envenomation, many authorities continue to recommend treatment directed against the snake's oral flora—i.e., the administration of broadly active agents such as ceftriaxone (1–2 g IV every 12–24 h) or ampicillin/sulbactam (1.5–3.0 g IV every 6 h).

Seal finger appears to respond to doxycycline (100 mg twice daily for a duration guided by the response to therapy).

Presumptive or Prophylactic Therapy

The use of antibiotics for patients presenting early (within 8 h) after bite injury is controversial. Although symptomatic infection frequently will not yet have manifested at this point, many early wounds will harbor

pathogens, and many will become infected. Studies of antibiotic prophylaxis for wound infections are limited and have often included only small numbers of cases in which various types of wounds have been managed according to various protocols. A meta-analysis of eight randomized trials of prophylactic antibiotics in patients with dog-bite wounds demonstrated a reduction in the rate of infection by 50% with prophylaxis. However, in the absence of sound clinical trials, many clinicians base the decision to treat bite wounds with empirical antibiotics on the species of the biting animal; the location, severity, and extent of the bite wound; and the existence of comorbid conditions in the host. All human- and monkey-bite wounds should be treated presumptively because of the high rate of infection. Most cat-bite wounds, particularly those involving the hand, should be treated. Other factors favoring treatment for bite wounds include severe injury, as in crush wounds; potential bone or joint involvement; involvement of the hands or genital region; host immunocompromise, including that due to diabetes mellitus, liver disease, or splenectomy; involvement of extremities with underlying venous and/or lymphatic compromise; and prior mastectomy on the side of an involved upper extremity. When prophylactic antibiotics are administered, they are usually given for 3–5 days.

Rabies and Tetanus Prophylaxis

Rabies prophylaxis, consisting of both passive administration of rabies immune globulin (with as much of the dose as possible infiltrated into and around the wound) and active immunization with rabies vaccine, should be given in consultation with local and regional public-health authorities for some animal bites and scratches as well as for certain nonbite exposures (Chap. 203). Rabies is endemic in a variety of animals, including dogs and cats in many areas of the world. In the United States, although the majority (90%) of rabid animals reported each year are wild (including raccoons, skunks, foxes, and bats), most people receive rabies prophylaxis because of close contact with domestic animals. Furthermore, more cats than dogs are reported rabid each year. Many local health authorities require the reporting of all animal bites.

A tetanus booster immunization should be given if the patient has undergone primary immunization but has not received a booster dose in the past 5 years. Patients who have not previously completed primary immunization should be immunized and should also receive tetanus immune globulin. Elevation of the site of injury is an important adjunct to antimicrobial therapy. Immobilization of the infected area, especially the hand, also is beneficial.

Hepatitis B Prophylaxis

Hepatitis B virus can be transmitted, albeit rarely, by exposure of non-intact skin to blood-free saliva. The mainstay of postexposure prophylaxis is active immunization with hepatitis B vaccine, but, in certain circumstances, hepatitis B immune globulin is recommended in addition to vaccine for added protection (Chap. 332).

TABLE 136-1

Management of Wound Infections Following Animal and Human Bites

Biting Species	Commonly Isolated Pathogens	Preferred Antibiotic(s)[a]	Alternative in Penicillin-Allergic Patient	Prophylaxis Advised for Early Uninfected Wounds	Other Considerations
Dog	*Staphylococcus aureus, Pasteurella multocida*, anaerobes, *Capnocytophaga canimorsus*	Amoxicillin/clavulanate (250–500 mg PO tid) or ampicillin/sulbactam (1.5–3.0 g IV q6h)	Clindamycin (150–300 mg PO qid) plus either TMP-SMX (1 DS tablet PO bid) or ciprofloxacin (500 mg PO bid)	Sometimes[b]	Consider rabies prophylaxis.
Cat	*P. multocida, S. aureus*, anaerobes	Amoxicillin/clavulanate or ampicillin/sulbactam as above	Clindamycin plus TMP-SMX as above or a fluoroquinolone	Usually	Consider rabies prophylaxis. Carefully evaluate for joint/bone penetration.
Human, occlusional	Viridans streptococci, *S. aureus, Haemophilus influenzae*, anaerobes	Amoxicillin/clavulanate or ampicillin/sulbactam as above	Erythromycin (500 mg PO qid) or a fluoroquinolone	Always	
Human, clenched-fist	As for occlusional, plus *Eikenella corrodens*	Ampicillin/sulbactam as above or imipenem (500 mg q6h)	Cefoxitin[c]	Always	Examine for tendon, nerve, or joint involvement.

Biting Species	Commonly Isolated Pathogens	Preferred Antibiotic(s)[a]	Alternative in Penicillin-Allergic Patient	Prophylaxis Advised for Early Uninfected Wounds	Other Considerations
Monkey	As for human bite	As for human bite	As for human bite	Always	For macaque monkeys, consider B virus prophylaxis with acyclovir.
Snake	*Pseudomonas aeruginosa, Proteus* spp., *Bacteroides fragilis, Clostridium* spp.	Ampicillin/sulbactam as above	Clindamycin plus TMP-SMX as above or a fluoroquinolone	Sometimes, especially with venomous snakes	Administer antivenin for venomous snakebite.
Rodent	*Streptobacillus moniliformis, Leptospira* spp., *P. multocida*	Penicillin VK (500 mg PO qid)	Doxycycline (100 mg PO bid)	Sometimes	
Aquatic animal (alligator, piranha, shark, moray eel, barracuda)	*Aeromonas hydrophila*, marine *Vibrio* spp. *(Vibrio vulnificus)*	Third-generation cephalosporin (e.g., ceftriaxone, 1 g IV q24h) plus doxycycline (100 mg PO bid)	Clindamycin plus levofloxacin (750 mg PO qd) plus doxycycline	Always	Obtain prompt surgical consultation, as risk for necrotizing infection is high with *Aeromonas* and *Vibrio* spp.

[a]Antibiotic choices should be based on culture data when available. These suggestions for empirical therapy need to be tailored to individual circumstances and local conditions. IV regimens should be used for hospitalized patients. A single IV dose of antibiotics may be given to patients who will be discharged after initial management. [b]Prophylactic antibiotics are suggested for severe or extensive wounds, facial wounds, and crush injuries; when bone or joint may be

involved; and when comorbidity is present (see text). [C]May be hazardous in patients with immediate-type hypersensitivity to penicillin.

Abbreviations: DS, double-strength; TMP-SMX, trimethoprim-sulfamethoxazole.

FURTHER READING

Abrahamian FM, Goldstein EJC: Microbiology of animal bite wound infections. Clin Microbiol Rev 24:231, 2011.

Brook I: Management of human and animal bite wounds: An overview. Adv Skin Wound Care 18:197, 2005.

Ellis R, Ellis C: Dog and cat bites. Am Fam Phys 90:239, 2014.

Fallouji MA: Traumatic love bites. Br J Surg 77:100, 1990.

Fleisher GR: The management of bite wounds. N Engl J Med 340:138, 1999.

Kaiser RM et al: Clinical significance and epidemiology of NO-1, an unusual bacterium associated with dog and cat bites. Emerg Infect Dis 8:171, 2002.

Kullberg BJ et al: Purpura fulminans and symmetrical peripheral gangrene caused by *Capnocytophaga canimorsus* (formerly DF-2) septicemia—a complication of dog bite. Medicine (Baltimore) 70:287, 1991.

Lohiya GS et al: Human bites: Bloodborne pathogen risk and postexposure follow-up algorithm. J Natl Med Assoc 105:92, 2013.

Morgan M, Palmer J: Dog bites. BMJ 334:413, 2007.

Oehler RL et al: Bite-related and septic syndromes caused by cats and dogs. Lancet Infect Dis 9:439, 2009.

Stevens DL et al: Practice guidelines for the diagnosis and management of skin and soft tissue infections. 2014 update by the Infectious Diseases Society of America. Clin Infect Dis 59:e10, 2014. [PubMed: 24973422]

Weber DJ et al: Infections resulting from animal bites. Infect Dis Clin North Am 5:663, 1991.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 137: Infections Acquired in Health Care Facilities

Robert A. Weinstein

FIGURE 137-1

INTRODUCTION

Health care–associated infections affect as many as 1.7 million patients at a cost of ~$10–33 billion and up to 99,000 lives in U.S. hospitals annually. Although efforts to lower infection risks are challenged by numbers of immunocompromised patients, antibiotic-resistant bacteria, and fungal and viral superinfections, a prevailing viewpoint—"zero tolerance"—is that health care–associated infections are avoidable with strict application of evidence-based prevention guidelines (Table 137-1). In fact, rates of most device-related infections—historically, the largest drivers of risk—have fallen steadily over the past few years. Unfortunately, at the same time, antimicrobial-resistant pathogens have risen in number and are estimated to contribute to ~23,000 deaths annually. This chapter reviews health care–associated and device-related infections as well as basic surveillance, prevention, control, and treatment activities.

TABLE 137-1

Sources of Infection Control Guidelines and Oversight

Organization	Role	Major Constituents	Website
Joint Commission CAP OSHA CMS	Regulatory Regulatory Regulatory Regulatory	Hospitals, long-term-care facilities, laboratories Laboratories Workers Medicare/Medicaid providers	*www.jointcommission.org* *www.cap.org* *www.osha.gov* *www.cms.hhs.gov*
PQRI HHS Action Plan	Regulatory and advisory Advisory and national action plan	Eligible professionals Health care and infection prevention personnel	*www.cms.hhs.gov/pqri/* *www.health.gov/hcq/prevent-hai.asp;* *https://health.gov/hcq/pdfs/hai-action-plan-acute-care-hospitals.pdf*
CDC DHQP NHSN HICPAC NIOSH AHRQ NQF National Academies Federal Influenza Planning Trust for America's Health PACCARB National Action Plan on CARB	 Advisory Surveillance Advisory Advisory Advisory Advisory Advisory Advisory Advisory Advisory Action plan	 Health care facilities and personnel Health care facilities and personnel Health care facilities and personnel Workers Broad (e.g., health care personnel) Broad (e.g., health care personnel) Broad (e.g., health care personnel) Health care and public health personnel Broad (e.g., the public) Health care and public health personnel Health care and public health personnel	 *www.cdc.gov/ncezid/dhqp/* *https://www.cdc.gov/nhsn/index.html* *www.cdc.gov/hicpac/* *www.cdc.gov/niosh/* *www.ahrq.gov; https://www.ahrq.gov/research/data/hcup/index.html* *www.qualityforum.org* *www.nationalacademies.org* *www.flu.gov/planning-preparedness/hospital* *www.healthyamericans.org* *www.hhs.gov/ash/advisory-committees/paccarb* *www.hhs.gov/ash/advisory-committees/paccarb/working-groups/national-action-plan*
CSTE	Advisory and professional society	Public health personnel	*www.cste.org*
IDSA SHEA HIS APIC BSAC	Professional society Professional society Professional society Professional society Professional society	Infectious disease physicians/researchers Health care epidemiologists Health care epidemiologists Infection preventionists Medical microbiologists	*www.idsociety.org* *www.shea-online.org* *www.his.org.uk* *www.apic.org* *www.bsac.org.uk*
MedQIC IHI Leapfrog Group NSQIP	Quality improvement Quality improvement Quality improvement Quality improvement	Broad (e.g., health care personnel) Broad (e.g., health care personnel) Broad (payers, consumers, employers, and health care personnel) Surgery services	*www.qualitynet.org* *www.ihi.org* *www.leapfroggroup.org* *www.facs.org/quality-programs/acs-nsqip*

Abbreviations: AHRQ, Agency for Healthcare Research and Quality; APIC, Association for Professionals in Infection Control and Epidemiology; BSAC, British Society for Antimicrobial Chemotherapy; CAP, College of American Pathologists; CARB, combating antibiotic-resistant bacteria; CDC, Centers for Disease Control and Prevention; CMS, Centers for Medicare & Medicaid Services; CSTE, Council of State and Territorial Epidemiologists; DHQP, Division of Healthcare Quality Promotion; HHS, Health and Human Services; HICPAC, Healthcare Infection Control Practices Advisory Committee; HIS, Hospital Infection Society; IDSA, Infectious Diseases Society of America; IHI, Institute for Healthcare Improvement; IOM, Institute of Medicine; MedQIC, Medicare Quality Improvement Community; NHSN, National Healthcare Safety Network; NIOSH, National Institute for Occupational Safety and Health; NQF, National Quality Forum; NSQIP, National Surgical Quality Improvement Program; OSHA, Occupational Safety & Health Administration; PACCARB, Presidential Advisory Council on Combating Antibiotic-Resistant Bacteria; PQRI, Physician Quality Reporting Initiative; SHEA, Society for Healthcare Epidemiology of America.

ORGANIZATION, RESPONSIBILITIES, AND SCRUTINY OF HEALTH CARE–ASSOCIATED INFECTION PROGRAMS

The standards of the Joint Commission require accredited hospitals to have active programs for surveillance, prevention, and control of nosocomial infections. Concerns over patient safety have led to federal legislation that prevents U.S. hospitals from upgrading Medicare charges to pay for hospital costs resulting from at least 14 specific nosocomial events, including some health care–associated infections (*www.cms.gov/hospitalacqcond*), and have prompted public reporting on processes of patient care (e.g., timely administration and appropriateness of perioperative antibiotic prophylaxis) and patient outcomes (e.g., surgical wound infection rates). Recommendations for ensuring a culture of safety continue to evolve (e.g., *www.npsf.org/free-from-harm*), but neither the carrot (pay-for-performance) nor the stick (nonpayment for preventable infections) appears to have had a major impact on infection rates. The effect of public attention may be more positive (*https://www.cdc.gov/hai/pdfs/stateplans/factsheets/us.pdf*) **(Fig. 137-1)**; the U.S. Department of Health and Human Services has updated its interagency Action Plan to Prevent Health Care–Associated Infections, with a good-progress midpoint 2014 evaluation and targets for the year 2020 **(Table 137-2)**.

TABLE 137-2

Summary of Progress Toward the National 2020 Targets for Elimination of Health Care–Associated Infections (U.S. Department of Health and Human Services) and Midpoint Evaluation

Metric (Source)	Original Target for 2013[a] (from 2009 Baseline)	Progress Made by 2014	2020 Target (from 2015 Baseline)
Central line–associated bloodstream infections (NHSN)	50% reduction	50% reduction	50% reduction
Catheter-associated urinary tract infections (NHSN)	25% reduction	No change	25% reduction
Invasive MRSA infections (NHSN)	50% reduction	36% reduction	50% reduction
Facility-onset MRSA infections (NHSN)	25% reduction	13% reduction	50% reduction
Clostridium difficile infections (NHSN)	30% reduction	8% reduction	30% reduction
Surgical-site infections (NHSN)	25% reduction	18% reduction	30% reduction
Clostridium difficile hospitalizations (HCUP)	30% reduction	18% reduction	30% reduction

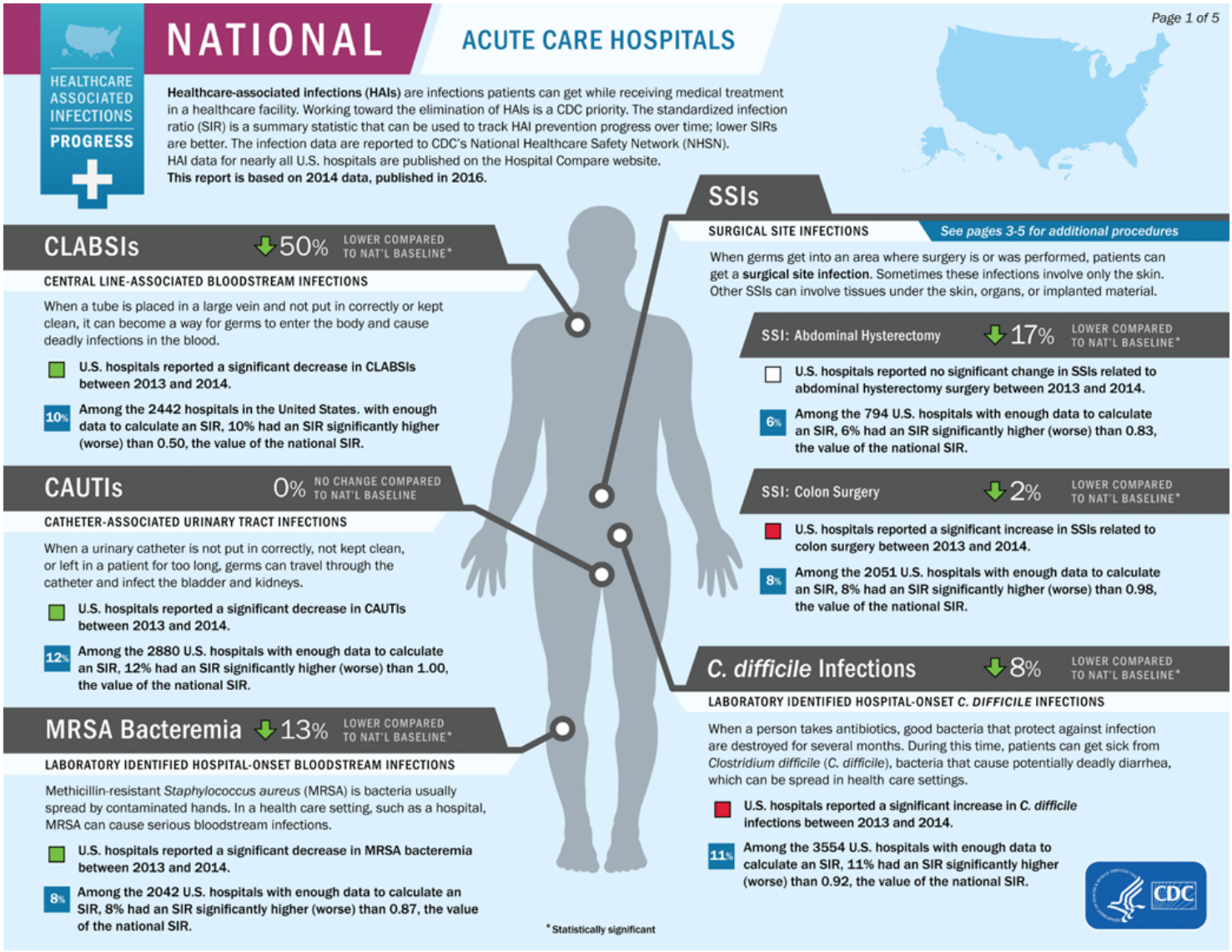
[a]Examples of changes as of 2013: Catheter-related bloodstream infections decreased to ~1.7/1000 catheter-days; *C. difficile* infections increased to ~11.2 cases/10,000 discharges; catheter-related urinary tract infections decreased to ~3.1/1000 catheter-days; and hospital-onset MRSA invasive infections decreased to ~4.5/100,000 persons.

Abbreviations: HCUP, Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project; MRSA, methicillin-resistant *Staphylococcus aureus*; NHSN, National Healthcare Safety Network (Centers for Disease Control and Prevention).

Source: Adapted from *www.health.gov/hcq/prevent-hai-measures.asp* (last accessed January 31, 2017).

FIGURE 137-1

Progress in control of selected common health care–associated infections in U.S. acute-care hospitals.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SURVEILLANCE

Traditionally, infection preventionists have surveyed inpatients for infections acquired in hospitals (defined as those neither present nor incubating at the time of admission). However, many infection-control programs have replaced manual surveillance of microbiology laboratory results and “shoe-leather” epidemiology on nursing wards with computerized algorithm-driven electronic surveillance of hospital databases (e.g., vascular catheter or surgical wound infections inferred from clinical microbiology data). Such approaches provide “house-wide” surveillance, remove observer bias, and display the potential power of newer computer techniques like machine learning algorithms. Although infection surveillance in many nursing homes and some long-term acute-care hospitals (LTACHs) is still in its formative stage, the role of these facilities in the transmission of antimicrobial-resistant pathogens puts a premium on their increased attention to surveillance and control.

In the spirit of “what is measured improves,” most states require public reporting of processes for prevention of health care–associated infection and/or patient outcomes. As a result, the surveillance pendulum is swinging back to use of “house-wide” surveillance, facilitated by computerized surveillance systems, and many states now require that hospitals use the Centers for Disease Control and Prevention’s (CDC’s) National Healthcare Safety Network (NHSN) reporting system to provide uniform definitions and to facilitate transmission of data. Increasing reliance on the NHSN by states to facilitate public reporting has led to participation by >20,000 facilities (~5000 of the ~5700 acute-care hospitals in the United States, ~540 LTACHs, ~1340 inpatient rehabilitation facilities, ~6800 outpatient dialysis facilities, ~4800 ambulatory surgery centers, and ~1900 long-term-care facilities). This level of participation provides a nationwide view of health care–associated infections and potential access to national rates of antimicrobial use and resistance.

Results of surveillance are expressed as rates, qualified when possible by duration of risk, site of infection, patient population, and exposure to risk factors. To account for some of these variables, the CDC now uses a standardized infection ratio (SIR; *www.cdc.gov/hai/national-annual-sir/*) as part of NHSN rate reporting. Meaningful denominators for infection rates include the number of patients exposed to a specific risk or the number of intervention days (e.g.,

1000 patient-days on a ventilator). As use of invasive devices such as indwelling bladder catheters has purposely been decreased, the denominators have become smaller, but patients who still require such devices often are at intrinsically higher risk (potential numerators)—a situation that may paradoxically increase rates when device-days account for the denominator. Temporal trends in rates should be reviewed, and rates should be compared with regional and national benchmarks that incorporate the SIR. Interhospital comparisons still may be misleading because of the wide range in risk factors and severity of underlying illnesses. Process measures (e.g., adherence to hand hygiene) usually do not require risk adjustment, and major morbidity and cost outcome measures (e.g., cardiac surgery wound-infection rates) can identify hospitals with outlier infection rates (e.g., in the top deciles) for further evaluation. Most importantly, temporal analysis of a hospital's infection rates—comparison to self over time—can help to determine whether control measures are succeeding and where increased efforts should be focused.

EPIDEMIOLOGIC BASIS AND GENERAL MEASURES FOR PREVENTION AND CONTROL

Nosocomial infections follow basic epidemiologic patterns that can help to direct prevention and control measures. Nosocomial pathogens have reservoirs, are transmitted by largely predictable routes, and require susceptible hosts. Reservoirs and sources exist in the inanimate environment (e.g., residual *Clostridium difficile* spores on frequently touched surfaces in patients' rooms) and in the animate environment (e.g., infected or colonized patients and hospital visitors). The mode of transmission usually is either cross-infection (e.g., indirect spread of pathogens from one patient to another on the inadequately cleaned hands of hospital personnel) or autoinoculation (e.g., aspiration of oropharyngeal flora into the lungs along an endotracheal tube). Occasionally, pathogens (e.g., group A streptococci and many respiratory viruses) are spread from person to person via large infectious droplets released by coughing or sneezing. Much less common—but often devastating in terms of epidemic risk—is true airborne spread of small or droplet nuclei (as in nosocomial chickenpox) or common-source spread (e.g., by contaminated IV fluids). Factors that increase host susceptibility include diabetes, renal insufficiency, and other underlying conditions; extremes of age; abnormalities of innate defense (e.g., due to genetic polymorphisms; see **Chap. 456**); and medical–surgical interventions that compromise host defenses.

Hospitals' infection-control programs must determine general and specific control measures. Given the prominence of cross-infection, hand hygiene is cited traditionally as the most important preventive measure. Health care workers' rates of adherence to hand-hygiene recommendations are abysmally low (often <50%). Reasons cited include inconvenience, time pressures, and skin damage from frequent washing. Sinkless alcohol rubs are quick and highly effective and may improve hand condition since they contain emollients and allow the retention of natural protective oils that would be removed with repeated rinsing. Use of alcohol hand rubs between patient contacts is recommended for all health care workers except when hands are visibly soiled or after care of a patient who is part of a health care facility outbreak of infection with *C. difficile*, whose spores resist killing by alcohol and require mechanical removal. In these cases, washing with soap and running water is recommended. A number of innovative electronic monitoring systems have been developed to track hand-hygiene adherence and to provide real-time feedback; although this approach is exciting, sustained improvements in rates remain to be seen.

NOSOCOMIAL AND DEVICE-RELATED INFECTIONS

The percentage of nosocomial infections that is due to invasive devices—25–50%—has fallen toward the lower end of that range in recent years; this decline reflects the marked improvements in the use and design of such devices. Intensive education, bundling of evidence-based interventions **(Table 137-3)**, and use of checklists to facilitate adherence have reduced infection rates (Table 137-2), largely through improved asepsis in handling and earlier removal of invasive devices. This progress demonstrates both the effectiveness of infection-control programs and the need to focus current surveillance and interventions on control of the other 50–75% of health care–associated infections.

TABLE 137-3

Examples of Selected Components of Evidence-Based Bundled Interventions to Prevent Common Health Care–Associated Infections and Other Adverse Events[a]

Prevention of Central Venous Catheter Infections
Catheter insertion bundle
• Educate personnel about catheter insertion and care. • Use chlorhexidine to prepare the insertion site. • Use maximal barrier precautions and asepsis during catheter insertion. • Consolidate insertion supplies (e.g., in an insertion kit or cart). • Use a checklist to enhance adherence to the insertion bundle. • Empower nurses to halt insertion if asepsis is breached.
Catheter maintenance bundle
• Cleanse patients daily with chlorhexidine. • Maintain clean, dry dressings. • Enforce hand hygiene among health care workers.
Ask daily: Is the catheter needed? Remove catheter if not needed or used.
Prevention of Ventilator-Associated Events
• Avoid mechanical ventilation whenever possible. • Elevate head of bed to 30–45°. • Decontaminate oropharynx regularly with chlorhexidine (controversial). • Give “sedation vacation” and assess readiness to extubate daily. • Use deep-vein thrombosis prophylaxis (unless contraindicated).
Prevention of Surgical-Site Infections
• Choose a surgeon wisely. • Administer prophylactic antibiotics within 1 h before surgery; discontinue within 24 h. • Limit any hair removal to the time of surgery; use clippers or do not remove hair at all. • Prepare surgical site with chlorhexidine-alcohol.
Prevention of Urinary Tract Infections
• Place bladder catheters only when absolutely needed (e.g., to relieve obstruction), not solely for the provider’s convenience. • Use aseptic equipment and technique for catheter insertion and urinary tract instrumentation. • Minimize manipulation or opening of drainage systems. • Ask daily: Is the bladder catheter needed? Remove catheter if not needed.
Prevention of Pathogen Cross-Transmission
• Cleanse hands with alcohol hand rub before and after all contacts with patients or their environments.

[a]See text for additional interventions to prevent device- and procedure-associated infections; checklists and personnel education have been recommended as management tools for each of the prevention bundles.

Source: Adapted from information presented at the following websites: *www.cdc.gov/hicpac/pubs.html; www.cdc.gov/HAI/index.html; www.ihi.org*.

URINARY TRACT INFECTIONS

Urinary tract infections (UTIs) account for ~14% of nosocomial infections; up to 3% of bacteriuric patients develop bacteremia. Although UTIs contribute marginally to prolongation of hospital stay and may have an attributable cost in the range of only $900, these infections are sources for spread of antibiotic-resistant bacteria. Most nosocomial UTIs are associated with preceding instrumentation or indwelling bladder catheters, which create a 3–7% risk of infection each day. UTIs generally are caused by pathogens that spread up the periurethral space from the patient's perineum or gastrointestinal tract—the most common pathogenesis in women—or via intraluminal contamination of urinary catheters, usually due to cross-infection by caregivers who are emptying drainage bags. Pathogens come occasionally from inadequately disinfected equipment and rarely from contaminated supplies.

A U.S. national demonstration project has shown that, with organized control programs, rates of catheter use and UTI can be reduced. Hospitals should monitor performance measures (Table 137-3). Prompts to clinicians to assess a patient's need for continued use of an indwelling bladder catheter can improve removal rates and lessen the risk of UTI. Guidelines for managing postoperative urinary retention (e.g., with bladder scanners) may limit use or duration of catheterization. Other prevention strategies have included the use of topical meatal antimicrobial agents, drainage bag disinfectants, and anti-infective catheters. None of the latter three measures is considered routine.

Irrigation of catheters, with or without antimicrobial agents, may actually increase the risk of infection. A condom catheter for men without bladder obstruction may be more acceptable than an indwelling catheter and may lessen the risk of UTI if maintained carefully. The role of suprapubic catheters in preventing infection is not well defined.

Treatment of UTIs is based on the results of quantitative urine cultures **(Chap. 130)**. The most common pathogens are *Escherichia coli*, nosocomial gram-negative bacilli, enterococci, and *Candida*. Several caveats apply in the treatment of institutionally acquired infection. First, in patients with chronic indwelling bladder catheters, especially those in long-term-care facilities, the catheter flora—microorganisms living on encrustations within the catheter lumen—may differ from actual urinary tract pathogens. Therefore, for suspected UTI in the setting of chronic catheterization (especially in women), it is useful to replace the bladder catheter and to obtain a freshly voided urine specimen. Second, as in all nosocomial infections, at the time treatment is initiated on the basis of a positive culture, it is useful to repeat the culture to verify the persistence of infection. Third, the frequency with which UTIs occur may lead to the erroneous assumption that the urinary tract alone is the source of infection in a febrile hospitalized patient. Fourth, recovery of *Staphylococcus aureus* from urine cultures may result from hematogenous seeding and indicate an occult systemic infection. Finally, although *Candida* is now the most common pathogen in nosocomial UTIs among patients on intensive care units (ICUs), treatment of candiduria is often unsuccessful and is recommended only when there is upper-pole or bladder-wall invasion, obstruction, neutropenia, or immunosuppression.

PNEUMONIA

Pneumonia accounts for ~24% of nosocomial infections; ventilator-associated pneumonia (VAP) occurs in ~10% of patients on ventilators, and these infections are reported as responsible for 12–14 extra hospital days and ~$40,000 in extra costs per episode. Most cases of bacterial nosocomial pneumonia are caused by aspiration of endogenous or hospital-acquired oropharyngeal (and occasionally gastric) flora. Nosocomial pneumonias have been associated with more deaths than have infections at any other body site. However, attributable mortality rates suggest that the risk of dying from nosocomial pneumonia is affected greatly by other factors, including comorbidities, inadequate antibiotic treatment, and the involvement of specific pathogens (particularly *Pseudomonas aeruginosa* or *Acinetobacter*). Surveillance and accurate diagnosis of pneumonia have been problematic in hospitals because many patients, especially those in the ICU, have abnormal chest roentgenographs, fever, and leukocytosis potentially attributable to multiple causes. This diagnostic uncertainty has led to questions about the reliability of surveillance data and a refocus from VAP to ventilator-associated events (VAEs), conditions, and complications, for which worsening physiologic parameters, such as oxygenation, are key metrics. VAEs occur in as many as 5–10% of patients using mechanical ventilators.

There is increasing interest in health care–associated pneumonia in patients who are on general medical and surgical wards and are not receiving mechanical ventilation. Viral pneumonias, which are particularly important in pediatric and immunocompromised patients, are discussed in the virology section and in **Chap. 121**.

Risk factors for nosocomial pneumonia include those events that increase colonization by potential pathogens (e.g., prior antimicrobial therapy, contaminated ventilator circuits or equipment, or decreased gastric acidity); those that facilitate aspiration of oropharyngeal contents into the lower respiratory tract (e.g., intubation, decreased levels of consciousness, or presence of a nasogastric tube); and those that reduce host defense mechanisms in the lung and permit overgrowth of aspirated pathogens (e.g., chronic obstructive pulmonary disease or upper abdominal surgery).

Control measures for pneumonia (Table 137-3) are aimed at frequent testing of readiness for extubation, which also can shorten ICU lengths of stay; remediation of risk factors in patient care (e.g., minimizing aspiration-prone supine positioning); and aseptic care of respirator equipment. Although the benefits of selective decontamination of the oropharynx and gut with nonabsorbable antimicrobial agents—a practice avoided in the United States because of concerns about antibiotic resistance—have been controversial, a randomized multicenter Dutch trial demonstrated lowered ICU mortality rates among patients on mechanical ventilation who underwent oropharyngeal decontamination.

Among the logical preventive measures that require further investigation are placement of endotracheal tubes that provide channels for subglottic drainage of secretions, which has been associated with reduced infection risks during short-term postoperative use, and noninvasive mechanical ventilation whenever feasible. It is noteworthy that reducing the rate of VAP often has not reduced overall ICU mortality; this fact suggests inadequacies of surveillance and indicates that this infection at times is a marker for patients with an otherwise-heightened risk of death.

The most likely pathogens for nosocomial pneumonia and treatment options are discussed in **Chap. 121**. Several considerations regarding diagnosis and treatment are worth emphasizing. First, clinical criteria for diagnosis (e.g., fever, leukocytosis, development of purulent secretions, new or changing radiographic infiltrates, and changes in oxygen requirement or ventilator settings) have high sensitivity but relatively low specificity. These criteria are useful for selecting patients for bronchoscopic or nonbronchoscopic procedures that yield lower respiratory tract samples protected from upper-tract contamination; quantitative cultures of such specimens have diagnostic sensitivities in the range of 80%. Second, early-onset nosocomial pneumonia, which manifests within the first 4 days of hospitalization, is most often caused by community-acquired pathogens such as *Streptococcus pneumoniae* and *Haemophilus* species, although some studies have challenged this view. Late-onset pneumonias most commonly are due to *S. aureus*, *P. aeruginosa*, *Enterobacter* species, *Klebsiella pneumoniae*, or *Acinetobacter*. Third, one multicenter study suggested that 8 days is an appropriate duration of therapy for nosocomial pneumonia and lessened emergence of resistant pathogens. Fourth, a controversial study of health care–associated pneumonia suggested that therapy based on guidelines from professional societies did not improve patient outcomes. Finally, in febrile patients (particularly those who have tubes inserted through the nares), occult bacterial sinusitis and otitis media should be considered.

SURGICAL WOUND INFECTIONS

Wound infections occur in 280,000 or more patients each year, account for ~24% of nosocomial infections, contribute up to 11 extra postoperative hospital days, and result in $3000–29,000 in extra costs, depending on the operative procedure and pathogen(s). The average wound infection has an incubation period of 5–7 days—longer than many postoperative stays. For this reason and because many procedures are now performed on an outpatient basis, the incidence of wound infections has become more difficult to assess. These infections usually are caused by the patient's endogenous or hospital-acquired skin and mucosal flora and occasionally are due to airborne spread of skin squames that may be shed into the wound from members of the operating-room team. True airborne spread of infection through droplet nuclei is rare in operating rooms unless there is a disseminator (e.g., of group A streptococci or staphylococci) among the staff. In general, the common risks for postoperative wound infection are related to the surgeon's technical skill, the patient's underlying conditions (e.g., diabetes mellitus, obesity) or advanced age, and inappropriate timing of antibiotic prophylaxis. Additional risks include the presence of drains, prolonged preoperative hospital stays, shaving of operative sites by razor the day before surgery, long duration of surgery, and infection at remote sites (e.g., untreated UTI).

The substantial global morbidity and costs associated with these infections have led to international guidelines (*http://www.who.int/gpsc/ssi-guidelines/en/*) in addition to existing national prevention programs and recommendations for bundling of preventive measures (Table 137-3). Other measures include attention to technical surgical issues (e.g., avoiding open or prophylactic drains), operating-room asepsis, and preoperative therapy for active infection. Reporting surveillance results to surgeons has been associated with reductions in infection rates. Preoperative administration of intranasal mupirocin to patients colonized with *S. aureus*, preoperative antiseptic bathing, intra- and postoperative oxygen supplementation, and attention to patients' blood glucose levels and body temperature have been controversial because of conflicting study results, but evidence seems mostly to favor these interventions.

The process of diagnosing and treating wound infections begins with a careful assessment of the surgical site. Diagnosis of infections of prosthetic devices, such as orthopedic implants, may be complicated when pathogens are cloistered in prosthesis-adherent biofilms; cultures of sonicates from explanted prosthetic joints have been more sensitive.

The most common pathogens in postoperative wound infections are *S. aureus*, coagulase-negative staphylococci, and enteric and anaerobic bacteria. In rapidly progressing postoperative infections manifesting within 24–48 h of a surgical procedure, the level of suspicion regarding group A streptococcal or clostridial infection **(Chaps. 143 and 149)** should be high. Treatment of postoperative wound infections requires adequate source control—i.e., drainage or surgical excision of infected or necrotic material—and antibiotic therapy aimed at the most likely or laboratory-confirmed pathogens.

INFECTIONS RELATED TO VASCULAR ACCESS AND MONITORING

Intravascular device–related bacteremias cause ~10–15% of nosocomial infections; central vascular catheters (CVCs) account for most of these bloodstream infections, although peripheral catheters may be under-appreciated as a source of nosocomial bacteremia. National estimates have indicated that ~72,000 primary bloodstream infections occur in the United States each year. CVC infections have had estimated attributable mortality rates of 12–25%, an excess length of hospital stay of 7–15 days, and an estimated cost of $31,000–65,000 per episode; one-third to one-half of these episodes occurred in ICUs. However, infection rates have dropped steadily (Table 137-2) since the publication of guidelines by the Healthcare Infection Control Practices Advisory Committee (HICPAC) in 2002. With increasing care of seriously ill patients in the community, vascular catheter–associated bloodstream infections acquired in outpatient settings are becoming more frequent. Broader surveillance for infections—outside ICUs and even outside hospitals—is more routine.

Catheter-related bloodstream infections derive largely from the cutaneous microflora of the insertion site, with pathogens migrating extraluminally to the catheter tip, usually during the first week after insertion—a risk that has been lessened greatly by use of bundled catheter-insertion guidelines. In addition, contamination of the hubs of CVCs or of the ports of needle-less systems may lead to intraluminal infection over longer periods, particularly with surgically implanted or cuffed catheters. Intrinsic (during the manufacturing process) or extrinsic (on-site in a health care facility) contamination of infusate, although rare, is the most common cause of epidemic device-related bloodstream infection. The most common pathogens isolated from vascular device–associated bacteremias include coagulase-negative staphylococci, *S. aureus* (with ≥50% of U.S. isolates resistant to methicillin), enterococci, nosocomial gram-

negative bacilli, and *Candida*. Many pathogens, especially staphylococci, produce extracellular polysaccharide biofilms that facilitate attachment to catheters and provide sanctuary from antimicrobial agents.

Evidence-based bundles of control measures (Table 137-3) have been strikingly effective, eliminating almost all CVC-associated infections in some ICUs. In a systematic review prompted by the global problem of CVC-associated infections in ICUs, prevention bundles were shown to be effective in low-income and middle-income, as well as in high-income, countries. Additional control measures for infections associated with vascular access include use of a chlorhexidine-impregnated patch at the skin-catheter junction; daily bathing of ICU patients with chlorhexidine; application of semitransparent access-site dressings (for ease of bathing and site inspection and protection of the site from secretions); avoidance of the femoral site for catheterization because of a higher risk of infection (most likely related to the density of the skin flora); rotation of peripheral catheters—an under-recognized cause of staphylococcal bacteremia—to a new site at specified intervals (e.g., every 72–96 h) rather than as clinically indicated (a controversial recommendation that may be facilitated by use of an IV therapy team); and use of aseptic technique when accessing pressure transducers or other vascular ports.

Unresolved issues include the role of gut translocation rather than vascular-access sites as a cause of primary bacteremia in immunocompromised patients and the implications for surveillance definitions; the best frequency for rotation of CVC sites (given that guidewire-assisted catheter changes at the same site do not lessen and can even increase infection risk); the relative risk posed by peripherally inserted central catheters (PICC lines); and the risk–benefit of prophylactic use of vancomycin, alcohol, or other solutions as catheter "locks"—i.e., concentrated anti-infective solutions instilled into the catheter lumen —for high-risk patients.

Vascular device–related infection is suspected on the basis of the appearance of the catheter site or the presence of fever or bacteremia without another source in patients with vascular catheters. The diagnosis is confirmed by the recovery of the same species of microorganism from peripheral-blood cultures (preferably two samples drawn from peripheral veins by separate venipunctures) and from semiquantitative or quantitative cultures of the vascular catheter tip. Potential complications of drawing cultures from the CVC for diagnostic purposes include false-positive results (e.g., due to catheter hub contamination) and increased risk of CVC infection as a result of interrupting a closed system. When infusion-related sepsis is considered (e.g., because of the abrupt onset of fever or shock temporally related to infusion therapy), a sample of the infusate or blood product should be retained for culture.

Therapy for vascular access–related infection is directed at the pathogen recovered from the blood and/or infected site. Important considerations in treatment are the need for an echocardiogram (to evaluate the patient for endocarditis), the duration of therapy, and the need to remove potentially infected catheters. In one report, approximately one-fourth of patients with intravascular catheter–associated *S. aureus* bacteremia who were studied by transesophageal echocardiography had evidence of endocarditis; this test may be useful in determining the appropriate duration of treatment.

Detailed consensus guidelines for the management of intravascular catheter–related infections have been published and recommend catheter removal in most cases of bacteremia or fungemia due to nontunneled CVCs. When attempting to salvage a potentially infected catheter, some clinicians use the "antibiotic lock" technique, which may facilitate penetration of infected biofilms, in addition to systemic antimicrobial therapy (see *www.idsociety.org/Other_Guidelines/*).

ISOLATION TECHNIQUES

Written policies for the isolation of infectious patients are standard for infection-control programs. The CDC has basic isolation guidelines for all components of health care, including acute-care hospitals and long-term, ambulatory, and home-care settings (see *www.cdc.gov/hicpac/pdf/isolation/Isolation2007.pdf*), as well as recommendations for the control of multidrug-resistant organisms.

Standard precautions are designed for the care of all patients in hospitals and aim to reduce the risk of transmission of microorganisms from both recognized and unrecognized sources. These precautions include gloving and hand cleansing for potential contact with (1) blood; (2) all other body fluids, secretions, and excretions, whether or not they contain visible blood; (3) nonintact skin; and (4) mucous membranes. Depending on exposure risks, standard precautions also include use of masks, eye protection, and gowns.

Precautions for the care of patients with potentially contagious clinical syndromes (e.g., acute diarrhea) or with suspected or diagnosed colonization or infection by transmissible pathogens are based on probable routes of transmission: *airborne*, *droplet*, or *contact*, for which personnel don, at a minimum, N95 respirators, surgical face masks, or glove and gown, respectively. Sets of precautions may be combined for diseases that have more than one route of transmission (e.g., contact and airborne isolation for varicella).

Some prevalent antibiotic-resistant pathogens, particularly those that colonize the gastrointestinal tract (e.g., vancomycin-resistant enterococci [VRE] and even multidrug-resistant gram-negative bacilli such as strains of *K. pneumoniae* and other Enterobacteriaceae that produce carbapenemases [carbapenem-resistant Enterobacteriaceae, or CRE]), may be present on *intact* skin of patients in hospitals (the "fecal patina"). This issue has led some experts to recommend gloving for all contact with patients who are acutely ill and/or in high-risk units, such as ICUs or LTACHs, and daily bathing of all ICU and LTACH patients with chlorhexidine to remove this veneer of antibiotic-resistant bacteria. Wearing gloves does not replace the need for hand hygiene because hands sometimes (in up to 20% of interactions) become contaminated during wearing or removal of gloves. To further lessen the risk of self-contamination, some personnel shun ties and long-sleeves, a practice that is understandable although not scientifically supported as a means of reducing the spread of resistant bacteria.

EPIDEMIC AND EMERGING PROBLEMS

Full-blown epidemics probably account for <5% of nosocomial infections, but mini-clusters of a few infections that result from time-limited gaps in asepsis may be more common. The investigation and control of nosocomial epidemics require that infection control personnel (1) develop a case definition, (2) confirm that an outbreak really exists (since apparent epidemics may actually be pseudo-outbreaks due to surveillance or laboratory artifacts), (3) review aseptic practices and disinfectant use, (4) determine the extent of the outbreak, (5) perform an epidemiologic investigation, which may require a case–control study to determine sources and modes of transmission, (6) work closely with microbiology personnel to culture for common sources or personnel carriers as appropriate and to provide molecular typing—by pulsed-field gel electrophoresis (PFGE) or whole-genome sequencing (WGS)—of epidemiologically important isolates, and (7) heighten surveillance to judge the effect of control measures. Control measures generally include reinforcing routine aseptic practices and hand hygiene, ensuring appropriate isolation of cases (and instituting cohort isolation and nursing if needed), and implementing further controls on the basis of the investigation's findings. Examples of some emerging and potential epidemic problems follow.

VIRAL RESPIRATORY INFECTIONS: PANDEMIC INFLUENZA

Infections caused by the severe acute respiratory syndrome (SARS)–associated coronavirus challenged health care systems globally in 2003 **(Chap. 194)**, and in 2012 Middle East respiratory syndrome coronavirus (MERS-CoV) emerged as a more geographically localized problem **(Chap. 194)**. For SARS, basic infection-control measures helped to keep the worldwide case and death counts at ~8000 and ~800, respectively. The epidemiology of SARS—spread largely in households once patients were ill or in hospitals—contrasts markedly with that of influenza **(Chap. 195)**, which is often contagious a day before symptom onset; can spread rapidly in the community among nonimmune persons; and, even in its seasonal variety, kills as many as 35,000 persons in the United States each year.

Control of seasonal influenza has depended on (1) the use of effective vaccines, with increasingly broad evidence-based recommendations for vaccination of children, the general public, and health care workers; (2) the use of antiviral medications for early treatment and for prophylaxis as part of outbreak control, especially for high-risk patients and in high-risk settings like nursing homes or hospitals; and (3) infection control (surveillance and droplet precautions) for symptomatic patients. Controversial infection-control issues have been the questionable role of airborne spread of influenza and the historical embarrassingly low rates of vaccination among health care workers, which have now markedly improved, in part as a result of mandated vaccination policies in many hospitals.

With the occurrence of localized outbreaks of avian influenza (due to H5N1 and other strains) in Asia over the past few years, concerns about potential pandemic influenza led to (1) recommendations for universal respiratory hygiene and cough etiquette (basically, "cover your cough"), as described and promoted in the CDC's *Guideline for Isolation Precautions*, and for source containment (e.g., use of face masks and spatial separation) for outpatients with potentially infectious respiratory illnesses; (2) re-examinations of the value in the 1918–1919 influenza pandemic of nonpharmacologic interventions, such as social distancing (e.g., closing of schools and community venues); and (3) debate about the level of respiratory protection required for health care workers (i.e., whether to use the higher-efficiency N95 respirators recommended for airborne isolation rather than the surgical masks used for droplet precautions).

In the spring of 2009, a novel strain of influenza virus—H1N1 (swine flu) virus—caused the first influenza pandemic in four decades. Recombinant events that create new strains (e.g., H7N9) continue to challenge global efforts at infection control and vaccine development **(Chap. 195)**.

EMERGING VIRAL PATHOGENS

The re-emergence of Ebola virus in West Africa has had a global impact on infection-control preparedness and isolation techniques, on situational awareness, and on vaccine development **(Chap. 205)**. The emergence of epidemic Zika virus disease in Brazil and its spread throughout Latin America to the United States has created a major threat to pregnant women and has added to the list yet another potential blood-borne pathogen that requires blood-bank screening **(Chap. 204)**.

NOSOCOMIAL DIARRHEA

Overall rates of *C. difficile*–associated diarrhea **(Chap. 129)** have increased, especially among older patients in U.S. hospitals during the past few years. This increase is related in part to a new, more virulent strain, NAP1/BI/027. There are ~250,000 incident cases, with 14,000 deaths, from *C. difficile*–associated diarrhea in the United States annually. In a CDC multistate survey, *C. difficile* was the most common nosocomial pathogen, causing 12% of health care–associated infections. Use of WGS is improving our understanding of *C. difficile* epidemiology. For now, control measures include judicious use of all antibiotics, especially fluoroquinolone antibiotics that have been implicated in driving outbreaks; heightened suspicion for atypical presentations (e.g., toxic megacolon or leukemoid reaction without diarrhea); enhanced disinfection of isolation rooms with sporicidal agents, such as bleach; and early diagnosis, treatment, and contact precautions. To improve diagnosis, use of more sensitive polymerase chain reaction–based rather than enzyme immunoassay–based testing of diarrheal stool is now recommended, with resultant artificial doubling of infection rates (as patients who are colonized but not clinically infected with *C. difficile* are being detected) in some hospitals. Preliminary data suggest a role for probiotics in the prevention of diarrhea in patients in whom systemic antibiotic therapy is being initiated. Fecal transplantation has had dramatic results in the treatment of relapsing cases of *C.*

difficile–associated diarrhea (Chap. 129). Successes with fecal transplants and probiotics have called attention to the potential role of manipulation of the intestinal microbiome as a broader infection-control strategy (e.g., to eliminate carriage of antibiotic-resistant gram-negative bacilli).

Reports of outbreaks of norovirus infection (Chap. 198) in U.S. and European health care facilities appear to continue to increase, with the virus often introduced by ill visitors or staff. This pathogen should be suspected when nausea and vomiting are prominent aspects of bacterial culture–negative diarrheal syndromes. Contact precautions may need to be augmented by aggressive environmental cleaning (given the persistence of norovirus on inanimate objects), prevention of secondary cases in members of the cleaning staff through an emphasis on the use of personal protective equipment and hand hygiene, and active exclusion of ill staff and visitors.

CHICKENPOX

Infection-control practitioners institute a varicella exposure investigation and control plan whenever health care workers have been exposed to chickenpox (Chap. 188) or have worked while having or during the 24 h before developing chickenpox. Fortunately, routine varicella vaccination of children and susceptible health care employees has made nosocomial spread less common.

MYCOBACTERIA

Important measures for the control of pulmonary tuberculosis (Chap. 173) include prompt recognition, isolation, and treatment of cases; recognition of atypical presentations (e.g., lower-lobe infiltrates without cavitation); use of negative-pressure, 100% exhaust, private isolation rooms with closed doors and at least 6–12 air changes per hour; use of N95 respirators by caregivers entering isolation rooms; possible use of high-efficiency particulate air-filter units and/or ultraviolet lights for disinfecting air when other engineering controls are not feasible or reliable; and follow-up testing of susceptible personnel who have been exposed to infectious patients before isolation. The use of serologic tests, rather than skin tests, in the diagnosis of latent tuberculosis for infection control purposes has become common, mostly for logistic reasons. As tuberculosis once again is on the decline in the United States, we need to remember that the price of freedom—in this instance, from a communicable disease—is eternal vigilance.

An unprecedented multicountry outbreak of postoperative invasive *Mycobacterium chimaera* infections has been traced to commercially contaminated heater–cooler devices used commonly during cardiac surgery. The implicated devices are manufactured by a single company that sells the majority of such devices used globally. Guidelines for diagnosis, treatment, and prevention are being formulated.

GROUP A STREPTOCOCCAL INFECTIONS

The potential for an outbreak of group A streptococcal infection (Chap. 143) should be considered when even one or two nosocomial cases occur. Most outbreaks involve surgical wounds and are due to the presence of an asymptomatic carrier in the operating room. Investigation can be confounded by carriage at extrapharyngeal sites such as the rectum and vagina.

FUNGAL INFECTIONS

When dusty areas—common sources of fungal spores—are disturbed during hospital repairs or renovation, the spores become airborne. Inhalation of spores by immunosuppressed (especially neutropenic) patients creates a risk of pulmonary and/or paranasal sinus infection and disseminated aspergillosis (Chap. 212). Routine surveillance among neutropenic patients for infections with filamentous fungi, such as *Aspergillus* and *Fusarium*, helps hospitals to assess environmental risks. As a matter of routine, hospitals should inspect and clean air-handling equipment; review all planned renovations with infection-control personnel and construct appropriate barriers; remove immunosuppressed patients from renovation sites; and consider the use of high-efficiency particulate air-intake filters for rooms housing immunosuppressed patients.

A major multistate iatrogenic outbreak of meningitis, localized spinal or paraspinal infection, and arthritis due to *Exserohilum rostratum* was recognized in 2012 and traced to contamination of an injectable preservative-free steroid product produced by a single compounding pharmacy (Chap. 212).

Candida auris, a pathogen first identified in Japan in 2009, is emerging globally as a cause of invasive health care–associated infections. The international strains are closely related within regions, often multidrug-resistant, and difficult to identify by routine laboratory testing.

LEGIONELLOSIS

Nosocomial *Legionella* pneumonia (Chap. 154) is most often due to contamination of potable water or of water used in decorative fountains. This disease predominantly affects immunosuppressed patients, particularly those receiving glucocorticoid medications. The risk varies greatly within and among geographic regions, depending on the extent of hospital water contamination and on specific hospital practices (e.g., the presence of decorative fountains in hospital lobbies or inappropriate use of nonsterile water in respiratory therapy equipment). The diagnosis of legionellosis should probably be considered more often than it is. If nosocomial cases are detected, environmental samples (e.g., tap water) should be cultured. If cultures yield *Legionella* and if typing of clinical and environmental isolates reveals a correlation, eradication measures should be pursued. An alternative approach is to periodically culture tap water in wards housing high-risk patients. If *Legionella* is found, a concerted effort should be made to culture samples from all patients with nosocomial pneumonia for *Legionella* and to introduce engineering controls to reduce or eliminate water-borne *Legionella* within the facility.

ANTIBIOTIC-RESISTANT BACTERIA: SURVEILLANCE, CONTROL, AND ANTIBIOTIC STEWARDSHIP

Emerging multidrug-resistant bacteria like CRE are harbingers of a potential postantibiotic era. Control of resistance depends on close laboratory surveillance, with early detection of problems; on aggressive reinforcement of routine asepsis; on implementation of barrier precautions for all colonized and/or infected patients; on use of patient-surveillance cultures to more fully ascertain the extent of patient colonization; on antimicrobial stewardship in humans and animals to lessen ecologic pressures; and on timely initiation of an epidemiologic investigation when rates increase. Advanced molecular diagnostics (e.g., PFGE and, most recently, WGS) can help differentiate an outbreak due to a single strain (which necessitates an emphasis on hand hygiene and an evaluation of potential common-source exposures) from a polyclonal outbreak (which requires an emphasis on antibiotic prudence and device bundles; Table 137-3). Continuing emergence of multidrug-resistant organisms suggests that control efforts have been insufficient and that heightened interventions and global strategies are needed urgently (see *www.cdc.gov/drugresistance/threat-report-2013/* and *https://www.cdc.gov/vitalsigns/protect-patients/index.html*); this need is highlighted by the creation of the U.S. Presidential Advisory Council on Combating Antibiotic-Resistant Bacteria and by the U.N. General Assembly's 2016 Declaration (see *https://www.hhs.gov/ash/advisory-committees/paccarb, http://www.un.org/pga/71/2016/09/21/press-release-hl-meeting-on-antimicrobial-resistance/, and www.gov.uk/government/publications/uk-5-year-antimicrobial-resistance-strategy-2013-to-2018/).*

To facilitate tracking of resistance problems, the CDC's innovative, interactive online Antibiotic Resistance Patient Safety Atlas allows users to search state-level resistance data (*http://www.cdc.gov/hai/surveillance/ar-patient-safety-atlas.html*). The European Centre for Disease Prevention and Control also has online resistance reporting (*http://ecdc.europa.eu/en/healthtopics/antimicrobial-resistance-and-consumption/antimicrobial_resistance/EARS-Net/Pages/EARS-Net.aspx*).

Currently, several antibiotic resistance problems are of particular concern. First, over the past decade or so, the emergence of community-associated methicillin-resistant *S. aureus* (CA-MRSA) has been dramatic in many countries, with as many as 50% of community-acquired "staph infections" in some U.S. cities now caused by strains resistant to β-lactam antibiotics **(Chap. 142)**. The incursion of CA-MRSA into hospitals is well documented and has impacted surveillance and control of nosocomial MRSA infections.

Second, in the global emergence of multidrug-resistant gram-negative bacilli, new problems include plasmid-mediated resistance to fluoroquinolones, metallo-β-lactamase-mediated resistance to carbapenems, CRE, and pan-resistant strains of *Acinetobacter*. The problematic plasmid-mediated New Delhi metallo-β-lactamase (NDM) has been highly successful in inter-genus transmission and quickly has become a threat worldwide. Many multidrug-resistant gram-negative bacilli are susceptible only to colistin, a drug that is consequently being "re-discovered," or to no available agents. The potential dependence on colistin has led to increased colistin susceptibility testing and recognition of plasmid-mediated resistance to colistin in *E. coli*—first in swine-associated strains from China and now in strains from many countries, including the United States.

Transmission of CRE has been traced in a number of outbreaks to exposure to duodenoscopes used for endoscopic retrograde cholangiopancreatography. The duodenoscope is more intricate than other endoscopes and has an "elevator mechanism" that can be difficult to clean and disinfect. In some—though not all—of these outbreaks, investigators identified breaches of approved cleaning protocols.

Third, there has been renewed recognition of the role of nursing homes, and now LTACHs, in the spread of resistant gram-negative bacilli such as CRE. In some LTACHs, as many as 30–50% of patients may be colonized with CREs. The frequent transfer of patients who are colonized or infected with antibiotic-resistant bacteria between long-term and acute-care facilities has led to studies of the regional spread of antibiotic resistance and the proposal of regional infection-control interventions (see *www.cdc.gov/vitalsigns/stop-spread*).

Fourth, there has been increasing community-based spread of *E. coli* strains harboring an enzyme, CTX-M, that renders them broadly resistant to β-lactam antibiotics. Given the community focus of spread, these strains may be seen as a gram-negative version of CA-MRSA.

Fifth, as a consequence of going abroad, international travelers, especially to Latin America, Asia, and Africa, may become gastrointestinal carriers of multidrug-resistant Enterobacteriaceae that express extended-spectrum β-lactamases (ESBLs) and encode resistance to many commonly used antibiotics. In one study, 34% of returning Dutch travelers had newly acquired resistant strains and were colonized for as long as 6–12 months until their pretravel gastrointestinal microbiomes rebounded. The frequency of clinical consequences—e.g., the emergence of antibiotic-resistant UTIs or local spread of resistant strains—is not yet known.

Finally, clinical infections with MRSA strains exhibiting high-level vancomycin resistance due to VRE-derived plasmids have been reported in a few patients—almost all in the United States and most in Michigan—in the setting of prolonged or repeated treatment with vancomycin and/or VRE colonization.

Colonized personnel who are implicated in nosocomial transmission of multidrug-resistant pathogens and patients who pose a threat may be decontaminated, depending on the pathogen and available decontamination regimens. In a few ICUs, nonabsorbed antimicrobial agents for gastrointestinal decontamination of patients have been used as a temporary emergency control measure for outbreaks of infection due to gram-negative bacilli. Manipulation of patients' intestinal microbiomes may prove to be a more durable strategy to control outbreaks of multidrug-resistant pathogens that have a gastrointestinal reservoir.

In several cluster-randomized controlled trials over the past ~15 years, source control—i.e., removal of patients' fecal patinas—by daily bathing with chlorhexidine has reduced the risk of bacteremia in ICU patients. "Search-and-destroy" methods—i.e., active surveillance cultures to detect and isolate the

"resistance iceberg" of patients colonized with MRSA—are credited with elimination of nosocomial MRSA in the Netherlands and Denmark. In a multicenter trial in the United States, universal source control with chlorhexidine and nasal mupirocin was significantly more effective in controlling MRSA than was a search-and-destroy approach and led to control of other pathogens as well, providing a broad (horizontal) rather than a narrower (vertical) intervention (see *www.ahrq.gov/professionals/systems/hospital/universal_icu_decolonization*). For some pathogens, such as VRE, enforcement of environmental cleaning also reduces cross-transmission risk.

Because the excessive use of broad-spectrum antibiotics underlies many resistance problems, antibiotic stewardship programs will be mandatory in acute care hospitals (see *https://www.cdc.gov/getsmart/healthcare/implementation/core-elements.html*) and are being promulgated actively for long-term care and outpatient facilities (see *https://www.cdc.gov/longtermcare/prevention/antibiotic-stewardship.html* and *https://www.cdc.gov/mmwr/volumes/65/rr/rr6506a1.htm*). The main tenets are to restrict the use of particular agents to narrowly defined indications in order to limit selective pressure on the nosocomial flora; to treat with the shortest efficacious courses; and, when broad-spectrum therapy is begun empirically in critically ill patients, to de-escalate treatment as soon as possible on the basis of the results of culture and susceptibility tests.

From a broader perspective, the One Health movement cites what many would call the overriding role of antimicrobial use in animal husbandry as a driver of resistance. This concern has led to recommendations for eliminating the use of antibiotics for growth promotion and as prophylaxis for feed animals. The United States lags behind some European countries in control of veterinary use of antimicrobial drugs.

Key to an understanding of the success of antibiotic stewardship initiatives is better surveillance of antibiotic use among humans and other animals at regional and national levels.

BIOTERRORISM AND OTHER SURGE-EVENT PREPAREDNESS

The horrific attack on the World Trade Center in New York City on September 11, 2001; subsequent mailings of anthrax spores in the United States; the Boston Marathon bombing in 2013; and ongoing terrorist activities globally have made bioterrorism a prominent source of concern to hospital infection-control programs **(Chap. S2)**. The essentials for hospital preparedness entail education, internal and external communication, and situational awareness. Up-to-date information is available from the CDC (see *www.emergency.cdc.gov/bioterrorism*).

EMPLOYEE HEALTH SERVICE ISSUES

An institution's employee health service is critical for infection control. New employees should be processed through the service, where a contagious-disease history can be taken; evidence of immunity to a variety of diseases, such as hepatitis B, chickenpox, measles, mumps, and rubella, can be sought; immunizations for hepatitis B, measles, mumps, rubella, varicella, and pertussis can be given as needed (an especially important step with the resurgence in the United States of vaccine-preventable diseases such as pertussis, mumps, and measles); baseline tuberculosis testing can be performed; and education about personal responsibility for infection control can be initiated.

The employee health service must have protocols for dealing with workers exposed to contagious diseases (e.g., influenza) and those percutaneously or mucosally exposed to the blood of patients infected with HIV or hepatitis B or C virus. Protocols are also needed for dealing with caregivers who have common contagious diseases (such as chickenpox, group A streptococcal infection, influenza or another respiratory infection, or infectious diarrhea) and for those who have less common but high-visibility public health problems (such as chronic hepatitis B or C or HIV infection) for which exposure-control guidelines have been published by the CDC and by the Society for Healthcare Epidemiology of America.

FURTHER READING

Allegranzi B et al: New WHO recommendations on preoperative measures for surgical site infection prevention: An evidence-based global perspective. Lancet Infect Dis 16:e276, 2016. [PubMed: 27816413]

Allegranzi B et al: New WHO recommendations on intraoperative and postoperative measures for surgical site infection prevention: An evidence-based global perspective. Lancet Infect Dis 16:e288, 2016. [PubMed: 27816414]

Magill SS et al: Multistate point-prevalence survey of health care–associated infections. N Engl J Med 370:1198, 2014.

Morency-Potvin P et al: Antimicrobial stewardship: How the microbiology laboratory can right the ship. Clin Microbiol Rev 30:381, 2016.

Saint S et al: A program to prevent catheter-associated urinary tract infection in acute care. N Engl J Med 374:2111, 2016.

Weiner LM et al: Antimicrobial-resistant pathogens associated with healthcare-associated infections: Summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2011–2014. Infect Control Hosp Epidemiol 37:1288, 2016.

Yokoe DS et al: A compendium of strategies to prevent healthcare-associated infections in acute care hospitals: 2014 updates. Infect Control Hosp Epidemiol 35:S21, 2014. [PubMed: 25376067]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 138: Infections in Transplant Recipients

Robert W. Finberg; Joyce D. Fingeroth

INTRODUCTION

This chapter considers aspects of infection unique to patients receiving transplanted tissue. The evaluation of infections in transplant recipients involves consideration of both the donor and the recipient of the transplanted cells or organ. Two central issues are of paramount importance: (1) infectious agents (particularly viruses, but also bacteria, fungi, and parasites) can be introduced into the recipient by the donor; and (2) treatment of the recipient with medicine to prevent rejection can suppress normal immune responses, greatly increasing susceptibility to infection. Thus, what might have been a latent or asymptomatic infection in an immunocompetent donor or in the recipient prior to therapy can become a life-threatening problem when the recipient becomes immunosuppressed. The pretransplantation evaluation of each patient should be guided by an analysis of both (1) what infections the recipient is currently harboring, since organisms that exist in a state of latency or dormancy before the procedure may cause fatal disease when the patient receives immunosuppressive treatment; and (2) what organisms are likely to be transmitted by the donor, particularly those to which the recipient may be naïve.

PRETRANSPLANTATION EVALUATION

The Donor

A variety of organisms have been transmitted by organ transplantation. Transmission of infections that may have been latent or not clinically apparent in the donor has resulted in the development of specific donor-screening protocols. Results from routine blood-bank studies, including those for antibodies to *Treponema pallidum* (syphilis), *Trypanosoma cruzi*, hepatitis B and C viruses, HIV-1 and -2, and human T-lymphotropic virus types 1 and 2 (HTLV-1 and -2), should be documented. Serologic studies should be ordered to identify latent infection with viruses such as herpes simplex virus types 1 and 2 (HSV-1, HSV-2), varicella-zoster virus (VZV), cytomegalovirus (CMV), Epstein-Barr virus (EBV), and Kaposi's sarcoma–associated herpesvirus (KSHV, also known as human herpesvirus type 8); acute infection with hepatitis A virus; and infection with the common parasite *Toxoplasma gondii*. Donors should be screened for parasites such as *Strongyloides stercoralis, T. cruzi*, and *Schistosoma* species if they have lived in endemic areas. Clinicians caring for prospective organ donors should examine chest radiographs for evidence of granulomatous disease (e.g., caused by mycobacteria or fungi) and should perform skin testing or obtain blood for immune cell-based assays that detect active or latent *Mycobacterium tuberculosis* infection. An investigation of the donor's dietary habits (e.g., consumption of raw meat or fish or of unpasteurized dairy products), occupations or avocations (e.g., gardening or spelunking), and travel history (e.g., travel to areas with endemic fungi causing infections such as blastomycosis, coccidioidomycosis, and histoplasmosis) also is indicated and may mandate additional testing. A number of unusual parasites (including *Balamuthia mandrillaris*) have been transplanted in kidneys. Uncommonly diagnosed viruses, including lymphocytic choriomeningitis virus (LCMV), West Nile virus, Zika virus, dengue virus, and rabies virus, can be transplanted in organs and are likely to be difficult to diagnose in recipients. If an unusual parasite or uncommon virus is identified in a

transplant recipient, the organ-donor organization and caregivers for recipients of other organs isolated from the same donor should be notified immediately.

Creutzfeldt-Jakob disease has been transmitted through corneal transplants; however, to what degree it can be transmitted by transfused blood is not known. Variant Creutzfeldt-Jakob disease can be transmitted with transfused non-leukodepleted blood, posing a theoretical risk to transplant recipients.

THE RECIPIENT

It is expected that the recipient will have been even more comprehensively assessed than the donor. Additional studies recommended for the recipient include evaluation for acute respiratory viruses and gastrointestinal pathogens in the immediate pretransplantation period. An important caveat is that, because of immune dysfunction resulting from chemotherapy or underlying chronic disease, serologic testing of the recipient may prove less reliable than usual.

THE DONOR CELLS/ORGAN

Careful attention to the sterility of the medium used to process the donor organ, combined with meticulous microbiologic evaluation, reduces rates of transmission of bacteria (or, rarely, yeasts) that may be present or grow in the organ culture medium. From 2 to >20% of donor kidneys are estimated to be contaminated with bacteria—in most cases, with the organisms that colonize the skin or grow in the tissue culture medium used to bathe the donor organ while it awaits implantation. The reported rate of bacterial contamination of transplanted stem cells (bone marrow, peripheral blood, cord blood) is as high as 17% but most commonly is ~1%. The use of enrichment columns and monoclonal antibody depletion procedures results in a higher incidence of contamination. In some cases, because of the clinical situation, contaminated cells have been infused, usually with concomitant administration of antimicrobial agents. In one series of patients receiving contaminated stem cells, 14% had fever or bacteremia, but none died. Results of cultures performed at the time of cryopreservation and at the time of thawing were helpful in guiding therapy for the recipient.

INFECTIONS IN HEMATOPOIETIC STEM CELL TRANSPLANT RECIPIENTS

Transplantation of hematopoietic stem cells (HSCs) from bone marrow or from peripheral or cord blood for cancer, immunodeficiency, or autoimmune disease is marked by a transient state of complete immunologic incompetence. Immediately after myeloablative chemotherapy and transplantation, both innate immune cells (phagocytes, dendritic cells, natural killer cells) and adaptive immune cells (T and B cells) are absent, and the host is extremely susceptible to infection. The reconstitution that follows transplantation has been likened to maturation of the immune system in neonates. The analogy does not entirely predict infections seen in HSC transplant recipients, however, because the stem cells mature in an old host who has several latent infections already. The choice among the current variety of methods for obtaining stem cells is determined by availability and by the need to optimize the chances of a cure for an individual recipient. One strategy is autologous HSC transplantation, in which the donor and the recipient are the same. After chemotherapy, stem cells are collected and are purged (ex vivo) of residual neoplastic populations. Allogeneic HSC transplantation has the advantage of providing a graft-versus-tumor effect. In this case, the recipient is matched to varying degrees for human leukocyte antigens (HLAs) with a donor who may be related or unrelated. In some individuals, nonmyeloablative therapy (mini-allotransplantation) is used and permits recipient cells to persist for some time after transplantation while preserving the graft-versus-tumor effect and sparing the recipient intense myeloablative therapy. Cord-blood transplantation is increasingly used in adults; two independent cord-blood units are

typically required for suitable neutrophil engraftment early after transplantation, even though only one of the units is likely to provide long-term engraftment. In each circumstance, a different balance is struck among the toxicity of conditioning therapy, the need for a maximal graft-versus-target effect, short-term and long-term infectious complications, and the risk of graft-versus-host disease (GVHD; acute versus chronic). The various approaches differ in terms of reconstitution speed, cell lineages introduced, and likelihood of GVHD—all factors that can produce distinct effects on the risk of infection after transplantation **(Table 138-1)**. Despite these caveats, most infections occur in a predictable time frame after transplantation **(Table 138-2)**.

TABLE 138-1

Risk of Infection, by Type of Hematopoietic Stem Cell Transplant

Type of Hematopoietic Stem Cell Transplant	Source of Stem Cells	Risk of Early Infection: Neutrophil Depletion	Risk of Late Infection: Impaired T and B Cell Function	Risk of Ongoing Infection: GVHD and Iatrogenic Immunosuppression	Graft vs Tumor Effect
Autologous	Recipient (self)	High risk; neutrophil recovery sometimes prolonged	~1 year	Minimal to no risk of GVHD and late-onset severe infection	None (–)
Syngeneic (genetic twin)	Identical twin	Low risk; 1–2 weeks for neutrophil recovery	~1 year	Minimal risk of GVHD and late-onset severe infection	+/–
Allogeneic related	Sibling	Low risk; 1–2 weeks for neutrophil recovery	~1 year	Minimal to moderate risk of GVHD and late-onset severe infection	++
Allogeneic related	Child/parent (haploidentical)	Intermediate risk; 2–3 weeks for neutrophil recovery	1–2 years	Moderate risk of GVHD and late-onset severe infection	++++
Allogeneic unrelated adult	Unrelated donor	Intermediate risk; 2–3 weeks for neutrophil recovery	1–2 years	High risk of GVHD and late-onset severe infection	++++
Allogeneic unrelated cord blood	Unrelated cord-blood units (×2)	Intermediate to high risk; neutrophil recovery sometimes prolonged	Prolonged	Minimal to moderate risk of GVHD and late-onset severe infection	++++
Allogeneic mini (nonmyeloablative)	Donor (transiently coexisting with recipient cells)	Low risk; neutrophil counts close to normal	1–2+ years	Variable risk of GVHD and late-onset severe infection[a]	++++ (but develops slowly)

[a]Depending on the disparity of the match (major and minor histocompatibility antigens), GVHD may be severe or mild, the requirement for immunosuppression intense or minimal, and the risk of severe late infections coordinate with the degree of immunosuppression.

Abbreviation: GVHD, graft-versus-host disease.

TABLE 138-2

Common Sources of Infection after Hematopoietic Stem Cell Transplantation

Infection Site	Period after Transplantation		
	Early (<1 Month)	**Middle (1–4 Months)**	**Late (>6 Months)**
Disseminated	Aerobic bacteria (gram-negative, gram-positive)	*Candida, Aspergillus,* EBV	Encapsulated bacteria (*Streptococcus pneumoniae, Haemophilus influenzae, Neisseria meningitidis*)
Skin and mucous membranes	HSV	HHV-6	VZV, HPV (warts)
Lungs	Aerobic bacteria (gram-negative, gram-positive), *Candida, Aspergillus,* other molds, HSV	CMV, seasonal respiratory viruses, *Pneumocystis, Toxoplasma*	*Pneumocystis, Nocardia, S. pneumoniae*
Gastrointestinal tract	*Clostridium difficile*	CMV, adenovirus	EBV, CMV
Kidney		BK virus, adenovirus	
Brain		HHV-6, *Toxoplasma*	*Toxoplasma,* JC virus (rare)
Bone marrow		CMV, HHV-6	CMV, HHV-6

Abbreviations: CMV, cytomegalovirus; EBV, Epstein-Barr virus; HHV-6, human herpesvirus type 6; HPV, human papillomavirus; HSV, herpes simplex virus; VZV, varicella-zoster virus.

BACTERIAL INFECTIONS

In the first month after HSC transplantation, infectious complications are similar to those in granulocytopenic patients receiving chemotherapy for acute leukemia **(Chap. 70)**. Because of the anticipated 1- to 4-week duration of neutropenia and the high rate of bacterial infection in this population, many centers give prophylactic antibiotics to patients upon initiation of myeloablative therapy. Quinolones decrease the incidence of gram-negative bacteremia among these patients. Bacterial infections are common in the first few days after HSC transplantation. The organisms involved are predominantly those found on skin, mucosa, or IV catheters (*Staphylococcus aureus*, coagulase-negative staphylococci, streptococci) or aerobic bacteria that colonize the bowel (*Escherichia coli, Klebsiella, Pseudomonas*). *Bacillus cereus*, although rare, has emerged as a pathogen early after transplantation and can cause meningitis, which is unusual in these patients. Chemotherapy, use of broad-spectrum antibiotics, and delayed reconstitution of humoral immunity place HSC transplant patients at risk for diarrhea and colitis caused by *Clostridium difficile* overgrowth and toxin

production. Reconstitution of the bowel with microbial flora from donors (“fecal transplantation”) has been successful in drug-resistant cases **(Chap. 129)**.

Beyond the first few days of neutropenia, infections with nosocomial pathogens (e.g., vancomycin-resistant enterococci, *Stenotrophomonas maltophilia*, *Acinetobacter* species, and extended-spectrum β-lactamase-producing gram-negative bacteria) as well as with filamentous bacteria (e.g., *Nocardia* species) become more common. Vigilance is indicated, particularly for patients with a history of active or known latent tuberculosis, even when they have been appropriately pretreated. A form of bacterial colitis among cord-blood recipients has occurred 90–300 days after transplantation, responds to antimicrobial agents such as metronidazole, and—as determined by polymerase chain reaction (PCR) testing of biopsy specimens—may be attributed to the common bacterium *Bradyrhizobium enterica* (related to *B. japonicum*). Episodes of bacteremia due to encapsulated organisms mark the late posttransplantation period (>6 months after HSC reconstitution); patients who have undergone splenectomy and those with persistent hypogammaglobulinemia are at particular risk.

FUNGAL INFECTIONS

Beyond the first week after HSC transplantation, fungal infections become increasingly common, particularly among patients who have received broad-spectrum antibiotics. As in most granulocytopenic patients, *Candida* infections are most commonly seen in this setting. However, with increased use of prophylactic fluconazole, infections with *Candida albicans* resistant to fluconazole and naturally fluconazole-resistant fungi—in particular, *Aspergillus* and other non-*Aspergillus* molds (*Rhizopus*, *Fusarium*, *Scedosporium*, *Penicillium*)—have become more common, prompting some centers to replace fluconazole with agents such as micafungin, voriconazole, isavuconazole, or posaconazole. Identification of *Candida auris* as a pathogen has made prophylaxis and treatment of fungal infections more difficult as these organisms are often resistant to most antifungal agents. The role of antifungal prophylaxis with these different agents, in contrast to empirical treatment for suspected infection that is based on a positive β-D-glucan assay or galactomannan antigen test, remains controversial **(Chap. 70)**. Documented infection should be aggressively treated, ideally with agents of proven activity. In patients with GVHD who require prolonged or indefinite courses of glucocorticoids and other immunosuppressive agents (e.g., cyclosporine, tacrolimus [FK506, Prograf], mycophenolate mofetil [CellCept], rapamycin [sirolimus, Rapamune], antithymocyte globulin, or anti-CD52 antibody [alemtuzumab, Campath—an antilymphocyte and antimonocyte monoclonal antibody]), there is a high risk of fungal infection (usually with *Candida* or *Aspergillus*) even after engraftment and resolution of neutropenia. These patients are also at high risk for reactivation of latent fungal infection (histoplasmosis, coccidioidomycosis, or blastomycosis) in areas where endemic fungi reside and after involvement in activities such as gardening or caving. Prolonged use of central venous catheters for parenteral nutrition (lipids) increases the risk of fungemia with *Malassezia*. Some centers administer prophylactic antifungal agents to these patients. Because of the high and prolonged risk of *Pneumocystis jirovecii* pneumonia (especially among patients being treated for hematologic malignancies), most patients receive maintenance prophylaxis with trimethoprim-sulfamethoxazole (TMP-SMX) starting after engraftment and continuing for at least 1 year.

PARASITIC INFECTIONS

The regimen just described for the fungal pathogen *Pneumocystis* may also protect patients seropositive for the parasite *T. gondii*, which can cause pneumonia, visceral disease (occasionally), and central nervous system (CNS) lesions (more commonly). The advantages of maintaining HSC transplant recipients on daily TMP-SMX for 1 year after transplantation include some protection against *Listeria monocytogenes* and nocardiosis as well as late infections with

Streptococcus pneumoniae and *Haemophilus influenzae* that stem from the inability of the immature immune system to respond to polysaccharide antigens.

With increasing international travel, parasitic diseases typically restricted to particular environmental niches may pose a risk of reactivation in certain patients after HSC transplantation. Thus, in recipients with an appropriate history who were not screened and/or treated before transplantation or in patients with recent exposures, evaluation for infection with *Strongyloides, Leishmania*, schistosomes, trypanosomes, or various parasitic causes of diarrheal illness (*Giardia, Entamoeba, Cryptosporidium*, microsporidia) may be warranted.

VIRAL INFECTIONS

HSC transplant recipients are susceptible to infection with a variety of viruses, including primary and reactivation syndromes caused by most human herpesviruses **(Table 138-3)** and acute infections caused by viruses that circulate in the community.

TABLE 138-3

Herpesvirus Syndromes in Transplant Recipients

Virus	Reactivation Disease
Herpes simplex virus type 1	Oral lesions Esophageal lesions Pneumonia (primarily HSC transplant recipients) Hepatitis (rare)
Herpes simplex virus type 2	Anogenital lesions Hepatitis (rare)
Varicella-zoster virus	Zoster (can disseminate)
Cytomegalovirus	Associated with graft rejection Fever and malaise Bone marrow failure Pneumonitis Gastrointestinal disease
Epstein-Barr virus	B cell lymphoproliferative disease/lymphoma Oral hairy leukoplakia (rare)
Human herpesvirus type 6	Fever Delayed monocyte/platelet engraftment Encephalitis (rare)
Human herpesvirus type 7	Undefined
Kaposi's sarcoma–associated virus (human herpesvirus type 8)	Kaposi's sarcoma Primary effusion lymphoma (rare) Multicentric Castleman's disease (rare) Marrow aplasia (rare)

Abbreviation: HSC, hematopoietic stem cell.

Herpes Simplex Virus

Within the first 2 weeks after transplantation, most patients who are seropositive for HSV-1 excrete the virus from the oropharynx. The ability to isolate HSV declines with time. Administration of prophylactic acyclovir (or valacyclovir) to seropositive HSC transplant recipients has been shown to reduce mucositis and prevent HSV pneumonia (a rare condition reported almost exclusively in allogeneic HSC transplant recipients). Both esophagitis (usually due to HSV-1)

and anogenital disease (commonly caused by HSV-2) may be prevented with acyclovir prophylaxis. **For further discussion, see Chap. 187.**

Varicella-Zoster Virus

Reactivation of VZV manifests as herpes zoster and may occur within the first month but more commonly occurs several months after transplantation. Reactivation rates are ~40% for allogeneic HSC transplant recipients and 25% for autologous recipients. Localized zoster can spread rapidly in an immunosuppressed patient. Fortunately, disseminated disease can usually be controlled with high doses of acyclovir. Because of frequent VZV dissemination among patients with skin lesions, acyclovir is given prophylactically in many centers to prevent severe disease. Low doses of acyclovir appear to be effective in preventing reactivation of VZV. However, acyclovir can also suppress the development of VZV-specific immunity. Thus, its administration for only 6 months after transplantation does not prevent zoster from occurring when treatment is stopped. Administration of low doses of acyclovir for an entire year after transplantation is effective and may eliminate most cases of posttransplantation zoster, even among cord-blood recipients. **For further discussion, see Chap. 188.**

Cytomegalovirus

The onset of CMV disease (interstitial pneumonia, bone marrow suppression, graft failure, hepatitis/colitis) usually begins 30–90 days after HSC transplantation, when the granulocyte count is adequate but immunologic reconstitution has not occurred. CMV disease rarely develops earlier than 14 days after transplantation and may become evident as late as 4 months or more after the procedure. It is of greatest concern in the second month after transplantation, particularly in allogeneic HSC transplant recipients. In cases in which the donor marrow is depleted of T cells (to prevent GVHD or eliminate a T cell tumor) and in cord-blood recipients, the disease may manifest earlier. The use of alemtuzumab to prevent GVHD in nonmyeloablative transplantation has been associated with an increase in CMV disease. Patients who receive ganciclovir for prophylaxis, preemptive treatment, or treatment (see below) may develop recurrent CMV infection even later than 4 months after transplantation, as treatment appears to delay the development of an effective immune response to CMV infection. Although CMV disease may present as isolated fever, granulocytopenia, thrombocytopenia, or gastrointestinal disease, the foremost cause of death from CMV infection in the setting of HSC transplantation is pneumonia.

With the standard use of CMV-negative or filtered blood products, CMV infection should be a major risk in allogeneic transplantation only when the recipient is CMV-seropositive and the donor is CMV-seronegative. This situation is the reverse of that in solid organ transplant recipients. CMV reactivates from latent reservoirs present in the recipient at a time when donor T cells (especially cord-blood T cells) are too immature to control CMV replication. If the T cells from the donor have never encountered CMV and the recipient carries the virus, the patient is at maximal risk of severe disease. Reactivation disease or superinfection with another strain from the donor also can occur in CMV-positive recipients, but clinical manifestations are typically less severe, presumably because of CMV-specific memory in transplanted donor T cells. Most patients infected with CMV who undergo HSC transplantation excrete virus, with or without clinical findings. Serious CMV disease is much more common among allogeneic than autologous recipients and is often associated with GVHD. In addition to pneumonia and marrow suppression (and, less often, graft failure), manifestations of CMV disease in HSC transplant recipients include fever with or without arthralgias, myalgias, hepatitis, and esophagitis. CMV ulcerations occur in both the lower and the upper gastrointestinal tract, and it may be difficult to distinguish diarrhea due to GVHD from that due to CMV infection. The finding of CMV in the liver of a patient with GVHD does not necessarily mean that CMV is responsible for hepatic enzyme abnormalities. It is interesting that

ocular and neurologic manifestations of CMV infections are common in patients with AIDS but uncommon in patients who develop CMV disease after transplantation.

Management of CMV disease in HSC transplant recipients includes strategies directed at prophylaxis, preemptive therapy (suppression of silent replication), and treatment of disease. Prophylaxis results in a lower incidence of disease at the cost of treating many patients who otherwise would not require therapy. Because of the high fatality rate associated with CMV pneumonia in these patients and the difficulty of early diagnosis of CMV infection, prophylactic IV ganciclovir or oral valganciclovir has been used in some centers and has been shown to prevent CMV disease during the period of maximal vulnerability (from engraftment to day 120 after transplantation). Ganciclovir also prevents HSV reactivation and reduces the risk of VZV reactivation; thus acyclovir prophylaxis should be discontinued when ganciclovir is administered. The foremost problem with the administration of ganciclovir relates to adverse effects, which include dose-related bone marrow suppression (thrombocytopenia, leukopenia, anemia, and pancytopenia). Because the frequency of CMV pneumonia is lower among autologous HSC transplant recipients (2–7%) than among allogeneic HSC transplant recipients (10–40%), prophylaxis in the former group will not become the rule until a less toxic oral antiviral agent becomes available. Several such agents are under study.

Preemptive treatment of CMV—that is, initiation of therapy with drugs only after CMV is detected in blood by a nucleic acid amplification test (NAAT)—is used at most centers. To limit variability between tests, the World Health Organization (WHO) has developed an international reference standard for measurement of CMV load by NAAT-based assays. Because of toxic drug side effects (e.g., neutropenia and bone marrow suppression), the preemptive approach has supplanted prophylactic therapy; it has also replaced treatment of all seropositive (recipient and/or donor) HSC transplants with an antiviral agent (typically ganciclovir). A positive test (or increasing viral load) prompts the initiation of preemptive therapy with ganciclovir. Preemptive approaches that target patients who have quantitative NAAT evidence of CMV infection can still lead to unnecessary treatment of many individuals with drugs that have adverse effects on the basis of a laboratory test that is not highly predictive of disease; however, invasive disease, particularly in the form of pulmonary infection, is difficult to treat and is associated with high mortality rates. When prophylaxis or preemptive therapy is stopped, late manifestations of CMV replication may occur, although by then the HSC transplant patient is often equipped with improved graft function and is better able to combat disease. Cord-blood transplant recipients are especially vulnerable to disease caused by members of the human herpesvirus family, including CMV. Implementation of the WHO standard for CMV load measurement will facilitate large-scale comparative studies and thus the establishment of optimal guidelines for distinct patient subsets.

CMV pneumonia in HSC transplant recipients (unlike that in other clinical settings) is often treated with both IV immunoglobulin (IVIg) and ganciclovir. In patients who cannot tolerate ganciclovir, foscarnet is a useful alternative, although it may produce nephrotoxicity and electrolyte imbalance. When neither ganciclovir nor foscarnet is clinically tolerated, cidofovir can be used; however, its efficacy is less well established, and its side effects include nephrotoxicity. A lipid-conjugate form of cidofovir, brincidofovir, appears to have more activity and less toxicity than cidofovir and is in clinical trials. Transfusion of CMV-specific T cells from the donor has decreased viral load in a small series of patients; this result suggests that immunotherapy (e.g., banked T cells) may play a role in the management of this disease in the future. **For further discussion, see Chap. 190.**

Human Herpesviruses 6 and 7

Human herpesvirus type 6 (HHV-6), the cause of roseola in children, is a ubiquitous herpesvirus that is reactivated (as determined by quantitative plasma PCR) in ~50% of HSC transplant recipients 2–4 weeks after transplantation. Reactivation is more common among patients requiring glucocorticoids for GVHD and among those receiving second

transplants. Reactivation of HHV-6, primarily type B, may be associated with delayed monocyte and platelet engraftment. Limbic encephalitis developing after transplantation has been associated with HHV-6 in cerebrospinal fluid (CSF). The causality of the association is not well defined; in several cases, plasma viremia was detected long before the onset of encephalitis. Nevertheless, most patients with encephalitis had very high viral loads in plasma at the time of CNS illness, and viral antigen has been detected in hippocampal astrocytes. HHV-6 DNA is sometimes found in lung samples after transplantation. However, its role in pneumonitis is unclear, as co-pathogens are frequently present. While HHV-6 is susceptible to foscarnet or cidofovir (and possibly to ganciclovir) in vitro, the efficacy of antiviral treatment has not been well studied. Little is known about the related herpesvirus HHV-7 or its role in posttransplantation infection. **For further discussion, see Chap. 190.**

Epstein-Barr Virus

Primary EBV infection can be fatal to HSC transplant recipients; EBV reactivation can cause EBV B cell lymphoproliferative disease (EBV-LPD), which may also be fatal to patients taking immunosuppressive drugs. Latent EBV infection of B cells leads to several interesting phenomena in HSC transplant recipients. The marrow ablation that occurs as part of the HSC transplantation procedure may sometimes eliminate latent EBV from the host. Infection can then be reacquired immediately after transplantation by transfer of infected donor B cells. Rarely, transplantation from a seronegative donor may result in a cure. The recipient is then at risk for a second primary infection.

EBV-LPD can develop in the recipient's B cells (if any survive marrow ablation) but is more likely to be a consequence of outgrowth of infected donor cells. Both lytic replication and latent replication of EBV are more likely during immunosuppression (e.g., they are associated with GVHD and the use of antibodies to T cells). Although less likely in autologous transplantation, reactivation can occur in T cell–depleted autologous recipients (e.g., patients being given antibodies to T cells for the treatment of T cell lymphoma with marrow depletion). EBV-LPD, which can become apparent as early as 1–3 months after engraftment, can cause high fevers and cervical adenopathy resembling the symptoms of infectious mononucleosis but more commonly presents as an extranodal mass. The incidence of EBV-LPD among allogeneic HSC transplant recipients is 0.6–1%, which contrasts with figures of ~5% for renal transplant recipients and up to 20% for cardiac transplant patients. In all cases, EBV-LPD is more likely to occur with high-dose, prolonged immunosuppression, especially that caused by the use of antibodies to T cells, glucocorticoids, and calcineurin inhibitors (e.g., cyclosporine, tacrolimus). Cord-blood recipients constitute another high-risk group because of delayed T cell function. Ganciclovir, administered to preempt CMV disease, may reduce EBV lytic replication and thereby diminish the pool of B cells that can become newly infected and give rise to LPD. Increasing evidence indicates that replacement of calcineurin inhibitors with mTOR inhibitors (e.g., rapamycin) exerts an antiproliferative effect on EBV-infected B cells that decreases the likelihood of development of LPD or unrelated proliferative disorders associated with transplant-related immunosuppression.

PCR can be used to monitor EBV production after HSC transplantation. High or increasing viral loads predict an enhanced likelihood of EBV-LPD development and should prompt rapid reduction of immunosuppression and a search for nodal or extranodal disease. If reduction of immunosuppression does not have the desired effect, administration of a monoclonal antibody to CD20 (e.g., rituximab) for the treatment of B-cell lymphomas that express this surface protein has elicited dramatic responses and currently constitutes first-line therapy for CD20-positive EBV-LPD. However, long-term suppression of new antibody responses accompanies therapy, and recurrences are common. Additional B cell–directed antibodies, including anti-CD22, are under study. The role of antiviral drugs is uncertain because no available agents have been documented to have activity against the different forms of latent EBV infection. Diminishing lytic replication and virion production in these patients would theoretically produce a statistical decrease in the frequency of

latent disease by decreasing the number of virions available to cause additional infection. In case reports and animal studies, ganciclovir and/or high-dose zidovudine, together with other agents, has been used to eradicate EBV-LPD and CNS lymphomas, another EBV-associated complication of transplantation. Both interferon and retinoic acid have been employed in the treatment of EBV-LPD, as has IVIg, but no large-scale prospective studies have assessed the efficacy of any of these agents. Several additional drugs are undergoing preclinical evaluation. Standard chemotherapeutic regimens are used if disease persists after reduction of immunosuppressive agents and administration of antibodies. EBV-specific T cells generated from the donor have been used experimentally to prevent and treat EBV-LPD in allogeneic recipients, and efforts are under way to increase the activity and specificity of ex vivo–generated T cells. **For further discussion, see Chap. 189.**

Human Herpesvirus 8 (KSHV)

The EBV-related gammaherpesvirus KSHV, which is causally associated with Kaposi's sarcoma, primary effusion lymphoma, and multicentric Castleman's disease, has rarely resulted in disease in HSC transplant recipients, although some cases of virus-associated marrow aplasia have been reported in the peritransplantation period. The relatively low seroprevalence of KSHV in the population and the limited duration of profound T cell suppression after HSC transplantation provide a plausible explanation for the currently low incidence of KSHV disease compared with that in recipients of solid organ transplants and patients with HIV infection. **For further discussion, see Chap. 190.**

Other (Non-Herpes) Viruses

The diagnosis of pneumonia in HSC transplant recipients poses special problems. Because patients have undergone treatment with multiple chemotherapeutic agents and sometimes irradiation, their differential diagnosis should include—in addition to bacterial and fungal pneumonia—CMV pneumonitis, pneumonia of other viral etiologies, parasitic pneumonia, diffuse alveolar hemorrhage, and chemical- or radiation-associated pneumonitis. Since fungi and viruses (e.g., influenza A and B viruses, respiratory syncytial virus [RSV], parainfluenza virus types 1–4, adenovirus, enterovirus, bocavirus, human metapneumovirus, coronavirus, and rhinovirus [increasingly detected by multiplex PCR]) also can cause pneumonia in this setting, it is important to obtain a specific diagnosis. Diagnostic modalities include Gram's stain, microbiologic culture, antigen testing, and—increasingly—multipathogen PCR and mass spectrometry assays.

GLOBAL CONSIDERATIONS

 *M. tuberculosis* has been an uncommon cause of pneumonia among HSC transplant recipients in Western countries (accounting for <0.1–0.2% of cases) but is common in Hong Kong (5.5%) and in countries where the prevalence of tuberculosis is high. The recipient's exposure history is clearly critical in an assessment of posttransplantation infections.

Both RSV and parainfluenza viruses, particularly type 3, can cause severe or even fatal pneumonia in HSC transplant recipients. Infections with both of these agents sometimes occur as disastrous nosocomial epidemics. Therapy with palivizumab or ribavirin for RSV infection remains controversial. New agents, some host-directed, are under study. Influenza also occurs in HSC transplant recipients and generally mirrors the presence of infection in the community. Progression to pneumonia is more common when infection occurs early after transplantation and when the recipient is lymphopenic. The neuraminidase inhibitors oseltamivir (oral) and zanamivir (aerosolized) are active against both influenza A virus and influenza B virus and are a reasonable treatment option. Parenteral forms of neuraminidase inhibitors such as peramivir (intravenous) are available in some countries, and several new oral agents are still being

assessed in trials. An important preventive measure is immunization of household members, hospital staff members, and other frequent contacts. Adenoviruses can be isolated from HSC transplant recipients at rates varying from 5 to ≥18%. Like CMV infection, adenovirus infection usually occurs in the first to third month after transplantation and is often asymptomatic, although pneumonia, hemorrhagic cystitis/nephritis, severe gastroenteritis with hemorrhage, and fatal disseminated infection have been reported and may be strain-specific. Banked virus-specific T cell therapy is under study for adenovirus infection (as well as for CMV and EBV infections).

Although diverse respiratory viruses can sometimes cause severe pneumonia and respiratory failure in HSC transplant recipients, mild or even asymptomatic infection may be more common. For example, rhinoviruses and coronaviruses are frequent co-pathogens in HSC transplant recipients; however, whether they independently contribute to significant pulmonary infection is not known. At present, the overall contribution of these viral respiratory pathogens to the burden of lower respiratory tract disease in HSC transplant recipients requires further study. Infections with parvovirus B19 (presenting as anemia or occasionally as pancytopenia) and disseminated enteroviruses (sometimes fatal) can occur. Parvovirus B19 infection can be treated with IVIg **(Chap. 192)**.

Rotaviruses, a cause of gastroenteritis in HSC transplant recipients, cause disease more frequently in children. Norovirus is a common cause of vomiting and diarrhea, and symptoms can be prolonged in HSC recipients. The BK virus (a polyomavirus) is found at high titers in the urine of patients who are profoundly immunosuppressed. BK viruria may be associated with hemorrhagic cystitis in these patients. In contrast to its incidence among patients with impaired T cell function due to AIDS (4–5%), progressive multifocal leukoencephalopathy caused by the related JC virus is relatively rare among HSC transplant recipients **(Chap. 133)**. When transmitted by mosquitoes or by blood transfusion, West Nile virus (WNV) can cause encephalitis and death after HSC transplantation.

INFECTIONS IN SOLID ORGAN TRANSPLANT RECIPIENTS

Rates of morbidity and mortality among recipients of solid organ transplants (SOTs) are reduced by the use of effective antibiotics. The organisms that cause acute infections in recipients of SOTs are different from those that infect HSC transplant recipients because SOT recipients do not go through a period of neutropenia. As the transplantation procedure involves major surgery, however, SOT recipients are subject to infections at anastomotic sites and to wound infections. Compared with HSC transplant recipients, SOT patients are immunosuppressed for longer periods (often permanently). Thus they are susceptible to many of the same organisms as patients with chronically impaired T cell immunity **(Chap. 70**, especially **Table 70-1)**. Moreover, the persistent HLA mismatch between recipient immune cells (e.g., effector T cells) and the donor organ (allograft) places the organ at permanently increased risk of infection.

During the early period (<1 month after transplantation; **Table 138-4**), infections are most commonly caused by extracellular bacteria (staphylococci, streptococci, enterococci, and *E. coli* and other gram-negative organisms, including nosocomial organisms with broad antibiotic resistance), which often originate in surgical-wound or anastomotic sites. The type of transplant largely determines the spectrum of infection. In subsequent weeks, the consequences of the administration of agents that suppress cell-mediated immunity become apparent, and acquisition —or, more commonly, reactivation—of viruses, mycobacteria, endemic fungi, and parasites (from the recipient or from the transplanted organ) can occur. CMV infection is often a problem, particularly in the first 6 months after transplantation, and may present as severe systemic disease or as infection of the transplanted organ. HHV-6 reactivation (assessed by plasma PCR) occurs within the first 2–4 weeks after transplantation and may be associated with fever, leukopenia, and very rare cases of encephalitis. Data suggest that replication of HHV-6 and HHV-7 may exacerbate CMV-induced disease. CMV is associated not only with generalized immunosuppression but also with organ-

specific, rejection-related syndromes: glomerulopathy in kidney transplant recipients, bronchiolitis obliterans in lung transplant recipients, vasculopathy in heart transplant recipients, and the vanishing bile duct syndrome in liver transplant recipients. A complex interplay between increased CMV replication and enhanced graft rejection is well established: elevated immunosuppression leads to increased CMV replication, which is associated with graft rejection. For this reason, considerable attention has been focused on the diagnosis, prophylaxis, and treatment of CMV infection in SOT recipients. Early transmission of WNV to transplant recipients from a donated organ or transfused blood has been reported; however, the risk of WNV acquisition has been reduced by implementation of screening procedures. In rare instances, rabies virus and lymphocytic choriomeningitis virus also have been acutely transmitted in this setting; although accompanied by distinct clinical syndromes, both viral infections have resulted in fatal encephalitis. As screening for unusual viruses is not routine, only vigilant assessment of the prospective donor is likely to prevent the use of an infected organ.

TABLE 138-4

Common Infections after Solid Organ Transplantation, by Site of Infection

	Period after Transplantation		
Infected Site	**Early (<1 Month)**	**Middle (1–4 Months)**	**Late (>6 Months)**
Donor organ	Bacterial and fungal infections of the graft, anastomotic site, and surgical wound	CMV infection	EBV infection (may present in allograft organ)
Systemic	Bacteremia and candidemia (often resulting from central venous catheter colonization)	CMV infection (fever, bone marrow suppression)	CMV infection, especially in patients given early posttransplantation prophylaxis; EBV proliferative syndromes (may occur in donor organs)
Lung	Bacterial aspiration pneumonia with prevalent nosocomial organisms associated with intubation and sedation (highest risk in lung transplantation)	*Pneumocystis* infection; CMV pneumonia (highest risk in lung transplantation); *Aspergillus* infection (highest risk in lung transplantation)	*Pneumocystis* infection; granulomatous lung diseases (nocardial and reactivated fungal and mycobacterial diseases)
Kidney	Bacterial and fungal (*Candida*) infections (cystitis, pyelonephritis) associated with urinary tract catheters (highest risk in kidney transplantation)	Kidney transplantation: BK virus infection (associated with nephropathy); JC virus infection	Kidney transplantation: bacterial infections (late urinary tract infections, usually not associated with bacteremia); BK virus infection (nephropathy, graft failure, generalized vasculopathy)
Liver and biliary tract	Cholangitis	CMV hepatitis	CMV hepatitis
Heart		*Toxoplasma gondii* infection (highest risk in heart transplantation); endocarditis (*Aspergillus* and gram-negative organisms more common than in general population)	*T. gondii* (highest risk in heart transplantation)

	Period after Transplantation		
Infected Site	**Early (<1 Month)**	**Middle (1–4 Months)**	**Late (>6 Months)**
Gastrointestinal tract	Peritonitis, especially after liver transplantation	Colitis secondary to *Clostridium difficile* infection (risk can persist)	Colitis secondary to *C. difficile* infection (risk can persist)
Central nervous system		*Listeria* infection (meningitis); *T. gondii* infection; CMV infection	Listerial meningitis; cryptococcal meningitis; nocardial abscess; JC virus-associated PML

Abbreviations: CMV, cytomegalovirus; EBV, Epstein-Barr virus; PML, progressive multifocal leukoencephalopathy.

Beyond 6 months after transplantation, infections characteristic of patients with defects in cell-mediated immunity—e.g., infections with *Listeria, Nocardia, Rhodococcus*, mycobacteria, various fungi, and other intracellular pathogens—may be a problem. International patients and global travelers may experience reactivation of dormant infections with trypanosomes, *Leishmania*, *Plasmodium*, *Strongyloides*, and other parasites. Reactivation of latent *M. tuberculosis* infection, while rare in Western nations, is far more common among persons from developing countries. The recipient is typically the source, although reactivation and spread from the donor organ can occur. While pulmonary disease remains most common, atypical sites can be involved and mortality rates can be high (up to 30%). Vigilance, prophylaxis/preemptive therapy (when indicated), and rapid diagnosis and treatment of infections can be lifesaving in SOT recipients, who, unlike most HSC transplant recipients, continue to be immunosuppressed.

SOT recipients are susceptible to EBV-LPD from as early as 2 months to many years after transplantation. The prevalence of this complication is increased by potent and prolonged use of T cell–suppressive drugs. Decreasing the degree of immunosuppression may in some cases reverse the condition. Among SOT patients, those with heart and lung transplants—who receive the most intensive immunosuppressive regimens—are most likely to develop EBV-LPD, particularly in the lungs. Although the disease usually originates in recipient B cells, several cases of donor origin, particularly in the transplanted organ, have been noted. High organ-specific content of B lymphoid tissues (e.g., bronchus-associated lymphoid tissue in the lung), anatomic factors (e.g., lack of access of host T cells to the transplanted organ because of disturbed lymphatics), and differences in major histocompatibility loci between the host T cells and the organ (e.g., lack of cell migration or lack of effective T cell/macrophage/dendritic cell cooperation) may result in defective elimination of EBV-infected B cells. SOT recipients are also highly susceptible to the development of Kaposi's sarcoma and, less frequently, to the B cell–proliferative disorders associated with KSHV, such as primary effusion lymphoma and multicentric Castleman's disease. Kaposi's sarcoma is 550–1000 times more common among SOT recipients than in the general population, can develop very rapidly after transplantation, and can also occur in the allograft. However, because the seroprevalence of KSHV is very low in Western countries, Kaposi's sarcoma is not common. Recipients (or donors) from Iceland, the Middle East, Mediterranean countries, and Africa are at highest risk of disease. Data suggest that a switch of immunosuppressive agents—from calcineurin inhibitors (cyclosporine, tacrolimus) to mTOR pathway–active agents (sirolimus, everolimus)—after adequate wound healing may significantly reduce the likelihood of development of Kaposi's sarcoma and perhaps of EBV-LPD and certain other posttransplantation malignancies.

KIDNEY TRANSPLANTATION

See Table 138-4.

Early Infections

Bacteria often cause infections that develop in the period immediately after kidney transplantation. There is a role for perioperative antibiotic prophylaxis, and many centers give cephalosporins to decrease the risk of postoperative complications. Urinary tract infections developing soon after transplantation are usually related to anatomic alterations resulting from surgery. Such early infections may require prolonged treatment (e.g., 6 weeks of antibiotic administration for pyelonephritis). Urinary tract infections that occur >6 months after transplantation may be treated for shorter periods because they do not seem to be associated with the high rate of pyelonephritis or relapse seen with infections that occur during the first 3 months.

Prophylaxis with TMP-SMX for the first 4–6 months after transplantation decreases the incidence of early and middle-period infections (see below, Table 138-4, and **Table 138-5**).

TABLE 138-5

Prophylactic Regimens Commonly Used to Decrease Risk of Infection in Transplant Recipients[a]

Risk Factor	Organism	Prophylactic Drug	Examination(s)[b]
Travel to or residence in area with known risk of endemic fungal infection	*Histoplasma*, *Blastomyces*, *Coccidioides*	Triazoles considered in context of clinical and laboratory assessment	Chest radiography, antigen testing, serology
Latent herpesviruses	HSV, VZV, CMV, EBV	Acyclovir after HSC transplantation to prevent HSV and VZV infection or reactivation; ganciclovir to prevent CMV infection, with possible effect on EBV/KSHV (HHV-8)/HHV-6 infections in some settings	Serologic tests for HSV, VZV, CMV, HHV-6, EBV, KSHV (HHV-8); PCR
Latent fungi and parasites	*Pneumocystis jirovecii*, *Toxoplasma gondii*	Trimethoprim-sulfamethoxazole (or alternatives)	Serologic test for *Toxoplasma*
History of exposure to active or latent tuberculosis	*Mycobacterium tuberculosis*	Isoniazid in patients with recent seroconversion or positive chest imaging and/or no previous treatment	Chest imaging; TST and/or cell-based assay

[a]For information on latent infection with hepatitis B or C virus, see **Chap. 334**. [b]Serologic examination, tuberculin skin test, and interferon assays may be less reliable after transplantation.

Abbreviations: CMV, cytomegalovirus; EBV, Epstein-Barr virus; HHV, human herpesvirus; HSC, hematopoietic stem cell; HSV, herpes simplex virus; KSHV, Kaposi's sarcoma–associated herpesvirus; PCR, polymerase chain reaction; TST, tuberculin skin test; VZV, varicella-zoster virus.

Middle-Period Infections

Because of continuing immunosuppression, kidney transplant recipients are predisposed to lung infections characteristic of those in patients with T-cell deficiency (i.e., infections with intracellular bacteria, mycobacteria, nocardiae, fungi, viruses, and parasites). A high mortality rate associated with *Legionella pneumophila* infection **(Chap. 154)** led to the closing of renal transplant units in hospitals with endemic legionellosis.

About 50% of all renal transplant recipients presenting with fever 1–4 months after transplantation have evidence of CMV disease; CMV itself accounts for the fever in more than two-thirds of cases and thus is the predominant pathogen during this period. CMV infection **(Chap. 190)** may also present as arthralgias, myalgias, or organ-specific symptoms. During this period, this infection may represent primary disease (in the case of a seronegative recipient of a kidney from

a seropositive donor) or may represent reactivation disease or superinfection. Patients may have atypical lymphocytosis. Unlike immunocompetent patients, however, they rarely have lymphadenopathy or splenomegaly. Therefore, clinical suspicion and laboratory confirmation are necessary for diagnosis. The clinical syndrome may be accompanied by bone marrow suppression (particularly leukopenia). CMV also causes glomerulopathy and is associated with an increased incidence of other opportunistic infections. Because of the frequency and severity of disease, a considerable effort has been made to prevent and treat CMV infection in renal transplant recipients. An immune globulin preparation enriched with antibodies to CMV was used by many centers in the past in an effort to protect the group at highest risk for severe infection (seronegative recipients of seropositive kidneys). However, with the development of effective oral antiviral agents, CMV immune globulin is no longer used. Ganciclovir (or valganciclovir) is beneficial for prophylaxis (when indicated) and for the treatment of serious CMV disease. The availability of valganciclovir has allowed most centers to move to oral prophylaxis for transplant recipients. Infection with the other herpesviruses may become evident within 6 months after transplantation or later. Early after transplantation, HSV may cause either oral or anogenital lesions that are usually responsive to acyclovir. Large ulcerating lesions in the anogenital area may lead to bladder and rectal dysfunction and may predispose the patient to bacterial infection. VZV may cause fatal disseminated infection in nonimmune kidney transplant recipients, but in immune patients reactivation zoster usually does not disseminate outside the dermatome; thus disseminated VZV infection is a less fearsome complication in kidney transplantation than in HSC transplantation. HHV-6 reactivation may take place and (although usually asymptomatic) may be associated with fever, rash, marrow suppression, or rare instances of renal impairment, hepatitis, colitis, or encephalitis.

EBV disease is more serious; it may present as an extranodal proliferation of B cells that invade the CNS, nasopharynx, liver, small bowel, heart, and other organs, including the transplanted kidney. The disease is diagnosed by the finding of a mass of proliferating EBV-positive B cells. The incidence of EBV-LPD is elevated among patients who acquire EBV infection from the donor and among patients given high doses of cyclosporine, tacrolimus, glucocorticoids, and anti–T cell antibodies. Disease may regress once immunocompetence is restored. KSHV infection can be transmitted with the donor kidney and result in the development of Kaposi's sarcoma, although it more often represents reactivation of latent infection in the recipient. Kaposi's sarcoma often appears within 1 year after transplantation, although the time of onset ranges widely (1 month to ~20 years). Avoidance of immunosuppressive agents that inhibit calcineurin has been associated with less Kaposi's sarcoma, less EBV disease, and even less CMV replication. The use of rapamycin (sirolimus) has independently led to regression of Kaposi's sarcoma.

The polyomaviruses BK virus and JC virus (polyomavirus hominis types 1 and 2) have been cultured from the urine of kidney transplant recipients (as they have from that of HSC transplant recipients) in the setting of profound immunosuppression. High levels of BK virus replication detected by PCR in urine and blood are predictive of pathology, especially in the setting of renal transplantation. JC virus may rarely cause similar disease in kidney transplantation. Urinary excretion of BK virus and BK viremia are associated with the development of ureteral strictures, polyomavirus-associated nephropathy (1–10% of renal transplant recipients), and (less commonly) generalized vasculopathy. Timely detection and early reduction of immunosuppression are critical and can reduce rates of graft loss related to polyomavirus-associated nephropathy from 90 to 10–30%. Therapeutic responses to IVIg, quinolones, leflunomide, and cidofovir have been reported, but the efficacy of these agents has not been substantiated through adequate clinical study. Most centers approach the problem by reducing immunosuppression in an effort to enhance host immunity and decrease viral titers. JC virus is associated with rare cases of progressive multifocal leukoencephalopathy. Adenoviruses may persist and cause hemorrhagic nephritis/cystitis with continued immunosuppression in these patients, but disseminated disease, like that seen in HSC transplant recipients, is much less common.

Kidney transplant recipients are also subject to infections with other intracellular organisms. These patients may develop pulmonary infections with *Mycobacterium, Aspergillus*, and *Mucor* species as well as infections with other pathogens in which the T cell/macrophage axis plays an important role. *L. monocytogenes* is a common cause of bacteremia ≥1 month after renal transplantation and should be seriously considered in renal transplant recipients presenting with fever and headache. Kidney transplant recipients may develop *Salmonella* bacteremia, which can lead to endovascular infections and require prolonged therapy. Pulmonary infections with *Pneumocystis* are common unless the patient is maintained on TMP-SMX prophylaxis. Acute interstitial nephritis caused by TMP-SMX is rare. However, because transient increases in creatinine (artifactual) and hyperkalemia (manageable) can occur, early discontinuation of prophylaxis, especially after kidney transplantation, is recommended by some groups. Although additional monitoring is indicated, the benefits of TMP-SMX in kidney transplant recipients may outweigh the risks; otherwise, second-line prophylactic agents should be used. *Nocardia* infection **(Chap. 169)** may present in the skin, bones, and lungs or in the CNS, where it usually takes the form of single or multiple brain abscesses. Nocardiosis generally occurs ≥1 month after transplantation and may follow immunosuppressive treatment for an episode of rejection. Pulmonary manifestations most commonly consist of localized disease with or without cavities, but the disease may be disseminated. The diagnosis is made by culture of the organism from sputum or from the involved nodule. As it is for *P. jirovecii* infection, prophylaxis with TMP-SMX is often efficacious in the prevention of nocardiosis.

Toxoplasmosis can occur in seropositive patients but is less common than in other transplantation settings, usually developing in the first few months after kidney transplantation. Again, TMP-SMX is helpful in prevention. In endemic areas, histoplasmosis, coccidioidomycosis, and blastomycosis may cause pulmonary infiltrates or disseminated disease.

Late Infections

Late infections (>6 months after kidney transplantation) may involve the CNS and include CMV retinitis as well as other CNS manifestations of CMV disease. Patients (particularly those whose immunosuppression has been increased) are at risk for subacute meningitis due to *Cryptococcus neoformans*. Cryptococcal disease may present in an insidious manner (sometimes as a skin infection before the development of clear CNS findings). *Listeria* meningitis may have an acute presentation and requires prompt therapy to avoid a fatal outcome. TMP-SMX prophylaxis may reduce the frequency of *Listeria* infections.

Patients who continue to take glucocorticoids are predisposed to ongoing infection. “Transplant elbow,” a recurrent bacterial infection in and around the elbow that is thought to result from a combination of poor tensile strength of the skin of steroid-treated patients and steroid-induced proximal myopathy, requires patients to push themselves up with their elbows to get out of chairs. Bouts of cellulitis (usually caused by *S. aureus*) recur until patients are provided with elbow protection.

Kidney transplant recipients are susceptible to invasive fungal infections, including those due to *Aspergillus* and *Rhizopus*, which may present as superficial lesions before dissemination. Mycobacterial infection (particularly that with *Mycobacterium marinum*) can be diagnosed by skin examination. Infection with *Prototheca wickerhamii* (an achlorophyllic alga) has been diagnosed by skin biopsy. Warts caused by human papillomaviruses (HPVs) are a late consequence of persistent immunosuppression; imiquimod or other forms of local therapy are usually satisfactory. Merkel cell carcinoma, a rare and aggressive neuroendocrine skin tumor whose frequency is increased fivefold in elderly SOT (especially kidney) recipients, is causally linked to a novel polyomavirus, Merkel cell polyomavirus.

Notably, although BK virus replication and virus-associated disease can be detected far earlier, polyomavirus-associated nephropathy is clinically diagnosed at a median of ~300 days and thus qualifies as a late-onset disease. With the establishment of better screening procedures (e.g., urine cytology, urine nucleic acid load, plasma PCR), disease onset is being detected earlier (see "Middle-Period Infections," above), and preemptive strategies (decrease or modification of immunosuppression) are being instituted more promptly, as the efficacy of antiviral therapy is not well established.

HEART TRANSPLANTATION

Early Infections

Sternal wound infection and mediastinitis are early complications of heart transplantation. An indolent course is common, with fever or a mildly elevated white blood cell count preceding the development of site tenderness or drainage. Clinical suspicion based on evidence of sternal instability and failure to heal may lead to the diagnosis. Common microbial residents of the skin (e.g., *S. aureus*, including methicillin-resistant strains, and *Staphylococcus epidermidis*) as well as gram-negative organisms (e.g., *Pseudomonas aeruginosa*) and fungi (e.g., *Candida*) are often involved. In rare cases, mediastinitis in heart transplant recipients can also be due to *Mycoplasma hominis* **(Chap. 183)**; since this organism requires an anaerobic environment for growth and may be difficult to see on conventional medium, the laboratory should be alerted that its involvement is suspected. *M. hominis* mediastinitis has been cured with a combination of surgical debridement (sometimes requiring muscle-flap placement) and the administration of clindamycin and tetracycline. Organisms associated with mediastinitis may sometimes be cultured from pericardial fluid.

Middle-Period Infections

T. gondii **(Chap. 223)** residing in the heart of a seropositive donor may be transmitted to a seronegative recipient. Thus serologic screening for *T. gondii* infection is important before and in the months after cardiac transplantation. Rarely, active disease can be introduced at the time of transplantation. The overall incidence of toxoplasmosis is so high in the setting of heart transplantation that some prophylaxis is always warranted. Although alternatives are available, the most frequently used agent is TMP-SMX, which prevents infection with *Pneumocystis* as well as with *Nocardia* and several other bacterial pathogens. CMV also has been transmitted by heart transplantation. *Toxoplasma*, *Nocardia*, and *Aspergillus* can cause CNS infections. *L. monocytogenes* meningitis should be considered in heart transplant recipients with fever and headache.

CMV infection is associated with poor outcomes after heart transplantation. The virus is usually detected 1–2 months after transplantation, causes early signs and laboratory abnormalities (usually fever and atypical lymphocytosis or leukopenia and thrombocytopenia) at 2–3 months, and can produce severe disease (e.g., pneumonia) at 3–4 months. An interesting observation is that seropositive recipients usually develop viremia faster than patients whose primary CMV infection is a consequence of transplantation. Between 40 and 70% of patients develop symptomatic CMV disease in the form of (1) CMV pneumonia, the form most likely to be fatal; (2) CMV esophagitis and gastritis, sometimes accompanied by abdominal pain with or without ulcerations and bleeding; and (3) the CMV syndrome, consisting of CMV in the bloodstream along with fever, leukopenia, thrombocytopenia, and hepatic enzyme abnormalities. Ganciclovir is efficacious in the treatment of CMV infection; prophylaxis with ganciclovir or possibly with other antiviral agents may reduce the overall incidence of CMV-related disease.

Late Infections

EBV infection usually presents as a lymphoma-like proliferation of B cells late after heart transplantation, particularly in patients maintained on intense immunosuppressive therapy. A subset of heart and heart–lung transplant recipients may develop early fulminant EBV-LPD (within 2 months). Treatment includes the reduction of immunosuppression (if possible), the use of glucocorticoid and calcineurin inhibitor–sparing regimens, and the consideration of therapy with anti–B cell antibodies (rituximab and possibly others). Immunomodulatory and antiviral agents continue to be studied. Ganciclovir prophylaxis for CMV disease may indirectly reduce the risk of EBV-LPD through reduced spread of replicating EBV to naïve B cells. Aggressive chemotherapy is a last resort, as discussed earlier for HSC transplant recipients. KSHV-associated disease, including Kaposi's sarcoma and primary effusion lymphoma, has been reported in heart transplant recipients. GVHD prophylaxis with sirolimus may decrease the risk of both rejection and outgrowth of KSHV-infected cells. Antitumor therapy is discussed in **Chap. 69**. Prophylaxis for *Pneumocystis* infection is required for these patients (see "Lung Transplantation, Late Infections," below).

LUNG TRANSPLANTATION

Early Infections

It is not surprising that lung transplant recipients are predisposed to the development of pneumonia. The combination of ischemia and the resulting mucosal damage, together with accompanying denervation and lack of lymphatic drainage, probably contributes to the high rate of pneumonia (66% in one series). The prophylactic use of high doses of broad-spectrum antibiotics for the first 3–4 days after surgery may decrease the incidence of pneumonia. Gram-negative pathogens (Enterobacteriaceae and *Pseudomonas* species) are troublesome in the first 2 weeks after surgery (the period of maximal vulnerability). Pneumonia can also be caused by *Candida* (including drug-resistant strains of *C. auris*), possibly as a result of colonization of the donor lung, and by *Aspergillus* and *Cryptococcus.* Many centers use antifungal prophylaxis (typically fluconazole or liposomal amphotericin B) for the first 1–2 weeks.

Mediastinitis may occur at an even higher rate among lung transplant recipients than among heart transplant recipients and most commonly develops within 2 weeks of surgery. In the absence of prophylaxis, pneumonitis due to CMV (which may be transmitted as a consequence of transplantation) usually presents between 2 weeks and 3 months after surgery, with primary disease occurring later than reactivation disease.

Middle-Period Infections

The incidence of CMV infection, either reactivated or primary, is 75–100% if either the donor or the recipient is seropositive for CMV. CMV-induced disease after solid organ transplantation appears to be most severe in recipients of lung and heart–lung transplants. Whether this severity relates to the mismatch in lung antigen presentation and host immune cells or is attributable to nonimmunologic factors is not known. More than half of lung transplant recipients with symptomatic CMV disease have pneumonia. Difficulty in distinguishing the radiographic picture of CMV infection from that of other infections or from organ rejection further complicates therapy. CMV can also cause bronchiolitis obliterans in lung transplants. The development of pneumonitis related to HSV has led to the prophylactic use of acyclovir. Such prophylaxis may also decrease rates of CMV disease, but ganciclovir is more active against CMV and is also active against HSV. The prophylaxis of CMV infection with IV ganciclovir—or increasingly with valganciclovir, the oral alternative—is recommended for lung transplant recipients. Antiviral alternatives are discussed in the earlier section on HSC transplantation. Although the overall incidence of serious disease is decreased during prophylaxis, late disease may occur when prophylaxis is stopped—a pattern observed increasingly in recent years. With recovery from peritransplantation complications and, in many cases, a decrease in immunosuppression, the recipient is often better equipped to combat late infection.

Late Infections

The incidence of *Pneumocystis* infection (which may present with a paucity of findings) is high among lung and heart–lung transplant recipients. Some form of prophylaxis for *Pneumocystis* pneumonia is indicated in all organ transplant situations (Table 138-5). Prophylaxis with TMP-SMX for 12 months after transplantation may be sufficient to prevent *Pneumocystis* disease in patients whose immunosuppression is not increased.

As in other transplant recipients, EBV infection in lung and heart–lung recipients may cause either a mononucleosis-like syndrome or EBV-LPD. The tendency of the B cell blasts to present in the lung appears to be greater after lung transplantation than after the transplantation of other organs, possibly because of a rich source of B cells in bronchus-associated lymphoid tissue. Reduction of immunosuppression and switching of regimens, as discussed in earlier sections, cause remission in some cases, but mTOR inhibitors such as rapamycin may contribute to lung toxicity. Airway compression can be fatal, and rapid intervention may, therefore, become necessary. The approach to EBV-LPD is similar to that described in other sections.

LIVER TRANSPLANTATION

Early Infections

As in other transplantation settings, early bacterial infections are a major problem after liver transplantation. Many centers administer systemic broad-spectrum antibiotics for the first 24 h or sometimes longer after surgery, even in the absence of documented infection. However, despite prophylaxis, infectious complications are common and correlate with the duration of the surgical procedure and the type of biliary drainage. An operation lasting >12 h is associated with an increased likelihood of infection. Patients who have a choledochojejunostomy with drainage of the biliary duct to a Roux-en-Y jejunal bowel loop have more fungal infections than those whose bile is drained via anastomosis of the donor common bile duct to the recipient common bile duct. Overall, liver transplant patients have a high incidence of fungal infections, and the occurrence of fungal (often candidal) infection in the setting of choledochojejunostomy correlates with re-transplantation, elevated creatinine levels, long procedures, transfusion of >40 units of blood, reoperation, preoperative use of glucocorticoids, prolonged treatment with antibacterial agents, and fungal colonization 2 days before and 3 days after surgery. Many centers give antifungal agents prophylactically in this setting.

Peritonitis and intraabdominal abscesses are common complications of liver transplantation. Bacterial peritonitis or localized abscesses may result from biliary leaks. Early leaks are especially common with live-donor liver transplants. Peritonitis in liver transplant recipients is often polymicrobial, frequently involving enterococci, aerobic gram-negative bacteria, staphylococci, anaerobes, or *Candida* and sometimes involving other invasive fungi. Only one-third of patients with intraabdominal abscesses have bacteremia. Abscesses within the first month after surgery may occur not only in and around the liver but also in the spleen, pericolic area, and pelvis. Treatment includes antibiotic administration and drainage as necessary. Not surprisingly, *C. difficile* colitis is also a problem in this setting **(Chap. 129)**.

Middle-Period Infections

The development of postsurgical biliary stricture predisposes patients to cholangitis. The incidence of strictures is increased in live-donor liver transplantation. Transplant recipients who develop cholangitis may have high spiking fevers and rigors but often lack the characteristic signs and symptoms of classic cholangitis, including abdominal pain and jaundice. Although these findings may suggest graft rejection, rejection is typically accompanied by marked elevation of liver function enzymes. In contrast, in cholangitis in transplant recipients, results of liver function tests

(with the possible exception of alkaline phosphatase levels) are often within the normal range. Definitive diagnosis of cholangitis in liver transplant recipients requires demonstration of aggregated neutrophils in bile duct biopsy specimens. Unfortunately, invasive studies of the biliary tract (either T-tube cholangiography or endoscopic retrograde cholangiopancreatography) may themselves lead to cholangitis. For this reason, many clinicians recommend an empirical trial of therapy with antibiotics covering gram-negative organisms and anaerobes before these procedures are undertaken as well as antibiotic coverage if procedures are eventually performed.

Reactivation of viral hepatitis is a common complication of liver transplantation **(Chap. 332)**. Recurrent hepatitis B and C infections, for which transplantation may be performed, are problematic. To prevent hepatitis B virus reinfection, prophylaxis with an optimal antiviral agent or combination of agents (lamivudine, adefovir, entecavir) and hepatitis B immune globulin is currently recommended, although the optimal dose, route, and duration of therapy remain controversial. Success in preventing reinfection with hepatitis B virus has increased in recent years. Complications related to hepatitis C infection are the most common reason for liver transplantation in the United States. Without treatment, reinfection of the graft with hepatitis C virus occurs in all patients, with a variable time frame. Recent studies employing direct-acting antivirals have provided impressive results in both the treatment of existing infections before transplantation and the prevention of infections after transplantation in patients with hepatitis C **(Chap. 332)**.

As in other transplantation settings, reactivation disease with herpesviruses is common (Table 138-3). Herpesviruses can be transmitted in donor organs. Although CMV hepatitis occurs in ~4% of liver transplant recipients, it is usually not so severe as to require re-transplantation. Without prophylaxis, CMV disease develops in the majority of seronegative recipients of organs from CMV-positive donors, but fatality rates are lower among liver transplant recipients than among lung or heart–lung transplant recipients. Disease due to CMV has also been associated with the vanishing bile duct syndrome after liver transplantation. Liver transplant recipients with high levels of CMV respond to treatment with ganciclovir; prophylaxis with oral forms of ganciclovir or valganciclovir decreases the frequency of disease. A role for HHV-6 reactivation in early posttransplantation fever and leukopenia has been proposed, although the more severe sequelae described in HSC transplantation are unusual. HHV-6 and HHV-7 appear to exacerbate CMV disease in this setting. EBV-LPD after liver transplantation shows a propensity for involvement of the liver, and such disease may be of donor origin. See previous sections for discussion of EBV infections in solid organ transplantation.

PANCREAS TRANSPLANTATION

Pancreas transplantation is most frequently performed together with or after kidney transplantation, although it may be performed alone. Transplantation of the pancreas can be complicated by early bacterial and yeast infections. Most pancreatic transplants are drained into the bowel, and the rest are drained into the bladder. A cuff of duodenum is used in the anastomosis between the pancreatic graft and either the gut or the bladder. Bowel drainage poses a risk of early intraabdominal and allograft infections with enteric bacteria and yeasts. These infections can result in loss of the graft. Bladder drainage causes a high rate of urinary tract infection and sterile cystitis; however, such infection can usually be cured with appropriate antimicrobial agents. In both procedures, prophylactic antimicrobial agents are commonly used at the time of surgery. Aggressive immunosuppression, especially when the patient receives a kidney and a pancreas from different donors, is associated with late-onset systemic fungal and viral infections; thus, many centers administer an antifungal drug and an antiviral agent (ganciclovir or a congener) for extended prophylaxis.

Issues related to the development of CMV infection, EBV-LPD, and infections with opportunistic pathogens in patients receiving a pancreatic transplant are similar to those in other SOT recipients.

COMPOSITE-TISSUE TRANSPLANTATION

Composite-tissue allotransplantation (CTA) is a new field in which, rather than a single organ, multiple tissue types composing a major body part are transplanted. The sites involved have included hands, feet, arms, legs, face, trachea, and abdominal wall. The numbers of recipients are limited. The different procedures and the associated infectious complications vary. Nevertheless, some early trends related to infectious complications have become apparent, as very intense and prolonged immunosuppression is typically required to prevent rejection. For example, in the early postoperative period, bacterial infections are especially frequent in facial transplant recipients. Perioperative prophylaxis is tailored to the organisms likely to complicate the different procedures. As in SOT recipients, complicated CMV infections have been observed in several CTA settings, particularly when the recipient is seronegative and the donor is seropositive. In some patients, anti-CMV immune globulin in addition to ganciclovir (as used in HSC transplant recipients with CMV pneumonia) was needed to control disease, and ganciclovir resistance requiring alternative therapies developed in several patients. Infectious complications from reactivation of other members of the human herpesvirus family and other latent viruses also caused significant morbidity, as discussed for SOT recipients. Prophylaxis for CMV infection, *P. jirovecii* infection, toxoplasmosis, and fungal infection is administered for several months on the basis of the limited studies available.

MISCELLANEOUS INFECTIONS IN SOLID ORGAN TRANSPLANTATION

Indwelling IV Catheter Infections

The prolonged use of indwelling IV catheters for administration of medications, blood products, and nutrition is common in diverse transplantation settings and poses a risk of local and bloodstream infections. Exit-site infection is most commonly caused by staphylococcal species. Bloodstream infection most frequently develops within 1 week of catheter placement or in patients who become neutropenic. Coagulase-negative staphylococci are the most common isolates from blood. Although infective endocarditis in HSC transplant recipients is uncommon, the incidence of endocarditis among SOT recipients has been estimated to be as high as 1%, and this infection is associated with excessive high mortality in this population. Although staphylococci predominate, the involvement of fungal and gram-negative organisms may be more common than in the general population. **For further discussion of differential diagnosis and therapeutic options, see Chap. 70.**

Tuberculosis

The incidence of tuberculosis within the first 12 months after solid organ transplantation is greater than that observed after HSC transplantation (0.23–0.79%) and ranges broadly worldwide (1.2–15%), reflecting the prevalence of tuberculosis in local populations. Lesions suggesting prior tuberculosis on chest radiography, older age, diabetes, chronic liver disease, GVHD, and intense immunosuppression are predictive of tuberculosis reactivation and development of disseminated disease in a host with latent disease. Tuberculosis has rarely been transmitted from the donor organ. In contrast to the low mortality rate among HSC transplant recipients, mortality rates among SOT recipients are reported to be as high as 30%. Vigilance is indicated, as the presentation of disease is often extrapulmonary (gastrointestinal, genitourinary, central nervous, endocrine, musculoskeletal, laryngeal) and atypical; tuberculosis in this setting sometimes manifests as fever of unknown origin. Careful elicitation of a history and direct evaluation of both the recipient and the donor prior to transplantation are optimal. Skin testing of the recipient with purified protein derivative may be unreliable because of chronic disease and/or immunosuppression. Cell-based assays that measure interferon γ and/or cytokine production may prove more sensitive in the future. Isoniazid toxicity has not been a significant problem except in the setting of liver transplantation. Therefore, appropriate prophylaxis should be used (see recommendations from the Centers for Disease Control and Prevention [CDC] at

www.cdc.gov/tb/topic/treatment/ltbi.htm). An assessment of the need to treat latent disease should include careful consideration of the possibility of a false-negative test result. Pending final confirmation of suspected tuberculosis, aggressive multidrug treatment in accordance with the guidelines of the CDC, the Infectious Diseases Society of America, and the American Thoracic Society is indicated because of the high mortality rates among these patients. Altered drug metabolism (e.g., upon coadministration of antituberculous medications and certain immunosuppressive agents) can be managed with careful monitoring of drug levels and appropriate dose adjustment. Close follow-up of hepatic enzymes is warranted. Drug-resistant tuberculosis is especially problematic in these individuals **(Chap. 173)**.

Virus-Associated Malignancies

In addition to malignancy associated with gammaherpesvirus infection (EBV, KSHV) and simple warts (HPV), other tumors that are virus-associated or suspected of being virus-associated are more likely to develop in transplant recipients, particularly those who require long-term immunosuppression, than in the general population. The interval to tumor development is usually >1 year. Transplant recipients develop nonmelanoma skin or lip cancers that, in contrast to de novo skin cancers, have a high ratio of squamous cells to basal cells. HPV may play a major role in these lesions. Cervical and vulvar carcinomas, which are quite clearly associated with HPV, develop with increased frequency in female transplant recipients. The frequency of Merkel cell carcinoma associated with Merkel cell polyomavirus is also increased among transplant recipients; however, it is unclear whether recipients infected with HTLV-1 are at increased risk of leukemia. Among renal transplant recipients, rates of melanoma are modestly increased and rates of cancers of the kidney and bladder are increased. Recommendations for dealing with these problems include vaccination against HPV, a switch from calcineurin inhibitors to mTOR inhibitors (see above), and reduction of immunosuppression to the lowest level possible without graft rejection.

VACCINATION OF TRANSPLANT RECIPIENTS

In addition to receiving antibiotic prophylaxis, transplant recipients should be vaccinated against likely pathogens **(Table 138-6), (See also Chap. 118)**. In the case of HSC transplant recipients, optimal responses cannot be achieved until after immune reconstitution, despite previous immunization of both donor and recipient. Recipients of an allogeneic HSC transplant must be reimmunized if they are to be protected against pathogens. The situation is less clear-cut in the case of autologous transplantation. T and B cells in the peripheral blood may reconstitute the immune response if they are transferred in adequate numbers. However, cancer patients (particularly those with Hodgkin's disease, in whom vaccination has been extensively studied) who are undergoing chemotherapy do not respond normally to immunization, and titers of antibodies to infectious agents fall more rapidly than in healthy individuals. Therefore, even immunosuppressed patients who have not undergone HSC transplantation may need booster vaccine injections. If memory cells are specifically eliminated as part of a stem cell "cleanup" procedure, it will be necessary to reimmunize the recipient with a new primary series. Optimal times for immunizations of different transplant populations are being evaluated. Yearly immunization of household and other contacts (including health care personnel) against influenza benefits the patient by preventing local spread.

TABLE 138-6

Vaccination of Hematopoietic Stem Cell Transplant (HSCT) and Solid Organ Transplant (SOT) Recipients

Vaccine	Type of Transplantation	
	HSCT	SOT[a]
Streptococcus pneumoniae, *Haemophilus influenzae*, *Neisseria meningitidis*	Immunize after transplantation with *H. influenzae* conjugate vaccine and *N. meningitidis* conjugate vaccine. *S. pneumoniae* vaccine is given in two steps.[b]	Immunize before transplantation with *H. influenzae* conjugate and *N. meningitidis* conjugate vaccines. *S. pneumoniae* vaccine is given in two steps.[b]
Influenza	Vaccinate in the fall. Vaccinate close contacts.	Vaccinate in the fall. Vaccinate close contacts.
Polio	Administer inactivated vaccine.	Administer inactivated vaccine.
Measles/mumps/rubella	Immunize 24 months after transplantation if GVHD is absent.	Immunize before transplantation.
Diphtheria, pertussis, tetanus	Reimmunize after transplantation with primary series, DTaP. See IDSA 2013 recommendations (*www.idsociety.org/Other_Guidelines/#immunizationFortheCompromisedHost*).	Immunize or boost before transplantation with Tdap; give Td boosters at 10-year intervals or as required.
Hepatitis B and A	Reimmunize after transplantation. See recommendations.	Immunize before transplantation.

Vaccine	Type of Transplantation	
	HSCT	SOT[a]
Human papillomavirus	3 doses for persons 9–26 years of age[c]	3 doses for persons 9–26 years of age[c]

[a]Immunizations should be given before SOT whenever possible. [b]Step 1: Administer one dose of the pneumococcal conjugate vaccine Prevnar® (13-valent pneumococcal vaccine, PCV13) to all transplant candidates. If a candidate has previously been vaccinated with Pneumovax® (23-valent pneumococcal vaccine, PPSV23), at least 6 months should have elapsed before Prevnar is administered. Step 2: Administer one dose of Pneumovax at least 8 weeks after vaccination with PCV13; follow with a booster dose of Pneumovax 5 years later. If the patient has previously been vaccinated with Pneumovax (i.e., before receiving Prevnar), at least 3 years should have elapsed before the second dose of Pneumovax is administered. [c]The second and third doses should follow the first dose after 1–2 months and 6 months, respectively. Centers for Disease Control (CDC) current recommendations: *https://www.cdc.gov/mmwr/volumes/65/wr/mm6549a5.htm*. Some authorities recommend vaccination of all transplant recipients because of the risk of new sexual contacts later in life and the potential for human papillomavirus replication in immunocompromised patients.

Note: Recommendations from the CDC should be checked regularly as they frequently change upon receipt of new clinical information and new formulations of specific vaccines.

Abbreviations: DTaP, full-level diphtheria and tetanus toxoids and acellular pertussis, adsorbed; GVHD, graft-versus-host disease; Tdap, tetanus toxoid, reduced diphtheria toxoid, and acellular pertussis; IDSA, Infectious Diseases Society of America.

In the absence of compelling data as to optimal timing, it is reasonable to administer the pneumococcal and *H. influenzae* type b conjugate vaccines to both autologous and allogeneic HSC transplant recipients beginning 12 months after transplantation. A series that includes both the 13-valent pneumococcal conjugate vaccine (Prevnar®) and the 23-valent pneumococcal polysaccharide vaccine (Pneumovax®) is now recommended (according to CDC guidelines). The pneumococcal and *H. influenzae* type b vaccines are particularly important for patients who have undergone splenectomy. The *Neisseria meningitidis* polysaccharide conjugate vaccine (Menactra® or Menveo®) also is recommended. In addition, diphtheria, tetanus, acellular pertussis, and inactivated polio vaccines can all be given at these same intervals (12 months and, as required, 24 months after transplantation). Some authorities recommend a new primary series for tetanus/diphtheria/pertussis and inactivated poliovirus vaccines beginning 12 months after transplantation. Vaccination to prevent hepatitis B and hepatitis A (both killed vaccines) also seems advisable. HPV vaccination, which can prevent genital warts as well as specific cancers, is recommended through age 26 for healthy young adults who previously have not been vaccinated or have not received the full series. Live-virus measles/mumps/rubella (MMR) vaccine can be given to autologous HSC transplant recipients 24 months after transplantation and to most allogeneic HSC transplant recipients at 24 months if they are not receiving maintenance therapy with immunosuppressive drugs and do not have ongoing GVHD. The risk of spread from a household contact is low for MMR vaccine. In parts of the world where live poliovirus vaccine is used, patients as well as contacts should be advised to receive only the killed vaccine. In the rare setting where both donor and recipient are VZV naïve and the recipient is no longer receiving acyclovir or ganciclovir prophylaxis, the patient should be counseled to receive varicella-

zoster immune globulin (VariZIG®) up to 10 days after exposure to a person with chickenpox or uncovered zoster; such patients should avoid close contact with persons recently vaccinated with Varivax®. Neither patients nor their household contacts should receive vaccinia vaccine unless they have been exposed to smallpox virus. Among patients who have active GVHD and/or are taking high maintenance doses of glucocorticoids, it may be prudent to avoid all live-virus vaccines.

In the case of SOT recipients, administration of all the usual vaccines and of the indicated booster doses should be completed before immunosuppression, if possible, to maximize responses. For patients taking immunosuppressive agents, the administration of pneumococcal vaccine should be repeated every 5 years. No data are available for the meningococcal vaccine, but it is probably reasonable to administer it along with the pneumococcal vaccine. *H. influenzae* conjugate vaccine is safe and should be efficacious in this population; therefore, its administration before transplantation is recommended. Booster doses of this vaccine are not recommended for adults. SOT recipients who continue to receive immunosuppressive drugs should not receive live-virus vaccines. A person in this group who is exposed to measles should be given measles immune globulin. Similarly, an immunocompromised patient who is seronegative for varicella and who comes into contact with a person who has chickenpox should be given varicella-zoster immune globulin as soon as possible (optimally within 96 h; up to 10 days after contact); if this is not possible, a 10- to 14-day course of acyclovir therapy should be started immediately. Upon the discontinuation of treatment, clinical disease may still occur in a small number of patients; thus vigilance is indicated. Rapid re-treatment with acyclovir should limit the symptoms of disease. Household contacts of transplant recipients can receive live attenuated VZV vaccine, but vaccinees should avoid direct contact with the patient if a rash develops. Virus-like particle vaccines have been licensed for the prevention of infection with several HPV serotypes most commonly implicated in cervical and anal carcinomas and in anogenital and laryngeal warts. These vaccines are not live; current CDC recommendations are for three doses in immunocompromised hosts.

Immunocompromised patients who travel may benefit from some but not all vaccines **(Chaps. 118 and 119)**. In general, these patients should receive any killed or inactivated vaccine preparation appropriate to the area they are visiting; this recommendation includes the vaccines for Japanese encephalitis, hepatitis A and B, poliomyelitis, meningococcal infection, and typhoid. Live typhoid vaccines are not recommended for use in most immunocompromised patients, but an inactivated or purified polysaccharide typhoid vaccine can be used. Live yellow fever vaccine should not be administered, nor should live cholera vaccine. On the other hand, primary immunization or boosting with the purified-protein hepatitis B vaccine is indicated. Inactivated hepatitis A vaccine should also be used in the appropriate setting **(Chap. 118)**. A vaccine is now available that provides dual protection against hepatitis A and hepatitis B. If hepatitis A vaccine is not administered, travelers should consider receiving passive protection with immune globulin (the dose depending on the duration of travel in the high-risk area).

FURTHER READING

Dantal J, Campone M: Daunting but worthy goal: Reducing the de novo cancer incidence after transplantation. Transplantation 100:2569, 2016.

Danziger-Isakov L et al: Vaccination in solid organ transplantation. Am J Transplant 13(Suppl 4):311, 2013. [PubMed: 23465023]

Fishman JA: Infection in organ transplantation. Am J Transplant 17:856, 2017.

Lantieri L et al: Face transplant: Long-term follow-up and results of a prospective open study. Lancet 388:1398, 2016.

Schuster MG et al: Infections in hematopoietic cell transplant recipients: Results from the Organ Transplant Infection Project, a multicenter, prospective, cohort study. Open Forum Infect Dis 4:ofx050, 2017. [PubMed: 28491889]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 139: Treatment and Prophylaxis of Bacterial Infections

David C. Hooper; Erica S. Shenoy; Christy A. Varughese

INTRODUCTION

Antimicrobial agents have had a major impact on human health. Together with vaccines, they have contributed to reduced mortality, extended lifespan, and enhanced quality of life. Among drugs used in human medicine, however, they are distinctive in that their use promotes the occurrence of drug resistance in the pathogens they are designed to treat as well as in other "bystander" organisms. Indeed, the history of antimicrobial development has been driven in large part by the medical need engendered by the emergence of resistance to each generation of agents. Thus, the careful and appropriate use of antimicrobial drugs is particularly important not only for optimizing efficacy and minimizing adverse effects but also for minimizing the risk of resistance and preserving the value of existing agents. Although this chapter focuses on antibacterial agents, the optimal use of all antimicrobials depends on an understanding of each drug's mechanism of action, spectrum of activity, mechanisms of resistance, pharmacology, and adverse effect profile. This information is applied in the context of the patient's clinical presentation, underlying conditions, and epidemiology to define the site and likely nature of the infection or other condition and thus to choose the best therapy. Gathering of microbiologic information is important for refining therapeutic choices on the basis of documented pathogen and susceptibility data whenever possible; this information also makes it possible to choose more targeted therapy, thereby reducing the risk of selection of resistant bacteria. Durations of therapy are chosen according to the nature of the infection and the patient's response to treatment and are informed by clinical studies when they are available, with the understanding that shorter courses are less likely than longer ones to promote the emergence of resistance. This chapter and the one that follows provide specific information that is necessary for making informed choices among antibacterial agents. The mechanisms of action of antibacterial agents are discussed in detail in the text of this chapter, and mechanisms of resistance are discussed in detail in **Chap. 140**. Both types of mechanisms are summarized for the most commonly used groups of agents in **Table 139-1**. A schematic of antibacterial targets is provided in **Fig. 139-1**.

TABLE 139-1

Mechanisms of Action of and Resistance to Antibacterial Agents

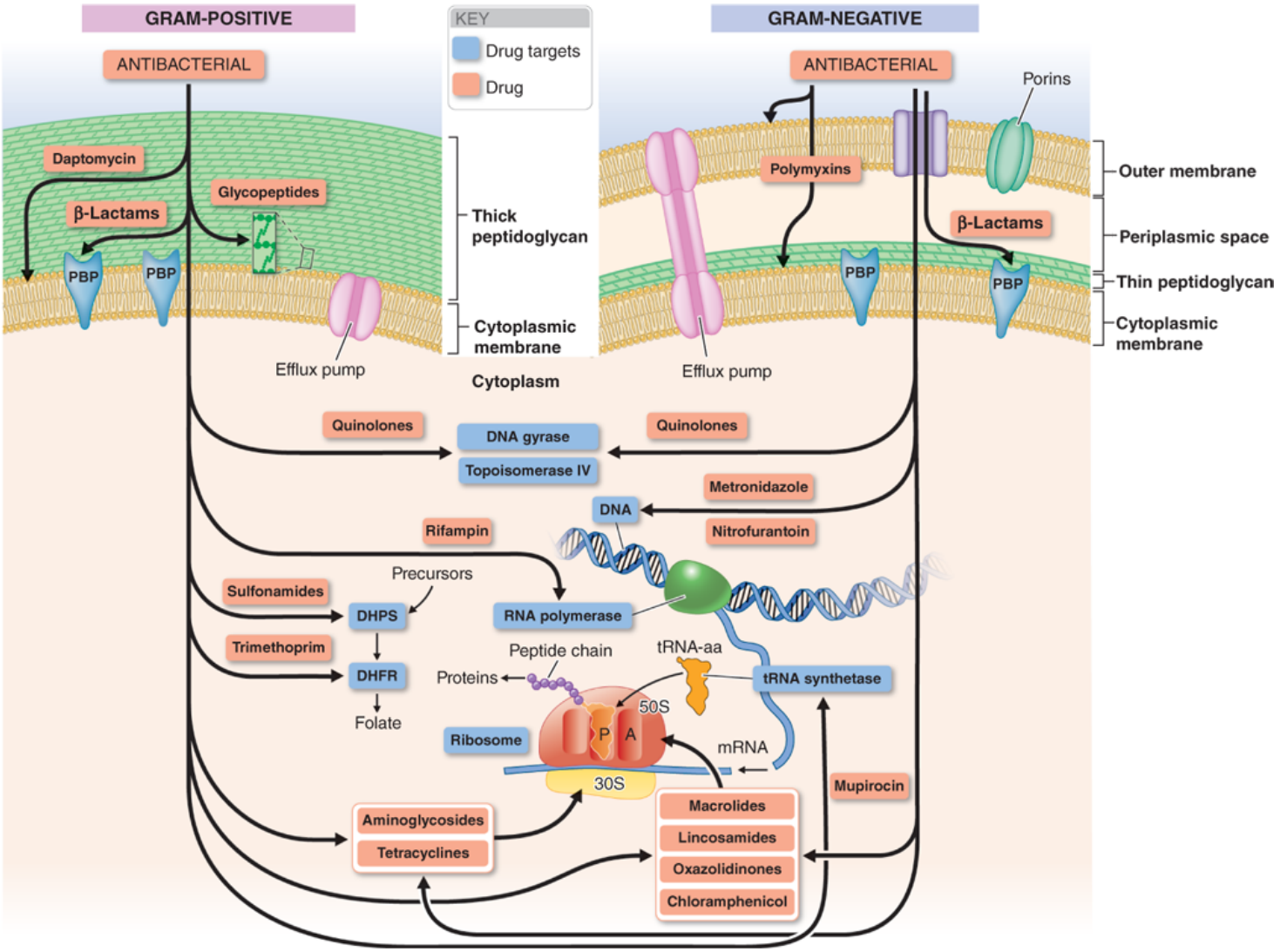
Antibacterial Agent(s)	Major Target	Mechanism(s) of Action	Mechanism(s) of Resistance
β-Lactams (penicillins, cephalosporins, monobactams, carbapenems)	Cell wall synthesis	Bind cell wall cross-linking enzymes (PBPs, transpeptidases)	1. Drug inactivation by β-lactamases 2. Altered PBP targets 3. Reduced diffusion through porin channels
Glycopeptides (vancomycin, teicoplanin, telavancin, dalbavancin, oritavancin)	Cell wall synthesis	Block cell wall glycosyltransferases by binding D-Ala-D-Ala stem-peptide terminus Teicoplanin, telavancin, dalbavancin, and oritavancin: affect membrane function	1. Altered D-Ala-D-Ala target (D-Ala-D-Lac) 2. Increased D-Ala-D-Ala target binding at sites distant from cell wall synthesis enzymes
Bacitracin	Cell wall synthesis	Blocks lipid carrier of cell wall precursors	Active drug efflux
Fosfomycin	Cell wall synthesis	Blocks linkage of stem peptide to NAG by enoyltransferase	1. Target enzyme overexpression 2. Drug-modifying enzymes
Aminoglycosides (gentamicin, tobramycin, amikacin)	Protein synthesis	Bind 30S ribosomal subunit Block translocation of peptide chain Cause misreading of mRNA	1. Drug-modifying enzymes 2. Methylation at ribosome binding site 3. Decreased permeation to target due to active efflux
Tetracyclines (tetracycline, doxycycline, minocycline)	Protein synthesis	Bind 30S ribosomal subunit Inhibit peptide elongation	1. Active drug efflux 2. Ribosomal protection proteins
Tigecycline	Protein synthesis	Same as tetracyclines	Active drug efflux (pumps different from those affecting tetracyclines)
Macrolides (erythromycin, clarithromycin, azithromycin) and ketolide (telithromycin)	Protein synthesis	Bind 50S ribosomal subunit Block peptide chain exit	1. Methylation at ribosome binding site 2. Active drug efflux
Lincosamides (clindamycin)	Protein synthesis	Bind 50S ribosomal subunit Block peptide bond formation	Methylation at ribosome binding site
Streptogramins (quinupristin, dalfopristin)	Protein synthesis	Same as macrolides	1. Same as macrolides 2. Drug-modifying enzymes
Chloramphenicol	Protein synthesis	Binds 50S ribosomal subunit Blocks aminoacyl tRNA positioning	Drug-modifying enzymes
Oxazolidinones (linezolid, tedizolid)	Protein synthesis	Bind 50S ribosomal subunit Inhibit initiation of peptide synthesis	1. Altered rRNA binding site 2. Methylation of ribosome binding site
Mupirocin	Protein synthesis	Blocks isoleucyl tRNA synthetase	1. Acquired resistant tRNA synthetase (drug bypass) 2. Altered native tRNA synthetase target

Antibacterial Agent(s)	Major Target	Mechanism(s) of Action	Mechanism(s) of Resistance
Sulfonamides (sulfadiazine, sulfisoxazole, and sulfamethoxazole)	Folate synthesis	Inhibit dihydropteroate synthetase	Acquired resistant dihydropteroate synthetase (drug bypass)
Trimethoprim	Folate synthesis	Inhibits DHFR	Acquired resistant DHFR (drug bypass)
Quinolones (norfloxacin, ciprofloxacin, ofloxacin, levofloxacin, moxifloxacin, gemifloxacin, delafloxacin)	DNA synthesis	Inhibit DNA gyrase and DNA topoisomerase IV Enzyme–DNA–drug complex: block DNA replication apparatus	1. Altered target(s) 2. Active efflux 3. Protection of target from drug 4. Drug-modifying enzyme (ciprofloxacin)
Rifamycins (rifampin, rifabutin, rifapentine)	RNA synthesis	Inhibit RNA polymerase	Altered target
Nitrofurantoin	Nucleic acid synthesis	Reduces reactive drug derivatives that damage DNA	Altered drug-activating enzymes
Metronidazole	Nucleic acid synthesis	Reduces reactive drug derivatives that damage DNA	1. Altered drug-activating enzyme 2. Acquired detoxifying enzymes 3. Active efflux
Polymyxins (polymyxin B and polymyxin E [colistin])	Cell membrane	Bind LPS and disrupt both outer and cytoplasmic membranes	Altered cell membrane charge with reduced drug binding
Daptomycin	Cell membrane	Produces membrane channel and membrane leakage	Altered cell membrane with reduced drug binding

Abbreviations: DHFR, dihydrofolate reductase; LPS, lipopolysaccharide; NAG, *N*-acetylglucosamine; PBP, penicillin-binding protein.

FIGURE 139-1

Antibacterial targets. A, aminoacyl site; DHFR, dihydrofolate reductase; DHPS, dihydropteroate synthetase; P, peptidyl site; PBP, penicillin-binding protein; tRNA-aa, aminoacyl tRNA.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MECHANISMS OF ACTION

Multiple essential components of bacterial cell structures and metabolism have been the targets of antibacterial agents used in clinical medicine, and the interaction of an agent with its target results in either inhibition of bacterial growth and replication (*bacteriostatic effect*) or bacterial killing (*bactericidal effect*). In general, targets have been chosen because they either do not exist in mammalian cells and physiology or are sufficiently different from their bacterial counterparts to allow selective bacterial targeting. Treatment with bacteriostatic agents is effective when the patient's host defenses are sufficient to contribute to eradication of the infecting pathogen. In patients with impaired host defenses (e.g., neutropenia) or infections at body sites with impaired or limited host defenses (e.g., meningitis and endocarditis), bactericidal agents are generally preferred.

INHIBITION OF CELL WALL SYNTHESIS

The bacterial cell wall, which is external to the cytoplasmic membrane and has no counterpart in mammalian cells, protects bacterial cells from lysis under low osmotic conditions. The cell wall is a cross-linked peptidoglycan composed of a polymer of alternating units of *N*-acetylglucosamine (NAG) and *N*-acetylmuramic acid (NAM), four-amino-acid stem peptides linked to each NAM, and a peptide cross-bridge that links adjacent stem peptides to form a netlike structure. Several steps in peptidoglycan synthesis are targets of antibacterial agents. Inhibition of cell wall synthesis generally results in a bactericidal effect that is linked to cell lysis. This effect results not only from the blocking of new cell-wall formation but from the uninhibited action of cell wall–remodeling enzymes called *autolysins*, which cleave peptidoglycan as part of normal cell-wall growth.

In gram-positive bacteria the peptidoglycan is the most external cell structure, but in gram-negative bacteria an asymmetric lipid outer membrane is external to the peptidoglycan and contains diffusion channels called *porins*. The space between the cytoplasmic membrane peptidoglycan and the outer membrane is referred to as the *periplasmic space*. Most antibacterial drugs enter the gram-negative bacterial cell through a porin channel, since the outer

membrane is a major diffusion barrier. Although the peptidoglycan layer is thicker in gram-positive (20–80 nm) than in gram-negative (1 nm) bacteria, peptidoglycan itself constitutes only a limited diffusion barrier for antibacterial agents.

β-Lactams

The β-lactam drugs, including penicillins, cephalosporins, monobactams, and carbapenems, target transpeptidase enzymes (also called *penicillin-binding proteins* or PBPs) involved in the stem-peptide cross-linking step. Inhibitors of β-lactamases—enzymes that can degrade β-lactams—are used in combination with some β-lactams to expand their spectrum of activity.

Glycopeptides and Lipoglycopeptides

The glycopeptides, including vancomycin and teicoplanin, and the lipoglycopeptides, including telavancin, dalbavancin, and oritavancin, bind the two terminal D-alanine residues of the stem peptide, hindering the glycosyltransferase involved in polymerizing NAG–NAM units as well as transpeptidases. Vancomycin also binds to the lipid II intermediate that delivers cell wall precursor subunits. The additional binding of teicoplanin, telavancin, dalbavancin, and oritavancin to the bacterial cytoplasmic membrane contributes to their increased potency. Both β-lactams and glycopeptides interact with their targets external to the cytoplasmic membrane.

Bacitracin (Topical) and Fosfomycin

These agents interrupt enzymatic steps in the production of peptidoglycan precursors in the cytoplasm.

INHIBITION OF PROTEIN SYNTHESIS

Most inhibitors of bacterial protein synthesis target bacterial ribosomes, whose difference from eukaryotic ribosomes allows selective antibacterial action. Some inhibitors bind to the 30S ribosomal subunit and others to the 50S subunit. Most protein synthesis–inhibiting agents are bacteriostatic; aminoglycosides are an exception and are bactericidal.

Aminoglycosides

Aminoglycosides (amikacin, gentamicin, kanamycin, netilmicin, streptomycin, and tobramycin) bind irreversibly to 16S ribosomal RNA (rRNA) of the 30S ribosomal subunit, blocking the translocation of peptidyl transfer RNA (tRNA) from the A (aminoacyl) to the P (peptidyl) site and, at low concentrations, causing misreading of messenger RNA (mRNA) codons and thus causing the introduction of incorrect amino acids into the peptide chain; at higher concentrations, translocation of the peptide chain is blocked. Cellular uptake of aminoglycosides is dependent on the electrochemical gradient across the bacterial membrane. Under anaerobic conditions, this gradient is reduced, with a consequent reduction in the uptake and activity of the aminoglycosides. Spectinomycin is a related aminocyclitol antibiotic that also binds to 16S rRNA of the 30S ribosomal subunit but at a different site. This drug inhibits translocation of the growing peptide chain but does not trigger codon misreading and produces only a bacteriostatic effect.

Tetracyclines and Glycylcyclines

Tetracyclines (doxycycline, minocycline, and tetracycline) bind reversibly to the 16S rRNA of the 30S ribosomal subunit and block the binding of aminoacyl tRNA to the ribosomal A site, thereby inhibiting peptide elongation. Active transport of tetracyclines into bacterial but not mammalian cells contributes to the selectivity of these agents. Tigecycline, a derivative of minocycline and the only available glycylcycline, acts similarly to the tetracyclines but is distinctive for its ability to circumvent the most common mechanisms of resistance to the tetracyclines.

Macrolides and Ketolides

In contrast to the aminoglycosides and tetracyclines, the macrolides (azithromycin, clarithromycin, and erythromycin) and ketolides (telithromycin) bind to the 23S rRNA of the 50S ribosomal subunit. These agents block translocation of the growing peptide chain by binding to the tunnel from which the chain exits the ribosome.

Lincosamides

Clindamycin is the only lincosamide in clinical use. It binds to the 23S rRNA of the 50S ribosomal subunit, interacting with both the ribosomal A and P sites and blocking peptide bond formation.

Streptogramins

The only streptogramin in clinical use is a combination of quinupristin, a group B streptogramin, and dalfopristin, a group A streptogramin. Both components bind to 23S rRNA of the 50S ribosome: dalfopristin binds to both the A and P sites of the peptidyl transferase center, and quinupristin binds to a site that overlaps the macrolide-binding site, blocking the emergence of nascent peptide from the ribosome. The combination is bactericidal, but macrolide-resistant bacteria exhibit cross-resistance to quinupristin, and the remaining activity of dalfopristin alone is only bacteriostatic.

Chloramphenicol

Chloramphenicol binds reversibly to the 23S rRNA of the 50S subunit in a manner that interferes with the proper positioning of the aminoacyl component of tRNA in the A site. This site of binding is near those of the macrolides and lincosamides.

Oxazolidinones

Linezolid and tedizolid are the only oxazolidinones in clinical use. They bind directly to the A site in the 23S rRNA of the 50S ribosomal subunit and block binding of aminoacyl tRNA, inhibiting the initiation of protein synthesis.

Mupirocin

Mupirocin (pseudomonic acid) is used topically. It competes with isoleucine for binding to isoleucyl tRNA synthetase, depleting stores of isoleucyl tRNA and thereby inhibiting protein synthesis.

INHIBITION OF BACTERIAL METABOLISM

Available inhibitors (antimetabolites) target the pathway for synthesis of folate, which is a cofactor in a number of one-carbon transfer reactions involved in the synthesis of some nucleic acids, including the pyrimidine thymidine and all purines (adenine and guanine), as well as some amino acids (methionine and serine) and acetyl coenzyme A. Two sequential steps in folate synthesis are targeted. The selective antibacterial effect stems from the inability of mammalian cells to synthesize folate; they depend instead on exogenous sources. Antibacterial activity, however, may be reduced in the presence of high exogenous concentrations of the end products of the folate pathway (e.g., thymidine and purines) that may occur in some infections, resulting from local breakdown of leukocytes and host tissues.

Sulfonamides

Sulfonamides, including sulfadiazine, sulfisoxazole, and sulfamethoxazole, inhibit dihydropteroate synthetase (DHPS), which adds *p*-aminobenzoic acid (PABA) to pteridine, producing dihydropteroate. Sulfonamides are structural analogues of PABA and act as competing enzyme substrates.

Trimethoprim

Subsequent steps in folate synthesis are catalyzed by dihydrofolate synthase, which adds glutamate to dihydropteroate, and dihydrofolate reductase (DHFR), which then generates the final product, tetrahydrofolate. Trimethoprim is a structural analogue of pteridine and inhibits DHFR. Trimethoprim is available alone but is most often used in combination products that also contain sulfamethoxazole and thus block two sequential steps in folate synthesis.

INHIBITION OF DNA AND RNA SYNTHESIS OR ACTIVITY

A variety of antibacterial agents act on these processes.

Quinolones

The quinolones include nalidixic acid, the first agent in the class, and newer, more widely used fluorinated derivatives (fluoroquinolones), including norfloxacin, ciprofloxacin, levofloxacin, moxifloxacin, gemifloxacin, and delafloxacin. The quinolones are synthetic compounds that inhibit bacterial DNA synthesis by interacting with the DNA complexes of two essential enzymes, DNA gyrase and DNA topoisomerase IV, which alter DNA topology. Quinolones trap enzyme–DNA complexes in such a way that they block movement of the DNA replication apparatus and can generate lethal double-strand breaks in DNA, resulting in bactericidal activity. Although mammalian cells also have type II DNA topoisomerases related to gyrase and topoisomerase IV, the structures of the mammalian enzymes are sufficiently different from those of the bacterial enzymes that quinolones have substantially selective antibacterial activity.

Rifamycins

Rifampin, rifabutin, and rifapentine are semisynthetic derivatives of rifamycin B and bind the β subunit of bacterial RNA polymerase, thereby blocking elongation of mRNA. Their action is highly selective for the bacterial enzyme over mammalian RNA polymerases.

Nitrofurantoin

The reduction of nitrofurantoin, a nitrofuran compound, by bacterial enzymes produces highly reactive derivatives that are thought to cause DNA strand breakage. Nitrofurantoin is used only for the treatment of lower urinary tract infections.

Metronidazole

Metronidazole is a synthetic nitroimidazole with activity limited to anaerobic bacteria and certain anaerobic protozoa. Reduction of its nitro group by the electron-transport system in anaerobic bacteria produces reactive intermediates that damage DNA and result in bactericidal activity. Both nitrofurantoin and metronidazole have selective antibacterial activity because the reducing activity needed to produce active derivatives is generated only by bacterial and not mammalian enzymes.

DISRUPTION OF MEMBRANE INTEGRITY

The integrity of the bacterial cytoplasmic membrane—and, in gram-negative bacteria, the outer membrane—is important for bacterial viability. Two bactericidal drugs have membrane targets.

Polymyxins

The polymyxins, including polymyxin B and polymyxin E (colistin), are cationic cyclic polypeptides that disrupt the cytoplasmic membrane and the outer membrane (the latter by binding lipopolysaccharide, which is negatively charged).

Daptomycin

Daptomycin is a lipopeptide that binds the cytoplasmic membrane of gram-positive bacteria in the presence of calcium, generating a channel that leads to leakage of cytoplasmic potassium ions and membrane depolarization.

PHARMACOKINETICS AND PHARMACODYNAMICS

The term *pharmacokinetics* describes the disposition of a drug in the body, whereas *pharmacodynamics* describes the determinants of drug action on the pathogen in relation to pharmacokinetic factors. An understanding of the principles governing these two areas is required for effective drug selection and dosing and for prevention of toxicities.

PHARMACOKINETICS

The process of drug disposition has four principal phases: absorption, distribution, metabolism, and excretion. These components determine the time course of drug concentrations in serum and subsequently the concentrations in other tissues and body fluids.

Absorption

When a drug is given by a particular route, *absorption* is defined as the percentage of the dose that reaches the systemic circulation. For example, since IV administration provides direct access to the systemic circulation, 100% of a drug dose given IV is usually absorbed. The level of absorption becomes more relevant when non-IV routes are used—e.g., the oral, IM, SC, and topical routes. The percentage of a drug that is absorbed is termed its *bioavailability*. Examples of antibacterial agents with a high oral bioavailability include metronidazole, levofloxacin, and linezolid. IV administration and oral dosing for highly bioavailable agents usually give equivalent results. Many factors can affect a drug's oral bioavailability, including the timing of food consumption relative to drug administration, drug-metabolizing enzymes, efflux transporters, concentration-dependent solubility, and acid degradation. Underlying conditions such as diarrhea or ileus can also affect the site of drug absorption and thereby alter bioavailability. Certain orally administered drugs have lower bioavailability because of the *first-pass effect*—the process by which drugs are absorbed in the small intestine through the portal circulation and then directly transported to the liver for metabolism.

Distribution

Distribution describes the process by which a drug transfers reversibly between the general circulation and the tissues. After absorption into the general circulation and the central compartment (the extensively perfused organs), the drug will also distribute into the peripheral compartment (less well-perfused tissues). The volume of distribution (Vd) is a pharmacokinetic parameter that describes the amount of drug in the body at a given time relative to the measured serum concentration. Properties such as the drug's lipophilicity, partition coefficient within different body tissues, and protein binding; blood flow; and pH can affect the Vd. Drugs with a small Vd are limited to certain areas within the body (typically extracellular fluid), whereas those with a higher Vd penetrate extensively into tissues throughout the body. Antibacterial drugs can bind to serum proteins, and a given drug is usually described as either poorly or highly protein bound. Only the unbound (free) drug is active and available to exert antibacterial effects. For example, because tigecycline is highly protein bound and also has a large Vd, concentrations of free drug in the serum are low.

Metabolism

Metabolism is the chemical transformation of a drug by the body. This modification can occur within several areas; the liver is the organ most commonly involved. Drugs are metabolized by enzymes, but enzyme systems have a finite capacity to metabolize a substrate drug. If a drug is given in a dose at which the concentration does not exceed the rate of metabolism, then the metabolic process is generally linear. If the dose exceeds the amount that can be metabolized, drug accumulation and potential toxicity may occur. Drugs are metabolized through phase I or phase II reactions. In phase I reactions, the

drug is made more polar through dealkylation, hydroxylation, oxidation, and deamination. Polarity facilitates drug removal from the body. Phase II reactions, which include glucuronidation, sulfation, and acetylation, result in compounds larger and more polar than the parent drug. Both phases usually inactivate the parent drug, but some drugs are rendered more active. The cytochrome P450 (CYP) enzyme system is responsible for phase I reactions and is generally found in the liver. CYP3A4 is a common subfamily within this system that is responsible for the majority of drug metabolism. Antibacterial drugs can be substrates, inhibitors, or inducers of a particular CYP enzyme. Inducers, such as rifampin, can increase the production of CYP enzymes and consequently increase the metabolism of other drugs. Inhibitors, such as quinupristin-dalfopristin, cause a decrease in enzyme activity (or competition for CYP substrate) and therefore an increase in the concentration of the interacting drug.

Excretion

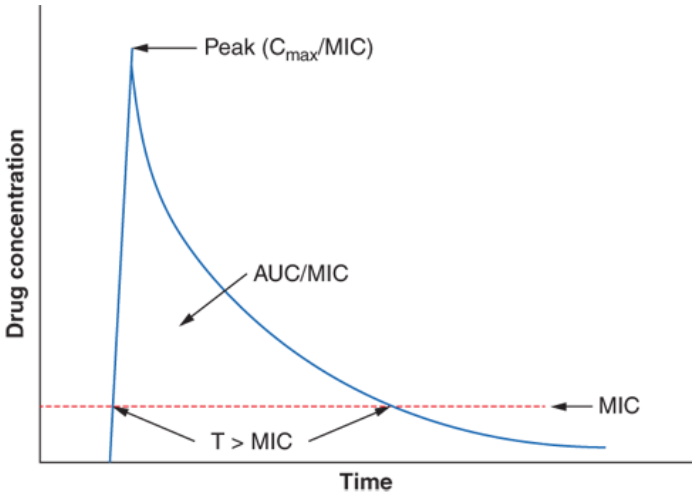
Excretion describes the body's mechanisms of drug elimination. Drugs can be eliminated through more than one mechanism. Renal clearance is the most common route and includes elimination through glomerular filtration, tubular secretion, and/or passive diffusion. Some agents have nonrenal clearance and rely on the biliary tree or the intestine for excretion. Excretion affects the half-life of a drug—i.e., the time it takes for the blood concentration of a drug to decrease by one-half. This value can range from minutes to days. Approximately five to seven half-lives are required for a drug to reach steady state when multiple doses are given in a time frame shorter than the half-life itself. Drug half-life and overall clearance can be extended if the organ responsible for clearance is impaired. Patients with renal or hepatic impairment may require dose adjustments that take delayed clearance into account and prevent toxicities from drug accumulation. For example, imipenem is cleared predominantly through glomerular filtration, and in the presence of renal impairment the dosing interval is typically increased to account for the increased half-life.

PHARMACODYNAMICS

The term *pharmacodynamics* describes the relationship between the serum concentrations that determine the efficacy of the drug and the serum concentrations that produce the toxic effects of the drug. For an antibacterial agent, the pharmacodynamic focus is the type of drug exposure needed for optimal antibacterial effect in relation to the minimal inhibitory concentration (MIC)—the lowest drug concentration that inhibits the growth of a microorganism under standardized laboratory conditions. Antibacterial effect usually correlates with one of the following parameters: (1) ratio of peak serum concentration to the MIC (C_{max}/MIC), (2) ratio of the area under the concentration–time curve to the MIC (AUC/MIC), or (3) duration of concentrations above the MIC (T > MIC) **(Fig. 139-2)**.

FIGURE 139-2

Pharmacokinetic and pharmacodynamic model predicting efficacy of antibacterial drugs. AUC, area under the time–concentration curve; C_{max}, peak serum concentration of drug; MIC, minimal inhibitory concentration; T > MIC, duration of drug concentrations above the MIC.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

For *concentration-dependent* killing agents, as the designation implies, the higher the drug concentration, the higher the rate and extent of bacterial killing. Aminoglycosides fit into the C_{max}/MIC model of pharmacodynamics activity, and a particular peak serum concentration is often targeted to achieve optimal killing. Fluoroquinolones exemplify antibacterial agents for which the AUC/MIC is a predictor of efficacy. For example, studies have found that an AUC/MIC ratio of >30 will maximize killing of *S. pneumoniae* by fluoroquinolones. In contrast, *time-dependent* killing agents reach a ceiling at which higher concentrations do not result in increased effect. Instead, these agents are active against bacteria only when the drug concentration is above the MIC. The T > MIC predicts clinical efficacy for all β-lactams. The longer the concentration of the β-lactam remains above the MIC for an infecting pathogen during the dosing interval, the greater the killing effect. For some drug classes, such as aminoglycosides, a *postantibiotic effect*—the delayed regrowth of surviving bacteria after exposure to an antibiotic—supports less frequent dosing.

APPROACH TO THERAPY

The approach to antibiotic therapy is driven by host factors, site of infection, and local resistance profiles of suspected or known pathogens. Further, national and local drug shortages and formulary restrictions can affect available therapies. Regular monitoring of the patient and collection of laboratory data should be undertaken to streamline antibacterial therapy as appropriate and to investigate the possibility of treatment failure if the patient fails to respond appropriately.

EMPIRICAL AND DIRECTED THERAPY

Therapy is considered *empirical* when the causative agent has yet to be determined and therapeutic decisions are based on the severity of illness, the clinician's assessment of likely pathogens in light of the clinical syndrome, the patient's medical conditions and prior therapy, and relevant epidemiologic factors. For patients with severe illness, empirical therapy often takes the form of an antibacterial combination that provides broad coverage of diverse agents and thus ensures adequate treatment of possible pathogens while additional data are being collected. *Directed* therapy is predicated on identification of the pathogen, determination of its susceptibility profile, and establishment of the extent of the infection. Directed therapy generally allows the use of more targeted and narrower-spectrum antibacterial agents than does empirical therapy.

Information on epidemiology, exposures, and local antibacterial susceptibility patterns can help guide empirical therapy. When empirical treatment is clinically appropriate, care should be taken to obtain clinical specimens for microbiologic analysis before the initiation of therapy and to de-escalate therapy as new information is obtained about the patient's clinical condition and the causal pathogens. De-escalation to the point of directed therapy can limit unnecessary risks to the patient as well as the risk of emergence of antibacterial resistance.

SITE OF INFECTION

The site of infection is a consideration in antibacterial therapy, largely because of the differing abilities of drugs to penetrate and achieve adequate concentrations at particular body sites. For example, to be effective in the treatment of meningitis, an agent must (1) be able to cross the blood–brain barrier and reach adequate concentrations in the cerebrospinal fluid (CSF) and (2) be active against the relevant pathogen(s). Dexamethasone, administered with or 15–20 min before the first dose of an antibacterial drug, has been shown to improve outcomes in patients with some types of acute bacterial meningitis, but its use may reduce penetration of some antibacterial agents, such as vancomycin, into the CSF. In this case, rifampin is added because its penetration is not reduced by dexamethasone. Infections at other sites where either pathogens are protected from normal host defenses or penetration of an antibacterial drug is suboptimal include osteomyelitis, prostatitis, intraocular infections, and abscesses. In such cases, consideration must be given to the mechanism of drug delivery (e.g., intravitreal injections) as well as to the role of interventions to drain, debride, or otherwise reduce the barriers to effective antibacterial therapy.

HOST FACTORS

Host factors, including immune function, pregnancy, allergies, age, renal and hepatic function, drug–drug interactions, comorbid conditions, and occupational or social exposures, should be considered.

Immune Dysfunction

Patients with deficits in immune function that blunt the response to bacterial infection, including neutropenia, deficient humoral immunity, and asplenia (either surgical or functional), are all at increased risk of severe bacterial infection. Such patients should be treated aggressively and often broadly in the early stages of suspected infection pending results of microbiologic tests. For asplenic patients, treatment should include coverage of encapsulated organisms, particularly *Streptococcus pneumoniae*, that may cause rapidly life-threatening infection.

Pregnancy

Pregnancy affects decisions regarding antibacterial therapy in two respects. First, pregnancy is associated with an increased risk of particular infections (e.g., those caused by *Listeria*). Second, the potential risks to the fetus that are posed by specific drugs must be considered. As for other drugs, the safety of the vast majority of antibacterial agents in pregnancy has not been established, and such agents are grouped in categories B and C by the U.S. Food and Drug Administration. Drugs in categories D and X are contraindicated in pregnancy or lactation due to established risks. The risks associated with antibacterial use in pregnancy and during lactation are summarized in **Table 139-2**.

TABLE 139-2

Risks Associated with Use of Antibacterial Drugs in Pregnancy and Lactation

Pregnancy Category[a]	Antibacterial Drug	Fetal Risk Recommendation[b]	Breast-Feeding Risk Recommendation[b]
B	Azithromycin	Limited human data. Animal data suggest low risk.	Limited human data; probably compatible
	Cephalosporins (including cephalexin, cefuroxime, cefixime, cefpodoxime, cefotaxime, ceftriaxone)	Compatible	Compatible
	Ceftazidime-avibactam	No human data; no fetal harm in animal studies	Ceftazidime is excreted into human milk in low concentrations. Avibactam is excreted into the milk of lactating rats; no human studies have been conducted.
	Ceftolozane-tazobactam	Compatible	Unknown
	Clindamycin	Compatible	Compatible
	Ertapenem	No human data; probably compatible	Limited human data; probably compatible
	Erythromycin	Compatible (except for estolate salt)	Compatible
	Meropenem and meropenem-vaborbactam	No human data. Animal data suggest low risk.	No human data; probably compatible
	Metronidazole	Human data suggest low risk.	Interrupt breast-feeding for 12–24 h after single 2-g dose. Limited human data; potential toxicity in divided doses
	Nitrofurantoin	Human data suggest risk in third trimester.	Limited human data; probably compatible. Higher risk associated with younger infants and those with G6PD deficiency
	Penicillins (including amoxicillin, ampicillin, cloxacillin)	Compatible	Compatible
	Quinupristin-dalfopristin	Compatible. Maternal benefit must far outweigh risk to embryo/fetus.	No human data; potential toxicity
	Vancomycin	Compatible	Limited human data; probably compatible
C	Chloramphenicol	Compatible	Limited human data; potential toxicity
	Fluoroquinolones	Human data suggest low risk.	Limited human data; probably compatible
	Clarithromycin	Limited human data. Animal data suggest high risk.	No human data; probably compatible
	Imipenem-cilastatin	Limited human data. Animal data suggest low risk.	Limited human data; probably compatible
	Linezolid	Compatible. Maternal benefit must far outweigh risk to embryo/fetus.	No human data; potential toxicity
	Telavancin	No human data. Animal studies have revealed evidence of teratogenicity.[c]	No human data. Animal studies have revealed evidence of teratogenicity.[c]

Pregnancy Category[a]	Antibacterial Drug	Fetal Risk Recommendation[b]	Breast-Feeding Risk Recommendation[b]
	Tedizolid	Limited data. Embryo-fetal studies in mice, rats, and rabbits have demonstrated fetal developmental toxicities. Use only if benefit outweighs risk.	Excreted in the breast milk of rats; unknown in humans; caution use
	Dalbavancin	Limited human data. At high doses in animal studies, delayed fetal maturation, increased embryo and offspring death. Use only if benefit outweighs risk.	Excreted in the breast milk of animals; unknown in humans; caution use
	Oritavancin	Limited human data. Studies in rats and rabbits demonstrated no harm at 25% of recommended human dose. Use only if benefit outweighs risk.	Excreted in the breast milk of rats; unknown in humans; caution use
C/D	Amikacin	Human data suggest low risk.	Compatible
	Gentamicin	Human data suggest low risk.	Compatible
D	Kanamycin	Human data suggest risk.	Limited human data; probably compatible
	Streptomycin	Human data suggest risk.	Compatible
	Sulfonamides	Human data suggest risk in third trimester.	Limited human data; potential toxicity. Avoid in ill, stressed, premature infants and in infants with hyperbilirubinemia or G6PD deficiency.
	Tetracyclines	Contraindicated in second and third trimesters	Compatible
	Tigecycline	Human data suggest risk in second and third trimesters.	No human data; potential toxicity

[a] *Category B: Either* animal reproduction studies have failed to demonstrate a risk to the fetus, and there are no adequate and well-controlled studies in pregnant women; *or* animal studies have shown an adverse effect, but adequate and well-controlled studies in pregnant women have failed to demonstrate a risk to the fetus in any trimester. *Category C:* Animal reproduction studies have shown an adverse effect on the fetus, and there are no adequate and well-controlled studies in humans, but potential benefits may warrant use of the drug in pregnant women despite potential risks. *Category D:* There is positive evidence of human fetal risk based on adverse-reaction data from investigational or marketing experience or studies in humans, but potential benefits may warrant use of the drug in pregnant women despite potential risks.[b] Fetal risk recommendation and breast-feeding risk recommendation adapted from GG Briggs et al (eds): *Drugs in Pregnancy and Lactation*, 9th ed. Philadelphia, Lippincott Williams and Wilkins, 2011; and the U.S. Food and Drug Administration (Drugs@FDA). [c] A registry has been established to monitor pregnancy outcomes of pregnant women exposed to telavancin. Physicians are encouraged to register pregnant patients, or pregnant women may enroll themselves by calling 1-855-633-8479.

Abbreviation: G6PD, glucose-6-phosphate dehydrogenase.

Allergies

Allergies to antibiotics are among the most common allergies reported, and an allergy history should be obtained whenever possible before therapy is chosen. A detailed allergy history can shed light on the type of reaction experienced previously and on whether rechallenge with the same or a related medication is advisable (and, if so, under what circumstances). Allergies to the penicillins are most common. Although as many as 10% of patients may report an allergy to penicillin, studies suggest that more than 90% of these patients could tolerate a penicillin or cephalosporin. Adverse effects **(Table 139-3)** should be distinguished from true allergies to ensure appropriate selection of antibacterial therapy.

TABLE 139-3

Common Adverse Reactions to Antibacterial Agents

Antibacterial(s)	Potential Adverse Effects	Comments
β-Lactams	Hypersensitivity reactions	Ranges from rash to anaphylaxis. Cross-reactivity among β-lactams is related to chemical structure and side chain similarity.
	Neurotoxicity	More commonly described with cefepime and imipenem, but likely a class effect. Risk is increased in patients with history of seizures, renal impairment, and advanced age.
	Neutropenia/hematologic reactions	May be related to high doses and prolonged duration
Vancomycin	Nephrotoxicity	Risk increases with vancomycin trough levels >20 μg/mL or concomitant administration with other potentially nephrotoxic agents. The effect is usually reversible.
	"Red man syndrome"	Can be managed with a slower vancomycin infusion and pretreatment with antihistamine
Telavancin	QT prolongation	
	Interference with coagulation tests	May falsely affect INR, PT, aPTT. Perform these tests before the next dose of telavancin (when serum drug levels are at their nadir).
	Taste disturbances	
	Nephrotoxicity	
Oritavancin	Interference with coagulation tests	May falsely affect INR, PT, aPTT. Perform these tests at least 24 h after the dose is administered.
	Gastrointestinal distress	
Dalbavancin	Gastrointestinal distress	
Daptomycin	Myopathy	Monitor CPK levels during therapy. Rhabdomyolysis has been reported but appears to be rare.
	Eosinophilic pneumonia	
Aminoglycosides	Nephrotoxicity	Associated with prolonged use; usually reversible
	Ototoxicity	Can cause both vestibular and cochlear toxicity. Ototoxicity may be irreversible.
Fluoroquinolones	QT_c prolongation	Moxifloxacin appears more likely than other quinolones to exert this effect. Risk of arrhythmia increases when these drugs are given concomitantly with other QT_c-prolonging agents.
	Tendinitis	Risk is greater among the elderly and patients receiving steroids.
	Dysglycemia	
	Exacerbation of myasthenia gravis	
Rifampin	Hepatotoxicity	Risk is greater when drug is given with other antituberculosis agents. When rifampin is given alone, LFT values may be transiently elevated without symptoms.
	Orange discoloration of body fluids	

Antibacterial(s)	Potential Adverse Effects	Comments
Tetracyclines and glycylcyclines	Photosensitivity	
	Gastrointestinal distress	High incidence of diarrhea, nausea, vomiting
Macrolides	Gastrointestinal distress	Erythromycin is occasionally used as a therapeutic agent for some gastric motility disorders.
	QT_c prolongation	Azithromycin use is associated with an increased risk of death from cardiovascular causes among patients at high baseline risk.
Metronidazole	Peripheral neuropathy	Associated with prolonged use
Clindamycin	Diarrhea and pseudomembranous colitis	
Linezolid, tedizolid	Myelosuppression	Associated with prolonged use
	Optic and peripheral neuropathy	Associated with prolonged use
	Lactic acidosis	
TMP-SMX	Hypersensitivity reactions	Allergy usually associated with sulfonamide moiety
	Nephrotoxicity	Associated with high doses
	Hematologic effects	Associated with prolonged use
Nitrofurantoin	Pneumonitis and other pulmonary reactions	Associated with prolonged use
	Peripheral neuropathy	Associated with accumulation of nitrofurantoin in renal failure. Avoid use in renal impairment.
Fosfomycin	Gastrointestinal effects	
Polymyxins	Nephrotoxicity	Associated with high dose
	Neurotoxicity	Neuromuscular blockade and muscle weakness are well described and usually reversible.
Quinupristin-dalfopristin	Arthralgias and myalgias	
Chloramphenicol	Bone marrow suppression	Aplastic anemia or hematopoietic toxicity

Note: All systemic antibiotics have the potential to alter abdominal flora and induce *Clostridium difficile* infection.

Abbreviations: aPTT, activated partial thromboplastin time; CPK, creatine phosphokinase; INR, international normalized ratio; LFT, liver function test; PT, prothrombin time; TMP-SMX, trimethoprim-sulfamethoxazole.

Drug–Drug Interactions

Patients commonly receive other drugs that may interact with antibacterial agents. A summary of the most common drug–drug interactions, by antibacterial class, is provided in **Table 139-4**.

TABLE 139-4

Significant Antibacterial Drug Interactions

Antibacterial(s)	Interacting Agent(s)	Potential Effect and Management
Nafcillin	Warfarin, cyclosporine, tacrolimus	Decreased levels of warfarin, cyclosporine via CYP3A4 induction. Monitor levels of affected drug closely if drugs are given concomitantly.
Ceftriaxone	Calcium-containing IV solutions	Concomitant use is contraindicated in neonates (<28 days); the combination can lead to precipitation of ceftriaxone-calcium particulate. Ceftriaxone and calcium-containing solutions can be given to infants >28 days of age provided they are given sequentially and the lines are thoroughly flushed between infusions.
Carbapenems	Valproic acid	Decreased levels of valproic acid. Monitor valproic acid levels closely if drugs are given concomitantly.
Linezolid, tedizolid	Serotonergic and adrenergic agents (e.g., SSRIs, vasopressors)	Increased levels of serotonergic and adrenergic agents. Monitor for serotonin syndrome. Tedizolid may have less potential than linezolid to cause this drug interaction.
Quinupristin-dalfopristin	Substrates of CYP3A4 (e.g., warfarin, ritonavir, cyclosporine, diazepam, verapamil)	Can result in increased levels of interacting drug
Fluoroquinolones	Theophylline[a]	Can result in theophylline toxicity
	Sucralfate; antacids containing aluminum, calcium, or magnesium; ferrous sulfate– and zinc-containing multivitamins	Can result in subtherapeutic fluoroquinolone levels. Administer fluoroquinolone 2 h before or 6 h after interacting drug.
	Tizanidine[a]	Can result in increased levels of tizanidine and hypotensive, sedative effects. Monitor for side effects if drugs are given concomitantly.
Rifampin	Substrates of CYP3A4 (e.g., warfarin, ritonavir, cyclosporine, diazepam, verapamil, protease inhibitors, voriconazole)	Can result in decreased levels of interacting drug. Avoid concomitant use if possible. If giving drugs concomitantly, monitor drug levels if possible.
	Substrates of CYP2C19 (e.g., omeprazole, lansoprazole)	
	Substrates of CYP2C9 (e.g., warfarin, tolbutamide)	
	Substrates of CYP2C8 (e.g., repaglinide, rosiglitazone)	
	Substrates of CYP2B6 (e.g., efavirenz)	
	Hormone therapy (e.g., norethindrone)	Can result in decreased levels of hormone. If oral contraceptive and rifampin are given concomitantly, use alternative form of birth control.
Tetracyclines	Antacids or drugs containing calcium, magnesium, iron, or aluminum	Can result in decreased absorption of tetracyclines. Administer tetracycline 2 h before or 6 h after interacting drug.
	Warfarin	Increased effect of warfarin. Monitor levels closely if drugs are given concomitantly.
Macrolides[b]	Substrates of CYP3A4 (e.g., warfarin, ritonavir, cyclosporine, diazepam, verapamil)	Avoid concomitant administration if possible.

Antibacterial(s)	Interacting Agent(s)	Potential Effect and Management
	QT_C-prolonging agents (e.g., fluoroquinolones, sotalol)	Increased risk of cardiotoxicity and arrhythmias. Monitor QT_C.
	Protease inhibitors (e.g., ritonavir)	Can result in increased levels of both macrolides and protease inhibitors. Avoid concomitant use if possible.
	Cimetidine	Cimetidine can increase levels of macrolides.
Metronidazole	Ethanol	Can result in disulfiram-like reaction. Ethanol may be present in some formulations of oral drug suspensions (e.g., ritonavir).
	Warfarin	Can increase warfarin levels. Monitor INR closely if drugs are given concomitantly.
TMP-SMX	Warfarin	Increased effect of warfarin. Monitor levels closely if drugs are given concomitantly.
	Phenytoin	Increased levels of phenytoin. Monitor levels closely if drugs are given concomitantly.
	Methotrexate	Increased levels of methotrexate. Monitor levels closely if drugs are given concomitantly.
Oritavancin	Substrates of CYP3A4 (e.g., cyclosporine, warfarin) and CYP2D6 (e.g., aripiprazole) Substrates of CYP2C19 (e.g., omeprazole) and CYP2C9 (e.g., warfarin)	Can result in decreased levels of interacting drug. Avoid concomitant use if possible. If giving drugs concomitantly, monitor drug levels if possible.

[a]Drug reaction described with ciprofloxacin only. [b]Clarithromycin and erythromycin are potent CYP3A4 inhibitors; the probability of a drug interaction with azithromycin is lower.

Abbreviations: INR, international normalized ratio; SSRI, selective serotonin-reuptake inhibitor; TMP-SMX, trimethoprim-sulfamethoxazole.

Exposures

Exposures, both occupational and social, may provide clues to likely pathogens. When relevant, inquiries about exposure to ill contacts, animals, insects, and water should be included in the history, along with sites of residence and travel.

Other Host Factors

Age, renal and hepatic function, and comorbid conditions are all considerations in the choice of and schedule for therapy. Dose adjustments should be made accordingly. In patients with decreased or unreliable oral absorption, IV therapy may be preferred to ensure adequate blood levels of drug and delivery of the antibacterial agent to the site of infection.

DURATION OF THERAPY

Whether empirical or directed, the duration of therapy should be determined in most clinical situations. Guidelines that synthesize available literature and expert opinion provide recommendations on therapy duration that are based on infecting organism, organ system, and patient factors. For example, the American Heart Association has published guidelines endorsed by the Infectious Diseases Society of America (IDSA) on diagnosis, antibacterial therapy, and management of complications of infective endocarditis. Similar guidelines from the IDSA exist for bacterial meningitis, urinary tract infections (including those that are catheter-associated), intraabdominal infections, community- and hospital-acquired pneumonia, skin and soft tissue infections, and other infections.

FAILURE OF THERAPY

If a patient does not respond to therapy, investigations often should include the collection of additional specimens for microbiologic testing and imaging as indicated. Failure to respond can be the result of an antibacterial regimen that does not address the underlying causative organism, the development of

resistance during therapy, or the existence of a focus of infection at a site poorly penetrated by systemic therapy. Some infections may also require surgical interventions (e.g., large abscesses, myonecrosis). Fever due to allergic drug reactions can sometimes complicate assessment of the patient's response to antibacterial treatment.

EXPERT GUIDANCE

Selected websites with the most up-to-date information and guidance for the clinician include the following:

Johns Hopkins ABX Guide (*www.hopkins-abxguide.org*)

IDSA Practice Guidelines (*http://www.idsociety.org/PracticeGuidelines/*)

Center for Disease Dynamics, Economics and Policy Resistance Map (*https://resistancemap.cddep.org/*)

Centers for Disease Control and Prevention Antibiotic/Antimicrobial Resistance (*www.cdc.gov/drugresistance/*)

CLINICAL USE OF ANTIBACTERIAL AGENTS

The clinical application of antibacterial therapy is guided by the spectrum of the agent and the suspected or known target pathogen. Infections for which specific antibacterial agents are among the drugs of choice are listed, along with associated pathogens and susceptibility data, in **Table 139-5**. Resistance rates of specific organisms are dynamic and should be taken into account in the approach to antibacterial therapy. While national resistance rates can serve as a reference, the most useful reference for the clinician is the most recent local laboratory antibiogram, which provides details on local resistance patterns, often on an annual or semiannual basis.

TABLE 139-5

Drug Indications for Specific Infections, Associated Pathogens, and Sample Susceptibility Rates

Antimicrobial(s)	Infections	Common Pathogens (% Susceptible); Resistance as Noted[a]
Penicillin G	Syphilis; yaws; leptospirosis; streptococcal infections; pneumococcal infections; actinomycosis; oral and periodontal infections; meningococcal meningitis and meningococcemia; viridans streptococcal endocarditis; clostridial myonecrosis; tetanus; rat-bite fever; *Pasteurella multocida* infections; erysipeloid (*Erysipelothrix rhusiopathiae*)	*Neisseria meningitidis*; viridans streptococci (69%); *Streptococcus pneumoniae* (96% nonmeningitis; 68% meningitis)
Ampicillin, amoxicillin	Salmonellosis; acute otitis media; *Haemophilus influenzae* meningitis and epiglottitis; *Listeria monocytogenes* meningitis; *Enterococcus faecalis* UTI	*Escherichia coli* (51%); *H. influenzae* (70%); *Salmonella* spp. (85%)
Nafcillin, oxacillin	MSSA bacteremia and endocarditis	*Staphylococcus aureus* (72%); coagulase-negative staphylococci (49%)
Piperacillin-tazobactam	Intraabdominal infections (facultative enteric gram-negative bacilli and obligate anaerobes); infections caused by mixed flora (aspiration pneumonia, diabetic foot ulcers); infections caused by *Pseudomonas aeruginosa*	*P. aeruginosa* (88%)[b]
Cefazolin	*E. coli* UTI; surgical prophylaxis; MSSA bacteremia and endocarditis	*E. coli* (82%)
Cefoxitin, cefotetan	Intraabdominal infections and pelvic inflammatory disease	*Bacteroides fragilis* (60%)[c]
Ceftriaxone	Gonococcal infections; pneumococcal meningitis; viridans streptococcal endocarditis; salmonellosis and typhoid fever; hospital-acquired infections caused by nonpseudomonal facultative gram-negative enteric bacilli	*S. pneumoniae* (93%);[d] *E. coli* (91%); *Klebsiella pneumoniae* (89%)
Ceftazidime, cefepime	Hospital-acquired infections caused by facultative gram-negative bacilli and *Pseudomonas* spp.	*P. aeruginosa* (90%)
Ceftaroline	CAP caused by *S. pneumoniae*, MSSA, *H. influenzae, K. pneumoniae, Klebsiella oxytoca*, and *E. coli*; acute bacterial skin and skin-structure *in*fections caused by MSSA, MRSA, *Streptococcus pyogenes, Streptococcus agalactiae, E. coli, K. pneumoniae*, and *Klebsiella oxytoca*	Mostly susceptible; four strains of MRSA with ceftaroline MICs >4 μg/mL reported in isolates from a single Greek hospital[e]

Antimicrobial(s)	Infections	Common Pathogens (% Susceptible); Resistance as Noted[a]
Ceftazidime-avibactam, meropenem-vaborbactam	Complicated UTIs (ceftazidime-avibactam and meropenem-vaborbactam) and complicated intraabdominal infections (ceftazidime-avibactam in combination with metronidazole) caused by resistant gram-negative organisms, including *Pseudomonas*, and some anaerobes	*P. aeruginosa* (84–97%)[f] MDR Enterobacteriaceae, including carbapenem-resistant Enterobacteriaceae that produce KPCs No activity against metallo-β-lactamases (e.g., NDM)
Ceftolozane-tazobactam	Complicated UTIs and complicated intraabdominal infections (in combination with metronidazole) caused by resistant gram-negative organisms, including *Pseudomonas*, and some anaerobes	*P. aeruginosa* (>86% overall; 60–80% of ceftazidime- and meropenem-resistant strains)[f] MDR Enterobacteriaceae No activity against KPC-producing organisms
Imipenem, meropenem	Intraabdominal infections, infections caused by *Enterobacter* spp. and ESBL-producing gram-negative bacilli	*P. aeruginosa* (84%); *Acinetobacter calcoaceticus-baumannii* complex (93%) (meropenem susceptibilities reported)
Ertapenem	CAP; complicated UTIs, including pyelonephritis; acute pelvic infections; complicated intraabdominal infections; complicated skin and skin-structure infections, excluding diabetic foot infections accompanied by osteomyelitis or caused by *P. aeruginosa*	*Enterobacter cloacae* (88%); *K. pneumoniae* (99%)
Aztreonam	HAIs caused by facultative gram-negative bacilli and *Pseudomonas* in penicillin-allergic patients	*P. aeruginosa* (74%)
Vancomycin	Bacteremia, endocarditis, and other invasive disease caused by MRSA; pneumococcal meningitis; oral formulation for CDAD	*S. aureus* (100%); *E. faecalis* (96%); *E. faecium* (33%)
Telavancin	Hospital- and ventilator-associated pneumonia or skin and soft tissue infections caused by MRSA	*S. aureus*: none reported
Dalbavancin, oritavancin	Complicated skin and soft tissue infections	*S. aureus*: rarely reported for dalbavancin,[g] none reported for oritavancin

Antimicrobial(s)	Infections	Common Pathogens (% Susceptible); Resistance as Noted[a]
Daptomycin	VRE infections; MRSA bacteremia	*E. faecalis* (99.9%);[h] *E. faecium* (99.7%);[h] *S. aureus* (99.9%)[g]
Gentamicin, tobramycin, amikacin	Combined with penicillin for staphylococcal, enterococcal, or streptococcal endocarditis; combined with β-lactam for gram-negative bacteremia; pyelonephritis	*E. coli* (gentamicin, 90%); *P. aeruginosa* (amikacin, 91%; gentamicin, 87%); *A. calcoaceticus-baumannii* complex (gentamicin, 94%)
Azithromycin, clarithromycin, erythromycin	*Legionella, Campylobacter*, and *Mycoplasma* infections; CAP; GAS pharyngitis in penicillin-allergic patients; bacillary angiomatosis; gastric infections due to *Helicobacter pylori*; MAI infections	*S. pneumoniae* (60%); group A streptococci (82%); *H. pylori* (75%)i
Clindamycin	Severe, invasive GAS infections (with β-lactam); infections caused by obligate anaerobes; infections caused by susceptible staphylococci	*S. aureus* (69%)
Doxycycline, minocycline	Acute bacterial exacerbations of chronic bronchitis; granuloma inguinale; brucellosis (with streptomycin); tularemia; glanders; melioidosis; spirochetal infections caused by *Borrelia* (Lyme disease and relapsing fever; doxycycline); infections caused by *Vibrio vulnificus*; some *Aeromonas* infections; infections due to *Stenotrophomonas* (minocycline); plague; ehrlichiosis; chlamydial infections (doxycycline); granulomatous infections due to *Mycobacterium marinum* (minocycline); rickettsial infections; mild CAP; skin and soft tissue infections caused by gram-positive cocci (e.g., CA-MRSA infections); leptospirosis; syphilis; and actinomycosis in the penicillin-allergic patient	*S. pneumoniae* (68%); *S. aureus* (94%)
Tigecycline	CAP caused by *S. pneumoniae, H. influenzae*, or *Legionella pneumophila*; complicated skin infections caused by *E. coli*, MRSA, MSSA, *S. pyogenes, Streptococcus anginosus, S. agalactiae, B. fragilis*; complicated intraabdominal infections caused by *E. coli*, vancomycin-susceptible *E. faecalis, Citrobacter freundii, Enterobacter cloacae, K. pneumoniae, K. oxytoca, Bacteroides* spp., *Clostridium perfringens*, and *Peptostreptococcus* spp.	Mostly susceptible, although case reports of resistance in *A. baumannii* and *K. pneumoniae*
TMP-SMX	Community-acquired UTI; CA-MRSA skin and soft tissue infections	*E. coli* (71%); *S. aureus* (95%)
Sulfonamides	Nocardial infections; leprosy (dapsone); toxoplasmosis (sulfadiazine)	Unknown
Ciprofloxacin, levofloxacin, moxifloxacin, delafloxacin	CAP (levofloxacin and moxifloxacin); UTI; bacterial gastroenteritis; hospital-acquired gram-negative enteric infections; *Pseudomonas* infections (ciprofloxacin and levofloxacin); skin and skin-structure infections (delafloxacin)	*S. pneumoniae* (99%); *E. coli* (80%); *P. aeruginosa* (ciprofloxacin, 77%; levofloxacin, 77%); *Salmonella* spp. (ciprofloxacin, 88%; levofloxacin, 98%)
Rifampin	Staphylococcal foreign body infections (in combination with other antistaphylococcal agents); *Legionella* pneumonia; *Mycobacterium tuberculosis*; atypical nontuberculous mycobacterial infection; pneumococcal meningitis when organisms are susceptible or response is delayed	*S. aureus* (99%), although staphylococci rapidly develop resistance with monotherapy

Antimicrobial(s)	Infections	Common Pathogens (% Susceptible); Resistance as Noted[a]
Metronidazole	Obligate anaerobic gram-negative bacteria (e.g., *Bacteroides* spp); abscess in lung, brain, or abdomen; bacterial vaginosis; CDAD	Mostly susceptible; resistance very rare
Linezolid, tedizolid	VRE; uncomplicated and complicated skin and soft tissue infections caused by MSSA and MRSA; CAP with concurrent bacteremia; hospital-acquired pneumonia	Mostly susceptible; resistance occasionally seen in VRE
Chloramphenicol	HAI due to gram-positive and gram-negative organisms resistant to standard alternatives (e.g., *Burkholderia*)	Unknown
Colistin	HAI due to gram-negative bacilli resistant to all other chemotherapy (e.g., *P. aeruginosa, Acinetobacter* spp., and *Stenotrophomonas maltophilia*)	*P. aeruginosa* (case reports, outbreaks)
Quinupristin-dalfopristin	VRE; complicated skin and skin-structure infections due to MSSA and *S. pyogenes*	*E. faecalis* (<20%);[j] *E. faecium* (>90%)j
Mupirocin	Topical application to nares for *S. aureus* decolonization	*S. aureus* (74–100%)[k]
Nitrofurantoin	UTI caused by most gram-negative bacilli and some gram-positive organisms; prophylaxis in recurrent cystitis	*E. coli* (95%); *E. faecalis* (99%)
Fosfomycin	UTI caused by most gram-negative bacilli and some gram-positive organisms; prophylaxis in recurrent cystitis	Unknown

[a]Unless otherwise noted, susceptibility rates are based on isolates from the Massachusetts General Hospital Clinical Microbiology Laboratory collected between January and December 2012. Local rates will vary. [b]Center for Disease Dynamics, Economics and Policy Resistance Map, Washington, DC. [c]JA Karlowsky et al: Antimicrob Agents Chemother 56:1247, 2012. [d]GV Doern et al: Clin Infect Dis 41:139, 2005. [e]RE Mendes et al: J Antimicrob Chemother 67:1321, 2012. [f]D Van Duin, RA Bonomo: Clin Infect Dis 63:234, 2016. [g]SP McCurdy et al: Antimicrob Agents Chemother 59:5007, 2015. [h]HS Sader et al: J Chemother 23:200, 2011. [i]J Torres et al: J Clin Microbiol 39:2677, 2001. [j]WS Oh et al: Antimicrob Agents Chemother 49:5176, 2005. [k]AE Simor et al: Antimicrob Agents Chemother 51:3880, 2007.

Abbreviations: CA-MRSA, community-acquired MRSA; CAP, community-acquired pneumonia; CA-UTI, community-acquired UTI; CDAD, *Clostridium difficile*–associated diarrhea; ESBL, extended-spectrum β-lactamase; GAS, group A streptococcal; HAI, hospital-acquired infection; KPCs, *Klebsiella pneumoniae* carbapenemases; MAI, *Mycobacterium avium-intracellulare*; MDR, multidrug-resistant; MIC, minimal inhibitory concentration; MRSA, methicillin-resistant *S. aureus*; MSSA, methicillin-susceptible *S. aureus*; NDM, New Delhi metallo-β-lactamase; TMP-SMX, trimethoprim-sulfamethoxazole; UTI, urinary tract infection; VRE, vancomycin-resistant *Enterococcus.*

β-LACTAMS

The β-lactam class of antibiotics consists of penicillins, cephalosporins, carbapenems, and monobactams. The term *β-lactam* reflects the drugs' four-membered lactam ring, which is their core structure. The differing side chains among the agents of this family determine the spectrum of activity. All β-lactams exert a bactericidal effect by inhibiting bacterial cell-wall synthesis. The β-lactams are classified as time-dependent killing agents; therefore, their clinical efficacy is best correlated with the proportion of the dosing interval during which drug levels remain above the MIC for the pathogenic organism.

Penicillins and β-Lactamase Inhibitors

Penicillin, the first β-lactam, was discovered in 1928 by Alexander Fleming. Natural penicillins, such as penicillin G, are active against non-β-lactamase-producing gram-positive and gram-negative bacteria, anaerobes, and some gram-negative cocci. Penicillin G is used for penicillin-susceptible streptococcal infections, pneumococcal and meningococcal meningitis, enterococcal endocarditis, and syphilis. The antistaphylococcal penicillins, which have potent activity against methicillin-susceptible *Staphylococcus aureus* (MSSA), include nafcillin, oxacillin, dicloxacillin, and flucloxacillin. Aminopenicillins, such as ampicillin and amoxicillin, provide added coverage beyond penicillin against gram-negative cocci, such as *Haemophilus influenzae*, and some Enterobacteriaceae, including *Escherichia coli*, *Proteus mirabilis*, *Salmonella*, and *Shigella.* The aminopenicillins are hydrolyzed by many common β-lactamases. These drugs are commonly used for infections caused by susceptible enterococcal and streptococcal species. IV ampicillin is commonly used in meningitis and endocarditis. Oral amoxicillin may be an option for otitis media, respiratory tract infections, and urinary tract infections. The antipseudomonal penicillins include ticarcillin and piperacillin. These penicillin groups generally offer adequate anaerobic coverage; the exceptions

are *Bacteroides* species (such as *Bacteroides fragilis*), which produce β-lactamases and are generally resistant. The rising prevalence of β-lactamase-producing bacteria has led to the increased use of β-lactam–β-lactamase inhibitor combinations, such as ampicillin-sulbactam, amoxicillin-clavulanate, ticarcillin-clavulanate, piperacillin-tazobactam, ceftolozane-tazobactam, ceftazidime-avibactam, and meropenem-vaborbactam. The β-lactamase inhibitors themselves do not have antibacterial activity (with the exception of sulbactam, which has activity against *Acinetobacter baumannii*) but typically inhibit the *S. aureus* class A β-lactamase, β-lactamases of *H. influenzae* and *Bacteroides* species, and a number of plasmid-encoded β-lactamases. These combination agents are typically used when broader-spectrum coverage is needed—e.g., in pneumonia and intraabdominal infections. Piperacillin-tazobactam is a useful agent for broad coverage in febrile neutropenic patients. Avibactam and vaborbactam inhibit a broader spectrum of β-lactamases than the other inhibitors, including extended-spectrum β-lactamases (ESBLs), AmpC β-lactamases, and some carbapenemases (**see Chap. 140**).

Cephalosporins

The cephalosporin drug class encompasses several generations determined by spectrum of antibacterial activity. The first generation (cefazolin, cefadroxil, and cephalexin) largely has activity against gram-positive bacteria, with some additional activity against *E. coli*, *P. mirabilis*, and *Klebsiella pneumoniae*. First-generation cephalosporins are commonly used for infections caused by MSSA and streptococci (e.g., skin and soft tissue infections). Cefazolin is a popular choice for surgical prophylaxis against skin organisms. The second generation (cefamandole, cefuroxime, cefaclor, cefprozil, cefuroxime axetil, cefoxitin, and cefotetan) has additional activity against *H. influenzae* and *Moraxella catarrhalis*. Cefoxitin and cefotetan have potent activity against anaerobes as well. Second-generation cephalosporins are used to treat community-acquired pneumonia because of their activity against *S. pneumoniae*, *H. influenzae*, and *M. catarrhalis*. They are also used for other mild or moderate infections, such as acute otitis media and sinusitis. The third-generation cephalosporins are characterized by greater potency against gram-negative bacilli and reduced potency against gram-positive cocci. These cephalosporins, which include cefoperazone, cefotaxime, ceftazidime, ceftriaxone, cefdinir, cefixime, and cefpodoxime, are used for infections caused by Enterobacteriaceae, although resistance is an increasing concern. Ceftriaxone penetrates the CSF and can be used to treat meningitis caused by *H. influenzae*, *N. meningitidis*, and susceptible strains of *S. pneumoniae*. It is also used for the treatment of later-stage Lyme disease, gonococcal infections, and streptococcal endocarditis. It is noteworthy that ceftazidime is the only third-generation cephalosporin with activity against *Pseudomonas aeruginosa* but lacks activity against gram-positive bacteria. This drug is frequently used for pulmonary infections in cystic fibrosis, postneurosurgical meningitis, and febrile neutropenia. The fourth generation of cephalosporins includes cefepime and cefpirome, broad-coverage agents with potent activity against both gram-negative bacilli, including *P. aeruginosa*, and gram-positive cocci. The fourth generation has clinical applications similar to those of the third generation and may offer additional activity over the first, second, and third generations in the presence of certain β-lactamases. These agents can be used in bacteremia, febrile neutropenia, and intraabdominal and urinary tract infections. Ceftaroline, a fifth-generation cephalosporin, differs from the other cephalosporins in its added activity against MRSA, which is resistant to all other β-lactams. Ceftaroline's gram-negative activity is similar to that of the third-generation cephalosporins but does not include *P. aeruginosa*. Ceftaroline may be used in community-acquired pneumonia and skin infections, and emerging data support its use in more severe infections such as bacteremia. Adverse reactions to ceftaroline have included hypersensitivity reactions and neutropenia. Ceftolozane-tazobactam and ceftazidime-avibactam are novel cephalosporin–β-lactamase inhibitor combinations with activity against gram-negative bacteria, including *Pseudomonas*, and some anaerobes. Both agents have been studied in complicated intraabdominal infections and complicated urinary tract infections. Ceftolozane-tazobactam is thought to be stable against many ESBL-producing organisms because of the tazobactam component. The addition of avibactam to ceftazidime yields a combination agent with activity against AmpC-, ESBL-, and KPC-producing organisms. These cephalosporin–β-lactamase inhibitor combinations may be of clinical benefit in multidrug-resistant gram-negative infections.

Carbapenems

Carbapenems, including doripenem, imipenem, meropenem, and ertapenem, offer the most reliable coverage for strains containing ESBLs. All carbapenems have broad activity against gram-positive cocci, gram-negative bacilli, and anaerobes. None is active against methicillin-resistant *S. aureus* (MRSA), but all are active against MSSA, *Streptococcus* species, and Enterobacteriaceae. Ertapenem is the only carbapenem that has poor activity against *P. aeruginosa* and *Acinetobacter*. Imipenem is active against penicillin-susceptible *Enterococcus faecalis* but not *Enterococcus faecium*. Carbapenems are not active against Enterobacteriaceae containing carbapenemases. *Stenotrophomonas maltophilia* and some *Bacillus* species are intrinsically resistant to carbapenems because of a zinc-dependent carbapenemase. Addition of vaborbactam to meropenem results in inhibition of AmpC β-lactamases, ESBLs, and *K. pneumoniae* carbapenemases (KPCs).

Monobactams

Aztreonam is the sole monobactam. Its activity is limited to gram-negative bacteria and includes *P. aeruginosa* and most other Enterobacteriaceae. This drug is inactivated by ESBLs and carbapenemases. The principal use for aztreonam is as an alternative to penicillins, cephalosporins, or carbapenems in patients with a serious β-lactam allergy. Aztreonam is structurally related to ceftazidime and should be used cautiously in individuals with a serious ceftazidime allergy. It is commonly used in febrile neutropenia and intraabdominal infections.

Adverse Reactions to β-Lactam Drugs

Agents within the β-lactam class are known for several adverse effects. Gastrointestinal side effects, mainly diarrhea, are common, but hypersensitivity reactions constitute the most common adverse effect of β-lactams. The reactions' severity can range from rash to anaphylaxis, but the rate of true anaphylactic reactions is only 0.05%. An individual with an accelerated IgE-mediated reaction to one β-lactam agent may still receive another agent within

the class, but caution should be used in choosing a β-lactam that has a dissimilar side chain and a low level of cross-reactivity. For example, the second-, third-, and fourth-generation cephalosporins and the carbapenems display very low cross-reactivity in patients with penicillin allergy. Aztreonam is the only β-lactam that has no cross-reactivity with the penicillin group. In cases of severe allergy, desensitization (a graded challenge) to the indicated β-lactam, with close monitoring, may be warranted if other antibacterial options are not suitable.

β-Lactams can rarely cause serum sickness, Stevens-Johnson syndrome, nephropathy, hematologic reactions, and neurotoxicity. Neutropenia appears to be related to high doses or prolonged use. Neutropenia and interstitial nephritis caused by β-lactams generally resolve upon discontinuation of the agent. Imipenem and cefepime are associated with an increased risk of seizure, but this risk is likely a class effect and related to high doses or doses that are not adjusted in renal impairment.

GLYCOPEPTIDES AND LIPOGLYCOPEPTIDES

Vancomycin is a glycopeptide antibiotic with activity against staphylococci (including MRSA and coagulase-negative staphylococci), streptococci (including *S. pneumoniae*), and enterococci. It is not active against gram-negative organisms. Vancomycin also displays activity against *Bacillus* species, *Corynebacterium jeikeium*, *Listeria monocytogenes*, and gram-positive anaerobes such as *Peptostreptococcus*, *Actinomyces*, *Clostridium*, and *Propionibacterium* species. Vancomycin has several important clinical uses. It is used for serious infections caused by MRSA, including health care–associated pneumonia, bacteremia, osteomyelitis, and endocarditis. It is also commonly used for skin and soft tissue infections. Oral vancomycin is not absorbed systemically and is reserved for the treatment of *Clostridium difficile* infection. Vancomycin is also an alternative for the treatment of infections caused by MSSA in patients who cannot tolerate β-lactams. Resistance to vancomycin is a rising concern. Strains of vancomycin-intermediate *S. aureus* (VISA) and vancomycin-resistant enterococci (VRE) are not uncommon. Vancomycin appears to be a concentration-dependent killer, with the AUC/MIC ratio being the best predictor of efficacy (Fig. 139-2). Guidelines recommend targeting a vancomycin trough level of 15–20 μg/mL in MRSA infections in order to maintain an AUC/MIC ratio >400. When using vancomycin, clinicians should monitor for nephrotoxicity. The risk increases when trough levels are >20 μg/mL. Concomitant therapy with other nephrotoxic agents, such as aminoglycosides, also increases the risk of nephrotoxicity. Ototoxicity was reported with early formulations of vancomycin but is currently uncommon because purer formulations are available. Both of these adverse effects are reversible upon discontinuation of vancomycin. Clinicians should be aware of the "red man syndrome," a common reaction that presents as a rapid onset of erythematous rash or pruritus on the head, face, neck, and upper trunk. This reaction is caused by histamine release from basophils and mast cells and can be treated with diphenhydramine and slowing of the vancomycin infusion.

Telavancin, dalbavancin, and oritavancin are structurally similar to vancomycin and are referred to as *lipoglycopeptides*. They have antibacterial activity against *S. aureus* (including MRSA and some strains of VISA and vancomycin-resistant *S. aureus* [VRSA]), streptococci, and enterococci. Oritavancin may have activity against some strains of VRE. These lipoglycopeptide agents also provide coverage against anaerobic gram-positive organisms except for *Lactobacillus* and some *Clostridium* species. The clinical efficacy of telavancin has been demonstrated in both skin and soft tissue infections and nosocomial pneumonia, and the efficacy of dalbavancin and oritavancin has been shown in skin and soft tissue infections. The vancomycin resistance phenotype may reduce the potency of all three lipoglycopeptides, but the rate of resistance to these drugs among *S. aureus* and enterococcal isolates has been low. Adverse effects of telavancin include nephrotoxicity, metallic taste, and gastrointestinal side effects. Clinicians should be aware of the potential for electrocardiographic QT_C prolongation that can increase the risk of cardiac arrhythmias when telavancin is used concomitantly with other QT_C-prolonging agents. Telavancin may interfere with certain coagulation tests (e.g., causing false elevations in prothrombin time). Dalbavancin and oritavancin have safety profiles similar to that of vancomycin, with common effects reported as headache and gastrointestinal side effects. These glycolipopeptides should be used cautiously in patients with hypersensitivity reactions to vancomycin, as cross-allergy may be possible.

LIPOPEPTIDES

Daptomycin is a lipopeptide antibiotic with activity against a broad range of gram-positive organisms. This drug is active against staphylococci (including MRSA and coagulase-negative staphylococci), streptococci, and enterococci. Daptomycin remains active against enterococci that are resistant to vancomycin. In addition, it exhibits activity against *Bacillus*, *Corynebacterium*, *Peptostreptococcus*, and *Clostridium* species. Daptomycin's pharmacodynamic parameter for efficacy is concentration-dependent killing. Resistance to daptomycin is rare, but MICs may be higher for VISA strains. Daptomycin can be used in skin and soft tissue infections, bacteremia, endocarditis, and osteomyelitis. It is an important alternative for MRSA and other gram-positive infections when bactericidal therapy is needed and vancomycin cannot be used. Daptomycin is generally well tolerated, and its main toxicity consists of elevation of creatine phosphokinase (CPK) levels and myopathy. CPK should be monitored during daptomycin treatment, and the drug should be discontinued if muscular toxicities occur. There have also been case reports of reversible eosinophilic pneumonia associated with daptomycin use.

AMINOGLYCOSIDES

The aminoglycosides are a class of antibacterial agents with concentration-dependent activity against most gram-negative organisms. The most commonly used aminoglycosides are gentamicin, tobramycin, and amikacin, although others, such as streptomycin, kanamycin, neomycin, and paromomycin, may be used in special circumstances. Aminoglycosides have a significant dose-dependent postantibiotic effect; i.e., they have an antibacterial effect even after serum drug levels are undetectable. The postantibiotic effect and concentration-dependent killing form the rationale behind extended-interval aminoglycoside dosing, in which a larger dose is given once daily rather than smaller doses multiple times daily. Aminoglycosides are active against gram-negative bacilli, such as Enterobacteriaceae, *P. aeruginosa*, and *Acinetobacter*. They also enhance the activity of cell wall–active agents

such as β-lactams or vancomycin against some gram-positive bacteria, including staphylococci and enterococci. This combination therapy is termed *synergistic* because the effect of both agents provides a killing effect greater than would be predicted from the effects of either agent alone. Amikacin and streptomycin have activity against *Mycobacterium tuberculosis*, and amikacin has activity against *Mycobacterium avium-intracellulare*. The aminoglycosides do not have activity against anaerobes, *S. maltophilia*, or *Burkholderia cepacia*. Aminoglycosides are used in clinical practice in a variety of infections caused by gram-negative organisms, including bacteremia and urinary tract infections. They are frequently used alone or in combination for the treatment of *P. aeruginosa* infection. When used in combination with a cell wall–active agent, gentamicin and streptomycin are also important for the treatment of gram-positive bacterial endocarditis. All aminoglycosides can cause nephrotoxicity and ototoxicity. The risk of nephrotoxicity is not well defined; however, some studies have indicated that the effect may be related to the duration of therapy as well as to the concomitant use of other nephrotoxic agents. Nephrotoxicity is usually reversible, but ototoxicity can be irreversible.

MACROLIDES AND KETOLIDES

The macrolides (azithromycin, clarithromycin, and erythromycin) and ketolides (telithromycin) are classes of antibiotics that inhibit protein synthesis. Compared with erythromycin (the older antibiotic), azithromycin and clarithromycin have better oral absorption and tolerability. Azithromycin, clarithromycin, and telithromycin all have broader spectra of activity than erythromycin, which is less frequently used. These agents are commonly used in the treatment of upper and lower respiratory tract infections caused by *S. pneumoniae*, *H. influenzae*, *M. catarrhalis*, and atypical organisms (e.g., *Chlamydia pneumoniae*, *Legionella pneumophila*, and *Mycoplasma pneumoniae*); group A streptococcal pharyngitis in penicillin-allergic patients; and nontuberculous mycobacterial infections (e.g., caused by *Mycobacterium marinum* and *Mycobacterium chelonae*) as well as in the prophylaxis and treatment of *M. avium-intracellulare* infection in patients with HIV/AIDS and in combination therapy for *Helicobacter pylori* infection and bartonellosis. Enterobacteriaceae, *Pseudomonas* species, and *Acinetobacter* species are intrinsically resistant to macrolides as a result of decreased membrane permeability, although azithromycin is active against gram-negative diarrheal pathogens. The major adverse effects of this drug class include nausea, vomiting, diarrhea and abdominal pain, prolongation of QT_C interval, exacerbation of myasthenia gravis, and tinnitus. Azithromycin specifically has been associated with an increased risk of death, especially among patients with underlying heart disease, because of the risk of QT_C interval prolongation and torsades de pointes. Erythromycin, clarithromycin, and telithromycin inhibit the CYP3A4 hepatic drug-metabolizing enzyme and can result in increased levels of coadministered drugs, including benzodiazepines, statins, warfarin, cyclosporine, and tacrolimus. Azithromycin does not inhibit CYP3A4 and therefore does not interact with these drugs.

CLINDAMYCIN

Clindamycin is a lincosamide antibiotic and is bacteriostatic against some organisms and bactericidal against others. It is used most often to treat bacterial infections caused by anaerobes (e.g., *B. fragilis*, *Clostridium perfringens*, *Fusobacterium* species, *Prevotella melaninogenicus*, and *Peptostreptococcus* species) and susceptible staphylococci and streptococci. Clindamycin is used for treatment of dental infections, anaerobic lung abscess, and skin and soft tissue infections. It is used together with bactericidal agents (penicillins or vancomycin) to inhibit new toxin synthesis in the treatment of streptococcal or staphylococcal toxic shock syndrome. Other uses include treatment of infections caused by *Capnocytophaga canimorsus*, combination therapy for malaria and babesiosis, and therapy for toxoplasmosis. Clindamycin has excellent oral bioavailability. Adverse effects include nausea, vomiting, diarrhea, *C. difficile*–associated diarrhea and pseudomembranous colitis, maculopapular rash, and (rarely) Stevens-Johnson syndrome.

TETRACYCLINES AND GLYCYLCYCLINES

The tetracyclines (doxycycline, minocycline, and tetracycline) and the glycylcyclines (tigecycline) inhibit protein synthesis and are bacteriostatic. These drugs have wide clinical uses. They are used in the treatment of skin and soft tissue infections caused by gram-positive cocci (including MRSA), spirochetal infections (e.g., Lyme disease, syphilis, leptospirosis, and relapsing fever), rickettsial infections (e.g., Rocky Mountain spotted fever), atypical pneumonia, sexually transmitted infections (e.g., *Chlamydia trachomatis* infection, lymphogranuloma venereum, and granuloma inguinale), infections with *Nocardia* and *Actinomyces*, brucellosis, tularemia, Whipple's disease, and malaria. Tigecycline, the only approved agent in the glycylcycline class, is a derivative of minocycline and is indicated in the treatment of complicated skin and soft tissue infections, complicated intraabdominal infections, and community-acquired bacterial pneumonia in adults. Tigecycline has activity against MRSA, vancomycin-sensitive enterococci, many Enterobacteriaceae, and *Bacteroides* species; it has no activity against *P. aeruginosa*. This drug has been used in combination with colistin for the treatment of serious infections with multidrug-resistant gram-negative organisms. A pooled analysis of 13 clinical trials found an increased risk of death and treatment failure among patients given tigecycline alone; as a result, the U.S. Food and Drug Administration mandated a black box warning. Tetracyclines have reduced absorption when orally coadministered with calcium- and iron-containing compounds, including milk, and doses should be spaced at least 2 h apart. The major adverse reactions to both of these classes are nausea, vomiting, diarrhea, and photosensitivity. Tetracyclines have been associated with fetal bone-growth abnormalities and should be avoided during pregnancy and in the treatment of children <8 years old.

TRIMETHOPRIM-SULFAMETHOXAZOLE

Trimethoprim-sulfamethoxazole (TMP-SMX) is an antibiotic with two components that inhibit folate synthesis and produce antibacterial activity. TMP-SMX is active against gram-positive bacteria such as staphylococci and streptococci; however, its use against MRSA is usually limited to community-acquired infections, and its activity against *Streptococcus pyogenes* may not be reliable. This drug is also active against many gram-negative bacteria, including *H.*

influenzae, *E. coli*, *P. mirabilis*, *Neisseria gonorrhoeae*, and *S. maltophilia*. TMP-SMX is not active against anaerobes or *P. aeruginosa*. It has many uses because of its wide spectrum of activity and high oral bioavailability. Urinary tract infections, skin and soft tissue infections, and respiratory tract infections are among the common uses. Another important indication is for both prophylaxis and treatment of *Pneumocystis jirovecii* infections in immunocompromised patients. Resistance to TMP-SMX has limited its use against many Enterobacteriaceae. Resistance rates among urinary isolates of *E. coli* are almost 25% in the United States. The most common adverse reactions associated with TMP-SMX are gastrointestinal effects such as nausea, vomiting, and diarrhea. In addition, rash is a common allergic reaction and may preclude the subsequent use of other sulfonamides. With prolonged use, leukopenia, thrombocytopenia, and granulocytopenia can develop. TMP-SMX can also cause nephrotoxicity, hyperkalemia, and hyponatremia, which are more common at high doses. TMP-SMX has several important interactions with other drugs (Table 139-4), including warfarin, phenytoin, and methotrexate.

FLUOROQUINOLONES

The fluoroquinolones include norfloxacin, ciprofloxacin, ofloxacin, levofloxacin, moxifloxacin, gemifloxacin, and delafloxacin. Ciprofloxacin and levofloxacin have the broadest spectrum of activity against gram-negative bacteria, including *P. aeruginosa* (similar to that of third-generation cephalosporins). Because of the risk of selection of resistance during fluoroquinolone treatment of serious pseudomonal infections, these agents are usually used in combination with an antipseudomonal β-lactam. Levofloxacin, moxifloxacin, gemifloxacin, and delafloxacin have additional gram-positive activity, including that against *S. pneumoniae* and some strains of MSSA, and, with the exception of delafloxacin, these agents are used for treatment of community-acquired pneumonia. Strains of MRSA are commonly resistant to all fluoroquinolones except delafloxacin. Moxifloxacin is used as one component of second-line regimens for multidrug-resistant tuberculosis. Fluoroquinolones exhibit concentration-dependent killing, are well absorbed orally, and have elimination half-lives that usually support once- or twice-daily dosing. Oral coadministration with compounds containing high concentrations of aluminum, magnesium, or calcium can reduce fluoroquinolone absorption. The penetration of fluoroquinolones into prostate tissue supports their use for bacterial prostatitis. Fluoroquinolones are generally well tolerated but can cause central nervous system (CNS) stimulatory effects, including seizures; peripheral neuropathy; glucose dysregulation; and tendinopathy associated with Achilles tendon rupture, particularly in older patients, organ transplant recipients, and patients taking glucocorticoids. Other potential effects on connective tissues include an association with increased risk of aortic aneurysm. Worsening of myasthenia gravis also has been associated with quinolone use. Moxifloxacin causes modest prolongation of the QT_C interval and should be used with caution in patients receiving other QT_C-prolonging drugs.

RIFAMYCINS

The rifamycins include rifampin, rifabutin, and rifapentine. Rifampin is the most commonly used rifamycin. For almost all therapeutic indications, it is used in combination with other agents to reduce the likelihood of selection of high-level rifampin resistance. Rifampin is used foremost in the treatment of mycobacterial infections—specifically, as a mainstay of combination therapy for *M. tuberculosis* infection or as a single agent in the treatment of latent *M. tuberculosis* infection. In addition, it is often used in the treatment of nontuberculous mycobacterial infection. Rifampin is used in combination regimens for the treatment of staphylococcal infections, particularly prosthetic-valve endocarditis and bone infections with retained hardware. It is a component of combination therapy for brucellosis (with doxycycline) and leprosy (with dapsone for tuberculoid leprosy and with dapsone and clofazimine for lepromatous disease). Rifampin can be used alone for prophylaxis in close contacts of patients with *H. influenzae* or *N. meningitidis* meningitis. The drug has high oral bioavailability, which is further enhanced when it is taken on an empty stomach. Rifampin has several adverse effects, including elevated aminotransferase levels (14%), rash (1–5%), and gastrointestinal events such as nausea, vomiting, and diarrhea (1–2%). Its many clinically relevant interactions with other drugs (Table 139-4) mandate the clinician's careful review of the patient's medications before rifampin initiation to assess safety and the need for additional monitoring, including monitoring of drug levels.

METRONIDAZOLE

Metronidazole is used in the treatment of anaerobic bacterial infections as well as infections caused by protozoa (e.g., amebiasis, giardiasis, trichomoniasis). It is the agent of choice as a component of combination therapy for polymicrobial abscesses in the lung, brain, or abdomen, the etiology of which often includes anaerobic bacteria, and for bacterial vaginosis, pelvic inflammatory disease, and anaerobic infections, such as those due to *Bacteroides*, *Fusobacterium*, and *Prevotella* species. This drug is an alternative agent for treatment of mild to moderate *C. difficile*–associated diarrhea. Metronidazole is bactericidal against anaerobic bacteria and exhibits concentration-dependent killing. It has high oral bioavailability and tissue penetration, including penetration of the blood–brain barrier. The majority of *Actinomyces*, *Propionibacterium*, and *Lactobacillus* species are intrinsically resistant to metronidazole. The major adverse effects include nausea, diarrhea, and a metallic taste. Concomitant ingestion of alcohol may result in a disulfiram-like reaction, and patients are usually instructed to avoid alcohol during treatment. Long-term treatment carries the risk of leukopenia, neutropenia, peripheral neuropathy, and CNS toxicity manifesting as confusion, dysarthria, ataxia, nystagmus, and ophthalmoparesis. Through metronidazole's effect on the CYP2C9 drug-metabolizing enzyme, its coadministration with warfarin can result in decreased metabolism and enhanced anticoagulant effects that require close monitoring. Concomitant administration of metronidazole with lithium can result in increased serum levels of lithium and associated toxicity; coadministration with phenytoin can result in phenytoin toxicity and possibly decreased levels of metronidazole.

OXAZOLIDINONES

Linezolid is a bacteriostatic agent and is indicated for serious infections due to resistant gram-positive bacteria, such as MRSA and VRE. The intrinsic resistance of gram-negative bacteria is mediated primarily by endogenous efflux pumps. Linezolid has excellent oral bioavailability. Adverse effects include myelosuppression and ocular and peripheral neuropathy with prolonged therapy. Peripheral neuropathy may be irreversible. Linezolid is a weak, reversible monoamine oxidase inhibitor, and coadministration with sympathomimetics and foods rich in tyramine should be avoided. Linezolid has been associated with serotonin syndrome when coadministered with selective serotonin-reuptake inhibitors. Tedizolid has properties similar to those of linezolid, but with lower dosing it may be less likely to cause adverse hematologic and neuropathic effects.

NITROFURANTOIN

Nitrofurantoin's antibacterial activity results from the drug's conversion to highly reactive intermediates that can damage DNA and other macromolecules. Nitrofurantoin is bactericidal, and its action is concentration dependent. It displays activity against a range of gram-positive bacteria, including *S. aureus*, *Staphylococcus epidermidis*, *Staphylococcus saprophyticus*, *E. faecalis*, *Streptococcus agalactiae*, group D streptococci, viridans streptococci, and corynebacteria, as well as gram-negative organisms, including *E. coli* and *Enterobacter*, *Neisseria*, *Salmonella*, and *Shigella* species. Nitrofurantoin is used primarily in the treatment of urinary tract infections and is preferred in the treatment of such infections in pregnancy. It may be used for the prevention of recurrent cystitis. Recently, there has been interest in the use of nitrofurantoin for treatment of urinary tract infections caused by ESBL-producing Enterobacteriaceae such as *E. coli*, although resistance has been growing in Latin America and parts of Europe. Coadministration with magnesium should be avoided because of decreased absorption, and patients should be encouraged to take the drug with food to increase its bioavailability and decrease the risk of adverse effects, which include nausea, vomiting, and diarrhea. Nitrofurantoin may also cause pulmonary fibrosis and drug-induced hepatitis. Because the risk of adverse reactions increases with age, the use of nitrofurantoin in elderly patients is not recommended. Patients with glucose-6-phosphate dehydrogenase (G6PD) deficiency are at elevated risk for nitrofurantoin-associated hemolytic anemia.

POLYMYXINS

Colistin and polymyxin B act by disrupting bacterial cell membrane integrity and are active against the nonenteric pathogens *P. aeruginosa* and *A. baumannii* but not against *Burkholderia*. These drugs also exhibit activity against many Enterobacteriaceae, with the exceptions of *Proteus*, *Providencia*, and *Serratia* species. They lack activity against gram-positive bacteria. Polymyxins are bactericidal and are available in IV formulations. Colistimethate is converted to the active form (colistin) in plasma. Polymyxins are most often used for infections due to pathogens resistant to multiple other antibacterial agents, including urinary tract infections, hospital-acquired pneumonia, and bloodstream infections. Nebulized formulations have been used for adjunctive treatment of refractory ventilator-associated pneumonia. The most important adverse effect is dose-dependent reversible nephrotoxicity. Neurotoxicity, including paresthesias, muscle weakness, and confusion, is reversible and less common than nephrotoxicity.

QUINUPRISTIN-DALFOPRISTIN

Quinupristin-dalfopristin is a member of the streptogramin class of antibiotics and kills bacteria by inhibiting protein synthesis. The antibacterial spectrum of quinupristin-dalfopristin includes staphylococci (including MRSA), streptococci, and *E. faecium* (but not *E. faecalis*). This drug is also active against *Corynebacterium* species and *L. monocytogenes*. Quinupristin-dalfopristin is not reliably active against gram-negative organisms. It exhibits concentration-dependent killing, with an AUC/MIC ratio predicting efficacy. The clinical use of quinupristin-dalfopristin is largely for infections due to vancomycin-resistant *E. faecium* and other gram-positive bacterial infections. The drug has demonstrated efficacy in a variety of infections, including urinary tract infections, bone and joint infections, and bacteremia. Adverse effects associated with quinupristin-dalfopristin include infusion-related reactions, arthralgias, and myalgias. The arthralgias and myalgias may be severe enough to warrant drug discontinuation. Quinupristin-dalfopristin inhibits the CYP3A4 drug-metabolizing enzyme, with consequent drug interactions (Table 139-4).

FOSFOMYCIN

Fosfomycin is a phosphonic acid antibiotic that has greater activity in acidic environments and is excreted in its active form in the urine. Thus, its use is primarily for prophylaxis and treatment of uncomplicated cystitis and should be avoided if there is concern about pyelonephritis. The drug is administered as a single 3-g dose that results in high urine concentrations for up to 48 h. Fosfomycin is active against *S. aureus*, vancomycin-susceptible enterococci and VRE, and a wide range of gram-negative organisms, including *E. coli*, *Enterobacter* species, *Serratia marcescens*, *P. aeruginosa*, and *K. pneumoniae*. Notably, the vast majority of ESBL-producing Enterobacteriaceae are susceptible to fosfomycin. *A. baumannii* and *Burkholderia* species are resistant. The emergence of resistance to fosfomycin has not been observed during treatment of cystitis but has been documented during treatment of respiratory tract infections and osteomyelitis. The few adverse effects that have been reported include nausea and diarrhea.

CHLORAMPHENICOL

The use of chloramphenicol is limited by its potentially serious toxicities. When other agents are contraindicated or ineffective, chloramphenicol represents an alternative treatment for infections, including meningitis caused by susceptible bacteria such as *N. meningitidis*, *H. influenzae*, and *S. pneumoniae*. It has also been used for the treatment of anthrax, brucellosis, *Burkholderia* infections, chlamydial infections, clostridial infections, erlichiosis, rickettsial infections, and typhoid fever. Adverse reactions include aplastic anemia, myelosuppression, and gray baby syndrome. Chloramphenicol inhibits the CYP2C19 and CYP3A4 drug-metabolizing enzymes and consequently increases levels of many classes of drugs.

APPROACH TO PROPHYLAXIS OF INFECTION

Antibacterial prophylaxis is indicated only in selected circumstances **(Table 139-6)** and should be supported by well-designed studies or expert panel recommendations. In all cases, the risk or severity of the infection to be prevented should be greater than the adverse consequences of antibacterial therapy, including the potential for selection of resistance. In addition, the timing and duration of antibacterial treatment should be targeted for maximal effect and minimal required exposure. Prophylaxis of surgical-site infections targets bacteria that may contaminate the wound during the surgical procedure, including the skin flora of the patient or operating team and the air in the operating room. Delivery of the antibacterial drug within 1 h before the surgical incision is most effective. For prolonged procedures, redosing may be necessary to maintain effective blood and tissue levels until the wound is closed. Additional dosing is not recommended after the incision is closed. In patients with nasal carriage of *S. aureus*, preoperative decolonization with nasal mupirocin reduces the rate of *S. aureus* surgical-site infections and is generally recommended for high-risk procedures such as cardiac surgery and orthopedic implantation of prosthetic devices. For dental procedures, preprocedure antibacterial drugs are given to prevent transient bacteremia and the seeding of certain high-risk cardiac lesions. Prophylaxis is also used in nonprocedural settings in certain patients who have recurrent infections or who are at risk of serious infection from a specific exposure (e.g., close contact with a patient with meningococcal meningitis). Extension of prophylaxis beyond the period of infection risk (24 h in the case of surgical procedures) does not add further benefit and may increase the risk of resistance selection or *C. difficile* disease.

TABLE 139-6

Prophylaxis of Bacterial Infections in Adults

Condition	Antibacterial Agents[a]	Timing or Duration of Prophylaxis
Surgical		
Clean (cardiac, thoracic, neurologic, orthopedic, vascular, plastic)	Cefazolin (vancomycin,[b] clindamycin)	1 h before incision; re-dose with long procedures
Clean (ophthalmic)	Topical neomycin–polymyxin B–gramicidin, topical moxifloxacin	Every 5–15 min for 5 doses immediately prior to procedure
Clean-contaminated (head and neck)	Cefazolin + metronidazole, ampicillin-sulbactam[c] (clindamycin)	1 h before incision; re-dose with long procedures
Clean-contaminated (hysterectomy, gastroduodenal, biliary, unobstructed small intestine, urologic)	Cefazolin, ampicillin-sulbactam[c] (clindamycin + aminoglycoside, aztreonam, or fluoroquinolone)	1 h before incision; re-dose with long procedures
Clean-contaminated (colorectal, appendectomy)	Cefazolin + metronidazole, ampicillin-sulbactam,[c] ertapenem (clindamycin + aminoglycoside, aztreonam, or fluoroquinolone)	1 h before incision; re-dose with long procedures
Dirty (ruptured viscus)	Therapeutic regimen directed at anaerobes and gram-negative bacteria (e.g., ceftriaxone + metronidazole)	1 h before incision; re-dose with long procedures; continue for 3–5 days after procedure
Dirty (traumatic wound)	Therapeutic regimen: cefazolin (clindamycin ± aminoglycoside, aztreonam, or fluoroquinolone)	1 h before incision; re-dose with long procedures; continue for 3–5 days after procedure
Nonsurgical		
Dental, oral, or upper respiratory procedures in patients with high-risk cardiac lesions (prosthetic valves, congenital heart defects, prior endocarditis)	Amoxicillin PO, ampicillin IM (clindamycin PO, IV)	Oral agents 1 h before procedure; injection 30 min before procedure
Recurrent *S. aureus* skin infections[d]	Mupirocin[e]	Intranasal application for 5 days
Recurrent cellulitis associated with lymphatic disruption[d]	Benzathine penicillin IM monthly, oral penicillin or erythromycin twice daily	Undefined
Recurrent cystitis in women[d]	Nitrofurantoin, TMP-SMX, fluoroquinolone	After sexual intercourse *or* 3 times weekly for up to 1 year
Bite wounds	Amoxicillin-clavulanate (doxycycline, moxifloxacin)	3–5 days
Recurrent spontaneous bacterial peritonitis in cirrhotic patients[d]	Fluoroquinolone[f]	Undefined
Recurrent pneumococcal meningitis in patient with CSF leak or humoral immune defect[d]	Penicillin	Undefined
Exposure to patient with meningococcal meningitis	Rifampin, ciprofloxacin	2 days (rifampin), single dose (ciprofloxacin)

Condition	Antibacterial Agents[a]	Timing or Duration of Prophylaxis
High-risk neutropenia (ANC, ≤100/μL for >7 days)[d]	Levofloxacin or ciprofloxacin[f]	Until neutropenia resolves or fever dictates use of other antibacterials

[a]Regimens in parentheses are alternatives for patients allergic to β-lactams. [b]Vancomycin may be given together with cefazolin to patients known to be colonized with methicillin-resistant *Staphylococcus aureus*. [c]Cefoxitin or cefotetan may also be considered. [d]Not considered routine for all patients, but an acceptable consideration among alternative approaches. [e]Usually coupled with bathing with chlorhexidine-containing skin antiseptic. [f]Choice of fluoroquinolone prophylaxis must be balanced against the risk of selection of resistance.

Abbreviations: ANC, absolute neutrophil count; CSF, cerebrospinal fluid; TMP-SMX, trimethoprim-sulfamethoxazole.

ANTIMICROBIAL STEWARDSHIP

In an era of increasing prevalence of multidrug-resistant bacteria and with a substantial amount of inappropriate antimicrobial use, the need for rational antimicrobial prescribing has never been greater **(Chap. 140)**. *Antimicrobial stewardship* describes the practice of promoting the selection of the appropriate drug, dosage, route, and duration of antimicrobial therapy. Antimicrobial stewardship programs implement a variety of strategies to (1) improve patient care through appropriate antimicrobial use; (2) decrease the development of resistance within patients and populations; (3) reduce the incidence of adverse effects; and (4) control costs.

Infections caused by resistant pathogens result in significant morbidity and mortality as well as increased health care costs. Antimicrobial stewardship programs are typically multidisciplinary and often include infectious disease physicians, clinical pharmacists (usually with special training in infectious disease), clinical microbiologists, information systems specialists, infection prevention and control practitioners, and epidemiologists. These teams employ a variety of approaches to achieving the program's goals.

Established strategies of antimicrobial stewardship programs include (1) prospective audit of antimicrobial use, with intervention and feedback; (2) formulary restriction; and (3) preauthorization. *Prospective audit and feedback* are usually undertaken by an infectious disease physician or a pharmacist. In this process, orders for broad-spectrum antimicrobials (e.g., carbapenems) or agents for which more cost-effective alternatives may exist (e.g., daptomycin, ceftazidime-avibactam) are reviewed on a regular basis for appropriateness. In circumstances in which an antimicrobial is used in the absence of an appropriate indication, the stewardship program team intervenes and recommends an alternative to the primary team caring for the patient. This process has been successful in several quasi-experimental studies, resulting in declines in use of broad-spectrum drugs and decreases in adverse events, such as *C. difficile* infection. *Formulary restriction* is the inclusion of a limited set of antimicrobial agents in a hospital formulary for the purpose of limiting indiscriminant use of antimicrobials in the absence of demonstrated benefit. Such restriction coincidentally serves to reduce costs. *Preauthorization* is the practice of requiring clinicians to obtain approval before using selected antimicrobials. Approval may be provided electronically with sophisticated Computerized Provider Order Entry (CPOE) software, after specific criteria for use are met, or after communication with an infectious disease specialist as designated by the stewardship program. These strategies have led to a decrease in *C. difficile* infections and to improvements in drug susceptibility patterns.

Additional strategies used in specific health care settings are guidelines and pathways, dose optimization, parenteral-to-oral conversion, antibiotic time-out, and de-escalation of therapy. Documentation of the indication for which each antimicrobial is prescribed is also encouraged. Antimicrobial stewardship is an evolving area and an increasingly active area of research aimed at identifying the best practices. The IDSA, in collaboration with several other professional organizations, has published guidelines for developing institutional antimicrobial stewardship programs (*www.idsociety.org/Antimicrobial_Agents/*).

FURTHER READING

Barlam TF et al: Implementing an antibiotic stewardship program: Guidelines by the Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America. Clin Infect Dis 62:e51, 2016. [PubMed: 27080992]

Bratzler DW et al: Clinical practice guidelines for antimicrobial prophylaxis in surgery. Surg Infect (Larchmt) 14:73, 2013.

Craig WA: Pharmacokinetic/pharmacodynamic parameters: Rationale for antibacterial dosing of mice and men. Clin Infect Dis 26:1, 1998.

Grayson ML et al (eds): *Kucers' The Use of Antibiotics. A Clinical Review of Antibacterial, Antifungal, Antiparasitic and Antiviral Drugs*, 6th ed. London, Hodder Arnold, 2010.

Infectious Diseases Society of America: Practice guidelines by organ system. Available at *http://www.idsociety.org/Organ_System/*. Accessed February 21, 2017.

Jeffries MN et al: Consequences of avoiding β-lactams in patients with β-lactam allergies. J Allergy Clin Immunol 137:1148, 2016.

Labreche MJ et al: Recent updates on the role of pharmacokinetics–pharmacodynamics in antimicrobial susceptibility testing as applied to clinical practice. Clin Infect Dis 61:1446, 2015.

Thomas Z et al: A multicenter evaluation of prolonged empiric antibiotic therapy in adult ICUs in the United States. Crit Care Med 43:2527, 2015.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 140: Bacterial Resistance to Antimicrobial Agents

David C. Hooper

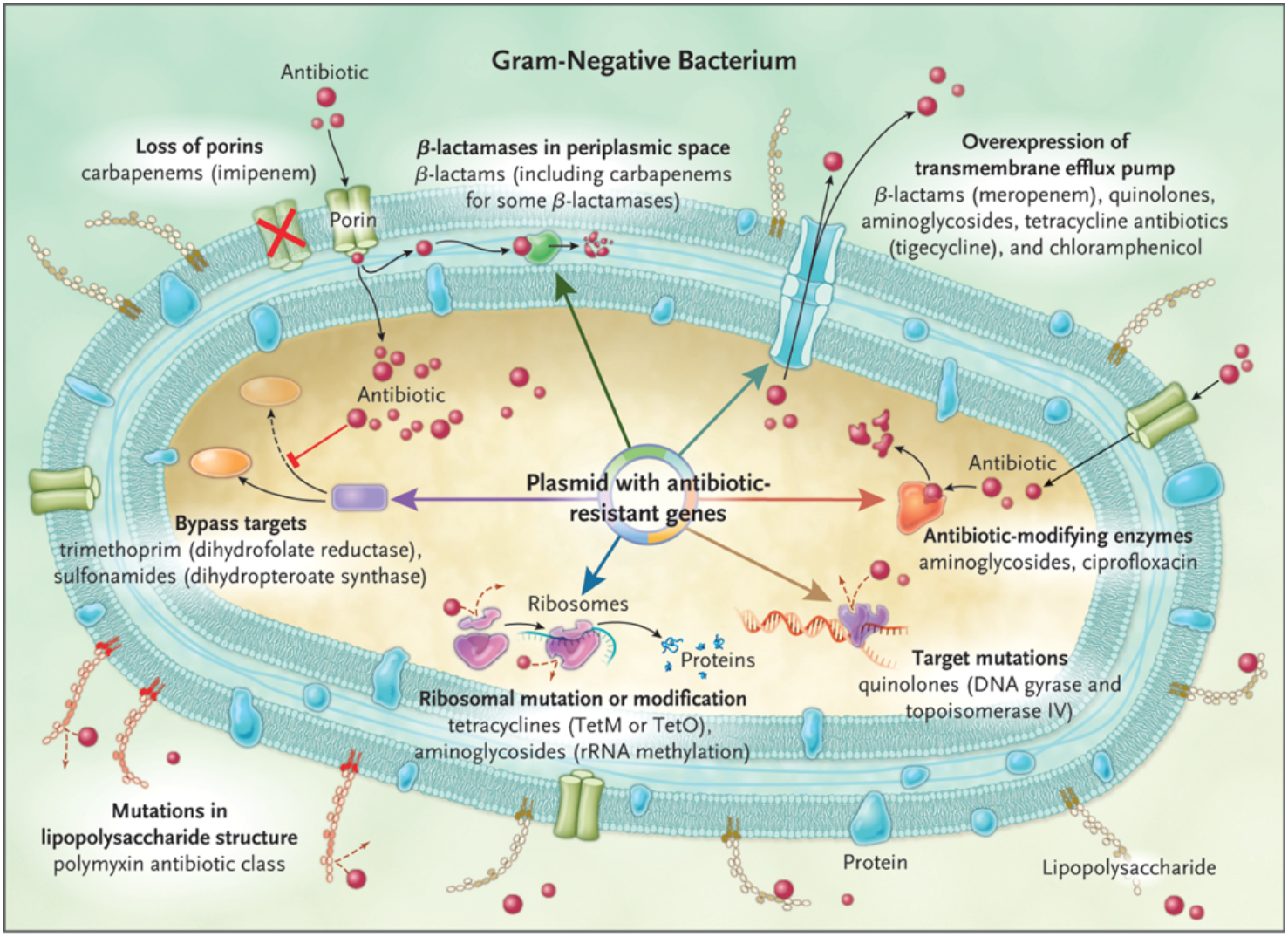
FIGURE 140-1

DEFINITION OF RESISTANCE

The action of antimicrobial agents on a range of targets within the bacterial cell can result in inhibition of bacterial growth or in killing of the bacterial cell (Chap. 139). Reduction in or loss of an agent's antibacterial effect is referred to as *resistance*, and the properties of or alterations in the bacterium that result in reduced antimicrobial activity are termed *resistance mechanisms*. Bacteria can be resistant to single or multiple antimicrobials, as detailed in the sections that follow. The occurrence and magnitude of resistance are often assessed in clinical microbiology laboratories by measurement of the lowest drug concentration that inhibits growth of a bacterium (minimal inhibitory concentration, or MIC) with a standardized inoculum and growth conditions. MIC values are generally interpreted as representing bacterial susceptibility, intermediate susceptibility, or resistance; the interpretation is based on correlations of the MIC values with the pharmacokinetics and delivery of a drug to the site of infection in the body as well as with data from clinical trials. Thus, a clinical laboratory result of "susceptible" for a bacterium predicts a likely clinical response to an appropriately dosed antimicrobial drug by a patient infected with that organism, whereas a result of "resistant" predicts poor or no clinical response to that drug. Breakpoint MIC values for categorization of bacteria as susceptible, intermediate, or resistant are generally developed by regulatory and advisory groups and are often based on the distribution of MIC values from a large collection of recent clinical bacterial isolates. Research studies on the mechanisms and epidemiology of resistance may in some cases use different and less rigid definitions of resistance based on determination of a reproducible increase in an MIC value relative to a baseline reference MIC, independent of clinical breakpoints.

FIGURE 140-1

Mechanisms of resistance to antibacterial agents, as illustrated in a gram-negative bacterium. Similar mechanisms are found in gram-positive bacteria, but their lack of an outer membrane causes β-lactamases to be excreted outside the cell, rather than into the periplasmic space between the inner and outer membranes, and reduces the efficiency of efflux pumps because exported drugs can re-enter the cell after crossing a single membrane, rather than the two membranes in gram-negative bacteria. Red spheres indicate antibiotics. *(From Peleg AY, Hooper DC: Hospital-acquired infections due to gram-negative bacteria. N Engl J Med 362:1084, 2010. © 2010 Massachusetts Medical Society. Reprinted with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MECHANISMS OF RESISTANCE

Bacteria use a wide variety of mechanisms to block or circumvent the activity of antibacterial agents (Table 140-1 and Fig. 140-1). Although myriad, these mechanisms can generally be grouped into three categories: (1) alteration or bypassing of targets that exhibit reduced binding of the drug, (2) altered access of the drug to its target by reductions in uptake or increases in active efflux, and (3) a modification of the drug that reduces its activity. These mechanisms result from either mutations in bacterial chromosomal genes occurring spontaneously during bacterial DNA replication or the acquisition of new genes by DNA transfer from other bacteria or uptake of exogenous DNA. New genes are most often acquired on self-replicating plasmids or other DNA elements transferred from other bacteria. However, some bacteria, such as *Streptococcus pneumoniae* and *Neisseria gonorrhoeae*, can also take up fragments of environmental DNA from related bacterial species and recombine that DNA directly into their own chromosomes, a process called *transformation*. Not uncommonly, resistant bacteria have combinations of resistance mechanisms either within one category or among categories, and many plasmids contain more than one resistance gene. Thus, plasmid acquisition itself can in many cases confer resistance to multiple antibacterial agents. Resistance to multiple, structurally unrelated antibiotics can also occur by mutations that cause increased expression of certain bacterial efflux pumps, some of which have broad substrate profiles.

TABLE 140-1

The Most Common Mechanisms of Resistance to Antibacterial Agents

Antibacterial Agent(S)	Major Target	Mechanism(S) of Action	Mechanism(S) of Resistance
β-Lactams (penicillins, cephalosporins, monobactams, carbapenems)	Cell wall synthesis	Bind cell-wall cross-linking enzymes (PBPs, transpeptidases)	1. Drug inactivation by β-lactamases 2. Altered PBP targets 3. Reduced diffusion through porin channels
Glycopeptides and lipoglycopeptides (vancomycin, teicoplanin, telavancin, dalbavancin, oritavancin)	Cell wall synthesis	Block cell wall glycosyltransferases by binding D-Ala-D-Ala stem-peptide terminus Teicoplanin, telavancin, dalbavancin, and oritavancin: affect membrane function	1. Altered D-Ala-D-Ala target (D-Ala-D-Lac) 2. Increased D-Ala-D-Ala target binding at sites distant from cell wall synthesis enzymes
Bacitracin	Cell wall synthesis	Blocks lipid carrier of cell wall precursors	Active drug efflux
Fosfomycin	Cell wall synthesis	Blocks linkage of stem peptide to NAG by enoyltransferase	1. Target enzyme overexpression 2. Drug-modifying enzymes
Aminoglycosides (gentamicin, tobramycin, amikacin)	Protein synthesis	Bind 30S ribosomal subunit Block translocation of peptide chain Cause misreading of mRNA	1. Drug-modifying enzymes 2. Methylation at ribosome binding site 3. Decreased permeation to target due to active efflux
Tetracyclines (tetracycline, doxycycline, minocycline)	Protein synthesis	Bind 30S ribosomal subunit Inhibit peptide elongation	1. Active drug efflux 2. Ribosomal protection proteins
Tigecycline	Protein synthesis	Same as tetracyclines	Active drug efflux (pumps different from those affecting tetracyclines)
Macrolides (erythromycin, clarithromycin, azithromycin) and the ketolide telithromycin	Protein synthesis	Bind 50S ribosomal subunit Block peptide chain exit	1. Methylation at ribosome binding site 2. Active drug efflux
Lincosamides (clindamycin)	Protein synthesis	Bind 50S ribosomal subunit Block peptide bond formation	Methylation at ribosome binding site
Streptogramins (quinupristin, dalfopristin)	Protein synthesis	Same as macrolides	1. Same as macrolides 2. Drug-modifying enzymes
Chloramphenicol	Protein synthesis	Binds 50S ribosomal subunit Blocks aminoacyl tRNA positioning	Drug-modifying enzymes
Oxazolidinones (linezolid, tedizolid)	Protein synthesis	Bind 50S ribosomal subunit Inhibit initiation of peptide synthesis	1. Altered rRNA binding site 2. Methylation of ribosome binding site

Antibacterial Agent(S)	Major Target	Mechanism(S) of Action	Mechanism(S) of Resistance
Mupirocin	Protein synthesis	Blocks isoleucyl tRNA synthetase	1. Acquired resistant tRNA synthetase (drug bypass) 2. Altered native tRNA synthetase target
Sulfonamides (sulfadiazine, sulfisoxazole, and sulfamethoxazole)	Folate synthesis	Inhibit dihydropteroate synthetase	Acquired resistant dihydropteroate synthetase (drug bypass)
Trimethoprim	Folate synthesis	Inhibits dihydrofolate reductase	Acquired resistant dihydrofolate reductase (drug bypass)
Quinolones (norfloxacin, ciprofloxacin, ofloxacin, levofloxacin, moxifloxacin, gemifloxacin, delafloxacin)	DNA synthesis	Inhibit DNA gyrase and DNA topoisomerase IV Enzyme–DNA–drug complex: blocks DNA replication apparatus	1. Altered target(s) 2. Active efflux 3. Protection of target from drug 4. Drug-modifying enzyme (ciprofloxacin)
Rifamycins (rifampin, rifabutin, rifapentine)	RNA synthesis	Inhibit RNA polymerase	Altered target
Nitrofurantoin	Nucleic acid synthesis	Reduces reactive drug derivatives that damage DNA	Altered drug-activating enzymes
Metronidazole	Nucleic acid synthesis	Reduces reactive drug derivatives that damage DNA	1. Altered drug-activating enzyme 2. Acquired detoxifying enzymes 3. Active efflux
Polymyxins (polymyxin B and polymyxin E [colistin])	Cell membrane	Bind LPS and disrupt both outer and cytoplasmic membranes	Altered cell-membrane charge with reduced drug binding
Daptomycin	Cell membrane	Produces membrane channel and membrane leakage	Altered cell-membrane charge with reduced drug binding

Abbreviations: LPS, lipopolysaccharide; NAG, *N*-acetylglucosamine; PBP, penicillin-binding protein.

Many antibacterial drugs are derived from natural products of environmental microbial species. Some genes encoding resistance to these drugs originate in the drug-producer organism to protect it from its product and have then been mobilized onto plasmids that spread into other organisms. Surviving non-producer bacteria in the exposed natural environment may also have evolved resistance under selection pressure that adds to the reservoir of resistance mechanisms. Exposure to antibacterial agents either in nature or during human or animal use then results in the selection of resistant strains within an otherwise susceptible bacterial population. In some cases, resistance mechanisms may confer disadvantages that render bacterial growth or survival fitness inferior to that of susceptible strains. In a number of examples, however, fitness defects are often mitigated over time by compensatory mutational mechanisms that make the bacteria both resistant and fit and thereby more likely to persist in a reservoir even in the absence of continued antimicrobial selection pressures. Discussed below are the major classes of antimicrobial agents currently in clinical use and the most important mechanisms of resistance encountered in clinical infections.

β-Lactams

β-lactams, the largest class of antibiotics, inhibit bacterial cell-wall synthesis by binding to cell wall transpeptidases, cross-linking enzymes that are also called penicillin-binding proteins (PBPs); PBPs are targets that are unique to bacteria and have no mammalian counterpart. The most common mechanism of resistance to β-lactams, particularly in gram-negative bacteria, is their degradation by β-lactamases, enzymes that break down the core β-lactam ring and destroy drug activity. β-Lactamases differ in the spectrum of β-lactams they can degrade. Some β-lactamases are encoded on the bacterial chromosome, and their activity contributes to the susceptibility profile of a particular species. Chromosomally encoded β-lactamases can be produced in varying amounts that affect the degree of resistance. In some cases, enzyme expression is physiologically induced by exposure to certain β-

lactams; in other cases, enzyme expression can become constant or constitutive through mutations in genes that encode the regulators of expression of a β-lactamase gene. Other β-lactamases are encoded by genes on acquired plasmids and are usually constitutively expressed. The resistance profiles due to plasmids may be present in some strains of a species but not others, depending on which plasmids the strain has acquired. In gram-positive bacteria β-lactamases are secreted into the extracellular environment, whereas in gram-negative bacteria these enzymes are secreted into the periplasmic space between the cytoplasmic and outer membranes—a limited space that permits the presence of high concentrations of β-lactamase. Thus, in gram-negative bacteria, access of β-lactams both to their target PBPs and to β-lactamases requires diffusion across the outer membrane, generally through the porin diffusion channels. Reductions in outer-membrane diffusion channels due to mutation can further augment the efficiency of β-lactamase degradation of β-lactams: slow diffusion acts together with the high enzyme concentrations in the periplasmic space to enhance drug degradation and resistance.

Most strains of *Staphylococcus aureus* produce a plasmid-encoded β-lactamase that degrades penicillin but not semisynthetic penicillins, such as oxacillin and nafcillin. The greatest diversity among β-lactamases, however, is found in gram-negative bacteria. The most common and earliest identified plasmid-encoded β-lactamases of gram-negative bacteria can inactivate all penicillins and most early-generation cephalosporins. Multiple extended-spectrum β-lactamase (ESBL) variants of these early enzymes have emerged and are now widely disseminated. These ESBLs can degrade later-generation cephalosporins (ceftriaxone, cefotaxime, ceftazidime) as well as the monobactam aztreonam, and some ESBLs also degrade the fourth-generation cephalosporin cefepime. Carbapenems (imipenem, meropenem, ertapenem, doripenem) generally are not degraded by ESBLs, but additional β-lactamases, called *carbapenemases*, which degrade carbapenems and most if not all other β-lactams, have emerged and are increasing in prevalence. In the United States, *Klebsiella pneumoniae* carbapenemases (KPCs), which are usually found in strains of *Escherichia coli* and *K. pneumoniae*, are most widespread, but New Delhi metallo-β-lactamases (NDM carbapenemases), which were found initially on the Indian subcontinent, have now appeared in several areas in the United States, as has an OXA group carbapenemase, OXA-48. In some cases, high levels of expression of an ESBL or an AmpC enzyme (see below), together with reduced porin diffusion channels, can also result in resistance to carbapenems. In *Pseudomonas aeruginosa*, resistance to carbapenems can occur by mutations that cause reductions in the OprD diffusion channel for imipenem or increased expression of efflux pumps that can remove meropenem from the bacterial cell.

The chromosomal β-lactamase of *K. pneumoniae* preferentially degrades penicillins but not cephalosporins. In contrast, the chromosomal β-lactamase of *Enterobacter* and related genera, AmpC, can degrade almost all cephalosporins but is normally expressed in only small amounts. Mutations in regulatory genes that cause increased amounts of AmpC to be produced confer full resistance to penicillins and cephalosporins; the exceptions are cefoxitin and cefepime, which are relatively stable to AmpC. Resistance to cefepime can develop, however, through the combined effects of increased AmpC production and decreased porin diffusion channels. Genes encoding AmpC have also been found on plasmids but are less common than plasmid-encoded ESBLs.

Inhibitors of β-lactamases such as clavulanate, sulbactam, tazobactam, avibactam, and vaborbactam have been developed and paired with amoxicillin and ticarcillin (clavulanate), ampicillin (sulbactam), piperacillin and ceftolozane (tazobactam), ceftazidime (avibactam), or meropenem (vaborbactam). These inhibitors have little or no antibacterial activity of their own but inhibit plasmid-mediated β-lactamases, including ESBLs. Only avibactam and vaborbactam inhibit AmpC enzymes and some carbapenemases (KPCs but not metallo-carbapenemases, such as NDM).

Resistance to β-lactams also occurs through alterations in the drugs' target transpeptidase enzymes (PBPs) involved in cross-linking of the bacterial cell-wall peptidoglycan structure. In *S. pneumoniae*, *N. gonorrhoeae*, and *Neisseria meningitidis*, resistance to penicillin occurs by recombination of transformed DNA from related species that results in mosaic PBPs with lower affinity for penicillin. A combination of increased expression of an efflux pump and a porin mutation also causes penicillin resistance in *N. gonorrhoeae*. In staphylococci, resistance to methicillin and other β-lactams occurs by acquisition of the *mec* gene, which encodes PBP2a with reduced drug affinity. PBP2a is a bypass target that can function in the presence of β-lactams, bypassing their effect on other PBPs. Ceftaroline is the only β-lactam that has an affinity for PBP2a and is thus active against methicillin-resistant staphylococcal strains. Resistance to ceftaroline can occur, however, by mutations in the gene encoding PBP2a that reduce its affinity for the drug.

Glycopeptides and Lipoglycopeptides

Glycopeptides and lipoglycopeptides inhibit bacterial cell-wall synthesis by binding to the terminal two D-alanine amino acids on the cell-wall peptidoglycan stem peptides, which are involved in peptidoglycan cross-links. In doing so, these drugs block the transpeptidase cross-linking enzymes and glycosyl transferases necessary for cell wall synthesis. Resistance to vancomycin in enterococci is due to the acquisition of a set of *van* genes that result in (1) the production of D-alanine-D-lactate—instead of the normal D-alanine-D-alanine—at the end of the peptidoglycan stem peptide and (2) the reduction of existing D-alanine-D-alanine–terminated peptides. Vancomycin binds D-alanine-D-lactate with a 1000-fold lower affinity than D-alanine-D-alanine. The *van* genes originated in the organisms that naturally produce vancomycin and have been mobilized and reorganized in transposon mobile genetic elements and onto plasmids, which can be transferred between enterococci. In rare cases, the *van* gene cassettes have been transferred from enterococci to *S. aureus*, with the consequent generation of full vancomycin resistance. In *S. aureus*, intermediate resistance to vancomycin is more common than full vancomycin resistance and is due to a different mechanism that results from a series of several chromosomal mutations leading to a thickened and poorly cross-linked cell wall. This modified cell wall contains additional D-alanine-D-alanine–terminated stem peptides that bind vancomycin at a site distant from the cell membrane, adjacent to which new peptidoglycan is synthesized and where vancomycin binding blocks

transpeptidase and transglycosylase enzymes. Thus, vancomycin's binding to these distant termini impedes its access to the proximal binding sites that result in inhibition of peptidoglycan synthesis. This intermediate-resistance phenotype was first recognized in patients receiving prolonged courses of vancomycin that created an opportunity for selection of the multiple mutations needed to produce the modified cell wall. Because of the energy costs of a thickened cell wall, this intermediate-resistance phenotype may be unstable, with strains returning to susceptibility in the absence of vancomycin selection pressure. Susceptibility to telavancin, dalbavancin, and oritavancin is also reduced in strains that exhibit resistance or intermediate susceptibility to vancomycin, although in some cases the drugs remain sufficiently active that the strains may still be classified as susceptible on the basis of standard clinical laboratory interpretive criteria.

Aminoglycosides

Aminoglycosides are one of several classes of antimicrobials that inhibit protein synthesis by binding to either the 30S or the 50S bacterial ribosomal subunit (both of which differ from eukaryotic ribosomal subunits), with consequent selective antibacterial activity. The aminoglycosides bind to the 30S subunit of the bacterial ribosome. The most common mechanism of resistance to aminoglycosides in gram-negative bacteria is due to acquisition of plasmid genes encoding transferase enzymes that modify aminoglycosides by the addition of acetyl, adenyl, or phosphate groups; these added groups decrease the drugs' binding affinity to their ribosomal target site. Transferases differ in which aminoglycosides they modify, and amikacin resistance occurs less often than resistance to gentamicin or tobramycin by these mechanisms. Another recently found mechanism of plasmid-mediated resistance is due to methylase enzymes that can methylate the site of aminoglycoside binding on the 16S ribosomal RNA of the 30S ribosomal subunit and reduce drug binding to its ribosome target, resulting in resistance to all aminoglycosides. For streptomycin, a single ribosomal protein mutation may also cause resistance. In *P. aeruginosa*, resistance can occur through mutations in regulatory genes causing increased expression of a chromosomally encoded efflux pump, MexXY, which reduces intracellular drug concentrations.

Tetracyclines and Glycylcyclines

These antibiotics bind the 16S ribosomal RNA of the 30S ribosomal subunit at a site distinct from the binding site of the aminoglycosides and inhibit bacterial protein synthesis. For tetracyclines, resistance is often plasmid mediated and attributable either to active efflux pumps, which are generally specific for tetracyclines, or to proteins that protect the ribosome from tetracycline action. A number of broad-spectrum, chromosomally encoded efflux pumps may also include tetracyclines among their substrates, and regulatory mutations that cause pump overexpression may confer tetracycline resistance together with resistance to other agents. Resistance to the glycylcycline tigecycline, which is not affected by the usual tetracycline resistance mechanisms, can occur through mutations that cause overexpression of some broad-spectrum efflux pumps, particularly in *Proteus* species. Plasmid-encoded tetracycline modification as a resistance mechanism has been described in *Bacteroides* species but is uncommon.

Macrolides, Ketolides, Lincosamides, and Streptogramins

These antibiotics are also inhibitors of bacterial protein synthesis, in this case through their binding to the 23S RNA of the 50S ribosomal subunit. They are generally active against gram-positive bacteria. Resistance to macrolides, clindamycin, and quinupristin is most often due to acquired Erm methylases that modify the drug-binding site on the ribosome, reducing binding. Resistance to quinupristin by this mechanism renders the quinupristin-dalfopristin combination bacteriostatic rather than bactericidal. Telithromycin, a ketolide structurally related to macrolides, has an additional binding site on the ribosome and remains active in the presence of some methylases. Methylase gene expression can be induced by exposure to most macrolides but generally not ketolides (e.g., telithromycin); however, bacterial strains constitutively expressing methylase genes can display resistance to both macrolides and ketolides. Acquired genes encoding active efflux pumps also can contribute to resistance to macrolides in streptococci and to resistance to macrolides, clindamycin, and dalfopristin in staphylococci. Plasmid-acquired, drug-modifying enzymes in staphylococci can also cause resistance to quinupristin and dalfopristin. Macrolide resistance due to 23S rRNA mutations at the site of drug binding is uncommon in staphylococci and streptococci because of the multiple copies of the rRNA genes on the chromosomes of these species; such resistance may occur more frequently, however, in mycobacteria, *Helicobacter pylori*, and *Treponema* species, which have only one or two chromosomal copies of these rRNA genes. Among gram-negative bacteria, many of which are not susceptible to current macrolides because of inadequate drug permeation, some strains with acquired genes for macrolide-modifying enzymes have been described.

Chloramphenicol

Chloramphenicol inhibits bacterial protein synthesis by binding to the 23S rRNA of the 50S subunit at a site that overlaps the macrolide-binding site. Chloramphenicol is uncommonly used in human medicine because of infrequent but potentially severe bone marrow toxicity. Resistance to chloramphenicol is most often due to plasmid-encoded, drug-modifying acetyltransferases that have been found in both gram-positive and gram-negative bacteria and whose expression can be induced by drug exposure. Among staphylococci, some resistant strains have been found to have a plasmid-encoded ribosomal methylase that confers resistance to chloramphenicol, clindamycin, and oxazolidinones. As is the case for macrolides, ribosomal mutations causing resistance to chloramphenicol are uncommon because of multiple copies of rRNA genes in most human pathogens. Plasmid-encoded efflux pumps affecting chloramphenicol specifically have been found in gram-negative bacteria, and other pumps affecting chloramphenicol and oxazolidinones have been found in gram-positive bacteria.

Oxazolidinones

Linezolid and tedizolid are the only members of the oxazolidinone class of antimicrobials in clinical use, and both are active against gram-positive bacteria only; lack of sufficient activity in gram-negative bacteria results from the ability of native efflux pumps in these bacteria to limit drug access to their cytoplasmic ribosome targets. Oxazolidinones target the bacterial ribosome and inhibit protein synthesis by binding to 23S rRNA of the 50S subunit at a distinct site that overlaps with the chloramphenicol-binding site. Resistance has been seen in enterococci more often than in staphylococci and, in both organisms, is most often due to mutations in multiple copies of the 23S rRNA genes that reduce drug binding to the ribosome. A plasmid-acquired ribosomal methylase gene that enables ribosomal alteration at a site that confers resistance to both linezolid and chloramphenicol has also been found in some strains of both *S. aureus* and coagulase-negative staphylococci but is not yet widespread. A plasmid-encoded active efflux pump conferring resistance to oxazolidinones (both linezolid and tedizolid) and chloramphenicol has been described in animal isolates and a small number of human isolates of *Enterococcus faecalis*.

Mupirocin

Mupirocin is used only in topical formulations, most often for elimination of nasal carriage of *S. aureus*. It targets bacterial leucyl-tRNA synthetase and inhibits protein synthesis. Resistance to mupirocin occurs by either mutation in the target leucyl-tRNA synthetase (low-level resistance) or the acquisition of a plasmid-encoded resistant tRNA synthetase (high-level resistance), which bypasses drug inhibition of the native, sensitive synthetase.

Sulfonamides and Trimethoprim

These agents inhibit the folate biosynthesis pathway at different steps. Sulfonamides are structurally similar to *para*-aminobenzoic acid (PABA) and competitively inhibit dihydropteroate synthetase, which, in an early step in the pathway, uses PABA to synthesize dihydropteroate, a precursor of dihydrofolate. Trimethoprim inhibits dihydrofolate reductase at a later step in the pathway that generates tetrahydrofolate. Clinical use of folate pathway inhibitors most often consists of the combination of sulfamethoxazole and trimethoprim; on occasion, however, trimethoprim or various sulfonamides are used individually. Resistance to both of these antimetabolites can result from mutation in their target enzymes or can be due to plasmid-acquired genes encoding resistant enzymes that bypass the inhibition of the native sensitive enzymes—a resistant dihydropteroate synthetase in the case of sulfonamides and a resistant dihydrofolate reductase in the case of trimethoprim. Resistance to the combination of sulfamethoxazole and trimethoprim requires that the bacterial strain have resistance mechanisms for both agents and yet is not uncommon. Resistance due to drug efflux or drug modification has not been problematic for either sulfonamides or trimethoprim.

Quinolones

Quinolones are synthetic inhibitors of bacterial DNA synthesis. They bind to two enzymes required for DNA synthesis: DNA gyrase and DNA topoisomerase IV. In addition to inhibiting the enzymes' catalytic functions of altering DNA topology, they stabilize enzyme–DNA complexes that form a barrier to the DNA replication machinery and are a precursor to lethal double-strand DNA breaks. Although related topoisomerase enzymes are involved in mammalian DNA synthesis, the mammalian and bacterial enzymes are sufficiently different from each other for quinolones to have selective activity against bacteria. Resistance to quinolones is most often due either to chromosomal mutations altering the target enzymes DNA gyrase and DNA topoisomerase IV, with consequent reduction in drug binding, or to mutations that increase the expression of native broad-spectrum efflux pumps for which quinolones (among other compounds) are substrates. In addition, three types of acquired genes can confer reduced susceptibility or low-level resistance by either protecting the target enzymes, modifying some quinolones (particularly ciprofloxacin and norfloxacin) to reduce their activity, or generating an efflux of quinolones. These genes are usually located on multidrug-resistance plasmids that have spread worldwide. Their presence can promote higher levels of quinolone resistance by enhancing selection of the mutations in chromosomal target genes with exposure to quinolones and can then link quinolone resistance to resistance to other antibacterial drugs that are encoded on the same plasmid.

Rifampin and Rifabutin

Antimicrobials of the rifamycin class target bacterial RNA polymerase and thereby inhibit transcription of messenger RNA and gene expression. Their activity is generally limited to gram-positive bacteria because native efflux pumps in most gram-negative bacteria reduce drug access to the cytoplasmic enzyme target. Single mutations in the β subunit of RNA polymerase constitute the principal mechanism of acquired rifampin resistance, which is high level. Thus, rifampin and other rifamycins are used for treatment of infections only in combination with other antibacterial drugs in order to reduce the likelihood of selection of high-level resistance.

Metronidazole

Metronidazole is actively taken up by most anaerobic bacteria and then converted to reactive drug derivatives that nonspecifically damage cytoplasmic proteins and nucleic acids. Thus, metronidazole lacks a specific cellular target. Acquired resistance to metronidazole in *Bacteroides* species is rare. Such resistance has been reported in strains that lack the endogenous activating nitroreductase or that have acquired *nim* genes responsible for further

reduction of DNA-damaging nitroso intermediates to an inactive derivative. Active efflux and enhanced DNA repair mechanisms also have been associated with resistance.

Nitrofurantoin

Nitrofurantoin is used only for treatment of lower urinary tract infections because adequate drug concentrations are found only in urine. Its mechanism of action is not fully understood but is thought to involve generation of reactive derivative molecules (as occurs with metronidazole) that damage DNA and ribosomes. Resistance to nitrofurantoin in *E. coli* can emerge through a series of mutations that progressively decrease the nitroreductase activity required for generating active nitrofuran metabolites. These mutants are also impaired in growth; this impairment possibly explains the infrequent occurrence of resistance with clinical use of nitrofurantoin.

Polymyxins

Because of emerging multidrug resistance in gram-negative bacteria, colistin and polymyxin B are being used increasingly for infections due to resistant Enterobacteriaceae, *P. aeruginosa*, and *Acinetobacter* species. Polymyxins are cationic cyclic peptide molecules that bind negatively charged lipopolysaccharides on the gram-negative bacterial outer membrane, with subsequent disruption and permeabilization of both outer-membrane and cytoplasmic-membrane structure. Thus, the polymyxins are bactericidal. Resistance is so far uncommon but can emerge during therapy through mutations that cause reductions in the negative charge of the gram-negative bacterial cell surface, thereby reducing binding of the positively charged colistin. Recently transferable plasmid-mediated colistin resistance has also been found to be due to *mcr-1*, a gene encoding a phosphoethanolamine transferase that also reduces the negative charge on the cell surface. *mcr-1*-containing enteric bacteria have now been identified in Asia, Europe, and the United States.

Daptomycin

Daptomycin is active against gram-positive bacteria and interacts with and disrupts the cytoplasmic membrane in a calcium-dependent manner, resulting in bactericidal activity. The mechanisms of resistance to daptomycin are complex and involve mutations in several genes that can alter cell membrane charge and structure and reduce daptomycin binding. Resistance to daptomycin is relatively infrequent but has emerged in some *S. aureus* strains with intermediate vancomycin susceptibility from patients treated with vancomycin and not exposed to daptomycin. In some strains of methicillin-resistant *S. aureus*, daptomycin resistance has been linked to acquired susceptibility to β-lactams; combinations of daptomycin with nafcillin or ceftaroline have been successful for treatment of patients infected with resistant strains when daptomycin alone or in combination with other agents has failed. The mechanism of this effect is not yet clear but may involve alteration in surface charge and increased daptomycin binding in the presence of β-lactams. Daptomycin resistance has also been reported in enterococci.

EPIDEMIOLOGY OF RESISTANCE AND REDUCTION OF ITS OCCURRENCE

Multidrug resistance in human bacterial infections has been increasing overall in recent years, substantially limiting the number of antibiotics that can be used to treat some infections. The prevalence of resistance to various antimicrobials among human pathogens can, however, vary greatly in different geographic areas and even at different institutions in the same area. Thus, specific local data on the occurrence of various types of resistance are an important component of the choice of antimicrobials for empirical treatment of infection until the responsible pathogen is identified and its specific susceptibilities are determined by the clinical microbiology laboratory. Prompt adjustment of the initially chosen antimicrobial on the basis of species and susceptibility data to best target therapy is equally important. These principles emphasize the importance of obtaining appropriate samples for culture or other diagnostic modalities and susceptibility testing—whenever possible, prior to administration of antimicrobials. They also highlight the importance of rapid and sensitive diagnostic methods and the prompt communication of their results to clinicians to inform best choices of antimicrobials.

The overall prevalence of resistance can be affected by a number of factors, including (1) the extent of resistance reservoirs in the patient population; (2) the selection pressures from use of antimicrobials that favor resistant strains over susceptible ones; and (3) the extent by which resistance is amplified by transmission of resistant strains to patients from their environment or other persons, either directly or indirectly via the contaminated hands of health care workers when hand hygiene and other infection control practices are inadequately followed. The likelihood that an individual patient will be infected with a resistant pathogen is likewise affected by his or her history. Studies have shown that prior antibiotic treatment, prior infection with resistant pathogens, and prior hospitalizations all increase this likelihood.

These factors emphasize the importance of the appropriate use of antimicrobials (particularly, the avoidance of their use in clinical conditions in which they are not needed), the use of the shortest courses of therapy sufficient for a successful clinical outcome, and the implementation of antimicrobial stewardship programs (Chap. 139) as well as careful and consistent infection control practices in short- and long-term-care institutions. Antimicrobial agents are distinct among drug classes in human medicine in that—despite their clear clinical value when used appropriately—the extent of their use can compromise their future utility because of resistance. The remarkable ability of pathogens to acquire resistance is inherent in their biology and

emphasizes the necessity for clinicians and institutions to pay careful attention to those factors that can be controlled through judicious antimicrobial use and rigorous infection control and prevention practices.

Efforts to address the problems caused by resistance are now being made worldwide. The U.S. Centers for Disease Control and Prevention (CDC) has recently estimated that >2 million resistant bacterial infections occur in the United States each year, with 23,000 deaths, and has identified particular resistant pathogens that are of greatest concern because of their overall effects on public health **(Table 140-2)**. Enteric bacteria (such as *E. coli*, *K. pneumoniae*, and *Enterobacter* species) that are resistant to carbapenems are included in the "urgent" category because of their increasing occurrence worldwide and because they are often highly resistant to multiple drugs, with few if any active antimicrobials available for treatment. Resistant *N. gonorrhoeae* is included in this category as well because of the ease with which gonorrhea can be spread from person to person and because few active agents are now available. Other resistances are common and also affect clinical care, often requiring use of alternatives to first-line agents that can be less effective and less well tolerated. To address the problems posed by resistance, the CDC has emphasized a set of four core actions. (1) *Preventing infections and preventing spread:* These efforts focus on implementation of evidence-based activities to reduce the risks and incidence of device-related infections overall and on improvement of compliance with infection control practices that prevent transmission of resistant pathogens from one person to another, such as hand hygiene and isolation precautions in health care and long-term-care settings. (2) *Tracking resistance patterns:* Efforts aim to increase the reporting and sharing of the occurrence of resistance to enhance epidemiologic data and inform targeting of preventive interventions. (3) *Improving use of existing antimicrobials:* Antimicrobial stewardship programs with specific components to track usage and educate clinicians on appropriate use have become required in hospitals, and the CDC has implemented efforts to reduce inappropriate use in outpatient settings, with particular attention to upper respiratory illnesses that often do not require antimicrobials because of their common self-limited viral causes. (4) *Developing new antimicrobials and diagnostic tests:* The U.S. Congress and the U.S. Food and Drug Administration have recently developed incentives and enhanced regulatory pathways for drug approval that pharmaceutical companies can use for development of antimicrobials that specifically address particular resistant pathogens. Both small and large companies have undertaken efforts in this area. New technologies for rapid detection of resistance and susceptibility are also being developed by multiple diagnostics companies in order to facilitate the appropriate choice of antimicrobials earlier in the course of illness, providing an important tool for antimicrobial stewardship programs.

TABLE 140-2

Antibiotic Resistance Threats in the United States, 2013

THREAT CATEGORY	ORGANISMS
Urgent	*Clostridium difficile* Carbapenem-resistant Enterobacteriaceae Drug-resistant *Neisseria gonorrhoeae*
Serious	Multidrug-resistant *Acinetobacter* Drug-resistant *Campylobacter* Extended-spectrum β-lactamase–producing Enterobacteriaceae Vancomycin-resistant *Enterococcus* Multidrug-resistant *Pseudomonas aeruginosa* Drug-resistant nontyphoidal *Salmonella* Drug-resistant *Salmonella* Typhi Drug-resistant *Shigella* Methicillin-resistant *Staphylococcus aureus* Drug-resistant *Streptococcus pneumoniae* Drug-resistant *Mycobacterium tuberculosis*
Concerning	Vancomycin-resistant *S. aureus* Erythromycin-resistant group A *Streptococcus* Clindamycin-resistant group B *Streptococcus*

Source: U.S. Centers for Disease Control and Prevention.

FURTHER READING

Centers for Disease Control and Prevention: Antibiotic resistance threats in the United States, 2013. Available at *www.cdc.gov/drugresistance/threat-report-2013/pdf/ar-threats-2013-508.pdf*. Accessed June 26, 2017.

French GL: Antimicrobial resistance and healthcare-associated infections, in *Hospital Epidemiology and Infection Control*, 4th ed. GC Mayhall (ed). Philadelphia, Lippincott Williams & Wilkins, 2012, pp 1297–1310.

Rice LB: Mechanisms of resistance and clinical relevance of resistance to β-lactams, glycopeptides, and fluoroquinolones. Mayo Clin Proc 87:198, 2012.

Silver LL, Bush K (eds): *Antibiotics and Antibiotic Resistance*. Cold Spring Harbor Perspectives in Medicine. New York, Cold Spring Harbor Laboratory Press, 2016. 404 pp.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 141: Pneumococcal Infections

David Goldblatt; Katherine L. O'Brien

INTRODUCTION

In the late nineteenth century, pairs of micrococci were first recognized in the blood of rabbits injected with human saliva by both Louis Pasteur, working in France, and George Sternberg, an American army physician. The important role of these micrococci in human disease was not appreciated at that time. By 1886, when the organism was designated "pneumokokkus" and *Diplococcus pneumoniae*, it had been isolated by many independent investigators, and its role in the etiology of pneumonia was well known. In the 1930s, pneumonia was the third leading cause of death in the United States (after heart disease and cancer) and was responsible for ~7% of all deaths both in the United States and in Europe. While pneumonia was caused by a host of pathogens, lobar pneumonia—a pattern more likely to be caused by the pneumococcus—accounted for approximately one-half of all pneumonia deaths in the United States in 1929. In 1974, the organism was reclassified as *Streptococcus pneumoniae*.

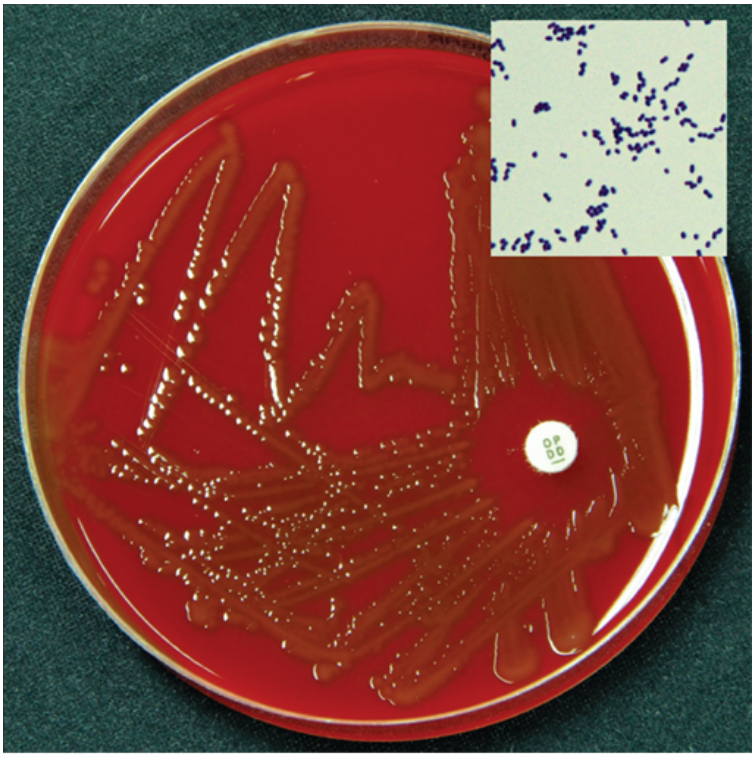
MICROBIOLOGY

Etiologic Agent

Pneumococci are spherical gram-positive bacteria of the genus *Streptococcus*. Within this genus, cell division occurs along a single axis, and bacteria grow in chains or pairs—hence the name *Streptococcus*, from the Greek *streptos*, meaning "twisted," and *kokkos*, meaning "berry." At least 22 streptococcal species are recognized and are divided further into groups based on their hemolytic properties. *S. pneumoniae* belongs to the α-hemolytic group that characteristically produces a greenish color on blood agar because of the reduction of iron in hemoglobin **(Fig. 141-1)**. The bacteria are fastidious and grow best in 5% CO_2 but require a source of catalase (e.g., blood) for growth on agar plates, where they develop mucoid (smooth/shiny) colonies. Pneumococci without a capsule produce colonies with a rough surface. Unlike that of other α-hemolytic streptococci, their growth is inhibited in the presence of optochin (ethylhydrocupreine hydrochloride), and they are bile soluble.

FIGURE 141-1

Pneumococci growing on blood agar, illustrating α hemolysis and optochin sensitivity (zone around optochin disk). *Inset:* Gram's stain, illustrating gram-positive diplococci. *(Photographs courtesy of Paul Turner, University of Oxford, United Kingdom.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

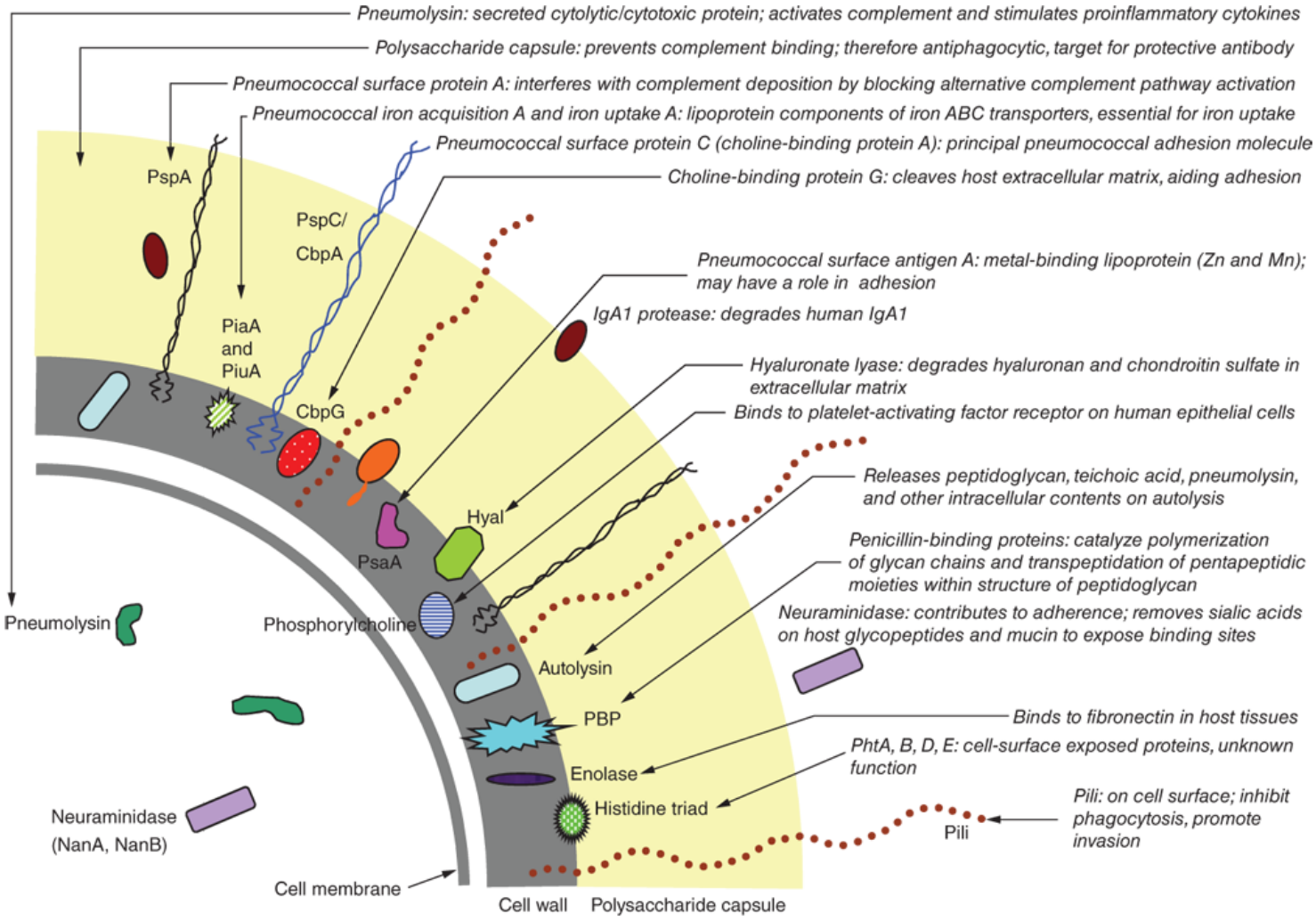
In common with other gram-positive bacteria, pneumococci have a cell membrane beneath a cell wall, which in turn is covered by a polysaccharide capsule. Pneumococci are divided into serogroups or serotypes based on capsular polysaccharide structure, as distinguished with rabbit polyclonal antisera; capsules swell in the presence of specific antiserum (the Quellung reaction). The most recently discovered serotypes—6C, 6D, 6F, 6G, 6H, 11E, 20A, and 20B—have been identified with monoclonal antibodies and by serologic, genetic, and biochemical means. The currently recognized 98 serotypes fall into 21 serogroups, and each serogroup contains two to eight serotypes with closely related capsules. In the absence of type-specific antibody, the capsule protects the bacteria from phagocytosis by host cells and is arguably the most important determinant of pneumococcal virulence. Unencapsulated variants are occasionally identified in cases of invasive pneumococcal disease; however, when their genotype is assessed, they often contain capsular genes. Thus it is likely that they were encapsulated in vivo and have stopped producing capsule during the laboratory steps of pathogen isolation.

Virulence Factors

Within the cytoplasm, cell membrane, and cell wall, many molecules that may play a role in pneumococcal pathogenesis and virulence have been identified **(Fig. 141-2)**. These proteins are often involved in direct interactions with host tissues or in concealment of the bacterial surface from host defense mechanisms. Pneumolysin is a secreted cytotoxin thought to result in cytolysis of cells and tissues, and LytA enhances pathogenesis. A number of cell wall proteins interfere with the complement pathway, thus inhibiting complement deposition and preventing lysis and/or opsonophagocytosis. The pneumococcal H inhibitor (Hic) impedes the formation of C3 convertase, while pneumococcal surface protein C (PspC), also known as choline-binding protein A (CbpA), binds factor H and is thought to accelerate the breakdown of C3. PspA and CbpA inhibit the deposition of or degrade C3b. The numerous pneumococcal proteins thought to be involved in adhesion include the ubiquitous surface-anchored sialidase (neuraminidase) NanA, which cleaves sialic acid on host cells and proteins, and pneumococcal surface adhesin A (PsaA). Pili recently recognized by electron microscopy also may play an important role in binding to cells. Some of the antigens mentioned above are potential vaccine candidates (see "Prevention," below). Biofilm production by pneumococci is now well recognized and is likely to be an important mechanism aiding survival of pneumococci in the upper respiratory tract and contributing to local disease manifestations such as otitis media.

FIGURE 141-2

Schematic diagram of the pneumococcal cell surface, with key antigens and their roles highlighted.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although the capsule surrounding the cell wall of *S. pneumoniae* is the basis for categorization by serotype, the behavior and pathogenic potential of a serotype may also be related to the genetic origin of the strain. Molecular typing is therefore of considerable interest. Initially, techniques such as pulsed-field gel electrophoresis were used to determine genetic relatedness; such techniques have been superseded by sequencing of housekeeping genes to define a clone (multilocus sequence typing, MLST). For *S. pneumoniae*, alleles at each of the loci *aroE*, *gdh*, *gki*, *recP*, *spi*, *xpt*, and *ddl* are sequenced and compared with all of the known alleles at that locus. Sequences identical to a known allele are assigned the same allele number, whereas those differing from any known allele—even at a single nucleotide site—are assigned new numbers. Software for assignment of alleles at each locus is available on the pneumococcal MLST website (*pubmlst.org/spneumoniae/*), and the allelic profile of each isolate and its consequent sequence type are generated. With the advent of high-throughput and relatively inexpensive sequencing techniques, whole-genome sequencing will soon supersede MLST. The first pneumococcal genome was sequenced in 2001 (a serotype 4 strain known as TIGR4), and to date more than 10,000 pneumococcal strains have been sequenced. The pneumococcal genome has ~2.2 million base pairs containing 2236 predicted coding regions. Genome sequence analysis has made major contributions to the understanding of pneumococcal biology and diversity.

EPIDEMIOLOGY

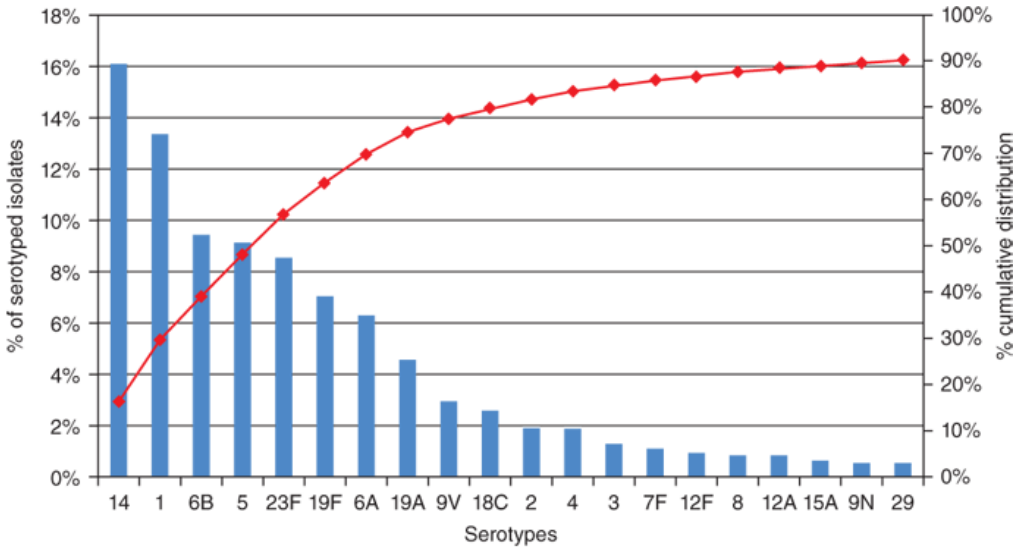
Pneumococcal infections remain a significant global cause of morbidity and death, particularly among children and the elderly **(See also "Global Health," below.)**. Rapid and dramatic changes in the epidemiology of this disease during the past 15 years in several developed countries followed the licensure and routine childhood administration of pneumococcal polysaccharide–protein conjugate vaccine (PCV). With PCV introduction in underdeveloped and middle-income countries, additional profound changes in pneumococcal ecology and disease epidemiology are occurring. The disease burden and serotype distribution in the PCV era are influenced not only by the reduction in disease caused by serotypes included in PCV but also by serotype replacement as a result of reductions in vaccine serotypes, concomitant secular trends in pneumococcal strains unrelated to vaccine use, the impact of antibiotic use on pneumococcal strain ecology, and surveillance system attributes that can themselves affect analysis of epidemiologic features of pneumococcal strains and disease.

Serotype Distribution

Not all pneumococcal serotypes are equally likely to cause disease; observed serotype distributions vary by age category, disease syndrome, and geography. Geographic differences may be driven by variations in the relative prevalence of syndromes causing disease rather than by true serotype distribution differences since certain serotypes are more common causes of some syndromes than others (e.g., pneumonia and meningitis). Most data on serotype distribution come from pediatric invasive pneumococcal disease (IPD, defined as infection of a normally sterile site); much less information on global or regional serotype distributions is available for disease in adults. In the era before PCV use, five to seven serotypes caused >60% of IPD cases among children <5 years of age in most parts of the world; seven serotypes (1, 5, 6A, 6B, 14, 19F, and 23F) accounted for ~60% of such cases in all areas of the world, but in any given region these seven serotypes may not all rank as the most common disease strains **(Fig. 141-3)**. Some serotypes (e.g., types 1 and 5) not only tend to cause disease in areas with a high disease burden but also cause waves of disease in lower-burden areas (e.g., Europe) or outbreaks (e.g., in military barracks; meningitis in sub-Saharan Africa). The broader range of serotypes causing disease among adults than among children is apparent from a comparison of the coverage of existing multiserotype vaccines in different age groups. For example, data from the United States for 2013–2014 on the serotypes causing IPD indicated that a polysaccharide vaccine containing 23 serotypes (PPSV23) would include the serotypes causing 59% of cases among children <5 years of age, 70% of those among persons 18–64 years of age, and 57% of those among persons ≥65 years of age.

FIGURE 141-3

Meta-analysis of available global pneumococcal serotype data, adjusted for regional disease incidence. The red line shows cumulative incidence, as indicated on the right-hand Y axis. *(Source: Global Serotype Project Report for the Pneumococcal Advance Market Commitment Target Product Profile; available at www.gavi.org/library/gavi-documents/amc/tpp-codebook/.)*



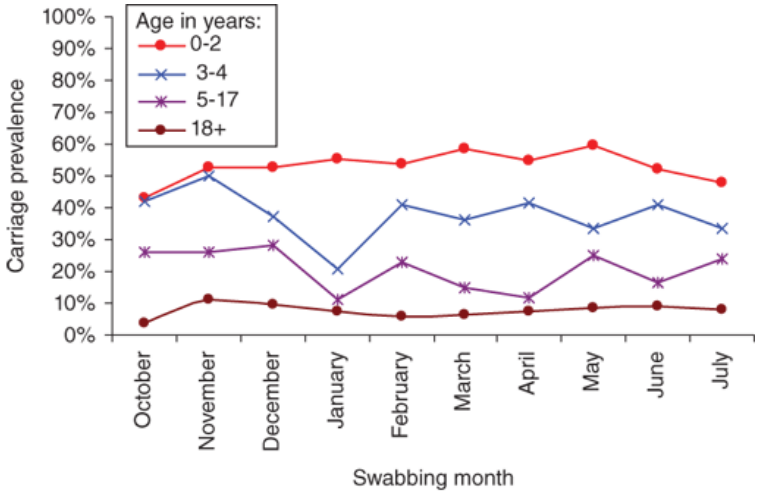
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Nasopharyngeal Carriage

Pneumococci are intermittent inhabitants of the healthy human nasopharynx and are transmitted by respiratory droplets. In children, pneumococcal nasopharyngeal ecology varies by geographic region, socioeconomic status, climate, degree of crowding, and particularly intensity of exposure to other children, with children in day-care settings having higher rates of colonization. In developed-world settings, children serve as the major vectors of pneumococcal transmission. By 1 year of age, ~50% of children have had at least one episode of pneumococcal colonization. Cross-sectional prevalence data show rates of pneumococcal carriage ranging from 20% to 50% among children <5 years of age and from 5% to 15% among young and middle-aged adults; **Fig. 141-4** shows relevant data from the United Kingdom. Data on colonization rates among healthy elderly individuals are limited. In developing-world settings, pneumococcal acquisition occurs much earlier, sometimes within the first few days after birth, and nearly all infants have had at least one episode of colonization by 2 months of age. Cross-sectional studies show that up to the age of 5 years, 70–90% of children carry *S. pneumoniae* in the nasopharynx, and a significant proportion of adults (sometimes >40%) also are colonized. Their high rates of colonization make adults an important source of transmission and may affect community transmission dynamics.

FIGURE 141-4

Prevalence of pneumococcal carriage in adults and children resident in the United Kingdom who had nasopharyngeal swabs collected monthly for 10 months (no seasonal trend; *t* test trend, >.05). *(Data adapted from D Goldblatt et al: J Infect Dis 192:387, 2005.)*



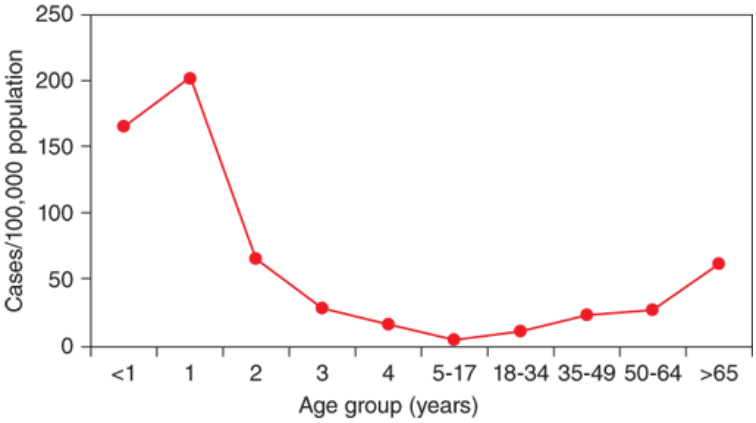
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Invasive Disease and Pneumonia

IPD develops when *S. pneumoniae* invades the bloodstream and seeds other organs or directly reaches the cerebrospinal fluid (CSF) by local extension. Pneumonia may follow aspiration of pneumococci, although only 10–30% of pneumococcal pneumonia cases are associated with a positive blood culture (and thus contribute to the measured burden of IPD). The substantial variation of IPD rates with age is illustrated by data from the United States for 1998–1999, a period prior to PCV introduction. Rates of IPD were highest among children <2 years of age and among adults ≥65 years of age (188 and 60 cases/100,000, respectively; **Fig. 141-5**). Since the introduction of PCV, IPD rates among infants and children in the United States have fallen by >75%, a decrease driven by the near elimination of vaccine-serotype IPD. A similar impact of PCV on vaccine-serotype IPD rates has been consistently observed in countries where PCV has been introduced into the routine pediatric vaccination schedule. However, the magnitudes of change in the non-vaccine-serotype IPD rate in various countries have been heterogeneous; the interpretation of this heterogeneity is a complex issue. In the United States, Canada, and Australia, rates of non-vaccine-serotype IPD have increased but the magnitude of the increase is generally small relative to the substantial reductions in vaccine-serotype IPD. In contrast, in other settings (e.g., Alaska Native communities and the United Kingdom), the reduction in vaccine-serotype IPD has been offset by notable increases in rates of disease caused by non-vaccine serotypes. Explanations for the heterogeneity of findings include replacement disease resulting from vaccine pressure, changes in clinical case investigation, secular trends unrelated to PCV use, antibiotic pressure selecting for resistant organisms, changes in surveillance or reporting systems, rapidity of PCV introduction, and inclusion of a catch-up campaign. A recent systematic review concludes that serotype replacement in IPD follows the use of PCV7 but that the magnitude of this phenomenon is small relative to the reduction in disease from vaccine serotypes. Furthermore, much of the replacement disease has been caused by serotype 19A. In contrast to the changes seen following implementation of PCV7, the replacing serotypes are more diverse and a single serotype is not dominant with the use of expanded-valency PCV products (i.e., PCV10 and PCV13, the latter of which includes serotype 19A). In spite of some serotype replacement in disease, the net effect of PCV is to reduce the rate of pneumococcal disease both in the age group targeted for vaccination and in unvaccinated age groups.

FIGURE 141-5

Rates of invasive pneumococcal disease before the introduction of pneumococcal conjugate vaccine, by age group: United States, 1998. *(Source: CDC, Active Bacterial Core Surveillance/Emerging Infectious Program Network, 2000. Data adapted from MMWR 49[RR-9], 2000.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pneumonia is the most common of the serious pneumococcal disease syndromes and poses special challenges from a clinical and public health perspective. Most cases of pneumococcal pneumonia are not associated with bacteremia, and in these cases a definitive etiologic diagnosis is difficult or impossible. As a result, estimates of disease burden focus primarily on IPD rates and fail to include the major portion of the burden of serious pneumococcal disease. Among children, PCV trials designed to collect efficacy data on syndrome-based outcomes (e.g., radiographically confirmed pneumonia, clinically diagnosed pneumonia) have revealed the burden of culture-negative pneumococcal pneumonia. These trials have provided the means to infer that only ~5–20% of pneumococcal pneumonia cases result in bacteremia. An important randomized controlled trial of PCV among the elderly in the Netherlands (the Capita trial) has revealed the small fraction of adult pneumococcal pneumonia patients who also have bacteremia. Use of high-quality sputum specimens and, in the case of adults with a low likelihood of colonization absent disease, urine antigen detection both contribute to the diagnosis of nonbacteremic pneumococcal pneumonia. Furthermore, evidence continues to accrue that pneumococcal pneumonia events are often the result of co-infection with viral or other bacterial pathogens. Thus a pneumonia case resulting from a pulmonary infection with a single pathogen is probably an uncommon event; rather, most cases of pneumonia likely result from the sequential or contemporaneous co-infection of a host with multiple pathogens, often both viruses and bacteria.

The case–fatality ratios (CFRs) for pneumococcal pneumonia and IPD vary by age, underlying medical condition, and access to care. In addition, the CFR for pneumococcal pneumonia varies with the severity of disease at presentation (rather than according to whether the pneumonia episode is associated with bacteremia) and with the patient's age (from <5% among hospitalized patients 18–44 years old to >12% among those >65 years old, even when appropriate and timely management is available). Notably, the likelihood of death in the first 24 h of hospitalization did not change substantially with the introduction of antibiotics; this surprising observation highlights the fact that the pathophysiology of severe pneumococcal pneumonia among adults reflects a rapidly progressive cascade of events that often unfolds irrespective of antibiotic administration. Management in an intensive care unit can provide critical support for the patient through the acute period, with lower CFRs, while antibiotics address the underlying infection.

Rates of pneumococcal disease vary by season, with higher rates in colder than in warmer months in temperate climates; by sex, with males more often affected than females; and by risk group, with risk factors including underlying medical conditions, behavioral issues (e.g., smoking), and ethnic group. In the United States, some Native American populations (including Alaska Natives) and African Americans have higher rates of disease than the general population; the increased risk is probably attributable to socioeconomic conditions and the prevalence of underlying risk factors for pneumococcal disease. Medical conditions that increase the risk of pneumococcal infection are listed in **Table 141-1**. Outbreaks of disease are well recognized in crowded settings with susceptible individuals, such as infant day-care facilities, military barracks, and nursing homes. Furthermore, there is a clear association between preceding viral respiratory disease (especially but not exclusively influenza) and risk of secondary pneumococcal infections. The significant role of pneumococcal pneumonia in the morbidity and mortality associated with seasonal and pandemic influenza is increasingly well recognized.

TABLE 141-1

Clinical Risk Groups for Pneumococcal Infection

Clinical Risk Group	Examples
Asplenia or splenic dysfunction	Sickle cell disease, celiac disease
Chronic respiratory disease	Chronic obstructive pulmonary disease, bronchiectasis, cystic fibrosis, interstitial lung fibrosis, pneumoconiosis, bronchopulmonary dysplasia, aspiration risk, neuromuscular disease (e.g., cerebral palsy), severe asthma
Chronic heart disease	Ischemic heart disease, congenital heart disease, hypertension with cardiac complications, chronic heart failure
Chronic kidney disease	Nephrotic syndrome, chronic renal failure, renal transplantation
Chronic liver disease	Cirrhosis, biliary atresia, chronic hepatitis
Diabetes mellitus	Diabetes mellitus requiring insulin or oral hypoglycemic drugs
Immunocompromise/immunosuppression	HIV infection, common variable immunodeficiency, leukemia, lymphoma, Hodgkin's disease, multiple myeloma, generalized malignancy, chemotherapy, organ or bone marrow transplantation, systemic glucocorticoid treatment for >1 month at a dose equivalent to ≥20 mg/d (children, ≥1 mg/kg per day)
Cochlear implants	…
Cerebrospinal fluid leaks	…
Miscellaneous	Infancy and old age; prior hospitalization; alcoholism; malnutrition; cigarette smoking; day-care center attendance; residence in military training camps, prisons, homeless shelters

Note: Groups for whom pneumococcal vaccines are recommended by the Advisory Committee on Immunization Practices can be found at *www.cdc.gov/vaccines/schedules/*.

Antibiotic Resistance

Reduced pneumococcal susceptibility to penicillin was first noted in 1967, but not until the 1990s did reduced antibiotic susceptibility emerge as a significant clinical and public health issue, with an increasing prevalence of pneumococcal isolates resistant to single or multiple classes of antibiotics and a rising absolute magnitude of minimal inhibitory concentrations (MICs). Strains with reduced susceptibility to penicillin G, cefotaxime, ceftriaxone, macrolides, and other antibiotics are now found worldwide and account for a significant proportion of disease-causing strains in many locations, especially among children. Vancomycin resistance has not yet been observed in clinical pneumococcal strains. Lack of antimicrobial susceptibility is clearly related to a subset of serotypes, many of which disproportionately cause disease among children. Resistance phenotypes are based on a diverse array of mutational events and inter- and intraspecies gene-transfer phenomena carried out by several types of mobile genetic elements, with consequent dissemination of successful resistant clones. The vicious cycle of antibiotic exposure, selection of resistant organisms in the nasopharynx, and transmission of these organisms within the community, leading to difficult-to-treat infections and increased antibiotic exposure, has been interrupted to some extent by the introduction and routine use of PCV. The clinical implications of pneumococcal antimicrobial nonsusceptibility are addressed below in the section on treatment.

PATHOGENESIS

Pneumococci colonize the human nasopharynx from an early age; colonization acquisition events are generally described as asymptomatic, but evidence exists to associate acquisition with mild respiratory symptoms, especially in the very young. Bacteria survive in the nasopharynx protected by a variety of factors, including their bacterial capsule and the formation of a biofilm. From the nasopharynx, the bacteria spread either via the bloodstream to distant sites (e.g., brain, joint, bones, peritoneal cavity) or locally to mucosal surfaces where they can cause otitis media or pneumonia. Direct spread from the nasopharynx to the central nervous system (CNS) can occur in rare cases of skull-base fracture, although most

cases of pneumococcal meningitis are secondary to hematogenous spread. The pneumococcus is not a static bacterium; rather, it modifies its expression of capsule in adaptation to the external environment. In the nasopharynx, the pneumococcus downregulates capsular expression, averting protective immunologic mechanisms that recognize capsule; the phenotype on culture are rough colonies. Upon invasion by traversal of the epithelium, the pneumococcus upregulates its capsular expression, transforming its appearance on culture to smooth colonies—a change illustrating the dynamic nature of the organism in response to the local environment. Pneumococci can cause disease in almost any organ or part of the body; however, otitis media, pneumonia, bacteremia, and meningitis are most common. Colonization is a relatively frequent event, yet disease is rare. In the nasopharynx, pneumococci survive in mucus secreted by epithelial cells and in a biofilm they create, where they can avoid local immune factors such as leukocytes and complement. The mucus itself is a component of local defense mechanisms, and the flow of mucus (driven in part by cilia in what is known as the *mucociliary escalator*) effects mechanical clearance of pneumococci. While many colonization episodes are of short duration, longitudinal studies in adults and children have revealed persistent colonization with a specific serotype over many months. Colonization eventually results in the development of capsule- and protein-specific serum IgG antibodies, which are thought to play a role in mediating clearance of bacteria from the nasopharynx. IgG antibodies to surface-exposed cell-wall or secreted proteins also appear in the circulation in an age-dependent fashion or after colonization; the biologic role of these antibodies is less clear. Recent acquisition of a new colonizing serotype is more likely to be associated with subsequent invasion, presumably as a result of the absence of type-specific immunity. Intercurrent viral infections make the host more susceptible to pneumococcal colonization, and pneumococcal disease in a colonized individual often follows perturbation of the nasopharyngeal mucosa by such infections. Local cytokine production after a viral infection is thought to upregulate adhesion factors in the respiratory epithelium, allowing pneumococci to adhere via a variety of surface adhesin molecules, including PsaA, PspA, CbpA, PspC, Hyl, pneumolysin, and the neuraminidases (Fig. 141-2). Adhesion coupled with inflammation induced by pneumococcal factors such as peptidoglycans and teichoic acids results in invasion. It is the inflammation induced by various bacterium-derived factors that is responsible for the pathology associated with pneumococcal infection. Pneumococcal cell wall–derived teichoic acids and peptidoglycans induce a variety of cytokines, including the proinflammatory cytokines interleukin (IL) 1, IL-6, and tumor necrosis factor, and activate complement via the alternative pathway. Polymorphonuclear leukocytes are thus attracted, and an intense inflammatory response is initiated. Pneumolysin also is important in local pathology, inducing proinflammatory cytokine production by local monocytes.

The pneumococcal capsule, consisting of polysaccharides with antiphagocytic properties (i.e., the capacity to resist complement deposition in the absence of type-specific antibody), plays an important role in pathogenesis. While most capsular types can cause human disease, certain capsular types are more commonly isolated from sites of infection. The reason for the dominance of some serotypes over others in IPD, as depicted in Fig. 141-3, is unclear.

HOST DEFENSE MECHANISMS

Innate Immunity

As described above, intact respiratory epithelium and a host of nonspecific or innate immune factors (e.g., mucus, splenic function, complement, neutrophils, and macrophages) constitute the first line of defense against pneumococci. Physical factors such as the cough reflex and the mucociliary escalator are important in clearing bacteria from the lungs. Immunologic factors are critical as well: C-reactive protein (CRP) binds phosphorylcholine in the pneumococcal cell wall, inducing complement activation and leading to bacterial clearance; Toll-like receptor 2 (TLR2) recognizes both pneumococcal lipoteichoic acid and cell wall peptidoglycan; and in animal models, the absence of host TLR2 leads to more severe infection and impaired clearance of nasopharyngeal colonization. TLR4 appears to be necessary for the proinflammatory effect of pneumolysin on macrophages. The importance of TLR recognition is underlined by descriptions of an inherited deficiency of human IL-1 receptor–associated kinase 4 (IRAK-4) that manifests as an unusual susceptibility to infection with bacteria, including *S. pneumoniae*. IRAK-4 is essential for the normal functioning of several TLRs. Other factors that interfere with these nonspecific mechanisms (e.g., viral infections, cystic fibrosis, bronchiectasis, complement deficiency, and chronic obstructive pulmonary disease) all predispose to the development of pneumococcal pneumonia. Patients who lack a spleen or have abnormal splenic function (e.g., persons with sickle cell disease) are at high risk of developing overwhelming pneumococcal disease.

Acquired Immunity

Acquired immunity induced following colonization or through exposure to cross-reactive antigens rests largely on the development of serum IgG antibody specific for the pneumococcal capsular polysaccharide. Nearly all polysaccharides are T cell–independent antigens; B cells can make antibodies to such antigens without T cell help. However, in children <1–2 years old, such B cell responses are poorly developed. This delayed ontogeny of capsule-specific IgG in young children is associated with susceptibility to pneumococcal infection (Fig. 141-5). The extremely high risk of pneumococcal infection in the absence of serum immunoglobulin (i.e., in conditions such as agammaglobulinemia) highlights the important role of capsular antibody in protection against disease. Each serotype's capsule is chemically distinct, even though for some serotypes the chemical distinction from another type may be a minor one; thus immunity tends to be serotype specific, although some cross-immunity exists. For example, conjugate vaccine–induced antibodies to serotype 6B prevent infection due to serotype 6A. However, cross-protection against serotypes within serogroups is not universal; for instance, antibodies to serotype 19F induced by some vaccines do not appear to confer protection against disease

caused by serotype 19A. Antibodies to surface-exposed or secreted pneumococcal proteins (such as pneumolysin, PsaA, and PspA) also appear in the circulation with increasing age of the host, but their functional significance remains unclear. Data from murine models suggest that CD4+ T cells may play a role in preventing pneumococcal colonization and disease, and experimental data derived from humans suggest that IL-17-secreting CD4+ T cells may be relevant.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Pneumococcal Infections

There is no pathognomonic presentation of pneumococcal disease; patients may present with one or more clinical syndromes (e.g., pneumonia, meningitis, sepsis). *S. pneumoniae* can infect nearly any body tissue, manifesting as disease ranging in severity from mild and self-limited to life-threatening. The differential diagnosis of common clinical syndromes such as pneumonia, otitis media, fever of unknown origin, and meningitis should always include pneumococcal infection. A microbiologically confirmed diagnosis is made in only a minority of pneumococcal cases since, in most circumstances (and especially in pneumonia and otitis media), fluid from the site of infection is not available for etiologic determination, and infection of body fluids distant from the site of infection (e.g., blood in the case of pneumonia) occurs in only a minority of true pneumococcal cases. Empirical therapy that includes appropriate treatment for *S. pneumoniae* is often indicated.

Algorithms for assessment and management of ill children (IMCI; Integrated Management of Childhood Illness) have been developed for use in the developing world or in other settings where evaluation by a trained physician may not be feasible. No such algorithms for the management of adults with suspected disease exist. Children who present with signs associated with increased risk of serious disease, such as an inability to drink, convulsions, lethargy, and severe malnutrition, are categorized as having very severe disease without further evaluation by the community health care worker; are given antibiotics; and are immediately referred to a hospital for diagnosis and management. Children who present with cough and tachypnea (the latter defined according to specific age strata) are further stratified into severity categories based on the presence or absence of lower chest wall indrawing and are managed accordingly either with antibiotics alone or with antibiotics and referral to a hospital facility. Children with cough but no tachypnea are categorized as having a nonpneumonia respiratory illness.

CLINICAL MANIFESTATIONS

The clinical manifestations of pneumococcal disease depend on the site of infection and the duration of illness. Clinical syndromes are classified as noninvasive (e.g., otitis media) or invasive (e.g., bacteremic pneumonia, meningitis) according to whether a normally sterile site is infected. The pathogenesis of noninvasive illness involves contiguous spread from the nasopharynx or skin; invasive disease involves infection of a normally sterile body fluid or follows bacteremia. Regardless of the mechanism, all pneumococcal infections result from nasopharyngeal acquisition of the organism.

Pneumonia

Pneumonia is the most common serious pneumococcal syndrome and is considered invasive when associated with a positive blood culture. Whether to categorize nonbacteremic pneumococcal pneumonia as invasive or noninvasive remains debatable.

Pneumococcal pneumonia can present as a mild community-acquired infection at one extreme and as a life-threatening disease requiring intubation and intensive support at the other.

Presenting Manifestations

The presentation of pneumococcal pneumonia does not reliably distinguish it from pneumonia of other etiologies. In a subset of cases, pneumococcal pneumonia is recognized at the outset as associated with a viral upper respiratory infection and is characterized by the abrupt onset of cough and dyspnea accompanied by fever, shaking chills, and myalgias. The cough evolves from nonpurulent to productive of sputum that is purulent and sometimes tinged with blood. Patients may describe stabbing pleuritic chest pain and significant dyspnea indicating involvement of the parietal pleura. Among the elderly, the presenting clinical symptoms may be less specific, with confusion or malaise but without fever or cough. In such cases, a high index of suspicion is required because failure to treat pneumococcal pneumonia promptly in an elderly patient is likely to result in rapid evolution of the infection, with increased severity, morbidity, and risk of death.

Findings on PHYSICAL Examination

The clinical signs associated with pneumococcal pneumonia among adults include tachypnea (defined as >30 breaths/min) and tachycardia, hypotension in severe cases, and fever in most cases (although not in all elderly patients). Respiratory signs are varied, including dullness to percussion in areas of the chest with significant consolidation, crackles on auscultation, reduced expansion of the chest in some cases as a result of

splinting to reduce pain, bronchial breathing in a minority of cases, pleural rub in occasional cases, and cyanosis in cases with significant hypoxemia. Among infants with severe pneumonia, chest wall indrawing and nasal flaring are common. Nonrespiratory findings can include upper abdominal pain if the diaphragmatic pleura is involved as well as mental status changes, particularly confusion in elderly patients.

Differential Diagnosis

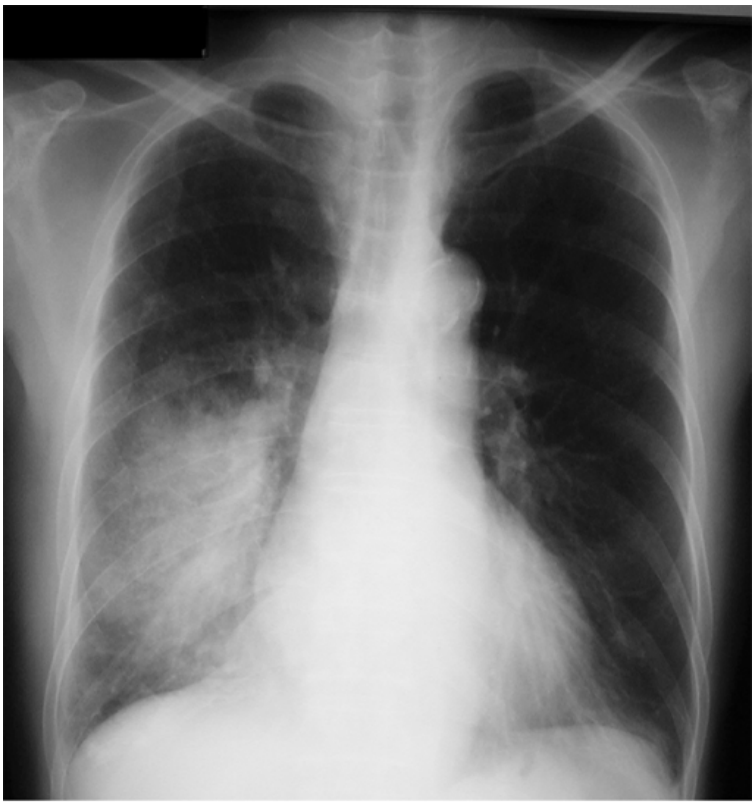
The differential diagnosis of pneumococcal pneumonia includes cardiac conditions such as myocardial infarction and heart failure with atypical pulmonary edema; pulmonary conditions such as atelectasis; and pneumonia caused by viral pathogens, mycoplasmas, *Haemophilus influenzae*, *Klebsiella pneumoniae*, *Staphylococcus aureus*, *Legionella*, or (in HIV-infected and otherwise immunocompromised hosts) *Pneumocystis*. In cases with abdominal symptoms, the differential diagnosis includes cholecystitis, appendicitis, perforated peptic ulcer disease, and subphrenic abscesses. The challenge in cases with abdominal symptoms is to remember to include pneumococcal pneumonia—a nonabdominal process—in the differential diagnosis.

Diagnosis

Some authorities advocate treating uncomplicated, nonsevere, community-acquired pneumonia without determining the microbiologic etiology, given that this information is unlikely to alter clinical management. However, efforts to identify the cause of pneumonia are important when the disease is more severe and when the diagnosis of pneumonia is not clearly established. The gold standard for etiologic diagnosis of pneumococcal pneumonia is pathologic examination of lung tissue. In lieu of that procedure, evidence of an infiltrate on chest radiography warrants a diagnosis of pneumonia. However, cases of pneumonia without radiographic evidence do occur. An infiltrate can be absent either early in the course of the illness or with dehydration; upon rehydration, an infiltrate usually appears. The radiographic appearance of pneumococcal pneumonia is varied; it classically consists of lobar or segmental consolidation **(Fig. 141-6)** but in some cases is patchy. More than one lobe is involved in ~30% of cases. Consolidation may be associated with a small pleural effusion or empyema in complicated cases. In children, “round pneumonia,” a distinctly spherical consolidation on chest radiography, is associated with a pneumococcal etiology. Round pneumonia is uncommon in adults. *S. pneumoniae* is not the only cause of such lesions; other causes, especially cancer, should be considered.

FIGURE 141-6

Chest radiograph depicting classic lobar pneumococcal pneumonia in the right lower lobe of an elderly patient’s lung.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Blood drawn from patients with suspected pneumococcal pneumonia can be used for supportive or definitive diagnostic tests. Blood cultures are positive for pneumococci in a minority (<30%) of cases of pneumococcal pneumonia, as evidenced especially by vaccine clinical trials, which provide an independent method to reveal the contribution of the pneumococcus to pneumonia cases. Nonspecific findings include an elevated

polymorphonuclear leukocyte count (>15,000/μL in most cases and upward of 40,000/μL in some), leukopenia in <10% of cases (a poor prognostic sign associated with a fatal outcome), and elevated values in liver function tests (e.g., both conjugated and unconjugated hyperbilirubinemia). Anemia, low serum albumin levels, hyponatremia, and elevated serum creatinine levels are all found in ~20–30% of patients.

Urinary pneumococcal antigen assays have facilitated etiologic diagnosis, but the application of the results is confounded by the fact that nasopharyngeal colonization with the pneumococcus, in the absence of disease, also results in a positive test. In adults, therefore, a positive pneumococcal urinary antigen test has a high predictive value for etiologic attribution of pneumonia because the prevalence of pneumococcal nasopharyngeal colonization is relatively low. In communities, particularly those in low-income countries, where colonization rates among adults are high, urine antigen assays may be less useful. The same issue holds for children, in whom a positive urinary antigen test is usually uninformative for etiologic attribution of their pneumonia illness because colonization rates are generally high. A recent advance is the development of quantitative serotype-specific urinary antigen detection assays; their application for adults and children holds promise, especially in detecting serotypes that are rarely identified in asymptomatic carriage (e.g., serotype 1), even among children.

Most cases of pneumococcal pneumonia in adults are diagnosed by Gram's staining and culture of sputum. The utility of a sputum specimen is directly related to its quality and the patient's antibiotic treatment status.

Complications

Empyema is the most common focal complication of pneumococcal pneumonia, occurring in <5% of cases. When fluid in the pleural space is accompanied by fever and leukocytosis (even low-grade) after 4–5 days of appropriate antibiotic treatment for pneumococcal pneumonia, empyema should be considered. Parapneumonic effusions are more common than empyema, representing a self-limited inflammatory response to pneumonia. Pleural fluid with frank pus, bacteria (detected by microscopic examination), or a pH of ≤7.1 indicates empyema and demands aggressive and complete drainage, usually through chest tube insertion.

Meningitis

Pneumococcal meningitis usually presents as a pyogenic condition that is clinically indistinguishable from meningitis of other bacterial etiologies. Meningitis can be the primary presenting pneumococcal syndrome or a complication of other conditions such as skull fracture, otitis media, bacteremia, or mastoiditis. Now that *H. influenzae* type b vaccine is routinely used in children, *S. pneumoniae* and *Neisseria meningitidis* are the most common bacterial causes of meningitis in both adults and children. Pyogenic meningitis, including that due to *S. pneumoniae*, is associated clinically with findings that include severe, generalized, gradual-onset headache, fever, and nausea as well as specific CNS manifestations such as stiff neck, photophobia, seizures, and confusion. Clinical signs include a toxic appearance, altered consciousness, bradycardia, and hypertension indicative of increased intracranial pressure. A small proportion of adult patients have Kernig's or Brudzinski's sign or cranial nerve palsies (particularly of the third and sixth cranial nerves).

A definitive diagnosis of pneumococcal meningitis rests on the examination of CSF for (1) evidence of turbidity (visual inspection); (2) elevated protein level, elevated white blood cell count, and reduced glucose concentration (quantitative measurement); and (3) specific identification of the etiologic agent (culture, Gram's staining, antigen testing, or polymerase chain reaction [PCR]). A blood culture positive for *S. pneumoniae* in conjunction with clinical manifestations of meningitis also is considered confirmatory. As discussed in the section on pneumonia, among adults, detection of pneumococcal antigen in urine is considered highly specific because of the low prevalence of nasopharyngeal colonization in this age group.

The mortality rate for pneumococcal meningitis is ~20%. In addition, up to 50% of survivors experience acute or chronic complications, including deafness, hydrocephalus, and mental retardation in children and diffuse brain swelling, subarachnoid bleeding, hydrocephalus, cerebrovascular complications, and hearing loss in adults.

Other Invasive Syndromes

S. pneumoniae can cause other invasive syndromes involving virtually any body site. These syndromes include primary bacteremia without other sites of infection (bacteremia without a source; occult bacteremia), osteomyelitis, septic arthritis, endocarditis, pericarditis, and peritonitis. The essential diagnostic approach is collection of fluid from the site of infection by sterile technique and examination by Gram's staining, culture, and—when relevant—capsular antigen assay or PCR. Hemolytic-uremic syndrome can complicate invasive pneumococcal disease.

Noninvasive Syndromes

The two major noninvasive syndromes caused by *S. pneumoniae* are sinusitis and otitis media; the latter is the most common pneumococcal syndrome and most often affects young children. The manifestations of otitis media include the acute onset of severe pain, fever, deafness, and tinnitus, most frequently in the setting of a recent upper respiratory tract infection. Clinical signs include a red, swollen, often bulging tympanic

membrane with reduced movement on insufflation or tympanography. Redness of the tympanic membrane is not sufficient for the diagnosis of otitis media.

Pneumococcal sinusitis is also a complication of upper respiratory tract infections and presents with facial pain, congestion, fever, and—in many cases—persistent nighttime cough. A definitive diagnosis is made by aspiration and culture of sinus material; however, presumptive treatment is most commonly initiated after application of a strict set of clinical diagnostic criteria.

TREATMENT

TREATMENT

Pneumococcal Infections

Historically, the activity of penicillin against pneumococci made parenteral penicillin G the drug of choice for disease caused by susceptible organisms, including community-acquired pneumonia. For susceptible strains, penicillin G remains the most commonly used agent, with daily doses ranging from 50,000 U/kg for minor infections to 300,000 U/kg for meningitis. Other parenteral β-lactam drugs, such as ampicillin, cefotaxime, ceftriaxone, and cefuroxime, can be used against penicillin-susceptible strains but offer little advantage over penicillin. Macrolides and cephalosporins are alternatives for penicillin-allergic patients. While agents such as clindamycin, tetracycline, and trimethoprim-sulfamethoxazole exhibit some activity against pneumococci, resistance to these agents is frequently encountered in different parts of the world.

Penicillin-resistant pneumococci were first described in the mid-1960s, at which point tetracycline- and macrolide-resistant strains had already been reported. Multidrug-resistant strains were first described in the 1970s, but it was during the 1990s that pneumococcal drug resistance reached pandemic proportions. The use of antibiotics selects for resistant pneumococci, and strains resistant to β-lactam agents and to multiple drugs are now found all over the world. The emergence of high rates of macrolide and fluoroquinolone resistance also has been described.

The molecular basis of penicillin resistance in *S. pneumoniae* is the alteration of penicillin-binding protein (PBP) genes by transformation and horizontal transfer of DNA from related streptococcal species. Such alteration of PBPs results in lower affinity for penicillins. Depending on the specific PBP(s) and the number of PBPs altered, the level of resistance ranges from intermediate to high. For many years, penicillin susceptibility breakpoints have been defined by MICs as follows: susceptible, ≤0.06 μg/mL; intermediate, 0.12–1.0 μg/mL; and resistant, ≥2.0 μg/mL. However, in vitro results often were not predictive of the response of a patient to treatment for pneumococcal diseases other than meningitis. Revised recommendations have been based on the penicillin G breakpoints established in 2008 by the Clinical and Laboratory Standards Institute. For IV treatment of meningitis with at least 24 million units per day in 8 divided doses, the susceptibility breakpoint remains ≤0.06 μg/mL, and MICs of ≥0.12 μg/mL indicate resistance. For IV treatment of nonmeningeal infections with 12 million units per day in 6 divided doses, the breakpoints are ≤2 μg/mL for susceptible organisms, 4 μg/mL for intermediate organisms, and ≥8 μg/mL for resistant organisms; a dosage of 18–24 million units per day is recommended for strains with MICs in the intermediate category. The original breakpoints remain the same for oral treatment of nonmeningeal infections with penicillin V.

Although guidelines for antibiotic therapy should be driven in part by local patterns of resistance, guidelines from national organizations in many countries (e.g., the Infectious Diseases Society of America/American Thoracic Society, the British Thoracic Society, and the European Respiratory Society) lay out evidence-based approaches. The following guidelines for the treatment of individual sepsis syndromes are based on those advocated by the American Academy of Pediatrics and published in the 2015 *Red Book*.

MENINGITIS LIKELY OR PROVEN TO BE DUE TO *S. PNEUMONIAE*

As a result of the increased prevalence of resistant pneumococci, first-line therapy for persons ≥1 month of age is a combination of vancomycin (adults, 30–60 mg/kg per day; infants and children, 60 mg/kg per day) and cefotaxime (adults, 8–12 g/d in 4–6 divided doses; children, 225–300 mg/kg per day in 1 dose or 2 divided doses) or ceftriaxone (adults, 4 g/d in 1 dose or 2 divided doses; children, 100 mg/kg per day in 1 dose or 2 divided doses). If children are hypersensitive to β-lactam agents (penicillins and cephalosporins), rifampin (adults, 600 mg/d; children, 20 mg/d in 1 dose or 2 divided doses) can be substituted for cefotaxime or ceftriaxone. A repeat lumbar puncture should be considered after 48 h if the organism is not susceptible to penicillin and information on cephalosporin sensitivity is not yet available, if the patient's clinical condition does not improve or deteriorates, or if dexamethasone has been administered and may be compromising clinical evaluation. When antibiotic sensitivity data become available, treatment should be modified accordingly. If the isolate is sensitive to penicillin, vancomycin can be discontinued and penicillin can replace the cephalosporin, or cefotaxime or ceftriaxone can be continued alone. If the isolate displays any resistance to penicillin but is susceptible to the cephalosporins, vancomycin can be discontinued and cefotaxime or ceftriaxone continued. If the isolate exhibits any resistance to penicillin and is not susceptible to cefotaxime and ceftriaxone, vancomycin and high-dose cefotaxime or ceftriaxone can be continued; rifampin may be added as well if the isolate is susceptible and the patient's clinical condition is worsening, if the CSF remains positive for bacteria, or if the MIC of the cephalosporin in question against the infecting strain is high. Some physicians advocate the use of glucocorticoids in children >6 months old, but this recommendation remains controversial and is not universally considered the standard of care. Glucocorticoids significantly reduce rates of mortality, severe hearing

loss, and neurologic sequelae in adults and should be administered to those with community-acquired bacterial meningitis. If dexamethasone is given to either adults or children, it should be administered before or in conjunction with the first antibiotic dose.

INVASIVE INFECTIONS (EXCLUDING MENINGITIS)

In previously well children with noncritical illness, therapy with a recommended antibiotic should be instigated at the following dosages: penicillin G, 250,000–400,000 units/kg per day (in divided doses 4–6 h apart); cefotaxime, 75–100 mg/d (doses 8 h apart); or ceftriaxone, 50–75 mg/d (doses 12–24 h apart). For critically ill children, including those who have myocarditis or multilobular pneumonia with hypoxia or hypotension, vancomycin may be added if the isolate may possibly be resistant to β-lactam drugs, with its use reviewed once susceptibility data become available. If the organism is resistant to β-lactam agents, therapy should be modified on the basis of clinical response and susceptibility to other antibiotics. Clindamycin or vancomycin can be used as a first-line agent for children with severe β-lactam hypersensitivity, but vancomycin should not be continued if the organism is shown to be sensitive to other non-β-lactam antibiotics.

For outpatient management, amoxicillin (1 g every 8 h) provides effective treatment for virtually all cases of pneumococcal pneumonia. Neither cephalosporins nor quinolones, which are far more expensive, offer advantages over amoxicillin. Levofloxacin (500–750 mg/d as a single dose) and moxifloxacin (400 mg/d as a single dose) also are highly likely to be effective in the United States except in patients who come from closed populations where these drugs are used widely or who have themselves been treated recently with a quinolone. Clindamycin (600–1200 mg/d every 6 h) is effective in 90% of cases and azithromycin (500 mg on day 1 followed by 250–500 mg/d) or clarithromycin (500–750 mg/d as a single dose) in 80% of cases. Treatment failure resulting in bacteremic disease due to macrolide-resistant isolates has been amply documented in patients given azithromycin empirically. As noted above, rates of resistance to all these antibiotics are relatively low in some countries and much higher in others; high-dose amoxicillin remains the best option worldwide.

The optimal duration of treatment for pneumococcal pneumonia is uncertain, but its continuation for at least 5 days once the patient becomes afebrile appears to be a prudent approach. Cases with a second focus of infection (e.g., empyema or septic arthritis) require longer therapy.

ACUTE OTITIS MEDIA

Amoxicillin (80–90 mg/kg per day) is recommended for children with acute otitis media except in situations where observation and symptom-based treatment without antibiotics are advocated. These situations include nonsevere illness and an uncertain diagnosis in children 6 months to 2 years of age and nonsevere illness (even if the diagnosis seems certain) in children >2 years of age. Although the optimal duration of therapy has not been conclusively established, a 10-day course is recommended for younger children and for children with severe disease at any age. For children >6 years old who have mild or moderate disease, a course of 5–7 days is considered adequate. Patients whose illness fails to respond should be reassessed at 48–72 h. If acute otitis media is confirmed and antibiotic treatment has not been started, administration of amoxicillin should be commenced. If antibiotic therapy fails, a change is indicated. Failure to respond to second-line antibiotics as well indicates that myringotomy or tympanocentesis may need to be undertaken in order to obtain samples for culture.

The above recommendations can also be followed for the treatment of sinusitis. Detailed information on the further management of these conditions in children has been published by the American Academy of Pediatrics and the American Academy of Family Physicians.

PREVENTION

Measures to prevent pneumococcal disease include vaccination against *S. pneumoniae* and influenza viruses, reduction of comorbidities that increase the risk of pneumococcal disease, and prevention of antibiotic overuse, which fuels pneumococcal resistance.

Capsular Polysaccharide Vaccines

The 23-valent pneumococcal polysaccharide vaccine (PPSV23), containing 25 μg of each capsular polysaccharide, has been licensed for use since 1983. Recommendations for its use vary by country. The U.S. Advisory Committee on Immunization Practices recommends PPSV23 for all persons ≥65 years of age and for those 2–64 years of age who have underlying medical conditions that put them at increased risk for pneumococcal disease or, if infected, disease of increased severity (Table 141-1; see also *www.cdc.gov/vaccines/schedules*). The committee updated their recommendations to include the combined use of PPSV23 and pneumococcal conjugate vaccine in at-risk individuals (see "Polysaccharide–Protein Conjugate Vaccines," below). Revaccination 5 years after the first dose is recommended for persons >2 years of age who have underlying medical conditions but not routinely for those whose only indication is an age of ≥65 years. PPSV23 does not induce an anamnestic response, and antibody concentrations wane over time; thus revaccination is particularly important for individuals with conditions resulting in loss of antibody. Concerns about repeated revaccination have focused on safety (i.e., local reactions) and the induction of immune hyporesponsiveness. Neither the clinical relevance nor the biologic basis of hyporesponsiveness is clear, but, given the possibility of its occurrence, more than one revaccination has not been recommended.

The effectiveness of PPSV23 against IPD, pneumococcal pneumonia, all-cause pneumonia, and death is controversial, with wide variation in observations. The many published meta-analyses of PPSV efficacy have often reached opposing conclusions with regard to a given clinical entity.

Generally, observational studies cite greater effectiveness than do controlled clinical trials. The consensus is that PPSV is effective against IPD but is less effective or ineffective against nonbacteremic pneumococcal pneumonia. However, the results of some published trials, observational studies, and meta-analyses contradict this view. Effectiveness is often lower in the elderly and in immunodeficient patients whose condition is associated with reduced antibody responses to vaccines than in younger, healthier populations. When PPSV is effective, the duration of protection following a single dose of vaccine is estimated to be ~5 years.

What is not disputed is that improved pneumococcal vaccines are needed for adults. Even in the setting of routine pneumococcal conjugate vaccination of infants (which indirectly protects adults from vaccine-serotype strains), disease caused by serotypes not represented in the conjugate vaccine continues to be a significant burden among adults.

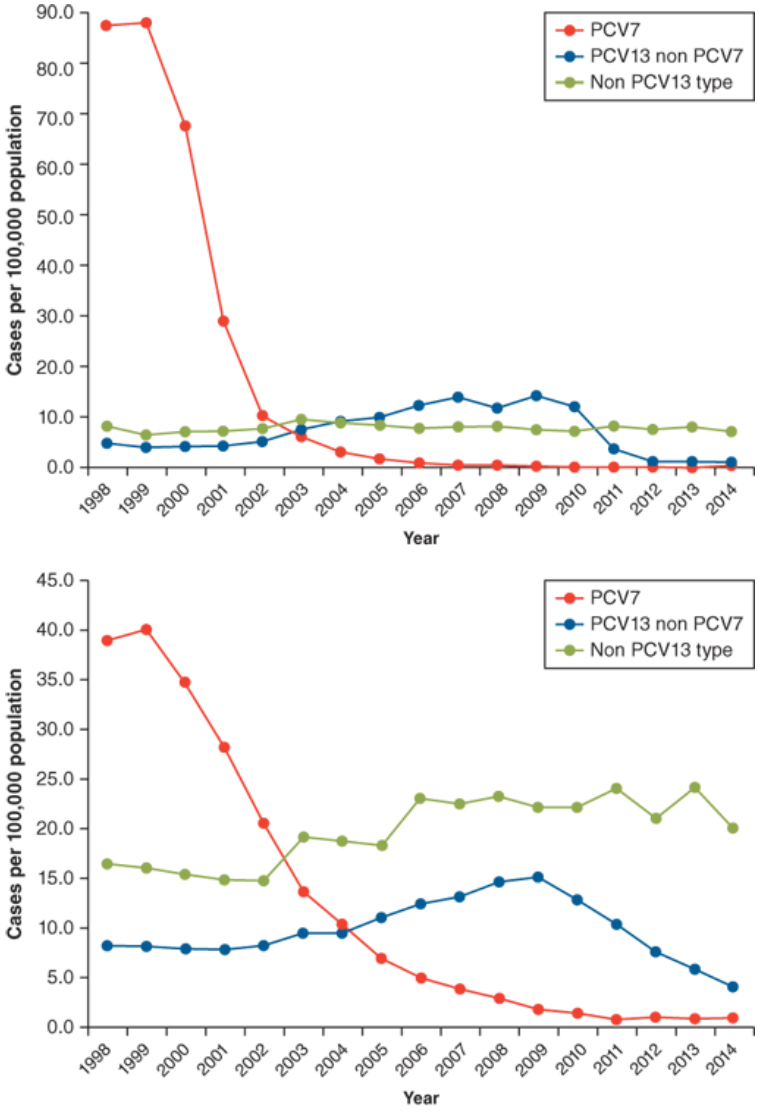
Polysaccharide–Protein Conjugate Vaccines

Infants and young children respond poorly to PPSV, which contains T cell–independent antigens. Consequently, another class of pneumococcal vaccines, the PCVs, were developed specifically for infants and young children. The first product, a 7-valent PCV, was licensed in 2000 in the United States. Two PCV products—containing 10 and 13 serotypes, respectively—are currently (2017) commercially available. The serotypes included in these PCV formulations are important causes of IPD and antibiotic resistance among young children. Randomized controlled trials have demonstrated a high degree of efficacy of PCVs against vaccine-serotype IPD as well as efficacy against pneumonia, otitis media, nasopharyngeal colonization, and all-cause mortality. PCVs are recommended by the World Health Organization for inclusion in routine childhood immunization schedules worldwide, especially in countries with high infant mortality rates.

The United States was the first country to introduce PCV (in 2000) and therefore has the longest experience with its community-wide effects. The introduction of PCV in the United States has resulted in a >90% reduction in vaccine-serotype IPD among the whole population **(Fig. 141-7)**. This decline has been noted not only in those age groups immunized but also in adults and is attributable to the near elimination of vaccine-serotype nasopharyngeal colonization in immunized infants, which reduces spread to adults. This protection of unimmunized community members through vaccination of a subset of the community is termed *the indirect effect*. Increases in colonization with—and concomitantly in disease due to—non-vaccine-serotype strains (i.e., replacement colonization and disease) have been seen; however, the absolute rate increases in IPD caused by non-vaccine serotypes are generally small, especially relative to decreases in vaccine-serotype IPD (see "Epidemiology," above). Since vaccine-serotype strains are more commonly resistant to antibiotics than are non-vaccine serotypes, use of PCV has also resulted in substantial declines in the proportion and absolute rates of drug-resistant pneumococcal disease. The recommendations of the Advisory Committee on Immunization Practices for the use of conjugate vaccines can be found at *www.cdc.gov/vaccines/hcp/acip-recs/vacc-specific/pneumo.html*. PCV has been shown to prevent pneumococcal infection in HIV-infected adults. In the United States, PCV13 followed by a dose of PPSV23 is now recommended for all immunocompromised children and adults. PCV13 is also effective in preventing vaccine sero type pneumococcal pneumonia in healthy adults over 65 years of age and is recommended for routine use in that age group (followed by a dose of PPSV23) in the United States (*www.cdc.gov/mmwr/preview/mmwrhtml/mm6337a4.htm*). This recent policy change will be reviewed in 2018 in light of ongoing and rapid changes in pneumococcal epidemiology with widespread PCV13 use in infants.

FIGURE 141-7

Changes in invasive pneumococcal disease incidence, by serotype group, among children <5 years old (top) and adult ≥65 years old (bottom), 1998–2014. 7-Valent pneumococcal conjugate vaccine (PCV7) was introduced in the United States for routine administration to infants and young children during the second half of 2000, while PCV13 was introduced in 2010. PCV7 includes serotypes 4, 6B, 9V, 14, 18C, 19F, and 23F as well as cross-reactive serotype 6A. PCV13 includes the PCV7 serotypes as well as serotypes 1, 3, 5, 6A, 7F, and 19A. *(Reprinted with permission from Dr. T. Pilishvili, Centers for Disease Control and Prevention.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Other Prevention Strategies

Pneumococcal disease can be averted through the prevention of illnesses that predispose individuals to pneumococcal infections. Relevant measures include influenza vaccination and improved management and control of diabetes, HIV infection, heart disease, and lung disease. Finally, the reduction of antibiotic misuse is a strategy for the prevention of pneumococcal disease in that antimicrobial resistance directly and indirectly perpetuates organism transmission and disease in the community.

GLOBAL HEALTH

Pneumococcal infections are estimated to cause ~330,000 annual deaths worldwide among children 1–59 months of age, accounting for 11% of the 3.2 million all-cause deaths in this age group in 2015. Reliable estimates of adult cases and deaths globally are more difficult to establish because of limited data from parts of the world where most disease occurs. Rates of pneumococcal disease and mortality vary substantially across geographic settings, with the highest rates in selected countries of sub-Saharan Africa and southern Asia, where risk factors for pneumococcal disease—including HIV infection, lack of breast feeding of infants and children, malnutrition, sickle cell disease, and limited access to medical care—are prevalent. Serotypes causing disease exhibit some heterogeneity across geographic settings, but a small number of serotypes universally account for the preponderance of disease in the absence of vaccination; accordingly, vaccine development and vaccination programs are globally relevant. Reductions in disease from pneumococcal infections are anchored in prevention through the inclusion of pneumococcal vaccines in infant immunization programs, timely assessment and appropriate treatment of persons with pneumococcal infections, and reduction of risk factors for pneumococcal disease. The availability of vaccines for the prevention of adult pneumococcal disease, particularly among the elderly, is currently restricted to high-income countries, with virtually no availability in low-income countries where most cases of disease exist.

FURTHER READING

Krone CL et al: Immunosenescence and pneumococcal disease: An imbalance in host–pathogen interactions. Lancet Respir Med 2:141, 2014.

Mackenzie GA et al: Effect of the introduction of pneumococcal conjugate vaccination on invasive pneumococcal disease in The Gambia: A population-based surveillance study. Lancet Infect Dis 16:703, 2016.

Moore MR et al: Effectiveness of 13-valent pneumococcal conjugate vaccine for prevention of invasive pneumococcal disease in children in the USA: A matched case–control study. Lancet Respir Med 4:399, 2016.

Van Der Poll T, Opal SM: Pathogenesis, treatment, and prevention of pneumococcal pneumonia. Lancet 374:1543, 2009.

World Health Organization: Pneumococcal vaccines: WHO position paper—2012. Wkly Epidemiol Rec 87:129, 2012.

WEBSITES

American Academy of Pediatrics RED BOOK: The report of the Committee on Infectious Diseases: *aapredbook.aappublications.org.*

Pneumococcal Regional Serotype Distribution for Pneumococcal AMC TPP: *www.gavialliance.org/library/documents/amc/tpp-codebook/*.

Chapter 142: Staphylococcal Infections

Franklin D. Lowy

INTRODUCTION

Staphylococcus aureus, the most virulent of the many staphylococcal species, has demonstrated its versatility by remaining a major cause of morbidity and mortality worldwide despite the availability of numerous effective antistaphylococcal antibiotics. *S. aureus* is a pluripotent pathogen, causing disease through both toxin- and non-toxin-mediated mechanisms. It is responsible for numerous nosocomial and community-based infections that range from relatively minor skin and soft tissue infections (SSTIs) to life-threatening systemic infections.

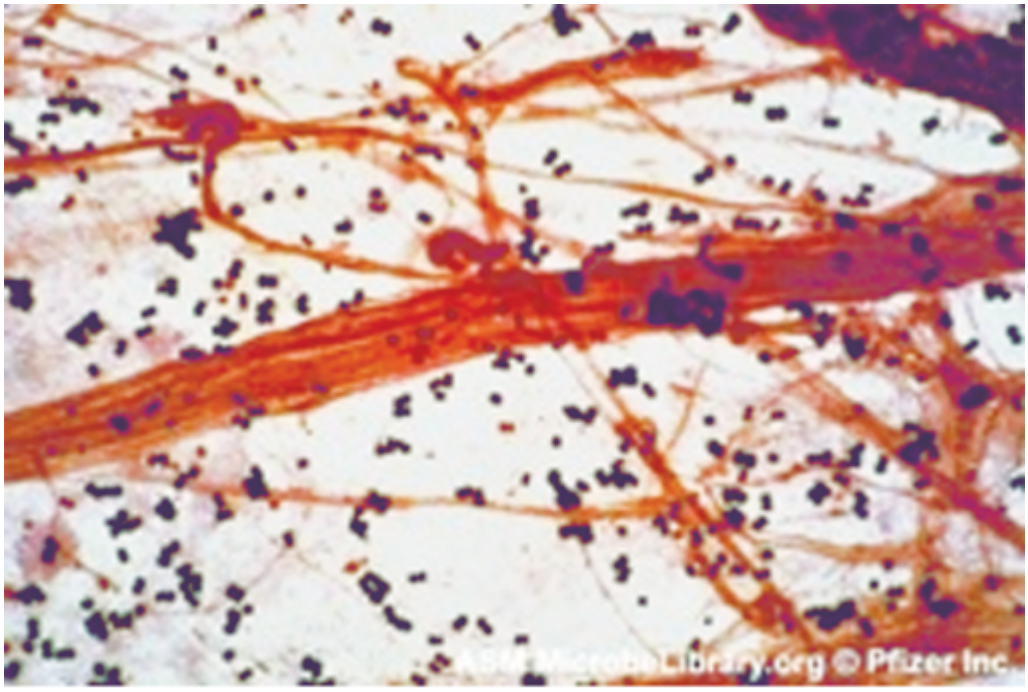
The “other” staphylococci, collectively designated coagulase-negative staphylococci (CoNS), are considerably less virulent than *S. aureus* but remain important pathogens in select settings, such as infections that involve prosthetic devices.

MICROBIOLOGY AND TAXONOMY

Staphylococci, gram-positive cocci in the family Micrococcaceae, form grapelike clusters on Gram’s stain **(Fig. 142-1)**. These organisms (~1 μm in diameter) are catalase-positive (unlike streptococcal species), nonmotile, aerobic, and facultatively anaerobic. They are capable of prolonged survival on environmental surfaces under varying conditions. Some species have a relatively broad host range, including mammals and birds, whereas the host range for others is quite narrow—i.e., limited to one or two closely related animals.

FIGURE 142-1

Gram’s stain of *S. aureus* in a sputum sample, illustrating staphylococcal clusters. *(From ASM MicrobeLibrary.org. © Pfizer, Inc.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

More than 30 staphylococcal species are pathogenic. Identification of the more clinically important species has generally relied on a series of biochemical tests. Automated diagnostic systems, kits for biochemical characterization, and DNA-based assays are available for species identification. With few exceptions, *S. aureus* is distinguished from other staphylococcal species by its production of coagulase, a surface enzyme that converts fibrinogen to fibrin. Latex kits that detect both protein A and clumping factor also distinguish *S. aureus* from most other staphylococcal species. *S. aureus* ferments mannitol, is positive for protein A, and produces DNAse. On blood agar plates, *S. aureus* tends to form golden β-hemolytic colonies; in contrast, CoNS form small white nonhemolytic colonies. Newer methods such as matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF) are increasingly being used for staphylococcal speciation in some clinical microbiology laboratories.

Determining whether multiple staphylococcal isolates from different patients are the same or different is often relevant when there is concern that a nosocomial outbreak is due to a common point source (e.g., a contaminated medical instrument). Molecular typing methods, such as pulsed-field gel electrophoresis and sequence-based techniques (e.g., staphylococcal protein A [SpA] typing), have increasingly been used for this purpose. More recently, whole-genome sequencing has dramatically enhanced the ability to discriminate among clinical isolates.

S. AUREUS INFECTIONS

EPIDEMIOLOGY

S. aureus is both a commensal and an opportunistic pathogen. Approximately 20–40% of healthy persons are colonized with *S. aureus*, with a smaller percentage (~10%) persistently colonized with the same strain. The

rate of colonization is elevated among insulin-dependent diabetics, HIV-infected patients, patients undergoing hemodialysis, injection drug users, and individuals with skin damage. The anterior nares and oropharynx are frequent sites of human colonization, although the skin (especially when damaged), vagina, axilla, and perineum are also often colonized. These colonization sites serve as potential reservoirs for future infections.

Most individuals who develop *S. aureus* infections become infected with a strain that is already a part of their own commensal flora. Breaches of the skin or mucosal membrane allow *S. aureus* to initiate infection. Person-to-person transmission of *S. aureus* also occurs, most frequently from direct personal contact with a colonized body site. Spread of staphylococci in aerosols of respiratory or nasal secretions from heavily colonized individuals also has been reported.

Some diseases increase the risk of *S. aureus* infection; diabetes, for example, combines an increased rate of *S. aureus* colonization and the use of injectable insulin with the possibility of impaired leukocyte function. Individuals with congenital or acquired qualitative or quantitative defects of polymorphonuclear leukocytes (PMNs) are at increased risk of *S. aureus* infections; this group includes neutropenic patients (e.g., those receiving chemotherapeutic agents), those with chronic granulomatous disease, and those with Job (autosomal dominant hyperimmunoglobulin E syndrome) or Chédiak-Higashi syndrome. Other groups at risk include individuals with end-stage renal disease, HIV infection, skin abnormalities, or prosthetic devices.

S. aureus is a leading cause of health care–associated infections **(Chap. 137)**. It is the most common cause of surgical wound infections and is second only to CoNS as a cause of primary bacteremia. These isolates are often resistant to multiple antibiotics; thus available therapeutic options may be limited. In the community, *S. aureus* remains an important cause of SSTIs, respiratory infections, and (among injection drug users) infective endocarditis. The increasing use of home infusion therapy also poses a risk of community-acquired staphylococcal infections.

 In the past three decades, there has been a dramatic change in the epidemiology of infections due to methicillin-resistant *S. aureus* (MRSA). In addition to its major role as a nosocomial pathogen, MRSA has become an established community-based pathogen. Numerous outbreaks of community-associated MRSA (CA-MRSA) infections have been reported in both rural and urban settings in widely separated regions throughout the world. This trend appears to be due in part to the dramatic increase in MRSA colonization found in the community in different parts of the world. Outbreaks of CA-MRSA infections have occurred among such diverse groups as children, prisoners, athletes, Native Americans, and drug users. Risk factors common to these outbreaks include poor hygienic conditions, close contact, contaminated material, and damaged skin. In some geographic regions of the world, the infections have been caused by a single CA-MRSA strain, while in others a variety of CA-MRSA strains have been responsible. In the United States, strain USA300 (defined by pulsed-field gel electrophoresis) has been the predominant clone. Although the majority of infections caused by these strains have involved the skin and soft tissue, 5–10% have been invasive and potentially life-threatening. CA-MRSA strains have also been responsible for an increasing number of

nosocomial infections. Of concern has been the apparently enhanced capacity of CA-MRSA to cause disease in immunocompetent individuals.

PATHOGENESIS

General Concepts

S. aureus is a pyogenic pathogen known for its capacity to induce abscess formation at sites of both local and metastatic infections. This classic pathologic response to *S. aureus* defines the framework within which the infection will progress. The bacteria elicit an inflammatory response characterized by an initial intense infiltration of PMNs and a subsequent infiltration of macrophages and fibroblasts. Either the host cellular response (including the deposition of fibrin and collagen) contains the infection, or infection spreads to the adjoining tissue or into the bloodstream.

In toxin-mediated staphylococcal disease, infection is not invariably present. For example, once the heat-stable enterotoxin has been elaborated into food, staphylococcal food poisoning can develop in the absence of viable bacteria. In staphylococcal toxic shock syndrome (TSS), conditions allowing toxin elaboration at colonization sites (e.g., the presence of a superabsorbent tampon) suffice for initiation of clinical illness.

The *S. aureus* Genome


The complete genomes of *S. aureus* strains are now readily sequenced. Among the interesting revelations are (1) the high degree of nucleotide sequence similarity of the core genomes of different strains; (2) the acquisition of a relatively large amount of genetic information by horizontal transfer from other bacterial species; and (3) the presence of unique “pathogenicity” or “genomic” islands—mobile genetic elements that contain clusters of enterotoxin and exotoxin genes and/or antimicrobial resistance determinants. Among the genes in these islands is *mecA*, the gene responsible for methicillin resistance. Methicillin resistance–containing islands have been designated staphylococcal cassette chromosome *mec* (SCC*mec*) types and range in size from ~20 to 60 kb. Among the more common SCC*mec* types, types 1–3 are traditionally associated with nosocomial MRSA isolates, whereas types 4–6 have been associated with epidemic CA-MRSA strains.


A relatively limited number of MRSA clones have been responsible for most community- and hospital-associated infections worldwide. A comparison of these strains with those from earlier outbreaks (e.g., the phage 80/81 strains from the 1950s) has revealed preservation of the nucleotide sequence over time. This observation suggests that these strains possess determinants that facilitate survival and spread.

Regulation of Virulence Gene Expression

 In both toxin-mediated and non-toxin-mediated diseases due to *S. aureus*, the expression of virulence determinants associated with infection depends on a series of regulatory genes (e.g., accessory gene regulator [*agr*] and staphylococcal accessory regulator [*sar*]) that coordinately control the expression of many virulence genes. The regulatory gene *agr* is part of a quorum-sensing signal transduction pathway that senses and responds to bacterial density. Staphylococcal surface proteins are synthesized during the bacterial exponential growth phase in vitro. In contrast, many secreted proteins, such as α toxin, the enterotoxins, and assorted enzymes, are released during the postexponential growth phase in response to transcription of the effector molecule of *agr*, RNAIII.

It has been hypothesized that these regulatory genes serve a similar function in vivo. Successful invasion requires the sequential expression of these different bacterial elements. Bacterial adhesins are needed to initiate colonization of host tissue surfaces. The subsequent release of various enzymes enables the colony to obtain nutritional support and permits bacteria to spread to adjacent tissues. Studies with strains in which these regulatory genes are inactivated show reduced virulence in several animal models of *S. aureus* infection.

Pathogenesis of Invasive *S. aureus* Infection

Staphylococci are opportunists. For these organisms to invade the host and cause infection, some or all of the following steps are necessary: contamination and colonization of host tissue surfaces, breach of cutaneous or mucosal barriers, establishment of a localized infection, invasion, evasion of the host response, and metastatic spread. Colonizing strains or strains transferred from other individuals are introduced into damaged skin, a wound, or the bloodstream. Recurrences of *S. aureus* infections are common, apparently because of the capacity of these pathogens to survive, to persist in a quiescent state in various tissues, and then to cause recrudescent infections when suitable conditions arise.

S. AUREUS COLONIZATION OF BODY SURFACES

The anterior nares are a primary site of staphylococcal colonization in humans. Colonization appears to involve the attachment of *S. aureus* to keratinized epithelial cells found in the anterior nares. Other factors that contribute to colonization include the influence of other resident nasal flora and their bacterial density, host factors, and nasal mucosal damage (e.g., that resulting from inhalational drug use). Other colonized body sites, such as damaged skin, the groin, and the oropharynx, may be particularly important reservoirs for CA-MRSA strains.

INOCULATION AND COLONIZATION OF TISSUE SURFACES

Staphylococci may be introduced into tissue as a result of minor abrasions (e.g., mosquito bites), administration of medications such as insulin, or establishment of IV access with catheters. After their introduction into a tissue site, bacteria replicate and colonize the host tissue surface. A family of structurally related *S. aureus* surface proteins referred to as MSCRAMMs (microbial surface components recognizing adhesive matrix molecules) plays an important role in mediating adherence to these different sites. By adhering to exposed matrix molecules (e.g., fibrinogen, fibronectin), MSCRAMMs, such as clumping factor and collagen-binding protein, enable the bacteria to colonize different host tissue surfaces; these proteins

contribute to the pathogenesis of invasive infections such as endocarditis and septic arthritis by facilitating the adherence of *S. aureus* to surfaces with exposed fibrinogen or collagen.

Although CoNS are classically known for their ability to elaborate biofilms and to colonize prosthetic devices, *S. aureus* also possesses the genes responsible for biofilm formation, such as the intercellular adhesion (*ica*) locus. Binding to these devices occurs in a stepwise fashion, involving staphylococcal adherence to serum constituents that have coated the device surface and subsequent biofilm elaboration. *S. aureus* is thus a frequent cause of biomedical device–related infections.

INVASION

After colonization, staphylococci replicate at the initial site of infection, elaborating enzymes that include serine proteases, hyaluronidases, thermonucleases, and lipases. These enzymes facilitate bacterial survival and local spread across tissue surfaces. The lipases may facilitate survival in lipid-rich areas such as the hair follicles, where *S. aureus* infections are often initiated. MSCRAMMs also appear to play an important role in the ability of *S. aureus* to spread and cause disease at other tissue sites.

Constitutional findings may result from either localized or systemic infections. The staphylococcal cell wall—consisting of alternating *N*-acetyl muramic acid and *N*-acetyl glucosamine units in combination with an additional cell wall component, lipoteichoic acid—can initiate an inflammatory response that includes the sepsis syndrome. Staphylococcal α toxin is a critical staphylococcal toxin. It causes pore formation in various eukaryotic cells and can also initiate an inflammatory response with findings suggestive of sepsis. The *S. aureus* toxin Panton-Valentine leukocidin is cytolytic to PMNs, macrophages, and monocytes. Strains elaborating this toxin have been epidemiologically linked with cutaneous and more serious infections caused by strains of CA-MRSA.

EVASION OF HOST DEFENSE MECHANISMS

Staphylococci have many immune evasion strategies that are crucial to their survival. They possess an antiphagocytic polysaccharide microcapsule. Most human *S. aureus* infections are due to strains with capsular types 5 and 8. The zwitterionic (both negatively and positively charged) *S. aureus* capsule also plays a critical role in the induction of abscess formation. Protein A, an MSCRAMM unique to *S. aureus*, acts as an Fc receptor, binding the Fc portion of IgG subclasses 1, 2, and 4 and preventing opsonophagocytosis by PMNs. Both chemotaxis inhibitory protein of staphylococci (CHIPS, a secreted protein) and extracellular adherence protein (EAP, a surface protein) interfere with PMN migration to sites of infection.

An additional potential mechanism of *S. aureus* evasion is its capacity for intracellular survival. Both professional and nonprofessional phagocytes internalize staphylococci. Internalization by these cells may provide a sanctuary that protects bacteria against the host's defenses. This phenomenon appears to be especially relevant for hepatic Kupffer cells during staphylococcal bacteremias. The intracellular environment favors the phenotypic expression of *S. aureus* small-colony variants, which are found in patients receiving antimicrobial therapy (e.g., with aminoglycosides) and in those with cystic fibrosis or osteomyelitis. These variants, whether intra- or extracellular, may facilitate prolonged staphylococcal survival in different tissue sites and enhance the likelihood of recurrences. Finally, *S. aureus* can survive within PMNs and may use these cells to spread and seed other tissue sites.

PATHOGENESIS OF COMMUNITY-ACQUIRED MRSA INFECTIONS

A number of virulence determinants have been identified that contribute to the pathogenesis of CA-MRSA infections. A strong epidemiologic association links the presence of the gene for the Panton-Valentine leukocidin with SSTIs and with necrotizing post-influenza pneumonia. Other determinants that play a role in the pathogenesis of these infections include the arginine catabolic mobile element (ACME), a cluster of unique genes that may facilitate evasion of host defense mechanisms; phenol-soluble modulins, a family of cytolytic peptides; and α toxin.

Host Response to *S. aureus* Infection

The primary host response to *S. aureus* infection is the recruitment of PMNs. These cells are attracted to infection sites by bacterial components such as formylated peptides or peptidoglycan as well as by the cytokines tumor necrosis factor (TNF) and interleukins (ILs) 1 and 6, which are released by activated macrophages and endothelial cells.

Although most individuals have antibodies to staphylococci, it is not clear that antibody levels are qualitatively or quantitatively sufficient to protect against infection. Anticapsular and anti-MSCRAMM antibodies facilitate opsonization in vitro and have been protective against infection in several animal models; however, they have not yet successfully prevented staphylococcal infections in clinical trials.

Pathogenesis of Toxin-Mediated Disease

S. aureus produces three types of toxin: cytotoxins, pyrogenic toxin superantigens, and exfoliative toxins. Both epidemiologic data and studies in animals suggest that antitoxin antibodies are protective against illness in TSS, staphylococcal food poisoning, and staphylococcal scalded-skin syndrome (SSSS). Illness develops after toxin synthesis and absorption and the subsequent toxin-initiated host response.

ENTEROTOXIN AND TOXIC SHOCK SYNDROME TOXIN 1 (TSST-1)

The pyrogenic toxin superantigens are a family of small-molecular-size, structurally similar proteins that are responsible for two diseases: TSS and food poisoning. TSS results from the ability of TSST-1 and enterotoxins to function as T cell mitogens. In the normal process of antigen presentation, the antigen is first processed within the cell, and peptides are then presented in the major histocompatibility complex (MHC) class II groove, initiating a measured T cell response. In contrast, enterotoxins bind directly to the invariant region of MHC—outside the MHC class II groove. The enterotoxins can then bind T cell receptors via the vβ chain; this binding results in a dramatic overexpansion of T cell clones (up to 20% of the total T cell population). The consequence of this T cell expansion is a “cytokine storm,” with the release of inflammatory mediators that include interferon γ, IL-1, IL-6, TNF-α, and TNF-β. The resulting multisystem disease produces a constellation of findings that mimic those found in endotoxin shock; however, the pathogenic mechanisms differ. The release of endotoxin from the gastrointestinal tract may synergistically enhance the toxin’s effects.

A different region of the enterotoxin molecule is responsible for the symptoms of food poisoning. The enterotoxins are heat stable and can survive conditions that kill the bacteria. Illness results from the

ingestion of preformed toxin. As a result, the incubation period is short (1–6 h). The toxin stimulates the vagus nerve and the vomiting center of the brain. It also appears to stimulate intestinal peristaltic activity.

EXFOLIATIVE TOXINS AND SSSS

The exfoliative toxins are responsible for SSSS, most commonly seen in newborns. The toxins that produce disease in humans are of two serotypes: ETA and ETB. These toxins are serine proteases, which cleave desmosomal cadherins in the superficial layer of the skin, triggering exfoliation. The result is a split in the epidermis at the granular level, which is responsible for the superficial desquamation of the skin that typifies this illness.

DIAGNOSIS

Staphylococcal infections are readily diagnosed by Gram's stain (Fig. 142-1) and microscopic examination of abscess contents or of infected tissue. Routine culture of infected material usually yields positive results, and blood cultures are sometimes positive even when infections are localized to extravascular sites. *S. aureus* is rarely a blood culture contaminant. Polymerase chain reaction (PCR)–based assays are now often used for the rapid diagnosis of *S. aureus* infection. A number of point-of-care tests are now available to screen patients for colonization with MRSA. Determining whether patients with documented *S. aureus* bacteremia also have infective endocarditis or a metastatic focus of infection remains a diagnostic challenge. Uniformly positive cultures of blood collected over time suggest an endovascular infection such as endocarditis (see "Bacteremia, Sepsis, and Infective Endocarditis," below).

CLINICAL SYNDROMES

(Table 142-1)

TABLE 142-1

Common Illnesses Caused by *Staphylococcus aureus*

Skin and Soft Tissue Infections
Folliculitis
Abscess, furuncle, carbuncle
Cellulitis
Impetigo
Mastitis
Surgical wound infections
Musculoskeletal Infections
Septic arthritis
Osteomyelitis (hematogenous or contiguous spread)
Pyomyositis
Psoas abscess
Respiratory Tract Infections
Ventilator-associated or nosocomial pneumonia
Septic pulmonary emboli
Postviral pneumonia (e.g., influenza)
Empyema
Bacteremia and Its Complications
Sepsis, septic shock
Metastatic foci of infection (kidney, joints, bone, lung)

Infective endocarditis
Infective Endocarditis
Injection drug use–associated
Native-valve
Prosthetic-valve
Nosocomial
Device-Related Infections (e.g., intravascular catheters, prosthetic joints)
Toxin-Mediated Illnesses
Toxic shock syndrome
Food poisoning
Staphylococcal scalded-skin syndrome
Invasive Infections Associated with Community-Acquired Methicillin-Resistant *S. aureus*
Necrotizing fasciitis
Waterhouse-Friderichsen syndrome
Necrotizing pneumonia
Purpura fulminans

Skin and Soft Tissue Infections

S. aureus causes a variety of cutaneous infections, many of which may also be caused by group A streptococci or (less commonly) other streptococcal species. Common factors predisposing to *S. aureus* cutaneous infection include chronic skin conditions (e.g., eczema), skin damage (e.g., insect bites, minor trauma), injections (e.g., in diabetes, injection drug use), and poor personal hygiene. These infections are characterized by the formation of pus-containing blisters, which often begin in hair follicles and spread to adjoining tissues. *Folliculitis* is a superficial infection that involves the hair follicle, with a central area of purulence (pus) surrounded by induration and erythema. *Furuncles* (boils) are more extensive, painful

lesions that tend to occur in hairy, moist regions of the body and extend from the hair follicle to become a true abscess with an area of central purulence. *Carbuncles* are most often located in the lower neck and are even more severe and painful, resulting from the coalescence of other lesions that extend to a deeper layer of the subcutaneous tissue. In general, furuncles and carbuncles are readily apparent, with pus often expressible or discharging from the abscess. Other cutaneous *S. aureus* infections include impetigo and cellulitis. *S. aureus* is one of the most common causes of surgical wound infection.

Mastitis develops in 1–3% of nursing mothers. This infection of the breast, which generally presents within 2–3 weeks after delivery, is characterized by findings that range from cellulitis to abscess formation. Systemic signs, such as fever and chills, are often present in more severe cases.

Musculoskeletal Infections

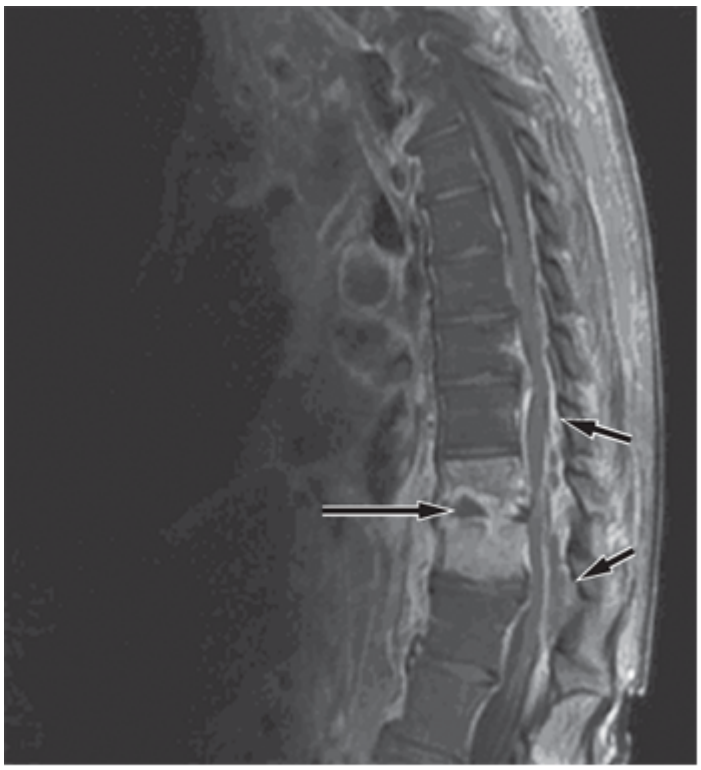
S. aureus is among the most common causes of bone infections—both those resulting from hematogenous dissemination and those arising from contiguous spread from a soft tissue site. *Hematogenous osteomyelitis* in children most often involves the long bones. Infections present with fever and bone pain or with a child's reluctance to bear weight. The white blood cell count and erythrocyte sedimentation rate are often elevated. Blood cultures are positive in ~50% of cases. When necessary, bone biopsies for culture and histopathologic examination are usually diagnostic.

In adults, hematogenous osteomyelitis involving the long bones is less common. However, *vertebral osteomyelitis* is among the more common clinical presentations. Vertebral bone infections are most often seen in patients with endocarditis, those undergoing hemodialysis, diabetics, and injection drug users. These infections may present as intense back pain with fever but may also be clinically occult, presenting as chronic back pain with low-grade fever. *S. aureus* is the most common cause of epidural abscess, a complication that can result in neurologic compromise. Patients report difficulty voiding or walking and radicular pain in addition to the symptoms associated with their osteomyelitis. Surgical intervention in this setting often constitutes a medical emergency.

MRI is the most reliable imaging modality to help establish the diagnosis of osteomyelitis. **(Fig. 142-2)**. Routine x-rays are an appropriate first step, but findings may be normal for up to 14 days after the onset of symptoms. If an MRI is not possible, CT is an acceptable alternative.

FIGURE 142-2

***S. aureus* vertebral osteomyelitis and epidural abscess involving the thoracic disk between T9 and T10.**
Sagittal post-contrast MRI of the spine illustrates destruction of the T9–T10 intervertebral space with enhancement (*long arrow*). There is impingement on the thoracic cord and an epidural collection extending from T9 through T11 (*short arrows*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Bone infections that result from contiguous spread tend to develop from soft tissue infections, such as those associated with diabetic or vascular ulcers, surgery, or trauma. Exposure of bone, a draining fistulous tract, failure to heal, or continued drainage suggests involvement of underlying bone. Bone involvement is established by bone culture and histopathologic examination (revealing evidence of PMN infiltration). Contamination of culture material from adjacent tissue can make the diagnosis of osteomyelitis difficult in the absence of pathologic confirmation. Samples obtained during surgery are more reliable. In addition, it is sometimes hard to distinguish radiologically between osteomyelitis and overlying soft tissue infection with underlying osteitis.

In both children and adults, *S. aureus* is the most common cause of *septic arthritis* in native joints. This infection is rapidly progressive and may be associated with extensive joint destruction if left untreated. It presents with intense pain on motion of the affected joint, swelling, and fever. Aspiration of the joint reveals turbid fluid, with >50,000 PMNs/μL and gram-positive cocci in clusters on Gram's stain (Fig. 142-1). In adults, septic arthritis may result from trauma, surgery, or hematogenous dissemination. The most commonly involved joints include the knees, shoulders, hips, and phalanges. Infection frequently develops in joints previously damaged by osteoarthritis or rheumatoid arthritis. Iatrogenic infections resulting from aspiration or injection of agents into the joint also occur. In these settings, the patient experiences increased pain and swelling in the involved joint in association with fever.

Pyomyositis is an unusual infection of skeletal muscles that is seen primarily in tropical climates but also occurs in immunocompromised and HIV-infected patients. It is believed to arise from occult bacteremia. Pyomyositis presents as fever, swelling, and pain overlying the involved muscle. Aspiration of fluid from the

involved tissue yields pus. Although a history of trauma may be associated with the infection, its pathogenesis is poorly understood.

Respiratory Tract Infections

Respiratory tract infections caused by *S. aureus* occur in selected clinical settings. *S. aureus* is a cause of serious respiratory tract infections in newborns and infants; these infections present with shortness of breath, fever, and respiratory failure. Chest x-ray may reveal pneumatoceles (shaggy, thin-walled cavities). Pneumothorax and empyema are recognized complications.

In adults, nosocomial *S. aureus* pulmonary infections are common among intubated patients in intensive care units. Nasally colonized patients are at increased risk of these infections. The clinical presentation is no different from that encountered in pulmonary infections caused by other bacterial pathogens. Patients produce increased volumes of purulent sputum and develop respiratory distress, fever, and new pulmonary infiltrates. Distinguishing bacterial pneumonia from respiratory failure or other causes of new pulmonary infiltrates in critically ill patients is difficult and relies on a constellation of clinical, radiologic, and laboratory findings.

Community-acquired respiratory tract infections due to *S. aureus* often follow viral infections—most commonly influenza. Patients may present with fever, bloody sputum production, and midlung-field pneumatoceles or multiple, patchy pulmonary infiltrates. Diagnosis is made by sputum Gram's stain and culture. Blood cultures, although useful, are usually negative.

Bacteremia, Sepsis, and Infective Endocarditis

S. aureus bacteremia may be complicated by sepsis, endocarditis, vasculitis, or metastatic seeding (establishment of suppurative collections at other tissue sites). Among the more commonly seeded tissue sites are bones, joints, kidneys, and lungs. The frequency of metastatic seeding during bacteremia has been estimated to be as high as 31%. The incidence of complications increases with the duration of the bacteremia.

Recognition of these complications by clinical and laboratory diagnostic methods alone is often difficult. Comorbid conditions that are frequently seen in association with *S. aureus* bacteremia and that increase the risk of complications include diabetes, HIV infection, and renal insufficiency. Other host factors associated with an increased risk of complications include presentation with community-acquired *S. aureus* bacteremia (except in injection drug users), lack of an identifiable primary focus of infection, and the presence of prosthetic devices or material.

Clinically, *S. aureus* sepsis presents in a manner similar to that documented for sepsis due to other bacteria. The well-described progression of hemodynamic changes—beginning with respiratory alkalosis and clinical findings of hypotension and fever—is commonly seen. The microbiologic diagnosis is established by positive blood cultures.

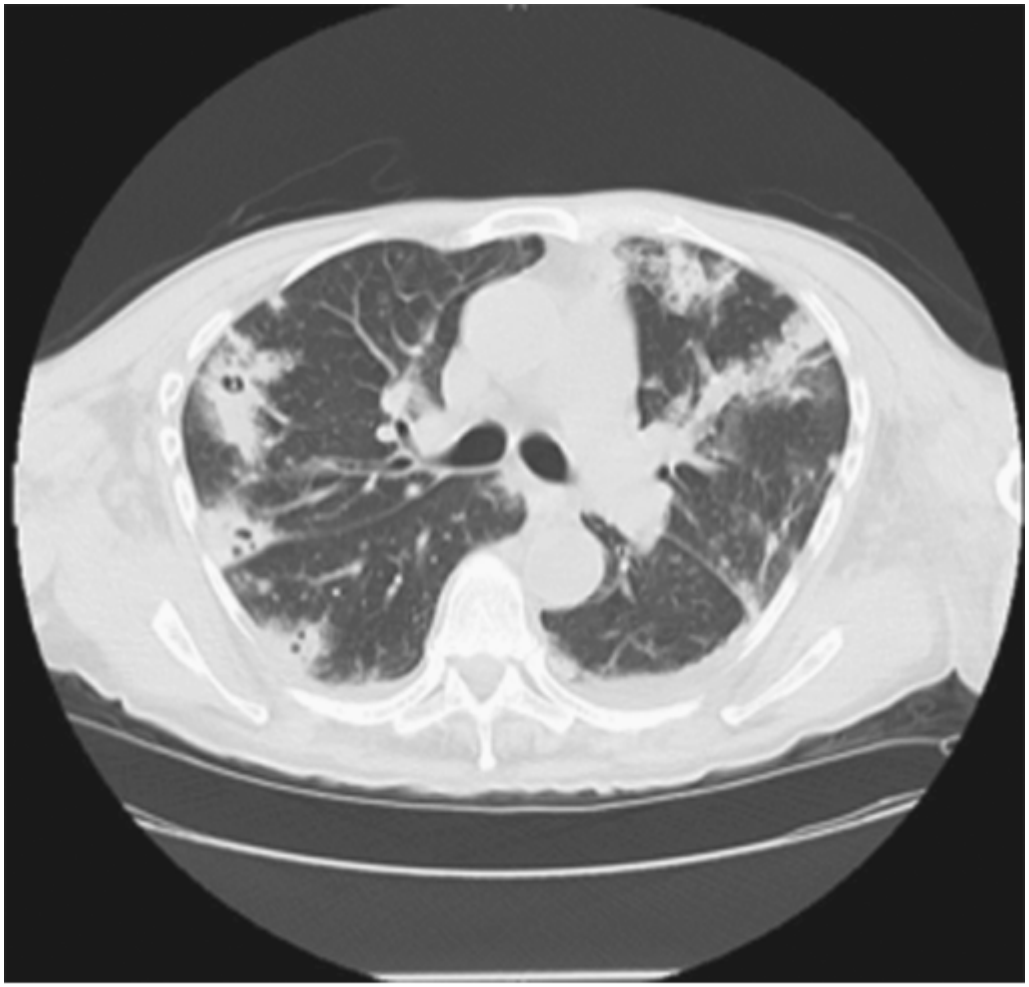
The overall incidence of *S. aureus* endocarditis has increased over the past 20 years. *S. aureus* is now the leading cause of endocarditis worldwide, accounting for 25–35% of cases. This increase is due, at least in part, to the increased use of intravascular devices. Studies using transesophageal echocardiography found an endocarditis incidence of ~25% among patients with intravascular catheter–associated *S. aureus* bacteremia. Other factors associated with an increased risk of endocarditis are injection drug use, hemodialysis, the presence of intravascular prosthetic devices at the time of bacteremia, and immunosuppression. Patients with implantable cardiac devices (e.g., permanent pacemakers) are at increased risk of endocarditis or device-related infections. Despite the availability of effective antibiotics, mortality rates from these infections continue to range from 20% to 40%, depending on both the host and the nature of the infection. Complications of *S. aureus* endocarditis include cardiac valvular insufficiency, peripheral emboli, metastatic seeding, vasculitis, and central nervous system (CNS) involvement (e.g., mycotic aneurysms, embolic strokes).

S. aureus endocarditis is encountered in four clinical settings: (1) right-sided endocarditis in association with injection drug use, (2) left-sided native-valve endocarditis, (3) prosthetic-valve endocarditis, and (4) nosocomial endocarditis. In each of these settings, the diagnosis is suspected from the patient's history and the recognition of clinical stigmata suggestive of endocarditis. These findings include cardiac manifestations, such as new or changing cardiac valvular murmurs; cutaneous evidence, such as vasculitic lesions, Osler's nodes, or Janeway lesions; evidence of right- or left-sided embolic disease; and a history suggesting a risk for *S. aureus* bacteremia. In the absence of antecedent antibiotic therapy, blood cultures are almost uniformly positive. Transthoracic echocardiography, while less sensitive than transesophageal echocardiography, is less invasive and identifies valvular vegetations. The Duke criteria **(see Table 123-3)** are now commonly used to help establish the likelihood of this diagnosis.

Acute right-sided tricuspid valvular *S. aureus* endocarditis is most often seen in injection drug users. The classic presentation includes a high fever, a toxic clinical appearance, pleuritic chest pain, and the production of purulent (sometimes bloody) sputum. Chest x-rays or CT scans reveal evidence of septic pulmonary emboli (small, peripheral, circular lesions that may cavitate with time) **(Fig. 142-3)**. A high percentage of affected patients have no history of antecedent valvular damage. At the outset of their illness, patients may present with fever alone, without cardiac or other localizing findings. As a result, a high index of clinical suspicion is essential for diagnosis.

FIGURE 142-3

CT scan illustrating septic pulmonary emboli in a patient with methicillin-resistant *Staphylococcus aureus* bacteremia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Individuals with antecedent cardiac valvular damage more commonly present with left-sided native-valve endocarditis involving the damaged valve. These patients tend to be older than those with right-sided endocarditis, their prognosis is worse, and their incidence of complications (including peripheral emboli, cardiac decompensation, and metastatic seeding) is higher.

S. aureus is one of the more common causes of prosthetic-valve endocarditis. This infection is especially fulminant in the early postoperative period and is associated with a high mortality rate. In most instances, medical therapy alone is not sufficient and urgent valve replacement is necessary. Patients are prone to develop valvular insufficiency or myocardial abscesses originating from the region of valve implantation.

The increased frequency of nosocomial endocarditis (15–30% of cases, depending on the series) reflects in part the increased use of intravascular devices. This form of endocarditis is most commonly caused by *S. aureus*. Because patients often are critically ill, are receiving antibiotics for various other indications, and have comorbid conditions, the diagnosis is often missed.

Prosthetic Device–Related Infections

S. aureus accounts for a large proportion of prosthetic device–related infections. These infections may involve intravascular catheters, prosthetic valves, orthopedic devices, peritoneal catheters, pacemakers, left-

ventricular-assist devices, or vascular grafts. In contrast with the more indolent presentation of CoNS infections, *S. aureus* device-related infections are more acute, have both local and systemic manifestations, and tend to progress more rapidly. It is relatively common for a pyogenic collection to be present at the device site. Aspiration of these collections and performance of blood cultures are important components in establishing a diagnosis. *S. aureus* infections tend to occur more commonly soon after implantation unless the device is used for access (e.g., intravascular or hemodialysis catheters). In the latter instance, infections can occur at any time. As in most prosthetic-device infections, successful therapy usually involves removal of the device. Left in place, the device serves as a potential nidus for either persistent or recurrent infections.

Urinary Tract Infections

Urinary tract infections (UTIs) are infrequently caused by *S. aureus*. The presence of *S. aureus* in the urine generally suggests hematogenous dissemination. Ascending *S. aureus* infections occasionally result from instrumentation of the genitourinary tract.

Infections Associated with Community-Acquired MRSA

Although the skin and soft tissues are by far the most common sites of infection associated with CA-MRSA, 5–10% of these infections are invasive and can be life-threatening. The latter unique infections, including necrotizing fasciitis, necrotizing pneumonia, and sepsis with Waterhouse-Friderichsen syndrome or purpura fulminans, were rarely associated with *S. aureus* prior to the emergence of CA-MRSA. These life-threatening infections reflect the increased virulence of CA-MRSA strains.

Toxin-Mediated Diseases

FOOD POISONING

S. aureus is among the most common causes of foodborne outbreaks in the United States. Staphylococcal food poisoning results from the inoculation of toxin-producing *S. aureus* into food by colonized food handlers. Toxin is then elaborated in such growth-promoting food as custards, potato salad, or processed meats. Even if the bacteria are killed by warming, the heat-stable toxin is not destroyed. The onset of illness is rapid, occurring within 1–6 h of ingestion. The illness is characterized by nausea and vomiting, although diarrhea, hypotension, and dehydration also may occur. The differential diagnosis includes diarrhea of other etiologies, especially that caused by similar toxins (e.g., the toxins elaborated by *Bacillus cereus*). The rapidity of onset, the absence of fever, and the epidemic nature of the presentation (without secondary spread) should arouse suspicion of staphylococcal food poisoning. Symptoms generally resolve within 8–10 h. The diagnosis can be established by the demonstration of bacteria or the documentation of enterotoxin in the implicated food. Treatment is entirely supportive.

TOXIC SHOCK SYNDROME

TSS gained attention in the early 1980s, when a nationwide outbreak occurred in the United States among young, otherwise healthy, menstruating women. Epidemiologic investigation demonstrated that these cases were associated with the use of a highly absorbent tampon that had recently been introduced to the market. Subsequent studies established the role of TSST-1 in these illnesses. Withdrawal of the tampon from the

market resulted in a rapid decline in the incidence of this disease. However, menstrual and nonmenstrual cases continue to be reported. Nonmenstrual cases are frequently seen in patients with surgical or postpartum wound infections, especially when packing of the wound occurs.

The clinical presentation is similar in menstrual and nonmenstrual TSS. Evidence of clinical *S. aureus* infection is not a prerequisite. TSS results from the elaboration of an enterotoxin or the structurally related enterotoxin-like TSST-1. More than 90% of menstrual cases are caused by TSST-1, whereas a high percentage of nonmenstrual cases are caused by enterotoxins (e.g., enterotoxin B). TSS begins with relatively nonspecific flulike symptoms. In menstrual cases, the onset usually comes 2 or 3 days after the start of menstruation. Patients present with fever, hypotension, and erythroderma of variable intensity. Mucosal involvement is common (e.g., conjunctival hyperemia). The illness can rapidly progress to symptoms that include vomiting, diarrhea, confusion, myalgias, and abdominal pain. These symptoms reflect the multisystemic nature of the disease, with involvement of the liver, kidneys, gastrointestinal tract, and/or CNS. Desquamation of the skin occurs during convalescence, usually 1–2 weeks after the onset of illness. Laboratory findings may include azotemia, leukocytosis, hypoalbuminemia, thrombocytopenia, and liver function abnormalities.

Diagnosis of TSS still depends on a constellation of findings rather than one specific finding and on a lack of evidence of other possible infections **(Table 142-2)**. Other diagnoses to be considered are drug toxicities, viral exanthems, Rocky Mountain spotted fever, sepsis, and Kawasaki disease. Illness occurs only in persons who lack antibody to TSST-1. Recurrences are possible if antibody fails to develop after the illness.

TABLE 142-2

Case Definition of *Staphylococcus aureus* Toxic Shock Syndrome

Clinical Criteria
An illness with the following clinical manifestations:
• Fever: temperature ≥102.0°F (≥38.9°C) • Rash: diffuse macular erythroderma • Desquamation: 1–2 weeks after rash onset • Hypotension: systolic blood pressure ≤90 mmHg for adults or less than the fifth percentile, by age, for children <16 years old • Multisystem involvement (≥3 of the following organ systems) ◦ Gastrointestinal: vomiting or diarrhea at illness onset ◦ Muscular: severe myalgia or creatine phosphokinase level at least twice ULN ◦ Mucous membrane: vaginal, oropharyngeal, or conjunctival hyperemia ◦ Renal: blood urea nitrogen or creatinine level at least twice ULN for laboratory or urinary sediment with pyuria (≥5 leukocytes per high-power field) in the absence of urinary tract infection ◦ Hepatic: total bilirubin or aminotransferase level at least twice ULN for laboratory ◦ Hematologic: platelet count $<10^5/\mu L$ ◦ Central nervous system: disorientation or alterations in consciousness without focal neurologic signs in the absence of fever and hypotension
Laboratory Criteria
Negative results in the following tests, if obtained: • Blood or cerebrospinal fluid cultures for another pathogen[a] • Serologic tests for Rocky Mountain spotted fever, leptospirosis, or measles
Case Classification
Probable: a case that meets the laboratory criteria and in which four of the five clinical criteria are fulfilled
Confirmed: a case that meets the laboratory criteria and in which all five of the clinical criteria are fulfilled, including desquamation (unless the patient dies before desquamation occurs)

[a]Blood cultures may be positive for *S. aureus*.

Abbreviation: ULN, upper limit of normal.

Source: Centers for Disease Control and Prevention (*https://www.cdc.gov/nndss/conditions/toxic-shock-syndrome-other-than-streptococcal/case-definition/2011/*).

STAPHYLOCOCCAL SCALDED-SKIN SYNDROME

SSSS primarily affects newborns and children. The illness may vary from a localized blister to exfoliation of much of the skin surface. The skin is usually fragile and often tender, with thin-walled, fluid-filled bullae. Gentle pressure results in rupture of the lesions, leaving denuded underlying skin. The mucous membranes are usually spared. In more generalized infection, there are often constitutional symptoms, including fever, lethargy, and irritability with poor feeding. Significant amounts of fluid can be lost in more extensive cases. Illness usually follows localized infection at one of a number of possible sites. SSSS is much less common among adults but can follow infections caused by exfoliative toxin–producing strains.

COAGULASE-NEGATIVE STAPHYLOCOCCAL INFECTIONS

Although considerably less virulent than *S. aureus*, CoNS are among the most common causes of prosthetic-device infections, including endocarditis. They also are increasingly a cause of native-valve endocarditis and life-threatening infections in neonates and in neutropenic patients. Approximately half of the identified CoNS species have been associated with human infections. Of these species, *Staphylococcus epidermidis* is the most common human pathogen. It is part of the normal human flora and is found on the skin (where it is the most abundant bacterial species) as well as in the oropharynx and vagina. *Staphylococcus saprophyticus*, a novobiocin-resistant species, is a common pathogen in UTIs.

PATHOGENESIS

S. epidermidis is the CoNS species most often associated with prosthetic-device infections. Infection is a two-step process, with initial adhesion to the device followed by colonization. *S. epidermidis* is uniquely adapted to colonize these devices because of its capacity to elaborate the extracellular polysaccharide (glycocalyx or slime) that facilitates formation of a protective biofilm on the device surface.

Implanted prosthetic material is rapidly coated with host serum or tissue constituents such as fibrinogen or fibronectin. These molecules serve as potential bridging ligands, facilitating initial bacterial attachment to the device surface. A number of staphylococcal surface-associated proteins, such as autolysin (AtlE), fibrinogen-binding protein, and accumulation-associated protein (AAP), appear to play a role in attachment to either modified or unmodified prosthetic surfaces. The polysaccharide intercellular adhesin facilitates subsequent staphylococcal colonization and accumulation on the device surface. Intercellular adhesin (*ica*) genes are more commonly found in strains of *S. epidermidis* that are associated with device infections than in strains associated with colonization of mucosal surfaces. Biofilm acts as a barrier, protecting bacteria from host defense mechanisms as well as from antibiotics while providing a suitable environment for bacterial maturation, survival, and potentially spread to other tissue sites.

Two additional CoNS species, *Staphylococcus lugdunensis* and *Staphylococcus schleiferi*, produce more serious infections (native-valve endocarditis and osteomyelitis) than do other CoNS. The basis for this

enhanced virulence is not known, although both species appear to share more virulence determinants with *S. aureus* (e.g., clumping factor and lipase) than do other CoNS.

The capacity of *S. saprophyticus* to cause UTIs in young women appears to be related to its enhanced capacity to adhere to uroepithelial cells. A 160-kDa hemagglutinin/adhesin may contribute to this affinity.

DIAGNOSIS

Although the detection of CoNS at sites of infection or in the bloodstream by standard microbiologic culture methods is not difficult, interpretation of these results is frequently problematic. Because these organisms are present in large numbers on the skin, they often contaminate cultures. It has been estimated that only 10–20% of blood cultures positive for CoNS reflect true bacteremia. Similar problems arise with cultures obtained from other sites. Among the clinical findings suggestive of true bacteremia are fever, evidence of local infection (e.g., erythema or purulent drainage at the IV catheter site), leukocytosis, and systemic signs of sepsis. Laboratory findings suggestive of true bacteremia include repeated isolation of the same strain (i.e., the same species with the same antibiogram or with a closely related DNA fingerprint) from separate cultures, growth of the strain within 48 h, and bacterial growth in both aerobic and anaerobic bottles.

CLINICAL SYNDROMES

CoNS cause a diverse array of prosthetic device–related infections, including those that involve prosthetic cardiac valves and joints, vascular grafts, intravascular devices, and CNS shunts. In all of these settings, the clinical presentation is similar. The signs of localized infection are often subtle, the rate of disease progression is slow, and the systemic findings are often limited. Signs of infection, such as purulent drainage, pain at the site, or loosening of prosthetic implants, are sometimes evident. Fever is frequently but not always present, and there may be mild leukocytosis. Acute-phase reactant levels, the erythrocyte sedimentation rate, and the C-reactive protein concentration may be elevated.

Infections that are not associated with prosthetic devices include, as noted, native-valve endocarditis due to CoNS, which accounts for ~5% of cases. Infections in preterm infants and neutropenic patients are often associated with the need for intravascular devices. *S. lugdunensis* appears to be a more aggressive pathogen in this setting, causing greater mortality and rapid valvular destruction with abscess formation.

TREATMENT

TREATMENT

Staphylococcal Infections

GENERAL PRINCIPLES OF THERAPY

Source control (e.g., incision and drainage of suppurative collections or removal of infected prosthetic devices), coupled with rapid institution of appropriate antimicrobial therapy, is essential for the

management of all staphylococcal infections. The emergence of MRSA in the community has increased the importance of culturing all collections in order to determine antimicrobial susceptibility.

DURATION OF ANTIMICROBIAL THERAPY

Therapy for *S. aureus* bacteremia is generally prolonged (4–6 weeks) because of the high risk of complications (e.g., endocarditis, metastatic foci of infection). Among the findings associated with complicated bacteremias are (1) persistently positive blood cultures 96 h after institution of therapy, (2) acquisition of the infection in the community, (3) failure to promptly remove or drain an identified focus of infection (i.e., an intravascular catheter), and (4) the presence of deep-seated infections. Patients with uncomplicated bacteremias are generally defined by a removable focus of infection, prompt response to antimicrobial therapy (i.e., no fever or positive blood cultures after 3–4 days), no evidence of metastatic foci of infection, and no implanted prostheses. In these infections, short-course therapy (2 weeks) can be given. Transesophageal echocardiography to rule out endocarditis is generally necessary because neither clinical nor laboratory findings can reliably detect cardiac involvement. A thorough radiologic investigation to identify potential metastatic collections is also indicated. All symptomatic body sites must be carefully evaluated.

CHOICE OF ANTIMICROBIAL AGENTS

The choice of antimicrobial agents to treat both coagulase-positive and coagulase-negative staphylococcal infections is often problematic because of the prevalence of multidrug-resistant strains and the absence of comparative clinical trials. Staphylococcal resistance to most antibiotic families, including β-lactams, aminoglycosides, fluoroquinolones, and (to a lesser extent) glycopeptides, has increased. This trend is even more apparent with CoNS; >80% of nosocomial isolates are resistant to methicillin, and these methicillin-resistant strains are often resistant to many other antibiotics. Because the selection of antimicrobial agents for *S. aureus* infections is similar to that for CoNS infections, treatment options for these pathogens are discussed together and are summarized in **Table 142-3**.

As a result of the widespread dissemination of plasmids containing the enzyme penicillinase, few strains of staphylococci (≤5%) remain susceptible to penicillin. Penicillin-resistant isolates are treated with semisynthetic penicillinase-resistant penicillins (SPRPs), such as oxacillin or nafcillin. Methicillin, the first of the SPRPs, is no longer used. Cephalosporins are alternative therapeutic agents for these infections. Second- and third-generation cephalosporins offer no therapeutic advantage over first-generation cephalosporins for the treatment of staphylococcal infections, and some third-generation cephalosporins (e.g., ceftazidime) have considerably less activity. The carbapenems also have excellent activity against methicillin-sensitive *S. aureus* but not against MRSA.

The isolation of MRSA was reported within 1 year of the introduction of methicillin. Since then, the prevalence of MRSA has steadily increased. In many U.S. hospitals, 40–50% of *S. aureus* isolates are resistant to methicillin. This trend has also been observed in many other countries. Resistance to methicillin indicates resistance to all SPRPs as well as to all cephalosporins (except ceftaroline). Production of a novel penicillin-binding protein (PBP2a) is responsible for methicillin resistance. This protein is synthesized by the *mecA* gene, which (as stated above) is part of a large mobile genetic element—a pathogenicity or genomic island—

called SCC*mec*. It is hypothesized that this genetic material was acquired via horizontal transfer from *Staphylococcus sciuri*, a related staphylococcal species. Phenotypic expression of methicillin resistance may be constitutive (i.e., expressed in all cells in a population) or heterogeneous (i.e., displayed by only a proportion of the total cell population). Detection of methicillin resistance is enhanced by growth of cultures at reduced temperatures (≤35°C for 24 h) and with increased concentrations of salt in the medium. Culture techniques are increasingly being replaced by PCR-based or other methods (e.g., latex agglutination) that allow the rapid detection of methicillin resistance.

Vancomycin or daptomycin is recommended as the drug of choice for the treatment of invasive MRSA infections. MRSA susceptibility to vancomycin has decreased in many areas of the world. It is important to note that vancomycin is less effective than SPRPs for the treatment of infections due to methicillin-susceptible strains. In patients with a history of serious β-lactam allergies, alternatives to SPRPs for the treatment of invasive infections should be used only after careful consideration. Desensitization to β-lactams remains an option for life-threatening infections.

Three types of staphylococcal resistance to vancomycin have emerged. (1) Minimal inhibitory concentration (MIC; an in vitro measure of susceptibility) “creep” refers to the incremental increase in vancomycin MICs that has been detected in various geographic areas. Studies suggest that morbidity and mortality are increased in infections due to *S. aureus* strains with vancomycin MICs of ≥1.5 µg/mL. (2) In 1997, an *S. aureus* strain with reduced susceptibility to vancomycin (vancomycin-intermediate *S. aureus* [VISA]) was reported from Japan. Subsequently, additional VISA clinical isolates were reported. These strains were resistant to methicillin and many other antimicrobial agents. The VISA strains appear to evolve (under vancomycin selective pressure) from strains that are susceptible to vancomycin but are heterogeneous, with a small proportion of the bacterial population expressing the resistance phenotype. The mechanism of VISA resistance is in part due to an abnormally thick cell wall. Vancomycin is trapped by the abnormal peptidoglycan cross-linking and is unable to gain access to its target site. Regulatory genes involved in cell wall metabolism appear to play an important role in this type of resistance. (3) In 2002, the first clinical isolate of fully vancomycin-resistant *S. aureus* (VRSA) was reported. Resistance in this and several additional clinical isolates was due to the presence of *vanA*, the gene responsible for expression of vancomycin resistance in enterococci. This observation suggested that resistance was acquired as a result of horizontal conjugal transfer from a vancomycin-resistant strain of *Enterococcus faecalis.* Several patients had both MRSA and vancomycin-resistant enterococci cultured from infection sites. The *vanA* gene is responsible for the synthesis of the dipeptide D-Ala-D-Lac in place of D-Ala-D-Ala. Vancomycin cannot bind to the altered peptide. While isolates with MICs of ≥1.5 µg/mL have been relatively common in some areas, VISA and VRSA isolates remain rare.

Daptomycin, a parenteral bactericidal agent with antistaphylococcal activity, is approved for the treatment of bacteremia (including right-sided endocarditis) and complicated skin infections. It is not effective in respiratory infections. This drug has a unique mechanism of action: it disrupts the cytoplasmic membrane. Staphylococcal resistance to daptomycin has been reported. Resistance can emerge during therapy; patients previously treated with vancomycin may have elevated MICs of daptomycin. Patients need to be monitored for rhabdomyolysis with creatine phosphokinase measurement and for eosinophilic pneumonia.

Linezolid—the first oxazolidinone—is bacteriostatic against staphylococci; it offers the advantage of comparable bioavailability after oral or parenteral administration. Cross-resistance with other inhibitors of protein synthesis has not been detected. Resistance to linezolid has been increasingly reported. Serious adverse reactions to linezolid include thrombocytopenia, occasional cases of neutropenia, and rare instances of lactic acidosis or peripheral and optic neuropathy. These reactions tend to occur after a relatively prolonged course of therapy.

Tedizolid, a second oxazolidinone released in 2014, is available as both oral and parenteral preparations. It exhibits enhanced in vitro activity against antibiotic-resistant gram-positive bacteria, including staphylococci. Tedizolid is administered once a day. Data on its efficacy for the treatment of deep-seated infections are limited.

Ceftaroline is a fifth-generation cephalosporin with bactericidal activity against MRSA (including strains with reduced susceptibility to vancomycin and daptomycin). It is generally well tolerated. Ceftaroline is approved for use in nosocomial pneumonias and for SSTIs.

Telavancin is a parenteral lipoglycopeptide derivative of vancomycin that is approved for the treatment of complicated SSTIs and for nosocomial pneumonias. The drug has two targets: the cell wall and the cell membrane. It remains active against VISA strains. Because of its potential nephrotoxicity, telavancin should be avoided in patients with renal disease.

The parenteral streptogramin antibiotic quinupristin/dalfopristin displays bactericidal activity against all staphylococci, including VISA strains. This drug, although infrequently used because of toxicity (infusion reactions), has been used successfully to treat serious MRSA infections. In cases of resistance to erythromycin or clindamycin, quinupristin/dalfopristin is bacteriostatic against staphylococci.

Dalbavancin and oritavancin are long-acting, parenterally administered lipoglycopeptides that have been used to treat complicated SSTIs. Because of their long half-lives, they can be administered on a weekly basis. Both have been used as single-dose regimens for the treatment of SSTIs. Data on their use for the treatment of invasive staphylococcal infections are limited.

Although the quinolones are active against staphylococci in vitro, the frequency of staphylococcal resistance to these agents has increased, especially among methicillin-resistant isolates. Of particular concern in MRSA is the possible emergence of quinolone resistance during therapy. Therefore, quinolones are not recommended for the treatment of MRSA infections. Resistance to the quinolones is most commonly chromosomal and results from mutations of the topoisomerase IV or DNA gyrase genes, although multidrug efflux pumps may also contribute. Although the newer quinolones exhibit increased in vitro activity against staphylococci, it is uncertain whether this increase translates into enhanced in vivo activity.

Tigecycline, a broad-spectrum minocycline analogue, has bacteriostatic activity against MRSA and is approved for use in SSTIs as well as intraabdominal infections caused by *S. aureus.* It is not recommended for the treatment of invasive infections. Other older antibiotics, such as minocycline, doxycycline, clindamycin, and trimethoprim-sulfamethoxazole, have been used successfully to treat MRSA infections.

Combinations of antistaphylococcal agents have been used to enhance bactericidal activity in the treatment of deep-seated infections, to shorten the duration of therapy (e.g., for right-sided endocarditis), or to optimize empirical therapy when the susceptibility of the isolate to methicillin is not yet known (e.g., using a β-lactam plus vancomycin). Among the additional antimicrobial agents often used are rifampin, gentamicin, or fusidic acid (not available in the United States). To date, clinical studies have not documented a therapeutic benefit from these different combinations; recent reports have raised concern about the potential nephrotoxicity of gentamicin and adverse reactions to/interactions with rifampin. As a result, the use of gentamicin in combination with β-lactams or other antimicrobial agents is no longer routinely recommended for the treatment of native-valve endocarditis. Rifampin continues to be used for the treatment of prosthetic device–related infections and for osteomyelitis.

ANTIMICROBIAL THERAPY FOR SELECTED SETTINGS

Empirical Therapy

Empirical coverage for MRSA is generally indicated. Addition of a β-lactam to vancomycin provides more effective initial therapy should the isolate prove to be methicillin susceptible and may offer synergy in MRSA infections. It remains uncertain at present whether daptomycin is preferable when elevated vancomycin MICs (>1.5 μg/mL) are common in a specific locale.

Salvage Therapy

Salvage therapy for complicated *S. aureus* infections is sometimes needed when the bacteremia persists (i.e., for more than 3 or 4 days) despite appropriate treatment. There is little high-quality evidence to serve as a guide to salvage therapy. The combination of daptomycin or vancomycin with a β-lactam antibiotic (e.g., ceftaroline) has been successfully used to treat patients with persistent MRSA bacteremia, even those with isolates displaying reduced susceptibility to these antimicrobial agents. This combination appears to enhance the bactericidal activity of daptomycin by reducing the bacterial cell-surface charge and thus allowing more daptomycin binding. For vancomycin, the combination may allow more strategic binding to the target site with reduced cell-wall thickness. Other combinations have included trimethoprim-sulfamethoxazole or rifampin combined with daptomycin. Linezolid or ceftaroline has also been used as a single alternative agent.

Endocarditis

S. aureus endocarditis is usually an acute, life-threatening infection. Thus, prompt collection of blood for cultures should be followed by immediate institution of empirical antimicrobial therapy. For native-valve endocarditis, therapy with a β-lactam is recommended. If a MRSA strain is isolated, vancomycin (15–20 mg/kg every 8–12 h, given in equal doses up to a total of 2 g, with the dose adjusted in the case of renal disease) or daptomycin (6–10 mg/kg every 24 h) is recommended. The vancomycin dose should be adjusted on the basis of trough drug levels. Patients are generally treated for 6 weeks. For prosthetic-valve endocarditis, surgery in addition to antibiotic therapy is often necessary. The combination of a β-lactam agent—or, if the isolate is β-lactam-resistant, vancomycin or daptomycin—with an aminoglycoside (gentamicin, 1 mg/kg IV every 8 h) for 2 weeks and rifampin (300 mg orally or IV every 8 h) for ≥6 weeks is recommended.

Bone and Joint Infections

For hematogenous osteomyelitis or septic arthritis in children, a 4-week course of therapy is usually adequate. In adults, treatment is often more prolonged. For chronic forms of osteomyelitis, surgical debridement is necessary in combination with antimicrobial therapy. For joint infections, a critical component of therapy is the repeated aspiration or arthroscopy of the affected joint to prevent damage from leukocytes. The combination of rifampin with ciprofloxacin has been used successfully to treat or suppress prosthetic-joint infections, especially when the device cannot be removed. The efficacy of this combination may reflect enhanced activity against staphylococci in biofilms as well as the attainment of effective intracellular concentrations.

Skin and Soft Tissue Infections

The increase in SSTIs caused by CA-MRSA has drawn attention to the need for initiation of appropriate empirical therapy. Many of these infections respond to incision and drainage alone without antibiotics. Antibiotics are selected depending on local antibiotic susceptibility data; a number of oral agents have been used to treat these infections, including clindamycin, trimethoprim-sulfamethoxazole, doxycycline, linezolid, and tedizolid. Parenteral therapy is reserved for more complicated infections.

Toxic Shock Syndrome

Supportive therapy with reversal of hypotension is the mainstay of therapy for TSS. Both fluids and pressors may be necessary. Tampons or other packing material should be promptly removed. Some investigators recommend therapy with a combination of clindamycin and a semisynthetic penicillin or (if the isolate is resistant to methicillin) vancomycin. Clindamycin is advocated because, as a protein synthesis inhibitor, it reduces toxin synthesis in vitro. Linezolid also appears to be effective. A semisynthetic penicillin or a glycopeptide is recommended to eliminate any potential focus of infection as well as to eradicate persistent carriage that might increase the likelihood of recurrent illness. Anecdotal reports document the successful use of IV immunoglobulin to treat TSS. Glucocorticoids are not recommended for the treatment of this disease.

Other Toxin-Mediated Diseases

Therapy for staphylococcal food poisoning is entirely supportive. For SSSS, antistaphylococcal therapy targets the primary site of infection.

TABLE 142-3

Antimicrobial Therapy for Staphylococcal Infections[a]

Sensitivity/Resistance of Isolate	Drug of Choice	Alternative(s)	Comments
Parenteral Therapy for Serious Infections			
Sensitive to penicillin	Penicillin G (4 mU q4h)	Nafcillin or oxacillin (2 g q4h), cefazolin (2 g q8h), vancomycin (15–20 mg/kg q8h[b])	Fewer than 5% of isolates are sensitive to penicillin. The clinical microbiology laboratory must verify that the strain is not a β-lactamase producer.
Sensitive to methicillin	Nafcillin or oxacillin (2 g q4h)	Cefazolin (2 g q8h), vancomycin (15–20 mg/kg q8h[b])	Patients with a penicillin allergy can be treated with a cephalosporin if the allergy does not involve an anaphylactic or accelerated reaction; desensitization to β-lactams may be indicated in selected cases of serious infection when maximal bactericidal activity is needed (e.g., prosthetic-valve endocarditis[c]). Type A β-lactamase may rapidly hydrolyze cefazolin and reduce its efficacy in endocarditis. Vancomycin is a less effective option than a β-lactam.

Sensitivity/Resistance of Isolate	Drug of Choice	Alternative(s)	Comments
Resistant to methicillin	Vancomycin (15–20 mg/kg q8–12h[b]), daptomycin (6–10 mg/kg IV q24h[b,d]) for bacteremia, endocarditis, and complicated skin infections	Linezolid (600 mg q12h PO or IV), ceftaroline (600 mg IV q8–12h), telavancin (7.5–10 mg/kg IV q24h)[b], TMP-SMX (5 mg [based on TMP]/kg IV q8–12h)[f] Newer agents include tedizolid (200 mg once daily IV or PO), oritavancin (single dose of 1200 mg), and dalbavancin (single dose of 1500 mg). These drugs are approved only for the treatment of skin and soft tissue infections.[g]	Sensitivity testing is necessary before an alternative drug is selected. The efficacy of adjunctive therapy is not well established in many settings. Linezolid, ceftaroline, and telavancin have in vitro activity against most VISA and VRSA strains. See footnote for treatment of prosthetic-valve endocarditis.[c]
Resistant to methicillin with intermediate or complete resistance to vancomycin[e]	Daptomycin (6–10 mg/kg q24h[b,d]) for bacteremia, endocarditis, and complicated skin infections	Same as for methicillin-resistant strains (check antibiotic susceptibilities) *or*	Same as for methicillin-resistant strains; check antibiotic susceptibilities. Ceftaroline is used either alone or in combination with daptomycin.

Sensitivity/Resistance of Isolate	Drug of Choice	Alternative(s)	Comments
		Ceftaroline (600 mg IV q8–12h) Newer agents include tedizolid (200 mg once daily IV or PO), oritavancin (single dose of 1200 mg), and dalbavancin (single dose of 1500 mg). These drugs are approved only for the treatment of skin and soft tissue infections.	
Not yet known (i.e., empirical therapy)	Vancomycin (15–20 mg/kg q8–12h[b]), daptomycin (6–10 mg/kg q24h[b,d]) for bacteremia, endocarditis, and complicated skin infections	—	Empirical therapy is given when the susceptibility of the isolate is not known. Vancomycin with or without a β-lactam is recommended for suspected community- or hospital-acquired *Staphylococcus aureus* infections because of the increased frequency of methicillin-resistant strains in the community. If isolates with an elevated MIC to vancomycin (≥1.5 μg/ml) are common in the community, daptomycin may be preferable.
Oral Therapy for Skin and Soft Tissue Infections			

Sensitivity/Resistance of Isolate	Drug of Choice	Alternative(s)	Comments
Sensitive to methicillin	Dicloxacillin (500 mg qid), cephalexin (500 mg qid), or cefadroxil (1 g q12h)	Minocycline or doxycycline (100 mg q12h[b]), TMP-SMX (1 or 2 ds tablets bid), clindamycin (300–450 mg tid), linezolid (600 mg PO q12h), tedizolid (200 mg PO q24h)	It is important to know the antibiotic susceptibility of isolates in the specific geographic region. All collections should be drained and drainage should be cultured.
Resistant to methicillin	Clindamycin (300–450 mg tid), TMP-SMX (1 or 2 ds tablets bid), minocycline or doxycycline (100 mg q12h[b]), linezolid (600 mg bid), or tedizolid (200 mg once daily)	Same options as under "Drug of Choice"	It is important to know the antibiotic susceptibility of isolates in the specific geographic region. All collections should be drained and drainage should be cultured.

[a]Recommended dosages are for adults with normal renal and hepatic function. [b]The dosage must be adjusted for patients with reduced creatinine clearance. [c]For the treatment of prosthetic-valve endocarditis, the addition of gentamicin (1 mg/kg q8h) and rifampin (300 mg PO q8h) is recommended, with adjustment of the gentamicin dosage if the creatinine clearance rate is reduced. [d]Daptomycin cannot be used for the treatment of pneumonia. [e]Vancomycin-resistant *S. aureus* isolates from clinical infections have been reported. [f]TMP-SMX may be less effective than vancomycin. [g]Limited data are available on the efficacy of dalbavancin, oritavancin, and tedizolid for the treatment of invasive infections.

Abbreviations: ds, double-strength; TMP-SMX, trimethoprim-sulfamethoxazole; VISA, vancomycin-intermediate *S. aureus;* VRSA, vancomycin-resistant *S. aureus.*

Source: Modified from C Liu et al: Clin Infect Dis 52:285, 2011; DL Stevens et al: Clin Infect Dis 59:148, 2014; DL Stevens et al: Med Lett Drugs Ther 56:39, 2014; and LM Baddour et al: Circulation 132:1435, 2015.

PREVENTION

Primary prevention of *S. aureus* infections in the hospital setting involves hand washing and careful attention to appropriate isolation procedures. Through careful screening for MRSA carriage and strict isolation practices, several Scandinavian countries have been remarkably successful at preventing the introduction and dissemination of MRSA in hospitals.

Decolonization strategies, using both universal and targeted approaches with topical agents (e.g., mupirocin) to eliminate nasal colonization and/or chlorhexidine to eliminate colonization of additional body sites with *S. aureus*, have been successful in some clinical settings where the risk of infection is high (e.g., intensive care units). An analysis of clinical trials suggests that the incidence of postsurgical infections may be reduced among persons who are nasally colonized with *S. aureus*.

"Bundling" (the application of selected medical interventions in a sequence of prescribed steps) has reduced rates of nosocomial infections related to such procedures as the insertion of intravenous catheters, in which staphylococci are among the most common pathogens **(see Table 137-3)**. A number of immunization strategies to prevent *S. aureus* infections—both active (e.g., capsular polysaccharide–protein conjugate vaccine) and passive (e.g., clumping factor antibody)—have been investigated. However, none has been successful for either prophylaxis or therapy in clinical trials.

Strategies to prevent recurrent *S. aureus* infections in the community have had limited success. Decolonization with intranasal mupirocin and chlorhexidine washes of the infected individual and the additional decolonization of household members combined with environmental cleaning of surfaces and personal items have all been studied. For individuals with extensive skin disease and recurrent infections, the use of bleach baths (e.g., one-half cup of bleach in a one-quarter-filled bathtub) may be useful.

FURTHER READING

Becker K et al: Coagulase-negative staphylococci. Clin Microbiol Rev 27:870, 2014.

DeLeo FR et al: Community-associated meticillin-resistant *Staphylococcus aureus*. Lancet 375:1557, 2010.

Howden BP et al: Reduced vancomycin susceptibility in *Staphylococcus aureus*, including vancomycin-intermediate and heterogeneous vancomycin-intermediate strains: Resistance mechanisms, laboratory detection, and clinical implications. Clin Microbiol Rev 23:99, 2010.

Knox J et al: *Staphylococcus aureus* infections: Transmission within households and the community. Trends Microbiol 23:437, 2015.

Kullar R et al: When sepsis persists: A review of MRSA bacteraemia salvage therapy. J Antimicrob Chemother 71:576, 2016.

Miller L et al: Clindamycin versus trimethoprim-sulfamethoxazole for uncomplicated skin infections. N Engl J Med 372:1093, 2015.

Tong SY et al: *Staphylococcus aureus* infections: Epidemiology, pathophysiology, clinical manifestations, and management. Clin Microbiol Rev 28:603, 2015.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 143: Streptococcal Infections

Michael R. Wessels

INTRODUCTION

Many varieties of streptococci are found as part of the normal flora colonizing the human respiratory, gastrointestinal, and genitourinary tracts. Several species are important causes of human disease. Group A *Streptococcus* (GAS, *Streptococcus pyogenes*) is responsible for streptococcal pharyngitis, one of the most common bacterial infections of school-age children, and for the postinfectious syndromes of acute rheumatic fever (ARF) and poststreptococcal glomerulonephritis (PSGN). Group B *Streptococcus* (GBS, *Streptococcus agalactiae*) is the leading cause of bacterial sepsis and meningitis in newborns and a major cause of endometritis and fever in parturient women. Viridans streptococci are the most common cause of bacterial endocarditis. Enterococci, which are morphologically similar to streptococci, are now considered a separate genus on the basis of DNA homology studies. Thus, the species previously designated as *Streptococcus faecalis* and *Streptococcus faecium* have been renamed *Enterococcus faecalis* and *Enterococcus faecium*, respectively. **The enterococci are discussed in Chap. 144**.

Streptococci are gram-positive, spherical to ovoid bacteria that characteristically form chains when grown in liquid media. Most streptococci that cause human infections are facultative anaerobes, although some are strict anaerobes. Streptococci are relatively fastidious organisms, requiring enriched media for growth in the laboratory. Clinicians and clinical microbiologists identify streptococci by several classification systems, including hemolytic pattern, Lancefield group, species name, and common or trivial name. Many streptococci associated with human infection produce a zone of complete (β) hemolysis around the bacterial colony when cultured on blood agar. The β-hemolytic streptococci that form large (≥0.5-mm) colonies on blood agar can be classified by the Lancefield system, a serologic grouping based on the reaction of specific antisera with bacterial cell-wall carbohydrate antigens. With rare exceptions, organisms belonging to Lancefield groups A, B, C, and G are all β-hemolytic, and each is associated with characteristic patterns of human infection. Other streptococci produce a zone of partial (α) hemolysis, often imparting a greenish appearance to the agar. These α-hemolytic streptococci are further identified by biochemical testing and include *Streptococcus pneumoniae* **(Chap. 141)**, an important cause of pneumonia, meningitis, and other infections, and the several species referred to collectively as the *viridans streptococci*, which are part of the normal oral flora and are important agents of subacute bacterial endocarditis. Finally, some streptococci are nonhemolytic, a pattern sometimes called γ *hemolysis*. Among the organisms classified serologically as group D streptococci, the enterococci are classified as a distinct genus **(Chap. 144)**. The classification of the major streptococcal groups causing human infections is outlined in **Table 143-1**.

TABLE 143-1

Classification of Streptococci

Lancefield Group	Representative Species	Hemolytic Pattern	Typical Infections
A	*S. pyogenes*	β	Pharyngitis, impetigo, cellulitis, scarlet fever
B	*S. agalactiae*	β	Neonatal sepsis and meningitis, puerperal infection, urinary tract infection, diabetic ulcer infection, endocarditis
C, G	*S. dysgalactiae* subsp. *equisimilis*	β	Cellulitis, bacteremia, endocarditis
D	Enterococci[a]: *E. faecalis, E. faecium*	Usually nonhemolytic	Urinary tract infection, nosocomial bacteremia, endocarditis
	Nonenterococci: *S. gallolyticus* (formerly *S. bovis*)	Usually nonhemolytic	Bacteremia, endocarditis
Variable or nongroupable	Viridans streptococci: *S. sanguis, S. mitis*	α	Endocarditis, dental abscess, brain abscess
	Intermedius or *milleri* group: *S. intermedius, S. anginosus, S. constellatus*	Variable	Brain abscess, visceral abscess
	Anaerobic streptococci[b]: *Peptostreptococcus magnus*	Usually nonhemolytic	Sinusitis, pneumonia, empyema, brain abscess, liver abscess

[a]See Chap. 144. [b]See Chap. 172.

GROUP A STREPTOCOCCI

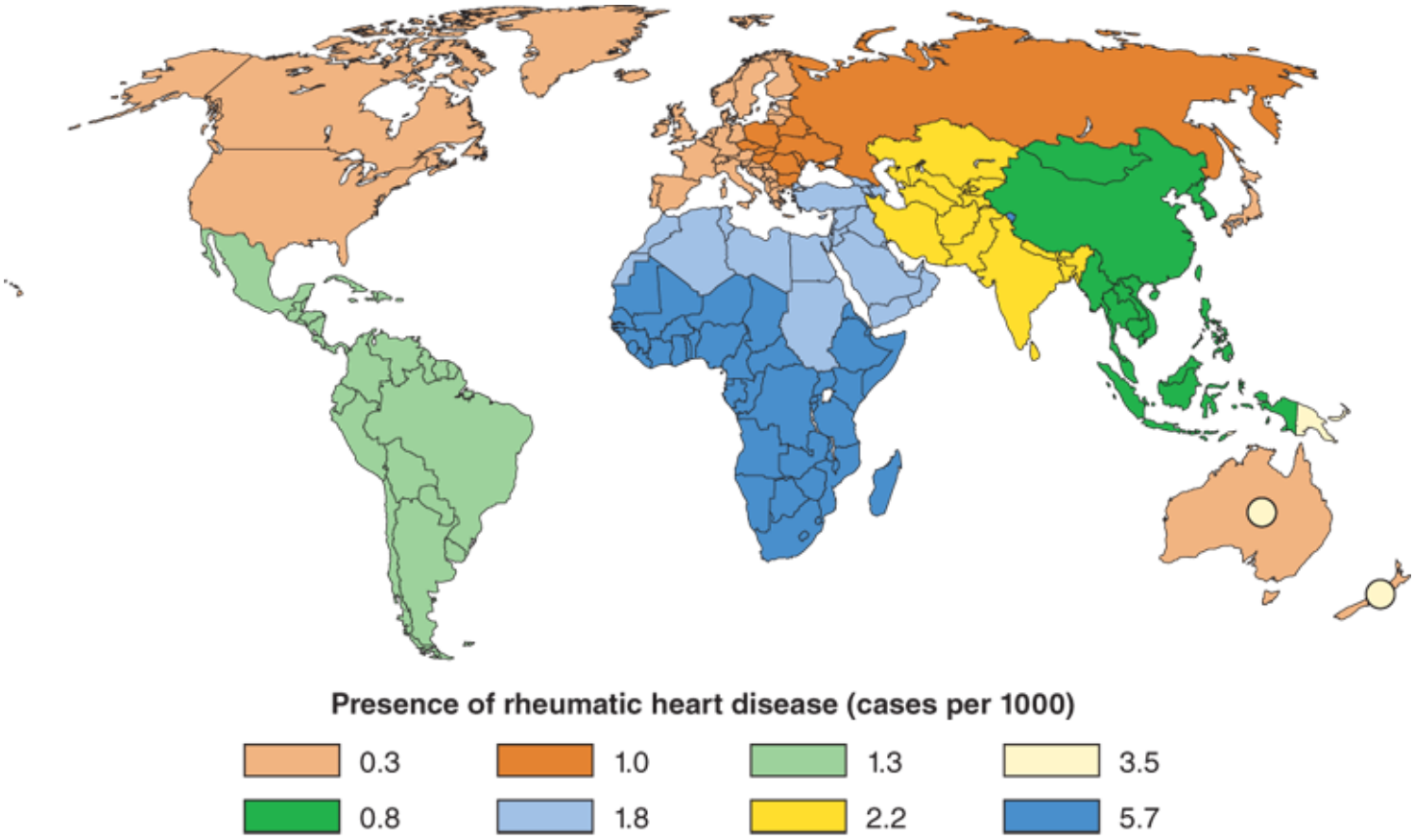
Lancefield's group A consists of a single species, *S. pyogenes*. As its species name implies, this organism is associated with a variety of suppurative infections. In addition, GAS can trigger the postinfectious syndromes of ARF (which is uniquely associated with *S. pyogenes* infection; Chap. 352) and PSGN (Chap. 308).

Worldwide, GAS infections and their postinfectious sequelae (primarily ARF and rheumatic heart disease) account for an estimated 500,000 deaths per year. Although data are incomplete, the incidence of all forms of GAS

infection and that of rheumatic heart disease are thought to be tenfold higher in resource-limited countries than in developed countries **(Fig. 143-1)**.

FIGURE 143-1

Prevalence of rheumatic heart disease in children 5–14 years old. The *circles* within Australia and New Zealand represent indigenous populations (and also Pacific Islanders in New Zealand). *(From JR Carapetis et al: Lancet Infect Dis 5:685, 2005, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOGENESIS

GAS elaborates a number of cell-surface components and extracellular products important in both the pathogenesis of infection and the human immune response. The cell wall contains a carbohydrate antigen that may be released by acid treatment. The reaction of such acid extracts with group A–specific antiserum is the basis for definitive identification of a streptococcal strain as *S. pyogenes*. The major surface protein of GAS is M protein, which is the basis for the serotyping of strains with specific antisera. The M protein molecules are fibrillar structures anchored in the cell wall of the organism that extend as hairlike projections away from the cell surface. The amino acid sequence of the distal or amino-terminal portion of the M protein molecule is variable, accounting for the antigenic variation of the different M types, while more proximal regions of the protein are relatively conserved. A newer technique for assignment of M type to GAS isolates uses the polymerase chain reaction to amplify the variable region of the *emm* gene, which encodes M protein. DNA sequence analysis of the amplified gene segment can be compared with an extensive database (developed at the Centers for Disease Control and Prevention [CDC]) for assignment of *emm* type. Use of *emm* typing has increased the number of identified *emm* types to more than

200. This method eliminates the need for typing sera, which are available in only a few reference laboratories. The presence of M protein on a GAS isolate correlates with its capacity to resist phagocytic killing in fresh human blood. This phenomenon appears to be due, at least in part, to the binding of plasma fibrinogen to M protein molecules on the streptococcal surface, which interferes with complement activation and deposition of opsonic complement fragments on the bacterial cell. This resistance to phagocytosis may be overcome by M protein–specific antibodies; thus individuals with antibodies to a given M type acquired as a result of prior infection are protected against subsequent infection with organisms of the same M type but not against that with different M types.

GAS also elaborates, to varying degrees, a polysaccharide capsule composed of hyaluronic acid. While most clinical isolates of GAS produce a hyaluronic acid capsule, strains of M type 4 or 22 lack a capsule, as do some isolates of M type 89. The fact that acapsular strains have been associated with pharyngitis and invasive infection implies that the capsule is not essential for virulence. The production of large amounts of capsule by certain strains imparts a characteristic mucoid appearance to the colonies. The capsular polysaccharide plays an important role in protecting GAS from ingestion and killing by phagocytes. In contrast to M protein, the hyaluronic acid capsule is a weak immunogen, and antibodies to hyaluronate have not been shown to be important in protective immunity. The presumed explanation is the apparent structural identity between streptococcal hyaluronic acid and the hyaluronic acid of mammalian connective tissues. The capsular polysaccharide may also play a role in GAS colonization of the pharynx by binding to CD44, a hyaluronic acid–binding protein expressed on human pharyngeal epithelial cells.

GAS produces a large number of extracellular products that may be important in local and systemic toxicity and in the spread of infection through tissues. These products include streptolysins S and O, toxins that damage cell membranes and account for the hemolysis produced by the organisms; streptokinase; DNAses; SpyCEP, a serine protease that cleaves and inactivates the chemoattractant cytokine interleukin 8, thereby inhibiting neutrophil recruitment to the site of infection; and several pyrogenic exotoxins. Previously known as erythrogenic toxins, the pyrogenic exotoxins cause the rash of scarlet fever. Since the mid-1980s, pyrogenic exotoxin–producing strains of GAS have been linked to unusually severe invasive infections, including necrotizing fasciitis and the streptococcal toxic shock syndrome (TSS). Several extracellular products stimulate specific antibody responses useful for serodiagnosis of recent streptococcal infection. Tests for antibodies to streptolysin O and DNase B are used most commonly for detection of preceding streptococcal infection in cases of suspected ARF or PSGN.

CLINICAL MANIFESTATIONS

Pharyngitis

Although seen in patients of all ages, GAS pharyngitis is one of the most common bacterial infections of childhood, accounting for 20–40% of all cases of exudative pharyngitis in children; it is rare among those under the age of 3. Younger children may manifest streptococcal infection with a syndrome of fever, malaise, and lymphadenopathy without exudative pharyngitis. Infection is acquired through contact with another individual carrying the organism. Respiratory droplets are the usual mechanism of spread, although other routes, including food-borne outbreaks, have been well described. The incubation period is 1–4 days. Symptoms include sore throat, fever and chills, malaise, and sometimes abdominal complaints and vomiting, particularly in children. Both symptoms and signs are quite variable, ranging from mild throat discomfort with minimal physical findings to high fever and severe sore throat associated with intense erythema and swelling of the pharyngeal mucosa and the presence of purulent

exudate over the posterior pharyngeal wall and tonsillar pillars. Enlarged, tender anterior cervical lymph nodes commonly accompany exudative pharyngitis.

The differential diagnosis of streptococcal pharyngitis includes the many other bacterial and viral etiologies (Table 143-2). Streptococcal infection is an unlikely cause when symptoms and signs suggestive of viral infection are prominent (conjunctivitis, coryza, cough, hoarseness, or discrete ulcerative lesions of the buccal or pharyngeal mucosa). Because of the range of clinical presentations of streptococcal pharyngitis and the large number of other agents that can produce the same clinical picture, diagnosis of streptococcal pharyngitis on clinical grounds alone is not reliable. The throat culture remains the diagnostic gold standard. Culture of a throat specimen that is properly collected (i.e., by vigorous rubbing of a sterile swab over both tonsillar pillars) and properly processed is the most sensitive and specific means of definitive diagnosis. A rapid diagnostic kit for latex agglutination or enzyme immunoassay of swab specimens is a useful adjunct to throat culture. While precise figures on sensitivity and specificity vary, rapid diagnostic kits generally are >95% specific. Thus a positive result can be relied upon for definitive diagnosis and eliminates the need for throat culture. However, because rapid diagnostic tests are less sensitive than throat culture (relative sensitivity in comparative studies, 55–90%), a negative result should be confirmed by throat culture.

TABLE 143-2

Infectious Etiologies of Acute Pharyngitis

Organism	Associated Clinical Syndrome(s)
Viruses	
Rhinovirus	Common cold
Coronavirus	Common cold
Adenovirus	Pharyngoconjunctival fever
Influenza virus	Influenza
Parainfluenza virus	Cold, croup
Coxsackievirus	Herpangina, hand-foot-and-mouth disease
Herpes simplex virus	Gingivostomatitis (primary infection)
Epstein-Barr virus	Infectious mononucleosis
Cytomegalovirus	Mononucleosis-like syndrome
HIV	Acute (primary) infection syndrome
Bacteria	
Group A streptococci	Pharyngitis, scarlet fever
Group C or G streptococci	Pharyngitis
Mixed anaerobes	Vincent’s angina
Arcanobacterium haemolyticum	Pharyngitis, scarlatiniform rash
Neisseria gonorrhoeae	Pharyngitis
Treponema pallidum	Secondary syphilis
Francisella tularensis	Pharyngeal tularemia
Corynebacterium diphtheriae	Diphtheria

Organism	Associated Clinical Syndrome(s)
Yersinia enterocolitica	Pharyngitis, enterocolitis
Yersinia pestis	Plague
Chlamydiae	
Chlamydia pneumoniae	Bronchitis, pneumonia
Chlamydia psittaci	Psittacosis
Mycoplasmas	
Mycoplasma pneumoniae	Bronchitis, pneumonia

TREATMENT

TREATMENT

GAS Pharyngitis

In the usual course of uncomplicated streptococcal pharyngitis, symptoms resolve after 3–5 days. The course is shortened little by treatment, which is given primarily to prevent suppurative complications and ARF. Prevention of ARF depends on eradication of the organism from the pharynx, not simply on resolution of symptoms, and requires 10 days of penicillin treatment **(Table 143-3)**. A first-generation cephalosporin, such as cephalexin or cefadroxil, may be substituted for penicillin in cases of penicillin allergy if the nature of the allergy is not an immediate hypersensitivity reaction (anaphylaxis or urticaria) or another potentially life-threatening manifestation (e.g., severe rash and fever).

Alternative agents are erythromycin and azithromycin. Azithromycin is more expensive but offers the advantages of better gastrointestinal tolerability, once-daily dosing, and a 5-day treatment course. Resistance to erythromycin and other macrolides is common among isolates from several countries, including Spain, Italy, Finland, Japan, and Korea. Macrolide resistance may be becoming more prevalent elsewhere with the increasing use of this class of antibiotics. In areas with resistance rates exceeding 5–10%, macrolides should be avoided unless results of susceptibility testing are known.

Follow-up culture after treatment is no longer routinely recommended but may be warranted in selected cases, such as those involving patients or families with frequent streptococcal infections or those occurring in situations in which the risk of ARF is thought to be high (e.g., when cases of ARF have recently been reported in the community).

TABLE 143-3

Treatment of Group A Streptococcal Infections

Infection	Treatment[a]
Pharyngitis	Benzathine penicillin G (1.2 mU IM) *or* penicillin V (250 mg PO tid or 500 mg PO bid) × 10 days
	(Children <27 kg: Benzathine penicillin G [600,000 units IM] *or* penicillin V [250 mg PO bid or tid] × 10 days)
Impetigo	Same as pharyngitis
Erysipelas/cellulitis	Severe: Penicillin G (1–2 mU IV q4h)
	Mild to moderate: Procaine penicillin (1.2 mU IM bid)
Necrotizing fasciitis/myositis	Surgical debridement *plus* penicillin G (2–4 mU IV q4h) *plus* clindamycin[b] (600–900 mg IV q8h)
Pneumonia/empyema	Penicillin G (2–4 mU IV q4h) *plus* drainage of empyema
Streptococcal toxic shock syndrome	Penicillin G (2–4 mU IV q4h) *plus* clindamycin[b] (600–900 mg IV q8h) *plus* IV immunoglobulin[b] (2 g/kg as a single dose)

[a]Penicillin allergy: A first-generation cephalosporin, such as cephalexin or cefadroxil, may be substituted for penicillin in cases of penicillin allergy if the nature of the allergy is not an immediate hypersensitivity reaction (anaphylaxis or urticaria) or another potentially life-threatening manifestation (e.g., severe rash and fever). Alternative agents for oral therapy are erythromycin (10 mg/kg PO qid, up to a maximum of 250 mg per dose) and azithromycin (a 5-day course at a dose of 12 mg/kg once daily, up to a maximum of 500 mg/d). Vancomycin is an alternative for parenteral therapy. [b]Efficacy unproven, but recommended by several experts. See text for discussion.

Complications

Suppurative complications of streptococcal pharyngitis have become uncommon with the widespread use of antibiotics for most symptomatic cases. These complications result from the spread of infection from the pharyngeal mucosa to deeper tissues by direct extension or by the hematogenous or lymphatic route and may include cervical lymphadenitis, peritonsillar or retropharyngeal abscess, sinusitis, otitis media, meningitis, bacteremia, endocarditis, and pneumonia. Local complications, such as peritonsillar or parapharyngeal abscess formation, should be considered in a patient with unusually severe or prolonged symptoms or localized pain associated with high fever and a toxic appearance. Nonsuppurative complications include ARF **(Chap. 351)** and PSGN **(Chap. 308)**, both of which are thought to result from immune responses to streptococcal infection. Penicillin treatment of streptococcal pharyngitis reduces the likelihood of ARF but not that of PSGN.

Bacteriologic Treatment Failure and the Asymptomatic Carrier State

Surveillance cultures have shown that up to 20% of individuals in certain populations may have asymptomatic pharyngeal colonization with GAS. There are no definitive guidelines for management of these asymptomatic carriers or of asymptomatic patients who still have a positive throat culture after a full course of treatment for symptomatic pharyngitis. A reasonable course of action is to give a single 10-day course of penicillin for symptomatic pharyngitis and, if positive cultures persist, not to re-treat unless symptoms recur. Studies of the natural history of streptococcal carriage and infection have shown that the risk both of developing ARF and of transmitting infection to others is substantially lower among asymptomatic carriers than among individuals with symptomatic pharyngitis. Therefore, overly aggressive attempts to eradicate carriage probably are not justified under most circumstances. An exception is the situation in which an asymptomatic carrier is a potential source of infection to others. Outbreaks of food-borne infection and nosocomial puerperal infection have been traced to asymptomatic carriers who may harbor the organisms in the throat, vagina, or anus or on the skin.

TREATMENT

TREATMENT

Asymptomatic Pharyngeal Colonization with GAS

When a carrier is transmitting infection to others, attempts to eradicate carriage are warranted. Data are limited on the best regimen to clear GAS after penicillin alone has failed. Regimens reported to have efficacy superior to that of penicillin alone for eradication of carriage include (1) a first-generation cephalosporin such as cephalexin (30 mg/kg; 500 mg maximum) twice daily for 10 days or (2) oral clindamycin (7 mg/kg; 300 mg maximum) three times daily for 10 days. A 10-day course of oral vancomycin (250 mg four times daily) and rifampin (600 mg twice daily) has eradicated rectal colonization.

Scarlet Fever

Scarlet fever consists of streptococcal infection, usually pharyngitis, accompanied by a characteristic rash **(Fig. 143-2)**. The rash arises from the effects of one of several toxins, currently designated *streptococcal pyrogenic exotoxins* and previously known as *erythrogenic* or *scarlet fever toxins*. In the past, scarlet fever was thought to reflect infection of an individual lacking toxin-specific immunity with a toxin-producing strain of GAS. Susceptibility to scarlet fever was correlated with results of the Dick test, in which a small amount of erythrogenic toxin injected intradermally produced local erythema in susceptible individuals but elicited no reaction in those with specific immunity. Subsequent studies have suggested that development of the scarlet fever rash may reflect a hypersensitivity reaction requiring prior exposure to the toxin. For reasons that are not clear, scarlet fever has become less common in recent years, although large outbreaks have occurred recently in China and the United Kingdom. The symptoms of scarlet fever are the same as those of pharyngitis alone. The rash typically begins on the first or second day of illness over the upper trunk, spreading to involve the extremities but sparing the palms and soles. The rash is made up of minute papules, giving a characteristic “sandpaper” feel to the skin. Associated findings include circumoral pallor, “strawberry tongue” (enlarged papillae on a coated tongue, which later may become denuded), and accentuation of the rash in skinfolds (*Pastia’s lines*). Subsidence of the rash in 6–9 days is followed after several days by desquamation of the palms and soles. The differential diagnosis of scarlet fever includes other causes of fever and generalized rash, such as measles and other viral exanthems, Kawasaki disease, TSS, and systemic allergic reactions (e.g., drug eruptions).

FIGURE 143-2

Scarlet fever exanthem. Finely punctate erythema has become confluent (scarlatiniform); petechiae can occur and have a linear configuration within the exanthem in body folds (Pastia's lines). *(From Fitzpatrick, Johnson, Wolff: Color Atlas and Synopsis of Clinical Dermatology, 4th ed, New York, McGraw-Hill, 2001, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Skin and Soft Tissue Infections

GAS—and occasionally other streptococcal species—can cause a variety of infections involving the skin, subcutaneous tissues, muscles, and fascia. While several clinical syndromes offer a useful means for classification of these infections, not all cases fit exactly into one category. The classic syndromes are general guides to predicting the level of tissue involvement in a particular patient, the probable clinical course, and the likelihood that surgical intervention or aggressive life support will be required.

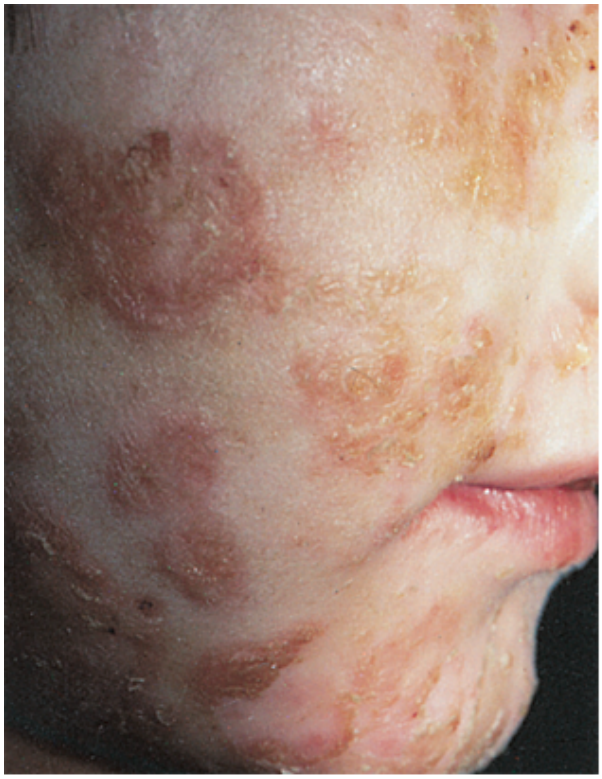
Impetigo (Pyoderma)

Impetigo, a superficial infection of the skin, is caused primarily by GAS and occasionally by other streptococci or *Staphylococcus aureus*. Impetigo is seen most often in young children, tends to occur during warmer months, and is more common in semitropical or tropical climates than in cooler regions. Infection is more common among children living under conditions of poor hygiene. Prospective studies have shown that colonization of unbroken skin with GAS precedes clinical infection. Minor trauma, such as a scratch or an insect bite, may then serve to inoculate organisms into the skin. Impetigo is best prevented, therefore, by attention to adequate hygiene. The usual sites of involvement are the face (particularly around the nose and mouth) and the legs, although lesions may occur at other locations. Individual lesions begin as red papules, which evolve quickly into vesicular and then

pustular lesions that break down and coalesce to form characteristic honeycomb-like crusts **(Fig. 143-3)**. Lesions generally are not painful, and patients do not appear ill. Fever is not a feature of impetigo and, if present, suggests either infection extending to deeper tissues or another diagnosis. The classic presentation of impetigo usually poses little diagnostic difficulty. Cultures of impetiginous lesions often yield *S. aureus* as well as GAS. In almost all cases, streptococci are isolated initially and staphylococci appear later, presumably as secondary colonizing flora. In the past, penicillin was nearly always effective against these infections. However, an increasing frequency of penicillin treatment failure suggests that *S. aureus* may have become more prominent as a cause of impetigo. *Bullous impetigo* due to *S. aureus* is distinguished from typical streptococcal infection by more extensive, bullous lesions that break down and leave thin paper-like crusts instead of the thick amber crusts of streptococcal impetigo. Other skin lesions that may be confused with impetigo include herpetic lesions—either those of orolabial herpes simplex or those of chickenpox or zoster. Herpetic lesions can generally be distinguished by their appearance as more discrete, grouped vesicles and by a positive Tzanck test. In difficult cases, cultures of vesicular fluid should yield GAS in impetigo and the responsible virus in herpesvirus infections.

FIGURE 143-3

Impetigo is a superficial streptococcal or *Staphylococcus aureus* infection consisting of honey-colored crusts and erythematous weeping erosions. Occasionally, bullous lesions may be seen. *(Courtesy of Mary Spraker, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Streptococcal Impetigo

Treatment of streptococcal impetigo is the same as that for streptococcal pharyngitis. In view of evidence that *S. aureus* has become a relatively frequent cause of impetigo, empirical regimens should cover both streptococci and

S. aureus. For example, either dicloxacillin or cephalexin can be given at a dose of 250 mg four times daily for 10 days. Topical mupirocin ointment is also effective. Culture may be indicated to rule out methicillin-resistant *S. aureus*, especially if the response to empirical treatment is unsatisfactory. ARF is not a sequela to streptococcal skin infections, although PSGN may follow either skin or throat infection. The reason for this difference is not known. One hypothesis is that the immune response necessary for development of ARF occurs only after infection of the pharyngeal mucosa. In addition, the strains of GAS that cause pharyngitis are generally of different M protein types than those associated with skin infections; thus the strains that cause pharyngitis may have rheumatogenic potential, while the skin-infecting strains may not.

Cellulitis

Inoculation of organisms into the skin may lead to *cellulitis*: infection involving the skin and subcutaneous tissues. The portal of entry may be a traumatic or surgical wound, an insect bite, or any other break in skin integrity. Often, no entry site is apparent. One form of streptococcal cellulitis, *erysipelas*, is characterized by a bright red appearance of the involved skin, which forms a plateau sharply demarcated from surrounding normal skin **(Fig. 143-4)**. The lesion is warm to the touch, may be tender, and appears shiny and swollen. The skin often has a *peau d'orange* texture, which is thought to reflect involvement of superficial lymphatics; superficial blebs or bullae may form, usually 2–3 days after onset. The lesion typically develops over a few hours and is associated with fever and chills. Erysipelas tends to occur on the malar area of the face (often with extension over the bridge of the nose to the contralateral malar region) or on the lower extremities. After one episode, recurrence at the same site—sometimes years later—is not uncommon. Classic cases of erysipelas, with typical features, are almost always due to β-hemolytic streptococci, usually GAS and occasionally group C or G. Often, however, the appearance of streptococcal cellulitis is not sufficiently distinctive to permit a specific diagnosis on clinical grounds. The anatomic area involved may not be typical for erysipelas, the lesion may be less intensely red than usual and may fade into surrounding skin, and/or the patient may appear only mildly ill. In such cases, it is prudent to broaden the spectrum of empirical antimicrobial therapy to include other pathogens, particularly *S. aureus*, that can produce cellulitis with the same appearance. Staphylococcal infection should be suspected if cellulitis develops around a wound or an ulcer.

FIGURE 143-4

Erysipelas is a streptococcal infection of the superficial dermis and consists of well-demarcated, erythematous, edematous, warm plaques.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Streptococcal cellulitis tends to develop at anatomic sites in which normal lymphatic drainage has been disrupted, such as sites of prior cellulitis, the arm ipsilateral to a mastectomy and axillary lymph node dissection, a lower extremity previously involved in deep venous thrombosis or chronic lymphedema, or the leg from which a saphenous vein has been harvested for coronary artery bypass grafting. The organism may enter via a dermal breach some distance from the eventual site of clinical cellulitis. For example, some patients with recurrent leg cellulitis following saphenous vein removal stop having recurrent episodes only after treatment of tinea pedis on the affected extremity. Fissures in the skin presumably serve as a portal of entry for streptococci, which then produce infection more proximally in the leg at the site of previous injury. Streptococcal cellulitis may also involve recent surgical wounds. GAS is among the few bacterial pathogens that typically produce signs of wound infection and surrounding cellulitis within the first 24 h after surgery. These wound infections are usually associated with a thin exudate and may spread rapidly, either as cellulitis in the skin and subcutaneous tissue or as a deeper tissue infection (see below). Streptococcal wound infection or localized cellulitis may also be associated with *lymphangitis*, manifested by red streaks extending proximally along superficial lymphatics from the infection site.

TREATMENT

TREATMENT

Streptococcal Cellulitis

See Table 143-3 and **Chap. 124.**

Deep Soft-Tissue Infections

Necrotizing fasciitis (*hemolytic streptococcal gangrene*) involves the superficial and/or deep fascia investing the muscles of an extremity or the trunk. The source of the infection is either the skin, with organisms introduced into

tissue through trauma (sometimes trivial), or the bowel flora, with organisms released during abdominal surgery or from an occult enteric source, such as a diverticular or appendiceal abscess. The inoculation site may be inapparent and is often some distance from the site of clinical involvement; e.g., the introduction of organisms via minor trauma to the hand may be associated with clinical infection of the tissues overlying the shoulder or chest. Cases associated with the bowel flora are usually polymicrobial, involving a mixture of anaerobic bacteria (such as *Bacteroides fragilis* or anaerobic streptococci) and facultative organisms (usually gram-negative bacilli). Cases unrelated to contamination from bowel organisms are most commonly caused by GAS alone or in combination with other organisms (most often *S. aureus*). Overall, GAS is implicated in ~60% of cases of necrotizing fasciitis. The onset of symptoms is usually quite acute and is marked by severe pain at the site of involvement, malaise, fever, chills, and a toxic appearance. The physical findings, particularly early on, may not be striking, with only minimal erythema of the overlying skin. Pain and tenderness are usually severe. In contrast, in more superficial cellulitis, the skin appearance is more abnormal, but pain and tenderness are only mild or moderate. As the infection progresses (often over several hours), the severity and extent of symptoms worsen, and skin changes become more evident, with the appearance of dusky or mottled erythema and edema. The marked tenderness of the involved area may evolve into anesthesia as the spreading inflammatory process produces infarction of cutaneous nerves.

Although myositis is more commonly due to *S. aureus* infection, GAS occasionally produces abscesses in skeletal muscles (*streptococcal myositis*), with little or no involvement of the surrounding fascia or overlying skin. The presentation is usually subacute, but a fulminant form has been described in association with severe systemic toxicity, bacteremia, and a high mortality rate. The fulminant form may reflect the same basic disease process seen in necrotizing fasciitis, but with the necrotizing inflammatory process extending into the muscles themselves rather than remaining limited to the fascial layers.

TREATMENT

TREATMENT

Deep Soft-Tissue Streptococcal Infections

Once necrotizing fasciitis is suspected, early surgical exploration is both diagnostically and therapeutically indicated. Surgery reveals necrosis and inflammatory fluid tracking along the fascial planes above and between muscle groups, without involvement of the muscles themselves. The process usually extends beyond the area of clinical involvement, and extensive debridement is required. Drainage and debridement are central to the management of necrotizing fasciitis; antibiotic treatment is a useful adjunct (Table 143-3), but surgery is life-saving. Treatment for streptococcal myositis consists of surgical drainage—usually by an open procedure that permits evaluation of the extent of infection and ensures adequate debridement of involved tissues—and high-dose penicillin (Table 143-3).

Pneumonia and Empyema

GAS is an occasional cause of pneumonia, generally in previously healthy individuals. The onset of symptoms may be abrupt or gradual. Pleuritic chest pain, fever, chills, and dyspnea are the characteristic manifestations. Cough is usually present but may not be prominent. Approximately one-half of patients with GAS pneumonia have an accompanying pleural effusion. In contrast to the sterile parapneumonic effusions typical of pneumococcal pneumonia, those complicating streptococcal pneumonia are almost always infected. The empyema fluid is usually visible by chest radiography on initial presentation, and its volume may increase rapidly. These pleural collections

should be drained early, as they tend to become loculated rapidly, resulting in a chronic fibrotic reaction that may require thoracotomy for removal.

Bacteremia, Puerperal Sepsis, and Streptococcal Toxic Shock Syndrome

In adults, GAS bacteremia is usually associated with an identifiable local infection, whereas children may have bacteremia without an associated focal infection. Bacteremia occurs rarely with otherwise uncomplicated pharyngitis, occasionally with cellulitis or pneumonia, and relatively frequently with necrotizing fasciitis. Bacteremia without an identified source raises the possibility of endocarditis, an occult abscess, or osteomyelitis. A variety of focal infections may arise secondarily from streptococcal bacteremia, including endocarditis, meningitis, septic arthritis, osteomyelitis, peritonitis, and visceral abscesses. GAS is occasionally implicated in infectious complications of childbirth, usually endometritis and associated bacteremia. In the preantibiotic era, puerperal sepsis was commonly caused by GAS; currently, it is more often caused by GBS. Several nosocomial outbreaks of puerperal GAS infection have been traced to an asymptomatic carrier, usually someone present at delivery. The site of carriage may be the skin, throat, anus, or vagina.

Beginning in the late 1980s, several reports described patients with GAS infections associated with shock and multisystem organ failure. This syndrome was called *streptococcal TSS* because it shares certain features with staphylococcal TSS. In 1993, a case definition for streptococcal TSS was formulated **(Table 143-4)**. The general features of the illness include fever, hypotension, renal impairment, and respiratory distress syndrome. Various types of rash have been described, but rash usually does not develop. Laboratory abnormalities include a marked shift to the left in the white blood cell differential, with many immature granulocytes; hypocalcemia; hypoalbuminemia; and thrombocytopenia, which usually becomes more pronounced on the second or third day of illness. In contrast to patients with staphylococcal TSS, the majority with streptococcal TSS are bacteremic. The most common associated infection is a soft tissue infection—necrotizing fasciitis, myositis, or cellulitis—although a variety of other associated local infections have been described, including pneumonia, peritonitis, osteomyelitis, and myometritis. Streptococcal TSS is associated with a mortality rate of ≥30%, with most deaths secondary to shock and respiratory failure. Because of its rapidly progressive and lethal course, early recognition of the syndrome is essential. Patients should receive aggressive supportive care (fluid resuscitation, pressors, and mechanical ventilation) in addition to antimicrobial therapy and, in cases associated with necrotizing fasciitis, should undergo surgical debridement. Exactly why certain patients develop this fulminant syndrome is not known. Early studies of the streptococcal strains isolated from these patients demonstrated a strong association with the production of pyrogenic exotoxin A. This association has been inconsistent in subsequent case series. Pyrogenic exotoxin A and several other streptococcal exotoxins act as superantigens to trigger release of inflammatory cytokines from T lymphocytes. Fever, shock, and organ dysfunction in streptococcal TSS may reflect, in part, the systemic effects of superantigen-mediated cytokine release.

TABLE 143-4

Proposed Case Definition for the Streptococcal Toxic Shock Syndrome[a]

I. Isolation of group A streptococci (*Streptococcus pyogenes*)
   A. From a normally sterile site
   B. From a nonsterile site
II. Clinical signs of severity
   A. Hypotension *and*
   B. ≥2 of the following signs
      1. Renal impairment
      2. Coagulopathy
      3. Liver function impairment
      4. Adult respiratory distress syndrome
      5. A generalized erythematous macular rash that may desquamate
      6. Soft tissue necrosis, including necrotizing fasciitis or myositis; *or* gangrene

[a]An illness fulfilling criteria IA, IIA, and IIB is defined as a *definite* case. An illness fulfilling criteria IB, IIA, and IIB is defined as a *probable* case if no other etiology for the illness is identified.

Source: Modified from Working Group on Severe Streptococcal Infections: JAMA 269:390, 1993.

TREATMENT

TREATMENT

Streptococcal Toxic Shock Syndrome

In light of the possible role of pyrogenic exotoxins or other streptococcal toxins in streptococcal TSS, treatment with clindamycin has been advocated by some authorities (Table 143-3), who argue that, through its direct action on protein synthesis, clindamycin is more effective in rapidly terminating toxin production than is penicillin—a cell-wall agent. Support for this view comes from studies of an experimental model of streptococcal myositis, in which mice given clindamycin had a higher rate of survival than those given penicillin. Comparable data on the treatment of human infections are not available, although retrospective analysis has suggested a better outcome when patients with invasive soft-tissue infection are treated with clindamycin rather than with cell wall–active antibiotics. Although clindamycin resistance in GAS is uncommon among U.S. isolates (<2%), resistance rates as high as 23% have been documented in Finland. Thus, if clindamycin is used for initial treatment of a critically ill patient, penicillin should be given as well until the antibiotic susceptibility of the streptococcal isolate is known. IV immunoglobulin has been used as adjunctive therapy for streptococcal TSS (Table 143-3). Pooled immunoglobulin preparations contain antibodies capable of neutralizing the effects of streptococcal toxins. Anecdotal reports and case series have suggested favorable clinical responses to IV immunoglobulin, but no adequately powered, prospective, controlled trials have been reported.

PREVENTION

No vaccine against GAS is commercially available. A formulation that consists of recombinant peptides containing epitopes of 26 M-protein types has undergone phase 1 and 2 testing in volunteers. Early results indicate that the vaccine is well tolerated and elicits type-specific antibody responses. Vaccines based on a conserved region of M protein or on a mixture of other conserved GAS protein antigens are in earlier stages of development.

Household contacts of individuals with invasive GAS infection (e.g., bacteremia, necrotizing fasciitis, or streptococcal TSS) are at greater risk of invasive infection than the general population. Asymptomatic pharyngeal colonization with GAS has been detected in up to 25% of persons with >4 h/d of same-room exposure to an index case. However, the CDC does not recommend antibiotic prophylaxis routinely for contacts of patients with invasive disease because such an approach (if effective) would require treatment of hundreds of contacts to prevent a single case. Prophylaxis may be considered for contacts of unusually severe cases or for individuals at increased risk for invasive infection.

STREPTOCOCCI OF GROUPS C AND G

Group C and group G streptococci are β-hemolytic bacteria that occasionally cause human infections similar to those caused by GAS. Strains that form small colonies on blood agar (<0.5 mm) are generally members of the *Streptococcus milleri* group (*Streptococcus intermedius*, *Streptococcus anginosus*; see "Viridans Streptococci," below). Large-colony group C and G streptococci of human origin are now considered a single species, *Streptococcus dysgalactiae* subspecies *equisimilis*. These organisms have been associated with pharyngitis, cellulitis and soft tissue infections, pneumonia, bacteremia, endocarditis, and septic arthritis. Puerperal sepsis, meningitis, epidural abscess, intraabdominal abscess, urinary tract infection, and neonatal sepsis have also been reported. Group C or G streptococcal bacteremia most often affects elderly or chronically ill patients and, in the absence of obvious local infection, is likely to reflect endocarditis. Septic arthritis, sometimes involving multiple joints, may complicate endocarditis or develop in its absence. Distinct streptococcal species of Lancefield group C cause infections in domesticated animals, especially horses and cattle; some human infections are acquired through contact with animals or consumption of unpasteurized milk. These zoonotic organisms include *Streptococcus equi* subspecies *zooepidemicus* and *S. equi* subspecies *equi*.

TREATMENT

TREATMENT

Group C or G Streptococcal Infection

Penicillin is the drug of choice for treatment of group C or G streptococcal infections. Antibiotic treatment is the same as for similar syndromes due to GAS (Table 143-3). Patients with bacteremia or septic arthritis should receive IV penicillin (2–4 mU every 4 h). All group C and G streptococci are sensitive to penicillin; nearly all are inhibited in vitro by concentrations of ≤0.03 μg/mL. Occasional isolates exhibit tolerance: although inhibited by low concentrations of penicillin, they are killed only by significantly higher concentrations. The clinical significance of tolerance is unknown. Because of the poor clinical response of some patients to penicillin alone, the addition of gentamicin (1 mg/kg every 8 h for patients with normal renal function) is recommended by some authorities for treatment of endocarditis or septic arthritis due to group C or G streptococci; however, combination therapy has

not been shown to be superior to penicillin treatment alone. Patients with joint infections often require repeated aspiration or open drainage and debridement for cure; the response to treatment may be slow, particularly in debilitated patients and those with involvement of multiple joints. Infection of prosthetic joints almost always requires prosthesis removal in addition to antibiotic therapy.

GROUP B STREPTOCOCCI

Identified first as a cause of mastitis in cows, streptococci belonging to Lancefield's group B have since been recognized as a major cause of sepsis and meningitis in human neonates. GBS is also a frequent cause of peripartum fever in women and an occasional cause of serious infection in nonpregnant adults. Since the widespread institution of prenatal screening for GBS in the 1990s, the incidence of neonatal infection per 1000 live births has fallen from ~2–3 cases to ~0.6 case. During the same period, GBS infection in adults with underlying chronic illnesses has become more common; adults now account for a larger proportion of invasive GBS infections than do newborns. Lancefield group B consists of a single species, *S. agalactiae*, which is definitively identified with specific antiserum to the group B cell wall–associated carbohydrate antigen. A streptococcal isolate can be classified presumptively as GBS on the basis of biochemical tests, including hydrolysis of sodium hippurate (in which 99% of isolates are positive), hydrolysis of bile esculin (in which 99–100% are negative), bacitracin susceptibility (in which 92% are resistant), and production of CAMP factor (in which 98–100% are positive). CAMP factor is a phospholipase produced by GBS that causes synergistic hemolysis with β lysin produced by certain strains of *S. aureus*. Its presence can be demonstrated by cross-streaking of the test isolate and an appropriate staphylococcal strain on a blood agar plate. GBS organisms causing human infections are encapsulated by one of ten antigenically distinct polysaccharides. The capsular polysaccharide is an important virulence factor. Antibodies to the capsular polysaccharide afford protection against GBS of the same (but not of a different) capsular type.

INFECTION IN NEONATES

Two general types of GBS infection in infants are defined by the age of the patient at presentation. *Early-onset infections* occur within the first week of life, with a median age of 20 h at onset. Approximately half of these infants have signs of GBS disease at birth. The infection is acquired during or shortly before birth from the colonized maternal genital tract. Surveillance studies have shown that 5–40% of women are vaginal or rectal carriers of GBS. Approximately 50% of infants delivered vaginally by carrier mothers become colonized, although only 1–2% develop clinically evident infection. Prematurity, prolonged labor, obstetric complications, and maternal fever are risk factors for early-onset infection. The presentation of early-onset infection is the same as that of other forms of neonatal sepsis. Typical findings include respiratory distress, lethargy, and hypotension. Essentially all infants with early-onset disease are bacteremic, one-third to one-half have pneumonia and/or respiratory distress syndrome, and one-third have meningitis.

Late-onset infections occur in infants 1 week to 3 months old and, in rare instances, in older infants (mean age at onset, 3–4 weeks). The infecting organism may be acquired during delivery (as in early-onset cases) or during later contact with a colonized mother, nursery personnel, or another source. Meningitis is the most common manifestation of late-onset infection and in most cases is associated with a strain of capsular type III. Infants present with fever, lethargy or irritability, poor feeding, and seizures. The various other types of late-onset infection

include bacteremia without an identified source, osteomyelitis, septic arthritis, and facial cellulitis associated with submandibular or preauricular adenitis.

TREATMENT

TREATMENT

Group B Streptococcal Infection in Neonates

Penicillin is the agent of choice for all GBS infections. Empirical broad-spectrum therapy for suspected bacterial sepsis, consisting of ampicillin and gentamicin, is generally administered until culture results become available. If cultures yield GBS, many pediatricians continue to administer gentamicin, along with ampicillin or penicillin, for a few days until clinical improvement becomes evident. Infants with bacteremia or soft tissue infection should receive penicillin at a dosage of 200,000 units/kg per day in divided doses. For meningitis, infants ≤7 days of age should receive 250,000–450,000 units/kg per day in three divided doses; infants >7 days of age should receive 450,000–500,000 units/kg per day in four divided doses. Meningitis should be treated for at least 14 days because of the risk of relapse with shorter courses.

Prevention

The incidence of GBS infection is unusually high among infants of women with risk factors: preterm delivery, early rupture of membranes (>24 h before delivery), prolonged labor, fever, or chorioamnionitis. Because the usual source of the organisms infecting a neonate is the mother's birth canal, efforts have been made to prevent GBS infections by the identification of high-risk carrier mothers and their treatment with various forms of antibiotic prophylaxis or immunoprophylaxis. Prophylactic administration of ampicillin or penicillin to such patients during delivery reduces the risk of infection in the newborn. This approach has been hampered by logistical difficulties in identifying colonized women before delivery; the results of vaginal cultures early in pregnancy are poor predictors of carrier status at delivery. The CDC recommends screening for anogenital colonization at 35–37 weeks of pregnancy by a swab culture of the lower vagina and anorectum; intrapartum chemoprophylaxis is recommended for culture-positive women and for women who, regardless of culture status, have previously given birth to an infant with GBS infection or have a history of GBS bacteriuria during pregnancy. Women whose culture status is unknown and who develop premature labor (<37 weeks), prolonged rupture of membranes (>18 h), or intrapartum fever or who have a positive intrapartum nucleic acid amplification test for GBS should also receive intrapartum chemoprophylaxis. The recommended regimen for chemoprophylaxis is a loading dose of 5 million units of penicillin G followed by 2.5 million units every 4 h until delivery. Cefazolin is an alternative for women with a history of penicillin allergy who are thought not to be at high risk for anaphylaxis. For women with a history of immediate hypersensitivity, clindamycin may be substituted, but only if the colonizing isolate has been demonstrated to be susceptible. If susceptibility testing results are not available or indicate resistance, vancomycin should be used in this situation.

Treatment of all pregnant women who are colonized or have risk factors for neonatal infection will result in exposure of up to one-third of pregnant women and newborns to antibiotics, with the attendant risks of allergic reactions and selection for resistant organisms. Although still in the developmental stages, a GBS vaccine may ultimately offer a better solution to prevention. Because transplacental passage of maternal antibodies produces protective antibody levels in newborns, efforts are under way to develop a vaccine against GBS that can be given to

childbearing-age women before or during pregnancy. Results of phase 1 clinical trials of GBS capsular polysaccharide–protein conjugate vaccines suggest that a multivalent conjugate vaccine would be safe and highly immunogenic.

INFECTION IN ADULTS

The majority of GBS infections in otherwise healthy adults are related to pregnancy and parturition. Peripartum fever, the most common manifestation, is sometimes accompanied by symptoms and signs of endometritis or chorioamnionitis (abdominal distention and uterine or adnexal tenderness). Blood and vaginal swab cultures are often positive. Bacteremia is usually transitory but occasionally results in meningitis or endocarditis. Infections in adults that are not associated with the peripartum period generally involve individuals who are elderly or have an underlying chronic illness, such as diabetes mellitus or a malignancy. Among the infections that develop with some frequency in adults are cellulitis and soft tissue infection (including infected diabetic skin ulcers), urinary tract infection, pneumonia, endocarditis, and septic arthritis. Other reported infections include meningitis, osteomyelitis, and intraabdominal or pelvic abscesses. Relapse or recurrence of invasive infection weeks to months after a first episode is documented in ~4% of cases.

TREATMENT

TREATMENT

Group B Streptococcal Infection in Adults

GBS is less sensitive to penicillin than GAS, requiring somewhat higher doses. Adults with serious localized infections (pneumonia, pyelonephritis, abscess) should receive doses of ~12 million units of penicillin G daily; patients with endocarditis or meningitis should receive 18–24 million units per day in divided doses. Vancomycin is an acceptable alternative for penicillin-allergic patients.

NONENTEROCOCCAL GROUP D STREPTOCOCCI

The main nonenterococcal group D streptococci that cause human infections were previously considered a single species, *Streptococcus bovis*. The organisms encompassed by *S. bovis* have been reclassified into two species, each of which has two subspecies: *Streptococcus gallolyticus* subspecies *gallolyticus*, *S. gallolyticus* subspecies *pasteurianus*, *Streptococcus infantarius* subspecies *infantarius*, and *S. infantarius* subspecies *coli.* Endocarditis caused by these organisms is often associated with neoplasms of the gastrointestinal tract—most frequently, a colon carcinoma or polyp—but is also reported in association with other bowel lesions. When occult gastrointestinal lesions are carefully sought, abnormalities are found in >60% of patients with endocarditis due to *S. gallolyticus* or *S. infantarius.* In contrast to the enterococci, nonenterococcal group D streptococci like these organisms are reliably killed by penicillin as a single agent, and penicillin is the agent of choice for the infections they cause.

VIRIDANS AND OTHER STREPTOCOCCI

VIRIDANS STREPTOCOCCI

Consisting of multiple species of α-hemolytic streptococci, the viridans streptococci are a heterogeneous group of organisms that are important agents of bacterial endocarditis **(Chap. 123)**. Several species of viridans streptococci, including *Streptococcus salivarius*, *Streptococcus mitis*, *Streptococcus sanguis*, and *Streptococcus mutans*, are part of the normal flora of the mouth, where they live in close association with the teeth and gingiva. Some species contribute to the development of dental caries.

Previously known as *Streptococcus morbillorum*, *Gemella morbillorum* has been placed in a separate genus, along with *Gemella haemolysans*, on the basis of genetic-relatedness studies. These species resemble viridans streptococci with respect to habitat in the human host and associated infections.

The transient viridans streptococcal bacteremia induced by eating, toothbrushing, flossing, and other sources of minor trauma, together with adherence to biologic surfaces, is thought to account for the predilection of these organisms to cause endocarditis **(see Fig. 123-1)**. Viridans streptococci are also isolated, often as part of a mixed flora, from sites of sinusitis, brain abscess, and liver abscess.

Viridans streptococcal bacteremia occurs relatively frequently in neutropenic patients, particularly after bone marrow transplantation or high-dose chemotherapy for cancer. Some of these patients develop a sepsis syndrome with high fever and shock. Risk factors for viridans streptococcal bacteremia include chemotherapy with high-dose cytosine arabinoside, prior treatment with trimethoprim-sulfamethoxazole or a fluoroquinolone, treatment with antacids or histamine antagonists, mucositis, and profound neutropenia.

The *S. milleri* group (also referred to as the *S. intermedius* or *S. anginosus* group) includes three species that cause human disease: *S. intermedius*, *S. anginosus*, and *Streptococcus constellatus*. These organisms are often considered viridans streptococci, although they differ somewhat from other viridans streptococci in both their hemolytic pattern (they may be α-, β-, or nonhemolytic) and the disease syndromes they cause. This group commonly produces suppurative infections, particularly abscesses of brain and abdominal viscera, and infections related to the oral cavity or respiratory tract, such as peritonsillar abscess, lung abscess, and empyema.

TREATMENT

TREATMENT

Infection with Viridans Streptococci

Isolates from neutropenic patients with bacteremia are often resistant to penicillin; thus these patients should be treated presumptively with vancomycin until the results of susceptibility testing become available. Viridans streptococci isolated in other clinical settings usually are sensitive to penicillin. Susceptibility testing should be performed to guide treatment of serious infections.

ABIOTROPHIA AND *GRANULICATELLA* SPECIES (NUTRITIONALLY VARIANT STREPTOCOCCI)

Occasional isolates cultured from the blood of patients with endocarditis fail to grow when subcultured on solid media. These *nutritionally variant streptococci* require supplemental thiol compounds or active forms of vitamin B_6 (pyridoxal or pyridoxamine) for growth in the laboratory. The nutritionally variant streptococci are generally grouped with the viridans streptococci because they cause similar types of infections. However, they have been

reclassified on the basis of 16S ribosomal RNA sequence comparisons into two separate genera: *Abiotrophia*, with a single species (*Abiotrophia defectiva*), and *Granulicatella*, with three species associated with human infection (*Granulicatella adiacens*, *Granulicatella para-adiacens*, and *Granulicatella elegans*).

TREATMENT

TREATMENT

Infection with Nutritionally Variant Streptococci

Treatment failure and relapse appear to be more common in cases of endocarditis due to nutritionally variant streptococci than in those due to the usual viridans streptococci. Thus the addition of gentamicin (1 mg/kg every 8 h for patients with normal renal function) to the penicillin regimen is recommended for endocarditis due to the nutritionally variant organisms.

OTHER STREPTOCOCCI

Streptococcus suis is an important pathogen in swine and has been reported to cause meningitis in humans, usually in individuals with occupational exposure to pigs. *S. suis* has been reported to be the most common cause of bacterial meningitis in Vietnam, and it has been responsible for outbreaks in China. Strains of *S. suis* associated with human infections have generally reacted with Lancefield group R typing serum and sometimes with group D typing serum as well. Isolates may be α- or β-hemolytic and are sensitive to penicillin. *Streptococcus iniae*, a pathogen of fish, has been associated with infections in humans who have handled live or freshly killed fish. Cellulitis of the hand is the most common form of human infection, although bacteremia and endocarditis have been reported. *Anaerobic streptococci*, or *peptostreptococci*, are part of the normal flora of the oral cavity, bowel, and vagina. **Infections caused by the anaerobic streptococci are discussed in Chap. 172.**

FURTHER READING

Bisno AL, Stevens DL: Streptococcal infections of skin and soft tissues. N Engl J Med 334:240, 1996.

Bruckner L, Gigliotti F: Viridans group streptococcal infections among children with cancer and the importance of emerging antibiotic resistance. Semin Pediatr Infect Dis 17:153, 2006.

Gerber MA et al: Prevention of rheumatic fever and diagnosis and treatment of acute streptococcal pharyngitis: A scientific statement from the American Heart Association Rheumatic Fever, Endocarditis, and Kawasaki Disease Committee of the Council on Cardiovascular Disease in the Young, the Interdisciplinary Council on Functional Genomics and Translational Biology, and the Interdisciplinary Council on Quality of Care and Outcomes Research: Endorsed by the American Academy of Pediatrics. Circulation 119:1541, 2009.

Gibbs RS et al: Perinatal infections due to group B streptococci. Obstet Gynecol 104:1062, 2004.

Phares CR et al: Epidemiology of invasive group B streptococcal disease in the United States, 1999–2005. JAMA 299:2056, 2008.

Chapter 144: Enterococcal Infections

Cesar A. Arias; Barbara E. Murray

FIGURE 144-1

INTRODUCTION

Enterococci have been recognized as potential human pathogens for more than a century, but only in recent years have these organisms acquired prominence as important causes of nosocomial infections. The ability of enterococci to survive and/or disseminate in the hospital environment and to acquire antibiotic resistance determinants makes the treatment of some enterococcal infections in critically ill patients a difficult challenge. Enterococci were first mentioned in the French literature in 1899; the “entérocoque” was found in the human gastrointestinal tract. The first pathologic description of an enterococcal infection dates to the same year. A clinical isolate from a patient who died as a consequence of endocarditis was initially designated *Micrococcus zymogenes*, was later named *Streptococcus faecalis* subspecies *zymogenes*, and would now be classified as *Enterococcus faecalis*. The ability of this isolate to cause severe disease in both rabbits and mice illustrated its potential lethality in the appropriate settings.

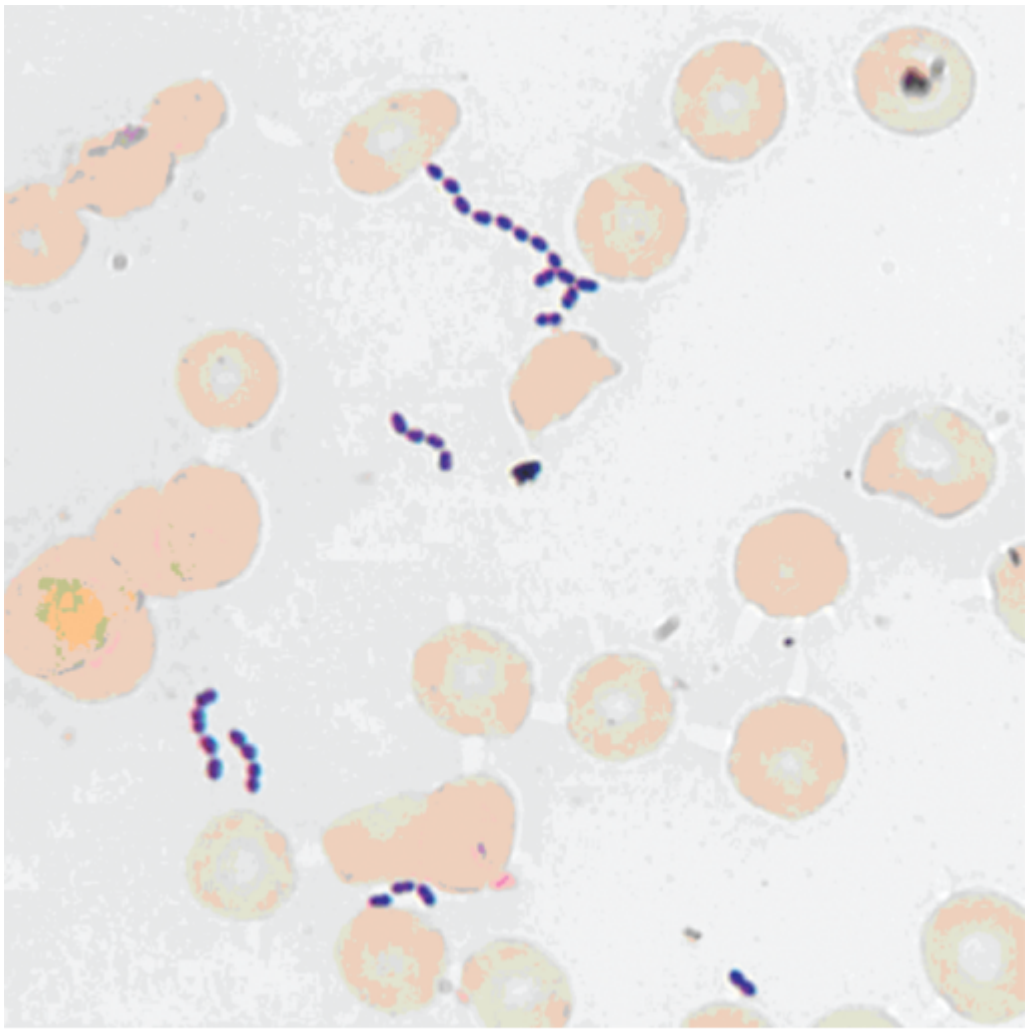
MICROBIOLOGY AND TAXONOMY

Enterococci are gram-positive organisms. In clinical specimens, they are usually observed as single cells, diplococci, or short chains **(Fig. 144-1)**, although long chains are noted with some strains. Enterococci were originally classified as streptococci because organisms of the two genera share many morphologic and phenotypic characteristics, including a generally negative catalase reaction. Only DNA hybridization studies and later 16S rRNA sequencing clearly demonstrated that enterococci should be grouped as a genus distinct from the streptococci. Nonetheless, unlike the majority of streptococci, enterococci hydrolyze esculin in the presence of 40% bile salts and grow at high salt concentrations (e.g., 6.5%) and at high temperatures (46°C). Enterococci are usually reported by the clinical laboratory to be nonhemolytic on the basis of their inability to lyse the ovine or bovine red blood cells (RBCs) commonly used in agar plates; however, some strains of *E. faecalis* do lyse RBCs from humans, horses, and rabbits. The majority of clinically relevant enterococcal species hydrolyze pyrrolidonyl-β-naphthylamide (PYR); this characteristic is helpful in differentiating enterococci from organisms of the *Streptococcus gallolyticus* group (formerly known as *S. bovis*), which includes *S. gallolyticus*, *S. pasteurianus*, and *S. infantarius*) and from *Leuconostoc* species. Although at least 18 species of enterococci have been isolated from human infections, the overwhelming majority of cases are

caused by two species, *E. faecalis* and *E. faecium*. Less frequently isolated species include *E. gallinarum*, *E. durans*, *E. hirae*, and *E. avium*.

FIGURE 144-1

Gram's stain of cultured blood from a patient with enterococcal bacteremia. Oval gram-positive bacterial cells are arranged as diplococci and short chains. *(Courtesy of Audrey Wanger, PhD.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOGENESIS

Enterococci are normal inhabitants of the large bowel of human adults, although they usually make up <1% of the culturable intestinal microbiota. In the healthy human host, enterococci are typical symbionts that coexist with other gastrointestinal bacteria; in fact, the utility of certain enterococcal strains as probiotics in the treatment of diarrhea suggests their possible role in maintaining the homeostatic equilibrium of the bowel. One of the most important factors that disrupts this equilibrium and promotes increased gastrointestinal colonization by enterococci is the administration of antimicrobial agents since enterococci are intrinsically resistant to a variety of commonly used antibacterial drugs. In particular, antibiotics that are excreted in the bile and have broad-spectrum activity (e.g., certain cephalosporins that target anaerobes and

gram-negative bacteria) are usually associated with the recovery of higher numbers of enterococci from feces. In the absence of antibiotics, hospital-associated lineages of *E. faecium* seem to be less adapted for survival in the gastrointestinal tract than are commensal *E. faecium* strains. However, in the presence of antimicrobial agents, the increased colonization by hospital-associated strains of *E. faecium* appears to be due not only to the simple filling of a "biological niche" after the eradication of competing components of the microbiota, but also (at least in mice) to the suppression—upon reduction of the gram-negative microflora by antibiotics—of important immunologic signals (e.g., by the lectin RegIIIγ) that contribute to the control of enterococcal counts in the normal human bowel. Several studies have shown that a higher level of gastrointestinal colonization is a critical factor in the pathogenesis of enterococcal infections. However, the mechanisms by which enterococci successfully colonize the bowel and gain access to the lymphatics and/or bloodstream remain incompletely understood. Recent data suggest that vancomycin-resistant enterococci (VRE) occupy distinct biological niches within the colonic lumen that fulfill their in vivo metabolic needs. Another factor that may contribute to enterococcal survival in the gastrointestinal tract is the production of bacteriocins (molecules that kill competing bacteria). Indeed, strains of *E. faecalis* harboring pheromone-producing plasmids that code for bacteriocins are capable of outcompeting enterococci lacking such plasmids. Furthermore, in vivo transfer of these plasmids occurs by conjugation, enhancing the survival of the recipients.

Several vertebrate, worm, and insect models have been developed to study the role of possible pathogenic determinants in both *E. faecalis* and *E. faecium*. Three main groups of virulence factors may increase the ability of enterococci to colonize the gastrointestinal tract and/or cause disease. The first group, *enterococcal secreted factors*, are molecules released outside the bacterial cell that contribute to the process of infection. The best-studied of these molecules include enterococcal hemolysin/cytolysin and two enterococcal proteases (gelatinase and serine protease). Enterococcal cytolysin is a heterodimeric toxin produced by some strains of *E. faecalis* that is capable of lysing human RBCs as well as polymorphonuclear leukocytes and macrophages. *E. faecalis* gelatinase and serine protease are thought to mediate virulence by several mechanisms, including the degradation of host tissues and the modification of critical components of the immune system. Mutants lacking the genes corresponding to these proteins are highly attenuated in experimental peritonitis, endocarditis, and endophthalmitis.

A second group of virulence factors, *enterococcal surface components*, includes adhesins and is thought to contribute to bacterial attachment to extracellular matrix molecules in the human host. Several molecules on the surface of enterococci have been characterized and shown to play a role in the pathogenesis of enterococcal infections. Among the characterized adhesins is *aggregation substance* of *E. faecalis*, which mediates the attachment of bacterial cells to each other, thereby facilitating conjugative plasmid exchange. Several lines of evidence indicate that aggregation substance and enterococcal cytolysin act synergistically to increase the virulence potential of *E. faecalis* strains in experimental endocarditis. The surface protein adhesin of collagen of *E. faecalis* (Ace) and its *E. faecium* homologue (Acm) are microbial surface components adhering to matrix molecules (MSCRAMMS); they recognize adhesive matrix molecules involved in bacterial attachment to host proteins such as collagen, fibronectin, and fibrinogen. Both Ace and Acm are important in the pathogenesis of experimental endocarditis. Pili of gram-positive bacteria are important mediators of

attachment to and invasion of host tissues and are considered potential targets for immunotherapy. Both *E. faecalis* and *E. faecium* have surface pili. Mutants of *E. faecalis* lacking pili are attenuated in biofilm production, experimental endocarditis, and urinary tract infections (UTIs). Other surface proteins that share structural homology with MSCRAMMs and appear to play a role in enterococcal attachment to the host and in virulence include the *E. faecalis* surface protein Esp and its *E. faecium* homologue Esp_{fm}, the second collagen adhesin of *E. faecium* (Scm), the surface proteins of *E. faecium* (Fms), SgrA (which binds to components of the basal lamina), and EcbA (which binds to collagen type V). Additional surface components apparently associated with pathogenicity include the Erl protein (a protein from the WxL family) and polysaccharides, which are thought to interfere with phagocytosis of the organism by host immune cells. Some *E. faecalis* strains appear to harbor at least three distinct classes of capsular polysaccharide; some of these polysaccharides play a role in virulence and are potential targets for immunotherapy. Teichoic acids on the enterococcal surface appear to be immunogenic, and antibodies to these molecules are protective in some animal models.

The third group of virulence factors has not been well characterized but includes the *E. faecalis* stress protein Gls24, which has been associated with enterococcal resistance to bile salts and appears to be important in the pathogenesis of endocarditis, and the hyl_{Efm}-containing plasmids of *E. faecium*, which are transferable between strains and increase gastrointestinal colonization by *E. faecium*. In mouse peritonitis, acquisition of these plasmids increased the lethality of a commensal strain of *E. faecium* and enhanced colonization of the uroepithelium. A gene encoding a regulator of oxidative stress (AsrR) has been identified as an important virulence factor of *E. faecium.*



The ability to sequence bacterial genomes has increased our understanding of bacterial diversity, evolution, pathogenesis, and mechanisms of antibiotic resistance. The genome sequences of more than 1000 enterococcal strains are currently available, and some have been entirely closed and annotated. Sequence analysis has shown that the genetic diversity of enterococci is related in large part to the acquisition of exogenous DNA and the mobilization of large chromosomal regions, resulting in recombination of the “core” genomes. In addition, analyses indicate that *E. faecium* harbors a malleable *accessory genome* incorporating a substantial content of exogenous elements, including DNA from phages. Indeed, a hospital-associated *E. faecium* clade that contains most clinical and outbreak-associated strains is the predominant genetic lineage circulating in hospitals around the world. This clade appears to be evolving rapidly, and genomic comparisons suggest that this lineage emerged 75–80 years ago—a time point that coincides with the introduction of antimicrobial drugs—and evolved, perhaps continuously, from animal strains, not from human commensal isolates. An initial genomic separation within *E. faecium* into human and animal commensals appears to have occurred ~3000 years ago, simultaneous with urbanization and domestication of animals. This genomic information provides new clues with regard to the evolution of enterococci from commensal organisms to important nosocomial pathogens.

EPIDEMIOLOGY

According to the National Healthcare Safety Network of the Centers for Disease Control and Prevention, enterococci are the second most common isolates (after staphylococci) from hospital-associated infections in the United States. Although *E. faecalis* remains the predominant species recovered from nosocomial infections, the isolation of *E. faecium* has increased substantially in the past 15–20 years. In fact, *E. faecium* is now almost as common as *E. faecalis* as an etiologic agent of hospital-associated infections. This point is important, since *E. faecium* is by far the most resistant and challenging enterococcal species to treat; indeed, more than 80% of *E. faecium* isolates recovered in U.S. hospitals are resistant to vancomycin, and more than 90% are resistant to ampicillin (historically the most effective β-lactam agent against enterococci). Resistance to vancomycin and ampicillin in *E. faecalis* isolates is much less common.

The dynamics of enterococcal transmission and dissemination in the hospital environment have been extensively studied, with a focus on VRE. These studies have revealed that VRE colonization of the gastrointestinal tract is a critical step in the development of enterococcal disease and that a substantial proportion of patients colonized with VRE remain colonized for prolonged periods (sometimes >1 year) and are more likely than patients without VRE colonization to develop an *Enterococcus*-related illness (e.g., bacteremia). Important factors associated with VRE colonization and persistence in the gut include prolonged hospitalization; long courses of antibiotic therapy; hospitalization in long-term-care facilities, surgical units, and/or intensive care units; organ transplantation; renal failure (particularly in patients undergoing hemodialysis) and/or diabetes; high APACHE scores; and physical proximity to patients infected or colonized with VRE or these patients' rooms. Once a patient becomes colonized with VRE, several key factors are involved in the organisms' dissemination in the hospital environment. VRE can survive exposure to heat and certain disinfectants and have been found on numerous inanimate objects in the hospital, including bed rails, medical equipment, doorknobs, gloves, telephones, and computer keyboards. Thus health care workers and the environment play pivotal roles in enterococcal transmission from patient to patient, and infection control measures are crucial in breaking the chain of transmission. Moreover, two meta-analyses have found that, independent of the patient's clinical status, VRE infection increases the risk of death over that among individuals infected with a glycopeptide-susceptible enterococcal strain.



The epidemiology of enterococcal disease and the emergence of VRE have followed slightly different trends in other parts of the world than in the United States. In Europe, the emergence of VRE in the mid-1980s was seen primarily in isolates recovered from animals and healthy humans rather than from hospitalized patients. The presence of VRE was associated with the use of the glycopeptide avoparcin as a growth promoter in animal feeds; this association prompted the European Union to ban the use of this compound in animal husbandry in 1996. However, after an initial decrease in the isolation of VRE from animals and humans, the prevalence of hospital-associated VRE infections has slowly increased in certain European countries, with important regional differences. For example, rates of vancomycin resistance among *E. faecium* clinical isolates in Europe are highest in Greece, the United Kingdom, and Portugal (10–30%), whereas rates in the Scandinavian countries and the Netherlands are <1%. These regional differences have been attributed in part to the implementation of aggressive "search-and-destroy" infection-control policies in countries such as the Netherlands; these policies have kept the frequency of nosocomial methicillin-resistant *Staphylococcus aureus* (MRSA) and VRE very low. In spite of regional differences, rates of VRE

continue to be much lower in Europe than in the United States. The reasons are not totally understood, although it has been postulated that this difference is related to the higher levels of human antibiotic use in the United States. Rates of enterococcal resistance to vancomycin in some Latin American countries are also lower (~4%) than those in the United States. Conversely, in Asia, rates of vancomycin resistance among enterococci appear to be similar to those in U.S. hospitals.



As mentioned above, genomic analyses of vancomycin-resistant *E. faecium* in different parts of the world suggest that the emergence and dissemination of these organisms in the hospital environment worldwide are due to the success of hospital-associated genetic lineages that acquired the genes responsible for vancomycin resistance as well as other antibiotic resistance determinants.

CLINICAL SYNDROMES

Urinary Tract Infection and Prostatitis

Enterococci are well-known causes of nosocomial UTI—the most common infection caused by these organisms (Chap. 130). Enterococcal UTIs are usually associated with indwelling catheterization, instrumentation, or anatomic abnormalities of the genitourinary tract, and it is often challenging to differentiate between true infection and colonization (particularly in patients with chronic indwelling catheters). The presence of leukocytes in the urine in conjunction with systemic manifestations (e.g., fever) or local signs and symptoms of infection with no other explanation and a positive urine culture ($\geq 10^5$ CFU/mL) suggests the diagnosis. Moreover, enterococcal UTIs often occur in critically or chronically ill patients whose comorbidities may obscure the diagnosis. In many cases, removal of the indwelling catheter may suffice to eradicate the organism without specific antimicrobial therapy. In rare circumstances, UTIs caused by enterococci may run a complicated course, with the development of pyelonephritis and perinephric abscesses that may be a portal of entry for bloodstream infections (see below). Enterococci are also known causes of chronic prostatitis, particularly in men whose urinary tract has been manipulated surgically or endoscopically. These infections can be difficult to treat since the agents most potent against enterococci (i.e., aminopenicillins and glycopeptides) penetrate prostatic tissue poorly. Chronic prostatic infection can be a source of recurrent enterococcal bacteremia.

Bacteremia and Endocarditis

Bacteremia without endocarditis is one of the most common presentations of enterococcal disease. Intravascular catheters and other devices are commonly associated with these bacteremic episodes (Chap. 137). Other well-known sources of enterococcal bacteremia include the gastrointestinal and hepatobiliary tracts; pelvic and intraabdominal foci; and, less frequently, wound infections, UTIs, and bone infections. In the United States, enterococci are ranked second (after coagulase-negative staphylococci) as etiologic agents of central line–associated bacteremia. Patients with enterococcal bacteremia usually have comorbidities and have been in the hospital for prolonged periods; they commonly have received several courses of antibiotics.

Several studies indicate that the isolation of *E. faecium* from the blood may lead to worse outcomes and higher mortality rates than when other enterococcal species are isolated; this finding may be related to the higher prevalence of vancomycin and ampicillin resistance in *E. faecium* than in other enterococcal species, with the consequent reduction of therapeutic options. In some cases (usually when the gastrointestinal tract is the source), enterococcal bacteremia may be polymicrobial, with gram-negative organisms isolated at the same time. In addition, several cases have been documented in which enterococcal bacteremia was associated with *Strongyloides stercoralis* hyperinfection syndrome in immunocompromised patients.

Enterococci are important causes of community- and health care–associated endocarditis, ranking second after staphylococci in the latter infections. The presumed initial source of bacteremia leading to endocarditis is the gastrointestinal or genitourinary tract—e.g., in patients who have malignant and inflammatory conditions of the gut or have undergone procedures in which these tracts are manipulated. The affected patients tend to be male and elderly and to have other debilitating diseases and heart conditions. Both prosthetic and native valves can be involved; mitral and aortic valves are affected most often. Community-associated endocarditis (usually caused by *E. faecalis*) also occurs in patients with no apparent risk factors or cardiac abnormalities. Endocarditis in women of childbearing age has been well described. The typical presentation of enterococcal endocarditis is a subacute course of fever, weight loss, malaise, and cardiac murmur; typical stigmata of endocarditis (e.g., petechiae, Osler's nodes, Roth's spots) are found in only a minority of patients. Atypical manifestations include arthralgias and manifestations of metastatic disease (splenic abscesses, hiccups, pain in the left flank, pleural effusion, and spondylodiscitis). Embolic complications are variable and can affect the brain. Heart failure is a common complication of enterococcal endocarditis, and valve replacement may be critical in curing this infection, particularly when multidrug-resistant organisms or major complications are involved. A recent clinical score (designated NOVA) has been proposed to help differentiate enterococcal bacteremia from true endocarditis. The duration of therapy is usually 4–6 weeks, with more prolonged courses suggested for multidrug-resistant isolates in the absence of valvular replacement.

Meningitis

Enterococcal meningitis is an uncommon disease (accounting for only ~4% of meningitis cases) that is usually associated with neurosurgical interventions and conditions such as shunts, central nervous system (CNS) trauma, and cerebrospinal fluid (CSF) leakage. In some instances—usually in patients with a debilitating condition, such as cardiovascular or congenital heart disease, chronic renal failure, malignancy, receipt of immunosuppressive therapy, or HIV/AIDS—presumed hematogenous seeding of the meninges is seen in infections such as endocarditis or bacteremia. Fever and changes in mental status are common, whereas overt meningeal signs are less so. CSF findings are consistent with bacterial infection—i.e., pleocytosis, with a predominance of polymorphonuclear leukocytes (average, ~500/μL), an elevated serum protein level (usually >100 mg/dL), and a decreased glucose concentration (average, 28 mg/dL). Gram's staining yields a positive result in about half of cases, with a high rate of organism recovery from CSF cultures; the most common species isolated are *E. faecalis* and *E. faecium*. Complications include

hydrocephalus, brain abscesses, and stroke. As mentioned before for bacteremia, an association with *Strongyloides* hyperinfection has also been documented.

Intraabdominal, Pelvic, and Soft Tissue Infections

As mentioned earlier, enterococci are part of the commensal microbiota of the gastrointestinal tract and can produce spontaneous peritonitis in cirrhotic individuals and in patients undergoing chronic ambulatory peritoneal dialysis (Chap. 127). These organisms are commonly found (usually along with other bacteria, including enteric gram-negative species and anaerobes) in clinical samples from intraabdominal and pelvic collections. The presence of enterococci in intraabdominal infections is sometimes considered to be of little clinical relevance. Several studies have shown that the role of enterococci in intraabdominal infections originating in the community and involving previously healthy patients is minor, since surgery and broad-spectrum antimicrobial drugs that do not target enterococci are often sufficient to treat these infections successfully. In the last few decades, however, these organisms have become prominent as a cause of intraabdominal infections in hospitalized patients because of the emergence and spread of vancomycin resistance among enterococci and the increase in rates of nosocomial infections due to multidrug-resistant *E. faecium* isolates. In fact, several studies have now documented treatment failures due to enterococci, with consequently increased rates of postoperative complications and death among patients with intraabdominal infections. Thus, anti-enterococcal therapy is recommended for nosocomial peritonitis in immunocompromised and severely ill patients who have had a prolonged hospital stay, have undergone multiple procedures, have persistent abdominal sepsis and collections, or have risk factors for the development of endocarditis (e.g., prosthetic or damaged heart valves). Conversely, specific treatment for enterococci in the first episode of intraabdominal infection originating in the community and affecting previously healthy patients with no important cardiac risk factors for endocarditis does not appear to be beneficial.

Enterococci are commonly isolated from soft tissue infections (Chap. 124), particularly those involving surgical wounds (Chap. 137). In fact, these organisms rank third as agents of nosocomial surgical-site infections, with *E. faecalis* the most frequently isolated species. The clinical relevance of enterococci in some of these infections—as in intraabdominal infections—is a matter of debate; differentiating between colonization and true infection is sometimes challenging, although in some cases enterococci have been recovered from lung, liver, and skin abscesses. Diabetic foot and decubitus ulcers are often colonized with enterococci and may be the portal of entry for bone infections.

Other Infections

Enterococci are well-known causes of neonatal infections, including sepsis (mostly late-onset), bacteremia, meningitis, pneumonia, and UTI. Outbreaks of enterococcal sepsis in neonatal units have been well documented. Risk factors for enterococcal disease in newborns include prematurity, low birth weight, indwelling devices, and abdominal surgery. Enterococci have also been described as etiologic agents of bone and joint infections, including vertebral osteomyelitis, usually in patients with underlying conditions such as

diabetes or endocarditis. Similarly, enterococci have been isolated from bone infections in patients who have undergone arthroplasty or reconstruction of fractures with the placement of hardware. Since enterococci can produce a biofilm that is likely to alter the efficacy of anti-enterococcal agents, treatment of infections that involve foreign material is challenging, and removal of the hardware may be necessary to eradicate the infection. Rare cases of enterococcal pneumonia, lung abscess, and spontaneous empyema have been described.

TREATMENT

TREATMENT

Enterococcal Infections

General Principles

Enterococci are intrinsically resistant and/or tolerant to several antimicrobial agents. (*Tolerance* is defined as lack of killing by drug concentrations 32 times higher than the minimal inhibitory concentration [MIC].) Monotherapy for endocarditis with a β-lactam antibiotic (to which many enterococci are tolerant) has produced disappointing results, with high relapse rates after the end of therapy. However, the addition of an aminoglycoside to a cell wall–active agent (a β-lactam or a glycopeptide) increases cure rates and eradicates the organisms; moreover, this combination is synergistic and bactericidal in vitro. Therefore, for many decades, combination therapy with a cell wall–active agent and an aminoglycoside was the standard of care for endovascular infections caused by enterococci. This synergistic effect can be explained, at least in part, by the increased penetration of the aminoglycoside into the bacterial cell, presumably as a result of cell wall alterations produced by the β-lactam (or glycopeptide). Nonetheless, attaining synergistic bactericidal activity in the treatment of severe enterococcal infections—particularly those caused by *E. faecium*—has become increasingly difficult because of the development of resistance to virtually all antibiotics available for this purpose.

The treatment of *E. faecalis* differs substantially from that of *E. faecium* (Tables 144-1 and 144-2), mainly because of differences in resistance profiles (see below). For example, resistance to ampicillin and vancomycin is rare in *E. faecalis*, whereas these antibiotics are only infrequently useful against current isolates of *E. faecium*. Moreover, as a consequence of the challenges and therapeutic limitations posed by the emergence of drug resistance in enterococci, valve replacement may need to be considered in the treatment of endocarditis caused by multidrug-resistant enterococci. Less severe infections are often related to indwelling intravascular catheters; removal of the catheter increases the likelihood of enterococcal eradication by a subsequent short course of appropriate antimicrobial therapy.

Choice of Antimicrobial Agents

Among the β-lactams, the most active are the aminopenicillins (ampicillin, amoxicillin) and ureidopenicillins (i.e., piperacillin); next most active are penicillin G and imipenem. For *E. faecium*, a combination of high-dose ampicillin (up to 30 g/d) plus an aminoglycoside has been suggested—even for ampicillin-resistant strains if the MIC is ≤64 μg/mL—since a plasma ampicillin concentration higher than this value can be achieved at high

doses. The only two aminoglycosides recommended for synergistic therapy in severe enterococcal infections are gentamicin and streptomycin. The use of amikacin is strongly discouraged, tobramycin should never be used for the treatment of *E. faecium* infection, and aminoglycoside monotherapy should not be employed. Vancomycin is an alternative to β-lactam drugs for the treatment of *E. faecalis* infections but is less useful against *E. faecium* because resistance is common.

As mentioned above, use of the aminoglycoside–ampicillin combination for *E. faecalis* infections has become increasingly problematic because of toxicity in critically ill patients and increased rates of high-level resistance to aminoglycosides. An observational, nonrandomized, comparative study encompassing a multicenter cohort was conducted in 17 Spanish hospitals and one Italian hospital; the results indicated that a 6-week course of ampicillin plus ceftriaxone is as effective as ampicillin plus gentamicin in the treatment of *E. faecalis* endocarditis, with less risk of toxicity. Therefore, this regimen should be considered in patients at risk for aminoglycoside toxicity and is now recommended as first-line therapy.

Linezolid is the only agent approved by the U.S. Food and Drug Administration (FDA) for the treatment of VRE infections (Table 144-2). (A prior approval for quinupristin/dalfopristin has been withdrawn.) Linezolid is not bactericidal, and its use in severe endovascular infections has produced mixed results; therefore, it is recommended only as an alternative to other agents for such infections. In addition, linezolid may cause significant toxicities (thrombocytopenia, peripheral neuropathy, and optic neuritis) when used in regimens given for >2 weeks. Nonetheless, linezolid may play a role in the treatment of enterococcal meningitis and other CNS infections, although clinical data are limited.

The lipopeptide daptomycin is a bactericidal antibiotic with potent in vitro activity against all enterococci. Although daptomycin is not approved by the FDA for the treatment of VRE or *E. faecium* infections, it has been used alone (at high dosage) or in combination with other agents (ampicillin, ceftaroline, and tigecycline) with apparent success against multidrug-resistant enterococcal infections (Tables 144-1 and 144-2). The main adverse reactions to daptomycin are elevated creatine phosphokinase levels and eosinophilic pneumonitis (rare). Daptomycin is not useful against pulmonary infections because the pulmonary surfactant inhibits its antibacterial activity.

The glycylcycline drug tigecycline is active in vitro against all enterococci, regardless of the isolates' vancomycin susceptibility. However, its use as monotherapy for endovascular or severe enterococcal infections is not recommended because of low attainable blood levels.

Telavancin, a lipoglycopeptide approved by the FDA for the treatment of skin and soft-tissue infections as well as hospital-associated pneumonia, is active against vancomycin-susceptible enterococci but not VRE. Oritavancin, a novel glycopeptide with activity against VRE, has been approved for the treatment of acute bacterial skin and soft-tissue infections caused by susceptible organisms, including vancomycin-susceptible *E. faecalis*. The MICs of oritavancin against VRE are low, and this compound may be a promising drug for VRE treatment in the future.

Lastly, tedizolid—a new oxazolidinone now available for clinical use—is approved only for the treatment of *E. faecalis* infections. Tedizolid is more potent than linezolid in vitro against VRE strains; however, its role in severe VRE infections remains to be determined.

Antimicrobial Resistance

Resistance to β-lactam agents continues to be observed only infrequently in *E. faecalis* but is characteristic of *E. faecium*. The mechanism of ampicillin resistance in *E. faecium* is related to a penicillin-binding protein (PBP) designated PBP5, which is the target of β-lactam antibiotics. PBP5 exhibits low affinity for ampicillin and can synthesize cell wall in the presence of this antibiotic, even when other PBPs are inhibited. The version of this protein found in ampicillin-resistant hospital-associated strains has multiple amino-acid differences that even further decrease the affinity of PBP5 for ampicillin; these changes and/or hyperproduction of PBP5 are the two most common mechanisms of high-level ampicillin resistance (e.g., MIC, >32 μg/mL) in clinical strains.

Vancomycin is a glycopeptide antibiotic that inhibits cell-wall peptidoglycan synthesis in susceptible enterococci and has been widely used against enterococcal infections in clinical practice when the utility of β-lactams is limited by resistance, allergy, or adverse reactions. This effect is mediated by binding of the antibiotic to peptidoglycan precursors (UDP-MurNAc-pentapeptides) upon their exit from the bacterial cell cytoplasm. The interaction of vancomycin with the peptidoglycan is specific and involves the last two D-alanine residues of the precursor. The first isolates of VRE were documented in 1986, and vancomycin resistance (particularly in *E. faecium*) has since increased considerably around the world. The mechanism involves the replacement of the last D-alanine residue of peptidoglycan precursors with D-lactate or D-serine, with consequent high- and low-level resistance, respectively. There is significant heterogeneity among isolates, but either substitution substantially decreases the affinity of vancomycin for the peptidoglycan; with the D-lactate substitution, the MIC is increased up to 1000-fold. Vancomycin-resistant organisms also produce enzymes that destroy the D-alanine-D-alanine ending precursors, ensuring that additional binding sites for vancomycin are not available.

High-level resistance to aminoglycosides (of which gentamicin and streptomycin are the only two tested by clinical laboratories) abolishes the synergism observed between cell wall–active agents and the aminoglycoside. This important phenotype is routinely sought by the clinical laboratory in isolates from serious infections (Tables 144-1 and 144-2). Genes encoding aminoglycoside-modifying enzymes are usually the cause of high-level resistance to these compounds and are widely disseminated among enterococci, decreasing the options for the treatment of severe enterococcal infections.

Resistance to daptomycin has now been well documented in both *E. faecalis* and *E. faecium*. This resistance seems to be related to activation of the cell-membrane stress response that, in enterococci, is mainly regulated by a three-component system designated LiaFSR, although other systems have also been implicated. In addition, changes in the enzymes responsible for phospholipid metabolism are involved. Most important, activation of the LiaFSR system leads to tolerance to daptomycin and minor increases in MICs. Therefore, isolates with MICs of 3–4 μg/mL (close to the breakpoint of 4 μg/mL) should be judged to be

potentially resistant, and alternative therapies (including combinations of daptomycin with ß-lactams; see above) should be considered.

Resistance to linezolid is usually due to mutations in the 23S rRNA genes or the presence of an rRNA methylase (designated *cfr*). A novel transferable gene (*optrA*) encoding a putative efflux pump has been implicated in linezolid resistance in enterococcal strains of human and animal origin.

Tigecycline resistance has been documented and appears to be related to changes in the S10 ribosomal protein.

TABLE 144-1

Suggested Regimens for the Management of Infections Caused by *Enterococcus faecalis*

Clinical Syndrome	Suggested Therapeutic Options[a]
Endovascular infections (including endocarditis)	• Ampicillin[b] (12 g/d IV in divided doses q4h or by continuous infusion) or penicillin (18–30 mU/d IV in divided doses q4h or by continuous infusion) *plus* an aminoglycoside[c] • Ampicillin[b] (12 g/d IV in divided doses q4h) plus ceftriaxone (2 g IV q12h) • Vancomycin[d] (15 mg/kg IV per dose) plus an aminoglycoside[c] • High-dose daptomycin[e] ± another active agent[f] • Ampicillin[b] plus imipenem
Non-endovascular bacteremia[g]	• Ampicillin (12 g/d IV in divided doses q4h) or penicillin (18 mU/d IV in divided doses q4h) ± an aminoglycoside[c] or ceftriaxone • Vancomycin[d] (15 mg/kg IV per dose) • High-dose daptomycin[e] ± another active agent[f] • Linezolid (600 mg IV/PO q12h)
Meningitis	• Ampicillin (20–24 g/d IV in divided doses q4h) or penicillin (24 mU/d IV in divided doses q4h) *plus* an aminoglycoside[c,h] and consider adding ceftriaxone (2 g IV q12h) • Vancomycin (500–750 mg IV q6h)[d] *plus* an aminoglycoside[c] or rifampin • Linezolid • High-dose daptomycin[e] (plus intrathecal daptomycin) ± another active agent[f]
Urinary tract Infections (uncomplicated)	• Fosfomycin (3 g PO, one dose)[i] • Ampicillin (500 mg IV or PO q6h) • Nitrofurantoin (100 mg PO q6h)

[a]Authors' preferences are underlined for each category; many of the regimens are off-label. [b]In rare cases, β-lactamase-producing isolates may be present. Because these isolates are not detected by conventional determination of the minimal inhibitory concentration, additional tests (e.g., the nitrocefin disk) are recommended for isolates from endocarditis. The use of ampicillin/sulbactam (12–24 g/d) is suggested in these cases. [c]Only if the organism does not exhibit high-level resistance (HLR) to aminoglycosides. This test is performed by the clinical microbiology laboratory

only for gentamicin or streptomycin (growth of enterococci on agar containing gentamicin [500 µg/mL] or streptomycin [2000 µg/mL]). If HLR is documented, the aminoglycoside will not act synergistically with the other agent in the combination. However, HLR to one of these aminoglycosides does not indicate resistance to the other agent (as reported individually). HLR to gentamicin implies lack of synergism with tobramycin and with amikacin. Gentamicin (1–1.5 mg/kg IV q8h) and streptomycin (15 mg/kg per day IV/IM in two divided doses) are the only two recommended aminoglycosides. [d]Vancomycin is recommended only as an alternative to β-lactam agents in cases of allergy or toxicity plus the inability to desensitize. Cerebrospinal fluid (CSF) concentrations in meningitis should be determined. Vancomycin-resistant strains of *E. faecalis* have been reported. [e]Consider doses of 10–12 mg/kg once daily if used in combination and 10–12 mg/kg day if used alone. Monitoring of creatine phosphokinase levels is recommended throughout therapy because of possible rhabdomyolysis. [f]Potentially active agents may include an aminoglycoside (if HLR is not detected), ampicillin, ceftaroline, tigecycline, or a fluoroquinolone (which, if the isolate is susceptible, may be favored in meningitis). The presence of mutations in *liaFSR* seems to increase susceptibility to ampicillin and ceftaroline, and combinations of daptomycin with these compounds are bactericidal in vitro against such strains. [g]In selected cases of catheter-associated bacteremia, removal of the catheter and a short course of therapy (~5–7 days) may be sufficient. A single positive blood culture that is likely to be associated with a catheter in a patient who is otherwise doing well may not require therapy after removal of the catheter. Patients at high risk for endovascular infections or with severe disease may benefit from synergistic combination therapy. [h]The addition of intrathecal or intraventricular therapy with gentamicin (2–10 mg/d) if the organism does not exhibit HLR or with vancomycin (10–20 mg/d) when the isolate is susceptible has been suggested by some authorities. The addition of systemic rifampin (a good CSF-penetrating agent) may be considered. The combination of ampicillin and ceftriaxone may have clinical benefit (by analogy with endocarditis), but no cases treated with this combination have been reported; the authors would use this combination. [i]Approved by the U.S. Food and Drug Administration only for uncomplicated urinary tract infections caused by vancomycin-susceptible *E. faecalis.*

TABLE 144-2

Suggested Regimens for the Management of Infections Caused by Vancomycin- and Ampicillin-Resistant *Enterococcus faecium*

Clinical Syndrome	Suggested Therapeutic Options[a]
Endovascular infections (including endocarditis)	• High-dose daptomycin[b] plus another agent[c] ± an aminoglycoside[d] • Linezolid (600 mg IV q12h) • High-dose ampicillin (if MIC is ≤64 µg/mL) ± an aminoglycoside[d] • Ampicillin plus imipenem (if the ampicillin MIC is ≤32 µg/mL) • Q/D[e] (22.5 mg/kg per day in divided doses q8h) ± another active agent[f]
Non-endovascular bacteremia[g]	• High-dose daptomycin[b] ± another agent[c] ± an aminoglycoside[d] • Linezolid (600 mg IV q12h) • Q/D (22.5 mg/kg per day in divided doses q8h) ± another active agent[f]
Meningitis	• Linezolid (600 mg IV q12h) ± another CSF-penetrating active agent[h] • High-dose daptomycin[b] (plus intraventricular daptomycin) ± another CSF-penetrating active agent[h,j] • Q/D (22.5 mg/kg per day in divided doses q8h plus intraventricular Q/D)[i] ± another active agent[h]
Urinary tract infections	• Fosfomycin (3 g PO, one dose)[k] • Nitrofurantoin (100 mg PO q6h) • Ampicillin or amoxicillin (2 g IV/PO q4–6h)[l]

[a]Authors' preferences are underlined for each category; many of these regimens are off-label. [b]Daptomycin at doses of 10–12 mg/kg once daily is suggested (off-label). Close monitoring of creatine phosphokinase levels is recommended throughout therapy because of possible rhabdomyolysis. [c]Potentially active agents may include ampicillin or ceftaroline (even if the infecting strain is resistant in vitro) or tigecycline. In vitro synergism of daptomycin with ampicillin or ceftaroline has been observed against some isolates that subsequently become nonsusceptible to daptomycin during therapy. The synergism of daptomycin and β-lactams is associated with mutations in *liaFSR*. Consider combination therapy if the minimal inhibitory concentration (MIC) of daptomycin is ≥3 µg/mL. [d]Only if the organism does not exhibit high-level resistance to aminoglycosides (see Table 144-1, footnote *c*). [e]Quinupristin-

dalfopristin (Q/D) lost U.S. Food and Drug Administration (FDA) approval for endocarditis due to VRE. [f]Agents that may be useful in combination with Q/D (if the isolate is susceptible to each agent) include doxycycline with rifampin (one reported case) or fluoroquinolones (one reported case). [g]In selected cases of catheter-associated bacteremia, removal of the catheter and a short course of therapy (~5–7 days) may be sufficient. A single positive blood culture that is likely to be associated with a catheter in a patient who is otherwise doing well may not require therapy after removal of the catheter. [h]Fluoroquinolones (e.g., moxifloxacin) and rifampin (if the isolate is susceptible to each agent) reach therapeutic levels in the cerebrospinal fluid. [i]Intrathecal Q/D (1–5 mg/d) has been used in combination with Q/D systemic therapy in meningitis. If Q/D is chosen, simultaneous use of both systemic and intrathecal therapy is suggested. [j]Intrathecal gentamicin (2–10 mg/d) if high-level resistance is not detected. Intraventricular daptomycin has been used in two cases of meningitis. [k]Approved by the FDA only for uncomplicated urinary tract infections caused by vancomycin-susceptible *E. faecalis*. [l]Concentrations of amoxicillin and ampicillin in urine far exceed those in serum and may be potentially effective even against isolates with high MICs. Doses up to 12 g/d are suggested for isolates with MICs of ≥64 µg/mL.

FURTHER READING

Bouza E et al: The NOVA score: A proposal to reduce the need for transesophageal echocardiography in patients with enterococcal bacteremia. Clin Infect Dis 60:528, 2015.

Brandl K et al: Vancomycin-resistant enterococci exploit antibiotic-induced innate immune deficits. Nature 455:804, 2008.

Kommineni S et al: Bacteriocin production augments niche competition by enterococci in the mammalian gastrointestinal tract. Nature 29:719, 2015.

Lebreton F et al: Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. MBio 20:e00534, 2013.

Miller WR et al: Vancomycin-resistant enterococci: Therapeutic challenges in the 21st century. Infect Dis Clin North Am 30:415, 2016.

Chapter 145: Diphtheria and Other Corynebacterial Infections

William R. Bishai; John R. Murphy

DIPHTHERIA

Diphtheria is a nasopharyngeal and skin infection caused by *Corynebacterium diphtheriae*. Toxigenic strains of *C. diphtheriae* produce a protein toxin that causes systemic toxicity, myocarditis, and polyneuropathy. The toxin is associated with the formation of pseudomembranes in the pharynx during respiratory diphtheria. While toxigenic strains most frequently cause pharyngeal diphtheria, nontoxigenic strains commonly cause cutaneous disease.

ETIOLOGY

C. diphtheriae is a gram-positive bacillus that is unencapsulated, nonmotile, and nonsporulating. The organism was first identified microscopically in 1883 by Klebs and a year later was isolated in pure culture by Löffler in Robert Koch's laboratory. The bacteria have a characteristic club-shaped bacillary appearance and typically form clusters of parallel rays, or *palisades*, that are referred to as "Chinese characters." The specific laboratory media recommended for the cultivation of *C. diphtheriae* rely upon tellurite, colistin, or nalidixic acid for the organism's selective isolation from other autochthonous pharyngeal microbes. *C. diphtheriae* may be isolated from individuals with both nontoxigenic (*tox*$^-$) and toxigenic (*tox*$^+$) phenotypes. Uchida and Pappenheimer demonstrated that corynebacteriophage beta carries the structural gene *tox,* which encodes diphtheria toxin, and that a family of closely related corynebacteriophages are responsible for toxigenic conversion of *tox*$^-$ *C. diphtheriae* to the *tox*$^+$ phenotype. Moreover, lysogenic conversion from a nontoxigenic to a toxigenic phenotype has been shown to occur in situ. Growth of toxigenic strains of *C. diphtheriae* under iron-limiting conditions leads to the optimal expression of diphtheria toxin and is believed to be a pathogenic mechanism during human infection. Less commonly, diphtheria-like disease may be caused by *Corynebacterium ulcerans* and *Corynebacterium pseudotuberculosis*, which express the same toxin and are considered members of the *C. diphtheriae* group (discussed below).

EPIDEMIOLOGY

While in many regions diphtheria has been controlled in recent years with effective vaccination, there have been sporadic outbreaks in the United States and Europe. Diphtheria is still common in the Caribbean,

Latin America, and the Indian subcontinent, where mass immunization programs are not enforced. Large-scale epidemics of diphtheria have occurred in the post–Soviet Union independent states. Additional outbreaks have recently been reported in Africa and Asia. In temperate regions, respiratory diphtheria occurs year-round but is most common during winter months.

C. diphtheriae is transmitted via the aerosol route, usually during close contact with an infected person. There are no significant reservoirs other than humans. The incubation period for respiratory diphtheria is 2–5 days, but disease onset has occurred as late as 10 days after exposure. Prior to the vaccination era, most individuals over the age of 10 were immune to *C. diphtheriae*; infants were protected by maternal IgG antibodies but became susceptible after ~6 months of age. Thus, the disease primarily affected children and nonimmune young adults.

The development of diphtheria antitoxin in 1898 by von Behring and of the diphtheria toxoid vaccine in 1924 by Ramon led to the near-elimination of diphtheria in Western countries. The annual incidence rate in the United States peaked in 1921, with 206,000 cases (191 cases per 100,000) and 15,520 deaths. In contrast, since 1980, the annual figure in the United States has been fewer than 5 cases per 100,000, with only two cases reported from 2004 through 2015. Nevertheless, pockets of colonization persist in North America, and groups or individuals who resist vaccination remain at risk. Immunity to diphtheria induced by childhood vaccination gradually decreases in adulthood. An estimated 30% of men 60–69 years old have antitoxin titers below the protective level. In addition to older age and lack of vaccination, risk factors for diphtheria outbreaks include alcoholism, low socioeconomic status, crowded living conditions, and Native American ethnic background. An outbreak of diphtheria in Seattle, Washington, between 1972 and 1982 comprised 1100 cases, most of which were cutaneous. During the 1990s in the states of the former Soviet Union, a much larger diphtheria epidemic included more than 140,000 cases and more than 4000 deaths; at its peak in 1995, more than 50,412 cases were reported. Clonally related toxigenic *C. diphtheriae* strains of the ET8 complex were associated with this outbreak. Beginning in 1998, this epidemic was controlled by mass vaccination programs, and between 2000 and 2009 the diphtheria incidence fell by >95%, with high-burden countries such as Latvia reporting fewer than 10 cases. During the epidemic, the incidence rate was high among individuals between 16 and 50 years of age. The epidemic was attributed to multiple factors, including socioeconomic instability, migration, deteriorating public health programs, unnecessary contraindications to vaccination, low-dose vaccine formulations, frequent vaccine and antitoxin shortages, delayed implementation of vaccination and treatment in response to cases, public mistrust, and lack of awareness.

Since 2010, significant outbreaks of diphtheria and diphtheria-related mortality have continued to be reported from many developing countries, including the Dominican Republic, Nigeria, India, Laos, Thailand, Indonesia, and Brazil. Statistics collected by the World Health Organization indicated that 7321 diphtheria cases were reported in 2014, but many more cases are likely to have gone unreported. Although 86% of the global population has been adequately vaccinated, only 28% of countries have successfully vaccinated >80% of individuals in all districts.

Cutaneous diphtheria is usually a secondary infection that follows a primary skin lesion due to trauma, allergy, or autoimmunity. Most often, these isolates lack the *tox* gene and thus do not express diphtheria

toxin. In tropical latitudes, cutaneous diphtheria is more common than respiratory diphtheria. In contrast to respiratory disease, cutaneous diphtheria is not reportable in the United States. Nontoxigenic strains of *C. diphtheriae* have been associated with pharyngitis in Europe, causing outbreaks among men who have sex with men and persons who use illicit IV drugs.

PATHOGENESIS AND IMMUNOLOGY

Diphtheria toxin produced by *tox*$^{+}$ strains of *C. diphtheriae* is the primary virulence factor in clinical disease. The toxin is synthesized in precursor form; is released as a 535-amino-acid, single-chain protein; and, in sensitive species (e.g., guinea pigs and humans, but not mice or rats), has a 50% lethal dose of ~100 ng/kg of body weight. The toxin is produced in the pseudomembranous lesion and is taken up in the bloodstream, from which it is distributed to all organ systems in the body. Once bound to its cell surface receptor (a heparin-binding epidermal growth factor–like precursor), the toxin is internalized by receptor-mediated endocytosis and enters the cytosol from an acidified early endosomal compartment. In vitro, the toxin may be separated into two chains by digestion with serine proteases: the N-terminal A fragment and the C-terminal B fragment. Delivery of the A fragment into the eukaryotic cell cytosol results in irreversible inhibition of protein synthesis by NAD^{+}-dependent ADP-ribosylation of elongation factor 2. The eventual result is the death of the cell.

In 1926, Ramon at the Institut Pasteur found that formalinization of diphtheria toxin resulted in the production of a nontoxic but highly immunogenic diphtheria toxoid. Subsequent studies showed that immunization with diphtheria toxoid elicited antibodies that neutralized the toxin and prevented most disease manifestations. In the 1930s, mass immunization of children and susceptible adults with diphtheria toxoid commenced in the United States and Europe.

Individuals with a diphtheria antitoxin titer of >0.01 U/mL are at low risk of disease. In populations where a majority of individuals have protective antitoxin titers, the carrier rate for toxigenic strains of *C. diphtheriae* decreases and the overall risk of diphtheria among susceptible individuals is reduced. Nevertheless, individuals with nonprotective titers may contract diphtheria through either travel or exposure to individuals who have recently returned from regions where the disease is endemic.

Characteristic pathologic findings of diphtheria include mucosal ulcers with a pseudomembranous coating composed of an inner band of fibrin and a luminal band of neutrophils. Initially white and firmly adherent, in advanced diphtheria the pseudomembranes turn gray or even green or black as necrosis progresses. Mucosal ulcers result from toxin-induced necrosis of the epithelium accompanied by edema, hyperemia, and vascular congestion of the submucosal base. A significant fibrinosuppurative exudate from the ulcer develops into the pseudomembrane. Ulcers and pseudomembranes in severe respiratory diphtheria may extend from the pharynx into medium-sized bronchial airways. Expanding and sloughing membranes may result in fatal airway obstruction.

APPROACH TO THE PATIENT

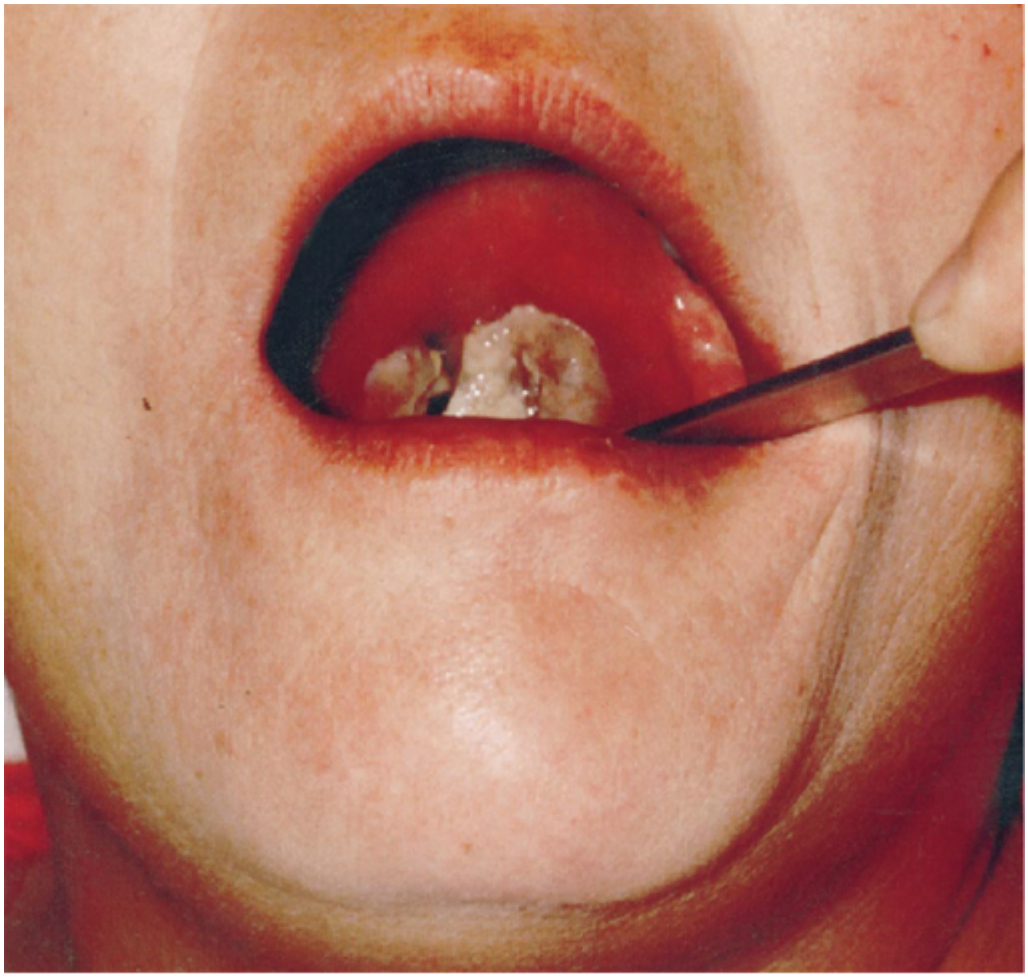
APPROACH TO THE PATIENT

Diphtheria

Diphtheria, although rare in the United States and other developed countries, should be considered when a patient has severe pharyngitis, particularly when there is difficulty swallowing, respiratory compromise, or signs of systemic disease (e.g., myocarditis or generalized weakness). The leading causes of pharyngitis are respiratory viruses (rhinoviruses, influenza viruses, parainfluenza viruses, coronaviruses, adenoviruses; ~25% of cases), group A streptococci (15–30%), group C streptococci (~5%), atypical bacteria such as *Mycoplasma pneumoniae* and *Chlamydia pneumoniae* (15–20% in some series), and other viruses such as herpes simplex virus (~4%) and Epstein-Barr virus (<1% in infectious mononucleosis). Less common causes are acute HIV infection, gonorrhea, fusobacterial infection (e.g., Lemierre's syndrome), thrush due to *Candida albicans* or other *Candida* species, and diphtheria. The presence of a pharyngeal pseudomembrane or an extensive exudate should prompt consideration of diphtheria **(Figure 145-1)**.

FIGURE 145-1

Respiratory diphtheria due to toxigenic *C. diphtheriae* producing exudative pharyngitis in a 47-year-old woman displaying neck edema and a pseudomembrane extending from the uvula to the pharyngeal wall. The characteristic white pseudomembrane is caused by diphtheria toxin–mediated necrosis of the respiratory epithelial layer, producing a fibrinous coagulative exudate. Submucosal edema adds to airway narrowing. The pharyngitis is acute in onset, and respiratory obstruction from the pseudomembrane may occur in severe cases. Inoculation of pseudomembrane fragments or submembranous swabs onto Löffler's or tellurite selective medium reveals *C. diphtheriae. (Photograph by P Strebel, MD, used by permission. From R Kadirova et al: J Infect Dis 181:S110, 2000. With permission of Oxford University Press.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CLINICAL MANIFESTATIONS

Respiratory Diphtheria

The clinical diagnosis of diphtheria is based on the constellation of sore throat; adherent tonsillar, pharyngeal, or nasal pseudomembranous lesions; and low-grade fever. In addition, diagnosis requires the isolation of *C. diphtheriae* or histopathologic isolation of compatible gram-positive organisms. The Centers for Disease Control and Prevention (CDC) recognizes *confirmed* respiratory diphtheria (laboratory proven or epidemiologically linked to a culture-confirmed case) and *probable* respiratory diphtheria (clinically compatible but not laboratory proven or epidemiologically linked). Carriers are defined as individuals who have positive cultures for *C. diphtheriae* and who either are asymptomatic or have symptoms but lack pseudomembranes. Most patients seek medical care for sore throat and fever several days into the illness. Occasionally, weakness, dysphagia, headache, and voice change are the initial manifestations. Neck edema and difficulty breathing are evident in more advanced cases and carry a poor prognosis.

The systemic manifestations of diphtheria stem from the effects of diphtheria toxin and include weakness as a result of neurotoxicity and cardiac arrhythmias or congestive heart failure due to myocarditis. Most commonly, the pseudomembranous lesion is located in the tonsillopharyngeal region. Less commonly, the

lesions are located in the larynx, nares, and trachea or bronchial passages. Large pseudomembranes are associated with severe disease and a poor prognosis. A few patients develop massive swelling of the tonsils and present with “bull-neck” diphtheria, which results from edema of the submandibular and paratracheal region and is further characterized by foul breath, thick speech, and stridorous breathing. The diphtheritic pseudomembrane is gray or whitish and sharply demarcated. Unlike the exudative lesion associated with streptococcal pharyngitis, the pseudomembrane in diphtheria is tightly adherent to the underlying tissues. Attempts to dislodge the membrane may cause bleeding. Hoarseness suggests laryngeal diphtheria, in which laryngoscopy may be diagnostically helpful.

Cutaneous Diphtheria

This dermatosis is characterized by punched-out ulcerative lesions with necrotic sloughing or pseudomembrane formation **(Figure 145-2)**. The diagnosis requires cultivation of *C. diphtheriae* from lesions, which most commonly occur on the lower and upper extremities, head, and trunk.

FIGURE 145-2

Cutaneous diphtheria due to nontoxigenic *C. diphtheriae* on the lower extremity. *(From the Centers for Disease Control and Prevention, Public Health Image Library [PHIL]. #1941.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Infections Due to Non-*diphtheriae Corynebacterium* Species and Nontoxigenic *C. diphtheriae*

Non-*diphtheriae* species of *Corynebacterium* and related genera (discussed below) as well as nontoxigenic strains of *C. diphtheriae* itself have been found in bloodstream and respiratory infections, often in individuals

with immunosuppression or chronic respiratory disease. These organisms can cause disease manifestations and should not necessarily be dismissed as colonizers.

Other Clinical Manifestations

C. diphtheriae causes rare cases of endocarditis and septic arthritis, most often in patients with preexisting risk factors, such as abnormal cardiac valves, injection drug use, or cirrhosis.

COMPLICATIONS

Airway obstruction poses a significant early risk in patients presenting with advanced diphtheria. Pseudomembranes may slough and obstruct the airway or may advance to the larynx or into the tracheobronchial tree. Children are particularly prone to obstruction because of their small airways.

Polyneuropathy and myocarditis are late toxic manifestations of diphtheria. During a diphtheria outbreak in the Kyrgyz Republic in 1999, myocarditis was found in 22% and neuropathy in 5% of 676 hospitalized patients. The mortality rate was 7% among patients with myocarditis as opposed to 2% among those without myocardial manifestations. The median time to death in hospitalized patients was 4.5 days. Myocarditis is typically associated with arrhythmias and dilated cardiomyopathy.

Polyneuropathy is seen 3–5 weeks after the onset of diphtheria and has a slow indolent course. However, patients may develop severe and prolonged neurologic abnormalities. The disorders typically occur in the mouth and neck, with lingual or facial numbness as well as dysphonia, dysphagia, and cranial nerve paresthesias. More ominous signs include weakness of respiratory and abdominal muscles and paresis of the extremities. Sensory manifestations and sensory ataxia also are observed. Cranial nerve dysfunction typically precedes disturbances of the trunk and extremities because of proximity to the site of infection. Autonomic dysfunction also is associated with polyneuropathy and can lead to hypotension. Polyneuropathy is typically reversible in patients who survive the acute phase.

Other complications of diphtheria include pneumonia, renal failure, encephalitis, cerebral infarction, pulmonary embolism, and serum sickness from antitoxin therapy.

DIAGNOSIS

The diagnosis of diphtheria is based on clinical signs and symptoms plus laboratory confirmation. Respiratory diphtheria should be considered in patients with sore throat, pharyngeal exudates, and fever. Other symptoms may include hoarseness, stridor, or palatal paralysis. The presence of a pseudomembrane should prompt strong consideration of diphtheria. Once a clinical diagnosis of diphtheria is made, diphtheria antitoxin should be obtained and administered as rapidly as possible.

Laboratory diagnosis of diphtheria is based either on cultivation of *C. diphtheriae* or toxigenic *C. ulcerans* from the site of infection or on the demonstration of local lesions with characteristic histopathology. *Corynebacterium pseudodiphtheriticum*, a nontoxigenic organism, is a common component of the normal

throat flora and does not pose a significant risk. Throat samples should be submitted to the laboratory for culture with the notation that diphtheria is being considered. This information should prompt cultivation on special selective medium and subsequent biochemical testing to differentiate *C. diphtheriae* from other nasopharyngeal commensal corynebacteria. All laboratory isolates of *C. diphtheriae,* including nontoxigenic strains, should be submitted to the CDC.

A diagnosis of cutaneous diphtheria requires laboratory confirmation since the lesions are not characteristic and are indistinguishable from other dermatoses. Diphtheritic ulcers occasionally—but not consistently—have a punched-out appearance (Fig. 145-2). Patients in whom cutaneous diphtheria is identified should have the nasopharynx cultured for *C. diphtheriae*. The laboratory medium for cutaneous diphtheria specimens is the same as that used for respiratory diphtheria: Löffler's or Tinsdale's selective medium in addition to nonselective medium such as blood agar. As has been mentioned, respiratory diphtheria remains a notifiable disease in the United States, whereas cutaneous diphtheria is not.

TREATMENT

TREATMENT

Diphtheria

DIPHTHERIA ANTITOXIN

Prompt administration of diphtheria antitoxin is critical in the management of respiratory diphtheria. Diphtheria antitoxin, a horse antiserum, is effective in reducing the extent of local disease as well as the risk of complications of myocarditis and neuropathy. Rapid institution of antitoxin therapy is also associated with a significant reduction in mortality risk. Because diphtheria antitoxin cannot neutralize cell-bound toxin, prompt initiation is important. This product, which is no longer commercially available in the United States, can be obtained from the CDC Emergency Operations Center at 770-488-7100 (website: *www.cdc.gov/diphtheria/dat.html*) after first contacting the state health department. The current protocol for the use of diphtheria antitoxin involves a test dose to rule out immediate hypersensitivity. Patients who demonstrate hypersensitivity require desensitization before a full therapeutic dose of antitoxin is administered.

Given that the world supply of equine anti–diphtheria toxin is limited, a human monoclonal antibody with the potential to provide a safer alternative to equine antitoxin therapy is being developed.

ANTIMICROBIAL THERAPY

Antibiotics are used in the management of diphtheria primarily to prevent transmission to susceptible contacts. Antibiotics also prevent further toxin production and reduce the severity of local infection. Recommended treatment options for patients with respiratory diphtheria are as follows:

Procaine penicillin G, 600,000 U IM q12h (for children: 12,500–25,000 U/kg IM q12h) until the patient can swallow comfortably; then oral penicillin V, 125–250 mg qid to complete a 14-day course

Erythromycin, 500 mg IV q6h (for children: 40–50 mg/kg per day IV in two or four divided doses) until the patient can swallow comfortably; then 500 mg PO qid to complete a 14-day course

A clinical study in Vietnam found that penicillin was associated with a more rapid resolution of fever and a lower rate of bacterial resistance than erythromycin; however, relapses were more common in the penicillin group. Erythromycin therapy targets protein synthesis and thus offers the presumed benefit of stopping toxin synthesis more quickly than a cell wall–active β-lactam agent. Alternative therapeutic agents for patients who are allergic to penicillin or cannot take erythromycin include rifampin and clindamycin. Other reasonable antibiotics are clarithromycin, azithromycin, linezolid, and vancomycin, although they have not been studied in comparison to the agents above.

Eradication of *C. diphtheriae* should be documented after antimicrobial therapy is complete. A repeat throat culture 2 weeks later is recommended. For patients in whom the organism is not eradicated after a 14-day course of erythromycin or penicillin, an additional 10-day course followed by repeat culture is recommended. Drug-resistant strains of *C. diphtheriae* exist, and several reports have described multidrug-resistant strains, predominantly in Southeast Asia. Drug resistance should be considered when efforts at pathogen eradication fail.

Cutaneous diphtheria should be treated as described above for respiratory disease. Individuals infected with toxigenic strains should receive antitoxin. It is important to treat the underlying cause of the dermatoses in addition to the superinfection with *C. diphtheriae*.

Patients who recover from respiratory or cutaneous diphtheria should have antitoxin levels measured. If diphtheria antitoxin has been administered, this test should be performed 6 months later. Patients who recover from respiratory or cutaneous diphtheria should receive the appropriate vaccine to ensure the development of protective antibody titers.

MANAGEMENT STRATEGIES

Patients in whom diphtheria is suspected should be hospitalized in respiratory isolation rooms, with close monitoring of cardiac and respiratory function. A cardiac workup is recommended to assess the possibility of myocarditis. In patients with extensive pseudomembranes, an anesthesiology or an ear, nose, and throat consultation is recommended because of the possible need for tracheostomy or intubation. In some settings, pseudomembranes can be removed surgically. Treatment with glucocorticoids has not been shown to reduce the risk of myocarditis or polyneuropathy.

PROGNOSIS

The mortality rate for diphtheria is 5–10% but may approach 20% among children <5 years old and adults >40 years of age. Fatal pseudomembranous diphtheria typically occurs in patients with nonprotective antibody titers and in unimmunized patients. The pseudomembrane may actually increase in size from the time it is first noted. Risk factors for death include bullneck diphtheria; myocarditis with ventricular tachycardia; atrial fibrillation; complete heart block; an age of >60 years or <6 months; alcoholism; extensive

pseudomembrane elongation; and laryngeal, tracheal, or bronchial involvement. Another important predictor of fatal outcome is the interval between the onset of local disease and the administration of antitoxin. Cutaneous diphtheria has a low mortality rate and is rarely associated with myocarditis or peripheral neuropathy.

PREVENTION

Vaccination

Sustained campaigns for vaccination of children and adequate boosting vaccination of adults are responsible for the exceedingly low incidence of diphtheria in most developed nations. Currently, diphtheria toxoid vaccine is coadministered with tetanus vaccine (with or without acellular pertussis). DTaP (full-level diphtheria toxoid, tetanus toxoid, and acellular pertussis vaccine) is currently recommended for children up to the age of 6; DTaP replaced the earlier whole-cell pertussis vaccine DTP in 1997. Tdap is a tetanus toxoid, reduced diphtheria toxoid, and acellular pertussis vaccine formulated for adolescents and adults. Tdap was licensed for use in the United States in 2005 and is recommended for children ≥7 years old and for adults. It is recommended that all adults (i.e., persons >19 years old) receive a single dose of Tdap if they have not received it previously, regardless of the interval since the last dose of Td (tetanus and reduced-dose diphtheria toxoids, adsorbed). Tdap vaccination is a priority for health care workers, pregnant women, adults anticipating contact with infants, and adults not previously vaccinated for pertussis. Adults who have received acellular pertussis vaccine should continue to receive decennial Td booster vaccinations. **The vaccine schedule is detailed in Chap. 118.**

Prophylaxis Administration to Contacts

Close contacts of diphtheria patients should undergo throat culture to determine whether they are carriers. After samples for throat culture are obtained, antimicrobial prophylaxis should be considered for all contacts, even those whose cultures are negative. The options are 7–10 days of oral erythromycin or one dose of IM benzathine penicillin G (1.2 million units for persons ≥6 years of age or 600,000 units for children <6 years of age).

Contacts of diphtheria patients whose immunization status is uncertain should receive the appropriate diphtheria toxoid–containing vaccine. The Tdap vaccine (rather than Td) is now the booster vaccine of choice for adults who have not recently received an acellular pertussis–-containing vaccine. Carriers of *C. diphtheriae* in the community should be treated and vaccinated when identified.

OTHER CORYNEBACTERIAL AND *RHODOCOCCUS* INFECTIONS

Nondiphtherial corynebacteria, referred to as *diphtheroids* or *coryneforms*, are frequently considered colonizers or contaminants; however, they have been associated with invasive disease, particularly in immunocompromised patients. These organisms have been isolated from the bloodstream, especially in association with catheter infection, endocarditis, prosthetic valve infection, meningitis, brain abscess,

osteomyelitis, and peritonitis. Risk factors include indwelling intravenous or peritoneal catheters and neurosurgical shunts. Patients infected with these organisms are often immunosuppressed or have significant medical comorbidities. The nondiphtherial coryneforms are a collection of bacteria that are taxonomically grouped together in the genus *Corynebacterium* on the basis of their 16S rDNA signature nucleotides. Despite the shared rDNA signatures, these isolates are quite diverse. For example, their guanine-cytosine content ranges from 45 to 70%. Several nondiphtheroid corynebacteria, including *Corynebacterium jeikeium* and *Corynebacterium urealyticum*, are associated with resistance to multiple antibiotics. *Rhodococcus equi* is associated with necrotizing pneumonia and granulomatous infection, particularly in immunocompromised individuals.

MICROBIOLOGY AND LABORATORY DIAGNOSIS

These organisms are non-acid-fast, catalase-positive, aerobic or facultatively anaerobic rods. Their colonial morphologies on blood agar vary widely; some species are small and α-hemolytic (similar to lactobacilli), whereas others form large white colonies (similar to yeasts). Many nondiphtherial coryneforms require special media, such as Löffler's, Tinsdale's, or tellurite medium. These cultivation idiosyncrasies have led to a complex taxonomic categorization of the organisms.

EPIDEMIOLOGY

Humans are the natural reservoirs for several nondiphtherial coryneforms, including *C. xerosis, C. pseudodiphtheriticum*, *C. striatum*, *C. minutissimum*, *C. jeikeium*, *C. urealyticum*, and *Arcanobacterium haemolyticum*. Animal reservoirs are responsible for carriage of *Arcanobacterium pyogenes*, *C. ulcerans*, and *C. pseudotuberculosis*. Soil is the natural reservoir for *R. equi*.

C. pseudodiphtheriticum is a component of the normal flora of the human pharynx and skin. *C. xerosis* is found on the skin, nasopharynx, and conjunctiva; *C. auris* in the external auditory canal; and *C. striatum* in the anterior nares and on the skin. *C. jeikeium* and *C. urealyticum* are found in the axilla, groin, and perineum, particularly in hospitalized patients. Infections with *C. ulcerans* and *C. pseudotuberculosis* have been associated with the consumption of raw milk from infected cattle.

C. ulcerans

This organism causes a diphtheria-like illness and produces both diphtheria toxin and a dermonecrotic toxin. The organism is a commensal in horses and cattle and has been isolated from cow's milk. In contrast to diphtheria, this infection is considered a zoonosis, and cases have been traced to contact with animal carriers, including dogs and pigs. *C. ulcerans* causes exudative pharyngitis, primarily during summer months, in rural areas, and among individuals exposed to animals. Treatment with antitoxin and antibiotics should be initiated when respiratory *C. ulcerans* is identified, and a contact investigation (including throat cultures to determine the need for antimicrobial prophylaxis and, in unimmunized contacts, administration of the appropriate diphtheria toxoid–containing vaccine) should be conducted. The organism grows on Löffler's, Tinsdale's, and telluride agars as well as blood agar. In addition to exudative pharyngitis, cutaneous disease

due to *C. ulcerans* has been reported. *C. ulcerans* is susceptible to a wide panel of antibiotics. Erythromycin and macrolides appear to be the first-line agents.

C. pseudotuberculosis (ovis)

Infection caused by *C. pseudotuberculosis* is rare and is reported predominantly from Australia. *C. pseudotuberculosis* causes suppurative granulomatous lymphadenitis and an eosinophilic pneumonia syndrome among individuals who handle sheep; horses, cattle, goats, deer, and raw milk have also been implicated. The organism is an important veterinary pathogen, causing suppurative lymphadenitis, abscesses, and pneumonia, but is rarely a human pathogen. Surgical excision of affected lymph nodes should be performed when feasible, and successful treatment with erythromycin or tetracycline has been reported. Some strains express diphtheria toxin and produce a diphtheria-like disease, which should be treated with antitoxin.

C. jeikeium (Group JK)

Originally described in American hospitals, *C. jeikeium* infection was subsequently reported in Europe. After a 1976 survey of diseases caused by nondiphtherial corynebacteria, CDC group JK emerged as an important opportunistic pathogen among neutropenic and HIV-infected patients. The organism has now been designated a separate species. *C. jeikeium* forms small, gray to white, glistening, nonhemolytic colonies on blood agar. It lacks urease and nitrate reductase and does not ferment most carbohydrates. The predominant syndrome associated with *C. jeikeium* is sepsis, sometimes with associated pneumonia, endocarditis, meningitis, osteomyelitis, or epidural abscess. Risk factors for *C. jeikeium* infection include hematologic malignancy, neutropenia from comorbid conditions, prolonged hospitalization, exposure to multiple antibiotics, and skin disruption. There is evidence that *C. jeikeium* is part of the inguinal, axillary, genital, and perirectal flora of hospitalized patients.

Broad-spectrum antimicrobial therapy appears to select for colonization. The organisms appear as gram-positive coccobacillary forms slightly resembling streptococci. *C. jeikeium* is resistant to the majority of antibiotic classes except oxazolidinones (e.g., linezolid) and glycopeptides (e.g., vancomycin). Effective therapy involves removal of the infectious source, whether a catheter, prosthetic joint, or prosthetic valve. Efforts have been made to prevent *C. jeikeium* infection with strict institution of infection control protocols for high-risk patients, particularly those in intensive care units.

C. urealyticum (Group D2)

Identified as a urease-positive nondiphtherial *Corynebacterium* in 1972, *C. urealyticum* is an opportunistic pathogen causing sepsis and urinary tract infection. *C. urealyticum* appears to be the etiologic agent of a severe urinary tract syndrome known as *alkaline-encrusted cystitis*, a chronic inflammatory bladder infection associated with deposition of ammonium magnesium phosphate on the surface and walls of ulcerating lesions in the bladder. In addition, *C. urealyticum* has been associated with pneumonia, peritonitis, endocarditis, osteomyelitis, and wound infection. It is similar to *C. jeikeium* in its resistance to

most antibiotics except oxazolidinones and glycopeptides. Vancomycin therapy has been used successfully in severe infections.

C. minutissimum (Erythrasma)

Erythrasma is a cutaneous infection producing reddish-brown, macular, scaly, pruritic intertriginous patches. The dermatologic presentation under the Wood's lamp is of coral red fluorescence. *C. minutissimum* appears to be a common cause of erythrasma, although there is evidence for a polymicrobial etiology in certain settings. This microbe has also been associated with bacteremia in patients with hematologic malignancy. Erythrasma responds to topical erythromycin, clarithromycin, clindamycin, or fusidic acid, although more severe infections may require oral macrolide therapy.

Other Nondiphtherial Corynebacterial

C. xerosis is a human commensal found in the conjunctiva, nasopharynx, and skin. This nontoxigenic organism is occasionally identified as a source of invasive infection in immunocompromised or postoperative patients and prosthetic joint recipients. *C. striatum* is found in the anterior nares, skin, face, and upper torso of healthy individuals. Also nontoxigenic, this organism has been associated with invasive opportunistic infections in severely ill or immunocompromised patients. *C. amycolatum* is isolated from human skin and is identified on the basis of a unique 16S ribosomal RNA sequence associated with opportunistic infection. *C. glucuronolyticum* is a nonlipophilic species that causes male genitourinary tract infections such as prostatitis and urethritis. These infections may be successfully treated with a wide variety of antibacterial agents, including β-lactams, rifampin, aminoglycosides, or vancomycin; however, the organism appears to be resistant to fluoroquinolones, macrolides, and tetracyclines. *C. imitans* has been identified in eastern Europe as a nontoxigenic cause of pharyngitis. *C. auris* has been identified in children with otitis media; it is susceptible to fluoroquinolones, rifampin, tetracycline, and vancomycin but resistant to penicillin G and variably susceptible to macrolides. *C. pseudodiphtheriticum* (*C. hoffmannii*) is a nontoxigenic species that is part of the normal human flora. Human infections—particularly endocarditis of either prosthetic or natural valves and invasive pneumonia—have been reported only rarely. Although *C. pseudodiphtheriticum* may be isolated from the nasopharynx of patients with suspected diphtheria, it is part of the normal flora and does not produce diphtheria toxin. *C. propinquum*, a close relative of *C. pseudodiphtheriticum*, is part of CDC group ANF-3 and has been isolated from the human respiratory tract and blood. *C. afermentans* subspecies *lipophilum* belongs to CDC group ANF-1 and has been isolated from human blood and abscesses. *C. accolens* has been isolated from wound drainage, throat swabs, and sputum and is typically identified as a satellite of staphylococcal organisms; this species has been associated with endocarditis. *C. bovis* is a veterinary commensal that has not been clearly associated with human disease. *C. aquaticum* is a water-dwelling organism that is occasionally isolated from patients using medical devices (e.g., for chronic ambulatory peritoneal dialysis or venous access).

Rhodococcus

Rhodococcus species are phylogenetically related to the corynebacteria. These gram-positive coccobacilli have been associated with tuberculosis-like infections in humans with granulomatous pathology. While *R. equi* is best known, other species have been identified, including *R.* (*Gordonia*) *bronchialis*, *R.* (*Tsukamurella*) *aurantiacus*, *R. luteus*, *R. erythropolis*, *R. rhodochrous*, and *R. rubropertinctus*.

R. equi has been recognized as a cause of pneumonia in horses since the 1920s and as a cause of related infections in cattle, sheep, and swine. It is found in soil as an environmental microbe. The organisms vary in length; appear as spherical to long, curved, clubbed rods; and produce large irregular mucoid colonies. *R. equi* cannot ferment carbohydrates or liquefy gelatin and is often acid fast. An intracellular pathogen of macrophages, *R. equi* can cause granulomatous necrosis and caseation. This organism has most commonly been identified in pulmonary infection, but infections of brain, bone, and skin also have been reported. Most commonly, *R. equi* disease manifests as nodular cavitary pneumonia of the upper lobe—a picture similar to that seen in tuberculosis or nocardiosis. Most patients are immunocompromised, often by HIV infection. Subcutaneous nodular lesions have also been identified. The involvement of *R. equi* should be considered when any patient presents with a tuberculosis-like syndrome.

Infection due to *R. equi* has been treated successfully with antibiotics that penetrate intracellularly, including macrolides, clindamycin, rifampin, and trimethoprim-sulfamethoxazole. β-Lactam antibiotics have not been useful. The organism is routinely susceptible to vancomycin, which is considered the drug of choice.

Other Related Species

Actinomyces pyogenes

This organism, a well-known pathogen of cattle, sheep, goats, and pigs, causes seasonal leg ulcers in rural Thailand. A few human cases of sepsis, endocarditis, septic arthritis, pneumonia, meningitis, and empyema have been reported. This species is susceptible to β-lactams, tetracyclines, aminoglycosides, and fluoroquinolones.

Arcanobacterium haemolyticum

A. haemolyticum was identified as an agent of wound infections in U.S. soldiers in the South Pacific during World War II. It appears to be a human commensal of the nasopharynx and skin, but has also been implicated in pharyngitis and chronic skin ulcers. In contrast to the much more common pharyngitis caused by *Streptococcus pyogenes*, *A. haemolyticum* pharyngitis is associated with a scarlatiniform rash on the trunk and proximal extremities in about half of cases; this illness is occasionally confused with toxic shock syndrome. Because *A. haemolyticum* pharyngitis primarily affects teenagers, it has been postulated that the rash–pharyngitis syndrome may represent co-pathogenicity, synergy, or opportunistic secondary infection with Epstein-Barr virus. *A. haemolyticum* has also been reported as a cause of bacteremia, soft tissue infections, osteomyelitis, and cavitary pneumonia, predominantly in the setting of underlying diabetes mellitus. The organism is susceptible to β-lactams, macrolides, fluoroquinolones, clindamycin, vancomycin, and doxycycline. Penicillin resistance has been reported.

FURTHER READING

Kadirova R et al: Clinical characteristics and management of 676 hospitalized diphtheria cases, Kyrgyz Republic, 1995. J Infect Dis 181(Suppl 1):S110, 2000. [PubMed: 10657201]

Kim R, Reboli AC: Other coryneform bacteria and *Rhodococcus,* in *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, 8th ed. JE Bennett et al (eds). Philadelphia, Elsevier Saunders, 2015, pp 2373–2382.

MacGregor RR: *Corynebacterium diphtheriae* (diphtheria), in *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, 8th ed. JE Bennett et al (eds). Philadelphia, Elsevier Saunders, 2015, pp 2366–2372.

Moore LS et al: *Corynebacterium ulcerans* cutaneous diphtheria. Lancet Infect Dis 15:1100, 2015.

Smith HL et al: Potency of a human monoclonal antibody to diphtheria toxin relative to equine diphtheria anti-toxin in a guinea pig intoxication model. Virulence 7:660, 2016.

Chapter 146: *Listeria monocytogenes* Infections

Elizabeth L. Hohmann; Daniel A. Portnoy

FIGURE 146-1

INTRODUCTION

Listeria monocytogenes is a food-borne pathogen that can cause serious infections, particularly in pregnant women and immunocompromised individuals. A ubiquitous saprophytic environmental bacterium, *L. monocytogenes* is also a facultative intracellular pathogen with a broad host range. Humans are probably accidental hosts for this microorganism. *L. monocytogenes* is of interest not only to clinicians but also to basic scientists as a model intracellular pathogen that is used to study basic mechanisms of microbial pathogenesis and host immunity.

MICROBIOLOGY

L. monocytogenes is a facultatively anaerobic, nonsporulating, gram-positive rod that grows over a broad temperature range, including refrigeration temperatures. This organism is motile during growth at low temperatures but much less so at 37°C. The vast majority of cases of human listerial disease can be traced to serotypes 1/2a, 1/2b, and 4. *L. monocytogenes* is weakly β-hemolytic on blood agar, and (as detailed below) its β-hemolysin is an essential determinant of its pathogenicity.

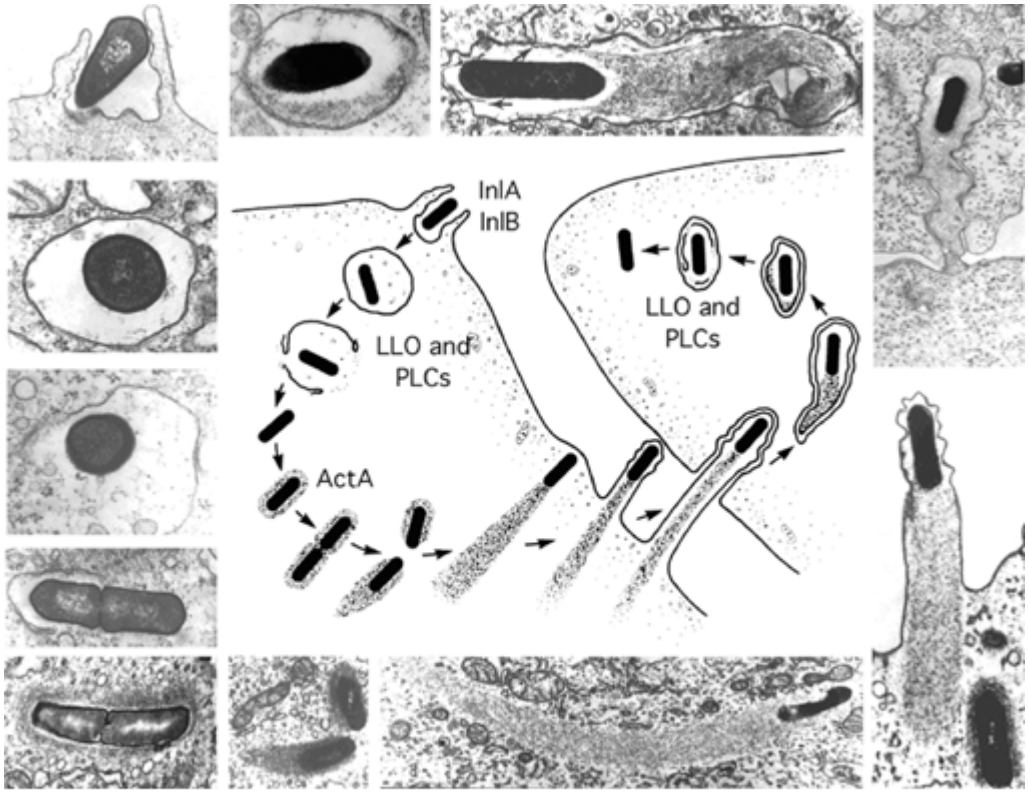
PATHOGENESIS

Infections with *L. monocytogenes* follow ingestion of contaminated food that contains the bacteria at high concentrations. The conversion from environmental saprophyte to pathogen involves the coordinate regulation of bacterial determinants of pathogenesis that mediate entry into cells, intracellular growth, and cell-to-cell spread. Many of the organism's pathogenic strategies can be examined experimentally in tissue culture models of infection **(Fig. 146-1)**. Like other enteric pathogens, *L. monocytogenes* induces its own internalization by cells that are not normally phagocytic. Its entry into cells is mediated by host surface proteins classified as internalins. Internalin-mediated entry is important in the crossing of intestinal, blood–brain, and fetoplacental barriers, although how *L. monocytogenes* traffics from the intestine to the brain or fetus is only beginning to be investigated. In a pregnant guinea pig model of infection, *L. monocytogenes* was shown to traffic from maternal organs to the placenta; surprisingly, however, it also trafficked from the

placenta back to maternal organs. These data are consistent with a model in which miscarriage can be viewed as a host defense strategy to eliminate a nidus of infection.

FIGURE 146-1

Stages in the intracellular life cycle of *Listeria monocytogenes.* The central diagram depicts cell entry, escape from a vacuole, actin nucleation, actin-based motility, and cell-to-cell spread. Surrounding the diagram are representative electron micrographs from which it was derived. ActA, surface protein mediating nucleation of host actin filaments to propel bacteria intra- and intercellularly; LLO, listeriolysin O; PLCs, phospholipases C; Inl, internalin. See text for further details. *(Adapted with permission from LG Tilney, DA Portnoy: J Cell Biol 109:1597, 1989. © Rockefeller University Press.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

An essential determinant of the pathogenesis of *L. monocytogenes* is its β-hemolysin, listeriolysin O (LLO). LLO is a pore-forming, cholesterol-dependent cytolysin. (Related cytolysins include streptolysin O, pneumolysin, and perfringolysin O, all of which are produced by extracellular pathogens.) LLO is largely responsible for mediating rupture of the phagosomal membrane that forms after phagocytosis of *L. monocytogenes*. LLO probably acts by insertion into an acidifying phagosome, which prevents the vesicle's maturation. In addition, LLO acts as a translocation pore for one or both of the *L. monocytogenes* phospholipases that also contribute to vacuolar lysis by blocking host cell autophagy. LLO synthesis and activity are controlled at multiple levels to ensure that its lytic activity is limited to acidic vacuoles and does not affect the cytosol. Mutations in LLO that influence its synthesis, cytosolic half-life, or pH optimum cause premature toxicity to infected cells. There is an inverse relationship between toxicity and virulence—i.e., the more cytotoxic the strain, the less virulent it is in animals. This relationship may seem paradoxical, but, as an intracellular pathogen, *L. monocytogenes* benefits from leaving its host cell unharmed.

Shortly after exposure to the mammalian-cell cytosol, *L. monocytogenes* expresses a surface protein, ActA, that mediates the nucleation of host actin filaments to propel the bacteria intra- and intercellularly. ActA mimics host proteins of the Wiskott-Aldrich syndrome protein (WASP) family by promoting the actin nucleation properties of the Arp2/3 complex. Thus, *L. monocytogenes* can enter the cytosol of almost any eukaryotic cell or cell extract and can exploit a conserved and essential actin-based motility system. Other pathogens as diverse as certain *Shigella*, *Mycobacterium*, *Rickettsia*, and *Burkholderia* species use a related pathogenic strategy that allows cell-to-cell spread without exposure to the extracellular milieu.

IMMUNE RESPONSE

The innate and acquired immune responses to *L. monocytogenes* have been studied extensively in mice. Shortly after IV injection, most bacteria are found in liver macrophages, with some organisms in splenic dendritic cells and macrophages. Listeriae that survive the bactericidal activity of initially infected macrophages grow in the cytosol and spread from cell to cell. *L. monocytogenes* triggers three innate immune pathways: a MyD88-dependent pathway that leads to inflammatory cytokine production; a STING/IRF3 pathway that is triggered by secreted bacterial cyclic di-adenosine monophosphate and leads to a type I interferon response; and low-level inflammasome activation that is triggered by DNA from infrequent bacteriolysis. Neutrophils are crucial to host defense during the first 24 h of infection, whereas an influx of activated macrophages from the bone marrow is critical subsequently. Mice that survive sublethal infection clear the organisms within a week, with consequent sterile immunity. Studies with knockout mice have been instrumental in dissecting the roles played by chemokines and cytokines during infection. For example, interferon γ, tumor necrosis factor, and CCR2 are essential in controlling infection. While innate immunity is sufficient to control infection, the acquired immune response is required for sterile immunity. Immunity is cell mediated; antibody plays no measurable role. The critical effector cells are cytotoxic (CD8+) T cells that recognize and lyse infected cells; the resulting extracellular bacteria are killed by circulating activated phagocytes.

A hallmark of the *L. monocytogenes* model is that killed vaccines do not provide protective immunity. The explanation for this fundamental observation is multifactorial, involving the generation of appropriate cytokines and the compartmentalization of bacterial proteins for antigen processing and presentation. Because the organism has the capacity to induce a robust cell-mediated immune response, attenuated strains have been engineered to express foreign antigens and are undergoing clinical studies as therapeutic vaccines for cancer.

EPIDEMIOLOGY

L. monocytogenes usually enters the body via the gastrointestinal tract in foods. Listeriosis is most often sporadic, although outbreaks do occur. No evidence supports person-to-person transmission (other than vertical transmission from mother to fetus) or waterborne infection. In line with its survival and multiplication at refrigeration temperatures, *L. monocytogenes* is commonly found in processed and unprocessed foods of animal and plant origin, especially soft cheeses, delicatessen meats, hot dogs, milk,

and cold salads; fresh fruits and vegetables also can transmit the organism. Because food supplies are increasingly centralized and normal hosts tolerate the organism well, outbreaks may not be immediately apparent.

DIAGNOSIS

Symptoms of listerial infection overlap greatly with those of other infectious diseases. Timely diagnosis requires that the illness be considered in groups at risk: pregnant women; elderly persons; neonates; individuals immunocompromised by organ transplantation, cancer, or treatment with tumor necrosis factor antagonists or glucocorticoids; and patients with a variety of chronic medical conditions, including alcoholism, diabetes, renal disease, and rheumatologic and hepatic illnesses. Meningitis in older adults (especially with parenchymal brain involvement or subcortical brain abscess) should trigger consideration of *L. monocytogenes* infection and treatment. Listeriosis occasionally affects healthy, young, nonpregnant individuals. HIV-infected patients are at risk; however, listeriosis seems to be prevented by trimethoprim-sulfamethoxazole (TMP-SMX) prophylaxis targeting other AIDS-related infections. The diagnosis is typically made by culture of blood, cerebrospinal fluid (CSF), or amniotic fluid. *L. monocytogenes* may be confused with “diphtheroids” or pneumococci in Gram-stained CSF or may be gram-variable and confused with *Haemophilus* species. Polymerase chain reaction diagnostics have been described but are not widely available, and serology is not clinically useful.

CLINICAL MANIFESTATIONS

Listerial infections present as several clinical syndromes, of which meningitis and septicemia are most common.

Gastroenteritis

Listerial gastroenteritis typically develops within 48 h of ingestion of a large inoculum of bacteria in contaminated foods. *L. monocytogenes* is neither sought nor found in routine fecal cultures, but its involvement should be considered in outbreaks when cultures for other likely pathogens are negative. Manifestations include fever, diarrhea, headache, and constitutional symptoms. The largest reported outbreak occurred in an Italian school system and included 1566 individuals; ~20% of patients were hospitalized, but only one person had a positive blood culture. Isolated gastrointestinal illness does not require antibiotic treatment. Surveillance studies show that 0.1–5% of healthy asymptomatic adults may have stool cultures positive for the organism.

Bacteremia

L. monocytogenes septicemia presents with fever, chills, and myalgias/arthralgias and cannot be differentiated from septicemia involving other organisms. Meningeal symptoms, focal neurologic findings, or mental status changes may suggest the diagnosis. Bacteremia is documented in 70–90% of cancer patients

with listeriosis. A nonspecific flulike illness with fever is a common presentation in pregnant women. A lumbar puncture is often prudent, although not necessary, in pregnant women without central nervous system (CNS) symptoms.

Meningitis

L. monocytogenes causes ~5–10% of all cases of community-acquired bacterial meningitis in adults in the United States. Case–fatality rates are reported to be 15–26% and do not appear to have changed over time. This diagnosis should be considered in all older or chronically ill adults with “aseptic” meningitis. The presentation is more frequently subacute (with illness developing over several days) than in meningitis of other bacterial etiologies, and nuchal rigidity and meningeal signs are less common. Photophobia is infrequent. Focal findings and seizures are common in some but not all series. The CSF profile in listerial meningitis most often shows white blood cell counts in the range of 100–5000/µL (rarely higher); 75% of patients have counts below 1000/µL, usually with a neutrophil predominance more modest than that in other bacterial meningitides. Low glucose levels and positive results on Gram’s staining are found ~30–40% of the time. Hydrocephalus can occur.

Meningoencephalitis and Focal CNS Infection

L. monocytogenes can directly invade the brain parenchyma, producing either cerebritis or focal abscess. Approximately 10% of cases of CNS infection are macroscopic abscesses resulting from bacteremic seeding; the affected patients often have positive blood cultures. Concurrent meningitis can exist, but the CSF may appear normal. Abscesses can be misdiagnosed as metastatic or primary tumors and, in rare instances, occur in the cerebellum and the spinal cord. Invasion of the brainstem results in a characteristic severe rhombencephalitis, usually in otherwise healthy older adults (although there are numerous other infectious and noninfectious causes of this syndrome). The presentation may be biphasic, with a prodrome of fever and headache followed by neurologic decline and focal findings. The subacute course and the often minimally abnormal CSF findings may delay the diagnosis, which may be suggested by MRI showing ring-enhancing lesions after gadolinium contrast. A pattern of multiple brain abscesses along white-matter fiber tracts may represent intra-axonal spread and suggest the diagnosis. MRI is superior to CT for the diagnosis of these infections.

Infection in Pregnant Women and Neonates

Listeriosis in pregnancy is a severe and important infection that can cause miscarriage and stillbirth. The usual presentation is a nonspecific acute or subacute febrile illness with myalgias, arthralgias, backache, and headache. Pregnant women with listeriosis are usually bacteremic. This syndrome should prompt blood cultures, especially if there is no other reasonable explanation. Involvement of the CNS is uncommon in pregnant women. Preterm delivery is a common complication, and the diagnosis may be made only postpartum. As many as 70–90% of fetuses from infected women can become infected. Prepartum treatment of bacteremic women enhances the chances of delivery of a healthy infant. Women usually do well after delivery: maternal deaths are very rare, even when the diagnosis is made late in pregnancy or postpartum.

Overall mortality rates for fetuses infected in utero approach 50% in some series; among live-born neonates treated with antibiotics, mortality rates are much lower (~20%). *Granulomatosis infantiseptica* is an overwhelming listerial fetal infection with miliary microabscesses and granulomas, most often in the skin, liver, and spleen. Less severe neonatal infection acquired in utero presents at birth. “Late-onset” neonatal illness typically develops ~10–30 days postpartum. Mothers of infants with late-onset disease are not ill. It is hypothesized that peripartum antibiotic prophylaxis for group B streptococcal infection may be decreasing the rates of neonatal listeriosis.

TREATMENT

TREATMENT

Infections Caused by *Listeria monocytogenes*

ANTIBIOTICS

No clinical trials have compared antimicrobial agents for the treatment of *L. monocytogenes* infections. Data from studies conducted in vitro and in animals as well as observational clinical data indicate that ampicillin is the drug of choice, although penicillin also is highly active. Adults should receive IV ampicillin at high doses (2 g every 6 h). Many experts recommend the addition of gentamicin for synergy (1.0–1.7 mg/kg every 8 h); retrospective uncontrolled trials are not conclusive, but one study suggests that gentamicin may not help. TMP-SMX, given IV, is the best alternative for the penicillin-allergic patient (15–20 mg of TMP/kg per day in divided doses every 6–8 h). The dosages recommended cover CNS infection and bacteremia (see below for duration); dosages must be reduced for patients with renal insufficiency. One small nonrandomized study supports a combination of ampicillin and TMP-SMX. Case reports document success with vancomycin, imipenem, meropenem, linezolid, tetracycline, and macrolides, although there are also reports of clinical failure or disease development with some of these agents. None of these agents has been demonstrated to be superior to ampicillin. Although not rigorously studied, adjunctive dexamethasone has not been shown to be advantageous in CNS infection. Acquired resistance to antimicrobial agents has been sought but not found in large strain collections. Cephalosporins are *not* effective and should not be used. Neonates should receive ampicillin and gentamicin at doses based on weight.

DURATION

The duration of therapy depends on the syndrome: 2 weeks for bacteremia, 3 weeks for meningitis, 6–8 weeks for brain abscess/encephalitis, and 4–6 weeks for endocarditis in both neonates and adults. Early-onset neonatal disease may be more severe and should be treated for >2 weeks.

COMPLICATIONS AND PROGNOSIS

Many individuals who are promptly diagnosed and treated recover fully, but permanent neurologic sequelae are common in patients with brain abscess or rhombencephalitis. Endocarditis and focal infections of visceral organs; the eye; the pleural, peritoneal, and pericardial spaces; the bones; and both native and

prosthetic joints and endovascular grafts have all been reported. Of 100 live-born, treated neonates in one series, 60% recovered fully, 24% died, and 13% had long-term neurologic or other complications.

PREVENTION

Healthy persons should take standard precautions to prevent food-borne illness: fully cooking meats, washing fresh vegetables, carefully cleaning utensils, and avoiding unpasteurized dairy products. In addition, persons at risk for listeriosis, including pregnant women, should avoid soft cheeses (hard cheeses and yogurt are not problematic) and should avoid or thoroughly reheat ready-to-eat and delicatessen foods. Entirely well pregnant women who report exposure to foods recalled for possible listerial contamination may be educated and followed expectantly.

FURTHER READING

Bakardjiev AI et al: Pathogens and the placental fortress. Curr Opin Microbiol 15:36, 2012.

Freitag NE et al: *Listeria monocytogenes*—from saprophyte to intracellular pathogen. Nat Rev Microbiol 9:623, 2009.

Lamont RF et al: Listeriosis in human pregnancy: A systematic review. J Perinat Med 39:227, 2011.

Le DT et al: Clinical development of *Listeria monocytogenes*–based immunotherapies. Semin Oncol 39:311, 2012.

Pamer ER: Immune responses to *Listeria monocytogenes*. Nat Rev Immunol 4:812, 2004.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 147: Tetanus

C. Louise Thwaites; Lam Minh Yen

FIGURE 147-1

INTRODUCTION

Tetanus is an acute disease manifested by skeletal muscle spasm and autonomic nervous system disturbance. It is caused by a powerful neurotoxin produced by the bacterium *Clostridium tetani* and is completely preventable by vaccination. *C. tetani* is found throughout the world, and tetanus commonly occurs where the vaccination coverage rate is low. In developed countries, the disease is seen occasionally in individuals who are incompletely vaccinated. In any setting, established tetanus is a severe disease with a high mortality rate.

DEFINITION

Tetanus is diagnosed on clinical grounds (sometimes with supportive laboratory confirmation of the presence of *C. tetani*; see "Diagnosis," below), and case definitions are often used to facilitate clinical and epidemiologic assessments. The Centers for Disease Control and Prevention (CDC) defines probable tetanus as "an acute illness with muscle spasms or hypertonia in the absence of a more likely diagnosis." *Neonatal* tetanus is defined by the World Health Organization (WHO) as "an illness occurring in a child who has the normal ability to suck and cry in the first 2 days of life but who loses this ability between days 3 and 28 of life and becomes rigid and has spasms." Given the unique presentation of neonatal tetanus, the history generally permits accurate classification of the illness with a high degree of probability. *Maternal* tetanus is defined by the WHO as tetanus occurring during pregnancy or within 6 weeks after the conclusion of pregnancy (whether with birth, miscarriage, or abortion).

ETIOLOGY

C. tetani is an anaerobic, gram-positive, spore-forming rod whose spores are highly resilient and can survive readily in the environment throughout the world. Spores resist boiling and many disinfectants. In addition, *C. tetani* spores and bacilli survive in the intestinal systems of many animals, and fecal carriage is common. The spores or bacteria enter the body through abrasions, wounds, or (in the case of neonates) the umbilical stump. Once in a suitable anaerobic environment, the organisms grow, multiply, and release tetanus toxin, an exotoxin that enters the nervous system and causes disease. Very low concentrations of this highly potent toxin can result in tetanus (minimal lethal human dose, 2.5 ng/kg).

In 20–30% of cases of tetanus, no puncture entry wound is found. Superficial abrasions to the limbs are the most common infection sites in adults. Deeper infections (e.g., attributable to open fracture, abortion, or drug injection) are associated with more severe disease and worse outcomes. In neonates, infection of the umbilical stump can result from inadequate umbilical-cord care; in some cultures, for example, the cord is cut with grass or animal dung is applied to the stump. Circumcision or ear-piercing also can result in neonatal tetanus.

EPIDEMIOLOGY

Tetanus is a rare disease in the developed world. Two cases of neonatal tetanus have occurred in the United States since 1989. In 2013, 26 cases of tetanus were reported to the U.S. national surveillance system. Most cases occur in incompletely vaccinated or unvaccinated individuals. Vaccination status is known in 50% of cases reported in the United States between 1972 and 2009; among these cases, only 16% of patients had had three or more doses of tetanus toxoid.

Persons >60 years of age are at greater risk of tetanus because antibody levels decrease over time. One-third of recent cases in the United States were in persons >65 years old. People who inject drugs—particularly those injecting heroin subcutaneously ("skin-popping")—are increasingly recognized as a high-risk group (15% of all cases in 2001–2008). In 2004, an outbreak of tetanus occurred in the United Kingdom, which had

previously reported low rates among drug users. The reasons for this outbreak remain unclear but are thought to involve a combination of heroin contamination, skin-popping, and incomplete vaccination. Since then, only sporadic cases have been reported in the United Kingdom.

The global incidence of tetanus among older children and adults is unknown, as few countries have good surveillance systems.

PATHOGENESIS

Genome sequencing of *C. tetani* has allowed identification of several exotoxins and virulence factors. Only those bacteria producing tetanus toxin (tetanospasmin) can cause tetanus. Although closely related to the botulinum toxins in structure and mode of action, tetanus toxin undergoes retrograde transport into the central nervous system (CNS) and thus produces clinical effects different from those caused by the botulinum toxins, which remain at the neuromuscular junction.

Tetanus toxin is intra-axonally transported to motor nuclei of the cranial nerves or ventral horns of the spinal cord. This toxin is produced as a single 150-kDa protein that is cleaved to produce heavy (100-kDa) and light (50-kDa) chains linked by a disulfide bond and noncovalent forces. The carboxy terminal of the heavy chain binds to specific membrane components in presynaptic α-motor nerve terminals; evidence suggests binding to both polysialogangliosides and membrane proteins. This binding results in toxin internalization and uptake into the nerves. Once inside the neuron, the toxin enters a retrograde transport pathway, whereby it is carried proximally to the motor neuron body in what appears to be a highly specific process. Unlike other components of the endosomal contents, which undergo acidification following internalization, tetanus toxin is transported in a carefully regulated pH-neutral environment that prevents an acid-induced conformational change that would result in light-chain expulsion into the surrounding cytosol.

The next stage in toxin trafficking is less clearly understood but involves tetanus toxin's escaping normal lysosomal degradation processes and undergoing translocation across the synapse to the GABA-ergic presynaptic inhibitory interneuron terminals. Here the light chain, which is a zinc-dependent endopeptidase, cleaves vesicle-associated membrane protein 2 (VAMP2, also known as *synaptobrevin*). This molecule is necessary for presynaptic binding and release of neurotransmitter; thus tetanus toxin prevents transmitter release and effectively blocks inhibitory interneuron discharge. The result is unregulated activity in the motor nervous system. Similar activity in the autonomic system accounts for the characteristic features of skeletal muscle spasm and autonomic system disturbance. The increased circulating catecholamine levels in severe tetanus are associated with cardiovascular complications.

Relatively little is known about the processes of recovery from tetanus. Recovery can take several weeks. Peripheral nerve sprouting is involved in recovery from botulism, and similar CNS sprouting may occur in tetanus. Other evidence suggests toxin degradation as a mechanism of recovery.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Tetanus

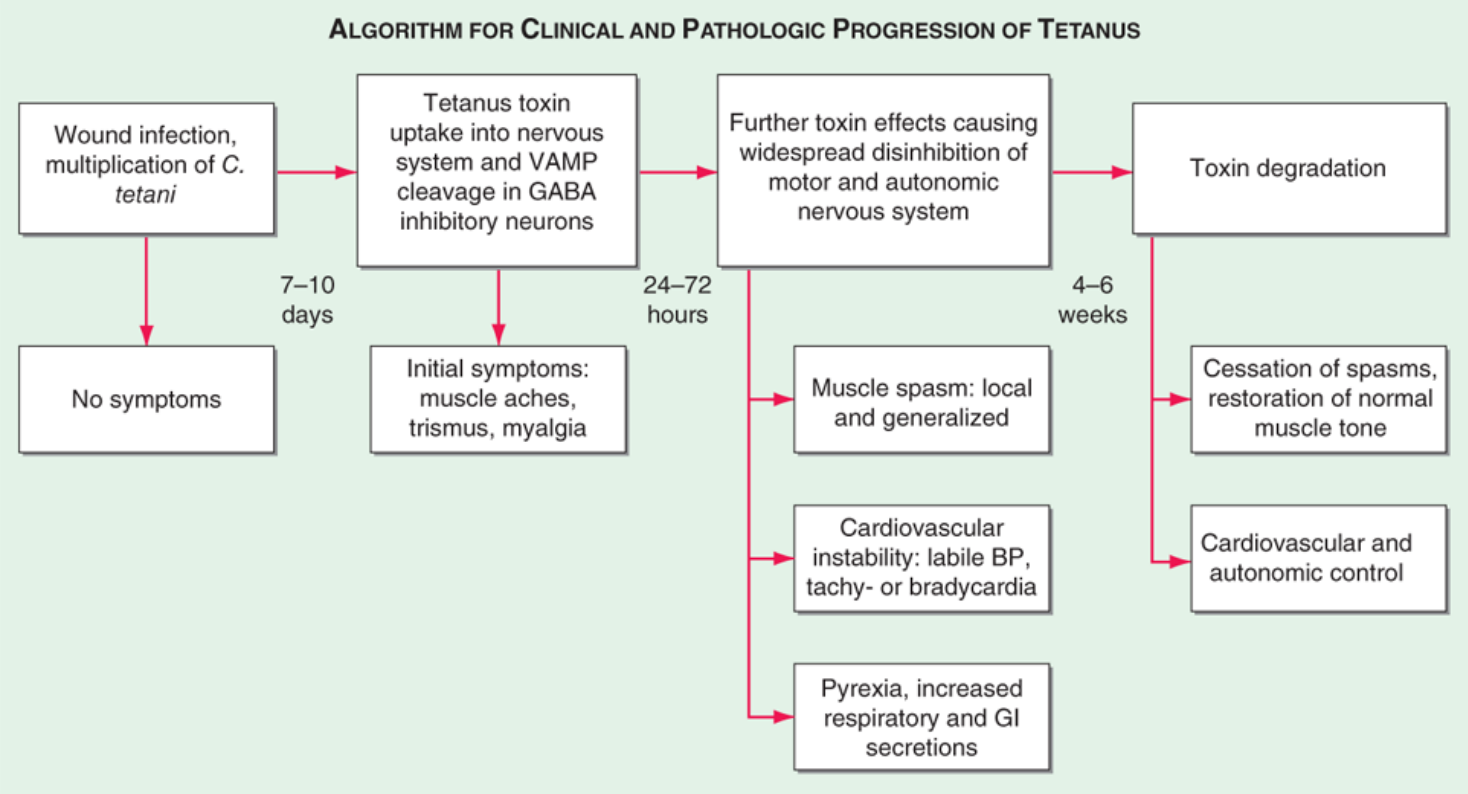
The clinical manifestations of tetanus occur only after tetanus toxin has reached presynaptic inhibitory nerves. Once these effects become apparent, there may be little that can be done to affect disease progression. Treatment should not be delayed while the results of laboratory tests are awaited. Management strategies aim to neutralize remaining unbound toxin and support vital functions until the effects of the toxin have worn off. Recent interest has focused on intrathecal methods of antitoxin administration to neutralize toxin within the CNS and limit disease progression (see "Treatment," below).

CLINICAL MANIFESTATIONS

Tetanus produces a wide spectrum of clinical features that are broadly divided into generalized (including neonatal) and local. In the usually mild form of local tetanus, only isolated areas of the body are affected and only small areas of local muscle spasm may be apparent. If the cranial nerves are involved in localized cephalic tetanus, the pharyngeal or laryngeal muscles may spasm, with consequent aspiration or airway obstruction, and the prognosis may be poor. In the typical progression of generalized tetanus **(Fig. 147-1)**, muscles of the face and jaw often are affected first, presumably because of the shorter distances toxin must travel up motor nerves to reach presynaptic terminals. Neonates typically present with an inability to suck.

FIGURE 147-1

Clinical and pathologic progression of tetanus. BP, blood pressure; GABA, γ-aminobutyric acid; GI, gastrointestinal; VAMP, vesicle-associated membrane protein (synaptobrevin).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In assessing prognosis, the speed at which tetanus develops is important. The incubation period (time from wound to first symptom) and the period of onset (time from first symptom to first generalized spasm) are of particular significance; shorter times are associated with worse outcome. In neonatal tetanus, the younger the infant is when symptoms occur, the worse the prognosis.

The most common initial symptoms are trismus (lockjaw), muscle pain and stiffness, back pain, and difficulty swallowing. In neonates, difficulty in feeding is the usual presentation. As the disease progresses, muscle spasm develops. Generalized muscle spasm can be very painful. Commonly, the laryngeal muscles are involved early or even in isolation. This is a life-threatening event as complete airway obstruction may ensue. Spasm of the respiratory muscles results in respiratory failure. Without ventilatory support, respiratory failure is the most common cause of death in tetanus. Spasms strong enough to produce tendon avulsions and crush fractures have been reported, but this outcome is rare.

Autonomic disturbance is maximal during the second week of severe tetanus, and death due to cardiovascular events becomes the major risk. Blood pressure is usually labile, with rapid fluctuations from high to low accompanied by tachycardia. Episodes of bradycardia and heart block also can occur. Autonomic involvement is evidenced by gastrointestinal stasis, sweating, increased tracheal secretions, and acute (often high-output) renal failure.

DIAGNOSIS

The diagnosis of tetanus is based on clinical findings. As stated above, treatment should not be delayed while laboratory tests are conducted. Culture of *C. tetani* from a wound provides supportive evidence. Serum anti-tetanus immunoglobulin G also may be measured in a sample taken before the administration of antitoxin or immunoglobulin; levels >0.1 IU/mL (measured by standard ELISA) are deemed protective and do not support the diagnosis of tetanus. If levels are below this threshold, a bioassay for serum tetanus toxin may be helpful, but a negative result does not exclude the diagnosis and these levels are not generally performed. Polymerase chain reaction also has been used for detection of tetanus toxin, but its sensitivity is unknown.

The few conditions that mimic generalized tetanus include strychnine poisoning and dystonic reactions to antidopaminergic drugs. Abdominal muscle rigidity is characteristically continuous in tetanus but is episodic in the latter two conditions. Cephalic tetanus can be confused with

trismus of other etiologies, such as oropharyngeal infection. Hypocalcemia and meningoencephalitis are included in the differential diagnosis of neonatal tetanus.

TREATMENT

TREATMENT

Tetanus

If possible, the entry wound should be identified, cleaned, and debrided of necrotic material in order to remove anaerobic foci of infection and prevent further toxin production. Metronidazole (400 mg rectally or 500 mg IV every 6 h for 7 days) is preferred for antibiotic therapy. An alternative is penicillin (100,000–200,000 IU/kg per day), although this drug theoretically may exacerbate spasms and in one study was associated with increased mortality. Failure to remove pockets of ongoing infection may result in recurrent or prolonged tetanus.

Antitoxin should be given early in an attempt to deactivate any circulating tetanus toxin and prevent its uptake into the nervous system. Two preparations are available: human tetanus immune globulin (TIG) and equine antitoxin. TIG is the preparation of choice, as it is less likely to be associated with anaphylactoid reactions. A single IM dose (3000–5000 IU) is given, with a portion injected around the wound. Equine-derived antitoxin is available widely and is used in low-income countries; after hypersensitivity testing, 10,000–20,000 U is administered IM as a single dose or as divided doses. Some evidence indicates that intrathecal administration of TIG inhibits disease progression and leads to a better outcome. The results of relevant studies have been supported by a meta-analysis of trials involving both adults and neonates, with TIG doses of 50–1500 IU administered intrathecally.

Spasms are controlled by heavy sedation with benzodiazepines. Chlorpromazine and phenobarbital are commonly used worldwide, and IV magnesium sulfate has been used as a muscle relaxant. A significant problem with all these treatments is that the doses necessary to control spasms also cause respiratory depression; thus, in resource-limited settings without mechanical ventilators, controlling spasms while maintaining adequate ventilation is problematic, and respiratory failure is a common cause of death. In locations with ventilation equipment, severe spasms are best controlled with a combination of sedatives or magnesium and relatively short-acting, cardiovascularly inert, nondepolarizing neuromuscular blocking agents that allow titration against spasm intensity. Infusions of propofol also have been used successfully to control spasms and provide sedation.

It is important to establish a secure airway early in severe tetanus. Ideally, patients should be nursed in calm, quiet environments because light and noise can trigger spasms. Tracheal secretions are increased in tetanus, and dysphagia due to pharyngeal involvement combined with hyperactivity of laryngeal muscles makes endotracheal intubation difficult. Patients may need ventilator support for several weeks. Thus tracheostomy is the usual method of securing the airway in severe tetanus.

Cardiovascular instability in severe tetanus is notoriously difficult to treat. Rapid fluctuations in blood pressure and heart rate can occur. Cardiovascular stability is improved by increasing sedation with IV magnesium sulfate (plasma concentration, 2–4 mmol/L or titrated against disappearance of the patella reflex), morphine, fentanyl, or other sedatives. In addition, drugs acting specifically on the cardiovascular system (e.g., esmolol, calcium antagonists, and inotropes) may be required. Short-acting drugs that allow rapid titration are preferred; particular care should be taken when longer-acting β antagonists are administered, as their use has been associated with hypotensive cardiac arrest.

Complications arising from treatment are common and include thrombophlebitis associated with diazepam injection, ventilator-associated pneumonia, central-line infections, and septicemia. In some centers, prophylaxis against deep-vein thrombosis and thromboembolism is routine.

Recovery from tetanus may take 4–6 weeks. Patients must be given a full primary course of immunization, as tetanus toxin is poorly immunogenic and the immune response following natural infection is inadequate.

PROGNOSIS

Rapid development of tetanus is associated with more severe disease and poorer outcome; it is important to note time of onset and length of incubation period. More sophisticated modeling has revealed other important predictors of prognosis **(Table 147-1)**. Few studies have formally addressed long-term outcomes of tetanus. However, it is generally accepted that recovery is typically complete unless periods of hypoventilation have been prolonged or other complications have ensued. Studies of children and neonates have suggested a higher incidence of neurologic sequelae. Neonates may be at increased risk of learning disabilities, behavioral problems, cerebral palsy, and deafness.

TABLE 147-1

Factors Associated with a Poor Prognosis in Tetanus

Adult Tetanus	Neonatal Tetanus
Age >70 years Incubation period <7 days Short time from first symptom to admission Puerperal, IV, postsurgery, burn entry site Period of onset[a] <48 h Heart rate >140 beats/min[b] Systolic blood pressure >140 mmHg[b] Severe disease or spasms[b] Temperature >38.5°C[b]	Younger age, premature birth Incubation period <6 days Delay in hospital admission Grass used to cut cord Low birth weight Fever on admission

[a]Time from first symptom to first generalized spasm. [b]At hospital admission.

PREVENTION

Tetanus is prevented by good wound care and immunization **(Chap. 118)**. In neonates, use of safe, clean delivery and cord-care practices as well as maternal vaccination are essential. The WHO guidelines for tetanus vaccination consist of a primary course of three doses in infancy, boosters at 4–7 and 12–15 years of age, and one booster in adulthood. In the United States, the CDC suggests an additional dose at 15–18 months with boosters every 10 years. “Catch-up” schedules recommend a three-dose primary course with 4 weeks between doses, followed by two boosters 6 months apart. For persons who have received a complete primary course in childhood but no further boosters, two doses at least 4 weeks apart are recommended.

Standard WHO recommendations for prevention of maternal and neonatal tetanus call for administration of two doses of tetanus toxoid at least 4 weeks apart to previously unimmunized pregnant women. However, in high-risk areas, a more intensive approach has been successful, with all women of childbearing age receiving a primary course along with education on safe delivery and postnatal practices.

Individuals sustaining tetanus-prone wounds should be immunized if their vaccination status is incomplete or unknown or if their last booster was given >10 years earlier. Patients with an inadequate vaccine status who sustain wounds not classified as clean or minor should also undergo passive immunization with TIG. It is recommended that tetanus toxoid be given in conjunction with diphtheria toxoid in a preparation with or without acellular pertussis: DTaP for children <7 years old, Td for 7- to 9-year-olds, and Tdap for children >9 years old and adults.

In the early 1980s, tetanus caused more than 1 million deaths a year, accounting for an estimated 5% of maternal deaths and 14% of all neonatal deaths. In 1989, the World Health Assembly adopted a resolution to eliminate neonatal tetanus by the year 2000; elimination was defined as <1 case/1000 live births in every district in every country. By 1999, elimination was still to be achieved in 57 countries and the deadline was extended until 2005, with the additional target of eliminating maternal tetanus (tetanus occurring during pregnancy or within 6 weeks of its end). Ratification of the Millennium Development Goals, in particular goal 4 (achieving a two-thirds reduction in the mortality rate among children under 5 by 2015), has further focused attention on reducing deaths from vaccine-preventable disease, particularly in the first 4 weeks of life.

Because vaccination reduces the incidence of neonatal tetanus by an estimated 94%, immunization of pregnant women with two doses of tetanus toxoid at least 4 weeks apart has been the primary method of maternal and neonatal tetanus elimination. In some areas, all women of childbearing age have been targeted as a means of increasing vaccination coverage. In addition, educational programs have focused on improving hygiene during the birth process, an intervention that in itself is estimated to reduce neonatal tetanus deaths by up to 40%.

The latest available data show that significant progress has been made: in recent years, 40 countries have achieved maternal and neonatal tetanus elimination, including China, India, and Indonesia. Worldwide, deaths from neonatal tetanus fell by 94% between 1990 and 2014; in the latter year, with 82% of newborns protected from the disease by maternal vaccination, there were an estimated 49,000 neonatal tetanus deaths,

mainly in Africa and Southeast Asia. Despite this relative success, immunization programs need to be ongoing as there is no herd immunity effect for tetanus and *C. tetani* contamination of soil and feces is widespread.

The rate of primary vaccination coverage in infancy (three doses of DTP) is 84%, but rates for the subsequent boosters necessary for long-term protection are unknown. Dedicated public health initiatives are lacking, and the continuing reports of sizable case series in the medical literature suggest that tetanus continues to pose a significant global health burden.

FURTHER READING

Borrow R et al: The immunological basis for immunization series. Module 3: Tetanus update 2006. Edited by Vaccines and Biologicals Immunization. World Health Organization, 2006.

Patel JC, Mehta BC: Tetanus: Study of 8,697 cases. Indian J Med Sci 53:393, 1999.

Rodrigo C et al: Pharmacological management of tetanus: An evidence-based review. Crit Care 18:217, 2014.

Thwaites CL et al: Predicting the clinical outcome of tetanus: The Tetanus Severity Score. Trop Med Int Health 11:279, 2006.

Thwaites CL et al: Maternal and neonatal tetanus. Lancet 385:362, 2014.

WEBSITES

Centers for Disease Control and Prevention: Pink Book. Tetanus. 1997. *www.cdc.gov/vaccines/pubs/pinkbook/downloads/tetanus.pdf*.

Health Protection Agency: Tetanus: Information for health professionals. 2013. *www.gov.uk/government/publications/tetanus-advice-for-health-professionals*.

World Health Organization: Maternal and neonatal tetanus (MNT) elimination. *www.who.int/immunization/diseases/MNTE_initiative/en/*.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

INTRODUCTION

Botulism, recognized at least since the eighteenth century, is a neuroparalytic disease caused by botulinum toxin, one of the most toxic substances known. While initially thought to be caused only by the ingestion of botulinum toxin in contaminated food (food-borne botulism), three additional forms caused by in situ toxin production after germination of spores in either a wound or the intestine are now recognized worldwide: wound botulism, infant botulism, and adult intestinal-colonization botulism. In addition to occurring in these recognized natural forms of the disease, botulism symptoms have been reported in patients receiving higher than recommended doses of botulinum toxin (iatrogenic botulism) for therapeutic or cosmetic purposes. Moreover, botulism was once reported after possible inhalation of botulinum toxin in a laboratory setting. Botulism manifests as a clinical syndrome of bilateral cranial-nerve palsies that may progress to respiratory compromise, a descending bilateral flaccid paralysis of voluntary muscles, and even death. Patients exposed to the same contaminated food may have a varying constellation of cranial nerve palsies and varying illness severity. Cases are known to be misdiagnosed or suspected late in a hospital course. The mainstays of therapy are meticulous intensive care and treatment with antitoxin early in the clinical course while alternative diagnoses are still being worked up. Early suspicion of botulism and empirical treatment are critical to favorable clinical outcomes.

ETIOLOGY AND PATHOGENESIS

Seven serologically distinct serotypes of botulinum toxin (A through G) have been confirmed. Botulinum toxin is produced by four recognized species of clostridia: *Clostridium botulinum* and rare strains of *Clostridium argentinense, Clostridium baratii*, and *Clostridium butyricum*. All these species are anaerobic gram-positive spore-forming organisms. The spores survive environmental conditions and ordinary cooking procedures. Toxin production, however, requires a rare confluence of product storage conditions: an anaerobic environment, a pH of >4.5, low salt and sugar concentrations, and temperatures of >3°C. Although commonly ingested, spores do not normally germinate and produce toxin in the adult human intestine.

Food-borne botulism is caused by consumption of foods contaminated with botulinum toxin; no confirmed host-specific factors are involved in the disease. *Wound botulism* is caused by toxin produced from germinating *C. botulinum* spores that contaminate an abscess or a wound. *Infant botulism* is caused by toxin

produced in situ by toxigenic clostridia colonizing the intestine of children <1 year of age. Colonization is thought to occur because the normal bowel microbiota is not yet fully established; this theory is supported by studies in animals. *Adult intestinal-colonization botulism*, a rare form that is poorly understood, is believed to have a pathology similar to that of infant botulism but occurs in adults; typically, patients have some anatomic or functional bowel abnormality or have recently used antibiotics that may help toxigenic clostridia compete more successfully against the normal bowel microbiota. Despite antitoxin treatment, relapse due to intermittent intraluminal production of toxin may be observed in patients with adult intestinal-colonization botulism. Relapse is a theoretical concern in wound botulism but has never been reported.

Regardless of how exposure occurs, botulinum neurotoxin enters the vascular system and is transported to peripheral cholinergic nerve terminals, including neuromuscular junctions, postganglionic parasympathetic nerve endings, and peripheral ganglia. Botulinum toxin is a zinc-endopeptidase protein of ~150 kDa, consisting of a 100-kDa heavy chain and a 50-kDa light chain. Steps in neurotoxin activity include (1) heavy-chain binding to nerve terminals, (2) internalization in endocytic vesicles, (3) translocation of the light chain to cytosol, and (4) light-chain serotype-specific cleavage of one of several proteins involved in the release of the neurotransmitter acetylcholine. Inhibition of acetylcholine release by any of the seven toxin serotypes results in characteristic flaccid paralysis.

All botulinum toxin serotypes have been demonstrated to cause botulism in nonhuman primates. Human cases associated with serotypes A, B, E, and F are reported each year. Serotype A produces the most severe syndrome, with the greatest proportion of patients requiring mechanical ventilation. Serotype B causes milder disease. Serotype E, most often associated with foods of aquatic origin (e.g., uneviscerated fish and seal oil), produces a syndrome of variable severity. Illnesses caused by toxin serotype F are infrequent and are characterized by rapid progression to quadriplegia and respiratory failure but also by relatively rapid recovery. Studies have shown that at least some serotypes can be differentiated into subtypes through neurotoxin gene sequencing; however, the impact of these subtype differences on clinical illness is not yet known.

EPIDEMIOLOGY

 Botulism occurs worldwide, but the number of cases reported varies among countries and regions. The variation may be due not only to actual differences in incidence but also to underreporting as a result of (1) limited awareness of the clinical presentation on which diagnosis depends; (2) lack of local specialized laboratory facilities to confirm botulism; (3) differences in requirements for disease reporting by physicians to public health agencies; and (4) lack of formal surveillance for botulism within a country. There is no universal surveillance system to capture worldwide botulism incidence. However, 30 countries currently participate in voluntary reporting of botulism cases to the European Union through an established surveillance system that includes standardized case definitions similar to those used in the United States and Canada. Some countries (e.g., the United States, Canada, Argentina, and Japan) maintain independent botulism surveillance, and some countries (e.g., Ethiopia) have no botulism surveillance.

Food-Borne Botulism

From 1899 to 2014, 1246 food-borne botulism events (single cases or outbreaks) were reported in the United States; from 1990 to 2014, a median of 20 cases were reported annually. Most such events (~80%) involve vegetables or fish/aquatic animals, usually home-preserved (canned, jarred). Native communities in both the United States (Alaska) and Canada have a high incidence of food-borne botulism due to traditional food-preparation practices; 85% of all cases in Canada occur in Native communities. Outbreaks typically involve three or four cases; however, one restaurant-associated outbreak in 1977 affected 59 persons. Worldwide, the highest incidence rate is reported from Georgia and Armenia in the southern Caucasus region, where illness is also associated with home-canning practices. Outbreaks in Asia are attributable to consumption of home-preserved fish or vegetable products such as bean curd and bamboo shoots. In parts of Europe, including Poland, France, and Germany, illness is often associated with home-preserved meat such as ham or sausage. Since 1950, commercial products have rarely been implicated in botulism in the United States and, when they have been implicated, cases have most often been attributed to consumer error in storage or cooking. However, manufacturer deficiencies do occur. In 2007, botulism developed in eight persons in the United States who consumed a widely distributed commercially canned hot-dog chili sauce. Significant deficiencies discovered by regulatory authorities involved 91 different products and resulted in the recall of 111 million cans of food. Lack of barriers to toxin production in commercially produced, non-canned products has also been documented. In 2006, commercially produced and internationally distributed carrot juice was implicated in six cases of botulism in the United States and Canada; inappropriate refrigeration in the setting of a lack of barriers to botulinum toxin production was believed to be the cause.

Wound Botulism

This form of disease was first recognized in 1951 as a result of a review of the clinical records on an accidental injury in 1943. Between 1943 and 2011, 491 cases of wound botulism were reported in the United States; 97% of cases reported after 1990 were associated with injection drug use. In the early 2000s, wound botulism associated with injection drug use emerged in Europe, and case clusters have been reported. The typical patient in the United States is a 30- to 50-year-old resident of the western United States with a long history of drug injection—specifically, of black-tar heroin. Some cases of wound botulism are due to contamination of traumatic wounds with *C. botulinum* spores. This category of botulism has been reported after injuries sustained during motor vehicle accidents.

Infant Botulism

More than 4400 infant botulism cases have been reported worldwide (80% of them in the United States) since this form of the disease was first recognized in 1976; ~80–100 cases (commonly caused by serotypes A and B) are reported annually in the United States.

Adult Intestinal-Colonization Botulism

This form of botulism is difficult to confirm because it is poorly understood. No clear criteria are available to differentiate cases of adult intestinal-colonization botulism from other adult botulism cases. Often these cases are caused by *C. baratii* type F, but the involvement of both *C. botulinum* type A and *C. butyricum* type E have been reported. Botulism following abdominal surgery or antibiotic use has sometimes been considered to be adult intestinal-colonization botulism.

Iatrogenic Botulism

Paralysis of variable severity has followed injection of high doses of licensed botulinum toxin products for treatment of refractory conditions involving hypertonicity of large muscle groups. Although some patients had symptoms consistent with botulism, no cases were laboratory-confirmed. Injection of approved doses of licensed products for cosmetic purposes has not been associated with botulism. However, four cases of laboratory-confirmed botulism in the United States resulted from illegal injection of a highly concentrated preparation of botulinum toxin that was not intended for clinical use; this preparation was privately administered for cosmetic purposes in 2004.

Inhalational Botulism

Inhalational botulism does not occur naturally. One report from Germany described botulism resulting from possible inhalational exposure to botulinum toxin in a laboratory incident.

Intentional Botulism

Botulinum toxin has been “weaponized” by governments and terrorist organizations. An attack might entail aerosolization of toxin or contamination of foods or beverages ranging in scope from small-scale tampering to contamination of a widely distributed food item. An unnatural event may be suggested by a very large number of ill persons, an unusual relationship between patients (e.g., a visit to the same building), exposure to foods not typically associated with botulism, or the involvement of atypical toxin serotypes (i.e., those not usually associated with human illness).

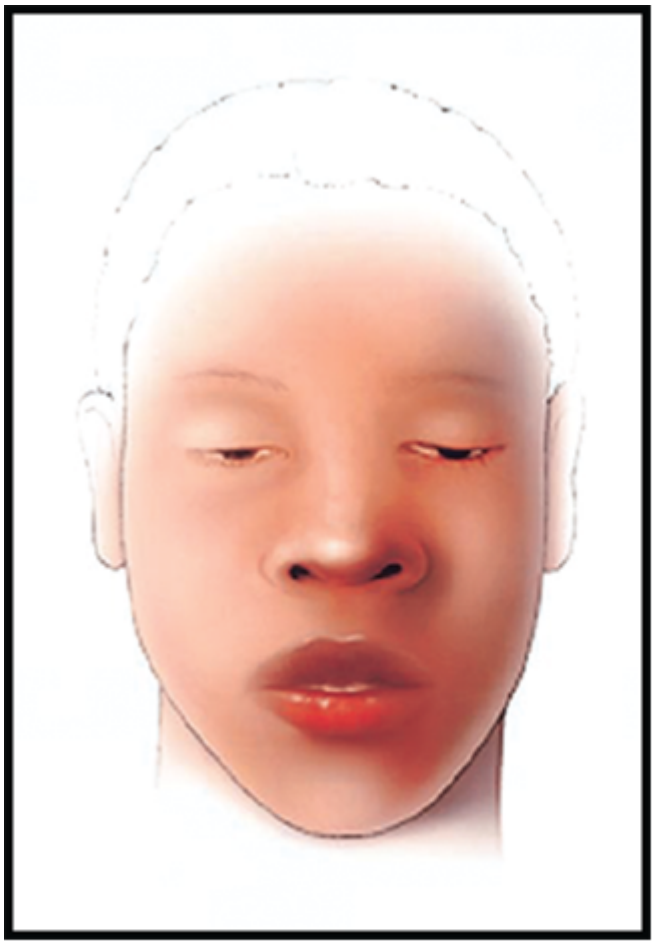
CLINICAL MANIFESTATIONS

The clinical syndrome of botulism consists of bilateral cranial-nerve palsies that may progress to respiratory compromise, a bilateral descending flaccid paralysis of voluntary muscles, and even death. The incubation period from ingestion of contaminated food to onset of symptoms in food-borne botulism is usually 8–36 h but can be as long as 10 days and is dose dependent. Incubation periods of 4–17 days have been documented in wound botulism associated with accidental injury. Estimation is difficult in cases involving injection drug users because patients may inject drugs several times daily. Similarly, the incubation period for infant botulism has not been established, but the fact that the illness has affected infants <3 days old suggests that this interval may be very short.

Cranial nerve deficits may manifest as some of the following: diplopia, dysarthria, dysphonia, ptosis, ophthalmoplegia, facial paralysis, and impaired gag reflex **(Fig. 148-1)**. Pupillary reflexes may be depressed, and fixed or dilated pupils are sometimes noted. Autonomic symptoms such as dizziness, dry mouth, and "sore throat" are common. Dysphagia may lead to increased pooling of secretions and the need for suctioning despite dry mouth. Constipation due to paralytic ileus is often noted in the days after illness onset, and urinary retention is also common. Patients are afebrile and remain alert and oriented, but dysarthria, ptosis, and paresis have sometimes led physicians to believe incorrectly that a patient's mental status is altered (Fig. 148-1). Either late or early in the illness, respiratory failure due either to paralysis of the diaphragm and accessory breathing muscles or to pharyngeal collapse secondary to cranial nerve paralysis may occur. Because of skeletal muscle paralysis, patients experiencing respiratory distress may appear placid and detached even as they near respiratory arrest.

FIGURE 148-1

Artist's rendition of an adult with mild botulism. The patient has ptosis and facial paralysis manifested as lethargy and expressionless facies. The patient is fully alert. *(Illustrator, James K. Archer; reprinted from J Sobel: Clin Infect Dis 41:1170, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Weakness descends from the head, often rapidly, to involve the neck, arms, thorax, and legs; weakness and some cranial nerve deficits can be asymmetric. Deep tendon reflexes typically are normal or may

progressively disappear. Paresthesias, while rare, have been reported. Ataxia, which has sometimes been reported, manifests not as cerebellar ataxia but rather as gait problems due to weakness or visual issues. The absence of cranial nerve palsies makes botulism highly unlikely, as does a lack of cranial nerve deficits at the onset of illness. Nausea, vomiting, and abdominal pain may precede or follow the onset of paralysis in food-borne botulism. Infants with botulism typically present with a reduced ability to suck and swallow, constipation, weakened voice, ptosis, sluggish pupils, hypotonia, lethargic appearance, and floppy neck; as in adults, illness can progress to generalized flaccidity and respiratory compromise.

Even when intubated, patients can sometimes respond to questions by moving their fingers or toes unless paralysis has affected the digits. Clinical improvement follows sprouting of new nerve terminals and may take weeks to months. Patients often require outpatient rehabilitation therapy and may experience residual deficits. Death in untreated botulism is usually due to airway obstruction from pharyngeal muscle paralysis and inadequate tidal volume resulting from paralysis of diaphragmatic and accessory respiratory muscles. Death can also result from hospital-associated infections and other sequelae of long-term paralysis, hospitalization, and mechanical ventilation.

A history of preparing improperly home-canned foods may assist with the diagnosis. Patients with wound botulism may or may not have a discernible wound or abscess. A history of injection drug use or the presence of track marks should prompt suspicion for the diagnosis. In a substantial number of cases, no epidemiologic clue is discerned at the time of clinical presentation.

DIAGNOSIS

Botulism is diagnosed primarily on clinical grounds, with laboratory confirmation by specific tests performed only in specialized public-health laboratories. In the setting of an outbreak in which multiple patients present to the same treatment facility, the diagnosis can still be challenging; patients in a case cluster may have different cranial nerve deficits, chief manifestations, illness severity, and signs and symptoms. Patients sometimes present with respiratory failure as their chief manifestation, and neurologic signs and symptoms are not immediately appreciated. The temporal occurrence of two or more cases with botulism-compatible symptoms is essentially pathognomonic because other illnesses that resemble botulism do not usually occur in clusters. Because of the rarity of this disease, few physicians in the United States have seen or will see botulism; this potential unfamiliarity contributes to delays in diagnosis.

A food history over the 10 days before illness onset, with emphasis on the 48 h before illness onset, is critical to the public health response. The names of contacts who may have shared foods should be obtained before illness progresses to respiratory failure. Specific questions should address the consumption of improperly home-canned foods, home-preserved and/or exotic foods, and products requiring refrigeration that have been left at room temperature in sealed plastic containers or bags. A history of recent consumption of home-canned food substantially enhances the probability of food-borne botulism.

Ascertainment of the patient's behavioral history related to injection drug use or the sustaining of a traumatic wound is important to the diagnosis of wound botulism. Caretakers' observations up to and

including the onset of symptoms are vital to the diagnosis of infant botulism. A history of recent abdominal surgery or antibiotic use may be important in the diagnosis of adult intestinal-colonization botulism.

DIFFERENTIAL DIAGNOSIS

The illnesses most commonly considered in the differential diagnosis of adult botulism cases include Guillain-Barré syndrome (GBS), myasthenia gravis, stroke syndromes, Eaton-Lambert syndrome, and tick paralysis. Less likely considerations are tetrodotoxin poisoning, shellfish poisoning, diphtheria, and tetanus. Botulism is sometimes confused with other illnesses because a neurologic examination is not adequately performed, particularly when a patient presents with symptoms that do not immediately indicate a neurologic illness—e.g., respiratory symptoms. A thorough history and a meticulous physical examination can effectively eliminate many alternative diagnoses, but a workup for other diagnoses should not delay treatment with botulinum antitoxin.

GBS, a rare autoimmune demyelinating polyneuropathy that often follows an acute infection, presents most often as an ascending paralysis. Case clusters are rare. Occasional GBS cases present as the Miller Fisher variant, which is a descending paralysis. The characteristic triad of ophthalmoplegia, ataxia, and areflexia in Miller Fisher variant GBS is easily mistaken for botulism. Protein levels in cerebrospinal fluid (CSF) can be elevated in GBS. In contrast, CSF findings are generally normal in botulism, although marginally elevated CSF protein concentrations have been reported. In experienced hands, electromyography may differentiate GBS from botulism but is limited by technical and interpretive expertise. The edrophonium (Tensilon) test is sometimes of value in distinguishing botulism (usually a negative result) from myasthenia gravis (usually a positive result), but results have been positive in botulism cases. In most cerebrovascular accidents, physical examination reveals unilateral paralysis and upper motor neuron signs. Brain imaging can reveal the rare basilar stroke that produces bilateral bulbar palsies. Eaton-Lambert syndrome usually manifests as proximal limb weakness in a patient already debilitated by cancer. Tick paralysis is a rare flaccid condition closely resembling botulism and caused by neurotoxins of certain ticks.

BOTULISM-SPECIFIC LABORATORY TESTS

Botulism is confirmed in specialized public health laboratories by demonstration of toxin in clinical specimens (e.g., serum, stool, gastric aspirate, and sterile-water enema samples) or in samples of ingested foods. Isolation of toxigenic clostridia from stool also provides evidence of botulism. Wound cultures yielding the organism are highly suggestive in symptomatic cases. The universally accepted method for confirmation of botulism is the mouse bioassay; no testing available in hospital or other clinical laboratories (e.g., a blood culture or culture-independent diagnostic test) can detect botulinum toxin or *C. botulinum*. Specific guidance about what specimens to collect should be obtained from the testing laboratory because the requirements vary with the form of botulism suspected. The earlier in the course of illness specimens are collected, the likelier they are to yield positive results, allowing confirmation of botulism. Clinical specimens collected early in the hospital admission process should be submitted for testing; toxin results are usually negative with specimens collected >7 days after symptom onset but have been positive weeks after illness

onset in high-level toxin exposures. Because of the extreme potency of botulinum toxin, serum toxin concentrations below the laboratory detection threshold can cause illness in a patient whose test yields a negative result; thus, a negative result does not rule out botulism. New laboratory tests for botulism are being developed but remain experimental. Standard blood work and radiologic studies are not useful in diagnosing botulism.

TREATMENT

TREATMENT

Botulism

The cornerstones of treatment for botulism are meticulous intensive care and administration of botulinum antitoxin. Because antitoxin is most beneficial early in the course of clinical illness, it should be administered empirically and before the time-consuming workup for other illnesses or laboratory confirmation is complete. Persons of all ages in whom botulism is suspected should be hospitalized immediately so that signs of respiratory failure—the usual cause of death—can be detected and managed. Vital capacity should be monitored frequently and mechanical ventilation provided as needed. Botulinum antitoxin can limit the progression of illness because it neutralizes toxin molecules in the circulation that have not yet bound to nerve endings. However, antitoxin does not reverse existing paralysis, which may take weeks to improve. In the United States, there are two licensed antitoxin products. Botulism Antitoxin Heptavalent® (BAT; Emergent Biosolutions, Rockville, MD) is an equine-derived heptavalent (A through G) product enzymatically de-speciated for treatment of all forms of adult botulism and for infant botulism not involving serotypes A and B. Botulism Immune Globulin Intravenous (BabyBIG®; California Department of Public Health, Sacramento, CA) is a human-derived product for treating infant botulism caused by serotype A and/or B only. Alternative antitoxins are available in some countries. Aminoglycosides and other medications that block the neuromuscular junction may potentiate botulism and should be avoided.

In wound botulism, suspect wounds and abscesses should be cleaned, debrided, and drained promptly. The role of penicillin and metronidazole in treatment and decolonization is unclear. It has been hypothesized that antimicrobial agents may increase circulating botulinum toxin from lysis of bacterial cells.

Person-to-person transmission of botulism does not occur. Universal precautions are the only infection-control measures required during inpatient care. Patients with botulism can acquire health care–associated infections, deep venous thromboses, and other ailments that occur among patients who are hospitalized and immobile for long periods.

NOTIFICATION, EXPERT CONSULTATION, AND ANTITOXIN PROVISION

Every botulism case is a public health emergency. Some countries maintain stockpiles of antitoxin for immediate response, whereas others must access supplies from other nations or commercial manufacturers when a suspected case occurs.

In the United States, clinicians must report every suspected case, regardless of form, on an emergency basis to their state health departments. The state health department may put the physician in contact with the 24/7 botulism consultation service at the Centers for Disease Control and Prevention (CDC) through the CDC Emergency Operations Center (770-488-7100) or a locally available service. The botulism consultant will review the case with the physician, and they will collaboratively determine whether botulism is likely. If indicated, the consultant will coordinate laboratory confirmation at appropriate testing facilities and facilitate emergency shipment of antitoxin for all adult cases and for infant cases not involving serotypes A and B. In the United States, botulinum antitoxin for non-infant cases is available exclusively from the CDC. Physicians who see suspected infant botulism cases should contact the California Department of Public Health Infant Botulism Treatment and Prevention Program (510-231-7600), which provides 24-h consultation and distributes antitoxin (BabyBIG®) for the treatment of infant botulism nationwide. Except in cases involving infants who reside in California, laboratory-testing requests must still be authorized by the state health department where the infant is located or by the CDC.

PREVENTION

No prophylaxis or licensed vaccine against botulism is available. Home-canning instructions and equipment have changed over the years. Up-to-date canning instructions from reliable sources (e.g., the U.S. Department of Agriculture or the U.S. Food and Drug Administration) should be followed to ensure food safety. Processed food should be stored properly and heated thoroughly before consumption. Because of the possible presence of spores, honey should not be given to infants (<12 months of age). Injection of illicit drugs increases the risk of botulism. All traumatic wounds should be meticulously cleaned to eliminate possible contamination with bacterial spores.

FURTHER READING

Angulo FJ et al: A large outbreak of botulism: The hazardous baked potato. J Infect Dis 178:172, 1998.

Hughes JM et al: Clinical features of types A and B food-borne botulism. Ann Intern Med 95:442, 1981.

McCarty CL et al: Large outbreak of botulism associated with a church potluck meal—Ohio, 2015. MMWR Morb Mortal Wkly Rep 64:802, 2015.

McCroskey LM, Hatheway CL: Laboratory findings in four cases of adult botulism suggest colonization of the intestinal tract. J Clin Microbiol 26:1052, 1988.

Morse DL et al: Garlic-in-oil associated botulism: Episode leads to product modification. Am J Public Health 80:1372, 1990.

Shapiro RL et al: Botulism in the United States: A clinical and epidemiologic review. Ann Intern Med 129:221, 1998.

Sobel J et al: Foodborne botulism in the United States, 1990–2000. Emerg Infect Dis 10:1606, 2004.

St Louis ME et al: Botulism from chopped garlic: Delayed recognition of a major outbreak. Ann Intern Med 108:363, 1988.

Werner SB et al: Wound botulism in California, 1951–1998: Recent epidemic in heroin injectors. Clin Infect Dis 31:1018, 2000.

Woodruff BA et al: Clinical and laboratory comparison of botulism from toxin types A, B, and E in the United States, 1975–1988. J Infect Dis 166:1281, 1992.

Chapter 149: Gas Gangrene and Other Clostridial Infections

Amy E. Bryant; Dennis L. Stevens

INTRODUCTION

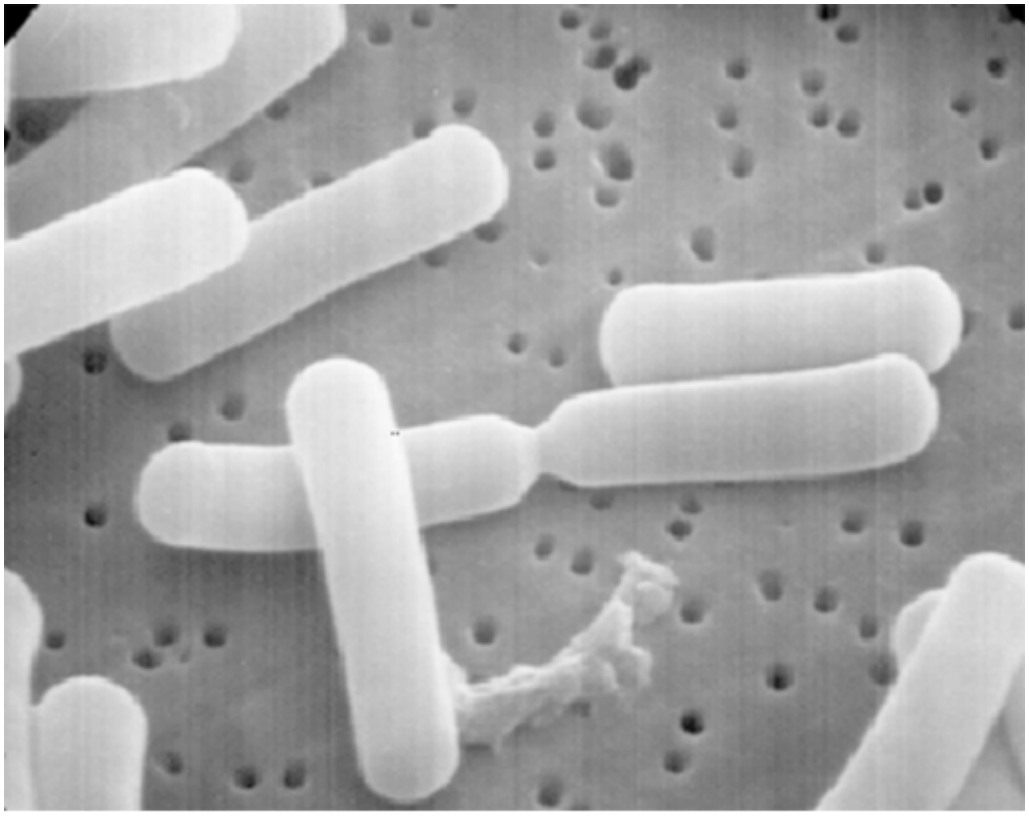
The genus *Clostridium* encompasses >60 species that may be commensals of the gut microflora or may cause a variety of infections in humans and animals through the production of a plethora of proteinaceous exotoxins. *C. tetani* and *C. botulinum*, for example, cause specific clinical disease by elaborating single but highly potent toxins. In contrast, *C. perfringens* and *C. septicum* cause aggressive necrotizing infections that are attributable to multiple toxins, including bacterial proteases, phospholipases, and cytotoxins.

ETIOLOGIC AGENT

Vegetative cells of *Clostridium* species are pleomorphic, rod-shaped, and arranged singly or in short chains (Fig. 149-1); the cells have rounded or sometimes pointed ends. Although clostridia stain gram-positive in the early stages of growth, they may appear to be gram-negative or gram-variable later in the growth cycle or in infected tissue specimens. Most strains are motile by means of peritrichous flagella; *C. septicum* swarms on solid media. Nonmotile species include *C. perfringens*, *C. ramosum*, and *C. innocuum*. Most species are obligately anaerobic, although clostridial tolerance to oxygen varies widely; some species (e.g., *C. septicum*, *C. tertium*) will grow but will not sporulate in air.

FIGURE 149-1

Scanning electron micrograph of *C. perfringens.*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clostridia produce more protein toxins than any other bacterial genus, and >25 clostridial toxins lethal to mice have been identified. These proteins include neurotoxins, enterotoxins, cytotoxins, collagenases, permeases, necrotizing toxins, lipases, lecithinases, hemolysins, proteinases, hyaluronidases, DNases, ADP-ribosyltransferases, and neuraminidases. Botulinum and tetanus neurotoxins are the most potent toxins known, with lethal doses of 0.2–10 ng/kg for humans. Epsilon toxin, a 33-kDa protein produced by *C. perfringens* types B and D, causes edema and hemorrhage in the brain, heart, spinal cord, and kidneys of animals. It is among the most lethal of the clostridial toxins and is considered a potential agent of bioterrorism. The genomic sequences of some pathogenic clostridia are now available and are likely to facilitate a comprehensive approach to understanding the virulence factors involved in clostridial pathogenesis.

EPIDEMIOLOGY AND TRANSMISSION

Clostridium species are widespread in nature, forming endospores that are commonly found in soil, feces, sewage, and marine sediments. The ecology of *C. perfringens* in soil is greatly influenced by the degree and duration of animal husbandry in a given location and is relevant to the incidence of gas gangrene caused by contamination of war wounds with soil. For example, the incidence of clostridial gas gangrene is higher in agricultural regions of Europe than in the Sahara Desert of Africa. Similarly, the incidences of tetanus and food-borne botulism are clearly related to the presence of clostridial spores in soil, water, and many foods. Clostridia are present in large numbers in the indigenous microbiota of the intestinal tract of humans and

animals, in the female genital tract, and on the oral mucosa. It should be noted that not all commensal clostridia are toxigenic.

Clostridial infections remain a serious public-health concern worldwide. In developing nations, food poisoning, necrotizing enterocolitis, and gas gangrene are common because large portions of the population are poor and have little or no immediate access to health care. These infections remain prevalent in developed countries as well. Gas gangrene commonly follows knife or gunshot wounds or vehicular accidents or develops as a complication of surgery or gastrointestinal carcinoma. Severe clostridial infections have emerged as a health threat to injection drug users and to women undergoing childbirth or abortion. Historically, clostridial gas gangrene has been the scourge of the battlefield. The global political situation portends another possible scenario involving mass casualties of war or terrorism, with extensive injuries conducive to gas gangrene. Thus there is an ongoing need to develop novel strategies to prevent or attenuate the course of clostridial infections in both civilians and military personnel. Vaccination against exotoxins important in pathogenesis would be of great benefit in developing nations and could also be used safely in at-risk populations such as the elderly, patients with diabetes who may require lower-limb surgery due to trauma or poor circulation, and those undergoing intestinal surgery. Moreover, a hyperimmune globulin would be a valuable tool for prophylaxis in victims of acute traumatic injury or for attenuation of the spread of infection in patients with established gas gangrene.

CLINICAL SYNDROMES

Life-threatening clostridial infections range from intoxications (e.g., food poisoning, tetanus) to necrotizing enteritis/colitis, bacteremia, myonecrosis, and toxic shock syndrome (TSS). **Tetanus and botulism are discussed in Chaps. 147 and 148, respectively. Colitis due to *C. difficile* is discussed in Chap. 129.**

CLOSTRIDIAL WOUND CONTAMINATION

Of open traumatic wounds, 30–80% reportedly are contaminated with clostridial species. In the absence of devitalized tissue, the presence of clostridia does not necessarily lead to infection. In traumatic injuries, clostridia are isolated with equal frequency from both suppurative and well-healing wounds. Thus, diagnosis and treatment of clostridial infection should be based on clinical signs and symptoms and not solely on bacteriologic findings.

POLYMICROBIAL INFECTIONS INVOLVING CLOSTRIDIA

Clostridial species may be found in polymicrobial infections also involving microbial components of the indigenous flora. In these infections, clostridia often appear in association with non-spore-forming anaerobes and facultative or aerobic organisms. Head and neck infections, conjunctivitis, brain abscess, sinusitis, otitis, aspiration pneumonia, lung abscess, pleural empyema, cholecystitis, septic arthritis, and bone infections all may involve clostridia. These conditions are often associated with severe local inflammation but may lack the characteristic systemic signs of toxicity and rapid progression seen in other clostridial infections. In addition, clostridia are isolated from ~66% of intraabdominal infections in which the

mucosal integrity of the bowel or respiratory system has been compromised. In this setting, *C. ramosum*, *C. perfringens*, and *C. bifermentans* are the most commonly isolated species. Their presence does not invariably lead to a poor outcome. Clostridia have been isolated from suppurative infections of the female genital tract (e.g., ovarian or pelvic abscess) and from diseased gallbladders. Although the most frequently isolated species is *C. perfringens*, gangrene is not typically observed; however, gas formation in the biliary system can lead to emphysematous cholecystitis, especially in diabetic patients. *C. perfringens* in association with mixed aerobic and anaerobic microbes can cause aggressive life-threatening type I necrotizing fasciitis or Fournier's gangrene.

The treatment of mixed aerobic/anaerobic infection of the abdomen, perineum, or gynecologic organs should be based on Gram's staining, culture, and antibiotic sensitivity information. Reasonable empirical treatment consists of ampicillin or ampicillin/sulbactam combined with either clindamycin or metronidazole **(Table 149-1)**. Broader gram-negative coverage may be necessary if the patient has recently been hospitalized or treated with antibiotics. Such coverage can be obtained by substituting ticarcillin/clavulanic acid, piperacillin/sulbactam, or a penem antibiotic for ampicillin or by adding a fluoroquinolone or an aminoglycoside to the regimen. Empirical treatment should be given for 10–14 days or until the patient's clinical condition improves.

TABLE 149-1

Treatment of Clostridial Infections

Condition	Antibiotic Treatment	Penicillin Allergy	Adjunctive Treatment/Note
Wound contamination	None	—	Treatment should be based on clinical signs and symptoms as listed below and not solely on bacteriologic findings.
Polymicrobial anaerobic infections involving clostridia (e.g., abdominal wall, gynecologic)	Ampicillin (2 g IV q4h) ***plus*** Clindamycin (600–900 mg IV q6–8h) ***plus*** Ciprofloxacin (400 mg IV q6–8 h)	Vancomycin (1 g IV q12h) ***plus*** Metronidazole (500 mg IV q6h) ***plus*** Ciprofloxacin (400 mg IV q6–8h)	Empirical therapy should be initiated. Therapy should be based on Gram's stain and culture results and on sensitivity data when available. Add gram-negative coverage if indicated (see text).
Clostridial sepsis	Penicillin (3–4 mU IV q4–6h) ***plus*** Clindamycin (600–900 mg IV q6–8h)	Clindamycin alone ***or*** Metronidazole (as above) ***or*** Vancomycin (as above)	Transient bacteremia without signs of systemic toxicity may be clinically insignificant.
Gas gangrene[a]	Penicillin G (4 mU IV q4–6 h) ***plus*** Clindamycin (600–900 mg IV q6–8h)	Cefoxitin (2 g IV q6h) ***plus*** Clindamycin (600–900 mg IV q6–8h)	Emergent surgical exploration and thorough debridement are extremely important. Hyperbaric oxygen therapy may be considered after surgery and antibiotic initiation.

[a] *C. tertium* is resistant to penicillin, cephalosporins, and clindamycin. Appropriate antibiotic therapy for *C. tertium* infection is vancomycin (1 g q12h IV) or metronidazole (500 mg q8h IV).

ENTERIC CLOSTRIDIAL INFECTIONS

C. perfringens type A is one of the most common bacterial causes of food-borne illness in the United States and Canada. The foods typically implicated include improperly cooked meat and meat products (e.g., gravy) in which residual spores germinate and proliferate during slow cooling or insufficient reheating. Illness results from the ingestion of food containing at least ~10^8 viable vegetative cells, which sporulate in the alkaline environment of the small intestine, producing *C. perfringens* enterotoxin in the process. The diarrhea that develops within 7–30 h of ingestion of contaminated food is generally mild and self-limiting; however, in the very young, the elderly, and the immunocompromised, symptoms are more severe and occasionally fatal. Enterotoxin-producing *C. perfringens* has been implicated as an etiologic agent of persistent diarrhea in elderly patients in nursing homes and tertiary-care institutions and has been considered to play a role in antibiotic-associated diarrhea without pseudomembranous colitis.

 *C. perfringens* strains associated with food poisoning possess the gene (*cpe*) coding for enterotoxin, which acts by forming pores in host cell membranes. *C. perfringens* strains isolated from non-food-borne diseases, such as antibiotic-associated and sporadic diarrhea, carry *cpe* on a plasmid that may be transmitted to other strains. Several methods have been described for the detection of *C. perfringens* enterotoxin in feces, including cell culture assay (Vero cells), enzyme-linked immunosorbent assay, reversed-phase latex agglutination, and polymerase chain reaction (PCR) amplification of *cpe*. Each method has its advantages and limitations.

 *Enteritis necroticans* (gas gangrene of the bowel) is a fulminating clinical illness characterized by extensive necrosis of the intestinal mucosa and wall. Cases can occur sporadically in adults or as epidemics in people of all ages. Enteritis necroticans is caused by α toxin– and β toxin–producing strains of *C. perfringens* type C; β toxin is located on a plasmid and is mainly responsible for pathogenesis. This life-threatening infection causes ischemic necrosis of the jejunum. In Papua New Guinea during the 1960s, enteritis necroticans (known in that locale as *pigbel*) was found to be the most common cause of death in childhood; it was associated with pig feasts and occurred both sporadically and in outbreaks. Intramuscular immunization against the β toxin resulted in a decreased incidence of the disease in Papua New Guinea, although the condition remains common. Enteritis necroticans has also been recognized in the United States, the United Kingdom, Germany (where it is known as *darmbrand*), and other developed nations; especially affected are adults who are malnourished or who have diabetes, alcoholic liver disease, or neutropenia.

Necrotizing enterocolitis, a disease resembling enteritis necroticans but associated with *C. perfringens* type A, has been found in North America in previously healthy adults. It is also a serious gastrointestinal disease of

low-birth-weight (premature) infants hospitalized in neonatal intensive care units. The etiology and pathogenesis of this disease have remained enigmatic for more than four decades. Pathologic similarities between necrotizing enterocolitis and enteritis necroticans include the pattern of small-bowel necrosis involving the submucosa, mucosa, and muscularis; the presence of gas dissecting the tissue planes; and the degree of inflammation. In contrast to enteritis necroticans, which most commonly involves the jejunum, necrotizing enterocolitis affects the ileum and frequently the ileocecal valve. Both diseases may manifest as intestinal gas cysts, although this feature is more common in necrotizing enterocolitis. The sources of the gas, which contains hydrogen, methane, and carbon dioxide, are probably the fermentative activities of intestinal bacteria, including clostridia. Epidemiologic data support an important role for *C. perfringens* or other gas-producing microorganisms (e.g., *C. neonatale*, certain other clostridia, or *Klebsiella* species) in the pathogenesis of necrotizing enterocolitis.

Patients with suspected clostridial enteric infection should undergo nasogastric suction and receive IV fluids. Pyrantel is given by mouth, and the bowel is rested by fasting. Benzylpenicillin (1 mU) is given IV every 4 h, and the patient is observed for complications requiring surgery. Patients with mild cases recover without surgical intervention. However, if surgical indications are present (gas in the peritoneal cavity, absent bowel sounds, rebound tenderness, abdominal rigidity), the mortality rate ranges from 35 to 100%; a fatal outcome is due in part to perforation of the intestine.

As pigbel continues to be a common disease in Papua New Guinea, consideration should be given to the use of a *C. perfringens* type C β toxoid vaccine in local areas. Two doses given 3–4 months apart are preventive.

CLOSTRIDIAL BACTEREMIA

Clostridium species are important causes of bloodstream infections. Molecular epidemiologic studies of anaerobic bacteremia have identified *C. perfringens* and *C. tertium* as the two most frequently isolated species; these organisms cause up to 79 and 5%, respectively, of clostridial bacteremias. Occasionally, *C. perfringens* bacteremia occurs in the absence of an identifiable infection at another site. When associated with myonecrosis, bacteremia has a grave prognosis.

C. septicum is also commonly associated with bacteremia. This species is isolated only rarely from the feces of healthy individuals but may be found in the normal appendix. More than 50% of patients whose blood cultures are positive for this organism have some gastrointestinal anomaly (e.g., diverticular disease) or underlying malignancy (e.g., carcinoma of the colon). In addition, a clinically important association of *C. septicum* bacteremia with neutropenia of any origin—and, more specifically, with neutropenic enterocolitis involving the terminal ileum or cecum—has been observed. Patients with diabetes mellitus, severe atherosclerotic cardiovascular disease, or anaerobic myonecrosis (gas gangrene) also may develop *C. septicum* bacteremia. *C. septicum* has been recovered from the bloodstream of cirrhotic patients, as have *C. perfringens*, *C. bifermentans*, and other clostridia. Infections of the bloodstream by *C. sordellii* and *C. perfringens* have been associated with TSS.

Bloodstream infection by *C. tertium*, either alone or in combination with *C. septicum* or *C. perfringens*, can be found in patients with serious underlying disease such as malignancy or acute pancreatitis, with or without neutropenic enterocolitis; the frequency has not been systematically studied. *C. tertium* may present special problems in terms of both identification and treatment. This organism may stain gram-negative; is aerotolerant; and is resistant to metronidazole, clindamycin, and cephalosporins.

Other clostridia from the *C. clostridioforme* group (including *C. clostridioforme*, *C. hathewayi*, and *C. bolteae*) can cause bacteremia.

The clinical importance of recognizing clostridial bacteremia—especially that due to *C. septicum*—and starting appropriate treatment immediately (Table 149-1) cannot be overemphasized. Patients with this condition usually are gravely ill, and infection may metastasize to distant anatomic sites, resulting in spontaneous myonecrosis (see next section). Alternative methods to identify bacteremia-causing clostridial species, such as PCR or other rapid diagnostic tests, are not currently available. Anaerobic blood cultures and Gram's stain interpretation remain the best diagnostic tests at this point.

CLOSTRIDIAL SKIN AND SOFT-TISSUE INFECTIONS

Histotoxic clostridial species such as *C. perfringens*, *C. histolyticum*, *C. septicum*, *C. novyi*, and *C. sordellii* cause aggressive necrotizing infections of the skin and soft tissues. These infections are attributable in part to the elaboration of bacterial proteases, phospholipases, and cytotoxins. Necrotizing clostridial soft-tissue infections are rapidly progressive and are characterized by marked tissue destruction, gas in the tissues, and shock; they frequently end in death. Severe pain, crepitus, brawny induration with rapid progression to skin sloughing, violaceous bullae, and marked tachycardia are characteristics found in the majority of patients.

Clostridial Myonecrosis (Gas Gangrene)

TRAUMATIC GAS GANGRENE

C. perfringens myonecrosis (gas gangrene) is one of the most fulminant gram-positive bacterial infections of humans. Even with appropriate antibiotic therapy and management in an intensive care unit, tissue destruction can progress rapidly. Gas gangrene is accompanied by bacteremia, hypotension, and multiorgan failure and is invariably fatal if untreated. Gas gangrene is a true emergency and requires immediate surgical debridement.

The development of gas gangrene requires an anaerobic environment and contamination of a wound with spores or vegetative organisms. Devitalized tissue, foreign bodies, and ischemia reduce locally available oxygen levels and favor outgrowth of vegetative cells and spores. Thus conditions predisposing to traumatic gas gangrene include crush-type injury, laceration of large or medium-sized arteries, and open fractures of long bones that are contaminated with soil or bits of clothing containing the bacterial spores. Gas gangrene of the abdominal wall and flanks follows penetrating injuries such as knife or gunshot wounds that are sufficient to compromise intestinal integrity, with resultant leakage of the bowel contents into the soft tissues. Proximity to fecal sources of bacteria is a risk factor for cases following hip surgery, adrenaline injections into the buttocks, or amputation of the leg for ischemic vascular disease. In the last decade,

cutaneous gas gangrene caused by *C. perfringens*, *C. novyi*, and *C. sordellii* has been described in the United States and northern Europe among persons injecting black-tar heroin subcutaneously.

The incubation period for traumatic gas gangrene can be as short as 6 h and is usually <4 days. The infection is characterized by the sudden onset of excruciating pain at the affected site and the rapid development of a foul-smelling wound containing a thin serosanguineous discharge and gas bubbles. Brawny edema and induration develop and give way to cutaneous blisters containing bluish to maroon-colored fluid. Such tissue later may become liquefied and slough. The margin between healthy and necrotic tissue often advances several inches per hour despite appropriate antibiotic therapy, and radical amputation remains the single best life-saving intervention. Shock and organ failure frequently accompany gas gangrene; when patients become bacteremic, the mortality rate exceeds 50%.

Diagnosis of traumatic gas gangrene is not difficult because the infection always begins at the site of significant trauma, is associated with gas in the tissue, and is rapidly progressive. Gram's staining of drainage or tissue biopsy is usually definitive, demonstrating large gram-positive (or gram-variable) rods, an absence of inflammatory cells, and widespread soft-tissue necrosis.

SPONTANEOUS (NONTRAUMATIC) GAS GANGRENE

Spontaneous gas gangrene generally occurs via hematogenous seeding of normal muscle with histotoxic clostridia—principally *C. perfringens*, *C. septicum*, and *C. novyi* and occasionally *C. tertium*—from a gastrointestinal tract portal of entry (as in colonic malignancy, inflammatory bowel disease, diverticulitis, necrotizing enterocolitis, cecitis, or distal ileitis or after gastrointestinal surgery, including colonoscopic polypectomy). These gastrointestinal pathologies permit bacterial access to the bloodstream; consequently, aerotolerant *C. septicum* can proliferate in normal tissues. Patients surviving bacteremia or spontaneous gangrene due to *C. septicum* should undergo aggressive diagnostic studies to rule out gastrointestinal pathology.

Additional predisposing host factors include leukemia, lymphoproliferative disorders, cancer chemotherapy, radiation therapy, and AIDS. Cyclic, congenital, or acquired neutropenia also is strongly associated with an increased incidence of spontaneous gas gangrene due to *C. septicum*; in such cases, necrotizing enterocolitis, cecitis, or distal ileitis is common, particularly among children.

The first symptom of spontaneous gas gangrene may be confusion followed by the abrupt onset of excruciating pain in the absence of trauma. These findings, along with fever, should heighten suspicion of spontaneous gas gangrene. However, because of the lack of an obvious portal of entry, the correct diagnosis is frequently delayed or missed. The infection is characterized by rapid progression of tissue destruction with demonstrable gas in the tissue **(Fig. 149-2)**. Swelling increases, and bullae filled with clear, cloudy, hemorrhagic, or purplish fluid appear. The surrounding skin has a purple hue, which may reflect vascular compromise resulting from the diffusion of bacterial toxins into surrounding tissues. Invasion of healthy tissue rapidly ensues, with quick progression to shock and multiple-organ failure. Mortality rates in this setting range from 67 to 100% among adults; among children, the mortality rate is 59%, with the majority of deaths occurring within 24 h of onset.

FIGURE 149-2

Radiograph of patient with spontaneous gas gangrene due to *C. septicum*, demonstrating gas in the affected arm and shoulder.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOGENESIS OF GAS GANGRENE

In traumatic gas gangrene, organisms are introduced into devitalized tissue. It is important to recognize that, for *C. perfringens* and *C. novyi*, trauma must be sufficient to interrupt the blood supply and thereby to establish an optimal anaerobic environment for growth of these species. These conditions are not strictly required for the more aerotolerant species such as *C. septicum* and *C. tertium*, which can seed normal tissues from gastrointestinal lesions. Once introduced into an appropriate niche, the organisms proliferate locally and elaborate exotoxins.

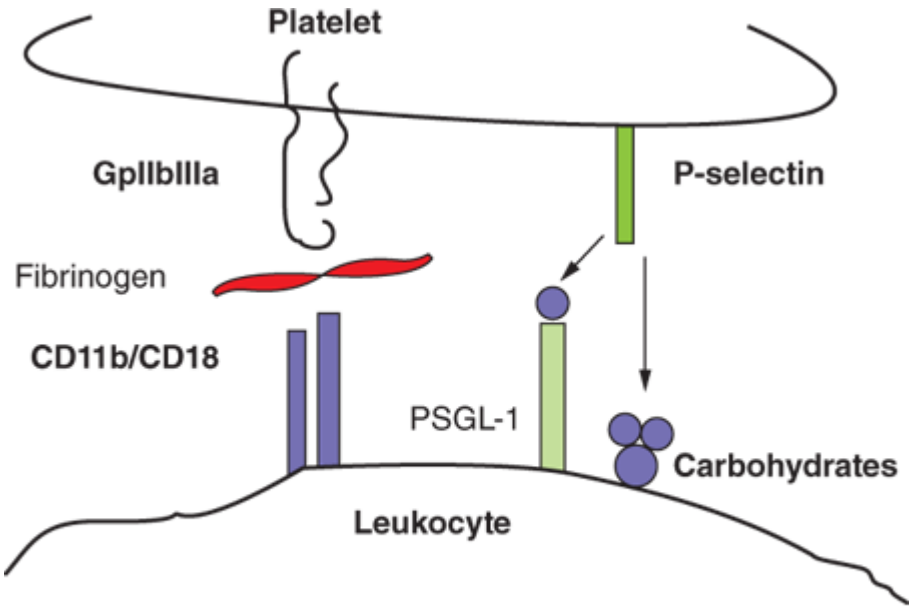
The major *C. perfringens* extracellular toxins implicated in gas gangrene are α toxin and θ toxin. A lethal hemolysin that has both phospholipase C and sphingomyelinase activities, α toxin has been implicated as the major virulence factor of *C. perfringens*: immunization of mice with the C-terminal domain of α toxin

provides protection against lethal challenge with *C. perfringens*, and isogenic α toxin–deficient mutant strains of *C. perfringens* are not lethal in a murine model of gas gangrene. Recently, a human single-chain recombinant antibody to α toxin that has significant preventive and therapeutic efficacy in mice has been developed.

It has been shown in experimental models that the severe pain, rapid progression, marked tissue destruction, and absence of neutrophils in *C. perfringens* gas gangrene are attributable in large part to α toxin–induced occlusion of blood vessels by heterotypic aggregates of platelets and neutrophils. The formation of these aggregates, which occurs within minutes, is largely mediated by α toxin's ability to activate the platelet adhesion molecule gpIIb/IIIa **(Fig. 149-3)**; the implication is that platelet glycoprotein inhibitors (e.g., eptifibatide, abciximab) may be therapeutic for maintaining tissue blood flow.

FIGURE 149-3

Schematic illustration of the molecular mechanisms of *C. perfringens* α toxin–induced platelet/neutrophil aggregates. Homotypic aggregates of platelets (not shown) and heterotypic aggregates of platelets and leukocytes are due to α toxin–induced activation of the platelet fibrinogen receptor gpIIb/IIIa and upregulation of leukocyte CD11b/CD18. Binding of fibrinogen (*red*) bridges the connection between these adhesion molecules on adjacent cells. An auxiliary role for α toxin–induced upregulation of platelet P-selectin and its binding to leukocyte P-selectin glycoprotein ligand 1 (PSGL-1) or other leukocyte surface carbohydrates also has been demonstrated.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

C. perfringens θ toxin (*perfringolysin*) is a member of the thiol-activated cytolysin family known as cholesterol-dependent cytolysins, which includes streptolysin O from group A *Streptococcus*, pneumolysin from *Streptococcus pneumoniae*, and several other toxins. Cholesterol-dependent cytolysins bind as oligomers to cholesterol in host cell membranes. At high concentrations, these toxins form ring-like pores resulting in cell lysis. At sub-lytic concentrations, θ toxin hyperactivates phagocytes and vascular endothelial cells.

Cardiovascular collapse and end-organ failure occur late in the course of *C. perfringens* gas gangrene and are largely attributable to both direct and indirect effects of α and θ toxins. In experimental models, θ toxin causes markedly reduced systemic vascular resistance but increased cardiac output (i.e., "warm shock"), probably via induction of endogenous mediators (e.g., prostacyclin, platelet-activating factor) that cause vasodilation. This effect is similar to that observed in gram-negative sepsis. In sharp contrast, α toxin directly suppresses myocardial contractility; the consequence is profound hypotension due to a sudden reduction in cardiac output. The roles of other endogenous mediators, such as cytokines (e.g., tumor necrosis factor, interleukin 1, interleukin 6) and vasodilators (e.g., bradykinin) have not been fully elucidated.

C. septicum produces three main toxins—α toxin (lethal, hemolytic, necrotizing activity), β toxin (DNase), and γ toxin (hyaluronidase)—as well as a protease and a neuraminidase. Unlike the α toxin of *C. perfringens*, that of *C. septicum* does not possess phospholipase activity. The mechanisms remain to be fully elucidated, but it is likely that each of these toxins contributes uniquely to *C. septicum* gas gangrene.

TREATMENT

TREATMENT

Gas Gangrene

Patients with suspected gas gangrene (either traumatic or spontaneous) should undergo prompt surgical inspection of the infected site. Direct examination of a Gram-stained smear of the involved tissues is of major importance. Characteristic histologic findings in clostridial gas gangrene include widespread tissue destruction, a paucity of leukocytes in infected tissues in conjunction with an accumulation of leukocytes in adjacent vessels **(Fig. 149-4)**, and the presence of gram-positive rods (with or without spores). CT and MRI are invaluable for determining whether the infection is localized or is spreading along fascial planes, and needle aspiration or punch biopsy may provide an etiologic diagnosis in at least 20% of cases. However, these techniques should not replace surgical exploration, Gram's staining, and histopathologic examination. When spontaneous gas gangrene is suspected, blood should be cultured since bacteremia usually precedes cutaneous manifestations by several hours.

For patients with evidence of clostridial gas gangrene, thorough emergent surgical debridement is of extreme importance. All devitalized tissue should be widely resected back to healthy viable muscle and skin so as to remove conditions that allow anaerobic organisms to continue proliferating. Closure of traumatic wounds or compound fractures should be delayed for 5–6 days until it is certain that these sites are free of infection.

Except for infection caused by *C. tertium* (see below), antibiotic treatment of traumatic or spontaneous gas gangrene (Table 149-1) consists of the administration of penicillin and clindamycin for 10–14 days. Penicillin is recommended on the basis of in vitro sensitivity data; clindamycin is recommended because of its superior efficacy over penicillin in animal models of *C. perfringens* gas gangrene and in some clinical reports. Controlled clinical trials comparing the efficacy of these agents in humans have not been performed. In the penicillin-allergic patient, clindamycin may be used alone. The superior efficacy of clindamycin is probably

due to its ability to inhibit bacterial protein toxin production, its insensitivity to the size of the bacterial load or the stage of bacterial growth, and its ability to modulate the host's immune response.

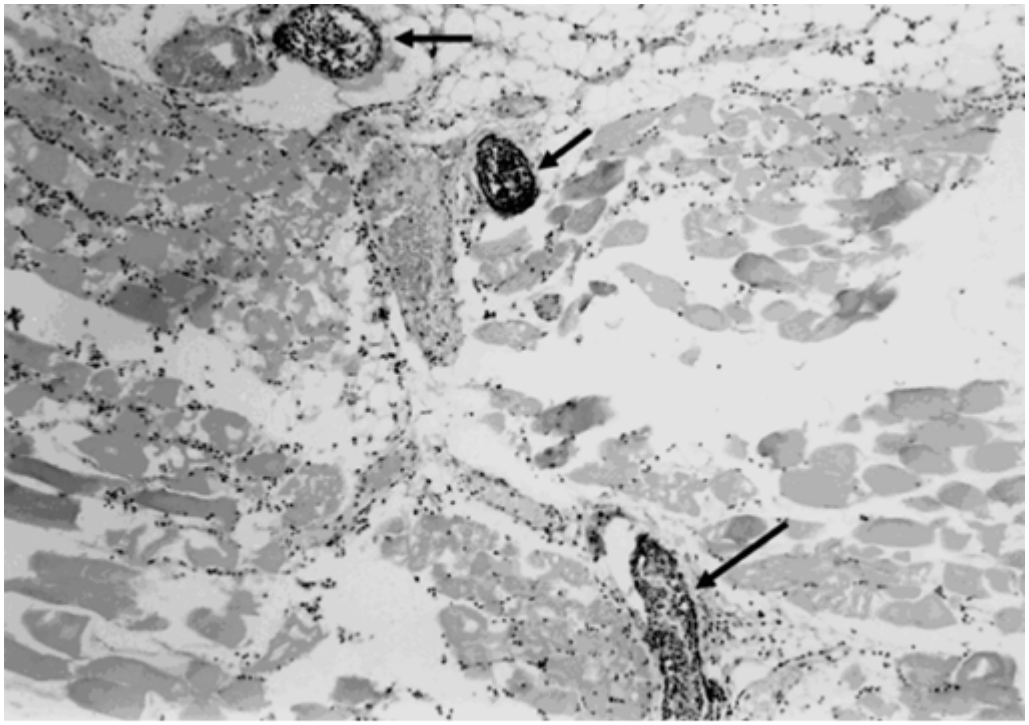
C. tertium is resistant to penicillin, cephalosporins, and clindamycin. Appropriate antibiotic therapy for *C. tertium* infection is vancomycin (1 g every 12 h IV) or metronidazole (500 mg every 8 h IV).

The value of adjunctive treatment with hyperbaric oxygen (HBO) for gas gangrene remains controversial. Basic-science studies suggest that HBO can inhibit the growth of *C. perfringens* but not that of the more aerotolerant *C. septicum*. In vitro, blood and macerated muscle inhibit the bactericidal potential of HBO. Numerous studies in animals demonstrate little efficacy of HBO alone, whereas antibiotics alone—especially those that inhibit bacterial protein synthesis—confer marked benefits. Addition of HBO to the therapeutic regimen provides some additional benefit, but only if surgery and antibiotic administration precede HBO treatment.

In conclusion, gas gangrene is a rapidly progressive infection whose outcome depends on prompt recognition, emergent surgery, and timely administration of antibiotics that inhibit toxin production. Gas gangrene associated with bacteremia probably represents a later stage of illness and is associated with the worst outcomes. Emergent surgical debridement is crucial to ensure survival, and ancillary procedures (e.g., CT or MRI) or transport to HBO units should not delay this intervention. Some trauma centers associated with HBO units may have special expertise in managing these aggressive infections, but proximity and speed of transfer must be carefully weighed against the need for haste.

FIGURE 149-4

Histopathology of experimental gas gangrene due to *C. perfringens*, demonstrating widespread muscle necrosis, a paucity of leukocytes in infected tissues, and accumulation of leukocytes in adjacent vessels (*arrows*). These features are due to the effects of α and θ toxins on muscle cells, platelets, leukocytes, and endothelial cells.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PROGNOSIS OF GAS GANGRENE

The prognosis for patients with gas gangrene is more favorable when the infection involves an extremity rather than the trunk or visceral organs, since debridement of the latter sites is more difficult. Gas gangrene is most likely to progress to shock and death in patients with associated bacteremia and intravascular hemolysis. Mortality rates are highest for patients in shock at the time of diagnosis. Mortality rates are relatively high among patients with spontaneous gas gangrene, especially that due to *C. septicum*. Survivors of gas gangrene may undergo multiple debridements and face long periods of hospitalization and rehabilitation.

PREVENTION OF GAS GANGRENE

Initial aggressive debridement of devitalized tissue can reduce the risk of gas gangrene in contaminated deep wounds. Interventions to be avoided include prolonged application of tourniquets and surgical closure of traumatic wounds; patients with compound fractures are at significant risk for gas gangrene if the wound is closed surgically. Vaccination against α toxin is protective in experimental animal models of *C. perfringens* gas gangrene but has not been investigated in humans. In addition, as mentioned above, a hyperimmune globulin would represent a significant advance for prophylaxis in victims of acute traumatic injury or for attenuation of the spread of infection in patients with established gas gangrene.

Toxic Shock Syndrome

Clostridial infection of the endometrium, particularly that due to *C. sordellii*, can develop after gynecologic procedures, childbirth, or abortion (spontaneous or elective, surgical or medical) and, once established, proceeds rapidly to TSS and death. Systemic manifestations, including edema, effusions, profound leukocytosis, and hemoconcentration, are followed by the rapid onset of hypotension and multiple-organ failure. Elevation of the hematocrit to 75–80% and leukocytosis of 50,000–200,000 cells/μL, with a left shift,

are characteristic of *C. sordellii* infection. Pain may not be a prominent feature, and fever is typically absent. In one series, 18% of 45 cases of *C. sordellii* infection were associated with normal childbirth, 11% with medically induced abortion, and 0.4% with spontaneous abortion; the case–fatality rate was 100% in these groups. Of the infections in this series that were not related to gynecologic procedures or childbirth, 22% occurred in injection drug users, and 50% of these patients died. Other infections followed trauma or surgery (42%), mostly in healthy persons, and 53% of these patients died. Overall, the mortality rate was 69% (31 of 45 cases). Of patients who succumbed, 85% died within 2–6 days after infection onset or following procedures. Rapidly fatal, spontaneous *C. bifermentans* necrotizing endometritis with toxic shock, leukemoid reaction, and capillary leak has also been described.

Early diagnosis of *C. sordellii* infections often proves difficult for several reasons. First, the prevalence of these infections is low. Second, the initial symptoms are nonspecific and frankly misleading. Early in the course, the illness resembles any number of infectious diseases, including viral syndromes. Given these vague symptoms and an absence of fever, physicians usually do not aggressively pursue additional diagnostic tests. The absence of local evidence of infection and the lack of fever make early diagnosis of *C. sordellii* infection particularly problematic in patients who develop deep-seated infection following childbirth, therapeutic abortion, gastrointestinal surgery, or trauma. Such patients are frequently evaluated for pulmonary embolization, gastrointestinal bleeding, pyelonephritis, or cholecystitis. Unfortunately, such delays in diagnosis increase the risk of death, and, as in most necrotizing soft-tissue infections, patients are hypotensive with evidence of organ dysfunction by the time local signs and symptoms become apparent. In contrast, infection is more readily suspected in injection drug users presenting with local swelling, pain, and redness at injection sites; early recognition probably contributes to the lower mortality rates in this group.

Physicians should suspect *C. sordellii* infection in patients who present within 2–7 days after injury, surgery, drug injection, childbirth, or abortion and who report pain, nausea, vomiting, and diarrhea but are afebrile. There is little information regarding appropriate treatment for *C. sordellii* infections. In fact, the interval between onset of symptoms and death is often so short that there is little time to initiate empirical antimicrobial therapy. Indeed, anaerobic cultures of blood and wound aspirates are time-consuming, and many hospital laboratories do not routinely perform antimicrobial sensitivity testing on anaerobes. Antibiotic susceptibility data from older studies suggest that *C. sordellii*, like most clostridia, is susceptible to β-lactam antibiotics, clindamycin, tetracycline, and chloramphenicol but is resistant to aminoglycosides and sulfonamides. Antibiotics that suppress toxin synthesis (e.g., clindamycin) may possibly prove useful as therapeutic adjuncts since they are effective in necrotizing infections due to other toxin-producing gram-positive organisms.

Other Clostridial Skin and Soft-Tissue Infections

Crepitant cellulitis (also called *anaerobic cellulitis*) occurs principally in diabetic patients and characteristically involves subcutaneous tissues or retroperitoneal tissues, whereas the muscle and fascia are not involved. This infection can progress to fulminant systemic disease.

Cases of *C. histolyticum* infection with cellulitis, abscess formation, or endocarditis have also been documented in injection drug users. Endophthalmitis due to *C. sordellii* or *C. perfringens* has been described. *C. ramosum* is also isolated frequently from clinical specimens, including blood and both intraabdominal and soft tissues. This species may be resistant to clindamycin and multiple cephalosporins.

FURTHER READING

Aldape MJ et al: *Clostridium sordellii* infection: Epidemiology, clinical findings, and current perspectives on diagnosis and treatment. Clin Infect Dis 43:1436, 2006.

Bodey GP et al: Clostridial bacteremia in cancer patients. A 12-year experience. Cancer 67:1928, 1991.

Bos J et al: Fatal necrotizing colitis following a foodborne outbreak of enterotoxigenic *Clostridium perfringens* type A infection. Clin Infect Dis 40:e78, 2005. [PubMed: 15844055]

Bryant AE et al: Clostridial gas gangrene II: Phospholipase C–induced activation of platelet gpIIb/IIIa mediates vascular occlusion and myonecrosis in *C. perfringens* gas gangrene. J Infect Dis 182:808, 2000.

Obladen M: Necrotizing enterocolitis—150 years of fruitless search for the cause. Neonatology 96:203, 2009.

Sayeed S et al: Beta toxin is essential for the intestinal virulence of *Clostridium perfringens* type C disease isolate CN3685 in a rabbit ileal loop model. Mol Microbiol 67:15, 2008.

Smith LDS, Williams BL: *The Pathogenic Anaerobic Bacteria*, 3rd ed. Springfield, IL: Charles C Thomas, 1984.

Stevens DL et al: *Clostridium*, in *Manual of Clinical Microbiology*, 11th ed, JH Jorgensen et al (eds). Washington, DC, ASM Press, 2015, pp 940–966. [PubMed: NBK368623] [PubMed: 4284726]

Stevens DL et al: Practice guidelines for the diagnosis and management of skin and soft tissue infections: 2014 update by the Infectious Diseases Society of America. Clin Infect Dis 59:e10, 2014. [PubMed: 24973422]

Wang C et al: Hyperbaric oxygen for treating wounds: A systematic review of the literature. Arch Surg 138:272, 2003.

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair

Chapter 150: Meningococcal Infections

Andrew J. Pollard

DEFINITION

Infection with *Neisseria meningitidis* most commonly manifests as asymptomatic colonization in the nasopharynx of healthy adolescents and adults. Invasive disease occurs rarely, usually presenting as either bacterial meningitis or meningococcal septicemia. Patients may also present with occult bacteremia, pneumonia, septic arthritis, conjunctivitis, and chronic meningococcemia.

ETIOLOGY AND MICROBIOLOGY

N. meningitidis is a gram-negative aerobic diplococcus that colonizes humans only and that causes disease after transmission to a susceptible individual. Several related neisserial organisms have been recognized, including the pathogen *N. gonorrhoeae* and the commensals *N. lactamica*, *N. flavescens*, *N. mucosa*, *N. sicca*, and *N. subflava*. *N. meningitidis* is a catalase- and oxidase-positive organism that utilizes glucose and maltose to produce acid.

Meningococci associated with invasive disease are usually encapsulated with polysaccharide, and the antigenic nature of the capsule determines an organism's capsular group (serogroup) (Table 150-1). In total, 12 capsular groups have been identified (A–C, X–Z, 29E, W, H–J, and L), but just six of these—A, B, C, X, Y, and W (formerly W135)—account for the majority of cases of invasive disease. Group D is often listed as the thirteenth capsular group but has recently been identified as an unencapsulated variant of group C. Meningococci are commonly isolated from the nasopharynx in studies of carriage; the lack of capsule often is a result of phase variation of capsule expression, but as many as 16% of isolates lack the genes for capsule synthesis and assembly. These "capsule-null" meningococci and those that express capsules other than A, B, C, X, Y, and W are only rarely associated with invasive disease and are most commonly identified in the nasopharynx of asymptomatic carriers.

TABLE 150-1

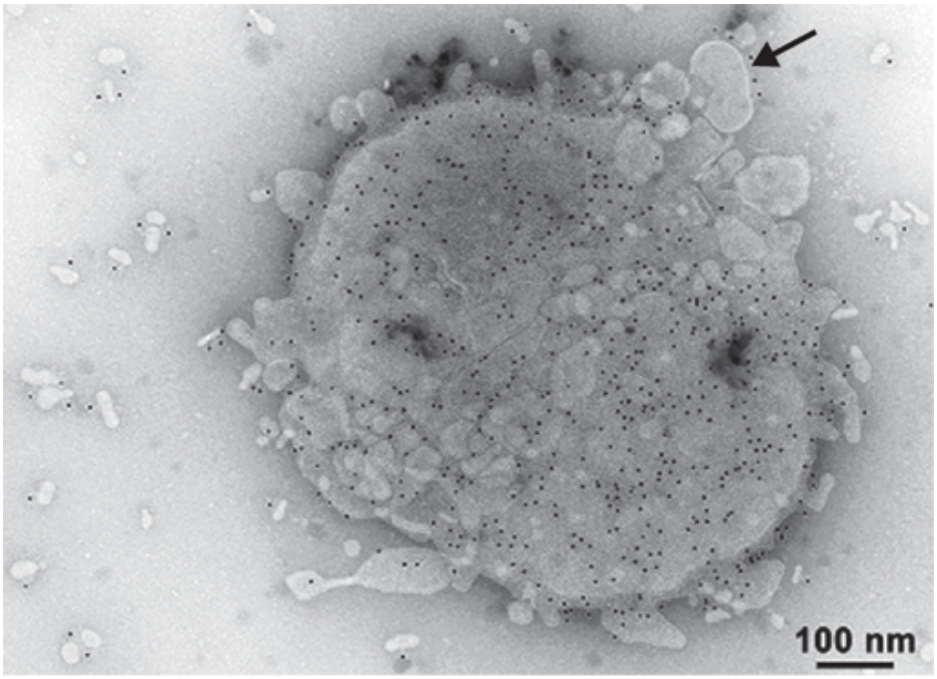
Structure of the Polysaccharide Capsule of Common Disease-Causing Meningococci

Meningococcal Capsular Group	Chemical Structure of Oligosaccharide	Current Disease Epidemiology
A	2-Acetamido-2-deoxy-D-mannopyranosyl phosphate	Epidemic disease mainly in sub-Saharan Africa; sporadic cases worldwide
B	α-2,8-*N*-acetylneuraminic acid	Sporadic cases worldwide; propensity to cause hyperendemic disease
C	α-2,9-*O*-acetylneuraminic acid	Small outbreaks and sporadic disease
Y	4-*O*-α-D-glucopyranosyl-*N*-acetylneuraminic acid	Sporadic disease and occasional small institutional outbreaks
W	4-*O*-α-D-galactopyranosyl-*N*-acetylneuraminic acid	Sporadic disease; outbreaks of disease associated with mass gatherings; epidemics in sub-Saharan Africa
X	(α1→4) *N*-acetyl-D-glucosamine-1-phosphate	Sporadic disease and large outbreaks in the meningitis belt of Africa

Beneath the capsule, meningococci are surrounded by an outer phospholipid membrane containing lipopolysaccharide (LPS, endotoxin) and multiple outer-membrane proteins **(Figs. 150-1 and 150-2)**. Antigenic variability in porins expressed in the outer membrane defines the serotype (PorB) and serosubtype (PorA) of the organism, and structural differences in LPS determine the immunotype. Serologic methods for typing of meningococci are restricted by the limited availability of serologic reagents that can distinguish among the organisms' highly variable surface proteins. Where available, high-throughput antigen gene sequencing has superseded serology for meningococcal typing. A large database of antigen gene sequences for the outer-membrane proteins PorA, PorB, FetA, Opa, and factor H–binding protein is available online (*www.neisseria.org*). The number of specialized iron-regulated proteins found in the meningococcal outer membrane (e.g., FetA and transferrin-binding proteins) highlights the organisms' dependence on iron from human sources. A thin peptidoglycan cell wall separates the outer membrane from the cytoplasmic membrane.

FIGURE 150-1

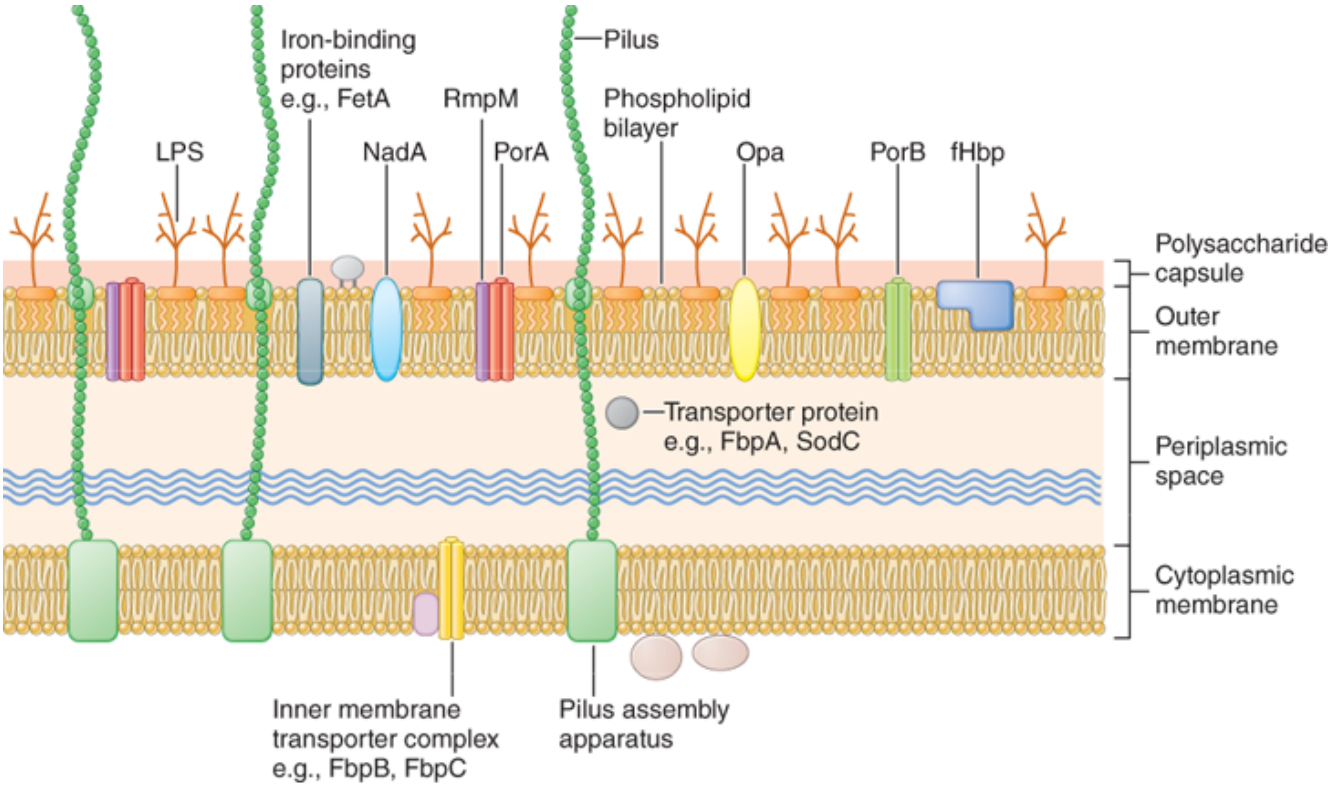
Electron micrograph of *Neisseria meningitidis.* Black dots are gold-labeled polyclonal antibodies binding surface opacity proteins. Blebs of outer membrane can be seen being released from the bacterial surface (*arrow*). *(Photo courtesy of D. Ferguson, Oxford University.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 150-2

Cross-section through surface structures of *Neisseria meningitidis.* LPS, lipopolysaccharide. *(Reprinted with permission from M Sadarangani, AJ Pollard: Lancet Infect Dis 10:112, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The structure of meningococcal populations involved in local and global spread has been studied with multilocus enzyme electrophoresis (MLEE), which characterizes isolates according to differences in the electrophoretic mobility of

cytoplasmic enzymes. However, this technique was replaced by multilocus sequence typing (MLST), in which meningococci are characterized by sequence types assigned on the basis of sequences of internal fragments of seven housekeeping genes. The online MLST database currently includes more than 12,000 unique *Neisseria* sequence types *(pubmlst.org/neisseria/)*. A limited number of hyperinvasive lineages of *N. meningitidis* have been recognized and are responsible for the majority of cases of invasive meningococcal disease worldwide. Hyperinvasive lineages may be associated with more than one capsular group. The apparent genetic stability of these meningococcal clones over decades and during wide geographic spread indicates that they are well adapted to the nasopharyngeal environment of the host and to efficient transmission. While MLST has become established as the main method for meningococcal genotyping in many reference laboratories over the past 15 years, whole-genome sequencing is gradually replacing this approach, with almost 3153 genomes already available in the United Kingdom's national library (*www.meningitis.org/genome-library*).

The group B meningococcal genome is >2 megabases in length and contains 2158 coding regions. Many genes undergo phase variation that makes it possible to control their expression; this capacity is likely to be important in meningococcal adaptation to the host environment and evasion of the immune response. Meningococci can obtain DNA from their environment and can acquire new genes—including the capsular operon—such that *capsule switching* from one capsular group to another can occur.

EPIDEMIOLOGY

Patterns of Disease


Up to 500,000 cases of meningococcal disease are thought to occur worldwide each year, though the numbers have been declining recently as a result of both immunization programs and secular trends. About 10% of the individuals affected die. There are several patterns of disease: epidemic, outbreak (small clusters of cases), hyperendemic, and sporadic or endemic.

Epidemics have continued since the original descriptions of meningococcal disease, especially affecting the sub-Saharan meningitis belt of Africa, where tens to hundreds of thousands of cases (caused mainly by capsular group A but also by capsular groups W and X) may be reported over a season and rates may be as high as 1000 cases per 100,000 population. Capsular group A epidemics took place in Europe and North America after the First and Second World Wars, and capsular group A outbreaks have been documented over the past 30 years in New Zealand, China, Nepal, Mongolia, India, Pakistan, Poland, and Russia. However, 81% of cases reported in the meningitis belt in 2014 were caused by capsular group W meningococci and followed an immunization campaign to control capsular group A outbreaks.

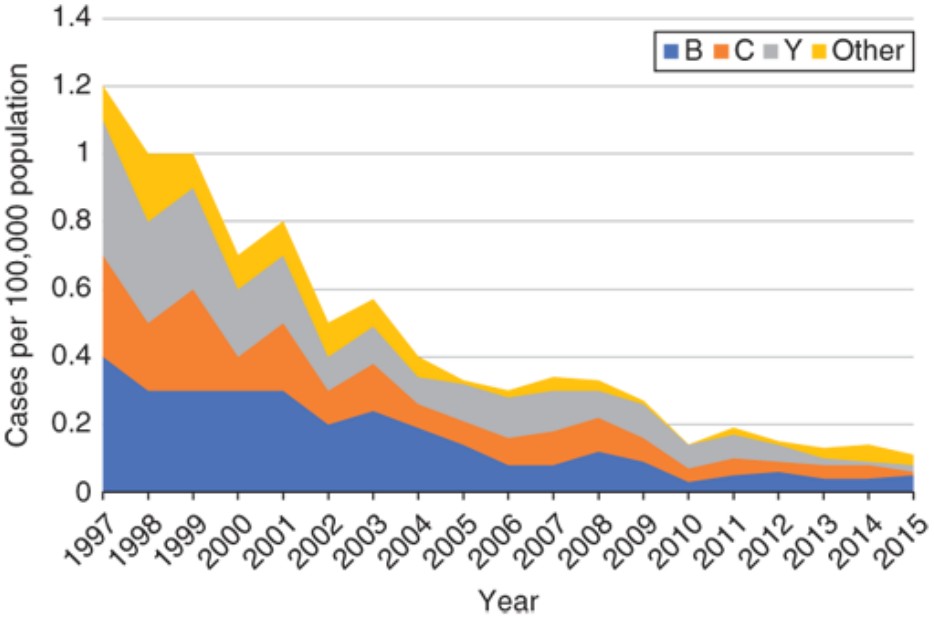
Clusters of cases occur where there is an opportunity for increased transmission—i.e., in closed or semi-closed communities such as schools, colleges, universities, military training centers, and refugee camps. Recently, such clusters have been especially strongly linked with a particular clone (sequence type 11) that is mainly associated with capsular group C or W but was first described in association with capsular group B. Clusters of capsular group W disease associated with the Hajj pilgrimage in 2000/2001 led to a requirement for vaccination against meningococcal disease for travel to Saudi Arabia. Wider and more prolonged community outbreaks (hyperendemic disease) due to single clones of capsular group B meningococci account for ≥10 cases per 100,000. Regions affected in the past decade include the U.S. Pacific Northwest, New Zealand (both islands), and the province of Normandy in France.

Most countries now experience predominantly sporadic cases (0.3–5 cases per 100,000 population), with many different disease-causing clones involved and usually no clear epidemiologic link between one case and another. The disease

rate and the distribution of meningococcal strains vary in different regions of the world and also in any one location over time. For example, in the United States, the rate of meningococcal disease fell from 1.2 cases per 100,000 population in 1997 to <0.15 case per 100,000 in 2012 **(Fig. 150-3)**. Meningococcal disease in the United States was previously dominated by capsular groups B and C; however, capsular group Y emerged during the 1990s and became more common than capsular group C in 2007 **(Fig. 150-4)**. In contrast, rates of disease in England and Wales rose to >5 cases per 100,000 during the 1990s because of an increase in cases caused by the ST11 capsular group C clone. As a result of a mass immunization program against capsular group C in 1999, the majority of cases in the United Kingdom are now attributed to capsular group B (Fig. 150-4). Over the last decade, most industrialized nations have seen a general decrease in meningococcal disease; this decrease is linked to immunization against capsular group C meningococci in Europe, Canada, and Australia and to adolescent immunization programs for capsular groups A, C, Y, and W in the United States. However, other factors, including changes in population immunity and prevalent clones of meningococci (factors that, in combination, probably explain the cyclic nature of meningococcal disease rates) as well as a reduction in smoking and passive exposure to tobacco smoke (driven by bans on smoking in buildings and public spaces) across wealthy countries, are likely to have contributed to the fall in cases. Recently, a hyperinvasive ST11 clone bearing a W capsule has emerged in South America and spread to the United Kingdom and has also emerged in other countries in Europe and in Australia, leading to a considerable increase in capsular group W cases. Increases in capsular group Y disease have also been noted in various countries in Europe, Canada, and South Africa.

FIGURE 150-3

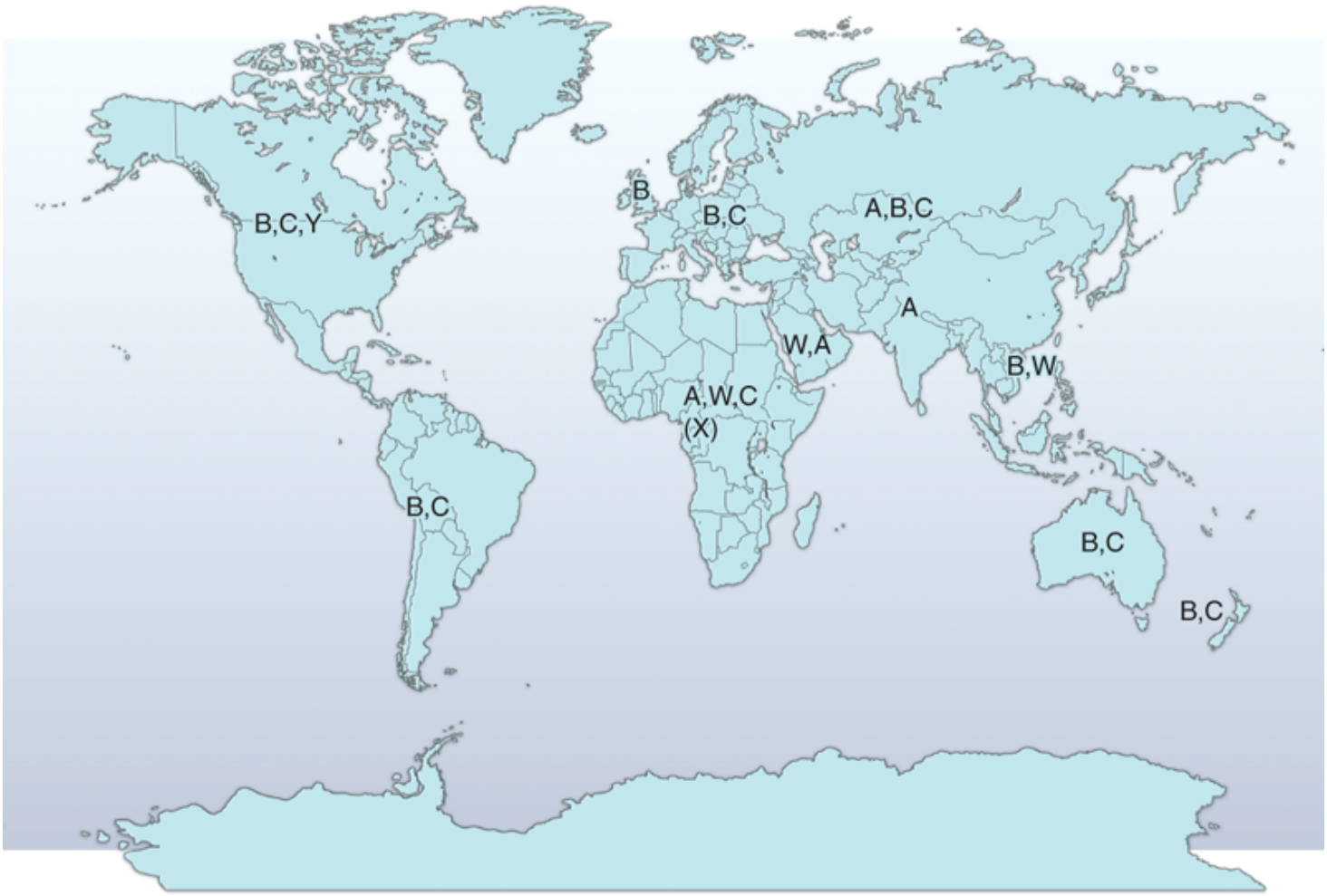
Meningococcal disease in the United States over time. ABCs, active bacterial cores. *(Adapted from ABC Surveillance data, Centers for Disease Control and Prevention; www.cdc.gov.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 150-4

Global distribution of meningococcal capsular groups, 1999–2009.



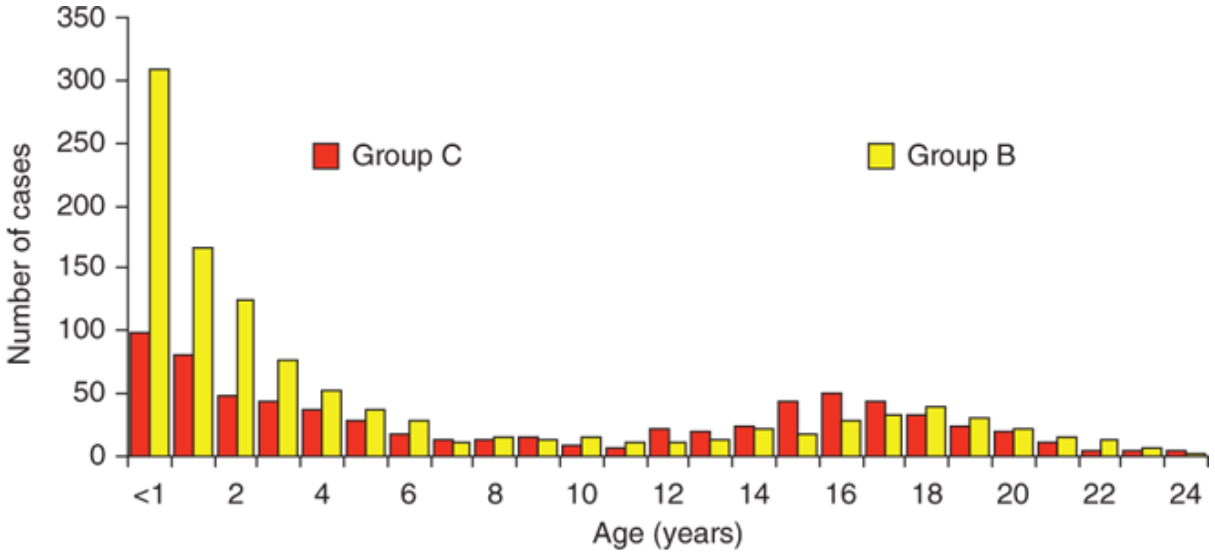
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Factors Associated with Disease Risk and Susceptibility

The principal determinant of disease susceptibility is age, with the peak incidence in the first year of life **(Fig. 150-5)**. The susceptibility of the very young presumably results from an absence of specific adaptive immunity in combination with very close contact with colonized individuals, including parents. Compared with other age groups, infants appear to be particularly susceptible to capsular group B disease: >30% of capsular group B cases in the United States occur during the first year of life. In the early 1990s in North America, the median ages for patients with disease due to capsular groups B, C, Y, and W were 6, 17, 24, and 33 years, respectively.

FIGURE 150-5

Age distribution of capsular groups B and C meningococcal disease in England and Wales, 1998–1999. *(Health Protection Agency, UK; www.hpa.org.uk.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

After early childhood, a second peak of disease occurs among adolescents and young adults (15–25 years of age) in Europe and North America. It is thought that this peak relates to social behaviors and environmental exposures in this age group, as discussed below. Most cases of infection with *N. meningitidis* in developed countries today are sporadic, and the rarity of the disease suggests that individual susceptibility may be important. A number of factors probably contribute to individual susceptibility, including the host's genetic constitution, environment, and contact with a carrier or a case.

The best-documented genetic association with meningococcal disease is complement deficiency, chiefly of the terminal complement components (C5–9), properdin, or factor D; such a deficiency increases the risk of disease by up to 600-fold and may result in recurrent attacks. Complement components are believed to be important for the bactericidal activity of serum, which is considered the principal mechanism of immunity against invasive meningococcal disease. However, when investigated, complement deficiency is found in only a very small proportion of individuals with meningococcal disease (0.3%). Conversely, 7–20% of persons whose disease is caused by the less common capsular groups (W, X, Y, Z, 29E) have a complement deficiency. Complement deficiency appears to be associated with capsular group B disease only rarely. Individuals with recurrences of meningococcal disease, particularly those caused by non-B capsular groups, should be assessed for complement deficiency by measurement of total hemolytic complement activity. There is also limited evidence that hyposplenism (through reduction in phagocytic capacity) and hypogammaglobulinemia (through absence of specific antibody) increase the risk of meningococcal disease. Genetic studies have revealed various associations with disease susceptibility, including complement and mannose-binding lectin deficiency, single-nucleotide polymorphisms in Toll-like receptor (TLR) 4 and complement factor H, and variants of Fc gamma receptors.

Factors that increase the chance of a susceptible individual's acquiring *N. meningitidis* via the respiratory route also increase the risk of meningococcal disease. Acquisition occurs through close contact with carriers as a result of overcrowding (e.g., in poor socioeconomic settings, in refugee camps, during the Hajj pilgrimage to Mecca, and during freshman-year residence in college dormitories) and certain social behaviors (e.g., attendance at bars and nightclubs, kissing). Secondary cases may occur in close contacts of an index case (e.g., household members and persons kissing the infected individual); the risk to these contacts may be as high as 1000 times the background rate in the population. Factors that damage the nasopharyngeal epithelium also increase the risk of both colonization with *N. meningitidis* and invasive disease. The most important of these factors are cigarette smoking (odds ratio, 4.1) and passive exposure to cigarette smoke. In addition, recent viral respiratory tract infection, infection with *Mycoplasma* species, and winter or

the dry season have been associated with meningococcal disease; all of these factors presumably either increase the expression of adhesion molecules in the nasopharynx, thus enhancing meningococcal adhesion, or facilitate meningococcal invasion of the bloodstream.

PATHOGENESIS

N. meningitidis has evolved as an effective colonizer of the human nasopharynx, with asymptomatic infection rates of >25% described in some series of adolescents and young adults and among residents of crowded communities. Point-prevalence studies reveal widely divergent rates of carriage for different types of meningococci. This variation suggests that some types may be adapted to a short duration of carriage with frequent transmission to maintain the population, while others may be less efficiently transmitted but may overcome this disadvantage by colonizing for a long period. Despite the high rates of carriage among adolescents and young adults, only ~10% of adults carry meningococci, and colonization is very rare in early childhood. Many of the same factors that increase the risk of meningococcal disease also increase the risk of carriage. Colonization of the nasopharynx involves a series of interactions of meningococcal adhesins (e.g., Opa proteins and pili) with their ligands on the epithelial mucosa. *N. meningitidis* produces an IgA1 protease that is likely to reduce interruption of colonization by mucosal IgA.

Colonization should be considered the normal state of meningococcal infection, with an increased risk of invasion being the unfortunate consequence (for both host and organism) of adaptations of hyperinvasive meningococcal lineages. The meningococcal capsule is an important virulence factor: acapsular strains only very rarely cause invasive disease. The capsule provides resistance to phagocytosis and may be important in preventing desiccation during transmission between hosts. Antigenic diversity in surface structures and an ability to vary levels of their expression have probably evolved as important factors in maintaining meningococcal populations within and between individual hosts.

Invasion through the mucosa into the bloodstream occurs rarely, usually within a few days of acquisition of an invasive strain by a susceptible individual. Only occasional cases of prolonged colonization prior to invasion have been documented. Once the organism is in the bloodstream, its growth may be limited if the individual is partially immune, although bacteremia may allow seeding of another site, such as the meninges or the joints. Alternatively, unchecked proliferation may continue, resulting in high bacterial counts in the circulation. During growth, meningococci release blebs of outer membrane (Fig. 150-1) containing outer-membrane proteins and LPS. Endotoxin binds cell-bound CD14 in association with TLR4 to initiate an inflammatory cascade with the release of high levels of various mediators, including tumor necrosis factor (TNF) α, soluble TNF receptor, interleukin (IL) 1, IL-1 receptor antagonist, IL-1β, IL-6, IL-8, IL-10, plasminogen-activator inhibitor 1 (PAI-1), and leukemia inhibitory factor. Soluble CD14-bound endotoxin acts as a mediator of endothelial activation. The severity of meningococcal disease is related both to the levels of endotoxin in the blood and to the magnitude of the inflammatory response. The latter is determined to some extent by polymorphisms in the inflammatory response genes (and their inhibitors), and the release of the inflammatory cascade heralds the development of meningococcal septicemia (meningococcemia). Endothelial injury is central to many clinical features of meningococcemia, including increased vascular permeability, pathologic changes in vascular tone, loss of thromboresistance, intravascular coagulation, and myocardial dysfunction. Endothelial injury leads to increased vascular permeability (attributed to loss of glycosaminoglycans and endothelial proteins), with subsequent gross proteinuria. Leakage of fluid and electrolytes into the tissues from capillaries (“capillary leak syndrome”) leads to hypovolemia, tissue edema, and pulmonary edema. Initial compensation results in vasoconstriction and tachycardia, although cardiac output eventually falls. While resuscitation fluids may restore circulating volume, tissue edema will continue to increase, and, in the lung, the consequence may be respiratory failure.

Intravascular thrombosis (caused by activation of procoagulant pathways in association with upregulation of tissue factor on the endothelium) occurs in some patients with meningococcal disease and results in purpura fulminans and infarction of areas of skin or even of whole limbs. At the same time, multiple anticoagulant pathways are downregulated through loss of endothelial thrombomodulin and protein C receptors and decreases in levels of antithrombin III, protein C, protein S, and tissue factor pathway inhibitor. Thrombolysis is also profoundly impaired in meningococcal sepsis through the release of high levels of PAI-1.

Shock in meningococcal septicemia appears to be attributable to a combination of factors, including hypovolemia, which results from the capillary leak syndrome secondary to endothelial injury, and myocardial depression, which is driven by hypovolemia, hypoxia, metabolic derangements (e.g., hypocalcemia), and cytokines (e.g., IL-6). Decreased perfusion of tissues as a result of intravascular thrombosis, vasoconstriction, tissue edema, and reduced cardiac output in meningococcal septicemia can cause widespread organ dysfunction, including renal impairment and—later in the disease—a decreased level of consciousness due to central nervous system involvement.

Bacteria that reach the meninges cause a local inflammatory response—with release of a spectrum of cytokines similar to that seen in septicemia—that presents clinically as meningitis and is thought to determine the severity of neuronal injury. Local endothelial injury may result in cerebral edema and rapid onset of raised intracranial pressure in some cases.

CLINICAL MANIFESTATIONS

As discussed above, the most common form of infection with *N. meningitidis* is asymptomatic carriage of the organism in the nasopharynx. Despite the location of infection in the upper airway, meningococcal pharyngitis is rarely reported; however, upper respiratory tract symptoms are common prior to presentation with invasive disease. It is not clear whether these symptoms relate to preceding viral infection (which may promote meningococcal acquisition) or to meningococcal acquisition itself. After acquiring the organism, susceptible individuals develop disease manifestations in 1–10 days (usually <4 days, although colonization for 11 weeks has been documented).

Along the spectrum of presentations of meningococcal disease, the most common clinical syndromes are meningitis and meningococcal septicemia. In fulminant cases, death may occur within hours of the first symptoms. Occult bacteremia is also recognized and, if untreated, progresses in two-thirds of cases to focal infection, including meningitis or septicemia. Meningococcal disease may also present as pneumonia, pyogenic arthritis or osteomyelitis, purulent pericarditis, endophthalmitis, conjunctivitis, primary peritonitis, or (rarely) urethritis. Perhaps because it is difficult to diagnose, pneumococcal pneumonia is not commonly reported but is associated with capsular groups Y, W, and Z and appears most often to affect individuals >10 years of age.

Rash

A nonblanching rash (petechial or purpuric) develops in >80% of cases of meningococcal disease; however, the rash is often absent early in the illness. Usually initially blanching in nature (macules, maculopapules, or urticaria) and indistinguishable from more common viral rashes, the rash of meningococcal infection becomes petechial or frankly purpuric over the hours after onset. In the most severe cases, large purpuric lesions develop (purpura fulminans; **Fig. A1-41**). Some patients (including those with overwhelming sepsis) may have no rash. While petechial rash and fever are important signs of meningococcal disease, fewer than 10% of children (and, in some clinical settings, fewer than 1% of patients) with this presentation are found to have meningococcal disease. Most patients presenting with a petechial or

purpuric rash have a viral infection (Table 150-2). The skin lesions exhibit widespread endothelial necrosis and occlusion of small vessels in the dermis and subcutaneous tissues, with a neutrophilic infiltrate.

TABLE 150-2

Common Causes of Petechial or Purpuric Rashes

Enteroviruses
Influenza and other respiratory viruses
Measles virus
Epstein-Barr virus
Cytomegalovirus
Parvovirus
Deficiency of protein C or S (including postvaricella protein S deficiency)
Platelet disorders (e.g., idiopathic thrombocytopenic purpura, drug effects, bone marrow infiltration)
Henoch-Schönlein purpura, connective tissue disorders, trauma (including nonaccidental injuries in children)
Pneumococcal, streptococcal, staphylococcal, or gram-negative bacterial sepsis

Meningitis

Meningococcal meningitis commonly presents as nonspecific manifestations, including fever, vomiting, and (especially in infants and young children) irritability, and is indistinguishable from other forms of bacterial meningitis unless there is an associated petechial or purpuric rash, which occurs in two-thirds of cases. Headache is rarely reported in early childhood but is more common in later childhood and adulthood. When headache is present, the following features, in association with fever or a history of fever, are suggestive of bacterial meningitis: neck stiffness, photophobia, decreased level of consciousness, seizures or status epilepticus, and focal neurologic signs. Classic signs of meningitis, such as neck stiffness and photophobia, are often absent in infants and young children with bacterial meningitis, who more usually present with fever and irritability and may have a bulging fontanelle.

While 30–50% of patients present with a meningitis syndrome alone, up to 40% of meningitis patients also present with some features of septicemia. Most deaths from meningococcal meningitis alone (i.e., without septicemia) are associated with raised intracranial pressure presenting as a reduced level of consciousness, relative bradycardia and hypertension, focal neurologic signs, abnormal posturing, and signs of brainstem involvement—e.g., unequal, dilated, or poorly reactive pupils; abnormal eye movement; and impaired corneal responses (Chap. 300).

Septicemia

Meningococcal septicemia alone accounts for up to 20% of cases of meningococcal disease. The condition may progress from early nonspecific symptoms to death within hours. Mortality rates among children with this syndrome have been high (25–40%), but early aggressive management (as discussed below) may reduce the figure to <10%. Early symptoms are nonspecific and suggest an influenza-like illness with fever, headache, and myalgia accompanied by vomiting and abdominal pain. As discussed above, the rash, if present, may appear to be viral early in the course until petechiae or purpuric lesions develop. Purpura fulminans occurs in severe cases **(Fig. A1-41)**, with multiple large purpuric lesions and signs of peripheral ischemia. Surveys of patients have indicated that limb pain, pallor (including a mottled appearance and cyanosis), and cold hands and feet may be prominent. Shock is manifested by tachycardia, poor peripheral perfusion, tachypnea, and oliguria. Decreased cerebral perfusion leads to confusion, agitation, or decreased level of consciousness. With progressive shock, multiorgan failure ensues; hypotension is a late sign in children, who more commonly present with compensated shock (tachycardia, poor peripheral perfusion, and normal blood pressure). Poor outcome is associated with an absence of meningism, hypotension, young age, coma, relatively low temperature (<38°C), leukopenia, and thrombocytopenia. Spontaneous hemorrhage (pulmonary, gastric, or cerebral) may result from consumption of coagulation factors and thrombocytopenia.

Chronic Meningococcemia

Chronic meningococcemia, which is rarely recognized, presents as repeated episodes of petechial rash **(Fig. A1-42)** associated with fever, joint pain, features of arthritis, and splenomegaly that may progress to acute meningococcal septicemia if untreated. During the relapsing course, bacteremia characteristically clears without treatment and then recurs. The differential diagnosis includes bacterial endocarditis, acute rheumatic fever, Henoch-Schönlein purpura, infectious mononucleosis, disseminated gonococcal infection, and immune-mediated vasculitis. This condition has been associated with complement deficiencies in some cases and with inadequate sulfonamide therapy in others.

 A study from the Netherlands found that half of isolates from patients with chronic meningococcemia had an underacylated lipid A (part of the surface LPS molecule) due to an *lpxL1* gene mutation, which markedly reduces the inflammatory response to endotoxin.

Postmeningococcal Reactive Disease

In a small proportion of patients, an immune complex disease develops ~4–10 days after the onset of meningococcal disease, with manifestations that include a maculopapular or vasculitic rash (2% of cases), arthritis (up to 8% of cases), iritis (1%), pericarditis, and/or polyserositis associated with fever. The immune complexes involve meningococcal polysaccharide antigen and result in immunoglobulin and complement deposition with an inflammatory infiltrate. These features resolve spontaneously without sequelae. It is important to recognize this condition since a new onset of fever and rash can lead to concerns about relapse of meningococcal disease and unnecessarily prolonged antibiotic treatment.

DIAGNOSIS

Like other invasive bacterial infections, meningococcal disease may produce elevations of the white blood cell (WBC) count and of values for inflammatory markers (e.g., C-reactive protein and procalcitonin levels or the erythrocyte sedimentation rate). Values may be normal or low in rapidly progressive disease, and a lack of rise in these laboratory test values does not exclude the diagnosis. However, in the presence of fever and a petechial rash, these elevations are

suggestive of meningococcal disease. In patients with severe meningococcal septicemia, common laboratory findings include hypoglycemia, acidosis, hypokalemia, hypocalcemia, hypomagnesemia, hypophosphatemia, anemia, and coagulopathy.

Although meningococcal disease is often diagnosed on clinical grounds, in suspected meningococcal meningitis or meningococcemia, blood should routinely be sent for culture to confirm the diagnosis and to facilitate public health investigations; blood cultures are positive in up to 75% of cases. Culture media containing sodium polyanethol sulfonate, which may inhibit meningococcal growth, should be avoided. Meningococcal viability is reduced if there is a delay in transport of the specimen to the microbiology laboratory for culture or in plating of cerebrospinal fluid (CSF) samples. In countries where treatment with antibiotics before hospitalization is recommended for meningococcal disease, the majority of clinically suspected cases are culture negative. Real-time polymerase chain reaction (PCR) analysis of whole-blood samples increases the diagnostic yield by >40%, and results obtained with this method may remain positive for several days after administration of antibiotics. Indeed, in the United Kingdom, more than half of clinically suspected cases are currently identified by PCR.

Unless contraindications exist (raised intracranial pressure, uncorrected shock, disordered coagulation, thrombocytopenia, respiratory insufficiency, local infection, ongoing convulsions), lumbar puncture should be undertaken to identify and confirm the etiology of suspected meningococcal meningitis, whose presentation cannot be distinguished from that of meningitis of other bacterial causes. Some authorities have recommended a CT brain scan prior to lumbar puncture because of the risk of cerebral herniation in patients with raised intracranial pressure. However, a normal CT scan is not uncommon in the presence of raised intracranial pressure in meningococcal meningitis, and the decision to perform a lumbar puncture should be made on clinical grounds. CSF features of meningococcal meningitis (elevated protein level and WBC count, decreased glucose level) are indistinguishable from those of other types of bacterial meningitis unless a gram-negative diplococcus is identified. (Gram's staining is up to 80% sensitive for meningococcal meningitis.) CSF should be submitted for culture (sensitivity, 90%) and (where available) PCR analysis. CSF antigen testing with latex agglutination is insensitive and should be replaced by molecular diagnosis when possible.

Lumbar puncture should generally be avoided in meningococcal septicemia, as positioning for the procedure may critically compromise the patient's circulation in the context of hypovolemic shock. Delayed lumbar puncture may still be useful when the diagnosis is uncertain, particularly if molecular diagnostic technology is available.

In other types of focal infection, culture and PCR analysis of normally sterile body fluids (e.g., synovial fluid) may aid in the diagnosis. Although some authorities have recommended cultures of scrapings or aspirates from skin lesions, this procedure adds little to the diagnostic yield when compared with a combination of blood culture and PCR analysis. Urinary antigen testing also is insensitive, and serologic testing for meningococcal infection has not been adequately studied. Because *N. meningitidis* is a component of the normal human nasopharyngeal flora, identification of the organism on throat swabs has limited diagnostic value, but strains identified in the nasopharynx in the context of a probable case are likely to be those responsible for disease.

TREATMENT

TREATMENT

Meningococcal Infections

Death from meningococcal disease is associated most commonly with hypovolemic shock (meningococcemia) and occasionally with raised intracranial pressure (meningococcal meningitis). Therefore, management should focus on the treatment of these urgent clinical issues in addition to the administration of specific antibiotic therapy. Delayed recognition of meningococcal disease or its associated physiologic derangements, together with inadequate emergency management, is associated with poor outcome. Since the disease is rare, protocols for emergency management have been developed (see *www.meningitis.org*).

Airway patency may be compromised if the level of consciousness is depressed as a result of shock (impaired cerebral perfusion) or raised intracranial pressure; this situation may require intervention. In meningococcemia, pulmonary edema and pulmonary oligemia (presenting as hypoxia) require oxygen therapy or elective endotracheal intubation. In cases with shock, aggressive fluid resuscitation (with replacement of the circulating volume several times in severe cases) and inotropic support may be necessary to maintain cardiac output. If shock persists after volume resuscitation at 40 mL/kg, the risk of pulmonary edema is high, and elective intubation is recommended to improve oxygenation and decrease the work of breathing. Metabolic derangements, including hypoglycemia, acidosis, hypokalemia, hypocalcemia, hypomagnesemia, hypophosphatemia, anemia, and coagulopathy, should be anticipated and corrected. In the presence of raised intracranial pressure, management includes correction of coexistent shock and neurointensive care to maintain cerebral perfusion.

Empirical antibiotic therapy for suspected meningococcal disease consists of a third-generation cephalosporin such as ceftriaxone (75–100 mg/kg per day [maximum, 4 g/d] in one or two divided IV doses) or cefotaxime (200 mg/kg per day [maximum, 8 g/d] in four divided IV doses) to cover the various other (potentially penicillin-resistant) bacteria that may produce an indistinguishable clinical syndrome. Although unusual in most isolates, reduced meningococcal sensitivity to penicillin (a minimal inhibitory concentration of 0.12–1.0 µg/mL) has been reported widely.

Both meningococcal meningitis and meningococcal septicemia are conventionally treated for 7 days, although courses of 3–5 days may be equally effective. Furthermore, a single dose of ceftriaxone or an oily suspension of chloramphenicol has been used successfully in resource-poor settings. No data are available to guide the duration of treatment for meningococcal infection at other foci (e.g., pneumonia, arthritis); antimicrobial therapy is usually continued until clinical and laboratory evidence of infection has resolved. Cultures usually become sterile within 24 h of initiation of appropriate antibiotic chemotherapy.

The use of glucocorticoids for adjunctive treatment of meningococcal meningitis remains controversial since no relevant studies have had sufficient power to determine true efficacy. One large study in adults did indicate a trend toward benefit, and in clinical practice a decision to use glucocorticoids usually must precede a definite microbiologic diagnosis. Therapeutic doses of glucocorticoids are not recommended in meningococcal septicemia, but many intensivists recommend replacement glucocorticoid doses for patients who have refractory shock in association with impaired adrenal gland responsiveness.

Various other adjunctive therapies for meningococcal disease have been considered, but few have been subjected to clinical trials and none can currently be recommended. An antibody to LPS (HA1A) failed to confer a demonstrable benefit. Recombinant bactericidal/permeability-increasing protein (which is not currently available) was tested in a study that had inadequate power to show an effect on mortality rates; however, there were trends toward lower mortality rates among patients who received a complete infusion, and this group also had fewer amputations, fewer blood-product transfusions, and a significantly improved functional outcome. Given that protein C concentrations are reduced in meningococcal disease, the use of activated protein C has been considered. A survival benefit was

demonstrated in adult sepsis trials; however, trials in pediatric sepsis (of particular relevance for meningococcal disease) found no benefit and indicated a potential risk of bleeding complications with use of activated protein C.

The postmeningococcal immune-complex inflammatory syndrome has been treated with nonsteroidal anti-inflammatory agents until spontaneous resolution occurs.

COMPLICATIONS

About 10% of patients with meningococcal disease die despite the availability of antimicrobial therapy and other intensive medical interventions. The most common complication of meningococcal disease (10% of cases) is scarring after necrosis of purpuric skin lesions, for which skin grafting may be necessary. The lower limbs are most often affected; next in frequency are the upper limbs, the trunk, and the face. On average, 13% of the skin surface area is involved. Amputations are necessary in 1–2% of survivors of meningococcal disease because of a loss of tissue viability after peripheral ischemia or compartment syndromes. Unless there is local infection, amputation should usually be delayed to allow the demarcation between viable and nonviable tissue to become apparent. Approximately 5% of patients with meningococcal disease suffer hearing loss, and 7% have neurologic complications. In one study pain was reported by 21% of survivors, and in a recent analysis of capsular group B meningococcal disease (the MOSAIC study) as many as one-quarter of survivors had psychological disorders. In some investigations, the rate of complications is higher for capsular group C disease (mostly associated with the ST11 clone) than for capsular group B disease. In patients with severe hypovolemic shock, renal perfusion may be impaired and prerenal failure is common, but permanent renal replacement therapy is rarely needed.

Several studies suggest adverse psychosocial outcomes after meningococcal disease, with reduced quality of life, lowered self-esteem, and poorer neurologic development, including increased rates of attention deficit/hyperactivity disorder and special educational needs. Other studies have not found evidence of such outcomes.

PROGNOSIS

Several prognostic scoring systems have been developed to identify patients with meningococcal disease who are least likely to survive. Factors associated with a poorer prognosis are shock; young age (infancy), old age, and adolescence; coma; purpura fulminans; disseminated intravascular coagulation; thrombocytopenia; leukopenia; absence of meningitis; metabolic acidosis; low plasma concentrations of antithrombin and proteins S and C; high blood levels of PAI-1; and a low erythrocyte sedimentation rate or C-reactive protein level. The Glasgow Meningococcal Septicaemia Prognostic Score (GMSPS) performs well and may be clinically useful for severity assessment in meningococcal disease. However, scoring systems do not direct the clinician to specific interventions, and the priority in management should be recognition of compromised airways, breathing, or circulation and direct, urgent intervention. Most patients improve rapidly with appropriate antibiotics and supportive therapy. Fulminant meningococcemia is more likely to result in death or ischemic skin loss than is meningitis; optimal emergency management may reduce mortality rates among the most severely affected patients.

PREVENTION

Since mortality rates in meningococcal disease remain high despite improvements in intensive care management, immunization is the only rational approach to prevention at a population level. Secondary cases are common among

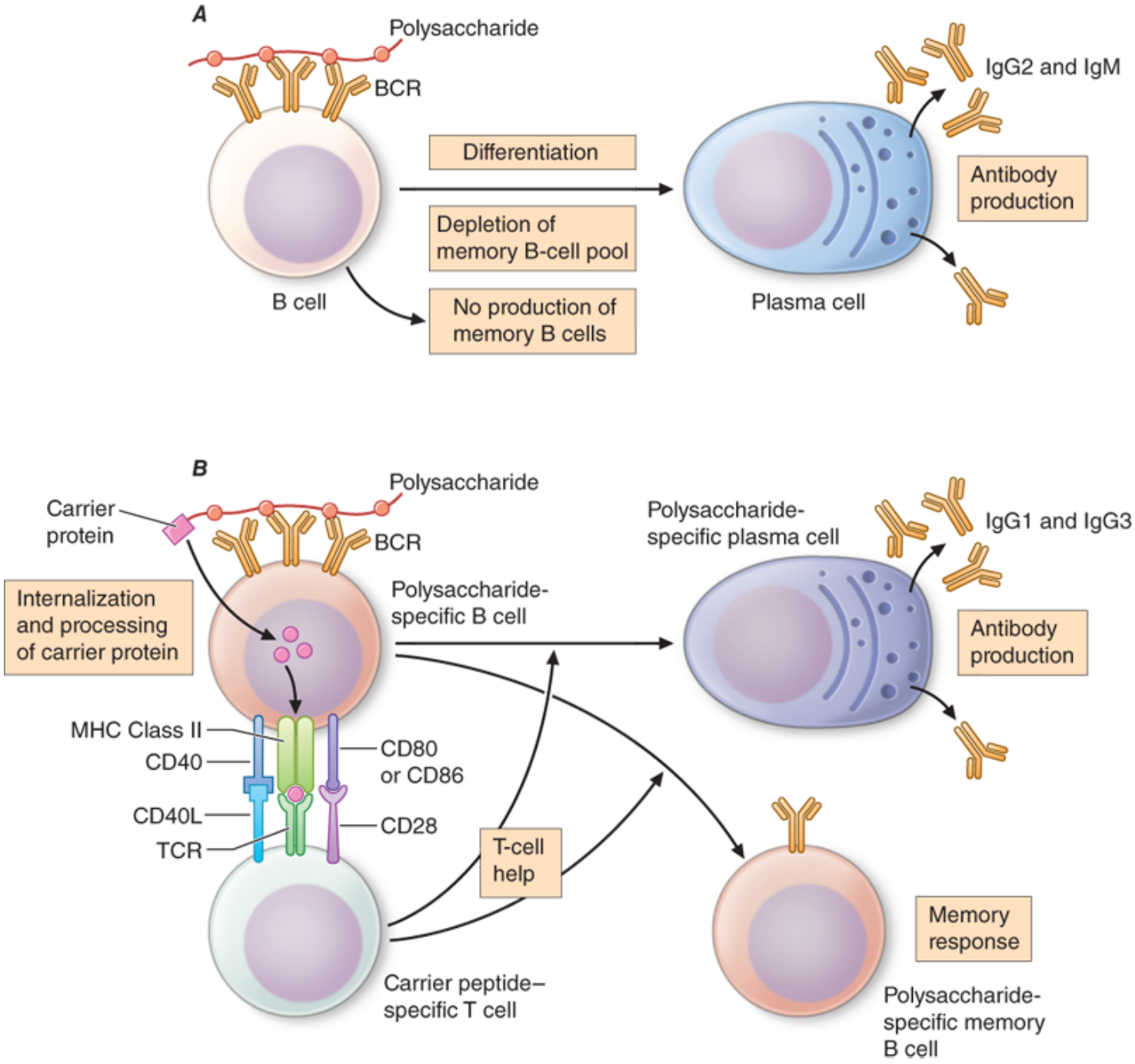
household and “kissing” contacts of cases, and secondary prophylaxis with antibiotics is widely recommended for these contacts (see below).

Polysaccharide Vaccines

Purified meningococcal capsular polysaccharide has been used for immunization since the 1960s. Meningococcal polysaccharide vaccines are currently formulated as either bivalent (capsular groups A and C) or quadrivalent (capsular groups A, C, Y, and W), with 50 μg of each polysaccharide per dose. Local reactions (erythema, induration, and tenderness) may occur in up to 40% of vaccinees, but serious adverse events (including febrile convulsions in young children) are very rarely reported. In adults, the vaccines are immunogenic, but immunity appears to be relatively short-lived (with antibody levels above baseline for only 2–10 years), and booster doses do not induce a further rise in antibody concentration. Indeed, a state of immunologic hyporesponsiveness has been widely reported to follow booster doses of plain polysaccharide vaccines. The repeating units of these vaccines cross-link B cell receptors to drive specific memory B cells to become plasma cells and produce antibody. Because meningococcal polysaccharides are T cell–independent antigens, no memory B cells are produced after immunization, and the memory B cell pool is depleted such that fewer polysaccharide-specific cells are available to respond to a subsequent dose of vaccine **(Fig. 150-6)**. The clinical relevance of hyporesponsiveness is unknown. Plain polysaccharide vaccines generally are not immunogenic in early childhood, possibly because marginal-zone B cells are involved in polysaccharide responses and maturation of the splenic marginal zone is not complete until 18 months to 2 years of age. The efficacy of the meningococcal capsular group C component is >90% in young adults; no efficacy data are available for the capsular group Y and W polysaccharides in this age group.

FIGURE 150-6

A. Polysaccharides from the encapsulated bacteria that cause disease in early childhood stimulate B cells by cross-linking the B cell receptor (BCR) and driving the production of immunoglobulins. There is no production of memory B cells, and the B cell pool may be depleted by this process such that subsequent immune responses are decreased. ***B.*** The carrier protein from protein–polysaccharide conjugate vaccines is processed by the polysaccharide-specific B cell, and peptides are presented to carrier peptide–specific T cells, with the consequent production of both plasma cells and memory B cells. MHC, major histocompatibility complex; TCR, T cell receptor. *(Reprinted from AJ Pollard et al: Nat Rev Immunol 9:213, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Group A meningococcal polysaccharides are exceptional in that they are effective in preventing disease at all ages. Two doses administered 2–3 months apart to children 3–18 months of age or a single dose administered to older children or adults has a protective efficacy rate of >95%. The vaccine was previously used widely in the control of outbreaks of meningococcal disease in the African meningitis belt. The duration of protection appears to be only 3–5 years. The plain polysaccharide vaccines have largely been superseded by protein–polysaccharide conjugate vaccines.

There is no meningococcal capsular group B plain polysaccharide vaccine because α-2,8-*N*-acetylneuraminic acid is expressed on the surface of neural cells in the fetus such that the B polysaccharide is perceived as "self" and therefore is not immunogenic in humans.

Conjugate Vaccines

The poor immunogenicity of plain polysaccharide vaccines in infancy has been overcome by chemical conjugation of the polysaccharides to a carrier protein (CRM_{197}, tetanus toxoid, or diphtheria toxoid). Conjugates that contain monovalent capsular group C polysaccharide and quadrivalent vaccines with A, C, Y, and W polysaccharides have been developed, as have vaccines including various other antigen combinations (e.g., tetanus conjugates with capsular group C and/or Y polysaccharide and *Haemophilus influenzae* type b polysaccharide). After immunization, peptides from the carrier protein are conventionally thought to be presented by polysaccharide-specific B cells to peptide-specific T cells in association with major histocompatibility complex (MHC) class II molecules. (Some recent data suggest that carrier protein peptide may actually be presented in association with an oligosaccharide and MHCII.) The result is a T cell–dependent immune response that allows production of antibody and generation of an expanded B cell memory pool. Unlike responses to booster doses of plain polysaccharides, responses to booster doses of conjugate vaccines have the characteristics of memory responses. Indeed, conjugate vaccines overcome the hyporesponsiveness induced by plain polysaccharides by replenishing the memory pool. The reactogenicity of conjugate vaccines is similar to that of plain polysaccharide vaccines.


The first widespread use of capsular group C meningococcal conjugate vaccine (MenC) came in 1999 in the United Kingdom after a rise in capsular group C disease. A mass vaccination campaign involving all individuals <19 years of age was undertaken, and the number of laboratory-confirmed capsular group C cases fell from 955 in 1998–1999 to just 29 in 2011–2012. The effectiveness of the immunization program was attributed both to direct protection of immunized persons and to reduced transmission of the organism in the population as a result of decreased rates of colonization among the immunized (i.e., herd immunity). Data on immunogenicity and effectiveness have shown that the duration of protection is short when the vaccine is administered in early childhood; thus booster doses are needed to maintain population immunity. In contrast, immunity after a dose of vaccine given in adolescence appears to be more prolonged.

In 2005, the first quadrivalent conjugate meningococcal vaccine containing A, C, Y, and W polysaccharides conjugated to diphtheria toxoid was initially recommended for all children >11 years of age in the United States and for persons 2–55 years of age in Canada. Such vaccines are now recommended by the Advisory Committee on Immunization Practices (ACIP) for routine administration to individuals 11–18 years of age, with a booster dose 3 years later; only a single dose is given to persons >16 years of age. These vaccines are also recommended for high-risk persons from 2 months to 55 years of age (see *www.cdc.gov/mmwr/preview/mmwrhtml/mm6324a2.htm*).

Uptake was slow initially, but current U.S. data suggest an efficacy rate of 82% in the first year after vaccination, with waning to 59% at 3–6 years after vaccination. Limited early data from the U.S. Vaccine Adverse Events Reporting System indicated that there might be a short-term increase in the risk of Guillain-Barré syndrome after immunization with the diphtheria conjugate vaccine; however, further investigation has not confirmed this finding. Quadrivalent conjugate vaccines with tetanus or CRM_{197} as carrier protein are now available in many countries and are used for high-risk groups and in routine programs for toddlers and adolescents.

A monovalent capsular group A vaccine, manufactured in India, was licensed in 2010 and rolled out to countries in the sub-Saharan African meningitis belt in a mass immunization campaign. There is strong evidence that this vaccine has been highly effective in controlling epidemic meningococcal disease in the region, with some evidence of a >90% reduction in disease in vaccinated populations. However, disease caused by capsular groups C, X, and W persists, and new-generation vaccines with wider coverage are being developed.

Vaccines Based on Subcapsular Antigens

The lack of immunogenicity of the group B capsule has led to the development of vaccines based on subcapsular antigens. Various surface components have been studied in early-phase clinical trials. Outer-membrane vesicles (OMVs) containing outer-membrane proteins, phospholipid, and LPS can be extracted from cultures of *N. meningitidis* by detergent treatment (Fig. 150-7). OMVs prepared in this way were used in efficacy trials with a Norwegian outbreak strain and reduced the incidence of group B disease among 14- to 16-year-old schoolchildren by 53%. Similarly, OMV vaccines constructed from local outbreak strains in Cuba and New Zealand have had reported efficacy rates of >70%. These OMV vaccines appear to produce strain-specific immune responses, with only limited cross-protection, and are therefore best suited to clonal outbreaks (e.g., those in Cuba and New Zealand as well as others in Norway and the province of Normandy in France).

FIGURE 150-7

Illustration of meningococcal outer-membrane vesicle containing outer-membrane structures.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Several purified surface proteins have been evaluated in phase 1 clinical trials but have not yet been developed further because of antigenic variability or poor immunogenicity (e.g., transferrin-binding proteins, neisserial surface protein A). Other vaccine candidates have been identified since sequencing of the meningococcal genome. The combination vaccine 4CMenB, which includes the New Zealand OMV vaccine and three recombinant proteins (neisserial adhesin A, factor H–binding protein, and neisserial heparin-binding antigen), is immunogenic from infancy and has been licensed for use in the United States, Canada, Europe, and Australia. This vaccine has been used with apparent success in the control of several university outbreaks in the United States and in a community outbreak in an area of Quebec, Canada. 4CMenB vaccine has an acceptable safety profile, with fever prominent among infants and injection-site pain frequently

reported among older children and adults. The vaccine is also being used in many countries for immunization of high-risk groups. In September 2015, 4CMenB was recommended for routine use in the United Kingdom for all infants born from May 2015 onward; a preliminary analysis has found an effectiveness rate of 82.9% (95% confidence interval, 24.1–95.2) against all capsular group B strains among infants receiving two or more doses, with a 50% reduction in cases below the number anticipated from comparison with trends in other age groups. Because the disease is so rare, the cost-effectiveness of capsular group B vaccine in infant immunization programs, as assessed with conventional thresholds, is borderline. Since infants are not commonly colonized with capsular group B meningococci, any impact on the total population burden of carried organisms will be small. It is therefore unlikely that an infant immunization program will provide additional value through induction of herd immunity. Rates of capsular group B carriage are higher among teenagers and young adults. Studies estimating the potential effect of 4CMenB on carriage of capsular group B meningococci among adolescents indicate that there is likely to be some impact. However, because these studies lack power, it remains uncertain whether the vaccine would have the substantial and sustained herd effects in this age group that could support widespread routine administration.

An immunogenic vaccine based on two variants of the lipoprotein factor H–binding protein (fHbp2) has been developed for use in adolescents and is licensed in the United States and Europe. The vaccine is immunogenic against representative indicator strains, inducing fourfold rises in bactericidal antibody titer in 50–92% of individuals. fHbp2 has an acceptable safety profile, with pain at the injection site, fatigue, and headache commonly reported. This vaccine can be used with a range of vaccines routinely administered in adolescence, including Tdap (tetanus–diphtheria–acellular pertussis), human papillomavirus, and MenACWY vaccines. fHbp2 has been used to control outbreaks of meningococcal disease in educational institutions in the United States, but no formal studies of its effectiveness have yet been undertaken.

Both of the new capsular group B meningococcal vaccines are licensed for use in the United States for persons 10–25 years of age. In addition, ACIP recommends their administration to individuals at high risk of capsular group B disease, with 4CMenB administered as two doses (1–2 months apart) and fHbp2 as three doses on a 0/1/6-month schedule.

MANAGEMENT OF CONTACTS

Close (household and kissing) contacts of individuals with meningococcal disease are at increased risk for developing secondary disease (up to 1000 times the rate for the general population); a secondary case follows as many as 3% of sporadic cases. About one-fifth of secondary cases are actually co-primary cases—i.e., cases that occur soon after the primary case and in which transmission is presumed to have originated from the same third party. The rate of secondary cases is highest during the week after presentation of the index case. The risk falls rapidly but remains above baseline for up to 1 year after the index case; 30% of secondary cases occur in the first week, 20% in the second week, and most of the remainder over the next 6 weeks. In outbreaks of meningococcal disease, mass prophylaxis has been used; however, limited data support population intervention, and significant concerns have arisen about adverse events and the development of resistance. For these reasons, prophylaxis is usually restricted to (1) persons at greatest risk who are intimate and/or household contacts of the index case and (2) health care workers who have been directly exposed to respiratory secretions. In most cases, members of wider communities (e.g., at schools or colleges) are not offered prophylaxis.

The aim of prophylaxis is to eradicate colonization of close contacts with the strain that has caused invasive disease in the index case. Prophylaxis should be given to all contacts at the same time to avoid recolonization by meningococci transmitted from untreated contacts and should also be used as soon as possible to treat early disease in secondary

cases. If the index patient is treated with an antibiotic that does not reliably clear colonization (e.g., penicillin), he or she should be given a prophylactic agent at the end of treatment to prevent relapse or onward transmission. Although rifampin has been most widely used and studied, it is not the optimal agent because it fails to eradicate carriage in 15–20% of cases, rates of adverse events have been high, compliance is affected by the need for four doses, and emerging resistance has been reported. Ceftriaxone as a single IM or IV injection is highly (97%) effective in carriage eradication and can be used at all ages and in pregnancy. Reduced susceptibility of isolates to ceftriaxone has occasionally been reported. Ciprofloxacin or ofloxacin is preferred in some countries; these agents are highly effective and can be administered by mouth but are not recommended in pregnancy. Resistance to fluoroquinolones has been reported in some meningococci in North America, Europe, and Asia.

In documented capsular group A, B, C, Y, or W disease, contacts may be offered immunization (with either the MenACWY conjugate vaccine or the MenB vaccine, as appropriate) in addition to chemoprophylaxis to provide protection beyond the duration of antibiotic therapy. Mass vaccination has been used successfully to control disease during outbreaks in closed communities (educational and military establishments) as well as during epidemics in open communities.

FURTHER READING

Christensen H et al: Meningococcal carriage by age: A systematic review and meta-analysis. Lancet Infect Dis 10:853, 2010.

Cohn AC et al: Prevention and control of meningococcal disease: Recommendations of the Advisory Committee on Immunization Practices (ACIP). MMWR Recomm Rep 62(RR-2):1, 2013. [PubMed: 23515099]

Gossger N et al: Immunogenicity and tolerability of recombinant serogroup B meningococcal vaccine administered with or without routine infant vaccinations according to different immunization schedules: A randomized controlled trial. JAMA 307:573, 2012.

Jafri RZ et al: Global epidemiology of invasive meningococcal disease. Popul Health Metr 11:17, 2013.

Pollard AJ et al: Maintaining protection against invasive bacteria with protein–polysaccharide conjugate vaccines. Nat Rev Immunol 9:213, 2009.

Read RC et al: Effect of a quadrivalent meningococcal ACWY glycoconjugate or a serogroup B meningococcal vaccine on meningococcal carriage: An observer-blind, phase 3 randomised clinical trial. Lancet 384:2123, 2014.

Vieusseux M: Memoire sur le maladie qui a regne a Geneva au printemps de 1805. J Med Clin Pharm 11:163, 1805.

Chapter 151: Gonococcal Infections

Sanjay Ram; Peter A. Rice

DEFINITION

Gonorrhea is a sexually transmitted infection (STI) of epithelium and commonly manifests as cervicitis, urethritis, proctitis, and conjunctivitis. If untreated, infections at these sites can lead to local complications such as endometritis, salpingitis, tuboovarian abscess, bartholinitis, peritonitis, and perihepatitis in female patients; periurethritis and epididymitis in male patients; and ophthalmia neonatorum in newborns. Disseminated gonococcemia is an uncommon event whose manifestations include skin lesions, tenosynovitis, arthritis, and (in rare cases) endocarditis or meningitis.

MICROBIOLOGY

Neisseria gonorrhoeae is a gram-negative, nonmotile, non-spore-forming organism that grows singly and in pairs (i.e., as monococci and diplococci, respectively). Exclusively a human pathogen, the gonococcus contains, on average, three genome copies per coccal unit; this polyploidy permits a high level of antigenic variation and the survival of the organism in its host. Gonococci, like all other *Neisseria* species, are oxidase positive. They are distinguished from other neisseriae by their ability to grow on selective media and to use glucose but not maltose, sucrose, or lactose.

EPIDEMIOLOGY

 The incidence of gonorrhea had been declining steadily in the United States, but in 2016 there were ~450,000 newly reported cases—up 46% since 2011. With 80 million cases estimated by the World Health Organization to have occurred globally in 2014, gonorrhea remains a major public health problem worldwide, is a significant cause of morbidity in developing countries, and may play a role in enhancing transmission of HIV.

Gonorrhea predominantly affects young, nonwhite, unmarried, less educated members of urban populations. The number of reported cases probably represents half of the true number of cases—a discrepancy resulting from under-reporting, self-treatment, and nonspecific treatment without a laboratory-proven diagnosis. The number of reported new cases of gonorrhea in the United States rose from ~250,000 in the early 1960s to a high of 1.01 million in 1978. The recorded incidence of gonorrhea in modern times peaked in 1975, with 468 reported new cases per 100,000 population in the United States. This peak was

attributable to the interaction of several variables, including improved accuracy of diagnosis, changes in patterns of contraceptive use, and changes in sexual behavior. A decline in the overall incidence of gonorrhea in the United States over the past quarter-century may reflect increased condom use resulting from public health efforts to curtail HIV transmission. Nevertheless, in 2016, 146 new cases per 100,000 population were reported in this country, representing a 1-year increase of 19%; this figure is the highest among industrialized countries. Simultaneously, antibiotic resistance is increasing in the United States and other countries, prompting the U.S. Centers for Disease Control and Prevention (CDC) to name antibiotic-resistant *N. gonorrhoeae* as one of the three most urgent threats of its kind. At present, the attack rate in the United States is highest among 15- to 24-year-old women and 20- to 29-year-old men; more than 70% of all reported cases occur in these two groups. From the standpoint of ethnicity, rates are highest among African Americans and lowest among persons of Asian descent.

The incidence of gonorrhea is higher in developing countries than in industrialized nations. The exact incidence of any STI is difficult to ascertain in developing countries because of limited surveillance and variable diagnostic criteria. Extremely high rates of gonorrhea have been reported among aboriginal populations in Namibia and Australia. Studies in Africa have clearly demonstrated that nonulcerative STIs such as gonorrhea (in addition to ulcerative STIs) are an independent risk factor for the transmission of HIV (Chap. 197).

Gonorrhea is transmitted from males to females more efficiently than in the opposite direction. The rate of transmission to a woman during a single unprotected sexual encounter with an infected man is ~50–70%. Oropharyngeal gonorrhea occurs in ~20% of women who practice fellatio with infected partners. Transmission in either direction by cunnilingus is rare.

In any population, there exists a small minority of individuals who have high rates of new-partner acquisition. These “core-group members” or “high-frequency transmitters” are vital in sustaining STI transmission at the population level. Another instrumental factor in sustaining gonorrhea in the population is the large number of infected individuals who are asymptomatic or have minor symptoms that are ignored. These persons, unlike symptomatic individuals, may not cease sexual activity and therefore may continue to transmit the infection. This situation underscores the importance of contact tracing and empirical treatment of the sex partners of index cases.

PATHOGENESIS, IMMUNOLOGY, AND ANTIMICROBIAL RESISTANCE

Outer-Membrane Proteins

PILI

Fresh clinical isolates of *N. gonorrhoeae* initially form piliated (fimbriated) colonies distinguishable on translucent agar. Pilus expression is rapidly switched off with unselected subculture because of rearrangements in pilus genes. This change is a basis for antigenic variation of gonococci. Piliated strains adhere better to cells derived from human mucosal surfaces and are more virulent in organ culture models

and human inoculation experiments than nonpiliated variants. In a fallopian tube explant model, pili mediate gonococcal attachment to nonciliated columnar epithelial cells. This event initiates gonococcal phagocytosis and transport through these cells to intercellular spaces near the basement membrane or directly into the subepithelial tissue. Pili are also essential for genetic competence and transformation of *N. gonorrhoeae*, which permit horizontal transfer of genetic material between different gonococcal lineages in vivo.

OPACITY-ASSOCIATED PROTEIN

Another gonococcal surface protein that is important in adherence to epithelial cells is opacity-associated protein (Opa, formerly called protein II). Opa contributes to intergonococcal adhesion, which is responsible for the opaque nature of gonococcal colonies on translucent agar and the organism's adherence to a variety of eukaryotic cells, including polymorphonuclear leukocytes (PMNs). Certain Opa variants promote invasion of epithelial cells, and this effect has been linked with the ability of Opa to bind vitronectin, glycosaminoglycans, and several members of the carcinoembryonic antigen–related cell adhesion molecule (CEACAM) receptor family. CEACAM5-binding gonococci prevent exfoliation of epithelium and may interfere with bacterial clearance. *N. gonorrhoeae* Opa proteins that bind CEACAM1, which is expressed by primary CD4+ T lymphocytes, suppress the activation and proliferation of these lymphocytes. Select Opa proteins can engage CEACAM3, which is expressed on neutrophils, with consequent nonopsonic phagocytosis (i.e., phagocytosis independent of antibody and complement) and killing of bacteria.

PORIN

Porin (previously designated protein I) is the most abundant gonococcal surface protein. Porin molecules exist as trimers that provide anion-transporting aqueous channels through the otherwise hydrophobic outer membrane. Porin exhibits stable interstrain antigenic variation and forms the basis for gonococcal serotyping. Two main serotypes have been identified; PorB.1A strains are often associated with disseminated gonococcal infection (DGI), whereas PorB.1B strains usually cause local genital infections only. DGI strains are generally resistant to the killing action of normal human serum and do not incite a significant local inflammatory response; therefore, they may not cause symptoms at genital sites. These characteristics may be related to the ability of PorB.1A strains to bind to complement-inhibitory molecules, resulting in a diminished inflammatory response. Porin can translocate to the cytoplasmic membrane of host cells—a process that could initiate gonococcal endocytosis and invasion.

OTHER OUTER-MEMBRANE PROTEINS

Other notable outer-membrane proteins include H.8, a lipoprotein that is present in high concentration on the surface of all gonococcal strains and is an excellent target for antibody-based diagnostic testing. Transferrin-binding proteins (Tbp1 and Tbp2) and lactoferrin-binding protein are required for scavenging iron from transferrin and lactoferrin in vivo. Transferrin and iron have been shown to enhance the attachment of iron-deprived *N. gonorrhoeae* to human endometrial cells. IgA1 protease is produced by *N. gonorrhoeae* and may protect the organism from the action of mucosal IgA.

Lipooligosaccharide

Gonococcal lipooligosaccharide (LOS) consists of a lipid A and a core oligosaccharide that lacks the repeating O-carbohydrate antigenic side chain seen in other gram-negative bacteria **(Chap. 116)**. Gonococcal LOS possesses marked endotoxic activity and contributes to the local cytotoxic effect in a fallopian tube model. LOS core sugars undergo a high degree of phase variation under different conditions of growth; this variation reflects genetic regulation and expression of glycotransferase genes that dictate the carbohydrate structure of LOS. These phenotypic changes may affect interactions of *N. gonorrhoeae* with elements of the humoral immune system (antibodies and complement) and may also influence direct binding of organisms to both professional phagocytes and nonprofessional phagocytes (epithelial cells). For example, gonococci that are sialylated at their LOS sites inhibit the classic pathway of complement by reducing binding of IgG and also bind complement factor H to inhibit the alternative pathway of complement. LOS sialylation may also decrease nonopsonic Opa-mediated association with neutrophils and inhibit the oxidative burst in PMNs. The binding of the unsialylated terminal lactosamine residue of LOS to an asialoglycoprotein receptor on male epithelial cells facilitates adherence and subsequent gonococcal invasion of these cells. Moreover, oligosaccharide structures in LOS can modulate host immune responses. For example, the terminal monosaccharide expressed by LOS determines the C-type lectin receptor on dendritic cells that is targeted by the bacteria. In turn, the specific C-type lectin receptor engaged influences whether a T_H1- or T_H2-type response is elicited; the latter response may be less favorable for clearance of gonococcal infection.

Host Factors

In addition to gonococcal structures that interact with epithelial cells, host factors seem to be important in mediating entry of gonococci into nonphagocytic cells. Activation of phosphatidylcholine-specific phospholipase C and acidic sphingomyelinase by *N. gonorrhoeae*, which results in the release of diacylglycerol and ceramide, is a requirement for the entry of *N. gonorrhoeae* into epithelial cells. Ceramide accumulation within cells leads to apoptosis, which may disrupt epithelial integrity and facilitate entry of gonococci into subepithelial tissue. Release of chemotactic factors as a result of complement activation contributes to inflammation, as does the toxic effect of LOS in provoking the release of inflammatory cytokines.

The importance of humoral immunity in host defenses against neisserial infections is best illustrated by the predisposition of persons deficient in terminal complement components (C5 through C9) to have recurrent bacteremic gonococcal infections and recurrent meningococcal meningitis or meningococcemia. Gonococcal porin induces T cell–proliferative responses in persons with urogenital gonococcal disease. A significant increase in porin-specific interleukin (IL) 4–producing CD4+ as well as CD8+ T lymphocytes is seen in individuals with mucosal gonococcal disease. A portion of these lymphocytes that show a porin-specific T_H2-type response could traffic to mucosal surfaces and play a role in immune protection against the disease. Few data clearly indicate that protective immunity is acquired from a previous gonococcal infection, although bactericidal and opsonophagocytic antibodies to porin and LOS may offer partial protection. On the other hand, women who are infected and acquire high levels of antibody to another outer-membrane

protein, Rmp (reduction modifiable protein, formerly called protein III), may be especially likely to become reinfected with *N. gonorrhoeae* because Rmp antibodies block the effect of bactericidal antibodies to porin and LOS. Rmp shows little, if any, interstrain antigenic variation; therefore, Rmp antibodies potentially may block antibody-mediated killing of all gonococci. The mechanism of blocking has not been fully characterized, but Rmp antibodies may noncompetitively inhibit binding of porin and LOS antibodies because of the proximity of these structures in the gonococcal outer membrane. In male volunteers who have no history of gonorrhea, the net effect of these events may influence the outcome of experimental challenge with *N. gonorrhoeae*. Because Rmp bears extensive homology to enterobacterial OmpA and meningococcal class 4 proteins, it is possible that these blocking antibodies result from prior exposure to cross-reacting proteins from these species and also play a role in first-time infection with *N. gonorrhoeae*.

Gonococcal Resistance to Antimicrobial Agents

It is no surprise that *N. gonorrhoeae*, with its remarkable capacity to alter its antigenic structure and adapt to changes in the microenvironment, has become resistant to numerous antibiotics. The first effective agents against gonorrhea were the sulfonamides, which were introduced in the 1930s and became ineffective within a decade. Penicillin was then used as the drug of choice for the treatment of gonorrhea. By 1965, 42% of gonococcal isolates had developed low-level resistance to penicillin G. Resistance due to the production of penicillinase arose later.


Gonococci become fully resistant to antibiotics either by chromosomal mutations or by acquisition of R factors (plasmids). Two types of chromosomal mutations have been described. The first type, which is drug specific, is a single-step mutation leading to high-level resistance. The second type involves mutations at several chromosomal loci that combine to determine the level as well as the pattern of resistance. Strains with mutations in chromosomal genes were first observed in the late 1950s. As recently as 2007, chromosomal mutations accounted for resistance to penicillin, tetracycline, or both in ~16% of strains surveyed in the United States.

β-Lactamase (penicillinase)–producing strains of *N. gonorrhoeae* (PPNG) carrying β-lactamase plasmids had rapidly spread worldwide by the early 1980s. *N. gonorrhoeae* strains with plasmid-borne tetracycline resistance (TRNG) can mobilize some β-lactamase plasmids, and PPNG and TRNG occur together, sometimes along with strains exhibiting chromosomally mediated resistance (CMRNG). Penicillin, ampicillin, and tetracycline are no longer reliable for the treatment of gonorrhea and should not be used.


Quinolone-containing regimens also were recommended for treatment of gonococcal infections; the fluoroquinolones offered the advantage of antichlamydial activity when administered for 7 days. However, quinolone-resistant *N. gonorrhoeae* (QRNG) appeared soon after these agents were first used to treat gonorrhea. QRNG is particularly common in the Pacific Islands (including Hawaii) and Asia, where, in certain areas, all gonococcal strains are now resistant to quinolones. At present, QRNG is also common in parts of Europe and the Middle East. In the United States, QRNG has been identified in all areas but predominantly in

states on the Pacific coast, where resistant strains were first seen. Alterations in DNA gyrase and topoisomerase IV have been implicated as mechanisms of fluoroquinolone resistance.

Resistance to spectinomycin, which has been used in the past as an alternative agent, has been reported. Because this agent usually is not associated with resistance to other antibiotics, spectinomycin can be reserved for use against multidrug-resistant strains of *N. gonorrhoeae*. Nevertheless, outbreaks caused by strains resistant to spectinomycin have been documented in Korea and England when the drug has been used for primary treatment of gonorrhea.

Third-generation cephalosporins have remained highly effective as single-dose therapy for gonorrhea, but the recent isolation of strains highly resistant to ceftriaxone (minimal inhibitory concentrations [MICs], 2 µg/mL) in Japan and some European countries is cause for concern. Even though the MICs of ceftriaxone against certain strains may reach 0.015–0.125 µg/mL (higher than the MICs of 0.0001–0.008 µg/mL for fully susceptible strains), these levels are greatly exceeded in the blood, the urethra, and the cervix when the routinely recommended parenteral dose of ceftriaxone is administered. The rising MICs of oral cefixime (the previously recommended alternative oral third-generation cephalosporin) against *N. gonorrhoeae*, combined with this drug's limited capacity to reach levels sufficiently higher than MICs in the blood, the urethra, the cervix, and especially the pharynx, have resulted in the removal of cefixime from the list of first-line agents for treatment of uncomplicated gonorrhea. *N. gonorrhoeae* strains with reduced susceptibility to ceftriaxone and cefixime (i.e., cephalosporin-intermediate/resistant strains) contain mutations in (1) the *penA* allele, which is the principal resistance determinant and encodes a penicillin-binding protein (PBP2) whose sequence can differ in up to 60–70 amino acids from that of wild-type PBP2; (2) the *m*ultiple *t*ransferable *r*esistance *r*egulator (*mtrR*) gene that results in increased drug efflux through the MtrCDE efflux pump; and (3) *penB*, which decreases drug influx through PorB.

Resistance to azithromycin can result from alterations of the ribosomal binding target by azithromycin and—as with cephalosporins—the over- and under-expression of efflux and influx systems. Combined resistance to cephalosporins and azithromycin could contribute to the failure of the currently recommended dual therapy for gonorrhea with these two antimicrobial agents. Indeed, clinical failures caused by organisms resistant to these agents have been reported on two occasions in infected heterosexual men treated with both agents.

CLINICAL MANIFESTATIONS

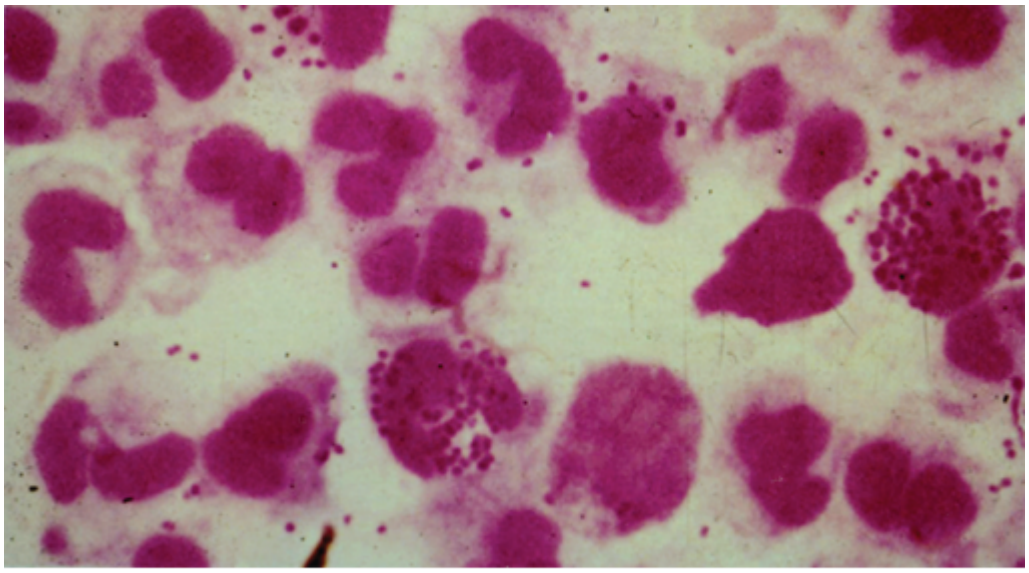
Gonococcal Infections in Men

Acute urethritis is the most common clinical manifestation of gonorrhea in male patients. The usual incubation period after exposure is 2–7 days, although the interval can be longer and most men remain asymptomatic. Strains of the PorB.1A serotype tend to cause a greater proportion of cases of mild and asymptomatic urethritis than do PorB.1B strains. When they occur, urethral discharge and dysuria, usually without urinary frequency or urgency, are the major symptoms. The discharge initially is scant and mucoid but becomes profuse and purulent within a day or two. Gram's staining of the urethral discharge may reveal PMNs and gram-negative intracellular monococci and diplococci **(Fig. 151-1)**. The clinical manifestations of

gonococcal urethritis are usually more severe and overt than those of nongonococcal urethritis, including urethritis caused by *Chlamydia trachomatis* (Chap. 184); however, exceptions are common, and it is often impossible to differentiate the causes of urethritis on clinical grounds alone. The majority of cases of urethritis seen in the United States today are not caused by *N. gonorrhoeae* and/or *C. trachomatis*. Although a number of other organisms may be responsible, many cases do not have a specific etiologic agent identified. Certain clones of *Neisseria meningitidis*, the second member of the pathogenic *Neisseria* species, have been associated with urethritis in men who have sex with men (MSM) in Europe and in heterosexual men in the southern and midwestern United States.

FIGURE 151-1

Gram's stain of urethral discharge from a male patient with gonorrhea shows gram-negative intracellular monococci and diplococci. *(From the Public Health Agency of Canada.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Most symptomatic men with gonorrhea seek treatment and cease to be infectious. The remaining men, who are largely asymptomatic, accumulate in number over time and constitute about two-thirds of all infected men at any point in time; together with men incubating the organism who shed the organism but are asymptomatic, they serve as the source of spread of infection. Before the antibiotic era, symptoms of urethritis persisted for ~8 weeks. Epididymitis is now an uncommon complication, and gonococcal prostatitis occurs rarely, if at all. Other unusual local complications of gonococcal urethritis include edema of the penis due to dorsal lymphangitis or thrombophlebitis, submucous inflammatory "soft" infiltration of the urethral wall, periurethral abscess or fistula, inflammation or abscess of Cowper's gland, and seminal vesiculitis. Balanitis may develop in uncircumcised men.

Gonococcal Infections in Women

GONOCOCCAL CERVICITIS

Mucopurulent cervicitis is a common STI diagnosis in American women and may be caused by *N. gonorrhoeae, C. trachomatis*, and other organisms, including *Mycoplasma genitalium* (Chap. 183). Cervicitis may coexist with candidal or trichomonal vaginitis. *N. gonorrhoeae* primarily infects the columnar epithelium of the cervical os. Bartholin's glands occasionally become infected.

Women infected with *N. gonorrhoeae* usually develop symptoms. However, women who either remain asymptomatic or have only minor symptoms may delay in seeking medical attention. These minor symptoms may include scant vaginal discharge issuing from the inflamed cervix (without vaginitis or vaginosis per se) and dysuria (often without urgency or frequency) that may be associated with gonococcal urethritis. Although the incubation period of gonorrhea is less well defined in women than in men, symptoms usually develop within 10 days of infection and are more acute and intense than those of chlamydial cervicitis.

The physical examination reveals a mucopurulent discharge (mucopus) issuing from the cervical os or a reddened (inflamed) cervix even in the absence of reported symptoms. Because Gram's stain is not sensitive for the diagnosis of gonorrhea in women, specimens should be submitted for culture or a nonculture assay (see "Laboratory Diagnosis," below). Edematous and friable cervical ectopy and endocervical bleeding induced by gentle swabbing are more often seen in chlamydial infection. Gonococcal infection may extend deep enough to produce dyspareunia and lower abdominal or back pain. In such cases, it is imperative to consider a diagnosis of pelvic inflammatory disease (PID) and to administer treatment for that disease (Chaps. 131 and 184).

N. gonorrhoeae may also be recovered from the urethra and rectum of women with cervicitis, but these are rarely the only infected sites. Urethritis in women may produce symptoms of internal dysuria, which is often attributed to "cystitis." Pyuria in the absence of bacteriuria visible on Gram's stain of unspun urine, accompanied by urine cultures that fail to yield $>10^2$ colonies of bacteria usually associated with urinary tract infection, signifies the possibility of urethritis due to *C. trachomatis*. Urethral infection with *N. gonorrhoeae* also may occur in this context, but in this instance urethral cultures are usually positive.

GONOCOCCAL VAGINITIS

The vaginal mucosa of healthy women is lined by stratified squamous epithelium and is rarely infected by *N. gonorrhoeae*. However, gonococcal vaginitis can occur in anestrogenic women (e.g., prepubertal girls and postmenopausal women), in whom the vaginal stratified squamous epithelium is often thinned down to the basilar layer, which can be infected by *N. gonorrhoeae*. The intense inflammation of the vagina makes the physical (speculum and bimanual) examination extremely painful. The vaginal mucosa is red and edematous, and an abundant purulent discharge is often present. Infection in the urethra and in Skene's and Bartholin's glands often accompanies gonococcal vaginitis. Inflamed cervical erosion or abscesses in nabothian cysts may also occur. Coexisting cervicitis may result in pus in the cervical os.

Anorectal Gonorrhea

Because the female anatomy permits the spread of cervical exudate to the rectum, *N. gonorrhoeae* is sometimes recovered from the rectum of women with uncomplicated gonococcal cervicitis. The rectum is the sole site of infection in only 5% of women with gonorrhea. Such women are usually asymptomatic but occasionally have acute proctitis manifested by anorectal pain or pruritus, tenesmus, purulent rectal discharge, and rectal bleeding. Among MSM, the frequency of gonococcal infection, including rectal infection, fell by ≥90% throughout the United States in the early 1980s, but a resurgence of gonorrhea among MSM has been documented in several cities since the 1990s. Gonococcal isolates from the rectum of MSM tend to be more resistant to antimicrobial agents than are gonococcal isolates from other sites. Gonococcal isolates with a mutation in *mtrR* or in the promoter region of the gene that encodes for this transcriptional regulator develop increased resistance to antimicrobial hydrophobic agents such as bile acids and fatty acids in feces and thus are found with increased frequency in MSM. This situation may have been responsible for higher rates of failure of treatment for rectal gonorrhea with older regimens consisting of penicillin or tetracyclines.

Pharyngeal Gonorrhea

Pharyngeal gonorrhea is usually mild or asymptomatic, although symptomatic pharyngitis does occasionally occur with cervical lymphadenitis. The mode of acquisition is oral–genital sexual exposure, with fellatio being a more efficient means of transmission than cunnilingus. In certain female adolescent populations in the United States, pharyngeal gonorrhea has become as common as genital gonorrhea. Most cases resolve spontaneously, and transmission from the pharynx to sexual contacts is rare. Pharyngeal infection almost always coexists with genital infection. Swabs from the pharynx should be plated directly onto gonococcal selective media. Pharyngeal colonization with *N. meningitidis* needs to be differentiated from that with other *Neisseria* species.

Ocular Gonorrhea in Adults

Ocular gonorrhea in an adult usually results from autoinoculation of *N. gonorrhoeae* from an infected genital site. As in genital infection, the manifestations range from severe to occasionally mild or asymptomatic disease. The variability in clinical manifestations may be attributable to differences in the ability of the infecting strain to elicit an inflammatory response. Infection may result in a markedly swollen eyelid, severe hyperemia and chemosis, and a profuse purulent discharge. The massively inflamed conjunctiva may be draped over the cornea and limbus. Lytic enzymes from the infiltrating PMNs occasionally cause corneal ulceration and rarely cause perforation.

Prompt recognition and treatment of this condition are of paramount importance. Gram's stain and culture of the purulent discharge establish the diagnosis. Genital cultures also should be performed.

Gonorrhea in Pregnant Women, Neonates, and Children

Gonorrhea in pregnancy can have serious consequences for both the mother and the infant. Recognition of gonorrhea early in pregnancy also identifies a population at risk for other STIs, particularly chlamydial infection, syphilis, and trichomoniasis. The risks of salpingitis and PID—conditions associated with a high

rate of fetal loss—are highest during the first trimester. Pharyngeal infection, most often asymptomatic, may be more common during pregnancy because of altered sexual practices. Prolonged rupture of the membranes, premature delivery, chorioamnionitis, funisitis (infection of the umbilical cord stump), and sepsis in the infant (with *N. gonorrhoeae* detected in the newborn's gastric aspirate during delivery) are common complications of maternal gonococcal infection at term. Other conditions and microorganisms, including *Mycoplasma hominis*, *Ureaplasma urealyticum*, *C. trachomatis*, and bacterial vaginosis (often accompanied by infection with *Trichomonas vaginalis*), have been associated with similar complications.

The most common form of gonorrhea in neonates is ophthalmia neonatorum, which results from exposure to infected cervical secretions during parturition. Ocular neonatal instillation of a prophylactic agent (e.g., 1% silver nitrate eye drops or ophthalmic preparations containing erythromycin or tetracycline) prevents ophthalmia neonatorum but is not effective for its treatment, which requires systemic antibiotics. The clinical manifestations are acute and usually begin 2–5 days after birth. An initial nonspecific conjunctivitis with a serosanguineous discharge is followed by tense edema of the eyelids, chemosis, and a profuse, thick, purulent discharge. Corneal ulcerations that result in nebulae or perforation may lead to anterior synechiae, anterior staphyloma, panophthalmitis, and blindness. Infections described at other mucosal sites in infants, including vaginitis, rhinitis, and anorectal infection, are likely to be asymptomatic. Pharyngeal colonization has been demonstrated in 35% of infants with gonococcal ophthalmia, and coughing is the most prominent symptom in these cases. Septic arthritis (see below) is the most common manifestation of systemic infection or DGI in the newborn. The onset usually comes at 3–21 days of age, and polyarticular involvement is common. Sepsis, meningitis, and pneumonia are seen in rare instances.

Any STI in children beyond the neonatal period raises the possibility of sexual abuse. Gonococcal vulvovaginitis is the most common manifestation of gonococcal infection in children beyond infancy. Anorectal and pharyngeal infections are common in these children and are frequently asymptomatic. The urethra, Bartholin's and Skene's glands, and the upper genital tract are rarely involved. All children with gonococcal infection should also be evaluated for chlamydial infection, syphilis, and possibly HIV infection.

Gonococcal Arthritis (DGI)

DGI (gonococcal arthritis) results from gonococcal bacteremia. In the 1970s, DGI occurred in ~0.5–3% of persons with untreated gonococcal mucosal infection. The lower incidence of DGI at present is probably attributable to a decline in the prevalence of particular strains that are likely to disseminate. DGI strains resist the bactericidal action of human serum and generally do not incite inflammation at genital sites, probably because of limited generation of chemotactic factors. Strains recovered from DGI cases in the 1970s were often of the PorB.1A serotype, were highly susceptible to penicillin, and had special growth requirements—including arginine, hypoxanthine, and uracil—that made the organism more fastidious and more difficult to isolate.

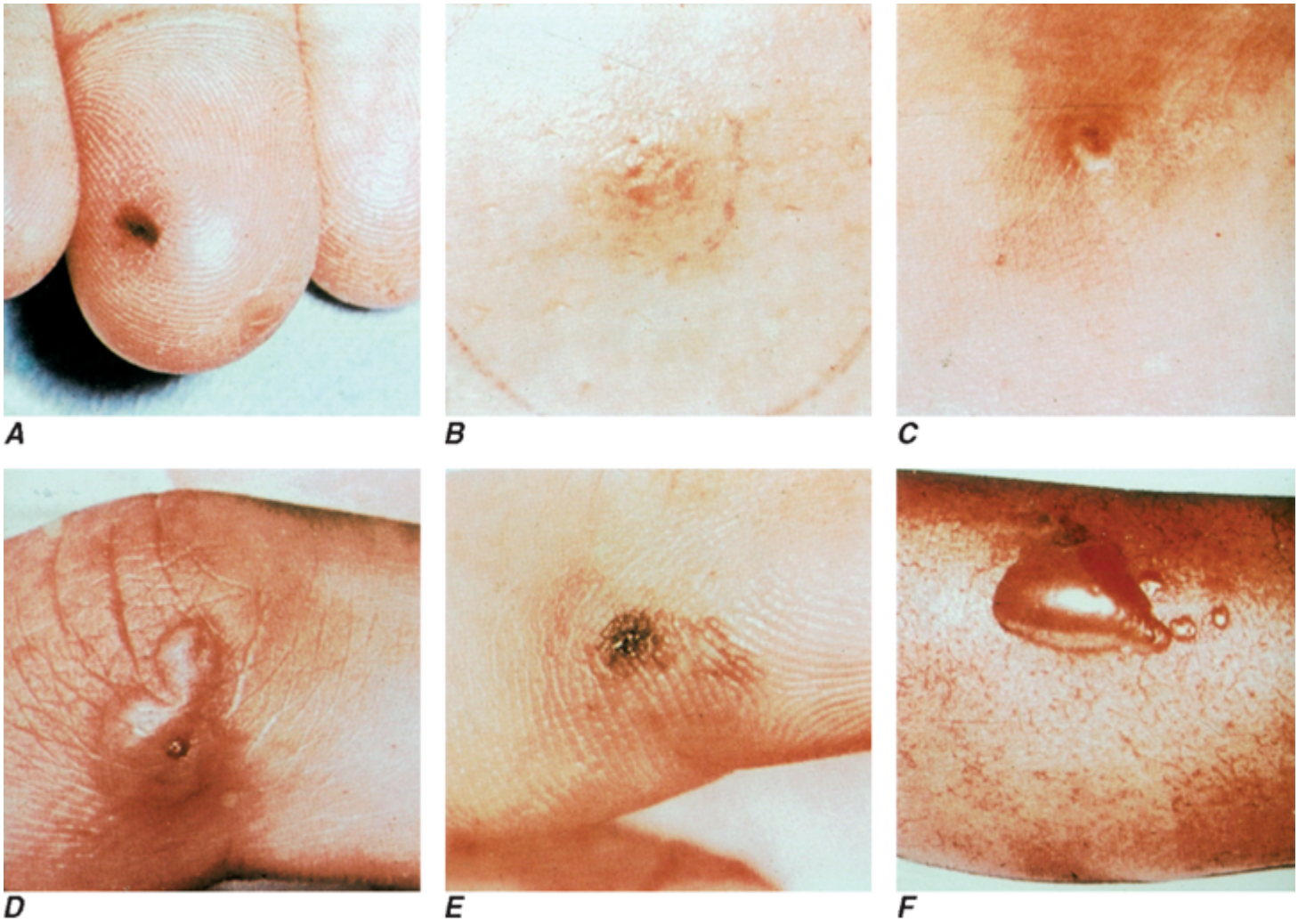
Menstruation is a risk factor for dissemination, and approximately two-thirds of cases of DGI are in women. In about half of affected women, symptoms of DGI begin within 7 days of onset of menses. Complement deficiencies, especially of the components involved in the assembly of the membrane attack complex (C5

through C9), predispose to neisserial bacteremia, and persons with more than one episode of DGI should be screened with an assay for total hemolytic complement activity.

The clinical manifestations of DGI have sometimes been classified into two stages: a bacteremic stage, which is less common today, and a joint-localized stage with suppurative arthritis. A clear-cut progression usually is not evident. Patients in the bacteremic stage have higher temperatures, and chills more frequently accompany their fever. Painful joints are common and often occur together with tenosynovitis and skin lesions. Polyarthralgias usually include the knees, elbows, and more distal joints; the axial skeleton is generally spared. Skin lesions are seen in ~75% of patients and include papules and pustules, often with a hemorrhagic component **(Fig. 151-2; see also Fig. A1-43)**. Other manifestations of noninfectious dermatitis, such as nodular lesions, urticaria, and erythema multiforme, have been described. These lesions are usually on the extremities and number between 5 and 40. The differential diagnosis of the bacteremic stage of DGI includes reactive arthritis, acute rheumatoid arthritis, sarcoidosis, erythema nodosum, drug-induced arthritis, and viral infections (e.g., hepatitis B and acute HIV infection). The distribution of joint symptoms in reactive arthritis differs from that in DGI **(Fig. 151-3)**, as do the skin and genital manifestations **(Chap. 355)**.

FIGURE 151-2

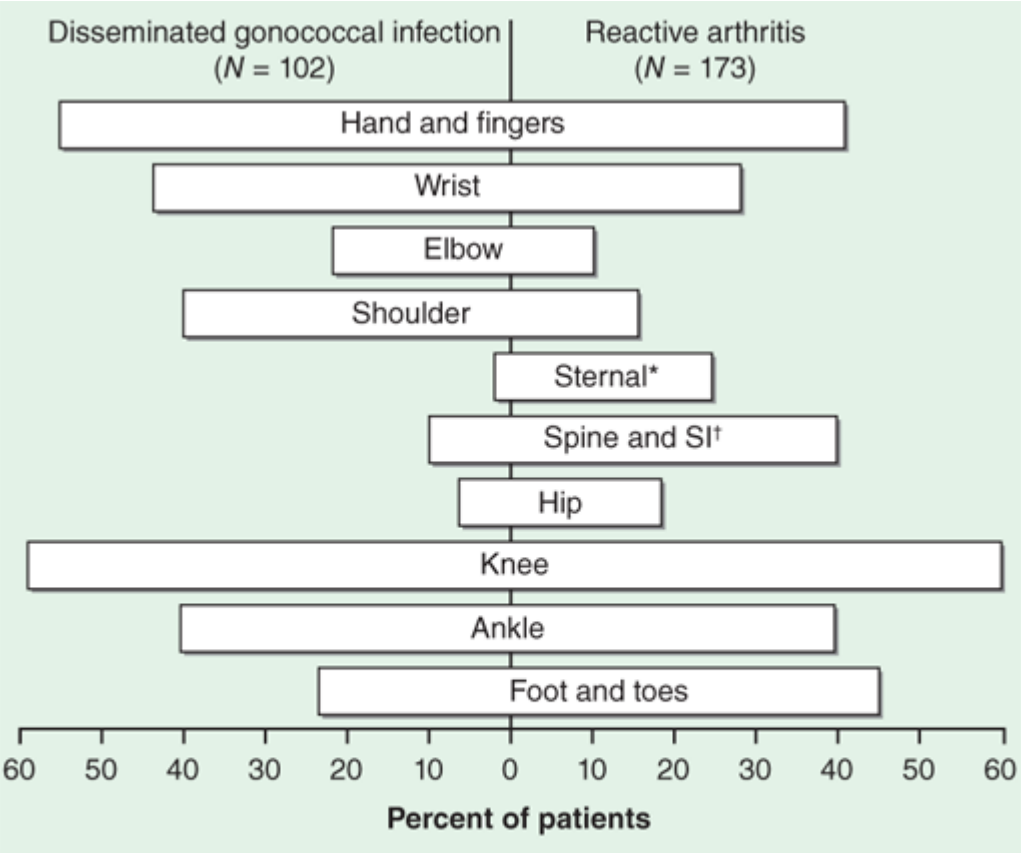
Characteristic skin lesions in patients with proven gonococcal bacteremia. The lesions are in various stages of evolution. ***A.*** Very early petechia on finger. ***B.*** Early papular lesion, 7 mm in diameter, on lower leg. ***C.*** Pustule with central eschar resulting from early petechial lesion. ***D.*** Pustular lesion on finger. ***E.*** Mature lesion with central necrosis (black) on hemorrhagic base. ***F.*** Bullae on anterior tibial surface. *(Reprinted with permission from KK Holmes et al: Disseminated gonococcal infection. Ann Intern Med 74:979, 1971.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 151-3

Distribution of joints with arthritis in 102 patients with disseminated gonococcal infection and 173 patients with reactive arthritis. *Includes the sternoclavicular joints. †SI, sacroiliac joint.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Suppurative arthritis involves one or two joints, most often the knees, wrists, ankles, and elbows (in decreasing order of frequency); other joints occasionally are involved. Most patients who develop gonococcal septic arthritis do so without prior polyarthralgias or skin lesions; in the absence of symptomatic genital infection, this disease cannot be distinguished from septic arthritis caused by other pathogens. The differential diagnosis of acute arthritis in young adults is discussed in **Chap. 125**. Rarely, osteomyelitis complicates septic arthritis involving small joints of the hand.

Gonococcal endocarditis, although rare today, was a relatively common complication of DGI in the preantibiotic era, accounting for about one-quarter of reported cases of endocarditis. Another unusual complication of DGI is meningitis.

Gonococcal Infections in HIV-Infected Persons

The association between gonorrhea and the acquisition of HIV has been demonstrated in several well-controlled studies, mainly in Kenya and Zaire. The nonulcerative STIs enhance the transmission of HIV three- to fivefold; transmission of HIV-infected immune cells and increased viral shedding by persons with urethritis or cervicitis may contribute **(Chap. 197)**. HIV has been detected by polymerase chain reaction (PCR) more commonly in ejaculates from HIV-positive men with gonococcal urethritis than in those from HIV-positive men with nongonococcal urethritis. PCR positivity diminishes twofold after appropriate therapy for urethritis. Not only does gonorrhea enhance the transmission of HIV, but it may also increase the individual's risk for acquisition of HIV. A proposed mechanism is the significantly greater number of CD4+ T lymphocytes

and dendritic cells that can be infected by HIV in endocervical secretions from women with nonulcerative STIs than in those from women with ulcerative STIs.

LABORATORY DIAGNOSIS

A rapid diagnosis of gonococcal infection in men may be obtained by Gram's staining of urethral exudates (Fig. 151-1). The detection of gram-negative intracellular monococci and diplococci is usually highly specific and sensitive in diagnosing gonococcal urethritis in symptomatic males but is only ~50% sensitive in diagnosing gonococcal cervicitis. Samples should be collected with Dacron or rayon swabs. Part of the sample should be inoculated onto a plate of modified Thayer-Martin or other gonococcal selective medium for culture. It is important to process all samples immediately because gonococci do not tolerate drying. If plates cannot be incubated immediately, they can be held safely for several hours at room temperature in candle extinction jars prior to incubation. If processing is to occur within 6 h, transport of specimens may be facilitated by the use of nonnutritive swab transport systems such as Stuart or Amies medium. For longer holding periods (e.g., when specimens for culture are to be mailed), culture media with self-contained CO_2-generating systems (such as the JEMBEC or Gono-Pak systems) may be used. Specimens should also be obtained for the diagnosis of chlamydial infection **(Chap. 184)**.

PMNs are often seen in the endocervix on a Gram's stain, and an abnormally increased number (≥30 PMNs per field in five 1000× oil-immersion microscopic fields) establishes the presence of an inflammatory discharge. Unfortunately, the presence or absence of gram-negative intracellular monococci or diplococci in cervical smears does not accurately predict which patients have gonorrhea, and the diagnosis in this setting should be made by culture or another suitable nonculture diagnostic method. The sensitivity of a single endocervical culture is ~80–90%. If a history of rectal sex is elicited, a rectal wall swab (uncontaminated with feces) should be cultured. A presumptive diagnosis of gonorrhea cannot be made on the basis of gram-negative diplococci in smears from the pharynx, where other *Neisseria* species are components of the normal flora.

Several nucleic acid amplification tests (NAATs), including the Roche COBAS AMPLICOR, Gen-Probe Aptima Combo 2, and BD ProbeTec ET, are now widely available on semiautomated or fully automated platforms and are the most commonly employed diagnostic tests for gonorrhea. These tests also detect *C. trachomatis* and are more sensitive than culture for identification of either *N. gonorrhoeae* or *C. trachomatis*. The Gen-Probe and BD tests offer the advantage that urine samples can be tested with a sensitivity similar to or greater than that obtained when urethral or cervical swab samples are assessed by other non-NAATs or culture, respectively. A disadvantage of non-culture-based assays is that *N. gonorrhoeae* cannot be grown from the transport systems. Thus a culture-confirmatory test and formal antimicrobial susceptibility testing, if needed, cannot be performed.

Because of the legal implications, the preferred method for the diagnosis of gonococcal infection in children is a standardized culture. Two positive NAATs, each targeting a different nucleic acid sequence, may be substituted for culture of the cervix or the urethra as legal evidence of infection in children. Although

nonculture tests for gonococcal infection have not been approved by the U.S. Food and Drug Administration for use with specimens obtained from the pharynx and rectum of infected children, NAATs from these sites are preferred for diagnostic evaluation in adult victims of suspected sexual abuse, especially if the NAATs have been evaluated by the local laboratory and found to be superior. Cultures should be obtained from the pharynx and anus of both girls and boys, the urethra of boys, and the vagina of girls; cervical specimens are not recommended for prepubertal girls. For boys with a urethral discharge, a meatal specimen of the discharge is adequate for culture. Presumptive colonies of *N. gonorrhoeae* should be identified definitively by at least two independent methods.

Blood should be cultured in suspected cases of DGI. The use of Isolator blood culture tubes may enhance the yield. The probability of positive blood cultures decreases after 48 h of illness. Synovial fluid should be inoculated into blood culture broth medium and plated onto chocolate agar rather than selective medium because this fluid is not likely to be contaminated with commensal bacteria. Gonococci are infrequently recovered from early joint effusions containing <20,000 leukocytes/μL but may be recovered from effusions containing >80,000 leukocytes/μL. The organisms are seldom recovered from blood and synovial fluid of the same patient.

TREATMENT

TREATMENT

Gonococcal Infections

Treatment failure can lead to continued transmission and the emergence of antibiotic resistance. The importance of adequate treatment with a regimen that the patient will adhere to cannot be overemphasized. Thus highly effective single-dose regimens have been developed for uncomplicated gonococcal infections. The 2015 treatment guidelines for gonococcal infections from the CDC are summarized in **Table 151-1**. Rising MICs of cefixime worldwide have led the CDC to discontinue its recommendation of this agent as first-line treatment for uncomplicated gonorrhea. The third-generation cephalosporin ceftriaxone in combination with azithromycin is recommended as treatment; dual therapy against gonorrhea could slow the development of resistance to either of these antimicrobial agents. Azithromycin, which also treats nongonococcal urethritis, is preferred to doxycycline because of its superior activity against gonorrhea and ease of use. The recommendations for uncomplicated gonorrhea apply to HIV-infected as well as HIV-uninfected patients.

The currently recommended regimen for the treatment of uncomplicated gonococcal infection of the urethra, cervix, rectum, or pharynx (a single IM dose of ceftriaxone plus a single dose of azithromycin taken orally) almost always results in an effective cure. Quinolone-containing regimens are no longer recommended in the United States as first-line treatment because of widespread resistance. A multicenter trial of treatment for uncomplicated gonorrhea in the United States showed ≥99.5% efficacy of two combination regimens: (1) gemifloxacin (320 mg, single oral dose) plus azithromycin (2 g, single oral dose) or (2) azithromycin (2 g, single oral dose) plus gentamicin (a single IM dose of 240 mg or, in individuals who

weigh ≤45 kg, 5 mg/kg). At this time, however, neither of these regimens is recommended as first-line treatment.

Co-infection with *C. trachomatis* occurs frequently; treatment of gonorrhea with ceftriaxone that also includes a single 1-g dose of azithromycin is effective against chlamydial infection. However, a 1-g dose of azithromycin used alone as treatment for gonorrhea in penicillin-allergic persons results in an unacceptably low cure rate (93%) for gonococcal infections and should not be used. A single 2-g dose of azithromycin, particularly in the extended-release microsphere formulation, delivers azithromycin to the lower gastrointestinal tract, thereby improving tolerability. Azithromycin is effective against sensitive strains, but this drug is expensive, causes gastrointestinal distress, and is not recommended for routine or first-line treatment of gonorrhea. Spectinomycin has been used as an alternative agent for the treatment of uncomplicated gonococcal infections in penicillin-allergic persons outside the United States but is not currently available in this country. Of note, the limited effectiveness of spectinomycin for the treatment of pharyngeal infection reduces its utility in populations among whom such infection is common, such as MSM.

Persons with uncomplicated infections who receive ceftriaxone and azithromycin do not need a test of cure; however, cultures for *N. gonorrhoeae* should be performed if symptoms persist after therapy with an established regimen, and any gonococci isolated should be tested for antimicrobial susceptibility. Persons given an alternative regimen should return for a test of cure targeting the infected anatomic site. This test ideally should be a culture. If culture is not readily available and NAAT is positive, every effort should be made to perform a confirmatory culture. All isolates from test-of-cure cultures should undergo antimicrobial susceptibility testing. Because of high rates of reinfection with *N. gonorrhoeae* and *C. trachomatis* within 6 months, repeat testing is recommended 3 months after treatment.

Symptomatic gonococcal pharyngitis is more difficult to eradicate than genital infection. Persons who cannot tolerate ceftriaxone and those in whom quinolones are contraindicated may be treated with spectinomycin if it is available, but this agent results in a cure rate of ≤52%. Persons given spectinomycin should have a pharyngeal sample cultured 3–5 days after treatment as a test of cure. A single 2-g dose of azithromycin may be used in areas where rates of resistance to azithromycin are low.

Treatments for gonococcal epididymitis and PID are discussed in **Chap. 131**. Ocular gonococcal infections in older children and adults should be managed with a single dose of ceftriaxone combined with saline irrigation of the conjunctivae (both undertaken expeditiously), and patients should undergo a careful ophthalmologic evaluation that includes a slit-lamp examination.

DGI may require higher dosages and longer durations of therapy (Table 151-1). Hospitalization is indicated if the diagnosis is uncertain, if the patient has localized joint disease that requires aspiration, or if the patient cannot be relied on to comply with treatment. Open drainage is necessary only occasionally—e.g., for management of hip infections that may be difficult to drain percutaneously. Nonsteroidal anti-inflammatory agents may be indicated to alleviate pain and hasten clinical improvement of affected joints.

Gonococcal meningitis and endocarditis should be treated in the hospital with high-dose IV ceftriaxone (1–2 g IV every 12–24 h); therapy should continue for 10–14 days for meningitis and for at least 4 weeks for endocarditis. All persons who experience more than one episode of DGI should be evaluated for complement deficiency.

TABLE 151-1

Recommended Treatment for Gonococcal Infections: Adapted from the 2015 Guidelines of the Centers for Disease Control and Prevention

Diagnosis	Treatment of Choice[a]
Uncomplicated gonococcal infection of the cervix, urethra, pharynx[b], or rectum	
First-line regimen	Ceftriaxone (250 mg IM, single dose)
	plus
	Azithromycin (1 g PO, single dose)
Alternative regimens[c]	Cefixime (400 mg PO, single dose) *or* ceftizoxime (500 mg IM, single dose) *or* cefotaxime (500 mg IM, single dose) *or* spectinomycin (2 g IM, single dose)[d,e] *or* cefotetan (1 g IM, single dose) plus probenecid (1 g PO, single dose)[d] *or* cefoxitin (2 g IM, single dose) plus probenecid (1 g PO, single dose)[d]
	plus
	Azithromycin (1 g PO, single dose)
Epididymitis	**See Chap. 131**
Pelvic inflammatory disease	**See Chap. 131**
Gonococcal conjunctivitis in an adult	Ceftriaxone (1 g IM, single dose)[f]
Ophthalmia neonatorum[g]	Ceftriaxone (25–50 mg/kg IV, single dose, not to exceed 125 mg)

Diagnosis	Treatment of Choice[a]
Disseminated gonococcal infection[h]	
Initial therapy[i]	
Patient tolerant of β-lactam drugs	Ceftriaxone (1 g IM or IV q24h; recommended) *or* cefotaxime (1 g IV q8h) *or* ceftizoxime (1 g IV q8h)
Patients allergic to β-lactam drugs	Spectinomycin (2 g IM q12h)[d]
Continuation therapy[i]	Cefixime (400 mg PO bid)
Meningitis or endocarditis	See text for specific recommendations[j]

[a]True failure of treatment with a recommended regimen is rare and should prompt an evaluation for reinfection, infection with a drug-resistant strain, or an alternative diagnosis. [b]Ceftriaxone and azithromycin are the only agents recommended for treatment of pharyngeal infection. [c]See text for follow-up of persons with infection who are treated with alternative regimens. [d]Spectinomycin, cefotetan, and cefoxitin, which are alternative agents, currently are unavailable or in short supply in the United States. [e]Spectinomycin may be ineffective for the treatment of pharyngeal gonorrhea. [f]Plus lavage of the infected eye with saline solution (once). [g]Prophylactic regimens are discussed in the text. [h]Hospitalization is indicated if the diagnosis is uncertain, if the patient has frank arthritis with an effusion, or if the patient cannot be relied on to adhere to treatment. [i]All initial regimens should also include azithromycin (1 g PO, single dose) and should be continued for 24–48 h after clinical improvement begins, at which time the switch may be made to an oral agent (e.g., cefixime) if antimicrobial susceptibility can be documented by culture of the causative organism. If no organism is isolated and the diagnosis is secure, then treatment with ceftriaxone should be continued for at least 1 week. [j]Hospitalization is indicated to exclude suspected meningitis or endocarditis.

PREVENTION AND CONTROL

Condoms, if properly used, provide effective protection against the transmission and acquisition of gonorrhea as well as other infections that are transmitted to and from genital mucosal surfaces. Spermicidal

preparations used with a diaphragm or cervical sponges impregnated with nonoxynol-9 offer some protection against gonorrhea and chlamydial infection. However, the frequent use of preparations that contain nonoxynol-9 is associated with mucosal disruption that paradoxically may enhance the risk of HIV infection in the event of exposure. All patients should be instructed to refer sex partners for evaluation and treatment. All sex partners of persons with gonorrhea should be evaluated and treated for *N. gonorrhoeae* and *C. trachomatis* infections if their last contact with the patient took place within 60 days before the onset of symptoms or the diagnosis of infection in the patient. If the patient's last sexual encounter was >60 days before onset of symptoms or diagnosis, the patient's most recent sex partner should be treated. Partner-delivered medications or prescriptions for medications to treat gonorrhea and chlamydial infection diminish the likelihood of reinfection (or relapse) in the infected patient. In states where it is not prohibited, this approach is an option for partner management. Patients should be instructed to abstain from sexual intercourse until therapy is completed and until they and their sex partners no longer have symptoms. Greater emphasis must be placed on prevention by public health education, individual patient counseling, and behavior modification, particularly the use of condoms. Sexually active persons, especially adolescents, should be offered screening for STIs. For male patients, NAAT of urine or a urethral swab may be used for screening. Preventing the spread of gonorrhea may help reduce the transmission of HIV. No effective vaccine for gonorrhea is yet available, but efforts to test several candidates are underway.

FURTHER READING

Bolan GA et al: The emerging threat of untreatable gonococcal infection. N Engl J Med 366:485, 2012.

Centers for Disease Control and Prevention: Sexually transmitted disease treatment guidelines, 2015. Morb Mortal Wkly Rep 64(RR3):1, 2015 (*www.cdc.gov/mmwr/preview/mmwrhtml/rr6403a1.htm*).

Golden MR et al: Effect of expedited treatment of sex partners on recurrent or persistent gonorrhea or chlamydial infections. N Engl J Med 352:676, 2005.

Rice PA: Gonococcal arthritis (disseminated gonococcal infection). Infect Dis Clin North Am 19:853, 2005.

Unemo M, Shafer WM: Antimicrobial resistance in *Neisseria gonorrhoeae* in the 21st century: Past, evolution, and future. Clin Microbiol Rev 27:587, 2014.

Wetzler LM et al: Summary and recommendations from the National Institute of Allergy and Infectious Diseases (NIAID) workshop on "Gonorrhea Vaccines: the Way Forward." Clin Vaccine Immunol 23:656, 2016.

Chapter 152: *Haemophilus* and *Moraxella* Infections

Timothy F. Murphy

HAEMOPHILUS INFLUENZAE

MICROBIOLOGY

Haemophilus influenzae was first recognized in 1892 by Pfeiffer, who erroneously concluded that the bacterium was the cause of influenza. *H. influenzae* is a small (1- × 0.3-μm) gram-negative organism of variable shape; thus, it is often described as a pleomorphic coccobacillus. In clinical specimens such as cerebrospinal fluid (CSF) and sputum, *H. influenzae* frequently stains only faintly with safranin and therefore can easily be overlooked.

H. influenzae grows both aerobically and anaerobically. Its aerobic growth requires two factors: hemin (X factor) and nicotinamide adenine dinucleotide (V factor). These requirements are used in the clinical laboratory to identify the bacterium. Caution must be used to distinguish *H. influenzae* from *Haemophilus haemolyticus*, a respiratory tract commensal that has identical growth requirements. *H. haemolyticus* has classically been distinguished from *H. influenzae* by the hemolysis of the former species on horse blood agar. However, a significant proportion of isolates of *H. haemolyticus* have now been recognized as nonhemolytic. Analysis of various genotypic and phenotypic markers, including16S ribosomal sequences, superoxide dismutase, outer-membrane protein P6, protein D, and fuculose kinase, can be used to distinguish these two species.

Six major serotypes of *H. influenzae* have been identified; designated *a* through *f*, they are based on antigenically distinct polysaccharide capsules. In addition, some strains lack a polysaccharide capsule and are referred to as *nontypable* strains. Type b and nontypable strains are the most relevant strains clinically (Table 152-1), although encapsulated strains other than type b can cause disease. *H. influenzae* was the first free-living organism to have its entire genome sequenced.

TABLE 152-1

Characteristics of Type b and Nontypable Strains of *Haemophilus influenzae*

Feature	Type b Strains	Nontypable Strains
Capsule	Ribosyl-ribitol phosphate	Unencapsulated
Pathogenesis	Invasive infections due to hematogenous spread	Mucosal infections due to contiguous spread
Clinical manifestations	Meningitis and invasive infections in incompletely immunized infants and children	Otitis media in infants and children; lower respiratory tract infections in adults with chronic bronchitis
Evolutionary history	Basically clonal	Genetically diverse
Vaccine	Highly effective conjugate vaccines	Under development

The antigenically distinct type b capsule is a linear polymer composed of ribosyl-ribitol phosphate. Strains of *H. influenzae* type b (Hib) cause disease primarily in infants and children <6 years of age. Nontypable strains are primarily mucosal pathogens but occasionally cause invasive disease.

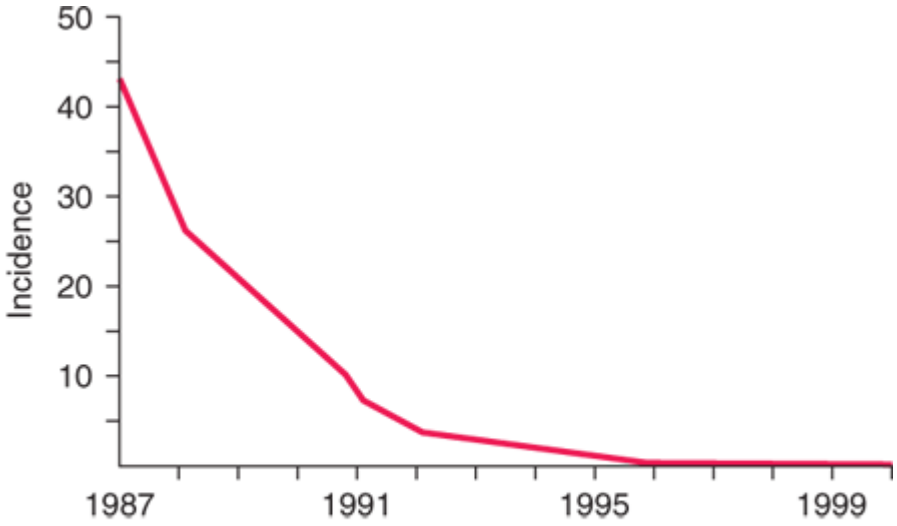
EPIDEMIOLOGY AND TRANSMISSION

H. influenzae, an exclusively human pathogen, is spread by airborne droplets or by direct contact with secretions or fomites. Colonization with nontypable *H. influenzae* is a dynamic process; new strains are acquired and other strains are replaced periodically.

The widespread use of Hib conjugate vaccines in many industrialized countries has resulted in striking decreases in the rate of nasopharyngeal colonization by Hib and in the incidence of Hib infection **(Fig. 152-1)**. Worldwide, invasive Hib disease occurs predominantly in unimmunized children and in those who have not completed the primary immunization series. Of 194 World Health Organization member countries, 99% have introduced Hib conjugate vaccination, but a large number of the world's children remain unimmunized. Certain groups have a higher incidence of invasive Hib disease than the general population, including African-American and Australian Aboriginal children and Native American groups. Although this increased incidence has not yet been accounted for, several factors may be relevant, including age at exposure to the bacterium, socioeconomic conditions, and genetic differences.

FIGURE 152-1

Estimated incidence (rate per 100,000) of invasive disease due to *Haemophilus influenzae* type b among children <5 years of age: 1987–2000. Fewer than 40 cases per year have been reported since 2000. *(Data from the Centers for Disease Control and Prevention.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOGENESIS

Hib strains cause systemic disease by invasion and hematogenous spread from the respiratory tract to distant sites such as the meninges, bones, and joints. The type b polysaccharide capsule is an important virulence factor affecting the bacterium's ability to avoid opsonization and cause systemic disease.

Nontypable strains cause disease by local invasion of mucosal surfaces. Otitis media results when bacteria reach the middle ear by way of the eustachian tube. Adults with chronic bronchitis experience recurrent lower respiratory tract infection due to nontypable strains. In addition, persistent nontypable *H. influenzae* colonization of the lower airways of adults with chronic obstructive pulmonary disease (COPD) contributes to the airway inflammation that is a hallmark of the disease. Nontypable strains that cause infection in adults with COPD differ in pathogenic potential and genome content from strains that cause otitis media. In the middle ear, nontypable strains form biofilms. More resistant to host clearance mechanisms and to antibiotics than are planktonic bacteria, biofilms are associated with chronic and recurrent otitis media. The incidence of invasive disease caused by nontypable strains is low. Strains that cause invasive disease are genetically and phenotypically diverse.

IMMUNE RESPONSE

Antibody to the capsule is important in protection from infection by Hib strains. The level of (maternally acquired) serum antibody to the capsular polysaccharide, which is a polymer of polyribitol ribose phosphate (PRP), declines from birth to 6 months of age and, in the absence of vaccination, remains low until ~2 or 3 years of age. The age at the antibody nadir correlates with that of the peak incidence of type b disease. Antibody to PRP then appears partly as a result of exposure to Hib or cross-reacting antigens. Systemic Hib

disease is unusual after the age of 6 years because of the presence of protective antibody. Vaccines in which PRP is conjugated to protein carrier molecules have been developed and are now used widely. These vaccines generate an antibody response to PRP in infants and effectively prevent invasive infections in infants and children.

Since nontypable strains lack a capsule, the immune response to infection is directed at noncapsular antigens. These antigens have generated considerable interest as immune targets and potential vaccine components. The human immune response to nontypable strains appears to be strain-specific, a characteristic that accounts in part for the propensity of these strains to cause recurrent otitis media and recurrent exacerbations of chronic bronchitis in immunocompetent hosts.

CLINICAL MANIFESTATIONS

Hib

The most serious manifestation of infection with Hib is *meningitis* (Chap. 133), which primarily affects children <2 years of age. The clinical manifestations of Hib meningitis are similar to those of meningitis caused by other bacterial pathogens. Fever and altered central nervous system function are the most common features at presentation. Nuchal rigidity may or may not be evident. Subdural effusion, the most common complication, is suspected when, despite 2 or 3 days of appropriate antibiotic therapy, the infant has seizures, hemiparesis, or continued obtundation. The overall mortality rate from Hib meningitis is ~5%, and the morbidity rate is high. Of survivors, 6% have permanent sensorineural hearing loss, and about one-fourth have a significant handicap of some type. If more subtle handicaps are sought, up to half of survivors are found to have some neurologic sequelae, such as partial hearing loss and delayed language development.

Epiglottitis (Chap. 31) is a life-threatening Hib infection involving cellulitis of the epiglottis and supraglottic tissues. It can lead to acute upper-airway obstruction. Its unique epidemiologic features are its occurrence in an older age group (2–7 years old) than other Hib infections and its absence among Navajo Native Americans and Alaskan Eskimos. Sore throat and fever rapidly progress to dysphagia, drooling, and airway obstruction. Epiglottitis also occurs in adults.

Cellulitis (Chap. 124) due to Hib occurs in young children. The most common location is on the head or neck, and the involved area sometimes takes on a characteristic bluish-red color. Most patients have bacteremia, and 10% have an additional focus of infection.

Hib causes *pneumonia* in infants. The infection is clinically indistinguishable from other types of bacterial pneumonia (e.g., pneumococcal pneumonia) except that Hib is more likely to involve the pleura. Several less common invasive conditions can be important clinical manifestations of Hib infection in children. These include osteomyelitis, septic arthritis, pericarditis, orbital cellulitis, endophthalmitis, urinary tract infection, abscesses, and bacteremia without an identifiable focus.

Non–type b encapsulated strains of *H. influenzae* (types a, c, d, e, and f) are unusual causes of invasive infection manifested predominantly by bacteremia and pneumonia. *H. influenzae* type a infections are seen with increased frequency in indigenous populations of North America. Most infections due to non–type b encapsulated strains occur in the setting of underlying conditions.

Nontypable *H. influenzae*

Nontypable *H. influenzae* is the most common bacterial cause of exacerbations of COPD; these exacerbations are characterized by increased cough, sputum production, and shortness of breath. Fever is low-grade, and no infiltrates are evident on chest x-ray. Nontypable strains also cause community-acquired bacterial pneumonia in adults, especially among patients with COPD or AIDS. The clinical features of *H. influenzae* pneumonia are similar to those of other types of bacterial pneumonia, including pneumococcal pneumonia.

Nontypable *H. influenzae* is one of the three most common causes of childhood otitis media (the other two being *Streptococcus pneumoniae* and *Moraxella catarrhalis*) **(Chap. 31)**. Infants are febrile and irritable, while older children report ear pain. Symptoms of viral upper-respiratory infection often precede otitis media. The diagnosis is made by pneumatic otoscopy. An etiologic diagnosis, although not routinely sought, can be established by tympanocentesis and culture of middle-ear fluid. Clinical features associated with *H. influenzae* otitis media include a history of recurrent episodes, treatment failure, concomitant conjunctivitis, bilateral otitis media, and recent antimicrobial therapy. The increasing use of pneumococcal polysaccharide conjugate vaccines in infants is resulting in a relative increase in the proportion of otitis media cases that are caused by *H. influenzae*.

Nontypable *H. influenzae* also causes puerperal sepsis and is an important cause of neonatal bacteremia. These nontypable strains, which are closely related to *H. haemolyticus*, tend to be of biotype IV and cause invasive disease after colonizing the female genital tract.

Nontypable *H. influenzae* causes sinusitis **(Chap. 31)** in adults and children. In addition, the bacterium is a less common cause of various invasive infections. These infections include empyema, adult epiglottitis, pericarditis, cellulitis, septic arthritis, osteomyelitis, endocarditis, cholecystitis, intraabdominal infections, urinary tract infections, mastoiditis, aortic graft infection, and bacteremia without a detectable focus. While most *H. influenzae* invasive infections in countries where Hib vaccines are used widely are caused by nontypable strains, there is no convincing evidence of an increased incidence of infection by nontypable *H. influenzae* as a result of the use of Hib vaccines. Continued monitoring will be important. Many patients with *H. influenzae* bacteremia have an underlying condition, such as HIV infection, cardiopulmonary disease, alcoholism, or cancer.

DIAGNOSIS

The most reliable method for establishing a diagnosis of Hib infection is recovery of the organism in culture. The presence of gram-negative coccobacilli in Gram-stained CSF is strong evidence for Hib meningitis. Recovery of the organism from CSF confirms the diagnosis. Cultures of other normally sterile body fluids,

such as blood, joint fluid, pleural fluid, pericardial fluid, and subdural effusion, are confirmatory in other infections.

Detection of PRP is an important adjunct to culture in rapid diagnosis of Hib meningitis. Immunoelectrophoresis, latex agglutination, coagglutination, and enzyme-linked immunosorbent assay are effective in detecting PRP. These assays are particularly helpful when patients have received prior antimicrobial therapy and thus are especially likely to have negative cultures.

Because nontypable *H. influenzae* is primarily a mucosal pathogen, it is a component of a mixed flora; thus etiologic diagnosis is challenging. Nontypable *H. influenzae* infection is strongly suggested by the predominance of gram-negative coccobacilli among abundant polymorphonuclear leukocytes in a Gram-stained sputum specimen from a patient in whom pneumonia is suspected. Although bacteremia is detectable in a small proportion of patients with pneumonia due to nontypable *H. influenzae*, most such patients have negative blood cultures.

A diagnosis of otitis media is based on the detection by pneumatic otoscopy of fluid in the middle ear. An etiologic diagnosis requires tympanocentesis but is not routinely sought. An invasive procedure is also required to determine the etiology of sinusitis; thus, treatment is often empirical once the diagnosis is suspected in light of clinical symptoms and sinus radiographs.

TREATMENT

TREATMENT

Haemophilus influenzae

Initial therapy for meningitis due to Hib should consist of a cephalosporin such as ceftriaxone or cefotaxime. For children, the dosage of ceftriaxone is 75–100 mg/kg daily given in two doses 12 h apart. The pediatric dosage of cefotaxime is 200 mg/kg daily given in four doses 6 h apart. Adult dosages are 2 g every 12 h for ceftriaxone and 2 g every 4–6 h for cefotaxime. An alternative regimen for initial therapy is ampicillin (200–300 mg/kg daily in four divided doses) plus chloramphenicol (75–100 mg/kg daily in four divided doses). Therapy should continue for a total of 1–2 weeks.

Administration of glucocorticoids to patients with Hib meningitis reduces the incidence of neurologic sequelae. The presumed mechanism is reduction of the inflammation induced by bacterial cell-wall mediators of inflammation when cells are killed by antimicrobial agents. Dexamethasone (0.6 mg/kg per day intravenously in four divided doses for 2 days) is recommended for the treatment of Hib meningitis in children >2 months of age.

Invasive infections other than meningitis are treated with the same antimicrobial agents. For epiglottitis, the dosage of ceftriaxone is 50 mg/kg daily, and the dosage of cefotaxime is 150 mg/kg daily, given in three divided doses 8 h apart. Epiglottitis constitutes a medical emergency, and maintenance of an airway is

critical. The duration of therapy is determined by the clinical response. A course of 1–2 weeks is usually appropriate.

Many infections caused by nontypable strains of *H. influenzae*, such as otitis media, sinusitis, and exacerbations of COPD, can be treated with oral antimicrobial agents. Approximately 20–35% of nontypable strains produce β-lactamase (with the exact proportion depending on geographic location), and these strains are resistant to ampicillin. Several agents have excellent activity against nontypable *H. influenzae*, including amoxicillin/clavulanic acid, various extended-spectrum cephalosporins, and the macrolides azithromycin and clarithromycin. Fluoroquinolones are highly active against *H. influenzae* and are useful in adults with exacerbations of COPD. However, fluoroquinolones are not currently recommended for the treatment of children or pregnant women because of possible effects on articular cartilage.

 In addition to β-lactamase production, alteration of penicillin-binding proteins—a second mechanism of ampicillin resistance—has been detected in isolates of *H. influenzae*. Although rare in the United States, these β-lactamase-negative ampicillin-resistant strains are common in Japan and are increasing in prevalence in Europe. Resistance to macrolides is also being observed with increasing frequency in many regions of the world. Continued monitoring of the evolving antimicrobial susceptibility patterns of *H. influenzae* will be important.

PREVENTION

Vaccination

 Three conjugate vaccines that prevent invasive infections with Hib in infants and children are licensed in the United States **(See also Chap. 118)**. In addition to eliciting protective antibody, these vaccines prevent disease by reducing rates of pharyngeal colonization with Hib. The widespread use of conjugate vaccines has dramatically reduced the incidence of Hib disease in developed countries. Even though the manufacture of Hib vaccines is costly, vaccination is cost-effective. The Global Alliance for Vaccines and Immunizations has recognized the underuse of Hib conjugate vaccines.

The disease burden has been reduced in developing countries that have implemented routine vaccination (e.g., The Gambia, Chile). An important obstacle to more widespread vaccination is the lack of data on the epidemiology and burden of Hib disease in many developing countries.

All children should be immunized with an Hib conjugate vaccine, receiving the first dose at ~2 months of age, the rest of the primary series at 2–6 months of age, and a booster dose at 12–15 months of age. Specific recommendations vary for the different conjugate vaccines. The reader is referred to the recommendations of the American Academy of Pediatrics (**Chap. 118** and *www.cispimmunize.org*).

Currently, no vaccines are available specifically for the prevention of disease caused by nontypable *H. influenzae*. However, a vaccine that contains protein D—a surface protein of *H. influenzae*—conjugated to pneumococcal polysaccharides is licensed in other countries and is used widely throughout the world. The

vaccine has shown partial efficacy in preventing *H. influenzae* otitis media in clinical trials. Additional progress in the development of vaccines against nontypable *H. influenzae* is anticipated.

Chemoprophylaxis

The risk of secondary disease is greater than normal among household contacts of patients with Hib disease. Therefore, all children and adults (except pregnant women) in households with an index case and at least one incompletely immunized contact <4 years of age should receive prophylaxis with oral rifampin. When two or more cases of invasive Hib disease have occurred within 60 days at a child-care facility attended by incompletely vaccinated children, administration of rifampin to all attendees and personnel is indicated, as it is for household contacts. Chemoprophylaxis is not indicated in nursery and child-care contacts of a single index case. The reader is referred to the recommendations of the American Academy of Pediatrics.

HAEMOPHILUS DUCREYI

Haemophilus ducreyi is the etiologic agent of chancroid **(Chap. 131)**, a sexually transmitted disease characterized by genital ulceration and inguinal adenitis. In addition to being a cause of morbidity in itself, chancroid is associated with HIV infection because of the role played by genital ulceration in HIV transmission. Chancroid increases the efficiency of transmission of and the degree of susceptibility to HIV infection. *H. ducreyi* has also been recognized as an important cause of non-sexually transmitted cutaneous ulcers.

MICROBIOLOGY

H. ducreyi is a highly fastidious coccobacillary gram-negative bacterium whose growth requires X factor (hemin). Although, in light of this requirement, the bacterium has been classified in the genus *Haemophilus*, DNA homology and chemotaxonomic studies have established substantial differences between *H. ducreyi* and other *Haemophilus* species. Taxonomic reclassification of the organism is likely in the future but awaits further study. Ulcers contain predominantly T cells. The fact that patients who have had chancroid may have repeated infections indicates that infection does not confer protection.

EPIDEMIOLOGY AND PREVALENCE

The prevalence of chancroid has declined in the United States and worldwide. However, prevalence data must be interpreted with caution because of the difficulty of establishing a diagnosis. The infection appears to be more common in developing countries. Transmission is predominantly heterosexual, and cases in males have outnumbered those in females by ratios of 3:1 to 25:1 during outbreaks. Contact with commercial sex workers and illicit drug use are strongly associated with chancroid.

H. ducreyi has emerged as a major cause of cutaneous ulcers in the South Pacific and Africa. Strains that cause cutaneous ulcers have genome sequences that are nearly identical to class I strains (of two related classes) of *H. ducreyi* that cause genital ulcers.

CLINICAL MANIFESTATIONS AND DIFFERENTIAL DIAGNOSIS

Infection is acquired as the result of a break in the epithelium during sexual contact with an infected individual. After an incubation period of 4–7 days, the initial lesion—a papule with surrounding erythema—appears. In 2 or 3 days, the papule evolves into a pustule, which spontaneously ruptures and forms a sharply circumscribed ulcer that generally is not indurated **(Fig. 152-2)**. The ulcers are painful and bleed easily; little or no inflammation of the surrounding skin is evident. Approximately half of patients develop enlarged, tender inguinal lymph nodes, which frequently become fluctuant and spontaneously rupture. Patients usually seek medical care after 1–3 weeks of painful symptoms.

FIGURE 152-2

Chancroid with characteristic penile ulcers and associated left inguinal adenitis (bubo).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The presentation of chancroid does not usually include all of the typical clinical features and is sometimes atypical. Multiple ulcers can coalesce to form giant ulcers. Ulcers can appear and then resolve, with inguinal adenitis (Fig. 152-2) and suppuration following 1–3 weeks later; this clinical picture can be confused with that of lymphogranuloma venereum **(Chap. 184)**. Multiple small ulcers can resemble folliculitis. Other differential diagnostic considerations include the various infections causing genital ulceration, such as primary syphilis, secondary syphilis (condyloma latum), genital herpes, and donovanosis. In rare cases, chancroid lesions become secondarily infected with bacteria; the result is extensive inflammation.

Non-sexually transmitted cutaneous ulcers caused by *H. ducreyi* resemble those of yaws caused by *Treponema pallidum* subspecies *pertenue*, which is endemic in regions where *H. ducreyi* cutaneous ulcers are seen. Ulcers caused by *H. ducreyi* are less likely than those of yaws to show central granulating tissue and less likely to have indurated edges, but substantial overlap in clinical characteristics exists.

DIAGNOSIS

Clinical diagnosis of chancroid is often inaccurate, and laboratory confirmation should be attempted in suspected cases. An accurate diagnosis of chancroid relies on culture of *H. ducreyi* from the lesion or from an aspirate of suppurative lymph nodes. Since the organism can be difficult to grow, the use of selective and supplemented media is necessary. No polymerase chain reaction (PCR) assay for *H. ducreyi* is commercially available; such tests can be performed by Clinical Laboratory Improvement Amendment (CLIA)–certified clinical laboratories that have developed their own assays.

A probable diagnosis of chancroid can be made when the following criteria are met: (1) one or more painful genital ulcers; (2) no evidence of *T. pallidum* infection by dark-field examination of ulcer exudate or by a negative serologic test for syphilis performed at least 7 days after ulcer onset; (3) a typical clinical presentation for chancroid; and (4) a negative test for herpes simplex virus in the ulcer exudate.

A serologic test for syphilis does not distinguish cutaneous ulcers due to *H. ducreyi* from those due to yaws. A PCR assay has been used in clinical studies to establish an *H. ducreyi* etiology, but, as stated above, no such assay is commercially available.

TREATMENT

TREATMENT

Haemophilus ducreyi

Treatment regimens recommended by the Centers for Disease Control and Prevention include (1) a single 1-g oral dose of azithromycin; (2) ceftriaxone (250 mg intramuscularly in a single dose); (3) ciprofloxacin (500 mg by mouth twice a day for 3 days); and (4) erythromycin base (500 mg by mouth three times a day for 7 days). Isolates from patients who do not respond promptly to treatment should be tested for antimicrobial resistance. In patients with HIV infection, healing may be slow and longer courses of treatment may be necessary. Clinical treatment failure in HIV-seropositive patients may reflect co-infection, especially with herpes simplex virus. Contacts of patients with chancroid should be identified and treated, whether or not symptoms are present, if they have had sexual contact with the patient during the 10 days preceding the patient's onset of symptoms.

MORAXELLA CATARRHALIS

MICROBIOLOGY

M. catarrhalis is an unencapsulated gram-negative diplococcus whose ecologic niche is the human respiratory tract. The organism was initially designated *Micrococcus catarrhalis*. Its name was changed to *Neisseria catarrhalis* in 1970 because of phenotypic similarities to commensal *Neisseria* species. On the basis of more rigorous analysis of genetic relatedness, *Moraxella catarrhalis* is now the widely accepted name for this species.

EPIDEMIOLOGY

Nasopharyngeal colonization by *M. catarrhalis* is common in infancy, with colonization rates ranging between 33% and 100% and depending on geographic location. Several factors probably account for this geographic variation, including living conditions, day-care attendance, hygiene, household smoking, and population genetics. The prevalence of colonization decreases steadily with age.

The widespread use of pneumococcal conjugate vaccines in some countries has resulted in alterations in patterns of nasopharyngeal colonization in resident populations. A relative increase in colonization by nonvaccine pneumococcal serotypes, nontypable *H. influenzae*, and *M. catarrhalis* has occurred. These changes in colonization patterns may be altering the distribution of pathogens of both otitis media and sinusitis in children.

PATHOGENESIS

M. catarrhalis causes mucosal infections of the respiratory tract by contiguous spread from its colonizing site in the upper airway. A preceding viral upper respiratory tract infection is a common inciting event for otitis media. In exacerbations of COPD, the acquisition of new strains is critical for pathogenesis. Strains exhibit substantial genetic diversity and differences in virulence properties.

The expression of several adhesin molecules with differing specificities for various host cell receptors reflects the importance of adherence to the respiratory epithelial surface in the pathogenesis of infection. *M. catarrhalis* invades multiple cell types. Its intracellular residence in lymphoid tissue provides a potential reservoir for persistence in the human respiratory tract. Like many gram-negative bacteria, *M. catarrhalis* sheds vesicles into the surrounding environment. The vesicles are internalized by host cells and mediate several virulence mechanisms, including induction of inflammation and delivery of β-lactamase, that can promote the survival of co-pathogens.

CLINICAL MANIFESTATIONS

In children, *M. catarrhalis* causes predominantly mucosal infections when the bacterium migrates from the nasopharynx to the middle ear or the sinuses (Chap. 31). The inciting event for both otitis media and sinusitis is often a preceding viral infection. Overall, cultures of middle-ear fluid obtained by tympanocentesis indicate that *M. catarrhalis* causes 15–20% of cases of acute otitis media. Acute otitis media caused by *M. catarrhalis* or nontypable *H. influenzae* is clinically milder than otitis media caused by *S. pneumoniae*, with less fever

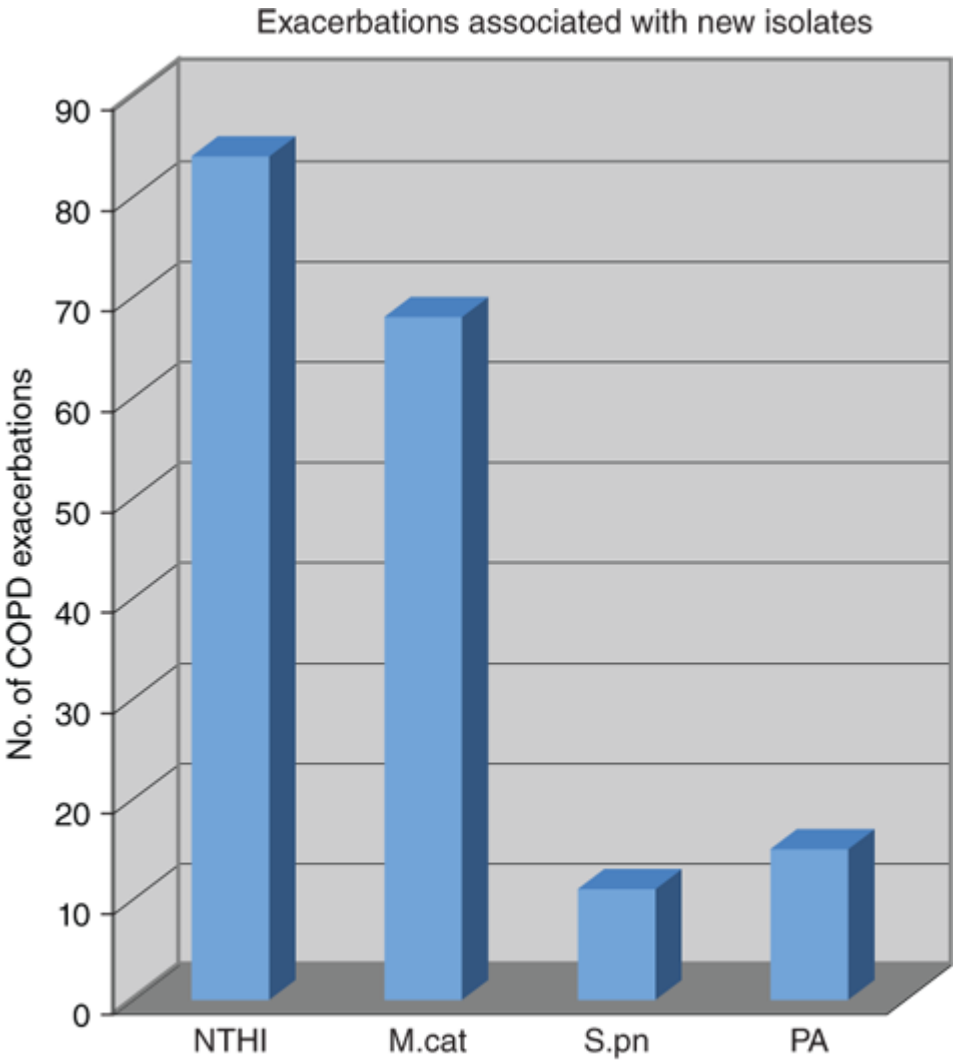
and a lower prevalence of a red bulging tympanic membrane. However, substantial overlap makes it impossible to predict etiology in an individual child on the basis of clinical features.

A small proportion of viral upper respiratory tract infections are complicated by bacterial sinusitis. Cultures of sinus puncture aspirates show that *M. catarrhalis* accounts for ~20% of cases of acute bacterial sinusitis in children and for a smaller proportion in adults.

M. catarrhalis is a common cause of exacerbations in adults with COPD. The bacterium has been overlooked in this clinical setting because it has long been considered to be a commensal and because it is easily mistaken for commensal *Neisseria* species in cultures of respiratory secretions (see "Diagnosis," below). Several independent lines of evidence have established *M. catarrhalis* as a pathogen in COPD. These include (1) the demonstration of *M. catarrhalis* in the lower airways during exacerbations, (2) the association of exacerbation with acquisition of new strains, (3) elevations of inflammatory markers in association with *M. catarrhalis*, and (4) the development of specific immune responses following infection. *M. catarrhalis* is the second most common bacterial cause of COPD exacerbations (after *H. influenzae*), as shown in a 10-year prospective study; the distribution of exacerbations associated with new-strain acquisitions is shown in **Fig. 152-3**. Not included are culture-negative cases or cases from which a pathogen had been previously isolated. With the application of rigorous clinical criteria for defining the etiology of exacerbations (both culture-positive and culture-negative), ~10% of all exacerbations in the same study were caused by *M. catarrhalis*. The clinical features of an exacerbation due to *M. catarrhalis* are similar to those of exacerbations due to other bacterial pathogens, including *H. influenzae* and *S. pneumoniae*. The cardinal symptoms are cough with increased sputum production, sputum purulence, and dyspnea in comparison with baseline symptoms.

FIGURE 152-3

Cumulative results of a prospective study (1994–2004) of bacterial infection in chronic obstructive pulmonary disease (COPD) showing etiology of exacerbations. The numbers of exacerbations shown indicate the acquisition of a new strain simultaneous with clinical symptoms of an exacerbation. NTHI, nontypable *H. influenzae*; M.cat, *M. catarrhalis*; S.pn, *Streptococcus pneumoniae*; PA, *Pseudomonas aeruginosa*. *(Adapted from TF Murphy, GI Parameswaran: Clin Infect Dis 49:124, 2009, with permission. © 2009 Infectious Diseases Society of America.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pneumonia due to *M. catarrhalis* occurs in the elderly, particularly in the setting of underlying cardiopulmonary disease, but is infrequent. Invasive infections, such as bacteremia, endocarditis, neonatal meningitis, and septic arthritis, are rare.

DIAGNOSIS

Tympanocentesis is required for etiologic diagnosis of otitis media, but this procedure is not performed routinely. Therefore, treatment of otitis media is generally empirical. Similarly, an etiologic diagnosis of sinusitis requires an invasive procedure and thus is usually not available to the clinician. Isolation of *M. catarrhalis* from an expectorated sputum sample from an adult experiencing clinical symptoms of an exacerbation is suggestive, but not diagnostic, of *M. catarrhalis* as the cause.

Upon culture, colonies of *M. catarrhalis* resemble those of commensal neisseriae that are part of the normal upper airway flora. As mentioned above, the difficulty in distinguishing colonies of *M. catarrhalis* from neisserial colonies in cultures of respiratory secretions explains in part why *M. catarrhalis* has been overlooked as a pathogen. In contrast to these *Neisseria* species, *M. catarrhalis* colonies can be slid across the agar surface without disruption (the “hockey puck sign”). In addition, after 48 h of growth, *M. catarrhalis*

colonies take on a pink color and tend to be larger than neisserial colonies. A variety of biochemical tests can distinguish *M. catarrhalis* from neisseriae. Kits that rely on these biochemical reactions are commercially available.

TREATMENT

TREATMENT

Moraxella catarrhalis

M. catarrhalis rapidly acquired β-lactamases during the 1970s and 1980s; antimicrobial susceptibility patterns have remained relatively stable since that time, with >90% of strains now producing β-lactamase and thus resistant to amoxicillin. Otitis media in children and exacerbations of COPD in adults are generally managed empirically with antimicrobial agents that are active against *S. pneumoniae, H. influenzae*, and *M. catarrhalis*. Most strains of *M. catarrhalis* are susceptible to amoxicillin/clavulanic acid, extended-spectrum cephalosporins, newer macrolides (azithromycin, clarithromycin), trimethoprim-sulfamethoxazole, and fluoroquinolones. However, recent reports from several centers in Asia show substantial resistance to macrolides and fluoroquinolones, indicating emerging resistance. Continued monitoring of global antimicrobial susceptibility patterns of *M. catarrhalis* will be critical.

FURTHER READING

Duell BL et al: Host–pathogen interactions of nontypeable *Haemophilus influenzae*: From commensal to pathogen. FEBS Lett 590:3840, 2016.
CrossRef

González-Beiras C et al: Epidemiology of *Haemophilus ducreyi* infections. Emerg Infect Dis 22:1, 2016.

Murphy TF: Vaccines for nontypeable *Haemophilus influenzae*: The future is now. Clin Vaccine Immunol 5:459, 2015.

Murphy TF, Parameswaran GI: *Moraxella catarrhalis*, a human respiratory tract pathogen. Clin Infect Dis 49:124, 2009.

Su YC et al: *Moraxella catarrhalis*: From interactions with the host immune system to vaccine development. Future Microbiol 7:1073, 2012.

Chapter 153: Infections Due to the HACEK Group and Miscellaneous Gram-Negative Bacteria

Tamar F. Barlam; Dennis L. Kasper

THE HACEK GROUP

HACEK organisms are a group of fastidious, slow-growing, gram-negative bacteria whose growth requires an atmosphere of carbon dioxide. These organisms do not grow on media routinely used for enteric bacteria (e.g., MacConkey agar). Species belonging to this group include several *Haemophilus* species, *Aggregatibacter* (formerly *Actinobacillus*) species, *Cardiobacterium* species, *Eikenella corrodens*, and *Kingella kingae*. HACEK bacteria normally reside in the oral cavity and have been associated with local infections in the mouth. They are also known to cause severe systemic infections—most often bacterial endocarditis, which can develop on either native or prosthetic valves (Chap. 123). HACEK bacteremia is strongly predictive of underlying infective endocarditis (overall positive predictive value, 60%). However, this association varies significantly by organism. For example, in one study, infective endocarditis was diagnosed in 100% of patients with *Aggregatibacter actinomycetemcomitans* bacteremia but in no patients with *Eikenella* bacteremia.

In large series, 0.8–6% of cases of infective endocarditis are attributable to HACEK organisms, most often *Aggregatibacter* species, *Haemophilus* species, and *Cardiobacterium hominis.* Invasive infection typically occurs in patients with a history of cardiac valvular disease or prosthetic valves, often in the setting of a recent dental procedure or nasopharyngeal infection. The aortic and mitral valves are most commonly affected. The clinical course of HACEK endocarditis tends to be subacute, particularly with *Aggregatibacter* or *Cardiobacterium*. However, *K. kingae* endocarditis may have a more aggressive presentation. Compared with non-HACEK endocarditis, HACEK endocarditis occurs in younger patients and is more frequently associated with embolic, vascular, and immunologic manifestations. Systemic embolization is common. The overall prevalence of major emboli associated with HACEK endocarditis ranges from 28 to 71% in different series. On echocardiography, valvular vegetations are seen in up to 85% of patients. *Aggregatibacter* and *Haemophilus* species cause mitral valve vegetations most often; *Cardiobacterium* is associated with aortic valve vegetations.

The microbiology laboratory should be alerted when a HACEK organism is being considered. Most cultures that ultimately yield a HACEK organism become positive within the first week, especially with improved culture systems such as BACTEC. Studies have not shown that prolonged incubation increases laboratory

recovery of clinically significant HACEK isolates. Polymerase chain reaction (PCR) techniques, such as gene amplification of 16S rRNA, can facilitate the diagnosis of HACEK infection of blood or cardiac valves. Other tools, such as matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry performed directly on agar colonies, can increase the accuracy and speed of diagnosis of HACEK infections.

Because of HACEK organisms' slow growth, antimicrobial susceptibility testing may be difficult, and β-lactamase production may not be detected. Resistance is most commonly noted in *Haemophilus* and *Aggregatibacter* species. Etest methodology may increase the accuracy of susceptibility testing. In recent studies, ceftriaxone and levofloxacin have been active against all isolates. The overall prognosis in both native-valve and prosthetic-valve HACEK endocarditis is excellent and is significantly better than that in endocarditis caused by non-HACEK pathogens.

Haemophilus Species

Haemophilus parainfluenzae is the most common *Haemophilus* species isolated from cases of HACEK endocarditis. Of patients with HACEK endocarditis due to *Haemophilus* species, 60% have been ill for <2 months before presentation, and 19–50% develop congestive heart failure. Mortality rates as high as 30–50% were reported in older series; however, more recent studies have documented mortality rates of <5%. *H. parainfluenzae* has been isolated from other infections, such as meningitis; brain, dental, pelvic, and liver abscess; pneumonia; urinary tract infection; and septicemia.

Aggregatibacter Species

Aggregatibacter species are the most common cause of HACEK endocarditis; the species most frequently involved are *A. actinomycetemcomitans*, *A.* (formerly *Haemophilus*) *aphrophilus*, and *A. paraphrophilus*. *Aggregatibacter* is associated with prosthetic-valve endocarditis more often than are *Haemophilus* species. *A. actinomycetemcomitans* can be isolated from soft tissue infections and abscesses in association with *Actinomyces israelii*. Typically, patients who develop *Aggregatibacter* endocarditis have periodontal disease or have recently undergone dental procedures in the setting of underlying cardiac valvular damage. The disease is insidious; patients may be sick for several months before diagnosis. Frequent complications include embolic phenomena, congestive heart failure, and renal failure.

A. actinomycetemcomitans has been isolated from patients with brain abscess, meningitis, endophthalmitis, parotitis, osteomyelitis, urinary tract infection, pneumonia, and empyema, among other infections. *A. aphrophilus* is often associated with bone and joint infection and is an important cause of brain abscess. In one series, *A. aphrophilus* was isolated from brain abscesses in 10% of cases—a rate that is disproportionate to its isolation from oral flora. This species has also been described as a cause of abscess in other organ systems.

Cardiobacterium Species

Cardiobacterium species, most often *C. hominis*, cause endocarditis primarily in patients with underlying valvular heart disease or with prosthetic valves. These organisms most frequently affect the aortic valve. Many patients have signs and symptoms of long-standing infection before diagnosis, with evidence of arterial embolization, vasculitis, cerebrovascular accidents, immune complex glomerulonephritis, or arthritis at presentation. Embolization, mycotic aneurysms, and congestive heart failure are common complications. A second species, *C. valvarum*, has been described in association with endocarditis.

Eikenella corrodens

E. corrodens is most frequently recovered from sites of infection in conjunction with other bacterial species. Clinical sources of *E. corrodens* include sites of human bite wounds (clenched-fist injuries), endocarditis, soft tissue infections, osteomyelitis, head and neck infections, respiratory infections, chorioamnionitis, gynecologic infections associated with intrauterine devices, meningitis, brain abscesses, and visceral abscesses. This organism is the least common cause of HACEK endocarditis.

Kingella kingae

More than half of cases of *K. kingae* infection are bone and joint infections; the majority of the remaining infections are infective endocarditis, bacteremia, and meningitis. Invasive *K. kingae* infections with bacteremia are associated with upper respiratory tract infections and stomatitis in 80% of cases. Rates of oropharyngeal colonization with *K. kingae* are highest in the first 3 years of life (detected in ~10% of children); colonization coincides with an increased incidence of skeletal infections and other invasive infections due to this organism from the age of 6 months to 4 years. *K. kingae* can be transmitted from child to child and has been the cause of outbreaks among young children. *K. kingae* bacteremia can present with a petechial rash similar to that seen in *Neisseria meningitidis* sepsis.

Because of improved microbiologic methodology and molecular methods such as real-time PCR, the isolation of *K. kingae* is increasingly common. Inoculation of clinical specimens (e.g., synovial fluid) into aerobic blood culture bottles enhances recovery of this organism. PCR studies of blood or joint fluid can identify *K. kingae* in culture-negative cases. Some studies have demonstrated that *K. kingae* has surpassed *Staphylococcus aureus* as the leading cause of septic arthritis and osteomyelitis in children.

Infective endocarditis, unlike other infections with *K. kingae*, occurs in older children and adults. The majority of patients have preexisting valvular disease. There is a high incidence of complications, including arterial emboli, cerebrovascular accidents, tricuspid insufficiency, and congestive heart failure with cardiovascular collapse.

TREATMENT

TREATMENT

HACEK Endocarditis

Ceftriaxone (2 g/d) is first-line therapy for HACEK endocarditis. Data on the use of levofloxacin (750 mg/d) for HACEK endocarditis remain limited, but this drug can be considered an alternative for treatment of patients intolerant of β-lactam therapy. Of note, *Eikenella* is resistant to clindamycin, metronidazole, and aminoglycosides **(Table 153-1)**.

Native-valve endocarditis should be treated for 4 weeks with antibiotics, whereas prosthetic-valve endocarditis requires 6 weeks of therapy. The cure rates for HACEK prosthetic-valve endocarditis appear to be high. Unlike prosthetic-valve endocarditis caused by other gram-negative organisms, HACEK endocarditis is often cured with antibiotic treatment alone—i.e., without surgical intervention.

TABLE 153-1

Treatment of Infections Caused by HACEK-Group and Other Fastidious Gram–Negative Organisms

Organisms	Preferred Therapy	Alternative Agents	Comments
Haemophilus spp. *Aggregatibacter* spp. *Cardiobacterium* spp. *Eikenella corrodens* *Kingella kingae*	Ceftriaxone (2 g/d)	Ampicillin/sulbactam (3 g of ampicillin q6h) Levofloxacin (750 mg/d)	Ampicillin/sulbactam resistance has been described in *Haemophilus* and *Aggregatibacter* spp. Data on use of levofloxacin for endocarditis therapy are limited. Fluoroquinolones are not recommended for treatment of patients <18 years of age. Penicillin (16–18 million units q4h) or ampicillin (2 g q4h) can be used if the organism is susceptible. However, because of the slow growth of HACEK bacteria, antimicrobial testing may be difficult, and β-lactamase production may not be detected.
Capnocytophaga spp.	Ampicillin/sulbactam (1.5–3 g of ampicillin q6h)	Ceftriaxone (2 g/d q12–24h)	Penicillin (12–18 million units q4h) should be used if the isolate is known to be susceptible.
Pasteurella multocida	Ampicillin/sulbactam (1.5–3 g of ampicillin q6h)	Ceftriaxone (1–2 g/d q12–24h)	Penicillin should be used if the isolate is known to be susceptible. *P. multocida* is also susceptible to tetracyclines and fluoroquinolones.

OTHER FASTIDIOUS GRAM-NEGATIVE BACTERIA

Capnocytophaga Species

Like HACEK organisms, this genus of fastidious, fusiform, gram-negative coccobacilli is facultatively anaerobic and requires an atmosphere enriched in carbon dioxide for optimal growth. *Capnocytophaga* species such as *C. ochracea*, *C. gingivalis*, *C. haemolytica*, and *C. sputigena* are part of the oral flora; most infections are contiguous with the oropharynx (e.g., periodontal disease, respiratory tract infections, cervical abscesses, endophthalmitis). These organisms have also been associated with sepsis in immunocompromised hosts, particularly neutropenic patients with oral ulcerations, meningitis, endocarditis, cellulitis, osteomyelitis, and septic arthritis. *Capnocytophaga* species have been isolated from many other sites as well, usually as part of a polymicrobial infection. There is a high prevalence of resistance to β-lactams and macrolides in *Capnocytophaga*; the oral cavity serves as a reservoir for resistance genes to those agents.

C. canimorsus and *C. cynodegmi* are endogenous to the canine and feline mouth **(Chap. 136)**. Patients infected with these species frequently have a history of dog or cat bites or of exposure without scratches or bites. Asplenia, glucocorticoid therapy, and alcohol abuse are predisposing conditions that can be associated with severe sepsis with shock and disseminated intravascular coagulation. Patients typically have a petechial rash that can progress from purpuric lesions to gangrene.

TREATMENT

TREATMENT

Capnocytophaga Infections

Because of increasing β-lactamase production, a penicillin derivative plus a β-lactamase inhibitor—such as ampicillin/sulbactam (1.5–3.0 g of ampicillin every 6 h)—is currently recommended for empirical treatment of infections caused by *Capnocytophaga* species (Table 153-1). If the isolate is known to be susceptible, infections with *C. canimorsus* should be treated with penicillin (12–18 million units every 4 h). *Capnocytophaga* is also susceptible to clindamycin (600–900 mg every 6–8 h) and third-generation cephalosporins such as ceftriaxone (2 g every 12–24 h). Antibiotics should be given prophylactically to asplenic patients who have sustained dog-bite injuries.

Pasteurella multocida

P. multocida is a fastidious, bipolar-staining, gram-negative coccobacillus that colonizes the respiratory and gastrointestinal tracts of domestic animals; oropharyngeal colonization rates are 70–90% in cats and 50–65% in dogs. *P. multocida* can be transmitted to humans through bites or scratches, via the respiratory tract from contact with contaminated dust or infectious droplets, or via deposition of the organism on injured skin or mucosal surfaces during licking. Most human infections affect skin and soft tissue; almost two-thirds of these infections are caused by cats. Patients at the extremes of age or with serious underlying disorders (e.g.,

cirrhosis, diabetes) are at increased risk for systemic manifestations, including meningitis, peritonitis, osteomyelitis and septic arthritis, endocarditis, septic shock, and purpura fulminans, and are more likely not to have evidence of an animal bite. However, cases have also occurred in healthy individuals of all ages. If inhaled, *P. multocida* can cause acute respiratory tract infection, particularly in patients with underlying sinus and pulmonary disease.

TREATMENT

TREATMENT

Pasteurella multocida Infections

P. multocida is susceptible to penicillin, ampicillin, ampicillin/sulbactam, second- and third-generation cephalosporins, tetracyclines, and fluoroquinolones (Table 153-1). β-Lactamase-producing strains have been reported.

OTHER GRAM-NEGATIVE BACTERIA

Achromobacter xylosoxidans

Achromobacter (previously *Alcaligenes*) *xylosoxidans* is an aerobic nonfermenting gram-negative organism that is probably part of the endogenous intestinal flora. It has been isolated from a variety of water sources, including well water, IV fluids, and humidifiers. Immunocompromised hosts, including patients with cancer and postchemotherapy neutropenia, cirrhosis, chronic renal failure, and cystic fibrosis, are at increased risk for infection. Nosocomial outbreaks and pseudo-outbreaks of *A. xylosoxidans* infection have been attributed to contaminated fluids, and clinical illness has been associated with isolates from many sites, including blood (often in the setting of intravascular devices). Community-acquired *A. xylosoxidans* bacteremia usually occurs in the setting of pneumonia. Metastatic skin lesions are present in one-fifth of cases. The reported mortality rate is as high as 67%—a figure similar to rates for other bacteremic gram-negative pneumonias.

TREATMENT

TREATMENT

Achromobacter xylosoxidans Infections

Treatment is based on in vitro susceptibility testing of all clinically relevant isolates; multidrug resistance is common **(Table 153-2)**. Carbapenems, tigecycline, and colistin are typically the most active agents.

TABLE 153-2

Treatment Options for Other Selected Gram-Negative Bacteria[a]

Organism	Treatment Options
Achromobacter xylosoxidans	Carbapenems, tigecycline, colistin
Aeromonas spp.	Fluoroquinolones, third- and fourth-generation cephalosporins, carbapenems, aminoglycosides
Elizabethkingia/Chryseobacterium spp.	Fluoroquinolones, minocycline, tigecycline, piperacillin/tazobactam
Rhizobium radiobacter	Fluoroquinolones, third- and fourth-generation cephalosporins, carbapenems
Shewanella spp.	Fluoroquinolones, third- and fourth-generation cephalosporins, β-lactam/β-lactamase inhibitors, carbapenems, aminoglycosides
Chromobacterium violaceum	Carbapenems, fluoroquinolones, trimethoprim-sulfamethoxazole

[a]Treatment should be based on in vitro susceptibility testing; multidrug resistance is common among these organisms.

Aeromonas Species

Aeromonas is a facultative anaerobic gram-negative bacterium. *Aeromonas* infections are most often caused by *A. hydrophila*, *A. caviae*, *A. veronii*, and *A. dhakensis*. *Aeromonas* proliferates in potable water, freshwater, and soil. It remains controversial whether *Aeromonas* is a cause of bacterial gastroenteritis; asymptomatic colonization of the intestinal tract with *Aeromonas* occurs frequently, and no clonally related diarrheal outbreak has been documented. However, rare cases of hemolytic-uremic syndrome following bloody diarrhea have been shown to be secondary to the presence of *Aeromonas*.

Aeromonas causes health care–associated sepsis and bacteremia in infants with multiple medical problems and in immunocompromised hosts, particularly those with cancer or hepatobiliary disease, including cirrhosis. *A. caviae* is associated with health care–related bacteremia. Community-acquired infections include bacteremia, spontaneous bacterial peritonitis, biliary tract infections, and skin and soft tissue infections. Severe soft tissue infections such as necrotizing fasciitis are more common in Taiwan than in Western countries; *Aeromonas* was the most common pathogen associated with skin and soft tissue infections after the tsunami in Thailand. *Aeromonas* infection and sepsis can occur in patients with trauma (including severe trauma with myonecrosis), patients with seawater-contaminated wounds, and burn

patients exposed to the organism by environmental (freshwater or soil) wound contamination. Reported mortality rates range from 25% among immunocompromised adults with sepsis to >90% among patients with myonecrosis. Patients with *A. dhakensis* bacteremia have higher 14-day mortality rates than do those whose bacteremia is attributable to other species. *Aeromonas* can produce ecthyma gangrenosum (hemorrhagic vesicles surrounded by a rim of erythema with central necrosis and ulceration; **see Fig. A1-34**) resembling the lesions seen in *Pseudomonas aeruginosa* infection. This organism causes nosocomial infections related to catheters, surgical incisions, or use of leeches. Other manifestations include meningitis, peritonitis, pneumonia, and ocular infections.

TREATMENT

TREATMENT

Aeromonas Infections

Aeromonas species are generally susceptible to fluoroquinolones (e.g., ciprofloxacin at a dosage of 500 mg every 12 h PO or 400 mg every 12 h IV), third- and fourth-generation cephalosporins, carbapenems, and aminoglycosides, but resistance to all those agents has been described (Table 153-2). Because *Aeromonas* can produce various β-lactamases, including carbapenemases, susceptibility testing must be used to guide therapy. Antibiotic prophylaxis (e.g., with ciprofloxacin) is indicated when medicinal leeches are used.

Elizabethkingia/Chryseobacterium Species

Elizabethkingia meningoseptica (formerly *Chryseobacterium meningosepticum*), a nonfastidious aerobic nonfermentative gram-negative bacillus, is an important cause of nosocomial infections, including outbreaks due to contaminated fluids (e.g., contaminated sinks, disinfectants, and aerosolized antibiotics) and sporadic infections due to indwelling devices, feeding tubes, and other fluid-associated apparatuses. Outbreaks due to this organism have persisted until extensive cleaning of environmental surfaces and equipment has been performed. Most published reports have originated from Taiwan. Nosocomial *E. meningoseptica* infection usually involves preterm neonates, patients with underlying immunosuppression (e.g., related to malignancy or diabetes), or patients exposed to antibiotics in intensive care. *E. meningoseptica* has been reported to cause meningitis (primarily in neonates), pneumonia, sepsis, endocarditis, bacteremia, eye infections, and soft tissue infections.

Chryseobacterium indologenes has caused bacteremia, sepsis, and pneumonia, typically in immunocompromised patients with indwelling devices. Mortality rates have been as high as 50% in some reports; it is unclear whether a poor prognosis is related to underlying comorbidities or to the multidrug-resistant phenotype of the organism.

TREATMENT

TREATMENT

Elizabethkingia/Chryseobacterium Infections

These organisms are often susceptible to fluoroquinolones, minocycline, tigecycline, and rifampin (Table 153-2). They may be susceptible to β-lactam/β-lactamase inhibitor combinations such as piperacillin/tazobactam but can possess extended-spectrum β-lactamases and metallo-β-lactamases. In vitro susceptibility testing often indicates activity of agents used against gram-positive bacteria (e.g., vancomycin), but it is unclear that those agents are reliable clinically. Combination therapy may be needed for successful treatment. Susceptibility testing should be performed to guide the choice of optimal agents.

MISCELLANEOUS ORGANISMS

Rhizobium (formerly *Agrobacterium) radiobacter* has usually been associated with infection in the presence of medical devices, including intravascular catheter–related infections, prosthetic-joint and prosthetic-valve infections, and peritonitis caused by dialysis catheters. Cases of endophthalmitis after cataract surgery also have been described. Most *R. radiobacter* infections occur in immunocompromised hosts, especially individuals with malignancy or HIV infection. Strains are usually susceptible to fluoroquinolones, third- and fourth-generation cephalosporins, and carbapenems (Table 153-2).

Shewanella species are ubiquitous nonfermentative gram-negative organisms found in seawater and marine environments. Human disease is caused primarily by *S. putrefaciens* and *S. algae*; *S. algae* may be the more virulent species. Most infections involve skin and soft tissue, ranging from impetigo to necrotizing fasciitis. Patients are exposed to the organism through contact of bites, open wounds, or devitalized tissue with seawater, marine animals, or fresh seafood or through ingestion of seawater or of raw or undercooked seafood, especially shellfish. *Shewanella* species also cause chronic ulcers of the lower extremities, osteomyelitis, biliary tract infections, pneumonia, bacteremia, sepsis, and potentially chronic otitis media. A fulminant course is associated with cirrhosis, hemochromatosis, diabetes mellitus, malignancy, or other severe underlying conditions. In one series of cases from Martinique, 13% of infections were fatal. These organisms are often susceptible to fluoroquinolones, third- and fourth-generation cephalosporins, β-lactam/β-lactamase inhibitors, carbapenems, and aminoglycosides (Table 153-2).

Chromobacterium violaceum is a facultative anaerobic organism found in soil and water in tropical or subtropical regions. After exposure, it can cause rare but serious—often fatal—skin and soft tissue infections of limbs. Life-threatening infections with severe sepsis and metastatic abscesses occur most often in patients with underlying illness, particularly in children with defective neutrophil function (e.g., those with chronic granulomatous disease). *C. violaceum* is frequently resistant to multiple drugs; carbapenems are most often used empirically. Fluoroquinolones and trimethoprim-sulfamethoxazole also can be active (Table 153-2).

Ochrobactrum anthropi causes infections related to central venous catheters in compromised hosts; other invasive infections such as bacteremia have been described. *Pseudomonas* (formerly *Flavimonas*) *oryzihabitans* can cause catheter-related bloodstream infections in immunocompromised patients. *Sphingobacterium* is a rare cause of human infection in immunocompromised hosts. It can colonize hospital water systems, respiratory tract equipment, and laboratory instruments. *Ralstonia* species also can contaminate water supplies, including hospital water systems. Cases of bacteremia, osteomyelitis,

pneumonia, and meningitis have been described. Other organisms implicated in human infections include *Weeksella* species; various CDC groups, such as Ve-1 and Ve-2; and *Oligella urethralis.* The reader is advised to consult subspecialty texts and references for further guidance on these organisms.

FURTHER READING

Chambers ST et al: HACEK infective endocarditis: Characteristics and outcomes from a large multi-national cohort. PLoS ONE 8:e63181, 2013. [PubMed: 23690995]

Jean SS et al: *Elizabethkingia meningoseptica*: An important emerging pathogen causing healthcare-associated infections. J Hosp Infect 86:244, 2014.

Norskov-Lauritsen N: Classification, identification, and clinical significance of *Haemophilus* and *Aggregatibacter* species with host specificity for humans. Clin Microbiol Rev 27:214, 2014.

Vignier N et al: Human infections with *Shewanella putrefaciens* and *S. algae*: Report of 16 cases in Martinique and review of the literature. Am J Trop Med Hyg 89:151, 2013.

Wu CJ et al: Clinical implications of species identification in monomicrobial *Aeromonas* bacteremia. PLoS One 10:e0117821, 2015. [PubMed: 25679227]

Chapter 154: *Legionella* Infections

Victor L. Yu; M. Luisa Pedro-Botet; Yusen E. Lin

INTRODUCTION

Legionellosis refers to the two clinical syndromes caused by bacteria of the genus *Legionella*. *Pontiac fever* is an acute, febrile, self-limited illness that has been serologically linked to *Legionella* species, whereas *Legionnaires' disease* is the designation for pneumonia caused by these species. Legionnaires' disease was first recognized in 1976, when an outbreak of pneumonia took place at a Philadelphia hotel during an American Legion convention.

MICROBIOLOGY

The family Legionellaceae comprises more than 50 species with more than 70 serogroups. The species *Legionella pneumophila* causes 80–90% of human infections and includes at least 16 serogroups; serogroups 1, 4, and 6 are most commonly implicated in human infections. To date, 18 species other than *L. pneumophila* have been associated with human infections, among which *L. micdadei* (Pittsburgh pneumonia agent), *L. bozemanii*, *L. dumoffii*, and *L. longbeachae* are the most common. Members of the Legionellaceae are aerobic gram-negative bacilli that do not grow on routine microbiologic media. Buffered charcoal yeast extract (BCYE) agar is the medium used to grow *Legionella*.

ECOLOGY AND TRANSMISSION

The natural habitats for *L. pneumophila* are aquatic bodies, including lakes and streams. *L. longbeachae* has been isolated from natural soil. Commercial potting soil has been suggested as the reservoir for *L. longbeachae* infections in Australia and New Zealand. *Legionella* can survive under a wide range of environmental conditions; for example, the organisms can live for years in refrigerated water samples. Natural bodies of water contain only small numbers of legionellae. However, once the organisms enter human-constructed aquatic reservoirs (such as drinking-water systems), they can grow and proliferate. Factors known to enhance colonization by and amplification of *Legionella* include warm temperatures (25–42°C) and the presence of scale and sediment. *L. pneumophila* can form microcolonies within biofilms; its eradication from drinking-water systems requires disinfectants that can penetrate the biofilm. The presence of symbiotic microorganisms, including algae, amebas, ciliated protozoa, and other water-dwelling bacteria, promotes the growth of *Legionella*. The organisms can invade and multiply within free-living protozoa.

 Heavy rainfall and flooding can result in the entry of high numbers of legionellae into water-distribution systems, leading to an upsurge of cases. Climate change may be a factor in the apparent increase in incidence of Legionnaires' disease worldwide.

Large buildings over three stories high are commonly colonized with *Legionella*. Sporadic community-acquired Legionnaires' disease has been linked to colonization of hotels, office buildings, factories, and even private homes. Drinking-water systems in hospitals and extended-care facilities have been the source for health care–associated Legionnaires' disease.

In contrast, cooling towers and evaporative condensers have been overestimated as sources of *Legionella* causing human illness. Early investigations that implicated cooling towers antedated the discovery that the organism could also exist in drinking water. In many outbreaks attributed to cooling towers, cases of Legionnaires' disease continued to occur despite disinfection of the cooling towers; drinking water was found to be the actual source. Koch's postulates have never been fulfilled for *Legionella* links to cooling tower–associated outbreaks as they have been for hospital-acquired Legionnaires' disease. Nevertheless, cooling towers have, in rare instances, been implicated in community-acquired outbreaks, including outbreaks in Murcia, Spain, and Bronx, New York. As mentioned above, *L. longbeachae* infections have been linked to potting soil, but the mode of transmission remains to be clarified.

Multiple modes of transmission of *Legionella* to humans exist, including aerosolization, aspiration, and direct instillation into the lungs during respiratory tract manipulations. Aspiration is now known to be the predominant mode of transmission, but it is unclear whether *Legionella* enters the lungs via oropharyngeal colonization or directly via the drinking of contaminated water. Oropharyngeal colonization with *Legionella* has been demonstrated in patients undergoing transplantation. Nasogastric tubes have been linked to hospital-acquired Legionnaires' disease; microaspiration of contaminated water was the hypothesized mode of transmission. Surgery with general anesthesia is a known risk factor that is consistent with aspiration. The incidence of postoperative Legionnaires' disease was 30% among patients undergoing head and neck surgery at a hospital with a contaminated water supply; aspiration is a recognized postoperative complication in such cases. Patients with hospital-acquired Legionnaires' disease

underwent endotracheal intubation significantly more often and for a significantly longer duration than patients with hospital-acquired pneumonias of other etiologies.

Aerosolization of *Legionella* by devices filled with tap water, including whirlpools, nebulizers, and humidifiers, have been linked to cases in patients. An ultrasonic mist machine in the produce section of a grocery store was the source in a community outbreak. Pontiac fever has been linked to *Legionella*-containing aerosols from water-using machinery, a cooling tower, air conditioners, and whirlpools.

EPIDEMIOLOGY

Community-Acquired Pneumonia

The incidence of Legionnaires' disease depends on the degree of contamination of the aquatic reservoir, the immune status of the persons exposed to water from that reservoir, the intensity of exposure, and the availability and use of specialized laboratory tests on which the correct diagnosis can be based. *Legionella* is an underestimated cause of community-acquired pneumonia; on the basis of a multihospital study of community-acquired pneumonia in Ohio, the Centers for Disease Control and Prevention (CDC) estimated that only 3% of community-acquired Legionnaires' disease cases are diagnosed. Observational studies of community-acquired pneumonia showed that Legionnaires' disease went largely unrecognized unless *Legionella* diagnostic testing was routinely applied to all patients with pneumonia; such studies in Spain, Germany, and Taiwan stimulated an upsurge in the detection of cases throughout Europe and Asia.

Hospital-Acquired Pneumonia

Legionella is responsible for 10–50% of cases of nosocomial pneumonia when a hospital's water system is colonized with the organisms. The incidence of hospital-acquired Legionnaires' disease depends on the degree of contamination of drinking water, as defined by the rate of positivity of distal water sites; in contrast, the use of quantitative criteria of the number of colony-forming units per milliliter has proven useless.

Proactive culture of the hospital water supply has increased the detection of hospital-acquired Legionnaires' disease and simultaneously allowed expeditious diagnosis resulting in early administration of antibiotic therapy. In the early years after its recognition, Legionnaires' disease was documented primarily in the United States. As diagnostic modalities (especially the urinary antigen test) became more widely used, cases subsequently appeared worldwide.

Risk factors for Legionnaires' disease include cigarette smoking, chronic lung disease, advanced age, and prior hospitalization with discharge within 10 days before onset of pneumonia symptoms. Immunosuppressive conditions that predispose to Legionnaires' disease include transplantation and treatment with glucocorticoids or tumor necrosis factor α antagonists. However, in a large prospective study of community-acquired pneumonia, 28% of patients with Legionnaires' disease did not have these classic risk factors. Hospital-acquired cases are now being recognized among neonates and immunosuppressed children.

Pneumonia in Transplant Recipients

Transplant recipients appear to be at unusually high risk of *Legionella* pneumonia. This elevated risk may be due to diagnostic bias, given the extensive workup for opportunistic pathogens as well as long-standing immunosuppression. Legionnaires' disease usually occurs during the 3 months after transplantation. Cavitation on chest radiograph is seen more frequently in transplant recipients, and mortality rates are higher.

Pontiac Fever

Pontiac fever occurs in epidemics. The high attack rate (>90%) reflects airborne transmission.

PATHOGENESIS AND IMMUNITY

Legionella enters the lungs through aspiration or direct inhalation. Attachment to host cells is mediated by bacterial type IV pili, heat-shock proteins, a major outer-membrane protein, and complement. Because the organism possesses pili that mediate adherence to respiratory tract epithelial cells, conditions that impair mucociliary clearance, including cigarette smoking, lung disease, or alcoholism, predispose to Legionnaires' disease.

Both innate and adaptive immune responses play a role in host defense. Toll-like receptors mediate recognition of *L. pneumophila* in alveolar macrophages and enhance early neutrophil recruitment to the site of infection. After phagocytosis, *L. pneumophila* evades intracellular killing by inhibiting phagosome–lysosome fusion. Although many legionellae are killed, some proliferate intracellularly until the cells rupture; the bacteria are then phagocytosed again by newly recruited phagocytes, and the cycle begins anew. The role of neutrophils in immunity appears to be minimal: neutropenic patients are not predisposed to Legionnaires' disease. Although *L. pneumophila* is susceptible to oxygen-dependent microbiologic systems in vitro, it resists killing by neutrophils. The humoral immune system is active against *Legionella*. Type-specific IgM and IgG antibodies are measurable

within weeks of infection. In vitro, antibodies promote killing of *Legionella* by phagocytes (neutrophils, monocytes, and alveolar macrophages). Immunized animals develop a specific antibody response, with subsequent resistance to *Legionella* challenge. However, antibodies neither enhance lysis by complement nor inhibit intracellular multiplication within phagocytes.



The genome of *L. pneumophila* has been sequenced. A broad range of membrane transporters within the genome are thought to optimize the use of nutrients in water and soil. Some *L. pneumophila* strains are clearly more virulent than others, although the precise factors mediating virulence remain uncertain. For example, although multiple strains may colonize water-distribution systems, only a few cause disease in patients exposed to water from these systems. At least one surface epitope of *L. pneumophila* serogroup 1 is associated with virulence. Monoclonal antibody subtype mAb2 has been linked to virulence. *L. pneumophila* serogroup 6 is more commonly involved in hospital-acquired Legionnaires' disease and is especially likely to be associated with a poor outcome.

CLINICAL AND LABORATORY FEATURES

Pontiac Fever

Pontiac fever is an acute, self-limiting, flu-like illness with an incubation period of 24–48 h. Pneumonia does not develop. Malaise, fatigue, and myalgias are the most common symptoms, occurring in 97% of cases. Fever (usually with chills) develops in 80–90% of cases and headache in 80%. Other symptoms (seen in <50% of cases) include arthralgias, nausea, cough, abdominal pain, and diarrhea. Modest leukocytosis with a neutrophilic predominance is sometimes detected. Complete recovery occurs within a few days; antibiotic therapy is unnecessary. A few patients may experience lassitude for some weeks after recovery. The diagnosis is established by antibody seroconversion. Pontiac fever due to *L. longbeachae* has also been reported in individuals exposed to potting soil.

Legionnaire's Disease (Pneumonia)

Legionnaires' disease is often included in the differential diagnosis of "atypical pneumonia," along with pneumonia due to *Chlamydia pneumoniae*, *Chlamydia psittaci*, *Mycoplasma pneumoniae*, *Coxiella burnetii*, and some viruses. The clinical similarities among "atypical" pneumonias include a nonproductive cough with a low frequency of grossly purulent sputum. The clinical manifestations of Legionnaires' disease are usually more severe than those of most "atypical" pneumonias. The course and prognosis of *Legionella* pneumonia more closely resemble those of bacteremic pneumococcal pneumonia than those of pneumonia due to other "atypical" pathogens. Patients with community-acquired Legionnaires' disease are significantly more likely than patients with pneumonia of other etiologies to be admitted to an intensive care unit on presentation.

The incubation period for Legionnaires' disease is usually 2–10 days, although slightly longer incubation periods have been documented. The presence of fever is almost universal. Temperatures in excess of 40°C (104°F) were seen in 20% of patients in one observational study. The symptoms and signs may range from a mild cough and a slight fever to stupor with widespread pulmonary infiltrates and multisystem failure. The mild cough of Legionnaires' disease is only slightly productive. Sometimes the sputum is streaked with blood. Chest pain—either pleuritic or nonpleuritic—can be a prominent feature and, when coupled with hemoptysis, can lead to an incorrect diagnosis of pulmonary embolism. Shortness of breath is reported by one-third to one-half of patients. Gastrointestinal difficulties are often pronounced; abdominal pain, nausea, and vomiting affect 10–20% of patients. Diarrhea (watery rather than bloody) is reported in 25–50% of cases. The most common neurologic abnormalities are confusion or changes in mental status; however, the multitudinous neurologic symptoms reported range from headache and lethargy to encephalopathy. Nonspecific symptoms—malaise, fatigue, anorexia, and headache—are seen early in the illness. Myalgias and arthralgias are uncommon but are prominent in a few patients. Upper respiratory symptoms, including coryza, are rare.

Relative bradycardia has been overemphasized as a useful diagnostic finding; it occurs primarily in older patients with severe pneumonia. Chest examination reveals rales early in the course and evidence of consolidation as the disease progresses. Abdominal examination may reveal generalized or local tenderness.

Although the clinical manifestations often considered classic for Legionnaires' disease may suggest the diagnosis (Table 154-1), prospective comparative studies have shown that clinical manifestations are generally nonspecific and that Legionnaires' disease is not readily distinguishable from pneumonia of other etiologies. In a review of 13 studies of community-acquired pneumonia, clinical manifestations that occurred significantly more often in Legionnaires' disease included diarrhea, neurologic findings (including confusion), and a temperature of >39°C (>102.2°F). Hyponatremia, elevated values in liver function tests, and hematuria also occurred more frequently in Legionnaires' disease. Other laboratory abnormalities include creatine phosphokinase elevation, hypophosphatemia, serum creatinine elevation, and proteinuria.

TABLE 154-1

Clinical Clues Suggestive of Legionnaires' Disease

Diarrhea
High fever (>40°C; >104°F)
Numerous neutrophils but no organisms revealed by Gram's staining of respiratory secretions
Hyponatremia (serum sodium level <131 mg/dL)
Failure to respond to β-lactam drugs (penicillins or cephalosporins) and aminoglycoside antibiotics
Occurrence of illness in an environment in which the potable water supply is known to be contaminated with *Legionella*
Onset of symptoms within 10 days after discharge from the hospital (hospital-acquired legionellosis manifesting after discharge or transfer)

Sporadic cases of Legionnaires' disease appear to be more severe than outbreak-associated and hospital-acquired cases, presumably because their diagnosis is delayed. Results of the German CAPNETZ Study showed that, among cases of community-acquired *Legionella* pneumonia, ambulatory patients were as common as hospitalized patients.

Extrapulmonary Legionellosis

Since the portal of entry for *Legionella* is the lung in virtually all cases, extrapulmonary manifestations usually result from bloodborne dissemination from the lung. *Legionella* has been identified in lymph nodes, spleen, liver, or kidneys in autopsied cases. Sinusitis, peritonitis, pyelonephritis, skin and soft-tissue infection, septic arthritis, and pancreatitis have developed predominantly in immunosuppressed patients. The most severe sequela, neurologic dysfunction, is rare but can be debilitating. The most common neurologic deficits in the long term—ataxia and speech difficulties—result from cerebellar dysfunction.

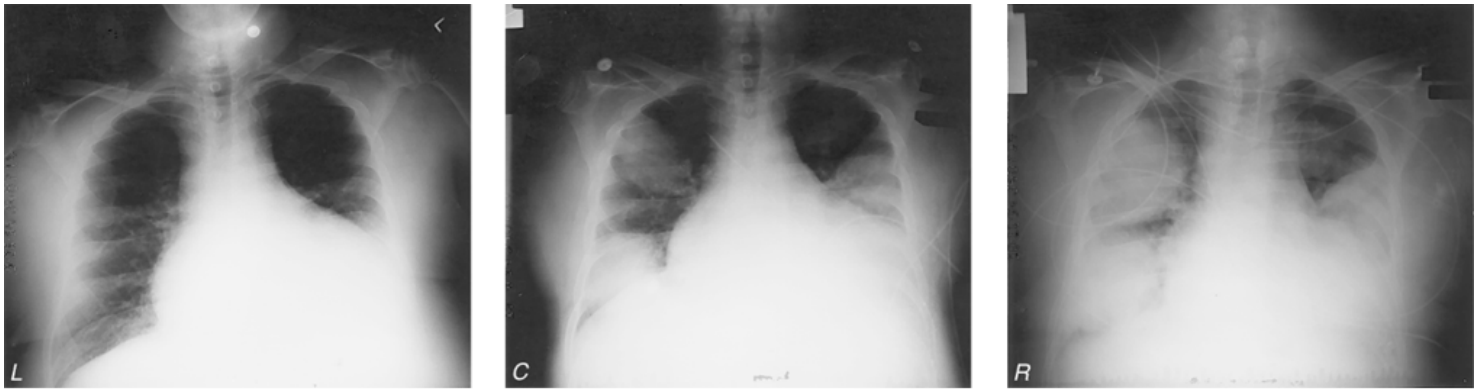
Cardiac abnormalities may occur without pneumonia; *Legionella*-contaminated water can enter through an IV injection site, chest tube, or surgical wound, with subsequent seeding of a prosthetic valve, the myocardium, or the pericardium.

Chest Radiography

Virtually all patients with Legionnaires' disease have abnormal chest radiographs showing pulmonary infiltrates at the time of clinical presentation. In a few cases of hospital-acquired disease, fever and respiratory tract symptoms have preceded the radiographic appearance of the infiltrate. Radiologic findings are nonspecific. Pleural effusion is evident in 28–63% of patients on hospital admission. In immunosuppressed patients, especially those receiving glucocorticoids, distinctive rounded nodular opacities may be seen; these lesions may expand and cavitate **(Fig. 154-1)**. Likewise, abscesses can occur in immunosuppressed hosts. The progression of infiltrates and pleural effusion on chest radiography despite appropriate antibiotic therapy within the first week is common, and radiographic improvement lags behind clinical improvement by several days. Complete clearing of infiltrates requires 1–4 months. CT scan is more sensitive than chest radiography and may show more extensive disease. A CT scan should be obtained if fever persists during treatment with presumably effective antibiotics **(Fig. 154-2)**.

FIGURE 154-1

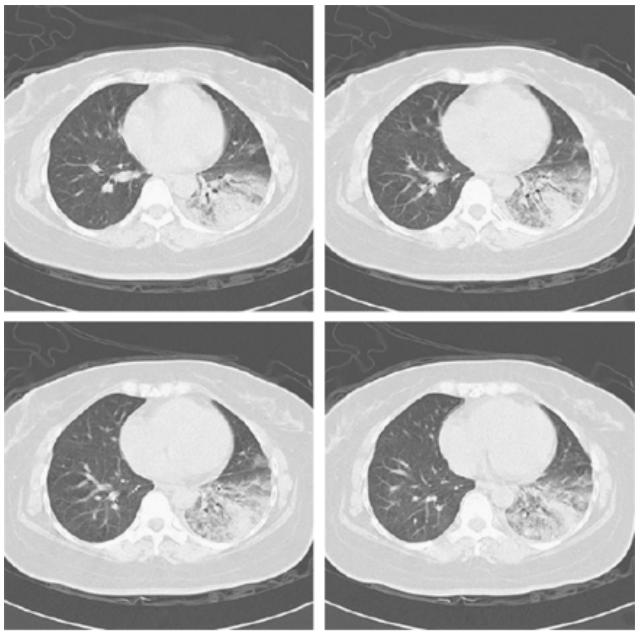
Chest radiographic findings in a 52-year-old man who presented with pneumonia subsequently diagnosed as Legionnaires' disease. The patient was a cigarette smoker with chronic obstructive pulmonary disease and alcoholic cardiomyopathy; he had received glucocorticoids. *Legionella pneumophila* was identified by direct fluorescent antibody staining and culture of sputum. ***Left:*** Baseline chest radiograph showing long-standing cardiomegaly. ***Center:*** Admission chest radiograph showing new rounded opacities. ***Right:*** Chest radiograph taken 3 days after admission, during treatment with erythromycin.



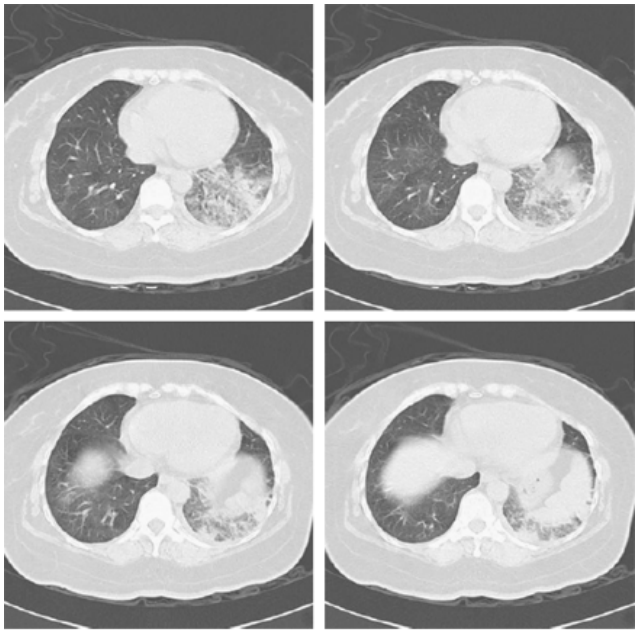
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 154-2

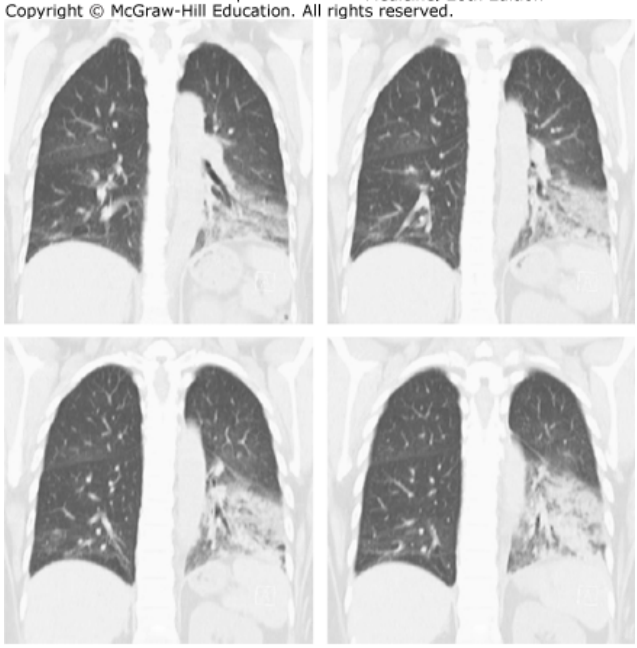
CT findings in a 49-year-old woman with no underlying conditions who presented with community-acquired pneumonia. CT revealed multilobar infiltrates, some of which were not as prominent on chest x-ray. Cultures of both the patient's sputum and her home water supply yielded *Legionella pneumophila* serogroup 1. *(Images courtesy of Dr. Wen-Chien Ko, National Cheng Kung University Hospital, Tainan, Taiwan.)*



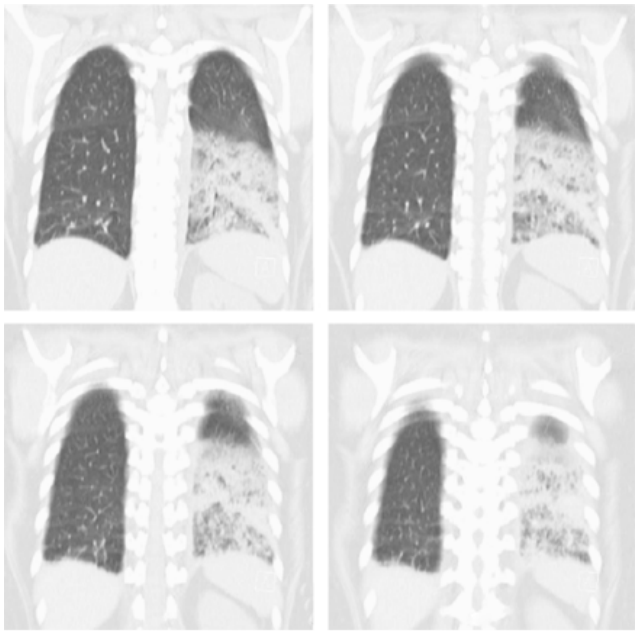
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS

Given the nonspecific clinical manifestations of Legionnaires' disease and the high mortality rates for untreated Legionnaires' disease, the use of *Legionella* testing—especially the *Legionella* urinary antigen test—is recommended for all patients with community-acquired pneumonia, including patients with ambulatory pneumonia and hospitalized children. *Legionella* cultures should be made more widely available since the urinary antigen test can diagnose only *L. pneumophila* serogroup 1. Hospitals in which the drinking water is known to be colonized with *Legionella* species should have *Legionella* cultures routinely available.

The diagnosis of Legionnaires' disease requires special microbiologic tests **(Table 154-2)**. The sensitivity of bronchoscopy specimens is similar to that of sputum samples for culture on selective media; if sputum is not available, bronchoscopy specimens may yield the organism. Bronchoalveolar lavage fluid gives higher yields than bronchial wash specimens. Thoracentesis should be performed if pleural effusion is found, and the fluid should be evaluated by direct fluorescent antibody (DFA) staining, culture, and the antigen assay designed for use with urine.

TABLE 154-2

Utility of Special Laboratory Tests for the Diagnosis of Legionnaires' Disease

Test	Sensitivity, %	Specificity, %
Culture		
Sputum[a]	80	100
Transtracheal aspirate	90	100
Direct fluorescent antibody staining of sputum	50–70	96–99
Urinary antigen testing[b]	70	100
Antibody serology[c]	40–60	96–99

[a]Use of multiple selective media with dyes. [b]Reliable only for *L. pneumophila* serogroup 1. [c]IgG and IgM testing of both acute- and convalescent-phase sera. A single titer of ≥1:256 is considered presumptive, while a fourfold rise in titer between the acute and convalescent phases is considered definitive. Titers peak at 3 months.

Stains

Gram's staining of material from normally sterile sites, such as pleural fluid or lung tissue, occasionally suggests the diagnosis; efforts to detect *Legionella* in sputum by Gram's staining typically reveal numerous leukocytes but no organisms. When they are visualized, the organisms appear as small, pleomorphic, faint, gram-negative bacilli. *L. micdadei* organisms can be detected as weakly or partially acid-fast bacilli in clinical specimens.

The DFA stain is rapid and highly specific but is less sensitive than culture because large numbers of organisms are required for microscopic visualization. This test is more likely to be positive in advanced than in early disease.

Culture

The "gold-standard" method for diagnosis of *Legionella* infection is isolation of the organism from respiratory secretions, although culture for 3–5 days—and sometimes up to 2 weeks for non-*pneumophila* species—is required. Antibiotics added to the medium suppress the growth of competing flora from nonsterile sites, and dyes color the colonies and assist in identification. The use of multiple selective BCYE media is necessary for maximal sensitivity. When culture plates are overgrown with other microflora, pretreatment of the specimen with acid or heat can markedly improve the yield. *L. pneumophila* is often isolated from sputum that is not grossly or microscopically purulent; sputum containing more than 25 epithelial cells per high-power field (a finding that classically suggests contamination) may still yield *L. pneumophila*.

Antibody Detection

Antibody testing of both acute- and convalescent-phase sera is necessary. A fourfold rise in titer is diagnostic; 12 weeks are often required for the detection of an antibody response. A single titer of 1:128 in a patient with pneumonia constitutes circumstantial evidence for Legionnaires' disease; a single titer of 1:256 constitutes presumptive evidence. The specificity of serology for *Legionella* species other than *L. pneumophila* is uncertain; there is cross-reactivity within *Legionella* species and some gram-negative bacilli. Positive serology serves as the criterion for the diagnosis of Pontiac fever.

Although serology has its limitations, it can still be useful for confirmation of Legionnaires' disease when isolation of the agent is impossible, and it can serve as a corroborating test. Serology is also useful for retrospective epidemiologic investigations and general seroprevalence determinations.

Urinary Antigen

The detection of *Legionella* soluble antigen in urine, which is often easier to collect than sputum, is a relatively low-cost test and is easy to perform. Urinary antigen can be detected shortly after clinical symptoms appear and for up to 10 months thereafter, even during antibiotic treatment. The test's specificity is 95–100%, and its sensitivity ranges from 70 to 90%. Its drawback is that it is reliable only for *L. pneumophila* serogroup 1, which causes ~80% of *Legionella* infections. Cross-reactivity with other *L. pneumophila* serogroups and other *Legionella* species has been detected in up to 22% of urine samples from patients with culture-proven cases.

Molecular Methods

Nucleic acid–based detection of *Legionella* offers advantages over serology and cultures because of its sensitivity and speed. DNA detection with the polymerase chain reaction (PCR) in both conventional and real-time thermal cyclers is highly specific (almost 100%) and sensitive. The sensitivity of DNA detection may be superior to that of culture in milder cases of Legionnaires' disease. DFA stains can identify a number of *Legionella* species. Both polyclonal and monoclonal antibody stains are commercially available. Procalcitonin can be used as an indicator of severity of illness in ICU patients. In addition, the clinical response to antibiotics can be monitored by procalcitonin levels.

TREATMENT

TREATMENT

Legionella Infection

Because *Legionella* is an intracellular pathogen, antibiotics that reach high intracellular concentrations are more likely to be effective. The dosages for various drugs used in the treatment of *Legionella* infection are listed in **Table 154-3**.

The macrolides (especially azithromycin) and the quinolones (especially levofloxacin or ciprofloxacin) are the antibiotics of choice and are effective as monotherapy. Compared with erythromycin, the newer macrolides have superior in vitro activity, display greater intracellular activity, reach higher concentrations in respiratory secretions and lung tissue, and have fewer adverse effects. Quinolones are the preferred antibiotics for transplant recipients because both macrolides and rifampin interact pharmacologically with cyclosporine and tacrolimus. Retrospective uncontrolled studies have shown that complications of pneumonia are fewer and clinical response is more rapid in patients receiving quinolones than in those receiving macrolides.

Alternative agents include tetracycline and its analogues: doxycycline and minocycline. Tigecycline is active in vitro, but clinical experience with this drug is minimal. Anecdotal reports have described both successes and failures with trimethoprim-sulfamethoxazole, imipenem, and clindamycin.

For treatment of critically ill patients, we use combinations of azithromycin, a quinolone, and/or rifampin. This practice is empirical and is not supported by comparative studies. Rifampin is highly active in vitro and in cell models. Adverse effects of rifampin can be minimized by limiting the duration of therapy to 3–5 days.

Initial antibiotic therapy should be given by the IV route because gastrointestinal symptoms are common in Legionnaires' disease. A clinical response usually occurs within 3–5 days, after which oral therapy can be substituted. The total duration of therapy in the immunocompetent host is 10–14 days; a longer course (3 weeks) may be appropriate for immunosuppressed patients. For azithromycin, with its long half-life, a 5- to 10-day course is sufficient.

Pontiac fever requires only symptom-based treatment, not antimicrobial therapy.

TABLE 154-3

Antibiotic Therapy for *Legionella* Infection

Antimicrobial Agent	Dosage[a]
Macrolides	
Azithromycin	500 mg[b] PO or IV[c] q24h
Clarithromycin	500 mg PO or IV[c] q12h
Quinolones	
Levofloxacin	750 mg IV q24h
	500 mg[b] PO q24h
Ciprofloxacin	400 mg IV q8h
	750 mg PO q12h
Moxifloxacin	400 mg[b] PO q24h
Ketolide	
Telithromycin	800 mg PO q24h
Tetracyclines	
Doxycycline	100 mg[b] PO or IV q12h
Minocycline	100 mg[b] PO or IV q12h
Tetracycline	500 mg PO or IV q6h
Tigecycline	100-mg IV load, then 50 mg IV q12h
Others	
Trimethoprim-sulfamethoxazole	160/800 mg IV q8h
Rifampin[d]	160/800 mg PO q12h
	300–600 mg PO or IV q12h

[a]Dosages are derived from clinical experience. [b]The authors recommend doubling the first dose. [c]The IV formulation is not available in some countries. [d]Rifampin should be used only in combination with a macrolide or a quinolone, and the duration of therapy should be limited to 3–5 days.

PROGNOSIS

Mortality rates for Legionnaires' disease vary with the patient's underlying disease, the patient's immune status, the severity of pneumonia, and the timing of administration of appropriate antimicrobial therapy. Mortality rates are highest (80%) among immunosuppressed patients who do not receive appropriate antimicrobial therapy early in the course of illness. With timely antibiotic treatment, mortality rates from community-acquired Legionnaires' disease among immunocompetent patients range from 3 to 11%; without treatment, the figure may be as high as 31%. In a study of

survivors of an outbreak of community-acquired Legionnaires' disease, sequelae of fatigue, neurologic symptoms, and weakness were found in 63–75% of patients 17 months after receipt of antibiotics.

PREVENTION

Routine environmental culture of hospital water supplies for *Legionella* is recommended as an approach to the prevention of hospital-acquired Legionnaires' disease. Guidelines mandating this proactive approach have been adopted throughout Europe, in Taiwan, and in several U.S. states. The presence of *Legionella* in the water supply mandates the use of specialized laboratory tests (especially culture on selective media and the urinary antigen test) for patients with hospital-acquired pneumonia. A 30% cutoff for the presence of *Legionella* in water from multiple sites in a hospital has been used as a prompt for an increased index of suspicion and consideration of empirical therapy for Legionnaires' disease. When the 30% cutoff is exceeded, *Legionella* diagnostic tests need to be performed for all patients with hospital-acquired pneumonia, and measures directed at the water supply may be considered. Quantitative criteria at a water site (CFU/mL) have proven unreliable and inconsistent in predicting disease.

Studies have shown that neither a high degree of outward cleanliness of the water system nor routine application of maintenance measures decreases the frequency or intensity of *Legionella* contamination. Thus, engineering guidelines and building codes, although routinely advocated as preventive measures, have little impact on the presence of *Legionella*.

Environmental cultures for *Legionella* from water taps, the hot-water recirculating line, and water storage tanks will reveal the source of hospital-acquired infections. Disinfection of hospital drinking-water systems is an effective preventive measure for hospital-acquired cases of Legionnaires' disease since these systems serve as the reservoir for *Legionella*. In geographic areas where the weather is semitropical, the cold-water lines also may be colonized by *Legionella*.

Copper–silver ionization has been shown to be a reliable and efficacious method for eradicating *Legionella* in hospitals worldwide. Unlike that of chlorine dioxide treatment and chlorination, the efficacy of ionization is not affected by high water temperatures. Ionization systems are easy to install, and the ions are odorless with minimal adverse effects. A comprehensive review of 10 published studies concluded that copper–silver ionization is effective in controlling *Legionella* as long as ion levels are monitored. If cold-water colonization by *Legionella* is the source of an outbreak, chlorine dioxide and monochloramine treatment is advantageous. Chlorine dioxide treatment is often the least expensive method, offers superior penetration into biofilms, and has less corrosive effects than chlorine. The major disadvantage of chlorine dioxide is the necessity for maintaining an effective residual throughout the drinking-water system, especially in the hot-water system. Eradication of *Legionella* by chlorine dioxide may require several months—a drawback in outbreak situations. Monochloramine appears to be effective. Hyperchlorination is no longer recommended because of its expense, carcinogenicity, corrosive effects on piping, and unreliable efficacy.

Point-of-use disposable water filters (pore size, 0.2 μm) may be an economical and effective option in high-risk areas (e.g., ICUs and transplant units). These filters can be used in an outbreak situation for a limited period.

Ineffective yet expensive methods of prevention are often promulgated. These methods include removal of stagnation (e.g., in dead-legs) within the water distribution system and replacement or disinfection/cleaning of distal outlets. Infection control personnel should oversee the selection of disinfection technology and apply evidence-based criteria for their selection. Health care facility managers should not be given the primary responsibility for selection and subsequent monitoring.

FURTHER READING

Barker KA et al: A review of guidelines for the primary prevention of legionellosis in long-term care facilities. J Am Med Dir Assoc 6:832, 2015.

Borella P et al: Hospital-acquired *Legionella* infections: An update on the procedures for controlling environmental contamination. Ann Ig 28:98, 2016.

Lin YSE et al: Controlling *Legionella* in hospital drinking water: An evidence-based review of disinfection methods. Infect Control Hosp Epidemiol 32:166, 2011.

Mercante JW, Winchell JM: Current and emerging *Legionella* diagnostics for laboratory and outbreak investigations. Clin Microbiol Rev 28:95, 2015.

Chapter 155: Pertussis and Other *Bordetella* Infections

Karina A. Top; Scott A. Halperin

INTRODUCTION

Pertussis is an acute infection of the respiratory tract caused by *Bordetella pertussis*. The word *pertussis* means “violent cough,” which aptly describes the most consistent and prominent feature of the illness. The inspiratory sound made at the end of an episode of paroxysmal coughing gives rise to the common name for the illness, “whooping cough.” However, this feature is variable: it is uncommon among infants ≤6 months of age and is frequently absent in older children and adults. The Chinese name for pertussis is “the 100-day cough,” which accurately describes the clinical course of the illness. The identification of *B. pertussis* was first reported by Bordet and Gengou in 1906, and vaccines were produced over the following two decades.

MICROBIOLOGY

Of the 10 identified species in the genus *Bordetella*, only four are of major medical significance. *B. pertussis* infects only humans and is the most important *Bordetella* species causing human disease. *B. parapertussis* causes an illness in humans that is similar to pertussis but is typically milder; co-infections with *B. parapertussis* and *B. pertussis* have been documented. With improved polymerase chain reaction (PCR) diagnostic methodology, up to 20% of patients with a pertussis-like syndrome have been found to be infected with *B. holmesii*, formerly thought to be an unusual cause of bacteremia. *B. bronchiseptica* is an important pathogen of domestic animals that causes kennel cough in dogs, atrophic rhinitis and pneumonia in pigs, and pneumonia in cats. Both respiratory infection and opportunistic infection due to *B. bronchiseptica* are reported occasionally in humans. *B. petrii*, *B. hinzii*, and *B. ansorpii* have been isolated from patients who are immunocompromised.



Bordetella species are gram-negative pleomorphic aerobic bacilli that share common genotypic characteristics. *B. pertussis* and *B. parapertussis* are the most similar of the species, but *B. parapertussis* does not express the gene coding for pertussis toxin. *B. pertussis* is a slow-growing fastidious organism that requires selective medium and forms small, glistening, bifurcated colonies. Suspicious colonies are presumptively identified as *B. pertussis* by direct fluorescent antibody testing or by agglutination with species-specific antiserum. *B. pertussis* is further differentiated from other *Bordetella* species by biochemical and motility characteristics.

B. pertussis produces a wide array of toxins and biologically active products that are important in its pathogenesis and in immunity. Most of these virulence factors are under the control of a single genetic locus that regulates their production, resulting in antigenic modulation and phase variation. Although these processes occur both in vitro and in vivo, their importance in the pathobiology of the organism is unknown; they may play a role in intracellular persistence and person-to-person spread. The organism’s most important virulence factor is *pertussis toxin*, which is composed of a B oligomer–binding subunit and an enzymatically active A protomer that ADP-ribosylates a guanine nucleotide–binding regulatory protein (G protein) in target cells, producing a variety of biologic effects. Pertussis toxin has important mitogenic activity, affects the circulation of lymphocytes, and serves as an adhesin for bacterial binding to respiratory ciliated cells. Other

important virulence factors and adhesins are *filamentous hemagglutinin*, a component of the cell wall, and *pertactin*, an outer-membrane protein. *Fimbriae*, bacterial appendages that play a role in bacterial attachment, are the major antigens against which agglutinating antibodies are directed. These agglutinating antibodies have historically been the primary means of serotyping *B. pertussis* strains. Other virulence factors include tracheal cytotoxin, which causes respiratory epithelial damage; adenylate cyclase toxin, which impairs host immune-cell function; dermonecrotic toxin, which may contribute to respiratory mucosal damage; and lipooligosaccharide, which has properties similar to those of other gram-negative bacterial endotoxins.

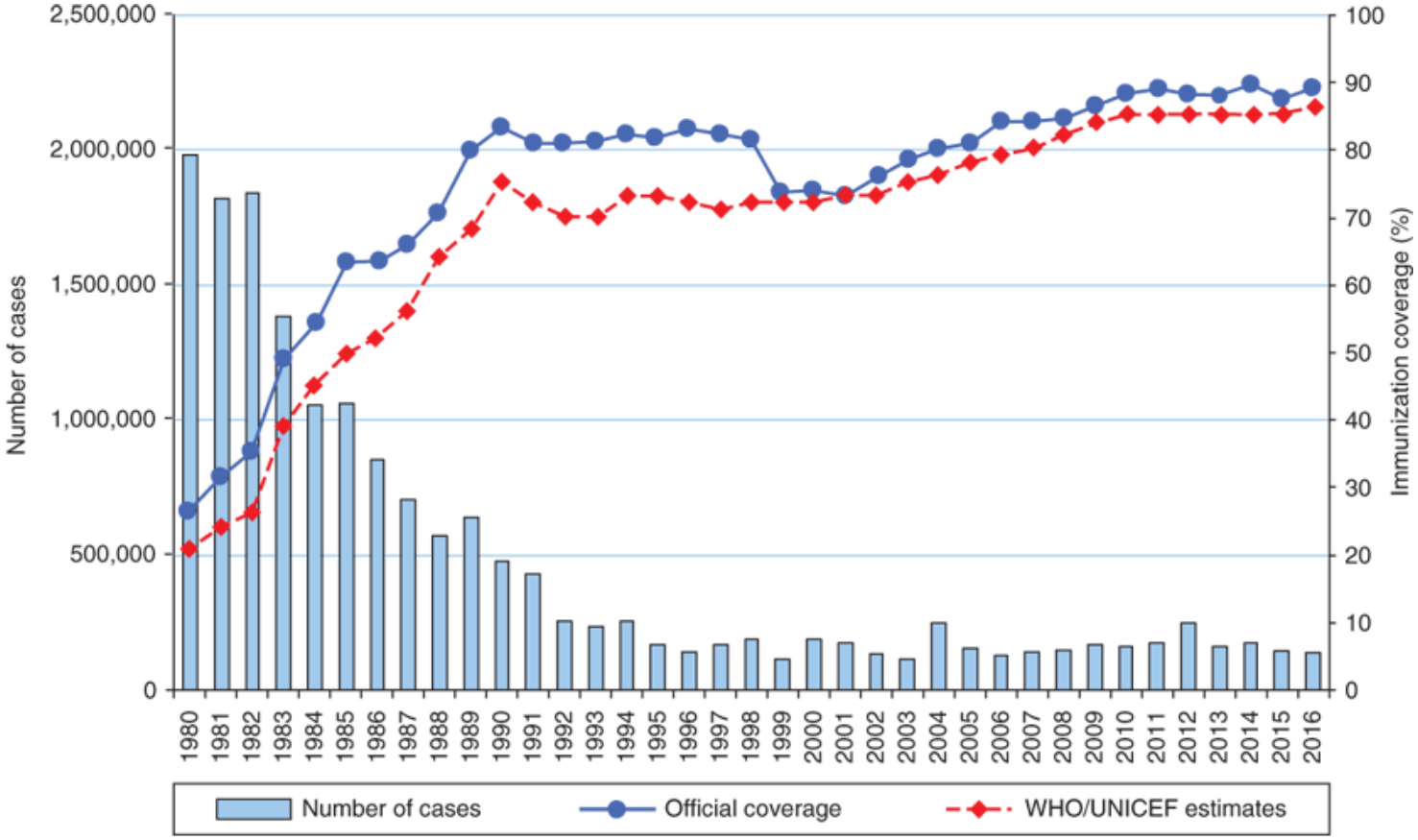
EPIDEMIOLOGY

Pertussis is a highly communicable disease, with attack rates of 80–100% among unimmunized household contacts and 20% within households in well-immunized populations. The infection has a worldwide distribution, with cyclical outbreaks every 3–5 years (a pattern that has persisted despite widespread immunization). Pertussis occurs in all months; however, in North America, its activity peaks in autumn and winter.

In developing countries, pertussis remains an important cause of infant morbidity and death. The reported incidence of pertussis worldwide has decreased as a result of improved vaccine coverage **(Fig. 155-1)**. However, coverage rates are still <60% in many developing nations; the World Health Organization (WHO) estimates that 90% of the burden of pertussis is in developing regions. In addition, over-reporting of immunization coverage and under-reporting of disease result in substantial underestimation of the global burden of pertussis. The WHO estimates that there were 63,000 deaths from pertussis among children <5 years of age in 2013.

FIGURE 155-1

Global annual reported cases of pertussis and rate of coverage with DTP3 (diphtheria toxoid, tetanus toxoid, and pertussis vaccine; 3 doses), 1980–2016. *(© World Health Organization, 2017. All rights reserved. From http://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/passive/pertussis_coverage.jpg?ua=1. Source: WHO/IVB database, 2017. Accessed January 8, 2018.)*

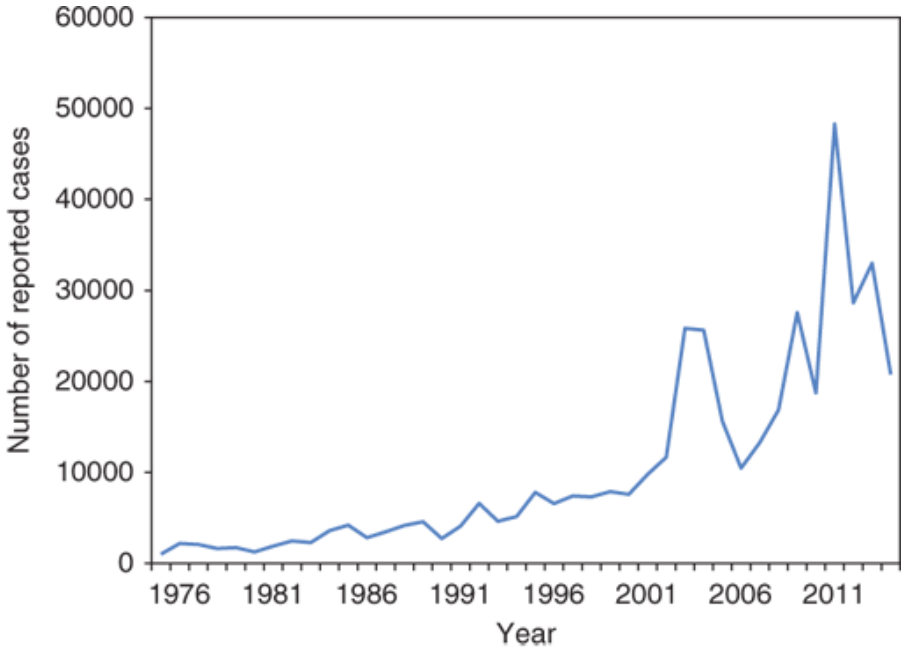


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Before the institution of widespread immunization programs in the developed world, pertussis was one of the most common infectious causes of morbidity and death. In the United States before the 1940s, between 115,000 and 270,000 cases of pertussis were reported annually, with an average yearly rate of 150 cases per 100,000 population. With universal childhood immunization, the number of reported cases fell by >95%, and mortality rates decreased even more dramatically. Only 1010 cases of pertussis were reported in 1976 **(Fig. 155-2)**. After that historic low, rates of pertussis slowly increased. In recent years, pertussis epidemics have been reported with increasing frequency in several developed countries, including Australia, the United Kingdom, and the United States. The United States experienced widespread outbreaks of pertussis in 2005, 2010, 2012, and 2014 at levels not seen in 40–50 years (>40,000 reported cases in 2012).

FIGURE 155-2

Reported cases of pertussis by year—United States, 1976–2015. *(From the Centers for Disease Control and Prevention, www.cdc.gov/pertussis/surv-reporting/cases-by-year.html. Accessed January 8, 2018.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although thought of as a disease of childhood, pertussis can affect people of all ages and is increasingly being identified as a cause of prolonged coughing illness in adolescents and adults. In unimmunized populations, pertussis incidence peaks during the preschool years, and well over half of children have the disease before reaching adulthood. In highly immunized populations such as those in North America, the peak incidence is among infants <1 year of age who have not completed the three-dose primary immunization series. An increase in pertussis incidence among adolescents and adults began in the late 1990s and led to the introduction of an adolescent booster dose across North America by 2006. While the disease burden among adolescents has started to decrease, children 7–10 years of age have recently emerged as a high-risk group. In major outbreaks in 2010 and 2012, the incidence of pertussis among children 10 years of age, most of whom were fully immunized, was as high as that among infants <6 months of age. Although adults contribute a smaller proportion of reported cases of pertussis than do children and adolescents, this difference may be related to a greater degree of under-recognition and under-reporting. A number of studies of prolonged coughing illness suggest that *B. pertussis* may be the etiologic agent in 12–30% of adults with cough that does not improve within 2 weeks. In one study of the efficacy of an acellular pertussis vaccine in adolescents and adults, the incidence of pertussis in the placebo group was 3.7–4.5 cases per 1000 person-years. Although this prospective cohort study yielded a lower estimate than the studies of cough illness, its results still translate to 600,000–800,000 cases of pertussis annually among adults in the United States.

Severe morbidity and high mortality rates, however, are restricted almost entirely to infants. In the United States between 1993 and 2004, all pertussis deaths and 86% of hospitalizations for pertussis involved infants ≤3 months of age. Although school-age children are the source of infection for most households, adults are often the source for cases in high-risk infants and may serve as the reservoir of infection between epidemic years.

PATHOGENESIS

Infection with *B. pertussis* is initiated by attachment of the organism to the ciliated epithelial cells of the nasopharynx. Attachment is mediated by surface adhesins (e.g., pertactin and filamentous hemagglutinin), which bind to the integrin family of cell-surface proteins, probably in conjunction with pertussis toxin. The role of fimbriae in adhesion and in maintenance of infection has not been fully delineated. At the site of attachment, the organism multiplies, producing a variety of other toxins that cause local mucosal damage (tracheal cytotoxin, dermonecrotic toxin). Impairment of host defense by *B. pertussis* is mediated by pertussis toxin and adenylate cyclase toxin. There is local cellular invasion, with

intracellular bacterial persistence; however, systemic dissemination does not occur. Systemic manifestations (lymphocytosis) result from the effects of the toxins.

The pathogenesis of the clinical manifestations of pertussis is poorly understood. It is not known what causes the hallmark paroxysmal cough. A pivotal role for pertussis toxin has been proposed. Proponents of this position point to the efficacy of preventing clinical symptoms with a vaccine containing only pertussis toxoid. Detractors counter that pertussis toxin is not the critical factor because paroxysmal cough also occurs in patients infected with *B. parapertussis*, which does not produce pertussis toxin. It is thought that neurologic events in pertussis, such as seizures and encephalopathy, are due to hypoxia from coughing paroxysms or apnea rather than to the effects of specific bacterial products. *B. pertussis* pneumonia, which occurs in up to 10% of infants with pertussis, is usually a diffuse bilateral primary infection. In older children and adults with pertussis, pneumonia is often due to secondary bacterial infection with streptococci or staphylococci. Deaths from pertussis among young infants are frequently associated with very high levels of leukocytosis and pulmonary hypertension.

IMMUNITY

Both humoral and cell-mediated immunity are thought to be important in pertussis. Antibodies to pertussis toxin, filamentous hemagglutinin, pertactin, and fimbriae are all protective in animal models. Pertussis agglutinins were correlated with protection in early studies of whole-cell pertussis vaccines. Serologic correlates of protection conferred by acellular pertussis vaccines have not been established, although antibody to pertactin, fimbriae, and (to a lesser degree) pertussis toxin correlated best with protection in two efficacy trials. The duration of immunity after whole-cell pertussis vaccination is short-lived, with little protection remaining after 10–12 years. Studies have demonstrated early waning of immunity—i.e., within 2–4 years after the fifth dose of acellular pertussis vaccine in children who received acellular pertussis vaccine for their primary series in infancy. These data suggest that boosters may be needed more frequently than every 10 years, as previously thought. Although immunity after natural infection was thought to be lifelong, seroepidemiologic evidence demonstrates that it clearly is not and that subsequent episodes of clinical pertussis are prevented by intermittent subclinical infection.

CLINICAL MANIFESTATIONS

Pertussis is a prolonged coughing illness with clinical manifestations that vary by age (Table 155-1). Although not uncommon among adolescents and adults, classic pertussis is most often seen in preschool and school-age children. After an incubation period averaging 7–10 days, an illness develops that is indistinguishable from the common cold and is characterized by coryza, lacrimation, mild cough, low-grade fever, and malaise. After 1–2 weeks, this *catarrhal phase* evolves into the *paroxysmal phase*: the cough becomes more frequent and spasmodic with repetitive bursts of 5–10 coughs, often within a single expiration. Post-tussive vomiting is frequent, with a mucous plug occasionally expelled at the end of an episode. The episode may be terminated by an audible whoop, which occurs upon rapid inspiration against a closed glottis at the end of a paroxysm. During a spasm, there may be impressive neck-vein distension, bulging eyes, tongue protrusion, and cyanosis. Paroxysms may be precipitated by noise, eating, or physical contact. Between attacks, the patient's appearance is normal but increasing fatigue is evident. The frequency of paroxysmal episodes varies widely, from several per hour to 5–10 per day. Episodes are often worse at night and interfere with sleep. Most complications occur during the paroxysmal stage. Fever is uncommon and suggests bacterial superinfection.

TABLE 155-1

Clinical Features of Pertussis, by Age Group and Diagnostic Status

Feature	Percentage of Patients		
	Adolescents and Adults		Children
	Laboratory Confirmation	No Laboratory Confirmation	
Cough	95–100	95–100	95–100
Prolonged	60–80	60–80	60–95
Paroxysmal	60–90	50–90	80–95
Sleep-disturbing	50–80	50–80	90–100
Whoop	10–40	5–30	40–80
Post-tussive vomiting	20–50	5–30	80–90

After 2–4 weeks, the coughing episodes become less frequent and less severe—changes heralding the onset of the *convalescent phase*. This phase can last 1–3 months and is characterized by gradual resolution of coughing episodes. For 6–12 months, intercurrent viral infections may be associated with a recrudescence of paroxysmal cough.

Not all individuals who develop pertussis have classic disease. The clinical manifestations in adolescents and adults are more often atypical. In a German study of pertussis in adults, more than two-thirds had paroxysmal cough and more than one-third had whoop. Adult illness in North America differs from this experience: the cough may be severe and prolonged but is less frequently paroxysmal, and a whoop is uncommon. Vomiting with cough is the best predictor of pertussis as the cause of prolonged cough in adults. Other predictive features are a cough at night, sweating episodes between paroxysms of coughing, and exposure to other individuals with a prolonged coughing illness.

COMPLICATIONS

Complications are frequently associated with pertussis and are more common among infants than among older children or adults. Subconjunctival hemorrhages, abdominal and inguinal hernias, pneumothoraces, and facial and truncal petechiae can result from increased intrathoracic pressure generated by severe fits of coughing. Weight loss can follow decreased caloric intake. Urinary incontinence, rib fracture, carotid artery aneurysm, and cough syncope have also been reported in adolescents and adults with pertussis. In a series of more than 1100 children <2 years of age who were hospitalized with pertussis, 27.1% had apnea, 9.4% had pneumonia, 2.6% had seizures, and 0.4% had encephalopathy; 10 children (0.9%) died. Pneumonia is reported in <5% of adolescents and adults and increases in frequency after 50 years of age. In contrast to the primary *B. pertussis* pneumonia that develops in infants, pneumonia in adolescents and adults with pertussis is usually caused by a secondary infection with encapsulated organisms such as *Streptococcus pneumoniae* or *Haemophilus influenzae*.

DIAGNOSIS

If the classic symptoms of pertussis are present, clinical diagnosis is not difficult. However, particularly in older children and adults, it is difficult to differentiate infections caused by *B. pertussis* and *B. parapertussis* from other respiratory tract infections on clinical grounds. Therefore, laboratory confirmation should be attempted in all cases. Lymphocytosis (absolute lymphocyte count, $>10^8$–10^9/L) is common among young children, in whom it is unusual with other infections, but not among adolescents and adults. Culture of nasopharyngeal secretions remains the gold standard of diagnosis, although DNA detection by PCR has replaced culture in many laboratories because of increased sensitivity and quicker results. Appropriate PCR methodology must include primers to differentiate among *B. pertussis*, *B. parapertussis*, and *B. holmesii*. The best specimen is collected by nasopharyngeal aspiration, in which a fine flexible plastic catheter attached to a 10-mL syringe is passed into the nasopharynx and withdrawn while gentle suction is applied. Since *B. pertussis* is highly sensitive to drying, secretions for culture should be inoculated without delay onto appropriate medium (Bordet-Gengou or Regan-Lowe), or the catheter should be flushed with a phosphate-buffered saline solution for culture and/or PCR. An alternative to the aspirate is a Dacron or rayon nasopharyngeal swab; again, inoculation of culture plates should be immediate or an appropriate transport medium (e.g., Regan-Lowe charcoal medium) should be used. Results of PCR can be available within hours; cultures become positive by day 5 of incubation. *B. pertussis* and *B. parapertussis* can be differentiated by agglutination with specific antisera or by direct immunofluorescence.

Nasopharyngeal cultures in untreated pertussis remain positive for a mean of 3 weeks after the onset of illness; these cultures become negative within 5 days of the institution of appropriate antimicrobial therapy. The duration of a positive PCR in untreated pertussis or after therapy is not known but exceeds that of positive cultures. Since much of the period during which the organism can be recovered from the nasopharynx falls into the catarrhal phase, when the etiology of the infection is not suspected, there is only a small window of opportunity for culture-proven diagnosis. Cultures from infants and young children are more frequently positive than those from older children and adults; this difference may reflect earlier presentation of the former age group for medical care. Direct fluorescent antibody tests of nasopharyngeal secretions for direct diagnosis may still be available in some laboratories but should not be used because of poor sensitivity and specificity. Pseudo-outbreaks of pertussis have been reported as a result of false-positive PCR results. Greater standardization of PCR methodology can alleviate this problem.

As a result of the difficulties with laboratory diagnosis of pertussis in adolescents, adults, and patients who have been symptomatic for >4 weeks, increasing attention is being given to serologic diagnosis. Enzyme immunoassays detecting IgA and IgG antibodies to pertussis toxin, filamentous hemagglutinin, pertactin, and fimbriae have been developed and assessed for reproducibility. Two- or fourfold increases in antibody titer are suggestive of pertussis, although cross-reactivity of some antigens (such as filamentous hemagglutinin and pertactin) among *Bordetella* species makes it difficult to depend diagnostically on seroconversion involving a single type of antibody. Late presentation for medical care and prior immunization also complicate serologic diagnosis because the first sample obtained may in fact be a convalescent-phase specimen. Criteria for serologic diagnosis based on comparison of results for a single serum specimen with established population values are gaining acceptance, and serologic measurement of antibody to pertussis toxin is becoming more widely standardized and available for diagnostic purposes, particularly in outbreak settings and for surveillance.

DIFFERENTIAL DIAGNOSIS

A child presenting with paroxysmal cough, post-tussive vomiting, and whoop is likely to have an infection caused by *B. pertussis* or *B. parapertussis*; lymphocytosis increases the likelihood of a *B. pertussis* etiology. Viruses such as respiratory syncytial virus and adenovirus have been isolated from patients with clinical pertussis but probably represent co-infection.

In adolescents and adults, who often do not have paroxysmal cough or whoop, the differential diagnosis of a prolonged coughing illness is more extensive. Pertussis should be suspected when any patient has a cough that does not improve within 14 days, a paroxysmal cough of any duration, a cough followed by vomiting (adolescents and adults), or any

respiratory symptoms after contact with a laboratory-confirmed case of pertussis. Other etiologies to consider include infections caused by *Mycoplasma pneumoniae*, *Chlamydia pneumoniae*, adenovirus, influenza virus, and other respiratory viruses. Use of angiotensin-converting enzyme (ACE) inhibitors, reactive airway disease, and gastroesophageal reflux disease are well-described noninfectious causes of prolonged cough in adults.

TREATMENT

TREATMENT

Pertussis

ANTIBIOTICS

The purpose of antibiotic therapy for pertussis is to eradicate the infecting bacteria from the nasopharynx; therapy does not substantially alter the clinical course unless given early in the catarrhal phase. Macrolide antibiotics are the drugs of choice for treatment of pertussis (Table 155-2); macrolide-resistant *B. pertussis* strains have been reported but are rare. Trimethoprim-sulfamethoxazole is recommended as an alternative for individuals allergic to macrolides.

SUPPORTIVE CARE

Young infants have the highest rates of complication and death from pertussis; therefore, most infants (and older children with severe disease) should be hospitalized. A quiet environment may decrease the stimulation that can trigger paroxysmal episodes. Use of β-adrenergic agonists and/or glucocorticoids has been advocated by some authorities but has not been proven to be effective. Cough suppressants are not effective and play no role in the management of pertussis.

INFECTION CONTROL MEASURES

Hospitalized patients with pertussis should be placed in respiratory isolation, with the use of precautions appropriate for pathogens spread by large respiratory droplets. Isolation should continue for 5 days after initiation of macrolide therapy or, in untreated patients, for 3 weeks (i.e., until nasopharyngeal cultures are consistently negative).

TABLE 155-2

Antimicrobial Therapy for Pertussis

Drug	Adult Daily Dose	Frequency	Duration, Days	Comments
Erythromycin estolate	1–2 g	3 divided doses	7–14	Frequent gastrointestinal side effects
Clarithromycin	500 mg	2 divided doses	7	—
Azithromycin	500 mg on day 1250 mg subsequently	1 daily dose	5	—
Trimethoprim-sulfamethoxazole	160 mg of trimethoprim, 800 mg of sulfamethoxazole	2 divided doses	14	For patients allergic to macrolides; data on effectiveness limited

PREVENTION

Chemoprophylaxis

Because the risk of transmission of *B. pertussis* within households is high, chemoprophylaxis is widely recommended for household contacts of pertussis cases. The effectiveness of chemoprophylaxis, although unproven, is supported by several epidemiologic studies of institutional and community outbreaks. In the only randomized, placebo-controlled study, erythromycin estolate (50 mg/kg per day in three divided doses; maximum dose, 1 g/d) was effective in reducing the incidence of bacteriologically confirmed pertussis by 67%; however, there was no decrease in the incidence of clinical disease. Despite these disappointing results, many authorities continue to recommend chemoprophylaxis, particularly in households with members at high risk of severe disease (children <1 year of age, pregnant women). Data on the use of the newer macrolides for chemoprophylaxis are not available, but these drugs are commonly used because of their increased tolerability and their effectiveness.

Immunization

The mainstay of pertussis prevention is active immunization **(See also Chap. 118)**. Pertussis vaccine became widely used in North America after 1940; the reported number of pertussis cases subsequently fell by >90%. Whole-cell pertussis vaccines are prepared through the heating, chemical inactivation, and purification of whole *B. pertussis* organisms. Despite their efficacy (average estimate, 85%; range for different products, 30–100%), whole-cell pertussis vaccines are associated with adverse events—both common (fever; injection-site pain, erythema, and swelling; irritability) and uncommon (febrile seizures, hypotonic-hyporesponsive episodes). Alleged associations of whole-cell pertussis vaccine with encephalopathy, sudden infant death syndrome, and autism, although not substantiated, have spawned an active anti-immunization lobby. The development of acellular pertussis vaccines, which are effective and less reactogenic, has greatly alleviated concerns about the inclusion of pertussis vaccine in the combined infant immunization series.

Although a wide variety of acellular pertussis vaccines were developed, only a few are still marketed widely; all contain pertussis toxoid and filamentous hemagglutinin. One acellular pertussis vaccine also contains pertactin, and another contains pertactin and two types of fimbriae. In phase 3 efficacy studies, multicomponent acellular pertussis vaccines were more efficacious than one- or two-component vaccines. However, epidemiologic studies in countries using one- and two-component acellular pertussis vaccines demonstrated high vaccine effectiveness against pertussis. Adult formulations of acellular pertussis vaccines have been shown to be safe, immunogenic, and efficacious in clinical trials in adolescents and adults and are now recommended for routine immunization of these groups in several countries.


Although whole-cell vaccines are still used extensively in developing regions of the world, acellular pertussis vaccines are used exclusively for childhood immunization in much of the developed world. In light of evidence of early waning of immunity among children who received acellular pertussis vaccine in infancy, the WHO Strategic Advisory Group of Experts (SAGE) now recommends using whole-cell pertussis vaccine for the primary infant immunization series. In countries using acellular pertussis vaccines in infancy, additional booster immunizations in older children, adolescents, and adults are recommended to prevent pertussis in high-risk infants. In North America, acellular pertussis vaccines for children are given as a three-dose primary series at 2, 4, and 6 months of age, with a reinforcing dose at 15–18 months of age and a booster dose at 4–6 years of age. Adolescents (11–18 years of age) and all unvaccinated adults should receive a dose of the adult-formulation diphtheria–tetanus–acellular pertussis vaccine. In addition, several countries, including the United States and the United Kingdom, recommend pertussis immunization specifically for health care workers and for women in the third trimester of pregnancy to increase passive transfer of maternal antibodies to the fetus. In a study from the United Kingdom, immunization of pregnant women ≥7 days prior to delivery was 91–93% effective at preventing pertussis in infants ≤3 months of age. Pertussis vaccine coverage among U.S. adolescents was 84.6% in 2012, and coverage among pregnant

women is increasing (from 12.7% in 2010 to 41.7% in 2013 in a cohort of privately insured women). Nevertheless, coverage among adults remains low (14.2% in 2012). Further improvements in adult vaccine coverage may permit better control of pertussis across the age spectrum, with collateral protection of infants too young to be immunized. However, more effective vaccines with longer-lasting protection will ultimately be needed to control this disease.

FURTHER READING

De Serres G et al: Morbidity of pertussis in adolescents and adults. J Infect Dis 182:174, 2000.

Kilgore PE et al: Pertussis: Microbiology, disease, treatment, and prevention. Clin Microbiol Rev 29:449, 2016.

Klein NP et al: Waning Tdap effectiveness in adolescents. Pediatrics 137:e20153326, 2016. [PubMed: 26908667]

Skoff TH: Sources of infant pertussis infection in the United States. Pediatrics 136:635, 2015.

World Health Organization: Pertussis vaccines: WHO position paper—August 2015. Wkly Epidemiol Rec 90:433, 2015.

Chapter 156: Diseases Caused by Gram-Negative Enteric Bacilli

Thomas A. Russo; James R. Johnson

FIGURE 156-1

GENERAL FEATURES AND PRINCIPLES

The post-antibiotic era has begun. For most people, this is the first time in their lives that an effective treatment for a bacterial infection may not exist. The Enterobacteriaceae are at the forefront of this evolving public health crisis. For example, the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) have designated carbapenem-resistant Enterobacteriaceae as representing a threat level of "urgent" and "priority one, critical," respectively. Enterobacteriaceae are responsible for a significant proportion of the deaths attributed to resistant bacteria, the number of which has been estimated at 23,000 and 25,000 annually in the United States and the European Union, respectively, with numbers three- to fivefold greater (per capita) in low- and middle-income countries (e.g., Thailand). These pathogens cause a wide variety of infections involving diverse anatomic sites in both healthy and compromised hosts. Therefore, a thorough knowledge of clinical presentations and appropriate therapeutic choices is necessary for optimal outcomes. *Escherichia coli*, *Klebsiella*, *Proteus*, *Enterobacter*, *Serratia*, *Citrobacter*, *Morganella*, *Providencia*, *Cronobacter*, and *Edwardsiella* are enteric gram-negative bacilli (GNB) that are members of the family Enterobacteriaceae. *Salmonella*, *Shigella*, and *Yersinia*, also in the family Enterobacteriaceae, are discussed in **Chaps. 160**, **161**, **and 166**, respectively.

EPIDEMIOLOGY



E. coli, *Klebsiella*, *Proteus*, *Enterobacter*, *Serratia*, *Citrobacter*, *Morganella*, *Providencia*, *Cronobacter*, and *Edwardsiella* are components of the normal animal and human colonic microbiota and/or the microbiota in various environmental habitats, including long-term-care facilities (LTCFs) and hospitals. As a result, except for certain pathotypes of intestinal pathogenic *E. coli*, these genera are global pathogens. The incidence of infection due to these agents is increasing because of the combination of an aging population and increasing antimicrobial resistance. In healthy humans, *E. coli* is the predominant species of GNB in the colonic microbiota; *Klebsiella* and *Proteus* are less prevalent. GNB (primarily *E. coli*, *Klebsiella*, and *Proteus*) colonize the oropharynx and skin of healthy individuals only transiently. By contrast, in LTCFs and hospital settings, a variety of GNB emerge as the dominant colonizers of both mucosal and skin surfaces, particularly in association with antimicrobial use, severe illness, and extended length of stay. LTCFs are emerging as an

important reservoir for resistant GNB. This colonization may lead to subsequent infection; for example, oropharyngeal colonization may lead to pneumonia, and colonic/perineal colonization may lead to urinary tract infection (UTI). The use of ampicillin or amoxicillin was associated with an increased risk of subsequent infection due to the hypervirulent pathotype of *Klebsiella pneumoniae* in Taiwan; this association suggests that changes in the quantity or prevalence of colonizing bacteria may significantly influence the risk of infection. *Serratia* and *Enterobacter* infection may be acquired directly through a variety of infusates (e.g., medications, blood products). *Edwardsiella* infections are acquired through freshwater and marine environment exposures and are most common in Southeast Asia.

STRUCTURE AND FUNCTION

Enteric GNB possess an extracytoplasmic outer membrane consisting of a lipid bilayer with associated proteins, lipoproteins, and polysaccharides (capsule, lipopolysaccharide). The outer membrane interfaces with the external environment, including the human host. A variety of components of the outer membrane are critical determinants in pathogenesis (e.g., capsule) and antimicrobial resistance (e.g., permeability barrier, efflux pumps). In addition, secreted products play an important role in both host infection (e.g., iron acquisition molecules) and environmental niche survival and colonization (e.g., type VI secretion systems).

PATHOGENESIS

Multiple bacterial virulence factors are required for the pathogenesis of infections caused by GNB. Possession of specialized virulence genes defines pathogens and enables them to infect the host efficiently. Hosts and their cognate pathogens have been co-adapting throughout evolutionary history. During the host–pathogen "chess match" over time, various and redundant strategies have emerged in both the pathogens and their hosts **(Table 156-1)**.

TABLE 156-1

Interactions of Extraintestinal Pathogenic *Escherichia coli* with the Human Host: A Paradigm for Extracellular, Extraintestinal Gram-Negative Bacterial Pathogens

Bacterial Goal	Host Obstacle	Bacterial Solution
Extraintestinal attachment	Flow of urine, mucociliary escalator	Multiple adhesins (e.g., type 1, S, and F1C fimbriae; P pili)
Nutrient acquisition for growth	Nutrient sequestration (e.g., iron via intracellular storage and extracellular scavenging via lactoferrin and transferrin)	Cellular lysis (e.g., hemolysin), multiple mechanisms for competing for iron (e.g., siderophores) and other nutrients
Initial avoidance of host bactericidal activity	Complement, phagocytic cells, antimicrobial peptides	Capsular polysaccharide, lipopolysaccharide
Dissemination (within host and between hosts)	Intact tissue barriers	Irritant tissue damage resulting in increased excretion (e.g., toxins such as hemolysin), invasion of brain endothelium
Late avoidance of host bactericidal activity	Acquired immunity (e.g., specific antibodies), treatment with antibiotics	Cell entry, acquisition of antimicrobial resistance

Intestinal pathogenic (diarrheagenic) mechanisms are discussed below. The members of the Enterobacteriaceae family that cause extraintestinal infections are primarily extracellular pathogens and therefore share certain pathogenic features. The principal components of host defense against Enterobacteriaceae, regardless of species, are innate immunity (including intact skin and mucosal barriers; the withholding of nutrients; and the activities of complement, antimicrobial peptides, and professional phagocytes) and humoral immunity. Both susceptibility to and severity of infection are increased with dysfunction or deficiencies of these host components. By contrast, the virulence traits of intestinal pathogenic *E. coli*—i.e., the distinctive strains that can cause diarrheal disease—are for the most part different from those of extraintestinal pathogenic *E. coli* (ExPEC) and other GNB that cause extraintestinal infections. This distinction reflects site-specific differences in host environments and defense mechanisms.

A given enterobacterial strain usually possesses multiple adhesins for binding to a variety of host cells (e.g., in *E. coli*: type 1, S, and F1C fimbriae; P pili). Nutrient acquisition (e.g., of iron via siderophores) requires many genes that are necessary but not sufficient for pathogenesis. The ability to resist the bactericidal activity of complement and phagocytes in the absence of antibody (e.g., as conferred by capsule or the O antigen component of lipopolysaccharide) is one of the defining traits of an extracellular pathogen. Tissue damage (e.g., as mediated by *E. coli* hemolysin) may facilitate nutrient acquisition and spread within the host. Without doubt, many important virulence genes await identification **(Chap. 116)**.

The ability to induce septic shock is another defining feature of these genera. GNB are the most common causes of this potentially lethal syndrome. Pathogen-associated molecular pattern molecules (PAMPs; e.g., the lipid A moiety of lipopolysaccharide) stimulate a proinflammatory host response via pattern recognition receptors (e.g., Toll-like or C-type lectin receptors) that activate host defense signaling pathways; if overly exuberant, this response results in shock **(Chap. 297)**. Direct bacterial damage of host tissue (e.g., by toxins) or collateral damage from the host response can result in the release of damage-associated molecular pattern molecules (DAMPs; e.g., HMGB1) that can propagate a detrimental proinflammatory host response.

Many antigenic variants (serotypes) exist in most genera of GNB. For example, *E. coli* has more than 150 O (somatic) antigens, 80 K (capsular) antigens, and 53 H (flagellar) antigens. This antigenic variability, which permits immune evasion and allows recurrent infection by different strains of the same species, has impeded vaccine development **(Chap. 118)**.

INFECTIOUS SYNDROMES

Depending on both the host and the pathogen, GNB can infect nearly every organ or body cavity. *E. coli* can cause either intestinal or extraintestinal infection, depending on the particular pathotype, and *Edwardsiella tarda* can cause both intestinal and extraintestinal infection. *Klebsiella* causes primarily extraintestinal infection, but a toxin-producing variant of *Klebsiella oxytoca* has been associated with hemorrhagic colitis.

E. coli and—to a lesser degree—*Klebsiella* account for most extraintestinal infections due to GNB. These species (for *K. pneumoniae*, primarily its hypervirulent pathotype) are the most virulent pathogens within this group, as demonstrated by their ability to cause severe infections in healthy, ambulatory hosts from the community. However, the other genera of GNB are also important extraintestinal pathogens, especially among LTCF residents and hospitalized patients, in large part because of the intrinsic or acquired antimicrobial resistance of these organisms and the increasing number of individuals with compromised host defenses. The mortality rate is substantial in many GNB infections and correlates with severity of illness and underlying host status. Especially problematic are pneumonia, sepsis, and septic shock (arising from any site of infection), for which the associated mortality rates are 20–60%.

DIAGNOSIS

Isolation of GNB from sterile sites almost always implies infection, whereas their isolation from nonsterile sites, particularly from open wounds and the respiratory tract, requires clinical correlation to differentiate

colonization from infection. Clinical microbiology laboratories are increasingly incorporating newer molecular-based methodologies (e.g., matrix-assisted laser desorption–ionization–time-of-flight mass spectrometry [MALDI-TOF-MS] and polymerase chain reaction [PCR]) to enhance the sensitivity, accuracy, and rapidity of reporting on pathogen identification and resistance genes (e.g., *bla*KPC, NDM, OXA, CTX). This information can be used to increase the timeliness of initiation and/or the accurate selection of empirical antimicrobial therapy, thereby improving outcomes.

TREATMENT

TREATMENT

Infections Caused by Gram-Negative Enteric Bacilli

 Initiation of appropriate empirical antimicrobial therapy early in the course of GNB infections (particularly serious ones) leads to improved outcomes **(See also Chap. 139)**. The ever-increasing prevalence of multidrug-resistant (MDR) and extensively drug-resistant (XDR) GNB; the lag between published and current resistance rates; and variations in antimicrobial susceptibility by species, geographic location, regional antimicrobial use, and hospital site (e.g., intensive care units [ICUs] versus wards) necessitate familiarity with evolving patterns of antimicrobial resistance for the selection of appropriate empirical therapy. Factors predictive of resistance in a given isolate include recent antimicrobial use, a health care association (e.g., recent or ongoing hospitalization, dialysis, residence in an LTCF), or international travel (e.g., to Asia, Latin America, Africa, Eastern Europe). Resistance rates will almost certainly increase over time and will likely be higher than shown here by the time this chapter is published. Data for 2008–2014 from the U.S. National Healthcare Safety Network indicates that the prevalence of the extended-spectrum β-lactamase (ESBL) phenotype among Enterobacteriaceae isolates varied by health care setting—i.e., 16% for short-term care, 38.6% for long-term care, and 10.7% for inpatient rehabilitation facilities—as did the prevalence of carbapenem resistance (2.8%, 12%, and 1.9%, respectively). Global ESBL rates for Enterobacteriaceae isolates from hospitalized patients were roughly similar in North America, Western Europe, Australia, and New Zealand and were higher in Latin America, Eastern Europe, and Asia. Perhaps even more concerning is the reported isolation of carbapenem-resistant Enterobacteriaceae (mediated primarily by New Delhi metallo-β-lactamase [NDM]) from ambulatory patients without known risk factors.

For appropriately selected patients, it may be prudent initially, pending antimicrobial susceptibility results, to use two potentially active agents as a way to increase the likelihood that at least one agent will be active against the patient's organism. If broad-spectrum treatment has been initiated, it is important to switch to the most appropriate narrower-spectrum agent once antimicrobial susceptibility results become available. Such responsible antimicrobial stewardship should help disrupt the ever-escalating cycle of selection for increasingly resistant bacteria, decrease the likelihood of *Clostridium difficile* infection, decrease costs, and maximize the useful longevity of available antimicrobial agents. Likewise, it is important to avoid treatment of patients who are colonized but not infected (e.g., who have a positive sputum culture without evidence of pneumonia or a positive urine culture without clinical manifestations of UTI).

At present, the most reliably active antimicrobial agents against GNB are the carbapenems (e.g., meropenem); the aminoglycoside amikacin; the fourth-generation cephalosporin cefepime; the β-lactamase inhibitor combination agents piperacillin-tazobactam, ceftolozane-tazobactam, and ceftazidime-avibactam; and the polymyxins (colistin and polymyxin B). However, it should be noted that *Proteus*, *Serratia*, *Morganella*, and *Providencia* are intrinsically resistant to the polymyxins. The number of antimicrobial agents effective against certain Enterobacteriaceae is shrinking, and truly pan-resistant GNB exist. Accordingly, the currently available antimicrobial drugs must be used judiciously.

β-Lactamases, which inactivate β-lactam agents, are the most important mediators of β-lactam resistance in GNB. Decreased permeability and/or active efflux of β-lactam agents, although less important and less potent, may occur alone or in combination with β-lactamase-mediated resistance.

Broad-spectrum β-*lactamases* (e.g., TEM, SHV), which mediate resistance to many penicillins and first-generation cephalosporins, are frequently expressed in enteric GNB. These enzymes are inhibited by β-lactamase inhibitors (e.g., clavulanate, sulbactam, tazobactam, avibactam). In their wild-type form, they do not hydrolyze third- and fourth-generation cephalosporins or cephamycins (e.g., cefoxitin). However, molecular variants of TEM and SHV that have amino acid replacements at certain critical positions in the peptide do exhibit such hydrolytic capability and thus are referred to as ESBLs, as discussed below.

 *ESBLs* (e.g., CTX-M, SHV, TEM) are modified broad-spectrum enzymes that hydrolyze third-generation cephalosporins, aztreonam, and (in some instances) fourth-generation cephalosporins in addition to the drugs hydrolyzed by broad-spectrum β-lactamases. GNB that express ESBLs may also possess porin mutations that result in decreased uptake of cephalosporins, β-lactam/β-lactamase inhibitor combinations, and carbapenems, thereby further reducing susceptibility to these β-lactam agents. The prevalence of acquired ESBL production, particularly of CTX-M-type enzymes, is increasing in GNB worldwide, in large part as a result of the presence of the corresponding genes on transferable (conjugal) plasmids variably linked to or associated with resistance to fluoroquinolones, trimethoprim-sulfamethoxazole (TMP-SMX), aminoglycosides, tetracyclines, and (more recently) fosfomycin.

To date, ESBLs are most prevalent in *E. coli*, *K. pneumoniae*, and *K. oxytoca*, but these enzymes can occur in all Enterobacteriaceae. The approximate regional prevalence of ESBL-producing GNB currently follows a descending gradient as follows: China > Eastern Europe > other parts of Asia > Latin America and Africa > Western Europe, the United States, Canada, and Australia. International travel to high-prevalence regions increases the likelihood of colonization with these strains.

ESBL-producing GNB were described initially in hospitals (ICUs > wards) and LTCFs, where outbreaks occurred in association with extensive use of third-generation cephalosporins. However, over the last decade, the incidence of UTI due to CTX-M ESBL-producing *E. coli* has increased worldwide (including in the United States), even among healthy ambulatory women without health care or antimicrobial exposure. Antimicrobial use in food animals has also been implicated in the rise of ESBLs.

Carbapenems are the most reliably active β-lactam agents against ESBL-expressing strains. Clinical experience with alternatives is more limited, but, for organisms susceptible to piperacillin-tazobactam (minimal inhibitory concentration [MIC], ≤4 µg/mL), this agent—dosed at 4.5 g q6h—may offer a carbapenem-sparing alternative, as may ceftazidime-avibactam and ceftolozane-tazobactam.

The role of tigecycline is unclear despite its excellent in vitro activity; *Proteus*, *Morganella*, and *Providencia* are inherently resistant, and attainable serum and urine levels are low. Therefore, caution is advisable, especially with serious infections, until more clinical data become available.

Oral options for the treatment of strains expressing ESBLs are limited. Fosfomycin and nitrofurantoin (for *E. coli*) and perhaps pivmecillinam (not available in the United States) are the most reliably active agents.

AmpC β-lactamases, when induced or stably derepressed to high levels of expression, confer resistance to the same substrates as do ESBLs as well as to the cephamycins (e.g., cefoxitin and cefotetan). The genes encoding these enzymes are primarily chromosomal and therefore may not exhibit the linked or associated resistance to fluoroquinolones, TMP-SMX, aminoglycosides, and tetracyclines that is common with ESBLs. These enzymes are problematic for the clinician: resistance may develop during therapy with third-generation cephalosporins and result in clinical failure, particularly in the setting of bacteremia.

Although chromosomal AmpC β-lactamases are present in nearly all members of the Enterobacteriaceae family, the risk of clinically significant induction of high-level expression or selection of stably derepressed mutants with cephalosporin treatment is greatest with *Enterobacter cloacae* and *Enterobacter aerogenes*, lower with *Serratia marcescens* and *Citrobacter freundii*, and lowest with *Providencia* and *Morganella morganii.* In addition, rare strains of *E. coli*, *K. pneumoniae*, and other Enterobacteriaceae have acquired plasmids containing inducible AmpC β-lactamase genes.

For AmpC-expressing strains, carbapenems are an appropriate treatment option. Ceftazidime-avibactam and ceftolozane-tazobactam are active in vitro, but clinical data are limited. The fourth-generation cephalosporin cefepime may be an appropriate option if the concomitant presence of an ESBL can be excluded (a task that currently exceeds the capability of most clinical microbiology laboratories) and source control is achieved. Although clinical data are limited, other carbapenem-sparing alternatives to consider if isolates are susceptible in vitro include fluoroquinolones, piperacillin-tazobactam, TMP-SMX, tigecycline, and aminoglycosides.

 *Carbapenemases*—e.g., class A (*Klebsiella pneumoniae* carbapenemase [KPC]); class B (NDM; Verona integron–encoded metallo-β-lactamase [VIM]; and imipenemase metallo-β-lactamase [IMP]); and class D [OXA-48])—confer resistance to the same drugs as do ESBLs as well as to cephamycins and carbapenems. As with ESBLs, carbapenemase-encoding genes may be present on transferable plasmids, which often encode linked resistance to fluoroquinolones, TMP-SMX, tetracyclines, and aminoglycosides. Transposon-mediated spread (e.g., TN4401 for KPC) is also important. Unfortunately, carbapenemase-producing Enterobacteriaceae are becoming increasingly common, particularly in Asia. Asymptomatic intestinal carriage may facilitate spread.

Carbapenemase production by Enterobacteriaceae is most prevalent in *K. pneumoniae* and, secondarily, in *E. coli*, but has been described in nearly all members of the family. Carbapenem resistance may also occur in the absence of carbapenemase production, mediated by production of an AmpC β-lactamase and/or ESBL along with modifications in permeability/efflux. Resistance to any carbapenem should prompt assessment for carbapenemase production via either genotypic or phenotypic tests, if available; the exception to this rule is isolated resistance to imipenem in *M. morganii*, *Proteus*, and *Providencia*, which exhibit intrinsic low-level resistance. Although the modified Hodge test is used widely for phenotypic confirmation of carbapenemase production, its limitations include false-positive results with *Enterobacter* species and false-negative results with NDM.

For treatment of infections due to carbapenem-resistant Enterobacteriaceae, tigecycline and colistin are the most reliably active parenteral agents in vitro. However, because tigecycline reaches only low serum and urine concentrations, caution is warranted in using it to treat bacteremia and perhaps UTI, although a few case reports describe some success with tigecycline therapy for UTI. Colistin has nephrotoxic and neurotoxic potential. The recent emergence of the colistin resistance gene *mcr-1* on a stable transferable plasmid is extremely concerning since polymyxins (polymyxin B and polymyxin E [colistin]) currently constitute a last line of defense against strains that produce metallo-carbapenemases (e.g., NDM-1). In addition, in a recent study, 13% of carbapenem-resistant *K. pneumoniae* isolates were co-resistant to colistin independent of *mcr-1*.

Ceftazidime-avibactam is active in vitro against the serine carbapenemases (e.g., KPC, OXA-48) but not the metallo-carbapenemases (e.g., NDM, VIM, IMP). However, limited clinical data from an uncontrolled retrospective study of ceftazidime-avibactam for the treatment of infection with carbapenem-resistant Enterobacteriaceae demonstrated suboptimal efficacy and development of resistance in 8% of the cohort. Aminoglycosides may have some utility for combination therapy if they are active in vitro. Fosfomycin is often active in vitro, but clinical data are limited. Furthermore, resistance may develop with monotherapy and increased use, plasmid-mediated resistance (via *fosA3*) has been described (raising concern about rapid dissemination), resistance is generally more prevalent among XDR strains, susceptibility testing may not be readily available, and no parenteral formulation is available in the United States. Aztreonam is active against the problematic metallo-carbapenemases but is hydrolyzed by ESBLs and AmpC β-lactamases, which often co-exist in XDR strains. Ongoing studies are assessing aztreonam plus avibactam, a promising combination for the treatment of pan-drug-resistant strains.

Extensive resistance to available agents may leave the clinician with few or no ideal therapeutic options. However, use of a regimen that takes into account the site of infection, achievable drug levels at that site (e.g., higher concentrations of many agents in urine), and pharmacodynamic factors (e.g., prolonged infusion of β-lactam agents to maintain drug levels above the MIC) may increase the chance for a successful outcome. Likewise, observational data suggest that combination therapy may be beneficial against carbapenem-resistant Enterobacteriaceae; randomized controlled trials are in progress.

Resistance to fluoroquinolones usually is due to alterations in or protection of the target sites in DNA gyrase and topoisomerase IV, with or without decreased permeability and active efflux. Fluoroquinolone resistance

is increasingly prevalent among GNB and is associated with resistance to other antimicrobial classes; for example, 20–80% of ESBL-producing enteric GNB are also resistant to fluoroquinolones. At present, fluoroquinolones should be considered unreliable as empirical therapy for GNB infections in critically ill patients.

In this era of increasing antimicrobial resistance, it is critical to culture the primary site of infection before initiating antimicrobial therapy and, for systemically ill patients, to obtain blood cultures. In vitro testing may not always detect antimicrobial resistance; therefore, it is important to assess the patient's clinical response to treatment. Moreover, as discussed above, resistance may emerge during therapy through the induction or stable derepression of AmpC β-lactamases. In addition, drainage of abscesses, resection of necrotic tissue, and removal of infected foreign bodies, sometimes referred to collectively as "source control," are often required for cure.

GNB are commonly involved in polymicrobial infections, in which the role of each individual pathogen is uncertain **(Chap. 172)**. Although some GNB are more pathogenic than others, it is usually prudent, if possible, to design an antimicrobial regimen active against all of the GNB identified, because each is capable of pathogenicity in its own right. Lastly, for patients treated initially with a broad-spectrum empirical regimen, the regimen should be de-escalated as expeditiously as possible once susceptibility results are known and the patient has responded to therapy.

PREVENTION

Certain measures are broadly applicable for decreasing infection risk **(See also Chap. 137)**. Antimicrobial stewardship programs should be instituted to facilitate appropriate antimicrobial use, which will minimize the development of resistance. Diligent adherence to hand-hygiene protocols by health care personnel and cleaning/disinfection or single-patient use of objects that come into contact with patients (e.g., stethoscopes and blood pressure cuffs) are essential. Indwelling devices (e.g., urinary and intravascular catheters) should be used only when necessary and inserted according to an appropriate protocol; protocols for daily-use evaluation and prompt removal should be implemented. Multi-use medication vials should be avoided if possible. Oral application of chlorhexidine decreases the incidence of pneumonia among patients on ventilators. Increasing data support the implementation of universal decolonization (e.g., chlorhexidine bathing) to prevent infection in ICU patients. The public health threat from carbapenem-resistant Enterobacteriaceae has resulted in additional recommendations, especially for carbapenemase-producing carbapenem-resistant Enterobacteriaceae, which are an even greater concern. These recommendations include contact precautions for patients colonized or infected with carbapenem-resistant Enterobacteriaceae, notification to the receiving facility from facilities transferring a patient colonized or infected with these organisms, and daily environmental cleaning. Screening of contacts and active surveillance for these bacteria may also be appropriate.

ESCHERICHIA COLI INFECTIONS

 All *E. coli* strains share a core genome of ~2000 genes. In contrast, an *E. coli* strain's ability to cause infection and the nature of such infections are defined largely by accessory (i.e., non-core, non-essential) genes that encode various virulence factors. The composition of the *E. coli* accessory genome is fluid and ongoing, as demonstrated by the recent evolution of Shiga toxin–producing enteroaggregative *E. coli*.

COMMENSAL STRAINS

Commensal *E. coli* variants are an important constituent of the normal intestinal microbiota that confer benefits to the host (e.g., resistance to colonization with pathogenic organisms). Such strains generally lack the specialized virulence traits that enable extraintestinal and intestinal pathogenic *E. coli* strains to cause disease outside and within the gastrointestinal tract, respectively. However, even commensal *E. coli* strains can be involved in extraintestinal infections in the presence of an aggravating factor, such as a foreign body (e.g., a urinary catheter), host compromise (e.g., local anatomic or functional abnormalities [including urinary or biliary tract obstruction] or systemic immunocompromise), or an inoculum that is large or contains a mixture of bacterial species (e.g., fecal contamination of the peritoneal cavity).

EXTRAINTESTINAL PATHOGENIC STRAINS

 ExPEC strains are the most common enteric GNB to cause community-acquired and health care–associated bacterial infections. The emerging propensity of these strains to acquire new mechanisms of antimicrobial resistance (e.g., ESBLs and carbapenemases) poses novel challenges in managing ExPEC infection. Several ExPEC clonal groups (e.g., ST131, ST95, ST69, and ST73) are recognized to have undergone global dissemination. The mechanisms underlying the epidemiologic success of such disseminated lineages presumably include superior biological fitness and acquisition of antimicrobial resistance, as demonstrated by members of the ST131 subclone *H*30-Rx, which are resistant to fluoroquinolones and usually express the ESBL CTX-M-15. Reservoirs and transmission pathways are an active area of study, but human-to-human, food-to-human (e.g., pork, turkey, and chicken), and perhaps environment-to-human are most likely.

Like commensal *E. coli* (but unlike intestinal pathogenic *E. coli*), ExPEC strains are often found in the intestinal microbiota of healthy individuals and do not cause gastroenteritis in humans. Entry from their site of colonization (e.g., the colon, vagina, or oropharynx) into a normally sterile extraintestinal site (e.g., the urinary tract, peritoneal cavity, or lungs) is the rate-limiting step for infection. ExPEC strains have acquired accessory genes encoding diverse extraintestinal virulence factors that enable the bacteria to cause infections outside the gastrointestinal tract in both normal and compromised hosts (Table 156-1). These virulence genes define ExPEC and, for the most part, are distinct from the virulence genes that enable intestinal pathogenic strains to cause diarrheal disease **(Table 156-2)**. All age groups, all types of hosts, and nearly all organs and anatomic sites are susceptible to infection by ExPEC. Even previously healthy hosts can become severely ill or die when infected with ExPEC; however, adverse outcomes are more common among hosts with comorbid illnesses and host defense abnormalities. The diversity and the medical and economic impact of ExPEC infections are evident from consideration of the following specific syndromes.

TABLE 156-2

Intestinal Pathogenic *E. coli*

Pathotype	Epidemiology	Clinical Syndrome[a]	Defining Molecular Trait	Responsible Genetic Element[b]
STEC/EHEC/ST-EAEC	Food, water, person-to-person; all ages, industrialized countries	Hemorrhagic colitis, hemolytic-uremic syndrome	Shiga toxin	Lambda-like Stx1- or Stx2-encoding bacteriophage
ETEC	Food, water; young children in and travelers to developing countries	Traveler's diarrhea	Heat-stable and labile enterotoxins, colonization factors	Virulence plasmid(s)
EPEC	Person-to-person; young children and neonates in developing countries	Watery diarrhea, persistent diarrhea	Localized adherence, attaching and effacing lesion on intestinal epithelium	EPEC adherence factor plasmid pathogenicity island (locus for enterocyte effacement [LEE])
EIEC	Food, water; children in and travelers to developing countries	Watery diarrhea, occasionally dysentery	Invasion of colonic epithelial cells, intracellular multiplication, cell-to-cell spread	Multiple genes contained primarily in a large virulence plasmid
EAEC	?Food, water; children in and travelers to developing countries; all ages, industrialized countries	Traveler's diarrhea, acute diarrhea, persistent diarrhea	Aggregative/diffuse adherence, virulence factors regulated by AggR	Chromosomal or plasmid-associated adherence and toxin genes

[a]Classic syndromes; see text for details on disease spectrum. [b]Pathogenesis involves multiple genes, including genes in addition to those listed.

Abbreviations: EAEC, enteroaggregative *E. coli;* EHEC, enterohemorrhagic *E. coli*; EIEC, enteroinvasive *E. coli*; EPEC, enteropathogenic *E. coli*; ETEC, enterotoxigenic *E. coli*; ST-EAEC, Shiga toxin–producing enteroaggregative *E. coli*; STEC, Shiga toxin–producing *E. coli*.

Extraintestinal Infectious Syndromes

URINARY TRACT INFECTION

The urinary tract is the site most frequently infected by ExPEC. UTI is an exceedingly common infection among ambulatory patients, accounting for 1% of ambulatory care visits in the United States and second only to lower respiratory tract infection among infections responsible for hospitalization. UTIs are best considered by clinical syndrome (e.g., cystitis, pyelonephritis, and catheter-associated UTI) and within the context of specific hosts (e.g., premenopausal women, compromised hosts; **Chap. 130**). *E. coli* is the single most common pathogen for all UTI syndrome/host group combinations. Each year in the United States, *E. coli* causes 80–90% of the estimated 6–8 million episodes of cystitis that occur in ambulatory, premenopausal women with an anatomically and functionally normal urinary tract (i.e., uncomplicated cystitis). Furthermore, 20% of women with an initial cystitis episode develop frequent recurrences.

Uncomplicated cystitis, the most common acute UTI syndrome, is characterized by dysuria, urinary frequency and urgency, and suprapubic pain. Fever and/or back pain suggests progression to pyelonephritis. Even when pyelonephritis is treated effectively, fever may take 5–7 days to resolve completely. Persistently elevated or increasing fever and neutrophil counts should prompt evaluation for intrarenal or perinephric abscess and/or obstruction. Pyelonephritis uncommonly causes renal parenchymal damage and loss of renal function, primarily in association with urinary obstruction, which can be preexisting or, rarely, occurs de novo in diabetic patients who develop renal papillary necrosis as a result of kidney infection. Pregnant women are at unusually high risk for developing pyelonephritis, which can adversely affect the outcome of pregnancy. As a result, prenatal screening for and treatment of asymptomatic bacteriuria during pregnancy are standard. Prostatic infection (prostatitis), a potential complication of UTI in men, can present in either an acute (severe) or a chronic (recurrent cystitis) manner. Acute pyelonephritis, acute prostatitis, and other systemic illnesses due to UTI can be designated collectively as *urosepsis*, *febrile UTI*, or *systemic UTI*. The diagnosis and treatment of UTI, as detailed in **Chap. 130**, should be tailored to the individual host, the nature and site of infection, and local patterns of antimicrobial susceptibility.

ABDOMINAL AND PELVIC INFECTION

The abdomen/pelvis is the second most common site of extraintestinal infection due to *E. coli*. A wide variety of clinical syndromes occur in this location, including acute peritonitis secondary to fecal contamination, spontaneous bacterial peritonitis, dialysis-associated peritonitis, diverticulitis, appendicitis, intraperitoneal or visceral abscesses (hepatic, pancreatic, splenic), infected pancreatic pseudocysts, and septic cholangitis and/or cholecystitis. In intraabdominal infections, *E. coli* can be isolated either alone or, as occurs more often, in combination with other facultative and/or anaerobic members of the intestinal microbiota **(Chap. 127)**.

PNEUMONIA

E. coli is not usually considered an important cause of pneumonia (Chap. 121). Indeed, enteric GNB account for only 1–3% of cases of community-acquired pneumonia, in part because these organisms colonize the oropharynx only transiently in a minority of healthy individuals. However, rates of oral colonization with *E. coli* and other GNB increase with severity of illness and antibiotic use. Consequently, GNB are a more common cause of pneumonia among residents of LTCFs and are the most common cause (60–70% of cases) of hospital-acquired pneumonia (Chap. 137), particularly among postoperative and ICU patients (e.g., ventilator-associated pneumonia).

Pulmonary infection is usually acquired by small-volume aspiration but occasionally occurs via hematogenous spread, in which case multifocal nodular infiltrates can be seen. Tissue necrosis, probably due to bacterial cytotoxins, is common. Despite significant institutional variation, *E. coli* is generally the third or fourth most commonly isolated type of GNB in hospital-acquired pneumonia, accounting for 5–8% of episodes in both U.S.-based and Europe-based studies. Regardless of the host, pneumonia due to ExPEC is a serious disease, with high crude and attributable mortality rates (20–60% and 10–20%, respectively).

MENINGITIS

E. coli is one of the two leading causes of neonatal meningitis, together with group B *Streptococcus* (See also Chap. 133). Most *E. coli* strains that cause neonatal meningitis possess the K1 capsular antigen and derive from a limited number of meningitis-associated clonal groups. Ventriculomegaly occurs commonly. After the first month of life, *E. coli* meningitis is uncommon, occurring predominantly in the setting of surgical or traumatic disruption of the meninges or hepatic cirrhosis. In patients with cirrhosis who develop meningitis, the meninges are presumably seeded as a result of poor hepatic clearance of portal vein bacteremia.

CELLULITIS/MUSCULOSKELETAL INFECTION

E. coli contributes frequently to infections of decubitus ulcers and occasionally to infections of lower-extremity ulcers and wounds in diabetic patients and other hosts with neurovascular compromise. Osteomyelitis secondary to contiguous spread can occur in these settings. *E. coli* also causes cellulitis or infections of burn sites and surgical wounds (accounting for ~10% of surgical site infections), particularly when the infection originates close to the perineum. *E. coli* causes hematogenously acquired osteomyelitis, especially of vertebral discs and bodies, accounting for up to 10% of cases in some series (Chap. 126). *E. coli* occasionally causes orthopedic device–associated infection or septic arthritis and rarely causes hematogenous myositis. Myositis or fasciitis of the thigh due to *E. coli* should prompt an evaluation for an abdominal source with contiguous spread.

ENDOVASCULAR INFECTION

Despite being one of the most common causes of bacteremia, *E. coli* rarely seeds native heart valves. When the organism does infect native valves, it usually does so in the setting of prior valvular disease. *E. coli* infections of aneurysms, the portal vein (*pylephlebitis*), and vascular grafts are quite uncommon.

MISCELLANEOUS INFECTIONS

E. coli can cause infection in nearly every organ and anatomic site. It occasionally causes postoperative mediastinitis or complicated sinusitis and uncommonly causes endophthalmitis, ecthyma gangrenosum, or brain abscess.

BACTEREMIA

E. coli bacteremia can arise from infection at any extraintestinal site. In addition, *E. coli* bacteremia can arise from percutaneous intravascular devices or transrectal prostate biopsy or from the increased intestinal mucosal permeability seen in neonates and in the settings of neutropenia, chemotherapy-induced mucositis, trauma, and burns. Roughly equal proportions of *E. coli* bacteremia cases originate in the community and in health care settings. Isolation of *E. coli* from the blood is almost always clinically significant and may be accompanied by the sepsis syndrome (dysfunction of at least one organ or system) or septic shock **(Chap. 297)**.

The urinary tract is the most common source for *E. coli* bacteremia, accounting for one-half to two-thirds of episodes. Bacteremia from a urinary tract source is particularly common among patients with pyelonephritis, urinary tract obstruction, or urinary instrumentation in the presence of infected urine. The abdomen is the second most common source, accounting for ~25% of episodes. Although many of these episodes result from biliary obstruction (stones, tumor) and overt bowel disruption, which typically are readily apparent, some abdominal sources (e.g., abscesses) are remarkably silent clinically and require identification via imaging studies (e.g., computed tomography). Therefore, especially given the high prevalence of asymptomatic bacteriuria among elderly and functionally compromised individuals, the physician should be cautious in attributing *E. coli* bacteremia to a urinary source in the absence of characteristic signs and symptoms of UTI. Soft tissue, bone, pulmonary, and intravascular catheter infections are other sources of *E. coli* bacteremia.

Diagnosis

Strains of *E. coli* that cause extraintestinal infections usually grow both aerobically and anaerobically within 24 h on standard diagnostic media and are identified readily by the clinical microbiology laboratory according to routine biochemical criteria. More than 90% of ExPEC strains are rapid lactose fermenters and are indole-positive.

TREATMENT

TREATMENT

Extraintestinal *E. coli* Infections

In the past, most *E. coli* isolates were highly susceptible to a broad range of antimicrobial agents. Unfortunately, this situation has changed. Although geographic differences exist, in general, the prevalence of resistance is >20% for ampicillin, amoxicillin-clavulanate, cefazolin, TMP-SMX, and fluoroquinolones, even in community-acquired infections. This resistance precludes empirical use of these agents for serious infections. Travel outside of the United States, prior exposure to an antimicrobial agent, or exposure to a health care setting increases the likelihood of resistance. Fortunately, >90% of isolates that cause uncomplicated cystitis remain susceptible to nitrofurantoin and fosfomycin.

ESBL-producing strains are increasingly prevalent (8–60%), with the highest prevalences in Eastern Europe, Asia, and health care settings. A growing number of reports describe community-acquired UTIs caused by *E. coli* strains that produce CTX-M ESBLs. Data suggest that in some locales acquisition of CTX-M-producing, fluoroquinolone-resistant strains may result from consumption of meat products from food animals treated with third- and fourth-generation cephalosporins and fluoroquinolones. Oral treatment options for such strains are limited; however, in vitro and limited clinical data indicate that fosfomycin, pivmecillinam, and nitrofurantoin often can be used for cystitis. Carbapenems, amikacin, piperacillin-tazobactam, ceftazidime-avibactam, and ceftolozane-tazobactam are the most predictably active agents overall. Carbapenemase-producing strains are also on the rise, but prevalences in most locales are <5%. Tigecycline and the polymyxins, with or without a second agent, have been used most frequently against these XDR isolates; however, the *mcr-1* polymyxin resistance gene has already been described in *E. coli*, and its presence calls into question the future reliability of polymyxins. Ceftazidime-avibactam is active in vitro against strains that produce serine carbapenemases such as KPC but not against those that produce metalloenzymes such as NDM-1; however, relevant clinical data are limited.

This evolving antimicrobial resistance—a source of serious concern—necessitates not only the increasing use of broad-spectrum agents for empirical therapy but also the use of appropriate narrower-spectrum agents for definitive therapy whenever possible as well as the avoidance of treatment of patients who are colonized but not infected.

INTESTINAL PATHOGENIC STRAINS

Pathotypes

Certain strains of *E. coli* are capable of causing diarrheal disease. **(Other important intestinal pathogens are discussed in Chaps. 128, 129, and 160–163.)** At least in the industrialized world, intestinal pathogenic *E. coli* strains are rarely encountered in the fecal flora of healthy persons and instead appear to be essentially obligate pathogens. These strains have evolved a special ability to cause enteritis, enterocolitis, and colitis when ingested in sufficient quantities by a naive host. At least five distinct pathotypes of intestinal pathogenic *E. coli* exist: (1) Shiga toxin–producing *E. coli* (STEC), which includes the subsets enterohemorrhagic *E. coli* (EHEC) and the recently evolved Shiga toxin–producing enteroaggregative *E. coli* (ST-EAEC); (2) enterotoxigenic *E. coli* (ETEC); (3) enteropathogenic *E. coli* (EPEC); (4) enteroinvasive *E. coli* (EIEC); and (5) enteroaggregative *E. coli* (EAEC). Diffusely adherent *E. coli* (DAEC) and cytodetaching *E. coli* are additional putative pathotypes. Lastly, a variant termed adherent invasive *E. coli* (AIEC) has been associated with Crohn's disease (although a causal role remains unproven) but does not cause acute diarrheal disease.

Contaminated food and water are the primary transmission vehicles for ETEC, STEC/EHEC/ST-EAEC, EIEC, and EAEC, whereas person-to-person spread (direct or indirect) is the primary transmission route for EPEC and a secondary transmission route for STEC/EHEC/ST-EAEC. Gastric acidity confers some protection against infection; therefore, persons with decreased stomach acid levels are especially susceptible. Humans are the major reservoir for such strains (except for STEC/EHEC, for which bovines are the main carriers); host range appears to be dictated by species-specific attachment factors. Although there is some overlap, each

pathotype possesses a distinctive combination of virulence traits that results in a pathotype-specific pathogenic mechanism (Table 156-2). With rare exceptions, these strains are largely incapable of causing disease outside the intestinal tract. Whereas disease due to STEC/EHEC/ST-EAEC occurs primarily in high-income countries, disease due to ETEC, EPEC, and EIEC occurs primarily in low- and middle-income countries in Asia, Africa, and Latin America, and disease due to EAEC occurs globally.

SHIGA TOXIN–PRODUCING *E. COLI*

STEC/EHEC/ST-EAEC strains constitute an emerging group of pathogens that can cause hemorrhagic colitis and the hemolytic-uremic syndrome (HUS). In contrast to other intestinal pathotypes, STEC/EHEC/ST-EAEC causes infections more frequently in industrialized countries than in developing regions. Several large outbreaks resulting from the consumption of fresh produce (e.g., lettuce, spinach, sprouts) and of undercooked ground beef have received significant media attention. In addition, a dramatic 2011 outbreak—mainly in Germany—involved an EAEC strain that acquired a Shiga toxin–encoding phage, resulting in a novel genotype, ST-EAEC (O104:H4). This strain was transmitted to the primary cases by sprouted fenugreek seeds, with subsequent human-to-human transmission, and resulted in >4000 cases and 54 deaths.

STEC strains are the fourth most commonly reported cause of bacterial diarrhea in the United States (after *Campylobacter*, *Salmonella*, and *Shigella*). O157:H7 is the most prominent serotype, but many other serogroups have been described, including O6, O26, O45, O55, O91, O103, O111, O113, O121, and O145. Domesticated ruminant animals, particularly cattle and young calves, serve as the major reservoir for STEC/EHEC. Ground or mechanically tenderized beef—the most common food source of STEC/EHEC strains—is often contaminated during processing. Furthermore, manure from cattle or other animals (including in the form of fertilizer) can contaminate produce (potatoes, lettuce, spinach, sprouts, fallen fruits, nuts, strawberries), and fecal runoff from these sources can contaminate water systems. Dairy products and petting zoos are additional sources of infection.

It is estimated that $<10^2$ colony-forming units (CFU) of STEC/EHEC/ST-EAEC can cause disease. Therefore, not only can low levels of food or environmental contamination (e.g., in water swallowed while swimming) result in disease, but person-to-person transmission (e.g., at day-care centers and in institutions) is an important route for secondary spread. Laboratory-associated infections also occur. Illness due to this group of pathogens occurs both as outbreaks and as sporadic cases, with a peak incidence in the summer months.



For STEC/EHEC/ST-EAEC, production of Shiga toxin (Stx2a-g and/or Stx1a,c,d) is a critical factor for occurrence of clinical disease, as demonstrated by the 2011 ST-EAEC outbreak. The *stx* gene is present on chromosomally integrated prophages, and various combinations of *stx* types and subtypes can occur in a given strain. *Shigella dysenteriae* strains that produce the closely related Shiga toxin Stx can also cause hemorrhagic colitis and HUS. Stx2 (especially Stx2a,c,d) appears to be more important than Stx1 in the development of HUS. All Shiga toxins studied to date are multimers comprising one A subunit that is enzymatically active and five identical B subunits that mediate binding to globosyl ceramides, which are membrane-associated glycolipids expressed on certain host cells. As in ricin, the Stx A subunit cleaves an

adenine from the host cell's 28S rRNA, thereby irreversibly inhibiting ribosomal function (i.e., protein synthesis) and potentially leading to apoptosis.

For full pathogenicity, STEC strains require additional properties such as acid tolerance and epithelial cell adherence. Most disease-causing isolates possess the chromosomal locus for enterocyte effacement (LEE). This pathogenicity island was first described in EPEC strains and contains genes that mediate adherence to intestinal epithelial cells and a system that subverts host cells by the translocation of bacterial proteins (type III secretion system). EHEC strains make up the subgroup of STEC strains that possess stx_1 and/or stx_2 as well as LEE. In contrast, the 2011 ST-EAEC outbreak strain lacked LEE, yet was associated with a higher proportion of patients developing HUS (22%) than the historical average for STEC/EHEC outbreaks (2–8%). Data support the essential role of the 2011 outbreak strain's EAEC-associated virulence factors (e.g., AAF/I fimbriae, serine proteases SigA, SepA) in adherence, increased inflammation, and disruption of the intestinal epithelial barrier, which in turn increased the systemic translocation of Stx2a.

After exposure to STEC/EHEC/ST-EAEC and a 3- to 4-day incubation period, colonization of the colon and perhaps the ileum results in symptoms. Colonic edema and an initial non-bloody secretory diarrhea may progress to the hallmark syndrome of grossly bloody diarrhea (identified by history or examination). Significant abdominal pain and fecal leukocytes are common (70% of cases), whereas fever is not; absence of fever can incorrectly lead to consideration of noninfectious conditions (e.g., intussusception and inflammatory or ischemic bowel disease). Occasionally, infections caused by *C. difficile*, *K. oxytoca* (see "Klebsiella Infections," below), *Campylobacter*, and *Salmonella* present in a similar fashion. STEC/EHEC disease is usually self-limited, lasting 5–10 days.

An uncommon but feared complication of infection with STEC/EHEC strains is HUS, which occurs 2–14 days after diarrhea, most often in very young or elderly patients; in contrast, with ST-EAEC strains, HUS occurs more commonly among non-elderly adults, especially young women. It is estimated that in the United States >50% of all HUS cases—and 90% of HUS cases in children—are caused by STEC/EHEC. HUS is mediated by the systemic translocation of Shiga toxins. Erythrocytes may serve as carriers of Stx to endothelial cells located in the small vessels of the kidney and brain. The subsequent development of thrombotic microangiopathy (perhaps with direct toxin-mediated effects on various nonendothelial cells) commonly produces some combination of fever, thrombocytopenia, renal failure, and encephalopathy. Stx-mediated complement activation may also play a role in the development of HUS. Although with dialysis support the mortality rate of HUS is <10%, survivors often have persisting renal and neurologic dysfunction.

ENTEROTOXIGENIC *E. COLI*

ETEC is a major cause of endemic diarrhea in low- and middle-income countries; it is responsible for an estimated 800 million cases annually. After weaning, children in these locales commonly experience several episodes of ETEC infection during the first 3 years of life. The incidence of disease diminishes with age, a pattern that correlates with the development of mucosal immunity to colonization factors (i.e., adhesins). In industrialized countries, infection usually follows travel to endemic areas, although occasional food-borne outbreaks occur.

ETEC is the most common agent of traveler's diarrhea, causing 25–75% of cases. The incidence of infection may be decreased by prudent avoidance of potentially contaminated fluids and foods, particularly items that are poorly cooked, unpeeled, or unrefrigerated **(Chap. 119)**. ETEC infection is uncommon in the United States, but outbreaks secondary to consumption of food products imported from endemic areas have occurred. A large inoculum (10^6–10^8 CFU) is needed to produce disease, which usually develops after an incubation period of 12–72 h.

After adherence of ETEC via colonization factors (e.g., CFA/I, CS), disease is mediated primarily by a heat-labile toxin (LT) and/or a heat-stable toxin (STa). Disease is less severe with strains that produce only LT. Both LT and STa cause net fluid secretion via activation of adenylate cyclase and/or guanylate cyclase C (STa) in the jejunum and ileum. The result is watery diarrhea accompanied by cramps.

LT consists of an A and a B subunit and is structurally and functionally similar to cholera toxin. Strong binding of the B subunit to the GM_1 ganglioside on intestinal epithelial cells leads to the intracellular translocation of the A subunit, which functions as an ADP-ribosyltransferase. Mature STa is an 18- or 19-amino-acid secreted peptide that leads to increased intracellular concentrations of cGMP. Characteristically absent in ETEC-mediated disease are histopathologic changes within the small bowel; mucus, blood, and inflammatory cells in stool; and fever.

The disease spectrum of ETEC infection ranges from mild illness to a life-threatening cholera-like syndrome. Although symptoms are usually self-limited (typically lasting for 3–5 days), infection may result in significant morbidity and mortality (>250,000 deaths annually, mostly from profound volume depletion) when access to health care or suitable rehydration fluids is limited and when small and/or undernourished children are affected.

ENTEROPATHOGENIC *E. COLI*



EPEC causes disease primarily in young children, including neonates. The first *E. coli* pathotype recognized as an agent of diarrheal disease, EPEC was responsible for outbreaks of infantile diarrhea (including in hospital nurseries) in industrialized countries in the 1940s and 1950s. At present, EPEC infection is uncommon in high-income countries, but among infants in low- and middle-income countries is an important cause of diarrhea (both sporadic and epidemic), often accompanied by vomiting and fever. Breast-feeding diminishes the incidence of EPEC infection. Rapid person-to-person spread may occur.

Symptoms develop after colonization of the small bowel and a brief incubation period (1 or 2 days). Initial localized adherence to enterocytes via type IV bundle-forming pili leads to a characteristic effacement of microvilli, with the formation of cuplike, actin-rich pedestals mediated by factors in the LEE. Diarrhea production is a complex and regulated process in which host cell modulation by a type III secretion system plays an important role. Strains lacking bundle-forming pili have been categorized as atypical EPEC (aEPEC); increasing data support a role for these strains as intestinal pathogens in all age groups and among HIV-infected individuals. Diarrheal stool often contains mucus but not blood. Although EPEC diarrhea is usually self-limited (lasting 5–15 days), it may persist for weeks.

ENTEROINVASIVE *E. COLI*

EIEC, a relatively uncommon (or perhaps under-recognized) cause of diarrhea, is rarely identified in the United States, although a few food-related outbreaks have been described. In low- and middle-income countries, sporadic disease is recognized infrequently in children and travelers.

EIEC shares many genetic and clinical features, as well as a common ancestor, with *Shigella*. Both are intracellular pathogens whose virulence is mediated by the presence of specific factors and by the loss or inactivation of other factors (antivirulence genes), which presumably occurred during these organisms' transition from an extracellular to an intracellular lifestyle.

Colonization and invasion of the colonic mucosa, followed by replication therein and cell-to-cell spread (in part via a type III secretion system), result in the development of inflammatory colitis. However, unlike *Shigella*, EIEC produces disease only with a large inoculum (10^8–10^{10} CFU) and is less virulent, typically causing only mild, self-limited (7–10 days), watery diarrhea. Onset generally follows an incubation period of 1–3 days. Occasionally, EIEC can cause a shigellosis-like (dysentery) syndrome characterized by fever, abdominal pain, tenesmus, and scant stool containing mucus, blood, and inflammatory cells.

ENTEROAGGREGATIVE AND DIFFUSELY ADHERENT *E. COLI*

EAEC has been described primarily in low- and middle-income countries and in young children. However, recent studies indicate that it also may be a relatively common cause of diarrhea in all age groups in industrialized countries. EAEC has been recognized increasingly as an important cause of traveler's diarrhea. It is highly adapted to humans—the probable reservoir. A large inoculum is required for infection, which usually manifests as watery and sometimes persistent diarrhea in healthy, malnourished, and HIV-infected hosts.

In vitro, EAEC cells exhibit a diffuse or "stacked-brick" pattern of adherence to small-intestine epithelial cells. Virulence factors that probably are necessary for disease are regulated in large part by the transcriptional activator AggR. The pathogenesis of EAEC disease begins with intestinal adherence, which results from stimulation of epithelial mucus production and bacterial biofilm formation, the latter mediated by fimbriae (AAF/I-III) and possibly the mucinase Pic and dispersin. Inflammation ensues, resulting in epithelial cell exfoliation, as does intestinal secretion mediated by the enterotoxins Pet, EAST-1, ShET1, and HlyE.

Some DAEC strains are capable of causing diarrheal disease, primarily in children 2–6 years of age in some developing countries, and may cause traveler's diarrhea. The Afa/Dr adhesins may contribute to the pathogenesis of such infections.

Diagnosis

Acute infectious diarrhea can be classified as noninflammatory (most commonly viral) or inflammatory (usually bacterial); the latter is suggested by grossly bloody or mucoid stools or a positive test for fecal leukocytes or lactoferrin **(Chap. 128)**. ETEC, EPEC, DAEC, and EAEC cause noninflammatory diarrhea. Identification of these agents requires specialized assays (e.g., PCR-based tests for pathotype-specific genes) that are not routinely available; however, it is rarely necessary to identify the organisms because the

associated diseases are self-limited. ETEC causes the majority and EAEC a minority of cases of noninflammatory traveler's diarrhea; here again, however, definitive diagnosis generally is not necessary for management (as discussed below). If diarrhea persists for >10 days despite treatment, *Giardia* or *Cryptosporidium* (or, in immunocompromised hosts, certain opportunistic pathogens) should be sought. The diagnosis of infection with EIEC, a rare cause of inflammatory diarrhea in the United States, also requires specialized assays.

Because of the considerable public-health importance of STEC/EHEC/ST-EAEC infections, including the threat of HUS, the CDC now recommends that all patients with community-acquired diarrhea, whether inflammatory or not, be evaluated for these pathogens by simultaneous culture (to provide an isolate for strain typing and for outbreak detection and control) and detection of Shiga toxin or its associated genes. The rationale for testing all cases of community-acquired diarrhea, regardless of clinical features, is that bloody stool and fecal white blood cells (or lactoferrin) are not reliably present with STEC/EHEC/ST-EAEC infection. In addition, the use of both tests increases diagnostic sensitivity over that with either test alone.

O157 STEC/EHEC may be identified via culture by screening for *E. coli* strains that do not ferment sorbitol, with subsequent serotyping and testing for Shiga toxin. Selective or screening media are not available for culture-based detection of non-O157 STEC/EHEC/ST-EAEC strains. Detection of Shiga toxins or toxin genes via DNA-based, enzyme-linked immunosorbent, and cytotoxicity assays offers the advantages of rapidity and detection of non-O157 STEC/EHEC/ST-EAEC strains. Specimens positive for toxin but culture-negative for O157 should be forwarded to the local or state public-health laboratory for specialized testing.

TREATMENT

TREATMENT

Intestinal *E. coli* Infections

The mainstay of treatment for all diarrheal syndromes is replacement of water and electrolytes. This measure is especially important for STEC/EHEC/ST-EAEC infection because appropriate volume expansion may decrease renal damage and improve outcome.

The use of prophylactic antibiotics to prevent traveler's diarrhea generally should be discouraged, especially in light of high rates of antimicrobial resistance. However, in selected patients (e.g., those who cannot afford a brief illness or are predisposed to infection), the use of rifaximin, which is nonabsorbable and is well tolerated, is reasonable.

When stools are free of mucus and blood, early patient-initiated treatment of traveler's diarrhea with a fluoroquinolone or azithromycin decreases the duration of illness, and the use of loperamide may halt symptoms within a few hours. Although dysentery caused by EIEC is self-limited, treatment hastens the resolution of symptoms, particularly in severe cases. In contrast, antimicrobial therapy for STEC/EHEC/ST-EAEC infection (the presence of which is suggested by grossly bloody diarrhea without fever) should be avoided because antibiotics may increase the incidence of HUS (possibly via increased production/release of

Stx). In the treatment of HUS, plasmapheresis has no benefit and the value of inhibition of C5 (via eculizumab) is unresolved.

KLEBSIELLA INFECTIONS

K. pneumoniae is the most important *Klebsiella* species from a medical standpoint, causing community-acquired, LTCF-acquired, and nosocomial infections. *K. oxytoca* is primarily a pathogen in LTCFs and hospitals. *Klebsiella* species are broadly prevalent in the environment and colonize the mucosal surfaces of mammals. In healthy humans, the prevalence of *K. pneumoniae* colonization is 5–35% in the colon and 1–5% in the oropharynx; skin is usually colonized only transiently.

Most *Klebsiella* infections in Western countries are caused by "classic" *K. pneumoniae* (cKP) and occur in hospitals and LTCFs. The most common clinical syndromes due to cKP are pneumonia, UTI, abdominal infection, intravascular device infection, surgical site infection, soft tissue infection, and secondary bacteremia. cKP strains have gained notoriety because their propensity for acquiring antimicrobial resistance determinants makes treatment challenging. Clonal group ST258, many members of which produce KPC, is undergoing international dissemination. The spread of NDM-1-producing strains from India in association with medical tourism has captured the attention of physicians and the lay press.

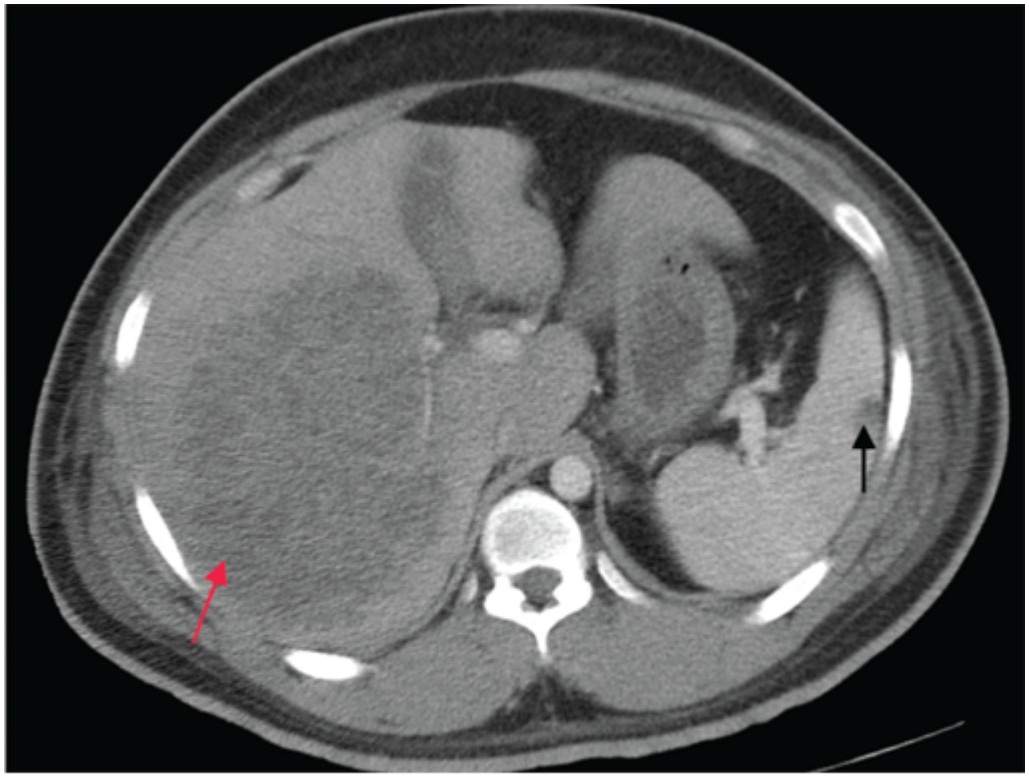
In addition, hypervirulent *K. pneumoniae* (hvKP) strains that are phenotypically and clinically distinct from cKP have emerged recently, having initially been recognized in Taiwan in 1986. Although hvKP infections have occurred globally in all ethnic groups, most cases have been reported in individuals of Asian ethnicity, mainly from the Asian Pacific Rim but also from other locales. Affected individuals often have diabetes mellitus. These demographics raise the possibility of a locale-specific distribution of the organism or an increased susceptibility of Asian hosts, especially those who are diabetic. In contrast to the usual health care–associated context for cKP infections in the West, hvKP is capable of causing serious life- and organ-threatening infections in younger, healthy individuals from the community and can spread metastatically to the eyes, central nervous system, and lungs from the primary site of infection.

hvKP infection initially was characterized and distinguished from traditional infections caused by cKP strains by its (1) presentation as community-acquired pyogenic liver abscess **(Fig. 156-1, *top*)**, (2) occurrence in patients lacking a history of hepatobiliary disease, and (3) propensity for metastatic spread to distant sites (11–80% of cases). More recently, the hvKP pathotype has been recognized as the cause of a variety of serious community-acquired extrahepatic abscesses and infections without liver involvement, including pneumonia, meningitis, endophthalmitis **(Fig. 156-1, *middle*)**, splenic abscess, and necrotizing fasciitis. Survivors often suffer catastrophic morbidity, such as vision loss and major neurologic sequelae.

FIGURE 156-1

Hypervirulent pathotype of *K. pneumoniae* (hvKP). ***Top:*** Abdominal CT scan of a previously healthy 24-year-old Vietnamese man shows a primary liver abscess (*red arrow*) with metastatic spread to the spleen (*black arrow*). *(Courtesy of Drs. Chiu-Bin Hsaio and Diana Pomakova.)* ***Middle:*** A previously healthy 33-year-old Chinese man presented with endophthalmitis. *(From AS Shon et al: Virulence 4:107, 2013.)* ***Bottom:*** A

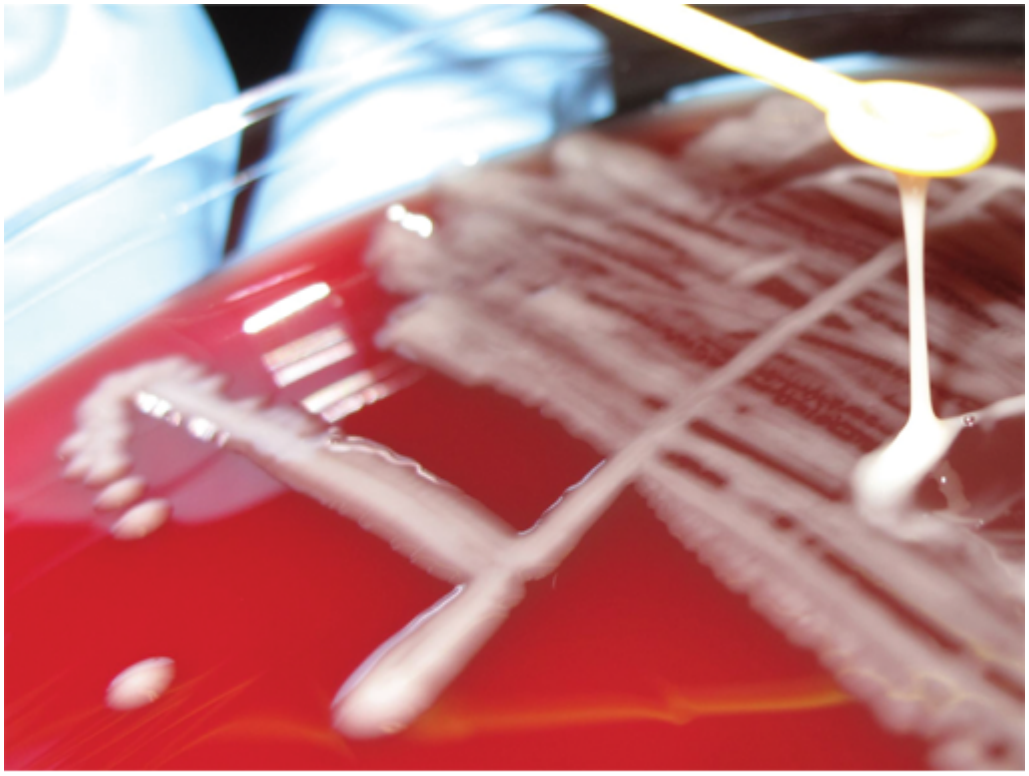
hypermucoviscous phenotype (which does not necessarily equate with a mucoid phenotype) has been associated with hvKP strains. A positive string test is shown. However, this test is not optimally sensitive or specific. A more sensitive and specific marker is needed.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

K. pneumoniae subspecies *rhinoscleromatis* is the causative agent of rhinoscleroma, a granulomatous mucosal upper-respiratory infection that progresses slowly (over months or years) and causes necrosis and occasionally obstruction of the nasal passages. *K. pneumoniae* subspecies *ozaenae* has been implicated as a cause of chronic atrophic rhinitis and rarely of invasive disease in compromised hosts. *K.* (*Calymmatobacterium*) *granulomatis* is sexually transmitted and is the causative agent of granuloma inguinale (donovanosis) that results in chronic genital ulcers (Chap. 168). These *Klebsiella* pathotypes are usually isolated from patients in tropical climates and are genomically distinct from both cKP and hvKP.

INFECTIOUS SYNDROMES

Pneumonia

Although cKP accounts for only a small proportion of cases of community-acquired pneumonia in Western countries (Chap. 121), cKP and *K. oxytoca* are common causes of pneumonia among LTCF residents and hospitalized patients because of increased rates of oropharyngeal colonization in such individuals. Mechanical ventilation is an important risk factor. In Asia and South Africa, community-acquired pneumonia due to hvKP is becoming increasingly common and often occurs in younger patients with no underlying disease. *Klebsiella* is also a common cause of pneumonia in severely malnourished children in developing countries.

As in all pneumonias due to enteric GNB, typical manifestations include production of purulent sputum and evidence of airspace disease. Presentation with earlier, less extensive infection is now more common than is the classically described lobar infiltrate, bulging fissure, and currant jelly sputum. Pulmonary infection due

to hvKP that has spread metastatically (e.g., from a hepatic abscess) usually includes nodular bilateral densities, more commonly in the lower lobes. Pulmonary necrosis, pleural effusion, and empyema can occur with disease progression.

UTI

cKP accounts for only 1–2% of UTI episodes among otherwise healthy adults but for 5–17% of episodes of UTI in patients with anatomical and functional abnormalities of the urinary tract, including indwelling urinary catheter use (complicated UTI). UTI due to hvKP presents more commonly as renal or prostatic abscess due to bacteremic spread than as ascending infection from the urethra and bladder.

Abdominal Infection

cKP causes a spectrum of abdominal infections similar to that caused by *E. coli* but is less frequently isolated from such infections than is *E. coli*. hvKP is a common cause of monomicrobial community-acquired pyogenic liver abscess; in the Asian Pacific Rim, it has been recovered with steadily increasing frequency over the past two decades, replacing *E. coli* as the most common pathogen causing this syndrome. hvKP also is increasingly described as a cause of spontaneous bacterial peritonitis and splenic abscess.

Other Infections

When cKP and *K. oxytoca* cause cellulitis or soft tissue infection, it most frequently involves devitalized tissue (e.g., decubitus and diabetic ulcers, burn wounds) and immunocompromised hosts. cKP and *K. oxytoca* cause some cases of surgical site infection and nosocomial sinusitis as well as occasional cases of osteomyelitis contiguous to soft tissue infection, nontropical myositis, and meningitis (during the neonatal period and after neurosurgery). By contrast, hvKP has become an important cause of community-acquired monomicrobial necrotizing fasciitis, meningitis, endophthalmitis (Fig. 156-1, *middle*), and abscesses within the brain, subdural space, and epidural space, particularly in the Asian Pacific Rim but also globally. Cytotoxin-producing strains of *K. oxytoca* have been implicated as a cause of non–*C. difficile* antibiotic-associated hemorrhagic colitis.

Bacteremia

Klebsiella infection at any site can produce bacteremia. Infections of the urinary tract, respiratory tract, and abdomen (especially hepatic abscess) each account for 15–30% of episodes of *Klebsiella* bacteremia. Intravascular device–related infections account for another 5–15% of episodes, and surgical site and miscellaneous infections account for the rest. *Klebsiella* is an occasional cause of sepsis in neonates and of bacteremia in neutropenic patients. However, like enteric GNB in general, *Klebsiella* rarely causes endocarditis or other endovascular infections.

DIAGNOSIS

Klebsiellae are readily isolated and identified in the laboratory. These organisms usually ferment lactose, although the subspecies *rhinoscleromatis* and *ozaenae* are nonfermenters and are indole-negative. hvKP usually possesses a hypermucoviscous phenotype (**Fig. 156-1, *bottom***), although the sensitivity and specificity of the string test is less than optimal. A better diagnostic test for hvKP is needed.

TREATMENT

TREATMENT

Klebsiella Infections

cKP and *K. oxytoca* have similar antibiotic resistance profiles. These species are intrinsically resistant to ampicillin and ticarcillin and are inconsistently susceptible to nitrofurantoin. The prevalence of resistance to amoxicillin-clavulanate, fluoroquinolones, and TMP-SMX is generally >20%. Increasing resistance is mediated primarily by plasmid-encoded ESBLs (6–70%) and carbapenemases (1–18%), with the highest prevalences in Eastern Europe and Asia and among health care–associated isolates. Furthermore, isolates of cKP that produce CTX-M ESBLs have been obtained from ambulatory patients with no recent health care contact. Oral treatment for infection due to ESBL-producing *Klebsiella* is more challenging than that for infection due to *E. coli* because of the poor activity of nitrofurantoin, the lesser activity—and perhaps lesser efficacy—of fosfomycin, and limited data on pivmecillinam. Empirical treatment of serious cKP and *K. oxytoca* infections with amikacin or a carbapenem may be prudent, depending on local susceptibility patterns and patient-specific risk factors.

Predictably, however, the ESBL-driven use of carbapenems has selected for strains of cKP and *K. oxytoca* that express carbapenemases. The limited treatment options for carbapenem-resistant *Klebsiella* are similar to those described for *E. coli*. Tigecycline, the polymyxins (e.g., colistin), and ceftazidime-avibactam are the most active agents in vitro. However, ceftazidime-avibactam is not active against metallo-carbapenemases (e.g., NDM), and resistance to polymyxins is emerging (e.g., *mcr-1*-mediated colistin resistance). A lethal infection due to a pan-resistant *K. pneumoniae* isolate has already been described in the United States. Combination therapy is often used in this setting, and consultation with relevant experts is advised.

PROTEUS INFECTIONS

Proteus species are part of the colonic flora of a wide variety of mammals, birds, fish, and reptiles. The ability of these GNB to generate histamine from contaminated fish has implicated them in the pathogenesis of scombroid (fish) poisoning **(Chap. 451)**.

Proteus mirabilis causes 90% of *Proteus* infections, which occur in the community, LTCFs, and hospitals. *Proteus vulgaris* and *Proteus penneri* are associated primarily with infections acquired in LTCFs or hospitals. *P. mirabilis* colonizes healthy humans (prevalence, 50%), whereas *P. vulgaris* and *P. penneri* are isolated primarily from individuals with underlying disease. By far the most common site of *Proteus* infection is the urinary tract, where the principal known urovirulence factors of *Proteus* include adhesins, flagella, IgA-IgG

protease, iron acquisition systems, and urease. *Proteus* less commonly causes infection at a variety of other extraintestinal sites.

INFECTIOUS SYNDROMES

UTI

P. mirabilis causes only 1–2% of UTIs in healthy women, and *Proteus* species collectively cause only 5% of hospital-acquired UTIs. However, *Proteus* is responsible for 10–15% of cases of complicated UTI, primarily those associated with catheterization; indeed, *Proteus* accounts for 20–45% of urine isolates from chronically catheterized patients. This high prevalence is due in part to bacterial production of urease, which hydrolyzes urea to ammonia and results in alkalization of the urine. Alkalization of urine, in turn, leads to precipitation of organic and inorganic compounds, which contributes to formation of struvite and carbonate–apatite crystals, formation of biofilms on catheters, and/or development of frank calculi. *Proteus* becomes associated with the stones and biofilms; thereafter, it usually cannot be eradicated without removal of the stones or catheter. Over time, staghorn calculi may form within the renal pelvis and lead to obstruction and renal failure. Although biologically plausible, clinical support is lacking for the concept that urine samples exhibiting unexplained alkalinity should be cultured, and isolation of a *Proteus* species (or other urea-splitting organism) should prompt consideration of an evaluation for urolithiasis.

Other Infections

Proteus occasionally causes pneumonia (primarily in LTCF residents or hospitalized patients), nosocomial sinusitis, intraabdominal abscesses, biliary tract infection, surgical site infection, soft tissue infection (especially decubitus and diabetic ulcers), and osteomyelitis (primarily contiguous); in rare cases, it causes nontropical myositis. In addition, *Proteus* uncommonly causes neonatal meningitis, with the umbilicus frequently implicated as the source; this disease is often complicated by development of a cerebral abscess. Otogenic brain abscess also occurs.

Bacteremia

Most episodes of *Proteus* bacteremia originate from the urinary tract; however, intravascular devices and any of the less common sites of *Proteus* infection are also potential sources. Endovascular infection is rare, but when endocarditis occurs it can be persistent and destructive. *Proteus* species are occasional agents of sepsis in neonates and of bacteremia in neutropenic patients.

DIAGNOSIS

Proteus is readily isolated and identified in the laboratory. Most strains are lactose-negative, produce H_2S, and demonstrate characteristic swarming motility on agar plates. *P. mirabilis* and *P. penneri* are indole-negative, whereas *P. vulgaris* is indole-positive. The inability to produce ornithine decarboxylase differentiates *P. penneri* from *P. mirabilis*.

TREATMENT

TREATMENT

Proteus Infections

The intrinsic resistance of *P. mirabilis* to tetracyclines, cefazolin, nitrofurantoin, polymyxins, and tigecycline renders treatment of XDR isolates problematic. Acquired resistance to ampicillin (prevalence range, 15–60%), fluoroquinolones (11–55%), and TMP-SMX (20–50%) is common. Ampicillin-sulbactam tends to be more active, with resistance prevalences of 6–18%. In the United States and Canada, the prevalence of ESBL production by *P. mirabilis* remains low (generally <5%). However, rates as high as 60% have been reported from Asia. Isolates of *P. mirabilis* that produce CTX-M ESBLs have been recovered from ambulatory patients with no recent health-care contact (see the section on the treatment of extraintestinal *E. coli* infections for treatment considerations). *P. vulgaris* and *P. penneri* exhibit more extensive drug resistance than does *P. mirabilis*, and induction or selection of *P. vulgaris* variants with stable derepression of chromosomal AmpC β-lactamase may occur. For critically ill patients, carbapenems, fourth-generation cephalosporins (e.g., cefepime), ceftazidime-avibactam, ceftolozane-tazobactam, and amikacin generally display excellent activity against *Proteus* species (90–100% of isolates susceptible).

ENTEROBACTER AND *CRONOBACTER* INFECTIONS

E. cloacae and *E. aerogenes* are responsible for most *Enterobacter* infections (65–75% and 15–25%, respectively); *Cronobacter sakazakii*, *Cronobacter malonaticus* (formerly *Enterobacter sakazakii*), and *Enterobacter gergoviae* are less commonly isolated (1% for each). *Enterobacter* species cause primarily health care–related infections. The organisms are widely prevalent in foods, environmental sources (including equipment at health care facilities), and a variety of animals.

These organisms colonize few healthy humans, but the percentage colonized increases significantly with LTCF residence or hospitalization. Although colonization is an important prelude to infection, direct introduction via IV lines (e.g., contaminated IV fluids or pressure monitors) also occurs. Extensive antibiotic resistance has developed in *Enterobacter* species and probably has contributed to the emergence of the organisms as prominent nosocomial pathogens. Individuals who have previously received antibiotic treatment, have comorbid disease, and are ICU residents are at greatest risk for infection. *Enterobacter* causes a spectrum of extraintestinal infections similar to that described for other GNB.

INFECTIOUS SYNDROMES

Pneumonia, UTI (particularly catheter-related), intravascular device–related infection, surgical site infection, and abdominal infection (primarily postoperative or related to devices such as biliary stents) are the most common syndromes encountered. Nosocomial sinusitis, meningitis related to neurosurgical procedures (including use of intracranial pressure monitors), osteomyelitis, and endophthalmitis after eye surgery are less frequent. Neonates (particularly those of low birth weight) are at risk for *C. sakazakii* infection, including

neonatal bacteremia, necrotizing enterocolitis, and meningitis (often complicated by brain abscess or ventriculitis). Contaminated powdered infant formula has been implicated as a source for such neonatal infections. The WHO recommends that, to reduce the initial number of bacteria, powdered infant formula should be reconstituted with hot water (>70°C) and, to limit replication of residual bacteria, the reconstituted formula should be stored at <5°C or its storage time minimized.

Enterobacter bacteremia can result from primary infection at any anatomic site. In bacteremia of unclear origin, the contamination of IV fluids or medications, blood components or plasma derivatives, catheter-flushing fluids, pressure monitors, and dialysis equipment should be considered, particularly in an outbreak setting. *Enterobacter* can also cause bacteremia in neutropenic patients. *Enterobacter* endocarditis is rare, occurring primarily in association with illicit IV drug use or prosthetic valves.

DIAGNOSIS

Enterobacter is readily isolated and identified in the laboratory. Most strains are lactose-positive and indole-negative.

TREATMENT

TREATMENT

Enterobacter Infections

Significant antimicrobial resistance exists among *Enterobacter* strains. Ampicillin, ampicillin-sulbactam, and first- and second-generation cephalosporins have little or no activity. Extensive use of third-generation cephalosporins can induce or select for variants with stable derepression of AmpC β-lactamase, which confers resistance to these agents, to monobactams (e.g., aztreonam), and—in many cases—to β-lactam/β-lactamase inhibitor combinations. Resistance may emerge during therapy; in one study, this phenomenon was documented in 20% of clinical isolates. De novo resistance should be considered when clinical deterioration follows initial improvement, and third-generation cephalosporins should be avoided in the treatment of serious *Enterobacter* infections.

Cefepime is stable in the presence of AmpC β-lactamases; thus, it is a suitable option for treatment of *Enterobacter* infections so long as no coexistent ESBL is present. Detection of ESBLs in *Enterobacter* is difficult because of the presence of AmpC β-lactamase; nonetheless, their prevalence (particularly in *E. cloacae*) is known to be variable worldwide but is generally increasing and is now 5–50% overall. This increase is evidenced by 2014 data from the National Healthcare Safety Network, which documented resistance to third- and fourth-generation cephalosporins in 36.1% of *Enterobacter* isolates from central line–associated bloodstream infections in the United States. The prevalence of resistance has ranged from 15 to 40% for piperacillin-tazobactam and from 5 to 15% for colistin; it is more variable but generally higher for the fluoroquinolones. Fortunately, carbapenems, ceftazidime-avibactam, ceftolozane-tazobactam, amikacin, and tigecycline have generally retained excellent activity (90–99% of isolates susceptible). Once susceptibility

data for a patient's isolate become available, it is advisable to de-escalate the antimicrobial regimen whenever possible.

SERRATIA INFECTIONS



S. marcescens causes the great majority (>90%) of *Serratia* infections, with other species isolated only occasionally. Serratiae are found primarily in the environment (including in health care institutions), particularly in moist settings. Serratiae have been isolated from a variety of animals, insects, and plants but only infrequently from healthy humans. In LTCFs or hospitals, reservoirs for the organisms include the hands and fingernails of health care personnel, food, milk (on neonatal units), sinks, medical equipment or devices, IV solutions or parenteral medications (particularly those generated by compounding pharmacies), prefilled syringes and multiple-access medication vials (e.g., heparin, saline), blood products (e.g., platelets), hand soaps and lotions, irrigation solutions, and even disinfectants.

Infection results from either direct inoculation (e.g., via contaminated IV fluid or injected medications or recreational drugs) or colonization (primarily of the respiratory tract). Sporadic infection is most common, but outbreaks (often involving MDR strains in adult and neonatal ICUs) also occur. Hygiene, medication-compounding standards, sterile technique, and infection control programs are critical measures to prevent infection.

The spectrum of extraintestinal infections caused by *Serratia* is similar to that for other GNB. *Serratia* species are usually considered causative agents of health care–associated infection and account for 1–3% of hospital-acquired infections. However, population-based laboratory surveillance studies in Canada and Australia have demonstrated that community-acquired *Serratia* infections occur more commonly than was previously appreciated. *Serratia* also is one of the pathogens associated with chronic granulomatous disease.

INFECTIOUS SYNDROMES

The respiratory tract, the genitourinary tract, intravascular devices, the eye (contact lens–associated keratitis and other ocular infections), surgical wounds, and the bloodstream (from contaminated infusions) are the most common sites of *Serratia* infection; the former five sites are the most common sources of *Serratia* bacteremia. Soft tissue infections (including myositis, fasciitis, mastitis), osteomyelitis, abdominal and biliary tract infections (postprocedural), and septic arthritis (primarily from intraarticular injections) occur less commonly. Serratiae are uncommon causes of neonatal or postsurgical meningitis and of bacteremia in neutropenic patients. Endocarditis is rare.

DIAGNOSIS

Serratiae are readily cultured and identified by the laboratory and are usually lactose- and indole-negative. The red pigmentation of some *S. marcescens* strains and *Serratia rubidaea* can produce distinctive clinical findings (e.g., pink breast milk or hypopyon; pseudohemoptysis).

TREATMENT

TREATMENT

Serratia Infections


Most *Serratia* strains (>80%) are resistant to ampicillin, amoxicillin-clavulanate, ampicillin-sulbactam, first-generation cephalosporins, cephamycins, nitrofurantoin, and colistin. Induction or selection of variants with stable derepression of chromosomal AmpC β-lactamases may develop during therapy. Both in the United States and globally, the prevalence of ESBL-producing isolates is generally low (<5%), but rates of 20–30% have been reported in Asia and Latin America. Acquisition of carbapenemase-encoding genes is uncommon but increasing. For critically ill patients, carbapenems, amikacin, ceftazidime-avibactam, and ceftolozane-tazobactam are most reliably active (>90% susceptible).

CITROBACTER INFECTIONS

C. freundii and *Citrobacter koseri* cause most human *Citrobacter* infections, which are epidemiologically and clinically similar to *Enterobacter* infections. *Citrobacter* species are commonly present in water, food, soil, and certain animals. *Citrobacter* is part of the normal fecal microbiota in a minority of healthy humans, but colonization rates are higher in LTCFs and hospitals—the settings in which nearly all *Citrobacter* infections occur. *Citrobacter* species account for 1–2% of nosocomial infections. The affected hosts are usually immunocompromised and/or have comorbid disease or disruption of skin or mucosal barriers. *Citrobacter* causes extraintestinal infections similar to those described for other GNB.

INFECTIOUS SYNDROMES

The urinary tract accounts for 40–50% of *Citrobacter* infections. Less commonly involved sites include the biliary tree (particularly with stones or obstruction), the respiratory tract, surgical sites, soft tissue (e.g., decubitus ulcers), the peritoneum, and intravascular devices. Osteomyelitis (usually from a contiguous focus), adult central nervous system infection (from neurosurgical or other types of meningeal disruption), and myositis occur rarely. *Citrobacter* (primarily *C. koseri*) also causes 1–2% of neonatal meningitis cases, of which 50–80% are complicated by brain abscess. Further, case reports in adults suggest that *C. koseri* infection has a predilection for abscess formation. Bacteremia is most often due to UTI, biliary/abdominal infection, or intravascular device infection. *Citrobacter* occasionally causes bacteremia in neutropenic patients. Endocarditis and endovascular infections are rare.

DIAGNOSIS

Citrobacter species are readily isolated and identified; 35–50% of isolates are lactose-positive, and 100% are oxidase-negative. *C. freundii* is indole-negative, whereas *C. koseri* is indole-positive.

TREATMENT

TREATMENT

Citrobacter Infections

C. freundii is more extensively antibiotic-resistant than is *C. koseri*. More than 90% of isolates are resistant to ampicillin and to first- and second-generation cephalosporins, and >50% of strains are resistant to ampicillin-sulbactam. *Citrobacter* species (except *C. koseri*) possess AmpC β-lactamases; induction or selection of variants with stable derepression may occur during therapy. The prevalence of resistance generally ranges from 15 to 35% for third-generation cephalosporins, from 5 to 15% for piperacillin-tazobactam and fluoroquinolones, and from 5 to 20% for TMP-SMX. The prevalence of ESBL production is <5%. Carbapenems, amikacin, cefepime, tigecycline, ceftazidime-avibactam, ceftolozane-tazobactam, fosfomycin, and colistin are most active (>90% of isolates susceptible).

MORGANELLA AND *PROVIDENCIA* INFECTIONS

M. morganii, Providencia stuartii, and (less frequently) *Providencia rettgeri* are the members of their respective genera that cause human infections. The epidemiologic associations, pathogenic properties, and clinical manifestations of these organisms resemble those of *Proteus* species. *Morganella* and *Providencia* occur more commonly among LTCF residents than among hospitalized patients, largely resulting from chronic urinary-catheter use. In settings with extensive use of polymyxins and tigecycline, these organisms may become increasingly common because of their intrinsic resistance to these agents.

INFECTIOUS SYNDROMES

These species are primarily urinary tract pathogens, causing UTIs that are most often associated with long-term (>30-day) catheterization. Such infections commonly lead to biofilm formation and catheter encrustation (sometimes causing catheter obstruction) or the development of struvite bladder or renal stones (sometimes causing renal obstruction and serving as foci for relapse). *Morganella* is also commonly isolated from snakebite infection.

Other, less common infectious syndromes include surgical site infection, soft tissue infection (primarily involving decubitus and diabetic ulcers), burn site infection, pneumonia (particularly ventilator-associated), intravascular device infection, and intraabdominal infection. Rarely, the other extraintestinal infections described for GNB also occur. Bacteremia is uncommon; when it does occur, any infected site can serve as the source, but the urinary tract accounts for most cases, with the next most common sources being surgical site, soft tissue, and hepatobiliary infections.

DIAGNOSIS

M. morganii and *Providencia* are readily isolated and identified. Nearly all isolates are lactose-negative and indole-positive.

TREATMENT

TREATMENT

Morganella and *Providencia* Infections

Morganella and *Providencia* may be extensively resistant to antibiotics. Most (or all) isolates are resistant to ampicillin, ampicillin-sulbactam, first-generation cephalosporins, nitrofurantoin, fosfomycin, tigecycline, and the polymyxins; treatment of XDR strains is especially challenging. The β-lactamase inhibitor tazobactam increases susceptibility to β-lactam agents, but sulbactam and clavulanic acid do not. *Morganella* and *Providencia* possess inducible AmpC β-lactamases; clinically significant induction or selection of stably derepressed mutants may develop during therapy. The prevalence of resistance generally ranges from 10 to 30% for the third-generation cephalosporins, from 10 to 40% for fluoroquinolones, and from 20 to 40% for TMP-SMX; the prevalence is more widely variable for piperacillin-tazobactam. The prevalence of ESBL-producing isolates is <5%. Carbapenems, amikacin, cefepime, ceftazidime-avibactam, and ceftolozane-tazobactam are the most active agents (>90% of isolates susceptible). Removal of a colonized urinary catheter or stone is critical for eradication of UTI.

EDWARDSIELLA INFECTIONS

E. tarda is the only member of the genus *Edwardsiella* that is associated with human disease. This organism is found predominantly in freshwater and marine environments and in the associated aquatic animal species. Human acquisition occurs primarily from interaction with these reservoirs or ingestion of inadequately cooked aquatic animals. *E. tarda* infection is rare in the United States, where acquisition occurs mainly along the Gulf of Mexico; recently reported cases are mostly from Asia. This pathogen shares clinical features with *Salmonella* species (as an intestinal pathogen; **Chap. 160**), *Vibrio vulnificus* (as an extraintestinal pathogen; **Chap. 163**), and *Aeromonas hydrophila* (as both an intestinal and an extraintestinal pathogen; **Chap. 153**).

INFECTIOUS SYNDROMES

Gastroenteritis is the predominant infectious syndrome (50–80% of infections). Self-limiting watery diarrhea is most common, but severe colitis also occurs. The most common extraintestinal infection is wound infection due to direct inoculation, which is often associated with freshwater, marine, or snakebite injuries. Other infectious syndromes result from invasion of the gastrointestinal tract and subsequent bacteremia. Most afflicted hosts have comorbidities (e.g., hepatobiliary disease, iron overload, cancer, or diabetes mellitus). A primary bacteremic syndrome, sometimes complicated by meningitis, has a 40% case–fatality rate. Visceral (primarily hepatic) and intraperitoneal abscesses also occur, as do endocarditis and empyema.

DIAGNOSIS

Although *E. tarda* can readily be isolated and identified, most laboratories do not routinely seek to identify it in stool samples. Production of hydrogen sulfide is a characteristic biochemical property.

TREATMENT

TREATMENT

Edwardsiella Infections

E. tarda is susceptible to most antimicrobial agents appropriate for use against GNB. Gastroenteritis is generally self-limiting, but treatment with a fluoroquinolone may hasten resolution. In the setting of severe sepsis, fluoroquinolones, third- and fourth-generation cephalosporins, carbapenems, and amikacin—either alone or in combination—are the safest choices pending susceptibility data.

INFECTIONS CAUSED BY MISCELLANEOUS GENERA

Species of *Hafnia*, *Kluyvera*, *Cedecea*, *Pantoea*, *Ewingella*, *Leclercia*, *Raoultella*, and *Photorhabdus* are occasionally isolated from diverse clinical specimens, including blood, sputum, urine, cerebrospinal fluid, joint fluid, bile, and wounds. These organisms are rare and usually cause infection in compromised hosts or in association with an invasive procedure or foreign body. Cephalosporinases from *Kluyvera* have been implicated as the progenitors of CTX-M ESBLs. *Kluyvera* and *Raoultella* may produce carbapenemases.

FURTHER READING

Boisen N et al: Shiga toxin 2a and enteroaggregative *Escherichia coli*—A deadly combination. Gut Microbes 6:272, 2015.

Chen L et al: Notes from the field: Pan-resistant New Delhi metallo-beta-lactamase–producing *Klebsiella pneumoniae*—Washoe County, Nevada, 2016. MMWR Morb Mortal Wkly Rep 66:33, 2017.

Croxen MA et al: Recent advances in understanding enteric pathogenic *Escherichia coli*. Clin Microbiol Rev 26:822, 2013.

Dale AP, Woodford N: Extra-intestinal pathogenic *Escherichia coli* (ExPEC): Disease, carriage and clones. J Infect 71:615, 2015.

Holy O, Forsythe S: *Cronobacter spp*. as emerging causes of healthcare-associated infection. J Hosp Infect 86:169, 2014.

Khatri A et al: Community-acquired pyelonephritis in pregnancy caused by KPC-producing *Klebsiella pneumoniae*. Antimicrob Agents Chemother 59:4375, 2015.

Lim C et al: Epidemiology and burden of multidrug-resistant bacterial infection in a developing country. Elife Sep 6:5, 2016. PMCID: PMC5030096.

Shields RK et al: Clinical outcomes, drug toxicity, and emergence of ceftazidime-avibactam resistance among patients treated for carbapenem-resistant Enterobacteriaceae infections. Clin Infect Dis 63:1615, 2016.

Shon AS et al: Hypervirulent (hypermucoviscous) *Klebsiella pneumoniae*: A new and dangerous breed. Virulence 4:107, 2013.

Tamma PD, Rodriguez-Bano J: The use of noncarbapenem β-lactams for the treatment of extended-spectrum β-lactamase infections. Clin Infect Dis 64:942, 2017.

Chapter 157: *Acinetobacter* Infections

Rossana Rosa; L. Silvia Munoz-Price

FIGURE 157-1

DEFINITION

Acinetobacter species were first described in 1911 and named *Micrococcus calcoaceticus*. Thereafter, the genus was renamed multiple times; since 1950, it has been known as *Acinetobacter*. *Acinetobacter* species are gram-negative, oxidase-negative, nonmotile, nonfermenting coccobacilli that are easily recovered on standard culture media. Differentiation among *Acinetobacter* species on the basis of phenotypic characteristics alone is very difficult. Molecular-based methods such as matrix-assisted laser desorption–ionization–time-of-flight mass spectrometry (MALDI-TOF-MS) and quantitative real-time polymerase chain reaction (PCR) are usually necessary to identify *Acinetobacter baumannii*, the most clinically relevant species of the genus.

ETIOLOGY AND EPIDEMIOLOGY

Acinetobacter species are naturally encountered in water and soil and have also been recovered from fruits and vegetables. In humans, *Acinetobacter* can be found on the skin and in the respiratory and gastrointestinal tracts. *A. baumannii* is capable of surviving environmental desiccation for weeks; this characteristic is important from an infection-control perspective as it allows this organism to persist in the hospital environment and on equipment.

Acinetobacter was historically considered a pathogen of hot and humid climates. In recent years, however, hospital outbreaks caused by *A. baumannii* have been reported worldwide, even in temperate climates. In the United States, the Centers for Disease Control and Prevention (CDC) estimates that 12,000 *Acinetobacter* infections occur every year, 7300 of which are caused by multidrug-resistant strains, with 500 attributable deaths. The increase in the number of infections with *A. baumannii* is suspected to be due to the rapid spread of certain genetically distinct lineages; of the three international clonal lineages (ICLs), ICL I and ICL II are multidrug resistant. The predominance of these lineages remains unexplained, although it has been proposed that this population structure is the result of two waves of expansion. The first wave followed a bottleneck (possibly linked to a restricted ecologic niche) that occurred in the distant past. The second wave is ongoing and is being driven by the rapid expansion of a limited number of multidrug-resistant clones.

 Analysis of the *A. baumannii* pangenome (the sum of the core and dispensable genomes) has shown that its organization is characterized by a small core genome and a large accessory or disposable genome. This organization reflects *A. baumannii's* high plasticity, which enables it to acquire exogenous genetic material. With few exceptions, gene functions associated with virulence are found in the core genome; this observation suggests a limited role for the acquisition of new virulence traits in the recent nosocomial expansion of *A. baumannii* clones. Genes associated with resistance to antimicrobial agents are found in both the species core genome and the accessory genome. In the accessory genome, these genes have been found in alien islands, often flanked by integrases, transposases, or insertion sequences. This pattern suggests possible acquisition by horizontal gene transfer from other *Acinetobacter* strains or even from different bacterial species present in the immediate environment. Acquisition of these antimicrobial resistance genes is hypothesized to have led to the recent rapid expansion of highly homogeneous clonal lineages, whose main difference from nonclonal *A. baumannii* appears to be their antimicrobial resistance.

Health Care–Associated Infections

Infections caused by *A. baumannii* occur frequently among patients admitted to intensive care units (ICUs). Risk factors for colonization and infection with this pathogen include nursing home residence, prolonged ICU stay, central venous catheterization, tracheostomy, mechanical ventilation, enteral feedings, and treatment with third-generation cephalosporins, fluoroquinolones, and carbapenems. Acquisition of carbapenem-resistant *A. baumannii* is most common among patients exposed to carbapenems. Spread of *A. baumannii* across different regions is facilitated by the movement of patients between health care systems and throughout the continuum of health care. Within the hospital, environmental spread of *A. baumannii* occurs as a result of inappropriate hand hygiene among workers providing health care for infected or colonized patients and the contamination of hospital equipment, such as respiratory therapy and ventilation equipment. The air surrounding the patient may also play a role in environmental colonization with *A. baumannii*, especially in inpatient areas without physical barriers between patients and with an inadequate number of air exchanges.

A. baumannii strains identified during hospital outbreaks are typically resistant to more antibiotic classes than strains from the community. The prevalence of colonization with *A. baumannii* at the time of admission or during a stay in a long-term acute-care hospital (LTACH) or nursing home is variable and depends on regional flora. Outbreaks of *A. baumannii* in acute-care hospitals and LTACHs that "share" patients have been described in Ohio, Michigan, Illinois, and Indiana.

Community-Acquired Infections

Community-acquired infections caused by *Acinetobacter* have been described in Australia and Asia. Few cases have been reported in regions with a temperate climate, and even those few cases have taken place during warm and humid months. Risk factors for community-acquired pneumonia due to this organism include a history of alcohol abuse, diabetes mellitus, smoking, and chronic lung disease.

War Zone–Associated Infections

Infections caused by *Acinetobacter* in war zones include skin and soft tissue infections associated with traumatic injuries and bloodstream infections. Outbreak investigations of *A. baumannii* infections among military personnel returning from Iraq and Afghanistan suggested the acquisition of *A. baumannii* in field hospitals rather than colonization of the skin before an injury. This view is supported by the recovery of *A. baumannii* isolates with similar genetic characteristics from inanimate surfaces in field hospitals and from patients.

Disaster Medicine

A. baumannii is linked to infections among victims of trauma during tsunamis, earthquakes, and terrorist attacks. The types of infections most frequently observed in these settings are soft tissue injuries, but bloodstream infections and pneumonia have also been reported. In addition, outbreaks of *A. baumannii* infection in ICUs caring for disaster victims have been described.

PATHOGENESIS

Mechanisms of pathogenesis and virulence in *Acinetobacter* species have not been fully elucidated. However, *A. baumannii* seems to have greater virulence potential than other *Acinetobacter* species, as evidenced by its ability to grow at 37°C and to resist uptake by macrophages.

Initial *A. baumannii* colonization of the host and the environment is facilitated by the organism's ability to adhere to surfaces and human cells and to create biofilms. The ability to form a biofilm is phenotypically associated with exopolysaccharide production and pilus formation. A quorum-sensing molecule encoded by the *abaI* autoinducer synthase gene has been implicated in *A. baumannii* biofilm formation on abiotic surfaces. Outer-membrane porins appear to mediate cell apoptosis. *A. baumannii* can survive in harsh environments within the host and on inanimate surfaces by modifying the structure of its lipid A, with a consequent decrease in susceptibility to antibiotics and antimicrobial peptides and an increase in survival upon desiccation.

Acinetobacter species produce an extracellular capsule that protects the bacteria from external threats, including complement-mediated killing. Studies of mouse models showed that *Acinetobacter* species can increase capsule production in the presence of subinhibitory levels of antibiotic—an ability that leads to increased resistance to complement-mediated killing and a hypervirulent phenotype.

Phospholipase C and phospholipase D have been identified as virulence factors in *A. baumannii.* These enzymes exert cytotoxic effects on epithelial cells and facilitate their invasion.

Iron-acquisition systems are also important virulence mechanisms in *A. baumannii*. Through secretion of siderophores (low-molecular-mass ferric-binding compounds), *A. baumannii* is able to grow despite iron deficiencies in the surrounding environment (e.g., in the human host).

Several protein-secretion systems have been identified in *A. baumannii*. The most recently described is a type II secretion system. The substrate for this system, the LipA lipase, is required for growth on medium containing lipids as a sole carbon source. Mutants lacking the genes for the type II secretion system or its substrate exhibit defective in vivo growth in a neutropenic murine model of bacteremia. *A. baumannii* also has a type VI secretion system whose primary function seems to be to secrete antibacterial toxins that kill competing bacteria, including other strains in the same species.

The type V autotransporter system has been characterized in *A. baumannii.* In a murine systemic model of *Acinetobacter* infection, the *Acinetobacter* trimeric autotransporter mediates biofilm formation and maintenance; adherence to extracellular matrix components such as collagen I, II, and IV; and virulence.

Outer-membrane vesicles (OMVs) play a special role in protein secretion. Many *A. baumannii* strains secrete OMVs containing various virulence factors, including outer-membrane protein A (OmpA), proteases, and phospholipases. The membrane proteins in OMVs are responsible for eliciting a potent innate immune response. Several studies have shown that *A. baumannii* OMVs could be used as an acellular vaccine to effectively control *A. baumannii* infections.

Nosocomial strains of *Acinetobacter* can deploy multiple mechanisms of resistance, including alterations in porins and efflux pumps and expression of β-lactamases. More specifically, *Acinetobacter* species can reduce the expression of porins, thus hindering the passage of β-lactam antibiotics into the periplasmic space. These species can overexpress bacterial efflux pumps and decrease the concentration of β-lactam antibiotics in the periplasmic space. Efflux pumps can also actively remove quinolones, tetracyclines, chloramphenicol, disinfectants, and tigecycline. *Acinetobacter* species possess chromosomally encoded cephalosporinases and are capable of acquiring β-lactamases, including serine and metallo-β-lactamases. AmpC β-lactamases are class C β-lactamases intrinsic to all *A. baumannii* strains. Although these enzymes are expressed at low levels and are not inducible, the addition of the insertion sequence *ISAba1* next to the AmpC gene increases β-lactamase production, with resulting resistance to cephalosporins.

Carbapenem resistance in *Acinetobacter* species is mostly tied to the emergence of Ambler class D oxacillinases of group 2d, some of which are intrinsic and chromosomal (e.g., OXA-51-like) while others are acquired and are found in plasmids or are chromosomally encoded (e.g., OXA-23-like, 24 [33-like, 40-like], 58-like, 143-like, and 235-like).

CLINICAL MANIFESTATIONS

Pneumonia

A. baumannii is a notorious cause of nosocomial pneumonia, most frequently among patients requiring prolonged mechanical ventilation. The onset of disease tends to be later than that caused by other gram-negative bacilli; however, clinical symptoms of hospital-acquired or ventilator-associated pneumonia due to *A. baumannii* are similar to those of nosocomial or ventilator-associated pneumonia due to other nosocomial pathogens. Thus, the most common indicators of infection include fever and increased sputum production.

The positivity of respiratory cultures in most cases may present a challenge for the clinician, since airway colonization with *A. baumannii* is a risk factor for infection itself. Radiologic findings are nonspecific and can include lobar consolidations and pleural effusions, but cavitations are rarely seen. The crude mortality rates associated with nosocomial pneumonia due to *A. baumannii* are reported to be as high as 65%. However, since these infections occur in debilitated patients, their attributable mortality has been difficult to establish.

Community-acquired pneumonia due to *A. baumannii* is a relatively rare entity. Its clinical presentation is characterized by fever, severe respiratory symptoms, and multiple-organ dysfunction. Patients frequently have a cough productive of purulent sputum, shortness of breath, and chest pain. Imaging studies usually show lobar consolidation. Mortality rates associated with this process are >50%.

Bloodstream Infections

Bloodstream infections due to *A. baumannii* are most frequent among ICU patients and usually occur in the presence of a central venous catheter or as a secondary complication of hospital-acquired or ventilator-associated pneumonia. Polymicrobial growth has been reported in 20–36% of bacteremia episodes. Fever is the most common sign of infection (developing in >95% of cases), and presentation with septic shock and disseminated intravascular coagulopathy has been described in as many as 25 and 30% of patients, respectively. *A. baumannii* bloodstream infections often result in higher hospitalization costs and longer ICU stays. Crude mortality rates from this infection are as high as 40%; however, rates can be as high as 70% from infections caused by carbapenem-resistant isolates. In patients with infections caused by extremely drug-resistant strains, poor outcomes are thought to be driven by delays in the initiation of adequate antimicrobial therapy.

Skin and Soft Tissue Infections

Acinetobacter species have been described as part of the skin flora, yet the majority of the organisms from this genus that colonize the skin are not those associated with nosocomial infections. Discerning infection from wound colonization is challenging. Gunshot wounds and the presence of orthopedic external-fixation devices are common among patients with combat trauma–associated *A. baumannii* skin and soft tissue infections. The report on a case series of eight U.S. military patients described the clinical presentation of their infections as evolving from an edematous *peau d'orange* appearance to a sandpaper appearance with overlying vesicles and then to a necrotizing process with hemorrhagic bullae. Other case series have also included necrotizing fasciitis. *A. baumannii* is an important pathogen in burn units worldwide. Large burns provide ideal conditions for *A. baumannii* and facilitate patient-to-patient transmission. The presence of *A. baumannii* in wounds contributes to healing delays and graft loss. In addition, wound colonization is a risk factor for bloodstream infections among patients with extensive burn injuries.

A. baumannii infections resulting from trauma to soft tissues in the setting of natural disasters, such as tsunamis and earthquakes, have been reported. The implication is that *A. baumannii* should be considered in the differential diagnosis of soft tissue infections following exposure to tropical and subtropical environments.

Urinary Tract Infections

A. baumannii is an infrequent cause of urinary tract infections. The majority of cases reported are catheter-associated infections, reflecting the ability of *A. baumannii* to form biofilms on these devices. A few reports have described community-acquired infections occurring in the setting of nephrolithiasis and after renal transplantation.

Meningitis

Central nervous system infections with *A. baumannii* have been reported in the context of outbreaks, traumatic injuries, neurosurgical procedures, and external ventricular drains. One case series described a petechial rash in up to 30% of patients. *Acinetobacter* species may look similar to *Neisseria meningitidis* on a Gram's stain of cerebrospinal fluid; both appear as gram-negative paired cocci.

Other Miscellaneous Infections

A few cases of *A. baumannii* keratitis associated with the use of contact lenses have been reported. Cases of native- and prosthetic-valve endocarditis have also been described.

TREATMENT

TREATMENT

Acinetobacter Infections

Treatment of *Acinetobacter* infections is challenging because *Acinetobacter* can develop resistance to most available antibiotics. Therefore, the choice of empirical therapy should be based on local epidemiology and the patient's colonization status. Definitive therapy should be determined by antimicrobial susceptibility testing. Antimicrobial options for the management of infections caused by *A. baumannii* are displayed in Table 157-1.

Acinetobacter species possess intrinsic β-lactamases that inactivate first- and second-generation cephalosporins. Through acquisition of extended-spectrum β-lactamases, the organisms can also become resistant to third- and fourth-generation cephalosporins. Nevertheless, when the isolate is susceptible, β-lactam agents are the drugs of choice for the treatment of *A. baumannii*. Among β-lactamase inhibitors, sulbactam is active against *A. baumannii* and is as effective as carbapenems and polymyxins.

Carbapenems have been the preferred drugs for treatment of invasive or hospital-acquired infections. Unfortunately, surveillance data from U.S. hospitals show that up to 50% of *A. baumannii* isolates recovered from ICUs are carbapenem resistant, and rates of carbapenem resistance are even higher around the world.

Aminoglycosides are of limited utility against *A. baumannii* because of toxicity and lack of lung penetration. Inhaled formulations of tobramycin have been used with variable success.

Polymyxins are cationic detergents that fell out of use as a result of nephrotoxicity and neurotoxicity. In vitro, they are the most active agents against carbapenem-resistant *A. baumannii*. Colistin has been used in both intravenous and inhaled formulations, although the optimal dosage has not yet been determined.

Tigecycline is a glycylcycline with clinical activity against *A. baumannii*. It reaches only low serum concentrations and therefore cannot be used for bloodstream infections. The susceptibility of isolates is variable, especially in outbreak settings, and the emergence of resistance during treatment has been reported.

Minocycline is a tetracycline that has a bacteriostatic effect on *A. baumannii*. Synergistic and bactericidal activity has been noted when minocycline is used in combination with colistin or a carbapenem.

Fosfomycin is an inhibitor of peptidoglycan synthesis that has no direct activity against *A. baumannii* but has been observed to be synergistic in vitro in combination with colistin or sulbactam. Clinical data have shown higher rates of microbiologic cure, but no differences in clinical response, with combinations of fosfomycin and colistin.

In vitro data favor combination therapy with colistin in many different regimens containing a carbapenem (imipenem, meropenem), rifampin, minocycline, ceftazidime, azithromycin, doxycycline, trimethoprim-sulfamethoxazole, or ampicillin-sulbactam. However, clinical data have not shown such combination therapy to be superior to colistin alone.

TABLE 157-1

Therapeutic Options for the Management of Multidrug-Resistant *Acinetobacter baumannii* Infections

Antibiotic	Dosing[a]	Comments
Sulbactam	3–9 g/d (9–27 g/d if given in combination with ampicillin)	Unavailable as single drug in many countries (including the United States)
Meropenem	2 g q8h	Prolonged infusion (3–4 h) has been used; limited data
Imipenem	500 mg q6h	Prolonged infusion (3–4 h) has been used; limited data
Colistin	Loading dose of 5 mg/kg followed by 2.5–5.0 mg/kg per day of colistin base given in 2–4 divided doses	Optimal dosing regimen unknown Inhaled formulation has been used as adjunct treatment in lung infections.
Polymyxin B	1.5–3 mg/kg q12h	
Tigecycline	100-mg loading dose followed by 50 mg q12h	Low serum concentrations and bacteriostatic activity limit use in bacteremia.
Minocycline	100 mg q12h	Loading dose of 200 mg IV has been used.
Amikacin	15 mg/kg qd	Inhaled formulation of tobramycin has been used as adjunct treatment in lung infections.
Rifampin	600 mg qd or 600 mg q12h	Use in combination therapy
Fosfomycin	4 g q12h PO	Use in combination therapy IV formulation not available in the United States

[a]All drugs are given by the IV route unless otherwise stated.

COMPLICATIONS AND PROGNOSIS

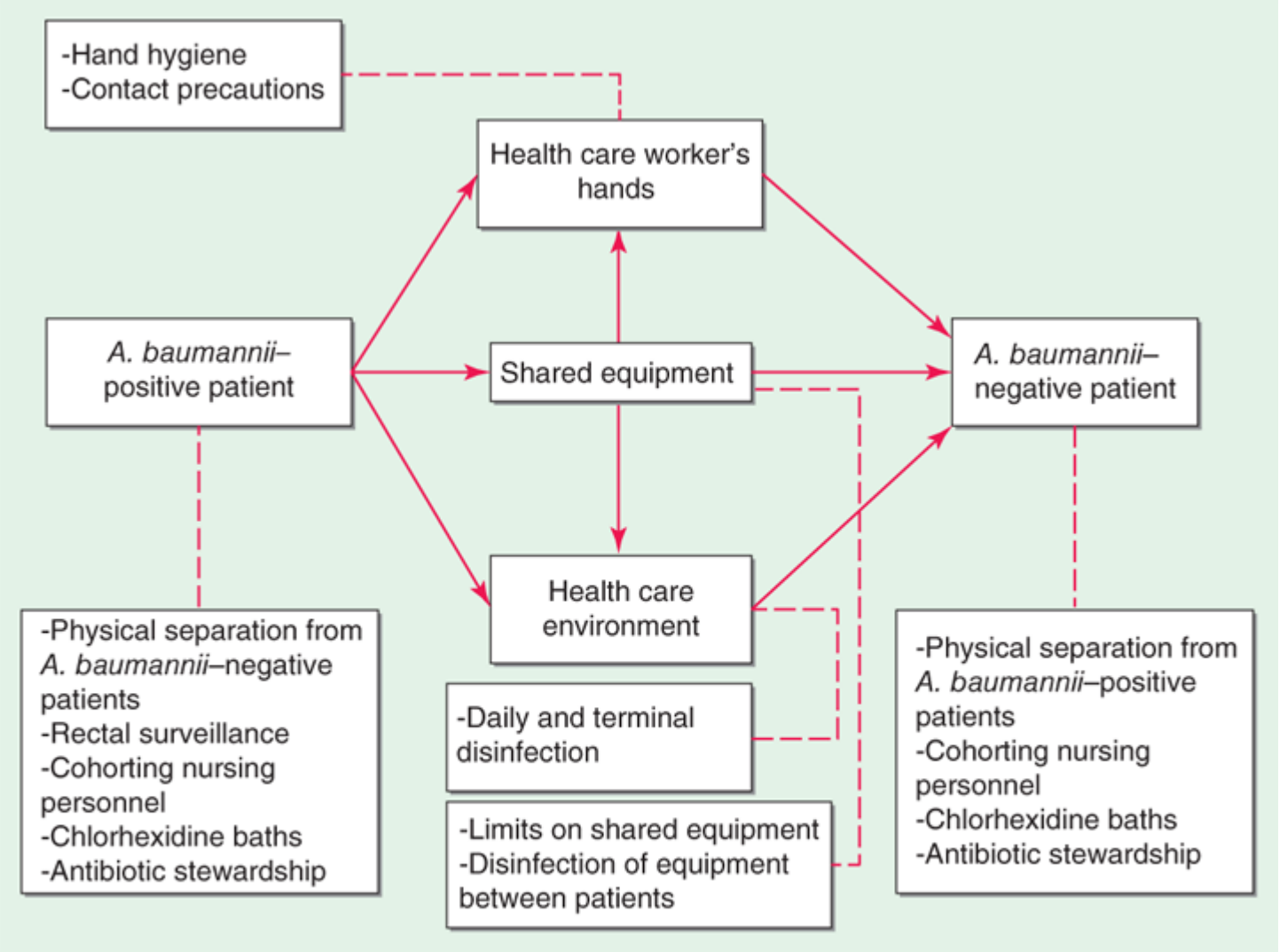
Infections caused by *A. baumannii* can be associated with high mortality rates. Factors contributing to higher mortality are thought to include severity of the patient's underlying illness and drug resistance in the infecting strain.

INFECTION CONTROL AND PREVENTION

Acinetobacter species are capable of surviving on hospital surfaces for prolonged periods. In the hospital environment, *A. baumannii* has been associated with establishment of a *fecal patina*; this term refers to a coating of enteric organisms that can cover the skin of colonized patients and extend to their surrounding environment. Concentrations of enteric organisms are highest in the colonized patient's rectum, with spread in a target-like concentric pattern covering the patient's body and the surrounding environment. High-frequency touch areas in rooms occupied by patients colonized with *A. baumannii* are more likely to be contaminated. The hands, gloves, and gowns of health care workers can be contaminated after entry into the room of a patient colonized with *A. baumannii* (Fig. 157-1).

FIGURE 157-1

Strategies for the prevention of dissemination of *Acinetobacter baumannii* in health care facilities.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Outbreaks caused by *A. baumannii* are frequently mono- or oligoclonal. A common source of infection has been identified in ~50% of outbreaks. These sources include respiratory therapy equipment, the hands of health care workers, bedside humidifiers, warm bathwater, hospital-prepared distilled water, bedpans, urine jugs, heparinized saline solution, mattresses, reusable pressure transducers in arterial lines, and fluids used for pressure lavage of wounds.

Control of multidrug-resistant *Acinetobacter* outbreaks starts with early recognition, with subsequent halting of the spread of infection throughout a facility and prevention of the establishment of an endemic strain. It is important to identify the outbreak strain and differentiate it from non-outbreak strains so that infection control activities can be better targeted. Traditionally, the strain was identified with phenotypic typing systems (biotyping) or by determination of antimicrobial susceptibility patterns. Molecular typing systems have ushered in an era of molecular epidemiology that allows more precise identification of outbreak strains through use of techniques such as ribotyping, pulse-field gel electrophoresis, repetitive sequence-based PCR, and multilocus sequence testing.

During outbreaks, the simultaneous introduction of multiple (“bundled”) measures makes it difficult to assess the impact of each individual measure. These interventions include aggressive cleaning of the general environment, active surveillance, contact isolation of colonized or infected patients, cohorting of medical staff, reinforcement of compliance with hand hygiene by health care workers, and use of aseptic care devices.

Colonization with *A. baumannii* is a strong predictor of subsequent clinical infection by this organism. Exposure to carbapenems is a risk factor for initial acquisition of this pathogen; therefore, efforts to curtail unnecessary use of antibiotics are fundamental to the prevention of *A. baumannii* colonization of patients and the organism’s establishment in health care facilities.

FURTHER READING

Adams-Haduch JM et al: Molecular epidemiology of carbapenem-nonsusceptible *Acinetobacter baumannii* in the United States. J Clin Microbiol 49:3849, 2011.

Antunes LC et al: *Acinetobacter baumannii:* evolution of a global pathogen. Pathog Dis 71:292, 2014.

Dexter C et al: Community-acquired *Acinetobacter baumannii*: Clinical characteristics, epidemiology and pathogenesis. Expert Rev Anti Infect Ther 13:567, 2015.

Garnacho-Montero J: Optimum treatment strategies for carbapenem-resistant *Acinetobacter baumannii*: Bacteremia. Expert Rev Anti Infect Ther 13:769, 2015.

Lee CR et al: Biology of *Acinetobacter baumannii*: Pathogenesis, antibiotic resistance mechanisms, and prospective treatment options. Front Cell Infect Microbiol 7:55, 2017.

Munoz-Price LS: Controlling multi-drug resistant gram-negative bacilli in your hospital: A transformational journey. J Hosp Infect 89:254, 2015.

Munoz-Price LS, Weinstein RA: *Acinetobacter* infection. N Engl J Med 358:1271, 2008.

Peleg AY et al: *Acinetobacter baumannii*: Emergence of a successful pathogen. Clin Microbiol Rev 21:538, 2008.

Tal-Jasper R et al: Clinical and epidemiological significance of carbapenem resistance in *Acinetobacter baumannii* infections. Antimicrob Agents Chemother 60:3127, 2016.

Wong D et al: Clinical and pathophysiological overview of *Acinetobacter* infections: A century of challenges. Clin Microbiol Rev 30:409, 2017.

Chapter 158: *Helicobacter pylori* Infections

John C. Atherton; Martin J. Blaser

INTRODUCTION



Helicobacter pylori colonizes the stomach in ~50% of the world's human population, essentially for life unless eradicated by antibiotic treatment. Colonization with this organism is the main risk factor for peptic ulceration **(Chap. 317)** as well as for gastric adenocarcinoma and gastric mucosa-associated lymphoid tissue (MALT) lymphoma **(Chap. 76)**. Treatment for *H. pylori* has revolutionized the management of peptic ulcer disease, providing a permanent cure in most cases. Such treatment also represents first-line therapy for patients with low-grade gastric MALT lymphoma. Treatment of *H. pylori* is of no benefit in the treatment of gastric adenocarcinoma, but prevention of *H. pylori* colonization or eradicative treatment could potentially prevent gastric malignancy and peptic ulceration. In contrast, increasing evidence indicates that lifelong *H. pylori* colonization may offer some protection against complications of gastroesophageal reflux disease (GERD), including esophageal adenocarcinoma. Recent research has focused on whether *H. pylori* colonization is also a risk factor for some extragastric diseases and whether it is protective against some recently emergent medical problems, such as childhood-onset asthma and other allergic and metabolic conditions.

ETIOLOGIC AGENT

Helicobacter pylori

H. pylori is a gram-negative bacillus that has naturally colonized humans for at least 100,000 years, and probably throughout human evolution. It lives in gastric mucus, with a proportion of the bacteria adherent to the mucosa and possibly a very small number of the organisms entering cells or penetrating the mucosa; the organism's distribution is never systemic. Its spiral shape and flagella render *H. pylori* motile in the mucus environment. The organism has several acid-resistance mechanisms, most notably a highly expressed urease that catalyzes urea hydrolysis to produce buffering ammonia. *H. pylori* is microaerophilic (i.e., requires low levels of oxygen), is slow-growing, and requires complex growth media in vitro.

Other *Helicobacter* Species

A very small proportion of gastric *Helicobacter* infections are due to species other than *H. pylori*, possibly acquired as zoonoses. These non-*pylori* gastric helicobacters are associated with low-level inflammation and occasionally with disease. In immunocompromised hosts, several nongastric (intestinal) *Helicobacter* species can cause disease with clinical features resembling those of *Campylobacter* infections; these species are covered in **Chap. 162**.

EPIDEMIOLOGY

Prevalence and Risk Factors


The prevalence of *H. pylori* among adults is <30% in most parts of the United States and in other developed countries as opposed to >80% in some developing countries. In the United States, prevalence varies with age: up to 50% of 60-year-old persons, ~20% of 30-year-old persons, and <10% of children are colonized. *H. pylori* is usually acquired in childhood. The age association is due mostly to a birth-cohort effect whereby current 60-year-olds were more commonly colonized as children than are current children. Spontaneous acquisition or loss of *H. pylori* in adulthood is uncommon. Childhood acquisition explains why the main risk factors for infection are markers of crowding and social deprivation in childhood.

Transmission


Humans are the only important reservoir of *H. pylori*. Children may acquire the organism from their parents (most often the primary caregiver) or from other children. The former is more common in developed countries and the latter in less developed countries. Whether transmission takes place more often by the fecal–oral or the oral–oral route is unknown, but *H. pylori* is easily cultured from vomitus and gastroesophageal refluxate and is less easily cultured from stool.

PATHOLOGY AND PATHOGENESIS

H. pylori colonization induces *chronic superficial gastritis*, a tissue response in the stomach that includes infiltration of the mucosa by both mononuclear and polymorphonuclear cells. (The term *gastritis* should be used specifically to describe histologic features; it has also been used to describe endoscopic appearances and even symptoms, but these features do not correlate with microscopic findings or even with the presence of *H. pylori*.) Although *H. pylori* is capable of numerous adaptations that prevent excessive stimulation of the immune system, colonization is accompanied by a considerable persistent local and systemic immune response, including the production of antibodies and cell-mediated responses. However, these responses are ineffective in clearing the bacterium. This inefficient clearing appears to be due in part to *H. pylori*'s downregulation of the immune system, which fosters its own persistence.

Most *H. pylori*–colonized persons do not develop clinical sequelae. That some persons develop overt disease whereas others do not is related to a combination of factors: bacterial strain differences, host susceptibility to disease, and environmental factors.

Bacterial Virulence Factors


Several *H. pylori* virulence factors are more common among strains that are associated with disease than among those that are not. The *cag* island is a group of genes that encodes a bacterial type IV secretion system. Through this system, an effector protein, CagA, is translocated into epithelial cells, where it may be activated by phosphorylation and induces host cell signal transduction; proliferative, cytoskeletal, and inflammatory changes in the cell result. The protein at the tip of the secretory apparatus, CagL, binds to integrins on the cell surface, transducing further signaling. Finally, soluble components of the peptidoglycan cell wall enter the cell, mediated by the same secretory system. These components are recognized by the intracellular bacterial receptor Nod1, which stimulates a proinflammatory

cytokine response resulting in an enhanced tissue response. Carriage of *cag*-positive strains increases the risk of peptic ulcer or gastric adenocarcinoma. A second major host-interaction factor is the vacuolating cytotoxin VacA, which forms pores in cell membranes. VacA is polymorphic, and carriage of more active forms also increases the risk of ulcer disease and gastric cancer. Other bacterial factors that are associated with increased disease risk include adhesins, such as BabA (which binds to blood group antigens on epithelial cells).

Host Genetic and Environmental Factors

The best-characterized host determinants of disease are genetic polymorphisms leading to enhanced activation of the innate immune response, including polymorphisms in cytokine genes and in genes encoding bacterial recognition proteins such as Toll-like receptors. For example, colonized people with polymorphisms in the interleukin 1 gene that increase the production of this cytokine in response to *H. pylori* infection are at increased risk of gastric adenocarcinoma. In addition, environmental cofactors are important in pathogenesis. Smoking increases the risks of duodenal ulcers and gastric cancer in *H. pylori*–positive individuals. Diets high in salt and preserved foods increase cancer risk, whereas diets high in antioxidants and vitamin C are modestly protective.

Distribution of Gastritis and Differential Disease Risk

The pattern of gastric tissue response is associated with disease risk: antral-predominant gastritis is most closely linked with duodenal ulceration, whereas pan-gastritis and corpus-predominant gastritis are linked with gastric ulceration and adenocarcinoma. This difference probably explains why patients with duodenal ulceration are not at high risk of developing gastric adenocarcinoma later in life, despite being colonized by *H. pylori*.

PATHOGENESIS OF DUODENAL ULCERATION

How gastric colonization causes duodenal ulceration is now becoming clearer. *H. pylori*–induced tissue responses in the gastric antrum diminish the number of somatostatin-producing D cells. Because somatostatin inhibits gastrin release, gastrin levels are higher than in *H. pylori*–negative persons, and these higher levels lead to increased meal-stimulated acid secretion from the relatively spared gastric corpus. How this situation increases duodenal ulcer risk remains controversial, but the increased acid secretion may contribute to the formation of the potentially protective gastric metaplasia found in the duodenum of duodenal ulcer patients. Gastric metaplasia in the duodenum may become colonized by *H. pylori* and subsequently inflamed and ulcerated.

PATHOGENESIS OF GASTRIC ULCERATION AND GASTRIC ADENOCARCINOMA

The pathogenesis of these conditions is less well understood, although both arise in association with pan- or corpus-predominant gastritis. The hormonal changes described above still occur, but the tissue responses in the gastric corpus mean that it produces less acid (hypochlorhydria) despite hypergastrinemia. Gastric ulcers commonly occur at the junction of antral and corpus-type mucosa, an area that is often particularly inflamed. Gastric cancer probably stems from progressive DNA damage and the survival of abnormal epithelial cell clones. The DNA damage is thought to be due principally to reactive oxygen and nitrogen species arising from inflammatory cells, perhaps in relation to other bacteria that survive in a hypochlorhydric stomach. Longitudinal analyses of gastric biopsy specimens taken years apart from the same patient show that the common *intestinal* type of gastric adenocarcinoma follows stepwise changes from simple gastritis to gastric atrophy, metaplasia, and dysplasia. A second, *diffuse* type of gastric

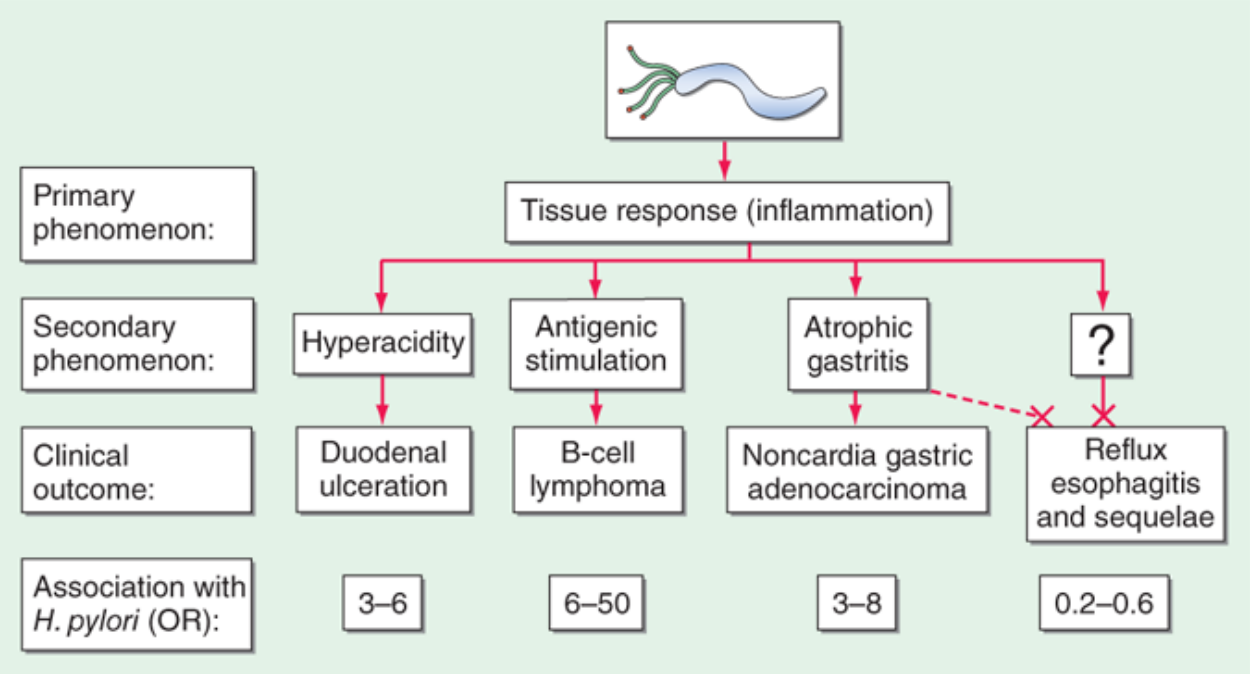
adenocarcinoma found more commonly in younger adults may arise directly from chronic gastritis without atrophic changes.

CLINICAL MANIFESTATIONS

Essentially all *H. pylori*–colonized persons have histologic gastritis, but only ~10–15% develop associated illnesses such as peptic ulceration, gastric adenocarcinoma, or gastric lymphoma **(Fig. 158-1)**. Despite similar rates of *H. pylori* colonization, rates of these diseases among women are less than half of those among men.

FIGURE 158-1

Schematic of the relationships between colonization with *Helicobacter pylori* and diseases of the upper gastrointestinal tract. Essentially all persons colonized with *H. pylori* develop a host response, which is generally termed *chronic gastritis*. The nature of the host's interaction with the particular bacterial population determines the clinical outcome. *H. pylori* colonization increases the lifetime risk of peptic ulcer disease, noncardia gastric cancer, and B-cell non-Hodgkin's gastric lymphoma (odds ratios [ORs] for all, >3). In contrast, a growing body of evidence indicates that *H. pylori* colonization (especially with *cagA*$^+$ strains) protects against adenocarcinoma of the esophagus (and the sometimes related gastric cardia) and premalignant lesions such as Barrett's esophagus (OR, <1). Although the incidences of peptic ulcer disease (cases not due to nonsteroidal anti-inflammatory drugs) and noncardia gastric cancer are declining in developed countries, the incidence of adenocarcinoma of the esophagus is increasing. *(Adapted from MJ Blaser: Hypothesis: The changing relationships of Helicobacter pylori and humans: Implications for health and disease. J Infect Dis 179:1523, 1999, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Peptic Ulcer Disease

Worldwide, ~70% of duodenal ulcers and ~50% of gastric ulcers are related to *H. pylori* colonization **(Chap. 317)**. However, in particular, the proportion of gastric ulcers caused by aspirin and nonsteroidal anti-inflammatory drugs

(NSAIDs) is increasing, and in many developed countries these drugs have overtaken *H. pylori* as a cause of gastric ulceration. The main lines of evidence supporting an ulcer-promoting role for *H. pylori* are that (1) the presence of the organism is a risk factor for the development of ulcers, (2) non-NSAID-induced ulcers rarely develop in the absence of *H. pylori*, (3) eradication of *H. pylori* virtually abolishes long-term ulcer relapse, and (4) experimental *H. pylori* infection of gerbils can cause gastric ulceration.

Gastric Adenocarcinoma and Lymphoma

Prospective nested case–control studies have shown that *H. pylori* colonization is a risk factor for adenocarcinomas of the distal (noncardia) stomach **(Chap. 76)**. Long-term experimental infection of gerbils also may result in gastric adenocarcinoma. Moreover, *H. pylori* may induce primary gastric lymphoma, although this condition is much less common. Many low-grade gastric B-cell lymphomas are dependent on *H. pylori* for continuing growth and proliferation, and these tumors may regress either fully or partially after *H. pylori* eradication. However, they require careful short- and long-term monitoring; any that are not confined to the superficial mucosa (and, indeed, some that are) require additional treatment with chemotherapeutic agents.

Functional Dyspepsia

Many patients have upper gastrointestinal symptoms but have normal results on upper gastrointestinal endoscopy (so-called functional or non-ulcer dyspepsia; **Chap. 317**). Because *H. pylori* is common, some of these patients will be colonized with the organism. *H. pylori* eradication leads to symptom resolution up to 15% more commonly than does placebo treatment. Whether such patients have peptic ulcers in remission at the time of endoscopy or whether a small subgroup of patients with “true” functional dyspepsia respond to *H. pylori* treatment is unclear. Either way, because functional dyspepsia is often persistent and difficult to treat, most guidelines recommend *H. pylori* eradication in these patients. If this advice is followed, it is important to realize that only a small subgroup of patients who are treated will benefit.

Protection Against Peptic Esophageal Disease, Including Esophageal Adenocarcinoma

Much interest has focused on a protective role for *H. pylori* against GERD **(Chap. 316)**, Barrett’s esophagus **(Chap. 316)**, and adenocarcinoma of the esophagus and gastric cardia **(Chap. 76)**. The main lines of evidence for this role are (1) that there is a temporal relationship between a falling prevalence of gastric *H. pylori* colonization and a rising incidence of these conditions; (2) that, in most studies, the prevalence of *H. pylori* colonization (especially with proinflammatory *cagA*$^+$ strains) is significantly lower among patients with these esophageal diseases than among control participants; and (3) that, in prospective nested studies (see above), the presence of *H. pylori* is inversely related to these cancers. The mechanism underlying this protective effect is likely *H. pylori*–induced hypochlorhydria. Because, at the individual level, GERD severity may decrease, worsen, or remain unchanged after *H. pylori* treatment, concerns about GERD should not affect decisions about whether to treat *H. pylori* in an individual patient when a clear-cut indication exists.

Other Pathologies

H. pylori has an increasingly recognized role in other gastric pathologies. It may predispose some patients to iron deficiency through occult blood loss and/or hypochlorhydria and reduced iron absorption. In addition, several extra-gastrointestinal pathologies have been linked with *H. pylori* colonization, although evidence of causality is less strong.

Studies of *H. pylori* treatment in idiopathic thrombocytopenic purpura have consistently described improvement in or even normalization of platelet counts. Potentially important but even more controversial (protective) associations are with ischemic heart disease and cerebrovascular disease. However, the strength of the latter associations is reduced if confounding factors are taken into account, and our present knowledge is incomplete. Most authorities consider the associations to be non-causal. An increasing number of studies have shown an inverse association of *cagA*$^{+}$ *H. pylori* with childhood-onset asthma, hay fever, and atopic disorders. These associations have been shown to be causal in animal models, but causality in humans and the size of any effect have not been established.

DIAGNOSIS

Tests for *H. pylori* fall into two groups: tests that require upper-gastrointestinal endoscopy and simpler tests that can be performed in the clinic **(Table 158-1)**.

TABLE 158-1

Tests Commonly Used to Detect *Helicobacter pylori*

Test	Advantages	Disadvantages
Tests Based on Endoscopic Biopsy		
Biopsy urease test	Quick, simple	Some commercial tests not fully sensitive before 24 h
Histology	May give additional histologic information	Sensitivity dependent on experience and use of special stains
Culture	Permits determination of antibiotic susceptibility	Sensitivity dependent on experience
Noninvasive Tests		
Serology	Inexpensive and convenient; not affected by recent antibiotics or proton pump inhibitors to the same extent as breath and stool tests	Cannot be used to monitor treatment success; some commercial kits inaccurate, and most less accurate than urea breath test
^{13}C urea breath test	Inexpensive and simpler than endoscopy; useful for follow-up after treatment	Requires fasting; not as convenient as blood or stool tests
Stool antigen test	Inexpensive and convenient; useful for follow-up after treatment; may be particularly useful in children	Stool-based tests disliked by people from some cultures

Endoscopy-Based Tests

Endoscopy is usually unnecessary in the initial management of young patients with simple dyspepsia but is commonly used to exclude malignancy and make a positive diagnosis in older patients or those with “alarm” symptoms. If endoscopy is performed, the most convenient biopsy-based test is the biopsy urease test, in which one large or two small gastric biopsy specimens are placed into a gel containing urea and an indicator. The presence of *H. pylori* urease leads to a pH alteration and therefore to a color change, which often occurs within minutes but can require up to 24 h. Histologic examination of biopsy specimens for *H. pylori* also is accurate, provided that a special stain (e.g., a modified Giemsa, silver, or immuno-stain) permitting optimal visualization of the organism is used. If biopsy specimens are obtained from both antrum and corpus, histologic study yields additional information, including the degree and pattern of inflammation and the presence of any atrophy, metaplasia, or dysplasia. Microbiologic culture is most specific but may be insensitive because of difficulty with *H. pylori* isolation. Once the organism is cultured, its identity as *H. pylori* can be confirmed by its typical appearance on Gram’s stain and its positive reactions in oxidase, catalase, and urease tests. Moreover, the organism’s susceptibility to antibiotics can be determined, and this information can be clinically useful in difficult cases. The occasional biopsy specimens containing the less common non-*pylori* gastric helicobacters give weakly positive results in the biopsy urease test. Positive identification of these bacteria requires visualization of the characteristic long, tight spirals in histologic sections; they cannot easily be cultured.

Noninvasive Tests

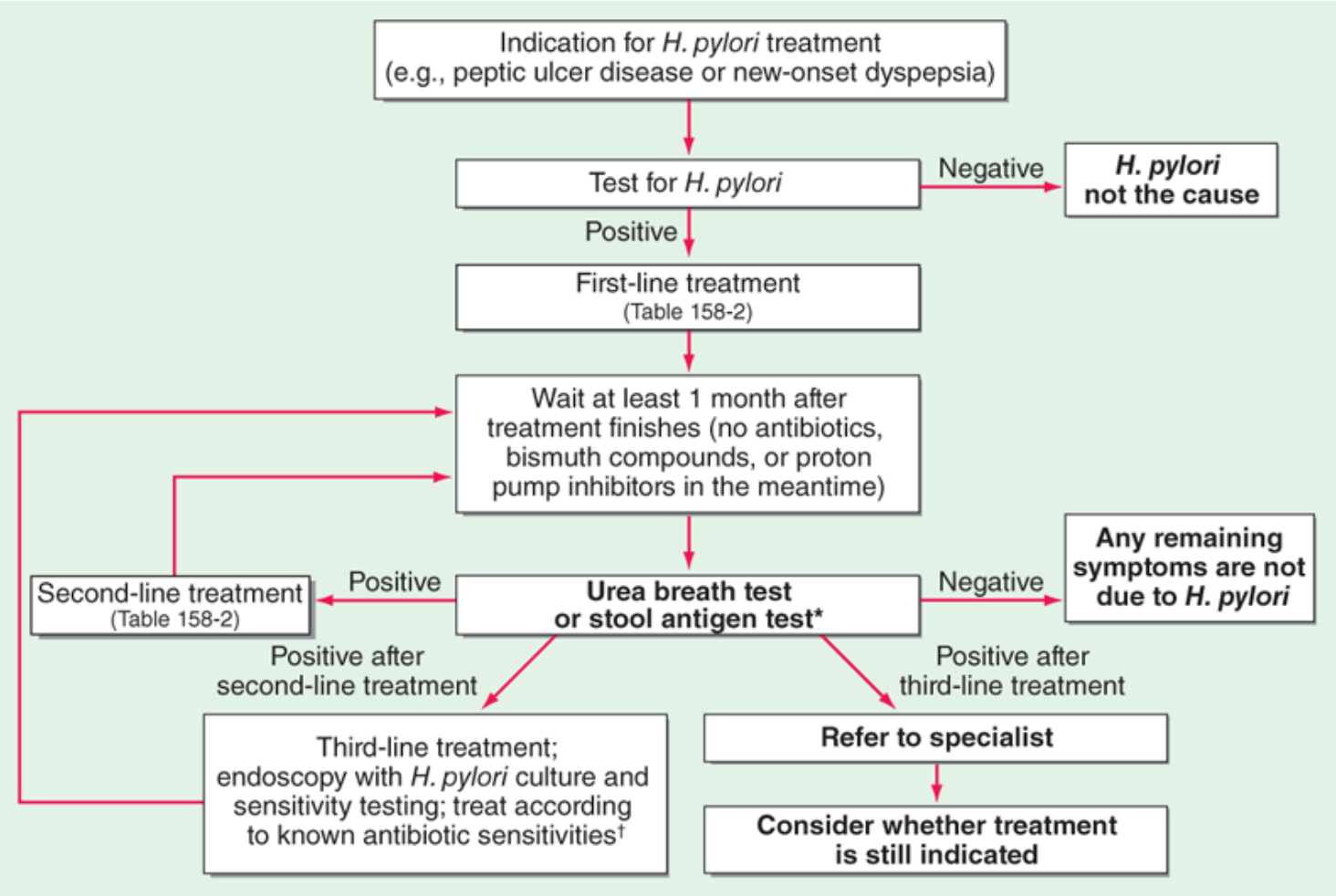
Noninvasive *H. pylori* testing is the norm if gastric cancer does not need to be excluded by endoscopy. The best-established test (and a very accurate one) is the *urea breath test*. In this simple test, the patient drinks a solution of urea labeled with the nonradioactive isotope ^{13}C and then blows into a tube. If *H. pylori* urease is present, the urea is hydrolyzed, and labeled carbon dioxide is detected in breath samples. The *stool antigen test*, a simple and accurate test using monoclonal antibodies specific for *H. pylori* antigens, is more convenient and potentially less expensive than the urea breath test, but some patients dislike sampling stool. The simplest tests for ascertaining *H. pylori* status are *serologic assays* measuring specific IgG levels in serum by enzyme-linked immunosorbent assay or immunoblot. The best of these tests are as accurate as other diagnostic methods, but many commercial tests—especially rapid office tests—do not perform well.

Use of Tests to Assess Treatment Success

The urea breath test, the stool antigen test, and biopsy-based tests can all be used to assess the success of treatment **(Fig. 158-2)**. However, because these tests are dependent on *H. pylori* load, their use <4 weeks after treatment may yield false-negative results. Early suppression of bacterial numbers may lead to false-negative results, since regrowth of the organism can result in its detection weeks later. For the same reason, these tests are unreliable if performed within 4 weeks of intercurrent treatment with antibiotics or bismuth compounds or within 2 weeks of the discontinuation of proton pump inhibitor (PPI) treatment. In the assessment of treatment success, noninvasive tests are normally preferred. However, after gastric ulceration, endoscopy should be repeated to ensure healing and to exclude gastric carcinoma by further histologic sampling; if PPIs have been stopped for at least 2 weeks and no antibiotics or bismuth compounds have been given for at least 4 weeks, there is an opportunity to assess treatment success with biopsy-based tests. Serologic tests are not used to monitor treatment success, as the gradual drop in titer of *H. pylori*–specific antibodies is too slow to be of practical use.

FIGURE 158-2

Algorithm for the management of *Helicobacter pylori* infection. *Note that either the urea breath test or the stool antigen test can be used in this algorithm. Occasionally, endoscopy and a biopsy-based test are used instead of either of these tests in follow-up after treatment. The main indication for these invasive tests is gastric ulceration; in this condition, as opposed to duodenal ulceration, it is important to check healing and to exclude underlying gastric adenocarcinoma. However, even in this situation, patients undergoing endoscopy may still be receiving proton pump inhibitor therapy, which precludes *H. pylori* testing. Thus, a urea breath test or a stool antigen test is still required at a suitable interval after the end of therapy to determine whether treatment has been successful (see text). †Some authorities use empirical third-line regimens, of which several have been described.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Helicobacter pylori Infection

INDICATIONS

The most clear-cut indications for treatment are *H. pylori*–related duodenal or gastric ulceration or low-grade gastric B-cell lymphoma. Whether or not the ulcers are currently active, *H. pylori* should be eradicated in patients with documented ulcer disease to prevent relapse (Fig. 158-2). Guidelines have recommended *H. pylori* treatment for

colonized patients with functional dyspepsia in case they are among the ~10% who will benefit from such therapy (beyond placebo effects). *H. pylori* eradication in the treatment of conditions not definitively known to respond has also been recommended but is not universally supported; such conditions include idiopathic thrombocytopenic purpura, vitamin B_{12} deficiency, and iron-deficiency anemia where other causes have been carefully excluded. For individuals with a strong family history of gastric cancer, treatment to eradicate *H. pylori* in the hope of reducing cancer risk is reasonable but of unproven value. For older dyspeptic patients in the community or those who have "alarm" symptoms (e.g., weight loss) associated with their dyspepsia, upper-gastrointestinal endoscopy is indicated to seek a diagnosis and test for *H. pylori*; the decision over whether to eradicate the organism can then be based on indication. Endoscopy is usually considered unnecessary for young dyspeptic patients in the community who have no alarm symptoms (with the precise age cutoff dependent on local guidelines). If the community prevalence of *H. pylori* is below ~20%, such patients are treated with a short course of acid suppression using a PPI. If these patients do not respond or relapse when treatment is stopped, or if the *H. pylori* community prevalence is >20%, all national guidelines recommend a strategy of testing for *H. pylori* noninvasively and eradicating it if it is found. This strategy will benefit patients who have peptic ulcers and the 10% of patients who have functional dyspepsia responsive to *H. pylori* eradication, but most patients will be treated unnecessarily. Currently, widespread community screening for and treatment of *H. pylori* as primary prophylaxis for gastric cancer and peptic ulcers are not recommended in most countries, mainly because the extent of the consequent reduction in cancer risk is not known. Several studies have found a modestly reduced cancer risk after treatment, but the period of follow-up is still fairly short, and the size of the effect in different populations remains unclear. Other reasons not to treat *H. pylori* in asymptomatic populations at present include (1) the adverse side effects (which are common and can be severe in rare cases) of the multiple-antibiotic regimens used; (2) antibiotic resistance, which may emerge in *H. pylori* or other incidentally carried bacteria; (3) the anxiety that may arise in otherwise healthy people, especially if treatment is unsuccessful; and (4) the existence of a subset of people who will develop GERD symptoms after treatment, although in most cases *H. pylori* treatment does not affect GERD symptoms or severity. Despite the absence of screening strategies, many doctors treat *H. pylori* if it is known to be present (particularly in children and younger adults), even when the patient is asymptomatic. The rationale is that it reduces patient concern and may reduce future gastric cancer risk and that any reduction in risk is likely to be greater in younger patients. However, such practices do not factor in any potential benefits of *H. pylori* colonization. Overall, despite widespread clinical activity in this area, most treatment of persons with asymptomatic *H. pylori* carriage is given without a firm evidence base.

REGIMENS

Although *H. pylori* is susceptible to a wide range of antibiotics in vitro, monotherapy is not usually successful, probably because of inadequate active antibiotic delivery to the colonization niche. Clinical failure of monotherapy prompted the development of multidrug regimens. Current regimens consist of a PPI and two or three antimicrobial agents given for 10–14 days **(Table 158-2)**. The optimal regimens vary in different parts of the world, depending on the known rates of primary antibiotic resistance in most *H. pylori* strains in a particular locale. For this reason, guidelines on optimal regimens for *H. pylori* eradication in individual countries are evolving, and physicians should refer to the most up-to-date local guideline.

 The two most important factors in successful *H. pylori* treatment are the patient's close compliance with the regimen and the use of drugs to which the patient's strain of *H. pylori* has not acquired resistance. Treatment failure following minor lapses in compliance is common and often leads to acquired resistance. To stress the importance of compliance, written instructions should be given to the patient, and minor side effects of the regimen should be explained. Increasing levels of primary *H. pylori* resistance to clarithromycin, levofloxacin, and—to a lesser extent—

metronidazole are of growing concern. In most parts of the world (the main exception being northwestern Europe), the rate of primary clarithromycin resistance is sufficiently high that regimens containing clarithromycin plus one other antibiotic often fail; regimens with clarithromycin and two other antibiotics remain an option as the other two antibiotics are likely to eradicate *H. pylori* even if the strain is clarithromycin resistant. When a patient is known to have been exposed—even distantly—to clarithromycin or a fluoroquinolone, these antibiotics usually should be avoided. Resistance to amoxicillin or tetracycline is unusual, even if these antibiotics have been given previously, and resistance to metronidazole is only partial; thus there is no need to avoid using these antibiotics whether or not they have been previously prescribed. Assessment of antibiotic susceptibilities before treatment would be optimal but is not usually undertaken because endoscopy and mucosal biopsy are necessary to obtain *H. pylori* for culture and because most microbiology laboratories are inexperienced in *H. pylori* culture. If initial *H. pylori* treatment fails, the usual approach is empirical re-treatment with another drug regimen (Table 158-2). The third-line approach ideally should be endoscopy, biopsy, and culture plus treatment based on documented antibiotic sensitivities. However, empirical third-line therapies are often used.

Non-*pylori* gastric helicobacters are treated in the same way as *H. pylori*. However, in the absence of trials, it is unclear whether a positive outcome always represents successful treatment or whether it is sometimes due to natural clearance of the bacteria.

TABLE 158-2

Commonly Recommended Treatment Regimens for *Helicobacter pylori*

Regimen[a] (Duration)	Drug 1	Drug 2	Drug 3	Drug 4
Regimen 1: OCM (14 days)[b]	Omeprazole (20 mg bid[c])	Clarithromycin (500 mg bid)	Metronidazole (500 mg bid)	—
Regimen 2: OCA (14 days)[b]	Omeprazole (20 mg bid[c])	Clarithromycin (500 mg bid)	Amoxicillin (1 g bid)	—
Regimen 3: OBTM (14 days)[d]	Omeprazole (20 mg bid[c])	Bismuth subsalicylate (2 tabs qid)	Tetracycline HCl (500 mg qid)	Metronidazole (500 mg tid)
Regimen 4: concomitant (14 days)[e]	Omeprazole (20 mg bid[c])	Amoxicillin (1 g bid)	Clarithromycin (500 mg bid)	Tinidazole (500 mg bid[f])
Regimen 5: OAL (10 days)[g]	Omeprazole (20 mg bid[c])	Amoxicillin (1 g bid)	Levofloxacin (500 mg bid)	—

[a]The recommended first-line regimens for most of the world are shown in **bold** type. [b]This regimen should be used only for populations in which the prevalence of clarithromycin-resistant strains is known to be <20%. In practice, this restriction limits the regimens' appropriate range mainly to northern Europe. [c]Many authorities and some guidelines recommend doubling this dose of omeprazole, as trials show resultant increased efficacy with some antibiotic combinations. Omeprazole may be replaced with any proton pump inhibitor at an equivalent dosage. Because extensive metabolizers of PPIs are prevalent among Caucasian populations, many authorities recommend esomeprazole (40 mg bid) or rabeprazole (20 mg bid), particularly for regimens 4 and 5. [d]Data supporting this regimen come mainly from Europe and are based on the use of bismuth subcitrate (1 tablet qid) and metronidazole (400 mg tid). This is a recommended first-line regimen in most countries and is the recommended second-line regimen in northern Europe. [e]This regimen may be used as an alternative to regimen 3. [f]Metronidazole (500 mg bid) may be used as an alternative. [g]This regimen is used as second-line treatment in many countries (particularly where quadruple or concomitant therapy is used as the first-line regimen) and as third-line treatment in others. It may be less effective where rates of fluoroquinolone use are high and is more likely to be ineffective if there is a personal history of fluoroquinolone use for previous treatment of other infections.

PREVENTION

Carriage of *H. pylori* has considerable public health significance in developed countries, where it is associated with peptic ulcer disease and gastric adenocarcinoma, and in developing countries, where gastric adenocarcinoma may be an even more common cause of cancer death late in life. If mass prevention were contemplated, vaccination would be the most obvious method: experimental immunization of animals has given promising results, and the first

reported trial in humans has shown some efficacy. Further trials are ongoing. However, given that *H. pylori* has co-evolved with its human host over millennia, preventing or eliminating colonization on a population basis may have biological and clinical costs. For example, lifelong absence of *H. pylori* is a risk factor for GERD complications, including esophageal adenocarcinoma. We have speculated that the disappearance of *H. pylori* may also be associated with an increased risk of other emergent diseases reflecting aspects of the current Western lifestyle, such as childhood-onset asthma and allergy.

FURTHER READING

Amieva M, Peek RM: Pathobiology of *Helicobacter pylori*–induced gastric cancer. Gastroenterology 150:64, 2016.

Arnold IC et al: *Helicobacter pylori* infection prevents allergic asthma in mouse models through the induction of regulatory T cells. J Clin Invest 121:3088, 2011.

Atherton JC, Blaser MJ: Co-adaptation of *Helicobacter pylori* and humans: Ancient history and modern implications. J Clin Invest 119:2475, 2009.

Chen Y, Blaser MJ: Inverse associations of *Helicobacter pylori* with asthma and allergies. Arch Intern Med 167:821, 2007.

Chen Y et al: Association between *Helicobacter pylori* and mortality in the NHANES II study. Gut 62:1262, 2013.

Chow WH et al: An inverse relation between *cagA*+ strains of *Helicobacter pylori* infection and risk of esophageal and gastric cardia adenocarcinoma. Cancer Res 58:588, 1998.

Ford AC et al: *Helicobacter pylori* eradication therapy to prevent gastric cancer in healthy asymptomatic infected individuals: Systematic review and meta-analysis of randomized controlled trials. BMJ 348:g3174, 2014. [PubMed: 24846275]

Linz B et al: An African origin for the intimate association between humans and *Helicobacter pylori*. Nature 445:915, 2007.

Maixner F et al: The 5300-year-old *Helicobacter pylori* genome of the Iceman. Science 351:162, 2016.

Marshall BJ, Warren JR: Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet 1:1311, 1984.

Plummer M et al: Global burden of gastric cancer attributable to *Helicobacter pylori*. Int J Cancer 136:487, 2015.

Chapter 159: Infections Due to *Pseudomonas*, *Burkholderia*, and *Stenotrophomonas* Species

Reuben Ramphal

FIGURE 159-1

INTRODUCTION

The pseudomonads are a heterogeneous group of gram-negative bacteria that have in common an inability to ferment lactose. Formerly classified in the genus *Pseudomonas*, the members of this group have been assigned to three medically important genera—*Pseudomonas, Burkholderia*, and *Stenotrophomonas*—whose biologic behaviors encompass both similarities and marked differences and whose genetic repertoires differ in many respects. The pathogenicity of most pseudomonads is based on opportunism; the exceptions are *Burkholderia pseudomallei* and *Burkholderia mallei*, which are primary pathogens.

Pseudomonas aeruginosa, the major pathogen of the group, is a significant cause of infections in hospitalized patients and in patients with cystic fibrosis (CF; Chap. 285). Cytotoxic chemotherapy, mechanical ventilation, and broad-spectrum antibiotic therapy set up conditions that predispose to colonization and infection of increasing numbers of hospitalized patients by this pathogen. The other members of the genus *Pseudomonas—Pseudomonas putida, Pseudomonas fluorescens*, and *Pseudomonas stutzeri*—infect humans infrequently.

The genus *Burkholderia* comprises >40 species, of which *Burkholderia cepacia* is most frequently encountered in Western countries. Like *P. aeruginosa, B. cepacia* is both a nosocomial pathogen and a cause of infection in CF. The other medically important members of this genus are *B. pseudomallei* and *B. mallei*, the etiologic agents of melioidosis and glanders, respectively.

The genus *Stenotrophomonas* contains one species of medical significance, *Stenotrophomonas maltophilia* (previously classified in the genera *Pseudomonas* and *Xanthomonas*). This organism is strictly an opportunist that “overgrows” in the setting of potent broad-spectrum antibiotic use.

PSEUDOMONAS AERUGINOSA

EPIDEMIOLOGY

P. aeruginosa is found in most moist environments. Soil, plants, vegetables, tap water, and countertops are all potential reservoirs for this microbe, as it has simple nutritional needs. Given the ubiquity of *P. aeruginosa*, it is clear that simple contact with the organism is not sufficient for colonization or infection. Clinical and experimental observations suggest that infection by *P. aeruginosa* occurs concomitantly with compromised host defenses, mucosal trauma, physiologic derangement, and antibiotic-mediated suppression of normal flora. Thus, it comes as no surprise that the majority of *P. aeruginosa* infections occur in intensive care units (ICUs), where these factors frequently converge. The organism is initially acquired from environmental sources, but patient-to-patient spread also occurs in clinics and families.

In the past, burned patients appeared to be unusually susceptible to *P. aeruginosa*. For example, in 1959–1963, *Pseudomonas* burn-wound sepsis was the principal cause of death in 60% of burned patients dying at the U.S. Army Institute of Surgical Research. For reasons that are unclear, *P. aeruginosa* infection in burns is no longer the major problem that it was during the 1950s and 1960s. Similarly, in the 1960s, *P. aeruginosa* appeared as a common pathogen in patients receiving cytotoxic chemotherapy at many institutions in the United States, but it has subsequently diminished in importance. Despite this subsidence, *P. aeruginosa* remains one of the most feared pathogens in this population because of its high attributable mortality.

 In some parts of Asia and Latin America, *P. aeruginosa* continues to be the most common cause of gram-negative bacteremia in neutropenic patients.

In contrast to the trends for burned patients and neutropenic patients in the United States, the incidence of *P. aeruginosa* infections among patients with CF has not changed. *P. aeruginosa* remains the most common contributing factor to respiratory failure in CF and is responsible for the majority of deaths among CF patients.

LABORATORY FEATURES

P. aeruginosa is a nonfastidious, motile, gram-negative rod that grows on most common laboratory media, including blood and MacConkey agars. It is easily identified in the laboratory on primary-isolation agar plates by pigment production that confers a yellow to dark green or even bluish appearance. Colonies have a shiny "gun-metal" appearance and a characteristic fruity odor. Two of the identifying biochemical characteristics of *P. aeruginosa* are an inability to ferment lactose on MacConkey agar and a positive reaction in the oxidase test. Most strains are identified on the basis of these readily detectable laboratory features even before extensive biochemical testing is done. Some isolates from CF patients are easily identified by their mucoid appearance, which is due to the production of large amounts of the mucoid exopolysaccharide or alginate.

PATHOGENESIS

Unraveling the mechanisms that underlie disease caused by *P. aeruginosa* has proved challenging. Of the common gram-negative bacteria, no other species produces such a large number of putative virulence factors (Table 159-1). Yet *P. aeruginosa* rarely initiates an infectious process in the absence of host injury or compromise, and few of its putative virulence factors have been shown definitively to be involved in disease

in humans. Despite its metabolic versatility and possession of multiple colonizing factors, *P. aeruginosa* exhibits no competitive advantage over enteric bacteria in the human gut; it is not a normal inhabitant of the human gastrointestinal tract, despite the host's continuous environmental exposure to the organism.

TABLE 159-1

Main Putative Virulence Factors of *Pseudomonas aeruginosa*

Substance/Organelle	Function	Virulence in Animal Disease
Pili	Adhesion to cells	?
Flagella	Adhesion, motility, inflammation	Yes
Lipopolysaccharide	Antiphagocytic activity, inflammation	Yes
Type III secretion system	Toxic activity (ExoU, ExoS)	Yes
Type II secretion system	Toxic activity	Yes
Proteases	Proteolytic activity	?
Phospholipases	Cytotoxicity	?
Exotoxin A	Cytotoxicity	?

Virulence Attributes Involved in Acute *P. aeruginosa* Infections

MOTILITY AND COLONIZATION

A general tenet of bacterial pathogenesis is that most bacteria must adhere to surfaces or colonize a host niche in order to initiate disease. Most gram-negative bacteria examined thus far possess adherence factors called *adhesins. P. aeruginosa* is no exception. Among its many adhesins are its pili, which demonstrate adhesive properties for a variety of cells and adhere best to injured cell surfaces. In the organism's flagellum, the flagellin molecule binds to cells, and the flagellar cap attaches to mucins through the recognition of glycan chains. Other *P. aeruginosa* adhesins include the outer core of the lipopolysaccharide (LPS) molecule, which binds to the cystic fibrosis transmembrane conductance regulator (CFTR) and aids in internalization of the organism, and the alginate coat of mucoid strains, which enhances adhesion to cells and mucins. In addition, membrane proteins and lectins have been proposed as colonization factors. The deletion of any given adhesin is not sufficient to abrogate the ability of *P. aeruginosa* to colonize surfaces. Motility is important in host invasion via mucosal surfaces in some animal models; however, nonmotile strains are not uniformly avirulent.

EVASION OF HOST DEFENSES

The transition from bacterial colonization to disease requires the evasion of host defenses followed by invasion by the microorganism. *P. aeruginosa* appears to be well equipped for evasion. Attached bacteria inject four known toxins (ExoS or ExoU, ExoT, and ExoY) via a type III secretion system that allows the bacteria to evade phagocytic cells either by direct cytotoxicity or by inhibition of phagocytosis. Clinical studies suggest that the mortality rate is higher among patients infected by strains that secrete the ExoU toxin. Another secretion system—the type II system—secretes toxins that can kill animals, and some of its secreted toxins, such as exotoxin A, have the potential to kill phagocytic cells. Multiple proteases secreted by this system may degrade host effector molecules, such as cytokines and chemokines, that are released in response to infection.

TISSUE INJURY

Among gram-negative bacteria, *P. aeruginosa* probably produces the largest number of substances that are toxic to cells and thus have the potential to injure tissues. The toxins secreted by the organism's type III secretion system are capable of injuring tissue. However, their delivery requires the adherence of the organism to cells. Thus, the effects of these toxins are likely to be local or to depend on the presence of vast numbers of bacteria. On the other hand, diffusible toxins, secreted by the organism's type II secretion system, can act freely wherever they come into contact with cells. Possible effectors include exotoxin A, at least four different proteases, and at least two phospholipases; in addition to these secreted toxins, rhamnolipids, pyocyanin, and hydrocyanic acid are produced by *P. aeruginosa* and are all capable of causing host injury.

INFLAMMATORY COMPONENTS

The inflammatory responses to the lipid A component of *Pseudomonas* LPS and to its flagellin, mediated through the Toll-like receptor (TLR) system (principally TLR4 and TLR5, respectively), are thought to represent important factors in disease causation. Although these inflammatory responses are required for successful defense against *P. aeruginosa* (i.e., in their absence, animals are defenseless against *P. aeruginosa* infection), florid responses are likely to result in disease. Thus, when the sepsis syndrome and septic shock develop in *P. aeruginosa* infection, they are probably the result of the host response to one or both of these substances, but injury to the lung by *Pseudomonas* toxins may also result in sepsis syndromes, possibly by causing cell death and the release of cellular components (e.g., heat-shock proteins) that may activate the TLR or another proinflammatory system.

Chronic *P. aeruginosa* Infections

Chronic infection due to *P. aeruginosa* occurs mainly in the lungs in the setting of structural pulmonary diseases. The classic example is CF; others include bronchiectasis and chronic relapsing panbronchiolitis, a disease seen in Japan and some Pacific Islands. A hallmark of these illnesses is altered mucociliary clearance leading to mucus stasis and mucus accumulation in the lungs. There is probably a common factor that selects for *P. aeruginosa* colonization in these lung diseases—perhaps the adhesiveness of *P. aeruginosa* for mucus, a phenomenon that is not noted for most other common gram-negative bacteria, and/or the ability of *P. aeruginosa* to evade host defenses in mucus. Furthermore, *P. aeruginosa* seems to evolve in ways that allow its prolonged survival in the lung without an early fatal outcome for the host. The strains found in CF patients exhibit minimal production of virulence factors. Many strains lose the ability to produce pili and

flagella, and most become complement-sensitive because of the loss of the O side chain of their LPS molecules. In addition, most strains found in CF patients overproduce a mucoid exopolysaccharide. These changes probably dampen the host response, allowing the organism to survive in CF mucus. *P. aeruginosa* is also believed to lose its ability to secrete many of its injectable toxins during growth in mucus. Although the alginate coat is thought to play a role in the organism's survival, alginate is not essential, as nonmucoid strains may predominate for long periods. In short, virulence in chronic infections may be mediated by the chronic but attenuated host inflammatory response, which injures the lungs over decades.

CLINICAL MANIFESTATIONS

P. aeruginosa causes infections at almost all sites in the body but shows a rather strong predilection for the lungs. The infections encountered most commonly in hospitalized patients are described below.

Bacteremia

Crude mortality rates exceeding 50% have been reported among patients with *P. aeruginosa* bacteremia. Consequently, this clinical entity has been much feared, and its management has been attempted with the use of multiple antibiotics. Recent publications report attributable mortality rates of 28–44%, with the precise figure depending on the adequacy of treatment and the seriousness of the underlying disease. In the past, the patient with *P. aeruginosa* bacteremia classically was neutropenic or had a burn injury. Today, however, a minority of such patients have bacteremic *P. aeruginosa* infections. Rather, *P. aeruginosa* bacteremia is seen most often in patients in ICUs.

The clinical presentation of *P. aeruginosa* bacteremia rarely differs from that of sepsis in general **(Chap. 297)**. Patients are usually febrile, but those who are most severely ill may be in shock or even hypothermic. The only point differentiating this entity from gram-negative sepsis of other causes may be the distinctive skin lesions (ecthyma gangrenosum) of *Pseudomonas* infection, which occur almost exclusively in markedly neutropenic patients and patients with AIDS. These small or large, painful, reddish, maculopapular lesions have a geographic margin; they are initially pink, then darken to purple, and finally become black and necrotic **(Fig. 159-1)**. Histopathologic studies indicate that the lesions are due to vascular invasion and are teeming with bacteria. Although similar lesions may occur in aspergillosis and mucormycosis, their presence suggests *P. aeruginosa* bacteremia as the most likely diagnosis.

FIGURE 159-1

Ecthyma gangrenosum in a neutropenic patient 3 days after onset.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

P. aeruginosa Bacteremia

Antimicrobial treatment of *P. aeruginosa* bacteremia has been controversial (Table 159-2). Before 1971, the outcome of *Pseudomonas* bacteremia in febrile neutropenic patients treated with the available agents—gentamicin and the polymyxins—was dismal. However, treatment with carbenicillin, with or without an aminoglycoside, significantly improved outcomes. Concurrently, several retrospective analyses suggested that the use of two agents that were synergistic against gram-negative pathogens in vitro resulted in better outcomes in neutropenic patients. Thus, combination therapy became the standard of care—first for *P. aeruginosa* bacteremia in febrile neutropenic patients and then for all *P. aeruginosa* infections in neutropenic or nonneutropenic patients.

With the introduction of newer antipseudomonal drugs, a number of studies have revisited the choice between combination treatment and monotherapy for *Pseudomonas* bacteremia. Although many experts still favor combination therapy, most recent observational studies indicate that a single modern antipseudomonal β-lactam agent to which the isolate is sensitive is as efficacious as a combination. Even in patients at greatest risk of early death from *P. aeruginosa* bacteremia (i.e., those with fever and neutropenia), empirical antipseudomonal monotherapy is deemed to be as efficacious as empirical combination therapy by the practice guidelines of the Infectious Diseases Society of America (IDSA). One firm conclusion is that monotherapy with an aminoglycoside is not optimal.

There are, of course, institutions and countries where rates of susceptibility of *P. aeruginosa* to first-line antibiotics are <80%. Thus, when a septic patient with a high probability of *P. aeruginosa* infection is encountered in such settings, empirical combination therapy should be administered until the pathogen is identified and susceptibility data become available. Thereafter, whether one or two agents should be continued remains a matter of individual preference. Recent studies suggest that extended infusions of β-

lactams such as cefepime, piperacillin/tazobactam, or meropenem may result in better outcomes of *Pseudomonas* bacteremia and possibly of *Pseudomonas* pneumonia.

TABLE 159-2

Antibiotic Treatment of Infections Due to *Pseudomonas aeruginosa* and Related Species

Infection	Antibiotics and Dosages	Other Considerations
Bacteremia		
Nonneutropenic host	Ceftazidime (2 g q8h IV) *or* cefepime (2 g q8h IV) *or* piperacillin/tazobactam (3.375 g q4h IV) *or* imipenem (500 mg q6h IV) *or* meropenem (1 g q8h IV) *or* doripenem (500 mg q8h IV) ***Optional:*** Amikacin (7.5 mg/kg q12h or 15 mg/kg q24h IV)	Add an aminoglycoside for patients in shock and in regions or hospitals where rates of resistance to the primary β-lactam agents are high. Tobramycin may be used instead of amikacin (susceptibility permitting). The duration of therapy is 7 days for nonneutropenic patients. Neutropenic patients should be treated until no longer neutropenic.
Neutropenic host	Cefepime (2 g q8h IV) *or* all the other agents above (except doripenem) in the above dosages	
Endocarditis	Antibiotic regimens as for bacteremia for 6–8 weeks	Resistance during therapy is common. Surgery is required for relapse.
Pneumonia	Drugs and dosages as for bacteremia, except that the available carbapenems should not be the sole primary drugs because of high rates of resistance during therapy.	IDSA guidelines recommend the addition of an aminoglycoside or ciprofloxacin. The duration of therapy is 7 days.
Bone infection, malignant otitis externa	Cefepime or ceftazidime at the same dosages as for bacteremia; aminoglycosides not a necessary component of therapy; ciprofloxacin (500–750 mg q12h PO) may be used	Duration of therapy varies with the drug used (e.g., 6 weeks for a β-lactam agent; at least 3 months for oral therapy except in puncture-wound osteomyelitis, for which the duration should be 2–4 weeks).
Central nervous system infection	Ceftazidime or cefepime (2 g q8h IV) *or* meropenem (1 g q8h IV)	Abscesses or other closed-space infections may require drainage. The duration of therapy is ≥2 weeks.

Infection	Antibiotics and Dosages	Other Considerations
Eye infection		
Keratitis/ulcer	Topical therapy with tobramycin/ciprofloxacin/levofloxacin eyedrops	Use maximal strengths available or compounded by pharmacy. Therapy should be administered for 2 weeks or until the resolution of eye lesions, whichever is shorter.
Endophthalmitis	Ceftazidime or cefepime as for central nervous system infection ***plus*** Topical therapy	
Urinary tract infection	Ciprofloxacin (500 mg q12h PO) *or* levofloxacin (750 mg q24h) *or* any aminoglycoside (total daily dose given once daily)	Relapse may occur if an obstruction or a foreign body is present. The duration of therapy for complicated UTI is 7–10 days (up to 2 weeks for pyelonephritis).
Multidrug-resistant *P. aeruginosa* infection	Ceftazidime/avibactam (2.5 g q8h, infused over 2 h) *or* ceftolozane/tazobactam (1.5 g q8h) *or* colistin (100 mg q12h IV for the shortest possible period to obtain a clinical response)	Higher doses of ceftolozane/tazobactam may be required for pneumonias. The colistin doses used have varied. Dosage adjustment for colistin is required in renal failure. Inhaled colistin may be added for pneumonia (100 mg q12h).
Burkholderia cepacia infection	Meropenem (1 g q8h IV) *or* TMP-SMX (1600/320 mg q12h IV) for 14 days	Resistance to both agents is increasing. Do not use them in combination because of possible antagonism.
Melioidosis (*B. pseudomallei*), glanders (*B. mallei*)	Ceftazidime (2 g q6h) *or* meropenem (1 g q8h) *or* imipenem (500 mg q6h) for 2 weeks ***followed by*** TMP-SMX (1600/320 mg q12h PO) for 3 months	

Infection	Antibiotics and Dosages	Other Considerations
Stenotrophomonas maltophilia infection	TMP-SMX (1600/320 mg q12h IV) *plus* ticarcillin/clavulanate (3.1 g q4h IV) for 14 days	Resistance to all agents is increasing. Levofloxacin or tigecycline may be alternatives, but there is little published clinical experience with these agents.

Abbreviations: IDSA, Infectious Diseases Society of America; TMP-SMX, trimethoprim-sulfamethoxazole.

Acute Pneumonia

Respiratory infections are the most common of all infections caused by *P. aeruginosa*. This organism appears first or second among the causes of ventilator-associated pneumonia (VAP). However, much debate centers on the actual role of *P. aeruginosa* in VAP. Many of the relevant data are based on cultures of sputum or endotracheal tube aspirates and may represent nonpathogenic colonization of the tracheobronchial tree, biofilms on the endotracheal tube, or simple tracheobronchitis.

Older reports of *P. aeruginosa* pneumonia described patients with an acute clinical syndrome of fever, chills, cough, and necrotizing pneumonia indistinguishable from other gram-negative bacterial pneumonias. The traditional accounts described a fulminant infection. Chest radiographs demonstrated bilateral pneumonia, often with nodular densities with or without cavities. This picture is now remarkably rare. Today, the typical patient is on a ventilator, has a slowly progressive infiltrate, and has been colonized with *P. aeruginosa* for days. While some cases may progress rapidly over 48–72 h, they are the exceptions. Nodular densities are not commonly seen. However, infiltrates may go on to necrosis. Necrotizing pneumonia has also been seen in the community (e.g., after inhalation of hot-tub water contaminated with *P. aeruginosa*). The typical patient has fever, leukocytosis, and purulent sputum, and the chest radiograph shows a new infiltrate or the expansion of a preexisting infiltrate. A sputum Gram's stain showing mainly polymorphonuclear leukocytes (PMNs) in conjunction with a culture positive for *P. aeruginosa* in this setting suggests a diagnosis of acute *P. aeruginosa* pneumonia. There is no consensus about whether an invasive procedure (e.g., bronchoalveolar lavage or protected-brush sampling of the distal airways) is superior to tracheal aspiration in obtaining samples for lung cultures in order to substantiate the occurrence of *P. aeruginosa* pneumonia and prevent antibiotic overuse.

TREATMENT

TREATMENT

Acute Pneumonia

Therapy for *P. aeruginosa* pneumonia has been unsatisfactory (Table 159-2). Reports suggest mortality rates of 40–80%, but how many of these deaths are attributable to underlying disease remains unknown. The

drugs of choice for *P. aeruginosa* pneumonia are similar to those given for bacteremia. A potent antipseudomonal β-lactam drug is the mainstay of therapy. Failure rates were high when aminoglycosides were used as single agents, possibly because of their poor penetration into the airways and their binding to airway secretions. Nonetheless, for the treatment of patients at high risk of death, some experts suggest the combination of a β-lactam agent and an antipseudomonal fluoroquinolone or aminoglycoside. As for the duration of therapy, recent IDSA/American Thoracic Society (ATS) guidelines recommend 7 days of treatment for hospital-acquired pneumonia or VAP, even when *P. aeruginosa* is the offending organism.

Chronic Respiratory Tract Infections

 *P. aeruginosa* is responsible for chronic infections of the airways associated with a number of underlying or predisposing conditions—most commonly CF **(Chap. 285)**. A state of chronic colonization beginning early in childhood is seen in some Asian populations with chronic or diffuse panbronchiolitis, a disease of unknown etiology. *P. aeruginosa* is one of the organisms that colonizes damaged bronchi in bronchiectasis, a disease secondary to multiple causes in which profound structural abnormalities of the airways result in mucus stasis.

TREATMENT

TREATMENT

Chronic Respiratory Tract Infections

Optimal management of chronic *P. aeruginosa* lung infection has not been determined. Patients respond clinically to antipseudomonal therapy, but the organism is rarely eradicated. Because eradication is unlikely, the aim of treatment for chronic infection is to quell exacerbations of inflammation. The regimens used are similar to those used for pneumonia, but an aminoglycoside is almost always added because resistance is common in chronic disease. However, it may be appropriate to use an inhaled aminoglycoside preparation in order to maximize airway drug levels.

Endovascular Infections

Infective endocarditis due to *P. aeruginosa* is a disease of IV drug users whose native valves are involved. This organism has also been reported to cause prosthetic-valve endocarditis. Sites of prior native-valve injury due to the injection of foreign material such as talc or fibers probably serve as niduses for bacterial attachment to the heart valve. The manifestations of *P. aeruginosa* endocarditis resemble those of other forms of endocarditis in IV drug users except that the disease is more indolent than *Staphylococcus aureus* endocarditis. While most disease involves the right side of the heart, left-sided involvement is not rare and multivalvular disease is common. Fever is a common manifestation, as is pulmonary involvement (due to septic emboli to the lungs). Thus, patients may also experience chest pain and hemoptysis. Involvement of the left side of the heart may lead to signs of cardiac failure, systemic emboli, and local cardiac involvement with sinus of Valsalva abscesses and conduction defects. Skin manifestations are rare in this disease, and

ecthyma gangrenosum is not seen. The diagnosis is based on positive blood cultures along with clinical signs of endocarditis.

TREATMENT

TREATMENT

Endovascular Infections

It has been customary to use synergistic antibiotic combinations in treating *P. aeruginosa* endocarditis because of the development of resistance during therapy with a single antipseudomonal β-lactam agent (Table 159-2). Which combination therapy is preferable is unclear, as all combinations have failed. Cases of *P. aeruginosa* endocarditis that relapse during or fail to respond to therapy are often caused by resistant organisms and may require surgical therapy. Other considerations for valve replacement are similar to those in other forms of endocarditis **(Chap. 123)**.

Bone and Joint Infections

P. aeruginosa is an infrequent cause of bone and joint infections. However, *Pseudomonas* bacteremia or infective endocarditis caused by the injection of contaminated illicit drugs has been documented to result in vertebral osteomyelitis and sternoclavicular joint arthritis. The clinical presentation of vertebral *P. aeruginosa* osteomyelitis is more indolent than that of staphylococcal osteomyelitis. The duration of symptoms in IV drug users with vertebral osteomyelitis due to *P. aeruginosa* varies from weeks to months. Fever is not uniformly present; when present, it tends to be low grade. There may be mild tenderness at the site of involvement. Blood cultures are usually negative unless there is concomitant endocarditis. The erythrocyte sedimentation rate (ESR) is generally elevated. Vertebral osteomyelitis due to *P. aeruginosa* has also been reported in the elderly, in whom it originates from urinary tract infections (UTIs). The infection generally involves the lumbosacral area because of a shared venous drainage (Batson's plexus) between the lumbosacral spine and the pelvis. Sternoclavicular septic arthritis due to *P. aeruginosa* is seen almost exclusively in IV drug users. This disease may occur with or without endocarditis, and a primary site of infection often is not found. Plain radiographs show joint or bone involvement. Treatment of these forms of disease is generally successful.

Pseudomonas osteomyelitis of the foot most often follows puncture wounds through sneakers and mostly affects children. The main manifestation is pain in the foot, sometimes with superficial cellulitis around the puncture wound and tenderness on deep palpation of the wound. Multiple joints or bones of the foot may be involved. Systemic symptoms are generally absent, and blood cultures are usually negative. Radiographs may or may not be abnormal, but the bone scan is usually positive, as are MRI studies. Needle aspiration usually yields a diagnosis. Prompt surgery, with exploration of the nail puncture tract and debridement of the involved bones and cartilage, is generally recommended in addition to antibiotic therapy.

Central Nervous System (CNS) Infections

CNS infections due to *P. aeruginosa* are relatively rare. Involvement of the CNS is almost always secondary to a surgical procedure or head trauma. The entity seen most often is postoperative or posttraumatic meningitis. Subdural or epidural infection occasionally results from contamination of these areas. Embolic disease arising from endocarditis in IV drug users and leading to brain abscesses has also been described. The cerebrospinal fluid (CSF) profile of *P. aeruginosa* meningitis is no different from that of pyogenic meningitis of any other etiology.

TREATMENT

TREATMENT

Central Nervous System Infections

Treatment of *Pseudomonas* meningitis is difficult; little information has been published (Table 159-2). However, the general principles involved in the treatment of meningitis apply, including the need for high doses of bactericidal antibiotics to attain high drug levels in the CSF. The agent with which there is the most published experience in *P. aeruginosa* meningitis is ceftazidime, but other antipseudomonal β-lactam drugs that reach reasonable CSF concentrations, such as cefepime, piperacillin/tazobactam, and meropenem, have also been used successfully. Other forms of *P. aeruginosa* CNS infection, such as brain abscesses and epidural and subdural empyema, generally require surgical drainage in addition to antibiotic therapy.

Eye Infections

Eye infections due to *P. aeruginosa* occur mainly as a result of direct inoculation into the tissue during trauma or surface injury by contact lenses. Keratitis and corneal ulcers are the most common types of eye disease and are often associated with contact lenses (especially the extended-wear variety). Keratitis can be slowly or rapidly progressive, but the classic description is disease progressing over 48 h to involve the entire cornea, with opacification and sometimes perforation. *P. aeruginosa* keratitis should be considered a medical emergency because of the rapidity with which it can progress to loss of sight. *P. aeruginosa* endophthalmitis secondary to bacteremia is the most devastating of *P. aeruginosa* eye infections. The disease is fulminant, with severe pain, chemosis, decreased visual acuity, anterior uveitis, vitreous involvement, and panophthalmitis.

TREATMENT

TREATMENT

Eye Infections

The usual therapy for keratitis is the administration of topical antibiotics (Table 159-2). Therapy for endophthalmitis includes the use of high-dose local and systemic antibiotics (to achieve higher drug concentrations in the eye) and vitrectomy.

Ear Infections

P. aeruginosa infections of the ears vary from mild swimmer's ear to serious life-threatening infections with neurologic sequelae. Swimmer's ear is common among children and results from infection of moist macerated skin of the external ear canal. Most cases resolve with treatment, but some patients develop chronic drainage. Swimmer's ear is managed with topical antibiotic agents (otic solutions). The most serious form of *Pseudomonas* infection involving the ear has been given various names: two of these designations, *malignant otitis externa* and *necrotizing otitis externa*, are now used for the same entity. This disease was originally described in elderly diabetic patients, in whom the majority of cases still occur. However, it has also been described in patients with AIDS and in elderly patients without underlying diabetes or immunocompromise. The usual presenting symptoms are decreased hearing and ear pain, which may be severe and lancinating. The pinna is usually painful, and the external canal may be tender. The ear canal almost always shows signs of inflammation, with granulation tissue and exudate. Tenderness anterior to the tragus may extend as far as the temporomandibular joint and mastoid process. A small minority of patients have systemic symptoms. Patients in whom the diagnosis is made late may present with cranial nerve palsies or even with cavernous venous sinus thrombosis. The ESR is invariably elevated (≥100 mm/h). The diagnosis is made on clinical grounds in severe cases; however, the "gold standard" is a positive technetium-99 bone scan in a patient with otitis externa due to *P. aeruginosa*. In diabetic patients, a positive bone scan constitutes presumptive evidence for this diagnosis and should prompt biopsy or empirical therapy.

TREATMENT

TREATMENT

Ear Infections

Given the infection of the ear cartilage, sometimes with mastoid or petrous ridge involvement, patients with malignant (necrotizing) otitis externa are treated as for osteomyelitis (Table 159-2).

Urinary Tract Infections

UTIs due to *P. aeruginosa* generally occur as a complication of a foreign body in the urinary tract, an obstruction in the genitourinary system, or urinary tract instrumentation or surgery. However, UTIs caused by *P. aeruginosa* have been described in pediatric outpatients without stones or evident obstruction.

TREATMENT

TREATMENT

Urinary Tract Infections

Most *P. aeruginosa* UTIs are considered complicated infections that must be treated longer than uncomplicated cystitis (Table 159-2). In general, a 7- to 10-day course of treatment suffices, with up to 2 weeks of therapy in cases of pyelonephritis. Urinary catheters, stents, or stones should be removed to prevent relapse, which is common and may be due not to resistance but rather to factors such as a foreign body that has been left in place or an ongoing obstruction.

Skin and Soft Tissue Infections

Besides pyoderma gangrenosum in neutropenic patients, folliculitis and other papular or vesicular lesions due to *P. aeruginosa* have been extensively described and are collectively referred to as *dermatitis*. Multiple outbreaks have been linked to whirlpools, spas, and swimming pools. To prevent such outbreaks, the growth of *P. aeruginosa* in the home and in recreational environments must be controlled by proper chlorination of water. Most cases of hot-tub folliculitis are self-limited, requiring only the avoidance of exposure to the contaminated source of water.

Toe-web infections occur especially often in the tropics, and the "green-nail syndrome" is caused by *P. aeruginosa* paronychia, which results from frequent submersion of the hands in water. In the latter entity, the green discoloration results from diffusion of pyocyanin into the nail bed. *P. aeruginosa* remains a prominent cause of burn wound infections in some parts of the world. The management of these infections is best left to specialists in burn wound care.

Infections in Febrile Neutropenic Patients

In febrile neutropenia, *P. aeruginosa* has historically been the organism against which empirical coverage is always essential. Although in Western countries these infections are now less common, their importance has not diminished because of persistently high mortality rates. In other parts of the world, *P. aeruginosa* continues to be a significant problem in febrile neutropenia, causing a larger proportion of infections in febrile neutropenic patients than any other single organism. For example, *P. aeruginosa* was responsible for 28% of documented infections in 499 febrile neutropenic patients in one study from the Indian subcontinent and for 31% of such infections in another. In a large study of infections in leukemia patients from Japan, *P. aeruginosa* was the most frequently documented cause of bacterial infection. In studies performed in North America, northern Europe, and Australia, the incidence of *P. aeruginosa* bacteremia in febrile neutropenia was quite variable. In a review of 97 reports published between 1987 and 1994, the incidence was reported to be 1–2.5% among febrile neutropenic patients given empirical therapy and 5–12% among patients with microbiologically documented infections. The most common clinical syndromes encountered were bacteremia, pneumonia, and soft tissue infections manifesting mainly as ecthyma gangrenosum.

TREATMENT

TREATMENT

Infections in Febrile Neutropenic Patients

Compared with rates three decades ago, improved rates of response to antibiotic therapy have been reported in many studies (Table 159-2). A study of 127 patients demonstrated a reduction in the mortality rate from 71 to 25% with the introduction of ceftazidime and imipenem. Because neutrophils—the normal host defenses against this organism—are absent in febrile neutropenic patients, maximal doses of

antipseudomonal β-lactam antibiotics should be used for the management of *P. aeruginosa* bacteremia in this setting.

Infections in Patients with AIDS

P. aeruginosa infections were documented in patients with AIDS before the advent of antiretroviral therapy. Since the introduction of protease inhibitors, *P. aeruginosa* infections in AIDS patients have been seen less frequently but still occur, particularly in the form of sinusitis. The clinical presentation of *Pseudomonas* infection (especially pneumonia and bacteremia) in AIDS patients is remarkable in that, although the illness may appear not to be severe, the infection may nonetheless be fatal. Patients with bacteremia may have only a low-grade fever and may present with ecthyma gangrenosum. Pneumonia, with or without bacteremia, is perhaps the most common type of *P. aeruginosa* infection. Patients with *P. aeruginosa* pneumonia exhibit the classic clinical signs and symptoms of pneumonia, such as fever, productive cough, and chest pain. The infection may be lobar or multilobar and shows no predisposition for any particular location. The most striking feature is the high frequency of cavitary disease.

TREATMENT

TREATMENT

Infections in Patients with AIDS

Therapy for any of these conditions in AIDS patients is no different from that in other patients. However, relapse is the rule unless the patient's CD4+ T cell count rises to >50/μL or suppressive antibiotic therapy is given. In attempts to achieve cures and prevent relapses, therapy tends to be more prolonged than in the case of an immunocompetent patient.

Gastrointestinal Infections

A poorly understood syndrome caused by *P. aeruginosa* has been described in the Far East and has been called *Shanghai fever* and *Pseudomonas enterocolitis*. This syndrome occurs in young children; its occurrence in adults appears to be rare. Shanghai fever manifests as severe enteric disease, sepsis with invasive disease, and complications, whereas *Pseudomonas* enterocolitis is characterized by prolonged fever with bloody or mucoid diarrhea mimicking bacterial enterocolitis. The mortality rate ranges between 23 and 89%, with ecthyma gangrenosum occurring in >50% of cases. Early recognition and treatment have led to a reduction in the mortality rate. There is an above-average occurrence of the *exoU* gene among *Pseudomonas* isolates from patients with this syndrome.

Multidrug-Resistant Infections

P. aeruginosa has a notorious propensity to develop antibiotic resistance (Table 159-2). During three decades, the impact of resistance was minimized by the rapid development of several potent antipseudomonal agents. However, the situation has changed, with the worldwide emergence of strains

carrying determinants that mediate resistance to multiple β-lactams, fluoroquinolones, and aminoglycosides. Physicians now resort to drugs such as colistin and polymyxin B, which were discarded decades ago. These alternative approaches to the management of multidrug-resistant *P. aeruginosa* infections were first used some time ago in CF patients, who receive colistin (polymyxin E) IV and by aerosol despite its renal toxicity. The clinical outcome of multidrug-resistant *P. aeruginosa* infections treated with colistin or polymyxin B is difficult to judge from case reports, especially given the many drugs used in the complicated management of these infections. Although earlier reports described marginal efficacy and serious nephrotoxicity and neurotoxicity, recent reports have been more encouraging as physicians learn how these agents should be dosed. Ceftolozane/tazobactam and ceftazidime/avibactam are welcome additions in the fight against multidrug-resistant *P. aeruginosa* strains as both agents are active against >50–80% of these strains in vitro. However, clinical data showing in vivo efficacy against multidrug-resistant strains are sparse at this time. The use of these agents will likely be restricted to certain niches because of cost constraints. A caveat is that these are β-lactam drugs, and *P. aeruginosa* has shown a remarkable ability to evade this class of antibiotics.

BURKHOLDERIA SPECIES

BURKHOLDERIA CEPACIA

The *B. cepacia* complex gained notoriety as the cause of a rapidly fatal syndrome of respiratory distress and septicemia (the “cepacia syndrome”) in CF patients. Previously, it had been recognized as an antibiotic-resistant nosocomial pathogen (then designated *Pseudomonas cepacia*) in ICU patients. Patients with chronic granulomatous disease are also predisposed to *B. cepacia* lung disease. The organism has been reclassified into nine subgroups, only some of which are common in CF. *B. cepacia* is an environmental organism that inhabits moist environments and is found in the rhizosphere. This organism possesses multiple virulence factors that may play roles in disease as well as colonizing factors that are capable of binding to lung mucus—an ability that may explain the predilection of *B. cepacia* for the lungs in CF. *B. cepacia* secretes elastase and possesses components of an injectable toxin-secretion system like that of *P. aeruginosa*; its LPS is among the most potent of all LPSs in stimulating an inflammatory response in the lungs. Inflammation may be the major cause of the lung disease seen in the cepacia syndrome. Besides infecting the lungs in CF, *B. cepacia* appears as an airway colonizer during broad-spectrum antibiotic therapy and is a cause of VAP, catheter-associated infections, and wound infections.

TREATMENT

TREATMENT

B. cepacia Infections

B. cepacia is intrinsically resistant to many antibiotics. Therefore, treatment must be tailored according to sensitivities. Trimethoprim-sulfamethoxazole (TMP-SMX), meropenem, and doxycycline are the most effective agents in vitro and may be started as first-line agents (Table 159-2). Some strains are susceptible to

third-generation ureidopenicillins, advanced cephalosporins, and fluoroquinolones, and these agents may be used against isolates known to be susceptible. Combination therapy for serious pulmonary infection (e.g., in CF) is suggested when multidrug-resistant strains are implicated; the combination of meropenem and TMP-SMX may be antagonistic, however. Resistance to all agents used has been reported during therapy.

BURKHOLDERIA PSEUDOMALLEI

 *B. pseudomallei* is the causative agent of melioidosis, a disease of humans and animals that is geographically restricted to Southeast Asia and northern Australia, with occasional cases in countries such as India and China. This organism may be isolated from individuals returning directly from these endemic regions and from military personnel who have served in endemic regions. Symptoms of this illness may develop only at a later date because of the organism's ability to cause latent infections. *B. pseudomallei* is found in soil and water. Humans and animals are infected by inoculation, inhalation, or ingestion; only rarely is the organism transmitted from person to person. Humans are not colonized without being infected. Among the pseudomonads, *B. pseudomallei* is perhaps the most virulent. Host compromise is not an essential prerequisite for disease, although many patients have common underlying medical diseases (e.g., diabetes or renal failure). *B. pseudomallei* is a facultative intracellular organism whose replication in PMNs and macrophages may be aided by the possession of a polysaccharide capsule. The organism also possesses elements of a type III secretion system that plays a role in its intracellular survival. During infection, there is a florid inflammatory response whose role in disease is unclear.

B. pseudomallei causes a wide spectrum of conditions, ranging from asymptomatic infection to abscesses, pneumonia, and disseminated disease. It is a significant cause of fatal community-acquired pneumonia and septicemia in endemic areas, with mortality rates as high as 44% reported in Thailand. Acute pulmonary infection is the most commonly diagnosed form of melioidosis. Pneumonia may be asymptomatic (with routine chest radiographs showing mainly upper-lobe infiltrates) or may present as severe necrotizing disease. *B. pseudomallei* also causes chronic pulmonary infections with systemic manifestations that mimic those of tuberculosis, including chronic cough, fever, hemoptysis, night sweats, and cavitary lung disease. Besides pneumonia, the other principal form of *B. pseudomallei* disease is skin ulceration with associated lymphangitis and regional lymphadenopathy. Spread from the lungs or skin, which is most often documented in debilitated individuals, gives rise to septicemic forms of melioidosis that carry a high mortality rate.

TREATMENT

TREATMENT

B. pseudomallei Infections

B. pseudomallei is susceptible to advanced penicillins and cephalosporins and to carbapenems (Table 159-2). Treatment is divided into two stages: an intensive 2-week phase of therapy with ceftazidime or a carbapenem followed by at least 12 weeks of oral TMP-SMX to eradicate the organism and prevent relapse.

The recognition of this bacterium as a potential agent of biologic warfare has stimulated interest in the development of a vaccine.

BURKHOLDERIA MALLEI

B. mallei causes the equine disease glanders in Africa, Asia, and South America. The organism was eradicated from Europe and North America decades ago. The last case seen in the United States occurred in 2001 in a laboratory worker; before that, *B. mallei* had last been seen in this country in 1949. In contrast to the other organisms discussed in this chapter, *B. mallei* is not an environmental organism and does not persist outside its equine hosts. Consequently, *B. mallei* infection is an occupational risk for handlers of horses, equine butchers, and veterinarians in areas of the world where it still exists. The polysaccharide capsule is a critical virulence determinant; diabetics are thought to be especially susceptible to infection by this organism. The organism is transmitted from animals to humans by inoculation into the skin, where it causes local infection with nodules and lymphadenitis. Regional lymphadenopathy is common. Respiratory secretions from infected horses are extremely infectious. Inhalation results in clinical signs of typical pneumonia but may also cause an acute febrile illness with ulceration of the trachea. The organism may disseminate from the skin or lungs to cause septicemia with signs of sepsis. The septicemic form is frequently associated with shock and a high mortality rate. The infection may also enter a chronic phase and present as disseminated abscesses. *B. mallei* infection may present as early as 1–2 days after inhalation or (in cutaneous disease) may not become evident for months.

TREATMENT

TREATMENT

B. mallei Infections

The antibiotic susceptibility pattern of *B. mallei* is similar to that of *B. pseudomallei*; in addition, the organism is susceptible to the macrolides azithromycin and clarithromycin. *B. mallei* infection should be treated with the same drugs and for the same duration as melioidosis.

STENOTROPHOMONAS MALTOPHILIA

S. maltophilia is the only potential human pathogen among a genus of ubiquitous organisms found in the rhizosphere (i.e., the soil that surrounds the roots of plants). The organism is an opportunist that is acquired from the environment but is even more limited than *P. aeruginosa* in its ability to colonize patients or cause infections. Immunocompromise is not sufficient to permit these events; rather, major perturbations of the human flora are usually necessary for the establishment of *S. maltophilia*. Accordingly, most cases of human infection occur in the setting of very broad-spectrum antibiotic therapy with agents such as advanced cephalosporins and carbapenems, which eradicate the normal flora and other pathogens. The remarkable ability of *S. maltophilia* to resist virtually all classes of antibiotics is attributable to the possession of antibiotic efflux pumps and of two β-lactamases (L1 and L2) that mediate β-lactam resistance, including that

to carbapenems. It is fortunate that the virulence of *S. maltophilia* appears to be limited. Although a serine protease is present in some strains, virulence is probably a result of the host's inflammatory response to components of the organism such as LPS and flagellin. *S. maltophilia* is most commonly found in the respiratory tract of ventilated patients, where the distinction between its roles as a colonizer and as a pathogen is often difficult to make. However, *S. maltophilia* does cause pneumonia and bacteremia in such patients, and these infections have led to septic shock. Also common is central venous line–associated infection (with or without bacteremia), which has been reported most often in patients with cancer. *S. maltophilia* is a rare cause of ecthyma gangrenosum in neutropenic patients. It has been isolated from ~5% of CF patients but is not believed to be a significant pathogen in this setting.

TREATMENT

TREATMENT

S. maltophilia Infections

The intrinsic resistance of *S. maltophilia* to most antibiotics renders infection difficult to treat. The antibiotics to which it is most often (although not uniformly) susceptible are TMP-SMX, ticarcillin/clavulanate, levofloxacin, and tigecycline (Table 159-2). Consequently, a combination of TMP-SMX and ticarcillin/clavulanate is recommended for initial therapy pending susceptibility testing. Catheters must be removed in the treatment of bacteremia. The treatment of VAP due to *S. maltophilia* is much more difficult than that of bacteremia, with frequent development of resistance during therapy.

FURTHER READING

Bauer KA et al: Extended-infusion cefepime reduces mortality in patients with *Pseudomonas aeruginosa* infections. Antimicrob Agents Chemother 57:2907, 2013.

Bowers DR et al: Outcomes of appropriate empiric combination versus monotherapy for *Pseudomonas aeruginosa* bacteremia. Antimicrob Agents Chemother 157:1270, 2013.

Brooke JS: *Stenotrophomonas maltophilia*: An emerging global opportunistic pathogen. Clin Microbiol Rev 25:2, 2012.

Cattaneo C et al: *P. aeruginosa* bloodstream infections among hematological patients: An old or new question? Ann Hematol 91:1299, 2012.

Chuang C-H et al: Shanghai fever: A distinct *Pseudomonas aeruginosa* enteric disease. Gut 63:736, 2014.

Kalil AC et al: Executive summary: Management of adults with hospital-acquired and ventilator-associated pneumonia: 2016 clinical practice guidelines by the Infectious Diseases Society of America and the American Thoracic Society. Clin Infect Dis 63:575, 2016.

Peña C et al: Influence of virulence genotype and resistance profile in the mortality of *Pseudomonas aeruginosa* bloodstream infections. Clin Infect Dis 60:539, 2015.

Van Zandt KE et al: Glanders: An overview of infections in humans. Orphanet J Rare Dis 8:131, 2013.

Chapter 160: Salmonellosis

David A. Pegues; Samuel I. Miller

INTRODUCTION

Bacteria of the genus *Salmonella* are highly adapted for growth in both humans and animals and cause a wide spectrum of diseases. The growth of serotypes *Salmonella* Typhi and *Salmonella* Paratyphi is restricted to human hosts, in whom these organisms cause enteric (typhoid) fever. The remaining serotypes (nontyphoidal *Salmonella*, or NTS) can colonize the gastrointestinal tracts of a broad range of animals, including mammals, reptiles, birds, and insects. More than 200 serotypes of *Salmonella* are pathogenic to humans, in whom they often cause gastroenteritis and can be associated with localized infections and/or bacteremia.

ETIOLOGY

This large genus of gram-negative bacilli within the family Enterobacteriaceae consists of two species: *Salmonella enterica*, which contains six subspecies, and *Salmonella bongori*. *S. enterica* subspecies I includes almost all the serotypes pathogenic for humans. Members of the seven *Salmonella* subspecies are classified into >2500 serotypes (serovars); for simplicity, *Salmonella* serotypes (most of which are named for the city where they were identified) are often used as the species designation. For example, the full taxonomic designation *S. enterica* subspecies *enterica* serotype Typhimurium can be shortened to *Salmonella* serotype Typhimurium or simply *S.* Typhimurium. Serotyping is based on the somatic O antigen (lipopolysaccharide cell-wall components), the surface Vi antigen (restricted to *S.* Typhi and *S.* Paratyphi C), and the flagellar H antigen.

Salmonellae are gram-negative, non-spore-forming, facultatively anaerobic bacilli that measure 2–3 μm by 0.4–0.6 μm. The initial identification of salmonellae in the clinical microbiology laboratory is based on growth characteristics. Salmonellae, like other Enterobacteriaceae, produce acid on glucose fermentation, reduce nitrates, and do not produce cytochrome oxidase. In addition, all salmonellae except *Salmonella* Gallinarum-Pullorum are motile by means of peritrichous flagella, and all but *S.* Typhi produce gas (H_2S) on sugar fermentation. Notably, only 1% of clinical isolates ferment lactose; a high level of suspicion must be maintained to detect these rare clinical lactose-fermenting isolates.

Although serotyping of all surface antigens can be used for formal identification, most laboratories perform a few simple agglutination reactions that define specific O-antigen serogroups, designated A, B, C_1, C_2, D, and

E. Strains in these six serogroups cause ~99% of *Salmonella* infections in humans and other warm-blooded animals. Molecular typing methods, including pulsed-field gel electrophoresis, multiple-locus variable-number tandem repeat analysis, and whole-genome sequencing, are used in epidemiologic investigations to differentiate *Salmonella* strains of a common serotype.

PATHOGENESIS

All *Salmonella* infections begin with ingestion of organisms, most commonly in contaminated food or water. The infectious dose ranges from 200 colony-forming units (CFU) to 10^6 CFU, and the ingested dose is an important determinant of incubation period and disease severity. Conditions that decrease either stomach acidity (an age of <1 year, acid suppression therapy, or achlorhydric disease) or intestinal integrity (inflammatory bowel disease, cytotoxic chemotherapy, prior gastrointestinal surgery, or alteration of the intestinal microbiome by antibiotic administration) increase susceptibility to *Salmonella* infection.

Once *S.* Typhi and *S.* Paratyphi reach the small intestine, they penetrate the mucus layer of the gut and traverse the intestinal layer through phagocytic microfold (M) cells that reside within Peyer's patches. Salmonellae can trigger the formation of membrane ruffles in normally nonphagocytic epithelial cells. These ruffles reach out and enclose adherent bacteria within large vesicles by *bacterium-mediated endocytosis*. This process is dependent on the direct delivery of *Salmonella* proteins into the cytoplasm of epithelial cells by the specialized bacterial type III secretion system. These bacterial proteins mediate alterations in the actin cytoskeleton that are required for *Salmonella* uptake.

After crossing the epithelial layer of the small intestine, *S.* Typhi and *S.* Paratyphi, which cause enteric (typhoid) fever, are phagocytosed by macrophages. These salmonellae survive the antimicrobial environment of the macrophage by sensing environmental signals that trigger alterations in regulatory systems of the phagocytosed bacteria. For example, PhoP/PhoQ (the best-characterized regulatory system) triggers the alteration of the outer membrane by increasing the synthesis and transport of different outer-membrane proteins, lipopolysaccharides, and glycerophospholipids, so that the altered bacterial surface can resist microbicidal activities and potentially alter host cell signaling. In addition, salmonellae encode a second type III secretion system that directly delivers bacterial proteins across the phagosome membrane into the macrophage cytoplasm. This secretion system functions to remodel the *Salmonella*-containing vacuole, promoting bacterial survival and replication.

Once phagocytosed, typhoidal salmonellae disseminate throughout the body in macrophages via the lymphatics and colonize reticuloendothelial tissues (liver, spleen, lymph nodes, and bone marrow). Patients have relatively few or no signs and symptoms during this initial incubation stage. Signs and symptoms, including fever and abdominal pain, probably result from secretion of cytokines by macrophages and epithelial cells in response to bacterial products that are recognized by innate immune receptors when a critical number of organisms have replicated. Over time, the development of hepatosplenomegaly is likely to be related to the recruitment of mononuclear cells and the development of a specific acquired cell-mediated immune response to *S.* Typhi colonization. The recruitment of additional mononuclear cells and

lymphocytes to Peyer's patches during the several weeks after initial colonization/infection can result in marked enlargement and necrosis of the Peyer's patches, which may be mediated by bacterial products that promote cell death as well as the inflammatory response. In the case of *S.* Typhi, many strains produce a toxin, which probably contributes to systemic symptoms as well as the unusual neuropsychiatric states that can be seen in severe typhoidal illness.

In contrast to enteric fever, which is characterized by an infiltration of mononuclear cells into the small-bowel mucosa, NTS gastroenteritis is characterized by massive polymorphonuclear leukocyte infiltration into both the large- and small-bowel mucosa. This response appears to depend on the induction of interleukin 8, a strong neutrophil chemotactic factor, which is secreted by intestinal cells as a result of nontyphoidal *Salmonella* colonization and translocation of bacterial proteins into host cell cytoplasm. The degranulation and release of toxic substances by neutrophils may result in damage to the intestinal mucosa, causing the inflammatory diarrhea observed with nontyphoidal gastroenteritis. An additional important factor in the persistence of NTS in the intestinal tract and the organism's capacity to compete with endogenous flora is the ability to utilize the sulfur-containing compound tetrathionate for metabolism in a microaerophilic environment. In the presence of intestinal inflammation, tetrathionate is generated from thiosulfate produced by epithelial cells through inflammatory cell production of reactive oxygen species.

ENTERIC (TYPHOID) FEVER

Enteric (typhoid) fever is a systemic disease characterized by fever and abdominal pain and caused by dissemination of *S.* Typhi or *S.* Paratyphi. The disease was initially called *typhoid fever* because of its clinical similarity to typhus. In the early 1800s, typhoid fever was clearly defined pathologically as a unique illness on the basis of its association with enlarged Peyer's patches and mesenteric lymph nodes. In 1869, given the anatomic site of infection, the term *enteric fever* was proposed as an alternative designation to distinguish typhoid fever from typhus. However, to this day, the two designations are used interchangeably.

EPIDEMIOLOGY

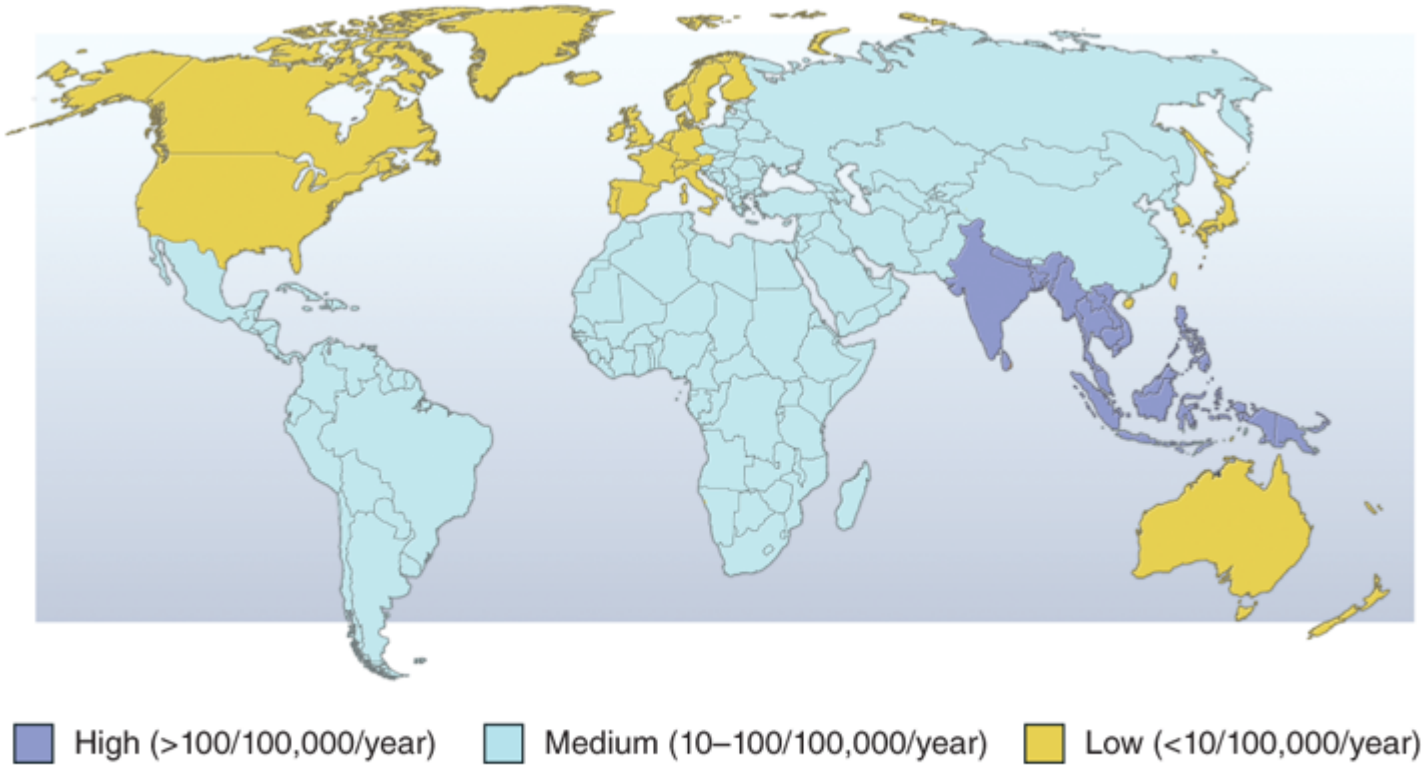
In contrast to other *Salmonella* serotypes, the etiologic agents of enteric fever—*S.* Typhi and *S.* Paratyphi serotypes A, B, and C—have no known hosts other than humans. Most commonly, food-borne or waterborne transmission results from fecal contamination by ill or asymptomatic chronic carriers. Sexual transmission between male partners has been described. Health care workers occasionally acquire enteric fever after exposure to infected patients or during processing of clinical specimens and cultures.

With improvements in food handling and water/sewage treatment, enteric fever has become rare in developed nations. Worldwide, however, there are an estimated 21–27 million cases of enteric fever, with 200,000–600,000 deaths annually. The annual incidence is highest (>100 cases/100,000 population) in South-Central and Southeast Asia; medium (10–100 cases/100,000) in the rest of Asia, Africa, Latin America, and Oceania (excluding Australia and New Zealand); and low in other parts of the world **(Fig. 160-1)**. A high incidence of enteric fever correlates with poor sanitation and lack of access to clean drinking water. In

endemic regions, enteric fever is more common in urban than rural areas and among young children and adolescents than among other age groups. Risk factors include contaminated water or ice, flooding, food and drinks purchased from street vendors, raw fruits and vegetables grown in fields fertilized with sewage, ill household contacts, lack of hand washing and toilet access, and evidence of prior *Helicobacter pylori* infection (an association probably related to chronically reduced gastric acidity). It is estimated that there is one case of paratyphoid fever for every four cases of typhoid fever, but the incidence of infection associated with *S.* Paratyphi A appears to be increasing, especially in India; this increase may be a result of vaccination for *S.* Typhi.

FIGURE 160-1

Annual incidence of typhoid fever per 100,000 population. *(Adapted from JA Crump et al: The global burden of typhoid fever. Bull World Health Organ 82:346, 2004.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Multidrug-resistant (MDR) strains of *S.* Typhi emerged in the 1980s in China and Southeast Asia and have since disseminated widely. These strains contain plasmids encoding resistance to chloramphenicol, ampicillin, and trimethoprim—antibiotics long used to treat enteric fever. With the increased use of fluoroquinolones to treat MDR enteric fever in the 1990s, MDR strains of *S.* Typhi and *S.* Paratyphi with decreased susceptibility to ciprofloxacin (DSC; minimal inhibitory concentration [MIC], 0.125–0.5 μg/mL) or ciprofloxacin resistance (MIC, ≥1 μg/mL) have emerged on the Indian subcontinent and have spread with human migration to southern Asia and now to eastern and southern Africa. These strains represent clone H58, which increasingly has been associated with clinical failure of quinolone treatment. Testing of isolates for resistance to the first-generation quinolone nalidixic acid detects many but not all strains with reduced

susceptibility to ciprofloxacin and is no longer recommended. Strains of *S.* Typhi and *S.* Paratyphi producing extended-spectrum β-lactamases have emerged, primarily in India and Nepal.

Approximately 300 cases of typhoid and 150 cases of paratyphoid fever are reported annually in the United States. Of 3499 cases of *S.* Typhi–associated enteric fever reported to the Centers for Disease Control and Prevention in 1999–2010, 82% were associated with recent international travel, most commonly to India, Pakistan, and Bangladesh, and occurred predominantly in young to middle-aged adults. Only 6% of travelers diagnosed with enteric fever had received *S.* Typhi vaccine. Overall, 15% of recent *S.* Typhi isolates in the United States were resistant to ampicillin, chloramphenicol, and trimethoprim-sulfamethoxazole (TMP-SMX), whereas 60% of isolates exhibited DSC. Infection with DSC *S.* Typhi was associated with travel to the Indian subcontinent. During this period, 18% of reported cases of enteric fever in the United States were domestically acquired, and these cases were less often due to MDR or DSC strains than were travel-associated cases. Most cases of domestically acquired enteric fever are sporadic, but outbreaks linked to contaminated food products and previously unrecognized chronic carriers continue to occur.

CLINICAL COURSE

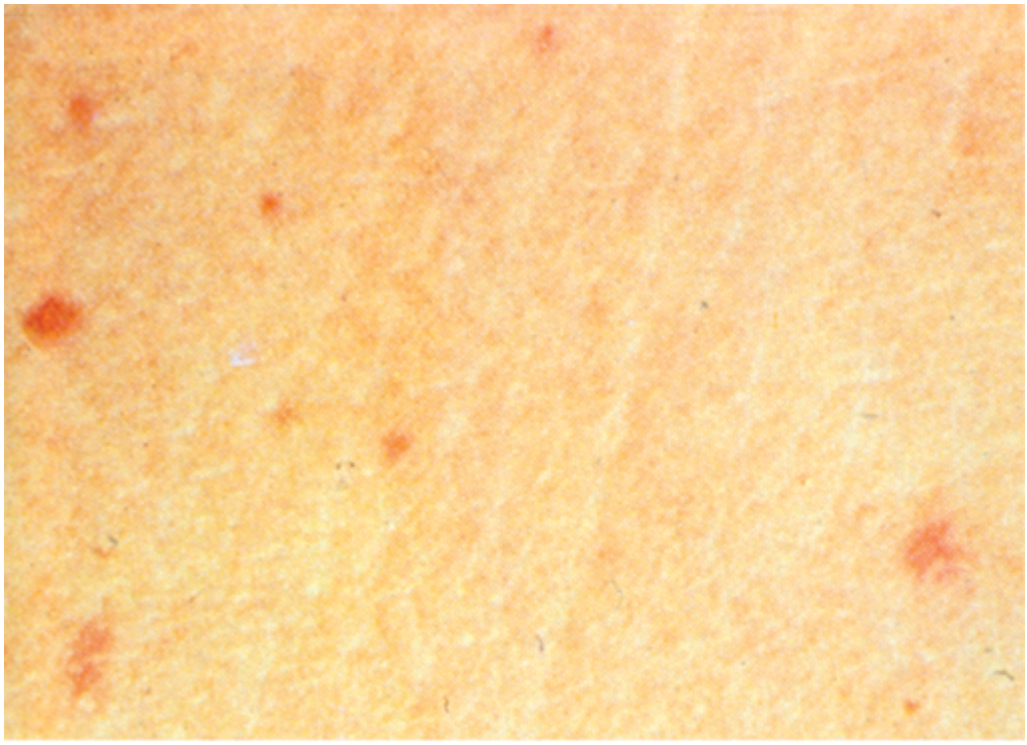
Enteric fever is a misnomer, in that the hallmark features of this disease—fever and abdominal pain—are variable. While fever is documented at presentation in >75% of cases, abdominal pain is reported in only 30–40%. Thus, a high index of suspicion for this potentially fatal systemic illness is necessary when a person presents with fever and a history of recent travel to a developing country.

The incubation period for *S.* Typhi averages 10–14 days but ranges from 5 to 21 days, depending on the inoculum size and the host's health and immune status. The most prominent symptom is prolonged fever (38.8°–40.5°C; 101.8°–104.9°F), which can continue for up to 4 weeks if untreated. *S.* Paratyphi A is thought to cause milder disease than *S.* Typhi, with predominantly gastrointestinal symptoms. However, a prospective study of 669 consecutive cases of enteric fever in Kathmandu, Nepal, found that the infections caused by these organisms were clinically indistinguishable. In this series, symptoms reported on initial medical evaluation included headache (80%), chills (35–45%), cough (30%), sweating (20–25%), myalgias (20%), malaise (10%), and arthralgia (2–4%). Gastrointestinal manifestations included anorexia (55%), abdominal pain (30–40%), nausea (18–24%), vomiting (18%), and diarrhea (22–28%) more commonly than constipation (13–16%). Physical findings included coated tongue (51–56%), splenomegaly (5–6%), and abdominal tenderness (4–5%).

Early physical findings of enteric fever include rash ("rose spots"; 30%), hepatosplenomegaly (3–6%), epistaxis, and relative bradycardia at the peak of high fever (<50%). Rose spots **(Fig. 160-2; see also Fig. A1-9)** make up a faint, salmon-colored, blanching, maculopapular rash located primarily on the trunk and chest. The rash is evident in ~30% of patients at the end of the first week and resolves without a trace after 2–5 days. Patients can have two or three crops of lesions, and *Salmonella* can be cultured from punch biopsies of these lesions. The faintness of the rash makes it difficult to detect in highly pigmented patients.

FIGURE 160-2

"Rose spots," the rash of enteric fever due to *Salmonella* Typhi or *Salmonella* Paratyphi.

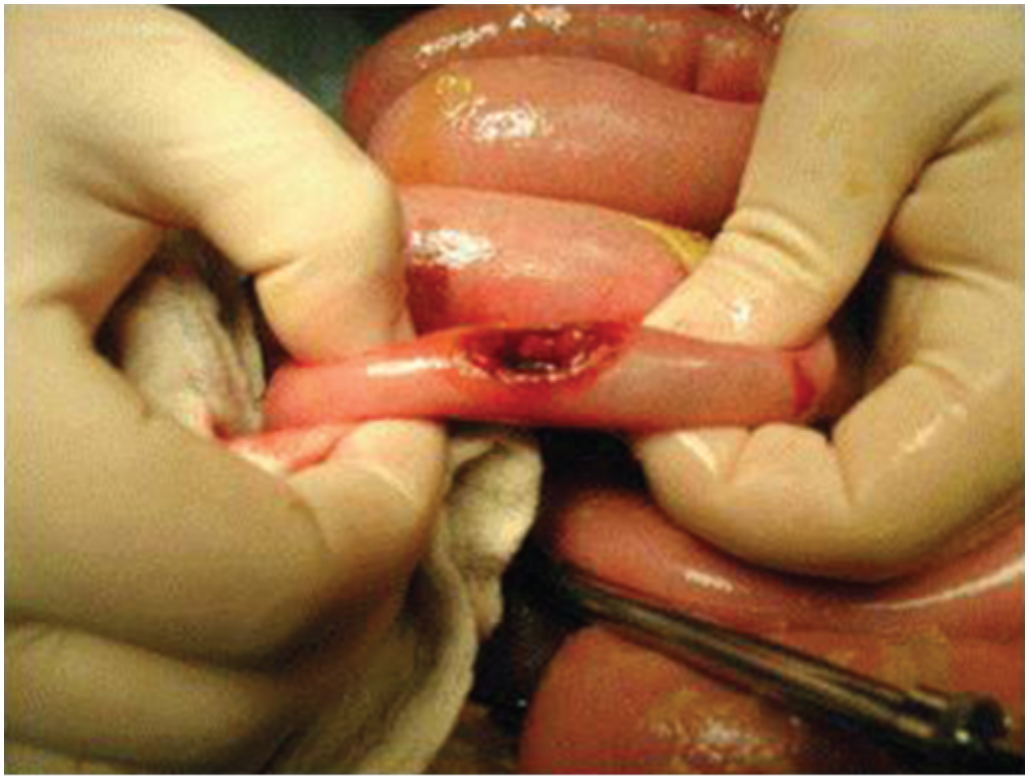


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The development of severe disease (which occurs in ~10–15% of patients) depends on host factors (host genetics, immunosuppression, acid suppression therapy, previous exposure, and vaccination), strain virulence and inoculum, and choice of antibiotic therapy. Gastrointestinal bleeding (10–20%) and intestinal perforation (1–3%) most commonly occur in the third and fourth weeks of illness and result from hyperplasia, ulceration, and necrosis of the ileocecal Peyer's patches at the initial site of *Salmonella* infiltration **(Fig. 160-3)**. Both complications are life-threatening and require immediate fluid resuscitation and surgical intervention, with broadened antibiotic coverage for polymicrobial peritonitis **(Chap. 127)** and treatment of gastrointestinal hemorrhages, including bowel resection. Neurologic manifestations occur in 2–40% of patients and include meningitis, Guillain-Barré syndrome, neuritis, and neuropsychiatric symptoms (described as "muttering delirium" or "coma vigil"), with picking at bedclothes or imaginary objects.

FIGURE 160-3

Typical ileal perforation associated with *Salmonella* Typhi infection. *(From JM Saxe, R Cropsey: Is operative management effective in treatment of perforated typhoid? Am J Surg 189:342, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Rare complications whose incidences are reduced by prompt antibiotic treatment include disseminated intravascular coagulation, hematophagocytic syndrome, pancreatitis, hepatic and splenic abscesses and granulomas, endocarditis, pericarditis, myocarditis, orchitis, hepatitis, glomerulonephritis, pyelonephritis and hemolytic-uremic syndrome, severe pneumonia, arthritis, osteomyelitis, endophthalmitis, and parotitis. Up to 10% of patients develop mild relapse, usually within 2–3 weeks of fever resolution and in association with the same strain type and susceptibility profile.

Up to 10% of untreated patients with typhoid fever excrete *S.* Typhi in the feces for up to 3 months, and 2–5% develop chronic asymptomatic carriage, shedding *S.* Typhi in either urine or stool for >1 year. Chronic carriage is more common among women, infants, and persons who have biliary abnormalities or concurrent bladder infection with *Schistosoma haematobium*. *S.* Typhi and other salmonellae are adapted to survive in the gallbladder environment by forming biofilms on gallstones and invading gallbladder epithelial cells. Chronic carriage is associated with an increased risk of gallbladder cancer, which is much more common in locales where *S.* Typhi is common, such as the Indian subcontinent.

DIAGNOSIS

Because the clinical presentation of enteric fever is relatively nonspecific, the diagnosis needs to be considered in any febrile traveler returning from a developing region, especially the Indian subcontinent, the Philippines, or Latin America. Other diagnoses that should be considered in these travelers include malaria, hepatitis, bacterial enteritis, dengue fever, rickettsial infections, leptospirosis, amebic liver abscesses, and acute HIV infection (Chap. 119). Other than a positive culture, no specific laboratory test is diagnostic for enteric fever. In 15–25% of cases, leukopenia and neutropenia are detectable. Leukocytosis is more common

among children, during the first 10 days of illness, and in cases complicated by intestinal perforation or secondary infection. Other nonspecific laboratory findings include moderately elevated values in liver function tests and muscle enzyme levels.

The definitive diagnosis of enteric fever requires the isolation of *S.* Typhi or *S.* Paratyphi from blood, bone marrow, other sterile sites, rose spots, stool, or intestinal secretions. The sensitivity of blood culture is only 40–80%, probably because of high rates of antibiotic use in endemic areas and the small number of *S.* Typhi organisms (i.e., <15/mL) typically present in the blood. Because almost all *S.* Typhi organisms in blood are associated with the mononuclear cell/platelet fraction, centrifugation of blood and culture of the buffy coat can substantially reduce the time to isolation of the organism but do not increase sensitivity.

Bone marrow culture is >80% sensitive, and, unlike that of blood culture, its yield is not reduced by up to 5 days of prior antibiotic therapy. Culture of intestinal secretions (best obtained by a noninvasive duodenal string test) can be positive despite a negative bone marrow culture. If blood, bone marrow, and intestinal secretions are all cultured, the yield is >90%. Stool cultures, although negative in 60–70% of cases during the first week, can become positive during the third week of infection in untreated patients.

The classic Widal serologic test for “febrile agglutinins” is simple and rapid but has limited sensitivity and specificity, especially in endemic regions. Rapid point-of-care tests that detect antibodies to outer-membrane proteins or to Vi or O:9 antigen are available for detection of *S.* Typhi; they are moderately sensitive and specific, but their cost and accuracy have limited their routine use in developing countries. More sensitive and highly specific nucleic acid amplification tests have been developed to detect *S.* Typhi and *S.* Paratyphi in blood, but they do not detect antibiotic resistance and remain impractical in many areas where enteric fever is endemic.

TREATMENT

TREATMENT

Enteric (Typhoid) Fever

Prompt administration of appropriate antibiotic therapy prevents severe complications of enteric fever and results in a case–fatality rate of <1%. The initial choice of antibiotics depends on the susceptibility of the *S.* Typhi and *S.* Paratyphi strains in the area of residence or travel **(Table 160-1)**. For treatment of drug-susceptible typhoid fever, fluoroquinolones are the most effective class of agents, with cure rates of ~98% and relapse and fecal carriage rates of <2%. Experience is most extensive with ciprofloxacin. Short-course ofloxacin therapy is similarly successful against infection caused by quinolone-susceptible strains. However, the high prevalence of DSC and ciprofloxacin-resistant *S.* Typhi and *S.* Paratyphi on the Indian subcontinent, in Nepal, and in some locales in Africa suggests that fluoroquinolones should no longer be used for empirical treatment of enteric fever in these regions. Patients infected with DSC strains of *S.* Typhi or *S.* Paratyphi should be treated with ceftriaxone or azithromycin. Alternatively, high-dose ciprofloxacin (750 mg twice daily

for 10–14 days) can be used to treat DSC strains but should not be used to treat ciprofloxacin-resistant enteric fever (MIC, ≥1 µg/mL) because of high failure rates.

Ceftriaxone, cefotaxime, and (oral) cefixime are effective for treatment of MDR enteric fever, including that caused by DSC and fluoroquinolone-resistant strains. These agents clear fever in ~1 week, with failure rates of ~5–10%, fecal carriage rates of <3%, and relapse rates of 3–6%. Oral azithromycin results in defervescence in 4–6 days, with rates of relapse and convalescent stool carriage of <3%. Against DSC strains, azithromycin is associated with lower rates of treatment failure and shorter durations of hospitalization than are fluoroquinolones. Despite efficient in vitro killing of *Salmonella*, first- and second-generation cephalosporins as well as aminoglycosides are ineffective in the treatment of clinical infections.

Most patients with uncomplicated enteric fever can be managed at home with oral antibiotics and antipyretics. Patients with persistent vomiting, diarrhea, and/or abdominal distension should be hospitalized and given supportive therapy as well as a parenteral third-generation cephalosporin or a fluoroquinolone, depending on the susceptibility profile. Therapy should be administered for at least 10 days or for 5 days after fever resolution.

 In a randomized, prospective, double-blind study of critically ill patients with enteric fever (i.e., those with shock and obtundation) in Indonesia in the early 1980s, the administration of dexamethasone (an initial dose of 3 mg/kg followed by eight doses of 1 mg/kg every 6 h) with chloramphenicol was associated with a substantially lower mortality rate than was treatment with chloramphenicol alone (10% vs 55%). Although this study has not been repeated in the “post-chloramphenicol era,” severe enteric fever remains one of the few indications for glucocorticoid treatment of an acute bacterial infection.

The 2–5% of patients who develop chronic carriage of *Salmonella* can be treated for 4 weeks with oral ciprofloxacin or other fluoroquinolones, with an eradication rate of ~80%. Treatment with oral amoxicillin or TMP-SMX is no longer recommended because of lower eradication rates than for fluoroquinolones and the high prevalence of MDR strains. In cases of anatomic abnormality (e.g., biliary or kidney stones), eradication often requires both antibiotic therapy and surgical correction.

TABLE 160-1

Antibiotic Therapy for Enteric Fever in Adults

Indication	Agent	Dosage (Route)	Duration, Days
Empirical Treatment			
	Ceftriaxone[a]	2 g/d (IV)	10–14
	Azithromycin[b]	1 g/d (PO)	5
Fully Susceptible			
Optimal treatment	Ciprofloxacin[c]	500 mg bid (PO) or 400 mg q12h (IV)	5–7
	Azithromycin	1 g/d (PO)	5
Alternative treatment	Amoxicillin	1 g tid (PO) or 2 g q6h (IV)	14
	Chloramphenicol	25 mg/kg tid (PO or IV)	14–21
	Trimethoprim-sulfamethoxazole	160/800 mg bid (PO)	7–14
Multidrug-Resistant			
Optimal treatment	Ceftriaxone[a]	2 g/d (IV)	10–14
	Azithromycin	1 g/d (PO)	5
Alternative treatment	Ciprofloxacin	500 mg bid (PO) or 400 mg q12h (IV)	5–14
Quinolone-Resistant			
Optimal treatment	Ceftriaxone	2 g/d (IV)	10–14
	Azithromycin	1 g/d (PO)	5

Indication	Agent	Dosage (Route)	Duration, Days
Alternative treatment	High-dose ciprofloxacin[d]	750 mg bid (PO) or 400 mg q8h (IV)	10–14

[a]Or another third-generation cephalosporin (e.g., cefotaxime, 2 g q8h IV; or cefixime, 400 mg bid PO). [b]Or 1 g on day 1 followed by 500 mg/d PO for 6 days. [c]Or ofloxacin, 400 mg bid PO for 2–5 days. [d]High-dose ciprofloxacin is an alternative for strains with reduced susceptibility to ciprofloxacin (MIC, 0.125–0.5 µg/mL) but should not be used for fully resistant strains (MIC, ≥1 µg/mL).

PREVENTION AND CONTROL


Theoretically, it is possible to eliminate the salmonellae that cause enteric fever because they survive only in human hosts and are spread by contaminated food and water. However, given the high prevalence of the disease in developing countries that lack adequate sewage disposal and water treatment, this goal is currently unrealistic. Thus, travelers to developing countries should be advised to monitor their food and water intake carefully and to strongly consider immunization against *S.* Typhi.

Two typhoid vaccines are commercially available: (1) Ty21a, an oral live attenuated *S.* Typhi vaccine (given on days 1, 3, 5, and 7, with revaccination with a full 4-dose series every 5 years); and (2) Vi CPS, a parenteral vaccine consisting of purified Vi polysaccharide from the bacterial capsule (given in a single dose, with a booster every 2 years). The old parenteral whole-cell typhoid/paratyphoid A and B vaccine is no longer licensed, largely because of significant side effects, especially fever. An acetone-killed whole-cell vaccine is available only for use by the U.S. military. The minimal age for vaccination is 6 years for Ty21a and 2 years for Vi CPS. In a recent meta-analysis of vaccines for preventing typhoid fever in populations in endemic areas, the cumulative efficacy was 58% for Ty21a at 2 years and 55% for Vi CPS at 3 years. Although data on typhoid vaccines in travelers are limited, recent evidence suggests moderate efficacy (80%) in U.S. travelers. Currently, there is no licensed vaccine for paratyphoid fever.

Vi CPS typhoid vaccine is poorly immunogenic in children <5 years of age because of T cell–independent properties. In the more recently developed Vi-rEPA vaccine, Vi is bound to a nontoxic recombinant protein that is identical to *Pseudomonas aeruginosa* exotoxin A. In 2- to 4-year-olds, two injections of Vi-rEPA induced stronger T cell responses and higher levels of serum IgG antibody to Vi than did Vi CPS in 5- to 14-year-olds. In a two-dose trial in 2- to 5-year-old children in Vietnam, Vi-rEPA provided 91% efficacy at 27 months and 89% efficacy at 46 months and was very well tolerated. This vaccine is not yet commercially available in the United States. Efforts to improve the immunogenicity and reduce the number of doses of live attenuated oral vaccines are ongoing.

Typhoid vaccine is not required for international travel, but it is recommended for travelers to areas where there is a moderate to high risk of exposure to *S.* Typhi, especially those who are traveling to southern Asia and other developing regions of Asia, Africa, the Caribbean, and Central and South America and who will be exposed to potentially contaminated food and drink. Typhoid vaccine should be considered even for persons planning <2 weeks of travel to high-risk areas. In addition, laboratory workers who deal with *S.* Typhi and household contacts of known *S.* Typhi carriers should be vaccinated. Because the protective efficacy of vaccine can be overcome by the high inocula that are commonly encountered in food-borne exposures, immunization is an adjunct and not a substitute for the avoidance of high-risk foods and beverages. Immunization is not recommended for adults residing in typhoid-endemic areas or for the management of persons who may have been exposed in a common-source outbreak.

Enteric fever is a notifiable disease in the United States. Individual health departments have their own guidelines for allowing ill or colonized food handlers or health care workers to return to their jobs. The reporting system enables public health departments to identify potential source patients and to treat chronic carriers in order to prevent further outbreaks. In addition, because 1–4% of patients with *S.* Typhi infection become chronic carriers, it is important to monitor patients (especially child-care providers and food handlers) for chronic carriage and to treat this condition if indicated.

NONTYPHOIDAL SALMONELLOSIS

EPIDEMIOLOGY


Worldwide, NTS causes ~93 million enteric infections and 155,000 deaths annually. In the United States, NTS causes ~12 million illnesses annually, and the incidence has remained relatively unchanged during the past two decades. In 2014, the incidence of NTS infection in the United States was 14.45 cases per 100,000 persons—the highest rate among the 10 food-borne enteric pathogens under active surveillance. Four serotypes accounted for more than half of U.S. infections in 2014: Enteritidis (21%), Typhimurium (12%), Newport (11%), and Javiana (10%).

The incidence of nontyphoidal salmonellosis is highest during the rainy season in tropical climates and during the warmer months in temperate climates—a pattern coinciding with the peak in food-borne outbreaks. Rates of morbidity and mortality associated with NTS are highest among the elderly, infants, and immunocompromised individuals, including those with hemoglobinopathies, HIV infection, or infections that cause blockade of the reticuloendothelial system (e.g., bartonellosis, malaria, schistosomiasis, histoplasmosis). NTS account for a significant majority of illnesses and hospitalizations associated with U.S. multistate food-borne outbreaks.

Unlike *S.* Typhi and *S.* Paratyphi, whose only reservoir is humans, NTS can be acquired from multiple animal reservoirs. Transmission is most commonly associated with food products of animal origin (especially eggs, poultry, undercooked ground meat, and dairy products), fresh produce contaminated with animal waste, and contact with animals or their environments.

S. Enteritidis infection associated with chicken eggs emerged as a major cause of food-borne disease during the 1980s and 1990s. *S.* Enteritidis infection of the ovaries and upper oviduct tissue of hens results in contamination of egg contents before shell deposition. Infection is spread to egg-laying hens from breeding flocks and through contact with rodents and manure. The number of *S.* Enteritidis outbreaks and the proportion attributable to egg-containing foods have continued to decline since the mid-1990s; these declines have coincided with interventions in the egg-producing and food service industries. Despite these control efforts, outbreaks of *S.* Enteritidis infection associated with shell eggs continue to occur. In 2010, a national outbreak of *S.* Enteritidis infection resulted in more than 1900 reported illnesses and the recall of 500 million eggs. Transmission via contaminated eggs can be prevented by cooking eggs until the yolk is solidified and pasteurizing egg products. However, a 2010 survey found that 44% of respondents reported consuming undercooked or runny eggs in the previous 12 months, suggesting that adherence to safe food-preparation practices has lagged in the home.

In the last 10 years in Europe (and to a lesser degree in the United States), nontyphoidal strains related to *S.* Typhimurium have emerged and are termed Serotype 4,5,12:i:-. These strains are genetically identical to serotype Typhimurium strains, but a mutation in the *fljA* and *fljB* genes encoding one of the flagellar antigens results in a different serotype that lacks flagellar phase 2 antigen. In Europe, these strains have been associated with pigs and their products, and the reason for their emergence is unknown.

Centralization of food processing and widespread food distribution have contributed to the increased incidence of NTS in developed countries. Manufactured foods to which recent multistate *Salmonella* outbreaks have been traced include peanut butter; milk products, including infant formula; and various processed foods, including packaged breakfast cereal, salsa, frozen prepared meals, and snack foods. Large outbreaks also have been linked to fresh produce, including alfalfa sprouts, nuts and seeds, cantaloupe, mangoes, papayas, and tomatoes; these items become contaminated by manure or water at a single site and then are widely distributed.

An estimated 6% of sporadic *Salmonella* infections in the United States are attributed to contact with reptiles or amphibians, especially iguanas, snakes, turtles, and lizards. Reptile-associated *Salmonella* infection more commonly leads to hospitalization and more frequently involves children, including infants, than do other *Salmonella* infections. Other pets, including hedgehogs, birds, rodents, baby chicks, ducklings, dogs, and cats, also are potential sources of NTS.

 Increasing antibiotic resistance in NTS species is a global problem and has been linked to the widespread use of antimicrobial agents in food animals and especially in animal feed. In the early 1990s, *S.* Typhimurium definitive phage type 104 (DT104), characterized by resistance to at least five antibiotics (ampicillin, chloramphenicol, streptomycin, sulfonamides, and tetracyclines; R-type ACSSuT), emerged worldwide. In 2014, resistance to at least ACSSuT was reported in 3.1% of U.S. NTS isolates, including 14.5% of *S.* Typhimurium isolates. Acquisition is associated with exposure to ill farm animals and to various meat products, including uncooked or undercooked ground beef. Although probably no more virulent than susceptible *S.* Typhimurium strains, DT104 strains are associated with an increased risk of bloodstream

infection and hospitalization. DSC and trimethoprim-resistant DT104 strains are emerging, especially in the United Kingdom.

Because of increased resistance to conventional antibiotics such as ampicillin and TMP-SMX, extended-spectrum cephalosporins and fluoroquinolones have emerged as the agents of choice for the treatment of MDR NTS infections. In 2014, 2.4% of all U.S. NTS strains were resistant to ceftriaxone. Most ceftriaxone-resistant isolates were from children <18 years of age, in whom ceftriaxone is the antibiotic of choice for treatment of invasive NTS infection. These strains contained plasmid-encoded AmpC β-lactamases that were probably acquired by horizontal genetic transfer from *Escherichia coli* strains in food-producing animals—an event linked to the widespread use of the veterinary cephalosporin ceftiofur. Most recently, carbapenemase-resistant NTS strains have been reported in Europe, North Africa, and southern Asia.

Over the last decade, strains of DSC NTS (MIC, 0.125–0.5 µg/mL) have emerged and have been associated with delayed response and treatment failure. In 2014, 4.3% of NTS isolates in the United States displayed decreased susceptibility or resistance to ciprofloxacin (MIC, ≥1 µg/mL), and the proportion is higher in Europe (~6%). These strains have diverse resistance mechanisms, including single and multiple mutations in the DNA gyrase genes *gyr*A and *gyr*B, mutations in the chromosomally encoded quinolone resistance–determining region, and plasmid-encoded quinolone resistance genes that are not reliably detected by nalidixic acid susceptibility testing or standard ciprofloxacin disk diffusion. In 2012, the U.S. Clinical Laboratory Standards Institute proposed a lower ciprofloxacin susceptibility breakpoint (≥0.06 µg/mL) for all *Salmonella* species to address this issue. Because commercial test systems do not contain ciprofloxacin concentrations low enough to allow use of this breakpoint, laboratories need to determine the ciprofloxacin MIC by Etest or another alternative method.

CLINICAL MANIFESTATIONS

Gastroenteritis

Infection with NTS most often results in gastroenteritis indistinguishable from that caused by other enteric pathogens. Nausea, vomiting, and diarrhea occur 6–48 h after the ingestion of contaminated food or water. Patients often experience abdominal cramping and fever (38–39°C; 100.5–102.2°F). Diarrheal stools are usually loose, nonbloody, and of moderate volume. However, large-volume watery stools, bloody stools, or symptoms of dysentery may occur. Rarely, NTS causes pseudoappendicitis or an illness that mimics inflammatory bowel disease.

Gastroenteritis caused by NTS is usually self-limited. Diarrhea resolves within 3–7 days and fever within 72 h. Stool cultures remain positive for 4–5 weeks after infection and—in rare cases of chronic carriage (<1%)—for >1 year. Persistent NTS infection and relapsing diarrhea have been described in a small fraction of patients and were associated with in-host single-nucleotide mutations in key virulence regulators. For acute NTS gastroenteritis, antibiotic treatment usually is not recommended and may prolong fecal carriage. Neonates, the elderly, and immunosuppressed patients (e.g., transplant recipients, HIV-infected persons) with NTS gastroenteritis are especially susceptible to dehydration and invasive infection and may require

hospitalization and antibiotic therapy. Acute NTS gastroenteritis was associated with a threefold increased risk of dyspepsia and irritable bowel syndrome at 1 year in a study from Spain.

Bacteremia and Endovascular Infections

Up to 8% of patients with NTS gastroenteritis develop bacteremia; of these, 5–10% develop localized infections. Bacteremia and metastatic infection are most common with *Salmonella* Choleraesuis and *Salmonella* Dublin and among infants, the elderly, and immunocompromised patients, especially those with HIV infection. NTS endovascular infection should be suspected in high-grade or persistent bacteremia, especially with preexisting valvular heart disease, atherosclerotic vascular disease, prosthetic vascular graft, or aortic aneurysm. Arteritis should be suspected in elderly patients with prolonged fever and back, chest, or abdominal pain developing after an episode of gastroenteritis. Endocarditis and arteritis are rare (<1% of cases) but are associated with potentially fatal complications, including valve perforation, endomyocardial abscess, infected mural thrombus, pericarditis, mycotic aneurysms, aneurysm rupture, aortoenteric fistula, and vertebral osteomyelitis.

 In some areas of sub-Saharan Africa, NTS is among the most common causes of bacteremia in children. NTS bacteremia among these children is not associated with diarrhea and has been associated with poor nutritional status, malaria, and HIV infection. The responsible strains form a specific clade that is associated with genome reduction; they also exhibit reduced resistance to oxidative and nitrogenous stress, and they have lost the ability to produce catalase and a flavohemoprotein involved in these processes.

Localized Infections

INTRAABDOMINAL INFECTIONS

Intraabdominal infections due to NTS are rare and usually manifest as hepatic or splenic abscesses or as cholecystitis. Risk factors include hepatobiliary anatomic abnormalities (e.g., gallstones), abdominal malignancy, and sickle cell disease (especially with splenic abscesses). Eradication of the infection often requires surgical correction of abnormalities and percutaneous drainage of abscesses.

CENTRAL NERVOUS SYSTEM INFECTIONS

NTS meningitis most commonly develops in infants 1–4 months of age and in adults with HIV infection. It often results in severe sequelae (including seizures, hydrocephalus, brain infarction, and mental retardation), with death in up to 60% of cases. Other rare central nervous system infections include ventriculitis, subdural empyema, and brain abscesses.

PULMONARY INFECTIONS

NTS pulmonary infections usually present as lobar pneumonia, and complications include lung abscess, empyema, and bronchopleural fistula formation. The majority of cases occur in patients with lung cancer, structural lung disease, sickle cell disease, or glucocorticoid use.

URINARY AND GENITAL TRACT INFECTIONS

Urinary tract infections caused by NTS present as either cystitis or pyelonephritis. Risk factors include malignancy, urolithiasis, structural abnormalities, HIV infection, and renal transplantation. NTS genital

infections are rare and include ovarian and testicular abscesses, prostatitis, and epididymitis. Like other focal infections, both genital and urinary tract infections can be complicated by abscess formation.

BONE, JOINT, AND SOFT TISSUE INFECTIONS

Salmonella osteomyelitis most commonly affects the femur, tibia, humerus, or lumbar vertebrae and is most often seen in association with sickle cell disease, hemoglobinopathies, or preexisting bone disease (e.g., fractures). Prolonged antibiotic treatment is recommended to decrease the risk of relapse and chronic osteomyelitis. Septic arthritis occurs in the same patient population as osteomyelitis and usually involves the knee, hip, or shoulder joints. Reactive arthritis can follow NTS gastroenteritis and is seen most frequently in persons with the HLA-B27 histocompatibility antigen. NTS rarely can cause soft tissue infections, usually at sites of local trauma in immunosuppressed patients.

DIAGNOSIS

The diagnosis of NTS infection is based on isolation of the organism from freshly passed stool or from blood or another ordinarily sterile body fluid. All salmonellae isolated in clinical laboratories should be sent to local public health departments for serotyping. Blood cultures should be done whenever a patient has prolonged or recurrent fever. Endovascular infection should be suspected if there is high-grade bacteremia (>50% of three or more blood cultures positive). Echocardiography, CT, and indium-labeled white cell scanning are used to identify localized infection. When another localized infection is suspected, joint fluid, abscess drainage, or cerebrospinal fluid should be cultured, as clinically indicated.

TREATMENT

TREATMENT

Nontyphoidal Salmonellosis

Antibiotics should not be used routinely to treat uncomplicated NTS gastroenteritis. The symptoms are usually self-limited, and the duration of fever and diarrhea is not significantly decreased by antibiotic therapy. In addition, antibiotic treatment has been associated with increased rates of relapse, prolonged gastrointestinal carriage, and adverse drug reactions. Dehydration secondary to diarrhea should be treated with fluid and electrolyte replacement.

Preemptive antibiotic treatment **(Table 160-2)** should be considered for patients at increased risk for invasive NTS infection, including neonates (probably up to 3 months of age); persons >50 years of age with known or suspected atherosclerosis; and patients with immunosuppression, cardiac valvular or endovascular abnormalities, or significant joint disease. Treatment should consist of an oral or IV antibiotic administered for 48–72 h or until the patient becomes afebrile. Immunocompromised persons may require up to 7–14 days of therapy. The <1% of persons who develop chronic carriage of NTS should receive a prolonged antibiotic course, as described above for chronic carriage of *S.* Typhi.

Because of the increasing prevalence of antibiotic resistance, empirical therapy for life-threatening NTS bacteremia or focal NTS infection should include a third-generation cephalosporin or a fluoroquinolone

(Table 160-2). If the bacteremia is low-grade (<50% of blood cultures positive), the patient should be treated for 7–14 days. Patients with HIV/AIDS and NTS bacteremia should receive 1–2 weeks of IV antibiotic therapy followed by 4 weeks of oral therapy with a fluoroquinolone. Patients whose infections relapse after this regimen should receive long-term suppressive therapy with a fluoroquinolone or TMP-SMX, as indicated by bacterial sensitivities.

If the patient has endocarditis or arteritis, treatment for 6 weeks with an IV β-lactam antibiotic (such as ceftriaxone or ampicillin) is indicated. IV ciprofloxacin followed by prolonged oral therapy is an option. Early surgical resection of infected aneurysms or other infected endovascular sites is recommended. Patients with infected prosthetic vascular grafts that cannot be resected have been maintained successfully on chronic suppressive oral therapy. For extraintestinal nonvascular infections, a 2- to 4-week course of antibiotic therapy (depending on the infection site) is usually recommended. In chronic osteomyelitis, abscess, or urinary or hepatobiliary infection associated with anatomic abnormalities, surgical resection or drainage may be required in addition to prolonged antibiotic therapy for eradication of infection.

TABLE 160-2

Antibiotic Therapy for Nontyphoidal *Salmonella* Infection in Adults

Indication	Agent	Dosage (Route)	Duration, Days
Preemptive Treatment[a]			
	Ciprofloxacin[b]	500 mg bid (PO)	2–3
Severe Gastroenteritis[c]			
	Ciprofloxacin	500 mg bid (PO) or 400 mg q12h (IV)	3–7
	Trimethoprim-sulfamethoxazole	160/800 mg bid (PO)	
	Amoxicillin	1 g tid (PO)	
	Ceftriaxone	1–2 g/d (IV)	
Bacteremia			
	Ceftriaxone[d]	2 g/d (IV)	7–14
	Ciprofloxacin	400 mg q12h (IV), then 500 mg bid (PO)	
Endocarditis or Arteritis			
	Ceftriaxone	2 g/d (IV)	42
	Ciprofloxacin	400 mg q8h (IV), then 750 mg bid (PO)	
	Ampicillin	2 g q4h (IV)	
Meningitis			
	Ceftriaxone	2 g q12 h (IV)	14–21
	Ampicillin	2 g q4h (IV)	
Other Localized Infection			

Indication	Agent	Dosage (Route)	Duration, Days
	Ceftriaxone	2 g/d (IV)	14–28
	Ciprofloxacin	500 mg bid (PO) or 400 mg q12h (IV)	
	Ampicillin	2 g q6h (IV)	

[a]Consider for neonates; persons >50 years of age with possible atherosclerotic vascular disease; and patients with immunosuppression, endovascular graft, or joint prosthesis. [b]Or ofloxacin, 400 mg bid (PO). [c]Consider on an individualized basis for patients with severe diarrhea and high fever who require hospitalization. [d]Or cefotaxime, 2 g q8h (IV).

PREVENTION AND CONTROL

Despite widespread efforts to prevent or reduce bacterial contamination of animal-derived food products and to improve food-safety education and training, recent declines in the incidence of NTS in the United States have been modest compared with those of other food-borne pathogens. This observation probably reflects the complex epidemiology of NTS. Identifying effective risk-reduction strategies requires monitoring of every step of food production, from handling of raw animal or plant products to preparation of finished foods. Contaminated food can be made safe for consumption by pasteurization, irradiation, or proper cooking. All cases of NTS infection should be reported to local public health departments because tracking and monitoring of these cases can identify the source(s) of infection and help authorities anticipate large outbreaks. Lastly, the prudent use of antimicrobial agents in both humans and animals is needed to limit the emergence of MDR *Salmonella*.

FURTHER READING

Andrews JR, Ryan ET: Diagnostics for invasive *Salmonella* infections: Current challenges and future directions. Vaccine 33(Suppl 3):C8, 2015. [PubMed: 25937611]

Anwar E et al: Vaccines for preventing typhoid fever. Cochrane Database Syst Rev CD001261:2014.

Ao TT et al: Global burden of invasive nontyphoidal *Salmonella* disease, 2010. Emerg Infect Dis 21:941, 2015.

Crump JA et al: Epidemiology, clinical presentation, laboratory diagnosis, antimicrobial resistance, and antimicrobial management of invasive *Salmonella* infections. Clin Microbiol Rev 28:90, 2015.

Wain J et al: Typhoid fever. Lancet 385:1136, 2015.

Chapter 161: Shigellosis

Philippe J. Sansonetti; Jean Bergounioux

FIGURE 161-1

INTRODUCTION

The discovery of *Shigella* as the etiologic agent of dysentery—a clinical syndrome of fever, intestinal cramps, and frequent passage of small, bloody, mucopurulent stools—is attributed to the Japanese microbiologist Kiyoshi Shiga, who isolated the Shiga bacillus (now known as *Shigella dysenteriae* type 1) from patients' stools in 1897 during a large and devastating dysentery epidemic. *Shigella* cannot be distinguished from *Escherichia coli* by DNA hybridization and remains a separate species only on historical and clinical grounds.

ETIOLOGIC AGENT

Shigella is a non-spore-forming, gram-negative bacterium that, unlike *E. coli*, is nonmotile and does not produce gas from sugars, decarboxylate lysine, or hydrolyze arginine. Some serovars produce indole, and occasional strains utilize sodium acetate. *Shigella dysenteriae*, *Shigella flexneri*, *Shigella boydii*, and *Shigella sonnei* (serogroups A, B, C, and D, respectively) can be differentiated on the basis of biochemical and serologic characteristics.


Genome sequencing of *E. coli* K12, *S. flexneri* 2a, *S. sonnei*, *S. dysenteriae* type 1, and *S. boydii* has revealed that these species have ~93% of genes in common. The three major genomic "signatures" of *Shigella* are (1) a 215-kb virulence plasmid that carries most of the genes required for pathogenicity (particularly invasive capacity); (2) the lack or alteration of genetic sequences encoding products (e.g., lysine decarboxylase) that, if expressed, would attenuate pathogenicity; and (3) in *S. dysenteriae* type 1, the presence of genes encoding Shiga toxin, a potent cytotoxin.

EPIDEMIOLOGY


The human intestinal tract represents the major reservoir of *Shigella*, which is also found (albeit rarely) in the higher primates. Because excretion of shigellae is greatest in the acute phase of disease, the bacteria are transmitted most efficiently by the fecal–oral route via hand carriage; however, some outbreaks reflect foodborne or waterborne transmission. In impoverished areas, *Shigella* can be transmitted by flies. The high-

level infectivity of *Shigella* is reflected by the very small inoculum required for experimental infection of volunteers (100 colony-forming units [CFU]), by the very high attack rates during outbreaks in day-care centers (33–73%), and by the high rates of secondary cases among family members of sick children (26–33%). Shigellosis can also be transmitted sexually.

Throughout history, *Shigella* epidemics have often occurred in settings of human crowding under conditions of poor hygiene—e.g., among soldiers in campaigning armies, inhabitants of besieged cities, groups on pilgrimages, and refugees in camps. Epidemics follow a cyclical pattern in areas such as the Indian subcontinent and sub-Saharan Africa. These devastating epidemics, which are most often caused by *S. dysenteriae* type 1, are characterized by high attack and mortality rates. In Bangladesh, for instance, an epidemic caused by *S. dysenteriae* type 1 was associated with a 42% increase in mortality rate among children 1–4 years of age. Apart from these epidemics, shigellosis is mostly an endemic disease, with 99% of cases occurring in the developing world and the highest prevalences in the most impoverished areas, where personal and general hygiene is below standard. *S. flexneri* isolates predominate in the least developed areas, whereas *S. sonnei* is more prevalent in economically emerging countries and in the industrialized world.

Prevalence in the Developing World

In a review published under the auspices of the World Health Organization (WHO), the total annual number of cases in 1966–1997 was estimated at 165 million, and 69% of these cases occurred in children <5 years of age. In this review, the annual number of deaths was calculated to range between 500,000 and 1.1 million. More recent data (2000–2004) from six Asian countries indicate that, even though the incidence of shigellosis remains stable, mortality rates associated with this disease may have decreased significantly, possibly as a result of improved nutritional status. However, extensive and essentially uncontrolled use of antibiotics, which may also account for declining mortality rates, has increased the rate of emergence of multidrug-resistant *Shigella* strains. A 2013 prospective matched case–control study of children <5 years of age emphasizes the importance of *Shigella* in the burden and etiology of diarrheal diseases in developing countries. *Shigella* is one of the top four pathogens associated with moderate to severe diarrhea and is now ranked first among children 12–59 months of age. These moderate to severe cases account for an 8.5-fold increase in mortality incidence over the average diarrheal disease–related mortality. The study's authors conclude that *Shigella* remains a major pathogen to be targeted by health care programs.

An often-overlooked complication of shigellosis is the short- and long-term impairment of the nutritional status of infected children in endemic areas. Combined with anorexia, the exudative enteropathy resulting from mucosal abrasions contributes to rapid deterioration of the patient's nutritional status. Shigellosis is thus a major contributor to stunted growth among children in developing countries.

Peaking in incidence in the pediatric population, endemic shigellosis is rare among young and middle-aged adults, probably because of naturally acquired immunity. Incidence then increases again in the elderly population.

Prevalence in the Industrialized World

In pediatric populations, local outbreaks occur when proper and adapted hygiene policies are not implemented in group facilities like day-care centers and institutions for the mentally retarded. In adults, as in children, sporadic cases occur among travelers returning from endemic areas, and rare outbreaks of varying size can follow waterborne or foodborne infections.

PATHOGENESIS AND PATHOLOGY

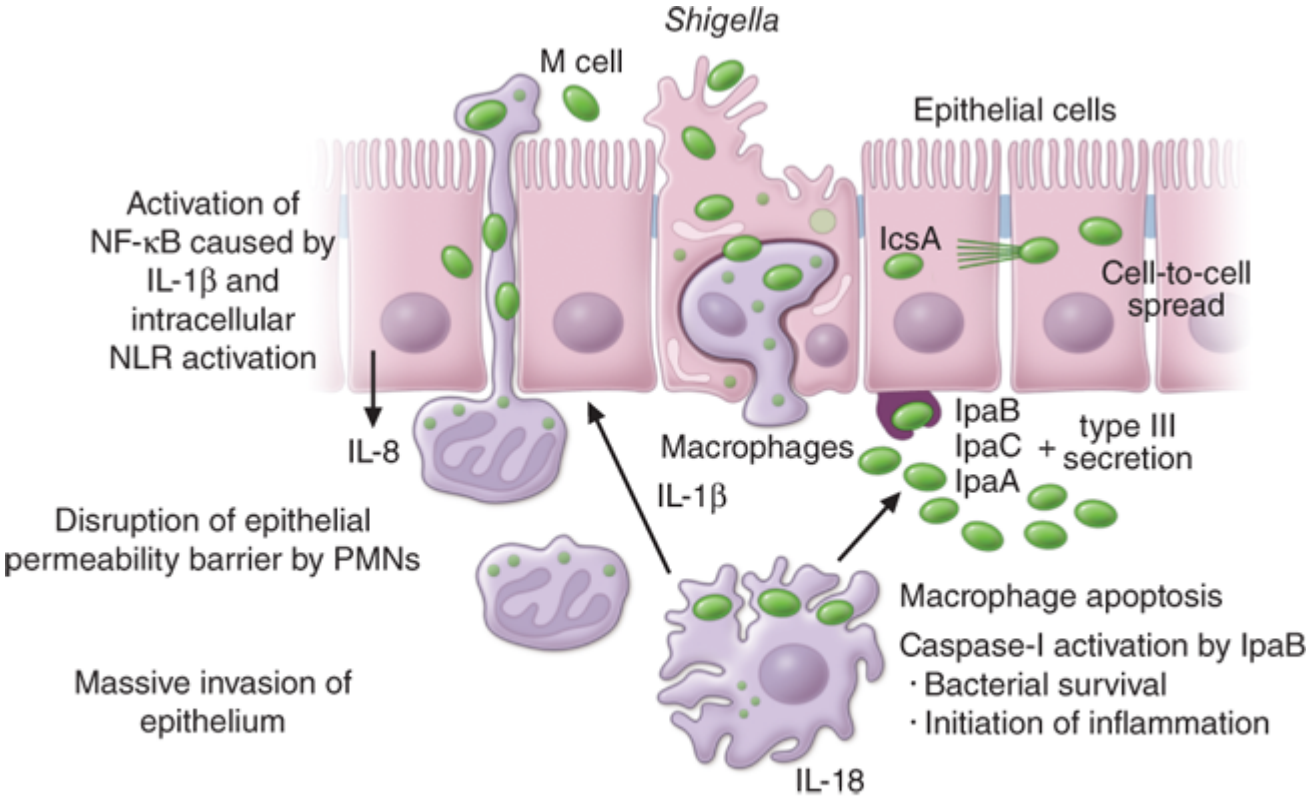
Shigella infection occurs essentially through oral contamination via direct fecal–oral transmission, the organism being poorly adapted to survive in the environment. Resistance to low-pH conditions allows shigellae to survive passage through the gastric barrier, an ability that may explain in part why a small inoculum (as few as 100 CFU) is sufficient to cause infection.

The watery diarrhea that usually precedes the dysenteric syndrome is attributable to active secretion and abnormal water reabsorption—a secretory effect at the jejunal level described in experimentally infected rhesus monkeys. This initial purge is probably due to the combined action of an enterotoxin (ShET-1) and mucosal inflammation. The dysenteric syndrome, manifested by bloody and mucopurulent stools, reflects invasion of the mucosa.


The pathogenesis of *Shigella* is essentially determined by a large virulence plasmid of 214 kb comprising ~100 genes, of which 25 encode a type III secretion system that inserts into the membrane of the host cell to allow effectors to transit from the bacterial cytoplasm to the host cell cytoplasm **(Fig. 161-1)**. Bacteria are thereby able to invade intestinal epithelial cells by inducing their own uptake after the initial crossing of the epithelial barrier through M cells (the specialized translocating epithelial cells in the follicle-associated epithelium that covers mucosal lymphoid nodules). The organisms induce apoptosis of subepithelial resident macrophages. Once inside the cytoplasm of intestinal epithelial cells, *Shigella* effectors trigger the cytoskeletal rearrangements necessary to direct uptake of the organism into the epithelial cell. The *Shigella*-containing vacuole is then quickly lysed, releasing bacteria into the cytosol.

FIGURE 161-1

Invasive strategy of *Shigella flexneri.* IL, interleukin; NF-κB, nuclear factor κB; NLR, NOD-like receptor; PMN, polymorphonuclear leukocyte.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Intracellular shigellae next use cytoskeletal components to propel themselves inside the infected cell; when the moving organism and the host cell membrane come into contact, cellular protrusions form and are engulfed by neighboring cells. This series of events permits bacterial cell-to-cell spread.

Cytokines released by a growing number of infected intestinal epithelial cells attract increased numbers of immune cells (particularly polymorphonuclear leukocytes [PMNs]) to the infected site, thus further destabilizing the epithelial barrier, exacerbating inflammation, and leading to the acute colitis that characterizes shigellosis. Evidence indicates that some type III secretion system–injected effectors can control the extent of inflammation, thus facilitating bacterial survival.

Shiga toxin produced by *S. dysenteriae* type 1 increases disease severity. This toxin belongs to a group of A1-B5 protein toxins whose B subunit binds to the receptor globotriaosylceramide on the target cell surface and whose catalytic A subunit is internalized by receptor-mediated endocytosis and interacts with the subcellular machinery to inhibit protein synthesis by expressing RNA *N*-glycosidase activity on 28S ribosomal RNA. This process leads to inhibition of binding of the amino-acyl-tRNA to the 60S ribosomal subunit and thus to a general shutoff of cell protein biosynthesis. Shiga toxins are translocated from the bowel into the circulation. After binding of the toxins to target cells in the kidney, pathophysiologic alterations may result in hemolytic-uremic syndrome (HUS; see below).

CLINICAL MANIFESTATIONS

The presentation and severity of shigellosis depend to some extent on the infecting serotype but even more on the age and the immunologic and nutritional status of the host. Poverty and poor standards of hygiene are strongly related to the number and severity of diarrheal episodes, especially in children <5 years old who have been weaned.

Shigellosis typically evolves through four phases: incubation, watery diarrhea, dysentery, and the postinfectious phase. The incubation period usually lasts 1–4 days but may be as long as 8 days. Typical initial manifestations are transient fever, limited watery diarrhea, malaise, and anorexia. Signs and symptoms may range from mild abdominal discomfort to severe cramps, diarrhea, fever, vomiting, and tenesmus. The manifestations are usually exacerbated in children, with temperatures up to 40°–41°C (104.0°–105.8°F) and more severe anorexia and watery diarrhea. This initial phase may represent the only clinical manifestation of shigellosis, especially in developed countries. Otherwise, dysentery follows within hours or days and is characterized by uninterrupted excretion of small volumes of bloody mucopurulent stools with increased tenesmus and abdominal cramps. At this stage, *Shigella* produces acute colitis involving mainly the distal colon and the rectum. Unlike most diarrheal syndromes, dysenteric syndromes rarely present with dehydration as a major feature. Endoscopy shows an edematous and hemorrhagic mucosa, with ulcerations and possibly overlying exudates resembling pseudomembranes. The extent of the lesions correlates with the number and frequency of stools and with the degree of protein loss by exudative mechanisms. Most episodes are self-limited and resolve without treatment in 1 week. With appropriate treatment, recovery takes place within a few days to a week, with no sequelae.

Acute life-threatening complications are seen most often in children <5 years of age (particularly those who are malnourished) and in elderly patients. Risk factors for death in a clinically severe case include nonbloody diarrhea, moderate to severe dehydration, bacteremia, absence of fever, abdominal tenderness, and rectal prolapse. Major complications are predominantly intestinal (e.g., toxic megacolon, intestinal perforations, rectal prolapse) or metabolic (e.g., hypoglycemia, hyponatremia, dehydration). Bacteremia is rare and is reported most frequently in severely malnourished and HIV-infected patients. Alterations of consciousness, including seizures, delirium, and coma, may occur, especially in children <5 years old, and are associated with a poor prognosis; fever and severe metabolic alterations are more often the major causes of altered consciousness than is meningitis or the Ekiri syndrome (toxic encephalopathy associated with bizarre posturing, cerebral edema, and fatty degeneration of viscera), which has been reported mostly in Japanese children. Pneumonia, vaginitis, and keratoconjunctivitis due to *Shigella* are rarely reported. In the absence of serious malnutrition, severe and very unusual clinical manifestations, such as meningitis, may be linked to genetic defects in innate immune functions (i.e., deficiency in interleukin 1 receptor–associated kinase 4 [IRAK-4]) and may require genetic investigation.

Two complications of particular importance are toxic megacolon and HUS. Toxic megacolon is a consequence of severe inflammation extending to the colonic smooth-muscle layer and causing paralysis and dilation. The patient presents with abdominal distention and tenderness, with or without signs of localized or generalized peritonitis. The abdominal x-ray characteristically shows marked dilation of the transverse colon (with the greatest distention in the ascending and descending segments); thumbprinting

caused by mucosal inflammatory edema; and loss of the normal haustral pattern associated with pseudopolyps, often extending into the lumen. Pneumatosis coli is an occasional finding. If perforation occurs, radiographic signs of pneumoperitoneum may be apparent. Predisposing factors (e.g., hypokalemia and use of opioids, anticholinergics, loperamide, psyllium seeds, and antidepressants) should be investigated.

Shiga toxin produced by *S. dysenteriae* type 1 has been linked to HUS in developing countries but rarely in industrialized countries, where enterohemorrhagic *E. coli* (EHEC) predominates as the etiologic agent of this syndrome. HUS is an early complication that most often develops after several days of diarrhea. Clinical examination shows pallor, asthenia, and irritability and, in some cases, bleeding of the nose and gums, oliguria, and increasing edema. HUS is a nonimmune (Coombs test–negative) hemolytic anemia defined by a diagnostic triad: microangiopathic hemolytic anemia (hemoglobin level typically <80 g/L [<8 g/dL]), thrombocytopenia (mild to moderate in severity; typically <60,000 platelets/μL), and acute renal failure due to thrombosis of the glomerular capillaries (with markedly elevated creatinine levels). Anemia is severe, with fragmented red blood cells (*schizocytes*) in the peripheral smear, high serum concentrations of lactate dehydrogenase and free circulating hemoglobin, and elevated reticulocyte counts. Acute renal failure occurs in 55–70% of cases; however, renal function recovers in most of these cases (up to 70% in various series). Leukemoid reactions, with leukocyte counts of 50,000/μL, are sometimes noted in association with HUS.

The postinfectious immunologic complication known as *reactive arthritis* can develop weeks or months after shigellosis, especially in patients expressing the histocompatibility antigen HLA-B27. About 3% of patients infected with *S. flexneri* later develop this syndrome, with arthritis, ocular inflammation, and urethritis—a condition that can last for months or years and can progress to difficult-to-treat chronic arthritis. Postinfectious arthropathy occurs only after infection with *S. flexneri* and not after infection with the other *Shigella* serotypes.

LABORATORY DIAGNOSIS

The differential diagnosis in patients with a dysenteric syndrome depends on the clinical and environmental context. In developing areas, infectious diarrhea caused by other invasive pathogenic bacteria (*Salmonella*, *Campylobacter jejuni*, *Clostridium difficile*, *Yersinia enterocolitica*) or parasites (*Entamoeba histolytica*) should be considered. Only bacteriologic and parasitologic examinations of stool can truly differentiate among these pathogens. A first flare of inflammatory bowel disease, such as Crohn's disease or ulcerative colitis (Chap. 319), should be considered in patients in industrialized countries. Despite the similarity in symptoms, anamnesis discriminates between shigellosis, which usually follows recent travel in an endemic zone, and these other conditions.

Microscopic examination of stool smears shows erythrophagocytic trophozoites with very few PMNs in *E. histolytica* infection, whereas bacterial enteroinvasive infections (particularly shigellosis) are characterized by high PMN counts in each microscopic field. However, because shigellosis often manifests only as watery diarrhea, systematic attempts to isolate *Shigella* are necessary.

The “gold standard” for the diagnosis of *Shigella* infection remains the isolation and identification of the pathogen from fecal material. One major difficulty, particularly in endemic areas where laboratory facilities are not immediately available, is the fragility of *Shigella* and its common disappearance during transport, especially with rapid changes in temperature and pH. In the absence of a reliable enrichment medium, buffered glycerol saline or Cary-Blair medium can be used as a holding medium, but prompt inoculation onto isolation medium is essential. The probability of isolation is higher if the portion of stools that contains bloody and/or mucopurulent material is directly sampled. Rectal swabs can be used, as they offer the highest rate of successful isolation during the acute phase of disease. Blood cultures are positive in fewer than 5% of cases but should be done when a patient presents with a clinical picture of severe sepsis.

In addition to quick processing, the use of several media increases the likelihood of successful isolation: a nonselective medium such as bromocresol-purple agar lactose; a low-selectivity medium such as MacConkey or eosin-methylene blue; and a high-selectivity medium such as Hektoen, *Salmonella-Shigella*, or xylose-lysine-deoxycholate agar. After incubation on these media for 12–18 h at 37°C (98.6°F), shigellae appear as non-lactose-fermenting colonies that measure 0.5–1 mm in diameter and have a convex, translucent, smooth surface. Suspected colonies on nonselective or low-selectivity medium can be subcultured on a high-selectivity medium before being specifically identified or can be identified directly by standard commercial systems on the basis of four major characteristics: glucose positivity (usually without production of gas), lactose negativity, H_2S negativity, and lack of motility. The four *Shigella* serogroups (A–D) can then be differentiated by additional characteristics. This approach adds time and difficulty to the identification process; however, after presumptive diagnosis, the use of serologic methods (e.g., slide agglutination, with group- and then type-specific antisera) should be considered. Group-specific antisera are widely available; in contrast, because of the large number of serotypes and subserotypes, type-specific antisera are rare and more expensive and thus are often restricted to reference laboratories.

TREATMENT

TREATMENT

Shigellosis

ANTIBIOTIC SUSCEPTIBILITY OF *SHIGELLA*



As an enteroinvasive disease, shigellosis requires antibiotic treatment. Since the mid-1960s, however, increasing resistance to multiple drugs has been a dominant factor in treatment decisions. Resistance rates are highly dependent on the geographic area. Clonal spread of particular strains and horizontal transfer of resistance determinants, particularly via plasmids and transposons, contribute to multidrug resistance. The current global status—i.e., high rates of resistance to classic first-line antibiotics such as amoxicillin—has led to a rapid switch to quinolones such as nalidixic acid. However, resistance to such early-generation quinolones has also emerged and spread quickly as a result of chromosomal mutations affecting DNA gyrase and topoisomerase IV; this resistance has necessitated the use of later-generation quinolones as first-line antibiotics in many areas. For instance, a review of the antibiotic resistance history of *Shigella* in India found

that, after their introduction in the late 1980s, the second-generation quinolones norfloxacin, ciprofloxacin, and ofloxacin were highly effective in the treatment of shigellosis, including cases caused by multidrug-resistant strains of *S. dysenteriae* type 1. However, investigations of subsequent outbreaks in India and Bangladesh detected resistance to norfloxacin, ciprofloxacin, and ofloxacin in 5% of isolates. The incidence of multidrug resistance parallels the widespread, uncontrolled use of antibiotics and calls for the rational use of effective drugs.

ANTIBIOTIC TREATMENT OF SHIGELLOSIS

Because of the ready transmissibility of *Shigella*, current public health recommendations in the United States are that every case be treated with antibiotics **(Table 161-1)**. Ciprofloxacin is recommended as first-line treatment. A number of other drugs have been tested and shown to be effective, including ceftriaxone, azithromycin, pivmecillinam, and some fifth-generation quinolones. Whereas infections caused by non-*dysenteriae Shigella* in immunocompetent individuals are routinely treated with a 3-day course of antibiotics, it is recommended that *S. dysenteriae* type 1 infections be treated for 5 days and that *Shigella* infections in immunocompromised patients be treated for 7–10 days.

Treatment for shigellosis must be adapted to the clinical context, with the recognition that the most fragile patients are children <5 years old, who represent two-thirds of all cases worldwide. There are few data on the use of quinolones in children, but *Shigella*-induced dysentery is a well-recognized indication for their use. The half-life of ciprofloxacin is longer in infants than in older individuals. The ciprofloxacin dose generally recommended for children is 30 mg/kg per day in two divided doses. Adults living in areas with high standards of hygiene are likely to develop milder, shorter-duration disease, whereas infants in endemic areas can develop severe, sometimes fatal, dysentery. In the former setting, treatment will remain minimal and bacteriologic proof of infection will often come after symptoms have resolved; in the latter setting, antibiotic treatment and more aggressive measures, possibly including resuscitation, are often required.

REHYDRATION AND NUTRITION

Shigella infection rarely causes significant dehydration. Cases requiring aggressive rehydration (particularly in industrialized countries) are uncommon. In developing countries, malnutrition remains the primary indicator for diarrhea-related death, highlighting the importance of nutrition in early management. Rehydration should be oral unless the patient is comatose or presents in shock. Because of the improved effectiveness of reduced-osmolarity oral rehydration solution (especially for children with acute noncholera diarrhea), the WHO and UNICEF now recommend a standard solution of 245 mOsm/L (sodium, 75 mmol/L; chloride, 65 mmol/L; glucose [anhydrous], 75 mmol/L; potassium, 20 mmol/L; citrate, 10 mmol/L). In shigellosis, the coupled transport of sodium and glucose may be variably affected, but oral rehydration therapy remains the easiest and most efficient form of rehydration, especially in severe cases.

Nutrition should be started as soon as possible after completion of initial rehydration. Early refeeding is safe, well tolerated, and clinically beneficial. Because breast-feeding reduces diarrheal losses and the need for oral rehydration in infants, it should be maintained in the absence of contraindications (e.g., maternal HIV infection).

NONSPECIFIC, SYMPTOM-BASED THERAPY

Antimotility agents have been implicated in prolonged fever in volunteers with shigellosis. These agents are suspected of increasing the risk of toxic megacolon and are thought to have been responsible for HUS in children infected by EHEC strains. For safety reasons, it is better to avoid antimotility agents in bloody diarrhea.

TREATMENT OF COMPLICATIONS

There is no consensus regarding the best treatment for toxic megacolon. The patient should be assessed frequently by both medical and surgical teams. Anemia, dehydration, and electrolyte deficits (particularly hypokalemia) may aggravate colonic atony and should be actively treated. Nasogastric aspiration helps to deflate the colon. Parenteral nutrition has not been proven to be beneficial. Fever persisting beyond 48–72 h raises the possibility of local perforation or abscess. Most studies recommend colectomy if, after 48–72 h, colonic distention persists. However, some physicians recommend continuation of medical therapy for up to 7 days if the patient seems to be improving clinically despite persistent megacolon without free perforation. Intestinal perforation, either isolated or complicating toxic megacolon, requires surgical treatment and intensive medical support.

Rectal prolapse must be treated as soon as possible. With the health care provider using surgical gloves or a soft warm wet cloth and the patient in the knee-chest position, the prolapsed rectum is gently pushed back into place. If edema of the rectal mucosa is evident (rendering reintegration difficult), it can be osmotically reduced by the application of gauze impregnated with a warm solution of saturated magnesium sulfate. Rectal prolapse often relapses but usually resolves along with the resolution of dysentery.

HUS must be treated by water restriction, including discontinuation of oral rehydration solution and potassium-rich alimentation. Hemofiltration is usually required.

TABLE 161-1

Recommended Antimicrobial Therapy for Shigellosis

	Treatment Schedule		
Antimicrobial Agent	**Children**	**Adults**	**Limitations**
First Line			
Ciprofloxacin	15 mg/kg	500 mg	
	2 times per day for 3 days, PO		
Second Line			
Pivmecillinam	20 mg/kg	100 mg	Cost
	4 times per day for 5 days PO		No pediatric formulation Frequent administration Emerging resistance
Ceftriaxone	50–100 mg/kg	–	Efficacy not validated Must be injected
	Once a day IM for 2–5 days		
Azithromycin	6–20 mg/kg	1–1.5 g	Cost
	Once a day for 1–5 days PO		Efficacy not validated Minimum inhibitory concentration near serum concentration Rapid emergence of resistance and spread to other bacteria

Source: WHO Library Cataloguing-in-Publication Data: Guidelines for the control of shigellosis, including epidemics due to *Shigella dysenteriae* type 1 (*www.who.int/cholera/publications/shigellosis/en/*).

PREVENTION

Hand washing after defecation or handling of children's feces and before handling of food is recommended. Stool decontamination (e.g., with sodium hypochlorite), together with a cleaning protocol for medical staff as well as for patients, has proven useful in limiting the spread of infection during *Shigella* outbreaks. Ideally, patients should have a negative stool culture before their infection is considered cured. Recurrences are rare if therapeutic and preventive measures are correctly implemented.

Although several live attenuated oral and subunit parenteral vaccine candidates have been produced and are undergoing clinical trials, no vaccine against shigellosis is currently available. Especially given the rapid progression of antibiotic resistance in *Shigella*, a vaccine is urgently needed.

FURTHER READING

Anderson M et al: *Shigella* diversity and changing landscape: Insights for the 21st century. Front Cell Infect Microbiol 6:45, 2016.

Bennish ML, Wojtyniak BJ: Mortality due to shigellosis: Community and hospital data. Rev Infect Dis 13(Suppl 4):S245, 1991. [PubMed: 2047645]

Cossart P, Sansonetti PJ: Bacterial invasion: The paradigms of enteroinvasive pathogens. Science 304:242, 2004.

Kotloff KL et al: Global burden of *Shigella* infections: Implications for vaccine development and implementation of control strategies. Bull World Health Organ 77:651, 1999.

Kotloff KL et al: Overview of live vaccine strategies against *Shigella*, in *New Generation Vaccines*, 3rd ed, MM Levine et al (eds). London, Informa Healthcare, 2004, pp 723–735.

Kotloff KL et al: Burden and aetiology of diarrhoeal diseases in infants and young children in developing countries (The Global Enteric Multicenter Study, GEMS): A prospective case–control study. Lancet 382:35, 2013.

Mani S et al: Status of vaccine research and development for *Shigella*. Vaccine 34:2887, 2016.

Niyogi SK: Shigellosis. J Microbiol 43:133, 2005.

Phalipon A, Sansonetti PJ: *Shigella*'s ways of manipulating the host intestinal innate and adaptive immune system: A tool box for survival? Immunol Cell Biol 85:119, 2007.

Traa BS et al: Antibiotics for the treatment of dysentery in children. Int J Epidemiol 39(Suppl 1):i70, 2010. [PubMed: 20348130]

World Health Organization: Guidelines for the control of shigellosis, including epidemics due to *Shigella dysenteriae* type 1. WHO Library Cataloguing-in-Publication Data. Available from *www.who.int/cholera/publications/shigellosis/en/.*

Chapter 162: Infections Due to *Campylobacter* and Related Organisms

Beth D. Kirkpatrick; Martin J. Blaser

DEFINITION

Bacteria of the genus *Campylobacter* and of the related genera *Arcobacter* and *Helicobacter* (Chap. 158) cause a variety of inflammatory conditions. Although acute diarrheal illnesses are most common, these organisms may cause infections in virtually all parts of the body, especially in compromised hosts, and these infections may have late nonsuppurative sequelae. The designation *Campylobacter* comes from the Greek for “curved rod” and refers to the organism’s vibrio-like morphology.

ETIOLOGY

Campylobacters are motile, non-spore-forming, curved, gram-negative rods. Originally known as *Vibrio fetus*, these bacilli were reclassified as a new genus in 1973 after their dissimilarity to other vibrios was recognized. More than 20 species have since been identified. These species are currently divided into three genera: *Campylobacter*, *Arcobacter*, and *Helicobacter*. Not all of the species are pathogens of humans. The human pathogens fall into two major groups: those that primarily cause diarrheal disease and those that cause extraintestinal infection. The principal diarrheal pathogen is *Campylobacter jejuni*, which accounts for 80–90% of all cases of recognized illness due to campylobacters and related genera. Other organisms that cause diarrheal disease include *Campylobacter coli, Campylobacter upsaliensis, Campylobacter lari, Campylobacter hyointestinalis, Campylobacter fetus, Arcobacter butzleri, Arcobacter cryaerophilus, Helicobacter cinaedi,* and *Helicobacter fennelliae.* The two *Helicobacter* species causing diarrheal disease, *H. cinaedi* and *H. fennelliae*, are intestinal rather than gastric organisms; in terms of the clinical features of the illnesses they cause, these species most closely resemble *Campylobacter* rather than *Helicobacter pylori* (Chap. 158) and thus are considered in this chapter. The pathogenic roles of *Campylobacter concisus, Campylobacter ureolyticus, Campylobacter troglodytis,* and *Campylobacter pyloridis* are uncertain. A new subspecies—*C. fetus* subspecies *testudinum*—has been described, chiefly in Asian patients; its close resemblance to strains isolated from reptiles suggests a food source.

The major species causing extraintestinal illnesses is *C. fetus*. However, any of the diarrheal agents listed above may cause systemic or localized infection as well, especially in compromised hosts. Neither aerobes nor strict anaerobes, these microaerophilic organisms are adapted for survival in the gastrointestinal mucous

layer. This chapter focuses on *C. jejuni* and *C. fetus* as the major pathogens in and prototypes for their groups. The key features of infection are listed by species (excluding *C. jejuni*, described in detail in the text below) in **Table 162-1**.

TABLE 162-1

Clinical Features Associated with Infection Due to "Atypical" *Campylobacter* and Related Species Implicated as Causes of Human Illness

Species	Common Clinical Features	Less Common Clinical Features	Additional Information
Campylobacter coli	Fever, diarrhea, abdominal pain	Bacteremia[a]	Clinically indistinguishable from *C. jejuni*
Campylobacter fetus	Bacteremia,[a] sepsis, meningitis, vascular infections	Diarrhea, relapsing fevers	Not usually isolated from media containing cephalothin or incubated at 42°C
Campylobacter upsaliensis	Watery diarrhea, low-grade fever, abdominal pain	Bacteremia, abscesses	Difficult to isolate because of cephalothin susceptibility
Campylobacter lari	Abdominal pain, diarrhea	Colitis, appendicitis	Seagulls frequently colonized; organism often transmitted to humans via contaminated water
Campylobacter hyointestinalis	Watery or bloody diarrhea, vomiting, abdominal pain	Bacteremia	Causes proliferative enteritis in swine
Helicobacter fennelliae	Chronic mild diarrhea, abdominal cramps, proctitis	Bacteremia[a]	Best treated with fluoroquinolones
Helicobacter cinaedi	Chronic mild diarrhea, abdominal cramps, proctitis	Bacteremia[a]	Best treated with fluoroquinolones; identified in healthy hamsters
Campylobacter jejuni subspecies *doylei*	Diarrhea	Chronic gastritis, bacteremia[b]	Uncertain role as human pathogen

Species	Common Clinical Features	Less Common Clinical Features	Additional Information
Arcobacter cryaerophilus	Diarrhea	Bacteremia	Cultured under aerobic conditions
Arcobacter butzleri	Fever, diarrhea, abdominal pain, nausea	Bacteremia, appendicitis	Cultured under aerobic conditions; enzootic in nonhuman primates
Campylobacter sputorum	Pulmonary, perianal, groin, and axillary abscesses; diarrhea	Bacteremia	Three clinically relevant biovars: *sputorum, faecalis*, and *paraureolyticus*

[a]In immunocompromised hosts, especially HIV-infected persons. [b]In children.

Source: Adapted from BM Allos, MJ Blaser: Clin Infect Dis 20:1092, 1995.

EPIDEMIOLOGY

Campylobacters are found in the gastrointestinal tract of many animals used for food (including poultry, cattle, sheep, and swine) and many household pets (including birds, dogs, and cats). These microorganisms usually do not cause illness in their animal hosts. In most cases, campylobacters are transmitted to humans in raw or undercooked food products or through direct contact with infected animals. In the United States and other developed countries, ingestion of contaminated poultry that has not been sufficiently cooked is the most common mode of acquisition (30–70% of cases). Other modes include ingestion of raw (unpasteurized) milk or untreated water, contact with infected household pets, travel to developing countries (campylobacters being a leading cause of traveler's diarrhea; **Chaps. 119 and 128**), oral–anal sexual contact, cross-contamination from any of these sources, and (occasionally) contact with an index case who is incontinent of stool.

Campylobacter infections are common. Active surveillance of foodborne infections in the United States estimates the incidence of diarrheal disease due to campylobacters at 11.8 cases per 100,000 persons—similar in incidence to *Salmonella* and more common than *Shigella*. Infections occur throughout the year, but the incidence peaks during summer and early autumn. Persons of all ages are affected; however, attack rates for *C. jejuni* are highest among young children and young adults, whereas those for *C. fetus* are highest at the extremes of age. Systemic infections due to *C. fetus* (and to other *Campylobacter* and related species) are most common among compromised hosts. Persons at increased risk include those with AIDS,

hypogammaglobulinemia, neoplasia, liver disease, diabetes mellitus, and generalized atherosclerosis as well as neonates and pregnant women. However, apparently healthy nonpregnant persons occasionally develop transient *Campylobacter* bacteremia as part of a gastrointestinal illness (0.1–1% of cases).

In contrast, in many developing countries, *C. jejuni* infections are hyperendemic, with the highest rates among children <2 years old. According to large prospective cohort studies in low- to middle-income countries, *Campylobacter* infections—even when asymptomatic—are associated with growth shortfalls (stunting). Rates of clinically apparent infection fall with age, as does the illness-to-infection ratio.

PATHOLOGY AND PATHOGENESIS

 *C. jejuni* infections may be subclinical, especially in hosts in developing countries who have had multiple prior infections and may be partially immune. Symptomatic infections mostly occur within 2–4 days (range, 1–7 days) of exposure to the organism. The sites of tissue injury include the jejunum, ileum, and colon. Biopsies show an acute nonspecific inflammatory reaction, with neutrophils, monocytes, and eosinophils in the lamina propria, as well as damage to the epithelium, including loss of mucus, glandular degeneration, and crypt abscesses. Biopsy findings may be consistent with Crohn's disease or ulcerative colitis, but these "idiopathic" chronic inflammatory diseases should not be diagnosed unless infectious colitis, *specifically including* that due to infection with *Campylobacter* species and related organisms, has been ruled out.

The components of protective immunity to *Campylobacter* in humans are poorly understood. The high frequency of *C. jejuni* infections and their severity and recurrence among hypogammaglobulinemic patients suggest that antibodies are important in protective immunity. Experience from field studies and human experimental infection models suggests that immune protection may be short-lived or incomplete in the absence of continuous exposure. Knowledge of the pathogenesis of infection is also incomplete. Both the motility of the strain and its capacity to adhere to host tissues appear to favor disease, but classic enterotoxins and cytotoxins (including cytolethal distending toxin) appear not to play substantial roles in tissue injury or disease production. The organisms have been visualized within the epithelium, albeit in low numbers. The documentation of a significant tissue response and occasionally of *C. jejuni* bacteremia further suggests that tissue invasion is clinically significant, and in vitro studies are consistent with this pathogenic feature.

The pathogenesis of *C. fetus* infections is better defined. Virtually all clinical isolates of *C. fetus* possess a proteinaceous capsule-like structure (an S-layer) that renders the organisms resistant to complement-mediated killing and opsonization. As a result, *C. fetus* can cause bacteremia and can seed sites beyond the intestinal tract. The ability of the organism to switch the S-layer proteins expressed—a phenomenon that results in antigenic variability—may contribute to the chronicity and high rate of recurrence of *C. fetus* infections in compromised hosts.

CLINICAL MANIFESTATIONS

The clinical features of infections due to *Campylobacter* and the related *Arcobacter* and intestinal *Helicobacter* species causing enteric disease appear to be highly similar. *C. jejuni* can be considered the prototype, in part because it is by far the most common enteric pathogen in the group. A prodrome of fever, headache, myalgia, and/or malaise often occurs 12–48 h before the onset of diarrheal symptoms. The most common signs and symptoms of the intestinal phase are diarrhea, abdominal pain, and fever. The degree of diarrhea varies from several loose watery stools to visibly bloody stools (~10% of cases in adults); most patients presenting for medical attention have ≥10 bowel movements on the worst day of illness. Abdominal pain usually consists of cramping and may be the most prominent symptom. Pain is usually generalized but may become localized; *C. jejuni* infection may cause pseudoappendicitis. Fever may be the only initial manifestation of *C. jejuni* infection, a situation mimicking the early stages of typhoid fever. Febrile young children may develop convulsions. *Campylobacter* enteritis is generally self-limited; however, symptoms persist for >1 week in 10–20% of patients seeking medical attention, and clinical relapses occur in 5–10% of untreated patients. Studies of common-source epidemics indicate that milder illnesses or asymptomatic infections may commonly occur.

C. fetus may cause a diarrheal illness similar to that due to *C. jejuni*, especially in immunocompetent hosts. This organism also may cause either intermittent diarrhea or nonspecific abdominal pain without localizing signs. Sequelae are uncommon, and the outcome is benign. *C. fetus* may also cause a prolonged relapsing systemic illness (with fever, chills, and myalgias) that has no obvious primary source; this manifestation is especially common among compromised hosts. Secondary seeding of an organ (e.g., meninges, brain, bone, urinary tract, or soft tissue) complicates the course, which may be fulminant. *C. fetus* infections have a tropism for vascular sites: endocarditis, mycotic aneurysm, and septic thrombophlebitis may all occur. Infection during pregnancy often leads to fetal death. A variety of *Campylobacter* species and *H. cinaedi* can cause recurrent cellulitis with fever and bacteremia in immunocompromised hosts.

COMPLICATIONS

Except in infection with *C. fetus*, bacteremia is uncommon, developing most often in immunocompromised hosts and at the extremes of age. Three patterns of extraintestinal infection have been noted: (1) transient bacteremia in a normal host with enteritis (benign course, no specific treatment needed); (2) sustained bacteremia or focal infection in a normal host (bacteremia originating from enteritis, with patients responding well to antimicrobial therapy); and (3) sustained bacteremia or focal infection in a compromised host. Enteritis may not be clinically apparent. Antimicrobial therapy, possibly prolonged, is necessary for suppression or cure of the infection.

Campylobacter, Arcobacter, and intestinal *Helicobacter* infections in patients with AIDS or hypogammaglobulinemia may be severe, persistent, and extraintestinal; relapse after cessation of therapy is common. Hypogammaglobulinemic patients also may develop osteomyelitis and an erysipelas-like rash or cellulitis.

Local suppurative complications of infection include cholecystitis, pancreatitis, and cystitis; distant complications include meningitis, endocarditis, arthritis, peritonitis, cellulitis, and septic abortion. All these complications are rare, except in immunocompromised hosts. Hepatitis, interstitial nephritis, and the hemolytic-uremic syndrome occasionally complicate acute infection. The two most common postinfectious sequelae are reactive arthritis and the Guillain-Barré syndrome. Reactive arthritis has been reported in up to 2.5% of cases, although nonspecific rheumatologic symptoms are more common (~10%). Reactive arthritis may develop several weeks after infection, especially in persons with the HLA-B27 phenotype. The knees are most frequently involved, but involvement of the ankles, wrists, and small joints of the hands is common, with an average of 3.2 joints affected. Guillain-Barré syndrome or its Miller Fisher (cranial polyneuropathy) variant follows symptomatic or asymptomatic *Campylobacter* infections uncommonly—i.e., in 1 of every 1000–2000 cases or, for certain *C. jejuni* serotypes (such as O19), in 1 of every 100–200 cases. Despite the low frequency of this complication, it is estimated that *Campylobacter* infections, because of their high incidence, may trigger 20–40% of all cases of Guillain-Barré syndrome. The presence of sialylated lipopolysaccharides on *C. jejuni* strains prompts a form of molecular mimicry that promotes autoimmune recognition of sialylated cell-surface molecules on axons. Immunoproliferative small-intestinal disease (*alpha chain disease*), a form of lymphoma that originates in small-intestinal mucosa-associated lymphoid tissue, has been associated with *C. jejuni*; antimicrobial therapy has led to marked clinical improvement.

DIAGNOSIS

In patients with *Campylobacter* enteritis, peripheral leukocyte counts reflect the severity of the inflammatory process. In addition, stools from nearly all *Campylobacter*-infected patients presenting for medical attention in the United States contain leukocytes or erythrocytes. Gram- or Wright-stained fecal smears should be examined in all suspected cases. When the diagnosis of *Campylobacter* enteritis is suspected on the basis of findings indicating inflammatory diarrhea (fever, fecal leukocytes), clinicians can ask the microbiology laboratory to attempt the visualization of organisms with characteristic vibrioid morphology by direct microscopic examination of stools with Gram's staining or to use phase-contrast or dark-field microscopy to identify the organisms' characteristic "darting" motility. Confirmation of the diagnosis of *Campylobacter* infection is based on identification of an isolate from cultures of stool, blood, or another site. *Campylobacter*-specific media should be used to culture stools from all patients with inflammatory or bloody diarrhea. Because all *Campylobacter* species are fastidious, they will not be isolated unless selective media or other selective techniques are used. Failure to isolate campylobacters from stool by culture does not entirely rule out their presence. Although culture remains the diagnostic gold standard, species-specific real-time polymerase chain reaction (PCR) techniques appear more sensitive than culture; although PCR techniques may detect nonviable bacteria, they are now used frequently to diagnose infection with *Campylobacter* and other enteric bacteria in clinical microbiology laboratories. The detection of the organisms in stool in the United States almost always implies active or recent infection. In contrast, *Campylobacter sputorum* and related organisms found in the oral cavity are commensals that only rarely have pathogenic significance. Because of the low levels of metabolic activity of *Campylobacter* species in standard blood culture media, *Campylobacter* bacteremia is difficult to detect.

DIFFERENTIAL DIAGNOSIS

The symptoms of *Campylobacter* enteritis are not sufficiently unusual to distinguish this illness from that due to *Salmonella, Shigella, Yersinia*, and other pathogens. The combination of fever and fecal leukocytes or erythrocytes is indicative of inflammatory diarrhea, and definitive diagnosis is based on culture, real-time PCR, or demonstration of the characteristic organisms on stained fecal smears. Similarly, extraintestinal *Campylobacter* illness is diagnosed by culture. Infection due to *Campylobacter* should be suspected in the setting of septic abortion, and that due to *C. fetus* should be suspected specifically in the setting of septic thrombophlebitis. It is important to reiterate that (1) the presentation of *Campylobacter* enteritis may mimic that of ulcerative colitis or Crohn's disease, (2) *Campylobacter* enteritis is much more common than either of the latter (especially among young adults), and (3) biopsy may not distinguish among these entities. Thus, a diagnosis of inflammatory bowel disease should not be made until *Campylobacter* infection has been ruled out, especially in persons with a history of foreign travel, significant animal contact, immunodeficiency, or exposure incurring a high risk of transmission.

TREATMENT

TREATMENT

Campylobacter Infection

Fluid and electrolyte replacement is central to the treatment of diarrheal illnesses **(Chap. 128)**. Even among patients presenting for medical attention with *Campylobacter* enteritis, not all clearly benefit from specific antimicrobial therapy. Indications for therapy include high fever, bloody diarrhea, severe diarrhea, persistence for >1 week, and worsening of symptoms. A 3-day course of azithromycin (500 mg once daily) is the regimen of choice. A 1-day regimen of azithromycin (1000 mg given as two 500-mg tablets) can also be used. Alternative regimens for adults consist of fluoroquinolones—ciprofloxacin (500 mg by mouth twice daily for 3 days) or levofloxacin (750 mg daily for 3 days)—but resistance to this class of agents as well as to tetracyclines is substantial; ~27% of U.S. human isolates of *Campylobacter* in 2014 were resistant to ciprofloxacin. Because macrolide resistance usually is much less common (<10%), these drugs are the empirical agents of choice. Patients infected with antibiotic-resistant strains are at increased risk of adverse outcomes. Use of antimotility agents, which may prolong the duration of symptoms and have been associated with toxic megacolon and with death, is not recommended. Of note, *C. jejuni* and *C. coli* are resistant to trimethoprim and β-lactam antibiotics, including penicillin and most cephalosporins.

For patients with immunocompromising conditions and uncomplicated enteritis caused by *C. jejuni*, therapy duration should be extended to 7–14 days. For systemic infections, treatment with a carbapenem (imipenem, 500 mg IV every 6 h; or meropenem, 1–2 g IV every 8 h) should be started empirically, and susceptibility testing should always be performed. For life-threatening illness, gentamicin (1.0–1.7 mg/kg IV every 8 h after a loading dose of 1.5–2 mg/kg) can be added. In the absence of endovascular involvement, therapy for systemic infections should be administered for 7–14 days. For immunocompromised patients with systemic infections due to *C. fetus* and for patients with endovascular infections due to any species, prolonged

therapy (up to 4 weeks) is usually necessary. For recurrent infections in immunocompromised hosts, lifelong therapy/prophylaxis is sometimes necessary.

PROGNOSIS

Nearly all patients recover fully from *Campylobacter* enteritis, either spontaneously or after antimicrobial therapy. Volume depletion probably contributes to the few deaths that are reported. As stated above, occasional patients develop reactive arthritis or Guillain-Barré syndrome or its variants. Systemic infection with *C. fetus* is much more often fatal than that due to related species; this higher mortality rate reflects in part the population affected. Prognosis depends on the rapidity with which appropriate therapy is begun. Otherwise healthy hosts usually survive *C. fetus* infections without sequelae. Compromised hosts often have recurrent and/or life-threatening infections due to a variety of *Campylobacter* species.

FURTHER READING

Amour C et al: Epidemiology and impact of *Campylobacter* infection in children in 8 low-resource settings: Results from the MAL-ED Study. Clin Infect Dis 63:1171, 2016.

Man SM: The clinical importance of emerging *Campylobacter* species. Nat Rev Gastroenterol Hepatol 8:669, 2011.

Marder EP et al: Incidence and trends of infections with pathogens transmitted commonly through food and the effect of increasing use of culture-independent diagnostic tests on surveillance—Foodborne Diseases Active Surveillance Network, 10 U.S. sites, 2013–2016. Morb Mortal Wkly Rep 66:397, 2017.

Riddle MS et al: ACG clinical guideline: Diagnosis, treatment, and prevention of acute diarrheal infections in adults. Am J Gastroenterol 111:602, 2016.

Ternhag A et al: A meta-analysis of the effects of antibiotic treatment on duration of symptoms caused by infection with *Campylobacter* species. Clin Infect Dis 44:696, 2007.

Chapter 163: Cholera and Other Vibrioses

Matthew K. Waldor; Edward T. Ryan

INTRODUCTION

Members of the genus *Vibrio* cause a number of important infectious syndromes. Classic among them is cholera, a devastating diarrheal disease caused by *Vibrio cholerae* that has been responsible for seven global pandemics and much suffering over the past two centuries. Epidemic cholera remains a significant public-health concern in the developing world today. Other vibrioses caused by other *Vibrio* species include syndromes of diarrhea, soft tissue infection, or primary sepsis. All *Vibrio* species are highly motile, facultatively anaerobic, curved gram-negative rods with one or more flagella. In nature, vibrios most commonly reside in tidal rivers and bays under conditions of moderate salinity. They proliferate in the summer months when water temperatures exceed 20°C. As might be expected, the illnesses they cause also increase in frequency during the warm months.

CHOLERA

DEFINITION

Cholera is an acute diarrheal disease that can, in a matter of hours, result in profound, rapidly progressive dehydration and death. Accordingly, *cholera gravis* (the severe form) is a much-feared disease, particularly in its epidemic presentation. Fortunately, prompt aggressive fluid repletion and supportive care can obviate the high mortality that is historically associated with cholera. Although the term *cholera* has occasionally been applied to any severely dehydrating secretory diarrheal illness, whether infectious in etiology or not, it now refers to disease caused by *V. cholerae* serogroup O1 or O139—i.e., the serogroups with epidemic potential.

MICROBIOLOGY AND EPIDEMIOLOGY

The species *V. cholerae* is classified into >200 serogroups based on the carbohydrate determinants of their lipopolysaccharide (LPS) O antigens. Although some non-O1 *V. cholerae* serogroups (strains that do not agglutinate in antisera to the O1 group antigen) have occasionally caused sporadic outbreaks of diarrhea, serogroup O1 was, until the emergence of serogroup O139 in 1992 (see below), the exclusive cause of epidemic cholera. Two biotypes of *V. cholerae* O1, classical and El Tor, are distinguished. Each biotype is further subdivided into two serotypes, termed *Inaba* and *Ogawa*.

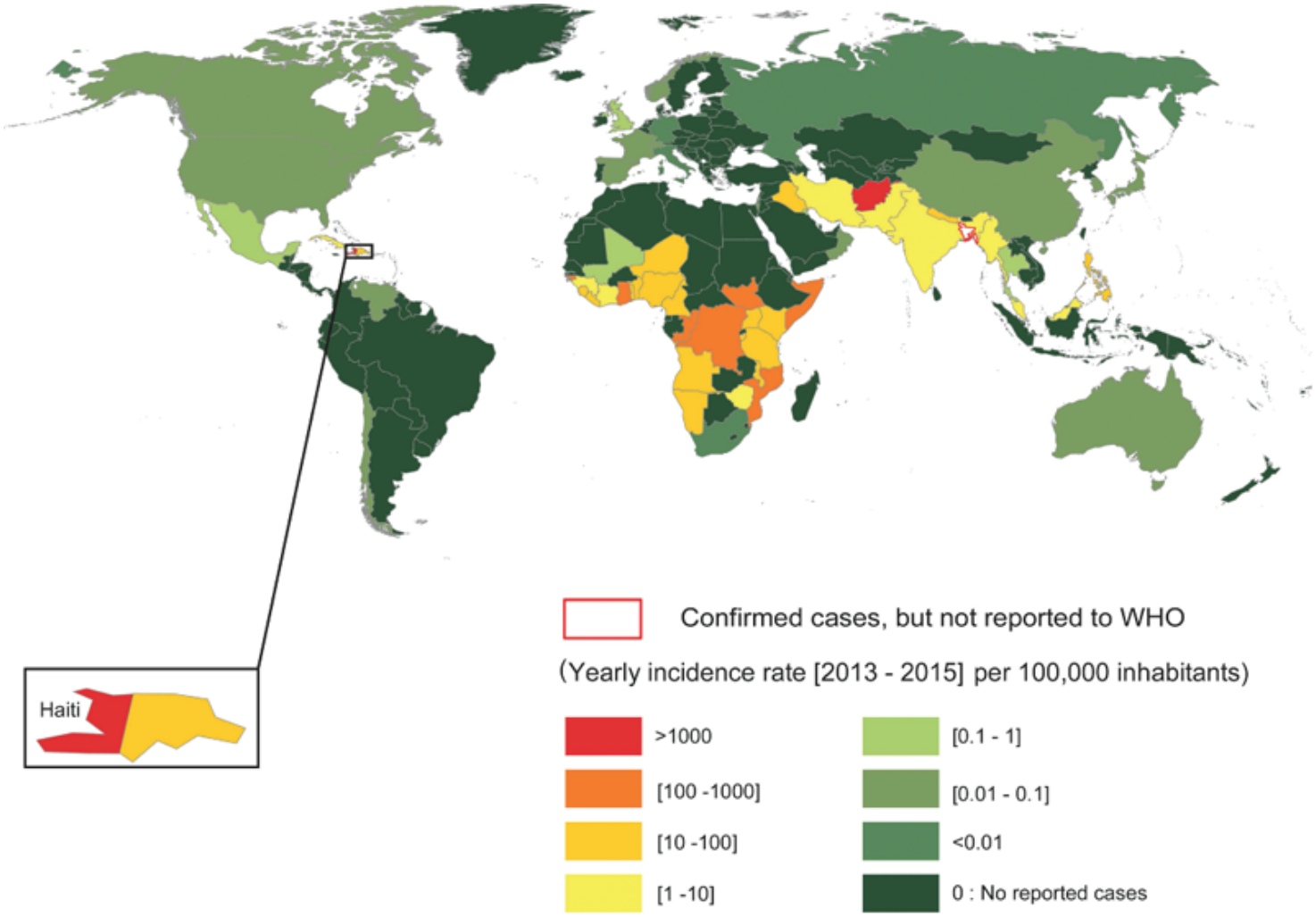
The natural habitat of *V. cholerae* is coastal salt water and brackish estuaries, where the organism lives in close relation to plankton. *V. cholerae* can also exist in freshwater in the presence of adequate nutrients and warmth. Humans become infected incidentally but, once infected, can act as vehicles for spread. Ingestion of water contaminated by human feces is the most common means of acquisition of *V. cholerae*. Consumption of contaminated food also can contribute to spread. There is no known animal reservoir. Although the infectious dose is relatively high, it is markedly

reduced in hypochlorhydric persons, in those using antacids, and when gastric acidity is buffered by a meal. Cholera is predominantly a pediatric disease in endemic areas, but it affects adults and children equally when newly introduced into a population. In endemic areas, the burden of disease is often greatest during “cholera seasons” associated with high temperatures, heavy rainfall, and flooding, but cholera can occur year-round. For unexplained reasons, susceptibility to cholera is significantly influenced by ABO blood group status; persons with type O blood are at greatest risk of severe disease if infected, whereas those with type AB are at least risk.

Cholera is native to the Ganges delta on the Indian subcontinent. Since 1817, seven global pandemics have occurred. The current (seventh) pandemic—the first due to the El Tor biotype—began in Indonesia in 1961 and spread in serial waves throughout Asia as *V. cholerae* El Tor displaced the endemic classical biotype, which is thought to have caused the previous six pandemics. In the early 1970s, El Tor cholera erupted in Africa, causing major epidemics before becoming a persistent endemic problem. Currently, >40% of cholera cases reported annually to the World Health Organization (WHO) are from Africa, >35% are from Asia, and >20% are from the Americas **(Fig. 163-1)**, but the true burden and distribution of cholera are unknown because the diagnosis is often syndromic and because many countries with endemic cholera do not report cholera to the WHO. It is possible that >2–3 million cases of cholera occur yearly (of which only ~200,000 are reported to the WHO) and that these cases result in >50,000–100,000 deaths annually (of which <2000 are reported to the WHO).

FIGURE 163-1

World distribution of cholera in 2013–2015. WHO, World Health Organization. *(Courtesy of Drs. M. and R. Piarroux, Université de la Méditerranée, France, with permission; map generated with Philcarto [http://philcarto.free.fr].)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

After a century without cholera in Latin America, the current cholera pandemic reached Central and South America in 1991. Following an initial explosive spread that affected millions, the burden of disease has markedly decreased in Latin America. In 2010, a severe cholera outbreak began in Haiti, a country with no recorded history of this disease. Several lines of evidence indicate that cholera was likely introduced into Haiti by United Nations security forces from Asia, raising the possibility that asymptomatic carriers of *V. cholerae* play an important role in transmitting cholera over long distances. To date, the outbreak has involved >800,000 individuals, resulting in thousands of deaths. The recent history of cholera has been punctuated by such severe outbreaks, especially among impoverished or displaced persons. These outbreaks are often precipitated by war or other circumstances that lead to the breakdown of public health measures. Such was the case in the camps for Rwandan refugees set up in 1994 around Goma, Zaire; in 2008–2009 in Zimbabwe; in 2015 in South Sudan and the Democratic Republic of the Congo; and in 2017 in Yemen.

Sporadic endemic infections due to *V. cholerae* O1 strains related to the seventh-pandemic strain have been recognized along the U.S. Gulf Coast of Louisiana and Texas. These infections are typically associated with the consumption of contaminated, locally harvested shellfish. Occasionally, cases in U.S. locations remote from the Gulf Coast have been linked to shipped-in Gulf Coast seafood.

In October 1992, a large-scale outbreak of clinical cholera caused by a new serogroup, O139, occurred in southeastern India. The organism appears to be a derivative of El Tor O1 but has a distinct LPS and an immunologically related O-antigen polysaccharide capsule. (O1 organisms are not encapsulated.) After an initial spread across 11 Asian countries,

V. cholerae O139 has once again been almost entirely replaced by O1 strains. The clinical manifestations of disease caused by *V. cholerae* O139 are indistinguishable from those of O1 cholera. Immunity to one, however, is not protective against the other.

PATHOGENESIS

In the final analysis, cholera is a toxin-mediated disease. The watery diarrhea characteristic of cholera is due to the action of cholera toxin, a potent protein enterotoxin elaborated by the organism in the small intestine. The toxin-coregulated pilus (TCP), so named because its synthesis is regulated in parallel with that of cholera toxin, is essential for *V. cholerae* to survive and multiply in (colonize) the small intestine. Cholera toxin, TCP, and several other virulence factors are coordinately regulated by ToxR. This protein modulates the expression of genes coding for virulence factors in response to environmental signals via a cascade of regulatory proteins. Additional regulatory processes, including bacterial responses to the density of the bacterial population (in a phenomenon known as *quorum sensing*), modulate the virulence of *V. cholerae*.

Once established in the human small bowel, the organism produces cholera toxin, which consists of a monomeric enzymatic moiety (the A subunit) and a pentameric binding moiety (the B subunit). The B pentamer binds to GM_1 ganglioside, a glycolipid on the surface of epithelial cells that serves as the toxin receptor and makes possible the delivery of the A subunit to its cytosolic target. The activated A subunit (A_1) irreversibly transfers ADP-ribose from nicotinamide adenine dinucleotide to its specific target protein, the GTP-binding regulatory component of adenylate cyclase. The ADP-ribosylated G protein upregulates the activity of adenylate cyclase; the result is the intracellular accumulation of high levels of cyclic adenosine monophosphate (AMP). In intestinal epithelial cells, cyclic AMP inhibits the absorptive sodium-transport system in villus cells and activates the secretory chloride-transport system in crypt cells, and these events lead to the accumulation of sodium chloride in the intestinal lumen. Because water moves passively to maintain osmolality, isotonic fluid accumulates in the lumen. When the volume of that fluid exceeds the capacity of the rest of the gut to resorb it, watery diarrhea results. Unless the wasted fluid and electrolytes are adequately replaced, shock (due to profound dehydration) and acidosis (due to loss of bicarbonate) follow. Although perturbation of the adenylate cyclase pathway is the primary mechanism by which cholera toxin causes excess fluid secretion, cholera toxin also enhances intestinal secretion via prostaglandins and/or neural histamine receptors.


The *V. cholerae* genome comprises two circular chromosomes. Lateral gene transfer has played a key role in the evolution of epidemic *V. cholerae*. The genes encoding cholera toxin (*ctxAB*) are part of the genome of a bacteriophage, CTXΦ. The receptor for this phage on the *V. cholerae* surface is the intestinal colonization factor TCP. Because *ctxAB* is part of a mobile genetic element (CTXΦ), horizontal transfer of this bacteriophage may account for the emergence of new toxigenic *V. cholerae* serogroups. Many of the other genes important for *V. cholerae* pathogenicity, including the genes encoding the biosynthesis of TCP, those encoding accessory colonization factors, and those regulating virulence gene expression, are clustered together in the *V. cholerae* pathogenicity island. Similar clustering of virulence genes is found in other bacterial pathogens. It is believed that pathogenicity islands are acquired by horizontal gene transfer. *V. cholerae* O139 is probably derived from an El Tor O1 strain that acquired the genes for O139 O-antigen synthesis by horizontal gene transfer.

CLINICAL MANIFESTATIONS

Individuals infected with *V. cholerae* O1 or O139 exhibit a range of clinical manifestations. Some individuals are asymptomatic or have only mild diarrhea; others present with the sudden onset of explosive and life-threatening diarrhea (*cholera gravis*). The reasons for the range in signs and symptoms of disease are incompletely understood but include the level of preexisting immunity, blood type, and nutritional status. In a nonimmune individual, after a 24- to 48-h incubation period, cholera characteristically begins with the sudden onset of painless watery diarrhea that may quickly become voluminous. Patients often vomit. In severe cases, volume loss can exceed 250 mL/kg in the first 24 h. If fluids and electrolytes are not replaced, hypovolemic shock and death may ensue. Fever is usually absent. Muscle cramps due to electrolyte disturbances are common. The stool has a characteristic appearance: a nonbilious, gray, slightly cloudy fluid with flecks of mucus, no blood, and a somewhat fishy, inoffensive odor. It has been called "rice-water" stool because of its resemblance to the water in which rice has been washed **(Fig. 163-2)**. Clinical symptoms parallel volume contraction: at losses of <5% of normal body weight, thirst develops; at 5–10%, postural hypotension, weakness, tachycardia, and decreased skin turgor are documented; and at >10%, oliguria, weak or absent pulses, sunken eyes (and, in infants, sunken fontanelles), wrinkled ("washerwoman") skin, somnolence, and coma are characteristic. Complications derive exclusively from the effects of volume and electrolyte depletion and include renal failure due to acute tubular necrosis. Thus, if the patient is adequately treated with fluid and electrolytes, complications are averted and the process is self-limited, resolving in a few days.

FIGURE 163-2

Rice-water cholera stool. Note floating mucus and gray watery appearance. *(Courtesy of Dr. A. S. G. Faruque, International Centre for Diarrhoeal Disease Research, Dhaka; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Laboratory data usually reveal an elevated hematocrit (due to hemoconcentration) in nonanemic patients; mild neutrophilic leukocytosis; elevated levels of blood urea nitrogen and creatinine consistent with prerenal azotemia; normal sodium, potassium, and chloride levels; a markedly reduced bicarbonate level (<15 mmol/L); and an elevated anion gap (due to increases in serum lactate, protein, and phosphate). Arterial pH is usually low (~7.2).

DIAGNOSIS

Cholera should be suspected when a patient ≥5 years of age develops acute watery diarrhea in an area known to have cholera or develops severe dehydration or dies from acute watery diarrhea, even in an area where cholera is not known to be present. The clinical suspicion of cholera can be confirmed by the identification of *V. cholerae* in stool; however, the organism must be specifically sought. With experience, it can be detected directly by dark-field microscopy on a wet mount of fresh stool, and its serotype can be discerned by immobilization with specific antiserum. Laboratory isolation of the organism requires the use of a selective medium such as taurocholate–tellurite–gelatin (TTG) agar or thiosulfate–citrate–bile salts–sucrose (TCBS) agar. If a delay in sample processing is expected, Carey-Blair transport medium and/or alkaline-peptone water-enrichment medium may be used as well. In endemic areas, there is little need for biochemical confirmation and characterization, although these tasks may be worthwhile in places where *V. cholerae* is an uncommon isolate. Standard microbiologic biochemical testing for Enterobacteriaceae will suffice for identification of *V. cholerae*. All vibrios are oxidase-positive. A point-of-care antigen-detection cholera dipstick assay is now commercially available for use in the field or where laboratory facilities are lacking.

TREATMENT

TREATMENT

Cholera

Death from cholera is due to hypovolemic shock; thus treatment of individuals with cholera first and foremost requires fluid resuscitation and management. In light of the level of dehydration **(Table 163-1)** and the patient's age and weight, euvolemia should first be rapidly restored, and adequate hydration should then be maintained to replace ongoing fluid losses **(Table 163-2)**. Administration of oral rehydration solution (ORS) takes advantage of the hexose-Na^+ co-transport mechanism to move Na^+ across the gut mucosa together with an actively transported molecule such as glucose (or galactose). Cl^- and water follow. This transport mechanism remains intact even when cholera toxin is active. ORS may be made by adding safe water to prepackaged sachets containing salts and sugar or by adding 0.5 teaspoon (i.e., a small spoonful) of table salt and 6 level teaspoons (i.e., 6 small spoonfuls) of table sugar to 1 L of safe water. Potassium intake in bananas or green coconut water should be encouraged. A number of ORS formulations are available, and the WHO now recommends "low-osmolarity" ORS for treatment of individuals with dehydrating diarrhea of any cause **(Table 163-3)**. If available, rice-based ORS is considered superior to standard ORS in the treatment of cholera. ORS can be administered via a nasogastric tube to individuals who cannot ingest fluid; however, optimal management of individuals with severe dehydration includes the administration of IV fluid and electrolytes. Because profound acidosis (pH <7.2) is common in this group, Ringer's lactate is the best choice among commercial products **(Table 163-4)**; it must be used with additional potassium supplements, preferably given by mouth. The total fluid deficit in severely dehydrated patients (>10% of body weight) can be replaced safely within the first 3–4 h of therapy, half within the first hour. Transient muscle cramps and tetany are common. Thereafter, oral therapy can usually be initiated, with the goal of maintaining fluid intake equal to fluid output. However, patients with continued large-volume diarrhea may require prolonged IV treatment to match gastrointestinal fluid losses. Severe hypokalemia can develop but will respond to

potassium given either IV or orally. In the absence of adequate staff to monitor the patient's progress, the oral route of rehydration and potassium replacement is safer than the IV route.

Although not necessary for cure, the use of an antibiotic to which the organism is susceptible diminishes the duration and volume of fluid loss and hastens clearance of the organism from the stool. Adjunctive antibiotics should therefore be administered to patients with moderate or severe dehydration due to cholera. In many areas, macrolides such as erythromycin (adults, 250 mg orally four times a day for 3 days; children, 12.5 mg/kg per dose four times a day for 3 days) or azithromycin (adults, a single 1-g dose; children, a single 20-mg/kg dose) are the agents of choice. Increasing resistance to tetracyclines is widespread; however, in areas with confirmed susceptibility, tetracycline (nonpregnant adults, 500 mg orally four times a day for 3 days; children >8 years old, 12.5 mg/kg per dose four times a day for 3 days) or doxycycline (nonpregnant adults, a 300-mg single dose; children >8 years old, a single dose of 4–6 mg/kg) may be used. Similarly, increasing resistance to fluoroquinolones is being reported, but in areas with confirmed susceptibility, a fluoroquinolone such as ciprofloxacin may be used (adults, 500 mg twice a day for 3 days; children, 15 mg/kg twice a day for 3 days). Oral administration of supplemental zinc is associated with decreased volume and severity of diarrhea in young children, including in those with cholera. Children <6 months of age with cholera should be treated with 10 mg of zinc daily for 10 days; children from 6 to <60 months of age should be treated with 20 mg of oral zinc daily for 10 days.

TABLE 163-1

Assessing the Degree of Dehydration in Patientswith Cholera

Degree of Dehydration	Clinical Findings
None or mild, but diarrhea	Thirst in some cases; <5% loss of total body weight
Moderate	Thirst, postural hypotension, weakness, tachycardia, decreased skin turgor, dry mouth/tongue, no tears; 5–10% loss of total body weight
Severe	Unconsciousness, lethargy, or "floppiness"; weak or absent pulse; inability to drink; sunken eyes (and, in infants, sunken fontanelles); >10% loss of total body weight

TABLE 163-2

Treatment of Cholera, Based on Degree of Dehydration[a]

Degree of Dehydration, Patient's Age (Weight)	Treatment[b]
None or Mild, but Diarrhea[c]	
<2 years	1/4–1/2 cup (50–100 mL) of ORS, to a maximum of 0.5 L/d
2–9 years	1/2–1 cup (100–200 mL) of ORS, to a maximum of 1 L/d
≥10 years	As much ORS as desired, to a maximum of 2 L/d
Moderate[c,d]	
<4 months (<5 kg)	200–400 mL of ORS
4–11 months (5–<8 kg)	400–600 mL of ORS
12–23 months (8–<11 kg)	600–800 mL of ORS
2–4 years (11–<16 kg)	800–1200 mL of ORS
5–14 years (16–<30 kg)	1200–2200 mL of ORS
≥15 years (≥30 kg)	2200–4000 mL of ORS
Severe[c]	
All ages and weights	Undertake IV fluid replacement with Ringer's lactate (or, if not available, normal saline). Give 100 mL/kg in the first 3-h period (or the first 6-h period for children <12 months old); start rapidly, then slow down. Give a total of 200 mL/kg in the first 24 h. Continue until the patient is awake, can ingest ORS, and no longer has a weak pulse.

[a]Adapted from World Health Organization: First steps for managing an outbreak of acute diarrhoea. Global Task Force on Cholera Control, 2009 (updated 2010; *http://www.who.int/cholera/publications/firststeps/en/*). [b]Continue normal feeding during treatment. [c]Reassess regularly; monitor stool and vomit output. [d]Volumes of ORS listed should be given within the first 4 h.

Abbreviation: ORS, oral rehydration solution.

TABLE 163-3

Composition of World Health Organization Reduced-Osmolarity Oral Rehydration Solution (ORS)[a,b]

Constituent	Concentration, mmol/L
Na^+	75
K^+	20
Cl^-	65
Citrate[c]	10
Glucose	75
Total osmolarity	245

[a]Contains (per package, to be added to 1 L of drinking water): NaCl, 2.6 g; $Na_3C_6H_5O_7 \cdot 2H_2O$, 2.9 g; KCl, 1.5 g; and glucose (anhydrous), 13.5 g. [b]If prepackaged ORS is unavailable, a simple homemade alternative can be prepared by combining 3.5 g (~1/2 teaspoon) of NaCl with either 50 g of precooked rice cereal or 6 teaspoons of table sugar (sucrose) in 1 L of drinking water. In that case, potassium must be supplied separately (e.g., in orange juice or coconut water). [c]10 mmol of citrate per liter, which supplies 30 mmol of HCO_3/L.

TABLE 163-4

Electrolyte Composition of Cholera Stool and of Intravenous Rehydration Solution

Substance	Concentration, mmol/L			
	Na^+	K^+	Cl^-	Base
Stool				
Adult	135	15	100	45
Child	100	25	90	30
Ringer's lactate	130	4[a]	109	28

[a]Potassium supplements, preferably administered by mouth, are required to replace the usual potassium losses from stool.

PREVENTION

Provision of safe water and of facilities for sanitary disposal of feces, improved nutrition, and attention to food preparation and storage in the household can significantly reduce the incidence of cholera. In addition, precautions should be taken to prevent the spread of cholera via infected and potentially asymptomatic persons from endemic to nonendemic regions of the world (as was probably the case in the ongoing outbreak in Haiti; see "Microbiology and Epidemiology," above).

Much effort has been devoted to the development of an effective cholera vaccine over the past few decades, with a particular focus on oral vaccine strains. In an attempt to maximize mucosal responses, two types of oral cholera vaccine have been developed: oral killed vaccines and live attenuated vaccines. Currently, three oral killed cholera vaccines have been prequalified by the WHO and are available internationally. BivWC (Shanchol™; Shantha Biotechnics, Hyderabad, India) contains several biotypes and serotypes of *V. cholerae* O1 and *V. cholerae* O139 without supplemental cholera toxin B subunit. A related vaccine is produced in South Korea (Euvichol™; Eubiologics, Seoul). WC-rBS (Dukoral®; Valneva, Lyon, France) contains several biotypes and serotypes of *V. cholerae* O1 supplemented with 1 mg of recombinant cholera toxin B subunit per dose. The vaccines are administered as a two- or three-dose regimen, with doses usually separated by 14 days. They provide ~60–85% protection for the first few months. Booster immunizations of WC-rBS are recommended after 2 years for individuals ≥6 years of age and after 6 months for children 2–5 years of age. For BivWC, which was developed more recently, no formal recommendation regarding booster immunizations exists. However, BivWC was associated with ~60% protection over 5 years among recipients of all ages in a study in Kolkata, India; the rate of protection among children ≤5 years of age approximated 40%. Models predict significant herd immunity when vaccination coverage rates exceed 50%. The killed vaccines have been safely administered among populations with high rates of HIV infection.

Oral live attenuated vaccines for *V. cholerae* O1 are also in development. These strains have in common their lack of the genes encoding cholera toxin. One such vaccine, CVD 103-HgR (Vaxchora™; PaxVax, Redwood City, CA), was approved

in 2016 by the U.S. Food and Drug Administration for use in travelers to cholera-endemic regions. The vaccine was 90 and 80% efficacious against severe cholera after experimental infection of North American volunteers 10 days and 90 days after vaccination, respectively. Vaxchora is approved for use in adults 18–64 years of age; no recommendations concerning the timing or need for booster vaccinations are currently available. Other live attenuated vaccine candidate strains have been prepared from El Tor and O139 *V. cholerae* and have been tested in studies of volunteers. An advantage of live attenuated cholera vaccines is that they may induce protection after a single oral dose; evaluation of single-dose regimens of oral killed cholera vaccines is underway. Conjugate and subunit cholera vaccines are also being developed. Recognizing that it may be decades before safe water and adequate sanitation become a reality for those most at risk of cholera, the WHO has recommended incorporation of cholera vaccination into comprehensive control strategies and has established an international stockpile of oral killed cholera vaccine to assist in outbreak responses. One million doses of oral killed cholera vaccine were released from this stockpile for use in Haiti following Hurricane Matthew in 2016. Since its inception in 2017, more than 5 million doses of oral cholera vaccine have been released from the global stockpile.

OTHER *VIBRIO* SPECIES



The genus *Vibrio* includes several human pathogens that do not cause cholera. Abundant in coastal waters throughout the world, noncholera vibrios can reach high concentrations in the tissues of filter-feeding mollusks. As a result, human infection commonly follows the ingestion of seawater or of raw or undercooked shellfish **(Table 163-5)**. Most noncholera vibrios can be cultured on blood or MacConkey agar, which contains enough salt to support the growth of these halophilic species. In the microbiology laboratory, the species of noncholera vibrios are distinguished by standard biochemical tests. The most important of these organisms are *Vibrio parahaemolyticus* and *Vibrio vulnificus*.

TABLE 163-5

Features of Selected Noncholera Vibrioses

Organism	Vehicle or Activity	Host at Risk	Syndrome
Vibrio parahaemolyticus	Shellfish, seawater	Normal	Gastroenteritis
	Seawater	Normal	Wound infection
Non-O1/O139 *Vibrio cholerae*	Shellfish, travel	Normal	Gastroenteritis
	Seawater	Normal	Wound infection, otitis media
Vibrio vulnificus	Shellfish	Immunosuppressed[a]	Sepsis, secondary cellulitis
	Seawater	Normal, immunosuppressed[a]	Wound infection, cellulitis
Vibrio alginolyticus	Seawater	Normal	Wound infection, cellulitis, otitis
	Seawater	Burned, other immunosuppressed	Sepsis

[a]Especially with liver disease or hemochromatosis.

Source: Table 161-3 in *Harrison's Principles of Internal Medicine*, 14th edition.

The two major types of syndromes for which these noncholera vibrios are responsible are gastrointestinal illness (due to *V. parahaemolyticus*, non-O1/O139 *V. cholerae*, *Vibrio mimicus*, *Vibrio fluvialis*, *Vibrio hollisae*, and *Vibrio furnissii*) and soft tissue infections (due to *V. vulnificus*, *Vibrio alginolyticus*, and *Vibrio damselae*). *V. vulnificus* is also a cause of primary sepsis in some compromised individuals.

SPECIES ASSOCIATED PRIMARILY WITH GASTROINTESTINAL ILLNESS

V. parahaemolyticus

Widespread in marine environments, the halophilic *V. parahaemolyticus* is the leading seafood-borne bacterial cause of enteritis worldwide. This species was originally implicated in enteritis in Japan in 1953, accounting for 24% of reported cases in one study—a rate that presumably was due to the common practice of eating raw seafood in that country. In the United States, common-source outbreaks of diarrhea caused by this organism have been linked to the consumption of undercooked or improperly handled seafood or of other foods contaminated by seawater. Since the mid-1990s, the incidence of *V. parahaemolyticus* infections has increased in several countries, including the United States. Serotypes O3:K6, O4:K68, and O1:K-untypable, which are genetically related to one another, account in part for

this increase. The enteropathogenicity of *V. parahaemolyticus* is linked to its ability to cause hemolysis on Wagatsuma agar (i.e., the *Kanagawa phenomenon*) via a thermostable direct hemolysin (Vp-TDH). Although the mechanisms by which the organism causes diarrhea are not fully defined, the genome sequence of *V. parahaemolyticus* contains two type III secretion systems, which directly inject toxic bacterial proteins into host cells. The activity of one of these secretion systems is required for intestinal colonization and virulence in animal models. *V. parahaemolyticus* should be considered a possible etiologic agent in all cases of diarrhea that can be linked epidemiologically to seafood consumption or to the sea itself. The incidence of *V. parahaemolyticus* infection in the United States may be increasing, with this species accounting for almost half of all *Vibrio* isolates reported in this country in 2014.

Infections with *V. parahaemolyticus* can result in two distinct gastrointestinal presentations. The more common of the two presentations (including nearly all cases in North America) is characterized by watery diarrhea, usually occurring in conjunction with abdominal cramps, nausea, and vomiting and accompanied in ~25% of cases by fever and chills. After an incubation period of 4 h to 4 days, symptoms develop and persist for a median of 3 days. Dysentery, the less common presentation, is characterized by severe abdominal cramps, nausea, vomiting, and bloody or mucoid stools. *V. parahaemolyticus* also causes rare cases of wound infection and otitis and very rare cases of sepsis.

Most cases of *V. parahaemolyticus*–associated gastrointestinal illness, regardless of the presentation, are self-limited. Fluid replacement should be stressed. The role of antimicrobial agents is uncertain, but they may be of benefit in moderate or severe disease. Doxycycline, fluoroquinolones, macrolides, or third-generation cephalosporins are usually used. Deaths are extremely rare among immunocompetent individuals. Severe infections are associated with underlying diseases, including diabetes, preexisting liver disease, iron-overload states, or immunosuppression.

Non-O1/O139 (Noncholera) *V. cholerae*

The heterogeneous non-O1/O139 *V. cholerae* organisms cannot be distinguished from *V. cholerae* O1 or O139 by routine biochemical tests but do not agglutinate in O1 or O139 antiserum. Non-O1/O139 strains have caused several well-studied food-borne outbreaks of gastroenteritis and have also been responsible for sporadic cases of otitis media, wound infection, and bacteremia; non-O1/O139 *V. cholerae* strains do not cause epidemics of cholera. Like other vibrios, non-O1/O139 *V. cholerae* organisms are widely distributed in marine environments. In most instances, recognized cases in the United States have been associated with the consumption of raw oysters or with recent travel. The broad clinical spectrum of diarrheal illness caused by these organisms is probably due to the group's heterogeneous virulence attributes.

In the United States, about half of all non-O1/O139 *V. cholerae* isolates are from stool samples. The typical incubation period for gastroenteritis due to these organisms is <2 days, and the illness lasts for ~2–7 days. Patients' stools may be copious and watery or may be partly formed, less voluminous, and bloody or mucoid. Diarrhea can result in severe dehydration. Many cases include abdominal cramps, nausea, vomiting, and fever. Like those with cholera, patients who are seriously dehydrated should receive oral or IV fluids; the value of antibiotics is not clear.

Extraintestinal infections due to non-O1/O139 *V. cholerae* commonly follow occupational or recreational exposure to seawater. Around 10% of non-O1/O139 *V. cholerae* isolates come from cases of wound infection, 10% from cases of otitis media, and 20% from cases of bacteremia (which is particularly likely to develop in patients with liver disease). Extraintestinal infections should be treated with antibiotics. Information to guide antibiotic selection and dosing is limited, but most strains are sensitive in vitro to tetracycline, ciprofloxacin, and third-generation cephalosporins.

SPECIES ASSOCIATED PRIMARILY WITH SOFT TISSUE INFECTION OR BACTEREMIA

(See also Chap. 124)

V. vulnificus

Infection with *V. vulnificus* is rare, but this organism is the most common cause of severe *Vibrio* infections in the United States. Like most vibrios, *V. vulnificus* proliferates in the warm summer months and requires a saline environment for growth. In the United States, infections in humans typically occur in coastal states between May and October and most commonly affect men >40 years of age. *V. vulnificus* has been linked to two distinct syndromes: primary sepsis, which usually occurs in patients with underlying liver disease, and primary wound infection, which generally affects people without underlying disease. (*Vulnificus* is Latin for "wound maker.") Some authors have suggested that *V. vulnificus* also causes gastroenteritis independent of other clinical manifestations. *V. vulnificus* is endowed with a number of virulence attributes, including a capsule that confers resistance to phagocytosis and to the bactericidal activity of human serum as well as a cytolysin. Measured as the 50% lethal dose in mice, the organism's virulence is considerably increased under conditions of iron overload; this observation is consistent with the propensity of *V. vulnificus* to infect patients who have hemochromatosis.

Primary sepsis most often develops in patients who have cirrhosis or hemochromatosis. However, *V. vulnificus* bacteremia can also affect individuals who have hematopoietic disorders or chronic renal insufficiency, those who are using immunosuppressive medications or alcohol, or (in rare instances) those who have no known underlying disease. After a median incubation period of 16 h, the patient develops malaise, chills, fever, and prostration. One-third of patients develop hypotension, which is often apparent at admission. Cutaneous manifestations develop in most cases (usually within 36 h of onset) and characteristically involve the extremities (the lower more often than the upper). In a common sequence, erythematous patches are followed by ecchymoses, vesicles, and bullae. In fact, sepsis and hemorrhagic bullous skin lesions suggest the diagnosis in appropriate settings. Necrosis and sloughing may also be evident. Laboratory studies reveal leukopenia more often than leukocytosis, thrombocytopenia, or elevated levels of fibrin-split products. *V. vulnificus* can be cultured from blood or cutaneous lesions. The mortality rate approaches 50%, with most deaths due to uncontrolled sepsis (Chap. 297). Accordingly, prompt treatment is critical and should include empirical antibiotic administration, aggressive debridement, and general supportive care. *V. vulnificus* is sensitive in vitro to a number of antibiotics, including tetracycline, fluoroquinolones, and third-generation cephalosporins. Data from animal models suggest that either a fluoroquinolone or the combination of a tetracycline and a third-generation cephalosporin should be used in the treatment of *V. vulnificus* septicemia.

V. vulnificus–associated soft tissue infection can complicate either a fresh or an old wound that comes into contact with seawater; the patient may or may not have underlying disease. After a short incubation period (4 h to 4 days; mean, 12 h), the disease begins with swelling, erythema, and (in many cases) intense pain around the wound. These signs and symptoms are followed by cellulitis, which spreads rapidly and is sometimes accompanied by vesicular, bullous, or necrotic lesions. Metastatic events are uncommon. Most patients have fever and leukocytosis. *V. vulnificus* can be cultured from skin lesions and occasionally from the blood. Prompt antibiotic therapy and debridement are usually curative.

V. alginolyticus

First identified as a pathogen of humans in 1973, *V. alginolyticus* occasionally causes eye, ear, and wound infections. This species is the most salt-tolerant of the vibrios and can grow in salt concentrations of >10%. Most clinical isolates come from superinfected wounds that presumably become contaminated at the beach. Although its severity varies, *V. alginolyticus* infection tends not to be serious and generally responds well to antibiotic therapy and drainage. A few

cases of otitis externa, otitis media, and conjunctivitis due to this pathogen have been described. Tetracycline treatment usually results in cure. *V. alginolyticus* is a rare cause of bacteremia in immunocompromised hosts.

ACKNOWLEDGMENT

The authors gratefully acknowledge the valuable contributions of Drs. Robert Deresiewicz and Gerald T. Keusch, coauthors of this chapter for previous editions.

FURTHER READING

Clemens JD et al: Cholera. Lancet 390:1539, 2017.

Kauffman RC et al: Single-cell analysis of the plasmablast response to *Vibrio cholerae* demonstrates expansion of cross-reactive memory B cells. MBio 7:e02021, 2016. [PubMed: 27999163]

Qadri F et al: Feasibility and effectiveness of oral cholera vaccine in an urban endemic setting in Bangladesh: A cluster randomised open-label trial. Lancet 386:1362, 2015.

Qadri F et al: Efficacy of a single-dose, inactivated oral cholera vaccine in Bangladesh. N Engl J Med 374:1723, 2016.

World Health Organization: Cholera, 2016. Wkly Epidemiol Rec 92:521, 2017.

Chapter 164: Brucellosis

Nicholas J. Beeching

DEFINITION

Brucellosis is a bacterial zoonosis transmitted directly or indirectly to humans from infected animals, predominantly domesticated ruminants and swine. The disease is known colloquially as *undulant fever* because of its remittent character. Although brucellosis commonly presents as an acute febrile illness, its clinical manifestations vary widely, and definitive signs indicative of the diagnosis may be lacking. Thus the clinical diagnosis usually must be supported by the results of bacteriologic and/or serologic tests.

ETIOLOGIC AGENTS

Human brucellosis is caused by strains of *Brucella*, a bacterial genus that was previously suggested, on genetic grounds, to comprise a single species, *B. melitensis*, with a number of biologic variants exhibiting particular host preferences. This view was challenged on the basis of detailed differences in chromosomal structure and host preference. The traditional classification into nomen species is now favored both because of these differences and because this classification scheme closely reflects the epidemiologic patterns of the infection. The nomen system recognizes *B. melitensis*, which is the most common cause of symptomatic disease in humans and for which the main sources are sheep, goats, and camels; *B. abortus*, which is usually acquired from cattle or buffalo; *B. suis*, which is generally acquired from swine but has one variant enzootic in reindeer and caribou and another in rodents; and *B. canis*, which is acquired most often from dogs. *B. ovis*, which causes reproductive disease in sheep, has not been clearly implicated in human disease, while rare human infections have been reported with *B. neotomae*, which is found in desert rodents. Two relatively new species, *B. ceti* and *B. pinnipedialis*, have been identified in marine mammals, including seals and dolphins. At least one case of laboratory-acquired human disease due to one of these species has been described, and several cases of natural human infection have been reported. As infections in marine mammals appear to be widespread, more cases of zoonotic infection in humans may be identified. Other newly reported species include *B. microti* (isolated from field voles), *B. papionis* (from baboons), *B. vulpis* (from foxes), and *B. inopinata* (from a patient with a breast implant). Additional novel strains have been described in diverse species, including frogs, bats, and various rodents, and the genus likely will expand further in forthcoming years. Moreover, it has become apparent that *Brucella* is closely related to the genus *Ochrobactrum*, which includes environmental bacteria sometimes associated with opportunistic infections. Genomics-based

studies are beginning to elucidate the pathway of evolution from free-living soil bacteria to highly successful intracellular pathogens.

All brucellae are small, gram-negative, unencapsulated, nonsporulating rods or coccobacilli. They grow aerobically on peptone-based medium incubated at 37°C; the growth of some types is improved by supplementary CO_2. In vivo, brucellae behave as facultative intracellular parasites. The organisms are sensitive to sunlight, ionizing radiation, and moderate heat; they are killed by boiling and pasteurization but are resistant to freezing and drying. Their resistance to drying renders brucellae stable in aerosol form, facilitating airborne transmission. The organisms can survive for up to 2 months in soft cheeses made from goat's or sheep's milk; for at least 6 weeks in dry soil contaminated with infected urine, vaginal discharge, or placental or fetal tissues; and for at least 6 months in damp soil or liquid manure kept in cool dark conditions. Brucellae are easily killed by a wide range of common disinfectants used under optimal conditions but are likely to be much more resistant at low temperatures or in the presence of heavy organic contamination.

EPIDEMIOLOGY


Brucellosis is a zoonosis whose occurrence and control are closely related to its prevalence in domesticated animals. Its distribution is worldwide apart from the few countries where it has been eradicated from the animal reservoir. The true global prevalence of human brucellosis is unknown because of the imprecision of diagnosis and the inadequacy of reporting and surveillance systems in many countries. Recently, there has been increased recognition of the high incidence of brucellosis in India and parts of China and of importations to countries in Oceania, such as Fiji, and in Asia, such as Thailand and Vietnam. In Europe, the incidence of brucellosis in a country is inversely related to gross domestic product, and, in both developed and less well-resourced settings, human brucellosis is related to rural poverty and inadequate access to medical care. Failure of veterinary control programs due to conflicts or for economic reasons contributes further to the emergence and re-emergence of disease, as seen currently in some eastern Mediterranean countries.

Even in well-resourced settings, the true incidence of brucellosis in domesticated animals may be 10–20 times higher than the reported figures. Bovine brucellosis has been the target of control programs in many parts of the world and has been eradicated from the cattle populations of much of northern Europe, Australia, New Zealand, and Canada, among other nations. Its incidence has been reduced to a low level in the United States and most Western European countries, with a varied picture in other parts of the world. Efforts to eradicate *B. melitensis* infection from sheep and goat populations have been much less successful. These efforts have relied heavily on vaccination programs, which have tended to fluctuate with changing economic and political conditions. In some countries (e.g., Israel), *B. melitensis* has caused serious outbreaks in cattle. Infections with *B. melitensis* still pose a major public health problem in Mediterranean countries; in western, central, and southern Asia; and in parts of Africa and South and Central America. Infections with *B. abortus* are common in cattle-rearing communities in African countries such as Kenya and Uganda.

Human brucellosis is usually associated with occupational or domestic exposure to infected animals or their products. Farmers, shepherds, goatherds, veterinarians, and employees in slaughterhouses and meat-processing plants in endemic areas are occupationally exposed to infection. Feral pig hunters are at risk of infection with *B. suis* in several countries, including Australia. Family members of individuals involved in animal husbandry may be at risk, although it is often difficult to differentiate food-borne infection from environmental contamination under these circumstances. Laboratory workers who handle cultures or infected samples also are at risk. Travelers and urban residents usually acquire the infection through consumption of contaminated foods. In countries that have eradicated the disease, new cases are most commonly acquired abroad. Dairy products, especially soft cheeses, unpasteurized milk, and ice cream, are the most frequently implicated sources of infection; raw meat and bone marrow may be sources under exceptional circumstances. Infections acquired through cosmetic treatments using materials of fetal origin have been reported. Person-to-person transmission is extremely rare, as is transfer of infection by blood or tissue donation. Although brucellosis is a chronic intracellular infection, there is no evidence for increased prevalence or severity among individuals with HIV infection or with immunodeficiency or immunosuppression of other etiologies.

Brucellosis may be acquired by ingestion, inhalation, or mucosal or percutaneous exposure. Accidental injection or ingestion of the live vaccine strains of *B. abortus* (S19 and RB51) and *B. melitensis* (Rev 1) can cause disease. *B. melitensis* and *B. suis* have historically been developed as biological weapons by several countries and could be exploited for bioterrorism **(Chap. S2)**. This possibility should be borne in mind in the event of sudden unexplained outbreaks.

IMMUNITY AND PATHOGENESIS

Exposure to brucellosis elicits both humoral and cell-mediated immune responses. The mechanisms of protective immunity against human brucellosis are presumed to be similar to those documented in laboratory animals, but such generalizations must be interpreted with caution. The response to infection and its outcome are influenced by the virulence, phase, and species of the infecting strain. Differences have been reported between *B. abortus* and *B. suis* in modes of cellular entry and subsequent compartmentalization and processing. Antibodies promote clearance of extracellular brucellae by bactericidal action and by facilitation of phagocytosis by polymorphonuclear and mononuclear phagocytes; however, antibodies alone cannot eradicate infection. Organisms taken up by macrophages and other cells can establish persistent intracellular infections. The key target cell is the macrophage, and bacterial mechanisms for suppressing intracellular killing and apoptosis result in very large intracellular populations. Opsonized bacteria are actively phagocytosed by neutrophilic granulocytes and by monocytes. In these and other cells, initial attachment takes place via specific receptors, including Fc, C3, fibronectin, and mannose-binding proteins. Opsonized—but not unopsonized—bacteria trigger an oxidative burst inside phagocytes. Unopsonized bacteria are internalized via similar receptors but at much lower efficiency. Smooth strains enter host cells via lipid rafts. Smooth lipopolysaccharide (LPS), β-cyclic glucan, and possibly an invasion–attachment protein (IalB) are involved in this process. Tumor necrosis factor α (TNF-α) produced early in the course of infection stimulates cytotoxic lymphocytes and activates macrophages, which can kill intracellular brucellae

(probably mainly through production of reactive oxygen and nitrogen intermediates) and may clear infection. However, virulent *Brucella* cells can suppress the TNF-α response, and control of infection in this situation depends on macrophage activation and interferon γ (IFN-γ) responses. Cytokines such as interleukin (IL) 12 promote production of IFN-γ, which drives T_H1-type responses and stimulates macrophage activation. Inflammatory cytokines, including IL-4, IL-6, and IL-10, downregulate the protective response. As in other types of intracellular infection, it is assumed that initial replication of brucellae takes place within cells of the lymph nodes draining the point of entry. Subsequent hematogenous spread may result in chronic localizing infection at almost any site, although the reticuloendothelial system, musculoskeletal tissues, and genitourinary system are most frequently targeted. Both acute and chronic inflammatory responses develop in brucellosis, and the local tissue response may include granuloma formation with or without necrosis and caseation. Abscesses may also develop, especially in chronic localized infection.

The determinants of pathogenicity of *Brucella* have not been fully characterized, and the mechanisms underlying the manifestations of brucellosis are incompletely understood. The organism is a “stealth” pathogen whose survival strategy is centered on several processes that avoid triggering innate immune responses and that permit survival within monocytic cells. These processes include evasion of intracellular destruction by restricting the fusion of type IV secretion system–dependent *Brucella*-containing vacuoles with lysosomal compartments, inhibition of apoptosis of infected mononuclear cells, and prevention of dendritic cell maturation, antigen presentation, and activation of naïve T cells. The smooth *Brucella* LPS, which has an unusual O-chain and core-lipid composition, has relatively low endotoxin activity and plays a key role in pyrogenicity and in resistance to phagocytosis and serum killing in the nonimmune host. In addition, LPS is believed to play a role in suppressing phagosome–lysosome fusion and diverting the internalized bacteria into vacuoles located in endoplasmic reticulum, where intracellular replication takes place. Specific exotoxins have not been isolated, but a type IV secretion system (VirB) that regulates intracellular survival and trafficking has been identified. In *B. abortus* this system can be activated extracellularly, but in *B. suis* it is activated (by low pH) only during intracellular growth. Brucellae then produce acid-stable proteins that facilitate the organisms’ survival in phagosomes and may enhance their resistance to reactive oxygen intermediates. A type III secretion system based on modified flagellar structures also has been inferred, although not yet confirmed. Virulent brucellae are resistant to defensins and produce a Cu-Zn superoxide dismutase that increases their resistance to reactive oxygen intermediates. A hemolysin-like protein may trigger the release of brucellae from infected cells.

CLINICAL FEATURES

Brucellosis almost invariably causes fever, which may be associated with profuse sweats, especially at night. In endemic areas, brucellosis may be difficult to distinguish from the many other causes of fever. However, two features recognized in the nineteenth century distinguish brucellosis from other tropical fevers, such as typhoid and malaria: (1) Left untreated, the fever of brucellosis shows an undulating pattern that persists for weeks before the commencement of an afebrile period that may be followed by relapse. (2) The fever of brucellosis is associated with musculoskeletal symptoms and signs in about one-half of all patients.

The clinical syndromes caused by the different nomen species are similar, although *B. melitensis* tends to be associated with a more acute and aggressive presentation and *B. suis* with focal abscess induction. *B. abortus* infections may be more insidious in onset and more likely to become chronic. *B. canis* infections are reported to present frequently with acute gastrointestinal symptoms.

The incubation period varies from 1 week to several months, and the onset of fever and other symptoms may be abrupt or insidious. In addition to experiencing fever and sweats, patients become increasingly apathetic and fatigued; lose appetite and weight; and have nonspecific myalgia, headache, and chills. Overall, the presentation of brucellosis often fits one of three patterns: febrile illness that resembles typhoid but is less severe; fever and acute monoarthritis, typically of the hip or knee, in a young child; and long-lasting fever, misery, and low-back or hip pain in an older man. In an endemic area (e.g., much of the Middle East), a patient with fever and difficulty walking into the clinic would be regarded as having brucellosis until it was proven otherwise.

Diagnostic clues in the patient's history include travel to an endemic area, employment in a diagnostic microbiology laboratory, consumption of unpasteurized milk products (including soft cheeses), contact with animals, accidental inoculation with veterinary *Brucella* vaccines, and—in an endemic setting—a history of similar illness in the family (documented in almost 50% of cases). Focal features are present in the majority of patients. The most common are musculoskeletal pain and physical findings in the peripheral and axial skeleton (~40% of cases). Osteomyelitis more commonly involves the lumbar and low thoracic vertebrae than the cervical and high thoracic spine. Individual joints that are most commonly affected by septic arthritis are the knee, hip, sacroiliac, shoulder, and sternoclavicular joints; the pattern may be one of monoarthritis or polyarthritis. Osteomyelitis may also accompany septic arthritis.

In addition to the usual causes of vertebral osteomyelitis or septic arthritis, the most important disease in the differential diagnosis is tuberculosis. This point influences the therapeutic approach as well as the prognosis, given that several antimicrobial agents used to treat brucellosis are also used to treat tuberculosis. Septic arthritis in brucellosis progresses slowly, starting with small pericapsular erosions. In the vertebrae, anterior erosions of the superior end plate are typically the first features to become evident, with eventual involvement and sclerosis of the whole vertebra. Anterior osteophytes eventually develop, but vertebral destruction or impingement on the spinal cord is rare and usually suggests tuberculosis **(Table 164-1)**.

TABLE 164-1

Radiology of the Spine: Differentiation of Brucellosis from Tuberculosis

	Brucellosis	Tuberculosis
Site	Lumbar and others	Dorsolumbar
Vertebrae	Multiple or contiguous	Contiguous
Diskitis	Late	Early
Body	Intact until late	Morphology lost early
Canal compression	Rare	Common
Epiphysitis	Anterosuperior (Pom's sign)	General: upper and lower disk regions, central, subperiosteal
Osteophyte	Anterolateral (parrot beak)	Unusual
Deformity	Wedging uncommon	Anterior wedge, gibbus
Recovery	Sclerosis, whole-body	Variable
Paravertebral abscess	Small, well-localized	Common and discrete loss, transverse process
Psoas abscess	Rare	More likely

Other systems may be involved in a manner that resembles typhoid. About one-quarter of patients have a dry cough, usually with few changes visible on the chest x-ray, although pneumonia, empyema, intrathoracic adenopathy, or lung abscess can occur. Sputum or pleural effusion cultures are rarely positive in such cases, which respond well to standard brucellosis treatment. One-quarter of patients have hepatosplenomegaly, and 10–20% have significant lymphadenopathy; the differential diagnosis includes glandular fever–like illness such as that caused by Epstein-Barr virus, *Toxoplasma*, cytomegalovirus, HIV, or *Mycobacterium tuberculosis*. Up to 10% of men have acute epididymo-orchitis, which must be distinguished from mumps and from surgical problems such as torsion. Prostatitis, inflammation of the seminal vesicles, salpingitis, and pyelonephritis all occur. There is an increased incidence of fetal loss among infected pregnant women,

although teratogenicity has not been described and the tendency toward abortion is much less pronounced in humans than in farm animals.

Neurologic involvement is common, with depression and lethargy whose severity may not be fully appreciated by either the patient or the physician until after treatment. A small proportion of patients develop lymphocytic meningoencephalitis that mimics neurotuberculosis, atypical leptospirosis, or noninfectious conditions and that may be complicated by intracerebral abscess, a variety of cranial nerve deficits, or ruptured mycotic aneurysms.

Endocarditis occurs in ~1% of cases, most often affecting the aortic valve (natural or prosthetic). Any site in the body may be involved in metastatic abscess formation or inflammation; the female breast and the thyroid gland are affected particularly often. Nonspecific maculopapular rashes and other skin manifestations are uncommon and are rarely noticed by the patient even if they develop.

DIAGNOSIS

Because the clinical picture of brucellosis is not distinctive, the diagnosis must be based on a history of potential exposure, a presentation consistent with the disease, and supporting laboratory findings. Results of routine biochemical assays are usually within normal limits, although serum levels of hepatic enzymes and bilirubin may be elevated. Peripheral leukocyte counts are usually normal or low, with relative lymphocytosis. Mild anemia may be documented. Thrombocytopenia and disseminated intravascular coagulation with raised levels of fibrinogen degradation products can develop. The erythrocyte sedimentation rate and C-reactive protein levels are often normal but may be raised.

In body fluids such as cerebrospinal fluid (CSF) or joint fluid, lymphocytosis and low glucose levels are the norm. Elevated CSF levels of adenosine deaminase cannot be used to distinguish tubercular meningitis, as they may also be found in brucellosis. Biopsied samples of tissues such as lymph node or liver may show noncaseating granulomas without acid/alcohol-fast bacilli. The radiologic features of bony disease develop late and are much more subtle than those of tuberculosis or septic arthritis of other etiologies, with less bone and joint destruction. Isotope scanning is more sensitive than plain x-ray and continues to give positive results long after successful treatment.

Isolation of brucellae from blood, CSF, bone marrow, or joint fluid or from a tissue aspirate or biopsy sample is definitive, and attempts at isolation are usually successful in 50–70% of cases. Duplicate cultures should be incubated for up to 6 weeks (in air and 10% CO_2, respectively). Concentration and lysis of buffy coat cells before culture may increase the isolation rate. Cultures in modern nonradiometric or similar signaling systems (e.g., Bactec) usually become positive within 7–10 days but should be maintained for at least 3 weeks before the results are declared negative. All cultures should be handled under containment conditions appropriate for dangerous pathogens. *Brucella* species may be misidentified as *Agrobacterium*, *Ochrobactrum*, or *Psychrobacter (Moraxella) phenylpyruvicus* by the gallery identification strips commonly used in the diagnostic laboratory. In recent years, matrix-assisted laser desorption ionization time-of-flight

spectrometry (MALDI-TOF MS) has emerged as a powerful tool in bacterial identification. The relative homogeneity of classical *Brucella* species makes identification beyond the genus level by routine approaches challenging, although further improvements may facilitate discrimination at the species level, particularly in reference laboratories. The place of this technique in routine diagnostic practice will depend on further refinements. Meanwhile, the author is aware of cases in which blood culture isolates have been identified incorrectly using MALDI-TOF MS.

The peripheral blood–based polymerase chain reaction has enormous potential to detect bacteremia, to predict relapse, and to exclude "chronic brucellosis." This method is probably more sensitive and is certainly quicker than blood culture, and it does not carry the attendant biohazard risk posed by culture. Nucleic acid amplification techniques are now quite widely used, although no single standardized procedure has been adopted. Primers for the spacer region between the genes encoding the 16S and 23S ribosomal RNAs (*rrs-rrl),* various outer-membrane protein–encoding genes, the insertion sequence *IS711*, and the protein BCSP31 are sensitive and specific. Blood and other tissues are the most suitable samples for analysis.

Serologic examination often provides the only positive laboratory findings in brucellosis. In acute infection, IgM antibodies appear early and are followed by IgG and IgA. All these antibodies are active in agglutination tests, whether performed by tube, plate, or microagglutination methods. The majority of patients have detectable agglutinins at this stage. As the disease progresses, IgM levels decline, and the avidity and subclass distribution of IgG and IgA change. The result is reduced or undetectable agglutinin titers. However, the antibodies are detectable by alternative tests, including the complement fixation test, Coomb's antiglobulin test, and enzyme-linked immunosorbent assay. There is no clear cutoff value for a diagnostic titer. Rather, serology results must be interpreted in the context of exposure history and clinical presentation. In endemic areas or in settings of potential occupational exposure, agglutinin titers of 1:320–1:640 or higher are considered diagnostic; in nonendemic areas, a titer of ≥1:160 is considered significant. Repetition of tests after 2–4 weeks may demonstrate a rising titer.

In most centers, the standard agglutination test (or derivatives such as the microagglutination test) is still the mainstay of serologic diagnosis. Some investigators rely on the Rose Bengal test, which has been only partially validated for human diagnostic use. Dipstick assays for anti-*Brucella* IgM are useful for the diagnosis of acute infection but are less sensitive for infection with symptoms of several months' duration. In an endemic setting, >90% of patients with acute bacteremia have standard agglutination titers of at least 1:320. Other screening tests are used in some centers.

Antibody to the *Brucella* LPS O chain—the dominant antigen—is detected by all the conventional tests that employ smooth *B. abortus* cells as antigen. Since *B. abortus* cross-reacts with *B. melitensis* and *B. suis*, there is no advantage in replicating the tests with these antigens. Cross-reactions also occur with the O chains of some other gram-negative bacteria, including *Yersinia enterocolitica* O:9, *Escherichia coli* O157, *Francisella tularensis*, *Salmonella enterica* group N, *Stenotrophomonas maltophilia*, and *Vibrio cholerae*. Cross-reactions do not occur with the cell-surface antigens of rough *Brucella* strains such as *B. canis* or *B. ovis*; serologic tests for these nomen species must employ an antigen prepared from either one. The live *B. abortus* vaccine strain RB51 does not elicit antibody responses in serologic tests that use smooth antigens,

and this fact must be taken into account if serologic tests are employed in attempts to identify or follow the course of infections in persons accidentally exposed to the vaccine.

TREATMENT

TREATMENT

Brucellosis

The broad aims of antimicrobial therapy are to treat and relieve the symptoms of current infection and to prevent relapse. Focal disease presentations may require specific intervention in addition to more prolonged and tailored antibiotic therapy. In addition, tuberculosis must always be excluded, or—to prevent the emergence of resistance—therapy must be tailored to specifically exclude drugs active against tuberculosis (e.g., rifampin used alone) or to include a full antituberculous regimen.

Early experience with streptomycin monotherapy showed that relapse was common; thus dual therapy with tetracyclines became the norm. This is still the most effective combination, but alternatives may be used, with the options depending on local or national policy about the use of rifampin for the treatment of nonmycobacterial infection. For the several antimicrobial agents that are active in vivo, efficacy can usually be predicted by in vitro testing. However, numerous *Brucella* strains show in vitro sensitivity to a whole range of antimicrobials that are therapeutically ineffective, including assorted β-lactams. Moreover, the use of fluoroquinolones remains controversial despite the good in vitro activity and white-cell penetration of most agents of this class. Low intravacuolar pH is probably a factor in the poor performance of these drugs.

For adults with acute nonfocal brucellosis (duration, <1 month), a 6-week course of therapy incorporating at least two antimicrobial agents is required. Complex or focal disease may necessitate ≥3 months of therapy. Adherence to the therapeutic regimen is very important, and poor adherence underlies almost all cases of apparent treatment failure; such failure is rarely due to the emergence of drug resistance, although increasing resistance to trimethoprim-sulfamethoxazole (TMP-SMX) has been reported at one center. There is good retrospective evidence that a 3-week course of two agents is as effective as a 6-week course for treatment and prevention of relapse in children, but this point has not yet been proven in prospective studies.

The gold standard for the treatment of brucellosis in adults is IM streptomycin (0.75–1 g daily for 14–21 days) together with doxycycline (100 mg twice daily for 6 weeks). In both clinical trials and observational studies, relapse follows such treatment in 5–10% of cases. The usual alternative regimen (and the current World Health Organization recommendation) is rifampin (600–900 mg/d) plus doxycycline (100 mg twice daily) for 6 weeks. The relapse/failure rate is ~10% in trial conditions but rises to >20% in many non-trial situations, possibly because doxycycline levels are reduced and clearance rates increased by concomitant rifampin administration. Patients who cannot tolerate or receive tetracyclines (children, pregnant women) can be given high-dose TMP-SMX instead (two or three standard-strength tablets twice daily for adults, depending on weight).

Increasing evidence supports the use of an aminoglycoside such as gentamicin (5–6 mg/kg per day for at least 2 weeks) instead of streptomycin. Shorter courses have been associated with high failure rates in adults. A 5- to 7-day course of therapy with gentamicin and a 3-week course of TMP-SMX may be adequate for children with uncomplicated disease, but prospective trials are still needed to support this recommendation. Early experience with fluoroquinolone monotherapy was disappointing, although it was suggested that ofloxacin or ciprofloxacin, given together with rifampin for 6 weeks, might be an acceptable alternative to the other 6-week regimens for adults. A substantial meta-analysis did not support the use of fluoroquinolones in first-line treatment regimens, and these drugs are not recommended by an expert consensus group (Ioannina) except in the context of well-designed clinical trials. However, a more recent meta-analysis is more supportive of the efficacy of these drugs, and an adequately powered prospective study will be needed to resolve their role in standard combination therapy. A triple-drug regimen—doxycycline and rifampin combined with an initial course of an aminoglycoside—was superior to double-drug regimens in a meta-analysis. The triple-drug regimen should be considered for all patients with complicated disease and for those for whom treatment adherence is likely to be a problem.

Significant neurologic disease due to *Brucella* species requires prolonged treatment (i.e., for 3–6 months), usually with ceftriaxone supplementation of a standard regimen. *Brucella* endocarditis is treated with at least three drugs (an aminoglycoside, a tetracycline, and rifampin), and many experts add ceftriaxone and/or a fluoroquinolone to reduce the need for valve replacement. Treatment is usually given for at least 4–6 months, and clinical end points for its discontinuation are often difficult to define. Surgery is still required for the majority of cases of infection of prosthetic heart valves and prosthetic joints.

There is no evidence base to guide prophylaxis after exposure to *Brucella* organisms (e.g., in the laboratory), inadvertent immunization with live vaccine intended for use in animals, or exposure to deliberately released brucellae. Most authorities have recommended the administration of rifampin plus doxycycline for 3 weeks after a low-risk exposure (e.g., an unspecified laboratory accident) and for 6 weeks after a major exposure to aerosol or injected material. However, such regimens are poorly tolerated, and doxycycline monotherapy of the same duration may be substituted. (Monotherapy is the standard recommendation in the United Kingdom but not in the United States.) Rifampin should be omitted after exposure to vaccine strain RB51, which is resistant to rifampin but sensitive to doxycycline. After significant brucellosis exposure, expert consultation is advised for women who are (or may be) pregnant.

PROGNOSIS AND FOLLOW-UP

Relapse occurs in up to 30% of poorly compliant patients. Thus patients should ideally be followed clinically for up to 2 years to detect relapse, which responds to a prolonged course of the same therapy used originally. The general well-being and the body weight of the patient are more useful guides than serology to lack of relapse. IgG antibody levels detected by the standard agglutination test and its variants can remain in the diagnostic range for >2 years after successful treatment. Complement fixation titers usually fall to normal within 1 year of cure. Immunity is not solid; patients can be reinfected after repeated exposures. Fewer than 1% of patients die of brucellosis. When the outcome is fatal, death is usually a consequence of cardiac

involvement; more rarely, it results from severe neurologic disease. Despite the low mortality rate, recovery from brucellosis is slow, and the illness can cause prolonged inactivity, with domestic and economic consequences.

The existence of a prolonged chronic brucellosis state after successful treatment remains controversial. Evaluation of patients in whom this state is considered (often those with work-related exposure to brucellae) includes careful exclusion of malingering, nonspecific chronic fatigue syndromes, and other causes of excessive sweating, such as alcohol abuse and obesity. In the future, the availability of more sensitive assays to detect *Brucella* antigen or DNA may help to identify patients with ongoing infection.

PREVENTION

Vaccines based on live attenuated *Brucella* strains, such as *B. abortus* strain 19BA or 104M, have been used in some countries to protect high-risk populations but have displayed only short-term efficacy and high reactogenicity. Subunit vaccines have been developed but are of uncertain value and cannot be recommended at present. Research in this area has been stimulated by interest in biodefense **(Chap. S2)** and may eventually yield new products. The mainstay of veterinary prevention is a national commitment to testing and slaughter of infected herds/flocks (with compensation for owners), control of animal movement, and active immunization of animals. These measures are usually sufficient to control human disease as well. In their absence, pasteurization of all milk products before consumption is sufficient to prevent non-occupational animal-to-human transmission. All cases of brucellosis in animals and humans should be reported to the appropriate public health authorities.

ACKNOWLEDGMENTS

Michael J. Corbel co-authored this chapter in several previous editions. The author thanks Adrian M. Whatmore for his review of the manuscript. Dr. Beeching is partially supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emerging and Zoonotic Infections, a partnership between the University of Liverpool, Liverpool School of Tropical Medicine, and Public Health England (PHE). The views expressed are those of the author and not necessarily those of the NHS, the NIHR, the Department of Health, or *Public Health England.*

FURTHER READING

Beeching NJ et al: Brucellosis. BMJ Best Practice, 2017. Available at *http://us.bestpractice.bmj.com.ezproxy1.library.usyd.edu.au/best-practice/monograph/911.html.*

Centers for Disease Control and Prevention: Brucellosis. Available at *https://www.cdc.gov/brucellosis/resources/index.html.*

Dean AS et al: Clinical manifestations of human brucellosis: A systematic review and metaanalysis. PLoS Negl Trop Dis 6:e1929, 2012. [PubMed: 23236528]

Norman FF et al: Imported brucellosis: A case series and literature review. Travel Med Infect Dis 14:182, 2016.

Yousefi-Nooraie R et al: Antibiotics for treating human brucellosis. Cochrane Database Syst Rev 10:CD007179, 2012. [PubMed: 23076931]

Chapter 165: Tularemia

Jeannine M. Petersen; Paul S. Mead

DEFINITION

Tularemia is a zoonotic disease caused by the bacterium *Francisella tularensis*. Human infection is rare but widespread and can be life-threatening. Sources of human infection include arthropod bites, agricultural aerosols, contaminated food or water, and contact with tissues of infected animals. Clinical diagnosis of tularemia can be challenging as the disease manifestations are diverse, with up to six distinct clinical syndromes. The prognosis is favorable when effective antimicrobial treatment is initiated early; however, complications are common if treatment is delayed.

ETIOLOGY

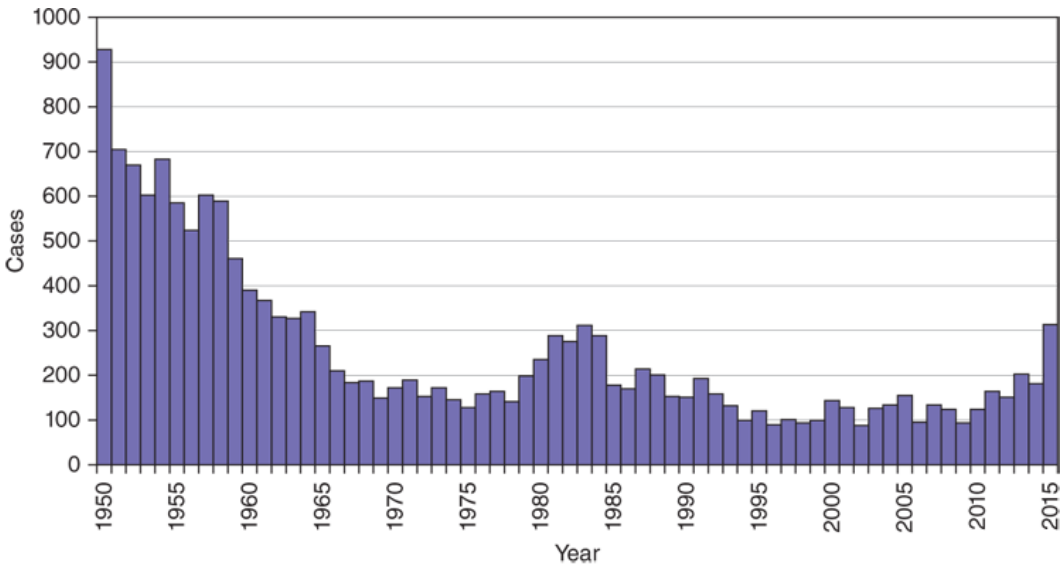
F. tularensis is a small (0.2 × 0.2–0.7 μm), aerobic, nonmotile, non-spore-forming, gram-negative coccobacillus. The bacterium is dependent on the invasion of host cells in vivo to multiply and cause disease. Genetically, it is not closely related to other known human pathogens. *F. tularensis* can enter the human body through the skin, mucous membranes, or respiratory tract. The infectious dose is low, with inhalation of 25 or fewer organisms sufficient to cause illness. Historically, *F. tularensis* was developed as a biological weapon and is currently classified as a Tier 1 select agent **(Chap. S2)**. Two subspecies of *F. tularensis*, subsp. *tularensis* (hereafter referred to as type A) and subsp. *holarctica* (hereafter referred to as type B) cause human tularemia in the United States.

EPIDEMIOLOGY

Tularemia occurs widely throughout the Northern Hemisphere. The disease is nationally notifiable in the United States, and cases have been reported in all U.S. states except Hawaii **(Fig. 165-1)**. States located in the south-central and midwestern regions—specifically, Arkansas, Kansas, Missouri, and Oklahoma—account for a disproportionate number of cases. Despite a substantial decrease during the mid-twentieth century, U.S. case counts have remained relatively stable since 1970 **(Fig. 165-2)**. During the 10-year period from 2006 through 2015, 93–314 cases were reported annually (average, 147 cases). The year 2015 marked a substantial increase in cases in Colorado, Wyoming, South Dakota, and Nebraska, with >100 cases reported among residents of these four states. The incidence was highest in Wyoming (35.8 cases per 1 million population), far higher than the national average of 1 case per 1 million population in 2015.

FIGURE 165-1

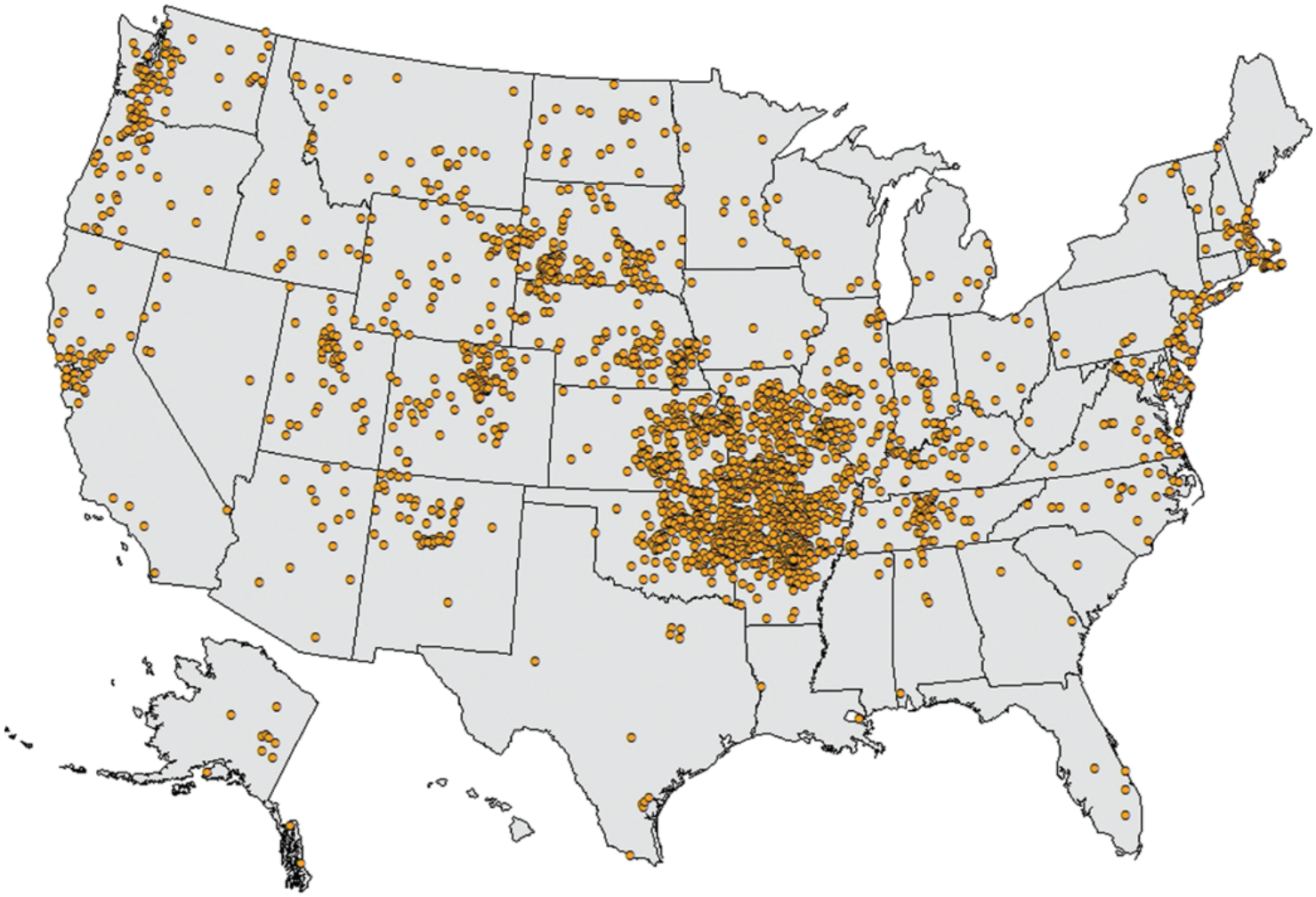
Yearly reported tularemia cases in the United States, 1950–2015.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 165-2

Map displaying reported U.S. tularemia cases in 2000–2015. One dot was plotted randomly within the county of residence for each reported case.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In nature, *F. tularensis* is maintained by arthropod and animal hosts. The bacterium is transmitted among animal hosts by arthropod bite or by direct exposure to contaminated materials in the environment. *F. tularensis* can infect and cause illness in an exceedingly broad variety of animals, having been isolated from >100 species, including domestic animals (cats and dogs). However, lagomorphs (wild hares and cottontail rabbits), terrestrial rodents (voles and meadow mice), aquatic rodents (muskrats and beavers), and ticks are thought to play a particularly significant role in propagating the organism in nature.

Humans are infected incidentally through various exposure routes—i.e., via the skin, mouth, lungs, or eyes (Table 165-1). Most commonly, *F. tularensis* enters through the skin by tick or deerfly bite or during handling of infected wildlife (e.g., while hunting or skinning). Domestic cats can transmit the bacterium by bite. The organism can be ingested by consumption of undercooked infected meat from wild animals or by drinking of untreated natural water contaminated by infected animals. It can directly enter the lungs upon inhalation of contaminated aerosols or dusts during farming or landscaping activities, especially when infected animals or carcasses are mowed over. Rarely, cases can occur by direct inoculation of contaminated materials into the mucous membranes of the eye. Person-to-person transmission of *F. tularensis* has not been reported. As in other zoonotic diseases, the risk of infection is associated with outdoor and occupational activities. Hunters, wildlife specialists, hikers, campers, veterinarians, and others with animal or arthropod exposure are at increased risk of infection. Laboratorians who handle cultures of *F. tularensis* without using personal protective equipment such as biosafety cabinets and N-95 respirators are also at elevated risk because of the high concentration of organisms being manipulated and the low infectious inhalation dose.

TABLE 165-1

Clinical Manifestations of Tularemia

Clinical Form	Portal of Entry	Transmission Source(s)	Symptoms[a]	Alternative Diagnoses
Ulceroglandular/glandular	Skin	Tick or deerfly bite; handling of infected animals	Fever; regional lymphadenopathy, with (ulceroglandular) or without (glandular) ulcer at site of entry	Cat-scratch disease, streptococcal or staphylococcal lymphadenitis, bubonic plague, lymphoma
Oropharyngeal	Mouth	Ingestion of contaminated food (undercooked game) or untreated water contaminated by infected animals	Fever; sore throat; marked unilateral cervical adenopathy; pharyngitis (which may be exudative)	Streptococcal infection, infectious mononucleosis, adenoviral infection
Oculoglandular	Eyes	Direct inoculation of contaminated materials into conjunctiva	Fever; unilateral conjunctivitis with mucopurulent discharge; eyelid swelling; regional lymphadenopathy	Cat-scratch disease
Pneumonic/respiratory[b]	Lungs	Inhalation of contaminated aerosols or agricultural dusts	Fever; dry paroxysmal cough; pleuritic or retrosternal pain and dyspnea	Mycoplasmal pneumonia, chlamydial pneumonia, Legionnaires' disease, pneumocystosis, histoplasmosis, tuberculosis, Q fever

[a]All clinical forms may also be marked by chills, headache, malaise, fatigue, myalgias, and arthralgias. [b]Pneumonia is also a secondary complication of other clinical forms of tularemia, in which the organism can spread to the lungs after initial replication at the entry site.

Tularemia cases are more common in the temperate months of May through September. Infections due to bites from infected ticks (*Dermacentor* and *Amblyomma* species) or deerflies (*Chrysops* species) occur during these months, whereas illness due to animal handling and hunting can develop at any time of the year. The principal animal sources of infection are the cottontail rabbit (*Sylvilagus* species), wild hares, and rodents (muskrats, beavers, voles). Human cases are most often sporadic and widespread; outbreaks are rare. Tularemia occurs more frequently in males than in females. Healthy persons of all ages are susceptible, with a higher incidence reported among children <10 years of age and middle-aged men. In children, the head and neck are often the primary sites of infection; ulcers or enlarged lymph nodes may be less obvious at these sites. The incidence of tularemia is roughly tenfold higher among some Native American populations, presumably because of higher rates of exposure.

The severity of clinical disease is influenced by the infecting strain, the route of exposure, and the length of delay before administration of effective treatment. Infections due to *F. tularensis* type A strains are typically more severe than those caused by type B strains. Although both of these subspecies cause tularemia throughout the United States, infections due to type B seem to predominate in the Pacific Northwest and along tributaries of the Mississippi, whereas infections due to type A prevail along the Atlantic seaboard and in south-central states. Virulence and geographic distribution differ among type A strains; subpopulations of strains localized in the eastern United States are associated with more severe disease.

GLOBAL CONSIDERATIONS

The distribution of *F. tularensis* type A is restricted to North America. In contrast, type B strains are widely dispersed throughout the Northern Hemisphere. *F. tularensis* type B, associated with opossums, has also been identified recently in Australia. Among Eurasian type B strains, a subpopulation displays natural resistance to erythromycin, whereas type B strains originating from the United States are uniformly susceptible to erythromycin. In Sweden, mosquito bite is a primary route of infection. This route of transmission is occasionally reported in other areas of northern and central Europe as well, but there is no evidence of mosquito-borne transmission of *F. tularensis* in North America. Water-borne cases of tularemia occur primarily in regions of the world where chlorinated municipal water systems are lacking, and water is an uncommon route of infection in the United States. A third subspecies of *F. tularensis*, subsp. *mediasiatica,* is localized to regions of central Asia, where it has been described in animals and ticks.

PATHOGENESIS

F. tularensis is an obligate intracellular pathogen that enters and replicates within the cytoplasm of various host cells, including (but not limited to) macrophages, dendritic cells, and polymorphonuclear neutrophils. Upon entry into the human body, the organism multiplies locally, producing a skin

ulcer after skin entry or destruction of bronchial tissues after inhalation and then spreading to local lymph nodes. Systemic spread can follow, with infected cells disseminating throughout the host to the liver, spleen, and lungs. Uncontrolled replication of *F. tularensis* leads to cell death, substantial tissue damage, and impairment of vital organs.

The ability of *F. tularensis* to survive and replicate within host cells is essential for the development of tularemia. The bacterium does not produce toxins. By evading recognition by the innate immune system, *F. tularensis* can enter and proliferate within host cells. The atypical lipopolysaccharide of *F. tularensis* plays a key role in this process, exhibiting reduced endotoxicity and therefore failing to stimulate the host's innate immunity. Rapid bacterial growth within host cells, followed by cell death, triggers a systemic inflammatory reaction that overwhelms the host defense system, culminating in extensive tissue injury.

Inflammatory cell infiltration and various degrees of necrosis are histopathologic hallmarks of lymph node involvement. In the early stages of disease, infected lymph nodes may be characterized by follicular hyperplasia and inflammatory cell infiltration, whereas granulomatous lesions with areas of focal caseous necrosis, which may resemble tuberculous nodes (granulomatous form), are detected later in disease. Histopathologic findings on autopsied tissues include granulomas and microabscesses in the liver and pyogranulomatous foci, often with central necrosis, in the spleen and lungs.

CLINICAL MANIFESTATIONS

Primary clinical manifestations of tularemia vary with the portal of entry, which may be through the skin (ulceroglandular, glandular), the eye (oculoglandular), the mouth (oropharyngeal), or the lung (pneumonic or respiratory) (Table 165-1). Systemic disease can follow entry by virtually any route. The incubation period for tularemia is typically 3–7 days (range, 1–14 days). A sudden onset of fever, with temperatures as high as 41°C (106°F), is characteristic in all forms of tularemia. All clinical forms of disease may also be marked by chills, headache, malaise, fatigue, myalgias, and arthralgias. Blood chemistries are of limited value in diagnosis, as infection is not commonly associated with distinctive changes. The white blood cell count may be normal or elevated, and the differential count usually shows a relative increase of mononuclear cells. A slight increase in C-reactive protein levels and liver enzymes may be observed.

Ulceroglandular/Glandular Tularemia

The most commonly reported clinical form is ulceroglandular tularemia, which follows a tick or deerfly bite or direct inoculation of the bacteria into the skin via handling of or biting by an infected animal. Initially, a small papule occurs at the site of organism entry, at or shortly after the onset of fever, and advances into an ulcer accompanied by painful regional lymphadenopathy in one or more adjacent lymph nodes. In ~30% of ulceroglandular tularemia cases, skin manifestations, including papular and maculopapular rash and erythema nodosum, are observed. Symptoms of glandular tularemia are similar to those of ulceroglandular tularemia, but without an ulcer. In both forms, suppuration of lymph nodes may occur in up to 40% of cases if there is a delay of >2 weeks in administration of an effective antibiotic. Children more often present with glandular disease, probably because of the increased frequency in this age group of the head and neck as the primary sites of infection; at these sites, ulcers may easily be overlooked or may resolve by the time tularemia is diagnosed.

Oropharyngeal and Oculoglandular Tularemia

Oropharyngeal infection occurs after ingestion of *F. tularensis* in contaminated water or inadequately cooked game meat and occasionally after inhalation of the organism. Patients present with fever, sore throat, marked cervical adenopathy (most often unilateral), and pharyngitis, which may be exudative or accompanied by a small ulcer. Oculoglandular tularemia results from entry of the organism into the eye, either by touching of the conjunctival sac with contaminated fingers or possibly by exposure to infectious aerosols. The patient typically presents with fever, unilateral conjunctivitis with mucopurulent discharge, eyelid swelling, and ulcers or pustules on the palpebral conjunctivae. Preauricular, submandibular, or cervical lymph nodes appear enlarged, red, and tender. In both clinical forms, lymph node suppuration may ensue if treatment is delayed.

Pneumonic Tularemia

Primary pneumonic tularemia develops from direct inhalation of *F. tularensis* and is the most severe form of the disease. Pneumonia is also a secondary complication of other clinical forms of tularemia in which the organism can spread to the lungs after initial replication at the entry site. Symptoms include dry paroxysmal cough, pleuritic or retrosternal pain, and dyspnea. Radiographic findings may include lobar and multilobar infiltrates, lung abscesses, and hilar adenopathy. Pleural effusions occur in ~20–30% of cases and appear to be exudative.

Typhoidal Tularemia

Typhoidal tularemia is a term that was originally used to designate those patients with systemic infections for which the portal of entry into the body was unclear. This designation dates back to the time when ingestion and inhalation were not recognized as routes of exposure to *F. tularensis.* The term is outdated and infrequently used.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Tularemia

As with many other rare diseases, failure to consider the diagnosis is the greatest obstacle to recognition and proper management. A careful history that reveals recent exposure to ticks or biting flies, hunting activity, or contact with secretions of domestic animals (e.g., cat bites, facial licking by dogs) is suggestive. Exposure to agricultural aerosols (i.e., during mowing or haying) should invariably trigger consideration of the diagnosis in patients with pneumonic involvement. Human cases often occur in the setting of an epizootic in which excess mortality among rabbits or rodents may have been noticed by the patient or a family member.

The differential diagnosis for tularemia is broad and reflects the diverse clinical forms of the infection (Table 165-1). In persons with glandular or ulceroglandular tularemia, alternative diagnoses include cat-scratch fever, streptococcal or staphylococcal lymphadenitis, bubonic plague, cutaneous anthrax, sporotrichosis, lymphoma, toxoplasmosis, mycobacterial infection, tick typhus, and viral infections such as herpetic whitlow or poxvirus infection. Involvement of the groin should lead to consideration of lymphogranuloma venereum, chancre, and chancroid. Oropharyngeal tularemia may be confused with stomatitis, pharyngitis, and cervical adenitis of other bacterial and viral etiologies, including infectious mononucleosis, streptococcal infection, adenoviral infection, mycobacterial infection, and diphtheria. Pneumonic tularemia can be severe but generally follows the pattern of atypical community-acquired pneumonia with scant sputum production and causes that include mycoplasmal and chlamydial pneumonia, Legionnaires' disease, Q fever, secondary pneumonic plague, pneumocystosis, histoplasmosis, and tuberculosis. Exudative pleural effusion is common, and cultures of pleural fluid can be helpful in confirming the diagnosis. Indolent cases presenting as weight loss, night sweats, and hilar adenopathy have been mistaken for malignancy. Conditions mimicking typhoidal tularemia include bacterial endocarditis, disseminated mycobacterial and fungal infection, typhoid fever, brucellosis, listeriosis, Q fever, and sepsis of other etiologies.

LABORATORY DIAGNOSIS

Confirmation of tularemia cases is based on recovery of an isolate from a clinical specimen or a rise in antibody titer between paired acute- and convalescent-phase serum specimens. The optimal specimen for diagnostic testing depends on the clinical form of disease, the duration of illness, and the treatment history. Serology is valuable for diagnosing all forms of tularemia, although it is of limited utility in acute illness because host immunologic responses to *F. tularensis* are not typically detectable until ~10 days after illness onset. Nonetheless, the diagnosis of tularemia may rely on serologic testing of convalescent-phase samples, as the slow-growing, fastidious characteristics of the organism reduce the likelihood of its recovery from clinical specimens. The primary serologic testing method used is an agglutination assay (micro- or tube), which detects total *F. tularensis* immunoglobulins.

For isolation of live bacteria in culture, collection of clinical specimens before antibiotic treatment is necessary. The appropriate specimens for culture are respiratory secretions, particularly pleural fluid or bronchial washes/aspirates (pneumonic disease); swabs of visible lesions or affected areas (ulceroglandular and oculoglandular disease); aspirates or biopsy samples of lymph nodes or lesions (ulceroglandular, glandular, oculoglandular, and oropharyngeal disease); and blood (all clinical forms). *F. tularensis* grows aerobically and requires cysteine-supplemented medium for optimal growth. Relevant media available in clinical laboratories include buffered charcoal yeast extract, Thayer-Martin, and chocolate agars. For growth of blood cultures, automated systems can be useful, with the caveat that the results may often be negative. If *F. tularensis* is suspected as the source of a patient's illness, the laboratory should be alerted so that cultures can be incubated for 7–10 days in order to maximize the chance of organism recovery. Notification of the laboratory that tularemia is suspected on clinical grounds is also critical for minimizing the risk to laboratory staff posed by aerosol exposure during the handling of cultures containing high concentrations of organisms. Because of the rarity of the organism, it often is not easily identified when it is cultured. On Gram's stain, it appears as a tiny gram-negative coccobacillus that counterstains faintly with safranin. Available *F. tularensis*–specific assays include the direct fluorescence assay (antigen detection) and polymerase chain reaction (PCR).

Misidentification of *F. tularensis* can occur, particularly when tularemia is not suspected. Commercial automated biochemical identification systems are not ideal for classification of *F. tularensis* because of the organism's slow growth and lack of biochemical reactivity. Some molecular methods may erroneously classify *Francisella novicida* as *F. tularensis* because of the high degree of genetic relatedness between the two organisms. In contrast to *F. tularensis*, *F. novicida* is an opportunistic pathogen, primarily causing illness in patients with an underlying compromising condition. Results from diagnostic testing should be interpreted in conjunction with the patient's symptoms and exposures.

TREATMENT

TREATMENT

Tularemia

Drugs approved by the U.S. Food and Drug Administration (FDA) for the treatment of tularemia include streptomycin, doxycycline, and tetracycline. Among these, streptomycin is the drug of choice on the basis of experience and lower relapse rates. Nevertheless, because streptomycin is difficult to obtain, gentamicin is often used as an acceptable alternative. Aminoglycosides are recommended for management of severe cases of tularemia; the primary drawback to their use is the potential for oto- and nephrotoxicity. The streptomycin dose is 1 g IM twice daily for adults and 15 mg/kg twice daily (up to a

maximal daily dose of 2 g) for children; the drug is given for 10 days. Gentamicin is given IV to adults in an initial dose of 1.5–2 mg/kg followed by 1–1.7 mg/kg IV or IM every 8 h or 5–7 mg/kg IV every 24 h for 10–14 days, depending on the nature and severity of the infection.

If aminoglycosides are contraindicated or are not readily available, tetracyclines are alternatives to streptomycin and gentamicin. Primary treatment failure and relapse occur at higher rates with tetracyclines, given their bacteriostatic mode of action; these agents are therefore recommended only for mild cases of tularemia. Tetracyclines should be administered for at least 14 days. For adults, the doxycycline dose is 100 mg by mouth twice daily and the tetracycline dose is 500 mg by mouth every 6 h.

Ciprofloxacin and other fluoroquinolones are not approved by the FDA for the treatment of tularemia. Nonetheless, they have displayed good efficacy for this indication in published case series, they are bactericidal, they are effective at low concentrations in vitro, and they have reasonably good tissue penetration. The Infectious Diseases Society of America's guidelines for treatment of skin and soft tissue infections due to *F. tularensis* recommend treatment with oral levofloxacin (500 mg daily) or ciprofloxacin (750 mg twice daily) for at least 14 days for mild to moderate cases.

Consensus-based treatment recommendations have been developed by the Working Group on Civilian Biodefense for use in the event that tularemia is used as a biological weapon. For postexposure prophylaxis, doxycycline and ciprofloxacin are the preferred choices and are administered for 14 days. Adults and children weighing >45 kg take 100 mg of doxycycline by mouth twice daily; children weighing <45 kg take 2.2 mg/kg by mouth twice daily. The oral dosages for ciprofloxacin in adults and children are 500 mg and 15 mg/kg twice daily, respectively.

β-Lactam antibiotics are ineffective for treatment of tularemia because of β-lactamase production by *F. tularensis* strains. Although the third-generation β-lactam ceftriaxone is active against *F. tularensis* in vitro, it is not used for management of tularemia cases because therapeutic failures are common. Likewise, macrolides are not used for treatment of tularemia because some Eurasian *F. tularensis* type B strains are naturally resistant to this class of antimicrobials.

Antimicrobial susceptibility testing of *F. tularensis* isolates worldwide indicates uniform susceptibility to antibiotics used for treatment of tularemia, including aminoglycosides, tetracyclines, and fluoroquinolones. Naturally occurring resistance to antibiotics used for treatment of tularemia has never been demonstrated in *F. tularensis*.

PROGNOSIS

Cases diagnosed early after illness onset respond rapidly to appropriate therapy, and fever most often abates within 24–72 h. Factors associated with poor outcome include a lack of timeliness in dispensing effective treatment, an underlying medical condition, the route of infection, and the infecting strain type. Prior to the advent of effective antimicrobials, mortality rates >60% were reported for primary pneumonic cases occurring via direct inhalation of *F. tularensis* type A strains. Among type A infections in the United States, those that occur along the Atlantic seaboard and in regions of the eastern United States are associated with the highest mortality rates. In Eurasia, where only *F. tularensis* type B causes tularemia, fatal cases are exceedingly rare, with or without antibiotic therapy. Overall, the expected mortality rate in the United States among appropriately treated patients is estimated at <1% overall and <5% for patients with pneumonic tularemia.

COMPLICATIONS

Complications of tularemia most often result from a delay in the initiation of treatment. Such a delay can be due to a failure to seek medical care until the later stages of disease progression, to a lack of clinical suspicion of tularemia, or to treatment with antibiotics ineffective against *F. tularensis*. Clinical chart review of 87 tularemia cases in Missouri (one of four states accounting for ~50% of U.S. cases) in 2000–2007 found that the disease was not suspected in more than half of cases until after the incidental isolation of *F. tularensis*.

The most commonly reported complication is suppuration of infected lymph nodes requiring surgical drainage(s). Other complications include hepatic abscesses, hepatitis, renal failure, rhabdomyolysis, adult respiratory distress syndrome, and loculated empyema requiring surgical decortication. Meningitis is a rare complication that can arise from untreated bacteremia, with only ~20 cases reported in the published literature; in two cases, cerebral microabscesses and brain lesions were documented. Other rare complications include endocarditis, pericarditis, and infections of prosthetic joints. Corneal edema and glaucoma have been reported as complications of oculoglandular tularemia.

TREATMENT FAILURES

Relapses are most often associated with the use of bacteriostatic antibiotics for shorter periods of treatment. The duration of treatment with a bacteriostatic agent needs to be sufficient to allow the development of a bactericidal cell-mediated host immune response to *F. tularensis*, which usually takes 2 weeks. Nonresponsiveness to antibiotics is most likely when there is a delay in the initiation of treatment after symptom onset. Poor antibiotic penetration of and accumulation in tissues, especially in the intracellular environment where *F. tularensis* multiplies, can also affect treatment efficacy. Lymph node suppuration, which is nonresponsive to all classes of antibiotics, occurs when disease progresses without the initiation of effective treatment; this nonresponsiveness to otherwise effective antibiotics is not due to resistance. For management of these cases, incision and surgical drainage or removal

are often required for clinical cure. As stated above, treatment failure due to resistance of *F. tularensis* to the antibiotics recommended for therapy has never been demonstrated.

PREVENTION

The most effective methods for preventing tularemia are based on the likely source of infection. Use of insect repellents (20–30% DEET [N,N-diethyl-meta-toluamide]) and the wearing of long pants, long sleeves, and long socks can reduce the risk of tick and deerfly bites. Arthropod exposure may be further reduced by the use of permethrin-treated clothing. If attached ticks are found on the body, they should be promptly removed with tweezers. During handling of potentially infected animals (carcasses, game), gloves should be worn to avoid bacterial invasion through cuts or abrasions on the hands. Game meat should be cooked thoroughly before eating and care taken to avoid cross-contamination from uncooked meat. Exclusion of rodents from food and water supplies and chlorination of drinking water can reduce the risk of oral ingestion. To limit the risk of inhalation of infectious aerosols, sick or dead animals should not be mowed over and the area should be checked for carcasses prior to mowing. Use of masks during mowing and other landscaping activities may also reduce the risk of inhaling the bacteria, but this measure has not been studied. To reduce the risk of laboratory-acquired infections, laboratory staff should be notified whenever tularemia is suspected. For management of patients with tularemia, standard hospital infection precautions are considered adequate, given that person-to-person transmission has not been documented. Bodies of patients who die of tularemia should be handled with standard precautions. Autopsy procedures likely to produce aerosols or droplets should be avoided. No vaccine is currently available for *F. tularensis*.

FURTHER READING

Centers for Disease Control and Prevention: Tularemia—United States, 2001–2010. MMWR Morb Mortal Wkly Rep 62:963, 2013.

Dennis DT et al: Tularemia as a biological weapon: Medical and public health management. JAMA 285:2763, 2001.

Fachinger P et al: Pulmonary tularaemia: All that looks like cancer is not necessarily cancer—Case report of four consecutive cases. BMC Pulm Med 15:27, 2015.

Stevens DL et al: Practice guidelines for the diagnosis and management of skin and soft tissue infections: 2014 update by the Infectious Diseases Society of America. Clin Infect Dis 59:e10, 2014. [PubMed: 24973422]

Thomas LD, Schaffner W: Tularemia pneumonia. Infect Dis Clin North Am 24:43, 2010.

Weber IB et al: Clinical recognition and management of tularemia in Missouri: A retrospective records review of 121 cases. Clin Infect Dis 55:1283, 2012.

Chapter 166: Plague and Other *Yersinia* Infections

Michael B. Prentice

PLAGUE

Plague is a systemic zoonosis caused by *Yersinia pestis*. It predominantly affects small rodents in rural areas of Africa, Asia, and the Americas and is usually transmitted to humans by an arthropod vector (the flea). Less often, infection follows contact with animal tissues or respiratory droplets. Plague is an acute febrile illness that is treatable with antimicrobial agents, but mortality rates among untreated patients are high. Ancient DNA studies have confirmed that both the fourteenth-century Black Death and the sixth-century Plague of Justinian in Europe were due to *Y. pestis* infection. Patients can present with the bubonic, septicemic, or pneumonic form of the disease. Although there is concern about epidemic spread of plague by the respiratory route, this is not the most common route of plague transmission, and established infection-control measures for respiratory plague exist. However, the fatalities associated with plague and the capacity for infection via the respiratory tract mean that *Y. pestis* fits the profile of a potential agent of bioterrorism **(Chap. S2)**. Consequently, measures have been taken to restrict access to the organism, including legislation affecting diagnostic and research procedures in some countries (e.g., the United States).

ETIOLOGY

The genus *Yersinia* comprises gram-negative bacteria of the family Enterobacteriaceae (gamma proteobacteria). Overwhelming taxonomic evidence showing *Y. pestis* strains as a clonal group within *Yersinia pseudotuberculosis* suggests recent evolution from the latter organism—an enteric pathogen of mammals that is spread by the fecal–oral route and thus has a phenotype distinctly different from that of *Y. pestis*. When grown in vivo or at 37°C, *Y. pestis* forms an amorphous capsule made from a plasmid-specified fimbrial protein, Caf or fraction 1 (F1) antigen, which is an immunodiagnostic marker of infection.

EPIDEMIOLOGY

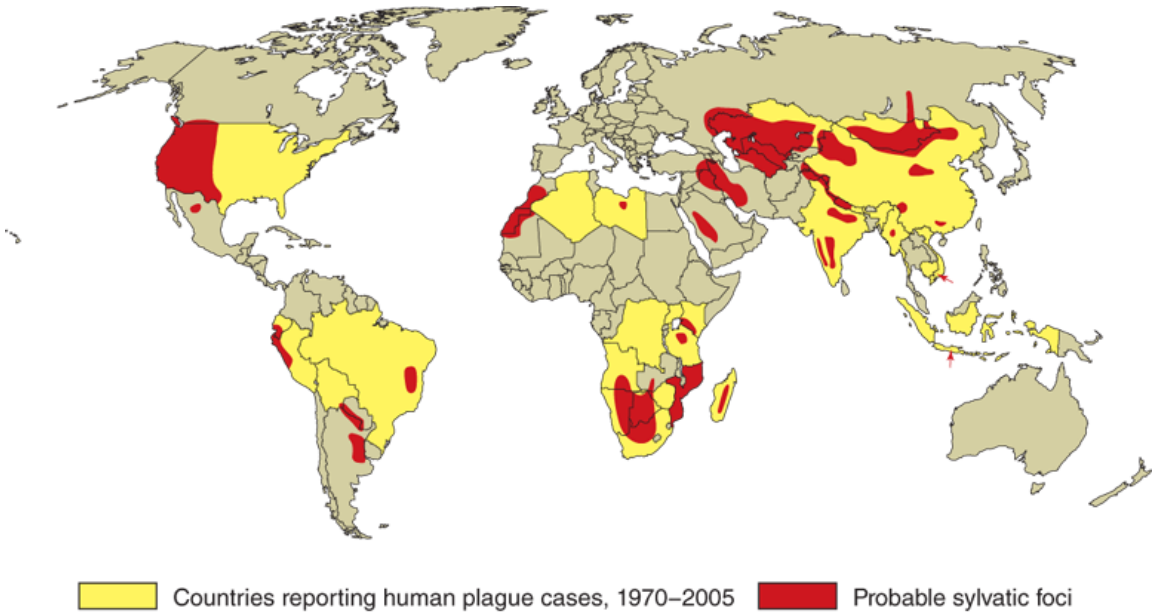
Human plague generally follows an outbreak in a host rodent population (epizootic). Mass deaths among the rodent primary hosts lead to a search by fleas for new hosts, with consequent incidental infection of other mammals. The precipitating cause for an epizootic may ultimately be related to climate or other environmental factors. The reservoir for *Y. pestis* causing enzootic plague in natural endemic foci between epizootics (i.e., when the organism may be difficult to detect in rodents or fleas) is a topic of ongoing research and may not be the same in all regions. The enzootic/epizootic pattern may be the result of complex dynamic interactions of host rodents that have different plague susceptibilities with different flea vectors; alternatively, an environmental reservoir may be important.

GLOBAL FEATURES

In general, the enzootic areas for plague are lightly populated regions of Africa, Asia, and the Americas **(Fig. 166-1)**. Between January 2010 and December 2015, 3248 cases of plague with a global case–fatality rate of 18% were recorded by the World Health Organization (WHO); these figures were based on cases notified under the International Health Regulations and on data from national surveillance programs and publications. More than 96% of these cases were in Africa. The majority of cases were reported from the island of Madagascar, which additionally in 2017 experienced an outbreak of more than 2300 cases (76% pneumonic), with a case–fatality rate of 8.6%. A decline in reports from the Democratic Republic of the Congo (DRC) may reflect ongoing conflict in that country affecting surveillance rather than a true decrease. In the past decade, outbreaks of pneumonic plague have been recorded in the DRC, Uganda, Algeria, Madagascar, China, and Peru.

FIGURE 166-1

Approximate global distribution of *Yersinia pestis*. *(Compiled from WHO, CDC, and country sources. Reprinted with permission from DT Dennis, GL Campbell: Plague and other Yersinia infections, in Harrison's Principles of Internal Medicine, 17th ed, AS Fauci et al [eds]. New York, McGraw-Hill, Chap. 152, 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Plague was introduced into North America via the port of San Francisco in 1900 as part of the Third Pandemic, which spread around the world from Hong Kong. The disease is presently enzootic on the western side of the continent from southwestern Canada to Mexico. Most human cases in the United States occur in two regions: "Four Corners" (the junction point of New Mexico, Arizona, Colorado, and Utah), especially northern New Mexico, northern Arizona, and southern Colorado; and further west in California, southern Oregon, and western Nevada *(http://www.cdc.gov/plague/maps/index.html)*. From 1990 to 2015, 185 cases of plague were reported in the United States, a mean of seven cases per year. Most cases occurred from May to October—the time of year when people are outdoors and rodents and their fleas are most plentiful. Infection is most often acquired by fleabite in peridomestic environments. Infection can also be transmitted during the handling of living or dead small mammals (e.g., rabbits, hares, and prairie dogs) or wild carnivores (e.g., wildcats, coyotes, or mountain lions). Dogs and cats may bring plague-infected fleas into the home, and infected cats or dogs may transmit plague directly to humans by the respiratory route. In 2014, an outbreak of nonfatal pneumonic plague in Colorado affected four people exposed to an infected dog, with possible interhuman transmission in one case. The most recent case of person-to-person transmission in the United States before this occurred in the Los Angeles pneumonic plague outbreak of 1924.

Plague most often develops in areas with poor sanitary conditions and infestations of rats—in particular, the widely distributed roof rat *Rattus rattus* and the brown rat *Rattus norvegicus* (which serves as a laboratory model of plague). Rat control in warehouses and shipping facilities has been recognized as important in preventing the spread of plague since the early twentieth century and features in the current WHO International Health Regulations. Urban rodents acquire infection from wild rodents, and the proximity of the former to humans increases the risk of transmission. The oriental rat flea *Xenopsylla cheopis* is the most efficient vector for transmission of plague among rats and onward to humans in Asia, Africa, and South America.

Worldwide, bubonic plague is the predominant form reported (80–95% of suspected cases), with mortality rates of 10–20%. The mortality rate is higher (22%) in the small proportion of patients (10–20%) with primary septicemic plague (i.e., systemic *Y. pestis* sepsis with no bubo; see "Clinical Manifestations," below) and is highest with primary pulmonary plague. The latter is generally the least common of the main plague presentations, but, as in the 2017 Madagascar outbreak, it is occasionally predominant. Mortality rates of 50% or more for primary pulmonary plague are reported with delayed antimicrobial treatment in small case series from the older literature. Rare outbreaks of pharyngeal plague following consumption of raw or undercooked camel or goat meat have been reported.

A total of 744 (82%) of the 913 plague cases with clinically documented features (out of 1006 cases reported in total) in the United States from 1900 to 2012 were bubonic disease, 87 (10%) were septicemic disease, and 74 (8%) were pneumonic disease; 6 cases (1%) were pharyngeal. Sixteen percent of cases were fatal in the postantibiotic era from 1942 onward compared with 66% in the period 1900–1941.

PATHOGENESIS

As mentioned earlier, genetic evidence suggests that *Y. pestis* is a clone derived from the enteric pathogen *Y. pseudotuberculosis* in the recent evolutionary past (9000–20,000 years ago). The change from infection by the fecal–oral route to a two-stage life cycle, with alternate parasitization of arthropod and mammalian hosts, followed the acquisition of two plasmids—pFra and pPst—and the inactivation of remarkably few *Y.*

pseudotuberculosis genes in conjunction with preexisting properties of the *Y. pseudotuberculosis* ancestor, including the presence of a third plasmid, pYV, and the capacity to cause septicemia. In the arthropod-parasitizing portion of its life cycle, *Y. pestis* multiplies and forms biofilm-embedded aggregates in the flea midgut after ingestion of a blood meal containing bacteria. In some fleas, biofilm-embedded bacteria eventually fill the proventriculus (a valve connecting the esophagus to the midgut) and block normal blood feeding. Both "blocked" fleas and those containing masses of biofilm-embedded *Y. pestis* without complete blockage inoculate *Y. pestis* into each bite site. The ability of *Y. pestis* to colonize and multiply in the flea requires phospholipase D encoded by the *ymt* gene on the pFra plasmid, and biofilm synthesis requires the chromosomal *hms* locus shared with *Y. pseudotuberculosis.* However, three *Y. pseudotuberculosis* genes inhibiting biofilm formation or promoting its degradation are inactivated in *Y. pestis*, together with urease, which causes acute flea gastrointestinal toxicity. Blockage takes days or weeks to come about after initial infection of the flea and is followed by the flea's death. Many flea vectors (including *X. cheopis*) are also able to transmit plague in an early-phase unblocked state for up to a week after feeding, but 10 fleas in this state are required to infect a mammalian host (mass transmission).

Y. pestis disseminates from the site of inoculation in the mammalian host in a process initially dependent on plasminogen activator Pla, which is encoded by the small pPst plasmid. This surface protease activates mammalian plasminogen, degrades complement, and adheres to the extracellular matrix component laminin. Pla is essential for the high-level virulence of *Y. pestis* in mice by subcutaneous or intradermal injection (laboratory proxies for fleabites) and for the development of primary pneumonic plague. When actual fleabite inoculation is used in mouse models, the fimbrial capsule-forming protein (Ca1 or fraction 1; F1 antigen) encoded on pFra increases the efficiency of transmission, and plasminogen activator is required for the formation of buboes. Macrophages, neutrophils, and dendritic cells are all involved in the innate immune response to flea-transmitted *Y. pestis*. The organism is taken up by macrophages but avoids being killed by autophagy and can also survive and replicate in neutrophils. Rapid transport of the bacteria to regional lymph nodes occurs. *Y. pestis* then undergoes extracellular replication with full expression of its antiphagocytic systems: the type III secretion machines and their effectors encoded by pYV as well as the F1 capsule. These factors prevent neutrophil uptake, and the type III secretion effectors also block extrusion of microbicidal DNA by neutrophils and trigger apoptotic cell death. Overproduction of the type III secretion substrate and translocation protein LcrV exerts an anti-inflammatory effect, reducing host immune responses. Likewise, *Y. pestis* lipopolysaccharide is modified to minimize stimulation of host Toll-like receptor 4, thereby reducing protective host inflammatory responses during peripheral infection and prolonging host survival with high-grade bacteremia—an effect that probably enhances the pathogen's subsequent transmission by fleabite.

Replication of *Y. pestis* in a regional lymph node results in the local swelling of the lymph node and periglandular region known as a *bubo*. On histology, the node is found to be hemorrhagic or necrotic, with thrombosed blood vessels, and the lymphoid cells and normal architecture are replaced by large numbers of bacteria and fibrin. Periglandular tissues are inflamed and also contain large numbers of bacteria in a serosanguineous, gelatinous exudate.

Continued spread through the lymphatic vessels to contiguous lymph nodes produces second-order primary buboes. Infection is initially contained in the infected regional lymph nodes, although transient bacteremia can be detected. As the infection progresses, spread via efferent lymphatics to the thoracic duct produces high-grade bacteremia. Hematogenous spread to the spleen, liver, and secondary buboes follows, with subsequent uncontrolled septicemia, endotoxic shock, and disseminated intravascular coagulation leading to death. In some patients, this septicemic phase occurs without obvious prior bubo development or lung disease (septicemic plague). Hematogenous spread to the lungs results in secondary plague pneumonia, with bacteria initially more prominent in the interstitium than in the air spaces (the reverse being the case in primary plague pneumonia). Hematogenous spread to other organs, including the meninges, can occur.

CLINICAL MANIFESTATIONS

Bubonic Plague

After an incubation period of 2–6 days, the onset of bubonic plague is sudden and is characterized by fever (>38°C), malaise, myalgia, dizziness, and increasing pain due to progressive lymphadenitis in the regional lymph nodes near the fleabite or other inoculation site. Lymphadenitis manifests as a tense, tender swelling (bubo) that, when palpated, has a boggy consistency with an underlying hard core. Generally, there is one painful and erythematous bubo with surrounding periganglionic edema. The bubo is most commonly inguinal but can also be crural, axillary **(Fig. 166-2)**, cervical, or submaxillary, depending on the site of the bite. Abdominal pain from intraabdominal node involvement can occur without other visible signs. Children are most likely to present with cervical or axillary buboes.

FIGURE 166-2

Plague patient in the southwestern United States with a left axillary bubo and an unusual plague ulcer and eschar at the site of the infective flea bite. *(Reprinted with permission from DT Dennis, GL Campbell: Plague and other Yersinia infections, in Harrison's Principles of Internal Medicine, 17th ed, AS Fauci et al [eds]. New York, McGraw-Hill, Chap. 152, 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The differential diagnosis includes acute focal lymphadenopathy of other etiologies, such as streptococcal or staphylococcal infection, tularemia, cat-scratch disease, tick typhus, infectious mononucleosis, or lymphatic filariasis. These infections do not progress as rapidly, are not as painful, and are associated with visible cellulitis or ascending lymphangitis—both of which are absent in plague.

Without treatment, *Y. pestis* dissemination occurs and causes serious illness, including pneumonia (secondary pneumonic plague) and meningitis. Secondary pneumonic plague can be the source of person-to-person transmission of respiratory infection by productive cough (droplet infection), with the consequent development of primary plague pneumonia. Appropriate treatment of bubonic plague results in fever resolution within 2–5 days, but buboes may remain enlarged for >1 week after initial treatment and can become fluctuant.

Primary Septicemic Plague

A minority (10–25%) of infections with *Y. pestis* present as gram-negative septicemia (hypotension, shock) without preceding lymphadenopathy. Septicemic plague occurs in all age groups, but persons >40 years of age are at elevated risk. Some chronic conditions may predispose to septicemic plague: in 2009 in the United States, a fatal laboratory-acquired infection with an attenuated *Y. pestis* strain manifested as septicemic plague in a 60-year-old researcher with diabetes mellitus and undiagnosed hemochromatosis. These conditions also carry an increased risk of septicemia with other pathogenic *Yersinia* species. The term *septicemic plague* can be confusing since most patients with buboes have detectable bacteremia at some stage, with or without systemic signs of sepsis. In laboratory experiments, however, septicemic disease without histologic changes in lymph nodes is seen in a minority of mice infected via fleabites.

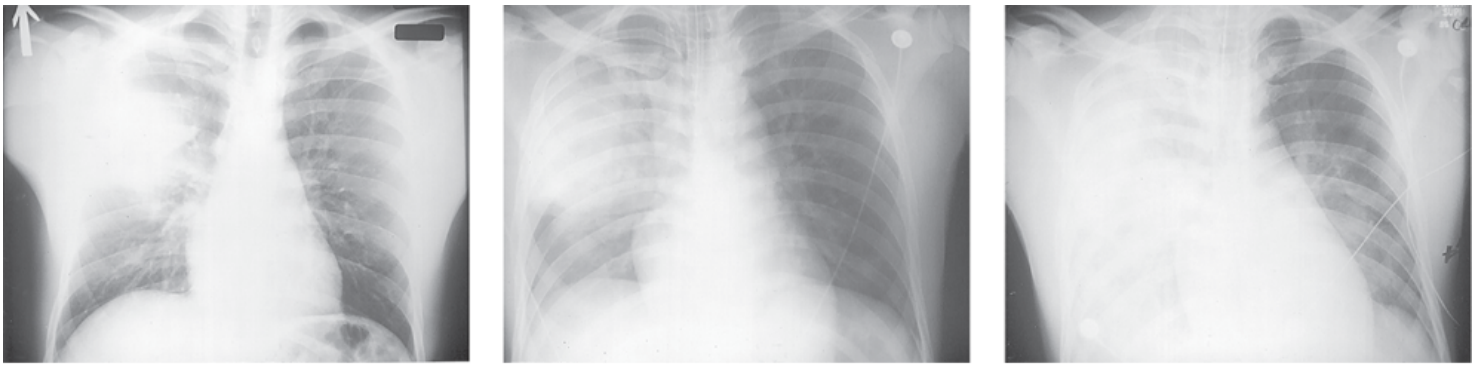
Pneumonic Plague

Primary pneumonic plague results from inhalation of infectious bacteria in droplets expelled from another person or an animal with primary or secondary plague pneumonia. This syndrome has a short incubation period, averaging from a few hours to 2–3 days (range, 1–7 days), and is characterized by a sudden onset of fever, headache, myalgia, weakness, nausea, vomiting, and dizziness. Respiratory signs—cough, dyspnea, chest pain, and sputum production with hemoptysis—typically arise after 24 h. Progression of initial segmental pneumonitis to lobar pneumonia and then to bilateral lung involvement may occur (Fig. 166-3). The possible release of aerosolized *Y. pestis* bacteria in a bioterrorist attack, manifesting as an outbreak of primary pneumonic plague in nonendemic regions or in an urban setting where plague is rarely seen, has been a source of public health concern. Secondary pneumonic plague is a consequence of bacteremia occurring in ~10–15% of patients with bubonic plague. Bilateral alveolar infiltrates are seen on chest x-ray, and diffuse interstitial pneumonitis with scanty sputum production is typical.

FIGURE 166-3

Sequential chest radiographs of a patient with fatal primary plague pneumonia. ***Left:*** Upright posteroanterior film taken at admission to hospital emergency department on third day of illness, showing segmental consolidation of right upper lobe. ***Center:*** Portable anteroposterior film taken 8 h after admission, showing extension of pneumonia to right middle and right lower lobes. ***Right:*** Portable anteroposterior film taken 13 h after

admission (when patient had clinical acute respiratory distress syndrome), showing diffuse infiltration throughout right lung and patchy infiltration of left lower lung. A cavity later developed at the site of initial right-upper-lobe consolidation. *(Reprinted with permission from DT Dennis, GL Campbell: Plague and other Yersinia infections, in Harrison's Principles of Internal Medicine, 17th ed. AS Fauci et al [eds]. New York, McGraw-Hill, Chap. 152, 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Meningitis

Meningeal plague is uncommon, occurring in ≤6% of plague cases reported in the United States. Presentation with headache and fever typically occurs >1 week after the onset of bubonic or septicemic plague and may be associated with suboptimal antimicrobial therapy (delayed therapy, penicillin administration, or low-dose tetracycline treatment) and cervical or axillary buboes.

Pharyngitis

Symptomatic plague pharyngitis can follow the consumption of contaminated meat from an animal dying of plague or contact with persons or animals with pneumonic plague. This condition can resemble tonsillitis, with peritonsillar abscess and cervical lymphadenopathy. Asymptomatic pharyngeal carriage of *Y. pestis* can also occur in close contacts of patients with pneumonic plague.

LABORATORY DIAGNOSIS

Because of the scarcity of laboratory facilities in regions where human *Y. pestis* infection is most common, and because of the potential significance of *Y. pestis* isolation in a nonendemic area or an area from which human plague has been absent for many years, the WHO recommends an initial presumptive diagnosis followed by reference laboratory confirmation **(Table 166-1)**. In the United States, comprehensive national diagnostic facilities for plague have been in place since a federal Laboratory Response Network (LRN; *https://emergency.cdc.gov/lrn/index.asp*) was set up in 1999 to detect possible use of biological terrorism agents, including *Y. pestis*. Routine diagnostic clinical microbiology laboratories that are included in this network as sentinel-level laboratories use joint protocols from the Centers for Disease Control and Prevention (CDC) and the American Society for Microbiology to identify suspected *Y. pestis* isolates and to refer these specimens to LRN reference laboratories for confirmatory tests (*https://www-asm-org.ezproxy1.library.usyd.edu.au/index.php/guidelines/sentinel-guidelines*). *Y. pestis* is designated a "Tier 1 select agent" under the Public Health Security and Bioterrorism Preparedness and Response Act of 2002 and subsequent executive orders; the provisions of this act, the Patriot Act of 2001, and related executive orders apply to all U.S. laboratories and individuals working with *Y. pestis*. Details of the applicable regulations are available from the CDC.

TABLE 166-1

World Health Organization Case Definitions of Plague

Suspected Case
Compatible clinical presentation ***and*** Consistent epidemiologic features, such as exposure to infected animals or humans and/or evidence of fleabites and/or residence in or travel to a known endemic focus within the previous 10 days
Presumptive Case
Meeting the definition of a suspected case ***plus*** **Putative new or reemerging focus:** ≥2 of the following tests positive • Microscopy: gram-negative coccobacilli in material from bubo, blood, or sputum; bipolar appearance of Wayson or Wright-Giemsa staining • F1 antigen detected in bubo aspirate, blood, or sputum • A single anti-F1 serology without evidence of previous *Yersinia pestis* infection or immunization • PCR detection of *Y. pestis* in bubo aspirate, blood, or sputum **Known endemic focus:** ≥1 of the following tests positive • Microscopic evidence of gram-negative or bipolar (Wayson, Wright-Giemsa) coccobacilli from bubo, blood, or sputum • A single anti-F1 serology without evidence of previous plague infection or immunization • F1 antigen detected in bubo aspirate, blood, or sputum
Confirmed Case
Meeting the definition of a suspected case ***plus*** • Identification of an isolate from a clinical sample as *Y. pestis* (colonial morphology and 2 of the following 4 tests positive: phage lysis of cultures at 20–25°C and 37°C; F1 antigen detection; PCR; *Y. pestis* biochemical profile) ***or*** • A fourfold rise in anti-F1 titer in paired serum samples ***or*** • In endemic areas when no other confirmatory test can be performed, a positive rapid diagnostic test with immunochromatography to detect F1 antigen

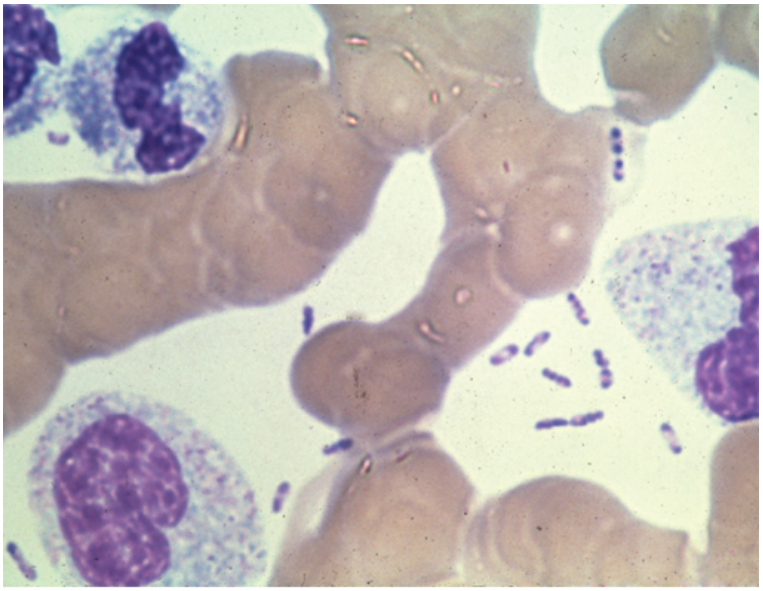
Abbreviation: PCR, polymerase chain reaction.

Source: Interregional Meeting on Prevention and Control of Plague, Antananarivo, Madagascar, 7–11 April 2006 *(www.who.int/entity/csr/resources/publications/WHO_HSE_EPR_2008_3w.pdf).*

Yersinia species are gram-negative coccobacilli (short rods with rounded ends) 1–3 µm in length and 0.5–0.8 µm in diameter. *Y. pestis* in particular appears bipolar (with a “closed safety pin” appearance) and pleomorphic when stained with a polychromatic stain (Wayson or Wright-Giemsa; **Fig. 166-4**). Its lack of motility distinguishes *Y. pestis* from other *Yersinia* species, which are motile at 25°C and nonmotile at 37°C. Transport medium (e.g., Cary-Blair medium) preserves the viability of *Y. pestis* if transport is delayed.

FIGURE 166-4

Peripheral-blood smear from a patient with fatal plague septicemia and shock, showing characteristic bipolar-staining *Yersinia pestis* bacilli (Wright’s stain, oil immersion). *(Reprinted with permission from DT Dennis, GL Campbell: Plague and other Yersinia infections, in Harrison’s Principles of Internal Medicine, 17th ed, AS Fauci et al [eds]. New York, McGraw-Hill, Chap. 152, 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The appropriate specimens for diagnosis of bubonic, pneumonic, and septicemic plague are bubo aspirate, bronchoalveolar lavage fluid or sputum, and blood, respectively. Culture of postmortem organ biopsy samples can also be diagnostic. A bubo aspirate is obtained by injection of 1 mL of sterile normal saline into a bubo under local anesthetic and aspiration of a small amount of (usually blood-stained) fluid. The WHO has provided interim guidance on how to aspirate buboes and collect sputum from patients with suspected pneumonic plague (*http://www.who.int/csr/disease/plague/collecting-pus-samples.PDF?ua=1*; *http://www.who.int/csr/disease/plague/collecting-sputum-samples.PDF?ua=1*). Gram's staining of these specimens may reveal gram-negative rods, which are shown by Wayson or Wright-Giemsa staining to be bipolar. These bacteria may even be visible in direct blood smears in septicemic plague (Fig. 166-4); this finding indicates very high numbers of circulating bacteria and a poor prognosis.

Y. pestis grows on nutrient agar and other standard laboratory media but forms smaller colonies than do other Enterobacteriaceae. Specimens should be inoculated onto nutrient-rich media such as sheep blood agar (SBA), into nutrient-rich broth such as brain-heart infusion broth, and onto selective agar such as MacConkey or eosin methylene blue (EMB) agar. *Yersinia*-specific CIN (cefsulodin, triclosan [Irgasan], novobiocin) agar can be useful for culture of contaminated specimens, such as sputum. Blood should be cultured in a standard blood culture system. The optimal growth temperature is <37°C (25–29°C), with pinpoint colonies only on SBA at 24 h. Slower growth occurs at 37°C. *Y. pestis* is oxidase-negative, catalase-positive, urea-negative, indole-negative, and lactose-negative. Automated biochemical identification systems can misidentify *Y. pestis* as *Y. pseudotuberculosis* or other bacterial species.

Reference laboratory tests for definitive identification of isolates include direct immunofluorescence for F1 antigen; specific polymerase chain reaction (PCR) for targets such as F1 antigen, the pesticin gene, and the plasminogen activator gene; and specific bacteriophage lysis. PCR can also be applied to diagnostic specimens, as can direct immunofluorescence for F1 antigen (produced in large amounts by *Y. pestis*) by slide microscopy. An immunochromatographic test strip for F1 antigen detection by monoclonal antibodies in clinical specimens has been devised in Madagascar. This method is effective for both laboratory and near-patient use and is now widely used in endemic countries. A similar test strip for Pla antigen has been developed and could be used to detect wild-type or engineered F1-negative virulent strains. Many other rapid diagnostic kits for possible bioterrorism pathogens, including *Y. pestis*, have been described in recent years, but none is widely used for primary or reference laboratory identification, and only one (a field real-time PCR for a range of potential bioterrorism agents) is approved by the U.S. Food and Drug Administration (FDA). Detailed phylogeographic DNA sequence data based on culture collections have been accumulated to trace plague evolution, and this approach could be adapted in the future to real-time clinical plague epidemiology.

In the absence of other positive laboratory diagnostic tests, a retrospective serologic diagnosis may be made on the basis of rising titers of hemagglutinating antibody to F1 antigen. Enzyme-linked immunosorbent assays (ELISAs) for IgG and IgM antibodies to F1 antigen are also available.

The white blood cell (WBC) count is generally raised (to 10,000–20,000/μL) in plague, with neutrophilic leukocytosis and a left shift (numerous immature neutrophils); in some cases, however, the WBC count is normal or leukopenia develops. WBC counts are occasionally very high, especially in children (>100,000/μL). Levels of fibrinogen degradation products are elevated in a majority of patients, but platelet counts are usually normal or low-normal. However, disseminated intravascular coagulation, with low platelet counts, prolonged prothrombin times, reduced fibrinogen, and elevated fibrinogen degradation product levels, occurs in a significant minority of patients.

TREATMENT

TREATMENT

Plague

Guidelines for the treatment of plague are given in **Table 166-2**. A 10- to 14-day course of antimicrobial therapy (or a course continued until 2 days after fever subsides) is recommended. Streptomycin has historically been the parenteral treatment of choice for plague and is approved for this indication by the FDA. Although not yet approved by the FDA for plague, gentamicin has proven safe and effective in clinical trials in Tanzania and Madagascar and in retrospective reviewed cases in the United States. In view of streptomycin's adverse-reaction profile and limited availability, some experts now recommend gentamicin over streptomycin. The FDA has approved levofloxacin, moxifloxacin, and ciprofloxacin for prophylaxis and treatment of plague (including septicemic and pneumonic plague) under a regulatory approach based on animal studies alone, known as the Animal Rule. Levofloxacin has more efficacy than ciprofloxacin in postexposure prophylaxis of inhalational anthrax in animal models and has also received FDA approval for this indication **(Chap. S2)**; thus it is a suitable agent for prophylaxis against two diseases in possible bioterrorism exposures.

While systemic chloramphenicol therapy is available in the resource-poor countries primarily affected by plague, it is less likely to be available or used in high-income countries because of its adverse-effect profile. Tetracyclines are also effective and can be given by mouth but are not generally recommended for children age <7 years because of tooth discoloration. Doxycycline is the tetracycline of choice; at an oral dosage of 100 mg twice daily, this drug was as effective as intramuscular gentamicin (2.5 mg/kg twice daily) in a trial in Tanzania. There is recent evidence that doxycycline does not cause dental staining in children because it binds calcium less readily than other tetracyclines.

Although *Y. pestis* is sensitive to β-lactam drugs in vitro and these drugs have been efficacious against plague in some animal models, the response to penicillins has been poor in some clinical cases; thus β-lactams and macrolides are not generally recommended as first-line therapy. Chloramphenicol, alone or in combination, is recommended for some focal complications of plague (e.g., meningitis, endophthalmitis, myocarditis) because of its tissue penetration properties. Fluoroquinolones, effective in vitro and in animal models, are recommended in guidelines for possible bioterrorism-associated pneumonic plague and are increasingly used in plague therapy.

TABLE 166-2

Guidelines for the Treatment of Plague

Drug	Daily Dose	Dosing Interval, h	Route
Gentamicin			
Adult	5 mg/kg[a]	24	IM/IV
	5 mg/kg	8[b]	IM/IV
Child	5 mg/kg[a]	24	IM/IV
	7.5 mg/kg	8[c]	IM/IV
Streptomycin			
Adult	2 g	12	IM
Child	30 mg/kg	12	IM
Levofloxacin			
Adult and child >50 kg	500 mg	24	PO/IV
Child <50 kg and ≥6 months of age	16 mg/kg (maximum, 250 mg/dose)	12	PO/IV
Ciprofloxacin			
Adult	1000–1500 mg	12	PO
	800–1200 mg	8–12	IV
Child	40 mg/kg (maximum, 500 mg/dose)	12	PO
	30 mg/kg (maximum, 400 mg/dose)	12	IV
Moxifloxacin			
Adult	400 mg	24	PO/IV
Doxycycline			
Adult and child ≥45 kg	200 mg	12 or 24	PO/IV
Child <45 kg	4.4 mg/kg (maximum, 100 mg/dose)	12	PO/IV
Tetracycline			
Adult	2 g	6	PO/IV
Child >8 yr	25–50 mg/kg	6	PO/IV
Chloramphenicol			
Adult	100 mg/kg	6	PO/IV

Drug	Daily Dose	Dosing Interval, h	Route
Child >2 yr	100 mg/kg (maximum, 4 g)	6	PO/IV

[a]Aminoglycoside dose is adjusted with impaired renal function. No trial data have been published for once-daily gentamicin therapy for plague in adults or children, but this regimen is efficacious in gram-negative sepsis of other etiologies and has been successful in a recent outbreak of pneumonic plague in the Democratic Republic of the Congo. Neonates (up to 1 week of age) and premature infants should receive gentamicin at 2.5 mg/kg IV twice daily. [b]A 2-mg/kg loading dose followed by 1.7 mg/kg 3 times daily, reduced. [c]2.5 mg/kg 3 times daily.

Source: TV Inglesby et al: Plague as a biological weapon: Medical and public health management. Working Group on Civilian Biodefense. JAMA 283:2281, 2000; and *https://www.cdc.gov/plague/healthcare/clinicians.html*.

PREVENTION

In endemic areas, the control of plague in humans is based on reduction of the likelihood of being bitten by infected fleas or exposed to infected droplets from either humans or animals with plague pneumonia. In the United States, residence and outdoor activity in rural areas of western states where epizootics occur are the main risk factors for infection. To assess potential risks to humans in specific areas, surveillance for *Y. pestis* infection among animal plague hosts and vectors is carried out regularly as well as in response to observed animal die-offs. Personal protective measures include avoidance of areas where a plague epizootic has been identified and publicized (e.g., by warning signs or closure of campsites). Sick or dead animals should not be handled by the general public. Hunters and zoologists should wear gloves when handling wild-animal carcasses in endemic areas. General measures to avoid rodent fleabite during outdoor activity are appropriate and include the use of insect repellant, insecticide, and protective clothing. General measures to reduce peridomestic and occupational human contact with rodents are advised and include rodent-proofing of buildings and food-waste stores and removal of potential rodent habitats (e.g., woodpiles and junk heaps). Flea control by insecticide treatment of wild rodents is an effective means of minimizing human contact with plague if an epizootic is identified in an area close to human habitation. Any attempt to reduce rodent numbers must be preceded by flea suppression to reduce the migration of infected fleas to human hosts. An oral F1-V subunit vaccine using raccoon poxvirus (RCN) as a vector (sylvatic plague vaccine) is partially protective against plague when administered to wild prairie dogs in field trials and may in the future provide a means of reducing the risk of human exposure to *Y. pestis*.

Patients in whom pneumonic plague is suspected should be managed in isolation (with negative pressure, if available), with droplet precautions observed until pneumonia is excluded or effective antimicrobial therapy has been given for 48 h. Review of the literature published before the advent of antimicrobial agents suggests that the main infective risk is posed by patients in the final stages of disease who are coughing up sputum with plentiful visible blood and/or pus. Cotton and gauze masks were protective in these circumstances. Current surgical masks capable of barrier protection against droplets, including large respiratory particles, are probably protective, but the differential diagnosis of fever and hemoptysis in plague-endemic areas includes aerosol-transmitted infections such as tuberculosis. In addition, WHO guidance recommends that personal protective equipment for potential aerosol-generating procedures (e.g., collection of respiratory samples from patients with suspected or confirmed plague) should include a fit-tested N95 face mask, a gown, gloves, and a face shield or goggles.

Antimicrobial Prophylaxis

Postexposure antimicrobial prophylaxis lasting 7 days is recommended following household, hospital, or other close contact with persons with untreated pneumonic plague. (*Close contact* is defined as contact with a patient at <2 m.) In animal aerosol-infection studies, levofloxacin and ciprofloxacin are associated with higher survival rates than doxycycline **(Table 166-3)**.

TABLE 166-3

Guidelines for Plague Prophylaxis

Drug	Daily Dose	Dosing Interval, h	Route
Doxycycline			
Adult	200 mg	12 or 24	PO
Child ≥8 y	≥45 kg: adult dose	12	PO
	≤45 kg: 2.2 mg/kg bid (maximum, 200 mg)	12	PO
Tetracycline			
Adult	1–2 g	6 or 12	PO
Child ≥8 y	25–50 mg/kg	6 or 12	PO
Levofloxacin			
Adult and child >50 kg)	500 mg	24	PO
Child <50 kg and ≥6 months of age	16 mg/kg (maximum, 250 mg/dose)	12	PO
Ciprofloxacin			
Adult	1 g	12	PO
Child	40 mg/kg	12	PO

Source: TV Inglesby et al: Plague as a biological weapon: Medical and public health management. Working Group on Civilian Biodefense. JAMA 283:2281, 2000; FDA Drug Product Label Reference ID:3382318; *https://www.cdc.gov/plague/healthcare/clinicians.html*; *www.accessdata.fda.gov/drugsatfda_docs/label/2013/020634s065,020635s071,021721s032lbl.pdf*.

Immunization

Studies with candidate plague vaccines in animal models show that neutralizing antibody provides protection against exposure but that cell-mediated immunity is critical for protection and clearance of *Y. pestis* from the host. A killed whole-cell vaccine used in humans required multiple doses, caused significant local and systemic reactions, and was not protective against pneumonic plague; this vaccine is not currently available in the United States. A live attenuated vaccine based on strain EV76 is still used in countries of the former Soviet Union and China but has significant side effects. The vaccines closest to licensing are subunit vaccines comprising recombinant F1 (rF1) and various recombinant V (rV) proteins produced in *Escherichia coli*, combined either as a fusion protein or as a mixture, purified, and adsorbed to aluminum hydroxide for injection. This combination protects mice and various nonhuman primates in laboratory models of bubonic and pneumonic plague and has been evaluated in phase 2 clinical trials. Special ethical considerations with controlled clinical studies involving plague in humans make prelicensing field-efficacy studies unlikely. In the United States, the FDA is therefore prepared to assess plague vaccines for human use under the Animal Rule, using efficacy data and other results from animal studies and antibodies and other correlates of immunity from human vaccinees *(www.fda.gov/emergencypreparedness/counterterrorism/medicalcountermeasures/mcmregulatoryscience/ucm391604.htm)*. Live attenuated *Y. pseudotuberculosis* and *Salmonella* strains expressing *Y. pestis*–specific antigens have been shown to be protective in laboratory animal models of bubonic and pneumonic plague and could be delivered by the oral route. A wide variety of other delivery mechanisms for *Y. pestis* antigens are being explored. Antigens other than F1 and V that could be added to subunit vaccines are being investigated. Providing impetus for exploration of these antigens are the recovery of F1-negative *Y. pestis* strains from natural sources and the observation that F1 antigen is not required for virulence in primate models of pneumonic plague.

YERSINIOSIS

Yersiniosis is a zoonotic infection with an enteropathogenic *Yersinia* species, usually *Y. enterocolitica* or *Y. pseudotuberculosis*. The usual hosts for these organisms are pigs and other wild and domestic animals; humans are usually infected by the oral route, and outbreaks from contaminated food occur. Yersiniosis is most common in childhood and in colder climates. Patients present with abdominal pain and sometimes with diarrhea (which may not occur in up to 50% of cases). *Y. enterocolitica* is more closely associated with terminal ileitis and *Y. pseudotuberculosis* with mesenteric adenitis, but both organisms may cause mesenteric adenitis and symptoms of abdominal pain and tenderness that result in pseudoappendicitis, with the surgical removal of a normal appendix. Diagnosis is based on culture of the organism or convalescent serology. *Y. pseudotuberculosis* and some rarer strains of *Y. enterocolitica* are especially likely to cause systemic infection, which is also more likely in patients with diabetes or iron overload. Systemic sepsis is treatable with antimicrobial agents, but postinfective arthropathy responds poorly to such therapy. Fourteen other *Yersinia* species lacking the virulence plasmid pYV common to *Y. pestis*, *Y. pseudotuberculosis*, and *Y. enterocolitica* are now recognized. These are, at most, opportunistic pathogens of humans (*Y. aldovae*, *Y. aleksiciae*, *Y. bercovieri, Y. entomophaga*, *Y. frederiksenii*, *Y. intermedia*, *Y. kristensenii*, *Y. massiliensis*, *Y. mollaretii*, *Y. nurmii*, *Y. pekkanenii*, *Y. rohdei*, *Y. similis,* and *Y. ruckeri*). Molecular phylogeny shows that *Y. enterocolitica* is more distantly related to *Y. pseudotuberculosis* than these other *Yersinia* species, and the similar virulence plasmid they share has probably been acquired independently by at least one of the two since the species diverged.

EPIDEMIOLOGY

Y. enterocolitica

Y. enterocolitica is found worldwide and has been isolated from a wide variety of wild and domestic animals and environmental samples, including samples of food and water. In vitro, *Y. enterocolitica* is resistant to predation by the protozoon *Acanthamoeba castellani* and can survive inside it, suggesting a possible mode of environmental persistence. Strains are differentiated by combined biochemical reactions (biovar) and serogroup. Most clinical infections are associated with serogroups O:3, O:9, and O:5,27, with a declining number of O:8 infections in North America. Some O:8 infections, previously confined to North America, have been reported from Europe and Japan in recent years, and O:8 infections caused a high percentage of yersiniosis cases in Poland in 2008–2011, with a subsequent decline. Yersiniosis, mostly due to *Y. enterocolitica*, is the third commonest zoonosis reported in Europe; most reports come from northern Europe, especially Germany and Scandinavia. The incidence is highest among children; children <4 years of age are more likely to present with diarrhea than are older children. Abdominal pain with mesenteric adenitis and terminal ileitis is more prominent among older children and adults. Septicemia is more likely in patients with preexisting conditions such as diabetes mellitus, liver disease, any condition involving iron overload (including thalassemia and hemochromatosis), advanced age, malignancy, or HIV/AIDS. As in enteritis of other bacterial etiologies, postinfective complications such as reactive arthritis occur mainly in individuals who are HLA-B27 positive. Erythema nodosum **(Fig. A1-39)** following *Yersinia* infection is not associated with HLA-B27 and is more common among women than among men.

Consumption or preparation of raw pork products (such as chitterlings) and some processed pork products is strongly linked with infection because a high percentage of pigs carry pathogenic *Y. enterocolitica* strains. Outbreaks of *Y. enterocolitica* infection have been associated with consumption of milk (pasteurized, unpasteurized, and chocolate-flavored) and various foods contaminated with springwater. Person-to-person transmission is suspected in a few cases (e.g., in nosocomial and familial outbreaks) but is much less likely with *Y. enterocolitica* than with other causes of gastrointestinal infection, such as *Salmonella*. A multivariate analysis indicates that contact with companion animals is a risk factor for *Y. enterocolitica* infection among children in Sweden, and low-level colonization of dogs and cats with *Y. enterocolitica* has been reported. Transfusion-associated septicemia due to *Y. enterocolitica*, while recognized as a very rare but frequently fatal event for >30 years, has been difficult to eradicate.

Y. pseudotuberculosis

Y. pseudotuberculosis is less frequently reported as a cause of human disease than *Y. enterocolitica*, and infection with *Y. pseudotuberculosis* is more likely to present as fever and abdominal pain due to mesenteric lymphadenitis. This organism is associated with wild mammals (rodents, rabbits, and deer), birds, and domestic pigs. Although outbreaks are generally rare, several have recently occurred in Finland in association with consumption of lettuce, raw carrots, or unpasteurized milk. Strains have historically been differentiated by combined biochemical reactions (biovar) and serogroup. Multilocus sequence typing and other phenotypic typing have revealed that some strains previously assigned to *Y. pseudotuberculosis* belong to the closely related but distinct species now called *Yersinia wautersii* (pathogenic) and *Yersinia similis* (nonpathogenic).

PATHOGENESIS

The usual route of infection is oral. Studies with both *Y. enterocolitica* and *Y. pseudotuberculosis* in animal models suggest that initial replication in the small intestine is followed by invasion of Peyer's patches of the distal ileum via M cells, with onward spread to mesenteric lymph nodes. The liver and spleen can also be involved after oral infection. The characteristic histologic appearance of enteropathogenic *Yersinia* after invasion of host tissues is as extracellular microabscesses surrounded by an epithelioid granulomatous lesion.

Experiments involving oral infection of mice with tagged *Y. enterocolitica* show that only a very small proportion of bacteria in the gut invade tissues. Individual bacterial clones from an orally inoculated pool give rise to each microabscess in a Peyer's patch, and the host restricts the invasion of previously infected Peyer's patches. A prior model positing progressive bacterial spread from Peyer's patches and mesenteric lymph nodes to the liver and spleen appears to be inaccurate: spread of *Y. pseudotuberculosis* and *Y. enterocolitica* to the liver and spleen of mice occurs independently of regional lymph node colonization and in mice lacking Peyer's patches.


Invasion requires the expression of several nonfimbrial adhesins, such as invasin (Inv) and—in *Y. pseudotuberculosis*—*Yersinia* adhesin A (YadA). Inv interacts directly with β1 integrins, which are expressed on the apical surfaces of M cells but not enterocytes. YadA of *Y. pseudotuberculosis* interacts with extracellular matrix proteins such as collagen and fibronectin to facilitate host cell integrin association and invasion. YadA of *Y. enterocolitica* lacks a crucial N-terminal region and binds collagen and laminin but not fibronectin and does not cause invasion. Inv is chromosomally encoded, whereas YadA is encoded on the virulence plasmid pYV. YadA also helps to confer serum resistance in *Y. enterocolitica* by binding host complement regulators such as factor H and C4-binding protein. Another chromosomal gene, *ail* (attachment and invasion locus), encodes the extracellular protein Ail, which is the main factor conferring serum resistance in *Y. pseudotuberculosis* by binding these complement regulators.

By binding to host cell surfaces, YadA allows targeting of immune effector cells by the pYV plasmid–encoded type III secretion system (injectisome). As a consequence, the host's innate immune response is altered; toxins (*Yersinia* outer proteins, or Yops) are injected into host macrophages, neutrophils, and dendritic cells, affecting signal transduction pathways, resulting in reduced phagocytosis and inhibited production of reactive oxygen species by neutrophils, and triggering apoptosis of macrophages. Other factors functional in invasive disease include yersiniabactin (Ybt), a siderophore produced by some strains of *Y. pseudotuberculosis* and *Y. enterocolitica* as well as other Enterobacteriaceae. Ybt allows bacteria to access iron from saturated lactoferrin during infection and reduces production of reactive oxygen species by innate immune effector cells, thereby decreasing bacterial killing. *Y. pseudotuberculosis* and *Y. pestis* make other siderophores apart from Ybt.

CLINICAL MANIFESTATIONS

Self-limiting diarrhea is the most common reported presentation in infection with pathogenic *Y. enterocolitica*, especially in children <4 years of age, who form the single largest group in most case series. Blood may be detected in diarrheal stool. Older children and adults are more likely than younger children to present with abdominal pain, which can be localized to the right iliac fossa—a situation that often leads to laparotomy for presumed appendicitis (pseudoappendicitis). Appendectomy is not indicated for *Yersinia* infection causing pseudoappendicitis. Thickening of the terminal ileum and cecum is seen on endoscopy and ultrasound, with elevated round or oval lesions that may overlie Peyer's patches. Mesenteric lymph nodes are enlarged. Ulcerations of the mucosa are noted on endoscopy. Gastrointestinal complications include granulomatous appendicitis, a chronic inflammatory condition affecting the appendix that is responsible for ≤2% of cases of appendicitis; *Yersinia* is involved in a minority of cases. *Y. enterocolitica* infection can present as acute pharyngitis with or without other gastrointestinal symptoms. Fatal *Y. enterocolitica* pharyngitis has been recorded. Mycotic aneurysm can follow *Y. enterocolitica* bacteremia, as can focal infection (abscess) in many other sites and body compartments (liver, spleen, kidney, bone, meninges, endocardium).


In all age groups, *Y. pseudotuberculosis* infection is more likely to present as abdominal pain and fever than as diarrhea. A superantigenic toxin —*Y. pseudotuberculosis* mitogen (YPM)—is produced by strains seen in eastern Russia in association with Far Eastern scarlet-like fever, a childhood illness with desquamating rash, arthralgia, and toxic shock. A similar illness is recognized in Japan (Izumi fever) and Korea. Similarities have been noted with Kawasaki disease, the idiopathic acute systematic vasculitis of childhood. There is an epidemiologic link between exposure of populations to superantigen-positive *Y. pseudotuberculosis* and an elevated incidence of Kawasaki disease.

Y. enterocolitica or *Y. pseudotuberculosis* septicemia presents as a severe illness with fever and leukocytosis, often without localizing features, and is significantly associated with predisposing conditions such as diabetes mellitus, liver disease, and iron overload. Hemochromatosis combines several of these risk factors. Administration of iron chelators like desferrioxamine, which provide iron accessible to *Yersinia* (and have an inhibitory effect on neutrophil function), may result in *Yersinia* septicemia in patients with iron overload who presumably have an otherwise mild gastrointestinal infection. HIV/AIDS has been associated with *Y. pseudotuberculosis* septicemia. The unusual phenomenon of transfusion-associated septicemia is linked to the ability of *Y. enterocolitica* to multiply at refrigerator temperature (psychrotrophy). Typically, the transfused unit has been stored for >20 days, and it is believed that small numbers of yersiniae from an apparently healthy donor with subclinical bacteremia are amplified to very high numbers by growth inside the bag at ≤4°C, with consequent septic shock after transfusion. A method for preventing this very rare event (i.e., a range of 1 case in 500,000 to 1 case in several million transfused units in countries such as the United States and France) without unacceptable restriction in the blood supply has not yet been devised.

POSTINFECTIVE PHENOMENA

As in other invasive intestinal infections (salmonellosis, shigellosis), reactive arthritis (articular arthritis of multiple joints developing within 2–4 weeks of a preceding infection) occurs as a result of autoimmune activity initiated by the deposition of bacterial components (not viable bacteria) in joints in combination with the immune response to invading bacteria. The majority of individuals affected by reactive arthritis due to *Yersinia* are HLA-B27 positive. Myocarditis with electrocardiographic ST-segment abnormalities may occur with *Yersinia*-associated reactive arthritis. Most *Yersinia*-associated cases follow *Y. enterocolitica* infection (presumably because it is more common than infection with other species), but *Y. pseudotuberculosis*–associated reactive arthritis is also well documented in Finland, where sporadic and outbreak infections with *Y. pseudotuberculosis* are more common than in other countries. Of infected individuals identified in a recent *Y. pseudotuberculosis* serotype O:3 outbreak in Finland, 12% developed reactive arthritis affecting the small joints of the hands and feet, knees, ankles, and shoulders and lasting >6 months in most cases. Erythema nodosum **(Fig. A1-39)** occurs after *Yersinia* infection (more commonly in women) with no evidence of HLA-B27 linkage.

There is a long-standing association between antithyroid and anti-*Yersinia* antibodies. Antibody evidence of prior *Y. enterocolitica* infection in Graves' disease and increased levels of antithyroid antibody in patients with *Y. enterocolitica* antibodies were first noted in the 1970s. *Y. enterocolitica* contains a thyroid-stimulating hormone (TSH)–binding site that is recognized by antibodies to TSH from Graves' disease patients. Raised titers of antibodies to *Y. enterocolitica* whole cells and Yops have been found in some series of Graves' disease patients but not in others. One Danish study of twins found no evidence of an association between asymptomatic *Yersinia* infection (as evidenced by titers of Yop antibody) and antithyroid antibodies in euthyroid individuals, while another Danish study of twins with and without Graves' disease found that increased titers of Yop antibody were associated with Graves' disease. It remains unclear whether this cross-reactivity is significant in the etiology of Graves' disease.

LABORATORY DIAGNOSIS

Standard laboratory culture methods can be used to isolate enteropathogenic *Yersinia* species from sterile samples, including blood and cerebrospinal fluid. Culture on specific selective media (CIN agar), with or without pre-enrichment in broth or phosphate-buffered saline at either 4°C or 16°C, is the basis of most schema for isolation of yersiniae from stool or other nonsterile samples. Outside known high-incidence areas, specific culture may be carried out by laboratories only upon request. Matrix-assisted laser desorption ionization time of flight (MALDI-TOF) mass spectrometry systems can speciate isolates of *Y. enterocolitica* and *Y. pseudotuberculosis*. Virulence plasmid–negative strains of *Y. enterocolitica* can be isolated from cultures of stool from asymptomatic individuals, especially after cold enrichment. These strains usually differ in biotype (typically biovar 1a) from virulence plasmid–possessing strains; although some display apparent pathogenicity in a mouse model and all are pathogenic in an insect model, virulence plasmid–negative strains are not commonly accepted as human pathogens. Because of the frequency with which the virulence plasmid is lost on laboratory subculture, combined biochemical identification (with biotyping according to a standard schema) and serologic identification are usually required to interpret the significance of an isolate of *Y. enterocolitica* from a nonsterile site. Most pathogenic *Y. enterocolitica* strains currently isolated from humans are of serogroup O:3/biovar 4 or serogroup O:9/biovar 2; this pattern holds even in the United States, where serogroup O:8/biovar 1B strains were previously predominant. Several CE-marked, FDA-approved multiplex real-time PCR kits now include *Y. enterocolitica* as one of the pathogens detectable in human feces; the precise assay targets are not disclosed. A standard for PCR detection of pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis* in food samples is available from the International Organization for Standardization.

Agglutinating or ELISA antibody titers to specific O-antigen types are used in the retrospective diagnosis of both *Y. enterocolitica* and *Y. pseudotuberculosis* infections. IgA and IgG antibodies persist in patients with reactive arthritis. Serologic cross-reactions between *Y. enterocolitica* serogroup O:9 and *Brucella* are due to the similarity of their lipopolysaccharide structures. Multiple assays are required to cover even the predominant serogroups (*Y. enterocolitica* O:3, O5,27, and O:9; *Y. pseudotuberculosis* O:1a, O:1b, and O:3), and these assays are generally available only in reference laboratories. ELISA and western blot tests for antibodies to Yops, which are expressed by all pathogenic strains of *Y. enterocolitica* and *Y. pseudotuberculosis*, are also available; most of the positivity in these assays probably relates to previous infection with *Y. enterocolitica*.

TREATMENT

TREATMENT

Yersiniosis

Most cases of diarrhea caused by enteropathogenic *Yersinia* are self-limiting. Data from clinical trials do not support antimicrobial treatment for adults or children with *Y. enterocolitica* diarrhea. Systemic infections with bacteremia or focal infections outside the gastrointestinal tract generally require antimicrobial therapy. Infants <3 months of age with documented *Y. enterocolitica* infection may require antimicrobial treatment because of the increased likelihood of bacteremia in this age group. *Y. enterocolitica* strains nearly always express β-lactamases. Because of the relative rarity of systemic *Y. enterocolitica* infection, there are no clinical trial data to guide antimicrobial choice or to suggest the optimal dose and duration of therapy. On the basis of retrospective case series and in vitro sensitivity data, fluoroquinolone therapy is effective for bacteremia in adults; for example, ciprofloxacin is given at a typical dose of 500 mg twice daily by mouth or 400 mg twice daily IV for at least 2 weeks (longer if positive blood cultures persist). A third-generation cephalosporin is an alternative—e.g., cefotaxime (typical dose, 6–8 g/d in 3 or 4 divided doses) or ceftriaxone. In

children, third-generation cephalosporins are effective; for example, cefotaxime is given to children ≥1 month of age at a typical dose of 75–100 mg/kg per day in 3 or 4 divided doses, with an increase to 150–200 mg/kg per day in severe cases (maximal daily dose, 8–10 g). Amoxicillin and amoxicillin/clavulanate have shown poor efficacy in case series. Trimethoprim-sulfamethoxazole, gentamicin, and imipenem are all active in vitro. *Y. pseudotuberculosis* strains do not express β-lactamase but are intrinsically resistant to polymyxin. Because human infection with *Y. pseudotuberculosis* is less common than that with *Y. enterocolitica*, less case information is available; however, studies in mice suggest that ampicillin is ineffective. Drugs similar to those used against *Y. enterocolitica* should be used. The best results have been obtained with a quinolone.

Some trials of treatment for reactive arthritis (with a large proportion of cases due to *Yersinia*) found that 3 months of oral ciprofloxacin therapy did not affect outcome. One trial in which the same therapy was given specifically for *Y. enterocolitica*–reactive arthritis found that, while outcome indeed was not affected, there was a trend toward faster remission of symptoms in the treated group. Follow-up 4–7 years after initial antibiotic treatment of reactive arthritis (predominantly following *Salmonella* and *Yersinia* infections) demonstrated apparent efficacy in the prevention of chronic arthritis in HLA-B27-positive individuals. A trial showing that azithromycin therapy did not affect outcome in reactive arthritis included cases thought to have followed yersiniosis, although no breakdown of cases was provided.

PREVENTION AND CONTROL

Current control measures are similar to those used against other enteric pathogens like *Salmonella* and *Campylobacter*, which colonize the intestine of food animals. The focus is on safe handling and processing of food. No vaccine is effective in preventing intestinal colonization of food animals by enteropathogenic *Yersinia*. Consumption of food made from raw pork (which is popular in Germany and Belgium) should be discouraged at present because it is not possible to eliminate contamination with the enteropathogenic *Yersinia* strains found worldwide in pigs. Exposure of infants to raw pig intestine during domestic preparation of chitterlings is inadvisable. Modification of abattoir technique in Scandinavian countries from the 1990s onward included the removal of pig intestines in a closed plastic bag; levels of carcass contamination with *Y. enterocolitica* were reduced, but such contamination was not eliminated. Experimental pig herds free of pathogenic *Y. enterocolitica* O:3 (and also of *Salmonella*, *Campylobacter*, *Toxoplasma*, and *Trichinella*) have been established by selective breeding in Norway but remain rare. In the food industry, vigilance is required because of the potential for large outbreaks if small numbers of enteropathogenic yersiniae contaminate any ready-to-eat food whose safe preservation is based on refrigeration before consumption.

The rare phenomenon of contamination of blood for transfusion has proved impossible to eradicate. However, leukodepletion is now practiced in most blood transfusion centers, primarily to prevent nonhemolytic febrile transfusion reactions and alloimmunization against HLA antigens. This measure reduces but does not eliminate the risk of *Yersinia* blood contamination.

Notification of yersiniosis is now obligatory in some countries.

FURTHER READING

Plague

Achtman M: How old are bacterial pathogens? Proc Biol Sci B 283:pii:20160990, 2016.

Atkinson S, Williams P: *Yersinia* virulence factors—A sophisticated arsenal for combating host defences. F1000Res 5:pii:F1000 Faculty Rev-1370, 2016. [PubMed: 27347390]

Bertherat E: Plague around the world, 2010–2015. Wkly Epidemiol Rec 91:89, 2016.

Hinnebusch BJ et al: Ecological opportunity, evolution, and the emergence of flea-borne plague. Infect Immun 84:1932, 2016.

Kool JL: Risk of person-to-person transmission of pneumonic plague. Clin Infect Dis 40:1166, 2005.

Kugeler KJ et al: Epidemiology of human plague in the United States, 1900–2012. Emerg Infect Dis 21:16, 2015.

Riddell SW et al: Sentinel level clinical laboratory guidelines for suspected agents of bioterrorism and emerging infectious diseases: *Yersinia pestis*. Available from *https://www-asm-org.ezproxy1.library.usyd.edu.au/images/PSAB/LRN/Ypestis316.pdf*. Accessed January 5, 2018.

Runfola JK et al: Outbreak of human pneumonic plague with dog-to-human and possible human-to-human transmission—Colorado, June–July 2014. MMWR Morb Mortal Wkly Rep 64:429, 2015.

Verma SK, Tuteja U: Plague vaccine development: Current research and future trends. Front Immunol 7:602, 2016.

Yersiniosis

Reuter S et al: Parallel evolution of pathogenicity within the genus *Yersinia*. Proc Natl Acad Sci USA 111:6768, 2014.

Chapter 167: *Bartonella* Infections, Including Cat-Scratch Disease

Michael Giladi; Moshe Ephros

INTRODUCTION

Bartonella species are fastidious, facultative intracellular, slow-growing, gram-negative bacteria that cause a broad spectrum of diseases in humans. This genus includes ~40 distinct species or subspecies, of which at least 16 have been recognized as confirmed or potential human pathogens; *Bartonella bacilliformis*, *Bartonella quintana*, and *Bartonella henselae* are most commonly identified **(Table 167-1)**. Most *Bartonella* species have successfully adapted to survival in specific domestic or wild mammals. Prolonged intraerythrocytic infection in these animals creates a niche where the bacteria are protected from both innate and adaptive immunity and which serves as a reservoir for human infections. *Bartonella* characteristically evades the host immune system by modification of its virulence factors (e.g., lipopolysaccharides or flagella) and by attenuation of the immune response. *B. bacilliformis* and *B. quintana*, which are not zoonotic, are exceptions. Arthropod vectors are often involved. Isolation and characterization of *Bartonella* species are difficult and require special techniques. Clinical presentation generally depends on both the infecting *Bartonella* species and the immune status of the infected individual. *Bartonella* species are susceptible to many antibiotics in vitro; however, clinical responses to therapy and studies in animal models suggest that the minimal inhibitory concentrations of many antimicrobial agents correlate poorly with the drugs' in vivo efficacies in patients with *Bartonella* infections.

TABLE 167-1

***Bartonella* Species Known or Suspected to Be Human Pathogens**

Bartonella Species[a]	Disease(s)[b]	Reservoir Host(s)[c]	Arthropod Vector
B. henselae	Cat-scratch disease, bacillary angiomatosis, bacillary peliosis, bacteremia, endocarditis	Cats, other felines	Cat fleas (*Ctenocephalides felis*): associated with cat-to-cat, but not with cat-to-human, transmission
B. quintana	Trench fever, chronic bacteremia, bacillary angiomatosis, endocarditis	Humans	Human body lice (*Pediculus humanus corporis*)
B. bacilliformis	Carrión's disease	Humans	Sandflies (*Lutzomyia verrucarum*)
B. elizabethae	Endocarditis	Rats, dogs	Unknown
B. grahamii[d]	Lymphadenopathy	Mice, voles	Fleas
B. vinsonii subsp. *arupensis*	Endocarditis, febrile illness	Mice, dogs	Ticks
B. vinsonii subsp. *berkhoffii*	Endocarditis	Domestic dogs, coyotes, gray foxes	Ticks
B. washoensis	Myocarditis, meningitis	Squirrels, possibly other rodents	Fleas
B. alsatica	Endocarditis, lymphadenitis	Rabbits	Fleas
B. koehlerae	Endocarditis	Cats	Unknown
B. clarridgeiae	Possibly cat-scratch disease	Cats	Unknown

Bartonella Species[a]	Disease(s)[b]	Reservoir Host(s)[c]	Arthropod Vector
B. rochalimae	Bacteremia, fever, splenomegaly	Unknown	Possibly fleas
B. tamiae	Bacteremia, fever, myalgia, rash	Unknown	Unknown
B. melophagi	Various clinical manifestations	Sheep	Sheep keds
B. ancashensis	Verruga peruana	Unknown	Unknown
Candidatus B. mayotimonensis[e]	Endocarditis	Bats	Unknown

[a]Many other *Bartonella* species exist but are not recognized as human pathogens. [b]Animal-associated *Bartonella* species (*B. henselae, B. doshiae, B. schoenbuchensis*, and *B. tribocorum*) were isolated from blood of patients who reported tick bites and chronic symptoms such as fatigue and myalgia. DNA of *B. henselae, B. vinsonii* subsp. *berkhoffii, B. koehlerae*, or *B. melophagi* or co-infection with more than one *Bartonella* species was detected by PCR in blood samples from patients with extensive arthropod and animal exposure who presented with chronic neurologic or neurocognitive syndromes. The causal relationship between bacteremia with these pathogens, tick bites, and clinical manifestations needs to be established. [c]Animals are implicated when existing evidence supports their infection with *Bartonella* species. Data supporting animal-to-human transmission may be lacking. [d]Retinitis may also be associated with *B. grahamii*. [e]*Candidatus* is a taxonomic status for bacteria that cannot be described in sufficient detail to warrant establishment of a novel taxon or cannot be cultured or propagated in culture media. The phylogenetic relatedness of these bacteria has been determined by gene amplification and sequence analysis.

CAT-SCRATCH DISEASE

DEFINITION AND ETIOLOGY

Usually a self-limited illness, cat-scratch disease (CSD) has two general clinical presentations. *Typical* CSD, the more common, is characterized by subacute regional lymphadenopathy; *atypical* CSD is the collective designation for numerous extranodal manifestations involving various organs. *B. henselae* is the principal etiologic agent of CSD. Rare cases have been associated with *Afipia felis* and other *Bartonella* species.

EPIDEMIOLOGY

CSD occurs worldwide, favoring warm and humid climates. In temperate climates, incidence peaks during fall and winter; in the tropics, disease occurs year-round. Adults are affected nearly as frequently as children. Intrafamilial clustering is rare, and person-to-person transmission does not occur. Apparently healthy cats constitute the major reservoir of *B. henselae*, and cat fleas (*Ctenocephalides felis*) may be responsible for cat-to-cat transmission. CSD usually follows contact with cats (especially kittens), but other animals (e.g., dogs) have been implicated as possible reservoirs in rare instances. In the United States, the estimated annual disease incidence is ~4–10 cases per 100,000 population. About 5–10% of patients are hospitalized.

PATHOGENESIS

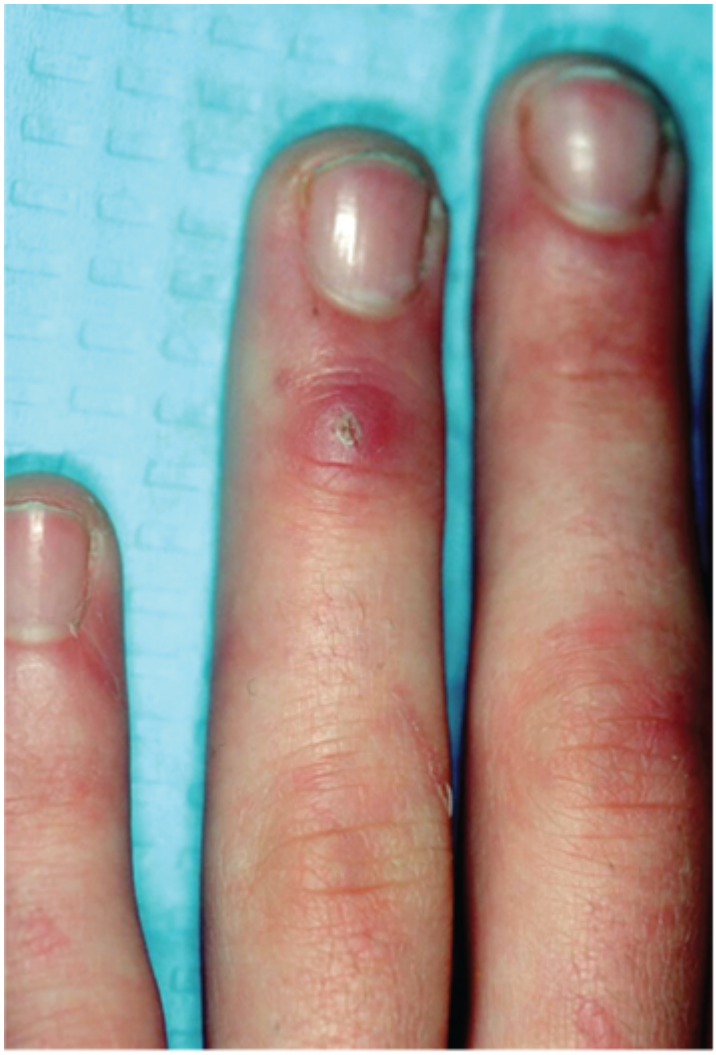
Inoculation of *B. henselae*, possibly via contaminated flea feces, usually results from a cat scratch or bite. Infection of mucous membranes or conjunctivae via droplets or licking may occur as well. With lymphatic drainage to one or more regional lymph nodes in immunocompetent hosts, a T_H1 response can result in necrotizing granulomatous lymphadenitis. Dendritic cells, along with their associated chemokines, play a role in the host inflammatory response and granuloma formation.

CLINICAL MANIFESTATIONS AND PROGNOSIS

Of patients with CSD, 85–90% have *typical* disease. The primary lesion, a small (0.3- to 1-cm) painless erythematous papule or pustule, develops at the inoculation site within days to 2 weeks in about one-third to two-thirds of patients **(Fig. 167-1A, B)**. Lymphadenopathy develops 1–3 weeks or longer after cat contact. The affected lymph node(s) are enlarged and usually painful, sometimes have overlying erythema, and suppurate in 10–15% of cases **(Fig. 167-1C, D, and E)**. Axillary/epitrochlear nodes are most commonly involved; next in frequency are head/neck nodes and then inguinal/femoral nodes. Approximately 50% of patients have fever, malaise, and anorexia. A smaller proportion experience weight loss and night sweats mimicking the presentation of lymphoma. Fever is usually low-grade but infrequently rises to ≥39°C. Resolution is slow, requiring weeks (for fever, pain, and accompanying signs and symptoms) to months (for node shrinkage).

FIGURE 167-1

Manifestations of cat-scratch disease. ***A.*** Primary inoculation lesion. Axillary and epitrochlear lymphadenitis appeared 2 weeks later. ***B.*** Primary inoculation lesion. Submental lymphadenitis appeared 10 days later. ***C.*** Axillary lymphadenopathy of 2 weeks' duration. The overlying skin appears normal. ***D.*** Cervical lymphadenopathy of 6 weeks' duration. The overlying skin is red. Thick, odorless pus (12 mL) was aspirated. ***E.*** Preauricular lymphadenopathy. ***F.*** Left-eye neuroretinitis. Note papilledema and stellate macular exudates ("macular star").



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



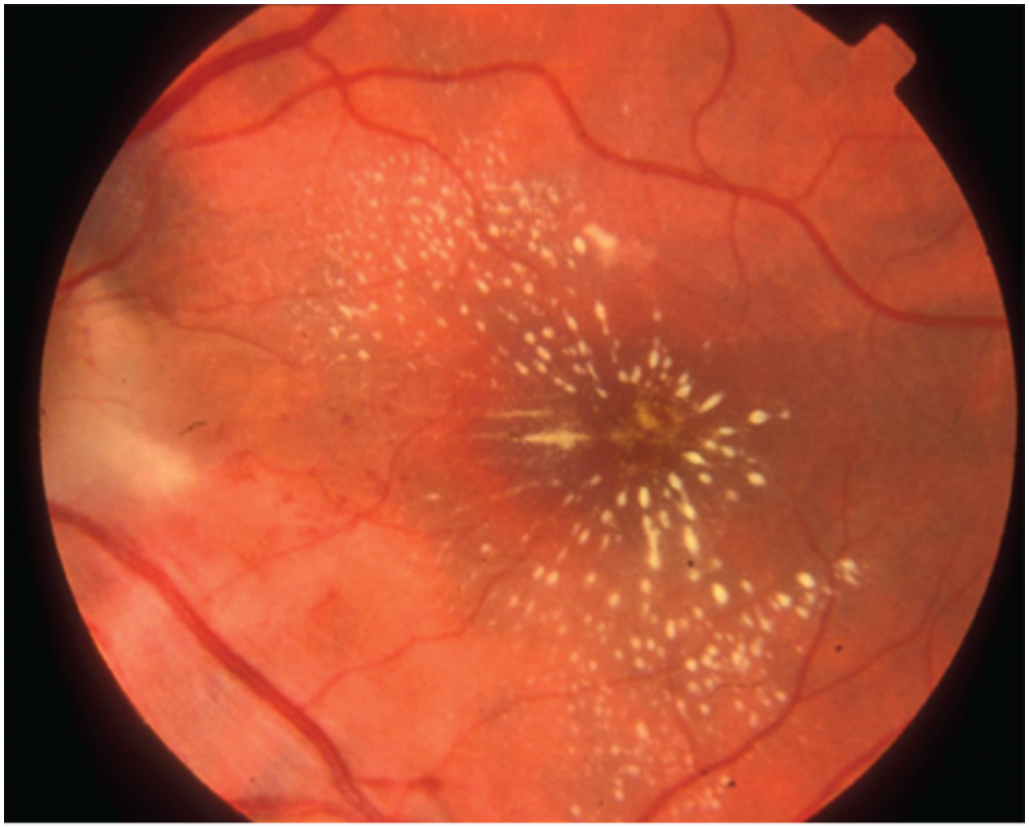
D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



E

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



F

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Atypical CSD occurs in 10–15% of patients as extranodal or complicated disease in the absence or presence of lymphadenopathy. Atypical disease includes Parinaud's oculoglandular syndrome (granulomatous conjunctivitis with ipsilateral preauricular lymphadenitis; Fig. 167-1E), granulomatous hepatitis/splenitis, neuroretinitis (often presenting as unilateral deterioration of vision; **Fig. 167-1F**), and other ophthalmologic manifestations. In addition, neurologic involvement (encephalopathy, seizures, myelitis, radiculitis, cerebellitis, facial and other cranial or peripheral palsies), fever of unknown origin, debilitating myalgia, arthritis or arthralgia (affecting mostly women >20 years old), osteomyelitis (including multifocal disease), tendinitis, neuralgia, and dermatologic manifestations (including erythema nodosum **[see Fig. A1-39]**, sometimes accompanying arthropathy) occur. Other manifestations and syndromes (pneumonitis, pleural effusion, idiopathic thrombocytopenic purpura, Henoch-Schönlein purpura, erythema multiforme **[see Fig. A1-24]**, hypercalcemia, glomerulonephritis, myocarditis) have also been associated with CSD. In elderly patients (>60 years old), lymphadenopathy is more often absent but encephalitis and fever of unknown origin are more common than in younger patients. In immunocompetent individuals, CSD—whether typical or atypical—usually resolves without treatment and without sequelae, although some of the ophthalmologic manifestations may occasionally result in moderate to severe vision loss. Lifelong immunity is the rule.

DIAGNOSIS

Routine laboratory tests usually yield normal or nonspecific results. Histopathology initially shows lymphoid hyperplasia and later demonstrates stellate granulomata with necrosis, coalescing microabscesses, and

occasional multinucleated giant cells—findings that, although nonspecific, may narrow the differential diagnosis. Serologic testing (immunofluorescence or enzyme immunoassay) is the most commonly used laboratory diagnostic approach, with variable sensitivity and specificity. CSD serodiagnosis is often based on the presence of IgG alone (i.e., in the absence of IgM), and seroconversion may take a few weeks; these two factors may pose difficulties in the interpretation of serologic results. Other tests are of low sensitivity (culture, Warthin-Starry silver staining), of low specificity (cytology, histopathology), or of limited availability in routine diagnostic laboratories (polymerase chain reaction [PCR], immunohistochemistry). PCR of pus aspirated from lymph nodes or the primary inoculation lesion is highly sensitive and specific and is particularly useful for definitive and rapid diagnosis in seronegative patients. PCR of a lymph node biopsy specimen may be less sensitive, perhaps because of sampling error.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Cat-Scratch Disease

A history of cat contact, a primary inoculation lesion, and regional lymphadenopathy—especially axillary/epitrochlear lymphadenopathy—are highly suggestive of CSD. A characteristic clinical course and corroborative laboratory tests make the diagnosis very likely. Conversely, when acute- and convalescent-phase sera are negative (as is the case in 10–20% of CSD patients), when spontaneous regression of lymph node size does not occur, and particularly when constitutional symptoms persist, malignancy must be ruled out. Pyogenic lymphadenitis, mycobacterial infection, brucellosis, syphilis, tularemia, plague, toxoplasmosis, sporotrichosis, and histoplasmosis should also be considered. In clinically suspected CSD in a seronegative individual, fine-needle aspiration may be adequate and PCR can confirm the diagnosis. When data are less supportive of CSD, lymph node biopsy rather than fine-needle aspiration is preferred. In seronegative CSD patients with lymphadenopathy and severe complications (e.g., encephalitis or neuroretinitis), early biopsy is important to establish a specific diagnosis.

TREATMENT

TREATMENT

Cat-Scratch Disease

Treatment regimens are based on only minimal data **(Table 167-2)**. Suppurative nodes should be drained by large-bore needle aspiration and not by incision and drainage in order to avoid chronic draining tracts. Immunocompromised patients must always be treated with systemic antimicrobials.

TABLE 167-2

Antimicrobial Therapy for Disease Caused by *Bartonella* Species in Adults

Disease	Antimicrobial Therapy
Typical cat-scratch disease	Not routinely indicated; for patients with extensive lymphadenopathy, consider azithromycin (500 mg PO on day 1, then 250 mg PO once a day for 4 days)
Cat-scratch disease neuroretinitis	Value of systemic antibiotics is controversial, particularly when visual acuity is not significantly compromised. For more severe cases, doxycycline (100 mg PO bid) *plus* rifampin (300 mg PO bid) for 4–6 weeks is given. Consider adding systemic glucocorticoids.
Other atypical cat-scratch disease manifestations[a]	As per neuroretinitis. Treatment duration should be individualized.
Trench fever or chronic bacteremia with *B. quintana*	Gentamicin (3 mg/kg IV once a day for 14 days) *plus* doxycycline (200 mg PO once a day or 100 mg PO bid for 6 weeks)
Suspected *Bartonella* endocarditis	Gentamicin[b] (1 mg/kg IV q8h for ≥14 days) *plus* doxycycline (100 mg PO/IV bid for 6 weeks[c]) *plus* ceftriaxone (2 g IV once a day for 6 weeks)
Confirmed *Bartonella* endocarditis	As for suspected *Bartonella* endocarditis *minus* ceftriaxone
Bacillary angiomatosis	Erythromycin[d] (500 mg PO qid for 3 months) *or* Doxycycline (100 mg PO bid for 3 months)
Bacillary peliosis	Erythromycin[d] (500 mg PO qid for 4 months) *or* Doxycycline (100 mg PO bid for 4 months)

Disease	Antimicrobial Therapy
Carrión's disease	
Oroya fever	Chloramphenicol (500 mg PO/IV qid for 14 days) *plus* another antibiotic (β-lactam preferred) *or* Ciprofloxacin (500 mg PO bid for 10 days)
Verruga peruana	Rifampin (10 mg/kg PO once a day, to a maximum of 600 mg, for 14 days) *or* Streptomycin (15–20 mg/kg IM once a day for 10 days)

[a]Data on treatment efficacy for encephalitis and hepatosplenic cat-scratch disease are lacking. Therapy similar to that given for neuroretinitis is reasonable. [b]Some experts recommend gentamicin at 3 mg/kg IV once a day. If gentamicin is contraindicated, rifampin (300 mg PO bid) can be added to doxycycline for documented *Bartonella* endocarditis. [c]Some experts recommend extending oral doxycycline therapy for 3–6 months. [d]Other macrolides are probably effective and may be substituted for erythromycin or doxycycline.

Source: Recommendations are modified from JM Rolain et al: Antimicrob Agents Chemother 48:1921, 2004.

PREVENTION

Avoiding cats (especially kittens) and instituting flea control are options for immunocompromised patients and for patients with valvular heart disease.

TRENCH FEVER AND CHRONIC BACTEREMIA

DEFINITION AND ETIOLOGY

Trench fever, also known as *5-day fever* or *quintan fever*, is a febrile illness caused by *B. quintana*. It was first described as an epidemic in the trenches of World War I; however, recent paleomicrobiologic studies have provided evidence that *B. quintana* has been associated with human infection for 4000 years. This infection recently reemerged as chronic bacteremia seen most often in homeless people, also referred to as *urban* or *contemporary trench fever*.

EPIDEMIOLOGY

In addition to epidemics during World Wars I and II, sporadic outbreaks of trench fever have been reported in many regions of the world. The human body louse has been identified as the vector and humans as the only known reservoir. After a hiatus of several decades during which trench fever was almost forgotten, small clusters of cases of *B. quintana* chronic bacteremia were reported sporadically, primarily from the United States and France, in HIV-uninfected homeless people. Alcoholism and louse infestation were identified as risk factors.

CLINICAL MANIFESTATIONS

The typical incubation period is 15–25 days (range, 3–38 days). "Classical" trench fever, as described in 1919, ranges from a mild febrile illness to a recurrent or protracted and debilitating disease. Onset may be abrupt or preceded by a prodrome of several days. Fever is often periodic, lasting 4–5 days with 5-day (range, 3- to 8-day) intervals between episodes. Other symptoms and signs include headache, back and limb pain, profuse sweating, shivering, myalgia, arthralgia, splenomegaly, a maculopapular rash in occasional cases, and nuchal rigidity in some cases. Untreated, the disease usually lasts 4–6 weeks. Death is rare. The clinical spectrum of *B. quintana* bacteremia in homeless people ranges from asymptomatic infection to a febrile illness with headache, severe leg pain, and thrombocytopenia. Endocarditis sometimes develops.

DIAGNOSIS

Definitive diagnosis requires isolation of *B. quintana* by blood culture. Some patients have positive blood cultures for several weeks. Patients with acute trench fever typically develop significant titers of antibody to *Bartonella*, whereas those with chronic *B. quintana* bacteremia may be seronegative. Patients with high titers of IgG antibodies should be evaluated for endocarditis. In epidemics, trench fever should be differentiated from epidemic louse-borne typhus and relapsing fever, which occur under similar conditions and share many features.

TREATMENT

TREATMENT

B. quintana Bacteremia

In a small, randomized, placebo-controlled trial involving homeless people with *B. quintana* bacteremia, therapy with gentamicin and doxycycline was superior to administration of placebo in eradicating bacteremia (Table 167–2). Treatment of bacteremia is important, even in clinically mild cases, to prevent endocarditis. Optimal therapy for trench fever without documented bacteremia is uncertain.

BARTONELLA ENDOCARDITIS

DEFINITION AND ETIOLOGY

Coxiella burnetii (Chap. 182) and *Bartonella* species are the most common pathogens in culture-negative endocarditis (Chap. 123). In France, for example, *Bartonella* species were identified as the etiologic agents in 28% of 348 cases of culture-negative endocarditis. Prevalence, however, varies by geographic location and epidemiologic setting. In addition to *B. quintana* and *B. henselae* (the most common *Bartonella* species implicated in endocarditis, the former more commonly than the latter), other *Bartonella* species have reportedly caused rare cases (Table 167-1).

EPIDEMIOLOGY

Bartonella endocarditis has been reported worldwide. Most patients are adults; more are male than female. Risk factors associated with *B. quintana* endocarditis include homelessness, alcoholism, and body louse infestation; however, individuals with no risk factors have had *Bartonella* endocarditis diagnosed as well. *B. henselae* endocarditis is associated with exposure to cats. Most cases involve native rather than prosthetic valves; the aortic valve accounts for ~60% of cases. Patients with *B. henselae* endocarditis usually have preexisting valvulopathy, whereas *B. quintana* often infects normal valves.

CLINICAL MANIFESTATIONS

Clinical manifestations are usually characteristic of subacute endocarditis of any etiology. However, a substantial number of patients have a prolonged, minimally febrile or even afebrile indolent illness, with mild nonspecific symptoms lasting weeks or months before the diagnosis is made. Initial echocardiography may not show vegetations. Acute, aggressive disease is rare.

DIAGNOSIS

Blood cultures, even with use of special techniques (lysis centrifugation or EDTA-containing tubes), are positive in only ~25% of cases—mostly those caused by *B. quintana* and only rarely those caused by *B. henselae*. Prolonged incubation of cultures (up to 6 weeks) is required. Serologic tests—either immunofluorescence or enzyme immunoassay—usually demonstrate high-titer IgG antibodies to *Bartonella*. Because of cross-antigenicity, routine serology does not distinguish between *B. quintana* and *B. henselae* and may also be low-titer cross-reactive with other pathogens, such as *C. burnetii* and *Chlamydia* species. Identification of *Bartonella* to the species level is usually accomplished by application of PCR and DNA sequencing methods to valve tissue.

TREATMENT

TREATMENT

Bartonella Endocarditis

For patients with culture-negative endocarditis suspected to be due to *Bartonella* species, empirical treatment consists of gentamicin, doxycycline, and ceftriaxone; the major role of ceftriaxone in this regimen

is to adequately treat other potential causes of culture-negative endocarditis, including members of the HACEK group **(Chap. 153)** (Table 167-2). Once a diagnosis of *Bartonella* endocarditis has been established, ceftriaxone is discontinued. Aminoglycosides, the only antibiotics known to be bactericidal against *Bartonella*, should be included in the regimen for ≥2 weeks. Indications for valvular surgery are the same as in subacute endocarditis due to other pathogens; however, the proportion of patients who undergo surgery (~60%) is high, probably as a consequence of delayed diagnosis.

BACILLARY ANGIOMATOSIS AND PELIOSIS

DEFINITION AND ETIOLOGY

Bacillary angiomatosis (sometimes called *bacillary epithelioid angiomatosis* or *epithelioid angiomatosis*) is a disease of severely immunocompromised patients, is caused by *B. henselae* or *B. quintana*, and is characterized by neovascular proliferative lesions involving the skin and other organs. Both species cause cutaneous lesions; hepatosplenic lesions are caused only by *B. henselae*, whereas subcutaneous and lytic bone lesions are more frequently associated with *B. quintana*. Bacillary peliosis is a closely related angioproliferative disorder caused by *B. henselae* and involving primarily the liver (peliosis hepatis) but also the spleen and lymph nodes. Bacillary peliosis is characterized by blood-filled cystic structures whose size ranges from microscopic to several millimeters.

EPIDEMIOLOGY

Bacillary angiomatosis and bacillary peliosis occur primarily in HIV-infected persons **(Chap. 197)** with CD4+ T cell counts of <100/μL but also affect other immunosuppressed patients and, in rare instances, immunocompetent patients. The incidence has decreased since the introduction of effective antiretroviral therapy and the routine use of rifabutin and macrolides to prevent *Mycobacterium avium* complex infection in AIDS patients. Contact with cats or cat fleas increases the risk of *B. henselae* infection. Risk factors for *B. quintana* infection are low income, homelessness, and body louse infestation.

CLINICAL MANIFESTATIONS

Bacillary angiomatosis presents most commonly as one or more cutaneous lesions that are not painful and that may be tan, red, or purple in color. Subcutaneous masses or nodules, superficial ulcerated plaques **(Fig. 167-2)**, and verrucous growths are also seen. Nodular forms resemble those seen in fungal or mycobacterial infections. Subcutaneous nodules are often tender. Painful osseous lesions, most often involving long bones, may underlie cutaneous lesions and occasionally develop in their absence. In rare cases, other organs are involved in bacillary angiomatosis. Patients usually have constitutional symptoms, including fever, chills, malaise, headache, anorexia, weight loss, and night sweats. In osseous disease, lytic lesions are generally seen on radiography, and technetium scan shows focal uptake. The differential diagnosis of cutaneous bacillary angiomatosis includes Kaposi's sarcoma, pyogenic granuloma, subcutaneous tumors, and verruga peruana. In bacillary peliosis, hypodense hepatic areas are usually evident on imaging. In patients with

advanced immunodeficiency, *B. henselae* and *B. quintana* are important causes of fever of unknown origin. Intermittent bacteremia with positive blood cultures can occur with or without endocarditis.

FIGURE 167-2

Nodular lesion of bacillary angiomatosis with superficial ulceration in an AIDS patient with advanced immunodeficiency. *(Reprinted with permission from DH Spach and E Darby: Bartonella Infections, Including Cat-Scratch Disease, in Harrison's Principles of Internal Medicine, 17th ed, AF Fauci et al [eds]. New York, McGraw-Hill, 2008, p 989.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOLOGY

Bacillary angiomatosis consists of lobular proliferations of small blood vessels lined by enlarged endothelial cells interspersed with mixed infiltrates of neutrophils and lymphocytes, with predominance of the former. Histologic examination of organs with bacillary peliosis reveals small blood-filled cystic lesions partially lined by endothelial cells that can be several millimeters in size. Peliotic lesions are surrounded by fibromyxoid stroma containing inflammatory cells, dilated capillaries, and clumps of granular material. Warthin-Starry silver staining of bacillary angiomatosis and peliosis lesions reveals clusters of bacilli. Cultures are usually negative.

DIAGNOSIS

Bacillary angiomatosis and bacillary peliosis are diagnosed on histologic grounds. Blood cultures may be positive.

TREATMENT

TREATMENT

Bacillary Angiomatosis and Peliosis

Prolonged therapy with a macrolide or doxycycline is recommended for both bacillary angiomatosis and bacillary peliosis (Table 167-2).

PREVENTION

Reasonable strategies for HIV-infected persons consist of control of cat-flea infestation and avoidance of cat scratches (for prevention of *B. henselae*) and avoidance and treatment of body louse infestation (for prevention of *B. quintana*). Primary prophylaxis is not recommended, but suppressive therapy with a macrolide or doxycycline is indicated in HIV-infected patients with bacillary angiomatosis or bacillary peliosis until CD4+ T cell counts are >200/μL. Relapse may necessitate lifelong suppressive therapy in individual cases.

CARRIÓN'S DISEASE (OROYA FEVER AND VERRUGA PERUANA)

DEFINITION AND ETIOLOGY

Carrión's disease is a biphasic disease caused by *B. bacilliformis*. *Oroya fever* is the initial, bacteremic, systemic form, and *verruga peruana* is its late-onset, eruptive manifestation.

EPIDEMIOLOGY AND PREVENTION

 Infection is endemic to the geographically restricted Andes valleys of Peru, Ecuador, and Colombia (~500–3200 m above sea level). Sporadic epidemics occur. The disease is transmitted by the phlebotomine sandfly *Lutzomyia verrucarum*. Humans are the only known reservoir of *B. bacilliformis*. Sandfly control measures (e.g., insecticides) and personal protection measures (e.g., repellents, screening, bed nets) may decrease the risk of infection.

PATHOGENESIS

After inoculation by the sandfly, bacteria invade the blood vessel endothelium and proliferate; the reticuloendothelial system and various organs may also be involved. Upon reentry into blood vessels, *B. bacilliformis* invades, replicates, and ultimately destroys erythrocytes, with consequent massive hemolysis and sudden, severe anemia. Microvascular thrombosis results in end-organ ischemia. Survivors sometimes develop cutaneous hemangiomatous lesions characterized by various inflammatory cells, endothelial proliferation, and the presence of *B. bacilliformis*.

CLINICAL MANIFESTATIONS

The incubation period is 3 weeks (range, 2–14 weeks). Oroya fever may present as a nonspecific bacteremic febrile illness without anemia or as an acute, severe hemolytic anemia with hepatomegaly and jaundice of rapid onset leading to vascular collapse and clouded sensorium. Myalgia, arthralgia, lymphadenopathy, and abdominal pain may develop. Temperature is elevated but not extremely so; high fever may suggest intercurrent infection. Subclinical asymptomatic infection also occurs. In verruga peruana, red, hemangioma-like, cutaneous vascular lesions of various sizes appear either weeks to months after systemic illness or with no previous suggestive history. These lesions persist for months up to 1 year. Mucosal and internal lesions may also develop.

DIAGNOSIS AND APPROACH TO THE PATIENT

Systemic illness (with or without anemia) or the development of cutaneous lesions in a person who has been to an endemic area raises the possibility of *B. bacilliformis* infection. Severe anemia with exuberant reticulocytosis—and sometimes thrombocytopenia—can occur. In systemic illness, Giemsa-stained blood films may show typical intraerythrocytic bacilli, and blood and bone marrow cultures are positive. Serologic assays may be helpful. Biopsy may be required to confirm the diagnosis of verruga peruana. Differential diagnosis includes the spectrum of coendemic systemic febrile illnesses (e.g., typhoid fever, malaria, brucellosis) as well as diseases producing cutaneous vascular lesions (e.g., hemangiomata, bacillary angiomatosis, Kaposi's sarcoma).

TREATMENT

TREATMENT

Carrión's Disease

Antibiotic therapy for systemic *B. bacilliformis* infection usually results in rapid defervescence (Table 167-2). Additional antibiotic treatment of intercurrent infection (particularly salmonellosis) is often required. Blood transfusion may be necessary. Treatment of verruga peruana usually is not required, although large lesions or those interfering with function may require excision. Patients with numerous lesions, especially lesions that have been present for only a short period, may respond well to antibiotic therapy.

COMPLICATIONS AND PROGNOSIS

Mortality rates associated with Oroya fever have been reported to be as high as 40% without treatment but are considerably lower (~10%) with treatment. Complications such as bacterial superinfection and neurologic and cardiac manifestations occur frequently. Generalized massive edema (anasarca) and petechiae are associated with poor outcome. Permanent immunity usually develops.

FURTHER READING

Fournier PE et al: Epidemiologic and clinical characteristics of *Bartonella quintana* and *Bartonella henselae* endocarditis: A study of 48 patients. Medicine (Baltimore) 80:245, 2001.

Gandhi TN et al: *Bartonella* including cat scratch disease, in *Principles and Practice of Infectious Diseases*, 8th ed, GL Mandell et al (eds). Philadelphia, Elsevier Saunders, 2015, pp 2649–63.

Harms A, Dehio C: Intruders below the radar: Molecular pathogenesis of *Bartonella* spp. Clin Microbiol Rev 25:42, 2012.

Koehler JE et al: Molecular epidemiology of *Bartonella* infections in patients with bacillary angiomatosis-peliosis. N Engl J Med 337:1876, 1997.

Maguina C et al: Bartonellosis (Carrión's disease) in the modern era. Clin Infect Dis 33:772, 2001.

Rolain JM et al: Recommendations for treatment of human infections caused by *Bartonella* species. Antimicrob Agents Chemother 48:1921, 2004.

Chapter 168: Donovanosis

Nigel O'Farrell

INTRODUCTION

Donovanosis is a chronic, progressive bacterial infection that usually involves the genital region. The condition is generally regarded as a sexually transmitted infection of low infectivity. This infection has been known by many other names, the most common being *granuloma inguinale*.

ETIOLOGY

The causative organism has been reclassified as *Klebsiella granulomatis comb nov* on the basis of phylogenetic analysis, although there is ongoing debate about this decision. Some authorities consider the original nomenclature (*Calymmatobacterium granulomatis*) to be more appropriate in light of analysis of 16S rRNA gene sequences.

Donovanosis was first described in Calcutta in 1882, and the causative organism was recognized by Charles Donovan in Madras in 1905. He identified the characteristic Donovan bodies, measuring 1.5 × 0.7 μm, in macrophages and the stratum malpighii. The organism was not reproducibly cultured until the mid-1990s, when its isolation in peripheral-blood monocytes and human epithelial cell lines was reported.

EPIDEMIOLOGY

Donovanosis has an unusual geographic distribution that includes Papua New Guinea, parts of southern Africa, India, the Caribbean, French Guyana, Brazil, and Aboriginal communities in Australia. In Australia, donovanosis has been almost entirely eliminated through a sustained program backed by strong political commitment and resources at the primary health care level. Although few cases are now reported in the United States, donovanosis was once prevalent in this country, with 5000–10,000 cases recorded in 1947. The largest epidemic recorded was in Dutch South Guinea, where 10,000 cases were identified in a population of 15,000 (the marind-anim people) between 1922 and 1952.

Donovanosis is associated with poor hygiene and is more common in lower socioeconomic groups than in those who are better off and in men than in women. Infection in sexual partners of index cases occurs to a limited extent. Donovanosis is a risk factor for HIV infection **(Chap. 197)**.

Globally, the incidence of donovanosis has decreased significantly in recent times. This decline probably reflects a greater focus on effective management of genital ulcers because of their role in facilitating HIV transmission.

CLINICAL FEATURES

A lesion starts as a papule or subcutaneous nodule that later ulcerates after trauma. The incubation period is uncertain, but experimental infections in humans indicate a duration of ~50 days. Four types of lesions have been described: (1) the classic ulcerogranulomatous lesion **(Fig. 168-1)**, a beefy red ulcer that bleeds readily when touched; (2) a hypertrophic or verrucous ulcer with a raised irregular edge; (3) a necrotic, offensive-smelling ulcer causing tissue destruction; and (4) a sclerotic or cicatricial lesion with fibrous and scar tissue.

FIGURE 168-1

Ulcerogranulomatous penile lesion of donovanosis, with some hypertrophic features.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The genitals are affected in 90% of patients and the inguinal region in 10%. The most common sites of infection are the prepuce, coronal sulcus, frenum, and glans in men and the labia minora and fourchette in women. Cervical lesions may mimic cervical carcinoma. In men, lesions are associated with lack of circumcision. Lymphadenitis is uncommon. Extragenital lesions occur in 6% of cases and may involve the lip, gums, cheek, palate, pharynx, larynx, and chest. Hematogenous spread with involvement of liver and bone has been reported. During pregnancy, lesions tend to develop more quickly and respond more slowly to treatment. Polyarthritis and osteomyelitis are rare complications. In newborn infants, donovanosis may present with ear infection. Cases in children have been attributed to sitting on the laps of infected adults. As

the incidence of donovanosis has decreased, the number of unusual case reports has appeared to be increasing.

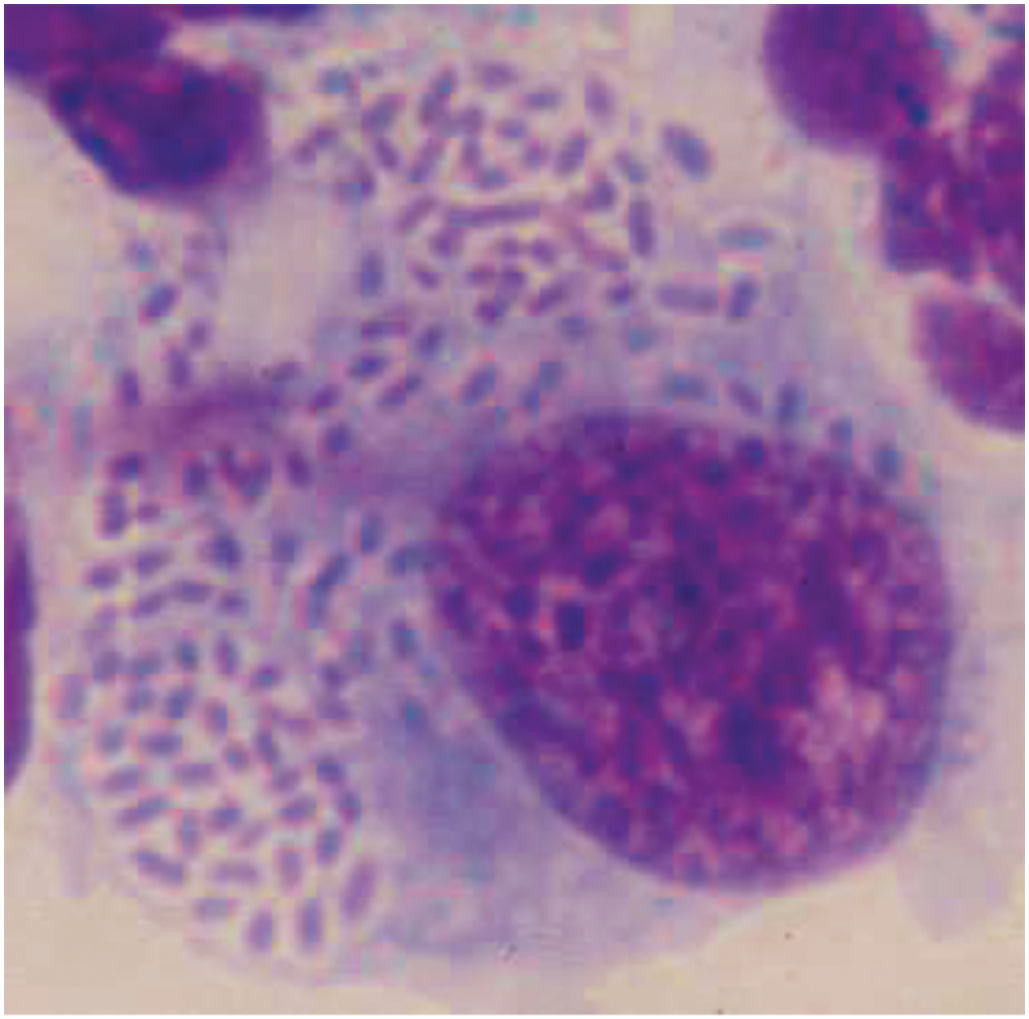
Complications include neoplastic changes, pseudoelephantiasis, and stenosis of the urethra, vagina, or anus.

DIAGNOSIS

A clinical diagnosis of donovanosis made by an experienced practitioner on the basis of the lesion's appearance usually has a high positive predictive value. The diagnosis is confirmed by microscopic identification of Donovan bodies **(Fig. 168-2)** in tissue smears. Preparation of a good-quality smear is important. If donovanosis is suspected on clinical grounds, the smear for Donovan bodies should be taken before swab samples are collected to be tested for other causes of genital ulceration so that enough material can be collected from the ulcer. A swab should be rolled firmly over an ulcer previously cleaned with a dry swab to remove debris. Smears can be examined in a clinical setting by direct microscopy with a rapid Giemsa or Wright's stain. Alternatively, a piece of granulation tissue crushed and spread between two slides can be used. Donovan bodies can be seen in large, mononuclear (Pund) cells as gram-negative intracytoplasmic cysts filled with deeply staining bodies that may have a safety-pin appearance. These cysts eventually rupture and release the infective organisms. Histologic changes include chronic inflammation with infiltration of plasma cells and neutrophils. Epithelial changes include ulceration, microabscesses, and elongation of rete ridges.

FIGURE 168-2

Pund cell stained by rapid Giemsa (RapiDiff) technique. Numerous Donovan bodies are visible.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A diagnostic polymerase chain reaction (PCR) test was based on the observation that two unique base changes in the *phoE* gene eliminate Hae111 restriction sites, enabling differentiation of *K. granulomatis comb nov* from related *Klebsiella* species. PCR analysis with a colorimetric detection system can now be used in routine diagnostic laboratories. A genital ulcer multiplex PCR that includes *K. granulomatis* has been developed. Serologic tests are only poorly specific and are not currently used.

The differential diagnosis of donovanosis includes primary syphilitic chancres, secondary syphilis (condylomata lata), chancroid, lymphogranuloma venereum, genital herpes, neoplasm, and amebiasis. Mixed infections are common. Histologic appearances should be distinguished from those of rhinoscleroma, leishmaniasis, and histoplasmosis.

TREATMENT

TREATMENT

Donovanosis

Many patients with donovanosis present quite late with extensive ulceration. They may be embarrassed and have low self-esteem related to their disease. Reassurance that they have a treatable condition is important, as are the administration of antibiotics and the monitoring of patients for an adequate interval (see below). Epidemiologic treatment of sexual partners and advice about how to improve genital hygiene are recommended.

The recommended drug regimens for donovanosis are shown in **Table 168-1**. Gentamicin can be added if the response is slow. Ceftriaxone, chloramphenicol, and norfloxacin are also effective. Patients treated for 14 days should be monitored until lesions have healed completely. Those treated with azithromycin probably do not need such rigorous follow-up.

Surgery may be indicated for very advanced lesions.

TABLE 168-1

Effective Antibiotics for the Treatment of Donovanosis

Antibiotic	Oral Dose
Azithromycin	1 g on day 1, then 500 mg daily for 7 days or 1 g weekly for 4 weeks
Trimethoprim-sulfamethoxazole	960 mg bid for 14 days
Doxycycline	100 mg bid for 14 days
Erythromycin	500 mg qid for 14 days (in pregnant women)
Tetracycline	500 mg qid for 14 days

CONTROL AND PREVENTION

Donovanosis is probably the cause of genital ulceration that is most readily recognizable clinically. Donovanosis is now limited to a few specific locations, and its global eradication is a distinct possibility.

FURTHER READING

O'Farrell N: Donovanosis, in *Sexually Transmitted Diseases*, 4th ed. KK Holmes et al (eds). McGraw-Hill, 2008, pp 701–708.

Rajam RV, Rangiah PN: Donovanosis (granuloma inguinale, granuloma venereum). Monogr Ser World Health Organ 24:1, 1954.

Sehgal VN, Prasad AL: Donovanosis. Current concepts. Int J Dermatol 5:8, 1986.

Chapter 169: Nocardiosis

Gregory A. Filice

INTRODUCTION

Nocardiosis results from infection with bacteria in the genus *Nocardia*, saprophytic aerobic actinomycetes that commonly reside in soil worldwide and contribute to the decay of organic matter. Nocardiae are relatively inactive in standard biochemical tests, and speciation with traditional biochemical methods is difficult. Since the year 2000, molecular phylogenetic techniques, primarily based on 16S rRNA gene sequences, have identified more than 100 *Nocardia* species, many of which are implicated in human disease.

In the past, the majority of isolates associated with pneumonia and systemic disease were identified biochemically as *Nocardia asteroides*, but the lineage of the type strain was muddled and most human isolates in fact belong to other species. Nine species or species complexes are commonly associated with human disease (Table 169-1). Most systemic disease involves *N. cyriacigeorgica*, *N. farcinica*, *N. pseudobrasiliensis*, and species in the *N. transvalensis* and *N. nova* complexes. *N. brasiliensis* is usually associated with disease limited to the skin. Actinomycetoma—an indolent, slowly progressive disease of skin and underlying tissues with nodular swellings and draining sinuses—is often associated with *N. brasiliensis*, *N. otitidiscaviarum*, *N. transvalensis* complex strains, or other actinomycetes. *N. asteroides* sensu stricto is rarely associated with human disease. However, most clinical laboratories cannot speciate isolates accurately and may identify them simply as *N. asteroides* or *Nocardia* species.

TABLE 169-1

***Nocardia* Species Most Commonly Associated with Human Disease and Their In Vitro Susceptibility Patterns**

Species	Susceptible to[a]	Resistant to[b]
N. abscessus	Amikacin, amoxicillin/clavulanate, ampicillin, ceftriaxone, gentamicin, linezolid, minocycline, tigecycline, tobramycin, TMP-SMX	Ciprofloxacin, clarithromycin (v), imipenem (v), moxifloxacin
N. brevicatena/paucivorans complex (*N. brevicatena*, *N. paucivorans*, *N. carnea*, others)	Amikacin, ampicillin, ceftriaxone, ciprofloxacin, clarithromycin (v), gentamicin, imipenem, linezolid, minocycline (v), moxifloxacin, tigecycline, tobramycin, TMP-SMX	Amoxicillin/clavulanate (v)
N. nova complex (*N. nova*, *N. veterana*, *N. africana*, *N. kruczakiae*, *N. elegans*, others)	Amikacin, ampicillin (v), ceftriaxone (v), clarithromycin, gentamicin (v), imipenem, linezolid, tigecycline (v), TMP-SMX	Amoxicillin/clavulanate, ciprofloxacin, minocycline, moxifloxacin, tobramycin
N. transvalensis complex (*N. blacklockiae*, *N. wallacei*, others)	Ceftriaxone (v), ciprofloxacin (v), linezolid, moxifloxacin, TMP-SMX (v)	Amikacin (v), amoxicillin/clavulanate (v), ampicillin, clarithromycin, gentamicin, imipenem, minocycline (v), tobramycin
N. farcinica	Amikacin, amoxicillin/clavulanate (v), linezolid, moxifloxacin (v), TMP-SMX	Ampicillin, ceftriaxone, ciprofloxacin (v), clarithromycin, gentamicin, imipenem (v), minocycline, tigecycline (v), tobramycin
N. cyriacigeorgica	Amikacin, ceftriaxone, gentamicin, linezolid, tigecycline, tobramycin, TMP-SMX	Amoxicillin/clavulanate, ampicillin, ciprofloxacin, clarithromycin, imipenem (v), minocycline, moxifloxacin
N. brasiliensis	Amikacin, amoxicillin/clavulanate, linezolid, tigecycline, tobramycin, TMP-SMX	Ampicillin, ceftriaxone (v), ciprofloxacin, clarithromycin, imipenem, minocycline (v), moxifloxacin
N. pseudobrasiliensis	Amikacin (v), ciprofloxacin, clarithromycin, linezolid, tobramycin, TMP-SMX (v)	Amoxicillin/clavulanate, ampicillin, ceftriaxone, imipenem, minocycline

Species	Susceptible to[a]	Resistant to[b]
N. otitidiscaviarum complex	Amikacin, gentamicin (v), linezolid, tobramycin (v), TMP-SMX	Amoxicillin/clavulanate, ampicillin, ceftriaxone, ciprofloxacin, clarithromycin, imipenem, minocycline (v), moxifloxacin (v)

[a]From 85 to 100% of isolates are susceptible unless the drug name is followed by (v), in which case 50–84% are susceptible. [b]From 0 to 15% of isolates are susceptible unless the drug name is followed by (v), in which case 16–49% are susceptible.

Abbreviations: TMP-SMX, trimethoprim-sulfamethoxazole; v, variable.

Source: Adapted from multiple sources.

EPIDEMIOLOGY

Pulmonary and/or systemic nocardiosis occurs worldwide. The annual incidence, estimated on three continents (North America, Europe, and Australia), is ~0.375 case per 100,000 persons and may be increasing. There is some geographic variation in species frequencies; for example, *N. asiatica* infections and *N. beijingensis* infections appear to be more commonly involved in cases from eastern Asia. However, exact species prevalences are difficult to determine precisely since nocardial infections are not reportable and most publications consist of case reports or case series.

Actinomycetoma occurs mainly in tropical and subtropical regions. Most cases are reported from Sudan, Mexico, and India. The most important risk factors are lower socioeconomic status and frequent contact with soil or vegetable matter; accordingly, many patients are laborers.

Pulmonary and/or systemic nocardiosis is more common among adults than among children and more common among males than among females. Nearly all cases are sporadic, but outbreaks have been associated with contamination of the hospital environment, cosmetic procedures, and parenteral illicit drug use. Person-to-person spread is not well documented. There is no known seasonality. In regions of the world where tuberculosis is relatively common, nocardiosis is diagnosed in 1–5% of patients in whom pulmonary tuberculosis is suspected, and tuberculosis and nocardiosis can occur in the same patient.

The majority of cases of pulmonary or disseminated disease occur in people with a host defense defect. Most have deficient cell-mediated immunity, especially that associated with lymphoma, transplantation, glucocorticoid therapy, or AIDS. The incidence is ~140-fold greater among patients with AIDS and ~340-fold greater among bone marrow transplant recipients than in general populations. In AIDS, nocardiosis usually affects persons with <250 CD4+ T lymphocytes/µL. Nocardiosis has also been associated with pulmonary alveolar proteinosis, tuberculosis and other mycobacterial diseases, chronic granulomatous disease, interleukin 12 deficiency, and autoantibodies to granulocyte-macrophage colony-stimulating factor (GM-CSF). Any child with nocardiosis and no known cause of immunosuppression should undergo tests to determine the adequacy of the phagocytic respiratory burst. Many cases have been associated with newer immunomodulating drugs—initially with tumor necrosis factor inhibitors and subsequently with a much broader array of these agents. *Nocardia* is frequently isolated from respiratory secretions of patients with cystic fibrosis and may be associated with deterioration of lung function, but this association has not been convincingly established.

PATHOLOGY AND PATHOGENESIS

Pneumonia and disseminated disease are both thought to follow inhalation of fragmented bacterial mycelia. The characteristic histologic feature of nocardiosis is an abscess with extensive neutrophil infiltration and prominent necrosis. Granulation tissue usually surrounds the lesions, but extensive fibrosis or encapsulation is uncommon.

Actinomycetoma is characterized by suppurative inflammation with sinus tract formation. Granules—microcolonies composed of dense masses of bacterial filaments extending radially from a central core—are occasionally observed in histologic preparations. The granules are frequently found in discharges from lesions of actinomycetoma but almost never in discharges from lesions in other forms of nocardiosis.

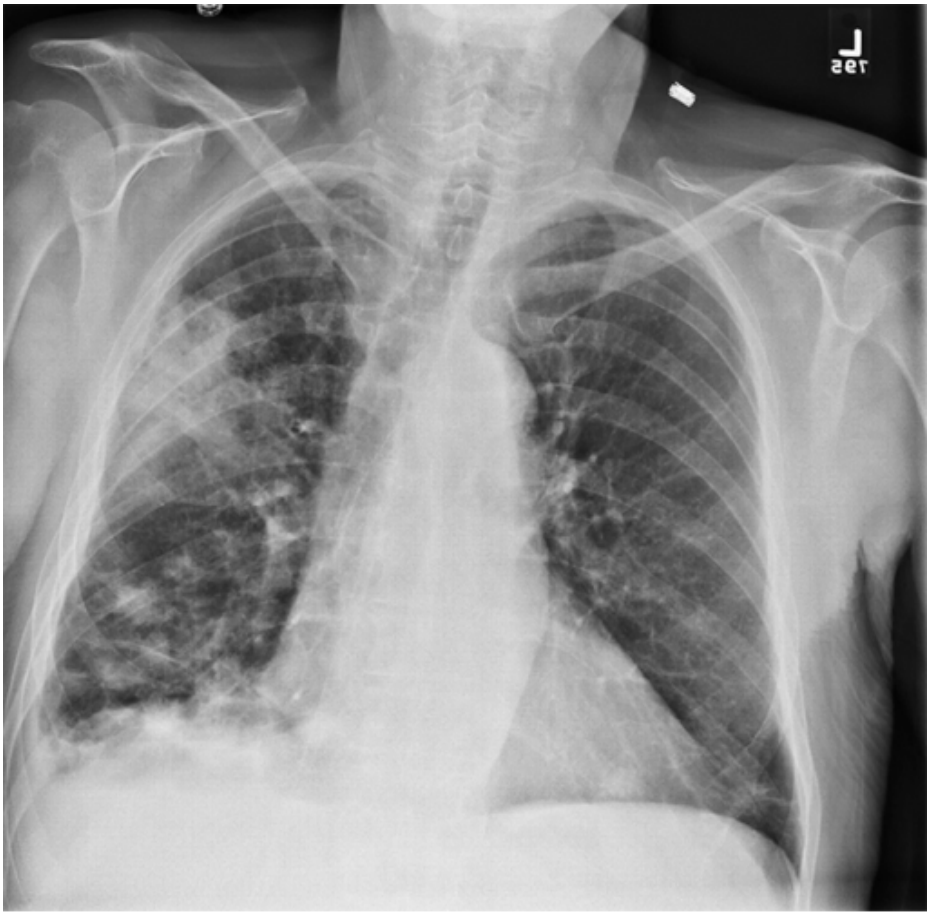
Nocardiae have evolved a number of properties that enable them to survive within phagocytes, including neutralization of oxidants, prevention of phagosome–lysosome fusion, and prevention of phagosome acidification. Neutrophils phagocytose the organisms and limit their growth but do not kill them efficiently. Cell-mediated immunity is important for definitive control and elimination of nocardiae. Nocardiae stimulate the production of GM-CSF in phagocytes in vitro, and nocardial infection has recently been observed in several patients with autoantibodies to GM-CSF. Antibodies to GM-CSF have been found in the majority of patients with alveolar proteinosis and appear to be central to the pathogenesis of this disease. These antibodies may explain the long-standing association of nocardiosis and alveolar proteinosis.

CLINICAL MANIFESTATIONS

Respiratory Tract Disease

Pneumonia, the most common form of nocardial disease in the respiratory tract, is typically subacute; symptoms have usually been present for days or weeks at presentation. The onset is occasionally more acute in immunosuppressed patients. Cough is prominent and produces small amounts of thick, purulent sputum that is not malodorous. Fever, anorexia, weight loss, and malaise are common; dyspnea, pleuritic pain, and hemoptysis are less common. Remissions and exacerbations over several weeks are frequent. Roentgenographic patterns vary, but some are highly suggestive of nocardial pneumonia. Infiltrates vary in size and are typically dense. Single or multiple nodules are common **(Figs. 169-1 and 169-2)**, sometimes suggesting tumors or metastases. Infiltrates and nodules tend to cavitate (Fig. 169-2). Empyema is present in one-quarter of cases.

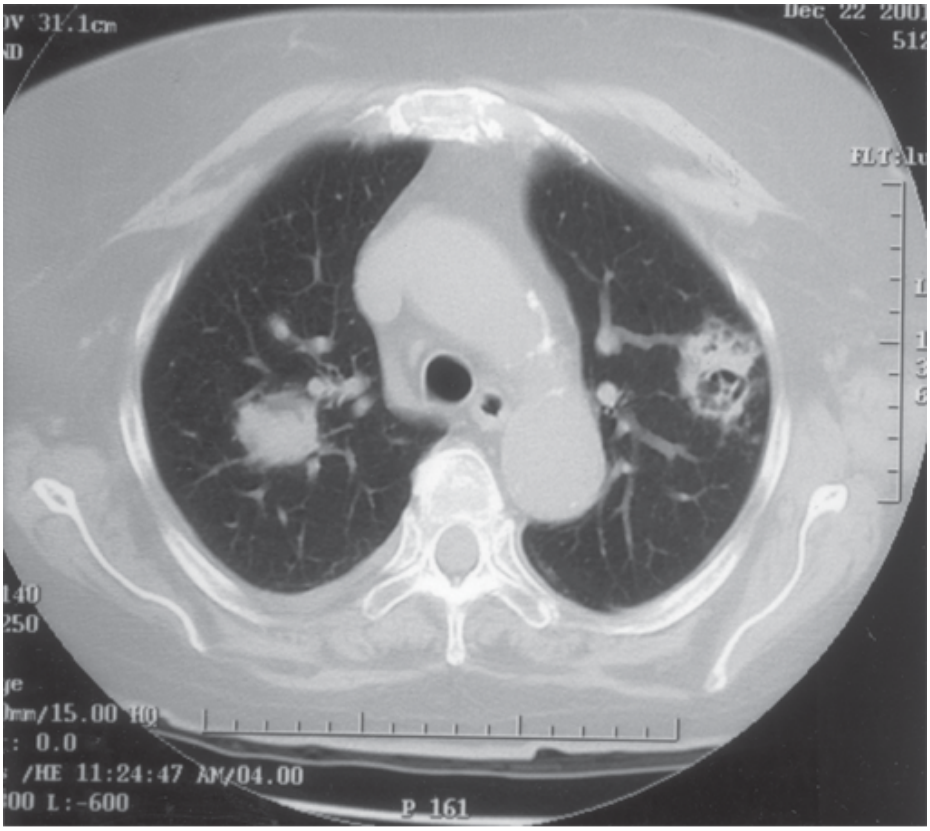
FIGURE 169-1
Nocardial pneumonia. A dense infiltrate with a possible cavity and several nodules are apparent in the right lung.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 169-2

Nocardial pneumonia. A computed tomography scan shows bilateral nodules, with cavitation in the nodule in the left lung.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Nocardiosis may spread directly from the lungs to adjacent tissues. Pericarditis, mediastinitis, and the superior vena cava syndrome have all been reported. Nocardial laryngitis, tracheitis, bronchitis, and sinusitis are much less common than pneumonia. In the major airways, disease often presents as a nodular or granulomatous mass. Nocardiae are sometimes isolated from respiratory secretions of persons without apparent nocardial disease, usually individuals who have underlying lung or airway abnormalities.

Extrapulmonary Disease

In half of all cases of pulmonary nocardiosis, disease appears outside the lungs. In one-fifth of cases of disseminated disease, lung disease is not apparent. The most common site of dissemination is the brain. Other common sites include the skin and supporting structures, kidneys, bones, muscles, and eyes, but almost any organ can be involved. Peritonitis has been reported in patients undergoing peritoneal dialysis. Nocardiae have been recovered from blood in a few cases of pneumonia, disseminated disease, or central venous catheter infection. Nocardial endocarditis occurs rarely and can affect either native or prosthetic valves.

The typical manifestation of extrapulmonary dissemination is a subacute abscess. A minority of abscesses outside the lungs or central nervous system (CNS) form fistulas and discharge small amounts of pus. In CNS infections, brain abscesses are usually supratentorial, are often multiloculated, and may be single or multiple (Fig. 169-3). Cases in the posterior fossa and spinal cord have been reported, but they are less common. Brain abscesses tend to burrow into the ventricles or extend out into the subarachnoid space. The symptoms and signs are somewhat more indolent than those of other types of bacterial brain abscess. Meningitis is uncommon and is usually due to spread from a nearby brain abscess. Nocardiae are not easily recovered from cerebrospinal fluid (CSF).

FIGURE 169-3
Nocardial abscesses in the right occipital lobe.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Disease Following Transcutaneous Inoculation

Disease that follows transcutaneous nocardial inoculation usually takes one of three forms: cellulitis, lymphocutaneous syndrome, or actinomycetoma.

Cellulitis generally begins 1–3 weeks after a recognized breach of the skin, often with soil contamination. Subacute cellulitis, with pain, swelling, erythema, and warmth, develops over days to weeks. The lesions are usually firm and not fluctuant. Disease may progress to involve underlying muscles, tendons, bones, or joints. Dissemination is rare. *N. brasiliensis* and species in the *N. otitidiscaviarum* complex are most common in cellulitis cases.

Lymphocutaneous disease usually begins as a pyodermatous nodule at the site of inoculation, with central ulceration and purulent or honey-colored drainage. Subcutaneous nodules often appear along lymphatics that drain the primary lesion. Most cases of nocardial lymphocutaneous syndrome are associated with *N. brasiliensis*. Similar disease occurs with other pathogens, most notably *Sporothrix schenckii* (Chap. 214) and *Mycobacterium marinum* (Chap. 175).

Actinomycetoma usually begins with a nodular swelling, sometimes at a site of local trauma. Lesions (Fig. 169-4A) typically develop on the feet or hands but may involve the posterior part of the neck, the upper back, the head, and other sites. The nodule eventually breaks down, and a fistula appears, typically followed by others. The fistulas tend to come and go, with new ones forming as old ones disappear. The discharge is serous or purulent, may be bloody, and often contains 0.1- to 2-mm white granules consisting of masses of mycelia (Figs. 169-4C and 169-4D). The lesions spread slowly along fascial planes to involve adjacent areas of skin, subcutaneous tissue, and bone. Over months or years, there may be extensive deformation of the affected part. Lesions involving soft tissues are only mildly painful; those affecting

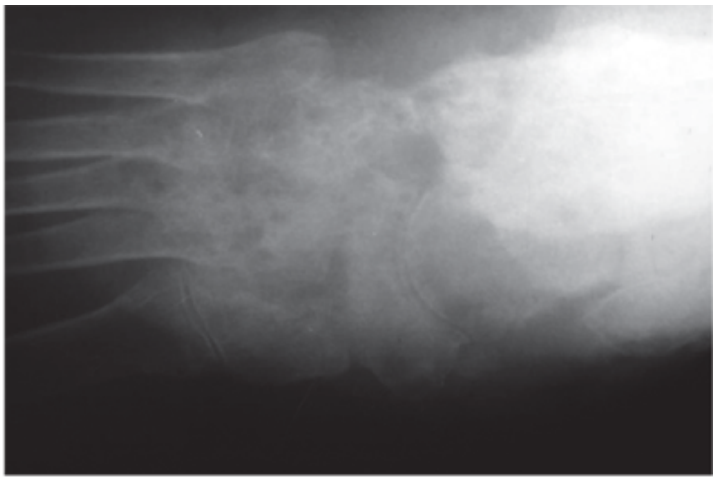
bones or joints are more so (Fig. 169-4B). Systemic symptoms are absent or minimal. Infection rarely disseminates from actinomycetoma, and lesions on the hands and feet usually cause only local disability. Lesions on the head, neck, and trunk can invade locally to involve deep organs, with consequent severe disability or death.

FIGURE 169-4

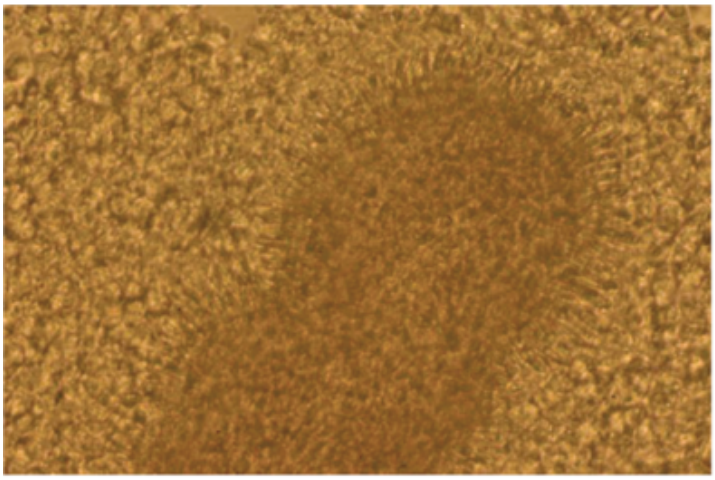
Nocardia brasiliensis **mycetoma.** ***A.*** Draining sinuses and giant white grains with a seropurulent discharge. ***B.*** Radiography of the foot showing marked soft tissue enlargement and bony lytic lesions. ***C.*** Direct microscopy of grains stained with Lugol's iodine (×40). ***D.*** Periodic acid–Schiff stain of skin biopsy (×40). *(Images provided by Roberto Arenas and Mahreen Ameen, St. John's Institute of Dermatology, Guy's & St Thomas' NHS Trust, London, UK. Reprinted from R Arenas, M Ameen: Lancet Infect Dis 10:66, 2010, with permission from Elsevier.)*



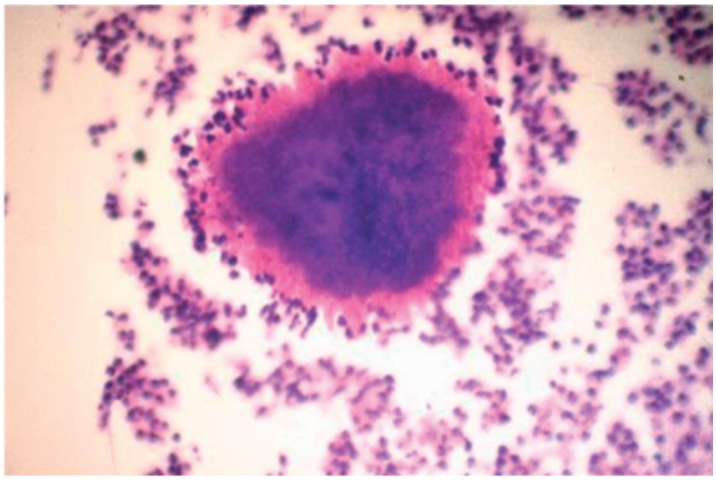
A



B



C



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Eye Infections

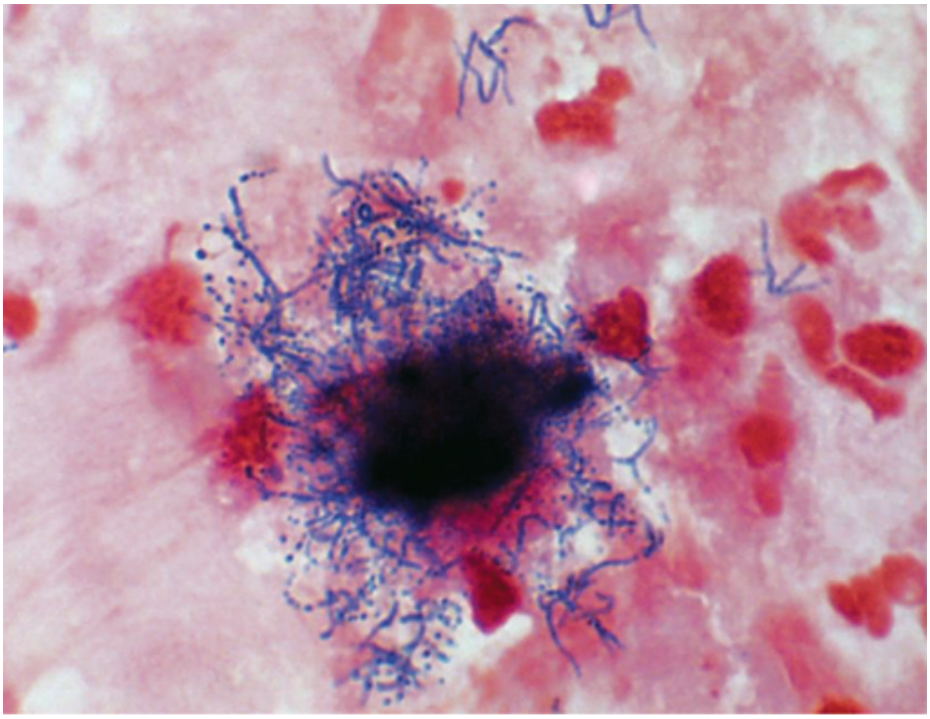
Nocardia species are uncommon causes of subacute keratitis, usually following eye trauma. Nocardial endophthalmitis can develop after eye surgery. In one series, nocardiae accounted for more than half of culture-proved cases of endophthalmitis after cataract surgery. Endophthalmitis can also occur during disseminated disease. Nocardial infection of lachrymal glands has been reported.

DIAGNOSIS

The first step in diagnosis is examination of sputum or pus for crooked, branching, beaded, gram-positive filaments 1 µm wide and up to 50 µm long **(Fig. 169-5)**. Most nocardiae are acid-fast in direct smears if a weak acid is used for decolorization (e.g., in the modified Kinyoun, Ziehl-Neelsen, and Fite-Faraco methods). The organisms often take up silver stains. Recovery from specimens containing a mixed flora can be improved with selective media (colistin–nalidixic acid agar, modified Thayer-Martin agar, or buffered charcoal–yeast extract agar). Nocardiae grow well on most fungal and mycobacterial media, but procedures used for decontamination of specimens for mycobacterial culture can kill nocardiae and should not be used when nocardiae are suspected.

FIGURE 169-5

Gram-stained sputum from a patient with nocardial pneumonia. *(Image provided by Charles Cartwright and Susan Nelson, Hennepin County Medical Center, Minneapolis, MN.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Nocardiae grow relatively slowly; colonies may take up to 2 weeks to appear and may not develop their characteristic appearance—white, yellow, or orange, with aerial mycelia and delicate, dichotomously branched substrate mycelia—for up to 4 weeks. Several blood culture systems support nocardial growth, although nocardiae may not be detected for up to 2 weeks. The growth of nocardiae is so different from that of more common pathogens that the laboratory should be alerted when nocardiosis is suspected in order to maximize the likelihood of isolation.

In nocardial pneumonia, sputum smears are often negative. Unless the diagnosis can be made in smear-negative cases by sampling lesions in more accessible sites, bronchoscopy or lung aspiration is usually necessary. To evaluate the possibility of dissemination in patients with nocardial pneumonia, a careful history should be obtained and a thorough physical examination performed. Suggestive symptoms or signs should be pursued with further diagnostic tests. Some authorities recommend brain imaging in all cases of pulmonary or disseminated disease. When clinically indicated, CSF or urine should be concentrated and then cultured. Actinomycetoma, eumycetoma (cases involving fungi; **Chap. 214**), and botryomycosis (cases involving cocci or bacilli, often *Staphylococcus aureus*) are difficult to distinguish clinically but

are readily distinguished with microbiologic testing or biopsy. Granules should be sought in any discharge. Suspect particles should be washed in saline, examined microscopically, and cultured. Granules in actinomycetoma cases are usually white, pale yellow, pink, or red. Viewed microscopically, they consist of tight masses of fine filaments (0.5–1 µm wide) radiating outward from a central core (Fig. 169-5). Granules from eumycetoma cases are white, yellow, brown, black, or green; under the microscope, they appear as masses of broader filaments (2–5 µm wide) encased in a matrix. Granules of botryomycosis consist of loose masses of cocci or bacilli. Organisms can also be seen in wound discharge or histologic specimens. The most reliable way to differentiate among the various organisms associated with mycetoma is by culture.

Isolation of nocardiae from sputum or blood occasionally represents colonization, transient infection, or contamination. In typical cases of respiratory tract colonization, Gram-stained specimens are negative and cultures are only intermittently positive. A positive sputum culture in an immunosuppressed patient usually reflects disease. When nocardiae are isolated from sputum of an immunocompetent patient without apparent nocardial disease, the patient should be observed carefully without treatment. A patient with a host-defense defect that increases the risk of nocardiosis should usually receive antimicrobial treatment.

Species are definitively determined by molecular techniques. In recent comparisons, the results were similar for species identification by molecular testing and matrix-assisted laser desorption/ionization/ time-of-flight (MALDI-TOF) mass spectrometry. MALDI-TOF is much more practical for clinical laboratories and is becoming common in laboratories in high-resource countries.

Because nocardiosis is uncommon, data on the relation between susceptibility test results for specific drugs and clinical outcomes in patients treated with these drugs are meager. Careful clinical monitoring is essential, and consultation with clinicians who have experience with nocardiosis is often needed. Susceptibility to antimicrobial agents in vitro can be determined with a Clinical Laboratory Standards Institute (CLSI)–approved broth dilution test. Determination of susceptibility by Etest and BACTEC radiometric methods appears to correlate well with that by broth microdilution. Nocardial growth is slower than the growth of most clinically important bacteria, and nocardiae tend to clump in suspension so that susceptibility-test end points are difficult to read; thus experience is necessary for reliable reading of results. If an isolate can be accurately speciated, its susceptibility to antimicrobial drugs can be predicted with a high degree of accuracy.

Speciation by molecular methods or MALDI-TOF is not practical in many resource-poor countries. As a result, therapy for nocardiosis is often initiated without definitive speciation or knowledge of susceptibility results. For mild or moderate cases, therapy with drugs known to be effective against most isolates is usually adequate. For severe cases or cases that do not respond promptly to antimicrobial therapy, isolates should be sent to a laboratory experienced with *Nocardia* for identification and susceptibility testing whenever possible.

TREATMENT

TREATMENT

Nocardiosis

Trimethoprim-sulfamethoxazole (TMZ-SMX) is the drug of choice for most cases (**Tables 169-1** and **169-2**). Reported rates of TMP-SMX susceptibility have varied widely, and controversy has ensued about the reliability of sulfonamides for therapy. However, clinical responses to appropriate sulfonamide treatment around the world are nearly always satisfactory. At the outset, 10–20 mg/kg of TMP and 50–100 mg/kg of SMX are given each day in two divided doses. Later,

daily doses can be decreased to as little as 5 mg/kg and 25 mg/kg, respectively. In persons with sulfonamide allergies, desensitization usually allows continuation of therapy with these effective and inexpensive drugs.

Clinical experience with other oral drugs is limited. Minocycline (100–200 mg twice a day) is often effective; other tetracyclines are usually less effective. Linezolid is the most consistently active antimicrobial agent, but adverse effects become common and limiting in many patients after 2–3 weeks. Amoxicillin (875 mg) combined with clavulanate (125 mg), given twice a day, has been effective but should be avoided in cases involving strains of the *N. nova* complex, in which clavulanate induces β-lactamase production. Among the quinolones, moxifloxacin and gemifloxacin appear to be most active.

Amikacin, the best-established parenteral drug except in cases involving the *N. transvalensis* complex, is given in doses of 5–7.5 mg/kg every 12 h or 15 mg/kg every 24 h. Serum drug levels should be monitored during prolonged therapy in patients with diminished renal function and in the elderly. Ceftriaxone and imipenem are usually effective except as indicated in Table 169-1. Tigecycline appears to be active in vitro against some species, but little clinical experience has been reported.

Patients with severe disease are initially treated with a combination including TMP-SMX, amikacin, and ceftriaxone or imipenem. Clinical improvement is usually noticeable after 1–2 weeks of therapy but may take longer, especially with CNS disease. After definite clinical improvement, therapy can be continued with a single oral drug, usually TMP-SMX. Some experts use two or more drugs for the entire course of therapy, but whether multiple drugs are better than a single agent is not known, and additional drugs increase the risk of toxicity. In patients with nocardiosis who need immunosuppressive therapy for an underlying disease or prevention of transplant rejection, immunosuppressive therapy should be continued.

Use of SMX and TMP in high-risk populations to prevent *Pneumocystis* disease or urinary tract infections appears to reduce but not eliminate the risk of nocardiosis. The incidence of nocardiosis is low enough that prophylaxis solely to prevent this disease is not recommended.

Surgical management of nocardial disease is similar to that of other bacterial diseases. Brain abscesses should be aspirated, drained, or excised if the diagnosis is unclear, if an abscess is large and accessible, or if an abscess fails to respond to chemotherapy. Small or inaccessible brain abscesses should be treated medically; clinical improvement should be noticeable within 1–2 weeks. Brain imaging should be repeated to document the resolution of lesions, although abatement on images often lags behind clinical improvement.

Antimicrobial therapy usually suffices for nocardial actinomycetoma. In deep or extensive cases, drainage or excision of heavily involved tissue may facilitate healing, but structure and function should be preserved whenever possible. Keratitis is treated with a topical sulfonamide or amikacin drops plus a sulfonamide or an alternative drug given by mouth.

Nocardial infections tend to relapse (particularly in patients with chronic granulomatous disease), and long courses of antimicrobial therapy are necessary (Table 169-2). If disease is unusually extensive or if the response to therapy is slow, the recommendations in Table 169-2 should be exceeded.

With appropriate treatment, the mortality rate for pulmonary or disseminated nocardiosis outside the CNS should be <5%. CNS disease carries a higher mortality rate. Patients should be followed carefully for at least 6 months after therapy has ended.

TABLE 169-2

Treatment Duration for Nocardiosis

Disease	Duration
Pulmonary or systemic Intact host defenses Deficient host defenses CNS disease	 6–12 months 12 months[a] 12 months[b]
Cellulitis, lymphocutaneous syndrome	2 months
Osteomyelitis, arthritis, laryngitis, sinusitis	4 months
Actinomycetoma	6–12 months after clinical cure
Keratitis	Topical: until apparent cure Systemic: until 2–4 months after apparent cure

[a]In some patients with AIDS and CD4+ T lymphocyte counts of <200/μL or with chronic granulomatous disease, therapy for pulmonary or systemic disease must be continued indefinitely. [b]If all apparent CNS disease has been excised, the duration of therapy may be reduced to 6 months.

FURTHER READING

Ameen M, Arenas R: Developments in the management of mycetomas. Clin Exp Dermatol 34:1, 2009.

Buckwalter SP et al: Evaluation of matrix-assisted laser desorption ionization–time of flight mass spectrometry for identification of *Mycobacterium* species, *Nocardia* species, and other aerobic actinomycetes. J Clin Microbiol 54:376, 2016.

Fahal A et al: A new model for management of mycetoma in the Sudan. PLoS Negl Trop Dis 8:e3271, 2014. [PubMed: 25356640]

Larruskain J et al: Susceptibility of 186 *Nocardia* sp. isolates to 20 antimicrobial agents. Antimicrob Agents Chemother 55:2995, 2011.

McTaggart LR et al: Antimicrobial susceptibility among clinical *Nocardia* species identified by multilocus sequence analysis. Antimicrob Agents Chemother 59:269, 2015.

Mei-Zahav M et al: The spectrum of *Nocardia* lung disease in cystic fibrosis. Pediatr Infect Dis J 34:909, 2015.

Rosen LB et al: *Nocardia*-induced granulocyte macrophage colony-stimulating factor is neutralized by autoantibodies in disseminated/extrapulmonary nocardiosis. Clin Infect Dis 60:1017, 2015.

Schlaberg R et al: Susceptibility profiles of *Nocardia* isolates based on current taxonomy. Antimicrob Agents Chemother 58:795, 2014.

Wang HK et al: Clinical characteristics, microbiology, and outcomes for patients with lung and disseminated nocardiosis in a tertiary hospital. J Formos Med Assoc 114:742, 2015.

Zijlstra EE et al: Mycetoma: A unique neglected tropical disease. Lancet Infect Dis 16:100, 2016.

Chapter 170: Actinomycosis

Thomas A. Russo

INTRODUCTION

Actinomycosis is uncommon, and most physicians' personal experience with its clinical presentations is limited. Laboratory identification of the etiologic agents from the order Actinomycetales is not routine. Thus actinomycosis remains a diagnostic challenge, even for a skilled clinician. However, this infection is usually curable with medical therapy alone. Therefore, an awareness of the full spectrum of clinical syndromes can expedite diagnosis and treatment and minimize unnecessary surgical interventions, morbidity, and mortality.

Actinomycosis is an indolent, slowly progressive infection caused by anaerobic or microaerophilic bacteria, primarily of the genus *Actinomyces*, that colonize the mouth, colon, and vagina. Mucosal disruption may lead to infection at virtually any site in the body. In vivo growth of actinomycetes usually results in the formation of characteristic clumps called *grains* or *sulfur granules*. The clinical presentations of actinomycosis are myriad. Common in the preantibiotic era, actinomycosis has diminished in incidence, as has its timely recognition. Actinomycosis has been called the most misdiagnosed disease, and it has been said that no disease is so often missed by experienced diagnosticians.

Three "classic" clinical presentations that should prompt consideration of this unique infection are (1) the combination of chronicity, progression across tissue boundaries, and mass-like features (mimicking malignancy, with which it is often confused); (2) the development of a sinus tract, which may spontaneously resolve and recur; and (3) a refractory or relapsing infection after a short course of therapy, since cure of established actinomycosis requires prolonged treatment.

ETIOLOGIC AGENTS

Actinomycosis is most commonly caused by *A. israelii, A. naeslundii, A. odontolyticus, A. viscosus, A. meyeri, A. graevenitzii*, and *A. gerencseriae*. Most if not all actinomycotic infections are polymicrobial. *Aggregatibacter (Actinobacillus) actinomycetemcomitans, Eikenella corrodens*, Enterobacteriaceae, and species of *Fusobacterium, Bacteroides, Capnocytophaga, Staphylococcus*, and *Streptococcus* are commonly isolated with actinomycetes in various combinations, depending on the site of infection. Their contribution to the pathogenesis of actinomycosis is uncertain.

Comparative 16S rRNA gene sequencing has led to the identification of an ever-expanding list of *Actinomyces* species and a reclassification of some species to other genera. At present, 47 species and 2 subspecies have been recognized (*http://www.bacterio.net/actinomyces.html*), with 25 species implicated as causes of human disease. *A. europaeus, A. neuii, A. radingae, A. turicensis, A. cardiffensis, A. urogenitalis, A. hongkongensis, A. georgiae, A. massiliensis, A. timonensis,* and *A. funkei* as well as two former *Actinomyces* species—*Trueperella* (*Arcanobacterium*) *pyogenes* and *Trueperella* (*Arcanobacterium*) *bernardiae*—and *Propionibacterium propionicum* are additional causes of human actinomycosis, albeit not always with a "classic" presentation.

EPIDEMIOLOGY

Actinomycosis has no geographic boundaries and occurs throughout life, with a peak incidence in the middle decades. Males have a threefold higher incidence than females, possibly because of poorer dental hygiene and/or more frequent trauma. Improved dental hygiene and the initiation of antimicrobial treatment before actinomycosis fully develops have probably contributed to a decrease in incidence since the advent of antibiotics. Individuals who do not seek or have access to health care, those who have an intrauterine contraceptive device (IUCD) in place for a prolonged period (see "Pelvic Disease," below), and those who receive bisphosphonate treatment (see "Oral–Cervicofacial Disease," below) are probably at higher risk.

PATHOGENESIS AND PATHOLOGY

The etiologic agents of actinomycosis are members of the normal oral flora and are often cultured from the bronchi, the gastrointestinal tract, and the female genital tract. The critical step in the development of actinomycosis is disruption of the mucosal barrier. Local infection may ensue. Once established, actinomycosis spreads contiguously in a slow, progressive manner, ignoring tissue planes. Although acute inflammation may initially develop at the infection site, the hallmark of actinomycosis is the characteristic chronic, indolent phase manifested by lesions that usually appear as single or multiple indurations. Central necrosis consisting of neutrophils and sulfur granules develops and is virtually diagnostic. The fibrotic walls of the mass are typically described as "wooden." The responsible bacterial and/or host factors have not been identified. Over time, sinus tracts to the skin, adjacent organs, or bone may develop. In rare instances, distant hematogenous seeding may occur; lymphatic spread and associated lymphadenopathy are uncommon. As mentioned above, these unique features of actinomycosis mimic malignancy, with which it is often confused.

Foreign bodies appear to facilitate infection. This association most frequently involves IUCDs. Reports have described an association of actinomycosis with HIV infection; transplantation; common variable immunodeficiency; chronic granulomatous disease; treatment with anti–tumor necrosis factor α agents, glucocorticoids, or bisphosphonates; and radio- or chemotherapy. Ulcerative mucosal infections (e.g., by herpes simplex virus or cytomegalovirus) may facilitate disease development.

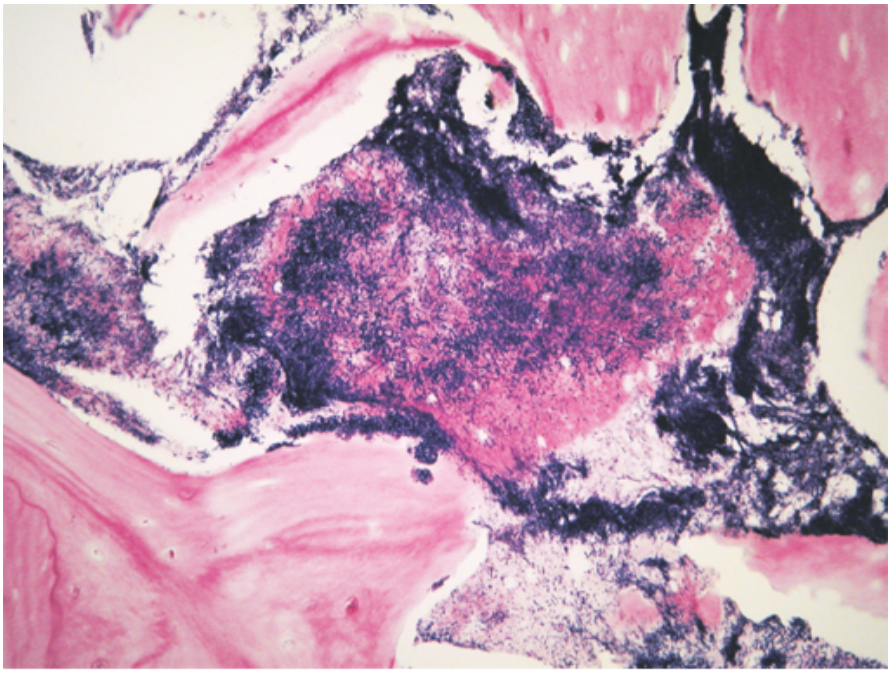
CLINICAL MANIFESTATIONS

Oral–Cervicofacial Disease

Actinomycosis occurs most frequently at an oral, cervical, or facial site, usually as a soft tissue swelling, abscess, mass, or ulcerative lesion that is often mistaken for a neoplasm. Dental diseases or procedures are common precipitating factors. The angle of the jaw is generally involved, but a diagnosis of actinomycosis should be considered with any mass lesion or relapsing infection in the head and neck **(Chap. 31)**. Radiation therapy and especially bisphosphonate treatment have been recognized as contributing to an increasing incidence of actinomycotic infection of the mandible and maxilla **(Fig. 170-1)**. Canaliculitis (commonly due to *P. propionicum*), otitis, sinusitis, and laryngeal disease also can develop. Pain, fever, and leukocytosis are variably reported. Contiguous extension to the cranium, cervical spine, or thorax is a potential sequela.

FIGURE 170-1

Bisphosphonate-associated maxillary osteomyelitis due to *Actinomyces viscosus.* A sulfur granule is seen within the bone. *(Reprinted with permission from NH Naik, TA Russo: Bisphosphonate related osteonecrosis of the jaw: The role of Actinomyces. Clin Infect Dis 49:1729, 2009. © 2009 Oxford University Press.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Thoracic Disease

Thoracic actinomycosis, which may be facilitated by aspirated foreign material, usually follows an indolent progressive course, with involvement of the pulmonary parenchyma and/or the pleural space. Chest pain, fever, and weight loss are common. A cough, when present, is variably productive. The usual radiographic finding is either a mass lesion or pneumonia. On CT, central areas of low attenuation and ring-like rim enhancement may be seen; cavitary disease may develop. More than 50% of cases include pleural thickening, effusion, or empyema **(Fig. 170-2)**. Rarely, pulmonary nodules or endobronchial lesions occur. Lesions suggestive of actinomycosis include those that cross fissures or pleura; extend into the mediastinum, contiguous bone, or chest wall (*empyema necessitatis*); or are associated with a sinus tract. In the absence of these findings, thoracic actinomycosis is usually mistaken for a neoplasm or pneumonia due to more usual causes.

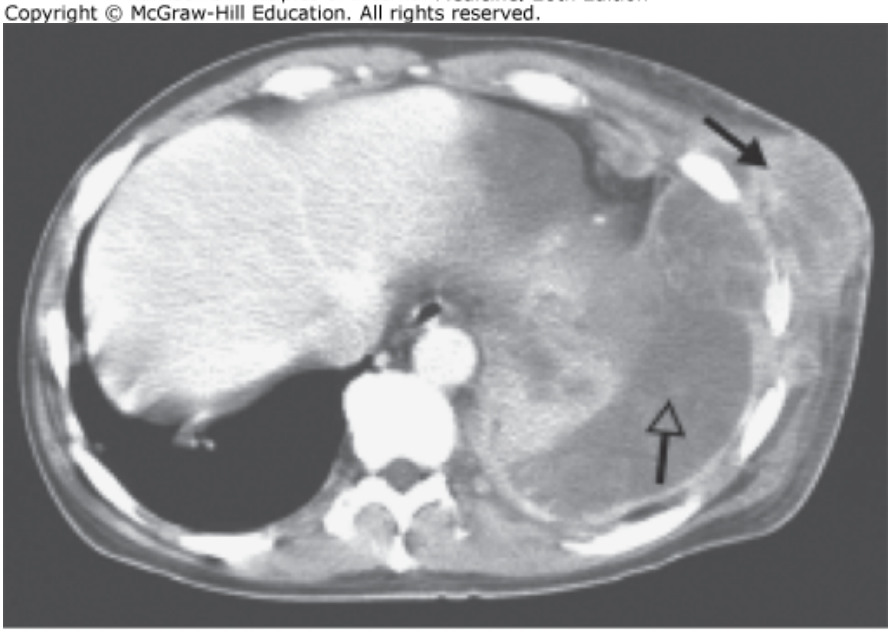
FIGURE 170-2

Thoracic actinomycosis. ***A.*** A chest wall mass from extension of pulmonary infection. ***B.*** Pulmonary infection is complicated by empyema (*open arrow*) and extension to the chest wall (*closed arrow*). *(Courtesy of Dr. C. B. Hsiao, Division of Infectious Diseases, Department of Medicine, State University of New York at Buffalo.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Mediastinal infection is uncommon, usually arising from thoracic extension but rarely from perforation of the esophagus, trauma, or extension of head and neck or abdominal disease. The structures within the mediastinum and the heart can be involved in various combinations; consequently, the possible presentations are diverse. Primary endocarditis (in which *A. neuii* has been increasingly described), esophageal infection, and isolated disease of the breast occur.

Abdominal Disease

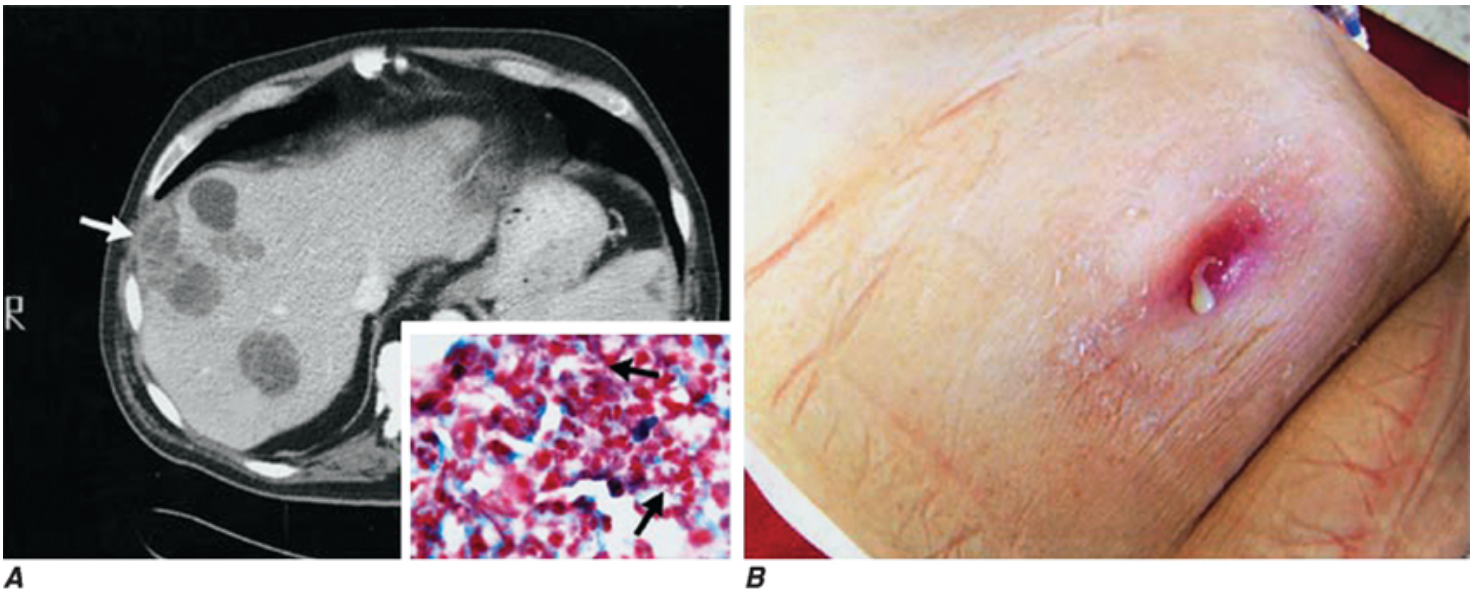
Abdominal actinomycosis poses a great diagnostic challenge. Months or years usually pass from the inciting event (e.g., appendicitis, diverticulitis, peptic ulcer disease, spillage of gall stones or bile during cholecystectomy, foreign-body perforation, bowel surgery, or ascension from IUCD-associated pelvic disease) to clinical recognition. Because of the flow of peritoneal fluid and/or the direct extension of primary disease, virtually any abdominal organ, region, or space can be involved. The disease usually presents as an abscess, a mass, or a mixed lesion that is often fixed to underlying tissue and

mistaken for a tumor. On CT, enhancement is most often heterogeneous and adjacent bowel is thickened. Sinus tracts to the abdominal wall, to the perianal region, or between the bowel and other organs may develop and mimic inflammatory bowel disease (Chap. 319). Recurrent disease or a wound or fistula that fails to heal suggests actinomycosis.

Hepatic infection usually presents as one or more abscesses or masses (Fig. 170-3). Isolated disease presumably develops via hematogenous seeding from cryptic foci. Imaging and percutaneous techniques have resulted in improved diagnosis and treatment.

FIGURE 170-3

Hepatic–splenic actinomycosis. ***A.*** Computed tomogram showing multiple hepatic abscesses and a small splenic lesion due to *Actinomyces israelii.* Arrow indicates extension outside the liver. *Inset:* Gram's stain of abscess fluid demonstrating beaded filamentous gram-positive rods. ***B.*** Subsequent formation of a sinus tract. *(Reprinted with permission from Saad M: Actinomyces hepatic abscess with cutaneous fistula. N Engl J Med 353:e16, 2005. © 2005 Massachusetts Medical Society. All rights reserved.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

All levels of the urogenital tract can be infected. Renal disease usually presents as pyelonephritis and/or renal and perinephric abscess. Bladder involvement, usually due to extension of pelvic disease, may result in ureteral obstruction or fistulas to bowel, skin, or uterus. *Actinomyces* can be detected in urine with appropriate stains and cultures.

Pelvic Disease

Actinomycotic involvement of the pelvis occurs most commonly in association with an IUCD but can also be associated with other foreign bodies, such as surgical mesh. When an IUCD is in place or has been used but removed, pelvic symptoms should prompt consideration of actinomycosis. The risk, although not quantified, appears small. The disease rarely develops when the IUCD has been in place for <1 year, but the risk increases with time. Symptoms are typically indolent; fever, weight loss, abdominal pain, and abnormal vaginal bleeding or discharge are the most common. The earliest stage of disease—often endometritis—commonly progresses to pelvic masses or a tuboovarian abscess (Fig. 170-4). Unfortunately, because the diagnosis is often delayed, a "frozen pelvis" mimicking malignancy or endometriosis can develop by the time of recognition, which may lead to unnecessary surgery. Cancer antigen 125 levels may be elevated, further contributing to misdiagnosis. In contrast to malignancy and tuberculosis, pelvic actinomycosis only uncommonly includes ascites and lymphadenopathy.

FIGURE 170-4

Computed tomogram showing pelvic actinomycosis associated with an intrauterine contraceptive device. The device is encased by endometrial fibrosis (*solid arrow*); also visible are paraendometrial fibrosis (*open triangular arrowhead*) and an area of suppuration (*open arrow*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Actinomyces-like organisms (ALOs), which are identified in Papanicolaou-stained specimens in (on average) 7% of women using an IUCD, have a low positive predictive value for diagnosis. The detection of ALOs in an asymptomatic patient warrants education and close follow-up but not removal of the IUCD unless a suitable contraceptive alternative is agreed on. In the presence of symptoms that cannot be accounted for, it seems prudent to remove the IUCD and—if advanced disease is excluded—to initiate a 14-day course of empirical treatment for possible early endometritis.

Central Nervous System Disease

Actinomycosis of the central nervous system (CNS) is rare. Single or multiple brain abscesses are most common. Individuals with hereditary hemorrhagic telangiectasia are at increased risk for brain abscess with *Actinomyces* as the potential etiologic agent. An abscess usually appears on CT as a ring-enhancing lesion with a thick wall that may be irregular or nodular. Magnetic resonance perfusion and spectroscopy findings have also been described, as have primary meningitis, epidural or subdural space infection, and cavernous sinus syndrome.

Musculoskeletal and Soft Tissue Infection

Actinomycotic infection of bones and joints is usually due to adjacent soft-tissue infection but may be associated with trauma, injections, surgery (e.g., prostheses), osteoradionecrosis and bisphosphonate osteonecrosis (limited to mandibular and maxillary bones), or hematogenous spread. Because of slow disease progression, new bone formation and bone destruction can be seen concomitantly. Infection of soft tissue is uncommon and is usually a result of trauma. An actinomycetoma results from progression over months to years with the development of granulation tissue and tumor-like features. The foot is most commonly involved. Skin, subcutaneous tissue, muscle, and bone are affected in various combinations, and cutaneous sinus tracts frequently develop.

Disseminated Disease

Hematogenous dissemination of disease from any location rarely results in multiple-organ involvement. *A. meyeri* is most commonly involved. The lungs and liver are most often affected, with the presentation of multiple nodules mimicking disseminated malignancy. The clinical presentation may be surprisingly indolent given the extent of disease.

DIAGNOSIS

The diagnosis of actinomycosis is rarely considered. All too often, actinomycosis is first mentioned by the pathologist after extensive surgery. Since medical therapy alone is frequently sufficient for cure, the challenge for the clinician is to consider the possibility of actinomycosis, to diagnose it in the least invasive fashion, and to avoid unnecessary surgery. The clinical and radiographic presentations that suggest actinomycosis are discussed above. Of note, hypermetabolism has been demonstrated by ^{18}F-fluorodeoxyglucose positron emission tomography (FDG-PET) in actinomycotic disease. Aspirations and biopsies (with or without CT or ultrasound guidance) are being used successfully to obtain clinical material for diagnosis, although surgery may be required. The microscopic identification of sulfur granules (an in vivo matrix of bacteria, calcium phosphate, and host material) in pus or tissues, which increases with the examination of additional histopathologic sections and the use of positively charged slides to optimize adhesion, is the most common means of diagnosis. Occasionally, these granules are identified grossly from draining sinus tracts or pus. Although sulfur granules are a defining characteristic of actinomycosis, granules also are found in mycetoma **(Chaps. 169 and 214)** and botryomycosis (a chronic suppurative bacterial infection of soft tissue or, in rare cases, visceral tissue that produces clumps of bacteria resembling granules). These entities can easily be differentiated from actinomycosis with appropriate histopathologic and microbiologic studies. Microbiologic identification of actinomycetes is often precluded by prior antimicrobial therapy or failure to perform appropriate microbiologic cultures. For optimal yield, the avoidance of even a single dose of antibiotics is mandatory. Although some species can grow aerobically, isolation is maximized under anaerobic conditions, usually requiring 5–7 days but potentially up to 2–4 weeks. The use of 16S rRNA gene amplification and sequencing by clinical microbiology laboratories is increasing and is enhancing diagnostic sensitivity and specificity. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) holds similar promise, but databases are still being optimized. Because actinomycetes are components of the normal oral and genital-tract flora, their identification in the absence of sulfur granules in sputum, bronchial washings, and cervicovaginal secretions is of little significance.

TREATMENT

TREATMENT

Actinomycosis

Decisions about treatment are based on the collective clinical experience of the past 70 years. Actinomycosis requires prolonged treatment with high doses of antimicrobial agents; suitable antimicrobial agents and those deemed unreliable are listed in **Table 170-1**. The need for intensive treatment is presumably due to the drugs' poor penetration of the thick-walled masses common in this infection and/or the sulfur granules themselves, which may represent a biofilm. Although therapy must be individualized, the IV administration of 18–24 million units of penicillin daily for 2–6 weeks, followed by oral therapy with penicillin or amoxicillin (total duration, 6–12 months), is a reasonable guideline for serious infections and bulky disease. For penicillin-allergic patients, tetracyclines, ceftriaxone, or carbapenems are reasonable alternatives. Less extensive disease, particularly that involving the oral–cervicofacial region or the isolation of *Actinomyces* in the absence of tissue changes associated with actinomycosis, may be cured with a shorter course. If therapy is extended beyond the resolution of measurable disease, the risk of relapse—a clinical hallmark of this infection—will be minimized; CT and MRI are generally the most sensitive and objective techniques by which to accomplish this goal. A similar approach is reasonable for immunocompromised patients, although refractory disease has been described in HIV-infected individuals. While the role

played by “companion” microbes in actinomycosis is unclear, many isolates are pathogens in their own right, and a regimen covering these organisms during the initial treatment course is reasonable. Isolation of *Actinomyces* from blood cultures in the absence of defined infection may represent contamination or transient bacteremia from a mucosal site of colonization, in which case treatment may not be necessary.

Combined medical–surgical therapy is still advocated in some reports. However, an increasing body of literature now supports an initial attempt at cure with medical therapy alone, even in extensive disease. CT and MRI should be used to monitor the response to therapy. In most cases, either surgery can be avoided or a less extensive procedure can be used. This approach is particularly valuable in sparing critical organs, such as the bladder or the reproductive organs in women of childbearing age. For a well-defined abscess, percutaneous drainage in combination with medical therapy is a reasonable approach. When a critical location is involved (e.g., the epidural space, the CNS), when there is significant hemoptysis, or when suitable medical therapy fails, surgical intervention may be appropriate. In the absence of optimal data, the combination of a prolonged course of antimicrobial therapy and resection—at least of necrotic bone for bisphosphonate-related osteonecrosis of the jaw (BRONJ)—is a reasonable approach.

TABLE 170-1

Appropriate and Inappropriate Antibiotic Therapy for Actinomycosis[a]

Category	Agent
Extensive successful clinical experience[b]	Penicillin: 3–4 million units IV q4h[c,d] Amoxicillin: 500 mg PO q6h Erythromycin: 500–1000 mg IV q6h or 500 mg PO q6h[c] Tetracycline: 500 mg PO q6h Doxycycline: 100 mg IV or PO q12h Minocycline: 100 mg IV or PO q12h Clindamycin: 900 mg IV q8h or 300–450 mg PO q6h[c]
Anecdotal successful clinical experience	Ceftriaxone[d] Ceftizoxime Imipenem-cilastatin Piperacillin-tazobactam
Agents that should be avoided	Metronidazole Aminoglycosides Oxacillin, dicloxacillin Cephalexin Fluoroquinolones
Agents predicted to be efficacious on the basis of in vitro activity	Vancomycin Linezolid Quinupristin-dalfopristin Rifampin Ertapenem[d] Tigecycline[d] Azithromycin[d]

[a]Additional coverage for concomitant "companion" bacteria may be required. [b]Controlled evaluations have not been performed. Dose and duration require individualization depending on the host, site, and extent of infection. As a general rule, a maximal parenteral antimicrobial dose for 2–6 weeks followed by oral therapy, for a total duration of 6–12 months, is required for serious infections and bulky disease, whereas a shorter course may suffice for less extensive disease, particularly in the oral–cervicofacial region. Monitoring the impact of therapy with CT or MRI is advisable when appropriate. [c]Recent in vitro data have demonstrated resistance in up to 33% of isolates. [d]This agent can be considered for at-home parenteral therapy; penicillin requires a continuous infusion pump.

FURTHER READING

Barberis C et al: Antimicrobial susceptibility of clinical isolates of *Actinomyces* and related genera reveals an unusual clindamycin resistance among *Actinomyces urogenitalis* strains. J Glob Antimicrob Resist 8:115, 2017.

Bonnefond S et al: Clinical features of actinomycosis: A retrospective, multicenter study of 28 cases of miscellaneous presentations. Medicine 95:e3923, 2016. [PubMed: 27311002]

Heo SH et al: Imaging of actinomycosis in various organs: A comprehensive review. Radiographics 34:19, 2014.

Jeffery-Smith A et al: Is the presence of *Actinomyces* spp. in blood culture always significant? J Clin Microbiol 54:1137, 2016.

Kononen E, Wade WG: *Actinomyces* and related organisms in human infections. Clin Microbiol Rev 28:419, 2015.

Lo Muzio L et al: The contribution of histopathological examination to the diagnosis of cervico-facial actinomycosis: A retrospective analysis of 68 cases. Eur J Clin Microbiol Infect Dis 33:1915, 2014.

Chapter 171: Whipple's Disease

Thomas A. Russo

FIGURE 171-1

INTRODUCTION

Whipple's disease, described by George Whipple in 1907, is a chronic infection caused by *Tropheryma whipplei*. Most commonly, years pass from the onset of symptoms to the recognition of the disease because of its rarity, its various manifestations mimicking other conditions, and the need to perform nonroutine diagnostic tests. The long-held belief that Whipple's disease is an infection was supported by observations on its responsiveness to antimicrobial therapy in the 1950s and the identification of bacilli via electron microscopy in small-bowel biopsy specimens in the 1960s. This hypothesis was finally confirmed by amplification and sequencing of a partial 16S rRNA polymerase chain reaction (PCR)–generated amplicon from duodenal tissue in 1991. The subsequent successful cultivation of *T. whipplei* enabled whole-genome sequencing and the development of additional diagnostic tests. The development of PCR-based diagnostics has broadened our understanding of both the epidemiology of and the clinical syndromes attributable to *T. whipplei*. Exposure to *T. whipplei*, which appears to be much more common than has been appreciated, can be followed by asymptomatic carriage, acute disease, or chronic infection. Chronic infection—Whipple's disease—is a rare development after exposure. "Classic" Whipple's disease is manifested by some combination of arthralgias/arthritis, weight loss, chronic diarrhea, abdominal pain, and fever. Variable involvement at other sites also occurs; neurologic and cardiac disease are most common. Acute infection and chronic organ disease in the absence of intestinal involvement (see "Isolated Infection," below) are described with increasing frequency. Since untreated Whipple's disease is often fatal and delayed diagnosis may lead to irreparable organ damage (e.g., in the central nervous system [CNS]), knowledge of the clinical scenarios in which Whipple's should be considered and of an appropriate diagnostic strategy is mandatory.

ETIOLOGIC AGENT



T. whipplei is a weakly staining gram-positive bacillus. Genomic sequence data have revealed that the organism has a small (<1-megabase) chromosome, with many biosynthetic pathways absent or incomplete. This finding is consistent with a host-dependent intracellular pathogen or a pathogen that requires a nutritionally rich extracellular environment. It is one of the slowest growing human pathogens, with a

doubling time of 18 days. A genotyping scheme based on a variable region has disclosed >100 genotypes to date. All genotypes appear to be capable of causing similar clinical syndromes.

EPIDEMIOLOGY

Whipple's disease is rare but has been increasingly recognized since the advent of PCR-based diagnostic tools. It occurs in all parts of the globe, with a prevalence estimated at 1–3 cases per 1 million population. Seroprevalence studies indicate that ~50% of Western Europeans and ~75% of Africans from rural Senegal have been exposed to *T. whipplei*. A predilection for chronic disease has been observed in middle-aged Caucasian men, who develop disease five to eight times more frequently than middle-aged women. Humans are the only known host. To date, no clear animal or environmental reservoir has been demonstrated. However, the organism has been identified by PCR in sewage water and human feces. Workers with direct exposure to sewage are more likely to be asymptomatically colonized than controls, a pattern suggesting fecal–oral spread. Fecal PCR detection rates of 38% among family members of carriers or patients with infection support oral–oral or fecal–oral spread, although a common environmental exposure cannot be excluded. Further, the development of acute *T. whipplei* pneumonia in children raises the possibility of droplet or airborne transmission.

PATHOGENESIS AND PATHOLOGY

Since rates of exposure to *T. whipplei*, as defined by seroprevalence, appear to be much higher than rates of chronic disease development (0.00001%), it has been hypothesized that chronically infected individuals possess a subtle host-defense abnormality. The human leukocyte antigen (HLA) alleles DRB1*13 and DQB1*06 are associated with an increased risk of infection. Chronic infection results in a general state of immunosuppression characterized by an impaired T_H1 response, enhanced production of anti-inflammatory cytokines, increased activity of regulatory T cells, M2 polarization of macrophages with diminished antimicrobial activity and impaired phagosome–lysosome fusion and ensuing apoptosis, and blunted development of *T. whipplei*–specific T cells. Immunosuppressive glucocorticoid treatment or anti–tumor necrosis factor α (anti-TNF-α) therapy appears to accelerate progression of chronic disease and perhaps enables infection in colonized individuals. Asymptomatic HIV-infected individuals have been found to have significantly higher levels of *T. whipplei* sequence in bronchoalveolar lavage fluid (BALF) than do non-HIV-infected individuals, and these levels decrease with antiretroviral therapy. A weak humoral response, perhaps due to bacterial glycosylation in patients with chronic disease, appears to differentiate persons who clear the bacillus from asymptomatic carriers. In the initiation of chronic infection, the relative importance of the host's genetic background versus the modulation of the host response by *T. whipplei* is unknown.

T. whipplei has a tropism for myeloid cells, which it invades and in which it can avoid being killed. Infiltration of infected tissue by large numbers of foamy macrophages containing periodic acid–Schiff (PAS)–staining inclusions (representing ingested bacteria) is a characteristic and most common finding. With disease progression, villus atrophy, lymphangiectasia, crypt hyperplasia, and apoptosis of surface epithelial cells are

observed in the small intestine, with resultant diarrhea due to decreased absorption and increased leak flux of water and solutes. Occasionally, involvement of lymphatic or hepatic tissue may manifest as noncaseating granulomas that can mimic sarcoid.

CLINICAL MANIFESTATIONS

Asymptomatic Colonization/Carriage

Studies using primarily PCR have detected *T. whipplei* sequence in stool, saliva, duodenal tissue, and (rarely) blood in the absence of symptoms. Although prevalence rates are still being defined, in Western European countries, detection in saliva (0.2%) is less common than that in stool (1–11%) and appears to occur only with concomitant fecal carriage. The prevalence of fecal carriage is elevated among individuals with exposure to waste water or sewage (12–26%) and among children living in tropical Africa and Asia (20–48%). A duration of carriage of 7 years for the same strain has been described in a sewer worker. Evolution of the carrier state into chronic disease is uncommon. Bacterial loads are lighter in asymptomatic carriage than in active disease.

Acute Infection

T. whipplei has been implicated as a cause of acute gastroenteritis in children. It was also detected via PCR in the blood of 4.6% of febrile patients (75% of whom were <15 years of age) from two rural villages in Senegal as opposed to 0.25% of healthy controls. Further, *T. whipplei* has been implicated as a cause of acute pneumonia. These data suggest that primary acquisition may result in symptomatic pulmonary or intestinal infection or a febrile syndrome, which perhaps are more common than is generally appreciated.

Chronic Infection

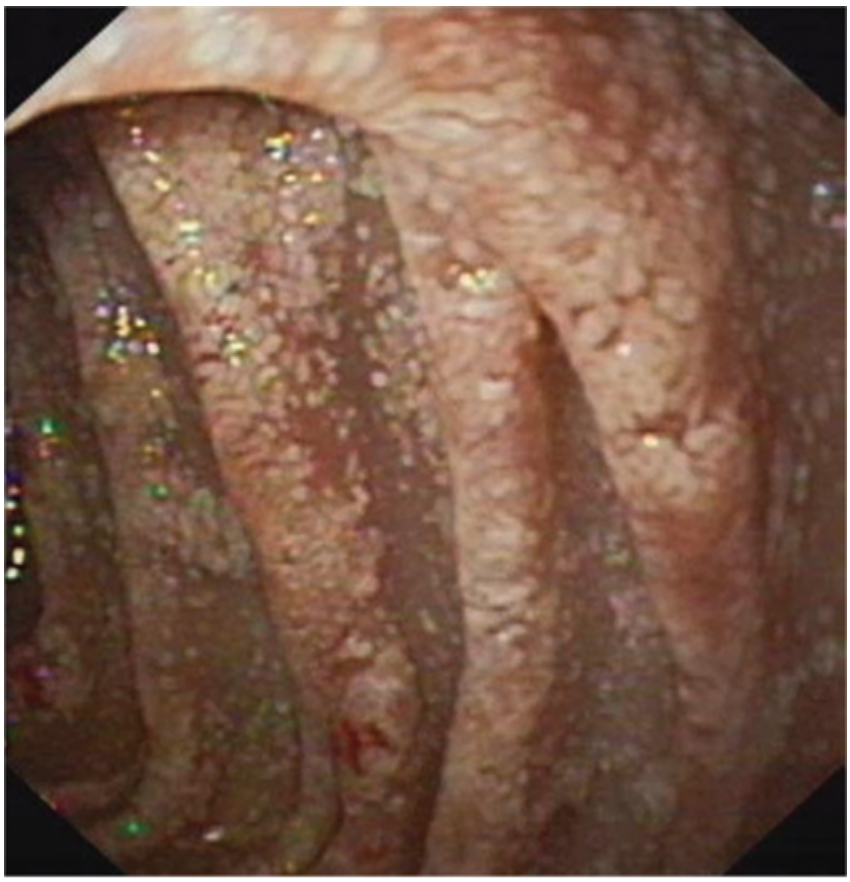
"CLASSIC" WHIPPLE'S DISEASE

So-called classic Whipple's disease was the initial clinical syndrome recognized, with consequent identification of *T. whipplei*. This chronic infection is defined by involvement of the duodenum and/or jejunum that develops over years. In most individuals, the initial phase of disease manifests primarily as intermittent, occasionally chronic, and rarely destructive migratory oligo- or polyarthralgias/seronegative arthritis involving the knees, wrists, and ankles most commonly. Less frequently, spondylitis, sacroiliitis, discitis, and prosthetic hip infection also have been described. Intermittent fever, myalgias, and skin nodules may accompany joint symptoms. Tests for rheumatoid factor and antinuclear antibody are usually negative. This initial stage is often confused with a variety of rheumatologic disorders and, on average, lasts 6–8 years before gastrointestinal symptoms commence. Treatment of presumed inflammatory arthritis with immunosuppressive agents (e.g., glucocorticoids, anti-TNF-α) can accelerate progression of the disease process; thus screening for Whipple's disease prior to initiation of immunosuppressant therapy may be appropriate, depending on the clinical scenario. Alternatively, antimicrobial therapy for another indication

may reduce symptoms, and this situation should also prompt consideration of Whipple's disease. The intestinal symptoms that develop in the majority of cases are characterized by diarrhea with accompanying weight loss and may be associated with fever and abdominal pain. Occult gastrointestinal blood loss, hepatosplenomegaly (10–15%), and ascites (10%) are less common. Anemia and hypereosinophilia may be detected. The most common finding on abdominal CT is mesenteric and/or retroperitoneal lymphadenopathy (usually raising concern about lymphoma). The endoscopic or video-capsule observation of pale, yellow, or shaggy mucosa with erythema or ulceration past the first portion of the duodenum suggests Whipple's disease **(Fig. 171-1)**. When endoscopy with duodenal biopsy is nondiagnostic, a video-capsule study may assist in identifying more distal lesions for subsequent biopsy. Diagnostic misdirection can be caused by co-infection with *Giardia lamblia*, which is occasionally identified. The intestinal phase can also be confused with Crohn's or celiac disease. In addition to rheumatologic and intestinal disease, neurologic (6–63%), cardiac (17–55%), pulmonary (10–50%), lymphatic (10–55%), ocular (5–10%), dermal (5–30%), and less commonly other sites are variably involved in classic Whipple's disease.

FIGURE 171-1

Endoscopic view of the jejunal mucosa demonstrating a thickened, granular mucosa and "white spots" due to dilated lacteals. *(Reprinted with permission from J Bureš et al: Whipple's disease: Our own experience and review of the literature. Gastroenterol Res Pract, 2013. http://dx.doi.org.ezproxy1.library.usyd.edu.au/10.1155/2013/478349.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Neurologic Disease

CNS disease, defined by PCR-based detection of *T. whipplei* in cerebrospinal fluid (CSF), develops in ~50% of patients, many of whom are asymptomatic. A variety of neurologic manifestations have been reported and portend a poor prognosis. The most common are cognitive changes progressing to dementia, personality and mood alterations, hypothalamic involvement (e.g., polyuria/polydipsia, sleep-cycle disorders), and supranuclear ophthalmoplegia. In addition, neuro-ophthalmologic manifestations of Whipple's disease include supranuclear gaze palsy (usually vertical), oculomasticatory and oculofacial myorhythmia (highly suggestive of Whipple's), nystagmus, and retrobulbar neuritis. Focal neurologic presentations (dependent on lesion location), seizures, ataxia, meningitis, encephalitis, hydrocephalus, myelopathy, myoclonus, choreiform movements, and distal polyneuropathy also have been described. Neurologic sequelae occur with CNS disease, and the mortality risk is significant.

MRI results may be normal. Identified lesions (solitary or multifocal) are usually T2 and fluid-attenuated inversion recovery (FLAIR) hyperintense and may enhance with gadolinium. All sites can be involved, and the nature of lesions is variable (e.g., nodular, infiltrative, tumor-like). Although imaging findings are myriad and are not diagnostic, the median temporal lobe, midbrain, hypothalamus, and thalamus are commonly affected. ^{18}F-Fluorodeoxyglucose positron emission tomography (FDG-PET) may reveal increased uptake. CSF analysis may be normal; when abnormal, leukocytosis (generally lymphocyte-predominant) and an elevated protein concentration are common. A low CSF glucose level has been reported.

Cardiac Disease

Endocarditis, which is increasingly recognized in Whipple's disease, presents as culture-negative infection and/or congestive heart failure; hypotension occurs rarely. Embolic events or various arrhythmias or conduction defects may also be noted. Fever is often absent, and the Duke clinical criteria are rarely met. Vegetations are identified by echocardiography in 50–75% of cases. All valves, alone or in combination, can be affected; most commonly involved are the aortic and mitral valves. Preexisting valvular disease is found in only a minority of cases, although infection of bioprosthetic valves has been described. Mural, myocardial, or pericardial disease also occurs alone or in combination with valvular involvement. Constrictive pericarditis develops infrequently. Diagnosis of cardiac disease is rarely made prior to surgical intervention.

Pulmonary Disease

Some combination of interstitial disease, nodules, parenchymal infiltrate, and pleural effusion is observed. An association with pulmonary hypertension has also been reported. The clinical significance of *T. whipplei* sequence identified in BALF from asymptomatic HIV-infected individuals or in a case of interstitial lung disease is unresolved but suggests caution in diagnosing "isolated" pneumonia on the basis of sequence alone.

Lymphatic Disease

Mesenteric and retroperitoneal lymphadenopathy are common with intestinal disease, and mediastinal adenopathy may be associated with pulmonary infection. Peripheral adenopathy is less common.

Ocular Disease (Non-neuro-Ophthalmologic)

Uveitis is the most common form of ocular disease, usually presenting as a change in vision or "floaters." Anterior (anterior chamber), intermediate (vitreous), and posterior (retina/choroid) uveitis can occur alone or in combination. Treatment with glucocorticoids alone can worsen uveitis and unmask extraocular disease.

Likewise, use of local or systemic glucocorticoids after ocular surgery can precipitate ocular infection, likely as a result of asymptomatic or subclinical disease. Keratitis, crystalline keratopathy, and optic neuritis also have been reported. Patients may be misdiagnosed with sarcoid or Behçet's disease prior to the recognition of Whipple's.

Dermatologic Disease

Skin hyperpigmentation, particularly in light-exposed areas in the absence of adrenal dysfunction, is suggestive of Whipple's disease. A variety of other cutaneous manifestations have been described, including erythematous macular lesions, nonthrombocytopenic purpura, subcutaneous nodules, and hyperkeratosis.

Miscellaneous Sites

Thyroid, renal, testicular, epididymal, gallbladder, skeletal muscle, and bone marrow involvement have all been described. In fact, almost any organ can be involved in classic Whipple's disease, with varying frequency, variable combinations, and myriad signs and symptoms. As a result, Whipple's disease should be considered in the setting of a chronic multisystemic process. Despite its rarity, the combination of rheumatologic and intestinal disease with weight loss, with or without neurologic and cardiac involvement, warrants heightened suspicion.

ISOLATED INFECTION

This entity has been defined as infection in the absence of intestinal symptoms, although an occasional small-bowel biopsy may be PCR-positive in this setting. "Isolated infection" is something of a misnomer since multiple nonintestinal sites of *T. whipplei* infection are not uncommon. Infection at the same nonintestinal sites (single or multiple) that are variably involved in classic Whipple's disease may also present as "isolated infection." Further, intestinal disease can subsequently develop. Endocarditis, neurologic disease, uveitis, rheumatologic manifestations, and pulmonary involvement are most commonly described. Signs and symptoms are similar to those described for *T. whipplei* infection of these sites in classic Whipple's disease. With enhanced PCR-based diagnostic capabilities, *T. whipplei* infection without concomitant intestinal involvement (of which endocarditis is the best example) will probably be diagnosed increasingly often.

REINFECTION/RELAPSING DISEASE/IMMUNE RECONSTITUTION INFLAMMATORY SYNDROME (IRIS)

It has been suggested that, if an underlying host immune defect places an individual at risk for chronic infection, then that person may be at risk for reinfection due to occupational exposure or contact with family members who are asymptomatically colonized. One case of apparent reinfection that was due to a different genotype supports this contention.

Optimal treatment regimens and durations are still being defined. However, it is clear, especially in the setting of occult or overt CNS disease, that treatment with oral tetracycline or trimethoprim-sulfamethoxazole (TMP-SMX) alone may result in disease relapse. Relapses or perhaps reinfections occurring years to decades after initial therapy have been described.

As in patients treated for HIV or mycobacterial disease, IRIS has been described in up to 17% of patients treated for *T. whipplei* infection. Prior immunosuppressive therapy increases the likelihood of IRIS, in which inflammation recurs after an initial clinical response to treatment and loss of PCR detection of *T. whipplei*. Manifestations include the development of fever, arthritis, skin lesions, subcutaneous nodules, pleuritis, uveitis, and orbital and periorbital inflammation; some cases have been fatal.

DIAGNOSIS

Considering *T. whipplei* infection and ensuring that the appropriate tests are performed are the critical steps in making the diagnosis, which otherwise will likely be missed. Serology is of little value since patients with active infection usually mount a poor IgM/IgG response to *T. whipplei* and a positive result most likely reflects prior exposure and clearance. The clinical presentation will in part dictate which clinical specimens are most likely to enable the diagnosis. In the presence (and perhaps the absence) of gastrointestinal symptoms, postbulbar duodenal biopsies should be performed, although a normal macroscopic appearance is common. As a general rule, diagnostic yield is greater for tissue specimens than for body fluids. Biopsy of normal-appearing skin may detect *T. whipplei* in the setting of classic Whipple's disease and serve as a minimally invasive means to establish the diagnosis. It is prudent to collect CSF even in the absence of CNS symptoms; asymptomatic disease is common, the CNS is the most common site for relapse, and thus the information gained by CSF examination could influence the design and duration of the treatment regimen.

The diagnosis of classic Whipple's disease was originally based on histologic findings in intestinal biopsy specimens, and this diagnostic procedure remains important. Infiltration of the lamina propria with macrophages containing PAS-positive inclusions that are resistant to diastase is observed. However, PAS is nonspecific, also yielding positive results with mycobacteria (which can be differentiated with Ziehl–Neelsen stain and culture), *Rhodococcus equi, Bacillus cereus*, *Corynebacterium* species, and *Histoplasma* species. *T. whipplei* can be detected by silver stain, Brown-Brenn (weakly positive), or acridine orange and is not stained by calcofluor. Staining of other tissues or fluids (e.g., ocular aspirations) for PAS-positive inclusions in macrophages can be performed to support the diagnosis. The sensitivity of identification of PAS-positive inclusions in Whipple's disease may be decreased by anti-TNF-α therapy. Electron microscopy can be used to identify the trilaminar cell wall of *T. whipplei*.

The development and implementation of PCR-based diagnostics have significantly increased the sensitivity and specificity of *T. whipplei* identification. PCR can be applied to affected tissues (with greater sensitivity for non-formalin-fixed than for formalin-fixed tissue) and to various body fluids (e.g., CSF; aqueous or vitreous humor; joint, pericardial, or pleural fluid; BALF; blood; feces). In some clinical scenarios, a generic 16S rRNA bacterial assay combined with amplicon sequencing can be used to detect and identify *T. whipplei* sequence. Delineation of the *T. whipplei* genomic sequence has enabled the development and broad availability of more sensitive and specific PCR-based assays. The interpretation of a PCR-based diagnostic approach must take into account limitations such as false-positive results due to sample contamination and false-negative results due to low organism load, poor sample quality, and inadequate DNA extraction. As with all diagnostic tests, consideration of pretest probability is critical for interpretation.

When available, immunohistochemistry has greater specificity and sensitivity than PAS staining and can be performed on archived fixed tissue. *T. whipplei* has been successfully cultured from blood, CSF, synovial fluid, BALF, valve tissue, duodenal tissue, skeletal muscle, and lymph nodes, but culture is not practical since it takes months to obtain a positive result.

Although histologic or cytologic detection of *T. whipplei* is less specific and sensitive than PCR, a positive result is strongly supportive within the appropriate clinical context and is definitive when combined with a second, more specific test (e.g., PCR, immunohistochemistry).

TREATMENT

TREATMENT

Whipple's Disease

Data on treatment are emerging, but questions persist regarding the optimal regimen and duration for chronic infection, which may depend on the sites involved (e.g., CNS and heart valve). Appropriate treatment usually results in a rapid—and at times remarkable—clinical response (e.g., in CNS disease), but eradication requires prolonged treatment. Maintenance of a durable response has been more challenging because of both relapse and host predisposition to reinfection.

Rates of relapse, particularly of CNS disease, were unacceptable with oral tetracycline or TMP-SMX monotherapy. Sequence data now indicate that TMP is not active against *T. whipplei* (given the absence of dihydrofolate reductase in *T. whipplei*) and that resistance to SMX and sulfadiazine can occur. However, a randomized controlled trial in 40 patients, who received either ceftriaxone (2 g IV q24h) or meropenem (1 g IV q8h) for 2 weeks followed by oral TMP-SMX (160/800 mg) twice a day for 1 year, demonstrated outstanding efficacy. The only case in which therapy failed—an asymptomatic CNS infection that was not eradicated by either regimen—was subsequently cured with oral minocycline and chloroquine (250 mg/d after a loading dose). A follow-up trial reported similar efficacy with a regimen of ceftriaxone (2 g IV q24h) for 2 weeks followed by oral TMP-SMX for 3 months. One issue in these trials was that the doses—and perhaps the duration of ceftriaxone and meropenem treatment as well—were not optimal for CNS infection. By contrast, in a small retrospective series, outcome was better in patients treated with oral doxycycline (100 mg twice a day) plus hydroxychloroquine (200 mg three times a day; to raise phagosome pH and increase drug activity in vitro) than in patients initially treated with TMP-SMX. A randomized trial comparing oral doxycycline/hydroxychloroquine with a TMP-SMX-based regimen is ongoing and may resolve these discrepancies. Until more data become available, it seems prudent—at least in asymptomatic/symptomatic CNS disease or cardiac infection—first to administer CNS-optimized doses of IV ceftriaxone (2 g q12h) or meropenem (2 g q8h) for 2–4 weeks and then to treat with oral doxycycline or minocycline plus hydroxychloroquine or chloroquine for at least 1 year, if tolerated. Although data on the use of PCR to guide therapy do not exist, it seems reasonable that continued *T. whipplei* detection by PCR, especially in the CSF, should dictate at least continuation of therapy and perhaps consideration of an alternative regimen. Data on isolated infection and certain site-specific treatment issues are even more limited. Anecdotal reports

describe successful treatment of uveitis with oral TMP-SMX with or without rifampin, whereas treatment with tetracycline alone has resulted in relapse. Although a role for adjunctive intraocular therapy has been reported, the data are unclear on this point. Surgery may be needed in the setting of endocarditis with significant valve dysfunction or myocardial abscess; however, timely recognition can result in cure with medical management alone. Although data on the treatment of foreign body-associated infection are virtually nonexistent, medical treatment for a prosthetic hip infection was apparently successful; however, follow-up was limited.

The occurrence of a Jarisch-Herxheimer reaction within 24 h of treatment initiation has been described, with rapid resolution. The addition of glucocorticoids may be beneficial in the management of IRIS, and thalidomide has been used in steroid-refractory cases.

Although data are completely lacking, lifelong suppressive therapy with doxycycline after completion of the initial treatment regimen has been advocated to prevent the occurrence of late relapse or reinfection. Regardless of the therapeutic approach chosen, an effort to ensure compliance and close follow-up for potential relapse or reinfection, which can occur many years after an apparent cure, will maximize the chances for a good outcome.

FURTHER READING

Balasa M et al: Clinical and neuropathological variability in clinically isolated central nervous system Whipple's disease. Brain Pathol 24:230, 2014.

Compain C et al: Central nervous system involvement in Whipple disease: Clinical study of 18 patients and long-term follow-up. Medicine 92:324, 2013.

Damaraju D et al: Clinical problem-solving: A surprising cause of chronic cough. N Engl J Med 373:561, 2015.

Epple HJ et al: Architectural and functional alterations of the small intestinal mucosa in classical Whipple's disease. Mucosal Immunol 10:1542, 2017.

Garcia-Alvarez L et al: *Tropheryma whipplei* endocarditis in Spain: Case reports of 17 prospective cases. Medicine 95:e4058, 2016. [PubMed: 27368042]

Gunther U et al: Gastrointestinal diagnosis of classical Whipple disease: Clinical, endoscopic, and histopathologic features in 191 patients. Medicine 94:e714, 2015. [PubMed: 25881849]

Lagier JC et al: Treatment of classic Whipple's disease: From in vitro results to clinical outcome. J Antimicrob Chemother 69:219, 2014.

Lagier JC et al: Acute infections caused by *Tropheryma whipplei*. Future Microbiol 12:247, 2017.

Meunier M et al: Rheumatic and musculoskeletal features of Whipple disease: A report of 29 cases. J Rheumatol 40:2061, 2013.

Watanuki S et al: Sutton's Law: Keep going where the money is. J Gen Intern Med 30:1711, 2015.

Chapter 172: Infections Due to Mixed Anaerobic Organisms

Neeraj K. Surana; Dennis L. Kasper

INTRODUCTION

Anaerobes comprise the predominant class of bacteria of the normal human microbiota that reside on mucous membranes and predominate in many infectious processes, particularly those arising from mucosal surfaces. These organisms generally cause disease subsequent to the breakdown of mucosal barriers and the leakage of the microbiota into normally sterile sites. Infections resulting from contamination by the microbiota are usually polymicrobial and involve both aerobic and anaerobic bacteria. However, the difficulties encountered in handling specimens in which anaerobes may be important and the technical challenges entailed in cultivating and identifying these organisms in clinical microbiology laboratories continue to leave the anaerobic etiology of an infectious process unproven in many cases. Therefore, an understanding of the types of infections in which anaerobes can play a role is crucial in selecting appropriate microbiologic tools to identify the organisms in clinical specimens and in choosing the most appropriate treatment, including antibiotics and surgical drainage or debridement of the infected site. This chapter focuses on infections caused by anaerobic bacteria other than *Clostridium* species, which are covered elsewhere **(Chaps. 129 and 149)**.

HISTORICAL PERSPECTIVE

Anaerobic organisms were first identified by Antonie van Leeuwenhoek in 1680—nearly a century before oxygen itself was discovered. Leeuwenhoek set up culture medium (crushed pepper powder and clean rainwater) in two glass tubes—one open to ambient air and the other sealed closed—that he incubated for several days. Although he did not expect to observe anything in the sealed tube, he was surprised to find “animalcules” in both tubes. He noted that these bacteria in the sealed tube were “bigger than the biggest sort” in the tube left open to air. It was not until the mid- to late nineteenth century that Leeuwenhoek’s findings were confirmed by Pasteur and others. However, these principles described by Leeuwenhoek underlie the basic pathogenesis of anaerobic infections: development of an anaerobic environment in a closed space is due to consumption of oxygen by aerobic organisms and results in the outgrowth of anaerobic organisms.

DIFFERENCES BETWEEN ANAEROBIC AND AEROBIC ORGANISMS

Anaerobic bacteria can be categorized as *obligate anaerobes* (killed in the presence of ≥0.5% oxygen), *aerotolerant organisms* (can tolerate the presence of oxygen but cannot use it for growth), and *facultative anaerobes* (can grow in the presence or absence of oxygen). Most clinically relevant anaerobes, such as *Bacteroides fragilis, Prevotella melaninogenica*, and *Fusobacterium nucleatum*, are relatively aerotolerant. These organisms contrast with *obligate aerobes*, which require high concentrations of oxygen for growth, and *microaerophilic organisms*, which are damaged by atmospheric concentrations of oxygen (~21%) but require low concentrations of oxygen (typically 2–10%) for growth. Given that molecular oxygen can reduce to superoxide (O_2^-) and hydrogen peroxide (H_2O_2), which are damaging to cells, the ability to tolerate the presence of oxygen is due, in part, to the expression of superoxide dismutase and catalase. The variation in anaerobic organisms tolerating anywhere from <0.5 to 8% O_2 may reflect the amount of these enzymes that is produced.

Furthermore, aerobic and anaerobic organisms differ in their energy metabolism. Cellular respiration requires establishment of an electrochemical gradient across the membrane, resulting in an electric potential (often related to a proton gradient) across the

membrane. In aerobic respiration, electrons are shuttled through an electron transport chain, with oxygen as the final electron acceptor. Anaerobic organisms can metabolize energy by either anaerobic respiration or fermentation. Given that the final electron acceptor in anaerobic respiration (e.g., sulfate, nitrate, carbon dioxide, or fumarate) is not as highly oxidizing as oxygen, this pathway is less efficient than aerobic respiration and produces less ATP per glucose molecule. In contrast, fermentation does not use an electrochemical gradient. Rather, it releases energy from an organic molecule (e.g., pyruvate and its derivatives) via substrate-level phosphorylation and is therefore a less efficient process than either aerobic or anaerobic respiration; for comparison, fermentation results in ~5% of the energy released by aerobic respiration. For these reasons, facultative anaerobes will preferentially utilize oxygen if it is available; in oxygen-limiting situations, organisms will use anaerobic respiration rather than fermentative processes, if possible.

ANAEROBES OF THE HUMAN MICROBIOTA

Most human mucocutaneous surfaces harbor a rich indigenous normal microbiota composed of aerobic and anaerobic bacteria. These surfaces are dominated by anaerobic bacteria, which often account for 99.0–99.9% of the cultivable microbiota and range in concentration from 10^3/mL in the nose to 10^{12}/mL in gingival scrapings and the colon **(Table 172-1)**. It is interesting that anaerobes inhabit many areas of the body that are exposed to air: skin, nose, mouth, and throat. Anaerobes are thought to reside in the portions of these sites that either are relatively well protected from oxygen (e.g., gingival crevices) or have a local anaerobic environment conferred by neighboring aerobic organisms (e.g., tooth surfaces). The ability to cultivate these organisms is improving, and—with strict attention to anaerobic conditions—more than 80% of the microscopic counts in fecal samples can be cultured. However, culture-independent approaches (e.g., sequencing of the 16S rDNA gene) show that the overwhelmingly diverse low-abundance bacterial species present in the microbiota remain uncultivated. Several projects, including the Human Microbiome Project (funded by the U.S. National Institutes of Health) and MetaHIT (financed by the European Commission), have characterized the normal microbiota of healthy individuals and have demonstrated the presence of >10,000 different bacterial species in the collective human microbiota. The human gut alone harbors >1000 bacterial species, with 100–200 species present in any given individual.

TABLE 172-1

The Anaerobic Human Microbiota: An Overview

Anatomic Site	Total Bacteria[a]	Anaerobic/Aerobic Ratio	Potential Pathogen(s)
Nose	10^3–10^4	2:1	*Peptostreptococcus* spp., *Prevotella* spp.
Oral cavity			
Saliva	10^8–10^9	10:1	*Fusobacterium nucleatum, Prevotella melaninogenica, Prevotella oralis* group, *Bacteroides ureolyticus* group, *Peptostreptococcus* spp.
Tooth surface	10^{10}–10^{11}	1:1	
Gingival crevices	10^{11}–10^{12}	10^3:1	
Gastrointestinal tract			
Stomach	0–10^5	1:1	*Bacteroides* spp. (principally members of the *B. fragilis* group), *Prevotella* spp., *Clostridium* spp., *Peptostreptococcus* spp.
Jejunum/ileum	10^4–10^7	1:1	
Cecum and colon	10^{11}–10^{12}	10^3:1	
Female genital tract	10^7–10^9	10:1	*Peptostreptococcus* spp., *Bacteroides* spp., *Prevotella bivia*
Skin	10^4–10^6	100:1	*Cutibacterium acnes*

[a]Per gram or milliliter.

The major reservoir of anaerobic bacteria is the lower gastrointestinal tract, but these organisms are also present in considerable numbers in the oral cavity, skin, and female genital tract (Table 172-1). In the oral cavity, the ratio of anaerobic to aerobic bacteria ranges from 1:1 on the surface of a tooth to 1000:1 in the gingival crevices. *Prevotella* and *Porphyromonas* species make up much of the indigenous oral anaerobic microbiota. *Fusobacterium* and *Bacteroides* (non–*B. fragilis* group) are present in lower numbers. Anaerobic bacteria are not found in appreciable numbers in the normal stomach and proximal small intestine. In the distal ileum, the microbiota begins to resemble that of the colon, where the ratio of anaerobes to aerobic species is high (~1000:1). The predominant anaerobes in the human intestine belong to the phyla Bacteroidetes and Firmicutes and include a number of *Prevotella* and *Bacteroides* species (e.g., members of the *B. fragilis* group, such as *B. fragilis, B. thetaiotaomicron, B. ovatus, B. vulgatus, B. uniformis*, and *Parabacteroides distasonis*) as well as various *Clostridium, Peptostreptococcus, Blautia*, and *Fusobacterium* species. In the female genital tract, there are ~10^9 organisms/mL of secretions, with an anaerobe-to-aerobe ratio of ~10:1. The predominant anaerobes in the female genital tract are *Prevotella, Bacteroides, Fusobacterium, Clostridium*, and the anaerobic *Lactobacillus* species. The skin microbiota contains anaerobes as well; *Cutibacterium acnes* (which was previously

Propionibacterium acnes and will be considered as one of the *Propionibacterium* species for the remainder of this chapter) is the predominant species, and other species of propionibacteria and peptostreptococci are present in lower numbers.

ANAEROBES AND HUMAN HEALTH

Commensal anaerobes have been implicated as crucial mediators of physiologic, metabolic, and immunologic functions in the mammalian host. The intestinal microbiota is essential for fermenting dietary carbohydrates into forms that are more usable by the host, among which polysaccharides are the most abundant biological source of energy. Of the organisms found within the intestines, *Bacteroides* species express the widest array of polysaccharide-degrading enzymes, providing important nutrients for both the host and other commensal organisms. For example, *B. thetaiotaomicron* expresses 172 glycosyl hydrolases. The anaerobic intestinal microbiota is also responsible for the production of secreted products that promote human health (e.g., vitamin K and bile acids useful in fat absorption and cholesterol regulation).

One of the most important roles that anaerobes serve as components of the normal colonic microbiota is the promotion of resistance to colonization. The presence of commensal bacteria effectively interferes with colonization by potentially pathogenic bacterial species through the depletion of oxygen and nutrients, the production of enzymes and toxic end products, and the modulation of the host's intestinal innate immune response. For example, the normal intestinal microbiota plays an important role in protection against enteric infections, including those due to *Salmonella enterica* serotype Typhimurium and *Clostridium difficile*.

The anaerobic intestinal microbiota also has immunomodulatory properties that help regulate the immune system. The first example of this role was demonstrated with *B. fragilis*, which can balance the effector functions of T cells in the peripheral immune system and induce colonic regulatory T cells via expression of polysaccharide A (PSA). Moreover, *B. fragilis* expresses a glycosphingolipid that regulates the number of colonic invariant natural killer T cells. There are now numerous examples of commensal anaerobes that can modulate different aspects of the intestinal and extraintestinal immune system—everything from specific effector T cells to dendritic cells to antimicrobial peptides.

Clearly, the gut microbiota confers many benefits, and its dysregulation may play a role in the pathogenesis of diseases characterized by inflammation and aberrant immune responses, such as inflammatory bowel disease, rheumatoid arthritis, multiple sclerosis, asthma, and type 1 diabetes. Furthermore, the gut microbiota has been associated with obesity and metabolic syndrome. A more complete discussion of the intersection between the microbiota and human health is covered elsewhere (Chap. 459).

ETIOLOGY

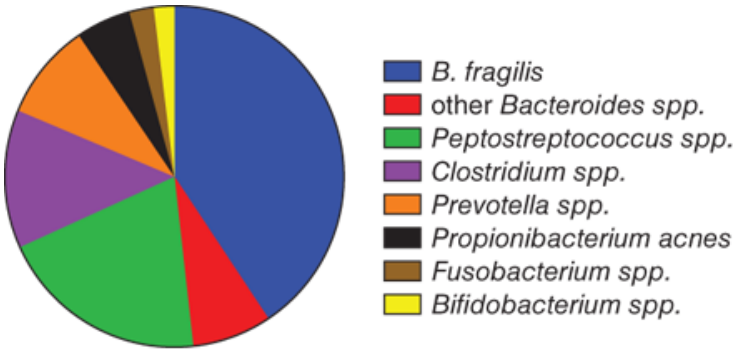
There are >10,000 species of bacteria—the overwhelming majority of which are anaerobes—in the human microbiota, with each individual colonized by hundreds of species. Anaerobic infections occur when the harmonious relationship between the host and the host's microbiota is disrupted. Any site in the body is susceptible to infection with these indigenous organisms if they are introduced into otherwise sterile tissue, either through disruption of mucosal surfaces (e.g., intestinal perforation, ischemia, surgery) or via direct inoculation of organisms into tissue (e.g., bite wounds, trauma). Because the sites that are colonized by anaerobes contain many species of bacteria, the resulting infections are often polymicrobial, involving multiple species of anaerobes in combination with synergistically acting facultative and/or microaerophilic organisms.

Despite the complex array of bacteria in the normal microbiota, relatively few genera are isolated commonly from human infections (Fig. 172-1). While the specific organisms identified vary with the site and source of infection, the etiologic agents typically reflect the neighboring microbiota. For example, organisms normally found in the oro- and nasopharyngeal microbiota (e.g., *P. melaninogenica*, *Fusobacterium necrophorum*, *F. nucleatum*, *Peptostreptococcus* species, *Porphyromonas gingivalis*, *Porphyromonas asaccharolytica*, and *Actinomyces* species) can cause disease in contiguous areas, including odontogenic infections, peripharyngeal space infections, chronic sinusitis, and pleuropulmonary infections. In female genital tract infections, organisms normally colonizing the vagina (e.g., *Prevotella bivia* and *Prevotella disiens*) are the most common isolates. *Escherichia coli* and *B. fragilis*, both of which are components of the intestinal microbiota, are the most commonly identified isolates from

intraabdominal abscesses. Indeed, the *B. fragilis* group, which encompasses 28 species and includes *B. thetaiotaomicron, B. vulgatus, B. uniformis*, and *B. ovatus*, contains the anaerobic organisms most frequently isolated from clinical infections.

FIGURE 172-1

Distribution of all anaerobic organisms identified at a single hospital over a 7-year period. *(Data from Y Park et al: Clinical features and prognostic factors of anaerobic infections: A 7-year retrospective study. Korean J Intern Med 24:13, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

It is useful to think about anaerobic infectious etiologies with regard not only to their anatomic location but also to their microbiologic features. While many anaerobic gram-negative bacilli cause disease (e.g., *Prevotella*, *Bacteroides*, *Fusobacterium*, and *Porphyromonas* species), *Veillonella* species, which are part of the oral and intestinal microbiota, are among the few anaerobic gram-negative cocci that have been implicated in human disease. Similarly, the peptostreptococci (e.g., *P. micros, P. asaccharolyticus*, and *P. anaerobius*) and *Finegoldia magnus* (which was previously *Peptostreptococcus magnus* and will be considered as part of the peptostreptococci for the remainder of this chapter) are the chief anaerobic gram-positive cocci that have pathogenic potential. *Clostridium* species are the primary anaerobic spore-forming gram-positive rods that produce human disease (Chap. 149). Uncommonly, anaerobic gram-positive non-spore-forming bacilli cause infection; *P. acnes*, a component of the skin microbiota and a cause of foreign-body infections, and *Actinomyces* species are relevant examples.

PATHOGENESIS

First and foremost, anaerobic infections require an anaerobic environment with a lowered oxidation-reduction potential. In some circumstances, this environment can occur directly—e.g., in tissue ischemia, trauma, surgery, or a perforated viscus. In many other situations, the infection is polymicrobial, and the facultative organisms maintain a lowered oxidation-reduction potential in the local microenvironment that allows for the propagation of obligate anaerobes. Once the proper anaerobic environment is established, the organisms must still contend with the host's immune defenses. Similar to aerobic organisms, anaerobes express an array of virulence factors that help evade host defenses, they can form abscesses as a protective measure, and they can act synergistically with other bacteria to better persist in the host.

Virulence factors associated with anaerobes typically confer the ability to evade host defenses, adhere to cell surfaces, produce toxins and/or enzymes, or display surface structures such as capsular polysaccharides and lipopolysaccharide that contribute to pathogenic potential. The ability of an organism to adhere to host tissues is often critical to the establishment of infection. Some oral species adhere to the epithelium in the oral cavity. *P. gingivalis*, a common isolate in periodontal disease, has fimbriae that facilitate attachment. In supragingival plaque, many oral anaerobes are able to attach directly to aerobic bacteria (e.g., *Streptococcus* species) that are adherent to the tooth's surface. *F. nucleatum* is a notable example of these secondary colonizers: it expresses receptors to which almost all oral bacteria can bind and serves as an important bridge between the primary colonizers and subsequent layers of bacteria. *B. fragilis* synthesizes pili, fimbriae, and hemagglutinins that aid in attachment to host cell surfaces in the intestine.

Anaerobic bacteria produce a number of exoproteins that can enhance the organisms' virulence. *P. gingivalis* produces a collagenase that enhances tissue destruction. Exotoxins produced by clostridial species, including botulinum toxins, tetanus toxin,

C. difficile toxins A and B, and five toxins produced by *Clostridium perfringens*, are among the most virulent bacterial toxins in mouse lethality assays. Anaerobic gram-negative bacteria, such as *B. fragilis, P. gingivalis*, and *Prevotella intermedia*, possess lipid A molecules (endotoxins) that are 100–1000 times less biologically potent than endotoxins associated with aerobic gram-negative bacteria; these differences may relate to variations in acylation status, length of fatty acids, and number of phosphate groups. This relative biologic inactivity may account for the lower frequency of disseminated intravascular coagulation and purpura in anaerobic gram-negative bacteremia than in facultative and aerobic gram-negative bacillary bacteremia. An exception is the lipopolysaccharide from *Fusobacterium*, which may account for the severity of Lemierre syndrome (see "Complications of Anaerobic Head and Neck Infections," below).

The most extensively studied virulence factor of the nonsporulating anaerobes is the capsular polysaccharide complex of *B. fragilis*. This organism is unique among anaerobes in its potential for virulence during growth at normally sterile sites. Although it constitutes only 0.5–1% of the normal colonic microbiota, *B. fragilis* is the anaerobe most commonly isolated from intraabdominal infections and bacteremia. In an animal model of intraabdominal sepsis, the capsular polysaccharide was identified as the major virulence factor of *B. fragilis*; this polymer plays a specific, central role in the induction of abscesses. A series of detailed biologic and molecular studies of this virulence factor showed that *B. fragilis* produces at least eight distinct capsular polysaccharides, far more than the number reported for any other encapsulated bacterium. *B. fragilis* can exhibit distinct surface polysaccharides either alone or in combination by regulating the expression of these different capsules in an on–off manner through a reversible inversion of DNA segments within the promoters for operons containing the genes required for polysaccharide synthesis. Structural analysis of two of these polysaccharides, PSA and polysaccharide B (PSB), revealed that each polymer consists of repeating units with positively charged free amino groups and negatively charged groups. This structural feature is rare among bacterial polysaccharides, and the ability of PSA—and, to a lesser extent, PSB—to induce abscesses in animals depends on this zwitterionic charge motif. Intraabdominal abscess induction is related to the capacity of PSA to stimulate macrophages to release cytokines and chemokines—in particular, interleukin (IL) 8, IL-17, and tumor necrosis factor α (TNF-α)—from resident peritoneal cells through a Toll-like receptor 2–dependent mechanism. The release of cytokines and chemokines results in the chemotaxis of polymorphonuclear neutrophils (PMNs) into the peritoneum, where they adhere to mesothelial cells induced by TNF-α to upregulate their expression of intercellular adhesion molecule 1 (ICAM-1). PMNs adherent to ICAM-1-expressing cells probably represent the nidus for an abscess. PSA also activates T cells to produce certain cytokines, including IL-17 and interferon γ, that are necessary for abscess formation.

These virulence factors not only promote persistence of the anaerobe that produces them but also aid in the survival of bystander organisms and result in bacterial synergies. Clinically, these synergies are evidenced by the fact that anaerobic infections typically involve three to six different organisms. Examples of this synergistic pathogenesis include creation of a favorable environment for growth (e.g., establishment and maintenance of an anaerobic environment by facultative organisms), inhibition of host defenses (e.g., production of short-chain fatty acids and succinic acid that inhibit the ability of phagocytes to clear facultative organisms), provision of necessary growth factors for other organisms (e.g., oral diphtheroids that produce vitamin K, which is needed by *P. melaninogenica*), and creation of tissue damage that promotes spread of the infection. In these ways, facultative and obligate anaerobes synergistically potentiate abscess formation.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Infections Due to Anaerobic Bacteria

The physician must consider several points when approaching the patient with a possible infection due to anaerobic bacteria.

1. The organisms colonizing mucosal sites are commensals, very few of which typically cause disease. When these organisms do cause disease, it often occurs in proximity to the mucosal site they colonize.

2. For anaerobes to cause tissue infection, they must spread beyond the normal mucosal barriers.

3. Conditions favoring the propagation of anaerobic bacteria, particularly a lowered oxidation-reduction potential, are necessary. These conditions exist at sites of trauma, tissue destruction, compromised vascular supply, and necrosis.

4. Frequently, a complex array of infecting microbes can be found, occasionally with >10 different species isolated from a suppurative site.

5. Anaerobic organisms tend to be found in abscess cavities or in necrotic tissue. The failure of an abscess to yield organisms on routine culture is a clue that the abscess is likely to contain anaerobic bacteria. Often smears of this "sterile pus" are found to be teeming with bacteria when Gram's stain is applied. Although some facultative organisms (e.g., *Staphylococcus aureus*) are also capable of causing abscesses, abscesses in organs or deeper body tissues should call anaerobic infection to mind.

6. Gas is found in many anaerobic infections of deep tissues but is not diagnostic because it can be produced by aerobic bacteria as well.

7. Although a putrid-smelling infection site or discharge is considered diagnostic for anaerobic infection, this manifestation usually develops late in the course and is present in only 30–50% of cases.

8. Some species (the best example being the *B. fragilis* group) require specific therapy. However, many synergistic infections can be cured with antibiotics directed at some but not all of the organisms involved. Antibiotic therapy, combined with debridement and drainage, disrupts the interdependent relationship among the bacteria, and some species that are resistant to the antibiotic do not survive without the co-infecting organisms.

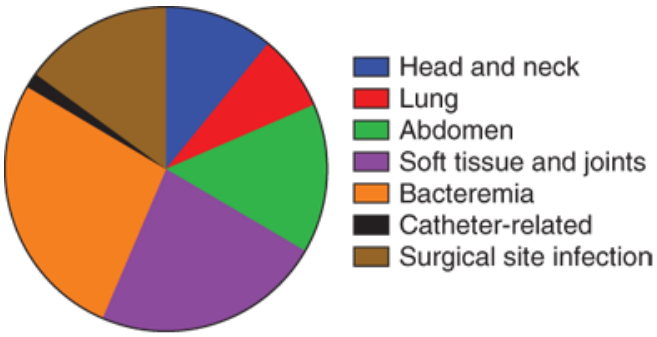
9. Manifestations of severe sepsis and disseminated intravascular coagulation are unusual in patients with purely anaerobic infection.

EPIDEMIOLOGY

Difficulties in the performance of appropriate cultures, contamination of cultures by components of the normal microbiota, and the lack of readily available, reliable culture techniques have made it challenging to obtain accurate data on the incidence or prevalence of anaerobic infections. However, anaerobic infections are encountered frequently, with anaerobes comprising 7–8% and 13–15% of bacteria isolated from inpatients and outpatients, respectively. Bacteremia and soft tissue infections are the most common types of anaerobic infection (Fig. 172-2). Typically, anaerobic bacteria account for <1% of all cases of bacteremia.

FIGURE 172-2

Distribution of types of infection from which anaerobic organisms were cultured at a single hospital over a 7-year period. Head and neck infections included sinusitis, otitis media, and retropharyngeal abscess; abdominal infections included liver abscess, biliary tract infection, bowel obstruction, and intraabdominal abscess; catheter-related infections included those related to peritoneal dialysis catheters and ventriculoperitoneal shunts. *(Data from Y Park et al: Clinical features and prognostic factors of anaerobic infections: A 7-year retrospective study. Korean J Intern Med 24:13, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CLINICAL MANIFESTATIONS

Although anaerobes can cause infection anywhere in the body, certain clinical findings and characteristics are commonly found. These include abscess formation, putrid purulence (due to volatile fatty acid byproducts), septic thrombophlebitis, tissue necrosis, and failure to respond clinically to broad-spectrum antibiotics that lack activity against anaerobes.

Anaerobic Infections of the Mouth, Head, and Neck

Anaerobic bacteria are commonly involved in infections of the mouth, head, and neck **(Chap. 31)**. The predominant isolates are components of the normal microbiota of the upper airways—mainly *Prevotella* species, *P. asaccharolytica*, *Fusobacterium* species, peptostreptococci, and microaerophilic streptococci.

OROFACIAL INFECTIONS

The most common oral infections are odontogenic and include dental caries and periodontal disease (gingivitis and periodontitis). While dental caries usually manifest with pain, sensitivity, and discoloration of the tooth, periodontal disease involves inflammation of the gums and underlying tissue. In its more severe forms, periodontitis can result in difficulty chewing, loose teeth, and occasionally tooth loss. Severe orofacial infections typically develop as a consequence of dental infection, and the infection can spread from the tooth to different anatomic areas that provide the least resistance, resulting in periapical, periodontal, or pericoronal infections. If the dental surface is completely breached, an endodontic infection (pulpitis) can occur. In late stages of pulpitis, the tooth is generally very sensitive to heat, but cold stimuli may provide relief. Left untreated, pulpitis can progress to invade the alveolar bone and develop into a periapical abscess. The abscesses, particularly those involving the second and third molars, can occasionally extend into the submandibular, sublingual, and submental spaces (*Ludwig's angina*). This infection results in marked local swelling of tissues, with pain, trismus, and superior and posterior displacement of the tongue. Submandibular swelling of the neck and obstruction by the tongue can impair swallowing and cause respiratory obstruction. In some cases, tracheotomy is lifesaving.

Microbiologically, dental caries begin with the binding of *Streptococcus mutans* and *Streptococcus sanguis* to the tooth surface, with subsequent further colonization by anaerobes. In contrast, periodontitis is typically associated with *P. gingivalis, Tannerella forsythensis, Aggregatibacter actinomycetemcomitans*, and *Treponema denticola*. *Fusobacterium*, *Actinomyces*, *Peptostreptococcus*, and *Bacteroides* species (other than *B. fragilis*) are the organisms most commonly isolated from periapical abscesses.

ACUTE NECROTIZING ULCERATIVE GINGIVITIS

Gingivitis may become a necrotizing infection (*trench mouth*, *Vincent's stomatitis*) **(Chap. 31)**. The onset of disease is usually sudden and is associated with painful bleeding gums, foul breath, and a bad taste. The gingival mucosa, especially the papillae between the teeth, becomes ulcerated and may be covered by a yellowish-white or gray "pseudomembrane," which is removable with gentle pressure. Patients may become systemically ill, developing fever, malaise, cervical lymphadenopathy, and leukocytosis. The infection can sometimes extend into the pharynx, resulting in an extremely sore throat, foul breath, and tonsillar pillars that are swollen, red, ulcerated, and covered by a pseudomembrane. *Prevotella*, *Treponema*, and *Fusobacterium* species have been implicated.


In some cases, acute necrotizing gingivitis can rapidly progress to noma (*cancrum oris*), a gangrenous infection that destroys the soft and hard tissues related to the oral cavity. Noma occurs most frequently in young children (1–4 years of age) who have immune dysfunction related to malnutrition and endemic infections (particularly measles). This infection occurs worldwide but is most common in sub-Saharan Africa, where the incidence is 1–7 cases per 1000 children. Although the pathogenesis is not fully understood, infection with *F. necrophorum* and *P. intermedia* are thought to be key drivers of this disease. Without treatment, the mortality rate is 70–90%.

PERIPHARYNGEAL SPACE INFECTIONS

These infections arise from the spread of organisms from the upper airways to potential spaces formed by the fascial planes of the head and neck. The etiology is typically polymicrobial and represents the normal microbiota of the mucosa of the originating site.

Peritonsillar abscess (*quinsy*) is the most common peripharyngeal infection and occurs as a complication of acute tonsillitis. Consistent with its association with tonsillitis, adolescents are most commonly affected. Patients present with a sore throat, dysphagia, peritonsillar swelling, muffled voice, and uvular deviation to the contralateral side. The abscess material typically grows group A *Streptococcus* in conjunction with obligate anaerobes (e.g., *Bacteroides*, *Prevotella*, and *Peptostreptococcus* species) (Chap. 31). Retropharyngeal abscesses typically occur in children 2–4 years of age, although they can occur at any age. Although a suppurative infection of the retropharyngeal lymph nodes is the usual precursor to these abscesses in children, foreign-body ingestion and/or local trauma is more commonly the inciting factor in adults. The clinical presentation shares many features with peritonsillar abscesses, but difficulty extending the neck and torticollis are more common with retropharyngeal abscesses. The etiologic agents are the same as in peritonsillar abscesses, with additional aerobic organisms (e.g., *S. aureus*, viridans streptococci) also playing a role.

SINUSITIS AND OTITIS

Anaerobic bacteria have been implicated in chronic sinusitis but play little role in acute sinusitis. Numerous studies related to the microbiology of chronic sinusitis have been conducted; on average, anaerobic bacteria have been found in two-thirds of patients, with many studies demonstrating their presence in >90% of patients. Anaerobic bacteria represent ~40% of all bacteria cultured, with *Peptostreptococcus*, *Prevotella,* and *Porphyromonas* species the most commonly isolated anaerobes. *S. aureus* and Enterobacteriaceae are the aerobes most commonly recovered in chronic sinusitis. Anaerobic bacteria have been isolated in ~60% of cases of chronic suppurative otitis media in children, but they are not involved in acute otitis media.

COMPLICATIONS OF ANAEROBIC HEAD AND NECK INFECTIONS

Direct extension of these infections into contiguous areas can result in additional disease manifestations. Cranial spread of these infections can result in osteomyelitis of the skull or mandible or in intracranial infections, such as brain abscess and subdural empyema. Caudal spread can produce mediastinitis or pleuropulmonary infection. Hematogenous complications can also result from anaerobic infections of the head and neck. Bacteremia, which occasionally is polymicrobial, can lead to endocarditis or other distant infections. Lemierre syndrome (Chap. 31), which is usually due to *F. necrophorum*, is an acute oropharyngeal infection with secondary septic thrombophlebitis of the internal jugular vein and frequent septic emboli, most commonly to the lung. This infection typically begins with pharyngitis, which is followed by local invasion in the lateral pharyngeal space, with resultant internal jugular vein thrombophlebitis.

Central Nervous System (CNS) Infections

CNS infections associated with anaerobic bacteria are brain abscess, epidural abscess, and subdural empyema, in which anaerobes are recovered in up to 30, 20, and 10% of cases, respectively. The frequency with which anaerobes are recovered depends in large part on the underlying reason for the infection. For example, brain abscesses are typically due to hematogenous seeding, contiguous spread, penetrating head trauma, or recent surgical intervention. Anaerobic bacteria are most commonly associated with brain abscesses resulting from contiguous spread (related to otogenic, odontogenic, and sinus infections), and the pathogens recovered are the same as in these antecedent infections. Facultative or microaerophilic streptococci and coliforms are often part of a mixed infecting flora in brain abscesses. The location of the abscess may also provide insight into the pathogens. Abscesses in the frontal lobe (often associated with sinusitis) are due to anaerobes, streptococci, and staphylococci; temporal lobe and cerebellar abscesses are often related to the oral microbiota and middle-ear pathogens.

Obligate anaerobes rarely cause meningitis. Only one obligate anaerobe was identified in a seminal study of 188 bacterial meningitis isolates, and a U.S. national surveillance study of 18,642 such isolates collected between 1977 and 1981 found only five

obligate anaerobes. This low incidence may be due, in part, to the fact that many clinical microbiology laboratories do not routinely culture cerebrospinal fluid (CSF) for anaerobes.

Pleuropulmonary Infections

The lungs are constantly seeded with organisms from the oral microbiota via subclinical microaspiration that normally occurs in all people. Even though the lung is the site of oxygen exchange and is therefore an overwhelmingly aerobic environment, the organisms most abundant in the lower respiratory tract (as assessed by culture-independent methods) include anaerobes such as *Prevotella* and *Veillonella* species, with oral microaerophilic streptococcal species (e.g., the *Streptococcus milleri* group) also present in significant abundances. In patients who have impaired bacterial clearance (due to decreased cough, dysfunctional mucociliary transport, or alcohol intoxication) and/or increased rates of aspiration (due to neurologic disorders, impaired consciousness, or swallowing dysfunction), these anaerobic bacteria can establish an infection and result in aspiration pneumonia, lung abscess, or empyema. These anaerobic infections have an indolent course that may serve as a clinical clue differentiating them from conditions with other etiologies (e.g., chemical pneumonitis, pneumococcal pneumonia) that often present more acutely.

ASPIRATION PNEUMONIA

Bacterial aspiration pneumonia must be distinguished from two other clinical syndromes associated with aspiration that are not of bacterial etiology. One syndrome results from aspiration of food or, rarely, other foreign bodies. Obstruction of major airways typically results in difficulty breathing, atelectasis, and moderate nonspecific inflammation. Therapy consists of removal of the foreign body. The second aspiration syndrome relates to chemical pneumonitis caused by inhalation or aspiration of alveolar irritants. Perhaps the most common cause of chemical pneumonitis is *Mendelson syndrome*, which results from regurgitation and aspiration of acidic gastric juices. Pulmonary inflammation—including the destruction of the alveolar lining, with transudation of fluid into the alveolar space—occurs with remarkable rapidity. This syndrome typically develops within 4–6 h, often following anesthesia when the gag reflex is depressed. The patient becomes tachypneic, tachycardic, and hypoxic, often in the absence of fever. The leukocyte count may rise, and the chest x-ray may evolve from normal to a complete bilateral “whiteout” within 8–24 h. Sputum production is minimal. The pulmonary signs and symptoms often resolve quickly with symptom-based therapy, but this condition can culminate in respiratory failure due, in part, to pulmonary edema. Antibiotic therapy is not indicated unless bacterial superinfection occurs.

In contrast to these syndromes, bacterial aspiration pneumonia develops over a period of several days or weeks rather than hours. The pathogenesis includes some combination of an increased bacterial burden, increased virulence of the organisms aspirated, and potential airway damage related to aspiration of gastric fluid. Patients generally report fever, malaise, and sputum production. In some patients, weight loss and anemia reflect a more chronic process. Usually the history reveals factors predisposing to aspiration, such as significant alcohol consumption or neurologic impairment due to a previous stroke. Severe dental disease is often associated with aspiration pneumonia, but it is not clear whether this association relates to an increased number of oral microbes and/or the presence of organisms with increased virulence. Sputum characteristically is not malodorous unless the process has been ongoing for at least a week. Chest x-rays show consolidation in dependent pulmonary segments: in the basilar segments of the lower lobes if the patient has aspirated while upright and in either the posterior segment of the upper lobe (usually on the right side, given that the right mainstem bronchus has a more vertical orientation) or the superior segment of the lower lobe if the patient has aspirated while supine.

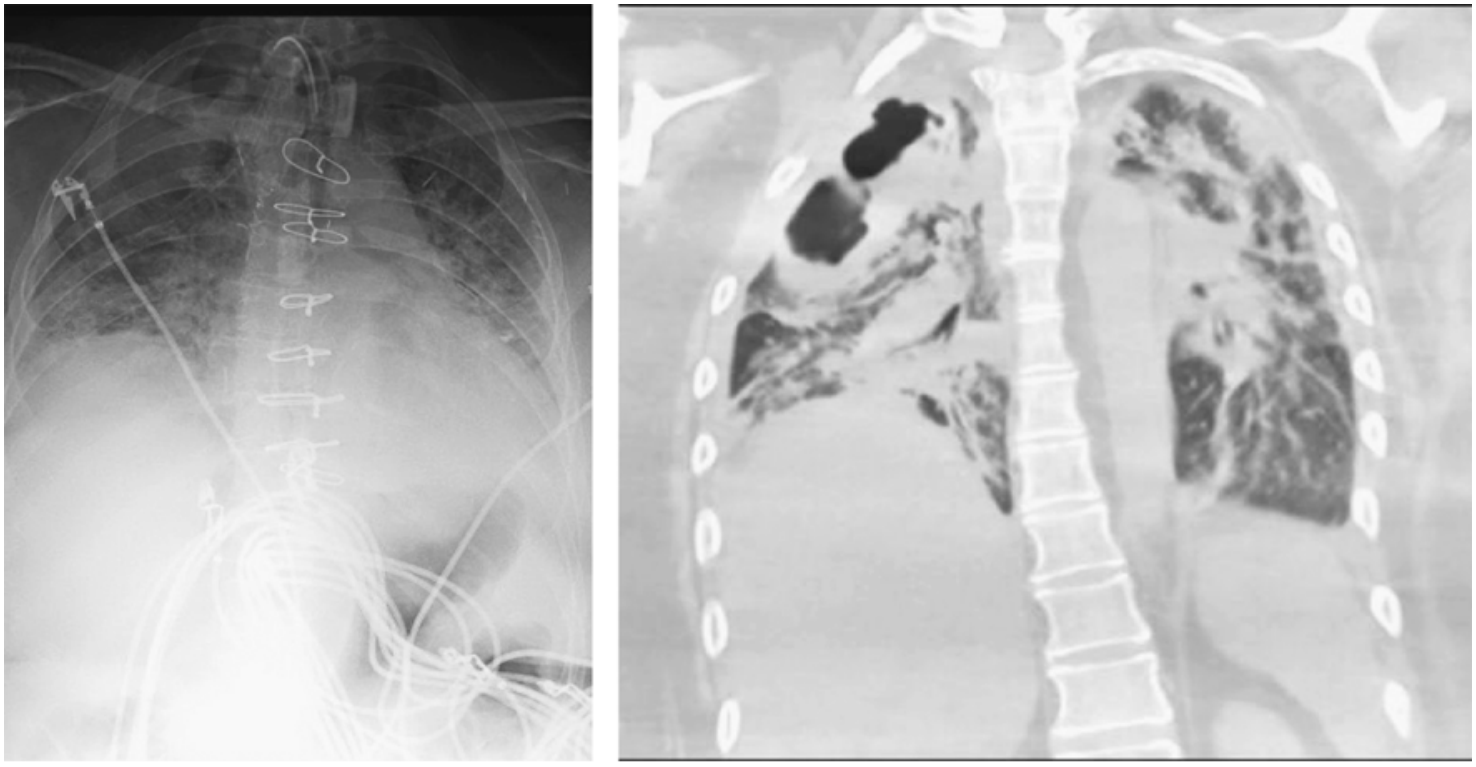
A mixed bacterial population with many PMNs is evident on Gram’s staining of sputum. Expectorated sputum is unreliable for anaerobic cultures because of inevitable contamination by the normal oral microbiota. Reliable specimens for culture can be obtained by transtracheal or transthoracic aspiration—techniques that are rarely used at present. Although the culture of protected-brush specimens or bronchoalveolar lavage fluid obtained by bronchoscopy is controversial, more recent data suggest that these approaches can be used without oropharyngeal contamination and can recover anaerobic organisms from the lower respiratory tract in a site-directed manner. Further research is needed to determine how these approaches compare with the previous gold standards.

ANAEROBIC LUNG ABSCESSES

These abscesses result from subacute anaerobic pulmonary infection **(See also Chap. 122)**. The clinical presentation typically involves a history of constitutional signs and symptoms (including malaise, weight loss, fever, night sweats, and foul-smelling sputum) that have typically persisted for 1–3 weeks prior to hospitalization. Patients who develop lung abscesses often, but not always, have an antecedent dental infection. Abscess cavities may be single or multiple and generally occur in dependent pulmonary segments **(Fig. 172-3)**. The differential diagnosis for lung abscesses includes pneumonia (including necrotizing pneumonia), a purulent pleural effusion with a bronchopleural fistula, and a pneumatocele. Of note, infection with some aerobic organisms, particularly *S. aureus*, can develop into a lung abscess without an anaerobic component. Approximately 90% of cases have an anaerobe identified—usually three to six isolates per sample—if careful attention is paid to handling and processing of the abscess sample. The most common isolates include peptostreptococci, *Prevotella* and *Porphyromonas* species, and *F. nucleatum*. An important finding is that ~90% of cultures also demonstrate the presence of aerobic organisms, such as *S. aureus*, *Streptococcus pneumoniae*, and *Klebsiella pneumoniae*. Consistent with the notion that anaerobes are contributing to disease, patients often do not improve clinically until they receive an antibiotic regimen that includes anaerobic coverage.

FIGURE 172-3

Chest radiograph (*left*) and CT image (*right*) of a lung abscess. The patient aspirated while supine and developed an abscess in the posterior segment of the right upper lobe. Cultures were pretreated and grew only *Klebsiella pneumoniae*. *(Images provided by Gita N. Mody, MD, MPH, Division of Thoracic Surgery, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EMPYEMA

Empyema is a manifestation of long-standing anaerobic pulmonary infection and manifests with thick, purulent material in the pleural space, often in association with a bronchopleural fistula. Alternatively, a subdiaphragmatic infection may extend into the pleural space and similarly result in an empyema. The clinical presentation resembles that of other anaerobic pulmonary infections and may include foul-smelling sputum, pleuritic chest pain, and marked chest-wall tenderness. This disease process must be differentiated from a parapneumonic effusion resulting from more routine causes of pneumonia (e.g., *S. pneumoniae*). In the latter instance, the fluid is a thin exudate that has a mean white blood cell (WBC) count of ~5000 cells/mL, a lactate dehydrogenase level

of >200 IU/L, and a pH of ~7.4. In contrast, empyema is characterized by foul-smelling thick pus with a mean WBC count of ~55,000 cells/mL, a lactate dehydrogenase level of >1000 IU/L, and a pH of <7.2 as well as loculations and a thick pleural peel on imaging. Drainage and occasionally decortication of the visceral and parietal pleura are required. Defervescence, a return to a feeling of well-being, and resolution of the process may require several months, particularly in the absence of surgical intervention.

Intraabdominal Infections

Breach of the gut mucosal surface (e.g., due to trauma, intestinal perforation, or malignancy) allows members of the microbiota to enter the normally sterile peritoneum. Accordingly, the offending organisms reflect the microbiota in the affected intestinal region. For example, recovery of *Candida* species from intraabdominal infections should prompt evaluation of the stomach and proximal small bowel for potential perforation. Furthermore, a study of patients with perforated and gangrenous appendicitis demonstrated that virtually all samples yielded *E. coli* and members of the *B. fragilis* group; peptostreptococci and *Bilophila wadsworthia*—additional components of the appendiceal and colonic microbiota—were also recovered from >50% of samples. Notably, some studies have identified an average of 10 different bacterial species, with an anaerobe-to-aerobe ratio of ~3:1. Given that >1000 bacterial species are present in the colonic microbiota, the dominance of such a limited repertoire of bacterial genera and species recovered in intraabdominal infections reflects a combination of two factors: the increased propensity of these organisms to result in intraabdominal abscesses and the difficulty faced by clinical microbiology laboratories in culturing the diverse organisms present in these samples. **See Chap. 127 for a complete discussion of intraabdominal infections.**

Neutropenic enterocolitis (typhlitis) involves marked thickening of the bowel wall (typically >4 mm) in the setting of neutropenia, abdominal pain, and fever. This condition most commonly affects the cecum and may extend to the neighboring terminal ileum and/or proximal colon, but any intestinal region may be involved. Typhlitis generally occurs after 1–2 weeks of chemotherapy-induced neutropenia associated with treatment of hematologic or, less commonly, solid tumor malignancies, but it can occur regardless of the cause of neutropenia. At least 5% of adults hospitalized for malignancy are thought to develop typhlitis, but this is likely an underestimate. Although the right lower quadrant is the most common location of abdominal pain and tenderness, these symptoms are absent in nearly half of cases; moreover, some patients, particularly those taking glucocorticoids, may not experience abdominal pain at all. Given the weakened integrity of the bowel wall and the associated neutropenia, patients often develop bacteremia due to one or more organisms related to the microbiota of the affected intestinal segment. Patients who develop bacteremia due to *Clostridium septicum* often have relatively severe disease, and identification of this organism is highly associated with the presence of malignancy—notably, colon cancer. Medical management including bowel rest, intestinal decompression, and broad-spectrum antibiotic administration is generally successful, although surgical intervention may be required in cases of persistent intestinal bleeding, necrotic bowel, or clinical deterioration suggestive of an ongoing intestinal process.

Pelvic Infections

Anaerobes are frequently encountered in pelvic inflammatory disease, pelvic abscess, endometritis, tubo-ovarian abscess, septic abortion, and postoperative or postpartum infections. These infections are often of mixed etiology, involving both anaerobes and coliforms; pure anaerobic infections without coliform or other facultative bacterial species occur more often in pelvic than in intraabdominal sites. The major anaerobic pathogens in pelvic abscesses are *P. bivia*, *P. disiens*, and the *B. fragilis* group, but many other anaerobes have also been implicated. **See Chap. 131 for a complete discussion of pelvic inflammatory disease.**

Anaerobic bacteria have been thought to be contributing factors in bacterial vaginosis. This syndrome of unknown etiology is characterized by a profuse malodorous discharge and a change in bacterial ecology that results in replacement of the *Lactobacillus*-dominated normal microbiota with an overgrowth of anaerobic bacterial species. Culture-based and culture-independent approaches have identified numerous organisms, including *Gardnerella vaginalis*, peptostreptococci, genital mycoplasmas, and species within the genera *Prevotella*, *Mobiluncus*, *Atopobium, Leptotrichia, Megasphaera*, and *Eggerthella*. This wide array of implicated bacteria may reflect differences in the overall disease spectrum of bacterial vaginosis and/or a shared physiologic response to these different organisms.

Skin and Soft Tissue Infections

Similar to other anatomic sites, skin or soft tissue injured by trauma, ischemia, or surgery creates a suitable environment for anaerobic infections. The infecting bacteria either are introduced directly (e.g., wounds associated with intestinal surgery, decubitus ulcers, or human bites) or originate in contiguous areas (e.g., cutaneous abscesses, rectal abscesses, and axillary sweat gland infections [*hidradenitis suppurativa*]). Anaerobes also are often cultured from foot ulcers of diabetic patients. The most common locations for anaerobic cellulitis include the neck, trunk, groin (including the genitalia), and legs. The deep soft-tissue infections associated with anaerobic bacteria are gas gangrene, synergistic cellulitis (both progressive and necrotizing), necrotizing fasciitis, and myositis **(Chaps. 124 and 149)**.

Gas gangrene (crepitus cellulitis) is most often due to *C. perfringens*, although other clostridial species have been implicated as well. This infection involves extensive gas formation in the tissue leading to crepitus and a thin, dark, occasionally malodorous discharge. True gas gangrene typically presents with fever and tenderness around the lesion and can rapidly spread; in contrast, there are somewhat more indolent forms of anaerobic cellulitis that may involve some gas formation but often present without fever or extensive local pain and can spread over the course of days rather than minutes.

Progressive bacterial synergistic gangrene (*Meleney gangrene*) is characterized by an area of exquisite pain, redness, and swelling followed by ulceration. As the ulcer enlarges, it is surrounded by a violaceous ring that fades into a pink edematous border. If it is not promptly treated, the ulcer continues to enlarge, and new, distant ulcers may emerge. Symptoms are limited to pain; fever is not typical. Peptostreptococci and microaerophilic streptococci are commonly found in the leading edge of the lesions, and *S. aureus* and *Proteus* species can be isolated from the ulcerated lesion. Treatment includes surgical removal of necrotic tissue and antimicrobial administration. In contrast, synergistic necrotizing cellulitis involves the deep fascia and occurs near the point of bacterial entry. Pain, fever, and systemic symptoms are common. If this form of cellulitis involves the scrotum, perineum, and anterior abdominal wall, it is referred to as *Fournier gangrene*. *S. aureus*, the *B. fragilis* group, *Peptostreptococcus* species, *Clostridium* species, *Fusobacterium* species, and members of the family Enterobacteriaceae are the predominant organisms identified.

Necrotizing fasciitis, a rapidly spreading destructive disease of the fascia, is usually attributed to group A streptococci **(Chap. 143)** but can also be a mixed infection involving anaerobes and aerobes. Polymicrobial necrotizing fasciitis differs from stereotypical group A streptococcal necrotizing fasciitis in that the initial erythematous, swollen, tender lesions progress over 3–5 days (as opposed to 1–3 days), with consequent skin breakdown and cutaneous gangrene. Fever, subcutaneous gas, development of anesthesia (often before skin necrosis), and a foul-smelling discharge are common. The particular clinical findings sometimes suggest the causative agent: regional lymphadenopathy suggests the *B. fragilis* group; necrosis and gangrene suggest *Clostridium* species, peptostreptococci, the *B. fragilis* group, and Enterobacteriaceae; bullous lesions suggest Enterobacteriaceae; a foul-smelling odor suggests *Bacteroides* and *Clostridium* species; and subcutaneous gas suggests peptostreptococci, *Clostridium* species, and the *B. fragilis* group. Moreover, diabetic infections are often associated with *Bacteroides* species, Enterobacteriaceae, and *S. aureus*, and infections related to trauma are associated with *Clostridium* species.

Although *S. aureus* is the typical cause of myositis, anaerobes—particularly *C. perfringens*—are often recovered from patients with pyogenic myositis. In anaerobic streptococcal myonecrosis, peptostreptococci are often identified along with group A streptococci or *S. aureus*. Patients typically present with fever, muscle pain, fatigue, and an elevated creatinine kinase level suggestive of muscle inflammation.

Bone and Joint Infections

A comprehensive review of the world literature on anaerobic bone infections included >650 cases. Of these, ~400 cases were caused by *Actinomyces* species; anaerobic cocci and *Bacteroides*, *Fusobacterium*, and *Clostridium* species were most commonly identified in the remaining cases. Actinomycotic involvement of the jaw was the most common bone infection, with the mandible involved four times as frequently as the maxilla. Patients with cervicofacial actinomycosis **Chap. 170)** are often described as having a “lumpy jaw” because of the prominent soft-tissue swelling that is sometimes mistaken for malignancy or granulomatous disease. These infections can be chronic in nature, can include the development of sinus tracts, can progress across normal tissue boundaries, and

can require prolonged antibiotic treatment to prevent relapse. The vertebrae are the second most common location for *Actinomyces* infection; involvement of the thorax, abdomen, or pelvis is much less frequent.

Osteomyelitis involving anaerobes other than *Actinomyces* species most commonly develops by extension of an adjacent infection (e.g., soft tissue, paranasal sinus, or middle-ear infection). For example, diabetic foot ulcers and decubitus ulcers may be complicated by mixed aerobic–anaerobic osteomyelitis **(Chap. 126)**. Hematogenous seeding of bone by anaerobes is uncommon and is thought to occur in fewer than 10% of cases. The most common sites of anaerobic osteomyelitis are the head (skull and jaw) and the extremities. Fusobacteria have been isolated in pure culture from infections of the mastoid, mandible, and maxilla. *Clostridium* species have been reported as anaerobic pathogens in cases of osteomyelitis of the long bones following trauma. Anaerobic and microaerophilic cocci are most frequently isolated from infections involving the skull or mastoid; usually, these organisms are present in mixed cultures.

In contrast to anaerobic osteomyelitis, anaerobic arthritis **(Chap. 125)** is uncommon, typically involving a single isolate, and most cases are secondary to hematogenous spread. Although *Fusobacterium* species accounted for nearly one-third of cases in the pre-antibiotic era, *P. acnes*, peptostreptococci, and *B. fragilis* are now among the more frequent causes of anaerobic septic arthritis. Peptostreptococci and *P. acnes* are often found in association with prosthetic joints, *Fusobacterium* species have a predilection for the sternoclavicular and sacroiliac joints, and clostridial arthritis is especially common after trauma. As a frequent cause of bacteremia, *B. fragilis* is a common cause of anaerobic septic arthritis; however, arthritis occurs in fewer than 5% of patients with *B. fragilis* bacteremia.

Bacteremia

B. fragilis is the anaerobe most commonly isolated from blood cultures. Although the frequency of positive cultures appeared to be decreasing in the 1980s, more recent evidence suggests that the rate is now increasing and that the increase may be related to changing demographics, with more patients who are elderly, immunocompromised, and/or receiving medications that may disrupt the mucosal barrier (e.g., chemotherapy). The source of bacteremia is most often an abscess in the abdomen, female genital tract, or soft tissue. At a large tertiary-care U.S. hospital, 0.8% of all positive blood cultures yielded an anaerobic gram-negative bacillus, with 0.5% yielding *B. fragilis*. A similar study in France revealed that 0.6% of all positive blood cultures yielded an anaerobic organism; 60% of these isolates were *Bacteroides* species, and 22% were *Clostridium* species. *Peptostreptococcus* and *Fusobacterium* species are also recovered with significant frequency.

Once the organism in the blood has been identified, both the portal of bloodstream entry and the underlying problem that probably led to seeding of the bloodstream can often be deduced from an understanding of the organism's normal site of residence. For example, mixed anaerobic bacteremia including *B. fragilis* usually implies a colonic pathology, with mucosal disruption from neoplasia, diverticulitis, or some other inflammatory lesion. The initial manifestations are determined by the portal of entry and reflect the localized condition. Although the clinical manifestations of *B. fragilis* bacteremia (e.g., rigors, hectic fevers) are similar to those of aerobic gram-negative bacillary bacteremia, the incidence of septic shock is lower with *B. fragilis*. This difference may be due to differences in the immunostimulatory effects of the different endotoxin structures.

In virtually all cases, isolation of a member of the *B. fragilis* group from blood indicates underlying infection that is associated with a mortality rate of 60% if untreated. It has been suggested that the mortality rate depends in part on the species recovered (*B. thetaiotaomicron* > *P. distasonis* > *B. fragilis*), but it is unclear whether differences in mortality rates relate to intrinsic differences in the virulence of these organisms, in their antimicrobial susceptibility profiles, and/or in the host's immune response. Case–fatality rates appear to increase with the increasing age of the patient (with reported rates of >66% among patients >60 years old), with the isolation of multiple species from the bloodstream, and with the failure to surgically remove a focus of infection.

Endocarditis

Although gram-negative anaerobic bacteria only rarely cause endocarditis, their involvement is associated with significant mortality rates (21–43%) **(See also Chap. 123)**. Members of the *B. fragilis* group are the most commonly identified gram-negative anaerobes in endocarditis. Anaerobic streptococci, which are often classified incorrectly, are likely responsible for this disease more

frequently than is generally appreciated. Compared to aerobic bacterial endocarditis, endocarditis due to *Bacteroides* species is less likely to be associated with a history of cardiovascular disease and more likely to involve thromboembolic complications.

DIAGNOSIS

There are three critical steps in the diagnosis of anaerobic infection: (1) proper collection of specimens; (2) rapid transport of the specimens to the microbiology laboratory, preferably in anaerobic transport media; and (3) proper handling of the specimens by the laboratory. Specimens must be collected by meticulous sampling of infected sites, with avoidance of contamination by the normal microbiota. Samples from sites known to harbor numerous anaerobes (e.g., the mouth, nose, vagina, feces) are not acceptable for anaerobic culture as the presence of the normal microbiota will complicate interpretation of the results in a clinically meaningful manner. In contrast, samples from normally sterile locations (e.g., blood, pleural fluid, peritoneal fluid, CSF, and aspirates or biopsy samples from normally sterile sites) are appropriate for anaerobic culture in clinical microbiology laboratories. As a general rule, liquid or tissue specimens are preferred; if swab specimens must be used, special anaerobic swab systems should be used to help maintain persistence of anaerobes. Liquid samples should be collected in an air-free syringe that is then capped, injected into anaerobic transport bottles, or quickly transported to the clinical microbiology laboratory for immediate culture.

Because of the time and difficulty involved in the isolation of anaerobic bacteria, the diagnosis of anaerobic infections must frequently be based on presumptive evidence. As mentioned previously, anaerobic infections are sometimes suggested by specific clinical factors, such as origins from a site with an anaerobic-rich microbiota (e.g., the intestinal tract, oropharynx), the presence of an abscess, involvement of sites with lowered oxidation-reduction potential (e.g., avascular necrotic tissues), a foul odor, and the presence of gas in tissues. None of these features is necessarily pathognomonic or required for the diagnosis of an anaerobic infection, but these are helpful clues to keep in mind when constructing a differential diagnosis.

When cultures of obviously infected sites or purulent material yield no growth, streptococci only, or a single aerobic species (such as *E. coli*) and Gram's staining reveals a mixed bacterial population, the involvement of anaerobes should be suspected; the implication is that the anaerobic microorganisms have failed to grow because of inadequate transport and/or culture techniques. It is also important to remember that prior antibiotic therapy reduces the cultivability of these bacteria. Failure of an infection to respond to antibiotics that are not active against anaerobes (e.g., aminoglycosides and—in some circumstances—penicillin, cephalosporins, or tetracyclines) suggests an anaerobic etiology.

TREATMENT

TREATMENT

Anaerobic Infections

Similar to successful therapy for other types of infection, treatment for anaerobic infections requires the administration of appropriate antibiotics, surgical debridement of devitalized tissues, and drainage of any large abscess. Any mucosal breach must be closed promptly to prevent ongoing infection.

ANTIBIOTIC THERAPY AND RESISTANCE

The antibiotics used to treat anaerobic infections should be active against both aerobic and anaerobic organisms because many of these infections are of mixed etiology. Antibiotic regimens can usually be selected empirically on the basis of the location of infection (which provides insight into the likely species involved), the severity of infection, and knowledge of local antimicrobial resistance patterns. Other factors influencing the selection of antibiotics include need for penetration into certain organs (such as the brain) and associated toxicity (Chap. 139). As with all infections, the general maxim is to use the least broad-spectrum agent(s) possible so as to minimize the impact on the normal microbiota and the development of resistance.

Because of the slow growth rate of many anaerobes, the lack of standardized testing methods and of clinically relevant standards for resistance, and the generally good results obtained with empirical therapy, the role of antibiotic susceptibility testing of these organisms has been limited in most clinical microbiology laboratories. Instead, isolates are sent to reference laboratories for

susceptibility testing when an infection is serious (e.g., brain abscess, meningitis, joint infection), is refractory, or requires prolonged therapy (e.g., osteomyelitis, prosthetic joint infection, endocarditis). Such testing should also be considered when a patient is not responding to antimicrobial therapy as expected; multidrug-resistant anaerobes have been reported. Antimicrobial susceptibility testing is also helpful in monitoring the activity of new drugs and recording current resistance patterns among anaerobic pathogens.

The need for susceptibility testing of anaerobic organisms is highlighted by increasing rates of antimicrobial resistance, geographic and institutional differences in susceptibility profiles, species-specific antibiograms, and the potential for worse clinical outcomes when ineffective antibiotics are used. These differences preclude making any sweeping generalizations regarding antibiotic therapy for anaerobic infections. For example, rates of resistance to piperacillin-tazobactam have remained low (≤1%) for all *Bacteroides* species in the United States, but *B. thetaiotaomicron* isolates in Korea have a notably higher resistance rate (17%). Clindamycin was historically effective against members of the *B. fragilis* group, but rates of resistance have increased to 30–43% in the United States and are >80% in some parts of the world. Furthermore, metronidazole is effective against many different anaerobic organisms and is considered a first-line agent for many anaerobic infections worldwide, but, in a population of Colombian patients with refractory periodontitis, 45% of *Fusobacterium* isolates and 25% of *Prevotella* and *Porphyromonas* strains were resistant to metronidazole; this finding underscores the importance of understanding the local antibiogram and of assessing susceptibility profiles in refractory disease.

Empirical Therapy

Not every anaerobe isolated must be specifically targeted by the antibiotic regimen. Given that infections involving anaerobes are typically polymicrobial, that the cultivation and identification of anaerobes are challenging (i.e., not all organisms may be recovered), and that organisms often depend on one another for persistence, clinical resolution of the infection is often achieved with empirical antibiotics targeting the bulk of the organisms recovered. Antibiotics that demonstrate no useful activity against anaerobes include aminoglycosides, monobactams, and trimethoprim-sulfamethoxazole. With the caveat that susceptibility profiles may change with time and geography, the antibiotics that are commonly used as empirical therapy against anaerobic bacteria include metronidazole, β-lactam/β-lactamase inhibitor combinations, clindamycin, carbapenems, and chloramphenicol **(Table 172-2)**.

Metronidazole is active against gram-negative anaerobes, including nearly all isolates of *Bacteroides* species, and gram-positive spore-forming organisms, such as *C. difficile* **(Chap. 129)** and other *Clostridium* species. Given intrinsically reduced susceptibility, metronidazole is clinically unreliable against gram-positive non-spore-forming organisms, such as *Actinomyces*, *Propionibacterium*, *Lactobacillus*, *Bifidobacterium*, *Eubacterium*, and *Peptostreptococcus*. Of note, a few metronidazole-resistant *Bacteroides* isolates have been identified in the United States, and rates of such resistance have been increasing in Europe. Moreover, the rate of resistance to metronidazole has probably been greatly underestimated in some countries (e.g., the United Kingdom) that use metronidazole susceptibility to discriminate between obligate and facultative anaerobes (with obligate anaerobes defined by their susceptibility). Although the majority of metronidazole-resistant isolates have been identified in patients who have been exposed to the drug, resistant organisms have also been found in metronidazole-naïve patients.

More than 90% of clinical isolates from the *B. fragilis* group produce β-lactamases that are predominantly active against cephalosporins and that are highly active, cell associated, and produced constitutively. Thus, members of the *B. fragilis* group are presumed to be resistant to penicillin and ampicillin but may remain susceptible to extended-spectrum penicillins, particularly in combination with a β-lactamase inhibitor (e.g., ampicillin-sulbactam, piperacillin-tazobactam). Rates of resistance to ampicillin-sulbactam are increasing, particularly in *P. distasonis*, which has a reported resistance rate of 21% in the United States. Because β-lactamase production is not common in *Clostridium* species, these combination agents are usually effective. Of note, the newer cephalosporin/β-lactamase inhibitors (e.g., ceftolozane-tazobactam, ceftazidime-avibactam) have limited anaerobic activity.

Clindamycin is active against many anaerobes. However, rates of resistance to clindamycin among *Bacteroides* species increased in the United States from 7% in 1981 to 35% in 2008–2009. Resistance to clindamycin among non-*Bacteroides* gram-negative anaerobes is much less common (<10%). Some *Clostridium* species are resistant to clindamycin, although *C. perfringens* typically is not.

Carbapenems (ertapenem, doripenem, meropenem, and imipenem) are active against anaerobes, with fewer than 5% of *Bacteroides* species resistant. There is little difference among resistance rates for specific species, and, of the carbapenems, imipenem typically has the lowest resistance rate. Although the β-lactamase produced by most *Bacteroides* species is unable to inactivate carbapenems, rare *B. fragilis* strains have been reported to produce a carbapenemase.

Resistance to chloramphenicol is rare in *Bacteroides* species. Nationwide surveys in the United States have identified no resistant organisms, but some isolates with elevated minimal inhibitory concentrations (MICs)—i.e., 16 μg/mL—have been noted. Although chloramphenicol has excellent in vitro activity against all clinically relevant anaerobes, some clinical failures have been documented. Therefore, this drug may be less preferable if other active agents are available.

Other antibiotics with more variable activity against anaerobes include the fluoroquinolones and tigecycline. Although many fluoroquinolones (e.g., ciprofloxacin, levofloxacin, ofloxacin) display reasonable activity against anaerobic organisms other than *Bacteroides* species, these agents exhibit poor activity against the *B. fragilis* group. Rates of resistance to moxifloxacin are relatively high (39–83%) among *Bacteroides* isolates obtained in the United States but are much lower among *B. fragilis* and *B. thetaiotaomicron* isolates collected in Korea (8 and 2%, respectively) or Taiwan (8 and 15%, respectively). Tigecycline is active against most anaerobic bacteria, although MICs are somewhat higher for *Clostridium* species. Tigecycline's efficacy for treatment of complicated intraabdominal infections is comparable to that of imipenem, and it is therefore recommended as single-agent therapy for these infections.

Infections at Specific Sites

In clinical situations, specific antibiotic regimens and durations must be tailored to the initial site of infection; the reader is referred to specific chapters on infections at specific sites for recommendations. In general, anaerobic infections are often broadly categorized as originating above or below the diaphragm. This distinction is clinically useful in that the predominant pathogens—and therefore the empirical antibiotic regimens—differ between these two categories of infection.

Infections above the diaphragm usually reflect the orodental microbiota, which includes *Prevotella*, *Porphyromonas*, *Fusobacterium*, and *Bacteroides* species other than the *B. fragilis* group along with streptococci (both aerobic and microaerophilic). Accordingly, antibiotic regimens should cover both aerobic and anaerobic bacteria. Given that >70% of these infections include a β-lactamase-producing organism, β-lactam drugs (penicillins and cephalosporins) are poor options as monotherapy. The recommended regimens include clindamycin, a β-lactam/β-lactamase inhibitor combination, or metronidazole in combination with a drug active against microaerophilic and aerobic streptococci (e.g., penicillin).

Anaerobic infections arising below the diaphragm (e.g., colonic and intraabdominal infections) must be treated specifically with agents active against *Bacteroides* species, including *B. fragilis*. Single agents suitable for this purpose include cefoxitin, moxifloxacin, a β-lactam/β-lactamase inhibitor combination, or a carbapenem. A two-drug regimen is an alternative, with one drug active against anaerobes and the other against coliforms (e.g., metronidazole with either a cephalosporin or a fluoroquinolone). In addition, if the clinician suspects that gram-positive facultative organisms such as enterococci are involved, therapeutic regimens should include ampicillin or vancomycin. Although clindamycin and cefotetan were previously considered acceptable options for intraabdominal infections involving anaerobes, these drugs are no longer recommended because of escalating rates of resistance in the *B. fragilis* group. Ampicillin-sulbactam is not recommended because of high rates of resistance among community-acquired strains of *E. coli* rather than because of resistance in anaerobic bacteria.

CNS infections involving anaerobic organisms may be treated with metronidazole, a carbapenem, chloramphenicol, or—if only gram-positive anaerobes are involved—penicillin. Clindamycin and cefoxitin have poor penetration into the CSF and should not be used. Cases of osteomyelitis in which a polymicrobial infection is identified from a bone biopsy specimen should be treated with a regimen that covers both aerobes and anaerobes, as some organisms that are often regarded as a contaminant (e.g., *P. acnes*) may have a pathogenic role. When an anaerobic organism is recognized as a major or sole pathogen infecting a joint, the duration of treatment should be similar to that used for arthritis caused by aerobic bacteria **(Chap. 125)**.

Although not every anaerobe needs to be covered with pathogen-directed therapy in most polymicrobial infections, several studies of *Bacteroides* bacteremia have clearly demonstrated that patients receiving effective therapy have lower mortality rates and more

rapid sterilization of blood cultures than patients receiving ineffective therapy.

FAILURE OF THERAPY

Anaerobic infections that fail to respond to treatment or that relapse should be reassessed. Potential causes include an uncontrolled source of infection (e.g., ongoing intestinal leak into the peritoneum), superinfection with a new organism, and/or antibiotic failure. Additional imaging may be useful to discern whether surgical drainage or debridement is warranted. Obtaining additional culture specimens will help identify whether an organism resistant to the antibiotics being used is present. Strong consideration should be given to obtaining susceptibility profiles for the isolates.

TABLE 172-2

Antimicrobial Therapy That Is Typically Active Against Commonly Encountered Anaerobes

Antibiotic(s)	Caveats
Metronidazole	This drug is clinically unreliable against gram-positive non-spore-forming anaerobes (e.g., *Actinomyces* spp., *Propionibacterium* spp., *Peptostreptococcus* spp.).
β-Lactam/β-lactamase inhibitor combinations (ampicillin-sulbactam, ticarcillin–clavulanic acid, piperacillin-tazobactam)	Rates of resistance are increasing in some gram-negative anaerobes. The newer cephalosporin/β-lactamase combinations have limited anaerobic activity.
Clindamycin	Rates of resistance are increasing in *Bacteroides* spp.
Carbapenems (meropenem, imipenem, ertapenem, doripenem)	Rates of resistance are currently very low (<5%), although some carbapenemase-producing strains have been identified.
Chloramphenicol	Some clinical failures have been noted, even when the isolate is found to be susceptible by in vitro testing.

ACKNOWLEDGMENT

The authors thank Ronit Cohen-Poradosu for her contributions to this chapter in previous editions.

FURTHER READING

Brook I: Antimicrobial therapy of anaerobic infections. J Chemother 28:143, 2016.

Citron DM, Appelbaum PC: How far should a clinical laboratory go in identifying anaerobic isolates, and who should pay? Clin Infect Dis 16(Suppl 4):S435, 1993. [PubMed: 8324163]

Finegold SM: Anaerobes: Problems and controversies in bacteriology, infections, and susceptibility testing. Rev Infect Dis 12(Suppl 2):S223, 1990. [PubMed: 2406874]

Styrt B, Gorbach SL: Recent developments in the understanding of the pathogenesis and treatment of anaerobic infections (2). N Engl J Med 321:240, 1989.

Wexler HM: *Bacteroides*: The good, the bad, and the nitty-gritty. Clin Microbiol Rev 20:593, 2007.

Chapter 173: Tuberculosis

Mario C. Raviglione

INTRODUCTION

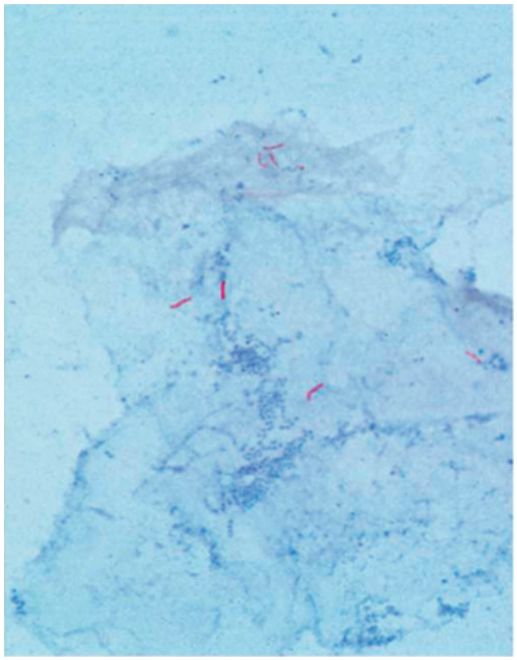
Tuberculosis (TB), which is caused by bacteria of the *Mycobacterium tuberculosis* complex, is one of the oldest diseases known to affect humans and the top cause of infectious death worldwide. Population genomic studies suggest that *M. tuberculosis* may have emerged ~70,000 years ago in Africa and subsequently disseminated along with anatomically modern humans, expanding globally during the Neolithic Age as human density started to increase. Progenitors of *M. tuberculosis* are likely to have affected prehominids. This disease most often affects the lungs, although other organs are involved in up to one-third of cases. If properly treated, TB caused by drug-susceptible strains is curable in the vast majority of cases. If untreated, the disease may be fatal within 5 years in 50–65% of cases. Transmission usually takes place through the airborne spread of droplet nuclei produced by patients with infectious pulmonary TB.

ETIOLOGIC AGENT

Mycobacteria belong to the family Mycobacteriaceae and the order Actinomycetales. Of the pathogenic species belonging to the *M. tuberculosis* complex, which comprises eight distinct subgroups, the most common and important agent of human disease by far is *M. tuberculosis* (*sensu stricto*). A closely related organism isolated from cases in West, Central, and East Africa is *M. africanum.* The complex includes some zoonotic members, such as *M. bovis* (the bovine tubercle bacillus—characteristically resistant to pyrazinamide, once an important cause of TB transmitted by unpasteurized milk, and currently responsible for ~150,000 human cases worldwide, half of them in Africa) and *M. caprae* (related to *M. bovis*). In addition, other organisms that have been reported rarely as causing TB include *M. pinnipedii* (a bacillus infecting seals and sea lions in the Southern Hemisphere and recently isolated from humans), *M. mungi* (isolated from banded mongooses in southern Africa), *M. orygis* (described in oryxes and other Bovidae in Africa and Asia and a potential cause of infection in humans), and *M. microti* (the “vole” bacillus, a less virulent organism). Finally, *M. canetti* is a rare isolate from East African cases that produces unusual smooth colonies on solid media and is considered closely related to a supposed progenitor type. There is no known environmental reservoir for any of these organisms.

M. tuberculosis is a rod-shaped, non-spore-forming, thin aerobic bacterium measuring 0.5 µm by 3 µm. Mycobacteria, including *M. tuberculosis*, are often neutral on Gram’s staining. However, once stained, the bacilli cannot be decolorized by acid alcohol; this characteristic justifies their classification as acid-fast bacilli (AFB; **Fig. 173-1**). Acid fastness is due mainly to the organisms’ high content of mycolic acids, long-chain cross-linked fatty acids, and other cell-wall lipids. Microorganisms other than mycobacteria that display some acid fastness include species of *Nocardia* and *Rhodococcus*, *Legionella micdadei*, and the protozoa *Isospora* and *Cryptosporidium*. In the mycobacterial cell wall, lipids (e.g., mycolic acids) are linked to underlying arabinogalactan and peptidoglycan. This structure results in very low permeability of the cell wall, thus reducing the effectiveness of most antibiotics. Another molecule in the mycobacterial cell wall, lipoarabinomannan, is involved in the pathogen–host interaction and facilitates the survival of *M. tuberculosis* within macrophages.

FIGURE 173-1
Acid-fast bacillus smear showing *M. tuberculosis* bacilli. *(Courtesy of the Centers for Disease Control and Prevention, Atlanta.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

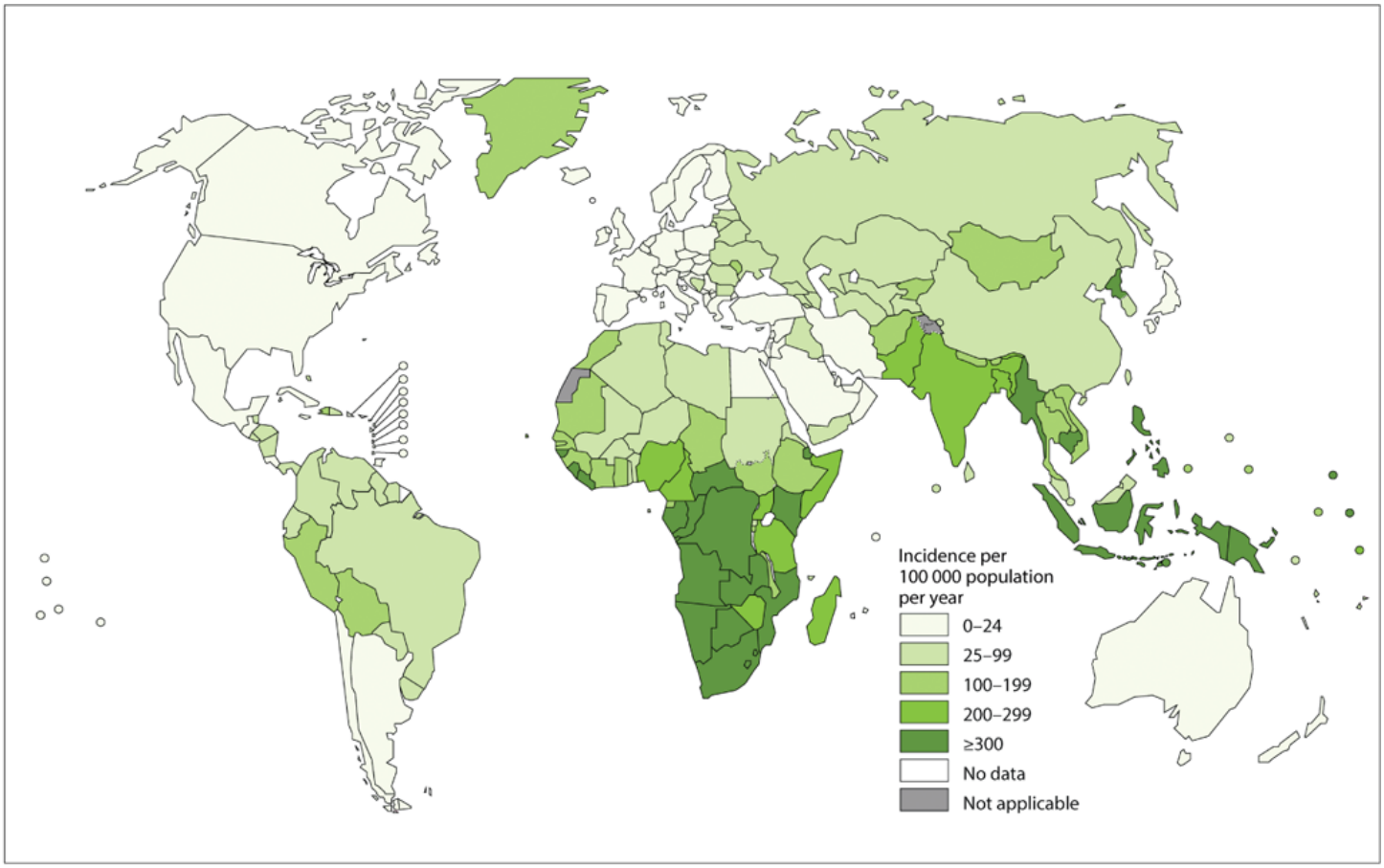
The complete genome sequence of *M. tuberculosis* comprises 4.4 million base pairs, 4043 genes encoding 3993 proteins, and 50 genes encoding RNAs; its high guanine-plus-cytosine content (65.6%) is indicative of an aerobic "lifestyle." A large proportion of genes are devoted to the production of enzymes involved in cell wall metabolism. Substantial genetic variability exists among strains from different parts of the world.

EPIDEMIOLOGY

In 2016, 6.3 million new cases of TB (all forms, both pulmonary and extrapulmonary) were reported to the World Health Organization (WHO) by its member states; 95% of cases were reported from developing countries. However, because of insufficient case detection and incomplete notification, reported cases may represent only about two-thirds of the total estimated cases. As a result, the WHO estimated that 10.4 million (range, 8.8–12.2 million) new (incident) cases of TB occurred worldwide in 2016, 95% of them in developing countries of Asia (6.5 million), Africa (2.6 million), the Middle East (0.77 million), and Latin America (0.26 million). Seven countries accounted for 64% of all new cases: India, Indonesia, China, the Philippines, Pakistan, Nigeria, and South Africa. Two-thirds of cases typically occur in male patients, and 1.04 million children are affected every year. It is further estimated that 1.7 million (range, 1.5–1.8 million) deaths from TB, including 0.37 million among people with HIV infection, occurred in 2016; 96% of these deaths were in developing countries. Estimates of TB incidence rates (per 100,000 population) and numbers of TB-related deaths in 2016 are depicted in **Figs. 173-2 and 173-3**, respectively. During the late 1980s and early 1990s, numbers of reported cases of TB increased in industrialized countries. These increases were related largely to immigration from countries with a high incidence of TB; the spread of the HIV epidemic; social problems, such as increased urban poverty, homelessness, and drug abuse; and dismantling of TB services. During the past few years, numbers of reported cases have begun to decline again or have stabilized in most industrialized nations. In the United States, with the re-establishment of stronger control programs, the decline resumed in 1993 and had since been maintained until 2015, when numbers increased over the previous year for the first time in more than two decades; in that year, 9557 cases of TB (3.0 cases/100,000 population) were reported to the U.S. Centers for Disease Control and Prevention (CDC). However, in 2016 a slight decline from 2015 was observed in incidence (2.9 cases/100,000 population) and number of cases (9287).

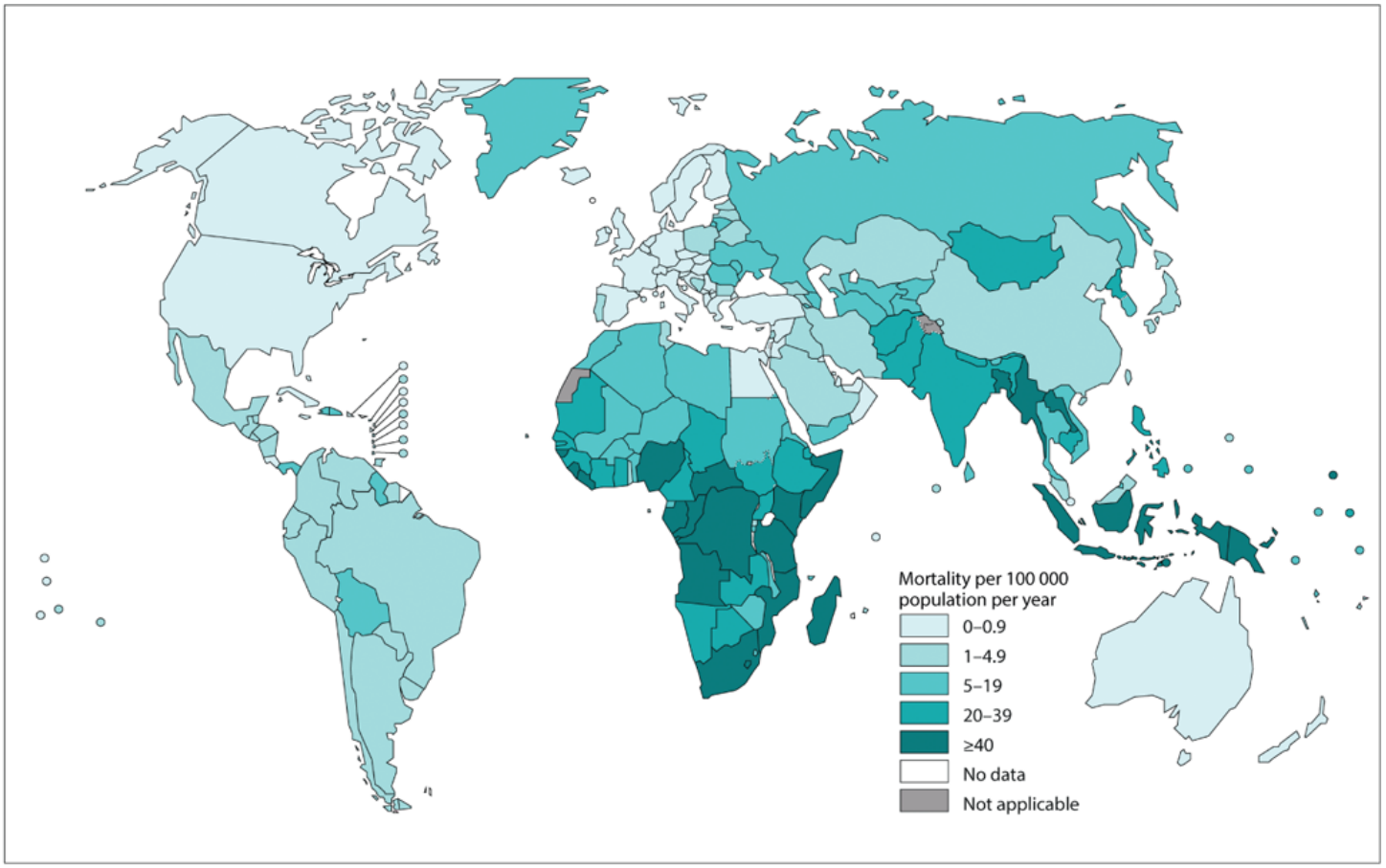
FIGURE 173-2

Estimated tuberculosis (TB) incidence rates (per 100,000 population) in 2016. The designations used and the presentation of material on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization (WHO) concerning the legal status of any country, territory, city, or area or of its authorities or concerning the delimitation of its frontiers or boundaries. *Dotted*, *dashed*, and *white lines* represent approximate border lines for which there may not yet be full agreement. *(Courtesy of the Global TB Programme, WHO; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.
FIGURE 173-3

Estimated tuberculosis mortality rates, excluding tuberculosis-related deaths among HIV-positive people, in 2016. *(See disclaimer in Fig. 173-2. Courtesy of the Global TB Programme, WHO; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In the United States, TB is uncommon among young adults of European descent, who have only rarely been exposed to *M. tuberculosis* infection during recent decades. In contrast, because of a high risk of transmission in the past, the prevalence of latent *M. tuberculosis* infection (LTBI) is relatively high among elderly whites. In general, adults ≥65 years of age have the highest incidence rate per capita (4.8 cases/100,000 population in 2016) and children <14 years of age the lowest (0.7 case/100,000 population). Among U.S.-born persons, blacks account for the highest proportion of cases (36%; 1062 cases in 2016). TB in the United States is also a disease of adult members of the HIV-infected population, the foreign-born population (68.5% of all cases in 2016), and disadvantaged/marginalized populations. Of the 6307 cases reported among foreign-born persons in the United States in 2016, 31% occurred in persons from the Americas and 47% in persons born in Asia. Overall, the highest rates per capita were among Asian Americans (18 cases/100,000 population). A total of 493 deaths were caused by TB in the United States in 2015. In Canada in 2015, 1639 TB cases were reported (4.6 cases/100,000 population); 71% (1169) of these cases occurred in foreign-born persons, and 17% (470 cases) occurred in members of the Canadian aboriginal peoples, whose per capita rate is disproportionately high (17.1 cases/100,000 population). The highest rate was found in the territory of Nunavut, at 119 cases/100,000 population—a rate similar to that in many highly endemic countries. Similarly, in Europe, TB has reemerged as an important public health problem, mainly as a result of cases among immigrants from high-incidence countries and among marginalized populations, often in large urban settings like London. In 2015, 39.4% of all cases reported from England occurred in London, and the rate per capita (26 cases/100,000 population) was similar to that in some middle-income countries. In most Western European countries, there are more cases annually among foreign-born than native populations.

Recent data on global trends indicate that in 2015 the TB incidence was stable or falling in most regions; this trend began in the early 2000s and appears to have continued, with an average annual decline of 1.5% globally. This global decrease is explained largely by the simultaneous reduction in TB incidence in sub-Saharan Africa, where rates had risen steeply since the 1980s as a result of the HIV epidemic and the lack of capacity of health systems and services to deal with the problem effectively, and in Eastern Europe, where incidence increased rapidly during the 1990s because of a deterioration in socioeconomic conditions and the health care infrastructure (although, after peaking in 2001, incidence in Eastern Europe has since declined slowly).

Of the estimated 10.4 million new cases of TB in 2016, 10% (1.03 million) were associated with HIV infection, and 74% of these HIV-associated cases occurred in Africa. An estimated 0.37 million persons with HIV-associated TB died in 2016. Furthermore, an estimated 500,000 cases of multidrug-resistant TB (MDR-TB)—a form of the disease caused by bacilli resistant at least to isoniazid and rifampin—and an additional 100,000 cases of rifampin-resistant TB (RR-TB), which also requires MDR-TB treatment (range for both forms together, 540,000–660,000), occurred in 2016. Only 25% of these cases were diagnosed because of a lack of culture and drug susceptibility testing (DST) capacity in many settings worldwide. As a consequence, 240,000 people with MDR/RR-TB died in 2016. The countries of the former Soviet Union have reported the highest proportions of MDR disease among new TB cases (up to 35% in some regions of Russia and Belarus). Overall, 47% of all MDR-TB cases occur in India, China, and the Russian Federation. Since 2006, 117 countries, including the United States, have reported cases of extensively drug-resistant TB (XDR-TB), in which MDR-TB is compounded by additional resistance to the most powerful

second-line anti-TB drugs (fluoroquinolones and at least one of the injectable drugs amikacin, kanamycin, and capreomycin). About 9.5% of the MDR-TB cases worldwide may be XDR-TB, but the vast majority of XDR-TB cases remain undiagnosed because reliable methods for DST are lacking and laboratory capacity is limited. Lately, a few cases deemed resistant to all anti-TB drugs have been reported; however, this information must be interpreted with caution because susceptibility testing for several second-line drugs is neither accurate nor reproducible.

FROM EXPOSURE TO INFECTION

M. tuberculosis is most commonly transmitted from a person with infectious pulmonary TB by droplet nuclei, which are aerosolized by coughing, sneezing, or speaking. The tiny droplets dry rapidly; the smallest (<5–10 μm in diameter) may remain suspended in the air for several hours and may reach the terminal air passages when inhaled. There may be as many as 3000 infectious nuclei per cough. Other routes of transmission of tubercle bacilli (e.g., through the skin or the placenta) are uncommon and of no epidemiologic significance. The risk of transmission and of subsequent acquisition of *M. tuberculosis* infection is determined mainly by exogenous factors. The probability of contact with a person who has an infectious form of TB, the intimacy and duration of that contact, the degree of infectiousness of the case, and the shared environment in which the contact takes place are all important determinants of the likelihood of transmission. Several studies of close-contact situations have clearly demonstrated that TB patients whose sputum contains AFB visible by microscopy (sputum smear–positive cases) are the most likely to transmit the infection. The most infectious patients have cavitary pulmonary disease or, much less commonly, laryngeal TB and produce sputum containing as many as 10^5–10^7 AFB/mL. Patients with sputum smear–negative/culture-positive TB are less infectious, although they have been responsible for up to 20% of transmission in some studies in the United States. Those with culture-negative pulmonary TB and extrapulmonary TB are essentially noninfectious. Because persons with both HIV infection and TB are less likely to have cavitations, they may be less infectious than persons without HIV co-infection. Crowding in poorly ventilated rooms is one of the most important factors in the transmission of tubercle bacilli because it increases the intensity of contact with a case. The virulence of the transmitted organism is also an important factor in establishing infection.

Because of delays in seeking care and in making a diagnosis, it has been estimated that, in high-prevalence settings, up to 20 contacts (or 3–10 people per year) may be infected by each AFB-positive case before the index case is diagnosed.

FROM INFECTION TO DISEASE

Unlike the risk of acquiring infection with *M. tuberculosis*, the risk of developing disease after being infected depends largely on endogenous factors, such as the individual's innate immunologic and nonimmunologic defenses and the level at which the individual's cell-mediated immunity is functioning. Clinical illness directly following infection is classified as *primary TB* and is common among children in the first few years of life and among immunocompromised persons. Although primary TB may be severe and disseminated, it generally is not associated with high-level transmissibility. When infection is acquired later in life, the chance is greater that the mature immune system will contain it at least temporarily. Bacilli, however, may persist for years before reactivating to produce *secondary* (or *postprimary*) *TB*, which, because of frequent cavitation, is more often infectious than is primary disease. Overall, it is estimated that up to 10% of infected persons will eventually develop active TB in their lifetime—half of them during the first 18 months after infection. The risk is much higher among HIV-infected persons. Reinfection of a previously infected individual, which is common in areas with high rates of TB transmission, may also favor the development of disease. At the height of the TB resurgence in the United States in the early 1990s, molecular typing and comparison of strains of *M. tuberculosis* suggested that up to one-third of cases of active TB in some inner-city communities were due to recent transmission rather than to reactivation of old latent infection. Age is an important determinant of the risk of disease after infection. Among infected persons, the incidence of TB is highest during late adolescence and early adulthood; the reasons are unclear. The incidence among women peaks at 25–34 years of age. In this age group, rates among women may be higher than those among men, whereas at older ages the opposite is true. The risk increases in the elderly, possibly because of waning immunity and comorbidity.

A variety of diseases and conditions favor the development of active TB **(Table 173-1)**. In absolute terms, the most potent risk factor for TB among infected individuals is clearly HIV co-infection, which suppresses cellular immunity. The risk that LTBI will proceed to active disease is directly related to the patient's degree of immunosuppression. In a study of HIV-infected, tuberculin skin test (TST)–positive persons, this risk varied from 2.6 to 13.3 cases/100 person-years and increased as the CD4+ T cell count decreased.

TABLE 173-1

Risk Factors for Active Tuberculosis in Persons Who Have Been Infected with Tubercle Bacilli

Factor	Relative Risk/Odds[a]
Recent infection (<1 year)	12.9
Fibrotic lesions (spontaneously healed)	2–20
Comorbidities and iatrogenic causes	
HIV infection	21–>30
Silicosis	30
Chronic renal failure/hemodialysis	10–25
Diabetes	2–4
IV drug use	10–30
Excessive alcohol use	3
Immunosuppressive treatment	10
Tumor necrosis factor α inhibitors	4–5
Gastrectomy	2–5
Jejunoileal bypass	30–60
Post-transplantation period (renal, cardiac)	20–70
Tobacco smoking	2–3
Malnutrition and severe underweight	2

[a]Old infection = 1.

NATURAL HISTORY OF DISEASE

Studies conducted in various countries before the advent of chemotherapy showed that untreated TB is often fatal. About one-third of patients died within 1 year after diagnosis, and >50% died within 5 years. The 5-year mortality rate among sputum smear–positive cases was 65%. Of the survivors at 5 years, ~60% had undergone spontaneous remission, while the remainder were still excreting tubercle bacilli. With effective, timely, and proper chemotherapy, patients have a very high chance of being cured. However, improper use of anti-TB drugs, while reducing mortality rates, may also result in large numbers of chronic infectious cases, often with drug-resistant bacilli.

PATHOGENESIS AND IMMUNITY

INFECTION AND MACROPHAGE INVASION

The interaction of *M. tuberculosis* with the human host begins when droplet nuclei containing viable microorganisms, propelled into the air by infectious patients, are inhaled by a close bystander. Although the majority of inhaled bacilli are trapped in the upper airways and expelled by ciliated mucosal cells, a fraction (usually <10%) reach the alveoli, a unique immunoregulatory environment. There, alveolar macrophages that have not yet been activated (prototypic alternatively activated macrophages) phagocytose the bacilli. Adhesion of mycobacteria to macrophages results largely from binding of the bacterial cell wall to a variety of macrophage cell-surface molecules, including complement receptors, the mannose receptor, the immunoglobulin G Fcγ receptor, and type A scavenger receptors. Surfactants may also play a role in the early phase of interaction between the host and the pathogen, and surfactant protein D can prevent phagocytosis. Phagocytosis is enhanced by complement activation leading to opsonization of bacilli with C3 activation

products such as C3b and C3bi. (Bacilli are resistant to complement-mediated lysis.) Binding of certain receptors, such as the mannose receptor, regulates postphagocytic events such as phagosome–lysosome fusion and inflammatory cytokine production. After a phagosome forms, the survival of *M. tuberculosis* within it seems to depend in part on reduced acidification due to lack of assembly of a complete vesicular proton-adenosine triphosphatase. A complex series of events is generated by the bacterial cell-wall lipoglycan lipoarabinomannan, which inhibits the intracellular increase of Ca^{2+}. Thus, the Ca^{2+}/calmodulin pathway (leading to phagosome–lysosome fusion) is impaired, and the bacilli survive within the phagosomes by blocking fusion. The *M. tuberculosis* phagosome inhibits the production of phosphatidylinositol 3-phosphate, which normally earmarks phagosomes for membrane sorting and maturation (including phagolysosome formation), which would destroy the bacteria. Bacterial factors block the host defense of autophagy, in which the cell sequesters the phagosome in a double-membrane vesicle (*autophagosome*) that is destined to fuse with lysosomes. If the bacilli are successful in arresting phagosome maturation, then replication begins and the macrophage eventually ruptures and releases its bacillary contents. This process is mediated by the ESX-1 secretion system that is encoded by genes contained in the region of difference 1 (RD1). Other uninfected phagocytic cells are then recruited to continue the infection cycle by ingesting dying macrophages and their bacillary content, thus, in turn, becoming infected themselves and expanding the infection.

VIRULENCE OF TUBERCLE BACILLI



M. tuberculosis must be viewed as a complex formed by a multitude of strains that differ in virulence and are capable of producing a variety of manifestations of disease. Since the elucidation of the *M. tuberculosis* genome in 1998, large mutant collections have been generated, and many bacterial genes that contribute to *M. tuberculosis* virulence have been found. Different patterns of virulence defects have been defined in various animal models—predominantly mice but also guinea pigs, rabbits, and nonhuman primates. The *katG* gene encodes for a catalase/peroxidase enzyme that protects against oxidative stress and is required for isoniazid activation and subsequent bactericidal activity. RD1 is a 9.5-kb locus that encodes two key small protein antigens—6-kDa early secretory antigen (ESAT-6) and culture filtrate protein-10 (CFP-10)—as well as a putative secretion apparatus that may facilitate their egress; the absence of this locus in the vaccine strain *M. bovis* bacille Calmette-Guérin (BCG) is a key attenuating mutation. In *M. marinum*, a mutation in the RD1 virulence locus encoding the ESX-1 secretion system impairs the capacity of apoptotic macrophages to recruit uninfected cells for further rounds of infection. The results are less replication and fewer new granulomas. These observations in *M. marinum* are similar in part to events related to the virulence of *M. tuberculosis*; however, ESX-1, although necessary, is probably insufficient to explain virulence, and other mechanisms may be in play. Mutants lacking key enzymes of bacterial biosynthesis become auxotrophic for the missing substrate and often are totally unable to proliferate in animals; these include the *leuCD* and *panCD* mutants, which require leucine and pantothenic acid, respectively. The isocitrate lyase gene (*icl1*) encodes a key step in the glyoxylate shunt that facilitates bacterial growth on fatty acid substrates; this gene is required for long-term persistence of *M. tuberculosis* infection in mice with chronic TB. *M. tuberculosis* mutants in regulatory genes such as sigma factor C and sigma factor H (*sigC* and *sigH*) are associated with normal bacterial growth in mice, but they fail to elicit full tissue pathology. Finally, the mycobacterial protein CarD (expressed by the *carD* gene) seems essential for the control of rRNA transcription that is required for mycobacterial replication and persistence in the host cell. Its loss exposes mycobacteria to oxidative stress, starvation, DNA damage, and ultimately sensitivity to killing by a variety of host mutagens and defense mechanisms.

INNATE RESISTANCE TO INFECTION



Several observations suggest that genetic factors play a key role in innate nonimmune resistance to infection with *M. tuberculosis* and the development of disease. The existence of this resistance, which is polygenic in nature, is suggested by the differing degrees of susceptibility to TB in different populations. This mechanism of elimination of the pathogen may be accompanied by negative results in the TST and interferon-γ (IFN-γ) release assays (IGRAs). In mice, a gene called *Nramp1* (natural resistance–associated macrophage protein 1) plays a regulatory role in resistance/susceptibility to mycobacteria. The human homologue NRAMP1, which maps to chromosome 2q, may play a role in determining susceptibility to TB, as is suggested by a study among West Africans. Studies of mouse genetics identified a novel host resistance gene, *ipr1*, that is encoded within the *sst1* locus; *ipr1* encodes an IFN-inducible nuclear protein that interacts with other nuclear proteins in macrophages primed with IFNs or infected by *M. tuberculosis*. In addition, polymorphisms in multiple genes, such as those encoding for various major histocompatibility complex alleles, IFN-γ, T cell growth factor β, interleukin (IL) 10, mannose-binding protein, IFN-γ receptor, Toll-like receptor 2, vitamin D receptor, and IL-1, have been associated with susceptibility to TB.

THE HOST RESPONSE, GRANULOMA FORMATION, AND "LATENCY"

In the initial stage of host–bacterium interaction, prior to the onset of an acquired cell-mediated immune (CMI) response, *M. tuberculosis* disseminates widely through the lymph vessels, spreading to other sites in the lungs and other organs, and undergoes a period of extensive growth within naïve unactivated macrophages; additional naïve macrophages are recruited to the early granuloma. How the bacillus accesses the parenchymal tissue remains to be elucidated: it may directly infect epithelial cells or transmigrate through infected macrophages across the epithelium. Infected dendritic cells or monocytes then begin to transport bacilli to the lymphatic system. Studies suggest that *M. tuberculosis* uses specific virulence mechanisms to subvert host cellular signaling and to elicit an early regulated proinflammatory response that promotes granuloma expansion and bacterial growth during this key early phase. A study of *M. marinum* infection in zebrafish has delineated one molecular mechanism by which mycobacteria induce granuloma formation. The mycobacterial protein ESAT-6 induces secretion of matrix metalloproteinase 9 (MMP9) by nearby epithelial cells that are in contact with infected macrophages. MMP9 in turn stimulates recruitment of naïve macrophages, thus inducing granuloma maturation and bacterial growth. Disruption of MMP9 function results in reduced bacterial growth. Another study has shown that *M. tuberculosis*–derived cyclic AMP is secreted from the phagosome into host

macrophages, subverting the cell's signal transduction pathways and stimulating an elevation in the secretion of tumor necrosis factor α (TNF-α) as well as further proinflammatory cell recruitment. Ultimately, the chemoattractants and bacterial products released during the repeated rounds of cell lysis and infection of newly arriving macrophages enable dendritic cells to access bacilli; these cells migrate to the draining lymph nodes and present mycobacterial antigens to T lymphocytes. At this point, the development of cell-mediated and humoral immunity begins. These initial stages of infection are usually asymptomatic.

About 2–4 weeks after infection, two host responses to *M. tuberculosis* develop: a macrophage-activating CMI response and a tissue-damaging response. The *macrophage-activating response* is a T cell–mediated phenomenon resulting in the activation of macrophages that are capable of killing and digesting tubercle bacilli. The *tissue-damaging response* is the result of a delayed-type hypersensitivity reaction to various bacillary antigens; it destroys unactivated macrophages that contain multiplying bacilli but also causes caseous necrosis of the involved tissues (see below). Although both of these responses can inhibit mycobacterial growth, it is the balance between the two that determines the forms of TB that will develop subsequently. With the development of specific immunity and the accumulation of large numbers of activated macrophages at the site of the primary lesion, granulomatous lesions (*tubercles*) are formed. These lesions consist of accumulations of lymphocytes and activated macrophages that evolve toward epithelioid and giant cell morphologies. Initially, the tissue-damaging response can limit mycobacterial growth within macrophages. As stated above, this response, mediated by various bacterial products, not only destroys macrophages but also produces early solid necrosis in the center of the tubercle. Although *M. tuberculosis* can survive, its growth is inhibited within this necrotic environment by low oxygen tension and low pH. At this point, some lesions may heal by fibrosis, with subsequent calcification, whereas inflammation and necrosis occur in other lesions. Some observations have challenged the traditional view that any encounter between mycobacteria and macrophages results in chronic infection. It is possible that an immune response capable of eradicating early infection may sometimes develop as a consequence, for instance, of disabling mutations in mycobacterial genomes rendering their replication ineffective. Individual granulomas that are formed during this phase of infection can vary in size and cell composition; some can contain the spread of mycobacteria, while others cannot. LTBI ensues as a result of this dynamic balance between the microorganism and the host. It has been speculated that *latency* may not be an accurate term because bacilli may remain active during this "latent" stage, forming biofilms in necrotic areas within which they temporarily hide. Thus, some have proposed the term *persister* as more accurate to indicate the behavior of the bacilli in this phase. It is important to recognize that latent infection and disease represent not a binary state but rather a continuum along which infection will eventually move in the direction of full containment or disease. The ability to predict, through systemic biomarkers, which affected individuals will progress toward disease would be of immense value in devising prophylactic interventions at scale.

MACROPHAGE-ACTIVATING RESPONSE

Cell-mediated immunity is critical at this early stage. In the majority of infected individuals, local macrophages are activated when bacillary antigens processed by macrophages stimulate T lymphocytes to release a variety of lymphokines. These activated macrophages aggregate around the lesion's center and effectively neutralize tubercle bacilli without causing further tissue destruction. In the central part of the lesion, the necrotic material resembles soft cheese (*caseous necrosis*)—a phenomenon that may also be observed in other conditions, such as neoplasms. Even when healing takes place, viable bacilli may remain dormant within macrophages or in the necrotic material for many years. These "healed" lesions in the lung parenchyma and hilar lymph nodes may later undergo calcification.

DELAYED-TYPE HYPERSENSITIVITY

In a minority of cases, the macrophage-activating response is weak, and mycobacterial growth can be inhibited only by intensified delayed hypersensitivity reactions, which lead to lung tissue destruction. The lesion tends to enlarge further, and the surrounding tissue is progressively damaged. At the center of the lesion, the caseous material liquefies. Bronchial walls and blood vessels are invaded and destroyed, and cavities are formed. The liquefied caseous material, containing large numbers of bacilli, is drained through bronchi. Within the cavity, tubercle bacilli multiply, spill into the airways, and are discharged into the environment through expiratory maneuvers such as coughing and talking. In the early stages of infection, bacilli are usually transported by macrophages to regional lymph nodes, from which they gain access to the central venous return; from there they reseed the lungs and may also disseminate beyond the pulmonary vasculature throughout the body via the systemic circulation. The resulting extrapulmonary lesions may undergo the same evolution as those in the lungs, although most tend to heal. In young children with poor natural immunity, hematogenous dissemination may result in fatal miliary TB or tuberculous meningitis.

ROLE OF MACROPHAGES AND MONOCYTES

While cell-mediated immunity confers partial protection against *M. tuberculosis*, humoral immunity plays a less well-defined role in protection (although evidence is accumulating on the existence of antibodies to lipoarabinomannan, which may prevent dissemination of infection in children). In cell-mediated immunity, two types of cells are essential: macrophages, which directly phagocytose tubercle bacilli, and T cells (mainly CD4+ T lymphocytes), which induce protection through the production of cytokines, especially IFN-γ. After infection with *M. tuberculosis*, alveolar macrophages secrete various cytokines responsible for a number of events (e.g., the formation of granulomas) as well as systemic effects (e.g., fever and weight loss). However, alternatively activated alveolar macrophages may be particularly susceptible to *M. tuberculosis* growth early on, given their more limited proinflammatory and bactericidal activity, which is related in part to being bathed in surfactant. New monocytes and macrophages attracted to the site are key components of the immune response. Their primary mechanism is probably related to production of oxidants (such as reactive oxygen intermediates or nitric oxide) that have antimycobacterial activity and increase the synthesis of cytokines such as TNF-α and IL-1, which in turn regulate the release of reactive oxygen intermediates

and reactive nitrogen intermediates. In addition, macrophages can undergo apoptosis—a defensive mechanism to prevent the release of cytokines and bacilli via their sequestration in the apoptotic cell. Recent work also describes the involvement of neutrophils in the host response, although the timing of their appearance and their effectiveness remain uncertain.

ROLE OF T LYMPHOCYTES

Alveolar macrophages, monocytes, and dendritic cells are also critical in processing and presenting antigens to T lymphocytes, primarily CD4+ and CD8+ T cells; the result is the activation and proliferation of CD4+ T lymphocytes, which are crucial to the host's defense against *M. tuberculosis*. Qualitative and quantitative defects of CD4+ T cells explain the inability of HIV-infected individuals to contain mycobacterial proliferation. Activated CD4+ T lymphocytes can differentiate into cytokine-producing T_H1 or T_H2 cells. T_H1 cells produce IFN-γ—an activator of macrophages and monocytes—and IL-2. T_H2 cells produce IL-4, IL-5, IL-10, and IL-13 and may also promote humoral immunity. The interplay of these various cytokines and their cross-regulation determine the host's response. The role of cytokines in promoting intracellular killing of mycobacteria, however, has not been entirely elucidated. IFN-γ may induce the generation of reactive nitrogen intermediates and regulate genes involved in bactericidal effects. TNF-α is also important. Although its precise mechanisms are complex and not yet fully clarified, a model has been suggested that foresees an ideal setting for TNF-α between excessive activation—with consequent worsening of immunopathological reactions—and insufficient activation—with resulting lack of containment—in the control of TB infection. Observations made originally in transgenic knockout mice and more recently in humans suggest that other T cell subsets, especially CD8+ T cells, may play an important role. CD8+ T cells have been associated with protective activities via cytotoxic responses and lysis of infected cells as well as with production of IFN-γ and TNF-α. Finally, natural killer cells act as co-regulators of CD8+ T cell lytic activities, and γδ T cells are increasingly thought to be involved in protective responses in humans.

MYCOBACTERIAL LIPIDS AND PROTEINS

Lipids are involved in mycobacterial recognition by the innate immune system, and lipoproteins (such as 19-kDa lipoprotein) trigger potent signals through Toll-like receptors present in blood dendritic cells. *M. tuberculosis* possesses various protein antigens. Some are present in the cytoplasm and cell wall; others are secreted. That the latter are more important in eliciting a T lymphocyte response is suggested by experiments documenting the appearance of protective immunity in animals after immunization with live, protein-secreting mycobacteria. Among the antigens that may play a protective role are the 30-kDa (or 85B) and ESAT-6 antigens. Protective immunity is probably the result of reactivity to many different mycobacterial antigens. These antigens are being incorporated into newly designed vaccines on various platforms.

SKIN TEST REACTIVITY

Coincident with the appearance of immunity, delayed-type hypersensitivity to *M. tuberculosis* develops. This reactivity is the basis of the TST, which is used primarily for the detection of *M. tuberculosis* infection in persons without symptoms. The cellular mechanisms responsible for TST reactivity are related mainly to previously sensitized CD4+ T lymphocytes, which are attracted to the skin-test site. There, they proliferate and produce cytokines. Although delayed hypersensitivity is associated with protective immunity (TST-positive persons are less susceptible to a new *M. tuberculosis* infection than TST-negative persons), it by no means guarantees protection against reactivation. In fact, cases of active TB are often accompanied by strongly positive skin-test reactions. There is also evidence of reinfection with a new strain of *M. tuberculosis* in patients previously treated for active disease. This evidence underscores the fact that previous latent or active TB may not confer fully protective immunity.

CLINICAL MANIFESTATIONS

TB is classified as pulmonary, extrapulmonary, or both. Depending on several factors linked to different populations and bacterial strains, extrapulmonary TB may occur in 10–40% of patients. Furthermore, up to two-thirds of HIV-infected patients with TB may have both pulmonary and extrapulmonary TB or extrapulmonary TB alone.

PULMONARY TB

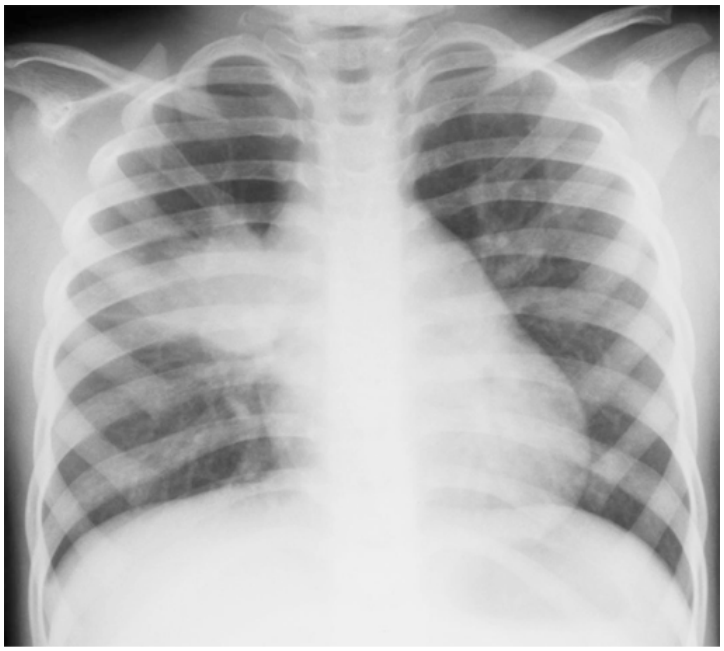
Pulmonary TB is conventionally categorized as primary or postprimary (adult-type, secondary). This distinction has been challenged by molecular evidence from TB-endemic areas indicating that a large percentage of cases of adult pulmonary TB result from recent infection (either primary infection or reinfection) and not from reactivation.

Primary Disease

Primary pulmonary TB occurs soon after the initial infection with tubercle bacilli. It may be asymptomatic or may present with fever and occasionally pleuritic chest pain. In areas of high TB transmission, this form of disease is often seen in children. Because most inspired air is distributed to the middle and lower lung zones, these areas are most commonly involved in primary TB. The lesion forming after initial infection (*Ghon focus*) is usually peripheral and accompanied by transient hilar or paratracheal lymphadenopathy, which may or may not be visible on standard chest radiography (CXR) **(Fig. 173-4)**. Some patients develop erythema nodosum on the legs **(see Fig. A1-39)** or phlyctenular conjunctivitis. In the majority of cases, the lesion heals spontaneously and becomes evident only as a small calcified nodule. Pleural reaction overlying a subpleural focus is also common. The Ghon focus, with or without overlying pleural reaction, thickening, and regional lymphadenopathy, is referred to as the *Ghon complex*.

FIGURE 173-4

Chest radiograph showing right hilar lymph node enlargement with infiltration into the surrounding lung tissue in a child with primary tuberculosis. *(Courtesy of Prof. Robert Gie, Department of Paediatrics and Child Health, Stellenbosch University, South Africa; with permission.)*

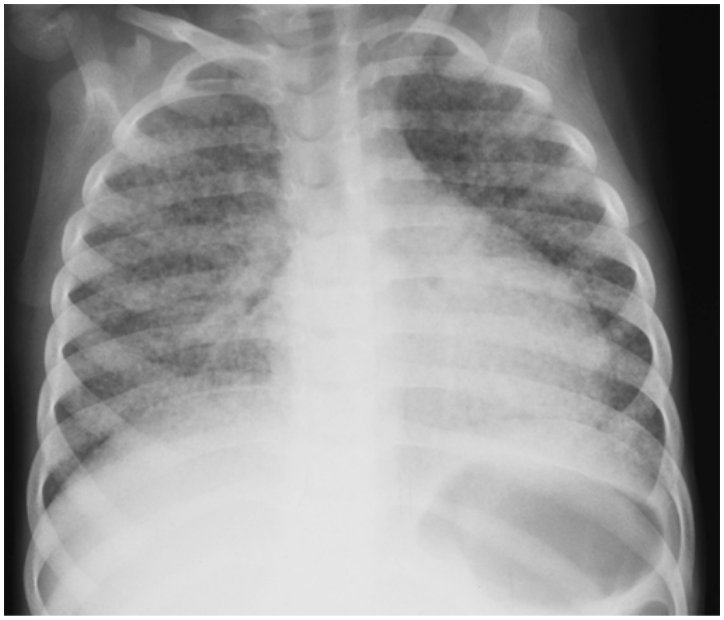


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In young children with immature cell-mediated immunity and in persons with impaired immunity (e.g., those with malnutrition or HIV infection), primary pulmonary TB may progress rapidly to clinical illness. The initial lesion increases in size and can evolve in different ways. Pleural effusion, which is found in up to two-thirds of cases, results from the penetration of bacilli into the pleural space from an adjacent subpleural focus. In severe cases, the primary site rapidly enlarges, its central portion undergoes necrosis, and cavitation develops (*progressive primary TB*). TB in young children is almost invariably accompanied by hilar or paratracheal lymphadenopathy due to the spread of bacilli from the lung parenchyma through lymphatic vessels. Enlarged lymph nodes may compress bronchi, causing total obstruction with distal collapse, partial obstruction with large-airway wheezing, or a ball-valve effect with segmental/lobar hyperinflation. Lymph nodes may also rupture into the airway with development of pneumonia, often including areas of necrosis and cavitation, distal to the obstruction. Bronchiectasis **(Chap. 284)** may develop in any segment/lobe damaged by progressive caseating pneumonia. Occult hematogenous dissemination commonly follows primary infection. However, in the absence of a sufficient acquired immune response, which usually contains the infection, disseminated or miliary disease may result **(Fig. 173-5)**. Small granulomatous lesions develop in multiple organs and may cause locally progressive disease or result in tuberculous meningitis; this is a particular concern in very young children and immunocompromised persons (e.g., patients with HIV infection).

FIGURE 173-5

Chest radiograph showing bilateral miliary (millet-sized) infiltrates in a child. *(Courtesy of Prof. Robert Gie, Department of Paediatrics and Child Health, Stellenbosch University, South Africa; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Postprimary (Adult-Type) Disease

Also referred to as *reactivation* or *secondary TB*, postprimary TB is probably most accurately termed *adult-type TB* because it may result from endogenous reactivation of distant LTBI or recent infection (primary infection or reinfection). It is usually localized to the apical and posterior segments of the upper lobes, where the substantially higher mean oxygen tension (compared with that in the lower zones) favors mycobacterial growth. The superior segments of the lower lobes are also more frequently involved. The extent of lung parenchymal involvement varies greatly, from small infiltrates to extensive cavitary disease. With cavity formation, liquefied necrotic contents are ultimately discharged into the airways and may undergo bronchogenic spread, resulting in satellite lesions within the lungs that may in turn undergo cavitation **(Figs. 173-6 and 173-7)**. Massive involvement of pulmonary segments or lobes, with coalescence of lesions, produces caseating pneumonia. While up to one-third of untreated patients reportedly succumb to severe pulmonary TB within a few months after onset (the classic "galloping consumption" of the past), others may undergo a process of spontaneous remission or proceed along a chronic, progressively debilitating course ("consumption" or *phthisis*). Under these circumstances, some pulmonary lesions become fibrotic and may later calcify, but cavities persist in other parts of the lungs. Individuals with such chronic disease continue to discharge tubercle bacilli into the environment. Most patients respond to treatment, with defervescence, decreasing cough, weight gain, and a general improvement in well-being within several weeks.

FIGURE 173-6

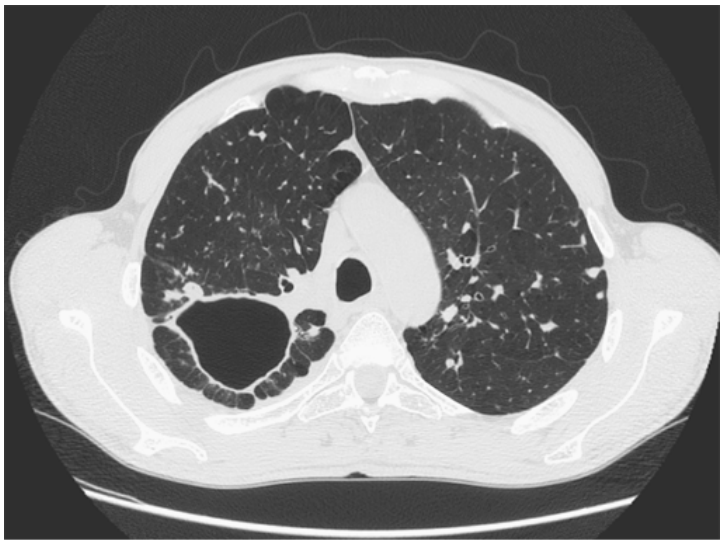
Chest radiograph showing a right-upper-lobe infiltrate and a cavity with an air-fluid level in a patient with active tuberculosis. *(Courtesy of Dr. Andrea Gori, Department of Infectious Diseases, S. Paolo University Hospital, Milan, Italy; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 173-7

CT scan showing a large cavity in the right lung of a patient with active tuberculosis. *(Courtesy of Dr. Elisa Busi Rizzi, National Institute for Infectious Diseases, Spallanzani Hospital, Rome, Italy; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Early in the course of disease, symptoms and signs are often nonspecific and insidious, consisting mainly of diurnal fever and night sweats due to defervescence, weight loss, anorexia, general malaise, and weakness. However, in up to 90% of cases, cough eventually develops—often initially nonproductive and limited to the morning and subsequently accompanied by the production of purulent sputum, sometimes with blood streaking. Hemoptysis develops in 20–30% of cases, and massive hemoptysis may ensue as a consequence of the erosion of a blood vessel in the wall of a cavity. Hemoptysis, however, may also result from rupture of a dilated vessel in a cavity (*Rasmussen's aneurysm*) or from aspergilloma formation in an old cavity. Pleuritic chest pain sometimes develops in patients with subpleural parenchymal lesions or pleural disease. Extensive disease may produce dyspnea and, in rare instances, adult respiratory distress syndrome. Physical findings are of limited use in pulmonary TB. Many patients have no abnormalities detectable by chest examination, whereas others have detectable rales in the involved areas during inspiration, especially after coughing. Occasionally, rhonchi due to partial bronchial obstruction and classic amphoric breath sounds in areas with large cavities may be heard. Systemic features include fever (often low-grade and intermittent) in up to 80% of cases and wasting. Absence of fever, however, does not exclude TB. In some cases, pallor and finger clubbing develop. The most common hematologic findings are mild anemia, leukocytosis, and thrombocytosis with a slightly elevated erythrocyte sedimentation rate and/or C-

reactive protein level. None of these findings is consistent or sufficiently accurate for diagnostic purposes. Hyponatremia due to the syndrome of inappropriate secretion of antidiuretic hormone has also been reported.

EXTRAPULMONARY TB

In descending order of frequency, the extrapulmonary sites most commonly involved in TB are the lymph nodes, pleura, genitourinary tract, bones and joints, meninges, peritoneum, and pericardium. However, virtually any organ system may be affected. As a result of hematogenous dissemination in HIV-infected individuals, extrapulmonary TB is seen more commonly today than in the past in settings of high HIV prevalence.

Lymph Node TB (Tuberculous Lymphadenitis)

The most common presentation of extrapulmonary TB in both HIV-seronegative individuals and HIV-infected patients (35% of cases worldwide and >40% of cases in the United States in recent series), lymph node disease is particularly frequent among HIV-infected patients and among children **(Fig. 173-8)**. In the United States, besides children, women (particularly non-Caucasians) seem to be especially susceptible. Once caused mainly by *M. bovis*, tuberculous lymphadenitis today is due largely to *M. tuberculosis*. Lymph node TB presents as painless swelling of the lymph nodes, most commonly at posterior cervical and supraclavicular sites (a condition historically referred to as *scrofula*). Lymph nodes are usually discrete in early disease but develop into a matted nontender mass over time; a fistulous tract draining caseous material may result. Associated pulmonary disease is present in fewer than 50% of cases, and systemic symptoms are uncommon except in HIV-infected patients. The diagnosis is established by fine-needle aspiration biopsy (with a yield of up to 80%) or surgical excision biopsy. Bacteriologic confirmation is achieved in the vast majority of cases, granulomatous lesions with or without visible AFBs are typically seen, and cultures are positive in 70–80% of cases. Among HIV-infected patients, granulomas are less well organized and are frequently absent entirely, but bacterial loads are heavier than in HIV-seronegative patients, with higher yields from microscopy and culture. Differential diagnosis includes a variety of infectious conditions, neoplastic diseases such as lymphomas or metastatic carcinomas, and rare disorders like Kikuchi's disease (necrotizing histiocytic lymphadenitis), Kimura's disease, and Castleman's disease.

FIGURE 173-8

Tuberculous lymphadenitis affecting the cervical lymph nodes in a 2-year-old child from Malawi. *(Courtesy of Prof. S. Graham, Centre for International Child Health, University of Melbourne, Australia; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pleural TB

Involvement of the pleura accounts for ~20% of extrapulmonary cases in the United States and elsewhere. Isolated pleural effusion usually reflects recent primary infection, and the collection of fluid in the pleural space represents a hypersensitivity response to mycobacterial antigens. Pleural disease may also result from contiguous parenchymal spread, as in many cases of pleurisy accompanying postprimary disease. Depending on the extent of reactivity, the effusion may be small, remain unnoticed, and resolve spontaneously or may be sufficiently large to cause symptoms such as fever, pleuritic chest pain, and dyspnea. Physical findings are those of pleural effusion: dullness to percussion and absence of breath sounds. CXR reveals the effusion and, in up to one-third of cases, also shows a parenchymal lesion. Thoracentesis is required to ascertain the nature of the effusion and to differentiate it from manifestations of other etiologies. The fluid is straw-colored and at times hemorrhagic; it is an exudate with a protein concentration >50% of that in serum (usually ~4–6 g/dL), a normal to low glucose concentration, a pH of ~7.3 (occasionally <7.2), and detectable white blood cells (usually 500–6000/μL). Neutrophils may predominate in the early stage, but lymphocyte predominance is the typical finding later. Mesothelial cells are generally rare or absent. AFBs are rarely seen on direct smear, and cultures often may be falsely negative for *M. tuberculosis*; positive cultures are more common among postprimary cases. Determination of the pleural concentration of adenosine deaminase may be a useful screening test, and TB may be excluded if the value is very low. Lysozyme is also

present in the pleural effusion. Measurement of IFN-γ, either directly or through stimulation of sensitized T cells with mycobacterial antigens, can be diagnostically helpful. Needle biopsy of the pleura is often required for diagnosis and is recommended over pleural fluid analysis; it reveals granulomas and/or yields a positive culture in up to 80% of cases. Pleural biopsy can yield a positive result in ~75% of cases when real-time automated nucleic acid amplification is used (the Xpert® MTB/RIF assay [Cepheid, Sunnyvale, CA]; see "Nucleic Acid Amplification Technology," below); testing of pleural fluid with this assay is not recommended because of low sensitivity. This form of pleural TB responds rapidly to chemotherapy and may resolve spontaneously. Concurrent glucocorticoid administration may reduce the duration of fever and/or chest pain but is not of proven benefit.

Tuberculous empyema is a less common complication of pulmonary TB. It is usually the result of the rupture of a cavity, with spillage of a large number of organisms into the pleural space. This process may create a bronchopleural fistula with evident air in the pleural space. CXR shows hydropneumothorax with an air-fluid level. The pleural fluid is purulent and thick and contains large numbers of lymphocytes. Acid-fast smears and mycobacterial cultures are often positive. Surgical drainage is usually required as an adjunct to chemotherapy. Tuberculous empyema may result in severe pleural fibrosis and restrictive lung disease. Removal of the thickened visceral pleura (*decortication*) is occasionally necessary to improve lung function.

TB of the Upper Airways

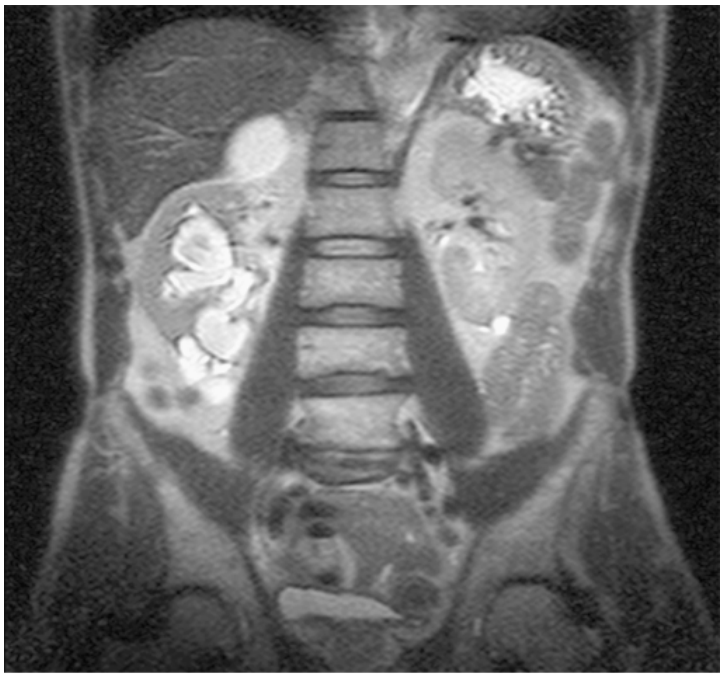
Nearly always a complication of advanced cavitary pulmonary TB, TB of the upper airways may involve the larynx, pharynx, and epiglottis. Symptoms include hoarseness, dysphonia, and dysphagia in addition to chronic productive cough. Findings depend on the site of involvement, and ulcerations may be seen on laryngoscopy. Acid-fast smear of the sputum is often positive, but biopsy may be necessary in some cases to establish the diagnosis. Carcinoma of the larynx may have similar features but is usually painless.

Genitourinary TB

Genitourinary TB, which accounts for ~10–15% of all extrapulmonary cases in the United States and elsewhere, may involve any portion of the genitourinary tract. Local symptoms predominate, and up to 75% of patients have abnormalities on CXR suggesting previous or concomitant pulmonary disease. Urinary frequency, dysuria, nocturia, hematuria, and flank or abdominal pain are common presentations. However, patients may be asymptomatic and their disease discovered only after severe destructive lesions of the kidneys have developed. Urinalysis gives abnormal results in 90% of cases, revealing pyuria and hematuria. The documentation of culture-negative pyuria in acidic urine should raise the suspicion of TB. IV pyelography, abdominal CT, or MRI **(Fig. 173-9)** may show deformities and obstructions; calcifications and ureteral strictures are suggestive findings. Culture of three morning urine specimens yields a definitive diagnosis in nearly 90% of cases. Severe ureteral strictures may lead to hydronephrosis and renal damage. Genital TB is diagnosed more commonly in female than in male patients. In female patients, it affects the fallopian tubes and the endometrium and may cause infertility, pelvic pain, and menstrual abnormalities. Diagnosis requires biopsy or culture of specimens obtained by dilation and curettage. In male patients, genital TB preferentially affects the epididymis, producing a slightly tender mass that may drain externally through a fistulous tract; orchitis and prostatitis may also develop. In almost half of cases of genitourinary TB, urinary tract disease is also present. Genitourinary TB responds well to chemotherapy.

FIGURE 173-9

MRI of culture-confirmed renal tuberculosis. T2-weighted coronary plane: coronal sections showing several renal lesions in both the cortical and the medullary tissues of the right kidney. *(Courtesy of Dr. Alberto Matteelli, Department of Infectious Diseases, University of Brescia, Italy; with permission.)*



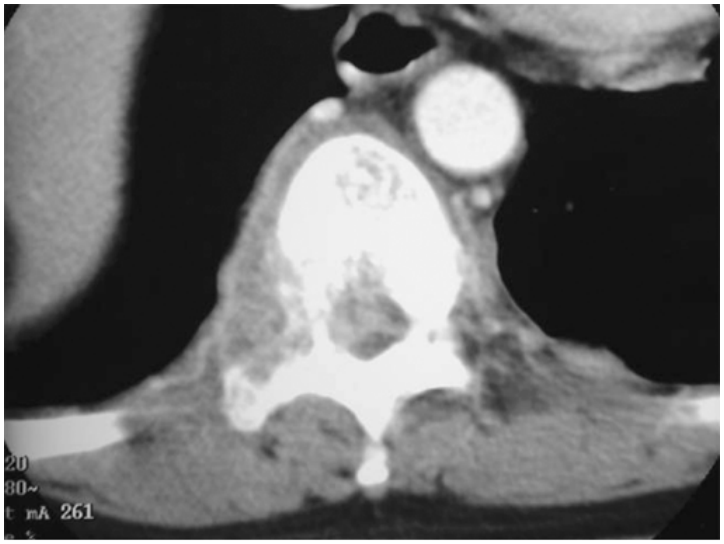
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Skeletal TB

In the United States, TB of the bones and joints is responsible for ~10% of extrapulmonary cases. In bone and joint disease, pathogenesis is related to reactivation of hematogenous foci or to spread from adjacent paravertebral lymph nodes. Weight-bearing joints (the spine in 40% of cases, the hips in 13%, and the knees in 10%) are most commonly affected. Spinal TB (Pott's disease or tuberculous spondylitis; **Fig. 173-10**) often involves two or more adjacent vertebral bodies. Whereas the upper thoracic spine is the most common site of spinal TB in children, the lower thoracic and upper lumbar vertebrae are usually affected in adults. From the anterior superior or inferior angle of the vertebral body, the lesion slowly reaches the adjacent body, later affecting the intervertebral disk. With advanced disease, collapse of vertebral bodies results in kyphosis (*gibbus*). A paravertebral "cold" abscess may also form. In the upper spine, this abscess may track to and penetrate the chest wall, presenting as a soft tissue mass; in the lower spine, it may reach the inguinal ligaments or present as a psoas abscess. CT or MRI reveals the characteristic lesion and suggests its etiology. The differential diagnosis includes tumors and other infections. Pyogenic bacterial osteomyelitis, in particular, involves the disk very early and produces rapid sclerosis. Aspiration of the abscess or bone biopsy confirms the tuberculous etiology, as cultures are usually positive and histologic findings highly typical. A catastrophic complication of Pott's disease is paraplegia, which is usually due to an abscess or a lesion compressing the spinal cord. Paraparesis due to a large abscess is a medical emergency and requires rapid drainage. TB of the hip joints, usually involving the head of the femur, causes pain; TB of the knee produces pain and swelling. If the disease goes unrecognized, the joints may be destroyed. Diagnosis requires examination of the synovial fluid, which is thick in appearance, with a high protein concentration and a variable cell count. Although synovial fluid culture is positive in a high percentage of cases, synovial biopsy and tissue culture may be necessary to establish the diagnosis. Skeletal TB responds to chemotherapy, but severe cases may require surgery.

FIGURE 173-10

CT scan demonstrating destruction of the right pedicle of T10 due to Pott's disease. The patient, a 70-year-old Asian woman, presented with back pain and weight loss and had biopsy-proven tuberculosis. *(Courtesy of Charles L. Daley, MD, University of California, San Francisco; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Tuberculous Meningitis and Tuberculoma

TB of the central nervous system accounts for ~5% of extrapulmonary cases in the United States. It is seen most often in young children but also develops in adults, especially those infected with HIV. Tuberculous meningitis results from the hematogenous spread of primary or postprimary pulmonary TB or from the rupture of a subependymal tubercle into the subarachnoid space. In more than half of cases, evidence of old pulmonary lesions or a miliary pattern is found on CXR. The disease often presents subtly as headache and slight mental changes after a prodrome of weeks of low-grade fever, malaise, anorexia, and irritability. If not recognized, tuberculous meningitis may evolve acutely with severe headache, confusion, lethargy, altered sensorium, and neck rigidity. Typically, the disease evolves over 1–2 weeks, a course longer than that of bacterial meningitis. Because meningeal involvement is pronounced at the base of the brain, paresis of cranial nerves (ocular nerves in particular) is a frequent finding, and the involvement of cerebral arteries may produce focal ischemia. The ultimate evolution is toward coma, with hydrocephalus and intracranial hypertension.

Lumbar puncture is the cornerstone of diagnosis. In general, examination of cerebrospinal fluid (CSF) reveals a high leukocyte count (up to 1000/μL), usually with a predominance of lymphocytes but sometimes with a predominance of neutrophils in the early stage; a protein content of 1–8 g/L (100–800 mg/dL); and a low glucose concentration. However, any of these three parameters can be within the normal range. AFBs are infrequently seen on direct smear of CSF sediment, and repeated lumbar punctures increase the yield. Culture of CSF is diagnostic in up to 80% of cases and remains the gold standard. Real-time automated nucleic acid amplification (the Xpert MTB/RIF assay) has a sensitivity of up to 80% and is the preferred initial diagnostic option. Treatment should be initiated immediately upon a positive Xpert MTB/RIF result. A negative result does not exclude a diagnosis of TB and requires further diagnostic workup. Imaging studies (CT and MRI) may show hydrocephalus and abnormal enhancement of basal cisterns or ependyma. If unrecognized, tuberculous meningitis is uniformly fatal. This disease responds to chemotherapy; however, neurologic sequelae are documented in 25% of treated cases, in most of which the

diagnosis has been delayed. Clinical trials have demonstrated that patients given adjunctive glucocorticoids may experience faster resolution of CSF abnormalities and elevated CSF pressure, resulting in lower rates of death or severe disability and relapse. In one study, adjunctive dexamethasone significantly enhanced the chances of survival among persons >14 years of age but did not reduce the frequency of neurologic sequelae. The dexamethasone schedule was (1) 0.4 mg/kg per day given IV with tapering by 0.1 mg/kg per week until the fourth week, when 0.1 mg/kg per day was administered; followed by (2) 4 mg/d given by mouth with tapering by 1 mg per week until the fourth week, when 1 mg/d was administered. The WHO now recommends that adjuvant glucocorticoid therapy with either dexamethasone or prednisolone, tapered over 6–8 weeks, should be used in central nervous system TB.

Tuberculoma, an uncommon manifestation of TB of the central nervous system, presents as one or more space-occupying lesions and usually causes seizures and focal signs. CT or MRI reveals contrast-enhanced ring lesions, but biopsy is necessary to establish the diagnosis.

Gastrointestinal TB

Gastrointestinal TB is uncommon, making up only 3.5% of extrapulmonary cases in the United States. Various pathogenetic mechanisms are involved: swallowing of sputum with direct seeding, hematogenous spread, or (largely in developing areas) ingestion of milk from cows affected by bovine TB. Although any portion of the gastrointestinal tract may be affected, the terminal ileum and the cecum are the sites most commonly involved. Abdominal pain (at times similar to that associated with appendicitis) and swelling, obstruction, hematochezia, and a palpable mass in the abdomen are common findings at presentation. Fever, weight loss, anorexia, and night sweats are also common. With intestinal-wall involvement, ulcerations and fistulae may simulate Crohn's disease; the differential diagnosis of this entity is always difficult. Anal fistulae should prompt an evaluation for rectal TB. Because surgery is required in most cases, the diagnosis can be established by histologic examination and culture of specimens obtained intraoperatively.

Tuberculous peritonitis follows either the direct spread of tubercle bacilli from ruptured lymph nodes and intraabdominal organs (e.g., genital TB in women) or hematogenous seeding. Nonspecific abdominal pain, fever, and ascites should raise the suspicion of tuberculous peritonitis. The coexistence of cirrhosis (Chap. 335) in patients with tuberculous peritonitis complicates the diagnosis. In tuberculous peritonitis, paracentesis reveals an exudative fluid with a high protein content and leukocytosis that is usually lymphocytic (although neutrophils occasionally predominate). The yield of direct smear and culture is relatively low; culture of a large volume of ascitic fluid can increase the yield, but peritoneal biopsy (with a specimen best obtained by laparoscopy) is often needed to establish the diagnosis.

Pericardial TB (Tuberculous Pericarditis)

Due either to direct extension from adjacent mediastinal or hilar lymph nodes or to hematogenous spread, pericardial TB has often been a disease of the elderly in countries with low TB prevalence. However, it also develops frequently in HIV-infected patients. Case–fatality rates are as high as 40% in some series. The onset may be subacute, although an acute presentation, with dyspnea, fever, dull retrosternal pain, and a pericardial friction rub, is possible. An effusion eventually develops in many cases; cardiovascular symptoms and signs of cardiac tamponade may ultimately appear (Chap. 265). In the presence of effusion, TB must be suspected if the patient belongs to a high-risk population (HIV-infected, originating in a high-prevalence country); if there is evidence of previous TB in other organs; or if echocardiography, CT, or MRI shows effusion and thickness across the pericardial space. A definitive diagnosis can be obtained by pericardiocentesis under echocardiographic guidance. The pericardial fluid must be submitted for biochemical, cytologic, and microbiologic evaluation. The effusion is exudative in nature, with a high count of lymphocytes and monocytes. Hemorrhagic effusion is common. Direct smear examination is very rarely positive. Culture of pericardial fluid reveals *M. tuberculosis* in up to two-thirds of cases, whereas pericardial biopsy has a higher yield. High levels of adenosine deaminase, lysozyme, and IFN-γ may suggest a tuberculous etiology.

Without treatment, pericardial TB is usually fatal. Even with treatment, complications may develop, including chronic constrictive pericarditis with thickening of the pericardium, fibrosis, and sometimes calcification, which may be visible on a chest radiograph. Systematic reviews and meta-analyses show a trend toward benefit from glucocorticoid treatment with regard to death and constrictive pericarditis. However, the largest and most recent study—the IMPI study—failed to show such a benefit. Of the patients enrolled in this trial, 67% were infected with HIV, and only a fraction were receiving antiretroviral treatment (ART). A supplemental analysis among HIV-negative people showed a small mortality benefit, as did another small study among HIV-infected people. The WHO currently recommends that, in patients with tuberculous pericarditis, initial adjuvant glucocorticoid therapy may be used. The 2016 guidelines of the American Thoracic Society (ATS), the CDC, and the Infectious Diseases Society of America (IDSA), on the other hand, suggest that glucocorticoid therapy should not be routinely administered.

Caused by direct extension from the pericardium or by retrograde lymphatic extension from affected mediastinal lymph nodes, tuberculous myocarditis is an extremely rare disease. Usually, it is fatal and is diagnosed post-mortem.

Miliary or Disseminated TB

Miliary TB is due to hematogenous spread of tubercle bacilli. Although in children it is often the consequence of primary infection, in adults it may be due to either recent infection or reactivation of old disseminated foci. The lesions are usually yellowish granulomas 1–2 mm in diameter that resemble millet seeds (thus the term *miliary*, coined by nineteenth-century pathologists). Clinical manifestations are nonspecific and protean, depending on the predominant site of involvement. Fever, night sweats, anorexia, weakness, and weight loss are presenting symptoms in the majority of cases. At times, patients have a cough and other respiratory symptoms due to pulmonary involvement as well as abdominal symptoms. Physical findings include hepatomegaly, splenomegaly,

and lymphadenopathy. Eye examination may reveal choroidal tubercles, which are pathognomonic of miliary TB, in up to 30% of cases. Meningismus occurs in fewer than 10% of cases.

A high index of suspicion is required for the diagnosis of miliary TB. Frequently, CXR (Fig. 173-5) reveals a miliary reticulonodular pattern (more easily seen on underpenetrated film), although no radiographic abnormality may be evident early in the course and among HIV-infected patients. Other radiologic findings include large infiltrates, interstitial infiltrates (especially in HIV-infected patients), and pleural effusion. Sputum-smear microscopy is negative in most cases. Various hematologic abnormalities may be seen, including anemia with leukopenia, lymphopenia, neutrophilic leukocytosis and leukemoid reactions, and polycythemia. Disseminated intravascular coagulation has been reported. Elevation of alkaline phosphatase levels and other abnormal values in liver function tests are detected in patients with severe hepatic involvement. TST results may be negative in up to half of cases, but reactivity may be restored during chemotherapy. Bronchoalveolar lavage and transbronchial biopsy are more likely to provide bacteriologic confirmation, and granulomas are evident in liver or bone-marrow biopsy specimens from many patients. If it goes unrecognized, miliary TB is lethal; with proper early treatment, however, it is amenable to cure. Glucocorticoid therapy has not proved beneficial.

A rare presentation seen in the elderly, *cryptic miliary TB* has a chronic course characterized by mild intermittent fever, anemia, and—ultimately—meningeal involvement preceding death. An acute septicemic form, *nonreactive miliary TB*, occurs very rarely and is due to massive hematogenous dissemination of tubercle bacilli. Pancytopenia is common in this form of disease, which is rapidly fatal. At postmortem examination, multiple necrotic but nongranulomatous ("nonreactive") lesions are detected.

Less Common Extrapulmonary Forms

TB may cause chorioretinitis, uveitis, panophthalmitis, and painful hypersensitivity-related phlyctenular conjunctivitis. Tuberculous otitis is rare and presents as hearing loss, otorrhea, and tympanic membrane perforation. In the nasopharynx, TB may simulate granulomatosis with polyangiitis. Cutaneous manifestations of TB include primary infection due to direct inoculation, abscesses and chronic ulcers, scrofuloderma, lupus vulgaris (a smoldering disease with nodules, plaques, and fissures), miliary lesions, and erythema nodosum. Tuberculous mastitis results from retrograde lymphatic spread, often from the axillary lymph nodes. Adrenal TB is a manifestation of disseminated disease presenting rarely as adrenal insufficiency. Finally, congenital TB results from transplacental spread of tubercle bacilli to the fetus or from ingestion of contaminated amniotic fluid. This rare disease affects the liver, spleen, lymph nodes, and various other organs.

Post-TB Complications

TB may cause persisting pulmonary damage in patients whose infection has been considered cured on clinical grounds. Chronic impairment of lung functions, bronchiectasis, aspergillomas, and chronic pulmonary aspergillosis have been associated with TB. Chronic pulmonary aspergillosis may manifest as simple aspergilloma (fungal ball) or chronic cavitary aspergillosis. Early studies revealed that, especially in the presence of large residual cavities, *Aspergillus fumigatus* may colonize the lesion and produce symptoms such as respiratory impairment, hemoptysis, persistent fatigue, and weight loss, often resulting in the erroneous diagnosis of TB recurrence. The detection of *Aspergillus* precipitins (IgG) in the blood suggests chronic pulmonary aspergillosis, as do radiographic abnormalities such as thickening of the pleura and cavitary walls or the presence of a fungal ball inside the cavity. Treatment is difficult. Recent preliminary studies on the use of itraconazole for ≥6 months indicate improvement or stabilization of 60–75% of the radiologic and clinical manifestations. Surgical removal of lesions is risky except in simple aspergilloma.

HIV-Associated TB

TB is one of the most common diseases among HIV-infected persons worldwide **(See also Chap. 197)**. Responsible for an estimated 20–25% of all HIV-related mortality (some 390,000 deaths per year), TB is likely the main cause of death in this population. In certain urban settings in some African countries, the prevalence of HIV infection among TB patients reaches 70–80%. A person with a positive TST who acquires HIV infection has a 3–13% annual risk of developing active TB, with the exact risk depending on the degree of immunosuppression when observation begins. Furthermore, a new TB infection acquired by an HIV-infected individual may evolve into active disease in a matter of weeks rather than months or years. TB can appear at any stage of HIV infection, and its presentation varies with the stage. When cell-mediated immunity is only partially compromised, pulmonary TB presents in a typical manner (Figs. 173-6 and 173-7), with upper-lobe infiltrates and cavitation and without significant lymphadenopathy or pleural effusion. In late stages of HIV infection, when the CD4+ T cell count is <200/μL, a primary TB–like pattern, with diffuse interstitial and subtle infiltrates, little or no cavitation, pleural effusion, and intrathoracic lymphadenopathy, is more common. However, these forms are becoming less common because of the expanded use of ART. Overall, sputum smears are less frequently positive among TB patients with HIV infection than among those without; thus, the diagnosis of TB with traditional technology may be difficult, especially in view of the variety of HIV-related pulmonary conditions mimicking TB. Extrapulmonary TB is common among HIV-infected patients. In various series, extrapulmonary TB—alone or in association with pulmonary disease—has been documented in 40–60% of all cases in HIV-co-infected individuals. The most common forms are lymphatic, disseminated, pleural, and pericardial. Mycobacteremia and meningitis are also common, particularly in advanced HIV disease. The diagnosis of TB in HIV-infected patients may be complicated not only by the increased frequency of sputum-smear negativity (up to 40% in culture-proven pulmonary cases) but also by atypical radiographic findings, a lack of classic granuloma formation in the late stages, and a negative TST. The Xpert MTB/RIF assay is the preferred initial diagnostic option, and therapy should be started on the basis of a positive result because treatment delays may be fatal. A negative Xpert MTB/RIF result, however, does not exclude a diagnosis of TB. Culture remains the gold standard. Recent assessment of a test based on the detection of mycobacterial lipoarabinomannan antigen in urine has shown favorable results in assisting with the detection of TB in HIV-positive people (see "Additional Diagnostic Procedures," below).

The *immune reconstitution inflammatory syndrome* (IRIS) or *TB immune reconstitution disease* consists of exacerbations in systemic manifestations (lymphadenopathy, fever) or respiratory signs (worsening of pulmonary infiltrations, pleural effusion) as well as laboratory or radiographic manifestations of TB. This syndrome has been associated with the administration of ART and occurs in ~10% of HIV-infected TB patients. Usually developing 1–3 months after initiation of ART, IRIS is more common among patients with advanced immunosuppression and extrapulmonary TB. "Unmasking IRIS" may develop after the initiation of ART in patients with undiagnosed subclinical TB. The earlier ART is started and the lower the baseline CD4+ T cell count, the greater the risk of IRIS. Death due to IRIS is relatively infrequent and occurs mainly among patients who have a high preexisting mortality risk. The presumed pathogenesis of IRIS consists of an immune response that is elicited by antigens released as bacilli are killed during effective chemotherapy and that is temporally associated with improving immune function. There is no diagnostic test for IRIS, and its confirmation relies heavily upon case definitions incorporating clinical and laboratory data; a variety of case definitions have been suggested. The first priority in the management of a possible case of IRIS is to ensure that the clinical syndrome does not represent a failure of TB treatment or the development of another infection. Mild paradoxical reactions can be managed with symptom-based treatment and do not worsen outcomes of treatment for TB. However, IRIS can result in serious neurologic complications or death in patients with central nervous system TB. Therefore, ART should not be initiated during the first 8 weeks of TB treatment in patients with TB meningitis. Glucocorticoids have been used for severe paradoxical reactions; prednisolone given for 4 weeks at a low dosage (1.5 mg/kg per day for 2 weeks and half that dose for the remaining 2 weeks) has reduced the need for hospitalization and therapeutic procedures and has hastened alleviation of symptoms, as reflected by Karnofsky performance scores, quality-of-life assessments, radiographic response, and C-reactive protein levels. The effectiveness of glucocorticoids in alleviating the symptoms of IRIS is probably linked to suppression of proinflammatory cytokine concentrations, as these medications reduce serum concentrations of IL-6, IL-10, IL-12p40, TNF-α, IFN-γ, and IFN-γ-inducible protein 10. Recommendations for the prevention and treatment of TB in HIV-infected individuals are provided below.

DIAGNOSIS

The key to the early diagnosis of TB is a high index of suspicion. Diagnosis is not difficult in persons belonging to high-risk populations who present with typical symptoms and a classic chest radiograph showing upper-lobe infiltrates with cavities (Fig. 173-6). On the other hand, the diagnosis can easily be missed in an elderly nursing-home resident or a teenager with a focal infiltrate. Often, the diagnosis is first entertained when the chest radiograph of a patient being evaluated for respiratory symptoms is abnormal. If the patient has no complicating medical conditions that cause immunosuppression, the chest radiograph may show typical upper-lobe infiltrates with cavitation (Fig. 173-6). The longer the delay between the onset of symptoms and the diagnosis, the more likely is the finding of cavitary disease. In contrast, immunosuppressed patients, including those with HIV infection, may have "atypical" findings on CXR—e.g., lower-zone infiltrates without cavity formation.

The several approaches to the diagnosis of TB require, above all, a well-organized laboratory network with an appropriate distribution of tasks at different levels of the health care system. Besides clinical assessment and radiography, screening and referral are the principal tasks at the peripheral and community levels. Diagnosis at a secondary level (e.g., a traditional district hospital in a high-incidence setting) can be accomplished nowadays through real-time automated nucleic acid amplification technology (e.g., the Xpert MTB/RIF assay, which also allows detection of drug resistance) or through traditional AFB microscopy, where new tools have not yet been introduced. At a tertiary level, additional technology is necessary, including molecular tests, rapid culture, and DST.

NUCLEIC ACID AMPLIFICATION TECHNOLOGY

Several test systems based on amplification of mycobacterial nucleic acid have become available in the past few years and are now the preferred first-line diagnostic tests. These tests are progressively replacing smear microscopy, as they ensure rapid confirmation of all types of TB. One system that permits rapid diagnosis of TB with high specificity and sensitivity (approaching that of liquid culture) is the fully automated, real-time nucleic acid amplification technology known as the Xpert MTB/RIF assay. Xpert MTB/RIF can simultaneously detect TB and rifampin resistance in <2 h and has minimal biosafety and training requirements. Therefore, it can be housed in nonconventional laboratory settings as long as a stable and uninterrupted power supply can be assured. The WHO recommends its use worldwide as the first-line diagnostic test in all adults and children with signs or symptoms of active TB. Given the test's high sensitivity, the WHO also recommends its use as the initial diagnostic test for people living with HIV in whom TB is suspected. Likewise, Xpert MTB/RIF should be the initial test applied to CSF from patients in whom TB meningitis is suspected as well as a replacement test (preferable to conventional microscopy, culture, and histopathology) for selected nonrespiratory specimens—those obtained by gastric lavage, fine-needle aspiration, or pleural or other biopsies—from patients in whom extrapulmonary TB is suspected. This test has a sensitivity of 98% among AFB-positive cases and ~70% among AFB-negative specimens. Recently, the new Xpert® MTB/RIF Ultra assay (Ultra), which uses the same GeneXpert® diagnostic platform, has been assessed by the WHO as non-inferior to the Xpert MTB/RIF assay. Overall, its sensitivity is 5% higher, with the greatest increases among smear-negative, culture-positive cases (+17%) and among HIV-infected persons (+12%). However, because of this greater sensitivity, the new Ultra cartridge also detects nonviable bacilli and consequently has 3.2% lower specificity than the original test. In this new assay, "trace calls" (i.e., the "noise" produced by detection of nonviable bacilli or fragments of bacilli) need to be evaluated according to risk/benefit considerations. For instance, trace calls in specimens from HIV-infected patients, children, and persons with extrapulmonary TB should be considered true positives, given the high risk of severe morbidity and premature death, while among other cases they warrant additional tests to confirm the diagnosis of TB and prevent over-treatment. Among patients with a recent history of TB, trace calls may represent false positivity. Accuracy in detection of rifampin resistance by Ultra is similar to that by the Xpert MTB/RIF assay.

Another recently introduced molecular test for detection of *M. tuberculosis* is based on the loop-mediated isothermal amplification (LAMP) temperature-independent technology that amplifies DNA, is relatively simple to use, and is interpreted through a visual display. The new TB-LAMP assay (LoopampTM *M.*

tuberculosis complex detection kit; Eiken Chemical Company, Japan) requires minimal laboratory infrastructure and has few biosafety requirements. It may be used as a replacement for sputum-smear microscopy for the diagnosis of adult pulmonary TB and as a follow-up test to smear microscopy for the further investigation of smear-negative specimens from adults with suspected pulmonary TB. The TB-LAMP assay should not replace rapid molecular tests that detect both TB and rifampin resistance, and its usefulness in HIV-infected people in whom TB is suspected remains unclear.

AFB MICROSCOPY

In many low- and middle-income settings, a presumptive diagnosis is still commonly based on the finding of AFB on microscopic examination of a diagnostic specimen, such as a smear of expectorated sputum or of tissue (e.g., a lymph node biopsy). Although inexpensive, AFB microscopy has relatively low sensitivity (40–60%) in culture-confirmed cases of pulmonary TB. The traditional method—light microscopy of specimens stained with Ziehl-Neelsen basic fuchsin dyes—is satisfactory, although time-consuming. Most modern laboratories processing large numbers of diagnostic specimens use auramine–rhodamine staining and fluorescence microscopy; this approach is more sensitive than the Ziehl-Neelsen method. However, it is expensive because it requires high-cost mercury vapor light sources and a dark room. Less expensive light-emitting diode (LED) fluorescence microscopes are now recommended by the WHO as the microscopy tool of choice. They are as sensitive as—or more sensitive than—traditional fluorescence microscopes. As a result, conventional light and fluorescence microscopes are being replaced with this more recent technology, especially in developing countries. For patients with signs or symptoms of pulmonary TB, it has been recommended that one or two sputum specimens, preferably collected early in the morning, should be submitted to the laboratory for AFB smear and mycobacterial culture. If tissue is obtained, it is critical that the portion of the specimen intended for culture not be put in preservation fluid such as formaldehyde. The use of AFB microscopy in examining urine or gastric lavage fluid is limited by the low numbers of organisms, which can cause false-negative results, or the presence of commensal mycobacteria, which can cause false-positive results.

MYCOBACTERIAL CULTURE

Definitive diagnosis depends on the isolation and identification of *M. tuberculosis* from a clinical specimen or the identification of specific DNA sequences in a nucleic acid amplification test. Commercial liquid-culture systems such as the mycobacterial growth indicator tube (MGIT) system (Becton Dickinson, Franklin Lakes, NJ) are recommended by the WHO as the reference standard for culture. The MGIT system uses a fluorescent compound sensitive to the presence of oxygen dissolved in the liquid medium. The appearance of fluorescence, detected by fluorometric technology, indicates active growth of mycobacteria. MGIT cultures usually become positive after a period ranging from 10 days to 2–3 weeks; the tubes are read weekly until the eighth week of incubation before the result is declared to be negative. Specimens may also be inoculated onto egg- or agar-based medium (e.g., Löwenstein-Jensen or Middlebrook 7H10 or 7H11) and incubated at 37°C (under 5% CO_2 for Middlebrook medium). Because most species of mycobacteria, including *M. tuberculosis*, grow slowly, 4–8 weeks may be required before growth is detected on these conventional culture media. Although *M. tuberculosis* may be identified presumptively on the basis of growth time and colony pigmentation and morphology, a variety of biochemical tests have traditionally been used to speciate mycobacterial isolates. In modern, well-equipped laboratories, commercial liquid culture for isolation and species identification by molecular methods or high-pressure liquid chromatography of mycolic acids has replaced isolation on solid media and identification by biochemical tests. A low-cost, rapid immunochromatographic lateral-flow assay based on detection of MTP64 antigen may also be used for species identification of the *M. tuberculosis* complex in culture isolates. These new methods, which are increasingly used in limited-resource settings, have decreased the time required for bacteriologic confirmation of TB to 2–3 weeks.

DRUG SUSCEPTIBILITY TESTING

Universal DST is considered by the WHO as the current standard of care for all TB patients and should consist in DST to at least rifampin for all initial isolates of *M. tuberculosis*, as rifampin resistance is an excellent proxy for MDR-TB. Susceptibility testing is particularly important if one or more risk factors for drug resistance are identified or if the patient either fails to respond to initial therapy or has a relapse after the completion of treatment (see "Treatment Failure and Relapse," below). In addition, expanded and rapid susceptibility testing for isoniazid and key second-line anti-TB drugs (especially the fluoroquinolones and the injectable drugs) is mandatory when RR-TB is found in order to guide selection of the appropriate treatment regimens. Susceptibility testing may be conducted directly (with the clinical specimen) or indirectly (with mycobacterial cultures) on solid or liquid medium. Results are obtained rapidly by direct susceptibility testing on liquid medium, with an average reporting time of 3 weeks. With indirect testing on solid medium, results may not be available for ≥8 weeks. Highly reliable genotypic methods for the rapid identification of genetic mutations in gene regions known to be associated with resistance to rifampin (such as those in *rpo*B) and isoniazid (such as those in *kat*G and *inh*A) have been developed and are being widely implemented for screening of patients at increased risk of drug-resistant TB. Apart from the Xpert MTB/RIF and Xpert MTB/RIF Ultra assays, which, as mentioned above, detect rifampin resistance, the most widely used tests are molecular line probe assays. After extraction of DNA from *M. tuberculosis* isolates or from clinical specimens, the resistance gene regions are amplified by polymerase chain reaction (PCR), and labeled and probe-hybridized PCR products are detected by colorimetric development. This assay reveals the presence of *M. tuberculosis* as well as mutations in target resistance-gene regions. Given the rapidity and accuracy of commercially available line probe assays, the WHO recommends that they (rather than phenotypic culture-based tests) may be used to detect resistance to isoniazid and rifampin when patients have sputum smear–positive specimens or a cultured isolate of *M. tuberculosis*. These recommendations do not eliminate the need for conventional culture-based testing to identify resistance to other drugs and to monitor emergence of additional drug resistance. A similar approach has been developed for second-line anti-TB drugs, such as the fluoroquinolones and the injectable drugs kanamycin, amikacin, and capreomycin. Therefore, second-line line probe assays (instead of phenotypic culture-based DST) are now recommended by the WHO as the initial test for rapid detection of resistance to the fluoroquinolones or the second-line injectable drugs in isolates from patients with confirmed RR-TB or MDR-TB. As with first-line line probe assays, these recommendations do not eliminate the need for conventional phenotypic, culture-based testing to identify resistance to other drugs and to

monitor for the emergence of additional resistance. Finally, a few noncommercial, inexpensive culture and susceptibility testing methods (e.g., microscopically observed drug susceptibility, nitrate reductase, and colorimetric redox indicator assays) have been used in resource-limited settings. Their use is restricted to national reference laboratories with proven proficiency and adequate external quality control as an interim solution while genotypic or automated liquid-culture technology is introduced.

RADIOGRAPHIC PROCEDURES

CXR is a rapid imaging technique that has historically been used as a primary tool to detect pulmonary TB. CXR has high sensitivity but poor specificity. Although TB may often present with typical patterns strongly suggesting the disease, some abnormalities seen in TB are also present in several other lung conditions. The initial suspicion of pulmonary TB is often based on abnormal CXR findings in a patient undergoing triage for respiratory symptoms. The presence of lesions suggestive of TB should prompt bacteriologic investigations in all cases, without exception. Although the "classic" picture is that of upper-lobe disease with infiltrates and cavities (Fig. 173-6), virtually any radiographic pattern—from a normal film or a solitary pulmonary nodule to diffuse alveolar infiltrates in a patient with adult respiratory distress syndrome—may be seen. In the era of HIV/AIDS, no radiographic pattern can be considered pathognomonic, but CXR can assist in diagnosing TB or ruling it out before initiation of treatment of latent infection. CXR is also helpful as a screening test used preceding rapid molecular assays (Xpert MTB/RIF and line probe assays) to improve their predictive value. Digital CXR technology, which allows display of images in a digital format on a computer screen instead of on x-ray film, offers several advantages: the procedure time is reduced, the running costs are lower, the imaging is of superior quality, and telemedicine assistance is available, including computer-aided detection and interpretation of findings. However, a recent systematic review of studies using computer-aided detection software that analyzes digital imaging for abnormalities compatible with TB concluded that the diagnostic accuracy of this technology is still limited.

CT (Fig. 173-7) may be useful in interpreting questionable findings on plain CXR and in diagnosing some forms of extrapulmonary TB (e.g., Pott's disease; Fig. 173-10). A recent study has shown the potential of positron emission tomography combined with CT for detection of subclinical disease that may be progressing toward full-blown TB in HIV-infected people. MRI is useful in the diagnosis of intracranial TB.

ADDITIONAL DIAGNOSTIC PROCEDURES

Other diagnostic tests may be used when pulmonary TB is suspected. Sputum induction by ultrasonic nebulization of hypertonic saline may be useful for patients who cannot produce a sputum specimen spontaneously. Frequently, patients with radiographic abnormalities that are consistent with other diagnoses (e.g., bronchogenic carcinoma) undergo fiberoptic bronchoscopy with bronchial brushings and endobronchial or transbronchial biopsy of the lesion. Bronchoalveolar lavage of a lung segment containing an abnormality may also be performed. In all cases, it is essential that specimens be submitted for molecular testing with the Xpert MTB/RIF assay, mycobacterial culture, and AFB smear. For the diagnosis of primary pulmonary TB in children, who often do not expectorate sputum, induced sputum specimens and specimens from early-morning gastric lavage may yield positive results in the Xpert MTB/RIF assay or on culture.

Invasive diagnostic procedures are indicated for patients with suspected extrapulmonary TB. In addition to testing of specimens from involved sites (e.g., CSF for tuberculous meningitis, pleural fluid and biopsy samples for pleural disease), biopsy and culture of bone marrow and liver tissue have a good diagnostic yield in disseminated (miliary) TB, particularly in HIV-infected patients, who also have a high frequency of positive blood cultures. Xpert MTB/RIF should always be the initial diagnostic test in patients where TB meningitis is suspected; any positive results should prompt immediate treatment initiation, while negative results should be followed up by additional testing. In some cases, the results of culture or Xpert MTB/RIF are negative but a clinical diagnosis of TB is supported by consistent epidemiologic evidence (e.g., a history of close contact with an infectious patient) and a compatible clinical and radiographic response to treatment. In the United States and other industrialized countries with low rates of TB, some patients with limited abnormalities on CXR and sputum positive for AFB are infected with nontuberculous mycobacteria, most commonly organisms of the *M. avium* complex or *M. kansasii* **(Chap. 175)**. Factors favoring the diagnosis of nontuberculous mycobacterial disease over TB include an absence of risk factors for TB and the presence of underlying chronic pulmonary disease.

Patients with HIV-associated TB pose several diagnostic problems (see "HIV-Associated TB," above). HIV-infected patients with sputum culture–positive, AFB-positive TB may present with a normal chest radiograph. The Xpert MTB/RIF assay is the preferred rapid diagnostic test in this population of patients because of its simplicity and increased sensitivity (~60–70% among AFB-negative, culture-positive cases and 97–98% among AFB-positive cases). With the advent of ART, the occurrence of disseminated *M. avium* complex disease that can be confused with TB has become much less common. A test based on the detection of mycobacterial lipoarabinomannan antigen in urine has emerged as a potentially useful point-of-care test for TB in HIV-infected persons with low CD4+ T cell counts. The lateral-flow urine lipoarabinomannan assay can be performed manually and read by eye. After a systematic review of the evidence, the WHO recommends that this assay be used to assist in the diagnosis of TB in HIV-positive adults who have signs and symptoms of TB and a CD4+ T cell count of ≤100 cells/μL or in HIV-positive patients who are seriously ill regardless of CD4+ T cell count or with an unknown CD4+ count. The WHO also recommends that this test not be used, pending information on recent promising technological test advances, for TB diagnosis or as a screening test for TB in any other patient categories.

SEROLOGIC AND OTHER DIAGNOSTIC TESTS FOR ACTIVE TB

A number of serologic tests based on detection of antibodies to a variety of mycobacterial antigens have been carefully assessed by the WHO and found not to be useful as diagnostic aids because of their low sensitivity and specificity and their poor reproducibility. In 2011, after a rigorous evaluation of these tests,

the WHO issued a “negative” recommendation in order to prevent their abuse in the private sector of many resource-limited countries. Various methods aimed at detection of mycobacterial antigens in diagnostic specimens are being investigated but are limited at present by low sensitivity. Determinations of adenosine deaminase and IFN-γ levels in pleural fluid may be useful adjunctive tests in the diagnosis of pleural TB; their utility in the diagnosis of other forms of extrapulmonary TB (e.g., pericardial, peritoneal, and meningeal) is less clear.

DIAGNOSIS OF LATENT *M. TUBERCULOSIS* INFECTION

Two tests currently exist for identification of individuals with LTBI: the TST and IGRAs. Both of these tests have limitations, especially in settings or populations with high TB and/or HIV prevalence.

Tuberculin Skin Testing

In 1891, Robert Koch discovered that components of *M. tuberculosis* in a concentrated liquid-culture medium, subsequently named “old tuberculin,” were capable of eliciting a skin reaction when injected subcutaneously into patients with TB. In 1932, Seibert and Munday purified this product by ammonium sulfate precipitation to produce an active protein fraction known as *tuberculin purified protein derivative* (PPD). In 1941, PPD-S, developed by Seibert and Glenn, was chosen as the international standard. Later, the WHO and UNICEF sponsored large-scale production of a master batch of PPD (RT23) and made it available for general use. The greatest limitation of PPD is its lack of mycobacterial species specificity, a property due to the large number of proteins in this product that are highly conserved in the various species. In addition, subjectivity of the skin-reaction interpretation, deterioration of the product, and batch-to-batch variations limit the usefulness of PPD.

The skin test with tuberculin PPD (TST) is most widely used in screening for LTBI. It probably measures the response to antigenic stimulation by T cells that reside in the skin rather than the response of recirculating memory T cells. The test is of limited value in the diagnosis of active TB because of its relatively low sensitivity and specificity and its inability to discriminate between LTBI and active disease. False-negative reactions are common in immunosuppressed patients and in those with overwhelming TB. False-positive reactions may be caused by infections with nontuberculous mycobacteria **(Chap. 175)** and by BCG vaccination. A repeated TST can produce larger reaction sizes due to either boosting or true conversion. The “boosting phenomenon” is a spurious TST conversion resulting from boosting of reactivity on a subsequent TST 1–5 weeks after the initial test. Distinguishing boosting from true conversion is difficult yet important and can be based on clinical and epidemiologic considerations. For instance, true conversions are likely after BCG vaccination in a previously TST-negative person or in a close contact of an infectious patient.

IFN-γ Release Assays

Two in vitro assays that measure T cell release of IFN-γ in response to stimulation with the highly TB-specific RD1-encoded antigens ESAT-6 and CFP-10 were introduced in the early 2000s and are commercially available. The T-SPOT®. TB test (Oxford Immunotec, Oxford, United Kingdom) is an enzyme-linked immunospot assay, and the QuantiFERON®-TB Gold test (Qiagen GmbH, Hilden, Germany) is a whole-blood enzyme-linked immunosorbent assay for measurement of IFN-γ. The QuantiFERON®-TB Gold In-Tube assay, which facilitates blood collection and initial incubation, also contains another specific antigen, TB7.7. These tests likely measure the response of recirculating memory T cells—normally part of a reservoir in the spleen, bone marrow, and lymph nodes—to persisting bacilli producing antigenic signals.

In settings or population groups with low TB and HIV burdens, IGRAs have previously been reported to be more specific than the TST as a result of less cross-reactivity with BCG vaccination and sensitization by nontuberculous mycobacteria; i.e., RD1 antigens are not encoded in the genome of either BCG strains or most nontuberculous mycobacteria. Recent studies suggest that IGRAs may not perform well in serial testing (e.g., among health care workers) and that interpretation of results depends on cutoff values used to define positivity. Potential advantages of IGRAs include logistical convenience, the need for fewer patient visits to complete testing, and the avoidance of somewhat subjective measurements (e.g., skin induration). However, IGRAs require that blood be drawn and then delivered to the laboratory in a timely fashion. IGRAs also require that testing be performed by specially trained technicians in a laboratory setting. These requirements pose challenges similar to those faced with the TST, including cold-chain requirements and batch-to-batch variations. Because of higher specificity and greater availability of resources, IGRAs have usually replaced the TST for LTBI diagnosis in low-incidence, high-income settings. However, in high-incidence TB and HIV settings and population groups, evidence about the performance and usefulness of IGRAs is still limited, and cost considerations may currently limit wider use.

A number of national guidelines on the use of IGRAs for LTBI testing have been issued. In the United States, an IGRA is preferred to the TST for most persons over the age of 5 years who are being screened for LTBI. However, for individuals at high risk of progression to active TB (e.g., HIV-infected persons), either test—or, to optimize sensitivity, both tests—may be used. Because of the paucity of data on the use of IGRAs in children, the TST is preferred for LTBI testing of children aged <5. In Canada and some European countries, a two-step approach for those with positive TSTs—i.e., an initial TST followed by an IGRA—is recommended. However, a TST may boost an IGRA response if the interval between the two tests exceeds 3 days.

In conclusion, both the TST and IGRAs, although useful as diagnostic aids, are imperfect tests for LTBI: while they can identify latently infected persons, they have low predictive value in identifying individuals with the highest risk of progression toward disease, cannot differentiate between active TB and LTBI, cannot distinguish new infections from reinfections, and display reduced sensitivity in immunocompromised patients.

TREATMENT

TREATMENT

Tuberculosis

The two main aims of TB treatment are (1) to prevent morbidity and death by curing TB while preventing the emergence of drug resistance and (2) to interrupt transmission by rendering patients noninfectious to others. Chemotherapy for TB became possible with the discovery of streptomycin in 1943. Randomized clinical trials clearly indicated that the administration of streptomycin to patients with chronic TB reduced mortality rates and led to cure in the majority of cases. However, monotherapy with streptomycin was soon associated with the development of resistance to this drug and the resulting failure of treatment. With the introduction into clinical practice of para-aminosalicylic acid (PAS) and isoniazid, it became axiomatic in the early 1950s that cure of TB required the concomitant administration of at least two agents to which the organism was susceptible. Furthermore, early clinical trials demonstrated that a long period of treatment—i.e., 12–24 months—was required to prevent recurrence. The introduction of rifampin (rifampicin) in the early 1970s heralded the era of effective short-course chemotherapy, with a treatment duration of <12 months. The discovery that pyrazinamide, which was first used in the 1950s, augmented the potency of isoniazid/rifampin regimens led to the use of a 6-month course of this triple-drug regimen as standard therapy. Streptomycin was added as the fourth drug mainly to prevent the emergence of drug resistance. These four drugs (with streptomycin eventually replaced by ethambutol) still form the basis of the optimal treatment regimen for rifampin-susceptible TB. The emergence of drug-resistant TB in the 1990s prompted attempts to standardize the approach to treatment of this condition mainly on the basis of expert opinion. This event has also stimulated research on and development of new anti-TB agents in the past 15 years. In 2013 and 2014, respectively, bedaquiline and delamanid—the first two drugs specifically developed for TB during nearly half a century—received conditional approval by the U.S. Food and Drug Administration (FDA) and other drug-regulatory authorities; approval was based on the results of phase 2b clinical trials in which the drugs were added to the 18- to 24-month WHO-recommended regimen for MDR-TB. Bedaquiline and delamanid are being used increasingly for treatment of MDR-TB under specific conditions.

DRUGS

Four major drugs are considered first-line agents for the treatment of TB: isoniazid, rifampin, pyrazinamide, and ethambutol. **Table 173-2** presents currently recommended dosages in adults and children. Some studies have suggested increased effectiveness when isoniazid, rifampin, and pyrazinamide are given at higher dosage; thus dosages may be revised in the future. These drugs are well absorbed after oral administration, with peak serum levels at 2–4 h and nearly complete elimination within 24 h. Except for ethambutol, these agents are recommended on the basis of their bactericidal activity (i.e., their ability to rapidly reduce the number of viable organisms and render patients noninfectious); in addition, all four agents are recommended in light of their sterilizing activity (i.e., their ability to kill all bacilli and thus sterilize the affected tissues, measured in terms of the ability to prevent relapses) and their low rate of induction of drug resistance by selection of mutant bacilli. Two additional rifamycins, rifapentine and rifabutin, are also available; however, the level of cross-resistance with rifampin is high. For a detailed discussion of the drugs used for the treatment of TB, see **Chap. 176**.

Because of a lower degree of effectiveness and a higher degree of intolerability and toxicity, several classes of second-line drugs are generally used only for the treatment of patients with drug-resistant TB. These agents are classified at the moment into four groups designated by letters: (A) the fluoroquinolones; (B) the second-line injectable aminoglycosides kanamycin, amikacin, and streptomycin and the injectable polypeptide capreomycin; (C) other oral agents (ethionamide and prothionamide, cycloserine and terizidone, linezolid, and clofazimine); and (D) add-on agents. Group D encompasses three subgroups: D1 (the first-line drugs pyrazinamide, ethambutol, and high-dose isoniazid); D2 (the two new drugs bedaquiline and delamanid); and D3 (PAS, imipenem-cilastatin, meropenem, amoxicillin-clavulanate, and amithiozone [thiacetazone]). Streptomycin, formerly a first-line agent, is now rarely used for drug-resistant TB because resistance levels worldwide are high and it is more toxic than the other drugs in the same class; however, the level of cross-resistance with the other injectable agents is not complete. Of the quinolones, later-generation agents such as levofloxacin (high-dose) and moxifloxacin are recommended; gatifloxacin can be considered as a good alternative with proper selection of patients and careful monitoring of safety. Group D2 includes the novel drugs belonging to two new classes of antituberculosis agents: the diarylquinoline bedaquiline and the nitroimidazole delamanid. These two compounds, which have been shown in phase 2b clinical trials to increase chances of cure among people with MDR-TB, must be used in accordance with international recommendations. However, recent information from the phase 3 clinical trial of delamanid added to an optimized WHO background regimen suggests that treatment success is similar to that of the optimized background regimen with lower than previously observed cardiac toxicity. At the moment, the future role of delamanid in MDR-TB treatment remains to be elucidated. Group D3 agents, the efficacy of which in MDR-TB regimens is not clearly defined, are used in the treatment of patients with TB resistant to most first- and second-line agents. Today, amithiozone is used very rarely because it has been associated with severe and at times fatal skin reactions, including Stevens-Johnson syndrome, among HIV-infected patients.

REGIMENS

Standard regimens are divided into an intensive (bactericidal) phase and a continuation (sterilizing) phase. During the intensive phase, the majority of tubercle bacilli are killed, symptoms resolve, and usually the patient becomes noninfectious. The continuation phase is required to eliminate persisting mycobacteria and prevent relapse.

The treatment regimen of choice for virtually all forms of drug-susceptible TB in adults consists of a 2-month initial (intensive) phase of isoniazid, rifampin, pyrazinamide, and ethambutol followed by a 4-month continuation phase of isoniazid and rifampin **(Table 173-3)**. This regimen can cure TB in >90% of patients. In children, most forms of TB in the absence of HIV infection or suspected isoniazid resistance can be safely treated without ethambutol in the intensive phase. Treatment should be given daily throughout the course. Systematic reviews have demonstrated that the use of an intermittent thrice-weekly regimen in the intensive phase is associated with increased risk of treatment failure, relapse, and acquisition of drug resistance. Furthermore, a thrice-weekly regimen in the continuation phase only has also been associated with increased rates of failure and relapse, while a twice-weekly regimen in the continuation phase increased the risk of acquisition of drug resistance as well as rates of failure and relapse. Therefore, the WHO now recommends that

TB treatment in all cases be administered daily. The 2016 guidelines by the ATS, the CDC, and the IDSA, while recommending daily administration of drugs, include a provision for use of intermittent thrice-weekly supervised regimens among patients who are not infected with HIV and are at low risk of relapse (i.e., have pulmonary TB caused by drug-susceptible organisms that, at the start of treatment, is noncavitary and/or sputum smear–negative). The same guidelines suggest that a 4-month regimen consisting of isoniazid, rifampin, pyrazinamide, and ethambutol may be adequate for treatment of HIV-negative adults with sputum smear–negative and culture-negative pulmonary TB (i.e., paucibacillary TB).

A continuation phase of once-weekly rifapentine and isoniazid is effective in HIV-seronegative patients without cavitation on CXR. In general, however, this regimen should be used with great caution. Patients with cavitary pulmonary TB and delayed sputum-culture conversion (i.e., those who remain culture-positive at 2 months) should be re-tested immediately for drug-resistant TB, and a change of regimen should be considered. A full course of therapy should not include interruptions of >4 weeks. In some developing countries where the ability to ensure adherence to treatment is limited, a continuation-phase regimen of daily isoniazid and ethambutol for 6 months has been used in the past. This regimen is clearly associated with a higher rate of relapse, treatment failure, and death, especially among HIV-infected patients, and is no longer recommended by the WHO. Several studies attempting to reduce treatment duration to 4 months by using fluoroquinolones (with moxifloxacin replacing ethambutol or isoniazid, or gatifloxacin replacing ethambutol) were conducted over the last decade. The main finding was that shorter (4-month) fluoroquinolone-containing regimens are associated with significantly higher rates of relapse at 18 months than the standard 6-month rifampin-containing regimen. In addition, the studies showed no reduction in adverse events with the fluoroquinolone-containing regimen and no difference in all-cause and TB-related mortality rates. Therefore, shortening of the treatment duration to 4 months through the use of fluoroquinolones is not recommended. Alternative regimens for patients who exhibit drug intolerance or adverse reactions are listed in Table 173-3. However, severe side effects prompting discontinuation of any of the first-line drugs and use of these alternative regimens are uncommon. To prevent isoniazid-related neuropathy, pyridoxine (10–25 mg/d) should be added to the regimen given to persons at high risk of vitamin B_6 deficiency (e.g., alcoholics; malnourished persons; pregnant and lactating women; and patients with conditions such as chronic renal failure, diabetes, and HIV infection, which are also associated with neuropathy).

PATIENT CARE AND SUPPORT

Poor adherence to treatment is one of the most important impediments to cure. Moreover, the tubercle bacilli harbored by patients who do not fully adhere to the prescribed regimen are likely to become resistant to the drugs to which they are irregularly exposed. Both patient- and provider-related factors may affect adherence. Patient-related factors include a lack of belief that the illness is worth the cost of adherence; the existence of concomitant medical conditions (notably alcohol or substance abuse); lack of social support; fear of the stigma and discrimination associated with TB; and poverty, with attendant joblessness and homelessness. Provider-related factors that may prevent adherence include lack of support, education, and encouragement of patients and inconvenient clinical services.

A variety of interventions to increase the chances of completion of the months-long treatment course are available. First, a package of social support interventions that are complementary and not mutually exclusive, consisting of educational, psychological, and material goods and services, may enable people with TB to address hurdles to treatment adherence. Health education and counseling on the disease's seriousness and solutions and on the importance of treatment adherence until cure should be provided to all patients at the start of and throughout the course of TB therapy. Psychological support (i.e., counseling sessions or peer-group support) can be particularly relevant in the context of the stigma and discrimination often affecting people with TB and their families. Material support (e.g., food or financial support in forms such as meals, food baskets, food supplements, food vouchers, transport subsidies, living allowances, housing incentives, or financial bonuses) reduces indirect costs incurred by patients or their attendants in accessing health services and mitigates the consequences of income loss related to the disease.

Second, it is paramount that health services be arranged to meet the needs and reasonable expectations of patients. Components of optimal health services include a suitable geographic location, a schedule responsive to patients' needs, functional channels of communication between patients and their health care providers (e.g., a telephone short-messaging system, audio/video call capability, home or workplace visits), and a staff willing and competent to care for people with TB, to address their concerns, and to base the care they provide on sound ethical standards.

Third, it is crucial to offer the patient a suitable option for treatment administration that minimizes the chance of non-adherence. Such options traditionally include unsupervised, self-administered therapy; in-person directly observed therapy (DOT); and non-daily DOT (e.g., supervision not for every dose but weekly or a few times per week) at a location mutually agreed on by patient and health care provider, with supervisory responsibility delegated to a qualified person. Direct supervision of adherence is crucial in view of the lack of tools to accurately predict adherence to self-administered treatment and of the public health importance of TB. The WHO, along with the ATS, the CDC, and the IDSA, states that ideally all patients should have their therapy directly supervised, especially during the initial phase, with proper social support based on a patient-centered approach as described above. In several countries, personnel to supervise therapy are usually available through TB control programs of local public health departments, often involving members of the community who are accepted by the patient and who have been properly trained and educated by health workers to undertake the supervisory role. Direct supervision with social support has been shown to significantly increase the proportion of patients completing treatment in all settings and to lessen the chances of treatment failure, relapse, and default. In general, community- or home-based DOT is recommended over health facility–based DOT or unsupervised treatment; DOT administered by trained lay providers or health care workers is recommended over DOT administered by family members. Recently, comparison of video-observed therapy with in-person DOT has shown similar outcomes. Video-observed therapy can replace DOT when internet access is good and video communication technology (e.g., smartphones, tablets, computers) is available. The system can be appropriately organized and operated by health care providers and patients. Other digital health tools can facilitate the monitoring of adherence, including digital medication monitors; these

monitors can register when the pill box is opened, with options to emit audio signals or a short message to remind patients to take medicines. These tools are customized to the needs and preferences of the individual patient and the provider.

In addition to the above measures promoting adherence, provision of fixed-dose combination products that reduce the number of tablets the patient needs to swallow is recommended over separate drug formulations. Various fixed-dose combination products are available (e.g., isoniazid/rifampin, isoniazid/rifampin/pyrazinamide, and isoniazid/rifampin/pyrazinamide/ethambutol). Fixed-dose combinations increase patient satisfaction and minimize the likelihood of prescription error or of development of drug resistance resulting from monotherapy if a drug is out of stock or the patient prefers one drug over others. In addition, these combinations facilitate programmatic management of procurement and supply. In the past, the bioavailability of rifampin was found to be substandard in some formulations of fixed-dose combinations. Medical regulatory authorities should ensure that combination products are of good quality; however, top standards for drug quality assurance are not always operative, especially in limited-resource countries. Prescribers should be aware of this potential problem.

MONITORING TREATMENT RESPONSE AND DRUG TOXICITY

Bacteriologic evaluation through commercial liquid-culture systems (or—when liquid-culture capacity is not yet available—through smear microscopy) is essential in monitoring the response to TB treatment. In addition, the patient's weight should be monitored regularly and the drug dosage adjusted with any significant weight change. Patients with pulmonary disease should have their sputum examined monthly until cultures become negative to allow early detection of treatment failure. With the recommended 6-month standard first-line regimen, >80% of drug-susceptible TB patients will have negative sputum cultures at the end of the second month of treatment. By the end of the third month, the sputum of virtually all patients should be culture negative. In some patients, especially those with extensive cavitary disease and large numbers of organisms, AFB smear conversion may lag behind culture conversion as a result of the expectoration and microscopic visualization of dead bacilli. Therefore, as capacity is built, smear microscopy should be progressively abandoned as a monitoring tool in favor of liquid culture. As noted above, patients with cavitary disease in whom sputum culture conversion does not occur by 2 months require immediate testing or re-testing for drug resistance. When a patient's sputum cultures or smears remain positive at ≥3 months despite good adherence, treatment failure caused by drug resistance is likely. The pattern of drug resistance should guide the choice of the best treatment option (see below). A sputum specimen should be collected at the end of treatment to document cure. In settings where mycobacterial cultures are not yet available, monitoring by AFB smear examination should be undertaken at 2, 5, and 6 months. Bacteriologic monitoring of patients with extrapulmonary TB is more difficult and often is not feasible. In these cases, the response to treatment must be assessed clinically with the help of medical imaging.

Monitoring of the response to chemotherapy by nucleic acid amplification technology, such as the Xpert MTB/RIF assay, is not suitable because these tests can produce positive results due to nonviable bacilli. Likewise, serial chest radiographs are not recommended because radiographic changes may lag behind bacteriologic response and are not highly sensitive. After the completion of treatment, neither sputum examination nor CXR is recommended for routine follow-up purposes. However, a chest radiograph obtained at the end of treatment may be useful for comparative purposes should the patient develop symptoms of recurrent TB months or years later. Patients should be instructed to report promptly for medical assessment if they develop any such symptoms.

During treatment, patients should also be monitored for drug toxicity. The most common adverse reaction of significance among people treated for drug-susceptible TB is hepatitis. Patients should be carefully educated about the signs and symptoms of drug-induced hepatitis (e.g., dark urine, loss of appetite, nausea) and should be instructed to discontinue treatment promptly and see their health care provider if these manifestations occur. Although biochemical monitoring is not routinely recommended, all adult patients should undergo baseline assessment of liver function (e.g., measurement of serum levels of hepatic aminotransferases and bilirubin). Older patients, those with concomitant diseases, those with a history of hepatic disease (especially hepatitis C), and those using alcohol daily should be monitored especially closely (i.e., monthly), with repeated measurements of aminotransferases, during the initial phase of treatment. Up to 20% of patients have small increases (up to three times the upper limit of normal) in serum levels of aspartate aminotransferase that are not accompanied by symptoms and are of no consequence. For patients with symptomatic hepatitis and those with marked (five- to sixfold) elevations in serum levels of aspartate aminotransferase, treatment should be stopped and drugs reintroduced one at a time after liver function has returned to normal. Hypersensitivity reactions usually require the discontinuation of all drugs and rechallenge to determine which agent is the culprit. Because of the variety of regimens available, it usually is not necessary—although it is possible—to desensitize patients. Hyperuricemia and arthralgia caused by pyrazinamide can usually be managed by the administration of acetylsalicylic acid; however, pyrazinamide treatment should be stopped if the patient develops gouty arthritis. Individuals who develop autoimmune thrombocytopenia secondary to rifampin therapy should not receive the drug thereafter. Similarly, the occurrence of optic neuritis with ethambutol is an indication for permanent discontinuation of this drug. Other common manifestations of drug intolerance, such as pruritus and gastrointestinal upset, can generally be managed without the interruption of therapy. Treatment with second-line agents for drug-resistant TB is associated with a variety of adverse drug reactions that are more frequent and severe than in patients receiving first-line TB regimens (see below). The likelihood of drug–drug interactions is also higher when second-line regimens are used.

TREATMENT FAILURE AND RELAPSE

As stated above, treatment failure should be suspected when a patient's cultures (or sputum smears, when cultures are not available) remain positive after 3 months of treatment. In the management of such patients, it is imperative that the current isolate be urgently re-tested (or tested for the first time if, for some reason, rapid molecular susceptibility testing was not performed at the start of treatment) for susceptibility to first-line agents and, if resistance to rifampin is detected, to second-line agents as well. The treatment approach should start with molecular testing for—at the least—resistance to rifampin and isoniazid. Since results are expected to become available within a few days, changes in the regimen can be postponed until that time. However, if the patient's clinical condition is deteriorating rapidly, an earlier change in regimen may be indicated. A cardinal rule in the latter situation is always to add more

than one drug, preferably two or three, at a time to a failing regimen; in practice, starting an empirical regimen for MDR-TB (see "Drug-Resistant TB," below) is warranted. The patient may continue to take isoniazid and rifampin along with these new agents pending the results of susceptibility tests.

Patients who experience a recurrence after apparently successful treatment (i.e., a relapse) are less likely to harbor drug-resistant strains than are patients in whom treatment has failed. Acquired resistance is uncommon among strains from patients in whom relapse follows the proper completion of a standard 6-month regimen. The treatment decision depends on a general assessment of the risk of drug resistance, the severity of the case, and the results of rapid susceptibility testing. Patients whose treatment has been interrupted and who have a high likelihood of MDR-TB should receive an empirical MDR-TB regimen that includes second-line agents (Table 173-3). Once drug susceptibility results are available, the regimen can be adjusted accordingly.

DRUG-RESISTANT TB



Strains of *M. tuberculosis* resistant to individual drugs arise by spontaneous point mutations in the mycobacterial genome that occur at low but predictable rates (10^{-7}–10^{-10} for the key drugs). Resistance to rifampin is associated with mutations in the *rpo*B gene in 95% of cases; that to isoniazid with mutations mainly in the *kat*G gene (50–95% of cases) and the *inh*A gene promotor region (up to 45%); that to pyrazinamide in the *pnc*A gene (up to 98%); that to ethambutol in the *emb*B gene (50–65%); that to the fluoroquinolones in the *gyr*A–*gyr*B genes (75–95%); and that to the aminoglycosides mainly in the *rrs* gene (up to 80%), with the C-12T mutation as the most common mutation in the *eis* promoter region associated with aminoglycoside resistance, especially in Eastern European countries. Because there is no cross-resistance among the commonly used classes of drugs, the probability that a strain will be resistant to two drug classes is the product of the probabilities of resistance to each drug and thus is low. The development of drug-resistant TB almost invariably follows monotherapy—i.e., the failure of the health care provider to prescribe at least two drugs to which tubercle bacilli are susceptible; of the patient to absorb or take properly prescribed therapy; or of poor-quality drugs to be adequately bioavailable. Drug-resistant TB may be either primary or acquired. In primary drug resistance, the patient is infected from the start by a drug-resistant strain. Acquired resistance develops in the infecting strain during treatment. In North America, Western Europe, most of Latin America, and the Persian Gulf States, rates of primary resistance are generally low and isoniazid resistance is most common. In the United States, although rates of primary isoniazid resistance have been stable at ~7–8%, the rate of primary MDR-TB has declined from 2.5% in 1993 to <1% since 2000. As described above, MDR-TB is an increasingly serious problem in some regions, especially in the countries of the former Soviet Union and some countries of Asia **(Fig. 173-11)**. Even more serious is the occurrence of XDR-TB due to MDR strains that are also resistant to any fluoroquinolones and to any of three second-line injectable agents (amikacin, kanamycin, and capreomycin). Creation of drug-resistant TB can be prevented by adherence to the principles of sound treatment: inclusion of at least two quality-assured, bactericidal drugs to which the organism is susceptible; use of fixed-dose combination products; supervision of treatment with patient support; and verification that patients complete the prescribed course. Transmission of drug-resistant strains can be prevented by implementation of respiratory infection-control measures (see below) and by early detection of people with active TB followed by immediate initiation of treatment with an effective regimen.

For the treatment of patients with isoniazid-resistant disease, it is recommended to use a combination of rifampin, ethambutol, pyrazinamide, and levofloxacin for 6 months. This fluoroquinolone-containing regimen should not be used until rifampin resistance has been excluded by a reliable diagnostic test to avoid inadvertent treatment of MDR-TB with an inadequate regimen. Ideally, a laboratory test for susceptibility should also be done for the fluoroquinolone and pyrazinamide. If the fluoroquinolone is contraindicated because of intolerance or resistance, the patient can be given a 6-month regimen of rifampin, ethambutol, and pyrazinamide. Isoniazid probably does not contribute to a successful outcome in these regimens but may be retained (also to facilitate treatment with the four-drug fixed-dose formulation). Injectable agents, such as streptomycin and kanamycin, are unlikely to play a role in the treatment of most isoniazid-resistant TB cases. MDR-TB, in which bacilli are resistant to (at least) isoniazid and rifampin, is more difficult to manage than is disease caused by drug-susceptible organisms because these two bactericidal drugs are the most potent agents available and because associated resistance to other first-line drugs as well (e.g., ethambutol) is not uncommon. Therapy for MDR-TB is therefore more complex: it is based on limited evidence, is lengthy and potentially toxic, uses drugs of limited efficacy, and is much more expensive than treatment of drug-susceptible disease. For treatment of MDR-TB (and all other rifampin-resistant TB cases in which isoniazid resistance is absent or unknown), two approaches are currently recommended by the WHO: a shorter standardized regimen of 9–12 months' duration and a longer regimen of 18–24 months' duration consisting of an optimal combination of drugs according to a standard design **(Fig. 173-12)**. In patients who have not been treated previously with second-line drugs and in whom resistance to fluoroquinolones and second-line injectable agents has been excluded or is considered highly unlikely, the WHO recommends the use of a shorter MDR-TB regimen comprising seven drugs and given for 9–12 months. The regimen consists of 4–6 months of kanamycin, moxifloxacin, clofazimine, prothionamide, pyrazinamide, high-dose isoniazid (10 mg/kg and up to 15 mg/kg), and ethambutol followed by 5 months of moxifloxacin, clofazimine, pyrazinamide, and ethambutol. This regimen has been tested in observational studies in a number of countries under operational and programmatic conditions and has been successful in a high percentage of cases, with low toxicity; results of an ongoing randomized, controlled clinical trial are expected by the end of 2018. This regimen is the first choice in the majority of MDR-TB cases worldwide, but it should be used only when the criteria shown in Fig. 173-12 are met. As with any anti-TB regimen, the risk of creating additional resistance is high if the regimen is used inappropriately.

In all other patients with MDR-TB or RR-TB, and especially in those presenting with—or presumed to have an infection with—a more complex pattern of resistance (e.g., resistance to the fluoroquinolones, the injectable drugs, or both [XDR-TB]), a regimen with at least five effective TB drugs during the intensive phase is recommended; this regimen should include pyrazinamide and four second-line TB agents—one from group A, one from group B, and at least two from group C (see "Drugs," above, and **Table 173-4**). If the minimal number of these effective TB drugs cannot be incorporated, an agent from subgroup D2 (see below) and other agents from subgroup D3 may be added to bring the total number of drugs to five. The regimen may be further strengthened with high-dose isoniazid and/or ethambutol; Table 173-4 summarizes the steps in designing such a regimen. Although the optimal duration of treatment is not known, a course of at least 20 months is recommended for previously untreated patients, including the initial phase of 5 months with an injectable agent,

which is usually discontinued 4 months after culture conversion. In late 2012, the FDA granted accelerated approval of bedaquiline, a diarylquinoline antibiotic included in group D2 (Table 173-4). This drug, when given in addition to an optimized MDR-TB regimen for the first 24 weeks (400 mg daily for 2 weeks followed by 200 mg thrice weekly for 22 weeks), accelerated sputum conversion in phase 2B clinical trials. Bedaquiline should be used with caution for people aged >65 years and for HIV-infected patients; its use is not advised for children and pregnant women. In early 2014, the European Medical Agency granted accelerated approval of another new agent, the nitroimidazole compound delamanid, which is also included in group D2. Data from a phase 2B clinical trial in which delamanid was added to the WHO-recommended longer MDR-TB regimen showed increased rates of culture conversion at 2 months. In 2016, the WHO expanded its recommendation for the use of delamanid to children aged 6–17 years who are not eligible for the shorter regimen. However, recent information from the phase 3 clinical trial of delamanid added to an optimized WHO background regimen shows that, although sputum conversion occurred 1–2 weeks earlier and cardiac toxicity was lower than predicted, treatment success was similar to that of the optimized background regimen. At the moment, therefore, the future role of delamanid in MDR-TB treatment remains to be elucidated. Pending phase 3 trial results and in view of the risk of adverse reactions (including QT interval prolongation), the WHO currently recommends the use of bedaquiline in cases of MDR-TB in which an effective WHO-recommended longer MDR-TB regimen cannot be designed because of known resistance, intolerability, or unavailability of groups A, B, and C second-line drugs. Informed consent should be sought from patients treated with bedaquiline or delamanid, and their treatment should be closely monitored. In particular, the QT interval should be monitored at the start of and during treatment; patients with a QTc interval >500 ms or a history of ventricular arrhythmias should not be given these drugs. Current information about simultaneous use of these two agents is insufficient to make a recommendation; it is therefore prudent to avoid the combination outside of clinical trials.

Patients with XDR-TB have fewer treatment options and a much poorer prognosis. As part of a patient-centered approach, palliative and end-of-life care should be provided to these patients as a priority when all treatment options have been exhausted; respiratory infection-control measures should be observed throughout. The design of XDR-TB regimens follows the same principles outlined in Table 173-4. The use of a novel regimen composed of the nitroimidazole compound pretomanid, bedaquiline, and linezolid recently resulted in a high cure rate among patients with XDR-TB in South Africa. In the future, this promising regimen may become an important therapeutic option. Observational studies have shown that aggressive management of XDR-TB cases, with early DST, rational combination of at least five effective drugs, strict DOT, monthly bacteriologic monitoring, and intensive patient support, may increase the chances of cure and avert death. For patients with localized disease and sufficient pulmonary reserve, lobectomy or wedge resection may be considered as part of treatment. Because the management of patients with MDR- and XDR-TB is complicated by both social and medical factors, care of seriously ill patients is ideally provided in specialized centers or, in their absence, in the context of programs with adequate resources and capacity, including community support. When possible, treatment and care on an ambulatory basis at a decentralized health care facility should be prioritized, as this approach may increase treatment success and reduce loss to follow-up. This approach should not, however, preclude hospitalization when it is necessary.

HIV-ASSOCIATED TB

Several observational studies and randomized controlled trials have shown that treatment of HIV-associated TB with anti-TB drugs and simultaneous use of ART are associated with significant reductions in mortality risk and AIDS-related events. Evidence from randomized controlled trials shows that early initiation of ART during anti-TB treatment is associated with a 34–68% reduction in mortality rates, with especially good results in patients with CD4+ T cell counts of <50/μL. Therefore, the main aim in the management of HIV-associated TB is to initiate anti-TB treatment and to immediately consider initiating or continuing ART. All HIV-infected TB patients, regardless of CD4+ T cell count, are candidates for ART, which optimally is initiated as soon as possible after the diagnosis of TB and within the first 8 weeks of anti-TB therapy; ART should be started within the first 2 weeks of TB treatment for profoundly immunosuppressed patients with CD4+ T cell counts of <50/μL. In general, the standard 6-month daily regimen is equally efficacious in HIV-negative and HIV-positive patients with drug-susceptible TB. However, in the uncommon situation where an HIV-infected patient cannot receive ART, prolongation of the continuation phase of TB treatment by 3 months can be considered. As in any other TB patient, intermittent regimens should not be used in HIV-infected people. As for any other adult living with HIV **(Chap. 197)**, first-line ART for TB patients consists of two nucleoside reverse transcriptase inhibitors plus a nonnucleoside reverse transcriptase inhibitor or an integrase inhibitor. Although TB treatment modalities are similar to those in HIV-negative patients, adverse drug reactions may be more pronounced in HIV-infected patients. In this regard, three important considerations are relevant: an increased frequency of paradoxical reactions, interactions between ART components and rifamycins, and development of rifampin monoresistance with intermittent treatment. IRIS—i.e., the exacerbation of symptoms and signs of TB—has been described above. Rifampin, a potent inducer of enzymes of the cytochrome P450 system, lowers serum levels of many HIV protease inhibitors and some nonnucleoside reverse transcriptase inhibitors—essential drugs used in ART. In such cases, rifabutin, which has much less enzyme-inducing activity, has been used in place of rifampin. However, dosage adjustments for rifabutin and protease inhibitors are still being assessed. Several clinical trials have found that patients with HIV-associated TB whose degree of immunosuppression is advanced (e.g., CD4+ T cell counts of <100/μL) are prone to treatment failure and relapse with rifampin-resistant organisms when treated with “highly intermittent” (i.e., once- or twice-weekly) rifamycin-containing regimens. Consequently, it is now recommended that all TB patients who are infected with HIV, like all other TB patients with rifampin-susceptible disease, receive a rifampin-containing regimen on a daily basis. Because recommendations are frequently updated, consultation of the following websites is advised: *www.who.int/hiv*, *www.who.int/tb*, *www.cdc.gov/hiv*, and *www.cdc.gov/tb*.

SPECIAL CLINICAL SITUATIONS

Although comparative clinical trials of treatment for extrapulmonary TB are limited, the available evidence indicates that most forms of disease can be treated with the 6-month regimen recommended for patients with pulmonary disease. For TB meningitis, the ATS, the CDC, and the IDSA recommend extension of the continuation phase for 7–10 months. The WHO and the American Academy of Pediatrics recommend that children with bone and joint TB, tuberculous meningitis, or miliary TB receive up to 12 months of treatment. Treatment for TB may be complicated by underlying medical problems that require special consideration. As a rule, patients with chronic renal failure should not receive aminoglycosides and should receive ethambutol only if serum drug levels can be monitored. Isoniazid, rifampin, and pyrazinamide may be given in the usual doses in cases of mild to moderate renal failure, but the

dosages of isoniazid and pyrazinamide should be reduced for all patients with severe renal failure except those undergoing hemodialysis. Patients with hepatic disease pose a special problem because of the hepatotoxicity of isoniazid, rifampin, and pyrazinamide. Patients with severe hepatic disease may be treated with ethambutol, streptomycin, and possibly another drug (e.g., a fluoroquinolone); if required, isoniazid and rifampin may be administered under close supervision. The use of pyrazinamide by patients with liver failure should be avoided. Silicotuberculosis necessitates the extension of therapy by at least 2 months.

The regimen of choice for pregnant women (Table 173-3) is 9 months of treatment with isoniazid and rifampin supplemented by ethambutol for the first 2 months. Although the WHO has recommended routine use of pyrazinamide for pregnant women in combination with isoniazid and rifampin, this drug has not been recommended for pregnant women in the United States because of insufficient data documenting its safety in pregnancy. Streptomycin is contraindicated because it is known to cause eighth-cranial-nerve damage in the fetus. The thioamides, bedaquiline, and delamanid should also be avoided in the treatment of pregnant women with MDR-TB. Treatment for TB is not a contraindication to breast-feeding; most of the drugs administered will be present in small quantities in breast milk, albeit at concentrations far too low to provide any therapeutic or prophylactic benefit to the child.

Medical consultation on difficult-to-manage cases is provided by the U.S. CDC Regional Training and Medical Consultation Centers (*www.cdc.gov/tb/education/rtmc/*).

TABLE 173-2

Recommended Dosage[a] for Initial Treatment of Tuberculosis in Adults and Children

Drug	Daily Dose	
	Adult	Pediatric
Isoniazid	5 mg/kg, max 300 mg	10 (7–15) mg/kg, max 300 mg
Rifampin	10 mg/kg, max 600 mg	15 (10–20) mg/kg, max 600 mg
Pyrazinamide	25 mg/kg, max 2 g	35 (30–40) mg/kg
Ethambutol[b]	15 mg/kg	20 (15–25) mg/kg

[a]The duration of treatment with individual drugs varies by regimen, as detailed in Table 173-3. [b]In certain settings, streptomycin (15 mg/kg daily, with a maximal dose of 1 g; or 25–30 mg/kg thrice weekly, with a maximal dose of 1.5 g) can replace ethambutol in the initial phase of treatment. However, streptomycin generally is no longer considered a first-line drug.

Source: Based on recommendations of the American Thoracic Society/Infectious Diseases Society of America/Centers for Disease Control and Prevention and the World Health Organization.

TABLE 173-3

Recommended Antituberculosis Treatment Regimens

INDICATION	INITIAL PHASE		CONTINUATION PHASE	
	DURATION, MONTHS	DRUGS	DURATION, MONTHS	DRUGS
New smear- or culture-positive cases	2	HRZE[a,b]	4	HR[a,c]
New culture-negative cases	2	HRZE[a]	4	HR[a,d]
Pregnancy	2	HRE[e]	7	HR
Relapses and treatment default[f]	← Tailored according to rapid drug susceptibility testing →			
Failures[f]	← Tailored according to rapid drug susceptibility testing →			
Resistance (or intolerance) to H	Throughout (6)	RZEQ		
Resistance (or intolerance) to R	← Same as for MDR-TB; see below →			
MDR-TB (resistance to at least H + R)	← See Fig. 173-12 and Table 173-4 →			
XDR-TB	← See Table 173-4 →			
Intolerance to Z	2	HRE	7	HR

[a]All drugs should be given daily. [b]Streptomycin was used in the past in place of ethambutol but is no longer considered a first-line drug. [c]A clinical trial showed that HIV-negative patients with noncavitary pulmonary tuberculosis who have negative sputum AFB smears after the initial phase of treatment can be given once-weekly rifapentine/isoniazid in the continuation phase. However, this regimen is rarely used. [d]The American Thoracic Society, the Centers for Disease Control and Prevention, and the Infectious Diseases Society of America suggest that a 2-month continuation phase could be used in HIV-seronegative patients with sputum smear–negative and culture-negative TB. [e]The 6-month regimen with pyrazinamide can probably be used safely during pregnancy and is recommended by the WHO and the International Union Against Tuberculosis and Lung Disease. If pyrazinamide is not included in the initial treatment regimen, the minimal duration of therapy is 9 months. [f]The availability of rapid molecular methods to identify drug resistance allows initiation of a proper regimen at the start of treatment.

Abbreviations: E, ethambutol; H, isoniazid; MDR-TB, multidrug-resistant tuberculosis; Q, a quinolone antibiotic; R, rifampin; WHO, World Health Organization; XDR-TB, extensively drug-resistant tuberculosis; Z, pyrazinamide.

TABLE 173-4

General Steps in Designing a Longer Regimen for the Treatment of Multidrug-Resistant Tuberculosis (MDR-TB)[a]

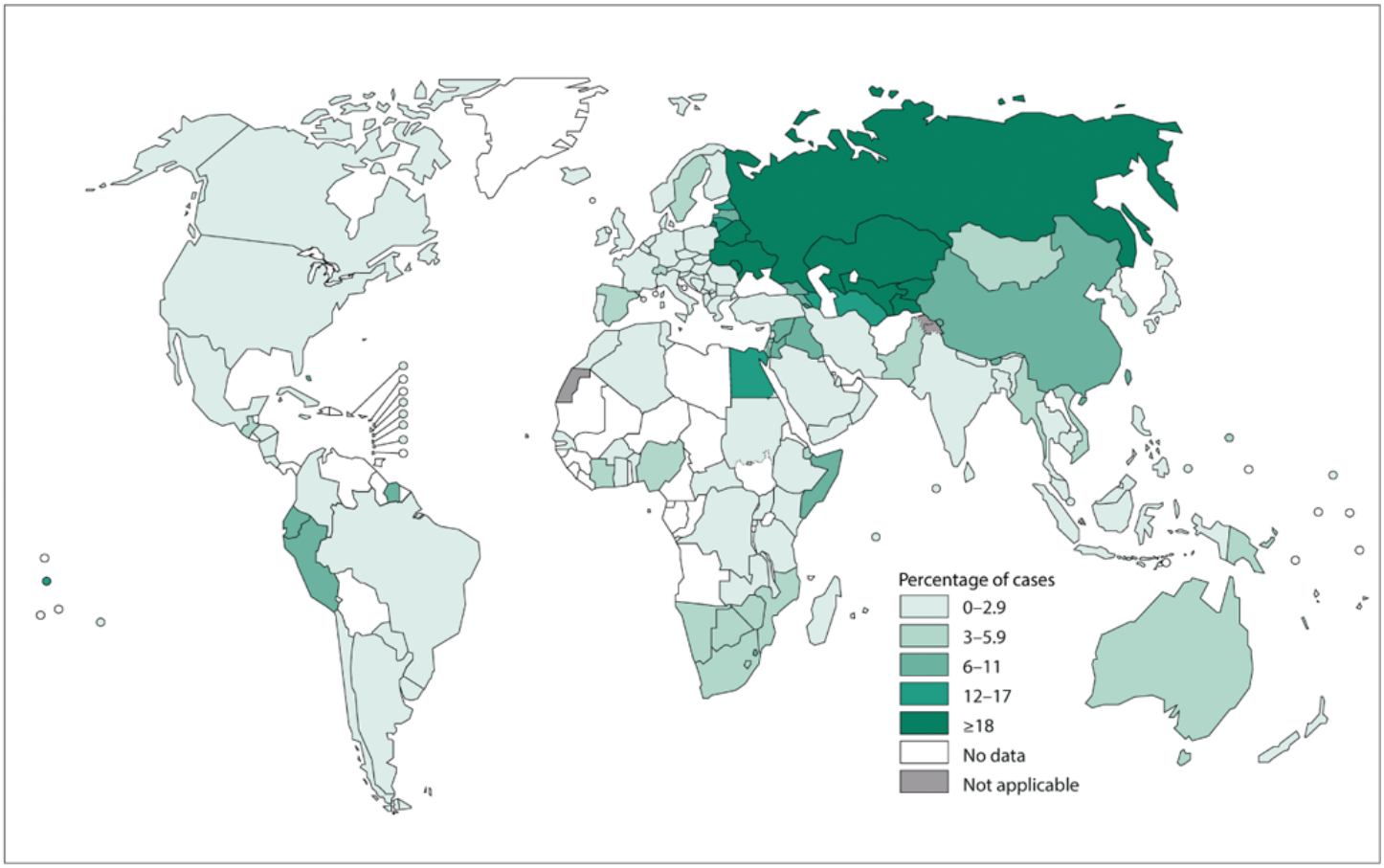
Step	Drug Group	Options
1. Add one later-generation fluoroquinolone.	A[b]	Levofloxacin Moxifloxacin Gatifloxacin
2. Add one second-line injectable agent.	B	Amikacin Capreomycin Kanamycin (Streptomycin[c])
3. Add two or more second-line agents.	C[b]	Ethionamide/prothionamide Cycloserine/terizidone Linezolid Clofazimine
4. Add pyrazinamide and any other first-line agent if they can help strengthen the regimen.	D1	Pyrazinamide Ethambutol High-dose isoniazid
5. Add bedaquiline or delamanid.[d]	D2	Bedaquiline Delamanid
6. Add any of these agents if the regimen cannot be composed otherwise.	D3	*p*-Aminosalicylic acid Imipenem-cilastatin[e] Meropenem[e] Amoxicillin-clavulanate[e] (Thioacetazone[f])

[a]This stepwise approach guides the design of individualized longer regimens for patients who are not eligible for the WHO-recommended shorter regimen. (The composition of the shorter MDR-TB regimen is standardized; see text). The aim is to combine at least five effective agents in the intensive phase; more agents may be included if they can safely increase the chances of cure. The choice of an agent is based on the likelihood of its effectiveness, on reliable information about drug resistance, and on the balance of expected benefits to risk. For instance, in case of nephrotoxicity or hearing loss, an injectable agent may be omitted and additional agents from group C or D2 included. [b]Drugs in groups A and C are shown in decreasing order of usual preference (subject to other considerations). [c]Streptomycin may be substituted for other injectable agents when the other three cannot be used. Resistance to streptomycin alone does not fulfill the definition of extensively drug-resistant TB. [d]Bedaquiline or delamanid may be added to the longer regimen to replace another second-line agent or to strengthen it in accordance with interim policies. [e]Carbapenems and clavulanate are used together. Clavulanate is available only in formulations combined with amoxicillin. [f]HIV status must be tested and confirmed to be negative before thioacetazone treatment is started.

Source: Adapted from D Falzon et al: World Health Organization treatment guidelines for drug-resistant tuberculosis, 2016 update. Eur Respir J 49:1602308, 2017.

FIGURE 173-11

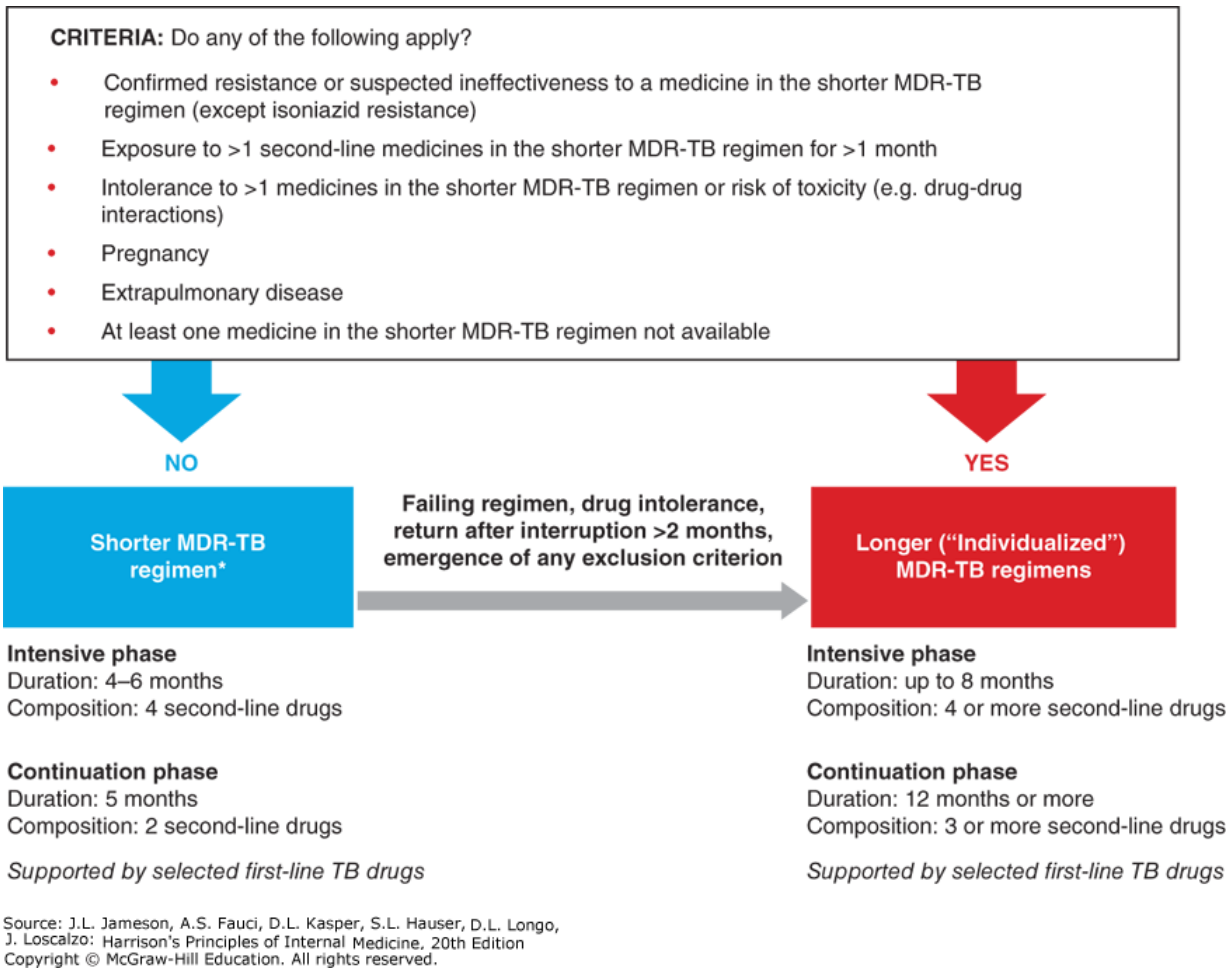
Percentage of new cases of multidrug-resistant/rifampin-resistant TB in all countries surveyed by the World Health Organization (WHO) Global Drug Resistance Surveillance Project during 1994–2016. Figures are based on the most recent year for which data have been reported, which varies among countries. Data reported before the year 2002 are not shown. *(See disclaimer in Fig. 173-2. Courtesy of the Global TB Programme, WHO; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 173-12

Choosing the treatment regimen for patients with confirmed rifampin-resistant/multidrug-resistant (RR/MDR) tuberculosis. *The intensive shorter regimen consists of 4–6 months of kanamycin, moxifloxacin, prothionamide, clofazimine, pyrazinamide, high-dose isoniazid, and ethambutol. The shorter continuation regimen consists of 5 months of moxifloxacin, clofazimine, pyrazinamide, and ethambutol. *(Source: World Health Organization.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PREVENTION

The primary way to prevent TB is to diagnose and isolate infectious cases rapidly and to administer appropriate treatment until patients are rendered noninfectious (usually 2–4 weeks after the start of proper treatment) and the disease is cured. Additional strategies include BCG vaccination and treatment of persons with LTBI who are at high risk of developing active disease.

BCG VACCINATION

BCG was derived from an attenuated strain of *M. bovis* and was first administered to humans in 1921. Many BCG vaccines are available worldwide; all are derived from the original strain, but the vaccines vary in efficacy, ranging from 80% to nil in randomized, placebo-controlled trials. A similar range of efficacy was found in observational studies (case–control, historic cohort, and cross-sectional) in areas where infants are vaccinated at birth. These studies and a meta-analysis also found higher rates of efficacy in the protection of infants and young children from serious disseminated forms of childhood TB, such as tuberculous meningitis and miliary TB. BCG vaccine is safe and rarely causes serious complications. The local tissue response begins 2–3 weeks after vaccination, with scar formation and healing within 3 months. Side effects—most commonly, ulceration at the vaccination site and regional lymphadenitis—occur in 1–10% of vaccinated persons. Some vaccine strains have caused osteomyelitis in ~1 case per million doses administered. Disseminated BCG infection ("BCGitis") and death have occurred in 1–10 cases per 10 million doses administered, although this problem is restricted almost exclusively to persons with impaired immunity, such as children with severe combined immunodeficiency syndrome or adults with HIV infection. BCG vaccination induces TST reactivity, which tends to wane with time. The presence or size of TST reactions after vaccination does not predict the degree of protection afforded.

BCG vaccine is recommended for routine use at birth in countries or among populations with high TB prevalence. However, because of the low risk of transmission of TB in the United States and other high-income countries, the variability in protection afforded by BCG, and its impact on the TST, the vaccine is not recommended for general use. HIV-infected adults and children should not receive BCG vaccine. Moreover, infants whose HIV status is unknown but who have signs and symptoms consistent with HIV infection or who are born to HIV-infected mothers should not receive BCG.

Over the past decade, renewed research and development efforts have been made toward a new TB vaccine, and several candidates have been developed and tested. The MVA-85A vaccine (a modified poxvirus-vectored vaccine that expresses the immune-dominant *M. tuberculosis* antigen 85A), developed at the University of Oxford, was the first new TB vaccine to be tested in a phase 2B proof-of-concept trial in infants in South Africa. The aim was to evaluate the efficacy of a new preventive TB vaccine candidate against clinical TB or *M. tuberculosis* infection. Results were published in early 2013: MVA-85A was well tolerated and modestly immunogenic but did not confer significant protection against clinical TB or *M. tuberculosis* infection.

As of late 2017, 12 candidate vaccines were in various stages of clinical trials. They included whole-cell or mycobacterial whole-cell or lysates, viral vector vaccines, and adjuvant recombinant protein vaccines. Several challenges must be faced in the development of a TB vaccine. For instance, the lack of predictive animal models and protection correlates renders trials long and expensive. Furthermore, the decision about whether a candidate vaccine should be developed for prevention of infection (pre-exposure) or prevention of reactivation (post-exposure) without an exact understanding of its precise mechanism of action is complex. Therefore, introduction of a new vaccine on a large scale is not likely in the near future. This step will require an intensified and much larger investment in research and development.

TREATMENT

TREATMENT

Latent Tuberculosis Infection

It is estimated that 1.7 billion people—more than one-quarter of the human population—have been infected with *M. tuberculosis.* Although only a small fraction of these infections will progress toward active disease in a lifetime, new active cases will continue to emerge from this pool of "latently" infected individuals. Unfortunately, at present, there is no gold-standard diagnostic test that can confirm true infection (as opposed to immunologic memory of previous exposure) or predict which individuals with LTBI will develop active TB. Therefore, latently infected individuals among persons in defined high-risk groups can only be presumably identified by TST or IGRA. For skin testing, five tuberculin units of polysorbate-stabilized PPD should be injected intradermally into the volar surface of the forearm (i.e., the Mantoux method). Multipuncture tests are not recommended. Reactions are read at 48–72 h as the transverse diameter (in millimeters) of induration; the diameter of erythema is not considered. In some persons, TST reactivity wanes with time but can be recalled by a second skin test administered ≥1 week after the first (i.e., two-step testing). For persons periodically undergoing the TST, such as health care workers and individuals admitted to long-term-care institutions, initial two-step testing may preclude subsequent misclassification of those who have boosted reactions as TST converters. The cutoff for a positive TST (and thus for treatment) is related both to the probability that the reaction represents true infection and to the likelihood that the individual, if truly infected, will develop TB. **Table 173-5** suggests possible conventional cutoff by risk group. Thus, positive reactions for persons with HIV infection, recent close contacts of infectious cases, organ transplant recipients, previously untreated persons whose chest radiograph shows fibrotic lesions consistent with old TB, and persons receiving drugs that suppress the immune system are defined as an area of induration ≥5 mm in diameter A 10-mm cutoff is used to define positive reactions in most other at-risk persons. For persons with a very low risk of developing TB if infected, a cutoff of 15 mm is used. (Except for employment purposes where longitudinal screening is anticipated, the TST is not indicated for these low-risk persons.) A positive IGRA is based on the manufacturer's recommendations; however, good clinical practice requires that epidemiologic and clinical factors also guide the decision to implement treatment for LTBI and that active TB be definitively excluded before the initiation of chemoprophylaxis. The WHO recommends systematic testing for and treatment of LTBI for the following high-risk groups: people living with HIV, adult and child contacts of patients with infectious pulmonary TB, patients preparing for organ or hematologic transplantation, patients with silicosis, patients starting anti-TNF treatment, and patients on dialysis. Systematic testing for and treatment of LTBI should also be considered for prisoners, health care workers, immigrants from countries with a high TB burden, homeless persons, and illicit drug users.

Some TST- and IGRA-negative individuals are also candidates for treatment. Once an appropriate clinical evaluation has excluded active TB, infants and children who have come into contact with infectious cases should be treated for presumed LTBI. HIV-infected persons who have been exposed to an infectious TB patient should receive treatment regardless of the TST result. Any HIV-infected candidate for LTBI treatment must be screened carefully to exclude active TB, which would necessitate full treatment. The use of a clinical algorithm based on four signs/symptoms (current cough, fever, weight loss, and night sweats) helps to define which HIV-infected person is a candidate for LTBI treatment. The absence of all four symptoms tends to exclude active TB. The presence of one of these four manifestations, on the other hand, warrants further investigation for active TB before treatment of LTBI is started. Although a TST is prudent, this test is not an absolute requirement—given the logistical challenges—among people living with HIV in high-TB-incidence and low-resource settings.

Among people living with HIV and receiving ART, conversion of the TST from negative to positive can occur during the first few months of treatment. Conversions (from negative to positive) and reversions (from positive to negative) are more common with IGRAs than with TSTs among serially tested health care workers in the United States.

Treatment of selected persons with LTBI aims at preventing active disease. Potential candidates for treatment of LTBI are listed in Table 173-5. This intervention (*preventive chemotherapy* or *chemoprophylaxis*) is based on the results of a large number of randomized, placebo-controlled clinical trials demonstrating that a 6- to 9-month course of isoniazid reduces the risk of active TB in infected people by up to 90%. Analysis of available data indicated that the optimal duration of treatment with this drug was ~9 months. In the absence of reinfection, the protective effect is believed to be lifelong. Clinical trials have shown that isoniazid reduces rates of TB among TST-positive persons with HIV infection. Studies in HIV-infected patients have also demonstrated the effectiveness of shorter courses of rifampin-based treatment. Several regimens **(Table 173-6)** can be used to treat LTBI. The most widely used is that based on isoniazid alone at a daily dose of 5 mg/kg (up to 300 mg/d) for 9 months. On the basis of cost–benefit analyses and concerns about feasibility, a 6-month period of treatment is currently recommended by the WHO. Isoniazid can be administered intermittently (twice weekly) at a dose of 15 mg/kg (up to 900 mg) but only as DOT. An alternative regimen for adults is 3–4 months of daily rifampin. A 3- to 4-month regimen of daily isoniazid and rifampin is used in some countries (e.g., the United Kingdom) for both adults and children who are known not to have HIV infection. A previously recommended 2-month regimen of rifampin and pyrazinamide has been associated with serious or even fatal hepatotoxicity and is not recommended. The rifampin-containing regimens should be considered for persons who are likely to have been infected with an isoniazid-resistant strain. A clinical trial showed that a regimen of isoniazid (900 mg)

and rifapentine (900 mg), given once weekly for 12 weeks, is as effective as the standard 9-month isoniazid regimen. This regimen was associated with higher rates of treatment completion (82% vs 69%) and less hepatotoxicity (0.4% vs 2.7%) than isoniazid alone, although the rate of permanent discontinuation due to an adverse event was higher (4.9 vs 3.7%).

Currently, the isoniazid–rifapentine regimen is not recommended for children <2 years of age or pregnant women. Rifampin and rifapentine are contraindicated in HIV-infected individuals receiving protease inhibitors and most nonnucleoside reverse transcriptase inhibitors. However, efavirenz can be used for simultaneous administration with a rifamycin. Clinical trials to assess the efficacy of long-term isoniazid administration (i.e., for at least 3 years) among people living with HIV in high-TB-transmission settings have shown that this regimen can be more effective than 9 months of isoniazid and is therefore recommended under those circumstances. Isoniazid should not be given to persons with active liver disease. All isoniazid recipients at increased risk of hepatotoxicity (e.g., those abusing alcohol daily and those with a history of liver disease) should undergo baseline and then monthly assessment of liver function; they should be carefully educated about hepatitis and instructed to discontinue use of the drug immediately should any symptoms develop. Moreover, these patients should be seen and questioned monthly during therapy about adverse reactions and should be given no more than a 1-month supply of drug at each visit.

Treatment of LTBI among persons likely to have been infected by a multidrug-resistant strain is a challenge because no regimens have yet been tested in clinical trials. Close observation for early signs of disease is one option. However, in selected high-risk household contacts of patients with MDR-TB (e.g., children, recipients of immunosuppressive therapy), preventive therapy may be considered on the basis of individualized risk assessment and clinical criteria. In the absence of evidence of efficacy of any regimen, important factors in the decision to treat include intensity of exposure, certainty about a source case, information on the drug resistance pattern of the index case, and potential adverse events. Drug selection should be based on the drug susceptibility profile of the index case. Confirmation of infection with LTBI testing is required.

It may be more difficult to ensure compliance when treating persons with LTBI than when treating those with active TB. If family members of active cases are being treated, compliance and monitoring may be easier. When feasible, supervised therapy may increase the likelihood of completion. As in active cases, the provision of incentives may also be helpful. Currently, no evidence shows that LTBI treatment leads to significant development of drug resistance. However, before treatment of LTBI begins, it is mandatory to carefully exclude active TB in order to prevent the development of resistance.

TABLE 173-5

Tuberculin Reaction Size and Treatment of Latent *Mycobacterium tuberculosis* Infection

Risk Group	Tuberculin Reaction Size, mm
HIV-infected persons	≥5
Recent contacts of a patient with TB	≥5[a]
Organ transplant recipients	≥5
Persons with fibrotic lesions consistent with old TB on chest radiography	≥5
Persons who are immunosuppressed—e.g., due to the use of glucocorticoids or tumor necrosis factor α inhibitors	≥5
Persons with high-risk medical conditions[b]	≥5
Recent immigrants (≤5 years) from high-prevalence countries	≥10
Injection drug users	≥10
Mycobacteriology laboratory personnel; residents and employees of high-risk congregate settings[c]	≥10
Children <5 years of age; children and adolescents exposed to adults in high-risk categories	≥10
Low-risk persons[d]	≥15

[a]Tuberculin-negative contacts, especially children, should receive prophylaxis for 2–3 months after contact ends and should then undergo repeat tuberculin skin testing (TST). Those whose results remain negative should discontinue prophylaxis. HIV-infected contacts should receive a full course of treatment regardless of TST results. [b]These conditions include silicosis and end-stage renal disease managed by hemodialysis. [c]These settings include correctional facilities, nursing homes, homeless shelters, and hospitals and other health care facilities. [d]Except for employment purposes where longitudinal TST screening is anticipated, TST is not indicated for these low-risk persons. A decision to treat should be based on individual risk/benefit considerations.

Source: Adapted from Centers for Disease Control and Prevention: TB elimination—treatment options for latent tuberculosis infection (2011). Available at *http://www.cdc.gov/tb/publications/factsheets/testing/skintestresults.pdf.*

TABLE 173-6

Recommended Regimens and Drug Dosages for Treatment of Latent *Mycobacterium tuberculosis* Infection[a]

Regimen	Dose	Adverse Events
Isoniazid alone for 6 or 9 months	Adults: 5 mg/kg (max, 300 mg) per day Children: 10 mg/kg per day	Drug-induced liver injury, nausea, vomiting, abdominal pain, skin rash, peripheral neuropathy, dizziness, drowsiness, seizure
Rifampin alone for 3–4 months	Adults: 10 mg/kg per day Children: 10 mg/kg (max: <45 kg, 450 mg; >45 kg, 600 mg) per day	Flu-like syndrome, skin rash, drug-induced liver injury, anorexia, nausea, abdominal pain, neutropenia, thrombocytopenia, renal reactions (e.g., acute tubular necrosis and interstitial nephritis)
Isoniazid plus rifampin for 3–4 months	As above	As above
Rifapentine plus isoniazid for 3 months	Adults and children: Isoniazid: 15 mg/kg (900 mg) weekly Rifapentine: 15–30 mg/kg (900 mg) weekly	Hypersensitivity reactions, petechial skin rash, drug-induced liver injury Anorexia, nausea, abdominal pain Hypotensive reactions

[a]See text for full description of evidence on and limitations of these regimens.

Source: World Health Organization.

PRINCIPLES OF TB CONTROL

The highest priority in any TB control program is the prompt detection of cases and the provision of chemotherapy to all TB patients under proper case-management conditions, including DOT and social support. In addition, screening of high-risk groups, including immigrants from high-prevalence countries, migrant workers, prisoners, homeless individuals, substance abusers, and HIV-seropositive persons, is recommended. TST- or IGRA-positive high-risk persons should be treated for LTBI as described above. Contact investigation is an important component of efficient TB control. In the United States and other countries worldwide, a great deal of attention has been given to the transmission of TB (particularly in association with HIV infection) in institutional settings such as hospitals, homeless shelters, and prisons. Measures to limit such transmission include respiratory isolation of persons with suspected TB until they are proven to be noninfectious (at least by sputum AFB smear negativity), proper ventilation in rooms of patients with infectious TB, use of ultraviolet irradiation in areas of increased risk of TB transmission, and periodic screening of personnel who may come into contact with known or unsuspected cases of TB. In the past, radiographic surveys, especially those conducted with portable equipment and miniature films, were advocated for case finding. Today, however, the prevalence of TB in industrialized countries is sufficiently low that “mass miniature radiography” is not cost-effective.

In high-prevalence countries, most TB control programs have made remarkable progress in reducing morbidity and mortality since the mid-1990s by adopting and implementing the standards and strategies internationally promoted by the WHO. Between 2000 and 2016, an estimated 52.5 million lives were saved. The essential elements of good TB care and control were established in the mid-1990s and consist of well-defined interventions that were the basis of the “DOTS strategy”: early detection of cases and bacteriologic confirmation of the diagnosis; administration of standardized short-course chemotherapy, with direct supervision to ensure adherence to treatment and the provision of social support to patients; availability of drugs of proven quality, with an effective supply and management system; and a monitoring and evaluation system, including assessment of treatment outcomes—e.g., cure, completion of treatment without bacteriologic proof of cure, death, treatment failure, and default—in all cases registered and notified as well as measurement of the impact of control methods on classical TB indicators such as mortality, incidence, prevalence, and drug resistance. In 2006, the WHO indicated that, besides pursuing these essential elements that remain the fundamental components of any control strategy, additional steps had to be undertaken in order to reach international TB control targets. These steps included addressing HIV-associated TB and MDR-TB with additional measures; operating in harmony with general health services; engaging all care providers beyond the public providers; empowering people with TB and their communities; and enabling and promoting research. Evidence-based International Standards for Tuberculosis Care—focused on diagnosis, treatment, and public health responsibilities—were introduced for wide adoption by medical and professional societies, academic institutions, and all practitioners worldwide.

Care and control of HIV-associated TB are particularly challenging in poor countries because existing interventions require collaboration between HIV/AIDS and TB programs as well as standard services. TB programs must test every patient for HIV in order to provide access to trimethoprim-sulfamethoxazole

prophylaxis against common infections and ART. HIV/AIDS programs must regularly screen persons living with HIV/AIDS for active TB, provide treatment for LTBI, and ensure infection control in settings where people living with HIV congregate.

Early and active case detection is considered an important intervention not only among persons living with HIV/AIDS but also among other vulnerable populations, as it reduces transmission in a community and provides early effective care. Additional measures are indicated for the management of MDR-TB, RR-TB, and other forms of drug-resistant TB; they include upgrades of laboratory capacity to perform rapid DST and ensure surveillance of drug resistance; availability of drug regimens that are recommended for RR/MDR-TB, with assured quality of drugs; and infection control measures in all settings where patients with drug-resistant forms of TB may congregate. In the new era of the United Nations Sustainable Development Goals (2016–2030), the approach to TB control and care needs to evolve further and become multisectoral and more holistic. Engagement beyond dedicated programs and even the health sector is now essential. Therefore, the new “End TB” strategy promoted by WHO since 2016 builds on three pillars and relies on increased investments and efforts by all governments, their national programs, and a multitude of partners within and beyond the health sector: (1) integrated, patient-centered care and prevention; (2) bold policies and supportive systems; and (3) intensified research and innovation. The first pillar incorporates all technological innovations, such as early diagnostic approaches (including universal DST and systematic screening of identified, setting-specific, high-risk groups); well-designed treatment regimens for all forms of TB; proper management of HIV-associated TB and other comorbidities; and preventive treatment of persons at high risk. The second pillar is fundamental and is normally beyond the scope of dedicated programs, relying on policies forged by the highest-level health and governmental authorities: availability of adequate human and financial resources; engagement of civil organizations and all relevant public and private providers to pursue proper care for all patients and prevention for all people at risk; a policy of universal health coverage (which, together with social protection, implies avoidance of catastrophic expenditures caused by TB among the poorest); regulatory frameworks for case notifications, vital registration, quality and rational use of medicines, and infection control; social protection mechanisms as part of poverty alleviation strategies; and promotion of interventions against the broader determinants of TB. Finally, the third pillar of the new strategy emphasizes intensification of research and development on new tools and interventions as well as optimal implementation and rapid adoption of new tools in endemic countries. Besides specific clinical care and control interventions as described in this chapter, elimination of TB in a society ultimately will require control and mitigation of the multitude of direct risk factors (e.g., HIV infection, smoking, alcohol abuse, diabetes) and socioeconomic determinants (e.g., extreme poverty, inadequate living conditions and poor housing, malnutrition, indoor air pollution) with clearly implemented policies within the health sector and other sectors linked to human development and welfare.

FURTHER READING

Falzon D et al: World Health Organization treatment guidelines for drug-resistant tuberculosis, 2016 update. Eur Respir J 49:1602308, 2017.

Getahun H et al: Latent *Mycobacterium tuberculosis* infection. N Engl J Med 372:2127, 2015.

Nahid P et al: Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America clinical practice guidelines: Treatment of drug-susceptible tuberculosis. Clin Infect Dis 63:853, 2016.

Pai M et al: Tuberculosis. Nat Rev Dis Primers 2:16076, 2016.

Tameris MD et al: Safety and efficacy of MVA85A, a new tuberculosis vaccine, in infants previously vaccinated with BCG: A randomized, placebo-controlled phase 2b trial. Lancet 381:1021, 2013.

Uplekar M et al: WHO’s new end TB strategy. Lancet 385:1799, 2015.

WEBSITES

World Health Organization: Guidance for national tuberculosis programmes on the management of tuberculosis in children. 2nd ed. Geneva, WHO, 2014. Available from http://www.who.int/tb/publications/childtb_guidelines/en/. Accessed December 12, 2017.

World Health Organization: Global tuberculosis report 2017. Geneva, WHO, 2017. Available from *http://www.who.int/tb/publications/global_report/en/*. Accessed December 12, 2017.

World Health Organization: Treatment of tuberculosis. Guidelines for treatment of drug-susceptible tuberculosis and patient care. 2017 update. Geneva, WHO, 2017. Available from *http://apps.who.int/iris/bitstream/10665/255052/1/9789241550000-eng.pdf?ua=1*. Accessed December 12, 2017.

Chapter 174: Leprosy

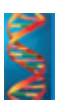
Robert H. Gelber

INTRODUCTION

Leprosy, first described in ancient Indian texts from the sixth century B.C., is a nonfatal, chronic infectious disease caused by *Mycobacterium leprae,* the clinical manifestations of which are largely confined to the skin, peripheral nervous system, upper respiratory tract, eyes, and testes. The unique tropism of *M. leprae* for peripheral nerves (from large nerve trunks to microscopic dermal nerves) and certain immunologically mediated reactional states are the major causes of morbidity in leprosy. The propensity of the disease, when untreated, to result in characteristic deformities and the recognition in most cultures that the disease is communicable from person to person have resulted historically in a profound social stigma. Today, with early diagnosis and the institution of appropriate and effective antimicrobial therapy, patients can lead productive lives in the community, and deformities and other visible manifestations can largely be prevented.

ETIOLOGY

M. leprae is an obligate intracellular bacillus (0.3–1 µm wide and 1–8 µm long) that is confined to humans, armadillos in certain locales, and sphagnum moss. The organism is acid-fast, indistinguishable microscopically from other mycobacteria, and ideally detected in tissue sections by a modified Fite stain. Strain variability has been documented in this organism. *M. leprae* produces no known toxins and is well adapted to penetrate and reside within macrophages, yet it may survive outside the body for months. In untreated patients, only ~1% of *M. leprae* organisms are viable. The morphologic index (MI), a measure of the number of acid-fast bacilli (AFB) in skin scrapings that stain uniformly bright, correlates with viability. The bacteriologic index (BI), a logarithmic-scaled measure of the density of *M. leprae* in the dermis, may be as high as 4–6+ in untreated patients and falls by 1 unit per year during effective antimicrobial therapy; the rate of decrease is independent of the relative potency of therapy. A rising MI or BI suggests relapse and perhaps —if the patient is being treated—drug resistance. Drug resistance can be confirmed or excluded in the mouse model of leprosy, and resistance to dapsone and rifampin can be documented by the recognition of mutant genes. However, the availability of these technologies is extremely limited.



As a result of reductive evolution, almost half of the *M. leprae* genome contains nonfunctional genes; only 1605 genes encode for proteins, and 1439 genes are shared with *Mycobacterium tuberculosis*. In contrast, *M. tuberculosis* uses 91% of its genome to encode for 4000 proteins. Among the lost genes in *M.*

leprae are those for catabolic and respiratory pathways; transport systems; purine, methionine, and glutamine synthesis; and nitrogen regulation. The genome of *M. leprae* provides a metabolic rationale for its obligate intracellular existence and reliance on host biochemical support, a template for targets of drug development, and ultimately a pathway to cultivation. The finding of strain variability among *M. leprae* isolates has provided a powerful tool with which to address anew the organism's epidemiology and pathobiology and to determine whether relapse represents reactivation or reinfection. The bacterium's complex cell wall contains large amounts of an *M. leprae*–specific phenolic glycolipid (PGL-1), which is detected in serologic tests. The unique trisaccharide of *M. leprae* binds to the basal lamina of Schwann cells; this interaction is probably related to the ability of *M. leprae*—unique among bacteria—to invade peripheral nerves.

Although it was the first bacterium to be etiologically associated with human disease, *M. leprae* remains one of the few bacterial species that still has not been cultivated on artificial medium or tissue culture. The regular multiplication of even a few *M. leprae* organisms in mouse footpads (albeit limited, with a doubling time of ~2 weeks) has provided a sensitive means to evaluate antimicrobial agents, monitor clinical trials, and screen vaccines. Several in vitro methods to assess the organisms' viability, though promising, are many times less sensitive than the mouse model in detecting viable *M. leprae*. *M. leprae* grows best in cooler tissues (the skin, peripheral nerves, anterior chamber of the eye, upper respiratory tract, and testes), sparing warmer areas of the skin (the axilla, groin, scalp, and midline of the back).

Another distinct and recently discovered mycobacterial species, *M. lepromatosis,* is genetically similar to *M. leprae* and evolved from a common mycobacterial ancestor. *M. lepromatosis* has been identified in tissue from a small number of leprosy patients in Mexico with diffuse lepromatosis/Lucio's phenomenon (see below) and in single leprosy patients from Singapore and Canada. Of six Mexican patients studied, four were infected with *M. leprae* and two with *M. lepromatosis*. Because some new leprosy patients harbor both *M. leprae* and *M. lepromatosis*, it is not entirely clear that the latter organism is, in fact, a causative agent of leprosy. Fortunately, like *M. leprae, M. lepromatosis* is generally sensitive to dapsone, rifampin, and fluoroquinolones.

EPIDEMIOLOGY

Demographics


Leprosy is almost exclusively a disease of the developing world, affecting areas of Asia, Africa, Latin America, and the Pacific. While Africa has the highest disease prevalence, Asia has the most cases. More than 80% of the world's cases occur in a few countries: India, China, Myanmar, Indonesia, Brazil, Nigeria, Madagascar, and Nepal. Within endemic locales, the distribution of leprosy is quite uneven, with areas of high prevalence bordering on areas with little or no disease. In Brazil the majority of cases occur in the Amazon basin and two western states, while in Mexico leprosy is mostly confined to the Pacific coast. Except as imported cases, leprosy is largely absent from the United States, Canada, and northwestern Europe. In the United States, ~4000 persons have leprosy and 100–200 new cases are reported annually, most of them in

California, Texas, New York, and Hawaii among immigrants from Mexico, Southeast Asia, the Philippines, and the Caribbean.

 The comparative genomics of single-nucleotide polymorphisms support the likelihood that four distinct strains exist, having originated in East Africa or Central Asia. A mutation spread to Europe and subsequently underwent two separate mutations that were then followed by spread to West Africa and the Americas.

The global prevalence of leprosy is difficult to assess, given that many of the locales with high prevalence lack a significant medical or public health infrastructure. Estimates range from 0.6 to 8 million affected individuals. The lower estimate includes only persons who have not completed chemotherapy, excluding those who may be physically or psychologically damaged from leprosy and who may yet relapse or develop immune-mediated reactions. The higher figure includes patients whose infections probably are already cured and many who have no leprosy-related deformity or disability. Although the figures on the worldwide prevalence of leprosy are debatable, incidence is not falling; there are still an estimated 500,000 new cases annually.

Leprosy is associated with poverty and rural residence. It appears not to be associated with AIDS, perhaps because of leprosy's long incubation period. Most individuals appear to be naturally immune to leprosy and do not develop disease manifestations after exposure. The time of peak onset is in the second and third decades of life.

 The most severe lepromatous form of leprosy is twice as common among men as among women and is rarely encountered in children. The frequency of the polar forms of leprosy in different countries varies widely and may in part be genetically determined; certain human leukocyte antigen (HLA) associations are known for both polar forms of leprosy (see below). Furthermore, variations in immunoregulatory genes are associated with an increased susceptibility to leprosy, particularly the multibacillary form. In India and Africa, 90% of cases are tuberculoid; in Southeast Asia, 50% are tuberculoid and 50% lepromatous; and in Mexico, 90% are lepromatous. (For definitions of disease types, see **Table 174-1** and "Clinical, Histologic, and Immunologic Spectrum," below.)

TABLE 174-1

Clinical, Bacteriologic, Pathologic, and Immunologic Spectrum of Leprosy

Feature	Tuberculoid (TT, BT) Leprosy	Borderline (BB, BL) Leprosy	Lepromatous (LL) Leprosy
Skin lesions	One or a few sharply defined annular asymmetric macules or plaques with a tendency toward central clearing, elevated borders	Intermediate between BT- and LL-type lesions; ill-defined plaques with an occasional sharp margin; few or many in number	Symmetric, poorly marginated, multiple infiltrated nodules and plaques or diffuse infiltration; xanthoma-like or dermatofibroma papules; leonine facies and eyebrow alopecia
Nerve lesions	Skin lesions anesthetic early; nerve near lesions sometimes enlarged; nerve abscesses most common in BT	Hypesthetic or anesthetic skin lesions; nerve trunk palsies, at times symmetric	Hypesthesia a late sign; nerve palsies variable; acral, distal, symmetric anesthesia common
Acid-fast bacilli (BI[a])	0–1+	3–5+	4–6+
Lymphocytes	2+	1+	0–1+
Macrophage differentiation	Epithelioid	Epithelioid in BB; usually undifferentiated but may have foamy changes in BL	Foamy changes the rule; may be undifferentiated in early lesions
Langerhans giant cells	1–3+	—	—
Lepromin skin test	+++	—	—
Lymphocyte transformation test	Generally positive	1–10%	1–2%

Feature	Tuberculoid (TT, BT) Leprosy	Borderline (BB, BL) Leprosy	Lepromatous (LL) Leprosy
CD4+/CD8+ T cell ratio in lesions	1.2	BB: NT; BL: 0.48	0.50
M. leprae PGL-1 antibodies	60%	85%	95%

[a]See text.

Abbreviations: BB, mid-borderline; BL, borderline lepromatous; BT, borderline tuberculoid; TT, polar tuberculoid; LL, polar lepromatous; BI, bacteriologic index; NT, not tested; PGL-1, phenolic glycolipid 1.

Transmission

The route of transmission of leprosy remains uncertain, and transmission routes may in fact be multiple. Nasal droplet infection, contact with infected soil, and amoeba insect vectors have been considered the prime candidates. Aerosolized *M. leprae* can cause infection in immunosuppressed mice, and a sneeze from an untreated lepromatous patient may contain $>10^{10}$ AFB. Furthermore, both IgA antibody to *M. leprae* and genes of *M. leprae*—demonstrable by polymerase chain reaction (PCR)—have been found in the nose of individuals from endemic areas who have no signs of leprosy and in 19% of occupational contacts of lepromatous patients. Several lines of evidence implicate soil transmission. (1) In endemic countries such as India, leprosy is primarily a rural and not an urban disease. (2) *M. leprae* products reside in soil in endemic locales. (3) Direct dermal inoculation (e.g., during tattooing) may transmit *M. leprae*, and common sites of leprosy in children are the buttocks and thighs, suggesting that microinoculation of infected soil may transmit the disease. Evidence for insect vectors of leprosy includes the demonstration that bedbugs and mosquitoes in the vicinity of leprosaria regularly harbor *M. leprae* and that experimentally infected mosquitoes can transmit the infection to mice. Red squirrels from Brownsea Island, England, have recently been found to be commonly infected with a strain of *M. leprae* that circulated in leprosy patients in medieval England. In addition, some red squirrels in England, Scotland, and Ireland have been found to be infected with *M. lepromatosis*. Both of these mycobacterial species have been found in overtly diseased red squirrels and in animals that appeared to be well. It is unclear what role, if any, these zoonoses might play in the propagation of human leprosy. Skin-to-skin contact generally is not considered an important route of transmission.

In endemic countries, ~50% of leprosy patients have a history of intimate contact with an infected person (often a household member), while, for unknown reasons, leprosy patients in nonendemic locales can identify such contact only 10% of the time. Moreover, household contact with an infected lepromatous case

carries an eventual risk of disease acquisition of ~10% in endemic areas as opposed to only 1% in nonendemic locales. Contact with a tuberculoid case carries a very low risk. Physicians and nurses caring for leprosy patients and the co-workers of these patients are not at risk for leprosy.


Although multilocus variable-number short-nucleotide tandem-repeat (VNTR) analyses have generally demonstrated considerable variability among isolates, highly similar and even identical VNTR results have been obtained with isolates from a limited number of families with multiple cases. Moreover, VNTR results have been similar for isolates within certain geographic locales and divergent for isolates within others. These findings suggest that genomic analyses may prove useful in the future for defining *M. leprae* transmission patterns.

M. leprae causes disease primarily in humans. However, in Texas and Louisiana, 15% of nine-banded armadillos are infected, and armadillo contact occasionally results in human disease. Armadillos develop disseminated infection after IV inoculation of live *M. leprae*.

CLINICAL, HISTOLOGIC, AND IMMUNOLOGIC SPECTRUM

The incubation period prior to manifestation of clinical disease can vary between 2 and 40 years, although it is generally 5–7 years in duration. This long incubation period is probably, at least in part, a consequence of the extremely long doubling time for *M. leprae* (14 days in mice versus in vitro doubling times of 1 day and 20 min for *M. tuberculosis* and *Escherichia coli*, respectively). Leprosy presents as a spectrum of clinical manifestations that have bacteriologic, pathologic, and immunologic counterparts. The spectrum from polar tuberculoid (TT) to borderline tuberculoid (BT) to mid-borderline (BB, which is rarely encountered) to borderline lepromatous (BL) to polar lepromatous (LL) disease is associated with an evolution from asymmetric localized macules and plaques to nodular and indurated symmetric generalized skin manifestations, an increasing bacterial load, and loss of *M. leprae*–specific cellular immunity (Table 174-1). Distinguishing dermatopathologic characteristics include the number of lymphocytes, giant cells, and AFB as well as the nature of epithelioid cell differentiation. Where a patient presents on the clinical spectrum largely determines prognosis, complications, reactional states, and the intensity of antimicrobial therapy required.

Tuberculoid Leprosy

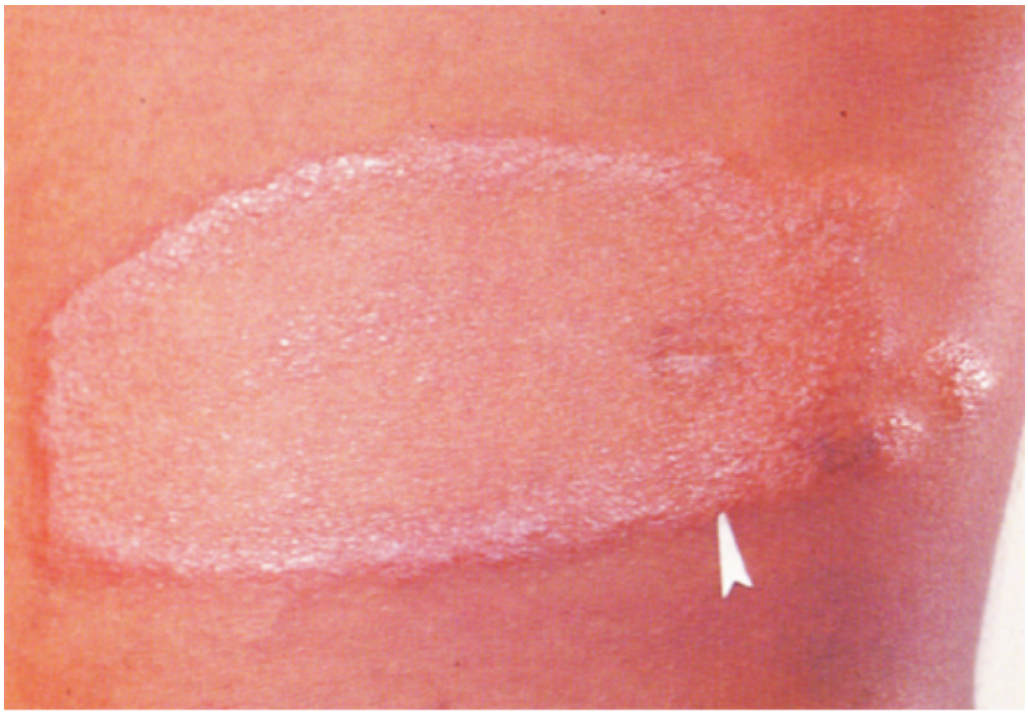

At the less severe end of the spectrum is tuberculoid leprosy, which encompasses TT and BT disease. In general, these forms of leprosy result in symptoms confined to the skin and peripheral nerves. TT/BT leprosy is the most common form encountered in India, and TT is most common in Africa, while TT is virtually absent in Southeast Asia, where BT leprosy is frequent.

The skin lesions of tuberculoid leprosy consist of one or a few hypopigmented macules or plaques **(Fig. 174-1)** that are sharply demarcated and hypesthetic, often have erythematous or raised borders, and are devoid of the normal skin organs (sweat glands and hair follicles) and thus are dry, scaly, and anhidrotic. AFB are

generally absent or few in number. Tuberculoid leprosy patients may have asymmetric enlargement of one or a few peripheral nerves. Indeed, leprosy and certain rare hereditary neuropathies are the only human diseases associated with peripheral-nerve enlargement. Although any peripheral nerve may be enlarged (including small digital and supraclavicular nerves), those most commonly affected are the ulnar, posterior auricular, peroneal, and posterior tibial nerves, with associated hypesthesia and myopathy.

FIGURE 174-1

Tuberculoid (TT) leprosy: a well-defined, hypopigmented, anesthetic macule with anhidrosis and a raised granular margin *(arrowhead).*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In tuberculoid leprosy, T cells breach the perineurium, and destruction of Schwann cells and axons may be evident, resulting in fibrosis of the epineurium, replacement of the endoneurium with epithelial granulomas, and occasionally caseous necrosis. Such invasion and destruction of nerves in the dermis by T cells are pathognomonic for leprosy.

Circulating lymphocytes from patients with tuberculoid leprosy readily recognize *M. leprae* and its constituent proteins, patients have positive lepromin skin tests (see "Diagnosis," below), and—because of a type 1 cytokine pattern in tuberculoid tissues—strong T cell and macrophage activation results in a localized infection. In tuberculoid leprosy tissue, there is a 2:1 predominance of helper CD4+ over CD8+ T lymphocytes. Tuberculoid tissues are rich in the mRNAs of the proinflammatory T_H1 family of cytokines: interleukin 2 (IL-2), interferon γ (IFN-γ), and IL-12; in contrast, IL-4, IL-5, and IL-10 mRNAs are scarce.

Lepromatous Leprosy

Lepromatous leprosy patients present with symmetrically distributed skin nodules **(Fig. 174-2)**, raised plaques, or diffuse dermal infiltration, which, when on the face, results in *leonine facies.* Late manifestations include loss of eyebrows (initially the lateral margins only) and eyelashes, pendulous earlobes, and dry scaling skin, particularly on the feet. In LL leprosy, bacilli are numerous in the skin (as many as 10^9/g), where they are often found in large clumps (*globi*), and in peripheral nerves, where they initially invade Schwann cells, resulting in foamy degenerative myelination and axonal degeneration and later in Wallerian degeneration. In addition, bacilli are plentiful in circulating blood and in all organ systems except the lungs and the central nervous system. Nevertheless, patients are afebrile, and there is no evidence of major organ system dysfunction.

FIGURE 174-2

Lepromatous (LL) leprosy: advanced nodular lesions.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Found almost exclusively in western Mexico and the Caribbean is a form of lepromatous leprosy without visible skin lesions but with diffuse dermal infiltration and a demonstrably thickened dermis, termed *diffuse lepromatosis*.

In lepromatous leprosy, nerve enlargement and damage tend to be symmetric, result from actual bacillary invasion, and are more insidious but ultimately more extensive than in tuberculoid leprosy. Patients with LL leprosy have acral, distal, symmetric peripheral neuropathy and a tendency toward symmetric nerve-trunk

enlargement. They may also have signs and symptoms related to involvement of the upper respiratory tract, the anterior chamber of the eye, and the testes.

In untreated LL patients, lymphocytes regularly fail to recognize either *M. leprae* or its protein constituents, and lepromin skin tests are negative (see "Diagnosis," below). This loss of protective cellular immunity appears to be antigen-specific, as patients are not unusually susceptible to opportunistic infections, cancer, or AIDS and maintain delayed-type hypersensitivity to *Candida*, *Trichophyton*, mumps virus, tetanus toxoid, and even purified protein derivative of tuberculin. At times, *M. leprae*–specific anergy is reversible with effective chemotherapy. In LL tissues, there is a 2:1 ratio of CD8+ to CD4+ T lymphocytes. LL patients have a predominant T_H2 response and hyperglobulinemia, and LL tissues demonstrate a T_H2 cytokine profile, being rich in mRNAs for IL-4, IL-5, and IL-10 and poor in those for IL-2, IFN-γ, and IL-12. It appears that cytokines mediate a protective tissue response in leprosy, as injection of IFN-γ or IL-2 into lepromatous lesions causes a loss of AFB and histopathologic conversion toward a tuberculoid pattern. Macrophages of lepromatous leprosy patients appear to be functionally intact; circulating monocytes exhibit normal microbicidal function and responsiveness to IFN-γ.

Reactional States

Lepra reactions comprise several common immunologically mediated inflammatory states that cause considerable morbidity. Some of these reactions precede diagnosis and the institution of effective antimicrobial therapy; indeed, these reactions may precipitate presentation for medical attention and diagnosis. Other reactions follow the initiation of appropriate chemotherapy; these reactions may cause patients to perceive that their leprosy is worsening and to lose confidence in conventional therapy. Only by warning patients of the potential for these reactions and describing their manifestations can physicians treating leprosy patients ensure continued credibility.

TYPE 1 LEPRA REACTIONS (DOWNGRADING AND REVERSAL REACTIONS)

Type 1 lepra reactions occur in almost half of patients with borderline forms of leprosy but not in patients with pure lepromatous disease. Manifestations include classic signs of inflammation within previously involved macules, papules, and plaques and, on occasion, the appearance of new skin lesions, neuritis, and (less commonly) fever—generally low-grade. The nerve trunk most frequently involved in this process is the ulnar nerve at the elbow, which may be painful and exquisitely tender. If patients with affected nerves are not treated promptly with glucocorticoids (see below), irreversible nerve damage may result in as little as 24 h. The most dramatic manifestation is footdrop, which occurs when the peroneal nerve is involved.

When type 1 lepra reactions precede the initiation of appropriate antimicrobial therapy, they are termed *downgrading reactions*, and the case becomes histologically more lepromatous; when they occur after the initiation of therapy, they are termed *reversal reactions*, and the case becomes more tuberculoid. Reversal reactions often occur in the first months or years after the initiation of therapy but may also develop several years thereafter.

Edema is the most characteristic microscopic feature of type 1 lepra lesions, whose diagnosis is primarily clinical. Reversal reactions are typified by a T_H1 cytokine profile, with an influx of CD4+ T helper cells and increased levels of IFN-γ and IL-2. In addition, type 1 reactions are associated with large numbers of T cells bearing γ/δ receptors—a unique feature of leprosy.

TYPE 2 LEPRA REACTIONS: ERYTHEMA NODOSUM LEPROSUM

Erythema nodosum leprosum (ENL) **(Fig. 174-3)** occurs exclusively in patients near the lepromatous end of the leprosy spectrum (BL/LL), affecting nearly 50% of this group. Although ENL may precede leprosy diagnosis and the initiation of therapy (sometimes, in fact, prompting the diagnosis), in 90% of cases it follows the institution of chemotherapy, generally within 2 years. The most common features of ENL are crops of painful erythematous papules that resolve spontaneously in a few days to a week but may recur; malaise; and fever that can be profound. However, patients may also experience neuritis, lymphadenitis, uveitis, orchitis, and glomerulonephritis and may develop anemia, leukocytosis, and abnormal liver function tests (particularly increased aminotransferase levels). Individual patients may have either a single bout of ENL or chronic recurrent manifestations. Bouts may be either mild or severe and generalized; in rare instances, ENL results in death. Skin biopsy of ENL papules reveals vasculitis or panniculitis, sometimes with many lymphocytes but characteristically with polymorphonuclear leukocytes as well.

FIGURE 174-3

Moderately severe skin lesions of erythema nodosum leprosum, some with pustulation and ulceration.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Elevated levels of circulating tumor necrosis factor (TNF) have been demonstrated in ENL; thus, TNF may play a central role in the pathobiology of this syndrome. ENL is thought to be a consequence of immune

complex deposition, given its T_H2 cytokine profile and its high levels of IL-6 and IL-8. However, in ENL tissue, the presence of HLA-DR framework antigen of epidermal cells—considered a marker for a delayed-type hypersensitivity response—and evidence of higher levels of IL-2 and IFN-γ than are usually seen in polar lepromatous disease suggest an alternative mechanism.

LUCIO'S PHENOMENON

Lucio's phenomenon is an unusual reaction seen exclusively in patients from the Caribbean and Mexico who have the diffuse lepromatosis form of lepromatous leprosy, most often those who are untreated. Patients with this reaction develop recurrent crops of large, sharply marginated, ulcerative lesions—particularly on the lower extremities—that may be generalized and, when so, are frequently fatal as a result of secondary infection and consequent septic bacteremia. Histologically, the lesions are characterized by ischemic necrosis of the epidermis and superficial dermis, heavy parasitism of endothelial cells with AFB, and endothelial proliferation and thrombus formation in the larger vessels of the deeper dermis. Like ENL, Lucio's phenomenon is probably mediated by immune complexes.

Complications

THE EXTREMITIES

Complications of the extremities in leprosy patients are primarily a consequence of neuropathy leading to insensitivity and myopathy. Insensitivity affects fine touch, pain, and heat receptors but generally spares position and vibration appreciation. The most commonly affected nerve trunk is the ulnar nerve at the elbow, whose involvement results in clawing of the fourth and fifth fingers, loss of dorsal interosseous musculature in the affected hand, and loss of sensation in these distributions. Median nerve involvement in leprosy impairs thumb opposition and grasp; radial nerve dysfunction, although rare in leprosy, leads to wrist drop. Tendon transfers can restore hand function but should not be performed until 6 months after the initiation of antimicrobial therapy and the conclusion of episodes of acute neuritis.

Plantar ulceration, particularly at the metatarsal heads, is probably the most common complication of leprous neuropathy. Therapy requires careful debridement; administration of appropriate antibiotics; avoidance of weight-bearing until ulcerations are healed, with slowly progressive ambulation thereafter; and wearing of special shoes to prevent recurrence.

Footdrop as a result of peroneal nerve palsy should be treated with a simple nonmetallic brace in the shoe or with surgical correction attained by tendon transfers. Although uncommon, Charcot's joints, particularly of the foot and ankle, may result from leprosy.

The loss of distal digits in leprosy is a consequence of insensitivity, trauma, secondary infection, and—in lepromatous disease—a poorly understood and sometimes profound osteolytic process. Conscientious protection of the extremities during cooking and work and the early institution of therapy have substantially reduced the frequency and severity of distal digit loss in recent times.

THE NOSE

In lepromatous leprosy, bacillary invasion of the nasal mucosa can result in chronic nasal congestion and epistaxis. Saline nose drops may relieve these symptoms. Long-untreated LL leprosy may further result in destruction of the nasal cartilage, with consequent saddle-nose deformity or anosmia (more common in the preantibiotic era than at present). Nasal reconstructive procedures can ameliorate significant cosmetic defects.

THE EYE

Arising from cranial nerve palsies, lagophthalmos and corneal insensitivity may complicate leprosy, resulting in trauma, secondary infection, and (without treatment) corneal ulcerations and opacities. For patients with these conditions, eye drops during the day and ointments at night provide some protection from such consequences. Furthermore, in LL leprosy, the anterior chamber of the eye is invaded by bacilli, and ENL may result in uveitis, with consequent cataracts and glaucoma. Thus leprosy is a major cause of blindness in the developing world. Slit-lamp evaluation of LL patients often reveals “corneal beading” that represents globi of *M. leprae*.

THE TESTES

M. leprae invades the testes, while ENL may cause orchitis. Thus males with lepromatous leprosy often manifest mild to severe testicular dysfunction, with an elevation of luteinizing and follicle-stimulating hormones, decreased testosterone, and aspermia or hypospermia in 85% of LL patients but in only 25% of BL patients. LL patients may become impotent and infertile. Impotence is sometimes responsive to testosterone replacement.

AMYLOIDOSIS

Secondary amyloidosis is a complication of LL leprosy and ENL that is encountered infrequently in the antibiotic era. This complication may result in abnormalities of hepatic and particularly renal function.

NERVE ABSCESSES

Patients with various forms of leprosy, but especially those with the BT form, may develop abscesses of nerves (most commonly the ulnar), with a cellulitic appearance of adjacent skin. In such conditions, the affected nerve is swollen and exquisitely tender. Although glucocorticoids may reduce signs of inflammation, rapid surgical decompression is necessary to prevent irreversible sequelae.

DIAGNOSIS

Leprosy most commonly presents with both characteristic skin lesions and skin histopathology. Thus the disease should be suspected when a patient from an endemic area has suggestive skin lesions or peripheral neuropathy. The diagnosis should be confirmed by histopathology. In tuberculoid leprosy, lesional areas—

preferably the advancing edge—must be biopsied because normal-appearing skin does not have pathologic features. In lepromatous leprosy, nodules, plaques, and indurated areas are optimal biopsy sites, but biopsies of normal-appearing skin also are generally diagnostic. Lepromatous leprosy is associated with diffuse hyperglobulinemia, which may result in false-positive serologic tests (e.g., Venereal Disease Research Laboratory, rheumatoid arthritis, and antinuclear antibody tests) and therefore may cause diagnostic confusion. On occasion, tuberculoid lesions may not (1) appear typical, (2) be hypesthetic, and (3) contain granulomas (instead containing only nonspecific lymphocytic infiltrates). In such instances, two of these three characteristics are considered sufficient for a diagnosis. It is preferable to overdiagnose leprosy rather than to allow a patient to remain untreated.

IgM antibodies to PGL-1 are found in 95% of patients with untreated lepromatous leprosy; the titer decreases with effective therapy. However, in tuberculoid leprosy—the form of disease most often associated with diagnostic uncertainty because of the absence or paucity of AFB—patients have significant antibodies to PGL-1 only 60% of the time; moreover, in endemic locales, exposed individuals without clinical leprosy may harbor antibodies to PGL-1. Thus PGL-1 serology is of little diagnostic utility in tuberculoid leprosy. Heat-killed *M. leprae* (lepromin) has been used as a skin test reagent. It generally elicits a reaction in tuberculoid leprosy patients, may do so in individuals without leprosy, and gives negative results in lepromatous leprosy patients; consequently, it is likewise of little diagnostic value. Unfortunately, PCR of the skin for *M. leprae*, although positive in LL and BL disease, yields negative results in 50% of tuberculoid cases, again offering little diagnostic assistance.

DIFFERENTIAL DIAGNOSIS

Included in the differential diagnosis of lesions that resemble leprosy are sarcoidosis, leishmaniasis, lupus vulgaris, dermatofibroma, histiocytoma, lymphoma, syphilis, yaws, granuloma annulare, and various other disorders causing hypopigmentation (notably pityriasis alba, tinea, and vitiligo). Sarcoidosis may result in perineural inflammation, but actual granuloma formation within dermal nerves is pathognomonic for leprosy. In lepromatous leprosy, sputum specimens may be loaded with AFB—a finding that can be incorrectly interpreted as representing pulmonary tuberculosis.

TREATMENT

TREATMENT

Leprosy

ANTIMICROBIAL THERAPY

Active Agents

Established agents used to treat leprosy include dapsone (50–100 mg/d), clofazimine (50–100 mg/d, 100 mg three times weekly, or 300 mg monthly), and rifampin (600 mg daily or monthly); see "Choice of Regimens," below. Of these drugs, only rifampin is bactericidal. The sulfones (folate antagonists), the foremost of which

is dapsone, were the first antimicrobial agents found to be effective for the treatment of leprosy and are still the mainstays of therapy. With sulfone treatment, skin lesions resolve and numbers of viable bacilli in the skin are reduced. Although primarily bacteriostatic, dapsone monotherapy results in a resistance-related relapse rate of only 2.5%. When dapsone monotherapy was discontinued in lepromatous patients treated for ≥18 years who had been smear-negative for several years, relapses began to occur in the first year after cessation and occurred in ~1% annually thereafter during the next nine years (total, 10%). Dapsone is generally safe and inexpensive. Individuals with glucose-6-phosphate dehydrogenase deficiency who are treated with dapsone may develop severe hemolysis; those without this deficiency also have reduced red cell survival and a hemoglobin decrease averaging 1 g/dL. Dapsone's usefulness is limited occasionally by allergic dermatitis and rarely by the sulfone syndrome (including high fever, anemia, exfoliative dermatitis, and a mononucleosis-type blood picture). When rifampin has been included in finite regimens to treat multibacillary leprosy (including World Health Organization [WHO] multidrug therapy), several studies have documented double-digit relapse rates, particularly frequently in patients with a high BI. Relapses following the discontinuation of rifampin-containing regimens (unlike those following the discontinuation of dapsone monotherapy) generally begin only after 6 years and most commonly occur after >10 years. It must be remembered that rifampin induces microsomal enzymes, necessitating increased doses of medications such as glucocorticoids and oral birth control regimens. Clofazimine is often cosmetically unacceptable to light-skinned leprosy patients because it causes a red-black skin discoloration that accumulates, particularly in lesional areas, and makes the patient's diagnosis obvious to members of the community.

Other antimicrobial agents active against *M. leprae* in animal models and at the usual daily doses used in clinical trials include ethionamide/prothionamide; the aminoglycosides streptomycin, kanamycin, and amikacin (but not gentamicin or tobramycin); minocycline; clarithromycin; and several fluoroquinolones, particularly ofloxacin and moxifloxacin. After rifampin, the most bactericidal agents against *M. leprae* in mice and patients appear to be minocycline, clarithromycin, and ofloxacin, but these drugs have not been used extensively in leprosy control programs. Most recently, rifapentine and moxifloxacin have been found to be especially potent against *M. leprae* in mice. In a clinical trial in lepromatous leprosy, moxifloxacin was profoundly bactericidal, matched in potency only by rifampin.

Choice of Regimens

Antimicrobial therapy for leprosy should be individualized, depending on the clinical/pathologic form of the disease encountered. Tuberculoid leprosy, which is associated with a low bacterial burden and a protective cellular immune response, is the easiest form to treat and can be cured reliably with a finite course of chemotherapy. In contrast, lepromatous leprosy may have a higher bacillary load than any other human bacterial disease, and the absence of a salutary T cell repertoire requires prolonged or even lifelong chemotherapy. Therefore, careful classification of disease prior to therapy is important.

A reasoned approach to the treatment of leprosy is confounded by several issues:

1. Even without therapy, TT leprosy may heal spontaneously, and dapsone monotherapy is generally curative.

2. In tuberculoid disease, it is common for no bacilli to be found in the skin prior to therapy. Thus there is no objective measure of therapeutic success. Furthermore, despite adequate treatment, TT and particularly BT lesions often resolve minimally or incompletely, while relapse and late type 1 lepra reactions can be difficult to distinguish.

3. LL leprosy patients commonly harbor viable *M. leprae* "persisters" that are the source of relapse if therapy is discontinued. Because leprosy may relapse many years after cessation of antibiotic therapy, prolonged follow-up after completion of treatment is recommended in order to prevent further disability and deformity.

4. Even though primary dapsone resistance is exceedingly rare and multidrug therapy is generally recommended (at least for lepromatous leprosy), there is a paucity of information from experimental animals and clinical trials on the optimal combination of antimicrobial agents, dosing schedule, and duration of therapy.

In 1982, the WHO made recommendations for leprosy chemotherapy administered in control programs. These recommendations recognized the limited resources available for leprosy care in the very areas where it is most prevalent and the frustration and discouragement of patients and program managers with the previous requirement for lifelong therapy for many leprosy patients. Thus, for the first time, and without supporting clinical-trial evidence (particularly data on long-term relapse frequency), the WHO advocated a finite duration of therapy for all forms of leprosy and—given the prohibitive cost of daily rifampin treatment in developing countries—encouraged the monthly administration of this agent as part of a multidrug regimen. The WHO treatment regimens were specifically meant for control programs where implementation of a finite regimen for all forms of leprosy would substantially decrease the operational burden of leprosy care; these regimens were not claimed to be optimal in locales where more considerable resources are available. Over the ensuing years, however, the WHO recommendations have been broadly implemented worldwide. For treatment purposes, the WHO originally classified patients with few bacteria in the dermis (BI <2) as paucibacillary and those with many bacteria (BI >2) as multibacillary. The WHO recommended that paucibacillary leprosy in adults be treated with 100 mg of dapsone daily and 600 mg of rifampin monthly (supervised) for 6 months **(Table 174-2)**. As an alternative for patients with single-lesion paucibacillary leprosy, the WHO recommended a single dose of rifampin (600 mg), ofloxacin (400 mg), and minocycline (100 mg). The recommendation for multibacillary leprosy in adults was 100 mg of dapsone plus 50 mg of clofazimine daily (unsupervised) and with 600 mg of rifampin plus 300 mg of clofazimine monthly (supervised). Originally, the WHO recommended that lepromatous patients be treated for 2 years or until smears became negative (generally in ~5 years). In subsequent years, the WHO mandated several modifications to their original multidrug-treatment recommendation, primarily related to leprosy classification criteria and treatment duration (see "The Leprosy Elimination Campaign: Modifications of the Multidrug Regimen and Their Consequences," below).

Several factors have caused many authorities to question the WHO recommendations and to favor a more intensive approach. Among these factors are—for multibacillary patients—a high (double-digit) relapse rate in several locales (reaching 20–40% in one locale, with the rate directly related to the initial bacterial burden)

and—for paucibacillary patients—demonstrable lesional activity for years in fully half of patients after the completion of therapy. The more intensive approach (Table 174-2) calls for tuberculoid leprosy to be treated with dapsone (100 mg/d) for 5 years and for lepromatous leprosy to be treated with rifampin (600 mg/d) for 3 years and with dapsone (100 mg/d) throughout life.

With effective antimicrobial therapy, new skin lesions as well as signs and symptoms of peripheral neuropathy cease appearing. Nodules and plaques of lepromatous leprosy noticeably flatten in 1–2 months and resolve in one or a few years, while tuberculoid skin lesions may disappear, ameliorate, or remain relatively unchanged. Although the peripheral neuropathy of leprosy may improve somewhat in the first few months of therapy, rarely is it significantly alleviated by treatment.

Although two of the three recommended drugs (dapsone and clofazimine) are only bacteriostatic against *M. leprae*, and although bactericidal agents have been identified since the WHO formulated its recommendations, significant studies employing the available alternatives in newly designed regimens have not been initiated. Given that moxifloxacin, like rifampin, is profoundly bactericidal in leprosy patients and that short-course chemotherapy for tuberculosis is possible only when two or more bactericidal agents are used, a moxifloxacin/rifamycin-based regimen including either minocycline or clarithromycin (each more potent than either dapsone or clofazimine) appears promising; such a regimen may prove to be more reliably curative than WHO-recommended multidrug therapy for lepromatous leprosy and may allow a considerably shorter course of treatment.

THE LEPROSY ELIMINATION CAMPAIGN: MODIFICATIONS OF THE MULTIDRUG REGIMEN AND THEIR CONSEQUENCES

The World Health Assembly resolution of 1991 proposed to “eliminate leprosy as a public health problem” in all countries by the year 2000, with success defined as <1 in 10,000 persons with leprosy not having completed a course of multidrug therapy. When leprosy elimination appeared not to be occurring in that time frame, several changes were made to WHO multidrug therapy recommendations to simplify leprosy diagnosis and treatment, ease operational requirements for control programs, and facilitate attainment of the elimination target:

1. The duration of treatment of multibacillary leprosy was reduced from 2 years to 1 year in 1995 and to only 6 months in 2002. Neither duration was supported by prior clinical trials, and neither has been demonstrated in trials with prolonged follow-up to provide a reliable cure. However, this reduction in duration of therapy for multibacillary leprosy had the effect of reducing for elimination purposes the number of multibacillary cases by a factor of four.

2. The recommendation for multibacillary treatment was first changed to mandate therapy for any patient whose skin smear was positive. By 1995, skin biopsies and even skin smears were no longer advocated for leprosy diagnosis and classification. For treatment purposes, classification henceforth was to be determined on clinical grounds alone: more than five anesthetic skin lesions or enlarged nerves were to be considered multibacillary and fewer lesions paucibacillary. Unfortunately, such classification often led to under-treatment for some patients and over-treatment for others. Although leprosy often can be diagnosed on clinical grounds alone, it is not infrequently confused with other dermatologic disorders,

and the diagnosis not uncommonly remains uncertain even for seasoned leprologists. Skin smears and biopsies often clarify a leprosy diagnosis and earmark those patients most prone to relapse after the completion of therapy. With these diagnostic tests abandoned, leprosy diagnosis has been profoundly compromised.

3. The WHO advocated the integration of leprosy into national general health services. In most countries where leprosy is endemic, those services are already overburdened, and the prospects that their leprosy patients will receive optimal care is most unlikely. Unfortunately, in endemic countries, medical education often ignores leprosy. Therefore, in general health services, a leprosy diagnosis is frequently delayed or missed, and a rise in leprosy disability and deformity has consequently been encountered worldwide.

4. The WHO now promotes “accompanied multidrug therapy,” which allows a newly diagnosed patient—if accompanied by a companion who will offer assistance and encourage drug compliance—to receive the full complement of the leprosy regimen at the time of diagnosis; thereafter, the patient is no longer counted as a case and often is no longer entitled to receive further leprosy services. Accompanied multidrug treatment entirely eliminated directly observed therapy—a cornerstone of effective treatment for tuberculosis—from the leprosy treatment regimen. Like that for all chronic diseases, including tuberculosis, compliance with leprosy treatment has proved consistently unreliable. Compliance was previously ameliorated to some extent by the directly observed monthly component of the WHO leprosy regimen.

5. The WHO now encourages “uniform multidrug therapy,” whereby all leprosy patients receive the same 6 months of treatment previously reserved for multibacillary cases. This treatment duration for multibacillary cases (or an even longer duration) has not proved reliable in preventing relapse. Moreover, patients whose cases were previously classified as paucibacillary receive clofazimine—a drug whose use is not necessary and is often cosmetically and psychosocially unacceptable, marking patients with a leprosy diagnosis.

6. Since 1995, the WHO went on record as not advocating for patient follow-up after completion of multidrug treatment. Thus, relapse is assuredly underreported, and its impact is not appreciated.

7. A by-product of the elimination campaign and the public perception that leprosy elimination is at hand is the substantial diminution of funding for both patient care and research. An older generation of leprologists is retiring, specialized leprosy facilities are disappearing, and recruitment of professionals for careers as leprosy clinicians and researchers has been substantially reduced. As a consequence, leprosy research has been largely abandoned. In particular, though there are real prospects for improving chemotherapy for leprosy and evaluating new antimicrobial agents both in mice and in clinical trials, almost no such efforts have commenced in the past decade. Furthermore, the mouse footpad laboratories required for such work worldwide are few and lack the capacity to monitor those efforts.

THERAPY FOR REACTIONS

Type 1

Type 1 lepra reactions are best treated with glucocorticoids (e.g., prednisone, initially at doses of 40–60 mg/d). As inflammation subsides, the glucocorticoid dose can be tapered, but steroid therapy must be continued for at least 3–6 months lest recurrence supervene. Because of the myriad toxicities of prolonged glucocorticoid therapy, the indications for its initiation are strictly limited to lesions whose intense inflammation poses a threat of ulceration; lesions at cosmetically important sites, such as the face; and cases in which neuritis is present. Mild to moderate lepra reactions that do not meet these criteria should be tolerated, with glucocorticoid treatment withheld. Thalidomide is ineffective against type 1 lepra reactions. Clofazimine (200–300 mg/d) is of questionable benefit but in any event is far less efficacious than glucocorticoids.

Type 2

Treatment of ENL must be individualized. If ENL is mild (i.e., if it occurs without fever or other organ involvement and with occasional crops of only a few skin papules), it may be treated with antipyretics alone. However, in cases with many skin lesions, fever, malaise, and other tissue involvement, brief courses (1–2 weeks) of glucocorticoid treatment (initially 40–60 mg/d) are often effective. With or without therapy, individual inflamed papules last for <1 week. Successful therapy is defined by the cessation of skin lesion development and the disappearance of other systemic signs and symptoms. If, despite two courses of glucocorticoid therapy, ENL appears to be recurring and persisting, treatment with thalidomide (100–300 mg nightly) should be initiated, with the dose depending on the initial severity of the reaction. Because even a single dose of thalidomide administered early in pregnancy may result in severe birth defects, including phocomelia, the use of this drug in the United States for the treatment of fertile female patients is tightly regulated and requires informed consent, prior pregnancy testing, and maintenance of birth control measures. Although the mechanism of thalidomide's dramatic action against ENL is not entirely clear, the drug's efficacy is probably attributable to its reduction of TNF levels and IgM synthesis and its slowing of polymorphonuclear leukocyte migration. After the reaction is controlled, lower doses of thalidomide (50–200 mg nightly) are effective in preventing relapses of ENL. Clofazimine in high doses (300 mg nightly) has some efficacy against ENL, but its use permits only a modest reduction of the glucocorticoid dose necessary for ENL control.

Lucio's Phenomenon

Neither glucocorticoids nor thalidomide is effective against this syndrome. Optimal wound care and therapy for bacteremia are indicated. Ulcers tend to be chronic and heal poorly. In severe cases, exchange transfusion may prove useful.

TABLE 174-2

Antimicrobial Regimens Recommended for the Treatment of Leprosy in Adults

Form of Leprosy	More Intensive Regimen	WHO-Recommended Regimen (1982)
Tuberculoid (paucibacillary)	Dapsone (100 mg/d) for 5 years	Dapsone (100 mg/d, unsupervised) *plus* rifampin (600 mg/month, supervised) for 6 months
Lepromatous (multibacillary)	Rifampin (600 mg/d) for 3 years *plus* dapsone (100 mg/d) indefinitely	Dapsone (100 mg/d) *plus* clofazimine (50 mg/d), unsupervised; *and* rifampin (600 mg) *plus* clofazimine (300 mg) monthly (supervised) for 1–2 years

Note: See text for discussion and comparison of the WHO recommendations with the more intensive approach as well as the alternative WHO regimen for single-lesion paucibacillary leprosy.

PREVENTION AND CONTROL

Multidrug Treatment and Leprosy Elimination

Between 2000 and 2006, the worldwide annual case-detection rate of leprosy fell from 700,000 to 280,000. Because India alone was estimated to carry 60% of the world's leprosy burden and was slow to reach the elimination goal, additional operational policies were instituted:

1. Single-lesion leprosy, which accounts for one-third of leprosy cases in India, was no longer classified as leprosy at all.
2. A case of leprosy was no longer counted if diagnosed by the treating physician but not verified (as required) by program managers at the district or state level—individuals who were under pressure to produce improved statistics.
3. Active case finding, which had been extensive and successful in India, was discouraged. Because of the stigma of leprosy, self-reporting often does not occur.

In India between 2000 and 2006, the annual leprosy case-detection rate fell from 560,000 to 140,000, and "elimination" was achieved in 2004. However, since the incubation period of leprosy in the large majority of cases is 5–7 years or longer, and since most new cases reported in 2006 were already incubating in 2000, the claim of a dramatic fall in the incidence of leprosy—and, as a consequence, of multidrug therapy—both in India and worldwide is epidemiologically unreasonable. Rather, millions of cases were undiagnosed and untreated, while new leprosy cases, disability, and deformity have been documented to be on the rise.

Vaccination

Vaccination at birth with bacille Calmette-Guérin (BCG) has proved variably effective in preventing leprosy: the results have ranged from total inefficacy to 80% efficacy. The addition of heat-killed *M. leprae* to BCG does not increase the effectiveness of the vaccine. Because whole mycobacteria contain large amounts of lipids and carbohydrates that have proved in vitro to be immunosuppressive for lymphocytes and macrophages, *M. leprae* proteins may prove to be superior vaccines. Data from a mouse model support this possibility.

Chemoprophylaxis

Chemoprophylaxis with dapsone may reduce the number of tuberculoid leprosy cases but not the number of lepromatous cases and therefore is not recommended, even for household contacts. In addition, single-dose rifampin prophylaxis is of doubtful efficacy.

Isolation

Because leprosy transmission appears to require close prolonged household contact, hospitalized patients need not be isolated.

LEPROSY: THE PRESENT SITUATION

During most of the twentieth century, nongovernmental organizations, particularly Christian missionaries, provided a medical infrastructure devoted to the care and treatment of leprosy patients—the envy of those with other medical priorities in the developing world. With the public perception that leprosy is near eradication, resources for patient care are rapidly being diverted, and the burden of patient care is being transferred to nonexistent or overloaded national health services and to health workers who lack the tools and skills needed for the disease's diagnosis and classification and for the selection of nuanced therapy (particularly in cases of reactional neuritis). Furthermore, after the completion of therapy, when a patient is no longer considered to represent a case, half of all patients continue to manifest disease activity for years; relapse rates (at least for multibacillary patients) are unacceptably high; disabilities and deformities go unchecked; and the social stigma of the disease persists. Thus the prerequisites for a salutary outcome increasingly go unmet.

FURTHER READING

Gelber RH et al: The relapse rate in MB leprosy patients treated with 2-years of WHO-MDT is not low. Int J Lepr Other Mycobact Dis 72:493, 2004.

Gelber RH, Grosset J: The chemotherapy of leprosy: An interpretive history. Lepr Rev 83:221, 2012.

Pardillo FE et al: Powerful bactericidal activity of moxifloxacin in human leprosy. Antimicrob Agents Chemother 52:3113, 2008.

Ridley DS, Jopling WH: Classification of leprosy according to immunity. A five-group system. Int J Lepr Other Mycobact Dis 34:255, 1964.

Smith WC et al: The missing millions: A threat to the elimination of leprosy. PLoS Negl Trop Dis 9:e0003658, 2015. [PubMed: 25905706]

Chapter 175: Nontuberculous Mycobacterial Infections

Steven M. Holland

INTRODUCTION

Several terms—nontuberculous mycobacteria (NTM), atypical mycobacteria, mycobacteria other than tuberculosis, and environmental mycobacteria—all refer to mycobacteria other than *Mycobacterium tuberculosis*, its close relatives (*M. bovis, M. caprae, M. africanum, M. pinnipedii, M. canetti*), and *M. leprae*. The number of identified species of NTM is growing and will continue to do so because of the use of DNA sequence typing for speciation. The number of known species currently exceeds 170. NTM are highly adaptable and can inhabit hostile environments, including industrial solvents.

EPIDEMIOLOGY

NTM are ubiquitous in soil and water. Specific organisms have recurring niches, such as *M. simiae* in certain aquifers, *M. fortuitum* in pedicure baths, and *M. immunogenum* in metalworking fluids. Most NTM cause disease in humans only rarely unless some aspect of host defense is impaired, as in bronchiectasis, or breached, as by inoculation (e.g., liposuction, trauma, cardiac surgery). There are few instances of human-to-human transmission of NTM, which occurs almost exclusively in cystic fibrosis. Because infections due to NTM are rarely reported to health agencies and because their identification is sometimes problematic, reliable data on incidence and prevalence are lacking. Disseminated disease denotes significant immune dysfunction (e.g., advanced HIV infection), whereas pulmonary disease, which is much more common, is highly associated with pulmonary epithelial defects but not with systemic immunodeficiency.

In the United States, the incidence and prevalence of pulmonary infection with NTM, mostly in association with bronchiectasis **(Chap. 284)**, have for many years been several-fold higher than the corresponding figures for tuberculosis, and rates of the former are increasing among the elderly as rates of tuberculosis continue to fall. Among patients with cystic fibrosis, who often have bronchiectasis, rates of clinical infection with NTM range from 3 to 15%, with even higher rates among older patients. Although NTM may be recovered from the sputa of many individuals, it is critical to differentiate active disease from commensal harboring of the organisms. A scheme to help with the proper diagnosis of pulmonary infection caused by NTM has been developed by the American Thoracic Society and is widely used. The bulk of nontuberculous mycobacterial disease in North America is due to *M. kansasii*, organisms of the *M. avium* complex (MAC), and *M. abscessus*.

In Europe, Asia, and Australia, the distribution of NTM in clinical specimens is roughly similar to that in North America, with MAC species and rapidly growing organisms such as *M. abscessus* encountered frequently. *M. xenopi* and *M. malmoense* are especially prominent in northern Europe. *M. ulcerans* causes the distinct clinical entity Buruli ulcer, which occurs throughout tropical zones, especially in western Africa. *M. marinum* is a common cause of cutaneous and tendon infections in coastal regions and among individuals exposed to fish tanks or swimming pools.

The true international epidemiology of infections due to NTM is hard to determine because the isolation of these organisms often is not reported and speciation often is not performed for *M. tuberculosis* or NTM. The latter issue poses an especially important problem during therapy for tuberculosis when smears positive for acid-fast bacilli are considered evidence of treatment failure. The increasing ease of identification and speciation of these organisms is already having a major impact on the description of the dynamic international epidemiology of tuberculosis and NTM infections.

PATHOBIOLOGY

Because exposure to NTM is essentially universal and disease is rare, it can be assumed that normal host defenses against these organisms must be strong and that otherwise healthy individuals in whom significant disease develops are highly likely to have specific susceptibility factors that permit NTM to become established, multiply, and cause disease. At the advent of HIV infection, CD4+ T lymphocytes were recognized as key effector cells against NTM; the development of disseminated MAC disease was highly correlated with a decline in CD4+ T lymphocyte numbers. Such a decrease has also been implicated in disseminated MAC infection in patients with idiopathic CD4+ T lymphocytopenia. Potent inhibitors of tumor necrosis factor α (TNF-α), such as infliximab, adalimumab, certolizumab, golimumab, and etanercept, neutralize this critical cytokine, with consequent inhibition of granuloma formation. The occasional result is severe mycobacterial or fungal infection; these associations indicate that TNF-α is a crucial element in mycobacterial control. However, in cases without the above risk factors, much of the genetic basis of susceptibility to disseminated infection with NTM is accounted for by specific mutations in the interferon γ (IFN-γ)/interleukin 12 (IL-12) synthesis and response pathways.

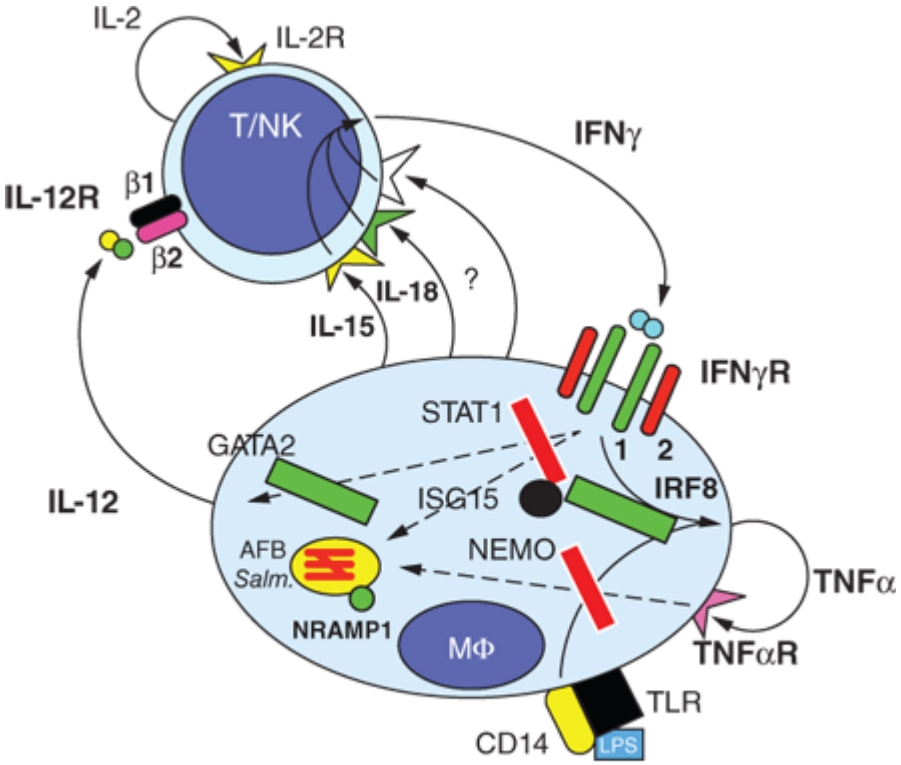


Mycobacteria are typically phagocytosed by macrophages, which respond with the production of IL-12, a heterodimer composed of IL-12p35 and IL-12p40 moieties that together make up IL-12p70. IL-12 activates T lymphocytes and natural killer cells through binding to its receptor (composed of IL-12Rβ1 and IL-12Rβ2/IL-23R), with consequent phosphorylation of STAT4. IL-12 stimulation of STAT4 leads to secretion of IFN-γ, which activates neutrophils and macrophages to produce reactive oxidants, to increase expression of the major histocompatibility complex and Fc receptors, and to concentrate certain antibiotics intracellularly. Signaling by IFN-γ through its receptor (composed of IFN-γR1 and IFN-γR2) leads to phosphorylation of STAT1, which in turn regulates IFN-γ-responsive genes, such as those coding for IL-12 and TNF-α. TNF-α signals through its own receptor via a downstream complex containing the nuclear factor κB (NF-κB) essential modulator (NEMO). Therefore, the positive feedback loop between IFN-γ and IL-12/IL-23 drives the

immune response to mycobacteria and other intracellular infections. These genes are known to be the critical ones in the pathway of mycobacterial control: specific Mendelian mutations have been identified in *IFNGR1, IFNGR2, STAT1, GATA2, ISG15, IRF8, IL-12A, IL-12RB1, IL-12RB2, CYBB* (which encodes the gp91^{phox} protein of the NADPH oxidase), and *IKBKG* (which encodes NEMO) **(Fig. 175-1)**. Despite the identification of genes associated with disseminated disease, only ~70% of cases of disseminated nontuberculous mycobacterial infections that are not associated with HIV infection have a genetic diagnosis; the implication is that more mycobacterial susceptibility genes and pathways remain to be identified.

FIGURE 175-1

Cytokine interactions of infected macrophages (MF) with T and natural killer (NK) lymphocytes. Infection of macrophages by mycobacteria (AFB) leads to the release of heterodimeric interleukin 12 (IL-12). IL-12 acts on its receptor complex (IL-12R), with consequent STAT4 activation and production of homodimeric interferon γ (IFNγ). Through its receptor (IFNγR), IFNγ activates STAT1, stimulating the production of tumor necrosis factor α (TNFα) and leading to the killing of intracellular organisms such as mycobacteria, salmonellae (*Salm.*), and some fungi. Homotrimeric TNFα acts through its receptor (TNFαR) and requires nuclear factor κB essential modulator (NEMO) to activate nuclear factor κB, which also contributes to the killing of intracellular bacteria. Both IFNγ and TNFα lead to upregulation of IL-12. TNFα-blocking antibodies work either by blocking the ligand (infliximab, adalimumab, certolizumab, golimumab) or by providing soluble receptor (etanercept). Mutations in *IFNGR1, IFNGR2, IL12B, IL12RB1, IL12RB2, STAT1, GATA2, ISG15, IRF8, CYBB,* and *IKBKG* (NEMO) have been associated with predisposition to mycobacterial infections. Other cytokines, such as IL-15 and IL-18, also contribute to IFNγ production. Signaling through the Toll-like receptor (TLR) complex and CD14 also upregulates TNFα production. IRF8, interferon regulatory factor 8; ISG15, interferon-stimulated gene 15; LPS, lipopolysaccharide; NRAMP1, natural resistance-associated macrophage protein 1.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In contrast to the recognized genes and mechanisms associated with disseminated nontuberculous mycobacterial infection, the best-recognized underlying condition for pulmonary infection with NTM is bronchiectasis (Chap. 284). Most of the well-characterized forms of bronchiectasis, including cystic fibrosis, primary ciliary dyskinesia, STAT3-deficient hyper-IgE syndrome, and idiopathic bronchiectasis, have high rates of association with nontuberculous mycobacterial infection. The precise mechanism by which bronchiectasis predisposes to locally destructive but not systemic involvement is unknown.

Unlike disseminated or pulmonary infection, "hot-tub lung" represents pulmonary hypersensitivity to NTM—most commonly MAC organisms—growing in underchlorinated water, often in indoor hot tubs.

CLINICAL MANIFESTATIONS

Disseminated Disease

Disseminated MAC or *M. kansasii* infections in patients with advanced HIV infection are now uncommon in North America because of effective antimycobacterial prophylaxis and improved treatment of HIV infection. When such mycobacterial disease was common, the portal of entry was the bowel, with spread to bone marrow and the bloodstream. Surprisingly, disseminated infections with rapidly growing NTM (e.g., *M. abscessus, M. fortuitum*) are very rare in HIV-infected patients, even in those with advanced HIV infection. Because these organisms are of low intrinsic virulence and disseminate only in conjunction with impaired immunity, disseminated disease can be indolent and progressive over weeks to months. Typical manifestations of malaise, fever, and weight loss are often accompanied by organomegaly, lymphadenopathy, and anemia. Because special cultures or stains are required to identify the organisms, the

most critical step in diagnosis is to suspect infection with NTM. Blood cultures may be negative, but involved organs typically have significant organism burdens, sometimes with a grossly impaired granulomatous response.

 In a child, disseminated involvement (i.e., involvement of two or more organs) without an underlying iatrogenic cause should prompt an investigation of the IFN-γ/IL-12 pathway. Recessive mutations in *IFNGR1* and *IFNGR2* typically lead to severe infection with NTM. In contrast, dominant negative mutations in *IFNGR1*, which lead to over-accumulation of a defective interfering mutant receptor on the cell surface, inhibit normal IFN-γ signaling and thus lead to nontuberculous mycobacterial osteomyelitis. Dominant negative mutations in *STAT1* and recessive mutations in *IL-12RB1* can produce variable phenotypes consistent with their residual capacities for IFN-γ synthesis and response. Male patients who have disseminated nontuberculous mycobacterial infections along with conical, peg, or missing teeth and an abnormal hair pattern should be evaluated for defects in the pathway that activates NF-κB through NEMO (*IKBKG*). These patients may have associated immune globulin defects as well. Patients with myelodysplasia and mycobacterial disease should be investigated for GATA2 deficiency. A recently recognized group of patients who often develop disseminated infections with rapidly growing NTM (predominantly *M. abscessus*) as well as other opportunistic infections have high-titer neutralizing autoantibodies to IFN-γ. Thus far, this syndrome has been reported most frequently in East Asian female patients.

IV catheters can become infected with NTM, usually as a consequence of contaminated water. *M. abscessus* and *M. fortuitum* sometimes infect deep indwelling lines as well as fluids used in eye surgery, subcutaneous injections, and local anesthetics. Infected catheters should be removed.

Pulmonary Disease

Lung disease is by far the most common form of nontuberculous mycobacterial infection in North America and the rest of the industrialized world. In North America, rates of NTM lung disease far exceed rates of tuberculosis. The clinical presentation typically consists of months or years of throat clearing, nagging cough, and slowly progressive fatigue. Patients will often have seen physicians multiple times and received symptom-based or transient therapy before the diagnosis is entertained and samples are sent for mycobacterial stains and cultures. Because not all patients can produce sputum, bronchoscopy may be required for diagnosis. The typical lag between onset of symptoms and diagnosis is ~5 years in older women. Predisposing factors include underlying lung diseases such as bronchiectasis **(Chap. 284)**, pneumoconiosis **(Chap. 283)**, chronic obstructive pulmonary disease **(Chap. 286)**, primary ciliary dyskinesia **(Chap. 284)**, α_1 antitrypsin deficiency **(Chap. 286)**, and cystic fibrosis **(Chap. 285)**. Bronchiectasis and nontuberculous mycobacterial infection often coexist and progress in tandem. This situation makes causality difficult to determine in a given index case, but bronchiectasis is certainly among the most critical predisposing factors that are exacerbated by infection.

MAC organisms are the most common causes of pulmonary nontuberculous mycobacterial infection in North America, but rates vary somewhat by region. MAC infection most commonly develops during the sixth or seventh decade of life in women who have had months or years of nagging intermittent cough and fatigue, with or without sputum production or chest pain. The constellation of pulmonary disease due to NTM in a tall and thin woman who may have chest wall abnormalities is often referred to as Lady Windermere syndrome, after an Oscar Wilde character of the same name. In fact, pulmonary MAC infection does afflict older nonsmoking white women more than men, with onset at ~60 years. Patients tend to be taller and thinner than the general population, with high rates of scoliosis, mitral valve prolapse, and pectus anomalies. Whereas male smokers with upper-lobe cavitary disease tend to carry the same single strain of MAC indefinitely, nonsmoking females with nodular bronchiectasis tend to carry several strains of MAC simultaneously, with changes over the course of their disease.

M. kansasii can cause a clinical syndrome that strongly resembles tuberculosis, consisting of hemoptysis, chest pain, and cavitary lung disease. The rapidly growing NTM, such as *M. abscessus*, have been associated with esophageal motility disorders such as achalasia. Patients with pulmonary alveolar proteinosis are prone to pulmonary nontuberculous mycobacterial and *Nocardia* infections; the underlying mechanism may be inhibition of alveolar macrophage function due to the autoantibodies to granulocyte-macrophage colony-stimulating factor found in many of these patients.

Cervical Lymph Nodes

The most common form of nontuberculous mycobacterial infection among young children in North America is isolated cervical lymphadenopathy, caused most frequently by MAC organisms but also by other NTM. The cervical swelling is typically firm and relatively painless, with a paucity of systemic signs. Because the differential diagnosis of painless adenopathy includes malignancy, many children have infection with NTM diagnosed inadvertently at biopsy; cultures and special stains may not have been requested because mycobacterial disease was not ranked high in the differential. Local fistulae usually resolve completely with resection and/or antibiotic therapy. Likewise, the entity of isolated pediatric intrathoracic nontuberculous mycobacterial infection, which is probably related to cervical lymph node infection, is usually mistaken for cancer. In neither isolated cervical nor isolated intrathoracic infections with NTM have children with underlying immune defects been commonly identified, nor do the affected children usually go on to develop other opportunistic infections.

Skin and Soft Tissue Disease

Cutaneous involvement with NTM usually requires a break in the skin for introduction of the bacteria. Pedicure bath–associated infection with *M. fortuitum* is more likely if skin abrasion (e.g., during leg shaving) has occurred just before the pedicure. Outbreaks of skin infection are often caused by rapidly growing NTM (especially *M. abscessus, M. fortuitum*, and *M. chelonae*) acquired via skin contamination from surgical instruments (especially in cosmetic surgery), injections, and other procedures. These infections are typically

accompanied by painful, erythematous, draining subcutaneous nodules, usually without associated fever or systemic symptoms.

M. marinum lives in many water sources and can be acquired from fish tanks, swimming pools, barnacles, and fish scales. This organism typically causes papules or ulcers ("fish-tank granuloma"), but the infection can progress to tendinitis with significant impairment of manual dexterity. Lesions appear days to weeks after inoculation of organisms by a typically minor trauma (e.g., incurred during the cleaning of boats or the handling of fish). Tender nodules due to *M. marinum* can advance up the arm in a pattern also seen with *Sporothrix schenckii* (*sporotricoid spread*). The typical carpal-tendon involvement may be the first presenting manifestation and may lead to surgical exploration or steroid injection. The index of suspicion for *M. marinum* infections must be high to ensure that proper specimens obtained during procedures are sent for culture.

 *M. ulcerans*, another waterborne skin pathogen, is found mainly in the tropics, especially in tropical areas of Africa. Infection follows skin trauma or insect bites that allow admission to contaminated water. The skin lesions are typically painless, clean ulcers that slough and can cause osteomyelitis. The toxin mycolactone accounts for the modest host inflammatory response and the painless ulcerations.

DIAGNOSIS

NTM can be detected on acid-fast or fluorochrome smears of sputum or other body fluids. When the organism burden is high, the organisms may appear as gram-positive beaded rods, but this finding is unreliable. (In contrast, nocardiae may appear as gram-positive and beaded but filamentous bacteria.) Again, the requisite and most sensitive step in the diagnosis of any mycobacterial disease is to think of including it in the differential. In almost all laboratories, mycobacterial sample processing, staining, and culture are conducted separately from routine bacteriologic tests; thus many infections go undiagnosed because of the physician's failure to request the appropriate test. In addition, mycobacteria usually require separate blood culture media. NTM are broadly differentiated into rapidly growing (<7 days) and slowly growing (≥7 days) forms. Because *M. tuberculosis* typically takes ≥2 weeks to grow, many laboratories refuse to consider culture results final until 6 weeks have elapsed. Newer techniques using liquid culture media permit more rapid isolation of mycobacteria from specimens than is possible with traditional media. Species more readily detected with incubation at 30°C include *M. marinum, M. haemophilum*, and *M. ulcerans. M. haemophilum* prefers iron supplementation or blood, whereas *M. genavense* requires supplemented medium with the additive mycobactin J. Bacterial formation of pigment in light conditions (*photochromogenicity*) or dark conditions (*scotochromogenicity*) or a lack of bacterial pigment formation (*nonchromogenicity*) was historically used to help categorize NTM. In contrast to NTM colonies, *M. tuberculosis* colonies are beige, rough, dry, and flat. Current identification schemes reliably use biochemical, nucleic acid, or cell wall composition, as assessed by high-performance liquid chromatography or mass spectrometry, for speciation. With the remarkable decline in U.S. cases of tuberculosis over recent decades, NTM have become the mycobacteria most commonly isolated from humans in North America. However, not all isolations of NTM, especially from the lung, reflect pathology and require treatment. Whereas identification of an organism in a

blood or organ biopsy specimen in a compatible clinical setting is diagnostic, the American Thoracic Society recommends that pulmonary infection due to NTM be diagnosed only when disease is clearly demonstrable —i.e., in an appropriate clinical and radiographic setting (nodules, bronchiectasis, cavities) and with repeated isolation of NTM from expectorated sputum or recovery of NTM from bronchoscopy or biopsy specimens. Given the large number of species of NTM and the importance of accurate diagnosis for the implementation of proper therapy, identification of these organisms is ideally taken to the species level.

The purified protein derivative (PPD) of tuberculin is delivered intradermally to evoke a memory T cell response to mycobacterial antigens. This test is variously referred to as the PPD test, the tuberculin skin test, and the Mantoux test, among other designations. Unfortunately, the cutaneous immune response to these tuberculosis-derived filtrate proteins does not differentiate well between infection with some NTM and that with *M. tuberculosis*. Because intermediate reactions (~10 mm) to PPD in latent tuberculosis and nontuberculous mycobacterial infections can overlap significantly, the progressive decline in active tuberculosis in the United States means that NTM probably account for increasing proportions of PPD reactivity. In addition, *bacille Calmette-Guérin* (BCG) can cause some degree of cross-reactivity, posing problems of interpretation for patients who have received BCG vaccine. Assays to measure the elaboration of IFN-γ in response to the relatively tuberculosis-specific proteins ESAT6 and CFP10 form the basis for IFN-γ-release assays (IGRAs). These assays can be performed with whole blood or on membranes. It is important to note that *M. marinum, M. kansasii*, and *M. szulgai* also have ESAT6 and CFP10 and may cause false-positive reactions in IGRAs. Despite cross-reactivity with NTM, large PPD reactions (>15 mm) most commonly signify tuberculosis.

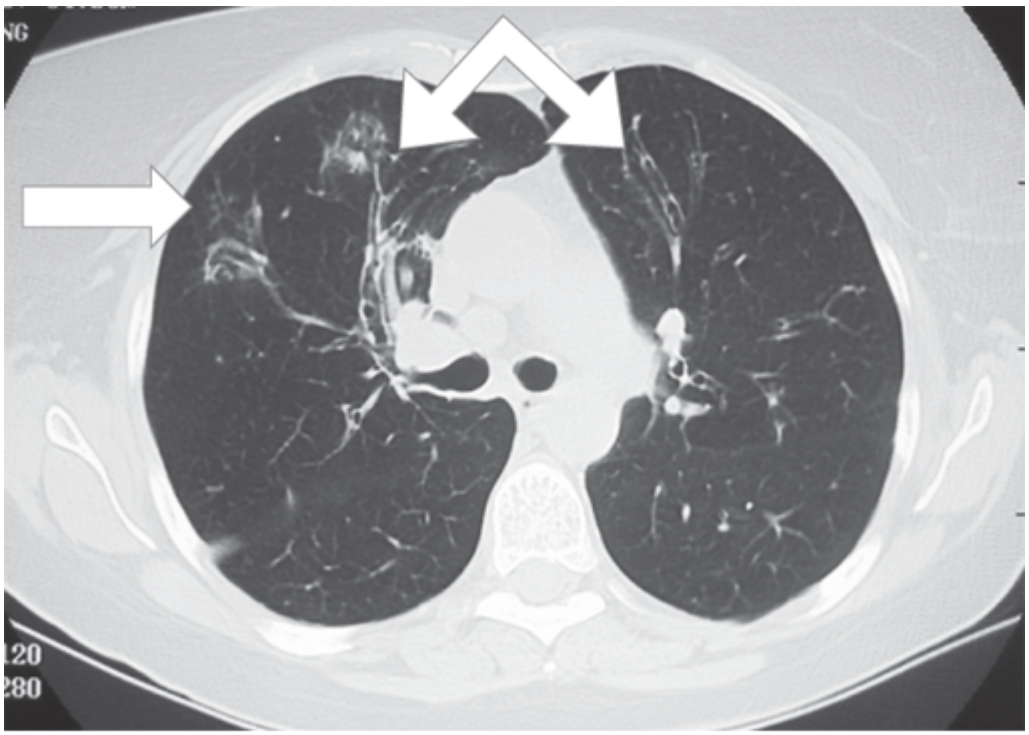
Isolation of NTM from blood specimens is clear evidence of disease. Whereas rapidly growing mycobacteria may proliferate in routine blood culture media, slow-growing NTM typically do not; thus it is imperative to suspect the diagnosis and to use the correct bottles for cultures. Isolation of NTM from a biopsy specimen constitutes strong evidence for infection, but cases of laboratory contamination do occur. Identification of organisms on stained sections of biopsy material confirms the authenticity of the culture. Certain NTM require lower incubation temperatures (*M. genavense*) or special additives (*M. haemophilum*) for growth. Some NTM (e.g., *M. tilburgii*) remain noncultivable but can be identified molecularly in clinical samples.

The radiographic appearance of nontuberculous mycobacterial disease in the lung depends on the underlying disease, the severity of the infection, and the imaging modality used. The advent and increase in the use of CT have allowed the identification of characteristic changes that are highly consistent with nontuberculous mycobacterial infection, such as the “tree-in-bud” pattern of bronchiolar inflammation **(Fig. 175-2)**. Involvement of the lingual and right-middle lobes is commonly seen on chest CT but is difficult to appreciate on plain film. Severe bronchiectasis and cavity formation are common in more advanced disease. Isolation of NTM from respiratory samples can be confusing. *M. gordonae* is often recovered from respiratory samples but is not usually seen on smear and is almost never a pathogen. Patients with bronchiectasis occasionally have NTM recovered from sputum culture with a negative smear. The American Thoracic Society has developed guidelines for the diagnosis of infection with MAC, *M. abscessus*, and *M. kansasii*. A positive diagnosis requires the growth of NTM from two of three sputum samples, regardless of smear findings; a

positive bronchoscopic alveolar sample, regardless of smear findings; or a pulmonary parenchyma biopsy sample with granulomatous inflammation or mycobacteria found on section and NTM found on culture. These guidelines probably apply to other NTM as well.

FIGURE 175-2

Chest CT of a patient with pulmonary *Mycobacterium avium* complex infection. *Arrows* indicate the "tree-in-bud" pattern of bronchiolar inflammation (peripheral right lung) and bronchiectasis (central right and left lungs).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although many laboratories use DNA probes to identify *M. tuberculosis*, MAC, *M. gordonae*, and *M. kansasii*, speciation of NTM helps determine the antimycobacterial therapy to be used. Only testing of MAC organisms for susceptibility to clarithromycin and of *M. kansasii* for susceptibility to rifampin is indicated; few data support other in vitro susceptibility tests, attractive though they appear. MAC isolates that have not been exposed to macrolides are almost always susceptible. NTM that have persisted beyond a course of antimicrobial therapy are often tested for antibiotic susceptibility, but the value and meaning of these tests are undetermined.

PREVENTION

Prophylaxis of MAC disease in patients infected with HIV is started when the CD4+ T lymphocyte count falls to <50/μL. Azithromycin (1200 mg weekly), clarithromycin (1000 mg daily), or rifabutin (300 mg daily) is effective. Macrolide prophylaxis in immunodeficient patients who are susceptible to NTM (e.g., those with defects in the IFN-γ/IL-12 axis) has not been prospectively validated but seems prudent.

TREATMENT

TREATMENT

Nontuberculous Mycobacteria

NTM cause chronic infections that evolve relatively slowly over a period of weeks to years. Therefore, it is rarely necessary to initiate treatment on an emergent basis before the diagnosis is clear and the infecting species is known. Treatment of NTM is complex, often poorly tolerated, and potentially toxic. Just as in tuberculosis, inadequate single-drug therapy is almost always associated with the emergence of antimicrobial resistance and relapse.

MAC infection often requires multidrug therapy, the foundation of which is a macrolide (clarithromycin or azithromycin), ethambutol, and a rifamycin (rifampin or rifabutin). For disseminated nontuberculous mycobacterial disease in HIV-infected patients, the use of rifamycins poses special problems—i.e., rifamycin interactions with protease inhibitors. For pulmonary MAC disease, thrice-weekly administration of a macrolide, a rifamycin, and ethambutol has been successful. Therapy is prolonged, generally continuing for 12 months after culture conversion; typically, a course lasts for at least 18 months. Other drugs with activity against MAC organisms include IV and aerosolized aminoglycosides, fluoroquinolones, and clofazimine. In elderly patients, rifabutin can exert significant toxicity. However, with only modest efforts, most antimycobacterial regimens are well tolerated by most patients. Resection of cavitary lesions or severely bronchiectatic segments has been advocated for some patients, especially those with macrolide-resistant infections. The success of therapy for pulmonary MAC infections depends on whether disease is nodular or cavitary and on whether it is early or advanced, ranging from 20 to 80%.

M. kansasii lung disease is similar to tuberculosis in many ways and is also effectively treated with isoniazid (300 mg/d), rifampin (600 mg/d), and ethambutol (15 mg/kg per day). Other drugs with very high-level activity against *M. kansasii* include clarithromycin, fluoroquinolones, and aminoglycosides. Treatment should continue until cultures have been negative for at least 1 year. In most instances, *M. kansasii* infection is easily cured. Bulky, severe, necrotizing *M. kansasii* lymphadenopathy, especially in the mediastinum, is strongly associated with GATA2 deficiency.

Rapidly growing mycobacteria pose special therapeutic problems. Extrapulmonary disease in an immunocompetent host is usually due to inoculation (e.g., via surgery, injections, or trauma) or to line infection and is often treated successfully with a macrolide and another drug (with the choice based on in vitro susceptibility), along with removal of the offending focus. In contrast, pulmonary disease, especially that caused by *M. abscessus*, is extremely difficult to cure. Repeated courses of treatment are usually effective in reducing the infectious burden and symptoms. Therapy generally includes a macrolide along with an IV-administered agent such as amikacin, a carbapenem, cefoxitin, or tigecycline. Other oral agents (used according to in vitro susceptibility testing and tolerance) include fluoroquinolones, doxycycline, and linezolid. Because nontuberculous mycobacterial infections are chronic, care must be taken in the long-term use of drugs with neurotoxicities, such as linezolid and ethambutol. Prophylactic pyridoxine has been

suggested in these cases. Durations of therapy for *M. abscessus* lung disease are difficult to predict because so many cases are chronic and require intermittent therapy. Expert consultation and management are strongly recommended.

Once recognized, *M. marinum* infection is highly responsive to antimicrobial therapy and is cured relatively easily with any combination of a macrolide, ethambutol, and a rifamycin. Therapy should be continued for 1–2 months after clinical resolution of isolated soft-tissue disease; tendon and bone involvement may require longer courses in light of clinical evolution. Other drugs with activity against *M. marinum* include sulfonamides, trimethoprim-sulfamethoxazole, doxycycline, and minocycline.

Treatment of the other NTM is less well defined, but macrolides and aminoglycosides are usually effective, with other agents added as indicated. Expert consultation is strongly encouraged for difficult or unusual infections due to NTM.

PROGNOSIS

The outcomes of nontuberculous mycobacterial infections are closely tied to the underlying condition (e.g., IFN-γ/IL-12 pathway defect, cystic fibrosis) and can range from recovery to death. With no or inadequate treatment, symptoms and signs can be debilitating, including persistent cough, fever, anorexia, and severe lung destruction. With treatment, patients typically regain strength and energy. The optimal duration of therapy when NTM persist in sputum is unknown, but treatment in this situation can be prolonged. In general, for severe underlying immunodeficiencies, hematopoietic stem cell transplantation is recommended and may be helpful in the resolution of severe mycobacterial disease.

GLOBAL CONSIDERATIONS

 In many countries, pulmonary tuberculosis is diagnosed by smear alone, which is also the method used for monitoring of response and relapse. However, examination of mycobacteria from the affected patients shows that a significant proportion of isolates are actually NTM. Overall, as rates of tuberculosis decline, the proportion of positive smears caused by NTM will increase. Advances in speciation will distinguish tuberculosis from nontuberculous mycobacterial infections and thereby affect rates of assumed relapse and resistance, leading to more targeted and appropriate therapy.

FURTHER READING

Aksamit TR et al: Adult patients with bronchiectasis: A first look at the US Bronchiectasis Research Registry. Chest 151:982, 2017.

Browne SK et al: Adult-onset immunodeficiency in Thailand and Taiwan. N Engl J Med 367:725, 2012.

Fleshner M et al: Mortality among patients with pulmonary non-tuberculous mycobacteria disease. Int J Tuberc Lung Dis 20:582, 2016.

Holland SM et al: Case 28-2017. A 13-month-old girl with pneumonia and a 33-year-old woman with hip pain. N Engl J Med 377:1077, 2017.

Kim RD et al: Pulmonary nontuberculous mycobacterial disease: Prospective study of a distinct preexisting syndrome. Am J Respir Crit Care Med 178:1066, 2008.

Lovell JP et al: Mediastinal and disseminated *Mycobacterium kansasii* disease in GATA2 deficiency. Ann Am Thorac Soc 13:2169, 2016.

Olivier KN et al: Inhaled amikacin for treatment of refractory pulmonary nontuberculous mycobacterial disease. Ann Am Thorac Soc 11:30, 2014.

Spinner MA et al: GATA2 deficiency: A protean disorder of hematopoiesis, lymphatics, and immunity. Blood 123:809, 2014.

Szymanski EP et al: Pulmonary nontuberculous mycobacterial infection. A multisystem, multigenic disease. Am J Respir Crit Care Med 192:618, 2015.

Wu UI, Holland SM: Host susceptibility to non-tuberculous mycobacterial infections. Lancet Infect Dis 15:968, 2015.

Chapter 176: Antimycobacterial Agents

Divya Reddy; Max R. O'Donnell

INTRODUCTION

Agents used for the treatment of mycobacterial infections, including tuberculosis (TB), leprosy, and infections due to nontuberculous mycobacteria (NTM), are administered in multiple-drug regimens for prolonged courses. Currently, >160 species of mycobacteria have been identified, the majority of which do not cause disease in humans. While the incidence of disease caused by *Mycobacterium tuberculosis* has been declining in the United States, TB remains a leading cause of morbidity and mortality in developing countries—particularly in sub-Saharan Africa and Asia, where the HIV epidemic rages. Not only effective drug regimens are needed; without a well-organized infrastructure for diagnosis and treatment of TB, therapeutic and control efforts are severely hampered **(Chaps. 460 and 462)**. Infections with NTM have gained in clinical prominence in the United States and other developed countries. These largely environmental organisms often establish infection in immunocompromised patients or in persons with structural lung disease.

TUBERCULOSIS

GENERAL PRINCIPLES

The earliest recorded human case of TB dates back 9000 years. Early treatment modalities, such as bloodletting, were replaced by sanatorium regimens in the late nineteenth century. The discovery of streptomycin in 1943 launched the era of antibiotic treatment for TB. Over subsequent decades, the discovery of additional agents and the use of multiple-drug regimens allowed progressive shortening of the treatment course from years to as little as 6 months for drug-susceptible TB. Latent TB infection (LTBI) and active TB disease are diagnosed by history, physical examination, radiographic imaging, tuberculin skin test, interferon γ release assays, acid-fast staining, mycobacterial cultures, and/or new molecular diagnostics. LTBI is treated with isoniazid (optimally daily or weekly for 9 months), rifampin (daily for 4 months), isoniazid plus rifampin (daily for 3 months), or isoniazid plus rifapentine (weekly for 3 months) **(Table 176-1)**.

TABLE 176-1

Regimens for the Treatment of Latent Tuberculosis Infection in Adults

Regimen	Schedule	Duration	Comments
Isoniazid	300 mg/d (5 mg/kg)	9 months (6 months acceptable)	Supplement with pyridoxine (25–50 mg daily).
	Alternative: 900 mg twice weekly (15 mg/kg)		Twice-weekly regimens require directly observed therapy.
Rifampin	600 mg/d (10 mg/kg)	4 months	Broader efficacy studies are needed.
Isoniazid plus rifampin	300 mg/d (5 mg/kg) plus 600 mg/d (10 mg/kg)	3 months	Broader efficacy studies are needed.
Isoniazid plus rifapentine	900 mg (15 mg/kg) weekly plus 900 mg (for weight >50 kg) weekly	3 months	Directly observed therapy is recommended for once-weekly treatment. This regimen may be supplemented with pyridoxine (25–50 mg/d).

Sources: D Menzies et al: Adverse events with 4 months of rifampin therapy or 9 months of isoniazid therapy for latent tuberculosis infection: A randomized trial. Ann Intern Med 149:689, 2008; American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America: Treatment of tuberculosis. Am J Respir Crit Care Med 167:603, 2003; T Sterling et al: Three months of rifapentine and isoniazid for latent tuberculosis infection. N Engl J Med 365: 2155, 2011.

For active or suspected TB disease, clinical factors, including HIV co-infection, symptom duration, radiographic appearance, and public health concerns about TB transmission, drive diagnostic testing and treatment initiation. Multiple-drug regimens are used for the treatment of TB disease **(Table 176-2)**. Initially, an intensive phase consisting of four drugs—isoniazid, rifampin, pyrazinamide, and ethambutol—given for 2 months is followed by a continuation phase of isoniazid and rifampin for 4 months, for a total treatment duration of 6 months. The continuation phase is extended to 7 months (for a total treatment duration of 9 months) for patients with cavitary disease; if the 2-month course of pyrazinamide is not completed; or if sputum cultures remain positive beyond 2 months of treatment (delayed culture conversion).

TABLE 176-2

Simplified Approach to Treatment of Active Tuberculosis (TB) in Adults

Culture results	Intensive Phase	Continuation Phase	Extension of Total Treatment
Culture-positive, drug-susceptible	HRZE for 2 months, daily[a] or 3 times per week (with dose adjustment)	HR for 4 months, daily or 5 days per week *or* HR for 4 months, 3 times per week[b] (with dose adjustment)	Continuation phase extended to 7 months if 2 months of Z is not completed, if the patient is infected with HIV and is not receiving antiretroviral therapy, or if culture conversion is prolonged and/or cavitation is evident on chest radiography[c]
Culture-negative	HRZE for 2 months	HR for 2 months, daily or 2 or 3 times per week[d]	Continuation phase extended to 4 months if the patient is infected with HIV
Extrapulmonary, drug-susceptible	HRZE for 2 months	HR for 4–7 months, daily or 5 days per week[e]	Continuation phase extended to 10 months in TB meningitis; 7 months recommended by some authorities for bone/joint TB
Resistant to H	QRZE[f] (or, less often, RZES) for 6 months	…	Prolonged culture conversion and/or evidence of cavitation on chest radiography
Resistant to R	HZEQ[f] (IA[g]) for 2 months	HEQ(S) for 10–16 months	Prolonged culture conversion, delayed response

Culture results	Intensive Phase	Continuation Phase	Extension of Total Treatment
Resistant to HR[h]	ZEQ[f] (IA[g]) ± alternative agents[i] for 18–24 months	…	Prolonged culture conversion

[a]Daily treatment is preferred; however, thrice-weekly therapy in the intensive phase (with or without an initial 2 weeks of daily therapy) may be considered in patients who are not infected with HIV and are at low risk of relapse (i.e., in pulmonary tuberculosis caused by drug-susceptible organisms that, at the start of treatment, is noncavitary and/or smear negative). [b]Use regimen with caution in HIV patients and/or those with cavitary disease, as missed doses can lead to treatment failure, relapse, and acquired drug resistance. [c]Culture conversion is prolonged if it occurs beyond 2 months. [d]Twice-weekly treatment regimens are not recommended in patients infected with HIV and those with cavitary pulmonary disease suspected to be TB. [e]Standard daily 6-month TB treatment regimen is considered to be adequate for most forms of extrapulmonary TB, including miliary TB. For TB meningitis, the addition of glucocorticoids is recommended. [f]Levofloxacin and moxifloxacin are the preferred fluoroquinolones. Gatifloxacin is associated with dysglycemia but may be an acceptable alternative. Ofloxacin and ciprofloxacin should generally be avoided because of resistance. [g]Injectable agents: streptomycin, amikacin, kanamycin, and capreomycin. [h]Multidrug-resistant TB should be managed by or in close consultation with an expert TB clinician. Surgical management should also be considered. [i]Alternative agents: cycloserine, ethionamide, para-aminosalicylic acid, clarithromycin, linezolid, and amoxicillin-clavulanate.

Abbreviations: E, ethambutol; H, isoniazid; IA, injectable agent; Q, fluoroquinolone; R, rifampin; S, streptomycin; and Z, pyrazinamide.

Sources: Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America: Clinical practice guidelines: Treatment of drug-susceptible tuberculosis. Clin Infect Dis 2016; 63:e147, 2016; C Mitnick et al: Comprehensive treatment of extensively drug-resistant tuberculosis. N Engl J Med 359:563, 2008; World Health Organization: 2011 Update: Guidelines for the programmatic management of drug-resistant tuberculosis *(www.who.int/tb/challenges/mdr/programmatic_guidelines_for_mdrtb/en/index.html)*.

Treatment of TB in patients co-infected with HIV poses significant challenges, but some progress is being made. To improve survival, current recommendations include initiation of antiretroviral therapy (ART) in HIV patients co-infected with *M. tuberculosis* within 2 weeks of the initiation of treatment for TB (except TB meningitis) if the CD4+ T cell count is ≤50/µL and by 8–12 weeks of TB treatment initiation if the CD4+ T cell count is ≥50/µL. Interactions of rifampin with protease inhibitors or non-nucleoside reverse transcriptase inhibitors are significant and require close monitoring and dose adjustments. Rifabutin is the alternative drug of choice in HIV patients co-infected with *M. tuberculosis*. The TB immune reconstitution inflammatory

syndrome (IRIS) may appear as early as 1 week after initiation of ART and manifests as paradoxical worsening or unmasking of existing TB infection. Conservative management consists of continued administration of ART and TB medications. However, severe or debilitating IRIS has been treated in reported case series with varying doses of glucocorticoids. Intermittent antimycobacterial therapy in patients infected with HIV and *M. tuberculosis* has been associated with low plasma levels of several key TB drugs and with higher rates of treatment failure or relapse; therefore, intermittent twice-weekly therapy for TB in HIV-co-infected individuals is not recommended.

Adherence to medications is critical in achieving a cure with antimycobacterial therapy. In addition to directly observed therapy (DOT) by trained staff, either in the clinic or at home, case management interventions such as patient education/counseling, field/home visits, and patient reminders are also recommended to improve treatment adherence. Use of mobile health technologies, including video DOT, text messaging, and next-generation electronic pillboxes, shows promise in promoting TB adherence. In drug-susceptible TB, monthly dispensing of TB medications is also advised for all patients to allow essential clinical monitoring for hepatotoxicity due to these medications. Clinical monitoring includes at least monthly assessment for symptoms (nausea, vomiting, abdominal discomfort, and unexplained fatigue) and signs (jaundice, dark urine, light stools, diffuse pruritus) of hepatotoxicity, although the latter represent comparatively late manifestations **(Table 176-3)**. The presence of such symptoms and signs mandates provisional discontinuation of potentially hepatotoxic agents; discontinuation at the onset of hepatitis symptoms reduces the risk of progression to fatal hepatitis. Biochemical testing of at least serum alanine aminotransferase (ALT) and total bilirubin levels and exclusion of other causes of these abnormalities are also indicated during treatment for those at risk for hepatotoxicity (Table 176-3). For patients with active TB, monthly mycobacterial cultures of sputum are recommended until it is certain that the organisms have been cleared and the patient has responded to therapy or until no sputum is available for culture.

TABLE 176-3

Monitoring and Clinical Management of Tuberculosis Treatment in Adults[a]

Drug	Assessment	Management
LTBI Treatment		
With hepatic risk factors[b], check ALT and bilirubin at baseline. If ALT is ≥3 × ULN or total bilirubin is >2 × ULN, defer treatment and reevaluate.		
Isoniazid	Determine whether hepatic risk factors are present. If so, obtain baseline and periodic ALT and bilirubin values.	If ALT is 5 × ULN (or 3 × ULN with symptoms)[c] or if bilirubin reaches jaundice levels (usually >2 × ULN), interrupt treatment. With normalization, consider an alternative agent.
Rifampin	Same as above	Same as above
TB Treatment		
Check ALT, bilirubin, platelets, creatinine, and hepatitis panel for all patients at baseline. If hepatic risk factors are present, check ALT and bilirubin monthly.		
Isoniazid	If ALT is >5 × ULN (or >3 × ULN with hepatitis symptoms)[c]	Obtain history of alcohol consumption and concomitant drug use. In most instances, discontinue H, Z, R, and other hepatotoxic drugs. Consider alternative agents. Obtain viral hepatitis serologies. Rechallenge: With normalization of liver enzymes, R and H may be sequentially reintroduced. With no recurrence of hepatotoxicity, Z is not resumed in many cases. Alternative rechallenge protocols have been used.
Rifampin	If primary elevation is in bilirubin and alkaline phosphatase, most likely due to rifampin	Discontinue R if total bilirubin reaches jaundice levels (usually >2 × ULN). May try to reintroduce; if not tolerated, may substitute Q

Drug	Assessment	Management
Ethambutol	Decrease in visual acuity or color vision on monthly screening	Discontinue ethambutol and repeat ocular exam. Peripheral neuropathy may be a precursor of ocular toxicity; if it occurs, consider repeat ocular exam.
Pyrazinamide	If ALT is >5 × ULN (or >3 × ULN with symptoms)[c]	Same as for H
Fluoroquinolone	If QTc prolongation is discovered incidentally on ECG	Asymptomatic QTc prolongation should prompt consideration of stopping known QT-prolonging drugs and/or close monitoring, depending on the clinical situation and degree of prolongation. Symptomatic QTc prolongation (e.g., palpitations or arrhythmias) should prompt discontinuation of drugs.
Aminoglycoside	Abnormal results on audiometry testing, BUN, creatinine, electrolytes at baseline or on monthly check	Discontinue aminoglycoside if not MDR-TB. Check audiometry and at least BUN and creatinine monthly. As appropriate, assess renal function, correct electrolytes, or seek ENT consultation.

[a]All regimens require monthly clinical monitoring. [b]Hepatic risk factors: chronic alcohol use, viral hepatitis, preexisting liver disease, pregnancy or ≤3 months postpartum, hepatotoxic medications. [c]Relevant manifestations include nausea, vomiting, abdominal pain, jaundice, or unexplained fatigue.

Abbreviations: ALT, alanine aminotransferase; BUN, blood urea nitrogen; ECG, electrocardiogram; ENT, ear, nose, and throat; H, isoniazid; LTBI, latent tuberculosis infection; MDR-TB, multidrug-resistant tuberculosis; Q, fluoroquinolone; QTc, corrected QT interval; R, rifampin; ULN, upper limit of normal; Z, pyrazinamide.

Sources: JJ Saukkonen et al: An official ATS statement: Hepatotoxicity of antituberculosis therapy. Am J Respir Crit Care Med 174:935, 2006; American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America: Treatment of tuberculosis. Am J Respir Crit Care Med 167:603, 2003.

If significant clinical improvement does not occur or the patient's condition deteriorates over the course of therapy, possibilities include treatment failure due to nonadherence, poor medication absorption, or the development of resistance. For patients co-infected with HIV and *M. tuberculosis*, IRIS, which is a diagnosis of exclusion, should also be a consideration. Drug susceptibility testing should be repeated at this point. If

resistance is documented or strongly suspected, at least two efficacious drugs to which the isolate is susceptible or which the patient has not already taken should be added to the therapeutic regimen.

Multidrug-resistant tuberculosis (MDR-TB) is defined as disease caused by a strain of *M. tuberculosis* that is resistant to both isoniazid and rifampin—the most efficacious of the first-line TB drugs. The risk of MDR-TB is elevated in patients presenting from geographic areas in which ≥5% of incident cases are MDR-TB and in patients previously treated for TB. Treatment regimens for MDR-TB generally include a late-generation fluoroquinolone and an injectable second-line agent (such as capreomycin, amikacin, or kanamycin). Regimens of at least five drugs are recommended for the treatment of MDR-TB. Both standardized and optimized/customized regimens are in use around the world. MDR-TB treatment should be initiated and monitored by clinicians with expertise in drug-resistant TB.

In 2016, the World Health Organization made a provisional recommendation for use of a regimen based on the "Bangladesh regimen" to treat specific patients with MDR-TB **(Table 176-4)**. Unlike a conventional MDR-TB treatment regimen, whose duration may be 18–20 months depending on patient response, the new Bangladesh regimen is 9–12 months in duration. Patients excluded from this short-course regimen include treatment-experienced MDR-TB patients, patients with extrapulmonary TB, and patients with phenotypic or genotypic resistance to pyrazinamide, fluoroquinolones, or second-line injectable agents. This regimen consists of a seven-drug intensive phase (kanamycin, prothionamide, isoniazid, fluoroquinolone, ethambutol, pyrazinamide, and clofazimine) and a four-drug continuation phase (fluoroquinolone, ethambutol, pyrazinamide, and clofazimine). A series of cohort studies in Bangladesh showed favorable outcomes in up to 90% of the treated MDR-TB cases. High-level fluoroquinolone resistance, particularly in the setting of initial pyrazinamide resistance, was the strongest risk factor for an unfavorable treatment outcome in these cohorts. A multicenter randomized control study with sites in South Africa, Ethiopia, Vietnam, and Mongolia is under way and aims to determine the safety and efficacy of a moxifloxacin-based Bangladesh regimen in a setting with diverse HIV and drug resistance prevalences.

TABLE 176-4

"Bangladesh Regimen" for the Treatment of Multidrug-Resistant Tuberculosis

Treatment Phase	Drugs[a]	Comments
Intensive phase	Isoniazid Ethambutol Pyrazinamide Fluoroquinolone[b] Kanamycin Prothionamide Clofazimine	All patients are hospitalized during the intensive phase, which is continued for at least 4 months or until there is sputum conversion or the patient is declared to have bacteriologic treatment failure.
Continuation phase	Fluoroquinolone Ethambutol Pyrazinamide Clofazimine	Patients are closely followed in the outpatient setting, and directly observed therapy is provided. The continuation phase is given for a total of 5 months.

[a]Daily drug doses were based on the patient's weight. [b]Ofloxacin and gatifloxacin have been the most studied fluoroquinolones in this regimen. Ofloxacin-based regimens appear to increase the prevalence of ofloxacin resistance over time from 0 to 10%. This effect was not observed with gatifloxacin-based regimens. Gatifloxacin, however, has been withdrawn from the market because of a high incidence of dysglycemia. A later-generation fluoroquinolone, such as moxifloxacin, is therefore currently preferred.

Sources: A Van Deun et al: Short, highly effective, and inexpensive standardized treatment of multidrug-resistant tuberculosis. Am J Respir Crit Care Med 182:684, 2010; KJ Augn et al: Successful "9-month Bangladesh regimen" for multidrug-resistant tuberculosis among over 500 consecutive patients. Int J Tuberc Lung Dis 18:1180, 2014; R Moodley, TR Godec: Short-course treatment for multidrug-resistant tuberculosis: The STREAM trials. Eur Respir Rev 25:29, 2016.

Extensively drug-resistant tuberculosis (XDR-TB) is defined as MDR-TB with additional resistance to any fluoroquinolone and at least one of the second-line injectable agents. Treatment of XDR-TB is individualized on the basis of complete phenotypic and, if possible, genotypic antimicrobial susceptibility testing. Therapeutic regimens for either MDR-TB or XDR-TB should be constructed with input from experienced clinicians, who should continue the management of the disease.

FIRST-LINE ANTITUBERCULOSIS DRUGS

The following discussion of individual anti-TB agents focuses on treatment of TB in adults, unless otherwise noted. Several agents are being actively investigated during the current remarkable period of drug

development for TB treatment.

Isoniazid

Isoniazid is a critical drug for treatment of both TB disease and LTBI. Isoniazid has excellent bactericidal activity against both intracellular and extracellular, actively dividing *M. tuberculosis*. This drug is bacteriostatic against slowly dividing organisms. In treatment of LTBI, isoniazid is considered the first-line agent because it is generally well tolerated, has well-established efficacy, and is inexpensive. In this setting, the drug is taken daily or intermittently (i.e., twice weekly) as DOT for 9 months. The 9-month course is more efficacious than the 6-month course (75–90% vs ≤65%), but extension of treatment to 12 months is not likely to provide further protection. A 6-month course of daily or intermittent isoniazid is considered second-line, but acceptable, therapy. A recent large open-label, multicenter, randomized controlled trial showed that weekly DOT with isoniazid and rifapentine, administered over 3 months, was not inferior to daily isoniazid given for 9 months and had a higher treatment completion rate than the single-drug regimen.

For treatment of TB disease, isoniazid is used in combination with other agents to ensure killing of both actively dividing *M. tuberculosis* and slowly growing "persister" organisms. Unless the organism is resistant, the standard regimen includes isoniazid, rifampin, ethambutol, and pyrazinamide (Table 176-2). Isoniazid is often given together with 25–50 mg of pyridoxine daily to prevent drug-related peripheral neuropathy.

MECHANISM OF ACTION

Isoniazid is a prodrug activated by the mycobacterial KatG catalase-peroxidase; isoniazid is coupled with reduced nicotinamide adenine dinucleotide (NADH). The resulting isonicotinic acyl–NADH complex blocks the mycobacterial ketoenoylreductase known as InhA, binding to its substrate and inhibiting fatty acid synthase and ultimately mycolic acid synthesis. Mycolic acids are essential components of the mycobacterial cell wall. KatG activation of isoniazid also results in the release of free radicals that have antimycobacterial activity, including nitric oxide.

The minimal inhibitory concentrations (MICs) of isoniazid for wild-type (untreated) susceptible strains are <0.1 μg/mL for *M. tuberculosis* and 0.5–2 μg/mL for *Mycobacterium kansasii*.

PHARMACOLOGY

Isoniazid is the hydrazide of isonicotinic acid, a small, water-soluble molecule. The usual adult oral daily dose of 300 mg results in peak serum levels of 3–5 μg/mL within 30 min to 2 h after ingestion—well in excess of the MICs for most susceptible strains of *M. tuberculosis*. Both oral and IM preparations of isoniazid reach effective levels in the body, although antacids and high-carbohydrate meals may interfere with oral absorption. Isoniazid diffuses well throughout the body, reaching therapeutic concentrations in body cavities and fluids, with concentrations in cerebrospinal fluid (CSF) comparable to those in serum.

Isoniazid is metabolized in the liver via acetylation by *N*-acetyltransferase 2 (NAT2) and hydrolysis. Both fast- and slow-acetylation phenotypes occur; patients who are "fast acetylators" may have lower serum levels of isoniazid, whereas "slow acetylators" may have higher levels and experience more toxicity. Satisfactory isoniazid levels are attained in the majority of homozygous fast NAT2 acetylators given a dose of 6 mg/kg and

in the majority of homozygous slow acetylators given only 3 mg/kg. Genotyping is increasingly being used to characterize isoniazid-related pharmacogenomic responses.

Isoniazid's interactions with other drugs are due primarily to its inhibition of the cytochrome P450 system. Among the drugs with significant isoniazid interactions are warfarin, carbamazepine, benzodiazepines, acetaminophen, clopidogrel, maraviroc, dronedarone, salmeterol, tamoxifen, eplerenone, and phenytoin.

DOSING

The recommended daily dose of isoniazid for the treatment of TB in the United States is 5 mg/kg for adults and 10–20 mg/kg for children, with a maximal daily dose of 300 mg for both. For intermittent therapy in adults (usually twice per week), the dose is 15 mg/kg, with a maximal daily dose of 900 mg. Isoniazid does not require dosage adjustment in patients with renal disease. When the 12-dose, 3-month weekly LTBI regimen is used, the dose of isoniazid is 15 mg/kg, with a maximal dose of 900 mg, and the drug is coadministered with rifapentine.

RESISTANCE

Although isoniazid, along with rifampin, is the mainstay of TB treatment regimens, ~7% of clinical *M. tuberculosis* isolates in the United States are resistant. Rates of primary isoniazid resistance among untreated patients are significantly higher in many populations born outside the United States. Five separate pathways for isoniazid resistance have been elucidated. Most strains have amino acid changes in either the catalase-peroxidase gene (*katG*) or the mycobacterial ketoenoylreductase gene (*inhA*). Less frequently, alterations in *kasA*, the gene for an enzyme involved in mycolic acid elongation, and loss of NADH dehydrogenase 2 activity confers isoniazid resistance. In 20–30% of isoniazid-resistant *M. tuberculosis* isolates, increased expression of efflux pump genes, such as *efpA*, *mmpL7*, *mmr*, *p55*, and the Tap-like gene *Rv1258c*, has been implicated as the underlying mechanism of resistance.

ADVERSE EFFECTS

Although isoniazid is generally well tolerated, drug-induced liver injury and peripheral neuropathy are significant adverse effects associated with this agent. Isoniazid may cause asymptomatic transient elevation of aminotransferase levels (often termed *hepatic adaptation*) in up to 20% of recipients. Other adverse reactions include rash (2%), fever (1.2%), anemia, acne, arthritic symptoms, a systemic lupus erythematosus–like syndrome, optic atrophy, seizures, and psychiatric symptoms. Symptomatic hepatitis occurs in fewer than 0.1% of persons treated with isoniazid alone for LTBI, and fulminant hepatitis with hepatic failure occurs in fewer than 0.01%. Isoniazid-associated hepatitis is idiosyncratic, but its incidence increases with age, with daily alcohol consumption, and in women who are within 3 months postpartum.

In patients who have liver disorders or HIV infection, who are pregnant or in the 3-month postpartum period, who have a history of liver disease (e.g., hepatitis B or C, alcoholic hepatitis, or cirrhosis), who use alcohol regularly, who have multiple medical problems, or who have other risk factors for chronic liver disease, the risks and benefits of treatment for LTBI should be weighed. If treatment is undertaken, these patients should have serum concentrations of ALT determined at baseline. Routine baseline hepatic ALT testing based solely on an age of >35 years is optional and depends on individual concerns. Monthly biochemical monitoring during isoniazid treatment is indicated for patients whose baseline liver function tests yield abnormal results

and for persons at risk for hepatic disease, including the groups just mentioned. Guidelines recommend that isoniazid be discontinued in the presence of hepatitis symptoms or jaundice and an ALT level three times the upper limit of normal or in the absence of symptoms with an ALT level five times the upper limit of normal (Table 176-3).

Peripheral neuropathy associated with isoniazid occurs in up to 2% of patients given 5 mg/kg. Isoniazid appears to interfere with pyridoxine (vitamin B_6) metabolism. The risk of isoniazid-related neurotoxicity is greatest for patients with preexisting disorders that also pose a risk of neuropathy, such as HIV infection; for those with diabetes mellitus, alcohol abuse, or malnutrition; and for those simultaneously receiving other potentially neuropathic medications, such as stavudine. These patients should be given prophylactic pyridoxine (25–50 mg/d).

Rifampin

Rifampin is a semisynthetic derivative of *Amycolatopsis rifamycinica* (formerly known as *Streptomyces mediterranei*). The most active antimycobacterial agent available, rifampin is the keystone of first-line treatment for TB. Introduced in 1968, this drug eventually permitted dramatic shortening of the TB treatment course. Rifampin has both sterilizing and bactericidal activity against dividing and nondividing *M. tuberculosis*. The drug is also active against an array of other organisms, including some gram-positive and gram-negative bacteria, *Legionella*, *M. kansasii*, and *Mycobacterium marinum*.

Rifampin, administered for 4 months, is also an alternative agent to isoniazid for the treatment of LTBI, although efficacy data are scant at this time. A 3-month course of rifampin alone has been found to be similar in efficacy to a 6-month course of isoniazid. Although the efficacy of the 4-month regimen of rifampin is still under study, comparison of this regimen with 9 months of isoniazid in randomized safety and tolerability studies suggests fewer adverse events, including hepatotoxicity; less treatment interruption; a higher completion rate; and greater cost-effectiveness with the rifampin regimen.

MECHANISM OF ACTION

Rifampin exerts both intracellular and extracellular bactericidal activities. Like other rifamycins, rifampin specifically binds to and inhibits mycobacterial DNA-dependent RNA polymerase, blocking RNA synthesis. Susceptible strains of *M. tuberculosis* as well as *M. kansasii* and *M. marinum* are inhibited by rifampin concentrations of 1 μg/mL.

PHARMACOLOGY

Rifampin is a fat-soluble, complex macrocyclic molecule readily absorbed after oral administration. Serum levels of 10–20 μg/mL are achieved 2.5 h after the usual adult oral dose of 10 mg/kg (given without food). Rifampin has a half-life of 1.5–5 h. The drug distributes well throughout most body tissues, including CSF. Rifampin turns body fluids such as urine, saliva, sputum, and tears a reddish-orange color—an effect that offers a simple means of assessing patients' adherence to this medication. Rifampin is excreted primarily through the bile and enters the enterohepatic circulation; <30% of a dose is renally excreted.

As a potent inducer of the hepatic cytochrome P450 system, rifampin can decrease the half-life of some drugs, such as digoxin, warfarin, phenytoin, prednisone, cyclosporine, methadone, oral contraceptives, clarithromycin, azole antifungal agents, quinidine, antiretroviral protease inhibitors, and non-nucleoside reverse transcriptase inhibitors. The Centers for Disease Control and Prevention (CDC) has issued guidelines for the management of drug interactions during treatment of HIV and *M. tuberculosis* co-infection *(www.cdc.gov/tb/)*.

DOSING

The daily dosage of rifampin is 10 mg/kg for adults and 10–20 mg/kg for children, with a maximum of 600 mg/d for both. The drug is given once daily, twice weekly, or three times weekly. No adjustments of dose or frequency are necessary in patients with renal insufficiency.

RESISTANCE

Resistance to rifampin in *M. tuberculosis, Mycobacterium leprae*, and other organisms is the consequence of spontaneous, mostly missense point mutations in a core region of the bacterial gene coding for the β subunit of RNA polymerase (*rpoB*). RNA polymerase altered in this manner is no longer subject to inhibition by rifampin. Most rapidly and slowly growing NTM harbor intrinsic resistance to rifampin, for which the mechanism has yet to be determined.

ADVERSE EFFECTS

Adverse events associated with rifampin are infrequent and generally mild. Hepatotoxicity due to rifampin alone is uncommon in the absence of preexisting liver disease and often consists of isolated hyperbilirubinemia rather than aminotransferase elevation. Other adverse reactions include rash, pruritus, gastrointestinal symptoms, and pancytopenia. Rarely, a hypersensitivity reaction may occur with intermittent therapy, manifesting as fever, chills, malaise, rash, and—in some instances—renal and hepatic failure.

Ethambutol

Ethambutol is a bacteriostatic antimycobacterial agent first synthesized in 1961. A component of the standard first-line regimen, ethambutol provides synergy with the other drugs in the regimen and is generally well tolerated. Susceptible species include *M. tuberculosis, M. marinum, M. kansasii*, and organisms of the *Mycobacterium avium* complex (MAC); however, among first-line drugs, ethambutol is the least potent against *M. tuberculosis*. This agent is also used in combination with other agents in the continuation phase of treatment when patients cannot tolerate isoniazid or rifampin or are infected with organisms resistant to either of the latter drugs.

MECHANISM OF ACTION

Ethambutol is bacteriostatic against *M. tuberculosis*. Its primary mechanism of action is the inhibition of the arabinosyltransferases involved in cell wall synthesis, which probably inhibits the formation of arabinogalactan and lipoarabinomannan. The MIC of ethambutol for susceptible strains of *M. tuberculosis* is 0.5–2 μg/mL.

PHARMACOLOGY AND DOSING

From a single dose of ethambutol, 75–80% is absorbed within 2–4 h of administration. Serum levels peak at 2–4 μg/mL after the standard adult daily dose of 15 mg/kg. Ethambutol is well distributed throughout the body except in the CSF; a dosage of 25 mg/kg is necessary for attainment of a CSF level half of that in serum. For intermittent therapy, the dosage is 25–35 mg/kg thrice weekly. To prevent toxicity, the dosage must be lowered and the frequency of administration reduced for patients with renal insufficiency.

ADVERSE EFFECTS

Ethambutol is usually well tolerated and has no significant interactions with other drugs. Optic neuritis, the most serious adverse effect reported, typically presents as reduced visual acuity, central scotoma, and loss of the ability to see green (or, less commonly, red). The cause of this neuritis is unknown, but it may be due to an effect of ethambutol on the amacrine and bipolar cells of the retina. Symptoms typically develop several months after initiation of therapy, but ocular toxicity soon after initiation of ethambutol has been described. The risk of ocular toxicity is dose dependent, with occurrence in 1–5% of patients, and can be increased by renal insufficiency. The routine use of ethambutol in younger children is not recommended because monitoring for visual complications can be difficult. If drug-resistant TB is suspected, ethambutol can be used for treatment of children.

All patients starting therapy with ethambutol should have a baseline test for visual acuity, visual fields, and color vision and should undergo an examination of the optic fundus. Visual acuity and color vision should be monitored monthly or less often as needed. Cessation of ethambutol in response to early symptoms of ocular toxicity usually results in reversal of the deficit within several months. Recovery of all visual function may take up to 1 year. In the elderly and in patients whose symptoms are not recognized early, deficits may be permanent. Some experts think that supplementation with hydroxycobalamin (vitamin B_{12}) is beneficial for patients with ethambutol-related ocular toxicity. Other adverse effects of ethambutol are rare. Peripheral sensory neuropathy occurs in rare instances.

RESISTANCE

Ethambutol resistance in *M. tuberculosis* and NTM is associated primarily with missense mutations in the *embB* gene that encodes for arabinosyltransferase. Mutations have been found in resistant strains at codon 306 in 50–70% of cases. Mutations at *embB*306 can cause significantly increased MICs of ethambutol, resulting in clinical resistance.

Pyrazinamide

A nicotinamide analog, pyrazinamide is an important bactericidal drug used in the initial phase of TB treatment. Its administration for the first 2 months of therapy with rifampin and isoniazid allows treatment duration to be shortened from 9 to 6 months and decreases rates of relapse.

MECHANISM OF ACTION

Pyrazinamide's antimycobacterial activity is essentially limited to *M. tuberculosis*. The drug is more active against slowly replicating organisms than against actively replicating organisms. Pyrazinamide is a prodrug that is converted by the mycobacterial pyrimidase to the active form, pyrazinoic acid (POA). This agent is active only in acidic environments (pH <6.0), as are found within phagocytes or granulomas. The exact

mechanism of action of POA is unclear, but fatty acid synthetase I may be the primary target in *M. tuberculosis*. Susceptible strains of *M. tuberculosis* are inhibited by pyrazinamide concentrations of 16–50 μg/mL at pH 5.5.

PHARMACOLOGY AND DOSING

Pyrazinamide is well absorbed after oral administration, with peak serum concentrations of 20–60 μg/mL at 1–2 h after ingestion of the recommended adult daily dose of 15–30 mg/kg (maximum, 2 g/d). It distributes well to various body compartments, including CSF, and is an important component of treatment for tuberculous meningitis. The serum half-life of the drug is 9–11 h with normal renal and hepatic function. Pyrazinamide is metabolized in the liver to POA, 5-hydroxypyrazinamide, and 5-hydroxy-POA. A high proportion of pyrazinamide and its metabolites (~70%) is excreted in the urine. The dosage must be adjusted according to the level of renal function in patients with reduced creatinine clearance.

ADVERSE EFFECTS

At the higher dosages used previously, hepatotoxicity was seen in as many as 15% of patients treated with pyrazinamide. However, at the currently recommended dosages, hepatotoxicity now occurs less commonly when this drug is administered with isoniazid and rifampin during the treatment of TB. Older age, active liver disease, HIV infection, and low albumin levels may increase the risk of hepatotoxicity. The use of pyrazinamide with rifampin for the treatment of LTBI is no longer recommended because of unacceptable rates of hepatotoxicity and death in this setting. Hyperuricemia is a common adverse effect of pyrazinamide therapy that usually can be managed conservatively. Clinical gout is rare.

Although pyrazinamide is recommended by international TB organizations for routine use in pregnancy, it is not recommended in the United States because of inadequate teratogenicity data.

RESISTANCE

The basis of pyrazinamide resistance in *M. tuberculosis* is a mutation in the *pncA* gene coding for pyrazinamidase, the enzyme that converts the prodrug to active POA. Resistance to pyrazinamide is associated with loss of pyrazinamidase activity, which prevents conversion of pyrazinamide to POA. Of pyrazinamide-resistant *M. tuberculosis* isolates, 72–98% have mutations in *pncA*. Conventional methods of testing for susceptibility to pyrazinamide may produce both false-negative and false-positive results because the high-acidity environment required for the drug's activation also inhibits the growth of *M. tuberculosis*. There is some controversy as to the clinical significance of in vitro pyrazinamide resistance.

OTHER FIRST-LINE DRUGS

Rifabutin

Rifabutin, a semisynthetic derivative of rifamycin S, inhibits mycobacterial DNA-dependent RNA polymerase. Rifabutin is recommended in place of rifampin for the treatment of TB in HIV-co-infected individuals who are taking protease inhibitors or nonnucleoside reverse transcriptase inhibitors, particularly nevirapine. A recent study in India showed better TB treatment outcomes in HIV-co-infected patients given daily rifabutin plus atazanavir/ritonavir than in those given thrice-weekly rifabutin plus atazanavir/ritonavir. Rifabutin's effect on hepatic enzyme induction is less pronounced than that of rifampin. Protease inhibitors may cause significant

increases in rifabutin levels through inhibition of hepatic metabolism. Rifabutin is more active in vitro than rifampin against MAC organisms and other NTM, but its clinical superiority has not been established.

PHARMACOLOGY

Like rifampin, rifabutin is lipophilic and is absorbed rapidly after oral administration, reaching peak serum levels 2–4 h after ingestion. Rifabutin distributes best to tissues, reaching levels 5–10 times higher than those in plasma. Unlike rifampin, rifabutin and its metabolites are partially cleared by the hepatic microsomal system. Rifabutin's slow clearance results in a mean serum half-life of 45 h—much longer than the 3- to 5-h half-life of rifampin. Clarithromycin (but not azithromycin) and fluconazole appear to increase rifabutin levels by inhibiting hepatic metabolism.

ADVERSE EFFECTS

The most common adverse effects of rifabutin treatment are gastrointestinal; other reactions include rash, headache, asthenia, chest pain, myalgia, and insomnia. Less common adverse reactions include fever, chills, a flu-like syndrome, anterior uveitis, hepatitis, *Clostridium difficile*–associated diarrhea, a diffuse polymyalgia syndrome, and yellow skin discoloration ("pseudo-jaundice"). Laboratory abnormalities include neutropenia, leukopenia, thrombocytopenia, and increased levels of liver enzymes. Rifabutin appears to be better tolerated by the majority (72%) of adult TB patients who have developed rifampin-related adverse effects. Female patients, those co-infected with hepatitis B or hepatitis C, and those with rifampin-related arthralgias, dermatologic reactions, and cholestasis are more likely to develop mild to severe rifabutin-related adverse effects.

RESISTANCE

Similar to rifampin resistance, rifabutin resistance is mediated by mutations in *rpoB*.

Rifapentine

Rifapentine is a semisynthetic cyclopentyl rifamycin, sharing a mechanism of action with rifampin. Rifapentine is lipophilic and has a prolonged half-life that permits weekly or twice-weekly dosing. Therefore, rifapentine is the subject of intensive clinical investigation aimed at determining optimal dosing and frequency of administration. Currently, it is an alternative to rifampin in the continuation phase of treatment for noncavitary drug-susceptible pulmonary TB in HIV-seronegative patients who have negative sputum smears at completion of the initial phase of treatment. When administered in these specific circumstances, rifapentine (10 mg/kg, up to 600 mg) is given once weekly with isoniazid. Because of higher rates of relapse, this regimen is not recommended for patients with TB disease and HIV co-infection; moreover, it has not been approved for children <12 years of age. In a phase 2 study, substituting daily rifapentine for rifampin yielded higher rates of sputum sterilization after 2 months of intensive treatment. Higher doses of rifapentine (20 mg/kg vs 10 mg/kg) had better results and were safe and well tolerated. Regimens containing high doses of rifapentine are being evaluated to see whether they can shorten the TB treatment course to <6 months.

A large randomized controlled trial demonstrated that, for LTBI, a 12-dose (3-month) regimen of weekly DOT with a weight-based dose of isoniazid and rifapentine was noninferior to daily isoniazid for 9 months. Although the rate of permanent drug discontinuation due to adverse events was higher with

rifapentine/isoniazid, this regimen had a higher treatment completion rate than daily isoniazid in this study. The efficacy of this combination regimen in HIV-infected individuals not receiving ART and in children <12 years of age is under study. A recent randomized controlled study of HIV-co-infected patients not receiving ART showed that weekly rifapentine/isoniazid or twice-weekly rifampin/isoniazid for 12 weeks or continuous isoniazid was not superior to 6 months of isoniazid. The regimen is not recommended for pregnant women, for persons with hypersensitivity reactions to isoniazid or rifampin, or for HIV-infected individuals taking ART.

PHARMACOLOGY

Rifapentine's absorption is improved when the drug is taken with food. After oral administration, rifapentine reaches peak serum concentrations in 5–6 h and achieves a steady state in 10 days. The half-life of rifapentine and its active metabolite, 25-desacetyl rifapentine, is ~13 h. The administered dose is excreted via the liver (70%).

ADVERSE EFFECTS

The adverse-effects profile of rifapentine is similar to that of other rifamycins. Rifapentine is teratogenic in animal models and is relatively contraindicated in pregnancy.

RESISTANCE

Rifapentine resistance is mediated by mutations in *rpoB*. Mutations that cause resistance to rifampin also cause resistance to rifapentine.

Streptomycin

Streptomycin was the first antimycobacterial agent used for the treatment of TB. Derived from *Streptomyces griseus*, streptomycin is bactericidal against dividing *M. tuberculosis* organisms but has only low-level early bactericidal activity. This drug is administered only by the IM and IV routes. In developed nations, streptomycin is used infrequently because of its toxicity, the inconvenience of injections, and drug resistance. In developing countries, however, streptomycin is used because of its low cost.

MECHANISM OF ACTION

Streptomycin inhibits protein synthesis by binding at a site on the 30S mycobacterial ribosome.

PHARMACOLOGY AND DOSING

Serum levels of streptomycin peak at 25–45 μg/mL after a 1-g dose. This agent penetrates poorly into the CSF, reaching levels that are only 20% of serum levels. The usual daily dose of streptomycin (given IM either daily or 5 days per week) is 15 mg/kg for adults and 20–40 mg/kg for children, with a maximum of 1 g/d for both. For patients ≥60 years of age, 10 mg/kg is the recommended daily dose, with a maximum of 750 mg/d. Because streptomycin is eliminated almost exclusively by the kidneys, its use in patients with renal impairment should be avoided or implemented with caution, with lower doses and less frequent administration.

ADVERSE EFFECTS

Adverse reactions occur frequently with streptomycin (10–20% of patients). Ototoxicity (primarily vestibulotoxicity), neuropathy, and renal toxicity are the most common and the most serious reactions. Renal toxicity, usually manifested as nonoliguric renal failure, is less common with streptomycin than with other

frequently used aminoglycosides, such as gentamicin. Manifestations of vestibular toxicity include loss of balance, vertigo, and tinnitus. Patients receiving streptomycin must be monitored carefully for these adverse effects, undergoing audiometry at baseline and monthly thereafter.

RESISTANCE

Spontaneous mutations conferring resistance to streptomycin are relatively common, occurring in 1 in 10^6 organisms. In the two-thirds of streptomycin-resistant *M. tuberculosis* strains exhibiting high-level resistance, mutations have been identified in one of two genes: a 16S rRNA gene (*rrs*) or the gene encoding ribosomal protein S12 (*rpsL*). Both targets are believed to be involved in streptomycin ribosomal binding. However, low-level resistance, which is seen in about one-third of resistant isolates, has no associated resistance mutation. A gene (*gidB*) that confers low-level resistance to streptomycin has been identified. Strains of *M. tuberculosis* resistant to streptomycin generally are not cross-resistant to capreomycin or amikacin. Streptomycin is not used for the treatment of MDR-TB or XDR-TB because of (1) the high prevalence of streptomycin resistance among strains resistant to isoniazid and (2) the unreliability of drug susceptibility testing.

SECOND-LINE ANTITUBERCULOSIS DRUGS

Second-line anti-TB agents are indicated for treatment of drug-resistant TB, for patients who are intolerant or allergic to first-line agents, and when first-line supplemental agents are unavailable.

Fluoroquinolones

Fluoroquinolones inhibit mycobacterial DNA gyrase and topoisomerase IV, preventing cell replication and protein synthesis, and are bactericidal. They are also being investigated for their potential to shorten the course of treatment for TB. A single randomized trial showed that a regimen of daily moxifloxacin/rifampin/pyrazinamide/ethambutol for 2 months followed by once-weekly 1200 mg rifapentine plus 400 mg of moxifloxacin for 4 continuation-phase months was associated with relapse rates similar to those documented with the standard 6-month regimen given daily in patients with drug-sensitive TB. Gatifloxacin has fallen out of favor because of significant dysglycemia. Ciprofloxacin and ofloxacin are no longer recommended for the treatment of TB because of poor efficacy. Despite documented resistance to early-generation fluoroquinolones (e.g., ofloxacin and ciprofloxacin), use of a later-generation fluoroquinolone in patients with XDR-TB has been associated with favorable outcomes. Fluoroquinolones are also considered safe alternatives for patients who develop treatment-limiting adverse effects from first-line agents. Levofloxacin and moxifloxacin have both been used effectively in the treatment of MDR-TB. The optimal dose of levofloxacin for this indication is being actively studied, but doses of at least 750 mg are commonly used.

The fluoroquinolones are well absorbed orally, reach high serum levels, and distribute well into body tissues and fluids. Their absorption is decreased by co-ingestion with products containing multivalent cations, such as antacids. Adverse effects are relatively infrequent (0.5–10% of patients) and include gastrointestinal intolerance, rashes, dizziness, and headache. Most studies of fluoroquinolone side effects have been based on relatively short-term administration for bacterial infections, but trials have now shown the relative safety

and tolerability of fluoroquinolones administered for months during TB treatment in adults. Although the potential to prolong the QTc interval, leading to cardiac arrhythmias, has been a source of concern with fluoroquinolones, cessation of treatment due to this adverse effect is rare. Because the benefits may outweigh the risks in treatment of drug-resistant TB, there is increasing interest in the use of fluoroquinolones in children, which has traditionally been avoided because of the risks of tendon rupture and cartilage damage.

Multiple courses of empirical fluoroquinolone therapy for presumed community-acquired pneumonia are associated with delayed diagnosis of active pulmonary TB and increased fluoroquinolone resistance in *M. tuberculosis*. Mutations in the genes encoding for DNA gyrase (*gyrA* and *gyrB*) are implicated in the majority of cases—but not all cases—of clinical resistance to fluoroquinolones.

Injectable Drugs

CAPREOMYCIN

Capreomycin, a cyclic peptide antibiotic derived from *Streptomyces capreolus*, is an important second-line agent used for treatment of MDR-TB and XDR-TB. Capreomycin is administered by the IM route; an inhaled preparation is under study. A dose of 15 mg/kg per day is given five to seven times per week (maximal daily dose, 1 g) and results in peak blood levels of 20–40 µg/mL. The dosage may be reduced to 1 g two or three times per week 2–4 months after mycobacterial cultures become negative. For individuals ≥60 years of age, the dose should be reduced to 10 mg/kg per day (maximal daily dose, 750 mg). For patients with renal insufficiency, the drug should be given intermittently and at lower dosage (12–15 mg/kg two or three times per week). A minimal duration of 3 months is recommended for MDR-TB treatment. Penetration of capreomycin into the CSF is believed to be poor.

The mechanism of capreomycin's action is not well understood but involves interference with the mycobacterial ribosome and inhibition of protein synthesis. Resistance to capreomycin is associated with mutations that inactivate a ribosomal methylase (*tlyA*) or that encode genes for the 16S ribosomal subunit (*rrs*). Cross-resistance to kanamycin and amikacin is common with *rrs* but not always with *tylA* mutations. The *rrs* A1401G mutation is now considered a strong predictor of cross-resistance among all three second-line injectable drugs: capreomycin, kanamycin, and amikacin.

Adverse effects of capreomycin are relatively common. Significant hypokalemia and hypomagnesemia as well as oto- and renal toxicity have been reported.

AMIKACIN AND KANAMYCIN

Amikacin and kanamycin are aminoglycosides that exert mycobactericidal activity by binding to the 16S ribosomal subunit. The spectrum of antibiotic activity for amikacin and kanamycin includes *M. tuberculosis*, several NTM species, and aerobic gram-negative and gram-positive bacteria. Although amikacin is highly active against *M. tuberculosis*, it is used only infrequently because of its significant side effects. The usual daily adult dosage of both amikacin and kanamycin is 15–30 mg/kg given IM or IV (maximal daily dose, 1 g), with a reduction to 10 mg/kg for patients ≥60 years old. For patients with renal insufficiency, the dose and frequency should be reduced (12–15 mg/kg two or three times per week). Mycobacterial resistance is due to

mutations in the genes encoding the 16S ribosomal RNA gene. Cross-resistance among kanamycin, amikacin, and capreomycin is common. Isolates resistant to streptomycin are frequently susceptible to amikacin or kanamycin. Adverse effects of amikacin include ototoxicity (in up to 10% of recipients, with auditory dysfunction occurring more commonly than vestibulotoxicity), nephrotoxicity, and neurotoxicity. Kanamycin has a similar side-effects profile, but adverse reactions are thought to be less frequent and less severe.

Other Second-Line Agents

ETHIONAMIDE

Ethionamide is a derivative of isonicotinic acid. Its mechanism of action is through inhibition of the *inhA* gene product enoyl–acyl carrier protein (acp) reductase, which is involved in mycolic acid synthesis. Ethionamide is bacteriostatic against metabolically active *M. tuberculosis* and some NTM. It is used in the treatment of drug-resistant TB, but its use is limited by severe gastrointestinal reactions (including abdominal pain, nausea, and vomiting) as well as significant central and peripheral neurologic side effects, reversible hepatitis (in ~5% of recipients), hypersensitivity reactions, and hypothyroidism. Ethionamide should be taken with food to reduce gastrointestinal effects and with pyridoxine (50–100 mg/d) to limit neuropathic side effects.

CYCLOSERINE

Cycloserine is an analog of the amino acid D-alanine and prevents bacterial cell-wall synthesis. It inhibits the action of enzymes, including alanine racemase, that are involved in the production of peptidoglycans. Cycloserine is active against a range of bacteria, including *M. tuberculosis*. Mechanisms of mycobacterial resistance are not well understood, but overexpression of alanine racemase can confer resistance in *Mycobacterium smegmatis*. Cycloserine is well absorbed after oral administration and is widely distributed throughout body fluids, including CSF. The usual adult dosage is 250 mg two or three times per day. Serious potential side effects include seizures and psychosis (with suicide in some cases), peripheral neuropathy, headache, somnolence, and allergic reactions. Drug levels are monitored to achieve optimal dosing and to reduce the risk of adverse effects, especially in patients with renal failure. Cycloserine should be administered as DOT only with caution and with support from experienced TB physicians to patients with epilepsy, active alcohol abuse, severe renal insufficiency, or a history of depression or psychosis.

PARA-AMINOSALICYLIC ACID

Para-aminosalicylic acid (PAS, 4-aminosalicylic acid) is an oral agent used in the treatment of MDR-TB and XDR-TB. Its bacteriostatic activity is due to inhibition of folate synthesis and of iron uptake. PAS has relatively little activity as an anti-TB agent. Adverse effects may include high-level nausea, vomiting, and diarrhea. PAS may cause hemolysis in patients with glucose-6-phosphate dehydrogenase deficiency. The drug should be taken with acidic foods to improve absorption. Enteric-coated PAS granules (4 g orally every 8 h) appear to be better tolerated than other formulations and produce higher therapeutic blood levels. PAS has a short half-life (1 h), and 80% of the dose is excreted in the urine.

CLOFAZIMINE

Clofazimine is a fat-soluble riminophenazine dye used primarily in the treatment of leprosy worldwide. It is currently gaining popularity in the management of MDR-TB and XDR-TB because of its low cost and its

intracellular and extracellular activity. By increasing reactive oxidant species and causing membrane destabilization, clofazimine may promote killing of antibiotic-tolerant *M. tuberculosis* persister organisms. In addition to antimicrobial activity, the drug has other pharmacologic activities, such as anti-inflammatory, pro-oxidative, and immunopharmacologic properties. Clofazimine has a half-life of ~70 days in humans, and average steady-state concentrations are achieved at ~1 month. Intake with fatty meals can improve its low and variable rates of absorption (45–62%). Common side effects include gastrointestinal intolerance, and reversible orange to brownish discoloration of skin, bodily fluids, and secretions. Dose adjustment may be necessary in patients with severe hepatic impairment. Clofazimine is being studied as part of a regimen developed in Bangladesh (Table 176-4) for potential shortening of the MDR-TB treatment course. A meta-analysis suggested that inclusion of clofazimine in a multidrug regimen for treatment of MDR-TB was associated with a favorable outcome. Newer analogues with improved pharmacokinetics and alternative formulations of clofazimine (liposomal, nanosuspension, inhalational) are being studied.

NEWER ANTITUBERCULOSIS DRUGS

Oxazolidinones

Linezolid is an oxazolidinone used primarily for the treatment of drug-resistant gram-positive bacterial infections. However, this drug is active in vitro against *M. tuberculosis* and NTM. Several case series have suggested that linezolid may help clear mycobacteria relatively rapidly when included in a regimen for the treatment of complex cases of MDR-TB and XDR-TB. Linezolid's mechanism of action is disruption of protein synthesis by binding to the 50S bacterial ribosome. Linezolid has nearly 100% oral bioavailability, with good penetration into tissues and fluids, including CSF. Clinical resistance to linezolid has been reported and is typically associated with mutations in the 23S rRNA and in two ribosomal proteins, L3 (*rplC*) and L4 (*rplD*). Adverse effects may include optic and peripheral neuropathy, pancytopenia, and lactic acidosis and are usually associated with higher doses. Linezolid is a weak monoamine oxidase inhibitor and can be associated with the serotonin syndrome when given concomitantly with serotonergic drugs (primarily antidepressants such as selective serotonin-reuptake inhibitors). It has been shown that ~80% of patients with MDR-TB or XDR-TB can be successfully treated with linezolid-containing, individualized anti-TB regimens based on drug sensitivity testing. Replacement of ethambutol with linezolid for 2–4 weeks during the intensive phase of treatment of drug-susceptible TB is currently being evaluated for possible faster sputum conversion and a shorter treatment regimen. For MDR-TB treatment, linezolid is usually administered at a dose of 600 mg (or less in some cases) once daily, which appears to be effective. A single daily dose is associated with fewer adverse events than twice-a-day dosing.

PNU 100480 and AZD 5847, modified versions of oxazolidinones and protein synthesis inhibitors, are undergoing phase 1 trials and appear to have greater efficacy than linezolid against *M. tuberculosis*. However, the adverse-effects profile of these compounds compared with that of linezolid needs further investigation.

Amoxicillin-Clavulanate and Carbapenems

β-Lactam agents are largely ineffective for the treatment of *M. tuberculosis* because of resistance conferred by a hydrolyzing class A β-lactamase. Because clavulanate may theoretically inhibit the β-lactamase, amoxicillin-clavulanate has been used in the treatment of MDR-TB; however, it is a comparatively weak agent. Carbapenems are poor substrates for class A β-lactamases found in *M. tuberculosis*. Accordingly, meropenem and imipenem have in vitro activity against *M. tuberculosis*, and their use to treat MDR-TB and XDR-TB has been reported anecdotally. Nevertheless, the need to administer these carbapenems intravenously and the lack of information on the drugs' long-term side effects have restricted their use to certain severe cases only. A newer agent, faropenem, has high oral bioavailability, is more resistant to hydrolysis by β-lactamases, and causes fewer adverse reactions. It is being evaluated in combination with linezolid and moxifloxacin as a first-line therapy for drug-resistant and drug-susceptible TB in children.

Diarylquinolines

Bedaquiline (TMC207 or R207910) is a new diarylquinoline with a novel mechanism of action: inhibition of the mycobacterial ATP synthetase proton pump. Bedaquiline is bactericidal for drug-susceptible and MDR strains of *M. tuberculosis*. Resistance has been reported and is due to point mutations in the *atpE* gene encoding for subunit c of ATP synthetase. A phase 2 randomized controlled clinical trial in MDR-TB patients demonstrated substantial improvement in 2-month culture-conversion rates as well as a reduction in acquired resistance to companion drugs. This drug is metabolized by the hepatic cytochrome CYP3A4. Rifampin lowers bedaquiline levels by 50%, and protease inhibitors also interact significantly with this drug. Because efavirenz induces CYP3A4, there is concern about lower bedaquiline levels with coadministration. In a study of co-treatment with bedaquiline and efavirenz in healthy volunteers, bedaquiline levels were reduced by only 20%; however, in a study modeling chronic coadministration of these two drugs, the reduction in bedaquiline levels was estimated to be 50%, leading many national TB programs to avoid efavirenz coadministration with bedaquiline.

The oral bioavailability of bedaquiline appears to be excellent. The dosage is 400 mg/d for the first 2 weeks and then 200 mg thrice weekly. The elimination half-life is long (>14 days). A single dose of this drug can inhibit the growth of *M. tuberculosis* for up to 1 week through a combination of long plasma half-life, high-level tissue penetration, and long tissue half-life. Bedaquiline added to a background regimen improved the 2-month sputum culture–conversion rate in multicenter, randomized placebo-controlled trials, and these results led to approval by the U.S. Food and Drug Administration (FDA). However, the death rate in one trial was higher in the bedaquiline arm than in the control arm (11.4% vs 2.5%); the result was a "black box" warning from the FDA, which also included QT prolongation. The CDC has made a provisional recommendation for the use of bedaquiline for 24 weeks in adults with laboratory-confirmed pulmonary MDR-TB when no other effective treatment regimen can be provided.

Nitroimidazoles

The prodrugs delamanid (OPC-67683) and PA 824 are novel nitro-dihydro-imidazooxazole derivatives that are activated by *M. tuberculosis*–specific flavin-dependent nitroreductases and whose antimycobacterial activity is attributable to inhibition of mycolic acid biosynthesis. These drugs are currently in phase 2 clinical trials and show potential in shortening treatment duration through their activity against nonreplicating drug-susceptible and drug-resistant mycobacteria. Delamanid was shown in a randomized, placebo-controlled,

multinational clinical trial to significantly improve the culture-conversion rate at 2 months. QT prolongation occurred significantly more often in delamanid-treated patients, but no clinically relevant events were reported.

Diamines

SQ109, an ethambutol analogue with a 1,2-diamine pharmacophore, is the most promising of the diamines for TB treatment. It is activated by mycobacterial cytochrome enzymes and inhibits mycobacterial cell-wall synthesis by an unknown mechanism. It has a high tissue protein-binding capacity with a very long half-life (~61 h) in humans. In vitro studies have demonstrated that SQ109 has low MICs against both susceptible and resistant *M. tuberculosis* strains as well as a synergistic effect when administered with isoniazid and rifampin. The drug is in clinical trials for TB treatment.

Pyrroles

LL3858, a pyrrole derivative, has entered clinical trials examining its utility in the treatment of drug-susceptible and drug-resistant TB. The drug's mechanism of action is unknown. However, because it is active against *M. tuberculosis* strains that are resistant to available anti-TB drugs, its target is thought to differ from those of currently used drugs.

NONTUBERCULOUS MYCOBACTERIA

More than 150 species of NTM have been identified. Only a minority of these environmental organisms, which are extensively found in soil and water, are important human pathogens. NTM cause extensive disease primarily in persons with preexisting pulmonary disease or immunocompromise but can also cause nodular/bronchiectatic disease in otherwise seemingly healthy hosts. Disseminated infections with NTM are common in immunocompromised individuals. NTM are also important causes of skin and soft-tissue infections in surgical settings. The two major classes of NTM are the slow-growing and rapidly growing species; subcultures of the latter grow within 1 week. The growth characteristics of NTM have diagnostic, therapeutic, and prognostic implications. The rate of growth can provide useful preliminary information within a specific clinical context, in that growth within 2–3 weeks is much more likely to indicate an NTM than *M. tuberculosis*. When NTM do grow from cultures, colonization should be distinguished from active disease in order to optimize the risk and benefit of prolonged treatment with multiple medications. According to the recommendations of the American Thoracic Society and the Infectious Diseases Society of America, significant clinical manifestations and/or radiographic evidence of progressive disease consistent with NTM infection as well as either reproducible sputum culture results or a single positive culture are required for the diagnosis of NTM pulmonary disease. Isolation of NTM from blood or from an infected-appearing extrapulmonary site, such as soft tissue or bone, is usually indicative of disseminated or local NTM infection (Chap. 175). Treatment of NTM disease is prolonged and requires multiple medications. Side effects of the regimens employed are common, and intermittent therapy is often used to mitigate these adverse events. Treatment regimens depend on the NTM species, the extent or type of disease, and—to some degree—drug susceptibility test results.

THERAPEUTIC CONSIDERATIONS FOR SPECIFIC NTM

Slowly Growing Mycobacteria

Slowly growing mycobacteria can be divided into three categories based on their pigment-producing capabilities and—if they do produce pigment—their requirement for light to do so. *Photochromogens*, including *M. marinum* and *M. kansasii*, can produce yellowish-orange pigment only when exposed to light. *Scotochromogens*, including *Mycobacterium gordonae* and *Mycobacterium scrofulaceum*, can make pigment regardless of light exposure. MAC organisms and *Mycobacterium ulcerans* are *nonchromogens*—i.e., are incapable of making pigment irrespective of light exposure.

MYCOBACTERIUM AVIUM COMPLEX

Among the NTM, MAC organisms most commonly cause human disease. In immunocompetent hosts, MAC species are most often found in conjunction with underlying significant lung disease, such as chronic obstructive pulmonary disease or bronchiectasis. For patients with nodular or bronchiectatic MAC lung disease, an initial regimen consisting of clarithromycin or azithromycin, rifampin or rifabutin (the latter is preferred for HIV patients receiving ART), and ethambutol is given three times per week. A daily regimen of these three drugs, with consideration of amikacin or streptomycin in the initial treatment phase, is recommended for patients with fibrocavitary MAC lung disease or severe nodular/bronchiectatic disease. Routine initial testing for macrolide resistance is recommended, as is testing at 6 months with a failing regimen (i.e., with cultures persistently positive for NTM).

In immunocompromised individuals, disseminated MAC infection is generally treated with clarithromycin, ethambutol, and rifabutin. Azithromycin may be substituted in patients unable to tolerate clarithromycin. Amikacin and fluoroquinolones are often used in salvage regimens. Treatment for disseminated MAC infection in AIDS patients may be lifelong in the absence of immune reconstitution. Therapy is recommended for at least 12 months after culture conversion and at least 6 months of effective immune reconstitution with ART (CD4+ T cell count, >100/μL).

Surgical resection should be considered for individuals whose infection is localized to one lung, who have adequate lung function to tolerate lung resection, who have had a poor response to medical therapy, and/or who have developed macrolide-resistant MAC disease.

MYCOBACTERIUM KANSASII

M. kansasii is the second most common NTM causing human disease. It is also the second most common cause of NTM pulmonary disease in the United States, where it is most commonly reported in the southeastern region. *M. kansasii* infection can be treated with isoniazid, rifampin, and ethambutol; therapy continues for at least 18 months or for 12 months after culture conversion. The American Thoracic Society and the Infectious Diseases Society of America recommend routine susceptibility testing to rifampin only. Resistance to isoniazid and ethambutol can be acquired during therapy but is usually associated with rifampin resistance as well. Rifampin-resistant *M. kansasii* is treated with a three-drug regimen of second-line agents such as clarithromycin, ethambutol, rifabutin, ciprofloxacin, amikacin, trimethoprim-sulfamethoxazole, and streptomycin after drug susceptibility testing.

MYCOBACTERIUM MARINUM

M. marinum is an NTM found in salt water and freshwater, including swimming pools and fish tanks. It is a cause of localized soft-tissue infections, which may require surgical management. Combination regimens include clarithromycin and either ethambutol or rifampin. Other agents with activity against *M. marinum* include doxycycline, minocycline, and trimethoprim-sulfamethoxazole. Drug susceptibility testing is recommended only if the sputum remains culture positive after 3 months of appropriate therapy.

Rapidly Growing Mycobacteria

Rapidly growing mycobacteria causing human disease include *Mycobacterium abscessus, Mycobacterium fortuitum*, and *Mycobacterium chelonae*. Treatment of these mycobacteria is complex and should be undertaken with input from experienced clinicians. It is important to note that testing rapidly growing mycobacteria for macrolide resistance is tricky, as an inducible *erm* gene may confer in vivo macrolide resistance to isolates that are susceptible in vitro.

M. abscessus is the third most common NTM pathogen in the United States. It is endemic in the southeastern states between Texas and Florida. Skin, soft tissue, and bone infections occur, usually after accidental trauma or surgery. This organism appears to have a predilection to cause lung infections in white nonsmoking women aged >60 who have no preexisting lung disease. *M. abscessus* isolates are usually resistant to standard anti-TB regimens. Skin and soft tissue infections are usually treated for a minimum of 4 months with a macrolide (clarithromycin or azithromycin) and a parenteral agent such as amikacin, cefoxitin, or imipenem. Bone infections are treated for at least 6 months. This regimen can be used for the treatment of lung infections but is often unsuccessful because of drug adverse effects and toxicities. A regimen comprising a combination of parenteral drugs is recommended on the basis of in vitro drug susceptibility testing. Surgical resection should be considered in all patients with good lung reserve and a localized infection.

DRUGS FOR THE TREATMENT OF NTM

Clarithromycin

Clarithromycin is a macrolide antibiotic with broad activity against many gram-positive and gram-negative bacteria as well as NTM. This drug is active against MAC organisms and many other NTM species, inhibiting protein synthesis by binding to the 50S mycobacterial ribosomal subunit. NTM resistance to macrolides is probably caused by overexpression of the gene *ermB*, with consequent methylation of the binding site. Clarithromycin is well absorbed orally and distributes well to tissues. It is cleared both hepatically and renally; the dosage should be reduced in renal insufficiency. Clarithromycin is a substrate for and inhibits cytochrome 3A4 and should not be administered with cisapride, pimozide, or terfenadine because cardiac arrhythmias may occur. Numerous drugs interact with clarithromycin through the CYP3A4 metabolic pathway. Rifampin lowers clarithromycin levels; conversely, rifampin levels are increased by clarithromycin. However, the clinical relevance of this interaction does not appear to be great.

For patients with nodular/bronchiectatic MAC infection, the dosage of clarithromycin is 500 mg, given morning and evening three times a week. For the treatment of fibrocavitary or severe nodular/bronchiectatic

MAC infection, a dose of 500–1000 mg is given daily. Disseminated MAC infection is treated with 1000 mg daily. Clarithromycin is used in combination regimens that typically include ethambutol and a rifamycin in order to avoid the development of macrolide resistance. Adverse effects include frequent gastrointestinal intolerance, hepatotoxicity, headache, rash, and rare instances of hypoglycemia. Clarithromycin is contraindicated during pregnancy because of its teratogenicity in animal models.

Azithromycin

Azithromycin is a derivative of erythromycin. Although technically an azalide and not a macrolide, it works similarly to macrolides, inhibiting protein synthesis through binding to the 50S ribosomal subunit. Resistance to azithromycin is almost always associated with complete cross-resistance to clarithromycin. Azithromycin is well absorbed orally, with good tissue penetration and a prolonged half-life (~48 h). The usual dosage for treatment of MAC infection is 250 mg daily or 500 mg three times per week. Azithromycin is used in combination with other agents to avoid the development of resistance. For prophylaxis against disseminated MAC infection in immunocompromised individuals, a dose of 1200 mg once per week is given. Because azithromycin is not metabolized by cytochrome P450, it interacts with few drugs. Adjustment of the dosage on the basis of renal function is not necessary.

Cefoxitin

Cefoxitin is a second-generation parenteral cephalosporin with activity against rapidly growing NTM, particularly *M. abscessus* and *M. chelonae*. Its mechanism of action against NTM is unknown but may involve inactivation of cell-wall synthesis enzymes. High doses are used for treatment of NTM: 200 mg/kg IV three or four times per day, with a maximal daily dose of 12 g. The half-life of cefoxitin is ~1 h, with primarily renal clearance that requires adjustment in renal insufficiency. Adverse effects are uncommon but include gastrointestinal manifestations, rash, eosinophilia, fever, and neutropenia.

Newer Drugs

Three newer class of drugs—the oxazolidinones, the glycylcyclines, and the ketolides—are currently being evaluated for possible use in the treatment of NTM infections, especially those caused by *M. abscessus*. Approximately 50% of *M. abscessus* isolates have shown some degree of susceptibility in vitro to linezolid, an oxazolidinone. Tigecycline, which is a glycylcycline and a tetracycline derivative, and telithromycin, a ketolide, also appear to have in vitro activity against *M. abscessus*. These drugs, however, have not yet been clinically tested in patients.

In addition, some anti-TB drugs, including clofazimine and bedaquiline, are being evaluated as alternative agents for the treatment of refractory NTM infections. In particular, clofazimine appears to act synergistically in combination with amikacin, bedaquiline, or tigecycline. Inhaled amikacin has a positive symptomatic and microbiologic impact, but its toxicity is still a problem. The exact role of these agents in the treatment of refractory NTM infections remains unclear. Suppressive therapy with periodic parenteral/oral drugs to limit disease progression and control symptoms may be an appropriate alternative to curative treatment.

CONCLUSION

Treatment of mycobacterial infections requires multiple-drug regimens that often exert significant side effects with the potential to limit tolerability. The prolonged duration of treatment has vastly improved results over those obtained in past decades, but drugs and regimens that will shorten treatment duration and limit adverse drug effects and interactions are needed.

FURTHER READING

Griffith DE et al: An official ATS/IDSA statement: Diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. Am J Respir Crit Care Med 175:367, 2007.

Nahid P et al: Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America clinical practice guidelines: Treatment of drug-susceptible tuberculosis. Clin Infect Dis 63:e147, 2016. [PubMed: 27516382]

Pontali E et al: Bedaquiline and multidrug-resistant tuberculosis: A systematic and critical analysis of the evidence. Eur Respir J 47:394, 2016.

Chapter 177: Syphilis

Sheila A. Lukehart

DEFINITION

Syphilis, a chronic systemic infection caused by *Treponema pallidum* subspecies *pallidum*, is usually sexually transmitted and is characterized by episodes of active disease interrupted by periods of latency. After an incubation period averaging 2–6 weeks, a primary lesion appears—often associated with regional lymphadenopathy—and then resolves without treatment. The secondary stage, with generalized mucosal and cutaneous lesions and generalized lymphadenopathy, is followed by a latent period of subclinical infection lasting years or decades. Central nervous system (CNS) involvement may occur early in infection and may be symptomatic or asymptomatic. In the preantibiotic era, one-third of untreated patients developed tertiary syphilis, characterized by destructive mucocutaneous, skeletal, or parenchymal lesions; aortitis; or late CNS manifestations.

ETIOLOGY

The Spirochaetales include four genera that are pathogenic for humans and for a variety of other animals: *Leptospira* species (leptospirosis, **Chap. 179)**; *Borrelia* species (relapsing fever and Lyme disease, **Chaps. 180 and 181)**; *Brachyspira* species (gastrointestinal infections); and *Treponema* species (syphilis and the endemic treponematoses; **see also Chap. 178)**. The *Treponema* subspecies include *T. pallidum* subsp. *pallidum* (venereal syphilis); *T. pallidum* subsp. *pertenue* (yaws); *T. pallidum* subsp. *endemicum* (endemic syphilis or bejel); and *T. carateum* (pinta). Until recently, the subspecies were distinguished primarily by the clinical syndromes they produce, but molecular signatures can now differentiate the three *T. pallidum* subspecies when assessed by polymerase chain reaction (PCR) or gene sequencing. The crossing of subspecies boundaries by some gene sequence “signatures” in certain strains demonstrates a genetic “continuum” among strains and subspecies of the pathogenic treponemes. Other *Treponema* species found in the human mouth, genital mucosa, and gastrointestinal tract have been associated with disease (e.g., periodontitis), but their role as primary etiologic agents is unclear.

T. pallidum subspecies are thin spiral organisms, with a cell body surrounded by a trilaminar cytoplasmic membrane, a delicate peptidoglycan layer, and a lipid-rich outer membrane. Endoflagella wind around the cell body in the periplasmic space and are responsible for motility.



The *T. pallidum* subspecies cannot be cultured in vitro. Genome sequencing revealed severely limited metabolic capabilities, including a lack of genes required for de novo synthesis of most amino acids, nucleotides, and lipids. Genes encoding enzymes of the Krebs cycle and oxidative phosphorylation are absent. The organisms contain numerous compensatory genes predicted to encode transporters of amino acids, carbohydrates, and lipids. In addition, genome analyses and other studies have revealed the existence of a 12-member gene family (*tpr*) with similarities to variable outer-membrane antigens of other spirochetes. One member, TprK, has discrete variable regions that undergo antigenic variation during infection, providing a mechanism for immune evasion.

The only known natural host for *T. pallidum* subsp. *pallidum* (referred to hereafter as *T. pallidum*) is the human. *T. pallidum* can infect many mammals, but only humans, higher apes, and a few laboratory animals regularly develop syphilitic lesions. Rabbits are used to propagate virulent strains of *T. pallidum* and serve as the animal model that best reflects human disease and immunopathology.

TRANSMISSION AND EPIDEMIOLOGY

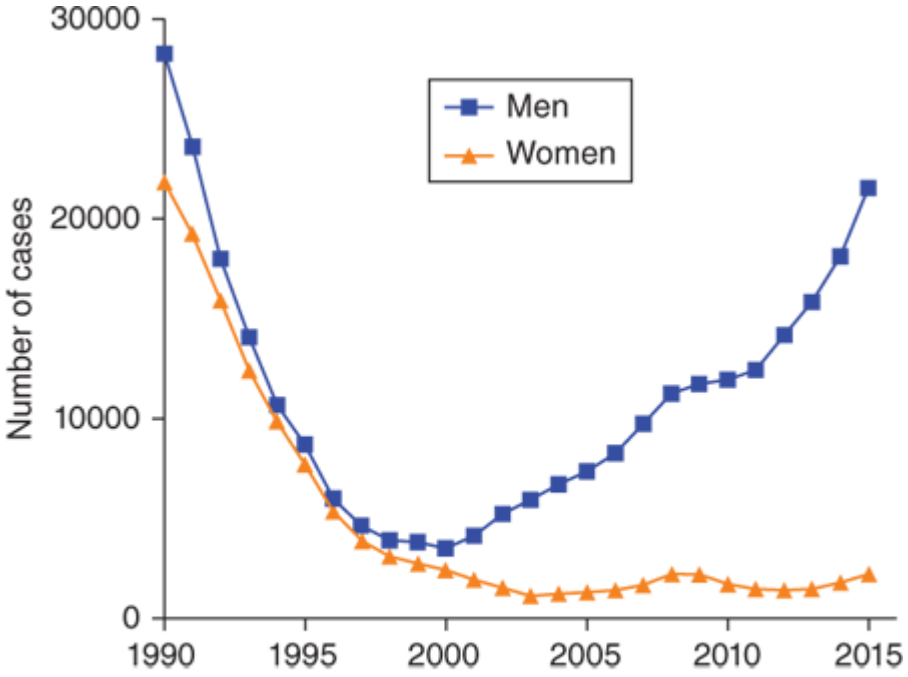
Nearly all cases of syphilis are acquired by sexual contact with infectious lesions (i.e., the chancre, mucous patch, skin rash, or condylomata lata; **see Fig. A1-20**). Less common modes of transmission include nonsexual personal contact, infection in utero, blood transfusion, and organ transplantation.

SYPHILIS IN THE UNITED STATES

With the advent of penicillin therapy in the 1940s, the total number of reported cases of syphilis of all stages in the United States declined 95% from 1943 to a low of 31,575 cases in 2000, with <6000 reported cases of infectious primary and secondary (P&S) syphilis. (P&S cases are a better indicator of disease activity than total syphilis cases.) Since 2000, the number of P&S cases has quadrupled, with 23,872 cases reported in 2015 **(Fig. 177-1)**. Nationally, ~90% of these cases were in men who have sex with men (MSM), ~50% of whom are co-infected with HIV (with exact rates varying by geographic location). From 2014 to 2015, P&S cases also rose among all men (19%) and among women (25%), and increases were seen in all geographic regions in the United States. The incidence of congenital syphilis roughly parallels that of infectious syphilis in women. In 2015, 487 cases in infants <1 year of age were reported, for an increase of 36% in the past 5 years.

FIGURE 177-1

Primary and secondary syphilis in the United States, 1990–2015, by sex. *(Data from the Centers for Disease Control and Prevention.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The populations at highest risk for acquiring syphilis have changed over time, with outbreaks among MSM in the pre-HIV era of the late 1970s and early 1980s as well as at present. It is speculated that recent increases in syphilis and other sexually transmitted infections in MSM may be due to unprotected sex between persons who are HIV concordant and to disinhibition permitted by highly effective antiretroviral therapies. The syphilis epidemic that peaked in 1990, predominantly among African-American heterosexual men and women, occurred largely in urban areas, where infectious syphilis was correlated with the exchange of sex for crack cocaine. Cases of P&S syphilis among African Americans increased 3.5-fold between 2003 and 2015, and the rate (21.4 per 100,000 population) remains higher than rates for other racial/ethnic groups, even though recent increases have been seen in all racial/ethnic groups.

Of individuals named as sexual contacts of persons with infectious syphilis, many have already developed manifestations of syphilis when they are first seen, and ~30% of asymptomatic contacts examined within 30 days of exposure actually have incubating infection and will later develop infectious syphilis if not treated. Thus, identification and treatment of all recently exposed sexual contacts continue to be important aspects of syphilis control.

GLOBAL SYPHILIS

Syphilis remains a significant health problem globally; the number of new infections is estimated at 11 million per year. The regions that are most affected include sub-Saharan Africa, South America, China, and Southeast Asia. During the past decade, the incidence rate for total syphilis in China reached 30 per 100,000, and rates of infectious syphilis have increased dramatically among MSM in many European countries. Worldwide, there are estimated to be 1.4 million cases of syphilis among pregnant women, with 500,000 adverse pregnancy outcomes annually.

NATURAL COURSE AND PATHOGENESIS OF UNTREATED SYPHILIS

T. pallidum rapidly penetrates intact mucous membranes or microscopic abrasions in skin and, within a few hours, enters the lymphatics and blood to produce systemic infection and metastatic foci long before the appearance of a primary lesion. Blood from a patient with incubating or early syphilis is infectious. The generation time of *T. pallidum* during early active disease in vivo is estimated to be ~30 h, and the incubation period of syphilis is inversely proportional to the number of organisms inoculated. The 50% infectious dose for intradermal inoculation in humans has been calculated to be 57 organisms, and the treponeme concentration generally reaches 10^7/g of tissue before a clinical lesion appears. The median incubation period in humans (~21 days) suggests an average inoculum of 500–1000 infectious organisms for naturally acquired disease; the incubation period rarely exceeds 6 weeks.

The primary lesion appears at the site of inoculation, usually persists for 4–6 weeks, and then heals spontaneously. Histopathologic examination shows perivascular infiltration, chiefly by CD4+ and CD8+ T lymphocytes, plasma cells, and macrophages, with capillary endothelial proliferation and subsequent obliteration of small blood vessels. The cellular infiltration produces a T_H1-type cytokine profile, consistent with the activation of macrophages. Phagocytosis of opsonized organisms by activated macrophages ultimately causes their destruction, resulting in spontaneous resolution of the chancre.

The generalized parenchymal, constitutional, mucosal, and cutaneous manifestations of secondary syphilis usually appear ~6–12 weeks after infection, although primary and secondary manifestations may occasionally overlap. In contrast, some patients may enter the latent stage without ever recognizing secondary lesions. The histopathologic features of secondary maculopapular skin lesions include hyperkeratosis of the epidermis, capillary proliferation with endothelial swelling in the superficial dermis, dermal papillae with transmigration of polymorphonuclear leukocytes, and—in the deeper dermis—perivascular infiltration by CD8+ T lymphocytes, CD4+ T lymphocytes, macrophages, and plasma cells. Treponemes are found in many tissues, including the aqueous humor of the eye and the cerebrospinal fluid (CSF). *T. pallidum* disseminates during the first weeks of infection, invading many tissues, including the CNS; CSF abnormalities are detected in as many as 40% of patients during the secondary stage. Clinical hepatitis and immune complex–induced glomerulonephritis are relatively rare but are recognized manifestations of secondary syphilis; however, liver function tests reveal the presence of infection and may yield abnormal results in up to one-quarter of cases of early syphilis. Generalized nontender lymphadenopathy is noted in 85% of patients with secondary syphilis. The paradoxical appearance of secondary manifestations, even after the development of an immune response that clears primary lesions, likely results from immune evasion due to antigenic variation of surface antigens. Secondary lesions generally subside within 2–6 weeks, and the infection enters the latent stage, which is detectable only by serologic testing. In the preantibiotic era, up to 25% of untreated patients experienced at least one cutaneous relapse of secondary lesions, usually during the first year. Therefore, identification and examination of sexual contacts are most important for patients with syphilis of <1 year’s duration.

In the preantibiotic era, about one-third of patients with untreated latent syphilis developed clinically apparent tertiary disease, the most common types being the gumma (a usually benign granulomatous lesion); cardiovascular syphilis (usually involving the vasa vasorum of the ascending aorta and resulting in aneurysm); and late symptomatic neurosyphilis (tabes dorsalis and paresis). In Western countries today, specific treatment for early and latent syphilis and coincidental therapy (i.e., therapy with antibiotics that are given for other conditions but are active against treponemes) have nearly eliminated tertiary syphilis. Asymptomatic CNS involvement, however, is still demonstrable in up to 40% of persons with early syphilis and 25% of patients with late latent syphilis, and modern cases of general paresis and tabes dorsalis are being reported from China. The factors that contribute to the development and progression of tertiary disease are unknown.

CLINICAL MANIFESTATIONS

PRIMARY SYPHILIS

The typical primary chancre usually begins as a single painless papule that rapidly becomes eroded and usually becomes indurated, with a characteristic cartilaginous consistency on palpation of the edge and base of the ulcer. Multiple primary lesions are seen in a minority of patients. In heterosexual men the chancre is usually located on the penis, where it is readily seen **(Fig. 177-2; see also Fig. A1-17)**, whereas in MSM it may also be found in the anal canal or rectum or in the mouth. Oral sex has been identified as the source of infection in some MSM. In women, common primary sites are the cervix and labia. Consequently, primary syphilis goes unrecognized in women and MSM more often than in heterosexual men.

FIGURE 177-2

Primary syphilis with a firm, nontender chancre.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Atypical primary lesions are common. A large inoculum produces a dark-field-positive ulcerative lesion in nonimmune volunteers but may produce a small dark-field-negative papule, an asymptomatic but seropositive latent infection, or no response at all in some individuals with a history of syphilis. A small inoculum may produce only a papular lesion, even in nonimmune individuals. Therefore, syphilis should be considered even in the evaluation of trivial or atypical dark-field-negative genital lesions. The lesions that most commonly must be differentiated from those of primary syphilis include those caused by herpes simplex virus infection **(Chap. 187)**, chancroid **(Chap. 152)**, traumatic injury, and donovanosis **(Chap. 168)**. Regional (usually inguinal) lymphadenopathy accompanies the primary syphilitic lesion, appearing within 1 week of lesion onset. The nodes are firm, nonsuppurative, and painless. Inguinal lymphadenopathy is bilateral and may occur with anal as well as with genital chancres. The chancre generally heals within 4–6 weeks (range, 2–12 weeks), but lymphadenopathy may persist for months.

SECONDARY SYPHILIS

The protean manifestations of the secondary stage usually include mucocutaneous or cutaneous lesions and generalized nontender lymphadenopathy. The healing primary chancre may still be present in ~15% of cases—more frequently in persons with concurrent HIV infection. The skin rash consists of macular, papular, papulosquamous, and occasionally pustular syphilides; often more than one form is present simultaneously. The eruption may be very subtle, and 25% of patients with a discernible rash may be unaware that they have dermatologic manifestations. Initial lesions are pale red or pink, nonpruritic, discrete macules distributed on the trunk and extremities; these macules progress to papular lesions that are distributed widely and that frequently involve the palms and soles (**Fig. 177-3; see also Figs. A1-18 and A1-19**). Rarely, severe necrotic lesions (*lues maligna*) may appear; they are more commonly reported in HIV-infected individuals. Involvement of the hair follicles may result in patchy alopecia of the scalp hair, eyebrows, or beard in up to 5% of cases.

FIGURE 177-3

Secondary syphilis. ***Left:*** Maculopapular truncal eruption. ***Middle:*** Papules on the palms. ***Right:*** Papules on the soles. *(Courtesy of Jill McKenzie and Christina Marra.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In warm, moist, intertriginous areas (commonly the perianal region, vulva, and scrotum), papules can enlarge to produce broad, moist, pink or gray-white, highly infectious lesions (*condylomata lata*; **see Fig. A1-20**) in 10% of patients with secondary syphilis. Superficial mucosal erosions (*mucous patches*) occur in 10–15% of patients and commonly involve the oral or genital mucosa **(see Fig. A1-21)**. The typical mucous patch is a painless silver-gray erosion surrounded by a red periphery.

Constitutional signs and symptoms that may accompany or precede secondary syphilis include sore throat (15–30%), fever (5–8%), weight loss (2–20%), malaise (25%), anorexia (2–10%), headache (10%), and meningismus (5%). *Acute meningitis* occurs in only 1–2% of cases, but CSF cell and protein concentrations are increased in up to 40% of early syphilis cases, and viable *T. pallidum* organisms have been recovered from CSF during primary and secondary syphilis in 30% of cases; the latter finding is often but not always

associated with other CSF abnormalities. Ocular findings associated with secondary (or later/unknown-stage) syphilis include pupillary abnormalities and optic neuritis as well as the classic iritis or uveitis. The diagnosis of ocular syphilis is often considered in affected patients only after they fail to respond to topical steroid therapy. Anterior uveitis has been reported in 5–10% of patients with secondary syphilis, and *T. pallidum* has been demonstrated in aqueous humor from such patients. Permanent blindness may result without prompt diagnosis and treatment. The recent publication of a number of reports of ocular syphilis reminds clinicians to inquire about neurologic manifestations in all stages of syphilis infection. In a recent retrospective study, 7.9% of patients with syphilis, when asked, reported recent vision or hearing changes, and more than half of those reporting these changes had abnormal CSF or ophthalmologic findings consistent with syphilis.

Less common complications of secondary syphilis include hepatitis, nephropathy, gastrointestinal involvement (hypertrophic gastritis, patchy proctitis, or a rectosigmoid mass), arthritis, and periostitis. Hepatic involvement is common in syphilis; although it is usually asymptomatic, up to 25% of patients may have abnormal liver function tests. Frank syphilitic hepatitis may be seen. Renal involvement usually results from immune complex deposition and produces proteinuria associated with an acute nephrotic syndrome. Like those of primary syphilis, most manifestations of the secondary stage resolve spontaneously, usually within 1–6 months.

LATENT SYPHILIS

Positive serologic tests for syphilis, together with a normal CSF examination and the absence of clinical manifestations of syphilis, indicate a diagnosis of latent syphilis in an untreated person. The diagnosis is often suspected on the basis of a history of primary or secondary lesions, a history of exposure to syphilis, or the delivery of an infant with congenital syphilis. A previous nonreactive serologic test or a history of lesions or exposure may help establish the duration of latent infection, which is an important factor in the selection of appropriate therapy. *Early latent* syphilis is limited to the first year after infection, whereas *late latent* syphilis is defined as that of ≥1 year's duration (or of unknown duration). *T. pallidum* may still seed the bloodstream intermittently during the latent stage, and latent syphilis in a pregnant woman may infect the fetus in utero. Moreover, syphilis has been transmitted through blood transfusion or organ donation from patients with latent syphilis. It was previously thought that untreated late latent syphilis had three possible outcomes: (1) persistent lifelong infection; (2) development of late syphilis; or (3) spontaneous cure, with reversion of serologic tests to negative. It is now apparent, however, that the more sensitive treponemal antibody tests rarely, if ever, become nonreactive without treatment. Although progression to clinically evident late syphilis is very rare today, the occurrence of spontaneous microbiologic cure is in doubt.

INVOLVEMENT OF THE CNS

Traditionally, neurosyphilis has been considered a late manifestation of syphilis, but this view is inaccurate. CNS syphilis represents a continuum encompassing early invasion (usually within the first weeks of

infection), months to years of asymptomatic involvement, and, in some cases, development of early or late neurologic manifestations.

Asymptomatic Neurosyphilis

The diagnosis of asymptomatic neurosyphilis is made in patients who lack neurologic symptoms and signs but who have CSF abnormalities, including mononuclear pleocytosis, increased protein concentrations, or reactivity in the CSF Venereal Disease Research Laboratory (VDRL) test. CSF abnormalities are demonstrated in up to 40% of cases of untreated primary or secondary syphilis and in 25% of cases of untreated latent syphilis. *T. pallidum* has been recovered by inoculation into rabbits of CSF from up to 30% of patients with primary or secondary syphilis but less frequently from patients with latent syphilis. The presence of *T. pallidum* in CSF is often associated with other CSF abnormalities, but organisms can be recovered from patients with otherwise normal CSF. Although the prognostic implications of these findings in early syphilis are uncertain, it may be appropriate to conclude that even patients with early syphilis who have such findings do indeed have asymptomatic neurosyphilis and should be treated for neurosyphilis; such treatment is particularly important in patients with concurrent HIV infection. Before the advent of penicillin, the risk of development of clinical neurosyphilis in untreated asymptomatic persons was roughly proportional to the intensity of CSF changes, with the overall cumulative probability of progression to clinical neurosyphilis ~20% in the first 10 years of infection but increasing with time. Most experts agree that neurosyphilis is more common among HIV-infected persons, while immunocompetent patients with untreated latent syphilis and normal CSF probably run a very low risk of subsequent neurosyphilis. In several large studies, neurosyphilis was associated with a rapid plasma reagin (RPR) titer of ≥1:32, regardless of clinical stage or HIV infection status.

Symptomatic Neurosyphilis

The major clinical categories of symptomatic neurosyphilis include meningeal, meningovascular, and parenchymatous syphilis. The last category includes general paresis and tabes dorsalis. The onset of symptoms usually occurs <1 year after infection for meningeal syphilis, up to 10 years after infection for meningovascular syphilis, at ~20 years for general paresis, and at 25–30 years for tabes dorsalis. Neurosyphilis is more frequently symptomatic in patients co-infected with HIV, particularly those with low CD4+ T lymphocyte counts. In addition, evidence suggests that syphilis infection worsens the cognitive impairment seen in HIV-infected persons and that this effect persists after treatment for syphilis.

Meningeal syphilis may present as headache, nausea, vomiting, neck stiffness, cranial nerve involvement, seizures, and changes in mental status. This condition may be concurrent with or may follow the secondary stage. Patients presenting with uveitis, iritis, or hearing loss often have meningeal syphilis, but these clinical findings can also be seen in patients with normal CSF.

Meningovascular syphilis reflects meningitis together with inflammatory vasculitis of small, medium, or large vessels. The most common presentation is a stroke syndrome involving the middle cerebral artery of a relatively young adult. However, unlike the usual thrombotic or embolic stroke syndrome of sudden onset,

meningovascular syphilis often becomes manifest after a subacute encephalitic prodrome (with headaches, vertigo, insomnia, and psychological abnormalities), which is followed by a gradually progressive vascular syndrome.

The manifestations of *general paresis* reflect widespread late parenchymal damage and include abnormalities corresponding to the mnemonic *paresis: p*ersonality, *a*ffect, *r*eflexes (hyperactive), *e*ye (e.g., Argyll Robertson pupils), *s*ensorium (illusions, delusions, hallucinations), *i*ntellect (a decrease in recent memory and in the capacity for orientation, calculations, judgment, and insight), and *s*peech. *Tabes dorsalis* is a late manifestation of syphilis that presents as symptoms and signs of demyelination of the posterior columns, dorsal roots, and dorsal root ganglia, including ataxia, foot drop, paresthesia, bladder disturbances, impotence, areflexia, and loss of positional, deep-pain, and temperature sensations. The small, irregular Argyll Robertson pupil, a feature of both tabes dorsalis and paresis, reacts to accommodation but not to light. *Optic atrophy* also occurs frequently in association with tabes.

OTHER MANIFESTATIONS OF LATE SYPHILIS

The slowly progressive inflammatory process leading to tertiary disease begins early during infection, although these manifestations may not become clinically apparent for years or decades. Early syphilitic aortitis first becomes evident soon after secondary lesions subside, and treponemes that trigger the development of gummas may have seeded the tissue years earlier.

Cardiovascular Syphilis

Cardiovascular manifestations, usually appearing 10–40 years after infection, are attributable to endarteritis obliterans of the vasa vasorum, which provide the blood supply to large vessels; *T. pallidum* DNA has been detected by PCR in aortic tissue. Cardiovascular involvement results in uncomplicated aortitis, aortic regurgitation, saccular aneurysm (usually of the ascending aorta), or coronary ostial stenosis. In the preantibiotic era, symptomatic cardiovascular complications developed in ~10% of persons with untreated late syphilis. Today, cardiovascular syphilis is rarely seen in the developed world.

Late Benign Syphilis (Gumma)

Gummas are usually solitary lesions ranging from microscopic to several centimeters in diameter. Histologic examination shows a granulomatous inflammation, with a central area of necrosis due to endarteritis obliterans. *T. pallidum* has been detected by PCR in these lesions, and penicillin treatment results in rapid resolution, confirming the treponemal stimulus for the inflammation. Common sites include the skin and skeletal system; however, any organ (including the brain) may be involved. Gummas of the skin produce indolent, painless, indurated nodular or ulcerative lesions that may resemble other chronic granulomatous conditions. Skeletal gummas may affect any bone or cartilage. Upper respiratory gummas can lead to perforation of the nasal septum or palate.

CONGENITAL SYPHILIS

Transmission of *T. pallidum* across the placenta from a syphilitic woman to her fetus may occur at any stage of pregnancy, but fetal damage generally does not occur until after the fourth month of gestation when fetal immunologic competence begins to develop. This timing suggests that the pathogenesis of congenital syphilis, like that of adult syphilis, depends on the host immune response rather than on a direct toxic effect of *T. pallidum*. The risk of fetal infection during untreated early maternal syphilis is ~75–95%, decreasing to ~35% for maternal syphilis of >2 years' duration. Adequate treatment of the woman before the 16th week of pregnancy should prevent fetal damage, and treatment before the third trimester should adequately treat the infected fetus. Untreated maternal infection may result in a rate of fetal loss of up to 40% with stillbirth (more common than abortion because of the late onset of fetal pathology), prematurity, neonatal death, or nonfatal congenital syphilis. Among infants born alive, only fulminant congenital syphilis is clinically apparent at birth, and these babies have a very poor prognosis. The most common clinical problem is the healthy-appearing baby born to a mother with a positive serologic test.

 Routine serologic testing for syphilis in early pregnancy is cost-effective in virtually all populations, even in areas with a low prenatal prevalence of syphilis. Low-tech point-of-care tests have been developed and widely implemented to facilitate antenatal testing in resource-poor settings. Globally, the past 5–8 years have seen a 38% reduction in maternal syphilis and a 39% reduction in congenital syphilis. Progress has been uneven, however, with major advances in Thailand, Cuba, several Baltic States, and India, but continuing high levels in Africa and China. Periodic lack of penicillin availability prevents treatment of seropositive women. Integration of programs to prevent congenital syphilis with programs to prevent maternal transmission of HIV would be highly cost-effective but is hampered by the restrictions placed on HIV-focused funds. Pregnant women should be screened at their first antenatal visit. Where the prevalence of syphilis is high or when the patient is at high risk of reinfection, serologic testing should be repeated in the third trimester and at delivery. Neonatal congenital syphilis must be differentiated from other generalized congenital infections, including rubella, cytomegalovirus or herpes simplex virus infection, and toxoplasmosis, as well as from erythroblastosis fetalis.

Manifestations of congenital syphilis may appear early (within the first 2 years of life, often at 2–10 weeks of age) or late (after 2 years). The earliest manifestations of congenital syphilis include rhinitis, or "snuffles" (23%); mucocutaneous lesions (35–41%); bone changes (61%), including periostitis detectable by x-ray examination of long bones; hepatosplenomegaly (50%); lymphadenopathy (32%); anemia (34%); jaundice (30%); thrombocytopenia; and leukocytosis. CNS invasion by *T. pallidum* is detectable in 22% of infected neonates. Neonatal death is usually due to pulmonary hemorrhage, secondary bacterial infection, or severe hepatitis. Late congenital syphilis (untreated after 2 years of age) is subclinical in 60% of cases; the clinical spectrum in the remainder of cases may include interstitial keratitis (which occurs at 5–25 years of age), eighth-nerve deafness, and recurrent arthropathy. Neurosyphilis was documented in about one-quarter of untreated patients with late congenital syphilis in the preantibiotic era. Gummatous periostitis occurs at 5–20 years of age and, as in nonvenereal endemic syphilis, tends to cause destructive lesions of the palate and nasal septum. Classic stigmata include *Hutchinson's teeth* (centrally notched, widely spaced, peg-shaped upper central incisors), "mulberry" molars (sixth-year molars with multiple, poorly developed cusps), saddle nose, and saber shins.

LABORATORY EXAMINATIONS

DEMONSTRATION OF THE ORGANISM

T. pallidum cannot be detected by culture. Historically, dark-field microscopy and immunofluorescence antibody staining have been used to identify this spirochete in samples from moist lesions such as chancres or condylomata lata, but these tests are rarely available outside of research laboratories. Sensitive and specific PCR tests have been developed but are not commercially available, although a number of laboratories perform in-house validated PCR testing.

T. pallidum can be found in tissue with appropriate silver stains, but these results should be interpreted with caution because artifacts resembling *T. pallidum* are often seen. Tissue treponemes can be demonstrated more reliably in research laboratories by PCR or by immunofluorescence or immunohistochemical methods using specific monoclonal or polyclonal antibodies to *T. pallidum*. *T. pallidum* DNA has been detected by PCR in lesion swabs, tissue samples, blood, CSF, ocular fluid, urine, and oropharyngeal swabs.

SEROLOGIC TESTS FOR SYPHILIS

Treponemal and Lipoidal Tests

There are two types of serologic tests for syphilis: lipoidal (so-called nontreponemal) and treponemal. Both are reactive in persons with any treponemal infection, including syphilis, yaws, pinta, and endemic syphilis.

The most widely used lipoidal antibody tests for syphilis are the RPR and VDRL tests, which measure IgG and IgM directed against a cardiolipin-lecithin-cholesterol antigen complex. The RPR test is easier to perform and uses unheated serum or plasma; it is the test of choice for rapid serologic diagnosis in a clinical setting. The VDRL test remains the standard for examining CSF and is superior to the RPR for this purpose. The RPR and VDRL tests are recommended for screening and for quantitation of serum antibody. The titer reflects disease activity, rising during the evolution of early syphilis, often exceeding 1:32 in secondary syphilis, and declining slowly thereafter without therapy. After treatment for early syphilis, a persistent fall by fourfold or more (e.g., a decline from 1:32 to 1:8) is considered an adequate response. VDRL titers do not correspond directly to RPR titers, and sequential quantitative testing (as for response to therapy) must employ a single test.

Treponemal tests measure antibodies to native or recombinant *T. pallidum* antigens and include the fluorescent treponemal antibody–absorbed (FTA-ABS) test and the *T. pallidum* particle agglutination (TPPA) test, both of which are more sensitive for primary syphilis than the lipoidal tests. When used to confirm reactive lipoidal test results, treponemal tests have a very high positive predictive value for diagnosis of syphilis.

Treponemal enzyme or chemiluminescence immunoassays (EIAs/CIAs), based largely on reactivity to recombinant antigens, are now widely used as screening tests by large laboratories. When used for screening, however, standard treponemal tests give false-positive results at rates as high as 1–2%, but the

rate is much higher with the EIA/CIA tests. A high proportion of sera that are reactive by EIA/CIA are nonreactive by lipoidal tests. Such sera should be examined in the TPPA test, which includes different antigens and a different platform. If the TPPA test is nonreactive, the patient is unlikely to have syphilis; if it is reactive, the patient is likely to have current or past syphilis. The rapid immunochromatographic tests described for antenatal screening in resource-poor settings are largely unavailable in the United States.

Both lipoidal and treponemal tests may be nonreactive in early primary syphilis, although treponemal tests are slightly more sensitive (85–90%) during this stage than lipoidal tests (~80%). All tests are reactive during secondary syphilis. (Fewer than 1% of patients with high titers have a lipoidal test that is nonreactive or weakly reactive with undiluted serum but is reactive with diluted serum—the *prozone phenomenon*.) VDRL and RPR sensitivity and titers may decline in untreated persons with late latent syphilis, but treponemal tests remain reactive in late syphilis. After treatment for early syphilis, lipoidal test titers will generally decline or the tests will become nonreactive, whereas treponemal tests often remain reactive after therapy and are not helpful in determining the infection status of persons with past syphilis.

Clinicians need to be familiar with three uses of serologic tests for syphilis recommended by the Centers for Disease Control and Prevention (CDC): (1) screening or diagnosis (RPR or VDRL), (2) quantitative measurement of antibody to assess clinical syphilis activity or to monitor response to therapy (RPR or VDRL), and (3) confirmation of a syphilis diagnosis in a patient with a reactive lipoidal test (FTA-ABS, TPPA, EIA/CIA). Whereas IgM titers appear to decline after therapy, the presence or absence of specific IgM does not strictly correlate with active *T. pallidum* infection. Moreover, no commercially available IgM test is recommended, even for evaluation of infants with suspected congenital syphilis.

False-Positive Serologic Tests for Syphilis

The lipid antigens of nontreponemal tests are similar to those found in human tissues, and the tests may be reactive (usually with titers ≤1:8) in persons without treponemal infection. Among patients being screened for syphilis because of risk factors, clinical suspicion, or history of exposure, ~1% of reactive tests are falsely positive. Modern VDRL and RPR tests are highly specific, and false-positive reactions are largely limited to persons with autoimmune conditions or injection drug use. The prevalence of false-positive results increases with advancing age. In a patient with a false-positive nontreponemal test, syphilis is excluded by a nonreactive treponemal test.

False-positive reactions may also occur with treponemal tests, particularly the very sensitive EIA/CIA tests. Screening a low-prevalence population for syphilis with a treponemal test may result in true-positive reactions' being outnumbered by false-positive reactions, leading to unnecessary treatment. Thus screening with lipoidal tests is highly recommended.

EVALUATION FOR NEUROSYPHILIS

Involvement of the CNS is detected by examination of CSF for mononuclear pleocytosis (>5 white blood cells/μL), increased protein concentration (>45 mg/dL), or CSF VDRL reactivity. Elevated CSF cell counts and

protein concentrations are not specific for neurosyphilis and may be confounded by HIV co-infection. Because CSF pleocytosis may also be due to HIV, some studies have suggested using a CSF white-cell cutoff of 20 cells/µL as diagnostic of neurosyphilis in HIV-infected patients with syphilis. The CSF VDRL test is highly specific and, when reactive, is considered diagnostic of neurosyphilis; however, this test is insensitive and may be nonreactive even in cases of symptomatic neurosyphilis. The RPR test should not be substituted for the VDRL test for CSF examination. The FTA-ABS test on CSF is reactive far more often than the CSF VDRL test in all stages of syphilis, but reactivity may reflect passive transfer of serum antibody into the CSF. A nonreactive FTA-ABS test on CSF, however, may be used to rule out asymptomatic neurosyphilis. Measuring CXCL13 in CSF has been demonstrated to distinguish between neurosyphilis and HIV-related CSF abnormalities.

All *T. pallidum*–infected patients who have signs or symptoms consistent with neurologic disease (e.g., meningitis, hearing loss) or ophthalmic disease (e.g., uveitis, iritis) should have a CSF examination, regardless of disease stage. The appropriate management of asymptomatic persons is less clear. Lumbar puncture on all asymptomatic patients with untreated syphilis is impractical and unnecessary. Because therapy with penicillin G benzathine fails to result in treponemicidal drug levels in CSF, however, it is important to identify those persons at higher risk for having or developing neurosyphilis so that appropriate therapy may be given. Viable *T. pallidum* has been isolated from the CSF of several patients (with and without HIV infection) after penicillin G benzathine therapy for early syphilis. Large-scale prospective studies have provided evidence-based guidelines for determining which syphilis patients may benefit most from CSF examination. Specifically, patients with RPR titers of ≥1:32 are at higher risk of having neurosyphilis (11-fold and 6-fold higher in HIV-infected and HIV-uninfected persons, respectively), as are HIV-infected patients with CD4+ T cell counts of ≤350/µL. Persons with active tertiary syphilis and those in whom treatment failure is suspected also should have their CSF examined.

EVALUATION OF HIV-INFECTED PATIENTS FOR SYPHILIS

Because persons at highest risk for syphilis are also at increased risk for HIV infection, these two infections frequently coexist. There is evidence that syphilis and other genital ulcer diseases are important risk factors for acquisition and transmission of HIV infection. Some manifestations of syphilis may be altered in patients with concurrent HIV infection, and multiple cases of neurologic relapse after standard therapy have been reported in these patients.

Persons with newly diagnosed HIV infection should be tested for syphilis; conversely, all patients with newly diagnosed syphilis should be tested for HIV infection. Some authorities, persuaded by reports of persistent *T. pallidum* in CSF of HIV-infected persons after standard therapy for early syphilis, recommend CSF examination for evidence of neurosyphilis for all co-infected patients, regardless of the stage of syphilis, with treatment for neurosyphilis if CSF abnormalities are found. Others, on the basis of their own clinical experience, think that standard therapy—without CSF examination—is sufficient for all cases of early syphilis in HIV-infected patients without neurologic signs or symptoms. As described above, RPR titer and CD4+ T cell count can be used to identify patients at higher risk of neurosyphilis for lumbar puncture, although some

cases of neurosyphilis will be missed even when these criteria are used. Serologic testing after treatment is important for all patients with syphilis, particularly for those also infected with HIV.

TREATMENT

TREATMENT

Syphilis

TREATMENT OF ACQUIRED SYPHILIS

 The CDC's 2015 guidelines for the treatment of syphilis are summarized in **Table 177-1** and are discussed below. Penicillin G is the drug of choice for all stages of syphilis. *T. pallidum* is killed by very low concentrations of penicillin G, although a long period of exposure to penicillin is required because of the unusually slow rate of multiplication of the organism. The efficacy of penicillin against syphilis remains undiminished after 70 years of use, and there is no evidence of penicillin resistance in *T. pallidum*. Other antibiotics effective in syphilis include the tetracyclines and the cephalosporins. Aminoglycosides and spectinomycin inhibit *T. pallidum* only in very large doses, and the sulfonamides and most quinolones are inactive. Azithromycin has shown significant promise as an effective oral agent against *T. pallidum*; however, strains harboring 23S rDNA mutations that confer macrolide resistance are widespread. Such strains represent >80–90% of recent isolates from large U.S., European, and Chinese cities, while rates of 23S mutation are much lower in some other locations. The prevalence of resistant strains varies by geographic location, and routine treatment of syphilis with azithromycin is not recommended. Careful follow-up of any patient treated for syphilis with azithromycin must be assured.

Early Syphilis Patients and Their Contacts

Penicillin G benzathine is the most widely used agent for the treatment of early syphilis; preventive treatment is also recommended for individuals who have been exposed to infectious syphilis within the previous 3 months. *The regimens recommended for prevention are the same as those recommended for early syphilis*. Penicillin G benzathine cures >95% of cases of early syphilis, although clinical relapse can follow treatment, particularly in patients with concurrent HIV infection. Because the risk of neurologic relapse may be higher in HIV-infected patients, CSF examination is recommended for HIV-seropositive individuals with syphilis of any stage, particularly those with a serum RPR titer of ≥1:32 or a CD4+ T cell count of ≤350/μL. Therapy appropriate for neurosyphilis should be given if there is any evidence of CNS infection.

Late Latent Syphilis or Syphilis of Unknown Duration

If the CSF is normal or is not examined, the recommended treatment is penicillin G benzathine (7.2 million units total; Table 177-1). If CSF abnormalities are found, the patient should be treated for neurosyphilis.

Tertiary Syphilis

CSF examination should be performed. If the CSF is normal, the recommended treatment is penicillin G benzathine (7.2 million units total; Table 177-1). If CSF is abnormal, the patient should be treated for neurosyphilis. The clinical response to treatment for benign tertiary syphilis is usually impressive, but

responses in cardiovascular syphilis are not dramatic because aortic aneurysm and aortic regurgitation cannot be reversed by antibiotics.

Syphilis in Penicillin-Allergic Patients

For penicillin-allergic patients with syphilis, a 2-week (early syphilis) or 4-week (late or late latent syphilis) course of therapy with doxycycline or tetracycline is recommended (Table 177-1). These regimens appear to be effective in early syphilis but have not been tested for late or late latent syphilis, and compliance may be problematic. Limited studies suggest that ceftriaxone (1 g/d, given IM or IV for 8–10 days) is effective for early syphilis. These nonpenicillin regimens have not been carefully evaluated in HIV-infected individuals and should be used with caution. If compliance and follow-up are not assured, penicillin-allergic HIV-infected persons with late latent or late syphilis should be desensitized and treated with penicillin.

Neurosyphilis

Penicillin G benzathine, even at high doses, does not produce treponemicidal concentrations of penicillin G in CSF and should not be used for treatment of neurosyphilis. Asymptomatic neurosyphilis may relapse as symptomatic disease after treatment with benzathine penicillin, and the risk of relapse may be higher in HIV-infected patients. Both symptomatic and asymptomatic neurosyphilis should be treated with aqueous penicillin (Table 177-1). Administration either of IV aqueous crystalline penicillin G or of IM aqueous procaine penicillin G plus oral probenecid in recommended doses is thought to ensure treponemicidal concentrations of penicillin G in CSF. The clinical response to penicillin therapy for meningeal syphilis is dramatic, but treatment of neurosyphilis with existing parenchymal damage may only arrest disease progression. No data suggest that additional therapy (e.g., penicillin G benzathine for 3 weeks) is beneficial after treatment for neurosyphilis.

The use of antibiotics other than penicillin G for the treatment of neurosyphilis has not been studied, although limited data suggest that ceftriaxone may be used. In patients with confirmed penicillin allergy, desensitization and treatment with penicillin are recommended.

Management of Syphilis in Pregnancy

Every pregnant woman should undergo a lipoidal screening test at her first prenatal visit and, if at high risk of exposure, again in the third trimester and at delivery. In the untreated pregnant patient with presumed syphilis, expeditious treatment appropriate to the stage of the disease is essential. Patients should be warned of the risk of a Jarisch-Herxheimer reaction, which may be associated with mild premature contractions but rarely results in premature delivery.

Penicillin is the only recommended agent for the treatment of syphilis in pregnancy. If the patient has a documented penicillin allergy, desensitization and penicillin therapy should be undertaken according to the CDC's 2015 guidelines. After treatment, a quantitative nontreponemal test should be repeated monthly throughout pregnancy to assess therapeutic efficacy. Treated women whose antibody titers rise by fourfold or whose titers do not decrease by fourfold over a 3-month period should be re-treated.

EVALUATION AND MANAGEMENT OF CONGENITAL SYPHILIS

Whether or not they are infected, newborn infants of women with reactive serologic tests may themselves have reactive tests because of transplacental transfer of maternal IgG antibodies. For asymptomatic infants born to women treated adequately with penicillin during the first or second trimester of pregnancy, monthly quantitative nontreponemal tests may be performed to monitor for appropriate reduction in antibody titers. Rising or persistent titers indicate infection, and the infant should be treated. Detection of neonatal IgM antibody may be useful, but no commercially available test is currently recommended.

An infant should be treated at birth if the treatment status of the seropositive mother is unknown; if the mother received inadequate or nonpenicillin therapy; if the mother received penicillin therapy in the third trimester; or if the infant may be difficult to follow. The CSF should be examined to obtain baseline values before treatment. Penicillin is the only recommended drug for the treatment of syphilis in infants. Specific recommendations for the treatment of infants and older children are included in the CDC's 2015 treatment guidelines.

JARISCH-HERXHEIMER REACTION

A dramatic although self-limited reaction consisting of fever, chills, myalgia, headache, tachycardia, increased respiratory rate, increased circulating neutrophil count, and vasodilation with mild hypotension may follow the initiation of treatment for syphilis. This reaction is thought to be a response to lipoproteins released by dying *T. pallidum* organisms. The Jarisch-Herxheimer reaction occurs in ~50% of patients with primary syphilis, 90% of those with secondary syphilis, and a lower proportion of persons with later-stage disease. Defervescence takes place within 12–24 h. In secondary syphilis, erythema and edema of the cutaneous lesions may increase. Patients should be warned to expect such developments, which can be managed with symptom-based treatment. Steroid therapy is not required for this mild transient reaction.

FOLLOW-UP EVALUATION OF RESPONSES TO THERAPY

Efficacy of treatment should be assessed by clinical evaluation and monitoring of the quantitative VDRL or RPR titer for a fourfold decline (e.g., from 1:32 to 1:8). Patients with primary or secondary syphilis should be examined 6 and 12 months after treatment, and persons with latent or late syphilis at 6, 12, and 24 months. More frequent clinical and serologic examination (3, 6, 9, 12, and 24 months) is recommended for patients concurrently infected with HIV, regardless of the stage of syphilis.

After successful treatment of seropositive first-episode primary or secondary syphilis, the VDRL or RPR titer progressively declines; the test becomes nonreactive by 12 months in 40–75% of seropositive primary cases and in 20–40% of secondary cases. In patients with HIV infection or a history of prior syphilis, VDRL and RPR tests are less likely to become nonreactive. Rates of decline of serologic titers appear to be slower, and serologically defined treatment failures more common, among HIV-infected patients than among those without HIV co-infection; however, effective antiretroviral therapy may reduce these differences. Re-treatment should be considered if serologic responses are not adequate or if clinical signs persist or recur. Because it is difficult to differentiate treatment failure from reinfection, the CSF should be examined, with treatment for neurosyphilis if CSF is abnormal and treatment for late latent syphilis if CSF is normal. A minority of patients treated for early syphilis may experience a one-dilution titer increase within 14 days after treatment; however, this early elevation does not significantly affect the serologic outcome at 6 months after

treatment. Patients treated for late latent syphilis frequently have low initial VDRL or RPR titers and may not have a fourfold decline after therapy with penicillin. In such patients, re-treatment is not warranted unless the titer rises or signs and symptoms of syphilis appear. Because treponemal tests may remain reactive despite treatment for seropositive syphilis, these tests are not useful in following the response to therapy.

The activity of neurosyphilis (symptomatic or asymptomatic) correlates best with CSF pleocytosis, and this measure provides the most sensitive index of response to treatment. Repeat CSF examinations should be performed every 6 months until the cell count is normal. An elevated CSF cell count falls to normal in 3–12 months in adequately treated HIV-uninfected patients. The persistence of mild pleocytosis in HIV-infected patients may be due to the presence of HIV in CSF; this scenario may be difficult to distinguish from treatment failure. Elevated levels of CSF protein fall more slowly, and the CSF VDRL titer declines gradually over several years. In patients treated for neurosyphilis, a fourfold reduction in serum RPR titer has been positively correlated with normalization of CSF abnormalities; this correlation is stronger in HIV-uninfected patients and in HIV-infected patients receiving effective antiretroviral therapy.

TABLE 177-1

Recommendations for the Treatment of Syphilis[a]

Stage of Syphilis	Patients without Penicillin Allergy	Patients with Confirmed Penicillin Allergy
Primary, secondary, or early latent	*CSF normal or not examined:* Penicillin G benzathine (single dose of 2.4 mU IM) *CSF abnormal:* Treat as neurosyphilis.	*CSF normal or not examined:* Tetracycline HCl (500 mg PO qid) or doxycycline (100 mg PO bid) for 2 weeks *CSF abnormal:* Treat as neurosyphilis.
Late latent (or latent of unknown duration), cardiovascular, or benign tertiary	*CSF normal or not examined:* Penicillin G benzathine (2.4 mU IM weekly for 3 weeks) *CSF abnormal:* Treat as neurosyphilis.	*CSF normal and patient not infected with HIV:* Tetracycline HCl (500 mg PO qid) or doxycycline (100 mg PO bid) for 4 weeks *CSF normal and patient infected with HIV:* Desensitize and treat with penicillin if compliance cannot be assured. *CSF abnormal:* Treat as neurosyphilis.
Neurosyphilis (asymptomatic or symptomatic)	Aqueous crystalline penicillin G (18–24 mU/d IV, given as 3–4 mU q4h or continuous infusion) for 10–14 days *or* Aqueous procaine penicillin G (2.4 mU/d IM) plus oral probenecid (500 mg qid), both for 10–14 days	Desensitize and treat with penicillin.
Syphilis in pregnancy	According to stage	Desensitize and treat with penicillin.

[a]See text for indications for CSF examination.

Abbreviations: CSF, cerebrospinal fluid; mU, million units.

Source: Adapted from the 2015 Sexually Transmitted Diseases Treatment Guidelines from the Centers for Disease Control and Prevention.

IMMUNITY TO SYPHILIS

The rate of development of acquired resistance to *T. pallidum* after natural or experimental infection depends on both the size of the infecting inoculum and the duration of infection before treatment. Both humoral and cellular responses are considered to be of major importance in the healing of early lesions. Cellular infiltration, predominantly by T lymphocytes and macrophages, produces an interferon γ–dominated cytokine milieu and results in the clearance of organisms by activated macrophages. Specific antibodies to surface antigens enhance phagocytosis. Antigenic variation of the TprK protein is thought to contribute to development of subsequent stages of syphilis, persistence of infection, and susceptibility to reinfection with another strain. Comparative genomic studies have revealed genes with sequence variations among *T. pallidum* strains, leading to development of molecular typing methods used to examine syphilis outbreaks. Recent work has demonstrated that immunization with the outer-membrane protein Tp0751 significantly reduces dissemination of *T. pallidum* during syphilis infection in an animal model. Vaccine studies with this and other antigens are underway.

FURTHER READING

Cameron CE, Lukehart SA: Current status of syphilis vaccine development: Need, challenges, prospects. Vaccine 32:1602, 2014.

Centers for Disease Control and Prevention: Discordant results from reverse sequence syphilis screening—five laboratories, United States, 2006–2010. MMWR Morb Mortal Wkly Rep 60:133, 2011.

Dombrowski JC et al: Prevalence estimates of complicated syphilis. Sex Transm Dis 42:702, 2015.

Hook EW III: Syphilis. Lancet 389:1550, 2017.

Workowski KA, Bolan GA: Sexually transmitted diseases treatment guidelines, 2015. MMWR Recomm Rep 64:34, 2015.

Chapter 178: Endemic Treponematoses

Sheila A. Lukehart

INTRODUCTION

The endemic treponematoses are chronic diseases that are transmitted by direct contact, usually during childhood, and, like syphilis, can cause severe late manifestations years after initial infection. These diseases are caused by very close relatives of Treponema pallidum subspecies pallidum, the etiologic agent of venereal syphilis **(Chap. 177)**. Yaws, pinta, and endemic syphilis (bejel) are traditionally distinguished from venereal syphilis by mode of transmission, age of acquisition, geographic distribution, and clinical features; however, there is some overlap for each of these factors. Our "knowledge" about these infections is based on observations by health care workers who have visited endemic areas. Except for recent pilot programs of mass drug administration (MDA) for yaws, virtually no well-designed studies of the natural history, diagnosis, or treatment of these infections have been conducted. The treponemal infections are compared and contrasted in **Table 178-1**.

TABLE 178-1

Comparison of the Treponemes and Associated Diseases

Feature	Venereal Syphilis	Yaws	Endemic Syphilis	Pinta
Organism	*T. pallidum* subsp. *pallidum*	*T. pallidum* subsp. *pertenue*	*T. pallidum* subsp. *endemicum*	*T. carateum*
Common modes of transmission	Sexual, transplacental	Skin-to-skin	Mouth-to-mouth or via shared drinking/eating utensils	Skin-to-skin
Usual age of acquisition	Sexual maturity or in utero	Early childhood	Early childhood	Late childhood
Primary lesion	Cutaneous ulcer (chancre)	Papilloma, often ulcerative	Mucosal papule, rarely seen	Nonulcerating papule with satellites, pruritic
Common location	Genital, oral, anal	Extremities	Oral	Extremities, face
Secondary lesions	Cutaneous rash and mucocutaneous lesions; condylomata lata	Cutaneous papillomatous or ulcerative lesions; condylomata lata, osteoperiostitis	Mucocutaneous lesions (mucous patch, split papule, condylomata lata); osteoperiostitis	Pintides, pigmented, pruritic
Infectious relapses	~25%	Common	Unknown	Unknown
Late complications	Gummas, cardiovascular and central nervous system involvement[a]	Destructive gummas of skin, bone, cartilage	Destructive gummas of skin, bone, cartilage	Nondestructive, dyschromic, achromic macules

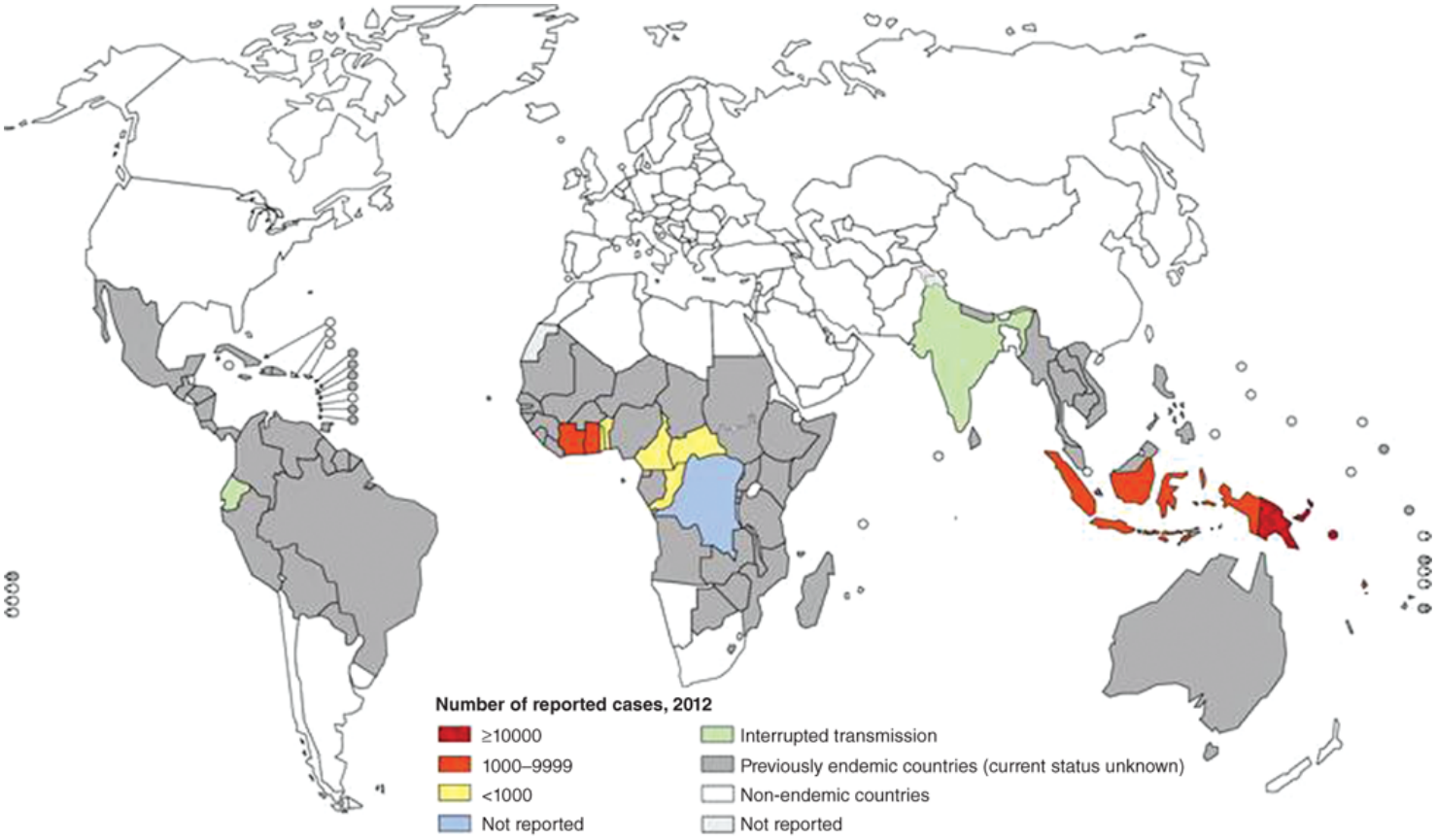
[a]Central nervous system involvement and congenital infection in the endemic treponematoses have been postulated by some investigators (see text).

EPIDEMIOLOGY

Generally, yaws flourishes in moist tropical areas **(Fig. 178-1)**; endemic syphilis has been found primarily in arid climates of West Africa and the Middle East; and pinta has been found in temperate foci in the Americas. Because no recent data are available for bejel and pinta, the extent of these infections today is unknown. The endemic treponematoses are usually limited to rural areas of developing nations and are seen in developed countries only among recent immigrants from endemic regions.

FIGURE 178-1

Geographic distribution of yaws in 2012. *(Reprinted with permission of the World Health Organization; http://www.who.int/yaws/epidemiology/Yaws_map_2012.png?ua=1.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In a World Health Organization (WHO)–sponsored mass eradication campaign from 1952 to 1969, >160 million people in Africa, Asia, and South America were examined for treponemal infections, and >50 million cases, contacts, and persons with latent infections were treated. This campaign reduced the prevalence of active yaws from >20% to <1% in many areas. In subsequent decades, lack of focused surveillance and diversion of resources resulted in documented resurgence of these infections in some regions. The most recent WHO global estimate (1995) suggested that there are 460,000 new cases per year (mostly yaws) and a prevalence of 2.5 million infected persons. In 2010–2013, a total of 256,000 cases were reported, primarily from countries in which focused yaws detection and treatment trials are ongoing. Areas of resurgent yaws morbidity include West Africa (Ivory Coast, Ghana, Togo, Benin), the Central African Republic, Nigeria, and the Democratic Republic of the Congo. The prevalence of endemic syphilis is estimated to be >10% in some regions of northern Ghana, Mali, Niger, Burkina Faso, and Senegal, although data are scarce. In Asia and the Pacific Islands, reports document active outbreaks of yaws in Indonesia, Papua New Guinea, the Solomon Islands, East Timor, and Vanuatu. India actively renewed its focus on yaws control in 1996, achieved zero-case status in 2003, declared elimination in 2006, and was declared yaws-free in 2016. In the Americas, foci of yaws had been thought to persist in Haiti and other Caribbean islands, Peru, Colombia, Ecuador, Brazil, Guyana, and Surinam, although recent data are lacking. Pinta is limited to Central America and northern South America, where it is found rarely and only in very remote villages. Evidence of yaws-like and genital manifestations, with treponemal seroreactivity, has been found in wild gorillas and baboons in both West and East Africa and has led to speculation that there may be an animal reservoir for yaws. Organisms very closely related to known *T. pallidum* subspecies *pertenue* isolates have been identified in lesions from affected baboons.

MICROBIOLOGY

The etiologic agents of the endemic treponematoses are listed in Table 178-1. These little-studied organisms are morphologically identical to *T. pallidum* subspecies *pallidum* (the agent of venereal syphilis), and no definitive antigenic differences among them have been identified to date. A

controversy has existed about whether the pathogenic treponemes are truly separate organisms, as genome sequencing indicates that yaws and syphilis treponemes are 99.8% identical. Three of the four etiologic agents are classified as subspecies of *T. pallidum*; the fourth (*T. carateum*) remains a separate species simply because no organisms have been available for genetic studies. Based on analysis of the small number of strains currently available, molecular signatures—assessed by restriction fragment length polymorphism and gene sequencing—have been identified that can differentiate the *T. pallidum* subspecies. Whether these genetic differences are related to distinct clinical characteristics of these diseases has not been determined. Full genome sequencing of a previously unclassified *Treponema* strain (Fribourg-Blanc), which was isolated from a baboon in 1966 and can cause experimental infection in humans, shows a very high degree of homology with available strains of *T. pallidum* subspecies *pertenue*. Recent genomic analyses of additional samples from nonhuman primates indicate a very close genetic relationship with known yaws isolates, but the importance of the nonhuman primate reservoir for human infection is not yet known.

CLINICAL FEATURES

All of the treponemal infections, including syphilis, are chronic and are characterized by defined disease stages, with a localized primary lesion, disseminated secondary lesions, periods of latency, and possible late lesions. Primary and secondary stages are more frequently overlapping in yaws and endemic syphilis than in venereal syphilis, and the late manifestations of pinta are very mild relative to the destructive lesions of the other treponematoses. The current preference is to divide the clinical course of the endemic treponematoses into “early” and “late” stages.

Historically, the major clinical distinctions made between venereal syphilis and the nonvenereal infections are the apparent lack of congenital transmission and of central nervous system (CNS) involvement in the nonvenereal infections. It is not known whether these distinctions are entirely accurate. Because of the high degree of genetic relatedness among the organisms, there is little biological reason to think that *T. pallidum* subspecies *endemicum* and *T. pallidum* subspecies *pertenue* would be unable to cross the blood–brain barrier or to invade the placenta. These organisms are like *T. pallidum* subspecies *pallidum* in that they obviously disseminate from the site of initial infection and can persist for decades. The lack of recognized congenital infection may be due to the fact that childhood infections often reach the latent stage (low bacterial load) before girls reach sexual maturity. Neurologic involvement may go unrecognized because of the lack of trained medical personnel in endemic regions, the delay of many years between infection and possible CNS manifestations, or a low rate of symptomatic CNS disease. Some published evidence supports congenital transmission as well as cardiovascular, ophthalmologic, and CNS involvement in yaws and endemic syphilis. Although the reported studies have been small, have failed to control for other causes of CNS abnormalities, and in some instances have not included serologic confirmation, it may be erroneous to accept unquestioningly the frequently repeated belief that these organisms fail to cause such manifestations.

Yaws

Also known as *pian*, *framboesia*, or *bouba*, yaws is characterized by the development of one or several primary lesions (“mother yaw”) followed by multiple disseminated skin lesions. All early skin lesions are infectious and may persist for many months; cutaneous relapses are common during the first 5 years. Late manifestations, affecting ~10% of untreated persons, are destructive lesions of skin, bone, and joints.

The infection is transmitted by direct contact with infectious lesions, often during play or group sleeping, and may be enhanced by disruption of the skin by insect bites or abrasions. While *T. pallidum* subspecies *pertenue* DNA has been detected on flies and fomites from endemic regions, there is not yet convincing evidence of insect or fomite transmission of infection. After an average of 3–4 weeks, the first lesion begins as a papule—usually on an extremity—and then enlarges (particularly during moist warm weather) to become ulcerated **(Fig 178-2A)** or papillomatous (“raspberry-like”—thus the name “framboesia”). Notably, recent data indicate that a large proportion of ulcerative lesions in yaws-endemic regions contain *Haemophilus ducreyi*, either as the sole etiologic agent or in combination with *T. pallidum* subspecies *pertenue*. (*H. ducreyi* DNA has also been detected on flies and fomites, as described above for *T. pallidum* subspecies *pertenue*.) Regional lymphadenopathy develops, and the lesion usually heals within 6 months; dissemination is thought to occur during the early weeks of infection. A generalized secondary eruption, accompanied by generalized lymphadenopathy, appears either concurrent with or after the primary lesion; may take several forms—macular, papular, or papillomatous **(Fig. 178-2B)**; and may become secondarily infected with other bacteria, including *H. ducreyi*. Painful papillomatous lesions on the soles of the feet result in a crablike gait (“crab yaws”), and periostitis may result in nocturnal bone pain and polydactylitis **(Fig. 178-2C)**. Late yaws is manifested by gummas of the skin and long bones, hyperkeratoses of the palms and soles, osteitis and periostitis, and hydrarthrosis. The late gummatous lesions are characteristically extensive. Destruction of the nose, maxilla, palate, and pharynx is termed *gangosa* and is similar to the destructive lesions seen in leprosy and leishmaniasis.

FIGURE 178-2

Clinical manifestations of early yaws. ***A.*** Primary ulcer. ***B.*** Secondary papillomata. ***C.*** Periostitis and polydactylitis. *(Photos were taken during a yaws elimination trial in Papua New Guinea and are published with permission from Dr. Oriol Mitjà.)*



A B C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Endemic Syphilis

The early lesions of endemic syphilis (*bejel*, *siti*, *dichuchwa*, *njovera*, *skerljevo*) are localized primarily to mucocutaneous and mucosal surfaces. The infection is reportedly transmitted by direct contact, by kissing, by premastication of food, or by sharing of drinking and eating utensils. A role for insects in transmission has been suggested but is unproven. The initial lesion, usually an intraoral papule, often goes unrecognized and is followed by mucous patches on the oral mucosa (Fig. 178-3A) and mucocutaneous lesions resembling the condylomata lata of secondary syphilis. This eruption may last for months or even years, and treponemes can readily be demonstrated in early lesions. Periostitis and regional lymphadenopathy are common. After a variable period of latency, late manifestations may appear, including osseous and cutaneous gummas. Destructive gummas, osteitis, and gangosa are more common in endemic syphilis than in yaws.

Pinta

Pinta (*mal del pinto*, *carate*, *azul*, *purupuru*) is the most benign of the treponemal infections. This disease has three stages that are characterized by marked changes in skin color (Fig. 178-3B), but pinta does not appear to cause destructive lesions or to involve tissues other than the skin. The initial papule is most often located on the extremities or face and is pruritic. After one to many months of infection, numerous disseminated secondary lesions (*pintides*) appear. These lesions are initially red but become deeply pigmented, ultimately turning a dark slate blue. The secondary lesions are infectious and highly pruritic and may persist for years. Late pigmented lesions are called *dyschromic macules* and contain treponemes. Over time, most pigmented lesions show varying degrees of depigmentation, becoming brown and eventually white and giving the skin a mottled appearance. White achromic lesions are characteristic of the late stage.

FIGURE 178-3
Clinical manifestations of endemic syphilis and pinta. ***A.*** Mucous patches of early endemic syphilis. ***B.*** Pigmented macules of early pinta. *(Photos reprinted with permission from the Handbook of Endemic Treponematoses, PL Perine et al, Geneva, World Health Organization, Color Plates 54, 60; 1984.)*



A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS

Diagnosis of the endemic treponematoses is based on clinical manifestations and, when available, dark-field microscopy and serologic testing. The same serologic tests that are used for venereal syphilis **(Chap. 177)** become reactive during all treponemal infections. Although several targets have been evaluated for specific serodiagnosis, to date there is no antibody test that can discriminate among the different infections. The nonvenereal treponemal infections should be considered in the evaluation of a reactive syphilis serology in any person who has emigrated from an endemic area. Sensitive polymerase chain reaction assays can be used to confirm treponemal infection and to identify the etiologic agent in research laboratories.

TREATMENT

TREATMENT

Endemic Treponematoses

The current WHO-recommended therapy for patients and their contacts includes either azithromycin (30 mg/kg, up to a maximum of 2 g) or benzathine penicillin G (1.2 million units IM for adults; 600,000 units for children <10 years old); these two drugs have been shown to be equivalent in a recent study. The recommended dose of benzathine penicillin is half of that recommended for early venereal syphilis, and no controlled efficacy studies have been conducted. Definitive evidence of resistance to penicillin is lacking, although relapsing lesions have been reported after penicillin treatment in Papua New Guinea.

The efficacy of single-dose azithromycin provided the WHO's revitalized yaws eradication program with a much easier regimen for use in mass treatment. Although macrolide resistance has become common in circulating strains of *T. pallidum* subspecies *pallidum* in many parts of the world, analysis of yaws samples from Papua New Guinea has only recently yielded evidence of mutations for resistance to macrolide antibiotics, including azithromycin, in a very small number of patients. Further surveillance is essential. Limited data suggest the efficacy of tetracycline for treatment of yaws, but no data exist for other endemic treponematoses. Solely on the basis of experience with venereal syphilis, it is thought that doxycycline or tetracycline (at doses appropriate for syphilis; **Chap. 177**) are alternatives, in addition to azithromycin, for patients allergic to penicillin. A Jarisch-Herxheimer reaction **(Chap. 177)** may follow treatment of endemic treponematoses. Nontreponemal serologic titers (in the Venereal Disease Research Laboratory [VDRL] slide test or the rapid plasma reagin [RPR] test) usually decline after effective therapy, but patients may not become seronegative.

CONTROL

Buoyed by the successful elimination of yaws in India and the availability of an inexpensive, single-dose oral drug for treatment, in 2012 the WHO renewed its efforts to eradicate yaws globally by 2020. Enthusiasm is high; several pilot programs of MDA have been conducted; and expansion of this approach to other regions is planned. Some caution is warranted: (1) Pilot studies indicate that a very high level of MDA coverage must be achieved, and mathematical modeling suggests that multiple rounds of MDA may be needed. Treatment must be followed by careful case detection and targeted treatment of cases and contacts. (2) The specter of increasing azithromycin resistance looms, and there may be only a short window of time during which countries can successfully use azithromycin for yaws eradication. Antibiotic resistance is of particular concern if multiple rounds of MDA are required. Further, given the ongoing campaigns against trachoma using low-dose azithromycin MDA, often in populations also at high risk for yaws, more widespread macrolide resistance seems inevitable. (3) Lastly, the possible animal reservoir needs be evaluated, particularly in Africa. Yaws elimination will require rapid implementation and scale-up of high-level drug coverage in endemic areas, and continued, careful surveillance by local health centers will be essential for success of this timely and important effort.

FURTHER READING

Giacani L, Lukehart SA: The endemic treponematoses. Clin Microbiol Rev 27:89, 2014.

Marks M et al: Mathematical modeling of programmatic requirements for yaws eradication. Emerg Infect Dis 23:22, 2017.

Mitjà O et al: Mass treatment with single-dose azithromycin for yaws. N Engl J Med 372:703, 2015.

Chapter 179: Leptospirosis

Jiři F. P. Wagenaar; Marga G. A. Goris

INTRODUCTION

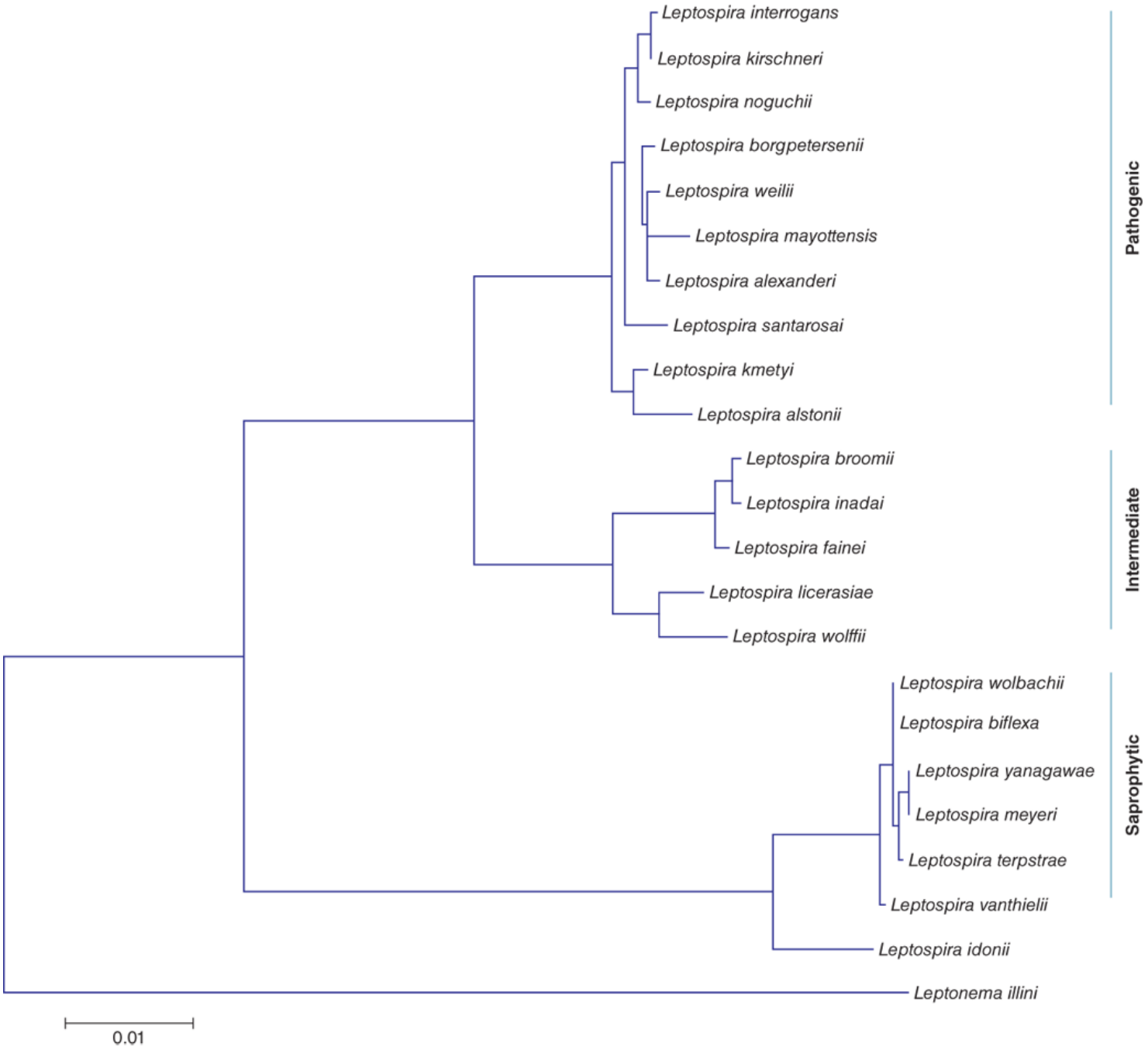
Leptospirosis is a globally important zoonotic disease whose apparent reemergence is illustrated by recent outbreaks on virtually all continents. The disease is caused by pathogenic *Leptospira* species and is characterized by a broad spectrum of clinical manifestations, varying from asymptomatic infection to fulminant, fatal disease. In its mild form, leptospirosis may present as nonspecific symptoms such as fever, headache, and myalgia. Severe leptospirosis, characterized by jaundice, renal dysfunction, and hemorrhagic diathesis, is often referred to as *Weil's syndrome*. With or without jaundice, severe pulmonary hemorrhage is increasingly recognized as an important presentation of severe disease.

ETIOLOGIC AGENT

Leptospira species are spirochetes belonging to the order Spirochaetales and the family Leptospiraceae. Traditionally, the genus *Leptospira* comprised two species: the pathogenic *L. interrogans* and the free-living *L. biflexa*, now designated *L. interrogans* sensu lato and *L. biflexa* sensu lato, respectively. Twenty-two *Leptospira* species with pathogenic (10 species), intermediate (5 species), and nonpathogenic (7 species) status have now been described on the basis of phylogenetic and virulence analyses **(Fig. 179-1)**. Genome sequences of five *Leptospira* species (*L. biflexa*, *L. interrogans*, *L. santarosai*, *L. borgpetersenii*, and *L. licerasiae*) have been published, and the availability of genome sequences of a wide variety of *Leptospira* strains will undoubtedly lead to a better understanding of the pathogenesis of leptospirosis. However, classification based on serologic differences better serves clinical, diagnostic, and epidemiologic purposes. Pathogenic *Leptospira* species are divided into serovars according to their antigenic composition. More than 250 serovars make up the 26 serogroups.

FIGURE 179-1

Differentiation of pathogenic, intermediate, and nonpathogenic (saprophytic) *Leptospira* species by molecular phylogenetic analysis using the *rrs* gene. Scale bar labeled 0.01 indicates the rate of nucleotide substitutions per base pair. *(Figure prepared and provided by Dr. A. Ahmed, Leptospirosis Reference Center, Academic Medical Center, Medical Microbiology, Amsterdam, The Netherlands.)*

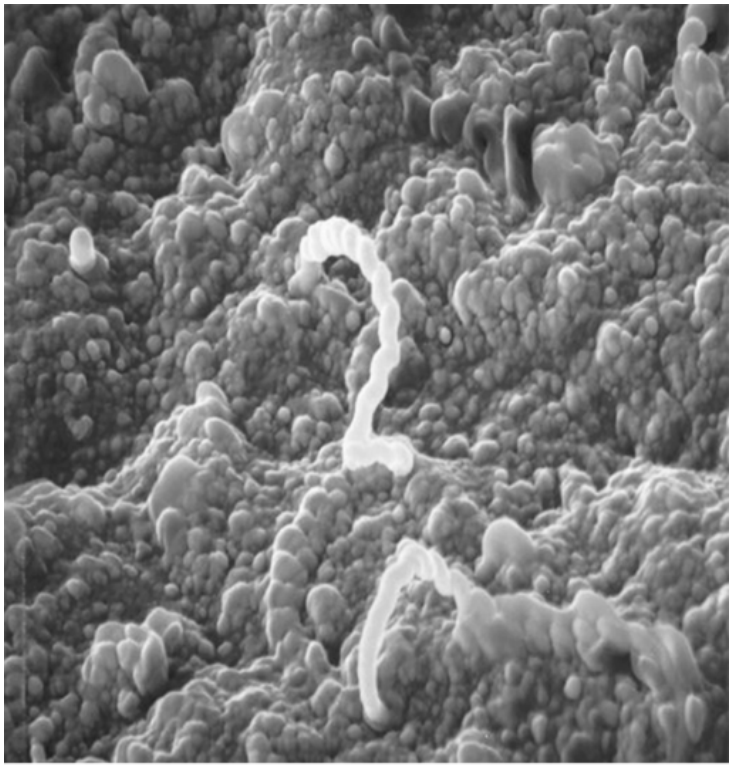


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Leptospires are coiled, thin, highly motile organisms that have hooked ends and two periplasmic flagella, with polar extrusions from the cytoplasmic membrane that are responsible for motility (Fig. 179-2). These organisms are 6–20 µm long and ~0.1 µm wide; they stain poorly but can be seen microscopically by dark-field examination and after silver impregnation staining of tissues. Leptospires require special media and conditions for growth; it may take weeks to months for cultures to become positive.

FIGURE 179-2

Transmission electron microscopic image of *Leptospira interrogans* invading equine conjunctival tissue. *(Image kindly provided by Dr. J. E. Nally, National Animal Disease Center, U.S. Department of Agriculture, Ames, IA.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPIDEMIOLOGY

Leptospirosis has a worldwide distribution but occurs most commonly in the tropics and subtropics because the climate and occasionally poor hygienic conditions favor the pathogen's survival and distribution. In most countries, leptospirosis is an underappreciated problem. Most cases occur in men, with a peak incidence during the summer and fall in both the Northern and Southern Hemispheres and during the rainy season in the tropics.

Reliable data on morbidity and mortality from leptospirosis have gradually started to appear. Current information on global human leptospirosis varies but indicates that ~1 million severe cases occur per year, with a mean case–fatality rate of nearly 10%.

As a zoonosis, leptospirosis affects almost all mammalian species and represents a significant veterinary burden. Rodents, especially rats, are the most important reservoir, although other wild mammals as well as domestic and farm animals may also harbor these microorganisms. Leptospires establish a symbiotic relationship with their host and can persist in the urogenital tract for years. Some serovars are generally associated with particular animals—e.g., Icterohaemorrhagiae and Copenhageni with rats, Grippotyphosa with voles, Hardjo with cattle, Canicola with dogs, and Pomona with pigs—but may occur in other animals as well.

Leptospirosis presents as both an endemic and an epidemic disease. Transmission of leptospires may follow direct contact with urine, blood, or tissue from an infected animal or, more commonly, exposure to environmental contamination. The dogma that human-to-human transmission is very rare is challenged by recent findings on household clustering, asymptomatic renal colonization, and prolonged excretion of leptospires. (Both of the latter features imply human infection sources that are not recognized.) Because leptospires can survive in a humid environment for many months, water is an important vehicle in their transmission. Epidemics of leptospirosis are not well understood. Outbreaks may result from exposure to flood waters contaminated by urine from infected animals, as has been reported from several countries. However, it is also true that outbreaks may occur without floods, and floods often occur without outbreaks.

The vast majority of infections with *Leptospira* cause no or only mild disease in humans. A small percentage of infections (~1%) lead to severe, potentially fatal complications. The proportion of leptospirosis cases that are mild is unknown because patients either do not seek or do not have access to medical care or because the nonspecific symptoms are interpreted as an influenza-like illness. Reported cases surely represent a significant underestimation of the total number. Certain occupational groups are at especially high risk, including veterinarians, agricultural workers, sewage workers, slaughterhouse employees, and workers in the fishing industry. Risk factors include direct or indirect contact with animals, including exposure to water and soil contaminated with animal urine. Leptospirosis has also been recognized in deteriorating inner cities and suburban areas where rat populations are expanding.

Recreational exposure and domestic-animal contact are prominent sources of leptospirosis. Recreational freshwater activities, such as canoeing, windsurfing, swimming, and waterskiing, place persons at risk for infection. Several outbreaks have followed sporting events. For example, an outbreak took place in 1998 among athletes after a triathlon in Springfield, Illinois. Ingestion of one or more swallows of lake water during the swimming leg of the triathlon was a prominent risk factor for illness. Heavy rains that preceded the triathlon, with consequent agricultural runoff, are likely to have increased the level of leptospiral contamination in the lake water. In another outbreak, 42% of participants contracted leptospirosis during the 2000 Eco-Challenge-Sabah multisport endurance race in Malaysian Borneo. Swimming in the Segama River was shown to be an independent risk factor.

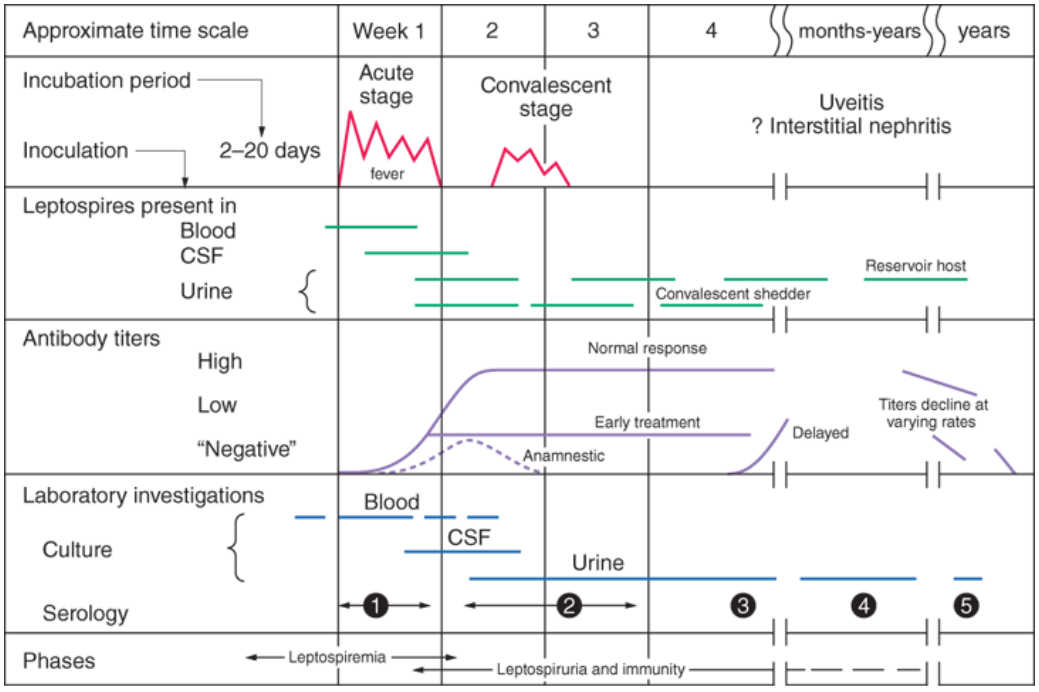
In addition, leptospirosis is a traveler's disease. Large proportions of patients acquire the infection while traveling in tropical countries, usually during adventurous activities such as whitewater rafting, jungle trekking, and caving. Transmission via laboratory accidents has been reported but is rare. New data indicate that leptospirosis may develop after unanticipated immersion in contaminated water (e.g., in an automobile accident) more frequently than has generally been thought and can also result from an animal bite.

PATHOGENESIS

Transmission occurs through cuts, abraded skin, or mucous membranes, especially the conjunctival and oral mucosa. After entry, the organisms proliferate, cross tissue barriers, and disseminate hematogenously to all organs (*leptospiremic phase*). During this initial incubation period, leptospires can be isolated from the bloodstream **(Fig. 179-3)**. The organisms are able to survive in the nonimmune host: they evade complement-mediated killing by binding factor H, a strong inhibitor of the complement system, on their surface. Moreover, pathogenic leptospires resist ingestion and killing by neutrophils, monocytes, and macrophages. During the *immune phase*, the appearance of antibodies coincides with the disappearance of leptospires from the blood. However, the bacteria persist in various organs, including liver, lung, kidney, heart, and brain. Autopsy findings illustrate the involvement of multiple organ systems in severe disease. Renal pathology shows both acute tubular damage and interstitial nephritis. Acute tubular lesions progress in time to interstitial edema and acute tubular necrosis. Severe nephritis is observed in patients who survive long enough to develop it and seems to be a secondary response to acute epithelial damage. The reported deregulation of the expression of several transporters along the nephron contributes to impaired sodium absorption, tubular potassium wasting, and polyuria. Histopathology of the liver shows focal necrosis (widespread hepatocellular necrosis usually is not found), foci of inflammation, and plugging of bile canaliculi. Hepatocyte apoptosis has also been documented. Experimental work showed infiltration of *Leptospira* in Disse's space and migration between hepatocytes with detachment of the intercellular junctions and disruption of bile canaliculi leading to bile leakage. Petechiae and hemorrhages are observed in the heart, lungs **(Fig. 179-4)**, kidneys (and adrenals), pancreas, liver, gastrointestinal tract (including retroperitoneal fat, mesentery, and omentum), muscles, prostate, testis, and brain (subarachnoid bleeding). Several studies show an association between hemorrhage and thrombocytopenia. Although the underlying mechanisms of thrombocytopenia have not been elucidated, it seems likely that platelet consumption plays an important role. A consumptive coagulopathy may occur, with elevated markers of coagulation activation (thrombin–antithrombin complexes, prothrombin fragments 1 and 2, D-dimer), diminished anticoagulant markers (antithrombin, protein C), and deregulated fibrinolytic activity. Overt disseminated intravascular coagulation (DIC) has been documented in several studies. Elevated plasma levels of soluble E-selectin and von Willebrand factor in patients with leptospirosis reflect endothelial cell activation. Experimental models show that pathogenic leptospires or leptospiral proteins are able to activate endothelial cells in vitro and to disrupt endothelial-cell barrier function, promoting dissemination. Platelets have been shown to aggregate on activated endothelium in the human lung, whereas histology reveals swelling of activated endothelial cells but no evident vasculitis or necrosis. Immunoglobulin and complement deposition have been demonstrated in lung tissue involved in pulmonary hemorrhage.

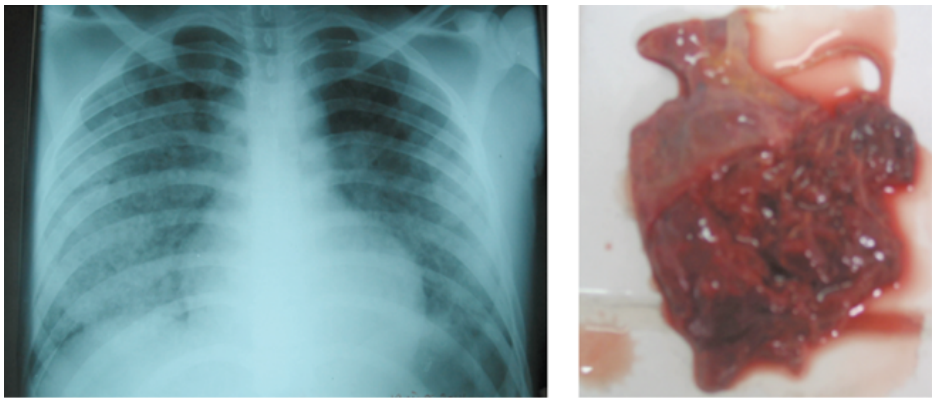
FIGURE 179-3

Biphasic nature of leptospirosis and relevant investigations at different stages of disease. *Note that an incubation period of up to 1 month has now been documented.* Specimens 1 and 2 for serology are acute-phase serum samples; specimen 3 is a convalescent-phase serum sample that may facilitate detection of a delayed immune response; and specimens 4 and 5 are follow-up serum samples that can provide epidemiologic information, such as the presumptive infecting serogroup. CSF, cerebrospinal fluid. *(Reprinted as adapted by PN Levett: Clin Microbiol Rev 14:296, 2001 [from LH Turner: Leptospirosis. BMJ 1:231, 1969] with permission from the American Society for Microbiology and the BMJ Publishing Group.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.
FIGURE 179-4

Severe pulmonary hemorrhage in leptospirosis. ***Left panel:*** Chest x-ray. ***Right panel:*** Gross appearance of right lower lobes of lung at autopsy. This patient, a 15-year-old from the Peruvian Amazonian city of Iquitos, died several days after presentation with acute illness, jaundice, and hemoptysis. Blood culture yielded *Leptospira interrogans* serovar Copenhageni/Icterohaemorrhagiae. *(Adapted with permission from E Segura et al: Clin Infect Dis 40:343, 2005. © 2005 by the Infectious Diseases Society of America.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Leptospira species have a typical double-membrane cell wall structure harboring a variety of membrane-associated proteins, including an unusually high number of lipoproteins. The peptidoglycan layer is located close to the cytoplasmic membrane. The lipopolysaccharide (LPS) in the outer membrane has an unusual structure with relatively low endotoxic potency. Pathogenic leptospires contain a variety of genes coding for proteins involved in motility and in cell and tissue adhesion and invasion that represent potential virulence factors. Many of these are surface-exposed outer-membrane proteins (OMPs). To date, the only leptospiral virulence factor shown to satisfy Koch's molecular postulates is *loa22* encoding a surface-exposed protein with an unknown function. However, the gene is not confined to pathogenic *Leptospira* species.

Immunity to *Leptospira* depends on the production of circulating antibodies to serovar-specific LPS. It is unclear whether other antigens play a significant role in protective humoral immunity. Moreover, immunity may not be confined to antibody responses; involvement of the innate-immune Toll-like receptor 2 (TLR2) and TLR4 activation pathways in controlling infection has been demonstrated, whereas in vaccinated cattle a cell-mediated immune response is correlated with protection. In recent work, the whole-blood transcriptome of patients from Brazil with severe leptospirosis (13 who survived and 3 who died) were studied. In fatal cases, expression of chemokines and the antimicrobial peptide cathelicidin was decreased, whereas transcription of proinflammatory cytokine pathways was more abundant. Previous studies had also highlighted the relation between an exaggerated proinflammatory immune response and death. Moreover, the recent study provided evidence that patients who die of leptospirosis fail to mount an adequate humoral immune response to leptospires.

It is likely that several surface-exposed proteins mediate leptospire–host cell interactions, and these proteins may represent candidate vaccine components. Although animal-model studies have shown various degrees of vaccine efficacy for various putative virulence-associated OMPs, it is not yet clear whether such proteins elicit acceptable levels of sterilizing immunity.

CLINICAL MANIFESTATIONS

Although leptospirosis is a potentially fatal disease with bleeding and multiorgan failure as its clinical hallmarks, the majority of cases are thought to be relatively mild, presenting as the sudden onset of a febrile illness. The incubation period is usually 1–2 weeks but ranges from 1 to 30 days. (Figure 179-3 indicates a slightly different range, but an incubation period of up to 1 month has now been documented.) Leptospirosis is classically described as biphasic. The acute *leptospiremic phase* is characterized by fever of 3–10 days' duration, during which time the organism can be cultured from blood and detected by polymerase chain reaction (PCR). During the *immune phase*, resolution of symptoms may coincide with the appearance of antibodies, and leptospires can be cultured from the urine. The distinction between the first and second phases is not always clear: milder cases do not always include the second phase, and severe disease may be monophasic and fulminant. The idea that distinct clinical syndromes are associated with specific serogroups has been refuted, although some serovars tend to cause more severe disease than others.

Mild Leptospirosis

Most patients are asymptomatic or only mildly ill and do not seek medical attention. Serologic evidence of past inapparent infection is frequently found in persons who have been exposed but have not become ill. Mild symptomatic leptospirosis usually presents as a flu-like illness of sudden onset, with fever, chills, headache, nausea, vomiting, abdominal pain, conjunctival suffusion (redness without exudate), and myalgia. Muscle pain is intense and especially affects the calves, back, and abdomen. The headache is intense, localized to the frontal or retroorbital region (resembling that occurring in dengue), and sometimes accompanied by photophobia. Aseptic meningitis may be present and is more common among children than among adults. Although *Leptospira* can be cultured from the cerebrospinal fluid (CSF) in the early phase, the majority of cases follow a benign course with regard to the central nervous system; symptoms usually disappear within a few days but may persist for weeks.

Physical examination may include any of the following findings, none of which is pathognomonic for leptospirosis: fever, conjunctival suffusion, pharyngeal injection, muscle tenderness, lymphadenopathy, rash, meningismus, hepatomegaly, and splenomegaly. If present, the rash is often transient; may be macular, maculopapular, erythematous, or hemorrhagic (petechial or ecchymotic); and may be misdiagnosed as due to scrub typhus or viral infection. Lung auscultation may reveal crackles. Mild jaundice may be present.

The natural course of mild leptospirosis usually involves spontaneous resolution within 7–10 days, but persistent symptoms have been documented. In the absence of a clinical diagnosis and antimicrobial therapy, the mortality rate in mild leptospirosis is low.

Severe Leptospirosis

Although the onset of severe leptospirosis may be no different from that of mild leptospirosis, severe disease is often rapidly progressive and is associated with a case–fatality rate ranging from 1 to 50%. Higher mortality rates are associated with an age >40, altered mental status, acute renal failure, respiratory insufficiency, hypotension, and arrhythmias. The classic presentation, often referred to as *Weil's syndrome*, encompasses the triad of hemorrhage, jaundice, and acute kidney injury.

Patients die of septic shock with multiorgan failure and/or severe bleeding complications that most commonly involve the lungs (pulmonary hemorrhage), gastrointestinal tract (melena, hemoptysis), urogenital tract (hematuria), and skin (petechiae, ecchymosis, and bleeding from venipuncture sites). Pulmonary hemorrhage (with or without jaundice) is now recognized as a widespread public health problem, presenting with cough, chest pain, respiratory distress, and hemoptysis that may not be apparent until patients are intubated.

Jaundice occurs in 5–10% of all patients with leptospirosis; it can be profound and give an orange cast to the skin but usually is not associated with fulminant hepatic necrosis. Physical examination may reveal an enlarged and tender liver.

Acute kidney injury is common in severe disease, presenting after several days of illness, and can be either nonoliguric or oliguric. Typical electrolyte abnormalities include hypokalemia and hyponatremia. Loss of magnesium in the urine is uniquely associated with leptospiral nephropathy. Hypotension is associated with acute tubular necrosis, oliguria, or anuria, requiring fluid resuscitation and sometimes vasopressor therapy. Hemodialysis can be lifesaving, with renal function typically returning to normal in survivors.

In severe leptospirosis, altered mental status may reflect leptospiral meningitis. The diagnosis of leptospiral meningitis may be challenging since patients may be anicteric or lack other diagnostic hallmarks of severe leptospirosis. Without proper antibiotic treatment, a mortality rate of 13% has been reported; in contrast, among patients treated with antibiotics, the mortality rate is 2%. Neurologic sequelae are described until months after acute illness.

Other syndromes include (necrotizing) pancreatitis, cholecystitis, skeletal muscle involvement, and rhabdomyolysis with moderately elevated levels of serum creatine kinase. Cardiac involvement is commonly reflected on the electrocardiogram as nonspecific ST- and T-wave changes. Repolarization

abnormalities and arrhythmias are considered poor prognostic factors. Myocarditis has been described. Rare hematologic complications include hemolysis, thrombotic thrombocytopenic purpura, and hemolytic-uremic syndrome.

Long-term symptoms following severe leptospirosis include fatigue, myalgia, malaise, and headache and may persist for years. Autoimmune-associated uveitis, a potentially chronic condition, is a recognized sequela of leptospirosis.

DIAGNOSIS

The clinical diagnosis of leptospirosis should be based on an appropriate exposure history combined with any of the protean manifestations of the disease. Returning travelers from endemic areas usually have a history of recreational freshwater activities or other mucosal or percutaneous contact with contaminated surface waters or soil. For nontravelers, recreational or accidental water/soil contact and occupational hazards that involve direct or indirect animal contact should be explored (see "Epidemiology," above).

Although biochemical, hematologic, and urinalysis findings in acute leptospirosis are nonspecific, certain patterns may suggest the diagnosis. Laboratory results usually show signs of a bacterial infection, including leukocytosis with a left shift and elevated markers of inflammation (C-reactive protein level, procalcitonin level, and erythrocyte sedimentation rate). Thrombocytopenia (platelet count $\leq 100 \times 10^9$/L) is common and is associated with bleeding and renal failure. In severe disease, signs of coagulation activation may be present, varying from borderline abnormalities to a serious derangement compatible with DIC as defined by international criteria. The kidneys are invariably involved in leptospirosis. Related findings range from urinary sediment changes (leukocytes, erythrocytes, and hyaline or granular casts) and mild proteinuria in mild disease to renal failure and azotemia in severe leptospirosis. Nonoliguric hypokalemic renal insufficiency (see "Clinical Manifestations," above) is characteristic of early leptospirosis. Serum bilirubin levels may be high, whereas rises in aminotransferase and alkaline phosphatase levels are usually moderate. Although clinical symptoms of pancreatitis are not a common finding, amylase levels are often elevated. When symptoms of meningitis develop, examination of the CSF shows pleocytosis that can range from a few cells to >1000 cells/μL, with a predominance of lymphocytes. Predominant polymorphonuclear pleocytosis has been reported. This phenomenon may be related to the timing of the lumbar puncture: polymorphonuclear cells are thought to be found in early disease and are later replaced by lymphocytes. Although protein and glucose levels in CSF are usually normal, protein levels may be slightly elevated.

In severe leptospirosis, pulmonary radiographic abnormalities are more common than would be expected on the basis of physical examination (Fig. 179-4). The most common radiographic finding is a patchy bilateral alveolar pattern that corresponds to scattered alveolar hemorrhage. These abnormalities predominantly affect the lower lobes. Other findings include pleura-based densities (representing areas of hemorrhage) and diffuse ground-glass attenuation typical of acute respiratory distress syndrome (ARDS).

A definitive diagnosis of leptospirosis is based on isolation of the organism from the patient, on a positive PCR result, or on seroconversion or a rise in antibody titer. In cases with strong clinical evidence of infection, a single antibody titer of 1:200–1:800 (depending on whether the case occurs in a low- or high-endemic area) in the microscopic agglutination test (MAT) is required. Preferably, a fourfold or greater rise in titer is detected between acute- and convalescent-phase serum specimens. Antibodies generally do not reach detectable levels until the second week of illness. The antibody response can be affected by early treatment with antibiotics.

The MAT, which uses a battery of live leptospiral strains, and the enzyme-linked immunosorbent assay (ELISA), which uses a broadly reacting antigen, are the standard serologic procedures. The MAT usually is available only in specialized laboratories and is used for determination of the antibody titer and for tentative identification of the involved leptospiral serogroup—and, when epidemiologic background information is available, the putative serovar. This point underscores the importance of testing antigens representative of the serovars prevalent in the particular geographic area. However, cross-reactions occur frequently, and thus definitive identification of the infecting serovar or serogroup is not possible without isolation of the causative organism. Because serologic testing lacks sensitivity in the early acute phase of the disease (up to day 5), it cannot be used as the basis for a timely decision about whether to start treatment.

In addition to the MAT and the ELISA, various rapid tests with diagnostic value have been developed, and some of these are commercially available. These rapid tests mainly apply lateral flow, (latex) agglutination, or ELISA methodology and are reasonably sensitive and specific, although results reported in the literature vary, probably as a consequence of differences in test interpretation, (re)exposure risks, serovar distribution, and the use of biased serum panels. These methods do not require culture or MAT facilities and are useful in settings that lack a strong medical infrastructure. PCR methodologies, notably real-time PCR, have become increasingly widely implemented. Compared with serology, PCR offers a great advantage: the capacity to confirm the diagnosis of leptospirosis with a high degree of accuracy during the first 5 days of illness.

DIFFERENTIAL DIAGNOSIS

The differential diagnosis of leptospirosis is broad, reflecting the diverse clinical presentations of the disease. Although leptospirosis transmission is more common in tropical and subtropical regions, the absence of a travel history does not exclude the diagnosis. When fever, headache, and myalgia predominate, influenza and other common and less common viral infections (e.g., dengue and chikungunya) should be considered. Malaria, typhoid fever, ehrlichiosis, viral hepatitis, and acute HIV infection may mimic the early stages of leptospirosis and are important to recognize. Rickettsial diseases as well as dengue and hantavirus infections (hemorrhagic fever with renal syndrome or hantavirus cardiopulmonary syndrome) share epidemiologic and clinical

features with leptospirosis. Dual infections have been reported. In this light, it is advisable to conduct serologic testing for rickettsiae, dengue virus, and hantavirus when leptospirosis is suspected. When bleeding is detected, dengue hemorrhagic fever and other viral hemorrhagic fevers, including hantavirus infection, yellow fever, Rift Valley fever, filovirus infections, and Lassa fever, should be considered.

TREATMENT

TREATMENT

Leptospirosis

Severe leptospirosis should be treated with IV penicillin **(Table 179-1)** as soon as the diagnosis is considered. Leptospires are highly susceptible to a broad range of antibiotics, including the β-lactam antibiotics, cephalosporins, aminoglycosides, and macrolides, but are not susceptible to vancomycin, rifampin, metronidazole, and chloramphenicol. Early intervention may prevent the development of major organ-system failure or lessen its severity. Although studies supporting antibiotic therapy have produced conflicting results, clinical trials are difficult to perform in settings where patients frequently present for medical care with late stages of disease. Antibiotics are less likely to benefit patients in whom organ damage has already occurred. Two open-label randomized studies comparing penicillin with parenteral cefotaxime, parenteral ceftriaxone, and doxycycline showed no significant differences among the antibiotics with regard to complications and mortality risk. Thus ceftriaxone, cefotaxime, or doxycycline is a satisfactory alternative to penicillin for the treatment of severe leptospirosis. Antimicrobial susceptibility testing is not routine practice in individual cases of leptospirosis; to date, however, antibiotic resistance has not been reported in isolates from patients or the environment.

In mild cases, oral treatment with doxycycline, azithromycin, ampicillin, or amoxicillin is recommended. In regions where rickettsial diseases are coendemic, doxycycline or azithromycin is the drug of choice. In rare instances, a Jarisch-Herxheimer reaction develops within hours after the initiation of antimicrobial therapy.

Aggressive supportive care for leptospirosis is essential and can be lifesaving. Patients with nonoliguric renal dysfunction require aggressive fluid and electrolyte resuscitation to prevent dehydration and precipitation of oliguric renal failure. Peritoneal dialysis or hemodialysis should be provided to patients with oliguric renal failure. Rapid initiation of hemodialysis has been shown to reduce mortality risk and typically is necessary only for short periods. Patients with pulmonary hemorrhage may have reduced pulmonary compliance (as seen in ARDS) and may benefit from mechanical ventilation with low tidal volumes to avoid high ventilation pressures. Evidence is contradictory for the use of glucocorticoids and desmopressin as adjunct therapy for pulmonary involvement associated with severe leptospirosis.

TABLE 179-1

Treatment and Chemoprophylaxis of Leptospirosis in Adults[a]

Indication	Regimen
Treatment	
Mild leptospirosis	Doxycycline[b] (100 mg PO bid) *or* Amoxicillin (500 mg PO tid) *or* Ampicillin (500 mg PO tid)
Moderate/severe leptospirosis	Penicillin (1.5 million units IV or IM q6h) *or* Ceftriaxone (2 g/d IV) *or* Cefotaxime (1 g IV q6h) *or* Doxycycline[b] (loading dose of 200 mg IV, then 100 mg IV q12h)
Chemoprophylaxis[c]	
	Doxycycline[b] (200 mg PO once a week) *or* Azithromycin (250 mg PO once or twice a week)

[a]All regimens are given for 7 days. [b]Doxycycline should not be given to pregnant women or children. [c]The efficacy of doxycycline prophylaxis in endemic or epidemic settings remains unclear. Experiments in animal models and a cost-effectiveness model indicate that azithromycin has a number of characteristics that may make it efficacious in treatment and prophylaxis.

PROGNOSIS

Most patients with leptospirosis recover. However, post-leptospirosis symptoms, mainly of a depression-like nature, may occur and persist for years after the acute disease. Mortality rates are highest among patients who are elderly and those who have severe disease (pulmonary hemorrhage, Weil's syndrome). Leptospirosis during pregnancy is associated with high fetal mortality rates. Long-term follow-up of patients with renal failure and hepatic dysfunction has documented good recovery of renal and hepatic function.

PREVENTION

Individuals who may be exposed to *Leptospira* through their occupations or their involvement in recreational freshwater activities should be informed about the risks. Measures for controlling leptospirosis include avoidance of exposure to urine and tissues from infected animals through proper eyewear, footwear, and other protective equipment. Targeted rodent-control strategies could be considered.

 Vaccines for agricultural and companion animals are generally available, and their use should be encouraged. The veterinary vaccine used in a given area should contain the serovars known to be present in that area. Unfortunately, some vaccinated animals still excrete leptospires in their urine. Vaccination of humans against a specific serovar prevalent in an area has been undertaken in some European and Asian countries and has proved effective. Although a large-scale trial of vaccine in humans has been reported from Cuba, no conclusions can be drawn about efficacy and adverse reactions because of insufficient details on study design. The efficacy of chemoprophylaxis with doxycycline (200 mg once a week) or azithromycin (in pregnant women and children) is being disputed, but focused pre- and postexposure administration is indicated in instances of well-defined short-term exposure (Table 179-1).

ACKNOWLEDGMENT

The authors thank Rudy A. Hartskeerl for his significant contributions to this chapter in the previous edition of the textbook.

FURTHER READING

Adler B: Vaccines against leptospirosis. Curr Top Microbiol Immunol 387:251, 2015.

Arean VM: The pathologic anatomy and pathogenesis of fatal human leptospirosis (Weil's disease). Am J Pathol 40:393, 1962.

Costa F et al: Global morbidity and mortality of leptospirosis: A systematic review. PLoS Negl Trop Dis 9:e0003898, 2015. [PubMed: 26379143]

Goris MG et al: Prospective evaluation of three rapid diagnostic tests for diagnosis of human leptospirosis. PLoS Negl Trop Dis 7:e2290, 2013. [PubMed: 23875034]

Goris MG et al: Towards the burden of human leptospirosis: Duration of acute illness and occurrence of post-leptospirosis symptoms of patients in the Netherlands. PLoS One 8:e76549, 2013. [PubMed: 24098528]

Haake DA, Levett PN: Leptospirosis in humans. Curr Top Microbiol Immunol 387:65, 2015.

Picardeau M: Virulence of the zoonotic agent of leptospirosis: Still terra incognita? Nat Rev Microbiol 15:297, 2017.

Chapter 180: Relapsing Fever

Alan G. Barbour

INTRODUCTION

Relapsing fever is caused by infection with any of several species of *Borrelia* spirochetes. Physicians in ancient Greece distinguished relapsing fever from other febrile disorders by its characteristic clinical presentation: two or more episodes of fever separated by varying periods of well-being. In the nineteenth century, relapsing fever was one of the first diseases to be associated with a specific microbe by virtue of its characteristic laboratory finding: the presence of large numbers of spirochetes of the genus *Borrelia* in the blood.

The host responds with systemic inflammation that results in an illness ranging from a flu-like syndrome to sepsis. Other manifestations are the consequences of central nervous system (CNS) involvement and coagulopathy. Antigenic variation of the spirochetes' surface proteins accounts for the infection's relapsing course. Acquired immunity follows the serial development of antibodies to each of the several variants appearing during an infection. Treatment with antibiotics results in rapid cure but at the risk of a moderate to severe Jarisch-Herxheimer reaction.


Louse-borne relapsing fever (LBRF) caused large epidemics well into the twentieth century and currently occurs in northeastern Africa and among migrants from that area. At present, however, most cases of relapsing fever are tick-borne in origin. Sporadic cases and small outbreaks are focally distributed on most continents, with Africa and Central Asia most affected. In North America, the majority of reports of relapsing fever have been from the western United States and Canada. Nevertheless, the recent discovery that another species in the relapsing fever group causes human disease in the same geographic distribution as Lyme disease **(Chap. 181)** confounds epidemiologic distinctions between the two infections.

ETIOLOGIC AGENT

Coiled, thin microscopic filaments that swim in one direction and then coil up before heading in another were first observed in the blood of patients with relapsing fever in the 1880s (*www.youtube.com/watch?v=VxDPV2lBd9U*). These microbes were categorized as spirochetes and assigned to the genus *Borrelia*. It was not until the 1960s that the organisms were isolated in pure culture. The breakthrough cultivation medium was rich in ingredients, ranging from simple (e.g., *N*-acetylglucosamine) to more complex (e.g., serum). The

limited biosynthetic capacity of *Borrelia* cells is accounted for by a genome content one-quarter that of *Escherichia coli*.

Like other spirochetes, the helix-shaped *Borrelia* cells have two membranes, the outer of which is more loosely secured than in other double-membrane bacteria, such as *E. coli*. As a consequence, fixed organisms with damaged membranes can assume a variety of morphologies in smears and histologic preparations. The flagella of spirochetes run between the two membranes and are not on the cell surface. Although technically gram-negative in their staining properties, the 10- to 20-µm-long *Borrelia* cells, with a diameter of 0.2–0.3 µm, are too narrow to be seen by bright-field microscopy.

EPIDEMIOLOGY

 The several species of *Borrelia* that cause relapsing fever have restricted geographic distributions **(Table 180-1)**. The exception is *Borrelia recurrentis*, which is also the only species transmitted by an insect. LBRF is usually acquired from a body louse (*Pediculus humanus corporis*), with humans serving as the reservoir. Acquisition occurs not from the bite itself but from either rubbing of the insect's feces into the bite site with the fingers in response to irritation or inoculation of feces into the conjunctivae or an open wound. Although LBRF transmission is currently limited to Ethiopia and adjacent countries, the disease has had a global distribution in the past, and that potential remains. Epidemics with thousands of cases of LBRF can occur under circumstances of famine, refugee migration, and war. Transmission of LBRF can occur in camps of migrants at a distance from their home countries.

TABLE 180-1

Relapsing Fever *Borrelia* Species, by Geographic Region, Vector, and Primary Reservoir

Species	Region(s)	Arthropod Vector(s)	Primary Reservoir
B. crocidurae	Africa	*Ornithodoros sonrai* (soft ticks)	Mammals
B. duttonii	Africa	*O. moubata*	Humans
B. hermsii	North America	*O. hermsi*	Mammals
B. hispanica	Europe, North Africa	*O. erraticus*	Mammals
B. mazzottii	Mexico, Central America	*O. talaje*	Mammals
B. miyamotoi	Eurasia, North America	*Ixodes* species (hard ticks)	Mammals
B. persica	Eurasia	*O. tholozani*	Mammals
B. recurrentis	Africa, global[a]	*Pediculus humanus corporis* (human body louse)	Humans
B. turicatae	North America	*O. turicata*	Mammals
B. venezuelensis	Central and South America	*O. rudis*	Mammals

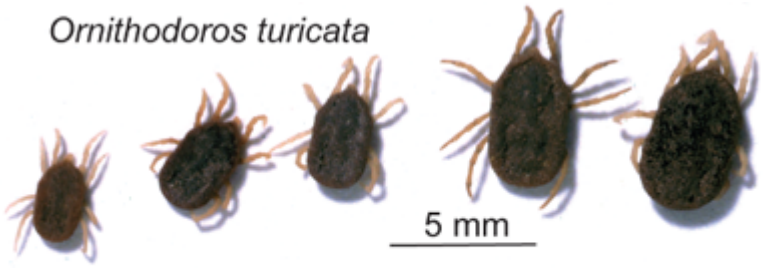
[a]Although transmission is currently limited to Ethiopia and adjacent countries, *B. recurrentis* infection has had a global distribution in the past, and that potential remains.

All other known species of relapsing fever agents are tick-borne—in most cases, by soft ticks of the genus *Ornithodoros* (Fig. 180-1). Tick-borne relapsing fever (TBRF) is found on most continents but is absent or rare in tropical, low-desert, or arctic environments. For most species, the reservoirs of infection are small to medium-sized mammals, usually rodents but sometimes pigs and other domestic animals living in or around human habitats. However, one species, *Borrelia duttonii* in sub-Saharan Africa, is largely maintained by tick transmission between human hosts. In North America, TBRF occurs as single cases or small case clusters through transient exposure of persons to infested buildings or caves in less populated areas where the rodent reservoirs have nests. The two main *Borrelia* species involved in North America are *Borrelia hermsii* (in the mountainous west) and *Borrelia turicatae* (in the southwestern and south-central regions). The soft

tick vectors typically feed for no more than 30 min, usually while the victim is sleeping. Transovarial transmission from one generation of ticks to the next means that infection risk may persist in an area long after incriminated mammalian reservoirs have been removed.

FIGURE 180-1

Ornithodoros turicata **soft ticks** of different ages.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A newly recognized pathogen, *Borrelia miyamotoi*, belongs to the same clade as relapsing fever species but is transmitted to humans from other mammals by the hard ticks (e.g., *Ixodes scapularis* in the eastern United States) that also transmit Lyme disease, babesiosis, anaplasmosis, and a viral encephalitis. Similar to *Borreliella*[*] *(Borrelia) burgdorferi* (Chap. 181), *B. miyamotoi* is acquired through contact with ticks in forested and shrubby areas during recreation, work, or outdoor activities around the home. In residents of areas where *B. miyamotoi* and *B. burgdorferi* coexist, the prevalence of antibodies to the former is about one-third of that to the latter. In contrast to that of *B. burgdorferi*, the transmission of *B. miyamotoi* to the host begins soon after the tick begins to feed.

[*] *Editors' note:* The nomenclature for the genus causing Lyme disease is currently being reviewed and is in transition. In this chapter, the proposed new genus designation, *Borreliella*, is used.

PATHOGENESIS AND IMMUNITY

Unlike LBRF spirochetes, TBRF spirochetes enter the body in the tick's saliva with the onset of feeding. From an inoculum of a few cells, the spirochetes proliferate in the blood, doubling every 6 h to numbers of $\geq 10^6$–10^7/mL. *Borrelia* species are extracellular pathogens; their presence inside cells connotes dead bacteria after phagocytosis. Binding of the spirochetes to erythrocytes leads to aggregation of red blood cells, their sequestration in the spleen and liver, and hepatosplenomegaly and anemia. A bleeding disorder is probably the consequence of thrombocytopenia, impaired hepatic production of clotting factors, and/or blockage of small vessels by aggregates of spirochetes, erythrocytes, and platelets. Some species are neurotropic and enter the brain, where they are comparatively sheltered from host immunity. Relapsing fever spirochetes can cross the maternal–fetal barrier and cause placental damage and inflammation, leading to intrauterine growth retardation and congenital infection.

Although *Borrelia* species do not have potent exotoxins or a lipopolysaccharide endotoxin, they have abundant lipoproteins whose binding by Toll-like receptors on host cells can lead to a proinflammatory process similar to that in endotoxemia, with elevations of tumor necrosis factor α, interleukin 6, and interleukin 8 concentrations.

IgM antibodies specific for the serotype-defining surface lipoprotein appear after a few days of infection and soon reach a concentration that causes lysis of bacteria in the blood through either direct bactericidal action or opsonization. The release of lipoproteins and other bacterial products from dying bacteria provokes a "crisis," during which there can be an increase in temperature, hypotension, and other signs of shock. A similar phenomenon occurring in some patients soon after the initiation of antibiotic treatment is characterized by an abrupt worsening of the patient's condition, termed a *Jarisch-Herxheimer reaction*.

CLINICAL MANIFESTATIONS

Relapsing fever presents with the sudden onset of fever. Febrile periods are punctuated by intervening afebrile periods of a few days; this pattern occurs at least twice. The patient's temperature is ≥39°C and may be as high as 43°C. The first fever episode often ends in a crisis lasting ~15–30 min and consisting of rigors, a further elevation in temperature, and increases in pulse and blood pressure. The crisis phase is followed by profuse diaphoresis, falling temperature, and hypotension, which usually persist for several hours. In LBRF, the first episode of fever is unremitting for 3–6 days; it is usually followed by a single milder episode. In TBRF, multiple febrile periods last 1–3 days each. In both forms, the interval between fevers ranges from 4 to 14 days, sometimes with symptoms of malaise and fatigue.

The symptoms that accompany the fevers are usually nonspecific. Headache, neck stiffness, arthralgia, myalgia, and nausea may accompany the first and subsequent febrile episodes. An enlarging spleen and liver cause abdominal pain. A nonproductive cough is common during LBRF and—in combination with fever and myalgias—may suggest influenza. Acute respiratory distress syndrome may occur during TBRF.

On physical examination, the patient may be delirious or apathetic. There may be body lice in the patient's clothes or signs of insect bites. In regions with *B. miyamotoi* infection, a hard tick may be embedded in the skin. Epistaxis, petechiae, and ecchymoses are common during LBRF but not during TBRF. Splenomegaly or spleen tenderness is common in both forms of relapsing fever. The majority of patients with LBRF and ~10% of patients with TBRF have discernible hepatomegaly.

Localizing neurologic findings are more common in TBRF than in LBRF. In North America, *B. turicatae* infection has neurologic manifestations more often than *B. hermsii* infection. Meningoencephalitis can result in residual hemiplegia or aphasia. Myelitis and radiculopathy may develop. Unilateral or bilateral Bell's palsy and deafness from seventh or eighth cranial nerve involvement are the most common forms of cranial neuritis and typically present in the second or third febrile episode, not the first. Visual impairment from unilateral or bilateral iridocyclitis or panophthalmitis may be permanent. In LBRF, neurologic manifestations such as altered mental state or stiff neck are thought to be secondary to systemic inflammation rather than to direct invasion of the nervous system.

Myocarditis appears to be common in both forms of relapsing fever and accounts for some deaths. Most commonly, myocarditis is evidenced by gallops on cardiac auscultation, a prolonged QT_c interval, and cardiomegaly and pulmonary edema on chest radiography.

General laboratory studies are not specific. Mild to moderate normocytic anemia is common, but frank hemolysis and hemoglobinuria do not develop. Leukocyte counts are usually in the normal range or only slightly elevated, and leukopenia can occur during the crisis. Platelet counts can fall below 50,000/μL. Laboratory evidence of hepatitis can be found, with elevated serum concentrations of unconjugated bilirubin and aminotransferases; the prothrombin and partial thromboplastin times may be moderately prolonged.

Analysis of the cerebrospinal fluid (CSF) is indicated in cases of suspected relapsing fever with signs of meningitis or meningoencephalitis. The presence of mononuclear pleocytosis and mildly to moderately elevated protein levels justifies intravenous antibiotic therapy in TBRF.

The manifestations and course of *B. miyamotoi* infection remain incompletely characterized. The most common presentation is fever without respiratory symptoms starting 1–2 weeks after a tick bite and recurring once or twice in some cases. Patients have been hospitalized with a presumptive diagnosis of undifferentiated sepsis. Meningoencephalitis, with spirochetes in the CSF, has been documented in immunodeficient adults.

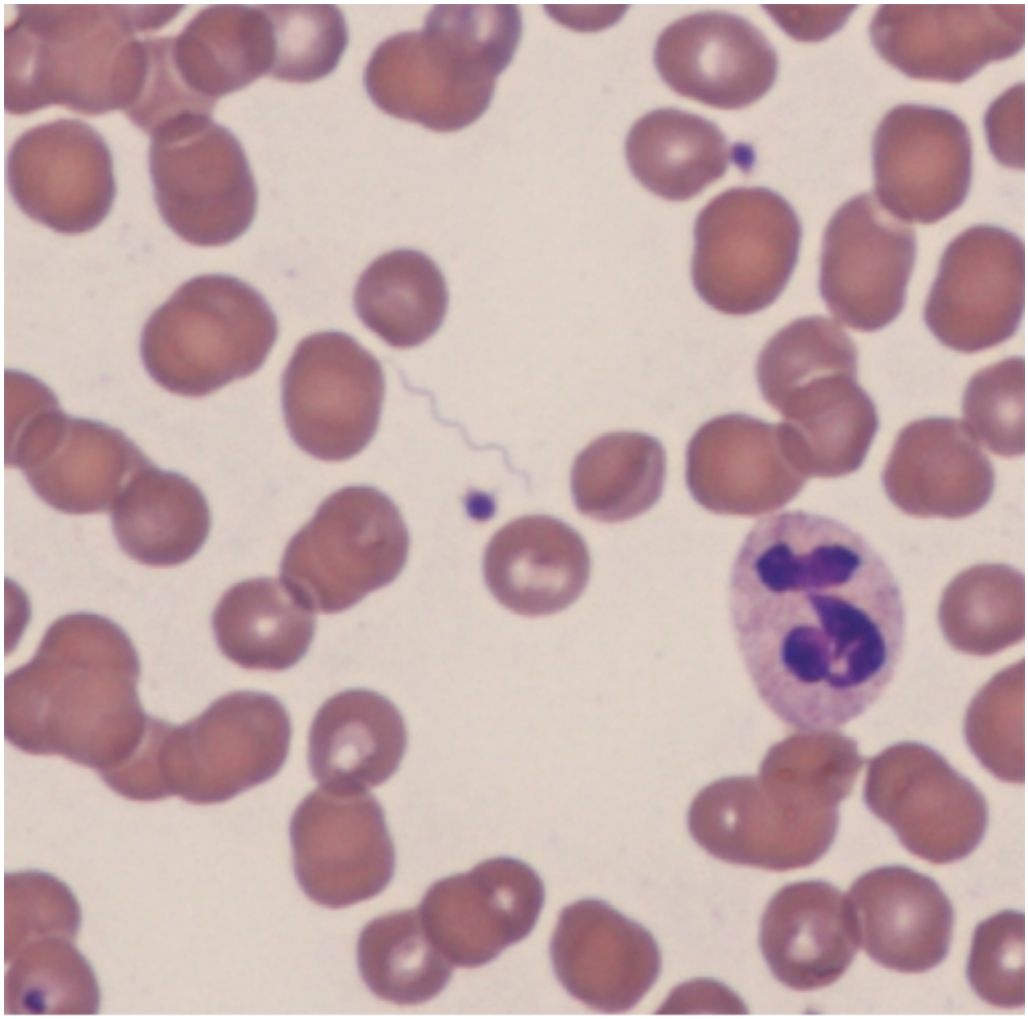
DIAGNOSIS

Relapsing fever should be considered in a patient with the characteristic fever pattern and a history of recent exposure—i.e., within 1–2 weeks before illness onset—to body lice, soft-bodied ticks, or *Ixodes* species hard-bodied ticks in geographic areas with documented current or past transmission. Because of the longevity of the ticks and the transovarial transmission of the pathogen in the ticks, a case of relapsing fever may be diagnosed many years after the last case reported in that locale.

The bedrock for laboratory diagnosis remains what it has been for a century: direct detection of the spirochetes by microscopy of the blood. Manual differential counts of white blood cells by Wright or Giemsa stain usually reveal spirochetes in thin blood smears if their concentration is $\geq 10^5$/mL and several oil-immersion fields are examined **(Fig. 180-2)**. The preferred time to obtain a blood specimen is between the fever's onset and its peak. Lower concentrations of spirochetes may be revealed by a thick blood smear that is either directly stained with acridine orange and then examined by fluorescence microscopy or treated with 0.5% acetic acid before Giemsa or Wright staining. An alternative is a wet mount of anticoagulated blood mixed with saline and examined by phase-contrast or dark-field microscopy for motile spirochetes.

FIGURE 180-2

Photomicrograph of tick-borne relapsing fever spirochete (*Borrelia turicatae*) in a Wright-Giemsa-stained thin blood smear. Included in the figure are a polymorphonuclear leukocyte and two platelets.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Polymerase chain reaction (PCR) and similar DNA amplification procedures are increasingly used for examination of blood or CSF. PCR may reveal spirochetes between febrile episodes, since there are already escape variants in the population when the first wave of bacteria is neutralized.

Culture of blood or CSF in Barbour-Stoenner-Kelly broth medium is an option for isolation of *Borrelia* species. However, few laboratories offer this service. An alternative for tick-borne *Borrelia* species, but not *B. recurrentis*, is inoculation of blood or CSF into immunodeficient mice and examination of the animal's blood after a few days.

Options for serologic confirmation of infection are limited. The few assays that are available commercially or in reference laboratories are based on whole cells of a single *Borrelia* species. These assays may not contain the antigens to which the patient is mainly responding or may yield false-positive results due to antibodies to cross-reactive antigens of related bacteria, including *B. burgdorferi*. The most promising assay is based on GlpQ, a protein antigen of all relapsing fever *Borrelia* species (including *B. miyamotoi*) but not of any Lyme disease species.

DIFFERENTIAL DIAGNOSIS

Depending on the patient's history of residential, occupational, travel, and recreational exposures, the differential diagnosis of relapsing fever includes one or more of the following infections that feature either periodicity in the fever pattern or an extended single febrile period with nonspecific constitutional symptoms: Colorado tick fever (which, along with dengue, can have a "saddleback" fever course); Rocky Mountain spotted fever, ehrlichiosis, anaplasmosis, and other rickettsial diseases; tick-borne arbovirus infection; rat bite fever; and babesiosis in North America, Europe, Russia, and northeastern Asia. Elsewhere in the Americas and Asia and in most of Africa, malaria, typhoid fever, typhus and other rickettsioses, dengue, brucellosis, and leptospirosis may also be considered. Co-infections—malaria, typhus, typhoid, or Lyme disease—may complicate relapsing fever.

TREATMENT

TREATMENT

Relapsing Fever

Penicillins and tetracyclines have been the antibiotics of choice for relapsing fever for several decades. Erythromycin has been a long-standing second choice. There is no evidence of acquired resistance to these antibiotics. *Borrelia* species are also susceptible to most cephalosporins and chloramphenicol, but there is less clinical experience with these drugs. Borreliae are relatively resistant to rifampin, sulfonamides, fluoroquinolones, and aminoglycosides. Spirochetes are no longer detectable in the blood within a few hours after the first dose of an effective antibiotic.

A single dose of antibiotic is usually sufficient for the treatment of LBRF **(Fig. 180-3)**. The recurrence rate after antibiotic treatment is ≤5%. For adults, a single dose of oral tetracycline (500 mg), oral doxycycline (200 mg), or intramuscular penicillin G procaine (400,000–800,000 units) is effective. The corresponding doses for children are oral tetracycline at 12.5 mg/kg, oral doxycycline at 5 mg/kg, and intramuscular penicillin G procaine at 200,000–400,000 units. When an adult patient is stuporous or nauseated, the intravenous dose of tetracycline is 250–500 mg. Tetracycline is contraindicated in pregnant and nursing women and in children <9 years old; for individuals in these groups who are allergic to penicillin, oral erythromycin (500 mg for adults and 12.5 mg/kg for children) is an alternative. Tetracycline is marginally superior to penicillin G in terms of time to fever clearance and relapse rate.

The accumulated anecdotal reports on TBRF therapy indicate a recurrence rate of ≥20% after single-dose treatment. This high rate of recurrence plausibly is due to the greater propensity of tick-borne species than of *B. recurrentis* to invade the CNS, from which they can reinvade the bloodstream after antibiotic levels decline. Accordingly, multiple antibiotic doses are recommended. The preferred treatment for adults is a 10-day course of tetracycline (500 mg or 12.5 mg/kg orally every 6 h) or doxycycline (100 mg twice daily). When tetracyclines are contraindicated, the alternative is erythromycin (500 mg or 12.5 mg/kg orally every 6 h) for 10 days. If a β-lactam antibiotic is given, it is preferably administered intravenously rather than orally, especially if CNS involvement is confirmed or suspected. For adults, the regimen is penicillin G (5 million units IV every 6 h) or ceftriaxone (2 g IV daily) for 10–14 days. Under conditions of limited resources and when

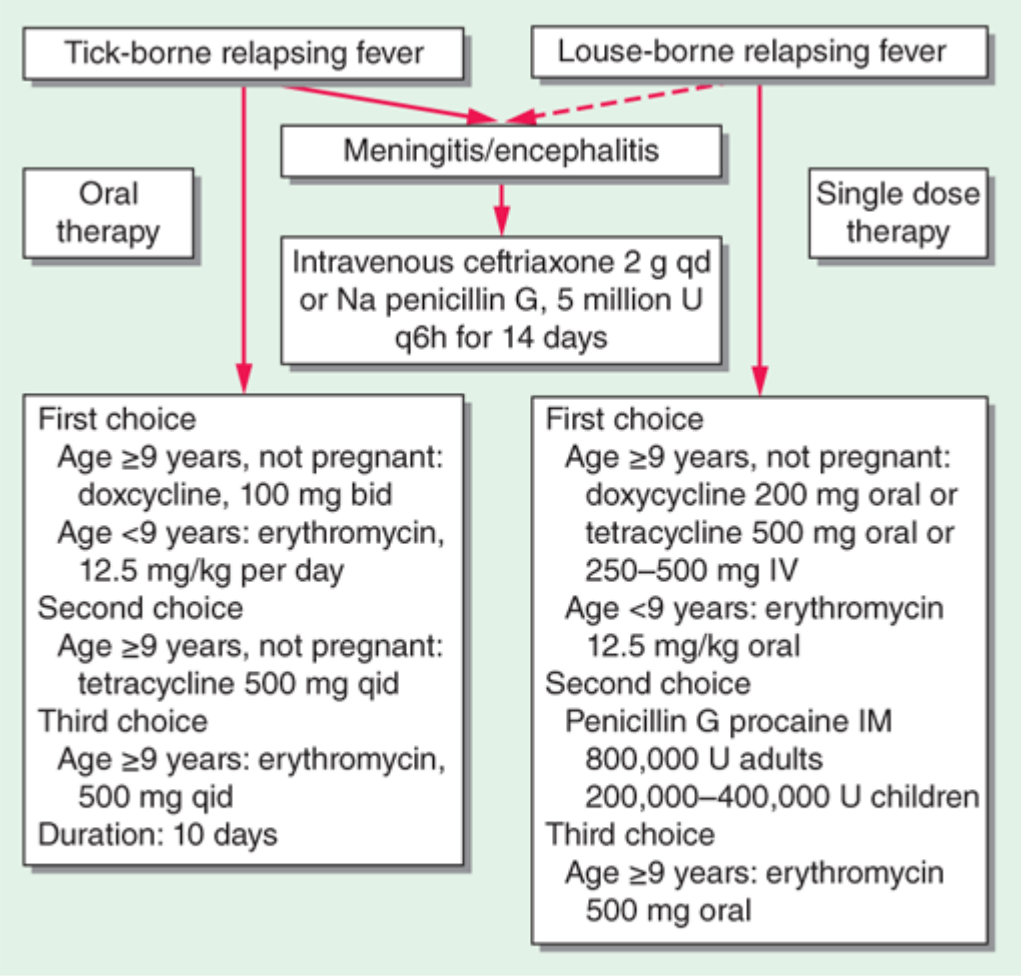
CNS involvement is not suspected, an oral penicillin in divided doses (e.g., penicillin V potassium or penicillin VK at 500 mg or 12.5 mg/kg every 8 h) for 10 days is used.

Experience with the treatment of *B. miyamotoi* infection is limited, but this organism likely has the same antibiotic susceptibilities as other *Borrelia* species. Until more is known about treatment efficacy, therapy for *B. miyamotoi* infection can follow the guidelines for Lyme disease—including parenteral therapy for CNS involvement—because it may be difficult to rule out co-infection.

The Jarisch-Herxheimer reaction during treatment of relapsing fever can be severe or even fatal if precautions are not in place for close monitoring and provision of cardiovascular and volume support as needed. Rigors, fever, and hypotension occur within 2–3 h of initiation of antibiotic treatment. The incidence of this reaction is ~80% in LBRF and ~50% in TBRF. Both penicillin and tetracycline can elicit the Jarisch-Herxheimer reaction.

FIGURE 180-3

Algorithm for treatment of relapsing fever. If it is not known whether the patient has tick-borne or louse-borne relapsing fever, the patient should be treated for the tick-borne form. The *dashed line* indicates that central nervous system invasion in louse-borne relapsing fever is uncommon.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PROGNOSIS

The mortality rates for untreated LBRF and TBRF are in the ranges of 10–70% and 4–10%, respectively, and are largely determined by coexisting conditions, such as malnutrition, dehydration, or malaria. Death from untreated relapsing fever is most common during the first fever episode. With prompt antibiotic treatment, the mortality rate is 2–5% for LBRF and <2% for TBRF. Features associated with a poor prognosis include concurrence with malaria, typhus, or typhoid; pregnancy; stupor or coma on admission; diffuse bleeding; poor liver function; myocarditis; and bronchopneumonia. The mortality rate from the Jarisch-Herxheimer reaction in LBRF, in the absence of adequate monitoring and resuscitation measures, is ~5%. Some patients have survived the crisis or the reaction only to die suddenly either later that day or on the next day. Relapsing fever during pregnancy frequently leads to abortion or stillbirth, but congenital malformations have not been reported. Although it is possible that spirochetes may persist in the CNS or other sequestered sites after bacteremia has resolved, chronic disability from a persistent infection has not been attributed to relapsing fever, including *B. miyamotoi* infection. Partial immunity against reinfection seems to develop in residents of endemic areas.

PREVENTION

There is no vaccine for either LBRF or TBRF. Reduction of exposure to lice and ticks is the key strategy for prevention. LBRF can be prevented through improved personal hygiene, reduction of crowding, better access to washing facilities, and selected use of pesticides. Infested clothing is an important factor in maintaining body lice. The risk of TBRF can be reduced by construction of houses with concrete or sealed plank floors and without thatched roofs or mud walls. Rustic cabins pose a particular risk in North America when rodents nest in the roof or beneath the house or porch. Buildings infested with *Ornithodoros* ticks can be treated with pesticides and then rodent-proofed. If residing in a high-risk environment, individuals should not sleep on the floor, and beds should be moved away from the wall. With an exposure to TBRF, postexposure treatment with doxycycline (200 mg on day 1 followed by 100 mg/d for 4 days) was efficacious in preventing infection in a placebo-controlled trial.

FURTHER READING

Barbour AG: Relapsing fever, in *Tick-Borne Diseases of Humans*, JL Goodman et al (eds). Washington, DC, ASM Press, 2005, p 268.
CrossRef

Cadavid D, Barbour AG: Neuroborreliosis during relapsing fever: Review of the clinical manifestations, pathology, and treatment of infections in humans and experimental animals. Clin Infect Dis 26:151, 1998.

Colin de Verdiere N et al: Tickborne relapsing fever caused by *Borrelia persica*, Uzbekistan and Tajikistan. Emerg Infect Dis 17:1325, 2011.

Dworkin MS et al: Tick-borne relapsing fever in North America. Med Clin North Am 86:417, 2002.

Guerrier G, Doherty T: Comparison of antibiotic regimens for treating louse-borne relapsing fever: A meta-analysis. Trans R Soc Trop Med Hyg 105:483, 2011.

Krause PJ et al: *Borrelia miyamotoi* infection in nature and in humans. Clin Microbiol Infect 21:631, 2015.

Lucchini A et al: Louseborne relapsing fever among East African refugees, Italy, 2015. Emerg Infect Dis 22:298, 2016.

Moran-Gilad J et al: Postexposure prophylaxis of tick-borne relapsing fever: Lessons learned from recent outbreaks in Israel. Vector Borne Zoonotic Dis 13:791, 2013.

Nordstrand A et al: Tickborne relapsing fever diagnosis obscured by malaria, Togo. Emerg Infect Dis 3:117, 2007.

Southern P, Sanford J: Relapsing fever: A clinical and microbiological review. Medicine (Baltimore) 48:129, 1969.

Chapter 181: Lyme Borreliosis

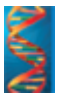
Allen C. Steere

DEFINITION

Lyme borreliosis is caused by a spirochete, *Borrelia burgdorferi sensu lato*, that is transmitted by ticks of the *Ixodes ricinus* complex. The infection usually begins with a characteristic expanding skin lesion, erythema migrans (EM; stage 1, localized infection). After several days or weeks, the spirochete may spread to many different sites (stage 2, disseminated infection). Possible manifestations of disseminated infection include secondary annular skin lesions, meningitis, cranial neuritis, radiculoneuritis, peripheral neuritis, carditis, atrioventricular nodal block, or migratory musculoskeletal pain. Months or years later (usually after periods of latent infection), intermittent or persistent arthritis, chronic encephalopathy or polyneuropathy, or acrodermatitis may develop (stage 3, persistent infection). Most patients experience early symptoms of the illness during the summer, but the infection may not become symptomatic until it progresses to stage 2 or 3.

Lyme disease was recognized as a separate entity in 1976 because of a geographic cluster of children in Lyme, Connecticut, who were thought to have juvenile rheumatoid arthritis. It became apparent that Lyme disease was a multisystemic illness that affected primarily the skin, nervous system, heart, and joints. Epidemiologic studies of patients with EM implicated certain *Ixodes* ticks as vectors of the disease. Early in the twentieth century, EM had been described in Europe and attributed to *I. ricinus* tick bites. In 1982, a previously unrecognized spirochete, now called *Borrelia burgdorferi*, was recovered from *Ixodes scapularis* ticks and then from patients with Lyme disease. The entity is now called Lyme disease or Lyme borreliosis.

ETIOLOGIC AGENT



B. burgdorferi, the causative agent of Lyme disease, is a fastidious microaerophilic bacterium. The spirochete's genome is quite small (~1.5 Mb) and consists of a highly unusual linear chromosome of 950 kb as well as 17–21 linear and circular plasmids. The most remarkable aspect of the *B. burgdorferi* genome is that there are sequences for more than 100 known or predicted lipoproteins—a larger number than in any other organism. The spirochete has few proteins with biosynthetic activity and depends on its host for most of its nutritional requirements. It has no sequences for recognizable toxins.



Currently, 20 closely related borrelial species are collectively referred to as *B. burgdorferi sensu lato* (i.e., "*B. burgdorferi* in the general sense"). The human infection Lyme borreliosis is caused primarily by three

pathogenic genospecies: *B. burgdorferi sensu stricto* ("*B. burgdorferi* in the strict sense," hereafter referred to simply as *B. burgdorferi*), *Borrelia garinii*, and *Borrelia afzelii*. *B. burgdorferi* is the sole cause of the infection in the United States; all three genospecies are found in Europe, and the latter two species occur in Asia.

Strains of *B. burgdorferi* have been subdivided according to several typing schemes: one based on sequence variation of outer-surface protein C (OspC), a second based on differences in the 16S–23S rRNA intergenic spacer region (RST or IGS), and a third called *multilocus sequence typing*. From these typing systems, it is apparent that strains of *B. burgdorferi* differ in pathogenicity. OspC type A (RST1) strains seem to be particularly virulent and may have played a role in the emergence of Lyme disease in epidemic form in the northeastern United States in the late twentieth century.

EPIDEMIOLOGY

The 20 known genospecies of *B. burgdorferi sensu lato* live in nature in enzootic cycles involving 14 species of ticks that are part of the *I. ricinus* complex. *I. scapularis* **(Fig. 452-1)** is the principal vector in the eastern United States from Maine to Georgia and in the midwestern states of Wisconsin, Minnesota, and Michigan. *I. pacificus* is the vector in the western states of California and Oregon. The disease is acquired throughout Europe (from Great Britain to Scandinavia to European Russia), where *I. ricinus* is the vector, and in Asian Russia, China, and Japan, where *I. persulcatus* is the vector. These ticks may transmit other agents as well. In the United States, *I. scapularis* also transmits *Babesia microti*; *Anaplasma phagocytophilum*; *Ehrlichia* species Wisconsin; *Borrelia miyamotoi*; *Borrelia mayonii*; and, in rare instances, Powassan encephalitis virus (the deer tick virus) (see "Differential Diagnosis," below). In Europe and Asia, *I. ricinus* and *I. persulcatus* also transmit tick-borne encephalitis virus.

Ticks of the *I. ricinus* complex have larval, nymphal, and adult stages. They require a blood meal at each stage. The risk of infection in a given area depends largely on the density of these ticks as well as their feeding habits and animal hosts, which have evolved differently in different locations. For *I. scapularis* in the northeastern United States, the white-footed mouse and certain other rodents are the preferred hosts of the immature larvae and nymphs. It is critical that both of the tick's immature stages feed on the same host because the life cycle of the spirochete depends on horizontal transmission: in early summer from infected nymphs to mice and in late summer from infected mice to larvae, which then molt to become the infected nymphs that will begin the cycle again the following year. It is the tiny nymphal tick that is primarily responsible for transmission of the disease to humans, which peaks during the early summer months. White-tailed deer, which are not involved in the life cycle of the spirochete, are the preferred host for the adult stage of *I. scapularis* and seem to be critical to the tick's survival.

Lyme disease is now the most common vector-borne infection in the United States and Europe. Since surveillance was begun by the Centers for Disease Control and Prevention (CDC) in 1982, the number of cases in the United States has increased dramatically. More than 30,000 new cases are now reported each summer, but the actual number of new cases is probably closer to 300,000 annually. In Europe, reported frequencies of the disease are highest in the middle of the continent and in Scandinavia.

PATHOGENESIS AND IMMUNITY

To maintain its complex enzootic cycle, *B. burgdorferi* must adapt to two markedly different environments: the tick and the mammalian host. The spirochete expresses outer-surface protein A (OspA) in the midgut of the tick, whereas OspC is upregulated as the organism travels to the tick's salivary gland. There, OspC binds a tick salivary-gland protein (Salp15), which is required for infection of the mammalian host. The tick usually must be attached for at least 24 h for transmission of *B. burgdorferi*.

After injection into the human skin, the spirochete downregulates OspC and upregulates the VlsE lipoprotein. This protein undergoes extensive antigenic variation, which is necessary for spirochetal survival. After several days to weeks, *B. burgdorferi* may migrate outward in the skin, producing EM, and may spread hematogenously or in the lymph to other organs. The only known virulence factors of *B. burgdorferi* are surface proteins that allow the spirochete to attach to mammalian proteins, integrins, glycosaminoglycans, or glycoproteins. For example, spread through the skin and other tissue matrices may be facilitated by the binding of human plasminogen and its activators to the surface of the spirochete. Some *Borrelia* strains bind complement regulator–acquiring surface proteins (FHL-1/reconectin, or factor H), which help to protect spirochetes from complement-mediated lysis. Dissemination of the organism in the blood is facilitated by binding to the fibrinogen receptor ($\alpha_{IIb}\beta_3$) on activated platelets and the vitronectin receptor ($\alpha_v\beta_3$) on endothelial cells. As the name indicates, spirochetal decorin-binding proteins A and B bind decorin, a glycosaminoglycan on collagen fibrils; this binding may explain why the organism is commonly aligned with collagen fibrils in the extracellular matrix in the heart, nervous system, or joints.

To control and eradicate *B. burgdorferi*, the host mounts both innate and adaptive immune responses, resulting in macrophage- and antibody-mediated killing of the spirochete. As part of the innate immune response, complement may lyse the spirochete in the skin. Cells at affected sites release potent proinflammatory cytokines, including interleukin 6, tumor necrosis factor α, interleukin 1β, and interferon γ. Patients who are homozygous for a Toll-like receptor 1 polymorphism (1805GG), particularly when infected with highly inflammatory *B. burgdorferi* RST1 strains, have exceptionally high levels of proinflammatory cytokines. The purpose of the adaptive immune response appears to be the production of specific antibodies, which opsonize the organism—a step necessary for optimal spirochetal killing. Studies with protein arrays expressing ~1200 *B. burgdorferi* proteins detected antibody responses to a total of 120 spirochetal proteins (particularly outer-surface lipoproteins) in a population of patients with Lyme arthritis. Histologic examination of all affected tissues reveals an infiltration of lymphocytes, macrophages, and plasma cells with some degree of vascular damage, including mild vasculitis or hypervascular occlusion. These findings suggest that the spirochete may have been present in or around blood vessels.

In enzootic infection, *B. burgdorferi* spirochetes must survive this immune assault only during the summer months before returning to larval ticks to begin the cycle again the following year. In contrast, infection of humans is a dead-end event for the spirochete. Within several weeks or months, innate and adaptive immune mechanisms—even without antibiotic treatment—control widely disseminated infection, and generalized systemic symptoms wane. However, without antibiotic therapy, spirochetes may survive in

localized niches for several more years. For example, *B. burgdorferi* infection in the United States may cause persistent arthritis or, in rare cases, subtle encephalopathy or polyneuropathy. Thus, immune mechanisms seem to succeed eventually in the near or total eradication of *B. burgdorferi* from selected niches, including the joints or nervous system, and symptoms resolve in most patients.

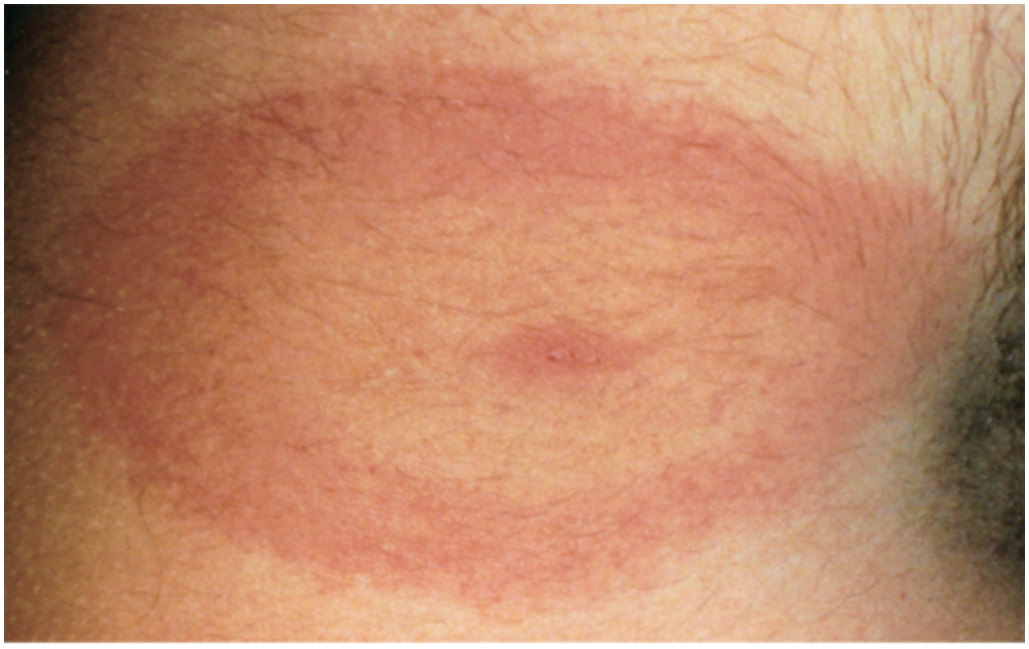
CLINICAL MANIFESTATIONS

Early Infection: Stage 1 (Localized Infection)

Because of the small size of nymphal ixodid ticks, most patients do not remember the preceding tick bite. After an incubation period of 3–32 days, EM usually begins as a red macule or papule at the site of the tick bite that expands slowly to form a large annular lesion **(Fig. 181-1)**. As the lesion increases in size, it often develops a bright red outer border and partial central clearing. The center of the lesion sometimes becomes intensely erythematous and indurated, vesicular, or necrotic. In other instances, the expanding lesion remains an even, intense red; several red rings are found within an outside ring; or the central area turns blue before the lesion clears. Although EM can be located anywhere, the thigh, groin, and axilla are particularly common sites. The lesion is warm but not often painful. Approximately 20% of patients do not exhibit this characteristic skin manifestation.

FIGURE 181-1

A classic erythema migrans lesion (9 cm in diameter) is shown near the right axilla. The lesion has partial central clearing, a bright red outer border, and a target center. *(Courtesy of Vijay K. Sikand, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Early Infection: Stage 2 (Disseminated Infection)

In cases in the United States, *B. burgdorferi* often spreads hematogenously to many sites within days or weeks after the onset of EM. In these cases, patients may develop secondary annular skin lesions similar in appearance to the initial lesion. Skin involvement is commonly accompanied by severe headache, mild stiffness of the neck, fever, chills, migratory musculoskeletal pain, arthralgias, and profound malaise and fatigue. Less common manifestations include generalized lymphadenopathy or splenomegaly, hepatitis, sore throat, nonproductive cough, conjunctivitis, iritis, or testicular swelling. Except for fatigue and lethargy, which are often constant, the early signs and symptoms of Lyme disease are typically intermittent and changing. Even in untreated patients, the early symptoms usually become less severe or disappear within several weeks. In ~15% of patients, the infection presents with these nonspecific systemic symptoms.

Symptoms suggestive of meningeal irritation may develop early in Lyme disease when EM is present but usually are not associated with cerebrospinal fluid (CSF) pleocytosis or an objective neurologic deficit. After several weeks or months, ~15% of untreated patients develop frank neurologic abnormalities, including meningitis, subtle encephalitic signs, cranial neuritis (including bilateral facial palsy), motor or sensory radiculoneuropathy, peripheral neuropathy, mononeuritis multiplex, cerebellar ataxia, or myelitis—alone or in various combinations. In children, the optic nerve may be affected because of inflammation or increased intracranial pressure, and these effects may lead to blindness. In the United States, the usual pattern consists of fluctuating symptoms of meningitis accompanied by facial palsy and peripheral radiculoneuropathy. Lymphocytic pleocytosis (~100 cells/µL) is found in CSF, often along with elevated protein levels and normal or slightly low glucose concentrations. In Europe and Asia, the first neurologic sign is characteristically radicular pain, which is followed by the development of CSF pleocytosis (meningopolyneuritis or *Bannwarth's syndrome*); meningeal or encephalitic signs are frequently absent. These early neurologic abnormalities usually resolve completely within months, but in rare cases chronic neurologic disease may occur later.

Within several weeks after the onset of illness, ~8% of patients develop cardiac involvement. The most common abnormality is a fluctuating degree of atrioventricular block (first-degree, Wenckebach, or complete heart block). Some patients have more diffuse cardiac involvement, including electrocardiographic changes indicative of acute myopericarditis, left ventricular dysfunction evident on radionuclide scans, or (in rare cases) cardiomegaly or fatal pancarditis. Cardiac involvement lasts for only a few weeks in most patients but may recur in untreated patients. Chronic cardiomyopathy caused by *B. burgdorferi* has been reported in Europe.

During this stage, musculoskeletal pain is common. The typical pattern consists of migratory pain in joints, tendons, bursae, muscles, or bones (usually without joint swelling) lasting for hours or days and affecting one or two locations at a time.

Late Infection: Stage 3 (Persistent Infection)

Months after the onset of infection, ~60% of patients in the United States who have received no antibiotic treatment develop frank arthritis. The typical pattern comprises intermittent attacks of oligoarticular arthritis in large joints (especially the knees), lasting for weeks or months in a given joint. A few small joints

or periarticular sites also may be affected, primarily during early attacks. The number of patients who continue to have recurrent attacks decreases each year. However, in a small percentage of cases, involvement of large joints—usually one or both knees—is persistent and may lead to erosion of cartilage and bone.

White cell counts in joint fluid range from 500 to 110,000/µL (average, 25,000/µL); most of these cells are polymorphonuclear leukocytes. Tests for rheumatoid factor or antinuclear antibodies usually give negative results. Examination of synovial biopsy samples reveals fibrin deposits, villous hypertrophy, vascular proliferation, microangiopathic lesions, and a heavy infiltration of lymphocytes and plasma cells.

Although most patients with Lyme arthritis respond well to antibiotic therapy, a small percentage in the northeastern United States have persistent (*postinfectious, antibiotic-refractory*) arthritis for months or even for several years after receiving oral and IV antibiotic therapy for 2 or 3 months. Although more often these patients are initially infected with RST1 strains of *B. burgdorferi*, this complication is not thought to result from persistent infection. Results of culture and polymerase chain reaction (PCR) for *B. burgdorferi* in synovial tissue obtained in the postantibiotic period have been uniformly negative. Rather, infection-induced autoimmunity, retained spirochetal antigens, or both may play a role in this outcome. Antibiotic-refractory arthritis is associated with a higher frequency of certain class II major histocompatibility complex molecules (particularly HLA-DRBI*0401 or *0101 molecules); the Toll-like receptor 1 polymorphism 1805GG, which leads to exceptionally high levels of cytokines and chemokines in affected joints; and low frequencies of FoxP3+ T regulatory cells in synovial fluid, which correlate with longer posttreatment durations of arthritis. Four autoantigens that are targets of T and B cell responses in patients with Lyme disease, particularly those with antibiotic-refractory arthritis, have now been identified: endothelial cell growth factor, matrix metalloproteinase-10, apolipoprotein B-100, and annexin A2. Additional, yet-to-be identified autoantigens may also have a role in antibiotic-refractory arthritis.

Although rare, chronic neurologic involvement also may become apparent from months to several years after the onset of infection, sometimes after long periods of latent infection. The most common form of chronic central nervous system involvement is subtle encephalopathy affecting memory, mood, or sleep, and the most common form of peripheral neuropathy is an axonal polyneuropathy manifested as either distal paresthesia or spinal radicular pain. Patients with encephalopathy frequently have evidence of memory impairment in neuropsychological tests and abnormal results in CSF analyses. In cases of polyneuropathy, electromyography generally shows extensive abnormalities of proximal and distal nerve segments. Encephalomyelitis or leukoencephalitis, a rare manifestation of Lyme borreliosis associated primarily with *B. garinii* infection in Europe, is a severe neurologic disorder that may include spastic paraparesis, upper motor-neuron bladder dysfunction, and, rarely, lesions in the periventricular white matter.

 *Acrodermatitis chronica atrophicans*, the late skin manifestation of Lyme borreliosis, has been associated primarily with *B. afzelii* infection in Europe and Asia. It has been observed especially often in elderly women. The skin lesions, which are usually found on the acral surface of an arm or leg, begin insidiously with reddish-violaceous discoloration; they become sclerotic or atrophic over a period of years.

The basic patterns of Lyme borreliosis are similar worldwide, but there are regional variations, primarily between the illness found in North America, which is caused exclusively by *B. burgdorferi*, and that found in Europe, which is caused primarily by *B. afzelii* and *B. garinii*. With each of the *Borrelia* species, the infection usually begins with EM. However, *B. burgdorferi* strains in the eastern United States often disseminate widely; they are particularly arthritogenic, and they may cause antibiotic-refractory arthritis. *B. garinii* typically disseminates less widely, but it is especially neurotropic and may cause borrelial encephalomyelitis. *B. afzelii* often infects only the skin but may persist in that site, where it may cause several different dermatoborrelioses, including acrodermatitis chronica atrophicans.

Post-Lyme Syndrome (Chronic Lyme Disease)

Despite resolution of the objective manifestations of the infection with antibiotic therapy, ~10% of patients (although the reported percentages vary widely) continue to have subjective pain, neurocognitive manifestations, or fatigue symptoms. These symptoms usually improve and resolve within months but may last for years. At the far end of the spectrum, the symptoms may be similar to or indistinguishable from chronic fatigue syndrome **(Chap. 442)** and fibromyalgia **(Chap. 366)**. Compared with symptoms of active Lyme disease, post-Lyme symptoms tend to be more generalized or disabling. They include marked fatigue, severe headache, diffuse musculoskeletal pain, multiple symmetric tender points in characteristic locations, pain and stiffness in many joints, diffuse paresthesias, difficulty with concentration, and sleep disturbances. Patients with this condition lack evidence of joint inflammation, have normal neurologic test results, and may exhibit anxiety and depression. In contrast, late manifestations of Lyme disease, including arthritis, encephalopathy, and neuropathy, are usually associated with minimal systemic symptoms. Currently, no evidence indicates that persistent subjective symptoms after recommended courses of antibiotic therapy are caused by active infection.

DIAGNOSIS

The culture of *B. burgdorferi* in Barbour-Stoenner-Kelly (BSK) medium permits definitive diagnosis, but this method has been used primarily in research studies. Moreover, with a few exceptions, positive cultures have been obtained only early in the illness—particularly from biopsy samples of EM skin lesions, less often from plasma samples, and occasionally from CSF samples. Later in the infection, PCR is greatly superior to culture for the detection of *B. burgdorferi* DNA in joint fluid; this is the major use for PCR testing in Lyme disease. However, because *B. burgdorferi* DNA may persist for at least weeks after spirochetal killing with antibiotics, detection of spirochetal DNA in joint fluid is not an accurate test of active joint infection in Lyme disease and cannot be used reliably to determine the adequacy of antibiotic therapy. The sensitivity of PCR determinations in CSF from patients with neuroborreliosis has been much lower than that in joint fluid. With current methods, there seems to be little if any role for PCR in the detection of *B. burgdorferi* DNA in blood or urine samples, although this is an area of active research. A potential drawback is that PCR must be carefully controlled to prevent contamination.

Because of the problems associated with direct detection of *B. burgdorferi*, Lyme disease is usually diagnosed by the recognition of a characteristic clinical picture accompanied by serologic confirmation. Although serologic testing may yield negative results during the first several weeks of infection, almost all patients have a positive antibody response to *B. burgdorferi* after that time when a two-test approach of enzyme-linked immunosorbent assay (ELISA) and western blot is used. The limitation of serologic tests is that they do not clearly distinguish between active and inactive infection. After antibiotic therapy, the amount of antibody declines but the results of western blot, a nonquantitative test, do not change much. Thus, patients with previous Lyme disease—particularly in cases progressing to late stages—often remain seropositive for years, even after adequate antibiotic therapy. In addition, ~10% of patients are seropositive because of asymptomatic infection. If individuals with past or asymptomatic *B. burgdorferi* infection subsequently develop another illness, the positive serologic test for Lyme disease may cause diagnostic confusion. According to an algorithm published by the American College of Physicians **(Table 181-1)**, serologic testing for Lyme disease is recommended only for patients with at least an intermediate pretest probability of Lyme disease, such as those with oligoarticular arthritis. It should not be used as a screening procedure in patients with pain or fatigue syndromes. In such patients, the probability of a false-positive serologic result is higher than that of a true-positive result.

TABLE 181-1

Algorithm for Testing for and Treating Lyme Disease

Pretest Probability	Example	Recommendation
High	Patients with erythema migrans	Empirical antibiotic treatment without serologic testing
Intermediate	Patients with oligoarticular arthritis	Serologic testing and antibiotic treatment if test results are positive
Low	Patients with nonspecific symptoms (myalgias, arthralgias, fatigue)	Neither serologic testing nor antibiotic treatment

Source: Adapted from the recommendations of the American College of Physicians (G Nichol et al: Ann Intern Med 128:37, 1998).

For serologic analysis of Lyme disease in the United States, the CDC recommends a two-step approach in which samples are first tested by ELISA, and equivocal or positive results are then tested by western blot. During the first weeks of infection, both IgM and IgG responses to the spirochete should be determined, preferably in both acute- and convalescent-phase serum samples. Approximately 20–30% of patients have a positive response detectable in acute-phase samples (usually only a positive IgM response), whereas ~70–

80% have a positive response during convalescence (2–4 weeks later). After 4–8 weeks of infection (by which time most patients with active Lyme disease have disseminated infection), the sensitivity and specificity of the IgG response to the spirochete are both very high—in the range of 99%—as determined by the two-test approach of ELISA and western blot. At this point and thereafter, a single test (that for IgG) is usually sufficient. In persons with illness of >2 months' duration, a positive IgM test result alone is likely to be false-positive and therefore should not be used to support the diagnosis.

According to current criteria adopted by the CDC, an IgM western blot is considered positive if two of the following three bands are present: 23, 39, and 41 kDa. However, the combination of two such bands may still represent a false-positive result. Misuse or misinterpretation of IgM blots has been a factor in the incorrect diagnosis of Lyme disease in patients with other illnesses. An IgG blot is considered positive if 5 of the following 10 bands are present: 18, 23, 28, 30, 39, 41, 45, 58, 66, and 93 kDa. In European cases, no single set of criteria for the interpretation of immunoblots results in high levels of sensitivity and specificity in all countries.

A promising new methodology, particularly for the determination of antibody responses during the first weeks of infection, is a two-test approach using two ELISAs rather than ELISA and western blot. One such method employs a whole–*B. burgdorferi* sonicate ELISA followed by a VlsE C6 peptide IgG ELISA. This approach, which gives simply a positive or a negative result, increases sensitivity during the first several weeks of infection without compromising specificity. For more complex cases, it is still valuable to determine antibody specificities to multiple spirochetal proteins, as is done with western blots. More recently, line immunoblots or other multiplexed antibody platforms have been developed as substitutes for western blots. These assays allow more objective interpretation, and some platforms can provide quantitative data about antibody responses to many spirochetal proteins. After successful antibiotic treatment, antibody titers decline slowly, but responses (including that to the VlsE C6 peptide) may persist for years. Moreover, not only the IgG but also the IgM response may persist for years after therapy. Therefore, even a positive IgM response cannot be interpreted as confirmation of recent infection or reinfection unless the clinical picture is appropriate.

DIFFERENTIAL DIAGNOSIS

Classic EM is a slowly expanding erythema, often with partial central clearing. If the lesion expands little, it may represent the red papule of an uninfected tick bite. If the lesion expands rapidly, it may represent cellulitis (e.g., streptococcal cellulitis) or an allergic reaction, perhaps to tick saliva. Patients with secondary annular lesions may be thought to have erythema multiforme, but neither the development of blistering mucosal lesions nor the involvement of the palms or soles is a feature of *B. burgdorferi* infection. In the eastern United States, an EM-like skin lesion, sometimes with mild systemic symptoms, may be associated with *Amblyomma americanum* tick bites. However, the cause of this Southern tick-associated rash illness (STARI) has not yet been identified. This tick may also transmit *Ehrlichia chaffeensis*, a rickettsial agent (Chap. 182).

As stated above, *I. scapularis* ticks in the United States may transmit not only *B. burgdorferi* but also *B. microti*, the red blood cell parasite causing babesiosis (Chap. 220); *A. phagocytophilum*, the agent of human granulocytotropic anaplasmosis (Chap. 182); *B. miyamotoi*, a relapsing fever spirochete (Chap. 180); *B. mayonii* and *Ehrlichia* species Wisconsin, newly recognized species that occur in the upper midwestern United States; or (rarely) Powassan encephalitis virus (the deer tick virus, which is closely related to European tick-borne encephalitis virus) (Chap. 204). Although babesiosis and anaplasmosis are most often asymptomatic, infection with any of these agents may cause nonspecific systemic symptoms, particularly in the young or the elderly, and co-infected patients may have more severe or persistent symptoms than patients infected with a single agent. Standard blood counts may yield clues regarding the presence of co-infection with *Anaplasma* or *Babesia*. Anaplasmosis may cause leukopenia or thrombocytopenia, and babesiosis may cause thrombocytopenia or (in severe cases) hemolytic anemia. IgM serologic responses may confuse the diagnosis. For example, *A. phagocytophilum* may elicit a positive IgM response to *B. burgdorferi*. The frequency of co-infection in different studies has been variable. In one prospective study, 4% of patients with EM had evidence of co-infection.

Facial palsy caused by *B. burgdorferi*, which occurs in the early disseminated phase of the infection (often in July, August, or September), is usually recognized by its association with EM. However, in rare cases, facial palsy without EM may be the presenting manifestation of Lyme disease. In such cases, both the IgM and the IgG responses to the spirochete are usually positive. The most common infectious agents that cause facial palsy are herpes simplex virus type 1 (Bell's palsy; Chap. 187) and varicella-zoster virus (Ramsay Hunt syndrome; Chap. 188).

Later in the infection, oligoarticular Lyme arthritis most resembles peripheral spondyloarthropathy in an adult or the pauciarticular form of juvenile idiopathic arthritis in a child. Patients with Lyme arthritis usually have the strongest IgG antibody responses seen in Lyme borreliosis, with reactivity to many spirochetal proteins.

The most common problem in diagnosis is to mistake Lyme disease for chronic fatigue syndrome (Chap. 442) or fibromyalgia (Chap. 366). This difficulty is compounded by the fact that a small percentage of patients do in fact develop these chronic pain or fatigue syndromes in association with or soon after Lyme disease. Moreover, a counterculture has emerged that ascribes pain and fatigue syndromes to chronic Lyme disease when there is little or no evidence of *B. burgdorferi* infection. In such cases, the term *chronic Lyme disease*, which is equated with chronic *B. burgdorferi* infection, is a misnomer, and the use of prolonged, dangerous, and expensive antibiotic treatment is not warranted.

TREATMENT

TREATMENT

Lyme Borreliosis

ANTIBIOTIC TREATMENT

As outlined in the algorithm in **Fig. 181-2**, the various manifestations of Lyme disease can usually be treated successfully with orally administered antibiotics; the exceptions are severe objective neurologic abnormalities and third-degree atrioventricular heart block, which are generally treated with IV antibiotics, and arthritis that does not respond to oral therapy. For early Lyme disease, doxycycline is effective and can be administered to men and nonpregnant women. An advantage of this regimen is that it is also effective against *A. phagocytophilum, B. miyamotoi*, and *B. mayonii*, which are transmitted by the same tick that transmits the Lyme disease agent. Amoxicillin, cefuroxime axetil, and erythromycin or its congeners are second-, third-, and fourth-choice alternatives, respectively, for the treatment of Lyme disease. In children, amoxicillin is effective (not >2 g/d); in cases of penicillin allergy, cefuroxime axetil or erythromycin may be used. In contrast to second- or third-generation cephalosporin antibiotics, first-generation cephalosporins, such as cephalexin, are not effective. For patients with infection localized to the skin, a 14-day course of therapy is generally sufficient; in contrast, for patients with disseminated infection, a 21-day course is recommended. Approximately 15% of patients experience a Jarisch-Herxheimer-like reaction during the first 24 h of therapy. In multicenter studies, more than 90% of patients whose early Lyme disease was treated with these regimens had satisfactory outcomes. Although some patients reported symptoms after treatment, objective evidence of persistent infection or relapse was rare, and re-treatment was usually unnecessary.

Oral administration of doxycycline or amoxicillin for 30 days is recommended for the initial treatment of Lyme arthritis in patients who do not have concomitant neurologic involvement. Among patients with arthritis who do not respond to oral antibiotics, re-treatment with IV ceftriaxone for 28 days is appropriate. In patients with arthritis in whom joint inflammation persists for months or even several years after both oral and IV antibiotics, treatment with nonsteroidal anti-inflammatory agents, therapy with disease-modifying antirheumatic drugs, or synovectomy may be successful.

In the United States, parenteral antibiotic therapy is usually used for severe objective neurologic abnormalities. Patients with such abnormalities are most commonly treated with IV ceftriaxone for 14–28 days, but IV cefotaxime or IV penicillin G for the same duration also may be effective. In Europe, similar results have been obtained with oral doxycycline and IV antibiotics in the treatment of acute neuroborreliosis. Although systematic trials have not been conducted in the United States, oral doxycycline is now used by some clinicians in this country for the treatment of patients with less severe neurologic abnormalities, such as facial palsy alone or uncomplicated Lyme meningitis. In patients with high-degree atrioventricular block or a PR interval of >0.3 s, IV therapy for at least part of the course and cardiac monitoring are recommended, but the insertion of a permanent pacemaker is not necessary.

It is unclear how and whether asymptomatic infection should be treated, but patients with such infection are often given a course of oral antibiotics. Because maternal–fetal transmission of *B. burgdorferi* seems to occur rarely (if at all), standard therapy for the manifestations of the illness is recommended for pregnant women. Long-term persistence of *B. burgdorferi* has not been documented in any large series of patients after treatment with currently recommended regimens. Although an occasional patient requires a second course of antibiotics, there is no indication for multiple, repeated antibiotic courses in the treatment of Lyme disease.

CHRONIC LYME DISEASE

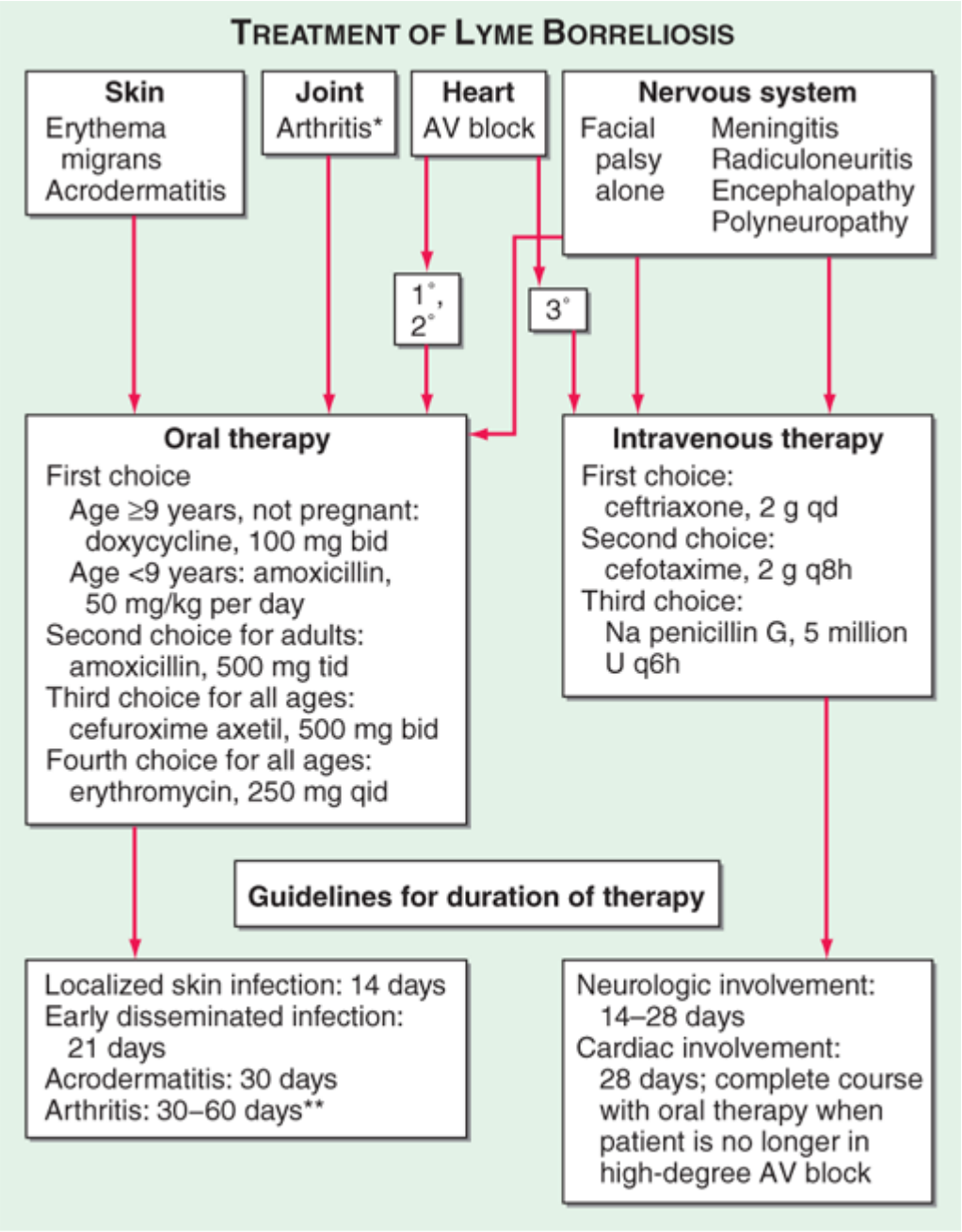
After appropriately treated Lyme disease, a small percentage of patients continue to have subjective symptoms, primarily musculoskeletal pain, neurocognitive difficulties, or fatigue. This *chronic Lyme disease* or *post-Lyme syndrome* is sometimes a disabling condition that is similar to chronic fatigue syndrome or fibromyalgia. Five double-blind, placebo-controlled trials conducted in the United States and Europe have failed to show benefit of further antibiotic therapy in these patients. For example, in a large study, one group of patients with post-Lyme syndrome received IV ceftriaxone for 30 days followed by oral doxycycline for 60 days, while another group received IV and oral placebo preparations for the same durations. No significant differences were found between groups in the numbers of patients reporting that their symptoms had improved, become worse, or stayed the same. Such patients are best treated for the relief of symptoms rather than with prolonged courses of antibiotics.

PROPHYLAXIS AFTER A TICK BITE

The risk of infection with *B. burgdorferi* after a recognized tick bite is so low that antibiotic prophylaxis is not routinely indicated. However, if an attached, engorged *I. scapularis* nymph is found or if follow-up is anticipated to be difficult, a single 200-mg dose of doxycycline, which usually prevents Lyme disease when given within 72 h after the tick bite, may be administered.

FIGURE 181-2

Algorithm for the treatment of the various early or late manifestations of Lyme borreliosis. AV, atrioventricular. *For arthritis, oral therapy should be tried first; if arthritis is unresponsive, IV therapy should be administered. **For Lyme arthritis, IV ceftriaxone (2 g given once a day for 14–28 days) also is effective and is necessary for patients who do not respond to oral therapy. However, compared with oral treatment, this regimen is less convenient to administer, has more side effects, and is more expensive.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PROGNOSIS

The response to treatment is best early in the disease. Later treatment of Lyme borreliosis is still effective, but the period of convalescence may be longer. Eventually, most patients recover with minimal or no residual deficits.

REINFECTION

Reinfection may occur after EM when patients are treated with antimicrobial agents. In such cases, the immune response is not adequate to provide protection from subsequent infection. However, patients who

develop an expanded immune response to the spirochete over a period of months (e.g., those with Lyme arthritis) have protective immunity for a period of years and rarely, if ever, acquire the infection again.

PREVENTION

Protective measures for the prevention of Lyme disease may include the avoidance of tick-infested areas, the use of repellents and acaricides, tick checks, and modification of landscapes in or near residential areas. Although a vaccine for Lyme disease used to be available, the manufacturer has discontinued its production. Another company is planning testing of a similar vaccine in both the United States and Europe. However, no vaccine is currently available commercially for the prevention of this infection.

FURTHER READING

Arvikar SL, Steere AC: Diagnosis and treatment of Lyme arthritis. Infect Dis Clin North Am 29:269, 2015.

Branda JA et al: Two-tiered antibody testing for Lyme disease with use of 2 enzyme immunoassays, a whole-cell sonicate enzyme immunoassay followed by a VlsE C6 peptide enzyme immunoassay. Clin Infect Dis 53:541, 2011.

Klempner MS et al: Two controlled trials of antibiotic treatment in patients with persistent symptoms and a history of Lyme disease. N Engl J Med 345:85, 2001.

Li X et al: Burden and viability of *Borrelia burgdorferi* in skin or joints of patients with erythema migrans or Lyme arthritis. Arthritis Rheum 63:2238, 2011.

Oschmann P et al: Stages and syndromes of neuroborreliosis. J Neurol 245:262, 1998.

Steere AC: Lyme disease. N Engl J Med 345:115, 2001.

Steere AC et al: Prospective study of serologic tests for Lyme disease. Clin Infect Dis 47:188, 2008.

Steere AC et al: Lyme borreliosis. Nat Rev Dis Primers 2:16090, 2016.

Wormser GP et al: The clinical assessment, treatment, and prevention of Lyme disease, human granulocytic anaplasmosis, and babesiosis: Clinical practice guidelines by the Infectious Diseases Society of America. Clin Infect Dis 43:1089, 2006.

Chapter 182: Rickettsial Diseases

David H. Walker; J. Stephen Dumler; Lucas S. Blanton; Thomas Marrie

INTRODUCTION

Rickettsiae are a heterogeneous group of small, obligately intracellular, gram-negative coccobacilli and short bacilli, most of which are transmitted by a tick, mite, flea, or louse vector. Except in the case of louse-borne typhus, humans are incidental hosts. Among rickettsiae, *Coxiella burnetii*, *Rickettsia prowazekii*, and *Rickettsia typhi* have the well-documented ability to survive for an extended period outside the reservoir or vector and to be extremely infectious: inhalation of a single *Coxiella* microorganism can cause pneumonia. High-level infectivity and severe illness after inhalation make *R. prowazekii*, *R. rickettsii*, *R. typhi*, *R. conorii*, and *C. burnetii* bioterrorism threats **(Chap. S2)**.

Clinical infections with rickettsiae can be classified according to (1) the taxonomy and diverse microbial characteristics of the agents, which belong to seven genera (*Rickettsia*, *Orientia*, *Ehrlichia*, *Anaplasma*, *Neorickettsia*, "*Candidatus* Neoehrlichia," and *Coxiella*); (2) epidemiology; or (3) clinical manifestations. The clinical manifestations of all the acute presentations are similar during the first 5 days: fever, headache, and myalgias with or without nausea, vomiting, and cough. As the course progresses, clinical manifestations—including a macular, maculopapular, or vesicular rash; eschar; pneumonitis; and meningoencephalitis—vary from one disease to another. Given the many etiologic agents with varied mechanisms of transmission, geographic distributions, and associated disease manifestations, the consideration of rickettsial diseases as a single entity poses complex challenges **(Table 182-1)**.

TABLE 182-1

Features of Selected Rickettsial Infections

Disease	Organism	Transmission	Geographic Range	Incubation Period, Days	Duration, Days	Rash, %	Eschar, %	Lymphadenopathy[a]
Rocky Mountain spotted fever	*Rickettsia rickettsii*	Tick bite: *Dermacentor andersoni, D. variabilis*	United States	2–14	10–20	90	<1	+
		Amblyomma cajennense sensu lato, *A. aureolatum*	Central/South America					
		Rhipicephalus sanguineus	Mexico, Brazil, United States					
Mediterranean spotted fever	*R. conorii*	Tick bite: *R. sanguineus, R. pumilio*	Southern Europe, Africa, Middle East, central Asia	5–7	7–14	97	50	+
African tick-bite fever	*R. africae*	Tick bite: *A. hebraeum, A. variegatum*	Sub-Saharan Africa, West Indies	4–10	4–19	50	90	+++
Maculatum disease	*R. parkeri*	Tick bite: *A. maculatum, A. triste, A. tigrinum*	United States, South America	2–10	6–16	88	94	++
Pacific Coast tick fever	*Rickettsia* 364D	Tick bite: *D. occidentalis*	United States	3–9	5–14	14	100	+++
Rickettsialpox	*R. akari*	Mite bite: *Liponyssoides sanguineus*	United States, Ukraine, Turkey, Mexico, Croatia	10–17	3–11	100	90	+++
Tick-borne lymphadenopathy	*R. slovaca*	Tick bite: *D. marginatus, D. reticularis*	Europe	7–9	17–180	5	100	++++
Flea-borne spotted fever	*R. felis*	Flea (mechanism undetermined): *Ctenocephalides felis*	Worldwide	8–16	8–16	80	15	—

Disease	Organism	Transmission	Geographic Range	Incubation Period, Days	Duration, Days	Rash, %	Eschar, %	Lymphadenopathy[a]
Epidemic typhus	*R. prowazekii*	Louse feces: *Pediculus humanus corporis*, fleas and lice of flying squirrels, or recrudescence	Worldwide	7–14	10–18	80	None	—
Murine typhus	*R. typhi*	Flea feces: *Xenopsylla cheopis, C. felis*, others	Worldwide	8–16	9–18	80	None	—
Human monocytotropic ehrlichiosis	*Ehrlichia chaffeensis*	Tick bite: *A. americanum, D. variabilis*	United States	1–21	3–21	26	None	++
Ewingii ehrlichiosis	*E. ewingii*	Tick bite*: A. americanum*	United States	1–21	4–21	0	None	
Unnamed ehrlichiosis	*E. muris eauclairensis*	Tick bite*: Ixodes scapularis*	United States	Unknown	3–14	12	None	
Human granulocytotropic anaplasmosis	*Anaplasma phagocytophilum*	Tick bite: *I. scapularis, I. ricinus, I. pacificus, I. persulcatus*	United States, Europe, Asia	4–8	3–14	Rare	None	—
Unnamed disease	*A. capra*	*I. persulcatus*	Northeastern China	Unknown	11–21	17	9	+
Neoehrlichiosis	*"Candidatus* Neoehrlichia mikurensis"	Tick bite: *I. ricinus, I. persulcatus, Haemaphysalis concinna*	Europe, China	≥8	11–75	10	None	
Scrub typhus	*Orientia tsutsugamushi*	Mite bite: *Leptotrombidium deliense*, others	Asia, Australia, Pacific and Indian Ocean islands	9–18	6–21	50	35	+++

Disease	Organism	Transmission	Geographic Range	Incubation Period, Days	Duration, Days	Rash, %	Eschar, %	Lymphadenopathy[a]
Q fever	*Coxiella burnetii*	Inhalation of aerosols of infected parturition material (goats, sheep, cattle, cats, others), ingestion of infected milk or milk products	Worldwide except New Zealand, Antarctica	3–30	5–57	<1	None	—

[a]++++, severe; +++, marked; ++, moderate; +, present in a small proportion of cases; —, not a noted feature.

Establishing the etiologic diagnosis of rickettsioses is very difficult during the acute stage of illness, and definitive diagnosis usually requires the examination of serum samples during the acute and convalescent phases of illness. Heightened clinical suspicion is based on epidemiologic data, history of exposure to vectors or reservoir animals, travel to endemic locations, clinical manifestations (sometimes including rash or eschar), and characteristic laboratory findings (including thrombocytopenia, normal or low white blood cell [WBC] counts, elevated hepatic enzyme levels, and hyponatremia). Such suspicion should prompt empirical treatment. Doxycycline is the empirical drug of choice for most of these infections. Only one agent, *C. burnetii*, has been documented to cause chronic illness. One other species, *R. prowazekii*, causes recrudescent illness (Brill-Zinsser disease) when latent infection is reactivated years after resolution of the acute illness.

Rickettsial infections dominated by fever may resolve without further clinical evolution. However, after nonspecific early manifestations, the illnesses can also evolve along one or more of several principal clinical lines: (1) development of a macular or maculopapular rash; (2) development of an eschar at the site of tick or mite feeding; (3) development of a vesicular rash (often in rickettsialpox and African tick-bite fever); (4) development of pneumonitis with chest radiographic opacities and/or rales (Q fever and severe cases of Rocky Mountain spotted fever [RMSF], Mediterranean spotted fever [MSF], louse-borne typhus, human monocytotropic ehrlichiosis [HME], human granulocytotropic anaplasmosis [HGA], scrub typhus, and murine typhus); (5) development of meningoencephalitis (louse-borne typhus and severe cases of RMSF, scrub typhus, HME, murine typhus, MSF, and [rarely] Q fever); and (6) progressive hypotension and multiorgan failure as seen with sepsis or toxic shock syndromes (RMSF, MSF, louse-borne typhus, murine typhus, scrub typhus, HME, HGA, and neoehrlichiosis).

Epidemiologic clues to the transmission of a particular pathogen include (1) environmental exposure to ticks, fleas, or mites during the season of activity of the vector species for the disease in the appropriate geographic region (spotted fever and typhus rickettsioses, scrub typhus, ehrlichiosis, anaplasmosis); (2) travel to or residence in an endemic geographic region during the incubation period (Table 182-1); (3) exposure to parturient ruminants, cats, and dogs (Q fever); (4) exposure to flying squirrels (*R. prowazekii* infection); and (5) history of previous louse-borne typhus (recrudescent typhus).

Clinical laboratory findings such as thrombocytopenia (particularly in spotted fever and typhus rickettsioses, ehrlichiosis, anaplasmosis, and scrub typhus), normal or low WBC counts, mild to moderate serum elevations of hepatic aminotransferases, and hyponatremia suggest some common pathophysiologic mechanisms.

Application of these clinical, epidemiologic, and laboratory principles requires consideration of a rickettsial diagnosis and knowledge of the individual diseases.

TICK-, MITE-, LOUSE-, AND FLEA-BORNE RICKETTSIOSES

These diseases, caused by organisms of the genera *Rickettsia* and *Orientia* in the family Rickettsiaceae, result from endothelial cell infection and increased vascular permeability. Pathogenic rickettsial species are very closely related, have small genomes (as a result of reductive evolution, which eliminated many genes for biosynthesis of intracellularly available molecules), and are traditionally separated into typhus

and spotted fever groups on the basis of lipopolysaccharide antigens. Some diseases and their agents (e.g., *R. africae*, *R. parkeri*, and *R. sibirica*) are too similar to require separate descriptions. Indeed, the similarities among MSF (*R. conorii* [all strains] and *R. massiliae*), North Asian tick typhus (*R. sibirica*), Japanese spotted fever (*R. japonica*), and Flinders Island spotted fever (*R. honei*) far outweigh their minor variations. The Rickettsiaceae that cause life-threatening infections are, in order of decreasing case–fatality rate, *R. rickettsii* (RMSF); *R. prowazekii* (louse-borne typhus); *Orientia tsutsugamushi* (scrub typhus); *R. conorii* (MSF); *R. typhi* (murine typhus); and, in rare cases, other spotted fever–group (SFG) organisms. Some agents (e.g., *R. parkeri*, *R. africae*, *Rickettsia* 364D, *R. akari*, *R. slovaca*, *R. honei*, *R. felis*, *R. massiliae*, *R. helvetica*, *R. heilongjiangensis*, *R. aeschlimannii*, and *R. monacensis*) have never been documented to cause a fatal illness. The most prevalent SFG rickettsia in the United States, *R. amblyommatis*, has been circumstantially associated with asymptomatic seroconversion in most persons and with self-limited illness in others.

ROCKY MOUNTAIN SPOTTED FEVER

Epidemiology

RMSF occurs in 47 states (with the highest prevalence in the south-central and southeastern states) as well as in Canada, Mexico, and Central and South America. The infection is transmitted by *Dermacentor variabilis*, the American dog tick, in the eastern two-thirds of the United States and California; by *D. andersoni*, the Rocky Mountain wood tick, in the western United States; by *Rhipicephalus sanguineus*, the brown dog tick, in Mexico, Arizona, and probably Brazil; and by *Amblyomma sculptum*, *A. mixtum*, *A. patinoi*, *A. cajennense*, *A. tonelliae*, and *A. aureolatum* in Central and/or South America. Maintained principally by transovarian transmission from one generation of ticks to the next, *R. rickettsii* can be acquired by uninfected ticks through the ingestion of a blood meal from rickettsemic small mammals or by co-feeding adjacent to an infected tick.

Humans become infected during tick season (in the Northern Hemisphere, from April to September), although some cases occur in winter. The mortality rate was 20–25% in the preantibiotic era and has been reported at ~3–5% in the postantibiotic era, principally because of delayed diagnosis and treatment. Recent reporting of a relatively low mortality rate (0.4%) is likely an artifact related to the abundance of less pathogenic SFG rickettsial species and to a relatively low proportion of diagnostically confirmed cases. Indeed, the reported case–fatality ratios in confirmed cases in the United States and in parts of Arizona, where *R. rickettsii* is the sole infecting SFG species, are 9% and 10%, respectively. The case–fatality ratio is highest among children (<10 years of age) and in the later decades of life (>70 years).

Pathogenesis

R. rickettsii organisms are inoculated into the dermis along with secretions of the tick's salivary glands after ≥6 h of feeding. The rickettsiae spread lymphohematogenously throughout the body and infect numerous foci of contiguous endothelial cells. The dose-dependent incubation period is ~1 week (range, 2–14 days). Occlusive thrombosis and ischemic necrosis are not the fundamental pathologic bases for tissue and organ injury. Instead, increased vascular permeability, with resulting edema, hypovolemia, and ischemia, is responsible. Consumption of platelets results in thrombocytopenia in 32–52% of patients, but disseminated intravascular coagulation (DIC) with hypofibrinogenemia is rare. Activation of platelets, generation of thrombin, and activation of the fibrinolytic system all appear to be homeostatic physiologic responses to endothelial injury by nonocclusive hemostatic plugs.

Clinical Manifestations

Early in the illness, when medical attention usually is first sought, RMSF is difficult to distinguish from many self-limiting viral illnesses. Fever, headache, malaise, myalgia, nausea, vomiting, and anorexia are the most common symptoms during the first 3 days. The patient becomes progressively more ill as vascular infection and injury advance. In one large series, only one-third of patients were diagnosed with presumptive RMSF early in the clinical course and treated appropriately as outpatients. In the tertiary-care setting, RMSF is all too often recognized only when late severe manifestations, developing at the end of the first week or during the second week of illness in patients without appropriate treatment, prompt return to a physician or hospital and admission to an intensive care unit.

The progressive nature of the infection is clearly manifested in the skin. Rash is evident in only 14% of patients on the first day of illness and in only 49% during the first 3 days. Macules (1–5 mm) appear first on the wrists and ankles and then on the remainder of the extremities and the trunk. Later, more severe vascular damage results in frank hemorrhage at the center of the maculopapule, producing a petechia that does not disappear upon compression (Fig. 182-1). This sequence of events is sometimes delayed or aborted by effective treatment. However, the rash is a variable manifestation, appearing on day 6 or later in 20% of cases and not appearing at all in 9–16% of cases. Petechiae occur in 41–59% of cases, appearing on or after day 6 in 74% of cases that manifest a rash. Involvement of the palms and soles,

often considered diagnostically important, usually develops relatively late in the course (after day 5 in 43% of cases) and does not develop at all in 18–64% of cases.

FIGURE 182-1

Top: Petechial lesions of Rocky Mountain spotted fever on the lower legs and soles of a young, previously healthy patient. ***Bottom:*** Close-up of lesions from the same patient. *(Photos courtesy of Dr. Lindsey Baden; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Hypovolemia leads to prerenal azotemia and (in 17% of cases) hypotension. Infection of the pulmonary microcirculation leads to noncardiogenic pulmonary edema; 12% of patients have severe respiratory disease, and 8% require mechanical ventilation. Cardiac involvement manifests as dysrhythmia in 7–16% of cases.

Besides respiratory failure, central nervous system (CNS) involvement is the other important determinant of the outcome of RMSF. Encephalitis, presenting as confusion or lethargy, is apparent in 26–28% of cases. Progressively severe encephalitis manifests as stupor or delirium in 21–26% of cases, ataxia in 18%, coma in 10%, and seizures in 8%. Numerous focal neurologic deficits have been reported. Meningoencephalitis results in cerebrospinal fluid (CSF) pleocytosis in 34–38% of cases; usually there are 10–100 cells/μL and a mononuclear predominance, but occasionally there are >100 cells/μL and a polymorphonuclear predominance. The CSF protein concentration is increased in 30–35% of cases, but the CSF glucose concentration is usually normal.

Renal failure, often reversible with rehydration, is caused by acute tubular necrosis in severe cases with shock. Hepatic injury with increased serum aminotransferase concentrations (38% of cases) is due to focal death of individual hepatocytes without hepatic failure. Jaundice is

recognized in 9% of cases and an elevated serum bilirubin concentration in 18–30%.

Life-threatening bleeding is rare. Anemia develops in 30% of cases and is severe enough to require transfusions in 11%. Blood is detected in the stool or vomitus of 10% of patients, and death has followed massive upper-gastrointestinal hemorrhage.

Other characteristic clinical laboratory findings include increased plasma levels of proteins of the acute-phase response (C-reactive protein, fibrinogen, ferritin, and others), hypoalbuminemia, and hyponatremia (in 56% of cases) due to the appropriate secretion of antidiuretic hormone in response to the hypovolemic state. Myositis occurs occasionally, with marked elevations in serum creatine kinase levels and multifocal rhabdomyonecrosis. Ocular involvement includes conjunctivitis in 30% of cases and retinal vein engorgement, flame hemorrhages, arterial occlusion, and papilledema with normal CSF pressure in some instances.

In untreated cases, the patient usually dies 8–15 days after onset. A rare presentation, fulminant RMSF, is fatal within 5 days after onset. This fulminant presentation is seen most often in male black patients with glucose-6-phosphate dehydrogenase (G6PD) deficiency and may be related to an undefined effect of hemolysis on the rickettsial infection. Although survivors of RMSF usually return to their previous state of health, permanent sequelae, including neurologic deficits and gangrene necessitating amputation of extremities, may follow severe illness.

Diagnosis

The diagnosis of RMSF during the acute stage is more difficult than is generally appreciated. The most important epidemiologic factor is a history of exposure to a potentially tick-infested environment within the 14 days preceding disease onset during a season of possible tick activity. However, only 60% of patients actually recall being bitten by a tick during the incubation period.

The differential diagnosis for early clinical manifestations of RMSF (fever, headache, and myalgia without a rash) includes influenza, enteroviral infection, infectious mononucleosis, viral hepatitis, leptospirosis, typhoid fever, gram-negative or gram-positive bacterial sepsis, HME, HGA, murine typhus, sylvatic flying-squirrel typhus, and rickettsialpox. Enterocolitis may be suggested by nausea, vomiting, and abdominal pain; prominence of abdominal tenderness has resulted in exploratory laparotomy. CNS involvement can masquerade as bacterial or viral meningoencephalitis. Cough, pulmonary signs, and chest radiographic opacities can lead to a diagnostic consideration of bronchitis or pneumonia.

At presentation during the first 3 days of illness, only 3% of patients exhibit the classic triad of fever, rash, and history of tick exposure. When a rash appears, a diagnosis of RMSF should be considered. However, many illnesses considered in the differential diagnosis also can be associated with a rash, including rubeola, rubella, meningococcemia, disseminated gonococcal infection, secondary syphilis, toxic shock syndrome, drug hypersensitivity, idiopathic thrombocytopenic purpura, thrombotic thrombocytopenic purpura, Kawasaki syndrome, and immune complex vasculitis. Conversely, any person in an endemic area with a provisional diagnosis of one of the above illnesses could have RMSF. Thus, if a viral infection is suspected during RMSF season in an endemic area, it should always be kept in mind that RMSF can mimic viral infection early in the course; if the illness worsens over the next couple of days after initial presentation, the patient should return for reevaluation.

The most common serologic test for confirmation of the diagnosis is the indirect immunofluorescence assay. Not until 7–10 days after onset is a diagnostic titer of ≥64 usually detectable. The sensitivity and specificity of the indirect immunofluorescence IgG assay are 89–100% and 99–100%, respectively. Detection of IgM is no more sensitive in early illness and is subject to nonspecific cross-reactivity. It is important to understand that serologic tests for RMSF are usually negative at the time of presentation for medical care and that treatment should not be delayed while a positive serologic result is awaited.

The only diagnostic test that has proven useful during the acute illness is immunohistologic examination of a cutaneous biopsy sample from a rash lesion for *R. rickettsii*. Examination of a 3-mm punch biopsy from such a lesion is 70% sensitive and 100% specific. Polymerase chain reaction (PCR) amplification for detection of *R. rickettsii* DNA in peripheral blood is not adequately sensitive. Although rickettsiae are present in large quantities in heavily infected foci of endothelial cells, there are relatively low quantities in the circulation. Cultivation of rickettsiae in cell culture is feasible but is seldom undertaken because of biohazard concerns. The recent dramatic increase in the reported incidence of RMSF correlates with the use of single-titer SFG cross-reactive enzyme immunoassay serology. Few cases are specifically determined to be caused by *R. rickettsii*. Currently, many febrile persons who do not have RMSF present with cross-reactive antibodies, possibly because of previous exposure to the highly prevalent SFG rickettsia *R. amblyommatis*.

TREATMENT

TREATMENT

Rocky Mountain Spotted Fever

The drug of choice for the treatment of both children and adults with RMSF is doxycycline. Because of the severity of RMSF, immediate empirical administration of doxycycline should be strongly considered for any patient with a consistent clinical presentation in the appropriate epidemiologic setting. Doxycycline is administered orally (or, with coma or vomiting, intravenously) at 100 mg twice daily. For children with suspected RMSF, up to five courses of doxycycline may be administered with minimal risk of dental staining. In patients with allergy to doxycycline, desensitization should be considered. Other regimens include oral tetracycline (500 mg four times daily). Treatment with chloramphenicol, a less effective drug, is advised only for patients who are pregnant. Although available in much of the world, chloramphenicol is difficult to obtain in the United States; when it is unavailable, doxycycline should be used. There is little evidence to support the occurrence of tetracycline-associated adverse events in mothers (hepatotoxicity) and fetuses (staining of deciduous teeth and teratogenicity) who receive doxycycline. The antirickettsial drug should be administered until the patient is afebrile and improving clinically —usually 3–5 days after defervescence. β-Lactam antibiotics, erythromycin, and aminoglycosides have no role in the treatment of RMSF, and sulfa-containing drugs are associated with more adverse outcomes than no treatment at all. There is little clinical experience with fluoroquinolones, clarithromycin, and azithromycin, which are not recommended. The most seriously ill patients are managed in intensive care units, with careful administration of fluids to achieve optimal tissue perfusion without precipitating noncardiogenic pulmonary edema. In some severely ill patients, hypoxemia requires intubation and mechanical ventilation; oliguric or anuric acute renal failure requires hemodialysis; seizures necessitate the use of antiseizure medication; anemia or severe hemorrhage necessitates transfusions of packed red blood cells; or bleeding with severe thrombocytopenia requires platelet transfusions.

Prevention

Avoidance of tick bites is the only available preventive approach. Use of protective clothing and tick repellents, inspection of the body once or twice a day, and removal of ticks before they inoculate rickettsiae reduce the risk of infection. Prophylactic doxycycline treatment of tick bites has no proven role in preventing RMSF.

MEDITERRANEAN SPOTTED FEVER (BOUTONNEUSE FEVER), AFRICAN TICK-BITE FEVER, AND OTHER TICK-BORNE SPOTTED FEVERS

Epidemiology and Clinical Manifestations

 *R. conorii* is prevalent in southern Europe, Africa, and southwestern and south-central Asia. Regional names for the disease caused by this organism include Mediterranean spotted fever, Kenya tick typhus, Indian tick typhus, Israeli spotted fever, and Astrakhan spotted fever. The disease is characterized by high fever, rash, and—in most geographic locales—an inoculation eschar (*tâche noire*) that appears before the onset of fever at the site of the tick bite. A severe form of the disease (mortality rate, 50%) occurs in patients with diabetes, alcoholism, or heart failure.

African tick-bite fever, caused by *R. africae*, occurs in rural areas of sub-Saharan Africa and in the Caribbean islands and is transmitted by *Amblyomma hebraeum* and *A. variegatum* ticks. The average incubation period is 4–10 days. The mild illness consists of headache, fever, eschar, and regional lymphadenopathy. *Amblyomma* ticks, a high portion of which are infected with *R. africae*, often feed in groups, with the consequent development of multiple eschars. Rash may be vesicular, sparse, or absent altogether. Because of tourism in sub-Saharan Africa, African tick-bite fever is the rickettsiosis most frequently imported into Europe and North America. Maculatum disease, a similar disease caused by the closely related species *R. parkeri*, is transmitted by *A. maculatum* and found in a low percentage of *A. americanum* ticks in the United States. It is also transmitted by *A. triste* in South America and Arizona as well as *A. tigrinum* in South America.

R. japonica causes Japanese spotted fever, which also occurs in Korea and China. Similar diseases in northern Asia are caused by *R. sibirica* and *R. heilongjiangensis.* Queensland tick typhus due to *R. australis* is transmitted by *Ixodes holocyclus* ticks. Flinders Island spotted fever, found on the island for which it is named as well as in Tasmania, mainland Australia, and Asia, is caused by *R. honei*. In Europe, patients infected with *R. slovaca* after a wintertime *Dermacentor* tick bite usually manifest an afebrile illness with an eschar (usually on the scalp) and painful regional lymphadenopathy.

Diagnosis

Diagnosis of these tick-borne spotted fevers is based on clinical and epidemiologic findings and is confirmed by serology, immunohistochemical demonstration of rickettsiae in skin biopsy specimens, cell-culture isolation of rickettsiae, or PCR of skin biopsy,

eschar biopsy or swab, or blood samples. Serologic diagnosis detects antibodies to antigens shared among SFG rickettsiae, hindering identification of the etiologic species. In an endemic area, a possible diagnosis of rickettsial spotted fevers should be considered when patients present with fever, rash, and/or a skin lesion consisting of a black necrotic lesion or a crust surrounded by erythema.

TREATMENT

TREATMENT

Tick-Borne Spotted Fevers

Successful therapeutic agents include doxycycline (100 mg bid orally for 1–5 days) and chloramphenicol (500 mg qid orally for 7–10 days). Pregnant patients may be treated with josamycin (3 g/d orally for 5 days). Data on the efficacy of treatment of mildly ill children with clarithromycin or azithromycin should not be extrapolated to adults or to patients with moderate or severe illness.

RICKETTSIALPOX

R. akari infects mice and their mites (*Liponyssoides sanguineus*), which maintain the organisms by transovarial transmission.

Epidemiology

Rickettsialpox is recognized principally in New York City, but cases have also been reported in other urban and rural locations in the United States and in Ukraine, Croatia, Mexico, and Turkey. Investigation of eschars suspected of representing bioterrorism-associated cutaneous anthrax revealed that rickettsialpox occurs more frequently than previously realized.

Clinical Manifestations

A papule forms at the site of the mite's feeding, develops a central vesicle, and becomes a 1- to 2.5-cm painless black crusted eschar surrounded by an erythematous halo **(Fig. 182-2)**. Enlargement of the regional lymph nodes draining the eschar suggests initial lymphogenous spread. After an incubation period of 10–17 days, during which the eschar and regional lymphadenopathy frequently go unnoticed, disease onset is marked by malaise, chills, fever, headache, and myalgia. A macular rash appears 2–6 days after onset and usually evolves sequentially into papules, vesicles, and crusts that heal without scarring **(Fig. 182-3)**; in some cases, the rash remains macular or maculopapular. Some patients develop nausea, vomiting, abdominal pain, cough, conjunctivitis, or photophobia. Without treatment, fever lasts 6–10 days.

FIGURE 182-2

Eschar at the site of the mite bite in a patient with rickettsialpox. *(Reprinted from A Krusell et al: Emerg Infect Dis 8:727, 2002. Photo obtained by Dr. Kenneth Kaye.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 182-3

Top: Papulovesicular lesions on the trunk of the patient with rickettsialpox shown in **Fig. 182-2**. ***Bottom:*** Close-up of lesions from the same patient. *(Reprinted from A Krusell et al: Emerg Infect Dis 8:727, 2002. Photos obtained by Dr. Kenneth Kaye.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Diagnosis and Treatment

Clinical, epidemiologic, and convalescent serologic data establish the diagnosis of an SFG rickettsiosis that is seldom pursued further. Doxycycline is the drug of choice for treatment.

FLEA-BORNE SPOTTED FEVER

An emerging rickettsiosis caused by *R. felis* occurs worldwide. Maintained transovarially in the geographically widespread cat flea *Ctenocephalides felis*, the infection has been described as moderately severe, with fever, rash, and headache as well as CNS, gastrointestinal, and pulmonary symptoms.

EPIDEMIC (LOUSE-BORNE) TYPHUS

Epidemiology

The human body louse (*Pediculus humanus corporis*) lives in clothing under poor hygienic conditions and usually in impoverished cold areas. Lice acquire *R. prowazekii* when they ingest blood from a rickettsemic patient. The rickettsiae multiply in the louse's midgut epithelial cells and are shed in its feces. The infected louse leaves a febrile person and deposits infected feces on its subsequent host during its blood meal; the patient autoinoculates the organisms by scratching. The louse is killed by the rickettsiae and does not pass *R. prowazekii* to its offspring.

 Epidemic typhus haunts regions afflicted by wars and disasters. An outbreak involved 100,000 people in refugee camps in Burundi in 1997. A small focus was documented in Russia in 1998, sporadic cases were reported from Algeria, and frequent outbreaks occurred in Peru and Rwanda. Eastern flying squirrels (*Glaucomys volans*) and their lice and fleas maintain *R. prowazekii* in a zoonotic cycle.

Brill-Zinsser disease is a recrudescent illness occurring years after acute epidemic typhus, probably as a result of waning immunity. *R. prowazekii* remains latent for years; its reactivation results in sporadic cases of disease in louse-free populations or in epidemics in louse-infested populations. Recrudescence has been documented after flying squirrel–associated typhus.

Rickettsiae are potential agents of bioterrorism **(Chap. S2)**. Infections with *R. prowazekii* and *R. rickettsii* have high case–fatality ratios. These organisms cause difficult-to-diagnose diseases and are highly infectious when inhaled as aerosols. Organisms resistant to tetracycline or chloramphenicol have been developed in the laboratory.

Clinical Manifestations

After an incubation period of ~1–2 weeks, the onset of illness is abrupt, with prostration, severe headache, and fever rising rapidly to 38.8°–40.0°C (102°–104°F). Cough is prominent, developing in 70% of patients. Myalgias are usually severe. A rash begins on the upper trunk, usually on the fifth day, and then becomes generalized, involving the entire body except the face, palms, and soles. Initially, this rash is macular; without treatment, it becomes maculopapular, petechial, and confluent. The rash often goes undetected on black skin; 60% of African patients have spotless epidemic typhus. Photophobia, with considerable conjunctival injection and eye pain, is common. The tongue may be dry, brown, and furred. Confusion and coma are common. Skin necrosis and gangrene of the digits as well as interstitial pneumonia may occur in severe cases. Untreated disease is fatal in 7–40% of cases, with outcome depending primarily on the condition of the host. Patients with untreated infections develop renal insufficiency and multiorgan involvement in which neurologic manifestations are frequently prominent. Overall, 12% of patients with epidemic typhus have neurologic involvement. Infection associated with North American flying squirrels is a milder illness; whether this milder disease is due to host factors (e.g., better health status) or attenuated virulence is unknown.

Diagnosis and Treatment

Epidemic typhus is sometimes misdiagnosed as typhoid fever in tropical countries **(Chap. 160)**. The means even for serologic studies are often unavailable in settings of louse-borne typhus. Epidemics can be recognized by the serologic or immunohistochemical diagnosis of a single case or by detection of *R. prowazekii* in a louse found on a patient. Doxycycline (100 mg bid) is administered orally or—if the patient is comatose or vomiting—intravenously and continued until 3–5 days after defervescence. Under epidemic conditions, a single 200-mg oral dose can be tried but fails in some cases. Pregnant patients should be evaluated individually and treated with chloramphenicol early in pregnancy or, if necessary, with doxycycline late in pregnancy.

Prevention

Prevention of epidemic typhus involves control of body lice. Clothes should regularly be changed and laundered in hot water, and insecticides can be used every 6 weeks to control the louse population.

ENDEMIC MURINE TYPHUS

Epidemiology

R. typhi is maintained in mammalian host–flea cycles, with rats (*Rattus rattus* and *R. norvegicus*) and the Oriental rat flea (*Xenopsylla cheopis*) as the classic zoonotic niche. Fleas acquire *R. typhi* from rickettsemic rats and carry the organism throughout their life span. Nonimmune rats and humans are infected when rickettsia-laden flea feces contaminate pruritic bite lesions; less frequently, the flea bite transmits the organisms. Transmission can also occur via inhalation of aerosolized rickettsiae from flea feces. Infected rats appear healthy, although they are rickettsemic for ~2 weeks.

Murine typhus occurs mainly in Texas and southern California, where the classic rat–flea cycle is absent and an opossum–cat flea (*C. felis*) cycle is prominent. Globally, endemic typhus occurs mainly in warm (often coastal) areas throughout the tropics and subtropics, where it is highly prevalent though often unrecognized. The incidence peaks from April through July in southern Texas and during the warm months of summer and early fall in other geographic locations. Patients seldom recall exposure to fleas, although exposure to animals such as cats, opossums, and rats is reported in nearly 40% of cases.

Clinical Manifestations

The incubation period of experimental murine typhus averages 11 days (range, 8–16 days). Headache, myalgia, arthralgia, nausea, and malaise develop 1–3 days before onset of chills and fever. Patients often experience nausea and vomiting.

The duration of untreated illness averages 12 days (range, 9–18 days). Rash is present in only 13% of patients at presentation for medical care (usually ~4 days after onset of fever), appearing an average of 2 days later in half of the remaining patients and never appearing in the others. The initial macular rash is often detected by careful inspection of the axilla or the inner surface of the arm. Subsequently, the rash becomes maculopapular, involving the trunk more often than the extremities; it is seldom petechial and rarely involves the face, palms, or soles. A rash is detected in only 20% of patients with darkly pigmented skin.

Pulmonary involvement is frequently prominent; 35% of patients have a hacking, nonproductive cough, and 23% of patients who undergo chest radiography have pulmonary densities due to interstitial pneumonia, pulmonary edema, and pleural effusions. Bibasilar rales are the most common pulmonary sign. Less common clinical manifestations include abdominal pain, confusion, stupor, seizures, ataxia, coma, and jaundice. Clinical laboratory studies frequently reveal anemia and leukopenia early in the course, leukocytosis late in the course, thrombocytopenia, hyponatremia, hypoalbuminemia, increased serum levels of hepatic aminotransferases, and prerenal azotemia. Complications can include respiratory failure, hematemesis, cerebral hemorrhage, and hemolysis. Severe illness necessitates the admission of 10% of hospitalized patients to an intensive care unit. Greater severity is generally associated with old age, underlying disease, and treatment with a sulfonamide; the case–fatality rate is 1%.

Diagnosis and Treatment

Serologic studies of acute- and convalescent-phase serum samples can provide a diagnosis, and an immunohistochemical method for identification of typhus group-specific antigens in biopsy samples has been developed. Cultivation is used infrequently and is not widely available. PCR of the blood is not adequately sensitive. When endemic typhus is suspected, patients should be treated empirically with doxycycline (100 mg twice daily by mouth for 7–15 days). Chloramphenicol and ciprofloxacin are less effective alternatives.

SCRUB TYPHUS

Epidemiology

O. tsutsugamushi differs substantially from *Rickettsia* species both genetically and in cell wall composition (i.e., it lacks lipopolysaccharide). *O. tsutsugamushi* is maintained by transovarial transmission in trombiculid mites. After hatching, infected larval mites (chiggers, the only stage that feeds on a host) inoculate organisms into the skin. Infected chiggers are particularly likely to be found in areas of heavy scrub vegetation during the wet season, when mites lay eggs.

Scrub typhus is endemic and reemerging in eastern and southern Asia, northern Australia, and islands of the western Pacific and Indian Oceans. Infections are prevalent in these regions; in some areas, >3% of the population is infected or reinfected each month. Immunity wanes over 1–3 years, and the organism exhibits remarkable antigenic diversity. Emerging cases in Chile and Africa challenge the classic epidemiology of scrub typhus.

Clinical Manifestations

Illness varies from mild and self-limiting to fatal. After an incubation period of 6–21 days, onset is characterized by fever, headache, myalgia, cough, and gastrointestinal symptoms. Some patients recover spontaneously after a few days. The classic case description includes an eschar where the chigger has fed, regional lymphadenopathy, and a maculopapular rash—signs that are seldom seen in indigenous patients. In fact, fewer than 50% of Westerners develop an eschar, and fewer than 40% develop a rash (on day 4–6 of illness). Severe cases typically

manifest with encephalitis and interstitial pneumonia due to vascular injury. The case–fatality rate for untreated classic cases is 7% but would probably be lower if all mild cases were diagnosed.

Diagnosis and Treatment

Serologic assays (indirect fluorescent antibody, indirect immunoperoxidase, and enzyme immunoassays) are the mainstays of laboratory diagnosis. PCR amplification of *Orientia* genes from eschars and blood also is effective. Patients are treated with oral doxycycline (100 mg twice daily for 7–15 days), azithromycin (500 mg for 3 days), or chloramphenicol (500 mg four times daily for 7–15 days).

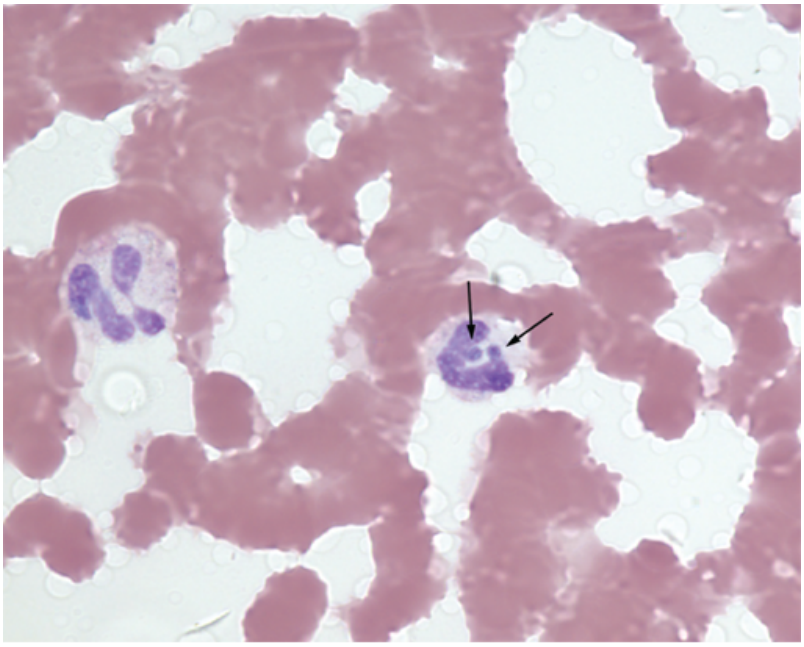
Some cases of scrub typhus in Thailand are poorly responsive to doxycycline or chloramphenicol but respond to azithromycin and rifampin.

EHRLICHIOSES AND ANAPLASMOSIS

Ehrlichioses are acute febrile infections caused by members of the family Anaplasmataceae, which is made up of obligately intracellular organisms of five genera: *Ehrlichia*, *Anaplasma*, *Wolbachia*, "*Candidatus* Neoehrlichia," and *Neorickettsia*. The bacteria reside in vertebrate reservoirs and target vacuoles of hematopoietic—and, for some species, endothelial—cells **(Fig. 182-4)**. Four *Ehrlichia* species, two *Anaplasma* species, and one *Neoehrlichia* species are transmitted by ticks to humans and cause infection that can be severe and prevalent. *E. chaffeensis*, the agent of HME, and *E. muris eauclairensis* infect predominantly mononuclear phagocytes; *E. ewingii* and *A. phagocytophilum* infect neutrophils. Infections with "*Candidatus* Neoehrlichia mikurensis" and *A. capra* are less well characterized, but human blood neutrophils and monocytes, respectively, are suspected targets.

FIGURE 182-4

Peripheral-blood smear from a patient with human granulocytotropic anaplasmosis. A neutrophil contains two morulae (vacuoles filled with *A. phagocytophilum). (Photo courtesy of Dr. J. Stephen Dumler.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ehrlichia, "*Candidatus* Neoehrlichia," and *Anaplasma* are maintained by horizontal tick–mammal–tick transmission, and humans are only inadvertently infected. Wolbachiae are associated with human filariasis, since they are important for filarial viability and pathogenicity; antibiotic treatment targeting wolbachiae is a strategy for filariasis control. Neorickettsiae parasitize flukes (trematodes) that in turn parasitize aquatic snails, fish, and insects. Only a single human neorickettsiosis has been described: sennetsu fever, an infectious mononucleosis–like illness first identified in 1953 in association with the ingestion of raw fish containing *N. sennetsu*–infected flukes.

HUMAN MONOCYTOTROPIC EHRLICHIOSIS

Epidemiology

More than 14,048 cases of *E. chaffeensis* infection had been reported to the U.S. Centers for Disease Control and Prevention (CDC) as of January 2017. However, active prospective surveillance documented an incidence as high as 414 cases per 100,000 population in some U.S. regions. Most *E. chaffeensis* infections are identified in the south-central, southeastern, and mid-Atlantic states, but cases have also been recognized in California, New York, and Wisconsin. All stages of the Lone Star tick (*A. americanum*) feed on white-tailed deer—a major reservoir. Dogs and coyotes also serve as reservoirs and often lack clinical signs. Tick bites and exposures are frequently reported by patients in rural areas, and 64% of infections occur in May through July. The median age of HME patients is 55 years; however, 11% of infections occur in children ≤19 years of age, and these include severe and fatal infections. Of patients with HME, 59% are male.

 *E. chaffeensis* has been detected in South and Central America, Africa, and Asia.

Clinical Manifestations

E. chaffeensis disseminates hematogenously from the dermal blood pool created by the feeding tick. After a median incubation period of 8 days, illness develops. Clinical manifestations are undifferentiated and include fever (97% of cases), headache (70%), myalgia (68%), and malaise (77%). Less frequently observed are nausea, vomiting, and diarrhea (28–57%); cough (30%); rash (29% overall, 6% at presentation); and confusion (20%). HME can be severe: 77% of patients with confirmed cases are hospitalized, and 2% die. Life-threatening complications include renal failure, meningoencephalitis, adult respiratory distress syndrome, a DIC-like syndrome, pneumonia, a septic shock–like syndrome, cardiac failure, hepatitis, hemorrhage, and—in immunocompromised patients—overwhelming ehrlichial infection; patients with diabetes, cancer, organ transplantation, asplenia, hepatitis C, or HIV infection have a 2.3 relative risk for death. Laboratory findings are valuable in the differential diagnosis of HME; 66% of patients have leukopenia (initially lymphopenia, later neutropenia), 86% have thrombocytopenia, and 89% have elevated serum levels of hepatic aminotransferases. Despite low blood cell counts, the bone marrow is hypercellular, and noncaseating granulomas can be present. Vasculitis is not a component of HME.

Diagnosis

HME can be fatal. If not given empirical doxycycline treatment, 39% and 40% of patients with HME require admission to an intensive care unit and mechanical ventilation, respectively; these measures are necessary in no patients receiving empirical treatment. In addition, hospital stay and illness duration are lengthened in untreated patients by 8 and 12 days, respectively. The diagnosis is suggested by fever, known tick exposure in the preceding 3 weeks, thrombocytopenia and/or leukopenia, and increased serum aminotransferase activities. Morulae are demonstrated in <10% of peripheral-blood smears. HME can be confirmed during active infection by PCR amplification of *E. chaffeensis* nucleic acids in blood obtained before the start of doxycycline therapy. Retrospective serodiagnosis requires a consistent clinical picture and a fourfold increase in *E. chaffeensis* antibody titer to ≥128 in paired serum samples obtained ~3 weeks apart. Separate specific diagnostic tests are necessary for HME and HGA (see below).

EWINGII EHRLICHIOSIS AND *EHRLICHIA MURIS EAUCLAIRENSIS* INFECTIONS

Ehrlichia ewingii resembles *E. chaffeensis* in its tick vector (*A. americanum*) and vertebrate reservoirs (white-tailed deer and dogs). *E. muris eauclairensis* causes human infections after *Ixodes scapularis* tick exposure in Wisconsin and Minnesota. *E. ewingii* and *E. muris* illnesses are similar to but less severe than HME. Many cases occur in immunocompromised patients. Human infections with *E. canis* have been documented as subclinical ehrlichemia. No specific serologic diagnostic tests for these other ehrlichiae are readily available, and *E. chaffeensis* serologic tests can be positive when the infecting agent is actually a different species of *Ehrlichia*.

"CANDIDATUS NEOEHRLICHIA MIKURENSIS" INFECTION

 *"Candidatus* Neoehrlichia mikurensis," a bacterium in a phylogenetic clade between *Ehrlichia* and *Anaplasma*, was originally identified in *Ixodes ricinus* ticks from the Netherlands and in mice and *Ixodes ovatus* ticks from Japan. By means of broad-range 16S rRNA gene amplification and sequence analysis, this organism was identified as the cause of severe and sometimes prolonged febrile illnesses in European immunocompromised patients with tick bites or exposures and in Chinese patients developing a mild febrile illness after being bitten by *Ixodes persulcatus* and *Haemaphysalis concinna* ticks. The clinical presentation is similar to those of HME and HGA. Specific diagnostic methods have been developed but are not widely available.

TREATMENT

TREATMENT

Ehrlichioses

Doxycycline is effective for HME as well as other ehrlichioses; the use of this drug in "*Candidatus* N. mikurensis" infection is associated with disease resolution. Therapy with doxycycline (100 mg given PO or IV twice daily) or tetracycline (250–500 mg given PO every 6 h) lowers hospitalization rates and shortens fever duration. *E. chaffeensis* is not susceptible to chloramphenicol in vitro, and the use of this drug is controversial. While a few reports document *E. chaffeensis* persistence in humans, this finding is rare; most infections are cured by short courses of doxycycline continuing for 3–5 days after defervescence. Although poorly studied for this indication, rifampin may be suitable when doxycycline is contraindicated.

Prevention

HME, *E. ewingii* ehrlichiosis, *E. muris* ehrlichiosis, and *"Candidatus* N. mikurensis" infection can be prevented by the avoidance of ticks in endemic areas. The use of protective clothing and tick repellents, careful postexposure tick searches, and prompt removal of attached ticks probably diminish infection risk.

HUMAN GRANULOCYTOTROPIC ANAPLASMOSIS

Epidemiology

 As of April 2013, 25,288 cases of HGA had been reported to the CDC, most in the upper-midwestern and northeastern United States. The global geographic distribution is similar to that of Lyme disease because of the shared *Ixodes* tick vectors. Natural reservoirs for *A. phagocytophilum* are white-footed mice, squirrels, and white-tailed deer in the United States and red deer in Europe. HGA incidence peaks in May through July, but the disease can occur throughout the year with exposure to *Ixodes* ticks. HGA often affects males (59%) and older persons (median age, 51 years).

Clinical Manifestations

Seroprevalence rates are high in endemic regions; thus it seems likely that most individuals develop subclinical infections. The incubation period for HGA is 4–8 days, after which the disease manifests as fever (75–100% of cases), myalgia (75%), headache (83%), and malaise (97%). A minority of patients develop nausea, vomiting, or diarrhea (21–39%); cough (29%); or confusion (17%). A rash in HGA (6%) almost invariably reflects co-infection with *Borrelia*, resulting in erythema migrans. Most patients develop thrombocytopenia (79%) and/or leukopenia (60%) with increased serum hepatic aminotransferase levels (91%).

Life-threatening complications occur most often in the elderly and include renal failure, adult respiratory distress syndrome, a toxic shock–like syndrome, pneumonia, and a DIC- or sepsis-like syndrome. Meningoencephalitis is rare in documented cases of HGA. Other documented neurologic sequelae include brachial plexopathy, cranial nerve involvement, and demyelinating polyneuropathy. Infection of patients with a preexisting immunocompromising condition (diabetes, immunosuppressive medications, asplenia, arthritis) is associated with a 3.0 relative risk for life-threatening complications. Of patients with HGA, 31% are hospitalized and 7% require intensive care. The case–fatality rate is 0.6%, but the relative risk for death is 16 if infection occurs with an immunosuppressive condition. Neither vasculitis nor granulomas are components of HGA. While patients can be co-infected with *Borrelia burgdorferi* and *Babesia microti* (transmitted by the same tick vector[s]), there is little evidence that these infections increase the severity or persistence of HGA. HGA transmitted by transfusion (including the transfusion of leukoreduced blood or platelets) has now been reported in nine cases.

Diagnosis

HGA should be included in the differential diagnosis of influenza-like illnesses during seasons with *Ixodes* tick activity (May through December), especially in the context of a known tick bite or exposure. Concurrent thrombocytopenia, leukopenia, or elevated serum levels of alanine or aspartate aminotransferase further increase the likelihood of HGA. Many HGA patients develop Lyme disease antibodies in the absence of clinical findings consistent with that diagnosis. Thus, HGA should be considered in the differential diagnosis of atypical severe Lyme disease presentations. Peripheral-blood film examination for neutrophil morulae can yield a diagnosis in 20–75% of infections. PCR testing of blood from patients with active disease before doxycycline therapy is sensitive and specific. Serodiagnosis is retrospective, requiring a fourfold increase in *A. phagocytophilum* antibody titer (to ≥160) in paired serum samples obtained 1 month apart. Since seroprevalence is high in some regions, a single acute-phase titer should not be used for diagnosis.

Anaplasma capra Infection

Human infection by *A. capra*, first isolated from goat blood, was identified in 28 patients from northeastern China. Patients presented with fever, headache, malaise, dizziness, myalgias, and chills, but these manifestations were less severe than in HGA. Hospitalization was recorded for 18% of patients, and 14% had underlying disorders, including hyperglycemia, hypertension, coronary heart disease, diabetes, and cancer. Five patients had severe manifestations, including one with encephalitic signs and *A. capra* DNA present in CSF. *A. capra* is found most often in *I. persulcatus* ticks in this region. All patients responded to doxycycline treatment and survived.

TREATMENT

TREATMENT

Human Granulocytotropic Anaplasmosis

No prospective studies of therapy for HGA have been conducted. However, doxycycline (100 mg PO twice daily) is effective. Rifampin therapy is associated with improvement of HGA in pregnant women and children. Most treated patients defervesce within 24–48 h.

Prevention

HGA prevention requires tick avoidance. Transmission can be documented as few as 4 h after a tick bite.

Q FEVER

The agent of Q fever is *C. burnetii*, a small intracellular prokaryote that only recently was grown in cell-free medium. *C. burnetii*, a pleomorphic coccobacillus with a gram-negative cell wall, survives in harsh environments; it escapes intracellular killing in macrophages by inhibiting the final step in phagosome maturation (cathepsin fusion) and has adapted to the acidic phagolysosome by producing superoxide dismutase. Infection with *C. burnetii* induces a range of immunomodulatory responses, from immunosuppression in chronic Q fever to the production of autoantibodies, particularly those to smooth muscle and phospholipids.

Q fever encompasses two broad clinical syndromes: acute and chronic infection. The host's immune response (rather than the particular strain) most likely determines whether chronic Q fever develops. *C. burnetii* survives in monocytes from patients with chronic Q fever but not in monocytes from patients with acute Q fever or from uninfected subjects. Impairment of the bactericidal activity of the *C. burnetii*–infected monocyte is associated with overproduction of interleukin 10. The CD4+/CD8+ ratio is decreased in Q fever endocarditis. Very few organisms and a strong cellular response are observed in patients with acute Q fever, while many organisms and a moderate cellular response occur in chronic Q fever. Immune control of *C. burnetii* is T cell–dependent, but 80–90% of bone marrow aspirates obtained years after recovery from Q fever contain *C. burnetii* DNA. *C. burnetii*'s ready multiplication within trophoblasts accounts for the high concentrations it can reach in the placenta.

Epidemiology

Q fever is a zoonosis. The primary sources of human infection are infected cattle, sheep, and goats. However, cats, rabbits, pigeons, kangaroos, and dogs also serve as sources for transmission of *C. burnetii* to humans. The wildlife reservoir is extensive and includes ticks, coyotes, gray foxes, skunks, raccoons, rabbits, deer, mice, bears, birds, opossums, and kangaroos. The three-legged sloth is important in the transmission of *C. burnetii* in French Guiana. In female animals, *C. burnetii* localizes to the uterus and mammary glands. Infection is reactivated during pregnancy and after radiotherapy in mouse models. High concentrations of *C. burnetii* are found in the placenta. At the time of parturition, the bacteria are released into the air, and infection follows inhalation of aerosolized organisms by a susceptible host. Windstorms can generate *C. burnetii* aerosols months after soil contamination that occurred during parturition. Individuals up to 18 km from the source have been infected. Because it is easily dispersed as an aerosol, *C. burnetii* is a potential agent of bioterrorism (Chap. S2), with a high infectivity rate and pneumonia as the major manifestation. Treatment with antibody to tumor necrosis factor α is a risk factor for development of Q fever.

Determining the source of an outbreak of Q fever can be challenging. An outbreak of Q fever at a horse-boarding ranch in Colorado in 2005 was due to spread of infection from two herds of goats that had been acquired by the owners. PCR testing confirmed the presence of *C. burnetii* in the soil and among the goats. Of 138 persons who lived within 1 mile of the ranch and who were also tested, 11 (8%) had evidence of *C. burnetii* infection, and 8 of these 11 individuals had no direct contact with the ranch.

Persons at risk for Q fever include abattoir workers, veterinarians, farmers, and other individuals who have contact with infected animals (particularly newborn animals) or products of conception. The organism is shed in milk for weeks to months after parturition. An outbreak of Q fever associated with ingestion of raw milk confirmed the oral route of transmission, although this route is uncommon. In rare instances, person-to-person transmission follows labor and childbirth in an infected woman, autopsy of an infected individual, or blood transfusion. Some evidence suggests that *C. burnetii* can be sexually transmitted among humans. Crushing an infected tick between the fingers has resulted in Q fever; the implication is that percutaneous transmission can occur. Some unusual modes of *C. burnetii* transmission to humans include treatment with live fetal sheep cells in Germany, which was responsible for cases in six persons (five from the United States and one from Canada).

 Infections due to *C. burnetii* occur in most geographic locations except New Zealand and Antarctica. Thus Q fever can be associated with travel. The number of reported cases of Q fever in the United States ranges from 28 to 54 per year. More than 70% of these cases occur in males, and April, May, and June are the most common months for acquisition. Q fever continues to be common in Australia, with 30 cases per 1 million population per year. Cases among abattoir workers in Australia declined dramatically as a result of a vaccination program. An outbreak of Q fever began in the Netherlands in 2007, and by 2010 more than 4000 cases had been reported. Pneumonia was a common manifestation in this outbreak. The outbreak was due to a combination of high-density goat farming in areas abutting large urban populations and environmental factors. Farms where spread did not occur had high vegetation densities and lower groundwater concentrations. Q fever is hyperendemic in Cayenne, capital of French Guiana, where it accounts for 24% of all cases of pneumonia.

The primary manifestations of acute Q fever differ geographically (e.g., pneumonia in Nova Scotia and granulomatous hepatitis in Marseille). These differences could reflect the route of infection (i.e., ingestion of contaminated milk for hepatitis and inhalation of contaminated aerosols for pneumonia) or strain differences. In the Netherlands outbreak, sequelae of infection in pregnant women were rare; this was not the case among pregnant women elsewhere.

Young age seems to be protective against disease caused by *C. burnetii*. In a large outbreak in Switzerland, symptomatic infection occurred five times more often among persons >15 years of age than among younger individuals. In many outbreaks, men are affected more commonly than women; the proposed explanation is that female hormones are partially protective.

Clinical Manifestations

ACUTE Q FEVER

After the usual incubation period of 3–30 days, 1070 patients with acute Q fever in southern France presented with hepatitis (40%), both pneumonia and hepatitis (20%), pneumonia (17%), isolated fever (14%), CNS involvement (2%), and pericarditis or myocarditis (1%). Acalculous cholecystitis, pancreatitis, lymphadenopathy, spontaneous rupture of the spleen, transient hypoplastic anemia, bone marrow necrosis, hemolytic anemia, histiocytic hemophagocytosis, optic neuritis, and erythema nodosum were less common manifestations.

The symptoms of acute Q fever are nonspecific; common among them are fever, extreme fatigue, photophobia, and severe headache that is frequently retro-orbital. Other symptoms include chills, sweats, nausea, vomiting, and diarrhea, each occurring in 5–20% of cases. Cough develops in about half of patients with Q fever pneumonia. Neurologic manifestations of acute Q fever are uncommon; however, in one outbreak in the United Kingdom, 23% of 102 patients had neurologic signs and symptoms as the major manifestation. A nonspecific rash may be evident in 4–18% of patients, and some patients have urticaria. The WBC count is usually normal. Thrombocytopenia occurs in ~25% of patients, and reactive thrombocytosis (with platelet counts exceeding $10^6/\mu L$) frequently develops during recovery. Chest radiography can show opacities similar to those seen in pneumonia caused by other pathogens, but multiple rounded opacities in patients in endemic areas suggest a diagnosis of Q fever pneumonia.

Acute Q fever occasionally complicates pregnancy. In one series, it resulted in premature birth in 35% of cases and in abortion or neonatal death in 43%. Neonatal death (previous or current) and lower infant birth weight are three times more likely among women seropositive for *C. burnetii*.

POST–Q FEVER FATIGUE SYNDROME

Prolonged fatigue can follow Q fever in up to 20% of cases and can be accompanied by a constellation of symptoms, including headaches, sweats, arthralgia, myalgias, blurred vision, muscle fasciculations, and enlarged and painful lymph nodes. Long-term persistence of a noninfective, nonbiodegraded complex of *Coxiella* cell components, with its antigens and specific lipopolysaccharide, has been detected in the affected persons. Patients who develop this syndrome have a higher frequency of carriage of HLA-DRB1*11 and of the 2/2 genotype of the interferon γ intron 1 microsatellite. When patients with Q fever fatigue syndrome were compared with those with chronic fatigue syndrome, the former patients were less likely to be female and less likely to have been treated for depression. Fatigue severity was the

same in both groups, and there were no differences in the presence of inflammatory markers in the two groups. Cognitive-based therapy shows some promise in patients with Q fever fatigue syndrome.

CHRONIC Q FEVER

Although it has recently been proposed that this entity be renamed *persistent Q fever*, we prefer the term *chronic Q fever*. Chronic Q fever most frequently is manifested as endocarditis and usually occurs in patients with previous valvular heart disease, immunosuppression, or chronic renal insufficiency. Fever is frequently absent or low grade. Valvular vegetations are detected in only 12% of patients with Q fever endocarditis by transthoracic echocardiography, but the rate of detection is higher (21–50%) with transesophageal echocardiography. The vegetations in chronic Q fever endocarditis differ from those in bacterial endocarditis, manifesting as endothelium-covered nodules on the valves. A high index of suspicion is necessary for timely diagnosis. Patients with chronic Q fever are often ill for >1 year before the diagnosis is made. The disease should be suspected in all patients with culture-negative endocarditis. In addition, all patients with valvular heart disease and an unexplained purpuric eruption, renal insufficiency, stroke, and/or progressive heart failure should be tested for *C. burnetii* infection. Patients with chronic Q fever have hepatomegaly and/or splenomegaly, which—in combination with rheumatoid factor, elevated erythrocyte sedimentation rate, high C-reactive protein level, and/or increased γ-globulin concentrations (up to 60–70 g/L)—suggests this diagnosis. Other manifestations of chronic Q fever include infection of vascular prostheses, infection of large-vessel aneurysms, lymphadenitis, bone infection, and chronic sternal wound infection. Unusual manifestations include chronic thrombocytopenia, mixed cryoglobulinemia, and livedo reticularis.

Diagnosis

Isolation of *C. burnetii* from buffy-coat blood samples or tissue specimens by a shell-vial technique is easy but requires a biosafety level 3 laboratory. PCR detects *C. burnetii* DNA in tissue specimens, including paraffin-embedded samples. Serology is the most commonly used diagnostic tool. Indirect immunofluorescence is sensitive and specific and is the method of choice. Rheumatoid factor should be adsorbed from the specimen before testing. With chronic infection, the titer to phase I antigen is usually much higher than that to phase II antigen (i.e., *C. burnetii* that has truncated lipopolysaccharide associated with gene deletions during laboratory passages), and the diagnosis should not be based on serology alone. Rather, the entire clinical setting must be taken into consideration. An anti–phase I IgG titer of ≥6400 would be considered a major criterion for the diagnosis of chronic Q fever, while a titer of ≥800 but ≤6400 would be a minor criterion. In acute Q fever, a fourfold rise in titer can be demonstrated between acute- and convalescent-phase serum samples. The phase II antibody titer is higher than the phase I antibody titer in acute Q fever.

Fluorodeoxyglucose positron emission tomography combined with CT (FDG-PET/CT) is useful in localizing the site of infection in chronic Q fever because it can detect not only valvular infection but also intravascular infection elsewhere, osteomyelitis, and lymphadenitis.

TREATMENT

TREATMENT

Q Fever

ANTIBIOTICS

Determining the antimicrobial susceptibility of intracellular microorganisms such as *C. burnetii* poses inherent methodologic difficulties. In general, *C. burnetii* is susceptible to tetracyclines, trimethoprim-sulfamethoxazole, and quinolones. In some areas (e.g., French Guiana), all isolates are resistant to erythromycin and azithromycin, and one of six isolates from French Guiana was resistant to telithromycin. There has been one report of the emergence of resistance to doxycycline during therapy for Q fever endocarditis.

Treatment of acute Q fever with doxycycline (100 mg twice daily for 14 days) is usually successful. Quinolones also are effective. When Q fever is diagnosed during pregnancy, treatment with trimethoprim-sulfamethoxazole (TMP-SMX) is recommended for the duration of the pregnancy; because TMP is a folic acid antagonist, folic acid supplementation should be given, especially to pregnant patients in the first trimester. One study showed no intrauterine fetal deaths and substantial reduction of obstetric complications in a group of Q fever patients treated with TMP-SMX.

The treatment of chronic Q fever is difficult and requires careful follow-up. Addition of hydroxychloroquine (to alkalinize the phagolysosome) renders doxycycline bactericidal against *C. burnetii*, and this combination is currently the favored regimen. Treatment for 18 months with doxycycline (100 mg twice daily) and hydroxychloroquine (200 mg three times daily, with the plasma concentration maintained at 0.8–1.2 μg/mL) is superior to a regimen of doxycycline and ofloxacin. Among 21 patients who received doxycycline and

hydroxychloroquine, one died of a surgical complication, two were still being treated at the end of the study, one was still being evaluated, and 17 had their infections cured. The mean duration of treatment was 31 months. In the ofloxacin and doxycycline group of 14 patients, one had died, one was still being treated, seven had experienced relapse, and five had been cured by the end of the study. Optimal management of Q fever endocarditis entails determination of the minimal inhibitory concentration (MIC) of doxycycline for the patient's isolate and measurement of serum doxycycline levels. A serum level–to–doxycycline MIC ratio of ≥1 is associated with a rapid decline in phase I antibodies with the doxycycline-hydroxychloroquine regimen. Patients treated with this regimen must be advised about photosensitivity and retinal toxicity risks. The doxycycline-hydroxychloroquine regimen was successful in one patient with HIV infection and Q fever endocarditis. The Jarisch-Herxheimer reaction occasionally complicates the treatment of chronic Q fever. Treatment of *C. burnetii*–infected aortic aneurysms is the same as that for Q fever endocarditis. Surgical intervention is often required.

If doxycycline-hydroxychloroquine cannot be used, the regimen chosen should include at least two antibiotics active against *C. burnetii*. Rifampin (300 mg once daily) combined with doxycycline (100 mg twice daily) or ciprofloxacin (750 mg twice daily) has been used successfully. The management of patients with Q fever endocarditis is complex and should preferably be undertaken by individuals with experience in managing this illness. Monitoring of antibody titers on a quarterly basis is an essential part of the management of these patients. Thus the laboratory should be contacted and asked to save all serum samples from such patients so that the current sample can be run with the previous one. There is incomplete agreement on the antibody titer at which therapy can be stopped. However, it is reasonable to discontinue treatment if levels of IgG antibody to phase I antigen have decreased by fourfold at 1 year, if IgM antibody to phase II antigen has disappeared, and if the patient is clinically stable.

Patients with acute Q fever and lesions of native heart valves (e.g., the bicuspid aortic valve), prosthetic valves, or prosthetic intravascular material should undergo serologic monitoring every 4 months for 2 years. If the phase I IgG titer is >800, further investigation is warranted. Some authorities recommend that patients with valvulopathy and acute Q fever receive doxycycline and hydroxychloroquine to prevent chronic Q fever. For women who exhibit a serologic profile of chronic Q fever after childbirth, hydroxychloroquine and doxycycline should be given for 1 year.

BIOLOGIC MODIFYING AGENTS

Interferon γ was successful in the treatment of a 3-year-old boy with prolonged fever, abdominal pain, and thrombocytopenia due to *C. burnetii* that had not been eradicated with conventional antibiotic therapy. Many patients with granulomatous hepatitis due to Q fever have a prolonged febrile illness that is unresponsive to antibiotics. For these individuals, treatment with prednisone (0.5 mg/kg) has resulted in defervescence within 2–15 days. After defervescence, the glucocorticoid dose is tapered over the next month.

Prevention

A whole-cell vaccine (Q-Vax) licensed in Australia effectively prevents Q fever in abattoir workers. Before administration of the vaccine, skin testing with intradermal diluted *C. burnetii* vaccine is performed, serologic testing is undertaken, and a history of possible Q fever is sought. Vaccine is given only to patients with no history of Q fever and negative results in serologic and skin testing.

Good animal-husbandry practices are important in preventing widespread contamination of the environment by *C. burnetii*. These practices include isolating aborting animals for up to 14 days, raising feed bunks to prevent contamination of feed by excreta, destroying aborted materials (by burning and burying fetal membranes and stillborn animals), and wearing masks and gloves when handling aborted materials. Vaccination of sheep and goats and a culling program were effective in the Netherlands outbreak. Only seronegative pregnant animals should be used in research settings, and only seronegative animals should be permitted in petting zoos.

During an outbreak of Q fever and for 4 weeks after it ceases, blood donations should not be accepted from individuals who live in the affected area.

FURTHER READING

Cross R et al: Revisiting doxycycline in pregnancy and early childhood—Time to rebuild its reputation? Expert Opin Drug Saf 15:367, 2016.

Drexler NA et al: National surveillance of spotted fever group rickettsioses in the United States 2008–2012. Am J Trop Med Hyg 94:26, 2016.

Dumler JS: Clinical disease: Current treatment and new challenges, in *Intracellular Pathogens II: Rickettsiales*. GH Palmer, AF Azad (eds). Washington, DC, ASM Press, 2012, pp 1–39.

Eldin C et al: From Q fever to *Coxiella burnetii* infection: A paradigm change. Clin Microbiol Rev 30:115, 2017.

Paris DH et al: Unresolved problems related to scrub typhus: A seriously neglected life-threatening disease. Am J Trop Med Hyg 89:301, 2013.

Chapter 183: Infections Due to Mycoplasmas

R. Doug Hardy

INTRODUCTION

Mycoplasmas are prokaryotes of the class Mollicutes. Their size (150–350 nm) is closer to that of viruses than to that of typical bacteria. Unlike viruses, however, mycoplasmas grow in cell-free culture media; in fact, they are the smallest organisms capable of independent replication.


The entire genomes of many *Mycoplasma* species have been sequenced and have been found to be among the smallest of all prokaryotic genomes. Sequencing information for these genomes has helped define the minimal set of genes necessary for cellular life. The absence of genes related to the synthesis of amino acids, fatty acid metabolism, and cholesterol dictates the mycoplasmas' parasitic or saprophytic dependence on a host for exogenous nutrients and necessitates the use of complex fastidious media to culture these organisms. Mycoplasmas lack a cell wall and are bound only by a cell membrane. The absence of a cell wall explains the inactivity of β-lactam antibiotics (penicillins and cephalosporins) against infections caused by these organisms.

At least 13 *Mycoplasma* species, two *Acholeplasma* species and two *Ureaplasma* species have been isolated from humans. Most of these species are thought to be normal inhabitants of oral and urogenital mucous membranes. *M. pneumoniae, M. hominis, M. genitalium, U. urealyticum*, and *U. parvum* have been shown conclusively to be pathogenic in immunocompetent humans. *M. pneumoniae* primarily infects the respiratory tract, while *M. hominis, M. genitalium, U. urealyticum*, and *U. parvum* are associated with a variety of genitourinary tract disorders and neonatal infections. Other mycoplasmas may cause disease in immunocompromised persons.

MYCOPLASMA PNEUMONIAE

PATHOGENESIS

M. pneumoniae is generally thought to act as an extracellular pathogen. Although the organism has been shown to exist and replicate within human cells, it is not known whether these intracellular events contribute to the pathogenesis of disease. *M. pneumoniae* attaches to ciliated respiratory epithelial cells by means of a complex terminal organelle at the tip of one end of the organism. Cytoadherence is mediated by interactive

adhesins and accessory proteins clustered on this organelle. After extracellular attachment, *M. pneumoniae* causes injury to host respiratory tissue. The mechanism of injury is thought to be mediated by the production of hydrogen peroxide and of an ADP-ribosylating and vacuolating cytotoxin of *M. pneumoniae* that has many similarities to pertussis toxin. Because mycoplasmas lack a cell wall, they also lack cell wall–derived stimulators of the innate immune system, such as lipopolysaccharide, lipoteichoic acid, and murein (peptidoglycan) fragments. However, lipoproteins from the mycoplasmal cell membrane appear to have inflammatory properties, probably acting through Toll-like receptors (primarily TLR2) on macrophages and other cells. Lung biopsy specimens from patients with *M. pneumoniae* respiratory tract infection reveal an inflammatory process involving the trachea, bronchioles, and peribronchial tissue, with a monocytic infiltrate that coincides with a luminal exudate of polymorphonuclear leukocytes.

Experimental evidence indicates that innate immunity provides most of the host's defense against mycoplasmal infection in the lungs, whereas cellular immunity may actually play an immunopathogenic role, exacerbating mycoplasmal lung disease. Humoral immunity appears to provide protection against dissemination of *M. pneumoniae* infection; patients with humoral immunodeficiencies do not have more severe lung disease than do immunocompetent patients in the early stages of infection but more often develop disseminated infection resulting in syndromes such as arthritis, meningitis, and osteomyelitis. The immunity that follows severe *M. pneumoniae* infections is more protective and longer-lasting than that following mild infections. Genuine second attacks of *M. pneumoniae* pneumonia have been reported infrequently.

EPIDEMIOLOGY

M. pneumoniae infection occurs worldwide. It is likely that the incidence of upper respiratory illness due to *M. pneumoniae* is up to 20 times that of pneumonia caused by this organism. Infection is spread from one person to another by respiratory droplets expectorated during coughing and results in clinically apparent disease in an estimated 80% of cases. The incubation period for *M. pneumoniae* is 2–4 weeks; therefore, the time-course of infection in a specific population may be several weeks long. Intrafamilial attack rates are as high as 84% among children and 41% among adults. Outbreaks of *M. pneumoniae* illness often occur in institutional settings such as military bases, boarding schools, and summer camps. Infections tend to be endemic, with sporadic epidemics every 4–7 years.

Most significantly, *M. pneumoniae* is a major cause of community-acquired respiratory illness in both children and adults and is often grouped with *Chlamydia pneumoniae* and *Legionella* species as one of the most important bacterial causes of "atypical" community-acquired pneumonia. For community-acquired pneumonia in adults, *M. pneumoniae* is the most frequently detected "atypical" organism. Analysis of 13 studies of community-acquired pneumonia published since 1995 (which included 6207 ambulatory and hospitalized adults) showed that the overall prevalence of *M. pneumoniae* was 22.7%; by comparison, the prevalence of *C. pneumoniae* was 11.7%, and that of *Legionella* species was 4.6%. *M. pneumoniae* pneumonia is also referred to as Eaton agent pneumonia (the organism having first been isolated in the early 1940s by Monroe Eaton), primary atypical pneumonia, and "walking" pneumonia.

CLINICAL MANIFESTATIONS

Upper Respiratory Tract Infections and Pneumonia

Acute *M. pneumoniae* infections generally manifest as pharyngitis, tracheobronchitis, reactive airway disease/wheezing, or a nonspecific upper respiratory syndrome. Little evidence supports the commonly held belief that this organism is an important cause of otitis media, with or without bullous myringitis. Pneumonia develops in 3–13% of infected individuals; its onset is usually gradual, occurring over several days, but may be more abrupt. Although *Mycoplasma* pneumonia may begin with a sore throat, the most common presenting symptom is cough. The cough is typically nonproductive, but some patients produce sputum. Headache, malaise, chills, and fever are noted in the majority of patients.

On physical examination, wheezes or rales are detected in ~80% of patients with *M. pneumoniae* pneumonia. In many patients, however, pneumonia can be diagnosed only by chest radiography. The most common radiographic pattern is that of peribronchial pneumonia with thickened bronchial markings, streaks of interstitial infiltration, and areas of subsegmental atelectasis. Segmental or lobar consolidation is not uncommon. While clinically evident pleural effusions are infrequent, lateral decubitus views reveal that up to 20% of patients have pleural effusions.

Overall, the clinical presentation of pneumonia in an individual patient is not useful for differentiating *M. pneumoniae* pneumonia from other types of community-acquired pneumonia. The possibility of *M. pneumoniae* infection deserves particular consideration when community-acquired pneumonia fails to respond to treatment with a penicillin or a cephalosporin—antibiotics that are ineffective against mycoplasmas. Symptoms usually resolve within 2–3 weeks after the onset of illness. Although *M. pneumoniae* pneumonia is generally self-limited, appropriate antimicrobial therapy significantly shortens the duration of clinical illness. Infection uncommonly results in critical illness and only rarely in death. In some patients, long-term recurrent wheezing or reactive airway disease may follow the resolution of acute pneumonia. The significance of chronic infection, especially as it relates to asthma, is an area of active investigation.

Extrapulmonary Manifestations

An array of extrapulmonary manifestations may develop during *M. pneumoniae* infection. The most significant are neurologic, dermatologic, cardiac, rheumatologic, and hematologic in nature. Extrapulmonary manifestations can be a result of disseminated infection, especially in patients with humoral immunodeficiencies (e.g., septic arthritis); postinfectious autoimmune phenomena (e.g., Guillain-Barré syndrome); or possibly ADP-ribosylating toxin. Overall, these manifestations are uncommon, given the frequency of *M. pneumoniae* infection. Notably, many patients with extrapulmonary *M. pneumoniae* disease do not have respiratory disease.

Skin eruptions described with *M. pneumoniae* infection include erythematous (macular or maculopapular), vesicular, bullous, petechial, and urticarial rashes. In some reports, 17% of patients with *M. pneumoniae*

pneumonia have had an exanthem. Erythema multiforme major (Stevens-Johnson syndrome) is the most clinically significant skin eruption associated with *M. pneumoniae* infection; it appears to occur more commonly with *M. pneumoniae* than with other infectious agents.

A wide spectrum of neurologic manifestations has been reported with *M. pneumoniae* infection. The most common are meningoencephalitis, encephalitis, Guillain-Barré syndrome, and aseptic meningitis. *M. pneumoniae* has been implicated as a likely etiologic agent in 5–7% of cases of encephalitis. Other neurologic manifestations may include cranial neuropathy, acute psychosis, cerebellar ataxia, acute demyelinating encephalomyelitis, cerebrovascular thromboembolic events, and transverse myelitis.

Hematologic manifestations of *M. pneumoniae* infection include hemolytic anemia, aplastic anemia, cold agglutinins, disseminated intravascular coagulation, and hypercoagulopathy. When anemia does occur, it generally develops in the second or third week of illness.

In addition, hepatitis, glomerulonephritis, pancreatitis, myocarditis, pericarditis, rhabdomyolysis, and arthritis (septic and reactive) have been convincingly ascribed to *M. pneumoniae* infection. Septic arthritis has been described most commonly in hypogammaglobulinemic patients.

DIAGNOSIS

Clinical findings, nonmicrobiologic laboratory tests, and chest radiography are not useful for differentiating *M. pneumoniae* pneumonia from other types of community-acquired pneumonia. In addition, since *M. pneumoniae* lacks a cell wall, it is not visible on Gram's stain. Although of historical interest, the measurement of cold agglutinin titers is no longer recommended for the diagnosis of *M. pneumoniae* infection because the findings are nonspecific and assays specific for *M. pneumoniae* are now available.

Acute *M. pneumoniae* infection can be diagnosed by polymerase chain reaction (PCR) detection of the organism in respiratory tract secretions or by isolation of the organism in culture **(Table 183-1)**. Oropharyngeal, nasopharyngeal, and pulmonary specimens are all acceptable for diagnosing *M. pneumoniae* pneumonia. Other bodily fluids, such cerebrospinal fluid, are acceptable for extrapulmonary infection. *M. pneumoniae* culture (which requires special media) is not recommended for routine diagnosis because the organism may take weeks to grow and is often difficult to isolate from clinical specimens. In contrast, PCR allows rapid, specific diagnosis earlier in the course of clinical illness.

TABLE 183-1

Diagnostic Tests for Respiratory *Mycoplasma pneumoniae* Infection[a]

Test	Sensitivity, %	Specificity, %
Respiratory culture	≤60	100
Respiratory PCR	65–90	90–100
Serologic studies[b]	55–100	55–100

[a]A combination of PCR and serology is suggested for routine diagnosis. If macrolide resistance is suspected, resistance testing by culture and/or PCR is available. [b]Acute- and convalescent-phase serum samples are recommended.

Abbreviation: PCR, polymerase chain reaction.

The diagnosis can also be established by serologic tests for IgM and IgG antibodies to *M. pneumoniae* in paired (acute- and convalescent-phase) serum samples; enzyme-linked immunoassay is the recommended serologic method. An acute-phase sample alone is not adequate for diagnosis, as antibodies to *M. pneumoniae* may not develop until 2 weeks into the illness; therefore, it is important to test paired samples. In addition, IgM antibody to *M. pneumoniae* can persist for up to 1 year after acute infection. Thus its presence may indicate recent rather than acute infection.

The combination of PCR of respiratory tract secretions and serologic testing constitutes the most sensitive and rapid approach to the diagnosis of *M. pneumoniae* infection.

TREATMENT

TREATMENT

Mycoplasma pneumoniae Infections

Although in the majority of untreated cases symptoms resolve within 2–3 weeks without significant associated morbidity, *M. pneumoniae* pneumonia can be a serious illness that responds to appropriate antimicrobial therapy **(Table 183-2)**. Randomized, double-blind, placebo-controlled trials in adults have demonstrated that antimicrobial treatment significantly decreases the duration of fever, cough, malaise, hospitalization, and radiologic abnormalities in *M. pneumoniae* pneumonia. Treatment options for acute *M. pneumoniae* infection include macrolides (e.g., oral azithromycin, 500 mg on day 1, then 250 mg/d on days 2–5), tetracyclines (e.g., oral doxycycline, 100 mg twice daily for 10–14 days), and respiratory fluoroquinolones. However, ciprofloxacin and ofloxacin are *not* recommended because of their high minimal

inhibitory concentrations against *M. pneumoniae* isolates and their poor performance in experimental studies. A 10- to 14-day course of quinolone therapy appears adequate.

 In Japan and China, very high levels (up to ≥90%) of *M. pneumoniae* resistance to macrolides have been reported. In Europe and to a lesser degree in the United States, macrolide-resistant *M. pneumoniae* is emerging. In investigated outbreaks of respiratory illness due to *M. pneumoniae* in the United States, macrolide resistance has been reported in 8–27% of isolates. Clinical studies have demonstrated that, when treated with macrolides, patients with community-acquired pneumonia due to macrolide-resistant *M. pneumoniae* experience a significantly longer duration of symptoms than do patients infected with macrolide-sensitive organisms; thus macrolide resistance in *M. pneumoniae* does appear to have clinical significance. If macrolide resistance is prominent in a particular geographic locale or is suspected, then a nonmacrolide antibiotic should be considered for treatment; in addition, in these instances, a respiratory sample may be sent to a mycoplasma reference laboratory for the detection of macrolide resistance by culture or PCR.

Clinical observations and experimental data suggest that the addition of glucocorticoids to an antibiotic regimen may be of value for the treatment of severe or refractory *M. pneumoniae* pneumonia. However, relevant clinical experience is limited. Even though appropriate antibiotic therapy significantly reduces the duration of respiratory illness, it does not appear to shorten the duration of detection of *M. pneumoniae* by culture or PCR; therefore, a test of cure or eradication is not suggested.

The roles of antimicrobial drugs, glucocorticoids, and IV immunoglobulin in the treatment of neurologic disease due to *M. pneumoniae* remain unknown.

TABLE 183-2

Antimicrobial Agents of Choice for *Mycoplasma* Infections[a]

Organism(s)	Drugs
Mycoplasma pneumoniae	Azithromycin, clarithromycin, erythromycin, doxycycline, levofloxacin, moxifloxacin, gemifloxacin (*not* ciprofloxacin or ofloxacin)
Ureaplasma urealyticum, Ureaplasma parvum	Azithromycin, clarithromycin, erythromycin, doxycycline
Mycoplasma hominis	Doxycycline, clindamycin
Mycoplasma genitalium	Azithromycin, moxifloxacin

[a]Antimicrobial resistance has been reported in mycoplasmas, as described in the text.

UROGENITAL MYCOPLASMAS

(See Also Chap. 131)

EPIDEMIOLOGY

M. hominis, M. genitalium, U. urealyticum, and *U. parvum* can cause urogenital tract disease. The significance of isolation of these organisms in a variety of other syndromes is unknown and in some cases is being investigated. *M. fermentans* has not been shown convincingly to cause human disease.

While urogenital mycoplasmas may be transmitted to a fetus during passage through a colonized birth canal, sexual contact is the major mode of transmission, and the risk of colonization increases dramatically with increasing numbers of sexual partners. In asymptomatic women, these mycoplasmas may be found throughout the lower urogenital tract. The vagina yields the largest number of organisms; next most densely colonized are the periurethral area and the cervix. Ureaplasmas are isolated less often from urine than from the cervix, but *M. hominis* is found with approximately the same frequency at these two sites. Ureaplasmas are isolated from the vagina of 40–80% of sexually active, asymptomatic women and *M. hominis* from 21–70%. The two microorganisms are found concurrently in 31–60% of women. In men, colonization with each organism is less prevalent. Mycoplasmas have been isolated from urine, semen, and the distal urethra of asymptomatic men.

CLINICAL MANIFESTATIONS

Urethritis, Pyelonephritis, and Urinary Calculi

In many episodes of *Chlamydia*-negative nongonococcal urethritis, ureaplasmas may be the causative agent. These organisms may also cause chronic voiding symptoms in women. The common presence of ureaplasmas in the urethra of asymptomatic men may suggest either that only certain serovars are pathogenic or that predisposing factors, such as lack of immunity, must exist in persons who develop symptomatic infection. Alternatively, disease may develop only upon initial exposure to ureaplasmas. Ureaplasmas have been implicated in epididymitis. *M. genitalium* also appears to cause urethritis. *M. genitalium* and ureaplasmas do not have a known role in prostatitis. *M. hominis* does not appear to play a primary etiologic role in urethritis, epididymitis, or prostatitis.

Evidence suggests that *M. hominis* causes up to 5% of cases of acute pyelonephritis. Ureaplasmas have not been associated with this disease.

Ureaplasmas play a limited role in the production of urinary calculi. The frequency with which ureaplasmas reach the kidney, the predisposing factors that allow them to do so, and the relative frequency of urinary tract calculi induced by this organism (compared with other organisms) are not known.

Pelvic Inflammatory Disease

M. hominis can cause pelvic inflammatory disease. In most episodes, *M. hominis* occurs as part of a polymicrobial infection, but the organism may play an independent role in a limited number of cases. Data also support an association of *M. genitalium* with pelvic inflammatory disease. Ureaplasmas are not thought to cause pelvic inflammatory disease.

Postpartum and Postabortal Infection

Studies implicate *M. hominis* as the primary pathogen in ~5–10% of women who have postpartum or postabortal fever; ureaplasmas have been implicated to a lesser degree. These infections are generally self-limited; however, if symptoms persist, specific antimicrobial therapy should be given. Ureaplasmas also appear to play a role in occasional postcesarean wound infections.

Nonurogenital Infection

In rare instances, *M. hominis* causes nonurogenital infections, such as brain abscess, wound infection, poststernotomy mediastinitis, endocarditis, and neonatal meningitis. These infections are most common among immunocompromised and hypogammaglobulinemic patients. Ureaplasmas and *M. hominis* can cause septic arthritis in immunodeficient patients. Ureaplasmas probably cause neonatal pneumonitis; their possible causal role in the development of bronchopulmonary dysplasia—the chronic lung disease of premature infants—has been extensively investigated, with most studies indicating at least a significant association. It is unclear whether ureaplasmas and *M. hominis* cause infertility, spontaneous abortion, premature labor, low birth weight, or chorioamnionitis.

DIAGNOSIS

Culture and PCR are both appropriate methods for the isolation of urogenital mycoplasmas. Culture of these organisms, however, requires special techniques and media that generally are available only at larger medical centers and reference laboratories. Serologic testing is not recommended for the clinical diagnosis of urogenital *Mycoplasma* infections.

TREATMENT

TREATMENT

Urogenital *Mycoplasma* Infections

Because colonization with urogenital mycoplasmas is common, it appears at present that their isolation from the urogenital tract in the absence of disease generally does not warrant treatment. Macrolides and doxycycline are considered the antimicrobial agents of choice for *Ureaplasma* infections (Table 183-2). *Ureaplasma* resistance to macrolides, doxycycline, quinolones, and chloramphenicol has been reported. *M. hominis* is resistant to macrolides. Doxycycline is generally the drug of choice for *M. hominis* infections, although resistance has been reported. Clindamycin is generally active against *M. hominis*. Quinolones are

active in vitro against *M. hominis*. For *M. genitalium*, the initial treatment of choice appears to be azithromycin; moxifloxacin has been successfully used to treat *M. genitalium* resistant to azithromycin.

FURTHER READING

Getman D et al: *Mycoplasma genitalium* prevalence, coinfection, and macrolide antibiotic resistance frequency in a multicenter clinical study cohort in the United States. J Clin Microbiol 54:2278, 2016.

Waites KB et al: *Mycoplasma pneumoniae* from the respiratory tract and beyond. Clin Microbiol Rev 30:747, 2017.

Workowski KA, Bolan GA: Sexually transmitted diseases treatment guidelines, 2015. MMWR Recomm Rep 64:1, 2015.

Zheng X et al: Macrolide-resistant *Mycoplasma pneumoniae*, United States. Emerg Infect Dis 21:1470, 2015.

Chapter 184: Chlamydial Infections

Charlotte A. Gaydos; Thomas C. Quinn

INTRODUCTION

Chlamydiae are obligate intracellular bacteria that cause a wide variety of diseases in humans and animals.

ETIOLOGIC AGENTS

The chlamydiae were originally classified as four species in the genus *Chlamydia*: *C. trachomatis, C. pneumoniae*, *C. psittaci*, and *C. pecorum* (the last species being found in ruminants). The *C. psittaci* group has been separated into three species: *C. psittaci*, *C. felis*, and *C. abortus.* The mouse pneumonitis strain (MoPn) is now classified as *C. muridarum*, and the guinea pig inclusion conjunctivitis strain (GPIC) is now designated *C. caviae*.

C. trachomatis is divided into two biovars: trachoma and LGV (lymphogranuloma venereum). The trachoma biovar causes two major types of disease in humans: ocular trachoma, the leading infectious cause of preventable blindness in the developing world; and urogenital infections, which are sexually or neonatally transmitted. The 18 serovars of *C. trachomatis* fall into three groups: the trachoma serovars A, B, Ba, and C; the oculogenital serovars D–K; and the LGV serovars L_1–L_3. Serovars can be distinguished by serologic typing with monoclonal antibodies or by molecular gene typing. However, serovar identification usually is not important clinically, since the antibiotic susceptibility pattern is the same for all three groups. The one exception applies when LGV is suspected on clinical grounds; in this situation, serovar determination is important because a longer treatment duration is required for LGV strains.

BIOLOGY, GROWTH CYCLE, AND PATHOGENESIS

BIOLOGY

During their intracellular growth, chlamydiae produce characteristic intracytoplasmic inclusions that can be visualized by direct fluorescent antibody or Giemsa staining of infected clinical material, such as conjunctival scrapings or cervical or urethral epithelial cells. Chlamydiae are nonmotile, gram-negative, obligate intracellular bacteria that replicate within the cytoplasm of host cells, forming the characteristic membrane-bound inclusions that are the basis for some diagnostic tests. Originally considered to be large viruses, chlamydiae differ from viruses in possessing RNA and DNA as well as a cell wall that is quite similar in

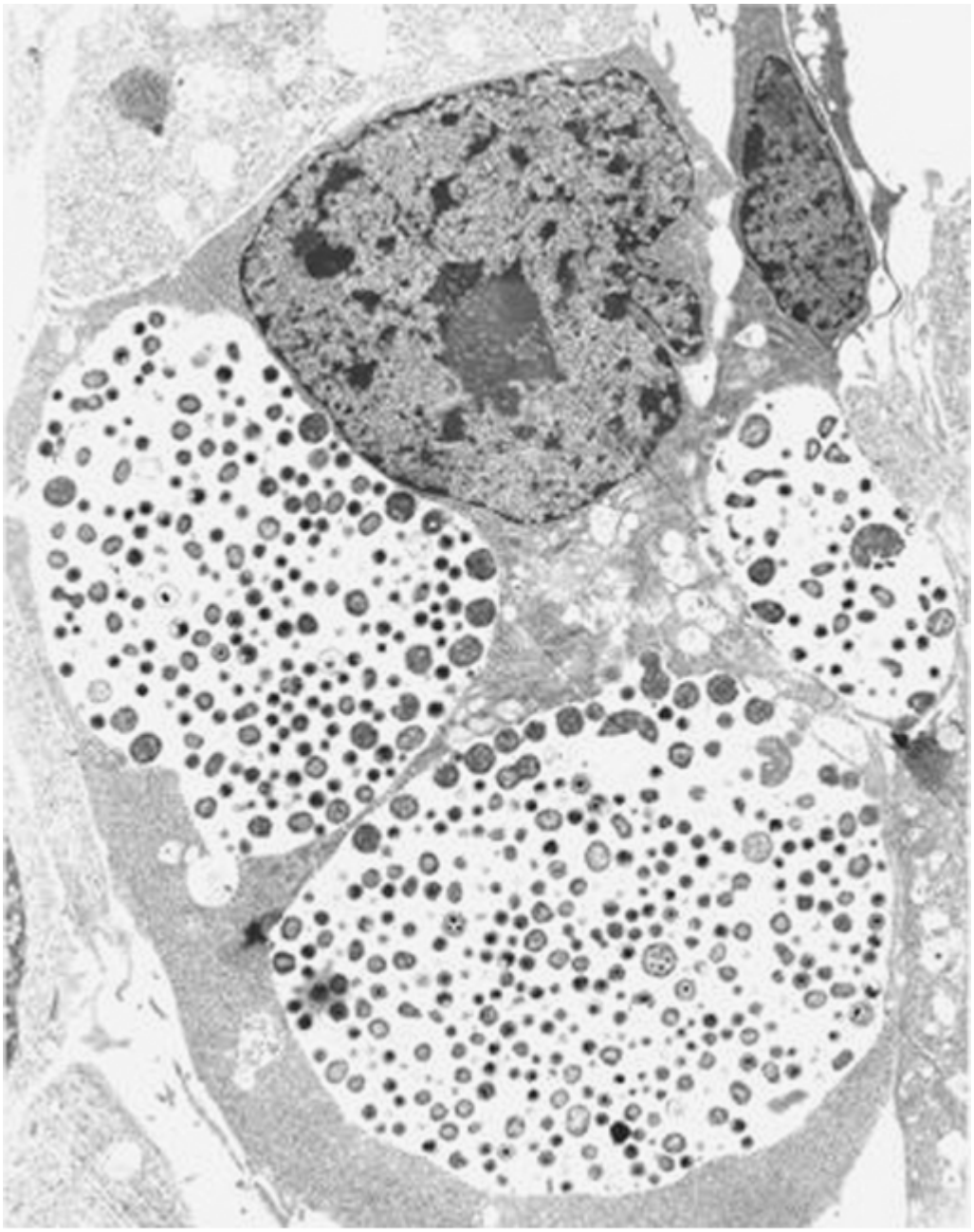
structure to the cell wall of typical gram-negative bacteria. However, chlamydiae lack peptidoglycan; their structural integrity depends on disulfide binding of outer-membrane proteins.

GROWTH CYCLE

Among the defining characteristics of chlamydiae is a unique growth cycle that involves alternation between two highly specialized morphologic forms (**Figs. 184-1 and 184-2**): the elementary body, which is the infectious form and is specifically adapted for extracellular survival, and the metabolically active and replicating reticulate body, which is not infectious, is adapted for an intracellular environment, and does not survive well outside the host cell. The biphasic growth cycle begins with attachment of the elementary body (diameter, 0.25–0.35 μm) at specific sites on the surface of the host cell. The elementary body enters the cell through a process similar to receptor-mediated endocytosis and resides in an inclusion, where the entire growth cycle is completed. The chlamydiae prevent phagosome–lysosome fusion. The inclusion membrane is modified by insertion of chlamydial antigens. Once the elementary body has entered the cell, it reorganizes into a reticulate body, which is larger (0.5–1 μm) and contains more RNA. After ~8 h, the reticulate body starts to divide by binary fission. The intracytoplasmic, membrane-bound inclusion body containing the reticulate bodies increases in size as the reticulate bodies multiply. Approximately 18–24 h after infection of the cell, these reticulate bodies begin to become elementary bodies by a reorganization or condensation process that is poorly understood. After rupture of the inclusion body, the elementary bodies are released to initiate another cycle of infection.

FIGURE 184-1

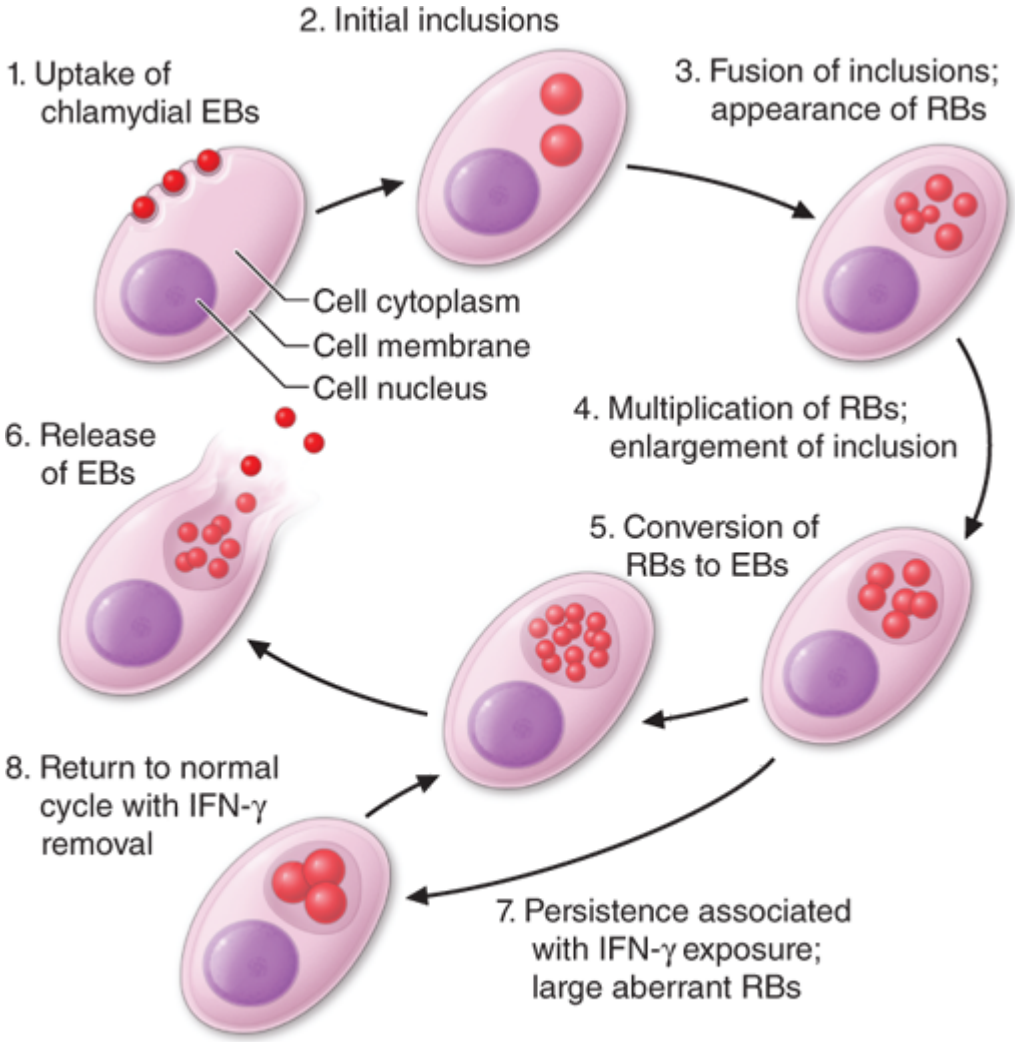
Chlamydial intracellular inclusions filled with smaller dense elementary bodies and larger reticulate bodies. *(Reprinted with permission from WE Stamm: Chlamydial infections, in Harrison's Principles of Internal Medicine, 17th ed, AS Fauci et al [eds]. New York, McGraw-Hill, 2008, p 1070.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 184-2

Chlamydial life cycle. EBs, elementary bodies; RBs, reticulate bodies; IFN-γ, interferon γ. *(Reprinted with permission from WE Stamm: Chlamydial infections, in Harrison's Principles of Internal Medicine, 17th ed, AS Fauci et al [eds]. New York, McGraw-Hill, 2008, p 1071.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Chlamydiae are susceptible to many broad-spectrum antibiotics and possess a number of enzymes, but they have a very restricted metabolic capacity. None of these metabolic reactions result in the production of energy. Chlamydiae have thus been considered to be energy parasites that use the ATP produced by the host cell for their own metabolic functions. Many aspects of chlamydial molecular biology are not well understood, but the sequencing of several chlamydial genomes and new proteomics research have provided researchers with many relevant tools for elucidating the biology of the life cycle.

PATHOGENESIS

Genital infections are mostly caused by *C. trachomatis* serovars D–K, with serovars D, E, and F involved most often. Molecular typing of the major outer-membrane protein gene (*omp1*) from which serovar differences arise has been used to demonstrate that polymorphisms can occur in isolates from patients who are exposed frequently to multiple infections, while less variation is observed in isolates from less sexually active populations. Polymorphisms in the major outer-membrane protein may provide antigenic variation, and the different forms allow persistence in the community because immunity to one is not protective against the others.

The trachoma biovar is essentially a parasite of squamocolumnar epithelial cells; the LGV biovar is more invasive and involves lymphoid cells. As is typical of chlamydiae, *C. trachomatis* strains are capable of causing chronic, clinically inapparent, asymptomatic infections. Because the duration of the chlamydial growth cycle is ~48–72 h, the incubation period of sexually transmitted chlamydial infections is relatively long—generally 1–3 weeks. *C. trachomatis* causes cell death as a result of its replicative cycle and can induce cell damage whenever it persists. However, few toxic effects are demonstrated, and cell death because of chlamydial replication is not sufficient to account for disease manifestations, the majority of which are due to immunopathologic mechanisms or nonspecific host responses to the organism or its by-products.

 In recent years, the entire genomes of various chlamydial species have been sequenced, the field of proteomics has become established, host innate immunity has been more precisely delineated, and innovative host cell–chlamydial interaction studies have been conducted. As a result, many insights have been gained into how chlamydiae adapt and replicate in their intracellular environment and produce disease. These insights into pathogenesis include information on the regulation of gene expression, protein localization, the type III secretion system, the roles of CD4+ and CD8+ T lymphocytes in the host response, and T lymphocyte trafficking.

The chlamydial heat-shock protein, which shares antigenic epitopes with similar proteins of other bacteria and with human heat-shock protein, may sensitize the host, and repeated infections may cause host cell damage. Persistent or recurrent chlamydial infections are associated with fibrosis, scarring, and complications following simple epithelial infections. A common endpoint of these late consequences is scarring of mucous membranes. Genital complications can lead to pelvic inflammatory disease (PID) and its late consequences of infertility, ectopic pregnancy, and chronic pelvic pain, while ocular infections may lead to blinding trachoma. High levels of antibody to human heat-shock protein have been associated with tubal factor infertility and ectopic pregnancy. Without adequate therapy, chlamydial infections may persist for several years, although symptoms—if present—usually abate.

Pathogenic mechanisms of *C. pneumoniae* have yet to be completely elucidated. The same is true for *C. psittaci*, except that this agent infects cells very efficiently and causes disease that may reflect direct cytopathic effects.

C. TRACHOMATIS INFECTIONS

GENITAL INFECTIONS

(See Also Chap. 131)

Spectrum

Although chlamydiae cause a number of human diseases, localized lower genital tract infections caused by *C. trachomatis* and the sequelae of such infections are the most important in terms of medical and economic

impact. Oculogenital infections due to *C. trachomatis* serovars D–K are transmitted during sexual contact or from mother to baby during childbirth and are associated with many syndromes, including cervicitis, salpingitis, acute urethral syndrome, endometritis, ectopic pregnancy, infertility, and PID in female patients; urethritis, proctitis, and epididymitis in male patients; and conjunctivitis and pneumonia in infants. Women bear the greatest burden of morbidity because of the serious sequelae of these infections. Untreated infections lead to PID, and multiple episodes of PID can lead to tubal factor infertility and chronic pelvic pain. Studies estimate that up to 80–90% of women and >50% of men with *C. trachomatis* genital infections lack symptoms; other patients have very mild symptoms. Thus, a large reservoir of infected persons continues to transmit infection to sexual partners.

As their designations reflect, the LGV serovars (L_1, L_2, and L_3) cause LGV, an invasive sexually transmitted disease (STD) characterized by acute lymphadenitis with bubo formation and/or acute hemorrhagic proctitis (see “LGV,” below).

Epidemiology

GLOBAL EPIDEMIOLOGY

C. trachomatis genital infections are global in distribution. The World Health Organization (WHO) estimated in 2008 that >106.4 million cases occur annually worldwide. This figure makes chlamydial infection the most prevalent bacterial sexually transmitted infection in the world. The associated morbidity is substantial, and the economic cost is high.

U.S. EPIDEMIOLOGY

In the United States, these infections are the most commonly reported of all infectious diseases. In 2015, 1,526,658 cases were reported to the U.S. Centers for Disease Control and Prevention (CDC); however, the CDC estimates that 2–3 million new cases occur per year, with substantial underreporting due to lack of screening in some populations. Rates of infection have increased every year; higher rates among women than among men reflect the focus on expansion of screening programs for women during the past 25 years. Use of increasingly sensitive diagnostic amplification tests, an increased emphasis on case reporting, and improvements in the information systems used have elevated the number of cases reported every year. The CDC and other professional organizations recommend annual screening of all sexually active women <25 years of age as well as rescreening of previously infected individuals at 3 months. The case count corresponds to 478.8 cases per 100,000 population, an increase of 5.9% compared with the rate in 2014. Young women have the highest infection rates (645.5 cases per 100,000)—more than twice the rate among men. Interestingly, with the increased availability of urine testing and extragenital testing, men—including gay, bisexual, and other men who have sex with men (MSM)—are increasingly being tested for chlamydial infection. From 2011 to 2015, rates of chlamydial infection in men increased by 20.0%, whereas rates in women rose by only 0.3% during this period. Chlamydial infection rates vary among different racial and ethnic minority populations. In 2015, rates among African Americans and American Indians/Alaska Natives were 5.9 and 3.8 times that among Caucasians, respectively. These disparities are important reflections of health inequities in the United States.

The aforementioned statistics are based on case reporting. Studies based on screening surveys estimate that the U.S. prevalence of *C. trachomatis* cervical infection is 5% among asymptomatic female college students and prenatal patients, >10% for women seen in family planning clinics, and >20% for women seen in STD clinics. The prevalence of genital *C. trachomatis* infections varies substantially by geographic locale, with the highest rates in the southeastern United States. The prevalence of *C. trachomatis* in the cervix of pregnant women is 5–10 times higher than that of *Neisseria gonorrhoeae*. The prevalence of genital infection with either agent is highest among women who are between the ages of 18 and 24, single, and non-Caucasian. Recurrent infections are common in these same risk groups and are often acquired from untreated sexual partners. The use of oral contraception and the presence of cervical ectopy also confer an increased risk. The proportion of infections that are asymptomatic appears to be higher for *C. trachomatis* than for *N. gonorrhoeae*, and symptomatic *C. trachomatis* infections are clinically less severe. Mild or asymptomatic *C. trachomatis* infections of the fallopian tubes nonetheless cause ongoing tubal damage and infertility. The costs of *C. trachomatis* infections and their complications to the U.S. health care system have recently been estimated to be >$516.7 million annually.

Clinical Manifestations

NONGONOCOCCAL AND POSTGONOCOCCAL URETHRITIS

C. trachomatis is the most common cause of nongonococcal urethritis (NGU) and postgonococcal urethritis (PGU). The designation *PGU* refers to NGU developing in men 2–3 weeks after treatment of gonococcal urethritis with single doses of agents such as penicillin or cephalosporins, which lack antimicrobial activity against chlamydiae. Current treatment regimens for gonorrhea have evolved and now include combination therapy with ceftriaxone and azithromycin; this current regimen is effective against concomitant chlamydial infection. Thus both the incidence of PGU and the causative role of *C. trachomatis* in this syndrome have declined.

In the United States, most of the estimated 2 million cases of acute urethritis are NGU, and *C. trachomatis* is implicated in 30–50% of these cases. The cause of most of the remaining cases of NGU is uncertain, but recent evidence suggests that *Mycoplasma genitalium, Trichomonas vaginalis*, and herpes simplex virus (HSV) cause some cases. The rate of involvement of *C. trachomatis* in urethral infection ranges from 3–7% among asymptomatic men to 15–20% among symptomatic men attending STD clinics. One recent multisite study of men in Baltimore, Seattle, Denver, and San Francisco reported an overall chlamydial prevalence of 7% in urine samples assessed by nucleic acid amplification tests (NAATs)—molecular tests that amplify the nucleic acids in clinical specimens. As in women, infection in men is age related, with young age as the greatest risk factor for chlamydial urethritis. The prevalence among men is highest at 20–24 years of age. In STD clinics, urethritis is usually less prevalent among MSM than among heterosexual men and is almost always much more common among black men than among white men. One study reported prevalences of 19 and 9% among nonwhite and white heterosexual men, respectively.

NGU is diagnosed by documentation of a leukocyte urethral exudate and by exclusion of gonorrhea by Gram's staining or culture. *C. trachomatis* urethritis is generally less severe than gonococcal urethritis, although in any individual patient these two forms of urethritis cannot reliably be differentiated solely on

clinical grounds. Symptoms include urethral discharge (often whitish and mucoid rather than frankly purulent), dysuria, and urethral itching. Physical examination may reveal meatal erythema and tenderness as well as a urethral exudate that is often demonstrable only by stripping of the urethra.

At least one-third of male patients with *C. trachomatis* urethral infection have no evident signs or symptoms of urethritis. The availability of NAATs for first-void urine specimens has facilitated broader-based testing for asymptomatic infection in male patients. As a result, asymptomatic chlamydial urethritis has been demonstrated in 5–10% of sexually active male adolescents screened at school-based clinics or community centers. Such patients generally have pyuria (≥15 leukocytes per 400× microscopic field in the sediment of first-void urine), a positive leukocyte esterase test, or an increased number of leukocytes on a Gram-stained smear prepared from a urogenital swab inserted 1–2 cm into the anterior urethra. When specific diagnostic tests for chlamydiae are not available, the examination of an endourethral specimen for increased leukocytes is useful in differentiating between true urethritis and functional symptoms in symptomatic patients or in making a presumptive diagnosis of *C. trachomatis* infection in high-risk but asymptomatic men (e.g., male patients in STD clinics, sex partners of women with nongonococcal salpingitis or mucopurulent cervicitis, fathers of children with inclusion conjunctivitis). Alternatively, urethritis can be assayed noninvasively by examination of a first-void urine sample for pyuria, either by microscopy or by the leukocyte esterase test. Urine (or a urethral swab) can also be tested directly for chlamydiae by DNA amplification methods (NAATs), as described below (see "Detection Methods").

EPIDIDYMITIS

Chlamydial urethritis may be followed by acute epididymitis, but this condition is rare, generally occurring in sexually active patients <35 years of age; in older men, epididymitis is usually associated with gram-negative bacterial infection and/or instrumentation procedures. An estimated 50–70% of cases of acute epididymitis are caused by *C. trachomatis*. The condition usually presents as unilateral scrotal pain with tenderness, swelling, and fever in a young man, often occurring in association with chlamydial urethritis. The illness may be mild enough to treat with oral antibiotics on an outpatient basis or severe enough to require hospitalization and parenteral therapy. Testicular torsion should be excluded promptly by radionuclide scan, Doppler flow study, or surgical exploration in a teenager or young adult who presents with acute unilateral testicular pain without urethritis. The possibility of testicular tumor or chronic infection (e.g., tuberculosis) should be excluded when a patient with unilateral intrascrotal pain and swelling does not respond to appropriate antimicrobial therapy.

REACTIVE ARTHRITIS

Reactive arthritis consists of conjunctivitis, urethritis (or, in female patients, cervicitis), arthritis, and characteristic mucocutaneous lesions. It may develop in 1–2% of cases of NGU and is thought to be the most common type of peripheral inflammatory arthritis in young men. *C. trachomatis* has been recovered from the urethra of 16–44% of patients with reactive arthritis and 69% of men who have signs of urogenital inflammation at the time of examination. Antibodies to *C. trachomatis* have also been detected in 46–67% of patients with reactive arthritis, and *Chlamydia*-specific cell-mediated immunity has been documented in 72%. In addition, *C. trachomatis* has been isolated from synovial biopsy samples from 15 of 29 patients in a number of small series and from a smaller proportion of synovial fluid specimens. Chlamydial nucleic acids

have been identified in synovial membranes and chlamydial elementary bodies in joint fluid. The pathogenesis of reactive arthritis is unclear, but this condition probably represents an abnormal host response to a number of infectious agents, including those associated with bacterial gastroenteritis (e.g., *Salmonella*, *Shigella*, *Yersinia*, or *Campylobacter*), or to infection with *C. trachomatis* or *N. gonorrhoeae*. Since >80% of affected patients have the HLA-B27 phenotype and since other mucosal infections produce an identical syndrome, chlamydial infection is thought to initiate an aberrant hyperreactive immune response that produces inflammation of the involved target organs in these genetically predisposed individuals. Evidence of exaggerated cell-mediated and humoral immune responses to chlamydial antigens in reactive arthritis supports this hypothesis. The finding of chlamydial elementary bodies and DNA in joint fluid and synovial tissue from patients with reactive arthritis suggests that chlamydiae may actually spread from genital to joint tissues in these patients—perhaps in macrophages.

NGU is the initial manifestation of reactive arthritis in 80% of patients, typically occurring within 14 days after sexual exposure. The urethritis may be mild and may even go unnoticed by the patient. Similarly, gonococcal urethritis may precede reactive arthritis, but co-infection with an agent of NGU is difficult to rule out. The urethral discharge may be purulent or mucopurulent, and patients may or may not report dysuria. Accompanying prostatitis, usually asymptomatic, has been described. Arthritis usually begins ~4 weeks after the onset of urethritis but may develop sooner or, in a small percentage of cases, may actually precede urethritis. The knees are most frequently involved; next most commonly affected are the ankles and small joints of the feet. Sacroiliitis, either symmetrical or asymmetrical, is documented in two-thirds of patients. Mild bilateral conjunctivitis, iritis, keratitis, or uveitis is sometimes present but lasts for only a few days. Finally, dermatologic manifestations occur in up to 50% of patients. The initial lesions—usually papules with a central yellow spot—most often involve the soles and palms and, in ~25% of patients, eventually epithelialize and thicken to produce keratoderma blenorrhagicum. Circinate balanitis is usually painless and occurs in fewer than half of patients. The initial episode of reactive arthritis usually lasts 2–6 months.

PROCTITIS

Primary anal or rectal infections with *C. trachomatis* have been described in women and MSM who practice anal intercourse. In these infections, rectal involvement is initially characterized by severe anorectal pain, a bloody mucopurulent discharge, and tenesmus. Oculogenital serovars D–K and LGV serovars L_1, L_2, and L_3 have been found to cause proctitis. The LGV serovars are far more invasive and cause much more severely symptomatic disease, including severe ulcerative proctocolitis that can be clinically confused with HSV proctitis. Histologically, LGV proctitis may resemble Crohn's disease in that giant cell formation and granulomas are detected. In the United States and Europe, cases of LGV proctitis occur almost exclusively in MSM, many of whom have HIV infection.

The less invasive non-LGV serovars of *C. trachomatis* cause mild proctitis. Many infected individuals are asymptomatic, and in these cases infection is diagnosed only by routine culture or NAAT of rectal swabs. The number of fecal leukocytes is usually abnormal in both asymptomatic and symptomatic cases. Sigmoidoscopy may yield normal findings or may reveal mild inflammatory changes or small erosions or follicles in the lower 10 cm of the rectum. Histologic examination of rectal biopsies generally shows anal crypts and prominent follicles as well as neutrophilic infiltration of the lamina propria. Chlamydial proctitis is

best diagnosed by isolation of *C. trachomatis* from the rectum and documentation of a response to appropriate therapy. NAATs are reportedly more sensitive than culture for diagnosis and are also specific.

MUCOPURULENT CERVICITIS

Although most women with chlamydial infections of the cervix have no symptoms, almost half generally have local signs of infection on examination. Cervicitis is usually characterized by the presence of a mucopurulent discharge, with >20 neutrophils per microscopic field visible in strands of cervical mucus in a thinly smeared, gram-stained preparation of endocervical exudate. Hypertrophic ectopy of the cervix may also be evident as an edematous area near the cervical os that is congested and bleeds easily on minor trauma (e.g., when a specimen is collected with a swab). A Papanicolaou smear shows increased numbers of neutrophils as well as a characteristic pattern of mononuclear inflammatory cells, including plasma cells, transformed lymphocytes, and histiocytes. Cervical biopsy shows a predominantly mononuclear cell infiltrate of the subepithelial stroma. Clinical experience and collaborative studies indicate that a cutoff of >30 polymorphonuclear leukocytes (PMNs)/1000× field in a gram-stained smear of cervical mucus correlates best with chlamydial or gonococcal cervicitis.

Clinical recognition of chlamydial cervicitis depends on a high index of suspicion and careful cervical examination. No genital symptoms are specifically correlated with chlamydial cervical infection. The differential diagnosis of a mucopurulent discharge from the endocervical canal in a young, sexually active woman includes gonococcal endocervicitis, salpingitis, endometritis, and intrauterine contraceptive device–induced inflammation. Diagnosis of cervicitis is based on the presence of PMNs on a cervical swab as noted above; the presence of chlamydiae is confirmed by either culture or NAAT.

PELVIC INFLAMMATORY DISEASE

Inflammation of sections of the fallopian tube is often referred to as salpingitis or PID. The proportion of acute salpingitis cases caused by *C. trachomatis* varies geographically and with the population studied. It has been estimated that *C. trachomatis* causes up to 50% of PID cases in the United States. PID occurs via ascending intraluminal spread of *C. trachomatis* or *N. gonorrhoeae* from the lower genital tract. Mucopurulent cervicitis is often followed by endometritis, endosalpingitis, and finally pelvic peritonitis. Evidence of mucopurulent cervicitis is often found in women with laparoscopically verified salpingitis. Similarly, endometritis, demonstrated by an endometrial biopsy showing plasma cell infiltration of the endometrial epithelium, is documented in most women with laparoscopy-verified chlamydial (or gonococcal) salpingitis. Chlamydial endometritis can also occur in the absence of clinical evidence of salpingitis. Histologic evidence of endometritis has been correlated with a syndrome consisting of vaginal bleeding, lower abdominal pain, and uterine tenderness in the absence of adnexal tenderness. Chlamydial salpingitis produces milder symptoms than gonococcal salpingitis and may be associated with less marked adnexal tenderness. Thus, mild adnexal or uterine tenderness in a sexually active woman with cervicitis suggests chlamydial PID.

Chronic untreated endometrial and tubal inflammation can result in tubal scarring, impaired tubal function, tubal occlusion, and infertility even among women who report no prior treatment for chlamydial infection. *C. trachomatis* has been particularly implicated in “subclinical” PID on the basis of a lack of history of PID among *Chlamydia*-seropositive women with tubal damage and detection of chlamydial DNA or antigen

among asymptomatic women with tubal infertility. These data suggest that the best method to prevent PID and its sequelae is surveillance and control of lower genital tract infections along with diagnosis and treatment of sex partners and prevention of reinfections. Promotion of early symptom recognition and health care presentation may reduce the frequency and severity of sequelae of PID.

PERIHEPATITIS

The Fitz-Hugh–Curtis syndrome was originally described as a complication of gonococcal PID. However, studies over the past several decades have suggested that chlamydial infection is more commonly associated with perihepatitis than is *N. gonorrhoeae*. Perihepatitis should be suspected in young, sexually active women who develop right-upper-quadrant pain, fever, or nausea. Evidence of salpingitis may or may not be found on examination. Frequently, perihepatitis is strongly associated with extensive tubal scarring, adhesions, and inflammation observed at laparoscopy, and high titers of antibody to the 57-kDa chlamydial heat-shock protein have been documented. Culture and/or serologic evidence of *C. trachomatis* is found in three-fourths of women with this syndrome.

URETHRAL SYNDROME IN WOMEN

In the absence of infection with uropathogens such as coliforms or *Staphylococcus saprophyticus*, *C. trachomatis* is the pathogen most commonly isolated from college women with dysuria, frequency, and pyuria. Screening studies can recover *C. trachomatis* at both the cervix and the urethra; in up to 25% of infected women, the organism is isolated only from the urethra. The urethral syndrome in women consists of dysuria and frequency in conjunction with chlamydial urethritis, pyuria, and no bacteriuria or urinary pathogens. Although symptoms of the urethral syndrome may develop in some women with chlamydial infection, the majority of women attending STD clinics for urethral chlamydial infection do not have dysuria or frequency. Even in women with chlamydial urethritis causing the acute urethral syndrome, signs of urethritis such as urethral discharge, meatal redness, and swelling are uncommon. However, mucopurulent cervicitis in a woman presenting with dysuria and frequency strongly suggests *C. trachomatis* urethritis. Other correlates of chlamydial urethral syndrome include a duration of dysuria of >7–10 days, lack of hematuria, and lack of suprapubic tenderness. Abnormal urethral Gram's stains showing >10 PMNs/1000× field in women with dysuria but without coliform bacteriuria support the diagnosis of chlamydial urethritis. Other possible diagnoses include gonococcal or trichomonal infection of the urethra.

INFECTION IN PREGNANCY AND THE NEONATAL PERIOD

Infections during pregnancy can be transmitted to infants during delivery. Approximately 20–30% of infants exposed to *C. trachomatis* in the birth canal develop conjunctivitis, and 10–15% subsequently develop pneumonia. Consequently, all newborn infants receive ocular prophylaxis at birth to prevent ophthalmia neonatorum. Without treatment, conjunctivitis usually develops at 5–19 days of life and often results in a profuse mucopurulent discharge. Roughly half of infected infants develop clinical evidence of inclusion conjunctivitis. However, it is impossible to differentiate chlamydial conjunctivitis from other forms of neonatal conjunctivitis (e.g., that due to *N. gonorrhoeae*, *Haemophilus influenzae*, *Streptococcus pneumoniae*, or HSV) on clinical grounds; thus laboratory diagnosis is required. Inclusions within epithelial cells are often detected in Giemsa-stained conjunctival smears, but these smears are considerably less sensitive than cultures or NAATs for chlamydiae. Gram-stained smears may show gonococci or occasional

small gram-negative coccobacilli in *Haemophilus* conjunctivitis, but smears should be accompanied by cultures or NAATs for these agents.

C. trachomatis has also been isolated frequently and persistently from the nasopharynx, rectum, and vagina of infected infants—occasionally for >1 year in the absence of treatment. In some cases, otitis media results from perinatally acquired chlamydial infection. Pneumonia may develop in infants from 2 weeks to 4 months of age. *C. trachomatis* is estimated to cause 20–30% of pneumonia cases in infants <6 months of age. Epidemiologic studies have linked chlamydial pulmonary infection in infants with increased occurrence of subacute lung disease (bronchitis, asthma, wheezing) in later childhood.

LYMPHOGRANULOMA VENEREUM



C. trachomatis serovars L_1, L_2, and L_3 cause LGV, an invasive systemic STD. The peak incidence of LGV corresponds with the age of greatest sexual activity: the second and third decades of life. The worldwide incidence of LGV is falling, but the disease is still endemic and a major cause of morbidity in parts of Asia, Africa, South America, and the Caribbean. LGV is rare in industrialized countries; for more than a decade, the reported incidence in the United States has been only 0.1 case per 100,000 population. In the Bahamas, an apparent outbreak of LGV was described in association with a concurrent increase in heterosexual infection with HIV. Reports of outbreaks with the newly identified variant L_{2b} in Europe, Australia, and the United States indicate that LGV is becoming more prevalent among MSM. These cases have usually presented as hemorrhagic proctocolitis in HIV-positive men. More widespread use of NAATs for identification of rectal infections may have enhanced case recognition.

LGV begins as a small painless papule that tends to ulcerate at the site of inoculation, often escaping attention. This primary lesion heals in a few days without scarring and is usually recognized as LGV only in retrospect. LGV strains of *C. trachomatis* have occasionally been recovered from genital ulcers and from the urethra of men and the endocervix of women who present with inguinal adenopathy; these areas may be the primary sites of infection in some cases. Proctitis is more common among people who practice receptive anal intercourse, and an elevated white blood cell count in anorectal smears may predict LGV in these patients. Ulcer formation may facilitate transmission of HIV infection and other sexually transmitted and blood-borne diseases.

As NAATs for *C. trachomatis* are being used more often, increasing numbers of cases of LGV proctitis are being recognized in MSM. Such patients present with anorectal pain and mucopurulent, bloody rectal discharge. Sigmoidoscopy reveals ulcerative proctitis or proctocolitis, with purulent exudate and mucosal bleeding. Histopathologic findings in the rectal mucosa include granulomas with giant cells, crypt abscesses, and extensive inflammation. These clinical, sigmoidoscopic, and histopathologic findings may closely resemble those of Crohn's disease of the rectum.

The most common presenting picture in heterosexual men and women is the *inguinal syndrome*, which is characterized by painful inguinal lymphadenopathy beginning 2–6 weeks after presumed exposure; in rare instances, the onset comes after a few months. The inguinal adenopathy is unilateral in two-thirds of cases, and palpable enlargement of the iliac and femoral nodes is often evident on the same side as the enlarged

inguinal nodes. The nodes are initially discrete, but progressive periadenitis results in a matted mass of nodes that becomes fluctuant and suppurative. The overlying skin becomes fixed, inflamed, and thin, and multiple draining fistulas finally develop. Extensive enlargement of chains of inguinal nodes above and below the inguinal ligament ("the sign of the groove") is not specific and, although not uncommon, is documented in only a minority of cases. Spontaneous healing usually takes place after several months; inguinal scars or granulomatous masses of various sizes persist for life. Massive pelvic lymphadenopathy may lead to exploratory laparotomy.

Constitutional symptoms are common during the stage of regional lymphadenopathy and, in cases of proctitis, may include fever, chills, headache, meningismus, anorexia, myalgias, and arthralgias. Other systemic complications are infrequent but include arthritis with sterile effusion, aseptic meningitis, meningoencephalitis, conjunctivitis, hepatitis, and erythema nodosum **(Fig. A1-39)**. Complications of untreated anorectal infection include perirectal abscess; anal fistulas; and rectovaginal, rectovesical, and ischiorectal fistulas. Secondary bacterial infection probably contributes to these complications. Rectal stricture is a late complication of anorectal infection and usually develops 2–6 cm from the anal orifice—i.e., at a site within reach on digital rectal examination. A small percentage of cases of LGV in men present as chronic progressive infiltrative, ulcerative, or fistular lesions of the penis, urethra, or scrotum. Associated lymphatic obstruction may produce elephantiasis. When urethral stricture occurs, it usually involves the posterior urethra and causes incontinence or difficulty with urination.

Diagnosis

DETECTION METHODS

Historically, chlamydiae were cultivated in the yolk sac of embryonated eggs. The organisms can be grown more easily in tissue culture, but cell culture—once considered the diagnostic gold standard—has been replaced by nonculture assays **(Table 184-1)**. In general, culture for chlamydiae in clinical specimens is now performed only in specialized laboratories. The first nonculture assays, such as direct fluorescent antibody staining of clinical material and enzyme immunoassay (EIA), have been replaced by NAATs, which are currently recommended by the CDC as the diagnostic assays of choice. At present, five NAAT assays cleared by the U.S. Food and Drug Administration (FDA) are commercially available, some of which are available as high-throughput robotic platforms. Point-of-care diagnostic assays are becoming available; they are of increasing interest since patients can potentially be treated before leaving the clinic.

TABLE 184-1

Diagnostic Tests for Sexually Transmitted and Perinatal *Chlamydia trachomatis* Infection

Infection	Suggestive Signs/Symptoms	Presumptive Diagnosis[a]	Confirmatory Test of Choice
Men			
NGU, PGU	Discharge, dysuria	Gram's stain with >4 neutrophils per oil-immersion field; no gonococci	Urine or urethral NAAT for *C. trachomatis*
Epididymitis	Unilateral intrascrotal swelling, pain, tenderness; fever; NGU	Gram's stain with >4 neutrophils per oil-immersion field; no gonococci; urinalysis with pyuria	Urine or urethral NAAT for *C. trachomatis*
Women			
Cervicitis	Mucopurulent cervical discharge, bleeding and edema of the zone of cervical ectopy	Cervical Gram's stain with ≥20 neutrophils per oil-immersion field in cervical mucus	Urine, cervical, or vaginal NAAT for *C. trachomatis*
Salpingitis	Lower abdominal pain, cervical motion tenderness, adnexal tenderness or masses	*C. trachomatis* always potentially present in salpingitis	Urine, cervical, or vaginal NAAT for *C. trachomatis*
Urethritis	Dysuria and frequency without hematuria	MPC; sterile pyuria; negative routine urine culture	Urine or urethral NAAT for *C. trachomatis*
Adults of Either Sex			

Infection	Suggestive Signs/Symptoms	Presumptive Diagnosis[a]	Confirmatory Test of Choice
Proctitis	Rectal pain, discharge, tenesmus, bleeding; history of receptive anorectal intercourse	Negative gonococcal culture and Gram's stain; at least 1 neutrophil per oil-immersion field in rectal Gram's stain	Rectal NAAT for *C. trachomatis* or culture
Reactive arthritis	NGU, arthritis, conjunctivitis, typical skin lesions	Gram's stain with >4 neutrophils per oil-immersion field; lack of gonococci indicative of NGU	Urine or urethral NAAT for *C. trachomatis*
LGV	Regional adenopathy, primary lesion, proctitis, systemic symptoms	None	Culture of LGV strain from node or rectum, occasionally from urethra or cervix; NAAT for *C. trachomatis* from these sites; LGV CF titer, ≥1:64; MIF titer, ≥1:512
Neonates			
Conjunctivitis	Purulent conjunctival discharge 6–18 days after delivery	Negative culture and Gram's stain for gonococci, *Haemophilus* spp., pneumococci, staphylococci	Conjunctival NAAT for *C. trachomatis*; FA-stained scraping of conjunctival material
Infant pneumonia	Afebrile, staccato cough, diffuse rales, bilateral hyperinflation, interstitial infiltrates	None	Chlamydial culture or NAAT of sputum, pharynx, eye, rectum; MIF antibody to *C. trachomatis*—fourfold change in IgG or IgM antibody titer

[a]A presumptive diagnosis of chlamydial infection is often made in the syndromes listed when gonococci are not found. A positive test for *Neisseria gonorrhoeae* does not exclude the involvement of *C. trachomatis*, which often is present in patients with gonorrhea.

Abbreviations: CF, complement-fixing; FA, fluorescent antibody; LGV, lymphogranuloma venereum; MIF, microimmunofluorescence; MPC, mucopurulent cervicitis; NAAT, nucleic acid amplification test; NGU, nongonococcal urethritis; PGU, postgonococcal urethritis.

Source: Reprinted with permission from WE Stamm: Chlamydial infections, in *Harrison's Principles of Internal Medicine*, 17th ed, AS Fauci et al (eds). New York, McGraw-Hill, 2008, p 1075.

CHOICE OF SPECIMEN

Cervical and urethral swabs have traditionally been used for the diagnosis of STDs in female and male patients, respectively. However, given the greatly increased sensitivity and specificity of NAATs, less invasive samples (e.g., urine for both sexes and vaginal swabs for women) can be used. For screening of asymptomatic women, the CDC now recommends that self-collected or clinician-collected vaginal swabs, which are slightly more sensitive than urine, be used. Urine screening tests are often used in outreach screening programs, however. For symptomatic women undergoing a pelvic examination, cervical swab samples are desirable because they have slightly higher chlamydial counts. For male patients, a urine specimen is the sample of choice, but self-collected penile-meatal swabs have been explored.

ALTERNATIVE SPECIMEN TYPES

Ocular samples from babies and adults can be assessed by NAATs. However, since commercial NAATs for this purpose have not yet been approved by the FDA, laboratories must perform their own verification studies. Samples from rectal and pharyngeal sites have been used successfully to detect chlamydiae by NAATs, but laboratories must perform validation studies to verify test performance.

OTHER DIAGNOSTIC ISSUES

Because NAATs detect nucleic acids instead of live organisms, they should be used with caution as test-of-cure assays. Residual nucleic acid from cells rendered noninfective by antibiotics may continue to yield a positive result in NAATs for as long as 3 weeks after therapy when viable organisms have actually been eradicated. Therefore, clinicians should not use NAATs for test of cure until after 3 weeks. The CDC currently does not recommend a test of cure after treatment for infection with *C. trachomatis*. However, because incidence studies have demonstrated that previous chlamydial infection increases the probability of becoming reinfected, the CDC does recommend that previously infected individuals be rescreened 3 months after treatment.

SEROLOGY

Serologic testing may be helpful in the diagnosis of LGV and neonatal pneumonia caused by *C. trachomatis.* The serologic test of choice is the microimmunofluorescence (MIF) test, in which high-titer purified elementary bodies mixed with embryonated chicken yolk sac material are affixed to a glass microscope slide to which dilutions of sera are applied. After incubation and washing, fluorescein-conjugated IgG or IgM antibody is applied. The test is read with an epifluorescence microscope, with the highest dilution of serum producing visible fluorescence designated as the titer. The MIF test is not widely available except in research laboratories and is highly labor intensive. Although the complement fixation (CF) test can also be used, it employs lipopolysaccharide (LPS) as the antigen and therefore identifies the pathogen only to the genus level. Single-point titers of >1:64 support a diagnosis of LGV, in which it is difficult to demonstrate rising antibody titers—i.e., paired serum samples are difficult to obtain since, by its very nature, the disease results

in the patient's being seen by the physician after the acute stage. Any antibody titer of above 1:16 is considered significant evidence of exposure to chlamydiae. However, serologic testing is never recommended for diagnosis of uncomplicated genital infections of the cervix, urethra, and lower genital tract or for *C. trachomatis* screening of asymptomatic individuals.

TREATMENT

TREATMENT

C. trachomatis Genital Infections

A 7-day course of tetracycline (500 mg four times daily), doxycycline (100 mg twice daily), erythromycin (500 mg four times daily), or a fluoroquinolone (ofloxacin, 300 mg twice daily; or levofloxacin, 500 mg/d) can be used for treatment of uncomplicated chlamydial infections. A single 1-g oral dose of azithromycin is as effective as a 7-day course of doxycycline for the treatment of uncomplicated genital *C. trachomatis* infections in adults. Azithromycin causes fewer adverse gastrointestinal reactions than do older macrolides such as erythromycin. The single-dose regimen of azithromycin has great appeal for the treatment of patients with uncomplicated chlamydial infection (especially those without symptoms and those with a likelihood of poor compliance) and of the sexual partners of infected patients. These advantages must be weighed against the considerably greater cost of azithromycin. Whenever possible, the single 1-g dose should be given as directly observed therapy. Although not approved by the FDA for use in pregnancy, this regimen appears to be safe and effective for this purpose. However, amoxicillin (500 mg three times daily for 7 days) can also be given to pregnant women. The fluoroquinolones are contraindicated in pregnancy. A 2-week course of treatment is recommended for complicated chlamydial infections (e.g., PID, epididymitis) and at least a 3-week course of doxycycline (100 mg orally twice daily) or erythromycin base (500 mg orally four times daily) for LGV. Failure of treatment with a tetracycline in genital infections usually indicates poor compliance or reinfection rather than involvement of a drug-resistant strain. To date, clinically significant drug resistance has not been observed in *C. trachomatis*.

Treatment or testing for chlamydiae should be considered among *N. gonorrhoeae*–infected patients because of the frequency of co-infection. Systemic treatment with erythromycin has been recommended for ophthalmia neonatorum and for *C. trachomatis* pneumonia in infants. For the treatment of adult inclusion conjunctivitis, a single 1-g dose of azithromycin was as effective as standard 10-day treatment with doxycycline. Recommended treatment regimens for both bubonic and anogenital LGV include tetracycline, doxycycline, or erythromycin for 21 days.

SEX PARTNERS

The continued high prevalence of chlamydial infections in most parts of the United States is due primarily to the failure to diagnose—and therefore treat—patients with symptomatic or asymptomatic infection and their sex partners. Urethral or cervical infection with *C. trachomatis* has been well documented in a high proportion of the sex partners of patients with NGU, epididymitis, reactive arthritis, salpingitis, and endocervicitis. If possible, confirmatory laboratory tests for chlamydiae should be undertaken in these individuals, but even persons without positive tests or evidence of clinical disease who have recently been exposed to proven or possible chlamydial infection (e.g., NGU) should be offered therapy. A novel approach

is partner-delivered therapy, in which infected patients receive treatment and are also provided with single-dose azithromycin to give to their sex partner(s).

NEONATES AND INFANTS

In neonates with conjunctivitis or infants with pneumonia, erythromycin ethylsuccinate or estolate can be given orally at a dosage of 50 mg/kg per day, preferably in four divided doses, for 2 weeks. Careful attention must be given to compliance with therapy—a frequent problem. Relapses of eye infection are common after topical treatment with erythromycin or tetracycline ophthalmic ointment and may also follow oral erythromycin therapy. Thus follow-up cultures should be performed after treatment. Both parents should be examined for *C. trachomatis* infection and, if diagnostic testing is not readily available, should be treated with doxycycline or azithromycin.

Prevention

Since many chlamydial infections are asymptomatic, effective control and prevention must involve periodic screening of individuals at risk. Selective cost-effective screening criteria have been developed. Among women, young age (generally <25 years) is a critical risk factor for chlamydial infections in nearly all studies. Other risk factors include mucopurulent cervicitis; multiple, new, or symptomatic male sex partners; and lack of barrier contraceptive use. In some settings, screening based on young age may be as sensitive as criteria that incorporate behavioral and clinical measures. Another strategy is universal testing of all patients in high-prevalence clinic populations (e.g., STD clinics, juvenile detention facilities, and family planning clinics).

The effectiveness of selective screening in reducing the prevalence of chlamydial infection among women has been demonstrated in several studies. In the Pacific Northwest, where extensive screening began in family planning clinics in 1998 and in STD clinics in 1993, the prevalence declined from 10% in the 1980s to <5% in 2000. Similar trends have occurred in association with screening programs elsewhere. In addition, screening can effect a reduction in upper genital tract disease. In Seattle, women at a large health maintenance organization who were screened for chlamydial infection on a routine basis had a lower incidence of symptomatic PID than did women who received standard care and underwent more selective screening.

In settings with low to moderate prevalence, the prevalence at which selective screening becomes more cost-effective than universal screening must be defined. Most studies have concluded that universal screening is preferable in settings with a chlamydial prevalence of >3–7%. Depending on the criteria used, selective screening is likely to be more cost-effective when prevalence falls below 3%. Nearly all regions of the United States have now initiated screening programs, particularly in family planning and STD clinics. Along with single-dose therapy, the availability of highly sensitive and specific diagnostic NAATs using urine specimens and self-obtained vaginal swabs makes it feasible to mount an effective nationwide *Chlamydia* control program, with screening of high-risk individuals in traditional health-care settings and in novel outreach and community-based settings. The U.S. Preventive Task Force has named *Chlamydia* screening as a Grade B recommendation, which means that private insurance and Medicare will cover the cost of screening under the Affordable Care Act.

TRACHOMA

Epidemiology

Trachoma—a sequela of ocular disease in developing countries—continues to be a leading cause of preventable infectious blindness worldwide. The WHO estimates that ~6 million people have been blinded by trachoma and that ~1.3 million people in developing countries still suffer from preventable blindness due to trachoma; certainly hundreds of millions live in trachoma-endemic areas. Foci of trachoma persist in Australia, the South Pacific, and Latin America. *C. trachomatis* serovars A, B, Ba, and C are isolated from patients with clinical trachoma in areas of endemicity in developing countries in Africa, the Middle East, Asia, and South America.

The trachoma-hyperendemic areas of the world are in northern and sub-Saharan Africa, the Middle East, drier regions of the Indian subcontinent, and Southeast Asia. In hyperendemic areas, the prevalence of trachoma is essentially 100% by the second or third year of life. Active disease is most common among young children, who are the reservoir for trachoma. By adulthood, active infection is infrequent but sequelae result in blindness. In such areas, trachoma constitutes the major cause of blindness.

Trachoma is transmitted through contact with discharges from the eyes of infected patients. Transmission is most common under poor hygienic conditions and most often takes place between family members or between families with shared facilities. Flies can also transfer the mucopurulent ocular discharges, carrying the organisms on their legs from one person to another. The International Trachoma Initiative founded by the WHO in 1998 aims to eliminate blinding trachoma globally by 2020.

Clinical Manifestations

Both endemic trachoma and adult inclusion conjunctivitis present initially as conjunctivitis characterized by small lymphoid follicles in the conjunctiva. In regions with hyperendemic classic blinding trachoma, the disease usually starts insidiously before the age of 2 years. Reinfection is common and probably contributes to the pathogenesis of trachoma. Studies using polymerase chain reaction (PCR) or other NAATs indicate that chlamydial DNA is often present in the ocular secretions of patients with trachoma, even in the absence of positive cultures. Thus, persistent infection may be more common than was previously thought.

The cornea becomes involved, with inflammatory leukocytic infiltrations and superficial vascularization (pannus formation). As the inflammation continues, conjunctival scarring eventually distorts the eyelids, causing them to turn inward so that the lashes constantly abrade the eyeball (trichiasis and entropion); eventually the corneal epithelium is abraded and may ulcerate, with subsequent corneal scarring and blindness. Destruction of the conjunctival goblet cells, lacrimal ducts, and lacrimal gland may produce a “dry-eye” syndrome, with resultant corneal opacity due to drying (xerosis) or secondary bacterial corneal ulcers.

Communities with blinding trachoma often experience seasonal epidemics of conjunctivitis due to *H. influenzae* that contribute to the intensity of the inflammatory process. In such areas, the active infectious process usually resolves spontaneously in affected persons at 10–15 years of age, but conjunctival scars continue to shrink, producing trichiasis and entropion with subsequent corneal scarring in adults. In areas with milder and less prevalent disease, the process may be much slower, with active disease continuing into adulthood; blindness is rare in these cases.

Eye infection with oculogenital *C. trachomatis* strains in sexually active young adults presents as an acute onset of unilateral follicular conjunctivitis and preauricular lymphadenopathy similar to that seen in acute conjunctivitis caused by adenovirus or HSV. If untreated, the disease may persist for 6 weeks to 2 years. It is frequently associated with corneal inflammation in the form of discrete opacities ("infiltrates"), punctate epithelial erosions, and minor degrees of superficial corneal vascularization. Very rarely, conjunctival scarring and eyelid distortion occur, particularly in patients treated for many months with topical glucocorticoids. Recurrent eye infections develop most often in patients whose sexual partners are not treated with antimicrobial agents.

Diagnosis

The clinical diagnosis of classic trachoma can be made if two of the following signs are present: (1) lymphoid follicles on the upper tarsal conjunctiva; (2) typical conjunctival scarring; (3) vascular pannus; or (4) limbal follicles or their sequelae, Herbert pits. The clinical diagnosis of endemic trachoma should be confirmed by laboratory tests in children with relatively marked degrees of inflammation. Intracytoplasmic chlamydial inclusions are found in 10–60% of Giemsa-stained conjunctival smears in such populations, but chlamydial NAATs are more sensitive and are often positive when smears or cultures are negative. Follicular conjunctivitis in European or American adults living in trachomatous regions is rarely due to trachoma.

TREATMENT

TREATMENT

Trachoma

Adult inclusion conjunctivitis responds well to treatment with the same regimens used in uncomplicated genital infections—namely, azithromycin (a 1-g single oral dose) or doxycycline (100 mg twice daily for 7 days). Simultaneous treatment of all sexual partners is necessary to prevent ocular reinfection and chlamydial genital disease. Topical antibiotic treatment is not required for patients who receive systemic antibiotics.

PSITTACOSIS

Psittacine birds and many other avian species act as natural reservoirs for *C. psittaci*–type organisms, common pathogens in domestic mammals and birds. The species *C. psittaci,* which now includes only avian strains, affects humans only as a zoonosis. (The other strains previously included in this species have been

placed into different species that reflect the animals they infect: *C. abortus*, *C. muridarum*, *C. suis, C. felis*, and *C. caviae*.) Although all birds are susceptible, pet birds (parrots, parakeets, macaws, and cockatiels) and poultry (turkeys and ducks) are most frequently involved in transmission of *C. psittaci* to humans. Exposure is greatest in poultry-processing workers and in owners of pet birds. Infectious forms of the organisms are shed from both symptomatic and apparently healthy birds and may remain viable for several months. *C. psittaci* can be transmitted to humans by direct contact with infected birds or by inhalation of aerosols from avian nasal discharges and from infectious avian fecal or feather dust. Transmission from person to person has never been demonstrated.

The diagnosis is usually established serologically. Psittacosis in humans may present as acute primary atypical pneumonia (which can be fatal in up to 10% of untreated cases); as severe chronic pneumonia; or as a mild illness or asymptomatic infection in persons exposed to infected birds.

EPIDEMIOLOGY

Fewer than 50 confirmed cases of psittacosis are reported in the United States each year, although many more cases probably occur than are reported. Control of psittacosis depends on control of avian sources of infection. A pandemic of psittacosis was once stopped by banning shipment or importation of psittacine birds. Birds can receive prophylaxis in the form of a tetracycline-containing feed. Imported birds are currently quarantined for 30 days of treatment.

CLINICAL MANIFESTATIONS

Typical symptoms include fever, chills, muscular aches and pains, severe headache, hepato- and/or splenomegaly, and gastrointestinal symptoms. Cardiac complications may involve endocarditis and myocarditis. Fatal cases were common in the preantibiotic era. As a result of quarantine of imported birds and improved veterinary-hygienic measures, outbreaks and sporadic cases of psittacosis are now rare. Severe pneumonia requiring management in an intensive care unit may develop. Endocarditis, hepatitis, and neurologic complications may occur, and fatal cases have been reported. The incubation period is usually 5–19 days but can last as long as 28 days.

DIAGNOSIS

Previously, the most widely used serologic test for diagnosing chlamydial infections was the genus-specific CF test, in which assay of paired serum specimens often shows fourfold or greater increases in antibody titer. The CF test remains useful, but the gold standard of serologic tests is now the MIF test, which is not widely available (see section on diagnosis of *C. trachomatis* genital infection, above). Any antibody titer above 1:16 is considered significant evidence of exposure to chlamydiae, and a fourfold titer rise in paired sera in combination with a clinically compatible syndrome can be used to diagnose psittacosis. Some commercially available serologic tests based on measurement of antibodies to LPS can be useful when the clinical diagnosis is consistent with bird exposure; however, since these tests are reactive for all chlamydiae (i.e., all chlamydiae contain LPS), caution must be used in their interpretation.

TREATMENT

TREATMENT

Psittacosis

The antibiotic of choice is tetracycline; the dosage for adults is 250 mg four times a day, continued for at least 3 weeks to avoid relapse. Severely ill patients may need cardiovascular and respiratory support. Erythromycin (500 mg four times a day by mouth) is an alternative therapy.

C. PNEUMONIAE INFECTIONS

C. pneumoniae is a common cause of human respiratory diseases, such as pneumonia and bronchitis. This organism reportedly accounts for as many as 10% of cases of community-acquired pneumonia, most of which are diagnosed by serology. Serologic studies have linked *C. pneumoniae* to atherosclerosis; isolation and PCR detection in cardiovascular tissues have also been reported. These findings suggest an expanded range of diseases and syndromes for *C. pneumoniae.* Large-scale case–cohort studies have demonstrated some association of *C. pneumoniae* with lung cancer, as evaluated by serology.

EPIDEMIOLOGY

Primary infection occurs mainly in school-aged children and reinfection in adults. Seroprevalence rates of 40–70% show that *C. pneumoniae* is widespread in both industrialized and developing countries. Seropositivity usually is first detected at school age, and rates generally increase by ~10% per decade. About 50% of individuals have detectable antibody at 30 years of age, and most have detectable antibody by the eighth decade of life. Although, as mentioned, serologic evidence suggests that *C. pneumoniae* may be associated with up to 10% of cases of community-acquired pneumonia, most of this evidence is based not on paired serum samples but rather on a single high IgG titer. Some doubt exists about the true prevalence and etiologic role of *C. pneumoniae* in atypical pneumonia, especially since reports of cross-reactivity have raised questions about the specificity of serology when only a single serum sample is used for diagnosis.

PATHOGENESIS

Little is known about the pathogenesis of *C. pneumoniae* infection. It begins in the upper respiratory tract and, in many persons, persists as a prolonged asymptomatic condition of the upper respiratory mucosal surfaces. However, evidence of replication within vascular endothelium and synovial membranes of joints shows that, in at least some individuals, the organism is transported to distant sites, perhaps within macrophages. A *C. pneumoniae* outer-membrane protein may induce host immune responses whose cross-reactivity with human proteins results in an autoimmune reaction.

The role of *C. pneumoniae* in the etiology of atherosclerosis has been discussed since 1988, when Finnish researchers presented serologic evidence of an association of this organism with coronary heart disease and acute myocardial infarction. Subsequently, the organism was identified in atherosclerotic lesions by culture,

PCR, immunohistochemistry, and transmission electron microscopy; however, discrepant study results (including those of animal studies) and failure of large-scale treatment studies have raised doubts as to the etiologic role of *C. pneumoniae* in atherosclerosis. Epidemiologic studies have demonstrated an association between serologic evidence of *C. pneumoniae* infection and atherosclerotic disease of the coronary and other arteries. In addition, *C. pneumoniae* has been identified in atherosclerotic plaques by electron microscopy, DNA hybridization, and immunocytochemistry. The organism has been recovered in culture from atheromatous plaques—a result indicating the presence of viable replicating bacteria in vessels. Evidence from animal models supports the hypothesis that *C. pneumoniae* infection of the upper respiratory tract is followed by recovery of the organism from atheromatous lesions in the aorta and that the infection accelerates the process of atherosclerosis, especially in hypercholesterolemic animals. Antimicrobial treatment of the infected animals reverses the increased risk of atherosclerosis. In humans, two small trials in patients with unstable angina or recent myocardial infarction suggested that antibiotics reduce the likelihood of subsequent untoward cardiac events. However, larger-scale trials have not documented an effect of various antichlamydial regimens on the risk of these events.

CLINICAL MANIFESTATIONS

C. pneumoniae was first reported as the etiologic agent of mild atypical pneumonia in military recruits and college students. The clinical spectrum of *C. pneumoniae* infection includes acute pharyngitis, sinusitis, bronchitis, and pneumonitis, primarily in young adults. The clinical manifestations of primary infection appear to be more severe and prolonged than those of reinfection. The pneumonitis of *C. pneumoniae* pneumonia resembles that of *Mycoplasma* pneumonia in that leukocytosis is frequently lacking and patients often have prominent antecedent upper respiratory tract symptoms, fever, nonproductive cough, mild to moderate illness, minimal findings on chest auscultation, and small segmental infiltrates on chest x-ray. In elderly patients, pneumonia due to *C. pneumoniae* can be especially severe and may necessitate hospitalization and respiratory support.

Chronic infection with *C. pneumoniae* has been reported among patients with chronic obstructive pulmonary disease and may also play a role in the natural history of asthma, including exacerbations. The clinical symptoms of respiratory infections caused by *C. pneumoniae* are nonspecific and do not differ from those caused by other agents of atypical pneumonia, such as *Mycoplasma pneumoniae*.

DIAGNOSIS

Serology, PCR amplification, and culture can be used to diagnose *C. pneumoniae* infection. Serology has been the traditional diagnostic method. The gold standard serologic test is the MIF test (see section on diagnosis of *C. trachomatis* genital infection, above). Any antibody titer >1:16 is considered significant evidence of exposure to chlamydiae. According to a CDC-sponsored expert working group, the diagnosis of acute *C. pneumoniae* infection requires demonstration of a fourfold rise in titer in paired serum samples. There are no official recommendations for diagnosis of chronic infections, although many research studies have used high titers of IgA as an indicator. The older CF tests and EIAs for LPS are not recommended, as they

are not specific for *C. pneumoniae* but identify the chlamydiae only to the genus level. The organism is very difficult to grow in tissue culture but has been cultivated in HeLa cells, HEp-2 cells, and HL cells. Although NAATs are commercially available for *C. trachomatis*, only research-based PCR assays are available for *C. pneumoniae.*

TREATMENT

TREATMENT

C. pneumoniae Infections

Although few controlled trials of treatment have been reported, *C. pneumoniae* is inhibited in vitro by erythromycin, tetracycline, azithromycin, clarithromycin, gatifloxacin, and gemifloxacin. Recommended therapy consists of 2 g/d of either tetracycline or erythromycin for 10–14 days. Other macrolides (e.g., azithromycin) and some fluoroquinolones (e.g., levofloxacin and gatifloxacin) also appear to be effective.

ACKNOWLEDGMENT

The authors acknowledge the late Walter E. Stamm, MD, for his significant contributions to the field of Chlamydia research. Dr. Stamm wrote the chapters on chlamydiae for previous editions of Harrison's Principles of Internal Medicine, and we thank the editors for permission to reproduce Figs. 169-1 and 169-2 and Table 169-1 from his chapter in the 17th edition. Dr. Stamm died on December 14, 2009, and this chapter is dedicated to him.

FURTHER READING

Association of Public Health Laboratories: *Guidelines for the Laboratory Testing of STDs, 2009*. Silver Spring, MD, Association of Public Health Laboratories, 2009.

Bavoil PM, Wyrick PB: *Chlamydia: Genomics and Pathogenesis.* Norfolk, UK, Horizon Bioscience, 2006, 542 pp.

Campbell LA et al: Chlamydial infections, in *Manual of Molecular and Clinical Laboratory Immunology,* 7th ed. B Detrick et al (eds). Washington, DC, ASM Press, 2006, pp 518–525.

Centers for Disease Control and Prevention: *Sexually Transmitted Disease Surveillance*, 2016. Atlanta, U.S. Department of Health and Human Services, 2017 (*www.cdc.gov/std/stats17/default.htm*).

Centers for Disease Control and Prevention: Sexually transmitted diseases treatment guidelines, 2015. MMWR Recomm Rep 59(RR-12):1, 2016.

Gaydos CA: Chlamydiae, in *Clinical Virology Manual*, 4th ed. S Spector et al (eds). Washington, DC, ASM Press, 2009, pp 630–640.

Gaydos CA, Essiq A: Chlamydiaceae, in *Manual of Clinical Microbiology*, 11th ed. JH Jorgensen et al (eds). Washington, DC, ASM Press, 2015, pp 1106–1121.

Kuypers J et al: Principles of laboratory diagnosis of STIs, in *Sexually Transmitted Diseases,* 4th ed. KK Holmes et al (eds). New York, McGraw-Hill, 2008, pp 937–948.

Schachter J, Stephens RS: Biology of *Chlamydia trachomatis, in Sexually Transmitted Diseases,* 4th ed. KK Holmes et al (eds). New York, McGraw-Hill, 2008, pp 555–574.

Scholes D et al: Prevention of pelvic inflammatory disease by screening for cervical chlamydial infection. N Engl J Med 334:1362, 1996.

Taylor HR: *Trachoma: A Blinding Scourge from the Bronze Age to the Twenty-First Century*. East Melbourne, Victoria, Australia, Centre for Eye Research Australia/Haddington Press, 2008, 282 pp.

Chapter 185: Medical Virology

Fred Wang; Elliott Kieff

DEFINING A VIRUS

Viruses are obligate intracellular parasites. They consist of a DNA or RNA genome surrounded by protein. They may also have an outer-membrane lipoprotein envelope. Viruses can replicate only within cells because their nucleic acid does not encode many enzymes necessary for the metabolism of proteins, carbohydrates, or lipids or for the generation of high-energy phosphates. Typically, viral nucleic acids encode messenger RNA (mRNA) and proteins necessary for replicating, packaging, and releasing progeny virus from infected cells.

Viruses differ from virusoids, viroids, and prions. *Virusoids* are nucleic acids that depend on cells and helper viruses for packaging their nucleic acids into virus-like particles. *Viroids* are naked, cyclical, mostly double-strand small RNAs that appear to be restricted to plants, spread from cell to cell, and are replicated by cellular RNA polymerase II. Prions (Chap. 430) are abnormal proteins that propagate and cause disease by altering the structure of a normal cell protein. Prions cause neurodegenerative diseases such as Creutzfeldt-Jakob disease, Gerstmann-Straüssler disease, kuru, and human or bovine spongiform encephalopathy (“mad cow disease”).

VIRUS STRUCTURE

Viral genomes may consist of single- or double-strand DNA, single- or double-strand RNA, single-strand or segmented antisense RNA, or double-strand segmented RNA. Viral nucleic acids may encode only a few genes or more than 100. Sense-strand viral RNA genomes can be translated directly into protein, whereas antisense RNAs must be copied into translatable RNA. Sense and antisense genomes are also referred to as *positive-strand* and *negative-strand genomes*, respectively. Viral nucleic acid is usually associated with virus-encoded nucleoprotein(s) in the virus core. Viral nucleic acids and nucleoproteins are almost always enclosed in a protein *capsid*. Because of the limited genetic complexity of viruses, their capsids are usually composed of multimers of identical *capsomeres* made up of one or a few proteins. Capsids have icosahedral or helical symmetry. Icosahedral capsid structures approximate spheres and have two-, three-, or fivefold axes of symmetry, whereas helical capsid structures have only a twofold axis of symmetry. The nucleic acid, nucleoprotein(s), and protein capsid together are called a *nucleocapsid*.

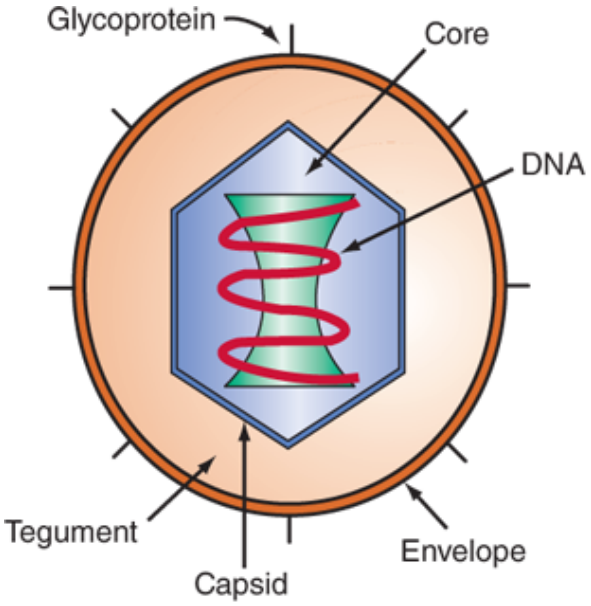
Many viruses are composed of a nucleic acid core and a capsid. For these viruses, the outer capsid surface mediates contact with uninfected cells’ plasma membranes. Other viruses are more complex and have an outer phospholipid, cholesterol, glycoprotein, and glycolipid envelope that is derived from virus-modified infected cell membranes. Cell nuclear, endoplasmic reticulum, Golgi, or plasma membranes that become parts of the viral envelope have usually been modified during infection by the insertion of virus-encoded glycoproteins, which

mediate contact of enveloped virus with uninfected cell surfaces. Matrix or tegument proteins may fill the space between the nucleocapsid and the outer envelope of the virus.

Enveloped viruses are usually sensitive to lipid solvents or detergents that can dissolve the envelope, whereas viruses with protein nucleocapsid exteriors may be somewhat detergent resistant. A schematic diagram of large and complex herpesviruses is shown in **Fig. 185-1**. Structures of prototypical pathogenic human viruses are described in **Table 185-1**. The relative sizes and structures of typical pathogenic human viruses are shown in **Fig. 185-2**.

FIGURE 185-1

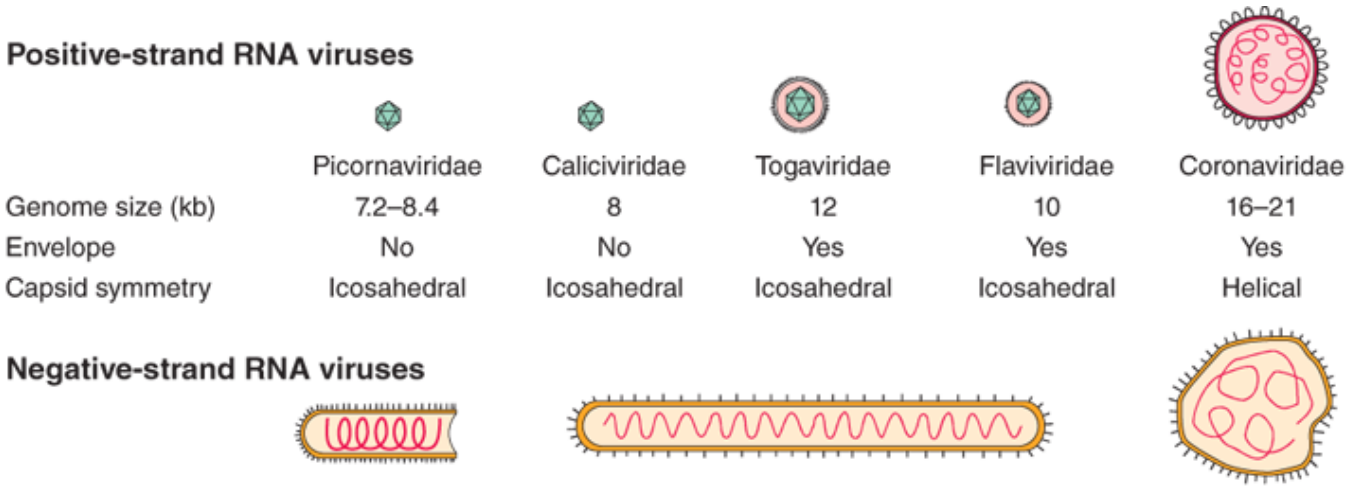
Schematic diagram of an enveloped herpesvirus with an icosahedral nucleocapsid. The approximate respective dimensions of the nucleocapsid and the enveloped particles are 110 and 180 nm. The capsid is composed of 162 capsomeres: 150 with sixfold and 12 with fivefold axes of symmetry.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 185-2

Schematic diagrams of the major virus families including species that infect humans. The viruses are grouped by genome type and are drawn approximately to scale. Prototype viruses of each family that cause human disease are listed in Table 185-1.



	Rhabdoviridae	Filoviridae	Paramyxoviridae
Genome size (kb)	13–16	13	16–20
Envelope	Yes	Yes	Yes
Capsid symmetry	Helical	Helical	Helical

Segmented negative-strand RNA viruses

Segmented double-strand RNA viruses






	Orthomyxoviridae	Bunyaviridae	Arenaviridae	Reoviridae
Genome size (kb)	14	13–21	10–14	16–27
Envelope	Yes	Yes	Yes	No
Capsid symmetry	Helical	Helical	Helical	Icosahedral

Retroviruses



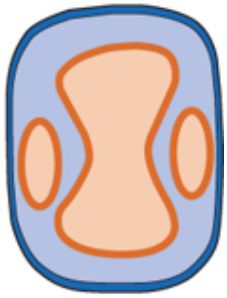
	Retroviridae
Genome size (kb)	3–9
Envelope	Yes
Capsid symmetry	Icosahedral

DNA viruses

100 nm





	Parvoviridae	Papillomaviridae Polyomaviridae	Hepadnaviridae	Adenoviridae	Herpesviridae	Poxviridae
Genome size (kb)	5	5–9	3*	36–38	100–250	240
Envelope	No	No	Yes	No	Yes	Yes
Capsid symmetry	Icosahedral	Icosahedral	Icosahedral	Icosahedral	Icosahedral	Complex

*partial RNA genome

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 185-1

Virus Families Pathogenic for Humans

Family	Representative Viruses	Type of RNA/DNA	Lipid Envelope
RNA Viruses			
Picornaviridae	Poliovirus	(+) RNA	No
	Coxsackievirus		
	Echovirus		
	Enterovirus		
	Rhinovirus		
	Hepatitis A virus		
Caliciviridae	Norovirus	(+) RNA	No
	Hepatitis E virus		
Togaviridae	Rubella virus	(+) RNA	Yes
	Eastern equine encephalitis virus		
	Western equine encephalitis virus		
Flaviviridae	Yellow fever virus	(+) RNA	Yes
	Dengue virus		
	St. Louis encephalitis virus		
	West Nile virus		
	Zika virus		
	Hepatitis C virus		
	Hepatitis G virus		

Family	Representative Viruses	Type of RNA/DNA	Lipid Envelope
Coronaviridae	Coronaviruses[a]	(+) RNA	Yes
Rhabdoviridae	Rabies virus	(–) RNA	Yes
	Vesicular stomatitis virus		
Filoviridae	Marburg virus	(–) RNA	Yes
	Ebola virus		
Paramyxoviridae	Parainfluenza virus	(–) RNA	Yes
	Respiratory syncytial virus		
	Newcastle disease virus		
	Mumps virus		
	Rubeola (measles) virus		
Orthomyxoviridae	Influenza A, B, and C viruses	(–) RNA, 8 segments	Yes
Bunyaviridae	Hantavirus	(–) RNA, 3 circular segments	Yes
	California encephalitis virus		
	Sandfly fever virus		
Arenaviridae	Lymphocytic choriomeningitis virus	(–) RNA, 2 circular segments	Yes
	Lassa fever virus		
	South American hemorrhagic fever virus		
Reoviridae	Rotavirus	ds RNA, 10–12 segments	No
	Reovirus		
	Colorado tick fever virus		
Retroviridae	Human T lymphotropic virus types 1 and 2	(+) RNA, 2 identical segments	Yes

Family	Representative Viruses	Type of RNA/DNA	Lipid Envelope
	Human immunodeficiency virus types 1 and 2		
DNA Viruses			
Hepadnaviridae	Hepatitis B virus	ds DNA with ss portions	Yes
Parvoviridae	Parvovirus B19	ss DNA	No
Papillomaviridae	Human papillomaviruses	ds DNA	No
Polyomaviridae	JC virus		
	BK virus		
	Merkel cell polyoma virus		
Adenoviridae	Human adenoviruses	ds DNA	No
Herpesviridae	Herpes simplex virus types 1 and 2[b]	ds DNA	Yes
	Varicella-zoster virus[c]		
	Epstein-Barr virus[d]		
	Cytomegalovirus[e]		
	Human herpesvirus 6		
	Human herpesvirus 7		
	Kaposi's sarcoma–associated herpesvirus[f]		
Poxviridae	Variola (smallpox) virus	ds DNA	Yes
	Orf virus		
	Molluscum contagiosum virus		

[a]Including the coronaviruses causing severe acute respiratory syndrome (SARS) and Middle Eastern respiratory syndrome (MERS). [b]Also called human herpesvirus 1 (HHV-1) and HHV-2, respectively. [c]Also called HHV-3. [d]Also called HHV-4. [e]Also called HHV-5. [f]Also called HHV-8.

Abbreviations: ds, double-strand; ss, single-strand.

TAXONOMY OF PATHOGENIC HUMAN VIRUSES

As is apparent from Table 185-1 and Fig. 185-2, the classification of viruses into orders and families is based on nucleic acid composition, nucleocapsid size and symmetry, and presence or absence of an envelope. Viruses of a single family have similar structures and may be morphologically indistinguishable in electron micrographs. Subclassification into genera depends on similarity in epidemiology, biologic effects, and nucleic acid sequence.

Most viruses that infect humans have a common name related to their pathologic effects or the circumstances of their discovery. In addition, formal species names—consisting of the name of the host followed by the family or genus of the virus and a number—have been assigned by the International Committee on Taxonomy of Viruses. This dual terminology can cause confusion when viruses are referred to by either name—e.g., varicella-zoster virus (VZV) or human herpesvirus 3 (HHV-3).

VIRAL INFECTION IN VITRO

STAGES OF VIRAL INFECTION OF CELLS IN CULTURE

Viral Interactions with Cell Surfaces and Cell Entry

To deliver its nucleic acid payload to the cell cytoplasm or nucleoplasm, a virus must overcome barriers posed by the cell's plasma and cytoplasmic membranes. Infection is frequently initiated by weak electrostatic or hydrophobic interactions with the cell surface. Subsequent stronger, more specific attachment to a cell plasma membrane protein, carbohydrate, glycolipid, heparan sulfate proteoglycan, or sialic acid enables stable binding to a specific cell surface "receptor" that mediates fusion with the cell plasma membrane **(see Table 116-1)**. Receptor binding is often augmented by a viral surface protein interaction with more than one cell surface protein or co-receptor. Receptors and co-receptors are important determinants of the species and cell type that a virus can infect. For example, the HIV envelope glycoprotein binds to the T cell surface protein CD4 and then engages a chemokine receptor that is the definitive co-receptor for the virus and mediates entry into the cell cytoplasm. The Epstein-Barr virus (EBV) glycoprotein gp350 binds to the B lymphocyte complement receptor CD21 and then uses a major histocompatibility complex (MHC) class II molecule as a co-receptor and an integrin for definitive entry.

Viruses have evolved a wide range of strategies to enter cells. Influenza virus has an outer-membrane hemagglutinin glycoprotein that binds to sialic acid on respiratory tract cell plasma membranes. The hemagglutinin mediates adsorption to cell membranes, receptor aggregation, and endocytosis. As the endosome pH decreases in the cell cytoplasm, the influenza hemagglutinin conformation changes, enabling hydrophobic helices, which are initially at the base of the hemagglutinin, to extend, interacting and fusing with the endosome

membrane and thereby releasing the viral genome into the cell cytoplasm. The influenza virus M2 membrane channel protein has a key role in lowering endosome pH and permitting virus and cell membrane fusion.

Nonenveloped viruses (e.g., human papillomaviruses [HPVs]) and some enveloped viruses have evolved to partially fuse with cell plasma membrane receptors and be internalized into endosomes. The low pH in an endosome can then trigger virus membrane or capsid fusion with the endocytic membrane, releasing viral DNA into the cytoplasm to initiate infection.

Hydrophobic interactions required for fusion can be susceptible to chemical inhibition or blockade. The HIV envelope glycoprotein gp120 is associated with gp41 on the viral surface. HIV gp120 binding to CD4 and then to specific chemokine receptors results in conformational changes that allow gp41 to initiate cell membrane fusion. The anti-HIV drug enfuvirtide is a small peptide derived from the gp41 structure. Enfuvirtide binds to gp41 and prevents conformational changes required for fusion. In contrast, maraviroc prevents virus entry by binding to the CCR5 receptor, thereby blocking gp120 binding to CCR5 and preventing gp120 fusion with CCR5.

Viral Gene Expression and Replication

After uncoating and release of viral nucleoprotein into the cytoplasm, the viral genome is transported to sites of expression and replication. To produce infectious progeny, viruses must produce proteins necessary for replicating their nucleic acids as well as structural proteins necessary for coating their nucleic acids and for assembling nucleic acids and proteins into progeny virus. Different viruses use different strategies and gene repertoires to accomplish these goals. Most DNA viruses, except for poxviruses, replicate their nucleic acid and assemble into nucleocapsids in the cell nucleus. RNA viruses, except for influenza viruses, transcribe and replicate their RNA and assemble in the cytoplasm before envelopment at the cell plasma membrane. The replication strategies of DNA and RNA viruses and of positive- and negative-strand RNA viruses are presented and discussed separately below. Medically important viruses of each group are used for illustrative purposes.

POSITIVE-STRAND RNA VIRUSES

RNA viruses of medical importance include positive-strand picornaviruses, flaviviruses, togaviruses, caliciviruses, and coronaviruses. Genome RNA from positive-strand RNA viruses is released into the cytoplasm without associated enzymes. Cell ribosomes recognize and associate with the viral genome's internal ribosome entry sequence and translate a virus-encoded polyprotein. Proteases within the polyprotein cleave out the viral RNA polymerase and other viral proteins necessary for replication. Antigenomic RNA is next transcribed from the genome RNA template. Positive-strand genomes and mRNAs are then transcribed from the antigenome RNA by the viral RNA polymerase and are translated into capsid proteins. Genomic RNA is encapsidated in the cytoplasm and released as the infected cell undergoes lysis.

NEGATIVE-STRAND RNA VIRUSES

Medically important negative-strand RNA viruses include rhabdoviruses, filoviruses, paramyxoviruses, orthomyxoviruses, and bunyaviruses. The genomes of negative-strand viruses are frequently segmented. Negative-strand RNA viral genomes are released into the cytoplasm with an associated RNA polymerase and one or more polymerase accessory proteins. The viral RNA polymerase transcribes mRNAs as well as full-length antigenome RNA, which is the template for genome RNA replication. Viral mRNAs encode the viral RNA polymerase and accessory factors as well as viral structural proteins. Except for influenza virus, which transcribes its mRNAs and

antigenome RNAs in the cell nucleus, negative-strand RNA viruses replicate entirely in the cytoplasm. All negative-strand RNA viruses, including influenza viruses, assemble in the cytoplasm.

DOUBLE-STRAND SEGMENTED RNA VIRUSES

Double-strand RNA viruses are taxonomically grouped in the family Reoviridae. The medically important viruses in this group are rotaviruses and Colorado tick fever virus. Reovirus genomes have 10–12 RNA segments. Reovirus particles contain an RNA polymerase complex. These viruses replicate and assemble in the cell cytoplasm.

DNA VIRUSES

Medically important DNA viruses include parvoviruses, which have small single-strand DNA genomes and cause transient arthritis, and polyomaviruses, including the smaller polyomaviruses such as JC virus, which causes progressive multifocal leukoencephalopathy in immunocompromised patients; BK virus; and Merkel cell polyomavirus. The larger HPVs cause warts as well as cervical, penile, and oral carcinomas. The next larger DNA viruses are adenoviruses, which mostly cause transient respiratory tract and ocular inflammatory disease. The herpesviruses include eight viruses that cause a wide range of inflammatory and malignant diseases in humans. EBV is an important cause of lymphomas and Hodgkin's disease in both immunocompromised and immunocompetent people and of nasopharyngeal carcinoma in southern Chinese and northern African populations. Cytomegalovirus (CMV) is an important cause of transplacental infections and neonatal neurologic impairment. Poxviruses, the largest DNA viruses and the largest viruses that infect humans (barely visible by light microscopy), cause smallpox, monkeypox, and molluscum contagiosum. Aside from those of poxviruses, other DNA virus genomes enter the cell nucleus and are transcribed by cellular RNA polymerase II.

After receptor binding and fusion with plasma membranes or endocytic vesicle membranes, herpesvirus nucleocapsids are released into the cytoplasm with tegument proteins and are transported along microtubules to a nuclear pore. Capsids then release DNA into the nucleus.

DNA virus transcription and mRNA processing depend on both viral and cellular proteins. For herpes simplex virus (HSV), a viral tegument protein enters the nucleus and activates immediate-early genes, the first genes expressed after infection. Transcription of immediate-early genes requires the viral tegument protein and cell transcription factors. HSV becomes nonreplicating, or latent, in neurons because essential cell transcription factors for expression of viral immediate-early genes are docked in the cytoplasm in neurons. Heat shock or other cell stresses can cause these cell factors to enter the nucleus, activate viral gene expression, and initiate replication. This information explains HSV-1 latency in neurons and activation of replicative infection.

For adenoviruses and herpesviruses, transcription of immediate-early genes results in expression of early proteins necessary for viral DNA replication. Viral DNA synthesis is required to turn on late-gene expression and production of viral structural components. The HPVs, polyomaviruses, and parvoviruses are not dependent on transactivators encoded from the viral genome for early-gene transcription. Instead, their early genes have upstream enhancing elements that bind cell transcription factors. The early genes encode proteins that are necessary for viral DNA synthesis and late-gene transcription. DNA virus late genes encode structural proteins necessary for viral assembly and for viral egress from the infected cell. Late-gene transcription is continuously dependent on DNA replication. Therefore, inhibitors of DNA replication also stop late-gene transcription.

Each DNA virus family uses unique mechanisms for replicating its DNA. Adenovirus and herpesvirus DNAs are linear in the virion. Adenovirus DNA remains linear in infected cells and replicates as a linear genome, using an initiator

protein–DNA complex. In contrast, herpesvirus DNA circularizes in the infected cell, and genomes replicate into linear concatemers through a “rolling-circle” mechanism. Full-length DNA genomes are cleaved and packaged into virus. Herpesviruses encode a DNA polymerase and at least six other viral proteins necessary for viral DNA replication. Acyclovir and ganciclovir prevent viral DNA synthesis when they are phosphorylated and incorporated into DNA by the viral polymerase. Herpesviruses also encode enzymes that increase the deoxynucleotide triphosphate pools. HPV and polyomavirus DNAs are circular both within the virus and in infected cells. These genomes are reproduced by cellular DNA replication enzymes and remain circular through replication and packaging. HPV and polyomavirus early proteins are necessary for DNA replication in both latent and viral replicative phases. Early viral proteins stimulate cells to remain in cycle, facilitating viral DNA replication.

Parvoviruses have negative single-strand DNA genomes and are the smallest DNA viruses. Their genomes are half the size of HPV genomes and include only two genes. The replication of autonomous parvoviruses, such as B19, depends on cellular DNA replication and requires the virus-encoded Rep protein. Other parvoviruses, such as adeno-associated virus (AAV), are not autonomous and require helper viruses of the adenovirus or herpesvirus family for their replication. AAV is being used as a potentially safe human gene therapy vector because its replication protein causes integration at a single chromosome site. The small genome size limits the range of proteins that can be expressed from AAV vectors.

As stated above, poxviruses are the largest DNA viruses. They are unique among DNA viruses in replicating and assembling in the cytoplasm. To accomplish cytoplasmic replication, poxviruses encode transcription factors, an RNA polymerase II orthologue, enzymes for RNA capping, enzymes for RNA polyadenylation, and enzymes for viral DNA synthesis. Poxvirus DNA also has a unique structure. The double-strand linear DNA is covalently linked at the ends, making a covalently closed double-strand circular genome. Replication of the circular genomes is initiated by nicking in inverted repeats at the ends of the linear DNA. During DNA replication, the genome is cleaved within the terminal inverted repeats, and the inverted repeats self-prime complementary-strand synthesis by the virus-encoded DNA polymerase. Like herpesviruses, poxviruses encode several enzymes that increase deoxynucleotide triphosphate precursor levels and thus facilitate viral DNA synthesis.

VIRUSES THAT USE BOTH RNA AND DNA GENOMES IN THEIR LIFE CYCLE

Retroviruses, including HIV, are RNA viruses that use a DNA intermediate to replicate their genomes. In contrast, hepatitis B virus (HBV) is a DNA virus that uses an RNA intermediate to replicate its genome. Thus these viruses are not purely RNA or DNA viruses. Retroviruses are RNA viruses with two identical sense-strand genomes and associated reverse transcriptase and integrase enzymes. Retroviruses differ from all other viruses in that they reverse-transcribe themselves into partially duplicated double-strand DNA copies and then routinely integrate into the host genome as part of their persistence and replication strategies. Inhibitors of reverse transcriptase (e.g., zidovudine) or integrase (e.g., raltegravir) are now commonly used as antiviral treatments for HIV infection. Integration of remnants and even complete copies of simple retrovirus DNAs into the human genome raises the possibility of replication-competent simple human retroviruses. However, endogenous human retrovirus replication has not been documented or associated with any disease. Integrated, replication-competent retroviral DNAs are also present in many animal species, such as pigs. These porcine retroviruses are a potential cause for concern in xenotransplantation because retrovirus replication could cause disease in humans.

Cellular RNA polymerase II and transcription factors regulate transcription from the integrated provirus DNA genome. Some retroviruses also encode regulators of transcription and RNA processing, such as Tax and Rex in

human T lymphotropic virus (HTLV) types 1 and 2. HIV-1 and HIV-2 have orthologous Tat and Rev genes as well as the additional accessory proteins Vpr, Vpu, and Vif, which are important for efficient infection and immune escape. Full-length proviral transcripts are made from a promoter in the viral terminal repeat and serve as both genome RNAs that are packaged in the nucleocapsids and differentially spliced mRNAs that encode for the virus Gag protein, polymerase/integrase protein, and envelope glycoprotein. The Gag protein includes a protease that cleaves it into several components, including a viral matrix protein that coats the viral RNA. Viral RNA polymerase/integrase, matrix protein, and cellular tRNAs are key components in the viral nucleocapsid. Protease inhibitors have been developed as effective agents against infections caused by HIV (e.g., saquinavir) or hepatitis C virus (HCV) (e.g., telaprevir).

HBV replication is unique in several respects. The HBV genome is a partially double-strand DNA genome that is repaired in infected cells to a fully double-strand circular DNA by the virion polymerase. Viral mRNAs are transcribed from the closed circular viral episome by the cellular RNA polymerase II and are translated to yield HBV proteins, including core protein, surface antigen, and polymerase. In addition, a full-genome-length mRNA is packaged into viral core particles in the cytoplasm of infected cells as an intermediate for viral DNA replication. This RNA associates with the viral polymerase, which also has reverse transcriptase activity and converts the full-length encapsidated RNA genome into partially double-strand DNA. Thus, nucleos(t)ide analogs that inhibit reverse transcription (e.g., tenofovir) are commonly used to treat HBV infection. HBV is believed to mature by budding through the cell's plasma membrane, which has been modified by the insertion of viral surface antigen protein.

Viral Assembly and Egress

For most viruses, nucleic acid and structural protein synthesis is accompanied by the assembly of protein and nucleic acid complexes. The assembly and egress of mature infectious virus mark the end of the *eclipse phase* of infection, during which infectious virus cannot be recovered from the infected cell. Nucleic acids from RNA viruses and poxviruses assemble into nucleocapsids in the cytoplasm. For all DNA viruses except poxviruses, viral DNA assembles into nucleocapsids in the nucleus. In general, the capsid proteins of viruses with icosahedral nucleocapsids can self-assemble into densely packed and highly ordered capsid structures. Herpesviruses require an assemblin protein as a scaffold for capsid assembly. Viral nucleic acid then spools into the assembled capsid. For herpesviruses, a full unit of the viral DNA genome is packaged into the capsid, and a capsid-associated nuclease cleaves the viral DNA at both ends. In the case of viruses with helical nucleocapsids, the protein component appears to assemble around the nucleic acid, which contributes to capsid organization.

Viruses must egress from the infected cell and not bind back to their receptor(s) on the outer surface of the plasma membrane. Viruses can acquire envelopes from cytoplasmic membranes or by budding through the cell's plasma membrane. Excess viral membrane glycoproteins are synthesized to saturate cell receptors and facilitate separation of the virus from the infected cell. Some viruses encode membrane proteins with enzymatic activity for receptor destruction. Influenza virus, for example, encodes a glycoprotein with neuraminidase activity. Neuraminidase destroys sialic acid on the infected cell's plasma membrane so that newly released virus does not get stuck to the dying cell. Oseltamivir and zanamivir are neuraminidase inhibitors that are used to treat or provide prophylaxis for influenza virus infection. Herpesvirus nucleocapsids acquire an initial envelope by assembling in the nucleus and then budding through the nuclear membrane into the endoplasmic reticular space. The initially enveloped herpesvirus is then de-enveloped and released from the cell either by exocytosis or by re-envelopment at the

plasma membrane. Nonenveloped viruses depend on the death and dissolution of the infected cell for their release.

FIDELITY OF VIRAL REPLICATION

Hundreds or thousands of progeny may be produced from a single virus-infected cell. Many particles partially assemble and never mature into virions. Many mature-appearing virions are imperfect and have only incomplete or nonfunctional genomes. Despite the inefficiency of assembly, a typical virus-infected cell releases 10–1000 infectious progeny. Some of these progeny may contain genomes that differ from those of the virus that infected the cell. Smaller, “defective” viral genomes have been noted with the replication of many RNA and DNA viruses. Virions with defective genomes can be produced in large numbers through packaging of incompletely synthesized nucleic acid. Adenovirus packaging is notoriously inefficient, and a high ratio of particle to infectious virus may limit the amount of recombinant adenovirus that can be administered for gene therapy since the immunogenicity of defective particles may contribute to adverse effects.

Changes in viral genomes can lead to mutant viruses of medical significance. In general, viral nucleic acid replication is more error-prone than cellular nucleic acid replication. RNA polymerases and reverse transcriptases are significantly more error-prone than DNA polymerases. Mutations can also be introduced into the HIV genome by APOBEC3G, a cellular protein that is packaged in the virion. APOBEC3G deaminates cytidine in the virion RNA to uridine. When reverse transcriptase subsequently uses the altered virion RNA as a template in the infected cell, a guanosine-to-adenosine mutation is introduced into the proviral DNA. Mutations resulting in less efficient viral growth, or fitness, may be detrimental to the virus. HIV-encoded Vif blocks APOBEC3G activity in the virion, inhibiting the debilitating effects of hypermutation on genetic integrity. Nevertheless, mutations resulting in evasion of the host immune response or resistance to antiviral drugs are preferentially selected in patients, with the consequent perpetuation of infection. Viral genomes can also be altered by recombination or reassortment between two related viruses in a single infected cell. Although this occurrence is unusual under most circumstances of natural infection, the genome changes can be substantial and can significantly alter virulence or epidemiology. Reassortment of the avian or mammalian influenza A hemagglutinin gene into a human influenza background can result in the emergence of new epidemic or pandemic influenza A strains.

VIRAL GENES NOT REQUIRED FOR VIRAL REPLICATION

Viruses frequently have genes encoding proteins that are not directly involved in replication or packaging of the viral nucleic acid, in virion assembly, or in regulation of the transcription of viral genes involved in those processes. Most of these proteins fall into five classes: (1) proteins that directly or indirectly alter cell growth; (2) proteins that inhibit cellular RNA or protein synthesis so that viral mRNA can be efficiently transcribed or translated; (3) proteins that promote cell survival or inhibit apoptosis so that progeny virus can mature and escape from the infected cell; (4) proteins that inhibit the host interferon response; and (5) proteins that downregulate host inflammatory or immune responses so that viral infection can proceed in an infected person to the extent consistent with the survival of the virus and its efficient transmission to a new host. More complex viruses of the poxvirus or herpesvirus family encode many proteins that serve these functions. Some of these viral proteins have motifs similar to those of cellular proteins, while others are quite novel. Virology has increasingly focused on these more sophisticated strategies evolved by viruses to permit the establishment of long-term infection in humans and other animals. These strategies often provide unique insights into the control of cell growth, cell survival,

macromolecular synthesis, proteolytic processing, immune or inflammatory suppression, immune resistance, cytokine mimicry, or cytokine blockade.

MicroRNAs (miRNAs) are small noncoding RNAs that can regulate gene expression at the posttranscriptional level by targeting—and usually silencing—mRNAs. miRNAs were initially discovered in plants and plant viruses, where they alter expression of cell defensins. Herpesviruses are especially rich in miRNAs; for example, at least 23 miRNAs have been identified in EBV and 11 in CMV. Adenovirus and polyomavirus miRNAs have also been described. Increasing data indicate that animal viruses encode miRNAs to alter the growth and survival of host cells and the innate and acquired immune responses.

HOST RANGE

The concept of host range was originally based on the cell types in which a virus replicates in tissue culture. For the most part, the host range is limited by specific cell-surface proteins required for viral adsorption or penetration—i.e., to the cell types that express receptors or co-receptors for a specific virus. Another common basis for host-range limitation is the degree of transcriptional activity from viral promoters in different cell types. Most DNA viruses depend not only on cellular RNA polymerase II and the basal components of the cellular transcription complex but also on activated components and transcriptional accessory factors, both of which differ among differentiated tissues, among cells at various phases of the cell cycle, and between resting and cycling cells.

The importance of host-range factors is illustrated by the effects of specific host determinants that limit the replication of influenza virus with avian or porcine hemagglutinins in humans. These viral proteins have adapted to bind avian or porcine sialic acids, and spread of avian or porcine influenza viruses in human populations is limited by their ability to infect human cells.

VIRAL CYTOPATHIC EFFECTS AND INHIBITORS OF APOPTOSIS

The replication of almost all viruses has adverse effects on the infected cell, inhibiting cellular synthesis of DNA, RNA, or proteins through efficient competition for key substrates and enzymatic processes. These general inhibitory effects enable viruses to nonspecifically limit components of innate host resistance, such as interferon (IFN) production. Viruses can specifically inhibit host protein synthesis by attacking a component of the translational initiation complex—frequently, a component that is not required for efficient translation of viral RNAs. Poliovirus protease 2A, for example, cleaves a cellular component of the complex that ordinarily facilitates translation of cellular mRNAs by interacting with their cap structure. Poliovirus RNA is efficiently translated without a cap because it has an internal ribosome entry sequence. Influenza virus inhibits the processing of mRNA by snatching cap structures from nascent cellular RNAs and using them as primers in the synthesis of viral mRNA. HSV has a virion tegument protein that inhibits cellular mRNA translation.

Apoptosis is the expected consequence of virus-induced inhibition of cellular macromolecular synthesis and viral nucleic acid replication. Although the induction of apoptosis may be important for the release of some viruses (particularly nonenveloped viruses), many viruses have acquired genes or parts of genes that enable them to forestall infected-cell death. This delay increases the yield from viral replication. Adenoviruses and herpesviruses encode analogues of the cellular Bc12 protein, which block mitochondrial enhancement of proapoptotic stimuli.

Poxviruses and some herpesviruses also encode caspase inhibitors. Many viruses, including HPVs and adenoviruses, encode proteins that inhibit p53 or its downstream proapoptotic effects.

VIRAL INFECTION IN VIVO

TRANSMISSION

The capsid and envelope of a virus protect the genome and enable efficient transmission of the virus from cell to cell and to new prospective hosts. Most common viral infections are spread by direct contact, by ingestion of contaminated water or food, or by inhalation of aerosolized particles. In all these situations, infection begins on an epithelial or mucosal surface and spreads along the mucosa and into deeper tissues. Infection may spread to cells that can enter blood vessels, lymphatics, or neural circuits. HBV, HCV, HTLV, and HIV are dependent on transmission by parenteral inoculation. Some viruses are transmitted only between humans. The dependence of smallpox virus and poliovirus infections on interhuman transmission makes it feasible to eliminate these viruses from human circulation by mass vaccination. Herpesviruses also survive by interhuman transmission but may be more difficult to eliminate because they establish persistent latent infection in humans and continuously reactivate to infect new and naïve generations.

Animals are also important reservoirs and vectors for transmission of viruses causing human disease. Insect vectors can mediate parenteral transfer of viruses that reach high titers in animal or human hosts. Arboviruses are parenterally transmitted from mammalian species to humans by mosquito vectors. Herpes B, monkeypox, rabies, and viral hemorrhagic fevers are other examples of zoonotic infections caused by direct contact with animals, animal tissues, or arthropod vectors.

PRIMARY INFECTION

Initial viral infections usually last for several days or weeks. During this period, the concentration of virus at sites of infection rises and then falls, usually to unmeasurable levels. The rise and fall of viral replication at a given site depend on local innate immune responses and the access of systemic antibody and cell immune effectors to the virus. Typically, primary infections with enteroviruses, mumps virus, measles virus, rubella virus, rotavirus, influenza virus, AAV, adenovirus, HSV, and VZV are cleared from almost all sites within 3–4 weeks. Some viruses are especially proficient in altering or evading innate and acquired immune responses. Primary infection with AAV, EBV, or CMV can last for several months. Characteristically, primary infections due to HBV, HCV, hepatitis D virus (HDV), HIV, HPV, and molluscum contagiosum virus (MCV) extend beyond several weeks. For some of these viruses (e.g., HPV, HBV, HCV, HDV, and MCV), the manifestations of primary infection are almost indistinguishable from the persistent phase.

Disease manifestations usually arise as a consequence of viral replication, infected-cell injury or death, and local inflammatory and innate immune responses. Disease severity may not necessarily correlate with the level of viral replication alone. For example, the clinical manifestations of intense primary infection with poliovirus, enterovirus, rabies virus, measles virus, mumps virus, or HSV at mucosal surfaces may be inapparent or relatively mild, whereas limited replication in neural cells can have dramatic consequences. Similarly, rubella virus or CMV infections in utero or neonatal HSV infections may have much more devastating effects than infections in adults.

Primary infections are cleared by nonspecific innate and specific adaptive immune responses. Thereafter, an immunocompetent host is usually immune to the disease manifestations of reinfection by the same virus. Immunity frequently does not prevent transient surface colonization on reexposure, persistent colonization, or even limited deeper infection.

PERSISTENT AND LATENT INFECTIONS

Relatively few viruses cause persistent or latent infections. HBV, HCV, rabies virus, measles virus, HIV, HTLV, HPV, HHVs, and MCV are notable exceptions. The mechanisms for persistent infection vary. HCV RNA polymerase and HIV reverse transcriptase are error-prone and generate variant genomes. Genome variation can be sufficient to permit evasion of host immune responses, thereby allowing persistent infection. HIV is also directly immunosuppressive, depleting CD4+ T lymphocytes and compromising CD8+ cytotoxic T cell immune responsiveness. Moreover, HIV encodes the Nef protein, which downmodulates MHC class I expression, rendering HIV-infected cells partially resistant to immune CD8+ T cell lysis.

DNA viruses have low mutation rates. Their persistence in human populations usually depends on their ability to establish latent infection in some cells, to reactivate from latency, and then to replicate at epithelial surfaces. *Latency* is defined as a state of infection in which virus is not replicating, viral genes associated with lytic infection are not expressed, and infectious virus is not made. The complete viral genome is present and may be replicated by cellular DNA polymerase in conjunction with replication of the cell's genome. HPVs establish latent infection in basal epithelial cells. The latently infected basal cell replicates, along with the HPV episome, by using cellular DNA polymerase. Some of the progeny cells provide new latently infected basal cells, whereas others go on to squamous differentiation. Infected cells that differentiate to squamous cells become permissive for lytic viral infection. Herpesviruses establish latent infection in nonreplicating neural cells (HSV and VZV) or in replicating cells of hematopoietic lineages (EBV, CMV, HHV-6, HHV-7, and Kaposi's sarcoma–associated herpesvirus [KSHV, also known as HHV-8]). In their latent stage, HPV and herpesvirus genomes are largely hidden from the normal immune response. Reactivated HPV and herpesvirus infections escape immediate and effective immune responses in highly immune hosts by inhibiting host innate immune and inflammatory responses. In addition, HPV, HSV, and VZV are somewhat protected because they replicate in the middle and upper layers of the squamous epithelium—sites not routinely visited by cells that mediate or amplify immune and inflammatory responses. HSV and CMV are also known to encode proteins that downregulate MHC class I expression and antigenic peptide presentation, enabling infected cells to escape recognition by and cytotoxic effects of CD8+ T lymphocytes.

Like other poxviruses, MCV cannot establish latent infection. This virus causes persistent infection in hypertrophic skin lesions that last for months or years. MCV encodes a chemokine homologue that probably blocks inflammatory responses, an MHC class I analogue that blocks cytotoxic T lymphocyte attack, and inhibitors of cell death that prolong infected-cell viability.

PERSISTENT VIRAL INFECTIONS AND CANCER

Persistent viral infection is estimated to be the root cause of as many as 20% of human malignancies. Cancer is an accidental and highly unusual or long-term effect of oncogenic human viral infection. With most "oncogenic viruses," infection is a critical and ultimately determinative early step in carcinogenesis. Latent HPV infection can block cell death and cause cervical cells to proliferate. A virus-infected cell with an integrated HPV genome

overexpressing E6 and E7 undergoes subsequent cellular genetic changes that enhance autonomous malignant cell growth.

Most hepatocellular carcinoma is believed to be caused by chronic inflammatory, immune, and regenerative responses to HBV or HCV infection. Epidemiologic data firmly link HBV and HCV infections to hepatocellular carcinoma. These infections elicit repetitive cycles of virus-induced liver injury followed by tissue repair and regeneration. Over decades, chronic viral infection, repetitive tissue regeneration, and acquired chromosomal changes can result in proliferative nodules. Further chromosomal mutations can lead to the degeneration of cells in a proliferating nodule into hepatocellular carcinoma. In rare instances, HBV DNA integrates into cellular DNA, promoting overexpression of a cell gene that can also contribute to oncogenesis.

Most cervical carcinoma is caused by persistent infection with “high-risk” HPV type 16 or 18. In contrast to HBV and HCV infections, which stimulate cell growth as a consequence of virus-induced cell death, HPV type 16 or 18 proteins E6 and E7 destroy p53 and pRB, respectively. Elimination of these key tumor-suppressive cell proteins increases cell growth, cell survival, and cell genome instability. However, like HBV and HCV infections, HPV infection alone is not sufficient for carcinogenesis. Cervical carcinoma is inevitably associated with persistent HPV infection and integration of the HPV genome into chromosomal DNA. Integrations that result in overexpression of E6 and E7 from HPV type 16 or 18 cause more profound changes in cell growth and survival and permit subsequent chromosomal changes that result in cervical carcinoma.

EBV is the most unusual oncogenic virus in that normal B cell infection results in latency with expression of viral proteins that can cause endless B lymphocyte growth. In almost all humans, strong CD4+ and CD8+ T cell immune responses to the antigenic EBV latent-infection nuclear proteins prevent uncontrolled B cell lymphoproliferation. However, when humans are severely immunosuppressed by transplantation-associated medication, HIV infection, or genetic immune deficiencies, EBV-induced B cell malignancies can emerge.

EBV infection also has a role in the long-term development of B lymphocyte and epithelial cell malignancies. Persistent EBV infection with expression of an EBV latency-associated integral membrane protein (LMP1) in latently infected epithelial cells appears to be a critical early step in the evolution of anaplastic nasopharyngeal carcinoma, a common malignancy in populations in southern China and northern Africa. Genomic instability and chromosomal abnormalities also contribute to the development of EBV-associated nasopharyngeal carcinoma. EBV is an important cause of Hodgkin’s lymphoma. High-level expression of LMP1 or LMP2 in Reed-Sternberg cells is a hallmark in up to 50% of Hodgkin’s lymphoma cases. LMP1-induced nuclear factor κB (NF-κB) activity may prolong the survival of defective B cells that are normally eliminated by apoptosis, thereby allowing other genetic changes leading to the development of malignant Reed-Sternberg cells.

The HTLV-1 Tax and Rex proteins are critical to the initiation of cutaneous adult T cell lymphoma/leukemias that occur long after primary HTLV-1 infection. Tax-induced NF-κB activation may contribute to cytokine production, infected-cell survival, and eventual outgrowth of malignant cells.

Molecular data confirm the presence of KSHV DNA in all Kaposi’s tumors, including those associated with HIV infection, transplantation, and familial transmission. KSHV infection is also etiologically implicated in pleural-effusion lymphomas and multicentric Castleman’s disease, which are more common among HIV-infected than among HIV-uninfected people. KSHV has a virus-encoded cyclin, an IFN regulatory factor, and a latency-associated nuclear antigen that are implicated in increased cell proliferation and survival.

Evidence supporting a causal role for viral infection in all of these malignancies includes (1) epidemiologic data, (2) the presence of viral DNA in all tumor cells, (3) the ability of the viruses to transform human cells in culture, (4) the results of in vitro cell culture–based assays that reveal transforming effects of specific viral genes on cell growth or survival, (5) pathologic data indicating the expression of transforming viral genes in premalignant or malignant cells in vivo, (6) the demonstration in animal models that these viral genes can cause malignant cell growth, and (7) the ability of virus-specific vaccines to reduce the incidence of virus-associated malignancy.

Virus-related malignancies provide an opportunity to expand our understanding of the biologic mechanisms important in the development of cancer. They also offer unique opportunities to develop diagnostics, vaccines, or therapeutics that could prevent or specifically treat cancers associated with viral infection. Widespread immunization against hepatitis B has resulted in a decreased prevalence of HBV-associated hepatitis and will probably prevent most HBV-related liver cancers. Current HPV vaccines can reduce rates of colonization with high-risk HPV strains and thereby decrease the risk of cervical cancer. The successful use of in vitro–expanded EBV-specific T cell populations to treat or prevent EBV-associated posttransplantation lymphoproliferative disease demonstrates the potential of immunoprevention or immunotherapy against virus-associated cancers.

RESISTANCE TO VIRAL INFECTIONS

Resistance to viral infections is initially provided by factors that are not virus-specific. Physical protection is afforded by the cornified layers of the skin and by mucous secretions that continuously sweep over mucosal surfaces. Once the first cell is infected, IFNs are induced and confer resistance to RNA virus replication. Viral infection may also trigger the release of other cytokines from infected cells. These cytokines may be chemotactic to inflammatory and immune cells. Viral protein epitopes expressed on the cell surface in the context of MHC class I and II proteins can stimulate the expansion of T cell populations with receptors that can recognize virus-encoded peptides presented on the cell surface by MHC class I proteins. Cytokines and antigens released by virus-induced cell death further attract inflammatory cells, dendritic cells, granulocytes, natural killer (NK) cells, and B lymphocytes to sites of infection and to draining lymph nodes. IFNs and NK cells are particularly important in containing viral infection for the first several days. Granulocytes and macrophages are also important in the phagocytosis and degradation of viruses, especially after an initial antibody response.

By 7–10 days after infection, virus-specific antibody responses, virus-specific human leukocyte antigen (HLA) class II–restricted CD4+ helper T lymphocyte responses, and virus-specific HLA class I–restricted CD8+ cytotoxic T lymphocyte responses develop. These responses, whose magnitude typically increases over the second and third weeks of infection, are important for rapid recovery. Also between the second and third weeks, the antibody type usually changes from IgM to IgG; IgG or IgA antibody can then be detected at infected mucosal surfaces. Antibody may directly neutralize virus by binding to its surface and preventing cell attachment or penetration. Complement can significantly enhance antibody-mediated virus neutralization. Antibody and complement can also lyse virus-infected cells that express viral membrane proteins on the cell surface. Cells infected with a replicating enveloped virus usually express the virus-envelope glycoproteins on the cell plasma membrane. Specific antibodies can bind to the glycoproteins, fix complement, and lyse the infected cell.

Antibody and CD4+/CD8+ T lymphocyte responses to viral infection can remain at high levels for several months after primary infection but usually wane over time. Low-level persistence of antibody-producing B lymphocytes and CD4+ or CD8+ T lymphocyte responses as memory cells can provide a rapid response to a second infection or

an early barrier to reinfection with the same virus. Redevelopment of T cell immunity may take longer than secondary antibody responses, particularly when many years have elapsed between primary infection and reexposure. However, persistent infections or frequent reactivations from latency can result in sustained high-level T cell responses. EBV and CMV typically induce high-level CD4+ and CD8+ T cell responses that are maintained for decades after primary infection.

Some viruses have genes that alter innate and acquired host defenses. Adenoviruses encode small RNAs that inhibit IFN-induced, protein kinase R (PKR)–mediated shutoff of infected-cell protein synthesis. Adenovirus E1A can also directly inhibit IFN-mediated changes in cell gene transcription. Moreover, adenovirus E3 proteins prevent tumor necrosis factor (TNF)–induced cytolysis and block HLA class I synthesis by the infected cell. HSV ICP47 and CMV US11 also block class I antigen presentation. EBV encodes an interleukin (IL) 10 homologue that inhibits NK and T cell responses. Vaccinia virus encodes a soluble receptor for IFN-α and binding proteins for IFN-γ, IL-1, IL-18, and TNF, which inhibit host innate and adaptive immune responses. Vaccinia virus also encodes a caspase inhibitor that inhibits the ability of CD8+ cytotoxic T cells to kill virus-infected cells. Some poxviruses and herpesviruses encode chemokine-binding proteins that inhibit cell inflammatory responses. The adoption of these strategies by viruses highlights the importance of the corresponding host resistance factors in containing viral infection and the importance of redundancy in host resistance.

The host inflammatory and immune responses to viral infection do not come without a price. These responses contribute to the symptoms, signs, and other pathophysiologic manifestations of viral infection. Inflammation at sites of viral infection can subvert an effective immune response and induce tissue death and dysfunction. Moreover, immune responses to viral infection could, in principle, result in immune attack upon cross-reactive epitopes on normal cells, with consequent autoimmunity.

INTERFERONS

All human cells can synthesize IFN-α or IFN-β in response to viral infection. These IFN responses are usually induced by the presence of double-strand viral RNA, which can be made by both RNA and DNA viruses and sensed by double-strand RNA binding proteins (e.g., PKR and RIG-I) in the cell cytoplasm. IFN-γ is not closely related to IFN-α or IFN-β and is produced mainly by NK cells and by immune T lymphocytes responding to IL-12. IFN-α and -β bind to the IFN-α receptor, whereas IFN-γ binds to a different but related receptor. Both receptors signal through receptor-associated JAK kinases and other cytoplasmic proteins, including “STAT” proteins, which are tyrosine-phosphorylated by JAK kinases, translocate to the nucleus, and activate promoters for specific cell genes. Three types of antiviral effects are induced by IFN at the transcriptional level. The first effect is attributable to the induction of 2′-5′ oligo(A) synthetases, which require double-strand RNA for their activation. Activated synthetase polymerizes oligo(A) and thereby activates RNAse L, which in turn degrades single-strand RNA. A second effect results from the induction of PKR, a serine and threonine kinase that is also activated by double-strand RNA. PKR phosphorylates and negatively regulates the translational initiation factor eIF2α, shutting down protein synthesis in the infected cell. A third effect is initiated through the induction of Mx proteins, a family of GTPases that is particularly important in inhibiting the replication of influenza virus and vesicular stomatitis virus. These IFN effects are mostly directed against the infected cell, causing virus and cell dysfunction and thereby limiting viral replication.

DIAGNOSTIC VIROLOGY

A wide variety of methods are used to diagnose viral infection. Serology and virus isolation in tissue culture remain important standards. Acute- and convalescent-phase sera with rising titers of antibody to virus-specific antigens and a shift from IgM to IgG antibodies are generally accepted as diagnostic of acute viral infection. Serologic diagnosis is based on a more than fourfold rise in IgG antibody concentration when acute- and convalescent-phase serum samples are analyzed at the same time.

Immunofluorescence, hemadsorption, and hemagglutination assays for antiviral antibodies are labor-intensive and have been replaced by enzyme-linked immunosorbent assays (ELISAs), which generally use the specific viral proteins most frequently targeted by the antibody response. The proteins are purified from virus-infected cells or produced by recombinant DNA technology and are attached to a solid phase, where they can be incubated with serum, washed to eliminate nonspecific antibodies, and allowed to react with an enzyme-linked reagent to detect human IgG or IgM antibody specifically adhering to the viral antigen. The amount of antibody can then be quantitated by the intensity of a color reaction mediated by the linked enzyme. ELISAs can be sensitive and automated. Western blots can simultaneously confirm the presence of antibody to multiple specific viral proteins. The proteins are separated by size and transferred to an inert membrane, where they are incubated with serum antibodies. Western blots have an internal specificity control because the level of reactivity for viral proteins can be compared with that for cellular proteins in the same sample. Western blots require individual evaluation and are inherently difficult to quantitate or automate.

Isolation of virus in tissue culture depends on infection and replication in susceptible cells. Growth of virus in cell cultures can frequently be identified by effects on cell morphology under light microscopy. For example, HSV produces a typical cytopathic effect in rabbit kidney cells within 3 days. Other viral cytopathic effects may not be as diagnostically distinctive. Identification usually requires confirmation by staining with virus-specific monoclonal antibodies. The efficiency and speed of virus identification can be enhanced by combining short-term culture with immune detection. In assays with “shell vials” of tissue culture cells growing on a coverslip, viral infection can be detected by staining with a monoclonal antibody to a specific viral protein expressed early in viral replication. Thus, virus-infected cells can be detected within hours or days of inoculation, whereas several rounds of infection would be required to produce visible cytopathic effects.

Isolation of virus in tissue culture also depends on the collection of specimens from appropriate sites and the rapid transport of these specimens in appropriate medium to the virology laboratory (**Chap. S13**). Rapid transport maintains viral viability and limits bacterial and fungal overgrowth. Enveloped viruses are generally more sensitive to freezing and thawing than nonenveloped viruses. The most appropriate site for culture depends on the pathogenesis of the virus in question. Nasopharyngeal, tracheal, or endobronchial aspirates are most appropriate for the identification of respiratory viruses. Sputum cultures generally are less appropriate because bacterial contamination and viscosity threaten tissue-culture cell viability. Aspirates of vesicular fluid are useful for isolation of HSV and VZV. Nasopharyngeal aspirates and stool specimens may be useful when the patient has fever and a rash and an enteroviral infection is suspected. Adenoviruses can be cultured from the urine of patients with hemorrhagic cystitis. CMV can frequently be isolated from cultures of urine or buffy coat. Biopsy material can be effectively cultured when viruses infect major organs, as in HSV encephalitis or adenovirus pneumonia.

The isolation of a virus does not necessarily establish disease causality. Viruses can persistently or intermittently colonize normal human mucosal surfaces. Saliva can be positive for herpesviruses, and normal urine samples can

be positive for CMV. Isolations from blood, cerebrospinal fluid (CSF), or tissue are more often diagnostic of significant viral infection.

Another method aimed at increasing the speed of viral diagnosis is direct testing for antigen or cytopathic effects. Virus-infected cells from the patient may be detected by staining with virus-specific monoclonal antibodies. For example, epithelial cells obtained by nasopharyngeal aspiration can be stained with a variety of specific monoclonal antibodies to identify the specific infecting respiratory virus. Antigen and serologic assays can be multiplexed to detect multiple analytes simultaneously by coupling of reagents to color-coded beads for each analyte and detection by flow cytometry.

Nucleic acid amplification techniques bring speed, sensitivity, and specificity to diagnostic virology. The ability to directly amplify minute amounts of viral nucleic acids in specimens means that detection no longer depends on viable virus and its replication. For example, amplification and detection of HSV nucleic acids in the CSF of patients with HSV encephalitis is a more sensitive detection method than culture of virus from CSF. The extreme sensitivity of these tests can be a problem, because subclinical infection or contamination can lead to false-positive results. Detection of viral nucleic acids does not necessarily indicate virus-induced disease.

Measurement of the amount of viral RNA or DNA in peripheral blood is an important means for determining whether a patient is at increased risk for virus-induced disease and for evaluating clinical responses to antiviral chemotherapy. Nucleic acid technologies for RNA quantification are routinely used in AIDS patients to evaluate responses to antiviral agents and to detect viral resistance or noncompliance with therapy. Virus-load measurements are also useful for evaluating the treatment of patients with HBV and HCV infections. Nucleic acid testing or direct staining with CMV-specific monoclonal antibodies to quantitate virus-infected cells in the peripheral blood (CMV antigenemia) is useful for identifying immunosuppressed patients who may be at risk for CMV-induced disease.

DRUG TREATMENT FOR VIRAL INFECTIONS

Multiple steps in the life cycles of viruses can be effectively targeted by antiviral drugs (**Chaps. 186** and **197**). Nucleoside and nonnucleoside reverse transcriptase inhibitors prevent HIV provirus synthesis, whereas protease inhibitors block maturation of the HIV and HCV polyprotein after infection of the cell. Enfuvirtide is a small peptide derived from HIV gp41 that acts before cell infection by preventing a conformational change required for initial fusion of the virus with the cell membrane. Raltegravir is an integrase inhibitor that is approved for use with other anti-HIV drugs. Amantadine and rimantadine inhibit the influenza M2 protein, preventing release of viral RNA early during infection, whereas zanamivir and oseltamivir inhibit the influenza neuraminidase, which is necessary for the efficient release of mature virions from infected cells.

Viral genomes can evolve resistance to drugs by mutation and selection, by recombination with a drug-resistant virus, or (in the case of influenza virus and other segmented RNA viral genomes) by reassortment. The emergence of drug-resistant strains can limit the efficacy of antiviral therapy. As in antibacterial therapy, excessive and inappropriate use of antiviral therapy can select for the emergence of drug-resistant strains. HIV genotyping is a rapid method for identifying drug-resistant viruses. Resistance to reverse transcriptase or protease inhibitors has been associated with specific mutations in the reverse transcriptase or protease genes. Identification of these mutations by polymerase chain reaction amplification and nucleic acid sequencing can be clinically useful for

determining which antiviral agents may still be effective. Drug resistance also can arise in herpesviruses but is a less common clinical problem.

IMMUNIZATION FOR THE PREVENTION OF VIRAL INFECTIONS

Viral vaccines are among the outstanding accomplishments of medical science. Smallpox has been eradicated except as a potential weapon of biological warfare or bioterrorism. Poliovirus eradication may soon follow. Measles can be contained or eliminated. Excess mortality due to influenza virus epidemics can be prevented, and the threat of influenza pandemics can be decreased by contemporary killed or live attenuated influenza vaccines. Mumps, rubella, and chickenpox are well controlled by childhood vaccination in the developed world. Reimmunization of mature adults can be used to control herpes zoster. New rotavirus vaccines can have a major impact on this leading cause of gastroenteritis and prominent cause of childhood death worldwide. Widespread HBV vaccination has dramatically lowered the frequency of acute and chronic hepatitis and is expected to lead to a dramatic decrease in the incidence of hepatocellular carcinoma. The HPV vaccine was the first vaccine specifically licensed to prevent virus-induced cancer. Use of purified proteins, genetically engineered live-virus vaccines, and recombinant DNA–based strategies will make it possible to immunize against severe infections with other viruses. The development of effective HIV and HCV vaccines is complicated by the high mutation rate of viral RNA polymerase and reverse transcriptase, the population-based and individual divergence of HIV or HCV genomes, and repeated high-level exposure in some populations. Concerns about the use of smallpox and other viruses as weapons necessitate maintenance of immunity to agents that are not encountered naturally.

VIRUSES AS NOVEL THERAPEUTIC TOOLS OR AGENTS

Viruses are being used experimentally to deliver biotherapeutic agents or novel vaccines. Foreign genes can be inserted into viral nucleic acids, and the recombinant virus vectors can be used to infect the patient or the patient's cells ex vivo. Retrovirus integration into the cell genome has been used to functionally replace the abnormal gene in T cells of patients with severe combined immunodeficiency, thereby restoring immune function. Recombinant adenovirus, AAV, and retroviruses are being explored for use in diseases due to single-gene defects, such as cystic fibrosis and hemophilia. AAV carrying a lipoprotein lipase gene is now being used in Europe to treat a rare lipid-processing disease and is the first gene therapy approved for clinical use. Recombinant poxviruses, adenoviruses, and influenza viruses are also being used experimentally as vaccine vectors. Viral vectors are being tested experimentally for the expression of cytokines that can enhance immunity against tumor cells or for the expression of proteins that can increase the sensitivity of tumor cells to chemotherapy. HSV deficient for replication in resting cells is being used to selectively kill proliferating glioblastoma cells after injections into CNS tumors. For improved safety, nonreplicating viruses are frequently used in clinical trials. Potential adverse events associated with virus-mediated gene transfer include the induction of inflammatory and antiviral immune responses. Instances of retrovirus-induced human malignancies have raised concerns about the safety of retroviral gene therapy vectors.

FURTHER READING

Chavali PL et al: Neurodevelopmental protein Musashi-1 interacts with the Zika genome and promotes viral replication. Science 357:83, 2017. (See also commentary: Griffin DE: Why are neurons susceptible to Zika virus? Science 357:33, 2017.)

Henao-Restrepo AM et al: Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: Final results from the Guinea ring vaccination, open-label, cluster-randomised trial. Lancet 389:505, 2017. (See also commentary: Geisbert TW: First Ebola virus vaccine to protect human beings? Lancet 389:479, 2017.)

Knipe DM, Howley PM (eds): *Fields Virology*, 6th ed. Philadelphia, Wolters Kluwer Health/Lippincott Williams & Wilkins, 2013.

Schulz F et al: Giant viruses with an expanded complement of translation system components. Science 356:82, 2017. (See also commentary: Leslie M: Cell-like giant viruses found. Science 356:15, 2017.)

Chapter 186: Antiviral Chemotherapy, Excluding Antiretroviral Drugs

Lindsey R. Baden

INTRODUCTION

The field of antiviral therapy—both the number of antiviral drugs and our understanding of their optimal use—historically has lagged behind that of antibacterial treatment, but significant progress has been made in recent years on new drugs for several viral infections. The development of antiviral drugs poses several challenges. Viruses replicate intracellularly and often use host cell enzymes, macromolecules, and organelles for synthesis of viral particles. Therefore, useful antiviral compounds must discriminate between host and viral functions with a high degree of specificity; agents without such selectivity are likely to be too toxic for clinical use.

Significant progress has also been made in the development of laboratory assays to assist clinicians in the appropriate use of antiviral drugs. Phenotypic and genotypic assays for resistance to antiviral drugs are becoming more widely available, and correlations of laboratory results with clinical outcomes are being better defined. Of particular note has been the development of highly sensitive and specific methods that measure the concentration of virus in blood (*viral load*) and permit direct assessment of the antiviral effect of a given drug regimen in that host site. Viral load measurements have been useful in recognizing the risk of disease progression in patients with viral infections and in identifying patients for whom antiviral chemotherapy might be of greatest benefit. As with any in vitro laboratory test, results are highly dependent on and likely vary with the laboratory techniques used.

Information regarding the pharmacodynamics of antiviral drugs, and particularly the relationship of concentration effects to efficacy, has been slow to develop but is also expanding. However, assays to measure concentrations of antiviral drugs, especially of their active moieties within cells, are still primarily research procedures not widely available to clinicians. Thus, there are limited guidelines for adjusting dosages of antiviral agents to maximize antiviral activity and minimize toxicity. Consequently, clinical use of antiviral drugs must be accompanied by particular vigilance for unanticipated adverse effects.

Like that of other infections, the course of viral infections is profoundly affected by interplay between the pathogen and a complex set of host defenses. The presence or absence of preexisting immunity, the ability to mount humoral and/or cell-mediated immune responses, and the stimulation of innate immunity are important determinants of the outcome of viral infections. The state of the host’s defenses needs to be considered when antiviral agents are used or evaluated.

As with any therapy, the optimal use of antiviral compounds requires a specific and timely diagnosis. For some viral infections, such as herpes zoster, the clinical manifestations are so characteristic that a diagnosis can be made on clinical grounds alone. For other viral infections, such as influenza A, epidemiologic information (e.g., the

documentation of a community-wide influenza outbreak) can be used to make a presumptive diagnosis with a high degree of accuracy. However, for most of the remaining viral infections, including herpes simplex encephalitis, cytomegalovirus (CMV) infections other than retinitis, and enterovirus infections, diagnosis on clinical grounds alone cannot be accomplished with certainty. For such infections, rapid viral diagnostic techniques are of great importance. Considerable progress has been made in recent years in the development of such tests, which are now widely available for a number of viral infections.

Despite these complexities, the efficacy of a number of antiviral compounds has been clearly established in rigorously conducted and controlled studies. As summarized in **Table 186-1**, this chapter reviews the antiviral drugs that are currently approved or are likely to be considered for approval in the near future for use against viral infections other than those caused by HIV. **Antiretroviral drugs are reviewed in Chap. 197**.

TABLE 186-1

Antiviral Chemotherapy and Chemoprophylaxis

Infection(s)	Drug	Route	Dosage	Comment
Influenza A and B: treatment	Oseltamivir	Oral	Adults: 75 mg bid × 5 d Children 1–12 years: 30–75 mg bid, depending on weight,[a] × 5 d	When started within 2 days of onset in uncomplicated disease, zanamivir and oseltamivir reduce symptom duration by 1.0–1.5 and 1.3 days, respectively. Their effectiveness in prevention or treatment of complications is unclear, although some analyses suggest that oseltamivir may reduce the frequency of respiratory tract complications and hospitalizations. Oseltamivir's side effects of nausea and vomiting can be reduced in frequency by drug administration with food. Zanamivir may exacerbate bronchospasm in patients with asthma. Amantadine and rimantadine are not recommended for routine use unless antiviral susceptibilities are known because of widespread resistance in A/H3N2 viruses since 2005–2006 and in pandemic A/H1N1 viruses in 2009–2010. Their efficacy in treatment of uncomplicated disease caused by sensitive viruses has been similar to that of neuraminidase inhibitors.
	Zanamivir	Inhaled orally	Adults and children ≥7 years: 10 mg bid × 5 d	
Influenza A: treatment	Amantadine[b]	Oral	Adults: 100 mg qd or bid × 5–7 d Children 1–9 years: 5 mg/kg per day (maximum, 150 mg/d) × 5–7 d	
	Rimantadine[b]	Oral	100 mg qd or bid × 5–7 d in adults	
Influenza A and B: prophylaxis	Oseltamivir	Oral	Adults: 75 mg/d Children ≥1 year: 30–75	Prophylaxis must be continued for the duration of exposure and can be administered simultaneously with

Infection(s)	Drug	Route	Dosage	Comment
			mg/d depending on weight[a]	inactivated vaccine. Unless the sensitivity of isolates is known, neither amantadine nor rimantadine is currently recommended for prophylaxis or therapy.
	Zanamivir	Inhaled orally	Adults and children ≥5 years: 10 mg/d	
Influenza A: prophylaxis	Amantadine[b] or rimantadine[b]	Oral	Adults: 200 mg/d Children 1–9 years: 5 mg/kg per day (maximum, 150 mg/d)	
RSV infection	Ribavirin	Small-particle aerosol	Administered 12–18 h/d from a reservoir containing 20 mg/mL × 3–6 d	Use of ribavirin is to be considered for treatment of infants and young children hospitalized with RSV pneumonia and bronchiolitis, according to the American Academy of Pediatrics.
CMV disease	Ganciclovir	IV	5 mg/kg bid × 14–21 d; then 5 mg/kg per day as maintenance dose	Ganciclovir, valganciclovir, foscarnet, and cidofovir are approved for treatment of CMV retinitis in patients with AIDS. They are also used for colitis, pneumonia, or "wasting" syndrome associated with CMV and for prevention of CMV disease in transplant recipients.
	Valganciclovir	Oral	900 mg bid × 21 d; then 900 mg/d as maintenance	Valganciclovir has largely supplanted oral ganciclovir and is frequently used in place of IV ganciclovir.

Infection(s)	Drug	Route	Dosage	Comment
			dose	
	Foscarnet	IV	60 mg/kg q8h × 14–21 d; then 90–120 mg/kg per day as maintenance dose	Foscarnet is not myelosuppressive and is active against acyclovir- and ganciclovir-resistant herpesviruses.
	Cidofovir	IV	5 mg/kg once weekly × 2 weeks, then once every other week; given with probenecid and hydration	—
	Fomivirsen	Intravitreal	330 mg on days 1 and 15 followed by 330 mg monthly as maintenance	Fomivirsen has reduced the rate of progression of CMV retinitis in patients in whom other regimens have failed or have not been well tolerated. The major form of toxicity is ocular inflammation.
Varicella: immunocompetent host	Acyclovir	Oral	20 mg/kg (maximum, 800 mg) 4 or 5 times daily × 5 d	Treatment confers modest clinical benefit when administered within 24 h of rash onset.
	Valacyclovir	Oral	Children 2–18 years: 20 mg/kg tid (not to	

Infection(s)	Drug	Route	Dosage	Comment
			exceed 1 g tid) × 5 d	
Varicella: immunocompromised host	Acyclovir	IV	10 mg/kg q8h × 7 d	A change to oral valacyclovir can be considered once fever has subsided if there is no evidence of visceral involvement.
Herpes simplex encephalitis	Acyclovir	IV	10 mg/kg q8h × 14–21 d	Results are optimal when therapy is initiated early. Some authorities recommend treatment for 21 d to prevent relapses.
Neonatal herpes simplex	Acyclovir	IV	20 mg/kg q8h × 14–21 d	Serious morbidity is common despite therapy. Prolonged oral administration after initial IV therapy has been suggested because of long-term sequelae associated with cutaneous recurrences of HSV infection.
Genital herpes simplex, primary: treatment	Acyclovir	IV	5 mg/kg q8h × 5–10 d	The IV route is preferred for infections severe enough to warrant hospitalization or with neurologic complications.
		Oral	400 mg tid or 200 mg 5 times daily × 7–10 d	The oral route is preferred for patients whose condition does not warrant hospitalization. Adequate hydration must be maintained.
		Topical	5% ointment; 4–6 applications	Topical use—largely supplemented by oral therapy—may obviate systemic administration to pregnant

Infection(s)	Drug	Route	Dosage	Comment
			daily × 7–10 d	women. Systemic symptoms and untreated areas are not affected.
	Valacyclovir	Oral	1 g bid × 7–10 d	Valacyclovir appears to be as effective as acyclovir but can be administered less frequently.
	Famciclovir	Oral	250 mg tid × 7–10 d[c]	Famciclovir appears to be similar in effectiveness to acyclovir.
Genital herpes simplex, recurrent: treatment	Acyclovir	Oral	400 mg tid × 5 d or 800 mg tid × 2 d	The clinical effect is modest and is enhanced if therapy is initiated early. Treatment does not affect recurrence rates.
	Famciclovir	Oral	125 mg bid × 5 d, 1000 mg bid × 1 d, or 500 mg once, then 250 mg PO bid × 3 doses	
	Valacyclovir	Oral	500 mg bid × 3 d or 1 g once a day × 5 d	
Genital herpes simplex, recurrent: suppression	Acyclovir	Oral	400 mg bid	Suppressive therapy is recommended only for patients with at least 6–10 recurrences per year. "Breakthrough" occasionally takes place, and asymptomatic shedding of virus occurs. The need for suppressive therapy should be reevaluated after 1 year. Suppression with valacyclovir reduces transmission of genital HSV
	Valacyclovir	Oral	500–1000 mg/d or 250–500 mg bid	
	Famciclovir	Oral	250 mg bid	

Infection(s)	Drug	Route	Dosage	Comment
				among virus-discordant couples.
Mucocutaneous herpes simplex in immunocompromised host: treatment	Acyclovir	IV	5 mg/kg q8h × 7–14 d	The choice of the IV or oral route and the duration of therapy depend on the severity of infection and the patient's ability to take oral medication. Oral or IV treatment has supplanted topical therapy except for small, easily accessible lesions. Foscarnet is used for acyclovir-resistant viruses.
		Oral	400 mg 5 times daily × 10–14 d	
		Topical	5% ointment; 4–6 applications daily × 7 d or until healed	
	Valacyclovir	Oral	1 g tid × 7–10 d[c]	
	Famciclovir	Oral	500 mg bid × 7–10 d[d]	
Mucocutaneous herpes simplex in immunocompromised host: prevention of recurrence during intense immunosuppression	Acyclovir	Oral	400 mg 2–5 times daily or 800 mg bid	Treatment is administered during periods when intense immunosuppression is expected—e.g., during antitumor chemotherapy or after transplantation—and is usually continued for 2–3 months.
		IV	5 mg/kg q12h	
	Valacyclovir	Oral	500 mg to 1 g bid or tid	
	Famciclovir	Oral	500 mg bid[c]	
Herpes simplex orolabialis, recurrent[e]	Penciclovir	Topical	1.0% cream applied q2h during waking hours	Treatment shortens healing time and symptom duration by 0.5–1.0 d (versus placebo).

Infection(s)	Drug	Route	Dosage	Comment
			× 4 d	
	Valacyclovir	Oral	2 g q12h × 1 d	Therapy begun at earliest symptom reduces disease duration by 1 d.
	Famciclovir[c]	Oral	1500 mg once or 750 mg bid × 1 d	Therapy begun within 1 h of prodrome decreases time to healing by 1.8–2.2 d.
	Docosanol[f]	Topical	10% cream 5 times daily until healed	Application at initial symptoms reduces healing time by 1 d.
Herpes simplex keratitis	Trifluridine	Topical	1 drop of 1% ophthalmic solution q2h while awake (maximum, 9 drops daily)	Therapy should be undertaken in consultation with an ophthalmologist.
	Vidarabine	Topical	0.5-in. ribbon of 3% ophthalmic ointment 5 times daily	
Herpes zoster: immunocompetent host	Valacyclovir	Oral	1 g tid × 7 d	Valacyclovir may be more effective than acyclovir for pain relief; otherwise, it has a similar effect on cutaneous lesions and should be given within 72 h of rash onset.
	Famciclovir	Oral	500 mg q8h × 7 d	The duration of postherpetic neuralgia is shorter than with placebo. Famciclovir showed overall efficacy similar to that

Infection(s)	Drug	Route	Dosage	Comment
				of acyclovir in a comparative trial. It should be given ≤72 h after rash onset.
	Acyclovir	Oral	800 mg 5 times daily × 7–10 d	Acyclovir causes faster resolution of skin lesions than placebo and provides some relief of acute symptoms if given within 72 h of rash onset. Combined with tapering doses of prednisone, acyclovir improves quality-of-life outcomes.
Herpes zoster: immunocompromised host	Acyclovir	IV	10 mg/kg q8h × 7 d	Effectiveness in localized zoster is most marked when treatment is given early. Foscarnet may be used for acyclovir-resistant VZV infections.
		Oral	800 mg 5 times daily × 7 d	
	Valacyclovir	Oral	1 g tid × 7 d[c]	
	Famciclovir	Oral	500 mg tid × 10 d[c]	
Herpes zoster ophthalmicus	Acyclovir	Oral	600–800 mg 5 times daily × 10 d	Treatment reduces ocular complications, including ocular keratitis and uveitis.
	Valacyclovir	Oral	1 g tid × 7 d	
	Famciclovir	Oral	500 mg tid × 7 d	

Condyloma acuminatum	IFN-α2b	Intralesional	1 million units per wart (maximum of	Intralesional treatment frequently results in regression of warts, but lesions often

Infection(s)	Drug	Route	Dosage	Comment
			5) thrice weekly × 3 weeks	recur. Parenteral administration may be useful if lesions are numerous.
	IFN-αn3	Intralesional	250,000 units per wart (maximum of 10) twice weekly × up to 8 weeks	
Chronic hepatitis B	IFN-α2b	SC	5 million units daily or 10 million units thrice weekly × 16–24 weeks	HBeAg and DNA are eliminated in 33–37% of cases. Histopathologic improvement is also seen.
	Pegylated IFN-α2a	SC	180 µg weekly × 48 weeks	ALT levels return to normal in 39% of patients, and histologic improvement occurs in 38%.
	Lamivudine	Oral	100 mg/d × 12–18 months; 150 mg bid as part of therapy for HIV infection	Lamivudine monotherapy is well tolerated and effective in reduction of HBV DNA levels, normalization of ALT levels, and improvement in histopathology. However, resistance develops in 24% of recipients when lamivudine is used as monotherapy for 1 year.
	Adefovir dipivoxil	Oral	10 mg/d × 48 weeks	A return of ALT levels to normal is documented in 48–72% of recipients and improved liver histopathology in 53–64%.

Infection(s)	Drug	Route	Dosage	Comment
				Adefovir is effective in lamivudine-resistant hepatitis B. Renal function including proteinuria should be monitored.
	Entecavir	Oral	0.5 mg/d × 48 weeks (1 mg/d if HBV is resistant to lamivudine)	Normalization of ALT is seen in 68–78% of recipients and loss of HBeAg in 21%. Entecavir is active against lamivudine-resistant HBV.
	Telbivudine	Oral	600 mg/d × 52 weeks	HBV DNA is reduced by >5 $\log_{10}$ copies/mL along with normalization of ALT levels in 74–77% of patients and improved histopathology in 65–67%. Resistance develops in 9–22% of patients after 2 years of therapy. Elevated CPK levels and myopathy may occur.
	Tenofovir disoproxil	Oral	300 mg/d × 48 weeks	ALT levels return to normal in 68–76% of patients, and liver histopathology improves in 72–74%. Resistance is uncommon with up to 2 years of therapy. Initial data suggest a better safety profile (renal and bone) for tenofovir alafenamide than for tenofovir disoproxil.
	Tenofovir alafenamide	Oral	25 mg/d × 48 weeks	
Chronic hepatitis C	Sofosbuvir[g]/ pegylated IFN-α2a or IFN-α2b/ribavirin	Oral/SC	HCV genotypes 1, 4, 5, and 6: sofosbuvir (400 mg qd) with daily weight-based ribavirin (1000 mg	Sofosbuvir is generally well tolerated, and most common side effects have been attributable to concomitantly administered IFN and ribavirin. Sofosbuvir is recommended in triple combination with pegylated IFN and ribavirin for genotypes 1, 4, 5, and 6, with

Infection(s)	Drug	Route	Dosage	Comment
			[<75 kg] to 1200 mg [>75 kg]) and weekly pegylated IFN (180 µg per week) for 12 weeks Genotypes 2 and 3: sofosbuvir (400 mg qd) with daily weight-based ribavirin for 12 and 24 weeks, respectively	SVRs in 89–97% of treatment-naïve patients.
	Simeprevir[g]/ pegylated IFN-α2b/ribavirin	Oral/SC/oral	Alternative regimen for genotypes 1 and 4: simeprevir (50 mg qd) for 12 weeks plus daily ribavirin and	Simeprevir has supplanted the first-generation protease inhibitors boceprevir and telaprevir. Its metabolism by cytochrome CYP3A can result in interactions with other drugs. Photosensitivity and reversible hyperbilirubinemia are associated toxicities.

Infection(s)	Drug	Route	Dosage	Comment
			weekly pegylated IFN for 24–28 weeks, respectively	Testing for the Q80K-resistant variant should be carried out since this variant is present in one-third of HCV genotype 1a infections. Triple combinations of simeprevir with pegylated IFN and ribavirin result in SVRs in 80% of genotype 1 infections without Q80K. Combination therapy results in SVR in up to 40–50% of recipients.
	IFN-α2a or IFN-α2b	SC	9 million units thrice weekly × 12 months	The overall efficacy and the optimal regimen and duration of therapy are not fully established. Sustained SVRs have been seen in 25–30% of patients for IFN-α and in 17–43% for pegylated IFN-α. The slower clearance of pegylated IFNs than of standard IFNs permits once-weekly administration. Pegylated formulations appear to be superior to standard IFNs in efficacy, both as monotherapy and in combination with ribavirin, and have largely supplanted standard IFNs in treatment of hepatitis C. SVRs were seen in 42–51% of patients infected with HCV genotype 1 and in 76–82% of those infected with genotype 2 or 3.
	Pegylated IFN-α2b	SC	1.5 μg weekly × 48 weeks	
	Pegylated IFN-α2a	SC	180 μg weekly × 48 weeks	
	Pegylated IFN-α2b/ribavirin	SC (IFN)/oral (ribavirin)	1.5 μg/kg weekly (IFN)/800–1400 mg daily (ribavirin) × 24–48 weeks	
	Pegylated IFN-α2a/ribavirin	SC (IFN)/oral (ribavirin)	180 μg weekly (IFN)/800–1200 mg daily (ribavirin) × 24–48 weeks	
	IFN-alfacon	SC	9–15 μg	Doses of 9 and 15 μg are

Infection(s)	Drug	Route	Dosage	Comment
			thrice weekly × 6–12 months	equivalent to IFN-α2a and IFN-α2b doses of 3 million units and 5 million units, respectively.
	Sofosbuvir[g]	Oral	HCV genotypes 1, 4, 5, and 6: 400 mg qd with daily weight-based ribavirin (1000 mg [<75 kg] to 1200 mg [>75 kg]) and weekly pegylated IFN for 12 weeks Genotypes 2 and 3: 400 mg qd with daily weight-based ribavirin for 12 and 24 weeks, respectively	Sofosbuvir is generally well tolerated, and most common side effects have been attributable to concomitantly administered IFN and ribavirin. Sofosbuvir is recommended in triple combination with pegylated IFN and ribavirin for genotypes 1, 4, 5, and 6, with SVRs in 89–97% of treatment-naïve patients, and in double combination with ribavirin for genotypes 2 and 3.
	Simeprevir[g]	Oral	Alternative regimen for genotypes 1 and 4: 150 mg qd for 12 weeks plus daily ribavirin and weekly	Simeprevir has supplanted the first-generation protease inhibitors boceprevir and telaprevir. Its metabolism by cytochrome CYP3A can result in interactions with other drugs. Photosensitivity and reversible hyperbilirubinemia are associated toxicities.

Infection(s)	Drug	Route	Dosage	Comment
			pegylated IFN for 24 weeks and for 24–48 weeks, respectively	Testing for the Q80K-resistant variant should be carried out since this variant is present in one-third of HCV genotype 1a infections. Triple combinations with pegylated IFN and ribavirin result in SVRs in 80% of genotype 1 infections without Q80K.
	Sofosbuvir/ledipasvir	Oral	Active against genotypes 1, 4, 5	These DAA regimens have largely supplanted prior HCV treatments because of their ease of administration, excellent tolerability, and high efficacy. Various durations of regimens—from 8 to 24 weeks—have been studied, depending on regimen, cirrhosis, genotype, and prior HCV treatment. These highly active regimens are IFN and ribavirin free. Monitoring for HBV reactivation is warranted. Please access *http://www.hcvguidelines.org/* for the latest recommendations.
	Sofosbuvir/daclatasvir	Oral	Active against genotypes 1, 2, 3	
	Sofosbuvir/velpatasvir	Oral	Active against genotypes 1, 2, 3, 4, 5	
	Elbasvir/grazoprevir	Oral	Active against genotypes 1, 4	
	Paritaprevir/ritonavir/ombitasvir +/- dasabuvir	Oral	Active against genotypes 1, 4	
Chronic hepatitis D	IFN-α2a or IFN-α2b	SC	9 million units thrice weekly × 12 months	The overall efficacy and the optimal regimen and duration of therapy are not fully established. Sustained SVRs

Infection(s)	Drug	Route	Dosage	Comment
	Pegylated IFN-α2b	SC	1.5 μg weekly × 48 weeks	have been seen in 25–30% of patients for IFN-α and in 17–43% for pegylated IFN-α.
	Pegylated IFN-α2a	SC	180 μg weekly × 48 weeks	

[a]For detailed weight recommendations and for children <1 year of age, see *www.cdc.gov/flu/professionals/antivirals/summary-clinicians.htm*. [b]Amantadine and rimantadine are not recommended for routine use because of widespread resistance in currently circulating A/H3N2 and pandemic A/H1N1 viruses. Their use may be reconsidered if sensitivities become reestablished. [c]Not approved for this indication by the U.S. Food and Drug Administration (FDA). [d]Approved by the FDA for treatment of HIV-infected individuals. [e]Acyclovir suspension (15 mg/kg PO, to a maximum of 200 mg/dose) given for 7 days has been reported to be effective in the treatment of primary herpetic gingivostomatitis in children. [f]Active ingredient: benzyl alcohol. Available without prescription. [g]Consult *www.hcvguidelines.org* for recommendations regarding treatment of null or partial responders to IFN regimens or of patients ineligible to receive IFN.

Abbreviations: ALT, alanine aminotransferase; CMV, cytomegalovirus; CPK, creatine phosphokinase; DAA, direct-acting antiviral agent; HBeAg, hepatitis B e antigen; HBV, hepatitis B virus; HCV, hepatitis C virus; HSV, herpes simplex virus; IFN, interferon; RSV, respiratory syncytial virus; SVR, sustained virologic response; VZV, varicella-zoster virus.

ANTIVIRAL DRUGS ACTIVE AGAINST RESPIRATORY INFECTIONS

(See Also Chaps. 194 and 195)

ZANAMIVIR, OSELTAMIVIR, PERAMIVIR, AND LANINAMIVIR

Zanamivir and oseltamivir are inhibitors of the influenza virus neuraminidase enzyme, which is essential for release of virus from infected cells and for its subsequent spread throughout the respiratory tract of the infected host. The enzyme cleaves terminal sialic acid residues and thus destroys the cellular receptors to which the viral hemagglutinin attaches. Zanamivir and oseltamivir are sialic acid transition-state analogues and are highly active and specific inhibitors of the neuraminidases of both influenza A and B viruses. The antineuraminidase activity of the two drugs is similar, although zanamivir has somewhat greater in vitro activity against influenza B virus. Zanamivir may also be active against certain strains of influenza A virus that are resistant to oseltamivir. Both zanamivir and oseltamivir act through competitive and reversible inhibition of the active site of influenza A and B viral neuraminidases and have relatively little effect on mammalian cell enzymes.

Oseltamivir phosphate is an ethyl ester prodrug that is converted to oseltamivir carboxylate by esterases in the liver. Orally administered oseltamivir has a bioavailability of >60% and a plasma half-life of 7–9 h. The drug is excreted unmetabolized, primarily by the kidneys. Zanamivir has low oral bioavailability and is administered orally via a hand-held inhaler. By this route, ~15% of the dose is deposited in the lower respiratory tract, and low plasma levels of the drug are detected. The toxicities most frequently encountered with orally administered oseltamivir are nausea, gastrointestinal discomfort, and (less commonly) vomiting. Gastrointestinal discomfort is usually transient and is less

likely if the drug is administered with food. Neuropsychiatric events (delirium, self-injury) have been reported in children who have been taking oseltamivir, primarily in Japan. Zanamivir is orally inhaled and is generally well tolerated, although exacerbations of asthma may occur. An IV formulation of zanamivir is under development and is available from GlaxoSmithKline as part of clinical trials.

Inhaled zanamivir and orally administered oseltamivir have been effective in the treatment of naturally occurring, uncomplicated influenza A or B in otherwise healthy adults. In placebo-controlled studies, illness has been shortened by 1.0–1.5 days of therapy with either of these drugs when treatment is administered within 2 days of onset of symptoms. Pooled analyses of clinical studies of oseltamivir suggest that treatment may reduce the likelihood of hospitalizations and of certain respiratory tract complications associated with influenza, and observational studies suggest that oseltamivir may reduce mortality rates associated with influenza A outbreaks **(Chap. 195)**. Once-daily inhaled zanamivir or once-daily orally administered oseltamivir can provide prophylaxis against laboratory-documented influenza A– and influenza B–associated illness.

Resistance to the neuraminidase inhibitors may develop by changes in the viral neuraminidase enzyme, by changes in the hemagglutinin that make it more resistant to the actions of the neuraminidase, or by both mechanisms. Isolates that are resistant to oseltamivir—most commonly through the H275Y mutation, which leads to a change from histidine to tyrosine at that residue in the neuraminidase—remain sensitive to zanamivir. Certain mutations impart resistance to both oseltamivir and zanamivir (e.g., I223R, which leads to a change from isoleucine to arginine). Because the mechanisms of action of the neuraminidase inhibitors differ from those of the adamantanes (see below), zanamivir and oseltamivir are active against strains of influenza A virus that are resistant to amantadine and rimantadine.

Appropriate use of antiviral agents against influenza viruses depends on a knowledge of the resistance patterns of circulating viruses. As of this writing, currently circulating influenza A/H1N1 and H3N2 viruses (2013–2014) were sensitive to zanamivir and oseltamivir, with a few exceptions for oseltamivir. Up-to-date information on patterns of resistance to antiviral drugs is available from the Centers for Disease Control and Prevention (CDC) at *www.cdc.gov/flu*.

Zanamivir and oseltamivir have been approved by the U.S. Food and Drug Administration (FDA) for treatment of influenza in adults and in children (those ≥7 years old for zanamivir and those ≥1 year old for oseltamivir) who have been symptomatic for ≤2 days. Oseltamivir is approved for prophylaxis of influenza in individuals ≥1 year of age and zanamivir for those ≥5 years of age (Table 186-1). Guidelines for the use of oseltamivir in children <1 year of age can be accessed through the CDC website (*www.cdc.gov/flu/professionals/antivirals/summary-clinicians.htm*).

Peramivir (BCX-1812) is a neuraminidase inhibitor that can be administered intravenously. It has been approved in Japan, China, and South Korea but not in the United States, where it has been available in clinical trials through BioCryst Pharmaceuticals and previously as part of an emergency use authorization in response to the influenza A(H1N1)pdm09 virus pandemic in 2009–2010. Oseltamivir-resistant viruses generally exhibit reduced sensitivity to peramivir.

Laninamivir octanoate is a neuraminidase that has been approved in Japan for the treatment and prevention of influenza A and B. It is the prodrug of laninamivir, which is administered by oral inhalation and has a prolonged half-life of ~3 days. In limited studies, it has been investigated as single-dose therapy for influenza; its effects were similar to those obtained with multiple doses of zanamivir or oseltamivir.

AMANTADINE AND RIMANTADINE

Amantadine and the closely related compound rimantadine are primary symmetric amines that have antiviral activity limited to influenza A viruses. Amantadine and rimantadine have a long history of efficacy in the prophylaxis and treatment of influenza A infections in humans. However, high frequencies of resistance to these drugs were noted among influenza A/H3N2 viruses in the 2005–2006 influenza season and continued to be seen in 2013–2014. The pandemic A/H1N1 viruses that circulated in 2009–2010 were also resistant to amantadine and rimantadine, and circulating influenza A/H1N1 viruses in the 2013–2014 season were largely resistant. Therefore, these agents are no longer recommended unless the sensitivity of the particular isolate of influenza A virus is known, in which case their use may be considered. Amantadine and rimantadine act through inhibition of the ion channel function of the influenza A M2 matrix protein, on which uncoating of the virus depends. A substitution of a single amino acid at critical sites in the M2 protein can result in a virus that is resistant to amantadine and rimantadine.

Amantadine and rimantadine have been shown to be effective in the prophylaxis of influenza A in large-scale studies of young adults and in less extensive studies of children and elderly persons. In such studies, efficacy rates of 55–80% in the prevention of influenza-like illness were noted, and even higher rates were reported when virus-specific attack rates were calculated. Amantadine and rimantadine have also been found to be effective in the treatment of influenza A infection in studies involving predominantly young adults and, to a lesser extent, children. Administration of these compounds within 24–72 h after the onset of illness has resulted in a reduction of the duration of signs and symptoms by ~50% compared with that in placebo recipients. The effect on signs and symptoms of illness is superior to that of commonly used antipyretic–analgesic agents. Only anecdotal reports are available concerning the efficacy of amantadine or rimantadine in the prevention or treatment of complications of influenza (e.g., pneumonia).

Amantadine and rimantadine are available only in oral formulations and are ordinarily administered to adults once or twice daily, with a dosage of 100–200 mg/d. Despite their structural similarities, the two compounds have different pharmacokinetics. Amantadine is not metabolized and is excreted almost entirely by the kidneys, with a half-life of 12–17 h and peak plasma concentrations of 0.4 μg/mL. In contrast, rimantadine is extensively metabolized to hydroxylated derivatives and has a half-life of 30 h. Only 30–40% of an orally administered dose of rimantadine is recovered in the urine. The peak plasma levels of rimantadine are approximately half those of amantadine, but rimantadine is concentrated in respiratory secretions to a greater extent than amantadine. For prophylaxis, the compounds must be administered daily for the period at risk (i.e., duration of the exposure). For therapy, amantadine or rimantadine is generally administered for 5–7 days.

Although these compounds are generally well tolerated, 5–10% of amantadine recipients experience mild central nervous system side effects consisting primarily of dizziness, anxiety, insomnia, and difficulty in concentrating. These effects are rapidly reversible upon cessation of the drug's administration. At a dose of 200 mg/d, rimantadine is better tolerated than amantadine; in a large-scale study of young adults, adverse effects were no more frequent among rimantadine recipients than among placebo recipients. Seizures and worsening of congestive heart failure have also been reported in patients treated with amantadine, although a causal relationship has not been established. The dosage of amantadine should be reduced to 100 mg/d in patients with renal insufficiency—i.e., a creatinine clearance rate (Cr_{Cl}) of <50 mL/min—and in the elderly. A rimantadine dose of 100 mg/d should be used for patients with a Cr_{Cl} of <10 mL/min and for the elderly.

RIBAVIRIN

Ribavirin is a synthetic nucleoside analogue that inhibits a wide range of RNA and DNA viruses. The mechanism of action of ribavirin is not completely defined and may be different for different groups of viruses. Ribavirin-5′-

monophosphate blocks the conversion of inosine-5′-monophosphate to xanthosine-5′-monophosphate and interferes with the synthesis of guanine nucleotides as well as with that of both RNA and DNA. Ribavirin-5′-monophosphate also inhibits capping of virus-specific messenger RNA in certain viral systems.

Ribavirin administered as a small-particle aerosol to young children hospitalized with respiratory syncytial virus (RSV) infection has been clinically beneficial and has improved oxygenation in some studies (7 of 11). Although ribavirin has been approved for treatment of infants hospitalized with RSV infection, the American Academy of Pediatrics has recommended that it be considered on an individual basis rather than used routinely in that setting. Aerosolized ribavirin has also been administered to older children and adults (including immunosuppressed patients) with severe RSV and parainfluenza virus infections and to older children and adults with influenza A or B infection, but the benefit of this treatment, if any, is unclear. In RSV infections in immunosuppressed patients, ribavirin has been given in combination with anti-RSV immunoglobulins.

Orally administered ribavirin has not been effective in the treatment of influenza A virus infections. IV or oral ribavirin has reduced mortality rates among patients with Lassa fever; it is thought to be more effective in this regard when given within the first 6 days of illness. IV ribavirin has been reported to be of clinical benefit in the treatment of hemorrhagic fever with renal syndrome caused by Hantaan virus and as therapy for Argentinean hemorrhagic fever. Oral ribavirin has also been recommended for the treatment and prophylaxis of Congo-Crimean hemorrhagic fever. Use of IV ribavirin in patients with hantavirus pulmonary syndrome in the United States has not been associated with clear-cut benefits.

Oral administration of ribavirin reduces serum aminotransferase levels in patients with chronic hepatitis C virus (HCV) infection; because it appears not to reduce serum HCV RNA levels, the mechanism of this effect is unclear. The drug provides added benefit when given by mouth in doses of 800–1200 mg/d in combination with interferon (IFN) α2b or α2a (see below), and the triple combination of ribavirin, IFN, and sofosbuvir or simeprevir has been approved for the treatment of patients with chronic HCV infection (see below). Recent data suggest that oral ribavirin may be beneficial in resolution of chronic hepatitis E infection (largely genotype 3) associated with organ transplantation. Larger oral doses of ribavirin (800–1000 mg/d) have been associated with reversible hematopoietic toxicity. This effect has not been observed with aerosolized ribavirin, apparently because little drug is absorbed systemically. Aerosolized administration of ribavirin is generally well tolerated but occasionally is associated with bronchospasm, rash, or conjunctival irritation. It should be administered under close supervision—particularly in the setting of mechanical ventilation, where precipitation of the drug is possible. Health care workers exposed to the drug have experienced minor toxicity, including eye and respiratory tract irritation. Because ribavirin is mutagenic, teratogenic, and embryotoxic, its use is generally contraindicated in pregnancy. Its administration as an aerosol poses a risk to pregnant health care workers. Because clearance of ribavirin is primarily renal, dose reduction is required in the setting of significant renal dysfunction.

AGENTS OF INVESTIGATIVE INTEREST

Favipiravir (T-705) is a viral RNA–dependent RNA polymerase inhibitor active against influenza viruses, including neuraminidase inhibitor–resistant strains. It is approved in Japan for the treatment of influenza. Given favipiravir's in vitro activity against a broad range of viruses, including arenaviruses, phleboviruses (e.g., Rift Valley Fever virus), hantaviruses, flaviviruses, and filoviruses (e.g., Ebola virus), its use has been considered in the context of outbreaks of disease caused by these viruses even though their clinical activity remains uncertain. DAS181 is an investigational antiviral agent with activity against influenza A and B and parainfluenza viruses. A sialidase fusion protein, DAS181

cleaves the terminal sialic acid residues on human respiratory cells, reducing the binding of influenza and parainfluenza viruses. It is interesting to note that this agent targets host cellular rather than microbial protein. DAS181 is administered by oral inhalation and is being evaluated in the treatment of parainfluenza type 3 infections in recipients of lung and stem cell transplants. Three investigational agents with activity against RSV are being studied: (1) GS-5806, a small molecule that blocks fusion of the viral envelope with the host cell membrane, thus inhibiting viral entry, and has shown promising activity in human challenge studies; (2) ALS-008176, a prodrug of a cytidine nucleoside analogue that inhibits RSV replication by means of chain termination; and (3) ALN-RSV01, which works via RNA interference and is directed against the conserved region encoding the nucleocapsid (N) protein.

ANTIVIRAL DRUGS ACTIVE AGAINST HERPESVIRUS INFECTIONS

ACYCLOVIR AND VALACYCLOVIR

Acyclovir is a highly potent and selective inhibitor of the replication of certain herpesviruses, including herpes simplex virus (HSV) types 1 and 2, varicella-zoster virus (VZV), and Epstein-Barr virus (EBV). This drug is relatively ineffective in the treatment of human CMV infections; however, some studies have indicated effectiveness (at higher doses) in the prevention of CMV-associated disease in immunosuppressed patients. Valacyclovir, the L-valyl ester of acyclovir, is converted almost entirely to acyclovir by intestinal and hepatic hydrolysis after oral administration. Valacyclovir offers pharmacokinetic advantages over orally administered acyclovir: it exhibits significantly greater oral bioavailability, results in higher blood levels, and can be given less frequently than acyclovir (two or three rather than five times daily).

The high degree of selectivity of acyclovir is related to its mechanism of action, which requires that the compound first be phosphorylated to acyclovir monophosphate. This phosphorylation occurs efficiently in herpesvirus-infected cells by means of a virus-coded thymidine kinase. In uninfected mammalian cells, little phosphorylation of acyclovir occurs, and the drug is therefore concentrated in herpesvirus-infected cells. Acyclovir monophosphate is subsequently converted by host cell kinases to a triphosphate that is a potent inhibitor of virus-induced DNA polymerase but has relatively little effect on host cell DNA polymerase. Acyclovir triphosphate can also be incorporated into viral DNA, with early chain termination.

Acyclovir is available in IV, oral, and topical forms, while valacyclovir is available in an oral formulation. IV acyclovir is effective in the treatment of mucocutaneous HSV infections in immunocompromised hosts, in whom it reduces time to healing, duration of pain, and virus shedding. When administered prophylactically during periods of intense immunosuppression (e.g., related to chemotherapy for leukemia or transplantation) and before the development of lesions, IV acyclovir reduces the frequency of HSV-associated disease. After prophylaxis is discontinued, HSV lesions recur. IV acyclovir is also effective in the treatment of HSV encephalitis.

Because VZV is generally less sensitive to acyclovir than is HSV, higher doses of acyclovir must be used to treat VZV infections. In immunocompromised patients with herpes zoster, IV acyclovir reduces the frequency of cutaneous dissemination and visceral complications and—in one comparative trial—was more effective than vidarabine. Acyclovir, administered at oral doses of 800 mg five times a day, had a modest beneficial effect on localized herpes zoster lesions in both immunocompromised and immunocompetent patients. Combination of acyclovir with a tapering regimen of prednisone appeared to be more effective than acyclovir alone in terms of quality-of-life outcomes in immunocompetent patients aged >50 years with herpes zoster. A comparative study of acyclovir (800 mg PO five times daily) and valacyclovir (1 g PO three times daily) in immunocompetent patients with herpes zoster

indicated that the latter drug may be more effective in eliciting the resolution of zoster-associated pain. Orally administered acyclovir (600 mg five times a day) reduced complications of herpes zoster ophthalmicus in a placebo-controlled trial.

In chickenpox, a modest overall clinical benefit is attained when oral acyclovir therapy is begun within 24 h of the onset of rash in otherwise healthy children (20 mg/kg, up to a maximum of 800 mg, four times a day) or adults (800 mg five times a day). IV acyclovir has also been reported to be effective in the treatment of immunocompromised children with chickenpox.

A common use of acyclovir is in the treatment of genital HSV infections. IV or oral acyclovir or oral valacyclovir has shortened the duration of symptoms, reduced virus shedding, and accelerated healing when used for the treatment of primary genital HSV infections. Oral acyclovir and valacyclovir have also had a modest effect in treatment of recurrent genital HSV infections. However, the failure of treatment of either primary or recurrent disease to reduce the frequency of subsequent recurrences has indicated that acyclovir is ineffective in eliminating latent infection. Documented chronic oral administration of acyclovir for up to 6 years or of valacyclovir for up to 1 year has reduced the frequency of recurrences markedly during therapy; once the drug is discontinued, lesions recur. In one study, suppressive therapy with valacyclovir (500 mg once daily for 8 months) reduced transmission of HSV-2 genital infections among discordant couples by 50%. A modest effect on herpes labialis (i.e., a reduction of disease duration by 1 day) was seen when valacyclovir was administered upon detection of the first symptom of a lesion at a dose of 2 g every 12 h for 1 day. In AIDS patients, chronic or intermittent administration of acyclovir has been associated with the development of HSV and VZV strains resistant to the action of the drug and with clinical failures. The most common mechanism of resistance is a deficiency of the virus-induced thymidine kinase. Patients with HSV or VZV infections resistant to acyclovir have frequently responded to foscarnet.

With the availability of the oral and IV forms, there are few indications for topical acyclovir, although treatment with this formulation has been modestly beneficial in primary genital HSV infections and in mucocutaneous HSV infections in immunocompromised hosts.

Overall, acyclovir is remarkably well tolerated and is generally free of toxicity. The most frequently encountered form of toxicity is renal dysfunction because of drug crystallization (which is pH dependent), particularly after rapid IV administration or with inadequate hydration. Central nervous system changes, including lethargy and tremors, are occasionally reported, primarily in immunosuppressed patients. However, whether these changes are related to acyclovir, to concurrent administration of other therapy, or to underlying infection remains unclear. Acyclovir is excreted primarily unmetabolized by the kidneys via both glomerular filtration and tubular secretion. Approximately 15% of a dose of acyclovir is metabolized to 9-Carboxymethoxymethylguanine or other minor metabolites. Reduction in dosage is indicated in patients with a Cr_{Cl} of <50 mL/min. The half-life of acyclovir is ~3 h in normal adults, and the peak plasma concentration after a 1-h infusion of a dose of 5 mg/kg is 9.8 µg/mL. Approximately 22% of an orally administered acyclovir dose is absorbed, and peak plasma concentrations of 0.3–0.9 µg/mL are attained after administration of a 200-mg dose. Acyclovir penetrates relatively well into the cerebrospinal fluid (CSF), with concentrations approaching half of those found in plasma.

Acyclovir causes chromosomal breakage at high doses, but its administration to pregnant women has not been associated with fetal abnormalities. Nonetheless, the potential risks and benefits of acyclovir should be carefully assessed before the drug is used in pregnancy.

Valacyclovir exhibits three to five times greater bioavailability than acyclovir. The concentration–time curve for valacyclovir, given as 1 g PO three times daily, is similar to that for acyclovir, given as 5 mg/kg IV every 8 h. The safety profiles of valacyclovir and acyclovir are similar, although thrombotic thrombocytopenic purpura/hemolytic-uremic syndrome has been reported in immunocompromised patients who have received high doses of valacyclovir (8 g/d). Valacyclovir is approved for the treatment of herpes zoster, of initial and recurrent episodes of genital HSV infection, and of herpes labialis in immunocompetent adults as well as for suppressive treatment of genital herpes. Although it has not been extensively studied in other clinical settings involving HSV or VZV infections, many consultants use valacyclovir rather than oral acyclovir in settings where only the latter has been approved because of valacyclovir's superior pharmacokinetics and more convenient dosing schedule.

CIDOFOVIR

Cidofovir is a phosphonate nucleotide analogue of cytosine. Its major use is in CMV infections, but it is active against a broad range of herpesviruses, including HSV, human herpesvirus (HHV) types 6A and 6B, HHV-8, and certain other DNA viruses such as polyomaviruses, papillomaviruses, adenoviruses, and poxviruses, including variola (smallpox) and vaccinia. Cidofovir does not require initial phosphorylation by virus-induced kinases; the drug is phosphorylated by host cell enzymes to cidofovir diphosphate, which is a competitive inhibitor of viral DNA polymerases and, to a lesser extent, of host cell DNA polymerases. Incorporation of cidofovir diphosphate slows or terminates nascent DNA chain elongation. Cidofovir is active against HSV isolates that are resistant to acyclovir because of absent or altered thymidine kinase and against CMV isolates that are resistant to ganciclovir because of UL97 phosphotransferase mutations. CMV isolates resistant to ganciclovir on the basis of UL54 mutations are usually resistant to cidofovir as well. Cidofovir is usually active against foscarnet-resistant CMV, although cross-resistance to foscarnet has been described, typically in the UL54 polymerase.

Cidofovir has poor oral availability and is administered intravenously. It is excreted primarily by the kidney and has a plasma half-life of 2.6 h. Cidofovir diphosphate's intracellular half-life of >48 h is the basis for the recommended dosing regimen of 5 mg/kg once a week for the initial 2 weeks and then 5 mg/kg every other week. The major toxic effect of cidofovir is proximal renal tubular injury, as manifested by elevated serum creatinine levels and proteinuria. The risk of nephrotoxicity can be reduced by vigorous saline hydration and by concomitant oral administration of probenecid. Neutropenia, rashes, and gastrointestinal tolerance may also occur.

IV cidofovir has been approved for the treatment of CMV retinitis in AIDS patients who are intolerant of ganciclovir or foscarnet or in whom those drugs have failed. In a controlled study, a maintenance dosage of 5 mg/kg per week administered to AIDS patients reduced the progression of CMV retinitis from that seen at 3 mg/kg. Intravitreal cidofovir has been used to treat CMV retinitis but has been associated with significant toxicity. IV cidofovir has been reported anecdotally to be effective for treatment of acyclovir-resistant mucocutaneous HSV infections. Likewise, topically administered cidofovir is reportedly beneficial against mucocutaneous HSV infections in HIV-infected patients. Anecdotal use of IV cidofovir has been described in disseminated adenoviral infections in immunosuppressed patients and in genitourinary infections with BK virus in renal transplant recipients; however, its efficacy, if any, in these circumstances is not established.

CMX-001 (brincidofovir) is an ester prodrug of cidofovir that can be administered orally and may be less nephrotoxic than IV cidofovir. It has not shown efficacy in the prevention of CMV infection in stem cell transplant recipients but is being studied for treatment of BK nephropathy and adenovirus infections.

FOMIVIRSEN

Fomivirsen is the first antisense oligonucleotide approved by the FDA for therapy in humans. This phosphorothioate oligonucleotide, 21 nucleotides in length, inhibits CMV replication through interaction with CMV messenger RNA. Fomivirsen is complementary to messenger transcripts of the major immediate early region 2 of CMV, which codes for proteins regulating viral gene expression. In addition to its antisense mechanism of action, fomivirsen may exert activity against CMV through inhibition of viral adsorption to cells as well as direct inhibition of viral replication. Because of its different mechanism of action, fomivirsen is active against CMV isolates that are resistant to nucleoside or nucleotide analogues, such as ganciclovir, foscarnet, or cidofovir.

Fomivirsen has been approved for intravitreal administration in the treatment of CMV retinitis in AIDS patients who have failed to respond to other treatments or cannot tolerate them. Injection of two doses of 330 mg 2 weeks apart, followed by maintenance doses of 330 mg monthly, significantly reduces the rate of progression of CMV retinitis. The major toxicity is ocular inflammation, including vitritis and iritis, which usually responds to topically administered glucocorticoids.

GANCICLOVIR AND VALGANCICLOVIR

An analogue of acyclovir, ganciclovir is active against HSV and VZV and is markedly more active than acyclovir against CMV. Ganciclovir triphosphate inhibits CMV DNA polymerase and can be incorporated into CMV DNA, whose elongation it eventually terminates. In HSV- and VZV-infected cells, ganciclovir is phosphorylated by virus-encoded thymidine kinases; in CMV-infected cells, it is phosphorylated by a viral kinase encoded by the UL97 gene. Ganciclovir triphosphate is present in tenfold higher concentrations in CMV-infected cells than in uninfected cells. Ganciclovir is approved for the treatment of CMV retinitis in immunosuppressed patients and for the prevention of CMV disease in transplant recipients. It is widely used for the treatment of other CMV-associated syndromes, including pneumonia, esophagogastrointestinal infections, hepatitis, and “wasting” illness.

Ganciclovir is available for IV or oral administration. Because its oral bioavailability is low (5–9%), relatively large doses (1 g three times daily) must be administered by this route. Oral ganciclovir has largely been supplanted by valganciclovir, which is the L-valyl ester of ganciclovir. Valganciclovir is well absorbed orally, with a bioavailability of 60%, and is rapidly hydrolyzed to ganciclovir in the intestine and liver. The area under the curve for a 900-mg dose of valganciclovir is equivalent to that for 5 mg/kg of IV ganciclovir, although peak serum concentrations are ~40% lower for valganciclovir. The serum half-life is 3.5 h after IV administration of ganciclovir and 4.0 h after PO administration of valganciclovir. Ganciclovir is excreted primarily by the kidneys in an unmetabolized form, and its dosage should be reduced in cases of renal failure. Ganciclovir therapy at the most commonly used initial IV dosage—i.e., 5 mg/kg every 12 h for 14–21 days—can be changed to valganciclovir (900 mg PO twice daily) when the patient can tolerate oral therapy. The maintenance dose is 5 mg/kg IV daily or five times per week for ganciclovir and 900 mg by mouth once a day for valganciclovir. Dose adjustment in patients with renal dysfunction is required. Intraocular ganciclovir, given by either intravitreal injection or intraocular implantation, has also been used to treat CMV retinitis.

Ganciclovir is effective as prophylaxis against CMV-associated disease in organ and bone marrow transplant recipients. Oral ganciclovir administered prophylactically to AIDS patients with CD4+ T cell counts of <100/μL has provided protection against the development of CMV retinitis. However, the long-term benefits of this approach to prophylaxis in AIDS patients have not been established, and most experts do not recommend the use of oral

ganciclovir for this purpose. As already mentioned, valganciclovir has supplanted oral ganciclovir in settings where oral prophylaxis or therapy is considered.

The administration of ganciclovir has been associated with bone marrow suppression, particularly neutropenia, which significantly limits the drug's use in many patients. Bone marrow toxicity is potentiated in the setting of renal dysfunction and when other bone marrow suppressants, such as zidovudine or mycophenolate mofetil, are used concomitantly. This toxicity is typically dose and duration sensitive and is reversible with cessation of ganciclovir use.

Resistance has been noted in CMV isolates obtained after therapy with ganciclovir, especially those from patients with AIDS or from patients receiving prolonged ganciclovir therapy after organ transplantation. Such resistance may develop through a mutation in either the viral UL97 gene or the viral DNA polymerase. Ganciclovir-resistant isolates are usually sensitive to foscarnet (see below) or may be sensitive to cidofovir, depending on the mechanism of resistance (see above).

FAMCICLOVIR AND PENCICLOVIR

Famciclovir is the diacetyl 6-deoxyester of the guanosine analogue penciclovir. This agent is well absorbed orally, has a bioavailability of 77%, and is rapidly converted to penciclovir by deacetylation and oxidation in the intestine and liver. Penciclovir's spectrum of activity and mechanism of action are similar to those of acyclovir. Thus, penciclovir usually is not active against acyclovir-resistant viruses. However, some acyclovir-resistant viruses with altered thymidine kinase or DNA polymerase substrate specificity may be sensitive to penciclovir. This drug is phosphorylated initially by a virus-encoded thymidine kinase and subsequently by cellular kinases to penciclovir triphosphate, which inhibits HSV-1, HSV-2, VZV, and EBV as well as hepatitis B virus (HBV). The serum half-life of penciclovir is 2 h, but the intracellular half-life of penciclovir triphosphate is 7–20 h—markedly longer than that of acyclovir triphosphate. The latter is the basis for the less frequent (twice-daily) dosing schedule for famciclovir than for acyclovir. Penciclovir is eliminated primarily in the urine by both glomerular filtration and tubular secretion. The usually recommended dosage interval should be adjusted for renal insufficiency.

Clinical trials involving immunocompetent adults with herpes zoster showed that famciclovir was superior to placebo in eliciting the resolution of skin lesions and virus shedding and in shortening the duration of postherpetic neuralgia; moreover, administered at 500 mg every 8 h, famciclovir was at least as effective as acyclovir administered at an oral dose of 800 mg five times daily. Famciclovir was also effective in the treatment of herpes zoster in immunosuppressed patients. Clinical trials have demonstrated its effectiveness in the suppression of genital HSV infections for up to 1 year and in the treatment of initial and recurrent episodes of genital herpes. Famciclovir is effective as therapy for mucocutaneous HSV infections in HIV-infected patients. Application of a 1% penciclovir cream reduces the duration of signs and symptoms of herpes labialis in immunocompetent patients (by 0.5–1 day) and has been approved for that purpose by the FDA. Famciclovir is generally well tolerated, with occasional headache, nausea, and diarrhea reported in frequencies similar to those among placebo recipients. The administration of high doses of famciclovir for 2 years was associated with an increased incidence of mammary adenocarcinomas in female rats, but the clinical significance of this effect is unknown.

FOSCARNET

Foscarnet (phosphonoformic acid) is a pyrophosphate-containing compound that potently inhibits herpesviruses, including CMV. This drug inhibits DNA polymerases at the pyrophosphate binding site at concentrations that have

relatively little effect on cellular polymerases. Foscarnet does not require phosphorylation to exert its antiviral activity and is therefore active against HSV and VZV isolates that are resistant to acyclovir because of deficiencies in thymidine kinase as well as against most ganciclovir-resistant strains of CMV. Foscarnet also inhibits the reverse transcriptase of HIV and is active against HIV in vivo.

Foscarnet is poorly soluble and must be administered intravenously via an infusion pump in a dilute solution over 1–2 h. The plasma half-life of foscarnet is 3–5 h and increases with decreasing renal function because the drug is eliminated primarily by the kidneys. It has been estimated that 10–28% of a dose may be deposited in bone, where it can persist for months. The most common initial dosage of foscarnet is 60 mg/kg every 8 h for CMV and 40 mg/kg every 8 h for HSV. Once the infection is controlled, a maintenance dose of 90–120 mg/kg once a day has been used by some.

Foscarnet is approved for the treatment of CMV retinitis in patients with AIDS and of acyclovir-resistant mucocutaneous HSV infections. In a comparative clinical trial, the drug appeared to be about as efficacious as ganciclovir against CMV retinitis but was associated with a longer survival period, possibly because of its activity against HIV. Intraocular foscarnet has been used to treat CMV retinitis. In addition, foscarnet has been employed to treat acyclovir-resistant HSV and VZV infections as well as ganciclovir-resistant CMV infections, although resistance to foscarnet has been reported in CMV isolates obtained during therapy. Foscarnet has also been used to treat HHV-6B infections in immunosuppressed patients.

The major form of toxicity associated with foscarnet is renal impairment. Thus renal function should be monitored closely, particularly during the initial phase of therapy. Because foscarnet binds divalent metal ions, hypocalcemia, hypomagnesemia, hypokalemia, and hypo- or hyperphosphatemia can develop. Saline hydration and slow infusion appear to protect the patient against nephrotoxicity and electrolyte disturbances. Although hematologic abnormalities have been documented (most commonly anemia), foscarnet is not generally myelosuppressive and can be administered concomitantly with myelosuppressive medications.

TRIFLURIDINE

Trifluridine is a pyrimidine nucleoside active against HSV-1, HSV-2, and CMV. Trifluridine monophosphate irreversibly inhibits thymidylate synthetase, and trifluridine triphosphate inhibits viral and, to a lesser extent, cellular DNA polymerases. Because of systemic toxicity, trifluridine's use is limited to topical therapy. Trifluridine is approved for treatment of HSV keratitis, against which trials have shown that it is more effective than topical idoxuridine but similar in efficacy to topical vidarabine. The drug has benefited some patients with HSV keratitis who have failed to respond to idoxuridine or vidarabine. Topical application of trifluridine to sites of acyclovir-resistant HSV mucocutaneous infection has also been beneficial in some cases.

VIDARABINE

Vidarabine is a purine nucleoside analogue with activity against HSV-1, HSV-2, VZV, and EBV. Vidarabine inhibits viral DNA synthesis through its 5′-triphosphorylated metabolite, although its precise molecular mechanisms of action are not completely understood. IV-administered vidarabine has been shown to be effective in the treatment of herpes simplex encephalitis, mucocutaneous HSV infections, herpes zoster in immunocompromised patients, and neonatal HSV infections. Its use has been supplanted by that of IV acyclovir, which is more effective and easier to administer.

Production of the IV preparation has been discontinued by the manufacturer, but vidarabine is available as an ophthalmic ointment, which is effective in the treatment of HSV keratitis.

AGENTS OF INVESTIGATIVE INTEREST

Maribavir is a benzimidazole that inhibits CMV and EBV. This drug inhibits the CMV UL97 kinase and does not require intracellular phosphorylation for its antiviral activity. Its mechanism of action involves blocking viral DNA synthesis and virion egress. Maribavir is orally administered and has been associated with taste disturbance and diarrhea. In phase 3 studies, it was not efficacious in the prevention of CMV infection in recipients of hematopoietic stem cell and adult liver transplants. However, when used at somewhat higher doses, it may be efficacious for the treatment of refractory or resistant CMV infections in transplant recipients.

Letermovir is an investigational drug with activity against CMV. It is a dihydroquinazoline that acts through inhibition of the viral terminase enzyme complex. This mechanism of action differs from that of ganciclovir, foscarnet, and cidofovir, which inhibit viral DNA polymerase; therefore, letermovir is active against CMV isolates that are resistant to those drugs. It is orally administered and is reportedly well tolerated. Letermovir demonstrated significant activity in preventing CMV reactivation in a recent phase 3 trial in adults undergoing hematopoietic stem cell transplantation and may be clinically available soon.

Inhibition of the helicase–primase heterotrimeric complex of HSV-1 and HSV-2 represents a novel mechanism of action of amenamevir and pritelivir. These drugs are being assessed for prevention and treatment of HSV genital infection. The efficacy of amenamevir, administered as a single oral dose of 1200 mg for recurrent genital herpes, was comparable to that of valacyclovir given for 3 days. Pritelivir has a longer half-life (up to 80 h) and was studied in a placebo-controlled trial of suppression of genital HSV infections. Compared with placebo, pritelivir—a loading dose followed by either a daily oral dose of 75 mg for 4 weeks or a weekly dose of 400 mg for 4 weeks—reduced HSV shedding and days of genital lesions. Additional clinical studies of the helicase–primase inhibitors of HSV are planned.

ANTIVIRAL DRUGS ACTIVE AGAINST HEPATITIS VIRUSES

LAMIVUDINE

Lamivudine is a pyrimidine nucleoside analogue that is used primarily in combination therapy against HIV infection (Chap. 197). Its activity against HBV is attributable to inhibition of the viral DNA polymerase. This drug has also been approved for the treatment of chronic HBV infection. At doses of 100 mg/d given for 1 year to patients positive for hepatitis B e antigen (HBeAg), lamivudine is well tolerated and results in suppression of HBV DNA levels, normalization of serum aminotransferase levels in 40–75% of patients, and reduction of hepatic inflammation and fibrosis in 50–60% of patients. Loss of HBeAg occurs in 30% of patients. Lamivudine also appears to be useful in the prevention or suppression of HBV infection associated with liver transplantation. Resistance to lamivudine develops in 24% of patients treated for 1 year and is associated with changes in the YMDD motif of HBV DNA polymerase. Because of the frequency of development of resistance, lamivudine has been largely supplanted by less-resistance-prone drugs for the treatment of HBV infection.

ADEFOVIR DIPIVOXIL

Adefovir dipivoxil is the oral prodrug of adefovir, an acyclic nucleotide analogue of adenosine monophosphate that is active against HBV, HIV, HSV, CMV, and poxviruses. It is phosphorylated by cellular kinases to the active triphosphate moiety, which is a competitive inhibitor of HBV DNA polymerase and results in chain termination after incorporation into nascent viral DNA. Adefovir is administered orally and is eliminated primarily by the kidneys, with a plasma half-life of 5–7.5 h. In clinical studies, therapy with adefovir at a dose of 10 mg/d for 48 weeks resulted in normalization of serum alanine aminotransferase (ALT) levels in 48–72% of patients and improved liver histology in 53–64%; it also resulted in a 3.5- to 3.9-$\log_{10}$ reduction in the number of HBV DNA copies per milliliter of plasma. Adefovir was effective in treatment-naïve patients as well as in those infected with lamivudine-resistant HBV. Resistance to adefovir appears to develop less readily than that to lamivudine, but adefovir resistance rates of 15–18% have been reported after 192 weeks of treatment and may reach 30% after 5 years. This agent is generally well tolerated. Significant nephrotoxicity attributable to adefovir is uncommon at the dose used in the treatment of HBV infections (10 mg/d) but is a treatment-limiting adverse effect at the higher doses used in therapy for HIV infections (30–120 mg/d). In any case, renal function should be monitored in patients taking adefovir, even at the lower dose. Adefovir is approved only for treatment of chronic HBV infection.

TENOFOVIR

Tenofovir disoproxil fumarate (TDF) is a prodrug of tenofovir, a nucleotide analogue of adenosine monophosphate with activity against both retroviruses and hepadnaviruses. In both immunocompetent and immunocompromised patients (including those co-infected with HIV and HBV), tenofovir given at a dose of 300 mg/d for 48 weeks reduced HBV replication by 4.6–6 $\log_{10}$, normalized ALT levels in 68–76% of patients, and improved liver histopathology in 72–74% of patients. Resistance develops uncommonly during ≥2 years of therapy, and tenofovir is active against lamivudine-resistant HBV. The safety profile of tenofovir is similar to that of adefovir, but nephrotoxicity has not been encountered at the dose used for HBV therapy. Tenofovir is approved for the treatment of HIV and chronic HBV infections. Tenofovir alafenamide (TAF) has recently been approved for use in the treatment of HIV and chronic HBV infections. TAF is dosed at 25 mg orally per day and has better renal and bone safety than TDF. Lactic acidosis is an important side effect associated with tenofovir use. **For a more detailed discussion of tenofovir, see Chap. 197**.

ENTECAVIR

Entecavir is a cyclopentyl 2′-deoxyguanosine analogue that inhibits HBV through interaction of entecavir triphosphate with several HBV DNA polymerase functions. At a dose of 0.5 mg/d given for 48 weeks, entecavir reduced HBV DNA copies by 5.0–6.9 $\log_{10}$, normalized serum aminotransferase levels in 68–78% of patients, and improved histopathology in 70–72% of patients. Entecavir inhibits lamivudine-resistant viruses that have M550I or M550V/L526M mutations but only at serum concentrations twenty- or thirtyfold higher than those obtained with the 0.5-mg/d dose. Thus, higher doses of entecavir (1 mg/d) are recommended for the treatment of patients infected with lamivudine-resistant HBV. Development of resistance to entecavir is uncommon in treatment-naïve patients but does occur at unacceptably high rates (43% after 4 years) in patients previously infected with lamivudine-resistant virus. Entecavir-resistant strains appear to be sensitive to adefovir and tenofovir.

Entecavir is highly bioavailable but should be taken on an empty stomach because food interferes with its absorption. The drug is eliminated primarily in unchanged form by the kidneys, and its dosage should be adjusted for patients with Cr_{Cl} values of <50 mL/min. Overall, entecavir is well tolerated, with a safety profile similar to that of lamivudine. As with other anti-HBV treatments, exacerbation of hepatitis may occur when entecavir therapy is stopped. Entecavir

is approved for treatment of chronic hepatitis B, including infection with lamivudine-resistant viruses, in adults. Entecavir has some activity against HIV-1 (median effective concentration, 0.026 to >10 μM) but should not be used as monotherapy in HIV-positive patients because of the potential for development of HIV resistance due to the M184V mutation.

TELBIVUDINE

Telbivudine is a β-L enantiomer of thymidine and is a potent, selective inhibitor of HBV. Its active form is telbivudine triphosphate, which inhibits HBV DNA polymerase and causes chain termination but has little or no activity against human DNA polymerase. Administration of telbivudine at an oral dose of 600 mg/d for 52 weeks to patients with chronic hepatitis B resulted in reduction of HBV DNA by 5.2–6.4 $\log_{10}$ copies/mL along with normalization of ALT levels in 74–77% of recipients and improved histopathology in 65–67% of patients. Telbivudine-resistant HBV is generally cross-resistant with lamivudine-resistant virus but is usually susceptible to adefovir. After 2 years of therapy, resistance to telbivudine was noted in isolates from 22% of HBeAg-positive patients and in those from 9% of HBeAg-negative patients.

Orally administered telbivudine is rapidly absorbed; because it is eliminated primarily by the kidneys, its dosage should be reduced in patients with a Cr_{Cl} value of <50 mL/min. Telbivudine is generally well tolerated, but increases in serum levels of creatinine kinases as well as fatigue and myalgias have been observed. As with other anti-HBV drugs, hepatitis may be exacerbated in patients who discontinue telbivudine therapy. Telbivudine has been approved for the treatment of adults with chronic hepatitis B who have evidence of viral replication and either persistently elevated serum aminotransferase levels or histopathologically active disease, but it has not been widely used because of the frequency of development of resistance, as noted above.

INTERFERONS

IFNs are cytokines that exhibit a broad spectrum of antiviral activities as well as immunomodulating and antiproliferative properties. IFNs are not available for oral administration but must be given IM, SC, or IV. Early studies with human leukocyte IFN demonstrated an effect in the prophylaxis of experimentally induced rhinovirus infections in humans and in the treatment of VZV infections in immunosuppressed patients. DNA recombinant technology has made available highly purified α, β, γ, and λ IFNs that have been evaluated in a variety of viral infections. Results of such trials have confirmed the effectiveness of intranasally administered IFN in the prophylaxis of rhinovirus infections, although its use has been associated with nasal mucosal irritation. Studies have also demonstrated a beneficial effect of intralesionally or systemically administered IFNs on genital warts. The effect of systemic administration consists primarily of a reduction in the size of the warts, and this mode of therapy may be useful in persons who have numerous warts that cannot easily be treated by individual intralesional injections. However, lesions frequently recur after either intralesional or systemic IFN therapy is discontinued.

IFNs have undergone extensive study in the treatment of chronic HBV infection. The administration of standard IFN-α2b (5 million units daily or 10 million units three times a week for 16–24 weeks) to patients with stable chronic HBV infection resulted in loss of markers of HBV replication, such as HBeAg and HBV DNA, in 33–37% of cases; 8% of patients also became negative for hepatitis B surface antigen. In most patients who lose HBeAg and HBV DNA markers, serum aminotransferases return to normal levels, and both short- and long-term improvements in liver histopathology have been described. Predictors of a favorable response to standard IFN therapy include low pretherapy levels of HBV

DNA, high pretherapy serum levels of ALT, a short duration of chronic HBV infection, and active inflammation in liver histopathology. Poor responses are seen in immunosuppressed patients, including those with HIV infection.

In pegylated IFNs, IFN alphas are linked to polyethylene glycol. This linkage results in slower absorption, decreased clearance, and more sustained serum concentrations, thereby permitting a more convenient, once-weekly dosing schedule; in many instances, pegylated IFN has supplanted standard IFN. After 48 weeks of treatment with 180 µg of pegylated IFN-α2a, HBV DNA was reduced by 4.1–4.5 $\log_{10}$ copies/mL, with normalization of serum ALT levels in 39% of patients and improved histology in 38%. Response rates were somewhat higher when lamivudine was administered with pegylated IFN-α2a. Adverse effects of IFN are common and include fever, chills, myalgia, fatigue, neurotoxicity (manifested primarily as somnolence, depression, anxiety, and confusion), and leukopenia. Autoantibodies (e.g., antithyroid antibodies) can also develop. IFN-α2b and pegylated IFN-α2a are approved for the treatment of patients with chronic hepatitis B. Data supporting the therapeutic efficacy of pegylated interferon-α2b in HBV infection have been published; the drug has not been approved for this indication in the United States but has been approved for treatment of chronic HBV infection in other countries.

Several IFN preparations, including IFN-α2a, IFN-α2b, IFN-alfacon-1, and IFN-αm1 (lymphoblastoid), have been studied as therapy for chronic HCV infections. A variety of monotherapy regimens have been studied, of which the most common for standard IFN is IFN-α2b or -α2a at 3 million units three times per week for 12–18 months. The addition of oral ribavirin to IFN-α2b—either as initial therapy or after failure of IFN therapy alone—results in significantly higher rates of sustained virologic and/or serum ALT responses (40–50%) than are obtained with monotherapy. Comparative studies indicate that pegylated IFN-α2b or -α2a therapy is more effective than standard IFN treatment against chronic HCV infection. The combination of SC pegylated IFN and oral ribavirin results in sustained virologic responses (SVRs) in 42–51% of patients with HCV genotype 1 infection and in 76–82% of patients with genotype 2 or 3 infection. Ribavirin appears to have a small antiviral effect in HCV infection but may also be working through an immunomodulatory effect in combination with IFN. Optimal results with ribavirin appear to be associated with weight-based dosing. Prognostic factors for a favorable response include an age of <40 years, a short duration of infection, low levels of HCV RNA, a lesser degree of liver histopathology, and infection with HCV genotypes other than 1. IFN-alfacon, a synthetic “consensus” α interferon, appears to produce response rates similar to those elicited by standard IFN-α2a or -α2b alone. In 2014, the approval of a polymerase inhibitor, sofosbuvir, and a second-generation protease inhibitor, simeprevir, as well as the successful development of other direct-acting antiviral agents (DAAs) active against HCV led to revised recommendations for treatment of hepatitis C with DAA regimens not requiring IFN or ribavirin in most cases. DAA regimens have been developed that are active against all HCV genotypes (see below and Table 186-1).

IFN-α and pegylated IFN-α are active against hepatitis D, but high doses are required (9 million units three times per week for 48 weeks). IFN-α elicited an SVR in 25–30% of patients, whereas pegylated IFN-α had a variable effect, evoking an SVR in 17–43% of patients. However, long-term biochemical and histologic improvements have been seen, even in the absence of sustained inhibition of viral replication.

POLYMERASE INHIBITORS

Sofosbuvir is the prodrug of a uridine nucleoside inhibitor of HCV RNA NS5B polymerase. Its metabolism to the active uridine nucleoside triphosphate results in chain termination. Sofosbuvir is active against all HCV genotypes (1–6) and has a median effective concentration (EC_{50}) of 0.7–2.6 µM against NS5B. Resistance to sofosbuvir is conferred by an

S282T substitution in NS5B, but clinically expressed resistance to treatment has only rarely been encountered in patients who receive sofosbuvir.

Sofosbuvir is administered orally and is unaffected by food. After oral administration, plasma concentrations of sofosbuvir and of its active metabolite peak in 0.5–2 h and 2–4 h, respectively. Approximately 61–65% of sofosbuvir is bound in plasma proteins, but very little of the active metabolite is bound. Both sofosbuvir and its active metabolite are cleared renally, with $t_{1/2}$ values of 0.4 and 27 h, respectively. Sofosbuvir is relatively free from clinically significant drug interactions, although P-glycoprotein inducers can reduce sofosbuvir concentrations.

Sofosbuvir is generally well tolerated and has not been associated with significant toxicity. The most common side effects in recipients of sofosbuvir have been attributable to concomitant administration of IFN and ribavirin in combination clinical trials (see below).

Sofosbuvir has been studied in a variety of controlled and open-label clinical trials. In late 2013, the results of these trials led to its recommendation—in triple combination with pegylated IFN and ribavirin—as first-line treatment for chronic hepatitis due to HCV genotypes 1, 4, 5, and 6, in which SVR rates among treatment-naïve patients were 89–97%. For HCV genotypes 2 and 3, IFN-free regimens consisting of sofosbuvir and ribavirin have been recommended, with SVR rates among treatment-naïve patients of 93% for genotype 2 and 61% for genotype 3.

PROTEASE INHIBITORS

BOCEPREVIR, TELAPREVIR

This drug class is specifically designed to inhibit the 3/4A (NS3/4A) HCV protease. These agents resemble the HCV polypeptide and, when processed by the viral protease, form a covalent bond with the catalytic NS3 serine residues, block further activity, and prevent proteolytic cleavage of the HCV polyprotein into NS4A, NS4B, NS5A, and NS5B proteins. Boceprevir and telaprevir are linear ketoamide compounds that are active against HCV genotype 1 (1b > 1a) and much less so against genotypes 2 and 3. These first-generation protease inhibitors received approval for combination therapy (with IFN and ribavirin) for genotype 1 infection. Neither boceprevir nor telaprevir is now recommended for the treatment of hepatitis C. These drugs have been supplanted by sofosbuvir and by simeprevir, a second-generation protease inhibitor with improved pharmacokinetic properties, fewer drug–drug interactions, and less overall toxicity (see below).

SIMEPREVIR

Simeprevir is a second-generation NS3/4A protease inhibitor with antiviral activity against HCV genotype 1 (1b > 1a); the EC_{50} is 9.4 nM in an HCV genotype 1b replicon. The NS3 polymorphism Q80K, which is present in approximately one-third of patients carrying HCV genotype 1b, increases the EC_{50} by elevenfold and results in clinical resistance to simeprevir. Thus testing for Q80K should be carried out if treatment with simeprevir is being considered. Cross-resistance occurs between simeprevir and the first-generation protease inhibitors boceprevir and telaprevir.

Simeprevir is orally administered as a 150-mg capsule, and its bioavailability is increased by administration with food. The serum concentration peaks 4–6 h after oral administration. The drug's elimination half-life is 10–13 h in healthy individuals and 41 h in patients with hepatitis C. Simeprevir is nearly entirely bound by plasma proteins and cleared by biliary excretion. Because there is no renal excretion, dose adjustments are not required in the presence of renal

dysfunction. Simeprevir is metabolized by hepatic CYP3A and therefore should not be administered to patients with decompensated liver function.

Because of its metabolism by cytochrome P450 3A (CYP3A), simeprevir interacts with drugs that induce or inhibit CYP3A, and these interactions may concomitantly increase or reduce plasma concentrations of simeprevir. Administration of simeprevir may also increase plasma concentrations of drugs that are substrates for hepatic organic anion-transporting polypeptide 1B1 or 1B3 or for P glycoprotein transporters.

Toxicity observed during clinical trials with simeprevir included photosensitivity (usually mild or moderate) in 28% of recipients and reversible hyperbilirubinemia (both conjugated and unconjugated), which was generally mild to moderate. Most of the other adverse effects seen in clinical trials with simeprevir were attributable to concomitant administration of IFN and ribavirin.

Simeprevir has been recommended as a component of alternative treatment—in combination with pegylated IFN and ribavirin—of chronic infection with HCV genotypes 1 and 4. Daily simeprevir, daily ribavirin, and weekly pegylated IFN for 12 weeks followed by another 12 weeks of pegylated IFN and ribavirin resulted in an SVR of 80% in the absence of the Q80K variant. In general, simeprevir-based triple therapy appeared to be 10% less likely to yield an SVR than sofosbuvir-based therapy and more likely to cause adverse effects. However, for prior nonresponders or partial responders to pegylated IFN, the IFN-free regimen of simeprevir, sofosbuvir, and ribavirin shows promise.

PARITAPREVIR/RITONAVIR AND GRAZOPREVIR

These drugs are more recently developed NS3/4A protease inhibitors. Paritaprevir is used with ritonavir and ombitasvir (an NS5A inhibitor; see below) as a fixed-dose combination and may be used with dasabuvir and ribavirin. This combination is active against HCV genotypes 1a and 1b. Ritonavir is used to increase the levels of paritaprevir. Ritonavir is a potent CYP3A inhibitor and may impact the metabolism of other medications handled by this pathway. Grazoprevir is used with elbasvir (an NS5A inhibitor; see below) in a fixed-dose combination and is approved by the FDA for treatment of HCV genotypes 1 and 4.

NS5A INHIBITORS

NS5A is a membrane-associated phosphoprotein that is part of the HCV RNA replication complex and is essential for viral replication and assembly. Ledipasvir, velpatasvir, daclatasvir, elbasvir, and ombitasvir are all NS5A inhibitors. Each of these agents has largely been developed and studied with specific partner drugs as noted above (Table 186-1).

Treatment of HCV has been associated with flaring of chronic HBV infection. Monitoring for HBV activation in this context is warranted. In the setting of significant renal dysfunction (Cr_{Cl}, <30 mL/min), few data are available to guide use of these newer DAAs. However, studies are ongoing to assess elbasvir/grazoprevir in this context, as these agents are eliminated through the feces and are not renally handled. Emergence of HCV resistance-associated substitutions to the DAAs have been documented. The impact on treatment is under active investigation and at this time is relevant mostly to those patients in whom prior treatment has failed.

These newer DAA regimens allow shorter courses of therapy, improved tolerability, and reduced resistance. For updated information, readers should consult *http://www.hcvguidelines.org/*.

ACKNOWLEDGMENT

The author thanks Raphael Dolin, MD, for his contributions to prior versions of this chapter and for years of mentorship.

FURTHER READING

American Association for the Study of Liver Disease/Infectious Diseases Society of America: Recommendations for testing, managing, and treating hepatitis C. Available from *http://www.hcvguidelines.org*. Accessed February 2, 2018.

crumpackercs CS: Ganciclovir. N Engl J Med 335:721, 1996.

Dienstag JL: Hepatitis B virus infection. N Engl J Med 359:1486, 2008.

Dobson J et al: Oseltamivir treatment for influenza in adults: A meta-analysis of randomized controlled trials. Lancet 385:1729, 2015.

Dolin R et al: A controlled trial of amantadine and rimantadine in the prophylaxis of influenza A infection. N Engl J Med 307:580, 1982.

Hurt AC et al: Overview of the 3rd isirv-Antiviral Group Conference—advances in clinical management. Influenza Other Respir Viruses 9:20, 2015.

Kamar N et al: Ribavirin for chronic hepatitis E virus infection in transplant recipients. N Engl J Med 370:1111, 2014.

Lai C-L et al: Entecavir versus lamivudine for patients with HBeAg-negative chronic hepatitis B. N Engl J Med 354:1011, 2006.

Marty FM et al: Letermovir prophylaxis for cytomegalovirus in hematopoietic-cell transplantation. N Engl J Med 377:2433, 2017.

Whitley RJ et al: Global assessment of resistance to neuraminidase inhibitors, 2008–2011: The Influenza Resistance Information Study (IRIS). Clin Infect Dis 58:1197, 2014.

Chapter 187: Herpes Simplex Virus Infections

Lawrence Corey

DEFINITION

Herpes simplex viruses (HSV-1, HSV-2; *Herpesvirus hominis*) produce a variety of infections involving mucocutaneous surfaces, the central nervous system (CNS), and—on occasion—visceral organs. Prompt recognition and treatment reduce the morbidity and mortality rates associated with HSV infections.

ETIOLOGIC AGENT



The genome of HSV is a 152-kb linear, double-stranded DNA molecule (molecular weight, ~100 × 10^6) that encodes >90 transcription units with 84 identified proteins. The genomic structures of the two HSV subtypes are similar. The overall genomic sequence homology between HSV-1 and HSV-2 is ~50%, whereas the proteome homology is >80%. The homologous sequences are distributed over the entire genome map, and most of the polypeptides specified by one viral type are antigenically related to polypeptides of the other viral type. Many type-specific regions unique to HSV-1 and HSV-2 proteins do exist, however, and a number of them appear to be important in host immunity. These type-specific regions have been used to develop serologic assays that distinguish between the two viral subtypes. Either restriction endonuclease analysis or sequencing of viral DNA can be used to distinguish between the two subtypes and among strains of each subtype. Recombinant viruses (HSV-1/HSV-2) do circulate in nature. The variability of nucleotide sequences from clinical strains of HSV-1 and HSV-2 is such that HSV isolates obtained from two individuals can be differentiated by restriction enzyme patterns or genomic sequences. Moreover, epidemiologically related sources, such as sexual partners, mother–infant pairs, or persons involved in a common-source outbreak, can be inferred from such patterns. Deep sequencing of sequential isolates suggests that more than one variant of HSV-1 or HSV-2 can be found in a single individual.

The viral genome is packaged in a regular icosahedral protein shell (capsid) composed of 162 capsomeres **(see Fig. 185-1)**. The outer covering of the virus is a lipid-containing membrane (envelope) acquired as the DNA-containing capsid buds through the inner nuclear membrane of the host cell. Between the capsid and lipid bilayer of the envelope is the tegument. Viral replication has both nuclear and cytoplasmic phases. Initial attachment to the cell membrane involves interactions of viral glycoproteins C and B with several cellular heparan sulfate–like surface receptors. Subsequently, viral glycoprotein D binds to cellular co-receptors that belong to the tumor necrosis factor receptor family of proteins, the immunoglobulin

superfamily (nectin family), or both. The ubiquity of these receptors contributes to the wide host range of herpesviruses. HSV replication is highly regulated. After fusion and entry, the nucleocapsid enters the cytoplasm and several viral proteins are released from the virion. Some of these viral proteins shut off host protein synthesis (by increasing cellular RNA degradation), whereas others "turn on" the transcription of early genes of HSV replication. These early gene products, designated *α genes*, are required for synthesis of the subsequent polypeptide group: the β polypeptides, many of which are regulatory proteins and enzymes required for DNA replication. Most current antiviral drugs interfere with β proteins, such as viral DNA polymerase. The third (γ) class of HSV genes requires viral DNA replication for expression and encodes most structural proteins specified by the virus. New antiviral drugs directed at viral assembly and release are under development.

After viral genome replication and structural protein synthesis, nucleocapsids are assembled in the cell's nucleus. Envelopment occurs as the nucleocapsids bud through the inner nuclear membrane into the perinuclear space. In some cells, viral replication in the nucleus forms two types of inclusion bodies: type A basophilic Feulgen-positive bodies that contain viral DNA and eosinophilic inclusion bodies that are devoid of viral nucleic acid or protein and represent a "scar" of viral infection. Enveloped virions are then transported via the endoplasmic reticulum and the Golgi apparatus to the cell surface.

Viral genomes are maintained by some neuronal cells in a repressed state called *latency*. Latency, which is associated with transcription of only a limited number of virus-encoded RNAs, accounts for the presence of viral DNA and RNA in neural tissue at times when infectious virus cannot be isolated. Maintenance and growth of neural cells from latently infected ganglia in tissue culture result in production of infectious virions (*explantation*) and in subsequent permissive infection of susceptible cells (*co-cultivation*). Activation of the viral genome may then occur, resulting in *reactivation*—the normal pattern of regulated viral gene expression and replication and HSV release. The release of virions from the neuron follows a complex process of anterograde transport down the length of neuronal axons. In experimental animals, ultraviolet light, systemic and local immunosuppression, and trauma to the skin or ganglia are associated with reactivation.

Three noncoding RNA latency-associated transcripts (LATs) are found in the nuclei of latently infected neurons. Microdissection plus real-time polymerase chain reaction (PCR) of individual neurons from cadaveric trigeminal ganglia explants revealed that many more neurons (2–10%) harbor HSV than would be predicted by in situ hybridization studies for LATs. Viral copy number is highly variable between neurons, with extremely high levels in certain neurons, and HSV DNA copy numbers are similar in LAT-positive and LAT-negative neurons. These findings add to the uncertainty about the role that LATs play in preventing reactivation. Deletion mutants of the LAT region exhibit reduced efficiency in their later reactivation. Substitution of HSV-1 LATs for HSV-2 LATs induces an HSV-1 reactivation pattern. These data indicate that LATs apparently maintain—rather than establish—latency. HSV-1 LATs promote the survival of acutely infected neurons, perhaps by inhibiting apoptotic pathways. LAT transcript abundance and low genome-copy number correlate with subnuclear positioning of HSV genomes around the centromere. Indeed, chromatization of HSV DNA appears to play a vital role in silencing expression of lytic replication genes. While the mechanism of latency and reactivation remains elusive, data suggest that viral micro-RNA appears to

silence expression of the key neurovirulence factor infected-cell protein 34.5 (ICP34.5) and to bind in an antisense configuration to the immediate-early protein ICP0 messenger RNA to prevent expression, which is vital to HSV reactivation. Although certain viral transcripts are known to be necessary for reactivation from latency, the molecular mechanisms of HSV latency are not fully understood, and strategies to interrupt or maintain latency in neurons are in developmental stages.

While latency is the predominant state of virus on a per-neuron basis, the high frequency of oral and genital tract reactivation for HSV-1 and HSV-2 suggests that the viruses are rarely quiescent within the entire biomass of ganglionic tissue. There is increasing recognition that HSV infection of the autonomic ganglia plays an important role in both initial and reactivation infections. In fact, deaths of animals from HSV-2 infection appear to be related to autonomic dysfunction of the bowel. Both HSV-1 and HSV-2 are shed subclinically. Most persons infected with HSV-2 and HSV-1 have frequent subclinical bursts of reactivation lasting 2–4 h, and the host tissue-based immune system can contain viral reactivation in the tissue before the development of clinical reactivation.

PATHOGENESIS

Exposure to HSV at mucosal surfaces or abraded skin sites permits entry of the virus into cells of the epidermis and dermis and initiation of viral replication therein. HSV infections are usually acquired subclinically. Whether clinical or subclinical, HSV acquisition is associated with sufficient viral replication to permit infection of sensory and/or autonomic nerve endings. On entry into the neuronal cell, the virus—or, more likely, the nucleocapsid—is transported intra-axonally to the nerve cell bodies in ganglia. Viral particles tether onto cellular proteins that motor along microtubules from axon tips (neurite endings) to neuronal cell bodies. In humans, the transit interval of spread to the ganglia after virus inoculation into peripheral tissue is unknown. During the initial phase of infection, viral replication occurs in ganglia and contiguous neural tissue. Virus then spreads to other mucocutaneous surfaces through centrifugal migration of infectious virions via peripheral sensory nerves. This mode of spread helps explain the large surface area involved, the high frequency of new lesions distant from the initial crop of vesicles that is characteristic in patients with primary genital or oral–labial HSV infection, and the ability to recover virus from neural tissue distant from neurons innervating the inoculation site. Contiguous spread of locally inoculated virus also may take place and allow further mucosal extension of disease. Recent studies have demonstrated HSV viremia—another mechanism for extension of infection throughout the body—in ~30–40% of persons with primary HSV-2 infection; latent infection with both viral subtypes in both sensory and autonomic ganglia has been demonstrated. For HSV-1 infection, trigeminal ganglia are most commonly infected, although extension to the inferior and superior cervical ganglia also occurs. With genital infection, sacral nerve root ganglia (S2–S5) are most commonly affected. Autonomic ganglia, pelvic nerves, and vaginal nerve roots are commonly infected.

After resolution of primary disease, infectious HSV can no longer be cultured from the ganglia; however, neuronal infection, as defined by the presence of viral DNA, persists in ganglionic cells in the anatomic regions of the initial infection. The mechanism of reactivation from latency is unknown, although

increasingly evidence of limited viral genes or microRNAs is identified in latently infected neurons. Evidence exists for viral antigen and activated host T cells at the ganglia and periphery, and immune responses in ganglia as well as peripheral tissue appear to influence the frequency and severity of HSV reactivation. HSV-specific T cells have been recovered from peripheral-nerve root ganglia. Many of these resident CD8+ T cells are juxtaposed with latently HSV-1-infected neurons in the trigeminal ganglia and can block reactivation with both interferon (IFN) γ release and granzyme B–mediated degradation of the immediate-early protein ICP4. In addition, there appears to be a latent viral load in the ganglia that correlates positively with the number of neurons infected and the rate of reactivation but inversely with the number of T cells present. It is not known whether reactivating stimuli transiently suppress these immune cells, independently upregulate transcription of lytic genes, or both. Moreover, host containment in the mucosa has been demonstrated. Once virus reaches the dermal–epidermal junction, there are three possible outcomes: (1) rapid host containment of infection near the site of reactivation; (2) spread of small amounts of virus into the epidermis, with a micro-ulceration associated with low-titer subclinical shedding; and (3) widespread replication and necrosis of epithelial cells and subsequent clinical recurrence (the latter defined clinically by a skin blister and ulceration). Histologically, herpetic lesions involve a thin-walled vesicle or ulceration in the basal region, multinucleated cells that may include intranuclear inclusions, necrosis, and an acute inflammatory infection. Re-epithelialization occurs once viral replication is restricted, almost always in the absence of a scar.

Analysis of the DNA from sequential isolates of HSV or from isolates from multiple infected ganglia in any one individual has revealed similar, if not identical, restriction endonuclease or DNA sequence patterns in most persons. As more sensitive genomic technologies are developed, evidence of multiple strains of the same subtype is increasingly being reported. For example, infection of individual neurons with multiple strains of drug-susceptible and drug-resistant virus in severely immunosuppressed patients indicates that ganglia can be reseeded during chronic infection. Because exposure to mucosal shedding is relatively common during a person's lifetime, current data suggest that exogenous infection with different strains of the same subtype does occur.

IMMUNITY

Host responses influence the acquisition of HSV disease, the severity of infection, resistance to the development of latency, the maintenance of latency, and the frequency of recurrences. Both antibody-mediated and cell-mediated reactions are clinically important. Immunocompromised patients with defects in cell-mediated immunity experience more severe and more extensive HSV infections than those with deficits in humoral immunity, such as agammaglobulinemia. Experimental ablation of lymphocytes indicates that T cells play a major role in preventing lethal disseminated disease, although antibodies help reduce titers of virus in neural tissue. Some clinical manifestations of HSV appear to be related to the host immune response (e.g., stromal opacities associated with recurrent herpetic keratitis). The surface viral glycoproteins have been shown to be targets of antibodies that mediate neutralization and immune-mediated cytolysis (antibody-dependent cell-mediated cytotoxicity). Monoclonal antibodies to HSV viral glycoproteins have, in experimental infections, conferred protection against subsequent neurologic disease or ganglionic latency. In humans, however, subunit glycoprotein vaccines have been largely ineffective in reducing acquisition of

infection. Multiple cell populations, including neutrophils, macrophages, and a variety of T lymphocytes, play a role in host defenses against HSV infections, as do lymphokines generated by T lymphocytes. In animals, passive transfer of primed lymphocytes confers protection from subsequent HSV challenge. Maximal protection usually requires the activation of multiple T-cell subpopulations, including cytotoxic T cells and T cells responsible for delayed hypersensitivity. The latter may confer protection by the antigen-stimulated release of lymphokines (e.g., IFNs), which in turn have a direct antiviral effect and both activate and enhance a variety of specific and nonspecific effector cells. The HSV virion contains a variety of genes that are directed at the inhibition of host responses. These include gene no. 12 (*US-12*), which can bind to the cellular transporter-activating protein TAP-1 and reduce the ability of this protein to bind HSV peptides to human leukocyte antigen class I, thereby reducing recognition of viral proteins by cytotoxic T cells of the host. This effect can be overcome by the addition of IFN-γ, but this reversal requires 24–48 h; thus, the virus has time to replicate and invade other host cells. Entry of infectious HSV-1 and HSV-2 inhibits several signaling pathways of both CD4+ and CD8+ T cells, leading to their functional impairment in killing and influencing the spectrum of their cytokine secretion.

HSV-specific CD8+ T-cell responses appear to be an important component in viral clearance from lesions. Immunosuppressed patients with frequent and prolonged HSV lesions have fewer functional CD8+ T cells directed at HSV. HSV-specific CD8+ T cells have been shown to persist in the genital skin at the dermal–epidermal junction contiguous to neuronal endings for months after lesion resolution. Even during clinical quiescence, these CD8+ T cells make both antiviral and cytotoxic proteins indicative of immune surveillance. These resident memory CD8+ T cells appear to be “first responders” capable of controlling viral reactivation at the site of viral release into the dermis. This rapid “on and off” interplay between the virus and the host helps explain the variability in clinical disease severity between episodes in any single individual. Differences of 30–60 min in host responses can result in 100- to 1000-fold differences in viral levels and can determine whether an episode of disease is subclinical or clinical.

There is a strong association between the magnitude of the CD8+ T lymphocyte response and the clearance of virus from genital lesions. The location, effectiveness, and longevity of the T lymphocytes (and perhaps of other immune effector cells) may be important in the expression of disease and the likelihood of transmission over time.

EPIDEMIOLOGY


Seroepidemiologic studies have documented HSV infections worldwide. The past 15 years have shown that the prevalence of HSV-2 is even higher in the developing than in the developed world. In sub-Saharan Africa, HSV-2 seroprevalence among pregnant women may approach 60%, and annual acquisition rates among teenage girls may verge on 20%. The global incidence has been estimated at ~23.6 million infections per year, with >400 million infected persons worldwide. As in the developed world, the rate of HSV-2 coital acquisition as well as the serologic prevalence is higher among women than among men. Most of this HSV-2 acquisition is preceded by acquisition of HSV-1; the frequency of genital HSV-1 in the developing world is low at present.

Infection with HSV-1 is acquired more frequently and earlier in life than infection with HSV-2. More than 90% of adults have antibodies to HSV-1 by the fifth decade of life. In populations of low socioeconomic status, most persons acquire HSV-1 infection before the third decade of life. Antibodies to HSV-2 are not detected routinely until puberty. Antibody prevalence rates correlate with past sexual activity and vary greatly among different population groups. There is evidence that the prevalence of HSV-2 has decreased slightly over the past decade or so in the United States. Serosurveys indicate that 15–20% of the U.S. population has antibodies to HSV-2. In most routine obstetric and family planning clinics, 25% of women have HSV-2 antibodies, although only 10% of those who are seropositive for HSV-2 report a history of genital lesions. As many as 50% of heterosexual adults attending sexually transmitted disease clinics have antibodies to HSV-2. A wide variety of serologic surveys has catalogued the widespread epidemic of HSV-2 in Central America, South America, and Africa. In Africa, HSV-2 seroprevalence has ranged from 40 to 70% in obstetric and other sexually experienced populations. Antibody prevalence rates average ~5–10% higher among women than among men.

Many studies continue to show that both incident and—more important—prevalent HSV-2 infection enhances the acquisition rate of HIV-1. More specifically, HSV-2 infection is associated on a population basis with a two- to fourfold increase in HIV-1 acquisition. This association has been amply demonstrated in heterosexual men and women in both the developed and the developing worlds. Epidemiologically, regions of the world with high HSV-2 prevalence and selected populations within such regions have a higher population-based incidence of HIV-1.

An important observation is that HSV-2 facilitates the spread of HIV into low-risk populations; prevalent HSV-2 appears to increase the risk of HIV infection by seven- to ninefold on a per-coital basis. Mathematical models suggest that ~33–50% of HIV-1 infections may be attributable to HSV-2 both in men who have sex with men (MSM) and in sub-Saharan Africa. In addition, HSV-2 is more frequently reactivated in and transmitted by persons co-infected with HIV-1 than in persons not co-infected. Thus, most areas of the world with a high HIV-1 prevalence also have a high HSV-2 prevalence. The shedding of HIV-1 virions from herpetic lesions in the genital region facilitates the spread of HIV through sexual contact. HSV-2 reactivation is associated with a localized persistent inflammatory response consisting of high concentrations of CCR5-enriched CD4+ T cells as well as inflammatory dendritic cells in the submucosa of the genital skin. These cells can support HIV infection and replication and thus are likely to account for the increased risk of HIV acquisition among persons with genital herpes. Unfortunately, antiviral therapy does not reduce this subclinical postreactivation inflammation, probably because of the inability of current antiviral agents to prevent the release of small amounts of HSV antigen into the genital mucosa.

Several studies suggest that many cases of “asymptomatic” genital HSV-2 infection are, in fact, simply unrecognized or confined to anatomic regions of the genital tract that are not easily visualized. Asymptomatic seropositive persons shed virus on mucosal surfaces almost as frequently as do those with symptomatic disease. This large reservoir of unidentified carriers of HSV-2 and the frequent asymptomatic reactivation of the virus from the genital tract have fostered the continued spread of genital herpes throughout the world.

HSV infections occur throughout the year. Transmission can result from contact with persons who have active ulcerative lesions or with persons who have no clinical manifestations of infection but who are shedding HSV from mucocutaneous surfaces. HSV reactivation on genital skin and mucosal surfaces is common. In fact, recent studies indicate that most HSV-1 and HSV-2 episodes last <4–6 h; thus, replication of the virus and clearance by the host are rapid. Even with once-daily sampling, HSV DNA can be detected on 20–30% of days by PCR. Corresponding figures for HSV-1 in oral secretions are similar. Rates of shedding are highest during the initial years after acquisition, with viral shedding occurring on as many as 30–50% of days during this period. Immunosuppressed patients shed HSV from mucosal sites at an even higher frequency (20–80% of days). These high rates of mucocutaneous reactivation suggest that exposure to HSV from sexual or other close contact (kissing, sharing of glasses or silverware) is common and help explain the continuing spread and high seroprevalence of HSV infections worldwide. Reactivation rates vary widely among individuals. Among HIV-positive patients, a low CD4+ T-cell count and a high HIV-1 load are associated with increased rates of HSV reactivation. Daily antiviral chemotherapy for HSV-2 infection can reduce shedding rates but does not eliminate shedding, as measured by PCR or culture.

CLINICAL SPECTRUM

HSV has been isolated from nearly all visceral and mucocutaneous sites. The clinical manifestations and course of HSV infection depend on the anatomic site involved, the age and immune status of the host, and the antigenic type of the virus. Primary HSV infections (i.e., first infections with either HSV-1 or HSV-2 in which the host lacks HSV antibodies in acute-phase serum) are frequently accompanied by systemic signs and symptoms. Compared with recurrent episodes, primary infections, which involve both mucosal and extramucosal sites, are characterized by a longer duration of symptoms and virus isolation from lesions. The incubation period ranges from 1 to 26 days (median, 6–8 days). Both viral subtypes can cause genital and oral–facial infections, and the infections caused by the two subtypes are clinically indistinguishable. However, the frequency of reactivation of infection is influenced by anatomic site and virus type. Genital HSV-2 infection is twice as likely to reactivate and recurs 8–10 times more frequently than genital HSV-1 infection. Conversely, oral–labial HSV-1 infection recurs more frequently than oral–labial HSV-2 infection. Asymptomatic shedding rates follow the same pattern.

Oral–Facial Infections

Gingivostomatitis and pharyngitis are the most common clinical manifestations of first-episode HSV-1 infection, whereas recurrent herpes labialis is the most common clinical manifestation of reactivation HSV-1 infection. HSV pharyngitis and gingivostomatitis usually result from primary infection and are most common among children and young adults. Clinical symptoms and signs, which include fever, malaise, myalgias, inability to eat, irritability, and cervical adenopathy, may last 3–14 days. Lesions may involve the hard and soft palate, gingiva, tongue, lip, and facial area. HSV-1 or HSV-2 infection of the pharynx usually results in exudative or ulcerative lesions of the posterior pharynx and/or tonsillar pillars. Lesions of the tongue, buccal mucosa, or gingiva may occur later in the course in one-third of cases. Fever lasting 2–7 days and cervical adenopathy are common. It can be difficult to differentiate HSV pharyngitis clinically from bacterial

pharyngitis, *Mycoplasma pneumoniae* infections, and pharyngeal ulcerations of noninfectious etiologies (e.g., Stevens-Johnson syndrome). No substantial evidence suggests that reactivation of oral–labial HSV infection is associated with symptomatic recurrent pharyngitis.

Reactivation of HSV from the trigeminal ganglia may be associated with asymptomatic virus excretion in the saliva, development of intraoral mucosal ulcerations, or herpetic ulcerations on the vermilion border of the lip or external facial skin. About 50–70% of seropositive patients undergoing trigeminal nerve-root decompression and 10–15% of those undergoing dental extraction develop oral–labial HSV infection a median of 3 days after these procedures. Clinical differentiation of intraoral mucosal ulcerations due to HSV from aphthous, traumatic, or drug-induced ulcerations is difficult.

In immunosuppressed patients, HSV infection may extend into mucosal and deep cutaneous layers. Friability, necrosis, bleeding, severe pain, and inability to eat or drink may result. The lesions of HSV mucositis are clinically similar to mucosal lesions caused by cytotoxic drug therapy, trauma, or fungal or bacterial infections. Persistent ulcerative HSV infections are among the most common infections in patients with AIDS. HSV and *Candida* infections often occur concurrently. Systemic antiviral therapy speeds the rate of healing and relieves the pain of mucosal HSV infections in immunosuppressed patients. The frequency of HSV reactivation during the early phases of transplantation or induction chemotherapy is high (50–90%), and prophylactic systemic antiviral agents such as IV acyclovir and penciclovir or the oral congeners of these drugs are used to reduce reactivation rates. Patients with atopic eczema may also develop severe oral–facial HSV infections (*eczema herpeticum*), which may rapidly involve extensive areas of skin and occasionally disseminate to visceral organs. Extensive eczema herpeticum has resolved promptly with the administration of IV acyclovir. Erythema multiforme may also be associated with HSV infections **(see Figs. 52-9 and A1-24)**; some evidence suggests that HSV infection is the precipitating event in ~75% of cases of cutaneous erythema multiforme. HSV antigen has been demonstrated both in circulatory immune complexes and in skin lesion biopsy samples from these cases. Patients with severe HSV-associated erythema multiforme are candidates for chronic suppressive oral antiviral therapy.

HSV-1 and varicella-zoster virus (VZV) have been implicated in the etiology of Bell's palsy (flaccid paralysis of the mandibular portion of the facial nerve). Some but not all trials have documented quicker resolution of facial paralysis with the prompt initiation of antiviral therapy, with or without glucocorticoids. However, other trials have shown little benefit. A recent Cochrane review indicates that there are advantages to the use of both antiviral drugs and glucocorticoids for moderate to severe Bell's palsy. Glucocorticoids alone are preferred for mild disease.

Genital Infections

First-episode primary genital herpes is characterized by fever, headache, malaise, and myalgias. Pain, itching, dysuria, vaginal and urethral discharge, and tender inguinal lymphadenopathy are the predominant local symptoms. Widely spaced bilateral lesions of the external genitalia are characteristic **(Fig. 187-1)**. Lesions may be present in varying stages, including vesicles, pustules, or painful erythematous ulcers. The cervix and urethra are involved in >80% of women with first-episode infections. First episodes of genital herpes in

patients who have had prior HSV-1 infection are associated with systemic symptoms in a few cases and with faster healing than primary genital herpes. Subclinical DNAemia has been found in ~30% of cases of true primary genital herpes. The clinical courses of acute first-episode genital herpes are similar for HSV-1 and HSV-2 infection. However, the recurrence rates of genital disease differ with the viral subtype: the 12-month recurrence rates among patients with first-episode HSV-2 and HSV-1 infections are ~90 and ~55%, respectively (median number of recurrences, 4 and <1, respectively). Recurrence rates for genital HSV-2 infections vary greatly among individuals and over time within the same individual. HSV has been isolated from the urethra and urine of men and women without external genital lesions. A clear mucoid discharge and dysuria are characteristics of symptomatic HSV urethritis. HSV has been isolated from the urethra of 5% of women with the dysuria–frequency syndrome. Occasionally, HSV genital tract disease is manifested by endometritis and salpingitis in women and by prostatitis in men. About 15% of cases of HSV-2 acquisition are associated with nonlesional clinical syndromes, such as aseptic meningitis, cervicitis, or urethritis. A more complete discussion of the differential diagnosis of genital herpes is presented in **Chap. 131**.

FIGURE 187-1

Genital herpes: primary vulvar infection, with multiple, extremely painful, punched-out, confluent, shallow ulcers on the edematous vulva and perineum. Micturition is often very painful. Associated inguinal lymphadenopathy is common. *(Reprinted with permission from K Wolff et al: Fitzpatrick's Color Atlas & Synopsis of Clinical Dermatology, 5th ed. New York, McGraw-Hill, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Both HSV-1 and HSV-2 can cause symptomatic or asymptomatic rectal and perianal infections. HSV proctitis is usually associated with rectal intercourse. However, subclinical perianal shedding of HSV is detected in women and men who report no rectal intercourse. This phenomenon is due to the establishment of latency in the sacral dermatome from prior genital tract infection, with subsequent reactivation in epithelial cells in the perianal region. Such reactivations are often subclinical. Symptoms of HSV proctitis include anorectal pain, anorectal discharge, tenesmus, and constipation. Sigmoidoscopy reveals ulcerative lesions of the distal 10 cm of the rectal mucosa. Rectal biopsies show mucosal ulceration, necrosis, polymorphonuclear and lymphocytic infiltration of the lamina propria, and (in occasional cases) multinucleated intranuclear inclusion-bearing cells. Perianal herpetic lesions are also found in immunosuppressed patients receiving cytotoxic therapy. Extensive perianal herpetic lesions and/or HSV proctitis is common among patients with HIV infection.

Herpetic Whitlow

Herpetic whitlow—HSV infection of the finger—may occur as a complication of primary oral or genital herpes by inoculation of virus through a break in the epidermal surface or by direct introduction of virus into the hand through occupational or some other type of exposure. Clinical signs and symptoms include abrupt-onset edema, erythema, and localized tenderness of the infected finger. Vesicular or pustular lesions of the fingertip that are indistinguishable from lesions of pyogenic bacterial infection are seen. Fever,

lymphadenitis, and epitrochlear and axillary lymphadenopathy are common. The infection may recur. Prompt diagnosis (to avoid unnecessary and potentially exacerbating surgical therapy and/or transmission) is essential. Antiviral chemotherapy is usually recommended (see below).

Herpes Gladiatorum

HSV may infect almost any area of skin. Mucocutaneous HSV infections of the thorax, ears, face, and hands have been described among wrestlers. Transmission of these infections is facilitated by trauma to the skin sustained during wrestling. Several recent outbreaks have illustrated the importance of prompt diagnosis and therapy to contain the spread of this infection.

Eye Infections

HSV infection of the eye is the most common cause of corneal blindness in the United States. HSV keratitis presents as an acute onset of pain, blurred vision, chemosis, conjunctivitis, and characteristic dendritic lesions of the cornea. Use of topical glucocorticoids may exacerbate symptoms and lead to involvement of deep structures of the eye. Debridement, topical antiviral treatment, and/or IFN therapy hastens healing. However, recurrences are common, and the deeper structures of the eye may sustain immunopathologic injury. Stromal keratitis due to HSV appears to be related to T cell–dependent destruction of deep corneal tissue. An HSV-1 epitope that is autoreactive with T cell–targeting corneal antigens has been postulated to be a factor in this infection. Chorioretinitis, usually a manifestation of disseminated HSV infection, may occur in neonates or in patients with HIV infection. HSV and VZV can cause acute necrotizing retinitis as an uncommon but severe manifestation.

Central and Peripheral Nervous System Infections

HSV accounts for 10–20% of all cases of sporadic viral encephalitis in the United States. The estimated incidence is ~2.3 cases per 1 million persons per year. Cases are distributed throughout the year, and the age distribution appears to be biphasic, with peaks at 5–30 and >50 years of age. HSV-1 causes >95% of cases.

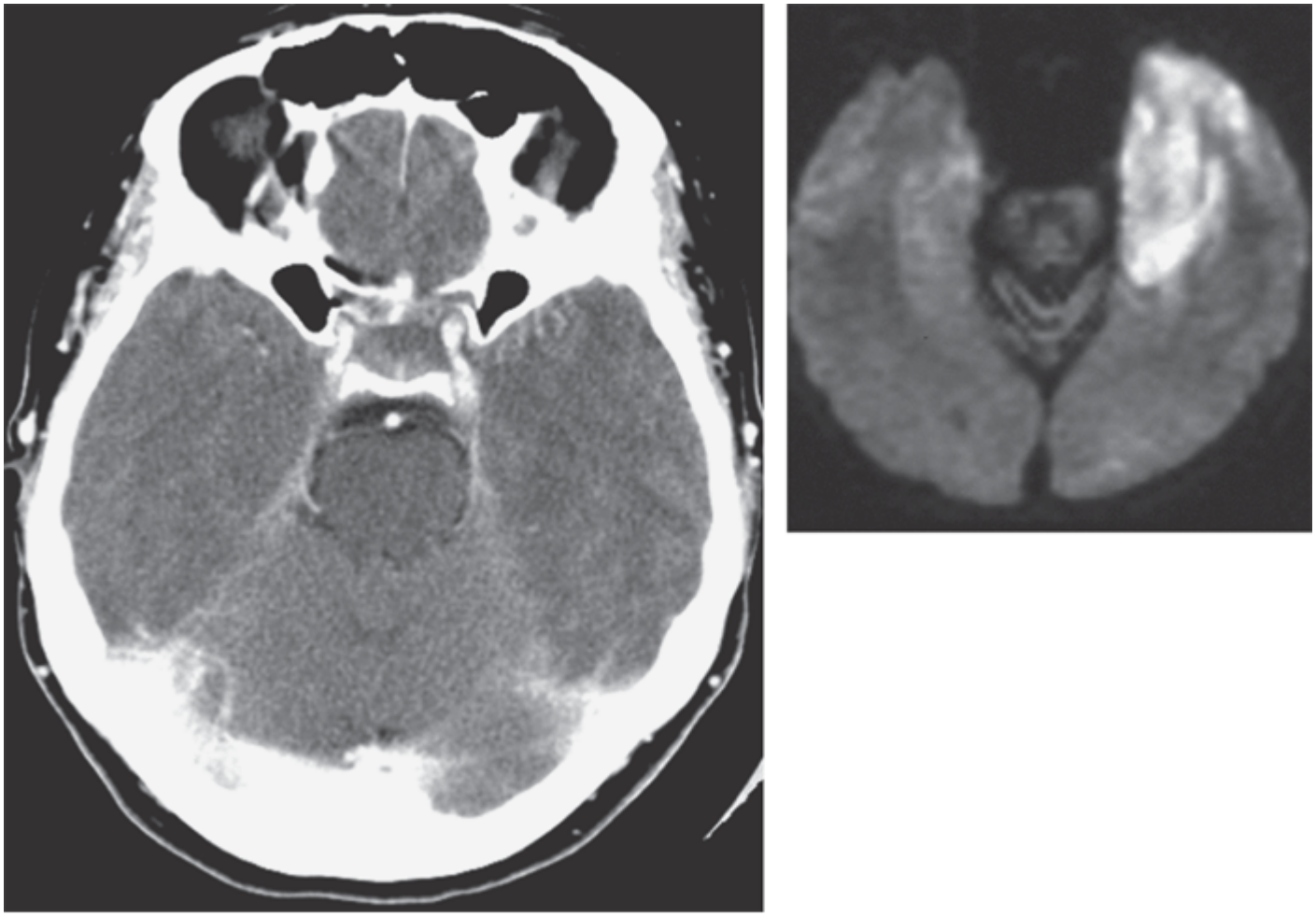
The pathogenesis of HSV encephalitis varies. In children and young adults, primary HSV infection may result in encephalitis; presumably, exogenously acquired virus enters the CNS by neurotropic spread from the periphery via the olfactory bulb. However, most adults with HSV encephalitis have clinical or serologic evidence of mucocutaneous HSV-1 infection before the onset of CNS symptoms. In ~25% of the cases examined, the HSV-1 strains from the oropharynx and brain tissue of the same patient differ; thus some cases may result from reinfection with another strain of HSV-1 that reaches the CNS. Two theories have been proposed to explain the development of actively replicating HSV in localized areas of the CNS in persons whose ganglionic and CNS isolates are similar. Reactivation of latent HSV-1 infection in trigeminal or autonomic nerve roots may be associated with extension of virus into the CNS via nerves innervating the middle cranial fossa. HSV DNA has been demonstrated by DNA hybridization in brain tissue obtained at autopsy—even from healthy adults. Thus, reactivation of long-standing latent CNS infection may be another mechanism for the development of HSV encephalitis.

 Recent studies have identified genetic polymorphisms among families with a high frequency of HSV encephalitis. Peripheral-blood mononuclear cells from these patients (predominantly children) appear to secrete reduced levels of IFN in response to HSV. Genetic mutations in *TLR3* documented in patients with HSV encephalitis suggest that some cases of sporadic HSV encephalitis may be related to host genetic determinants.

The clinical hallmark of HSV encephalitis has been the acute onset of fever and focal neurologic symptoms and signs, especially in the temporal lobe **(Fig. 187-2)**. Clinical differentiation of HSV encephalitis from other viral encephalitides, focal infections, or noninfectious processes is difficult. Elevated cerebrospinal fluid (CSF) protein levels, leukocytosis (predominantly lymphocytes), and red blood cell counts due to hemorrhagic necrosis are common. While brain biopsy has been the gold standard for defining HSV encephalitis, a highly sensitive and specific PCR for detection of HSV DNA in CSF has largely replaced biopsy for defining CNS infection. Although titers of antibody to HSV in CSF and serum increase in most cases of HSV encephalitis, they rarely do so earlier than 10 days into the illness and therefore, although useful in retrospect, generally are not helpful in establishing an early clinical diagnosis. In rare cases, demonstration of HSV antigen, HSV DNA, or HSV replication in brain tissue obtained by biopsy is highly sensitive; examination of such tissue also provides the opportunity to identify alternative, potentially treatable causes of encephalitis. Antiviral chemotherapy with acyclovir reduces the rate of death from HSV encephalitis. Most authorities recommend the administration of IV acyclovir to patients with presumed HSV encephalitis until the diagnosis is confirmed or an alternative diagnosis is made. All confirmed cases should be treated with IV acyclovir (30 mg/kg per day in three divided doses for 14–21 days). After the completion of therapy, the clinical recurrence of encephalitis requiring more treatment has been reported. For this reason, some authorities prefer to treat initially for 21 days, and many continue therapy until HSV DNA has been eliminated from the CSF. Even with therapy, neurologic sequelae are common, especially among persons >50 years of age.

FIGURE 187-2

Computed tomography and diffusion-weighted magnetic resonance imaging scans of the brain of a patient with left-temporal-lobe herpes simplex virus encephalitis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HSV DNA has been detected in CSF from 3 to 15% of persons presenting to the hospital with aseptic meningitis. HSV meningitis, which is usually seen in association with primary genital HSV infection, is an acute, self-limited disease manifested by headache, fever, and mild photophobia and lasting 2–7 days. Lymphocytic pleocytosis in the CSF is characteristic. Neurologic sequelae of HSV meningitis are rare. HSV is the most commonly identified cause of recurrent lymphocytic meningitis (*Mollaret's meningitis*). Demonstration of HSV antibodies in CSF or persistence of HSV DNA in CSF can establish the diagnosis. For persons with frequent recurrences of HSV meningitis, daily antiviral therapy has reduced the frequency of recurrent episodes of symptomatic meningitis.

Autonomic nervous system dysfunction, especially of the sacral region, has been reported in association with both HSV and VZV infections. Numbness, tingling of the buttocks or perineal areas, urinary retention, constipation, CSF pleocytosis, and (in males) impotence may occur. Symptoms appear to resolve slowly over days or weeks. Occasionally, hypoesthesia and/or weakness of the lower extremities persists for many months. Transitory hypoesthesia of the area of skin innervated by the trigeminal nerve and vestibular system dysfunction (as measured by electronystagmography) are the predominant signs of disease. Whether antiviral chemotherapy can abort these signs or reduce their frequency and severity is not yet known. Rarely, transverse myelitis, manifested by a rapidly progressive symmetric paralysis of the lower extremities or

Guillain-Barré syndrome, follows HSV infection. Similarly, peripheral nervous system involvement (Bell's palsy) or cranial polyneuritis may be related to reactivation of HSV-1 infection.

Visceral Infections

HSV infection of visceral organs usually results from viremia, and multiple-organ involvement is common. Occasionally, however, the clinical manifestations of HSV infection involve only the esophagus, lung, or liver. HSV esophagitis may result from direct extension of oral–pharyngeal HSV infection into the esophagus or may occur de novo by reactivation and spread of HSV to the esophageal mucosa via the vagus nerve. The predominant symptoms of HSV esophagitis are odynophagia, dysphagia, substernal pain, and weight loss. Multiple oval ulcerations appear on an erythematous base with or without a patchy white pseudomembrane. The distal esophagus is most commonly involved. With extensive disease, diffuse friability may spread to the entire esophagus. Neither endoscopic nor barium examination can reliably differentiate HSV esophagitis from *Candida* esophagitis or from esophageal ulcerations due to thermal injury, radiation, or corrosives. Endoscopically obtained secretions—for cytologic examination and culture or DNA detection by PCR—provide the most useful material for diagnosis. Systemic antiviral chemotherapy usually reduces the severity and duration of symptoms and heals esophageal ulcerations.

HSV pneumonitis is uncommon except in severely immunosuppressed patients and may result from extension of herpetic tracheobronchitis into lung parenchyma. Focal necrotizing pneumonitis usually ensues. Hematogenous dissemination of virus from sites of oral or genital mucocutaneous disease may also occur, producing bilateral interstitial pneumonitis. Bacterial, fungal, and parasitic pathogens are commonly present in HSV pneumonitis. The mortality rate from untreated HSV pneumonia in immunosuppressed patients is high (>80%). HSV has also been isolated from the lower respiratory tract of persons with acute respiratory distress syndrome and prolonged intubation. Most authorities believe that the presence of HSV in tracheal aspirates in such settings is due to reactivation of HSV in the tracheal region and localized tracheitis in persons with long-term intubation. Such patients should be evaluated for extension of HSV infection into the lung parenchyma. Controlled trials assessing the role of antiviral agents used against HSV in morbidity and mortality associated with acute respiratory distress syndrome have not been conducted. The role of lower respiratory tract HSV infection in overall rates of morbidity and mortality associated with these conditions is unclear. HSV is an uncommon cause of hepatitis in immunocompetent patients. HSV infection of the liver is associated with fever, abrupt elevations of bilirubin and serum aminotransferase levels, and leukopenia (<4000 white blood cells/μL). Disseminated intravascular coagulation may also develop.

Other reported complications of HSV infection include monarticular arthritis, adrenal necrosis, idiopathic thrombocytopenia, and glomerulonephritis. Disseminated HSV infection in immunocompetent patients is rare. In immunocompromised patients, burn patients, or malnourished individuals, HSV occasionally disseminates to other visceral organs, such as the adrenal glands, pancreas, small and large intestines, and bone marrow. Rarely, primary HSV infection in pregnancy disseminates and may be associated with the death of both mother and fetus. This uncommon event is usually related to the acquisition of primary

infection in the third trimester. Disseminated HSV infection is best detected by the presence of HSV DNA in plasma or blood.

Neonatal HSV Infections

Of all HSV-infected populations, neonates (infants <6 weeks) have the highest frequency of visceral and/or CNS infection. Without therapy, the overall rate of death from neonatal herpes is 65%; <10% of neonates with CNS infection develop normally. Although skin lesions are the most commonly recognized features of disease, many infants do not develop lesions at all or do so only well into the course of disease. Neonatal infection is usually acquired perinatally from contact with infected genital secretions at delivery. Congenitally infected infants have been reported. Of neonatal HSV infections, 30–50% are due to HSV-1 and 50–70% to HSV-2. The risk of developing neonatal HSV infection is 10 times higher for an infant born to a mother who has recently acquired HSV than for other infants. Neonatal HSV-1 infections may also be acquired through postnatal contact with immediate family members who have symptomatic or asymptomatic oral–labial HSV-1 infection or through nosocomial transmission within the hospital. All neonates with presumed herpes should be treated with IV acyclovir and then placed on maintenance oral antiviral therapy for the first 6–12 months of life. Antiviral chemotherapy with high-dose IV acyclovir (60 mg/kg per day) has reduced the mortality rate from neonatal herpes to ~15%. However, rates of morbidity, especially among infants with HSV-2 infection involving the CNS, are still very high.

HSV in Pregnancy

In the United States, 22% of all pregnant women and 55% of non-Hispanic black pregnant women are seropositive for HSV-2. However, the risk of mother-to-child transmission of HSV in the perinatal period is highest when the infection is acquired near the time of labor—that is, in previously HSV-seronegative women. The clinical manifestations of recurrent genital herpes—including the frequency of subclinical versus clinical infection, the duration of lesions, pain, and constitutional symptoms—are similar in pregnant and nonpregnant women. Recurrences increase in frequency over the course of pregnancy. However, when women are seropositive for HSV-2 at the outset of pregnancy, no effect on neonatal outcomes (including birth weight and gestational age) is seen. First-episode infections in pregnancy have more severe consequences for mother and infant. Maternal visceral dissemination during the third trimester occasionally occurs, as does premature birth or intrauterine growth retardation. The acquisition of primary disease in pregnancy, whether related to HSV-1 or HSV-2, carries the risk of transplacental transmission of virus to the neonate and can result in spontaneous abortion, although this outcome is relatively uncommon. For newly acquired genital HSV infection during pregnancy, most authorities recommend treatment with acyclovir (400 mg three times daily) or valacyclovir (500–1000 mg twice daily) for 7–10 days. However, the impact of this intervention on transmission is unknown. The high HSV-2 prevalence rate in pregnancy and the low incidence of neonatal disease (1 case per 6000–20,000 live births) indicate that only a few infants are at risk of acquiring HSV. Therefore, cesarean section is not warranted for all women with recurrent genital disease. Because intrapartum transmission of infection accounts for the majority of cases, abdominal delivery need be considered only for women who are shedding HSV at delivery. Several studies have shown no correlation

between recurrence of viral shedding before delivery and viral shedding at term. Hence, weekly virologic monitoring and amniocentesis are not recommended.

The frequency of transmission from mother to infant is markedly higher among women who acquire HSV near term (30–50%) than among those in whom HSV-2 infection is reactivated at delivery (<1%). Although maternal antibody to HSV-2 is protective, antibody to HSV-1 offers little or no protection against neonatal HSV-2 infection. Primary genital infection with HSV-1 leads to a particularly high risk of transmission during pregnancy and accounts for an increasing proportion of neonatal HSV cases. Moreover, during reactivation, HSV-1 appears more transmissible to the neonate than HSV-2. Only 2% of women who are seropositive for HSV-2 have HSV-2 isolated from cervical secretions at delivery, and only 1% of infants exposed in this manner develop infection, presumably because of the protective effects of maternally transferred antibodies and perhaps lower viral titers during reactivation. Despite the low frequency of transmission of HSV in this setting, 30–50% of infants with neonatal HSV are born to mothers with established genital herpes.

Isolation of HSV by cervicovaginal swab at the time of delivery is the greatest risk factor for intrapartum HSV transmission (relative risk = 346); however, culture-negative, PCR-positive cases of intrapartum transmission are well described. New acquisition of HSV (odds ratio [OR] = 49), isolation of HSV-1 versus HSV-2 (OR = 35), cervical versus vulvar HSV detection (OR = 15), use of fetal scalp electrodes (OR = 3.5), and young maternal age confer further risk of transmission, whereas cesarean delivery is protective (OR = 0.14). Physical examination poorly predicts the absence of shedding, and PCR far exceeds culture in terms of sensitivity and speed. Therefore, PCR detection at the onset of labor should be used to aid clinical decision-making for women with HSV-2 antibody. Because cesarean section appears to be an effective means of reducing maternal–fetal transmission, patients with recurrent genital herpes should be encouraged to come to the hospital early at the time of delivery for careful examination of the external genitalia and cervix as well as collection of a swab sample for viral isolation. Women who have no evidence of lesions can have a vaginal delivery. The presence of active lesions on the cervix or external genitalia is an indication for cesarean delivery.

If first-episode exposure has occurred (e.g., if HSV serologies show that the mother is seronegative or if the mother is HSV-1-seropositive and the isolate at delivery is found to be HSV-2), many authorities would initiate antiviral therapy for the infant with IV acyclovir. At a minimum, samples for viral cultures and PCR should be obtained from the throat, nasopharynx, eyes, and rectum of these infants immediately and at 5- to 10-day intervals. Lethargy, skin lesions, or fever should be evaluated promptly. All infants from whom HSV is isolated 24 h after delivery should be treated with IV acyclovir at recommended doses.

DIAGNOSIS

Both clinical and laboratory criteria are useful for diagnosing HSV infections. A clinical diagnosis can be made accurately when characteristic multiple vesicular lesions on an erythematous base are present. However, herpetic ulcerations may resemble skin ulcerations of other etiologies. Mucosal HSV infections may also present as urethritis or pharyngitis without cutaneous lesions. Thus, laboratory studies to confirm the

diagnosis and to guide therapy are recommended. While staining of scrapings from the base of the lesions with Wright's, Giemsa's (Tzanck preparation), or Papanicolaou's stain to detect giant cells or intranuclear inclusions of *Herpesvirus* infection is a well-described procedure, few clinicians are skilled in this technique, the sensitivity of staining is low (<30% for mucosal swabs), and these cytologic methods do not differentiate between HSV and VZV infections.

HSV infection is best confirmed in the laboratory by detection of virus, viral antigen, or viral DNA in scrapings from lesions. HSV DNA detection by PCR is the most sensitive laboratory technique for detecting mucosal or visceral HSV infections and is the recommended test for laboratory confirmation of a diagnosis. HSV causes a discernible cytopathic effect in a variety of cell culture systems, and this effect can be identified within 48–96 h after inoculation. Spin-amplified culture with subsequent staining for HSV antigen has shortened the time needed to identify HSV to <24 h. Culture is indicated when antiviral sensitivity testing is required. The sensitivity of all detection methods depends on the stage of the lesions (with higher sensitivity for vesicular than for ulcerative lesions), on whether the patient has a first or a recurrent episode of the disease (with higher sensitivity in first than in recurrent episodes), and on whether the sample is from an immunosuppressed or an immunocompetent patient (with more antigen or DNA in immunosuppressed patients). Laboratory confirmation permits subtyping of the virus; information on subtype may be useful epidemiologically and may help to predict the frequency of reactivation after first-episode oral–labial or genital HSV infection.

Both type-specific and type-common antibodies to HSV develop during the first several weeks after infection and persist indefinitely. Serologic assays with whole-virus antigen preparations, such as complement fixation, neutralization, indirect immunofluorescence, passive hemagglutination, radioimmunoassay, and enzyme-linked immunosorbent assay, are useful for differentiating uninfected (seronegative) persons from those with past HSV-1 or HSV-2 infection, but they do not reliably distinguish between the two viral subtypes. Serologic assays that identify antibodies to the type-specific glycoprotein G of the two viral subtypes (G1 and G2) are available commercially and can distinguish reliably between the human antibody responses to HSV-1 and HSV-2. Point-of-care assays that provide results from capillary blood or serum during a clinic visit are available. A western blot assay that can detect several HSV type-specific proteins can also be used. The presence of type-specific HSV-2 antibody implies past HSV-2 infection—i.e., latent infection and likely subclinical reactivation.

Acute- and convalescent-phase serum samples can be useful in demonstrating seroconversion during primary HSV-1 or HSV-2 infection. However, few available tests report titers, and increases in index values do not reflect first episodes in all patients. Serologic assays based on type-specific proteins should be used to identify asymptomatic carriers of HSV-1 or HSV-2. No reliable IgM method for defining acute HSV infection is available.

Several studies have shown that persons with previously unrecognized HSV-2 infection can be taught to identify symptomatic reactivations. Individuals seropositive for HSV-2 should be told about the high frequency of subclinical reactivation on mucosal surfaces that are not visible to the eye (e.g., cervix, urethra, perianal skin) or in microscopic ulcerations that may not be clinically symptomatic. Transmission of infection

during such episodes is well established. HSV-2-seropositive persons should be educated about the high likelihood of subclinical shedding and the role that condoms (male or female) may play in reducing transmission. Antiviral therapy with valacyclovir (500 mg once daily) has been shown to reduce the transmission of HSV-2 between sexual partners.

TREATMENT

TREATMENT

Herpes Simplex Virus Infections

Many aspects of mucocutaneous and visceral HSV infections are amenable to antiviral chemotherapy. For mucocutaneous infections, acyclovir and its congeners famciclovir and valacyclovir have been the mainstays of therapy. Several antiviral agents are available for topical use in HSV eye infections: idoxuridine, trifluorothymidine, topical vidarabine, and cidofovir. For HSV encephalitis and neonatal herpes, IV acyclovir is the treatment of choice.

All licensed antiviral agents for use against HSV inhibit the viral DNA polymerase. One class of drugs, typified by the drug acyclovir, is made up of substrates for the HSV enzyme thymidine kinase (TK). Acyclovir, ganciclovir, famciclovir, and valacyclovir are all selectively phosphorylated to the monophosphate form in virus-infected cells. Cellular enzymes convert the monophosphate form of the drug to the triphosphate, which is then incorporated into the viral DNA chain. Acyclovir is the agent most frequently used for the treatment of HSV infections and is available in IV, oral, and topical formulations. Valacyclovir, the valyl ester of acyclovir, offers greater bioavailability than acyclovir and thus can be administered less frequently. Famciclovir, the oral formulation of penciclovir, is clinically effective in the treatment of a variety of HSV-1 and HSV-2 infections. Ganciclovir is active against both HSV-1 and HSV-2; however, it is more toxic than acyclovir, valacyclovir, and famciclovir and generally is not recommended for the treatment of HSV infections. Anecdotal case reports suggest that ganciclovir may also be less effective than acyclovir for the treatment of HSV infections. All three recommended compounds—acyclovir, valacyclovir, and famciclovir—have proved effective in shortening the duration of symptoms and lesions of mucocutaneous HSV infections in both immunocompromised and immunocompetent patients **(Table 187-1)**. IV and oral formulations prevent reactivation of HSV in seropositive immunocompromised patients during induction chemotherapy or in the period immediately after bone marrow or solid organ transplantation. Chronic daily suppressive therapy reduces the frequency of reactivation disease among patients with frequent genital or oral–labial herpes. Only valacyclovir has been subjected to clinical trials that demonstrated reduced transmission of HSV-2 infection between sexual partners. IV acyclovir (30 mg/kg per day, given as a 10-mg/kg infusion over 1 h at 8-h intervals) is effective in reducing rates of death and morbidity from HSV encephalitis. Early initiation of therapy is a critical factor in outcome. The major side effect associated with IV acyclovir is transient renal insufficiency, usually due to crystallization of the compound in the renal parenchyma. This adverse reaction can be avoided if the medication is given slowly over 1 h and the patient is well hydrated. Because CSF levels of acyclovir average only 30–50% of plasma levels, the dosage of acyclovir used for treatment of CNS infection (30 mg/kg per day) is double that used for treatment of mucocutaneous or visceral disease (15

mg/kg per day). Even higher doses of IV acyclovir are used for neonatal HSV infection (60 mg/kg per day in three divided doses). Antiviral drugs neither eradicate latent infection nor affect the risk, frequency, or severity of subclinical or clinical recurrence after the drug is discontinued.

Increasingly, shorter courses of therapy are being used for recurrent mucocutaneous infection with HSV-1 or HSV-2 in immunocompetent patients. One-day courses of famciclovir and valacyclovir are clinically effective, more convenient, and generally less costly than longer courses of therapy (Table 187-1). These short-course regimens should be reserved for immunocompetent hosts.

SUPPRESSION OF MUCOCUTANEOUS HERPES

Recognition of the high frequency of subclinical reactivation provides a well-accepted rationale for the use of daily antiviral therapy to suppress reactivations of HSV, especially in persons with frequent clinical reactivations (e.g., those with recently acquired genital HSV infection). Immunosuppressed persons, including those with HIV infection, may also benefit from daily antiviral therapy. Daily acyclovir and valacyclovir reduce the frequency of HSV reactivations among HIV-positive persons. Regimens used include acyclovir (400–800 mg twice daily), famciclovir (500 mg twice daily), and valacyclovir (500 mg twice daily); valacyclovir at a dose of 4 g/d was associated with thrombotic thrombocytopenic purpura in one study of HIV-infected persons. Daily acyclovir therapy is associated with a modest reduction in the titer of HIV RNA in plasma (0.5-$\log_{10}$ reduction) and in the genital mucosa (0.33-$\log_{10}$ reduction).

REDUCED HSV TRANSMISSION TO SEXUAL PARTNERS

Once-daily valacyclovir (500 mg) has been shown to reduce transmission of HSV-2 between sexual partners. Transmission rates are higher from males to females and among persons with frequent HSV-2 reactivation. Serologic screening can be used to identify at-risk couples. Daily valacyclovir appears to be more effective at reducing subclinical shedding than daily famciclovir.

ACYCLOVIR RESISTANCE

Clinically relevant acyclovir-resistant strains of HSV do occur. Most of these strains have an altered substrate specificity for phosphorylating acyclovir. Thus, cross-resistance to famciclovir and valacyclovir is usually found. Occasionally, an isolate with altered TK specificity arises and is sensitive to famciclovir but not to acyclovir. In some patients infected with TK-deficient virus, higher doses of acyclovir are associated with clearing of lesions. In others, clinical disease progresses despite high-dose therapy. Almost all clinically significant acyclovir resistance has been seen in immunocompromised patients, and HSV-2 isolates are more often resistant than HSV-1 strains. A study by the Centers for Disease Control and Prevention indicated that ~5% of HSV-2 isolates from HIV-positive persons exhibit some degree of in vitro resistance to acyclovir. Of HSV-2 isolates from immunocompetent patients attending sexually transmitted disease clinics, <0.5% show reduced in vitro sensitivity to acyclovir. The lack of appreciable change in the frequency of detection of such isolates in the past 30 years probably reflects the reduced transmission of TK-deficient mutants. Isolation of HSV from lesions persisting despite adequate dosages and blood levels of acyclovir should raise the suspicion of acyclovir resistance. Clinical management of acyclovir resistance is challenging. Therapy with the antiviral drug foscarnet (40–80 mg/kg IV every 8 h until clinical resolution) is useful in acyclovir-resistant

cases (Chap. 186). Because of its toxicity and cost, this drug is usually reserved for patients with extensive mucocutaneous infections. Cidofovir is a nucleotide analogue and exists as a phosphonate or monophosphate form. Most TK-deficient strains of HSV are sensitive to cidofovir. Cidofovir ointment speeds healing of acyclovir-resistant lesions. No well-controlled trials of systemic cidofovir have been reported. Occasional cases may respond to topical imiquimod. True TK-negative variants of HSV appear to have a reduced capacity to spread because of altered neurovirulence—a feature important in the relatively infrequent presence of such strains in immunocompetent populations, even with increasing use of antiviral drugs. A new class of drugs that inhibit HSV-specific helicase/primase activity (pritelivir) is under clinical investigation and may offer a better toxicity profile for the treatment of acyclovir-resistant strains of HSV.

ACYCLOVIR EFFICACY IN THE DEVELOPING WORLD



Initial studies of acyclovir-like drugs were performed solely in the developed world. While acyclovir, valacyclovir, and famciclovir are effective in the developing world, their clinical and virologic benefits, especially in reducing the frequency of genital lesions among patients in Africa, seem reduced from those in European and U.S. populations. The mechanism of this phenomenon is uncertain. Acyclovir therapy does not reduce the rate of HIV acquisition; however, HIV load among MSM in the United States decreased by 1.3 $\log_{10}$ in contrast to 0.9 $\log_{10}$ among Peruvian MSM and 0.5 $\log_{10}$ among African women. Curiously, the anti-HIV drug tenofovir reduces HSV-2 acquisition among women in Africa although it has no demonstrable clinical benefit or antiviral effects among persons with established HSV-2 infection in studies in the United States. The reasons for these disparate results are unclear.

TABLE 187-1

Antiviral Chemotherapy for Herpes Simplex Virus (HSV) Infection

I. **Mucocutaneous HSV infections**

A. *Infections in immunosuppressed patients*

1. *Acute symptomatic first or recurrent episodes:* IV acyclovir (5 mg/kg q8h) or oral acyclovir (400 mg qid), famciclovir (500 mg bid or tid), or valacyclovir (500 mg bid) is effective. Treatment duration may vary from 7 to 14 days. IV therapy may be given for 2–7 days until clinical improvement and followed by oral therapy.
2. *Suppression of reactivation disease (genital or oral–labial):* IV acyclovir (5 mg/kg q8h) or oral valacyclovir (500 mg bid) or acyclovir (400–800 mg 3–5 times per day) prevents recurrences during the 30-day period immediately after transplantation. Longer-term HSV suppression is often used for persons with continued immunosuppression. In bone marrow and renal transplant recipients, oral valacyclovir (2 g/d) is also effective in reducing cytomegalovirus infection. Oral valacyclovir at a dose of 4 g/d has been associated with thrombotic thrombocytopenic purpura after extended use in HIV-positive persons. In HIV-infected persons, oral acyclovir (400–800 mg bid), valacyclovir (500 mg bid), or famciclovir (500 mg bid) is effective in reducing clinical and subclinical reactivations of HSV-1 and HSV-2.

B. *Infections in immunocompetent patients*

1. *Genital herpes*
   a. *First episodes:* Oral acyclovir (200 mg 5 times per day or 400 mg tid), valacyclovir (1 g bid), or famciclovir (250 mg bid) for 7–14 days is effective. IV acyclovir (5 mg/kg q8h for 5 days) is given for severe disease or neurologic complications such as aseptic meningitis.
   b. *Symptomatic recurrent genital herpes:* Short-course (1- to 3-day) regimens are preferred because of low cost, likelihood of adherence, and convenience. Oral acyclovir (800 mg tid for 2 days), valacyclovir (500 mg bid for 3 days), or famciclovir (750 or 1000 mg bid for 1 day, a 1500-mg single dose, or 500 mg stat followed by 250 mg q12h for 2 days) effectively shortens lesion duration. Other options include oral acyclovir (200 mg 5 times per day), valacyclovir (500 mg bid), and famciclovir (125 mg bid for 5 days).
   c. *Suppression of recurrent genital herpes:* Oral acyclovir (400–800 mg bid) or valacyclovir (500 mg daily) is given. Patients with >9 episodes per year should take oral valacyclovir (1 g daily or 500 mg bid) or famciclovir (250 mg bid or 500 mg bid).
2. *Oral–labial HSV infections*
   a. *First episode:* Oral acyclovir is given (200 mg 5 times per day or 400 mg tid); an oral acyclovir suspension can be used (600 mg/m^2 qid). Oral famciclovir (250 mg bid) or valacyclovir (1 g bid) has been used clinically. The duration of therapy is 5–10 days.
   b. *Recurrent episodes:* If initiated at the onset of the prodrome, single-dose or 1-day therapy effectively reduces pain and speeds healing. Regimens include oral famciclovir (a 1500-mg single dose or 750 mg bid for 1 day) or valacyclovir (a 2-g single dose or 2 g bid for 1 day).

Self-initiated therapy with 6-times-daily topical penciclovir cream effectively speeds healing of oral–labial HSV infection. Topical acyclovir cream has also been shown to speed healing.

c. *Suppression of reactivation of oral–labial HSV:* If started before exposure and continued for the duration of exposure (usually 5–10 days), oral acyclovir (400 mg bid) prevents reactivation of recurrent oral–labial HSV infection associated with severe sun exposure.

3. *Surgical prophylaxis of oral or genital HSV infection:* Several surgical procedures, such as laser skin resurfacing, trigeminal nerve-root decompression, and lumbar disk surgery, have been associated with HSV reactivation. IV acyclovir (3–5 mg/kg q8h) or oral acyclovir (800 mg bid), valacyclovir (500 mg bid), or famciclovir (250 mg bid) effectively reduces reactivation. Therapy should be initiated 48 h before surgery and continued for 3–7 days.
4. *Herpetic whitlow:* Oral acyclovir (200 mg) is given 5 times daily (alternative: 400 mg tid) for 7–10 days.
5. *HSV proctitis:* Oral acyclovir (400 mg 5 times per day) is useful in shortening the course of infection. In immunosuppressed patients or in patients with severe infection, IV acyclovir (5 mg/kg q8h) may be useful.
6. *Herpetic eye infections:* In acute keratitis, topical trifluorothymidine, vidarabine, idoxuridine, acyclovir, penciclovir, and interferon are all beneficial. Debridement may be required. Topical steroids may worsen disease.

II. **Central nervous system HSV infections**

A. *HSV encephalitis:* IV acyclovir (10 mg/kg q8h; 30 mg/kg per day) is given for 10 days or until HSV DNA is no longer detected in cerebrospinal fluid.

B. *HSV aseptic meningitis:* No studies of systemic antiviral chemotherapy exist. If therapy is to be given, IV acyclovir (15–30 mg/kg per day) should be used.

C. *Autonomic radiculopathy:* No studies are available. Most authorities recommend a trial of IV acyclovir.

III. **Neonatal HSV infections:** IV acyclovir (60 mg/kg per day, divided into 3 doses) is given. The recommended duration of IV treatment is 21 days. Monitoring for relapse should be undertaken. Continued suppression with oral acyclovir suspension should be given for 3–4 months.

IV. **Visceral HSV infections**

A. *HSV esophagitis:* IV acyclovir (15 mg/kg per day) is given. In some patients with milder forms of immunosuppression, oral therapy with valacyclovir or famciclovir is effective.

B. *HSV pneumonitis:* No controlled studies exist. IV acyclovir (15 mg/kg per day) should be considered.

V. **Disseminated HSV infections:** No controlled studies exist. IV acyclovir (5 mg/kg q8h) should be tried. Adjustments for renal insufficiency may be needed. No definite evidence indicates that therapy will decrease the risk of death.

VI. **Erythema multiforme associated with HSV:** Anecdotal observations suggest that oral acyclovir (400 mg bid or tid) or valacyclovir (500 mg bid) will suppress erythema multiforme.

VII. **Infections due to acyclovir-resistant HSV:** IV foscarnet (40 mg/kg IV q8h) should be given until lesions heal. The optimal duration of therapy and the usefulness of its continuation to suppress lesions are unclear. Some patients may benefit from cutaneous application of trifluorothymidine or 1% cidofovir gel, both of

which must be compounded at a pharmacy. These preparations should be applied once daily for 5–7 days. Topical imiquimod can be considered. The helicase primase inhibitor pritelivir is being studied for treatment of acyclovir-resistant HSV infection. IV cidofovir (5 mg/kg weekly) may be considered.

VIII. **Acyclovir and pregnancy:** No adverse effects to the fetus or newborn have been attributable to acyclovir. Acyclovir can be used in all stages of pregnancy and among women who are breastfeeding (the drug can be found in breast milk). Suppressive acyclovir treatment in late pregnancy reduces the frequency of cesarean delivery among women with recurrent genital herpes. Such treatment may not protect against transmission to neonates.

PREVENTION

Efforts to control HSV disease on a population basis through suppressive antiviral chemotherapy and/or educational programs have been limited. Barrier forms of contraception (especially condoms) decrease the likelihood of transmission of HSV infection, particularly during periods of asymptomatic viral excretion. When lesions are present, HSV infection may be transmitted by skin-to-skin contact despite the use of a condom. Nevertheless, the available data suggest that consistent condom use is an effective means of reducing the risk of genital HSV-2 transmission. Chronic daily antiviral therapy with valacyclovir can also be partially effective in reducing acquisition of HSV-2, especially among susceptible women. There are no comparative efficacy studies of valacyclovir versus condom use. Most authorities suggest both approaches. The need for a vaccine to prevent acquisition of HSV infection is great, especially in light of the role HSV-2 plays in enhancing the acquisition and transmission of HIV-1.

A substantial portion of neonatal HSV cases could be prevented by reducing the acquisition of HSV by women in the third trimester of pregnancy. Neonatal HSV infection can result from either the acquisition of maternal infection near term or the reactivation of infection at delivery in the already-infected mother. Women without known genital herpes should be counseled to abstain from vaginal intercourse during the third trimester with partners known to have or suspected of having genital herpes. Some authorities have recommended that antiviral therapy with acyclovir or valacyclovir be given to HSV-2-infected women in late pregnancy as a means of reducing reactivation of HSV-2 at term. Data are not available to support the efficacy of this approach, and the high treatment-to-prevention ratio makes this a difficult if not dubious public health strategy, even though it can reduce the frequency of HSV-associated cesarean delivery.

FURTHER READING

Gagyor I et al: Antiviral treatment for Bell's palsy (idiopathic facial paralysis). Cochrane Database Syst Rev 7:CD001869, 2015.

Gnann JW jr, Whitley RJ: Herpes simplex encephalitis: An update. Curr Infect Dis Rep 19:13, 2017.

Looker KJ et al: Global estimates of prevalent and incident herpes simplex virus type 2 infections in 2012. PLoS One 10:e114989, 2015. [PubMed: 25608026]

Looker KJ et al: Effect of HSV-2 infection on subsequent HIV acquisition: An updated systematic review and meta-analysis. Lancet Infect Dis 17:1303, 2017.

Wald A et al: Helicase-primase inhibitor pritelivir for HSV-2 infection. N Engl J Med 370:201, 2014.

Zhu J et al: Immune surveillance by CD8αα+ skin-resident T cells in human herpes virus infection. Nature 497:494, 2013.

Chapter 188: Varicella-Zoster Virus Infections

Richard J. Whitley

DEFINITION

Varicella-zoster virus (VZV) causes two distinct clinical syndromes: varicella (chickenpox) and herpes zoster (shingles). Chickenpox, a ubiquitous and extremely contagious infection, is usually a benign illness of childhood characterized by an exanthematous vesicular rash. With reactivation of latent VZV (which is most common after the sixth decade of life), herpes zoster presents as a dermatomal vesicular rash and is usually associated with severe pain.

ETIOLOGY

Early in the twentieth century, similarities in the histopathologic features of skin lesions resulting from varicella and herpes zoster were demonstrated. Viral isolates from patients with both of these diseases produced similar pathology in tissue culture—specifically, the appearance of eosinophilic intranuclear inclusions and multinucleated giant cells. These results suggested that the viruses were biologically similar. Restriction endonuclease analyses of viral DNA from a patient with chickenpox who subsequently developed herpes zoster verified the molecular identity of the two viruses responsible for these different clinical presentations.

VZV is a member of the family Herpesviridae, sharing with other members such structural characteristics as a lipid envelope surrounding a nucleocapsid with icosahedral symmetry, a total diameter of ~180–200 nm, and centrally located double-stranded DNA that is ~125,000 bp in length.

PATHOGENESIS AND PATHOLOGY

Primary Infection

Transmission occurs readily by the respiratory route; the subsequent localized replication of the virus at an undefined site (presumably the nasopharynx) leads to seeding of the lymphatic/reticuloendothelial system and ultimately to the development of viremia. Viremia in patients with chickenpox is reflected in the diffuse and scattered nature of the skin lesions and can be confirmed in selected cases by the recovery of VZV from the blood or routinely by the detection of viral DNA in either blood or lesions by polymerase chain reaction (PCR). Vesicles involve the corium and dermis, with degenerative changes characterized by ballooning, the

presence of multinucleated giant cells, and eosinophilic intranuclear inclusions. Infection may involve localized blood vessels of the skin, resulting in necrosis and epidermal hemorrhage. With the evolution of disease, the vesicular fluid becomes cloudy because of the recruitment of polymorphonuclear leukocytes and the presence of degenerated cells and fibrin. Ultimately, the vesicles either rupture and release their fluid (which includes infectious virus) or are gradually reabsorbed.

Recurrent Infection

The mechanism of reactivation of VZV that results in herpes zoster is unknown. The virus infects dorsal root ganglia during chickenpox, where it remains latent until reactivated. Histopathologic examination of representative dorsal root ganglia during active herpes zoster demonstrates hemorrhage, edema, and lymphocytic infiltration. Latent virus been detected in sensory (dorsal, cranial, and enteric) ganglia.

Active replication of VZV in other organs, such as the lung or the brain, can occur during either chickenpox or herpes zoster but is uncommon in the immunocompetent host. Pulmonary involvement is characterized by interstitial pneumonitis, multinucleated giant cell formation, intranuclear inclusions, and pulmonary hemorrhage. Central nervous system (CNS) infection leads to histopathologic evidence of perivascular cuffing similar to that encountered in measles and other viral encephalitides. Focal hemorrhagic necrosis of the brain, characteristic of herpes simplex virus (HSV) encephalitis, develops infrequently in VZV infection.

EPIDEMIOLOGY AND CLINICAL MANIFESTATIONS

Chickenpox

Humans are the only known reservoir for VZV. Chickenpox is highly contagious, with an attack rate of at least 90% among susceptible (seronegative) individuals. Persons of both sexes and all races are infected equally. The virus is endemic in the population at large; however, it becomes epidemic among susceptible individuals during seasonal peaks—namely, late winter and early spring in the temperate zone. Much of our knowledge of the disease's natural history and incidence predates the licensure of the chickenpox vaccine in 1995. Historically, children 5–9 years old were most commonly affected, accounting for 50% of all cases. Most other cases involved children 1–4 and 10–14 years old. Approximately 10% of the population of the United States over the age of 15 was susceptible to infection. VZV vaccination during the second year of life has dramatically changed the epidemiology of infection, causing a significant decrease in the annualized incidence of chickenpox, as noted below.

The incubation period of chickenpox ranges from 10 to 21 days but is usually 14–17 days. Secondary attack rates in susceptible siblings within a household are 70–90%. Patients are infectious ~48 h before the onset of the vesicular rash, during the period of vesicle formation (which generally lasts 4–5 days), and until all vesicles are crusted.

Clinically, chickenpox presents as a rash, low-grade fever, and malaise, although a few patients develop a prodrome 1–2 days before onset of the exanthem. In the immunocompetent patient, chickenpox is usually a

benign illness associated with lassitude and with body temperatures of 37.8°–39.4°C (100°–103°F) of 3–5 days' duration. The skin lesions—the hallmark of the infection—include maculopapules, vesicles, and scabs in various stages of evolution **(Fig. 188-1; see also Fig. A1-30)**. These lesions, which evolve from maculopapules to vesicles over hours to days, appear on the trunk and face and rapidly spread to involve other areas of the body. Most are small and have an erythematous base with a diameter of 5–10 mm. Successive crops appear over a 2- to 4-day period. Lesions can also be found on the mucosa of the pharynx and/or the vagina. Their severity varies from one person to another. Some individuals have very few lesions, while others have as many as 2000. Younger children tend to have fewer vesicles than older individuals. Within families, secondary and tertiary cases are associated with a larger number of vesicles than the first case. Immunocompromised patients—both children and adults, particularly those with leukemia—have lesions (often with a hemorrhagic base) that are more numerous and take longer to heal than those of immunocompetent patients. Immunocompromised individuals are also at greater risk for visceral complications, which occur in 30–50% of cases and are fatal 15% of the time in the absence of antiviral therapy.

FIGURE 188-1

Varicella lesions at various stages of evolution: vesicles on an erythematous base, umbilical vesicles, and crusts.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The most common infectious complication of varicella is secondary bacterial superinfection of the skin, which is usually caused by *Streptococcus pyogenes* or *Staphylococcus aureus,* including strains that are

methicillin-resistant. Skin infection results from excoriation of lesions after scratching. Gram's staining of skin lesions should help clarify the etiology of unusually erythematous and pustulated lesions.

The most common extracutaneous site of involvement in children is the CNS. The syndrome of acute cerebellar ataxia and meningeal inflammation generally appears ~21 days after onset of the rash and rarely develops in the pre-eruptive phase. The cerebrospinal fluid (CSF) contains lymphocytes and elevated levels of protein. CNS involvement is a benign complication of VZV infection in children and generally does not require hospitalization. Aseptic meningitis, encephalitis, transverse myelitis, and Guillain-Barré syndrome also can occur. Encephalitis is reported in 0.1–0.2% of children with chickenpox. Reye's syndrome can occur in children concomitantly treated with aspirin, which therefore is no longer used. Other than supportive care, no specific therapy (e.g., acyclovir administration) has proved efficacious for patients with CNS involvement.

Varicella pneumonia, the most serious complication following chickenpox, develops more often in adults (up to 20% of cases) than in children and is particularly severe in pregnant women. Pneumonia due to VZV usually has its onset 3–5 days into the illness and is associated with tachypnea, cough, dyspnea, and fever. Cyanosis, pleuritic chest pain, and hemoptysis are frequently noted. Roentgenographic evidence of disease consists of nodular infiltrates and interstitial pneumonitis. Resolution of pneumonitis parallels improvement of the skin rash; however, patients may have persistent fever and compromised pulmonary function for weeks.

Other complications of chickenpox include myocarditis, corneal lesions, nephritis, arthritis, bleeding diatheses, acute glomerulonephritis, and hepatitis. Hepatic involvement, distinct from Reye's syndrome and usually asymptomatic, is common in chickenpox and is generally characterized by elevated levels of liver enzymes, particularly aspartate and alanine aminotransferases.

Perinatal varicella is associated with mortality rates as high as 30% when maternal disease develops within 5 days before delivery or within 48 h thereafter. Illness in this setting is unusually severe because the newborn does not receive protective transplacental antibodies and has an immature immune system. *Congenital varicella*, with clinical manifestations of limb hypoplasia, cicatricial skin lesions, and microcephaly at birth, is extremely uncommon.

Herpes Zoster

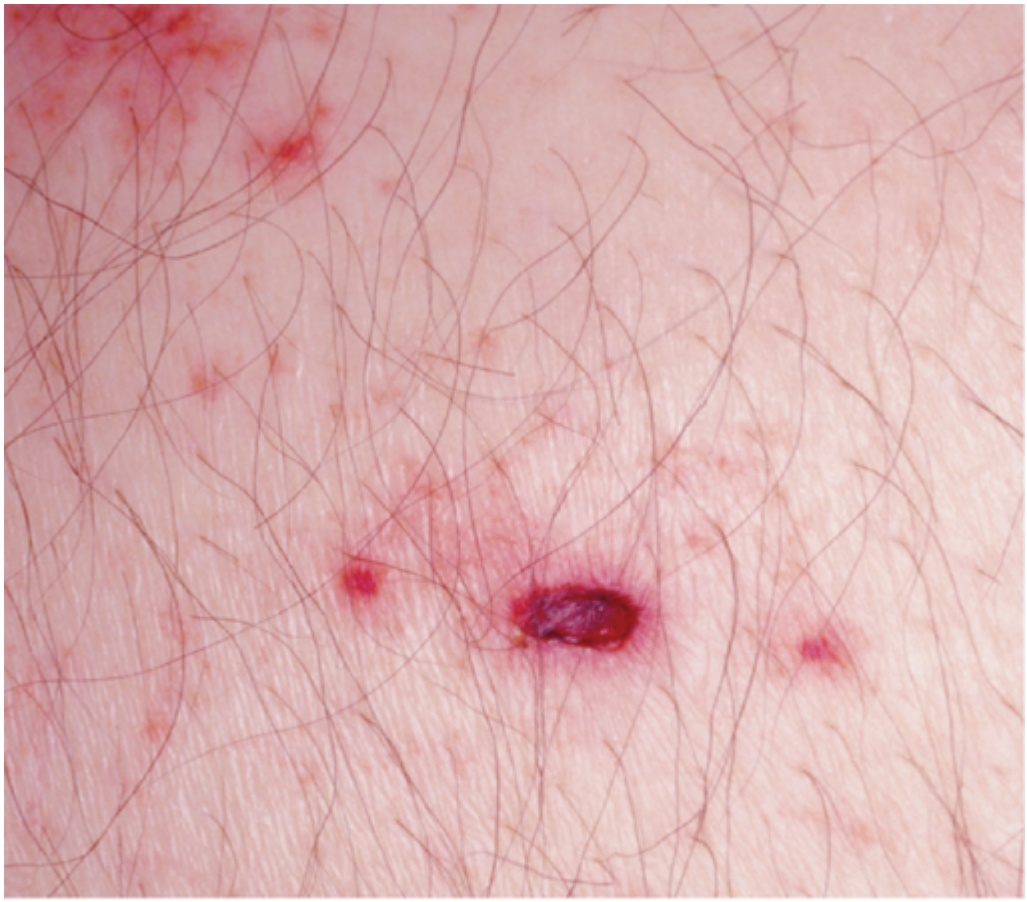
Herpes zoster (shingles) is a sporadic disease that results from reactivation of latent VZV from dorsal root ganglia. Most patients with shingles have no history of recent exposure to other individuals with VZV infection. Herpes zoster occurs at all ages, but its incidence is highest (5–10 cases per 1000 persons) among individuals in the sixth decade of life and beyond. Data suggest that at least 1.2 million cases occur annually in the United States. Recurrent herpes zoster is exceedingly rare except in immunocompromised hosts, especially those with AIDS.

Herpes zoster is characterized by a unilateral vesicular dermatomal eruption, often associated with severe pain. The dermatomes from T3 to L3 are most frequently involved. If the ophthalmic branch of the trigeminal

nerve is involved, *zoster ophthalmicus* results. The factors responsible for the reactivation of VZV are not known. In children, reactivation is usually benign; in adults, it can be debilitating because of pain. The onset of disease is heralded by pain within the dermatome, which may precede lesions by 48–72 h; an erythematous maculopapular rash evolves rapidly into vesicular lesions **(Fig. 188-2)**. In the normal host, these lesions may remain few in number and continue to form for only 3–5 days. The total duration of disease is generally 7–10 days; however, it may take as long as 2–4 weeks for the skin to return to normal. Patients with herpes zoster can transmit infection to seronegative individuals, with resulting chickenpox. In a few patients, characteristic localization of pain to a dermatome with serologic evidence of herpes zoster has been reported in the absence of skin lesions, an entity known as *zoster sine herpetica*. When branches of the trigeminal nerve are involved, lesions may appear on the face, in the mouth, in the eye, or on the tongue. *Zoster ophthalmicus* is usually a debilitating condition that can result in blindness in the absence of antiviral therapy. In *Ramsay Hunt syndrome*, pain and vesicles appear in the external auditory canal, and patients lose their sense of taste in the anterior two-thirds of the tongue while developing ipsilateral facial palsy. The geniculate ganglion of the sensory branch of the facial nerve is involved.

FIGURE 188-2

Close-up of lesions of disseminated zoster. Note lesions at different stages of evolution, including pustules and crusting. *(Photo courtesy of Lindsey Baden; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In both normal and immunocompromised hosts, the most debilitating complication of herpes zoster is pain associated with acute neuritis and postherpetic neuralgia. Postherpetic neuralgia is uncommon in young individuals; however, at least 50% of patients over age 50 report some degree of pain in the involved dermatome for months after the resolution of cutaneous disease. Changes in sensation in the dermatome, resulting in either hypo- or hyperesthesia, are common.

CNS involvement may follow localized herpes zoster. Many patients without signs of meningeal irritation have CSF pleocytosis and moderately elevated levels of CSF protein. Symptomatic meningoencephalitis is characterized by headache, fever, photophobia, meningitis, and vomiting. A rare manifestation of CNS involvement is granulomatous angiitis with contralateral hemiplegia, which can be diagnosed by cerebral arteriography. Other neurologic manifestations include transverse myelitis with or without motor paralysis.

Like chickenpox, herpes zoster is more severe in immunocompromised than immunocompetent individuals. Lesions continue to form for >1 week, and scabbing is not complete in most cases until 3 weeks into the illness. Patients with Hodgkin's disease and non-Hodgkin's lymphoma are at greatest risk for progressive herpes zoster. Cutaneous dissemination (Fig. 188-3) develops in ~40% of immunocompromised patients. Among patients with cutaneous dissemination, the risk of pneumonitis, meningoencephalitis, hepatitis, and other serious complications is increased by 5–10%. However, even in immunocompromised patients, disseminated zoster is rarely fatal.

FIGURE 188-3

Herpes zoster in an HIV-infected patient is seen as hemorrhagic vesicles and pustules on an erythematous base grouped in a dermatomal distribution.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Recipients of hematopoietic stem cell transplants are at particularly high risk of VZV infection. Of all cases of post-transplantation VZV infection, 30% occur within 1 year (50% of these within 9 months); 45% of the patients involved have cutaneous or visceral dissemination. The mortality rate in this situation is 10%. Postherpetic neuralgia, scarring, and bacterial superinfection are especially common in VZV infections occurring within 9 months of transplantation. Among infected patients, concomitant graft-versus-host disease increases the chance of dissemination and/or death.

Differential Diagnosis

The diagnosis of chickenpox is not difficult. The characteristic rash and a history of recent exposure should lead to a prompt diagnosis. Other viral infections that can mimic chickenpox include disseminated HSV infection in patients with atopic dermatitis and the disseminated vesiculopapular lesions sometimes associated with coxsackievirus infection, echovirus infection, or atypical measles. However, these rashes are more commonly morbilliform with a hemorrhagic component rather than vesicular or vesiculopustular. Rickettsialpox (Chap. 182) is sometimes confused with chickenpox; however, rickettsialpox can be distinguished easily by detection of the “herald spot” at the site of the mite bite and the development of a more pronounced headache. Serologic testing is also useful in differentiating rickettsialpox from varicella and can confirm susceptibility in adults unsure of their chickenpox history. Monkeypox can be considered in travelers returning from endemic areas (Chap. 191). Concern about smallpox has recently increased because of the threat of bioterrorism (Chap. S2). The lesions of smallpox are larger than those of chickenpox and are all at the same stage of evolution at any given time.

Unilateral vesicular lesions in a dermatomal pattern should lead rapidly to the diagnosis of herpes zoster, although the occurrence of shingles without a rash has been reported. Both HSV and coxsackievirus infections can cause dermatomal vesicular lesions. Supportive diagnostic virology and fluorescent staining of skin scrapings with monoclonal antibodies are helpful in ensuring the proper diagnosis. In the prodromal stage of herpes zoster, the diagnosis can be exceedingly difficult and may be made only after lesions have appeared or by retrospective serologic assessment.

LABORATORY FINDINGS

Unequivocal confirmation of the diagnosis is possible only through the isolation of VZV in susceptible tissue-culture cell lines, the demonstration of either seroconversion or a fourfold or greater rise in antibody titer between acute-phase and convalescent-phase serum specimens, or the detection of VZV DNA by PCR. Specimens for detection of VZV DNA by PCR include lesions, blood, and saliva. A rapid impression can be obtained by a Tzanck smear, with scraping of the base of the lesions in an attempt to demonstrate multinucleated giant cells; however, the sensitivity of this method is low (~60%). PCR technology for the detection of viral DNA in vesicular fluid is available in many diagnostic laboratories. Direct immunofluorescent staining of cells from the lesion base or detection of viral antigens by other assays (such as the immunoperoxidase assay) also is useful, although these tests are not commercially available. The most frequently employed serologic tools for assessing host response are the immunofluorescent detection of antibodies to VZV membrane antigens, the fluorescent antibody to membrane antigen (FAMA) test, immune adherence hemagglutination, and enzyme-linked immunosorbent assay (ELISA). The FAMA test and the ELISA appear to be most sensitive.

TREATMENT

TREATMENT

Varicella-Zoster Virus Infections

Medical management of chickenpox in the immunologically normal host is directed toward the prevention of avoidable complications. Obviously, good hygiene includes daily bathing and soaks. Secondary bacterial infection of the skin can be avoided by meticulous skin care, particularly with close cropping of fingernails. Pruritus can be decreased with topical dressings or the administration of antipruritic drugs. Tepid water baths and wet compresses are better than drying lotions for the relief of itching. Administration of aspirin to children with chickenpox should be avoided because of the association of aspirin derivatives with the development of Reye's syndrome. Acyclovir (800 mg by mouth five times daily), valacyclovir (1 g three times daily), or famciclovir (250 mg three times daily) for 5–7 days is recommended for adolescents and adults with chickenpox of ≤24 h duration. (Valacyclovir is licensed for use in children and adolescents. Famciclovir is recommended but not licensed for varicella.) Likewise, acyclovir therapy may be of benefit to children <12 years of age if initiated early in the disease (<24 h) at a dose of 20 mg/kg every 6 h. The advantages (i.e., pharmacokinetics) of the second-generation agents valacyclovir and famciclovir are described in **Chap. 186**.

Aluminum acetate soaks for the management of herpes zoster can be both soothing and cleansing. Patients with herpes zoster benefit from oral antiviral therapy, as evidenced by accelerated healing of lesions and resolution of zoster-associated pain with acyclovir, valacyclovir, or famciclovir. Acyclovir is administered at a dosage of 800 mg five times daily for 7–10 days. However, valacyclovir and famciclovir are superior in terms of pharmacokinetics and pharmacodynamics and should be used preferentially. Famciclovir, the prodrug of penciclovir, is at least as effective as acyclovir and perhaps more so; the dose is 500 mg by mouth three times daily for 7 days. Valacyclovir, the prodrug of acyclovir, accelerates healing and resolution of zoster-associated pain more promptly than acyclovir. The dose is 1 g by mouth three times daily for 5–7 days. Compared with acyclovir, both famciclovir and valacyclovir offer the advantage of less frequent administration. All three of these drugs are now available as generic products.

In severely immunocompromised hosts (e.g., transplant recipients, patients with lymphoproliferative malignancies), both chickenpox and herpes zoster (including disseminated disease) should be treated, at least at the outset, with IV acyclovir, which reduces the occurrence of visceral complications but has no effect on healing of skin lesions or pain. The dose is 10 mg/kg every 8 h for 7 days. For low-risk immunocompromised hosts, oral therapy with valacyclovir or famciclovir appears beneficial. If medically feasible, it is desirable to decrease immunosuppressive treatment concomitant with the administration of IV acyclovir.

Patients with varicella pneumonia typically require ventilatory support. Persons with zoster ophthalmicus should be referred immediately to an ophthalmologist. Therapy for this condition consists of the administration of analgesics for severe pain and the use of atropine. Acyclovir, valacyclovir, and famciclovir all accelerate healing. Decisions about the use of glucocorticoids should be made by the ophthalmologist.

The management of acute neuritis and/or postherpetic neuralgia can be particularly difficult. In addition to the judicious use of analgesics ranging from non-narcotics to narcotic derivatives, drugs such as gabapentin, pregabalin, amitriptyline hydrochloride, lidocaine (patches), and fluphenazine hydrochloride are reportedly beneficial for pain relief. In one study, glucocorticoid therapy administered early in the course of localized herpes zoster significantly accelerated such quality-of-life improvements as a return to usual activity and termination of analgesic medications. The dose of prednisone administered orally was 60 mg/d on days 1–7, 30 mg/d on days 8–14, and 15 mg/d on days 15–21. This regimen is appropriate only for relatively healthy elderly persons with moderate or severe pain at presentation. Patients with osteoporosis, diabetes mellitus, glycosuria, or hypertension may not be appropriate candidates. Glucocorticoids should not be used without concomitant antiviral therapy.

PREVENTION

Three methods are used for the prevention of VZV infections. First, a live attenuated varicella vaccine (Oka) is recommended for all children >1 year of age (up to 12 years of age) who have not had chickenpox and for adults known to be seronegative for VZV. Two doses are recommended for all children: the first at 12–15 months of age and the second at ~4–6 years of age. VZV-seronegative persons >13 years of age should receive

two doses of vaccine at least 1 month apart. The vaccine is both safe and efficacious. Breakthrough cases are mild and may result in spread of the vaccine virus to susceptible contacts. The universal vaccination of children has resulted in a decreased incidence of chickenpox in sentinel communities. Furthermore, inactivation of the vaccine virus significantly decreases the occurrence of herpes zoster after hematopoietic stem-cell transplantation.

In individuals >50 years of age, a VZV vaccine with 18 times the viral content of the Oka vaccine (ZostaVax) decreased the incidence of shingles by 51%, the burden of illness by 61%, and the incidence of postherpetic neuralgia by 66%. The Advisory Committee on Immunization Practices has therefore recommended that persons in this age group be offered this vaccine in order to reduce the frequency of shingles and the severity of postherpetic neuralgia. Of note, vaccine immunity wanes over time, and reassessment of current recommendations or the use of a promising inactivated vaccine in development will be required.

A second approach is to administer varicella-zoster immune globulin (VZIG) to individuals who are susceptible, are at high risk for developing complications of varicella, and have had a significant exposure. This product should be given within 96 h (preferably within 72 h) of the exposure. Indications for administration of VZIG appear in Table 188-1.

TABLE 188-1

Recommendations for Vzig Administration

Exposure Criteria
1. Significant exposure to a person with chickenpox or zoster a. Household: residence in the same household b. Playmate: face-to-face indoor play c. Hospital Varicella: same 2- to 4-bed room or adjacent beds in a large ward, face-to-face contact with an infectious staff member or patient, visit by a person deemed contagious Zoster: intimate contact (e.g., touching or hugging) with a person deemed contagious d. Newborn infant: onset of varicella in the mother ≤5 days before delivery or ≤48 h after delivery; VZIG not indicated if the mother has zoster 2. Patient should receive VZIG as soon as possible but not >96 h after exposure.
Candidates (Provided They Have Significant Exposure) Include
1. Immunocompromised susceptible children without a history of varicella or varicella immunization 2. Susceptible pregnant women 3. Newborn infants whose mother had onset of chickenpox within 5 days before or within 48 h after delivery 4. Hospitalized premature infant (≥28 weeks of gestation) whose mother lacks a reliable history of chickenpox or serologic evidence of protection against varicella 5. Hospitalized premature infant (<28 weeks of gestation or ≤1000-g birth weight), regardless of maternal history of varicella or VZV serologic status

Note: Table is adapted from the American Academy of Pediatrics *Red Book.*

Lastly, antiviral therapy can be given as prophylaxis to individuals at high risk who are ineligible for vaccination or who are beyond the 96-h window after direct contact. While the initial studies have used acyclovir, similar benefit can be anticipated with either valacyclovir or famciclovir. Therapy is instituted 7 days after intense exposure. At this time, the host is midway into the incubation period. This approach significantly decreases disease severity, if not totally preventing disease.

FURTHER READING

Arvin A: Aging, immunity, and the varicella-zoster virus. N Engl J Med 352:2266, 2005.

Gershon AA: Varicella zoster virus infection. Nat Rev Dis Primers 1:15016, 2015.

Gnann JW, Whitley RJ: Herpes zoster. N Engl J Med 347:340, 2002.

Hata A et al: Use of an inactivated varicella vaccine in recipients of hematopoietic-cell transplants. N Engl J Med 347:26, 2002.

Kimberlin DW, Whitley RJ: Varicella-zoster vaccine for the prevention of herpes zoster. N Engl J Med 356:1338, 2007.

Lal H et al: Efficacy of an adjuvanted herpes zoster subunit vaccine in older adults. N Engl J Med 372:2087, 2015.

Morrison VA et al: Long-term persistence of zoster vaccine efficacy. Clin Infect Dis 60:900, 2015.

Nguyen HQ et al: Decline in mortality due to varicella after implementation of varicella vaccination in the United States. N Engl J Med 352:450, 2005.

Oxman MN et al: A vaccine to prevent herpes zoster and postherpetic neuralgia in older adults. N Engl J Med 352:2271, 2005.

Seward JF et al: Varicella disease after introduction of varicella vaccine in the United States, 1995–2000. JAMA 287:606, 2002.

Seward JF et al: Contagiousness of varicella in vaccinated cases: A household contact study. JAMA 292:704, 2004.

Chapter 189: Epstein-Barr Virus Infections, Including Infectious Mononucleosis

Jeffrey I. Cohen

DEFINITION

Epstein-Barr virus (EBV) is the cause of heterophile-positive infectious mononucleosis (IM), which is characterized by fever, sore throat, lymphadenopathy, and atypical lymphocytosis. EBV is also associated with several tumors, including nasopharyngeal and gastric carcinoma, Burkitt's lymphoma, Hodgkin's disease, T cell lymphoma, and (in patients with immunodeficiencies) B cell lymphoma and smooth muscle tumors. The virus is a member of the family Herpesviridae. The two types of EBV that are widely prevalent in nature are not distinguishable by conventional serologic tests.

EPIDEMIOLOGY

 EBV infections occur worldwide. These infections are most common in early childhood, with a second peak during late adolescence. By adulthood, more than 90% of individuals have been infected and have antibodies to the virus. IM is usually a disease of young adults. In lower socioeconomic groups and in areas of the world with deficient standards of hygiene (e.g., developing regions), EBV tends to infect children at an early age, and IM is uncommon. In areas with higher standards of hygiene, infection with EBV is often delayed until adulthood, and IM is more prevalent.

EBV is spread by contact with oral secretions. The virus is frequently transmitted from asymptomatic adults to infants and among young adults by transfer of saliva during kissing. Transmission by less intimate contact is rare. EBV has been transmitted by blood transfusion and by bone marrow transplantation. More than 90% of asymptomatic seropositive individuals shed the virus in oropharyngeal secretions. Shedding is increased in immunocompromised patients and those with IM.

PATHOGENESIS

EBV is transmitted by salivary secretions. The virus infects the epithelium of the oropharynx and the salivary glands and is shed from these cells. While B cells may become infected after contact with epithelial cells, studies suggest that lymphocytes in the tonsillar crypts can be infected directly. The virus then spreads through the bloodstream. The proliferation and expansion of EBV-infected B cells along with reactive T cells

during IM result in enlargement of lymphoid tissue. Polyclonal activation of B cells leads to the production of antibodies to host-cell and viral proteins. During the acute phase of IM, up to 1 in every 100 B cells in the peripheral blood is infected by EBV; after recovery, 1–50 in every 1 million B cells is infected. During IM, there is an inverted CD4+/CD8+ T cell ratio. The percentage of CD4+ T cells decreases, while there are large clonal expansions of CD8+ T cells; up to 40% of CD8+ T cells are directed against EBV antigens during acute infection. Memory B cells, not epithelial cells, are the reservoir for EBV in the body. When patients are treated with acyclovir, shedding of EBV from the oropharynx stops but the virus persists in B cells.

The EBV receptor (CD21) on the surface of B cells is also the receptor for the C3d component of complement. Another EBV receptor (CD35) on B cells binds to CD21. Human leukocyte antigen class II serves as a co-receptor for EBV entry into B cells. EBV infection of epithelial cells occurs by virus binding to integrins and results in viral replication and production of virions. When B cells are infected by EBV in vitro, they become transformed and can proliferate indefinitely. During latent infection of B cells, the EBV nuclear antigens (EBNAs), latent membrane proteins (LMPs), multiple microRNAs, and small EBV RNAs (EBERs) are expressed in vitro. EBV-transformed B cells secrete immunoglobulin; only a small fraction of these cells produce virus.

Cellular immunity is more important than humoral immunity in controlling EBV infection. In the initial phase of infection, suppressor T cells, natural killer cells, and nonspecific cytotoxic T cells are important in controlling the proliferation of EBV-infected B cells. Levels of markers of T cell activation and serum interferon γ are elevated. Later in infection, human leukocyte antigen–restricted cytotoxic T cells that recognize EBNAs and LMPs and destroy EBV-infected cells are generated.

If T cell immunity is compromised, EBV-infected B cells may begin to proliferate. When EBV is associated with lymphoma in immunocompetent persons, virus-induced proliferation is but one step in a multistep process of neoplastic transformation. In many EBV-containing tumors, LMP-1 mimics members of the tumor necrosis factor receptor family (e.g., CD40), transmitting growth-proliferating signals.

CLINICAL MANIFESTATIONS

Signs and Symptoms

Most EBV infections in infants and young children either are asymptomatic or present as mild pharyngitis with or without tonsillitis. In contrast, ~75% of infections in adolescents present as IM. IM in the elderly often presents with nonspecific symptoms, including prolonged fever, fatigue, myalgia, and malaise. In contrast, pharyngitis, lymphadenopathy, splenomegaly, and atypical lymphocytes are relatively rare in elderly patients.

The incubation period for IM in young adults is ~4–6 weeks. A prodrome of fatigue, malaise, and myalgia may last for 1–2 weeks before the onset of fever, sore throat, and lymphadenopathy. Fever is usually low-grade and is most common in the first 2 weeks of the illness; however, it may persist for >1 month. Common signs and symptoms are listed along with their frequencies in **Table 189-1**. Lymphadenopathy and pharyngitis are most prominent during the first 2 weeks of the illness, while splenomegaly is more prominent during the

second and third weeks. Lymphadenopathy most often affects the posterior cervical nodes but may be generalized. Enlarged lymph nodes are frequently tender and symmetric but are not fixed in place. Pharyngitis, often the most prominent sign, can be accompanied by enlargement of the tonsils with an exudate resembling that of streptococcal pharyngitis. A morbilliform or papular rash, usually on the arms or trunk, develops in ~5% of cases **(Fig. 189-1)**. Earlier studies reported that many patients treated with penicillin derivatives develop a macular rash; penicillin-associated rashes are not predictive of future adverse reactions to penicillins. More recent studies suggest that EBV-associated rashes may occur with similar frequency in those exposed to penicillin derivatives and those not taking these drugs. Erythema nodosum **(Fig. A1-39)** and erythema multiforme **(Fig. A1-24)** also have been described **(Chap. 54)**. The severity of the disease correlates with the levels of CD8+ T cells and EBV DNA in the blood. Most patients have symptoms for 2–4 weeks, but nearly 10% have fatigue that persists for ≥6 months.

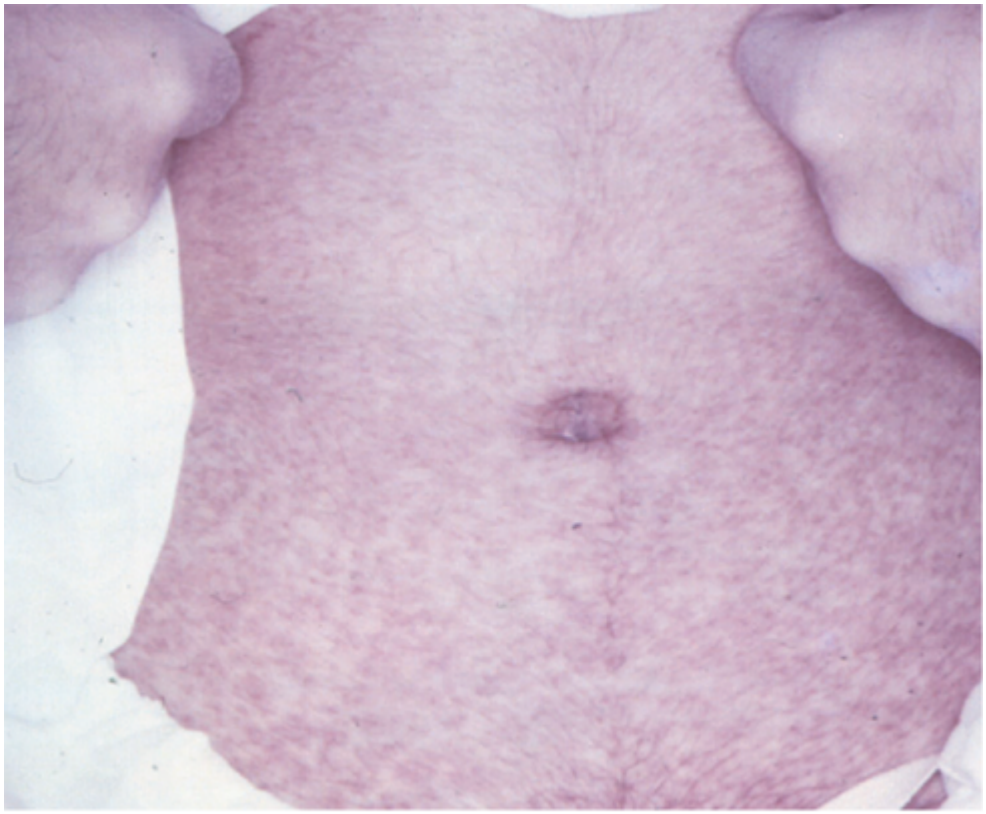
TABLE 189-1

Signs and Symptoms of Infectious Mononucleosis

Manifestation	Median Percentage of Patients (Range)
Symptoms	
Sore throat	75 (50–87)
Malaise	47 (42–76)
Headache	38 (22–67)
Abdominal pain, nausea, or vomiting	17 (5–25)
Chills	10 (9–11)
Signs	
Lymphadenopathy	95 (83–100)
Fever	93 (60–100)
Pharyngitis or tonsillitis	82 (68–90)
Splenomegaly	51 (43–64)
Hepatomegaly	11 (6–15)
Rash	10 (0–25)
Periorbital edema	13 (2–34)
Palatal enanthem	7 (3–13)
Jaundice	5 (2–10)

FIGURE 189-1

Rash in a patient with infectious mononucleosis due to Epstein-Barr virus. *(Courtesy of Maria Turner, MD; with permission.)*



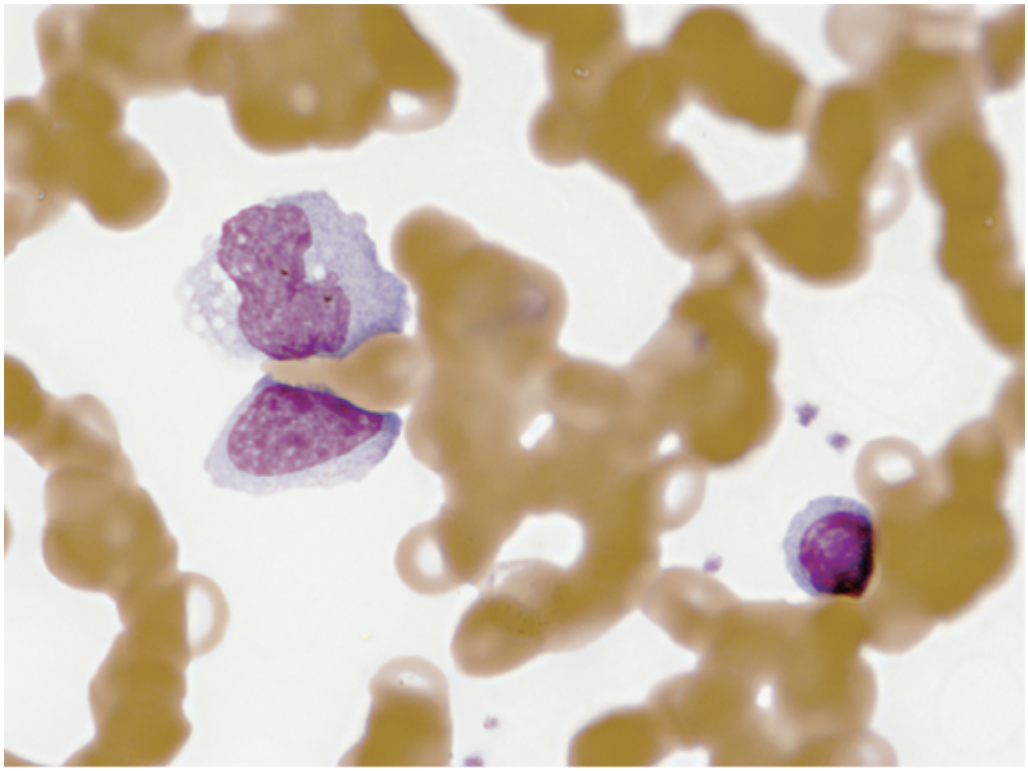
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Laboratory Findings

The white blood cell count is usually elevated and peaks at 10,000–20,000/μL during the second or third week of illness. Lymphocytosis is usually demonstrable, with >10% atypical lymphocytes. The latter cells are enlarged lymphocytes that have abundant cytoplasm, vacuoles, and indentations of the cell membrane **(Fig. 189-2)**. CD8+ T cells predominate among the atypical lymphocytes. Low-grade neutropenia and thrombocytopenia are common during the first month of illness. Liver function is abnormal in >90% of cases. Serum levels of aminotransferases and alkaline phosphatase are usually mildly elevated. The serum concentration of bilirubin is elevated in ~40% of cases.

FIGURE 189-2

Atypical lymphocytes from a patient with infectious mononucleosis due to Epstein-Barr virus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Complications

Most cases of IM are self-limited. Deaths are very rare and are most often due to central nervous system (CNS) complications, splenic rupture, upper-airway obstruction, or bacterial superinfection.

When CNS complications develop, they usually do so during the first 2 weeks of EBV infection; in some patients, especially children, they are the only clinical manifestations of acute infection. Heterophile antibodies and atypical lymphocytes may be absent. Meningitis and encephalitis are the most common neurologic abnormalities, and patients may present with headache, meningismus, or cerebellar ataxia. Acute hemiplegia and psychosis also have been described. The cerebrospinal fluid contains mainly lymphocytes, with occasional atypical lymphocytes. Most cases resolve without neurologic sequelae. Acute EBV infection has also been associated with cranial nerve palsies (especially those involving cranial nerve VII), Guillain-Barré syndrome, acute transverse myelitis, and peripheral neuritis.

Autoimmune hemolytic anemia occurs in ~2% of cases during the first 2 weeks. In most cases, the anemia is Coombs-positive, with cold agglutinins directed against the red blood cell antigen. Most patients with hemolysis have mild anemia that lasts for 1–2 months, but some patients have severe disease with hemoglobinuria and jaundice. Nonspecific antibody responses may also include rheumatoid factor, antinuclear antibodies, anti–smooth muscle antibodies, antiplatelet antibodies, and cryoglobulins. IM has been associated with red-cell aplasia, severe granulocytopenia, thrombocytopenia, pancytopenia, and hemophagocytic lymphohistiocytosis. The spleen ruptures in <0.5% of cases. Splenic rupture is more

common among male than female patients and may manifest as abdominal pain, referred shoulder pain, or hemodynamic compromise.

Hypertrophy of lymphoid tissue in the tonsils or adenoids can result in upper-airway obstruction, as can inflammation and edema of the epiglottis, pharynx, or uvula. About 10% of patients with IM develop streptococcal pharyngitis after their initial sore throat resolves.

Other rare complications associated with acute EBV infection include hepatitis (which can be fulminant), myocarditis or pericarditis, pneumonia with pleural effusion, interstitial nephritis, genital ulcerations, and vasculitis.

EBV-Associated Diseases Other Than IM

EBV-associated lymphoproliferative disease has been described in patients with congenital or acquired immunodeficiency, including those with severe combined immunodeficiency, patients with AIDS, and recipients of bone marrow or organ transplants who are receiving immunosuppressive drugs (especially cyclosporine). Proliferating EBV-infected B cells infiltrate lymph nodes and multiple organs, and patients present with fever and lymphadenopathy or gastrointestinal symptoms. Pathologic studies show B cell hyperplasia or poly- or monoclonal lymphoma.



X-linked lymphoproliferative disease is a recessive disorder of young boys who have a normal response to childhood infections but develop fatal lymphoproliferative disorders after infection with EBV. The protein associated with most cases of this syndrome (SAP) binds to a protein that mediates interactions of B and T cells. Most patients with this syndrome die of acute IM. Others develop hypogammaglobulinemia, malignant B cell lymphomas, aplastic anemia, or agranulocytosis. Disease resembling X-linked lymphoproliferative disease, but with more prominent hemophagocytosis, has also been associated with mutations in XIAP. Mutations in *ITK*, *MAGT1*, *CORO1A*, or *CD27* are associated with inability to control EBV and lymphoma. Mutations in other genes, such as *GATA2*, *PIK3CD*, *CTPS1*, and several genes associated with severe combined immunodeficiency, also can predispose to severe or fatal EBV disease as well as other infections. Moreover, IM has proved fatal to some patients with no obvious preexisting immune abnormality.

Oral hairy leukoplakia **(Fig. 189-3)** is an early manifestation of infection with HIV in adults **(Chap. 197)**. Most patients present with raised, white corrugated lesions on the tongue (and occasionally on the buccal mucosa) that contain EBV DNA. Children infected with HIV can develop lymphoid interstitial pneumonitis; EBV DNA is often found in lung tissue from these patients.

FIGURE 189-3

Oral hairy leukoplakia often presents as white plaques on the lateral surface of the tongue and is associated with Epstein-Barr virus infection.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Patients with chronic fatigue syndrome may have titers of antibody to EBV that are elevated but are not significantly different from those in healthy EBV-seropositive adults. While some patients have malaise and fatigue that persist for weeks or months after IM, persistent EBV infection is not a cause of chronic fatigue syndrome. Chronic active EBV infection is very rare and is distinct from chronic fatigue syndrome. The affected patients have an illness lasting >6 months, with elevated levels of EBV DNA in the blood (in T cells, NK cells, or B cells); high titers of antibody to EBV; and evidence of organ involvement, including hepatosplenomegaly, lymphadenopathy, and pneumonitis, uveitis, or neurologic disease.

EBV is associated with several malignancies. About 15% of cases of Burkitt's lymphoma in the United States and ~90% of those in Africa are associated with EBV (**Chap. 104**). African patients with Burkitt's lymphoma have high levels of antibody to EBV, and their tumor tissue usually contains viral DNA. Malaria in African patients may impair cellular immunity to EBV and induce polyclonal B cell activation with an expansion of EBV-infected B cells. In addition, malaria may target B cells and result in expansion of germinal centers, with consequently increased activity of activation-induced cytidine deaminase, which can mutate DNA. These changes may enhance the proliferation of B cells with elevated EBV DNA in the bloodstream, thereby increasing the likelihood of a *c-myc* translocation—the hallmark of Burkitt's lymphoma. EBV-containing Burkitt's lymphoma also occurs in patients with AIDS.

Anaplastic nasopharyngeal carcinoma is common in southern China and is uniformly associated with EBV; the affected tissues contain viral DNA and antigens. Patients with nasopharyngeal carcinoma often have elevated titers of antibody to EBV (**Chap. 73**). High levels of EBV plasma DNA before treatment or detectable levels of EBV DNA after radiation therapy correlate with lower rates of overall survival and relapse-free survival among patients with nasopharyngeal carcinoma.

Worldwide, the most common EBV-associated malignancy is gastric carcinoma. About 9% of these tumors are EBV-positive (**Chap. 76**).

EBV has been associated with Hodgkin's disease, especially the mixed-cellularity type (Chap. 105). Patients with Hodgkin's disease often have elevated titers of antibody to EBV. In about half of cases in the United States, viral DNA and antigens are found in Reed-Sternberg cells. The risk of EBV-positive Hodgkin's disease is significantly increased in young adults for several years after EBV-seropositive IM. About 50% of non-Hodgkin's lymphomas in patients with AIDS are EBV-positive.

EBV is present in B cells of lesions from patients with lymphomatoid granulomatosis. In some cases, EBV DNA has been detected in tumors from immunocompetent patients with angiocentric nasal NK/T cell lymphoma, T cell lymphoma, and CNS lymphoma. Studies have demonstrated viral DNA in leiomyosarcomas from AIDS patients and in smooth-muscle tumors from organ transplant recipients. Virtually all CNS lymphomas in AIDS patients are associated with EBV. Studies have found that a history of IM and higher levels of antibodies to EBV before the onset of disease is more common in persons with multiple sclerosis than in the general population; additional research on a possible causal relationship is needed.

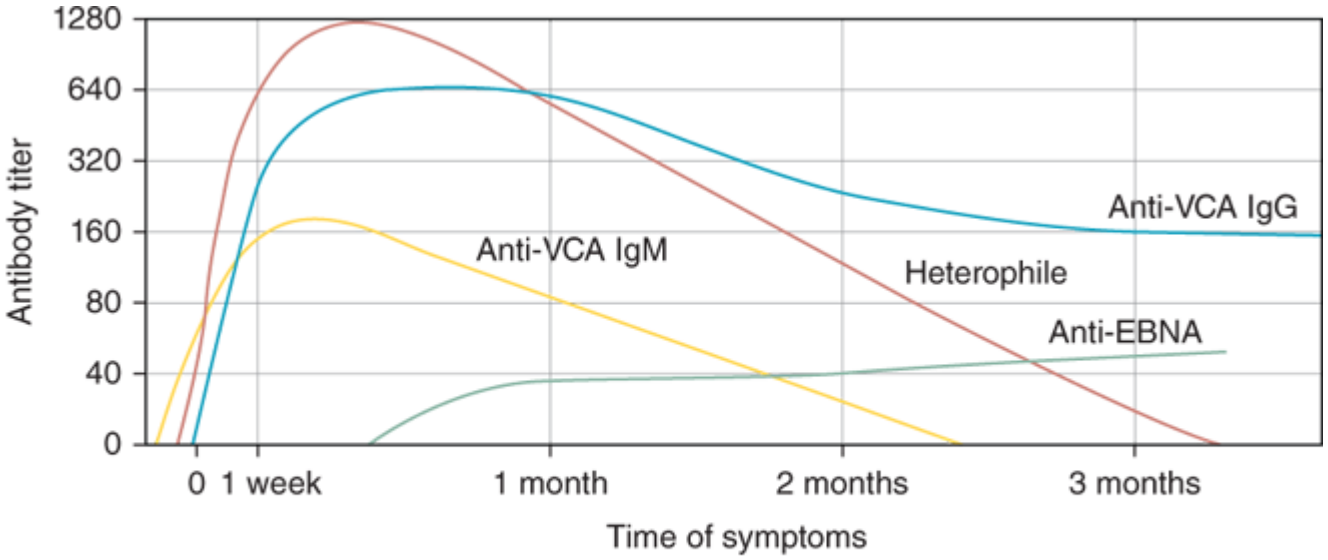
DIAGNOSIS

Serologic Testing

The heterophile test is used for the diagnosis of IM in children and adults (Fig. 189-4). In the test for this antibody, human serum is absorbed with guinea pig kidney, and the heterophile titer is defined as the greatest serum dilution that agglutinates sheep, horse, or cow erythrocytes. The heterophile antibody does not interact with EBV proteins. A titer of ≥40 is diagnostic of acute EBV infection in a patient who has symptoms compatible with IM and atypical lymphocytes. Tests for heterophile antibodies are positive in 40% of patients with IM during the first week of illness and in 80–90% during the third week. Therefore, repeated testing may be necessary, especially if the initial test is performed early. Tests usually remain positive for 3 months after the onset of illness, but heterophile antibodies can persist for up to 1 year. These antibodies usually are not detectable in children <5 years of age, in the elderly, or in patients presenting with symptoms not typical of IM. The commercially available monospot test for heterophile antibodies is somewhat more sensitive than the classic heterophile test. The monospot test is ~75% sensitive and ~90% specific compared with EBV-specific serologies (see below). False-positive monospot results are more common among persons with connective tissue disease, lymphoma, viral hepatitis, and malaria.

FIGURE 189-4

Pattern of Epstein-Barr virus (EBV) serology during acute infection. EBNA, Epstein-Barr nuclear antigen; VCA, viral capsid antigen. *(From JI Cohen, in NS Young et al [eds]: Clinical Hematology. Philadelphia, Mosby, 2006.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EBV-specific antibody testing is used for patients with suspected acute EBV infection who lack heterophile antibodies and for patients with atypical infections. Titers of IgM and IgG antibodies to viral capsid antigen (VCA) are elevated in the serum of more than 90% of patients at the onset of disease. IgM antibody to VCA is most useful for the diagnosis of acute IM because it is present at elevated titers only during the first 2–3 months of the disease; in contrast, IgG antibody to VCA usually is not useful for diagnosis of IM but often is used to assess past exposure to EBV because it persists for life. Seroconversion to EBNA positivity also is useful for the diagnosis of acute infection with EBV. Antibodies to EBNA become detectable relatively late (3–6 weeks after the onset of symptoms) in nearly all cases of acute EBV infection and persist for the lifetime of the patient. These antibodies may be lacking in immunodeficient patients and in those with chronic active EBV infection.

Titers of other antibodies also may be elevated in IM; however, these elevations are less useful for diagnosis. Antibodies to early antigens are detectable 3–4 weeks after the onset of symptoms in patients with IM. About 70% of individuals with IM have antibodies to early antigen diffuse (EA-D) during the illness; the presence of EA-D antibodies is especially likely in patients with relatively severe disease. These antibodies usually persist for only 3–6 months. Levels of EA-D antibodies are also elevated in patients with nasopharyngeal carcinoma or chronic active EBV infection. Antibodies to early antigen restricted (EA-R) are only occasionally detected in patients with IM but are often found at elevated titers in patients with African Burkitt's lymphoma or chronic active EBV infection. IgA antibodies to EBV antigens have proved useful for the identification of patients with nasopharyngeal carcinoma and of persons at high risk for the disease.

Other Studies

Detection of EBV DNA, RNA, or proteins has been valuable in demonstrating the association of the virus with various malignancies. The polymerase chain reaction has been used to detect EBV DNA in the cerebrospinal fluid of some AIDS patients with CNS lymphomas and to monitor the amount of EBV DNA in the blood of patients with lymphoproliferative disease. Detection of high levels of EBV DNA in blood for a few days to

several weeks after the onset of IM may be useful if serologic studies yield equivocal results. Culture of EBV from throat washings or blood is not helpful in the diagnosis of acute infection, since EBV persists in the oropharynx and in B cells for the lifetime of the infected individual.

Differential Diagnosis

Whereas ~90% of cases of IM are due to EBV, 5–10% of cases are due to cytomegalovirus (CMV) **(Chap. 190)**. CMV is the most common cause of heterophile-negative mononucleosis; less common causes of IM and differences from IM due to EBV are shown in **Table 189-2**.

TABLE 189-2

Differential Diagnosis of Infectious Mononucleosis

	Sign or Symptom				
Etiology	**Fever**	**Adenopathy**	**Sore Throat**	**Atypical Lymphocytes**	**Differences from EBV Mononucleosis**
EBV infection	+	+	+	+	—
CMV infection	+	±	±	+	Older age at presentation, longer duration of fever
HIV infection	+	+	+	±	Diffuse rash, oral/genital ulcers, aseptic meningitis
Toxoplasmosis	+	+	±	±	Less splenomegaly, exposure to cats or raw meat
HHV-6 infection	+	+	+	+	Older age at presentation
Streptococcal pharyngitis	+	+	+	–	No splenomegaly, less fatigue
Viral hepatitis	+	±	–	±	Higher aminotransferase levels
Rubella	+	+	±	±	Maculopapular rash, no splenomegaly
Lymphoma	+	+	+	+	Fixed, nontender lymph nodes
Drugs[a]	+	+	–	±	Occurs at any age

[a]Most commonly phenytoin, carbamazepine, sulfonamides, or minocycline.

Abbreviations: CMV, cytomegalovirus; EBV, Epstein-Barr virus; HHV, human herpesvirus.

Treatment

TREATMENT

Therapy for IM consists of supportive measures, with rest and analgesia. Excessive physical activity during the first month should be avoided to reduce the possibility of splenic rupture, which often necessitates splenectomy. Glucocorticoid therapy is not indicated for uncomplicated IM and in fact may predispose to bacterial superinfection. Prednisone (40–60 mg/d for 2–3 days, with subsequent tapering of the dose over 1–2 weeks) has been used for the prevention of airway obstruction in patients with severe tonsillar hypertrophy, for autoimmune hemolytic anemia, for hemophagocytic lymphohistiocytosis, and for severe thrombocytopenia. Glucocorticoids have also been administered to rare patients with severe malaise and fever and to patients with severe CNS or cardiac disease.

Acyclovir has had no significant clinical impact on IM in controlled trials. In one study, the combination of acyclovir and prednisolone had no significant effect on the duration of symptoms of IM.

Acyclovir, at a dosage of 400–800 mg five times daily, has been effective for the treatment of oral hairy leukoplakia (despite common relapses). Post-transplantation EBV lymphoproliferative disease **(Chap. 138)** generally does not respond to antiviral therapy. When possible, therapy should be directed toward reduction of immunosuppression. Antibody to CD20 (rituximab) has been effective in some cases. Infusions of donor lymphocytes are often effective for stem cell transplant recipients, although graft-versus-host disease can occur. Infusions of EBV-specific cytotoxic T cells have been used to prevent EBV lymphoproliferative disease in high-risk settings as well as to treat the disease. Interferon α administration, cytotoxic chemotherapy, and radiation therapy (especially for CNS lesions) also have been used. Infusion of autologous EBV-specific cytotoxic T lymphocytes has shown promise in small studies of patients with nasopharyngeal carcinoma and Hodgkin's disease. Treatment of several cases of X-linked lymphoproliferative disease with antibody to CD20 resulted in a successful outcome of what otherwise would probably have been fatal acute EBV infection.

PREVENTION

The isolation of patients with IM is unnecessary. A vaccine directed against the major EBV glycoprotein reduced the frequency of IM but did not affect the rate of asymptomatic infection in a phase 2 trial. Additional vaccines are under development.

FURTHER READING

Cohen JI: Primary immunodeficiencies associated with EBV disease. Curr Top Microbiol Immunol 390:241, 2015.

Dharnidharka VR et al: Post-transplant lymphoproliferative disorders. Nat Rev Dis Primers 2:15088, 2016.

Grimm JM et al: Prospective studies of infectious mononucleosis in university students. Clin Transl Immunol 5:e94, 2016.

CrossRef

Tzannou I et al: Off-the-shelf virus-specific T cells to treat BK virus, human herpesvirus 6, cytomegalovirus, Epstein-Barr virus, and adenovirus infections after allogeneic hematopoietic stem-cell transplantation. J Clin Oncol 35:3547, 2017.

Chapter 190: Cytomegalovirus and Human Herpesvirus Types 6, 7, and 8

Camille Nelson Kotton; Martin S. Hirsch

FIGURE 190-1

CYTOMEGALOVIRUS

DEFINITION

Cytomegalovirus (CMV), which was initially isolated from patients with congenital cytomegalic inclusion disease, is now recognized as an important pathogen in all age groups. In addition to inducing severe birth defects, CMV causes a wide spectrum of disorders in older children and adults, ranging from an asymptomatic subclinical infection to a mononucleosis syndrome in healthy individuals to disseminated disease in immunocompromised patients. Human CMV is one of several related species-specific viruses that cause similar diseases in various animals. All are associated with the production of characteristic enlarged cells—hence the name *cytomegalovirus*.

CMV, a β-herpesvirus, has double-stranded DNA, four species of mRNA, a protein capsid, and a lipoprotein envelope. Like other herpesviruses, CMV demonstrates icosahedral symmetry, replicates in the cell nucleus, and can cause either a lytic and productive or a latent infection. CMV can be distinguished from other herpesviruses by certain biologic properties, such as host range and type of cytopathology. Viral replication is associated with the production of large intranuclear inclusions and smaller cytoplasmic inclusions. CMV appears to replicate in a variety of cell types in vivo; in tissue culture it grows preferentially in fibroblasts. Although there is little evidence that CMV is oncogenic in vivo, it does transform fibroblasts in rare instances, and genomic transforming fragments have been identified.

EPIDEMIOLOGY

CMV has a worldwide distribution. In many regions of the world, nearly all adults are seropositive for CMV, whereas only half of adults in the United States and Canada are seropositive. In regions where the prevalence of CMV antibody is high, immunocompromised adults are more likely to undergo reactivation disease rather than primary infection. Data generated in specific regions should be considered in the context of local seropositivity rates, when appropriate.

Of newborns in the United States, ~1% are infected with CMV; the percentages are higher in less developed regions. Communal living and poor personal hygiene facilitate spread. Perinatal and early childhood infections are common. CMV may be present in breast milk, saliva, feces, and urine. Transmission has occurred among young children in day-care centers and has been traced from infected toddler to pregnant mother to developing fetus. When an infected child introduces CMV into a household, 50% of susceptible family members seroconvert within 6 months.

CMV is not readily spread by casual contact but rather requires repeated or prolonged intimate exposure for transmission. In late adolescence and young adulthood, CMV is often transmitted sexually, and asymptomatic carriage in semen or cervical secretions is common. Antibody to CMV is present at detectable levels in a high proportion of sexually active men and women, who may harbor several strains simultaneously. Transfusion of blood products containing viable leukocytes may transmit CMV, with a frequency of 0.14–10% per unit transfused. Transfusion of leukocyte-reduced or CMV-seronegative blood significantly decreases the risk of CMV transmission.

Once infected, an individual generally carries CMV for life. The infection usually remains silent. CMV reactivation syndromes develop more frequently, however, when T lymphocyte–mediated immunity is compromised—for example, after organ transplantation, with lymphoid neoplasms and certain acquired immunodeficiencies (in particular, HIV infection; **Chap. 197**), or during critical illness in intensive care units. Most primary CMV infections in organ transplant recipients **(Chap. 138)** result from transmission via the graft. In CMV-seropositive transplant recipients, infection results from reactivation of latent virus or from infection by a new strain from the donor. CMV infection may also be associated with diseases as diverse as coronary artery stenosis and malignant gliomas, although these associations require further validation.

PATHOGENESIS

Congenital CMV infection can result from either primary or reactivation infection of the mother. However, clinical disease in the fetus or newborn is related largely to primary maternal infection **(Table 190-1)**. The major factors determining the severity of congenital infection are unclear, although a deficient capacity to produce precipitating antibodies and to mount T cell responses to CMV is associated with relatively severe disease.

TABLE 190-1

Cmv Disease in the Immunocompromised Host

Population	Risk Factors	Principal Syndrome(s)	Treatment	Prevention
Fetus	Primary maternal infection/early pregnancy	Cytomegalic inclusion disease	Ganciclovir followed by valganciclovir for symptomatic neonates	Avoidance of exposure; possible maternal treatment with CMV immunoglobulin during pregnancy
Organ transplant recipient	Seropositivity of donor and/or recipient; potent immunosuppressive regimen; treatment of rejection	Febrile leukopenia; gastrointestinal disease; pneumonia	Ganciclovir or valganciclovir	Prophylaxis with ganciclovir or valganciclovir or preemptive therapy
Hematopoietic stem cell transplant recipient	Graft-vs-host disease; older age of recipient; seropositive recipient; viremia	Pneumonia; gastrointestinal disease	Ganciclovir or valganciclovir or foscarnet, ± CMV immunoglobulin	Prophylaxis with letermovir, ganciclovir, or valganciclovir or preemptive therapy
Person with HIV	<50 CD4+ T cells/μL; CMV seropositivity	Retinitis; gastrointestinal disease; neurologic disease	Ganciclovir, valganciclovir, foscarnet, or cidofovir	Oral valganciclovir

Primary infection with CMV in late childhood or adulthood is often associated with a vigorous T lymphocyte response that may contribute to the development of a mononucleosis syndrome similar to the sequelae of infection with Epstein-Barr virus **(Chap. 189)**. The hallmark of such infection is the appearance of atypical lymphocytes in the peripheral blood; these cells are predominantly activated CD8+ T lymphocytes. Polyclonal activation of B cells by CMV contributes to the development of rheumatoid factors and other autoantibodies during mononucleosis.

Once acquired, CMV persists indefinitely in host tissues. The sites of persistent infection may include multiple cell types and various organs. Transmission via blood transfusion or organ transplantation is due primarily to

silent infections in these tissues. If the host's T cell responses become compromised by disease or by iatrogenic immunosuppression, latent virus can reactivate to cause a variety of syndromes. Chronic antigenic stimulation in the presence of immunosuppression (for example, after organ transplantation) appears to be an ideal setting for CMV activation and CMV disease. Certain particularly potent suppressants of T cell immunity (e.g., antithymocyte globulin, alemtuzumab) are associated with a high rate of clinical CMV syndromes. CMV may itself contribute to further T lymphocyte hyporesponsiveness, which often precedes superinfection with other opportunistic pathogens such as bacteria, molds, and *Pneumocystis*.

PATHOLOGY

Cytomegalic cells in vivo (presumed to be infected epithelial cells) are two to four times larger than surrounding cells and often contain an 8- to 10-μm intranuclear inclusion that is eccentrically placed and is surrounded by a clear halo, producing an "owl's eye" appearance. Smaller granular cytoplasmic inclusions are demonstrated occasionally. Cytomegalic cells are found in a wide variety of organs, including the salivary gland, lung, liver, kidney, intestine, pancreas, adrenal gland, and central nervous system.

The cellular inflammatory response to infection consists of plasma cells, lymphocytes, and monocyte-macrophages. Granulomatous reactions occasionally develop, particularly in the liver. Immunopathologic reactions may contribute to CMV disease. Immune complexes have been detected in infected infants, sometimes in association with CMV-related glomerulopathies. Immune-complex glomerulopathy has also been observed in some CMV-infected patients after renal transplantation.

CLINICAL MANIFESTATIONS

Congenital CMV Infection

Fetal infections range from subclinical to severe and disseminated. CMV seroconversion rates during pregnancy range from 1% to 7%. Of infants born to mothers with primary CMV infections during pregnancy, 5–20% will develop clinical manifestations, with a mortality rate of ~5%. Petechiae, hepatosplenomegaly, and jaundice are the most common presenting features (60–80% of cases). Microcephaly with or without cerebral calcifications, intrauterine growth retardation, and prematurity are reported in 30–50% of cases. Inguinal hernias and chorioretinitis are less common. Laboratory abnormalities include elevated alanine aminotransferase levels in serum, thrombocytopenia, conjugated hyperbilirubinemia, hemolysis, and elevated protein levels in cerebrospinal fluid. The prognosis for severely infected infants is poor, and few survivors escape intellectual or hearing difficulties later in childhood. The differential diagnosis of cytomegalic inclusion disease in infants includes syphilis, toxoplasmosis, bacterial sepsis, and infection with a variety of viruses, including rubella, Zika, or herpes simplex virus.

Most congenital CMV infections are clinically inapparent at birth. Of asymptomatically infected infants, 5–25% develop significant psychomotor, hearing, ocular, or dental abnormalities over the next several years.

Perinatal CMV Infection

The newborn may acquire CMV at delivery by passage through an infected birth canal or by postnatal contact with infected breast milk or other maternal secretions. Of infants who are breast-fed for >1 month by seropositive mothers, 40–60% become infected. Iatrogenic transmission can result from blood transfusion; use of leukocyte-reduced or CMV-seronegative blood products for transfusion into low-birth-weight seronegative infants or seronegative pregnant women decreases risk.

The great majority of infants infected at or after delivery remain asymptomatic. However, protracted interstitial pneumonitis has been associated with perinatally acquired CMV infection, particularly in premature infants, and occasionally has been accompanied by infection with *Chlamydia trachomatis*, *Pneumocystis*, or *Ureaplasma urealyticum*. Poor weight gain, adenopathy, rash, hepatitis, anemia, and atypical lymphocytosis may also be found, and CMV excretion often persists for months or years.

CMV Mononucleosis

The most common clinical manifestation of CMV infection in immunocompetent hosts beyond the neonatal period is a heterophile antibody–negative mononucleosis syndrome, which may develop spontaneously or follow transfusion of leukocyte-containing blood products. Although the syndrome occurs at all ages, it most often involves sexually active young adults. With incubation periods of 20–60 days, the illness generally lasts for 2–6 weeks. Prolonged high fevers, sometimes with chills, profound fatigue, and malaise, characterize this disorder. Myalgias, headache, and splenomegaly are common, but in CMV mononucleosis (as opposed to Epstein-Barr virus mononucleosis), exudative pharyngitis and cervical lymphadenopathy are rare. Occasional patients develop rubelliform rashes, often after exposure to ampicillin or certain other antibiotics. Less common are interstitial or segmental pneumonia, myocarditis, pleuritis, arthritis, and encephalitis. In rare cases, Guillain-Barré syndrome complicates CMV mononucleosis. The characteristic laboratory abnormality of CMV mononucleosis is relative lymphocytosis in peripheral blood, with >10% atypical lymphocytes. Total leukocyte counts may be low, normal, or markedly elevated. Although significant jaundice is uncommon, serum aminotransferase and alkaline phosphatase levels are often moderately elevated. Heterophile antibodies are absent; however, transient immunologic abnormalities are common and may include the presence of cryoglobulins, rheumatoid factors, cold agglutinins, and antinuclear antibodies. Hemolytic anemia, thrombocytopenia, and granulocytopenia complicate recovery in rare instances.

Most patients recover without sequelae, although postviral asthenia may persist for months. The excretion of CMV in urine, genital secretions, and/or saliva often continues for months or years. Rarely, CMV infection is fatal in immunocompetent hosts; survivors can have recurrent episodes of fever and malaise, sometimes associated with autonomic nervous system dysfunction (e.g., attacks of sweating or flushing).

CMV Infection in the Immunocompromised Host

(Table 190-1) CMV is the most common viral pathogen complicating organ transplantation **(Chap. 138)**. In recipients of kidney, heart, lung, liver, pancreas, and vascularized composite (hand, face, other) transplants, CMV infection may result in a variety of clinical manifestations, including fever and leukopenia, hepatitis,

colitis, pneumonitis, esophagitis, gastritis, and retinitis. CMV disease is an independent risk factor for both graft loss and death. Without prophylaxis, the period of maximal risk is between 1 and 4 months after transplantation. Disease likelihood and viral replication levels generally are greater after primary infection than after reactivation. Molecular studies indicate that seropositive organ transplant recipients are susceptible to infection with donor-derived, genotypically variant CMV. Reactivation infection, although common, is less likely than primary infection to be clinically significant. The overall risk of clinical disease is related to various factors, such as serologic mismatch (donor seropositive, recipient seronegative), degree of immunosuppression, use of antilymphocyte antibodies, lack of anti-CMV prophylaxis, and co-infection with other pathogens. The transplanted organ is particularly vulnerable as a target for CMV infection; thus there is a tendency for CMV hepatitis to follow liver transplantation and for CMV pneumonitis to follow lung transplantation.

CMV viremia occurs in roughly one-third of hematopoietic stem cell transplant recipients; the risk of severe disease may be reduced by prophylaxis or preemptive therapy with antiviral drugs. The risk is greatest in the first 100 days after transplantation, and identified risk factors include certain types of immunosuppressive therapy, an allogeneic (rather than an autologous) graft, acute graft-versus-host disease, older age, and pretransplantation recipient seropositivity.

CMV is an important pathogen in patients with advanced HIV infection **(Chap. 197)**, in whom it may cause retinitis or disseminated disease, particularly when peripheral-blood CD4+ T cell counts fall below 50/μL. As treatment for underlying HIV infection has improved, the incidence of serious CMV infections (e.g., retinitis) has decreased. During the first few weeks after institution of highly active antiretroviral therapy, however, acute flare-ups of CMV retinitis may occur secondary to an immune reconstitution inflammatory syndrome.

Syndromes produced by CMV in immunocompromised hosts ("CMV syndrome") often begin with fatigue, fever, malaise, anorexia, night sweats, and arthralgias or myalgias. Liver function abnormalities, leukopenia, thrombocytopenia, and atypical lymphocytosis may be observed during these episodes. Without treatment, CMV infection may progress to more severe end-organ disease. The development of tachypnea, hypoxemia, and nonproductive cough signals respiratory involvement. Radiologic examination of the lung often shows bilateral interstitial or reticulonodular infiltrates that begin in the periphery of the lower lobes and spread centrally and superiorly; localized segmental, nodular, or alveolar patterns are less common. The differential diagnosis includes *Pneumocystis* infection; other viral, bacterial, or fungal infections; pulmonary hemorrhage; and injury secondary to irradiation or to treatment with cytotoxic drugs.

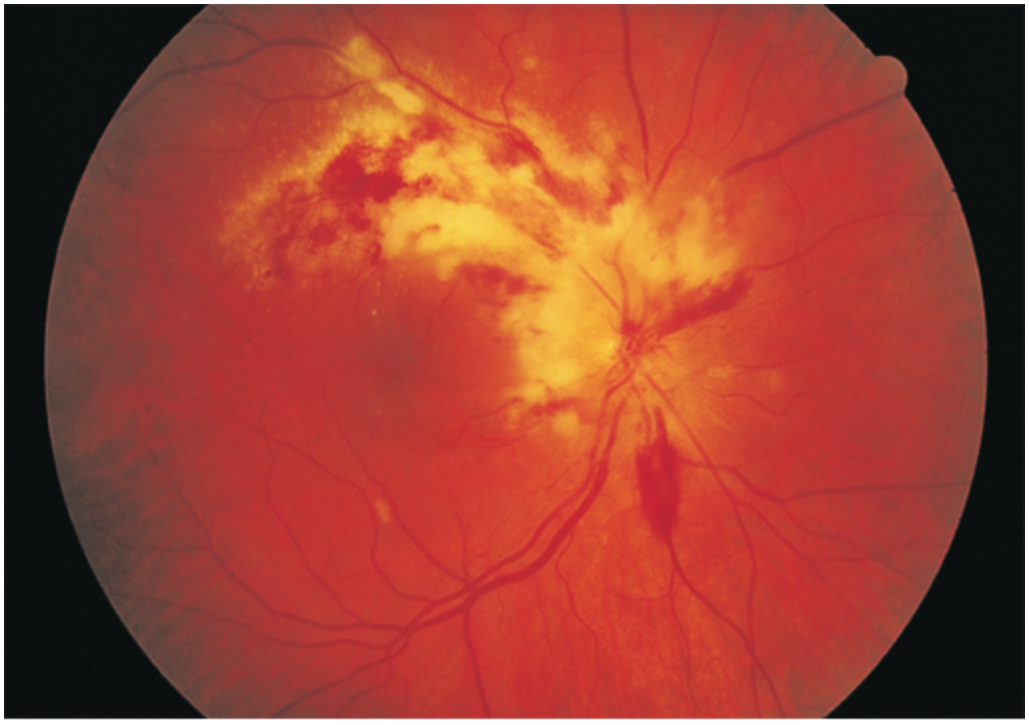
Gastrointestinal CMV involvement may be localized or extensive and almost exclusively affects immunocompromised hosts. Colitis is the most common clinical manifestation in organ transplant recipients. Ulcers of the esophagus, stomach, small intestine, or colon may result in bleeding or perforation. Clinicians should be aware that blood tests such as CMV antigenemia and viral load testing may yield negative results in the setting of intestinal disease. CMV infection may lead to exacerbations of underlying ulcerative colitis. Hepatitis occurs frequently, particularly after liver transplantation. Acalculous cholecystitis and adrenalitis also have been described.

CMV rarely causes meningoencephalitis in otherwise healthy individuals. Two forms of CMV encephalitis are seen in people with HIV. One resembles HIV encephalitis and presents as progressive dementia; the other is a ventriculoencephalitis characterized by cranial-nerve deficits, nystagmus, disorientation, lethargy, and ventriculomegaly. In immunocompromised patients, CMV can also cause subacute progressive polyradiculopathy, which is often reversible if recognized and treated promptly.

CMV retinitis is an important cause of blindness in immunocompromised patients, particularly patients with advanced AIDS **(Chap. 197)**. Early lesions consist of small, opaque, white areas of granular retinal necrosis that spread in a centrifugal manner and are later accompanied by hemorrhages, vessel sheathing, and retinal edema **(Fig. 190-1)**. CMV retinopathy must be distinguished from that due to other conditions, including toxoplasmosis, candidiasis, and herpes simplex virus infection.

FIGURE 190-1

Cytomegalovirus infection in a patient with AIDS may appear as an arcuate zone of retinitis with hemorrhages and optic disk swelling. Often CMV is confined to the retinal periphery, beyond view of the direct ophthalmoscope.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Fatal CMV infections are often associated with persistent viremia and the involvement of multiple organ systems. Progressive pulmonary infiltrates, pancytopenia, hyperamylasemia, and hypotension are characteristic features that are frequently found in conjunction with a terminal bacterial, fungal, or protozoan superinfection. Extensive adrenal necrosis with CMV inclusions is often documented at autopsy, as is CMV involvement of many other organs.

DIAGNOSIS

CMV infection usually cannot be diagnosed reliably on clinical grounds alone. Isolation of CMV or detection of its antigens or DNA in appropriate clinical specimens is the preferred approach. The most common method of detection is quantitative nucleic acid testing (QNAT) for CMV by polymerase chain reaction (PCR) technology, for which blood or other specimens can be used; some centers use a CMV antigenemia test, an immunofluorescence assay that detects CMV antigens (pp65) in peripheral-blood leukocytes. Such assays may yield a positive result several days earlier than culture methods. QNAT may predict the risk for disease progression, particularly in immunocompromised hosts. CMV DNA in cerebrospinal fluid is useful in the diagnosis of CMV encephalitis or polyradiculopathy. Recent introduction of an international testing standard has helped reduce variation in viral load test results.

Virus excretion and/or viremia is readily detected by culture of appropriate specimens on human fibroblast monolayers. If CMV titers are high, as is common in congenital disseminated infection and in AIDS, characteristic cytopathic effects may be detected within a few days. However, in some situations (e.g., CMV mononucleosis), viral titers are low, and cytopathic effects may take several weeks to appear. Many laboratories expedite diagnosis with an overnight tissue-culture method (shell vial assay) involving centrifugation and an immunocytochemical detection technique employing monoclonal antibodies to an immediate-early CMV antigen. Isolation of virus from urine, stool, or saliva does not, by itself, constitute proof of acute infection, since excretion from these sites may continue for months or years after illness. Detection of viremia by QNAT or antigenemia testing is a better predictor of acute infection.

A variety of serologic assays detect antibody to CMV. An increased level of IgG antibody to CMV may not be detectable for up to 4 weeks after primary infection. Detection of CMV-specific IgM is sometimes useful in the diagnosis of recent or active infection; however, circulating rheumatoid factors may result in occasional false-positive IgM tests. Serology is more helpful when used to predict risk of CMV infection and disease in transplant recipients rather than to diagnose acute disease.

PREVENTION

Prevention of CMV infection and disease in organ and hematopoietic stem cell transplant recipients is usually based on one of two methods: universal prophylaxis or preemptive therapy. With universal prophylaxis, antiviral drugs are used for a defined period, often 3 or 6 months. One clinical trial demonstrated that, in CMV-seronegative kidney transplant recipients with seropositive donors, prophylaxis with (val)ganciclovir was more effective at prevention when given for 200 days rather than 100 days. With preemptive therapy, patients are monitored weekly for CMV viremia, and antiviral treatment is initiated once viremia is detected. Because of the bone marrow–suppressive effects of universal prophylaxis, preemptive therapy has been more commonly employed in hematopoietic stem cell transplant recipients; letermovir, which has recently been approved, allows prophylaxis in higher-risk patients. For patients with HIV infection, CMV end-organ disease is best prevented by using antiretroviral therapy sufficient to maintain CD4+ T cell counts above 100/μL. Primary prophylaxis with ganciclovir or valganciclovir is not recommended.

Several additional measures are useful for the prevention of CMV transmission to CMV-naïve, high-risk patients. The use of CMV-seronegative or leukocyte-depleted blood significantly decreases the rate of

transfusion-associated transmission. In a placebo-controlled trial, a CMV glycoprotein B vaccine reduced infection rates among 464 CMV-seronegative women; this outcome raises the possibility that this experimental vaccine will reduce rates of congenital infection, but further studies must validate this approach. A CMV glycoprotein B vaccine with MF59 adjuvant appeared effective in reducing the risk and duration of viremia in both seropositive and seronegative renal transplant recipients at risk for CMV infection. CMV immune globulin has been studied in a variety of clinical situations (primary CMV infection in pregnancy, hematopoietic stem cell transplantation, solid organ transplantation), with conflicting results.

Prophylactic acyclovir or valacyclovir at high doses may reduce rates of CMV infection and disease in renal transplant recipients; neither drug is effective in the treatment of active CMV disease.

TREATMENT

TREATMENT

Cytomegalovirus Infection

Ganciclovir is a guanosine derivative that has considerably more activity against CMV than its congener acyclovir. After intracellular conversion by a viral phosphotransferase encoded by CMV gene region UL97, ganciclovir triphosphate is a selective inhibitor of CMV DNA polymerase. Several clinical studies have indicated response rates of 70–90% among people with HIV who are given ganciclovir for the treatment of CMV retinitis or colitis. In severe infections (e.g., CMV pneumonia in hematopoietic stem cell transplant recipients), ganciclovir is sometimes combined with CMV immune globulin. Prophylactic or suppressive ganciclovir may be useful in high-risk hematopoietic stem cell or organ transplant recipients (e.g., those who are CMV-seropositive before transplantation). In many people with HIV, persistently low CD4+ T cell counts, and CMV disease, clinical and virologic relapses occur promptly if treatment with ganciclovir is discontinued. Therefore, prolonged maintenance regimens are recommended for such patients. Resistance to ganciclovir is more common among patients treated for >3 months and is usually related to mutations in the CMV UL97 gene (or, less commonly, the UL54 gene). The advent of CMV genotyping for resistance mutations has made it possible to rapidly obtain information regarding optimal treatment approaches against clinically resistant virus.

Valganciclovir is an orally bioavailable prodrug that is rapidly metabolized to ganciclovir in intestinal tissues and the liver. Approximately 60–70% of an oral dose of valganciclovir is absorbed. An oral valganciclovir dose of 900 mg results in ganciclovir blood levels similar to those obtained with an IV ganciclovir dose of 5 mg/kg. Valganciclovir appears to be as effective as IV ganciclovir for both CMV induction (treatment) and maintenance regimens, also offering the advantage of oral dosing. Furthermore, the adverse-event profiles and rates of resistance for the two drugs are similar.

Ganciclovir or valganciclovir therapy for CMV disease consists of a 14- to 21-day induction course (5 mg/kg IV twice daily for ganciclovir or 900 mg PO twice daily for valganciclovir), sometimes followed by maintenance therapy (e.g., valganciclovir, 900 mg/d). Peripheral-blood neutropenia develops in roughly one-quarter of treated patients but may be ameliorated by granulocyte colony-stimulating factor or granulocyte-

macrophage colony-stimulating factor. Whether to use maintenance therapy should depend on the overall level of immunocompromise and the risk of recurrent disease. Discontinuation of maintenance therapy should be considered in people with HIV who, while receiving antiretroviral therapy, have a sustained (3- to 6-month) increase in CD4+ T cell counts to >100/μL. Compared with shorter (6-week) courses, prolonged (6-month) courses of valganciclovir had beneficial effects on hearing and developmental outcomes in infants with congenital CMV infection.

For treatment of CMV retinitis, some clinicians prefer intravitreal injections of ganciclovir or foscarnet (see below) plus oral valganciclovir to intravenous ganciclovir, although no clinical trials have compared these approaches. *Foscarnet* (sodium phosphonoformate) inhibits CMV DNA polymerase. Because this agent does not require phosphorylation to be active, it is also effective against most ganciclovir-resistant isolates. Foscarnet is less well tolerated than ganciclovir and causes considerable toxicity, including renal dysfunction, hypomagnesemia, hypokalemia, hypocalcemia, genital ulcers, dysuria, nausea, and paresthesia. Moreover, foscarnet administration requires the use of an infusion pump and close clinical monitoring. With aggressive hydration and dose adjustments for renal dysfunction, the toxicity of foscarnet can be reduced. The use of foscarnet should be avoided when a saline load cannot be tolerated (e.g., in cardiomyopathy). The approved induction regimen is 60 mg/kg every 8 h for 2 weeks, although 90 mg/kg every 12 h is equally effective and no more toxic. Maintenance infusions should deliver 90–120 mg/kg once daily. No oral preparation is available. Foscarnet-resistant virus may emerge during extended therapy. This drug is used more frequently after hematopoietic stem cell transplantation than in other situations to avoid the myelosuppressive effects of ganciclovir; in general, foscarnet is also the first choice for infections with ganciclovir-resistant CMV.

Cidofovir is a nucleotide analogue with a long intracellular half-life that allows intermittent IV administration. Induction regimens of 5 mg/kg weekly for 2 weeks are followed by maintenance regimens of 3–5 mg/kg every 2 weeks. Cidofovir can cause severe nephrotoxicity through dose-dependent proximal tubular cell injury; however, this adverse effect can be tempered somewhat by saline hydration and probenecid. Cidofovir is used primarily for ganciclovir-resistant virus.

HUMAN HERPESVIRUS (HHV) TYPES 6, 7, AND 8

HHV-6 AND HHV-7

 HHV-6 and -7 seropositivity rates are generally high throughout the world. HHV-6 was first isolated in 1986 from peripheral-blood leukocytes of six persons with various lymphoproliferative disorders. Two genetically distinct variants (HHV-6A and HHV-6B) are now recognized. HHV-6 appears to be transmitted by saliva and possibly by genital secretions.

Infection with HHV-6 frequently occurs during infancy as maternal antibody wanes. The peak age of acquisition is 9–21 months; by 24 months, seropositivity rates approach 80%. Older siblings appear to serve as a source of transmission. In addition, congenital infection may occur, and ~1% of newborns are infected

with HHV-6; placental infection with HHV-6 has been described. Congenital infection is generally asymptomatic, although subtle neurologic defects have been described. Most postnatally infected children develop symptoms (fever, fussiness, and diarrhea). A minority develop exanthem subitum (roseola infantum; **see Fig. A1-5**), a common illness characterized by fever with subsequent rash. In addition, ~10–20% of febrile seizures without rash during infancy are caused by HHV-6. After initial infection, HHV-6 persists in peripheral-blood mononuclear cells as well as in the central nervous system, salivary glands, and female genital tract.

In older age groups, HHV-6 has been associated with mononucleosis syndromes; in immunocompromised hosts, encephalitis, pneumonitis, syncytial giant-cell hepatitis, and disseminated disease are seen. In transplant recipients, HHV-6 infection may also be associated with graft dysfunction. Acute HHV-6-associated limbic encephalitis has been reported in hematopoietic stem cell transplant recipients and is characterized by memory loss, confusion, seizures, hyponatremia, and abnormal electroencephalographic and MRI results. High plasma loads of HHV-6 DNA in hematopoietic stem cell transplant recipients are associated with allelic-mismatched donors, use of glucocorticoids, delayed monocyte and platelet engraftment, development of limbic encephalitis, and increased all-cause mortality rates. Mesial temporal lobe epilepsy has been associated with HHV-6 infections, and, like many other viruses, HHV-6 has been implicated in the pathogenesis of multiple sclerosis, although further study is needed to distinguish between association and etiology.

HHV-7 was isolated in 1990 from T lymphocytes from the peripheral blood of a healthy 26-year-old man. The virus is frequently acquired during childhood, albeit at a later age than HHV-6. HHV-7 is commonly present in saliva, which is presumed to be the principal source of infection; breast milk and cervical secretions may also carry the virus. Viremia can be associated with either primary or reactivation infection. The most common clinical manifestations of childhood HHV-7 infections are fever and seizures. Some children present with respiratory or gastrointestinal signs and symptoms. An association has been made between HHV-7 and pityriasis rosea, but evidence is insufficient to indicate a causal relationship.

Clustering of HHV-6, HHV-7, and CMV infections in transplant recipients can make it difficult to sort out the roles of the various agents in individual clinical syndromes. HHV-6 and HHV-7 appear to be susceptible to ganciclovir and foscarnet, although definitive evidence of clinical response is lacking.

HHV-8

Unique herpesvirus-like DNA sequences were reported during 1994 and 1995 in tissues derived from Kaposi's sarcoma (KS) and body cavity–based lymphoma occurring in people with HIV. The virus from which these sequences were derived is designated HHV-8 or Kaposi's sarcoma–associated herpesvirus (KSHV). HHV-8, which infects B lymphocytes, macrophages, and both endothelial and epithelial cells, appears to be causally related not only to KS and a subgroup of AIDS-related B cell body cavity–based lymphomas (primary effusion lymphomas) but also to multicentric Castleman disease, a lymphoproliferative disorder of B cells. The association of HHV-8 with several other diseases has been reported but not confirmed.


HHV-8 seropositivity occurs worldwide, with areas of high endemicity influencing rates of disease. Unlike other herpesvirus infections, HHV-8 infection is much more common in some geographic areas (e.g., central and southern Africa) than in others (North America, Asia, northern Europe). In high-prevalence areas, infection occurs in childhood, seropositivity is associated with having a seropositive mother or (to a lesser extent) older sibling, and HHV-8 may be transmitted in saliva. In low-prevalence areas, infections typically occur in adults, probably with sexual transmission. Concurrent epidemics of HIV-1 and HHV-8 infections among certain populations (e.g., men who have sex with men) in the late 1970s and early 1980s appear to have resulted in the frequent association of AIDS and KS. Transmission of HHV-8 may also be associated with organ transplantation, injection drug use, and blood transfusion; however, transmission via organ transplantation or blood transfusion in the United States appears to be quite rare.

Primary HHV-8 infection in immunocompetent children may manifest as fever and maculopapular rash. Among individuals with intact immunity, chronic asymptomatic infection is the rule, and neoplastic disorders generally develop only after subsequent immunocompromise. Immunocompromised persons with primary infection may present with fever, splenomegaly, lymphoid hyperplasia, pancytopenia, or rapid-onset KS. Quantitative analysis of HHV-8 DNA suggests a predominance of latently infected cells in KS lesions and frequent lytic replication in multicentric Castleman disease. The KS-associated herpesvirus inflammatory cytokine syndrome (KICS)—consisting of fever and high levels of HHV-8, human and viral interleukin 6, and human interleukin 10—has been described in some HIV-infected patients and is associated with a high mortality rate.

Effective antiretroviral therapy for HIV-infected individuals has led to a marked reduction in rates of KS among persons dually infected with HHV-8 and HIV in resource-rich areas. HHV-8 itself is susceptible in vitro to ganciclovir, foscarnet, and cidofovir. A small, randomized, double-blind, placebo-controlled, crossover trial suggested that oral valganciclovir administered once daily reduced HHV-8 replication. However, clinical benefits of valganciclovir or other drugs in HHV-8 infection have not yet been demonstrated. Sirolimus inhibits the progression of dermal KS in kidney transplant recipients while providing effective immunosuppression.

FURTHER READING

Erard V et al: Reduced mortality of cytomegalovirus pneumonia after hematopoietic cell transplantation due to antiviral therapy and changes in transplantation practices. Clin Infect Dis 61:31, 2015.

Kaminski H, Fishman JA: The cell biology of cytomegalovirus: Implications for transplantation. Am J Transplant 16:2254, 2016.

Kawamura J et al: Pathogenic role of human herpes virus 6B with mesial temporal lobe epilepsy. J Infect Dis 212:1014, 2015.

Kimberlin DW et al: Valganciclovir for symptomatic congenital cytomegalovirus disease. N Engl J Med 372:933, 2015.

Kotton CN et al: Updated international consensus guidelines on the management of cytomegalovirus in solid-organ transplantation. Transplantation 96:331, 2013.

Chapter 191: Molluscum Contagiosum, Monkeypox, and Other Poxvirus Infections

Fred Wang

FIGURE 191-1

INTRODUCTION

The poxvirus family includes a large number of related DNA viruses that infect various vertebrate hosts. The poxviruses responsible for infections in humans, the geographic locations in which these infections are found, the host reservoirs, and the main manifestations are listed in **Table 191-1**. Infections with orthopoxviruses—e.g., smallpox (variola major) virus **(Chap. S2)** or the zoonotic monkeypox virus—can result in systemic, potentially lethal human disease. Other poxvirus infections cause primarily localized skin disease in humans.

TABLE 191-1

Poxviruses and Human Infections

Genus	Species	Geographic Location	Host Reservoir	Human Disease
Orthopoxvirus	Variola[a]	Extinct	Humans	Smallpox, systemic
	Monkeypox	Africa	Rodents	Smallpox-like, systemic
	Cowpox	Europe	Rodents	Local pox lesion, occasionally systemic
	Buffalopox	Indian subcontinent	Water buffalo	Local pox lesion, mild illness
	Cantagalo and Araçatuba	South America	Cattle	Local pox lesion, mild illness
	Vaccinia	—	—	Local pox lesions (smallpox vaccine)
Molluscipoxvirus	Molluscum contagiosum	Worldwide	Humans	Multiple cutaneous lesions
Parapoxvirus	Orf	Worldwide	Sheep, goats	Local pox lesions (contagious pustular dermatitis)
	Pseudocowpox (paravaccinia)	Worldwide	Cattle	Local pox lesions (milker's nodule)
	Bovine papular stomatitis	Worldwide	Cattle	Local pox lesions
	Deerpox	Deer herds	Deer	Local pox lesions
	Sealpox	Seal colonies	Seals	Local pox lesions
Yatapoxvirus	Tanapox	Africa	Monkeys	Local pox lesions

[a]See Chap. S2.

MOLLUSCUM CONTAGIOSUM

Molluscum contagiosum virus is an obligate human pathogen that causes distinctive proliferative skin lesions. These lesions measure 2–5 mm in diameter and are pearly, flesh-colored, and umbilicated, with a characteristic dimple at the center (Fig. 191-1). A relative lack of inflammation and necrosis distinguishes these proliferative lesions from other poxvirus lesions. Lesions may be found—singly or in clusters—anywhere on the body except on the palms and soles and may be associated with an eczematous rash.

FIGURE 191-1

Molluscum contagiosum is a cutaneous poxvirus infection characterized by multiple umbilicated flesh-colored or hypopigmented papules.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Molluscum contagiosum is highly prevalent among children and is the most common human disease resulting from poxvirus infection. Swimming pools are a common vector for transmission. Atopy and compromise of skin integrity increase the risk of infection. Genital lesions are more common in adults, to whom the virus may be transmitted by sexual contact. The incubation period ranges from 2 weeks to 6 months, with an average of 2–7 weeks. In most cases, the disease is self-limited and regresses spontaneously after 3–4 months in immunocompetent hosts. There are no systemic complications, but skin lesions may persist for 3–5 years. Molluscum contagiosum can be associated with immunosuppression and is frequently seen among HIV-infected patients (Chap. 197). The disease can be more generalized, severe, and persistent in AIDS patients than in other groups. Moreover, molluscum contagiosum can be exacerbated in the immune reconstitution inflammatory syndrome (IRIS) associated with the initiation of antiretroviral therapy.

The diagnosis of molluscum contagiosum is typically based on its clinical presentation and can be confirmed by histologic demonstration of the cytoplasmic eosinophilic inclusions (*molluscum bodies*) that are characteristic of poxvirus replication. Molluscum contagiosum virus cannot be propagated in vitro, but electron microscopy and molecular studies can be used for its identification.

There is no specific systemic treatment for molluscum contagiosum, and lesions are likely to resolve spontaneously. Treatment may be more pressing in HIV-positive patients, but initiation of effective antiretroviral therapy may still be the best option. Regardless of the patient's HIV status, strong clinical-trial evidence of superiority for physical ablation or drugs is lacking.

MONKEYPOX

 Although monkeypox virus was named after the animal from which it was originally isolated, rodents are the primary viral reservoir. Human infections with monkeypox virus typically occur in Africa when humans come into direct contact with infected animals. Human-to-human transmission of monkeypox infection has been rare; however, such transmission may increase as cross-reactive immunity from smallpox vaccination wanes. Such an increase was observed during a 2013 monkeypox outbreak in the Democratic Republic of the Congo. Human disease is characterized by a systemic illness and vesicular rash similar to those of variola. The clinical presentation of monkeypox can be confused with that of the more common varicella-zoster virus infection **(Chap. 188)**. Compared with the lesions of this herpesvirus infection, monkeypox lesions tend to be more uniform (i.e., in the same stage of development), diffuse, and peripheral in distribution. Lymphadenopathy is a prominent feature of monkeypox infection.

The first outbreak of human monkeypox infection in the Western Hemisphere occurred during 2003, when more than 70 cases were reported in the midwestern United States. The outbreak was linked to contact with pet prairie dogs that had become infected while being housed with rodents imported from Ghana. Patients presented most frequently with fever, rash, and lymphadenopathy ~14 days after exposure. The risk of human disease from animal orthopoxvirus infections is increasing as cross-reactive smallpox immunity wanes in the general population and exposure to wild animals in the rainforest and to exotic animals as household pets grows.

OTHER ZOONOTIC POXVIRUS INFECTIONS

Cowpox and buffalopox are rare zoonotic infections characterized by cutaneous poxlike lesions and mild systemic illness. Outbreaks of similar poxlike lesions among cattle and farm workers in Brazil have been due to Cantagalo and Araçatuba viruses, which are virtually identical to vaccinia virus and may have become established in cattle during smallpox vaccination programs.

Parapoxviruses are widely scattered among animal species, but only a few are known to cause human disease via direct contact with infected animals. Parapoxviruses are antigenically distinct from

orthopoxviruses and share no cross-immunity. *Tanapox* virus belongs to a separate, antigenically distinct genus and usually causes a single nodular lesion on the exposed area after contact with infected monkeys.

FURTHER READING

Kantele A et al: Emerging diseases—the monkeypox epidemic in the Democratic Republic of the Congo. Clin Microbiol Infect 22:658, 2016.

Martin P: Interventions for molluscum contagiosum in people infected with human immunodeficiency virus: A systematic review. Int J Dermatol 55:956, 2016.

Shisler JL: Immune evasion strategies of molluscum contagiosum virus. Adv Virus Res 92:201, 2015.

Sterling J: Treatment of warts and molluscum: What does the evidence show? Curr Opin Pediatr 28:490, 2016.

Chapter 192: Parvovirus Infections

Kevin E. Brown

INTRODUCTION

Parvoviruses, members of the family Parvoviridae, are small (diameter, ~22 nm), nonenveloped, icosahedral viruses with a linear single-strand DNA genome of ~5000 nucleotides. These viruses are dependent on either rapidly dividing host cells or helper viruses for replication. At least five groups of parvoviruses infect humans: parvovirus B19 (B19V), dependoparvoviruses (adeno-associated viruses; AAVs), human tetraparvoviruses (PARV4 and PARV5), human bocaparvoviruses (HBoVs), and human protoparvoviruses (bufavirus and tusavirus). Human dependoparvoviruses are nonpathogenic and will not be considered further in this chapter.

PARVOVIRUS B19

DEFINITION

B19V is the type member of the genus *Erythroparvovirus*. On the basis of viral sequence, B19V is divided into three genotypes (designated 1, 2, and 3), but only a single B19V antigenic type has been described. Genotype 1 is predominant in most parts of the world; genotype 2 is rarely associated with active infection; and genotype 3 appears to predominate in parts of western Africa.

EPIDEMIOLOGY

B19V exclusively infects humans, and infection is endemic in virtually all parts of the world. Transmission occurs predominantly via the respiratory route and is followed by the onset of rash and arthralgia. By the age of 15 years, ~50% of children have detectable IgG antibody to B19V; this figure rises to >90% among the elderly. In pregnant women, the estimated annual seroconversion rate is ~1%. Within households, secondary infection rates approach 50%.

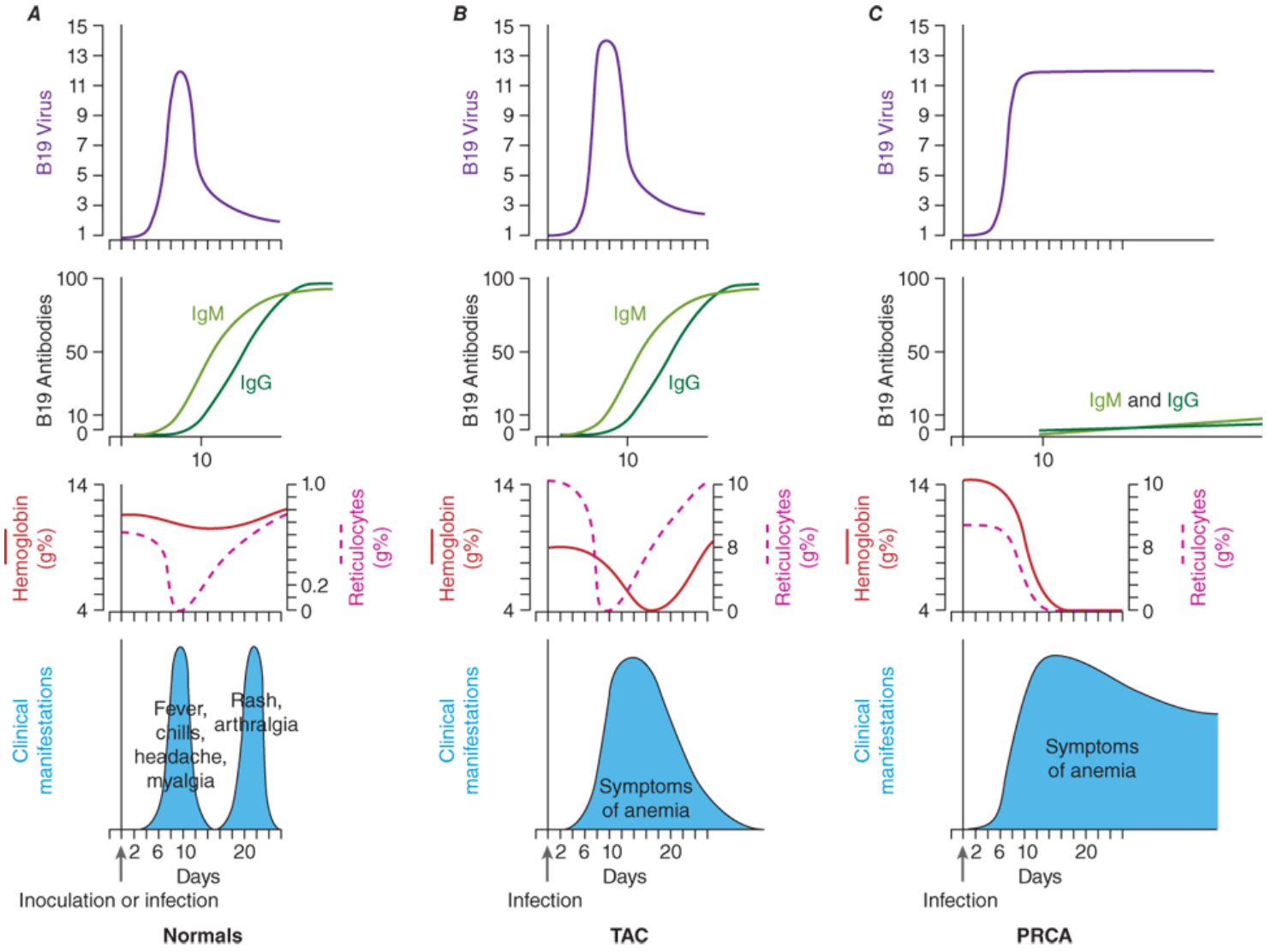
Detection of high-titer B19V in blood is not unusual (see "Pathogenesis," below). Transmission can occur as a result of transfusion, most commonly of pooled components. To reduce the risk of transmission, plasma pools are screened by nucleic acid amplification technology, and high-titer pools are discarded. B19V is resistant to both heat and solvent-detergent inactivation.

PATHOGENESIS

B19V replicates primarily in erythroid progenitors. This specificity is due in part to the limited tissue distribution of the primary B19V receptor, blood group P antigen (globoside). Infection leads to high-titer viremia, with >10^{12} virus particles (or IU)/mL detectable in the blood at the apex **(Fig. 192-1)**, and virus-induced cytotoxicity results in cessation of red cell production. In immunocompetent individuals, viremia and arrest of erythropoiesis are transient and resolve as the IgM and IgG antibody response is mounted. In individuals with normal erythropoiesis, there is only a minimal drop in hemoglobin levels; however, in those with increased erythropoiesis (especially with hemolytic anemia), this cessation of red cell production can induce a transient crisis with severe anemia (Fig. 192-1). Similarly, if an individual (or, after maternal infection, a fetus) does not mount a neutralizing antibody response and halt the lytic infection, erythroid production is compromised and chronic anemia develops (Fig. 192-1).

FIGURE 192-1

Schematic of the time course of parvovirus B19 infection in (***A***) normals (erythema infectiosum), (***B***) transient aplastic crisis (TAC), and (***C***) chronic anemia/pure red-cell aplasia (PRCA). *(Reprinted with permission from NS Young, KE Brown: N Engl J Med 350:586, 2004. © 2004 Massachusetts Medical Society. All rights reserved.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The immune-mediated phase of illness, which begins 2–3 weeks after infection as the IgM response peaks, manifests as the rash of fifth disease together with arthralgia and/or frank arthritis. Low-level B19V DNA can be detected by polymerase chain reaction (PCR) in blood and tissues for months to years after acute infection. The B19V receptor is found in a variety of other cells and tissues, including megakaryocytes, endothelial cells, placenta, myocardium, and liver. Infection of these tissues by B19V may be responsible for some of the more unusual presentations of the infection. Rare individuals who lack P antigen are naturally resistant to B19V infection.

CLINICAL MANIFESTATIONS

Erythema Infectiosum

Most B19V infections are asymptomatic or are associated with only a mild nonspecific illness. The main manifestation of symptomatic B19V infection is erythema infectiosum, also known as *fifth disease* or *slapped-cheek disease* (Figs. 192-2 and A1-1A). Infection begins with a minor febrile prodrome ~7–10 days after exposure, and the classic facial rash develops several days later; after 2–3 days, the erythematous macular rash may spread to the extremities in a lacy reticular pattern. However, its intensity and distribution vary, and B19V-induced rash is difficult to distinguish from other viral exanthems. Adults typically do not exhibit the "slapped-cheek" phenomenon but present with arthralgia, with or without the macular rash.

FIGURE 192-2
Young child with erythema infectiosum, or fifth disease, showing typical "slapped-cheek" appearance.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Polyarthropathy Syndrome

Although uncommon among children, arthropathy occurs in ~50% of adults and is more common among women than among men. The distribution of the affected joints is often symmetrical, with arthralgia affecting the small joints of the hands and occasionally the ankles, knees, and wrists. Resolution usually occurs within a few weeks, but recurring symptoms can continue for months. The illness may mimic rheumatoid arthritis, and rheumatoid factor can often be detected in serum. B19V infection may trigger rheumatoid disease in some patients and has been associated with juvenile idiopathic arthritis.

Transient Aplastic Crisis

Asymptomatic transient reticulocytopenia occurs in most individuals with B19V infection. However, in patients who depend on continual rapid production of red cells, infection can cause transient aplastic crisis (TAC). Affected individuals include those with hemolytic disorders, hemoglobinopathies, red cell enzymopathies, and autoimmune hemolytic anemias. Patients present with symptoms of severe anemia (sometimes life-threatening) and a low reticulocyte count, and bone marrow examination reveals an absence of erythroid precursors and characteristic giant pronormoblasts. As its name indicates, the illness is transient, and anemia resolves with the cessation of cytopathic infection in the erythroid progenitors.

Pure Red-Cell Aplasia/Chronic Anemia

Chronic B19V infection has been reported in a wide range of immunosuppressed patients, including those with congenital immunodeficiency, AIDS **(Chap. 197)**, lymphoproliferative disorders (especially acute lymphocytic leukemia), and transplantation **(Chap. 138)**. Patients have persistent anemia with reticulocytopenia, absent or low levels of B19V IgG, high titers of B19V DNA in serum, and—in many cases—scattered giant pronormoblasts in bone marrow. Rarely, nonerythroid hematologic lineages also are affected. Transient neutropenia, lymphopenia, and thrombocytopenia (including idiopathic thrombocytopenic purpura) have been observed. B19V occasionally causes a hemophagocytic syndrome.

Studies in Papua New Guinea, Gabon, and Ghana, where malaria is endemic, suggest that co-infection with *Plasmodium* and B19V plays a major role in the development of severe anemia in young children. Further studies must determine whether B19V infection contributes to severe anemia in other malarial regions.

Hydrops Fetalis

B19V infection during pregnancy can lead to hydrops fetalis and/or fetal loss. The risk of transplacental fetal infection is ~30%, and the risk of fetal loss (predominantly early in the second trimester) is ~9%. The risk of congenital infection is <1%. Although B19V does not appear to be teratogenic, anecdotal cases of eye damage and central nervous system (CNS) abnormalities have been reported. Cases of congenital anemia have also been described. B19V probably causes 10–20% of all cases of nonimmune hydrops.

Unusual Manifestations

B19V infection may rarely cause hepatitis, vasculitis, myocarditis, glomerulosclerosis, or meningitis. A variety of other cardiac manifestations, CNS diseases, and autoimmune infections have also been reported. However, B19V DNA can be detected by PCR for years in many tissues; this finding is of no known clinical significance, but its interpretation may cause confusion regarding B19V disease association.

DIAGNOSIS

Diagnosis of B19V infection in immunocompetent individuals is generally based on detection of B19V IgM antibodies **(Table 192-1)**. IgM can be detected at the time of rash in erythema infectiosum and by the third day of TAC in patients with hematologic disorders; these antibodies remain detectable for ~3 months. B19V IgG is detectable by the seventh day of illness and persists throughout life. Quantitative detection of B19V DNA should be used for the diagnosis of early TAC or chronic anemia. Although B19V levels fall rapidly with the development of the immune response, DNA can be detectable by PCR for months or even years after infection, even in healthy individuals; therefore, quantitative PCR should be used. In acute infection at the height of viremia, $>10^{12}$ B19V DNA IU/mL of serum can be detected; however, titers fall rapidly within 2 days. Patients with aplastic crisis or B19V-induced chronic anemia generally have $>10^{5}$ B19V DNA IU/mL.

TABLE 192-1

Diseases Associated with Human Parvovirus B19 Infection and Methods of Diagnosis

Disease	Hosts	IgM	IgG	PCR	Quantitative PCR
Fifth disease	Healthy children	Positive	Positive	Positive	$>10^{4}$ IU/mL
Polyarthropathy syndrome	Healthy adults (more often women)	Positive within 3 months of onset	Positive	Positive	$>10^{4}$ IU/mL
Transient aplastic crisis	Patients with increased erythropoiesis	Negative/positive	Negative/positive	Positive	Often $>10^{12}$ IU/mL, but rapidly decreases
Persistent anemia/pure red-cell aplasia	Immunodeficient or immunocompetent patients	Negative/weakly positive	Negative/weakly positive	Positive	Often $>10^{12}$ IU/mL, but should be $>10^{6}$ in the absence of treatment
Hydrops fetalis/congenital anemia	Fetuses (<20 weeks)	Negative/positive	Positive	Positive amniotic fluid or tissue	n/a

Abbreviations: IU, international units (1 IU equals ~1 genome); n/a, not applicable; PCR, polymerase chain reaction.

TREATMENT

TREATMENT

No antiviral drug effective against B19V is available, and treatment of B19V infection often targets symptoms only. TAC precipitated by B19V infection frequently necessitates symptom-based treatment with blood transfusions. In patients receiving chemotherapy, temporary cessation of treatment may result in an immune response and resolution. If this approach is unsuccessful or not applicable, commercial immune globulin (IVIg; Gammagard, Sandoglobulin) from healthy blood donors can cure or ameliorate persistent B19V infection in immunosuppressed patients. Generally, the dose used is 400 mg/kg daily for 5–10 days. Like patients with TAC, immunosuppressed patients with persistent B19V infection should be considered infectious. Administration of IVIg is not beneficial for erythema infectiosum or B19V-associated polyarthropathy. Intrauterine blood transfusion can prevent fetal loss in some cases of fetal hydrops.

PREVENTION

No vaccine has been approved for the prevention of B19V infection, although vaccines based on B19V virus-like particles expressed in insect cells are known to be highly immunogenic. Phase 1 trials of a putative vaccine were discontinued because of adverse side effects.

HUMAN TETRAPARVOVIRUSES (PARV4/5)

DEFINITION

The PARV4 viral sequence was initially detected in a patient with an acute viral syndrome. Similar sequences, including the related PARV5 sequence, have been detected in pooled plasma collections. The DNA sequence of PARV4/5 is distinctly different from that of all other parvoviruses, and this virus is now classified as a member of the newly described genus *Tetraparvovirus.*

EPIDEMIOLOGY

PARV4 DNA is commonly found in plasma pools but at lower concentrations than the levels of B19V DNA found before in plasma pools prior to screening. The higher levels of PARV4 DNA and IgG antibody in tissues (bone marrow and lymphoid tissue) and sera from IV drug users than in the corresponding specimens from control patients suggest that the virus is transmitted predominantly by parenteral means in the United States and Europe. Evidence for nonparenteral transmission in other parts of the world is limited.

CLINICAL MANIFESTATIONS

To date, PARV4/5 infection has been associated only with mild clinical disease (rash and/or transient aminotransferase elevation).

HUMAN BOCAPARVOVIRUSES

DEFINITION

Animal bocaparvoviruses are associated with mild respiratory symptoms and enteritis in young animals. Human bocavirus 1 (HBoV1) was originally identified in the respiratory tract of young children with lower respiratory tract infections. More recently, HBoV1 and the related viruses HBoV2, HBoV3, and HBoV4 have all been identified in human fecal samples.

EPIDEMIOLOGY

Seroepidemiologic studies with HBoV virus-like particles suggest that HBoV infection is common. Worldwide, most individuals are infected before the age of 5 years.

CLINICAL MANIFESTATIONS

HBoV1 DNA is found in respiratory secretions from 2–20% of children with acute respiratory infection, often in the presence of other pathogens; in these circumstances, the role of HBoV1 in disease pathogenesis is unknown. Clinical disease due to HBoV1 is associated with evidence of primary infection (IgG seroconversion or the presence of IgM), HBoV1 DNA in serum, or high-titer HBoV1 DNA (>10^4 genome copies/mL) in respiratory secretions. Symptoms are not dissimilar from those of other viral respiratory infections, and cough and wheezing are commonly reported. There is no specific treatment for HBoV infection. The role of HBoVs in childhood gastroenteritis remains to be established.

HUMAN PROTOPARVOVIRUSES

DEFINITION

Bufavirus and tusavirus were both identified in clinical samples by a metagenomics approach used for identifying new pathogens. These viruses are classified as members of the protoparvovirus group along with the original prototype member of the Parvoviridae, minute virus of mice.

EPIDEMIOLOGY


Little is known about the epidemiology of either virus, but bufavirus DNA has been found in 0.2–4% of stools from children and adults with diarrhea in many countries; tusavirus has been identified in only a single patient with diarrhea in Tunisia.

CLINICAL MANIFESTATIONS

Although bufavirus DNA is found in patients with diarrhea, often it is detected in conjunction with other viruses. The role of bufavirus in childhood gastroenteritis remains to be confirmed.

FURTHER READING

Crabol Y et al: Intravenous immunoglobulin therapy for pure red cell aplasia related to human parvovirus B19 infection: A retrospective study of 10 patients and review of the literature. Clin Infect Dis 56:968, 2013.

Jartti T et al: Human bocavirus—the first 5 years. Rev Med Virol 22:46, 2012.

Maple PA et al: Identification of past and recent parvovirus B19 infection in immunocompetent individuals by quantitative PCR and enzyme immunoassays: A dual-laboratory study. J Clin Microbiol 52:947, 2014.

Matthews PC et al: PARV4: An emerging tetraparvovirus. PLoS Pathog 10:e100403, 2014.
CrossRef

Söderlund-Venermo M et al: Human parvoviruses, in *Clinical Virology*, 4th ed. DD Richman et al (eds). Washington, DC, ASM Press, 2016, pp 679–700.

Chapter 193: Human Papillomavirus Infections

Darron R. Brown; Aaron Ermel

INTRODUCTION

Interest in human papillomavirus (HPV) infection began in earnest in the 1980s after Harold zur Hausen postulated that infection with these viruses was associated with cervical cancer. It is now recognized that HPV infection of the human genital tract is extremely common and causes clinical conditions ranging from asymptomatic infection to genital warts (condylomata acuminata); dysplastic lesions and invasive cancers of the anus, penis, vulva, vagina, and cervix; and a subset of oropharyngeal cancers. This chapter describes the epidemiology of HPV as a virus and a pathogen, the natural history of HPV infections and associated cancers, strategies to prevent infection and HPV-associated disease, and treatment modalities for some conditions caused by HPV.

PATHOGENESIS

Overview



HPV is an icosahedral, nonenveloped, 8000-base-pair, double-stranded DNA virus with a diameter of 55 nm. Like the genomes of other papillomaviruses, HPV's genome consists of an early (E) gene region, a late (L) gene region, and a noncoding region, which contains regulatory elements. The E1, E2, E5, E6, and E7 proteins are expressed early in the growth cycle and are necessary for viral replication and cellular transformation. The E6 and E7 proteins are responsible for malignant transformation, targeting the human cell-cycle regulatory molecules p53 and Rb (retinoblastoma protein) for degradation, respectively. Translation of the L1 and L2 transcripts and splicing of an E1^E4 transcript occur later. The L1 gene encodes the 54-kDa major capsid protein that makes up the majority of the virus shell; the 77-kDa L2 minor protein contributes a smaller percentage of the capsid mass.

More than 125 HPV types have been identified and are numerically designated on the basis of a unique L1 gene sequence. Approximately 40 HPV types are regularly identified in the anogenital tract; these types are subdivided into high-risk and low-risk categories depending on the associated risk of cervical cancer. For example, HPV types 6 and 11 cause genital warts and ~10% of low-grade cervical lesions and are thus designated low risk. HPV types 16 and 18 cause dysplastic lesions and a high percentage of invasive cancers of the cervix and are therefore considered high risk.

HPV targets basal keratinocytes after microtrauma allows exposure of these cells to the virus. The HPV replication cycle is completed as keratinocytes undergo differentiation. Virions are assembled in the nuclei of differentiated keratinocytes and can be detected by electron microscopy. Infection is transmitted by contact with virus contained in these desquamated keratinocytes (or with free virus) from an infected individual.

The Immune Response to HPV Infection

Unlike many viral infections, HPV infection has no viremic phase. This lack of viremia may account for the incomplete antibody response to HPV infection. Natural HPV infection of the genital tract gives rise to a serum antibody response in 60–70% of individuals. Significant, although incomplete, protection against type-specific reinfection is associated with the presence of neutralizing antibodies. Serum antibodies likely reach the cervical epithelium and secretions by transudation and exudation. Therefore, protection against infection relates to the amount of neutralizing antibody at the site of infection and lasts as long as sufficient levels of neutralizing antibodies are present.

A cell-mediated immune response plays an important role in controlling progression of HPV infection. Histologic examination of lesions in individuals who experience regression of genital warts demonstrates infiltration by T cells and macrophages. CD4+ T cell regulation is particularly important in controlling HPV infections, as evidenced by the higher rates of infection and disease in immunosuppressed individuals, particularly those who are infected with HIV. Specific T-cell responses may be measured against HPV proteins, the most important of which appear to be the E2 and E6 proteins. In women with HPV type 16 cervical infection, a strong T-cell response to type 16–derived E2 protein is associated with a lack of progression of cervical disease. However, measurable changes occur in the innate and adaptive immune systems of patients with HPV-associated cancers. There is suppression of the antigen-presentation process as well as suppression of antitumor activity. The end result is a reduction of HPV-specific antitumor immune responses and an increase in immunosuppressive cellular responses.

THE NATURAL HISTORY OF HPV-ASSOCIATED MALIGNANCY

HPV is transmitted by vaginal or anal intercourse, oral sex, and probably by touching a partner’s genitalia. In cross-sectional and longitudinal studies, ~40% of young women demonstrate evidence of HPV infection, with peaks during the teens and early twenties, soon after first coital experience. The number of lifetime sexual partners correlates with the likelihood of HPV infection and the subsequent risk of HPV-associated malignancy. HPV infection may occur in a monogamous person if that person’s partner is infected.

Most HPV infections become undetectable after 6–9 months, a phenomenon known as “clearance.” However, with prolonged follow-up and frequent sampling, the same HPV types may again be detected months or even years later. It is still debated whether such episodic detection indicates viral latency followed by reactivation or represents reinfection with an identical HPV type.

While HPV is the causative agent of several cancers, most attention has focused on cervical cancer, which is the second most common cancer in women worldwide. More than 500,000 women are diagnosed and 275,000 die from invasive cervical cancer annually. More than 85% of all cervical cancer cases, as well as deaths, occur in women living in low-income countries, especially countries in sub-Saharan Africa, Asia, and South and Central America.

Twenty-five years of evidence shows that HPV causes nearly 100% of cervical cancers. Persistent HPV infection is the most significant risk factor for cervical cancer; relative risks range from 10 to 20 and exceed 100 in prospective and case–control studies, respectively. The time from HPV infection to cervical cancer may exceed 20 years. Cervical cancer peaks in the fifth and sixth decades of life for women living in developed countries and as much as a decade earlier for women living in resource-poor countries. Persistent carriers of oncogenic HPV types are at greatest risk for high-grade cervical dysplasia and cancer.

Why HPV infections in some women but not others eventually lead to malignancy is not clear. Although oncogenic HPV infection is necessary for the development of cervical malignancy, only ~3–5% of infected women will ever develop this cancer, even in the absence of cytologic screening. Biomarkers that can predict which women will develop cervical cancer are not available. Immunosuppression in general plays a significant role in redetection/reactivation of HPV infections, while other factors, such as smoking, hormonal changes, chlamydial infection, and nutritional deficits, have an impact on viral persistence and cancer.

The International Agency for Research on Cancer has concluded that HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59 are carcinogenic in the uterine cervix. HPV type 16 is particularly virulent and causes 50% of cervical cancers. Worldwide, HPV types 16 and 18 cause at least 70% of cervical squamous cell carcinomas and 85% of cervical adenocarcinomas. Oncogenic types other than 16 or 18 cause the remaining 30% of cervical cancers. HPV types 16 and 18 also cause nearly 90% of anal cancers worldwide.

In addition to cervical and anal cancer, other HPV-associated cancers include vulvar and vaginal cancer (caused by HPV in 50–70% of cases), penile cancer (caused by HPV in 50% of cases), and at least 65% of oropharyngeal squamous cell carcinomas (OPSCCs). Over the past two decades, an epidemic of OPSCC related to oncogenic HPV infection, primarily HPV type 16, has developed. Rates of OPSCC in the United States have been increasing in men from a low of 0.27 case per 100,000 in 1973 to 0.57 case per 100,000 per year in 2004; rates in women have remained relatively stable at ~0.17 per 100,000 per year. The greatest increase in the incidence of OPSCC is among white men 40–50 years of age. Nearly 14,000 new cases were diagnosed in the United States in 2013. OPSCCs of the base of the tongue and tonsil cancer have increased annually by rates of 1.3 and 0.6%, respectively. Few data are available from developing countries about OPSCC.

THE EFFECTS OF HIV ON HPV-ASSOCIATED DISEASE

HIV infection accelerates the natural history of HPV infections. HIV-infected individuals are more likely than other individuals to develop genital warts, and their lesions are more recalcitrant to treatment. HIV infection

has been consistently associated with precancerous cervical lesions, including low-grade cervical intraepithelial lesions (CIN) and CIN 3, the immediate precursor to cervical cancer. Women with HIV/AIDS have significantly higher rates of cervical cancer as well as subsets of some vulvar, vaginal, and oropharyngeal tumors than women in the general population. Studies indicate a direct relationship between low CD4+ T lymphocyte count and the risk of cervical cancer. Some studies show a reduced likelihood of HPV infection and precancerous lesions of the cervix in HIV-infected women given antiretroviral therapy (ART). However, the incidence of cervical cancer in HIV-infected women has not changed significantly since ART was introduced, possibly because of preexisting oncogenic HPV infections that occurred before ART was initiated.

The burden of HPV-associated cancers is expected to increase in HIV-infected patients, given the prolonged life expectancies provided with ART. For women living in developing countries where cervical cancer screening is not widely available, this trend will have significant consequences. Thus, elucidating the interactions of HIV infection and cervical cancer with cofactors such as diet, other sexually transmitted infections, and environmental exposures is an important focus of research that impacts women living in low- and middle-income countries.

Similar to that of cervical cancer, the incidence of anal cancer is strongly influenced by HIV infection. HIV-infected men who have sex with men (MSM) and HIV-infected women have much higher rates of anal cancer than HIV-uninfected populations. Specifically, the incidence among HIV-infected MSM has been found to be as high as 130 cases per 100,000, as opposed to 5 cases per 100,000 among HIV-negative MSM. The advent of ART has not impacted the incidence of anal cancer and high-grade anal intraepithelial neoplasia in the HIV-infected patient population.

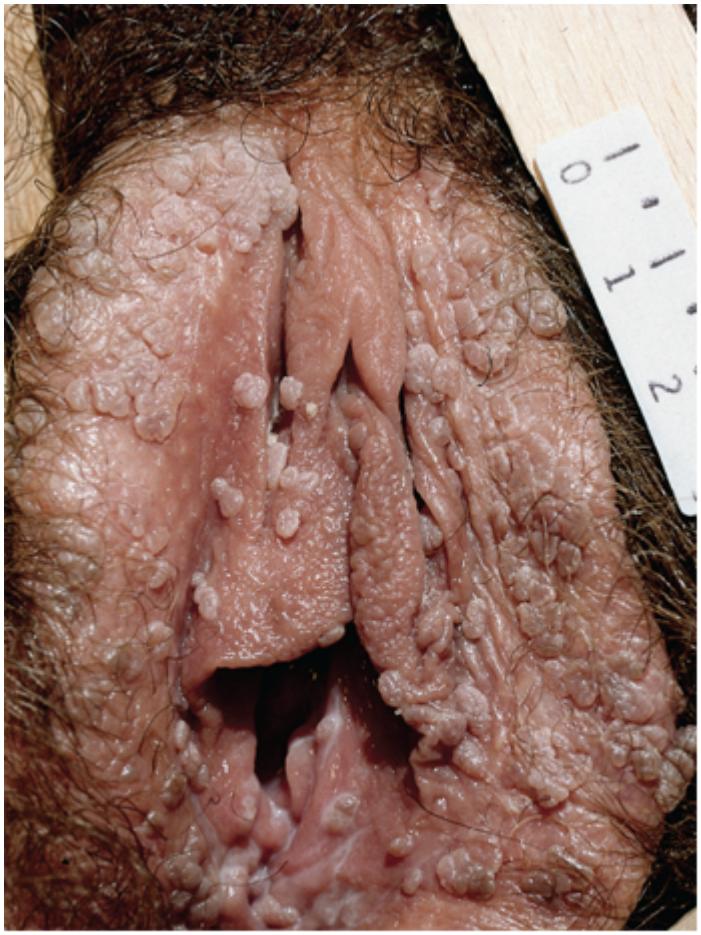
More information regarding screening, prevention, and treatment in the HIV-infected population can be found at the Department of Health and Human Services website (*aidsinfo.nih.gov/guidelines*).

CLINICAL MANIFESTATIONS OF HPV INFECTION

HPV infects the male urethra, penis, and scrotum and the female vulva, vagina, and cervix. Perianal, anal, and oropharyngeal infections occur in both genders. Genital warts are caused primarily by HPV type 6 or 11 and appear as soft sessile growths with a surface that is either smooth or rough with multiple finger-like projections. Penile genital warts are usually 2–5 mm in diameter and often occur in groups. A second type of penile lesion, the keratotic plaque, is slightly raised above normal epithelium and has a rough, often pigmented surface. **Figs. 193-1, 193-2, 193-3** show vulvar and vaginal, penile, and perianal warts, respectively.

FIGURE 193-1

Warts of the vulva and vagina caused by human papillomavirus. *(Reprinted with permission from K Wolff et al: Fitzpatrick's Color Atlas and Synopsis of Clinical Dermatology, 7th ed. New York: McGraw-Hill, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 193-2

Penile genital warts caused by human papillomavirus. *(Reprinted with permission from K Wolff et al: Fitzpatrick's Color Atlas and Synopsis of Clinical Dermatology, 7th ed. New York: McGraw-Hill, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 193-3

Perianal warts caused by human papillomavirus. *(Reprinted with permission from K Wolff et al: Fitzpatrick's Color Atlas and Synopsis of Clinical Dermatology, 7th ed. New York: McGraw-Hill, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Vulvar warts are soft, whitish papules that are either sessile or have multiple fine, finger-like projections. These lesions are most often located in the introitus and labia. In nonmucosal areas, vulvar lesions are similar in appearance to those in men: dry and keratotic. Vulvar lesions can appear as smooth, sometimes pigmented papules that may coalesce. Vaginal lesions appear as multiple areas of elongated papillae. Biopsy of vulvar or vaginal lesions may reveal malignancy; differentiation based on clinical exam is not always reliable.

Subclinical cervical HPV infections are common, and the cervix may appear normal on examination. Cervical lesions often appear as papillary proliferations near the transformation zone. Irregular vascular loops are present beneath the surface epithelium. Patients who develop cervical cancer from HPV infection may present with a variety of symptoms. Early carcinomas appear eroded and bleed easily. More advanced carcinomas present as ulcerated lesions or as an exophytic cervical mass. Some cervical carcinomas are located in the cervical canal and may be difficult to see. Bleeding, symptoms of a mass lesion in late stages, and metastatic disease that may manifest as bowel or bladder obstruction due to direct extension of the tumor have also been described.

Patients with squamous cell cancer of the anus have more variable presentations. The most common presentations include rectal bleeding and pain or a mass sensation. Twenty percent of patients who are

diagnosed with anal cancer may not present with any specific symptoms at the time of diagnosis, and the lesion is found fortuitously.

PREVENTION OF HPV INFECTION AND DISEASE

Behaviors That Can Reduce Exposure to HPV

HPV infections are transmitted through direct contact with infected genital skin or mucosal surfaces and secretions. Does abstinence reduce HPV infections? For both men and women, numerous studies indicate that HPV infection and HPV-associated diseases correlate with the number of lifetime sexual partners, and people with no history of sexual intercourse have a lower detection rate of HPV. Fewer studies look at nonpenetrative sex and the risk of HPV infection and disease, but several studies indicate that HPV can be spread by *any* sexual intimacy, including touching, oral sex, or use of sex toys. It is therefore possible that individuals who have not partaken in penetrative sex can become infected.

Use of latex condoms reduces the risk of HPV infection and HPV-associated disease, such as genital warts and cervical precancers. Correct and consistent condom use has also been associated with regression of CIN in women and regression of HPV-associated penile lesions in men. As a preventive measure, condom use should be considered partially effective at best, not a substitute for cervical cancer screening or vaccination against HPV.

HPV Vaccines

The development of HPV vaccines effective in preventing infection and HPV-associated disease represents a major development in the last decade. The vaccines use virus-like particles (VLPs) that consist of the HPV L1 major capsid protein. The L1 protein self-assembles into VLPs when expressed in eukaryotic cells (i.e., yeast for the Merck vaccine or insect cells for the GlaxoSmithKline vaccine). These VLPs contain the same epitopes as the HPV virion. However, they do not contain genetic material and cannot transmit infection. The immunogenicity of the HPV vaccines relies on development of conformational neutralizing antibodies to epitopes displayed on viral capsids.

Several large vaccine trials have been completed and demonstrate the high degree of safety and efficacy of HPV vaccines. The evidence to date shows high and sustained efficacy against disease caused by those HPV types represented in the vaccines (HPV types 6, 11, 16, and 18 in the Merck vaccine and HPV types 16 and 18 in the GlaxoSmithKline vaccine). However, no therapeutic effect of either vaccine against active infection or disease has been documented.

BIVALENT VACCINE (CERVARIX)

The bivalent L1 VLP vaccine (HPV types 16 and 18) marketed under the name Cervarix (GlaxoSmithKline) is administered by intramuscular injection at months 0, 1, and 6. This vaccine was tested in 18,644 women 15–25 years of age who resided in the United States, South America, Europe, and Asia. The primary endpoints of

the study included vaccine efficacy against persistent infections with HPV types 16 and 18. Investigators also assessed vaccine efficacy against CIN grade 2 or higher due to HPV 16 and 18 in women who had no evidence of HPV 16 or 18 infection at baseline. Vaccine efficacy related to HPV type 16 or 18 was 94.9% (95% confidence interval [CI], 87.7–98.4) against CIN 2 or worse; 91.7% (95% CI, 66.6–99.1) against CIN 3 or worse; and 100% (95% CI, –8.6 to 100) against adenocarcinoma in situ (AIS). Adverse events associated with the bivalent vaccine were evaluated in phase 3 trials in a subset of 3077 women who received vaccine and 3080 women who received hepatitis A vaccine. Injection-site adverse events (pain, redness, and swelling) and systemic adverse events (fatigue, headache, and myalgia) were reported more frequently in the HPV vaccine group than in the control group. Serious adverse events, new-onset chronic disease, or medically significant conditions occurred in the same proportion (3.5%) of HPV vaccine recipients and control vaccine recipients. The bivalent vaccine is approved in the United States for prevention of cervical cancer, CIN 2 or worse, AIS, and CIN 1 caused by HPV types 16 and 18. This vaccine is approved for females 9–25 years of age.

QUADRIVALENT VACCINE (GARDASIL)

The quadrivalent L1 VLP vaccine (HPV types 6, 11, 16, and 18) marketed under the name Gardasil (Merck) is administered intramuscularly at months 0, 2, and 6. A combined efficacy analysis based on data from four randomized double-blind clinical studies including >20,000 participants was performed; results demonstrated that vaccine efficacy against external genital warts was 98.9% (95% CI, 93.7–100). Vaccine protective efficacy was 95.2% (95% CI, 87.2–98.7) against CIN; 100% (95% CI, 92.9–100) against type 16– or 18–related CIN 2/3 or AIS; and 100% (95% CI, 55.5–100.0) against type 16– or 18–related vulvar intraepithelial neoplasia (VIN) grades 2 and 3 and against vaginal intraepithelial neoplasia (VaIN) grades 2 and 3.

Safety data on the quadrivalent HPV vaccine are available from seven clinical trials including nearly 12,000 women 9–26 years of age who received the vaccine and ~10,000 women who received aluminum-containing or saline placebo. A larger proportion of young women reported injection-site adverse events in the vaccine groups than in the placebo groups. Systemic adverse events were reported by similar proportions of vaccine and placebo recipients and were described as mild or moderate for most participants. The types of serious adverse events reported were similar for the two groups. Ten persons who received the quadrivalent vaccine and seven persons who received placebo died during the trials; no deaths were considered to be vaccine related.

During the course of studies on the quadrivalent vaccine, surveillance data on the development of new medical conditions were collected for up to 4 years after vaccination. No statistically significant differences in the incidence of any medical conditions between vaccine and placebo recipients were demonstrated; this result indicated a very high safety profile for the vaccine. A recent safety review by the U.S. Food and Drug Administration (FDA) and the Centers for Disease Control and Prevention (CDC) examined events related to Gardasil that had been reported to the Vaccine Adverse Events Reporting System (VAERS). The adverse events were consistent with what was seen in previous safety studies of the vaccine. Of note, rates of syncope and venous thrombotic events were higher with Gardasil than those usually observed with other vaccines.

The quadrivalent vaccine is approved for (1) vaccination of girls and women ages 9–26 years of age to prevent genital warts and cervical cancer caused by HPV types 6, 11, 16, and 18; (2) vaccination of the same population to prevent precancerous or dysplastic lesions, including cervical AIS, CIN 2/3, VIN 2/3, VaIN 2/3, and CIN 1; (3) vaccination of boys and men 9–26 years of age to prevent genital warts caused by HPV types 6 and 11; and (4) vaccination of people 9–26 years of age to prevent anal cancer and associated precancerous lesions due to HPV types 6, 11, 16, and 18. While the duration of protection has not been established, no evidence of waning protection has been found after a three-dose series of quadrivalent HPV vaccine, even after 10 years of follow-up from clinical trials.

NINE-VALENT VACCINE (GARDASIL-9)



HPV types 16 and 18 together cause up to 80% of all cervical cancers worldwide, and worldwide data show that HPV types 31, 33, 35, 45, 52, and 58 are the next most frequently detected types in invasive cervical cancers. In 2014, the FDA approved a new nine-valent L1 VLP vaccine that targets HPV types 6, 11, 16, and 18 (the types also targeted by the quadrivalent HPV vaccine) as well as five additional oncogenic HPV types (31, 33, 45, 52, and 58). The nine-valent vaccine is marketed under the name Gardasil-9 (Merck) and is administered intramuscularly at months 0, 2, and 6.

In clinical studies of girls and women 16–26 years of age, the nine-valent HPV vaccine generated a noninferior antibody response to HPV types 6, 11, 16, and 18 compared to the quadrivalent vaccine. Bridging immunologic studies in male and female vaccine recipients 9–15 years of age and in boys and men 16–26 years of age indicated that the lower bound of the 95% CIs of the geometric mean titer ratio and seroconversion rates met criteria for noninferiority for all HPV types represented in the vaccine. In female recipients 16–26 years of age, vaccine efficacy against the combined endpoint of high-grade cervical, vulvar, or vaginal disease caused by any of the five additional oncogenic HPV types was 96.7% (95% CI, 80.9–99.8%). Like the other available HPV vaccines, the nine-valent HPV vaccine is safe and extremely well tolerated.

Mathematical models estimate that the level of protection conferred by the nine-valent HPV vaccine against all HPV-associated squamous cell cancers worldwide could be raised to 90%. However, the true impact will depend on vaccine uptake, especially in poor countries where cervical cancer is common and the vaccine is currently prohibitively expensive.

CROSS-PROTECTION OF HPV VACCINES

Women who receive any of the available HPV vaccines produce neutralizing antibodies to virus types that are closely related to type 16 or 18. Analyses of data from clinical trials suggest that the HPV vaccines may offer limited cross-protection against nonvaccine virus types. Over short periods, the bivalent vaccine appears more efficacious against HPV types 31, 33, and 45 than the quadrivalent vaccine, but differences in study design make direct comparisons difficult, if not impossible. In addition, in the bivalent vaccine trials, vaccine efficacy against persistent infections with HPV types 31 and 45 waned over time, whereas efficacy against persistent infection with HPV type 16 or 18 remained stable. These results suggest that cross-protection is likely to be shorter lived than efficacy against infection and disease caused by vaccine types.

TWO-DOSE VERSUS THREE-DOSE SCHEDULE FOR HPV VACCINATION

In an effort to simplify the dosing schedule and potentially reduce costs and improve vaccine uptake, a two-dose schedule has been considered. In several randomized vaccine trials among adolescent girls, geometric mean concentrations (GMCs) of antibodies to HPV types 16 and 18 were shown to be noninferior up to 24 months after a two-dose schedule to GMCs after a three-dose schedule. Numerous countries have adopted a two-dose HPV vaccination schedule. In the United States, the CDC now recommends two doses of HPV vaccine (at 0 and 6–12 months) for persons starting the vaccination series before the 15th birthday, as the immunologic response is rigorous in this age group. Three doses of HPV vaccine (at 0, 1–2, and 6 months) are recommended for persons starting the vaccination series on or after the fifteenth birthday and for persons with certain immunocompromising conditions, including HIV/AIDS.

RECOMMENDATIONS FOR HPV VACCINATION

The most recent guidelines for HPV vaccination from the Advisory Committee on Immunization Practices (ACIP) are summarized below and provided in detail at *https://www.cdc.gov/mmwr/volumes/65/wr/mm6549a5.htm.*

For both girls and boys, the ACIP recommends that routine HPV vaccination be initiated at age 11 or 12 years, although the vaccination series can be started beginning at age 9 years. An individual can begin a vaccines series with one HPV vaccine and then complete the series with another.

For female recipients, vaccination is recommended through 26 years of age with the bivalent, quadrivalent (if still available), or nine-valent HPV vaccine. Papanicolaou (Pap) smear testing and screening for HPV DNA are not recommended before vaccination. After vaccination, cervical cancer screening should be conducted according to age-specific guidelines (see below).

For male recipients, vaccination with either the bivalent vaccine or (as long as this formulation is available) the nine-valent vaccine is recommended through 21 years of age to those who have not been vaccinated previously or who have not completed the three-dose series. Men 22–26 years of age may also be vaccinated.

SCREENING FOR HPV-ASSOCIATED CANCER

Once HPV infection occurs, prevention of HPV-associated disease relies on screening. At present, screening for cervical cancer is widely accepted as cost-effective in preventing cervical cancer, and anal screening is accepted for screening in high-risk groups. In resource-rich countries, the primary method of cervical cancer screening is cytology via Pap smear. The American Society of Colposcopy and Cervical Pathology guidelines recommend initiation of cervical cancer screening at age 21, no matter the age of sexual debut. Women 21–29 years old should have a Pap smear every 3 years if their initial and subsequent Pap smears are normal. Although adolescent and young women often test positive for HPV DNA, they are at very low risk of cervical cancer. Because the presence of HPV DNA does not correlate with the presence of high-grade squamous

intraepithelial neoplasia, co-testing (testing for HPV DNA at the time of Pap smear) is not recommended for women in this age group.

As a method of determining the need for colposcopy, HPV DNA co-testing is recommended for women 25–29 years of age in whom cytology detects abnormal squamous cells of undetermined significance. Women 30–65 years of age should have a Pap smear every 3 years if testing for HPV DNA is not performed. The screening interval for women in this age group can be extended to every 5 years if HPV DNA co-testing is performed and results are negative. HPV testing is not recommended for partners of women with HPV or for screening of conditions other than cervical cancer.

The role of HPV DNA testing as a primary screen for cervical cancer is changing. In 2014, the FDA approved the Roche Cobas HPV test for primary screening of cervical cancer. This test can be used to detect HPV DNA in specimens obtained from the cervix without cervical cytology for women ≥25 years of age. A positive result for HPV type 16 or 18 has a high enough positive predictive value in the general population that these women should have colposcopy performed. If high-risk HPV types other than 16 or 18 are detected, then cytology can be used. Ongoing studies of other HPV DNA testing methods are evaluating their use as a primary screening tool. The complete set of algorithms for appropriate age-specific screening guidelines, HPV DNA testing, and the management of abnormal Pap smears is available through the American Society of Colposcopy and Cervical Pathology at *http://www.asccp.org/asccp-guidelines*.

For women ≤30 years of age who are infected with HIV, cervical cytology is the preferred method of cervical cancer screening, and HPV DNA co-testing is not recommended. Cervical cancer screening should begin within 1 year of diagnosis of HIV infection, regardless of the mode of HIV transmission. If the first Pap smear is normal, then subsequent Pap smears should be performed annually until three negative tests are obtained. Cytology can then be conducted every 3 years. For women >30 years old, Pap testing is performed in the same manner as for younger women. However, HPV DNA co-testing can be used in women of this age group. If cytology and HPV DNA co-testing are negative, the next exam can be performed in 3 years. Positive results of HPV DNA co-testing are treated in the same manner as in HIV-uninfected women.

Women residing in developing countries with a lack of access to cervical screening programs have a higher rate of cervical cancer and poorer cancer-specific survival. Approximately 75% of women living in developed countries have been screened in the past 5 years, as opposed to ~5% of women living in developing countries. Economic and logistic obstacles likely impede routine cervical cancer screening for these populations. Many poor countries rely on an alternative method—visual inspection with acetic acid (VIA)—for cervical cancer screening. While some studies show a reduction in cervical cancer mortality in communities where VIA is widely used, other studies do not. In addition, the low specificity of VIA is problematic. As newer methods that use detection of oncogenic HPV DNA become available, even resource-limited countries may be able to replace VIA with these methods and achieve a reduction in cervical cancers as a result.

Currently, there is no broad consensus regarding screening for anal cancer and its precursors, including high-grade anal intraepithelial lesions. The reason is a lack of understanding of optimal treatment for low- or high-grade anal dysplasia found during cytologic screening. Current HIV treatment guidelines suggest that there

may be a benefit to screening, but an effect on the associated morbidity and mortality of anal squamous cell cancer has not been consistently demonstrated.

TREATMENT

TREATMENT

HPV-Associated Disease

A variety of treatment modalities are available for various HPV infections, but none has been proven to eliminate HPV from tissue adjacent to the destroyed and infected tissue. Treatment efficacies are limited by frequent recurrences, presumably due to reinfection from an infected partner, reactivation of latent virus, or autoinoculation from nearby infected cells. The goals of treatment include prevention of viral transmission, eradication of premalignant lesions, and reduction of symptoms.

Therapies are generally successful in eliminating visible lesions and grossly diseased tissue. Different therapies are indicated for genital warts, vaginal and cervical disease, and perianal and anal disease.

THERAPEUTIC OPTIONS

Imiquimod

Imiquimod (5 or 3.75% cream) is a patient-applied topical immunomodulatory agent thought to activate immune cells by binding to a Toll-like receptor that leads to an inflammatory response. Imiquimod 5% cream is applied to genital warts at bedtime three times per week for up to 16 weeks. Warts are cleared in ~56% of patients, more often in women than in men; recurrence rates approach 13%. Local inflammatory side effects are common. Rates of clearance of genital warts are not as high with the 3.75% formulation as with the 5% preparation, but the duration of treatment is shorter (daily application required for a maximum of 8 weeks) and fewer local and systemic adverse reactions occur. Imiquimod should not be used to treat vaginal, cervical, or anal lesions. The safety of imiquimod during pregnancy has not been established.

Interferon

Recombinant interferon α is used for intralesional treatment of genital warts, including perianal lesions. The recommended dosage is 1.0×10^6 IU of interferon into each lesion three times weekly for 3 weeks. Interferon therapy causes clearance of infected cells by immune-boosting effects. Adverse events include headache, nausea, vomiting, fatigue, and myalgia. Interferon therapy is costly and should be reserved for severe cases that do not respond to less expensive treatments. Interferon should not be used to treat vaginal, cervical, or anal lesions.

Cryotherapy

Cryotherapy (liquid nitrogen treatment) for HPV-associated lesions causes cellular death. Genital warts usually disappear after two or three weekly sessions but often recur. Cryotherapy, which is nontoxic and is not associated with significant adverse reactions, can also be used for diseased cervical tissue. Local pain occurs frequently.

Surgical Methods

Exophytic lesions can be surgically removed after intradermal injection of 1% lidocaine. This treatment is well tolerated but can cause scarring and requires hemostasis. Genital warts can also be destroyed by electrocautery, in which no additional hemostasis is required.

Laser Therapy

Laser treatment affords destruction of exophytic lesions and other HPV-infected tissue while preserving normal tissue. Local anesthetics are generally adequate. Efficacy for genital lesions is at least equal to that of other therapies (60–90%), with low recurrence rates (5–10%). Complications include local pain, vaginal discharge, periurethral swelling, and penile or vulvar swelling. Laser therapy has also been used successfully for cervical dysplasia and anal disease caused by HPV.

Therapeutic Vaccines

The innate and adaptive immune systems are altered in patients with HPV-associated cancers. Antitumor immune responses are blunted by specific viral mechanisms. Numerous therapeutic vaccines that are being developed are designed to enhance the cell-mediated response to the HPV E6 and E7 oncoproteins, which are expressed in HPV-associated cancers. Such vaccines would enhance the ability to treat HPV-associated cancers, conditions that are very difficult to treat with current modalities. However, while progress has been made, no HPV vaccine is currently available for treatment of HPV infection or HPV-associated disease.

Other Therapies

Both trichloroacetic acid and bichloroacetic acid are caustic agents that destroy warts by coagulation of proteins. Neither of these agents is recommended for treatment. Sinecatechins (15% ointment) and podophyllotoxin (0.05% solution or gel and 0.15% cream) are occasionally used for external genital warts, but other modalities listed above are as or more effective and are better tolerated.

RECOMMENDATIONS FOR TREATMENT

Table 193-1 lists available treatments for genital warts. An optimal therapy for HPV-related genital tract disease that combines high efficacy, low toxicity, low cost, and low recurrence is not available. For genital warts of the penis or vulva, cryotherapy is the safest, least expensive, and most effective modality. However, all available modalities for treatment of genital warts carry high rates of recurrence. Guidelines for the treatment of genital warts can be found on the CDC website (*http://www.cdc.gov/std/treatment/2010/genital-warts.htm*).

Women with vaginal lesions should be referred to a gynecologist experienced in colposcopy and treatment of these lesions. Treatment of cervical disease involves careful inspection, biopsy, and histopathologic grading to determine the severity and extent of disease. Women with evidence of HPV-associated cervical disease should be referred to a gynecologist familiar with HPV and experienced in colposcopy. Optimal follow-up of these patients includes colposcopic examination of the cervix and vagina on a yearly basis. Guidelines from the American College of Obstetricians and Gynecologists are available for the treatment of cervical dysplasia and cancer.

For anal or perianal lesions, cryotherapy or surgical removal is safest and most effective. Anoscopy and/or sigmoidoscopy should be performed in patients with perianal lesions, and suspicious lesions should be biopsied to rule out malignancy.

TABLE 193-1

Recommended Treatments for Genital Warts Caused by Human Papillomavirus[a]

Treatment	Imiquimod	Cryotherapy	Interferon	Surgical removal	Laser
Effectiveness	Good	Good	Good	Excellent	Excellent
Recurrence	Frequent	Frequent	Frequent	Frequent	Frequent
Adverse effects	Frequent, mild to moderate	Mild, well tolerated	Frequent, moderately severe	Mild, well tolerated	Mild to moderate, well tolerated
Availability	Fair	Good	Fair	Good	Fair
Cost	Expensive	Inexpensive	Very expensive	Moderately expensive	Very expensive

[a]Imiquimod can be self-administered. All other treatments must be administered by a clinician.

COUNSELING PATIENTS REGARDING HPV DISEASE

Most sexually active adults will be infected with HPV during their lives. The only way to avoid acquiring an HPV infection is to abstain from sexual activity, including intimate touching and oral sex. Practicing safe sex (partner reduction, use of condoms) may help reduce HPV transmission. Most HPV infections will be controlled by the immune system and cause no symptoms or disease. Some infections lead to genital warts and cervical precancers. Genital warts can be treated for cosmetic reasons and to prevent spread of infection to others. Even after resolution of genital warts, latent HPV may persist in normal-appearing skin or mucosa and thus theoretically may be transmitted to uninfected partners. Precancerous cervical lesions should be treated to prevent progression to cancer.

FURTHER READING

Clifford GM et al: Carcinogenicity of human papillomavirus (HPV) types in HIV-positive women: A meta-analysis from HPV infection to cervical cancer. Clin Infect Dis 64:1228, 2017.

de Sanjosé S et al: The natural history of human papillomavirus infection. Best Pract Res Clin Obstet Gynaecol 47:2, 2018.

Garland SM et al: Impact and effectiveness of the quadrivalent human papillomavirus vaccine: A systematic review of 10 years of real-world experience. Clin Infect Dis 63:519, 2016.

Giuliano AR et al: Efficacy of quadrivalent HPV vaccine against HPV infection and disease in males. N Engl J Med 364:401, 2011.

Gravitt PE, Winer RL: Natural history of HPV infection across the lifespan: Role of viral latency. Viruses, Sept 21, 2017. doi: 10.3390/v9100267.

Lopalco PL: Spotlight on the 9-valent HPV vaccine. Drug Des Devel Ther 11:35, 2016.

Schiffman M et al: Carcinogenic human papillomavirus infection. Nat Rev Dis Primers, Dec 1, 2016. doi: 10.1038/nrdp.2016.86.

Serrano B et al: Epidemiology and burden of HPV-related disease. Best Pract Res Clin Obstet Gynaecol 47:14, 2018.

Small W JR et al: Cervical cancer: A global health crisis. Cancer 123:2404, 2017.

Wentzensen N et al: Eurogin 2016 roadmap: How HPV knowledge is changing screening practice. Int J Cancer 140:2192, 2017.

Chapter 194: Common Viral Respiratory Infections

James E. Crowe, Jr.

INTRODUCTION

The most common and frequent infections in humans are respiratory virus infections. Some classical respiratory viruses (e.g., rhinoviruses) enter the body through the respiratory tract, replicating and causing disease only in cells of the respiratory epithelium. Other, more systemic viruses (e.g., measles virus and severe acute respiratory syndrome [SARS] coronavirus) spread via the bloodstream and cause systemic disease; however, they also may enter through and cause disease in the respiratory tract. Although infections with systemic viruses often induce lifelong immunity against disease, respiratory viruses that do not cause viremia usually can reinfect the same host many times throughout life. Reinfection with the same virus is common because of incomplete or waning immunity after natural infection. Hundreds of different viruses cause infection of the respiratory tract, and within each virus type there can be a nearly unlimited diversity of field strains that vary antigenically, geographically, and over time (e.g., antigenically drifting influenza viruses). Specific antiviral treatment options are limited, and only a few licensed vaccines are available. **For further discussion of common respiratory virus infections, see Chap. 31 and syndrome-specific chapters.**

Common viral respiratory infections can be categorized in several ways, including by site of anatomic involvement, disease syndrome, or etiologic agent.

ANATOMIC SITES IN THE HUMAN RESPIRATORY TRACT

The type of respiratory disease that develops during virus infection is dictated to a large degree by the cell types and tissue organization in the respiratory tract. The vocal cords mark the transition between the upper and lower respiratory tracts. The upper respiratory tract is a complex anatomic system with interconnected structures, including the sinuses, middle-ear spaces, Eustachian tubes, conjunctiva, nasopharynx, oropharynx, and larynx. The tonsils and the adenoids are large collections of lymphoid tissue in the pharynx that participate in immunity but also are susceptible to infections. The lower respiratory tract structures include the trachea, bronchi, bronchioles, alveolar spaces, and lung tissue, including epithelial cells and blood vessels. The epithelial cell types that line the respiratory tract are varied in morphology and function, and their susceptibility to different virus infections varies. The principal types of cells in the major airways are ciliated or nonciliated epithelial cells, goblet cells, and Clara cells. Smooth-muscle cells form major tissue structures around the epithelial structures of the large airways of the lower respiratory tract down to the level of the bronchioles, and these cells are reactive to intrinsic and extrinsic signals, including viral infection

or exposure to allergens or pollutants. The pathologic process of wheezing is driven by smooth-muscle contraction and obstruction of airways caused by mucous accumulation and epithelial sloughing in the lumen. Reactive airways causing wheezing are most often due to constriction of lumen size at the level of the bronchioles (which have the narrowest lumen diameter of the airways). The lung does not have smooth-muscle or ciliated cells, but instead possesses pneumocytes of types I and II. Pneumonia (Chap. 121) is an infection of the pneumocytes in the lung tissue and the alveolar spaces. The alveolar spaces also contain cells of the monocyte lineage, such as macrophages, which patrol the air spaces.

DISEASE SYNDROMES

Since different respiratory viruses tend to have a predilection for replication in differing cells or regions of the respiratory tract, it is possible for the well-trained clinician with epidemiologic information to understand the most likely associations of viruses with clinical syndromes. The clinical diagnoses for virus infections in the upper respiratory tract are rhinitis or the common cold, sinusitis, otitis media, conjunctivitis, pharyngitis, tonsillitis, and laryngitis. In reality, some upper respiratory tract infections affect more than one upper respiratory tract anatomic site during a single infection, such as the classical pattern of pharyngoconjunctival fever during adenovirus infection. Lower respiratory tract syndromes also can be associated easily with anatomic region, including tracheitis, bronchitis, bronchiolitis, pneumonia, and exacerbations of reactive airway disease or asthma. Bronchiolitis is a disease condition characterized by trapping of air in the lungs with difficulty in expiration (i.e., wheezing); it is caused by inflammation or infection of the bronchioles, the smallest and most highly resistant airways. Again, mixed syndromes occur, such as laryngotracheitis, usually termed croup. Croup, a disease condition characterized by difficulty in inspiration associated with a barky cough, is caused by inflammation or infection of the larynx, trachea, and bronchi. When respiratory symptoms occur in the context of a respiratory viral illness with significant systemic signs, infection with particular agents can be suspected (e.g., influenza, measles, SARS, or hantavirus pulmonary syndrome [HPS]), with exposure history taken into account.

ETIOLOGIC AGENTS

RESPIRATORY VIRUSES CAUSING DISEASE IN IMMUNOCOMPETENT HOSTS

Children have more frequent respiratory virus infections than adults; thus it was natural that many early discoveries about the viral causes of respiratory infections came from pediatric studies. The principal causes of acute viral respiratory infections were determined in large epidemiologic studies in the 1960s and 1970s, when cell culture of infectious agents became available. More recently, studies of viral epidemiology have been conducted in adults, especially in special populations such as the elderly, nursing home residents, and immunocompromised individuals. Rapid antigen detection tests (based on immunoassays for detection of viral proteins) became available for respiratory syncytial virus (RSV) and influenza virus in the 1980s. With the availability of sensitive and specific molecular tests, such as reverse transcription combined with the polymerase chain reaction (RT-PCR), studies in the past two decades have greatly increased the extent to which we understand the causes of viral respiratory infections. Multiplex panels of RT-PCR tests capable of

detecting a dozen or more viruses are commonly available for clinical testing of respiratory secretions. These sensitive tests have been especially helpful in studies of infection in adults, who often shed much lower concentrations of virus in secretions than do children. Influenza viruses, RSV, and human metapneumovirus (hMPV) are the most common causes of serious lower respiratory tract disease in otherwise healthy subjects; parainfluenza viruses (PIVs) and adenoviruses also cause substantial disease. Rhinoviruses (the most common cause of the common cold syndrome) have been increasingly associated with lower respiratory tract syndromes. Rhinovirus infection is so common, even in asymptomatic individuals, that it has been hard to establish clear figures for the role of rhinovirus in lower respiratory disease. Generally, about two-thirds of cases of respiratory illness in a research setting can be associated with a specific viral agent. Besides the viruses mentioned above (and discussed below), a number of additional viruses identified with molecular tools have been associated with respiratory illness. Still, it is fair to say that our diagnostic tools remain suboptimal since a specific infectious agent is not identified in approximately one-third of clinical respiratory illnesses in large surveillance studies. It is likely that in most of these cases pathogens are not detected because of the very low titers of virus in patient samples at the time of clinical presentation, which may occur after the period of peak virus shedding. It is also possible that novel agents are yet to be identified. As emerging tools for microbiome and “virome” studies (with sequencing of all nucleic acids in a sample) are applied in these settings in coming years, new agents and new associations with disease will probably be discovered.

RESPIRATORY VIRUSES CAUSING DISEASE IN IMMUNOCOMPROMISED HOSTS

Special populations of patients are susceptible not only to the conventional respiratory viruses discussed above but also to agents causing symptoms during reactivation of latent viruses or new infections with opportunistic agents. Most prominently, reactivating latent viruses, such as herpes simplex virus (HSV) and cytomegalovirus (CMV) and adenoviruses, cause disease in immunocompromised humans. Patients at most risk are those with hematopoietic stem cell or solid organ transplantation, leukopenia caused by chemotherapy, or advanced HIV-AIDS. In immunosuppressed patients with pneumonia, CMV is the virus recovered most frequently during deep respiratory tract diagnostic procedures such as bronchoalveolar lavage. These patients also are highly susceptible to more frequent and more severe disease caused by common respiratory viruses, including RSV, hMPV, PIVs, influenza viruses, rhinoviruses, and adenoviruses. Conventional acute respiratory viruses can cause chronic and sometimes fatal infections in these populations. Nosocomial transmission of respiratory viruses occurs in hematopoietic stem cell transplantation units, and the frequency of transmission can be high, with entire units affected.

SPECIFIC VIRAL CAUSES OF RESPIRATORY DISEASE

Orthomyxoviridae: Influenza Viruses

Influenza virus infection and influenza syndrome usually are associated with fever, myalgias, fatigue, sore throat, headache, and cough **(See also Chap. 195)**. Influenza causes severe and even fatal pneumonia, particularly in elderly patients, nursing home residents, immunocompromised persons, and very young

children. Influenza pneumonia has an unusually high rate of complication by bacterial superinfection, with staphylococcal and streptococcal bacterial pneumonia occurring in as many as 10% of cases in some clinical series.

Influenza is a single-stranded, segmented, negative-sense, RNA genome virus of the family Orthomyxoviridae. There are three (sero)types of influenza viruses: A, B, and C. Influenza A and C viruses infect multiple species, while influenza B virus infects humans almost exclusively. Type A viruses appear to be the most virulent for humans and most commonly cause severe disease manifestations, although type B viruses cause substantial morbidity. On the basis of antibody response, influenza A viruses can be subdivided into 18 different hemagglutinin (H) surface protein subtypes and 11 neuraminidase (N) surface protein subtypes. The subtypes that have caused major pandemics in humans are H1N1, which caused the 1918 pandemic; H2N2, which caused the 1957 pandemic; H3N2, which caused the 1968 pandemic; and H1N1pdm2009, which caused the 2009 pandemic. Currently, type A subtypes H1N1 and H3N2 and type B viruses cause annual seasonal epidemics.

Major pandemics caused by new influenza viruses are always possible. Many highly pathogenic influenza viruses circulate in aquatic birds. Occasionally, avian viruses infect humans directly after close contact with infected wild birds or poultry. Co-housing of pigs (which have both avian and human influenza virus receptors) with poultry may increase the risk of reassortment of human and animal or bird viruses; reassortment can make the zoonotic viruses more fit for replication in humans. Several outbreaks of avian influenza have occurred in limited numbers of humans to date, and there is the risk of a worldwide pandemic with avian influenza viruses if a strain acquires the potential to spread efficiently from human to human. H5N1 influenza virus infection of humans, predominantly by direct chicken-to-human transmission, occurred during an epizootic in Hong Kong's poultry population in 1997. The disease affected many types of wild and domestic birds and caused a high rate of systemic disease and death in infected humans. This virus, carried in the gastrointestinal tract of wild birds, has spread throughout Asia and beyond and continues to evolve antigenically. Avian H7N7 and H7N9 viruses also have caused zoonotic outbreaks. A significant outbreak of H7N9 virus infection began in China in March 2013, with high mortality, and seasonal outbreaks that have subsequently occurred nearly yearly threaten to cause a pandemic. H1N2 virus is endemic in pigs and affects humans with close contact. An H3N2 variant virus that differs antigenically from seasonal human viruses is endemic in pigs and occasionally infects children who have close contact with pigs in the United States. Rare human cases caused by H6, H7, H9, and H10 viruses have been reported. Type B influenza viruses co-circulate in humans during seasonal epidemics. Type B viruses mutate less frequently than type A viruses, and there is only one influenza B subtype. The slower evolution of type B viruses is probably linked to the fact that they are almost exclusively human pathogens. There is some antigenic diversity in these strains, however, and two major lineages have been designated B/Shanghai-like and B/Fujian-like strains.

Paramyxoviridae

RESPIRATORY SYNCYTIAL VIRUS

RSV is a single-stranded, negative-sense, nonsegmented, RNA genome virus of the genus *Pneumovirus* in the family Paramyxoviridae. Infection is ubiquitous, affecting most humans in the first several years of life and

causing reinfections throughout life. RSV is among the most transmissible viruses of humans. Disease epidemics occur yearly, typically between October or November and March in temperate regions. RSV is one of the most common viral causes of severe lower respiratory tract illness in the elderly and in children; it is among the most important causes of hospitalization of elderly and infant patients throughout the world. There is only one serotype of RSV, but antigenic variability does occur in circulating field strains. In immune serum reciprocal cross-neutralization studies, the two antigenic subgroups, A and B, appear to be ~25% antigenically related; this relatedness may partially explain the susceptibility of humans to reinfection, which is very common and can be caused by viruses of the same subgroup or even the same strain. However, reinfection in otherwise healthy adults usually is associated with mild disease confined to the upper respiratory tract. Severe lower respiratory tract disease is common in the elderly, especially in frail institutionalized elderly populations. Immunocompromised patients of any age also are at risk of severe or prolonged disease, especially recipients of hematopoietic stem cell transplants. Wheezing is common with primary infection in children (bronchiolitis), and there is a strong association of RSV infection early in life and subsequent asthma, although it is unclear whether severe childhood RSV causes asthma or is the first manifestation of reactive airway disease. RSV causes exacerbations of asthma and is associated with acute exacerbations of chronic obstructive pulmonary disease (COPD), also referred to as acute exacerbations of chronic bronchitis (AECB).

HUMAN PARAINFLUENZA VIRUSES

The human PIVs are a group of four distinct serotypes (designated 1–4) of single-stranded, negative-sense RNA viruses belonging to the family Paramyxoviridae. PIV3 most commonly causes severe disease, and repeated infection is common throughout life, although secondary infections often are mild or asymptomatic. Primary infections in children manifest as laryngotracheitis (croup), while subsequent infections typically are limited to the upper respiratory tract. PIVs are detected with sensitive RT-PCR tests or, more classically, by cell culture with immunofluorescent microscopy or hemadsorption in reference laboratories.

HUMAN METAPNEUMOVIRUS

hMPV was discovered only in 2001 but probably has always been present in human populations. Infection occurs first in early childhood, and reinfections are common throughout life. This virus is similar in many respects to RSV. It belongs to the family Paramyxoviridae and is a member of the genus *Pneumovirus*. It causes both upper and lower respiratory disease. It appears to be somewhat less virulent than RSV, causing about half as much severe lower respiratory tract disease, probably because it does not possess the nonstructural genes that RSV expresses in infected cells to abrogate the effect of host innate immune effectors like interferons. The clinical features of lower respiratory tract infections caused by hMPV are similar to those of such infections caused by other paramyxoviruses, most often including cough, coryza, and wheezing. Like RSV, hMPV plays an important role in exacerbations of asthma or COPD and causes pneumonia or wheezing in frail and institutionalized elderly individuals and immunocompromised patients.

MEASLES VIRUS

Measles virus is also a paramyxovirus but of the genus *Morbillivirus* **(See also Chap. 200)**. This virus causes a systemic infection known as rubeola but also can manifest with respiratory symptoms. Measles virus probably is the most contagious respiratory virus infection of humans: it is transmitted efficiently not only by

direct contact with infected persons or fomites (like other respiratory viruses) but also by small-particle aerosols. Measles virus infection is preventable by vaccination but is so infectious that cases are inevitable—even in the United States—whenever vaccination rates fall below 90–95% in a population. The virus causes systemic illness, sometimes including severe pneumonia, when primary infection occurs in an unvaccinated adult or an immunocompromised person of any age. Therefore, vigilance in maintaining high vaccination rates is critical. With primary infection, the illness in children is typically milder; however, mortality rates in lower-resource countries are high, especially among persons with underlying risk factors, including malnutrition.

Symptoms of measles include ≥3 days of high fever and a classical set of upper and lower respiratory tract symptoms sometimes termed "the 3 Cs": cough, coryza, and conjunctivitis. Unlike most respiratory viruses, measles virus circulates in the bloodstream and thus causes disseminated infection with systemic manifestations. Usually, a characteristic diffuse maculopapular rash appears within days of fever onset. Koplik's spots **(see Fig. A1-2)**—typical mucosal lesions in the mouth that appear briefly—are considered diagnostic of measles infection in the setting of the typical rash and fever.

Picornaviridae

A wide variety of picornaviruses cause respiratory disease, including non-polio enteroviruses, rhinoviruses, and parechoviruses **(Chap. 199)**. The designations of these viruses can be confusing: the enterovirus, rhinovirus, and parechovirus species names were changed (with the approval of the International Committee on Taxonomy of Viruses in February 2013) to remove references to host species names. These changes are summarized in **Table 194-1**. The genus *Enterovirus* consists of 13 species, including enteroviruses A through D and rhinoviruses A through C. The genus *Parechovirus* contains two species, one of which—Parechovirus A—encompasses 19 types: human parechovirus (HPeV) 1 through 19. These viruses exhibit seasonal patterns that differ from those of most other acute respiratory viruses. Rhinovirus infections occur year-round. Enterovirus infections occur most commonly in the summer months in temperate areas.

TABLE 194-1

Enterovirus, Rhinovirus, and Parechovirus Species Name Changes Made in Order to Remove References to Host Species Names and Approved by the International Committee on Taxonomy of Viruses in February 2013

Genus	Current Species Name	Former Species Name
Enterovirus (now 13 species)	Enterovirus A: consists of 25 serotypes, including coxsackieviruses and some non-polio enteroviruses that cause respiratory disease	Human enterovirus A
	Enterovirus B: consists of 63 serotypes, including some coxsackieviruses, echoviruses, and non-polio enteroviruses	Human enterovirus B
	Enterovirus C: consists of 23 serotypes, including the polioviruses	Human enterovirus C
	Enterovirus D: consists of 5 serotypes and includes enterovirus D68	Human enterovirus D
	Rhinoviruses A–C	Human rhinoviruses A–C
Parechovirus (2 species)	Parechovirus A: consists of 19 types (1–19). Human parechoviruses (HPeVs) 1 and 2 are common human pathogens.	HPeV-1 and HPeV-2 were formerly classified in the genus *Enterovirus* as echoviruses 22 and 23, respectively.

RHINOVIRUSES

Rhinoviruses have single-stranded, positive-sense RNA genomes. Rhinoviruses A through C represent species in the *Enterovirus* genus of the family Picornaviridae. Rhinoviruses are the most common viral infective agents in humans and the most frequent cause of the common cold. Field isolates of rhinovirus are exceptionally diverse; they can be classified by serotyping into more than 100 serotypes or alternatively by genotyping into a large number of genotypes that cause cold symptoms. At the time of writing in 2017, the species Rhinovirus A contained 80 types, Rhinovirus B had 32 types, and Rhinovirus C had 55 types. The viral particles are icosahedral in structure and are non-enveloped. Rhinoviruses are responsible for at least half of all cases of the common cold. Rhinovirus-induced common colds may be complicated in children by otitis media and in adults by sinusitis. Most adults, in fact, have radiographic evidence of sinusitis during the common cold, which resolves without therapy. Therefore, the primary disease is probably best termed *rhinosinusitis*. Rhinovirus infection is associated with exacerbations of reactive airway disease in children and asthma in adults. It is not clear whether rhinovirus is restricted to the upper respiratory tract and only indirectly induces inflammatory responses that affect the lower respiratory tract or whether the viruses

spread to the lower respiratory tract. In the past, it was thought that these viruses did not often replicate or cause disease in the lower respiratory tract. However, recent studies have discerned strong epidemiologic associations of rhinoviruses with wheezing and asthma exacerbations, including episodes severe enough to require hospitalization. Rhinoviruses likely can infect the lower airways to some degree, inducing a local inflammatory response. Another possibility is that significant local infection of the upper respiratory tract may induce regional elaboration of mediators that causes lower airway disease. The association of rhinovirus infection with lower respiratory tract illness is difficult to study because diagnosis by cell culture is not sensitive. RT-PCR diagnostic tests are difficult to interpret because they are often positive for prolonged periods and even asymptomatic individuals may have a positive test. Comprehensive serologic studies to confirm infection are difficult because of the large number of serotypes. Nevertheless, most experts believe rhinoviruses are a common cause of serious lower respiratory tract illness.

ENTEROVIRUSES

Non-polio enteroviruses are common and distributed worldwide. Although infection often is asymptomatic, these viruses cause outbreaks of clinical respiratory disease, sometimes with fatal consequences. The species Enterovirus A consists of 25 serotypes, including coxsackieviruses and some non-polio enteroviruses that cause respiratory disease. Coxsackieviruses cause oral lesions and often are associated in children with hand-foot-and-mouth disease. The pharyngitis associated with this infection characteristically manifests with herpangina, a clinical syndrome of ulcers or small vesicles on the palate that often involves the tonsillar fossa and is associated with fever, difficulty swallowing, and throat pain. Outbreaks commonly occur in young children during the summer. Enterovirus A71 also causes large outbreaks of hand-foot-and-mouth disease, especially in Asia, sometimes leading to neurologic complications and even death. The species Enterovirus B consists of 63 serotypes, including the echoviruses (*echo* being an acronym for *e*nteric *c*ytopathic *h*uman *o*rphan, which may be an archaic notion since most echoviruses are associated with human diseases, most commonly in children). Echoviruses can be isolated from many children with upper respiratory tract infections during the summer months. Echovirus 11 has been associated with laryngotracheitis or croup. Epidemiologic studies also have associated echoviruses with epidemic pleurodynia, an acute illness characterized by sharp chest pain and fever. The species Enterovirus C consists of 23 serotypes, including the polioviruses. The species Enterovirus D consists of five serotypes, including enterovirus D68, which has been associated with wheezing and some severe syndromes in children.

PARECHOVIRUSES

The genus *Parechovirus* comprises two species, one of which is Parechovirus A. The most common member of the genus *Parechovirus*, human parechovirus 1, is a frequent human pathogen. The genus also includes the closely related human parechovirus 2. Human parechoviruses usually cause mild respiratory or gastrointestinal illness. Most infections occur in young children. The seroprevalence of parechoviruses 1 and 2 is high among adults.

Adenoviridae

Viruses of the family Adenoviridae infect both humans and animals. As their designation indicates, adenoviruses were first isolated in human lymphoid tissues from surgically removed adenoids. In fact, some

serotypes establish persistent asymptomatic infections in tonsil and adenoid tissues, and virus shedding can occur for months or years. These double-stranded DNA viruses are <100 nm in diameter and have non-enveloped icosahedral morphology. The large double-stranded DNA genome is linear and nonsegmented. The seven major human adenovirus species (designated A through G) fall into 57 immunologically distinct serotypes. Human respiratory tract infections are caused mainly by the B and C species. Adenovirus infections can occur throughout the year. Many serotypes cause sporadic outbreaks, while others appear to be endemic in particular locations. Respiratory illnesses include mild disease such as the common cold and lower respiratory tract illnesses including croup, bronchiolitis, and pneumonia. Conjunctivitis is associated with infection by the B and D species. A particular constellation of symptoms referred to as pharyngoconjunctival fever is frequently associated with acute adenovirus infection. In contrast, gastroenteritis has been associated most frequently with serotypes 40 and 41 virus of species F. Immunocompromised patients are highly susceptible to severe disease during infection with respiratory adenoviruses. The syndrome of acute respiratory disease (ARD), especially common in stressful or crowded living conditions, was first recognized among military recruits during World War II and has continued to be a problem when vaccination has been suspended temporarily because of lapses in vaccine supply. ARD is most often associated with adenovirus types 4 and 7.

Coronaviridae

Members of the genus *Coronavirus* also contribute to respiratory illness, including severe disease. Dozens of coronaviruses affect animals. In the twentieth century, only two representative strains of human coronaviruses were known to cause disease: 229E (HCoV-229E) and OC43 (HCoV-OC43). An outbreak of infection with SARS-associated coronavirus (SARS-CoV) showed that animal coronaviruses have the potential to cross from other species to humans, with devastating effects. The one major epidemic to date (November 2002 through July 2003) encompassed more than 8000 cases, with mortality rates approaching 10%. SARS-CoV causes a systemic illness with a respiratory route of entry. SARS is a unique form of viral pneumonia. In contrast to most other viral pneumonias, SARS lacks upper respiratory symptoms, although cough and dyspnea occur in most patients. Typically, patients present with a nonspecific illness manifesting as fever, myalgia, malaise, and chills or rigors; watery diarrhea may occur as well. Investigators have reported the identification of a fourth human coronavirus, HCoV-NL63. Evidence is emerging that this new group 1 coronavirus is a common respiratory pathogen of humans, causing both upper and lower respiratory tract illness. HCoV-HKU1 was first described in January 2005 after its detection in a patient with pneumonia. Several cases of respiratory illness have been associated with this virus, but its infrequent identification suggests that this putative group 2 coronavirus has caused a low incidence of illness to date. The Middle East respiratory syndrome coronavirus (MERS-CoV), first isolated in 2012, causes severe disease in humans, with 35% mortality. MERS-CoV is a zoonotic virus (transmitted between animals and people). The virus may have emerged from bats in the Middle East. Studies have shown that humans are infected through direct or indirect contact with infected dromedary camels.

Herpesviridae

Several herpesviruses cause upper respiratory infections, especially infection of the oral cavity. Herpes simplex pharyngitis is associated with characteristic clinical findings, such as acute ulcerative stomatitis and ulcerative pharyngitis. HSV types 1 and 2—also called human herpesvirus (HHV) 1 and 2, respectively—both cause oral lesions **(Chap. 187)**, although >90% of oral infections are caused by HSV-1. Primary oral disease can be severe, especially in young children, who sometimes are admitted for rehydration therapy as a result of poor oral intake. A significant proportion of individuals suffer recurrences of symptomatic disease consisting of vesicles on the lips. Epstein-Barr virus (EBV) mononucleosis syndrome **(Chap. 189)** is often marked by acute or subacute exudative pharyngitis; in some cases, tonsillar swelling in EBV pharyngitis is so severe that airway occlusion appears imminent. Most of the viruses in the family Herpesviridae—including CMV **(Chap. 190)**; EBV; varicella-zoster virus (VZV; **Chap. 188**); and HHV-6, -7, and -8 **(Chap. 190)**—can cause severe disease in immunocompromised patients, especially hematopoietic stem cell transplant recipients.

Parvoviridae: Human Bocavirus

A new virus was recently identified in respiratory samples from children with lower respiratory tract disease in Sweden. Sequence analysis of the genome revealed that the virus is highly related to canine minute virus and bovine parvovirus and is a member of the genus *Bocavirus* (subfamily Parvovirinae, family Parvoviridae). This virus, tentatively named human bocavirus (HBoV), has been identified as the sole agent in a limited number of respiratory samples from children hospitalized with respiratory tract disease. Whether the virus causes or is merely associated with disease remains controversial.

Retroviridae: HIV

Pharyngitis occurs with primary HIV infection and may be associated with mucosal erosions and lymphadenopathy.

Papovaviridae: Polyomaviruses

Polyomaviruses are small, double-stranded, DNA-genome, non-enveloped icosahedral viruses that may be oncogenic. Two major polyomaviruses, JC and BK viruses, are known to infect humans. Of adults in the United States, ≥80% are seropositive for these viruses. JC virus can infect the respiratory system, kidneys, or brain. BK virus infection causes a mild respiratory infection or pneumonia and can involve the kidneys of immunosuppressed transplant recipients.

EPIDEMIOLOGY

AGE

Age (along with the associated factor of prior exposure history) is a major determinant of risk for symptomatic disease during respiratory virus infection. Primary infection with most of the acute respiratory viruses often is more severe than secondary infection. Indeed, reinfection with most of these viruses occurs throughout life, but primary infection is much more likely to be associated with severe lower respiratory tract

disease, while secondary infection typically is asymptomatic or associated with upper respiratory tract symptoms only. As these infections are ubiquitous, most primary infections (and thus many of the severe cases) occur during the first few years of life. Later, exposure to young children (in populations such as parents of young children and daycare workers) is a risk factor for frequent reinfection. Despite a lifetime of previous exposures, the risk of severe disease increases with age in the elderly, probably because of immune senescence and general medical decline.

SEASON

Infections with most of the conventional respiratory viruses (e.g., influenza virus, RSV, and hMPV) occur in winter. Typically, there is one dominant virus sweeping through a local community at any one time, a pattern that suggests some population-level interference with transmission. However, outbreaks can be closely spaced, and co-circulation of different viruses or antigenically diverse strains of one virus does occur. In the United States, some regional differences in seasonality have been noted; for example, RSV often appears in Florida and other southeastern states first. Seasons are, of course, reversed in the Northern and Southern hemispheres, so that winter epidemics occur roughly from November to March in the United States but from April to August in Australia; therefore, “winter” epidemics are almost always occurring somewhere in the world. Seasonal variances differ in the tropics, where acute respiratory viral infections are more common in the rainy season.

RISK FACTORS FOR DISEASE

Infection with these viruses is nearly universal, but disease expression varies among individuals infected with identical viruses. Therefore, investigators have sought to identify risk factors for severe disease. Most single risk factors identified have a moderate effect on the incidence of severe disease, but an accumulation of factors is associated with high risk. Underlying lung disease is a major factor, especially diseases associated with the need for chronic oxygen supplementation. COPD is one of the most profound risk factors. Other severe underlying medical conditions, especially cardiovascular disease, also enhance risk. Smoking (or exposure to wood smoke), low socioeconomic status, and male gender all contribute to a minor increase in the risk of lower respiratory tract illness. Close exposure to infected people is a major factor. For instance, living in close quarters (e.g., housing for military trainees, college dormitories, or nursing homes) puts groups of individuals at risk for rapid outbreaks. The U.S. military has instituted an adenovirus vaccination program to prevent severe or fatal adenovirus respiratory infections that can occur during outbreaks when new recruits are brought together. A breakdown in isolation and hand-washing compliance procedures can lead to cycles of nosocomial transmission of infection in hospital inpatient wards and intensive care units. In assessments of severe lower respiratory tract illness, a history of travel to an area with unusual agents should be considered carefully (e.g., exposure to avian influenza outbreaks in Asia, exposure to MERS-CoV in the Middle East).

TRANSMISSION

Respiratory viruses are transmitted by two principal modes: fomites or large-particle aerosols of respiratory droplets spread directly from person to person by coughing or sneezing. Fomite transmission occurs indirectly when infected respiratory droplets are deposited on the hands or on inanimate objects and surfaces, with subsequent transfer of secretions to a susceptible person's nose or conjunctiva. Most respiratory viruses do not spread by small-particle aerosols across rooms or down halls, although measles virus and VZV do spread in this manner. Therefore, contact and droplet precautions are sufficient to prevent transmission in most settings; hand washing is especially critical in health care settings during the winter.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Common Viral Respiratory Infections

The principal interventions that make a difference in the care of patients with acute respiratory virus infections are supportive, and these factors should be managed meticulously. Hypoxia is managed with supplemental oxygen and respiratory failure with mechanical ventilation. Because the tachypnea and fever that often accompany pneumonia and wheezing frequently result in dehydration, fluid management is important. The astute clinician can narrow the etiologic possibilities on the basis of epidemiologic knowledge; information about viruses circulating in the community (widely available from local reference laboratories, county and state health departments, and the U.S. Centers for Disease Control and Prevention [CDC]); and the patient's exposure history, age, and immunologic status, including vaccination status. Proper use of rapid diagnostic tests is important. When diagnostic tests are applied only to samples from individuals at high risk of exposure to an infectious agent in the appropriate season, the positive predictive value of the test is increased. A central medical decision is whether or not to use a specific antibacterial or antiviral agent to treat a respiratory infection. Antibiotics do not improve the outcome of uncomplicated respiratory virus infections in otherwise healthy subjects. Some viral infections, especially influenza, can be complicated by secondary bacterial infection. There are only a limited number of licensed antiviral drugs, which should be used when a specific viral etiology is determined. Antiviral treatment generally is effective only when administered early in the course of illness.

CLINICAL MANIFESTATIONS

The common cold is characterized by nasal congestion, sneezing, rhinorrhea, cough, and sore throat. Laryngitis is accompanied by hoarseness or dysphonia. Acute bronchitis is characterized by a dry or productive cough of <3 weeks' duration (most prevalent in winter) in the absence of signs and symptoms of pneumonia and of evidence of pneumonia on chest radiography and is primarily caused by viruses. Bacteria play a more prominent role in chronic bronchitis. Bronchiolitis is an acute illness with wheezing and evidence of upper respiratory infection, most commonly seen in the winter in infants and young children. The typical clinical manifestations of acute pneumonia include cough, sputum production, dyspnea, and chest pain. More systemic signs and symptoms also occur in pneumonia, including fever, fatigue, sweats, headache, myalgia, and occasionally nausea, abdominal pain, and diarrhea.

DIAGNOSIS

The clinical diagnosis of a respiratory syndrome and the anatomic location of infection is based on history, physical examination, and radiography. A specific viral etiology can be determined by specific diagnostic tests. The gold standard for diagnosing a respiratory viral infection is virus isolation, performed by inoculation of cell cultures with fresh secretions and use of multiple cell types in a reference laboratory staffed by experienced technologists. Direct or indirect fluorescent antibody detection can be used to visualize virus-infected cells in nasal secretions. Rapid antigen-based diagnostic tests are used to detect influenza virus or RSV proteins in nasopharyngeal secretions. The most sensitive tests typically are RT-PCR molecular diagnostic tests that amplify and detect the presence of viral genomic RNA or DNA in respiratory secretions. Multiplex panels assaying a sample for a dozen or more common respiratory viruses are available. These tests must be used and interpreted carefully because of their extreme sensitivity. If care is not taken, it is relatively easy to contaminate a PCR test in the laboratory with small amounts of DNA from a previous reaction. In addition, because a viral genome can sometimes persist in nasal secretions for weeks after an infection resolves, a positive test may indicate a recently resolved rather than a currently acute infection. Despite these limitations, PCR tests generally are considered the most sensitive and specific tests available. Chest radiographs should be obtained for all patients with suspected pneumonia.

TREATMENT

TREATMENT

Common Viral Respiratory Infections

NFLUENZA

A number of drugs are licensed in the United States for the treatment or prophylaxis of influenza (See also Chap. 195). Neuraminidase inhibitors act on both influenza A and B viruses by serving as transition-state analogs of the viral neuraminidase that is needed to release newly budded virion progeny from the surface of infected cells. The cell surface normally is coated heavily with the viral receptor sialic acid. Oseltamivir is administered orally and is effective for the prevention or treatment of uncomplicated influenza in otherwise healthy adults. Observational studies indicate that oseltamivir also may be beneficial during serious illness. The drug is generally well tolerated, with primarily gastrointestinal toxicity. Zanamivir, a powder that is administered through oral inhalation, exhibits effectiveness similar to that of oseltamivir. Moreover, zanamivir is active against some influenza virus strains that are resistant to oseltamivir. Inhalation of zanamivir powder may cause bronchospasm in patients with COPD or asthma. Peramivir is a newer drug that is administered intravenously as a single 600-mg dose. It is efficacious in acute, uncomplicated influenza and is approved for treatment of individuals who cannot take oral or inhaled medications. Its efficacy in severe influenza requiring hospitalization has not yet been demonstrated. Laninamivir is a new drug that is approved in Japan for prophylaxis and treatment of influenza. It is a polymeric zanamivir conjugate that is delivered by oral inhalation, and it exhibits greater potency and longer retention times than conventional zanamivir. The adamantanes amantadine and rimantidine have been used for the treatment of influenza A

infection. These drugs interfere with the ion channel activity caused by the M2 protein of influenza A viruses, which is needed for viral particle uncoating after endocytosis. These agents were commonly used in the past, but widespread resistance has been found in many currently circulating influenza A viruses. Therefore, the adamantanes should not be used unless isolate sensitivity is demonstrated, and, in many influenza seasons, the CDC advises against their use. When they are used, they are administered orally and display efficacy against uncomplicated influenza A caused by susceptible strains. The effectiveness of these drugs in serious illness has not been established. Toxicity with rimantadine generally manifests as gastrointestinal intolerance. Toxicity with amantadine is primarily associated with central nervous system symptoms.

RSV INFECTION

Ribavirin is a nucleoside antimetabolite prodrug whose activation by kinases in the cell results in a 5′-triphosphate nucleotide form that inhibits RNA replication. The drug was licensed in an aerosol formula in the United States in 1986 for treatment of children with severe RSV-induced lower respiratory tract infection. The efficacy of aerosolized ribavirin therapy remains uncertain despite a number of clinical trials. Most centers use it infrequently, if ever, in otherwise healthy infants with severe RSV disease. Intravenous ribavirin has been used for adenovirus, hantavirus, measles virus, PIV, and influenza virus infections, although a good risk/benefit profile has not been clearly established for any of these uses.

OTHER VIRAL TARGETS

Pleconaril, an oral drug with good bioavailability for treatment of infections caused by picornaviruses, has been tested for treatment of rhinovirus infection. This drug acts by binding to a hydrophobic pocket in the VP1 protein and stabilizing the protein capsid, preventing release of viral RNA into the cell. Pleconaril reduces mucus secretions and other symptoms and is being further examined for this indication. Acyclovir and related compounds are guanine-analog antiviral drugs used in the treatment of herpesvirus infections. HSV stomatitis in immunocompromised patients is treated with famciclovir or valacyclovir, and immunocompetent patients with severe oral disease compromising oral intake are sometimes treated with these agents. These compounds have also been used prophylactically to prevent the recurrence of outbreaks, with mixed results. Intravenous acyclovir is effective against HSV or VZV pneumonia in immunocompromised patients. Ganciclovir, given together with human immunoglobulin, may reduce the mortality rates associated with CMV pneumonia in hematopoietic stem cell transplant recipients and has been used as monotherapy in other patient groups. Cidofovir is a nucleotide analog with activity against a large number of viruses, including adenoviruses. Intravenous cidofovir has been effective in the management of severe adenoviral infection in immunocompromised patients but may cause serious nephrotoxicity.

COMPLICATIONS: CO-INFECTIONS

Co-infections with two or more viruses can occur because of the overlap in the winter season of these viruses in temperate areas. In general, in careful studies using cell culture techniques for virus isolation, two or more viruses were isolated from respiratory secretions of otherwise healthy adults with acute respiratory illness in ~5–10% of cases. There is little evidence that more severe disease occurs during co-infections. The incidence

of positive results in two molecular diagnostic tests (generally RT-PCR for these RNA viruses) is expected to be higher than that of culture because, as discussed above, molecular tests can remain positive for an extended period after shedding of infectious virus has ended.

PREVENTION

VACCINES

Numerous vaccines against influenza viruses have been licensed. In the United States, trivalent and quadrivalent inactivated intramuscular vaccines (covering H3N2, H1N1, and one or two B antigens) and a live attenuated trivalent vaccine for intranasal administration are available, although in 2017 the CDC stopped recommending the latter vaccine. Vaccines are effective when the vaccine strains chosen for inclusion are highly related antigenically to the epidemic strain, but occasional antigenic mismatches cause negligible efficacy of a vaccine component. Antigenic drift caused by point mutations in the H and N molecules leads to antigenic divergence, requiring the production of new vaccines each year. The segmented influenza genome allows reassortment of two viruses during co-infection of one individual or animal; sometimes the consequence is a major antigenic shift resulting in a pandemic. On average, pandemics occur every 20–30 years. There is current concern about the potential for an H5N1 or H7N9 pandemic, and experimental vaccines are being tested for these viruses.

Vaccines were developed for adenovirus serotypes 4 and 7 and were approved for prevention of epidemic respiratory illness among military recruits. Essentially, these vaccines consisted of unmodified viruses given by the enteric route in capsules instead of by the respiratory route—the natural route of infection leading to disease. Inoculation by the altered route resulted in an immunizing asymptomatic infection. Most U.S. military recruits are vaccinated against adenovirus, and epidemic disease recurs in the absence of vaccination.

Live attenuated and subunit vaccine candidates against RSV are under development and are being tested in clinical trials. Subunit RSV vaccines are being tested for maternal immunization and in the elderly. There are no licensed vaccines against rhinoviruses; as there is little or no cross-protection between serotypes, it will be challenging to develop a vaccine covering >100 serotypes. Efforts to develop coronavirus vaccines are in the preclinical stage.

PASSIVE PROTECTION WITH IMMUNOTHERAPY

Palivizumab, a humanized mouse monoclonal antibody to the F protein of RSV, is licensed for prevention of RSV hospitalization in high-risk infants, in half or more of whom it is effective. Experimental treatment of both immunocompetent and immunocompromised RSV-infected individuals has been reported, but the efficacy of this approach has not been established. Next-generation antibodies with higher potency and an extended half-life of ~90 days are being tested.

ISOLATION PROCEDURES, PERSONAL PROTECTIVE EQUIPMENT, AND HAND WASHING

Most respiratory viruses are spread by direct contact—i.e., body-surface to body-surface contact and physical transfer of microorganisms between a susceptible person and an infected person. Poor hand hygiene is probably the most common cause of contact transmission of viruses, which occurs often in family, school, and workplace settings. Transmission between health care workers and patients also takes place when hand-washing compliance is low. Fomites (objects or substances capable of carrying infectious organisms), including instruments, stethoscopes, and other objects in medical environments, can contribute to transmission. Airborne transmission can occur but is probably not the dominant mode of transmission for most respiratory viruses. Particle size affects the epidemiology of airborne pathogens. The composition and size distribution of the generated particles affect the duration of suspension of the infectious agents in the air, the distance across which they can be transported, the interval during which the virus remains infectious, and the site of deposition in the airway of a susceptible host. Direct exposure to large-particle aerosols (e.g., exposure at close range—up to 3 ft—to a cough or sneeze) causes some transmission. Particles of small size can remain suspended in the air for long periods; for instance, particles of ~1 μm can remain suspended for hours. However, in general, only a few respiratory viruses are thought to be transmitted by small-particle aerosols. Protection from transmission in health care environments can be achieved by proper implementation of and adherence to established procedures for the appropriate level of precaution.

Standard and Contact Precautions

Standard precautions, the basic level of infection control that is used in the care of all patients at all times, reduces the risk of transmission of viruses from respiratory tract secretions and mucous membranes. Contact precautions, the second level, require a single room for the patient when possible and the use of additional personal protective equipment, including the wearing of clean, nonsterile gloves when touching a patient or coming into contact with secretions. Fluid-resistant nonsterile gowns are used to protect skin and clothing during activities where contact with secretions is anticipated, and providers should wear each gown for the care of only one patient. A face mask is used when there is potential for direct contact with respiratory secretions. Eye protection (goggles or face shields) is worn in anticipation of potential splashing of respiratory secretions. Good hand hygiene should always follow any patient contact, including washing for 20 seconds with soap and warm water or cleaning with an alcohol-based hand rub. Providers should attempt to avoid the contamination of clothing and the transfer of microorganisms to other patients, surfaces, or environments.

Droplet Precautions

Large-particle droplets are generated during sneezing and coughing and during the performance of some medical procedures, such as airway suctioning in critical care units or bronchoscopy. Such droplets may contain viruses, but their range is usually limited to about 3 ft. Transmission of large-particle droplets occurs when they are deposited on the nasal mucosa or conjunctivae. To prevent transmission in these settings, providers should implement droplet precautions. They should wear a face mask, such as a surgical mask, for close contact (within 3 ft of the patient). Patients also should wear a face mask when exiting the examination room and should avoid coming into close contact with other patients.

Airborne Precautions

Airborne transmission occurs through the dissemination of airborne droplet nuclei (particles of ≤5 µm) or evaporated droplets containing viruses that can remain suspended in the air for long periods. Certain viruses that are carried by the airborne route can be inhaled by a susceptible host in the same room or over a long distance from the source patient, depending on environmental factors such as temperature and ventilation. Viruses transmitted by this route are SARS-CoV, measles virus, and VZV. Patients with these infections should be managed with personal respiratory protection and special ventilation and air handling. Providers should wear an N95 respirator selected with fit-testing, which must be repeated annually. Powered air-purifying respirators (PAPRs) are used in some cases. The patient should be housed in an airborne-infection isolation room—a negative-pressure room that has a minimum of six air exchanges per hour and exhausts through high-efficiency particulate air (HEPA) filtration or directly to the outside.

GLOBAL CONSIDERATIONS

HENDRA AND NIPAH VIRUSES

 These emerging paramyxoviruses, which are grouped in their own new genus (*Henipavirus*), may not be respiratory pathogens in a conventional sense, but they probably infect humans by the respiratory route. Nipah virus is a newly recognized zoonotic virus, named after the location in Malaysia where it was first identified in 1999. It has caused disease in humans who have had contact with infectious animals. Hendra virus (formerly called equine morbillivirus) is another closely related zoonotic paramyxovirus and was first isolated in Australia in 1994. The viruses have caused only a few localized outbreaks, but their wide host range and ability to cause high mortality raise concerns for the future. The natural host of these viruses is thought to be a certain species of fruit bat present in Australia and the Pacific. Pigs may be an intermediate host for transmission to humans in Nipah infection and horses in Hendra infection. Although the mode of transmission from animals to humans is not defined, inoculation of infected materials onto the respiratory tract probably plays a role. The clinical presentation usually appears to be an influenza-like syndrome that progresses to encephalitis, includes respiratory illness, and causes death in about half of identified cases.

BUNYAVIRIDAE: HANTAVIRUS

Intermittent outbreaks of hantavirus infection occur in South America and cause a severe lung infection: HPS. In addition, more than 400 cases of HPS have been reported in the United States. The disease was first recognized during an outbreak in 1993. About one-third of recognized cases end in death. The Four Corners outbreak (at the intersection of the northwestern corner of New Mexico, the northeastern corner of Arizona, the southeastern corner of Utah, and the southwestern corner of Colorado) is well known; however, cases now have been reported in a total of 32 states. Patients with HPS usually present with an influenza-like illness, including fever. Findings on physical examination are nonspecific, often consisting only of fever and elevated respiratory and heart rates. In addition to respiratory symptoms, abdominal pain is common.

Diagnosis is often delayed until illness becomes severe, at which point intubation and mechanical ventilation may be required for respiratory support.

SUMMARY

Viruses are the leading causes of acute lower respiratory tract infection in most populations. Influenza virus and RSV are the most common pathogens; hMPV, PIV3, and rhinoviruses account for most other acute viral respiratory infections. Infection in otherwise healthy adults generally leads to partial immunity to these pathogens, with protection against severe lower respiratory disease. However, reinfection, with upper respiratory tract illness, is common throughout life. Special populations such as immunocompromised patients, institutionalized frail elderly patients, and patients with COPD are at highest risk for severe disease.

FURTHER READING

Branche AR et al: Detection of respiratory viruses in sputum from adults by use of automated multiplex PCR. J Clin Microbiol 52:3590, 2014.

Englund JA et al: Common emergence of amantadine- and rimantadine-resistant influenza A viruses in symptomatic immunocompromised adults. Clin Infect Dis 26:1418, 1998.

Falsey AR et al: Bacterial complications of respiratory tract viral illness: A comprehensive evaluation. J Infect Dis 208:432, 2013.

Fendrick AM et al: The economic burden of non-influenza-related viral respiratory tract infection in the United States. Arch Intern Med 163:487, 2013.

Fry AM et al: Seasonal trends of human parainfluenza viral infections: United States, 1990–2004. Clin Infect Dis 43:1016, 2006.

Johnston SL et al: The relationship between upper respiratory infections and hospital admissions for asthma: A time-trend analysis. Am J Respir Crit Care Med 154:654, 1996.

Monto AS, Cavallaro JJ: The Tecumseh study of respiratory illness. II. Patterns of occurrence of infection with respiratory pathogens, 1965–1969. Am J Epidemiol 94:280, 1971.

Walsh EE et al: Clinical impact of human coronaviruses 229E and OC43 infection in diverse adult populations. J Infect Dis 208:1634, 2013.

Williams JV et al: Human metapneumovirus infection plays an etiologic role in acute asthma exacerbations requiring hospitalization in adults. J Infect Dis 192:1149, 2005.

Williams JV et al: The role of human metapneumovirus in upper respiratory tract infections in children: A 20-year experience. J Infect Dis 193:387, 2006.

Chapter 195: Influenza

Peter F. Wright

DEFINITION

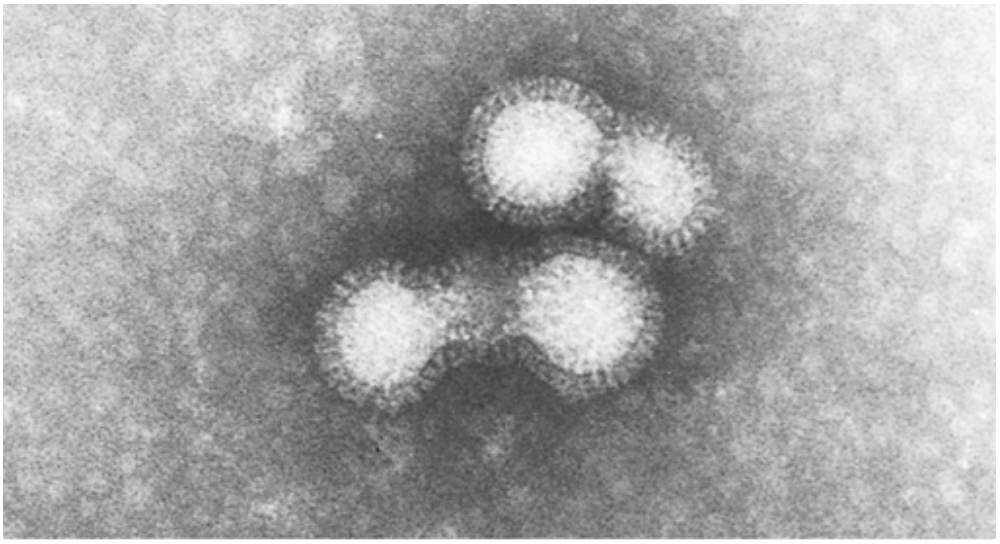
The term *influenza* represents both a clinically defined respiratory illness accompanied by systemic symptoms of fever, malaise, and myalgia and the name of the orthomyxoviruses that cause this syndrome. Although this term is sometimes used more generally to denote any viral respiratory illness, many features distinguish influenza from these other illnesses, most particularly its systemic symptoms, its propensity to cause sharply peaked winter epidemics, and its capacity to spread rapidly among close contacts. The morbidity and mortality associated with influenza epidemics are documented closely in the United States by the Centers for Disease Control and Prevention (CDC), which records clinical cases of influenza-like illness, cases of virologically documented influenza, and excess deaths due to pneumonia and influenza combined.

ETIOLOGIC AGENTS

Three influenza viruses occur in humans: A, B, and C. These viruses are irregularly circular in shape, measure 80–120 nm in diameter, and have a lipid envelope and prominent spikes that are formed by the two surface glycoproteins, hemagglutinin (H) and neuraminidase (N) **(Fig. 195-1)**. The hemagglutinin functions as the viral attachment protein, binding to sialic acid receptors on the cells that line the superficial epithelium of the respiratory tract. The neuraminidase cleaves the virus from the cell membrane to facilitate its release from the cell and prevents self-aggregation of viruses. Influenza A viruses have eight single-strand negative-sense RNA segments in their genomes that encode hemagglutinin and neuraminidase as well as internal genes, including polymerase, matrix, nucleoprotein, and nonstructural genes. The segmented nature of the genome allows gene *reassortment*; an analogy for reassortment is the shuffling of a deck of cards. Reassortment takes place when a single cell is infected with two different strains.

FIGURE 195-1

An electron micrograph of influenza A virus (x40,000). *(Reprinted with permission from YZ Cohen, R Dolin: Influenza, in Harrison's Principles of Internal Medicine, 19th ed. DL Kasper et al [eds]. New York: McGraw-Hill, 2015, p 1209.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Among the influenza viruses, the A viruses are of paramount importance for several reasons: (1) the plasticity of their genomes, which enables them to react to the prevailing immunity in the community by modifying their immunogenic epitopes, particularly on the hemagglutinin surface protein (*antigenic drift*); (2) the segmentation of their genomes, which allows genes coding both surface and internal proteins to be reassorted between influenza A variants (*antigenic shift*); and (3) their extensive mammalian and avian reservoirs, in which multiple variants with distinct hemagglutinin and neuraminidase genes lie in wait. As a result of all of these factors, influenza A virus has the ability, particularly after an antigenic shift, to cause a worldwide epidemic (*pandemic*). The most severe influenza A pandemic in modern history took place in 1918; ~50 million deaths were attributed to the culpable influenza A H1N1 virus in the years surrounding 1918.

The influenza A viruses are further classified by their surface glycoproteins (H and N), the geographic location of their isolation, their sequential number among isolated viruses, and their year of isolation. Thus, the influenza vaccine for the 2017–2018 season in the Northern Hemisphere was formulated to provide protection against influenza A/Michigan/45/2015 (H1N1)pdm09–like virus, influenza A/Hong Kong/4801/2014 (H3N2)–like virus, and two lineages in the influenza B family: B/Brisbane/60/2008–like virus (Victoria lineage) and B/Phuket/3073/2013–like virus (Yamagata lineage).

EPIDEMIOLOGY

Influenza virus causes outbreaks during the cooler months of the year and thus has a mirror-image season in the antipodes compared with that in the Northern Hemisphere. The circulation of strains in the Southern Hemisphere has some predictive value for vaccine composition in the Northern Hemisphere, and vice versa. This information is important as the degree of antigenic drift is one determinate of vaccine efficacy. Vaccine composition typically must change in at least one component yearly in anticipation of the predicted circulating strains.

A typical outbreak begins in early winter and lasts 4–5 weeks in a given community, although its impact on the country as a whole will be of considerably longer duration. When excess mortality occurs, an influenza outbreak

is classified as an *epidemic*. Influenza's impact is reflected in increased school and work absenteeism, increased visits to emergency rooms and primary care physicians, and increased hospitalizations, particularly of elderly patients and individuals with underlying cardiopulmonary disease. The impact often is most easily recognized in the pediatric population, whose school absenteeism quickly peaks. Despite efforts to limit influenza spread through vaccination, cohorting, use of masks, and hand washing, long-term-care facilities house another sentinel population, including many elderly patients who are at increased risk of complicated disease.

Influenza is largely spread by small- and large-particle droplets; spread is undoubtedly facilitated by the coughing and sneezing that accompany the illness. Within families, the illness is often introduced by a preschool or school-aged child.

 Influenza's global spread and causative strain(s) in a given year are well documented by the surveillance networks of the World Health Organization (WHO) and the CDC. The severity of an epidemic depends on the transmissibility and virulence of the viral strain, the susceptibility of the population, the adaptation of the virus to its human host, and the degree of antigenic match to the recommended vaccine. None of these parameters is totally predictable for influenza A.

Influenza A Viruses

When a major shift in the hemagglutinin and/or the neuraminidase occurs, with introduction of a new serotype from an animal or avian reservoir, an influenza A strain has the potential to cause a pandemic. In modern influenza history, such shifts occurred in 1918 (H1N1), 1957 (H2N2), 1968 (H3N2), 1977 (H1N1), and 2009 (H1N1pdm) **(Table 195-1)**. On the basis of seroarchaeology (the analysis of serum antibody profiles in the elderly), epidemics that took place in the 1890s have been attributed to H3N2 and H2N2 viruses. Epidemics typical of influenza have been documented throughout recorded history.

TABLE 195-1

Emergence of Antigenic Subtypes of Influenza A Virus Associated with Pandemic or Epidemic Disease

Years	Subtype	Extent of Outbreak
1889–1890	H2N8[a]	Severe pandemic
1900–1903	H3N8[a]	? Moderate epidemic
1918–1919	H1N1[b] (formerly HswN1)	Severe pandemic
1933–1935	H1N1[b] (formerly H0N1)	Mild epidemic
1946–1947	H1N1	Mild epidemic
1957–1958	H2N2	Severe pandemic
1968–1969	H3N2	Moderate pandemic
1977–1978[c]	H1N1	Mild pandemic
2009–2010[d]	H1N1	Pandemic

[a]As determined by retrospective serologic survey of individuals alive during those years ("seroarchaeology").

[b]Hemagglutinins formerly designated as Hsw and H0 are now classified as variants of H1.

[c]From this time until 2016–2017, viruses of the H1N1 and H3N2 subtypes circulated in alternating years or concurrently.

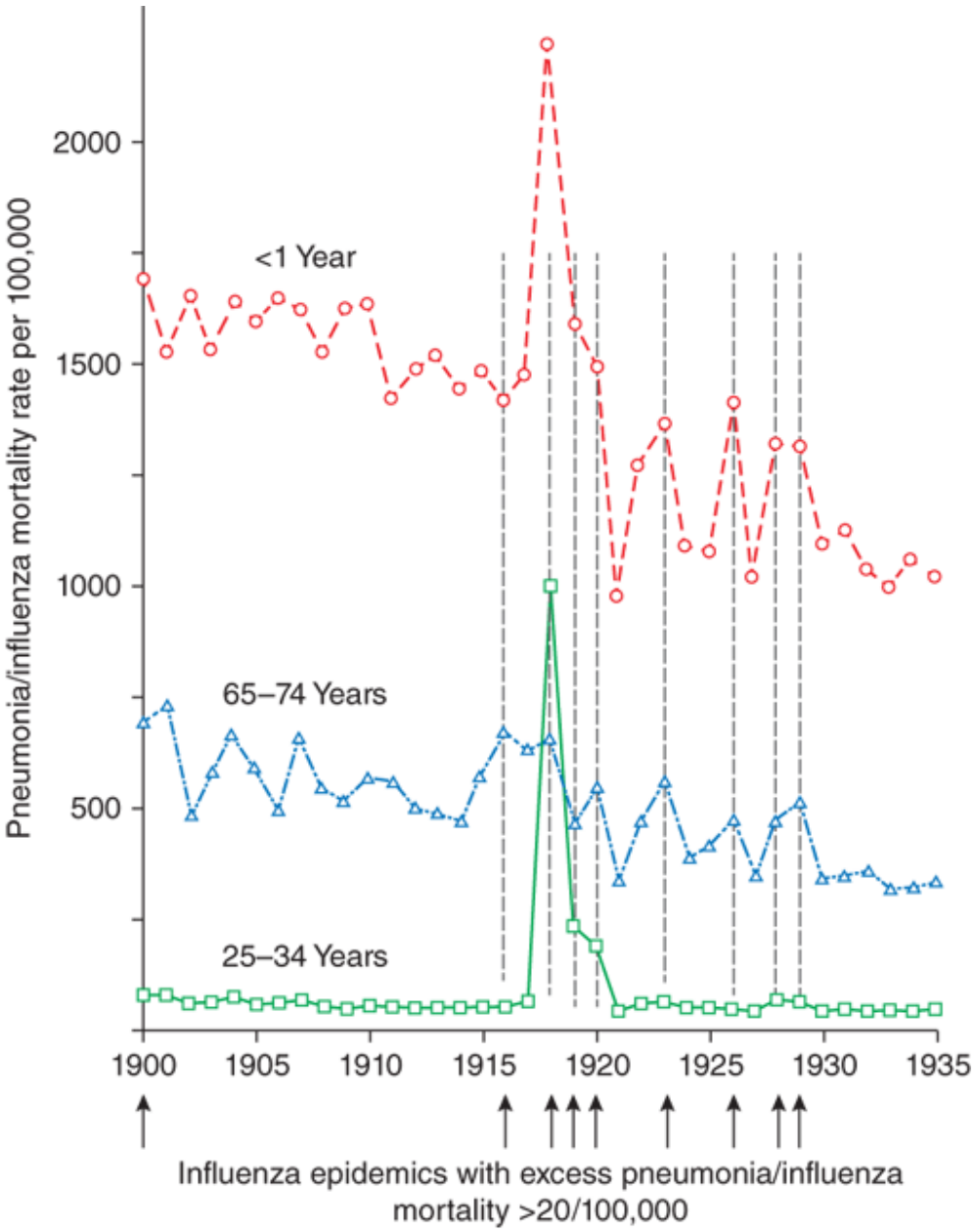
[d]A novel influenza A/H1N1 virus emerged to cause this pandemic.

Source: Adapted from YZ Cohen, R Dolin: Influenza, in *Harrison's Principles of Internal Medicine*, 19th ed. DL Kasper et al (eds). New York, McGraw-Hill, 2015, p 1209.

In some epidemics, a younger age group proves especially susceptible. This is the case with current H1N1 epidemics, where individuals born before 1968 had likely been exposed to related viral strains and thus were relatively protected against the current strain. The 1918 epidemic was striking in this regard: the most severely infected individuals were infants and previously healthy young adults—the latter being a group not typically found to have high influenza mortality **(Fig. 195-2)**. The 1918 epidemic increased all-cause mortality and led to more deaths than all military losses in World War I. In spite of the attention paid to the risk and impact of pandemic disease, it is generally appreciated that—with the exception of 1918—cumulatively more illness occurs during yearly epidemics combined than in pandemics.

FIGURE 195-2

Excess pneumonia/influenza deaths in 1900–1953, demonstrating the dramatic peaks of deaths among young infants and young adults (25–34 years of age). *(Data are from public health records collated by the author.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

All of the annual influenza A epidemics in the past 50 years have been caused by H1N1 and/or H3N2 strains. H2N2 strains circulated between 1957 and 1968, and H1N1 strains circulated prior to that, including in 1918. However, potentially pandemic viruses continue to emerge, mostly in Asia, with higher-numbered hemagglutinins (e.g., H5, H6, H7, H9) reflecting some of the 16 distinct H and 9 distinct N subtypes in avian reservoirs. Most cases of these potentially pandemic illnesses have occurred in individuals who have had direct contact with domesticated birds or who have visited live-bird markets, which are common in Asia. In addition to the global aeronautic movement of infected people, bird migration is one mechanism for rapid global spread. It is not clear why higher-numbered avian hemagglutinin strains have not acquired the degree of transmissibility necessary to cause pandemic disease.

Avian and Swine Influenza Viruses

The full panoply of influenza viruses is found in domestic and migratory wild birds. It is postulated that epithelial cells in the swine respiratory tract may play a specific role as a “mixing vessel,” allowing the reassortment of genes from avian and human sources and thereby permitting the transition of avian viruses to humans. The nature of the sialic acid receptors for influenza virus hemagglutinin partially accounts for host preference. Humans have largely α-2,6-galactose receptors, while birds have α-2,3-galactose receptors. Swine have both types of receptors on their respiratory epithelial cells—hence their postulated role in facilitating reassortment and host adaptation of avian strains to growth in humans. Strains such as 2009 H1N1pdm (pandemic) had genes of avian, swine, and human origin. Some avian strains—notably H5 strains—are highly pathogenic in humans, as was the 1918 strain. The reasons for the high pathogenicity of certain strains are not entirely clear. Virulence and transmissibility often appear to be separate genetic traits.

After the sequencing of the 1918 virus recovered from the lungs of bodies buried in the Arctic permafrost, the virus was genetically reconstructed under carefully controlled isolation conditions. In animal studies of this viable 1918 virus, both the hemagglutinin and the ribonucleoprotein contributed to high levels of replication accompanied by an abnormally enhanced innate immune response characterized by proinflammatory cytokines. Perhaps this “cytokine storm” is the best explanation for the enhanced illness occurring in young, immunologically vigorous individuals in the 1918 pandemic. Sequencing demonstrated that the 1918 virus was of avian origin. Although the 1918 virus was first identified in military camps in the United States, its impact cannot be attributed to the disruption of war: the illness was well documented in countries such as Iceland that were not directly involved in World War I.

The same concerns about a “cytokine storm” have been raised with regard to the H5N1 viruses that first emerged in Hong Kong in 1996. These viruses exhibited high pathogenicity in individuals who had direct contact with domestic fowl, with mortality rates close to 50%, but also displayed poor human-to-human transmissibility. Pathogenicity appears to be a function not just of the viruses’ surface proteins, but also of an optimal gene constellation including all eight segmented influenza genes. However, unlike the 1918 strain, the H5N1 viruses have, to date, caused only sporadic disease, as have other limited clusters of a highly pathogenic H7N9 virus.

Influenza B and C Viruses

The influenza B viruses are more genetically stable than the influenza A viruses and have no animal reservoirs. Two lineages of influenza B have circulated for the past 40 years (B/Yamagata-like and B/Victoria-like viruses), and it has proven very difficult to predict which strain will be dominant in a given year. This issue has led to the incorporation of representatives of both influenza B lineages plus influenza A/H1N1 and H3N2 viruses into a quadrivalent vaccine.

Influenza C viruses cause intermittent mild disease and have attracted little attention. These viruses have been the subject of fewer than 10 publications annually since the year 2000.

Influenza-Associated Morbidity and Mortality

Although epidemics vary in severity and in the age groups most affected, certain high-risk groups are seen in all epidemics **(Table 195-2)**. These groups are assigned the highest priority for vaccination and other preventive and therapeutic measures. Their caregivers and close contacts are also prioritized targets of interventions. A generalization is that the relative impact of an epidemic is seen in the youngest age group with the least prior exposure—and therefore the least immunity—to influenza. The impact of influenza can be depicted as a pyramid of illnesses, medical visits, hospitalizations, and deaths **(Fig. 195-3)**.

TABLE 195-2

High-Risk Groups Who Should be Assigned a High Priority for Influenza Immunization and Treatment[a]

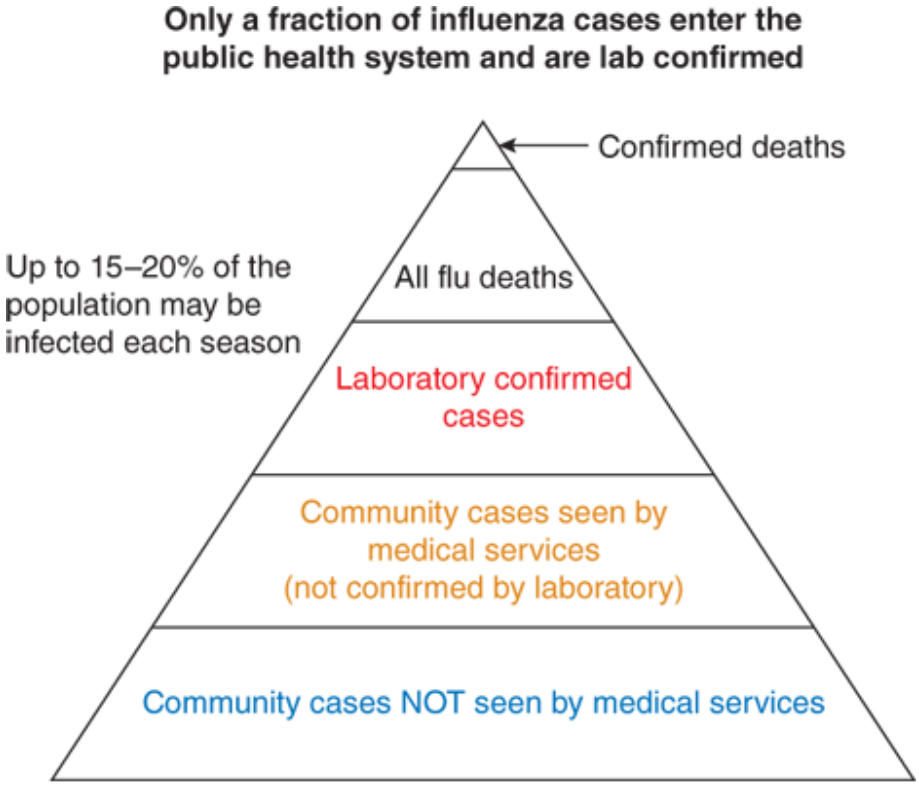
High-Risk Group
Children 6–59 months of age Adults ≥50 years of age Persons with chronic pulmonary (including asthma), cardiovascular (except isolated hypertension), renal, hepatic, neurologic, hematologic, or metabolic disorders (including diabetes mellitus) Persons who are immunocompromised (any cause, including medications or HIV infection) Women who are or plan to be pregnant during the influenza season Children and adolescents (6 months through 18 years of age) who are receiving aspirin- or salicylate-containing medications and who might be at risk for Reye syndrome Residents of nursing homes and other long-term-care facilities Native Americans, including Alaska Natives Persons who are extremely obese (BMI ≥40)
Contacts and Caregivers
Caregivers and contacts of those at risk: health care personnel in inpatient and outpatient care settings, medical emergency-response workers, employees of nursing home and long-term-care facilities who have contact with patients or residents, and students in these professions who have contact with patients Household contacts and caregivers of children ≤59 months (i.e., <5 years) of age (particularly contacts of infants <6 months old) and adults ≥50 years of age Household contacts and caregivers of persons who are in a high-risk group

[a]No hierarchy is implied by order of listing.

Source: Centers for Disease Control and Prevention 2017–2018 summary of recommendations for influenza vaccine (*https://www.cdc.gov/vaccines/hcp/acip-recs/vacc-specific/flu.html*).

FIGURE 195-3

Pyramid of impact of influenza illness. *(From http://4.bp.blogspot.com/-ZjKKbq-d5O4/VClasJsjOGI/AAAAAAAADVo/UdpfdzH1JYc/s1600/WHO-flu-pyramid.jpg; accessed October 15, 2017.)*

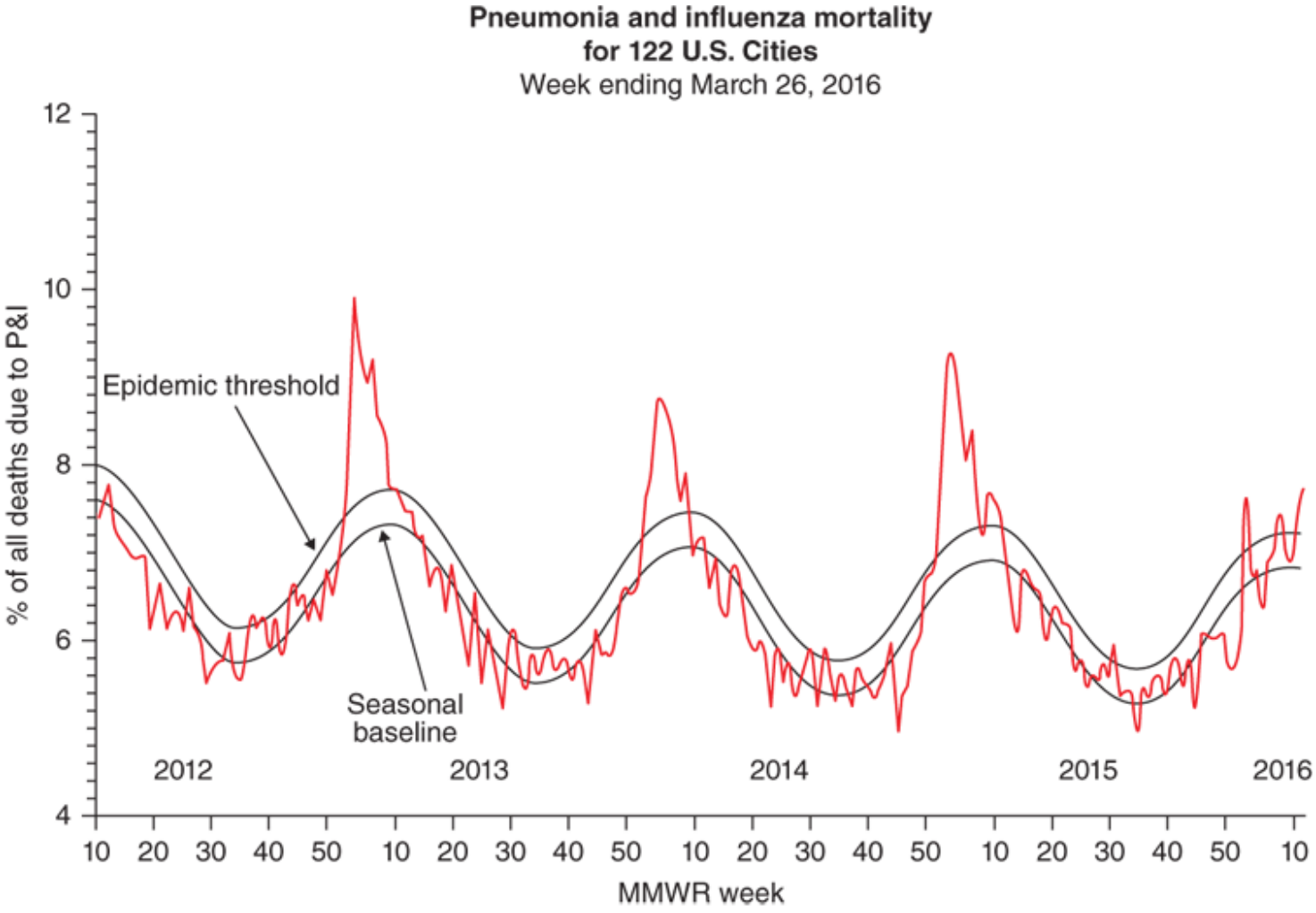


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pneumonia and influenza mortality, reported as excess over the anticipated sine-wave curve of deaths during the year, is seen in the CDC's data for 2012–2017 **(Fig. 195-4)**. In addition to excess respiratory deaths directly attributed to influenza, an increase in circulatory deaths also occurs during an influenza epidemic.

FIGURE 195-4

Morbidity and mortality attributable to influenza. P&I, pneumonia and influenza; MMWR, *Morbidity and Mortality Weekly Report*. *(From https://www.cdc.gov/flu/weekly/weeklyarchives2015-2016/images/122CMRS12_small.gif; accessed February 13, 2018.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOGENESIS AND IMMUNITY

At a cellular level, influenza virus binds to sialic acid receptors and enters the epithelial cell through receptor-mediated endocytosis. The virus then enters an endosome, where acidification promotes proteolytic cleavage of the hemagglutinin, exposing a fusion domain. The influenza hemagglutinin undergoes a marked structural reorganization in this cleavage step. Hemagglutinin cleavage may be one of the factors that restrict viral growth to epithelial cells, as a unique protease in the respiratory milieu is required for this cleavage to occur. The fusion domain allows the viral RNA to enter the cytoplasm. The nucleoprotein is transported into the nucleus of the cell, where transcription to a positive-sense RNA and replication take place. Viral proteins then assemble on the apical surface of the infected cell and, after incorporation of cellular membrane, bud from the membrane back into the mucosal milieu.

Influenza infection is initiated in the upper respiratory tract via aerosolized virus. The cells infected with influenza virus are primarily the ciliated cells of the respiratory tract. Denudation of the superficial epithelium probably accounts for much of the symptomatology and may predispose to secondary bacterial infections. The onset of symptoms follows an incubation period that, for a viral illness, is very short: 48–72 h. The infection spreads to the lungs but, even there, remains confined to the epithelial layer.

Uniquely among respiratory viruses, influenza virus is associated with systemic symptoms of fever, malaise, and myalgia. These manifestations are presumed to be mediated by cytokines, and excess cytokine production has been implicated in the acute toxicity of H5N1 and other highly pathogenic influenza viruses.

The immune response to influenza virus occurs at the systemic and mucosal levels and involves both T and B cells. The B cell responses are directed primarily toward antigenic epitopes on the two surface glycoproteins—i.e., hemagglutinin and neuraminidase. At a structural level, the four recognized epitopes on the hemagglutinin are largely confined to the globular head of the protein, which collectively constitute the targets for hemagglutination inhibition (HAI) antibodies. HAI and neutralizing antibodies are highly correlated; HAI antibody levels are used as a measure of susceptibility to clinical infection and thus as a measure of vaccine-induced protection. In a child or an adult without prior vaccination or with the emergence of a distinctly new strain, serum HAI antibody is a surrogate for protection. However, in individuals with both vaccine-induced and natural immunity, the protective efficacy of a vaccine based on serum HAI antibody is more difficult to predict.

Studies with improved collection methods and assays that more sensitively and reproducibly measure mucosal antibody suggest that mucosal neutralizing IgA antibody more accurately reflects susceptibility to infection. Perhaps the patterns of immune protection are best shown in a murine model, in which passively administered IgA antibody to influenza virus protects animals from initiation of infection and epithelial damage in the upper respiratory tract, while infused IgG antibody to the virus is protective in the lungs.

There is now considerable research interest in the induction and protective role of broadly neutralizing antibodies that recognize less antigenically variable regions on the stalk of the hemagglutinin. The results of these studies have led to talk of a universal influenza vaccine, although no such vaccines are yet available in clinical practice.

The role of T-cell immunity, which primarily recognizes internal protein epitopes, remains unclear in humans. However, T-cell immunity is thought to play a role in clearance of an influenza infection that quite reproducibly develops 8–10 days after exposure. A role for T cells in protection against acquisition of infection has also been proposed.

CLINICAL MANIFESTATIONS

Influenza is primarily a respiratory illness causing rhinorrhea, sore throat, conjunctivitis, and cough. The illness has a sudden onset and is epidemiologically linked to close contact with persons who have similar symptoms and often to community-wide respiratory illness. What distinguishes influenza from other respiratory illnesses is the degree of accompanying fever, fatigue, myalgia, and malaise. The symptoms typically begin within 48–72 h of exposure. The constellation of symptoms caused by an H3N2 viral strain, A/Port Chalmers 1/73, was followed prospectively in young seronegative children. Although these data involve children and a viral strain circulating 45 years ago, they present a representative picture of influenza today except that irritability in a young child is more specifically recognized as malaise, myalgia, and headache in an adult **(Table 195-3)**.

TABLE 195-3

Clinical Observations in 24 Seronegative Children Examined during Influenza A/Port Chalmers Infection

Condition/Event	No. of Patients
Coryza	22
Fever (temperature >38.4°C [>101°F])	21
Cough	21
Pharyngitis	20
Irritability	20
Fever (temperature >39.5°C [>103°F])	13
Anorexia	12
Tonsillitis	8
Vomiting	7
Otitis	6
Pneumonia	6
Diarrhea	6
Hoarseness	4
Croup	1

Respiratory symptoms, particularly recurrent cough, persist well beyond the 2–5 days of systemic symptoms. There is a postinfectious delay in return to normal levels of activity. Pulmonary function is persistently decreased after acute influenza. Persons with a regular exercise routine (e.g., runners) note a decrease from their prior level of performance that typically lasts for a month or more. In the elderly, the respiratory presentation may be less prominent, but there is often a decline in baseline activity and a loss of appetite.

The variation among epidemics in symptom severity must be stressed. The circulation of an early H1N1 virus, A/USSR/77, in young children followed prospectively was detected only by comparison of pre- and post-season serum samples, which documented that 31% of 195 children had experienced infection with this virus without

developing associated clinical illness. In the same population in the same year, influenza A/Texas/H3N2 virus infected 37% of children and was clearly associated with a typical influenza illness. In recent years, when H1N1 and H3N2 viruses have co-circulated, the latter has caused more severe epidemics.

On physical examination, the patient with influenza appears ill and rheumy, with sweating, coughing, nonpurulent conjunctivitis, and diffuse pharyngeal erythema. Pulmonary examination typically reveals nonlocalizing scattered rales, rhonchi, and wheezes. When present, localized pulmonary findings suggest relatively complicated pneumonia with a bacterial component. Muscle pain may be elicited by pressure, particularly in the calves and thighs. There are rare gastrointestinal findings. No rash is associated with influenza.

COMPLICATIONS

Complications of influenza occur most commonly in persons >65 years of age, those with underlying cardiopulmonary disease, those with immunosuppression, and women who are in the second or third trimester of pregnancy. In recent years, there has been mortality attributable to influenza among often previously healthy children <5 years of age in the United States, with ~100 deaths per year.

Respiratory Complications

Pneumonia characterized by progressive air hunger, localized pulmonary findings on physical examination, and radiographic findings of infiltrates and consolidation is the most common complication of influenza. Pneumonia in influenza can be primary influenza viral pneumonia, secondary bacterial pneumonia, or mixed viral and bacterial pneumonia. Primary viral pneumonia is characterized by increasing dyspnea, persistent fever, and—in more severe cases—cyanosis. Primary influenza pneumonia was typical in the 1918 pandemic and occurs with H5N1 virus, as initially described in Hong Kong in 1997. Pathologically, a marked inflammatory reaction in the alveolar septa is characterized by infiltration of monocytes, lymphocytes, and macrophages, with variable numbers of neutrophils. Destruction and hemorrhage are seen in the respiratory epithelium. Large amounts of virus can be recovered from the lungs.

In secondary bacterial pneumonia or mixed viral and bacterial pneumonia, illness may be biphasic, with evidence of recovery from the primary influenza illness followed by recrudescence of fever and pulmonary symptoms. Localizing findings may be detected on pulmonary examination and/or x-ray. The development of secondary bacterial infection is not surprising, as influenza de-epithelializes the airways and destroys ciliary function, allowing bacterial contamination. Another proposed mechanism for bacterial/viral enhancement is the production by *Staphylococcus* and *Pseudomonas* of proteases that enhance cleavage of the influenza hemagglutinin and thereby facilitate viral replication. The risk of secondary bacterial disease is greatest in elderly patients and those with chronic obstructive pulmonary disease.

Some influenza strains cause laryngotracheobronchitis or croup in children. Otitis media—a common accompaniment to influenza in children—may also be due to a combination of influenza virus and bacteria.

Extrapulmonary Complications

Although influenza is believed to spread only rarely beyond the respiratory epithelial cells, where unique endogenous proteases facilitate hemagglutinin cleavage and productive infection, this disease causes not only prominent systemic complaints but also a variety of extrapulmonary manifestations. The most common extrapulmonary manifestation of influenza is myositis, which is seen more often in influenza B and is characterized by severe muscle pain, elevated creatinine phosphokinase levels, and myoglobinuria that can lead to renal failure. The muscles are extremely tender to touch. Myo/pericarditis is seen less frequently and has been reported only in selected epidemics, notably the pandemic of 1918. However, a consistent epidemiologic link exists between influenza epidemics and excess cardiovascular hospitalizations.

Postinfectious acute demyelinating encephalomyelitis can follow influenza as well as other viral infections. The literature is mixed on the benefit and reliability of efforts to establish a polymerase chain reaction (PCR)–based diagnosis in this condition. MRI shows distinctive multifocal, symmetric brain lesions affecting the thalamus, brainstem tegmentum, cerebral periventricular white matter, and cerebellar medulla. Encephalitis and transverse myelitis may accompany influenza infection. Guillain-Barré syndrome can develop after influenza and was reported after a widespread influenza vaccination effort in the fall of 1976 that was undertaken in anticipation of a swine influenza epidemic (which never materialized). Until aspirin was recognized as a co-factor in its precipitation, Reye syndrome, an acute hepatic decompensation, was seen commonly in children and adolescents with influenza, particularly those infected with influenza B virus. Subsequently, the use of aspirin for fever control and symptom relief in children with viral infections was strongly discouraged, and Reye syndrome has virtually disappeared from clinical practice.

LABORATORY FINDINGS AND DIAGNOSIS

There is a strong argument for establishing a microbiologic diagnosis from both an individual-patient and a public-health perspective. This information is particularly valuable early in the season, when the extent of influenza and the precise circulating strain(s) are uncertain; in the management of complicated cases in hospitalized patients; and in settings such as long-term-care facilities and hospitals, where the institution of specific infection-control measures is appropriate.

Influenza virus is most easily recovered from nasopharygeal specimens. These samples are most effectively collected with a flocked swab that is inserted 1–2 inches into the nose (following the course of the inferior meatus), twirled, placed in viral transport medium that supports viral viability, and transported on ice to the laboratory as promptly as possible. The available rapid tests based on antigen detection vary in complexity and cost. Some are point-of-care tests, and others require laboratory support. These tests are highly specific but have a sensitivity of only 50–70%. Their sensitivity is strongly dependent on sample collection early in the course of illness—ideally within 48 h of the onset of symptoms. Traditionally, viruses have been isolated in tissue culture or with the use of embryonated eggs, and infection has been confirmed with hemadsorption or hemagglutination.

The most useful clinical approach today is to use a PCR-based molecular probe that amplifies specific segments of the influenza genome. Not only is this the most sensitive and specific method; it also provides opportunities to identify the strain with some specificity. Testing by multiplex PCR can simultaneously identify multiple respiratory pathogens—an advantage in the ill hospitalized patient. Serologic confirmation of infection is also

possible but requires paired serum samples, with the convalescent-phase sample obtained 2 weeks after infection. Mucosal antibody assays that are now being developed can detect strain-specific antibodies in paired mucosal specimens and yield insights into the importance of mucosal immunity in protection against influenza.

Other laboratory tests are of limited value. Mild leukopenia is seen in influenza, and a white blood cell count above 15,000/μL suggests a secondary bacterial component in influenzal pneumonia.

DIFFERENTIAL DIAGNOSIS

The distinctive nature of influenza is such that clinical diagnoses by experienced pediatric physicians are 85% concordant with microbiologic etiologic confirmation. Respiratory syncytial virus often co-circulates with influenza virus; it particularly affects the youngest children, causing bronchiolitis, but it can also infect the elderly, leading to an influenza-like nonspecific respiratory illness and a decline in mobility, nutrition, and pulmonary function, with resultant hospitalization.

PROPHYLAXIS

The major intervention to limit influenza illness is vaccination, which is conducted on a yearly basis because of variation in the predicted circulating strains for the coming year and which enhances existing immunity in advance of the influenza season. The decision about vaccine composition must be made ~10 months before the seasonal peak in influenza virus circulation; this decision is made by committees at the WHO and the CDC. This timing can result in a mismatch of vaccine composition with the viral strains that are actually prevalent in the upcoming season. The overall accuracy of the prediction is at least 70% for all strains in the recommended vaccine. Influenza vaccine is unique in being given seasonally in the months immediately preceding an outbreak. In the United States, vaccine is typically available from September.

The currently available vaccines are all based on purified subunit inactivated virus produced in eggs, in tissue culture, or through a baculovirus-expressed hemagglutinin protein. They are all calibrated to hemagglutinin content. Depending on the vaccine, they are administered intramuscularly or intradermally. Some influenza vaccines include MF-59 as an adjuvant or have elevated hemagglutinin content for enhanced immunogenicity in the elderly. The recommendations for use, the approved age range of each product, the route of administration, and the anticipated side effects are published annually by the CDC (*https://www.cdc.gov/vaccines/hcp/acip-recs/vacc-specific/flu.html*). The protection (and licensure) of all influenza vaccines depend on their stimulation of antibodies to the hemagglutinin; all are 50–75% effective at preventing clinical influenza. No effort is made to standardize the neuraminidase content. The contraindications to inactivated influenza vaccine administration are limited to individuals who have experienced a Guillain-Barré reaction within 6 weeks of a prior influenza vaccination. Egg allergy is not considered a contraindication to vaccination.

In recent years, a live, attenuated, intranasally administered vaccine (LAIV) has been used in children and has exhibited an efficacy exceeding that of injected inactivated vaccines. However, beginning in 2014–2015 and continuing in the following season, LAIV had no demonstrable efficacy assignable to the vaccine's H1N1 component. This change led advisory committees in the United States not to recommend the use of LAIV despite

its prior effectiveness, ease of administration, and theoretical advantage of stimulating mucosal immunity by the topical route. Active research is aimed at discerning the reason(s) for this lack of efficacy.

In the United States, the recommendation is that all individuals >6 months of age receive inactivated influenza vaccine yearly and that two doses of vaccine be given to children <9 years of age who are getting their first or second yearly vaccination. Groups at special risk of experiencing or transmitting influenza for whom influenza immunization is a particularly high priority are listed in Table 195-2.

Especially in hospital settings, considerable attention is paid to hand washing and the use of masks by persons with respiratory symptoms and those who are at particular risk of acquisition (typically, immunocompromised patients). Studies have demonstrated the benefit of face masks and hand hygiene in the hospital setting.

TREATMENT

TREATMENT

Influenza

Antiviral therapy for influenza has been limited by the paucity of available drugs, the short duration of symptoms in uncomplicated influenza, and the changing patterns of drug resistance in influenza viral strains. In the past, influenza A infection could be treated with the M-2 channel blockers amantadine and rimantadine. Widespread resistance has currently relegated these compounds to historical interest only.

The currently available class of drugs for treatment of influenza A and B viruses consists of neuraminidase inhibitors. As their name implies, these drugs act by inhibiting the influenza neuraminidase and thus limiting the egress of influenza virus from an infected cell. They are most effective in patients whose influenza illness is recognized early and confirmed by rapid antigen detection or on the basis of clinical and epidemiologic evidence. In experimental trials, these drugs hasten the resolution of symptoms if given within 48 h of infection. There are indications for their use both prophylactically—either throughout the season or, when a case is recognized in a close contact, in the short term—and therapeutically. The anticipated effect of early administration is the resolution of symptoms 1–2 days sooner than without treatment. The use of neuraminidase inhibitors is recommended for complicated influenza infections in hospitalized patients in the absence of formal proof of efficacy and when diagnosis may have been delayed. All the available neuraminidase inhibitors carry a risk of development of resistance, particularly with prolonged administration (e.g., to an immunodeficient individual with persistent recovery of influenza virus). Resistance to neuraminidase inhibitors is not widespread among currently circulating influenza A or B strains, but its development has been demonstrated in the laboratory, and clinical resistance could influence the utility of these drugs.

The defined risk groups who can benefit from neuraminidase inhibitors include children <2 years of age, adults >65 years of age, patients with chronic conditions, immunosuppressed individuals, pregnant women, women who have delivered infants ≤2 weeks previously, patients <19 years old who are receiving long-term aspirin treatment, Native Americans (including Alaska Natives), morbidly obese individuals, and residents of nursing homes or chronic-care facilities. This list resembles that of candidates whose vaccination is a high priority (Table

195-2). Use of neuraminidase inhibitors should be considered in selected high-risk cases despite a history of vaccination.

The available neuraminidase inhibitors are oral oseltamivir, nasal-spray zanamivir, and intravenous peramivir and zanamivir. Oseltamivir, which is most widely used, is an orally absorbed drug that is converted to its active component, oseltamivir carboxylate, in the liver. Gastrointestinal symptoms, especially nausea, may accompany the administration of oseltamivir. Because zanamivir is not orally bioavailable, it is given as an inhaled dry powder dispersed through a Diskhaler device.

The usual duration of therapy with either oral oseltamivir or intranasal zanamivir is 5 days, with twice-a-day dosing. Oseltamivir is preferred for treatment of pregnant women and is approved for children ≥1 year of age. Poor oral intake or absorption is a contraindication to the use of oseltamivir, although this drug can also be given by oro/nasal tube. Asthma and chronic obstructive pulmonary disease are relative contraindications to the use of intranasal zanamivir; this agent is approved for children ≥5 years of age. The use of the intravenous preparations of peramivir and zanamivir is indicated in severely ill patients. Peramivir is licensed for individuals >18 years of age, and intravenous zanamivir may be available through the manufacturer via an individual Emergency Investigational New Drug request. The most current recommendations and details on influenza antiviral drug use and release are available through the CDC (*https://www.cdc.gov/flu/professionals/antivirals/summary-clinicians.htm*).

Other critical aspects of treatment include maintenance of fluid and electrolyte balance, oxygen supplementation, fever control with nonsteroidal anti-inflammatory drugs, and treatment of suspected secondary bacterial complications with antibiotics. Appropriate respiratory isolation of patients should be practiced in accordance with local hospital guidelines.

FURTHER READING

Barry JM: *The Great Influenza: The Story of the Deadliest Pandemic in History*. New York, Penguin Books, 2005.

Fineberg HV: Pandemic preparedness and response—lessons from the H1N1 influenza of 2009. N Engl J Med 370:1335, 2014.

Jackson ML et al: Influenza vaccine effectiveness in the United States during the 2015–2016 season. N Engl J Med 377:534, 2017.

Kash JC, Taubenberger JK: The role of viral, host, and secondary bacterial factors in influenza pathogenesis. Am J Pathol 185:1528, 2015.

Krammer F: Novel universal influenza virus vaccine approaches. Curr Opin Virol 17:95, 2016.

Nichol KL et al: Influenza vaccination and reduction in hospitalizations for cardiac disease and stroke among the elderly. N Engl J Med 348:1322, 2003.

Treanor JJ: Influenza vaccination. N Engl J Med 375:1261, 2016.

Uyeki TM et al: Novel influenza A viruses and pandemic threats. Lancet 398:2172, 2017.

Watanabe T et al: 1918 influenza virus hemaggutinin (HA) and the viral RNA polymerase complex enhance viral pathogenicity, but only HA induces aberrant host responses in mice. J Virol 87:5239, 2013.

Wright PF et al: Correlates of immunity to influenza as determined by challenge of children with live, attenuated influenza vaccine. Open Forum Infect Dis 3(2):ofw108, 2016. [PubMed: 27419180]

Chapter 196: The Human Retroviruses

Dan L. Longo; Anthony S. Fauci

INTRODUCTION

The retroviruses, which make up a large family (Retroviridae), infect mainly vertebrates. These viruses have a unique replication cycle whereby their genetic information is encoded by RNA rather than DNA. Retroviruses contain an RNA-dependent DNA polymerase (a reverse transcriptase) that directs the synthesis of a DNA form of the viral genome after infection of a host cell. The designation *retrovirus* denotes that information in the form of RNA is transcribed into DNA in the host cell—a sequence that overturned a central dogma of molecular biology: that information passes unidirectionally from DNA to RNA to protein. The observation that RNA was the source of genetic information in the causative agents of certain animal tumors led to a number of paradigm-shifting biologic insights regarding not only the direction of genetic information passage but also the viral etiology of certain cancers and the concept of oncogenes as normal host genes scavenged and altered by a viral vector.

The family Retroviridae includes seven subfamilies **(Table 196-1)**. Members of two of the families infect humans with pathologic consequences: the deltaretroviruses, of which human T-cell lymphotropic virus (HTLV) type 1 is the most important in humans; and the lentiviruses, of which HIV is the most important in humans.

TABLE 196-1

Classification of Retroviruses: The Family Retroviridae

Genus	Example(s)	Feature
Alpharetrovirus	Rous sarcoma virus	Contains *src* oncogene
Betaretrovirus	Mouse mammary tumor virus	Exogenous or endogenous
Gammaretrovirus	Abelson murine leukemia virus	Contains *abl* oncogene
Deltaretrovirus	HTLV-1	Causes T-cell lymphoma and neurologic disease
Epsilonretrovirus	Walleye dermal sarcoma virus	Not known to be pathogenic in humans
Lentivirus	HIV-1, HIV-2	Causes AIDS
Spumavirus	Simian foamy virus	Not known to be pathogenic in humans

The wide variety of interactions of a retrovirus with its host range from completely benign events (e.g., silent carriage of endogenous retroviral sequences in the germline genome of many animal species) to rapidly fatal infections (e.g., exogenous infection with an oncogenic virus such as Rous sarcoma virus in chickens). The ability of retroviruses to acquire and alter the structure and function of host cell genetic sequences has revolutionized our understanding of molecular carcinogenesis. The viruses can insert into the germline genome of the host cell and behave as a transposable or movable genetic element. They can activate or inactivate genes near the site of integration into the genome. They can rapidly alter their own genome by recombination and mutation under selective environmental stimuli.

Most human viral diseases occur as a consequence of tissue destruction either directly by the virus itself or indirectly by the host's response to the virus. Although these mechanisms are operative in retroviral infections, retroviruses have additional mechanisms of inducing disease, including the malignant transformation of an infected cell and the induction of an immunodeficiency state through selective destruction or dysfunction of immune-competent cells that renders the host susceptible to opportunistic diseases (infections and neoplasms; **Chap. 197**).

STRUCTURE AND LIFE CYCLE

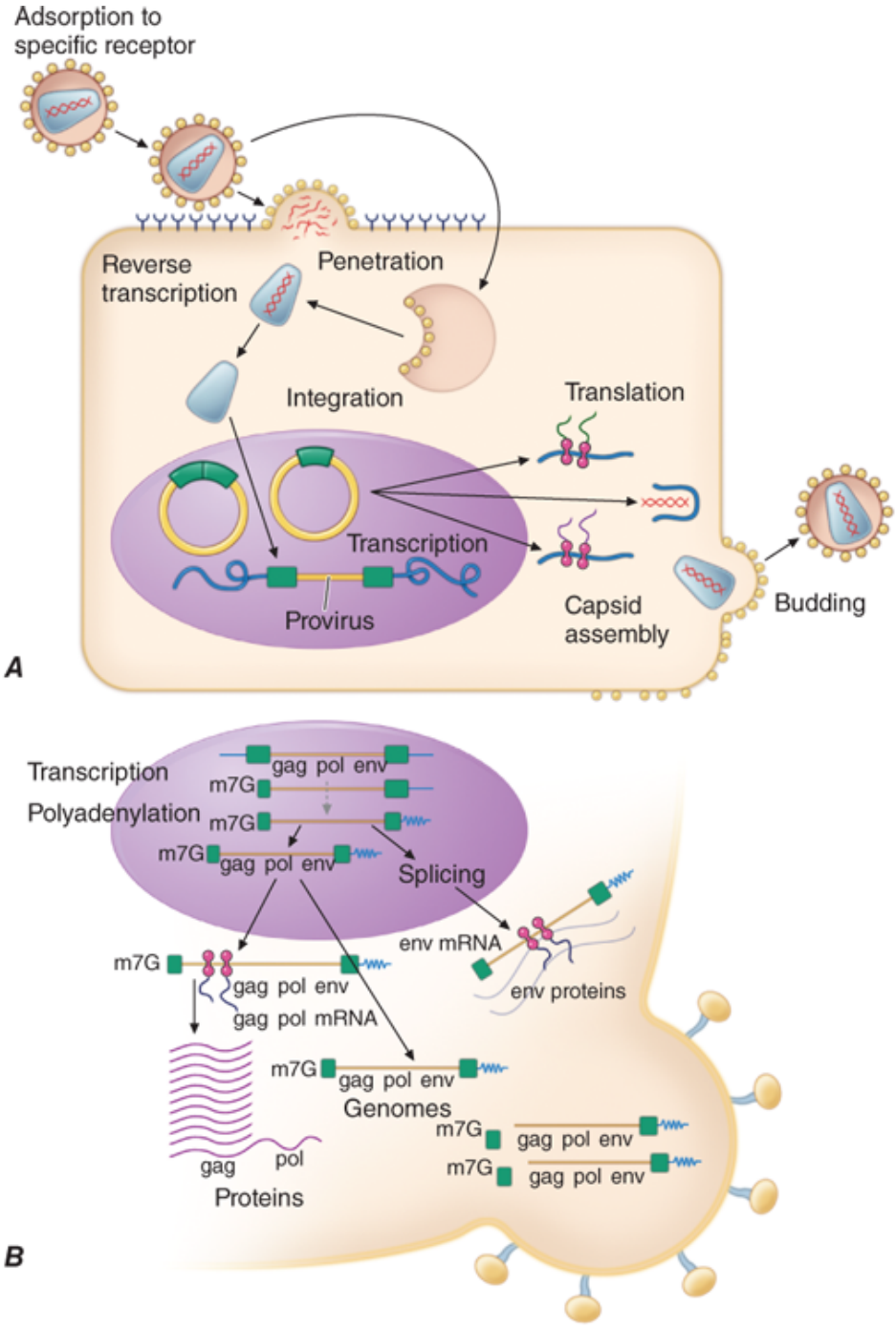
All retroviruses are similar in structure, genome organization, and mode of replication. Retroviruses are 70–130 nm in diameter and have a lipid-containing envelope surrounding an icosahedral capsid with a dense inner core. The core contains two identical copies of the single-strand RNA genome. The RNA molecules are 8–10 kb long and are complexed with reverse transcriptase and tRNA. Other viral proteins, such as integrase, are also components of the virion particle. The RNA has features usually found in mRNA: a cap site at the 5′ end of the molecule, which is important in the initiation of mRNA translation, and a polyadenylation site at the 3′ end, which influences mRNA turnover (i.e., messages with shorter polyA tails turn over faster than messages with longer polyA tails). However, the retroviral RNA is not translated; instead it is transcribed into DNA. The DNA form of the retroviral genome is called a *provirus*.

The replication cycle of retroviruses proceeds in two phases **(Fig. 196-1)**. In the first phase, the virus enters the cytoplasm after binding to one or more specific cell-surface receptors; the viral RNA and reverse transcriptase synthesize a double-strand DNA version of the RNA template; and the provirus moves into the nucleus and integrates into the host cell genome. This proviral integration is permanent. Although some animal retroviruses integrate into a single specific site of the host genome in every infected cell, the human retroviruses integrate randomly. This first phase of replication depends entirely on gene products in the virus. The second phase includes the synthesis and processing of viral genomes, mRNAs, and proteins using host cell machinery, often under the influence of viral gene products. Virions are assembled and released from the cell by budding from the membrane; host cell membrane proteins are frequently incorporated into the envelope of the virus. Proviral integration occurs during the S-phase of the cell cycle; thus, in general, nondividing cells are resistant to retroviral infection. Only the lentiviruses are able to infect nondividing cells. Once a host cell is infected, it is infected for the life of the cell.

FIGURE 196-1

The life cycle of retroviruses. ***A.*** Overview of virus replication. The retrovirus enters a target cell by binding to a specific cell-surface receptor; once the virus is internalized, its RNA is released from the nucleocapsid and is reverse-transcribed into proviral DNA. The provirus is inserted into the genome and then transcribed into RNA; the RNA is translated; and virions assemble and are extruded from the cell membrane by budding. ***B.*** Overview of retroviral gene expression. The provirus is transcribed, capped, and polyadenylated. Viral RNA molecules then have one of three

fates: they are exported to the cytoplasm, where they are packaged as the viral RNA in infectious viral particles; they are spliced to form the message for the envelope polyprotein; or they are translated into Gag and Pol proteins. Most of the messages for the Pol protein fail to initiate Pol translation because of a stop codon before its initiation; however, in a fraction of the messages, the stop codon is missed and the Pol proteins are translated. *(Modified from JM Coffin, in BN Fields, DM Knipe [eds]: Fields Virology. New York, Raven, 1990; with permission.)*



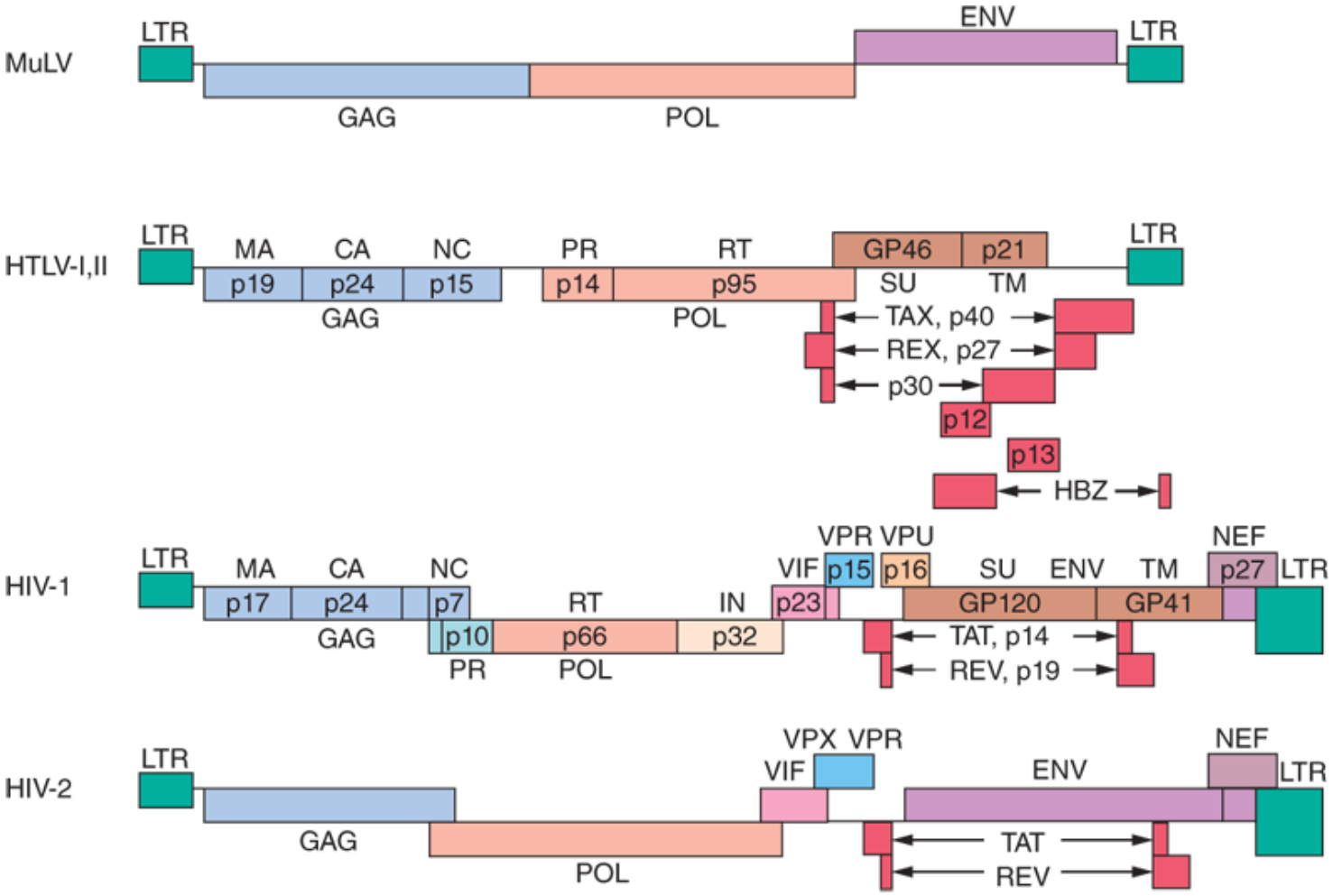
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Retroviral genomes include both coding and noncoding sequences (Fig. 196-2). In general, noncoding sequences are important recognition signals for DNA or RNA synthesis or processing events and are located in the 5′ and 3′ terminal regions of the genome. All retroviral genomes are terminally redundant, containing identical sequences

called *long terminal repeats* (LTRs). The ends of the retroviral RNA genome differ slightly in sequence from the integrated retroviral DNA. In the latter, the LTR sequences are repeated in both the 5′ and the 3′ terminus of the virus. The LTRs contain sequences involved in initiating the expression of the viral proteins, the integration of the provirus, and the polyadenylation of viral RNAs. The primer binding site, which is critical for the initiation of reverse transcription, and the viral packaging sequences are located outside the LTR sequences. The coding regions include the *gag* (group-specific antigen, core protein), *pol* (RNA-dependent DNA polymerase), and *env* (envelope) genes. The *gag* gene encodes a precursor polyprotein that is cleaved to form three to five capsid proteins; a fraction of the Gag precursor proteins also contain a protease responsible for cleaving the Gag and Pol polyproteins. A Gag-Pol polyprotein gives rise to the protease that is responsible for cleaving the Gag-Pol polyprotein. The *pol* gene encodes three proteins: the reverse transcriptase, the integrase, and the protease. The reverse transcriptase copies the viral RNA into the double-strand DNA provirus, which inserts itself into the host cell DNA via the action of integrase. The protease cleaves the Gag-Pol polyprotein into smaller protein products. The *env* gene encodes the envelope glycoproteins: one protein that binds to specific surface receptors and determines what cell types can be infected and a smaller transmembrane protein that anchors the complex to the envelope. **Fig. 196-3** shows how the retroviral gene products make up the virus structure.

FIGURE 196-2

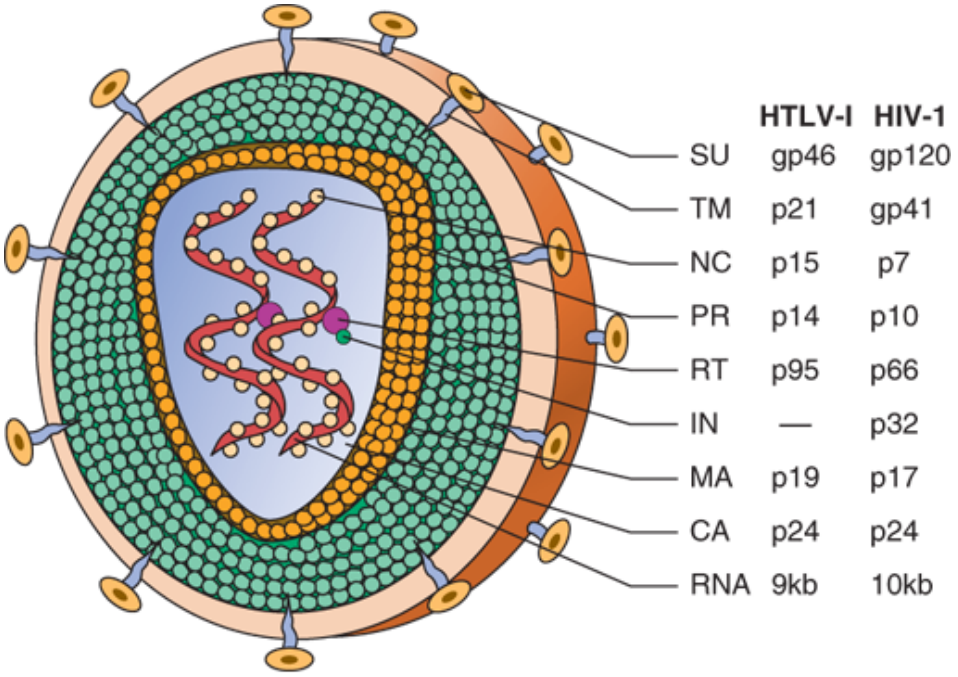
Genomic structure of retroviruses. The murine leukemia virus MuLV has the typical three structural genes: *gag, pol*, and *env*. The *gag* region gives rise to three proteins: matrix (MA), capsid (CA), and nucleic acid-binding (NC) proteins. The *pol* region encodes both a protease (PR) responsible for cleaving the viral polyproteins and a reverse transcriptase (RT). In addition, HIV *pol* encodes an integrase (IN). The *env* region encodes a surface protein (SU) and a small transmembrane protein (TM). The human retroviruses have additional gene products translated in each of the three possible reading frames. HTLV-1 and HTLV-2 have *tax* and *rex* genes with exons on either side of the *env* gene. HIV-1 and HIV-2 have six accessory gene products: *tat, rev, vif, nef, vpr*, and either *vpu* (in HIV-1) or *vpx* (in HIV-2). The genes for these proteins are located mainly between the *pol* and *env* genes. GP, glycoprotein; HBZ, HTLV-1 basic leucine zipper domain–containing protein; LTR, long terminal repeat.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 196-3

Schematic structure of human retroviruses. The surface glycoprotein (SU) is responsible for binding to receptors of host cells. The transmembrane protein (TM) anchors SU to the virus. NC is a nucleic acid–binding protein found in association with the viral RNA. A protease (PR) cleaves the polyproteins encoded by the *gag, pol*, and *env* genes into their functional components. RT is reverse transcriptase, and IN is an integrase present in some retroviruses (e.g., HIV-1) that facilitates insertion of the provirus into the host genome. The matrix protein (MA) is a Gag protein closely associated with the lipid of the envelope. The capsid protein (CA) forms the major internal structure of the virus, the core shell.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HTLVs have a region between *env* and the 3′ LTR that encodes several proteins and transcripts in overlapping reading frames (Fig. 196-2). Tax is a 40-kDa protein that does not bind to DNA but induces the expression of host cell transcription factors that alter host cell gene expression and is capable of inducing cell transformation under certain circumstances. Rex is a 27-kDa protein that regulates the expression of viral mRNAs. Other transcripts from this region (p12, p13, and p30) tend to restrict expression of viral genes and diminish the immunogenicity of infected cells. The protein of *HBZ*, a product of the complementary proviral DNA strand, interacts with many cellular transcription factors and signaling proteins. It stimulates proliferation of infected cells and is the only viral product universally expressed in HTLV-1-infected tumor cells. These proteins are produced from messages that are similar but that are spliced differently from overlapping but distinct exons.

The lentiviruses in general, and HIV-1 and -2 in particular, contain a larger genome than other pathogenic retroviruses. They contain an untranslated region between *pol* and *env* that encodes portions of several proteins, varying with the reading frame into which the mRNA is spliced. Tat is a 14-kDa protein that augments the expression of virus from the LTR. The Rev protein of HIV-1, similar to the Rex protein of HTLV, regulates RNA splicing and/or RNA transport. The Nef protein downregulates CD4, the cellular receptor for HIV; alters host T cell–activation pathways; and enhances viral infectivity. The Vif protein is necessary for the proper assembly of the HIV nucleoprotein core in many types of cells; without Vif, proviral DNA is not efficiently produced in these infected cells. In addition, the Vif protein targets APOBEC (apolipoprotein B mRNA-editing enzyme catalytic polypeptide, a cytidine deaminase that mutates the viral sequence) for proteasomal degradation, thus blocking its virus-suppressing effect. Vpr, Vpu (HIV-1 only), and Vpx (HIV-2 only) are viral proteins encoded by translation of the same message in different reading frames. As noted above, oncogenic retroviruses depend on cell proliferation for their replication; lentiviruses can infect nondividing cells, largely through effects mediated by Vpr. Vpr facilitates transport of the provirus into the nucleus and can induce other cellular changes, such as G_2 growth arrest and differentiation of some target cells. Vpx is structurally related to Vpr, but its functions are not fully defined. Vpu promotes the degradation of CD4 in the endoplasmic reticulum and stimulates the release of virions from infected cells.

Retroviruses can be either exogenously acquired (by infection with an infected cell or a free virion capable of replication) or transmitted in the germline as endogenous virus. Endogenous retroviruses are often replication defective. The human genome contains endogenous retroviral sequences, but there are no known replication-competent endogenous retroviruses in humans.

In general, viruses that contain only the *gag, pol*, and *env* genes either are not pathogenic or take a long time to induce disease; these observations indicate the importance of the other regulatory genes in viral disease pathogenesis. The pathogenesis of neoplastic transformation by retroviruses relies on the chance integration of the provirus at a spot in the genome resulting in the expression of a cellular gene (protooncogene) that becomes transforming by virtue of its unregulated expression. For example, avian leukosis virus causes B-cell leukemia by inducing the expression of *myc*. Some retroviruses possess captured and altered cellular genes near their integration site, and these viral oncogenes can transform the infected host cell. Viruses that have oncogenes often have lost a portion of their genome that is required for replication. Such viruses need helper viruses to reproduce, a feature that may explain why these acute transforming retroviruses are rare in nature. All human retroviruses identified to date are exogenous and are not acutely transforming (i.e., they lack a transforming oncogene).

These remarkable properties of retroviruses have led to experimental efforts to use them as vectors to insert specific genes into particular cell types, a process known as *gene therapy* or *gene transfer*. The process could be used to repair a genetic defect or to introduce a new property that could be used therapeutically; for example, a gene (e.g., thymidine kinase) that would make a tumor cell susceptible to killing by a drug (e.g., ganciclovir) could be inserted. One source of concern about the use of retroviral vectors in humans is that replication-competent viruses might rescue endogenous retroviral replication, with unpredictable results. This concern is not merely hypothetical: the detection of proteins encoded by endogenous retroviral sequences on the surface of cancer cells implies that the genetic events leading to the cancer were able to activate the synthesis of these usually silent genes.

HUMAN T-CELL LYMPHOTROPIC VIRUS

HTLV-1 was isolated in 1980 from a T-cell lymphoma cell line from a patient originally thought to have cutaneous T-cell lymphoma. Later it became clear that the patient had a distinct form of lymphoma (originally reported in Japan) called *adult T-cell leukemia/lymphoma* (ATL). Serologic data have determined that HTLV-1 is the cause of at least two important diseases: ATL and tropical spastic paraparesis, also called *HTLV-1-associated myelopathy* (HAM). HTLV-1 may also play a role in infective dermatitis, arthritis, uveitis, and Sjögren's syndrome.

Two years after the isolation of HTLV-1, HTLV-2 was isolated from a patient with an unusual form of hairy-cell leukemia that affected T cells. Epidemiologic studies of HTLV-2 failed to reveal a consistent disease association. Similarly, HTLV-3 and HTLV-4 have been identified but have no known disease association.

BIOLOGY AND MOLECULAR BIOLOGY

Because the biology of HTLV-1 and that of HTLV-2 are similar, the following discussion will focus on HTLV-1.

Human glucose transporter protein 1 (GLUT-1) functions as a receptor for HTLV-1, probably acting together with neuropilin-1 (NRP1) and heparan sulfate proteoglycans. Generally, only T cells are productively infected, but infection of B cells and other cell types is occasionally detected. The most common outcome of HTLV-1 infection is latent carriage of randomly integrated provirus in CD4+ T cells. HTLV-1 does not contain an oncogene and does not insert

into a unique site in the genome. Indeed, most infected cells express no viral gene products. The only viral gene product that is routinely expressed in tumor cells transformed by HTLV-1 in vivo is *hbz.* The *tax* gene is thought to be critical to the transformation process but is not expressed in the tumor cells of many ATL patients, possibly because of the immunogenicity of *tax*-expressing cells. Cells transformed in vitro, by contrast, actively transcribe HTLV-1 RNA and produce infectious virions. Most HTLV-1-transformed cell lines are the result of the infection of a normal host T cell in vitro. It is difficult to establish cell lines derived from authentic ATL cells.

Although *tax* does not itself bind to DNA, it does induce the expression of a wide range of host cell gene products, including transcription factors (especially c-rel/nuclear factor κB [NF-κB], ets-1 and -2, and members of the fos/jun family), cytokines (e.g., interleukin [IL] 2, granulocyte-macrophage colony-stimulating factor, and tumor necrosis factor), membrane proteins and receptors (major histocompatibility [MHC] molecules and IL-2 receptor α), and chromatin remodeling complexes. The genes activated by *tax* are generally controlled by transcription factors of the c-rel/NF-κB and cyclic AMP response element binding (CREB) protein families. It is unclear how this induction of host gene expression leads to neoplastic transformation; *tax* can interfere with G_1 and mitotic cell-cycle checkpoints, block apoptosis, inhibit DNA repair, and promote antigen-independent T-cell proliferation. Induction of a cytokine-autocrine loop has been proposed; however, IL-2 is not the crucial cytokine. The involvement of IL-4, IL-7, and IL-15 has been proposed.

In light of the irregular expression of *tax* in ATL cells, it has been suggested that *tax* is important in the early phases of transformation but is not essential for the maintenance of the transformed state. The maintenance role is thought to be due to *hbz* expression. As is clear from the epidemiology of HTLV-1 infection, transformation of an infected cell is a rare event and may depend on heterogeneous second, third, or fourth genetic hits. No consistent chromosomal abnormalities have been described in ATL; however, aneuploidy is common, and individual cases with p53 mutations and translocations involving the T-cell receptor genes on chromosome 14 have been reported. *Tax* may repress certain DNA repair enzymes, permitting the accumulation of genetic damage that would normally be repaired. However, the molecular pathogenesis of HTLV-1-induced neoplasia is not fully understood.

FEATURES OF HTLV-1 INFECTION

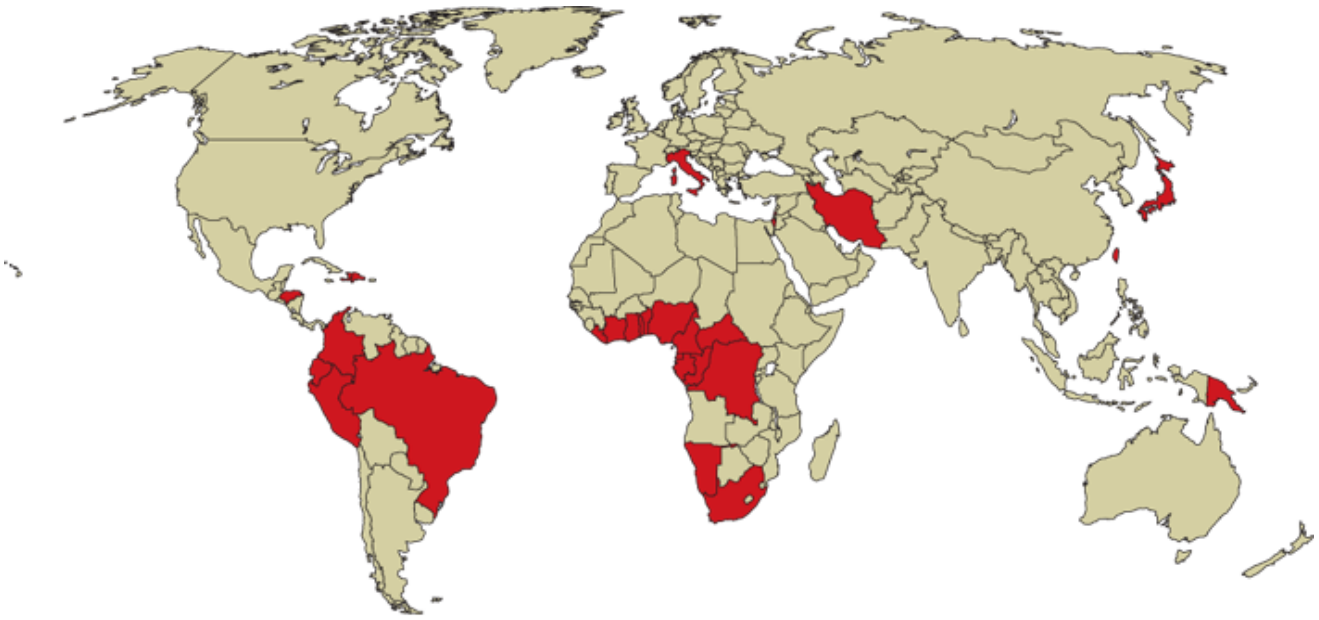
Epidemiology

HTLV-1 infection is transmitted in at least three ways: from mother to child, especially via breast milk; through sexual activity, more commonly from men to women; and through the blood—via contaminated transfusions or contaminated needles. The virus is most commonly transmitted perinatally. Compared with HIV, which can be transmitted in cell-free form, HTLV-1 is less infectious, and its transmission usually requires cell-to-cell contact.

HTLV-1 is endemic in southwestern Japan and Okinawa, where >1 million persons are infected. Antibodies to HTLV-1 are present in the serum of up to 35% of Okinawans, 10% of residents of the Japanese island of Kyushu, and <1% of persons in nonendemic regions of Japan. Despite this high prevalence of infection, only ~500 cases of ATL are diagnosed in this area each year. Clusters of infection have been noted in other areas of eastern Asia, such as Taiwan; in the Caribbean basin, including northeastern South America; in northwestern South America; in central and southern Africa; in Italy, Israel, Iran, and Papua New Guinea; in the Arctic; and in the southeastern part of the United States (Fig. 196-4). An estimated 5–10 million persons have HTLV-1 infection worldwide.

FIGURE 196-4

Global distribution of HTLV-1 infection. Countries with a prevalence of HTLV-1 infection of 1–5% are shaded darkly. Note that the distribution of infected patients is not uniform in endemic countries. For example, the people of southwestern Japan and northeastern Brazil are more commonly affected than those in other regions of those countries.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Progressive spastic or ataxic myelopathy developing in an individual who is HTLV-1 positive (i.e., who has serum antibodies to HTLV-1) may be due to direct infection of the nervous system with the virus, but destruction of the pyramidal tracts appears to involve HTLV-1-infected CD4+ T cells; a similar disorder may result from infection with HIV or HTLV-2. In rare instances, patients with HAM are seronegative but have detectable antibody to HTLV-1 in cerebrospinal fluid (CSF).

The cumulative lifetime risk of developing ATL is 3% among HTLV-1-infected patients, with a threefold greater risk among men than among women; a similar cumulative risk is projected for HAM (4%), but with women more commonly affected than men. The distribution of these two diseases overlaps the distribution of HTLV-1, with >95% of affected patients showing serologic evidence of HTLV-1 infection. The latency period between infection and the emergence of disease is 20–30 years for ATL. For HAM, the median latency period is ~3.3 years (range, 4 months to 30 years). The development of ATL is rare among persons infected by blood products; however, ~20% of patients with HAM acquire HTLV-1 from contaminated blood. ATL is more common among perinatally infected individuals, whereas HAM is more common among persons infected via sexual transmission.

Associated Diseases

ATL

Four clinical types of HTLV-1-induced neoplasia have been described: acute, lymphomatous, chronic, and smoldering. All of these tumors are monoclonal proliferations of CD4+ postthymic T cells with clonal proviral integrations and clonal T-cell receptor gene rearrangements.

ACUTE ATL

About 60% of patients who develop malignancy have classic acute ATL, which is characterized by a short clinical prodrome (~2 weeks between the first symptoms and the diagnosis) and an aggressive natural history (median

survival period, 6 months). The clinical picture is dominated by rapidly progressive skin lesions, pulmonary involvement, hypercalcemia, and lymphocytosis with cells containing lobulated or "flower-shaped" nuclei **(see Fig. 104-10)**. The malignant cells have monoclonal proviral integrations and express CD4, CD3, and CD25 (low-affinity IL-2 receptors) on their surface. Serum levels of CD25 can be used as a tumor marker. Anemia and thrombocytopenia are rare. The skin lesions may be difficult to distinguish from those in mycosis fungoides. Lytic bone lesions, which are common, do not contain tumor cells but rather are composed of osteolytic cells, usually without osteoblastic activity. Despite the leukemic picture, bone marrow involvement is patchy in most cases.

The hypercalcemia of ATL is multifactorial; the tumor cells produce osteoclast-activating factors (tumor necrosis factor α, IL-1, lymphotoxin) and can also produce a parathyroid hormone–like molecule. Affected patients have an underlying immunodeficiency that makes them susceptible to opportunistic infections similar to those seen in patients with AIDS **(Chap. 197)**. The pathogenesis of the immunodeficiency is unclear. Pulmonary infiltrates in ATL patients reflect leukemic infiltration half the time and opportunistic infections with organisms such as *Pneumocystis* and other fungi the other half. Gastrointestinal symptoms are nearly always related to opportunistic infection. *Strongyloides stercoralis* is a gastrointestinal parasite that has a pattern of endemic distribution similar to that of HTLV-1. HTLV-1-infected persons also infected with this parasite may develop ATL more often or more rapidly than those without *Strongyloides* infections. Serum concentrations of lactate dehydrogenase and alkaline phosphatase are often elevated in ATL. About 10% of patients have leptomeningeal involvement leading to weakness, altered mental status, paresthesia, and/or headache. Unlike other forms of central nervous system (CNS) lymphoma, ATL may be accompanied by normal CSF protein levels. The diagnosis depends on finding ATL cells in the CSF **(Chap. 104)**.

LYMPHOMATOUS ATL

The lymphomatous type of ATL occurs in ~20% of patients and is similar to the acute form in its natural history and clinical course, except that circulating abnormal cells are rare and lymphadenopathy is evident. The histology of the lymphoma is variable but does not influence the natural history. In general, the diagnosis is suspected on the basis of the patient's birthplace (see "Epidemiology," above) and the presence of skin lesions and hypercalcemia. The diagnosis is confirmed by the detection of antibodies to HTLV-1 in serum.

CHRONIC ATL

Patients with the chronic form of ATL generally have normal serum levels of calcium and lactate dehydrogenase and no involvement of the CNS, bone, or gastrointestinal tract. The median duration of survival for these patients is 2 years. In some cases, chronic ATL progresses to the acute form of the disease.

SMOLDERING ATL

Fewer than 5% of patients have the smoldering form of ATL. In this form, the malignant cells have monoclonal proviral integration; <5% of peripheral-blood cells exhibit typical morphologic abnormalities; hypercalcemia, adenopathy, and hepatosplenomegaly do not develop; the CNS, the bones, and the gastrointestinal tract are not involved; and skin lesions and pulmonary lesions may be present. The median survival period for this small subset of patients appears to be ≥5 years.

HAM (TROPICAL SPASTIC PARAPARESIS)

In contrast to ATL, in which there is a slight predominance of male patients, HAM affects female patients disproportionately. HAM resembles multiple sclerosis in certain ways **(Chap. 436)**. The onset is insidious. Symptoms include weakness or stiffness in one or both legs, back pain, and urinary incontinence. Sensory changes are usually mild, but peripheral neuropathy may develop. The disease generally takes the form of slowly progressive and unremitting thoracic myelopathy; one-third of patients are bedridden within 10 years of diagnosis, and one-half are

unable to walk unassisted by this point. Patients display spastic paraparesis or paraplegia with hyperreflexia, ankle clonus, and extensor plantar responses. Cognitive function is usually spared; cranial nerve abnormalities are unusual.

MRI reveals lesions in both the white matter and the paraventricular regions of the brain as well as in the spinal cord. Pathologic examination of the spinal cord shows symmetric degeneration of the lateral columns, including the corticospinal tracts; some cases involve the posterior columns as well. The spinal meninges and cord parenchyma contain an inflammatory infiltrate that includes CD8+ T cells with myelin destruction.

HTLV-1 is not usually found in cells of the CNS but may be detected in a small population of lymphocytes present in the CSF. In general, HTLV-1 replication is greater in HAM than in ATL, and patients with HAM have a stronger immune response to the virus. Antibodies to HTLV-1 are present in the serum and appear to be produced in the CSF of HAM patients, where titers are often higher than in the serum. The pathophysiology of HAM may involve the induction of autoimmune destruction of neural cells by T cells with specificity for viral components such as Tax or Env proteins. One theory is that susceptibility to HAM may be related to the presence of human leukocyte antigen (HLA) alleles capable of presenting viral antigens in a fashion that leads to autoimmunity. Insufficient data are available to confirm an HLA association. However, antibodies in the sera of HAM patients have been shown to bind a neuron-specific antigen (heteronuclear ribonuclear protein A1 [hnRNP A1]) and to interfere with neurotransmission in vitro.

It is unclear what factors influence whether HTLV-1 infection will cause disease and, if it does, whether it will induce a neoplasm (ATL) or an autoimmune disorder (HAM). Differences in viral strains, the susceptibility of particular MHC haplotypes, the route of HTLV-1 infection, the viral load, and the nature of the HTLV-1-related immune response are putative factors, but few definitive data are available.

OTHER PUTATIVE HTLV-1-RELATED DISEASES

Even in the absence of the full clinical picture of HAM, bladder dysfunction is common in HTLV-1-infected women. In areas where HTLV-1 is endemic, diverse inflammatory and autoimmune diseases have been attributed to the virus, including uveitis, dermatitis, pneumonitis, rheumatoid arthritis, and polymyositis. However, a causal relationship between HTLV-1 and these illnesses has not been established.

Prevention

Women in endemic areas should not breast-feed their children, and blood donors should be screened for serum antibodies to HTLV-1. As in the prevention of HIV infection, the practice of safe sex and the avoidance of needle sharing are important.

TREATMENT

TREATMENT

HTLV-1 Infection

For the small number of patients who develop HTLV-1-related disease, therapies are not curative. In patients with the acute and lymphomatous types of ATL, the disease progresses rapidly. Hypercalcemia is generally controlled by glucocorticoid administration and cytotoxic therapy directed against the neoplasm. The tumor is highly responsive to combination chemotherapy that is used against other forms of lymphoma; however, patients are susceptible to overwhelming bacterial and opportunistic infections, and ATL relapses within 4–10 months after remission in most cases. The combination of interferon α and zidovudine may extend survival. Because viral replication is not clearly

associated with ATL progression, zidovudine is probably effective through its cytotoxic effects (as a chain-terminating thymidine analogue) rather than its antiviral effects. Selected series have reported high rates of response and a 40% rate of 5-year survival; however, this level of response has not been universal. LSG15, a multidrug chemotherapy program developed in Japan, induces complete responses in about one-third of patients, about half of whom survive for >2 years; however, the median survival time is about 13 months. High-dose therapy with bone marrow transplantation has been widely tested in Japan. Median survival has not been influenced by this treatment; however, up to 25% of patients survive free of disease for 4 years. Lenalidomide has been reported to have a 42% response rate in patients with relapsed ATL, extending median survival to 20 months despite a short 4-month progression-free survival period. A pilot trial suggested that treatment with mogamulizumab, an antibody to CCR4 (a receptor for a number of chemokines, including RANTES and TARC), improved response rates when added to chemotherapy. An experimental approach using an yttrium 90-labeled or toxin-conjugated antibody to the IL-2 receptor appears promising but is not widely available. Patients with the chronic or smoldering form of ATL may be managed with an expectant approach: treat any infections, and watch and wait for signs of progression to acute disease.

Patients with HAM may obtain some benefit from the use of glucocorticoids to reduce inflammation. Antiretroviral regimens have not been effective. In one study, danazol (200 mg three times daily) produced significant neurologic improvement in five of six treated patients, with resolution of urinary incontinence in two cases, decreased spasticity in three, and restoration of the ability to walk after confinement to a wheelchair in two. Antibody to IL-15 receptor β chain has been tested with some promising clinical effects in small numbers of patients. Physical therapy and rehabilitation are important components of management.

FEATURES OF HTLV-2 INFECTION

Epidemiology

 HTLV-2 is endemic in certain Native American tribes and in Africa. It is generally considered to be a New World virus that was brought from Asia to the Americas 10,000–40,000 years ago during the migration of infected populations across the Bering land bridge. The mode of transmission of HTLV-2 is probably the same as that of HTLV-1 (see above). HTLV-2 may be less readily transmitted sexually than HTLV-1.

Studies of large cohorts of injection drug users with serologic assays that reliably distinguish HTLV-1 from HTLV-2 indicated that the vast majority of HTLV-positive cohort members were infected with HTLV-2. The seroprevalence of HTLV in a cohort of 7841 injection drug users from drug treatment centers in Baltimore, Chicago, Los Angeles, New Jersey (Asbury Park and Trenton), New York City (Brooklyn and Harlem), Philadelphia, and San Antonio was 20.9%, with >97% of cases due to HTLV-2. The seroprevalence of HTLV-2 was higher in the Southwest and the Midwest than in the Northeast. In contrast, the seroprevalence of HIV-1 was higher in the Northeast than in the Southwest or the Midwest. Approximately 3% of the cohort members were infected with both HTLV-2 and HIV-1. The seroprevalence of HTLV-2 increased linearly with age. Women were significantly more likely than men to be infected with HTLV-2; the virus is thought to be more efficiently transmitted from male to female than from female to male.

Associated Diseases

Although HTLV-2 was isolated from a patient with a T-cell variant of hairy-cell leukemia, this virus has not been consistently associated with a particular disease and in fact has been thought of as “a virus searching for a disease.” However, evidence is accumulating that HTLV-2 may play a role in certain neurologic, hematologic, and dermatologic

diseases. These data require confirmation, particularly in light of the previous confusion regarding the relative prevalences of HTLV-1 and HTLV-2 among injection drug users.

Prevention

Avoidance of needle sharing, adherence to safe-sex practices, screening of blood (by assays for HTLV-1, which also detect HTLV-2), and avoidance of breast-feeding by infected women are important principles in the prevention of spread of HTLV-2.

HUMAN IMMUNODEFICIENCY VIRUS

HIV-1 and HIV-2 are members of the lentivirus subfamily of Retroviridae and are the only lentiviruses known to infect humans. The lentiviruses are slower-acting than viruses that cause acute infection (e.g., influenza virus) but not than other retroviruses. The features of acute primary infection with HIV resemble those of more classic acute infections. The characteristic chronicity of HIV disease is consistent with the designation *lentivirus*. **For a detailed discussion of HIV, see Chap. 197.**

FURTHER READING

Ishida T et al: Multicenter phase II study of lenalidomide in relapsed or recurrent adult T-cell leukemia-lymphoma: ATLL-002. J Clin Oncol 34:4086, 2016.

Katsuya H et al: Treatment and survival among 1594 patients with ATL. Blood 126:2570, 2015.

Ma G et al: Multifaceted functions and roles of HBA in HTLV-1 pathogenesis. Retrovirology 13:16, 2016.

Moir S et al: Pathogenic mechanisms of HIV disease. Annu Rev Pathol 6:223, 2011.

Panfil AR et al: Human T-cell leukemia virus–associated malignancy. Curr Opin Virol 20:40, 2016.

Chapter 197: Human Immunodeficiency Virus Disease: AIDS and Related Disorders

Anthony S. Fauci; Gregory K. Folkers; H. Clifford Lane

INTRODUCTION

AIDS was first recognized in the United States in the summer of 1981, when the U.S. Centers for Disease Control and Prevention (CDC) reported the unexplained occurrence of *Pneumocystis jirovecii* (formerly *P. carinii*) pneumonia in five previously healthy homosexual men in Los Angeles and of Kaposi's sarcoma (KS) with or without *P. jirovecii* pneumonia and other opportunistic infections in 26 previously healthy homosexual men in New York, San Francisco, and Los Angeles. The disease was soon recognized in male and female injection drug users; in hemophiliacs and blood transfusion recipients; among female sexual partners of men with AIDS; and among infants born to mothers with AIDS. In 1983, human immunodeficiency virus (HIV) was isolated from a patient with lymphadenopathy, and by 1984 it was demonstrated clearly to be the causative agent of AIDS. In 1985, a sensitive enzyme-linked immunosorbent assay (ELISA) was developed; this led to an appreciation of the scope and evolution of the HIV epidemic at first in the United States and other developed nations and ultimately among developing nations throughout the world (see "HIV Infection and AIDS Worldwide," below). The staggering worldwide evolution of the HIV pandemic has been matched by an explosion of information in the areas of HIV virology, pathogenesis (both immunologic and virologic), treatment of HIV disease, treatment and prophylaxis of the opportunistic diseases associated with HIV infection, and prevention of HIV infection. The information flow related to HIV disease is enormous and continues to expand, and it has become almost impossible for the health care generalist to stay abreast of the literature. The purpose of this chapter is to present the most current information available on the scope of the pandemic; on its pathogenesis, treatment, and prevention; and on prospects for vaccine development. Above all, the aim is to provide a solid scientific basis and practical clinical guidelines for a state-of-the-art approach to the HIV-infected patient.

DEFINITION

The current CDC classification system for HIV infection and AIDS categorizes patients based on clinical conditions associated with HIV infection together with the level of the CD4+ T lymphocyte count. A confirmed HIV case can be classified in one of five HIV infection stages (0, 1, 2, 3, or unknown). If there was a negative HIV test within 6 months of the first HIV infection diagnosis, the stage is 0, and remains 0 until 6 months after diagnosis. Advanced HIV disease (AIDS) is classified as stage 3 if one or more specific opportunistic illness has been diagnosed **(Table 197-1)**. Otherwise, the stage is determined by CD4+ T lymphocyte test results and immunologic criteria **(Table 197-2)**. If none of these criteria apply (e.g., because of missing information on CD4+ T lymphocyte test results), the stage is U (unknown).

TABLE 197-1

CDC Stage 3 (AIDS)-Defining Opportunistic Illnesses in HIV Infection

Bacterial infections, multiple or recurrent[a]
Candidiasis of bronchi, trachea, or lungs
Candidiasis of esophagus
Cervical cancer, invasive[b]
Coccidioidomycosis, disseminated or extrapulmonary
Cryptococcosis, extrapulmonary
Cryptosporidiosis, chronic intestinal (>1 month's duration)
Cytomegalovirus disease (other than liver, spleen, or nodes), onset at age >1 month
Cytomegalovirus retinitis (with loss of vision)
Encephalopathy attributed to HIV
Herpes simplex: chronic ulcers (>1 month's duration) or bronchitis, pneumonitis, or esophagitis (onset at age >1 month)
Histoplasmosis, disseminated or extrapulmonary
Isosporiasis, chronic intestinal (>1 month's duration)
Kaposi's sarcoma
Lymphoma, Burkitt's (or equivalent term)
Lymphoma, immunoblastic (or equivalent term)
Lymphoma, primary, of brain
Mycobacterium avium complex or *Mycobacterium kansasii*, disseminated or extrapulmonary
Mycobacterium tuberculosis of any site, pulmonary,[b] disseminated, or extrapulmonary
Mycobacterium, other species or unidentified species, disseminated or extrapulmonary
Pneumocystis jirovecii (previously known as *Pneumocystis carinii*) pneumonia
Pneumonia, recurrent[b]
Progressive multifocal leukoencephalopathy
Salmonella septicemia, recurrent
Toxoplasmosis of brain, onset at age >1 month
Wasting syndrome attributed to HIV

[a]Only among children age <6 years. [b]Only among adults, adolescents, and children age ≥6 years.

Source: MMWR 63(RR-03), April 11, 2014.

TABLE 197-2

CDC HIV Infection Stages 1–3 Based on Age-Specific CD4+ T Lymphocyte Count or CD4+ T Lymphocyte Percentage of Total Lymphocytes[a]

	Age on Date of CD4 T+ Lymphocyte Test					
	<1 Year		1–5 Years		6 Years through Adult	
Stage[a]	Cells/µL	%	Cells/µL	%	Cells/µL	%
1	≥1500	≥34	≥1000	≥30	≥500	≥26
2	750–1499	26–33	500–999	22–29	200–499	14–25
3	<750	<26	<500	<22	<200	<14

[a]The stage is based primarily on the CD4+ T lymphocyte count; the CD4+ T lymphocyte count takes precedence over the CD4+ T lymphocyte percentage, and the percentage is considered only if the count is missing.

Source: MMWR 63(RR-03), April 11, 2014.

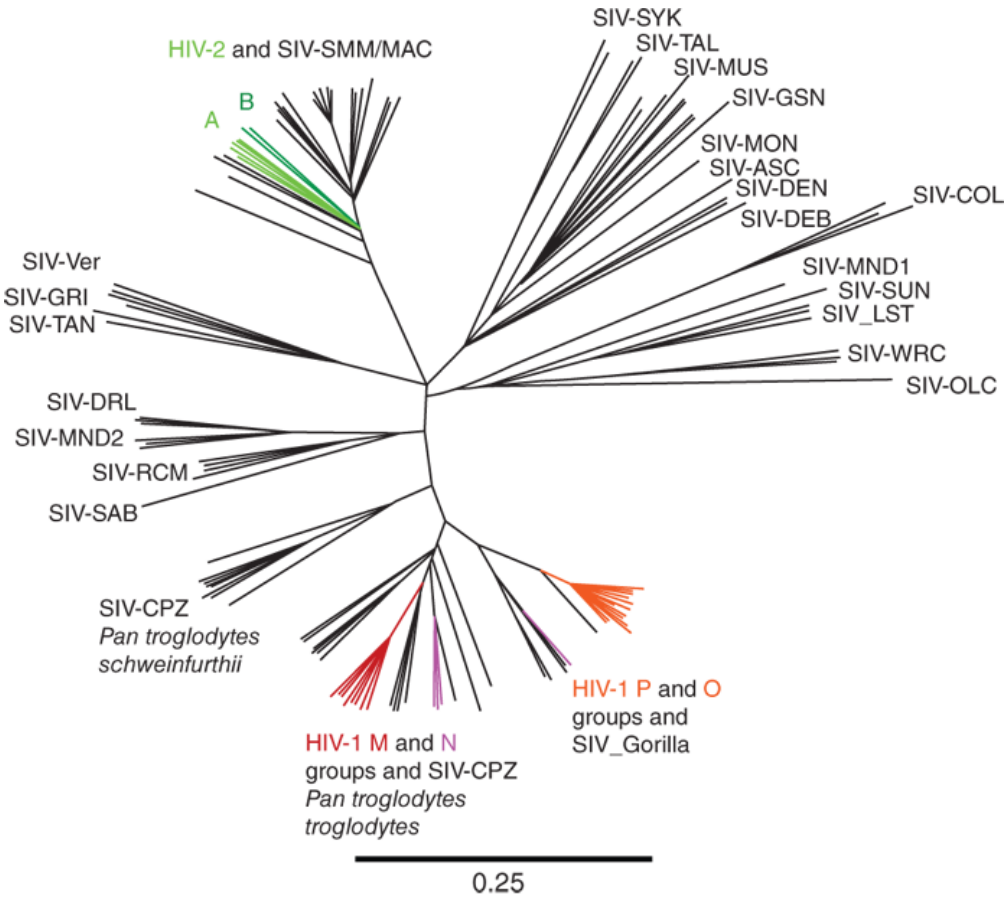
The definition and staging criteria of AIDS are complex and comprehensive and were established for surveillance purposes rather than for the practical care of patients. Thus, the clinician should not focus on whether the patient fulfills the strict definition of AIDS, but should view HIV disease as a spectrum ranging from primary infection, with or without the acute syndrome, to the relatively asymptomatic stage, to advanced stages associated with opportunistic diseases (see "Pathophysiology and Pathogenesis," below).

ETIOLOGIC AGENT

HIV is the etiologic agent of AIDS; it belongs to the family of human retroviruses (Retroviridae) and the subfamily of lentiviruses **(Chap. 196)**. Nononcogenic lentiviruses cause disease in other animal species, including sheep, horses, goats, cattle, cats, and monkeys. The four retroviruses known to cause human disease belong to two distinct groups: the human T lymphotropic viruses (HTLV)-1 and HTLV-2, which are transforming retroviruses; and the human immunodeficiency viruses, HIV-1 and HIV-2, which cause cytopathic effects either directly or indirectly **(Chap. 196)**. The most common cause of HIV disease throughout the world, and certainly in the United States, is HIV-1, which comprises several subtypes with different geographic distributions (see "Molecular Heterogeneity of HIV-1," below). HIV-2 was first identified in 1986 in West African patients and was originally confined to West Africa. However, cases traced to West Africa or to sexual contacts with West Africans have been identified throughout the world. The currently defined groups of HIV-1 (M, N, O, P) and the HIV-2 groups A through H each are likely derived from a separate transfer to humans from a nonhuman primate reservoir. HIV-1 viruses likely came from chimpanzees and/or gorillas, and HIV-2 from sooty mangabeys. The AIDS pandemic is primarily caused by the HIV-1 M group viruses. Although HIV-1 group O and HIV-2 viruses have been found in numerous countries, including those in the developed world, they have caused much more localized epidemics. The taxonomic relationship between primate lentiviruses is shown in **Fig. 197-1**.

FIGURE 197-1

A phylogenetic tree based on the nearly complete genomes (*gag* through *nef* genes) of primate immunodeficiency viruses. The scale (0.25) indicates a 25% phylogenetically corrected genetic distance at the nucleotide level. Clades in color represent viruses (HIV-1, HIV-2) identified in humans after relatively recent transfers from chimpanzee, gorilla, and sooty mangabey reservoirs. *(Prepared by Brian Foley, PhD, of the HIV Sequence Database, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory; additional information at www.hiv.lanl.gov/content/sequence/HelpDocs/subtypes.html.)*



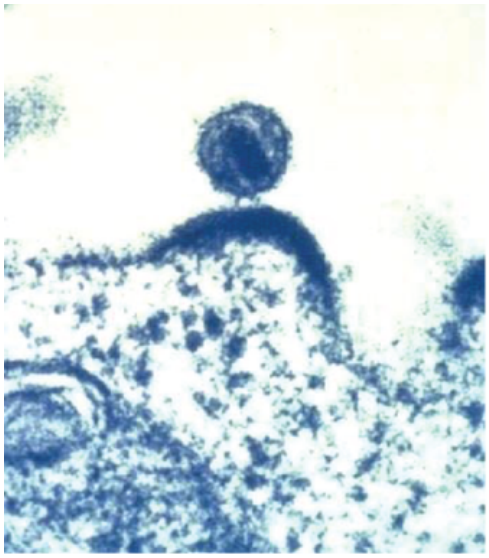
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MORPHOLOGY OF HIV

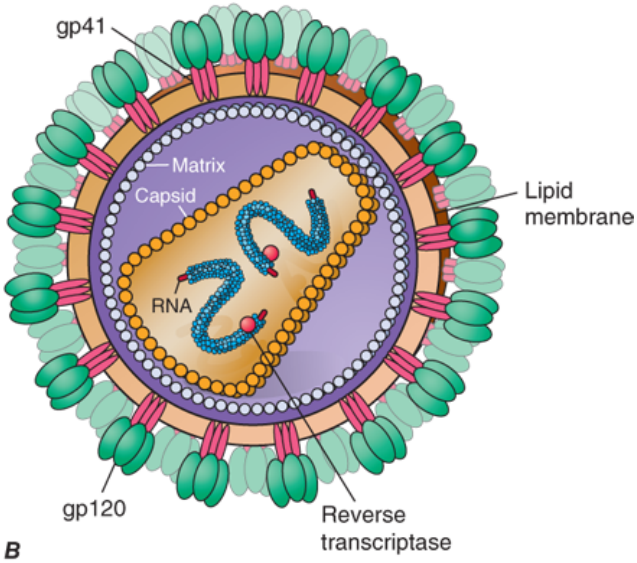
Electron microscopy shows that the HIV virion is an icosahedral structure **(Fig. 197-2)** containing numerous external spikes formed by the two major envelope proteins, the external gp120 and the transmembrane gp41. The HIV envelope exists as a trimeric heterodimer. The virion buds from the surface of the infected cell **(Fig. 197-2A)** and incorporates a variety of host cellular proteins into its lipid bilayer. The structure of HIV-1 is schematically diagrammed in **Fig. 197-2B**.

FIGURE 197-2

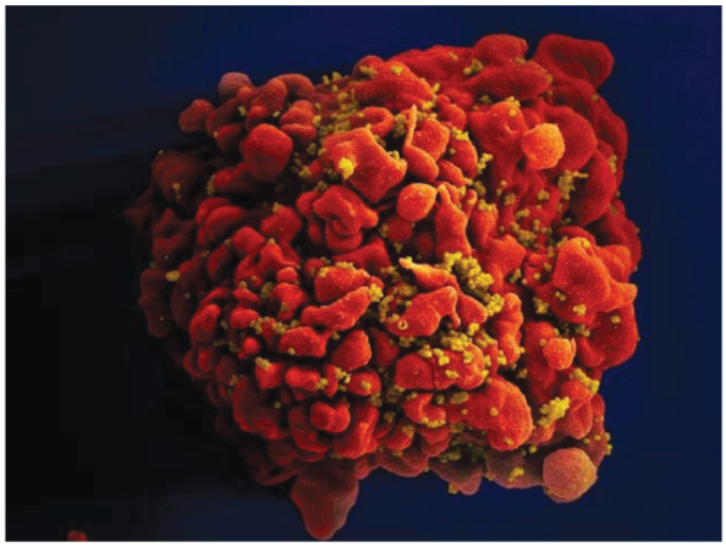
A. Electron micrograph of HIV. Figure illustrates a typical virion following budding from the surface of a CD4+ T lymphocyte, together with two additional incomplete virions in the process of budding from the cell membrane. ***B.*** Structure of HIV-1, including the gp120 envelope, gp41 transmembrane components of the envelope, genomic RNA, enzyme reverse transcriptase, p18(17) inner membrane (matrix), and p24 core protein (capsid). *(Copyright by George V. Kelvin.) (Adapted from RC Gallo: Sci Am 256:46, 1987.)* ***C.*** Scanning electron micrograph of HIV-1 virions infecting a human CD4+ T lymphocyte. The original photograph was imaged at 20,000× magnification. Cell is approximately 10 microns in diameter, and the HIV particles are approximately 120 nanometers. *(Courtesy of Elizabeth R. Fischer, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases; with permission.)*



A



B



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

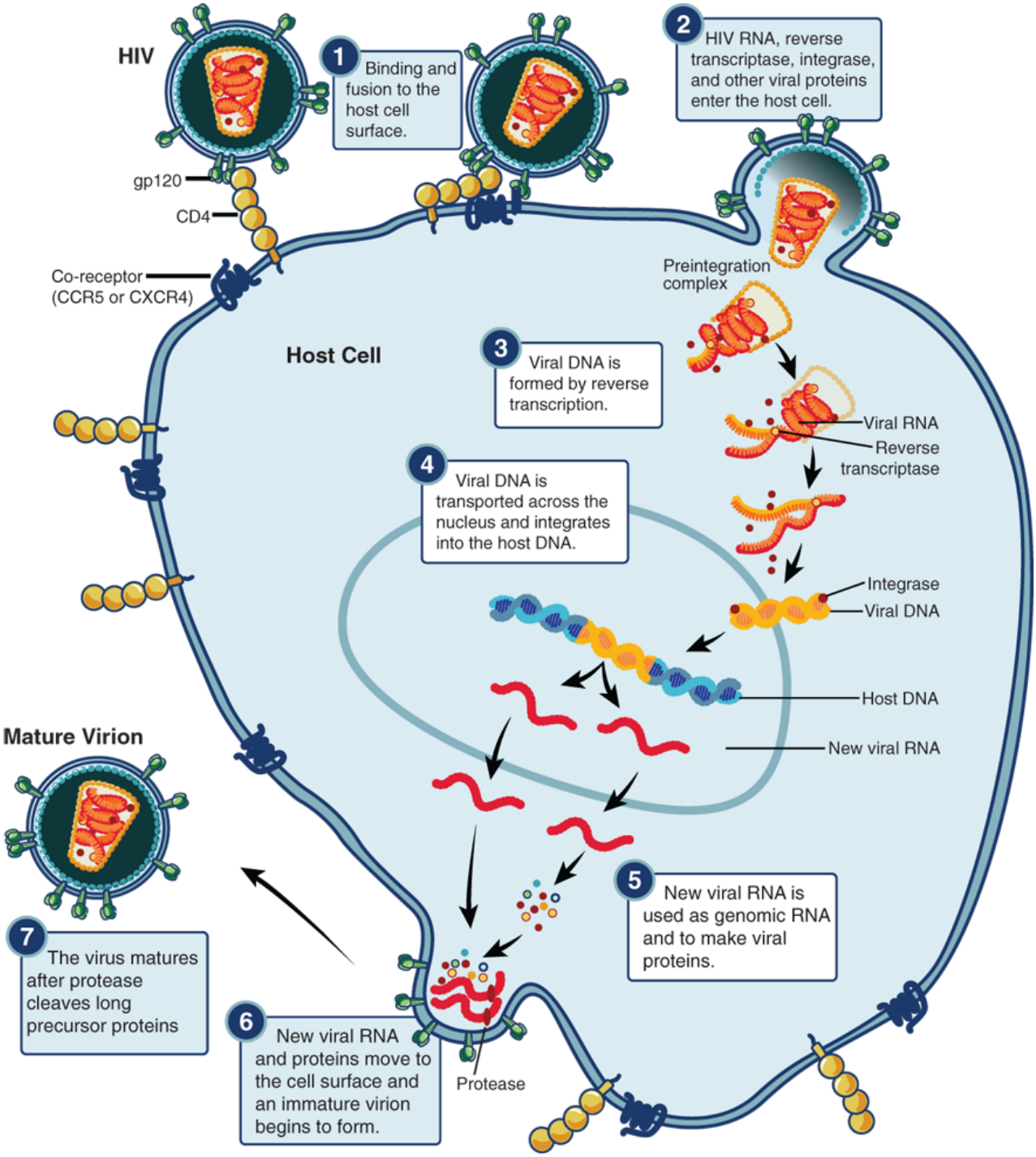
REPLICATION CYCLE OF HIV

HIV is an RNA virus whose hallmark is the reverse transcription of its genomic RNA to DNA by the enzyme *reverse transcriptase*. The replication cycle of HIV begins with the high-affinity binding via surface-exposed residues within the gp120 protein to its receptor on the host cell surface, the CD4 molecule **(Fig. 197-3)**. The CD4 molecule is a 55-kDa protein found predominantly on a subset of T lymphocytes that are responsible for helper function in the immune system **(Chap. 342)**. Once it binds to CD4, the gp120 protein undergoes a conformational change that facilitates binding to one of two major co-receptors. The two major co-receptors for HIV-1 are CCR5 and CXCR4. Both receptors belong to the family of seven-transmembrane-domain G protein–coupled cellular receptors, and the use of one or the other or both receptors by the virus for entry into the cell is an important determinant of the cellular tropism of the virus. Cell-to-cell spread is also facilitated by accessory molecules such as the C-type lectin receptor *DC-SIGN* expressed on certain dendritic cells (DCs) that bind to the HIV gp120 envelope protein, allowing virus captured on DCs to spread to CD4+ T cells. Following binding of the envelope protein to the CD4 molecule associated with the above-mentioned conformational change in the viral envelope gp120, fusion with the host cell membrane occurs via the newly exposed gp41 molecule penetrating the plasma membrane of the target cell and then coiling upon itself to bring the virion and target cell together **(Fig. 197-4)**. Following fusion, uncoating of the capsid protein shell is initiated—a step that facilitates reverse transcription and leads to formation of the preintegration complex, composed of viral RNA, enzymes, and accessory proteins and surrounded by capsid and matrix proteins (Fig. 197-3). As the preintegration complex traverses the cytoplasm to reach the nucleus, the viral reverse transcriptase enzyme catalyzes the reverse transcription of the genomic RNA into DNA, resulting in the formation of double-stranded proviral HIV DNA. At several steps of the replication cycle, the virus is vulnerable to various cellular factors that can block the progression of infection. The cytoplasmic tripartite motif-containing protein 5-α (TRIM5-α) is a host restriction factor that interacts with retroviral capsids, causing their premature disassembly and induction of innate immune responses. While early studies with laboratory strains found the HIV-1 capsid bound weakly to the human form of TRIM5-α, capsids of primary isolates

appear to be more susceptible to TRIM5-α-mediated disassembly. The apolipoprotein B mRNA editing enzyme (catalytic polypeptide-like 3 [APOBEC3]) family of cellular proteins also inhibits progression of virus infection after virus has entered the cell and prior to entering the nucleus. APOBEC3 proteins, which are incorporated into virions and released into the cytoplasm of a newly infected cell, bind to the single minus-strand DNA intermediate and deaminate viral cytidine, causing hypermutation of retroviral genomes. HIV has evolved a powerful strategy to protect itself from APOBEC. The viral protein Vif targets APOBEC3 for proteasomal degradation. SAMHD1 is another post-entry host factor that prevents reverse transcription by depleting pools of deoxynucleotides (dNTPs). The type I interferon (IFN)-induced myxovirus resistance protein 2 (MX2) is another restriction factor associated with innate immunity that inhibits HIV-1 nuclear entry.

FIGURE 197-3

The replication cycle of HIV. See text for description. *(From the National Institute of Allergy and Infectious Diseases.)*

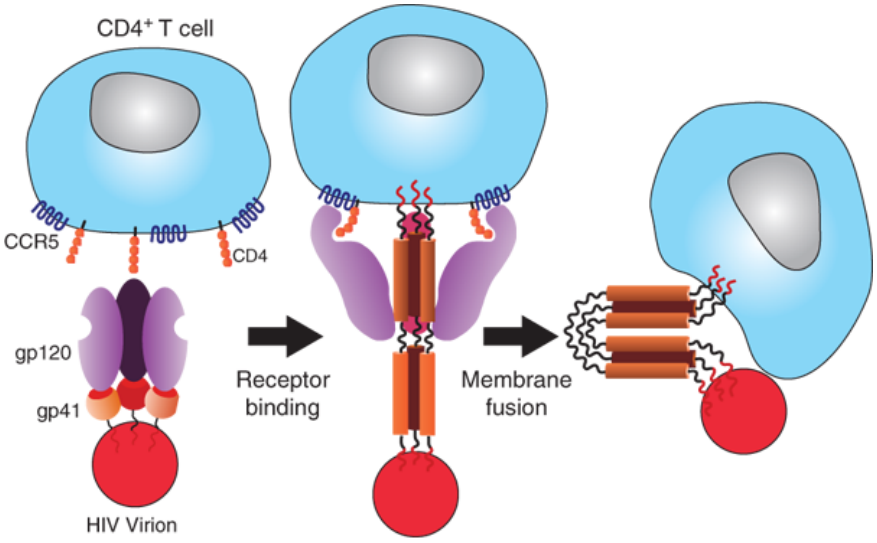


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 197-4

Binding and fusion of HIV-1 with its target cell. HIV-1 binds to its target cell via the CD4 molecule, leading to a conformational change in the gp120 molecule that allows it to bind to the co-receptor CCR5 (for R5-using viruses). The virus then firmly attaches to the host cell membrane in a coiled-spring

fashion via the newly exposed gp41 molecule. Virus-cell fusion occurs as the transitional intermediate of gp41 undergoes further changes to form a hairpin structure that draws the two membranes into close proximity (see text for details). *(Adapted from D Montefiori, JP Moore: Science 283:336, 1999; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

With activation of the cell, the viral DNA accesses the nuclear pore and is transferred from the cytoplasm to the nucleus, where it is integrated into the host cell chromosomes through the action of another virally encoded enzyme, *integrase* (Fig. 197-3). HIV proviral DNA integrates into the host genomic DNA preferentially in regions of active transcription and regional hotspots. This provirus may remain transcriptionally inactive (latent) or it may manifest varying levels of gene expression, up to active transcription and production of virus depending on the metabolic state of the infected cell.

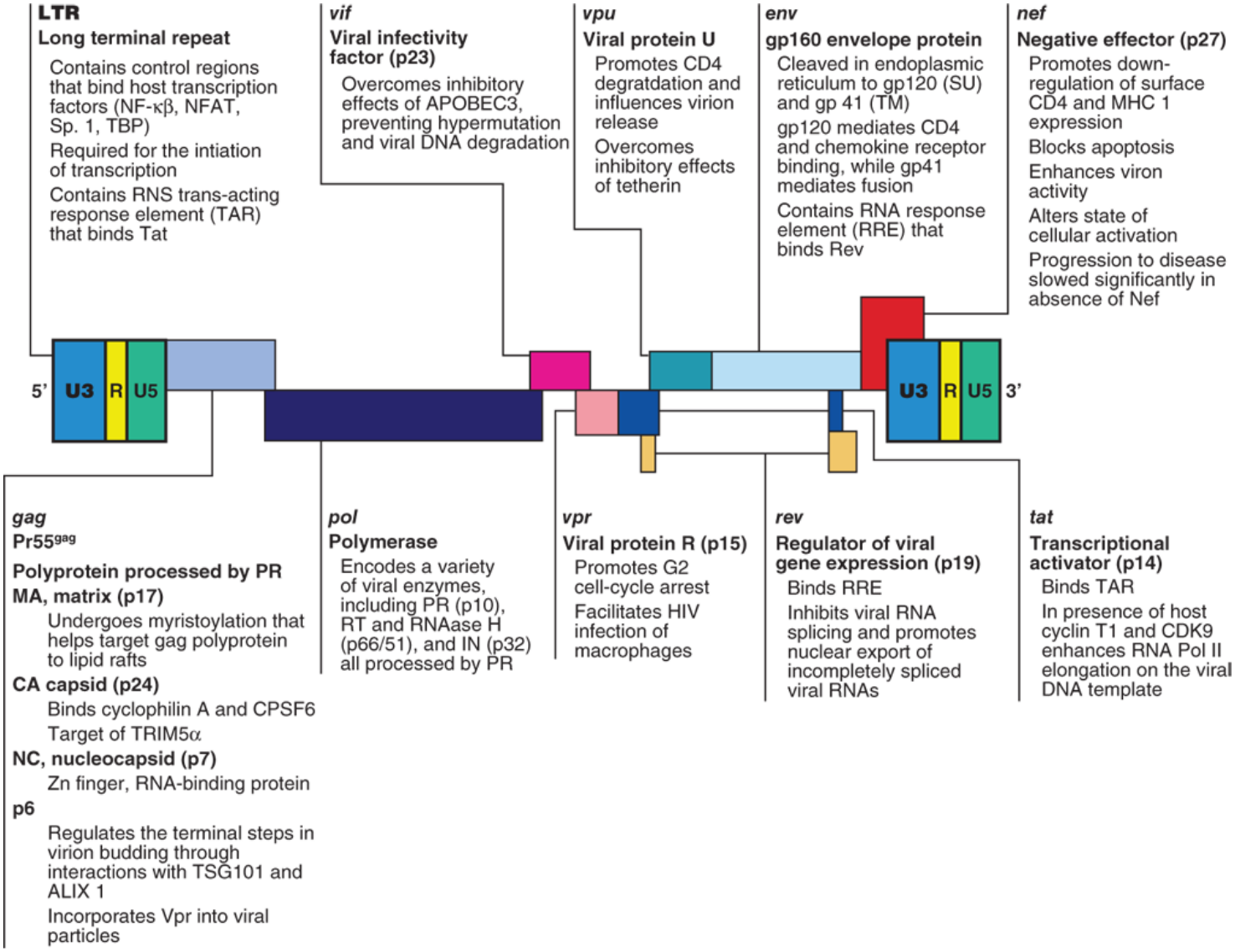
Cellular activation plays an important role in the replication cycle of HIV and is critical to the pathogenesis of HIV disease (see "Pathogenesis and Pathophysiology," below). Following initial binding, fusion, and internalization of the nucleic acid contents of virions into the target cell, incompletely reverse-transcribed DNA intermediates are labile in quiescent cells and do not integrate efficiently into the host cell genome unless cellular activation occurs shortly after infection. Furthermore, some degree of activation of the host cell is required for the initiation of transcription of the integrated proviral DNA into either genomic RNA or mRNA. This latter process may not necessarily be associated with the detectable expression of the classic cell-surface markers of activation. In this regard, activation of HIV expression from the latent state depends on the interaction of a number of cellular and viral factors. Following transcription, HIV mRNA is translated into proteins that undergo modification through glycosylation, myristoylation, phosphorylation, and cleavage. The viral particle is formed by the assembly of HIV proteins, enzymes, and genomic RNA at the plasma membrane of the cells. Budding of the progeny virion through the lipid bilayer of the host cell membrane is the point at which the core acquires its external envelope and where the host restriction factor tetherin can inhibit the release of budding particles. Tetherin is an IFN-induced type II transmembrane protein that interferes with virion detachment, although the HIV accessory protein Vpu counteracts this effect through direct interactions with tetherin. During or soon after budding, the virally encoded protease catalyzes the cleavage of the gag-pol precursor to yield the mature virion. Progression through the virus replication cycle is profoundly influenced by a variety of viral regulatory gene products. Likewise, each point in the replication cycle of HIV is a real or potential target for therapeutic intervention. Thus far, the reverse transcriptase, protease, and integrase enzymes as well as the process of virus–target cell binding and fusion have proved to be susceptible to pharmacologic disruption.

HIV GENOME

Figure 197-5 illustrates schematically the arrangement of the HIV genome. Like other retroviruses, HIV-1 has genes that encode the structural proteins of the virus: *gag* encodes the proteins that form the core of the virion (including p24 antigen); *pol* encodes the enzymes responsible for protease processing of viral proteins, reverse transcription, and integration; and *env* encodes the envelope glycoproteins. However, HIV-1 is more complex than other retroviruses, particularly those of the nonprimate group, in that it also contains at least six other regulatory genes (*tat*, *rev*, *nef*, *vif*, *vpr*, and *vpu*), which code for proteins involved in the modification of the host cell to enhance virus growth and the regulation of viral gene expression. Several of these proteins are thought to play a role in the pathogenesis of HIV disease; their various functions are listed in **Fig. 197-5**. Flanking these genes are the long terminal repeats (LTRs), which contain regulatory elements involved in gene expression (Fig. 197-5). The major difference between the genomes of HIV-1 and HIV-2 is the fact that HIV-2 lacks the *vpu* gene and has a *vpx* gene not contained in HIV-1.

FIGURE 197-5

Organization of the genome of the HIV provirus together with a summary description of its 9 genes encoding 15 proteins. *(Adapted from WC Greene, BM Peterlin: Nat Med 8:673, 2002.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MOLECULAR HETEROGENEITY OF HIV-1

Molecular analyses of HIV isolates reveal varying levels of sequence diversity over all regions of the viral genome. For example, the degree of difference in the coding sequences of the viral envelope protein ranges from a few percent (very close, among isolates from the same infected individual) to more than 50% (extreme diversity, between isolates from the different groups of HIV-1: M, N, O, and P). The changes tend to cluster in hypervariable regions. HIV can evolve by several means, including simple base substitution, insertions and deletions, recombination, and gain and loss of glycosylation sites. HIV sequence diversity arises directly from the limited fidelity of the reverse transcriptase, i.e., a tendency toward copying errors. The balance of immune pressure and functional constraints on proteins influences the regional level of variation within proteins. For example, Envelope, which is exposed on the surface of the virion and is under immune selective pressure from both antibodies and cytolytic T lymphocytes, is extremely variable, with clusters of mutations in hypervariable domains. In contrast, reverse transcriptase, with important enzymatic functions, is relatively conserved, particularly around the active site. The extraordinary variability of HIV-1 contrasts markedly with the relative stability of HTLV-1 and 2.

The four groups (M, N, O and P) of HIV-1 are the result of four separate chimpanzee-to-human (or possibly gorilla-to-human for groups O and P) transfers. Group M (major), which is responsible for most of the infections in the world, has diversified into subtypes and intersubtype recombinant forms, due to "sub-epidemics" within humans after one of those transfers.

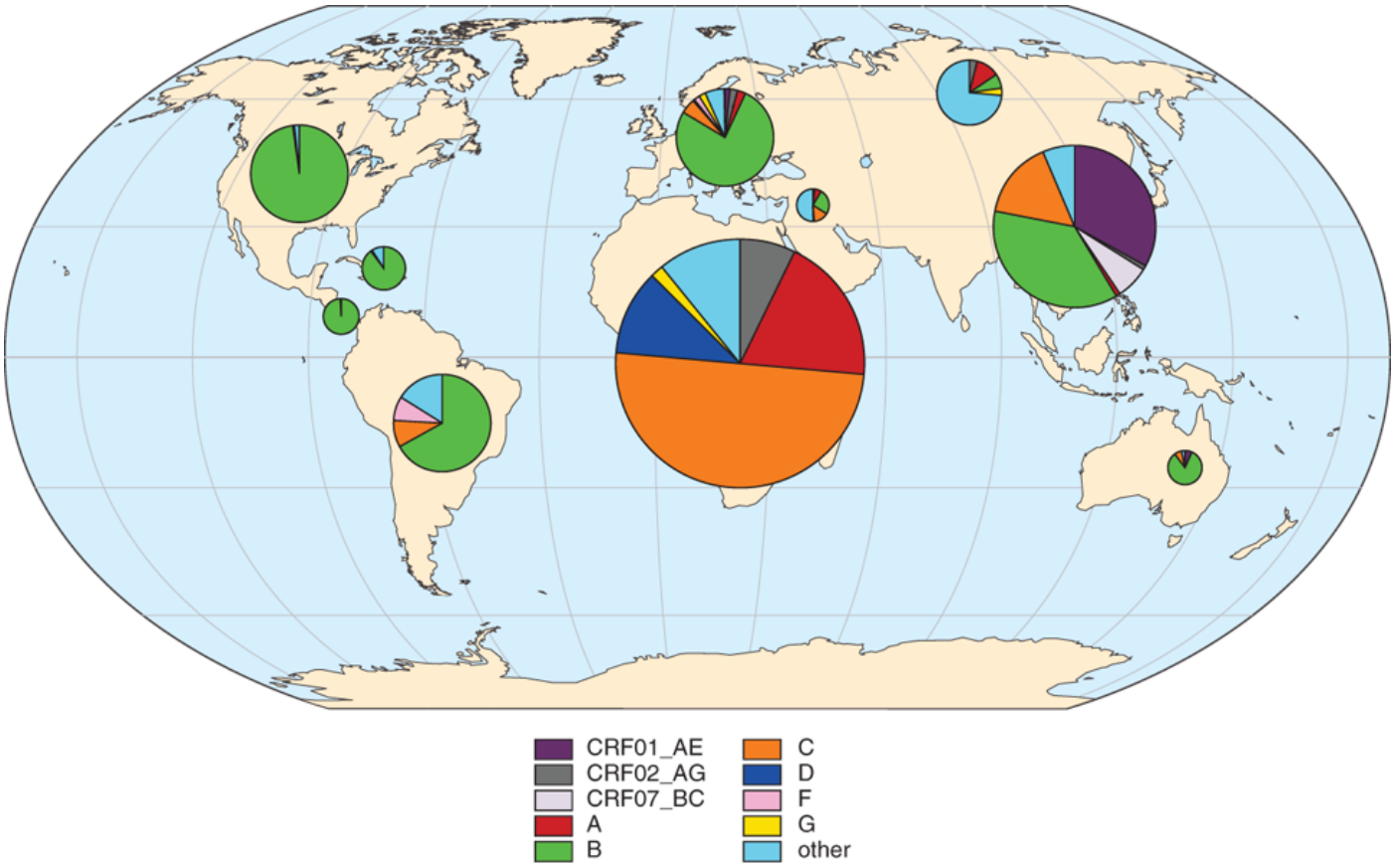
Among primate lentiviruses, HIV-1 is most closely related to viruses isolated from chimpanzees and gorillas (Fig. 197-1). The chimpanzee subspecies *Pan troglodytes troglodytes* has been established to be the natural reservoir of the HIV-1 M and N groups. The rare viruses of the HIV-1 O and P groups are most closely related to viruses found in Cameroonian gorillas. The M group comprises nine subtypes, or *clades*, designated A, B, C, D, F, G, H, J, and K, as well as more than 90 known circulating recombinant forms (CRFs) and numerous unique recombinant forms. Intersubtype recombinants are generated by infection of an individual with two subtypes that then recombine and create a virus with a selective advantage. These CRFs range from highly

prevalent forms such as CRF01_AE, common in southeast Asia, and CRF02_AG from west and central Africa, to a large number of CRFs that are relatively rare, either because they are of a more recent origin (newly recombined) or because they have not broken out into a major population. The subtypes and CRFs create the major lineages of the M group of HIV-1. HIV-1 M group subtype C dominates the global pandemic, and although there is much speculation that it is more transmissible than other subtypes, solid data on variations in transmissibility between subtypes are lacking. Human population densities, access to prevention and treatment, prevalence of genital ulcers, iatrogenic transmissions, and other confounding host factors are all possible reasons why one subtype has spread more than another.

Figure 197-6 schematically diagrams the worldwide distribution of HIV-1 subtypes by region. Nine strains account for the vast majority of HIV infections globally: HIV-1 subtypes A, B, C, D, F, G and three of the CRFs, CRF01_AE, CRF02_AG, and CRF07_BC. Subtype C viruses (of the M group) are by far the most common form worldwide, likely accounting for ~50% of prevalent infections worldwide. In sub-Saharan Africa, home to approximately two-thirds of all individuals living with HIV/AIDS, most infections are caused by subtype C, with smaller proportions of infections caused by subtype A, subtype D, CRF02_AG, and other subtypes and recombinants. In South Africa, the country with the largest number of prevalent infections (7.1 million in 2016), >98% of the HIV-1 isolates sequenced are of subtype C. In Asia, HIV-1 isolates of the CRF01_AE lineage and subtypes B and C predominate. CRF01_AE accounts for most infections in south and southeast Asia, while >95% of infections in India, home to an estimated 2.1 million HIV-infected individuals, are of subtype C (see "HIV Infection and AIDS Worldwide," below). Subtype B viruses are the overwhelmingly predominant viruses seen in the United States, Canada, certain countries in South America, western Europe, and Australia. It is thought that, purely by chance, subtype B was seeded into the United States and Europe in the late 1970s, thereby establishing an overwhelming founder effect. Many countries have co-circulating viral subtypes that are giving rise to new CRFs. Sequence analyses of HIV-1 isolates from infected individuals indicate that recombination among viruses of different clades likely occurs when an individual is infected with viruses of more than one subtype, particularly in geographic areas where subtypes overlap, and more often in sub-epidemics driven by IV drug use than in those driven by sexual transmission.

FIGURE 197-6

Global geographic distribution of HIV-1 subtypes and recombinant forms. Distributions derived from relative frequency of subtypes among >710,000 HIV genomic sequences in the Los Alamos National Laboratory HIV Sequence Database. (Additional information available at *www.hiv.lanl.gov/components/sequence/HIV/geo/geo.comp*.)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The extraordinary diversity of HIV, reflected by the presence of multiple subtypes, circulating recombinant forms, and continuous viral evolution, has implications for possible differential rates of transmission, rates of disease progression, and the development of resistance to antiretroviral drugs. This diversity may also prove to be a formidable obstacle to HIV vaccine development, as a broadly useful vaccine would need to induce protective responses against a wide range of viral strains.

TRANSMISSION

HIV is transmitted primarily by sexual contact (both heterosexual and male to male); by blood and blood products; and by infected mothers to infants intrapartum, perinatally, or via breast milk. After more than 35 years of experience and observations, there is no evidence that HIV is transmitted by any other modality. Table 197-3 lists the estimated risk of HIV transmission for various types of exposures.

TABLE 197-3

Estimated Per-Act Probability of Acquiring HIV From an Infected Source, By Exposure Act

Type of Exposure	Risk per 10,000 Exposures
Parenteral	
Blood transfusion	9250
Needle-sharing during injection drug use	63
Percutaneous (needle-stick)	23
Sexual	
Receptive anal intercourse	138
Insertive anal intercourse	11
Receptive penile-vaginal intercourse	8
Insertive penile-vaginal intercourse	4
Receptive oral intercourse	Low
Insertive oral intercourse	Low
Other[a]	
Biting	Negligible
Spitting	Negligible
Throwing body fluids (including semen or saliva)	Negligible
Sharing sex toys	Negligible

[a]HIV transmission through these exposure routes is technically possible but unlikely and not well documented.

Sources: CDC, *www.cdc.gov/hiv/risk/estimates/riskbehaviors.html*; P Patel: AIDS 28:1509, 2014.

SEXUAL TRANSMISSION

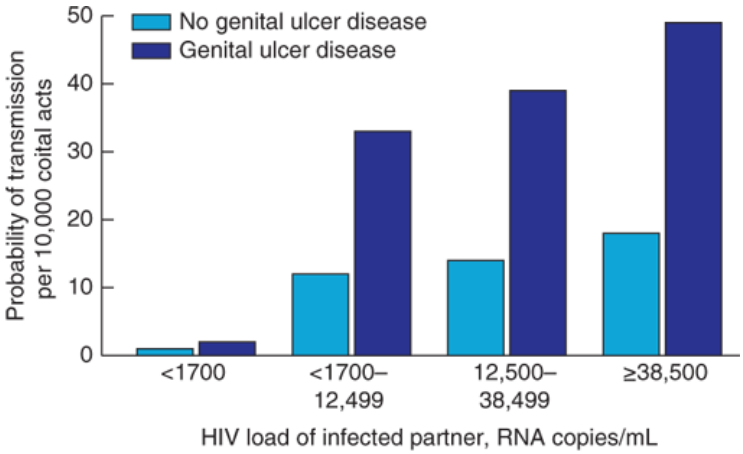
HIV infection is predominantly a sexually transmitted infection (STI) worldwide. By far the most common mode of infection, particularly in developing countries, is heterosexual transmission, although in many western countries male-to-male sexual transmission dominates. Although a wide variety of factors including viral load and the presence of ulcerative genital diseases influence the efficiency of heterosexual transmission of HIV, such transmission is generally inefficient. A recent systemic review found a low per-act risk of heterosexual transmission in the absence of antiretrovirals: 0.04% for female-to-male transmission and 0.08% for male-to-female transmission during vaginal intercourse in the absence of antiretroviral therapy or condom use (Table 197-3).

HIV has been demonstrated in seminal fluid both within infected mononuclear cells and in cell-free material. The virus appears to concentrate in the seminal fluid, particularly in situations where there are increased numbers of lymphocytes and monocytes in the fluid, as seen in genital inflammatory states such as urethritis and epididymitis, conditions closely associated with other STIs. The virus has also been demonstrated in cervical smears and vaginal fluid. There is an elevated risk of HIV transmission associated with unprotected receptive anal intercourse (URAI) among both men and women compared to the risk associated with unprotected receptive vaginal intercourse. Although data are limited, the per-act risk for HIV transmission via URAI has been estimated to be ~1.4% (Table 197-3). The risk of HIV acquisition associated with URAI is higher than that seen in penile-vaginal intercourse probably because only a thin, fragile rectal mucosal membrane separates the deposited semen from potentially susceptible cells in and beneath the mucosa, and micro-trauma of the mucosal membrane has been associated with anal intercourse. Anal douching and sexual practices that traumatize the rectal mucosa also increase the likelihood of infection. It is likely that anal intercourse provides at least two modalities of infection: (1) direct inoculation into blood in cases of traumatic tears in the mucosa; and (2) infection of susceptible target cells, such as Langerhans cells, in the mucosal layer in the absence of trauma. Insertive anal intercourse also confers an increased risk of HIV acquisition compared to insertive vaginal intercourse in the receptive partner since the vaginal mucosa is several layers thicker than the rectal mucosa and less likely to be traumatized during intercourse. Nonetheless, the virus can be transmitted to either partner through vaginal intercourse. As noted in Table 197-3, male-to-female HIV transmission is more efficient than female-to-male transmission. The differences in reported transmission rates between men and women may be due in part to the prolonged exposure to infected seminal fluid of the vaginal and cervical mucosa, as well as the endometrium (when semen enters through the cervical os). By comparison, the penis and urethral orifice of the uninfected male partner are exposed relatively briefly to infected vaginal fluid. Among various cofactors examined in studies of heterosexual HIV transmission, the presence of other STIs has been strongly associated with HIV transmission. In this regard, there is a close association between genital ulcerations and transmission, owing to both susceptibility to infection and infectivity. Infections with microorganisms such as *Treponema pallidum* **(Chap. 177)**, *Haemophilus ducreyi* **(Chap. 152)**, and herpes simplex virus (HSV; **Chap. 187**) are important causes of genital ulcerations linked to transmission of HIV. In addition, pathogens responsible for non-ulcerative inflammatory STIs such as those caused by *Chlamydia trachomatis* **(Chap. 184)**, *Neisseria gonorrhoeae* **(Chap. 151)**, and *Trichomonas vaginalis* **(Chap. 224)** also are associated with an increased risk of transmission of HIV infection. Bacterial vaginosis, an infection related to sexual behavior, but not strictly an STI, also may be linked to an increased risk of transmission of HIV infection. Several studies suggest that treating other STIs and genital tract syndromes may help prevent transmission of HIV. This effect is most prominent in populations in which the prevalence of HIV infection is relatively low. It is noteworthy that this principle may not apply to the treatment of HSV infections since it has been shown that even following anti-HSV therapy with resulting healing of HSV-related genital ulcers, HIV acquisition is not reduced. Biopsy studies revealed that the likely explanation is that HIV receptor–positive inflammatory cells persisted in the genital tissue despite the healing of ulcers, and so HIV-susceptible targets remained at the site.

The quantity of HIV-1 in plasma (viral load) is a primary determinant of the risk of HIV-1 transmission. In a cohort of heterosexual couples in Uganda discordant for HIV infection and not receiving antiretroviral therapy, the mean serum HIV RNA level was significantly higher among HIV-infected subjects whose partners seroconverted than among those whose partners did not seroconvert. In fact, transmission was rare when the infected partner had a plasma level of <1700 copies of HIV RNA per milliliter, even when genital ulcer disease was present **(Fig. 197-7)**. The rate of HIV transmission per coital act was highest during the early stage of HIV infection when plasma HIV RNA levels were high and in advanced disease with high viral set points.

FIGURE 197-7

Probability of HIV transmission per coital act among monogamous, heterosexual, HIV-serodiscordant couples in Uganda. *(From RH Gray et al: Lancet 357:1149, 2001.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Antiretroviral therapy dramatically reduces plasma viremia in most HIV-infected individuals (see "Treatment," below) and is associated with a dramatic reduction in risk of transmission. In a large study of serodiscordant couples, earlier treatment of the HIV-infected partner with antiretroviral therapy rather than treatment delayed until the CD4+ T cell counts fell below 250 cells per μL was associated with a 96% reduction in HIV transmission to the uninfected partner. This approach has been widely referred to as *treatment as prevention* or TasP. Recent cohort studies have indicated that if the viral

load of the infected partner is decreased to below detectable levels by antiretroviral therapy, there is essentially no chance of sexual transmission to the uninfected partner.

A number of studies including large, randomized, controlled trials clearly have indicated that male *circumcision* is associated with a lower risk of acquisition of HIV infection for heterosexual men. Studies also suggest that circumcision is protective in those men who have sex with men who are insertive only. The benefit of circumcision may be due to increased susceptibility of uncircumcised men to ulcerative STIs, as well as to other factors such as microtrauma to the foreskin and glans penis. In addition, the highly vascularized inner layer of foreskin tissue contains a high density of Langerhans cells as well as increased numbers of CD4+ T cells, macrophages, and other cellular targets for HIV. Finally, the moist environment under the foreskin may promote the presence or persistence of microbial flora that, via inflammatory changes, may lead to even higher concentrations of target cells for HIV in the foreskin. In addition, randomized clinical trials have demonstrated that male circumcision also reduces herpes simplex virus (HSV) type 2, human papillomavirus virus (HPV), and genital ulcer disease in men as well as HPV, genital ulcer disease, bacterial vaginosis, and *Trichomonas vaginalis* infections among female partners of circumcised men. Thus, there may be an added indirect benefit of diminution of risk for HIV acquisition to the female sexual partners of circumcised men.

In some studies the use of oral contraceptives was associated with an increase in incidence of HIV infection over and above that which might be expected by not using a condom for birth control. This phenomenon may be due to drug-induced changes in the cervical mucosa, rendering it more vulnerable to penetration by the virus. Adolescent girls might also be more susceptible to infection upon exposure due to the properties of an immature genital tract with increased cervical ectopy or exposed columnar epithelium.

Oral sex is a much less efficient mode of transmission of HIV than is anal intercourse or vaginal intercourse (Table 197-3). A number of studies have reported that the incidence of transmission of infection by oral sex among couples discordant for HIV was extremely low. However, there have been well-documented reports of HIV transmission that likely resulted from fellatio or cunnilingus. Therefore, the assumption that oral sex is completely safe is not warranted.

The association of alcohol consumption and illicit drug use with unsafe sexual behavior, both homosexual and heterosexual, leads to an increased risk of sexual transmission of HIV. Methamphetamine and other so-called club drugs (e.g., MDMA, ketamine, and gamma hydroxybutyrate), sometimes taken in conjunction with PDE-5 inhibitors such as sildenafil (Viagra), tadalafil (Cialis), or vardenafil (Levitra), have been associated with risky sexual practices and increased risk of HIV infection, particularly among men who have sex with men.

TRANSMISSION THROUGH INJECTION DRUG USE

HIV can be transmitted to injection drug users (IDUs) who are exposed to HIV while sharing injection paraphernalia such as needles, syringes, the water in which drugs are mixed, or the cotton through which drugs are filtered. Parenteral transmission of HIV during injection drug use does not require IV puncture; subcutaneous (“skin popping”) or intramuscular (“muscling”) injections can transmit HIV as well, even though these behaviors are sometimes erroneously perceived as low-risk. Among IDUs, the risk of HIV infection increases with the duration of injection drug use; the frequency of needle sharing; the number of partners with whom paraphernalia are shared, particularly in the setting of “shooting galleries” where drugs are sold and large numbers of IDUs may share a limited number of “works”; comorbid psychiatric conditions such as antisocial personality disorder; the use of cocaine in injectable form or smoked as “crack”; and the use of injection drugs in a geographic location with a high prevalence of HIV infection. As noted in Table 197-3, the per-act risk of transmission from injection drug use with a contaminated needle has been estimated to be approximately 0.6%.

TRANSMISSION BY TRANSFUSED BLOOD AND BLOOD PRODUCTS

HIV can be transmitted to individuals who receive HIV-contaminated blood transfusions, blood products, or transplanted tissue. The vast majority of HIV infections acquired via contaminated blood transfusions, blood components, or transplanted tissue in resource-rich countries occurred prior to the spring of 1985, when mandatory testing of donated blood for HIV-1 was initiated. It is estimated that >90% of individuals exposed to HIV-contaminated blood products become infected (Table 197-3). Transfusions of whole blood, packed red blood cells, platelets, leukocytes, and plasma are all capable of transmitting HIV infection. In contrast, hyperimmune gamma globulin, hepatitis B immune globulin, plasma-derived hepatitis B vaccine, and Rh_0 immune globulin have not been associated with transmission of HIV infection. The procedures involved in processing these products either inactivate or remove the virus.

Currently, in the United States and in most developed countries, the following measures have made the risk of transmission of HIV infection by transfused blood or blood products extremely small: the screening of blood donations for antibodies to HIV-1 and HIV-2 and determination of the presence of HIV nucleic acid usually in minipools of several specimens; the careful selection of potential blood donors with health history questionnaires to exclude individuals with risk behavior; and opportunities for self-deferral and the screening out of HIV-negative individuals with serologic testing for infections that have shared risk factors with HIV, such as hepatitis B and C and syphilis. The chance of infection of a hemophiliac via clotting factor concentrates has essentially been eliminated because of standard screening of blood together with the added layer of safety resulting from heat treatment of the concentrates. It is currently estimated that the risk of infection with HIV in the United States via transfused screened blood is approximately 1 in 1.5 million units. Therefore, since nearly 21 million blood components are transfused in the United States each year, despite the best

efforts of science, one cannot completely eliminate the risk of transfusion-related transmission of HIV. In this regard, a case of transfusion-related transmission of HIV was reported in the United States in 2010, which was tracked to a blood donation in 2008; this was the first such reported case since 2002 and only the third in that decade. Transmission of HIV (both HIV-1 and HIV-2) by blood or blood products is still an ongoing threat in certain developing countries where routine screening of blood is not universally practiced. In 2013, 108 out of 167 countries (65%) had specific legislation covering the safety and quality of blood transfusion, including 79% of high-income countries, 64% of middle-income countries, and 41% of low-income countries. Furthermore, there have been reports in certain countries of sporadic breakdowns in routinely available screening procedures in which contaminated blood was allowed to be transfused, resulting in small clusters of patients becoming infected.

OCCUPATIONAL TRANSMISSION OF HIV: HEALTH CARE WORKERS, LABORATORY WORKERS, AND THE HEALTH CARE SETTING

There is a small but definite occupational risk of HIV transmission to health care workers and laboratory personnel and potentially others who work with HIV-containing materials, particularly when sharp objects are used. An estimated 600,000 to 800,000 health care workers are stuck with needles or other sharp medical instruments in the United States each year. The global number of HIV infections among health care workers attributable to sharps injuries has been estimated to be 1000 cases (range, 200–5000) per year. In the United States, 58 documented cases of occupational HIV transmission to health care workers, and 150 possible transmissions have been reported by the CDC. There have been no confirmed cases reported since 1999.

Exposures that place a health care worker at potential risk of HIV infection are percutaneous injuries (e.g., a needle stick or cut with a sharp object) or contact of mucous membrane or nonintact skin (e.g., exposed skin that is chapped, abraded, or afflicted with dermatitis) with blood, tissue, or other potentially infectious body fluids. Large, multi-institutional studies have indicated that the risk of HIV transmission following skin puncture from a needle or a sharp object that was contaminated with blood from a person with documented HIV infection is ~0.23% and after a mucous membrane exposure it is 0.09% (see "HIV and the Health Care Worker," below) if the injured and/or exposed person is not treated within 24 h with antiretroviral drugs. The risk of hepatitis B virus (HBV) infection following a similar type of exposure is ~6–30% in nonimmune individuals; if a susceptible worker is exposed to HBV, postexposure prophylaxis with hepatitis B immune globulin and initiation of HBV vaccine is >90% effective in preventing HBV infection. The risk of HCV infection following percutaneous injury is ~1.8% **(Chap. 332)**.

Rare HIV transmission after nonintact skin exposure has been documented, but the average risk for transmission by this route has not been precisely determined; however, it is estimated to be less than the risk for mucous membrane exposure. Transmission of HIV through intact skin has not been documented. Currently in developed countries, virtually all puncture wounds and mucous membrane exposures in health care workers involving blood from a patient with documented HIV infection are treated prophylactically with combination antiretroviral therapy (cART). This practice, referred to as *postexposure prophylaxis* or PEP, has dramatically reduced the occurrence of puncture-related transmissions of HIV to health care workers.

In addition to blood and visibly bloody body fluids, semen and vaginal secretions also are considered potentially infectious; however, they have not been implicated in occupational transmission from patients to health care workers. The following fluids also are considered potentially infectious: cerebrospinal fluid, synovial fluid, pleural fluid, peritoneal fluid, pericardial fluid, and amniotic fluid. The risk for transmission after exposure to fluids or tissues other than HIV-infected blood also has not been quantified, but it is probably considerably lower than the risk after blood exposures. Feces, nasal secretions, saliva, sputum, sweat, tears, urine, and vomitus are not considered potentially infectious for HIV unless they are visibly bloody. Rare cases of HIV transmission via human bites have been reported, but not in the setting of occupational exposure.

An increased risk for HIV infection following percutaneous exposures to HIV-infected blood is associated with exposures involving a relatively large quantity of blood, as in the case of a device visibly contaminated with the patient's blood, a procedure that involves a hollow-bore needle placed directly in a vein or artery, or a deep injury. Factors that might be associated with mucocutaneous transmission of HIV include exposure to an unusually large volume of blood and prolonged contact. In addition, the risk increases for exposures to blood from untreated patients with high levels of HIV in the blood. Since the beginning of the HIV epidemic, there have been rare instances where transmission of infection from a health care worker to patients seemed highly probable. Despite these small number of documented cases, the risk of HIV transmission involving health care workers (infected or not) to patients is extremely low in developed countries—in fact, too low to be measured accurately. In this regard, several retrospective epidemiologic studies have been performed tracing thousands of patients of HIV-infected dentists, physicians, surgeons, obstetricians, and gynecologists, and no other cases of HIV transmission that could be linked to the health care providers were identified.

Breaches in infection control and the reuse of contaminated syringes, failure to properly sterilize surgical instruments, and/or hemodialysis equipment also have resulted rarely in the transmission of HIV from patient to patient in hospitals, nursing homes, and outpatient settings. Finally, these very rare occurrences of transmission of HIV as well as HBV and HCV to and from health care workers in the workplace underscore the importance of the use of universal precautions when caring for all patients (see below and **Chap. 137**).

MOTHER-TO-CHILD TRANSMISSION OF HIV

HIV infection can be transmitted from an infected mother to her fetus during pregnancy, during delivery, or by breast-feeding. This remains an important form of transmission of HIV infection in some developing countries. Virologic analyses of aborted fetuses indicate that HIV can be transmitted to the fetus during the first or second trimesters of pregnancy. However, maternal transmission to the fetus occurs most commonly in the perinatal period. Two

studies performed in Rwanda and the Democratic Republic of Congo (then called Zaire) indicated that among all mother-to-child transmissions of HIV, the relative proportions were 23–30% before birth, 50–65% during birth, and 12–20% via breast-feeding.

In the absence of antiretroviral therapy for the mother during pregnancy, labor, and delivery, and for the fetus prophylactically following birth, the probability of transmission of HIV from mother to infant/fetus ranges from 15% to 25% in industrialized countries and from 25% to 35% in developing countries. These differences may relate to the adequacy of prenatal care as well as to the stage of HIV disease and the general health of the mother during pregnancy. Higher rates of transmission have been reported to be associated with many factors—the best documented of which is the presence of high maternal levels of plasma viremia, with the risk increasing linearly with the level of maternal plasma viremia. It is very unlikely that mother-to-child transmission will occur if the mother's level of plasma viremia is <1000 copies of HIV RNA/mL of blood and extremely unlikely if the level is undetectable (i.e., <50 copies/mL). However, there may not be a lower "threshold" below which transmission never occurs, since certain studies have reported rare transmission by women with viral RNA levels <50 copies/mL. Increased mother-to-child transmission is also correlated with closer human leukocyte antigen (HLA) match between mother and child. A prolonged interval between membrane rupture and delivery is another well-documented risk factor for transmission. Other conditions that are potential risk factors, but that have not been consistently demonstrated, are the presence of chorioamnionitis at delivery; STIs during pregnancy; illicit drug use during pregnancy; cigarette smoking; preterm delivery; and obstetric procedures such as amniocentesis, amnioscopy, fetal scalp electrodes, and episiotomy. Today, the rate of mother-to-child transmission has fallen to 1% or less in pregnant women who are receiving cART for their HIV infection. Such treatment, combined with cesarean section delivery, has rendered mother-to-child transmission of HIV an unusual event in the United States and other developed nations. In this regard, both the United States Public Health Service and the World Health Organization guidelines recommend that all pregnant HIV-infected women should receive life-long cART for the health of the mother and to prevent perinatal transmission regardless of plasma HIV RNA copy number or CD4+ T cell counts.

Breast-feeding is an important modality of transmission of HIV infection in certain developing countries, particularly where mothers continue to breast-feed for prolonged periods. The risk factors for mother-to-child transmission of HIV via breast-feeding include detectable levels of HIV in breast milk, the presence of mastitis, low maternal CD4+ T cell counts, and maternal vitamin A deficiency. The risk of HIV infection via breast-feeding is highest in the early months of breast-feeding. In addition, exclusive breast-feeding has been reported to carry a lower risk of HIV transmission than mixed feeding. In developed countries, breast feeding of babies by an HIV-infected mother is contraindicated since alternative forms of adequate nutrition, i.e., formulas, are readily available. In developing countries, where breast-feeding may be essential for the overall health of the infant, the continuation of cART in the infected mother during the period of breastfeeding markedly diminishes the risk of transmission of HIV to the infant. In fact, once cART has been initiated in a pregnant woman, it should be continued for life.

TRANSMISSION OF HIV BY OTHER BODY FLUIDS

Although HIV can be isolated typically in low titers from saliva of a small proportion of infected individuals, there is no convincing evidence that saliva can transmit HIV infection, either through kissing or through other exposures, such as occupationally to health care workers. Saliva contains endogenous antiviral factors; among these factors, HIV-specific immunoglobulins of IgA, IgG, and IgM isotypes are detected readily in salivary secretions of infected individuals. It has been suggested that large glycoproteins such as mucins and thrombospondin 1 sequester HIV into aggregates for clearance by the host. In addition, a number of soluble salivary factors inhibit HIV to various degrees in vitro, probably by targeting host cell receptors rather than the virus itself. Perhaps the best studied of these, secretory leukocyte protease inhibitor (SLPI), blocks HIV infection in several cell culture systems, and it is found in saliva at levels that approximate those required for inhibition of HIV in vitro. In this regard, higher salivary levels of SLPI in breast-fed infants were associated with a decreased risk of HIV transmission through breast milk. It has also been suggested that submandibular saliva reduces HIV infectivity by stripping gp120 from the surface of virions, and that saliva-mediated disruption and lysis of HIV-infected cells occurs because of the hypotonicity of oral secretions. There have been outlier cases of suspected transmission by saliva, but these have probably been blood-to-blood transmissions. Transmission of HIV by a human bite can occur but is a rare event. Although virus can be identified, if not isolated, from virtually any body fluid, there is no evidence that HIV transmission can occur as a result of exposure to tears, sweat, or urine. However, there have been isolated cases of transmission of HIV infection by body fluids that may or may not have been contaminated with blood. Most of these situations occurred in the setting of a close relative providing intensive nursing care for an HIV-infected person without observing universal precautions, underscoring the importance of adhering to such precautions in the handling of body fluids and wastes from HIV-infected individuals.

EPIDEMIOLOGY

HIV INFECTION AND AIDS WORLDWIDE

HIV infection/AIDS is a global pandemic, with cases reported from virtually every country. At the end of 2016, an estimated 36.7 million individuals were living with HIV infection, according to the Joint United Nations Programme on HIV/AIDS (UNAIDS). An estimated 95% of people living with HIV/AIDS reside in low- and middle-income countries; ~50% are female, and 2.1 million are children <15 years. The regional distribution of these cases is illustrated in **Fig. 197-8**. The estimated number of people living with HIV—i.e., the global prevalence—has increased more than fourfold since 1990, reflecting the combined effects of continued high rates of new HIV infections and the life-prolonging impact of antiretroviral therapy **(Fig. 197-9)**. In 2016,

the global prevalence of HIV infection among people aged 15–49 years was 0.8%, with rates varying widely by country and region as illustrated in **Fig. 197-10**.

FIGURE 197-8

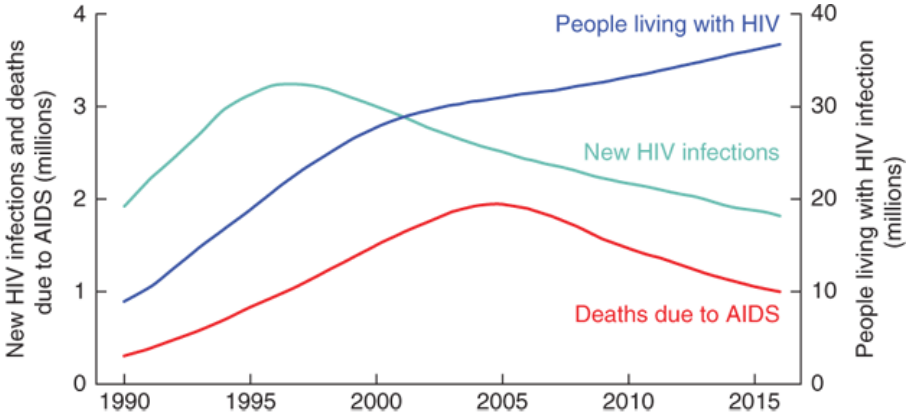
Estimated number of adults and children living with HIV infection as of December, 2016. Total: 36.7 million (30.8 million–42.9 million). *(From Joint United Nations Programme on HIV/AIDS [UNAIDS].)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 197-9

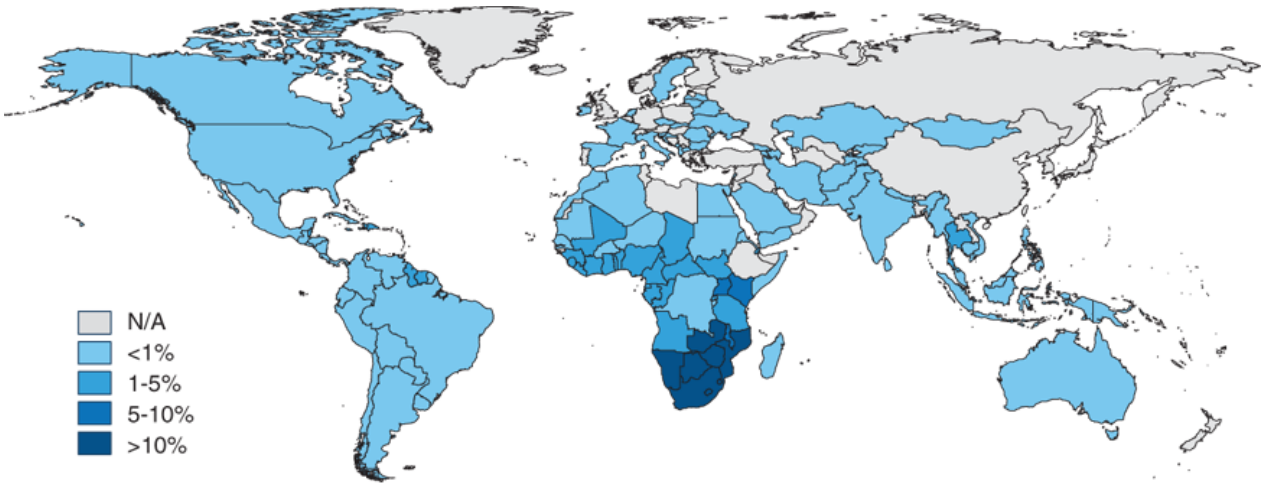
Global estimates of HIV incidence, AIDS deaths, and HIV prevalence 1990–2016. *(From UNAIDS.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 197-10

Adult HIV prevalence rates by country, 2016. Data are estimates for adults age 15–49 years. *(From UNAIDS.)*

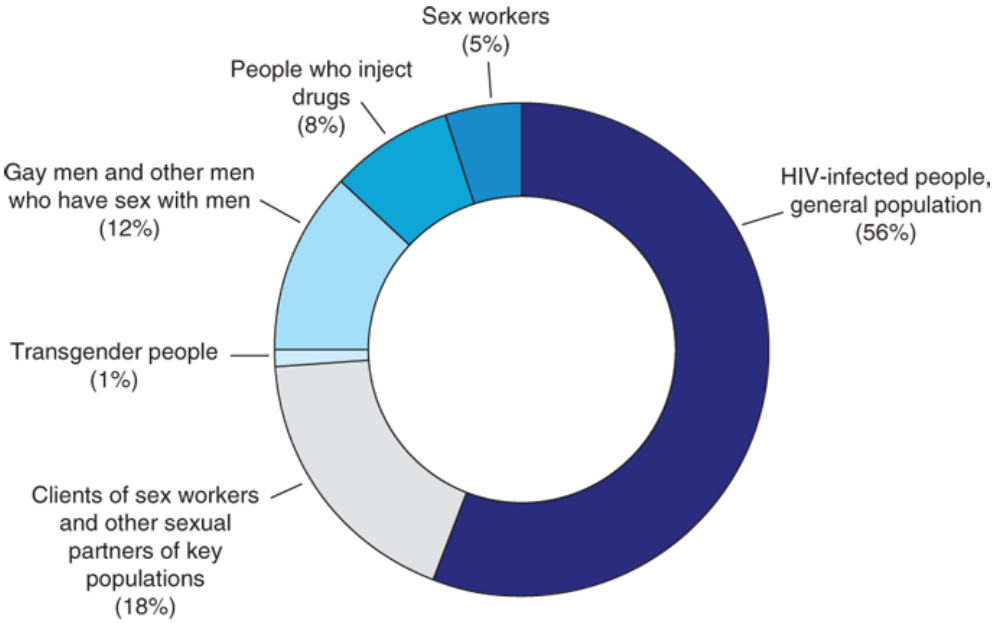


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In 2016, an estimated 1.8 million new cases of HIV infection occurred worldwide, including 160,000 among children <15 years; about one-third of new infections were among people age 15–24 years. Globally, the majority of new HIV infections are due to heterosexual transmission. Members of certain high-risk populations are disproportionately affected. Sex workers, people who inject drugs, transgender people, prisoners, gay men, other men who have sex with men, and their sexual partners accounted for 34% of all new HIV infections in 2015 **(Fig. 197-11)**.

FIGURE 197-11

Global distribution of new HIV infections by population. Data for 2015. *(From UNAIDS.)*



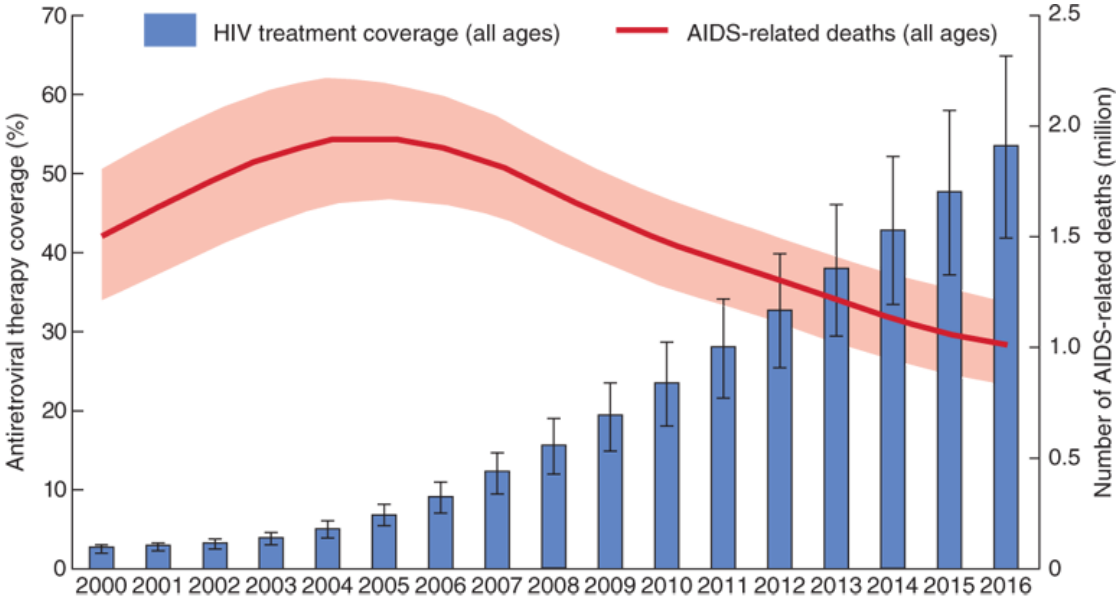
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Between 2000 and 2016, the estimated annual number of new HIV infections globally fell by 40% (Fig. 197-9). These reductions in global HIV incidence likely reflect progress with HIV prevention efforts and the increased provision to HIV-infected people of antiretroviral therapy, which makes them much less likely to transmit the virus to sexual partners. Among adults, the estimated incidence declined by 11% from 2010 to 2016. From 2010 to 2016 there was a ~47% reduction in HIV infections among children <15 years, progress that is due largely to the increasing availability of antiretroviral medications to prevent the transmission of HIV from mother to infant.

In 2016, global AIDS deaths totaled 1.0 million (including 120,000 children <15 years), a 48% decrease since 2005 that coincides with a rapid expansion of access to antiretroviral therapy **(Fig. 197-12)**. Since the beginning of the pandemic, an estimated 35 million people have died of an AIDS-related illness.

FIGURE 197-12

Global antiretroviral therapy coverage and number of AIDS-related deaths, 2000–2016. *(From UNAIDS).*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The HIV epidemic has occurred in "waves" in different regions of the world, each wave having somewhat different characteristics depending on the demographics of the country and region in question and the timing of the introduction of HIV into the population. Although the AIDS epidemic was first recognized in the United States and shortly thereafter in Western Europe, it very likely began in sub-Saharan Africa (see above), which has been particularly devastated by the epidemic. East and Southern Africa is the region hardest hit by HIV. The region is home to 6.2% of the world's population but has 19.4 million people living with HIV, >50% of the global total (Fig. 197-8). In eight countries in the region, >10% of the adult population age 15–49 is HIV-infected (Fig. 197-10). South Africa has the highest number of people living with HIV in the world (7.1 million); Swaziland has the highest adult HIV prevalence in the world (27.2%). Among high-risk individuals, rates are much higher than in the general population. HIV prevalence among sex workers varies between 50% and 70% in several countries in the region. Recent data offer promising signs of declining HIV incidence and prevalence in many countries in the region, although frequently at levels that remain high. Heterosexual exposure is the primary mode of HIV transmission in most countries in sub-Saharan Africa. Women and girls account for ~60 percent of all HIV infections in that region.

The 25 countries of West and Central Africa are home to 6.1 million people living with HIV, of whom half a million are children. HIV prevalence in most of the countries is relatively low compared with East and Southern Africa. HIV prevalence among adults across the region overall stands at 2.2% although there is wide variation between countries, ranging from 0.5% in Niger and Senegal to 4.9% in Equatorial Guinea. An estimated 60% of new infections in the region in 2015 occurred in Nigeria. As in East and Southern Africa, heterosexual transmission accounts for most HIV transmission West and Central Africa.

The Middle East and North Africa region has one of the lowest HIV prevalence rates in the world (0.1%). In 2016, an estimated 230,000 people were living with HIV in the region. Cases are largely concentrated among IDUs, men who have sex with men, and sex workers and their clients.

In Asia and the Pacific, an estimated 5.1 million people were living with HIV at the end of 2016. In this region of the world, HIV prevalence is highest in southeast Asian countries, with wide variation in trends between different countries. Among countries in Asia, only Thailand has an adult seroprevalence rate of >1%. However, the populations of many Asian nations are so large that even low infection and seroprevalence rates result in large numbers of people living with HIV. In this regard, three populous countries—China, India and Indonesia—account for around three-quarters of all people living with HIV in the region. Although the HIV epidemic in Asia has long been concentrated among specific populations—sex workers and their clients, men who have sex with men, and IDUs—it is expanding to the heterosexual partners of those most at risk.

Eastern Europe and Central Asia is the only region in the world where the HIV epidemic continues to expand rapidly, with a >60% increase in annual new HIV infections between 2010 and 2016. The Russian Federation and Ukraine account for the majority of HIV cases in the region, where the epidemic has been driven by injection drug use and increasingly by heterosexual transmission.

Approximately 2.1 million people were living with HIV/AIDS in Latin America and the Caribbean at the end of 2016. The rate of new HIV infections in the region held steady from 2010 to 2016. Brazil is home to the largest number of HIV-infected persons (830,00) in the region, and the Bahamas has the region's highest prevalence (3.3%). Men who have sex with men account for the largest proportion of HIV infections in Central and South America. In the Caribbean, heterosexual transmission, often tied to sex work, is the main driver of transmission.

Approximately 2.1 million people were living with HIV/AIDS in North America and western and central Europe at the end of 2016. While modes of transmission vary greatly by country, HIV disproportionately affects men who have sex with men. Over the past decade, the number of HIV diagnoses

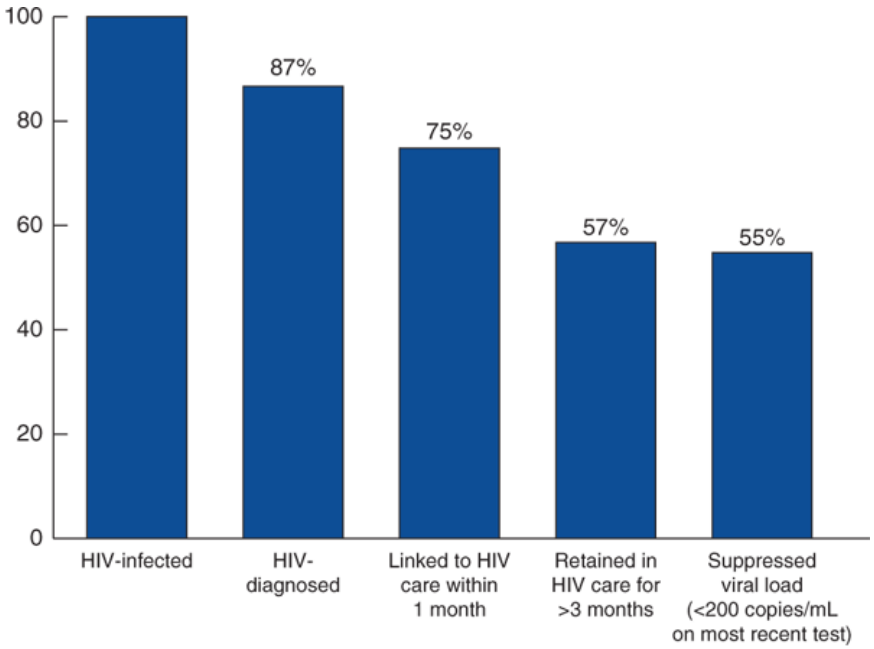
decreased dramatically in all risk groups in western Europe but increased slightly in central Europe. North America saw a decrease in HIV diagnoses overall but a small increase among gay and bisexual men.

HIV INFECTION AND AIDS IN THE UNITED STATES

About 1.8 million people have been infected with HIV in the United States since the beginning of the epidemic, of whom ~693,000 have died. Approximately 1.1 million individuals in the United States are living with HIV infection, ~15% of whom are unaware of their infection, according to recent estimates. As illustrated in **Fig. 197-13**, only about half of HIV-infected people in the United States have been able to negotiate the steps in the HIV "care continuum," from diagnosis, to entering into care and receiving antiretroviral therapy, and ultimately to achieving a suppressed viral load (see "Treatment," below).

FIGURE 197-13

Estimated percentage of HIV-infected people engaged at selected stages of the continuum of HIV care in the United States. *(From Centers for Disease Control and Prevention [CDC]: HIV Surveillance Supplemental Report 21[No. 7], 2016.)*



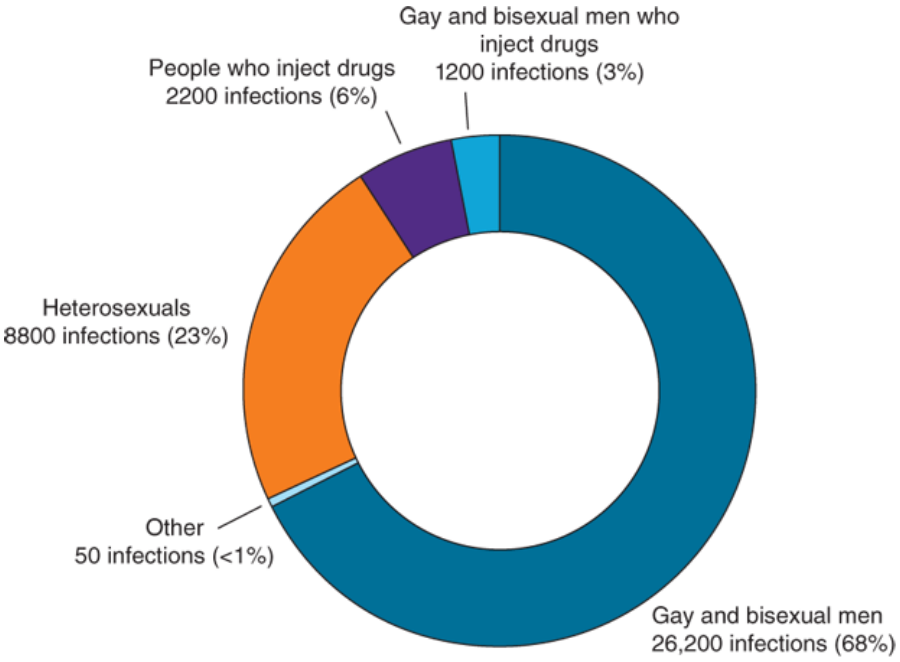
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

More than 60% of people living with HIV in the United States are Black/African American or Hispanic/Latino, and more than half are men who have sex with men. The estimated HIV seroprevalence rate among all individuals age 13 years or older in the United States is ~0.5%. Approximately 2% of Black/African-American adults are HIV-infected in the United States, higher than any other group.

The number of new HIV infections in the United States, *HIV incidence*, peaked at about 130,000 per year in the late 1980s, followed by declines. After remaining stable since the mid-1990s, the estimated number of annual HIV infections in the United States fell ~15% between 2008 and 2015 (from 45,200 to 38,500). The distribution of incident HIV cases in 2015 is shown in **Fig. 197-14**. Gay and bisexual men account for more than two-thirds of incident infections and were the only group that did not experience an overall decline in annual HIV infections from 2008 to 2015. While infections among white gay and bisexual men and men age 15–24 years fell during that period, these declines were offset by increases among 25- to 34-year-old gay and bisexual men, and among Hispanic/Latino gay and bisexual males.

FIGURE 197-14

Estimated distribution of new HIV infections in the United States. Total: 38,500. Incidence estimate for 2015. *(From CDC.)*

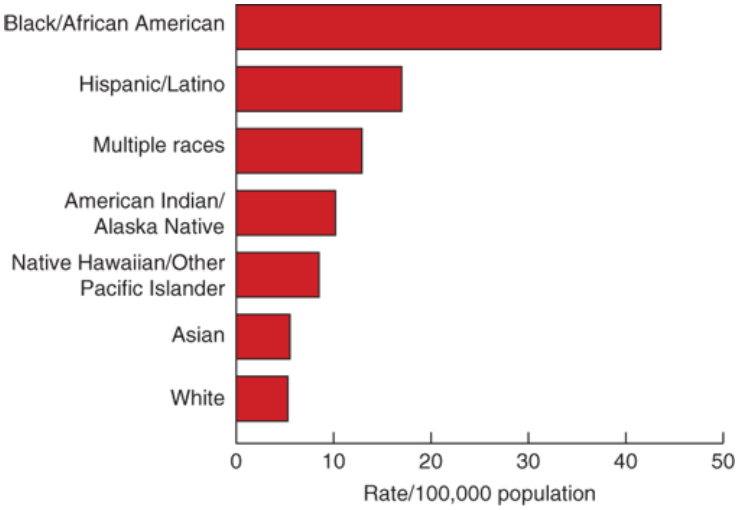


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In the United States, the burden of HIV and AIDS is not evenly distributed across states and regions. In most areas of the country, HIV is concentrated in urban areas. In the southern United States, larger percentages of diagnoses are in smaller metropolitan and nonmetropolitan areas. HIV infection and AIDS have disproportionately affected minority populations in the United States in both urban and rural areas. Among those diagnosed with HIV (regardless of stage of infection) in 2016, 44% percent were Blacks/African Americans, a group that constitutes only 12% of the U.S. population. The estimated rate of new HIV diagnoses in 2016 by race/ethnicity per 100,000 population in the United States is shown in Fig. 197-15.

FIGURE 197-15

Estimated rate of HIV infections (including children) diagnosed during 2016 in the United States, by race/ethnicity (per 100,000 population). *(From CDC.)*



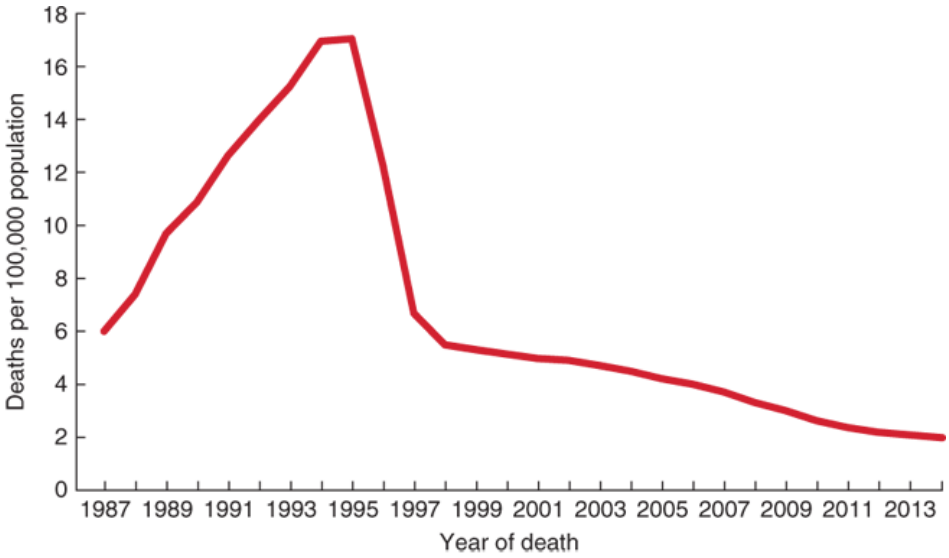
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Perinatal HIV transmission, from an HIV-infected mother to her baby, has declined significantly in the United States, largely due to the implementation of guidelines for the universal counseling and voluntary HIV testing of pregnant women and the use of antiretroviral therapy for pregnant women and newborn infants to prevent infection. In 2016, 122 children were newly diagnosed with HIV infection in the United States, down from a peak of ~1750 in 1991.

The rate of HIV-related deaths in the United States rose steadily through the 1980s and peaked in 1995. Since then, the HIV death rate has fallen fourfold (Fig. 197-16). This trend is likely due to several factors, including improved prophylaxis and treatment of opportunistic infections, growing experience among the health professions in caring for HIV-infected individuals, improved access to health care, and a decrease in new infections. However, the most influential factor clearly has been the increased use of potent antiretroviral drugs, generally administered in a combination of three or four agents.

FIGURE 197-16

Trends in annual age-adjusted rates of death due to HIV infection, United States, 1987–2014. Age distribution based on 2000 population. *(From CDC.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

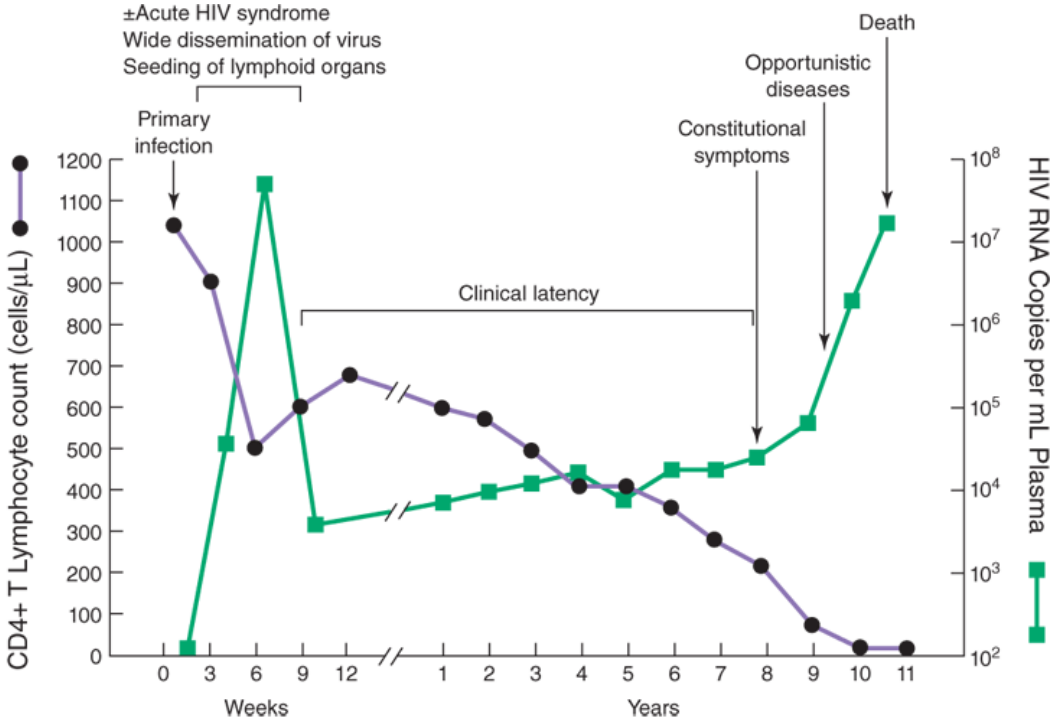
PATHOPHYSIOLOGY AND PATHOGENESIS

The hallmark of HIV disease is a profound immunodeficiency resulting primarily from a progressive quantitative and qualitative deficiency of the subset of T lymphocytes referred to as *helper T cells* occurring in a setting of polyclonal immune activation. The *helper* subset of T cells is defined phenotypically by the presence on its surface of the CD4 molecule **(Chap. 342)**, which serves as the primary cellular receptor for HIV. A co-receptor also must be present together with CD4 for efficient binding, fusion, and entry of HIV-1 into its target cells (Figs. 197-3 and 197-4). HIV-1 uses two major co-receptors, CCR5 and CXCR4, for fusion and entry; these co-receptors are also the primary receptors for certain chemoattractive cytokines termed *chemokines* and belong to the seven-transmembrane-domain G protein–coupled family of receptors. A number of mechanisms responsible for cellular depletion and/or immune dysfunction of CD4+ T cells have been demonstrated in vitro; these include direct infection and destruction of these cells by HIV, as well as indirect effects such as immune clearance of infected cells, cell death associated with aberrant immune activation, and immune exhaustion due to aberrant cellular activation with resulting cellular dysfunction. Patients with CD4+ T cell levels below certain thresholds are at high risk of developing a variety of opportunistic diseases, particularly the infections and neoplasms that are AIDS-defining illnesses. Some features of AIDS, such as Kaposi's sarcoma and certain neurologic abnormalities, cannot be explained completely by the immunodeficiency caused by HIV infection, since these complications may occur prior to the development of severe immunologic impairment.

The combination of viral pathogenic and immunopathogenic events that occurs during the course of HIV disease from the moment of initial (primary) infection through the development of advanced-stage disease is complex and varied. It is important to appreciate that the pathogenic mechanisms of HIV disease are multifactorial and multiphasic and are different at different stages of the disease. Therefore, it is essential to consider the typical clinical course of an untreated HIV-infected individual in order to more fully appreciate these pathogenic events **(Fig. 197-17)**.

FIGURE 197-17

Typical course of an untreated HIV-infected individual. See text for detailed description. *(From G Pantaleo et al: N Engl J Med 328:327, 1993. Copyright 1993 Massachusetts Medical Society. All rights reserved.)*



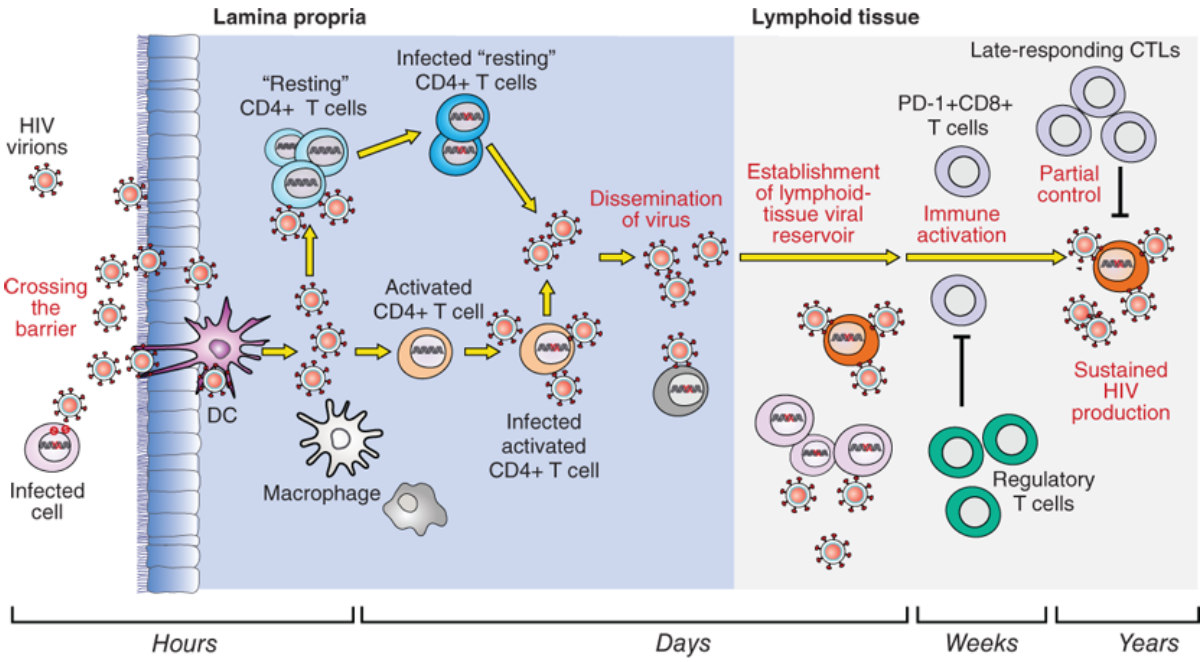
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EARLY EVENTS IN HIV INFECTION: PRIMARY INFECTION AND INITIAL DISSEMINATION OF VIRUS

Using rectal or vaginal mucosal transmission in nonhuman primates as a model, the earliest events (within hours) that occur following exposure of HIV to the mucosal surface determine whether an infection will be established or aborted as well as the subsequent course of events following infection. Although the mucosal barrier is relatively effective in limiting access of HIV to susceptible targets in the submucosal tissue, the virus can cross the barrier by transport on Langerhans cells, an epidermal type of DC, just beneath the surface or through microscopic rents in the mucosa. Significant disruptions in the mucosal barrier as seen in ulcerative genital disease facilitate viral entry and increase the efficiency of infection. Virus then seeks susceptible targets, which are primarily CD4+ T cells that are spatially dispersed in the mucosa. This spatial dispersion of targets provides a significant obstacle to the establishment of infection. Such obstacles account for the low efficiency of sexual transmission of HIV (see "Sexual Transmission," above). Both "partially" resting CD4+ T cells and activated CD4+ T cells serve as early amplifiers of infection. Resting CD4+ T cells are more abundant; however, activated CD4+ T cells produce larger amounts of virus. In order for infection to become established, the basic reproductive rate (R_0) must become equal to or greater than 1, i.e., each infected cell would infect at least one other cell. Once infection is established, the virus replicates in lymphoid cells in the mucosa, the submucosa, and to some extent the lymphoreticular tissues that drain the gut or genital tissues. For a variable period of time ranging from a few to several days, the virus cannot yet be detected in the plasma. This period is referred to as the "eclipse" phase of infection. As more virus is produced within several days to weeks, it is disseminated, first to the draining lymph nodes and then to other lymphoid compartments where it has easy access to dense concentrations of CD4+ T cell targets, allowing for a burst of high-level viremia that is readily detectable by currently available assays (Fig. 197-18). The gut-associated lymphoid tissue (GALT) is a major target of HIV infection and the location where large numbers of CD4+ T cells (usually memory cells) are infected and depleted, both by direct viral effects and by activation-associated apoptosis. Once virus replication reaches this threshold and virus is widely disseminated, infection is firmly established throughout the lymphoid tissues of the body and the process is irreversible. It is important to point out that the initial infection of susceptible cells may vary somewhat with the route of infection. Virus that enters directly into the bloodstream via infected blood or blood products (i.e., transfusions, use of contaminated needles for injection drugs, sharp-object injuries, maternal-to-fetal transmission either intrapartum or perinatally, or sexual intercourse where there is enough trauma to cause bleeding) is likely first cleared from the circulation to the spleen and other lymphoid organs, where primary focal infections begin, followed by wider dissemination throughout other lymphoid tissues as described above.

FIGURE 197-18

Summary of early events in HIV infection. See text for detailed description. CTLs, cytolytic T lymphocytes; HIV, human immunodeficiency virus. *(Adapted from AT Haase: Nat Rev Immunol 5:783, 2005.)*

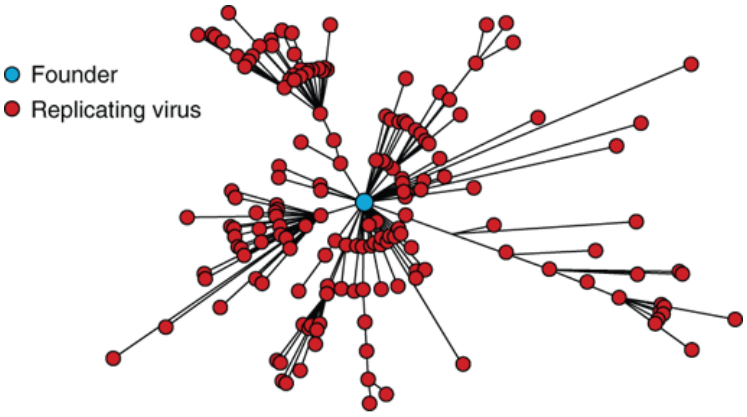


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

It has been demonstrated that sexual transmission of HIV is the result of a single infectious event and that a viral genetic bottleneck exists for transmission with selective transmission of certain viruses. In this regard, certain characteristics of the HIV envelope glycoprotein have a major influence on transmission, at least in subtype A and C viruses. Transmitting viruses, often referred to as "founder viruses," are usually underrepresented in the circulating viremia of the transmitting partner and are less-diverged viruses with signature sequences including shorter V1–V2 loop sequences and fewer predicted N-linked glycosylation sites relative to the major circulating variants. These viruses are almost exclusively R5 strains and are usually sensitive to neutralizing antibody. Once replication proceeds in the newly infected partner, the founder virus diverges and accumulates glycosylation sites, becoming progressively more resistant to neutralization **(Fig. 197-19)**.

FIGURE 197-19

As HIV diverges from founder to chronically replicating virus, it accumulates N-linked glycosylation sites. See text for detailed description. *(Adapted from CA Derdeyn et al: Science 303:2019, 2004; B Chohan et al: J Virol 79:6528, 2005; and BF Keele et al: Proc Natl Acad Sci USA 105:7552, 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The acute burst of viremia and wide dissemination of virus in primary HIV infection may be associated with an *acute HIV syndrome*, which occurs to varying degrees in ~50% of individuals within 2 to 4 weeks of initial infection (see below). This syndrome is usually associated with high levels of viremia measured in millions of copies of HIV RNA per milliliter of plasma that last for several weeks. Acute mononucleosis-like symptoms are well correlated with the presence of viremia. Virtually all patients develop some degree of viremia during primary infection, which contributes to virus dissemination throughout the lymphoid tissue, even though they may remain asymptomatic or not recall experiencing symptoms. The initial level of plasma viremia in primary HIV infection does not necessarily determine the rate of disease progression; however, the set point of the level of steady-state plasma viremia after ~1 year seems to correlate with the slope of disease progression in the untreated patient. The strikingly high levels of viremia observed in many patients with acute HIV infection is felt to be associated with a higher likelihood of transmission of the virus to others by a variety of routes including sexual transmission, shared needles and syringes, and mother-to-child transmission intrapartum, perinatally, or via breast milk.

ESTABLISHMENT OF CHRONIC AND PERSISTENT INFECTION

Persistence of Virus Replication

HIV infection is unique among human viral infections. Despite the robust cellular and humoral immune responses that are mounted following primary infection (see "Immune Response to HIV," below), once infection has been established the virus succeeds in escaping complete immune-mediated clearance, paradoxically seems to thrive on immune activation, and is never eliminated completely from the body. Rather, a chronic infection develops and persists with varying degrees of continual virus replication in the untreated patient for a median of ~10 years before the patient becomes clinically ill (see "Advanced HIV Disease," below). It is this establishment of a chronic, persistent infection that is the hallmark of HIV disease. Throughout the often-protracted course of chronic infection, virus replication can invariably be detected in untreated patients by widely available assays that measure copies of virion-associated HIV RNA in plasma (copies per milliliter). Levels of virus vary greatly in most untreated patients, usually ranging from several thousand to a few million copies of HIV RNA per milliliter of plasma. Studies using highly sensitive molecular techniques have demonstrated that even in certain patients in whom plasma viremia is suppressed to below detection (lower limit, 20–50 copies of HIV RNA per milliliter depending on assay kit manufacturer) by cART, there is a continual low level of virion production in the majority of infected patients. In other human viral infections, with very few exceptions, if the host survives, the virus is completely cleared from the body and a state of immunity against subsequent infection develops. HIV infection very rarely kills the host during primary infection. Certain viruses, such as HSV **(Chap. 187)**, are not completely cleared from the body after infection, but instead enter a latent state; in these cases, clinical latency is accompanied by microbiologic latency. This is not the case with HIV infection as described above. Chronicity associated with persistent virus replication can also be seen in certain cases of HBV and HCV infections **(Chap. 334)**; however, in these infections the immune system is not a target of the virus.

Escape of HIV from Effective Immune System Control

Inherent to the establishment of chronicity of HIV infection is the ability of the virus to evade adequate control and elimination by both the cellular and humoral immune responses. There are a number of mechanisms whereby the virus accomplishes this evasion. Paramount among these is the establishment of a sustained level of replication associated with the generation of viral diversity via mutation and recombination. The selection of mutants that escape control by CD8+ cytolytic T lymphocytes (CTLs) is critical to the propagation and progression of HIV infection. The high rate of virus replication associated with inevitable mutations also contributes to the inability of antibody to clear the autologous virus. Furthermore, for reasons that remain unclear, the humoral immune system does not readily produce classic neutralizing antibodies against the HIV envelope and does so only after years of persistent virus replication and after the infection is firmly established (see below). Extensive analyses of sequential HIV isolates and host responses have demonstrated that viral escape from B cell and CD8+ T cell responses occurs early after infection and allows the virus to stay one step ahead of effective immune responses. Virus-specific CD8+ CTLs expand greatly during primary HIV infection, and they likely represent the high-affinity responses that would be expected to be most efficient in eliminating virus-infected cells; however, viral control is generally incomplete as viral replication persists at relatively high levels in the majority of individuals. In addition to viral escape from CTLs through high rates of mutation, it is thought that the initially strong response becomes qualitatively dysfunctional owing to the overwhelming immune activation associated with persistent viral replication, leading to immune "exhaustion" that affects both arms of adaptive immunity. Several studies have indicated that exhaustion of HIV-specific CD8+ T cells during prolonged immune activation is associated with upregulation of several inhibitory receptors, such as programmed death (PD) 1 molecule (of the B7-CD28 family of molecules), T cell immunoreceptor with Ig and ITIM domains (TIGIT), T cell immunoglobulin and mucin-domain containing molecule 3 (Tim-3), and lymphocyte activating gene 3 (Lag-3), collectively referred to as *immune-checkpoint receptors*. Upregulation of these surface proteins restricts polyreactivity and proliferative capacity, functional attributes of CD8+ T cells that are essential for effective killing of pathogens. Another mechanism contributing to the evasion by HIV of immune system control is the downregulation of HLA class I molecules on the surface of HIV-infected cells by the viral proteins Nef, Tat, and Vpu, resulting in the lack of ability of CD8+ CTLs to recognize and kill infected target cells. Although this downregulation of HLA class I molecules would seem to favor elimination of HIV-infected cells by natural killer (NK) cells, this latter mechanism does not remove HIV-infected cells effectively (see below). Another potential means of escape of HIV-infected cells from elimination by CD8+ CTLs is the sequestration of infected cells in immunologically privileged sites such as the central nervous system (CNS), as well as the low frequency of virus-specific CD8+ CTLs in areas of lymphoid tissues, namely germinal centers, where HIV actively replicates.

The principal targets of neutralizing antibodies against HIV are the envelope proteins gp120 and gp41. HIV employs at least three mechanisms to evade neutralizing antibody responses: hypervariability in the primary sequence of the envelope, extensive glycosylation of the envelope, and conformational masking of neutralizing epitopes. Several studies that have followed the evolution of the humoral immune response to HIV from the earliest points after primary infection indicate that the virus continually mutates to escape the emerging antibody response such that the sequential antibodies that are induced do not neutralize the currently autologous virus. *Broadly neutralizing antibodies* capable of neutralizing a wide range of primary HIV isolates in vitro occur in only about 20% of HIV-infected individuals, and, when they do occur, 2 to 3 years of infection with continual virus replication are generally required to drive the affinity maturation of the antibodies. Unfortunately, by the time these broadly neutralizing antibodies are formed, they are ineffective in containing the virus currently replicating in the patient. Persistent viremia also results in exhaustion of B cells similar to the exhaustion reported for CD8+ T cells, adding to the defects in the humoral response to HIV.

CD4+ T cell help is essential for the integrity of antigen-specific immune responses, both humoral and cell-mediated. HIV preferentially infects activated CD4+ T cells including HIV-specific CD4+ T cells, and so this loss of viral-specific helper T cell responses has profoundly negative consequences for the

immunologic control of HIV replication. Furthermore, this loss occurs early in the course of infection, and animal studies indicate that 40–70% of all memory CD4+ T cells in the GALT are eliminated during acute infection. During chronic HIV viremia, CD4+ T cells also exhibit evidence of exhaustion, including by upregulation of the cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4), also a member of the B7-CD28 family.

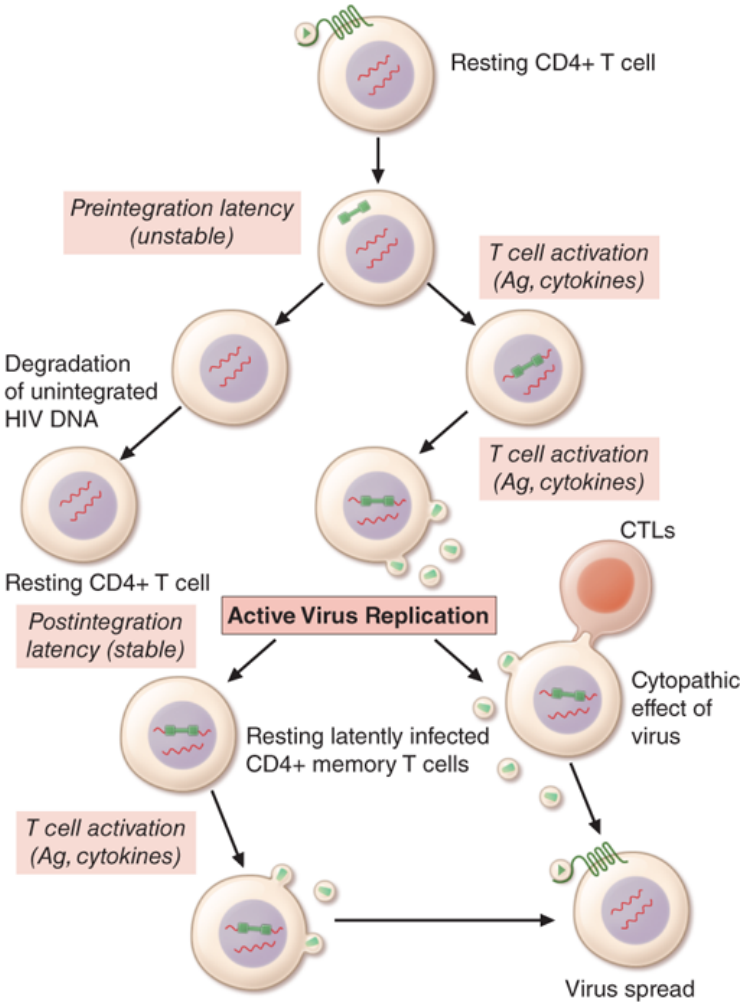
Finally, the escape of HIV from immune-mediated elimination during primary infection allows the formation of a pool of latently infected cells, referred to as the *viral reservoir*, that may not be recognized or completely eliminated by virus-specific CTLs or by ART (see below). Thus, despite a potent immune response and the marked downregulation of virus replication following primary HIV infection, HIV succeeds in establishing a state of chronic infection with a variable degree of persistent virus replication. During this period most patients make the clinical transition from acute primary infection to variable periods of clinical latency or smoldering disease activity (see below).

The HIV Reservoir: Obstacles to the Eradication of Virus

A pool of latently infected, resting CD4+ T cells that serves as at least one component of the persistent reservoir of virus exists in virtually all HIV-infected individuals, including those who are receiving cART. Such cells carry an integrated form of HIV DNA in the genome of the host and can remain in this state until an activation signal drives the expression of HIV transcripts. Only a small fraction of the latently infected cells in the viral reservoir contain replication-competent virus with the overwhelming majority of cells containing defective proviruses incapable of a full replication cycle. However, upon activation of the reservoir variable degrees of sustained virus replication invariably occur. This form of latency is to be distinguished from preintegration latency, in which HIV enters a resting CD4+ T cell and, in the absence of an activation signal, reverse transcription of the HIV genome occurs to a certain extent but the resulting proviral DNA fails to integrate into the host genome. This period of preintegration latency may last hours to days, and if no activation signal is delivered to the cell, the proviral DNA loses its capacity to initiate a productive infection. If these cells do become activated prior to decay of the preintegration complex, reverse transcription proceeds to completion and the virus continues along its replication cycle (see above and **Fig. 197-20**). The pool of cells that are in the postintegration state of latency is established early during the course of primary HIV infection. Despite the suppression of plasma viremia to undetectable levels by potent regimens of cART administered over several years, this pool of latently infected cells persists and can give rise to replication-competent virus upon cellular activation ex vivo. Modeling studies built on projections of decay curves have estimated that in such a setting of prolonged viral suppression, it would require a few to several years for the pool of latently infected cells to be completely eliminated. This has not been documented to occur spontaneously in any patients very likely because the latent viral reservoir is long-lived and is continually replenished by the low levels of persistent virus replication that may remain below the limits of detection of current assays (see below) as well as by the expansion by proliferation of the pool of latently infected cells (Fig. 197-20), even in patients who for the most part are treated successfully. Reservoirs of HIV-infected cells, latent or otherwise, can exist in a number of compartments including the lymphoid tissue, peripheral blood, and the CNS (likely in cells of the monocyte/macrophage lineage) as well as in other unidentified locations. Over the past several years attempts have been made to eliminate HIV in the latent viral reservoir using agents that activate resting CD4+ T cells during the course of cART; however, such attempts, referred to as “shock and kill,” have been unsuccessful. Thus, this persistent reservoir of infected cells and/or low levels of persistent virus replication remain major obstacles to the goal of eradication of virus from infected individuals and hence a “cure,” despite the favorable clinical outcomes that have resulted from cART.

FIGURE 197-20

Generation of latently infected, resting CD4+ T cells in HIV-infected individuals. See text for details. Ag, antigen; CTLs, cytolytic T lymphocytes. *(Courtesy of TW Chun; with permission.)*

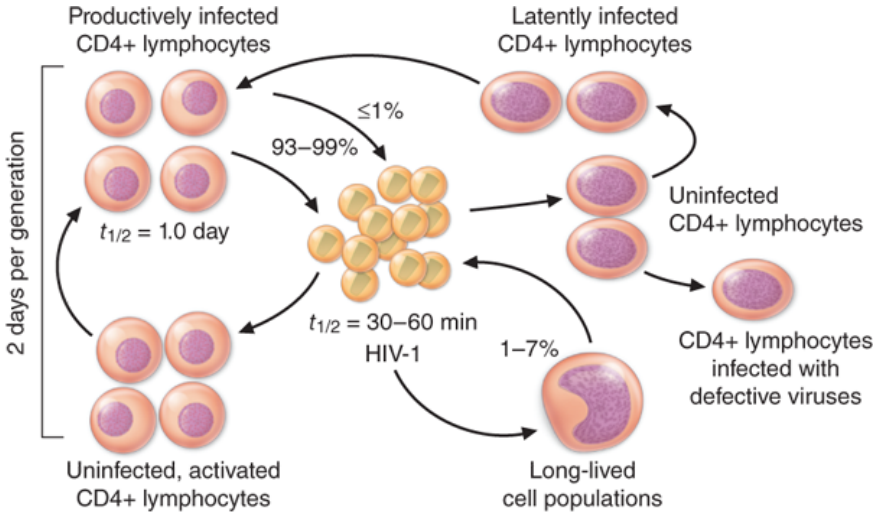


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Viral Dynamics

The dynamics of viral production and turnover have been quantified using mathematical modeling in the setting of the administration of reverse transcriptase and protease inhibitors to HIV-infected individuals in clinical studies. Treatment with these drugs resulted in a precipitous decline in the level of plasma viremia, which typically fell by well over 90% within 2 weeks. It was determined on the basis of modeling the kinetics of viral decline and the emergence of resistant mutants during therapy that 93–99% of the circulating virus originated from recently infected, rapidly turning over CD4+ T cells and that ~1–7% of circulating virus originated from longer-lived cells, likely monocytes/macrophages. A negligible amount of circulating virus originated from the pool of latently infected cells **(Fig. 197-21)**. It was also determined that the half-life of a circulating virion was ~30–60 min and that of productively infected cells was 1 day. Given the relatively steady level of plasma viremia and of infected cells, it appears that extremely large amounts of virus (~10^{10}–10^{11} virions) are produced and cleared from the circulation each day. In addition, data suggest that the minimal duration of the HIV-1 replication cycle in vivo is ~2 days. Other studies have demonstrated that the decrease in plasma viremia that results from cART correlates closely with a decrease in virus replication in lymph nodes, further confirming that lymphoid tissue is the main site of HIV replication and the main source of plasma viremia.

FIGURE 197-21
Dynamics of HIV infection in vivo. See text for detailed description. *(From AS Perelson et al: Science 271:1582, 1996.)*

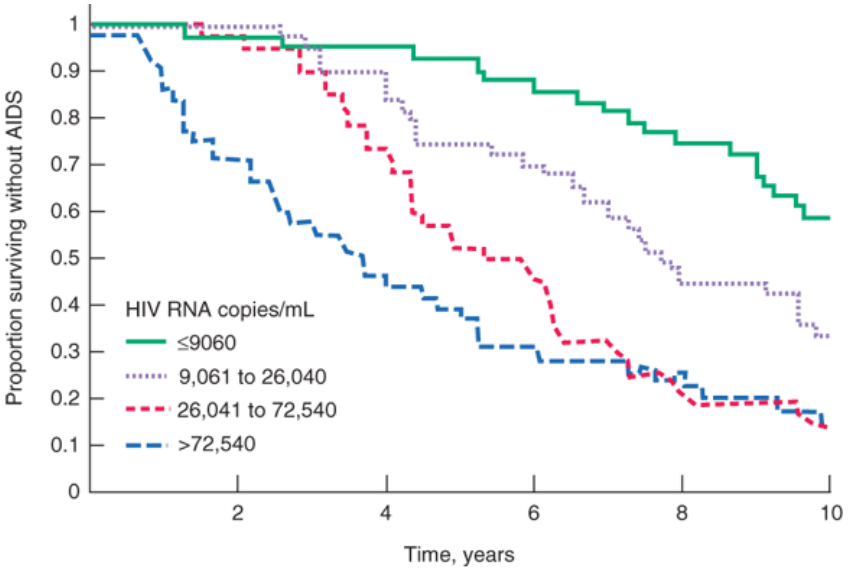


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The level of steady-state viremia, called the viral *set point*, at ~1 year following acquisition of HIV infection has important prognostic implications for the progression of HIV disease in the untreated patient. It has been demonstrated that, as a group, untreated HIV-infected individuals who have a low set point at 6 months to 1 year following infection progress to AIDS much more slowly than do individuals whose set point is very high at that time **(Fig. 197-22)**.

FIGURE 197-22

Relationship between levels of virus and rates of disease progression. Kaplan-Meier curves for AIDS-free survival stratified by baseline HIV-1 RNA categories (copies per milliliter). *(From JW Mellors et al: Science 272:1167, 1996.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical Latency versus Microbiologic Latency

With the exception of certain long-term nonprogressors and "elite controllers" of HIV replication, the level of CD4+ T cells in the blood inevitably decreases progressively in viremic HIV-infected individuals in the absence of cART. The decline in CD4+ T cells may be gradual or abrupt, the latter usually reflecting a significant spike in the level of plasma viremia. Most patients are relatively asymptomatic while this progressive decline is taking place (see below) and are often described as being in a state of *clinical latency*. However, this term is misleading; it does not mean disease latency, since progression, although slow in many cases and often without symptoms, is generally relentless during this period. Furthermore, clinical latency should not be confused with microbiologic latency, since varying levels of virus replication inevitably occur during this period of clinical latency. Even in those rare patients who have <50 copies of HIV RNA per milliliter in the absence of therapy, there is virtually always some degree of low-level ongoing virus replication.

ADVANCED HIV DISEASE

In untreated patients or in patients in whom therapy has not adequately controlled virus replication, after a variable period, usually measured in years, the CD4+ T cell count falls below a critical level (<200/μL) and the patient becomes highly susceptible to opportunistic disease (Fig. 197-17). For this reason, the CDC case definition of AIDS includes all HIV-infected individuals >5 years of age with CD4+ T cell counts below this level (Table 197-2). Patients may experience constitutional signs and symptoms or may develop an opportunistic disease abruptly without any prior symptoms. The depletion of CD4+ T cells continues to be progressive and unrelenting in this phase. It is not uncommon for CD4+ T cell counts in the untreated patient to drop to as low as 10/μL or even to zero. In countries where cART and prophylaxis and treatment for opportunistic infections are readily accessible to such patients, survival is increased dramatically even in those patients with advanced HIV disease. In contrast, untreated patients who progress to this severest form of immunodeficiency usually succumb to opportunistic infections or neoplasms (see below).

LONG-TERM SURVIVORS, LONG-TERM NONPROGRESSORS, AND ELITE CONTROLLERS

It is important to distinguish between the terms *long-term survivor* and *long-term nonprogressor*. Long-term nonprogressors are by definition long-term survivors; however, the reverse is not always true. Predictions from one study that antedated the availability of effective cART estimated that ~13% of homosexual/bisexual men who were infected at an early age may remain free of clinical AIDS for >20 years. Many of these individuals may have progressed in their degree of immune deficiency; however, they certainly survived for a considerable period of time. With the advent of effective cART, the survival of HIV-infected individuals has dramatically increased. Early in the AIDS epidemic, prior to the availability of therapy, if a patient presented with a life-threatening opportunistic infection, the median survival was 26 weeks from the time of presentation. Currently, an HIV-infected 20-year-old individual in a high-income country who is appropriately treated with cART can expect to live at least 50 years according to mathematical model projections. In the face of cART, long-term survival is becoming commonplace. Definitions of long-term nonprogressors have varied considerably over the years, and so such individuals constitute a heterogeneous group. Long-term nonprogressors were first described in the 1990s. Originally, individuals were considered to be long-term nonprogressors if they had been infected with HIV for a long period (≥10 years), their CD4+ T cell counts were in the normal range, and they remained stable over years without receiving cART. Approximately 5–15% of HIV-infected individuals fell into this broader nonprogressor category. However, this group was rather heterogenous and over time a significant proportion of these individuals progressed and ultimately required therapy. From this broader group, a much smaller subgroup of “elite” controllers or nonprogressors was identified, and they constituted a fraction of 1% of HIV-infected individuals. These elite controllers, by definition, have extremely low levels of plasma viremia that is often undetectable by standard assays and normal CD4+ T cell counts. It is noteworthy that certain of their HIV-specific immune responses are robust and clearly superior to those of HIV-infected progressors. In this group of elite controllers certain HLA class I haplotypes are overrepresented, particularly HLA-B57-01 and HLA-B27-05. Outside of the subgroup of elite controllers, a number of other genetic factors have been shown to be involved to a greater or lesser degree in the control of virus replication and thus in the rate of HIV disease progression (see “Genetic Factors in HIV-1 and AIDS Pathogenesis,” below).

LYMPHOID ORGANS AND HIV PATHOGENESIS

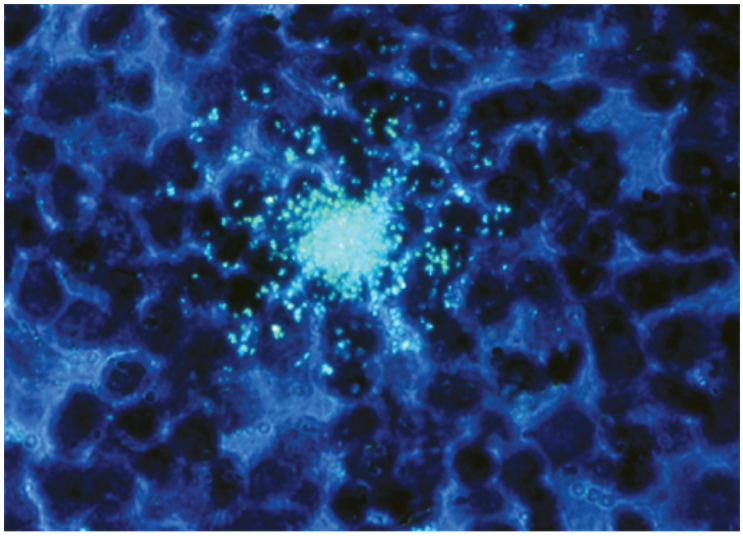
Regardless of the portal of entry of HIV, lymphoid tissues are the major anatomic sites for the establishment and propagation of HIV infection. Despite the use of measurements of plasma viremia to determine the level of disease activity, virus replication occurs mainly in lymphoid tissue and not in blood; indeed, the level of plasma viremia directly reflects virus production in lymphoid tissue.

Some patients experience progressive generalized lymphadenopathy early in the course of the infection; others experience varying degrees of transient lymphadenopathy. Lymphadenopathy reflects the cellular activation and immune response to the virus in the lymphoid tissue, which is generally characterized by follicular or germinal center hyperplasia. Lymphoid tissue involvement is a common denominator of virtually all patients with HIV infection, even those without easily detectable lymphadenopathy.

Examinations of lymph tissue and peripheral blood in patients and monkeys during various stages of HIV and SIV infection, respectively, have led to substantial insight into the pathogenesis of HIV disease. In most of the original human studies, peripheral lymph nodes have been used predominantly as the source of lymphoid tissue. More recent studies in monkeys and humans have also focused on the GALT, where the earliest burst of virus replication occurs associated with marked depletion of CD4+ T cells. In detailed studies of peripheral lymph node tissue that utilized a variety of molecular techniques to measure the level of HIV DNA and RNA and imaging techniques to visualize virus and cells, the following picture has emerged. During acute HIV infection resulting from mucosal transmission, virus replication progressively amplifies from scattered lymphoid cells in the lamina propria of the gut to draining lymphoid tissue, leading to high levels of plasma viremia. The GALT plays a major role in the amplification of virus replication, and virus is disseminated from replication in the GALT to peripheral lymphoid tissue. A profound degree of cellular activation occurs within lymphoid tissue (see below) and is reflected in follicular or germinal center hyperplasia. At this time copious amounts of extracellular virions (both infectious and defective) are trapped on the processes of the follicular dendritic cells (FDCs) in the germinal centers of the lymph nodes. Virions that have bound complement components on their surfaces attach to the surface of FDCs via interactions with complement receptors and likely via Fc receptors that bind to antibodies that are attached to the virions. In situ hybridization reveals expression of virus in individual cells of the paracortical area and, to a lesser extent, the germinal center **(Fig. 197-23)**. The persistence of trapped virus likely reflects a steady state whereby trapped virus turns over and is replaced by fresh virions that are continually produced. The trapped virus, either as whole virion or shed envelope, serves as a continual activator of CD4+ T cells, thus driving further virus replication.

FIGURE 197-23

HIV in the lymph node of an HIV-infected individual. An individual cell infected with HIV shown expressing HIV RNA by in situ hybridization using a radiolabeled molecular probe. Original ×500. *(Adapted from G Pantaleo, AS Fauci et al: Nature 362:355, 1993.)*



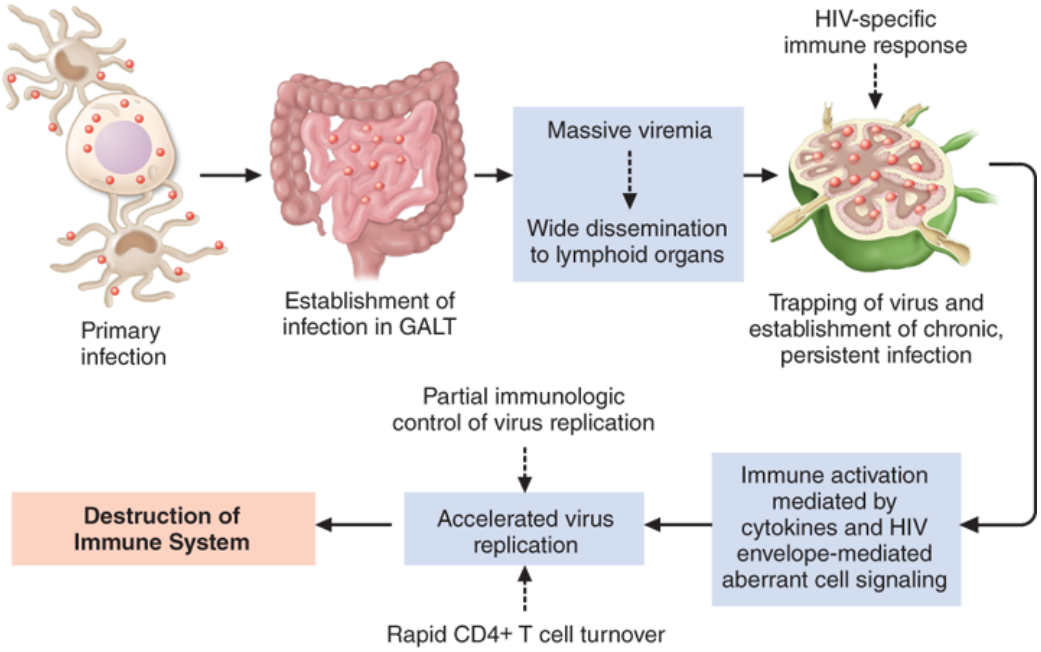
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

During the early stages of HIV disease, the architecture of lymphoid tissues is generally preserved and may even be hyperplastic owing to an increased presence of B cells and specialized CD4+ T cells called follicular helper CD4+ T cells (TF_H) in prominent germinal centers. Extracellular virions can be seen by electron microscopy attached to FDC processes. The trapping of antigen is a physiologically normal function for the FDCs, which present antigen to B cells and, along with stimulatory factors produced by TF_H cells, contribute to the generation of B cell memory. However, in the case of HIV, persistent cellular activation, resulting in a shift to secretion of proinflammatory cytokines such as interleukin (IL) 1β, tumor necrosis factor (TNF) α, IFN-γ, and IL-6, can induce viral replication (see below) and diminish the effectiveness of the immune response against the virus. In addition, the CD4+ TF_H cells that are recruited into the germinal center to provide help to B cells in the generation of an HIV-specific immune response are highly susceptible to infection and may be an important component of the HIV reservoir. Thus, in HIV infection, a normal physiologic function of the immune system that contributes to the clearance of virus, as well as to the generation of a specific immune response, can also have deleterious consequences.

As HIV disease progresses, the architecture of lymphoid tissues begins to show disruption. Confocal microscopy reveals destruction of the fibroblastic reticular cell (FRC) and FDC networks in the T cell zone and B cell follicles, respectively. The mechanisms of destruction are not completely understood, but they are thought to be associated with collagen deposition causing fibrosis and loss of production of cytokines such as IL-7 and lymphotoxin-α, which are critical to the maintenance of lymphoid tissues and their lymphocyte constituents. As the disease progresses to an advanced stage, there is complete disruption of the architecture of the lymphoid tissues, accompanied by dissolution of the FRC and FDC networks. At this point, the lymph nodes are “burnt out.” This destruction of lymphoid tissue compounds the immunodeficiency of HIV disease and contributes both to the inability to control HIV replication and to the inability to mount adequate immune responses against opportunistic pathogens. The events from primary infection to the ultimate destruction of the immune system are illustrated in **Fig. 197-24**. More recently, nonhuman primate studies and some human studies have examined GALT at various stages of HIV disease. Within the GALT, the basal level of cellular activation combined with virus-mediated activation results in the infection and elimination of an estimated 50–90% of CD4+ T cells in the gut. The extent of this early damage to GALT, which constitutes a major component of lymphoid tissue in the body, may play a role in determining the potential for immunologic recovery of the memory cell subset.

FIGURE 197-24

Events that transpire from primary HIV infection through the establishment of chronic persistent infection to the ultimate destruction of the immune system. See text for details. CTLs, cytolytic T lymphocytes; GALT, gut-associated lymphoid tissue.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE ROLE OF IMMUNE ACTIVATION AND INFLAMMATION IN HIV PATHOGENESIS

Activation of the immune system and variable degrees of inflammation are essential components of any appropriate immune response to a foreign antigen. However, immune activation and inflammation, which can be considered aberrant in HIV-infected individuals, play a critical role in the pathogenesis of HIV disease as well as other chronic conditions associated with HIV infection. Immune activation and inflammation in the HIV-infected individual contribute substantially to (1) the replication of HIV, (2) the induction of immune dysfunction, and (3) the increased incidence of chronic conditions such as premature cardiovascular disease (Table 197-4).

TABLE 197-4

Conditions Associated with Persistent Immune Activation and Inflammation in Patients with HIV Infection

Accelerated aging syndrome
Bone fragility
Cancers
Cardiovascular disease
Diabetes
Kidney disease
Liver disease
Neurocognitive dysfunction

INDUCTION OF HIV REPLICATION BY ABERRANT IMMUNE ACTIVATION

The immune system is normally in a state of homeostasis, awaiting perturbation by foreign antigenic stimuli. Once the immune response deals with and clears the antigen, the system returns to relative quiescence (Chap. 342). This is generally not the case in HIV infection where, in the untreated patient, virus replication is invariably persistent with very few exceptions and immune activation is persistent. HIV replicates most efficiently in activated CD4+ T cells; in HIV infection, chronic activation provides the cell substrates necessary for persistent virus replication throughout the course of HIV disease, particularly in the untreated patient. Even in certain patients receiving cART whose levels of plasma viremia are suppressed to below the level of

detection by standard assays, there are low but detectable degrees of virus replication that drives persistent immune activation. From a virologic standpoint, although quiescent CD4+ T cells can be infected, albeit inefficiently, with HIV, reverse transcription, integration, and virus spread are much more efficient in activated cells. Furthermore, cellular activation induces expression of virus in cells latently infected with HIV. In essence, immune activation and inflammation provide the engine that drives HIV replication. In addition to endogenous factors such as cytokines, a number of exogenous factors such as other microbes that are associated with heightened cellular activation can enhance HIV replication and thus may play a role in HIV pathogenesis. Co-infection with a range of viruses, such as HSV types 1 and 2, cytomegalovirus (CMV), human herpesvirus (HHV) 6, Epstein-Barr virus (EBV), HBV, HCV, adenovirus, and HTLV-1 have been shown to upregulate HIV expression. In addition, infestation with nematodes has been shown to be associated with a heightened state of immune activation that facilitates HIV replication; in certain studies deworming of the infected host has resulted in a decrease in plasma viremia. Two diseases of extraordinary global health significance, malaria and tuberculosis (TB), have been shown to increase HIV viral load in dually infected individuals. Globally, *Mycobacterium tuberculosis* is the most common opportunistic infection in HIV-infected individuals **(Chap. 173)**. In addition to the fact that HIV-infected individuals are more likely to develop active TB after exposure and to reactivate latent TB, it has been demonstrated that active TB can accelerate the course of HIV infection. It has also been shown that levels of plasma viremia are greatly elevated in HIV-infected individuals with active TB who are not receiving cART, compared with pre-TB levels and levels of viremia after successful treatment of the active TB. The situation is similar in the interaction between HIV and malaria parasites **(Chap. 219)**. Acute infection of HIV-infected individuals with *Plasmodium falciparum* increases HIV viral load, and the increased viral load is reversed by effective treatment of malaria.

MICROBIAL TRANSLOCATION AND PERSISTENT IMMUNE ACTIVATION

One proposed mechanism of persistent immune activation involves the disruption of the mucosal barrier in the gut due to HIV replication in submucosal lymphoid tissue. As a result of this disruption, there is an increase in the products, particularly lipopolysaccharide (LPS), of bacteria that translocate from the bowel lumen through the damaged mucosa to the circulation, leading to persistent systemic immune activation and inflammation. This effect can persist even after the HIV viral load is brought to <50 copies/mL by cART. Depletion in the GALT of IL-17–producing T cells, which are responsible for defense against extracellular bacteria and fungi, also is thought to contribute to HIV pathogenesis.

PERSISTENT IMMUNE ACTIVATION AND INFLAMMATION INDUCE IMMUNE DYSFUNCTION

The immune activated state in HIV infection is reflected by hyperactivation of B cells leading to hypergammaglobulinemia; increased lymphocyte turnover; activation of monocytes; expression of activation markers and immune checkpoint receptors on CD4+ and CD8+ T cells; increased activation-associated cellular apoptosis; lymph node hyperplasia, particularly early in the course of disease; increased secretion of proinflammatory cytokines, particularly IL-6; elevated levels of high-sensitivity C-reactive protein, fibrinogen, D-dimer, neopterin, β_2-microglobulin, acid-labile interferon, soluble (s) IL-2 receptors (R), sTNFR, sCD27, and sCD40L; and autoimmune phenomena (see "Autoimmune Phenomena," below). Even in the absence of direct infection of a target cell, HIV envelope proteins can interact with cellular receptors (CD4 molecules and chemokine receptors) to deliver potent activation signals resulting in calcium flux, the phosphorylation of certain proteins involved in signal transduction, co-localization of cytoplasmic proteins including those involved in cell trafficking, immune dysfunction, and, under certain circumstances, apoptosis. From an immunologic standpoint, chronic exposure of the immune system to a particular antigen over an extended period of time may ultimately lead to an inability to sustain an adequate immune response to the antigen in question. In many chronic viral infections, including HIV infection, persistent viremia is associated with "functional exhaustion" of virus-specific T cells, decreasing their capacity to proliferate and perform effector functions. It has been demonstrated that this phenomenon of immune exhaustion may be mediated, at least in part, by the upregulation of inhibitory receptors on HIV-specific T cells, such as PD-1 and Tim-3 that are shared by both CD4+ and CD8+ T cells, as well as CTLA-4 on CD4+ and 2B4 and CD106 on CD8+ T cells. Furthermore, the ability of the immune system to respond to a broad spectrum of non-HIV antigens may be compromised if immunocompetent bystander cells are maintained in a state of chronic activation.

The deleterious effects of chronic immune activation on the progression of HIV disease are well established. As in most conditions of persistent antigen exposure, the host must maintain sufficient activation of antigen (HIV)-specific responses but must also prevent excessive activation and potential immune-mediated damage to tissues. Certain studies suggest that normal immunoregulatory mechanisms that act to keep hyperimmune activation in check, particularly CD4+, FoxP3+, and CD25+ regulatory T cells (T-regs), may be dysfunctional or depleted in the context of advanced HIV disease.

Apoptosis

Apoptosis is a form of programmed cell death that is a normal mechanism for the elimination of effete cells in organogenesis as well as in the cellular proliferation that occurs during a normal immune response **(Chap. 342)**. Apoptosis can occur by intrinsic or extrinsic pathways, the latter of which is largely dependent on cellular activation, and the aberrant cellular activation associated with HIV disease is correlated with a heightened state of apoptosis. HIV can trigger activation-induced cell death through the upregulation of the death receptors, such as Fas/CD95, TNFR1, or TNF-related apoptosis-inducing ligand (TRAIL) receptors 1 and 2. Their corresponding ligands FasL, TNF, and TRAIL also are upregulated in HIV disease. HIV-induced stress and alterations in homeostasis also can trigger intrinsic apoptosis due to the downregulation of antiapoptotic proteins such as Bcl-2. Other mechanisms of HIV-induced cell death have been described, including autophagy, necrosis, necroptosis, and pyroptosis. The phenomenon of pyroptosis, an inflammatory form of cell death involving the upregulation of the proinflammatory enzyme caspase 1 and release of the proinflammatory cytokines IL-1β and IL-18, has been linked to a bystander effect of HIV replication on CD4+ T cells. The process of pyroptosis generates multimeric complexes called inflammasomes, which can also be activated by LPS. Certain viral gene products have been associated with enhanced susceptibility to apoptosis; these include Env, Tat, and Vpr. In contrast, Nef has been shown to possess antiapoptotic properties. The intensity of apoptosis correlates

with the general state of activation of the immune system and not with the stage of disease or with viral burden. A number of studies, including those examining lymphoid tissue, have demonstrated that the rate of apoptosis is elevated in HIV infection and that apoptosis is seen in "bystander" cells such as CD8+ T cells and B cells as well as in uninfected CD4+ T cells. It is likely that this bystander apoptosis of immunocompetent cells related to immune activation contributes to the general immune abnormalities in HIV disease.

MEDICAL CONDITIONS ASSOCIATED WITH PERSISTENT IMMUNE ACTIVATION AND INFLAMMATION IN HIV DISEASE

It has become clear, as the survival of HIV-infected individuals has increased, that a number of previously unrecognized medical complications are associated with HIV disease—and that these complications relate to chronic immune activation and inflammation (Table 197-4). These complications can appear even after patients have experienced years of adequate control of viral replication (plasma viremia below detectable levels) for several years. Of particular note are endothelial cell dysfunction and its relationship to cardiovascular disease. Other chronic conditions that have been reported include bone fragility, certain cancers, diabetes, kidney and liver disease, and neurocognitive dysfunction, thus presenting an overall picture of accelerated aging.

Autoimmune Phenomena

Autoimmune phenomena are commonly observed in HIV-infected individuals and they reflect, at least in part, chronic immune activation and the dysregulation of B and T cells. Although these phenomena usually occur in the absence of autoimmune disease, a wide spectrum of clinical manifestations that may be associated with autoimmunity have been described (see "Immunologic and Rheumatologic Diseases," below). Autoimmune phenomena include antibodies against autoantigens expressed on intact lymphocytes and other cells, or against proteins released from dying cells. Antiplatelet antibodies have some clinical relevance in that they may contribute to the thrombocytopenia of HIV disease (see below). Antibodies to nuclear and cytoplasmic components of cells have been reported, as have antibodies to cardiolipin and phospholipids; CD4 molecules; CD43 molecules; C1q-A; variable regions of the T cell receptor α, β, and γ chains; Fas; denatured collagen; and IL-2. In addition, autoantibodies to a range of serum proteins, including albumin, immunoglobulin, and thyroglobulin, have been reported. Molecular mimicry, either from opportunistic pathogens or from HIV itself, also is a trigger or cofactor in autoimmunity. Antibodies against the HIV envelope proteins, especially gp41, often cross-react with host proteins; the best known examples are antibodies directed against the membrane-proximal external region (MPER) of gp41 that also react with phospholipids and cardiolipin. The phenomenon of polyreactive HIV-specific antibodies may be beneficial to the host (see "Immune Response to HIV," below).

The increased occurrence and/or exacerbation of certain autoimmune diseases have been reported in HIV infection; these diseases include psoriasis, idiopathic thrombocytopenic purpura, Graves' disease, antiphospholipid syndrome, and primary biliary cirrhosis. The majority of these manifestations were described prior to the advent of cART and have decreased in frequency since its widespread use. However, with increasing availability of cART, an *immune reconstitution inflammatory syndrome* (IRIS) has been increasingly observed in infected individuals, particularly those with low CD4+ T cell counts (see below). IRIS is an autoimmune-like phenomenon characterized by a paradoxical deterioration of clinical condition, which is usually compartmentalized to a particular organ system, in individuals in whom cART has recently been initiated. It is associated with a decrease in viral load and at least partial recovery of immune competence, which is usually associated with increases in CD4+ T cell counts. The immunopathogenesis of this syndrome is felt to be related to an increase in immune response against the presence of residual antigens that are usually microbial and is most commonly seen with underlying *Mycobacterium tuberculosis* and cryptococcosis. This syndrome is discussed in more detail below.

CYTOKINES AND OTHER SOLUBLE FACTORS IN HIV PATHOGENESIS

The immune system is homeostatically regulated by a complex network of immunoregulatory cytokines, which are pleiotropic and redundant and operate in an autocrine and paracrine manner. They are expressed continuously, even during periods of apparent quiescence of the immune system. On perturbation of the immune system by antigenic challenge, the expression of cytokines increases to varying degrees **(Chap. 342)**. Cytokines that are important components of this immunoregulatory network are thought to play major roles in HIV disease, during both the early and chronic phases of infection. A potent proinflammatory "cytokine storm" is induced during the acute phase of HIV infection, likely a response by inflammatory cells to virus replicating at very high levels. Cytokines that are induced during this early phase include IFN-α, IL-15, and the CXC chemokine IP-10 (CXCL10), followed by IL-6, IL-12, and TNF-α, and a delayed peak of the anti-inflammatory cytokine IL-10. Soluble factors of innate immunity also are induced shortly after infection, including neopterin and β-microglobulin. Several of these early-expressed cytokines and factors are not downregulated following the early phase of HIV infection, as seen in other self-resolving viral infections, and persist during the chronic phase of infection and contribute to maintaining high levels of immune activation. Among the cytokines and factors associated with early innate immune responses, they are intended to contain viral replication, although paradoxically most are potent inducers of HIV expression/replication because of their ability to induce immune activation that leads to enhanced viral production and an increase in readily available target cells for HIV (activated CD4+ T cells). The induction of IFN-α, one of the first cytokines induced during primary HIV infection and an important element of innate immune sensing, is thought to play a particularly important role in HIV pathogenesis by inducing a large number of IFN-associated genes that activate the immune system, alter the homeostasis of CD4+ T cells, and influence the virus variants that are selected during the HIV transmission bottleneck. Other cytokines that are elevated during the chronic phase of HIV infection and linked to immune activation include IFN-γ, the *CC-chemokine* RANTES (CCL5), macrophage inflammatory protein (MIP)-1β (CCL4), and IL-18.

Several specific cytokines and soluble factors have been associated with HIV pathogenesis at various stages of disease, in various tissues or organs, and in the regulation of HIV replication. Plasma levels of IP-10 are predictive of disease progression, whereas the proinflammatory cytokine IL-6, soluble CD14 (sCD14), and coagulation marker D-dimer are associated with increased risk of all-cause mortality in HIV-infected individuals. In particular, IL-6, sCD14, and D-dimer are associated with increased risk of cardiovascular disease and other causes of death, even in individuals receiving cART. IL-18 has also been shown to play a role in the development of the HIV-associated lipodystrophy syndrome, whereas increased levels of transforming growth factor (TGF)-β are associated with the induction of collagen deposition in lymph nodes (see above). Elevated levels of TNF-α and IL-6 have been demonstrated in plasma and cerebrospinal fluid (CSF), and increased expression of TNF-α, IL-1β, IFN-γ, and IL-6 has been demonstrated in the lymph nodes of HIV-infected individuals. RANTES, MIP-1α (CCL3), and MIP-1β (CCL4) **(Chap. 342)** inhibit infection by and spread of R5 HIV-1 strains, while *stromal cell–derived factor* (SDF) 1 inhibits infection by and spread of X4 strains. The mechanisms whereby the CC-chemokines RANTES (CCL5), MIP-1α (CCL3), and MIP-1β (CCL4) inhibit infection of R5 strains of HIV, or SDF-1 blocks X4 strains of HIV, involve blocking of the binding of the virus to its co-receptors, the CC-chemokine receptor CCR5 and the CXC-chemokine receptor CXCR4, respectively. Other soluble factors that have not yet been fully characterized, such as soluble CD8 antiviral factor (CAF), also have been shown to suppress HIV replication, independent of co-receptor usage.

LYMPHOCYTE TURNOVER IN HIV INFECTION

The immune systems of patients with HIV infection are characterized by a profound increase in lymphocyte turnover that is immediately reduced with effective cART. Studies utilizing in vivo or in vitro labeling of lymphocytes in the S-phase of the cell cycle have demonstrated a tight correlation between the degree of lymphocyte turnover and plasma levels of HIV RNA. This increase in turnover is seen in CD4+ and CD8+ T lymphocytes as well as B lymphocytes and can be observed in peripheral blood and lymphoid tissue. Mathematical models derived from these data suggest that one can view the lymphoid pool as consisting of dynamically distinct subpopulations of cells that are differentially affected by HIV infection. A major consequence of HIV infection appears to be a shift in cells from a more quiescent pool to a pool with a higher turnover rate. It is likely that a consequence of a higher rate of turnover is a higher rate of cell death. It has been suggested that the more rapid decline in CD4+ compared with CD8+ T cells may be linked to alterations in inflammatory and homeostatic cytokines that cause increased activation-induced death without replenishment of CD4+ but not CD8+ T cells. (See **Table 197-5** for additional mechanisms of depletion.)

TABLE 197-5

Proposed Mechanisms of CD4+ T Cell Dysfunction and Depletion

Direct Mechanisms	Indirect Mechanisms
Loss of plasma membrane integrity due to viral budding	Aberrant intracellular signaling events
Accumulation of unintegrated viral DNA Activation of DNA-dependent protein kinase during viral integration into host genome	Autoimmunity
Interference with cellular RNA processing	Innocent bystander killing of viral antigen–coated cells
Intracellular gp120-CD4 autofusion events	Apoptosis, pyroptosis (caspase-1 associated inflammation), autophagy
Syncytia formation	Inhibition of lymphopoiesis from reduced survival cytokines and lymphoid tissue integrity Activation-induced cell death Elimination of HIV-infected cells by virus-specific immune responses

THE ROLE OF VIRAL RECEPTORS AND CO-RECEPTORS IN HIV PATHOGENESIS

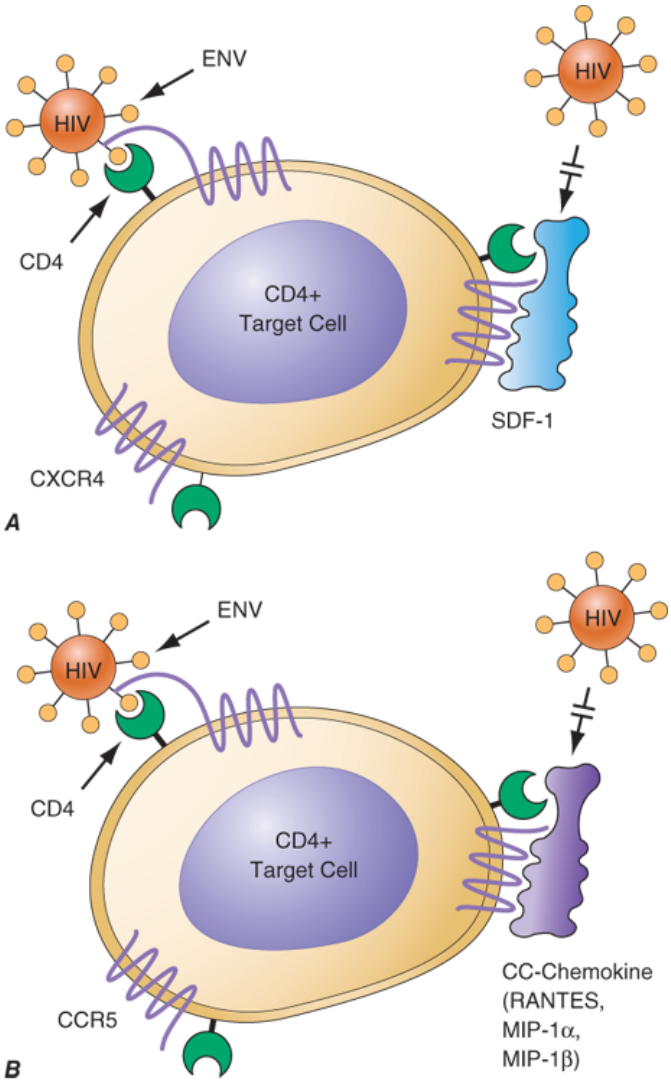
CCR5 AND CXCR4

As mentioned above, HIV-1 utilizes two major co-receptors along with CD4 to bind to, fuse with, and enter target cells; these co-receptors are CCR5 and CXCR4, which are also receptors for certain endogenous chemokines. Strains of HIV that utilize CCR5 as a co-receptor are referred to as *R5 viruses*. Strains of HIV that utilize CXCR4 are referred to as *X4 viruses*. Many virus strains are *dual tropic* in that they utilize both CCR5 and CXCR4; these are referred to as *R5X4 viruses*.

The natural chemokine ligands for the major HIV co-receptors can readily block entry of HIV. For example, the CC-chemokines RANTES (CCL5), MIP-1α (CCL3), and MIP-1β (CCL4), which are the natural ligands for CCR5, block entry of R5 viruses, whereas SDF-1, the natural ligand for CXCR4, blocks entry of X4 viruses. The mechanism of inhibition of viral entry is a steric inhibition of binding that is not dependent on signal transduction **(Fig. 197-25)**.

FIGURE 197-25

Model for the role of co-receptors CXCR4 and CCR5 in the efficient binding and entry of X4 (***A***) and R5 (***B***) strains of HIV-1, respectively, into CD4+ target cells. Blocking of this initial event in the virus life cycle can be accomplished by inhibition of binding to the co-receptor by the normal ligand for the receptor in question. The ligand for CXCR4 is stromal cell–derived factor (SDF-1); the ligands for CCR5 are RANTES, MIP-1α, and MIP-1β.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The transmitting virus is almost invariably an R5 virus that predominates during the early stages of HIV disease. In the absence of cART or in cases of therapy failure, there is a transition to a predominantly X4 virus in approximately half of individuals infected with subtype B. The transition is often preceded by dual R5X4 strains, and detection of X4 variants is associated with a relatively rapid decline in CD4+ T cell counts, increased HIV plasma viremia, and progression of disease. However, the other half of infected individuals progress in their disease while maintaining predominance of an R5 virus, and individuals infected with subtype C rarely switch from CCR5 tropism to CXCR4 tropism. The reason for this difference is unclear.

The basis for the tropism of different envelope glycoproteins for either CCR5 or CXCR4 relates to the ability of the HIV envelope, including the third variable region (V3 loop) of gp120, to interact with these co-receptors. In this regard, binding of gp120 to CD4 induces a conformational change in gp120 that increases its affinity for the relevant co-receptor. Finally, R5 viruses are more efficient in infecting monocytes/macrophages and microglial cells of the brain (see "Neuropathogenesis in HIV Disease," below).

THE INTEGRIN ALPHA 4 BETA 7

The integrin *α4β7* is an accessory receptor for HIV. It is not essential for the binding and infection of a CD4+ T cell by HIV; however, it likely plays an important role in the transmission of HIV at mucosal surfaces such as the genital tract and gut and contributes somewhat to the pathogenesis of HIV disease. The integrin α4β7, which is the gut homing receptor for peripheral T cells, binds in its activated form to a specific tripeptide in the V2 loop of gp120, resulting in rapid activation of leukocyte function–associated antigen 1 (LFA-1), the central integrin in the establishment of virologic synapses,

which facilitate efficient cell-to-cell spread of HIV. It has been demonstrated that $\alpha4\beta7^{high}$ CD4+ T cells are more susceptible to productive infection than are $\alpha4\beta7^{low–neg}$ CD4+ T cells because this cellular subset is enriched with metabolically active CD4+ T cells that are $CCR5^{high}$. These cells are present at the gut and genital tract mucosal surfaces. Importantly, it has been demonstrated that the virus that is transmitted during sexual exposure binds much more efficiently to α4β7 than does the virus that diversifies from the transmitting virus over time by mutation, particularly involving the accumulation of glycogens on the surface of the HIV envelope (see "Early Events in HIV Infection: Primary Infection and Initial Dissemination of Virus," above).

CELLULAR TARGETS OF HIV

CD4+ T lymphocytes and to a lesser extent CD4+ cells of the myeloid lineage are the principal targets of HIV and are the only cells that can be productively infected with HIV. Circulating DCs have been reported to express low levels of CD4, although high expression of the restriction factor SAMHD1 in myeloid (mDC) and plasmacytoid (pDC) DCs limits HIV replication in these cells by depleting intracellular pools of dNTPs and directly degrading viral RNA. Epidermal Langerhans cells express CD4 and have been infected by HIV in vivo, although, they too restrict replication by high expression of the host restriction factor, langerin. As has been shown in vivo for DCs, FDCs, and B cells, Langerhans cells are more likely to bind and transfer virus to activated CD4+ T cells than to be productively infected themselves.

Of potential clinical relevance is the demonstration that thymic precursor cells, which were assumed to be negative for CD3, CD4, and CD8 molecules, actually do express low levels of CD4 and can be infected with HIV in vitro. In addition, human thymic epithelial cells transplanted into an immunodeficient mouse can be infected with HIV by direct inoculation of virus into the thymus. Since these cells may play a role in the normal regeneration of CD4+ T cells, it is possible that their infection and depletion contribute, at least in part, to the impaired ability of the CD4+ T cell pool to completely reconstitute itself in certain infected individuals in whom cART has suppressed plasma viremia to below the level of detection (see below). In addition, CD34+ monocyte precursor cells have been shown to be infected in vivo in patients with advanced HIV disease. It is likely that these cells express low levels of CD4, and therefore it is not essential to invoke CD4-independent mechanisms to explain the infection. The clinical relevance of this finding is unclear.

QUALITATIVE AND QUANTITATIVE ABNORMALITIES OF MONONUCLEAR CELLS

CD4+ T Cells

The primary immunopathogenic lesion in HIV infection involves CD4+ T cells, and the range of CD4+ T cell abnormalities in advanced HIV infection is broad. The defects are both quantitative and qualitative and ultimately impact virtually every limb of the immune system, indicating the critical dependence of the integrity of the immune system on the inducer/helper function of CD4+ T cells. In advanced HIV disease, most of the observed immune defects can ultimately be explained by the quantitative depletion of CD4+ T cells. However, T cell dysfunction can be demonstrated in patients early in the course of infection, even when the CD4+ T cell count is in the low-normal range. The degree and spectrum of dysfunctions increase as the disease progresses, reflecting the range of CD4+ T cell functional heterogeneity, especially in lymphoid tissues. One of the first sites of intense HIV replication is in the GALT where CD4+ T_H17 cells reside; they are important for host defense against extracellular pathogens in the intestinal mucosa and help maintain the integrity of the gut epithelium. In HIV infection, they are depleted by direct and indirect effects of viral replication and cause loss of gut homeostasis and integrity, as well as a shift toward a T_H1 phenotype. Studies have shown that even after many years of cART, normalization of the CD4+ T cells in the GALT remains incomplete. In lymph nodes, HIV perturbs another important subset of the CD4+ helper T lineage, namely TF_H cells (see "Lymphoid Organs and HIV Pathogenesis," above). TF_H cells, which are derived either directly from naïve CD4+ T cells or from other T_H precursors, migrate into B follicles during germinal center reactions and provide help to antigen-specific B cells through cell–cell interactions and secretion of cytokines to which B cells respond, the most important of which is IL-21. As with T_H17 cells, TF_H cells are highly susceptible to HIV infection. However, in contrast to T_H17 and most other CD4+ T cell subsets, the number of TF_H cells is increased in lymph nodes of HIV-infected individuals, especially those who are viremic. It is unclear whether this increase is helpful to responding B cells, although the likely outcome is that the increase in numbers is detrimental to the quality of the humoral immune response against HIV (see "Immune Response to HIV," below). In addition, defects of central memory cells are a critical component of HIV immunopathogenesis. The progressive loss of antigen-specific CD4+ T cells has important implications for the control of HIV infection. In this regard, there is a correlation between the maintenance of HIV-specific CD4+ T cell proliferative responses and improved control of infection. Essentially every T cell function has been reported to be abnormal at some stage of HIV infection. Loss of polyfunctional HIV-specific CD4+ T cells, especially those that produce IL-2, occurs early in disease, whereas IFN-producing CD4+ T cells are maintained longer and do not correlate with control of HIV viremia. Other abnormalities include impaired expression of IL-2 receptors, defective IL-2 production, reduced expression of the IL-7 receptor (CD127), and a decreased proportion of CD4+ T cells that express CD28, a major co-stimulatory molecule necessary for the normal activation of T cells, which is also depleted as a result of aging. Cells lacking expression of CD28 do not respond normally to activation signals and may express markers of terminal activation including HLA-DR, CD38, and CD45RO. As mentioned above ("The Role of Immune Activation and Inflammation in HIV Pathogenesis"), a subset of CD4+ T cells referred to as *T regulatory cells*, or T-regs, may be involved in damping aberrant immune activation that propagates HIV replication. The presence of these T-reg cells correlates with lower viral loads and higher CD4+/CD8+ T cell ratios. A loss of this T-reg capability with advanced disease may be detrimental to the control of virus replication.

It is difficult to explain completely the profound immunodeficiency noted in HIV-infected individuals solely on the basis of direct infection and quantitative depletion of CD4+ T cells. This is particularly apparent during the early stages of HIV disease, when CD4+ T cell numbers may be only marginally decreased. In this regard, it is likely that CD4+ T cell dysfunction results from a combination of depletion of cells due to direct infection of the cell and a number of virus-related but indirect effects on the cell such as elimination of "innocent bystander cells" (Table 197-5). Several of these effects have been demonstrated ex vivo and/or by the analysis of cells isolated from the peripheral blood. Soluble viral proteins, particularly gp120, can bind with high affinity to the CD4 molecules on uninfected T cells and monocytes; in addition, virus and/or viral proteins can bind to DCs or FDCs. HIV-specific antibody can recognize these bound molecules and potentially collaborate in the elimination of the cells by ADCC. HIV envelope glycoproteins gp120 and gp160 manifest high-affinity binding to the CD4 molecule as well as to various chemokine receptors. Intracellular signals transduced by gp120 through both CD4 and CCR5/CXCR4 have been associated with a number of immunopathogenic processes including anergy, apoptosis, and abnormalities of cell trafficking. The molecular mechanisms responsible for these abnormalities include dysregulation of the T cell receptor–phosphoinositide pathway, p56lck activation, phosphorylation of focal adhesion kinase, activation of the MAP kinase and ras signaling pathways, and downregulation of the co-stimulatory molecules CD40 ligand and CD80.

The inexorable decline in CD4+ T cell counts that occurs in most HIV-infected individuals may result in part from the inability of the immune system to regenerate over an extended period of time the rapidly turning over CD4+ T cell pool efficiently enough to compensate for both HIV-mediated and naturally occurring attrition of cells. In this regard, the degree and duration of decline of CD4+ T cells at the time of initiation of therapy is an important predictor of the restoration of these cells. A person who maintains a very low CD4+ T cell count for a considerable period of time before the initiation of cART almost invariably has an incomplete reconstitution of such cells. At least two major mechanisms may contribute to the failure of the CD4+ T cell pool to reconstitute itself adequately over the course of HIV infection. The first is the destruction of lymphoid precursor cells, including thymic and bone marrow progenitor cells; the other is the gradual disruption of the lymphoid tissue architecture and microenvironment, which is essential for efficient regeneration of immunocompetent cells. Finally, during the advanced stages of CD4+ T lymphopenia, there are increased serum levels of the homeostatic cytokine IL-7. It was initially felt that this elevation was a homeostatic response to the lymphopenia; however, recent findings suggest that the increase in serum IL-7 was a result of reduced utilization of the cytokine related to the loss of cells expressing the IL-7 receptor, CD127, which serves as a normal physiologic regulator of IL-7 production.

CD8+ T Cells

A relative CD8+ T lymphocytosis is generally associated with high levels of HIV plasma viremia and likely reflects an immune response to the virus as well as dysregulated homeostasis associated with generalized immune activation. During the late stages of HIV infection, there may be a significant reduction in the numbers of CD8+ T cells despite the presence of high levels of viremia. HIV-specific CD8+ CTLs have been demonstrated in HIV-infected individuals early in the course of disease, and their emergence often coincides with a decrease in plasma viremia—an observation that is a factor in the proposal that virus-specific CTLs can control HIV disease for a finite period of time in a certain percentage of infected individuals. However, emergence of HIV escape mutants that ultimately evade these HIV-specific CD8+ T cells has been described in the majority of HIV-infected individuals who are not receiving cART. In addition, as the disease progresses, the functional capability of these cells gradually decreases, at least in part due to the persistent nature of HIV infection that causes functional exhaustion via the upregulation of inhibitory receptors such as PD-1 and TIGIT on HIV-specific CD8+ T cells (see "The Role of Immune Activation and Inflammation in HIV Pathogenesis," above). As chronic immune activation persists, there are also systemic effects on CD8+ T cells, such that as a population they assume an abnormal phenotype characterized by expression of activation markers such as HLA-DR and CD38 with an absence of expression of the IL-2 receptor (CD25) and a reduced expression of the IL-7 receptor (CD127). In addition, CD8+ T cells lacking CD28 expression are increased in HIV disease, reflecting a skewed expansion of a less differentiated CD8+ T cell subset. This skewing of subsets is also associated with diminished polyfunctionality, a qualitative difference that distinguishes elite controllers from progressors. Elite controllers can also be distinguished from progressors by the maintenance in the former of a high proliferative capacity of their HIV-specific CD8+ T cells coupled to increases in perforin expression and elimination of infected targets, characteristics that are markedly diminished in advanced HIV disease. It has been reported that the phenotype of CD8+ T cells in HIV-infected individuals may be of prognostic significance. Those individuals whose CD8+ T cells developed a phenotype of HLA-DR+/CD38– following seroconversion had stabilization of their CD4+ T cell counts, whereas those whose CD8+ T cells developed a phenotype of HLA-DR+/CD38+ had a more aggressive course and a poorer prognosis. In addition to the defects in HIV-specific CD8+ CTLs, functional defects in other MHC-restricted CTLs, such as those directed against influenza and CMV, have been demonstrated. CD8+ T cells secrete a variety of soluble factors that inhibit HIV replication, including the CC-chemokines RANTES (CCL5), MIP-1α (CCL3), and MIP-1β (CCL4) as well as potentially a number of as yet unidentified factors. The presence of high levels of HIV viremia in vivo as well as exposure of CD8+ T cells in vitro to HIV envelope, both of which are associated with aberrant immune activation, have been shown to be associated with a variety of cellular functional abnormalities. Furthermore, since the integrity of CD8+ T cell function depends in part on adequate inductive signals from CD4+ T cells, the defect in CD8+ CTLs is likely compounded by the quantitative loss and qualitative dysfunction of CD4+ T cells.

B Cells

The predominant defect in B cells from HIV-infected individuals is one of aberrant cellular activation, which is reflected by increased propensity to terminal differentiation and immunoglobulin secretion, as well as increased expression of markers of activation and exhaustion. As a result of activation and differentiation in vivo, B cells from HIV viremic patients manifest a decreased capacity to mount a proliferative response to ligation of the B cell

antigen receptor and other B cell stimuli in vitro. B cells from HIV-infected individuals manifest enhanced spontaneous secretion of immunoglobulins in vitro, a process that reflects their highly differentiated state in vivo. There is also an increased incidence of EBV-related B cell lymphomas in HIV-infected individuals that are likely due to combined effects of defective T cell immune surveillance and increased B cell turnover that increases the risk of oncogenesis. Untransformed B cells cannot be infected with HIV, although HIV or its products can activate B cells directly. B cells from patients with high levels of viremia bind virions to their surface via the CD21 complement receptor. It is likely that in vivo activation of B cells by replication-competent or defective virus as well as viral products during the viremic state accounts at least in part for their activated phenotype. B cell subpopulations from HIV-infected individuals undergo a number of changes over the course of HIV disease, including the attrition of resting memory B cells and replacement with several aberrant memory and differentiated B cell subpopulations that collectively express reduced levels of CD21 and either increased expression of activation markers or inhibitory receptors associated with functional exhaustion. The more activated and differentiated B cells are also responsible for increased secretion of immunoglobulins and increased susceptibility to Fas-mediated apoptosis. In more advanced disease, there is also the appearance of immature B cells associated with CD4+ T cell lymphopenia. Despite increased frequencies of germinal center B cells and CD4+ TF_H cells, both of which are required for effective humoral immunity, cognate B cell–CD4+ T cell interactions in lymphoid tissues are perturbed in HIV-infected individuals, especially those with persistent viremia. In vivo, the aberrant activated state of B cells manifests itself by hypergammaglobulinemia and by the presence of circulating immune complexes and autoantibodies. HIV-infected individuals respond poorly to primary and secondary immunizations with protein and polysaccharide antigens. Using immunization with influenza vaccine, it has been demonstrated that there is a memory B cell defect in HIV-infected individuals, particularly those with high levels of HIV viremia. There is also evidence that responses to HIV and non-HIV antigens in infected individuals, especially those who remain viremic, are enriched in abnormal subsets of B cells that either are highly prone to apoptosis or show signs of functional exhaustion. Taken together, these B cell defects are likely responsible at least in part for the inadequate humoral response to HIV as well as to decreased response to vaccinations and the increase in certain bacterial infections seen in advanced HIV disease in adults. In addition, they likely contribute to the inadequacy of host defenses against bacterial infections that play a role in the increased morbidity and mortality of HIV-infected children. The absolute number of circulating B cells also may be depressed in HIV infection; this phenomenon likely reflects increased activation-induced apoptosis as well as a redistribution of cells out of the circulation and into the lymphoid tissue—phenomena that are associated with ongoing viral replication.

Monocytes/Macrophages

Circulating monocytes are generally normal in number in HIV-infected individuals; however, there is evidence of increased activation within this lineage. The increased level of sCD14 and other biomarkers (see above) reported in HIV-infected individuals is an indirect marker of monocyte activation in vivo. A number of other abnormalities of circulating monocytes have been reported in HIV-infected individuals, many of which may be related directly or indirectly to aberrant in vivo immune activation. In this regard, increased levels of lipopolysaccharide (LPS) are found in the sera of HIV-infected individuals due, at least in part, to translocation across the gut mucosal barrier (see above). LPS is a highly inflammatory bacterial product that preferentially binds to macrophages through CD14 and Toll-like receptors, resulting in cellular activation. Functional abnormalities of monocyte/macrophages in HIV disease include decreased secretion of IL-1 and IL-12; increased secretion of cytokines such as IL-10 and IL-18 and markers of coagulation such as D-dimer; defects in antigen presentation and induction of T cell responses due to decreased MHC class II expression; and abnormalities of Fc receptor function, C3 receptor–mediated clearance, oxidative burst responses, and certain cytotoxic functions such as ADCC, possibly related to low levels of expression of Fc and complement receptors. Monocytes express the CD4 molecule and several co-receptors for HIV on their surface, and thus are potential targets of HIV infection. However, in vivo infection of circulating monocytes is difficult to demonstrate, although infection of tissue macrophages and macrophage-lineage cells in the brain (infiltrating macrophages or resident microglial cells) and lung (pulmonary alveolar macrophages) can be demonstrated easily. Tissue macrophages are an important source of HIV during the inflammatory response associated with opportunistic infections and can serve as persistent reservoirs of HIV infection, thus representing an obstacle to the eradication of HIV by antiretroviral drugs. Infection of monocyte precursors in the bone marrow may directly or indirectly be responsible for certain of the hematologic abnormalities in HIV-infected individuals. However, as with DCs, monocytes and macrophages express high levels of host restriction factors that likely help explain the low contribution of myeloid cells to the overall viral burden in HIV-infected individuals.

Dendritic and Langerhans Cells

DCs and Langerhans cells are not productively infected with HIV, but they are thought to play an important role in the initiation of HIV infection by virtue of the ability of HIV to bind to cell-surface C-type lectin receptors, particularly DC-SIGN (see above) and langerin. However, while langerin provides a host barrier for replication by trafficking HIV to acidic compartments for degradation, DC-SIGN retains HIV in early endosomal compartments. This allows efficient presentation of intact virus to CD4+ T cell targets that become infected; complexes of infected CD4+ T cells and DCs provide an optimal microenvironment for virus replication. Furthermore, pDCs secrete large amounts of IFN-α in response to viral infections and as such play an important role in innate sensing of HIV during early phase of infection. The numbers of circulating pDCs are decreased in HIV infection through mechanisms that remain unclear, although several studies have shown increased lymphoid tissue recruitment of DCs associated with lymphoid hyperplasia and inflammation. The mDCs or conventional DCs are also involved in the initiation of adaptive immunity in draining lymph nodes by presenting antigen to T cells and B cells, as well as by secreting cytokines such as IL-12, IL-15, and IL-18 that activate other immune cells. There are also indications that the relatively low infectibility of DCs may be associated with the expression of host restriction factors, including APOBEC3G and SAMHD1 (see above).

Natural Killer Cells

The role of NK cells is to provide immunosurveillance against virus-infected cells, certain tumor cells, and allogeneic cells (Chap. 342). There are no convincing data that HIV productively infects NK cells in vivo; however, functional abnormalities in NK cells have been observed throughout the course of HIV disease, and the severity of these abnormalities increases as disease progresses. NK cells are part of the innate immune system and act by direct killing of infected cells and secretion of antiviral cytokines and chemokines. In early HIV infection there is an increase in the activation of NK cells, and the capacity to secrete IFN-γ is maintained, although they manifest reduced cytotoxic function. During chronic HIV infection, both NK cell cytotoxicity and cytokine secretion become impaired. Given that HIV infection of target cells downregulates HLA-A and B, but not HLA-C and D molecules, this may explain in part the relative inability of NK cells to kill HIV-infected target cells. However, the NK cell impairments, especially in patients with high levels of virus replication, are associated with an expansion of an "anergic" CD56–/CD16+ NK cell subset. This abnormal subset of NK cells manifests an increased expression of inhibitory NK cell receptors (iNKRs) and a substantial decrease in expression of natural cytotoxicity receptors (NCRs) and shows a markedly impaired lytic activity. The overrepresentation of this abnormal subset of NK cells may explain in part the observed defects in NK cell function in HIV-infected individuals and likely begins to occur during primary infection. The relative expression of iNKRs and NCRs—as well as their ligands, which include HLA class I molecules—has an impact on the antiviral functions associated with NK cells, including direct killing and ADCC. Polymorphisms in iNKR and NCR alleles have been linked to HIV-1 disease outcomes. NK cells also serve as important sources of HIV-inhibitory CC-chemokines. NK cells isolated from HIV-infected individuals constitutively produce high levels of MIP-1α (CCL3), MIP-1β (CCL4), and RANTES (CCL5), although the impact of these chemokines on HIV replication in vivo is unclear. Finally, NK cell–DC interactions are important for normal immune function. NK cells and DCs reciprocally modulate each other's activation and maturation. These interactions are markedly impaired in HIV-infected individuals with high levels of plasma viremia.

GENETIC FACTORS IN HIV-1 AND AIDS PATHOGENESIS

Candidate gene approaches and genome-wide association studies (GWAS) have identified polymorphisms in host genes that contribute to inter-individual variation in (1) the risk of acquiring HIV, (2) the steady-state levels of HIV that are established soon after infection (virologic set point), (3) the rate at which untreated HIV-infected patients progress to AIDS as well as risk of developing specific AIDS-defining illnesses (e.g., renal and neurologic diseases), and (4) the level of immune reconstitution (e.g., CD4+ counts) achieved after initiation of virally suppressive ART. The key polymorphisms that influence these four traits are summarized in Table 197-6, and their identification has greatly advanced our understanding of the genes that influence HIV-AIDS pathogenesis. Of particular interest are polymorphisms in two chromosomal regions, as they are associated with consistent effects on HIV acquisition, virologic set point, and/or rates of HIV disease progression: the region in chromosome 3 that includes the gene that encodes the HIV co-receptor *CC chemokine receptor 5* (*CCR5*) and the *major histocompatibility locus* (MHC) in chromosome 6 (Fig. 197-26).

TABLE 197-6

Host Genetic Factors That Influence Risk of HIV-1 Acquisition and Rates of HIV-1 Disease Progression

Gene[a]	Genetic Variation	Mechanisms[b]	Genetic Effect on HIV-AIDS[c]
Genes in MHC Locus			
HLA-B	B*27 and B*57 B*35 HLA-Bw4	Presentation of specific HIV antigens Restriction of specific HIV peptide presentation Providing ligands for activating KIR	Slower progression to AIDS; lower viral load Faster progression to AIDS; higher viral load Slower progression to AIDS
HLA class I allele	Homozygosity of *HLA*-class I alleles Shared donor-recipient *HLA* alleles Rare *HLA* alleles	Reduced repertoire for epitope recognition Preadaptation of HIV strains Limited adaptation of HIV strains; less frequent escape mutants	Faster progression to AIDS; increased risk of mother-to-child transmission Faster disease progression to AIDS Protection against HIV infection
HLA class II allele	*HLA-DRB1* alleles	Influencing protein specificity of CD4+ T cell responses to HIV Gag and Nef proteins	HLA-DRB1*15:02—lower viral load HLA-DRB1*03:01—higher viral load
HLA extended haplotype	A1-B8-DR3-DQ2 (AH 8.1)	Increased proinflammatory responses; higher TNF-α production	Faster progression to AIDS
HLA-C	35 kb upstream, rs9264942-C	Increased expression of HLA-C	Decreased viral load set point
HCP5	rs2395029-G	Linkage disequilibrium with HLA-B*57:01	Lower viral load
MICA	Noncoding SNP near *MICA*, rs4418214-T	May affect HLA class I peptide presentation—linkage with protective *HLA-B* alleles	Enriched in HIV-1 controllers
PSORS1C3	rs3131018-A	May affect HLA class I peptide presentation	Enriched in HIV-1 controllers
ZNRD1	rs9261174-C	Possible interference in processing of HIV transcripts; influencing ZNRD1 expression; linkage disequilibrium with *HLA-A10*	Slower disease progression to AIDS
Chemokine Receptors			
CCR5	32-bp deletion in the ORF (Δ32), rs333 Promoter SNPs, haplotypes (HHA to HHG*2)	Truncated CCR5 protein Altered CCR5 expression, e.g., HHE haplotype correlates with high CCR5 expression	Δ32/Δ32: resistance to acquiring HIV infection Δ32/wild type: delays AIDS onset; improves immune reconstitution during ART HHE/HHE: increased HIV/AIDS susceptibility
CCR2	SNP in ORF (64 V⟶I), rs1799864	Possibly due to linkage with polymorphism in *CCR5* promoter	64I: delayed AIDS onset
CCRL2	Coding SNP (167 Y⟶F) rs3204849 rs1015164	Possibly due to linkage with *CCR5* haplotype Possibly due to linkage with *CCR5* haplotype	167F is associated with accelerated progression to AIDS and more rapid development of PCP Associated with high viral load set point

Gene[a]	Genetic Variation	Mechanisms[b]	Genetic Effect on HIV-AIDS[c]
CXCR6	rs2234358 G→T in the 3'UTR	Trafficking of effector T cells and activation of NK T cells; minor HIV co-receptor	Prevalence of rs2234358-T was lower in long-term nonprogressors and viremic controllers
CX3CR1	SNPs in ORF (249 V→I, rs3732379; and 280 T→M, rs3732378)	280M reduces receptor expression and binding of fractalkine, the CX3CR1 ligand	249I and 280M associated with faster AIDS onset in some Caucasian cohorts; inconsistent effects detected in other cohorts
DARC	African-specific promoter SNP (–46T→C), rs2814778	–46C/C associates with low neutrophil counts; influences circulating chemokine levels; alters HIV binding to RBCs and transinfection of HIV-1	–46C/C: increased risk of acquiring HIV but slower HIV disease progression; Duffy-null-associated low neutrophil trait associated with increased HIV risk in persons of African descent
Chemokines			
CCL3L, CCL4L	Gene copy number of *CCL3L* and *CCL4L*	High numbers of *CCL3L* and *CCL4L* gene-containing segmental duplications correlate with high CCL3L and CCL4L levels	Gene copy number lower than population median associated with increased HIV/AIDS susceptibility and reduced immune reconstitution during ART
CCL5	Promoter SNPs	Altered gene expression	Altered HIV-AIDS susceptibility
CCL2	Promoter SNP (–2578 T→G), rs1024611	–2578G allele: increased CCL2 expression and monocyte recruitment	–2578G/G associated with increased risk of developing HIV-1-associated dementia and rapid AIDS onset
Cytokines			
IL-6	Promoter SNP (–174 G→C), rs1800795	–174G/G associated with increased IL-6 and CRP levels	–174G/G associated with high risk of KS development and variable recovery of CD4 cells during ART
IL-7RA	Coding SNP (244 T→I), rs6897932	244 I/I associated with increased signal transduction and proliferation in response to IL-7	244 I/I associated with faster CD4+ T cell recovery after ART initiation
IL-10	Promoter SNP (–592 C→A), rs1800872	–592A results in decreased IL-10 levels	–592A associated with increased HIV-AIDS susceptibility
Innate Immunity Genes			
MBL	Coding alleles (O) X allele (promoter SNP –221)	Low plasma concentration and structural variation of MBL protein Decreased levels of MBL protein	Slow progression to AIDS in heterozygous subjects (A/O) Faster progression to AIDS in homozygous X/X subjects
Apobec-3G	ORF SNP (186 H→R), rs8177832	Reduced anti-HIV-1 activity	186R associated with rapid AIDS onset in African Americans
Apobec-3F	Haplotype tagged by ORF SNP (231 I→V), rs2076101	231V variant may influence Vif-mediated Apobec-3F degradation	231V associated with lower VL, slower progression to AIDS and delayed progression to PCP
TLR7	ORF SNP (Gln11Leu), rs179008	Decreased expression of TLR7 leading to lack of recognition of HIV-infected cells	Leu-containing protein associated with higher viral load and faster progression to AIDS

Gene[a]	Genetic Variation	Mechanisms[b]	Genetic Effect on HIV-AIDS[c]
PARD3B	rs11884476 (C→G), near exon 20	Direct interaction with HIV, signaling through SMAD family of proteins	rs11884476-G associated with slower progression to AIDS
IFNL4	Frameshift mutation (TT→ΔG), rs368234815	Functional polymorphism in *IFNL4* gene, possibly in linkage with IL28B variant and regulates IL-28B levels	rs368234815-ΔG associated with higher prevalence of AIDS-defining illnesses and potentially increased HIV-1 infection risk
Others			
ApoE	E4 allele	E4 enhances HIV cell entry in vitro	ApoE4/E4 associated with rapid AIDS onset and dementia
ApoL1/ MYH9	Several risk haplotypes, including G1	Unknown	Increased risk for HIV-associated nephropathy
RYR3	ORF SNP (A →G), rs2229116	Unknown, potential impact on calcium signaling and homeostasis	rs2229116-G associated with subclinical atherosclerosis
PROX1	rs17762192-G, 36kb upstream of *PROX1*	Unknown, presumably due to its impact on PROX1 expression, which is a negative regulator of IFN-γ	rs17762192-G: reduced rate of disease progression
Gene–Gene Interaction			
KIR+HLA	KIR3DS1 + HLA-Bw4-80Ile HLA-C1 + KIR2DL3	Altered NK cell activity required to eliminate HIV-infected cells Reduction of inhibitory KIR likely results in increased immune activation; impaired killing of latently infected cells; and a higher proviral burden	KIR3DS1 associated with HLA-Bw4-80Ile; +: delayed AIDS onset HLA-C1+/KIR2DL3+: better immune recovery after viral load suppression on ART
LILRB2+HLA	LILRB2 + HLA class I	Regulation of dendritic cells by LILRB2-HLA engagement	Control of HIV-1
CCL3L1+ CCR5	Low *CCL3L1* gene copies + detrimental *CCR5* genotypes	Low CCL3L1 and high CCR5 expression	Increased HIV/AIDS susceptibility and reduced immune reconstitution during ART

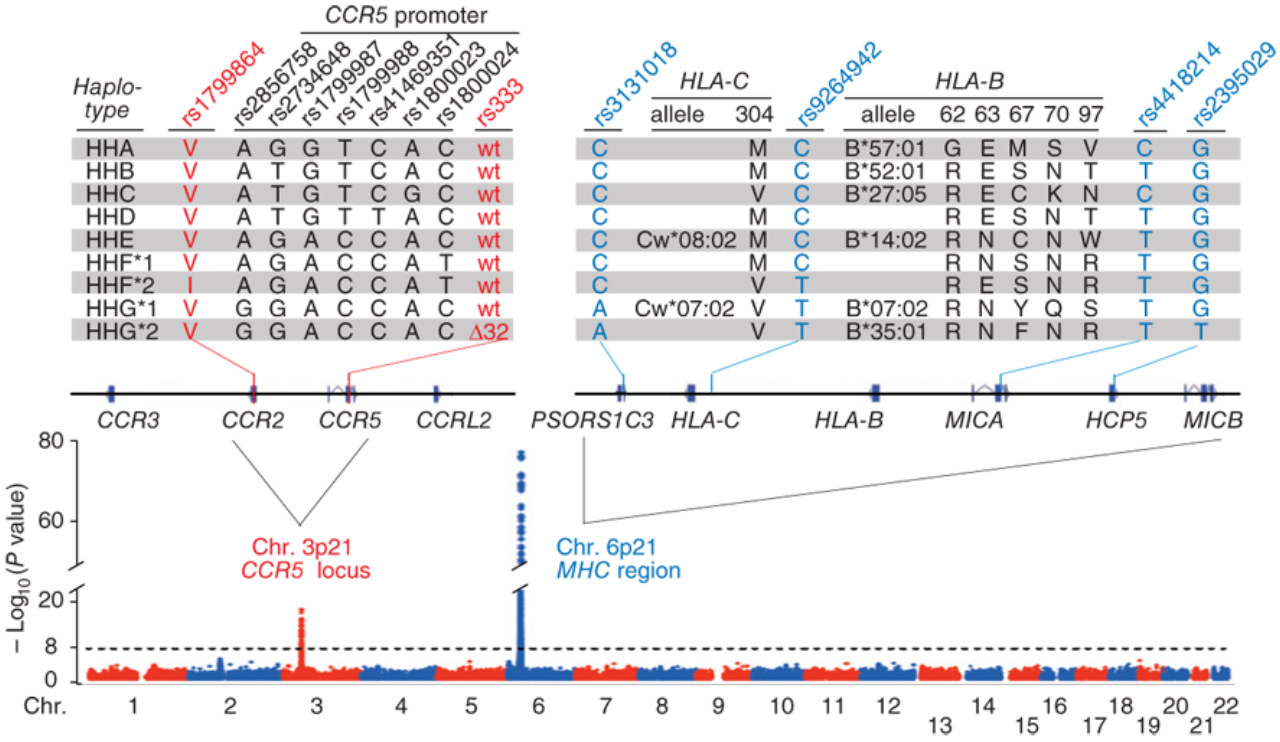
[a]Representative genes and polymorphisms and [b]possible mechanisms are listed. [c]Some of the associations are population specific and may display cohort-specific effects.

Note: Apobec, apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like; ApoE, apolipoprotein E; ART, antiretroviral therapy; CCL, CC ligand; CCL3L, CCL3-like; CCR5, CC chemokine receptor 5; CCRL2: CC chemokine receptor like 2; CRP, C-reactive protein; CXCR6, chemokine (C-X-C motif) receptor 6; DARC, Duffy antigen receptor for chemokines; HCP5, HLA class I histocompatibility antigen protein P5; HHE, human haplogroup E; HLA, human leukocyte antigen; IFN, interferon; IFNL4, interferon λ4 gene; IL, interleukin; IL-7RA, interleukin 7 receptor-α; KIR, killer cell immunoglobulin-like receptors; KS, Kaposi's sarcoma; LILRB2, leukocyte immunoglobulin-like receptor B2; MBL, mannose-binding lectin; MHC, major histocompatibility complex; MICA, MHC class I polypeptide-related sequence A; NK, natural killer; ORF, open reading frame; PARD3B, par-3 family cell polarity regulator beta; PCP, *Pneumocystis* pneumonia; PROX1, prospero homeobox 1; PSORS1C3, psoriasis susceptibility 1 candidate 3; SMAD, Mothers against decapentaplegic homolog; SNP, single nucleotide polymorphism; rs#, SNP identification number; TNF-α, tumor necrosis factor-α; UTR, untranslated region; VL, viral load; ZNRD1, zinc ribbon domain containing 1; +, present, –, absent.

Sources: Sunil K. Ahuja, MD, Weijing He, MD, Kristen Rogers, MS. Reviews for additional information: P An et al: Trends Genet 26:119, 2010; J Fellay: Antivir Ther 14:731, 2009; RA Kaslow et al: J Infect Dis 191:S68, 2005; D van Manen et al: Retrovirology 9:70, 2012; MP Martin et al: Immunol Rev 254:245, 2013; S Limou et al: Front Immunol 4:118, 2013; PJ McLaren et al: Curr Opin HIV AIDS 10:110, 2015; PJ McLaren et al: Proc Natl Acad Sci USA 112:14658, 2015; PJ McLaren, M Carrington: Nat Immunol 16:577, 2015; P An et al: PLoS Genet 12:e1005921, 2016; F Pereyra et al: Science 330:1551, 2010; I Bartha et al: PLoS Comput Biol 13:e1005339, 2017.

FIGURE 197-26

CCR5 and MHC loci: two key regions that influence HIV-AIDS pathogenesis. ***Bottom:*** Manhattan plot of genome-wide association from >6000 HIV-infected Europeans. *(Adapted from PJ McLaren et al: Proc Acad Natl Acad Sci 112:14658, 2015.)* ~8 million common variants were tested for their association with HIV viral load set point using linear regression. Genome-wide signals of association ($p < 5 \times 10^{-8}$, dotted line) were observed on chromosomes (Chr.) 6p21 in the MHC region and 3p21 in the *CCR5* locus. ***Upper left:*** schema depicting composition of *CCR5* haplotypes. *(Adapted from E Gonzalez et al: Proc Natl Acad Sci USA 96:12004, 1999.)* The 9 human haplotypes (HH) designated as HHA to HHG*2 were derived from the coding polymorphism of CCR2 64 V → I; 7 single nucleotide polymorphisms (SNPs) in the *CCR5* promoter region as well as the Δ32 coding mutation in *CCR5*; wt, wild type. Carrington and colleagues refer to the *CCR5-HHE* haplotype as the P1 haplotype (*MP Martin et al: Science 282:1907, 1998*). ***Upper right:*** Representative haplotypes at Chr. 6p21 MHC region in Europeans. *(Adapted from F Pereyra et al: Science 330:1551, 2010.)* Haplotypes defined by the SNPs identified by genome-wide association studies, classic HLA alleles and amino acids in HLA-B and HLA-C that are associated with HIV viral control. Figure is not to scale and full references are in Table 197-6.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GENETICS OF CCR5: FROM BENCH TO BEDSIDE

While the discovery of CCR5 as a major co-receptor for cell entry of HIV-1 was established by in vitro studies, genetic association studies were required to establish its seminal role in HIV pathogenesis. Initial in vitro studies revealed that a 32-bp deletion (*Δ32*) in the coding region of CCR5 contributes to resistance to CCR5-using R5 strains of HIV. The *CCR5 Δ32* allele encodes a truncated protein that is not expressed on the cell surface. Congruently, genotype-phenotype association studies in large cohorts demonstrated that individuals homozygous for the *CCR5 Δ32* allele *(Δ32/Δ32)* lack CCR5 surface expression and are highly resistant to acquiring HIV infection; heterozygosity for the *CCR5 Δ32* allele is associated with a lower risk of acquiring HIV.

The distribution of the *CCR5 Δ32* allele is population specific. Approximately 1% of individuals of European ancestry are homozygous for the *CCR5 Δ32* allele. Depending on the geographic region in Europe, up to 18% of individuals are heterozygous for the *CCR5 Δ32* allele. The *CCR5 Δ32* allele is rare in other populations. The evolutionary pressure that resulted in the emergence of the *CCR5 Δ32* allele in the European population remains unknown and has been speculated to be secondary to an ancestral pandemic, such as the plague.

Subsequent studies identified single nucleotide polymorphisms (SNPs) in the promoter (regulatory) region of *CCR5* that influence gene expression levels. Alleles bearing specific cassettes of linked polymorphisms (haplotypes) were identified and designated as human haplogroups A to G*2 (HHA to HHG*2) (Fig. 197-26). The *CCR5 Δ32* polymorphism is found on the HHG*2 haplotype. *CCR5* haplotypes A–D vs. E–G*2 differ by bearing GT versus AC at polymorphic sites rs1799987 and rs1799988 (Fig. 197-26). *CCR5-HHA* haplotype represents the ancestral haplotype (found in chimpanzees) and is associated with lower CCR5 gene expression, whereas the *CCR5-HHE* haplotype is associated with higher CCR5 expression. Methylation of DNA is a common epigenetic signaling mechanism that cells use to lock genes in the "off" position, and polymorphisms in *CCR5* haplotypes may mediate their effects by influencing DNA methylation levels in the *CCR5* locus. The *CCR5-HHE* and *CCR5-HHA* haplotypes are more sensitive and resistant, respectively, to T cell activation–induced demethylation of the *CCR5* locus.

In worldwide populations, HHE and HHC are more prevalent, whereas the ancestral HHA haplotype is more common in persons of African ancestry. The associations of *CCR5* haplotypes with HIV acquisition and/or HIV disease course are largely consistent with their effects on *CCR5* gene expression. For example, homozygosity for the *CCR5-HHE* haplotype is associated with an increased risk of acquiring HIV, progressing rapidly to AIDS, and reduced immune recovery while the patient is on ART. The HHA haplotype is associated with slower disease progression in African populations and has been speculated to be a basis for why chimpanzees (who all carry the ancestral *CCR5* HHA haplotype) naturally infected with simian immunodeficiency virus (SIV) may resist disease progression. The pairing of the HHC and *CCR5 Δ32*-bearing HHG*2 haplotypes (HHC/HHG*2 genotype) is associated with a lower risk of acquiring HIV infection and slower rate of HIV disease progression, whereas the pairing of the HHE haplotype with the HHG*2 haplotype is associated with the opposite effects. The *CCR2*-64I-bearing HHF*2 haplotype is associated with a slower HIV disease course.

Consistent with these genetic associations, polymorphisms in genes encoding ligands for CCR5 have also been demonstrated to associate with variable HIV susceptibility and disease progression rates. Examples include copy number variations of *CCL3L1* and SNPs in *CCL5*. The sum of these studies established a pivotal role of CCR5 and its ligands in HIV-AIDS pathogenesis and, potentially, immune recovery.

The discovery that the *CCR5 Δ32/Δ32* genotype is associated with strong resistance to HIV infection, and that uninfected Caucasians bearing this genotype did not appear to have impaired immunity, led to the development of two kinds of novel therapies. First, it spurred the development of a new class of therapies approved by the U.S. Food and Drug Administration (FDA), i.e., entry inhibitors (e.g., maraviroc) that block the interaction of CCR5 with the HIV envelope. Second, it led to the evaluation of novel experimental cellular therapies. An HIV-infected patient with acute myelogenous leukemia was given an allogeneic stem cell transplantation from an HLA-compatible person whose cells lacked expression of CCR5 due to the *Δ32/Δ32* genotype. There has been no evidence of HIV-1 infection in the patient who underwent the transplant thus far (~10 years). This observation provided a “proof of concept” for an HIV cure and led to the development of additional novel cellular therapies involving autologous transplantation of CD4+ T cells in which the *CCR5* gene is inactivated ex vivo using new gene editing procedures.

DISCOVERY OF HLA CLASS I ALLELES THAT ASSOCIATE WITH VIROLOGIC CONTROL OF HIV INFECTION

There is a strong association between variations within the *HLA-B* gene with protective (e.g., HLA-B*57 and B*27 alleles) or detrimental (e.g., HLA-B*35 allele) outcomes during HIV infection. Carriage of the HLA-B*57 and/or HLA-B*27 alleles is associated with slower disease progression. The beneficial effects of these alleles may relate in part to their consistent associations with a lower virologic set point as well as to higher cell-mediated immunity in HIV-infected persons. The protective effect of the HLA-B*57 and B*27 alleles on the HIV disease course is underscored by the finding that the prevalence of these alleles is higher among long-term nonprogressors and persons who control HIV replication spontaneously (elite controllers). In contrast, the HLA-B*35 allele has been associated with faster progression to AIDS and higher viral load. The prevalence of the *HLA-B* alleles differs between populations. HLA-B*57:01 in Europeans and HLA-B*57:03 in African Americans are the protective alleles. In some populations (e.g., Japanese) where the HLA-B*57/-B*27 alleles are absent, HLA-B*51 is associated with a protective phenotype.

Possession of the protective *HLA-B* alleles is associated with broader and stronger CD8+ T cell responses to HIV epitopes. The mechanisms underlying the differential effects of the *HLA-B* alleles on the course of HIV disease may relate to differences in the ability of antigen-presenting cells to present immunodominant HIV epitopes to T helper or cytotoxic T lymphocytes in the context of MHC-encoded molecules. This may result in differential immune responses that influence viral replication. In this regard, the *HLA-B* alleles that impact the course of HIV disease differ in their amino acid residues in the HLA-B peptide-binding groove; this may play a critical role in virologic control.

Investigators have also examined the influence of extended *HLA* haplotypes (linked alleles) on the course of HIV disease. The extended *HLA* ancestral haplotype (AH) 8.1 is defined by the presence of HLA-A1, HLA-B8, and HLA-DR3 alleles. AH 8.1 is the most common ancestral haplotype in Caucasians (present in 10%) and is associated with multiple autoimmune diseases in HIV-uninfected persons. These associations of AH 8.1 are thought to be due to a genetically determined hyperresponsiveness characterized by high TNF-α production and lack of complement C4A. Strong epidemiologic data indicate that carriage of AH 8.1 in HIV-infected persons is associated with a rapid decline in CD4+ T cells and faster progression to AIDS development. Gene–gene interactions between HLA alleles and other genes (e.g., killer cell immunoglobulin-like receptors) also may influence HIV disease progression rates.

POLYMORPHISMS IDENTIFIED BY GWAS THAT ASSOCIATE WITH HIV-1 ACQUSITION AND VIROLOGIC CONTROL

GWAS have not identified additional genetic variations that associate with risk of HIV-1 acquisition. By contrast, large-scale GWAS have identified SNPs, especially in the MHC, that influence HIV viral load, including in a large group of individuals termed “HIV controllers” who spontaneously (without ART) control viral replication. GWAS in HIV-infected persons of European ancestry identified four SNPs in genes in the HLA class I loci that associated with virologic control. These SNPs are within or in the vicinity of *PSORS1C3*, *HLA-C*, *MICA*, and *HCP5* genes (Fig. 197-26). As noted in this figure, the individual effects of these alleles are difficult to discern because of linkage disequilibrium. The protective effects of the SNPs in *HCP5* and *MICA* may relate to their linkage with known protective *HLA-B* alleles. The protective *HCP5* allele is in linkage disequilibrium with the HLA-B*57:01 allele, and the protective *MICA* allele tags with the HLA-B*57:01 and B*27:05 alleles. The protective *HLA-C* SNP is associated with higher HLA-C expression, and this effect is thought to be due to the altered binding of a microRNA to the *HLA-C* mRNA. Higher HLA-C expression has been associated with beneficial HIV phenotypes. The mechanism associated with the SNP in *PSORS1C3* is unknown. GWAS in African Americans identified a SNP that tags the HLA-B*57:03 allele that is known to associate with a lower virologic set point and slower disease course. Together, these GWAS data underscore the importance of variations in HLA class I loci in control of viral replication. A recent GWAS study suggested that an allele in the gene encoding *CCRL2* influences the HIV viral load set point. *CCRL2* is on chromosome 3p21 and resides ~30 kb downstream of the *CCR5* loci; its effect could potentially be due to its linkage with the *CCR5*

haplotypes. Mathematical modeling revealed that variations in host genes may explain about 10% of the observed variability in HIV viral load, whereas viral genetic diversity may explain 29% of the variability.

GENETIC ASSOCIATIONS WITH SPECIFIC AIDS AND NON-AIDS CONDITIONS

Carotid artery disease

Many of the non-AIDS events in HIV-infected individuals resemble those related to immune senescence and those found in the HIV-uninfected aging population. A functional SNP in the ryanodine receptor 3 (*RYR3*) gene was found to be associated with an increased risk of common carotid intima–media thickness (cIMT), which is a surrogate for subclinical atherosclerosis. Functional studies on *RYR3* and its isoforms demonstrate a major role of these receptors in modulating endothelial function and atherogenesis via calcium signaling pathways, providing a biologically plausible mechanism by which the SNP in *RYR3* may associate with increased cIMT risk.

Renal disease

HIV-1-associated nephropathy (HIVAN) is a form of focal sclerosing glomerulonephritis caused by direct infection of kidney epithelial cells with HIV. HIVAN is more common in persons of African descent. There is evidence that polymorphisms in the *MYH9* gene and in the neighboring *APOL1* gene are a strong determinant of susceptibility to HIVAN in African Americans. The effect of carrying two *APOL1* risk alleles explains nearly 35% of HIVAN. The mechanisms by which *MYH9*/*APOL1* variants predispose to HIVAN are currently unknown.

HIV-associated neurocognitive disorder

HIV-associated neurocognitive disorder (HAND) comprises a spectrum of neurocognitive deficits due to HIV infection. Variations in the apolipoprotein E (ApoE) gene have strong associations with Alzheimer's disease in the HIV-uninfected population. In HIV-infected persons, possession of the *ApoE4* allele has been associated with several cognitive outcomes, including dementia, peripheral neuropathy, and impairment in cognition and immediate and delayed verbal memory. Macrophage recruitment and activation play a central role in the development of many of the HAND syndromes. Variations in chemokines that play an influential role in macrophage activation and recruitment, namely *CCL2* (MCP-1) and *CCL3* (MIP-1α), have been shown to alter the risk of developing HAND. Variations in mitochondrial genes also have been associated with risk of AIDS and HAND. A GWAS identified a polymorphism in chromosome 14 in the T cell receptor α locus that may influence neurocognitive outcomes.

HIV-1 associated Pneumocystis pneumonia

Human Apobec3 cytidine deaminases are intrinsic resistance factors to HIV-1. However, HIV-1 encodes a viral infectivity factor (Vif) that degrades Apobec3 proteins. Association studies suggest a role of the genetic variation in the Apobec3 family in HIV disease. A common haplotype derived from 6 SNPs in the *Apobec-3F* gene and tagged by a codon-changing variant is associated with significantly lower viral load set point, slower rate of progression to AIDS, and delayed development of *Pneumocystis* pneumonia (PCP). In addition, a coding SNP in the *CCRL2* gene is associated with accelerated progression to AIDS and more rapid development of PCP.

associations with ART-related adverse events

Abacavir, an effective antiretroviral agent, is associated with significant risk of hypersensitivity reactions (2–9% of cases). Interestingly, while the *HLA-B*57:01* allele is associated with a slower HIV disease course, possession of this allele is associated with a higher risk of abacavir-associated hypersensitivity. Pharmacogenetic screening for the *HLA-B*57:01* allele is recommended before initiation of abacavir treatment.

NEUROPATHOGENESIS IN HIV DISEASE

While there has been a remarkable decrease in the incidence of the severe forms of HIV encephalopathy among those with access to treatment in the era of effective cART, HIV-infected individuals can still experience milder forms of neurocognitive impairment despite adequate cART. Factors that contribute to the neurocognitive decline include lack of complete control of HIV replication in the brain; production of HIV proteins that may be neurotoxic; low CD4+ T cell nadir; chronic immune activation; comorbidities such as drug abuse, microvascular disease, older age, and diabetes; and the potential for neurotoxicity of certain antiretroviral drugs. HIV has been demonstrated in the brain and CSF of infected individuals with and without neuropsychiatric abnormalities. As opposed to lymphoid tissues, there are no resident lymphocytes in the brain. The main cell types that are infected in the brain in vivo are the perivascular macrophages and the microglial cells, which can sometimes form syncytia resulting in multinucleated giant cells; low-level viral replication is also seen in perivascular astrocytes. It has been proposed that monocytes that have already been infected in the blood can migrate into the brain, where they then reside as macrophages, or macrophages can be directly infected while residing within the brain. The precise mechanisms whereby HIV enters the brain are unclear; however, they are thought to relate, at least in part, to the ability of virus-infected and immune-activated macrophages to induce adhesion molecules such as E-selectin and vascular cell adhesion molecule 1 (VCAM-1) on brain endothelium. Other studies have demonstrated that HIV gp120 enhances the expression of intercellular adhesion molecule 1 (ICAM-1) in glial cells and HIV Tat protein can disrupt the tight junctions of the brain endothelial cells to facilitate entry of HIV-infected cells into the CNS. Virus isolates from the brain are preferentially R5 strains as opposed to X4 strains; in this regard, HIV-infected individuals who are heterozygous for *CCR5-Δ32* appear to be relatively protected against the development of HIV encephalopathy. Once HIV enters the brain due to pressures of the local environment, it evolves to develop distinct sequences in the env, tat, and LTR genes. These unique sequences have been associated with neurocognitive dysfunction; however, it is unclear if they are causal (see below).

HIV-infected individuals may manifest white matter lesions as well as neuronal loss. The white matter lesions are due to axonal injury and a disruption of the blood-brain barrier and not due to demyelination. Given the absence of evidence of HIV infection of neurons, HIV-mediated effects on neurons are

thought to involve indirect pathways whereby viral proteins, particularly gp120 and Tat, trigger the release of endogenous neurotoxins from macrophages and to a lesser extent from astrocytes. In addition, it has been demonstrated that both HIV-1 Nef and Tat can induce chemotaxis of leukocytes, including monocytes, into the CNS. Neurotoxins can be released from monocytes as a consequence of infection and/or immune activation. Monocyte-derived neurotoxic factors have been reported to kill neurons via a variety of mechanisms including activation of the *N*-methyl-D-aspartate (NMDA) receptors and induction of oxidative stress. In addition, HIV gp120 shed by virus-infected monocytes could cause neurotoxicity by antagonizing the function of vasoactive intestinal peptide (VIP), by elevating intracellular calcium levels, and by decreasing neurotrophic factor levels in the cerebral cortex. A variety of monocyte-derived cytokines can contribute directly or indirectly to the neurotoxic effects in HIV infection; these include TNF-α, IL-1, IL-6, TGF-β, IFN-γ, platelet-activating factor, and endothelin. Furthermore, among the CC-chemokines, elevated levels of monocyte chemotactic protein-1 (MCP-1 or CCL-2) in the brain and CSF have been shown to correlate best with the presence and degree of HIV encephalopathy in ART-naïve patients. In addition, infection and/or activation of monocyte-lineage cells can result in increased production of eicosanoids, quinolinic acid, nitric oxide, excitatory amino acids such as L-cysteine and glutamate, arachidonic acid, platelet activating factor, free radicals, TNF-α, and TGF-β, which may contribute to neurotoxicity. Astrocytes may play diverse roles in HIV neuropathogenesis. Reactive gliosis or astrocytosis has been demonstrated in the brains of HIV-infected individuals, and TNF-α and IL-6 have been shown to induce astrocyte proliferation. In addition, astrocyte-derived IL-6 can induce HIV expression in infected cells in vitro. Furthermore, it has been suggested that astrocytes may downregulate macrophage-produced neurotoxins. Evidence of neuronal injury can be demonstrated by measuring neurofilament levels in CSF. Treatment with cART leads to improvement in neuropsychiatric manifestations and a decrease in these cytokine levels in CSF, suggesting that they are driven by the virus or by its products. However, even in patients on long-term cART, there may be evidence of persistently activated lymphocytes in the CSF. It is unclear if these lymphocytes may contribute to neuronal injury in the brain or are critical for controlling the CNS viral reservoir. However, some individuals may develop a subacute encephalitis due to an IRIS reaction (see below). This often occurs weeks or a few months after initiation of cART in individuals with low CD4+ T cell counts. It is thought that the recovery of CD4+ T cells causes a lymphocyte response to the CNS HIV reservoir. The contribution of host genetic factors to development of neuropsychiatric manifestations of HIV infection has not been well studied. However, evidence supports the role of several genetic factors including the E4 allele for apoE in an increased risk of HIV-associated neurocognitive disorders and peripheral neuropathy.

It has also been suggested that the CNS may serve as a relatively sequestered site for a reservoir of latently infected cells that might be a barrier for the eradication of virus by cART (see "The HIV Reservoir: Obstacles to the Eradication of Virus," above).

PATHOGENESIS OF KAPOSI'S SARCOMA

There are at least four distinct epidemiologic forms of KS: (1) the classic form that occurs in older men of predominantly Mediterranean or eastern European Jewish backgrounds with no recognized contributing factors; (2) the equatorial African form that occurs in all ages, also without any recognized precipitating factors; (3) the form associated with organ transplantation and its attendant iatrogenic immunosuppressed state; and (4) the form associated with HIV-1 infection. In the latter two forms, KS is an opportunistic disease; in HIV-infected individuals, unlike typical opportunistic infections, its occurrence is not strictly related to the level of depression of CD4+ T cell counts. The pathogenesis of KS is complex; fundamentally, it is an angioproliferative disease that is not a true neoplastic sarcoma, at least not in its early stages. It is a manifestation of excessive proliferation of spindle cells that are believed to be of vascular origin and have features in common with endothelial and smooth-muscle cells. In HIV disease the development of KS is dependent on the interplay of a variety of factors including HIV-1 itself, human herpes virus 8 (HHV-8), immune activation, and cytokine secretion. A number of epidemiologic and virologic studies have clearly linked HHV-8, which is also referred to as *Kaposi's sarcoma–associated herpesvirus* (KSHV), to KS not only in HIV-infected individuals but also in individuals with the other forms of KS. HHV-8 is a γ-herpesvirus related to EBV and *herpesvirus saimiri*. It encodes a homologue to human IL-6 and, in addition to KS, has been implicated in the pathogenesis of body cavity lymphoma, multiple myeloma, and monoclonal gammopathy of undetermined significance. Sequences of HHV-8 are found universally in the lesions of KS, and patients with KS are virtually all seropositive for HHV-8. HHV-8 DNA sequences can be found in the B cells of 30–50% of patients with KS and 7% of patients with AIDS without clinically apparent KS.

Between 1% and 2% of eligible blood donors are positive for antibodies to HHV-8, while the prevalence of HHV-8 seropositivity in HIV-infected men is 30–35%. The prevalence of HHV-8 seropositivity in HIV-infected women is ~4%. This finding is reflective of the lower incidence of KS in women. It has been debated whether HHV-8 is actually the transforming agent in KS; the bulk of the cells in the tumor lesions of KS are not neoplastic cells. However, it has been demonstrated that endothelial cells can be transformed in vitro by HHV-8. In this regard, HHV-8 possesses a number of genes, including homologues of the IL-8 receptor, Bcl-2, and cyclin D, that can potentially transform the host cell. Despite the complexity of the pathogenic events associated with the development of KS in HIV-infected individuals, HHV-8 is the etiologic agent of this disease. The initiation and/or propagation of KS requires an activated state and is mediated, at least in part, by cytokines. A number of factors, including TNF-α, IL-1β, IL-6, granulocyte-macrophage colony-stimulating factor (GM-CSF), basic fibroblast growth factor, and oncostatin M, function in an autocrine and paracrine manner to sustain the growth and chemotaxis of the KS spindle cells. In this regard, KSHV-derived IL-6 has been demonstrated to induce proliferation of lymphoma cells and to inhibit the cytostatic effects of IFN-α on KSHV-infected lymphoma cells.

IMMUNE RESPONSE TO HIV

As detailed above and below, following the initial burst of viremia during primary infection, HIV-infected individuals mount robust immune responses that in most cases substantially curtail the levels of plasma viremia and likely contribute to delaying the ultimate development of clinically apparent disease for a median of 10 years in untreated individuals. This immune response contains elements of both humoral and cell-mediated immunity involving both adaptive and innate immune responses (**Table 197-7**; **Fig. 197-27**). It is directed against multiple antigenic determinants of the HIV virion as well as against viral proteins expressed on the surface of infected cells. Ironically, those CD4+ T cells with T cell receptors specific for HIV are theoretically those CD4+ T cells most likely to be activated—and thus to serve as early targets for productive HIV infection and the cell death or dysfunction associated with infection. Thus, an early consequence of HIV infection is interference with and decrease of the helper T cell population needed to generate an effective immune response.

TABLE 197-7

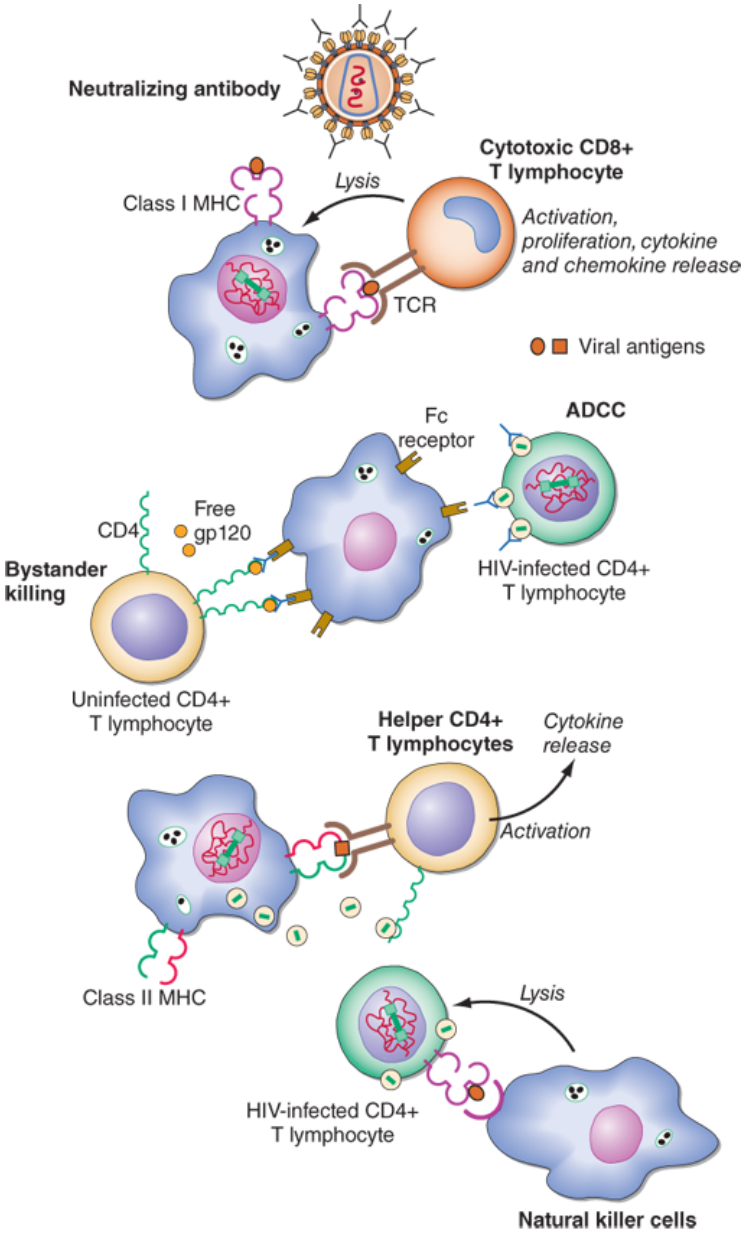
Elements of the Immune Response to HIV

Humoral immunity
Binding antibodies
Neutralizing antibodies
Type specific
Group specific
Broadly neutralizing
Antibodies participating in antibody-dependent cellular cytotoxicity (ADCC)
Protective
Pathogenic (bystander killing)
Enhancing antibodies
Complement
Cell-mediated immunity
Helper CD4+ T lymphocytes
Class I MHC–restricted cytotoxic CD8+ T lymphocytes
CD8+ T cell–mediated inhibition (noncytolytic)
ADCC
Natural killer cells

Abbreviation: MHC, major histocompatibility complex.

FIGURE 197-27

Schematic representation of the different immunologic effector mechanisms thought to be active in the setting of HIV infection. Detailed descriptions are given in the text. ADCC, antibody-dependent cellular cytotoxicity; MHC, major histocompatibility complex; TCR, T cell receptor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although a great deal of investigation has been directed toward delineating and better understanding the components of this immune response, it remains unclear which immunologic effector mechanisms are most important in delaying progression of infection and which, if any, play a role in the pathogenesis of HIV disease. This lack of knowledge has also hampered the ability to develop an effective vaccine for HIV disease.

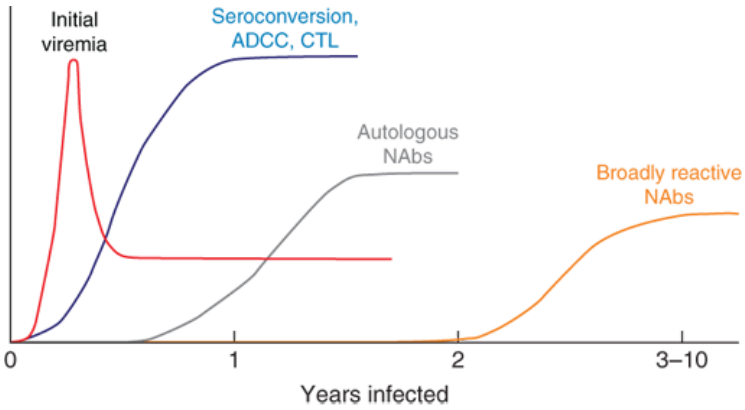
HUMORAL IMMUNE RESPONSE

Antibodies to HIV usually appear within 3–6 weeks and almost invariably within 12 weeks of primary infection **(Fig. 197-28)**; rare exceptions are in individuals who have defects in the ability to produce HIV-specific antibodies. Detection of these antibodies forms the basis of most diagnostic screening tests for HIV infection. The appearance of HIV-binding antibodies detected by ELISA and western blot assays occurs prior to the appearance of neutralizing antibodies; the latter generally appear following the initial decreases in plasma viremia and are more closely related to the appearance of HIV-specific CD8+ T lymphocytes. The first antibodies detected are those directed against the immunodominant region of the envelope gp41, followed by the appearance of antibodies to the structural or gag protein p24 and the gag precursor p55. Antibodies to p24 gag are followed by the appearance of antibodies to the outer envelope glycoprotein (gp120), the gag protein p17, and the products of the *pol* gene (p31 and p66). In addition, one may see antibodies to the low-molecular-weight regulatory proteins encoded by the HIV genes *vpr*, *vpu*, *vif*, *rev*, *tat*, and *nef*. On rare occasion, levels of HIV-specific antibodies may decline during treatment of acute HIV infection.

FIGURE 197-28

Relationship between initial HIV viremia and the development of antibodies to HIV. Within 3 to 6 weeks of initial HIV infection, non-neutralizing antibodies to HIV appear. These antibodies are capable of mediating antibody-dependent cellular cytotoxicity (ADCC). The decline in plasma viremia

generally correlates with the appearance of cytotoxic T lymphocytes (CTL). After approximately 3 months, autologous neutralizing antibodies (NAbs) capable of neutralizing prior circulating strains of HIV appear. After 2 or more years, broadly reactive NAbs appear. *(Adapted from JT Mascola, DC Montefiori: Annu Rev Immunol 28:413, 2010.)*



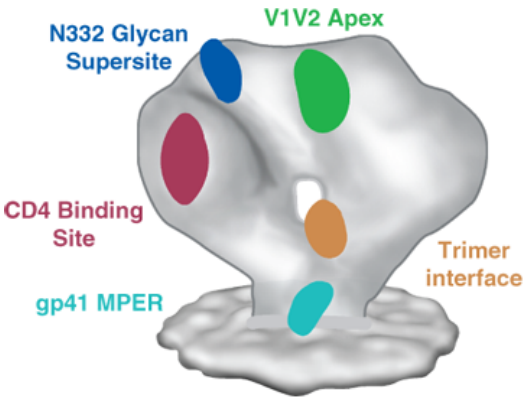
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

While antibodies to multiple antigens of HIV are produced, the precise functional significance of these different antibodies is unclear. The only viral proteins that elicit neutralizing antibodies are the envelope proteins gp120 and gp41. Antibodies directed toward the envelope proteins of HIV have been characterized both as being protective and as possibly contributing to the pathogenesis of HIV disease. Among the protective antibodies are those that function to neutralize HIV directly and prevent the spread of infection to additional cells, as well as those that participate in ADCC. The first neutralizing antibodies are directed against the autologous infecting virus and appear after approximately 12 to 24 weeks of infection. Due to its high rate of mutation the virus is usually able to quickly escape these (and subsequent) neutralizing antibodies. One important mechanism of immune escape is the addition of N-linked glycosylation sites, forming a glycan shield that interferes with envelope recognition by these initial antibodies.

A number of broad and potent HIV-neutralizing envelope-specific antibodies have been isolated from HIV-infected individuals in studies designed to better understand the host response to HIV infection. Approximately 20% of patients develop antibodies capable of neutralizing highly diverse strains. These usually appear 2 or more years following infection in the face of continual viremia. These studies have revealed at least five major sites within the HIV envelope trimer that are able to elicit broadly neutralizing antibodies. These sites include antibodies directed toward the CD4 binding site (CD4bs) of gp120, those binding glycan-dependent epitopes in the V1/V2 region of gp120, those near the base of the V3 region of gp120, those binding to the gp120/gp41 bridge, and those binding to the membrane-proximal region of gp41 **(Fig. 197-29)**. Several of these antibodies contain unique features including high levels of somatic hypermutation, selective germline gene usage (especially for CD4bs antibodies), and long heavy chain complementary determining regions (especially CDRH3). Of note, while these antibodies are broadly neutralizing in vitro, their precise in vivo significance is unclear and the patients from whom they were derived demonstrate evidence of ongoing viral replication unless treated with cART.

FIGURE 197-29

Known targets of broadly neutralizing antibodies against HIV-1. *(Adapted from PD Kwong, JR Mascola: Immunity 37:412, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The other major class of protective antibodies are those that participate in ADCC, a form of cell-mediated immunity **(Chap. 342)** in which NK cells that bear Fc receptors are armed with specific anti-HIV antibodies that bind to the NK cells via their Fc portion. These armed NK cells then bind to and destroy cells expressing HIV antigens. The levels of anti-envelope antibodies capable of mediating ADCC are highest in the earlier stages of HIV infection. Antibodies to both gp120 and gp41 have been shown to participate in ADCC-mediated killing of HIV-infected cells. In vitro, IL-2 can augment ADCC-mediated killing.

In addition to playing a role in host defense, HIV-specific antibodies have also been implicated in disease pathogenesis. Antibodies directed to gp41, when present in low titer, have been shown in vitro to be capable of facilitating infection of cells through an Fc receptor–mediated mechanism known as *antibody enhancement*. Thus, the same regions of the envelope protein of HIV that give rise to antibodies capable of mediating ADCC can also elicit the production of antibodies that can facilitate infection of cells in vitro. In addition, it has been postulated that anti-gp120 antibodies that participate in the ADCC killing of HIV-infected cells might also kill uninfected CD4+ T cells if the uninfected cells had bound free gp120, a phenomenon referred to as *bystander killing*.

One of the most primitive components of the humoral immune system is the complement system **(Chap. 342)**. This element of innate immunity consists of ~30 proteins that are found circulating in blood or associated with cell membranes. While HIV alone is capable of directly activating the complement cascade, the resulting lysis is weak due to the presence of host cell regulatory proteins captured in the virion envelope during budding. It is possible that complement-opsonized HIV virions have increased infectivity in a manner analogous to antibody-mediated enhancement.

CELLULAR IMMUNE RESPONSE

Given that T cell–mediated immunity is known to play a major role in host defense against most viral infections **(Chap. 342)**, it is generally thought to be an important component of the host immune response to HIV. T cell immunity can be divided into two major categories: that mediated by *helper/inducer CD4+ T cells* and that mediated by *cytotoxic/immunoregulatory CD8+ T cells*.

HIV-specific CD4+ T cells can be detected in the majority of HIV-infected patients through the use of flow cytometry to measure intracellular cytokine production in response to MHC class II tetramers pulsed with HIV peptides or through lymphocyte proliferation assays utilizing HIV antigens such as p24. These cells likely play a critical role in the orchestration of the immune response to HIV by providing help to HIV-specific B cells and CD8+ T cells. They may also be capable of directly killing HIV-infected cells. HIV-specific CD4+ T cells may be preferential targets of HIV infection by HIV-infected antigen-presenting cells during the generation of an immune response to HIV (Fig. 197-27). However, they also are likely to undergo clonal expansions in response to HIV antigens and thus survive as a population of cells. No clear correlations exist between levels of HIV-specific CD4+ T lymphocytes and plasma HIV RNA levels; however, in the setting of high viral loads, CD4+ T cell responses to HIV antigens appear to shift from one of proliferation and IL-2 production to one of IFN-γ production. Thus, while a reverse correlation exists between the level of p24-specific proliferation and levels of plasma HIV viremia, the nature of the causal relationship between these parameters is unclear.

MHC class I–restricted, HIV-specific CD8+ T cells have been identified in the peripheral blood of patients with HIV-1 infection. These cells include CTLs that produce perforins and T cells that can be induced by HIV antigens to express an array of cytokines such as IFN-γ, IL-2, MIP-1β, and TNF-α. CTLs have been identified in the peripheral blood of patients within weeks of HIV infection and prior to the appearance of plasma virus. The selective pressure they exert on the evolution of the population of circulating viruses reflects their potential role in control of HIV infection. These CD8+ T lymphocytes, through their HIV-specific antigen receptors, bind to and cause the lytic destruction of target cells bearing autologous MHC class I molecules presenting HIV antigens. Two types of CTL activity can be demonstrated in the peripheral blood or lymph node mononuclear cells of HIV-infected individuals. The first type directly lyses appropriate target cells in culture without prior in vitro stimulation (*spontaneous CTL activity*). The other type of CTL activity reflects the *precursor frequency of CTLs* (CTLp); this type of CTL activity can be demonstrated by stimulation of CD8+ T cells in vitro with a mitogen such as phytohemagglutinin or anti-CD3 antibody.

In addition to CTLs, CD8+ T cells capable of being induced by HIV antigens to express cytokines such as IFN-γ also appear in the setting of HIV-1 infection. It is not clear whether these are the same or different effector pools compared with those cells mediating cytotoxicity; in addition, the relative roles of each in host defense against HIV are not fully understood. It does appear that these CD8+ T cells are driven to in vivo expansion by HIV antigen. There is a direct correlation between levels of CD8+ T cells capable of producing IFN-γ in response to HIV antigens and plasma levels of HIV-1 RNA. Thus, while these cells are clearly induced by HIV-1 infection, their overall ability to control infection remains unclear. Multiple HIV antigens, including Gag, Env, Pol, Tat, Rev, and Nef, can elicit CD8+ T cell responses. Among patients who control viral replication in the absence of antiretroviral drugs are a subset of patients referred to as *elite nonprogressors* (see "Long-Term Survivors, Long-Term Nonprogressors, and Elite Controllers," above) whose peripheral blood contains a population of CD8+ T cells that undergo substantial in vitro proliferation and perforin expression in response to HIV antigens. It is possible that these cells play an important role in HIV-specific host defense.

At least three other forms of cell-mediated immunity to HIV have been described: non-cytolytic CD8+ T cell–mediated suppression of HIV replication, ADCC, and NK cell activity. *Non-cytolytic CD8+ T cell–mediated suppression of HIV replication* refers to the ability of CD8+ T cells from an HIV-infected patient to inhibit the replication of HIV in tissue culture without killing infected targets. There is no requirement for HLA compatibility between the CD8+ T cells and the HIV-infected cells. This effector mechanism is thus nonspecific and appears to be mediated by soluble factor(s) including the CC-chemokines RANTES (CCL5), MIP-1α (CCL3), and MIP-1β (CCL4). These CC-chemokines are potent suppressors of HIV replication and operate at least in part via blockade of the HIV co-receptor (*CCR5*) for R5 (macrophage-tropic) strains of HIV-1 (see above). *ADCC*, as described above in relation to humoral immunity, involves the killing of HIV-expressing cells by NK cells armed with specific antibodies directed against HIV antigens. Finally, *NK cells* alone have been shown to be capable of killing HIV-infected target cells in tissue culture. This primitive cytotoxic mechanism of host defense is directed toward nonspecific surveillance for neoplastic transformation and viral infection through recognition of altered class I MHC molecules.

DIAGNOSIS AND LABORATORY MONITORING OF HIV INFECTION

The establishment of HIV as the causative agent of AIDS and related syndromes early in 1984 was followed by the rapid development of sensitive screening tests for HIV infection. By March 1985, blood donors in the United States were routinely screened for antibodies to HIV. In 1996, blood banks in the United States added the p24 antigen capture assay to the screening process to help identify the rare infected individuals who were donating blood in the time (up to 3 months) between infection and the development of antibodies. In 2002, the ability to detect early infection with HIV was further enhanced by the licensure of nucleic acid testing (NAT) as a routine part of blood donor screening. These refinements decreased the interval between infection and detection (window period) from 22 days for antibody testing to 16 days with p24 antigen testing and subsequently to 12 days with NAT. The development of sensitive assays for monitoring levels of plasma viremia ushered in a new era of being able to monitor the progression of HIV disease more closely. Utilization of these tests, coupled with the measurement of levels of CD4+ T lymphocytes in peripheral blood, is essential in the management of patients with HIV infection.

DIAGNOSIS OF HIV INFECTION

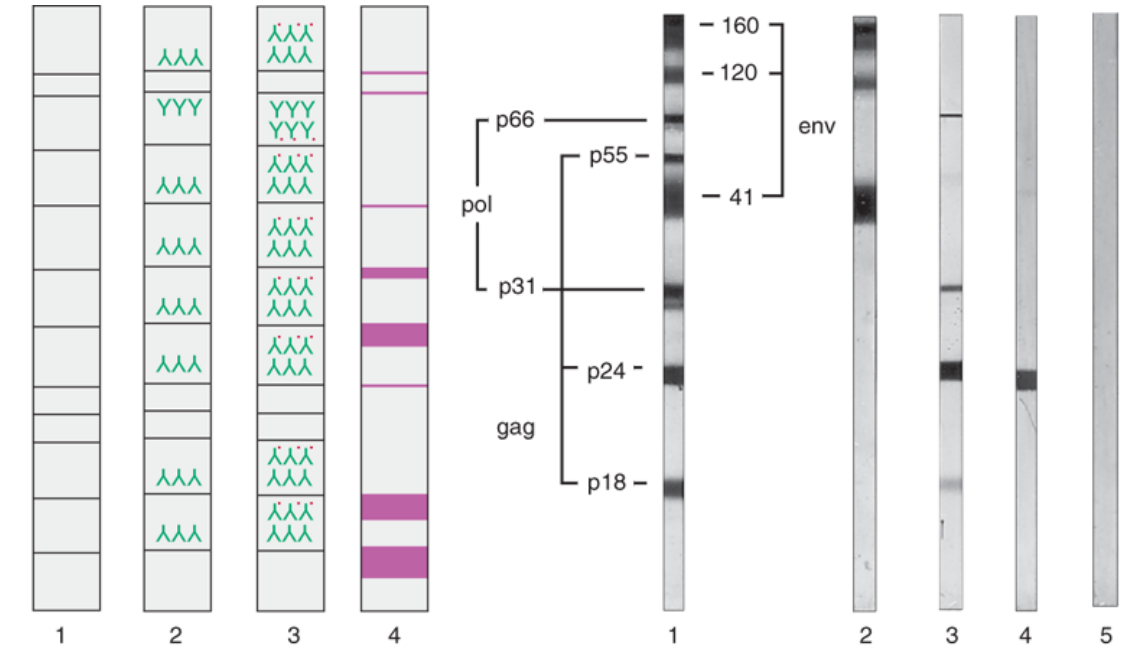
The CDC has recommended that screening for HIV infection be performed as a matter of routine health care. The diagnosis of HIV infection depends on the demonstration of antibodies to HIV and/or the direct detection of HIV or one of its components. As noted above, antibodies to HIV generally appear in the circulation 3–12 weeks following infection.

The standard blood screening tests for HIV infection are based on the detection of antibodies to HIV. A common platform is the ELISA, also referred to as an *enzyme immunoassay* (EIA). This solid-phase assay is an extremely good screening test with a sensitivity of >99.5%. Most diagnostic laboratories use commercial kits that contain antigens from both HIV-1 and HIV-2 and thus are able to detect antibodies to either. These kits use both natural and recombinant antigens and are continuously updated to increase their sensitivity to newly discovered species, such as group O viruses (Fig. 197-1). The fourth-generation EIA tests combine detection of antibodies to HIV with detection of the p24 antigen of HIV. EIA tests are generally scored as positive (highly reactive), negative (nonreactive), or indeterminate (partially reactive). While the EIA is an extremely sensitive test, it is not optimal with regard to specificity. This is particularly true in studies of low-risk individuals, such as volunteer blood donors. In this latter population, only 10% of EIA-positive individuals are subsequently confirmed to have HIV infection. Among the factors associated with false-positive EIA tests are antibodies to class II antigens (such as may be seen following pregnancy, blood transfusion, or transplantation), autoantibodies, hepatic disease, recent influenza vaccination, and acute viral infections. For these reasons, anyone suspected of having HIV infection based on a positive or inconclusive fourth-generation EIA result should have the result confirmed with a more specific assay such as an HIV-1- or HIV-2-specific antibody immunoassay, a western blot, or a plasma HIV RNA level. One can estimate whether an individual has a recent infection with HIV-1 by comparing the results on a standard EIA test that will score positive for all infected individuals with the results on an assay modified to be less sensitive (“detuned assay”) that will score positive for individuals with established HIV infection and negative for individuals with recent infection. In rare instances, an HIV-infected individual treated early in the course of infection may revert to a negative EIA. This does *not* indicate clearing of infection; rather, it signifies levels of ongoing exposure to virus or viral proteins insufficient to maintain a measurable antibody response. When these individuals have discontinued therapy, viruses and antibodies have reappeared.

While current CDC recommendations indicate that a positive fourth-generation assay confirmed by a second HIV-1- or HIV-2-specific immunoassay is adequate for diagnosis, many feel it is prudent to confirm diagnosis with a second platform test such as the western blot or HIV plasma RNA level. The western blot **(Fig. 197-30)** assay takes advantage of the fact that multiple HIV antigens of different, well-characterized molecular weights elicit the production of specific antibodies. These antigens can be separated on the basis of molecular weight, and antibodies to each component can be detected as discrete bands on the western blot. A negative western blot is one in which no bands are present at molecular weights corresponding to HIV gene products. In a patient with a positive or indeterminate EIA and a negative western blot, one can conclude with certainty that the EIA reactivity was a false positive. On the other hand, a western blot demonstrating antibodies to products of all three of the major genes of HIV (*gag, pol*, and *env*) is conclusive evidence of infection with HIV. Criteria established by the FDA in 1993 state that a western blot result is considered positive if antibodies exist to two of the three HIV proteins: p24, gp41, and gp120/160. Using these criteria, ~10% of all blood donors deemed positive for HIV-1 infection lacked an antibody band to the *pol* gene product p31. Some 50% of these blood donors were subsequently found to be false positives. Thus, the absence of the p31 band should increase the suspicion that one may be dealing with a false-positive test result. In this setting it is prudent to obtain additional confirmation with an RNA-based test for HIV-1 and/or a follow-up western blot. By definition, western blot patterns of reactivity that do not fall into the positive or negative categories are considered “indeterminate.” There are two possible explanations for an indeterminate western blot result. The most likely explanation in a low-risk individual is that the patient being tested has antibodies that cross-react with one of the proteins of HIV. The most common patterns of cross-reactivity are antibodies that react with p24 and/or p55. The least likely explanation in this setting is that the individual is infected with HIV and is in the process of mounting a classic antibody response. In either instance, the western blot should be repeated in 1 month to determine whether the indeterminate pattern is a pattern in evolution. In addition, one may attempt to confirm a diagnosis of HIV infection with one of the tests for HIV RNA (discussed below). While the western blot is an excellent confirmatory test for HIV infection in patients with a positive or indeterminate EIA, it is a poor screening test. Among individuals with a negative EIA and PCR for HIV, 20–30% may show one or more bands on western blot. While these bands are usually faint and represent cross-reactivity, their presence creates a situation in which other diagnostic modalities (such as DNA PCR, RT-PCR, or p24 antigen capture) must be employed to ensure that the bands do not indicate early HIV infection.

FIGURE 197-30

Western blot assay for detection of antibodies to HIV. ***A.*** Schematic representation of how a western blot is performed. ***B.*** Examples of patterns of western blot reactivity. In each instance the western blot strip contains antigens to HIV-1. The serum from the patient immunized to the HIV-1 envelope gp160 contains only antibodies to the HIV-1 envelope proteins. The serum from the patient with HIV-2 infection cross-reacts with both *reverse transcriptase* and *gag* gene products of HIV-1.



A

1. Virus digested: digest separated into components by molecular weight
2. Proteins transferred to filter paper: reaction with test serum
3. Enzyme-conjugated antihuman antibody added
4. Substrate added and color noted

B

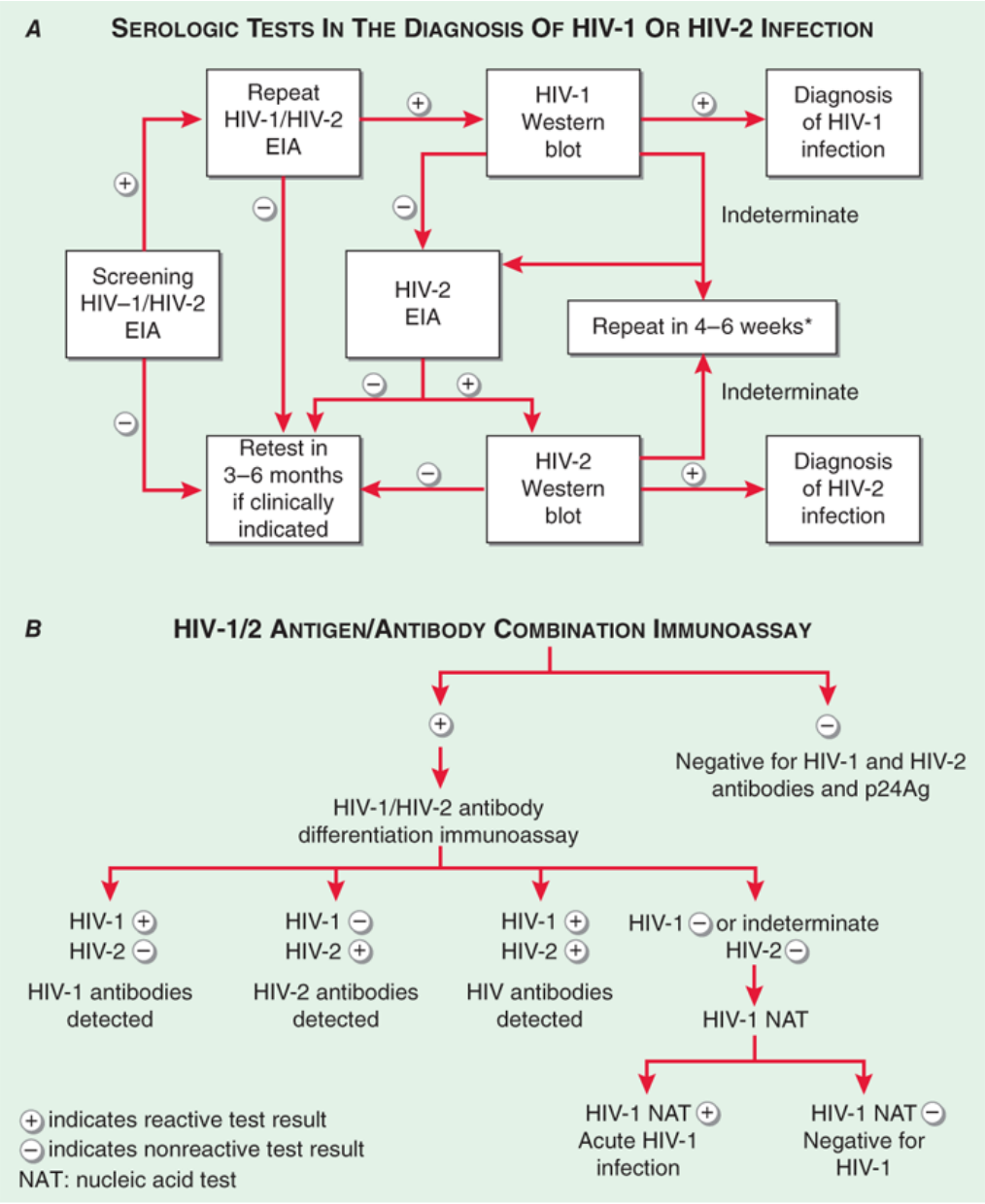
1. Positive HIV-1 infection
2. gp 160 immunization
3. Indeterminate (HIV-2 infection)
4. Indeterminate (cross-reacting antibody to p24)
5. Negative

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A guideline for the use of these serologic tests in attempting to make a diagnosis of HIV infection is depicted in Fig. 197-31. In patients in whom HIV infection is suspected, the appropriate initial test is the EIA. If the result is negative, unless there is strong reason to suspect early HIV infection (as in a patient exposed within the previous 3 months), the diagnosis is ruled out and retesting should be performed only as clinically indicated. If the EIA is indeterminate or positive, the test should be repeated. If the repeat is negative on two occasions, one can assume that the initial positive reading was due to a technical error in the performance of the assay and that the patient is negative. If the repeat is indeterminate or positive, one should proceed to the HIV-1 western blot. If the western blot is positive, the diagnosis is HIV-1 infection. If the western blot is negative, the EIA can be assumed to have been a false positive for HIV-1 and the diagnosis of HIV-1 infection is ruled out. It would also be prudent at this point to perform specific serologic testing for HIV-2 following the same type of algorithm. If the western blot for HIV-1 is indeterminate, it should be repeated in 4–6 weeks; in addition, one may proceed to a specific HIV-1 or HIV-2 antibody differentiation assay, HIV-1 RNA assay, or HIV-1 DNA PCR. If the HIV RNA assays are negative and there is no progression in the western blot, a diagnosis of HIV-1 is ruled out. If either HIV-1 RNA assay is positive and/or the HIV-1 western blot shows progression, a tentative diagnosis of HIV-1 infection can be made and later confirmed with a follow-up western blot demonstrating a positive pattern. In addition to these standard laboratory-based assays for detecting antibodies to HIV, a series of point-of-care tests can provide results in 1–60 min. Among the most popular of these is the OraQuick Rapid HIV-1 antibody test that can be run on blood, plasma, or saliva. The sensitivity and specificity of this test is ~99% when run on whole blood. Specificity remains the same but sensitivity drops to 98% when the test is run on saliva. While negative results from this test are adequate to rule out a diagnosis of HIV infection, a positive finding should be considered preliminary and confirmed with standard serologic testing, as described above. Two rapid test kits are licensed for home use. They are the OraQuick HIV test and the Home Access HIV-1 test system. A positive result with either of these tests should be followed with confirmatory testing by a healthcare professional.

FIGURE 197-31

Serologic tests for the diagnosis of HIV-1 or HIV-2 infection. ***A.*** Algorithm including the use of a western blot. *Stable indeterminate western blot 4–6 weeks later makes HIV infection unlikely. However, it should be repeated twice at 3-month intervals to rule out HIV infection. Alternatively, one may test for HIV-1 p24 antigen or HIV RNA. EIA, enzyme immunoassay. ***B.*** CDC algorithm not including the use of a western blot. *(Adapted from stacks.cdc.gov/view/cdc/23446.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A variety of laboratory tests are available for the direct detection of HIV or its components (Table 197-8). These tests may be of considerable help in making a diagnosis of HIV infection when the antibody determination assays or western blot results are indeterminate. In addition, the tests detecting levels of HIV RNA can be used to determine prognosis and to assess the response to antiretroviral therapies. The simplest, least expensive, and most rarely used of the direct detection tests is the *p24 antigen capture assay*. This is an EIA-type assay in which the solid phase consists of antibodies to the p24 antigen of HIV. It detects the viral protein p24 in the blood of HIV-infected individuals where it exists either as free antigen or complexed to anti-p24 antibodies. Overall, ~30% of individuals with untreated HIV infection have detectable levels of free p24 antigen. This increases to ~50% when samples are treated with a weak acid to dissociate antigen-antibody complexes. Throughout the course of HIV infection, an equilibrium exists between p24 antigen and anti-p24 antibodies. During the first few weeks of infection, before an immune response develops, there is a brisk rise in p24 antigen levels. After the development of anti-p24 antibodies, these levels decline. Late in the course of infection, when circulating levels of virus are high, p24 antigen levels also increase, particularly when detected by techniques involving dissociation of antigen-antibody complexes. The p24 antigen capture assay has its greatest use as a screening test for HIV infection in patients suspected of having the acute HIV syndrome (see below), as high levels of p24 antigen are present prior to the development of antibodies. Its use as a stand-alone test for routine blood donor screening for HIV infection has been replaced by use of NAT or "fourth-generation" assays that combine antigen and antibody testing. The ability to measure and monitor levels of HIV RNA in the plasma of patients with HIV infection has been of extraordinary value in furthering our understanding of the pathogenesis of HIV infection, in monitoring the response to cART, and in providing a diagnostic tool in settings where measurements of anti-HIV antibodies may be misleading, such as in acute infection and neonatal infection. Four assays are predominantly used for this purpose. They are reverse transcriptase PCR (*RT-PCR*; Amplicor and RealTime); branched DNA (*bDNA*; VERSANT); transcription-mediated amplification (*TMA*; APTIMA); and nucleic acid sequence–based amplification (*NASBA*; NucliSENS). These tests are of value in making a diagnosis of HIV infection, in establishing initial prognosis, and in monitoring the effects of

therapy. In addition to the commercially available tests for measuring HIV RNA, *DNA PCR* assays are also employed by research laboratories for making a diagnosis of HIV infection by amplifying HIV proviral DNA from peripheral blood mononuclear cells. The commercially available RNA detection tests have a sensitivity of 40–80 copies of HIV RNA per milliliter of plasma. Research laboratory–based RNA assays can detect as few as one HIV RNA copy per milliliter, while the DNA PCR tests can detect proviral DNA at a frequency of one copy per 10,000–100,000 cells. Thus, these tests are extremely sensitive. One frequent consequence of a high degree of sensitivity is some loss of specificity, and false-positive results have been reported with each of these techniques. For this reason, a positive EIA with a confirmatory western blot or HIV RNA assay remains the "gold standard" for a diagnosis of HIV infection, and the interpretation of other test results must be done with this in mind.

TABLE 197-8

Characteristics of Tests for Direct Detection of HIV

Test	Technique	Sensitivity[a]	Cost/Test[b]
Immune complex–dissociated p24 antigen capture assay	Measurement of levels of HIV-1 core protein in an EIA-based format following dissociation of antigen-antibody complexes by weak acid treatment	Positive in 50% of patients; detects down to 15 pg/mL of p24 protein	$1–2
HIV RNA by PCR	Target amplification of HIV-1 RNA via reverse transcription followed by PCR	Reliable to 40 copies/mL of HIV RNA	$75–150
HIV RNA by bDNA	Measurement of levels of particle-associated HIV RNA in a nucleic acid capture assay employing signal amplification	Reliable to 50 copies/mL of HIV RNA	$75–150
HIV RNA by TMA	Target amplification of HIV-1 RNA via reverse transcription followed by T7 RNA polymerase	Reliable to 100 copies/mL of HIV RNA	$225
HIV RNA by NASBA	Isothermal nucleic acid amplification with internal controls	Reliable to 80 copies/mL of HIV RNA	$75–150

[a]Sensitivity figures refer to those approved by the U.S. Food and Drug Administration. [b]Prices may be lower in large-volume settings.

Abbreviations: bDNA, branched DNA; cDNA, complementary DNA; EIA, enzyme immunoassay; NASBA, nucleic acid sequence–based amplification; PCR, polymerase chain reaction; TMA, transcription-mediated amplification.

In the RT-PCR technique, following DNAse treatment, a cDNA copy is made of all RNA species present in plasma. Because HIV is an RNA virus, this will result in the production of DNA copies of the HIV genome in amounts proportional to the amount of HIV RNA present in plasma. This cDNA is then amplified and characterized using standard PCR techniques, employing primer pairs that can distinguish genomic cDNA from messenger cDNA. The bDNA assay involves the use of a solid-phase nucleic acid capture system and signal amplification through successive nucleic acid hybridizations to detect small quantities of HIV RNA. Both tests can achieve a tenfold increase in sensitivity to 40–50 copies of HIV RNA per milliliter with a preconcentration step in which plasma undergoes ultracentrifugation to pellet the viral particles. In the TMA assay, a cDNA copy of viral RNA is made using primers that contain a promoter sequence for T7 RNA polymerase. T7 polymerase is then added to produce multiple copies of RNA amplicon from the DNA template. It is qualified at 100 copies/mL. The NASBA technique involves the isothermal amplification of a sequence within the gag region of HIV in the presence of internal standards and employs the production of multiple RNA copies through the action of T7-RNA polymerase. The resulting RNA species are quantitated through hybridization with a molecular beacon DNA probe that is quenched in the absence of hybridization. The lower limit of detection for the NucliSENS assay is 80 copies/mL.

In addition to being a diagnostic and prognostic tool, RT-PCR and DNA-PCR are also useful for amplifying defined areas of the HIV genome for sequence analysis and have become an important technique for studies of sequence diversity and microbial resistance to antiretroviral agents. In patients with a positive or indeterminate EIA test and an indeterminate western blot, and in patients in whom serologic testing may be unreliable (such as patients with hypogammaglobulinemia or advanced HIV disease), these tests for quantitating HIV RNA in plasma or detecting proviral DNA in peripheral blood mononuclear cells are valuable tools for making a diagnosis of HIV infection; however, they should be used for diagnosis only when standard serologic testing has failed to provide a definitive result.

LABORATORY MONITORING OF PATIENTS WITH HIV INFECTION

The epidemic of HIV infection and AIDS has provided the clinician with new challenges for integrating clinical and laboratory data to effect optimal patient management. The close relationship between clinical manifestations of HIV infection and CD4+ T cell count has made measurement of CD4+ T

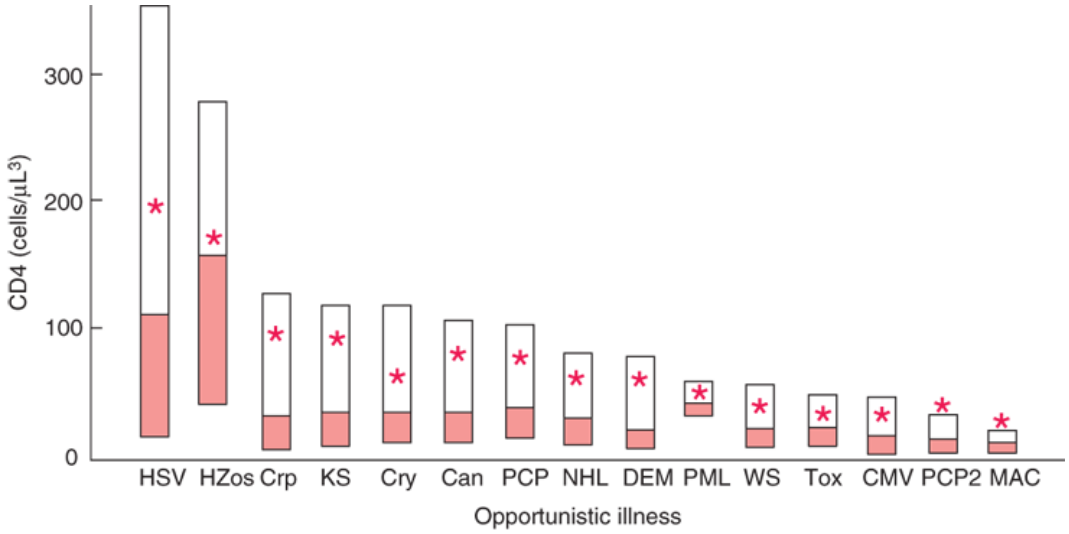
cell numbers a routine part of the evaluation of HIV-infected individuals. The discovery of HIV as the cause of AIDS led to the development of sensitive tests that allow one to monitor the levels of HIV in the blood. Determinations of peripheral blood CD4+ T cell counts and measurements of the plasma levels of HIV RNA provide a powerful set of tools for determining prognosis and monitoring response to therapy.

CD4+ T Cell Counts

The CD4+ T cell count is the laboratory test generally accepted as the best indicator of the immediate state of immunologic competence of the patient with HIV infection. This measurement, which can be made directly or calculated as the product of the percentage of CD4+ T cells (determined by flow cytometry) and the total lymphocyte count (determined by the white blood cell count [WBC] multiplied by the lymphocyte differential percentage), has been shown to correlate very well with the level of immunologic competence. Patients with CD4+ T cell counts <200/µL are at high risk of disease from *P. jirovecii*, while patients with CD4+ T cell counts <50/µL are also at high risk of disease from CMV, mycobacteria of the *M. avium* complex (MAC), and/or *T. gondii* **(Fig. 197-32)**. Once the CD4+ T cell count is <200/µL, patients should be placed on a regimen for *P. jirovecii* prophylaxis, and once the count is <50/µL, primary prophylaxis for MAC infection is indicated. As with any laboratory measurement, one may wish to obtain two determinations prior to any significant changes in patient management based on CD4+ T cell count alone. Patients with HIV infection should have CD4+ T cell measurements performed at the time of diagnosis and every 3–6 months thereafter. More frequent measurements should be made if a declining trend is noted. For patients who have been on cART for at least 2 years with HIV RNA levels persistently <50 copies/mL and CD4 counts >500/µl, the monitoring of the CD4 count is felt by many to be optional. There are a handful of clinical situations in which the CD4+ T cell count may be misleading. Patients with HTLV-1/HIV co-infection may have elevated CD4+ T cell counts that do not accurately reflect their degree of immune competence. In patients with hypersplenism or those who have undergone splenectomy, and in patients receiving medications that suppress the bone marrow such as IFN-α, the CD4+ T cell percentage may be a more reliable indication of immune function than the CD4+ T cell count. A CD4+ T cell percentage of 15 is comparable to a CD4+ T cell count of 200/µL.

FIGURE 197-32

Relationship between CD4+ T cell counts and the development of opportunistic diseases. Boxplot of the median (line inside the box), first quartile (bottom of the box), third quartile (top of the box), and mean (asterisk) CD4+ lymphocyte count at the time of the development of opportunistic disease. Can, candidal esophagitis; CMV, cytomegalovirus infection; Crp, cryptosporidiosis; Cry, cryptococcal meningitis; DEM, AIDS dementia complex; HSV, herpes simplex virus infection; HZos, herpes zoster; KS, Kaposi's sarcoma; MAC, *Mycobacterium avium* complex bacteremia; NHL, non-Hodgkin's lymphoma; PCP, primary *Pneumocystis jirovecii* pneumonia; PCP2, secondary *P. jirovecii* pneumonia; PML, progressive multifocal leukoencephalopathy; Tox, *Toxoplasma gondii* encephalitis; WS, wasting syndrome. *(From RD Moore, RE Chaisson: Ann Intern Med 124:633, 1996.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HIV RNA Determinations

Facilitated by highly sensitive techniques for the precise quantitation of small amounts of nucleic acids, the measurement of serum or plasma levels of HIV RNA has become an essential component in the monitoring of patients with HIV infection. As discussed in "Diagnosis of HIV Infection," above, the most commonly used technique is the RT-PCR assay. This assay generates data in the form of number of copies of HIV RNA per milliliter of serum or plasma and can reliably detect as few as 40 copies of HIV RNA per milliliter of plasma. Research-based assays can detect down to one copy per milliliter. While it is common practice to describe levels of HIV RNA below these cut-offs as "undetectable," this is a term that should be avoided as it is imprecise and leaves the false impression that the level of virus is 0. By utilizing more sensitive, nested PCR techniques and by studying tissue levels of virus as well as plasma levels, HIV RNA can be detected in virtually every patient with HIV infection. The one notable exception to this is a patient who underwent cytoreductive therapy followed by a bone marrow transplant from a *CCR5Δ32* homozygous donor.

Measurements of changes in HIV RNA levels over time have been of great value in delineating the relationship between levels of virus and rates of disease progression (Fig. 197-22), the rates of viral turnover, the relationship between immune system activation and viral replication, and the time to development of drug resistance. HIV RNA measurements are greatly influenced by the state of activation of the immune system and may fluctuate greatly in the setting of secondary infections or immunization. For these reasons, decisions based on HIV RNA levels should never be made on a single determination. Measurements of plasma HIV RNA levels should be made at the time of HIV diagnosis and every 3–6 months thereafter in the untreated patient. Following the initiation of therapy or any change in therapy, plasma HIV RNA levels should be monitored approximately every 4 weeks until the effectiveness of the therapeutic regimen is determined by the development of a new steady-state level of HIV RNA. In most instances of effective antiretroviral therapy the plasma level of HIV RNA will drop to <50 copies/mL within 6 months of the initiation of treatment. During therapy, levels of HIV RNA should be monitored every 3–6 months to evaluate the continuing effectiveness of therapy.

HIV Resistance Testing

The availability of multiple antiretroviral drugs as treatment options has generated a great deal of interest in the potential for measuring the sensitivity of an individual's HIV viral quasispecies to different antiretroviral agents. HIV resistance testing can be done through either genotypic or phenotypic measurements. In the genotypic assays, sequence analyses of the HIV genomes obtained from patients are compared with sequences of viruses with known antiretroviral resistance profiles. In the phenotypic assays, the in vivo growth of viral isolates obtained from the patient is compared with the growth of reference strains of the virus in the presence or absence of different antiretroviral drugs. A modification of this phenotypic approach utilizes a comparison of the enzymatic activities of the reverse transcriptase, protease, or integrase genes obtained by molecular cloning of patients' isolates to the enzymatic activities of genes obtained from reference strains of HIV in the presence or absence of different drugs targeted to these genes. These tests are quite good in identifying those antiretroviral agents that have been utilized in the past and suggesting agents that may be of future value in a given patient. Resistance testing is recommended at the time of initial diagnosis and, if therapy is not initiated at that time, at the time of initiation of cART. Drug resistance testing is also indicated in the setting of virologic failure and should be performed while the patient is still on the failing regimen because of the propensity for the pool of HIV quasispecies to rapidly revert to wild-type in the absence of the selective pressures of cART. In the hands of experts, resistance testing enhances the short-term ability to decrease viral load by ~0.5 log compared with changing drugs merely on the basis of drug history. In addition to the use of resistance testing to help in the selection of new drugs in patients with virologic failure, it may also be of value in selecting an initial regimen for treatment of therapy-naïve individuals. This is particularly true in geographic areas with a high level of background resistance. The patient needs to have an HIV-1 RNA level above 500–1000 copies/mL for an accurate resistance determination. Resistance assays lose their consistency at lower levels of plasma viremia.

Co-Receptor Tropism Assays

Following the licensure of maraviroc as the first CCR5 antagonist for the treatment of HIV infection (see below), it became necessary to be able to determine whether a patient's virus was likely to respond to this treatment. Patients tend to have CCR5-tropic virus early in the course of infection, with a trend toward CXCR4 viruses later in disease. The antiretroviral agent maraviroc is effective only against CCR5-tropic viruses. Because the genotypic determinants of cellular tropism are poorly defined, a phenotypic assay is necessary to determine this property of HIV. Two commercial assays, the Trofile assay (Monogram Biosciences) and the Phenoscript assay (VIRalliance), are available to make this determination. These assays clone the envelope regions of the patient's virus into an indicator virus that is then used to infect target cells expressing either CCR5 or CXCR4 as their co-receptor. These assays take weeks to perform and are expensive. Another, less costly option is to obtain a genotypic assay of the V3 region of HIV-1 and then employ a computer algorithm to predict viral tropism from the sequence. While this approach is less expensive than the classic phenotypic assay, there are fewer data to validate its predictive value.

Other Tests

A variety of other laboratory tests have been studied as potential markers of HIV disease activity. Among these are quantitative culture of replication-competent HIV from plasma, peripheral blood mononuclear cells, or resting memory CD4+ T cells; circulating levels of β_2-microglobulin, soluble IL-2 receptor, IgA, acid-labile endogenous IFN, or TNF-α; and the presence or absence of activation markers such as CD38, HLA-DR, and PD-1 on CD4+ or CD8+ T cells. Nonspecific serologic markers of inflammation and/or coagulation such as IL-6, D-dimer, and sCD14 have been shown to have a high correlation with all-cause mortality **(Table 197-9)**. While these measurements have value as markers of disease activity and help to increase our understanding of the pathogenesis of HIV disease, they do not currently play a major role in the monitoring of patients with HIV infection.

TABLE 197-9

Association Between High-Sensitivity CRP, Il-6, and D-Dimer with All-Cause Mortality in Patients with HIV Infection

	Unadjusted		Adjusted	
Marker	**Odds Ratio (Fourth/First)**	***p***	**Odds Ratio (Fourth/First)**	***p***
Hs-CRP	2.0	.05	2.8	.03
IL-6	8.3	<.0001	11.8	<.0001
D-dimer	12.4	<.0001	26.5	<.0001

Abbreviations: Hs-CRP, high-sensitivity C-reactive protein; IL-6, interleukin 6.

Source: From LH Kuller et al: PLoS Med 5:e203, 2008.

CLINICAL MANIFESTATIONS

The clinical consequences of HIV infection encompass a spectrum ranging from an acute syndrome associated with primary infection to a prolonged asymptomatic state to advanced disease. It is best to regard HIV disease as beginning at the time of primary infection and progressing through various stages. As mentioned above, active virus replication and progressive immunologic impairment occur throughout the course of HIV infection in most patients. With the exception of the rare, true, "elite" virus controllers or long-term nonprogressors (see "Long-Term Survivors, Long-Term Nonprogressors, and Elite Controllers," above), HIV disease in untreated patients inexorably progresses even during the clinically latent stage. Since the mid-1990s, cART has had a major impact on preventing and reversing the progression of disease over extended periods of time in a substantial proportion of adequately treated patients. Today, a person diagnosed with HIV infection and treated with cART has a close to normal life expectancy.

ACUTE HIV INFECTION

It is estimated that 50–70% of individuals with HIV infection experience an acute clinical syndrome ~3–6 weeks after primary infection **(Fig. 197-33)**. Varying degrees of clinical severity have been reported, and although it has been suggested that symptomatic seroconversion leading to the seeking of medical attention indicates an increased risk for an accelerated course of disease, there does not appear to be a correlation between the level of the initial burst of viremia in acute HIV infection and the subsequent course of disease. The typical clinical findings in the acute HIV syndrome are listed in **Table 197-10**; they occur along with a burst of plasma viremia. It has been reported that several symptoms of the acute HIV syndrome (fever, skin rash, pharyngitis, and myalgia) occur less frequently in those infected by injection drug use compared with those infected by sexual contact. The syndrome is typical of an acute viral syndrome and has been likened to acute infectious mononucleosis. Symptoms usually persist for one to several weeks and gradually subside as an immune response to HIV develops and the levels of plasma viremia decrease. Opportunistic infections have been reported during this stage of infection, reflecting the immunodeficiency that results from reduced numbers of CD4+ T cells and likely also from the dysfunction of CD4+ T cells owing to viral protein and endogenous cytokine-induced perturbations of cells (Table 197-5) associated with the extremely high levels of plasma viremia. The Fiebig staging system has been used to describe the different stages of acute HIV infection, ranging from Stage 1 (HIV RNA positive alone) to Stage VI (HIV RNA and full western blot positive). A number of immunologic abnormalities accompany the acute HIV syndrome, including multiphasic perturbations of the numbers of circulating lymphocyte subsets. The number of total lymphocytes and T cell subsets (CD4+ and CD8+) are initially reduced. An inversion of the CD4+/CD8+ T cell ratio occurs later because of a rise in the number of CD8+ T cells. In fact, there may be a selective and transient expansion of CD8+ T cell subsets, as determined by T cell receptor analysis (see above). The total circulating CD8+ T cell count may remain elevated or return to normal; however, CD4+ T cell levels usually remain somewhat depressed, although there may be a slight rebound toward normal. Lymphadenopathy occurs in ~70% of individuals with primary HIV infection. Most patients recover spontaneously from this syndrome and many are left with only a mildly depressed CD4+ T cell count that remains stable for a variable period before beginning its progressive decline; in some individuals, the CD4+ T cell count returns to the normal range. Approximately 10% of patients manifest a fulminant course of immunologic and clinical deterioration after primary infection, even after the disappearance of initial symptoms. In most patients, primary infection with or without the acute syndrome is followed by a prolonged period of clinical latency or smoldering low disease activity.

TABLE 197-10

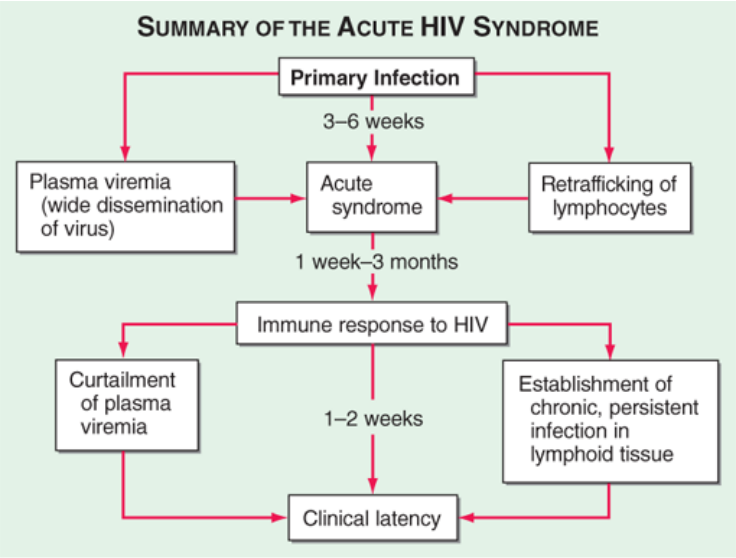
Clinical Findings in the Acute HIV Syndrome

General	Neurologic
Fever	Meningitis
Pharyngitis	Encephalitis
Lymphadenopathy	Peripheral neuropathy
Headache/retroorbital pain	Myelopathy
Arthralgias/myalgias	Dermatologic
Lethargy/malaise	Erythematous maculopapular rash
Anorexia/weight loss	Mucocutaneous ulceration
Nausea/vomiting/diarrhea	

Source: From B Tindall, DA Cooper: AIDS 5:1, 1991.

FIGURE 197-33

The acute HIV syndrome. See text for detailed description. *(Adapted from G Pantaleo et al: N Engl J Med 328:327, 1993. Copyright 1993 Massachusetts Medical Society. All rights reserved.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE ASYMPTOMATIC STAGE—CLINICAL LATENCY

Although the length of time from initial infection to the development of clinical disease varies greatly, the median time for untreated patients is ~10 years. As emphasized above, HIV disease with active virus replication is ongoing and progressive during this asymptomatic period. The rate of disease progression is directly correlated with HIV RNA levels. Patients with high levels of HIV RNA in plasma progress to symptomatic disease faster than do patients with low levels of HIV RNA (Fig. 197-22). Some patients referred to as *long-term nonprogressors* show little if any decline in CD4+ T cell counts over extended periods of time. These patients generally have extremely low levels of HIV RNA; a subset, referred to as *elite nonprogressors*, exhibits HIV RNA levels <50 copies/mL. Certain other patients remain entirely asymptomatic despite the fact that their CD4+ T cell counts show a steady progressive decline to extremely low levels. In these patients, the appearance of an opportunistic disease may be the first manifestation of HIV infection. During the asymptomatic period of HIV infection, the average rate of CD4+ T cell decline is ~50/μL per year in an untreated patient. When the CD4+ T cell count falls

to <200/µL, the resulting state of immunodeficiency is severe enough to place the patient at high risk for opportunistic infections and neoplasms and, hence, for clinically apparent disease.

SYMPTOMATIC DISEASE

Symptoms of HIV disease can appear at any time during the course of HIV infection. Generally speaking, the spectrum of illnesses that one observes changes as the CD4+ T cell count declines. The more severe and life-threatening complications of HIV infection occur in patients with CD4+ T cell counts <200/µL. A diagnosis of AIDS is made in any individual age 6 years and older with HIV infection and a CD4+ T cell count <200/µL (Stage 3, Table 197-2) and in anyone with HIV infection who develops one of the HIV-associated diseases considered to be indicative of a severe defect in cell-mediated immunity (Table 197-1). While the causative agents of the secondary infections are characteristically opportunistic organisms such as *P. jirovecii*, atypical mycobacteria, CMV, and other organisms that do not ordinarily cause disease in the absence of a compromised immune system, they also include several common bacterial and mycobacterial pathogens. Following the widespread use of cART and implementation of guidelines for the prevention of opportunistic infections **(Table 197-11)**, the incidence of these secondary infections has decreased dramatically **(Fig. 197-34)**. Overall, the clinical spectrum of HIV disease is constantly changing as patients live longer and new and better approaches to treatment and prophylaxis are developed. In addition to the classic AIDS-defining illnesses, patients with HIV infection also have an increase in several serious non-AIDS illnesses, including non-AIDS related cancers and cardiovascular, renal, and hepatic disease. Non-AIDS events dominate the disease burden for patients with HIV infection receiving cART (Table 197-4). In developed countries, AIDS-related illnesses are responsible for only ~25% of deaths in patients with HIV infection. A similar percentage of deaths are due to non-AIDS-defining malignancies, with cardiovascular disease and liver disease each accounting for approximately 15% of deaths. The physician providing care to a patient with HIV infection must be well versed in general internal medicine as well as HIV-related opportunistic diseases. In general, it should be stressed that a key element of treatment of symptomatic complications of HIV disease, whether they are primary or secondary, is achieving good control of HIV replication through the use of cART and instituting primary and secondary prophylaxis for opportunistic infections as indicated.

TABLE 197-11

NIH/CDC/IDSA 2013 Guidelines for the Prevention of Opportunistic Infections in Persons Infected with HIV

Pathogen	Indications	First Choice(s)	Alternatives
Recommended as Standard of Care for Primary and Secondary Prophylaxis			
Pneumocystis jirovecii	CD4+ T cell count <200/µL *or* Oropharyngeal candidiasis *or* Prior bout of PCP	Trimethoprim-sulfamethoxazole (TMP-SMX), 1 DS tablet qd PO *or* TMP-SMX, 1 SS tablet qd PO	Dapsone 50 mg bid PO or 100 mg/d PO *or* Dapsone 50 mg/d PO + Pyrimethamine 50 mg/week PO + Leucovorin 25 mg/week PO *or* (Dapsone 200 mg PO + Pyrimethamine 75 mg PO + Leucovorin 25 mg weekly PO) *or* Aerosolized pentamidine, 300 mg via Respirgard II nebulizer every month *or* Atovaquone 1500 mg/d PO *or* TMP-SMX 1 DS tablet 3×/week PO
	May stop prophylaxis if CD4+ T cell count >200/µL for ≥3 months		
Mycobacterium tuberculosis Isoniazid sensitive	Skin test >5 mm *or* Positive IFN-γ release assay *or* Prior positive test without treatment *or* Close contact with case of active pulmonary TB Same with high probability of exposure to drug-resistant TB	(Isoniazid 300 mg PO + Pyridoxine 25 mg PO) qd × 9 months *or* Isoniazid 900 mg PO twice weekly + Pyridoxine 25 mg PO daily × 9 months	Rifabutin (dose adjusted based on cART regimen) or rifampin 600 mg PO qd × 4 months
Drug resistant	Consult local public health authorities		
Mycobacterium-avium complex	CD4+ T cell count <50/µL	Azithromycin 1200 mg weekly PO or 600 mg twice weekly PO *or* Clarithromycin 500 mg bid PO	Rifabutin (dose adjusted based on cART regimen)

Pathogen	Indications	First Choice(s)	Alternatives
	Prior documented disseminated disease	Clarithromycin 500 mg bid PO + Ethambutol 15 (mg/kg)/d PO	Azithromycin 500–600 mg/d PO + Ethambutol 15 (mg/kg)/d PO
	May stop prophylaxis if CD4+ T cell count >100/µL for ≥6 months		
Toxoplasma gondii	TOXO IgG antibody positive and CD4+ T cell count <100/µL	TMP-SMX 1 DS tablet PO qd	TMP-SMX 1 DS 3× weekly PO *or* TMP-SMX, 1 SS PO daily *or* Dapsone 50 mg/d PO + Pyrimethamine 50 mg weekly PO + Leucovorin 25 mg weekly PO *or* (Dapsone 200 mg PO + Pyrimethamine 75 mg PO + Leucovorin 25 mg PO) weekly *or* Atovaquone 1500 mg PO daily ± (Pyrimethamine 25 mg PO + Leucovorin 10 mg PO) daily
	Prior toxoplasmic encephalitis and CD4+ T cell count <200/µL	Sulfadiazine 2000–4000 mg in 2–4 divided doses daily PO + Pyrimethamine 25–50 mg/d PO + Leucovorin 10–25 mg/d PO	Clindamycin 600 mg q8h PO + Pyrimethamine 25–50 mg/d PO + Leucovorin 10–25 mg/d PO *or* TMP-SMX 1 DS tablet bid *or* Atovaquone 750–1500 mg PO bid ± (Pyrimethamine 25 mg/d PO + Leucovorin 10 mg/d PO) or Sulfadiazine 2000–4000 mg/d (in 2–4 divided doses) PO
Toxoplasma gondii	May stop prophylaxis if CD4+ T cell count >200/µL for ≥3 months		

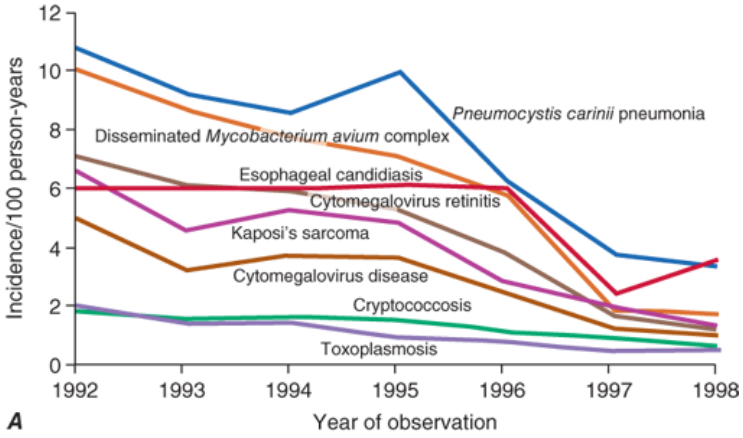
Pathogen	Indications	First Choice(s)	Alternatives
Varicella zoster virus	Significant exposure to chickenpox or shingles in a patient with no history of immunization or prior exposure to either	Varicella zoster immune globulin, IM, within 10 d of exposure (800-843-7477)	Acyclovir 800 mg PO 5 × day for 5–7 days *or* Valacyclovir 1 g PO tid for 5–7 days
Cryptococcus neoformans	Prior documented disease	Fluconazole 200 mg/d PO	Itraconazole 200 mg/d PO
	May stop prophylaxis if CD4+ T cell count >100/µL, no evidence of active fungal infection, and HIV RNA levels <500 copies/mL for >3 months		
Histoplasma capsulatum	Prior documented disease or CD4+ T cell count <150µL and high risk (endemic area or occupational exposure)	Itraconazole 200 mg bid PO	Fluconazole 400 mg/d PO
	May stop prophylaxis after 1 year if CD4+ T cell count >150/µL and patient on cART for ≥6 months		
Coccidioides immitis	Prior documented disease *or* positive serology and CD4+ T cell count <250/µL if from a disease endemic area. (For this indication prophylaxis can be stopped if CD4+ T cell count ≥250 for 6 months.)	Fluconazole 400 mg/d PO	
Penicillium marneffei	Prior documented disease Patients with CD4+T cell counts <100 who live or stay in northern Thailand, Southern China, or Vietnam	Itraconazole 200 mg/d PO	Fluconazole 400 mg PO once weekly
	May stop secondary prophylaxis in patients on ARV therapy with CD4+ T cell count >100/µL for ≥6 months		
Salmonella species	Prior recurrent bacteremia	Ciprofloxacin 500 mg bid PO for ≥6 months	
Bartonella	Prior infection	Doxycycline 200 mg/d PO *or* Azithromycin 1200 mg weekly PO *or* Clarithromycin 500 mg bid PO	
	May stop if CD4+ T cell count >200/µL for >3 months		
Cytomegalovirus	Prior end-organ disease	Valganciclovir 900 mg bid PO	Cidofovir 5 mg/kg every other week IV + Probenecid *or* Foscarnet 90–120 (mg/kg)/d IV
	May stop prophylaxis if CD4+ T cell count >100/µL for 6 months and no evidence of active CMV disease Restart if prior retinitis and CD4+ T cells <100/µL		
Immunizations Generally Recommended			

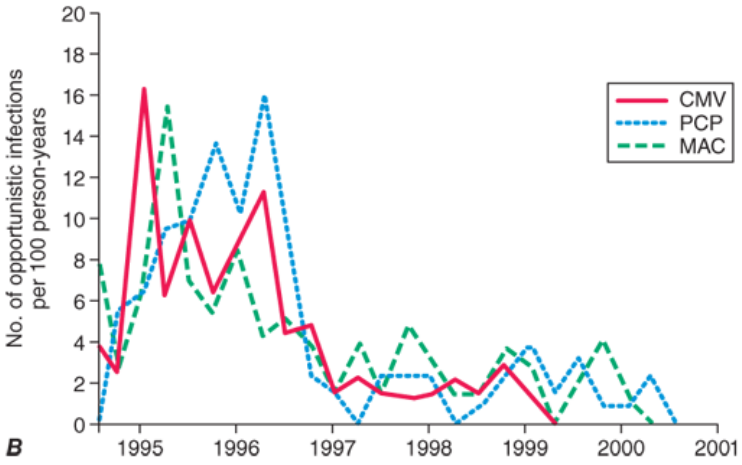
Pathogen	Indications	First Choice(s)	Alternatives
Hepatitis B virus	All susceptible (anti-HBc- and anti-HBs-negative) patients	Hepatitis B vaccine: 3 doses	
Hepatitis A virus	All susceptible (anti-HAV-negative) patients	Hepatitis A vaccine: 2 doses	
Influenza virus	All patients annually	Inactivated trivalent influenza virus vaccine 1 dose yearly	Oseltamivir 75 mg PO qd *or* Rimantadine or amantadine 100 mg PO bid (influenza A only)
Streptococcus pneumoniae	All patients, preferably before CD4+ T cell count ≤200/μL	Pneumococcal conjugated vaccine (13) 0.5 mL IM × 1 followed in 8 weeks or more by pneumococcal polysaccharide vaccine (23) if CD4+ T cell count >200/μL	
	Patients initially immunized at a CD4+ T cell count <100/μL whose CD4+ T cell count then increases to>200/μL	Reimmunize	
Human papillomavirus	All patients 13–26 years of age	HPV vaccine; 3 doses	
Recommended for Prevention of Severe or Frequent Recurrences			
Herpes simplex	Frequent/severe recurrences	Valacyclovir 500 mg bid PO *or* Acyclovir 400 mg bid PO *or* Famciclovir 500 mg bid PO	
Candida	Frequent/severe recurrences	Fluconazole 100–200 mg/d PO	Posaconazole 400 mg bid PO

Abbreviations: ARV, antiretroviral; bid, twice daily; cART, combination antiretroviral therapy; DS, double-strength; IM, intramuscular; PCP, *Pneumocystis jirovecii* pneumonia; PO, by mouth; qd, daily; SS, single-strength; TB, tuberculosis; tid, three times a day.

FIGURE 197-34

A. Decrease in the incidence of opportunistic infections and Kaposi's sarcoma in HIV-infected individuals with CD4+ T cell counts <100/μL from 1992 through 1998. *(Adapted and updated from FJ Palella et al: N Engl J Med 338:853, 1998, and JE Kaplan et al: Clin Infect Dis 30[S1]:S5, 2000, with permission.)* ***B.*** Quarterly incidence rates of cytomegalovirus (CMV), *Pneumocystis jirovecii* pneumonia (PCP), and *Mycobacterium avium* complex (MAC) from 1995 to 2001. *(From FJ Palella et al: AIDS 16:1617, 2002.)*





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Diseases of the Respiratory System

Acute bronchitis and sinusitis are prevalent during all stages of HIV infection. The most severe cases tend to occur in patients with lower CD4+ T cell counts. Sinusitis presents as fever, nasal congestion, and headache. The diagnosis is made by CT or MRI. The maxillary sinuses are most commonly involved; however, disease is also frequently seen in the ethmoid, sphenoid, and frontal sinuses. While some patients may improve without antibiotic therapy, radiographic improvement is quicker and more pronounced in patients who have received antimicrobial therapy. It is postulated that this high incidence of sinusitis results from an increased frequency of infection with encapsulated organisms such as *H. influenzae* and *Streptococcus pneumoniae*. In patients with low CD4+ T cell counts one may see mucormycosis infections of the sinuses. In contrast to the course of this infection in other patient populations, mucormycosis of the sinuses in patients with HIV infection may progress more slowly. In this setting aggressive, frequent local debridement in addition to local and systemic amphotericin B may result in effective treatment.

Pulmonary disease is one of the most frequent complications of HIV infection. The most common manifestation of pulmonary disease is pneumonia. Three of the 10 most common AIDS-defining illnesses are recurrent bacterial pneumonia, tuberculosis, and pneumonia due to the unicellular fungus *P. jirovecii*. Other major causes of pulmonary infiltrates include other mycobacterial infections, other fungal infections, nonspecific interstitial pneumonitis, KS, and lymphoma.

Bacterial pneumonia is seen with an increased frequency in patients with HIV infection, with 0.8–2.0 cases per 100 person-years. Patients with HIV infection are particularly prone to infections with encapsulated organisms. *S. pneumoniae* (Chap. 141) and *H. influenzae* (Chap. 152) are responsible for most cases of bacterial pneumonia in patients with AIDS. This may be a consequence of altered B cell function and/or defects in neutrophil function that may be secondary to HIV disease (see above). Pneumonias due to *S. aureus* (Chap. 142) and *P. aeruginosa* (Chap. 159) also are reported to occur with an increased frequency in patients with HIV infection. *S. pneumoniae* (pneumococcal) infection may be the earliest serious infection to occur in patients with HIV disease. This can present as pneumonia, sinusitis, and/or bacteremia. Patients with untreated HIV infection have a sixfold increase in the incidence of pneumococcal pneumonia and a 100-fold increase in the incidence of pneumococcal bacteremia. Pneumococcal disease may be seen in patients with relatively intact immune systems. In one study, the baseline CD4+ T cell count at the time of a first episode of pneumococcal pneumonia was ~300/μL. Of interest is the fact that the inflammatory response to pneumococcal infection appears proportional to the CD4+ T cell count. Due to this high risk of pneumococcal disease, immunization with the conjugated pneumococcal vaccine followed by booster immunization with the 23-valent pneumococcal polysaccharide vaccine is one of the generally recommended prophylactic measures for patients with HIV infection. This is likely most

effective if given while the CD4+ T cell count is >200/μL and, if given to patients with lower CD4+ T cell counts, should be repeated once the count has been above 200 for 6 months. Although clear guidelines do not exist, it also makes sense to repeat immunization every 5 years. The incidence of bacterial pneumonia is cut in half when patients quit smoking.

Pneumocystis pneumonia (PCP), once the hallmark of AIDS, has dramatically declined in incidence following the development of effective prophylactic regimens and the widespread use of cART. It is, however, still the single most common cause of pneumonia in patients with HIV infection in the United States and can be identified as a likely etiologic agent in 25% of cases of pneumonia in patients with HIV infection, with an incidence in the range of 2–3 cases per 100 person-years. Approximately 30% of cases of HIV-associated PCP occur in patients who are unaware of their HIV status. The risk of PCP is greatest among those who have experienced a previous bout of PCP and those who have CD4+ T cell counts of <200/μL. Overall, 79% of patients with PCP have CD4+ T cell counts <100/μL and 95% of patients have CD4+ T cell counts <200/μL. Recurrent fever, night sweats, thrush, and unexplained weight loss also are associated with an increased incidence of PCP. For these reasons, it is strongly recommended that all patients with CD4+ T cell counts <200/μL (or a CD4 percentage <15) receive some form of PCP prophylaxis. The incidence of PCP is approaching zero in patients with known HIV infection receiving appropriate cART and prophylaxis. In the United States, primary PCP is now occurring at a median CD4+ T cell count of 36/μL, while secondary PCP is occurring at a median CD4+ T cell count of 10/μL. Patients with PCP generally present with fever and a cough that is usually nonproductive or productive of only scant amounts of white sputum. They may complain of a characteristic retrosternal chest pain that is worse on inspiration and is described as sharp or burning. HIV-associated PCP may have an indolent course characterized by weeks of vague symptoms and should be included in the differential diagnosis of fever, pulmonary complaints, or unexplained weight loss in any patient with HIV infection and <200 CD4+ T cells/μL. The most common finding on chest x-ray is either a normal film, if the disease is suspected early, or a faint bilateral interstitial infiltrate. The classic finding of a dense perihilar infiltrate is unusual in patients with AIDS. In patients with PCP who have been receiving aerosolized pentamidine for prophylaxis, one may see an x-ray picture of upper lobe cavitary disease, reminiscent of TB. Other less common findings on chest x-ray include lobar infiltrates and pleural effusions. Thin-section CT may demonstrate a patchy ground-glass appearance. Routine laboratory evaluation is usually of little help in the differential diagnosis of PCP. A mild leukocytosis is common, although this may not be obvious in patients with prior neutropenia. Elevation of lactate dehydrogenase is common. Arterial blood-gases may indicate hypoxemia with a decline in Pa_{O2} and an increase in the arterial-alveolar (a–A) gradient. Arterial blood-gas measurements not only aid in making the diagnosis of PCP but also provide important information for staging the severity of the disease and directing treatment (see below). A definitive diagnosis of PCP requires demonstration of the organism in samples obtained from induced sputum, bronchoalveolar lavage, transbronchial biopsy, or open-lung biopsy. PCR has been used to detect specific DNA sequences for *P. jirovecii* in clinical specimens where histologic examinations have failed to make a diagnosis.

In addition to pneumonia, a number of other clinical problems have been reported in HIV-infected patients as a result of infection with *P. jirovecii*. Otic involvement may be seen as a primary infection, presenting as a polypoid mass involving the external auditory canal. In patients receiving aerosolized pentamidine for prophylaxis against PCP, one may see a variety of extrapulmonary manifestations of *P. jirovecii*. These include ophthalmic lesions of the choroid, a necrotizing vasculitis that resembles Buerger disease, bone marrow hypoplasia, and intestinal obstruction. Other organs that have been involved include lymph nodes, spleen, liver, kidney, pancreas, pericardium, heart, thyroid, and adrenals. Organ infection may be associated with cystic lesions that may appear calcified on CT or ultrasound.

The standard treatment for PCP or disseminated pneumocystosis is trimethoprim-sulfamethoxazole (TMP-SMX). A high (20–85%) incidence of side effects, particularly skin rash and bone marrow suppression, is seen with TMP-SMX in patients with HIV infection. Alternative treatments for mild to moderate PCP include dapsone/trimethoprim, clindamycin/primaquine, and atovaquone. IV pentamidine is the treatment of choice for severe disease in the patient unable to tolerate TMP-SMX. For patients with a Pa_{O2} <70 mmHg or with an a–A gradient >35 mmHg, adjunct glucocorticoid therapy should be used in addition to specific antimicrobials. Overall, treatment should be continued for 21 days and followed by secondary prophylaxis. Prophylaxis for PCP is indicated for any HIV-infected individual who has experienced a prior bout of PCP, any patient with a CD4+ T cell count of <200/μL or a CD4 percentage <15, any patient with unexplained fever for >2 weeks, and any patient with a recent history of oropharyngeal candidiasis. The preferred regimen for prophylaxis is TMP-SMX, one double-strength tablet daily. This regimen also provides protection against toxoplasmosis and some bacterial respiratory pathogens. For patients who cannot tolerate TMP-SMX, alternatives for prophylaxis include dapsone plus pyrimethamine plus leucovorin, aerosolized pentamidine administered by the Respirgard II nebulizer, and atovaquone. Primary or secondary prophylaxis for PCP can be discontinued in those patients treated with cART who maintain good suppression of HIV (<50 copies/mL) and CD4+ T cell counts >200/μL for at least 3 months.

M. tuberculosis, once thought to be on its way to extinction in the United States, experienced a resurgence associated with the HIV epidemic **(Chap. 173)**. Worldwide, approximately one-third of all AIDS-related deaths are associated with TB, and TB is the primary cause of death for 10–15% of patients with HIV infection. In the United States ~5% of AIDS patients have active TB. Patients with HIV infection are more likely to have active TB by a factor of 100 when compared with an HIV-negative population. For an asymptomatic HIV-negative person with a positive purified protein derivative (PPD) skin test, the risk of reactivation TB is around 1% per year. For the patient with untreated HIV infection, a positive PPD skin test, and no signs or symptoms of TB, the rate of reactivation TB is 7–10% per year. Untreated TB can accelerate the course of HIV infection. Levels of plasma HIV RNA increase in the setting of active TB and decline in the setting of successful TB treatment. Active TB is most common in patients 25–44 years of age, in African Americans and Hispanics, in patients in New York City and Miami, and in patients in developing countries. In these demographic groups, 20–70% of the new cases of active TB are in patients with HIV infection. The epidemic of TB embedded in the epidemic of HIV infection probably represents the greatest health risk to the general public and the health care profession associated with the HIV epidemic. In contrast to infection with atypical mycobacteria such as MAC,

active TB often develops relatively early in the course of HIV infection and may be an early clinical sign of HIV disease. In one study, the median CD4+ T cell count at presentation of TB was 326/μL. The clinical manifestations of TB in HIV-infected patients are quite varied and generally show different patterns as a function of the CD4+ T cell count. In patients with relatively high CD4+ T cell counts, the typical pattern of pulmonary reactivation occurs: patients present with fever, cough, dyspnea on exertion, weight loss, night sweats, and a chest x-ray revealing cavitary apical disease of the upper lobes. In patients with lower CD4+ T cell counts, disseminated disease is more common. In these patients the chest x-ray may reveal diffuse or lower-lobe bilateral reticulonodular infiltrates consistent with miliary spread, pleural effusions, and hilar and/or mediastinal adenopathy. Infection may be present in bone, brain, meninges, GI tract, lymph nodes (particularly cervical lymph nodes), and viscera. Some patients with advanced HIV infection and active TB may have no symptoms of illness, and thus screening for TB should be part of the initial evaluation of every patient with HIV infection. Approximately 60–80% of HIV-infected patients with TB have pulmonary disease, and 30–40% have extrapulmonary disease. Respiratory isolation and a negative-pressure room should be used for patients in whom a diagnosis of pulmonary TB is being considered. This approach is critical to limit nosocomial and community spread of infection. Culture of the organism from an involved site provides a definitive diagnosis. Blood cultures are positive in 15% of patients. This figure is higher in patients with lower CD4 +T cell counts. In the setting of fulminant disease one cannot rely on the accuracy of a negative PPD skin test to rule out a diagnosis of TB. In addition, IFN-γ release assays may be difficult to interpret due to high backgrounds as a consequence of HIV-associated immune activation. TB is one of the conditions associated with HIV infection for which cure is possible with appropriate therapy. Therapy for TB is generally the same in the HIV-infected patient as in the HIV-negative patient **(Chap. 173)**. Due to the possibility of multidrug-resistant or extensively drug-resistant TB, drug susceptibility testing should be performed to guide therapy. Due to pharmacokinetic interactions, adjusted doses of rifabutin and/or changes in cART are required when treating TB in the setting of HIV infection. Treatment is most effective in programs that involve directly observed therapy. Initiation of cART and/or anti-TB therapy may be associated with clinical deterioration due to immune reconstitution inflammatory syndrome (IRIS) reactions. These are most common in patients initiating both treatments at the same time, may occur as early as 1 week after initiation of cART therapy, and are seen more frequently in patients with advanced HIV disease. For these reasons it is recommended that initiation of cART be delayed in antiretroviral-naïve patients with CD4 counts >50 cells/μL until 2–4 weeks following the initiation of treatment for TB. For patients with lower CD4 counts the benefits of more immediate cART outweigh the risks of IRIS, and cART should be started as soon as possible in those patients. Effective prevention of active TB can be a reality if the health care professional is aggressive in looking for evidence of latent or active TB by making sure that all patients with HIV infection receive a PPD skin test or evaluation with an IFN-γ release assay. Anergy testing is not of value in this setting. Since these tests rely on the host mounting an immune response to *M. tuberculosis*, patients with CD4+ T cell counts <200 cells/μL should be retested if their CD4+ T cell counts rise to persistently above 200. Patients at risk of continued exposure to TB should be tested annually. HIV-infected individuals with a skin-test reaction of >5 mm, those with a positive IFN-γ release assay, or those who are close household contacts of persons with active TB should receive treatment with 9 months of isoniazid and pyridoxine.

Atypical mycobacterial infections are also seen with an increased frequency in patients with HIV infection. Infections with at least 12 different mycobacteria have been reported, including *M. bovis* and representatives of all four Runyon groups. The most common atypical mycobacterial infection is with *M. avium* or *M. intracellulare* species—the *Mycobacterium avium* complex (MAC). Infections with MAC are seen mainly in patients in the United States and are rare in Africa. It has been suggested that prior infection with *M. tuberculosis* decreases the risk of MAC infection. MAC infections probably arise from organisms that are ubiquitous in the environment, including both soil and water. There is little evidence for person-to-person transmission of MAC infection. The presumed portals of entry are the respiratory and GI tracts. MAC infection is a late complication of HIV infection, occurring predominantly in patients with CD4+ T cell counts of <50/μL. The average CD4+ T cell count at the time of diagnosis is 10/μL. The most common presentation is disseminated disease with fever, weight loss, and night sweats. At least 85% of patients with MAC infection are mycobacteremic, and large numbers of organisms can often be demonstrated on bone marrow biopsy. The chest x-ray is abnormal in ~25% of patients, with the most common pattern being that of a bilateral, lower-lobe infiltrate suggestive of miliary spread. Alveolar or nodular infiltrates and hilar and/or mediastinal adenopathy also can occur. Other clinical findings include endobronchial lesions, abdominal pain, diarrhea, and lymphadenopathy. Anemia and elevated liver alkaline phosphatase are common. The diagnosis is made by the culture of blood or involved tissue. The finding of two consecutive sputum samples positive for MAC is highly suggestive of pulmonary infection. Cultures may take 2 weeks to turn positive. Therapy consists of a macrolide, usually clarithromycin, with ethambutol. Some physicians elect to add a third drug from among rifabutin, ciprofloxacin, or amikacin in patients with extensive disease. Therapy is continued until resolution of clinical signs and symptoms, negative cultures, and CD4+ T cell counts >100/μL for 3–6 months in the setting of cART. Primary prophylaxis for MAC is indicated in patients with HIV infection and CD4+ T cell counts <50/μL (Table 197-11). This may be discontinued in patients in whom cART induces a sustained suppression of viral replication and an increase in CD4+ T cell count to >100/μL for ≥6 months.

Rhodococcus equi is a gram-positive, pleomorphic, acid-fast, non-spore-forming bacillus that can cause pulmonary and/or disseminated infection in patients with advanced HIV infection. Fever and cough are the most common presenting signs. Radiographically one may see cavitary lesions and consolidation. Blood cultures are often positive. Treatment is based on antimicrobial sensitivity testing.

Fungal infections of the lung, in addition to PCP, can be seen in patients with AIDS. Patients with pulmonary cryptococcal disease present with fever, cough, dyspnea, and, in some cases, hemoptysis. A focal or diffuse interstitial infiltrate is seen on chest x-ray in >90% of patients. In addition, one may see lobar disease, cavitary disease, pleural effusions, and hilar or mediastinal adenopathy. More than half of patients are fungemic, and 90% of patients have concomitant CNS infection. *Coccidioides immitis* is a mold that is endemic in the southwest United States. It can cause a reactivation pulmonary syndrome in patients with HIV infection. Most patients with this condition will have CD4+ T cell counts <250/μL. Patients present with fever, weight loss,

cough, and extensive, diffuse reticulonodular infiltrates on chest x-ray. One may also see nodules, cavities, pleural effusions, and hilar adenopathy. While serologic testing is of value in the immunocompetent host, serologies are negative in 25% of HIV-infected patients with coccidioidal infection. Invasive aspergillosis is not an AIDS-defining illness and is generally not seen in patients with AIDS in the absence of neutropenia or administration of glucocorticoids. When it does occur, *Aspergillus* infection may have an unusual presentation in the respiratory tract of patients with AIDS, where it gives the appearance of a pseudomembranous tracheobronchitis. Primary pulmonary infection of the lung may be seen with *histoplasmosis*. The most common pulmonary manifestation of histoplasmosis, however, is in the setting of disseminated disease, presumably due to reactivation. In this setting respiratory symptoms are usually minimal, with cough and dyspnea occurring in 10–30% of patients. The chest x-ray is abnormal in ~50% of patients, showing either a diffuse interstitial infiltrate or diffuse small nodules, and the urine will often be positive for *Histoplasma* antigen.

Two forms of *idiopathic interstitial pneumonia* have been identified in patients with HIV infection: lymphoid interstitial pneumonitis (LIP) and nonspecific interstitial pneumonitis (NIP). LIP, a common finding in children, is seen in about 1% of adult patients with untreated HIV infection. This disorder is characterized by a benign infiltrate of the lung and is thought to be part of the polyclonal activation of lymphocytes seen in the context of HIV and EBV infections. Transbronchial biopsy is diagnostic in 50% of the cases, with an open-lung biopsy required for diagnosis in the remainder of cases. This condition is generally self-limited and no specific treatment is necessary. Severe cases have been managed with brief courses of glucocorticoids. Although rarely a clinical problem since the use of cART, evidence of NIP may be seen in up to half of all patients with untreated HIV infection. Histologically, interstitial infiltrates of lymphocytes and plasma cells in a perivascular and peribronchial distribution are present. When symptomatic, patients present with fever and nonproductive cough occasionally accompanied by mild chest discomfort. Chest x-ray is usually normal or may reveal a faint interstitial pattern. Similar to LIP, NIP is a self-limited process for which no therapy is indicated other than appropriate management of the underlying HIV infection. HIV-related pulmonary arterial hypertension (HIV-PAH) is seen in ~0.5% of HIV-infected individuals. Patients may present with an array of symptoms including shortness of breath, fatigue, syncope, chest pain, and signs of right-sided heart failure. Chest x-ray reveals dilated pulmonary vessels and right-sided cardiomegaly with right ventricular hypertrophy seen on electrocardiogram. cART does not appear to be of clear benefit, and the prognosis is quite poor with a median survival in the range of 2 years.

Neoplastic diseases of the lung including KS and lymphoma are discussed below in the section on neoplastic diseases.

Diseases of the Cardiovascular System

Heart disease is a relatively common postmortem finding in HIV-infected patients (25–75% in autopsy series). The most common form of heart disease is coronary heart disease. In one large series the overall rate of myocardial infarction (MI) was 3.5/1000 patient-years, 28% of these events were fatal, and MI was responsible for 7% of all deaths in the cohort. In patients with HIV infection, cardiovascular disease may be associated with classic risk factors such as smoking, a direct consequence of HIV infection, or a complication of cART. Patients with HIV infection have higher levels of triglycerides, lower levels of high-density lipoprotein cholesterol, and a higher prevalence of smoking than cohorts of individuals without HIV infection. The finding that the rate of cardiovascular disease events was lower in patients on antiretroviral therapy than in those randomized to undergo a treatment interruption identified a clear association between HIV replication and risk of cardiovascular disease. In one study, a baseline CD4+ T cell count of <500/μL was found to be an independent risk factor for cardiovascular disease comparable in magnitude to that attributable to smoking. While the precise pathogenesis of this association remains unclear, it is likely related to the immune activation and increased propensity for coagulation seen as a consequence of HIV replication. Exposure to HIV protease inhibitors and certain reverse transcriptase inhibitors has been associated with increases in total cholesterol and/or risk of MI. Any increases in the risk of death from MI resulting from the use of certain antiretrovirals must be balanced against the marked increases in overall survival brought about by these drugs.

Another form of heart disease associated with HIV infection is a dilated cardiomyopathy associated with congestive heart failure (CHF) referred to as *HIV-associated cardiomyopathy*. This generally occurs as a late complication of HIV infection and, histologically, displays elements of myocarditis. For this reason some have advocated treatment with IV immunoglobulin (IVIg). HIV can be directly demonstrated in cardiac tissue in this setting, and there is debate over whether it plays a direct role in this condition. Patients present with typical findings of CHF including edema and shortness of breath. Patients with HIV infection may also develop cardiomyopathy as side effects of IFN-α or nucleoside analogue therapy. These are reversible once therapy is stopped. KS, cryptococcosis, Chagas' disease, and toxoplasmosis can involve the myocardium, leading to cardiomyopathy. In one series, most patients with HIV infection and a treatable myocarditis were found to have myocarditis associated with toxoplasmosis. Most of these patients also had evidence of CNS toxoplasmosis. Thus, MRI or double-dose contrast CT scan of the brain should be included in the workup of any patient with advanced HIV infection and cardiomyopathy.

A variety of other cardiovascular problems are found in patients with HIV infection. Pericardial effusions may be seen in the setting of advanced HIV infection. Predisposing factors include TB, CHF, mycobacterial infection, cryptococcal infection, pulmonary infection, lymphoma, and KS. While pericarditis is quite rare, in one series 5% of patients with HIV disease had pericardial effusions that were considered to be moderate or severe. Tamponade and death have occurred in association with pericardial KS, presumably owing to acute hemorrhage. Nonbacterial thrombotic endocarditis has been reported and should be considered in patients with unexplained embolic phenomena. IV pentamidine, when given rapidly, can result in hypotension as a consequence of cardiovascular collapse.

Diseases of the Oropharynx and Gastrointestinal System

Oropharyngeal and GI diseases are common features of HIV infection. They are most frequently due to secondary infections. In addition, oral and GI lesions may occur with KS and lymphoma.

Oral lesions, including *thrush*, *hairy leukoplakia*, and *aphthous ulcers* **(Fig. 197-35)**, are particularly common in patients with untreated HIV infection. Thrush, due to *Candida* infection, and oral hairy leukoplakia, presumed due to EBV, are usually indicative of fairly advanced immunologic decline; they generally occur in patients with CD4+ T cell counts of <300/μL. In one study, 59% of patients with oral candidiasis went on to develop AIDS in the next year. Thrush appears as a white, cheesy exudate, often on an erythematous mucosa in the posterior oropharynx. While most commonly seen on the soft palate, early lesions are often found along the gingival border. The diagnosis is made by direct examination of a scraping for pseudohyphal elements. Culturing is of no diagnostic value, as patients with HIV infection may have a positive throat culture for *Candida* in the absence of thrush. Oral hairy leukoplakia presents as white, frondlike lesions, generally along the lateral borders of the tongue and sometimes on the adjacent buccal mucosa (Fig. 197-35). Despite its name, oral hairy leukoplakia is not considered a premalignant condition. Lesions are associated with florid replication of EBV. While usually more disconcerting as a sign of HIV-associated immunodeficiency than a clinical problem in need of treatment, severe cases have been reported to respond to topical podophyllin or systemic therapy with anti-herpesvirus agents. Aphthous ulcers of the posterior oropharynx also are seen with regularity in patients with untreated HIV infection (Fig. 197-35). These lesions are of unknown etiology and can be quite painful and interfere with swallowing. Topical anesthetics provide immediate symptomatic relief of short duration. The fact that thalidomide is an effective treatment for this condition suggests that the pathogenesis may involve the action of tissue-destructive cytokines. Palatal, glossal, or gingival ulcers may also result from cryptococcal disease or histoplasmosis.

FIGURE 197-35

Various oral lesions in HIV-infected individuals. ***A.*** Thrush. ***B.*** Hairy leukoplakia. ***C.*** Aphthous ulcer. ***D.*** Kaposi's sarcoma.



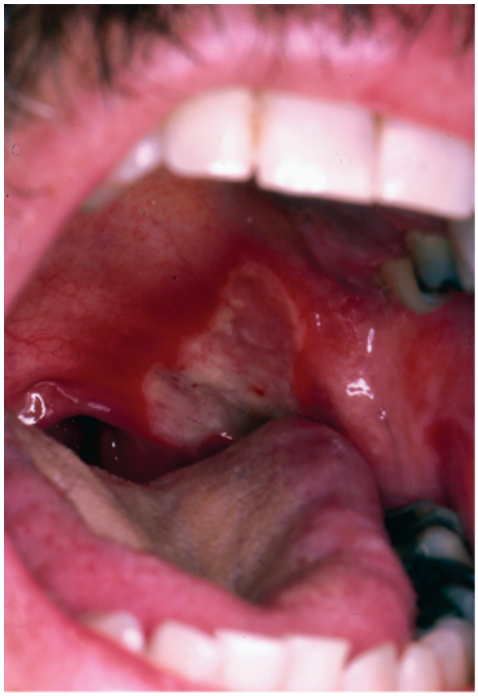
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



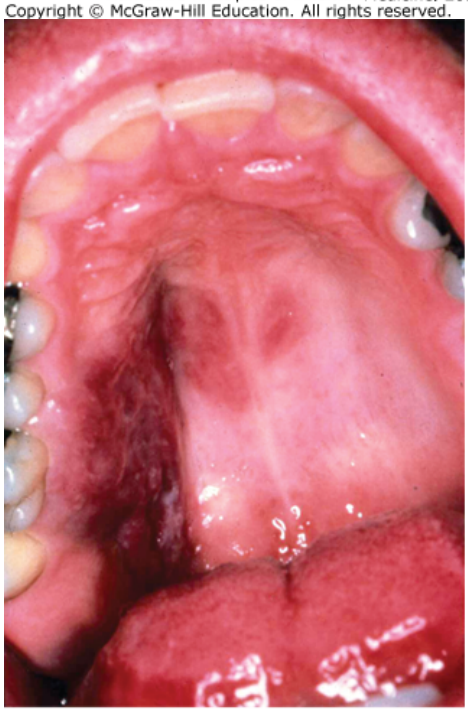
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Esophagitis **(Fig. 197-36)** may present with odynophagia and retrosternal pain. Upper endoscopy is generally required to make an accurate diagnosis. Esophagitis may be due to *Candida*, CMV, or HSV. While CMV tends to be associated with a single large ulcer, HSV infection is more often associated with multiple small ulcers. The esophagus may also be the site of KS and lymphoma. Like the oral mucosa, the esophageal mucosa may have large, painful ulcers of unclear etiology that may respond to thalidomide. While achlorhydria is a common problem in patients with HIV infection, other gastric problems are generally rare. Among the neoplastic conditions involving the stomach are KS and lymphoma.

FIGURE 197-36

Barium swallow of a patient with *Candida* esophagitis. The flow of barium along the mucosal surface is grossly irregular.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Infections of the small and large intestine leading to diarrhea, abdominal pain, and occasionally fever are among the most significant GI problems in HIV-infected patients. They include infections with bacteria, protozoa, and viruses.

Bacteria may be responsible for secondary infections of the GI tract. Infections with enteric pathogens such as *Salmonella*, *Shigella*, and *Campylobacter* are more common in men who have sex with men and are often more severe and more apt to relapse in patients with HIV infection. Patients with untreated HIV have approximately a 20-fold increased risk of infection with *S. typhimurium*. They may present with a variety of nonspecific symptoms including fever, anorexia, fatigue, and malaise of several weeks' duration. Diarrhea is common but may be absent. Diagnosis is made by culture of blood and stool. Long-term therapy with ciprofloxacin is the recommended treatment. HIV-infected patients also have an increased incidence of *S. typhi* infection in areas of the world where typhoid is a problem. *Shigella* spp., particularly *S. flexneri*, can cause severe intestinal disease in HIV-infected individuals. Up to 50% of patients will develop bacteremia. *Campylobacter* infections occur with an increased frequency in patients with HIV infection. While *C. jejuni* is the strain most frequently isolated, infections with many other strains have been reported. Patients usually present with crampy abdominal pain, fever, and bloody diarrhea. Infection may also present as proctitis. Stool examination reveals the presence of fecal leukocytes. Systemic infection can occur, with up to 10% of infected patients exhibiting bacteremia. Most strains are sensitive to erythromycin. Abdominal pain and diarrhea may be seen with MAC infection.

Fungal infections may also be a cause of diarrhea in patients with HIV infection. Histoplasmosis, coccidioidomycosis, and penicilliosis have all been identified as a cause of fever and diarrhea in patients with HIV infection. Peritonitis has been seen with *C. immitis*.

Cryptosporidia, microsporidia, and *Isospora belli* **(Chap. 224)** are the most common opportunistic protozoa that infect the GI tract and cause diarrhea in HIV-infected patients. Cryptosporidial infection may present in a variety of ways, ranging from a self-limited or intermittent diarrheal illness in patients in the early stages of HIV infection to a severe, life-threatening diarrhea in severely immunodeficient individuals. In patients with untreated HIV infection and CD4+ T cell counts of <300/μL, the incidence of cryptosporidiosis is ~1% per year. In 75% of cases the diarrhea is accompanied by crampy abdominal pain, and 25% of patients have nausea and/or vomiting. Cryptosporidia may also cause biliary tract disease in the HIV-infected patient, leading to cholecystitis with or without accompanying cholangitis and pancreatitis secondary to papillary stenosis. The diagnosis of cryptosporidial diarrhea is made by stool examination or biopsy of the small intestine. The diarrhea is noninflammatory, and the characteristic finding is the presence of oocysts that stain with acid-fast dyes. Therapy is predominantly supportive, and marked improvements have been reported in the setting of effective cART. Treatment with up to 2000 mg/d of nitazoxanide (NTZ) is associated with improvement in symptoms or a decrease in shedding of organisms in about half of patients. Its overall role in the management of this condition remains unclear. Patients can minimize their risk of developing cryptosporidiosis by avoiding contact with human and animal feces, by not drinking untreated water from lakes or rivers, and by not eating raw shellfish.

Microsporidia are small, unicellular, obligate intracellular parasites that reside in the cytoplasm of enteric cells **(Chap. 224)**. The main species causing disease in humans is *Enterocytozoon bieneusi*. The clinical manifestations are similar to those described for cryptosporidia and include abdominal pain, malabsorption, diarrhea, and cholangitis. The small size of the organism may make it difficult to detect; however, with the use of chromotrope-based stains, organisms can be identified in stool samples by light microscopy. Definitive diagnosis generally depends on electron-microscopic examination of a stool specimen, intestinal aspirate, or intestinal biopsy specimen. In contrast to cryptosporidia, microsporidia have been noted in a variety of

extraintestinal locations, including the eye, brain, sinuses, muscle, and liver, and they have been associated with conjunctivitis and hepatitis. The most effective way to deal with microsporidia in a patient with HIV infection is to restore the immune system by treating the HIV infection with cART. Albendazole, 400 mg bid, has been reported to be of benefit in some patients.

I. belli is a coccidian parasite **(Chap. 224)** most commonly found as a cause of diarrhea in patients from tropical and subtropical regions. Its cysts appear in the stool as large, acid-fast structures that can be differentiated from those of cryptosporidia on the basis of size, shape, and number of sporocysts. The clinical syndromes of *Isospora* infection are identical to those caused by cryptosporidia. The important distinction is that infection with *Isospora* is generally relatively easy to treat with TMP-SMX. While relapses are common, a thrice-weekly regimen of TMP-SMX appears adequate to prevent recurrence.

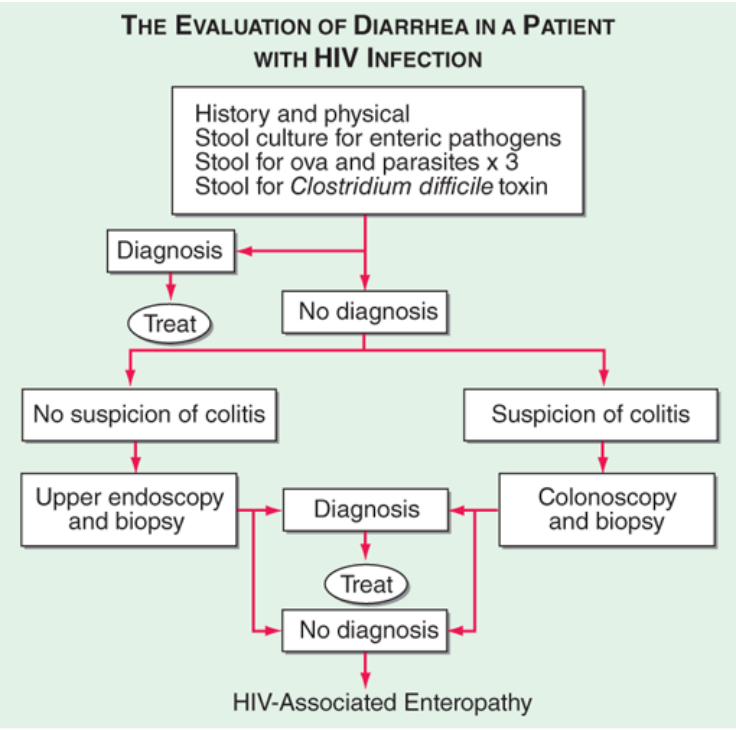
CMV colitis was once seen as a consequence of advanced immunodeficiency in 5–10% of patients with AIDS. It is much less common with the advent of cART. CMV colitis presents as diarrhea, abdominal pain, weight loss, and anorexia. The diarrhea is usually nonbloody, and the diagnosis is achieved through endoscopy and biopsy. Multiple mucosal ulcerations are seen at endoscopy, and biopsies reveal characteristic intranuclear and cytoplasmic inclusion bodies. Secondary bacteremias may result as a consequence of thinning of the bowel wall. Treatment is with either ganciclovir or foscarnet for 3–6 weeks. Relapses are common, and maintenance therapy is typically necessary in patients whose HIV infection is poorly controlled. Patients with CMV disease of the GI tract should be carefully monitored for evidence of CMV retinitis.

In addition to disease caused by specific secondary infections, patients with HIV infection may also experience a chronic diarrheal syndrome for which no etiologic agent other than HIV can be identified. This entity is referred to as *AIDS enteropathy* or *HIV enteropathy*. It is most likely a direct result of HIV infection in the GI tract. Histologic examination of the small bowel in these patients reveals low-grade mucosal atrophy with a decrease in mitotic figures, suggesting a hyporegenerative state. Patients often have decreased or absent small-bowel lactase and malabsorption with accompanying weight loss.

The initial evaluation of a patient with HIV infection and diarrhea should include a set of stool examinations, including culture, examination for ova and parasites, and examination for *Clostridium difficile* toxin. Approximately 50% of the time this workup will demonstrate infection with pathogenic bacteria, mycobacteria, or protozoa. If the initial stool examinations are negative, additional evaluation, including upper and/or lower endoscopy with biopsy, will yield a diagnosis of microsporidial or mycobacterial infection of the small intestine ~30% of the time. In patients for whom this diagnostic evaluation is nonrevealing, a presumptive diagnosis of HIV enteropathy can be made if the diarrhea has persisted for >1 month. An algorithm for the evaluation of diarrhea in patients with HIV infection is given in **Fig. 197-37**.

FIGURE 197-37

Algorithm for the evaluation of diarrhea in a patient with HIV infection. HIV-associated enteropathy is a diagnosis of exclusion and can be made only after other, generally treatable, forms of diarrheal illness have been ruled out.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Rectal lesions are common in HIV-infected patients, particularly the perirectal ulcers and erosions due to the reactivation of HSV **(Fig. 197-38)**. These lesions may appear quite atypical, as denuded skin without vesicles. They typically respond well to treatment with valacyclovir, famciclovir, or foscarnet. Other rectal lesions encountered in patients with HIV infection include condylomata acuminata, KS, and intraepithelial neoplasia (see below).

FIGURE 197-38

Severe, erosive perirectal herpes simplex in a patient with AIDS.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Hepatobiliary Diseases

Diseases of the hepatobiliary system are a major problem in patients with HIV infection. It has been estimated that approximately 15% of the deaths of patients with HIV infection are related to liver disease. While this is predominantly a reflection of the problems encountered in the setting of co-infection with hepatitis B or C, it is also a reflection of the hepatic injury, ranging from hepatic steatosis to hypersensitivity reactions to immune reconstitution, that can be seen in the context of cART.

The prevalence of co-infection with HIV and hepatitis viruses varies by geographic region. In the United States, ~90% of HIV-infected individuals have evidence of infection with HBV; 6–14% have chronic HBV infection; 5–50% of patients are co-infected with HCV; and co-infections with hepatitis D, E, and/or G viruses are common. Among IV drug users with HIV infection, rates of HCV infection range from 70% to 95%. HIV infection has a significant impact on the course of hepatitis virus infection. It is associated with approximately a threefold increase in the development of persistent hepatitis B surface antigenemia. Patients infected with both HBV and HIV have decreased evidence of inflammatory liver disease. The presumption that this is due to the immunosuppressive effects of HIV infection is supported by the observations that this situation can be reversed, and one may see the development of more severe hepatitis following the initiation of effective cART. In studies of the impact of HIV on HBV infection, four- to tenfold increases in liver-related mortality rates have been noted in patients with HIV and active HBV infection compared to rates in patients with either infection alone. There is, however, only a slight increase in overall mortality rate in HIV-infected individuals who are also hepatitis B surface antigen (HBsAg)–positive. IFN-α is less successful as treatment for HBV in patients with HIV co-infection. Lamivudine, emtricitabine, adefovir/tenofovir/entecavir, and telbivudine alone or in combination are useful in the treatment of hepatitis B in patients with HIV infection. It is important to remember that all the above-mentioned drugs also have activity against HIV and should not be used alone in patients with HIV infection, in order to avoid the emergence of quasispecies of HIV resistant to these drugs. For this reason, the treatment of hepatitis B infection in a patient with HIV infection should always be done in the setting of cART. HCV infection is more severe in the patient with HIV infection; it does not appear to affect overall mortality rates in HIV-infected individuals when other variables such as age, baseline CD4+ T cell count, and use of cART are taken into account. In the setting of HIV and HCV co-infection, levels of HCV are approximately tenfold higher than in the HIV-negative patient with HCV infection. There is a 50% higher overall mortality rate with a five-fold increased risk of death due to liver disease in patients chronically infected with both HCV and HIV. Use of directly acting agents for the treatment of HCV leads to cure rates approaching 100%, even in patients with HIV co-infection. Successful treatment of HCV in HIV-infected patients

decreases mortality. Hepatitis A virus infection is not seen with an increased frequency in patients with HIV infection. It is recommended that all patients with HIV infection who have not experienced natural infection be immunized with hepatitis A and/or hepatitis B vaccines. Infection with hepatitis G virus, also known as GB virus C, is seen in ~50% of patients with HIV infection. For reasons that are currently unclear, there are data to suggest that patients with HIV infection co-infected with this virus have a decreased rate of progression to AIDS.

A variety of other infections also may involve the liver. Granulomatous hepatitis may be seen as a consequence of mycobacterial or fungal infections, particularly MAC infection. Hepatic masses may be seen in the context of TB, peliosis hepatis, or fungal infection. Among the fungal opportunistic infections, *C. immitis* and *Histoplasma capsulatum* are those most likely to involve the liver. Biliary tract disease in the form of papillary stenosis or sclerosing cholangitis has been reported in the context of cryptosporidiosis, CMV infection, and KS. When no diagnosis can be made, the term *AIDS cholangiopathy* is used. Hemophagocytic lymphohistiocytosis of the liver has been seen in the setting of Hodgkin's disease and may occur prior to diagnosis of the underlying neoplasm.

Many of the drugs used to treat HIV infection are metabolized by the liver and can cause liver injury. Fatal hepatic reactions have been reported with a wide array of antiretrovirals including nucleoside analogues, nonnucleoside analogues, and protease inhibitors. Nucleoside analogues work by inhibiting DNA synthesis. This can result in toxicity to mitochondria, which can lead to disturbances in oxidative metabolism. This may manifest as hepatic steatosis and, in severe cases, lactic acidosis and fulminant liver failure. It is important to be aware of this condition and to watch for it in patients with HIV infection receiving nucleoside analogues. It is reversible if diagnosed early and the offending agent(s) discontinued. Nevirapine has been associated with at times fatal fulminant and cholestatic hepatitis, hepatic necrosis, and hepatic failure. Indinavir may cause mild to moderate elevations in serum bilirubin in 10–15% of patients in a syndrome similar to Gilbert's syndrome. A similar pattern of hepatic injury may be seen with atazanavir. In the patient receiving cART with an unexplained increase in hepatic transaminases, strong consideration should be given to drug toxicity.

Pancreatic injury is most commonly a consequence of drug toxicity, notably that secondary to pentamidine or dideoxynucleosides. While up to half of patients in some series have biochemical evidence of pancreatic injury, <5% of patients show any clinical evidence of pancreatitis that is not linked to a drug toxicity.

Diseases of the Kidney and Genitourinary Tract

Diseases of the kidney or genitourinary tract may be a direct consequence of HIV infection, due to an opportunistic infection or neoplasm, or related to drug toxicity. Overall, microalbuminuria is seen in ~20% of untreated HIV-infected patients; significant proteinuria is seen in closer to 2%. The presence of microalbuminuria has been associated with an increase in all-cause mortality rate. *HIV-associated nephropathy* (HIVAN) was first described in IDUs and was initially thought to be IDU nephropathy in patients with HIV infection; it is now recognized as a true direct complication of HIV infection. Although the majority of patients with this condition have CD4+ T cell counts <200/μL, HIV-associated nephropathy can be an early manifestation of HIV infection and is also seen in children. Over 90% of reported cases have been in African-American or Hispanic individuals; the disease is not only more prevalent in these populations but also more severe and is the third leading cause of end-stage renal failure among African Americans age 20–64 in the United States. Proteinuria is the hallmark of this disorder. Edema and hypertension are rare. Ultrasound examination reveals enlarged, hyperechogenic kidneys. A definitive diagnosis is obtained through renal biopsy. Histologically, focal segmental glomerulosclerosis is present in 80%, and mesangial proliferation in 10–15% of cases. Prior to effective antiretroviral therapy, this disease was characterized by relatively rapid progression to end-stage renal disease. Patients with HIV-associated nephropathy should be treated for their HIV infection. Treatment with angiotensin-converting enzyme (ACE) inhibitors and/or prednisone, 60 mg/d, also has been reported to be of benefit in some cases. The incidence of this disease in patients receiving adequate cART has not been well defined; however, the impression is that it has decreased in frequency and severity. It is the leading cause of end-stage renal disease in patients with HIV infection.

Among the drugs commonly associated with renal damage in patients with HIV disease are pentamidine, amphotericin, adefovir, cidofovir, tenofovir, and foscarnet. TMP-SMX may compete for tubular secretion with creatinine and cause an increase in the serum creatinine level. The pharmacokinetic booster cobicistat, a component of several fixed-drug cART formulations, inhibits renal tubular secretion of creatinine leading to increased serum creatinine levels without a true decline in glomerular filtration rate. Sulfadiazine may crystallize in the kidney and result in an easily reversible form of renal shutdown, while indinavir or atazanavir may form renal calculi. Adequate hydration is the mainstay of treatment and prevention for these latter two conditions.

Genitourinary tract infections are seen with a high frequency in patients with HIV infection; they present with skin lesions, dysuria, hematuria, and/or pyuria and are managed in the same fashion as in patients without HIV infection. Infections with HSV are covered below ("Dermatologic Diseases"). Infections with *T. pallidum*, the etiologic agent of *syphilis*, play an important role in the HIV epidemic. In HIV-negative individuals, genital syphilitic ulcers as well as the ulcers of chancroid are major predisposing factors for heterosexual transmission of HIV infection. While most HIV-infected individuals with syphilis have a typical presentation, a variety of formerly rare clinical problems may be encountered in the setting of dual infection. Among them are *lues maligna*, an ulcerating lesion of the skin due to a necrotizing vasculitis; unexplained fever; nephrotic syndrome; and neurosyphilis. The most common presentation of syphilis in the HIV-infected patient is that of *condylomata lata*, a form of secondary syphilis. Neurosyphilis may be asymptomatic or may present as acute meningitis, neuroretinitis, deafness, or stroke. The rate of neurosyphilis may be as high as 1% in patients with HIV infection, and one should consider a lumbar puncture to look for neurosyphilis in all patients with HIV infection and secondary syphilis. As a

consequence of the immunologic abnormalities seen in the setting of HIV infection, diagnosis of syphilis through standard serologic testing may be challenging. On the one hand, a significant number of patients have false-positive Venereal Disease Research Laboratory (VDRL) tests due to polyclonal B cell activation. On the other hand, the development of a new positive VDRL may be delayed in patients with new infections, and the anti–fluorescent treponemal antibody (anti-FTA) test may be negative due to immunodeficiency. Thus, dark-field examination of appropriate specimens should be performed in any patient in whom syphilis is suspected, even if the patient has a negative VDRL. Similarly, any patient with a positive serum VDRL test, neurologic findings, and an abnormal spinal fluid examination should be considered to have neurosyphilis and treated accordingly, regardless of the CSF VDRL result. In any setting, patients treated for syphilis need to be carefully monitored to ensure adequate therapy. Approximately one-third of patients with HIV infection will experience a Jarisch-Herxheimer reaction upon initiation of therapy for syphilis.

Vulvovaginal candidiasis is a common problem in women with HIV infection. Symptoms include pruritus, discomfort, dyspareunia, and dysuria. Vulvar infection may present as a morbilliform rash that may extend to the thighs. Vaginal infection is usually associated with a white discharge, and plaques may be seen along an erythematous vaginal wall. Diagnosis is made by microscopic examination of the discharge for pseudohyphal elements in a 10% potassium hydroxide solution. Mild disease can be treated with topical therapy. More serious disease can be treated with fluconazole. Other causes of vaginitis include *Trichomonas* and mixed bacteria.

Diseases of the Endocrine System and Metabolic Disorders

A variety of endocrine and metabolic disorders are seen in the context of HIV infection. These may be a direct consequence of HIV infection, secondary to opportunistic infections or neoplasms, or related to medication side effects. Between 33% and 75% of patients with HIV infection receiving thymidine analogues or protease inhibitors as a component of cART develop a syndrome often referred to as *lipodystrophy*, consisting of elevations in plasma triglycerides, total cholesterol, and apolipoprotein B, as well as hyperinsulinemia and hyperglycemia. Many of the patients have been noted to have a characteristic set of body habitus changes associated with fat redistribution, consisting of truncal obesity coupled with peripheral wasting **(Fig. 197-39)**. Truncal obesity is apparent as an increase in abdominal girth related to increases in mesenteric fat, a dorsocervical fat pad ("buffalo hump") reminiscent of patients with Cushing's syndrome, and enlargement of the breasts. The peripheral wasting, or lipoatrophy, is particularly noticeable in the face and buttocks and by the prominence of the veins in the legs. These changes may develop at any time ranging from ~6 weeks to several years following the initiation of cART. Approximately 20% of the patients with HIV-associated lipodystrophy meet the criteria for the *metabolic syndrome* as defined by The International Diabetes Federation or The U.S. National Cholesterol Education Program Adult Treatment Panel III. The lipodystrophy syndrome has been reported in association with regimens containing a variety of different drugs, and while initially reported in the setting of protease inhibitor therapy, it appears that similar changes can also be induced by protease-sparing regimens. It has been suggested that the lipoatrophy changes are particularly severe in patients receiving the thymidine analogues stavudine and zidovudine. Current treatment guidelines avoid these drugs and recommend drugs with fewer of these side effects. National Cholesterol Education Program (NCEP) guidelines should be followed in the management of these lipid abnormalities **(Chap. 400)**, and consideration should be given to changing the components of cART with avoidance of thymidine analogues (azidothymidine and stavudine) and offending protease inhibitors. Due to concerns regarding drug interactions, the most commonly utilized lipid-lowering agents in this setting are gemfibrozil and atorvastatin. In addition, lactic acidosis is associated with cART. This is most commonly seen with nucleoside analogue reverse transcriptase inhibitors and can be fatal.

FIGURE 197-39

Characteristics of lipodystrophy. ***A.*** Truncal obesity and buffalo hump. ***B.*** Facial wasting. ***C.*** Accumulation of intraabdominal fat on CT scan.



A B



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Patients with advanced HIV disease may develop hyponatremia due to the syndrome of inappropriate antidiuretic hormone (vasopressin) secretion (SIADH) as a consequence of increased free-water intake and decreased free-water excretion. SIADH is usually seen in conjunction with pulmonary or CNS disease. Low serum sodium may also be due to adrenal insufficiency; a concomitant high serum potassium should alert one to this possibility. Hyperkalemia may be secondary to adrenal insufficiency; HIV nephropathy; or medications, particularly trimethoprim and pentamidine. Hypokalemia may be seen in the setting of tenofovir or amphotericin therapy. Adrenal gland disease may be due to mycobacterial infections, CMV disease, cryptococcal disease, histoplasmosis, or ketoconazole toxicity. Iatrogenic Cushing's syndrome with suppression of the hypothalamic-pituitary-adrenal axis may be seen with the use of local glucocorticoids (injected or inhaled) in patients receiving ritonavir. This is due to inhibition of the hepatic enzyme CYP3A4 by ritonavir leading to prolongation of the glucocorticoid half-life.

Thyroid function may be altered in 10–15% of patients with HIV infection. Both hypo- and hyperthyroidism may be seen. The predominant abnormality is subclinical hypothyroidism. In the setting of cART, up to 10% of patients have been noted to have elevated thyroid-stimulating hormone levels, suggesting that this may be a manifestation of immune reconstitution. Immune-reconstitution Graves' disease may occur as a late (9–48 months) complication of cART. In advanced HIV disease, infection of the thyroid gland may occur with opportunistic pathogens, including *P. jirovecii*, CMV, mycobacteria, *Toxoplasma gondii*, and *Cryptococcus neoformans*. These infections are generally associated with a nontender, diffuse enlargement of the thyroid gland. Thyroid function is usually normal. Diagnosis is made by fine-needle aspirate or open biopsy.

Depending on the severity of disease, HIV infection is associated with *hypogonadism* in 20–50% of men. While this is generally a complication of underlying illness, testicular dysfunction may also be a side effect of ganciclovir therapy. In some surveys, up to two-thirds of patients report decreased libido and one-third complain of erectile dysfunction. Androgen-replacement therapy should be considered in patients with symptomatic hypogonadism. HIV infection does not seem to have a significant effect on the menstrual cycle outside the setting of advanced disease.

Immunologic and Rheumatologic Diseases

Immunologic and rheumatologic disorders are common in patients with HIV infection and range from excessive immediate-type hypersensitivity reactions (Chap. 347) to an increase in the incidence of reactive arthritis (Chap. 355) to conditions characterized by a diffuse infiltrative lymphocytosis. The occurrence of these phenomena is an apparent paradox in the setting of the profound immunodeficiency and immunosuppression that characterizes HIV infection and reflects the complex nature of the immune system and its regulatory mechanisms.

Drug allergies are the most significant allergic reactions occurring in HIV-infected patients and appear to become more common as the disease progresses. They occur in up to 65% of patients who receive therapy with TMP-SMX for PCP. In general, these drug reactions are characterized by erythematous, morbilliform eruptions that are pruritic, tend to coalesce, and are often associated with fever. Nonetheless, ~33% of patients can be maintained on the offending therapy, and thus these reactions are not an immediate indication to stop the drug. Anaphylaxis is extremely rare in patients with HIV infection, and patients who have a cutaneous reaction during a single course of therapy can still be considered candidates for future treatment or prophylaxis with the same agent. The one exception to this is the nucleoside analogue abacavir, where fatal hypersensitivity reactions have been reported with rechallenge. This hypersensitivity is strongly associated with the HLA-B5701 haplotype, and a hypersensitivity reaction to abacavir is an absolute contraindication to future therapy. For other agents, including TMP-SMX, desensitization regimens are moderately successful. While the mechanisms underlying these allergic-type reactions remain unknown, patients with HIV infection have been noted to have elevated IgE levels that increase as the CD4+ T cell count declines. The numerous examples of patients with multiple drug reactions suggest that a common pathway is involved.

HIV infection shares many similarities with a variety of autoimmune diseases, including a substantial polyclonal B cell activation that is associated with a high incidence of antiphospholipid antibodies, such as anticardiolipin antibodies, VDRL antibodies, and lupus-like anticoagulants. In addition, HIV-infected individuals have an increased incidence of antinuclear antibodies. Despite these serologic findings, there is no evidence that HIV-infected individuals have an increase in two of the more common autoimmune diseases, i.e., systemic lupus erythematosus and rheumatoid arthritis. In fact, it has been observed that these diseases may be somewhat ameliorated by the concomitant presence of HIV infection, suggesting that an intact CD4+ T cell limb of the immune response plays an integral role in the pathogenesis of these conditions. Similarly, there are anecdotal reports of patients with common variable immunodeficiency (Chap. 344), characterized by hypogammaglobulinemia, who have had a normalization of Ig levels following the development of HIV infection, suggesting a possible role for overactive CD4+ T cell immunity in certain forms of that syndrome. The one autoimmune disease that may occur with an increased frequency in patients with HIV infection is a variant of primary Sjögren's syndrome (Chap. 354). Patients with HIV infection may develop a syndrome consisting of parotid gland enlargement, dry eyes, and dry mouth that is associated with lymphocytic infiltrates of the salivary gland and lung. One also can see peripheral neuropathy, polymyositis, renal tubular acidosis, and hepatitis. In contrast to Sjögren's syndrome, in which the lymphocytic infiltrates are composed predominantly of CD4+ T cells, in patients with HIV infection the infiltrates are composed predominantly of CD8+ T cells. In addition, while patients with Sjögren's syndrome are mainly women who have autoantibodies to Ro and La and who frequently have HLA-DR3 or B8 MHC haplotypes, HIV-infected individuals with this syndrome are usually African-American men who do not have anti-Ro or anti-La and who most often are HLA-DR5. This syndrome appears to be less common with the increased use of effective cART. The term *diffuse infiltrative lymphocytosis syndrome* (DILS) is used to describe this entity and to distinguish it from Sjögren's syndrome.

Approximately one-third of HIV-infected individuals experience arthralgias; furthermore, 5–10% are diagnosed as having some form of reactive arthritis, such as Reiter's syndrome or psoriatic arthritis as well as undifferentiated spondyloarthropathy (Chap. 355). These syndromes occur with increasing frequency as the competency of the immune system declines. This association may be related to an increase in the number of infections with organisms that may trigger a reactive arthritis with progressive immunodeficiency or to a loss of important regulatory T cells. Reactive arthritides in HIV-infected individuals generally respond well to standard treatment; however, therapy with methotrexate has been associated with an increase in the incidence of opportunistic infections and should be used with caution and only in severe cases.

HIV-infected individuals also experience a variety of joint problems without obvious cause that are referred to generically as *HIV-* or *AIDS-associated arthropathy*. This syndrome is characterized by subacute oligoarticular arthritis developing over a period of 1–6 weeks and lasting 6 weeks to 6 months. It generally involves the large joints, predominantly the knees and ankles, and is nonerosive with only a mild inflammatory response. X-rays are nonrevealing. Nonsteroidal anti-inflammatory drugs are only marginally helpful; however, relief has been noted with the use of intraarticular glucocorticoids. A second form of arthritis also thought to be secondary to HIV infection is called *painful articular syndrome*. This condition, reported as occurring in as many as 10% of AIDS patients, presents as an acute, severe, sharp pain in the affected joint. It affects primarily the knees, elbows, and shoulders; lasts 2–24 h; and may be severe enough to require narcotic analgesics. The cause of this arthropathy is unclear; however, it is thought to result from a direct effect of HIV on the joint. This condition is reminiscent of the fact that other lentiviruses, in particular the caprine arthritis-encephalitis virus, are capable of directly causing arthritis.

A variety of other immunologic or rheumatologic diseases have been reported in HIV-infected individuals, either de novo or in association with opportunistic infections or drugs. Using the criteria of widespread musculoskeletal pain of at least 3 months' duration and the presence of at least 11 of 18 possible tender points by digital palpation, 11% of an HIV-infected cohort containing 55% IDUs were diagnosed as having *fibromyalgia* (Chap. 366). While the incidence of frank arthritis was less in this population than in other studied populations that consisted predominantly of men who have sex with men, these data support the concept that there are musculoskeletal problems that occur as a direct result of HIV infection. In addition there have been reports of leukocytoclastic vasculitis in the setting of zidovudine therapy. CNS angiitis and polymyositis also have been reported in HIV-infected

individuals. Septic arthritis is surprisingly rare, especially given the increased incidence of staphylococcal bacteremias seen in this population. When septic arthritis has been reported, it has usually been due to *Staphylococcus aureus*, systemic fungal infection with *C. neoformans*, *Sporothrix schenckii*, or *H. capsulatum* or to systemic mycobacterial infection with *M. tuberculosis*, *M. haemophilum*, *M. avium*, or *M. kansasii*.

Patients with HIV infection treated with cART have been found to have an increased incidence of osteonecrosis or avascular necrosis of the hip and shoulders. In a study of asymptomatic patients, 4.4% were found to have evidence of osteonecrosis on MRI. While precise cause-and-effect relationships have been difficult to establish, this complication has been associated with the use of lipid-lowering agents, systemic glucocorticoids, and testosterone; bodybuilding exercise; alcohol consumption; and the presence of anticardiolipin antibodies. Osteoporosis has been reported in 7% of women with HIV infection, with 41% of women demonstrating some degree of osteopenia. Several studies have documented decreases in bone mineral density of 2–6% in the first 2 years following the initiation of cART. This may be particularly apparent with tenofovir-containing regimens.

Immune Reconstitution Inflammatory Syndrome (IRIS)

Following the initiation of effective cART, a paradoxical worsening of preexisting, untreated, or partially treated opportunistic infections may be noted. One may also see exacerbations of pre-existing autoimmune conditions or the development of new autoimmune conditions following the initiation of antiretrovirals **(Table 197-12)**. IRIS related to a known pre-existing infection or neoplasm is referred to as *paradoxical IRIS*, while IRIS associated with a previously undiagnosed condition is referred to as *unmasking IRIS*. The term *immune reconstitution disease (IRD)* is sometimes used to distinguish IRIS manifestations related to opportunistic diseases from IRIS manifestations related to autoimmune diseases. IRD is particularly common in patients with underlying untreated mycobacterial or fungal infections. Some form of IRIS is seen in 10–30% of patients, depending on the clinical setting, and is most common in patients starting therapy with CD4+ T cell counts <50 cells/μL who have a precipitous drop in HIV RNA levels following the initiation of cART. Signs and symptoms may appear anywhere from 2 weeks to 2 years after the initiation of cART and can include localized lymphadenitis, prolonged fever, pulmonary infiltrates, hepatitis, increased intracranial pressure, uveitis, sarcoidosis, and Graves' disease. The clinical course can be protracted, and severe cases can be fatal. The underlying mechanism appears to be related to a phenomenon similar to type IV hypersensitivity reactions and reflects the immediate improvements in immune function that occur as levels of HIV RNA drop and the immunosuppressive effects of HIV infection are controlled. In severe cases, the use of immunosuppressive drugs such as glucocorticoids may be required to blunt the inflammatory component of these reactions while specific antimicrobial therapy takes effect.

TABLE 197-12

Characteristics of Immune Reconstitution Inflammatory Syndrome (IRIS)

Paradoxical worsening of an existing clinical condition or abrupt appearance of a new clinical finding (unmasking) is seen following the initiation of antiretroviral therapy
Occurs weeks to months following the initiation of antiretroviral therapy
Is most common in patients starting therapy with a CD4+ T cell count <50/μL who experience a precipitous drop in viral load
Is frequently seen in the setting of tuberculosis; particularly when cART is starting soon after initiation of anti-TB therapy
Can be fatal

Diseases of the Hematopoietic System

Disorders of the hematopoietic system including lymphadenopathy, anemia, leukopenia, and/or thrombocytopenia are common throughout the course of HIV infection and may be the direct result of HIV, manifestations of secondary infections and neoplasms, or side effects of therapy **(Table 197-13)**. Direct histologic examination and culture of lymph node or bone marrow tissue are often diagnostic. A significant percentage of bone marrow aspirates from patients with HIV infection have been reported to contain lymphoid aggregates, the precise significance of which is unknown. Initiation of cART will lead to reversal of most hematologic complications that are the direct result of HIV infection.

TABLE 197-13

Causes of Bone Marrow Suppression in Patients with HIV Infection

HIV infection	Medications
Mycobacterial infections	Zidovudine
Fungal infections	Dapsone
B19 parvovirus infection	Trimethoprim/sulfamethoxazole
Lymphoma	Pyrimethamine
	5-Flucytosine
	Ganciclovir
	Interferon α
	Trimetrexate
	Foscarnet

Some patients, otherwise asymptomatic, may develop *persistent generalized lymphadenopathy* as an early clinical manifestation of HIV infection. This condition is defined as the presence of enlarged lymph nodes (>1 cm) in two or more extrainguinal sites for >3 months without an obvious cause. The lymphadenopathy is due to marked follicular hyperplasia in the node in response to HIV infection. The nodes are generally discrete and freely movable. This feature of HIV disease may be seen at any point in the spectrum of immune dysfunction and is not associated with an increased likelihood of developing AIDS. Paradoxically, a loss in lymphadenopathy or a decrease in lymph node size outside the setting of cART may be a prognostic marker of disease progression. In patients with CD4+ T cell counts >200/μL, the differential diagnosis of lymphadenopathy includes KS, TB, Castleman's disease, and lymphoma. In patients with more advanced disease, lymphadenopathy may also be due to atypical mycobacterial infection, toxoplasmosis, systemic fungal infection, or bacillary angiomatosis. While indicated in patients with CD4+ T cell counts <200/μL, lymph node biopsy is not indicated in patients with early-stage disease unless there are signs and symptoms of systemic illness, such as fever and weight loss, or unless the nodes begin to enlarge, become fixed, or coalesce. Monoclonal gammopathy of unknown significance (MGUS) **(Chap. 107)**, defined as the presence of a serum monoclonal IgG, IgA, or IgM in the absence of a clear cause, has been reported in 3% of patients with HIV infection. The overall clinical significance of this finding in patients with HIV infection is unclear, although it has been associated with other viral infections, non-Hodgkin's lymphoma, and plasma cell malignancy.

Anemia is the most common hematologic abnormality in HIV-infected patients and, in the absence of a specific treatable cause, is independently associated with a poor prognosis. While generally mild, anemia can be quite severe and require chronic blood transfusions. Among the specific reversible causes of anemia in the setting of HIV infection are drug toxicity, systemic fungal and mycobacterial infections, nutritional deficiencies, and parvovirus B19 infections. Zidovudine may block erythroid maturation prior to its effects on other marrow elements. A characteristic feature of zidovudine therapy is an elevated mean corpuscular volume (MCV). Another drug used in patients with HIV infection that has a selective effect on the erythroid series is dapsone. This drug can cause a serious hemolytic anemia in patients who are deficient in glucose-6-phosphate dehydrogenase and can create a functional anemia in others through induction of methemoglobinemia. Folate levels are usually normal in HIV-infected individuals; however, vitamin B_{12} levels may be depressed as a consequence of achlorhydria or malabsorption. True autoimmune hemolytic anemia is rare, although ~20% of patients with HIV infection may have a positive direct antiglobulin test as a consequence of polyclonal B cell activation. Infection with parvovirus B19 may also cause anemia. It is important to recognize this possibility given the fact that it responds well to treatment with IVIg. Erythropoietin levels in patients with HIV infection and anemia are generally lower than expected given the degree of anemia. Treatment with erythropoietin may result in an increase in hemoglobin levels. An exception to this is a subset of patients with zidovudine-associated anemia in whom erythropoietin levels may be quite high.

During the course of HIV infection, neutropenia may be seen in approximately half of patients. In most instances it is mild; however, it can be severe and can put patients at risk of spontaneous bacterial infections. This is most frequently seen in patients with severely advanced HIV disease and in patients receiving any of a number of potentially myelosuppressive therapies. In the setting of neutropenia, diseases that are not commonly seen in HIV-infected patients, such as aspergillosis or mucormycosis, may occur. Both granulocyte colony-stimulating factor (G-CSF) and GM-CSF increase neutrophil counts

in patients with HIV infection regardless of the cause of the neutropenia. Earlier concerns about the potential of these agents to also increase levels of HIV were not confirmed in controlled clinical trials.

Thrombocytopenia may be an early consequence of HIV infection. Approximately 3% of patients with untreated HIV infection and CD4+ T cell counts ≥400/μL have platelet counts <150,000/μL. For untreated patients with CD4+ T cell counts <400/μL, this incidence increases to 10%. In patients receiving antiretrovirals, thrombocytopenia is associated with hepatitis C, cirrhosis, and ongoing high-level HIV replication. Thrombocytopenia is rarely a serious clinical problem in patients with HIV infection and generally responds well to successful cART. Clinically, it resembles the thrombocytopenia seen in patients with idiopathic thrombocytopenic purpura **(Chap. 111)**. Immune complexes containing anti-gp120 antibodies and anti-anti-gp120 antibodies have been noted in the circulation and on the surface of platelets in patients with HIV infection. Patients with HIV infection have also been noted to have a platelet-specific antibody directed toward a 25-kDa component of the surface of the platelet. Other data suggest that the thrombocytopenia in patients with HIV infection may be due to a direct effect of HIV on megakaryocytes. Whatever the cause, it is very clear that the most effective medical approach to this problem has been the use of cART. For patients with platelet counts <20,000/μL, a more aggressive approach combining IVIg or anti-Rh Ig for an immediate response and cART for a more lasting response is appropriate. Rituximab has been used with some success in otherwise refractory cases. Splenectomy is a rarely needed option and is reserved for patients refractory to medical management. Because of the risk of serious infection with encapsulated organisms, all patients with HIV infection about to undergo splenectomy should be immunized with pneumococcal polysaccharide. It should be noted that, in addition to causing an increase in the platelet count, removal of the spleen will result in an increase in the peripheral blood lymphocyte count, making CD4+ T cell counts unreliable markers of immunocompetence. In this setting, the clinician should rely on the CD4+ T cell percentage for making diagnostic decisions with respect to the likelihood of opportunistic infections. A CD4+ T cell percentage of 15 is approximately equivalent to a CD4+ T cell count of 200/μL. In patients with early HIV infection, thrombocytopenia has also been reported as a consequence of classic thrombotic thrombocytopenic purpura **(Chap. 111)**. This clinical syndrome, consisting of fever, thrombocytopenia, hemolytic anemia, and neurologic and renal dysfunction, is a rare complication of early HIV infection. As in other settings, the appropriate management is the use of salicylates and plasma exchange. Other causes of thrombocytopenia include lymphoma, mycobacterial infections, and fungal infections.

The incidence of venous thromboembolic disease such as deep-vein thrombosis or pulmonary embolus is approximately 1% per year in patients with HIV infection. This is approximately 10 times higher than that seen in an age-matched population. Factors associated with an increased risk of clinical thrombosis include age over 45, history of an opportunistic infection, lower CD4 count, and estrogen use. Abnormalities of the coagulation cascade, including decreased protein S activity, increases in factor VIII, anticardiolipin antibodies, PAR-1 expression on T cells, or lupus-like anticoagulant, have been reported in more than 50% of patients with HIV infection. The clinical significance of this increased propensity toward thromboembolic disease is likely reflected in the observation that elevations in D-dimer are strongly associated with all-cause mortality in patients with HIV infection (Table 197-9).

Dermatologic Diseases

Dermatologic problems occur in >90% of patients with HIV infection. From the macular, roseola-like rash seen with the acute seroconversion syndrome to extensive end-stage KS, cutaneous manifestations of HIV disease can be seen throughout the course of HIV infection. Among the more common nonneoplastic problems are seborrheic dermatitis, folliculitis, and opportunistic infections. Extrapulmonary pneumocystosis may cause a necrotizing vasculitis. Neoplastic conditions are covered in a separate section below.

Seborrheic dermatitis occurs in 3% of the general population and in up to 50% of patients with HIV infection. Seborrheic dermatitis increases in prevalence and severity as the CD4+ T cell count declines. In HIV-infected patients, seborrheic dermatitis may be aggravated by concomitant infection with *Pityrosporum*, a yeastlike fungus; use of topical antifungal agents has been recommended in cases refractory to standard topical treatment.

Folliculitis is among the most prevalent dermatologic disorders in patients with HIV infection and is seen in ~20% of patients. It is more common in patients with CD4+ T cell counts <200 cells/μL. *Pruritic papular eruption* is one of the most common pruritic rashes in patients with HIV infection. It appears as multiple papules on the face, trunk, and extensor surfaces and may improve with cART. *Eosinophilic pustular folliculitis* is a rare form of folliculitis that is seen with increased frequency in patients with HIV infection. It presents as multiple, urticarial perifollicular papules that may coalesce into plaquelike lesions. Skin biopsy reveals an eosinophilic infiltrate of the hair follicle, which in certain cases has been associated with the presence of a mite. Patients typically have an elevated serum IgE level and may respond to treatment with topical anthelmintics. Pruritus is a common symptom in patients with HIV infection and can lead to *prurigo nodularis*. Patients with HIV infection have also been reported to develop a severe form of *Norwegian scabies* with hyperkeratotic psoriasiform lesions.

Both *psoriasis* and *ichthyosis*, although they are not reported to be increased in frequency, may be particularly severe when they occur in patients with HIV infection. Preexisting psoriasis may become guttate in appearance and more refractory to treatment in the setting of HIV infection.

Reactivation herpes zoster (*shingles*) is seen in 10–20% of patients with HIV infection. This reactivation syndrome of varicella-zoster virus indicates a modest decline in immune function and may be the first indication of clinical immunodeficiency. In one series, patients who developed shingles did so an average of 5 years after HIV infection. In a cohort of patients with HIV infection and localized zoster, the subsequent rate of the development of AIDS was 1% per month. In that study, AIDS was more likely to develop if the outbreak of zoster was associated with severe pain, extensive skin involvement, or involvement of cranial or cervical dermatomes. The clinical manifestations of reactivation zoster in HIV-infected patients, although indicative of immunologic compromise, are not as severe as those seen in other immunodeficient conditions. Thus, while lesions may extend over several

dermatomes, involve the spinal cord, and/or be associated with frank cutaneous dissemination, visceral involvement has not been reported. In contrast to patients without a known underlying immunodeficiency state, patients with HIV infection tend to have recurrences of zoster with a relapse rate of ~20%. Valacyclovir, acyclovir, or famciclovir is the treatment of choice. Foscarnet may be of value in patients with acyclovir-resistant virus.

Infection with *herpes simplex virus* in HIV-infected individuals is associated with recurrent orolabial, genital, and perianal lesions as part of recurrent reactivation syndromes **(Chap. 187)**. As HIV disease progresses and the CD4+ T cell count declines, these infections become more frequent and severe. Lesions often appear as beefy red, are exquisitely painful, and have a tendency to occur high in the gluteal cleft (Fig. 197-37). Perirectal HSV may be associated with proctitis and anal fissures. HSV should be high in the differential diagnosis of any HIV-infected patient with a poorly healing, painful perirectal lesion. In addition to recurrent mucosal ulcers, recurrent HSV infection in the form of *herpetic whitlow* can be a problem in patients with HIV infection, presenting with painful vesicles or extensive cutaneous erosion. Valacyclovir, acyclovir, or famciclovir is the treatment of choice in these settings. It is noteworthy that even subclinical reactivation of herpes simplex may be associated with increases in plasma HIV RNA levels.

Diffuse skin eruptions due to *Molluscum contagiosum* may be seen in patients with advanced HIV infection. These flesh-colored, umbilicated lesions resemble those of *Penicillium marnefei* or *Cryptococcosis.* They tend to regress with effective cART and can also be treated with local therapy. Similarly, *condyloma acuminatum* lesions may be more severe and more widely distributed in patients with low CD4+ T cell counts. Imiquimod cream may be helpful in some cases. Atypical mycobacterial infections may present as erythematous cutaneous nodules, as may fungal infections, *Bartonella*, *Acanthamoeba*, and KS. Cutaneous infections with *Aspergillus* have been noted at the site of IV catheter placement.

The skin of patients with HIV infection is often a target organ for drug reactions **(Chap. 56)**. Although most skin reactions are mild and not necessarily an indication to discontinue therapy, patients may have particularly severe cutaneous reactions, including erythroderma, *Stevens-Johnson syndrome*, and toxic epidermal necrolysis, as a reaction to drugs—particularly sulfa drugs, nonnucleoside reverse transcriptase inhibitors, abacavir, amprenavir, darunavir, fosamprenavir, and tipranavir. Similarly, patients with HIV infection are often quite photosensitive and burn easily following exposure to sunlight or as a side effect of radiation therapy **(Chap. 57)**.

HIV infection and its treatment may be accompanied by cosmetic changes of the skin that are not of great clinical importance but may be troubling to patients. Yellowing of the nails and straightening of the hair, particularly in African-American patients, have been reported as a consequence of HIV infection. Zidovudine therapy has been associated with elongation of the eyelashes and the development of a bluish discoloration to the nails, again more common in African-American patients. Therapy with clofazimine may cause a yellow-orange discoloration of the skin and urine.

Neurologic Diseases

Clinical disease of the nervous system accounts for a significant degree of morbidity in a high percentage of patients with HIV infection **(Table 197-14)**. The neurologic problems that occur in HIV-infected individuals may be either primary to the pathogenic processes of HIV infection or secondary to opportunistic infections or neoplasms. Among the more frequent opportunistic diseases that involve the CNS are toxoplasmosis, cryptococcosis, progressive multifocal leukoencephalopathy, and primary CNS lymphoma. Other less common problems include mycobacterial infections; syphilis; and infection with CMV, herpes zoster, HTLV-1, *Trypanosoma cruzi*, or *Acanthamoeba*. Overall, secondary diseases of the CNS have been reported to occur in approximately one-third of patients with AIDS. These data antedate the widespread use of cART, and this frequency is considerably lower in patients receiving effective antiretroviral drugs. Primary processes related to HIV infection of the nervous system are reminiscent of those seen with other lentiviruses, such as the maedi-visna virus of sheep.

TABLE 197-14

Neurologic Diseases in Patients with HIV Infection

Opportunistic infections	HIV-1 infection
Toxoplasmosis	Aseptic meningitis
Cryptococcosis	HIV-associated neurocognitive disorders (HAND), including HIV encephalopathy/AIDS dementia complex
Progressive multifocal leukoencephalopathy	Myelopathy
Cytomegalovirus	Vacuolar myelopathy
Syphilis	Pure sensory ataxia
Mycobacterium tuberculosis	Paresthesia/dysesthesia
HTLV-1 infection	Peripheral neuropathy
Amebiasis	Acute inflammatory demyelinating polyneuropathy (Guillain-Barré syndrome)
	Chronic inflammatory demyelinating polyneuropathy (CIDP)
Neoplasms	Mononeuritis multiplex
Primary CNS lymphoma	Distal symmetric polyneuropathy
Kaposi's sarcoma	Myopathy

Neurologic problems directly attributable to HIV occur throughout the course of infection and may be inflammatory, demyelinating, or degenerative in nature. The term *HIV-associated neurocognitive disorders* (HAND) is used to describe a spectrum of disorders that range from asymptomatic neurocognitive impairment (ANI) to minor neurocognitive disorder (MND) to clinically severe dementia. The most severe form, *HIV-associated dementia* (HAD), also referred to as the *AIDS dementia complex*, or *HIV encephalopathy*, is considered an AIDS-defining illness. Most HIV-infected patients have some neurologic problem during the course of their disease. Even in the setting of suppressive cART, approximately 50% of HIV-infected individuals can be shown to have mild to moderate neurocognitive impairment using sensitive neuropsychiatric testing. As noted in the section on pathogenesis, damage to the CNS may be a direct result of viral infection of the CNS macrophages or glial cells or may be secondary to the release of neurotoxins and potentially toxic cytokines such as IL-1β, TNF-α, IL-6, and TGF-β. It has been reported that HIV-infected individuals with the E4 allele for apoE are at increased risk for AIDS encephalopathy and peripheral neuropathy. Virtually all patients with HIV infection have some degree of nervous system involvement with the virus. This is evidenced by the fact that CSF findings are abnormal in ~90% of untreated patients, even during the asymptomatic phase of HIV infection. CSF abnormalities include pleocytosis (50–65% of patients), detection of viral RNA (~75%), elevated CSF protein (35%), and evidence of intrathecal synthesis of anti-HIV antibodies (90%). It is important to point out that evidence of infection of the CNS with HIV does not imply impairment of cognitive function. The neurologic function of an HIV-infected individual should be considered normal unless clinical signs and symptoms suggest otherwise.

Aseptic meningitis may be seen in any but the very late stages of HIV infection. In the setting of acute primary infection, patients may experience a syndrome of headache, photophobia, and meningismus. Rarely, an acute encephalopathy due to encephalitis may occur. Cranial nerve involvement may be seen, predominantly cranial nerve VII but occasionally V and/or VIII. CSF findings include a lymphocytic pleocytosis, elevated protein level, and normal glucose level. This syndrome, which cannot be clinically differentiated from other viral meningitides **(Chap. 134)**, usually resolves spontaneously within 2–4 weeks; however, in some patients, signs and symptoms may become chronic. Aseptic meningitis may occur any time in the course of HIV infection; however, it is rare following the development of AIDS. This suggests that clinical aseptic meningitis in the context of HIV infection is an immune-mediated disease.

Cryptococcus is the leading infectious cause of meningitis in patients with AIDS **(Chap. 210)**. While the vast majority of these are due to *C. neoformans*, up to 12% may be due to *C. gattii*. Cryptococcal meningitis is the initial AIDS-defining illness in ~2% of patients and generally occurs in patients with CD4+ T cell counts <100/μL. Cryptococcal meningitis is particularly common in untreated patients with AIDS in Africa, occurring in ~5% of patients. Most patients present with a picture of subacute meningoencephalitis with fever, nausea, vomiting, altered mental status, headache, and meningeal signs. The incidence of seizures and focal neurologic deficits is low. The CSF profile may be normal or may show only modest elevations in WBC or protein levels and decreases in glucose. The opening pressure in the CSF is usually elevated. In addition to meningitis, patients may develop cryptococcomas and cranial nerve involvement. Approximately one-third of patients also have pulmonary disease. Uncommon manifestations of cryptococcal infection include skin lesions that resemble *molluscum contagiosum*, lymphadenopathy, palatal and glossal ulcers, arthritis, gastroenteritis, myocarditis, and prostatitis. The prostate gland may serve as a reservoir for smoldering cryptococcal infection. The diagnosis of cryptococcal meningitis is made by identification of organisms in spinal fluid with india ink examination or by the detection of cryptococcal antigen. Blood cultures for fungus are often positive. A biopsy may be needed to make a diagnosis of CNS cryptococcoma. Treatment is with IV amphotericin B 0.7 mg/kg daily, or liposomal amphotericin 4–6 mg/kg daily, with flucytosine 25 mg/kg qid for at least 2 weeks if possible, continuing with amphotericin alone ideally until the CSF culture turns negative. Decreases in renal function in association with amphotericin can lead to increases in flucytosine levels and subsequent bone marrow suppression. Amphotericin is followed by fluconazole 400 mg/d PO for 8 weeks, and then fluconazole 200 mg/d until the CD4+ T cell count has increased to >200 cells/μL for 6 months in response to cART. Repeated lumbar puncture may be required to manage increased intracranial pressure.

Symptoms may recur with initiation of cART as an immune reconstitution syndrome (see above). Other fungi that may cause meningitis in patients with HIV infection are *C. immitis* and *H. capsulatum*. Meningoencephalitis has also been reported due to *Acanthamoeba* or *Naegleria*.

HIV-associated dementia consists of a constellation of signs and symptoms of CNS disease. While this is generally a late complication of HIV infection that progresses slowly over months, it can be seen in patients with CD4+ T cell counts >350 cells/μL. A major feature of this entity is the development of dementia, defined as a decline in cognitive ability from a previous level. It may present as impaired ability to concentrate, increased forgetfulness, difficulty reading, or increased difficulty performing complex tasks. Initially these symptoms may be indistinguishable from findings of situational depression or fatigue. In contrast to "cortical" dementia (such as Alzheimer's disease), aphasia, apraxia, and agnosia are uncommon, leading some investigators to classify HIV encephalopathy as a "subcortical dementia" characterized by defects in short-term memory and executive function (see below). In addition to dementia, patients with HIV encephalopathy may also have motor and behavioral abnormalities. Among the motor problems are unsteady gait, poor balance, tremor, and difficulty with rapid alternating movements. Increased tone and deep tendon reflexes may be found in patients with spinal cord involvement. Late stages may be complicated by bowel and/or bladder incontinence. Behavioral problems include apathy, irritability, and lack of initiative, with progression to a vegetative state in some instances. Some patients develop a state of agitation or mild mania. These changes usually occur without significant changes in level of alertness. This is in contrast to the finding of somnolence in patients with dementia due to toxic/metabolic encephalopathies.

HIV-associated dementia is the initial AIDS-defining illness in ~3% of patients with HIV infection and thus only rarely precedes clinical evidence of immunodeficiency. Clinically significant encephalopathy eventually develops in ~25% of untreated patients with AIDS. As immunologic function declines, the risk and severity of HIV-associated dementia increases. Autopsy series suggest that 80–90% of patients with HIV infection have histologic evidence of CNS involvement. Several classification schemes have been developed for grading HIV encephalopathy; a commonly used clinical staging system is outlined in **Table 197-15**.

TABLE 197-15

Clinical Staging of Hand According to Frascati Criteria

Stage	Neurocognitive Status[a]	Functional Status[b]
Asymptomatic	1 SD below mean in 2 cognitive domains	No impairments in activities of daily living
Mild neurocognitive disorder	1 SD below mean in 2 cognitive domains	Impairments in activities of daily living
HIV-associated dementia	2 SD below mean in 2 cognitive domains	Notable impairments in activities of daily living

[a]Neurocognitive testing should include assessment of at least 5 domains, including attention-information processing, language, abstraction-executive, complex perceptual motor skills, memory (including learning and recall), simple motor skills, or sensory perceptual skills. Appropriate norms must be available to establish the number of domains in which performance is below 1 SD. [b]Functional status is typically assessed by self-reporting but might be corroborated by a collateral source. No agreed measures exist for HIV-associated neurocognitive disorder criteria. Note that, for diagnosis of HIV-associated neurocognitive disorder, other causes of dementia must be ruled out and potential confounding effects of substance use or psychiatric illness should be considered.

Source: Adapted from A Antinori et al: Neurology 69:1789, 2007.

The precise cause of HIV-associated dementia remains unclear, although the condition is thought to be a result of a combination of direct effects of HIV on the CNS and associated immune activation. HIV has been found in the brains of patients with HIV encephalopathy by Southern blot, in situ hybridization, PCR, and electron microscopy. Multinucleated giant cells, macrophages, and microglial cells appear to be the main cell types harboring virus in the CNS. Histologically, the major changes are seen in the subcortical areas of the brain and include pallor and gliosis, multinucleated giant cell encephalitis, and vacuolar myelopathy. Less commonly, diffuse or focal spongiform changes occur in the white matter. Areas of the brain involved in motor function, language, and judgment are most severely affected.

There are no specific criteria for a diagnosis of HIV-associated dementia, and this syndrome must be differentiated from a number of other diseases that affect the CNS of HIV-infected patients (Table 197-14). The diagnosis of dementia depends on demonstrating a decline in cognitive function. This can be accomplished objectively with the use of a Mini-Mental Status Examination (MMSE) in patients for whom prior scores are available. For this reason, it is advisable for all patients with a diagnosis of HIV infection to have a baseline MMSE. However, changes in MMSE scores may be absent in patients with mild HIV encephalopathy. Imaging studies of the CNS, by either MRI or CT, often demonstrate evidence of cerebral atrophy **(Fig. 197-40)**. MRI may also reveal small areas of increased density on T2-weighted images. Lumbar puncture is an important element of the evaluation of patients with HIV infection and neurologic abnormalities. It is generally most helpful in ruling out or making a diagnosis of opportunistic infections. In HIV encephalopathy, patients may have the nonspecific findings of an increase in CSF cells and protein level. While HIV RNA can often be detected in the spinal fluid and HIV can be cultured from the CSF, this finding is not specific for HIV encephalopathy. There appears to be no correlation between the

presence of HIV in the CSF and the presence of HIV encephalopathy. Elevated levels of macrophage chemoattractant protein (MCP-1), β_2-microglobulin, neopterin, and quinolinic acid (a metabolite of tryptophan reported to cause CNS injury) have been noted in the CSF of patients with HIV encephalopathy. These findings suggest that these factors as well as inflammatory cytokines may be involved in the pathogenesis of this syndrome.

FIGURE 197-40

AIDS dementia complex. Postcontrast CT scan through the lateral ventricles of a 47-year-old man with AIDS, altered mental status, and dementia. The lateral and third ventricles and the cerebral sulci are abnormally prominent. Mild white matter hypodensity is seen adjacent to the frontal horns of the lateral ventricles.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Combination antiretroviral therapy is of benefit in patients with HIV-associated dementia. Improvement in neuropsychiatric test scores has been noted for both adult and pediatric patients treated with antiretrovirals. The rapid improvement in cognitive function noted with the initiation of cART suggests that at least some component of this problem is quickly reversible, again supporting at least a partial role of soluble mediators in the pathogenesis. It should also be noted that these patients have an increased sensitivity to the side effects of neuroleptic drugs. The use of these drugs for symptomatic treatment is associated with an increased risk of extrapyramidal side effects; therefore, patients with HIV encephalopathy who receive these agents must be monitored carefully. It is felt by many physicians that the decrease in the prevalence of severe cases of HAND brought about by cART has resulted in an increase in the prevalence of milder forms of this disorder.

Seizures may be a consequence of opportunistic infections, neoplasms, or HIV encephalopathy **(Table 197-16)**. The seizure threshold is often lower than normal in patients with advanced HIV infection due in part to the frequent presence of electrolyte abnormalities. Seizures are seen in 15–40% of patients with cerebral toxoplasmosis, 15–35% of patients with primary CNS lymphoma, 8% of patients with cryptococcal meningitis, and 7–50% of patients with HIV encephalopathy. Seizures may also be seen in patients with CNS tuberculosis, aseptic meningitis, and progressive multifocal leukoencephalopathy. Seizures may be the presenting clinical symptom of HIV disease. In one study of 100 patients with HIV infection presenting with a first seizure, cerebral mass lesions were the most common cause, responsible for 32 of the 100 new-onset seizures. Of these 32 cases, 28 were due to toxoplasmosis and 4 to lymphoma. HIV encephalopathy accounted for an additional 24 new-onset seizures. Cryptococcal meningitis was the third most common diagnosis, responsible for 13 of the 100 seizures. In 23 cases, no cause could be found, and it is possible that these cases represent a subcategory of HIV encephalopathy. Of these 23 cases, 16 (70%) had 2 or more seizures, suggesting that anticonvulsant therapy is indicated in all patients with HIV infection and seizures unless a rapidly correctable cause is found. Due to a variety of drug-drug interactions between antiseizure medications and antiretrovirals, drug levels need to be monitored carefully.

TABLE 197-16

Causes of Seizures in Patients with HIV Infection

Disease	Overall Contribution to First Seizure, %	Fraction of Patients Who Have Seizures, %
HIV encephalopathy	24–47	7–50
Cerebral toxoplasmosis	28	15–40
Cryptococcal meningitis	13	8
Primary central nervous system lymphoma	4	15–30
Progressive multifocal leukoencephalopathy	1	20

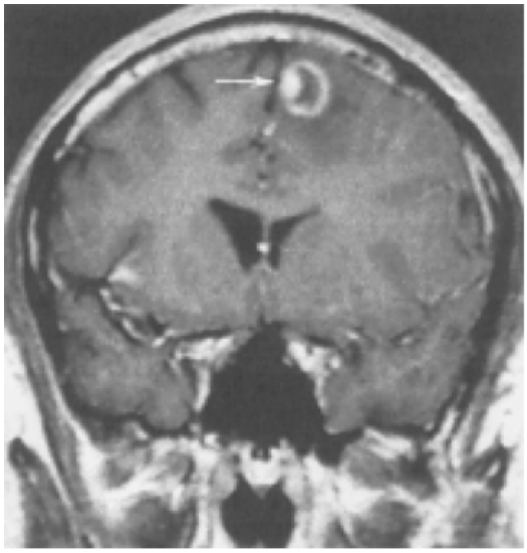
Source: From DM Holtzman et al: Am J Med 87:173, 1989.

Patients with HIV infection may present with *focal neurologic deficits* from a variety of causes. The most common causes are toxoplasmosis, progressive multifocal leukoencephalopathy, and CNS lymphoma. Other causes include cryptococcal infections (discussed above; also **Chap. 210**), stroke, and reactivation of Chagas' disease.

Toxoplasmosis has been one of the most common causes of secondary CNS infections in patients with AIDS, but its incidence is decreasing in the era of cART. It is most common in patients from the Caribbean and from France, where the seroprevalence of *T. gondii* is around 50%. This figure is closer to 15% in the United States. Toxoplasmosis is generally a late complication of HIV infection and usually occurs in patients with CD4+ T cell counts <200/μL. Cerebral toxoplasmosis is thought to represent a reactivation of latent tissue cysts. It is 10 times more common in patients with antibodies to the organism than in patients who are seronegative. Patients diagnosed with HIV infection should be screened for IgG antibodies to *T. gondii* during the time of their initial workup. Those who are seronegative should be counseled about ways to minimize the risk of primary infection including avoiding the consumption of undercooked meat and careful hand washing after contact with soil or changing the cat litter box. The most common clinical presentation of cerebral toxoplasmosis in patients with HIV infection is fever, headache, and focal neurologic deficits. Patients may present with seizure, hemiparesis, or aphasia as a manifestation of these focal deficits or with a picture more influenced by the accompanying cerebral edema and characterized by confusion, dementia, and lethargy, which can progress to coma. The diagnosis is usually suspected on the basis of MRI findings of multiple lesions in multiple locations, although in some cases only a single lesion is seen. Pathologically, these lesions generally exhibit inflammation and central necrosis and, as a result, demonstrate ring enhancement on contrast MRI **(Fig. 197-41)** or, if MRI is unavailable or contraindicated, on double-dose contrast CT. There is usually evidence of surrounding edema. In addition to toxoplasmosis, the differential diagnosis of single or multiple enhancing mass lesions in the HIV-infected patient includes primary CNS lymphoma and, less commonly, TB or fungal or bacterial abscesses. The definitive diagnostic procedure is brain biopsy. However, given the morbidity rate that can accompany this procedure, it is usually reserved for the patient who has failed 2–4 weeks of empiric therapy for toxoplasmosis. If the patient is seronegative for *T. gondii*, the likelihood that a mass lesion is due to toxoplasmosis is <10%. In that setting, one may choose to be more aggressive and perform a brain biopsy sooner. Standard treatment is sulfadiazine and pyrimethamine with leucovorin as needed for a minimum of 4–6 weeks. Alternative therapeutic regimens include clindamycin in combination with pyrimethamine; atovaquone plus pyrimethamine; and azithromycin plus pyrimethamine plus rifabutin. Relapses are common, and it is recommended that patients with a history of prior toxoplasmic encephalitis receive maintenance therapy with sulfadiazine, pyrimethamine, and leucovorin as long as their CD4+ T cell counts remain <200 cells/μL. Patients with CD4+ T cell counts <100/μL and IgG antibody to *Toxoplasma* should receive primary prophylaxis for toxoplasmosis. Fortunately, the same daily regimen of a single double-strength tablet of TMP-SMX used for *P. jirovecii* prophylaxis provides adequate primary protection against toxoplasmosis. Secondary prophylaxis/maintenance therapy for toxoplasmosis may be discontinued in the setting of effective cART and increases in CD4+ T cell counts to >200/μL for 6 months.

FIGURE 197-41

Central nervous system toxoplasmosis. A coronal postcontrast T1-weighted MRI scan demonstrates a peripheral enhancing lesion in the left frontal lobe, associated with an eccentric nodular area of enhancement (*arrow*); this so-called eccentric target sign is typical of toxoplasmosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

JC virus, a human polyomavirus that is the etiologic agent of *progressive multifocal leukoencephalopathy* (PML), is an important opportunistic pathogen in patients with AIDS **(Chap. 133)**. While ~80% of the general adult population has antibodies to JC virus, indicative of prior infection, <10% of healthy adults show any evidence of ongoing viral replication. PML is the only known clinical manifestation of JC virus infection. It is a late manifestation of AIDS and is seen in ~1–4% of patients with AIDS. The lesions of PML begin as small foci of demyelination in subcortical white matter that eventually coalesce. The cerebral hemispheres, cerebellum, and brainstem may all be involved. Patients typically have a protracted course with multifocal neurologic deficits, with or without changes in mental status. Approximately 20% of patients experience seizures. Ataxia, hemiparesis, visual field defects, aphasia, and sensory defects may occur. Headache, fever, nausea, and vomiting are rarely seen. Their presence should suggest another diagnosis. MRI typically reveals multiple, nonenhancing white matter lesions that may coalesce and have a predilection for the occipital and parietal lobes. The lesions show signal hyperintensity on T2-weighted images and diminished signal on T1-weighted images. The measurement of JC virus DNA levels in CSF has a diagnostic sensitivity of 76% and a specificity of close to 100%. Prior to the availability of cART, the majority of patients with PML died within 3–6 months of the onset of symptoms. Paradoxical worsening of PML has been seen with initiation of cART as an immune reconstitution syndrome. There is no specific treatment for PML; however, a median survival of 2 years and survival of >15 years have been reported in patients with PML treated with cART for their HIV disease. Despite having a significant impact on survival, only ~50% of patients with HIV infection and PML show neurologic improvement with cART. Studies with other antiviral agents such as cidofovir have failed to show clear benefit. Factors influencing a favorable prognosis for PML in the setting of HIV infection include a CD4+ T cell count >100/μL at baseline and the ability to maintain an HIV viral load of <500 copies/mL. Baseline HIV-1 viral load does not have independent predictive value of survival. PML is one of the few opportunistic infections that continues to occur with some frequency despite the widespread use of cART.

Reactivation American trypanosomiasis may present as acute meningoencephalitis with focal neurologic signs, fever, headache, vomiting, and seizures. Accompanying cardiac disease in the form of arrhythmias or heart failure should increase the index of suspicion. The presence of antibodies to *T. cruzi* supports the diagnosis. In South America, reactivation of *Chagas' disease* is considered to be an AIDS-defining condition and may be the initial AIDS-defining condition. The majority of cases occur in patients with CD4+ T cell counts <200 cells/μL. Lesions appear radiographically as single or multiple hypodense areas, typically with ring enhancement and edema. They are found predominantly in the subcortical areas, a feature that differentiates them from the deeper lesions of toxoplasmosis. *T. cruzi* amastigotes, or trypanosomes, can be identified from biopsy specimens or CSF. Other CSF findings include elevated protein and a mild (<100 cells/μL) lymphocytic pleocytosis. Organisms can also be identified by direct examination of the blood. Treatment consists of benzimidazole (2.5 mg/kg bid) or nifurtimox (2 mg/kg qid) for at least 60 days, followed by maintenance therapy for the duration of immunodeficiency with either drug at a dose of 5 mg/kg three times a week. As is the case with cerebral toxoplasmosis, successful therapy with antiretrovirals may allow discontinuation of therapy for Chagas' disease.

Stroke may occur in patients with HIV infection. In contrast to the other causes of focal neurologic deficits in patients with HIV infection, the symptoms of a stroke are sudden in onset. Patients with HIV infection have an increased prevalence of many classic risk factors associated with stroke, including smoking and diabetes. It has been reported that HIV infection itself can lead to an increase in carotid artery stiffness. The relative increase in risk for stroke as a consequence of HIV infection is more pronounced in women and in individuals between the ages of 18 and 29. Among the secondary infectious diseases in patients with HIV infection that may be associated with stroke are vasculitis due to cerebral varicella zoster or neurosyphilis and septic embolism in association with fungal infection. Other elements of the differential diagnosis of stroke in the patient with HIV infection include atherosclerotic cerebral vascular disease, thrombotic thrombocytopenic purpura, and cocaine or amphetamine use.

Primary CNS lymphoma is discussed below in the section on neoplastic diseases.

Spinal cord disease, or myelopathy, is present in ~20% of patients with AIDS, often as part of HIV-associated neurocognitive disorder. In fact, 90% of the patients with HIV-associated myelopathy have some evidence of dementia, suggesting that similar pathologic processes may be responsible for both conditions. Three main types of spinal cord disease are seen in patients with AIDS. The first of these is a vacuolar myelopathy, as mentioned above. This condition is pathologically similar to subacute combined degeneration of the cord, such as that occurring with pernicious anemia. Although vitamin B_{12} deficiency can be seen in patients with AIDS as a primary complication of HIV infection, it does not appear to be responsible for the majority of the cases of myelopathy seen in patients with HIV infection. Vacuolar myelopathy is characterized by a subacute onset and often presents with gait disturbances, predominantly ataxia and spasticity; it may progress to include bladder and bowel dysfunction. Physical findings include evidence of increased deep tendon reflexes and extensor plantar responses. The second form of spinal cord disease involves the dorsal columns and presents as a pure sensory ataxia. The third form is also sensory in nature and presents with paresthesias and dysesthesias of the lower extremities. In contrast to the cognitive problems seen in patients with HIV encephalopathy, these spinal cord syndromes do not respond well to antiretroviral drugs, and therapy is mainly supportive.

One important disease of the spinal cord that also involves the peripheral nerves is a *myelopathy* and *polyradiculopathy* seen in association with CMV infection. This entity is generally seen late in the course of HIV infection and is fulminant in onset, with lower extremity and sacral paresthesias, difficulty in walking, areflexia, ascending sensory loss, and urinary retention. The clinical course is rapidly progressive over a period of weeks. CSF examination reveals a predominantly neutrophilic pleocytosis, and CMV DNA can be detected by CSF PCR. Therapy with ganciclovir or foscarnet can lead to rapid improvement, and prompt initiation of foscarnet or ganciclovir therapy is important in minimizing the degree of permanent neurologic damage. Combination therapy with both drugs should be considered in patients who have been previously treated for CMV disease. Other diseases involving the spinal cord in patients with HIV infection include HTLV-1-associated myelopathy (HAM) **(Chap. 196)**, neurosyphilis **(Chap. 177)**, infection with herpes simplex **(Chap. 187)** or varicella-zoster **(Chap. 188)**, TB **(Chap. 173)**, and lymphoma **(Chap. 104)**.

Peripheral neuropathies are common in patients with HIV infection. They occur at all stages of illness and take a variety of forms. Early in the course of HIV infection, an acute inflammatory demyelinating polyneuropathy resembling Guillain-Barré syndrome may occur **(Chap. 439)**. In other patients, a progressive or relapsing-remitting inflammatory neuropathy resembling chronic inflammatory demyelinating polyneuropathy (CIDP) has been noted. Patients commonly present with progressive weakness, areflexia, and minimal sensory changes. CSF examination often reveals a mononuclear pleocytosis, and peripheral nerve biopsy demonstrates a perivascular infiltrate suggesting an autoimmune etiology. Plasma exchange or IVIg has been tried with variable success. Because of the immunosuppressive effects of glucocorticoids, they should be reserved for severe cases of CIDP refractory to other measures. Another autoimmune peripheral neuropathy seen in patients with AIDS is mononeuritis multiplex **(Chaps. 439 and 356)** due to a necrotizing arteritis of peripheral nerves. The most common peripheral neuropathy in patients with HIV infection is a *distal sensory polyneuropathy* (DSPN) also referred to as painful sensory neuropathy (HIV-SN), predominantly sensory neuropathy, or distal symmetric peripheral neuropathy. This condition may be a direct consequence of HIV infection or a side effect of dideoxynucleoside therapy. It is more common in taller individuals, older individuals, and those with lower CD4 counts. Two-thirds of patients with AIDS may be shown by electrophysiologic studies to have some evidence of peripheral nerve disease. Presenting symptoms are usually painful burning sensations in the feet and lower extremities. Findings on examination include a stocking-type sensory loss to pinprick, temperature, and touch sensation and a loss of ankle reflexes. Motor changes are mild and are usually limited to weakness of the intrinsic foot muscles. Response of this condition to antiretrovirals has been variable, perhaps because antiretrovirals are responsible for the problem in some instances. When due to dideoxynucleoside therapy, patients with lower extremity peripheral neuropathy may complain of a sensation that they are walking on ice. Other entities in the differential diagnosis of peripheral neuropathy include diabetes mellitus, vitamin B_{12} deficiency, and side effects from metronidazole or dapsone. For distal symmetric polyneuropathy that fails to resolve following the discontinuation of dideoxynucleosides, therapy is symptomatic; gabapentin, carbamazepine, tricyclics, or analgesics may be effective for dysesthesias. Treatment-naïve patients may respond to cART.

Myopathy may complicate the course of HIV infection; causes include HIV infection itself, zidovudine, and the generalized wasting syndrome (discussed below). HIV-associated myopathy may range in severity from an asymptomatic elevation in creatine kinase levels to a subacute syndrome characterized by proximal muscle weakness and myalgias. Quite pronounced elevations in creatine kinase may occur in asymptomatic patients, particularly after exercise. The clinical significance of this as an isolated laboratory finding is unclear. A variety of both inflammatory and noninflammatory pathologic processes have been noted in patients with more severe myopathy, including myofiber necrosis with inflammatory cells, nemaline rod bodies, cytoplasmic bodies, and mitochondrial abnormalities. Profound muscle wasting, often with muscle pain, may be seen after prolonged zidovudine therapy. This toxic side effect of the drug is dose-dependent and is related to its ability to interfere with the function of mitochondrial polymerases. It is reversible following discontinuation of the drug. Red ragged fibers are a histologic hallmark of zidovudine-induced myopathy.

Ophthalmologic Diseases

Ophthalmologic problems occur in ~50% of patients with advanced HIV infection. The most common abnormal findings on funduscopic examination are cotton-wool spots. These are hard white spots that appear on the surface of the retina and often have an irregular edge. They represent areas of retinal ischemia secondary to microvascular disease. At times they are associated with small areas of hemorrhage and thus can be difficult to distinguish from CMV retinitis. In contrast to CMV retinitis, however, these lesions are not associated with visual loss and tend to remain stable or improve over time.

One of the most devastating consequences of HIV infection is CMV retinitis. Patients at high risk of CMV retinitis (CD4+ T cell count <100/μL) should undergo an ophthalmologic examination every 3–6 months. The majority of cases of CMV retinitis occur in patients with a CD4+ T cell count <50/μL. Prior to the availability of cART, this CMV reactivation syndrome was seen in 25–30% of patients with AIDS. In the cART era this has dropped to close to 2%. CMV retinitis usually presents as a painless, progressive loss of vision. Patients may also complain of blurred vision, "floaters," and scintillations. The disease is usually bilateral, although typically it affects one eye more than the other. The diagnosis is made on clinical grounds by an experienced ophthalmologist. The characteristic retinal appearance is that of perivascular hemorrhage and exudate. In situations where the diagnosis is in doubt due to an atypical presentation or an unexpected lack of response to therapy, vitreous or aqueous humor sampling with molecular diagnostic techniques may be of value. CMV infection of the retina results in a necrotic inflammatory process, and the visual loss that develops is irreversible. CMV retinitis may be complicated by rhegmatogenous retinal detachment as a consequence of retinal atrophy in areas of prior inflammation. Therapy for CMV retinitis consists of oral valganciclovir, IV ganciclovir, or IV foscarnet, with cidofovir as an alternative. Combination therapy with ganciclovir and foscarnet has been shown to be slightly more effective than either ganciclovir or foscarnet alone in the patient with relapsed CMV retinitis. A 3-week induction course is followed by maintenance therapy with oral valganciclovir. If CMV disease is limited to the eye, intravitreal injections of ganciclovir or foscarnet may be considered. Intravitreal injections of cidofovir are generally avoided due to the increased risk of uveitis and hypotony. Maintenance therapy is continued until the CD4+ T cell count remains >100 μL for >6 months. The majority of patients with HIV infection and CMV disease develop some degree of uveitis with the initiation of cART. The etiology of this is unknown; however, it has been suggested that this may be due to the generation of an enhanced immune response to CMV as an IRIS (see above). In some instances this has required the use of topical glucocorticoids.

Both HSV and varicella zoster virus can cause a rapidly progressing, bilateral, necrotizing retinitis referred to as the *acute retinal necrosis syndrome*, or *progressive outer retinal necrosis* (PORN). This syndrome, in contrast to CMV retinitis, is associated with pain, keratitis, and iritis. It is often associated with orolabial HSV or trigeminal zoster. Ophthalmologic examination reveals widespread pale gray peripheral lesions. This condition is often complicated by retinal detachment. It is important to recognize and treat this condition with IV acyclovir as quickly as possible to minimize the loss of vision.

Several other secondary infections may cause ocular problems in HIV-infected patients. *P. jirovecii* can cause a lesion of the choroid that may be detected as an incidental finding on ophthalmologic examination. These lesions are typically bilateral, are from half to twice the disc diameter in size, and appear as slightly elevated yellow-white plaques. They are usually asymptomatic and may be confused with cotton-wool spots. Chorioretinitis due to toxoplasmosis can be seen alone or, more commonly, in association with CNS toxoplasmosis. KS may involve the eyelid or conjunctiva, while lymphoma may involve the retina. Syphilis may lead to a uveitis that is highly associated with the presence of neurosyphilis.

Additional Disseminated Infections and Wasting Syndrome

Infections with species of the small, gram-negative, *Rickettsia*-like organism *Bartonella* **(Chap. 167)** are seen with increased frequency in patients with HIV infection. While it is not considered an AIDS-defining illness by the CDC, many experts view infection with *Bartonella* as indicative of a severe defect in cell-mediated immunity. It is usually seen in patients with CD4+ T cell counts <100/μL and is a significant cause of unexplained fever in patients with advanced HIV infection. Among the clinical manifestations of *Bartonella* infection are bacillary angiomatosis, cat-scratch disease, and trench fever. *Bacillary angiomatosis* is usually due to infection with *B. henselae* and is linked to exposure to flea-infested cats. It is characterized by a vascular proliferation that leads to a variety of skin lesions that have been confused with the skin lesions of KS. In contrast to the lesions of KS, the lesions of bacillary angiomatosis generally blanch, are painful, and typically occur in the setting of systemic symptoms. Infection can extend to the lymph nodes, liver (peliosis hepatis), spleen, bone, heart, CNS, respiratory tract, and GI tract. *Cat-scratch disease* also is due to *B. henselae* and generally begins with a papule at the site of inoculation. This is followed several weeks later by the development of regional adenopathy and malaise. Infection with *B. quintana* is transmitted by lice and has been associated with case reports of trench fever, endocarditis, adenopathy, and bacillary angiomatosis. The organism is quite difficult to culture, and diagnosis often relies on identifying the organism in biopsy specimens using the Warthin-Starry or similar stains. Treatment is with either doxycycline or erythromycin for at least 3 months.

Histoplasmosis is an opportunistic infection that is seen most frequently in patients in the Mississippi and Ohio River valleys, Puerto Rico, the Dominican Republic, and South America. These are all areas in which infection with *H. capsulatum* is endemic **(Chap. 207)**. Because of this limited geographic distribution, the percentage of AIDS cases in the United States with histoplasmosis is only ~0.5. Histoplasmosis is generally a late manifestation of HIV infection; however, it may be the initial AIDS-defining condition. In one study, the median CD4+ T cell count for patients with histoplasmosis and AIDS was 33/μL. While disease due to *H. capsulatum* may present as a primary infection of the lung, disseminated disease, presumably due to reactivation, is the most common presentation in HIV-infected patients. Patients usually present with a 4- to 8-week history of fever and weight loss. Hepatosplenomegaly and lymphadenopathy are each seen in about 25% of patients. CNS disease, either meningitis or a mass lesion, is seen in 15% of patients. Bone marrow involvement is common, with thrombocytopenia, neutropenia, and anemia occurring in 33% of patients. Approximately 7% of patients have mucocutaneous lesions consisting of a maculopapular rash and skin or oral ulcers. Respiratory symptoms are usually mild, with chest x-ray showing a diffuse infiltrate or diffuse small nodules in ~50% of cases. The gastrointestinal tract may be involved. Diagnosis is made by silver staining of tissue, by culturing the organisms from blood, bone marrow, or tissue, or by detecting antigen in blood or urine. Treatment is typically with liposomal amphotericin B followed by maintenance therapy with oral itraconazole until the serum histoplasma antigen is <2 units, the patient has been on antiretrovirals for at least 6 months, and the CD4 count is >150 cells/μL. In the setting of mild infection, it may be appropriate to initiate therapy with itraconazole alone.

Following the spread of HIV infection to southeast Asia, disseminated infection with the fungus *Penicillium marneffei* was recognized as a complication of HIV infection and is considered an AIDS-defining condition in those parts of the world where it occurs. *P. marneffei* is the third most common AIDS-defining illness in Thailand, following TB and cryptococcosis. It is more frequently diagnosed in the rainy than the dry season. Clinical features include fever, generalized lymphadenopathy, hepatosplenomegaly, anemia, thrombocytopenia, and papular skin lesions with central umbilication resembling the lesions of *Molluscum contagiosum*. Treatment is with amphotericin B followed by itraconazole until the CD4+ T cell count is >100 cells/μL for at least 6 months.

Visceral leishmaniasis **(Chap. 221)** is recognized with increasing frequency in patients with HIV infection who live in or travel to areas endemic for this protozoal infection transmitted by sandflies. The clinical presentation is one of hepatosplenomegaly, fever, and hematologic abnormalities. Lymphadenopathy and other constitutional symptoms may be present. A chronic, relapsing course is seen in two-thirds of co-infected patients. Organisms can be isolated from cultures of bone marrow aspirates. Histologic stains may be negative, and antibody titers are of little help. Patients with HIV infection usually respond well initially to standard therapy with amphotericin B or pentavalent antimony compounds. Eradication of the organism is difficult, however, and relapses are common.

Patients with HIV infection are at a slightly increased risk of clinical malaria. This is particularly true for patients from nonendemic areas and thus at risk for primary infection and in patients with lower CD4+ T cell counts. HIV-positive individuals with CD4+ T cell counts <300 cells/μL have a poorer response to malaria treatment than others. Co-infection with malaria is associated with a modest increase in HIV viral load. The risk of malaria may be decreased with TMP-SMX prophylaxis.

Generalized wasting is an AIDS-defining condition; it is defined as involuntary weight loss of >10% associated with intermittent or constant fever and chronic diarrhea or fatigue lasting >30 days in the absence of a defined cause other than HIV infection. Prior to the widespread use of cART it was the initial AIDS-defining condition in ~10% of patients with AIDS in the United States. Generalized wasting is rarely seen today with the earlier initiation of antiretrovirals. A constant feature of this syndrome is severe muscle wasting with scattered myofiber degeneration and occasional evidence of myositis. Glucocorticoids may be of some benefit; however, this approach must be carefully weighed against the risk of compounding the immunodeficiency of HIV infection. Androgenic steroids, growth hormone, and total parenteral nutrition have been used as therapeutic interventions with variable success.

Neoplastic Diseases

The neoplastic diseases considered to be AIDS-defining conditions are Kaposi's sarcoma, non-Hodgkin's lymphoma, and invasive cervical carcinoma. In addition, there is also an increase in the incidence of a variety of non-AIDS-defining malignancies including Hodgkin's disease; multiple myeloma; leukemia; melanoma; and cervical, brain, testicular, oral, lung, gastric, liver, renal, and anal cancers. Since the introduction of potent cART, there has been a marked reduction in the incidence of KS (Fig. 197-34). The non-AIDS-defining malignancies now account for more morbidity and mortality in patients with HIV infection than the AIDS-defining malignancies and are responsible for approximately 25% of the deaths in patients with HIV infection. Rates of non-Hodgkin's lymphoma have declined; however, this decline has not been as dramatic as the decline in rates of KS. In contrast, cART has had little effect on human papillomavirus (HPV)-associated malignancies. As patients with HIV infection live longer, a wider array of cancers is seen in this population. While some may only reflect known risk factors (e.g., smoking, alcohol consumption, co-infection with other viruses such as hepatitis B) that are increased in patients with HIV infection, some may be a direct consequence of HIV and are clearly increased in patients with lower CD4+ T cell counts.

Kaposi's sarcoma is a multicentric neoplasm consisting of multiple vascular nodules appearing in the skin, mucous membranes, and viscera. The clinical course of KS ranges from indolent, with only minor skin or lymph node involvement, to fulminant, with extensive cutaneous and visceral involvement. In the initial period of the AIDS epidemic, KS was a prominent clinical feature of the first cases of AIDS, occurring in 79% of the patients diagnosed in 1981. By 1989 it was seen in only 25% of cases, by 1992 the number had decreased to 9%, and by 1997 the number was <1%. HHV-8 (KSHV) has been strongly implicated as a viral cofactor in the pathogenesis of KS.

Clinically, KS has varied presentations and may be seen at any stage of HIV infection, even in the presence of a normal CD4+ T cell count. The initial lesion may be a small, raised reddish-purple nodule on the skin **(Fig. 197-42)**, a discoloration on the oral mucosa (Fig. 197-34D), or a swollen lymph node. Lesions often appear in sun-exposed areas, particularly the tip of the nose, and have a propensity to occur in areas of trauma (Koebner phenomenon). Because of the vascular nature of the tumors and the presence of extravasated red blood cells in the lesions, their colors range from reddish to purple to brown and often take the appearance of a bruise, with yellowish discoloration and tattooing. Lesions range in size from a few millimeters to several centimeters in diameter and may be either discrete or confluent. KS lesions most commonly appear as raised macules; however, they can also be papular, particularly in patients with higher CD4+ T cell counts. Confluent lesions may give rise to surrounding lymphedema and may be disfiguring when they involve the face and disabling when they involve the lower extremities or the surfaces of joints. Apart from skin, the lymph nodes, GI tract, and lung are the organ systems most commonly affected by KS. Lesions have been reported in virtually every organ, including the heart and the CNS. In contrast to most malignancies, in which lymph node involvement implies metastatic spread and a poor prognosis, lymph node involvement may be seen very early in KS and is of no special clinical significance. In fact, some patients may present with disease limited to the lymph nodes. These are generally patients with relatively intact immune function and thus the patients with the best prognosis. Pulmonary involvement with KS generally presents with shortness of breath. Some 80% of patients with pulmonary KS also have cutaneous lesions. The chest x-ray characteristically shows

bilateral lower lobe infiltrates that obscure the margins of the mediastinum and diaphragm **(Fig. 197-43)**. Pleural effusions are seen in 70% of cases of pulmonary KS, a fact that is often helpful in the differential diagnosis. GI involvement is seen in 50% of patients with KS and usually takes one of two forms: (1) mucosal involvement, which may lead to bleeding that can be severe; these patients sometimes also develop symptoms of GI obstruction if lesions become large; and (2) biliary tract involvement. KS lesions may infiltrate the gallbladder and biliary tree, leading to a clinical picture of obstructive jaundice similar to that seen with sclerosing cholangitis. Several staging systems have been proposed for KS. One in common use was developed by the National Institute of Allergy and Infectious Diseases AIDS Clinical Trials Group; it distinguishes patients on the basis of tumor extent, immunologic function, and presence or absence of systemic disease **(Table 197-17)**.

TABLE 197-17

National Institute of Allergy and Infectious Diseases AIDS Clinical Trials Group TIS Staging System For Kaposi's Sarcoma

Parameter	Good Risk (Stage 0): All of the Following	Poor Risk (Stage 1): Any of the Following
Tumor (T)	Confined to skin and/or lymph nodes and/or minimal oral disease	Tumor-associated edema or ulceration Extensive oral lesions GI lesions Nonnodal visceral lesions
Immune system (I)	CD4+ T cell count ≥200/μL	CD4+ T cell count <200/μL
Systemic illness (S)	No B symptoms[a]	B symptoms[a] present
	Karnofsky performance status ≥70	Karnofsky performance status <70
	No history of opportunistic infection, neurologic disease, lymphoma, or thrush	History of opportunistic infection, neurologic disease, lymphoma, or thrush

[a]Defined as unexplained fever, night sweats, >10% involuntary weight loss, or diarrhea persisting for more than 2 weeks.

FIGURE 197-42

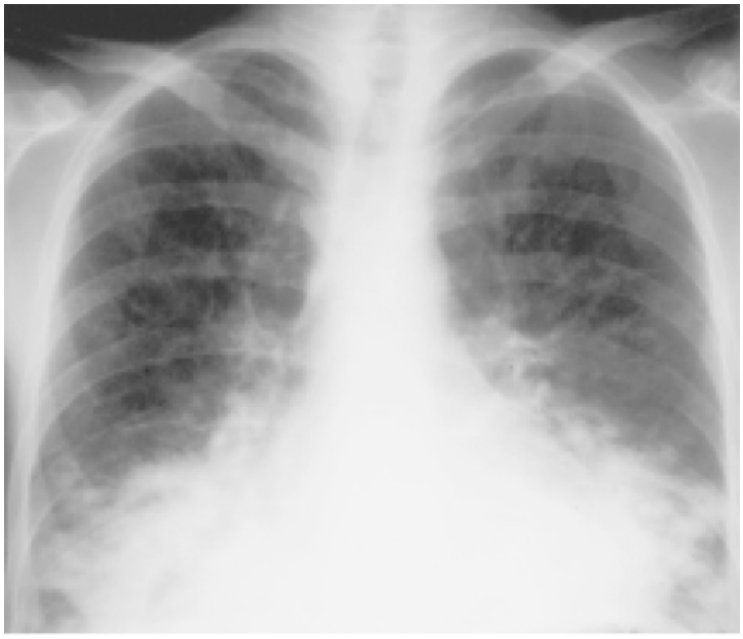
Kaposi's sarcoma in three patients with AIDS demonstrating *(**A**)* periorbital edema and bruising; *(**B**)* classic truncal distribution of lesions; and *(**C**)* upper extremity lesions.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 197-43

Chest x-ray of a patient with AIDS and pulmonary Kaposi's sarcoma. The characteristic findings include dense bilateral lower lobe infiltrates obscuring the heart borders and pleural effusions.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A diagnosis of KS is based on biopsy of a suspicious lesion. Histologically one sees a proliferation of spindle cells and endothelial cells, extravasation of red blood cells, hemosiderin-laden macrophages, and, in early cases, an inflammatory cell infiltrate. Included in the differential diagnosis are lymphoma (particularly for oral lesions), bacillary angiomatosis, and cutaneous mycobacterial infections.

Management of KS **(Table 197-18)** should be carried out in consultation with an expert since definitive treatment guidelines do not exist. In the majority of cases, effective cART will go a long way in achieving control. Antiretroviral therapy has been associated with the spontaneous regression of KS lesions. Paradoxically, it has also been associated with the initial appearance of KS as a form of IRIS. For patients in whom tumor persists or is compromising vital functions or in whom control of HIV replication is not possible, a variety of options exist. In some cases, lesions remain quite indolent, and many of these patients can be managed with no specific treatment. Fewer than 10% of AIDS patients with KS die as a consequence of their malignancy, and death from secondary infections is considerably more common. Thus, whenever possible one should avoid treatment regimens that may further suppress the immune system and increase susceptibility to opportunistic infections. Treatment is indicated under two main circumstances. The first is when a single lesion or a limited number of lesions are causing significant discomfort or cosmetic problems, such as with prominent facial lesions, lesions overlying a joint, or lesions in the oropharynx that interfere with swallowing or breathing. Under these circumstances, treatment with localized radiation, intralesional vinblastine, topical 9-*cis*-retinoic acid, or cryotherapy may be helpful. It should be noted that patients with HIV infection are particularly sensitive to the side effects of radiation therapy. This is especially true with respect to the development of radiation-induced mucositis; doses of radiation directed at mucosal surfaces, particularly in the head and neck region, should be adjusted accordingly. The use of systemic therapy, either IFN-α or chemotherapy, should be considered in patients with a large number of lesions or in patients with visceral involvement. The single most important determinant of response appears to be the CD4+ T cell count. This relationship between response rate and baseline CD4+ T cell count is particularly true for IFN-α. The response rate to IFN-α for patients with CD4+ T cell counts >600/μL is ~80%, while the response rate for patients with counts <150/μL is <10%. In contrast to the other systemic therapies, IFN-α provides an added advantage of having antiretroviral activity; thus, it may be the appropriate first choice for single-agent systemic therapy for early patients with disseminated disease. A variety of chemotherapeutic agents also have been shown to have activity against KS. Four of them—liposomal daunorubicin, liposomal doxorubicin, vinblastine, and paclitaxel—have been approved by the FDA for this indication. Liposomal daunorubicin is approved as first-line therapy for patients with advanced KS. It has fewer side effects than conventional chemotherapy. In contrast, liposomal doxorubicin and paclitaxel are approved only for KS patients who have failed standard chemotherapy. Response rates vary from 23% to 88%, appear to be comparable to what had been achieved earlier with combination chemotherapy regimens, and are greatly influenced by CD4+ T cell count.

TABLE 197-18

Management of AIDS-Associated Kaposi's Sarcoma

Observation and optimization of antiretroviral therapy
Single or limited number of lesions
Radiation
Intralesional vinblastine
Cryotherapy
Extensive disease
Initial therapy
Interferon α (if CD4+ T cells >150/μL)
Liposomal daunorubicin
Subsequent therapy
Liposomal doxorubicin
Paclitaxel
Combination chemotherapy with low-dose doxorubicin, bleomycin, and vinblastine (ABV)
Targeted radiation

Lymphomas occur with an increased frequency in patients with congenital or acquired T cell immunodeficiencies **(Chap. 344)**. AIDS is no exception; at least 6% of all patients with AIDS develop lymphoma at some time during the course of their illness. This is a 120-fold increase in incidence compared with the general population. In contrast to the situation with KS, primary CNS lymphoma, and most opportunistic infections, the incidence of AIDS-associated systemic lymphomas has not experienced a dramatic decrease as a consequence of the widespread use of effective cART. Lymphoma occurs in all risk groups, with the highest incidence in patients with hemophilia and the lowest incidence in patients from the Caribbean or Africa with heterosexually acquired infection. Lymphoma is a late manifestation of HIV infection, generally occurring in patients with CD4+ T cell counts <200/μL. As HIV disease progresses, the risk of lymphoma increases. The attack rate for lymphoma increases exponentially with increasing duration of HIV infection and decreasing level of immunologic function. At 3 years following a diagnosis of HIV infection, the risk of lymphoma is 0.8% per year; by 8 years after infection, it is 2.6% per year. As individuals with HIV infection live longer as a consequence of improved cART and better treatment and prophylaxis of opportunistic infections, it is anticipated that the incidence of lymphomas may increase.

Three main categories of lymphoma are seen in patients with HIV infection: grade III or IV immunoblastic lymphoma, Burkitt's lymphoma, and primary CNS lymphoma. Approximately 90% of these lymphomas are B cell in phenotype; more than half contain EBV DNA. Some are associated with KSHV. These tumors may be either monoclonal or oligoclonal in nature and are probably in some way related to the pronounced polyclonal B cell activation seen in patients with AIDS.

Immunoblastic lymphomas account for ~60% of the cases of lymphoma in patients with AIDS. The majority of these are diffuse large B cell lymphomas (DLBCL). They are generally high grade and would have been classified as diffuse histiocytic lymphomas in earlier classification schemes. This tumor is more common in older patients, increasing in incidence from 0% in HIV-infected individuals <1 year old to >3% in those >50 years of age. Two variants of immunoblastic lymphoma that are seen primarily in HIV-infected patients are primary effusion lymphoma (PEL) and its solid variant, plasmacytic lymphoma of the oral cavity. PEL, also referred to as body cavity lymphoma, presents with lymphomatous pleural, pericardial, and/or peritoneal effusions in the absence of discrete nodal or extranodal masses. The tumor cells do not express surface markers for B cells or T cells and are felt to represent a preplasmacytic stage of differentiation. While both HHV-8 and EBV DNA sequences have been found in the genomes of the malignant cells from patients with body cavity lymphoma, KSHV is felt to be the driving force behind the oncogenesis (see above).

Small noncleaved cell lymphoma (*Burkitt's lymphoma*) accounts for ~20% of the cases of lymphoma in patients with AIDS. It is most frequent in patients 10–19 years old and usually demonstrates characteristic c-*myc* translocations from chromosome 8 to chromosome 14 or 22. Burkitt's lymphoma is not commonly seen in the setting of immunodeficiency other than HIV-associated immunodeficiency, and the incidence of this particular tumor is more than 1000-fold higher in the setting of HIV infection than in the general population. In contrast to African Burkitt's lymphoma, where 97% of the cases contain EBV genome, only 50% of HIV-associated Burkitt's lymphomas are EBV-positive.

Primary CNS lymphoma accounts for ~20% of the cases of lymphoma in patients with HIV infection. In contrast to HIV-associated Burkitt's lymphoma, primary CNS lymphomas are usually positive for EBV. In one study, the incidence of Epstein-Barr positivity was 100%. This malignancy does not have a predilection for any particular age group. The median CD4+ T cell count at the time of diagnosis is ~50/μL. Thus, CNS lymphoma generally presents at a later stage of HIV infection than does systemic lymphoma. This may explain, at least in part, the poorer prognosis for this subset of patients.

The clinical presentation of lymphoma in patients with HIV infection is quite varied, ranging from focal seizures to rapidly growing mass lesions in the oral mucosa **(Fig. 197-44)** to persistent unexplained fever. At least 80% of patients present with extranodal disease, and a similar percentage have B-type symptoms of fever, night sweats, or weight loss. Virtually any site in the body may be involved. The most common extranodal site is the CNS, which is involved in approximately one-third of all patients with lymphoma. Approximately 60% of these cases are primary CNS lymphoma. Primary CNS lymphoma generally presents with focal neurologic deficits, including cranial nerve findings, headaches, and/or seizures. MRI or CT generally reveals a limited number (one to three) of 3- to 5-cm lesions **(Fig. 197-45)**. The lesions often show ring enhancement on contrast administration and may occur in any location. Contrast enhancement is usually less pronounced than that seen with toxoplasmosis. Locations that are most commonly involved with CNS lymphoma are deep in the white matter. The main diseases in the differential diagnosis are cerebral toxoplasmosis and cerebral Chagas' disease. In addition to the 20% of lymphomas in HIV-infected individuals that are primary CNS lymphomas, CNS disease is also seen in HIV-infected patients with systemic lymphoma. Approximately 20% of patients with systemic lymphoma have CNS disease in the form of leptomeningeal involvement. This fact underscores the importance of lumbar puncture in the staging evaluation of patients with systemic lymphoma.

FIGURE 197-44

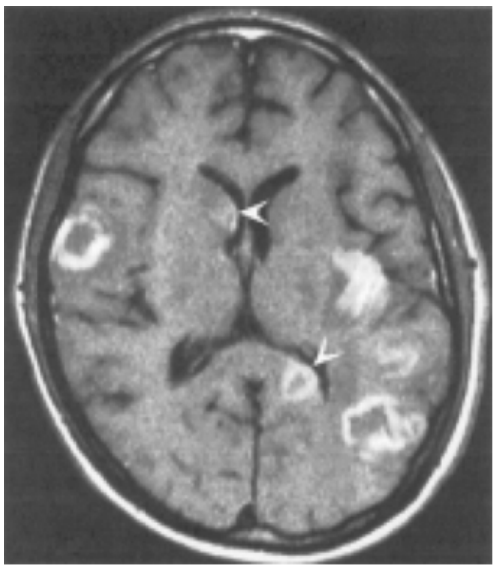
Immunoblastic lymphoma involving the hard palate of a patient with AIDS.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 197-45

Central nervous system lymphoma. Postcontrast T1-weighted MRI scan in a patient with AIDS, altered mental status, and hemiparesis. Multiple enhancing lesions, some ring-enhancing, are present. The left sylvian lesion shows gyral and subcortical enhancement, and the lesions in the caudate and splenium (*arrowheads*) show enhancement of adjacent ependymal surfaces.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Systemic lymphoma is seen at earlier stages of HIV infection than primary CNS lymphoma. In one series the mean CD4+ T cell count was 226/μL. In addition to lymph node involvement, systemic lymphoma may commonly involve the GI tract, bone marrow, liver, and lung. GI tract involvement is seen in ~25% of patients. Any site in the GI tract may be involved, and patients may complain of difficulty swallowing or abdominal pain. The diagnosis is

usually suspected on the basis of CT or MRI of the abdomen. Bone marrow involvement is seen in ~20% of patients and may lead to pancytopenia. Liver and lung involvement are each seen in ~10% of patients. Pulmonary disease may present as a mass lesion, multiple nodules, or an interstitial infiltrate.

Both conventional and unconventional approaches have been employed in an attempt to treat HIV-related lymphomas. Systemic lymphoma is generally treated by the oncologist with combination chemotherapy. Earlier disappointing figures are being replaced with more optimistic results for the treatment of systemic lymphoma following the availability of more effective cART and the use of rituximab in CD20+ tumors. While there is some controversy regarding the use of antiretrovirals during chemotherapy, there is no question that their use overall in patients with HIV lymphoma has improved survival. Concerns regarding synergistic bone marrow toxicities with chemotherapy and cART are mitigated with the use of cART regimens that avoid bone marrow–toxic antiretrovirals. As in most situations in patients with HIV disease, those with higher CD4+ T cell counts tend to fare better. Response rates as high as 72% with a median survival of 33 months and disease-free intervals up to 9 years have been reported. Treatment of primary CNS lymphoma remains a significant challenge. Treatment is complicated by the fact that this illness usually occurs in patients with advanced HIV disease. Palliative measures such as radiation therapy provide some relief. The prognosis remains poor in this group, with a 2-year survival of 29%.

Multicentric Castleman's disease is a KSHV-associated lymphoproliferative disorder that is seen with an increased frequency in patients with HIV infection. While not a true malignancy, it shares many features with lymphoma including generalized lymphadenopathy, hepatosplenomegaly, and systemic symptoms of fever, fatigue, and weight loss. Pulmonary symptoms may be seen in ~50% of patients. KS is present in 75–82% of cases. Lymph node biopsies reveal a predominance of interfollicular plasma cells and/or germinal centers with vascularization and an "onionskin" (hyaline vascular) appearance. Prior to the availability of cART, HIV-infected patients with multicentric Castleman's disease had a 15-fold increased risk of developing non-Hodgkin's lymphoma compared with HIV-infected patients in general. Treatment typically involves chemotherapy. Anecdotal reports of success with rituximab suggest that more specific treatment may be successful, although, in one series treatment with rituximab was associated with worsening of coexisting KS. The median survival of patients with treated multicentric Castleman's disease pre-cART was initially reported as 14 months. This has increased to a 2-year survival of more than 90% in the era of cART.

Evidence of infection with *human papillomavirus* (HPV), associated with *intraepithelial dysplasia of the cervix* or *anus*, is approximately twice as common in HIV-infected individuals as in the general population and can lead to intraepithelial neoplasia and eventually invasive cancer. In a series of studies, HIV-infected men were examined for evidence of anal dysplasia, and Papanicolaou (Pap) smears were found to be abnormal in 20–80%. These changes tend to persist and are generally not affected by cART, raising the possibility of a subsequent transition to a more malignant condition. While the incidence of an abnormal Pap smear of the cervix is ~5% in otherwise healthy women, the incidence of abnormal cervical smears in women with HIV infection is 30–60%, and *invasive cervical cancer* is included as an AIDS-defining condition. While only small increases in the absolute numbers of cervical or anal cancers have been seen as a consequence of HIV infection, the relative risk of these conditions when one compares HIV-infected to noninfected men and women is on the order of 10- to 100-fold. Given the high rates of dysplasia and relative risks for cervical and anal cancer, a comprehensive gynecologic and rectal examination, including Pap smear, is indicated at the initial evaluation and 6 months later for all patients with HIV infection. If these examinations are negative at both time points, the patient should be followed with yearly evaluations. If an initial or repeat Pap smear shows evidence of severe inflammation with reactive squamous changes, the next Pap smear should be performed at 3 months. If, at any time, a Pap smear shows evidence of squamous intraepithelial lesions, colposcopic examination with biopsies as indicated should be performed. The 2-year survival rate for HIV-infected patients with invasive cervical cancer is 64% compared with 79% in non-HIV-infected patients. In addition to rectal and cervical lesions, HPV can also lead to head and neck cancers. In one study of men who have sex with men, 25% were found to have oral HPV; high-risk HPV genotypes were three times more common in the HIV-infected men. The most common HPV genotypes in the general population and the genotypes upon which current HPV vaccines are based are 6, 11, 16, and 18. This is not the case in the HIV-infected population, where other genotypes such as 58 and 53 also are prominent. This raises concerns about the level of effectiveness of the current HPV vaccines for HIV-infected patients. Despite this, it is recommended that patients with HIV infection be vaccinated against HPV.

IDIOPATHIC CD4+ T LYMPHOCYTOPENIA

A syndrome was recognized in 1992 characterized by an absolute CD4+ T cell count of <300/μL or <20% of total T cells on a minimum of two occasions at least 6 weeks apart; no evidence of HIV-1, HIV-2, HTLV-1, or HTLV-2 on testing; and the absence of any defined immunodeficiency or therapy associated with decreased levels of CD4+ T cells. By mid-1993, ~100 patients had been described. After extensive multicenter investigations, a series of reports were published in early 1993, which together allowed a number of conclusions. Idiopathic CD4+ lymphocytopenia (ICL) is a very rare syndrome, as determined by studies of blood donors and cohorts of HIV-seronegative men who have sex with men. Cases were clearly identified as early as 1983 and were remarkably similar to the clinical features of ICL that had been identified decades earlier. The definition of ICL based on CD4+ T cell counts coincided with the ready availability of testing for CD4+ T cells in patients suspected of being immunodeficient. Although, as a result of immune deficiency, certain patients with ICL develop some of the opportunistic diseases (particularly cryptococcosis, nontuberculous mycobacterial infections, and cervical dysplasia) seen in HIV-infected patients, the syndrome is demographically, clinically, and immunologically unlike HIV infection and AIDS. Fewer than half of the reported ICL patients had risk factors for HIV infection, and there were wide geographic and age distributions. The fact that a significant proportion of patients did have risk factors probably reflects a selection bias, in that physicians who take care of HIV-infected patients are more likely to monitor CD4+ T cells. Approximately half of the patients are women, compared with approximately one-third among HIV-infected individuals in the United States. Many patients with ICL remained clinically stable, and their condition did not deteriorate progressively as is common

with seriously immunodeficient HIV-infected patients. Approximately 15% of patients with ICL experience spontaneous reversal of the CD4+ T lymphocytopenia. Immunologic abnormalities in ICL are somewhat different from those of HIV infection. ICL patients often have increases in CD4+ T cell activation with decreases in CD8+ T cells and B cells. Furthermore, immunoglobulin levels are either normal or, more commonly, decreased in patients with ICL, compared with the usual hypergammaglobulinemia of HIV-infected individuals. Virologic studies of these patients have revealed no evidence of HIV-1, HIV-2, HTLV-1, or HTLV-2 or of any other mononuclear cell–tropic virus. Furthermore, there has been no epidemiologic evidence to suggest that a transmissible microbe was involved. The cases of ICL have been widely dispersed, with no clustering. Close contacts and sexual partners who were studied were clinically well and were serologically, immunologically, and virologically negative for HIV. ICL is a heterogeneous syndrome, and it is highly likely that there is no common cause; however, there may be common causes among subgroups of patients that are currently unrecognized.

Patients who present with laboratory data consistent with ICL should be worked up for underlying diseases that could be responsible for the immune deficiency. If no underlying cause is detected, no specific therapy should be initiated. However, if opportunistic diseases occur, they should be treated appropriately (see above). Depending on the level of the CD4+ T cell count, patients should receive prophylaxis for the commonly encountered opportunistic infections.

TREATMENT

TREATMENT

AIDS and Related Disorders

GENERAL PRINCIPLES OF PATIENT MANAGEMENT

The CDC guidelines call for the testing for HIV infection to be a part of routine medical care. It is recommended that the patient be informed of the intention to test, as is the case with other routine laboratory determinations, and be given the opportunity to “opt out.” Such an approach is critical to the goal of identifying as many infected individuals as possible since 15% of the 1.1 million individuals in the United States who are HIV-infected are not aware of their status. Under these circumstances of routine testing, although it is desirable, pretest counseling may not always be built into the testing process. However, no matter how well prepared a patient is for adversity, the discovery of a diagnosis of HIV infection is a devastating event. Thus, physicians should be sensitive to this fact and, where possible, execute some degree of pretest counseling to at least partially prepare the patient should the results demonstrate the presence of HIV infection. Following a diagnosis of HIV infection, the health care provider should be prepared to immediately activate support systems for the newly diagnosed patient and initiate cART therapy. These supports should include individuals who can spend time talking to the newly diagnosed person and ensuring that he or she is emotionally stable and ready to begin therapy. Most communities have HIV support centers that can be of great help in these difficult situations.

The treatment of patients with HIV infection requires not only a comprehensive knowledge of the possible disease processes that may occur and up-to-date knowledge of and experience with cART, but also the ability to deal with the problems of a chronic, potentially life-threatening illness. A comprehensive knowledge of internal medicine is required to deal with the changing spectrum of illnesses associated with HIV infection, many of which are similar to a state of accelerated aging. Great advances have been made in the treatment of patients with HIV infection. The appropriate use of potent cART and other treatment and prophylactic interventions are of critical importance in providing each patient with the best opportunity to live a long and healthy life with HIV infection. In contrast to the earlier days of this epidemic, a diagnosis of HIV infection need no longer be equated with having an inevitably fatal disease. In addition to medical interventions, the health care provider has a responsibility to provide each patient with appropriate counseling and education concerning their disease as part of a comprehensive care plan. Patients must be educated about the potential transmissibility of their infection and about the fact that while health care providers may refer to levels of the virus as “undetectable,” this is only a reflection of the sensitivity of the assay being used to measure the virus, rather than a comment on the presence or absence of the virus. It is important for patients to be aware that the virus is still present in virtually all patients who have ever been diagnosed with HIV infection and capable of being transmitted at all stages of HIV disease. Thus, there must be frank discussions concerning sexual practices and the sharing of syringes and other paraphernalia used in illicit drug use. The treating physician not only must be aware of the latest medications available for patients with HIV infection but also must educate patients concerning the natural history of their illness and listen and be sensitive to their fears and concerns. As with other diseases, therapeutic decisions should be made in consultation with the patient, when possible, and with the patient’s proxy if the patient is incapable of making decisions. In this regard, it is recommended that all patients with HIV infection, and in particular those with CD4+ T cell counts <200/μL, designate a trusted individual with durable power of attorney to make medical decisions on their behalf, if necessary.

Following a diagnosis of HIV infection, there are several examinations and laboratory studies that should be performed to help determine the extent of disease and provide baseline standards for future reference (Table 197-19). In addition to routine chemistry, fasting lipid profile, aspartate aminotransferase, alanine aminotransferase, total and direct bilirubin, fasting glucose and hematology screening panels, Pap smear, urinalysis, and chest x-ray, one should also obtain a CD4+ T cell count, a plasma HIV RNA level, an HIV resistance test, a rapid plasma reagin or VDRL test, an anti-*Toxoplasma* antibody titer, and serologies for hepatitis A, B, and C. A PPD test or IFN-γ release assay should be done and an MMSE performed and recorded. A pregnancy test should be done in women in whom the drug efavirenz is being considered, and HLA-B5701 testing should be done in all patients in whom the drug abacavir is being considered. Patients should be immunized with pneumococcal polysaccharide, with annual influenza shots, and, if seronegative for these viruses, with HPV, hepatitis A, and hepatitis B vaccines. The status of hepatitis C infection should be determined. In

addition, patients should be counseled with regard to sexual practices and needle sharing, and counseling should be offered to those whom the patient knows or suspects may also be infected. Once these baseline activities are performed, short- and long-term medical management strategies should be developed based on the most recent information available and modified as new information becomes available. The field of HIV medicine is changing rapidly, and it is difficult to remain fully up-to-date. Fortunately there are a series of excellent sites on the Internet that are frequently updated, and they provide the most recent information on a variety of topics, including consensus panel reports on treatment **(Table 197-20)**.

ANTIRETROVIRAL THERAPY

Combination antiretroviral therapy (cART), also referred to as highly active antiretroviral therapy (HAART), is the cornerstone of management of patients with HIV infection. Following the initiation of widespread use of cART in the United States in 1995–1996, marked declines were noted in the incidence of most AIDS-defining conditions (Fig. 197-34). Suppression of HIV replication is an important component in prolonging life as well as in improving the quality of life in patients with HIV infection. Adequate suppression requires strict adherence to prescribed regimens of antiretroviral drugs. This has been facilitated by the coformulations of antiretrovirals and the development of once-daily regimens. Unfortunately, many of the most important questions related to the treatment of HIV disease currently lack definitive answers. Among the decisions that need to be made in the context of prescribing cART are when therapy should be started, selection of the best initial regimen, when a given regimen should be changed, and what it should be changed to when a change is made. The care provider and patient must come to a mutually agreeable plan based on the best available data. In an effort to facilitate this process, the U.S. Department of Health and Human Services makes available on the Internet (*www.aidsinfo.nih.gov*) a series of periodically updated guidelines, including *"Guidelines for the Use of Antiretroviral Agents in HIV-Infected Adults and Adolescents*" and "*Guidelines for the Prevention of Opportunistic Infections in Persons Infected with Human Immunodeficiency Virus*." At present, an extensive clinical trials network, involving both clinical investigators and patient advocates, is in place attempting to develop improved approaches to therapy. Consortia comprising representatives of academia, industry, independent foundations, and the federal government are involved in the process of drug development, including a wide-ranging series of clinical trials. As a result, new therapies and new therapeutic strategies are continually emerging. New drugs are often available through expanded-access programs prior to official licensure. Given the complexity of this field, decisions regarding cART are best made in consultation with experts.

Currently available drugs for the treatment of HIV infection as part of a combination regimen fall into four categories: those that inhibit the viral reverse transcriptase enzyme (nucleoside and nucleotide reverse transcriptase inhibitors; nonnucleoside reverse transcriptase inhibitors), those that inhibit the viral protease enzyme (protease inhibitors), those that inhibit the viral integrase enzyme (integrase inhibitors), and those that interfere with viral entry (fusion inhibitors; CCR5 antagonists) **(Table 197-21; Fig. 197-46)**. Numerous formulations combining two or more of these antiretroviral drugs have been licensed **(Table 197-22)**. Prior to initiation of therapy and at any time a change in therapy due to treatment failure is being considered, drug resistance testing should be performed to help guide the selection of drugs to be used in combination. A summary of known resistance mutations for antiretroviral drugs is shown in **Fig. 197-47**.

The FDA-approved reverse transcriptase inhibitors include the *nucleoside analogues* zidovudine, didanosine, zalcitabine, stavudine, lamivudine, abacavir, and emtricitabine; the *nucleotide analogues* tenofovir disoproxil and tenofovir alafenamide; and the *nonnucleoside reverse transcriptase inhibitors* nevirapine, delavirdine, efavirenz, etravirine, and rilpivirine (Table 197-21). These represent the first class of drugs licensed for the treatment of HIV infection. They are indicated for this use as part of combination regimens. It should be stressed that none of these drugs should be used as monotherapy for HIV infection due to the relative ease with which drug resistance may develop under such circumstances. Thus, when lamivudine, emtricitabine, or tenofovir is used to treat hepatitis B infection in the setting of HIV infection, one should ensure that the patient is also on additional antiretroviral medication. The reverse transcriptase inhibitors block the HIV replication cycle at the point of RNA-dependent DNA synthesis, the reverse transcription step. While the nonnucleoside reverse transcriptase inhibitors are quite selective for the HIV-1 reverse transcriptase, the nucleoside and nucleotide analogues inhibit a variety of DNA polymerases in addition to those of the HIV-1 reverse transcriptase. For this reason, serious side effects are more varied with the nucleoside analogues and include mitochondrial damage that can lead to hepatic steatosis and lactic acidosis as well as peripheral neuropathy and pancreatitis. The use of either of the thymidine analogues zidovudine and stavudine has been associated with a syndrome of hyperlipidemia, glucose intolerance/insulin resistance, and fat redistribution often referred to as *lipodystrophy syndrome* (discussed in "Diseases of the Endocrine System and Metabolic Disorders," above). The reverse transcriptase inhibitors preferred for use in combination regimens according to the DHHS Panel on the use of antiretroviral drugs are lamivudine, emtricitabine, abacavir, tenofovir disoproxil, and tenofovir alafenamide.

The HIV-1 protease inhibitors (saquinavir, indinavir, ritonavir, nelfinavir, amprenavir, fosamprenavir, lopinavir/ritonavir, atazanavir, tipranavir, and darunavir) are an important part of the therapeutic armamentarium of antiretrovirals. While possessing antiviral properties of its own, ritonavir is typically used as a pharmacokinetic enhancer due to its high affinity for several isoforms of cytochrome P450 (3A4, 2D6) leading to large increases in the plasma concentrations of co-administered drugs metabolized by these pathways As in the case of reverse transcriptase inhibitors, resistance to protease inhibitors can develop rapidly in the setting of monotherapy, and thus these agents should be used only as part of combination therapeutic regimens. Based upon superior efficacy and side-effect profile, ritonavir-boosted darunavir in combination with emtricitabine and tenofovir disoproxil or tenofovir alafenamide is the protease inhibitor strategy preferred for initial therapy according to the DHHS Panel on the use of antiretroviral drugs.

Integrase inhibitors act by blocking the action of the HIV integrase enzyme and thus preventing integration of the HIV provirus into the host cell genome. They are among the most potent and safest of the antiretroviral drugs and frequently part of initial combination regimens. The four licensed integrase inhibitors are raltegravir, elvitegravir, dolutegravir, and bictegravir. Elvitegravir is always given in combination with cobicistat, which acts to boost the

concentrations of elvitegravir. Cobicistat also inhibits tubular secretion of creatinine, resulting in increases in serum creatinine, and is not recommended for patients with estimated creatinine clearances <70 mL/min. Bictegravir is available only in combination with tenofovir alafenamide and emtricitabine.

Entry inhibitors act by interfering with the binding of HIV to its receptor or co-receptor or by interfering with the process of fusion (see above). The first drug in this class to be licensed was the fusion inhibitor *enfuvirtide*, or T-20, followed by the CCR5 antagonist *maraviroc*. The anti-CD4 monoclonal antibody ibalizumab was licensed in 2018, and a variety of additional small molecules that bind to HIV-1 co-receptors are currently in clinical trials.

PRINCIPLES OF THERAPY

The principles of therapy for HIV infection have been articulated by a panel sponsored by the U.S. Department of Health and Human Services as a working group of the NIH Office of AIDS Research Advisory Council. These principles are summarized in **Table 197-23**. As noted in these guidelines, cART of HIV infection does not lead to eradication or cure of HIV. The single possible exception to this is an individual with HIV infection who received an allogeneic stem cell transplant for treatment of acute myelogenous leukemia. His conditioning regimen included cytotoxic chemotherapy, total-body irradiation, and antithymocyte immunoglobulin. The donor cells were homozygous for the *CCR5Δ32* mutation (see above) and thus resistant to HIV infection. Despite cART being stopped the day of the transplant, the patient has exhibited no signs of active HIV infection for more than 8 years.

Treatment decisions must take into account the fact that one is dealing with a chronic infection that requires daily therapy. Patients initiating antiretroviral therapy must be willing to commit to life-long treatment and understand the importance of adherence to their prescribed regimen. The importance of adherence is illustrated by the observation that treatment interruption is associated with rapid increases in HIV RNA levels, rapid declines in CD4+ T cell counts, and an increased risk of clinical progression. While it seems reasonable to assume that the complications associated with cART could be minimized by intermittent treatment regimens designed to minimize exposure to the drugs in question, all efforts to do so have paradoxically been associated with an increase in serious adverse events in the patients randomized to intermittent therapy, suggesting that some "non-AIDS-associated" serious adverse events such as heart attack and stroke may be linked to HIV replication. Thus, unless contraindicated for reasons of toxicity, patients started on cART should remain on cART.

At present, the U.S. Department of Health and Human Services Guidelines panel recommends that everyone with HIV infection be treated with cART and that therapy be initiated a soon as possible after diagnosis. Therapy has been associated with a decrease in disease progression in patients at all stages of HIV infection and leads to a decrease in the risk of transmission of infection. In addition, one may wish to administer a 6-week course of therapy to uninfected individuals immediately following a high-risk exposure to HIV. The combination of tenofovir and emtricitabine is also indicated for pre-exposure prophylaxis in individuals at high risk of HIV infection. For patients diagnosed with an opportunistic infection and HIV infection at the same time, one may consider a 2- to 4-week delay in the initiation of antiretroviral therapy during which time treatment is focused on the opportunistic infection. This delay may decrease the severity of any subsequent immune reconstitution inflammatory syndrome by lowering the antigenic burden of the opportunistic infection. This is particularly true for patients with TB or cryptococcal infections. For patients with advanced HIV infection (CD4+ <50 cells/μL), however, cART should be initiated as soon as possible.

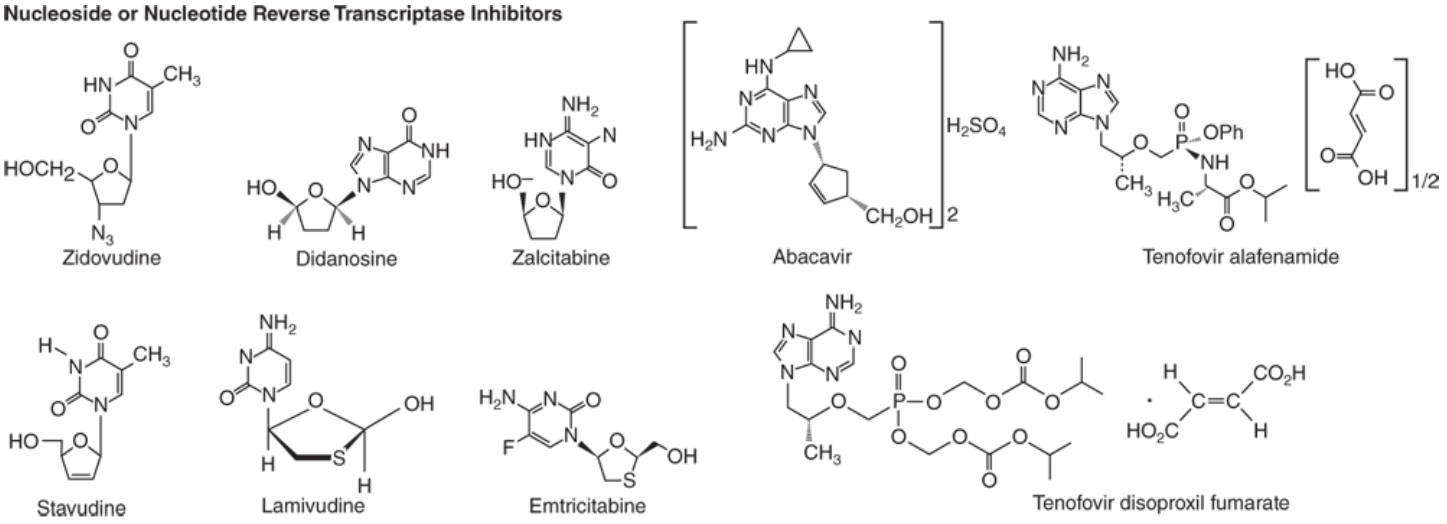
Once the decision has been made to initiate therapy, the health care provider must decide which drugs to use as the first regimen. The decision regarding choice of drugs not only will affect the immediate response to therapy but also will have implications regarding options for future therapeutic regimens. The initial regimen is usually the most effective insofar as the virus has yet to develop significant resistance. HIV is capable of rapidly developing resistance to any single agent, and therapy must be given as a multidrug combination. Given that patients can be infected with viruses that harbor drug resistance mutations, it is recommended that a viral genotype be done prior to the initiation of therapy to optimize the selection of antiretroviral agents. The combination regimens currently recommended for initial therapy in most treatment-naïve patients are listed in **Table 197-24**. It is currently debated whether treatment-naïve individuals with <50 copies/mL of HIV RNA benefit from cART. While these individuals are at low risk of disease progression in the short term, they do have evidence of persistent immune activation that may have long-term consequences. Following the initiation of therapy one should expect a rapid, at least 1-log (tenfold) reduction in plasma HIV RNA levels within 1–2 months and then a slower decline in plasma HIV RNA levels to <50 copies/mL within 6 months. During this same time there should be a rise in the CD4+ T cell count of 100–150/cells μL that is also particularly brisk during the first month of therapy. Subsequently, one should anticipate a CD4+ T cell count increase of 50–100 cells/year until numbers approach normal. Many clinicians feel that failure to achieve these endpoints is an indication for a change in therapy. Other reasons for a change in therapy include a persistently declining CD4+ T cell count, a consistent increase in HIV RNA levels to >200 copies/mL, clinical deterioration, or drug toxicity **(Table 197-25)**. As in the case of initiating therapy, changing therapy may have a lasting impact on future therapeutic options. When changing therapy because of treatment failure (clinical progression or worsening laboratory parameters), it is important to attempt to provide a regimen with at least two new active drugs. This decision can be guided by resistance testing (see below). In the patient in whom a change is made for reasons of drug toxicity, a simple replacement of one drug is reasonable. It should be stressed that in attempting to sort out a drug toxicity it may be advisable to hold all therapy for a period of time to distinguish between drug toxicity and disease progression. Drug toxicity will usually begin to show signs of reversal within 1–2 weeks. Prior to changing a treatment regimen because of drug failure, it is important to ensure that the patient has been adherent to the prescribed regimen. As in the case of initial therapy, the simpler the new therapeutic regimen, the easier it is for the patient to be compliant. Plasma HIV RNA levels should be monitored every 3–6 months during therapy and more frequently if one is contemplating a change in regimen due to an increase in viral load or immediately following a change in regimen.

In order to determine an optimal therapeutic regimen for initial therapy or for a patient on a failing regimen, one may attempt to measure antiretroviral drug susceptibility through genotyping or phenotyping of HIV quasispecies and to determine adequacy of dosing through measurement of drug levels. Genotyping may be done through cDNA sequencing. Phenotypic assays typically measure the enzymatic activity of viral enzymes in the presence or absence of different concentrations of different drugs and have also been used to determine co-receptor tropism. These assays will generally detect quasispecies present at a frequency of ≥10%. Next-generation sequencing may allow detection of quasispecies at frequencies down to 1%. It is generally recommended that resistance testing be used in selecting initial therapy in settings where the risk of transmission of resistant virus is high (such as the United States and Europe) and in determining new regimens for patients experiencing virologic failure while on therapy. Resistance testing may be of particular value in distinguishing drug-resistant virus from poor patient compliance. Due to the rapid rate at which drug-resistant viruses revert to wild-type, it is recommended that resistance testing performed in the setting of drug failure be carried out while the patient is still on the failing regimen. Measurement of plasma drug levels can also be used to tailor an individual treatment. The inhibitory quotient, defined as the trough blood level/IC_{50} of the patient's virus, is used by some to determine the adequacy of dosing of a given treatment regimen. Despite the best of efforts there will still be patients with ongoing high levels of HIV replication while receiving the best available therapy. These patients will receive benefit from remaining on antiretroviral therapy even though it is not fully suppressive.

In addition to the licensed medications discussed above, a large number of experimental agents are being evaluated as possible therapies for HIV infection. Therapeutic strategies are being developed to interfere with virtually every step of the replication cycle of the virus (Fig. 197-3) and in an attempt to eliminate the reservoir of infected cells to "cure" HIV infection. In addition to directly acting antiviral drugs, other strategies, generically referred to as "immune-based therapies," are being developed as a complement to antiviral therapy. Among the antiviral agents in early clinical trials are additional nucleoside and nucleotide analogues, protease inhibitors, fusion inhibitors, receptor and co-receptor antagonists, and integrase inhibitors—as well as new antiviral strategies including antisense nucleic acids and maturation inhibitors. Among the immune-based therapies being evaluated are monoclonal antibodies, IFN-α, bone marrow transplantation, adoptive transfer of lymphocytes genetically modified to resist infection or enhance HIV-specific immunity, active immunotherapy with inactivated HIV or its components, IL-7, and IL-15. Strategies directed toward cure are examining the role of latency-reversing agents such as histone-deacetylase inhibitors.

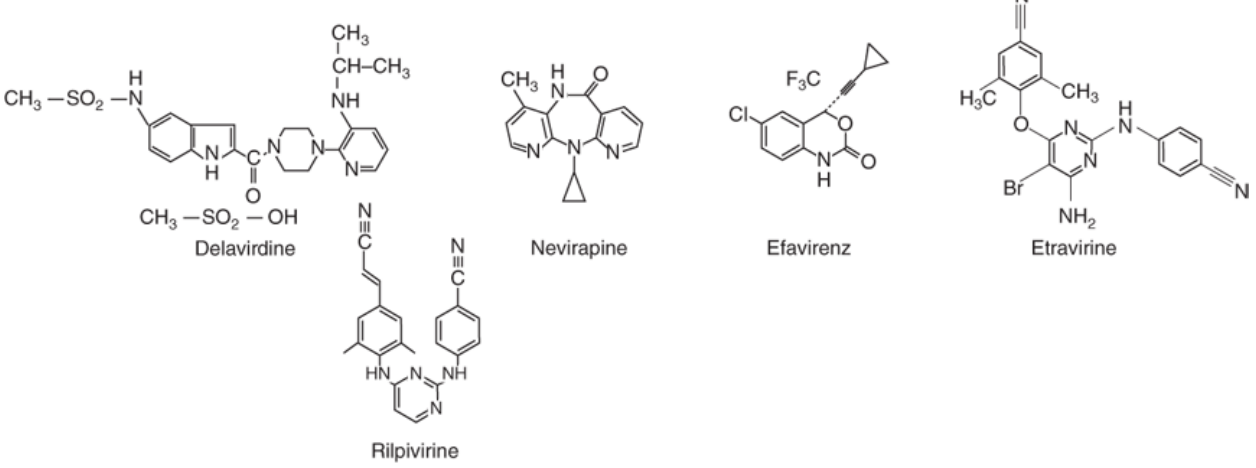
FIGURE 197-46

Molecular structures of antiretroviral agents.



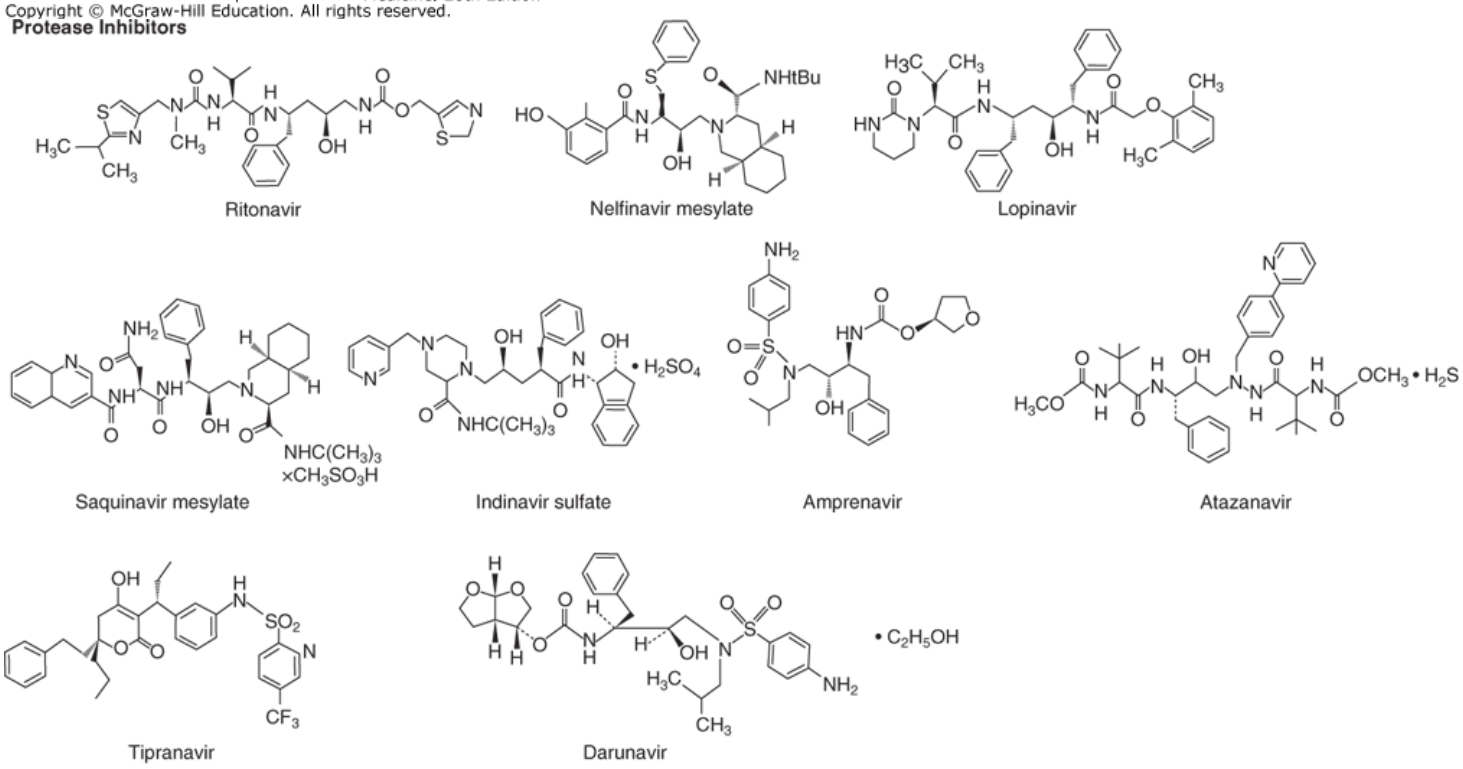
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Nonnucleoside Reverse Transcriptase Inhibitors



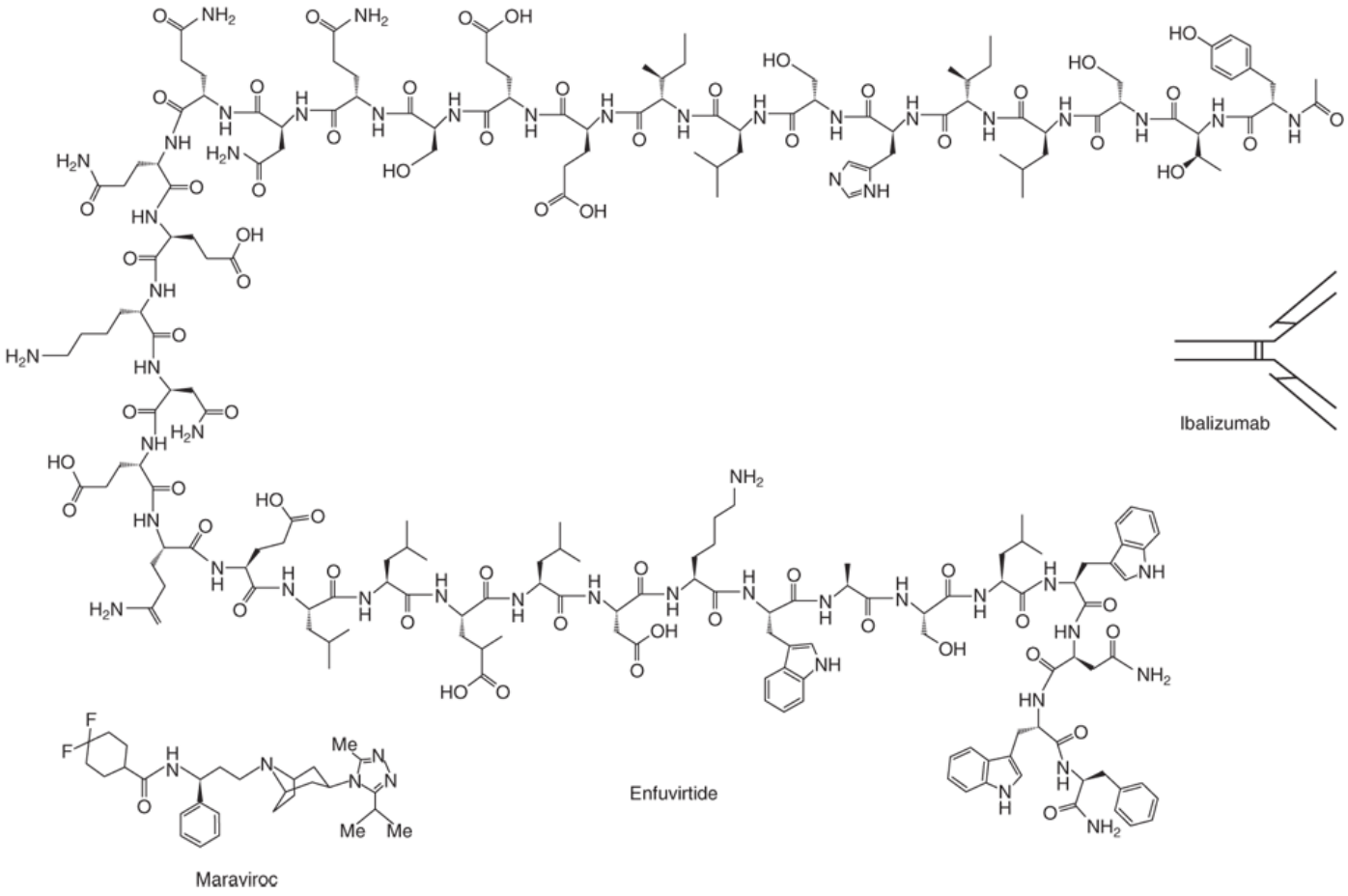
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Protease Inhibitors



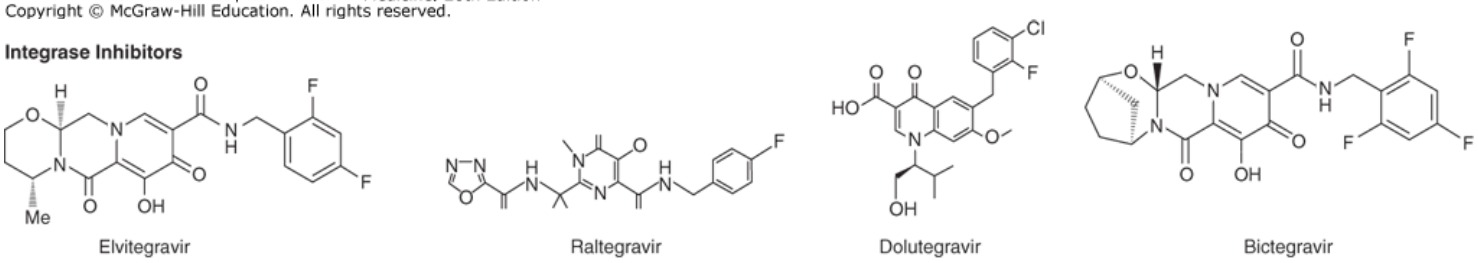
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Entry Inhibitors



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Integrase Inhibitors



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 197-47

Amino acid substitutions conferring resistance to antiretroviral drugs. For each amino acid residue, the letter above the bar indicates the amino acid associated with wild-type virus and the letter(s) below indicate the substitution(s) that confer viral resistance. The number shows the position of the mutation in the protein. Mutations selected by protease inhibitors in Gag cleavage sites are not listed. HR1, first heptad repeat; NAMs, nRTI-associated mutations; nRTI, nucleoside reverse transcriptase inhibitor; NNRTI, nonnucleoside reverse transcriptase inhibitor; PI, protease inhibitor. Amino acid abbreviations: A, alanine; C, cysteine; D, aspartate; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine. *(Reprinted with permission from the International Antiviral Society—USA. AM Wensing, V Calvez, HR Günthard et al: 2014 Update of the Drug Resistance Mutations in HIV-1. Top Antivir Med 22:642, 2014. Updated information [and thorough explanatory notes] available at www.iasusa.org.)*

MUTATIONS IN THE REVERSE TRANSCRIPTASE GENE ASSOCIATED WITH RESISTANCE TO REVERSE TRANSCRIPTASE INHIBITORS

Nucleoside and Nucleotide Analogue Reverse Transcriptase Inhibitors (nRTIs)[a]

69 Insertion Complex[b] (affects all nRTIs currently approved by the US FDA)

	M	A	▼	K	L	T	K
Multi-nRTI Resistance	41 L	62 V	69 Insert	70 R	210 W	215 Y F	219 Q E

151 Complex[c] (affects all nRTIs currently approved by the US FDA except tenofovir)

	A	V	F	F	Q
Multi-nRTI Resistance	62 V	75 I	77 L	116 Y	151 M

Thymidine Analogue-Associated Mutations[d,e] (TAMs; affect all nRTIs currently approved by the US FDA other than emtricitabine and lamivudine)

	M	D	K	L	T	K
Multi-nRTI Resistance	41 L	67 N	70 R	210 W	215 Y F	219 Q E

	K	L	Y	M
Abacavir[f,g]	65 R E N	74 V	115 F	184 V
Didanosine[g,h]	65 R E N	74 V		
Emtricitabine	65 R E N			184 V I
Lamivudine	65 R E N			184 V I

	M	K	D	K	L	T	K
Stavudine[d,e,g,i,j,k]	41 L	65 R E N	67 N	70 R	210 W	215 Y F	219 Q E

	K	K
Tenofovir[l]	65 R E N	70 E

	M	D	K	L	T	K
Zidovudine[d,e,j,k]	41 L	67 N	70 R	210 W	215 Y F	219 Q E

Nonnucleoside Analogue Reverse Transcriptase Inhibitors (NNRTIs)[a,m]

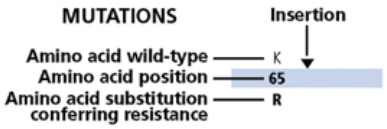
	L	K	K	V	V	Y	Y	G	P	M
Efavirenz	100 I	101 P	103 N S	106 M	108 I	181 C I	188 L	190 S A	225 H	230 L

	V	A	L	K	V	E	V	Y	G	M
Etravirine[n]	90 I	98 G	100 I	101 E H P	106 I	138 A G K Q	179 D F T	181 C I V	190 S A	230 L

	L	K	K	V	V	Y	Y	G	M
Nevirapine	100 I	101 P	103 N S	106 A M	108 I	181 C I	188 C L H	190 A	230 L

	L	K	E	V	Y	Y	H	F	M
Rilpivirine[o]	100 I	101 E P	138 A G K Q R	179 L	181 C I V	188 L	221 Y	227 C	230 I L

Amino acid abbreviations: A, alanine; C, cysteine; D, aspartate; E, glutamate; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MUTATIONS IN THE PROTEASE GENE ASSOCIATED WITH RESISTANCE TO PROTEASE INHIBITORS[p,q,r]

	L 10	G 16	K 20	L 24	V 32	L 33	E 34	M 36	M 46	G 48	I 50	F 53	I 54	D 60	I 62	I 64	A 71	G 73	V 82	I 84	I 85	N 88	L 90	I 93
Atazanavir +/- ritonavir[s]	I F V C	E	R M T V	I	I	I F V	Q	I L V	I L	V	L	L Y	L V M T A	E	V	L M V	V I T L	C S T A	A T F I	V	V	S	M	L M
Darunavir/ ritonavir[t]	V 11 I				V 32 I	L 33 F			I **47** V		I **50** V		I **54** M L					T 74 P	L **76** V	I **84** V		L 89 V		
Fosamprenavir/ ritonavir	L 10 F I R V				V 32 I				M 46 I L	I 47 V	I **50** V		I 54 L V M				G 73 S		L 76 V	V 82 A F S T	I **84** V		L 90 M	
Indinavir/ ritonavir[u]	L 10 I R V		K 20 M R	L 24 I	V 32 I			M 36 I	M **46** I L				I 54 V				A 71 V T	G 73 S A	L 76 V	V 77 I	V **82** A F T	I **84** V		L 90 M	
Lopinavir/ ritonavir[v]	L 10 F I R V		K 20 M R	L 24 I	V **32** I	L 33 F			M 46 I L	I **47** V A	I 50 V	F 53 L	I 54 V L A M T S		L 63 P		A 71 V T	G 73 S	L **76** V	V **82** A F T S	I 84 V		L 90 M	
Nelfinavir[u,w]	L 10 F I			D **30** N		M 36 I			M 46 I L								A 71 V T		V 77 I	V 82 A F T S	I 84 V	N 88 D S	L **90** M	
Saquinavir/ ritonavir[u]	L 10 I R V			L 24 I						G **48** V			I 54 V L		I 62 V		A 71 V T	G 73 S	V 77 I	V 82 A F T S	I 84 V		L **90** M	
Tipranavir/ ritonavir	L 10 V				L 33 F	M 36 I L V		K 43 T	M 46 L	I **47** V			I 54 A M V	Q **58** E		H 69 K R		T **74** P		V **82** L T	N 83 D	I **84** V	L 89 I M V		

MUTATIONS IN THE ENVELOPE GENE ASSOCIATED WITH RESISTANCE TO ENTRY INHIBITORS

	G 36	I 37	V 38	Q 39	Q 40	N 42	N 43
Enfuvirtide[x]	**D S**	**V**	**A M E**	**R**	**H**	**T**	**D**

Maraviroc[y] See User Note

MUTATIONS IN THE INTEGRASE GENE ASSOCIATED WITH RESISTANCE TO INTEGRASE STRAND TRANSFER INHIBITORS[z]

	T 66	L 74	E 92	T 97	F 121	E 138	G 140	Y 143	S 147	Q 148	N 155	R 263
Dolutegravir[aa]					F 121 Y	E 138 A K	G 140 A S			Q **148** **H K R**	N 155 H	R 263 K
Elvitegravir[bb]	T **66** **I** A K		E **92** **Q** G	T 97 A	F **121** **Y**				S **147** **G**	Q **148** **H K R**	N **155** **H**	R 263 K
Raltegravir[cc]		L 74 M	E 92 Q	T 97 A	F **121** **Y**	E 138 A K	G 140 A S	Y **143** **R H C**		Q **148** **H K R**	N **155** **H**	R 263 K

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 197-19

Initial Evaluation of the Patient with HIV Infection

History and physical examination
Routine chemistry and hematology
AST, ALT, alkaline phosphatase, direct and indirect bilirubin
Lipid profile and fasting glucose
CD4+ T lymphocyte count
Plasma HIV RNA level
HIV resistance testing
HLA-B5701 screening
RPR or VDRL test
Anti-*Toxoplasma* antibody titer
Urinalysis
PPD skin test or IFN-γ release assay
Mini-Mental Status Examination
Serologies for hepatitis A, hepatitis B, and hepatitis C
Immunization with pneumococcal polysaccharide; influenza; HPV as indicated
Immunization with hepatitis A and hepatitis B if seronegative
Counseling regarding natural history and transmission
Help contacting others who might be infected

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; PPD, purified protein derivative; RPR, rapid plasma reagin; VDRL, Venereal Disease Research Laboratory.

TABLE 197-20

HIV Disease Resources Available on the World Wide Web

www.aidsinfo.nih.gov	AIDSinfo, a service of the U.S. Department of Health and Human Services, posts federally approved treatment guidelines for HIV and AIDS; provides information on federally funded and privately funded clinical trials and CDC publications and data
www.cdcnpin.org	Updates on epidemiologic data and prevention information from the CDC

Abbreviation: CDC, Centers for Disease Control and Prevention.

TABLE 197-21

Antiretroviral Drugs Most Commonly Used in the Treatment of HIV Infection

Drug	Status	Indication	Dose in Combination	Supporting Data	Toxicity
Nucleoside or Nucleotide Reverse Transcriptase Inhibitors					
Zidovudine (AZT, azidothymidine, *Retrovir, 3'azido-3'-deoxythymidine)	Licensed	Treatment of HIV infection in combination with other antiretroviral agents	200 mg q8h or 300 mg bid	19 vs 1 death in original placebo-controlled trial in 281 patients with AIDS or ARC	Anemia, granulocytopenia, myopathy, lactic acidosis, hepatomegaly with steatosis, headache, nausea, nail pigmentation, lipid abnormalities, lipoatrophy, hyperglycemia
		Prevention of maternal-fetal HIV transmission		In pregnant women with CD4+ T cell count ≥200/μL, AZT PO beginning at weeks 14–34 of gestation plus IV drug during labor and delivery plus PO AZT to infant for 6 weeks decreased transmission of HIV by 67.5% (from 25.5% to 8.3%); n = 363	
Lamivudine (Epivir, 2'3'-dideoxy-3'-thiacytidine, 3TC)	Licensed	In combination with other antiretroviral agents for the treatment of HIV infection	150 mg bid 300 mg qd	In combination with AZT superior to AZT alone with respect to changes in CD4+ T cell counts in 495 patients who were zidovudine-naïve and 477 patients who were zidovudine-experienced; overall CD4+ T cell counts for the zidovudine group were at baseline by 24 weeks, while in the group treated with zidovudine plus lamivudine, they were 10–50 cells/μL above baseline; 54% decrease in progression to AIDS/death compared with AZT alone	Flare of hepatitis in HBV-co-infected patients who discontinue drug
Emtricitabine (FTC, Emtriva)	Licensed	In combination with other antiretroviral agents for the treatment of HIV infection	200 mg qd	Comparable to lamivudine in combination with stavudine and nevirapine/efavirenz	Hepatotoxicity in HBV-co-infected patients who discontinue drug, skin discoloration
Abacavir (Ziagen)	Licensed	For treatment of HIV infection in combination with other antiretroviral agents	300 mg bid	Abacavir + AZT + 3TC equivalent to indinavir + AZT + 3TC with regard to viral load suppression (~60% in each group with <400 HIV RNA copies/mL plasma) and CD4+ T cell increase (~100/μL in each group) at 24 weeks	Hypersensitivity reaction In HLA-B5701+ individuals (can be fatal); fever, rash, nausea, vomiting, malaise or fatigue, and loss of appetite

Drug	Status	Indication	Dose in Combination	Supporting Data	Toxicity
Tenofovir disoproxil fumarate (Viread)	Licensed	For use in combination with other antiretroviral agents when treatment is indicated	300 mg qd	Reduction of ~0.6 log in HIV-1 RNA levels when added to background regimen in treatment-experienced patients	Renal, osteomalacia, flare of hepatitis in HBV-co-infected patients who discontinue drug
Tenofovir alafenamide (Vemlidy)	Licensed	In combination with emtricitabine and other antiretroviral agents for treatment of HIV-1 infection	25 mg qd	92% of patients treated in combination with emtricitabine, elvitegravir, and cobicistat had HIV-1 RNA levels <50 copies/mL	Nausea, less renal toxicity than tenofovir disoproxil fumarate
Non-Nucleoside Reverse Transcriptase Inhibitors					
Nevirapine (Viramune)	Licensed	In combination with other antiretroviral agents for treatment of progressive HIV infection	200 mg/d × 14 days then 200 mg bid *or* 400 mg extended release qd	Increase in CD4+ T cell count, decrease in HIV RNA when used in combination with nucleosides	Skin rash, hepatotoxicity
Efavirenz (Sustiva)	Licensed	For treatment of HIV infection in combination with other antiretroviral agents	600 mg qhs	Efavirenz + AZT + 3TC comparable to indinavir + AZT + 3TC with regard to viral load suppression (a higher percentage of the efavirenz group achieved viral load <50 copies/mL, but the discontinuation rate in the indinavir group was unexpectedly high, accounting for most treatment “failures”); CD4 cell increase (~140/μL in each group) at 24 weeks	Rash, dysphoria, elevated liver function tests, drowsiness, abnormal dreams, depression, lipid abnormalities, potentially teratogenic
Etravirine (Intelence)	Licensed	In combination with other antiretroviral agents in treatment-experienced patients whose HIV is resistant to nonnucleoside reverse transcriptase inhibitors and other antiretroviral medications	200 mg bid	Higher rates of HIV RNA suppression to <50 copies/mL (56% vs 39%); greater increases in CD4+ T cell count (89 vs 64 cells) compared to placebo when given in combination with an optimized background regimen	Rash, nausea, hypersensitivity reactions
Rilpivirine (Edurant)	Licensed	In combination with other drugs in previously untreated patients when treatment is indicated.	25 mg qd	Noninferior to efavirenz with respect to suppression at week 48 in 1368 treatment-naive individuals, except in patients with pretherapy HIV RNA levels >100,000 where it was inferior	Nausea, dizziness, somnolence, vertigo, less CNS toxicity and rash than efavirenz
Protease Inhibitors					

Drug	Status	Indication	Dose in Combination	Supporting Data	Toxicity
Ritonavir (Norvir)	Licensed	In combination with other antiretroviral agents for treatment of HIV infection when treatment is warranted	600 mg bid (also used in lower doses as pharmacokinetic booster)	Reduction in the cumulative incidence of clinical progression or death from 34% to 17% in patients with CD4+ T cell count <100/μL treated for a median of 6 months	Nausea, abdominal pain, hyperglycemia, fat redistribution, lipid abnormalities, may alter levels of many other drugs, paresthesias, hepatitis
Atazanavir (Reyataz)	Licensed	For treatment of HIV infection in combination with other antiretroviral agents	400 mg qd or 300 mg qd + ritonavir 100 mg qd when given with efavirenz	Comparable to efavirenz when given in combination with AZT + 3TC in a study of 810 treatment-naïve patients; comparable to nelfinavir when given in combination with stavudine + 3TC in a study of 467 treatment-naïve patients	Hyperbilirubinemia, PR prolongation, nausea, vomiting, hyperglycemia, fat maldistribution, rash transaminase elevations, renal stones
Darunavir (Prezista)	Licensed	In combination with 100 mg ritonavir for combination therapy in treatment-experienced adults	600 mg + 100 mg ritonavir twice daily with food	At 24 weeks, patients with prior extensive exposure to antiretrovirals treated with a new combination including darunavir showed a –1.89-log change in HIV RNA levels and a 92-cell increase in CD4+ T cells compared with –0.48 log and 17 cells in the control arm	Diarrhea, nausea, headache, skin rash, hepatotoxicity, hyperlipidemia, hyperglycemia
Entry Inhibitors					
Enfuvirtide (Fuzeon)	Licensed	In combination with other agents in treatment-experienced patients with evidence of HIV-1 replication despite ongoing anti-retroviral therapy	90 mg SC bid	In treatment of experienced patients, superior to placebo when added to new optimized background (37% vs 16% with <400 HIV RNA copies/mL at 24 weeks; + 71 vs + 35 CD4+ T cells at 24 weeks)	Local injection reactions, hypersensitivity reactions, increased rate of bacterial pneumonia
Maraviroc (Selzentry)	Licensed	In combination with other antiretroviral agents in adults infected with only CCR5-tropic HIV-1	150–600 mg bid depending on concomitant medications (see text)	At 24 weeks, among 635 patients with CCR5-tropic virus and HIV-1 RNA >5000 copies/mL despite at least 6 months of prior therapy with at least 1 agent from 3 of the 4 antiretroviral drug classes, 61% of patients randomized to maraviroc achieved HIV RNA levels <400 copies/mL compared with 28% of patients randomized to placebo	Hepatotoxicity, nasopharyngitis, fever, cough, rash, abdominal pain, dizziness, musculoskeletal symptoms
Ibalizumab (Trogarzo)	Licensed	In combination with other antiretroviral agents in patients with multidrug-resistant HIV-1	Single loading dose of 2000 mg followed by a maintenance dose of 800 mg every 2 weeks	At 25 weeks, 50% of patients with multi-drug resistant HIV-1 with HIV-1 RNA >1000 copies/mL treated with an optimized background of 1 active drug and ibalizumab achieved HIV RNA levels <200 copies/mL	Rash, diarrhea, nausea
Integrase Inhibitor					

Drug	Status	Indication	Dose in Combination	Supporting Data	Toxicity
Raltegravir (Isentress)	Licensed	In combination with other antiretroviral agents	400 mg bid	At 24 weeks, among 436 patients with 3-class drug resistance, 76% of patients randomized to receive raltegravir achieved HIV RNA levels <400 copies/mL compared with 41% of patients randomized to receive placebo	Nausea, headache, diarrhea, CPK elevation, muscle weakness, rhabdomyolysis
Elvitegravir (Available only in combination with cobicistat, tenofovir, and emtricitabine [Stribild])	Licensed	Fixed-dose combination	1 tablet daily	Noninferior to raltegravir or atazanavir/ritonavir in treatment-experienced patients.	Diarrhea, nausea, upper respiratory infections, headache
Dolutegravir (Tivicay)	Licensed	In combination with other antiretroviral agents	50 mg daily for treatment-naïve patients 50 mg twice daily for treatment-experienced patients or those also receiving efavirenz or rifampin	Noninferior to raltegravir, superior to efavirenz or darunavir/ritonavir	Insomnia, headache, hypersensitivity reactions, hepatotoxicity
Bictegravir (Available only in combination with tenofovir alafenamide and emtricitabine [Biktarvy]	Licensed	For treatment of HIV infection in adults	50 mg bictegravir/25 mg tenofovir alafenamide/200 mg emtricitabine qd	Non-inferior to dolutegravir/tenofovir/emtricitabine and non-inferior to dolutegravir/abacavir/lamivudine	Nausea, diarrhea, headache

*Initial trade names are provided. Generic forms may be available.

Abbreviations: ARC, AIDS-related complex; NRTIs, nonnucleoside reverse transcriptase inhibitors.

TABLE 197-22

Combination Formulations of Antiretroviral Drugs

Name	Combination
Atripla*	Tenofovir disoproxil fumarate + emtricitabine + efavirenz
Biktarvy*	Tenofovir alafenamide + emtricitabine + bictegravir
Cimduo	Tenofovir disoproxil fumarate + lamivudine
Combivir	Zidovudine + lamivudine
Complera*	Tenofovir disoproxil fumarate+ emtricitabine + rilpivirine
Descovy	Tenofovir alafenamide + emtricitabine
Dutrebis	Raltegravir + lamivudine
Epzicom	Abacavir + lamivudine
Evotaz	Atazanavir + cobicistat
Genvoya*	Tenofovir alafenamide + emtricitabine + elvitegravir + cobicistat
Juluca	Dolutegravir + rilpivirine
Kaletra	Lopinavir + ritonavir
Odefsey*	Tenofovir alafenamide +emtricitabine + rilpivirine
Prezcobix	Darunavir + cobicistat
Stribild*	Tenofovir disoproxil fumarate + emtricitabine + elvitegravir + cobicistat
Symfi*	Tenofovir disoproxil fumarate + lamivudine + efavirenz (600 mg)
Symfi Lo*	Tenofovir disoproxil fumarate + lamivudine + efavirenz (400 mg)
Triumeq*	Abacavir + lamivudine + dolutegravir
Truvada	Tenofovir disoproxil fumarate + emtricitabine
Trizivir	Zidovudine + lamivudine + abacavir

*Complete, once-daily, single tablet regimens.

TABLE 197-23

Principles of Therapy of HIV Infection

1. Ongoing HIV replication leads to immune system damage, progression to AIDS, and systemic immune activation.
2. Plasma HIV RNA levels indicate the magnitude of HIV replication and the rate of CD4+ T cell destruction. CD4+ T cell counts indicate the current level of competence of the immune system.
3. Maximal suppression of viral replication is a goal of therapy; the greater the suppression the less likely the appearance of drug-resistant quasispecies.
4. The most effective therapeutic strategies involve the simultaneous initiation of combinations of effective anti-HIV drugs with which the patient has not been previously treated and that are not cross-resistant with antiretroviral agents that the patient has already received.
5. The antiretroviral drugs used in combination regimens should be used according to optimum schedules and dosages.
6. The number of available drugs is limited. Any decisions on antiretroviral therapy have a long-term impact on future options for the patient.
7. Women should receive optimal antiretroviral therapy regardless of pregnancy status.
8. The same principles apply to children and adults. The treatment of HIV-infected children involves unique pharmacologic, virologic, and immunologic considerations.
9. Compliance is an important part of ensuring maximal effect from a given regimen. The simpler the regimen, the easier it is for the patient to be compliant.

Source: Modified from *Principles of Therapy of HIV Infection*, USPHS, and the Henry J. Kaiser Family Foundation.

TABLE 197-24

Initial Combination Regimens Recommended for Most Treatment-Naïve Patients Regardless of HIV RNA Level or CD4 Count

Dolutegravir + tenofovir* + emtricitabine**
Raltegravir + tenofovir* + emtricitabine**
Bictegravir + tenofovir*+ emtricitabine**
Elvitegravir + cobicistat + tenofovir* + emtricitabine**
Dolutegravir + abacavir + lamivudine** (only for those HLA-B*5701 negative)

*Tenofovir alafenamide and tenofovir disoproxil fumarate are two forms of tenofovir approved by FDA. Tenofovir alafenamide has fewer bone and renal toxicities while tenofovir disoproxil fumarate is associated with lower lipid levels.

**Lamivudine may substitute for emtricitabine and vice versa.

Source: Guidelines for the Use of Antiretroviral Agents in HIV-Infected Adults and Adolescents, USPHS.

TABLE 197-25

Indications for Changing Antiretroviral Therapy in Patients with HIV Infection[a]

Less than a 1-log drop in plasma HIV RNA by 4 weeks following the initiation of therapy
A reproducible significant increase (defined as threefold or greater) from the nadir of plasma HIV RNA level not attributable to intercurrent infection, vaccination, or test methodology
Persistently declining CD4+ T cell numbers
Clinical deterioration
Side effects

[a]Generally speaking, a change should involve the initiation of at least two drugs felt to be effective in the given patient. The exception to this is when change is being made to manage toxicity, in which case a single substitution is reasonable.

Source: Guidelines for the Use of Antiretroviral Agents in HIV-Infected Adults and Adolescents, USPHS.

HIV AND THE HEALTH CARE WORKER

Health care workers, especially those who deal with large numbers of HIV-infected patients, have a small but definite risk of becoming infected with HIV as a result of professional activities (see "Occupational Transmission of HIV: Health Care Workers, Laboratory Workers, and the Health Care Setting," above).

In the United States 58 health care workers for whom case investigations have been completed have had documented seroconversions to HIV following occupational exposures. Only one of these has occurred since 1999. Approximately 85% of the exposures resulting in infection have been due to percutaneous (puncture/cut injury) exposures to HIV-infected blood. The individuals with documented seroconversions included 19 laboratory workers (16 of whom were clinical laboratory workers), 24 nurses, 16 clinical laboratory technicians, 6 physicians, 4 nonclinical laboratory technicians, 2 housekeepers, 2 surgical technicians, 1 dialysis technician, 1 respiratory therapist, 1 health aide, 1 embalmer/morgue technician, and 1 unknown. In addition, at least 150 possible cases of occupationally acquired HIV infection have been reported among health care personnel in the United States. The number of these workers who actually acquired their infection through occupational exposures is not known. Taken together, data from several large studies suggest that the risk of HIV infection following a percutaneous exposure to HIV-contaminated blood is ~0.2323%, and after a mucous membrane exposure, ~0.09%. Although episodes of HIV transmission after nonintact skin exposure have been documented, the average risk for transmission by this route has not been precisely quantified but is estimated to be less than the risk for mucous membrane exposures. The risk for transmission after exposure to body fluids or tissues other than HIV-infected blood also has not been quantified but is probably considerably lower than for blood exposures. A seroprevalence survey of 3420 orthopedic surgeons, 75% of whom practiced in an area with a relatively high prevalence of HIV infection and 39% of whom reported percutaneous exposure to patient blood, usually through an accident involving a suture needle, failed to reveal any cases of possible occupational infection, suggesting that the risk of infection with a suture needle may be considerably less than that with a blood-drawing (hollow-bore) needle.

Most cases of health care worker seroconversion occur as a result of needle-stick injuries. When one considers the circumstances that result in needle-stick injuries, it is immediately obvious that adhering to the standard guidelines for dealing with sharp objects would result in a significant decrease in this type of accident. In one study, 27% of needle-stick injuries resulted from improper disposal of the needle (over half of these were due to recapping the needle), 23% occurred during attempts to start an IV line, 22% occurred during blood drawing, 16% were associated with an IM or SC injection, and 12% were associated with giving an IV infusion.

Occupational exposures to HIV should be considered as a medical emergency to ensure timely postexposure management and administration of postexposure antiretroviral prophylaxis (PEP). Recommendations regarding PEP must take into account that a variety of circumstances determine the risk of transmission of HIV following occupational exposure. In this regard, several factors have been associated with an increased risk for occupational transmission of HIV infection, including deep injury, the presence of visible blood on the instrument causing the exposure, injury with a device that had been placed in the vein or artery of the source patient, and advanced HIV disease in the source patient. Other important considerations when considering PEP in the health care worker include known or suspected pregnancy or breast-feeding, the possibility of exposure to drug-resistant virus, and the toxicities of different PEP regimens. Regardless of the decision to use PEP, the wound should be cleansed immediately and antiseptic applied. If a decision is made to offer PEP, U.S. Public Health Service guidelines recommend that PEP regimens contain 3 (or more) antiretroviral drugs administered for a 4-week duration for all occupational exposures to HIV. Detailed guidelines are available from the *Updated U.S. Public Health Service Guidelines for the Management of Occupational Exposures to HIV and Recommendations for Postexposure Prophylaxis* (CDC, 2013). The report emphasizes the importance of adherence to PEP when it is indicated, and close follow-up of exposed workers should be provided including counseling, baseline and follow-up HIV testing, and monitoring for drug toxicity. Follow-up appointments should begin within 72 h of an HIV exposure, and if newer fourth-generation combination HIV p24 antigen-HIV antibody test is utilized for follow-up HIV testing of the exposed health care worker, HIV testing may be concluded 4 months after exposure; if a newer testing platform is not available, follow-up HIV testing is typically concluded 6 months after an HIV exposure. For consultation on the treatment of occupational exposures to HIV and other bloodborne pathogens, the clinician managing the exposed patient can call the National Clinicians' Post-Exposure Prophylaxis Hotline (PEPline) at 888-448-4911. This service is available 24 hours a day at no charge. (Additional information on the Internet is available at *www.nccc.ucsf.edu*.) PEPline support may be especially useful in challenging situations, such as when drug-resistant HIV strains are suspected or the health care worker is pregnant.

Health care workers can minimize their risk of occupational HIV infection by following the CDC guidelines of June 2015, which include adherence to universal precautions, assuming that blood and other body fluids from all patients are potentially infectious. Therefore, the following infection control precautions should be adhered to at all times: (1) routinely use barriers (such as gloves and/or goggles) when anticipating contact with blood or body fluids; (2) immediately wash hands and other skin surfaces after contact with blood or body fluids; and (3) carefully handle and dispose of sharp instruments during and after use. For further information contact the CDC at 800-CDC-INFO (232-4636) or see *www.cdc.gov/info*. In attempting to put this small but definite occupational risk of HIV infection to the health care worker in perspective, it is important to point out that ~200 health care workers die each year as a result of occupationally acquired hepatitis B infection. The tragedy in this instance is that these infections and deaths due to HBV could be greatly decreased by more extended use of the HBV vaccine. The risk of HBV infection following a needle-stick injury from a hepatitis antigen–positive patient is much higher than the risk of HIV infection (see "Transmission," above). There are multiple examples of needle-stick injuries where the patient was positive for both HBV and HIV and the health care worker became infected only with HBV. For these reasons, it is advisable, given

the high prevalence of HBV infection in HIV-infected individuals, that all health care workers dealing with HIV-infected patients be immunized with the HBV vaccine.

TB is another infection common to HIV-infected patients that can be transmitted to the health care worker. For this reason, all health care workers should know their PPD status, have it checked yearly, and, where appropriate, receive 6 months of isoniazid treatment if their skin test converts to positive. In addition, all patients in whom a diagnosis of pulmonary TB is being entertained should be placed immediately in respiratory isolation, pending results of the diagnostic evaluation. The emergence of drug-resistant organisms, including extensively drug-resistant TB strains, has made TB an increasingly important problem for health care workers. This is particularly true for the health care worker with preexisting HIV infection.

One of the most charged issues ever to come between health care workers and patients is that of transmission of infection from HIV-infected health care workers to their patients. This is discussed in “Occupational Transmission of HIV: Health Care Workers, Laboratory Workers, and the Health Care Setting,” above. Theoretically, the same universal precautions that are used to protect the health care worker from the HIV-infected patient will also protect the patient from the HIV-infected health care worker.

VACCINE FOR THE PREVENTION OF HIV INFECTION

There is currently no safe and effective vaccine that has been approved for the prevention of HIV infection. Successful vaccines are predicated on the assumptions that the body can mount an adequate immune response to the microbe or virus in question during natural infection and that the vaccine will mimic the natural response to infection. Even with serious diseases, such as smallpox, poliomyelitis, measles, and influenza among others, the body in the vast majority of cases clears the infectious agent and provides protection, which is usually life-long against future exposure against the same pathogen. Unfortunately, this is not the case with HIV infection since the natural immune response to HIV infection is unable to clear the virus from the body and cases of superinfection are not uncommon. Some of the factors that contribute to the problematic nature of development of a preventive HIV vaccine are (1) the high mutability of the virus; (2) the fact that the infection can be transmitted by cell-free or cell-associated virus; (3) the fact that the HIV provirus integrates itself into the genome of the target cell and may remain in a latent form unexposed to the immune system; (4) the likely need for the development of effective mucosal immunity; and (5) the fact that it has been difficult to establish the precise correlates of protective immunity to HIV infection. A fraction of a percent of HIV-infected individuals are “elite controllers” in that they maintain extremely low and even undetectable levels of viremia in the absence of cART, and a number of individuals have been exposed to HIV multiple times but remain uninfected; these facts suggest that there are elements of host defense or an HIV-specific immune response that have the potential to be protective against acquisition of infection. Early attempts to develop a vaccine with the envelope protein gp120 aimed at inducing neutralizing antibodies in humans were unsuccessful; the elicited antisera failed to neutralize primary isolates of HIV cultured and tested in fresh peripheral blood mononuclear cells. In this regard, two phase 3 trials were undertaken in the United States and Thailand using soluble gp120, and the vaccines failed to protect human volunteers from HIV infection. In addition, two separate vaccine trials aimed at eliciting CD8+ T cell responses to prevent infection and, if unsuccessful in preventing infection, to control postinfection viremia, also failed at both goals. In 2009, a vaccine using a poxvirus vector prime expressing various viral proteins followed by an envelope protein boost was tested in a 16,000-person clinical trial (RV144) conducted in Thailand among predominantly low-HIV-prevalence heterosexuals. The vaccine provided the first positive, albeit very modest, signal ever reported in an HIV vaccine trial, showing 31% protection against acquisition of infection. Such a result is certainly not sufficient justification for clinical use of the vaccine, but it served as an important first step in the direction of the development of a safe and effective vaccine against HIV infection. Follow-up studies of RV144 indicate that nonneutralizing or weakly neutralizing antibody responses against certain constant epitopes in the otherwise highly variable V1-V2 region of the HIV envelope may be associated with the modest degree of protection observed in that clinical trial. Additional similar studies are being conducted in high-HIV-prevalence countries in sub-Saharan Africa in attempts to improve on the results of RV144 by a variety of approaches, including increasing the number of vaccine boosts with envelope protein and the addition of adjuvant.

An area of HIV vaccine research that is currently being actively pursued is the attempt to induce broadly neutralizing antibodies by developing as immunogens for vaccination certain epitopes on the HIV envelope that are the targets of naturally occurring broadly neutralizing antibodies during HIV infection. It is curious that only about 20% of HIV-infected individuals develop broadly neutralizing antibodies in response to natural infection and they do so only after 2–3 years of ongoing infection. By the time these antibodies appear, they can neutralize a broad range of primary HIV isolates, but they appear to be ineffective against the autologous virus in the infected subject. Upon close examination, these broadly neutralizing antibodies manifest a high degree of somatic mutations that were accumulated over time and are responsible for their affinity maturation and broadly neutralizing capacity. The goal of current efforts is to develop the conformationally correct HIV envelope epitopes that, when used as immunogens, would direct the immune response of an uninfected individual to the production of broadly neutralizing antibodies over a reasonable time frame by sequential immunizations. It remains to be seen whether this approach will be feasible.

PREVENTION OF HIV ACQUISITION

Education, counseling, and behavior modification are the cornerstones of any HIV prevention strategy. A major problem in the United States and elsewhere is that many infections are passed on by those who do not know that they are infected. Of the ~1.1 million persons in the United States who are HIV-infected, it is estimated that ~15% do not know their HIV status and approximately 23% of all new infections are transmitted by those people

who are not aware that they are infected. In this regard, the CDC has recommended that HIV testing become part of routine medical care and that all individuals between the ages of 13 and 64 years be tested at least one time. These individuals should be informed of the testing and be tested without the need for written informed consent. Each individual can "opt out" of testing, but testing should otherwise be routinely administered. Individuals who are practicing high-risk behavior should be tested more often. In addition to identifying individuals who might benefit from cART, information gathered from such an approach should serve as the basis for behavior-modification programs, both for infected individuals who may be unaware of their HIV status and who could infect others and for uninfected individuals practicing high-risk behavior. The practice of "safer sex" is the most effective way for sexually active uninfected individuals to avoid contracting HIV infection and for infected individuals to avoid spreading infection. Abstinence from sexual relations is the only absolute way to prevent sexual transmission of HIV infection. However, for most individuals this is not feasible, and there are a number of relatively safe practices that can markedly decrease the chances of transmission of HIV infection. Partners engaged in monogamous sexual relationships who wish to be assured of safety should both be tested for HIV antibody. If both are negative, it must be understood that any divergence from monogamy puts both partners at risk; open discussion of the importance of honesty in such relationships should be encouraged. When the HIV status of either partner is not known, or when one partner is positive, there are a number of options. Use of condoms can markedly decrease the chance of HIV transmission. It should be remembered that condoms are not 100% effective in preventing transmission of HIV infection, and there is a ~10% failure rate of condoms used for contraceptive purposes. Most condom failures result from breakage or improper usage, such as not wearing the condom for the entire period of intercourse. Latex condoms are preferable, since virus has been shown to leak through natural skin condoms. Petroleum-based gels should never be used for lubrication of the condom, since they increase the likelihood of condom rupture. Some men who have sex with men practice exclusively fellatio as a "minimal-risk" activity compared with anal intercourse. It should be emphasized that receptive fellatio is not guaranteed safe sex, and although the incidence of transmission via fellatio is considerably less than that of rectal or vaginal intercourse, there has been documentation of transmission of HIV where receptive fellatio was the only sexual act performed (see "Transmission," above). Topical microbicides composed of gels containing antiretroviral drugs have been shown to be only modestly and variably efficacious in preventing acquisition of HIV infection in women engaging in vaginal intercourse. The considerable degree of variability in efficacy relates to the generally poor adherence of participants to the use of the intervention. Pre-exposure prophylaxis (PrEP) using oral antiretroviral drugs such as Truvada (tenofovir + emtricitabine) as a single daily pill in uninfected men who have sex with men and transgender women is highly efficacious in preventing acquisition of HIV infection. The degree of efficacy can be greater than 95% if subjects adhere strictly to the regimen. CDC estimates that approximately 1.2 million people in the United States are at "substantial risk" for HIV infection and should be counseled about PrEP. As of 2016, however, fewer than half of primary physicians and nurses have heard of PrEP.

Adult male circumcision, which has been shown to result in a 50–65% reduction in HIV acquisition in the circumcised subject, is currently being pursued, particularly in developing nations, as a component of HIV prevention. The most effective way to prevent transmission of HIV infection among IDUs is to stop the use of injectable drugs. Unfortunately, that is extremely difficult to accomplish unless the individual enters a treatment program. For those who will not or cannot participate in a drug treatment program and who will continue to inject drugs, the avoidance of sharing of needles and other paraphernalia ("works") is the next best way to avoid transmission of infection. However, the cultural and social factors that contribute to the sharing of paraphernalia are complex and difficult to overcome. Under these circumstances, paraphernalia should be cleaned after each usage with a virucidal solution, such as undiluted sodium hypochlorite (household bleach). Needle exchange programs have been highly successful in decreasing HIV transmission among injection drug users without increasing the use of injection drugs. Oral PrEP also is effective in preventing acquisition of HIV infections among IDUs. It is important for IDUs to be tested for HIV infection and counseled to avoid transmission to their sexual partners. Prevention of transmission through blood or blood products and prevention of mother-to-child transmission are discussed in "Transmission," above.

FURTHER READING

Fauci AS, Marston HD: Ending AIDS—is an HIV vaccine necessary? N Engl J Med 370:495, 2014.

Fauci AS, Marston HD: Ending the HIV-AIDS pandemic—follow the science. N Engl J Med 373:2197, 2015.

Gray GE et al: Which new health technologies do we need to achieve an end to HIV/AIDS? PLoS Biol 14:e1002372, 2016. [PubMed: 26933962]

Haase AT: Perils at mucosal front lines for HIV and SIV and their hosts. Nat Rev Immunol 5:783, 2005.

Haynes BF et al: HIV-Host interactions: Implications for vaccine design. Cell Host Microbe 19:292, 2016.

Kanters S et al: Comparative efficacy and safety of first-line antiretroviral therapy for the treatment of HIV infection: A systematic review and network meta-analysis. Lancet HIV 3:e510, 2016. [PubMed: 27658869]

Kuhar DT et al: Updated US Public Health Service guidelines for the management of occupational exposures to human immunodeficiency virus and recommendations for postexposure prophylaxis. Infect Control Hosp Epidemiol 34:875, 2013.

Marcus JL et al: Narrowing the gap in life expectancy between HIV-infected and HIV-uninfected individuals with access to care. J Acquir Immune Defic Syndr 73:39, 2016.

Moir S, Fauci AS: B-cell responses to HIV infection. Immunol Rev 275:33, 2017.

Rerks-Ngarm S et al: Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med 361:2209, 2009.

UN Joint Programme On HIV/AIDS (UNAIDS): Prevention Gap Report, 2016. Available at *www.unaids.org/en/resources/documents/2016/prevention-gap.*

World Health Organization: Blood safety and availability Fact Sheet. Available at *www.who.int/mediacentre/factsheets/fs279/en/.*

Chapter 198: Viral Gastroenteritis

Umesh D. Parashar; Roger I. Glass

INTRODUCTION

Acute infectious gastroenteritis is a common illness that affects persons of all ages worldwide. It is a leading cause of death among children in developing countries, accounting for an estimated 0.6 million deaths each year, and is responsible for up to 10–12% of all hospitalizations among children in industrialized countries, including the United States. Elderly persons, especially those with debilitating health conditions, also are at risk of severe complications and death from acute gastroenteritis. Among healthy young adults, acute gastroenteritis is rarely fatal but incurs substantial medical and social costs, including those of time lost from work.

Several enteric viruses have been recognized as important etiologic agents of acute infectious gastroenteritis (**Table 198-1**, **Fig. 198-1**). Although most viral gastroenteritis is caused by RNA viruses, the DNA viruses that are occasionally involved (e.g., adenovirus types 40 and 41) are included in this chapter. Illness caused by these viruses is characterized by the acute onset of vomiting and/or diarrhea, which may be accompanied by fever, nausea, abdominal cramps, anorexia, and malaise. As shown in **Table 198-2**, several features can help distinguish gastroenteritis caused by viruses from that caused by bacterial agents. However, the distinction based on clinical and epidemiologic parameters alone is often difficult, and laboratory tests are required to confirm the diagnosis.

TABLE 198-1

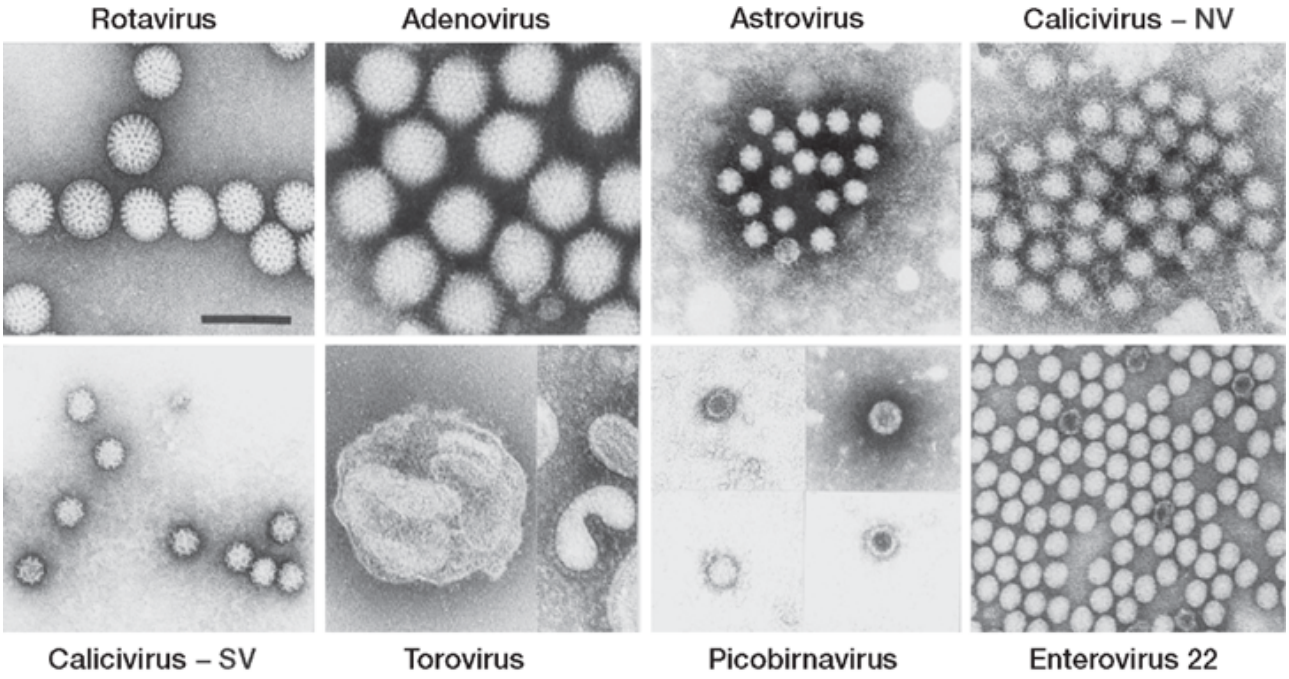
Viral Causes of Gastroenteritis Among Humans

Virus	Family	Genome	Primary Age Group at Risk	Clinical Severity	Detection Assays
Group A rotavirus	Reoviridae	Double-strand segmented RNA	Children <5 years	+ + +	EM, EIA (commercial), PAGE, RT-PCR
Norovirus	Caliciviridae	Positive-sense single-strand RNA	All ages	+ +	EM, RT-PCR
Sapovirus	Caliciviridae	Positive-sense single-strand RNA	Children <5 years	+	EM, RT-PCR
Astrovirus	Astroviridae	Positive-sense single-strand RNA	Children <5 years	+	EM, EIA, RT-PCR
Adenovirus (mainly types 40 and 41)	Adenoviridae	Double-strand DNA	Children <5 years	+/+ +	EM, EIA (commercial), PCR

Abbreviations: EIA, enzyme immunoassay; EM, electron microscopy; PAGE, polyacrylamide gel electrophoresis; PCR, polymerase chain reaction; RT-PCR, reverse-transcription PCR.

FIGURE 198-1

Viral agents of gastroenteritis. NV, norovirus; SV, sapovirus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 198-2

Characteristics of Gastroenteritis Caused by Viral and Bacterial Agents

Feature	Viral Gastroenteritis	Bacterial Gastroenteritis
Setting	Incidence similar in developing and developed countries	More common in settings with poor hygiene and sanitation
Infectious dose	Low (10–100 viral particles) for most agents	High (>10^5 bacteria) for *Escherichia coli, Salmonella, Vibrio;* medium (10^2–10^5 bacteria) for *Campylobacter jejuni;* low (10–100 bacteria) for *Shigella*
Seasonality	In temperate climates, winter seasonality for most agents; year-round occurrence in tropical areas	More common in summer or rainy months, particularly in developing countries with a high disease burden
Incubation period	1–3 days for most agents; can be shorter for norovirus	1–7 days for common agents (e.g., *Campylobacter, E. coli, Shigella, Salmonella);* a few hours for bacteria producing preformed toxins (e.g., *Staphylococcus aureus*, *Bacillus cereus*)
Reservoir	Primarily humans	Depending on bacterial species, human (e.g., *Shigella, Salmonella*), animal (e.g., *Campylobacter*, *Salmonella, E. coli*), and water (e.g., *Vibrio*) reservoirs exist
Fever	Common with rotavirus and norovirus; uncommon with other agents	Common with agents causing inflammatory diarrhea (e.g., *Salmonella*, *Shigella*)
Vomiting	Prominent and can be the only presenting feature, especially in children	Common with bacteria producing preformed toxins; less prominent in diarrhea due to other agents
Diarrhea	Common; non-bloody in almost all cases	Prominent and occasionally bloody with agents causing inflammatory diarrhea
Duration	1–3 days for norovirus and sapovirus; 2–8 days for other viruses	1–2 days for bacteria producing preformed toxins; 2–8 days for most other bacteria
Diagnosis	This is often a diagnosis of exclusion in clinical practice. Commercial enzyme immunoassays are available for detection of rotavirus and adenovirus, but identification of other agents is limited to research and public health laboratories.	Fecal examination for leukocytes and blood is helpful in differential diagnosis. Culture of stool specimens, sometimes on special media, can identify several pathogens. Molecular techniques are useful epidemiologic tools but are not routinely used in most laboratories.

Feature	Viral Gastroenteritis	Bacterial Gastroenteritis
Treatment	Supportive therapy to maintain adequate hydration and nutrition should be given. Antibiotics and antimotility agents are contraindicated.	Supportive hydration therapy is adequate for most patients. Antibiotics are recommended for patients with dysentery caused by *Shigella* or diarrhea caused by *Vibrio cholerae* and for some patients with *Clostridium difficile* colitis.

HUMAN CALICIVIRUSES

Etiologic Agent

The Norwalk virus is the prototype strain of a group of small (27–40 nm), nonenveloped, round, icosahedral viruses with relatively amorphous surface features on visualization by electron microscopy. Molecular cloning and characterization have demonstrated that the viruses have a single, positive-strand RNA genome ~7.5 kb in length and possess a single virion-associated protein—similar to that of typical caliciviruses—with a molecular mass of 60 kDa. On the basis of these molecular characteristics, these viruses are presently classified into two genera belonging to the family Caliciviridae: the *noroviruses* and the *sapoviruses* (previously called Norwalk-like viruses and Sapporo-like viruses, respectively). Human noroviruses can be classified into three genogroups: GI, GII, and GIV, which include 9, 22, and 1 genotype, respectively.

Epidemiology

Infections with the Norwalk and related human caliciviruses are common worldwide, and most adults have antibodies to these viruses. Antibody is acquired at an earlier age in developing countries—a pattern consistent with the presumed fecal–oral mode of transmission. Infections occur year-round, although, in temperate climates, a distinct increase has been noted in cold-weather months. Noroviruses may be the most common infectious agents of mild gastroenteritis in the community and affect all age groups, whereas sapoviruses primarily cause gastroenteritis in children. Noroviruses also cause traveler's diarrhea, and outbreaks have occurred among military personnel deployed to various parts of the world. The limited data available indicate that norovirus may be the second most common viral agent (after rotavirus) among young children and the most common agent among older children and adults. In the United States, with the decline in severe rotavirus disease following implementation of a rotavirus vaccination program, norovirus has become the leading cause of medically attended gastroenteritis in young children. Noroviruses are also recognized as the major cause of epidemics of gastroenteritis worldwide. In the United States, ~50% of all reported outbreaks of gastroenteritis are caused by noroviruses.

Virus is transmitted predominantly by the fecal–oral route but is also present in vomitus. Because an inoculum with very few viruses can be infectious, transmission can occur by aerosolization, by contact with contaminated fomites, and by person-to-person contact. Viral shedding and infectivity are greatest during the acute illness, but challenge studies with Norwalk virus in volunteers indicate that viral antigen may be shed by asymptomatically infected persons and also by symptomatic persons before the onset of symptoms and for several weeks after the resolution of illness. Viral shedding can be prolonged in immunocompromised individuals.

Pathogenesis

The exact sites and cellular receptors for attachment of viral particles have not been determined. Data suggest that carbohydrates that are similar to human histo-blood group antigens and are present on the gastroduodenal epithelium of individuals with the secretor phenotype may serve as ligands for the attachment of Norwalk virus. Additional studies must more fully elucidate norovirus–carbohydrate interactions, including potential strain-specific variations. After the infection of volunteers, reversible lesions are noted in the upper jejunum, with broadening and blunting of the villi, shortening of the microvilli, vacuolization of the lining epithelium, crypt hyperplasia, and infiltration of the lamina propria by polymorphonuclear neutrophils and lymphocytes. The lesions persist for at least 4 days after the resolution of symptoms and are associated with malabsorption of carbohydrates and fats and a decreased level of brush-border enzymes. Adenylate cyclase activity is not altered. No histopathologic changes are seen in the stomach or colon, but gastric motor function is delayed, and this alteration is believed to contribute to the nausea and vomiting that are typical of this illness.

Clinical Manifestations

Gastroenteritis caused by Norwalk and related human caliciviruses has a sudden onset following an average incubation period of 24 h (range, 12–72 h). The illness generally lasts 12–60 h and is characterized by one or more of the following symptoms: nausea, vomiting, abdominal cramps, and diarrhea. Vomiting is more prevalent among children, whereas a greater proportion of adults develop diarrhea. Constitutional symptoms are common, including headache, fever, chills, and myalgias. The stools are characteristically loose and watery, without blood, mucus, or leukocytes. White cell counts are generally normal; rarely, leukocytosis with relative lymphopenia may be observed. Death is a rare outcome and usually results from severe dehydration in vulnerable persons (e.g., elderly patients with debilitating health conditions).

Immunity

Approximately 50% of persons challenged with Norwalk virus become ill and acquire short-term immunity against the infecting strain. Immunity to Norwalk virus appears to correlate inversely with level of antibody; i.e., persons with higher levels of preexisting antibody to Norwalk virus are more susceptible to illness. This observation suggests that some individuals have a genetic predisposition to illness. Specific ABO, Lewis, and secretor blood group phenotypes influence susceptibility to norovirus infection.

Diagnosis

Cloning and sequencing of the genomes of Norwalk and several other human caliciviruses have allowed the development of assays based on polymerase chain reaction (PCR) for detection of virus in stool and vomitus. Virus-like particles (VLPs) produced by expression of capsid proteins in a recombinant baculovirus vector have been used to develop enzyme immunoassays (EIAs) for detection of virus in stool or a serologic response to a specific viral antigen. These newer diagnostic techniques are considerably more sensitive than previous detection methods, such as electron microscopy, immune electron microscopy, and EIAs based on reagents derived from humans. However, no currently available single assay can detect all human caliciviruses because of their great genetic and antigenic diversity. In addition, the assays are still cumbersome and are available primarily in research laboratories, although they are increasingly being adopted by public health laboratories for routine screening of fecal specimens from patients affected by outbreaks of gastroenteritis. Commercial EIA kits have limited sensitivity and usefulness in clinical practice and are of greatest utility in outbreaks, in which many specimens are tested and only a few need be positive to identify norovirus as the cause.

TREATMENT

TREATMENT

Infections with Norwalk and Related Human Caliciviruses

The disease is self-limited, and oral rehydration therapy is generally adequate. If severe dehydration develops, IV fluid therapy is indicated. No specific antiviral therapy is available.

Prevention

Epidemic prevention relies on situation-specific measures, such as control of contamination of food and water, exclusion of ill food handlers, and reduction of person-to-person spread through good personal hygiene and disinfection of contaminated fomites. The role of immunoprophylaxis is not clear, given the lack of long-term immunity from natural disease, but efforts to develop norovirus vaccines are ongoing. Vaccines based on VLPs are being tested in human volunteers. In a proof-of-concept trial, the efficacy of a monovalent GI.1 VLP vaccine was 47% among volunteers who received the vaccine intranasally and were then challenged with a homologous strain. In a second trial, norovirus disease severity was reduced in volunteers who received a bivalent G1.1/GII.4 VLP vaccine intramuscularly (with the GII.4 component including a consensus sequence from three different GII.4 strains) and were subsequently challenged with a GII.4 norovirus strain. While these initial data are encouraging, key issues to be further studied include the duration of protection and the level of heterotypic protection against antigenically distinct strains, particularly given the continuing and rapid natural evolution leading to the emergence of novel norovirus strains.

ROTAVIRUS

Etiologic Agent

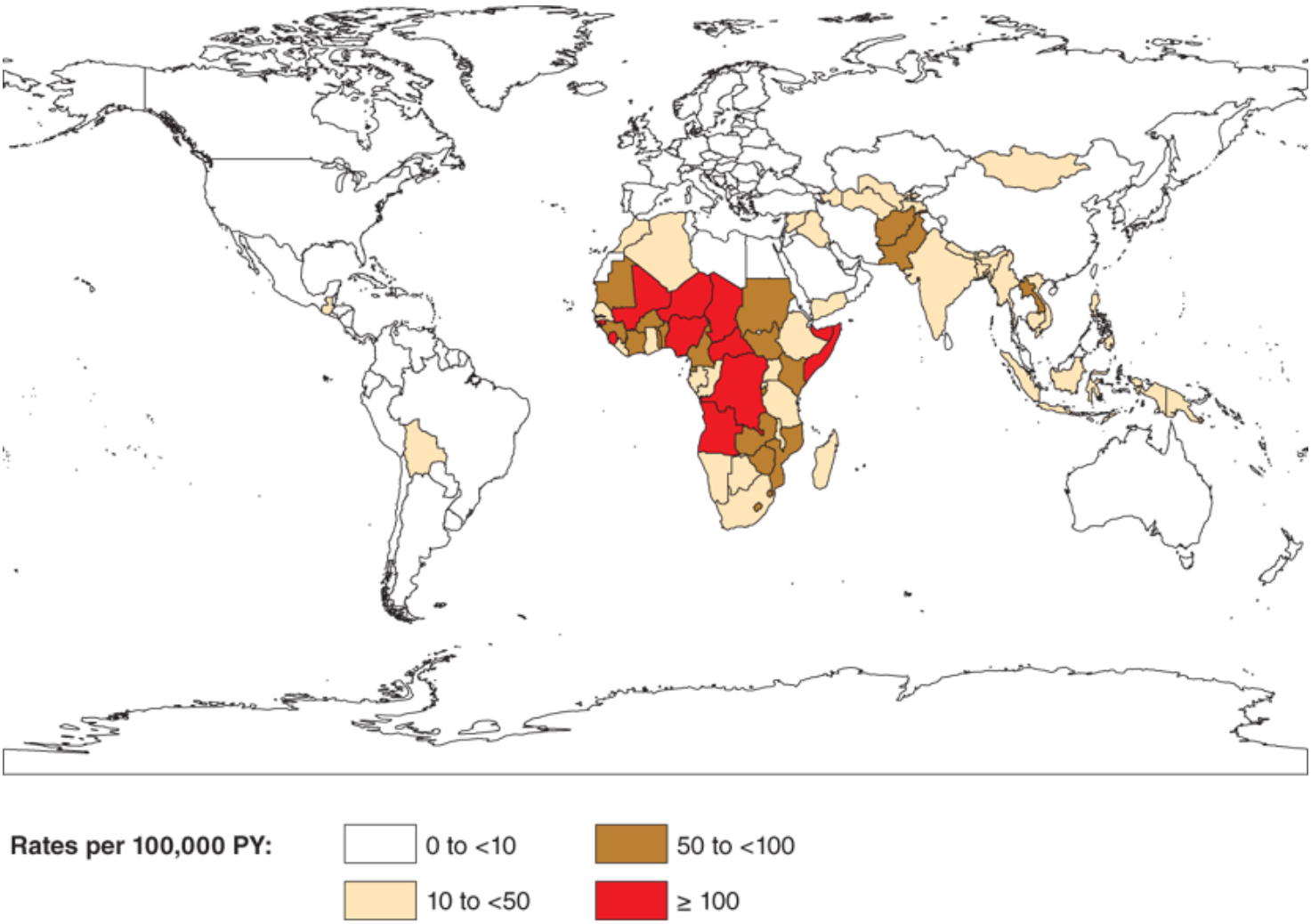
Rotaviruses are members of the family Reoviridae. The viral genome consists of 11 segments of double-strand RNA that is enclosed in a triple-layered, nonenveloped, icosahedral capsid 75 nm in diameter. Viral protein 6 (VP6), the major structural protein, is the target of commercial immunoassays and determines the group specificity of rotaviruses. There are seven major groups of rotavirus (A through G); human illness is caused primarily by group A and, to a much lesser extent, by groups B and C. Two outer-capsid proteins, VP7 (G-protein) and VP4 (P-protein), determine serotype specificity, induce neutralizing antibodies, and form the basis for binary classification of rotaviruses (G and P types). The segmented genome of rotavirus allows genetic reassortment (i.e., exchange of genome segments between viruses) during co-infection—a property that plays a role in viral evolution and that has been utilized in the development of reassortant animal/human rotavirus–based vaccines.

Epidemiology

Worldwide, nearly all children are infected with rotavirus by 3–5 years of age. Neonatal infections are common but are often asymptomatic or mild, presumably because of protection by maternal antibody or breast milk. Compared with rotavirus disease in industrialized countries, disease in developing countries occurs at a younger age, is less seasonal, and is more frequently caused by uncommon rotavirus strains. Moreover, because of suboptimal access to hydration therapy, rotavirus is a leading cause of diarrheal death among children in the developing world, with the highest mortality rates among children in sub-Saharan Africa and southern Asia **(Fig. 198-2)**.

FIGURE 198-2

Rotavirus mortality rates by country, per 100,000 children <5 years of age. *(Reproduced with permission from JE Tate et al: Clin Infect Dis 62 (Suppl 2):S96, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

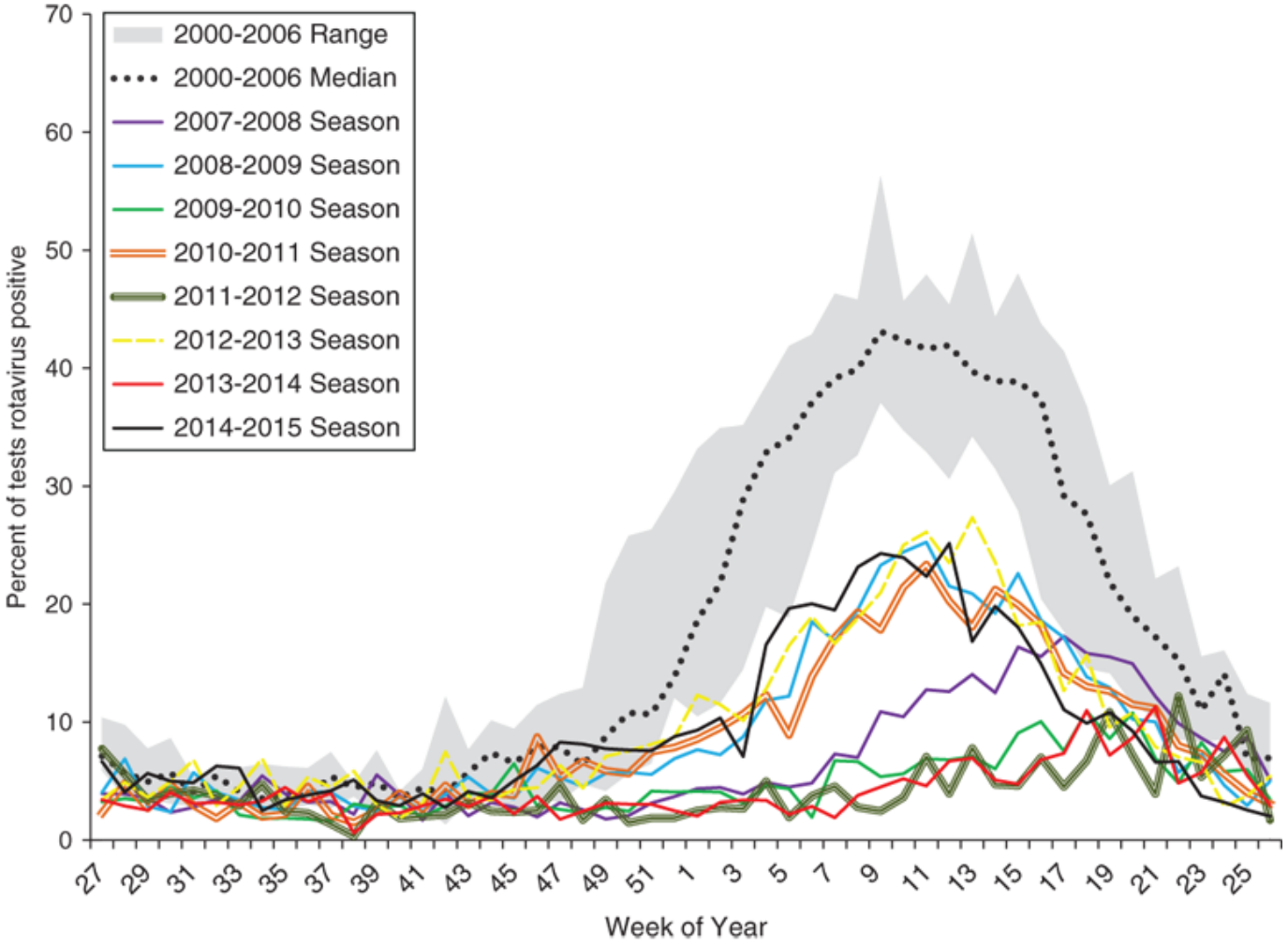
First infections after 3 months of age are likely to be symptomatic, and the incidence of disease peaks among children 4–23 months of age. Reinfections are common, but the severity of disease decreases with each repeat infection. Therefore, severe rotavirus infections are less common among older children and adults than among younger individuals. Nevertheless, rotavirus can cause illness in parents and caretakers of children with rotavirus diarrhea, immunocompromised persons, travelers, and elderly individuals and should be considered in the differential diagnosis of gastroenteritis among adults.

In tropical settings, rotavirus disease occurs year-round, with less pronounced seasonal peaks than in temperate settings, where rotavirus disease occurs predominantly during the cooler fall and winter months. Before the introduction of rotavirus vaccine in the United States, the rotavirus season each year began in the Southwest during the autumn and early winter (October through December) and migrated across the continent, peaking in the Northeast during late winter and spring (March through May). The reasons for this characteristic pattern are not clear but may be correlated with state-specific differences in birth rates, which could influence the rate of accumulation of susceptible infants after each rotavirus season. After the implementation of routine vaccination of U.S. infants against rotavirus in 2006, the characteristic prevaccine geotemporal pattern of U.S. rotavirus was dramatically altered, and these changes were accompanied by substantial declines in rotavirus detections by a national network of sentinel laboratories (Fig. 198-3). During the seven seasons since vaccine introduction (spanning 2008–2014), the number of rotavirus detections has declined by 58–90% from the prevaccine baseline, and the annual proportion of rotavirus tests positive in postvaccine seasons has ranged from 4% to

11%, in contrast to a prevaccine baseline median of 26%. In addition, a pattern of biennial increases in rotavirus activity has emerged during postvaccine seasons.

FIGURE 198-3

Percentage of rotavirus tests with positive results, by week of year, July–June, 2000–2015. The maximal or minimal percentage of rotavirus-positive tests for 2000–2006 may have occurred during any of the six baseline seasons. Data are from the National Respiratory and Enteric Virus Surveillance System. *(Adapted from Centers for Disease Control and Prevention, 2015.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

During episodes of rotavirus-associated diarrhea, virus is shed in large quantities in stool (10^7–10^{12}/g). Viral shedding detectable by EIA usually subsides within 1 week but may persist for >30 days in immunocompromised individuals; it may be detected for longer periods by sensitive molecular assays, such as PCR. The virus is transmitted predominantly through the fecal–oral route. Spread through respiratory secretions, person-to-person contact, or contaminated environmental surfaces has been postulated to explain the rapid acquisition of antibody in the first 3 years of life, regardless of sanitary conditions.

At least 10 different G serotypes of group A rotavirus have been identified in humans, but only 5 types (G1 through G4 and G9) are common. While human rotavirus strains that possess a high degree of genetic homology with animal strains have been identified, animal-to-human transmission appears to be uncommon.

Group B rotaviruses have been associated with several large epidemics of severe gastroenteritis among adults in China since 1982 and have also been identified in India. Group C rotaviruses have been associated with a small proportion of pediatric gastroenteritis cases in several countries worldwide.

Pathogenesis

Rotaviruses infect and ultimately destroy mature enterocytes in the villous epithelium of the proximal small intestine. The loss of absorptive villous epithelium, coupled with the proliferation of secretory crypt cells, results in secretory diarrhea. Brush-border enzymes characteristic of differentiated cells are reduced, and this change leads to the accumulation of unmetabolized disaccharides and consequent osmotic diarrhea. Studies in mice indicate that a nonstructural rotavirus protein, NSP4, functions as an enterotoxin and contributes to secretory diarrhea by altering epithelial cell function and permeability. In addition, rotavirus may evoke fluid secretion through activation of the enteric nervous system in the intestinal wall. Data indicate that rotavirus antigenemia and viremia are common among children with acute rotavirus infection, although the antigen and RNA levels in serum are substantially lower than those in stool.

Clinical Manifestations

The clinical spectrum of rotavirus infection ranges from subclinical infection to severe gastroenteritis leading to life-threatening dehydration. After an incubation period of 1–3 days, the illness has an abrupt onset, with vomiting frequently preceding the onset of diarrhea. Up to one-third of patients may have a temperature of >39°C. The stools are characteristically loose and watery and only infrequently contain red or white cells. Gastrointestinal symptoms generally resolve in 3–7 days.

Respiratory and neurologic features in children with rotavirus infection have been reported, but causal associations have not been proven. Moreover, rotavirus infection has been associated with a variety of other clinical conditions (e.g., sudden infant death syndrome, necrotizing enterocolitis, intussusception, Kawasaki disease, and type 1 diabetes), but no causal relationship has been confirmed with any of these syndromes.

Rotavirus does not appear to be a major opportunistic pathogen in children with HIV infection. In severely immunodeficient children, rotavirus can cause protracted diarrhea with prolonged viral excretion and, in rare instances, can disseminate systemically. Persons who are immunosuppressed for bone marrow transplantation also are at risk for severe or even fatal rotavirus disease.

Immunity

Protection against rotavirus disease is correlated with the presence of virus-specific secretory IgA antibodies in the intestine and, to some extent, the serum. Because virus-specific IgA production at the intestinal surface is short-lived, complete protection against disease is only temporary. However, each infection and subsequent reinfection confers progressively greater immunity; thus severe disease is most common among young children with first or second infections. Immunologic memory is believed to be important in the attenuation of disease severity upon reinfection.

Diagnosis

Illness caused by rotavirus is difficult to distinguish clinically from that caused by other enteric viruses. Because large quantities of virus are shed in feces, the diagnosis can usually be confirmed by a wide variety of commercially available EIAs or by techniques for detecting viral RNA, such as gel electrophoresis, probe hybridization, or PCR.

TREATMENT

TREATMENT

Rotavirus Infections

Rotavirus gastroenteritis can lead to severe dehydration. Thus appropriate treatment should be instituted early. Standard oral rehydration therapy is successful for most children who can take fluids by mouth, but IV fluid replacement may be required for patients who are severely dehydrated or are unable to tolerate oral therapy because of frequent vomiting. The therapeutic roles of probiotics, bismuth subsalicylate, enkephalinase inhibitors, and nitazoxanide have been evaluated in clinical studies but are not clearly defined. Antibiotics and antimotility agents should be avoided. In immunocompromised children with chronic symptomatic rotavirus disease, orally administered immunoglobulins or colostrum may result in the resolution of symptoms, but the best choices regarding agents and their doses have not been well studied, and treatment decisions are often empirical.

Prevention


Efforts to develop rotavirus vaccines were pursued because it was apparent—given the similar rates in less developed and industrialized nations—that improvements in hygiene and sanitation were unlikely to reduce disease incidence. The first rotavirus vaccine licensed in the United States in 1998 was withdrawn from the market within 1 year because it was linked with a low incidence of intussusception, a form of bowel obstruction.

In 2006, promising safety and efficacy results for two new rotavirus vaccines were reported from large clinical trials conducted in North America, Europe, and Latin America. Both vaccines are now recommended for routine immunization of all U.S. infants, and their use has rapidly led to a >70–80% decline in rotavirus hospitalizations and emergency department visits at hospitals across the United States. Somewhat unexpectedly, rotavirus vaccination of young infants has also resulted in the added benefit of declines in rotavirus disease among children who miss vaccination and even among older children and adults who are not eligible for vaccination in some settings. The reason is likely to be a reduction in community transmission of rotavirus because of vaccination—i.e., herd protection. In April 2009, the World Health Organization recommended the use of rotavirus vaccines in all countries worldwide. As of December 2016, a total of 82 countries, including 38 low-income countries in Africa and Asia, have incorporated rotavirus vaccine into their national childhood immunization programs. In Mexico and Brazil, a decline in deaths from childhood diarrhea following introduction of rotavirus vaccines has been documented. Postmarketing surveillance has identified a low risk of intussusception in some countries; however, the benefits of vaccination exceed the risks, and no changes in vaccine administration policy have been implemented.

The different epidemiology of rotavirus disease and the greater prevalence of co-infection with other enteric pathogens, of comorbidities, and of malnutrition in developing countries may adversely affect the performance of oral rotavirus vaccines, as is the case with oral vaccines against poliomyelitis, cholera, and typhoid in these regions. Therefore, evaluation of the efficacy of rotavirus vaccines in resource-poor settings of Africa and Asia was specifically recommended, and these trials have now been completed. As anticipated, the efficacy of rotavirus vaccines was moderate (50–65%) in these settings when compared with that in industrialized countries. Despite modest efficacy, routine use of rotavirus vaccines in low-income African countries with a heavy disease burden has yielded substantial public health benefits.

Several manufacturers in emerging markets, including India, China, Vietnam, Indonesia, and Brazil, are developing candidate rotavirus vaccines. In 2014, India licensed an indigenously manufactured rotavirus vaccine that showed 56% efficacy against severe rotavirus gastroenteritis during the first year of life. The vaccine has been recommended for inclusion in the Universal Immunization Program of India, and its use was initially implemented in four Indian states in 2015.

OTHER VIRAL AGENTS OF GASTROENTERITIS

Enteric *adenoviruses* of serotypes 40 and 41 belonging to subgroup F are 70- to 80-nm viruses with double-strand DNA that cause ~2–12% of all diarrhea episodes in young children. Unlike adenoviruses that cause respiratory illness, enteric adenoviruses are difficult to cultivate in cell lines, but they can be detected with commercially available EIAs. Adenovirus types 31 and 42–49 have been linked to diarrhea in HIV-infected and other immunocompromised persons.

Astroviruses are 28- to 30-nm viruses with a characteristic icosahedral structure and a positive-sense, single-strand RNA. At least seven serotypes have been identified, of which serotype 1 is most common. Astroviruses are primarily pediatric pathogens, causing ~2–10% of cases of mild to moderate gastroenteritis in children. The availability of simple immunoassays to detect virus in fecal specimens and of molecular methods to confirm and characterize strains will permit more comprehensive assessment of the etiologic role of these agents.

Toroviruses are 100- to 140-nm, enveloped, positive-strand RNA viruses that are recognized as causes of gastroenteritis in horses (Berne virus) and cattle (Breda virus). Their role as a cause of diarrhea in humans is still unclear, but studies from Canada have demonstrated associations between torovirus excretion and both nosocomial gastroenteritis and necrotizing enterocolitis in neonates. These associations require further evaluation.

Picobirnaviruses are small, bisegmented, double-strand RNA viruses that cause gastroenteritis in a variety of animals. Their role as primary causes of gastroenteritis in humans remains unclear, but several studies have found an association between picobirnaviruses and gastroenteritis in HIV-infected adults.

Several other viruses (e.g., enteroviruses, reoviruses, pestiviruses, and parvovirus B) have been identified in the feces of patients with diarrhea, but their etiologic role in gastroenteritis has not been proven. Diarrhea has also been noted as a manifestation of infection with recently recognized viruses that primarily cause severe respiratory illness: the severe acute respiratory syndrome–associated coronavirus (SARS-CoV), influenza A/H5N1 virus, and the current pandemic strain of influenza A/H1N1 virus.

FURTHER READING

Burnett E et al: Rotavirus vaccines: Current global impact and future perspectives. Future Virol 11:699, 2016.

Lopman BA et al: The vast and varied global burden of norovirus: Prospects for prevention and control. PLoS Med 13:e1001999, 2016. [PubMed: 27115709]

Tate JE et al: Global, regional, and national estimates of rotavirus mortality in children <5 years of age, 2000–2013. Clin Infect Dis 62(Suppl 2):S96, 2016. [PubMed: 27059362]

Yen C et al: Rotavirus vaccination and intussusception—science, surveillance, and safety: A review of evidence and recommendations for future research priorities in low and middle income countries. Hum Vaccin Immunother 12:2580, 2016.

Chapter 199: Enterovirus, Parechovirus, and Reovirus Infections

Jeffrey I. Cohen

ENTEROVIRUSES

CLASSIFICATION AND CHARACTERIZATION

Enteroviruses, members of the family Picornaviridae, are so designated because of their ability to multiply in the gastrointestinal tract. Despite their name, these viruses are not a prominent cause of gastroenteritis. Enteroviruses encompass more than 115 human serotypes: 3 serotypes of poliovirus, 21 serotypes of coxsackievirus A, 6 serotypes of coxsackievirus B, 28 serotypes of echovirus, enteroviruses 68–71, and multiple new enteroviruses (beginning with enterovirus 73) that have been identified by molecular techniques. Human enteroviruses have been reclassified into four species designated A–D. Echoviruses 22 and 23 have been reclassified as parechoviruses 1 and 2 on the basis of low nucleotide homology and differences in viral proteins. Enterovirus and parechovirus surveillance conducted in the United States by the Centers for Disease Control and Prevention (CDC) in 2009–2013 showed that the most common enteroviruses and parechoviruses were coxsackievirus A6 and human parechovirus 3, followed in frequency by echoviruses 11 and 18, coxsackieviruses A9 and B4, and echoviruses 6 and 30. Together, these eight viruses accounted for 58% of all isolates.

Human enteroviruses contain a single-stranded RNA genome surrounded by an icosahedral capsid comprising four viral proteins. These viruses have no lipid envelope and are stable in acidic environments, including the stomach. They are susceptible to chlorine-containing cleansers but resistant to inactivation by standard disinfectants (e.g., alcohol, detergents) and can persist for days at room temperature.

PATHOGENESIS AND IMMUNITY

Much of what is known about the pathogenesis of enteroviruses has been derived from studies of poliovirus infection. After ingestion, poliovirus is thought to infect epithelial cells in the mucosa of the gastrointestinal tract and then to spread to and replicate in the submucosal lymphoid tissue of the tonsils and Peyer's patches. The virus next spreads to the regional lymph nodes, a viremic phase ensues, and the virus replicates in organs of the reticuloendothelial system. In some cases, a second episode of viremia occurs and the virus replicates further in various tissues, sometimes causing symptomatic disease.

It is uncertain whether poliovirus reaches the central nervous system (CNS) during viremia or whether it also spreads via peripheral nerves. Since viremia precedes the onset of neurologic disease in humans, it has been assumed that the virus enters the CNS via the bloodstream. The poliovirus receptor is a member of the immunoglobulin superfamily. Poliovirus infection is limited to primates, largely because their cells express the viral receptor. Studies demonstrating the poliovirus receptor in the end-plate region of muscle at the neuromuscular junction suggest that, if the virus enters the muscle during viremia, it could travel across the neuromuscular junction up the axon to the anterior horn cells. Studies of monkeys and of transgenic mice expressing the poliovirus receptor show that, after IM injection, poliovirus does not reach the spinal cord if the sciatic nerve is cut. Taken together, these findings suggest that poliovirus can spread directly from muscle to the CNS by neural pathways.

Poliovirus can usually be cultured from the blood 3–5 days after infection, before the development of neutralizing antibodies. While viral replication at secondary sites begins to slow 1 week after infection, it continues in the gastrointestinal tract. Poliovirus is shed from the oropharynx for up to 3 weeks after infection and from the gastrointestinal tract for as long as 12 weeks; hypogammaglobulinemic patients can shed poliovirus for >20 years. During replication in the gastrointestinal tract, attenuated oral poliovirus can mutate, reverting to a more neurovirulent phenotype within a few days; however, additional mutations are probably required for full neurovirulence. One patient with hypogammaglobulinemia who had been infected 12 years earlier and was receiving IV immune globulin suddenly developed quadriplegia and respiratory muscle paralysis and died; analysis showed that the virus had reverted to a more wild-type sequence.

Humoral and secretory immunity in the gastrointestinal tract is important for the control of enterovirus infections. Enteroviruses induce specific IgM, which usually persists for <6 months, and specific IgG, which persists for life. Capsid protein VP1 is the predominant target of neutralizing antibody, which generally confers lifelong protection against subsequent disease caused by the same serotype but does not prevent infection or virus shedding. Enteroviruses also induce cellular immunity whose significance is uncertain. Patients with impaired cellular immunity are not known to develop unusually severe disease when infected with enteroviruses. In contrast, the severe infections in patients with agammaglobulinemia emphasize the importance of humoral immunity in controlling enterovirus infections. Disseminated enterovirus infections have occurred in hematopoietic cell transplant recipients. IgA antibodies are instrumental in reducing poliovirus replication in and shedding from the gastrointestinal tract. Breast milk contains IgA specific for enteroviruses and can protect humans from infection.

EPIDEMIOLOGY

 Enteroviruses have a worldwide distribution. More than 50% of nonpoliovirus enterovirus infections and more than 90% of poliovirus infections are subclinical. When symptoms do develop, they are usually nonspecific and occur in conjunction with fever; only a minority of infections are associated with specific clinical syndromes. The incubation period for most enterovirus infections ranges from 2 to 14 days but usually is <1 week.

Enterovirus infection is more common in socioeconomically disadvantaged areas, especially in those where conditions are crowded and in tropical areas where hygiene is poor. Infection is most common among infants and young children; serious illness develops most often during the first few days of life and in older children and adults. In developing countries, where children are infected at an early age, poliovirus infection has less often been associated with paralysis; in countries with better hygiene, older children and adults are more likely to be seronegative, become infected, and develop paralysis. Passively acquired maternal antibody reduces the risk of symptomatic infection in neonates. Young children are the most frequent shedders of enteroviruses and are usually the index cases in family outbreaks. In temperate climates, enterovirus infections occur most often in the summer and fall; no seasonal pattern is apparent in the tropics.

Most enteroviruses are transmitted primarily by the fecal-oral or oral-oral route. Patients are most infectious shortly before and after the onset of symptomatic disease, when virus is present in the stool and throat. The ingestion of virus-contaminated food or water also can cause disease. Certain enteroviruses (such as enterovirus 70, which causes acute hemorrhagic conjunctivitis) can be transmitted by direct inoculation from the fingers to the eye. Airborne transmission is important for some viruses that cause respiratory tract disease, such as coxsackievirus A21. Enteroviruses can be transmitted across the placenta from mother to fetus, causing severe disease in the newborn. The transmission of enteroviruses through blood transfusions or insect bites has not been documented. Nosocomial spread of coxsackievirus and echovirus has taken place in hospital nurseries.

CLINICAL FEATURES

Poliovirus Infection

Most infections with poliovirus are asymptomatic. After an incubation period of 3–6 days, ~5% of patients present with a minor illness (abortive poliomyelitis) manifested by fever, malaise, sore throat, anorexia, myalgias, and headache. This condition usually resolves in 3 days. About 1% of patients present with aseptic meningitis (nonparalytic poliomyelitis). Examination of cerebrospinal fluid (CSF) reveals lymphocytic pleocytosis, a normal glucose level, and a normal or slightly elevated protein level; CSF polymorphonuclear leukocytes may be present early. In some patients, especially children, malaise and fever precede the onset of aseptic meningitis.

PARALYTIC POLIOMYELITIS

The least common presentation is that of paralytic disease. After one or several days, signs of aseptic meningitis are followed by severe back, neck, and muscle pain and by the rapid or gradual development of motor weakness. In some cases the disease appears to be biphasic, with aseptic meningitis followed first by apparent recovery but then (1–2 days later) by the return of fever and the development of paralysis; this form is more common among children than among adults. Weakness is generally asymmetric, is proximal more than distal, and may involve the legs (most commonly); the arms; or the abdominal, thoracic, or bulbar muscles. Paralysis develops during the febrile phase of the illness and usually does not progress after defervescence. Urinary retention also may occur. Examination reveals weakness, fasciculations, decreased muscle tone, and reduced or absent reflexes in affected areas. Transient hyperreflexia sometimes precedes

the loss of reflexes. Patients frequently report sensory symptoms, but objective sensory testing usually yields normal results. Bulbar paralysis may lead to dysphagia, difficulty in handling secretions, or dysphonia. Respiratory insufficiency due to aspiration, involvement of the respiratory center in the medulla, or paralysis of the phrenic or intercostal nerves may develop, and severe medullary involvement may lead to circulatory collapse. Most patients with paralysis recover some function weeks to months after infection. About two-thirds of patients have residual neurologic sequelae.

Paralytic disease is more common among older individuals, pregnant women, and persons exercising strenuously or undergoing trauma at the time of CNS symptoms. Tonsillectomy predisposes to bulbar poliomyelitis, and IM injections increase the risk of paralysis in the involved limb(s).

VACCINE-ASSOCIATED POLIOMYELITIS

The risk of developing poliomyelitis after oral vaccination is estimated at 1 case per 2.5 million doses. The risk is ~2000 times higher among immunodeficient persons, especially persons with hypo- or agammaglobulinemia. Before 1997, an average of eight cases of vaccine-associated poliomyelitis occurred—in both vaccinees and their contacts—in the United States each year. With the change in recommendations first to a sequential regimen of inactivated poliovirus vaccine (IPV) and oral poliovirus vaccine (OPV) in 1997 and then to an all-IPV regimen in 2000, the number of cases of vaccine-associated polio declined. From 1997 to 1999, six such cases were reported in the United States; no cases have been reported since 1999.

POSTPOLIO SYNDROME

The *postpolio syndrome* presents as a new onset of weakness, fatigue, fasciculations, and pain with additional atrophy of the muscle group involved during the initial paralytic disease 20–40 years earlier. The syndrome is more common among women and with increasing time after acute disease. The onset is usually insidious, and weakness occasionally extends to muscles that were not involved during the initial illness. The prognosis is generally good; progression to further weakness is usually slow, with plateau periods of 1–10 years. The postpolio syndrome is thought to be due to progressive dysfunction and loss of motor neurons that compensated for the neurons lost during the original infection and not to persistent or reactivated poliovirus infection.

Other Enteroviruses

An estimated 5–10 million cases of symptomatic disease due to enteroviruses other than poliovirus occur in the United States each year. Among neonates, enteroviruses are the most common cause of aseptic meningitis and nonspecific febrile illnesses. Certain clinical syndromes are more likely to be caused by certain serotypes **(Table 199-1)**.

TABLE 199-1

Manifestations Commonly Associated with Enterovirus Serotypes

	Serotype(S) of Indicated Virus	
Manifestation	**Coxsackievirus**	**Echovirus (E) and Enterovirus (Ent)**
Acute hemorrhagic conjunctivitis	A24	E70
Aseptic meningitis	A2, 4, 7, 9, 10; B1–5	E4, 6, 7, 9, 11, 13, 16, 18, 19, 30, 33; Ent70, 71
Encephalitis	A9; B1–5	E3, 4, 6, 7, 9, 11, 18, 25, 30; Ent71
Exanthem	A4, 5, 6, 9, 10, 16; B1, 3–5	E4–7, 9, 11, 16–19, 25, 30; Ent71
Generalized disease of the newborn	B1–5	E4–7, 9, 11, 14, 16, 18, 19
Hand-foot-and-mouth disease	A5–7, 9, 10, 16; B1, 2, 5	Ent71
Herpangina	A1–10, 16, 22; B1–5	E6, 9, 11, 16, 17, 25, 30; Ent71
Myocarditis, pericarditis	A4, 9, 16; B1–5	E6, 9, 11, 22
Paralysis	A4, 7, 9; B1–5	E2–4, 6, 7, 9, 11, 18, 30; Ent70, 71
Pleurodynia	A1, 2, 4, 6, 9, 10, 16; B1–6	E1–3, 6, 7, 9, 11, 12, 14, 16, 19, 24, 25, 30
Pneumonia	A9, 16; B1–5	E6, 7, 9, 11, 12, 19, 20, 30; EntD68, 71

NONSPECIFIC FEBRILE ILLNESS (SUMMER GRIPPE)

The most common clinical manifestation of enterovirus infection is a nonspecific febrile illness. After an incubation period of 3–6 days, patients present with an acute onset of fever, malaise, and headache. Occasional cases are associated with upper respiratory symptoms, and some cases include nausea and vomiting. Symptoms often last for 3–4 days, and most cases resolve in a week. While infections with other respiratory viruses occur more often from late fall to early spring, febrile illness due to enteroviruses frequently occurs in the summer and early fall.

GENERALIZED DISEASE OF THE NEWBORN

Most serious enterovirus infections in infants develop during the first week of life, although severe disease can occur up to 3 months of age. Neonates often present with an illness resembling bacterial sepsis, with fever, irritability, and lethargy. Laboratory abnormalities include leukocytosis with a left shift, thrombocytopenia, elevated values in liver function tests, and CSF pleocytosis. The illness can be complicated by myocarditis and hypotension, fulminant hepatitis and disseminated intravascular coagulation, meningitis or meningoencephalitis, or pneumonia. It may be difficult to distinguish neonatal enterovirus infection from bacterial sepsis, although a history of a recent virus-like illness in the mother provides a clue.

ASEPTIC MENINGITIS AND ENCEPHALITIS

In children and young adults, enteroviruses are the cause of up to 90% of cases of aseptic meningitis in which an etiologic agent can be identified. Patients with aseptic meningitis typically present with an acute onset of fever, chills, headache, photophobia, and pain on eye movement. Nausea and vomiting also are common. Examination reveals meningismus without localizing neurologic signs; drowsiness or irritability also may be apparent. In some cases, a febrile illness may be reported that remits but returns several days later in conjunction with signs of meningitis. Other systemic manifestations may provide clues to an enteroviral cause, including diarrhea, myalgias, rash, pleurodynia, myocarditis, and herpangina. Examination of the CSF invariably reveals pleocytosis; the CSF cell count shows a shift from neutrophil to lymphocyte predominance within 1 day of presentation, and the total cell count does not exceed 1000/µL. The CSF glucose level is usually normal (in contrast to the low CSF glucose level in mumps), with a normal or slightly elevated protein concentration. Partially treated bacterial meningitis may be particularly difficult to exclude in some instances. Enteroviral meningitis is more common in summer and fall in temperate climates, while viral meningitis of other etiologies is more common in winter and spring. Symptoms ordinarily resolve within a week, although CSF abnormalities can persist for several weeks. Enteroviral meningitis is often more severe in adults than in children. Neurologic sequelae are rare, and most patients have an excellent prognosis.

Enteroviral encephalitis is much less common than enteroviral aseptic meningitis. Occasional highly inflammatory cases of enteroviral meningitis may be complicated by a mild form of encephalitis that is recognized on the basis of progressive lethargy, disorientation, and sometimes seizures. Less commonly, severe primary encephalitis may develop. An estimated 10–35% of cases of viral encephalitis are due to enteroviruses. Immunocompetent patients generally have a good prognosis.

Patients with hypogammaglobulinemia, agammaglobulinemia, or severe combined immunodeficiency may develop chronic meningitis or encephalitis; about half of these patients have a dermatomyositis-like syndrome, with peripheral edema, rash, and myositis. They may also have chronic hepatitis. Patients may develop neurologic disease while receiving immunoglobulin replacement therapy. Echoviruses (especially echovirus 11) are the most common pathogens in this situation.

Paralytic disease due to enteroviruses other than poliovirus occurs sporadically and is usually less severe than poliomyelitis. Most cases are due to enterovirus 70 or 71 or to coxsackievirus A7 or A9. Guillain-Barré syndrome is also associated with enterovirus infection. While earlier studies suggested a link between

enteroviruses and chronic fatigue syndrome, most recent studies have not demonstrated such an association.

PLEURODYNIA (BORNHOLM DISEASE)

Patients with pleurodynia present with an acute onset of fever and spasms of pleuritic chest or upper abdominal pain. Chest pain is more common in adults, and abdominal pain is more common in children. Paroxysms of severe, knifelike pain usually last 15–30 min and are associated with diaphoresis and tachypnea. Fever peaks within an hour after the onset of paroxysms and subsides when pain resolves. The involved muscles are tender to palpation, and a pleural rub may be detected. The white blood cell count and chest x-ray results are usually normal. Most cases are due to coxsackievirus B and occur during epidemics. Symptoms resolve in a few days, and recurrences are rare. Treatment includes the administration of nonsteroidal anti-inflammatory agents or the application of heat to the affected muscles.

MYOCARDITIS AND PERICARDITIS

Enteroviruses are estimated to cause up to one-third of cases of acute myocarditis. Coxsackievirus B and its RNA have been detected in pericardial fluid and myocardial tissue in some cases of acute myocarditis and pericarditis. Most cases of enteroviral myocarditis or pericarditis occur in newborns, adolescents, or young adults. More than two-thirds of patients are male. Patients often present with an upper respiratory tract infection that is followed by fever, chest pain, dyspnea, arrhythmias, and occasionally heart failure. A pericardial friction rub is documented in half of cases, and the electrocardiogram shows ST-segment elevations or ST- and T-wave abnormalities. Serum levels of myocardial enzymes are often elevated. Neonates commonly have severe disease, while most older children and adults recover completely. Up to 10% of cases progress to chronic dilated cardiomyopathy. Chronic constrictive pericarditis also may be a sequela.

EXANTHEMS

Enterovirus infection is the leading cause of exanthems in children in the summer and fall. While exanthems are associated with many enteroviruses, certain types have been linked to specific syndromes. Echoviruses 9 and 16 have frequently been associated with exanthem and fever. Rashes may be discrete or confluent, beginning on the face and spreading to the trunk and extremities. Echovirus 9 is the most common cause of a rubelliform (discrete) rash. Unlike the rash of rubella, the enteroviral rash occurs in the summer and is not associated with lymphadenopathy. Roseola-like rashes develop after defervescence, with macules and papules on the face and trunk. The Boston exanthem, caused by echovirus 16, is a roseola-like rash. A variety of other rashes have been associated with enteroviruses, including erythema multiforme **(see Fig. A1-24)** and vesicular, urticarial, petechial, bullous, or purpuric lesions. Enanthems also occur, including lesions that resemble the Koplik's spots seen with measles **(see Fig. A1-2)**.

HAND-FOOT-AND-MOUTH DISEASE (Fig. 199-1)

After an incubation period of 4–6 days, patients with hand-foot-and-mouth disease present with fever, anorexia, and malaise; these manifestations are followed by the development of sore throat and vesicles **(see Fig. A1-22)** on the buccal mucosa and often on the tongue and then by the appearance of tender vesicular lesions on the dorsum of the hands, sometimes with involvement of the palms. The vesicles may form bullae and quickly ulcerate. About one-third of patients also have lesions on the palate, uvula, or tonsillar pillars,

and one-third have a rash on the feet (including the soles) or on the buttocks. Generalized rashes also have been reported. The disease is highly infectious, with attack rates of close to 100% among young children. The lesions usually resolve in 1 week. Most cases are due to coxsackievirus A16 or enterovirus 71.

FIGURE 199-1

Vesicular eruptions of the hand (*A*), knee (*B*), and mouth (*C*) of a 6-year-old boy with coxsackievirus A6 infection. Several of his fingernails were shed 2 months later (*D*). *(Images reprinted courtesy of Centers for Disease Control and Prevention/Emerging Infectious Diseases.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

An epidemic of enterovirus 71 infection in Taiwan in 1998 resulted in thousands of cases of hand-foot-and-mouth disease or herpangina (see below). Severe complications included CNS disease, myocarditis, and pulmonary hemorrhage. About 90% of those who died were children ≤5 years old, and death was associated with pulmonary edema or pulmonary hemorrhage. CNS disease included aseptic meningitis, flaccid paralysis (similar to that seen in poliomyelitis), and rhombencephalitis with myoclonus and tremor or ataxia. The mean age of patients with CNS complications was 2.5 years, and MRI in cases with encephalitis usually showed brain-stem lesions. Follow-up of children at 6 months showed persistent dysphagia, cranial nerve

palsies, hypoventilation, limb weakness, and atrophy; at 3 years, persistent neurologic sequelae were documented, with delayed development and impaired cognitive function.

Yearly epidemics of enterovirus 71 infection have occurred in China since 2008, with hundreds of thousands of cases and hundreds of deaths each year. Infections have been associated with fever, rash, brain-stem encephalitis with myoclonic jerks, and limb trembling; some cases have progressed to seizures and coma. Lung findings include pulmonary edema and hemorrhage. While the level of creatine kinase MB is sometimes elevated, myocardial necrosis generally is not found.

Cyclic epidemics occur every 2–3 years in other Asian countries. However, the virus circulates at lower rates in the United States, Europe, and Africa. In the United States, hand-foot-and-mouth disease is most commonly associated with coxsackievirus A16. Between November 2011 and February 2012, outbreaks of hand-foot-and-mouth disease due to coxsackievirus A6 occurred in several U.S. states, and 19% of the affected persons were hospitalized.

HERPANGINA

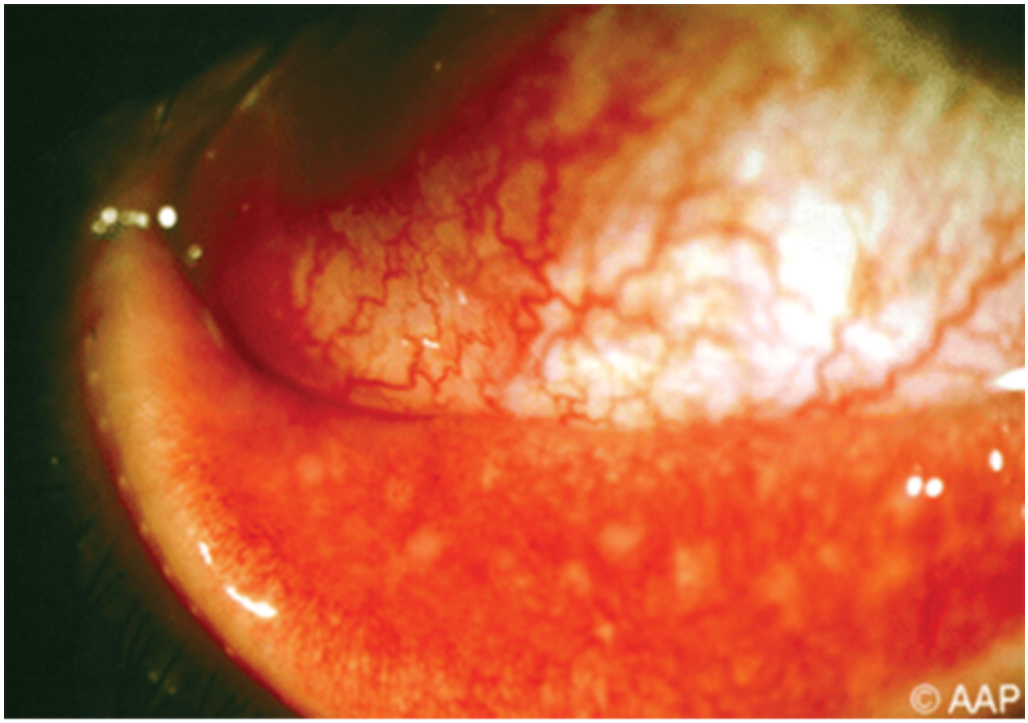
Herpangina is usually caused by coxsackievirus A and presents as acute-onset fever, sore throat, odynophagia, and grayish-white papulovesicular lesions on an erythematous base that ulcerate. The lesions can persist for weeks; are present on the soft palate, anterior pillars of the tonsils, and uvula; and are concentrated in the posterior portion of the mouth. In contrast to herpes stomatitis, enteroviral herpangina is not associated with gingivitis. Acute lymphonodular pharyngitis associated with coxsackievirus A10 presents as white or yellow nodules surrounded by erythema in the posterior oropharynx. The lesions do not ulcerate.

ACUTE HEMORRHAGIC CONJUNCTIVITIS

Patients with acute hemorrhagic conjunctivitis present with an acute onset of severe eye pain, blurred vision, photophobia, and watery discharge from the eye. Examination reveals edema, chemosis, and subconjunctival hemorrhage and often shows punctate keratitis and conjunctival follicles as well **(Fig. 199-2)**. Preauricular adenopathy is often found. Epidemics and nosocomial spread have been associated with enterovirus 70 and coxsackievirus A24. Recent outbreaks have been due to coxsackievirus A24 in China and India (2010), Japan (2011), and Thailand (2014). Systemic symptoms, including headache and fever, develop in 20% of cases, and recovery is usually complete in 10 days. The sudden onset and short duration of the illness help to distinguish acute hemorrhagic conjunctivitis from other ocular infections, such as those due to adenovirus and *Chlamydia trachomatis*. Paralysis has been associated with some cases of acute hemorrhagic conjunctivitis due to enterovirus 70 during epidemics.

FIGURE 199-2

Acute hemorrhagic conjunctivitis due to enterovirus 70. *(Image reprinted courtesy of Jerri Ann Jenista, MD.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

OTHER MANIFESTATIONS

Enteroviruses are an infrequent cause of childhood pneumonia and the common cold. From mid-August 2014 to January 2015, enterovirus D68 infection was confirmed in more than 1000 persons with mild to severe respiratory illnesses in 49 U.S. states. Nearly all reported cases were in children, many of whom had asthma. Enterovirus D68 has been detected in upper respiratory tract samples and very rarely in stool and serum from patients with acute flaccid myelitis. While epidemiologic evidence in 2014 suggested that enterovirus D68 may be associated with paralysis, the link was not definitively established, and more recent cases of acute flaccid myelitis in children generally have not been associated with enterovirus D68. Coxsackievirus B has been isolated at autopsy from the pancreas of a few children presenting with type 1 diabetes mellitus; however, most attempts to isolate the virus have been unsuccessful. Other diseases that have been associated with enterovirus infection include parotitis, bronchitis, bronchiolitis, croup, infectious lymphocytosis, polymyositis, acute arthritis, and acute nephritis.

DIAGNOSIS

Isolation of enterovirus in cell culture is the traditional diagnostic procedure. While cultures of stool, nasopharyngeal, or throat samples from patients with enterovirus diseases are often positive, isolation of the virus from these sites does not prove that it is directly associated with disease because these sites are frequently colonized for weeks in patients with subclinical infections. Isolation of virus from the throat is more likely to be associated with disease than is isolation from the stool since virus is shed for shorter periods from the throat. Cultures of CSF, serum, fluid from body cavities, or tissues are positive less frequently, but a positive result is indicative of disease caused by enterovirus. In some cases, the virus is isolated only from the blood or only from the CSF; therefore, it is important to culture multiple sites. Cultures are more likely to be positive earlier than later in the course of infection. Most human enteroviruses can be

detected within a week after inoculation of cell cultures. Cultures may be negative because of the presence of neutralizing antibody, lack of susceptibility of the cells used, or inappropriate handling of the specimen. Coxsackievirus A may require inoculation into special cell-culture lines or into suckling mice.

Identification of the enterovirus serotype is useful primarily for epidemiologic studies and, with a few exceptions, has little clinical utility. It is important to identify serious infections with enterovirus during epidemics and to distinguish the vaccine strain of poliovirus from the other enteroviruses in the throat or in the feces. Stool and throat samples for culture as well as acute- and convalescent-phase serum specimens should be obtained from all patients with suspected poliomyelitis. In the absence of a positive CSF culture, a positive culture of stool obtained within the first 2 weeks after the onset of symptoms is most often used to confirm the diagnosis of poliomyelitis. If poliovirus infection is suspected, two or more fecal and throat swab samples should be obtained at least 1 day apart and cultured for enterovirus as soon as possible. If poliovirus is isolated, it should be sent to the CDC for identification as either wild-type or vaccine virus.

Reverse-transcription polymerase chain reaction (PCR) has been used to amplify viral nucleic acid from CSF, serum, urine, stool, conjunctiva, throat swabs, and tissues. A pan-enterovirus PCR assay can detect all human enteroviruses. With the proper controls, PCR of the CSF is highly sensitive (70–100%) and specific (>80%) and is more rapid than culture. PCR of the CSF is less likely to be positive when patients present ≥3 days after the onset of meningitis or with enterovirus 71 infection; in these cases, PCR of throat or rectal swabs—although less specific than PCR of CSF—should be considered.

PCR of serum is also highly sensitive and specific in the diagnosis of disseminated disease. PCR may be particularly helpful for the diagnosis and follow-up of enterovirus disease in immunodeficient patients receiving immunoglobulin therapy, whose viral cultures may be negative. Antigen detection is less sensitive than PCR.

Serologic diagnosis of enterovirus infection is limited by the large number of serotypes and the lack of a common antigen. Demonstration of seroconversion may be useful in rare cases for confirmation of culture results, but serologic testing is usually limited to epidemiologic studies. Serum should be collected and frozen soon after the onset of disease and again ~4 weeks later. Measurement of neutralizing titers is the most accurate method for antibody determination; measurement of complement-fixation titers is usually less sensitive. Titers of virus-specific IgM are elevated in both acute and chronic infection.

TREATMENT

TREATMENT

Enterovirus Infections

Most enterovirus infections are mild and resolve spontaneously; however, intensive supportive care may be needed for cardiac, hepatic, or CNS disease. IV, intrathecal, or intraventricular immunoglobulin has been used with apparent success in some cases for the treatment of chronic enterovirus meningoencephalitis and dermatomyositis in patients with hypogammaglobulinemia or agammaglobulinemia. The disease may

stabilize or resolve during therapy; however, some patients decline inexorably despite therapy. IV immunoglobulin often prevents severe enterovirus disease in these patients. IV administration of immunoglobulin with high titers of antibody to the infecting virus has been used in some cases of life-threatening infection in neonates, who may not have maternally acquired antibody. In one trial involving neonates with enterovirus infections, immunoglobulin containing very high titers of antibody to the infecting virus reduced rates of viremia; however, the study was too small to show a substantial clinical benefit. The level of enteroviral antibodies varies with the immunoglobulin preparation. A phase 2 trial of pleconaril for neonatal enterovirus sepsis showed that the time to serum PCR negativity was reduced and the survival rate increased in newborns who had confirmed enterovirus infections and were treated with the drug, although in this small study the differences did not reach significance; as of this writing, the drug is not available on a compassionate-use basis. Pocapavir and vapendavir are also being tested for enterovirus infections. Glucocorticoids are contraindicated.

Good hand-washing practices and the use of gowns and gloves are important in limiting nosocomial transmission of enteroviruses during epidemics. Enteric precautions are indicated for 7 days after the onset of enterovirus infections. Inactivated enterovirus 71 vaccines have been shown to be efficacious in large clinical trials; they are not yet licensed.

PREVENTION AND ERADICATION OF POLIOVIRUS

(See also Chap. 118) After a peak of 57,879 cases of poliomyelitis in the United States in 1952, the introduction of IPV in 1955 and of OPV in 1961 ultimately eradicated disease due to wild-type poliovirus in the Western Hemisphere. Such disease has not been documented in the United States since 1979, when cases occurred among religious groups who had declined immunization. In the Western Hemisphere, paralysis due to wild-type poliovirus was last documented in 1991.

In 1988, when ~350,000 cases of polio occurred in 125 countries, the World Health Organization adopted a resolution to eradicate poliomyelitis by the year 2000. From 1988 to 2001, the number of cases worldwide decreased by >99%, with only 496 confirmed cases reported in 2001. Wild-type poliovirus type 2 has not been detected in the world since 1999 and wild-type poliovirus type 3 has not been circulating since 2012. The Americas were certified free of indigenous wild-type poliovirus transmission in 1994, the Western Pacific Region in 2000, the European Region in 2002, and Southeast Asia in 2014. After the nadir of 496 cases in 2001, 21 countries that had previously been free of polio reported cases imported from 6 polio-endemic countries in 2002–2005. By 2006, polio transmission had been reduced in most of these 21 countries. In 2016, there were 37 cases of wild-type polio; all of these cases were from Nigeria, Pakistan, and Afghanistan, the only countries where polio remains endemic **(Table 199-2)**. In 2017, wild-type poliovirus was detected in many sewage samples in Pakistan. As of early 2018, 22 cases of wild-type polio occurring in 2017 had been reported in Afghanistan and Pakistan. Polio is a source of concern for unimmunized or partially immunized travelers. While importation of poliovirus accounted for nearly 50% of cases in 2013 and also occurred in 2014, it has not been reported recently. Clearly, global eradication of polio is necessary to eliminate the risk of importation of wild-type virus. Outbreaks are thought to have been facilitated by suboptimal rates of

vaccination, isolated pockets of unvaccinated children, poor sanitation and crowding, improper vaccine-storage conditions, and a reduced level of response to one of the serotypes in the vaccine. While the global eradication campaign has markedly reduced the number of cases of endemic polio, doubts have been raised as to whether eradication is a realistic goal, given the large number of asymptomatic infections and the political instability in developing countries.

TABLE 199-2

Laboratory-Confirmed Cases of Poliomyelitis in 2016

Country	Endemic Cases	Vaccine-Derived Cases
Pakistan	20	1
Afghanistan	13	—
Nigeria	4	1
Laos	—	3
Total	37	5

Use of OPV, especially in areas with low vaccination rates, has been associated with vaccine-derived polio due to mutations that result in restoration of viral fitness and neurovirulence during prolonged replication in individuals or person-to-person transmission. Vaccine-derived polio was recognized in Egypt in 1983–1993, and hundreds of cases have been reported in many countries, including 385 cases in Nigeria in 2005–2012. Epidemics have been rapidly terminated after intensive vaccination with OPV. In 2005, a case of vaccine-derived polio occurred in an unvaccinated U.S. woman returning from a visit to Central and South America. In the same year, an unvaccinated immunocompromised infant in Minnesota was found to be shedding vaccine-derived poliovirus; further investigation identified 4 of 22 infants in the same community who were shedding the virus. All 5 infants were asymptomatic. These outbreaks emphasize the need for maintaining high levels of vaccine coverage and continued surveillance for circulating virus. From 2010 to 2014, 60–70 cases of vaccine-derived polio were reported annually. In 2016, only 5 cases were reported (in Nigeria, Pakistan, and Laos). As of early 2018, 92 cases of vaccine-derived polio occurring in 2017 had been reported in Syria and the Democratic Republic of the Congo. Of the 721 cases of vaccine-derived polio reported from January 2006 to May 2016, 94% were due to type 2 virus. Cessation of vaccination with type 2 OPV should reduce the number of these cases further.

IPV is used in most industrialized countries and OPV in most developing countries, including those in which polio still is or recently was endemic. While IM injections of other vaccines (live or attenuated) can be given concurrently with OPV, unnecessary IM injections should be avoided during the first month after OPV

vaccination because they increase the risk of vaccine-associated paralysis. Since 1988, an enhanced-potency inactivated poliovirus vaccine has been available in the United States.

After several doses of OPV alone, the seropositivity rate for individual poliovirus serotypes may still be suboptimal for children in developing countries; one or more supplemental doses of IPV can increase the rate of seropositivity for these serotypes. Against a given serotype, monovalent OPV containing only that serotype is more immunogenic than trivalent vaccine because of a lack of interference from other serotypes. Given the eradication of wild-type poliovirus type 2 and the establishment of OPV type 2 as the primary cause of vaccine-derived polio, bivalent OPV (types 1 and 3), which had been shown to be superior to trivalent OPV in inducing antibodies to types 1 and 3, replaced trivalent OPV vaccine in April 2016. Addition of at least one dose of trivalent IPV after immunization with bivalent OPV will reduce the risk of vaccine-derived polio associated with type 2 virus and enhance immunity to poliovirus types 1 and 3. Accordingly, in 2016, ~90% of countries included trivalent IPV in their immunization schedules. As the frequency of wild-type polio declines and reports of polio associated with circulating vaccine-derived viruses increase, the World Health Organization is investigating whether IPV can be produced from OPV strains that require less biocontainment, ultimately replacing OPV.

OPV and IPV induce antibodies that persist for at least 5 years. Both vaccines induce IgG and IgA antibodies. Compared with recipients of IPV, recipients of OPV shed less virus and less frequently develop reinfection with wild-type virus after exposure to poliovirus. Although IPV is safe and efficacious, OPV offers the advantages of ease of administration, lower cost, and induction of intestinal immunity resulting in a reduction in the risk of community transmission of wild-type virus. Because of progress toward global eradication of polio and the continued occurrence of cases of vaccine-associated polio, an all-IPV regimen was recommended in 2000 for childhood poliovirus vaccination in the United States, with vaccine administration at 2, 4, and 6–18 months and 4–6 years of age. The risk of vaccine-associated polio should be discussed before OPV is administered. Recommendations for vaccination of adults are listed in **Table 199-3**.

TABLE 199-3

Recommendations for Poliovirus Vaccination of Adults

1. Most adults in the United States have little risk for exposure to polioviruses, and most are immune as a result of vaccination during childhood. Vaccination is recommended for those at greater risk for exposure to polioviruses than the general population:
   a. travelers to areas or countries where polio is epidemic or endemic or where wild-type virus is known to have been circulating in the past year;
   b. members of communities or specific population groups with disease caused by wild-type polioviruses;
   c. laboratory workers who handle specimens that might contain polioviruses; and
   d. health care workers who have close contact with patients who might be excreting wild-type polioviruses.
2. Adults who are unvaccinated or whose vaccination status is unknown and who are at increased risk should receive three doses of IPV. Two doses of IPV should be administered at intervals of 4–8 weeks; a third dose should be administered 6–12 months after the second.
3. Adults who have had a primary series of polio vaccine and who are at increased risk should receive another dose of IPV. Currently, data do not indicate a need for more than a single lifetime booster dose with IPV for adults. However, adults who will be in a polio-infected or polio-exporting country for >4 weeks and whose booster dose of polio vaccine was administered >1 year earlier should receive an additional booster dose of vaccine before departing for that country.

Abbreviation: IPV, inactivated poliovirus vaccine.

Source: Modified from Centers for Disease Control and Prevention: MMWR Recomm Rep 46(RR-3):1, 1997; and Wallace et al: MMWR Morb Mortal Wkly Rep 63(27):591, 2014.

There are concerns about discontinuing vaccination in the event that endemic spread of poliovirus is eliminated. Among the reasons for these concerns are that poliovirus is shed from some immunocompromised persons for >25 years, that vaccine-derived poliovirus can circulate and cause disease, and that wild-type poliovirus is present in research laboratories and vaccine manufacturing facilities. Antivirals and monoclonal antibodies are in development to reduce or terminate shedding of poliovirus by long-term virus excretors. Pocapavir was recently shown to reduce shedding of OPV type 1 in a clinical trial, but rapid development of resistance with virus transmission, despite reduced shedding, indicates that combination therapy with antivirals and/or monoclonal antibodies will be needed.

PARECHOVIRUSES

Human parechoviruses (HPeVs), like enteroviruses, are members of the family Picornaviridae. The 16 serotypes of HPeV commonly cause infections in early childhood. Infections with HPeV type 1 (HPeV-1) occur throughout the year, while other parechovirus infections occur more commonly in summer and fall. Infections with HPeVs present similarly to those due to enteroviruses and may cause generalized disease of the newborn, aseptic meningitis, encephalitis, seizures, transient paralysis, exanthems, respiratory tract disease, rash, hepatitis, and gastroenteritis. While HPeV-1 is the most common serotype and generally causes mild disease, deaths of infants in the United States have been associated with HPeV-1, HPeV-3, and HPeV-6. HPeVs can be isolated from the same sites as enteroviruses, including the nasopharynx, stool, and respiratory tract secretions. PCR using pan-enterovirus primers does not detect HPeVs, and while PCR assays are performed by the CDC and research laboratories, many commercial laboratories do not perform the test. Pleconaril is not active against parechoviruses.

REOVIRUSES

Reoviruses are double-stranded RNA viruses encompassing three serotypes. Serologic studies indicate that most humans are infected with reoviruses during childhood. Most infections either are asymptomatic or cause mild upper respiratory tract symptoms. Reovirus is considered a rare cause of mild gastroenteritis or meningitis in infants and children. Speculation regarding an association of reovirus type 3 with idiopathic neonatal hepatitis and extrahepatic biliary atresia is based on an elevated prevalence of antibody to reovirus in some affected patients and the detection of viral RNA by PCR in hepatobiliary tissues in some studies. New orthoreoviruses have been associated with human disease—e.g., Melaka and Kampar viruses with fever and acute respiratory disease in Malaysia and Nelson Bay virus with acute respiratory disease in a traveler from Bali.

FURTHER READING

Abzug MJ et al: A randomized, double-blind, placebo-controlled trial of pleconaril for the treatment of neonates with enterovirus sepsis. J Pediatric Infect Dis Soc 5:53, 2016.

Holm-Hansen CC et al: Global emergence of enterovirus D68: A systematic review. Lancet Infect Dis 16:e64, 2016. [PubMed: 26929196]

Patel M, Orenstein W: A world free of polio—the final steps. N Engl J Med 374:501, 2016.

Pons-Salort M et al: The epidemiology of non-polio enteroviruses: Recent advances and outstanding questions. Curr Opin Infect Dis 28:479, 2015.

Sutter RW et al: Immunogenicity of a new routine vaccination schedule for global poliomyelitis prevention: An open-label, randomised controlled trial. Lancet 386:2413, 2015.

Sutter RW et al: Completing polio eradication: The case for antiviral drugs. J Infect Dis 215:333, 2017.

Chapter 200: Measles (Rubeola)

Kaitlin Rainwater-Lovett; William J. Moss

FIGURE 200-1

DEFINITION

Measles is a highly contagious viral disease that is characterized by a prodromal illness of fever, cough, coryza, and conjunctivitis followed by the appearance of a generalized maculopapular rash. Before the widespread use of measles vaccines, it was estimated that measles caused >2 million deaths worldwide each year.

GLOBAL CONSIDERATIONS

Remarkable progress has been made in reducing global measles incidence and mortality rates through measles vaccination. In the Americas, intensive vaccination and surveillance efforts—based in part on the successful Pan American Health Organization strategy of periodic nationwide measles vaccination campaigns (supplementary immunization activities, or SIAs)—and high levels of routine measles vaccine coverage interrupted endemic transmission of measles virus. The World Health Organization's (WHO's) Region of the Americas was declared to have eliminated measles in September 2016—the first region in the world to do so. In the United States, high-level coverage with two doses of measles vaccine eliminated endemic measles virus transmission in 2000. More recently, progress has been made in reducing measles incidence and mortality rates in sub-Saharan Africa and Asia as a consequence of increasing routine measles vaccine coverage and provision of a second dose of measles vaccine through mass measles vaccination campaigns and childhood immunization programs.

In 2003, the World Health Assembly endorsed a resolution urging member countries to reduce the number of deaths attributed to measles by 50% (compared with 1999 estimates) by the end of 2005. This target was met. Global measles mortality continued to decline and, by 2015, there were an estimated 134,200 deaths due to measles (uncertainty bounds: 74,400 and 353,600 deaths). These achievements attest to the enormous public-health significance of measles vaccination. However, recent large outbreaks of measles in Europe and Africa illustrate the challenges faced in sustaining measles control: in these outbreaks, measles was imported into countries that had eliminated indigenous transmission of measles virus.

The Measles and Rubella Initiative, a partnership led by the American Red Cross, the United Nations Foundation, UNICEF, the U.S. Centers for Disease Control and Prevention (CDC), and the WHO, is playing an

important role in reducing global measles incidence and mortality rates. Since its inception in 2001, the Initiative has provided governments and communities in more than 80 countries with technical and financial support for routine immunization activities, mass vaccination campaigns, and disease surveillance systems. Through its 2012–2020 Global Measles and Rubella Strategic Plan, the Initiative aimed to reduce measles deaths by 95% (compared with year 2000 estimates) by 2015 and to eliminate measles from at least five of the six WHO regions by 2020. The mortality reduction goal was not met, but an increasing number of countries have introduced a second dose of measles-containing vaccine into their routine immunization schedule. All six WHO regions have adopted goals for measles elimination by 2020 or earlier, and global measles eradication is likely to become a public health goal in the near future.

ETIOLOGY

Measles virus is a spherical, nonsegmented, single-stranded, negative-sense RNA virus and a member of the *Morbillivirus* genus in the family Paramyxoviridae. Measles was originally a zoonotic infection, arising from animal-to-human transmission of an ancestral morbillivirus ~10,000 years ago, when human populations had attained sufficient size to sustain virus transmission. Although RNA viruses typically have high mutation rates, measles virus is considered to be an antigenically monotypic virus; i.e., the surface proteins responsible for inducing protective immunity have retained their antigenic structure across time and distance. The public health significance of this stability is that measles vaccines developed decades ago from a single strain of measles virus remain protective worldwide. Measles virus is killed by ultraviolet light and heat, and attenuated measles vaccine viruses retain these characteristics, necessitating a cold chain for vaccine transport and storage.

EPIDEMIOLOGY

Measles virus is one of the most highly contagious directly transmitted pathogens. Outbreaks can occur in populations in which <10% of persons are susceptible. Chains of transmission are common among household contacts, school-age children, and health care workers. There are no latent or persistent measles virus infections that result in prolonged contagiousness, nor are there animal reservoirs for the virus. Thus, measles virus can be maintained in human populations only by an unbroken chain of acute infections, which requires a continuous supply of susceptible individuals. Newborns become susceptible to measles virus infection when passively acquired maternal antibody is lost; when not vaccinated, these infants account for the bulk of new susceptible individuals.

 Endemic measles has a typical temporal pattern characterized by yearly seasonal epidemics superimposed on longer epidemic cycles of 2–5 years or more. In temperate climates, annual measles outbreaks typically occur in the late winter and early spring. These annual outbreaks are probably attributable to social networks facilitating transmission (e.g., congregation of children at school) and environmental factors favoring the viability and transmission of measles virus. Measles cases continue to occur during interepidemic periods in large populations, but at low incidence. The longer epidemic cycles

occurring every several years result from the accumulation of susceptible persons over successive birth cohorts and the subsequent decline in the number of susceptibles following an outbreak.

Secondary attack rates among susceptible household and institutional contacts generally exceed 90%. The average age at which measles occurs depends on rates of contact with infected persons, protective maternal antibody decline, and vaccine coverage. In densely populated urban settings with low-level vaccination coverage, measles is a disease of infants and young children. The cumulative incidence can reach 50% by 1 year of age, with a significant proportion of children acquiring measles before 9 months—the age of routine vaccination in many countries, in line with the schedule recommended by the WHO's Expanded Programme on Immunization. As measles vaccine coverage increases or population density decreases, the age distribution shifts toward older children. In such situations, measles cases predominate in school-age children. Infants and young children, although susceptible if not protected by vaccination, are not exposed to measles virus at a rate sufficient to cause a heavy disease burden in this age group. As vaccination coverage increases further, the age distribution of cases may be shifted into adolescence and adulthood; this distribution is seen in measles outbreaks in the United States and necessitates targeted measles vaccination programs for these older age groups. Many countries have a bimodal distribution, with measles cases predominantly in young infants and adults.

Persons with measles are infectious for several days before and after the onset of rash, when levels of measles virus in blood and body fluids are highest and when cough, coryza, and sneezing, which facilitate virus spread, are most severe. The contagiousness of measles before the onset of recognizable disease hinders the effectiveness of quarantine measures. Viral shedding by children with impaired cell-mediated immunity can be prolonged.

Medical settings are well-recognized sites of measles virus transmission. Children may present to health care facilities during the prodrome, when the diagnosis is not obvious although the child is infectious and is likely to infect susceptible contacts. Health care workers can acquire measles from infected children and transmit measles virus to others. Nosocomial transmission can be reduced by maintenance of a high index of clinical suspicion, use of appropriate isolation precautions when measles is suspected, administration of measles vaccine to susceptible children and health care workers, and documentation of health care workers' immunity to measles (i.e., proof of receipt of two doses of measles vaccine or detection of antibodies to measles virus).

As efforts at measles control are increasingly successful, public perceptions of the risk of measles as a disease diminish and are replaced by concerns about possible adverse events associated with measles vaccine. As a consequence, numerous measles outbreaks have occurred because of opposition to vaccination on religious or philosophical grounds or unfounded fears of serious adverse events (see "Active Immunization," below).

PATHOGENESIS

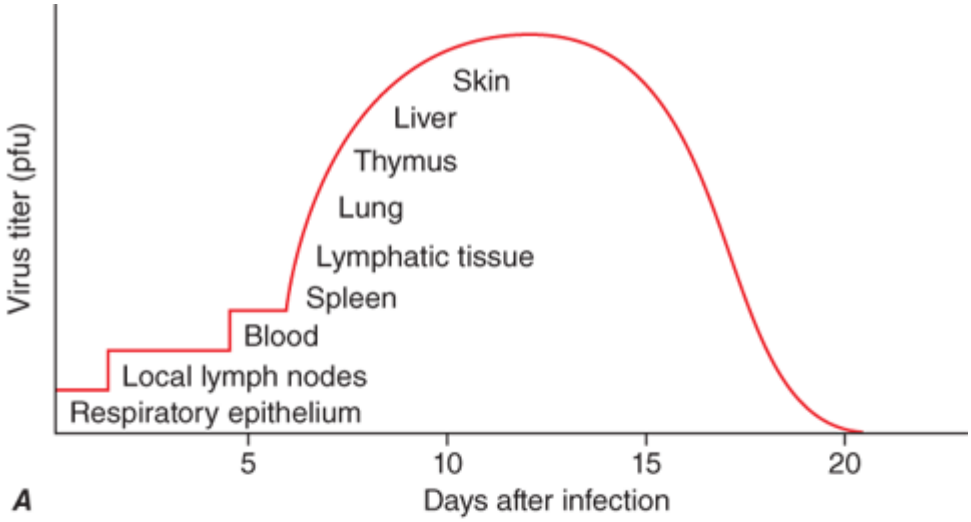
Measles virus is transmitted primarily by respiratory droplets over short distances and, less commonly, by small-particle aerosols that remain suspended in the air for long periods. Airborne transmission appears to

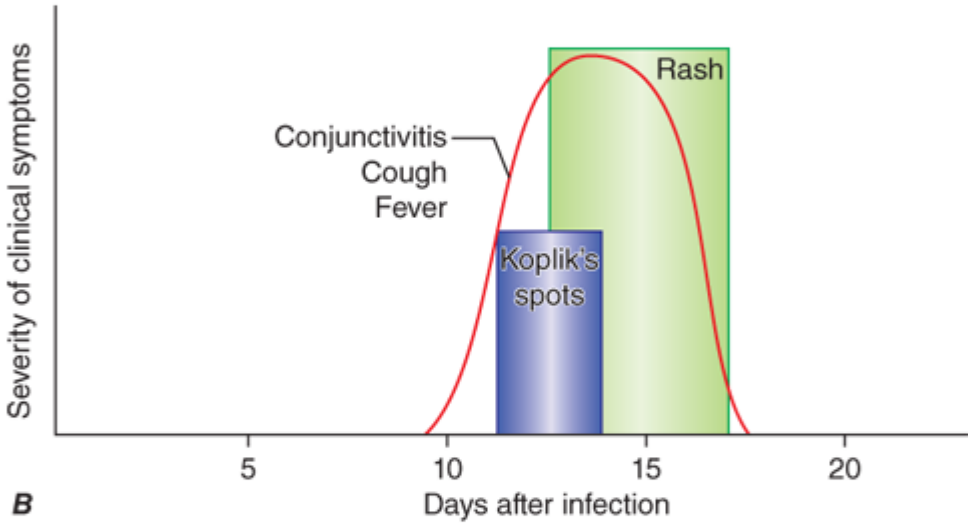
be important in certain settings, including schools, physicians' offices, hospitals, and enclosed public places. The virus can be transmitted by direct contact with infected secretions but does not survive for long on fomites.

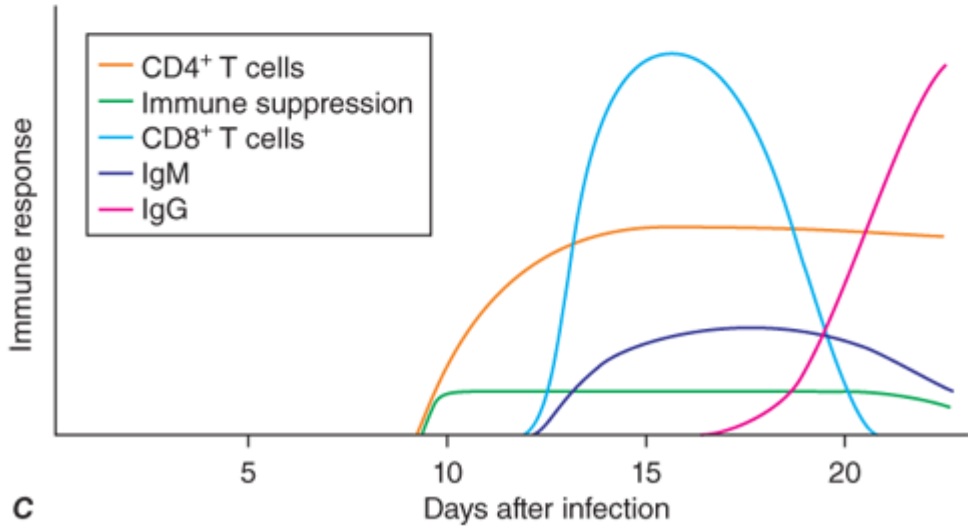
The incubation period for measles is ~10 days to fever onset and 14 days to rash onset. This period may be shorter in infants and longer (up to 3 weeks) in adults. Infection is initiated when measles virus is deposited in the respiratory tract, oropharynx, or conjunctivae **(Fig. 200-1A)**. During the first 2–4 days after infection, measles virus proliferates locally in the respiratory mucosa, primarily in dendritic cells and lymphocytes, and spreads to draining lymph nodes. Virus then enters the bloodstream in infected leukocytes, producing the primary viremia that disseminates infection throughout the reticuloendothelial system. Further replication results in secondary viremia that begins 5–7 days after infection and disseminates measles virus throughout the body. Replication of measles virus in the target organs, together with the host's immune response, is responsible for the signs and symptoms of measles that occur 8–12 days after infection and mark the end of the incubation period **(Fig. 200-1B)**.

FIGURE 200-1

Measles virus infection: pathogenesis, clinical features, and immune responses. ***A.*** Spread of measles virus, from initial infection of the respiratory tract through dissemination to the skin. ***B.*** Appearance of clinical signs and symptoms, including Koplik's spots and rash. ***C.*** Antibody and T cell responses to measles virus. The signs and symptoms of measles arise coincident with the host immune response. *(Source: Modified from WJ Moss, DE Griffin: Nat Rev Microbiol 4:900, 2006.)*







Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IMMUNE RESPONSES

Host immune responses to measles virus are essential for viral clearance, clinical recovery, and the establishment of long-term immunity **(Fig. 200-1C)**. Early nonspecific (innate) immune responses during the prodromal phase include activation of natural killer cells and increased production of antiviral proteins. The adaptive immune responses consist of measles virus–specific antibody and cellular responses. The protective efficacy of antibodies to measles virus is illustrated by the immunity conferred to infants from passively acquired maternal antibodies and the protection of exposed, susceptible individuals after administration of anti–measles virus immunoglobulin. The first measles virus–specific antibodies produced after infection are of the IgM subtype, with a subsequent switch to predominantly IgG1 and IgG4 isotypes. The IgM antibody response is typically absent following reexposure or revaccination and serves as a marker of primary infection.

The importance of cellular immunity to measles virus is demonstrated by the ability of children with agammaglobulinemia (congenital inability to produce antibodies) to recover fully from measles and the contrasting picture for children with severe defects in T lymphocyte function, who often develop severe or fatal disease **(Chap. 344)**. The initial predominant T_H1 response (characterized by interferon γ) is essential for viral clearance, and the later T_H2 response (characterized by interleukin 4) promotes the development of measles virus–specific antibodies that are critical for protection against reinfection.

The duration of protective immunity following wild-type measles virus infection is generally thought to be lifelong. Immunologic memory to measles virus includes both continued production of measles virus–specific antibodies and circulation of measles virus–specific $CD4^+$ and $CD8^+$ T lymphocytes.

However, the intense immune responses induced by measles virus infection are paradoxically associated with depressed responses to unrelated (non–measles virus) antigens, which persist for several weeks to months beyond resolution of the acute illness. This state of immune suppression enhances susceptibility to secondary infections with bacteria and viruses that cause pneumonia and diarrhea and is responsible for a substantial proportion of measles-related morbidity and deaths. Delayed-type hypersensitivity responses to recall antigens, such as tuberculin, are suppressed, and cellular and humoral responses to new antigens are impaired. Reactivation of tuberculosis and remission of autoimmune diseases after measles have been described and are attributed to this period of immune suppression.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Measles

Clinicians should consider measles in persons presenting with fever and generalized erythematous rash, particularly when measles virus is known to be circulating or the patient has a history of travel to endemic areas. Appropriate precautions must be taken to prevent nosocomial transmission. The diagnosis requires

laboratory confirmation except during large outbreaks in which an epidemiologic link to a confirmed case can be established. Care is largely supportive and consists of the administration of vitamin A and antibiotics (see "Treatment," below). Complications of measles, including secondary bacterial infections and encephalitis, may occur after acute illness and require careful monitoring, particularly in immunocompromised persons.

CLINICAL MANIFESTATIONS

In most persons, the signs and symptoms of measles are highly characteristic (Fig. 200-1B). Fever and malaise beginning ~10 days after exposure are followed by cough, coryza, and conjunctivitis. These signs and symptoms increase in severity over 4 days. Koplik's spots **(see Fig. A1-2)** develop on the buccal mucosa ~2 days before the rash appears. The characteristic rash of measles **(see Fig. A1-3)** begins 2 weeks after infection, when the clinical manifestations are most severe, and signal the host's immune response to the replicating virus. Headache, abdominal pain, vomiting, diarrhea, and myalgia may be present.

Koplik's spots are pathognomonic of measles and consist of bluish white dots ~1 mm in diameter surrounded by erythema. The lesions appear first on the buccal mucosa opposite the lower molars but rapidly increase in number and may involve the entire buccal mucosa. They fade with the onset of rash.

The rash of measles begins as erythematous macules behind the ears and on the neck and hairline. The rash progresses to involve the face, trunk, and arms, with involvement of the legs and feet by the end of the second day. Areas of confluent rash appear on the trunk and extremities, and petechiae may be present. The rash fades slowly in the same order of progression as it appeared, usually beginning on the third or fourth day after onset. Resolution of the rash may be followed by desquamation, particularly in undernourished children.

Because the characteristic rash of measles is a consequence of the cellular immune response, it may not develop in persons with impaired cellular immunity (e.g., those with AIDS; **Chap. 197**). These persons have a high case–fatality rate and frequently develop giant-cell pneumonitis caused by measles virus. T lymphocyte defects due to causes other than HIV-1 infection (e.g., cancer chemotherapy) also are associated with increased severity of measles.

A severe atypical measles syndrome was observed in recipients of a formalin-inactivated measles vaccine (used in the United States from 1963 to 1967 and in Canada until 1970) who were subsequently exposed to wild-type measles virus. The atypical rash began on the palms and soles and spread centripetally to the proximal extremities and trunk, sparing the face. The rash was initially erythematous and maculopapular but frequently progressed to vesicular, petechial, or purpuric lesions.

DIFFERENTIAL DIAGNOSIS

The differential diagnosis of measles includes other causes of fever, rash, and conjunctivitis, including rubella, Kawasaki disease, infectious mononucleosis, roseola, scarlet fever, Rocky Mountain spotted fever,

enterovirus or adenovirus infection, and drug sensitivity. Rubella is a milder illness without cough and with distinctive lymphadenopathy. The rash of roseola (exanthem subitum) **(see Fig. A1-5)** appears after fever has subsided. The atypical lymphocytosis in infectious mononucleosis contrasts with the leukopenia commonly observed in children with measles.

DIAGNOSIS

Measles is readily diagnosed on clinical grounds by clinicians familiar with the disease, particularly during outbreaks. Koplik's spots are especially helpful because they appear early and are pathognomonic. Clinical diagnosis is more difficult (1) during the prodromal illness; (2) when the rash is attenuated by passively acquired antibodies or prior immunization; (3) when the rash is absent or delayed in immunocompromised children or severely undernourished children with impaired cellular immunity; and (4) in regions where the incidence of measles is low and other pathogens are responsible for the majority of illnesses with fever and rash. The CDC case definition for measles requires (1) a generalized maculopapular rash of at least 3 days' duration; (2) fever of at least 38.3°C (101°F); and (3) cough, coryza, or conjunctivitis.

Serology is the most common method of laboratory diagnosis. The detection of measles virus–specific IgM in a single specimen of serum or oral fluid is considered diagnostic of acute infection, as is a fourfold or greater increase in measles virus–specific IgG antibody levels between acute- and convalescent-phase serum specimens. Primary infection in the immunocompetent host results in antibodies that are detectable within 1–3 days of rash onset and reach peak levels in 2–4 weeks. Measles virus–specific IgM antibodies may not be detectable until 4–5 days or more after rash onset and usually fall to undetectable levels within 4–8 weeks of rash onset.

Several methods for measurement of antibodies to measles virus are available. Neutralization tests are sensitive and specific, and the results are highly correlated with protective immunity; however, these tests require propagation of measles virus in cell culture and thus are expensive and laborious. Commercially available enzyme immunoassays are most frequently used. Measles can also be diagnosed by isolation of the virus in cell culture from respiratory secretions, nasopharyngeal or conjunctival swabs, blood, or urine. Direct detection of giant cells in respiratory secretions, urine, or tissue obtained by biopsy provides another method of diagnosis.

For detection of measles virus RNA by reverse-transcription polymerase chain reaction amplification of RNA extracted from clinical specimens, primers targeted to highly conserved regions of measles virus genes are used. Extremely sensitive and specific, this assay may also permit identification and characterization of measles virus genotypes for molecular epidemiologic studies and can distinguish wild-type from vaccine virus strains.

TREATMENT

TREATMENT

Measles

There is no specific antiviral therapy for measles. Treatment consists of general supportive measures, such as hydration and administration of antipyretic agents. Because secondary bacterial infections are a major cause of morbidity and death attributable to measles, effective case management involves prompt antibiotic treatment for patients who have clinical evidence of bacterial infection, including pneumonia and otitis media. *Streptococcus pneumoniae* and *Haemophilus influenzae* type b are common causes of bacterial pneumonia following measles; vaccines against these pathogens probably lower the incidence of secondary bacterial infections following measles.

Vitamin A is effective for the treatment of measles and can markedly reduce rates of morbidity and mortality. The WHO recommends administration of once-daily doses of 200,000 IU of vitamin A for 2 consecutive days to all children with measles who are ≥12 months of age. Lower doses are recommended for younger children: 100,000 IU per day for children 6–12 months of age and 50,000 IU per day for children <6 months old. A third dose is recommended 2–4 weeks later for children with evidence of vitamin A deficiency. While such deficiency is not a widely recognized problem in the United States, many American children with measles do, in fact, have low serum levels of vitamin A, and these children experience increased measles-associated morbidity. The Committee on Infectious Diseases of the American Academy of Pediatrics recommends that the administration of two consecutive daily doses of vitamin A be considered for children who are hospitalized with measles and its complications as well as for children with measles who are immunodeficient; who have ophthalmologic evidence of vitamin A deficiency, impaired intestinal absorption, or moderate to severe malnutrition; or who have recently immigrated from areas with high measles mortality rates. Parenteral and oral formulations of vitamin A are available.

Anecdotal reports have described the recovery of previously healthy pregnant and immunocompromised patients with measles pneumonia and of immunocompromised patients with measles encephalitis after treatment with aerosolized and IV ribavirin. However, the clinical benefits of ribavirin in measles have not been conclusively demonstrated in clinical trials.

COMPLICATIONS

Most complications of measles involve the respiratory tract and include the effects of measles virus replication itself and secondary bacterial infections. Acute laryngotracheobronchitis (croup) can occur during measles and may result in airway obstruction, particularly in young children. Giant-cell pneumonitis due to replication of measles virus in the lungs can develop in immunocompromised children, including those with HIV-1 infection. Many children with measles develop diarrhea, which contributes to undernutrition.

Most complications of measles result from secondary bacterial infections of the respiratory tract that are attributable to a state of immune suppression lasting for several weeks to months after acute measles. Otitis media and bronchopneumonia are most common and may be caused by *S. pneumoniae*, *H. influenzae* type

b, or staphylococci. Recurrence of fever or failure of fever to subside with the rash suggests secondary bacterial infection.

Rare but serious complications of measles involve the central nervous system (CNS). Post-measles encephalomyelitis complicates ~1 in 1000 cases, affecting mainly older children and adults. Encephalomyelitis occurs within 2 weeks of rash onset and is characterized by fever, seizures, and a variety of neurologic abnormalities. The finding of periventricular demyelination, the induction of immune responses to myelin basic protein, and the absence of measles virus in the brain suggest that post-measles encephalomyelitis is an autoimmune disorder triggered by measles virus infection. Other CNS complications that occur months to years after acute infection are measles inclusion body encephalitis (MIBE) and subacute sclerosing panencephalitis (SSPE). In contrast to post-measles encephalomyelitis, MIBE and SSPE are caused by persistent measles virus infection. MIBE is a rare but fatal complication that affects individuals with defective cellular immunity and typically occurs months after infection. SSPE is a slowly progressive disease characterized by seizures and progressive deterioration of cognitive and motor functions, with death occurring 5–15 years after measles virus infection. SSPE most often develops in persons infected with measles virus at <2 years of age.

PROGNOSIS

Most persons with measles recover and develop long-term protective immunity to reinfection. Measles case–fatality proportions vary with the average age of infection, the nutritional and immunologic status of the population, measles vaccine coverage, and access to health care. Among previously vaccinated persons who do become infected, disease is less severe and mortality rates are significantly lower. In developed countries, <1 in 1000 children with measles dies. In endemic areas of sub-Saharan Africa, the measles case–fatality proportion may be 5–10% or even higher. Measles is a major cause of childhood deaths in refugee camps and in internally displaced populations, where case–fatality proportions have been as high as 20–30%.

PREVENTION

Passive Immunization

Human immunoglobulin given shortly after exposure can attenuate the clinical course of measles. In immunocompetent persons, administration of immunoglobulin within 72 h of exposure usually prevents measles virus infection and almost always prevents clinical measles. Administered up to 6 days after exposure, immunoglobulin will still prevent or modify the disease. Prophylaxis with immunoglobulin is recommended for susceptible household and nosocomial contacts who are at risk of developing severe measles, particularly children <1 year of age, immunocompromised persons (including HIV-infected persons previously immunized with live attenuated measles vaccine), and pregnant women. Except for premature infants, children <6 months of age usually will be partially or completely protected by passively acquired maternal antibody. Infants born to women with vaccine-induced measles immunity become susceptible to

measles at a younger age than infants born to women with acquired immunity from natural infection. If measles is diagnosed in a household member, all unimmunized children in the household should receive immunoglobulin. The recommended dose is 0.25 mL/kg given intramuscularly. Immunocompromised persons should receive 0.5 mL/kg. The maximal total dose is 15 mL. IV immunoglobulin contains antibodies to measles virus; the usual dose of 100–400 mg/kg generally provides adequate prophylaxis for measles exposures occurring as long as 3 weeks or more after IV immunoglobulin administration.

Active Immunization

The first live attenuated measles vaccine was developed by passage of the Edmonston strain in chick embryo fibroblasts to produce the Edmonston B virus, which was licensed in 1963 in the United States. Further passage of Edmonston B virus produced the more attenuated Schwarz vaccine that currently serves as the standard in much of the world. The Moraten ("more attenuated Enders") strain, which was licensed in 1968 and is used in the United States, is genetically closely related to the Schwarz strain.

Lyophilized measles vaccines are relatively stable, but reconstituted vaccine rapidly loses potency. Live attenuated measles vaccines are inactivated by light and heat and lose about half their potency at 20°C and almost all their potency at 37°C within 1 h after reconstitution. Therefore, a cold chain must be maintained before and after reconstitution. Antibodies first appear 12–15 days after vaccination, and titers peak at 1–3 months. Measles vaccines are often combined with other live attenuated virus vaccines, such as those for mumps and rubella (MMR) and for mumps, rubella, and varicella (MMR-V).

The recommended age of first vaccination varies from 6 to 15 months and represents a balance between the optimal age for seroconversion and the probability of acquiring measles before that age. The proportions of children who develop protective levels of antibody after measles vaccination approximate 85% at 9 months of age and 95% at 12 months. Common childhood illnesses concomitant with vaccination may reduce the level of immune response, but such illness is not a valid reason to withhold vaccination. Measles vaccines have been well tolerated and immunogenic in HIV-1-infected children and adults, although antibody levels may wane. Because of the potential severity of wild-type measles virus infection in HIV-1-infected children, routine measles vaccination is recommended except for those who are severely immunocompromised. Measles vaccination is contraindicated in individuals with other severe deficiencies of cellular immunity because of the possibility of disease due to progressive pulmonary or CNS infection with the vaccine virus.

The duration of vaccine-induced immunity is at least several decades, if not longer. Rates of secondary vaccine failure 10–15 years after immunization have been estimated at ~5%, but are probably lower when vaccination takes place after 12 months of age. Decreasing antibody concentrations do not necessarily imply a complete loss of protective immunity: a secondary immune response usually develops after reexposure to measles virus, with a rapid rise in antibody titers in the absence of overt clinical disease.

Standard doses of currently licensed measles vaccines are safe for immunocompetent children and adults. Fever to 39.4°C (103°F) occurs in ~5% of seronegative vaccine recipients, and 2% of vaccine recipients

develop a transient rash. Mild transient thrombocytopenia has been reported, with an incidence of ~1 case per 40,000 doses of MMR vaccine.

Since the publication of a report in 1998 hypothesizing that MMR vaccine may cause a syndrome of autism and intestinal inflammation, much public attention has focused on this purported association. The events that followed publication of this report led to diminished vaccine coverage in the United Kingdom and provide important lessons in the misinterpretation of epidemiologic evidence and the communication of scientific results to the public. The publication that incited the concern was a case series describing 12 children with a regressive developmental disorder and chronic enterocolitis; 9 of these children had autism. In 8 of the 12 cases, the parents associated onset of the developmental delay with MMR vaccination. This simple temporal association was misinterpreted and misrepresented as a possible causal relationship, first by the lead author of the study and then by elements of the media and the public. Subsequently, many comprehensive reviews and additional epidemiologic studies refuted evidence of a causal relationship between MMR vaccination and autism.

PROSPECTS FOR MEASLES ERADICATION

Progress in global measles control has renewed discussion of measles eradication. In contrast to poliovirus eradication, the eradication of measles virus will not entail challenges posed by prolonged shedding of potentially virulent vaccine viruses and environmental viral reservoirs. However, in comparison with smallpox eradication, higher levels of population immunity will be necessary to interrupt measles virus transmission, more highly skilled health care workers will be required to administer measles vaccines, and containment through case detection and ring vaccination will be more difficult for measles virus because of infectivity before rash onset. New tools, such as microneedle patches to deliver measles vaccine, will facilitate mass vaccination campaigns. Despite enormous progress, measles remains a leading vaccine-preventable cause of childhood mortality worldwide and continues to cause outbreaks in communities with low vaccination coverage rates in industrialized nations.

FURTHER READING

Bester JC: Measles and measles vaccination: A review. JAMA Pediatr 170:1209, 2016.

Durrheim DN et al: Measles—the epidemiology of elimination. Vaccine 32:6880, 2014.

Goodson JL, Seward JF: Measles 50 years after use of measles vaccine. Infect Dis Clin North Am 29:725, 2015.

Griffin DE: The immune response in measles: Virus control, clearance and protective immunity. Viruses 8:282, 2016.

Laksono BM et al: Measles virus host invasion and pathogenesis. Viruses 8:210, 2016.

Moss WJ: Measles. Lancet 380:2490, 2017.

Phadke VK et al: Association between vaccine refusal and vaccine-preventable diseases in the United States: A review of measles and pertussis. JAMA 315:1149, 2016.

Rota PA et al: Measles. Nat Rev Dis Primers 2:16049, 2016.

World Health Organization: Measles vaccines: WHO position paper—April 2017. Wkly Epidemiol Rec 92:205, 2017.

World Health Organization: Progress towards regional measles elimination—worldwide, 2000–2016. Wkly Epidemiol Rec 92:649, 2017.

Chapter 201: Rubella (German Measles)

Laura A. Zimmerman; Susan E. Reef

INTRODUCTION

Rubella was historically viewed as a variant of measles or scarlet fever. Not until 1962 was a separate viral agent for rubella isolated. After an epidemic of rubella in Australia in the early 1940s, the ophthalmologist Norman Gregg noticed the occurrence of congenital cataracts among infants whose mothers had reported rubella during early pregnancy, and congenital rubella syndrome (CRS; see "Clinical Manifestations," below) was first described.

ETIOLOGY

Rubella virus is a member of the Togaviridae family and the only member of the genus *Rubivirus*. This single-strand RNA enveloped virus measures 40–80 nm in diameter. Its core protein is surrounded by a single-layer lipoprotein envelope with spike-like projections containing two glycoproteins, E1 and E2. There is only one antigenic type of rubella virus, and humans are its only known reservoir.

PATHOGENESIS AND PATHOLOGY

Although the pathogenesis of postnatal (acquired) rubella has been well documented, data on pathology are limited because of the mildness of the disease. Rubella virus is spread from person to person via respiratory droplets. Primary implantation and replication in the nasopharynx are followed by spread to the lymph nodes. Subsequent viremia occurs, which in pregnant women often results in infection of the placenta. Placental virus replication may lead to infection of fetal organs. The pathology of CRS in the infected fetus is well defined, with almost all organs found to be infected; however, the pathogenesis of CRS is only poorly delineated. In tissue, infections with rubella virus have diverse effects, ranging from no obvious impact to cell destruction. The hallmark of fetal infection is chronicity, with persistence throughout fetal development in utero and for up to 1 year after birth.

Individuals with acquired rubella may shed virus from 7 days before rash onset to ~5–7 days thereafter. Both clinical and subclinical infections are considered contagious. Infants with CRS may shed large quantities of virus from bodily secretions, particularly from the throat and in the urine, up to 1 year of age. Outbreaks of rubella, including some in nosocomial settings, have originated with index cases of CRS. Thus only individuals immune to rubella virus should have contact with infants who have CRS or who are congenitally infected with rubella virus but are not showing signs of CRS.

EPIDEMIOLOGY

The largest recent rubella epidemic in the United States took place in 1964–1965, when an estimated 12.5 million cases occurred, resulting in ~20,000 cases of CRS. Since the introduction of the routine rubella vaccination program in the United States in 1969, the number of rubella cases reported each year has dropped by >99%; the rate of vaccination coverage with rubella-containing vaccine (RCV) has been >90% among children 19–35 months old since 1995 and >95% for kindergarten and first-grade entrants since 1980. In the United States, a goal for the elimination of rubella and CRS was set in 1989. Interruption of endemic transmission of rubella virus was achieved by 2001. In 2004, a panel of experts agreed unanimously that rubella was no longer an endemic disease in the United States. The criteria used to document lack of endemic transmission included low disease incidence, high nationwide rubella antibody seroprevalence, outbreaks that were few and contained (i.e., small numbers of cases), and lack of endemic virus transmission (as assessed by genetic sequencing). Although interruption of endemic transmission has been sustained since 2001, rubella virus importations continue to occur and cases continue to develop among susceptible persons. During 2004–2015, 60% of rubella cases occurred in persons 20–49 years old—an age group that includes women of childbearing age. Cases of CRS in infants whose mothers have acquired rubella abroad (including three cases in 2012, one case in 2016, and one case thus far in 2017) continue to be identified and reported in the United States. Therefore, health care providers should remain vigilant, considering the possibility of rubella virus infection in adults (especially those emigrating or returning from countries without rubella control programs) and recognizing the potential for CRS among their infants.

 The Global Vaccine Action Plan 2011–2020 calls for the elimination of rubella in five of the six World Health Organization (WHO) regions by 2020. Although rubella and CRS are no longer endemic in the WHO Region of the Americas, they remain important public health problems globally. The number of rubella cases reported worldwide in 2000 was ~700,000; this figure declined to 22,427 in 2015. However, numbers of rubella cases may be underestimated because cases are often mild and may not be reported and, in some countries, are identified through measles surveillance systems that are not specific for rubella. Despite an increase in the number of countries with rubella vaccination programs, more than half of the world's children remained unvaccinated against rubella in 2015. In 2010, it was estimated that 105,000 cases of CRS occurred globally.

CLINICAL FEATURES

Acquired Rubella

Acquired rubella commonly presents with a generalized maculopapular rash that usually lasts for up to 3 days **(Fig. 201-1)**, although as many as 50% of cases may be subclinical or without rash. When it occurs, the rash is usually mild and may be difficult to detect in persons with darker skin. In children, rash is usually the first sign of illness. However, in older children and adults, a 1- to 5-day prodrome often precedes the rash and may include low-grade fever, malaise, and upper respiratory symptoms. The incubation period is 14 days (range, 12–23 days).

FIGURE 201-1

Mild maculopapular rash of rubella in a child.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Lymphadenopathy, particularly occipital and postauricular, may be noted during the second week after exposure. Although acquired rubella is usually thought of as a benign disease, arthralgia and arthritis are common in infected adults, particularly women. Thrombocytopenia and encephalitis are less common complications.

Congenital Rubella Syndrome

The most serious consequence of rubella virus infection can develop when a woman becomes infected during pregnancy, particularly during the first trimester. The resulting complications may include miscarriage, fetal death, premature delivery, or live birth with congenital defects. Infants infected with rubella virus in utero may have myriad physical defects **(Table 201-1)**, which most commonly relate to the eyes, ears, and heart. This constellation of severe birth defects is known as *CRS*. In addition to permanent manifestations, there are a host of transient physical manifestations, including thrombocytopenia with purpura/petechiae (e.g., dermal erythropoiesis, “blueberry muffin syndrome”). Some infants may be born with congenital rubella virus infection but have no apparent signs or symptoms of CRS and are referred to as “infants with congenital rubella virus infection only.”

TABLE 201-1 Common Transient and Permanent Manifestations in Infants with Congenital Rubella Syndrome

Transient Manifestations	Permanent Manifestations
Hepatosplenomegaly	Hearing impairment/deafness
Interstitial pneumonitis	Congenital heart defects (patent ductus arteriosus, pulmonary arterial stenosis)
Thrombocytopenia with purpura/petechiae (e.g., dermal erythropoiesis or “blueberry muffin syndrome”)	Eye defects (cataracts, cloudy cornea, microphthalmos, pigmentary retinopathy, congenital glaucoma)
Hemolytic anemia	Microcephaly
Bony radiolucencies	Central nervous system sequelae (mental and motor delay, autism)
Intrauterine growth retardation	
Adenopathy	
Meningoencephalitis	

DIAGNOSIS

Acquired Rubella

Clinical diagnosis of acquired rubella is difficult because of the mimicry of many illnesses with rashes, the varied clinical presentations, and the high rates of subclinical and mild disease. Illnesses that may be similar to rubella in presentation include scarlet fever, roseola, toxoplasmosis, fifth disease, measles, Zika, and illnesses with suboccipital and postauricular lymphadenopathy. Thus, laboratory documentation of rubella virus infection is considered the only reliable way to confirm acute disease.

Laboratory assessment of rubella virus infection is conducted by serologic and virologic methods. For acquired rubella, serologic diagnosis is most common and depends on the demonstration of IgM antibodies in an acute-phase serum specimen or a fourfold rise in IgG antibody titer between acute- and convalescent-phase specimens. The enzyme-linked immunosorbent assay IgM capture technique is considered most accurate for serologic diagnosis, but the indirect IgM assay also is acceptable. After rubella virus infection, IgM antibody may be detectable for up to 6 weeks. In case of a negative result for IgM in specimens taken earlier than day 5 after rash onset, serologic testing should be repeated. Although uncommon, reinfection with rubella virus is possible, and IgM antibodies may be present. To detect a rise in IgG antibody titer indicative of acute disease, the acute-phase serum specimen should be collected within 7–10 days after onset of illness and the convalescent-phase specimen ~14–21 days after the first specimen.

IgG avidity testing is used in conjunction with IgG testing. Low-avidity antibodies indicate recent infection. Mature (high-avidity) IgG antibodies most likely indicate an infection occurring at least 2 months previously. This test helps distinguish primary infection from reinfection. Avidity testing may be particularly useful in diagnosing rubella in pregnant women and assessing the risk of CRS.

Rubella virus can be isolated from the blood and nasopharynx during the prodromal period and for as long as 2 weeks after rash onset. However, as the secretion of virus in individuals with acquired rubella is maximal just before or up to 4 days after rash onset, this is the optimal time frame for collecting specimens for viral cultures. Rubella can also be diagnosed by viral RNA detection in a reverse-transcriptase polymerase chain reaction (RT-PCR) assay.

Congenital Rubella Syndrome

The classic triad of CRS—clinical manifestations of cataracts, hearing impairment, and heart defects—is seen in ~10% of infants with CRS. Infants may present with different combinations of defects depending on when infection occurs during gestation. Hearing impairment is the most common single defect of CRS. However, as with acquired rubella, laboratory diagnosis of congenital infection is highly recommended, particularly because most features of the clinical presentation are nonspecific and may be associated with other intrauterine infections. Early diagnosis of CRS facilitates appropriate medical intervention for specific disabilities and prompts implementation of infection control measures.

Diagnostic tests used to confirm CRS include serologic assays and virus detection. In an infant with congenital infection, serum IgM antibodies are normally present for up to 6 months but may be detectable for up to 1 year after birth. In some instances, IgM may not be detectable until 1 month of age; thus infants who have symptoms consistent with CRS but who test negative shortly after birth should be retested at 1 month. A rubella serum IgG titer persisting beyond the time expected after passive transfer of maternal IgG antibody (i.e., a rubella titer that does not decline at the expected rate of a twofold dilution per month) is another serologic criterion used to confirm CRS.

In congenital infection, rubella virus is isolated most commonly from throat swabs and less commonly from urine and cerebrospinal fluid. Infants with congenital rubella may excrete virus for up to 1 year, but specimens for virus isolation are most likely to be positive if obtained within the first 6 months after birth. Rubella virus in infants with CRS can also be detected by RT-PCR.

Rubella Diagnosis in Pregnant Women

In the United States, screening for rubella IgG antibodies is recommended as part of routine prenatal care. Pregnant women with a positive IgG antibody serologic test are considered immune. Susceptible pregnant women should be vaccinated postpartum.

A susceptible pregnant woman exposed to rubella virus should be tested for IgM antibodies and/or a fourfold rise in IgG antibody titer between acute- and convalescent-phase serum specimens to determine whether she was infected during pregnancy. Pregnant women with evidence of acute infection must be clinically monitored, and gestational age at the time of maternal infection must be determined to assess the possibility of risk to the fetus. Among women infected with rubella virus during the first 11 weeks of gestation, up to 90% deliver an infant with CRS; with maternal infection during the first 20 weeks of pregnancy, the CRS rate is 20%. Because of the potential for false-positive results, rubella IgM antibody testing is not recommended for pregnant women with no history of illness or contact with a rubella-like illness.

TREATMENT

TREATMENT Rubella

No specific therapy is available for rubella virus infection. Symptom-based treatment for various manifestations, such as fever and arthralgia, is appropriate. Immunoglobulin does not prevent rubella virus infection after exposure and therefore is not recommended as routine postexposure prophylaxis. Although immunoglobulin may modify or suppress symptoms, it can create an unwarranted sense of security: infants with congenital rubella have been born to women who received immunoglobulin shortly after exposure. Administration of immunoglobulin should be considered only if a pregnant woman who has been exposed to a person with rubella will not consider termination of the pregnancy under any circumstances. In such cases, IM administration of 20 mL of immunoglobulin within 72 h of rubella exposure may reduce—but does not eliminate—the risk of rubella.

PREVENTION

After the isolation of rubella virus in the early 1960s and the occurrence of a devastating pandemic in 1964–1965, a vaccine for rubella was developed and licensed in 1969. The majority of RCVs used worldwide are combined measles and rubella (MR) or measles, mumps, and rubella (MMR) formulations. A tetravalent measles, mumps, rubella, and varicella (MMRV) vaccine is available but is not widely used.

The public health burden of rubella virus infection is measured primarily through the occurrence of CRS cases among women who were infected during pregnancy. The 1964–1965 rubella epidemic in the United States resulted in >30,000 infections during pregnancy. CRS occurred in ~20,000 infants born alive, including >11,000 infants who were deaf, >3500 infants who were blind, and almost 2000 infants who were mentally retarded. The medical cost of this epidemic exceeded $1.5 billion. In 1983, the lifetime cost per child with CRS was estimated at $200,000.

In some countries, there are few data to document the epidemiology of CRS, but clusters of CRS cases have been reported in developing countries. Before the introduction of routine immunization against rubella in the United States, the incidence of CRS was 0.1–0.2 case per 1000 live births during endemic periods and 1–4 cases per 1000 live births during epidemic periods. Where rubella virus is circulating and women of childbearing age are susceptible, CRS cases will continue to occur.

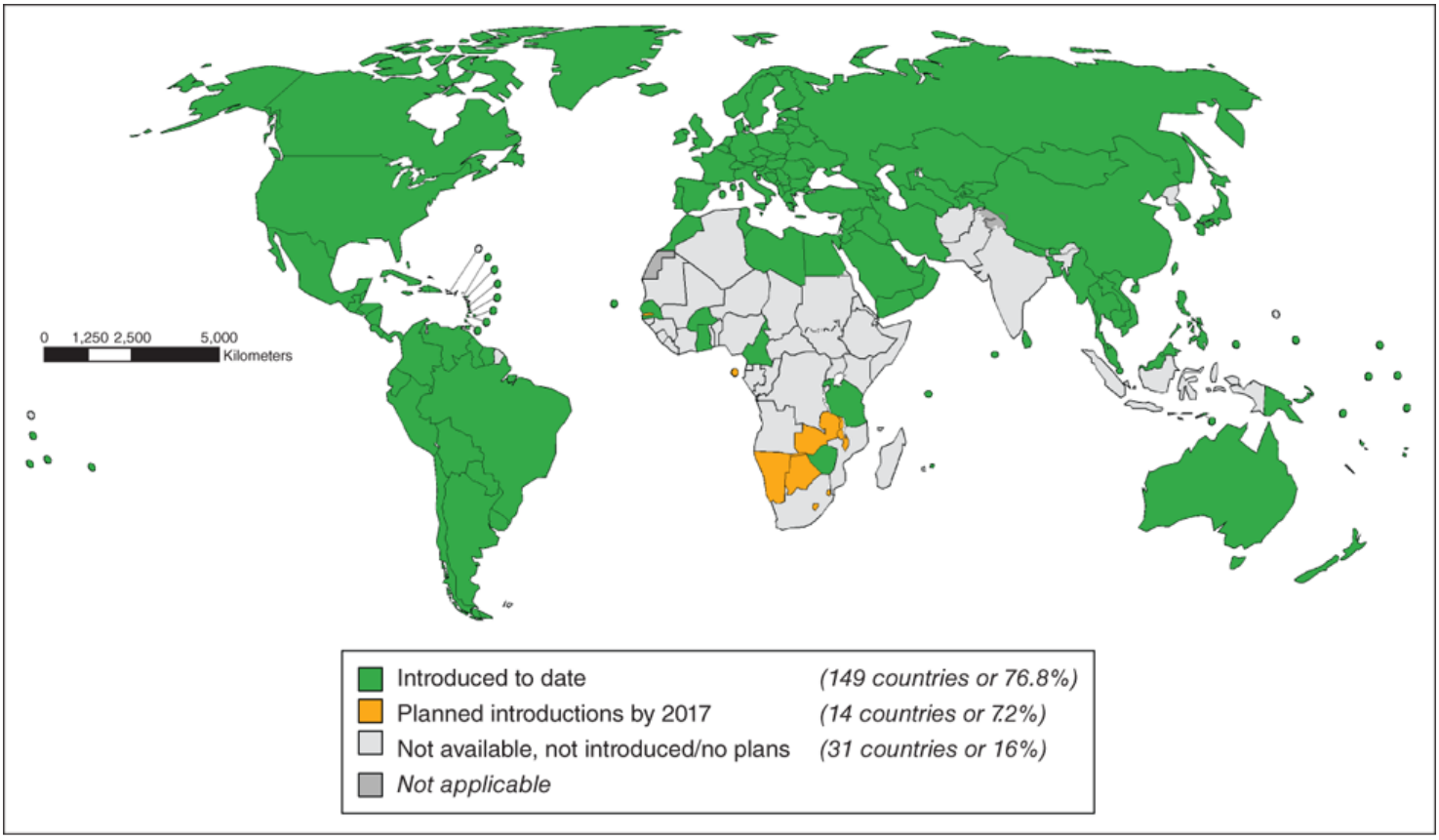
The most effective method of preventing acquired rubella and CRS is through vaccination with an RCV. One dose induces seroconversion in ≥95% of persons ≥1 year of age. Immunity is considered long-term and is probably lifelong. The most commonly used vaccine globally is the RA27/3 virus strain. The recommendation for routine rubella vaccination schedules in the United States is a first dose of MMR vaccine at 12–15 months of age and a second dose at 4–6 years. Target groups for rubella vaccine include children ≥1 year of age, adolescents and adults without documented evidence of immunity, individuals in congregate settings (e.g., college students, military personnel, child care and health care workers), and susceptible women before and after pregnancy.

Because of the theoretical risk of transmission of live attenuated rubella vaccine virus to the developing fetus, women known to be pregnant should not receive RCV. In addition, pregnancy should be avoided for 28 days after receipt of RCV. In follow-up studies of ~3000 unknowingly pregnant women who received rubella vaccine, no infant was born with CRS. Receipt of RCV during pregnancy is not ordinarily a reason to consider termination of the pregnancy.

In 2017, 149 (77%) of the 194 member countries of the WHO recommend inclusion of RCV in the routine childhood vaccination schedule; by the end of 2017, 14 more countries will follow suit **(Fig. 201-2)**. Goals for control or elimination of rubella and CRS have been established in the WHO American, European, South-East Asia, and Western Pacific regions. The Eastern Mediterranean and African regions have not yet set such goals.

FIGURE 201-2

Countries using rubella vaccine in national childhood immunization schedules, 2017. *(From the World Health Organization.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Centers for Disease Control and Prevention: Control and prevention of rubella: Evaluation and management of suspected outbreaks, rubella in pregnant women, and surveillance for congenital rubella syndrome. MMWR Morb Mortal Wkly Rep 50:1, 2001. [PubMed: 11215787]
Centers for Disease Control and Prevention: Notice to readers: Revised ACIP recommendation for avoiding pregnancy after receiving a rubella-containing vaccine. MMWR Morb Mortal Wkly Rep 50:1117, 2001. [PubMed: 11794623]
Centers for Disease Control and Prevention: Rubella, in *Manual for the Surveillance of Vaccine-Preventable Diseases*, 5th ed, SW Roush et al (eds). Atlanta, Centers for Disease Control and Prevention, 2012, Chapter 14. Available at *https://www.cdc.gov/vaccines/pubs/surv-manual/index.html*. Accessed December 4, 2017.
Centers for Disease Control and Prevention: Rubella, in *Epidemiology and Prevention of Vaccine Preventable Diseases*, 13th ed, Jennifer Hamborsky et al (eds). Washington, DC, Public Health Foundation, 2015, Chapter 18. Available at *https://www.cdc.gov/vaccines/pubs/pinkbook/index.html*. Accessed December 4, 2017.
Papania M et al: Nosocomial measles, mumps, rubella, and other viral infections, in *Hospital Epidemiology and Infection Control*, 3rd ed, CG Mayhall (ed). Philadelphia, Lippincott Williams and Wilkins, 2004, pp 828–49.
Reef S, Plotkin SA: Rubella vaccine, in *Vaccines*, SA Plotkin and WA Orenstein (eds). Philadelphia, Saunders, 2013, pp 688–717.
Vynnycky E et al: Using seroprevalence and immunisation coverage data to estimate the global burden of congenital rubella syndrome, 1996–2010: A systematic review. PLoS One 11:e0149160, 2016. [PubMed: 26962867]
World Health Organization: Rubella vaccines: WHO position paper. Wkly Epidemiol Rec 86:301, 2011. Available at *http://www.who.int/wer/2011/wer8629.pdf?ua=1*. Accessed December 4, 2017. [PubMed: 21766537]
World Health Organization: Rubella, module 11, in *The Immunological Basis for Immunization Series*. Geneva, WHO, 2008. Available at *http://www.who.int/immunization/documents/ISBN9789241596848/en/index.html*. Accessed December 4, 2017.
World Health Organization: Global Vaccine Action Plan 2011–2020. Geneva, WHO, 2013. Available at *http://www.who.int/immunization/global_vaccine_action_plan/GVAP_doc_2011_2020/en/*. Accessed December 4, 2017.

Chapter 202: Mumps

Steven A. Rubin

DEFINITION

Mumps is an illness characterized by acute-onset unilateral or bilateral tender, self-limited swelling of the parotid or other salivary gland(s) that lasts at least 2 days and has no other apparent cause.

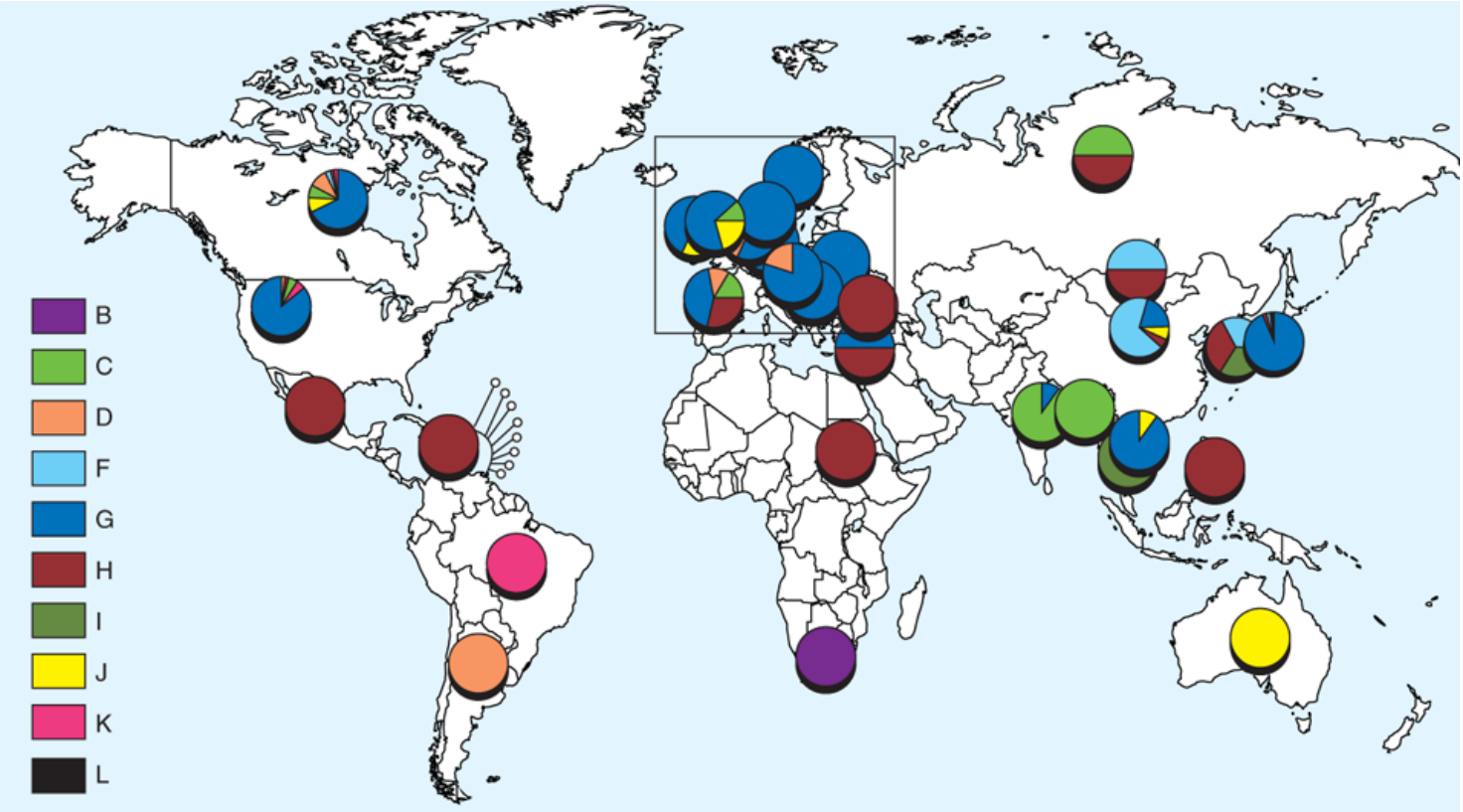
ETIOLOGIC AGENT


Mumps is caused by a paramyxovirus with a negative-strand, nonsegmented RNA genome of 15,384 bases encoding at least 8 proteins: the nucleo- (N), phospho- (P), V, matrix (M), fusion (F), small hydrophobic (SH), hemagglutinin–neuraminidase (HN), and large (L) proteins. The N, P, and L proteins together provide the polymerase activity responsible for genome transcription and replication. The viral genome is surrounded by a host cell–derived lipid bilayer envelope containing the M, F, SH, and HN proteins. The M protein is involved in viral assembly, whereas the HN and F proteins are responsible for cell attachment and entry and are the major targets of virus-neutralizing antibody. The V and SH proteins are accessory proteins, acting as antagonists of the host antiviral response; the former interferes with the interferon response and the latter with the tumor necrosis factor α (TNF-α)–mediated apoptotic signaling pathway. Because of the hypervariability of the SH gene, its nucleotide sequence is used to genotype the virus for molecular epidemiologic purposes. Thus far, 12 mumps virus genotypes have been assigned by SH gene sequence and are designated A–N (with the exclusion of E and M, which have been merged with genotypes C and K, respectively).


Nucleotide sequencing of clinical isolates shows that virus genotypes D and G circulate predominantly in the Western Hemisphere; genotypes F, C, and I in the Asia–Pacific region; and genotypes B, H, J, and K in the Southern Hemisphere **(Fig. 202-1)**. Although numerous mumps virus genotypes have been identified and some vary antigenically from others, only one serotype exists, and there is no evidence to suggest that certain circulating virus strains are more virulent or contagious than others.

FIGURE 202-1

Distribution of reported mumps genotypes, 2005–2016. Pie-slice size represents the frequency of the reported genotype. *(Figure courtesy of WHO, with permission. This figure has been updated by the author to include data on mumps virus genotypes reported since 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPIDEMIOLOGY

Mumps is endemic worldwide, with epidemics every 3–5 years in unvaccinated populations. These epidemics typically occur in locations where children and young adults congregate, such as schools, military barracks, and other institutions. In countries without national mumps vaccination programs, the estimated annual global incidence is 100–1000 cases per 100,000 population. After the introduction of mumps vaccine in the United States in 1967, the number of reported cases declined dramatically. Within a few years of the 1977 recommendation by the Advisory Committee on Immunization Practices (ACIP) for use of mumps-containing vaccine (measles–mumps–rubella; MMR) in the national immunization schedule, the number of reported cases declined by >97%. In the years following implementation of a two-dose schedule in 1989 and the passage of school immunization laws in 1991, fewer than 400 cases were being reported annually—a 99.8% reduction from prevaccine-era levels. Mumps incidence remained at historic lows in the United States until 2006, when 6584 cases were reported—the largest outbreak since 1987. At the time of the 2006 outbreak, the disease was resurging globally, even in populations with high-level vaccination coverage. Since this time, the mean number of mumps cases reported annually has increased by more than threefold over numbers in the prior decade (1996–2005). Vaccine coverage among kindergartners has remained relatively stable over the past 20 years at 94–95% (according to data compiled from the *Morbidity and Mortality Weekly Report*). This picture suggests that the resurgence in cases is not due to a failure to vaccinate—a notion supported by outbreak investigations showing that nearly all patients had a history of mumps vaccination.

Although historically a disease of childhood, with the largest proportion of cases occurring in children 5–9 years of age, mumps now most frequently occurs in older age groups, primarily college students. This shift in age distribution and the occurrence of mumps in vaccinated populations suggest waning of vaccine immunity. Thus the introduction of a third dose of MMR vaccine into the national immunization schedule to boost antibody titers has been investigated. However, in a study of 656 subjects, this additional dose only transiently raised levels of neutralizing antibody to mumps virus: levels returned to baseline within a year. While this temporary increase in antibody titers could reduce mumps attack rates during outbreaks, the value of the third dose for routine use in vaccinated populations was determined to be low. Therefore, this option is not being pursued.

PATHOGENESIS

Humans are the only natural hosts for mumps virus, although a virus with >90% sequence identity to mumps virus has recently been identified in bats. The incubation period of mumps is ~19 days (range, 7–23 days). The virus is transmitted by the respiratory route via droplets, saliva, and fomites. Mumps virus is typically shed from 1 week before to 1 week after symptom onset, although this window appears to be narrower in vaccinated individuals. Persons are most contagious 1–2 days before onset of clinical symptoms. Inference from related respiratory diseases and animal studies indicates that primary replication likely occurs in the nasal mucosa or upper respiratory mucosal epithelium. Mononuclear cells and cells within regional lymph nodes can become infected; such infection facilitates the development of viremia, posing a risk for a wide array of acute inflammatory reactions, most commonly parotitis and

orchitis. Other common sites of virus dissemination include the kidneys (reflected in the frequency of viruria) and the central nervous system (CNS). Less common sites include the pancreas, heart, ovaries, mammary glands, perilymphatic fluid within the cochlea, and (during pregnancy) the fetus.

Little is known about the pathology of mumps since the disease is rarely fatal. Affected salivary glands contain perivascular and interstitial mononuclear-cell infiltrates and exhibit hemorrhage with prominent edema. Necrosis of acinar and epithelial duct cells is evident in the salivary glands and in the germinal epithelium of the seminiferous tubules of the testes. The virus probably enters cerebrospinal fluid (CSF) through the choroid plexus or via transiting mononuclear cells during plasma viremia. Although relevant data are limited, in most cases mumps encephalitis appears to be secondary to respiratory spread and is probably a parainfectious or postinfectious process (as suggested by perivenous demyelination and perivascular mononuclear-cell inflammation) rather than a result of direct virus damage to brain tissue. However, although rare, primary mumps encephalitis does occur, as shown by mumps virus isolation from brain tissue. Infection of the perilymphatic fluid likely develops via retrograde penetration by the virus from the cervical lymph nodes following viremia, but infection could also occur via the CSF in cases of mumps CNS infection, given that the perilymph communicates with the CSF. Virus in the perilymph can result in infection of the cochlea and damage to the organ of Corti and the tectorial membrane, leading to transient or permanent deafness. Evidence of placental and intrauterine spread has been found in both early and late gestation. Mumps during the first trimester of pregnancy increases the risk of miscarriage, but there is no evidence that mumps during pregnancy increases the risk of premature delivery or birth defects.

CLINICAL MANIFESTATIONS

Up to half of mumps virus infections are asymptomatic or lead to nonspecific respiratory symptoms. Inapparent infections are more common among adults than among children. The prodrome of mumps consists of low-grade fever, malaise, myalgia, headache, and anorexia. Mumps parotitis or swelling of other salivary glands usually occurs within 24 h of prodromal symptoms but sometimes as long as 1 week thereafter. Parotitis is generally bilateral, although the two sides may not be involved synchronously. Unilateral involvement occurs in about one-third of cases. Swelling of the parotid is accompanied by tenderness and obliteration of the space between the earlobe and the angle of the mandible **(Figs. 202-2 and 202-3)**. The patient frequently reports an earache and finds it difficult to eat, swallow, or talk. The orifice of the parotid duct is commonly red and swollen. The submaxillary and sublingual glands are involved less often than the parotid gland and are almost never involved alone. Glandular swelling increases for a few days and then gradually subsides, disappearing within 1 week. Recurrent sialadenitis is a rare sequela of mumps parotitis. In ~6% of mumps cases, obstruction of lymphatic drainage secondary to bilateral salivary-gland swelling may lead to presternal pitting edema, associated often with submandibular adenitis and rarely with the more life-threatening supraglottic edema.

FIGURE 202-2

The same person before mumps acquisition (*A*) and on day 3 of acute bilateral parotitis (*B*). *(Courtesy of patient C.M. From JD Shanley: The resurgence of mumps in young adults and adolescents. Cleve Clin J Med 74:42, 2007. Reprinted with permission. Copyright © 2007 Cleveland Clinic Foundation. All rights reserved.)*

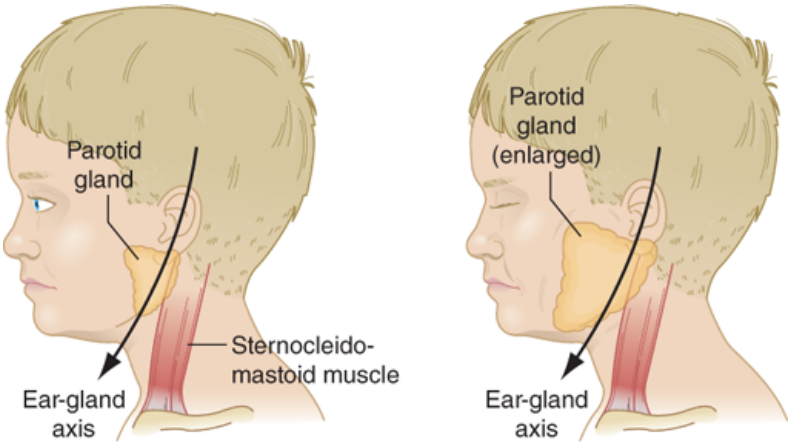


A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 202-3

Schematic drawings of a normal parotid gland *(left)* and a parotid gland infected with mumps virus (*right*). An enlarged cervical lymph node is usually posterior to the imaginary line. *(Reprinted with permission from A Gershon et al: Mumps, in Krugman's Infectious Diseases of Children, 11th ed. Philadelphia, Elsevier, 2004, p 392.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Epididymo-orchitis is the next most common manifestation of mumps, developing in 15–30% of cases in postpubertal males, with bilateral involvement in 10–30% of those cases. Orchitis, accompanied by fever, typically occurs during the first week of parotitis but can develop up to 6 weeks after parotitis or in its absence. The testis is painful and tender and can be enlarged to several times its normal size; this condition usually resolves within 1 week. Testicular atrophy develops in one-half of affected men. Sterility after mumps is rare, although subfertility is estimated to occur in 13% of cases of unilateral orchitis and in 30–87% of cases of bilateral orchitis. Oophoritis occurs in ~5% of women with mumps and may be associated with lower abdominal pain and vomiting, but has only rarely been associated with sterility or premature menopause. Mumps infection in postpubertal women may also present with mastitis.

Documented CSF pleocytosis indicates that the virus invades the CNS in ~50% of cases; however, symptomatic CNS disease, typically in the form of aseptic meningitis, occurs in <10% of cases, with a male predominance. Symptoms of aseptic meningitis include stiff neck, headache, and drowsiness and typically appear ~5 days after parotitis but can occur in the absence of parotid involvement. Mumps meningitis is a self-limited manifestation without significant risk of death or long-term sequelae. In ~0.1% of infections, mumps virus may cause encephalitis, which presents as high fever with marked changes in the level of consciousness, seizures, and focal neurologic symptoms. Electroencephalographic abnormalities may be seen. Permanent sequelae are sometimes identified in survivors, and adult infections more commonly have poor outcomes than do pediatric infections. The mortality rate associated with mumps encephalitis is ~1.5%. Other CNS problems occasionally associated with mumps include cerebellar ataxia, facial palsy, transverse myelitis, hydrocephalus, Guillain-Barré syndrome, flaccid paralysis, and behavioral changes. Mumps deafness, which may or may not be related to CNS infection, occurs in 1 in 1000 to 1 in 100,000 mumps cases.

Mumps pancreatitis, which may present as abdominal pain, occurs in ~4% of infections but is difficult to diagnose because an elevated serum amylase level can be associated with either parotitis or pancreatitis. An etiologic association of mumps virus and juvenile diabetes mellitus remains controversial. Myocarditis and endocardial fibroelastosis are rare and self-limited but may represent severe complications of mumps infection; however, mumps-associated electrocardiographic abnormalities have been reported in up to 15% of cases. Other unusual complications include thyroiditis, nephritis, arthritis, hepatic disease, keratouveitis, and thrombocytopenic purpura. Abnormal renal function is common, but severe, life-threatening nephritis is rare.

DIFFERENTIAL DIAGNOSIS

During a mumps outbreak, the diagnosis is made easily in patients with parotitis and a history of recent exposure; however, when disease incidence is low, other causes of parotitis should be considered and laboratory testing is required for case confirmation. In two recent studies, 13–15% of patients with suspected mumps parotitis who tested negative for mumps virus by polymerase chain reaction (PCR) tested positive for influenza A virus. Other viruses known to cause parotitis include HIV, coxsackievirus, parainfluenza virus type 3, Epstein-Barr virus, adenovirus, parvovirus B19, lymphocytic choriomeningitis virus, and human herpesvirus 6. Gram-positive bacteria, atypical mycobacteria, and *Bartonella* species can cause parotitis. Parotitis can also develop in the setting of sarcoidosis, Sjögren's syndrome, Mikulicz's syndrome, Parinaud's syndrome, uremia, diabetes mellitus, laundry starch ingestion, malnutrition, cirrhosis, and some drug treatments. Unilateral parotitis can be caused by ductal obstruction, cysts, and tumors. In the absence of parotitis or other salivary gland enlargement, symptoms of other visceral-organ and/or CNS involvement may predominate, and a laboratory diagnosis is required. Other entities should be considered when manifestations consistent with mumps appear in organs other than the parotid. For example, testicular torsion may produce a painful scrotal mass resembling that seen in mumps orchitis, and a number of viruses (e.g., enteroviruses) can cause aseptic meningitis that is clinically indistinguishable from that due to mumps virus.

LABORATORY DIAGNOSIS

Laboratory diagnosis is primarily based on detection of viral RNA by reverse-transcriptase PCR (RT-PCR) or on serology. Detection of viral antigens (e.g., via mumps virus–specific immunofluorescent staining of cultured clinical specimens) is comparatively inefficient and is no longer commonly performed.

For RT-PCR-based testing, viral RNA can be extracted directly from clinical samples. Buccal swabs appear to be the best specimens for virus detection, particularly when obtained within 2 days of clinical onset; however, mumps virus can also be detected readily in throat swabs and saliva and, in cases of meningitis, in CSF. Despite the apparent high frequency of viremia during mumps, mumps virus has rarely been detected in blood. The ability to detect viral RNA in clinical samples rapidly diminishes beyond the first week after symptom onset, and in several studies rates of virus detection were substantially lower in recipients of two doses of mumps-containing vaccine than in unvaccinated persons or recipients of one dose of vaccine. The rate of false-negative RT-PCR findings can be quite high, approaching 70% in some studies. Nonetheless, RT-PCR-based testing is more sensitive than other methods and is the preferred means of case confirmation. While serologic testing is commonly performed (typically via enzyme-linked immunosorbent assay [ELISA]), its diagnostic value is complicated by the fact that most people are vaccinated and may not mount a detectable IgM response upon reinfection since IgM is not a major component of the anamnestic response. Thus, a negative IgM result does not necessarily rule out mumps. In addition, regardless of vaccination status, IgM may not be detectable if serum is assayed too early (prior to day 3 of symptom onset) or too late (beyond 6 weeks after symptom onset) in the course of disease. Reliance on a rise in IgG titer in paired acute- and convalescent-phase sera also is problematic: IgG titers in convalescent-phase sera may be only nominally greater than those in acute-phase sera. Traditional and labor-intensive serologic tests such as complement fixation, hemagglutination inhibition, and virus neutralization are now performed only rarely. The main downside to replacement of these functional serologic assays with the more rapid ELISA method is the latter's detection of all virus-specific antibodies, including those that are non-neutralizing and may not be protective. Thus, an individual who is seropositive by ELISA may lack protective levels of antibody. While there is a strong association between the presence of neutralizing antibody to mumps virus and protection from disease, an antibody titer predictive of serologic protection is lacking; in this respect, mumps differs from other respiratory infections, such as measles and influenza.

TREATMENT

TREATMENT Mumps

Mumps is generally a benign, self-resolving illness. Therapy for parotitis and other clinical manifestations is symptom based and supportive. The administration of analgesics and the application of warm or cold compresses to the parotid area may be helpful. Testicular pain may be minimized by the local application of cold compresses and gentle support for the scrotum. Anesthetic blocks also may be used. Neither the administration of glucocorticoids nor incision of the tunica albuginea is of proven value in severe orchitis. Anecdotal information on a small number of patients with orchitis suggests that SC administration of interferon α2b may help preserve the organ and fertility. Lumbar puncture is occasionally performed to relieve headache associated with meningitis. Mumps immune globulin has not been consistently effective in preventing mumps and is not recommended for treatment or postexposure prophylaxis.

PREVENTION

Vaccination is the only practical control measure. Nearly all developed countries use mumps-containing vaccines, but in many countries mumps is not a notifiable disease and vaccination is voluntary. However, where used, mumps vaccination has had a tremendous impact, with reductions in incidence and morbidity typically exceeding 90%. Despite the tremendous success of mumps vaccination programs, large mumps outbreaks continue to occur globally, even in settings of high-level two-dose vaccine coverage.

In the United States, the benefit–cost ratios for mumps vaccination alone are >13 for direct costs (e.g., medical expenses) and >24 for societal costs (including productivity losses for patients and caregivers). Several mumps virus vaccines are used throughout the world; in the United States, only the live attenuated Jeryl Lynn strain is used. Current recommendations are that mumps vaccine be administered as part of the combined trivalent MMR vaccine (M-M-R®$_{II}$) or the quadrivalent measles–mumps–rubella–varicella vaccine (ProQuad®). Monovalent vaccine is no longer produced for the U.S. market but is available in other countries.

Before administering mumps-containing vaccine, physicians should consult the latest recommendations from the ACIP. Current recommendations for children specify two doses of mumps-containing vaccine: the first dose given on or after the first birthday and the second dose administered no earlier than 28 days after the first. In the United States, children often receive the second dose between the ages of 4 and 6 years.

In 2009, the ACIP revised its recommendations for evidence of mumps immunity in health care personnel to include (1) documented administration of two doses of a preparation containing live mumps vaccine, (2) laboratory evidence of immunity or laboratory confirmation of disease, or (3) birth date before 1957. For unvaccinated health care personnel born before 1957 who lack laboratory evidence of mumps immunity or laboratory confirmation of mumps, health care facilities should consider two doses of MMR vaccine separated by the appropriate interval; during a mumps outbreak, vaccination of these individuals is recommended.

Mumps vaccine contains live attenuated virus. It is not recommended for pregnant women, for individuals who have had a life-threatening allergic reaction to components of the vaccine, for persons with evidence of severe immunosuppression, or for persons receiving chemotherapy or long-term immunosuppressive therapy. (For details, see the ACIP guidelines on the CDC's website: *www.cdc.gov/vaccines/hcp/acip-recs/*.) Occasionally, febrile reactions and parotitis have been reported soon after mumps vaccination. Allergic reactions after vaccination (e.g., rash and pruritus) are uncommon and are usually mild and self-limited. More serious complications, such as aseptic meningitis, have been causally associated with certain vaccine strains but not with the Jeryl Lynn strain.

Immunity to mumps is associated with the development of neutralizing antibody, although a specific correlate of protection has not been established. Seroconversion occurs in ~95% of recipients of the Jeryl Lynn strain; however, the vaccine efficacy rate is ~78% for

one dose and 88% for two doses. In several studies, seropositivity rates and vaccine efficacy have declined with time since vaccination. Administration of an additional dose of vaccine results in a transient increase in antibody levels, but the restorative effect of this increase on vaccine efficacy remains to be demonstrated. The effectiveness of vaccine use during mumps outbreaks was studied in school-based outbreaks in Guam, New York, and Illinois. In all, the intervention was linked with reduced attack rates, but either statistical significance could not be established because of the small number of cases recorded or it was impossible to evaluate effectiveness because the intervention was initiated after the outbreak started to decline. The role of the cellular arm of the immune response is unclear, but there is evidence that it may help limit the spread of mumps virus and the development of complications.

ACKNOWLEDGMENT

The authors thank and acknowledge Drs. Anne Gershon and Kathryn M. Carbone, authors of this chapter in earlier editions.

FURTHER READING

Dayan GH et al: Recent resurgence of mumps in the United States. N Engl J Med 358:1580, 2008. [PubMed: 18403766]
Fiebelkorn AP et al: Mumps postexposure prophylaxis with a third dose of measles-mumps-rubella vaccine, Orange County, New York, USA. Emerg Infect Dis 19:1411, 2013. [PubMed: 23965729]
Huang AS et al: Risk factors for mumps at a university with a large mumps outbreak. Public Health Rep 124:419, 2009. [PubMed: 19445418]
Lebaron CW et al: Persistence of mumps antibodies after 2 doses of measles-mumps-rubella vaccine. J Infect Dis 199:552, 2009. [PubMed: 19113988]
Polgreen PM et al: The duration of mumps virus shedding after the onset of symptoms. Clin Infect Dis 46:1447, 2008. [PubMed: 18419451]
Rota JS et al: Comparison of the sensitivity of laboratory diagnostic methods from a well-characterized outbreak of mumps in New York city in 2009. Clin Vaccine Immunol 20:391, 2013. [PubMed: 23324519]

Chapter 203: Rabies and Other Rhabdovirus Infections

Alan C. Jackson

RABIES

Rabies is a rapidly progressive, acute infectious disease of the central nervous system (CNS) in humans and animals that is caused by infection with rabies virus. The infection is normally transmitted from animal vectors. Rabies has encephalitic and paralytic forms that progress to death.

ETIOLOGIC AGENT

Rabies virus is a member of the family Rhabdoviridae. Two genera in this family, *Lyssavirus* and *Vesiculovirus*, contain species that cause human disease. Rabies virus is a lyssavirus that infects a broad range of animals and causes serious neurologic disease when transmitted to humans. This single-strand RNA virus has a nonsegmented, negative-sense (antisense) genome that consists of 11,932 nucleotides and encodes 5 proteins: nucleocapsid protein, phosphoprotein, matrix protein, glycoprotein, and a large polymerase protein. Rabies virus variants, which can be characterized by distinctive nucleotide sequences, are associated with specific animal reservoirs. Six other non–rabies virus species in the *Lyssavirus* genus have been reported to cause a clinical picture similar to rabies. Vesicular stomatitis virus, a vesiculovirus, causes vesiculation and ulceration in cattle, horses, and other animals and causes a self-limited, mild, systemic illness in humans (see "Other Rhabdoviruses," below).

EPIDEMIOLOGY

Rabies is a zoonotic infection that occurs in a variety of mammals throughout the world except in Antarctica and on some islands. Rabies virus is usually transmitted to humans by the bite of an infected animal. Worldwide, endemic canine rabies is estimated to cause 59,000 human deaths annually. Most of these deaths occur in Asia and Africa, with rural populations and children most frequently affected. Thus, in many resource-poor and resource-limited countries, canine rabies continues to be a threat to humans. However, in Latin America, rabies control efforts in dogs have been quite successful in recent years. Endemic canine rabies has been eliminated from the United States and most other resource-rich countries. Rabies is endemic in wildlife species, and a variety of animal reservoirs have been identified in different countries of the world **(Fig. 203-1)**. Surveillance data from 2015 identified 5508 confirmed animal cases of rabies in the United States and Puerto Rico. Only 7.6% of these cases were in domestic animals, including 244 cases in cats, 67 in dogs, and 85 in cattle. In North American wildlife reservoirs, including bats, raccoons, skunks, and foxes, the infection is endemic, with involvement of one or more rabies virus variants in each reservoir species **(Fig. 203-2)**. "Spillover" of rabies to other wildlife species and to domestic animals occurs. Bat rabies virus variants are present in every state except Hawaii and are responsible for most indigenously acquired human rabies cases in the United States. Raccoon rabies is endemic along the entire eastern coast of the United States. Skunk rabies is present in the midwestern states, with another focus in California. Rabies in foxes occurs in Texas, New Mexico, Arizona, and Alaska.

FIGURE 203-1

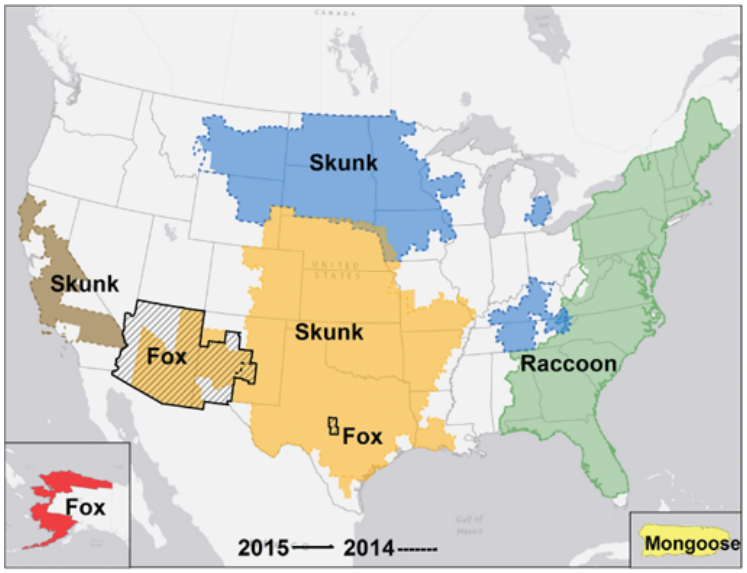
Distribution of global rabies vectors. *(Courtesy of the Centers for Disease Control and Prevention.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 203-2

Distribution of the major rabies virus variants among wild terrestrial reservoirs in the United States and Puerto Rico, 2009–2015. Black diagonal lines represent gray fox rabies variants. Solid borders represent 5-year rabies virus variant aggregates for 2011 through 2015; dashed borders represent the previous 5-year aggregates for 2010 through 2014. *(From MG Birhane et al: J Am Vet Med Assoc 250:1117, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In Canada and Europe, epizootics of rabies in red foxes have been well controlled with the use of baits containing rabies vaccine. A similar approach, along with additional measures, is used in Canada to control incursions of raccoon rabies.

Rabies virus variants isolated from humans or other mammalian species can be identified by reverse-transcription polymerase chain reaction (RT-PCR) amplification and sequencing or by characterization with monoclonal antibodies. These techniques are helpful in human cases with no known history of exposure. Worldwide, most human rabies is transmitted from dogs in countries with endemic canine rabies and dog-to-dog transmission, and human cases can be imported by travelers returning from these regions. In North America, human disease is usually associated with transmission from bats; there may be no known history of bat bite or other bat exposure in these cases. Most human cases are due to a bat rabies virus variant associated with silver-haired and tricolored bats. These are small bats whose bite may not be recognized, and the virus has adapted for replication at skin temperature and in cell types that are present in the skin.

Transmission from nonbite exposures is relatively uncommon. Aerosols generated in the laboratory or in caves containing millions of Brazilian free-tail bats have rarely caused human rabies. Transmission has resulted from corneal transplantation and also from solid-organ

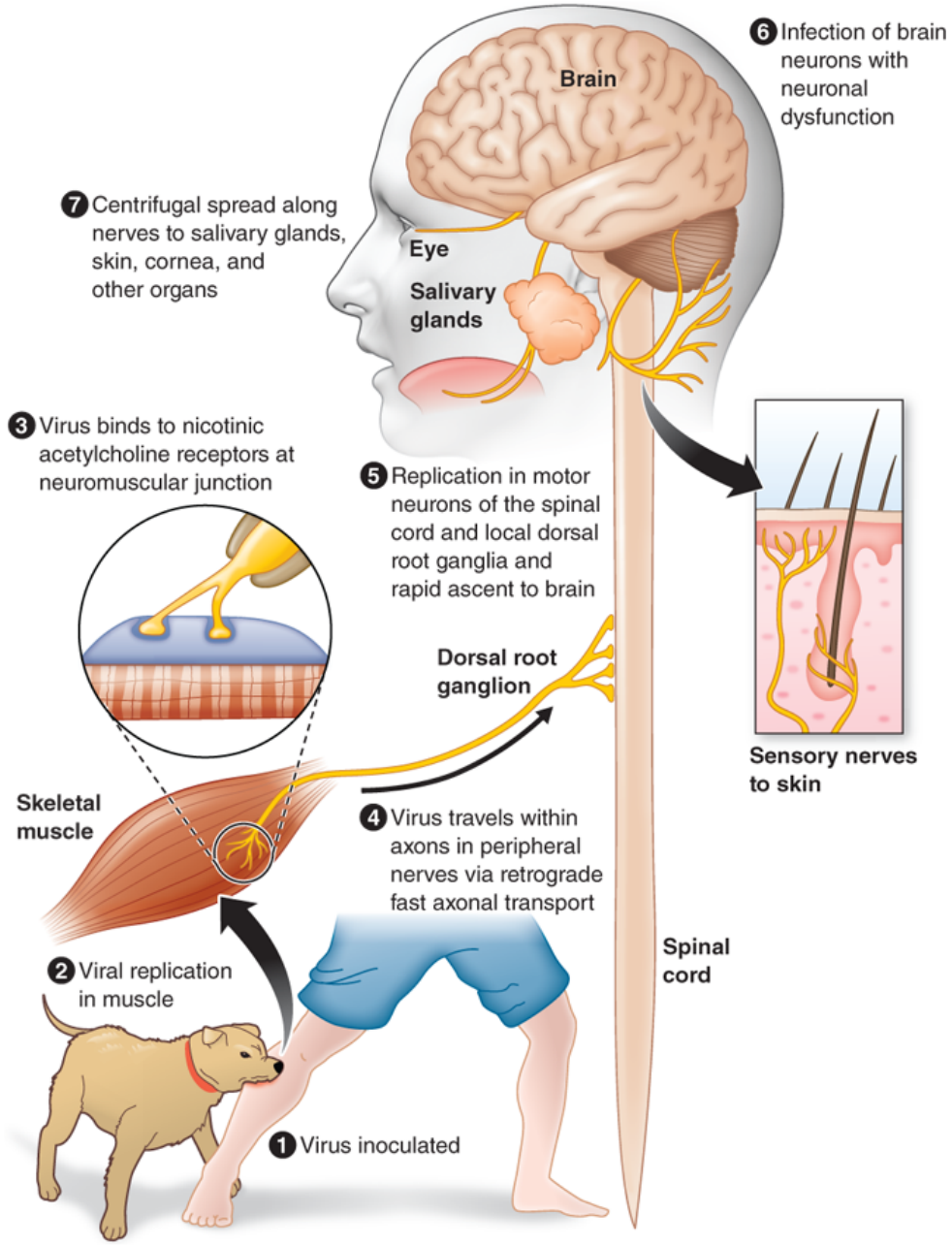
transplantation and a vascular conduit (for a liver transplant) from undiagnosed donors with rabies in Texas, Florida, Germany, Kuwait, and China. Human-to-human transmission is extremely rare, although hypothetical concern about transmission to health care workers has prompted the implementation of barrier techniques to prevent exposures from patients with rabies.

PATHOGENESIS

The incubation period of rabies (defined as the interval between exposure and the onset of clinical disease) is usually 20–90 days, but in rare cases is either as short as a few days or >1 year. During most of the incubation period, rabies virus is thought to be present at or close to the site of inoculation **(Fig. 203-3)**. In muscles, the virus is known to bind to nicotinic acetylcholine receptors on postsynaptic membranes at neuromuscular junctions, but the exact details of viral entry into the skin and SC tissues have not yet been clarified. Rabies virus spreads centripetally along peripheral nerves toward the spinal cord or brainstem via retrograde fast axonal transport (rate, up to ~250 mm/d), with delays at intervals of ~12 h at each synapse. Once the virus enters the CNS, it rapidly disseminates to other regions of the CNS via fast axonal transport along neuroanatomic connections. Neurons are prominently infected in rabies; infection of astrocytes is unusual. After CNS infection becomes established, there is centrifugal spread along sensory and autonomic nerves to other tissues, including the salivary glands, heart, adrenal glands, and skin. Rabies virus replicates in acinar cells of the salivary glands and is secreted in the saliva of rabid animals that serve as vectors of the disease. There is no well-documented evidence for hematogenous spread of rabies virus.

FIGURE 203-3

Schematic representation of events in rabies pathogenesis following peripheral inoculation of rabies virus by an animal bite. *(Adapted from AC Jackson: Human disease, in Rabies: Scientific basis of the disease and its management, 3rd ed. AC Jackson [ed], Oxford, UK, Elsevier Academic Press, 2013, pp 269–298; adapted with permission.)*

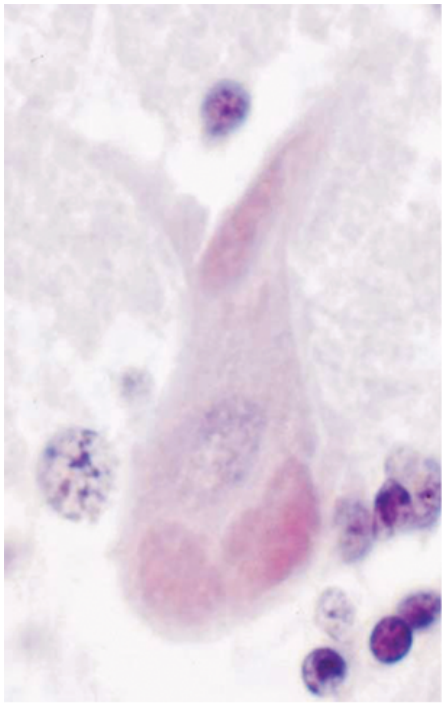


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pathologic studies show mild inflammatory changes in the CNS in rabies, with mononuclear inflammatory infiltration in the leptomeninges, perivascular regions, and parenchyma, including microglial nodules called *Babes nodules*. Degenerative neuronal changes usually are not prominent, and there is little evidence of neuronal death; neuronophagia is observed occasionally. The pathologic changes are surprisingly mild in light of the clinical severity and fatal outcome of the disease. The most characteristic pathologic finding in rabies is the *Negri body* **(Fig. 203-4)**. Negri bodies are eosinophilic cytoplasmic inclusions in brain neurons that are composed of rabies virus proteins and viral RNA. These inclusions occur in a minority of infected neurons, are commonly observed in Purkinje cells of the cerebellum and in pyramidal neurons of the hippocampus, and are less frequently seen in cortical and brainstem neurons. Negri bodies are not observed in all cases of rabies. The lack of prominent degenerative neuronal changes has led to the concept that neuronal dysfunction—rather than neuronal death—is responsible for clinical disease in rabies. The basis for behavioral changes, including the aggressive behavior of rabid animals, is not well understood but may be related to infection of serotonergic neurons in the brainstem.

FIGURE 203-4

Three large Negri bodies in the cytoplasm of a cerebellar Purkinje cell from an 8-year-old boy who died of rabies after being bitten by a rabid dog in Mexico. *(From AC Jackson, E Lopez-Corella: N Engl J Med 335:568, 1996. © Massachusetts Medical Society.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CLINICAL MANIFESTATIONS

In rabies, the emphasis must be on postexposure prophylaxis (PEP) initiated after a recognized exposure and before any symptoms or signs develop. Rabies should usually be suspected on the basis of the clinical presentation. The disease generally presents as atypical encephalitis with relative preservation of consciousness. Rabies may be difficult to recognize late in the clinical course when progression to coma has occurred. A minority of patients present with acute flaccid paralysis. There are prodromal, acute neurologic, and comatose phases that usually progress to death despite aggressive therapy **(Table 203-1)**.

TABLE 203-1 Clinical Stages of Rabies

Stage	Typical Duration	Symptoms and Signs
Incubation period	20–90 days	None
Prodrome	2–10 days	Fever, malaise, anorexia, nausea, vomiting; paresthesias, pain, or pruritus at the wound site
Acute Neurologic Disease		
Encephalitic (80%)	2–7 days	Anxiety, agitation, hyperactivity, bizarre behavior, hallucinations, autonomic dysfunction, hydrophobia
Paralytic (20%)	2–10 days	Flaccid paralysis in limb(s) progressing to quadriparesis with facial paralysis
Coma, death[a]	0–14 days	

[a]Recovery is rare.

Source: MAW Hattwick: Rabies virus, in *Principles and Practice of Infectious Diseases,* GL Mandell et al (eds). New York, Wiley, 1979, pp 1217–1228. Adapted with permission from Elsevier.

Prodromal Features

The clinical features of rabies begin with nonspecific prodromal manifestations, including fever, malaise, headache, nausea, and vomiting. Anxiety or agitation may also occur. The earliest specific neurologic symptoms of rabies include paresthesias, pain, or pruritus near the site of the exposure, one or more of which occur in 50–80% of patients and strongly suggest rabies. The wound has usually healed by this point, and these symptoms probably reflect infection with associated inflammatory changes in local dorsal root or cranial sensory ganglia.

Encephalitic Rabies

Two acute neurologic forms of rabies are seen in humans: the encephalitic (furious) form in 80% and the paralytic form in 20%. Some of the manifestations of encephalitic rabies, including fever, confusion, hallucinations, combativeness, and seizures, may be seen in other viral encephalitides as well. Autonomic dysfunction is common and may result in hypersalivation, gooseflesh, cardiac arrhythmia, and priapism. In encephalitic rabies, episodes of hyperexcitability are typically followed by periods of complete lucidity that become shorter as the disease progresses. Rabies encephalitis is distinguished by early brainstem involvement, which results in the classic features of *hydrophobia* (involuntary, painful contraction of the diaphragm and accessory respiratory, laryngeal, and pharyngeal muscles in response to swallowing liquids) **(Fig. 203-5)** and *aerophobia* (the same features caused by stimulation from a draft of air). These symptoms are probably due to dysfunction of infected brainstem neurons that normally inhibit inspiratory neurons near the nucleus ambiguus, resulting in exaggerated defense reflexes that protect the respiratory tract. The combination of hypersalivation and pharyngeal dysfunction is also responsible for the classic appearance of “foaming at the mouth”. Brainstem dysfunction progresses rapidly, and coma—followed within days by death—is the

rule unless the course is prolonged by supportive measures. With such measures, late complications can include cardiac and/or respiratory failure, disturbances of water balance (syndrome of inappropriate antidiuretic hormone secretion or diabetes insipidus), noncardiogenic pulmonary edema, and gastrointestinal hemorrhage. Cardiac arrhythmias may be due to dysfunction affecting vital centers in the brainstem or to myocarditis. Multiple-organ failure is common in patients treated aggressively in critical care units.

FIGURE 203-5

Hydrophobic spasm of inspiratory muscles associated with terror in a patient with encephalitic (furious) rabies who is attempting to swallow water. *(Copyright DA Warrell, Oxford, UK; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Paralytic Rabies

About 20% of patients have paralytic rabies in which muscle weakness predominates and cardinal features of encephalitic rabies (hyperexcitability, hydrophobia, and aerophobia) are lacking. There is early and prominent flaccid muscle weakness, often beginning in the bitten extremity and spreading to produce quadriparesis and facial weakness. Sphincter involvement is common, sensory involvement is usually mild, and these cases are commonly misdiagnosed as Guillain-Barré syndrome. Patients with paralytic rabies generally survive a few days longer than those with encephalitic rabies, but multiple-organ failure nevertheless ensues.

LABORATORY INVESTIGATIONS

Most routine laboratory tests in rabies yield normal results or show nonspecific abnormalities. Complete blood counts are usually normal. Examination of cerebrospinal fluid (CSF) often reveals mild mononuclear-cell pleocytosis with a mildly elevated protein level. Severe pleocytosis (>1000 white cells/μL) is unusual and should prompt a search for an alternative diagnosis. Imaging is usually performed to exclude other diagnostic possibilities. CT head scans are usually normal in rabies. MRI brain scans may show signal abnormalities in the brainstem or other gray-matter areas, but these findings are variable and nonspecific. Electroencephalograms typically show only nonspecific abnormalities. Of course, important tests in suspected cases of rabies include those that may identify an alternative, potentially treatable diagnosis (see "Differential Diagnosis," below).

DIAGNOSIS

In North America, a diagnosis of rabies often is not considered until relatively late in the clinical course, even with a typical clinical presentation. This diagnosis should be considered in patients presenting with acute atypical encephalitis or acute flaccid paralysis, including those in whom Guillain-Barré syndrome is suspected. The absence of an animal-bite history is common in North America. The lack of hydrophobia is not unusual in rabies. Once rabies is suspected, rabies-specific laboratory tests should be performed to confirm the diagnosis. Diagnostically useful specimens include serum, CSF, fresh saliva, skin biopsy samples from the neck, and brain tissue (rarely obtained before death). Because skin biopsy relies on the demonstration of rabies virus antigen in cutaneous nerves at the base of hair follicles, samples are usually taken from hairy skin at the nape of the neck. Corneal impression smears are of low diagnostic yield and are generally not performed. Negative antemortem rabies-specific laboratory tests never exclude a diagnosis of rabies, and tests may need to be repeated after an interval for diagnostic confirmation.

Rabies Virus–Specific Antibodies

In a previously unimmunized patient, serum neutralizing antibodies to rabies virus are diagnostic. However, because rabies virus infects immunologically privileged neuronal tissues, serum antibodies may not develop until late in the disease. Antibodies may be detected within a few days after the onset of symptoms, but some patients die without detectable antibodies. The presence of rabies virus–specific neutralizing antibodies in the CSF suggests rabies encephalitis, regardless of immunization status. A diagnosis of rabies is questionable in patients who recover from their illness without developing serum neutralizing antibodies to rabies virus.

RT-PCR Amplification

Detection of rabies virus RNA by RT-PCR is highly sensitive and specific. This technique can detect virus in fresh saliva samples, skin biopsy specimens, CSF, and brain tissues. In addition, RT-PCR with genetic sequencing can distinguish among rabies virus variants, permitting identification of the probable source of an infection.

Direct Fluorescent Antibody Testing

Direct fluorescent antibody (DFA) testing with rabies virus antibodies conjugated to fluorescent dyes is highly sensitive and specific for the detection of rabies virus antigen in tissues; the test can be performed quickly and applied to skin biopsy and brain tissue samples. In skin biopsy samples, rabies virus antigen may be detected in cutaneous nerves at the base of hair follicles.

DIFFERENTIAL DIAGNOSIS

The diagnosis of rabies may be difficult without a history of animal exposure, and no exposure to an animal (e.g., a bat) may be recalled. The presentation of rabies is usually quite different from that of acute viral encephalitis due to most other causes, including herpes simplex encephalitis and arboviral (e.g., West Nile) encephalitis. Early neurologic symptoms may occur at the site of the bite, and there may be early features of brainstem involvement with preservation of consciousness. Anti–*N*-methyl-D-aspartate receptor (anti-NMDA) encephalitis occurs in young patients (especially females) and is characterized by behavioral changes, autonomic instability, hypoventilation, and seizures. Many other antibodies are associated with autoimmune encephalitis. Postinfectious (immune-mediated) encephalomyelitis may follow influenza, measles, mumps, and other infections; it may also occur as a sequela of immunization with rabies vaccines derived from neural tissues, which are used only in resource-limited and resource-poor countries. Rabies may present with unusual neuropsychiatric symptoms and may be misdiagnosed as a psychiatric disorder. Rabies hysteria (now classified as a somatic symptom disorder) may occur as a psychological response to the fear of rabies and is often characterized by a shorter incubation period than rabies, aggressive behavior, inability to communicate, and a long course with recovery.

As previously mentioned, paralytic rabies may mimic Guillain-Barré syndrome. In these cases, fever, bladder dysfunction, a normal sensory examination, and CSF pleocytosis favor a diagnosis of rabies. Conversely, Guillain-Barré syndrome may occur as a complication of rabies vaccination with a neural tissue–derived product (e.g., suckling mouse brain vaccine) and may be mistaken for paralytic rabies (i.e., vaccine failure).

TREATMENT

TREATMENT Rabies

There is no established treatment for rabies. Aggressive management with supportive care in critical care units has resulted in the survival of more than 15 patients with rabies. There have been many recent treatment failures (~40) with the combination of antiviral drugs, ketamine, and therapeutic (induced) coma—measures that were used in a healthy survivor in whom neutralizing antibodies to rabies virus were detected at presentation. Expert opinion is recommended before a course of experimental therapy is embarked upon. A palliative approach may be appropriate for many patients.

PROGNOSIS

Rabies is an almost uniformly fatal disease but is nearly always preventable after recognized exposures with appropriate postexposure therapy during the early incubation period (see below). All but one of more than 15 documented survivors of rabies received one or more doses of rabies vaccine before disease onset. The single survivor who had not received vaccine had neutralizing antibodies to rabies virus in serum and CSF at clinical presentation. Most patients with rabies die within several days of the onset of illness, despite aggressive care in a critical care unit.

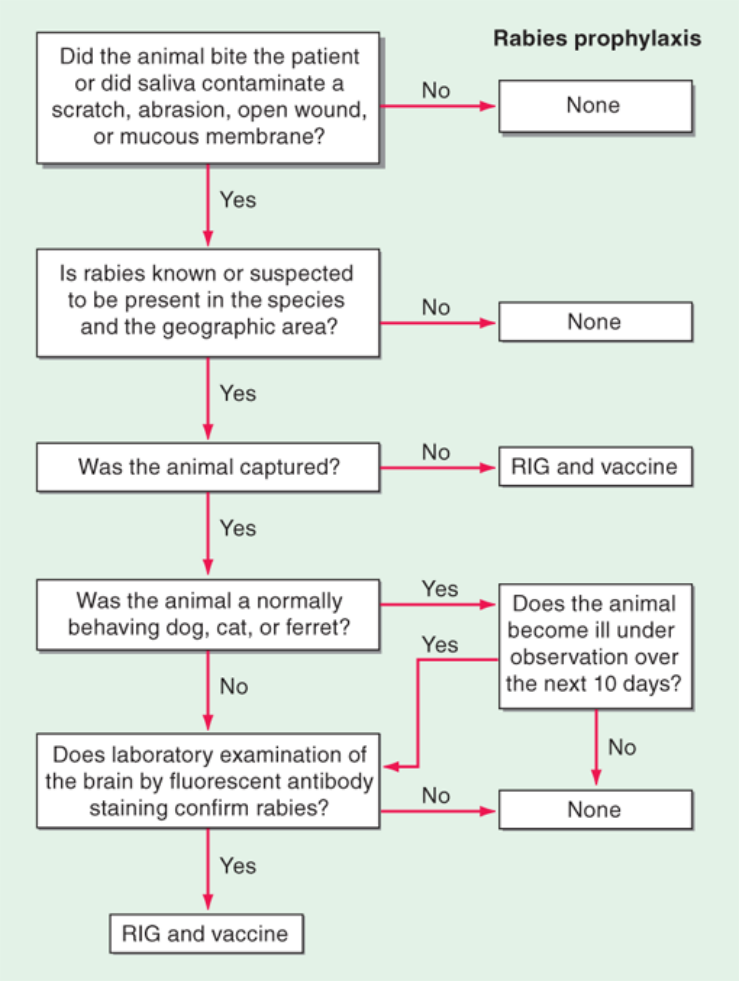
PREVENTION

Postexposure Prophylaxis

Since there is no effective therapy for rabies, it is extremely important to prevent the disease after an animal exposure. **Figure 203-6** shows the steps involved in making decisions about PEP. On the basis of the exposure history and local epidemiologic information, the physician must decide whether initiation of PEP is warranted. Healthy dogs, cats, or ferrets may be confined and observed for 10 days. PEP is not necessary if the animal remains healthy. If the animal develops signs of rabies during the observation period, it should be euthanized immediately; the head should be transported to the laboratory under refrigeration, rabies virus should be sought by DFA testing, and viral isolation should be attempted by cell culture and/or mouse inoculation. Any animal other than a dog, cat, or ferret should be euthanized immediately and the head submitted for laboratory examination. In high-risk exposures and in areas where canine rabies is endemic, rabies prophylaxis should be initiated without waiting for laboratory results. If the laboratory results prove to be negative, it may safely be concluded that the animal's saliva did not contain rabies virus, and immunization should be discontinued. If an animal escapes after an exposure, it must be considered rabid, and PEP must be initiated unless information from public health officials indicates otherwise (i.e., there is no endemic rabies in the area). Although controversial, the use of PEP may be warranted when a person (e.g., a small child or a sleeping adult) has been present in the same space as a bat and an unrecognized bite cannot be reliably excluded.

FIGURE 203-6

Algorithm for rabies postexposure prophylaxis. RIG, rabies immune globulin. (*From L Corey, in Harrison's Principles of Internal Medicine, 15th ed. E Braunwald et al [eds]: New York, McGraw-Hill, 2001; adapted with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PEP includes local wound care and both active and passive immunization. It is important that current recommendations are followed very closely because minor deviations can lead to failure of prophylactic measures. Local wound care is essential and may greatly decrease the risk of rabies virus infection. Wound care should not be delayed, even if the initiation of immunization is postponed pending the results of the 10-day observation period. All bite wounds and scratches should be washed thoroughly with soap and water. Devitalized tissues should be debrided, tetanus prophylaxis given, and antibiotic treatment initiated whenever indicated.

All previously unvaccinated persons (but not those who have previously been immunized) should be passively immunized with rabies immune globulin (RIG). If RIG is not immediately available, it should be administered no later than 7 days after the first vaccine dose. After day 7, endogenous antibodies are being produced, and passive immunization may actually be counterproductive. If anatomically feasible, the entire dose of RIG (20 IU/kg) should be infiltrated at the site of the bite, and any RIG remaining after infiltration of the bite site should be administered IM at a distant site. With multiple or large wounds, the RIG preparation may need to be diluted in order to obtain a sufficient volume for adequate infiltration of all wound sites. If the exposure involves a mucous membrane, the entire dose should be administered IM. Rabies vaccine and RIG should never be administered at the same site or with the same syringe. Commercially available RIG in the United States is purified from the serum of hyperimmunized human donors. These human RIG preparations are much better tolerated than are the equine-derived preparations still in use in some countries (see below). Serious adverse effects of human RIG are uncommon. Local pain and low-grade fever may occur.

Two purified inactivated rabies vaccines are available for rabies PEP in the United States. They are highly immunogenic and remarkably safe compared with earlier vaccines. Four 1-mL doses of rabies vaccine should be given IM in the deltoid area. (The anterolateral aspect of the thigh also is acceptable in children.) Gluteal injections, which may not always reach muscle, should not be given and have been associated with rare vaccine failures. Ideally, the first dose should be given as soon as possible after exposure; failing that, it should be given without further delay. The three additional doses should be given on days 3, 7, and 14; a fifth dose on day 28 is no longer recommended. Pregnancy is not a contraindication for immunization. Glucocorticoids and other immunosuppressive medications may interfere with the development of active immunity and should not be administered during PEP unless they are essential. Routine measurement of serum neutralizing antibody titers is not required, but titers should be measured 2–4 weeks after immunization in immunocompromised persons. Local reactions (pain, erythema, edema, and pruritus) and mild systemic reactions (fever, myalgias, headache, and nausea) are common; anti-inflammatory and antipyretic medications may be used, but immunization should not be discontinued. Systemic allergic reactions are uncommon, but

anaphylaxis does occur rarely and can be treated with epinephrine and antihistamines. The risk of rabies development should be carefully considered before the decision is made to discontinue vaccination because of an adverse reaction.

Most of the burden of rabies PEP is borne by persons with the fewest resources. In addition to the rabies vaccines discussed above, vaccines grown in either primary cell lines (hamster or dog kidney) or continuous cell lines (Vero cells) are satisfactory and are available in many countries outside the United States. Less expensive vaccines derived from neural tissues are still used in a diminishing number of developing countries; however, these vaccines are associated with serious neuroparalytic complications, including postinfectious encephalomyelitis and Guillain-Barré syndrome. The use of these vaccines should be discontinued as soon as possible, and progress has been made in this regard. Worldwide, more than 10 million individuals receive postexposure rabies vaccine each year.

If human RIG is unavailable, purified equine RIG can be used in the same manner at a dose of 40 IU/kg. The incidence of anaphylactic reactions and serum sickness has been low with recent equine RIG products.

Preexposure Rabies Vaccination

Preexposure rabies prophylaxis should be considered for people with an occupational or recreational risk of rabies exposures and also for certain travelers to rabies-endemic areas. The primary schedule consists of three doses of rabies vaccine given on days 0, 7, and 21 or 28. Serum neutralizing antibody tests help determine the need for subsequent booster doses. When a previously immunized individual is exposed to rabies, two booster doses of vaccine should be administered on days 0 and 3. Wound care remains essential. As stated above, RIG should not be administered to previously vaccinated persons.

OTHER RHABDOVIRUSES

OTHER LYSSAVIRUSES

A growing number of lyssaviruses other than rabies virus have been discovered to infect bat populations in Europe, Africa, Asia, and Australia. Six of these viruses have produced a very small number of cases of a human disease indistinguishable from rabies: European bat lyssaviruses 1 and 2, Australian bat lyssavirus, Irkut virus, and Duvenhage virus. Mokola virus, a lyssavirus that has been isolated from shrews with an unknown reservoir species in Africa, may also produce human disease indistinguishable from rabies.

VESICULAR STOMATITIS VIRUS

Vesicular stomatitis is a viral disease of cattle, horses, pigs, and some wild mammals. Vesicular stomatitis virus is a member of the genus *Vesiculovirus* in the family Rhabdoviridae. Outbreaks of vesicular stomatitis in horses and cattle occur sporadically in the southwestern United States. The animal infection is associated with severe vesiculation and ulceration of oral tissues, teats, and feet and may be clinically indistinguishable from the more dangerous foot-and-mouth disease. Epidemics are usually seasonal, typically beginning in the late spring, and are probably due to arthropod vectors. Direct animal-to-animal spread can also occur, although the virus cannot penetrate intact skin. Transmission to humans usually results from direct contact with infected animals (particularly cattle) and occasionally follows laboratory exposure. In human disease, early conjunctivitis is followed by an acute influenza-like illness with fever, chills, nausea, vomiting, headache, retrobulbar pain, myalgias, substernal pain, malaise, pharyngitis, and lymphadenitis. Small vesicular lesions may be present on the buccal mucosa or on the fingers. Encephalitis is very rare. The illness usually lasts 3–6 days, with complete recovery. Subclinical infections are common. A serologic diagnosis can be made on the basis of a rise in titer of complement-fixing or neutralizing antibodies. Therapy is symptom-based.

FURTHER READING

Fooks AR et al: Current status of rabies and prospects for elimination. Lancet 384:1389, 2014. [PubMed: 24828901]
Jackson AC (ed): *Rabies: Scientific Basis of the Disease and Its Management*, 3rd ed. Oxford, UK, Elsevier Academic Press, 2013.
Letchworth GJ et al: Vesicular stomatitis. Vet J 157:239, 1999. [PubMed: 10328837]
Manning SE et al: Human rabies prevention—United States, 2008: Recommendations of the Advisory Committee on Immunization Practices. MMWR Recomm Rep 57(RR-3):1, 2008. [PubMed: 18496505]
World Health Organization: *WHO Expert Consultation on Rabies: Second Report* (WHO Technical Report Series No. 982). Geneva, World Health Organization, 2013. (Available at *http://apps.who.int/iris/bitstream/10665/85346/1/9789240690943_eng.pdf*. Accessed January 2, 2018.)

Chapter 204: Arthropod-Borne and Rodent-Borne Virus Infections

Jens H. Kuhn; Rémi N. Charrel

INTRODUCTION

This chapter summarizes the major features of selected arthropod-borne and rodent-borne viruses. Numerous viruses of this category are transmitted in nature among animals without ever infecting humans. Other viruses incidentally infect humans, but only a proportion of these viruses induce human disease. In addition, certain viral agents are regularly introduced into human populations or spread among humans by certain arthropods (specifically, insects and ticks) or by chronically infected rodents. These zoonotic viruses are taxonomically diverse and therefore differ fundamentally from one another in terms of virion morphology, replication strategies, genomic organization, and genome sequence. While a virus's classification in a taxon is enlightening with regard to natural maintenance strategies, sensitivity to antiviral agents, and particular aspects of pathogenesis, the classification does not necessarily predict which clinical signs and symptoms (if any) the virus will cause in humans. Zoonotic viruses are evolving, and "new" zoonotic viruses are regularly discovered. The epizootiology and epidemiology of zoonotic viruses continue to change because of environmental alterations affecting vectors, reservoirs, wildlife, livestock, and humans. Zoonotic viruses are most numerous in the tropics but are also found in temperate and even frigid climates. The distribution and seasonal activity of a zoonotic virus may vary, and the rate at which they change is likely to depend largely on ecologic conditions (e.g., rainfall and temperature), which can affect the density of virus vectors and reservoirs and the development of infection.

Arthropod-borne viruses (arboviruses) infect their vectors after ingestion of a blood meal from a viremic, usually nonhuman vertebrate; some arthropods may also become infected by saliva-activated transmission. The arthropod vectors then develop chronic, systemic infection as the viruses penetrate the gut and spread throughout the body to the salivary glands; such virus dissemination, referred to as *extrinsic incubation*, typically lasts 1–3 weeks in mosquitoes. At this point, if the salivary glands become involved, the arthropod vector is competent to continue the chain of transmission by infecting a vertebrate during a subsequent blood meal. An alternative mechanism for virus maintenance in its arthropod vector is *transovarial transmission*. The arthropod generally is unharmed by the infection, and the natural vertebrate partner usually has only transient viremia with no overt disease.

Rodent-borne viruses are maintained in nature by transmission between rodents, which become chronically infected. Usually a high degree of rodent–virus specificity is observed, and overt disease in the reservoir host is rare.

ETIOLOGY

Arthropod-borne and rodent-borne zoonotic viruses belong to the proposed orders "*Articulavirales*" (family *Orthomyxoviridae*), the order *Bunyavirales* (families *Arenaviridae*, *Hantaviridae, Nairoviridae, Peribunyaviridae, Phenuiviridae*), and the order *Mononegavirales* (family *Rhabdoviridae*) and to the unassigned families *Flaviviridae, Reoviridae,* and *Togaviridae* **(Table 204-1)**.

TABLE 204-1 Zoonotic Arthropod- and Rodent-Borne Viruses That Infect Humans

Virus Group	Virus (Abbreviation)	Principal Reservoir Host(s)	Vector(s)	Syndrome[a]
Alphaviruses (Barmah Forest serocomplex)	Barmah Forest virus (BFV)	Horses, marsupials	Biting midges (*Culicoides marksi*), mosquitoes (*Aedes camptorhynchus, A. normanensis, A. notoscriptus, A. vigilax, Culex annulirostris*)	A/R
Alphaviruses (Semliki Forest serocomplex)	Chikungunya virus (CHIKV)	Bats, nonhuman primates	Mosquitoes (*Aedes, Culex* spp.)	A/R[b]
	Mayaro virus (MAYV)	Nonhuman primates, possums, rodents	Mosquitoes (predominantly *Haemagogus* spp.)	A/R
	O'nyong-nyong virus (ONNV)	Unknown	Mosquitoes (in particular *Anopheles gambiae*, *A. funestus*, *Mansonia* spp.)	A/R
	Una virus (UNAV)	Birds, horses, rodents	Mosquitoes (*Aedes*, *Anopheles*, *Coquillettidia*, *Culex*, *Ochlerotatus, Psorophora* spp.)	F/M
	Ross River virus (RRV)	Macropods, rodents	Mosquitoes (*Aedes normanensis*, *A. vigilax*, *Culex annulirostris*)	A/R
	Semliki Forest virus (SFV)	Birds, rodents	Mosquitoes (*Aedes*, *Culex* spp.)	A/R
Alphaviruses (eastern equine encephalitis serocomplex)	Eastern equine encephalitis virus (EEEV)	Freshwater swamp birds	Mosquitoes (*Aedes*, *Coquillettidia*, *Culex* spp.; *Culiseta melanura*, *Mansonia perturbans*, *Psorophora* spp.)	E
Alphaviruses (Venezuelan equine encephalitis serocomplex)	Everglades virus (EVEV)	Hispid cotton rats (*Sigmodon hispidus*)	Mosquitoes (*Culex cedecei*)	F/M, E
	Mucambo virus (MUCV)	Nonhuman primates, rodents	Mosquitoes (*Culex*, *Ochlerotatus* spp.)	F/M, E
	Tonate virus (TONV)	Suriname crested oropendolas (*Psarocolius decumanus*)	Mosquitoes (*Culex portesi*)	F/M, E
	Venezuelan equine encephalitis virus (VEEV)	Horses, rodents	Mosquitoes (*Aedes*, *Culex* spp.; *Psorophora confinnis*)	F/M, E
Alphaviruses (western equine encephalitis serocomplex)	Sindbis virus (SINV)	Birds	Mosquitoes (*Culex*, *Culiseta* spp.)	A/R
	Western equine encephalitis virus (WEEV)	Lagomorphs, passerine birds	Mosquitoes (*Culex tarsalis*)	E
"Banyangviruses" (Bhanja serocomplex)	Bhanja virus[c] (BHAV)	Cattle, four-toed hedgehog (*Atelerix albiventris*), goats, sheep, striped ground squirrels (*Xerus erythropus*)	Ixodid ticks (*Amblyomma*, *Dermacentor*, *Haemaphysalis*, *Hyalomma*, *Rhipicephalus* spp.)	E, F/M
	Heartland virus	Cattle, deer, elk, goats, sheep?	Ixodid ticks (*Amblyomma americanum*)	F/M
	Severe fever with thrombocytopenia syndrome virus (SFTSV)[d]	Cattle, chicken, dogs, goats, rodents, sheep?	Ixodid ticks (*Haemaphysalis longicornis*, *Rhipicephalus microplus*)	F/M, VHF
Bunyaviruses (family and genus undetermined)	Bangui virus (BGIV)	Unknown	Unknown	F/M
Coltiviruses	Colorado tick fever virus (CTFV)	Bushy-tailed woodrats (*Neotoma cinerea*), Columbian ground squirrels (*Spermophilus columbianus*), deer mice (*Peromyscus maniculatus*), golden-mantled ground squirrels (*Spermophilus lateralis*), least chipmunks (*Tamias minimus*), North American porcupines (*Erethizon dorsata*), yellow pine chipmunks (*Tamias amoenus*)	Ixodid ticks (predominantly *Dermacentor andersoni*)	E, F/M
	Eyach virus (EYAV)	Lagomorphs, rodents	Ixodid ticks (*Ixodes ricinus*, *I. ventalloi*)	E, F/M
	Salmon River virus (SRV)	Unknown	Ixodid ticks (*Ixodes* spp.)	E, F/M
Flaviviruses (mosquito-borne)	Dengue viruses 1–4 (DENV 1–4)	Nonhuman primates	Mosquitoes (predominantly *Aedes aegypti*, *A. albopictus*)	F/M, VHF

Virus Group	Virus (Abbreviation)	Principal Reservoir Host(s)	Vector(s)	Syndrome[a]
	Japanese encephalitis virus (JEV)	Ardeid wading birds (in particular herons), horses, pigs	Mosquitoes (*Culex* spp., in particular *C. tritaeniorhynchus*)	E
	Kokobera virus (KOKV)	Macropods, horses	Mosquitoes (*Culex* spp.)	A/R
	Murray Valley encephalitis virus (MVEV)	Birds	Mosquitoes (predominantly *Culex annulirostris*)	E
	Rocio virus (ROCV)	Rufous-collared sparrows (*Zonotrichia capensis*)	Mosquitoes (*Aedes*, *Culex*, *Psorophora* spp.)	E
	St. Louis encephalitis virus (SLEV)	Columbiform and passeriform birds (finches, sparrows)	Mosquitoes (predominantly *Culex* spp., in particular *C. nigripalpus*, *C. pipiens*, *C. quinquefasciatus*, *C. tarsalis*)	E
	Usutu virus (USUV)	Passerine birds	Mosquitoes (*Culex* spp., in particular *C. pipiens*)	(E)
Flaviviruses (mosquito-borne)	West Nile virus (WNV)[e]	Passerine birds (blackbirds, crows, finches, sparrows), small mammals, horses	Mosquitoes (*Culex* spp., in particular *C. pipiens*, *C. quinquefasciatus*, *C. restuans*, *C. tarsalis*)	E
	Yellow fever virus (YFV)	Nonhuman primates (*Alouatta*, *Ateles*, *Cebus*, *Cercopithecus*, *Colobus* spp.)	Mosquitoes (*Aedes* spp., in particular *A. aegypti*)	VHF
	Zika virus (ZIKV)	Nonhuman primates (*Macaca*, *Pongo* spp.)	Mosquitoes (*Aedes* spp.)	A/R, F/M
Flaviviruses (tick-borne)	Kyasanur Forest disease virus (KFDV)[f]	Indomalayan vandeleurias (*Vandeleuria oleracea*), roof rats (*Rattus rattus*)	Ixodid ticks (predominantly *Haemaphysalis spinigera*), sand tampans (*Ornithodoros savignyi*)	VHF
	Omsk hemorrhagic fever virus (OHFV)	Migratory birds, rodents	Ixodid ticks (predominantly *Dermacentor* spp.)	VHF
	Powassan virus (POWV)	Red squirrels (*Tamiasciurus hudsonicus*), white-footed deer mice (*Peromyscus leucopus*), woodchucks (*Marmota monax*), other small mammals	Ixodid ticks (in particular *Ixodes cookei*, other *Ixodes* spp., *Dermacentor* spp.)	E
	Tick-borne encephalitis virus (TBEV)	Passerine birds, deer, eulipotyphla, goats, grouse, small mammals, rodents, sheep	Ixodid ticks (*Ixodes gibbosus*, *I. persulcatus*, *I. ricinus*; sporadically *Dermacentor*, *Haemaphysalis*, *Hyalomma* spp.)	E, F/M, (VHF)
Mammarenaviruses (Old World)	Lassa virus (LASV)	Natal mastomys (*Mastomys natalensis*)	None	F/M, VHF
	Lujo virus (LUJV)	Unknown	None	VHF
	Lymphocytic choriomeningitis virus (LCMV)	House mice (*Mus musculus*)	None	E, F/M, (VHF)
Mammarenaviruses (New World)	Chapare virus (CHAPV)	Unknown	None	VHF
	Guanarito virus (GTOV)	Short-tailed zygodonts (*Zygodontomys brevicauda*)	None	VHF
	Junín virus (JUNV)	Drylands lauchas (*Calomys musculinus*)	None	VHF
	Machupo virus (MACV)	Big lauchas (*Calomys callosus*)	None	VHF
	Sabiá virus (SABV)	Unknown	None	VHF
	Whitewater Arroyo virus (WWAV)[g]	White-throated woodrats (*Neotoma albigula*)	None	(E)
Orbiviruses	Kemerovo virus (KEMV)	Birds, rodents	Ixodid ticks (*Ixodes persulcatus*)	E, F/M
	Lebombo virus (LEBV)	Unknown	Mosquitoes (*Aedes*, *Mansonia* spp.)	F/M
	Orungo virus (ORUV)	Camels, cattle, goats, nonhuman primates, sheep	Mosquitoes (*Aedes*, *Anopheles*, *Culex* spp.)	E, F/M
	Tribeč virus (TRBV)[h]	Bank voles (*Myodes glareolus*), birds, common pine voles (*Microtus subterraneus*), goats, hares	Ixodid ticks (*Ixodes persulcatus*, *I. ricinus*)	F/M

Virus Group	Virus (Abbreviation)	Principal Reservoir Host(s)	Vector(s)	Syndrome[a]
Orthobunyaviruses (Anopheles A serogroup)	Tacaiuma virus (TCMV)	Nonhuman primates	Mosquitoes (*Anopheles*, *Haemagogus* spp.)	F/M
Orthobunyaviruses (Bunyamwera serogroup)	Batai virus (BATV)[i]	Birds, camels, cattle, goats, rodents, sheep	Mosquitoes (*Aedes abnormalis*, *A. curtipes*, *Anopheles barbirostris*, *Culex gelidus*, other spp.)	F/M
	Bunyamwera virus (BUNV)	Birds, cows, goats, horses, sheep	Mosquitoes (*Aedes* spp.)	F/M
	Cache Valley virus (CVV)	Cattle, deer, foxes, horses, nonhuman primates, raccoons	Mosquitoes (*Aedes*, *Anopheles*, *Culiseta* spp.)	F/M
	Fort Sherman virus (FSV)	Unknown	Mosquitoes?	F/M
	Germiston virus (GERV)	Rodents	Mosquitoes (*Culex* spp.)	F/M
	Guaroa virus (GROV)	Unknown	Mosquitoes (*Anopheles* spp.)	F/M
	Ilesha virus (ILEV)	Unknown	Mosquitoes (*Anopheles gambiae*)	F/M, (VHF)
	Maguari virus (MAGV)	Birds, cattle, horses, sheep, water buffalo	Mosquitoes (*Aedes*, *Anopheles*, *Culex*, *Psorophora*, *Wyeomyia* spp.)	F/M
	Ngari virus (NRIV)	Unknown	Mosquitoes (*Aedes*, *Anopheles* spp.)	F/M, VHF
	Shokwe virus (SHOV)	Rodents	Mosquitoes (*Aedes*, *Anopheles*, *Mansonia* spp.)	F/M
	Xingu virus (XINV)	Unknown	Unknown	F/M
Orthobunyaviruses (Bwamba serogroup)	Bwamba virus (BWAV)	Unknown	Mosquitoes (*Aedes*, *Anopheles*, *Mansonia* spp.)	F/M
	Pongola virus (PGAV)	Cattle, donkeys, goats, sheep	Mosquitoes (*Aedes*, *Anopheles*, *Mansonia* spp.)	F/M
Orthobunyaviruses (California serogroup)	California encephalitis virus (CEV)	Lagomorphs, rodents	Mosquitoes (*Aedes*, *Culex*, *Culiseta*, *Psorophora* spp.)	E, F/M
	Inkoo virus (INKV)	Cattle, foxes, hares, moose, rodents	Mosquitoes (*Aedes* spp.)	E, F/M
	Jamestown Canyon virus (JCV)	Bison, deer, elk, moose	Mosquitoes (*Aedes*, *Culiseta*, *Ochlerotatus* spp.)	E, F/M
	La Crosse virus (LACV)	Chipmunks, squirrels	Mosquitoes (*Ochlerotatus triseriatus*)	E, F/M
	Lumbo virus (LUMV)	Unknown	Mosquitoes (*Aedes pembaensis*)	E, F/M
	Snowshoe hare virus (SSHV)	Snowshoe hares, squirrels, other small mammals	Mosquitoes (*Aedes*, *Culiseta*, *Ochlerotatus* spp.)	E, F/M
	Ťahyňa virus (TAHV)	Cattle, dogs, eulipotyphla, foxes, hares, horses, pigs, rodents	Mosquitoes (*Aedes*, *Culex*, *Culiseta* spp.)	E, F/M
Orthobunyaviruses (group C serogroup)	Apeú virus (APEUV)	Bare-tailed woolly opossums (*Caluromys philander*) and other opossums; rodents; tufted capuchins (*Cebus apella*)	Mosquitoes (*Aedes*, *Culex* spp.)	F/M
	Caraparú virus (CARV)	Rodents, tufted capuchins (*C. apella*)	Mosquitoes (*Culex* spp.)	F/M
	Itaquí virus (ITQV)	Capuchins (*Cebus* spp.), opossums, rodents	Mosquitoes (*Culex* spp.)	F/M
	Madrid virus (MADV)	Capuchins (*Cebus* spp.), opossums, rodents	Mosquitoes (*Culex* spp.)	F/M
	Marituba virus (MTBV)	Capuchins (*Cebus* spp.), opossums, rodents	Mosquitoes (*Culex* spp.)	F/M
	Murutucú virus (MURV)	Capuchins (*Cebus* spp.), opossums, pale-throated sloths (*Bradypus tridactylus*), rodents	Mosquitoes (*Coquillettidia*, *Culex* spp.)	F/M
	Nepuyo virus (NEPV)	Bats (*Artibeus* spp.), rodents	Mosquitoes (*Culex* spp.)	F/M
	Oriboca virus (ORIV)	Capuchins (*Cebus* spp.), opossums, rodents	Mosquitoes (*Aedes*, *Culex*, *Mansonia*, *Psorophora* spp.)	F/M
	Ossa virus (OSSAV)	Rodents	Mosquitoes (*Culex* spp.)	F/M

Virus Group	Virus (Abbreviation)	Principal Reservoir Host(s)	Vector(s)	Syndrome[a]
	Restan virus (RESV)	Unknown	Mosquitoes (*Culex* spp.)	F/M
	Zungarococha virus (ZUNV)	Unknown	Unknown	F/M
Orthobunyaviruses (Guamá serogroup)	Catú virus (CATUV)	Bats, capuchins (*Cebus* spp.), opossums, rodents	Mosquitoes (*Culex* spp.)	F/M
	Guamá virus (GMAV)	Bats, capuchins (*Cebus* spp.), howlers (*Alouatta* spp.), marsupials, rodents	Mosquitoes (*Aedes*, *Culex*, *Limatus*, *Mansonia*, *Psorophora*, *Trichoprosopon* spp.)	F/M
Orthobunyaviruses (Mapputta serogroup)	Gan Gan virus (GGV)	Unknown	Mosquitoes (*Aedes*, *Culex* spp.)	A/R
	Trubanaman virus (TRUV)	Unknown	Mosquitoes (*Anopheles*, *Culex* spp.)	(A/R)
Orthobunyaviruses (Nyando serogroup)	Nyando virus (NDV)	Unknown	Mosquitoes (*Aedes*, *Anopheles* spp.), sandflies (*Lutzomyia* spp.)	F/M
	Iquitos virus (IQTV)	Unknown	Unknown	F/M
Orthobunyaviruses (Simbu serogroup)	Oropouche virus (OROV)	Marmosets (*Callithrix* spp.), pale-throated sloths (*B. tridactylus*)	Biting midges (*Culicoides paraensis*), mosquitoes (*Coquillettidia venezuelensis*, *Culex quinquefasciatus*, *Mansonia* spp., *Ochlerotatus serratus*)	F/M
Orthobunyaviruses (Wyeomyia serogroup)	Wyeomyia virus (WYOV)	Unknown	Mosquitoes (*Wyeomyia* spp.)	F/M
Orthobunyaviruses (serogroup undetermined)	Tataguine virus (TATV)	Unknown	Mosquitoes (*Anopheles* spp.)	F/M
Orthohantaviruses (Old World)	Amur/Soochong virus (ASV)	Korean field mice (*Apodemus peninsulae*)	None	VHF
	Dobrava-Belgrade virus (DOBV)	Caucasus field mice (*Apodemus ponticus*), striped field mice (*Apodemus agrarius*), yellow-necked field mice (*Apodemus flavicollis*)	None	VHF
	Go¯u virus (GOUV)	Brown rats (*Rattus norvegicus*), roof rats (*R. rattus*), Oriental house rats (*Rattus tanezumi*)	None	VHF
	Hantaan virus (HTNV)	Striped field mice (*A. agrarius*)	None	VHF
	Kurkino virus (KURV)	Striped field mice (*A. agrarius*)	None	VHF
	Muju virus (MUJV)	Korean red-backed voles (*Myodes regulus*)	None	VHF
	Puumala virus (PUUV)	Bank voles (*Myodes glareolus*)	None	(P), VHF
	Saaremaa virus (SAAV)	Striped field mice (*A. agrarius*)	None	VHF
	Seoul virus (SEOV)	Brown rats (*R. norvegicus*), roof rats (*R. rattus*)	None	VHF
	Sochi virus	Caucasus field mice (*A. ponticus*)	None	VHF
	Tula virus (TULV)	Common voles (*Microtus arvalis*), East European voles (*Microtus levis*), field voles (*Microtus agrestis*)	None	(P), VHF
Orthohantaviruses (New World)	Anajatuba virus (ANJV)	Fornes' colilargos (*Oligoryzomys fornesi*)	None	P
	Andes virus (ANDV)	Long-tailed colilargos (*Oligoryzomys longicaudatus*)	None	P
	Araraquara virus (ARAV)	Hairy-tailed akodonts (*Necromys lasiurus*)	None	P
	Araucária virus (ARAUV)	Black-footed colilargos (*Oligoryzomys nigripes*)	None	P
	Bayou virus (BAYV)	Marsh rice rats (*Oryzomys palustris*)	None	P
	Bermejo virus (BMJV)	Chacoan colilargos (*Oligoryzomys chacoensis*)	None	P
	Black Creek Canal virus (BCCV)	Hispid cotton rats (*S. hispidus*)	None	P

Virus Group	Virus (Abbreviation)	Principal Reservoir Host(s)	Vector(s)	Syndrome[a]
	Blue River virus (BRV)	White-footed deer mice (*P. leucopus*)	None	P
	Castelo dos Sonhos virus (CASV)	Brazilian colilargos (*Oligoryzomys eliurus*)	None	P
	Choclo virus (CHOV)	Fulvous colilargos (*Oligoryzomys fulvescens*)	None	F/M
	El Moro Canyon virus (ELMCV)	Sumichrast's harvest mice (*Reithrodontomys sumichrasti*), western harvest mice (*Reithrodontomys megalotis*)	None	P
	Juquitiba virus (JUQV)	Black-footed colilargos (*O. nigripes*)	None	P
	Laguna Negra virus (LANV)	Little lauchas (*Calomys laucha*)	None	P
	Lechiguanas virus (LECV)	Flavescent colilargos (*Oligoryzomys flavescens*)	None	P
	Maciel virus (MCLV)	Dark-furred akodonts (*Necromys obscurus*)	None	P
	Maripa virus (MARV)	Unknown	None	P
	Monongahela virus (MGLV)	North American deer mice (*P. maniculatus*)	None	P
	Muleshoe virus (MULV)	Hispid cotton rats (*S. hispidus*)	None	P
	New York virus (NYV)	White-footed deer mice (*P. leucopus*)	None	P
	Orán virus (ORNV)	Long-tailed colilargos (*O. longicaudatus*)	None	P
	Paranoá virus	Unknown	None	P
	Pergamino virus (PRGV)	Azara's akodonts (*Akodon azarae*)	None	P
	Río Mamoré virus (RIOMV)	Common bristly mice (*Neacomys spinosus*)	None	P
	Sin nombre virus (SNV)	North American deer mice (*P. maniculatus*)	None	P
	Tunari virus (TUNV)	Unknown	None	P
Orthonairoviruses (Crimean-Congo hemorrhagic fever virus group)	Crimean-Congo hemorrhagic fever virus (CCHFV)	Cattle, dogs, goats, hares, hedgehogs, mice, ostriches, sheep	Predominantly ixodid ticks (*Hyalomma* spp.)	VHF
Orthonairoviruses (Dugbe virus group)	Dugbe virus (DUGV)	Northern giant pouched rats (*Cricetomys gambianus*), Zébu cattle (*Bos primigenius*)	Biting midges (*Culicoides* spp.), ixodid ticks (*Amblyomma, Hyalomma, Rhipicephalus* spp.)	F/M
	Nairobi sheep disease virus[j] (NSDV)	Sheep	Ixodid ticks (*Haemaphysalis*, *Rhipicephalus* spp.), mosquitoes (*Culex* spp.)	F/M
Orthonairoviruses (Sakhalin virus group)	Avalon virus (AVAV)	European herring gulls (*Larus argentatus*)	Ixodid ticks (*Ixodes uriae*)	(Polyradiculoneuritis?)
Orthonairoviruses (Thiafora virus group)	Erve virus (ERVEV)	Greater white-toothed shrews (*Crocidura russula*)	?	(Thunderclap headache?)
Phleboviruses (Candiru serocomplex)	Alenquer virus (ALEV)	Unknown	Unknown	F/M
	Candiru virus (CDUV)	Unknown	Unknown	F/M
	Escharate virus (ESCV)	Unknown	Unknown	F/M
	Maldonado virus (MLOV)	Unkown	Unknown	F/M
	Morumbi virus (MRBV)	Unknown	Unknown	F/M
	Serra Norte virus (SRNV)	Unknown	Unknown	F/M
Phleboviruses (Punta Toro serocomplex)	Coclé virus (CCLV)	Unknown	Sandlfies	F/M

Virus Group	Virus (Abbreviation)	Principal Reservoir Host(s)	Vector(s)	Syndrome[a]
	Punta Toro virus (PTV)	Unknown	Sandflies (*Lutzomyia* spp.)	F/M
Phleboviruses (sandfly fever serocomplex)	Chagres virus (CHGV)	Unknown	Sandflies (*Lutzomyia* spp.)	F/M
	Chios virus	Unknown	Unknown	E
	Granada virus (GRV)	Unknown	Sandflies	F/M
	Rift Valley fever virus (RVFV)	Cattle, sheep	Mosquitoes (*Aedes*, *Anopheles*, *Coquillettidia*, *Culex*, *Eretmapodites*, *Mansonia* spp.)	E, F/M, VHF
	Sandfly fever Cyprus virus (SFCV)	Unknown	Unknown	F/M
	Sandfly fever Ethiopia virus (SFEV)	Unknown	Sandflies	F/M
	Sandfly fever Naples virus (SFNV)	Unknown	Sandflies (*Phlebotomus papatasi*, *P. perfiliewi*, *P. perniciosus*)	F/M
	Sandfly fever Sicilian virus (SFSV)	Eulipotyphla, least weasles (*Mustela nivalis*), rodents	Sandflies (particularly *Phlebotomus papatasi*)	F/M
	Sandfly fever Turkey virus (SFTV)	Unknown	Sandflies (*Phlebotomus* spp.)	F/M
	Toscana virus (TOSV)	Unknown	Sandflies (*Phlebotomus papatasi*, *P. perfiliewi*)	E, F/M
Phleboviruses (Salehabad serocomplex)	Adria virus (ADRV)	Unknown	Sandlfies	E
Phleboviruses (Uukuniemi serocomplex)	Uukuniemi virus (UUKV)	Birds, cattle, rodents	Ixodid ticks (*Ixodes* spp.)	F/M
Quaranjaviruses	Quaranfil virus (QRFV)	Birds	Ixodid ticks (*Argas arboreus*)	F/M
Seadornaviruses	Banna virus (BAV)	Cattle, pigs	Mosquitoes (*Aedes*, *Anopheles*, *Culiseta* spp.)	E
Thogotoviruses	Bourbon virus (BRBV)	Unknown	Ticks?	F/M
	Dhori virus (DHOV)[k]	Bats, camels, horses	Mosquitoes (*Aedes*, *Anopheles*, *Culex* spp.), ixodid ticks (*Dermacentor*, *Hyalomma*, *Ornithodoros* spp.)	E, F/M
	Thogoto virus (THOV)	Camels, cattle	Ixodid ticks (*Amblyomma*, *Hyalomma*, *Rhipicephalus* spp.)	E, F/M
Vesiculoviruses	Chandipura virus (CHPV)	Hedgehogs	Mosquitoes (*Aedes aegypti*), sandflies (*Phlebotomus*, *Sergentomyia* spp.)	E, F/M
	Isfahan virus (ISFV)	Great gerbils (*Rhombomys opimus*)	Sandflies (*Phlebotomus papatasi*)	F/M
	Piry virus (PIRYV)	Gray four-eyed opossums (*Philander opossum*)	Mosquitoes (*Aedes*, *Culex*, *Toxorhynchites* spp.)	F/M
	Vesicular stomatitis Indiana virus (VSIV)	Cattle, horses, pigs	Sandflies (*Lutzomyia* spp.)	F/M
	Vesicular stomatitis New Jersey virus (VSNJV)	Cattle, horses, pigs	Biting midges (*Culicoides* spp.), chloropid flies, mosquitoes (*Culex*, *Mansonia* spp.), muscoid flies (*Musca* spp.), simuliid flies	F/M

[a]Abbreviations refer to the syndrome most commonly associated with the virus: A/R, arthritis/rash; E, encephalitis; F/M, fever/myalgia; P, pulmonary; VHF, viral hemorrhagic fever. Abbreviations are placed in parentheses when cases are either extremely rare or controversial.

[b]In the older literature, chikungunya virus often is also listed as a causative agent of VHF. However, later studies revealed that, in most cases, people with “chikungunya hemorrhagic fever” were co-infected with one or more dengue viruses, an observation suggesting that the VHF was actually severe

dengue.

[c]Also known as Palma virus.

[d]Alternatives used in the literature are Huaiyangshan virus (HYSV) and Henan fever virus (HNFV).

[e]Also includes Kunjin virus.

[f]Includes the recently described Alkhurma/Alkhumra and Nanjianyin variants of Kyasanur Forest disease virus.

[g]Whitewater Arroyo virus is often listed as a causative agent of VHF in the literature, but convincing data associating this virus with VHF have not been published.

[h]Also known as Brezová virus, Cvilín virus, Kharagysh virus, Koliba virus, or Lipovník virus.

[i]Also known as čalovo virus or Chittoor virus.

[j]Also known as Ganjam virus.

[k]Also known as Astra virus and Batken virus.

"ARTICULAVIRALES": ORTHOMYXOVIRIDAE

The family *Orthomyxoviridae* includes two genera of medically relevant arthropod-borne viruses: *Quaranjavirus* and *Thogotovirus*. Quaranjaviruses are transmitted among birds by ixodid ticks, whereas hogotoviruses have a predilection for mammalian host reservoirs and can be transmitted by both ixodid ticks and mosquitoes.

BUNYAVIRALES: ARENAVIRIDAE

The members of the family *Arenaviridae* that infect humans are all assigned to the genus *Mammarenavirus*. The members of this genus are divided into two main phylogenetic branches: Old World viruses (the Lassa–lymphocytic choriomeningitis serocomplex) and New World viruses (the Tacaribe serocomplex). Mammarenaviruses form spherical, oval, or pleomorphic enveloped and spiked virions (~50–300 nm in diameter) that bud from the infected cell's plasma membrane. The particles contain two genomic single-stranded RNAs (S, ~3.5 kb; and L, ~7.5 kb) encoding structural proteins in an ambisense orientation. Most mammarenaviruses persist in nature by chronically infecting rodents. The human Old World mammarenaviruses are maintained by murid rodents that often are persistently viremic and commonly transmit viruses vertically and horizontally. One Old World mammarenavirus that has been associated with human infections is maintained by shrews. Human New World mammarenaviruses are found in cricetid rodents; horizontal transmission is typical, vertical infection may occur, and persistent viremia may be observed. Strikingly, each mammarenavirus is predominantly adapted to one particular type of rodent. Humans usually become infected through inhalation of or direct contact with infected rodent excreta or secreta (e.g., aerosols of rodents in harvesting machines; aerosolized dried rodent urine or feces in barns or houses; direct contact with rodents in traps). Person-to-person transmission of mammarenaviruses is uncommon.

BUNYAVIRALES: HANTAVIRIDAE, NAIROVIRIDAE, PERIBUNYAVIRIDAE,* AND *PHENUIVIRIDAE

The members of all these families that infect humans form spherical- to-pleomorphic enveloped virions containing three genomic single-stranded RNAs (S, ~1–2 kb; M, 3.6–5.3 kb; and L, 6.4–12.3 kb) of negative (hantaviruses, nairoviruses, peribunyaviruses) or ambisense (phenuiviruses) polarity. These bunyaviruses mature into particles ~80–120 nm in diameter in the Golgi complex of infected cells and exit these cells by exocytosis.

Hantaviruses that infect humans are classified in the genus *Orthohantavirus* and are maintained in nature by rodents that chronically shed virions. Old World orthohantaviruses are harbored by murid and cricetid rodents, and New World orthohantaviruses are maintained by cricetid rodents. As with mammarenaviruses, individual orthohantaviruses usually are specifically adapted to a particular type of rodent. However, orthohantaviruses do not cause chronic viremia in their rodent hosts and are transmitted only horizontally from rodent to rodent. Similar to mammarenaviruses, hantaviruses infect humans primarily through inhalation of or direct contact with rodent excreta or secreta, and person-to-person transmission is not a common event (with the notable exception of Andes virus). Although there is overlap, the human Old World orthohantaviruses usually are the etiologic agents of hemorrhagic fever with renal syndrome (HFRS), whereas the New World orthohantaviruses usually cause hantavirus (cardio)pulmonary syndrome.

Nairoviruses that infect humans are classified in the genus *Orthonairovirus*. These orthonairoviruses are maintained by ixodid ticks, which vertically (transovarially and transstadially) transmit these viruses to progeny tick generations and horizontally spread them through viremic vertebrate hosts. Humans are usually infected via a tick bite or during handling of infected vertebrates.

Peribunyaviruses of one genus (*Orthobunyavirus*) infect humans. Orthobunyaviruses are largely mosquito-borne and rarely midge-borne and have viremic vertebrate intermediate hosts. Many orthobunyaviruses are also transovarially transmitted in their mosquito hosts. Numerous orthobunyaviruses have been associated with human infection and disease. They have been considered to be members of ~19 serogroups based on antigenic cross-reactions, but this grouping is currently undergoing revision with the accumulation of new genomic data and phylogenetic analyses. Humans are infected by viruses in at least nine serogroups.

Phenuiviruses are transmitted vertically (transovarially) in their arthropod hosts and horizontally through viremic vertebrate hosts. Human phenuiviruses are found in two genera: "*Banyangvirus*" and *Phlebovirus*. "Banyangviruses" and viruses of the phlebovirus Uukuniemi group are transmitted by ticks, whereas those of the phlebovirus sandfly fever group are transmitted by sandflies. Phleboviruses are assigned to at least 10 serocomplexes; human pathogens are found in at least four of these serocomplexes.

MONONEGAVIRALES: RHABDOVIRIDAE

Rhabdoviruses have linear, typically nonsegmented, single-stranded RNA genomes of negative polarity (~11–15 kb) and form bullet-shaped to pleomorphic enveloped particles (100–430 nm long and 45–100 nm wide). Only the genus *Vesiculovirus* includes confirmed human arthropod-borne viruses, all of which are transmitted by insects (biting midges, mosquitoes, and sandflies). **The general properties of rhabdoviruses are discussed in more detail in Chap. 203.**

FLAVIVIRIDAE

The family *Flaviviridae* currently includes only one genus (*Flavivirus*) that comprises arthropod-borne human viruses. Flaviviruses sensu stricto have single-stranded positive-sense RNA genomes (~11 kb) and form spherical enveloped particles 40–60 nm in diameter. The flaviviruses discussed here belong to two phylogenetically and antigenically distinct groups that are transmitted among vertebrates by mosquitoes and ixodid ticks, respectively. Vectors are usually infected when they feed on viremic hosts; as in the case of most other viruses discussed here, humans are accidental hosts who usually are infected by arthropod bites. Arthropods maintain flavivirus infections horizontally, although transovarial transmission has been documented. Under certain circumstances, flaviviruses can also be transmitted by aerosols or via contaminated food products; in particular, raw milk can transmit tick-borne encephalitis virus.

REOVIRIDAE

The family *Reoviridae* contains viruses with linear, multisegmented, double-stranded RNA genomes (~16–29 kb in total). These viruses produce particles that have icosahedral symmetry and are 60–80 nm in diameter. In contrast to all other virions discussed here, reovirions are not enveloped and thus are insensitive to detergent inactivation. Human arthropod-borne viruses are found among the genera *Coltivirus* (subfamily *Spinareovirinae*), *Orbivirus*, and *Seadornavirus* (subfamily *Sedoreovirinae*). Arthropod-borne coltiviruses possess 12 genome segments. Coltiviruses are transmitted by numerous tick types transstadially but not transovarially. Overall maintenance of the transmission cycle, therefore, involves viremic mammalian hosts infected by tick bites. Arthropod-borne orbiviruses have 10 genome segments and are transmitted by mosquitoes or ixodid ticks, whereas relevant seadornaviruses have 12 genome segments and are transmitted exclusively by mosquitoes.

TOGAVIRIDAE

The members of the family *Togaviridae* have linear, single- and positive-stranded RNA genomes (~9.7–11.8 kb) and form enveloped icosahedral virions (~60–70 nm in diameter) that bud from the plasma membrane of the infected cell. The togaviruses discussed here are all members of the genus *Alphavirus* and are transmitted among vertebrates by mosquitoes.

EPIDEMIOLOGY

The distributions of arthropod-borne and rodent-borne viruses are restricted by the areas inhabited by their reservoir hosts and/or vectors. Consequently, a patient's geographic origin or travel history can provide important clues in the differential diagnosis. **Table 204-2** lists the approximate geographic distribution of most arthropod-borne and rodent-borne infections. Many of these diseases can be acquired in either rural or urban settings; these diseases include yellow fever, dengue (previously called dengue fever), severe dengue (previously called dengue hemorrhagic fever and dengue shock syndrome), chikungunya virus disease, HFRS caused by Seoul virus, sandfly fever caused by sandfly fever Naples and Sicilian viruses, and Oropouche virus disease.

TABLE 204-2 Geographic Distribution (United Nations Geoscheme) of Zoonotic Arthropod-Borne or Rodent-Borne Viral Diseases

Area	Type of Disease[a]						
	Arenaviral	Bunyaviral	Flaviviral	Orthomyxoviral	Reoviral	Rhabdoviral	Togaviral
Africa	Lassa fever; Lujo virus infection	Bangui, Batai, Bhanja, Bunyamwera, and Bwamba virus infections; Crimean-Congo hemorrhagic fever; Dugbe, Germiston, Ilesha virus infections; Nairobi sheep disease virus infection; Ngari, Nyando, and Pongola virus infections; Rift Valley fever, sandfly fever/Pappataci fever/phlebotomus fever; Shokwe, Tataguine virus infections	Dengue/severe dengue; (Usutu virus infection); West Nile virus infection; yellow fever; Zika virus disease	Dhori, Quaranfil, and Thogoto virus infections	Lebombo, Orungo, and Tribeč virus infections	—	Chikungunya virus disease; o'nyong-nyong fever; Semliki Forest and Sindbis virus infections
Central Asia	—	Bhanja virus infection; Crimean-Congo hemorrhagic fever	Tick-borne viral encephalitis	Dhori virus infections	—	Isfahan virus infections	—
Eastern Asia	—	Crimean-Congo hemorrhagic fever; hemorrhagic fever with renal syndrome; sandfly fever/Pappataci fever/phlebotomus fever; severe fever with thrombocytopenia syndrome	Dengue/severe dengue; Japanese encephalitis; Kyasanur Forest disease; tick-borne viral encephalitis	—	Banna virus infections	—	—
Southern Asia	—	Batai and Bhanja virus infections; Crimean-Congo hemorrhagic fever; hemorrhagic fever with renal syndrome; Nairobi sheep disease virus infection; sandfly fever/Pappataci fever/phlebotomus fever	Dengue/severe dengue; Japanese encephalitis; Kyasanur Forest disease; West Nile virus infection; Zika virus disease	Dhori, Quaranfil, and Thogoto virus infections	—	Chandipura and Isfahan virus infections	Chikungunya virus disease
South-Eastern Asia	—	Batai virus infection; hemorrhagic fever with renal syndrome	Dengue/severe dengue; Japanese encephalitis; West Nile virus infection; Zika virus disease	—	—	—	Chikungunya virus disease
Western Asia	—	Batai and Bhanja virus infections; Crimean-Congo hemorrhagic fever; hemorrhagic fever with renal syndrome; sandfly fever/Pappataci fever/phlebotomus fever	Dengue/severe dengue; Kyasanur Forest disease; tick-borne viral encephalitis; West Nile virus infection	Dhori and Quaranfil virus infections	—	—	Chikungunya virus disease

Area	Type of Disease[a]						
	Arenaviral	Bunyaviral	Flaviviral	Orthomyxoviral	Reoviral	Rhabdoviral	Togaviral
Latin/Central America and the Caribbean	“Brazilian hemorrhagic fever”; Chapare virus infection; Junín/Argentinian hemorrhagic fever; lymphocytic choriomeningitis/meningoencephalitis; Machupo/Bolivian hemorrhagic fever; “Venezuelan hemorrhagic fever”	Alenquer, Apeú, Bunyamwera, Cache Valley, Candiru, Caraparú, Catú, Chagres, Coclé, Escharate, Fort Sherman, Guamá, and Guaroa virus infections; hantavirus (cardio)pulmonary syndrome; Itaquí, Juquitiba, Madrid, Maguari, Maldonado, Marituba, Mayaro, Morumbi, Murutucú, Nepuyo, and Oriboca virus infections; Oropouche virus disease; Ossa, Punta Toro, Restan, Serra Norte, Tacaiuma, Trinidad, Wyeomyia, Xingu, and Zungarococha virus infections	Dengue/severe dengue; Rocio virus disease; yellow fever; Zika virus disease	—	—	Piry virus disease; vesicular stomatitis virus disease/Indiana fever	Chikungunya virus disease; Mayaro virus infection; Mucambo, Tonate, and Una virus infections; Venezuelan equine fever
Northern America	Lymphocytic choriomeningitis/meningoencephalitis; (Whitewater Arroyo virus infection)	(Avalon and) Cache Valley virus infections; California (meningo)encephalitis; hantavirus (cardio)pulmonary syndrome; Heartland virus and Nepuyo virus infections	Dengue/severe dengue; Powassan virus disease; St. Louis encephalitis; West Nile virus infection; Zika virus disease	Bourbon virus infection	Colorado tick fever; Salmon River virus infection	Vesicular stomatitis virus disease/Indiana fever	Eastern equine encephalitis; Everglades virus infection; western equine encephalitis
Europe	Lymphocytic choriomeningitis/meningoencephalitis	(Adria, Avalon, and) Bhanja virus infections; California (meningo)encephalitis; Crimean-Congo hemorrhagic fever; (Erve virus infection); hemorrhagic fever with renal syndrome; Inkoo virus infection; sandfly fever/Pappataci fever/phlebotomus fever; Uukuniemi virus infection	Dengue/severe dengue; tick-borne viral encephalitis; Omsk hemorrhagic fever; (Usutu virus infection); West Nile virus infection	Dhori and Thogoto virus infections	Eyach, Kemerovo, and Tribeč virus infections	—	Chikungunya virus disease; Sindbis virus infection
Oceania	—	Gan Gan (and Trubanaman virus) infections	Australian encephalitis; dengue/severe dengue; Japanese encephalitis; Kokobera virus infection; Murray Valley encephalitis; West Nile virus infection; Zika virus disease	—	—	—	Barmah Forest virus infection; Ross River disease; Sindbis virus infection

[a]Quotation marks indicate common usage in the absence of International Classification of Disease version 10 (ICD-10) recognition. Diseases not acknowledged by the ICD-10 are designated as “virus infection(s).” Disease names are placed in parentheses when cases are either extremely rare or controversial.

DIAGNOSIS

In patients with suspected viral infection, a recognized history of mosquito bite(s) has little diagnostic significance, but a history of tick bite(s) is more diagnostically useful. Exposure to rodents is sometimes reported by persons infected with mammarenaviruses or orthohantaviruses. Laboratory diagnosis is required in all cases, although epidemics occasionally provide enough clinical and epidemiologic clues for a presumptive etiologic diagnosis. For most arthropod-borne and rodent-borne viruses, acute-phase serum samples (collected within 3 or 4 days of onset) have yielded isolates. Paired serum samples have been used to demonstrate rising antibody titers. Intensive efforts to develop rapid tests for viral hemorrhagic

fevers (VHFs) have resulted in reliable antigen-detection enzyme-linked immunosorbent assays (ELISAs), IgM-capture ELISAs, and multiplex polymerase chain reaction (PCR) assays. These tests can provide a diagnosis based on a single serum sample within a few hours and are particularly useful in patients with severe disease. More sensitive reverse-transcription PCR (RT-PCR) assays may yield diagnoses based on samples without detectable antigen and may also provide useful genetic information about the etiologic agent.

Orthohantavirus infections differ from other viral infections discussed here in that severe acute disease is immunopathologic; patients present with serum IgM that serves as the basis for a sensitive and specific test. At diagnosis, patients with encephalitides generally are no longer viremic or antigenemic and usually do not have virions in cerebrospinal fluid (CSF). In this situation, the value of serologic methods for IgM determination and RT-PCR is high. IgM-capture ELISA is increasingly used for the simultaneous testing of serum and CSF. IgG ELISA or classic serology is useful in the evaluation of past exposure to viruses, many of which circulate in areas with minimal medical infrastructures and sometimes cause only mild or subclinical infections.

SYNDROMES

The spectrum of possible human responses to infection with arthropod- or rodent-borne viruses is wide, and knowledge of the outcome of most of these infections is limited. People infected with these viruses may not develop signs of illness. If viral disease is recognized, it can usually be grouped into one of five broad categories: arthritis and rash, encephalitis, fever and myalgia, pulmonary disease, or VHF **(Table 204-3)**. These categories often overlap. For example, infections with West Nile and Venezuelan equine encephalitis viruses are discussed here as encephalitides, but during epidemics many patients present with much milder febrile syndromes. Similarly, Rift Valley fever virus is best known as a cause of VHF, but the attack rates for febrile disease are far higher, and encephalitis and blindness occasionally occur as well. Lymphocytic choriomeningitis virus is classified here as a cause of fever and myalgia because this syndrome is the most common disease manifestation. Even when central nervous system (CNS) disease evolves during infection with this virus, neural manifestations are usually mild and are preceded by fever and myalgia. However, this virus may also cause fetal microcephaly. Infection with any dengue virus (1, 2, 3, or 4) is considered as a cause of fever and myalgia because this syndrome is by far the most common manifestation worldwide. However, severe dengue is a VHF with a complicated pathogenesis that is of tremendous importance in pediatric practice in certain areas of the world. Unfortunately, most of the known arthropod- or rodent-borne viral diseases have not been studied in detail with modern medical approaches; thus available data may be incomplete or biased. The reader must be aware that data on geographic distribution are often fuzzy: the literature frequently is not clear as to whether the data pertain to the distribution of a particular virus or to the areas where human disease has been observed. In addition, the designations for viruses and viral diseases have changed multiple times over decades. Here, virus and taxon names are in line with the latest reports of the International Committee on Taxonomy of Viruses, and disease names are in accordance with the World Health Organization’s International Classification of Disease version 10 (ICD-10) and more recent updates.

TABLE 204-3 Clinical Syndromes Caused by Zoonotic Arthropod-Borne or Rodent-Borne Viruses

Syndrome	Virus[a]
Arthritis and rash (A/R)	*Flaviviridae:* Kokobera and Zika viruses
	Peribunyaviridae: Gan Gan (and Trubanaman) viruses
	Togaviridae: Barmah Forest, chikungunya, Mayaro, o'nyong-nyong, Ross River, Semliki Forest, and Sindbis viruses
	Arenaviridae: lymphocytic choriomeningitis (and Whitewater Arroyo) viruses
Encephalitis (E)	*Flaviviridae:* Japanese encephalitis, Murray Valley encephalitis, Powassan, Rocio, St. Louis encephalitis, tick-borne encephalitis, (Usutu), and West Nile viruses
	Orthomyxoviridae: Dhori and Thogoto viruses
	Peribunyaviridae: California encephalitis, Inkoo, Jamestown Canyon, La Crosse, Lumbo, snowshoe hare, and Ťahyňa viruses
	Phenuiviridae: Adria, Bhanja, Chios, Rift Valley fever, and Toscana viruses
	Reoviridae: Banna, Colorado tick fever, Eyach, Kemerovo, Orungo, and Salmon River viruses
	Rhabdoviridae: Chandipura virus
	Togaviridae: eastern equine encephalitis, Everglades, Mucambo, Tonate, Venezuelan equine encephalitis, and western equine encephalitis viruses
	Arenaviridae: Lassa and lymphocytic choriomeningitis viruses
	Bunyavirales (unclassified): Bangui virus
	Flaviviridae: dengue 1–4, tick-borne encephalitis, and Zika viruses
	Hantaviridae*:* Choclo virus
	Nairoviridae: Dugbe and Nairobi sheep disease viruses
	Orthomyxoviridae: Bourbon, Dhori, and Thogoto viruses
Fever and myalgia (F/M)	*Peribunyaviridae:* Apeú, Batai, Bunyamwera, Bwamba, Cache Valley, California encephalitis, Caraparú, Catú, Fort Sherman, Germiston, Guamá, Guaroa, Ilesha, Inkoo, Iquitos, Itaquí, Jamestown Canyon, La Crosse, Lumbo, Madrid, Maguari, Marituba, Nepuyo, Ngari, Nyando, Oriboca, Oropouche, Ossa, Pongola, Restan, Shokwe, snowshoe hare, Tacaiuma, Ťahyňa, Tataguine, Wyeomyia, Xingu, and Zungarococha viruses
	Phenuiviridae*:* Alenquer, Bhanja, Candiru, Chagres, Escharate, Heartland, Maldonado, Morumbi, Punta Toro, Rift Valley fever, sandfly fever Cyprus, sandfly fever Ethiopia, sandfly fever Naples, sandfly fever Sicilian, sandfly fever Turkey, Serra Norte, severe fever with thrombocytopenia syndrome, Toscana, and Uukuniemi viruses
	Reoviridae: Colorado tick fever, Eyach, Kemerovo, Lebombo, Orungo, Salmon River, and Tribeč viruses
	Rhabdoviridae: Chandipura, Isfahan, Piry, vesicular stomatitis Indiana, and vesicular stomatitis New Jersey viruses
	Togaviridae: Everglades, Mucambo, Tonate, Una, and Venezuelan equine encephalitis viruses
Pulmonary disease (P)	*Hantaviridae:* Anajatuba, Andes, Araucária, bayou, Bermejo, Black Creek Canal, Blue River, Castelo dos Sonhos, El Moro Canyon, Juquitiba, Laguna Negra, Lechiguanas, Maciel, Monongahela, Muleshoe, New York, Orán, Paranoá, Pergamino, (Puumala), Río Mamoré, sin nombre, (Tula), and Tunari viruses
	Arenaviridae: Chapare, Guanarito, Junín, Lassa, Lujo, (lymphocytic choriomeningitis), Machupo, and Sabiá viruses
	Hantaviridae: Amur/Soochong, Dobrava-Belgrade, Go¯u, Hantaan, Kurkino, Muju, Puumala, Saaremaa, Seoul, Sochi, and Tula viruses
Viral hemorrhagic fever (VHF)	*Nairoviridae:* Crimean-Congo hemorrhagic fever virus
	Peribunyaviridae: (Ilesha and) Ngari viruses
	Phenuiviridae: Rift Valley fever and severe fever with thrombocytopenia syndrome viruses
	Flaviviridae: dengue 1–4, Kyasanur Forest disease, Omsk hemorrhagic fever, (tick-borne encephalitis), and yellow fever viruses

[a]Virus names are placed in parentheses when human infections are either extremely rare or controversial.

ARTHRITIS AND RASH

Arthritides are common accompaniments of several viral diseases, such as hepatitis B, parvovirus B19 infection, and rubella, and occasionally accompany infection due to adenoviruses, enteroviruses, herpesviruses, or mumps virus. Two orthobunyaviruses—Gan Gan virus and Trubanaman virus—and the flavivirus Kokobera virus have been associated with single cases of polyarthritic disease. Arthropod-borne alphaviruses are also common causes of arthritides—usually acute febrile diseases accompanied by the development of a maculopapular rash. Rheumatic involvement includes arthralgia alone, periarticular swelling, and (less commonly) joint effusions. Most alphavirus infections are less severe and have fewer articular manifestations in children than in adults. In temperate climates, these ailments are summer diseases. No specific therapies or licensed vaccines exist. The most important alphavirus arthritides are Barmah Forest virus infection, chikungunya virus disease, Ross River disease, and Sindbis virus infection. A large (>2 million cases), albeit isolated, epidemic was caused by o'nyong nyong virus in 1959–1961 (o'nyong nyong fever). Mayaro, Semliki Forest, and Una viruses caused isolated cases or limited and infrequent epidemics (30 to several hundred cases per year). Signs and symptoms of infections with these viruses often are similar to those observed with chikungunya virus disease.

Chikungunya Virus Disease

Disease caused by chikungunya virus is endemic in rural areas of Africa. Intermittent epidemics take place in towns and cities of both Africa and Asia. Yellow fever mosquitoes (*Aedes aegypti*) are the usual vectors for the disease in urban areas. In 2004, a massive epidemic began in the Indian Ocean region (in particular on the islands of Réunion and Mauritius) and was most likely spread by travelers. The Asian tiger mosquito (*Aedes albopictus*) was identified as the major vector of chikungunya virus during that epidemic. From 2013 and 2014, several thousand chikungunya virus infections were reported (and several tens to hundreds of thousands of cases were suspected) from Caribbean islands. The virus was imported to Italy, France, and the United States by travelers from the Caribbean. Chikungunya virus poses a threat to the continental United States as suitable vector mosquitoes are present in southern states. The disease is most common among adults, in whom the clinical presentation may be dramatic. The abrupt onset of chikungunya virus disease follows an incubation period of 2–10 days. Fever (often severe) with a saddleback pattern and severe arthralgia are accompanied by chills and constitutional symptoms and signs, such as abdominal pain, anorexia, conjunctival injection, headache, nausea, and photophobia. Migratory polyarthritis mainly affects the small joints of the ankles, feet, hands, and wrists, but the larger joints are not necessarily spared. Rash may appear at the outset or several days into the illness; its development often coincides with defervescence, which occurs around day 2 or 3 of the disease. The rash is most intense on the trunk and limbs and may desquamate. Young children develop less prominent signs and are therefore less frequently hospitalized. Children also often develop a bullous rather than a maculopapular/petechial rash. Maternal–fetal transmission has been reported and, in some cases, has led to fetal death. Recovery may require weeks, and some elderly patients may continue to experience joint pain, recurrent effusions, or stiffness for several years. This persistence of signs and symptoms may be especially common in human leukocyte antigen B27 subtype (HLA-B27)–positive patients. In addition to arthritis, petechiae are occasionally seen, and epistaxis is not uncommon, but chikungunya virus should not be considered a VHF agent. A few patients develop leukopenia. Elevated concentrations of aspartate aminotransferase (AST) and C-reactive protein have been described, as have mildly decreased platelet counts. Treatment of chikungunya virus disease relies on nonsteroidal anti-inflammatory drugs and sometimes chloroquine for refractory arthritis.

Ross River Disease and Barmah Forest Virus Infection

Ross River virus and Barmah Forest virus cause diseases that are indistinguishable on clinical grounds alone (hence the previously common disease designation *epidemic polyarthritis* for both infections). Ross River virus has caused epidemics in Australia, Papua New Guinea, and the South Pacific since the beginning of the twentieth century. In 1979–1980, the virus swept through the Pacific Islands, causing more than 500,000 infections. In 1991–2011, Ross River virus caused a total of 92,559 infections or disease in rural and suburban areas. Ross River virus is predominantly transmitted by *Aedes normanensis*, *Aedes vigilax*, and *Culex annulirostris* mosquitoes. Wallabies and rodents are probably the main vertebrate hosts. Barmah Forest virus infections have been on the rise in recent years. For instance, in 1991–2011, 21,815 cases were recorded in Australia. Barmah Forest virus is transmitted by both *Aedes* and *Culex* mosquitoes and has been isolated from biting midges. The vertebrate hosts remain to be determined, but serologic studies implicate horses and possums.

Of the human Barmah Forest and Ross River virus infections surveyed, 55–75% were asymptomatic; however, these viral diseases can be debilitating. The incubation period is 7–9 days; the onset of illness is sudden, and disease is usually ushered in by disabling symmetrical joint pain. A non-itchy, diffuse, maculopapular rash (more common in Barmah Forest virus infection) generally develops coincidentally or follows shortly, but in some patients rash can precede joint pain by several days. Constitutional symptoms such as low-grade fever, asthenia, headache, myalgia, and nausea are not prominent or are absent in many patients. Most patients are incapacitated for considerable periods (≥6 months) by joint involvement, which interferes with grasping, sleeping, and walking. Ankle, interphalangeal, knee, metacarpophalangeal, and wrist joints are most often involved, although elbows, shoulders, and toes may also be affected. Periarticular swelling and tenosynovitis are common, and one-third of patients have true arthritis (more common in Ross River disease). Myalgia and nuchal stiffness may accompany joint pains. Only half of all patients with arthritis can resume normal activities within 4 weeks, and 10% still must limit their activity after 3 months. Occasional patients are symptomatic for 1–3 years but without progressive arthropathy.

In the diagnosis of either infection, clinical laboratory values are normal or variable. Tests for rheumatoid factor and antinuclear antibodies are negative, and the erythrocyte sedimentation rate is acutely elevated. Joint fluid contains 1000–60,000 mononuclear cells/μL, and viral antigen can usually be detected in macrophages. IgM antibodies are valuable in the diagnosis of this infection, although such antibodies occasionally persist for years. Isolation of the virus from blood after mosquito inoculation or growth of the virus in cell culture is possible early in the illness. Because of the great economic impact of annual epidemics in Australia, an inactivated Ross River virus vaccine is under advanced development; phase 3 trials were completed in 2015. Nonsteroidal anti-inflammatory drugs, such as naproxen or acetylsalicylic acid, are effective for treatment.

Sindbis Virus Infection

Sindbis virus is transmitted among birds by infected mosquitoes. Infections with northern European or southern African variants are particularly likely in rural environments. After an incubation period of <1 week, Sindbis virus infection begins with rash and arthralgia. Constitutional clinical signs are not marked, and fever is modest or lacking altogether. The rash, which lasts ~1 week, begins on the trunk, spreads to the extremities, and evolves from macules to papules that often vesiculate. The arthritis is multiarticular, migratory, and incapacitating, with resolution of the acute phase in a few days. The ankles, elbows, knees, phalangeal joints, wrists, and—to a much lesser extent—proximal and axial joints are involved. Persistence of joint pain and occasionally of arthritis is a major problem and may continue for months or even years despite lack of deformities.

Zika Virus Disease

Zika virus is an emerging pathogen that is transmitted among nonhuman primates and humans by *Aedes* mosquitoes. The virus was discovered 1947 in a sentinel rhesus monkey (*Macaca mulatta*) and *Aedes africanus* mosquitoes in the Zika Forest in what was then the British Protectorate of Uganda. Human Zika virus infection was first documented during a yellow fever outbreak in 1954 in Nigeria. Later, Zika virus infections were recognized in south-Eastern and Southern Asia. Prior to 2007, only 14 clinically identified cases of Zika virus disease had been reported. In recent years, the number of Zika virus infections reported has increased steadily and rapidly, with large, but generally mild, disease outbreaks on Yap Island, Micronesia (2007), and in Cambodia (2010), the Philippines (2012), and French Polynesia (2013–2014). Invasion of the New World was first reported in 2014 on Easter Island in Chile and in 2015 in Brazil. At the end of May 2017, Zika virus infections had been recorded on the five continents in 85 countries, including Mexico and the United States. An estimated 440,000 to 1.3 million cases had occurred in Brazil by the end of 2015.

Phylogenetic analysis of all available African Zika virus isolates revealed two geographically overlapping clades (western and eastern Africa). A descendant Asian lineage, represented by viruses collected from mosquitoes trapped in homes in Malaysia, was first reported in 1969. All Zika virus isolates causing human cases outside of Africa trace back to this Asian lineage.

Human infections are usually asymptomatic or benign and self-resolving and are most likely misdiagnosed as dengue or influenza. Zika virus disease is typically characterized by low-grade fever, headache, and malaise. An itchy maculopapular rash, nonpurulent conjunctivitis, myalgia, and arthralgia usually accompany or follow those manifestations. Vomiting, hematospermia, and hearing impairments are relatively common clinical signs. In severe

cases, Zika virus infection is associated with serious complications such as Guillain-Barré syndrome or fetal microcephaly after congenital transmission. Other neurologic complications of Zika virus infection are encephalitis, meningoencephalitis, transverse myelitis, peripheral neuropathies, retinopathies, and neurologic birth defects. Although most human Zika virus infections are acquired after bites by infected female mosquitoes, transmission may also occur perinatally or via heterosexual or homosexual contact with an infected person, breastfeeding, or transfusion of blood products. Specifically, viral persistence in the testes, which can last up to at least 160 days, is worrisome, as sexual virus transmission be may be possible throughout that period. Unfortunately, antiviral treatments (curative or preventive) and licensed vaccines against Zika virus are not yet available.

ENCEPHALITIS

The major encephalitis viruses are found in the families *Flaviviridae, Peribunyaviridae, Rhabdoviridae,* and *Togaviridae*. However, individual agents of other families, including Dhori virus and Thogoto virus *(Orthomyxoviridae)* and Banna virus *(Reoviridae)*, have been known to cause isolated cases of encephalitis as well. Arboviral encephalitides are seasonal diseases, commonly occurring in the warmer months. Their incidence varies markedly with time and place, depending on ecologic factors. The causative viruses differ substantially in terms of case–infection ratio (i.e., the ratio of clinical to subclinical infections), lethality, and residual disease. Humans are not important amplifiers of these viruses.

All the viral encephalitides discussed in this section have a similar pathogenesis. An infected arthropod ingests blood from a human and thereby initiates infection. The initial viremia is thought to originate from the lymphoid system. Viremia leads to multifocal entry into the CNS, presumably through infection of olfactory neuroepithelium, with passage through the cribriform plate; “Trojan horse” entry with infected macrophages; or infection of brain capillaries. During the viremic phase, there may be little or no recognizable disease except in tick-borne flavivirus encephalitides, which may manifest with clearly delineated phases of fever and systemic illness.

CNS lesions arise partly from direct neuronal infection and subsequent damage and partly from edema, inflammation, and other indirect effects. The usual pathologic features of arboviral encephalitides are focal necroses of neurons, inflammatory glial nodules, and perivascular lymphoid cuffing. Involved areas display the “luxury perfusion” phenomenon, with normal or increased total blood flow and low oxygen extraction. The typical patient presents with a prodrome of nonspecific constitutional signs and symptoms, including fever, abdominal pain, sore throat, and respiratory signs. Headache, meningeal signs, photophobia, and vomiting follow quickly. The severity of human infection varies from an absence of signs/symptoms to febrile headache, aseptic meningitis, and full-blown encephalitis. The proportions and severity of these manifestations vary with the infecting virus. Involvement of deeper brain structures in less severe cases may be signaled by lethargy, somnolence, and intellectual deficit (as disclosed by the mental status examination). More severely affected patients are obviously disoriented and may become comatose. Tremors, loss of abdominal reflexes, cranial nerve palsies, hemiparesis, monoparesis, difficulty swallowing, limb-girdle syndrome, and frontal lobe signs are all common. Spinal and motor neuron diseases are documented after West Nile and Japanese encephalitis virus infections. Seizures and focal signs may be evident early or may appear during the course of the disease. Some patients present with an abrupt onset of fever, convulsions, and other signs of CNS involvement. The acute encephalitis usually lasts from a few days to as long as 2–3 weeks. The infections may be fatal, or recovery may be slow, with weeks or months required for the return of maximal recoupable function, or incomplete, with persisting long-term deficits. Difficulty concentrating, fatigability, tremors, and personality changes are common during recovery.

The diagnosis of arboviral encephalitides depends on the careful evaluation of a febrile patient with CNS disease and the performance of laboratory studies to determine etiology. Clinicians should (1) consider empirical acyclovir treatment for herpesvirus meningoencephalitis and antibiotic treatment for bacterial meningitis until test results are received; (2) exclude intoxination and metabolic or oncologic causes, including paraneoplastic syndromes, hyperammonemia, liver failure, and anti-*N*-methyl-D-aspartate (NMDA) receptor encephalitis; and (3) rule out a brain abscess or a stroke. Leptospirosis, neurosyphilis, Lyme disease, cat-scratch disease, and more recently described viral encephalitides (e.g., Nipah virus infection), among others, should be considered if epidemiologically relevant. CSF examination usually shows a modest increase in leukocyte counts—in the tens or hundreds or perhaps a few thousand. Early in the process, a significant proportion of these leukocytes may be polymorphonuclear, but mononuclear cells are usually predominant later. CSF glucose concentrations are generally normal. There are exceptions to this pattern of findings: in eastern equine encephalitis, for example, polymorphonuclear leukocytes may predominate during the first 72 h of disease, and hypoglycorrhachia may be detected. In lymphocytic choriomeningitis/meningoencephalitis, lymphocyte counts may be in the thousands, and glucose concentrations may be diminished. A humoral immune response is usually detectable at or near the onset of disease. Both serum (acute- or convalescent-phase) and CSF should be examined for IgM antibodies, and viruses should be detected by plaque-reduction neutralization assay and/or (RT)-PCR. Virus generally cannot be isolated from blood or CSF, although Japanese encephalitis virus has been recovered from CSF of patients with severe disease. RT-PCR analysis of CSF may yield positive results. Viral antigen is present in brain tissue, although its distribution may be focal. Electroencephalography usually shows diffuse abnormalities and is not directly helpful.

Experience with medical imaging is still evolving. Both CT and MRI scans may be normal except for evidence of preexisting conditions or occasional diffuse edema. Imaging is generally nonspecific, as most patients do not present with pathognomonic lesions, but it can be used to rule out other suspected causes of disease. It is important to remember that imaging may yield negative results if done early in the disease course but may later detect lesions. For example, eastern equine encephalitis (focal abnormalities) and severe Japanese encephalitis (hemorrhagic bilateral thalamic lesions) have caused lesions detectable by medical imaging.

Comatose patients may require management of intracranial pressure elevations, inappropriate secretion of antidiuretic hormone, respiratory failure, or seizures. Specific therapies for these viral encephalitides are not available. The only practical preventive measures are vector management and personal protection against the arthropod transmitting the virus. For Japanese encephalitis or tick-borne viral encephalitis, vaccination should be considered in certain circumstances (see relevant sections below).

Flaviviruses

The most important flavivirus encephalitides are Japanese encephalitis, St. Louis encephalitis, tick-borne encephalitis, and West Nile virus infection. Australian encephalitis (Murray Valley encephalitis) and Rocio virus infection resemble Japanese encephalitis but are documented only occasionally in Australia and Brazil, respectively. Powassan virus has caused ~77 cases of often-severe disease (lethality, ~10%), frequently occurring among children in eastern Canada and the United States. Usutu virus has caused only individual cases of human infection, but such infections may be underdiagnosed.

JAPANESE ENCEPHALITIS

Japanese encephalitis is the most important viral encephalitis in Asia. Each year ~68,000 cases and ~13,600–20,400 deaths are reported. Japanese encephalitis virus is found throughout Asia, including in far eastern Russia, Japan, China, India, Pakistan, and south-Eastern Asia, and causes occasional epidemics on western Pacific islands. The virus has been detected in the Torres Strait islands, and five human encephalitis cases have been identified on the nearby Australian mainland. The virus is particularly common in areas where irrigated rice fields attract the natural avian

vertebrate hosts and provide abundant breeding sites for *Culex tritaeniorhynchus* mosquitoes, which transmit the virus to humans. Additional amplification by pigs, which suffer abortion, and horses, which develop encephalitis, may be significant as well. Vaccination of these additional amplifying hosts may reduce the transmission of the virus.

Clinical signs of Japanese encephalitis emerge after an incubation period of 5–15 days and range from an unspecific febrile presentation (nausea, vomiting, diarrhea, cough) to aseptic meningitis, meningoencephalitis, acute flaccid paralysis, and severe encephalitis. Common findings are cerebellar signs, cranial nerve palsies, and cognitive and speech impairments. A Parkinsonian presentation and seizures are typical in severe cases. Effective vaccines are available. Vaccination is indicated for summer travelers to rural Asia, where the risk of acquiring Japanese encephalitis is considered to be about 1 per 5000 to 1 per 20,000 travelers per week if travel duration exceeds 3 weeks. Usually two intramuscular doses of the vaccine are given 28 days apart, with the second dose administered at least 1 week prior to travel.

ST. LOUIS ENCEPHALITIS

St. Louis encephalitis virus is transmitted between mosquitoes and birds. This virus causes a low-level endemic infection among rural residents of the western and central United States, where *Culex tarsalis* mosquitoes serve as vectors (see "Western Equine Encephalitis," below). The more urbanized mosquitoes (*Culex pipiens* and *Culex quinquefasciatus*) have been responsible for epidemics resulting in hundreds or even thousands of cases in cities of the central and eastern United States. Most cases occur in June through October. The urban mosquitoes breed in accumulations of stagnant water and sewage with high organic content and readily feed on humans in and around houses at dusk. The elimination of open sewers and trash-filled drainage systems is expensive and may not be possible. However, screening of houses and implementation of personal protective measures may be effective approaches to the prevention of infection. The rural mosquito vector is most active at dusk and outdoors; its bites can be avoided by modification of activities and use of repellents.

Disease severity increases with age. St. Louis encephalitis virus infections that result in aseptic meningitis or mild encephalitis are concentrated among children and young adults, whereas severe and fatal cases primarily affect the elderly. Infection rates are similar in all age groups; thus, the greater susceptibility of older persons to disease is a biologic consequence of aging. St. Louis encephalitis has an abrupt onset after an incubation period of 4–21 days, sometimes following a prodrome, and begins with fever, lethargy, confusion, and headache. In addition, nuchal rigidity, hypotonia, hyperreflexia, myoclonus, and tremors are common. Severe cases can include cranial nerve palsies, hemiparesis, and seizures. Patients often report dysuria and may have viral antigen in urine as well as pyuria. The overall lethality is generally ~7% but may reach 20% among patients >60 years of age. Recovery is slow. Emotional lability, difficulties with concentration and memory, asthenia, and tremors are commonly prolonged in older convalescent patients. The CSF of patients with St. Louis encephalitis usually contains tens to hundreds of leukocytes, with a lymphocytic predominance and a left shift. The CSF glucose concentration is normal in these patients.

TICK-BORNE VIRAL ENCEPHALITIS

Tick-borne encephalitis viruses are currently subdivided into four groups: the western/European subtype (previously called central European encephalitis virus), the (Ural-)Siberian subtype (previously called Russian spring–summer encephalitis virus), the Far Eastern subtype, and the louping ill subtype (previously called louping ill virus or, in Japan, Negishi virus). Small mammals and grouse, deer, and sheep are the vertebrate amplifiers for these viruses, which are transmitted by ticks. The risk of infection varies by geographic area and can be highly localized within a given area. Human infections usually follow either outdoor activities resulting in tick bites or consumption of raw (unpasteurized) milk from infected goats or, less commonly, from other infected animals (cows, sheep). Milk seems to represent the main transmission route for louping ill–subtype viruses, which cause disease only very rarely. The western/European-subtype viruses are transmitted mainly by *Ixodes ricinus* ticks from Scandinavia to the Ural Mountains. (Ural-)Siberian viruses are transmitted predominantly by *Ixodes persulcatus* ticks from Europe across the Ural Mountains to the Pacific Ocean. Louping ill–subtype viruses seem to be confined primarily to Great Britain. Several thousand infections with tick-borne encephalitis virus are recorded each year among people of all ages. Human tick-borne viral encephalitis occurs between April and October, with a peak in June and July.

Western/European viruses classically caused bimodal disease. After an incubation period of 7–14 days, the illness begins with a *fever–myalgia phase* (arthralgia, fever, headaches, myalgia, nausea) that lasts for 2–4 days and is thought to correlate with viremia. A subsequent remission for several days is followed by the recurrence of fever and the onset of meningeal signs. The *CNS phase* (7–10 days before onset of improvement) varies from mild aseptic meningitis, which is more common among younger patients, to severe (meningo)encephalitis with coma, seizures, tremors, and motor signs. Spinal and medullary involvement can lead to typical limb-girdle paralysis and respiratory paralysis. Most patients with western/European virus infections recover (lethality, 1%), and only a minority of patients have significant deficits. However, the lethality from (Ural-)Siberian virus infections reaches 7–8%.

Infections with Far Eastern viruses generally run a more abrupt course. The encephalitic syndrome caused by these viruses sometimes begins without a remission from the fever–myalgia phase and has more severe manifestations than the western/European syndrome. Lethality is high (20–40%), and major sequelae—most notably, lower motor neuron paralyses of the proximal muscles of the extremities, trunk, and neck—are common, developing in approximately one-half of patients. Thrombocytopenia sometimes develops during the initial febrile illness, resembling the early hemorrhagic phase of some other tick-borne flavivirus infections, such as Kyasanur Forest disease. In the early stage of the illness, virus may be isolated from the blood. In the CNS phase, IgM antibodies are detectable in serum and/or CSF.

Diagnosis of tick-borne viral encephalitis primarily relies on serology and detection of viral genomes by RT-PCR. There is no specific therapy for infection. However, effective alum-adjuvanted, formalin-inactivated virus vaccines are produced in Austria, Germany, and Russia in chicken embryo cells (FSME-Immun® and Encepur®). Two doses of the Austrian vaccine separated by an interval of 1–3 months appear to be effective in the field, and antibody responses are similar when vaccine is given on days 0 and 14. Because rare cases of postvaccination Guillain-Barré syndrome have been reported, vaccination should be reserved for persons likely to experience rural exposure in an endemic area during the season of transmission. Cross-neutralization for the western/European and Far Eastern variants has been established, but there are no published field studies on cross-protection among formalin-inactivated vaccines.

Because 0.2–4% of ticks in endemic areas may be infected, the use of immunoglobulin prophylaxis of tick-borne viral encephalitis has been raised. Prompt administration of high-titered specific antibody preparations should probably be undertaken, although no controlled data are available to prove the efficacy of this measure. Immunoglobulins should be considered because of the risk of antibody-mediated enhancement of infection or antigen–antibody complex deposition in tissues.

WEST NILE VIRUS INFECTION

West Nile virus is now the primary cause of arboviral encephalitis in the United States. From 1999 to 2015, 20,265 cases of neuroinvasive disease (e.g., meningitis, encephalitis, acute flaccid paralysis), with 1783 deaths, and 23,672 cases of non-neuroinvasive infection, with 128 deaths, were reported. West Nile virus was initially described as being transmitted among wild birds by *Culex* mosquitoes in Africa, Asia, and Southern Europe. In

addition, the virus has been implicated in severe and fatal hepatic necrosis in Africa. West Nile virus was introduced into New York City in 1999 and subsequently spread to other areas of the northeastern United States, causing die-offs among crows, exotic zoo birds, and other birds. The virus has continued to spread and is now found in almost all U.S. states as well as in Canada, Mexico, South America, and the Caribbean islands. *C. pipiens* mosquitoes remain the major vectors in the northeastern United States, but mosquitoes of several other *Culex* species and *A. albopictus* mosquitoes are also involved. Jays compete with crows and other corvids as amplifiers and lethal targets in other areas of the country.

West Nile virus is a common cause of febrile disease without CNS involvement (incubation period, 3–14 days), but it occasionally causes aseptic meningitis and severe encephalitis, particularly among the elderly. The fever–myalgia syndrome caused by West Nile virus differs from that caused by other viruses in terms of the frequent—rather than occasional—appearance of a maculopapular rash concentrated on the trunk (especially in children) and the development of lymphadenopathy. Back pain, fatigue, headache, myalgia, retroorbital pain, sore throat, nausea and vomiting, and arthralgia (but not arthritis) are common accompaniments that may persist for several weeks. Encephalitis, sequelae, and death are all more common among elderly, diabetic, and hypertensive patients and among patients with previous CNS insults. In addition to the more severe motor and cognitive sequelae, milder findings may include tremor, slight abnormalities in motor skills, and loss of executive functions. Intense clinical interest and the availability of laboratory diagnostic methods have made it possible to define a number of unusual clinical features. Such features include chorioretinitis, flaccid paralysis with histologic lesions resembling poliomyelitis, and initial presentation with fever and focal neurologic deficits in the absence of diffuse encephalitis. Immunosuppressed patients may have fulminant courses or develop persistent CNS infection. Virus transmission through both transplantation and blood transfusion has necessitated screening of blood and organ donors by nucleic acid–based tests. Occasionally, pregnant women infect their fetuses with West Nile virus.

Peribunyaviruses

CALIFORNIA (MENINGO)ENCEPHALITIS

The isolation of California encephalitis virus established California serogroup orthobunyaviruses as causes of encephalitides. However, California encephalitis virus has been implicated in only a very few cases of encephalitis, whereas its close relative, La Crosse virus, is the major cause of encephalitis in this serogroup (~80–100 cases per year in the United States). California (meningo)encephalitis due to La Crosse virus infection is most commonly reported from the upper midwestern United States but is also found in other areas of the central and eastern parts of the country, such as West Virginia, Tennessee, North Carolina, and Georgia. The serogroup includes 13 other viruses, some of which (e.g., Inkoo, Jamestown Canyon, Lumbo, snowshoe hare, and Ťahyňa viruses) also cause human disease. Transovarial transmission is a strong component of transmission of the California serogroup viruses in *Aedes* and *Ochlerotatus* mosquitoes. The vector of La Crosse virus is the *Ochlerotatus triseriatus* mosquito. In addition to transovarial transmission, acquisition through feeding on viremic chipmunks and other mammals and venereal transmission can result in infection of this mosquito. *O. triseriatus* breeds in sites such as tree holes and abandoned tires and bites during daylight hours. The habits of this mosquito correlate with the risk factors for human cases: recreation in forested areas, residence at a forest's edge, and the presence of water-containing abandoned tires around the home. Intensive environmental modification based on these findings has reduced the incidence of disease in a highly endemic area in the midwestern United States.

Most humans are infected from July through September. *A. albopictus* mosquitoes efficiently transmit La Crosse virus to mice and also transmit the agent transovarially in the laboratory. This aggressive anthropophilic mosquito has the capacity to urbanize, and its possible impact on transmission of virus to humans is of concern. The prevalence of antibody to La Crosse virus in humans is ≥20% in endemic areas, a figure indicating that infection is common but often asymptomatic. CNS disease has been recognized primarily in children <15 years of age.

The illness from La Crosse virus varies from aseptic meningitis accompanied by confusion to severe and occasionally fatal encephalitis (lethality, <0.5%). The incubation period is ~3–7 days. Although there may be prodromal symptoms/signs, the onset of CNS disease is sudden, with fever, headache, and lethargy often joined by nausea and vomiting, convulsions (in one-half of patients), and coma (in one-third of patients). Focal seizures, hemiparesis, tremor, aphasia, chorea, Babinski signs, and other evidence of significant neurologic dysfunction are common, but residual disease is not. Approximately 10% of patients have recurrent seizures in the succeeding months. Other serious sequelae of La Crosse virus infection are rare, although a decrease in scholastic standing among children has been reported, and mild personality change has occasionally been suggested.

The blood leukocyte count is commonly elevated in patients with La Crosse virus infection, sometimes reaching 20,000/μL, and is usually accompanied by a left shift. CSF leukocyte counts are typically 30–500/μL, usually with a mononuclear cell predominance (although 25–90% of cells are polymorphonuclear in some patients). The blood protein concentration is normal or slightly increased, and the glucose concentration is normal. Specific virologic diagnosis based on IgM-capture assays of serum and CSF is efficient. The only human anatomic site from which virus has been isolated is the brain.

Treatment is supportive over a 1- to 2-week acute phase during which status epilepticus, cerebral edema, and inappropriate secretion of antidiuretic hormone are important concerns. A phase 2B clinical trial of IV ribavirin in children with La Crosse virus infection was discontinued during dose escalation because of adverse effects.

Jamestown Canyon virus has been implicated in several cases of encephalitis in adults, usually with a significant respiratory illness at onset. Human infection with this virus has been documented in New York, Wisconsin, Ohio, Michigan, Ontario, and other areas of North America where the vector mosquito (*Aedes stimulans*) feeds on its main host, the white-tailed deer (*Odocoileus virginianus*). Ťahyňa virus can be found in central Europe, Russia, China, and Africa. The virus is a prominent cause of febrile disease but can also cause pharyngitis, pulmonary syndromes, aseptic meningitis, or meningoencephalitis.

Rhabdoviruses

CHANDIPURA VIRUS INFECTION

Chandipura virus is an emerging and increasingly important human virus in India, where it is transmitted among hedgehogs by mosquitoes and sandflies. In humans, the disease begins as an influenza-like illness, with fever, headache, abdominal pain, nausea, and vomiting. These manifestations are followed by neurologic impairment and infection-related or autoimmune-mediated encephalitis. Chandipura virus infection is characterized by high lethality in children. Several hundred cases of infection are recorded in India every year. Infections with other arthropod-borne rhabdoviruses (Isfahan, Piry, vesicular stomatitis Indiana, vesicular stomatitis New Jersey viruses) may imitate the early febrile stage of Chandipura virus infection.

Togaviruses

EASTERN EQUINE ENCEPHALITIS

This disease is encountered primarily in swampy foci along the eastern coast of the United States, with a few inland foci as far removed as Michigan. Infected humans present for medical care from June through October. During this period, the bird–*Culiseta* mosquito cycle spills over into other vectors such as *Aedes sollicitans* or *Aedes vexans* mosquitoes, which are more likely to feed on mammals. There is concern over the potential role of introduced *A. albopictus* mosquitoes, which have been found to be infected with eastern equine encephalitis virus and are an effective experimental vector in the laboratory. Horses are a common target for the virus. Contact with unvaccinated horses may be associated with human disease, but horses probably do not play a significant role in amplification of the virus.

Eastern equine encephalitis is one of the most destructive of the arboviral diseases, with a sudden onset after an incubation period of ~5–10 days, rapid progression, 50–75% lethality, and frequent sequelae in survivors. This severity is reflected in the extensive necrotic lesions and polymorphonuclear infiltrates found at postmortem examination of the brain. Acute polymorphonuclear CSF pleocytosis, often occurring during the first 1–3 days of disease, is another indication of severity. In addition, leukocytosis with a left shift is a common feature. A formalin-inactivated vaccine has been used to protect laboratory workers but is not generally available or applicable.

VENEZUELAN EQUINE FEVER

Venezuelan equine encephalitis viruses are separated into epizootic viruses (subtypes IA/B and IC) and enzootic viruses (subtypes ID, IE, and IF). Closely related enzootic viruses are Everglades virus, Mucambo virus, and Tonate virus. Enzootic viruses are found primarily in humid tropical-forest habitats and are maintained between culicoid mosquitoes and rodents. These viruses cause acute febrile human disease but are not pathogenic for horses and do not cause epizootics. Everglades virus has caused encephalitis in humans in Florida in the United States. Extrapolation from the rate of genetic change suggests that Everglades virus may have been introduced into Florida <200 years ago. Everglades virus is most closely related to the ID-subtype viruses that appear to have given evolutionary rise to the epizootic variants active in South America.

Epizootic viruses have an unknown natural cycle but periodically cause extensive epizootics/epidemics in equids and humans in the Americas. These epizootics/epidemics are the result of high-level viremia in horses and mules, which transmit the infection to several types of mosquitoes. Infected mosquitoes in turn infect humans and perpetuate virus transmission. Humans also have high-level viremia, but their role in virus transmission is unclear. Relatively restricted epizootics of Venezuelan equine fever occurred repeatedly in South America at intervals of ≤10 years from the 1930s until 1969, when a massive epizootic, including tens of thousands of equine and human infections, spread throughout Central America and Mexico, reaching southern Texas in 1971. Genetic sequencing suggested that the virus from that outbreak originated from residual “un-inactivated” IA/B-subtype virus in veterinary vaccines. The outbreak was terminated in Texas with a live attenuated vaccine (TC-83) originally developed for human use by the U.S. Army; the epizootic virus was then used for further production of inactivated veterinary vaccines. No further major epizootic disease outbreaks occurred until 1995 and 1996, when large epizootics of Venezuelan equine fever occurred in Colombia/Venezuela and Mexico, respectively. Of the more than 85,000 clinical cases, 4% (with a higher proportion among children than adults) included neurologic symptoms/signs, and 300 cases ended in death. The viruses involved in these epizootics as well as previously epizootic IC viruses are close phylogenetic relatives of known enzootic ID viruses. This finding suggests that active evolution and selection of epizootic viruses are underway in South America.

During epizootics, extensive human infection is typical, with clinical disease occurring in 10–60% of infected individuals. Most infections result in notable acute febrile disease, whereas relatively few infections (5–15%) result in neurologic disease. A low rate of CNS invasion is supported by the absence of encephalitis among the many infections resulting from exposure to aerosols in the laboratory setting or from vaccination accidents.

The prevention of epizootic Venezuelan equine fever depends on vaccination of horses with the attenuated TC-83 vaccine or with an inactivated vaccine prepared from that variant. Enzootic viruses are genetically and antigenically different from epizootic viruses, and protection against the former with vaccines prepared from the latter is relatively ineffective. Humans can be protected by immunization with similar vaccines prepared from Everglades virus, Mucambo virus, and Venezuelan equine encephalitis virus, but the use of the vaccines is restricted to laboratory personnel because of reactogenicity, possible fetal pathogenicity, and limited availability.

WESTERN EQUINE ENCEPHALITIS

The primary maintenance cycle of western equine encephalitis virus in the United States is between *C. tarsalis* mosquitoes and birds, principally sparrows and finches. Equids and humans become infected, and both suffer encephalitis without amplifying the virus in nature. St. Louis encephalitis virus is transmitted in a similar cycle in the same regions harboring western equine encephalitis virus; disease caused by the former occurs about a month earlier than that caused by the latter (July through October). Large epidemics of western equine encephalitis occurred in the western and central United States and Canada during the 1930s through 1950s, but in recent years the disease has been uncommon. From 1964 through 2010, only 640 cases were reported in the United States. This decline in incidence may reflect in part the integrated approach to mosquito management that has been employed in irrigation projects and in part the increasing use of agricultural pesticides. The decreased incidence of western equine encephalitis almost certainly reflects the increased tendency for humans to be indoors behind closed windows at dusk—the peak biting period by the major vector.

After an incubation period of ~5–10 days, western equine encephalitis virus causes a typical diffuse viral encephalitis, with an increased attack rate and increased morbidity among the young, particularly children <2 years old. In addition, lethality is high among the young and the very elderly (3–7% overall). One-third of individuals who have convulsions during the acute illness have subsequent seizure activity. Infants <1 year old—particularly those in the first months of life—are at serious risk of motor and intellectual damage. Twice as many males as females develop clinical encephalitis after 5–9 years of age. This difference in incidence may be related to greater outdoor exposure of boys to the vector but may also be due in part to biologic differences. A formalin-inactivated vaccine has been used to protect laboratory workers but is not generally available.

FEVER AND MYALGIA

The fever and myalgia syndrome is most commonly associated with zoonotic virus infection. Many of the numerous viruses listed in Table 204-1 probably cause at least a few cases of this syndrome, but only some of these viruses have prominent associations with the syndrome and are of biomedical importance. The fever and myalgia syndrome typically begins with the abrupt onset of fever, chills, intense myalgia, and malaise. Patients may also report joint or muscle pains, but true arthritis is not found. Anorexia is characteristic and may be accompanied by nausea or even vomiting. Headache is common and may be severe, with photophobia and retroorbital pain. Physical findings are minimal and are usually confined to conjunctival injection with pain on palpation of muscles or the epigastrium. The duration of symptoms/signs is quite variable (generally 2–5 days), with a biphasic course in some instances. The spectrum of disease varies from subclinical to temporarily incapacitating. Less constant findings include a nonpruritic maculopapular rash. Epistaxis may occur but does not necessarily indicate a bleeding diathesis. A minority of patients may develop aseptic meningitis. This diagnosis is difficult to make in remote areas, given patients’ photophobia and myalgia as well as the lack of opportunity to examine the CSF. Although pharyngitis or radiographic evidence of pulmonary infiltrates is found in some patients, the agents causing this syndrome are not primary respiratory pathogens.

The differential diagnosis includes anicteric leptospirosis, rickettsial diseases, and the early stages of other syndromes discussed in this chapter. The fever and myalgia syndrome is often described as “influenza-like,” but the usual absence of cough and coryza makes influenza an unlikely confounder except at the earliest stages. Treatment is supportive, but acetylsalicylic acid is avoided because of the potential for exacerbated bleeding or Reye’s syndrome. Complete recovery is the general outcome for people with this syndrome, although prolonged asthenia and nonspecific symptoms have been described in some patients, particularly after infection with lymphocytic choriomeningitis virus or dengue viruses 1–4.

Efforts for preventing viral infection are best based on vector control, which, however, may be expensive or impossible. For mosquito control, destruction of breeding sites is generally the most economically and environmentally sound approach. Emerging containment technologies include the release of genetically modified mosquitoes and the spread of *Wolbachia* bacteria to limit mosquito multiplication rates. Depending on the vector and its habits, other possible approaches include the use of screens or other barriers (e.g., permethrin-impregnated bed nets) to prevent the vector from entering dwellings, judicious application of arthropod repellents such as *N*,*N*,-diethyltoluamide (DEET) to the skin, use of long-sleeved and ideally permethrin-impregnated clothing, and avoidance of the vectors’ habitats and times of peak activity.

Bunyaviruses

Numerous bunyaviruses cause fever and myalgia. Many of these viruses cause individual infections and usually do not result in epidemics. These viruses include arenaviruses, such as lymphocytic choriomeningitis virus; hantaviruses, such as the orthohantavirus Choclo virus; nairoviruses, such as the orthonairoviruses Dugbe virus and Nairobi sheep disease virus; peribunyaviruses, such as the viruses of the orthobunyavirus Anopheles A serogroup (e.g., Tacaiuma virus), the Bunyamwera serogroup (Bunyamwera, Batai, Cache Valley, Fort Sherman, Germiston, Guaroa, Ilesha, Ngari, Shokwe, and Xingu viruses), the Bwamba serogroup (Bwamba virus, Pongola virus), the Guamá serogroup (Catú virus, Guamá virus), the Nyando serogroup (Nyando virus), the Wyeomyia serogroup (Wyeomyia virus), and the ungrouped orthobunyavirus Tataguine virus; and phenuiviruses, such as the “banyangvirus” Bhanja complex (Bhanja virus, Heartland virus) and the phlebovirus Candiru complex (Alenquer, Candiru, Escharate, Maldonado, Morumbi, and Serra Norte viruses).

ARENAVIRUSES

Lymphocytic choriomeningitis/meningoencephalitis is the only human mammarenavirus infection resulting predominantly in fever and myalgia. Lymphocytic choriomeningitis virus is transmitted to humans from the common house mouse (*Mus musculus*) by aerosols of excreta or secreta. The virus is maintained in the mouse mainly by vertical transmission from infected dams. The vertically infected mouse remains viremic and sheds virus for life, with high concentrations of virus in all tissues. Infected colonies of pet hamsters also can serve as a link to humans. Infections among scientists and animal caretakers can occur because the virus is widely used in immunology laboratories as a model of T cell function and can silently infect cell cultures and passaged tumor lines. In addition, patients may have a history of residence in rodent-infested housing or other exposure to rodents. An antibody prevalence of ~5–10% has been reported among adults from Argentina, Germany, and the United States.

Lymphocytic choriomeningitis/meningoencephalitis differs from the general syndrome of fever and myalgia in that the onset is gradual. Conditions occasionally associated with the disease are orchitis, transient alopecia, arthritis, pharyngitis, cough, and maculopapular rash. An estimated one-fourth of patients (or fewer) experience a febrile phase of 3–6 days. After a brief remission, many develop renewed fever accompanied by severe headache, nausea and vomiting, and meningeal signs lasting for ~1 week (the CNS phase). These patients virtually always recover fully, as do the rare patients with clear-cut signs of encephalitis. Recovery may be delayed by transient hydrocephalus. During the initial febrile phase, leukopenia and thrombocytopenia are common, and virus can usually be isolated from blood. During the CNS phase, the virus may be found in the CSF, and antibodies are present in the blood. The pathogenesis of lymphocytic choriomeningitis/meningoencephalitis is thought to resemble manifestations following direct intracranial inoculation of the virus into adult mice. The onset of the immune response leads to T cell–mediated immunopathologic meningitis. During the meningeal phase, CSF mononuclear-cell counts range from the hundreds to the low thousands per microliter, and hypoglycorrhachia is found in one-third of patients.

IgM-capture ELISA, immunochemistry, and RT-PCR are used in the diagnosis of lymphocytic choriomeningitis/meningoencephalitis. IgM-capture ELISA of serum and CSF usually yields positive results; RT-PCR assays have been developed for probing CSF. Because patients who have fulminant infections transmitted by recent organ transplantation do not mount an immune response, immunohistochemistry or RT-PCR is required for diagnosis. Infection should be suspected in acutely ill febrile patients with marked leukopenia and thrombocytopenia. In patients with aseptic meningitis, any of the following suggests lymphocytic choriomeningitis/meningoencephalitis: a well-marked febrile prodrome, adult age, occurrence in the autumn, low CSF glucose levels, or CSF mononuclear-cell counts of >1000/μL. In pregnant women, infection may lead to fetal invasion with consequent congenital hydrocephalus, microcephaly, and/or chorioretinitis. Because the maternal infection may be mild, causing only a short febrile illness, antibodies to the virus should be sought in both the mother and the fetus under suspicious circumstances, particularly in TORCH (toxoplasmosis, rubella, cytomegalovirus, herpes simplex, and HIV)–negative neonatal hydrocephalus.

ORTHOBUNYAVIRUS GROUP C SEROGROUP

Apeú, Caraparú, Itaquí, Madrid, Marituba, Murutucú, Nepuyo, Oriboca, Ossa, Restan, and Zungarococha viruses are among the most common causes of arboviral infection in humans entering South American jungles. These viruses cause acute febrile disease and are transmitted by mosquitoes in neotropical forests.

ORTHOBUNYAVIRUS SIMBU SEROGROUP

Oropouche virus is transmitted in Central and South America by biting midges (*Culicoides paraensis*), which often breed to high density in cacao husks and other vegetable detritus found in towns and cities. Explosive epidemics involving thousands of patients have been reported from several towns in Brazil and Peru. Rash and aseptic meningitis have been detected in a number of patients. Iquitos virus, a recently discovered reassortant and close relative of Oropouche virus, causes disease that is easily mistaken for Oropouche virus disease; its overall epidemiologic significance remains to be determined.

PHLEBOVIRUS SANDFLY FEVER GROUP

The phlebovirus sandfly fever group consists of numerous viruses that may cause human infection. Sandfly fever Cyprus virus, sandfly fever Ethiopia virus, sandfly fever Sicilian virus, and sandfly fever Turkey virus (and the encephalitis-causing Chios virus) are very closely related genetically and antigenically; they likely belong to the same species, in which they may constitute variants of the same virus. In contrast, sandfly fever Naples virus (SFNV) is genetically and antigenically distantly related to these viruses. SFNV has not been detected in sandflies, humans, or nonhuman vertebrates since the 1980s and therefore may be extinct. SFNV is the prototypic member of the species *Sandfly fever Naples phlebovirus* that includes other human viruses such as Granada and Toscana viruses. Toscana virus is thus far the only phlebovirus transmitted by sandflies that is known to cause

diseases affecting the central and the peripheral nervous systems, such as encephalitis, meningitis, myositis, or polymyeloradiculopathy. *Phlebotomus* sandflies transmit the virus, probably by biting small mammals and humans. Female sandflies may be infected by the oral route as they take a blood meal and may transmit the virus to offspring when they lay their eggs after a second blood meal. This prominent transovarial transmission confounds virus control.

Sandfly fever is found in the circum-Mediterranean area, extending to the east through the Balkans into parts of China as well as into Western Asia. Sandflies are found in both rural and urban settings and are known for their short flight ranges and their small sizes; the latter enables them to penetrate standard mosquito screens and netting. Epidemics have been described in the wake of natural disasters and wars. After World War II, extensive spraying in parts of Europe to control malaria greatly reduced sandfly populations and SFNV transmission; the incidence of sandfly fever continues to be low.

A common pattern of disease in endemic areas consists of high attack rates among travelers and military personnel and little or no disease in the local population, who are protected after childhood infection. Toscana virus infection is common during the summer among rural residents and vacationers, particularly in Italy, Spain, and Portugal; a number of cases have been identified in travelers returning to Germany and Scandinavia. The disease may manifest as an uncomplicated febrile illness but is often associated with aseptic meningitis, with virus isolated from the CSF.

Coclé virus and Punta Toro virus are phleboviruses that are not part of the sandfly fever serocomplex but that, like the members of this complex, are transmitted by sandflies. These two viruses cause a sandfly fever–like disease in Latin American and Caribbean tropical forests, respectively, where the vectors rest on tree buttresses. Epidemics have not been reported, but antibody prevalence among inhabitants of villages in endemic areas indicates a cumulative lifetime exposure rate of >50% in the case of Punta Toro virus.

Flaviviruses

The most clinically important flaviviruses that cause the fever and myalgia syndrome are dengue viruses 1–4. In fact, dengue is probably the most important arthropod-borne viral disease worldwide, with ~390 million infections occurring per year, of which ~96 million cause signs of disease. Year-round transmission of dengue viruses 1–4 occurs between latitudes 25°N and 25°S, but seasonal forays of the viruses into the United States and Europe have been documented. All four viruses have *A. aegypti* mosquitoes as their principal vectors. Through increasing spread of mosquitoes throughout the tropics and subtropics and international travel of infected humans, large areas of the world have become vulnerable to the introduction of dengue viruses. Thus, dengue and severe dengue (see “Viral Hemorrhagic Fevers,” below) are becoming increasingly common. For instance, conditions favorable to dengue virus 1–4 transmission via *A. aegypti* mosquitoes exist in Hawaii and the southern United States. The range of a lesser dengue virus vector (*A. albopictus*) now extends from Asia to the United States, the Indian Ocean, parts of Europe, and Hawaii. *A. aegypti* mosquitoes typically breed near human habitation, using relatively fresh water from sources such as water jars, vases, discarded containers, coconut husks, and old tires. These mosquitoes usually inhabit dwellings and bite during the day. Bursts of dengue cases are to be expected in the southern United States, particularly along the Mexican border, where containers of water may be infested with *A. aegypti* mosquitoes. Closed habitations with air-conditioning may inhibit transmission of many arboviruses, including dengue viruses 1–4.

Dengue begins after an incubation period averaging 4–7 days, when the typical patient experiences the sudden onset of fever, frontal headache, retroorbital pain, and back pain along with severe myalgias. These symptoms gave rise to the colloquial designation of dengue as “break-bone fever.” Often a transient macular rash appears on the first day, as do adenopathy, palatal vesicles, and scleral injection. The illness may last a week, with additional symptoms and clinical signs usually including anorexia, nausea or vomiting, and marked cutaneous hypersensitivity. Near the time of defervescence on days 3–5, a maculopapular rash begins on the trunk and spreads to the extremities and the face. Epistaxis and scattered petechiae are often noted in uncomplicated dengue, and preexisting gastrointestinal lesions may bleed during the acute illness.

Laboratory findings of dengue include leukopenia, thrombocytopenia, and, in many cases, elevations of serum aminotransferase concentrations. The diagnosis is made by IgM ELISA or paired serology during recovery or by antigen-detection ELISA or RT-PCR during the acute phase. Virus is readily isolated from blood in the acute phase if mosquito inoculation or mosquito cell culture is used.

Orthomyxoviruses

Bourbon virus was recently identified as the cause of a severe and sometimes fatal febrile disease of humans in the midwestern and southern United States.

Reoviruses

Several orbiviruses (Lebombo, Kemerovo, Orungo, and Tribeč viruses) and coltiviruses (Colorado tick fever, Eyach, and Salmon River viruses) can cause fever and myalgia in humans. With the exception of Lebombo and Orungo viruses, all of these viruses are transmitted by ticks. The most important reoviral arthropod-borne disease is Colorado tick fever. Several hundred patients with this disease are reported annually in the United States. The infection is acquired between March and November through the bite of an infected ixodid tick, the Rocky Mountain wood tick (*Dermacentor andersoni*), in mountainous western regions at altitudes of 1200–3000 m. Small mammals serve as amplifying hosts. The most common presentation is fever and myalgia; meningoencephalitis is not uncommon, and hemorrhagic disease, pericarditis, myocarditis, orchitis, and pulmonary presentations have also been reported. Rash develops in a minority of patients. Leukopenia and thrombocytopenia are also noted. The disease usually lasts 7–10 days and is often biphasic. The most important differential diagnostic considerations since the beginning of the twentieth century have been Rocky Mountain spotted fever (although Colorado tick fever is much more common in Colorado) and tularemia. Colorado tick fever virus replicates for several weeks in erythropoietic cells and can be found in erythrocytes. This feature, detected in erythroid smears stained by immunofluorescence, can be diagnostically helpful and is important during screening of blood donors.

PULMONARY DISEASE

Hantavirus (cardio)pulmonary syndrome, or H(C)PS, was first described in 1993, but retrospective identification of cases by immunohistochemistry (1978) and serology (1959) support the idea that H(C)PS is a recently discovered rather than a truly new disease. The causative agents are orthohantaviruses of a distinct phylogenetic lineage that is associated with the cricetid rodent subfamily *Sigmodontinae*. Sin nombre virus, which chronically infects North American deer mice (*Peromyscus maniculatus*), is the most important agent of H(C)PS in the United States. Several other related viruses (Anajatuba, Andes, Araraquara, Araucária, bayou, Bermejo, Black Creek Canal, Blue River, Castelo dos Sonhos, El Moro Canyon, Juquitiba, Laguna Negra, Lechiguanas, Maciel, Monongahela, Muleshoe, New York, Orán, Paranoá, Pergamino, Río Mamoré, and Tunari viruses) cause the disease in North and South America. Andes virus is unusual in that it has been implicated in human-to-human transmission. H(C)PS particularly affects rural residents living in dwellings permeable to rodent entry or working in occupations that pose a risk of rodent exposure. Each type of rodent has its own particular habits; in the case of deer mice, these behaviors include living in and around human habitation.

H(C)PS begins with a prodrome of ~3–4 days (range, 1–11 days) comprising fever, malaise, myalgia, and—in many cases—gastrointestinal disturbances such as abdominal pain, nausea, and vomiting. Dizziness is common, and vertigo is occasional. Severe prodromal symptoms/signs may bring some patients to medical attention, but most cases are recognized as the pulmonary phase begins. Typical signs are slightly lowered blood pressure, tachycardia, tachypnea, mild hypoxemia, thrombocytopenia, and early radiographic signs of pulmonary edema. Physical findings in the chest are often surprisingly scant. The conjunctival and cutaneous signs of vascular involvement seen in hantavirus VHFs (see below) are uncommon. During the next few hours, decompensation may progress rapidly to severe hypoxemia and respiratory failure.

The H(C)PS differential diagnosis includes abdominal surgical conditions and pyelonephritis as well as rickettsial disease, sepsis, meningococcemia, plague, tularemia, influenza, and relapsing fever. A specific diagnosis is best made by IgM antibody testing of acute-phase serum, which has yielded positive results even in the prodrome. Tests using a sin nombre virus antigen detect antibodies to the related H(C)PS-causing hantaviruses. Occasionally, heterotypic viruses will react only in the IgG ELISA, but such a finding is highly suspicious given the very low seroprevalence of these viruses in normal populations. RT-PCR is usually positive when used to test blood clots obtained in the first 7–9 days of illness and when used to test tissues. This assay is useful in identifying the infecting virus in areas outside the home range of deer mice and in atypical cases.

During the prodrome, the differential diagnosis of H(C)PS is difficult, but by the time of presentation or within 24 h thereafter, a number of diagnostically helpful clinical features become apparent. Cough usually is not present at the outset. Interstitial edema is evident on a chest x-ray. Later, bilateral alveolar edema with a central distribution develops in the setting of a normal-sized heart; occasionally, the edema is initially unilateral. Pleural effusions are often seen. Thrombocytopenia, circulating atypical lymphocytes, and a left shift (often with leukocytosis) are almost always evident; thrombocytopenia is a particularly important early clue. Hemoconcentration, hypoalbuminemia, and proteinuria should also be sought for diagnosis. Although thrombocytopenia virtually always develops and prolongation of the partial thromboplastin time is the rule, clinical evidence for coagulopathy or laboratory indications of disseminated intravascular coagulation (DIC) are found in only a minority of severely ill patients. Patients with severe illness also have acidosis and elevated serum lactate concentrations. Mildly increased values in renal function tests are common, but patients with severe H(C)PS often have markedly elevated serum creatinine concentrations. Some New World hantaviruses other than sin nombre virus (e.g., Andes virus) have been associated with more kidney involvement, but few such cases have been studied.

Management of H(C)PS during the first few hours after presentation is critical. The goal is to prevent severe hypoxemia by oxygen therapy, with intubation and intensive respiratory management if needed. During this period, hypotension and shock with increasing hematocrit invite aggressive fluid administration, but this intervention should be undertaken with great caution. Because of low cardiac output with myocardial depression and increased pulmonary vascular permeability, shock should be managed expectantly with vasopressors and modest infusion of fluid guided by pulmonary capillary wedge pressure. Mild cases can be managed by frequent monitoring and oxygen administration without intubation. Many patients require intubation to manage hypoxemia and developing shock. Extracorporeal membrane oxygenation is instituted in severe cases, ideally before the onset of shock. The procedure is indicated in patients who have a cardiac index of 2.3 L/min/m^2 or an arterial oxygen tension/fractional inspired oxygen (PaO_2/FIO_2) ratio of <50 and who are unresponsive to conventional support. Lethality remains at ~30–40% even with good management, but most patients surviving the first 48 h of hospitalization are extubated and discharged within a few days with no apparent long-term residua. The antiviral drug ribavirin inhibits hantaviruses in vitro but did not have a marked effect on patients treated in an open-label study.

VIRAL HEMORRHAGIC FEVER

VHF is a constellation of findings based on vascular instability and decreased vascular integrity. An assault, direct or indirect, on the microvasculature leads to increased permeability and (particularly when platelet function is decreased) to actual disruption and local hemorrhage (a positive tourniquet sign). Blood pressure is decreased, and in severe cases shock supervenes. Cutaneous flushing and conjunctival suffusion are examples of common, observable abnormalities in the control of local circulation. Hemorrhage occurs infrequently. In most patients, hemorrhage is an indication of widespread vascular damage rather than a life-threatening loss of blood volume. In some VHFs, specific organs may be particularly impaired. For instance, the kidneys are primary targets in HFRS, and the liver is a primary target in yellow fever and filovirus diseases. However, in all of these diseases, generalized circulatory disturbance is critically important. The pathogenesis of VHF is poorly understood and varies among the viruses regularly implicated in the syndrome. In some viral infections, direct damage to the vascular system or even to parenchymal cells of target organs is an important factor; in other viral infections, soluble mediators are thought to play a major role in the development of hemorrhage or fluid redistribution.

The acute phase in most cases of VHF is associated with ongoing virus replication and viremia. VHFs begin with fever and myalgia, usually of abrupt onset. (Mammarenavirus infections are the exceptions as they often develop gradually.) Within a few days, the patient presents for medical attention because of increasing prostration that is often accompanied by abdominal or chest pain, anorexia, dizziness, severe headache, hyperesthesia, photophobia, and nausea or vomiting and other gastrointestinal disturbances. Initial examination often reveals only an acutely ill patient with conjunctival suffusion, tenderness to palpation of muscles or abdomen, and borderline hypotension or postural hypotension, perhaps with tachycardia. Petechiae (often best visualized in the axillae), flushing of the head and thorax, periorbital edema, and proteinuria are common. AST concentrations are usually elevated at presentation or within a day or two thereafter. Hemoconcentration from vascular leakage, which is usually evident, is most marked in HFRS and in severe dengue. The seriously ill patient progresses to more severe clinical signs and develops shock and other findings typical of the causative virus. Shock, multifocal bleeding, and CNS involvement (encephalopathy, coma, seizures) are all poor prognostic signs.

One of the major diagnostic clues to VHF is travel to an endemic area within the incubation period for a given syndrome. Except in infections with Seoul, dengue, and yellow fever viruses, which have urban hosts/vectors, travel to a rural setting is especially suggestive of a diagnosis of VHF. In addition, several diseases considered in the differential diagnosis—falciparum malaria, shigellosis, typhoid fever, leptospirosis, relapsing fever, and rickettsial diseases—are treatable and potentially lethal.

Early recognition of VHF is important because of the need for virus-specific therapy and supportive measures. Such measures include prompt, atraumatic hospitalization; judicious fluid therapy that takes into account the patient's increased capillary permeability; administration of cardiotonic drugs; use of vasopressors to maintain blood pressure at levels that will support renal perfusion; treatment of the relatively common secondary bacterial (and the more rare fungal) infections; replacement of clotting factors and platelets as indicated; and the usual precautionary measures used in the treatment of patients with hemorrhagic diatheses. DIC should be treated only if clear laboratory evidence of its existence is found and if laboratory monitoring of therapy is feasible; there is no proven benefit of such therapy. The available evidence suggests that VHF patients have decreased cardiac output and will respond poorly to fluid loading as it is often practiced in the treatment of shock associated with bacterial sepsis. Specific therapy is available for several of the VHFs. Strict barrier nursing and other precautions against infection of medical staff and visitors are indicated when VHFs are encountered except when the illness is due to dengue viruses, hantaviruses, Rift Valley fever virus, or yellow fever virus.

Novel VHF-causing agents are still being discovered. Besides the viruses listed below, the latest additions are the "banyangvirus" severe fever with thrombocytopenia syndrome virus, which is continuing to cause VHF cases in China, Korea, and Japan, and possibly the tibrovirus Bas-Congo virus, which has been associated with three cases of VHF in the Democratic Republic of the Congo. However, Koch's postulates have not yet been fulfilled to prove cause and effect in the case of Bas-Congo virus.

Bunyaviruses

The most important VHF-causing bunyaviruses are arenaviruses (Junín, Lassa, and Machupo viruses), hantaviruses, nairoviruses (Crimean-Congo hemorrhagic fever virus), and phenuiviruses (Rift Valley fever and severe fever with thrombocytopenia syndrome viruses). Other bunyaviruses—e.g., the Garissa variant of Ngari virus and Ilesha virus (both orthobunyaviruses) or Chapare, Guanarito, Lujo, and Sabiá viruses (all mammarenaviruses)—have caused sporadic VHF outbreaks.

JUNÍN/ARGENTINIAN AND MACHUPO/BOLIVIAN HEMORRHAGIC FEVERS

These severe diseases (with lethality reaching 15–30%) are caused by Junín virus and Machupo virus, respectively. Their clinical presentations are similar, but their epidemiology differs because of the distribution and behavior of the viruses' rodent reservoirs. Junín/Argentinian hemorrhagic fever has thus far been recorded only in rural areas of Argentina, whereas Machupo/Bolivian hemorrhagic fever seems to be confined to rural Bolivia. Infection with the causative agents almost always results in disease, and all ages and both sexes are affected. Person-to-person or nosocomial transmission is rare but has occurred. The transmission of Junín/Argentinian hemorrhagic fever from convalescing men to their wives suggests the need for counseling of patients with mammarenavirus hemorrhagic fever concerning the avoidance of intimate contacts for several weeks after recovery. In contrast to the pattern in Lassa fever (see below), thrombocytopenia—often marked—is the rule, hemorrhage is common, and CNS dysfunction (e.g., marked confusion, tremors of the upper extremities and tongue, and cerebellar signs) is much more common in disease caused by Junín virus and Machupo virus. Some cases follow a predominantly neurologic course, with a poor prognosis.

The clinical laboratory is helpful in diagnosis since thrombocytopenia, leukopenia, and proteinuria are typical findings. Junín/Argentinian hemorrhagic fever is readily treated with convalescent-phase plasma given within the first 8 days of illness. In the absence of passive antibody therapy, IV ribavirin in the dose recommended for Lassa fever is likely to be effective in all the South American VHFs caused by mammarenaviruses. A safe, effective, live attenuated vaccine exists for Junín/Argentinian hemorrhagic fever. After vaccination of more than 250,000 high-risk persons in the endemic area, the incidence of this VHF decreased markedly. In experimental animals, this vaccine is cross-protective against Machupo/Bolivian hemorrhagic fever.

LASSA FEVER

Lassa virus is known to cause endemic and epidemic disease in Nigeria, Sierra Leone, Guinea, and Liberia, although it is probably more widely distributed in western Africa. In countries where Lassa virus is endemic, Lassa fever can be a prominent cause of febrile disease. For example, in one hospital in Sierra Leone, laboratory-confirmed Lassa fever is consistently responsible for one-fifth of admissions to the medical wards. In western Africa alone, probably tens of thousands of Lassa virus infections occur annually. Lassa virus can be transmitted by close person-to-person contact. The virus is often present in urine during convalescence and is suspected to be present in seminal fluid early in recovery. Nosocomial spread has occurred but is uncommon if proper sterile parenteral techniques are used. All ages and both sexes are affected; the incidence of disease is highest in the dry season, but transmission takes place year-round.

Among the VHF agents, only mammarenaviruses are typically associated with a gradual onset of illness, which begins after an incubation period of 5–16 days. Hemorrhage is seen in only ~15–30% of Lassa fever patients; a maculopapular rash is often noted in light-skinned patients. Effusions are common, and male-dominant pericarditis may develop late in infection. Maternal lethality is higher than the usual 15–30% and is especially increased during the last trimester. Fetal lethality reaches 90%. Excavation of the uterus may increase survival rates of pregnant women, but data on Lassa fever and pregnancy are still sparse. These figures suggest that interruption of the pregnancy of Lassa virus–infected women should be considered. White blood cell counts are normal or slightly elevated, and platelet counts are normal or somewhat low. Deafness coincides with clinical improvement in ~20% of patients and is permanent and bilateral in some patients. Reinfection may occur but has not been associated with severe disease.

High-level viremia or a high serum AST concentration statistically predicts a fatal outcome. Thus, patients with an AST concentration of >150 IU/mL should be treated with IV ribavirin. This antiviral nucleoside analogue appears to be partially effective in reducing lethality from that documented among retrospective controls. However, possible side effects, such as reversible anemia (which usually does not require transfusion), dependent hemolytic anemia, and bone marrow suppression, need to be kept in mind. Ribavirin should be given by slow IV infusion in a dose of 32 mg/kg; this dose should be followed by 16 mg/kg every 6 h for 4 days and then by 8 mg/kg every 8 h for 6 days. Inactivated Lassa virus vaccines failed in preclinical studies, but several promising vaccine platforms are currently under experimental evaluation.

HEMORRHAGIC FEVER WITH RENAL SYNDROME

HFRS is the most important VHF today, with more than 100,000 cases of severe disease in Asia annually and milder infections numbering in the thousands in Europe. The disease is widely distributed in Eurasia. The major causative viruses are Puumala virus (Europe), Dobrava-Belgrade virus (the Balkans), and Hantaan virus (Eastern Asia). Amur/Soochong, Go¯u, Kurkino, Muju, Saaremaa, Sochi, and Tula viruses also cause HFRS but much less frequently and in more geographically confined areas determined by the distribution of reservoir hosts. Seoul virus is exceptional in that it is associated with brown rats (*Rattus norvegicus*); therefore, the virus has a worldwide distribution because of the migration of these rodents on ships. Despite the wide distribution of Seoul virus, only mild or moderate HFRS occurs in Asia, and human disease has been difficult to identify in many areas of the world. Most cases of HFRS occur in rural residents or vacationers; the exception is Seoul virus infection, which may be acquired in an urban or rural setting or from contaminated laboratory-rat colonies. Classic Hantaan virus infection in Korea and in rural China is most common in the spring and fall and is related to rodent density and agricultural practices. Human infection is acquired primarily through aerosols of rodent urine, although virus is also present in rodent saliva and feces. Patients with HFRS are not infectious.

Severe cases of HFRS evolve in four identifiable stages. The *febrile stage* lasts 3 or 4 days and is identified by the abrupt onset of fever, headache, severe myalgia, thirst, anorexia, and often nausea and vomiting. Photophobia, retroorbital pain, and pain on ocular movement are common, and the vision may become blurred with ciliary body inflammation. Flushing over the face, the V area of the neck, and the back is characteristic, as are pharyngeal injection, periorbital edema, and conjunctival suffusion. Petechiae often develop in areas of pressure, the conjunctivae, and the axillae. Back pain and tenderness to percussion at the costovertebral angle reflect massive retroperitoneal edema. Laboratory evidence of mild to moderate DIC is present. Other laboratory findings of HFRS include proteinuria and active urinary sediment. The *hypotensive stage* lasts from a few hours to 48 h and begins with falling blood pressure and sometimes shock. The relative bradycardia typical of the febrile phase is replaced by tachycardia. Kinin activation is marked. The rising hematocrit reflects increasing vascular leakage. Leukocytosis with a left shift develops, and thrombocytopenia continues. Atypical lymphocytes—which in fact are activated CD8+ and, to a lesser extent, CD4+ T cells—circulate. Proteinuria is marked, and the urine's specific gravity falls to 1.010. Renal circulation is congested and compromised from local and systemic circulatory changes resulting in necrosis of tubules, particularly at the corticomedullary junction, and oliguria. During the *oliguric stage*, hemorrhagic tendencies continue, probably in large part because of uremic bleeding defects. Oliguria persists for 3–10 days before the return of renal function marks the onset of the *polyuric stage* (diuresis and hyposthenuria), which carries the danger of dehydration and electrolyte abnormalities.

Mild cases of HFRS may be much less stereotypical. The presentation may include only fever, gastrointestinal abnormalities, and transient oliguria followed by hyposthenuria. Infections with Puumala virus, the most common cause of HFRS in Europe (*nephropathia epidemica*), result in a much-attenuated picture but the same general presentation. Bleeding manifestations are found in only 10% of patients, hypotension rather than shock is usually documented, and oliguria is present in only about half of patients. The dominant features may be fever, abdominal pain, proteinuria, mild oliguria, and sometimes blurred vision or glaucoma followed by polyuria and hyposthenuria in recovery. Lethality is <1%.

HFRS should be suspected in patients with rural exposure in an endemic area. Prompt recognition of the disease permits rapid hospitalization and expectant management of shock and renal failure. Useful clinical laboratory parameters include leukocytosis, which may be leukemoid and is associated with a left shift; thrombocytopenia; and proteinuria. HFRS is readily diagnosed by an IgM-capture ELISA that is positive at admission or within 24–48 h thereafter. The isolation of hantaviruses is difficult, but RT-PCR of a blood clot collected early in the clinical course or of tissues obtained postmortem should give positive results. Such testing is usually undertaken if definitive identification of the infecting virus is required.

Mainstays of therapy are management of shock, reliance on vasopressors, modest crystalloid infusion, IV human serum albumin administration, treatment of renal failure with prompt dialysis to prevent overhydration that may result in pulmonary edema, and control of hypertension that increases the possibility of intracranial hemorrhage. Use of IV ribavirin has reduced lethality and morbidity in severe cases, provided treatment is begun within the first 4 days of illness. Lethality may be as high as 15%, but with proper therapy lethality should be <5%. Sequelae have not been definitively established.

CRIMEAN-CONGO HEMORRHAGIC FEVER (CCHF)

This severe VHF has a wide geographic distribution, potentially emerging wherever virus-bearing ticks occur. Because of the propensity of CCHF virus–transmitting ticks to feed on domestic livestock and certain wild mammals, veterinary serosurveys are the most effective mechanism for the monitoring of virus circulation in a particular region. Human infections are acquired via tick bites or during the crushing of infected ticks. Domestic animals do not become ill but do develop viremia. Thus, risk of acquiring CCHF occurs during sheep shearing, slaughter, and contact with infected hides or carcasses from recently slaughtered, infected animals. Nosocomial epidemics are common and are usually related to extensive blood exposure or needlesticks.

Although generally similar to other VHFs, CCHF causes extensive liver damage, resulting in jaundice in some patients. Clinical laboratory values indicate DIC and elevations in concentrations of AST, creatine phosphokinase, and bilirubin. Patients who do not survive generally have more distinct changes than survivors in the concentrations of these markers, even in the early days of illness, and also develop leukocytosis rather than leukopenia. In addition, thrombocytopenia is more marked and develops earlier in patients who do not survive than in survivors. The benefit of IV ribavirin for treatment remains hotly debated and unproven. Clinical experience and retrospective comparison of patients with ominous clinical laboratory values support a contention that ribavirin may be efficacious, but a randomized clinical trial was not supportive of a benefit in lowering lethality rates. No human or veterinary vaccines are recommended.

RIFT VALLEY FEVER

The natural range of Rift Valley fever virus was previously confined to sub-Saharan Africa, with circulation of the virus markedly enhanced by substantial rainfall. The El Niño Southern Oscillation phenomenon of 1997 facilitated subsequent spread of Rift Valley fever to the Arabian Peninsula, with epidemic disease in 2000. The virus has also been found in Madagascar and introduced into Egypt, where it caused major epidemics in 1977–1979, 1993, and thereafter. Rift Valley fever virus is maintained in nature by transovarial transmission in floodwater *Aedes* mosquitoes and presumably also has a vertebrate amplifier. Increased transmission during particularly heavy rains leads to epizootics characterized by high-level viremia in cattle, goats, or sheep. Numerous types of mosquitoes then feed on these animals and become infected, thereby increasing the possibility of human infections. Remote sensing via satellite can detect the ecologic changes associated with high rainfall that predict the likelihood of Rift Valley fever virus transmission. High-resolution satellites can also detect the special depressions in floodwaters from which the mosquitoes emerge. In addition, the virus can be transmitted by contact with blood or aerosols from domestic animals. Transmission risk is therefore high during birthing, and both abortuses and placentas need to be handled with caution. Slaughtered animals are not infectious because anaerobic glycolysis in postmortem tissues results in an acidic environment that rapidly inactivates bunyaviruses. Neither person-to-person nor nosocomial transmission of Rift Valley fever has been documented.

Rift Valley fever virus is unusual in that it causes several clinical syndromes. Most infections are manifested as the fever–myalgia syndrome. A small proportion of infections result in VHF with especially prominent liver involvement or encephalitis. Renal failure and DIC are also common features. Perhaps 10% of otherwise mild infections lead to retinal vasculitis, and some patients have permanently impaired vision. Funduscopic examination reveals edema, hemorrhages, and infarction of the retina as well as optic nerve degeneration. In a small proportion of patients (<1 in 200), retinal vasculitis is followed by viral encephalitis.

No proven therapy exists for Rift Valley fever. Both retinal disease and encephalitis occur after the acute febrile syndrome has resolved and serum neutralizing antibody has developed—events suggesting that only supportive care need be given. Epidemic disease is best prevented by vaccination of livestock. The ability of this virus to propagate after introduction into Egypt suggests that other potentially receptive areas, including the United States, should develop response plans. Rift Valley fever, like Venezuelan equine fever, is likely to be controlled only with adequate stocks of an effective live attenuated vaccine, but such global stocks are unavailable. A formalin-inactivated vaccine confers immunity in humans, but quantities are limited, and three injections are required. This vaccine is recommended for potentially exposed laboratory workers and for veterinarians working in sub-Saharan Africa. A new live attenuated vaccine, MP-12, is being tested in humans (phase 2 trials have been completed). The vaccine is safe and licensed for use in sheep and cattle. In addition, several vaccines are being developed specifically for use in animals.

SEVERE FEVER WITH THROMBOCYTOPENIA SYNDROME

This recently described tick-borne disease is caused by severe fever with thrombocytopenia syndrome virus. Numerous human infections have been reported during the past few years from China, and several cases have also been detected in Japan and South Korea. The clinical presentation ranges from mild nonspecific fever to severe VHF with a high (>12%) lethality.

Flaviviruses

The most important flaviviruses that cause VHF are the mosquito-borne dengue viruses 1–4 and yellow fever virus. These viruses are widely distributed and cause tens to hundreds of thousands of infections each year. Kyasanur Forest disease virus and Omsk hemorrhagic fever virus are geographically very restricted but important tick-borne flaviviruses that cause VHF, sometimes with subsequent viral encephalitis. Tick-borne encephalitis virus has caused VHF in a few patients. There is currently no therapy for these VHFs, but an inactivated vaccine has been used in India to prevent Kyasanur Forest disease.

SEVERE DENGUE

Several weeks after convalescence from infection with dengue virus 1, 2, 3, or 4, the transient protection conferred by that infection against reinfection with a heterotypic dengue virus usually wanes. Heterotypic reinfection may result in classic dengue or, less commonly, in severe dengue. In the past 20 years, *A. aegypti* mosquitoes have progressively reinvaded Latin America and other areas, and frequent travel by infected individuals has introduced multiple variants of dengue viruses 1–4 from many geographic areas. Thus, the pattern of hyperendemic transmission of multiple dengue virus serotypes established in the Americas and the Caribbean has led to the emergence of severe dengue as a major problem. Among the millions of dengue virus 1–4 infections, ~500,000 cases of severe dengue occur annually, with a lethality of ~2.5%. The induction of vascular permeability and shock depends on multiple factors, such as the presence or absence of enhancing and nonneutralizing antibodies, age (susceptibility to severe dengue drops considerably after 12 years of age), sex (females are more often affected than males), race (whites are more often affected than blacks), nutritional status (malnutrition is protective), and sequence of infections (e.g., dengue virus 1 infection followed by dengue virus 2 infection seems to be more dangerous than dengue virus 4 infection followed by dengue virus 2 infection). In addition, considerable heterogeneity exists among each dengue virus population. For instance, South-Eastern Asian dengue virus 2 variants have more potential to cause severe dengue than do other variants.

Severe dengue is identified by the detection of bleeding tendencies (tourniquet test, petechiae) or overt bleeding in the absence of underlying causes, such as preexisting gastrointestinal lesions. Shock may result from increased vascular permeability. In milder cases of severe dengue, restlessness, lethargy, thrombocytopenia (<100,000/μL), and hemoconcentration are detected 2–5 days after the onset of typical dengue, usually at the time of defervescence. The maculopapular rash that often develops in dengue may also appear in severe dengue. In more severe cases, frank shock is apparent, with low pulse pressure, cyanosis, hepatomegaly, pleural effusions, and ascites; in some patients, severe ecchymoses and gastrointestinal bleeding develop. The period of shock lasts only 1 or 2 days.

A virologic diagnosis of severe dengue can be made by the usual means. However, multiple flavivirus infections result in broad immune responses to several members of the genus, and this situation may result in a lack of virus specificity of the IgM and IgG immune responses. A secondary antibody response can be sought with tests against several flavivirus antigens to demonstrate the characteristic wide spectrum of reactivity.

Most patients with shock respond promptly to close monitoring, oxygen administration, and infusion of crystalloid or—in severe cases—colloid. Lethality varies greatly with case ascertainment and quality of treatment. However, most patients with severe dengue respond well to supportive therapy, and the overall lethality at an experienced center in the tropics is probably as low as 1%.

The key to control of both dengue and severe dengue is the control of *A. aegypti* mosquitoes, which also reduces the risk of urban yellow fever and chikungunya virus circulation. Control efforts have been handicapped by the presence of nondegradable tires and long-lived plastic containers in trash repositories (perfect mosquito breeding grounds when filled with water during rainfall) and by insecticide resistance. Urban poverty and an inability of the public health community to mobilize the populace to respond to the need to eliminate mosquito breeding sites are also factors in lack of mosquito control. A tetravalent live attenuated dengue vaccine based on the attenuated yellow fever virus 17D platform is under advanced development (phase 1 to phase 3 trials for various platforms in Latin America, Asia, and Australia). At least two live attenuated candidate vaccines based on modified recombinant dengue viruses have been evaluated in phase 1 clinical studies, but the results have not been promising.

YELLOW FEVER

Yellow fever virus had caused major epidemics in Africa and Europe before its transmission by *A. aegypti* mosquitoes was discovered in 1900. Urban yellow fever became established in the New World as a result of colonization with *A. aegypti*—originally an African mosquito. Subsequently, different types of mosquitoes and nonhuman primates were found to maintain yellow fever virus in Africa and also in Central and South American jungles. Transmission to humans is incidental, occurring via bites from mosquitoes that have fed on viremic monkeys. After the identification of *A. aegypti* mosquitoes as vectors of yellow fever, containment strategies were aimed at increased mosquito control. Today, urban yellow fever transmission occurs only in some African cities, but the threat exists in the great cities of South America, where reinfestation by *A. aegypti* mosquitoes has taken place, and dengue virus 1–4 transmission by the same mosquito is common. Despite the existence of a highly effective and safe vaccine, several hundred jungle yellow fever cases occur annually in South America, and 84,000–170,000 severe jungle and urban cases, including 29,000–60,000 deaths, occurred in 2013 in Africa.

Yellow fever is a typical VHF accompanied by prominent hepatic necrosis. A period of viremia, typically lasting 3 or 4 days, is followed by a period of “intoxication.” During the latter phase in severe cases, characteristic jaundice, hemorrhages, black vomit, anuria, and terminal delirium occur, perhaps related in part to extensive hepatic involvement. Blood leukocyte counts may be normal or reduced and are often high in terminal stages. Albuminuria is usually noted and may be marked. As renal function fails in terminal or severe cases, the concentration of blood urea nitrogen rises proportionately. Abnormalities detected in liver function tests range from modest elevations of AST concentrations in mild cases to severe derangement.

Urban yellow fever can be prevented by the control of *A. aegypti* mosquitoes. The continuing sylvatic cycles require vaccination of all visitors to areas of potential transmission with live attenuated variant 17D vaccine virus, which cannot be transmitted by mosquitoes. With few exceptions, reactions to the vaccine are minimal; immunity is provided within 10 days and lasts for at least 25–35 years. An egg allergy mandates caution in vaccine administration. Although there are no documented harmful effects of the vaccine on fetuses, pregnant women should be immunized only if they are definitely at risk of exposure to yellow fever virus. Because vaccination has been associated with several cases of encephalitis in children <6 months of age, it is contraindicated in this age group, nor is it recommended for infants 6–8 months of age unless the risk of exposure is very high. Rare, serious, multisystemic adverse reactions (occasionally fatal) have been reported, particularly affecting the elderly, and risk-to-benefit should be weighed prior to vaccine administration to individuals ≥60 years of age. Nevertheless, the number of deaths of unvaccinated travelers with yellow fever exceeds the number of deaths from vaccination, and a liberal vaccination policy for travelers to involved areas should be pursued. Timely information on changes in yellow fever distribution and yellow fever vaccine requirements can be obtained from the U.S. Centers for Disease Control and Prevention (*http://www.cdc.gov/vaccines/vpd-vac/yf/default.htm*).

ACKNOWLEDGMENT

The authors gratefully acknowledge the major contributions of Clarence J. Peters to this chapter in previous editions.

FURTHER READING

Centers for Disease Control and Prevention: Arbovirus catalog. Available at *https://wwwn.cdc.gov/arbocat/*. Accessed February 14, 2018.
King AMQ et al (eds): *Virus Taxonomy—Ninth Report of the International Committee on Taxonomy of Viruses*. London, Elsevier/Academic Press, 2011.
Knipe DM, Howley PM (eds): *Fields Virology*, 6th ed. Philadelphia, Wolters Kluwer/Lippincott Williams & Wilkins, 2013.

Lvov DK et al: *Zoonotic Viruses of Northern Eurasia: Taxonomy and Ecology*. London, Elsevier/Academic Press, 2015.
Singh SK, Ruzek D (eds): *Viral Hemorrhagic Fevers*. Boca Raton, FL, CRC Press, 2013.
CrossRef
Vasilakis N, Gubler DJ (eds): *Arboviruses: Molecular Biology, Evolution and Control*. Haverhill, UK, Caister Academic Press, 2016.

WEBSITE

International Committee on Taxonomy of Viruses (ICTV): *https://talk.ictvonline.org/*. Accessed February 14, 2018.

Chapter 205: Ebolavirus and Marburgvirus Infections

Jens H. Kuhn

INTRODUCTION

Several viruses of the family *Filoviridae* cause severe and frequently fatal infections in humans. Introduction of filoviruses into human populations is an extremely rare event that most likely occurs by direct or indirect contact with healthy filovirus hosts or by contact with infected, sick, or deceased mammals. Filoviruses are highly infectious but not especially contagious. Human-to-human transmission takes place through direct person-to-person (usually skin-to-skin) contact or exposure to infected bodily fluids and tissues; no evidence of such transmission by aerosol or respiratory droplets in natural outbreak settings is available. Infections progress rapidly from influenza-like to gastrointestinal manifestations and coagulopathy, typically culminating in multiple-organ dysfunction syndrome and shock. The occurrence of primary subclinical infections is controversial, but a small percentage of survivors may be subclinically and persistently infected. Treatment of filovirus infections is entirely supportive in nature because no specific efficacious antiviral agents or vaccines are yet licensed.

Filoviruses are categorized as World Health Organization (WHO) Risk Group 4 Pathogens. Consequently, all work with material suspected of containing replicating filoviruses should be conducted only in maximal containment (biosafety level 4) laboratories, or the viruses should be properly inactivated prior to further analysis in biosafety level 2 laboratories. Experienced personnel handling these viruses must wear appropriate personal protective equipment (see "Control and Prevention," below) and follow rigorous standard operating procedures. The proper national authorities and WHO reference laboratories should be contacted immediately when filovirus infections are suspected.

ETIOLOGY

The family *Filoviridae* includes three genera: *Cuevavirus*, *Ebolavirus*, and *Marburgvirus* **(Table 205-1 and Fig. 205-1)**. The available data suggest that the only known cuevavirus, Lloviu virus, does not infect humans, and that one ebolavirus, Reston virus, may infect but not cause disease in humans. The remaining four ebolaviruses—Bundibugyo virus, Ebola virus, Sudan virus, and Taï Forest virus—cause Ebola virus disease (EVD; International Classification of Disease, Tenth Revision [ICD-10], code A98.4). The two marburgviruses, Marburg virus and Ravn virus, are the etiologic agents of Marburg virus disease (MVD; ICD-10 code A98.3).

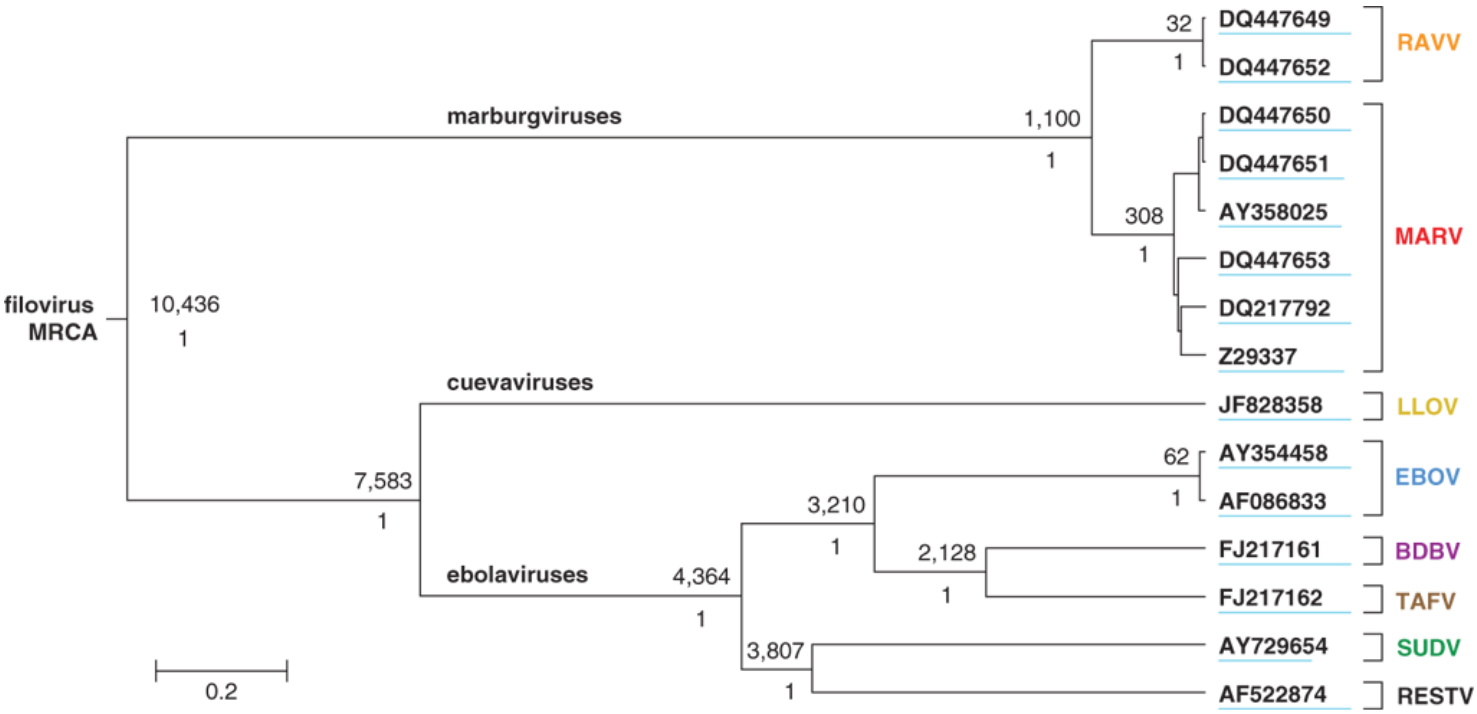
TABLE 205-1 Current Filovirus Taxonomy

Order *Mononegavirales*

- Family *Filoviridae*
  - Genus *Marburgvirus*
    - Species *Marburg marburgvirus*
      - Virus 1: Marburg virus (MARV)
      - Virus 2: Ravn virus (RAVV)
    - Genus *Ebolavirus*
      - Species *Bundibugyo ebolavirus*
        - Virus: Bundibugyo virus (BDBV)
      - Species *Reston ebolavirus*
        - Virus: Reston virus (RESTV)
      - Species *Sudan ebolavirus*
        - Virus: Sudan virus (SUDV)
      - Species *Tai Forest ebolavirus*
        - Virus: Taï Forest virus (TAFV)
      - Species *Zaire ebolavirus*
        - Virus: Ebola virus (EBOV)
    - Genus *Cuevavirus*
      - Species *Lloviu cuevavirus*
        - Virus: Lloviu virus (LLOV)

FIGURE 205-1

Filovirus phylogeny/evolution. Bayesian coalescent analysis of representative variants of all known filovirus clades (represented by underlined GenBank accession numbers). The maximal clade credibility tree is shown with the most recent common ancestor (MRCA) at each node. Posterior probability values are shown beneath MRCA estimates in years. Scale is in substitutions/site based on an analysis performed by Dr. Serena Carroll, US Centers for Disease Control and Prevention. BDBV, Bundibugyo virus; EBOV, Ebola virus; LLOV, Lloviu virus; MARV, Marburg virus; RAVV, Ravn virus; RESTV, Reston virus; SUDV, Sudan virus; TAFV, Taï Forest virus.

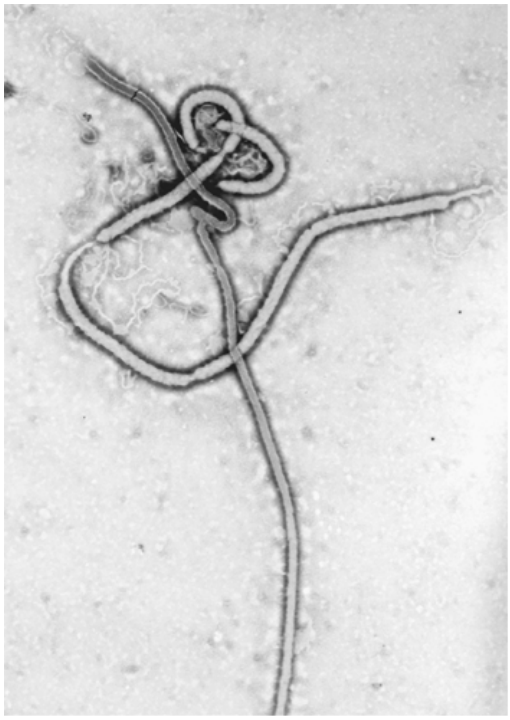


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Filoviruses have linear, nonsegmented, single-stranded, negative-sense RNA genomes that are ~19 kb in length. These genomes contain six or seven genes that encode the following seven structural proteins: nucleoprotein (NP), polymerase cofactor (VP35), matrix protein (VP40), glycoprotein ($GP_{1,2}$), transcriptional cofactor/viral protein 30 (VP30), nucleocapsid-associated protein (VP24), and RNA-dependent RNA polymerase (L). Cuevaviruses and ebolaviruses, but not marburgviruses, also encode three nonstructural proteins of unknown function (sGP, ssGP, and Δ-peptide). Filovirions are unique among human virus particles in that they are predominantly pleomorphic filaments but also assume torus- or 6-like shapes (width, ~80 nm; average length, ≥790 nm). These enveloped virions contain helical ribonucleocapsids and are covered with $GP_{1,2}$ spikes **(Fig. 205-2)**.

FIGURE 205-2

Ebola virus particle: The first transmission electron micrograph of an Ebola virion in a culture of grivet (*Chlorocebus aethiops*) Vero cells inoculated with a blood sample from a patient from the 1976 Zaire outbreak of Ebola virus disease. Shown is the typical and unique filamentous and pleomorphic structure of filovirions. *(PHIL ID#1833, taken by Dr. Fredrick A. Murphy, US Centers for Disease Control and Prevention.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

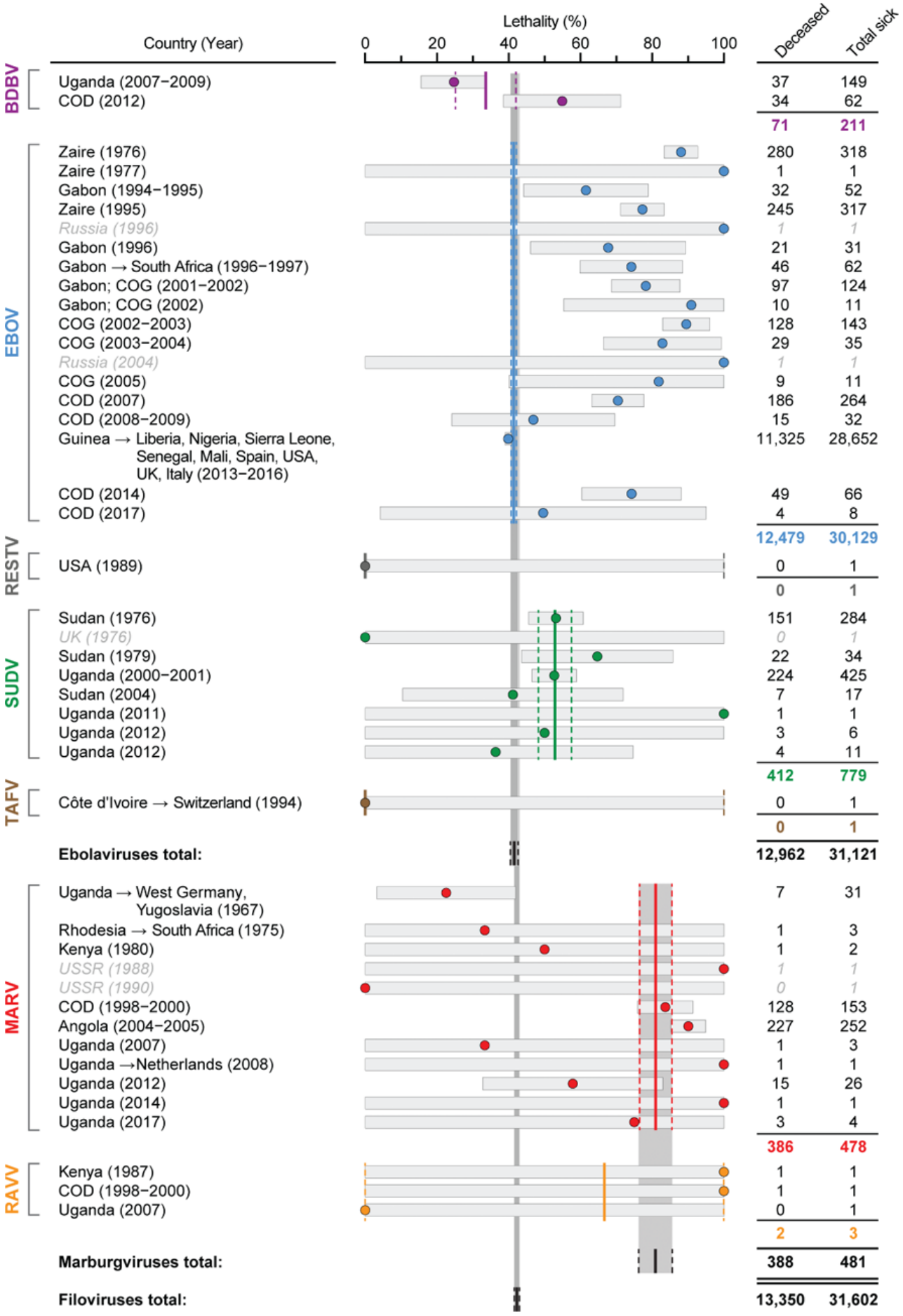
EPIDEMIOLOGY

As of October 1, 2017, a total of 31,602 human filovirus infections and 13,350 filovirus deaths had been recorded **(Fig. 205-3)**. Of those, 28,652 cases and 11,325 deaths occurred in a single outbreak, the 2013–2016 EVD outbreak in western Africa. The total numbers emphasize both the high degree of lethality (number of deaths per number of sick people; 41.7%) and the overall low mortality (impact on the healthy population) of filovirus infections. Until 2013, natural filovirus infections were not considered a global threat. However, the large 2013–2016 EVD outbreak challenged this assessment. Filoviruses pathogenic for humans appear to be exclusively endemic to equatorial (western, Middle, and Eastern) Africa **(Fig. 205-4)**, although this distribution may change if natural or artificial environmental alterations lead to filovirus host migration and increased contacts between nonhuman hosts and humans.

The majority of recorded EVD and MVD outbreaks, including the 2013–2016 EVD outbreak, can be traced back to single index cases who transmitted the infection to others. These chains of contacts suggest that only around 50 natural host-to-human spillover events have occurred since the discovery of filoviruses in 1967. Outbreak frequency, case numbers, and overall lethality probably depend on the particular etiologic agent, the geographic location and socioeconomic conditions of the affected country, and local customs. In particular, the accessibility of health-care centers and the availability of personal protective equipment and reusable medical equipment, such as syringes and needles, have affected overall case numbers in the past. Outbreaks have been contained when local burial practices, such as ritual washing, have been either prevented or altered by the use of gloves. The incidence of EVD and MVD may have increased over the past two decades (Figs. 205-3 and 205-4), but debate continues about whether the observed change is due to increased filovirus activity, more frequent contact between filovirus hosts and humans, or continuous improvement in surveillance capabilities.

FIGURE 205-3

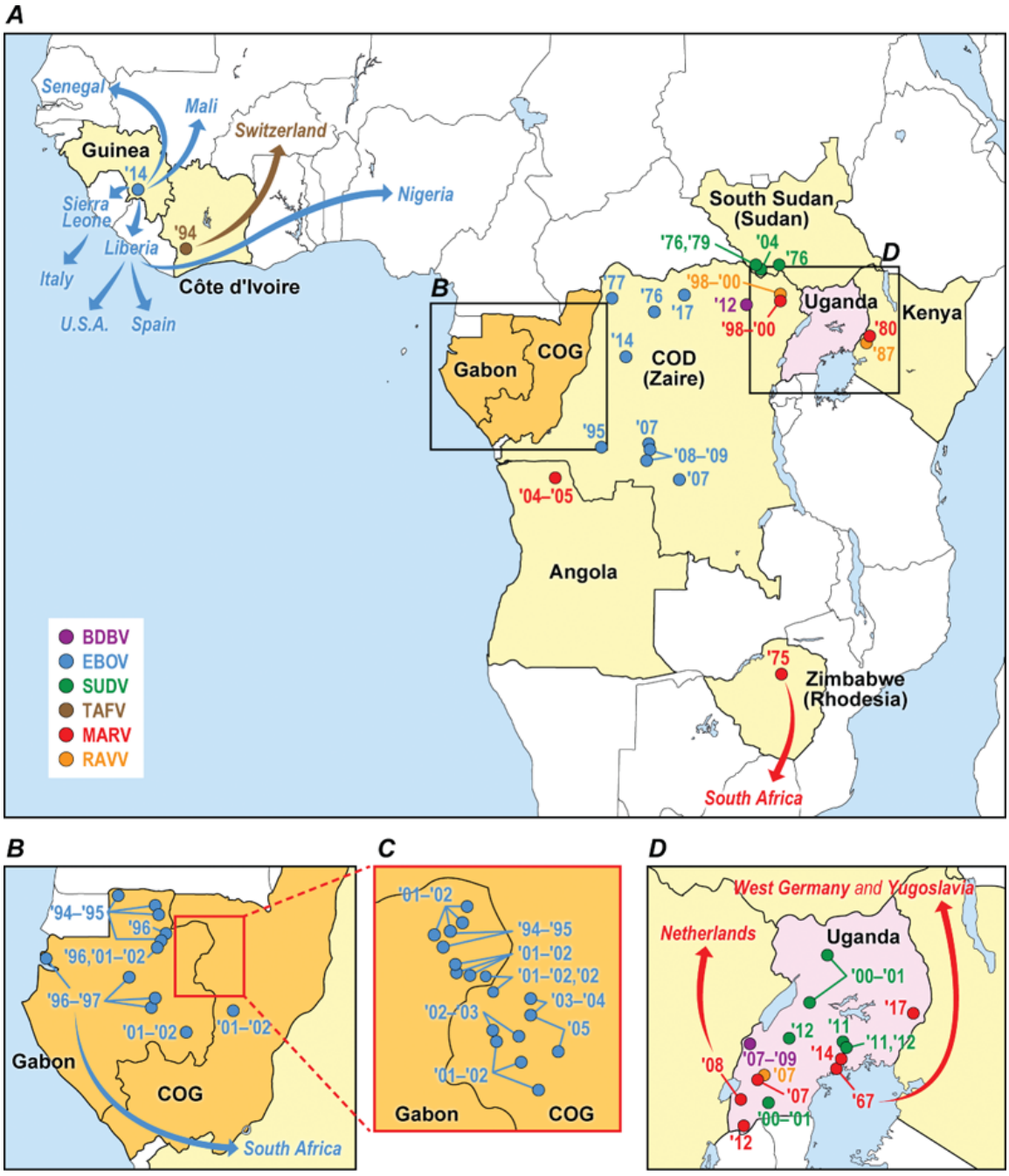
Characteristics of outbreaks of human filovirus disease. Seven of eight known filoviruses have caused infections in humans. Outbreaks are listed by virus in chronological order in the left column. Laboratory infections are *shaded gray and italicized*. *Arrows* indicate international case exportation. The total number of cases and lethal cases are summarized in the middle column. The lethality or case–fatality rate (*colored dots*) for each outbreak is plotted on a 0–100% scale along with 99% confidence intervals (*gray horizontal bars*). The overall case–fatality rate for disease caused by a particular virus is delineated by *vertical bold-colored lines*, with *vertical bold-colored dashed lines* indicating the corresponding 99% confidence intervals. The overall case–fatality rates for all ebolavirus infections, all marburgvirus infections, and all filovirus infections are shown by (overlapping) *vertical gray bars*. BDBV, Bundibugyo virus; COD, Democratic Republic of the Congo (formerly Zaire); COG, Republic of the Congo; EBOV, Ebola virus; MARV, Marburg virus; RAVV, Ravn virus; RESTV, Reston virus; SUDV, Sudan virus; TAFV, Taï Forest virus; UK, United Kingdom; USSR, Union of Soviet Socialist Republics (today Russia).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 205-4

Geographic distribution of human filovirus disease outbreaks and years of occurrence. *Arrows* indicate international case exportation. BDBV, Bundibugyo virus; COD, Democratic Republic of the Congo (formerly Zaire); COG, Republic of the Congo; EBOV, Ebola virus; MARV, Marburg virus; RAVV, Ravn virus; SUDV, Sudan virus; TAFV, Taï Forest virus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EVD and MVD outbreaks are associated with distinct meteorological and geographic conditions and are probably associated with distinct hosts or reservoirs. The four ebolaviruses that cause disease in humans appear to be endemic in humid rainforests. EVD outbreaks have often been associated with hunting or contact with bushmeat (i.e., meat from apes, other nonhuman primates, duikers, or bush pigs) in forests. Ecologic studies indicate that Ebola virus may play a role in extensive and frequently fatal epizootics among wild ape populations. However, replicating isolates of ebolaviruses from wild nonhuman primates thus far have not been obtained. The marburgviruses, Marburg virus and Ravn virus, on the other hand, seem to infect hosts inhabiting arid woodlands. MVD outbreaks have almost always been epidemiologically linked to visits to or work in natural or artificial caves or mines. A pteropodid (fruit) bat, the cave-dwelling Egyptian rousette (*Rousettus aegyptiacus*), serves as a natural and subclinically infected reservoir for both Marburg virus and Ravn virus. Although bats are suspected to be the hosts for ebolaviruses as well, definitive proof is lacking. In fact, thus far, only Ebola virus and Reston virus have been loosely connected to frugivorous and insectivorous bats by means of antibody or genome fragment detection, whereas the hosts of Bundibugyo virus, Sudan virus, and Taï Forest virus are enigmatic.

PATHOGENESIS

Human infections typically occur through direct exposure of skin lesions or mucosal surfaces to contaminated bodily fluids or material or by parenteral inoculation (e.g., via accidental needlesticks or reuse of needles in poorly equipped hospitals). Numerous studies, both in vitro and in vivo (in several animal models of human disease), have shed light on key pathogenetic events that evolve subsequent to filovirion exposure. The $GP_{1,2}$ spikes on the surface of filovirions determine their cell and tissue tropism by engaging yet-unidentified cell-surface molecules and the intracellular receptor Niemann-Pick C1.

One of the pathogenetic hallmarks of filovirus infection is a pronounced modulation of the immune system. The first targets of filovirions are local macrophages, monocytes, and dendritic cells. Several structural proteins of filovirions (i.e., VP35, VP40, and/or VP24) then suppress intrinsic and innate immune responses by, for instance, inhibiting the interferon pathways and enabling a productive filovirus infection. The result is the secretion of copious numbers of progeny virions, as evidenced by high titers in the bloodstream (>10^6 plaque-forming units [pfu]/mL of serum in humans) and the lymphatics, and dissemination to most tissues. Filovirions then infect additional phagocytic cells, including other macrophages (alveolar, peritoneal, and pleural macrophages; Kupffer cells in the liver; and microglia). Other targets, such as adrenal cortical cells, fibroblasts, hepatocytes, endothelial cells, and a variety of epithelial cells, are also infected. Infection leads to the secretion of soluble signaling molecules (varying with the cell type) that most likely are crucial factors in immune response modulation and development of multiorgan dysfunction syndrome. For instance, infected macrophages react by secreting proinflammatory cytokines, a response that leads to further recruitment of macrophages to the site of infection. In contrast, infected dendritic cells are not activated to secrete cytokines, and expression of major histocompatibility class II antigens is partially suppressed. Immunosuppression occurs in part by massive lymphoid depletion in lymph nodes, spleen, and thymus in the absence of reactive inflammatory cellular responses. Results from animal studies suggest that depletion is a direct consequence of considerable bystander apoptosis of lymphocytes; this explanation would also account for the severe lymphopenia that develops in patients. The consequence of these events is not only florid filovirus dissemination but also a proclivity of the patient for secondary bacterial and fungal infections.

Other pathogenetic hallmarks of filovirus infections are a severe disturbance of the clotting system and the impairment of vascular integrity. Disseminated intravascular coagulation is the cause of the severe imbalance in the clotting system of filovirus-infected patients. Thrombocytopenia, increased concentrations of tissue factor, consumption of clotting factors, increased concentrations of fibrin degradation products (D-dimers), and declining concentrations of protein C are typical features of infection. Consequently, the occlusion of small vessels by widely distributed microthrombi leads to extensive necroses/hypoxic infarcts in target tissues (particularly the gonads, kidneys, liver, and spleen) in the absence of marked inflammatory responses. In addition, petechiae, ecchymoses, extensive visceral effusions, and other hemorrhagic signs are observed in internal organs, mucous membranes, and skin. Actual severe blood loss, however, is a rare event (although it frequently occurs during or after childbirth). Aberrance in cytokines or other factors such as nitric oxide and direct infection and activation of endothelial cells most likely are responsible for upregulated permeability of blood-vessel endothelia. This upregulation leads to fluid redistribution (*third spacing*); interstitial and myocardial edema and hypovolemic shock are common developments. Clinical improvement is possible and is usually characterized by falling viral titers during the development of a virus-specific immune response.

CLINICAL MANIFESTATIONS

MVD and EVD cannot be differentiated by mere observation of clinical manifestations and for all practical purposes may be considered the same disease. The incidence of clinical signs does not differ significantly among infections caused by disparate filoviruses **(Table 205-2)**, although, apart from the patients in the 2013–2016 EVD outbreak, the numbers of thoroughly observed patients are very low. The incubation period ranges from 3 to 25 days, after which infected people develop a biphasic syndrome with a 1- to 2-day relative remission separating the two phases. The first phase (disease onset until around day 5–7) resembles influenza and is characterized by sudden onset of fever and chills, severe headaches, cough, myalgia, pharyngitis, arthralgia of the larger joints, development of a maculopapular rash, and other signs/symptoms (Table 205-2). The second phase (~5–7 days after disease onset and thereafter) involves the gastrointestinal tract (abdominal pain with vomiting and/or diarrhea), respiratory tract (chest pain, cough), vascular system (postural hypotension, edema), and central nervous system (confusion, coma, headache). Hemorrhagic manifestations such as subconjunctival injection, epistaxis, hematemesis, hematuria, and melena are typical (Table 205-2). Relapses are extremely rare, but, when they do occur, their course resembles that of the primary disease.

TABLE 205-2 Distribution of Clinical Signs/Symptoms of Filovirus-Infected Patients in Three Representative Outbreaks

Sign/Symptom	Frequency (%) Among Survivors			Frequency (%) Among Fatal Cases		
	BDBV (2007–2009)	EBOV (1995)[a]	MARV (1998–2000)	BDBV (2007–2009)	EBOV (1995)[a]	MARV (1998–2000)
Abdominal pain	88	68	59	93	62	57
Abortion	NR	5	NR	NR	2	NR
Anorexia	83	47	77	80	43	72
Anuria	NR	0	NR	NR	7	NR
Arthralgia or myalgia	83	79	55	86	50	55
Asthenia	NR	95	NR	NR	85	NR
Bleeding from puncture sites	NR	5	0	NR	8	7
Bleeding from the gums	NR	0	23	NR	15	36
Bleeding from any site	NR	NR	59	NR	NR	71
Bloody stools	NR	5	NR	NR	7	NR
Chest pain	NR	5	18	NR	10	4
Conjunctival injection	NR	47	14	NR	42	42
Convulsions	NR	0	NR	NR	2	NR
Cough	NR	26	9	NR	7	5
Diarrhea	92	84	59	87	86	56
Difficulty breathing	26	NR	36	57	NR	58
Dysesthesia	NR	5	NR	NR	0	NR
Epistaxis	NR	0	18	NR	2	34
Fever	100	95	100	100	93	92
Headaches	84	74	73	93	52	79
Hearing loss	NR	11	NR	NR	5	NR
Hematemesis	NR	0	68	NR	13	76
Hematoma	NR	0	0	NR	2	3
Hematuria	NR	16	NR	NR	7	NR
Hemoptysis	NR	11	9	NR	0	4
Hepatomegaly (without jaundice)	NR	5	NR	NR	2	NR
Hiccups	17	5	18	40	17	44
Lumbar pain	NR	26	5	NR	12	8
Maculopapular rash	35	16	NR	33	14	NR
Malaise or fatigue	96	NR	86	100	NR	83
Melena	NR	16	41	NR	8	58
Nausea and vomiting	92	68	77	87	73	76
Petechiae	NR	0	9	NR	8	7
Sore throat, odynophagia, or dysphagia	43	58	43	60	56	43
Splenomegaly	NR	5	NR	NR	2	NR
Tachypnea	NR	0	NR	NR	31	NR
Tinnitus	NR	11	NR	NR	1	NR

[a]In contrast to the 1995 outbreak (317 cases), the 2013–2016 EVD outbreak (28,652 cases) would yield much more statistically robust numbers. However, to date, only very few reports have been published on patient cohorts larger than a few hundred cases. Metastudies comparing these reports are largely absent or focus on very few symptoms or signs.

Abbreviations: BDBV, Bundibugyo virus; EBOV, Ebola virus; MARV, Marburg virus; NR, not reported.

Typical laboratory findings are leukopenia (with cell counts as low as 1000/μL) with a left shift prior to leukocytosis, thrombocytopenia (with counts as low as 50,000/μL), increased concentrations of liver and pancreatic enzymes (aspartate aminotransferase > alanine aminotransferase, γ-glutamyltransferase, serum amylase), hypokalemia, hypoproteinemia, increased creatinine and urea concentrations with proteinuria, and prolonged prothrombin and partial thromboplastin times.

Patients usually succumb to disease 4–14 days after infection. Patients who survive often report prolonged and sometimes incapacitating arthralgia, asthenia, iridocyclitis, hearing loss, myalgia, orchitis, parotitis, psychosis, recurrent hepatitis, transverse myelitis, or uveitis, but clinical research is still ongoing to ascertain that these are true sequelae. Temporary hair loss and desquamation of skin areas previously affected by a typical maculopapular rash are visible consequences of the disease. Rarely, filoviruses can persist in the brain, eyes, liver, or testicles of survivors and may cause recurrent disease after convalescence and/or sexual transmission.

DIAGNOSIS

Filovirus infections cannot be diagnosed on the basis of clinical presentation alone. Numerous diseases typical for equatorial Africa need to be considered in the differential diagnosis of a febrile patient. Almost all of these diseases occur at a much higher incidence than filovirus infections and are therefore the more likely candidates during differential diagnostic deliberations. The most important of the infectious diseases that closely mimic EVD and MVD are falciparum malaria and typhoid fever; also important are enterohemorrhagic *Escherichia coli* enteritis,

gram-negative septicemia (including shigellosis), meningococcal septicemia, rickettsial infections, fulminant viral hepatitis, leptospirosis, measles, and all other viral hemorrhagic fevers (in particular, Lassa and yellow fevers). Other ailments, such as venomous snakebites, warfarin intoxication, and the many transient or inherited platelet and vascular disorders, also must be considered. Visits to caves or mines and direct contact with bats, nonhuman primates, or bushmeat should raise suspicion of filovirus infection, as should admission to or treatment in rural hospitals or direct contact with severely ill local residents.

If EVD or MVD is suspected on the basis of epidemiologic history, exposure history, and/or clinical manifestations, infectious disease specialists and the proper public health authorities, including the WHO, should be notified immediately. Laboratory diagnosis of EVD and MVD is relatively straightforward but ideally requires maximal containment (biosafety level 4), which usually is not available in filovirus-endemic countries. Alternatively, laboratory diagnosis is performed using inactivated samples in lower-containment settings by on-site personnel trained in the use of diagnostic assays adapted for field use. Consequently, diagnostic samples should be collected with great caution and with use of proper personal protective equipment and strict barrier nursing techniques. With adherence to established biosafety precautionary measures, samples should be sent in suitable transport media to national or international WHO reference laboratories. Acute-phase blood/serum is the preferred diagnostic specimen because it usually contains high titers of filovirions and filovirion-specific antibodies.

The current methods of choice for the diagnosis of filovirus infection are reverse-transcription polymerase chain reaction (typical detection limit, 1000–5000 pfu per milliliter of serum, depending on the assay) and antigen capture enzyme-linked immunosorbent assay (ELISA) for the detection of filovirus genomes and filovirion proteins, respectively. Direct IgM and IgG–capture or IgM-capture ELISA is used for the detection of filovirion-targeting antibodies from patients in later stages of disease—i.e., those who have been able to mount a detectable immune response, including survivors. All these assays can be conducted on samples treated with guanidinium isothiocyanate (for polymerase chain reaction) or cobalt-60 irradiation (for ELISA) or subjected to other effective measures that render filoviruses noninfectious. Virus isolation in cell culture and plaque assays for quantification or diagnostic confirmation are relatively easy but must be performed in maximal-containment laboratories. If available, electron microscopic examination of properly inactivated samples or cultures can further support the diagnosis because filovirions have unique filamentous shapes (Fig. 205-2). Formalin-fixed skin biopsies and possibly skin swabs can be useful for safe postmortem diagnoses.

TREATMENT

TREATMENT Filovirus Infections

Any treatment of patients with suspected or confirmed filovirus infection must be administered under increased safety precautions by experienced specialists using appropriate personal protective equipment (see "Control and Prevention," below). Treatment of EVD and MVD is entirely supportive because no accepted/approved, efficacious, specific antiviral agents or vaccines are yet licensed. Exceptions are hyperimmune equine immunoglobulin, which has been approved in Russia for emergency treatment of laboratory infections, and the anti–Ebola virus monoclonal antibody cocktail ZMapp, which is on its way to becoming available under U.S. Emergency Use Authorization. However, convincing efficacy data are still missing for both medical countermeasures. Given the high lethality of filoviruses, special protocols may be established by ad hoc expert groups to outline treatment of exposed individuals with one of several regimens that have shown promise in experimental nonhuman primate models. Current options include post-exposure vaccination with filovirus $GP_{1,2}$-expressing recombinant replicating vesicular stomatitis Indiana virus or administration of filovirus-specific antibodies or antibody cocktails (convalescent sera have not yet been proven effective), synthetic adenosine analogs (galidesivir/BCX4430, GS-5734) that act as non-obligate RNA chain terminators, or favipiravir. Regardless of the availability of these experimental agents, measures to stabilize patients include those generally recommended for severe septicemia/sepsis/shock **(Chap. 297)** and should be applied with an emphasis on fluid and electrolyte replacement. Countermeasures should address hypotension and hypoperfusion, vascular leakage in the systemic and pulmonary circulatory system, disseminated intravascular coagulation and overt hemorrhaging, acute kidney failure, and electrolyte (especially potassium) imbalances. Pain management and administration of antipyretics, antiemetics, and antidiarrheal agents should be considered. Aggressive supportive measures, including mechanical ventilatory support and renal replacement therapy, may shore up patients with severe EVD until their immune systems respond and virus is cleared.

COMPLICATIONS

Secondary infections should be kept in mind and appropriately treated as early as possible. Pregnancy and labor cause severe and frequently fatal complications in filovirus infections due to clotting factor consumption, fetal loss, and/or severe blood loss during birth.

PROGNOSIS

The prognosis of filovirus infections is generally poor, although outcome probably depends somewhat on which particular virus causes the infection (Fig. 205-3). Convalescence may take months, with skin peeling, alopecia, prostration, weight loss, orchitis, amnesia, confusion, and anxiety as typical sequelae. Rarely, filoviruses persist in apparently healthy survivors and are either reactivated by unknown means at a later point or transmitted sexually. Abstinence from sexual activity for at least 12 months after disappearance of clinical signs is recommended for survivors unless testing proves semen to be free of filoviruses. The use of condoms is generally recommended for all sexual activities.

CONTROL AND PREVENTION

Currently, licensed filovirus vaccines are not available. Prevention of filovirus infection in nature is difficult because the ecology of the viruses is not completely understood. At present, to prevent marburgvirus infection, avoidance of direct or indirect contact with Egyptian rousettes is the most useful advice to people entering or living in areas where these animals can be found. Prevention seems to be more difficult in the case of ebolaviruses, for which definite reservoirs have not yet been pinpointed. EVD outbreaks have been associated not so much with bats as with hunting or consumption of nonhuman primates. The mechanism of introduction of ebolaviruses into nonhuman primate populations is unclear. Therefore, the best advice to locals and travelers is to avoid contact with bushmeat, nonhuman primates, and bats.

Relatively simple barrier nursing techniques, vigilant use of proper personal protective equipment, and quarantine measures (including contact tracing) usually suffice to terminate or at least contain filovirus disease outbreaks. Isolation of filovirus-infected people and their contacts and

avoidance of direct person-to-person contact without proper personal protective equipment usually suffice to prevent further spread as the pathogens are not transmitted through droplets or aerosols under natural conditions. Typical protective gear sufficient to prevent filovirus infections consists of disposable gloves, gowns, and shoe covers and a face shield and/or goggles. If available, N-95 or N-100 respirators may be used to further limit infection risk. Positive air pressure respirators should be considered for high-risk medical procedures such as intubation or suctioning. Medical equipment used in the care of a filovirus-infected patient, such as gloves or syringes, should never be reused. Because filovirions are enveloped, disinfection with detergents, such as 1% sodium deoxycholate, diethyl ether, or phenolic compounds, is relatively straightforward. Bleach solutions of 1:100 or 1:10 are recommended for surface disinfection and application to excreta or corpses, respectively. Whenever possible, potentially contaminated materials should be autoclaved, irradiated, or destroyed.

FURTHER READING

Burk R et al: Neglected filoviruses. FEMS Microbiol Rev 40:494, 2016. [PubMed: 27268907]
Caviness K et al: Ebola virus persistence as a new focus in clinical research. Curr Opin Virol 23:43, 2017. [PubMed: 28340374]
Cnops L et al: Essentials of filoviral load quantification. Lancet Infect Dis 16:e134, 2016. [PubMed: 27296694]
Cross RW et al: Post-exposure treatments for Ebola and Marburg virus infections. Nat Rev Drug Discov, 2018 (in press).
Dudas G et al: Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544:309, 2017. [PubMed: 28405027]
Feldmann H et al: *Filoviridae*: Marburg and Ebola viruses, in *Fields Virology*, 6th ed, vol 1. DM Knipe, PM Howley (eds). Philadelphia, Wolters Kluwer/Lippincott Williams & Wilkins, 2013, pp 923–956.
Kuhn JH: *Filoviruses: A Compendium of 40 Years of Epidemiological, Clinical, and Laboratory Studies*. Arch Virol Supplementa, vol 20. New York, SpringerWien, 2008.
Kuhn JH et al: Family *Filoviridae*, in *Virus Taxonomy—Ninth Report of the International Committee on Taxonomy of Viruses*. AMQ King et al (eds). London, Elsevier/Academic Press, 2011, pp 665–671.
Martins KA et al: Ebola virus disease candidate vaccines under evaluation in clinical trials. Expert Rev Vaccines 15:1101, 2016. [PubMed: 27160784]
Prescott JB et al: Immunobiology of Ebola and Lassa virus infections. Nat Rev Immunol 17:195, 2017. [PubMed: 28111475]

Chapter 206: Diagnosis and Treatment of Fungal Infections

John E. Edwards, Jr.

TERMINOLOGY AND MICROBIOLOGY

Traditionally, fungal infections have been classified into specific categories based on both anatomic location and epidemiology. The most common general anatomic categories are mucocutaneous and deep organ infection; the most common general epidemiologic categories are endemic and opportunistic infection. Although *mucocutaneous infections* can cause serious morbidity, they are rarely fatal. *Deep organ infections* also cause severe illness in many cases and, in contrast to mucocutaneous infections, are often fatal. The *endemic mycoses* (e.g., coccidioidomycosis) are caused by fungal organisms that are not part of the normal human microbiota, but rather are acquired from environmental sources. In contrast, *opportunistic mycoses* are caused by organisms (e.g., *Candida* and *Aspergillus*) that commonly are components of the normal human microbiota and whose ubiquity in nature renders them easily acquired by the immunocompromised host **(Table 206-1)**. Opportunistic fungi cause serious infections when the immunologic response of the host becomes ineffective, allowing the organisms to transition from harmless commensals to invasive pathogens. Frequently, the diminished effectiveness of the immune system is a result of advanced modern therapies that coincidentally either cause an imbalance in the host's microbiota or directly interfere with immunologic responses. Endemic mycoses usually cause more severe illness in immunocompromised patients than in immunocompetent individuals.

TABLE 206-1 Endemic and Opportunistic Mycoses

Endemic Mycoses[a]	Opportunistic Mycoses
Coccidioidomycosis	Candidiasis
Histoplasmosis	Aspergillosis
Blastomycosis	Cryptococcosis
Phaeohyphomycosis	Mucormycosis (zygomycosis)
Penicilliosis	Scedosporiosis
Sporotrichosis	Trichosporonosis
Paracoccidioidomycosis	Fusariosis
Adiaspiromycosis	Pneumocystosis

[a]The endemic mycoses can also occur as opportunistic infections.

Patients acquire deep organ infection with endemic fungi almost exclusively by inhalation. Cutaneous infections result either from hematogenous dissemination or, more often, from direct contact with soil—the natural reservoir for the vast majority of endemic mycoses. The dermatophytic fungi may be acquired by human-to-human transmission, but the majority of infections result from environmental contact. In contrast, the opportunistic fungus *Candida* invades the host from normal sites of colonization, usually the mucous membranes of the gastrointestinal tract. In general, innate immunity

is the primary defense mechanism against fungi. Although antibodies are formed during many fungal infections (and even during commensalism), they generally do not constitute the primary mode of host defense. Nevertheless, in selected infections, as discussed below, measurement of antibody titers may be a useful diagnostic test.

Three other terms frequently used in clinical discussions of fungal infections are *yeast*, *mold*, and *dimorphic fungus*. *Yeasts* are seen as rounded single cells or as budding organisms. *Molds* grow as filamentous forms called *hyphae* both at room temperature and in invaded tissue. *Aspergillus*, *Rhizopus* (the genus that causes mucormycosis, also known as zygomycosis), and fungi commonly infecting the skin to cause ringworm and related cutaneous conditions are classified as molds. Variations occur within this classification of yeasts and molds. For instance, when *Candida* infects tissue, both yeasts and filamentous forms may be present (except with *Candida glabrata*, which forms only yeasts in tissue); in contrast, *Cryptococcus* exists only in yeast form. *Dimorphic* is the term used to describe fungi that grow as yeasts or large spherical structures in tissue but as filamentous forms at room temperature in the environment. Classified in this group are the organisms causing candidiasis, blastomycosis, paracoccidioidomycosis, coccidioidomycosis, histoplasmosis, and sporotrichosis as well as *Emonsia* and *Ustilago*.

The incidence of nearly all fungal infections has risen substantially. Opportunistic infections have increased in frequency as a consequence of intentional immunosuppression in organ and stem cell transplantation and other disorders, the administration of cytotoxic chemotherapy for cancers, the liberal use of antibacterial agents, and, more recently, the increasing use of monoclonal antibodies.

 Within a global context, the incidence of endemic mycoses has increased in geographic locations where there has been substantial population growth. When advances in medical care (e.g., more aggressive treatment of cancer or organ transplantation) are introduced into a given area, the opportunistic mycoses increase in incidence.

DIAGNOSIS

The definitive diagnosis of any fungal infection requires histopathologic identification of the fungus invading tissue and accompanying evidence of an inflammatory response. The identification of an inflammatory response has been especially important with regard to *Aspergillus* infection. *Aspergillus* is ubiquitous and can float in the air onto biopsy material. Therefore, in rare but important instances, this fungus is an ex vivo contaminant during processing of a specimen for microscopy, with a consequent incorrect diagnosis. The stains most commonly used to identify fungi are periodic acid–Schiff and Gomori methenamine silver. *Candida*, unlike other fungi, is visible on gram-stained tissue smears. Hematoxylin and eosin stain is not sufficient to identify *Candida* in tissue specimens. When positive, an india ink preparation of cerebrospinal fluid (CSF) is diagnostic for cryptococcosis. Most laboratories now use calcofluor white staining coupled with fluorescent microscopy to identify fungi in fluid specimens. Matrix-assisted laser desorption ionization–time of flight mass spectometry (MALDI-TOF MS) will likely be used extensively for detection and speciation in the future. Point-of-care, lateral-flow testing techniques are being developed for many fungal infections.

Extensive investigations of the diagnosis of deep organ fungal infections have yielded a variety of tests with different degrees of specificity and sensitivity. The most reliable tests are the detection of antibody to *Coccidioides immitis* in serum and CSF; of *Histoplasma capsulatum* antigen in urine, serum, and CSF; and of cryptococcal polysaccharide antigen in serum and CSF. Commercially available antibody detection systems can now be used for blastomycosis, histoplasmosis, cryptococcosis, aspergillosis, and β-glucan detection. These tests have a general sensitivity and specificity of 90%; however, because of variability among laboratories, testing on multiple occasions is advisable. The test for galactomannan has been used extensively in Europe and is now approved in the United States for diagnosis of aspergillosis. Sources of concern regarding galactomannan are the

incidence of false-negative results and the need for multiple serial tests to reduce this incidence. This test is most useful when applied to bronchoalveolar lavage fluid. The β-glucan test for *Candida* is also under evaluation but, like the galactomannan test, requires additional validation; this test has a negative predictive value of ~90%. Both of these tests are being used with increasing frequency, especially for guiding the timing of initiation and the duration of therapy and for following principles of antimicrobial stewardship. One of the most useful applications of these non-culture-based tests has been the detection of *Histoplasma* antigens in serum and urine. T2 magnetic resonance is also being extensively evaluated for *Candida* and other organisms. It has been approved by the U.S. Food and Drug Administration (FDA) for detection of *Candida* in serum. Numerous polymerase chain reaction assays to detect antigens are in the developmental stages, as are nucleic acid hybridization techniques; currently, these tests are not widely available.

Of the fungal organisms, *Candida* is by far the most frequently recovered from blood. Although *Candida* species can be detected with any of the automated blood culture systems widely used at present, the lysis–centrifugation technique increases the sensitivity of blood cultures for *Candida* and for less common organisms (e.g., *H. capsulatum*). Lysis–centrifugation should be used when disseminated fungal infection is suspected.

Except in the cases of coccidioidomycosis, cryptococcosis, and histoplasmosis, there are no fully validated and widely used tests for serodiagnosis of disseminated fungal infection. Skin tests for the endemic mycoses are no longer available.

TREATMENT

TREATMENT Fungal Infections

This discussion is intended as a brief overview of general strategies for the use of antifungal agents in the treatment of fungal infections. Regimens, schedules, and choices are detailed in the chapters on specific mycoses that follow in this section. The doses cited here are standard doses for adults with invasive infection.

Since fungal organisms are eukaryotic cells that contain most of the same organelles (with many of the same physiologic functions) as human cells, the identification of drugs that selectively kill or inhibit fungi, but are not toxic to human cells, has been highly problematic. Far fewer antifungal than antibacterial agents have been introduced into clinical medicine.

AMPHOTERICIN B

The introduction of amphotericin B (AmB) in the late 1950s revolutionized the treatment of fungal infections in deep organs. Before AmB became available, cryptococcal meningitis and other disseminated fungal infections were nearly always fatal. For nearly a decade after AmB was introduced, it was the only effective agent for the treatment of life-threatening fungal infections. AmB remains the broadest-spectrum antifungal agent but carries several disadvantages, including significant nephrotoxicity, lack of an oral preparation, and unpleasant side effects (fever, chills, and nausea) during treatment. To circumvent nephrotoxicity and infusion side effects, lipid formulations of AmB were developed and have virtually replaced the original colloidal deoxycholate formulation in clinical use (although the older formulation is still available). The lipid formulations include liposomal AmB (L-AmB; 3–5 mg/kg per day) and AmB lipid complex (ABLC; 5 mg/kg per day). A third preparation, AmB colloidal dispersion (ABCD; 3–4 mg/kg per day), is rarely used because of the high incidence of side effects associated with infusion.

The lipid formulations of AmB have the disadvantage of being considerably more expensive than the deoxycholate formulation. Experience is still accumulating on the comparative efficacy, toxicity, and advantages of the different formulations for specific clinical fungal infections, including central nervous

system (CNS) infection. Whether there is a clinically significant difference in these drugs with respect to CNS penetration or nephrotoxicity remains controversial. Despite these issues and despite the expense, the lipid formulations are now much more commonly used than AmB deoxycholate in developed countries. In developing countries, AmB deoxycholate is still preferred because of the expense of the lipid formulations.

AZOLES

This class of antifungal drugs offers important advantages over AmB: the azoles cause little or no nephrotoxicity and are available in oral formulations. Early azoles included ketoconazole and miconazole, which have been replaced by newer agents for the treatment of deep organ fungal infections. The azoles' mechanism of action is inhibition of ergosterol synthesis in the fungal cell wall. Unlike AmB, these drugs are considered fungistatic, not fungicidal.

Fluconazole

Since its introduction, fluconazole has played an extremely important role in the treatment of a wide variety of serious fungal infections. Its major advantages are the availability of both oral and IV formulations, a long half-life, satisfactory penetration of most body fluids (including ocular fluid and CSF), and minimal toxicity (especially relative to that of AmB). Its disadvantages include (usually reversible) hepatotoxicity and—at high doses—alopecia, muscle weakness, and dry mouth with a metallic taste. Fluconazole is not effective for the treatment of aspergillosis, mucormycosis, or *Scedosporium apiospermum* infections. It is less effective than the newer azoles against *C. glabrata* and *Candida krusei.*

Fluconazole has become the agent of choice for the treatment of coccidioidal meningitis, although relapses have followed therapy with this drug. In addition, fluconazole is useful as both consolidation and maintenance therapy for cryptococcal meningitis. This agent has been shown to be as efficacious as AmB in the treatment of candidemia. The effectiveness of fluconazole and the echinocandins (see below) in candidemia and the drugs' relatively minimal toxicity, in conjunction with the inadequacy of diagnostic tests for widespread hematogenously disseminated candidiasis, have led to a change in the paradigm for candidemia management. The standard of care is now to treat all candidemic patients with an antifungal agent and to change all their intravascular lines, if feasible, rather than merely removing a singular suspect intravascular line and then observing the patient. The usual fluconazole regimen for treatment of candidemia is a loading dose of 12 mg/kg on day 1 and then 6 mg/kg per day until 2 weeks after the last positive blood culture.

Fluconazole is considered effective as fungal prophylaxis in bone marrow transplant recipients and high-risk liver transplant patients. Its general use for prophylaxis in patients with leukemia, in AIDS patients with low CD4+ T cell counts, and in patients on surgical intensive care units remains controversial. Many centers are now using posaconazole for prophylaxis in neutropenic patients (see below). Because of concern about the possibility of infection due to resistant *Candida* species and of infection with *Aspergillus* species in neutropenic patients, many clinicians are initiating therapy with an echinocandin, which is replaced by fluconazole once a susceptible *Candida* species is recovered and concern about *Aspergillus* is diminished. In a recent report, even low doses of fluconazole used for the treatment of vulvovaginal candidiasis were associated with an increased incidence of miscarriage in pregnant patients.

Voriconazole

Voriconazole, which is available in both oral and IV formulations, has a broader spectrum than fluconazole against *Candida* species (including *C. glabrata* and *C. krusei*) and is active against *Aspergillus*, *Scedosporium*, *Fusarium*, and *Coccidioides*. It is generally considered the first-line drug of choice for treatment of aspergillosis. Case reports have shown voriconazole to be effective in individual patients with coccidioidomycosis, blastomycosis, and histoplasmosis; however, because the

data are limited, this agent is not generally recommended for primary treatment of the endemic mycoses. Among the disadvantages of voriconazole (compared with fluconazole) are its more numerous interactions with many of the drugs used in patients predisposed to fungal infections. Hepatotoxicity, skin rashes (including photosensitivity), and visual disturbances are relatively common. Skin cancer surveillance is now recommended for patients taking voriconazole. In addition, voriconazole is considerably more expensive than fluconazole. Moreover, it is advisable to monitor voriconazole levels in certain patients since (1) this drug is completely metabolized in the liver by CYP2C9, CYP3A4, and CYP2C19; and (2) human genetic variability in CYP2C19 activity exists. Dosages should be reduced accordingly in patients with liver failure. Dose adjustments for renal insufficiency are not necessary; however, because the IV formulation is prepared in cyclodextrin, it should not be given to patients with severe renal insufficiency.

Itraconazole

Itraconazole is available in IV and oral (capsule and suspension) formulations. Varying blood levels among patients taking oral itraconazole reflect a disadvantage compared with the other azoles. Itraconazole is the drug of choice for mild to moderate histoplasmosis and blastomycosis and has often been used for chronic mucocutaneous candidiasis. It has been approved by the FDA for use in febrile neutropenic patients; however, most centers use other azoles in neutropenic patients for both prophylaxis and treatment. Itraconazole has also proved useful for the treatment of chronic coccidioidomycosis, sporotrichosis, and *S. apiospermum* infection. The mucocutaneous and cutaneous fungal infections that have been treated successfully with itraconazole include oropharyngeal candidiasis (especially in AIDS patients), tinea versicolor, tinea capitis, and onychomycosis. Disadvantages of itraconazole include its poor penetration into CSF, the use of cyclodextrin in both the oral suspension and the IV formulation, the variable absorption of the drug in capsule form, and the need for monitoring of blood levels in patients taking capsules for disseminated mycoses. Reported cases of severe congestive heart failure in patients taking itraconazole have been a source of concern. Like the other azoles, itraconazole can cause hepatic toxicity.

Posaconazole

Posaconazole is approved by the FDA for prophylaxis of aspergillosis and candidiasis in patients at high risk for developing these infections because of severe immunocompromise. It has also been approved for the treatment of oropharyngeal candidiasis and has been evaluated as therapy for zygomycosis, fusariosis, aspergillosis, cryptococcosis, and various other forms of candidal infection except candidemia. The relevant studies of posaconazole in zygomycosis, fusariosis, and aspergillosis have examined salvage therapy. A study of >90 patients whose zygomycosis was refractory to other therapy yielded encouraging results. No trials of posaconazole for the treatment of candidemia have yet been reported. Case reports have described the drug’s efficacy in coccidioidomycosis and histoplasmosis. Controlled trials have shown its effectiveness as a prophylactic agent in patients with acute leukemia and in bone marrow transplant recipients. In addition, posaconazole has been found to be effective against fluconazole-resistant *Candida* species. The results of a large-scale study of the use of posaconazole as salvage therapy for aspergillosis indicated that it is an alternative to other agents for salvage therapy; however, that study predated the use of voriconazole and the echinocandins.

Isavuconazole

Isavuconazole is the newest of the azoles to be approved by the FDA. It is approved for invasive aspergillosis and invasive mucormycosis. Because of the paucity of drugs effective in mucormycosis and the high mortality rate from this infection, isavuconazole was approved on the basis of an open-label, non-comparative trial in 37 patients. Future experience will more definitively determine its place in the antifungal armamentarium.

ECHINOCANDINS

The echinocandins, including the FDA-approved drugs caspofungin, anidulafungin, and micafungin, have added considerably to the stock of available antifungal drugs. All three of these agents inhibit β-1,3-glucan synthase, which is necessary for cell wall synthesis in fungi and is not a component of human cells. None of these agents is currently available in an oral formulation. The echinocandins are considered fungicidal for *Candida* and fungistatic for *Aspergillus*. Their greatest use to date is against candidal infections. They offer two advantages: broad-spectrum activity against all *Candida* species and relatively low toxicity. The minimal inhibitory concentrations (MICs) of all the echinocandins are highest against *Candida parapsilosis*; it is not clear whether these higher MIC values represent less clinical effectiveness against this species. The echinocandins are among the safest antifungal agents.

In controlled trials, *caspofungin* has been at least as efficacious as AmB for the treatment of candidemia and invasive candidiasis and as efficacious as fluconazole for the treatment of candidal esophagitis. In addition, caspofungin has been efficacious as salvage therapy for aspergillosis. *Anidulafungin* has been approved by the FDA as therapy for candidemia in nonneutropenic patients and for *Candida* esophagitis, intraabdominal infection, and peritonitis. In controlled trials, anidulafungin has been shown to be noninferior and possibly superior to fluconazole against candidemia and invasive candidiasis. It is as efficacious as fluconazole against candidal esophagitis. When anidulafungin is used with cyclosporine, tacrolimus, or voriconazole, no dosage adjustment is required for either drug in the combination. *Micafungin* has been approved for the treatment of esophageal candidiasis and candidemia and for prophylaxis in patients receiving stem cell transplants. In a head-to-head trial, micafungin was noninferior to caspofungin for the treatment of candidemia. Studies thus far have shown that coadministration of micafungin and cyclosporine does not require dose adjustments for either drug. When micafungin is given with sirolimus, the area under the plasma drug concentration–time curve rises for sirolimus, usually necessitating a reduction in its dose. In open-label trials, favorable results have been obtained with micafungin for the treatment of deep-seated *Aspergillus* and *Candida* infections.

FLUCYTOSINE (5-FLUOROCYTOSINE)

The use of flucytosine has diminished as newer antifungal drugs have been developed. This agent is now used most commonly in combination with AmB (deoxycholate or lipid formulations) for the initial treatment of cryptococcal meningitis. Flucytosine has a unique mechanism of action based on intrafungal conversion to 5-fluorouracil, which is toxic to the fungal cell. Development of resistance to the compound has limited its use as a single agent. Flucytosine is nearly always used in combination with AmB. Its good penetration into the CSF makes it attractive for use with AmB for treatment of cryptococcal meningitis. Flucytosine has also been recommended for the treatment of candidal meningitis in combination with AmB; comparative trials with AmB alone have not been done. Significant and frequent bone marrow depression is seen with flucytosine when this drug is used with AmB.

GRISEOFULVIN AND TERBINAFINE

Historically, griseofulvin has been useful primarily for ringworm infection. This agent is usually given for relatively long periods. Terbinafine has been used primarily for onychomycosis but also for ringworm. In comparative studies, terbinafine has been as effective as itraconazole and more effective than griseofulvin for both conditions.

TOPICAL ANTIFUNGAL AGENTS

A detailed discussion of the agents used for the treatment of cutaneous fungal infections and onychomycosis is beyond the scope of this chapter; the reader is referred to **Chap. 214** and the dermatology literature. Many classes of compounds have been used to treat the common fungal infections of the skin. Among the azoles used are clotrimazole, econazole, miconazole, oxiconazole, sulconazole, ketoconazole, tioconazole, butoconazole, and terconazole. In general, topical treatment of vaginal candidiasis has been successful. Since little difference is thought to exist in the efficacy of

the various vaginal preparations, the choice of agent is made by the physician and/or the patient on the basis of preference and availability. Fluconazole given orally at 150 mg has the advantage of not requiring repeated intravaginal application. Nystatin is a polyene that has been used for both oropharyngeal thrush and vaginal candidiasis. Useful agents in other classes include ciclopirox olamine, haloprogin, terbinafine, naftifine, tolnaftate, and undecylenic acid.

FURTHER READING

Bennett JE: *Introduction to the mycoses, in Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, 8th ed. JE Bennett et al (eds). Philadelphia, Elsevier Saunders, 2015, p 2874.
Busca A et al: Reviewing the importance and evolution of fungal infections and potential antifungal resistance in haematological patients. J Glob Antimicrob Resist 3:237, 2015. [PubMed: 27842866]
Miceli MH et al: Risk factors and outcomes of invasive fungal infections in allogeneic hematopoietic cell transplant recipients. Mycopathologia 182:495, 2017.
Ruhnke M, Schwartz S: Recent developments in the management of invasive fungal infections in patients with oncohematological diseases. Ther Adv Hematol 7:345, 2016. [PubMed: 27904738]
Suleyman G, Alangaden GJ: Nosocomial fungal infections: Epidemiology, infection control, and prevention. Infect Dis Clin North Am 30:1023, 2016. [PubMed: 27816138]

Chapter 207: Histoplasmosis

Chadi A. Hage; L. Joseph Wheat

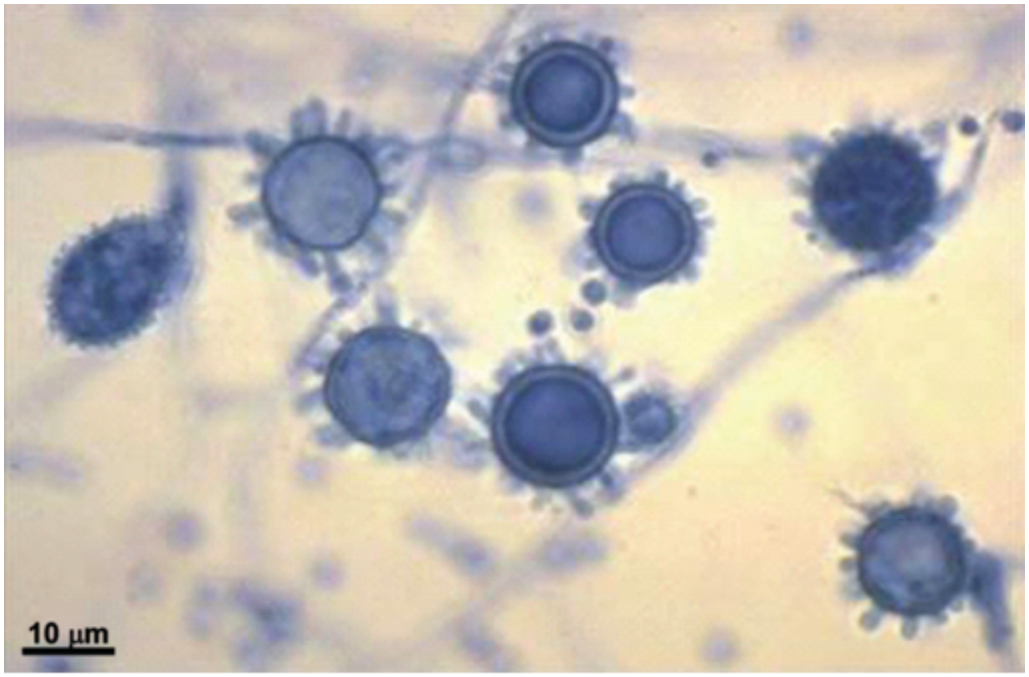
ETIOLOGY

Histoplasma capsulatum, a thermal dimorphic fungus, is the etiologic agent of histoplasmosis. In most endemic areas, *H. capsulatum* var. *capsulatum* is the causative agent. In Central and South America, histoplasmosis is common and is caused by genetically different clades of *H. capsulatum* var. *capsulatum*. In Africa, *H. capsulatum* var. *duboisii* is also found. Yeasts of var. *duboisii* are larger than those of var. *capsulatum*.

Mycelia—the naturally infectious form of *Histoplasma*—have a characteristic appearance, with microconidial and macroconidial forms **(Fig. 207-1)**. Microconidia are oval and are small enough (2–4 μm) to reach the terminal bronchioles and alveoli. Shortly after infecting the host, mycelia transform into the yeasts that are found inside macrophages and other phagocytes. The yeast forms are characteristically small (2–5 μm), with occasional narrow budding **(Fig. 207-2)**. In the laboratory, mycelia are best grown at room temperature, whereas yeasts are grown at 37°C on enriched media.

FIGURE 207-1

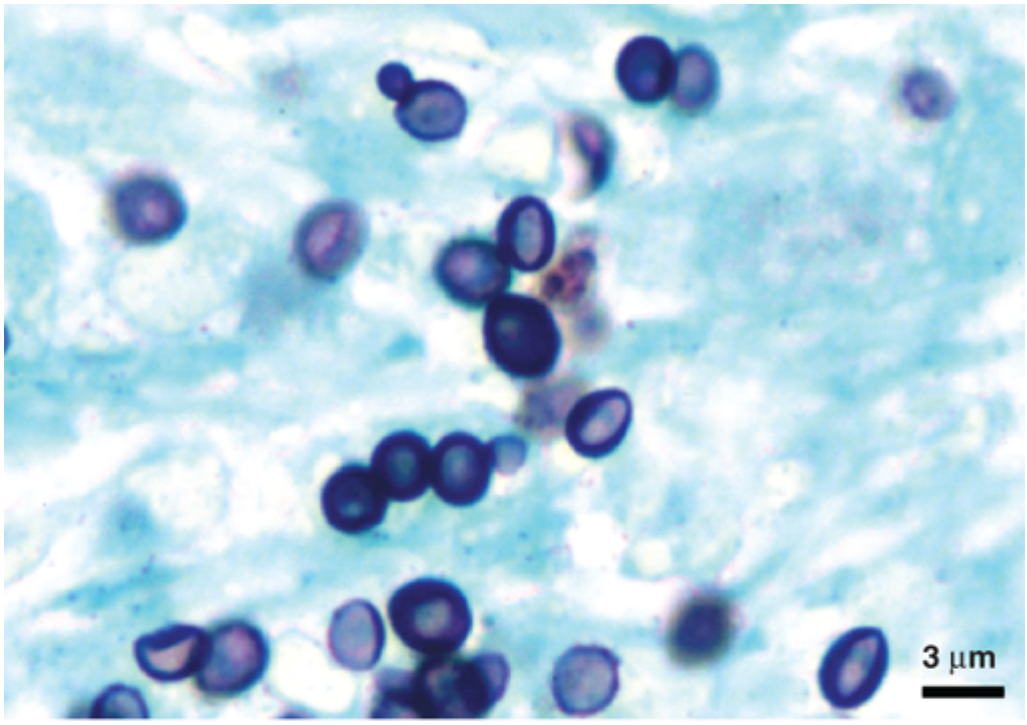
Spiked spherical conidia of *H. capsulatum* (lacto-phenol cotton blue stain).



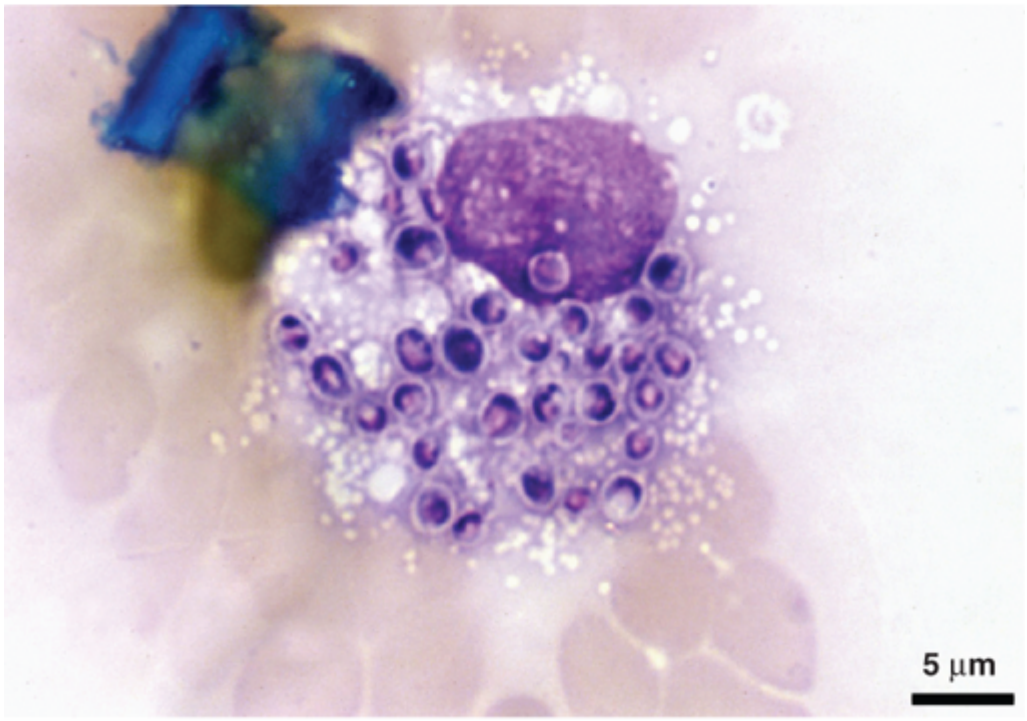
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 207-2

A. Small (2–5 μm) narrow budding yeasts of *H. capsulatum* from bronchoalveolar lavage fluid (Grocott's methenamine silver stain). ***B.*** Intracellular yeasts of *H. capsulatum* within an alveolar macrophage from a patient with AIDS and disseminated histoplasmosis (Giemsa stain).



A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPIDEMIOLOGY

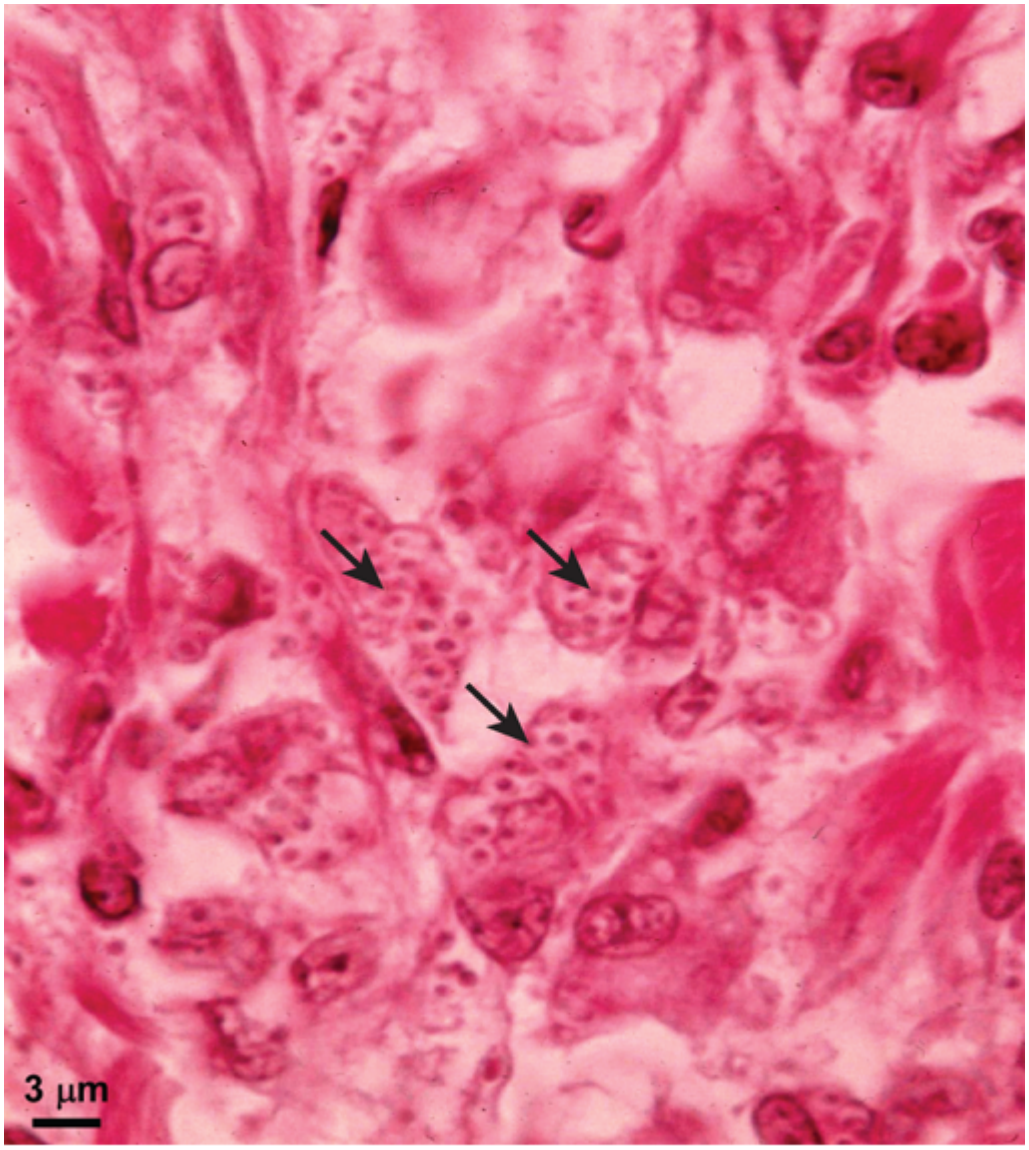
Histoplasmosis is the most prevalent endemic mycosis in North America. Although this fungal disease has been reported throughout the world, its endemicity is particularly notable in the Ohio and Mississippi river valleys of North America and in certain parts of Central and South America, Africa, and Asia. In Europe, histoplasmosis is diagnosed fairly often, mostly in emigrants from or travelers to endemic areas on other continents. The geographic distribution of histoplasmosis is related to the humid and acidic nature of the soil in the endemic areas. Soil enriched with bird or bat droppings promotes the growth and sporulation of *Histoplasma.* Disruption of soil containing the organism leads to aerosolization of the microconidia and exposure of humans nearby. Activities associated with high-level exposure include spelunking, excavation, cleaning of chicken coops, demolition and remodeling of old buildings, and cutting of dead trees. Most cases seen outside of highly endemic areas represent imported disease—e.g., cases reported in Europe after travel to the Americas, Africa, or Asia. The epidemiology of histoplasmosis is changing with the continued expansion of at-risk populations and the acceleration of intercontinental and international travel that brings this infection to areas of the world that are not known to be endemic. The population at risk for histoplasmosis continues to grow as a result of increasing numbers of patients receiving immunosuppressive therapies for autoimmune disorders, cancers, and organ transplants.

PATHOGENESIS AND PATHOLOGY

Infection follows inhalation of microconidia (Fig. 207-1). Once they reach the alveolar spaces, microconidia are rapidly recognized and engulfed by alveolar macrophages. At this point, the microconidia transform into budding yeasts **(Fig. 207-2)**, a process that is integral to the pathogenesis of histoplasmosis and is dependent on the availability of calcium and iron inside the phagocytes. The yeasts are capable of multiplying inside resting macrophages. Neutrophils and then lymphocytes are attracted to the site of infection. Before the development of cellular immunity, yeasts use the phagosomes as a vehicle for translocation to local draining lymph nodes, whence they spread hematogenously throughout the reticuloendothelial system. Adequate cellular immunity develops ~2 weeks after infection. T cells produce interferon γ to assist the macrophages in killing the organism and controlling the progression of disease. Interleukin 12 and tumor necrosis factor α (TNF-α) play an essential role in cellular immunity to *H. capsulatum*. In the immunocompetent host, macrophages, lymphocytes, and epithelial cells eventually organize and form granulomas that contain the organisms. These granulomas typically fibrose and calcify; calcified lung nodules, mediastinal lymph nodes, and hepatosplenic calcifications are frequently found in healthy individuals from endemic areas. In immunocompetent hosts, infection with *H. capsulatum* confers some immunity to reinfection. In patients with impaired cellular immunity, the infection is not properly contained and can disseminate throughout the reticuloendothelial system. Progressive disseminated histoplasmosis (PDH) can involve multiple organs, most commonly the lungs, bone marrow, spleen, liver **(Fig. 207-3)**, adrenal glands, and mucocutaneous membranes. Unlike latent tuberculosis, latent histoplasmosis rarely reactivates.

FIGURE 207-3

Intracellular yeasts (*arrows*) of *H. capsulatum* in a liver biopsy specimen (hematoxylin and eosin stain).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Structural lung disease (e.g., emphysema) impairs the clearance of pulmonary histoplasmosis, and chronic pulmonary disease can result. This chronic process is characterized by progressive inflammation, tissue necrosis, and fibrosis mimicking cavitary tuberculosis.

CLINICAL MANIFESTATIONS

The clinical spectrum of histoplasmosis ranges from asymptomatic infection to life-threatening illness. The attack rate and the extent and severity of the disease depend on the intensity of exposure, the immune status of the exposed individual, and the underlying lung architecture of the host.

In immunocompetent individuals with low-level exposure, most *Histoplasma* infections are either asymptomatic or mild and self-limited. Of adults residing in endemic areas, 50–80% have skin-test and/or radiographic evidence of previous infection without clinical manifestations. Asymptomatic lung nodules representing controlled histoplasmosis are frequently found on chest CT scans obtained during screening for lung cancer in smokers from endemic areas. When symptoms of acute histoplasmosis develop, they usually appear 1–4 weeks after exposure. Heavy exposure leads to a flulike illness with fever, chills, sweats, headache, myalgia, anorexia, cough, dyspnea, and chest pain.

Chest radiographs usually show signs of pneumonitis with prominent hilar or mediastinal adenopathy. Pulmonary infiltrates may be focal with light exposure or diffuse with heavy exposure. Rheumatologic symptoms of arthralgia or arthritis, often associated with erythema nodosum, occur in 5–10% of patients with acute histoplasmosis. Pericarditis may also develop. These manifestations represent inflammatory responses to the acute pulmonary infection rather than extrapulmonary spread. Affected hilar or mediastinal lymph nodes may undergo necrosis and coalesce to form large mediastinal masses that can cause compression of great vessels, proximal airways, and the esophagus. These necrotic lymph nodes may also rupture and create fistulas between mediastinal structures (e.g., bronchoesophageal fistulas).

PDH is typically seen in immunocompromised individuals, who account for ~70% of cases. Common risk factors include AIDS (CD4+ T cell count, <200/μL), extremes of age, the administration of immunosuppressive medications to prevent or treat rejection following transplantation (e.g., prednisone, mycophenolate, calcineurin inhibitors), and the use of methotrexate, anti-TNF-α agents, and other biological response modifiers for autoimmune disorders.

The clinical spectrum of PDH ranges from an acute, rapidly fatal course—with diffuse interstitial or reticulonodular lung infiltrates causing respiratory failure, shock, coagulopathy, and multiorgan failure—to a more subacute course with a focal organ distribution. Common manifestations include fever, weight loss, hepatosplenomegaly, and thrombocytopenia. Other findings may include meningitis or focal brain lesions, ulcerations of the oral mucosa, gastrointestinal ulcerations and bleeding, and adrenal insufficiency. Prompt recognition of this devastating illness is of paramount importance in patients with more severe manifestations or with underlying immunosuppression, especially that due to AIDS **(Chap. 197)**.

Chronic cavitary histoplasmosis is seen in smokers who have structural lung disease (e.g., bullous emphysema). This chronic illness is characterized by productive cough, dyspnea, low-grade fever, night sweats, and weight loss. Chest radiographs usually show upper-lobe infiltrates, cavitation, and pleural thickening—findings resembling those of tuberculosis. Without treatment, the course is slowly progressive.

Fibrosing mediastinitis is an uncommon and serious complication of histoplasmosis. In certain patients, acute infection is followed for unknown reasons by progressive fibrosis around the hilar and mediastinal lymph nodes. Involvement may be unilateral or bilateral; bilateral involvement carries a worse prognosis. Major manifestations include superior vena cava syndrome, obstruction of pulmonary vessels, and airway obstruction. Patients may experience recurrent pneumonia, hemoptysis, or respiratory failure. Fibrosing mediastinitis is fatal in up to one-third of cases.

In healed histoplasmosis, calcified mediastinal nodes or lung parenchymal nodules may erode through the walls of the airways and cause hemoptysis and expectoration of calcified material. This condition is called *broncholithiasis*.

 The clinical features and management of histoplasmosis caused by the genetically different clades in Central and South America are similar to those of the disease in North America. African histoplasmosis caused by var. *duboisii* is clinically distinct and is characterized by frequent skin and bone involvement.

DIAGNOSIS

Recommendations for the diagnosis and treatment of histoplasmosis are summarized in **Table 207-1**. Once suspected, the diagnosis of histoplasmosis is usually straightforward as many diagnostic tools are now available. This is not the case in resource-limited endemic regions of Central America, South America, and Africa, where the diagnosis is often delayed, with consequently poor outcomes.

Fungal culture remains the gold standard diagnostic test for histoplasmosis. However, culture results may not be known for up to 1 month, and cultures are often negative in less severe cases. Cultures are positive in ~75% of cases of PDH and chronic pulmonary histoplasmosis. Cultures of bronchoalveolar lavage (BAL) fluid are positive in about half of patients with acute pulmonary histoplasmosis causing diffuse infiltrates with hypoxemia. In PDH, the culture yield is highest for BAL fluid, bone marrow aspirate, and blood. Cultures of sputum or bronchial washings are usually positive in chronic pulmonary histoplasmosis. Cultures are typically negative, however, in other forms of histoplasmosis.

Fungal stains of cytopathology or biopsy materials showing structures resembling *Histoplasma* yeasts are helpful in the diagnosis of PDH, yielding positive results in about half of cases. Yeasts can be seen in BAL fluid (Fig. 207-2) from patients with diffuse pulmonary infiltrates, in bone marrow biopsy samples, and in biopsy specimens of other involved organs (e.g., the adrenal glands). Occasionally, yeasts are seen within circulating phagocytes on blood smears from patients with severe PDH. However, staining artifacts and other fungal elements sometimes stain positively and may be misidentified as *Histoplasma* yeasts.

TABLE 207-1 Recommendations for the Diagnosis and Treatment of Histoplasmosis

Type of Histoplasmosis	Diagnostic Tests	Treatment Recommendations	Comments
Acute pulmonary, moderate to severe illness with diffuse infiltrates and/or hypoxemia	*Histoplasma* antigen (BAL fluid, serum, urine) Cytopathology on and fungal culture of BAL fluid	Lipid AmB (3–5 mg/kg per day) ± glucocorticoids for 1–2 weeks; then itraconazole (200 mg bid) for 12 weeks. Monitor renal and hepatic function.	Patients with mild cases usually recover without therapy, but itraconazole should be considered if the patient's condition has not improved after 1 month.
Chronic/cavitary pulmonary	*Histoplasma* serology (immunodiffusion and complement fixation) Fungal culture of sputum or BAL fluid	Itraconazole (200 mg qd or bid) for at least 12 months. Monitor hepatic function.	Continue treatment until radiographic findings show no further improvement. Monitor for relapse after treatment is stopped.
Progressive disseminated	*Histoplasma* antigen (serum, urine) Fungal culture of blood or bone marrow aspirate Cytopathology on biopsy of affected organ	Lipid AmB (3–5 mg/kg per day) for 1–2 weeks; then itraconazole (200 mg bid) for at least 12 months. Monitor renal and hepatic function.	Liposomal AmB is preferred, but the AmB lipid complex may be used because of cost. Chronic maintenance therapy may be necessary if the degree of immunosuppression cannot be reduced.
Central nervous system	*Histoplasma* antigen and serology of CSF Fungal culture of CSF	Liposomal AmB (5 mg/kg per day) for 4–6 weeks; then itraconazole (200 mg bid or tid) for at least 12 months. Monitor renal and hepatic function.	A longer course of lipid AmB is recommended because of the high risk of relapse. Itraconazole should be continued until CSF or CT abnormalities clear.

Abbreviations: AmB, amphotericin B; BAL, bronchoalveolar lavage; CSF, cerebrospinal fluid.

The detection of *Histoplasma* antigen in body fluids is extremely useful in the diagnosis of PDH and acute diffuse pulmonary histoplasmosis. The sensitivity of this technique is >95% in patients with PDH and >80% in patients with acute pulmonary histoplasmosis if both urine and serum are tested. Antigen levels correlate with severity of illness in PDH and can be used to follow disease progression, as levels predictably decrease with effective therapy. Increased antigen levels also predict relapse. Antigen can be detected in cerebrospinal fluid from patients with meningitis and in BAL fluid from those with pneumonia. Cross-reactivity occurs with African histoplasmosis, blastomycosis, coccidioidomycosis, paracoccidioidomycosis, and *Penicillium marneffei* infection.

Serologic tests, including immunodiffusion and complement fixation, are useful for the diagnosis of histoplasmosis in immunocompetent patients. At least 1 month is required for the production of antibodies after the onset of infection; thus the utility of serology for early diagnosis of acute histoplasmosis is limited. The antibody titer may rise by fourfold in patients with acute histoplasmosis.

Serologic tests are especially useful for the diagnosis of chronic pulmonary histoplasmosis. Limitations of serology, however, include insensitivity early in the course of infection and in immunosuppressed patients and the persistence of detectable antibody for several years after infection. Positive results from past infection may lead to a misdiagnosis of active histoplasmosis in a patient with another disease process.

TREATMENT

TREATMENT Histoplasmosis

Treatment is indicated for all patients with PDH or chronic pulmonary histoplasmosis as well as for symptomatic patients with acute pulmonary histoplasmosis causing diffuse infiltrates, especially with hypoxemia. In most cases of pulmonary histoplasmosis, treatment is not recommended because the immune system of the host is intact and the degree of exposure is not heavy; the infection is asymptomatic or symptoms are mild, subacute, and not progressive; and the illness resolves without therapy.

The preferred treatments for histoplasmosis (Table 207-1) include the lipid formulations of amphotericin B in more severe cases and itraconazole in others. Liposomal amphotericin B is more effective and better tolerated than the deoxycholate formulation for treatment of PDH in patients with AIDS. The deoxycholate formulation of amphotericin B is an alternative to a lipid formulation for patients at low risk for nephrotoxicity. Voriconazole, posaconazole, and isavuconazole are alternatives for patients who cannot take itraconazole.

In severe cases requiring hospitalization, a lipid formulation of amphotericin B is used first, followed by itraconazole. In patients with meningitis, a lipid formulation of amphotericin B should be given for 4–6 weeks before switching to itraconazole. In immunosuppressed patients, the degree of immunosuppression should be reduced if possible, although immune reconstitution inflammatory syndrome (IRIS) may ensue. Antiretroviral treatment improves the outcome of PDH in patients with AIDS and is recommended; however, whether antiretroviral treatment should be delayed to avoid IRIS is unknown.

Blood levels of itraconazole should be monitored to ensure adequate drug exposure, with target concentrations of the parent drug and its hydroxy metabolites measuring 1–5 μg/mL. Drug interactions should be carefully assessed: itraconazole not only is cleared by cytochrome P450 metabolism but also inhibits cytochrome P450. This profile causes interactions with many other medications.

The duration of treatment for acute pulmonary histoplasmosis is 6–12 weeks, while that for PDH and chronic pulmonary histoplasmosis is ≥1 year. Antigen levels in urine and serum should be monitored during and for at least 1 year after therapy for PDH. Stable or rising antigen levels suggest treatment failure or relapse.

Lifelong itraconazole maintenance therapy is recommended for patients with persistently suppressed immunity but not for those with immune recovery—e.g., patients with AIDS who respond well to antiretroviral therapy, with CD4+ T cell counts of at least 150/μL (preferably >250/μL); who complete at least 1 year of itraconazole therapy; and who exhibit neither clinical evidence of active histoplasmosis nor an antigenuria level of >2 ng/mL. Similarly, maintenance therapy is not necessary in patients receiving immunosuppressive treatment if the degree of immunosuppression can be reduced along with effective control of the infection.

Fibrosing mediastinitis, which represents a chronic fibrotic reaction to past mediastinal histoplasmosis rather than an active infection, does not respond to antifungal therapy. While treatment is often prescribed for patients with pulmonary histoplasmosis who have not recovered within 1 month and for

those with persistent mediastinal lymphadenopathy, the effectiveness of antifungal therapy in these situations is unknown.

FURTHER READING

Bahr NC et al: Histoplasmosis infections worldwide: Thinking outside of the Ohio River valley. Curr Trop Med Rep 2:70, 2015. [PubMed: 26279969]
Hage CA et al: A multicenter evaluation of tests for diagnosis of histoplasmosis. Clin Infect Dis 53:448, 2011. [PubMed: 21810734]
Hage CA et al: Histoplasmosis: Up-to-date evidence-based approach to diagnosis and management. Semin Respir Crit Care Med 36:729, 2015. [PubMed: 26398539]

Chapter 208: Coccidioidomycosis

Neil M. Ampel

FIGURE 208-1

DEFINITION AND ETIOLOGY

Coccidioidomycosis, commonly known as Valley fever (see "Epidemiology," below), is caused by dimorphic soil-dwelling fungi of the genus *Coccidioides*. Genetic analysis has demonstrated the existence of two species, *C. immitis* and *C. posadasii*. These species are indistinguishable with regard to the clinical disease they cause and their appearance on routine laboratory media. Thus, the organisms will be referred to simply as *Coccidioides* for the remainder of this chapter.

EPIDEMIOLOGY

Coccidioidomycosis is confined to the Western Hemisphere between the latitudes of 40°N and 40°S. In the United States, areas of high endemicity include the southern portion of the San Joaquin Valley of California (hence the sobriquet "Valley fever") and the south-central region of Arizona. However, infection may be acquired in other areas of the southwestern United States, including the southern coastal counties in California, southern Nevada, southwestern Utah, southern New Mexico, and western Texas (including the Rio Grande Valley). The recent acquisition of cases well outside the recognized areas, including in eastern Washington state and in northeastern Utah, suggests that the endemic region may be expanding. Outside the United States, coccidioidomycosis is endemic to northern Mexico as well as to localized regions of Central America. In South America, there are endemic foci in Colombia, Venezuela, northeastern Brazil, Paraguay, Bolivia, and north-central Argentina.

The risk of infection is increased by direct exposure to soil harboring *Coccidioides*. Because of difficulty in isolating *Coccidioides* from the soil, the precise characteristics of potentially infectious soil are not known. In the United States, several outbreaks of coccidioidomycosis have been associated with soil from archaeologic excavations of Amerindian sites both within and outside of the recognized endemic region. These cases often involved alluvial soils in regions of relative aridity with moderate temperature ranges. When found, *Coccidioides* is isolated 2–20 cm below the surface; it is not found in soil at greater depths, nor is it usually isolated from cultivated soil.

In endemic areas, many cases of *Coccidioides* infection occur without obvious soil or dust exposure. Climatic factors appear to increase the infection rate in these regions. In particular, periods of aridity following rainy seasons have been associated with marked increases in the number of symptomatic cases. Overall, the incidence within the United States increased substantially over the past decade, with nearly 43 cases per 100,000 residents of the endemic region in 2011. Most of that increase occurred in south-central Arizona, where most of that state's population resides, and in the southern

San Joaquin Valley of California, a less populated region. The factors causing this increase have not been fully elucidated; however, an influx of older individuals without prior coccidioidal infection appears to be involved. Other variables, such as climate change, construction activity, and increased awareness and reporting, may also be contributors. Health care providers should consider coccidioidomycosis when evaluating persons with pneumonia who live in or have traveled to endemic areas.

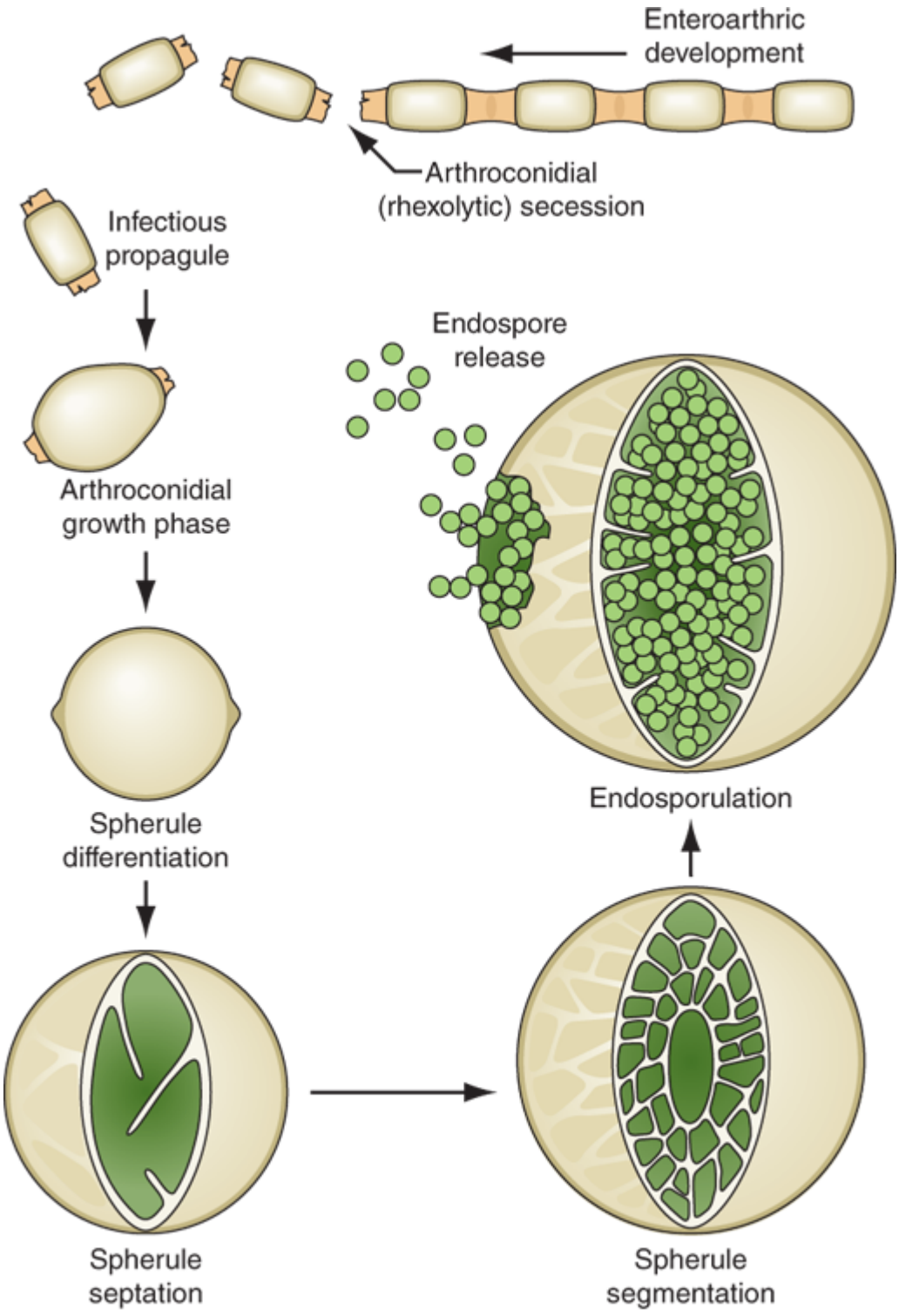
PATHOGENESIS, PATHOLOGY, AND IMMUNE RESPONSE

On agar media and in the soil, *Coccidioides* organisms exist as filamentous molds. Within this mycelial structure, individual filaments (*hyphae*) elongate and branch, some growing upward. Alternating cells within the hyphae degenerate, leaving barrel-shaped viable elements called *arthroconidia*. Measuring ~2 μm by 5 μm, arthroconidia may become airborne for extended periods. Their small size allows them to evade initial mechanical mucosal defenses and reach deep into the bronchial tree, where infection is initiated in the nonimmune host.

Once in a susceptible host, the arthroconidia enlarge, become rounded, and develop internal septations. The resulting structures, called *spherules* **(Fig. 208-1)**, may attain sizes of 200 μm and are unique to *Coccidioides*. The septations encompass uninuclear elements called *endospores*. Spherules may rupture and release packets of endospores that can themselves develop into spherules, thus propagating infection locally. If returned to artificial media or the soil, the fungus reverts to its mycelial stage.

FIGURE 208-1

Life cycle of *Coccidioides.* *(From TN Kirkland, J Fierer: Emerg Infect Dis 2:192, 1996.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical observations and data from studies of animals strongly support the critical role of a robust cellular immune response in the host's control of coccidioidomycosis. Necrotizing granulomas containing spherules are typically identified in patients with resolved pulmonary infection. In disseminated disease, granulomas are generally poorly formed or do not develop at all, and a polymorphonuclear leukocyte response occurs frequently. In patients who are asymptomatic or in whom the initial pulmonary infection resolves, delayed-type hypersensitivity to coccidioidal antigens has been routinely documented.

CLINICAL AND LABORATORY MANIFESTATIONS

Among infected individuals, 60% are completely asymptomatic, and the remaining 40% have symptoms that are related principally to pulmonary infection, including fever, cough, and pleuritic chest pain. The risk of symptomatic illness increases with age. Coccidioidomycosis is commonly misdiagnosed as community-acquired bacterial pneumonia.

There are several cutaneous manifestations of primary pulmonary coccidioidomycosis. Toxic erythema consisting of a maculopapular rash has been noted in some cases. Erythema nodosum **(see Fig. A1-39)**—typically over the lower extremities—or erythema multiforme **(see Fig. A1-24)**—usually in a necklace distribution—may occur; these manifestations are seen especially often in women. Arthralgias and arthritis may develop. The diagnosis of primary pulmonary coccidioidomycosis is particularly suggested by a history of night sweats or profound fatigue as well as by peripheral-blood eosinophilia and hilar or mediastinal lymphadenopathy on chest radiography. While pleuritic chest pain is common, pleural effusions occur in fewer than 10% of cases. Such effusions are invariably associated with a pulmonary infiltrate on the same side. The cellular content of these effusions is mononuclear in nature; *Coccidioides* is rarely isolated on culture of effusions.

In most patients, primary pulmonary coccidioidomycosis usually resolves without sequelae in weeks. However, several pneumonic complications may arise. Pulmonary nodules are residua of primary pneumonia. Generally single, frequently located in the upper lobes, and ≤4 cm in diameter, nodules are often discovered on a routine chest radiograph in an asymptomatic patient. Calcification is uncommon. Coccidioidal pulmonary nodules can be difficult to distinguish radiographically from pulmonary malignancies. Like malignancies, coccidioidal nodules often enhance on positron emission tomography. However, unlike malignancies, routine CT often demonstrates multiple nodules in coccidioidomycosis. Biopsy is often required to distinguish between these two conditions.

Pulmonary cavities occur when a nodule extrudes its contents into the bronchus, resulting in a thin-walled shell. These cavities can be associated with persistent cough, hemoptysis, and pleuritic chest pain. Rarely, a cavity may rupture into the pleural space, causing pyopneumothorax. In such cases, patients present with acute dyspnea, and the chest radiograph reveals a collapsed lung with a pleural air-fluid level. Chronic or persistent pulmonary coccidioidomycosis manifests with prolonged fever, cough, and weight loss and is radiographically associated with pulmonary scarring, fibrosis, and cavities. It occurs most commonly in patients who already have chronic lung disease due to other etiologies.

In some cases, primary pneumonia presents as a diffuse reticulonodular pulmonary process (detected by plain chest radiography) in association with dyspnea and fever. Primary diffuse coccidioidal pneumonia may occur in settings of intense environmental exposure or profoundly suppressed cellular immunity (e.g., in patients with AIDS), with unrestrained fungal growth that is frequently associated with fungemia.

Clinical dissemination outside the thoracic cavity occurs in fewer than 1% of infected individuals. Dissemination is more likely to occur in male patients, particularly those of African-American or Filipino ancestry, and in persons with depressed cellular immunity, including patients with HIV infection and peripheral-blood CD4+ T cell counts of <250/μL; those receiving chronic glucocorticoid therapy; those with allogeneic solid-organ transplants; and those being treated with tumor necrosis factor α antagonists. Women who acquire infection during the second or third trimester of pregnancy or postpartum also are at risk for disseminated disease. Common sites for dissemination include the skin, bones, joints, soft tissues, and meninges. Dissemination may follow symptomatic or asymptomatic pulmonary infection and may involve only one site or multiple anatomic foci. When it occurs, clinical dissemination is usually evident within the first 6 months after primary pulmonary infection.

Coccidioidal meningitis, if untreated, is uniformly fatal. Patients usually present with a persistent headache, which is sometimes accompanied by lethargy and confusion. Nuchal rigidity, if present, is not severe. Examination of cerebrospinal fluid (CSF) demonstrates lymphocytic pleocytosis with profound hypoglycorrhachia and elevated protein levels. CSF eosinophilia is occasionally documented. With or without appropriate therapy, patients may develop hydrocephalus, either communicating or non-communicating, which presents clinically as a marked decline in mental status, often with gait disturbances.

DIAGNOSIS

As mentioned above, coccidioidomycosis is often misdiagnosed as community-acquired bacterial pneumonia. Clues that suggest a diagnosis of coccidioidomycosis include peripheral-blood eosinophilia, hilar or mediastinal adenopathy on radiographic imaging, marked fatigue, and failure to improve with antibiotic therapy.

Serology plays an important role in establishing a diagnosis of coccidioidomycosis. Several techniques are available, including the traditional tube-precipitin (TP) and complement-fixation (CF) assays, immunodiffusion TP and CF (IDTP and IDCF), and enzyme immunoassay (EIA) to detect IgM and IgG antibodies. TP and IgM antibodies are found in serum soon after infection and persist for weeks. They are not useful for gauging disease progression and are not found in the CSF. The CF and IgG antibodies occur later in the course of the disease and persist longer than TP and IgM antibodies. Rising CF titers are associated with clinical progression, and the presence of CF antibody in CSF is indicative of coccidioidal meningitis. Antibodies disappear over time in persons whose clinical illness resolves.

Because of its commercial availability, the coccidioidal EIA is frequently used as a screening tool for coccidioidal serology. There has been concern that the IgM EIA is occasionally falsely positive, particularly in asymptomatic individuals. In addition, while the sensitivity and specificity of the IgG EIA appear to be higher than those of the CF and IDCF assays, the optical density obtained in the EIA does not correlate with the serologic titer of either of the latter tests.

Coccidioides grows within 3–7 days at 37°C on a variety of artificial media, including blood agar. Therefore, it is always useful to obtain samples of sputum or other respiratory fluids and tissues for culture in suspected cases of coccidioidomycosis. The clinical laboratory should be alerted to the possibility of this diagnosis, since *Coccidioides* poses a significant laboratory hazard if it is inadvertently inhaled. The organism can also be identified directly. While treatment of samples with potassium hydroxide is rarely fruitful in establishing the diagnosis, examination of sputum or other respiratory fluids after Papanicolaou, Gomori methenamine silver, or calcofluor white staining reveals spherules in a significant proportion of patients with pulmonary coccidioidomycosis. For fixed tissues (e.g., those obtained from biopsy specimens), spherules with surrounding inflammation can be demonstrated with hematoxylin-eosin or Gomori methenamine silver staining.

A commercially available test for coccidioidal antigenuria and antigenemia has been developed and appears to be particularly useful in immunosuppressed patients with severe or disseminated disease. False-positive results may occur in cases of histoplasmosis or blastomycosis. Some laboratories offer genomic detection by polymerase chain reaction; this assay has not been shown to be more sensitive than culture.

TREATMENT

TREATMENT Coccidioidomycosis

Currently, two main classes of antifungal agents are useful for the treatment of coccidioidomycosis **(Table 208-1)**. While once prescribed routinely, amphotericin B in all its formulations is now reserved for only the most severe cases of dissemination and for intrathecal or intraventricular administration to patients with coccidioidal meningitis in whom triazole antifungal therapy has failed. The original formulation of amphotericin B, which is dispersed with deoxycholate, is usually administered intravenously in doses of 0.7–1.0 mg/kg either daily or three times per week. The newer lipid-based formulations are associated with less renal toxicity but have not been demonstrated to lead to better improvement than the deoxycholate formulation in coccidioidomycosis. The lipid dispersions are administered intravenously at doses of 3–5 mg/kg daily or three times per week.

Triazole antifungals are the principal drugs now used to treat most cases of coccidioidomycosis. Clinical trials have demonstrated the usefulness of both fluconazole and itraconazole. Evidence indicates that itraconazole is efficacious against bone and joint disease. Because of its demonstrated penetration into CSF, fluconazole is the azole of choice for the treatment of coccidioidal meningitis, but itraconazole also is effective. For both drugs, a minimal oral adult dosage of 400 mg/d should be used. The maximal dose of itraconazole is 200 mg three times daily, but higher doses of fluconazole may be given. The newer triazole antifungals, posaconazole and especially voriconazole, appear to be useful against clinical disease, including meningitis, in which prior fluconazole therapy has failed. High-dose triazole therapy may be teratogenic during the first trimester of pregnancy; thus, amphotericin B should be considered as therapy for coccidioidomycosis in pregnant women during this period. Isavuconazole has been used in limited circumstances in coccidioidomycosis.

Most patients with focal primary pulmonary coccidioidomycosis do not require antifungal therapy. Patients for whom antifungal therapy should be considered include those with underlying cellular immunodeficiencies and those with prolonged symptoms and signs of extensive disease. Specific criteria include symptoms persisting for ≥2 months, night sweats occurring for >3 weeks, weight loss of >10%, a serum CF antibody titer of >1:16, and extensive pulmonary involvement apparent on chest radiography. When antifungal therapy is used, either fluconazole or itraconazole at 400 mg daily for 6 months is considered appropriate.

Diffuse pulmonary coccidioidomycosis represents a special situation. Because most patients with this form of disease are profoundly hypoxemic and critically ill, many clinicians favor beginning therapy with an amphotericin B formulation and switching to an oral triazole antifungal once clinical improvement occurs.

The nodules that may follow primary pulmonary coccidioidomycosis do not require treatment. As noted above, these nodules are not easily distinguished from pulmonary malignancies by means of radiographic imaging. Close clinical follow-up and biopsy may be required to distinguish between these two entities. Most pulmonary cavities do not require therapy. Antifungal treatment should be considered in patients with persistent cough, pleuritic chest pain, and hemoptysis. Occasionally, pulmonary coccidioidal cavities become secondarily infected. This development is usually manifested by an air-fluid level within the cavity. Bacterial flora or *Aspergillus* species are commonly involved, and therapy directed at these organisms should be considered. Surgery is rarely required in cases of persistent hemoptysis or pyopneumothorax. In addition, cavities >4 cm in diameter are unlikely to resolve spontaneously, and their surgical extirpation should be considered. For chronic pulmonary coccidioidomycosis, prolonged antifungal therapy—lasting for at least 1 year—is usually required, with monitoring of symptoms, radiographic changes, sputum cultures, and serologic titers.

Most cases of disseminated coccidioidomycosis require prolonged antifungal therapy. Duration of treatment is based on clinical improvement in conjunction with a significant decline in serum CF antibody titer. Such therapy routinely is continued for at least several years. Relapse occurs in 15–30% of individuals once therapy is discontinued.

Coccidioidal meningitis poses a special challenge. While most patients with this form of disease respond to treatment with oral triazoles, 80% experience relapse when therapy is stopped. Thus,

lifelong therapy is recommended. In cases of triazole failure, intrathecal or intraventricular amphotericin B may be used. Installation requires considerable expertise and should be undertaken only by an experienced health care provider. Shunting of CSF in addition to appropriate antifungal therapy is required in cases of meningitis complicated by hydrocephalus. It is prudent to obtain expert consultation in all cases of coccidioidal meningitis.

TABLE 208-1 Clinical Presentations of Coccidioidomycosis, Their Frequency, and Recommended Initial Therapy for the Immunocompetent Host

Clinical Presentation	Frequency, %	Recommended Therapy
Asymptomatic infection	60	None
Primary pneumonia (focal)	40	In most cases, none[a]
Diffuse pneumonia	<1	Amphotericin B followed by prolonged oral triazole therapy
Pulmonary sequelae		
• Nodule	5	None
• Cavity		In most cases, none[b]
• Chronic pneumonia		Prolonged triazole therapy
Disseminated disease		
• Skin, bone, joint, soft tissue disease	≤1	Prolonged triazole therapy[c]
• Meningitis		Lifelong triazole therapy[d]

[a]Treatment is indicated for hosts with depressed cellular immunity as well as for those with prolonged symptoms and signs of increased severity, including night sweats for >3 weeks, weight loss of >10%, a complement-fixation titer of >1:16, and extensive pulmonary involvement on chest radiography.

[b]Treatment (usually with the oral triazoles fluconazole and itraconazole) is recommended for persistent symptoms.

[c]In severe cases, some clinicians would use amphotericin B as initial therapy.

[d]Intraventricular or intrathecal amphotericin B is recommended in cases of triazole failure. Hydrocephalus may occur, requiring a CSF shunt.

Note: See text for dosages and durations.

PREVENTION

There are no proven methods to reduce the risk of acquiring coccidioidomycosis among residents of an endemic region, but avoidance of direct contact with uncultivated soil or with visible dust containing soil is a reasonable measure. For individuals with suppressed cellular immunity, the risk of developing symptomatic coccidioidomycosis is greater than that in the general population. Among those about to undergo allogeneic solid-organ transplantation, antifungal therapy is appropriate when there is evidence of active or recent coccidioidomycosis. Some transplant centers in the endemic region are

providing universal antifungal prophylaxis for 6 months to 1 year after solid organ transplantation. Several cases of donor-transmitted coccidioidomycosis have occurred during transplantation. If possible, donors from an endemic region should be screened for coccidioidomycosis before transplantation. Data on the use of antifungal agents for prophylaxis in other situations are limited. The administration of prophylactic antifungals is not recommended for HIV-1-infected patients who live in an endemic region. Most experts would administer a triazole antifungal to patients with a history of active coccidioidomycosis or a positive coccidioidal serology in whom therapy with tumor necrosis factor α antagonists is being initiated.

FURTHER READING

Centers for Disease Control and Prevention: Increase in reported coccidioidomycosis—United States, 1998–2011. MMWR Morb Mortal Wkly Rep 62:217, 2013. [PubMed: 23535687]
Galgiani JN et al: 2016 Infectious Diseases Society of America (IDSA) clinical practice guideline for the treatment of coccidioidomycosis. Clin Infect Dis 63:e112, 2016. [PubMed: 27470238]
Kahn A et al: Universal fungal prophylaxis and risk of coccidioidomycosis in liver transplant recipients living in an endemic area. Liver Transpl 21:353, 2015. [PubMed: 25482428]
Kusne S et al: Coccidioidomycosis transmission through organ transplantation: A report of the OPTN Ad Hoc Disease Transmission Advisory Committee. Am J Transplant 16:3562, 2016. [PubMed: 27376472]
Litvintseva AP et al: Valley fever: Finding new places for an old disease: *Coccidioides immitis* found in Washington State soil associated with recent human infection. Clin Infect Dis 60:e1, 2015. [PubMed: 25165087]

Chapter 209: Blastomycosis

S. Travis King; Rathel L. Nolan, III

FIGURE 209-1

INTRODUCTION

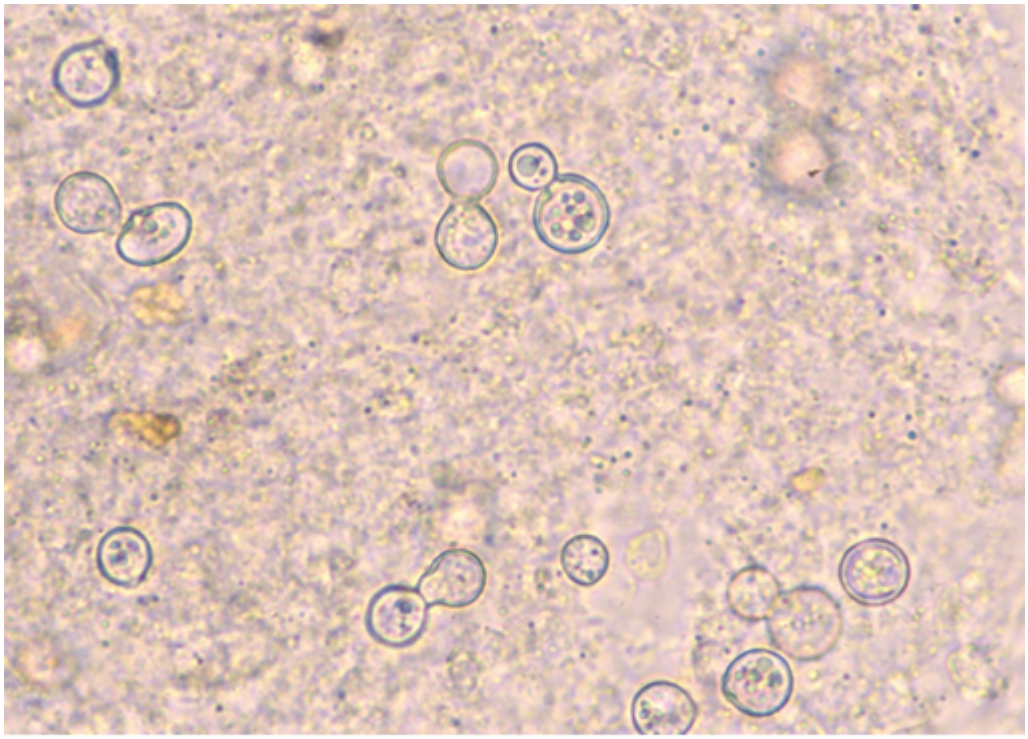
Blastomycosis is a systemic pyogranulomatous infection, involving primarily the lungs, that follows inhalation of the conidia of *Blastomyces dermatitidis*. Pulmonary blastomycosis varies from an asymptomatic infection to acute or chronic pneumonia. Hematogenous dissemination to skin, bones, and the genitourinary system is common; however, almost any organ can be involved.

ETIOLOGIC AGENT

B. dermatitidis is the asexual state of *Ajellomyces dermatitidis*. Two serotypes have been identified on the basis of the presence or absence of the A antigen. Distinct genotypic groups have been differentiated by rDNA polymerase chain reaction restriction fragment length polymorphisms and microsatellite markers. *B. dermatitidis* exhibits thermal dimorphism, growing as the mycelial phase at room temperature and as the yeast phase at 37°C. Primary isolation in the laboratory is most dependable for the mycelial phase incubated at 30°C. Definitive identification usually requires conversion to the yeast phase at 37°C or—now more commonly—the use of nucleic acid amplification techniques that detect mycelial-phase growth. Under the microscope, the yeast cells are usually 8–15 μm in diameter, have thick refractile cell walls, are multinucleate, and exhibit a single, large, broad-based bud **(Fig. 209-1)**.

FIGURE 209-1

***Blastomyces dermatitidis* broad-based budding yeast in the aspirate of a chest wall abscess.** Note the presence of multiple nuclei, the thickened cell wall, and the broad-based bud.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPIDEMIOLOGY

Most cases of blastomycosis have been reported in North America. Endemic areas include the southeastern and south-central states bordering the Mississippi and Ohio river basins, the midwestern states, and the Canadian provinces bordering the Great Lakes. A small endemic area exists in New York and Canada along the St. Lawrence River. Acute blastomycosis is typically found only in North America, and the clinical presentation of blastomycosis in nonendemic areas is as a chronic disease.

Outside North America, blastomycosis occurs sporadically in Nigeria, Zimbabwe, Tunisia, Saudi Arabia, Israel, Lebanon, and India. The disease has been reported most frequently in Africa.

Early studies indicated that middle-aged men with outdoor occupations were at greatest risk. Reported outbreaks, however, do not suggest a predilection according to sex, age, race, occupation, or season. The specific niche in nature in which the organism resides remains uncertain; *B. dermatitidis* probably grows as microfoci in the warm, moist soil of wooded areas rich in organic debris. Inhalation of conidia following exposure to soil, whether related to work or recreation, appears to be the common factor associated with infection. Outbreaks of human disease may be preceded by the occurrence of disease in simultaneously exposed dogs. Zoonotic transmission is rare but has been reported in association with dog bites, pet kinkajou bites, cat scratches, and animal necropsies.

PATHOGENESIS

Alveolar macrophages and polymorphonuclear leukocytes are critical for phagocytosis and killing of the inhaled conidia of *B. dermatitidis*. The interaction of these mediators of the innate immune response with local host factors, such as lung surfactant, plays a significant role in inhibiting conversion to the pathogenic yeast form. This inhibition prevents the establishment of symptomatic disease and may account for the high frequency of asymptomatic infections in outbreaks. Once

conversion to the thick-walled yeast form has occurred, phagocytosis and killing are much more difficult, and the development of clinically apparent infection is much more likely. Ultimately, the T lymphocyte response—specifically, a T_H1 response—is the primary factor in limiting infection and dissemination. Moreover, yeast-phase conversion results in the expression of yeast phase–specific proteins such as the 120-kDa glycoprotein adhesin BAD-1 and the *Blastomyces* yeast phase–specific protein 1 (BYS1). BAD-1 has been well characterized as a virulence factor and is the major epitope for humoral and cellular immunity. The role of BYS1, putatively identified as a signal peptide, has not been determined.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Blastomycosis

Blastomycosis most commonly presents as acute or chronic pneumonia that has been refractory to therapy with antibacterial drugs. Whether acute or chronic, blastomycosis may mimic many other disease processes. For example, acute pulmonary blastomycosis may present with signs and symptoms indistinguishable from those of bacterial pneumonia or influenza, and chronic pulmonary blastomycosis may mimic malignancy or tuberculosis. Skin lesions are often misdiagnosed as basal cell or squamous cell carcinoma, pyoderma gangrenosum, or keratoacanthoma. Laryngeal lesions are frequently mistaken for squamous cell carcinoma. Thus, the clinician must maintain a high index of suspicion and ensure that secretions or biopsy materials from patients who live in or have visited regions endemic for blastomycosis are subjected to careful histologic evaluation. This diligence is especially important in caring for individuals with pneumonia who fail to respond to treatment with antibacterial agents.

CLINICAL MANIFESTATIONS

Acute pulmonary infection is often diagnosed in association with point-source outbreaks. Typical symptoms include the abrupt onset of fever, chills, pleuritic chest pain, arthralgias, and myalgias. Cough is initially nonproductive but frequently becomes purulent as disease progresses. Chest radiographs usually reveal alveolar infiltrates with consolidation. Pleural effusions and hilar adenopathy are uncommon. Most patients diagnosed with pulmonary blastomycosis have chronic indolent pneumonia with signs and symptoms of fever, weight loss, productive cough, and hemoptysis. The most common radiologic findings are alveolar infiltrates with or without cavitation, mass lesions that mimic bronchogenic carcinoma, and fibronodular infiltrates. Hematogenous dissemination to the skin, bones, and genitourinary tract occurs most often in association with chronic pulmonary disease. Although blastomycosis is not considered an opportunistic infection, immunosuppression has been recognized as a risk factor for more serious pulmonary involvement, including respiratory failure (adult respiratory distress syndrome) associated with miliary disease or diffuse pulmonary infiltrates. In the late stages of AIDS, mortality rates of ≥50% have been documented. Most deaths occur within the first few days of therapy. Solid-organ transplant recipients with endemic fungal infections, including both histoplasmosis and blastomycosis, frequently have more severe pulmonary disease as well as dissemination. Blastomycosis has been associated with a mortality rate of 36% in these patients.

 In Africa, pulmonary cases typically include bony involvement (frequently of the vertebrae), with subcutaneous abscesses of the chest wall or legs. All of the manifestations seen in African patients fall within the spectrum of blastomycosis observed in North America. The increased prevalence of chronic and disseminated bone disease in these patients may reflect a delay in diagnosis in regions where spinal disease is often treated empirically as tuberculosis.

Skin disease is the most common extrapulmonary manifestation of blastomycosis. Two types of skin lesions occur: verrucous (more common) and ulcerative. Osteomyelitis occurs in as many as one-

fourth of *B. dermatitidis* infections. The vertebrae, pelvis, sacrum, skull, ribs, and long bones are most frequently involved. Patients with *B. dermatitidis* osteomyelitis often present with contiguous soft-tissue abscesses or chronic draining sinuses. In men, blastomycosis may involve the prostate and epididymis. Central nervous system (CNS) disease occurs in fewer than 5% of immunocompetent patients with blastomycosis. A recent multicenter review identified 22 patients with CNS disease, of whom 12 (54%) met at least one criterion for immunosuppression; although most cases of CNS blastomycosis are associated with infection at other sites, 22.7% of the reviewed cases had only CNS involvement. CNS disease, usually presenting as a brain abscess, has been reported in ~40% of cases in patients with AIDS. Less common forms of CNS disease are cranial or spinal epidural abscess and meningitis.

DIAGNOSIS

Definitive diagnosis of blastomycosis requires growth of the organism from sputum, bronchial washings, pus, or biopsy material. Specimens should be inoculated onto a fungal medium such as Sabouraud dextrose agar, with or without chloramphenicol. *B. dermatitidis* is generally visible in 5–10 days but may require incubation for up to 30 days if only a few organisms are present in the specimen. A presumptive diagnosis may be based on demonstration of the characteristic broad-based budding yeast by microscopic examination of wet preps of sputum in pneumonia or of skin-lesion scrapings.

An assay that detects *Blastomyces* antigen in urine and serum is commercially available and is reasonably sensitive and specific (MiraVista Diagnostics, Indianapolis, IN). Antigen detection appears to be more sensitive in urine than in serum. This antigen test may be useful for monitoring of patients during therapy or for early detection of relapse. Recently, a *Blastomyces* antibody enzyme immunoassay targeting the BAD-1 protein was developed (MiraVista Diagnostics). In combination with antigen testing, the sensitivity was >95% in blastomycosis patients, with 94% specificity in histoplasmosis patients. Chemiluminescent DNA probes (AccuProbe; GenProbe Inc., San Diego, CA) are commonly used to confirm identification of *B. dermatitidis* once growth has been detected in culture. Repetitive sequence–based PCR is available (DiversiLab System; bioMérieux, Durham, NC). Molecular identification techniques are currently used only to supplement traditional diagnostic methods.

TREATMENT

TREATMENT Blastomycosis

The Infectious Diseases Society of America has published guidelines for the treatment of blastomycosis. Selection of an appropriate therapeutic regimen must be based on the clinical form and severity of the disease, the immune status of the patient, and the toxicity of the antifungal agent **(Table 209-1)**. Although spontaneous cures of acute pulmonary infection are well documented, there are no criteria by which to distinguish patients whose disease will progress or resolve without treatment. Thus all patients with blastomycosis should be treated.

Itraconazole is the agent of choice for immunocompetent patients with mild to moderate pulmonary or non-CNS extrapulmonary disease. Therapy is continued for 6–12 months. Amphotericin B (AmB) is preferred for initial treatment of patients who are severely immunocompromised, who have life-threatening disease or CNS disease, or whose disease progresses during treatment with itraconazole. Although not rigorously studied, lipid formulations of AmB provide an alternative for patients who cannot tolerate AmB deoxycholate. Most patients with non-CNS disease whose clinical condition improves after an initial course of AmB (usually 2 weeks in duration) can be switched to itraconazole to complete 6–12 months of therapy. Fluconazole, because of its excellent penetration of the CNS, is useful in the treatment of patients with brain abscess or meningitis after an initial course of AmB.

Voriconazole has been used successfully to treat refractory blastomycosis, blastomycosis in immunosuppressed patients, and—given its good penetration of cerebrospinal fluid—CNS disease. Posaconazole has also been used for refractory pulmonary disease. The echinocandins have variable activity against *B. dermatitidis* and therefore are not used in the treatment of blastomycosis.

Posaconazole's role in the management of blastomycosis is unclear. Case reports have detailed success in the management of osseous blastomycosis; however, until recently, only a pharmacokinetically unfavorable suspension has been available, and the lack of a preferable formulation has likely limited the drug's utility. Approval of delayed-release tablets and intravenous formulations may enhance the role of posazonazole in salvage therapy.

In a recent study including only three patients, isavuconazole, the most recent addition to the azole antifungal class, showed variable clinical efficacy (i.e., a 33% success rate) in the management of pulmonary and non-CNS disseminated blastomycosis. Its role in the management of blastomycosis is uncertain at this time, but its consideration may be warranted for salvage therapy in milder pulmonary disease.

TABLE 209-1 Treatment of Blastomycosis[a]

Disease	Primary Therapy	Alternative Therapy
Immunocompetent Patient/Life-Threatening Disease		
Pulmonary	Lipid AmB, 3–5 mg/kg qd, *or* AmB deoxycholate, 0.7–1.0 mg/kg qd (total dose: 1.5–2.5 g)	Itraconazole, 200–400 mg/d (once patient's condition has stabilized)
Disseminated CNS	Lipid AmB, 3–5 mg/kg qd, *or* AmB deoxycholate, 0.7–1.0 mg/kg qd (total dose: at least 2 g)	Fluconazole, 800 mg/d (if patient is intolerant to full course of AmB)
Non-CNS	Lipid AmB, 3–5 mg/kg qd, *or* AmB deoxycholate, 0.7–1.0 mg/kg qd (total dose: 1.5–2.5 g)	Itraconazole, 200–400 mg/d (once patient's condition has stabilized)
Immunocompetent Patient/Non-Life-Threatening Disease		
Pulmonary or disseminated (non-CNS)	Itraconazole, 200–400 mg/d, *or* Lipid AmB, 3–5 mg/kg qd, *or* AmB deoxycholate, 0.5–0.7 mg/kg qd (in patients intolerant to itraconazole or whose disease progresses despite therapy)	Fluconazole, 400–800 mg/d, *or* Ketoconazole, 400–800 mg/d
Immunocompromised Patient[b]		
All infections	Lipid AmB, 3–5 mg/kg qd, *or* AmB deoxycholate, 0.7–1.0 mg/kg qd (total dose: 1.5–2.5 g)	Itraconazole, 200–400 mg/d (non-CNS disease, once clinically improved)

[a]Therapy is generally given for 6–12 months. For bone and joint disease, a 12-month course is usually necessary. [b]Suppressive therapy with itraconazole may be considered for patients whose immunocompromised state continues. Fluconazole (800 mg/d) may be useful for patients who have CNS disease or cannot tolerate itraconazole.

Abbreviations: AmB, amphotericin B; CNS, central nervous system.

PROGNOSIS

Cure rates are 90–95% among compliant immunocompetent patients given itraconazole for mild to moderate pulmonary and extrapulmonary disease without CNS involvement. Bone and joint disease usually requires 12 months of therapy. The fewer than 5% of infections that relapse after an initial course of itraconazole usually respond well to a second treatment course.

GLOBAL CONSIDERATIONS

 *Blastomyces* remains a prominent cause of dimorphic fungal infections worldwide. In the United States, changes in mandatory reporting requirements have probably blurred estimates of the current incidence of blastomycosis. Diagnostic delays continue to contribute to increased use of health care resources and administration of unnecessary antibacterial courses and probably contribute to global resistance as well.

ACKNOWLEDGMENT

The authors thank Drs. Stanley W. Chapman and Donna C. Sullivan, Professors Emeriti, University of Mississippi, for their continued help and support and for their contributions to this chapter in an earlier edition.

FURTHER READING

Richer SM et al: Development of a highly sensitive and specific blastomycosis antibody enzyme immunoassay using *Blastomyces dermatitidis* surface protein BAD-1. Clin Vaccine Immunol 21:143, 2014. [PubMed: 24285817]
Thompson GR 3rd: Isavuconazole treatment of cryptococcosis and dimorphic mycoses. Clin Infect Dis 63:356, 2016. [PubMed: 27169478]

Chapter 210: Cryptococcosis

Arturo Casadevall

DEFINITION AND ETIOLOGY

Cryptococcus, a genus of yeast-like fungi, is the etiologic agent of cryptococcosis. Until recently, cryptococcal strains were separated into two species, *Cryptococcus neoformans* and *Cryptococcus gattii*, both of which can cause cryptococcosis in humans. The two varieties of *C. neoformans—grubii* and *neoformans*—correlate with serotypes A and D, respectively. *C. gattii*, although not divided into varieties, also is antigenically diverse, encompassing serotypes B and C. However, genome sequencing studies have now revealed tremendous diversity among isolates previously assigned to each species, suggesting that some may be reclassified as new species. Most clinical microbiology laboratories do not routinely distinguish between *C. neoformans* and *C. gattii* or among varieties, but rather identify and report all isolates simply as *C. neoformans*.

EPIDEMIOLOGY

Cryptococcosis was first described in the 1890s but remained relatively rare until the mid-twentieth century, when advances in diagnosis and increases in the number of immunosuppressed individuals markedly raised its reported prevalence. Although serologic evidence of cryptococcal *infection* is common among immunocompetent individuals, cryptococcal *disease* (cryptococcosis) is relatively rare in the absence of impaired immunity. Individuals at high risk for disease due to *C. neoformans* include patients with hematologic malignancies, recipients of solid organ transplants who require ongoing immunosuppressive therapy, persons whose medical conditions necessitate glucocorticoid therapy, and patients with advanced HIV infection and CD4+ T lymphocyte counts of <200/μL. In contrast, *C. gattii*–related disease is not associated with specific immune deficits and often occurs in immunocompetent individuals.

 Cryptococcal infection is acquired from the environment. *C. neoformans* and *C. gattii* inhabit different ecologic niches. *C. neoformans* is frequently found in soils contaminated with avian excreta and can easily be recovered from shaded and humid soils contaminated with pigeon droppings. In contrast, *C. gattii* is not found in bird feces. Instead, it inhabits a variety of arboreal species, including several types of eucalyptus tree. *C. neoformans* strains are found throughout the world; however, var. *grubii* (serotype A) strains are far more common than var. *neoformans* (serotype D) strains among both clinical and environmental isolates. The geographic distribution of *C. gattii* was thought to be largely limited to tropical regions until an outbreak of cryptococcosis caused by a new serotype B strain began in Vancouver in 1999. This outbreak has extended into the United States, and *C. gattii* infections are being encountered increasingly in several states in the Pacific Northwest.

The global burden of cryptococcosis was recently estimated at ~1 million cases, with >600,000 deaths annually. Thus cryptococci are important human pathogens. Since the onset of the HIV pandemic in the early 1980s, the overwhelming majority of cryptococcosis cases have occurred in patients with AIDS **(Chap. 197)**. To comprehend the impact of HIV infection on the epidemiology of cryptococcosis, it is instructive to note that in the early 1990s there were >1000 cases of cryptococcal meningitis each

year in New York City—a figure far exceeding that for all cases of bacterial meningitis. With the advent of effective antiretroviral therapy, the incidence of AIDS-related cryptococcosis has been sharply reduced among treated individuals. Therefore, most cases of cryptococcosis now occur in resource-limited regions of the world. The disease remains distressingly common in regions where antiretroviral therapy is not readily available (e.g., parts of Africa and Asia); in these regions, up to one-third of patients with AIDS have cryptococcosis. Among HIV-infected persons, those with a decreased percentage of memory B cells expressing IgM may be at greater risk for cryptococcosis.

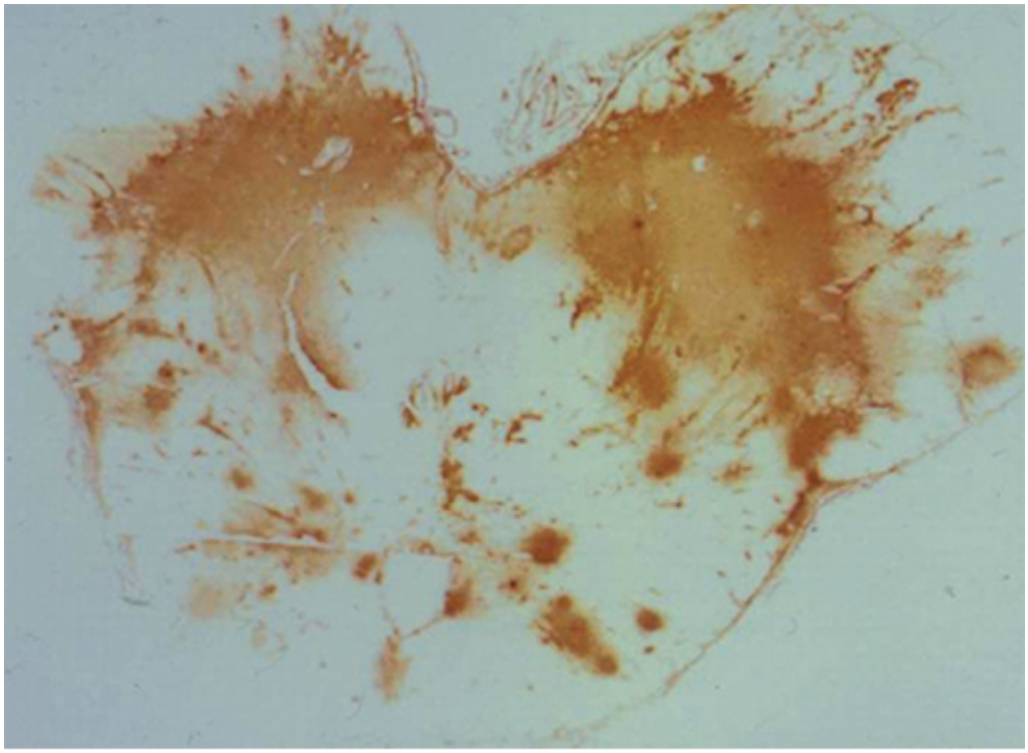
PATHOGENESIS

Cryptococcal infection is acquired by inhalation of aerosolized infectious particles. The exact nature of these particles is not known; the two leading candidate forms are small desiccated yeast cells and basidiospores. Little is known about the pathogenesis of initial infection. Serologic studies have shown that cryptococcal infection is acquired in childhood, but it is not known whether the initial infection is symptomatic. Given that cryptococcal infection is common while disease is rare, the consensus is that pulmonary defense mechanisms in immunologically intact individuals are highly effective at containing this fungus. It is not clear whether initial infection leads to a state of immunity or whether most individuals are subject throughout life to frequent and recurrent infections that resolve without clinical disease. However, evidence indicates that some human cryptococcal infections lead to a state of latency in which viable organisms are harbored for prolonged periods, possibly in granulomas. Thus the inhalation of cryptococcal cells and/or spores can be followed by either clearance or establishment of the latent state. The consequences of prolonged harboring of cryptococcal cells in the lung are not known, but evidence from animal studies indicates that the organisms' prolonged presence could alter the immunologic milieu in the lung and predispose to allergic airway disease.

Cryptococcosis usually presents clinically as chronic meningoencephalitis. The mechanisms by which the fungus undergoes extrapulmonary dissemination and enters the central nervous system (CNS) remain poorly understood. The mechanism by which cryptococcal cells cross the blood–brain barrier is a subject of intensive study. Current evidence suggests that both direct fungal-cell migration across the endothelium and fungal-cell carriage inside macrophages as "Trojan horse" invaders can occur. *Cryptococcus* species have well-defined virulence factors that include the expression of the polysaccharide capsule, the ability to make melanin, and the elaboration of enzymes (e.g., phospholipase and urease) that enhance the survival of fungal cells in tissue. Among these virulence factors, the capsule and melanin production have been most extensively studied. The cryptococcal capsule is antiphagocytic, and the capsular polysaccharide has been associated with numerous deleterious effects on host immune function. Cryptococcal infections can elicit little or no tissue inflammatory response. The immune dysfunction seen in cryptococcosis has been attributed to the release of copious amounts of capsular polysaccharide into tissues, where it probably interferes with local immune responses **(Fig. 210-1)**. In clinical practice, the capsular polysaccharide is the antigen that is measured as a diagnostic marker of cryptococcal infection.

FIGURE 210-1

Cryptococcal antigen in human brain tissue, as revealed by immunohistochemical staining. Brown areas show polysaccharide deposits in the midbrain of a patient who died of cryptococcal meningitis. *(Reprinted with permission from SC Lee et al: Am J Pathol 148; 1267, 1996.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Cryptococcosis

Cryptococcosis should be included in the differential diagnosis when any patient presents with findings suggestive of chronic meningitis. Concern about cryptococcosis is heightened by a history of headache and neurologic symptoms in a patient with an underlying immunosuppressive disorder or state that is associated with an increased incidence of cryptococcosis, such as advanced HIV infection or solid organ transplantation.

CLINICAL MANIFESTATIONS

The clinical manifestations of cryptococcosis reflect the site of fungal infection. The spectrum of disease caused by *Cryptococcus* species consists predominantly of meningoencephalitis and pneumonia, but skin and soft tissue infections also occur; in fact, cryptococcosis can affect any tissue or organ. CNS involvement usually presents as signs and symptoms of chronic meningitis, such as headache, fever, lethargy, sensory deficits, memory deficits, cranial nerve paresis, vision deficits, and meningismus. Cryptococcal meningitis differs from bacterial meningitis in that many *Cryptococcus*-infected patients present with symptoms of several weeks' duration. In addition, classic characteristics of meningeal irritation, such as meningismus, may be absent in cryptococcal meningitis. Indolent cases can present as subacute dementia. Meningeal cryptococcosis can lead to sudden catastrophic vision loss.

Pulmonary cryptococcosis usually presents as cough, increased sputum production, and chest pain. Patients infected with *C. gattii* can present with granulomatous pulmonary masses known as *cryptococcomas*. Fever develops in a minority of cases. Like CNS disease, pulmonary cryptococcosis can follow an indolent course, and the majority of cases probably do not come to clinical attention. In fact, many cases are discovered incidentally during the workup of an abnormal chest radiograph

obtained for other diagnostic purposes. Pulmonary cryptococcosis can be associated with antecedent diseases such as malignancy, diabetes, and tuberculosis.

Skin lesions are common in patients with disseminated cryptococcosis and can be highly variable, including papules, plaques, purpura, vesicles, tumor-like lesions, and rashes. The spectrum of cryptococcosis in HIV-infected patients is so varied and has changed so much since the advent of antiretroviral therapy that a distinction between HIV-related and HIV-unrelated cryptococcosis is no longer pertinent. In patients with AIDS and solid organ transplant recipients, the lesions of cutaneous cryptococcosis often resemble those of molluscum contagiosum **(Fig. 210-2; Chap. 191)**.

FIGURE 210-2

Disseminated fungal infection. A liver transplant recipient developed six cutaneous lesions similar to the one shown. Biopsy and serum antigen testing demonstrated *Cryptococcus*. Important features of the lesion include a benign-appearing fleshy papule with central umbilication resembling molluscum contagiosum. *(Photo courtesy of Dr. Lindsey Baden; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS

A diagnosis of cryptococcosis requires the demonstration of yeast cells in normally sterile tissues. Visualization of the capsule of fungal cells in cerebrospinal fluid (CSF) mixed with India ink is a useful rapid diagnostic technique. Cryptococcal cells in India ink have a distinctive appearance because their capsules exclude ink particles. However, the CSF India ink examination may yield negative results in patients with a low fungal burden. This examination should be performed by a trained individual, since leukocytes and fat globules can sometimes be mistaken for fungal cells. Cultures of CSF and blood that are positive for cryptococcal cells are diagnostic for cryptococcosis. In cryptococcal meningitis, CSF examination usually reveals evidence of chronic meningitis with mononuclear cell pleocytosis

and increased protein levels. A particularly useful test is cryptococcal antigen (CRAg) detection in CSF and blood. The assay is based on serologic detection of cryptococcal polysaccharide and is both sensitive and specific. A positive CRAg test provides strong presumptive evidence for cryptococcosis; however, because the result is often negative in pulmonary cryptococcosis, the test is less useful in the diagnosis of pulmonary disease and is of only limited usefulness in monitoring the response to therapy.

 In areas of Africa where there is a high prevalence of HIV infection, routine screening of blood for CRAg in HIV-infected patients with low CD4+ T lymphocyte counts may identify individuals at high risk of cryptococcal disease who are candidates for antifungal therapy. CRAg screening has shown that a significant proportion of HIV-infected patients hospitalized with pneumonia in Thailand harbor cryptococcal infection. Inexpensive point-of-care CRAg tests are under development and could be of great diagnostic benefit in resource-limited regions.

TREATMENT

TREATMENT Cryptococcosis

Both the site of infection and the immune status of the host must be considered in the selection of therapy for cryptococcosis. The disease has two general patterns of manifestation: (1) pulmonary cryptococcosis, with no evidence of extrapulmonary dissemination; and (2) extrapulmonary (systemic) cryptococcosis, with or without meningoencephalitis. Pulmonary cryptococcosis in an immunocompetent host sometimes resolves without therapy. However, given the propensity of *Cryptococcus* species to disseminate from the lung, the inability to gauge the host's immune status precisely, and the availability of low-toxicity therapy in the form of fluconazole, the current recommendation is for pulmonary cryptococcosis in an immunocompetent individual to be treated with fluconazole (200–400 mg/d for 3–6 months). Extrapulmonary cryptococcosis without CNS involvement in an immunocompetent host can be treated with the same regimen, although amphotericin B (AmB; 0.5–1 mg/kg daily for 4–6 weeks) may be required for more severe cases. In general, extrapulmonary cryptococcosis without CNS involvement requires less intensive therapy, with the caveat that morbidity and death in cryptococcosis are associated with meningeal involvement. Thus the decision to categorize cryptococcosis as "extrapulmonary without CNS involvement" should be made only after careful evaluation of the CSF reveals no evidence of cryptococcal infection. For CNS involvement in a host without AIDS or obvious immune impairment, most authorities recommend initial therapy with AmB (0.5–1 mg/kg daily) during an induction phase, which is followed by prolonged therapy with fluconazole (400 mg/d) during a consolidation phase. For cryptococcal meningoencephalitis without a concomitant immunosuppressive condition, the recommended regimen is AmB (0.5–1 mg/kg) plus flucytosine (100 mg/kg) daily for 6–10 weeks. Alternatively, patients can be treated with AmB (0.5–1 mg/kg) plus flucytosine (100 mg/kg) daily for 2 weeks and then with fluconazole (400 mg/d) for at least 10 weeks. Patients with immunosuppression are treated with the same initial regimens except that consolidation therapy with fluconazole is given for a prolonged period to prevent relapse.

Cryptococcosis in patients with HIV infection always requires aggressive therapy and is considered incurable unless immune function improves. Consequently, therapy for cryptococcosis in the setting of AIDS has two phases: induction therapy (intended to reduce the fungal burden and alleviate symptoms) and lifelong maintenance therapy (to prevent a symptomatic clinical relapse). Pulmonary and extrapulmonary cryptococcosis without evidence of CNS involvement can be treated with fluconazole (200–400 mg/d). In patients who have more extensive disease, flucytosine (100 mg/kg per day) may be added to the fluconazole regimen for 10 weeks, with lifelong fluconazole maintenance therapy thereafter. For HIV-infected patients with evidence of CNS involvement, most authorities recommend induction therapy with AmB. An acceptable regimen is AmB (0.7–1 mg/kg) plus flucytosine (100 mg/kg) daily for 2 weeks followed by fluconazole (400 mg/d) for at least 10 weeks and then by lifelong maintenance therapy with fluconazole (200 mg/d). Fluconazole (400–800 mg/d) plus

flucytosine (100 mg/kg per day) for 6–10 weeks followed by fluconazole (200 mg/d) as maintenance therapy is an alternative. Newer triazoles like voriconazole and posaconazole are highly active against cryptococcal strains and appear to be clinically effective, but clinical experience with these agents in the treatment of cryptococcosis is limited. Lipid formulations of AmB can be substituted for AmB deoxycholate in patients with renal impairment. Neither caspofungin nor micafungin is effective against *Cryptococcus* species; consequently, neither drug has a role in the treatment of cryptococcosis. Cryptococcal meningoencephalitis is often associated with increased intracranial pressure, which is believed to be responsible for damage to the brain and cranial nerves. Appropriate management of CNS cryptococcosis requires careful attention to the management of intracranial pressure, including the reduction of pressure by repeated therapeutic lumbar puncture and the placement of shunts. Studies suggest that the addition of a short course of interferon γ to antifungal therapy in patients with HIV infection increases clearance of cryptococci from the CSF.

In HIV-infected patients with previously treated cryptococcosis who are receiving fluconazole maintenance therapy, it may be possible to discontinue antifungal drug treatment if antiretroviral therapy results in immunologic improvement.

PROGNOSIS AND COMPLICATIONS

Even with antifungal therapy, cryptococcosis is associated with high rates of morbidity and death. For the majority of patients with cryptococcosis, the most important prognostic factors are the extent and the duration of the underlying immunologic deficits that predisposed them to develop the disease. Therefore, cryptococcosis is often curable with antifungal therapy in individuals with no apparent immunologic dysfunction, but, in patients with severe immunosuppression (e.g., those with AIDS), the best that can be hoped for is that antifungal therapy will induce remission, which can then be maintained with lifelong suppressive therapy. Before the advent of antiretroviral therapy, the median overall survival period for AIDS patients with cryptococcosis was <1 year. Cryptococcosis in patients with underlying neoplastic disease has a particularly poor prognosis. For CNS cryptococcosis, poor prognostic markers are a CSF assay positive for yeast cells on initial India ink examination (evidence of a heavy fungal burden), high CSF pressure, low CSF glucose levels, low CSF pleocytosis (<2/μL), recovery of yeast cells from extraneural sites, absence of antibody to capsular polysaccharide, a CSF or serum cryptococcal antigen level of ≥1:32, and concomitant glucocorticoid therapy or hematologic malignancy. A response to treatment does not guarantee cure since relapse of cryptococcosis is common even among patients with relatively intact immune systems. Complications of CNS cryptococcosis include cranial nerve deficits, vision loss, and cognitive impairment.

IMMUNE RECONSTITUTION INFLAMMATORY SYNDROME

The frequent chronicity of cryptococcal infections and their common occurrence in settings of changing immunity can result in new clinical syndromes, such as the immune reconstitution inflammatory syndrome (IRIS). IRIS occurs when immunity rebounds in the setting of treated cryptococcosis (or an undiagnosed asymptomatic infection) and the immune response to cryptococcal antigens in tissue triggers an inflammatory response that can be difficult to distinguish from a relapsing infection. IRIS can occur when patients with AIDS and treated cryptococcosis are given antiretroviral therapy that results in improved immunity. Apart from the difficulties in distinguishing IRIS from cryptococcal relapse, the management of this syndrome is complex because it is caused by the desirable outcome of improving immunity, which is important in controlling cryptococcal infection and preventing relapses. The approach to the patient with IRIS must attempt to balance resurgent immunity against immune-mediated damage. Currently, management of IRIS is individualized and can involve the use of glucocorticoids to reduce inflammation.

PREVENTION

No vaccine is available for cryptococcosis. In patients at high risk (e.g., those with advanced HIV infection and CD4+ T lymphocyte counts of <200/μL), primary prophylaxis with fluconazole (200 mg/d) is effective in reducing the prevalence of disease. Since antiretroviral therapy raises the CD4+ T lymphocyte count, it constitutes an immunologic form of prophylaxis.

FURTHER READING

Kwon-Chung KJ et al: The case for adopting the “species complex” nomenclature for the etiologic agents of cryptococcosis. mSphere 2 pii:e00357, 2017.
Maziarz EK, Perfect JR: Cryptococcosis. Infect Dis Clin North Am 30:179, 2016. [PubMed: 26897067]
Robertson EJ et al: *Cryptococcus neoformans* ex vivo capsule size is associated with intracranial pressure and host immune response in HIV-associated cryptococcal meningitis. J Infect Dis 209:74, 2014. [PubMed: 23945372]
Saag MS et al: Practice guidelines for the management of cryptococcal disease. Infectious Diseases Society of America. Clin Infect Dis 30:710, 2000. [PubMed: 10770733]
Srichatrapimuk S, Sungkanuparph S: Integrated therapy for HIV and cryptococcosis. AIDS Res Ther 13:42, 2016. [PubMed: 27906037]

Chapter 211: Candidiasis

John E. Edwards, Jr.

INTRODUCTION

The genus *Candida* encompasses >150 species, only a few of which cause disease in humans. With rare exceptions (although the exceptions are increasing in number), the human pathogens are *C. albicans*, *C. guilliermondii*, *C. krusei*, *C. parapsilosis*, *C. tropicalis*, *C. kefyr*, *C. lusitaniae*, *C. dubliniensis*, *C. glabrata*, and *C. auris.* Ubiquitous in nature, they inhabit the gastrointestinal tract (including the mouth and oropharynx), the female genital tract, and the skin. Although cases of candidiasis have been described since antiquity in debilitated patients, the advent of *Candida* species as common human pathogens dates to the introduction of modern therapeutic approaches that suppress normal host-defense mechanisms. Of these relatively recent advances, the most important is the use of antibacterial agents that alter the normal human microbiota and allow nonbacterial species to become more prevalent in the commensal flora. With the introduction of antifungal agents, the causes of *Candida* infections shifted from an almost complete dominance of *C. albicans* to the common involvement of *C. glabrata* and the other species listed above. The non-*albicans* species now account for approximately half of all cases of candidemia and hematogenously disseminated candidiasis. Recognition of this change is clinically important, since the various species differ in susceptibility to the newer antifungal agents.

Candida is a small, thin-walled, ovoid yeast that measures 4–6 μm in diameter and reproduces by budding. Organisms of this genus occur in three forms in tissue: blastospores, pseudohyphae, and hyphae. *Candida* grows readily on simple medium; lysis centrifugation enhances its recovery from blood. Species are identified by biochemical testing (currently with automated devices) or on special agar (e.g., CHROMagar).

EPIDEMIOLOGY

 *Candida* are present in humans as commensals, in animals, in foods, and on inanimate objects. In developed countries, where contemporary medical therapeutics are commonly used, *Candida* species are now among the most common nosocomial pathogens. In the United States, these species are the fourth most common isolates from the blood of hospitalized patients. In a recent point-prevalence study, *Candida* species were the most common organisms infecting the bloodstream of hospitalized patients. In regions where advanced medical care is rarely available, mucocutaneous *Candida* infections, such as thrush, are more common than deep-organ infections, which rarely occur. However, the incidence of deep-organ candidiasis increases steadily as advances in health care—such as therapy with broad-spectrum antibiotics, more aggressive treatment of cancer, and the use of immunosuppression for sustaining organ transplants—are implemented. In aggregate, the global incidence of infections due to *Candida* species has risen steadily over the past few decades. Of great recent concern has been the global emergence of *C. auris*; certain strains of this organism are resistant to all classes of antifungal agents, and mortality rates from infection have been very high.

PATHOGENESIS

In the most serious form of *Candida* infection, the organisms disseminate hematogenously and form microabscesses and small macroabscesses in major organs. Although the exact mechanism is not known, *Candida* probably enters the bloodstream from mucosal surfaces after growing to large numbers as a consequence of bacterial suppression by antibacterial drugs; alternatively, in some instances, the organism may enter from the skin. A change from the blastospore stage to the pseudohyphal and hyphal stages is generally considered integral to *Candida*'s penetration into tissue. However, *C. glabrata* can cause extensive infection even though it does not transform into pseudohyphae or hyphae. Adherence to both epithelial and endothelial cells is thought to be the first step in invasion and infection; several adhesins have been identified as well as a mucosal toxin, candidalysin. Biofilm formation also is considered important in pathogenesis. Numerous reviews of cases of hematogenously disseminated candidiasis have identified the predisposing factors or conditions associated with disseminated disease **(Table 211-1)**. Women who receive antibacterial agents may develop vaginal candidiasis.

TABLE 211-1 Well-Recognized Factors and Conditions Predisposing to Hematogenously Disseminated Candidiasis

Antibacterial agents	Abdominal and thoracic surgery
Indwelling intravenous catheters	Cytotoxic chemotherapy
Hyperalimentation fluids	Immunosuppressive agents for organ transplantation
Indwelling urinary catheters	Respirators
Parenteral glucocorticoids	Neutropenia
Severe burns	Low birth weight (neonates)
HIV-associated low CD4+ T-cell counts	Diabetes

Innate immunity is the most important defense mechanism against hematogenously disseminated candidiasis, and the neutrophil is the most important component of this defense. Macrophages also play an important defensive role. STAT1, Dectin-1, CARD9, and T_H1 and T_H17 lymphocytes contribute significantly to innate defense (see "Clinical Manifestations," below). Although many immunocompetent individuals have antibodies to *Candida*, the role of these antibodies in defense against the organism is not clear. Multiple genetic polymorphisms that predispose to disseminated candidiasis will most likely be identified in future studies.

CLINICAL MANIFESTATIONS

Mucocutaneous Candidiasis

Thrush is characterized by white, adherent, painless, discrete or confluent patches in the mouth, on the tongue, or in the esophagus, occasionally with fissuring at the corners of the mouth. This form of disease caused by *Candida* can also occur at points of contact with dentures. Organisms are identifiable in gram-stained scrapings from lesions. The occurrence of thrush in a young, otherwise healthy-appearing person should prompt an investigation for underlying HIV infection. More commonly, thrush is seen as a nonspecific manifestation of severe debilitating illness. Vulvovaginal candidiasis is accompanied by pruritus, pain, and vaginal discharge, which is usually thin but may contain whitish "curds" in severe cases. A subset of patients with recurrent vulvovaginitis have a deficiency in the surface expression of Dectin-1, a major recognition factor for β-glucan on *Candida*.

This deficiency leads to suboptimal functioning of the CARD9 pathway, which ultimately increases the propensity for recurrent vaginal infections.

Other *Candida* skin infections include *paronychia*, a painful swelling at the nail–skin interface; *onychomycosis*, a fungal nail infection rarely caused by this genus; *intertrigo*, an erythematous irritation with redness and pustules in the skin folds; *balanitis*, an erythematous-pustular infection of the glans penis; *erosio interdigitalis blastomycetica*, an infection between the digits of the hands or toes; *folliculitis*, with pustules developing most frequently in the area of the beard; *perianal candidiasis*, a pruritic, erythematous, pustular infection surrounding the anus; and *diaper rash*, a common erythematous, pustular perineal infection in infants. *Generalized disseminated cutaneous candidiasis*, another form of infection that occurs primarily in infants, is characterized by widespread eruptions over the trunk, thorax, and extremities. The diagnostic macronodular lesions of hematogenously disseminated candidiasis **(Fig. 211-1)** indicate a high probability of dissemination to multiple organs as well as the skin. While the lesions are seen predominantly in immunocompromised patients treated with cytotoxic drugs, they may also develop in patients without neutropenia.

FIGURE 211-1

Macronodular skin lesions associated with hematogenously disseminated candidiasis. *Candida* organisms are usually but not always visible on histopathologic examination. The fungi grow when a portion of the biopsied specimen is cultured. Therefore, for optimal identification, both histopathology and culture should be performed. *(Image courtesy of Dr. Noah Craft and the Victor Newcomer collection at UCLA, archived by Logical Images, Inc.; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Chronic mucocutaneous candidiasis is a heterogeneous infection of the hair, nails, skin, and mucous membranes that persists despite intermittent therapy. The onset of disease usually comes in infancy or within the first two decades of life, but in rare cases comes in later life. The condition may be mild and limited to a specific area of the skin or nails, or it may take a severely disfiguring form (*Candida*

granuloma) characterized by exophytic outgrowths on the skin. Chronic mucocutaneous candidiasis is usually associated with specific immunologic dysfunction; most frequently reported is a failure of T lymphocytes to proliferate or to excrete cytokines in response to stimulation by *Candida* antigens in vitro. A subset of the affected patients have mutations in the STAT1 gene resulting in an insufficiency of interferon γ, interleukin 17, and interleukin 22.

Approximately half of patients with chronic mucocutaneous candidiasis have associated endocrine abnormalities that together are designated the *a*utoimmune *p*oly*e*ndocrinopathy–*c*andidiasis–*e*ctodermal *d*ystrophy (APECED) syndrome. This syndrome is due to mutations in the autoimmune regulator (*AIRE*) gene and is most prevalent among Finns, Iranian Jews, Sardinians, northern Italians, and Swedes. Conditions that usually follow the onset of the disease include hypoparathyroidism, adrenal insufficiency, autoimmune thyroiditis, Graves' disease, chronic active hepatitis, alopecia, juvenile-onset pernicious anemia, malabsorption, and primary hypogonadism. In addition, dental enamel dysplasia, vitiligo, pitted nail dystrophy, and calcification of the tympanic membranes may occur. Patients with chronic mucocutaneous candidiasis rarely develop hematogenously disseminated candidiasis, probably because their neutrophil function remains intact.

Deeply Invasive Candidiasis

Deeply invasive *Candida* infections may or may not be due to hematogenous seeding. Deep esophageal infection may result from penetration by organisms from superficial esophageal erosions; joint or deep-wound infection from contiguous spread of organisms from the skin; kidney infection from catheter-initiated spread of organisms through the urinary tract; infection of intraabdominal organs and the peritoneum from perforation of the gastrointestinal tract; and gallbladder infection from retrograde migration of organisms from the gastrointestinal tract into the biliary drainage system.

However, far more commonly, deeply invasive candidiasis results from hematogenous seeding of various organs as a complication of candidemia. Once the organism gains access to the intravascular compartment (either from the gastrointestinal tract or, less often, from the skin through the site of an indwelling intravascular catheter), it may spread hematogenously to a variety of deep organs. The brain, chorioretina **(Fig. 211-2)**, heart, and kidneys are most commonly infected and the liver and spleen less commonly so (most often in neutropenic patients). In fact, nearly any organ can become involved, including the endocrine glands, pancreas, heart valves (native or prosthetic), skeletal muscle, joints (native or prosthetic), bones, and meninges. *Candida* organisms can also spread hematogenously to the skin and cause classic macronodular lesions (Fig. 211-1). Frequently, painful muscular involvement is evident beneath the area of affected skin. Chorioretinal involvement and skin involvement are highly significant, since both findings are associated with a very high probability of abscess formation in multiple deep organs as a result of generalized hematogenous seeding. Ocular involvement (Fig. 211-2) may require specific treatment (e.g., partial vitrectomy or intraocular injection of antifungal agents) to prevent permanent blindness. An ocular examination is indicated for all patients with candidemia, whether or not they have ocular manifestations.

FIGURE 211-2

Hematogenous *Candida* endophthalmitis. A classic off-white lesion projecting from the chorioretina into the vitreous causes the surrounding haze. The lesion is composed primarily of inflammatory cells rather than organisms. Lesions of this type may progress to cause extensive vitreal inflammation and eventual loss of the eye. Partial vitrectomy, combined with IV and possibly intravitreal antifungal therapy, may be helpful in controlling the lesions. *(Image courtesy of Dr. Gary Holland; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS

The diagnosis of *Candida* infection is established by visualization of pseudohyphae or hyphae on wet mount (saline and 10% KOH), tissue Gram's stain, periodic acid–Schiff stain, or methenamine silver stain in the presence of inflammation. Absence of organisms on hematoxylin-eosin staining does not reliably exclude *Candida* infection. The most challenging aspect of diagnosis is determining which patients with *Candida* isolates have hematogenously disseminated candidiasis. For instance, recovery of *Candida* from sputum, urine, or peritoneal catheters may indicate mere colonization rather than deep-seated infection, and *Candida* isolation from the blood of patients with indwelling intravascular catheters may reflect inconsequential seeding of the blood from or growth of the organisms on the catheter. Despite extensive research into both antigen and antibody detection systems, there is currently no widely available and validated diagnostic test to distinguish patients with inconsequential seeding of the blood from those whose positive blood cultures represent hematogenous dissemination to multiple organs. Many studies are under way to establish the utility of the β-glucan test; at present, its greatest utility is its negative predictive value (~90%). Meanwhile, the presence of ocular or macronodular skin lesions is highly suggestive of widespread infection of multiple deep organs. Despite extensive tests for hematogenous dissemination, such as polymerase chain reaction and T2 technology, no test is fully validated or widely available at present. Matrix-assisted laser desorption–ionization–time-of-flight mass spectometry (MALDI-TOF MS) will likely be used extensively for detection and speciation in the future.

TREATMENT

TREATMENT *Candida* Infections MUCOCUTANEOUS *CANDIDA* INFECTION

The treatment of mucocutaneous candidiasis is summarized in **Table 211-2**.

CANDIDEMIA AND SUSPECTED HEMATOGENOUSLY DISSEMINATED CANDIDIASIS

All patients with candidemia are treated with a systemic antifungal agent. A certain percentage of patients, including many of those who have candidemia associated with an indwelling intravascular catheter, probably have “benign” candidemia rather than deep-organ seeding. However, because there is no reliable way to distinguish benign candidemia from deep-organ infection, and because antifungal drugs less toxic than amphotericin B are available, antifungal treatment for candidemia—with or without clinical evidence of deep-organ involvement—has become the standard of practice. In addition, if an indwelling intravascular catheter is present, it is best to remove or replace the device whenever feasible.

The drugs used for the treatment of candidemia and suspected disseminated candidiasis are listed in **Table 211-3**. Various lipid formulations of amphotericin B, three echinocandins, and the azoles fluconazole and voriconazole are used; no agent within a given class has been clearly identified as superior to the others. Most institutions choose an agent from each class on the basis of their own specific microbial epidemiology, strategies to minimize toxicities, and cost considerations. There is a trend to treat with an echinocandin until sensitivities or speciation is determined. In stable patients, many centers then switch to fluconazole if a sensitive strain is identified and there is no evidence of hematogenous disemination. For hemodynamically unstable or neutropenic patients, initial treatment with broader-spectrum agents is desirable; these drugs include polyenes, echinocandins, or later-generation azoles such as voriconazole. Once the clinical response has been assessed and the pathogen specifically identified, the regimen can be altered accordingly. At present, the vast majority of *C. albicans* isolates are sensitive to fluconazole. Isolates of *C. glabrata* and *C. krusei* are less sensitive to fluconazole and more sensitive to polyenes and echinocandins. *C. parapsilosis* is less sensitive to echinocandins in vitro; however, this lesser sensitivity is considered insignificant. Posaconazole has been approved for prophylaxis, including that against *Candida*, in neutropenic patients. Itraconazole is rarely used for *Candida*, and isavuconazole has not been approved for this indication to date.

Some generalizations exist regarding the management of specific *Candida* infections. Recovery of *Candida* from sputum is almost never indicative of underlying pulmonary candidiasis and does not by itself warrant antifungal treatment. Similarly, *Candida* in the urine of a patient with an indwelling bladder catheter may represent colonization only rather than bladder or kidney infection; however, the threshold for systemic treatment is lower in severely ill patients in this category since it is impossible to distinguish colonization from lower or upper urinary tract infection. If the isolate is *C. albicans*, most clinicians use oral fluconazole rather than a bladder washout with amphotericin B, which was more commonly used in the past. Caspofungin has been used with success; although echinocandins are poorly excreted into the urine, they may be an option, especially for non-*albicans* isolates. The doses and duration are the same as for disseminated candidiasis. The significance of the recovery of *Candida* from abdominal drains in postoperative patients is unclear, but again the threshold for treatment is generally low because most of the affected patients have been subjected to factors predisposing to disseminated candidiasis. In addition, there has been a considerable increase in the recognition and diagnosis of intraabdominal candidiasis.

Removal of the infected valve and long-term antifungal administration constitute appropriate treatment for *Candida* endocarditis. Although definitive studies are not available, patients usually are treated for weeks with a systemic antifungal agent (Table 211-3) and then given chronic suppressive therapy for months or years (sometimes indefinitely) with an oral azole (usually fluconazole at 400–800 mg/d).

Hematogenous *Candida* endophthalmitis is a special problem requiring ophthalmologic consultation. When lesions are expanding or are threatening the macula, an IV polyene combined with flucytosine (25 mg/kg four times daily) has been the regimen of choice, although comparative studies with other regimens have not yet been reported. As more data on the azoles (e.g., voriconazole) and the echinocandins become available, new strategies involving these agents are developing. Of paramount importance is the decision to perform a partial vitrectomy. This procedure debulks the infection and can preserve sight, which may otherwise be lost due to vitreal scarring. All patients with candidemia

should undergo ophthalmologic examination because of the relatively high frequency of this ocular complication. Not only can this examination detect a developing eye lesion early in its course; in addition, identification of a lesion signifies a probability of ~90% of deep-organ abscesses and may prompt prolongation of therapy for candidemia beyond the recommended 2 weeks after the last positive blood culture. Although the basis for the consensus is a very small data set, the recommended treatment for *Candida* meningitis is a polyene (Table 211-3) plus flucytosine (25 mg/kg four times daily). Successful treatment of *Candida*-infected prosthetic material (e.g., an artificial joint) nearly always requires removal of the infected material followed by long-term administration of an antifungal agent selected on the basis of the isolate's sensitivity and the logistics of administration.

TABLE 211-2 Treatment of Mucocutaneous Candidal Infections

Disease	Preferred Treatment	Alternatives
Cutaneous	Topical azole	Topical nystatin
Vulvovaginal	Oral fluconazole (150 mg) or azole cream or suppository	Nystatin suppository
Oral (thrush)	Clotrimazole troches	Nystatin, fluconazole
Esophageal	Fluconazole tablets (100–200 mg/d) or itraconazole solution (200 mg/d)	Caspofungin, micafungin, or amphotericin B

TABLE 211-3 Agents for the Treatment of Disseminated Candidiasis

Agent	Route of Administration	Dose[a]	Comment
Amphotericin B deoxycholate	IV only	0.5–1.0 mg/kg daily	Being replaced by lipid formulations
Amphotericin B lipid formulations			
Liposomal (AmBiSome, Abelcet)	IV only	3.0–5.0 mg/kg daily	Not approved as primary therapy by the U.S. Food and Drug Administration, but used commonly because less toxic than amphotericin B deoxycholate
Lipid complex (ABLC)	IV only	3.0–5.0 mg/kg daily	
Colloidal dispersion (ABCD)	IV only	3.0–5.0 mg/kg daily	Associated with frequent infusion reactions
Azoles[b]			
Posaconazole	IV and oral	300 mg/d (IV) 200 mg tid (oral)	Approved for prophylaxis
Fluconazole	IV and oral		Most commonly used
Voriconazole	IV and oral	400 mg/d	Multiple drug interactions
		400 mg/d	Approved for candidemia in nonneutropenic patients
Echinocandins			Broad spectrum against *Candida* species; approved for disseminated candidiasis
Caspofungin	IV only	50 mg/d	
Anidulafungin	IV only	100 mg/d	
Micafungin	IV only	100 mg/d	

[a]For loading doses and adjustments in renal failure, see Pappas PG et al: Clinical practice guidelines for the management of candidiasis: 2016 update by the Infectious Diseases Society of America. Clin Infect Dis 62:e1, 2016. The recommended duration of therapy is 2 weeks beyond the last positive blood culture and the resolution of signs and symptoms of infection. [b]Although ketoconazole is

approved for the treatment of disseminated candidiasis, it has been replaced by the newer agents listed in this table. Posaconazole has been approved for prophylaxis in neutropenic patients and for oropharyngeal candidiasis.

PROPHYLAXIS

The use of antifungal agents to prevent *Candida* infections has been controversial, but some general principles have emerged. Most centers administer prophylactic fluconazole (400 mg/d) to recipients of allogeneic stem cell transplants. High-risk liver transplant recipients also are given fluconazole prophylaxis in most centers. The use of prophylaxis for neutropenic patients has varied considerably from center to center; many centers that elect to give prophylaxis to this population use either fluconazole (200–400 mg/d) or a lipid formulation of amphotericin B (AmBiSome, 1–2 mg/d). Caspofungin (50 mg/d) also has been recommended. Some centers have used itraconazole suspension (200 mg/d). Posaconazole (200 mg three times daily) has been approved by the U.S. Food and Drug Administration for prophylaxis in neutropenic patients; it is gaining in popularity and may replace fluconazole.

Prophylaxis is sometimes given to surgical patients at very high risk. The widespread use of prophylaxis for nearly all patients in general surgical or medical intensive care units is not—and should not be—a common practice for three reasons: (1) the incidence of disseminated candidiasis is relatively low, (2) the cost–benefit ratio is suboptimal, and (3) increased resistance with widespread prophylaxis is a valid concern.

Prophylaxis for oropharyngeal or esophageal candidiasis in HIV-infected patients is not recommended unless there are frequent recurrences.

FURTHER READING

Edwards JE Jr: *Candida* species, in *Mandell, Douglas, and Bennett's Principles of Infectious Diseases*, 8th ed. JE Bennett et al (eds). Philadelphia, Elsevier, 2015, pp 2879–2894.
Pappas PG et al: Clinical practice guideline for the management of candidiasis. 2016 update by the Infectious Diseases Society of America. Clin Infect Dis 62:e1, 2016. [PubMed: 26679628]
Uppuluri P et al: Current trends in candidiasis, in *Candida albicans: Cellular and Molecular Biology*, 2nd ed. R Prasad (ed). Cham, Switzerland, Springer, 2017, pp 5–24.

Chapter 212: Aspergillosis

David W. Denning

INTRODUCTION

Aspergillosis is the collective term used to describe all disease entities caused by any one of ~50 pathogenic and allergenic species of *Aspergillus*. Only those species that grow at 37°C can cause invasive infection, although some species without this ability can cause allergic syndromes. Each common pathogenic species is actually a complex of many species (many of them cryptic), but is referred to as a single species here for simplicity. *A. fumigatus* is responsible for most cases of invasive aspergillosis, almost all cases of chronic aspergillosis, and most allergic syndromes. *A. flavus* is more prevalent in some hospitals and causes a higher proportion of cases of sinus infections, cutaneous infections, and keratitis than *A. fumigatus*. *A. niger* can cause invasive infection but more commonly colonizes the respiratory tract and causes external otitis. *A. terreus* causes only invasive disease, usually with a poor prognosis. *A. nidulans* occasionally causes invasive infection, primarily in patients with chronic granulomatous disease.

EPIDEMIOLOGY AND ECOLOGY

Aspergillus has a worldwide distribution, most commonly growing in decomposing plant materials (i.e., compost) and in bedding. This hyaline (nonpigmented), septate, branching mold produces vast numbers of conidia (spores) on stalks above the surface of mycelial growth. Aspergilli are found in indoor and outdoor air, on surfaces, and in water from surface reservoirs. Daily exposures vary from a few to many millions of conidia; high numbers of conidia are encountered in hay barns and other very dusty environments. The required size of the infecting inoculum is uncertain; however, only intense exposures (e.g., during construction work, handling of moldy bark or hay, or composting) are sufficient to cause disease—acute community-acquired pulmonary aspergillosis—in healthy immunocompetent individuals. Allergic syndromes may be exacerbated by continuous antigenic exposure arising from sinus or airway colonization or from nail infection. High-efficiency particulate air (HEPA) filtration is often protective against infection; thus HEPA filters should be installed and monitored for efficiency in operating rooms and in areas of the hospital that house high-risk patients.

The incubation period of invasive aspergillosis after exposure is highly variable, extending in documented cases from 2 to 90 days. Thus community acquisition of an infecting strain frequently manifests as invasive infection during hospitalization, although nosocomial acquisition is also common. Outbreaks usually are directly related to a contaminated air source in the hospital.

 Global aspergillosis incidence and prevalence have been estimated **(Table 212-1)**. The frequency of different manifestations of aspergillosis varies considerably with geographic location; most notably, chronic granulomatous sinusitis is rare outside the Middle East and India. Fungal (mycotic) keratitis is particularly common in Nepal, Myanmar, Bhutan, and India but occurs globally. Chronic pulmonary aspergillosis follows pulmonary tuberculosis in ~6–10% of treated cases and also mimics pulmonary tuberculosis as smear-negative or “clinically diagnosed” tuberculosis.

TABLE 212-1 Disease Frequency and Diagnostic Sensitivity for Different Manifestations of Aspergillosis

	Type of Disease		
Parameter	**Invasive**	**Chronic**	**Allergic**
Incidence/100,000[a]	3.5	10.4	?[b]
Prevalence/100,000[a]	—	32.8	286[c]
Global burden[a]	~250,000	~3,000,000	~10,000,000
Mortality rate without treatment	~100%	~50%	<1%
Respiratory Diagnostic Sensitivity[d]			
Culture	✓	✓	✓
Microscopy[e]	✓	✓	?
Antigen	✓✓✓	✓✓	✓✓✓
Real-time PCR	✓✓✓	✓✓✓	✓✓✓
Blood Diagnostic Sensitivity[d]			
Culture	×	×	×
Antigen	✓✓	✓	×
β-D-glucan	✓✓	✓	?
Real-time PCR	✓✓	×	×
IgG antibody	✓	✓✓✓✓	✓✓
IgE antibody	×	✓✓	✓✓✓✓

[a] *http://www.gaffi.org/roadmap/*.

[b] Allergic fungal disease can develop at any age, usually in adulthood; the annual frequency with which it occurs is not known.

[c] Allergic bronchopulmonary aspergillosis and severe asthma with fungal sensitization.

[d] Key for sensitivity: 1 check = limited (as the text indicates, 10–30% for culture); 2 checks = higher; 3 checks = >80%; and 4 checks = ~95%.

[e] High-volume fungal culture increases sensitivity to the same level as PCR.

Abbreviation: PCR, polymerase chain reaction.

RISK FACTORS AND PATHOGENESIS

The primary risk factors for invasive aspergillosis are profound neutropenia and glucocorticoid use; risk increases with longer duration of these conditions. Higher doses of glucocorticoids increase the risk of both acquisition of invasive aspergillosis and death from the infection. Neutrophil and/or phagocyte dysfunction also is an important risk factor, as evidenced by aspergillosis in chronic granulomatous disease, advanced HIV infection, and relapsed leukemia. An increasing incidence of invasive aspergillosis in medical intensive-care units suggests that, in patients who are not immunocompromised, temporary abrogation of protective responses as a result of glucocorticoid use or a general anti-inflammatory state is a significant risk factor. Many patients have some evidence of prior pulmonary disease—typically, a history of pneumonia or chronic obstructive pulmonary disease.

Many new immunomodulating agents, such as infliximab and ibrutinib, increase the risk of invasive aspergillosis, as do severe liver disease and high levels of stored iron in bone marrow. Influenza infection and extracorporeal membrane oxygenation therapy have recently been recognized as risk factors.

Patients with chronic pulmonary aspergillosis have a wide spectrum of underlying pulmonary disease, including tuberculosis, prior pneumothorax, or chronic obstructive pulmonary disease. These patients are immunocompetent except for some cytokine-regulation defects, most of which are consistent with an inability to mount an inflammatory immune (T_H1-like) response or to control it adequately. Glucocorticoids accelerate disease progression.

Allergic bronchopulmonary aspergillosis (ABPA) complicates asthma and cystic fibrosis. Many genetic associations indicate a strong basis for the development of a T_H2-like and "allergic" response to *A. fumigatus*. Remarkably, high-dose glucocorticoid treatment for exacerbations of ABPA almost never leads to invasive aspergillosis. Most patients with *Aspergillus* bronchitis have bronchiectasis, including those with cystic fibrosis.

 Different genetic traits are associated with invasive, chronic, and allergic aspergillosis; the majority of people probably are not at risk for aspergillosis. Multiple gene variants appear to be necessary for susceptibility to each form of aspergillosis.

CLINICAL FEATURES AND APPROACH TO THE PATIENT

(Table 212-2)

TABLE 212-2 Major Manifestations of Aspergillosis

	Type of Disease			
Organ	**Invasive (Acute and Subacute)**	**Chronic**	**Saprophytic**	**Allergic**
Lung	Angioinvasive (in neutropenia), non-angioinvasive, granulomatous	Chronic cavitary, *Aspergillus* nodule, chronic fibrosing	Aspergilloma (single), airway colonization	Allergic bronchopulmonary, severe asthma with fungal sensitization, extrinsic allergic alveolitis
Sinus	Acute invasive	Chronic invasive, chronic granulomatous	Maxillary fungal ball	Allergic fungal sinusitis, eosinophilic fungal rhinosinusitis
Brain	Abscess, hemorrhagic infarction, meningitis	Granulomatous, meningitis	None	None
Skin	Acute disseminated, locally invasive (trauma, burns, IV access)	External otitis, onychomycosis	None	None
Heart	Endocarditis (native or prosthetic), pericarditis	None	None	None
Eye	Keratitis, endophthalmitis	None	None	None described

Invasive Pulmonary Aspergillosis

Both the frequency of invasive disease and the pace of its progression increase with greater degrees of immunocompromise. Invasive aspergillosis is arbitrarily classified as acute and subacute, with courses of ≤1 month and 1–3 months, respectively. More than 80% of cases of invasive aspergillosis involve the lungs. The most common clinical features are no symptoms at all, fever, cough (sometimes productive), nondescript chest discomfort, trivial hemoptysis, and shortness of breath. Although the fever often responds to glucocorticoids, the disease progresses. The keys to early diagnosis in at-risk patients are a high index of suspicion, screening for circulating antigen (in leukemia), and urgent CT of the thorax. Invasive aspergillosis is one of the most common diagnostic errors revealed at autopsy.

Invasive Sinusitis

The sinuses are involved in 5–10% of cases of invasive aspergillosis, especially affecting patients with leukemia and recipients of hematopoietic stem cell transplants. In addition to fever, the most common features are nasal or facial discomfort, blocked nose, and nasal discharge (sometimes bloody). Endoscopic examination of the nose reveals pale, dusky or necrotic-looking tissue in any location. CT or MRI of the sinuses is essential but does not distinguish invasive *Aspergillus* sinusitis from preexisting allergic or bacterial sinusitis early in the disease process.

Tracheobronchitis

Occasionally, only the airways are infected by *Aspergillus*. The resulting manifestations seen on bronchoscopy range from acute or chronic bronchitis to ulcerative or pseudomembranous tracheobronchitis. These entities are particularly common among lung transplant recipients and patients on artificial ventilation. Obstruction with mucous plugs may occur and is called obstructing bronchial aspergillosis in immunocompromised patients and mucous impaction in other patients, such as those with ABPA.

Aspergillus Bronchitis

Recurrent chest infections that only partially improve with antibiotic treatment and are associated with significant breathlessness or coughing up of thick sputum plugs are typical features of *Aspergillus* bronchitis. Patients are not significantly immunocompromised and usually have bronchiectasis or cystic fibrosis. Occasional patients present with respiratory failure because of airway obstruction with mucus. Concurrent bacterial bronchitis is common. The diagnosis rests on recurrent detection of *Aspergillus* in the airway by microscopy, culture, or polymerase chain reaction (PCR). *Aspergillus* IgG is usually detectable.

Disseminated Aspergillosis

In the most severely immunocompromised patients, *Aspergillus* disseminates from the lungs to multiple organs—most often to the brain but also to the skin, thyroid, bone, kidney, liver, gastrointestinal tract, eye (endophthalmitis), and heart valve. Aside from cutaneous lesions, the most common features are gradual clinical deterioration over 1–3 days, with low-grade fever and features of mild sepsis, and nonspecific abnormalities in laboratory tests. In most cases, at least one localization becomes apparent before death. Blood cultures are almost always negative.

Cerebral Aspergillosis

Hematogenous dissemination to the brain is a devastating complication of invasive aspergillosis. Single or multiple lesions may develop. In acute disease, hemorrhagic infarction is most typical, and cerebral abscess is common. Rarer manifestations include meningitis, mycotic aneurysm, and cerebral granuloma (mimicking a brain tumor). Local spread from cranial sinuses also occurs.

Postoperative infection develops rarely and is exacerbated by glucocorticoids, which are often given after neurosurgery. The presentation can be either acute or subacute, with mood changes, focal signs, seizures, and decline in mental status. MRI is the most useful immediate investigation; unenhanced CT of the brain is usually nonspecific, and contrast is often contraindicated because of poor renal function.

Endocarditis

Most cases of *Aspergillus* endocarditis are prosthetic-valve infections resulting from contamination during surgery. Native-valve disease is reported, especially as a feature of disseminated infection and in persons using illicit IV drugs. Culture-negative endocarditis with large vegetations is the most common presentation; embolectomy occasionally reveals the diagnosis.

Cutaneous Aspergillosis

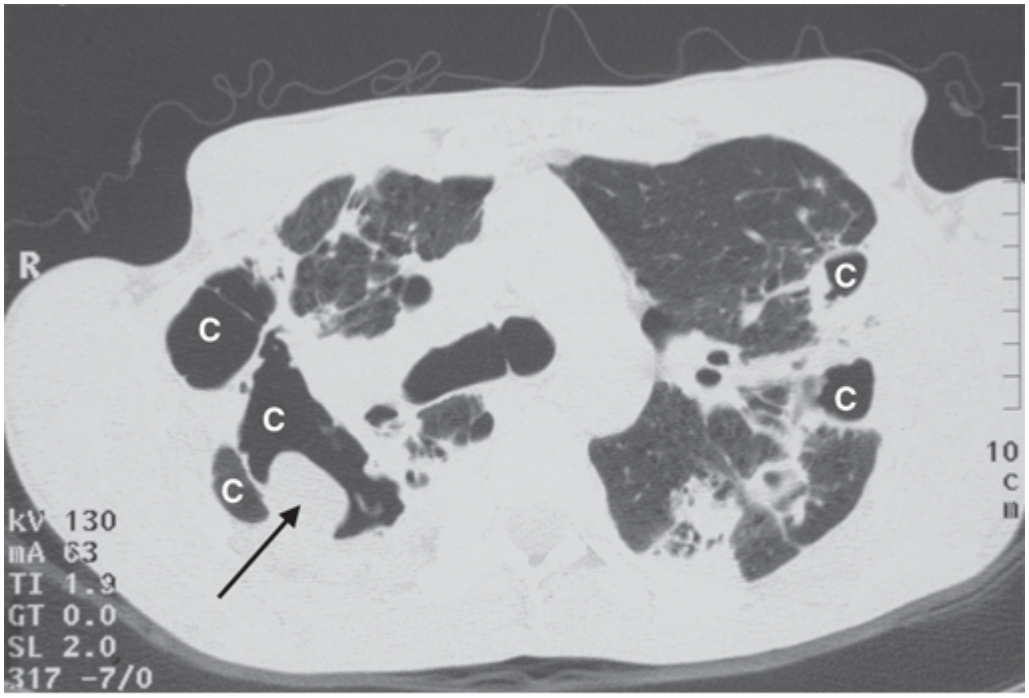
Dissemination of *Aspergillus* occasionally results in cutaneous features, usually an erythematous or purplish nontender area that progresses to a necrotic eschar. Direct invasion of the skin occurs in neutropenic patients at the site of IV catheter insertion and in burn patients. Surgical, burn, and trauma wounds may become infected with *Aspergillus* (especially *A. flavus*).

Chronic Pulmonary Aspergillosis

The hallmark of chronic cavitary pulmonary aspergillosis (CPA; also called semi-invasive aspergillosis, chronic necrotizing aspergillosis, or complex aspergilloma) **(Fig. 212-1)** is one or more pulmonary cavities expanding over a period of months or years in association with pulmonary symptoms and systemic manifestations such as fatigue and weight loss. Often mistaken initially for tuberculosis, more than 90% of CPA cases occur in patients with prior pulmonary disease (e.g., tuberculosis, atypical mycobacterial infection, sarcoidosis, rheumatoid lung disease, pneumothorax, bullae) or lung surgery. The onset is insidious, and systemic features may be more prominent than pulmonary symptoms. Cavities may have a fluid level or a well-formed fungal ball, but pericavitary infiltrates and multiple cavities—with or without pleural thickening—are typical. An irregular internal cavity surface and thickened cavity walls are indicative of disease activity. CPA is usually caused by *A. fumigatus*, but *A. niger* has been implicated, particularly in diabetic patients, and is linked to oxalosis with renal dysfunction. IgG antibodies to *Aspergillus* are detectable in ~95% of patients with CPA. Some patients have concurrent infections—even without a fungal ball—with atypical mycobacteria and/or other bacterial pathogens. The most significant complication is life-threatening hemoptysis, which may be the presenting manifestation.

FIGURE 212-1

CT scan image of the chest in a patient with long-standing bilateral chronic cavitary pulmonary aspergillosis. This patient had a history of several bilateral pneumothoraces and had required bilateral pleurodesis in 1990. CT then demonstrated multiple bullae, and sputum cultures grew *A. fumigatus*. The patient had initially weakly and later strongly positive serum IgG *Aspergillus* antibody tests. This scan (2003) shows a mixture of thick- and thin-walled cavities in both lungs (each marked with *C*), with a probable fungal ball (*black arrow*) protruding into the large cavity on the patient's right side (*R*). There is also considerable pleural thickening bilaterally.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A recently recognized form of chronic pulmonary aspergillosis is the *Aspergillus* nodule, which may resemble early lung carcinoma and may cavitate. Nodules may be single or multiple and 5–50 mm in diameter. Larger mass lesions are rarely seen. Nodules are usually avid on positron emission tomography. IgG antibodies to *Aspergillus* are detectable in ~65% of patients with an *Aspergillus* nodule. If untreated, chronic cavitary pulmonary aspergillosis typically progresses (sometimes relatively rapidly) to unilateral or upper-lobe fibrosis. This end-stage entity is termed *chronic fibrosing pulmonary aspergillosis*.

Aspergilloma

Aspergilloma (fungal ball) is a late manifestation of CPA, but some patients are asymptomatic. The inside of a pulmonary cavity allows growth that peels off, forming the layers of the fungal ball. Signs and symptoms associated with single (simple) aspergillomas are minor, including cough (sometimes productive), hemoptysis, wheezing, and mild fatigue. More significant signs and symptoms are associated with chronic cavitary pulmonary aspergillosis and should be treated as such. About 10% of fungal balls resolve spontaneously (by being coughed up), but the cavity may still be infected.

Chronic *Aspergillus* Sinusitis

Three entities are subsumed under this broad designation: fungal ball of the sinus, chronic invasive sinusitis, and chronic granulomatous sinusitis. Fungal ball of the sinus is limited to the maxillary sinus (except in rare cases involving the sphenoid sinus) and consists of a chronic saprophytic entity in which the sinus cavity is filled with a fungal ball. Maxillary disease is associated with prior upper-jaw root canal work and chronic (bacterial) sinusitis. About 90% of CT scans show focal hyperattenuation related to concretions; on MRI scans, the T2-weighted signal is decreased, whereas it is increased in bacterial sinusitis. Removal of the fungal ball is curative. No tissue invasion is demonstrable histologically or radiologically.

In contrast, chronic invasive sinusitis is a slowly destructive process that most commonly affects the ethmoid and sphenoid sinuses. Patients are usually but not always immunocompromised to some

degree (e.g., as a result of diabetes or HIV infection). Imaging of the cranial sinuses shows opacification of one or more sinuses, local bone destruction, and invasion of local structures. The differential diagnosis is wide, including other infections. Apart from a history of chronic nasal discharge and blockage, loss of the sense of smell, and persistent headache, the usual presenting features are related to local involvement of critical structures. The orbital apex syndrome (blindness and proptosis) is characteristic. Facial swelling, cavernous sinus thrombosis, carotid artery occlusion, pituitary fossa, and brain and skull-base invasion are complications.

Chronic granulomatous sinusitis due to *Aspergillus* is most commonly seen in the Middle East and India and is often caused by *A. flavus*. It typically presents late, with facial swelling and unilateral proptosis. The prominent granulomatous reaction histologically distinguishes this disease from chronic invasive sinusitis, in which tissue necrosis with a low-grade mixed-cell infiltrate is typical. IgG antibodies to *A. flavus* are usually detectable.

Allergic Bronchopulmonary Aspergillosis

In almost all cases, ABPA represents a hypersensitivity reaction to *A. fumigatus*; rare cases are due to other aspergilli and other fungi. ABPA occurs in ~2.5% of patients with asthma who are referred to secondary care and in up to 15% of teenagers with cystic fibrosis. Episodes of bronchial obstruction with mucous plugs leading to coughing fits, “pneumonia,” consolidation, and breathlessness are typical. Many patients report coughing up thick sputum casts. Eosinophilia commonly develops before systemic glucocorticoids are given. The cardinal diagnostic test is detection of *Aspergillus*-specific IgE (or a positive skin-prick test in response to *A. fumigatus* extract) together with an elevated serum level of total IgE (usually >1000 IU/mL). The presence of hyperattenuated mucus in airways is highly specific. Bronchiectasis is characteristic, and some patients develop chronic cavitary pulmonary aspergillosis.

Severe Asthma with Fungal Sensitization (SAFS)

Many adults with severe asthma do not fulfill the criteria for ABPA and yet are allergic to fungi. Although *A. fumigatus* is a common allergen, numerous other fungi (e.g., *Cladosporium* and *Alternaria* species) are implicated by skin-prick testing and/or specific IgE testing. Serum total IgE concentrations are <1000 IU/mL, and bronchial-wall thickening is common. ABPA and SAFS are referred to as *fungal asthma*.

Allergic Fungal Rhinosinusitis

Like the lungs, the sinuses manifest allergic responses to *Aspergillus* and other fungi. The affected patients present with chronic (i.e., perennial) sinusitis that is relatively unresponsive to antibiotics. Many of these patients have nasal polyps, and all have congested nasal mucosae and sinuses full of mucoid material. The histologic hallmarks of allergic fungal sinusitis are local eosinophilia and Charcot-Leyden crystals. Removal of abnormal mucus and polyps, with local and occasionally systemic administration of glucocorticoids, usually leads to resolution. Persistent or recurrent signs and symptoms may require more extensive surgery (ethmoidectomy) and possibly antifungal therapy. Recurrence is common, often after another bacterial or viral infection.

Superficial Aspergillosis

Aspergillus can cause keratitis and otitis externa. The former may be difficult to diagnose early enough to save the patient’s sight. Treatment requires local surgical debridement as well as intensive topical antifungal therapy with natamycin (5%). Otitis externa usually resolves with debridement and local application of antifungal agents.

DIAGNOSIS

Several techniques are required to establish the diagnosis of any form of aspergillosis with confidence (Table 212-1).

Acute Invasive Aspergillosis

Patients with acute invasive aspergillosis have a relatively heavy load of fungus in the affected organ; thus culture, molecular diagnosis, antigen detection, and histopathology usually confirm the diagnosis. However, the pace of progression leaves only a narrow window for making the diagnosis without losing the patient, and some invasive procedures are not possible because of coagulopathy, respiratory compromise, and other factors. Currently, ~40% of cases of invasive aspergillosis are missed clinically and are diagnosed only at autopsy. Histologic examination of affected tissue reveals either infarction, with invasion of blood vessels by many fungal hyphae, or acute necrosis, with limited inflammation and fewer hyphae. *Aspergillus* hyphae are hyaline, narrow, and septate, with branching at 45°; no yeast forms are present in infected tissue. Hyphae can be seen in cytology or microscopy preparations, which therefore provide a rapid means of presumptive diagnosis.

A positive culture supports the diagnosis, given that multiple other (rarer) fungi can mimic *Aspergillus* species histologically, but only 10–30% of patients with invasive aspergillosis have a positive culture. Bacterial agar is less sensitive than fungal media for culture; thus, if physicians do not request fungal culture, the diagnosis may be missed. A positive culture may represent noninvasive forms of aspergillosis or airway colonization. Both antigen detection and real-time PCR are faster and much more sensitive than culture of respiratory samples and blood.

The *Aspergillus* antigen test relies on detection of galactomannan release from *Aspergillus* organisms during growth. Positive serum antigen results usually precede clinical or radiologic features by several days. The sensitivity of antigen detection is reduced by antifungal prophylaxis and empirical therapy.

Definitive confirmation of a diagnosis of invasive aspergillosis requires (1) a positive culture of a sample taken directly from an ordinarily sterile site (e.g., a brain abscess) or (2) positive results of both histologic testing and culture of a sample taken from an affected organ (e.g., sinuses or skin). Most diagnoses of invasive aspergillosis are inferred from fewer data, including the presence of the *halo sign* on a high-resolution thoracic CT scan, in which a localized ground-glass appearance representing hemorrhagic infarction surrounds a nodule or consolidation. Halo signs are present for ~7 days early in the course of infection in neutropenic patients and are a good prognostic feature, reflecting an early diagnosis. Other characteristic radiologic features of invasive pulmonary aspergillosis include nodules and pleural-based infarction or cavitation, but nonspecific consolidation is common.

Chronic Aspergillosis

For chronic aspergillosis, *Aspergillus* antibody testing combined with characteristic imaging is sufficient for the diagnosis. Biopsy of *Aspergillus* nodules reveals hyphae surrounded by cells of chronic inflammation and sometimes granulomas. Antibody titers fall slowly with successful therapy. Cultures are infrequently positive but are important in checking for azole resistance. Real-time PCR of sputum is often strongly positive. Some patients with chronic pulmonary aspergillosis also have elevated titers of total serum IgE and *Aspergillus*-specific IgE.

ABPA, SAFS, and Allergic *Aspergillus* Sinusitis

ABPA and SAFS are diagnosed serologically with elevated specific and total serum IgE levels or with skin-prick tests. Allergic *Aspergillus* sinusitis is usually diagnosed histologically, although

measurement of IgE antibodies in blood also may be useful.

TREATMENT

TREATMENT Aspergillosis

Antifungal drugs active against *Aspergillus* include voriconazole, itraconazole, posaconazole, isavuconazole, caspofungin, micafungin, and amphotericin B (AmB). Possible interactions with other drugs must be considered before azoles are prescribed. In addition, plasma azole concentrations vary substantially from one patient to another, and many authorities recommend monitoring levels to ensure that drug concentrations are adequate but not excessive, especially with itraconazole and voriconazole. Initial IV administration is preferred for acute invasive aspergillosis and oral administration for all other diseases that require antifungal therapy. Current recommendations are shown in **Table 212-3**.

Voriconazole and isavuconazole are the preferred agents for invasive aspergillosis; caspofungin, posaconazole, micafungin, and lipid-associated AmB are second-line agents. AmB is not active against *A. terreus* or *A. nidulans*; multi-azole resistance in *A. fumigatus* is present in <5% of isolates but is increasing; and *A. niger* is resistant to itraconazole and isavuconazole. An infectious disease consultation is advised for patients with invasive disease, given the complexity of management. Immune reconstitution can complicate recovery. The duration of therapy for invasive aspergillosis varies from ~3 months to several years, depending on the patient's immune status and response to therapy. Relapse occurs if the response is suboptimal and immune reconstitution is not complete.

Itraconazole is currently the preferred oral agent for chronic and allergic forms of aspergillosis. Voriconazole or posaconazole can be substituted when failure, emergence of resistance, or adverse events occur. An itraconazole dose of 200 mg twice daily is recommended, with monitoring of drug concentrations in the blood. Acute exacerbations of ABPA respond well to a short course of glucocorticoids. Because chronic cavitary pulmonary aspergillosis responds slowly, therapy for >6 months is necessary, and disease control may require years of treatment, whereas the duration of treatment for other forms of chronic and allergic aspergillosis requires case-by-case evaluation. Glucocorticoids should be used in chronic cavitary pulmonary aspergillosis only if covered by adequate antifungal therapy. Antifungal response in *Aspergillus* bronchitis is gratifying, but relapse after 4 months of therapy is common.

 Resistance in *A. fumigatus* to one or more azoles, although uncommon, is increasingly found globally. Resistance may be derived from azole fungicide use for crops. In addition, resistance arising from multiple mechanisms may develop during long-term treatment, and a positive culture during antifungal therapy is an indication for susceptibility testing.

Surgical treatment is important in several forms of aspergillosis, including fungal ball of the sinus and single aspergillomas, in which surgery is curative; invasive aspergillosis involving bone, heart valve, sinuses, and proximal areas of the lung; brain abscess; keratitis; and endophthalmitis. In allergic fungal sinusitis, removal of abnormal mucus and polyps, with local and occasionally systemic glucocorticoid treatment, usually leads to resolution. Persistent or recurrent signs and symptoms may require more extensive surgery (ethmoidectomy) and possibly antifungal therapy. Surgery is problematic in chronic cavitary pulmonary aspergillosis, usually resulting in serious complications. Bronchial artery embolization is preferred for problematic hemoptysis.

TABLE 212-3 Treatment of Aspergillosis[a]

Indication	Primary Treatment	Precautions	Secondary Treatment	Comments
Invasive disease[b]	Voriconazole, isavuconazole	Drug interactions (especially with rifampin and carbamazepine)[c]	AmB, caspofungin, posaconazole, micafungin	As primary therapy, voriconazole and isavuconazole have a 20% higher response rate than AmB. Therapeutic drug monitoring is recommended for voriconazole.
Prophylaxis	Posaconazole tablet, itraconazole solution	Diarrhea and vomiting with itraconazole, vincristine interaction	Micafungin, aerosolized AmB	Some centers monitor plasma levels of itraconazole and posaconazole.
Single aspergilloma	Surgery	Multicavity disease: poor outcome of surgery, medical therapy preferable	Itraconazole, voriconazole, intracavity AmB	Single large cavities with an aspergilloma are best resected.
Chronic pulmonary disease[b]	Itraconazole, voriconazole	Poor absorption of itraconazole capsules with proton pump inhibitors or H_2 blockers	Posaconazole, IV AmB, IV micafungin	Resistance may emerge during treatment, especially if plasma drug levels are subtherapeutic.
ABPA/SAFS ("fungal asthma")	Itraconazole	Some glucocorticoid interactions, including with inhaled formulations	Voriconazole, posaconazole	Long-term therapy is helpful in most cases. No evidence indicates whether therapy modifies progression to bronchiectasis/fibrosis.

[a]For information on duration of therapy and drug resistance in certain *Aspergillus* species, see text. [b]An infectious disease consultation is appropriate for these patients. [c]Online drug-interaction resource: *www.aspergillus.org.uk/content/antifungal-drug-interactions*.

Note: After loading doses, the oral dose is usually 200 mg bid for voriconazole and itraconazole, 300 mg qd for posaconazole tablets, and 200 mg qd for isavuconazole. The IV dose of voriconazole for adults is 6 mg/kg twice at 12-h intervals (loading doses) followed by 4 mg/kg q12h; a larger dose is required for children and teenagers; a lower dose may be safer for persons >70 years of age. Plasma monitoring is helpful in optimizing the dosage. The IV dose of isavuconazole is 200 mg tid for 2 days (loading dose) followed by 200 mg qd. Caspofungin is given as a single loading dose of 70 mg and then at 50 mg/d; some authorities use 70 mg/d for patients weighing >80 kg, and lower doses are required with hepatic dysfunction. Micafungin is given as 50 mg/d for prophylaxis and as at least 150 mg/d for treatment; this drug has not yet been approved by the U.S. Food and Drug Administration (FDA) for this indication. AmB deoxycholate is given at a daily dose of 1 mg/kg if tolerated. Several strategies are available for minimizing renal dysfunction. Lipid-associated AmB is given at 3 mg/kg (AmBisome) or 5 mg/kg (Abelcet). Different regimens are available for aerosolized AmB, but none is FDA approved. Other considerations that may alter dose selection or route include age; concomitant medications; renal, hepatic, or intestinal dysfunction; and drug tolerability.

Abbreviations: AmB, amphotericin B; ABPA, allergic bronchopulmonary aspergillosis; SAFS, severe asthma with fungal sensitization.

PROPHYLAXIS

In situations in which moderate or high risk is predicted (e.g., after induction therapy for acute myeloid leukemia), the need for antifungal prophylaxis for superficial and systemic candidiasis and for invasive aspergillosis is generally accepted. Fluconazole is commonly used in these situations but has no activity against *Aspergillus* species. Itraconazole capsules are ineffective, and itraconazole solution offers only modest efficacy. Posaconazole tablets are more effective in reducing infection rates and the need for empirical antifungal therapy. Some data support the use of IV micafungin. No prophylactic regimen is completely successful.

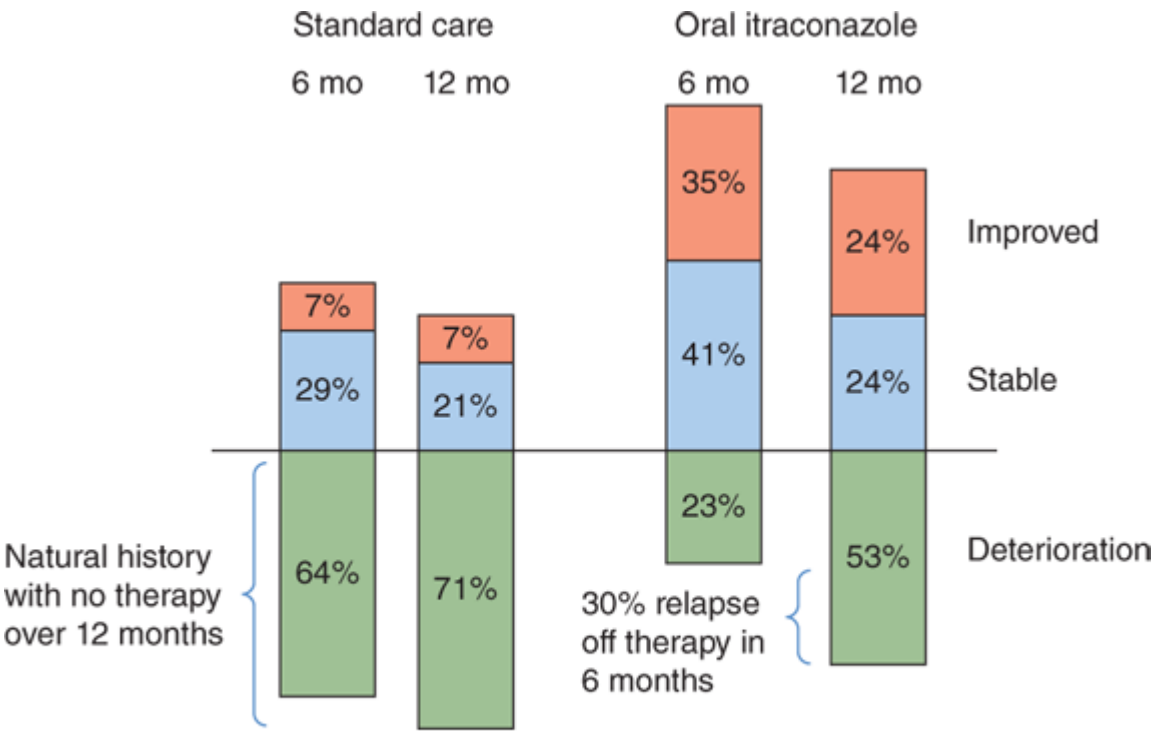
OUTCOME

Invasive aspergillosis is curable if immune reconstitution occurs, whereas allergic and chronic forms are not. The mortality rate for invasive aspergillosis is 30–70% if the infection is treated but is 100% if the diagnosis is missed. Infection with a voriconazole-resistant strain carries a mortality rate of 90%. Cerebral aspergillosis, *Aspergillus* endocarditis, and bilateral extensive invasive pulmonary aspergillosis have very poor outcomes, as does invasive infection in persons with late-stage AIDS or relapsed uncontrolled leukemia.

The mortality rate for chronic cavitary pulmonary aspergillosis is ~40% over 5 years if the patient is actively treated with antifungal agents. After 12 months with no antifungal therapy, 70% of patients have deteriorated and 30% are stable **(Fig. 212-2)**. Therapy fails in ~30% of recipients of antifungal therapy and still more often if azole resistance is present.

FIGURE 212-2

Comparison of the impact of itraconazole therapy (400 mg/d) and that of standard care on chronic cavitary pulmonary aspergillosis at 6 and 12 months. *(After R Agarwal et al: Itraconazole in chronic cavitary pulmonary aspergillosis: A randomised controlled trial and systematic review of literature. Mycoses 56:559, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Both ABPA and SAFS patients respond to antifungal therapy; ~60% respond to itraconazole and ~80% to voriconazole and posaconazole (if tolerated). Inhaled amphotericin B is effective in and tolerated by ~15% of patients. If the severity of asthma declines, the inhaled glucocorticoid dose can be reduced and oral glucocorticoids can be stopped. Relapse after discontinuation is common but not universal.

FURTHER READING

Denning DW et al: Fungal allergy in asthma—State of the art and research needs. Clin Transl Allergy 4:14, 2014. [PubMed: 24735832]
Denning DW et al: Chronic pulmonary aspergillosis—Rationale and clinical guidelines for diagnosis and management. Eur Resp J 47:45, 2016.
Gregg KS, Kauffman CA: Invasive aspergillosis: Epidemiology, clinical aspects, and treatment. Semin Respir Crit Care Med 36:662, 2015. [PubMed: 26398533]
Muldoon EG et al: *Aspergillus* nodules; another presentation of chronic pulmonary aspergillosis. BMC Pulm Med 16:123, 2016. [PubMed: 27538521]

Chapter 213: Mucormycosis

Brad Spellberg; Ashraf S. Ibrahim

FIGURE 213-1

INTRODUCTION

Mucormycosis represents a group of life-threatening infections caused by fungi of the order Mucorales of the subphylum Mucoromycotina (formerly known as the class Zygomycetes). Mucormycosis is highly invasive and relentlessly progressive, resulting in higher rates of morbidity and mortality than many other infections. Although mortality rates from mucormycosis have declined in recent years as a result of early initiation of more effective antifungal therapies, they remain high overall.

ETIOLOGY

Fungi of the order Mucorales belong to seven families **(Table 213-1)**, all of which can cause mucormycosis. Among the Mucorales, *Rhizopus oryzae* and *R. delemar* (a related, more recently recognized species)—both in the family Mucoraceae—are by far the most common causes of mucormycosis in the Western Hemisphere. Less frequently isolated species of the Mucoraceae that cause a similar spectrum of infections include *Rhizopus microsporus*, *Rhizomucor pusillus, Lichtheimia corymbifera* (formerly *Absidia corymbifera*), *Apophysomyces elegans*, and *Mucor* species. Increasing numbers of cases of mucormycosis due to infection with *Cunninghamella* species (family Cunninghamellaceae) have also been reported, particularly in highly immunocompromised patients. Only rare case reports have demonstrated the ability of fungi in the remaining families of the Mucorales to cause mucormycosis, although other Mucorales can be the major cause of disease in certain geographic areas (e.g., *A. elegans* in India and *Mucor irregularis* in China).

TABLE 213-1 Taxonomy of Fungi Causing Mucormycosis (Subphylum Mucoromycotina, Order Mucorales)

Family	Genus (Species Listed for Some)
Mucoraceae	*Rhizopus oryzae* *Rhizopus delemar* *Rhizopus microsporus* *Rhizomucor* *Mucor* *Actinomucor*
Lichtheimiaceae	*Lichtheimia* (formerly *Mycocladus*, formerly *Absidia*)
Cunninghamellaceae	*Cunninghamella*
Thamnidiaceae	*Cokeromyces*
Mortierellaceae	*Mortierella*
Saksenaceae	*Saksenaea* *Apophysomyces*
Syncephalastraceae	*Syncephalastrum*

PATHOGENESIS

The Mucorales are ubiquitous environmental fungi to which humans are constantly exposed. These fungi cause infection primarily in patients with diabetes, defects in phagocytic function (e.g., neutropenia or glucocorticoid treatment), and/or elevated levels of free iron, which supports fungal growth in serum and tissues. In the past, iron-overloaded patients with end-stage renal failure who were treated with deferoxamine had a high risk of developing rapidly fatal disseminated mucormycosis; deferoxamine is an iron chelator for the human host, but it serves as a fungal siderophore, directly delivering iron to the Mucorales. Furthermore, patients with diabetic ketoacidosis (DKA) are at high risk of developing rhinocerebral mucormycosis. The acidosis causes dissociation of iron from sequestering proteins, resulting in enhanced fungal survival and virulence. The ketoacid β-hydroxybutyrate increases expression of host and fungal receptors that result in fungal adherence and penetration into tissues.

Nevertheless, the majority of diabetic patients who present with mucormycosis are not acidotic, and, even absent acidosis, hyperglycemia directly contributes to the risk of mucormycosis by at least four likely mechanisms: (1) hyperglycation of iron-sequestering proteins, disrupting normal iron sequestration; (2) upregulation of a mammalian cell receptor (GRP78) that binds to Mucorales, enabling tissue penetration (due to both a direct effect of hyperglycemia and increasing levels of free iron, which independently enhances GRP78 expression); (3) induction of poorly characterized defects in phagocytic function; and (4) enhanced expression of CotH, a Mucorales-specific protein that mediates host cell invasion by binding to GRP78 (due to hyperglycemia and the resulting free iron).

EPIDEMIOLOGY

Mucormycosis typically occurs in patients with diabetes mellitus, solid organ or hematopoietic stem cell transplantation (HSCT), prolonged neutropenia, or malignancy. The majority of diabetic patients are not acidotic on presentation with mucormycosis. Furthermore, patients often have no previously

recognized history of diabetes mellitus when they present with mucormycosis. In these instances, presentation for mucormycosis may result in the first clinical recognition of hyperglycemia, which often has been unmasked by recent glucocorticoid use. Thus a high index of suspicion of mucormycosis must be maintained, even in the absence of a known history of diabetes, if hyperglycemia is present. In patients undergoing HSCT, mucormycosis develops at least as commonly during nonneutropenic as during neutropenic periods, probably because of glucocorticoid treatment of graft-versus-host disease. Mucormycosis can occur as isolated cutaneous or subcutaneous infection in immunologically normal individuals after traumatic implantation of soil or vegetation (e.g., due to natural disasters, motor vehicle accidents, or—in soldiers—severe injuries during combat operations) or in nosocomial settings via direct access through IV catheters, SC injections, or maceration of the skin by a moist dressing.

Patients receiving antifungal prophylaxis with either itraconazole or voriconazole may be at increased risk of mucormycosis. These patients typically present with disseminated mucormycosis, the most lethal form of disease. Breakthrough mucormycosis also has been described in patients receiving posaconazole or echinocandin prophylaxis.

CLINICAL MANIFESTATIONS

Mucormycosis can be divided into at least six clinical syndromes: rhino-orbital-cerebral, pulmonary, cutaneous, gastrointestinal, disseminated, and miscellaneous. Patients with specific defects in host defense tend to develop specific syndromes. For example, patients with diabetes mellitus and/or DKA typically develop the rhino-orbital-cerebral form and much more rarely develop pulmonary or disseminated disease. In contrast, pulmonary mucormycosis occurs most commonly in leukemic patients who are receiving chemotherapy and in patients undergoing HSCT.

Rhino-Orbital-Cerebral Disease

Rhino-orbital-cerebral mucormycosis continues to be the most common form of the disease. Most cases occur in patients with diabetes, although such cases are increasingly being described in the transplantation setting, often along with glucocorticoid-induced diabetes mellitus. The initial symptoms of rhino-orbital-cerebral mucormycosis are nonspecific and include eye or facial pain and facial numbness followed by the onset of conjunctival suffusion and blurry vision. Fever may be absent in up to half of cases. White blood cell counts are typically elevated as long as the patient has functioning bone marrow. If untreated, infection usually spreads from the ethmoid sinus to the orbit, resulting in compromise of extraocular muscle function and proptosis, typically with chemosis. From the orbit, spread often takes place via hematogenous or contiguous dissemination to the frontal lobe of the brain and/or via venous drainage to the cavernous sinus. Onset of signs and symptoms in the contralateral eye, with resulting bilateral proptosis, chemosis, vision loss, and ophthalmoplegia, is ominous, suggesting the development of cavernous sinus thrombosis.

Upon visual inspection, infected tissue may appear to be normal during the earliest stages of fungal spread, then progressing through an erythematous phase, with or without edema, before the onset of a violaceous appearance and finally the development of a black necrotic eschar. Infection can sometimes extend from the sinuses into the mouth and produce painful necrotic ulcerations of the hard palate, but this is a late finding that suggests extensive, well-established infection.

One common misperception about mucormycosis is that it is always rapidly progressive. In fact, the rate of progression is extremely variable and is possibly dependent on the immune status of the patient and the causative Mucorales species, some of which are more virulent and/or have faster growth rates than others. Patients may go from initial symptoms to death in days; alternatively, it can take months or even a year or more for lethal progression to occur.

Pulmonary Disease

Pulmonary mucormycosis is the second most common manifestation. Symptoms include dyspnea, cough, and chest pain; fever is often but not invariably present. Angioinvasion results in necrosis, cavitation, and/or hemoptysis. Lobar consolidation, isolated masses, nodular disease, cavities, or wedge-shaped infarcts may be seen on chest radiography. High-resolution chest CT is the best method for determining the extent of pulmonary mucormycosis and may demonstrate evidence of infection before it is seen on chest x-ray. In the setting of cancer, where mucormycosis may be difficult to differentiate from aspergillosis, the presence of ≥10 pulmonary nodules, pleural effusion, or concomitant sinusitis makes mucormycosis more likely. It is critical to distinguish mucormycosis from aspergillosis as rapidly as possible because treatments for these infections differ. Indeed, voriconazole—the first-line treatment for aspergillosis—exacerbates mucormycosis in mouse and fly models of infection.

Cutaneous Disease

Cutaneous mucormycosis may result from external implantation of the fungus or from hematogenous dissemination. External implantation–related infection has been described in the setting of soil exposure from trauma (e.g., in a motor vehicle accident, a natural disaster, or combat-related injuries), penetrating injury with plant material (e.g., a thorn), injections of medications (e.g., insulin), catheter insertion, contamination of surgical dressings, and use of tape to secure endotracheal tubes. Cutaneous disease can be highly invasive, penetrating into muscle, fascia, and even bone. Necrotizing fasciitis caused by mucormycosis carries a mortality rate approaching 80%. Necrotic cutaneous lesions in the setting of hematogenous dissemination also are associated with an extremely high mortality rate. However, with prompt, aggressive surgical debridement, isolated cutaneous mucormycosis has a favorable prognosis and a low mortality rate.

Gastrointestinal Disease

In the past, gastrointestinal mucormycosis occurred primarily in premature neonates in association with disseminated disease and necrotizing enterocolitis. However, there has been a marked increase in case reports describing adults with neutropenia, glucocorticoid use, or other immunocompromising conditions. In addition, gastrointestinal disease has been reported as a nosocomial process following administration of medications mixed with contaminated wooden applicator sticks. Nonspecific abdominal pain and distention associated with nausea and vomiting are the most common symptoms. Gastrointestinal bleeding is common, and fungating masses may be seen in the stomach at endoscopy. The disease may progress to visceral perforation, with extremely high mortality rates.

Disseminated and Miscellaneous Forms of Disease

Hematogenously disseminated mucormycosis may originate from any primary site of infection. The most common site of dissemination is the brain, but metastatic lesions may also be found in any other organ. Mortality rates for widely disseminated mucormycosis exceed 90%; however, these high rates are likely to be due in part to the underlying predisposing condition leading to the infection. Miscellaneous forms of mucormycosis may affect any body site, including bones, mediastinum, trachea, kidneys, and peritoneum (in association with dialysis); even isolated infection of teeth has been reported.

DIAGNOSIS

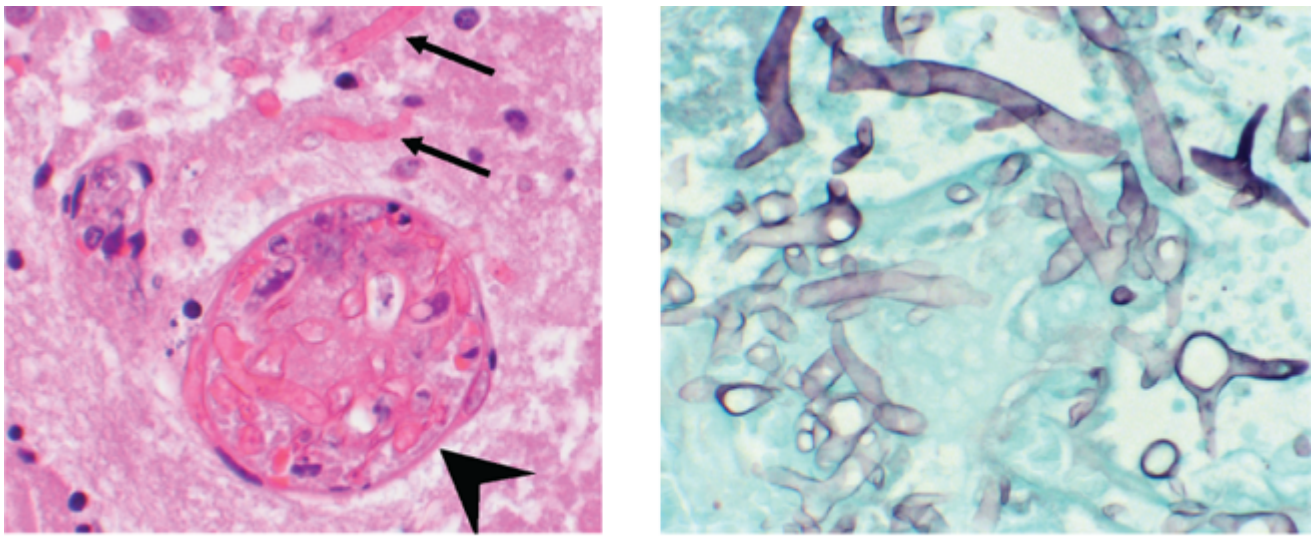
A high index of suspicion is required for diagnosis of mucormycosis. Unfortunately, autopsy series have shown that up to half of cases are diagnosed only post-mortem. Because the Mucorales are environmental isolates, definitive diagnosis requires a positive culture from a sterile site (e.g., a needle aspirate, a tissue biopsy specimen, or pleural fluid) or histopathologic evidence of invasive

mucormycosis. A probable diagnosis of mucormycosis can be established by culture from a nonsterile site (e.g., sputum or bronchoalveolar lavage) or the detection of Mucorales on the surface of histopathology samples (without visualization of evidence of invasion) when a patient has appropriate risk factors as well as clinical and radiographic evidence of disease. In such cases, given the urgency of administering therapy early, the patient should be treated while confirmation of the diagnosis is awaited.

Biopsy with histopathologic examination remains the most sensitive and specific modality for definitive diagnosis **(Fig. 213-1)**. Biopsy reveals characteristic wide (≥6- to 30-μm), thick-walled, ribbon-like, aseptate hyphal elements that branch at right angles. Other fungi, including *Aspergillus, Fusarium*, and *Scedosporium* species, have septa, are thinner, and branch at acute angles. Because artificial septa may result from folding of tissue during processing (which may also alter the appearance of the angle of branching), the width and the ribbon-like form of the fungus are the most reliable features distinguishing mucormycosis. The Mucorales are visualized most effectively with periodic acid–Schiff or hematoxylin and eosin; in contrast to many other fungi, methenamine silver may not result in optimal staining. While histopathology can identify the Mucorales, species can be identified only by culture. Polymerase chain reaction (PCR) is being investigated as a diagnostic tool for mucormycosis but is not yet approved by the U.S. Food and Drug Administration (FDA) for this purpose and is not generally available.

FIGURE 213-1

Histopathology sections of *Rhizopus delemar* in infected brain. *A.* Broad, ribbon-like, nonseptate hyphae in the parenchyma (*arrows*) and a thrombosed blood vessel with extensive intravascular hyphae (*arrowhead*) (hematoxylin and eosin). ***B.*** Extensive, broad, ribbon-like hyphae invading the parenchyma (Gomori methenamine silver).



A ***B***

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Unfortunately, cultures are positive in fewer than half of cases of mucormycosis. Nevertheless, the Mucorales are not fastidious organisms and tend to grow quickly (i.e., within 48–96 h) on culture media. The likely explanation for the low sensitivity of culture is that the Mucorales form long filamentous structures that are killed by tissue homogenization—the standard method for preparing tissue cultures in the clinical microbiology laboratory. Thus the laboratory should be advised when a diagnosis of mucormycosis is suspected, and the tissue should be cut into sections and placed in the

center of culture dishes rather than homogenized. Because there is also substantial variability among isolates in optimal growth temperature, growth at both room temperature and 37°C is advisable.

Imaging techniques often yield subtle findings that underestimate the extent of disease. For example, the most common finding on CT or MRI of the head or sinuses of a patient with rhino-orbital mucormycosis is sinusitis that is indistinguishable from bacterial sinusitis. It is also common to detect no abnormalities in sinus bones despite clinical evidence of progressive disease. MRI is more sensitive (~80%) for detecting orbital and CNS disease than is CT. High-risk patients should always undergo endoscopy and/or surgical exploration, with biopsy of the areas of suspected infection. If mucormycosis is suspected, initial empirical therapy with a polyene antifungal agent should be initiated while the diagnosis is being confirmed.

DIFFERENTIAL DIAGNOSIS

Other mold infections, including aspergillosis, scedosporiosis, fusariosis, and infections caused by the dematiaceous fungi (brown-pigmented soil organisms), can cause clinical syndromes identical to mucormycosis. Histopathologic examination usually allows distinction of the Mucorales from these other organisms, and a positive culture permits definitive species identification. As stated above, it is important to distinguish the Mucorales from these other fungi, as the preferred antifungal treatments differ (i.e., polyenes for the Mucorales vs expanded-spectrum triazoles for most septate molds). The entomophthoromycoses caused by *Basidiobolus* and *Conidiobolus* also can cause identical clinical syndromes. These fungi cannot be readily distinguished from the Mucorales by histopathology but can be reliably distinguished by culture. Fortunately, entomophthoromycoses are uncommon in developed countries and can be treated with polyenes; in this setting, it is not urgent to distinguish them from mucormycosis.

In a patient with sinusitis and proptosis, orbital cellulitis and cavernous sinus thrombosis caused by bacterial pathogens (most commonly *Staphylococcus aureus*, but also streptococcal and gram-negative species) must be excluded. *Klebsiella rhinoscleromatis* is a rare cause of an indolent facial rhinoscleroma syndrome that may appear similar to mucormycosis. Finally, the Tolosa-Hunt syndrome causes painful ophthalmoplegia, ptosis, headache, and cavernous sinus inflammation; biopsies and clinical follow-up may be needed to distinguish the Tolosa-Hunt syndrome from mucormycosis by the lack of progression of the former entity.

TREATMENT

TREATMENT Mucormycosis GENERAL PRINCIPLES

Optimizing the chances for successful treatment of mucormycosis requires three steps: (1) early initiation of therapy; (2) rapid reversal of underlying predisposing risk factors, if possible; and (3) surgical debridement, when possible. Maintaining a high index of suspicion for patients at risk for mucormycosis is critical. Multiple studies have found that earlier initiation of polyene-based therapy improves survival of patients with mucormycosis. Because the disease can present subtly at first and because confirmation of the diagnosis can take days, therapy often must be started empirically before the diagnosis is established. When there is a reasonable suspicion of mucormycosis, clinicians should not hesitate to initiate therapy with a lipid polyene as soon as possible, since the toxicity of lipid polyenes (unlike that of amphotericin B [AmB] deoxycholate) is rarely substantial after one or two doses.

It is also crucial to rapidly reverse (or prevent) underlying defects in host defense during treatment (e.g., by stopping or reducing the dosage of immunosuppressive medications or by rapidly restoring euglycemia and normal acid–base status). Indeed, a recent study confirmed that resolution of acidosis in mice with DKA via the administration of sodium bicarbonate (in lieu of insulin) improved survival.

Administration of glucocorticoids predisposes animals to death from mucormycosis in experimental models. Similarly, iron administration to patients with active mucormycosis should be avoided, as iron exacerbates infection in experimental models. Blood transfusion typically results in some liberation of free iron due to hemolysis, so a conservative approach to red blood cell transfusions is advisable.

Blood vessel thrombosis and resulting tissue necrosis during mucormycosis can result in poor penetration of antifungal agents to the site of infection. Therefore, debridement of all necrotic tissues is critical for eradication of disease. Surgery has been found (by logistic regression and in multiple case series) to be an independent variable for favorable outcome in patients with mucormycosis. Limited data from a retrospective study support the use of intraoperative frozen sections to delineate the margins of infected tissues, with sparing of tissues lacking evidence of infection. A multidisciplinary team, including an internist, an infectious disease specialist, and surgical specialists whose expertise is relevant to the sites of infection, is typically required for the management of mucormycosis.

ANTIFUNGAL THERAPY

Primary therapy for mucormycosis should be based on a polyene antifungal agent **(Table 213-2)**, except perhaps in mild localized infection (e.g., isolated suprafascial cutaneous infection) that has been eradicated surgically in an immunocompetent patient. Lipid formulations of AmB are significantly less nephrotoxic than AmB deoxycholate, can be administered at higher doses, and are probably more effective for this purpose. Liposomal amphotericin B (LAmB) is preferred to amphotericin B lipid complex (ABLC) for management of central nervous system (CNS) infection on the basis of retrospective survival data and superior brain penetration; there is no clear efficacy advantage of either agent for non-CNS infections, although LAmB may be less nephrotoxic than ABLC.

Starting dosages of 1 mg/kg per day for AmB deoxycholate and 5 mg/kg per day for LAmB and ABLC are commonly given to adults and children to treat mucormycosis. Dose escalation of LAmB to 7.5 or 10 mg/kg per day for CNS mucormycosis may be considered in light of the limited penetration of polyenes into the brain. Because of auto-induction of metabolism, which results in paradoxically lower drug levels, there is no advantage to escalating the LAmB dose above 10 mg/kg per day, and doses of 5 mg/kg per day are probably adequate for non-CNS infections. ABLC dose escalation above 5 mg/kg per day is not advisable given the lack of relevant data and the drug's potential toxicity.

In multiple studies, various combinations of lipid polyenes (both ABLC and LAmB) plus echinocandins (e.g., caspofungin, micafungin, and anidulafungin) improved survival rates among mice with disseminated mucormycosis (including CNS disease). Furthermore, combination lipid polyene–echinocandin therapy was associated with significantly better outcomes than polyene monotherapy in a retrospective clinical study involving patients with rhino-orbital-cerebral mucormycosis (including CNS disease). The effect of echinocandins appears to be to down-modulate the virulence of the fungus and reduce tissue necrosis and destruction from fungal invasion. On the basis of such data, some experts prefer combination lipid polyene–echinocandin therapy as a first-line option. However, definitive clinical trials are needed to establish whether the combination is superior in efficacy to monotherapy for mucormycosis. When used, echinocandins should be administered at standard, FDA-approved doses, since dose escalation has resulted in paradoxical loss of efficacy in preclinical models.

In contrast to deferoxamine, the iron chelator deferasirox is fungicidal against clinical isolates of the Mucorales. In mice with DKA and disseminated mucormycosis, combination deferasirox–LAmB therapy resulted in synergistic improvement of survival rates and reduced the fungal burden in the brain. Unfortunately, a randomized, double-blind, phase 2 safety clinical trial of adjunctive therapy with deferasirox (plus LAmB) documented excess mortality among patients treated with deferasirox. Of note, the study population included primarily patients with active malignancy, and few patients in the study had diabetes mellitus as their only risk factor. Deferasirox is therefore contraindicated as therapy in patients with active malignancy, but its role in patients who have diabetes mellitus without malignancy (the setting in which its preclinical efficacy was optimal) remains uncertain.

Posaconazole and isavuconazole are the only FDA-approved azoles with in vitro activity against the Mucorales. However, posaconazole has been found to be inferior in efficacy to AmB for the treatment of murine mucormycosis and was not superior to placebo. Moreover, posaconazole–polyene combination therapy was not superior to polyene monotherapy for mucormycosis in mice, and no comparative data are available for combination therapy in humans. Thus no data support the use of combination posaconazole–polyene regimens. Although the minimal inhibitory concentrations (MICs) of isavuconazole against the Mucorales are four- to eightfold higher than those of posaconazole, blood levels may be higher with standard isavuconazole dosing than with posaconazole. Therefore, neither azole is clearly preferable to the other as a therapeutic option. Isavuconazole is approved for the treatment of mucormycosis on the basis of a small, historically controlled study. Given this limited dataset, many experts continue to think that lipid polyenes are first-line options and that isavuconazole, like posaconazole, is best reserved for oral step-down therapy in patients whose condition has substantially improved on polyene-based therapy or for salvage therapy in patients who are intolerant of polyene-based regimens or whose infection is refractory to these regimens. As for posaconazole, no data support the use of combination isavuconazole–polyene regimens in lieu of polyene monotherapy or polyene–echinocandin combination regimens. Some experts use triple therapy with a polyene, echinocandin, and either posaconazole or isavuconazole for patients who have extensive disease or whose disease has progressed on prior therapy. Empirical, dual lipid polyene–azole therapy is a rational choice in a patient with likely invasive mold infections when septate molds and mucormycosis are both in the differential diagnosis and the etiologic agent has not yet been confirmed.

The roles of recombinant cytokines and neutrophil transfusions in the primary treatment of mucormycosis are not clear, although it is intuitive that earlier recovery of neutrophil counts should improve survival rates. Limited data from uncontrolled studies support the use of hyperbaric oxygen in centers with the appropriate technical expertise and facilities; its efficacy remains undefined. As mentioned previously, one study in mice with DKA found that administration of sodium bicarbonate improved survival from mucormycosis; however, because insulin was not administered to the mice, it is unclear whether the therapeutic effect is clinically relevant.

In general, antifungal therapy for mucormycosis should be continued until resolution of clinical signs and symptoms of infection and resolution of underlying immunosuppression. However, after several weeks of daily therapy in a patient who is clinically improving, it is reasonable to consider switching to thrice-weekly lipid polyene doses—with ultimate weaning down to twice-weekly doses—for maintenance therapy. For patients with mucormycosis who are receiving immunosuppressive medications, secondary antifungal prophylaxis is typically continued for as long as the immunosuppressive regimen is administered.

One common vexing problem encountered in long-term management is the role of radiographic follow-up. Analysis of data from the phase 2 DEFEAT Mucor study indicated that early radiographic progression (within the first 2 weeks) did not predict long-term survival. Caution should be used in reacting to short-term, serial radiographic results, and greater emphasis should be placed on clinical response, particularly within the first 2–4 weeks after initiation of therapy.

TABLE 213-2 Antifungal Options for the Treatment of Mucormycosis[a]

Drug	Recommended Dosage	Advantages and Supporting Studies	Disadvantages
First-Line Antifungal Therapy			
AmB deoxycholate	1.0–1.5 mg/kg once per day	• >5 decades of clinical experience • Inexpensive • FDA-approved for treatment of mucormycosis	• Highly toxic • Poor CNS penetration
LAmB	5–10 mg/kg once per day	• Less nephrotoxic than AmB deoxycholate • Better CNS penetration than AmB deoxycholate or ABLC • Better outcomes than with AmB deoxycholate in murine models and a retrospective clinical review	• Expensive
ABLC	5 mg/kg once per day	• Less nephrotoxic than AmB deoxycholate • Murine and retrospective clinical data suggest benefit of combination therapy with echinocandins	• Expensive • Possibly less efficacious than LAmB for CNS infection
Second-Line/Salvage Option			

Drug	Recommended Dosage	Advantages and Supporting Studies	Disadvantages
Isavuconazole	200 mg of isavuconazole (372 mg of isavuconazonium sulfate), load q8h × 6 followed by once-daily dosing	• Efficacy similar to that of LAmB in mouse models • FDA-approved for treatment of mucormycosis • May be a rational empirical option when septate mold vs. mucormycosis is not yet established	• Much less clinical experience; concern about a more slowly cidal agent than lipid polyenes • Clinical study supporting approval was small and historically controlled.
Posaconazole	200 mg four times per day	• In vitro activity against the Mucorales, with lower MICs than isavuconazole • Retrospective data for salvage therapy in mucormycosis	• Substantially lower blood levels than isavuconazole • No data on initial therapy for mucormycosis, and no evidence for combination therapy with posaconazole • Experience limited, potential use for salvage therapy
Combination Therapy[b]			
Echinocandin plus lipid polyene	Standard echinocandin doses	• Favorable toxicity profile • Synergistic in murine disseminated mucormycosis • Retrospective clinical data suggest superior outcomes for rhino-orbital-cerebral mucormycosis.	• Limited clinical data on combination therapy
Lipid polyene plus azole (posaconazole or isavuconazole)	Standard doses	• Favorable toxicity profile	• Limited efficacy data, with no available evidence of superiority vs. monotherapy

Drug	Recommended Dosage	Advantages and Supporting Studies	Disadvantages
Triple therapy (lipid polyene plus echinocandin plus azole)	Standard doses	• Maximal aggressiveness	• Limited efficacy data, with no available evidence for superiority vs. monotherapy or dual therapy

[a]Primary therapy should generally include a polyene. Non-polyene-based regimens may be appropriate for patients who refuse polyene therapy or for relatively immunocompetent patients with mild disease (e.g., isolated suprafascial cutaneous infection) that can be surgically eradicated.
[b]Prospective randomized trials are necessary to confirm the suggested benefit (from animal and small retrospective human studies) of combination therapy for mucormycosis. Dose escalation of any echinocandin is not recommended because of a paradoxical loss of benefit of combination therapy at echinocandin doses of ≥3 mg/kg per day.

Abbreviations: ABLC, AmB lipid complex; AmB, amphotericin B; CNS, central nervous system; FDA, U.S. Food and Drug Administration; LAmB, liposomal AmB; MIC, minimal inhibitory concentration.

Source: Modified from B Spellberg et al: Clin Infect Dis 48:1743, 2009.

PROGNOSIS

Over the past two decades, the prognosis of mucormycosis has substantially improved with aggressive antifungal therapy. Even CNS infection is often successfully treated. The key driver of outcome may now be control of the patient's predisposing condition. In the past, experts often recommended delaying chemotherapy in infected patients with cancer in order to try to eradicate the fungus. However, cure of mucormycosis is not likely to be effected until underlying malignancy is controlled. Thus, a balanced approach is required. It may be far more harmful to long-term success to withhold chemotherapy than to try to treat the patient with antifungal agents during chemotherapy; some consideration can be given to moderating the aggressiveness of the chemotherapy and the resulting duration and depth of neutropenia.

FURTHER READING

Cornely OA et al: ESCMID and ECMM joint clinical guidelines for the diagnosis and management of mucormycosis 2013. Clin Microbiol Infect 20(S3):5, 2014. [PubMed: 24479848]
Kennedy KJ et al: Mucormycosis in Australia: Contemporary epidemiology and outcomes. Clin Microbiol Infect 22:775, 2016. [PubMed: 26806139]
Kontoyiannis DP et al: Prospective antifungal therapy (PATH) alliance: Focus on mucormycosis. Mycoses 57:240, 2014. [PubMed: 24147728]
Pettrikos G et al: Epidemiology of mucormycosis in Europe. Clin Microbiol Infect 20(S3):67, 2014. [PubMed: 24476100]
Roden MM et al: Epidemiology and outcome of zygomycosis: A review of 929 reported cases. Clin Infect Dis 41:634, 2005. [PubMed: 16080086]
Spellberg B et al: Novel perspectives on mucormycosis: Pathophysiology, presentation, and management. Clin Microbiol Rev 18:556, 2005. [PubMed: 16020690]
Spellberg B et al: Recent advances in the management of mucormycosis: From bench to bedside. Clin Infect Dis 48:1743, 2009. [PubMed: 19435437]
Spellberg B et al: Combination therapy for mucormycosis: Why, what, and how? Clin Infect Dis 54(S1):S732012. [PubMed: 22247449]

Spellberg B et al: Risk factors for mortality in patients with mucormycosis. Med Mycol 50:611, 2012. [PubMed: 22435877]

Chapter 214: Superficial Mycoses and Less Common Systemic Mycoses

Carol A. Kauffman

ENDEMIC MYCOSES (DIMORPHIC FUNGI)

Dimorphic fungi exist in discrete environmental niches as molds that produce conidia, which are their infectious form. In tissues and at temperatures of >35°C, the mold converts to the yeast form. **Other endemic mycoses—histoplasmosis, coccidioidomycosis, and blastomycosis—are discussed in Chaps. 207, 208, and 209, respectively**.

SPOROTRICHOSIS

Etiologic Agent, Epidemiology, and Pathogenesis

Sporothrix schenckii is a thermally dimorphic fungus that is found worldwide in sphagnum moss, decaying vegetation, and soil. Sporotrichosis most commonly affects persons who participate in outdoor activities such as landscaping, gardening, and tree farming. Infected animals can transmit *S. schenckii* to humans. A large ongoing outbreak of sporotrichosis in Rio de Janeiro has been traced to cats, which are highly susceptible to this infection. Sporotrichosis is primarily a localized infection of skin and subcutaneous tissues that follows traumatic inoculation of conidia. Osteoarticular sporotrichosis is uncommon, occurring most often in middle-aged men who abuse alcohol, and pulmonary sporotrichosis occurs almost exclusively in persons with chronic obstructive pulmonary disease who have inhaled the organism from the environment. Dissemination occurs rarely, almost always affecting markedly immunocompromised patients, especially those with AIDS.

Clinical Manifestations and Differential Diagnosis

Days or weeks after inoculation, a papule develops at the site and then usually ulcerates but is not very painful. Similar lesions develop sequentially along the lymphatic channels proximal to the original lesion **(Fig. 214-1)**. Some patients develop a fixed cutaneous lesion that can be verrucous or ulcerative and that remains localized without lymphatic extension. The differential diagnosis of lymphocutaneous sporotrichosis includes nocardiosis, tularemia, nontuberculous mycobacterial infection (especially that due to *Mycobacterium marinum*), and leishmaniasis. Osteoarticular sporotrichosis can present as chronic synovitis or septic arthritis. Pulmonary sporotrichosis must be differentiated from tuberculosis and from other fungal pneumonias. Numerous ulcerated skin lesions, with or without spread to visceral organs (including the central nervous system [CNS]), are characteristic of disseminated sporotrichosis.

FIGURE 214-1

Several nodular lesions that developed after a young boy pricked his index finger with a thorn. A culture yielded *S. schenckii. (Courtesy of Dr. Angela Restrepo.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Diagnosis

S. schenckii usually grows readily as a mold on Sabouraud's agar when material from a cutaneous lesion is incubated at room temperature. Histopathologic examination of biopsy material shows a mixed granulomatous and pyogenic reaction, and tiny oval or cigar-shaped yeasts sometimes can be seen with special stains. In cases in which the organism has not grown, polymerase chain reaction (PCR) of tissue samples can sometimes be helpful.

Treatment and Prognosis

Guidelines for the management of the various forms of sporotrichosis have been published by the Infectious Diseases Society of America **(Table 214-1)**. Itraconazole is the drug of choice for lymphocutaneous sporotrichosis. Fluconazole is less effective, voriconazole is not effective, and posaconazole has been used successfully in a few instances. Saturated solution of potassium iodide (SSKI) continues to be used for lymphocutaneous infection because it costs much less than itraconazole. However, SSKI is poorly tolerated because of adverse reactions, including metallic taste, salivary gland swelling, rash, and fever. High-dose terbinafine may be effective for lymphocutaneous infection. Treatment for lymphocutaneous sporotrichosis is continued for 2–4 weeks after all lesions have resolved, usually for a total of 3–6 months. The success rate for treatment of lymphocutaneous sporotrichosis is 90–100%.

Pulmonary and osteoarticular forms of sporotrichosis are treated with itraconazole for at least 1 year. Severe pulmonary infection and disseminated sporotrichosis, including that involving the CNS, should be treated initially with amphotericin B (AmB), with a switch to itraconazole after improvement has been noted. Lifelong suppressive therapy with itraconazole often is required for AIDS patients. These forms of sporotrichosis respond poorly to antifungal therapy.

TABLE 214-1 Suggested Treatment for Endemic Mycoses

Disease	First-Line Therapy	Alternatives/Comments
Sporotrichosis		
Cutaneous, lymphocutaneous	Itraconazole, 200 mg/d until 2–4 weeks after lesions resolve	SSKI, increasing doses[a] Terbinafine, 500 mg bid
Pulmonary, osteoarticular	Itraconazole, 200 mg bid for 12 months	Lipid AmB[b] for severe pulmonary disease until stable; then itraconazole
Disseminated, central nervous system	Lipid AmB[b] for 4–6 weeks	Itraconazole, 200 mg bid after AmB for 12 months AIDS patients: itraconazole maintenance, 200 mg/d until CD4+ T cell count is >200/μL for ≥12 months
Paracoccidioidomycosis		
Chronic (adult form)	Itraconazole, 100–200 mg/d for 6–12 months	Voriconazole, 200 mg bid for 6–12 months Posaconazole, 300 mg/d for 6–12 months TMP-SMX, 160/800 mg bid for 12–36 months
Acute (juvenile form)	AmB[c] or lipid AmB[b] until improvement	Itraconazole, 200 mg bid after AmB for 12 months Voriconazole or posaconazole at doses noted above may be used
Talaromycosis (Penicilliosis)		
Mild or moderate	Itraconazole, 200 mg bid for 12 weeks	Voriconazole, 200 mg bid
Severe	Lipid AmB[b] or AmB[c] until improvement	Itraconazole, 200 mg bid after AmB for 12 weeks
Maintenance therapy (AIDS)	200 mg/d until CD4+ T cell count is >100/μL for ≥6 months	

[a]The starting dosage is 5–10 drops tid in water or juice. The dosage is increased weekly by 10 drops per dose, as tolerated, up to 40–50 drops tid.

[b]The dosage of lipid AmB is 3–5 mg/kg daily; the higher dosage should be used when the central nervous system is involved.

[c]The dosage of AmB deoxycholate is 0.6–1.0 mg/kg daily.

Abbreviations: AmB, amphotericin B; SSKI, saturated solution of potassium iodide; TMP-SMX, trimethoprim-sulfamethoxazole.

PARACOCCIDIOIDOMYCOSIS

Etiologic Agent, Epidemiology, and Pathogenesis

Paracoccidioides brasiliensis is a thermally dimorphic fungus that is found in humid areas of Central and South America and is highly endemic in Brazil. A striking male-to-female ratio varies from 14:1 to as high as 70:1 in various reports. Most patients are middle-aged or elderly men from rural

areas. Paracoccidioidomycosis develops after the inhalation of aerosolized conidia encountered in the environment. For most patients, disease rarely develops at the time of the initial infection but appears years later, presumably after reactivation of a latent infection.

Clinical Manifestations

Two major syndromes are associated with paracoccidioidomycosis: the acute or juvenile form and the chronic or adult form. The acute form is uncommon, occurs mostly in persons <30 years old, and manifests as disseminated infection of the reticuloendothelial system. Immunocompromised individuals also manifest this type of rapidly progressive disease. The chronic form of paracoccidioidomycosis accounts for ~90% of cases and predominantly affects older men. The primary manifestation is progressive pulmonary disease, primarily in the lower lobes, with fibrosis. Ulcerative and nodular mucocutaneous lesions in the nares and mouth—another common manifestation of chronic paracoccidioidomycosis—must be differentiated from leishmaniasis **(Chap. 221)** and squamous cell carcinoma **(Chap. 72)**.

Diagnosis

The diagnosis is established by growth of the mold form of *P. brasiliensis* in culture at room temperature. A presumptive diagnosis can be made by detection of the distinctive thick-walled yeast, which has multiple narrow-necked buds attached circumferentially, in purulent material or tissue biopsies.

Treatment and Prognosis

Itraconazole is the treatment of choice for paracoccidioidomycosis (Table 214-1). Ketoconazole is also effective but more toxic; voriconazole and posaconazole also appear to be effective. Sulfonamides have been used for years and are the least costly agents; however, the response is slower and the relapse rate higher. Seriously ill patients should be treated with AmB initially. Patients with paracoccidioidomycosis have an excellent response to therapy, but pulmonary fibrosis can be progressive in those with chronic disease.

TALAROMYCOSIS (PENICILLIOSIS)

Etiologic Agent, Epidemiology, and Pathogenesis

 *Talaromyces marneffei (formerly Penicillium marneffei)* is a thermally dimorphic fungus that is endemic in the soil in certain areas of Vietnam, Thailand, and other southeastern Asian countries. The epidemiology of talaromycosis is linked to bamboo rats that are infected with the fungus but rarely manifest disease. The disease occurs most often among persons living in rural areas in which the rats are found, but there is no evidence for transmission of the infection directly from rats to humans. Infection is rare in immunocompetent hosts, and most cases are reported in persons who have advanced AIDS. Infection results from the inhalation of conidia from the environment. The organism converts to the yeast phase in the lungs and then spreads hematogenously to the reticuloendothelial system.

Clinical Manifestations

The clinical manifestations of talaromycosis mimic those of disseminated histoplasmosis and include fever, fatigue, weight loss, dyspnea, diarrhea (in some cases), lymphadenopathy, hepatosplenomegaly, and skin lesions, which appear as papules that often umbilicate and resemble molluscum contagiosum **(Chap. 191)**.

Diagnosis

Talaromycosis is diagnosed by culture of *T. marneffei* from blood or from biopsy samples of skin, bone marrow, or lymph node. The organism usually grows within 1 week as a mold producing a distinctive red pigment that diffuses into the agar. Histopathologic examination of tissues and smears of blood or material from skin lesions shows oval or elliptical yeast-like organisms with central septation and can quickly establish a presumptive diagnosis.

Treatment and Prognosis

For mild or moderate infection, itraconazole is the drug of choice; voriconazole can also be used. Severe infection should be treated with AmB until improvement occurs; then therapy can be changed to itraconazole (Table 214-1). For patients with AIDS, suppressive therapy with itraconazole is recommended until the CD4+ T cell count has been >100 cells/μL for at least 6 months. Disseminated talaromycosis is usually fatal if not treated. With treatment, the mortality rate is ~10%.

PHAEOHYPHOMYCOSES

Dematiaceous or brown-black fungi, the common soil organisms that cause phaeohyphomycoses, contain melanin, which causes the hyphae and conidia to be darkly pigmented. The term *phaeohyphomycosis* is used to describe any infection with a pigmented mold. This definition encompasses two specific syndromes—eumycetoma and chromoblastomycosis—as well as all other types of infections caused by these organisms. It is important to note that eumycetomas can be caused by hyaline molds as well as by brown-black molds and that only about half of all mycetomas are due to fungi. Actinomycetes cause the remainder **(Chap. 169)**. Most dematiaceous fungi cause localized subcutaneous infections after direct inoculation, but disseminated infections and serious focal visceral infections do occur, especially in immunocompromised patients.

Etiologic Agents

A large number of pigmented molds can cause human infection. Most are found in the soil or on plants, and some cause economically important plant diseases. Disseminated infection and focal visceral infections are caused by a variety of dematiaceous fungi; *Alternaria*, *Exophiala*, *Curvularia*, and *Wangiella* species are among the more common molds reported to cause human infection. In 2012, *Exserohilum* species caused a large outbreak of severe, sometimes fatal CNS and osteoarticular infections after the injection of methylprednisolone contaminated with this fungus. The most common cause of eumycetoma is *Madurella* species. *Fonsecaea* and *Cladophialophora* species are responsible for most cases of chromoblastomycosis.

Epidemiology and Pathogenesis

Infections with dematiaceous molds are acquired by traumatic inoculation into the eye or through the skin, by inhalation, or by injection of contaminated medication. Melanin is a virulence factor for all the pigmented molds. Several organisms, specifically *Cladophialophora bantiana* and *Rhinocladiella mackenziei*, are neurotropic and likely to cause CNS infection. When a patient is immunocompromised or when a pigmented mold is injected directly into a deep structure, these organisms become opportunists, invading blood vessels and mimicking better-known opportunistic infections, such as aspergillosis. Eumycetoma and chromoblastomycosis are acquired by inoculation through the skin; these two syndromes are seen almost entirely in tropical and subtropical areas and occur mostly in rural laborers who are frequently exposed to the organisms.

Clinical Manifestations

Dematiaceous molds are the most common cause of allergic fungal sinusitis and a less common cause of invasive fungal sinusitis. Keratitis occurs with traumatic corneal inoculation. Even in many immunocompromised patients, inoculation through the skin generally produces only localized cyst-like, nodular lesions at the entry site. However, other immunocompromised patients develop pneumonia, brain abscess, or disseminated infection. In the outbreak mentioned above, epidural injection of *Exserohilum*-contaminated glucocorticoids led to meningitis, basilar stroke, epidural abscess and phlegmon, vertebral osteomyelitis, and arachnoiditis.

Eumycetoma is a chronic subcutaneous and cutaneous infection that usually occurs on the lower extremities and that is characterized by swelling, the development of sinus tracts, and the appearance of grains that are actually colonies of fungi discharged from the sinus tract. As the infection progresses, adjacent fascia and bony structures become involved. The disease is indolent and disfiguring, progressing slowly over years. Complications include fractures of infected bone and bacterial superinfection.

Chromoblastomycosis is an indolent subcutaneous infection characterized by nodular, verrucous, or plaque-like painless lesions that occur predominantly on the lower extremities and grow slowly over months to years. There is hardly ever extension to adjacent structures, as is seen with eumycetoma. Long-term consequences include bacterial superinfection, chronic lymphedema, and (rarely) the development of squamous cell carcinoma.

Diagnosis

The specific diagnosis of infection with a pigmented mold is established by growth of the organism in culture, which is essential to differentiate infection with a hyaline mold (e.g., *Aspergillus* or *Fusarium*) from that due to a pigmented mold. A tentative clinical diagnosis of mycetoma can be made when a patient presents with a lesion characterized by swelling, sinus tracts, and grains. Histopathologic examination and culture are necessary to confirm that the etiologic agent is a mold and not an actinomycete. In chromoblastomycosis, the diagnosis rests on the histologic demonstration of sclerotic bodies (dark brown, thick-walled, septate fungal forms that resemble large yeasts) in the tissues; culture establishes which pigmented mold is causing the infection. PCR assays are increasingly used in the diagnosis of infection due to dematiaceous molds but are available only through fungal reference laboratories.

Treatment and Prognosis

The choice of antifungal agent to treat disseminated and focal visceral infections with brown-black molds is based on the location and extent of the infection, in vitro test results, and clinical experience with the specific infecting organism. AmB is not effective against many of these organisms but has been used successfully against some species **(Table 214-2)**. Itraconazole, voriconazole, or posaconazole can be used in the treatment of localized infections. Voriconazole is preferred when infections involve the CNS because this drug reaches adequate concentrations at that site. Voriconazole or posaconazole could be used for disseminated infection; these agents are available as both IV and well-absorbed oral formulations. Disseminated and focal visceral infections, especially those involving the CNS, are associated with high mortality rates.

TABLE 214-2 Suggested Treatment for Phaeohyphomycoses and Opportunistic Infections

Disease	First-Line Therapy	Alternatives/Comments
Phaeohyphomycoses	Voriconazole, 200 mg bid Itraconazole, 200 mg bid Posaconazole, 300 mg/d	Lipid AmB may be effective against some mold species.
Fusariosis	Voriconazole, 200–300 mg bid Lipid AmB, 5 mg/kg per day	Lipid AmB plus voriconazole or posaconazole is used by some physicians for initial therapy.
Scedosporiosis	Posaconazole, 300 mg/d Voriconazole, 200–300 mg bid	Not susceptible to AmB
	Posaconazole, 300 mg/d	*Scedosporium prolificans* is resistant to almost all antifungal drugs.
Trichosporonosis	Voriconazole, 200–300 mg bid	Posaconazole, 300 mg/d

Abbreviation: AmB, amphotericin B.

Treatment of eumycetoma and chromoblastomycosis involves both surgical extirpation of the lesion and use of antifungal agents. Surgical removal of the lesions is most effective if performed before extensive spread has occurred. In chromoblastomycosis, cryosurgery and laser therapy have been used with variable success. The antifungal agents of choice are itraconazole, voriconazole, and posaconazole. The most experience has accrued with itraconazole; less experience has been gained with the newer azoles, which are active in vitro and have been reported to be effective in a few patients. Flucytosine and terbinafine also have been used to treat chromoblastomycosis. Chromoblastomycosis and eumycetoma are chronic indolent infections that are difficult to cure but are not life-threatening.

OPPORTUNISTIC FUNGAL INFECTIONS

Two genera of hyaline (nonpigmented) molds, *Fusarium* and *Scedosporium*, and one yeast-like genus, *Trichosporon*, have become prominent pathogens among immunocompromised patients. Infections caused by *Fusarium* and *Scedosporium* species overlap with invasive aspergillosis in their clinical manifestations; when seen in tissues, these organisms appear similar to *Aspergillus*. In the immunocompetent host, these fungi cause localized infections of skin, skin structures, and subcutaneous tissues, but their role as causes of infection in immunocompromised patients will be emphasized in this section.

FUSARIOSIS

Etiologic Agent, Epidemiology, and Pathogenesis

Fusarium species, which are found worldwide in soil and on plants, have emerged as major opportunists in markedly immunocompromised patients. Most human infections follow inhalation of conidia, but ingestion and direct inoculation also can lead to disease. An outbreak of severe *Fusarium* keratitis among soft contact lens wearers was traced back to a particular brand of contact lens solution and individual contact lens cases that had been contaminated. Disseminated infection is reported most often in patients who have a hematologic malignancy, are neutropenic, have received a hematopoietic cell or solid organ transplant, or have a severe burn.

Clinical Manifestations

In immunocompetent persons, *Fusarium* species cause localized infections of various organs. These organisms commonly cause fungal keratitis, which can extend into the anterior chamber of the eye; cause loss of vision; and require corneal transplantation. Onychomycosis due to *Fusarium* species, while basically an annoyance in immunocompetent patients, is a source of subsequent hematogenous dissemination and should be aggressively sought and treated in neutropenic patients. In profoundly immunocompromised patients, fusariosis is angioinvasive, and clinical manifestations mimic those of aspergillosis. Pulmonary infection is characterized by multiple nodular lesions. Sinus infection is likely to lead to invasion of adjacent structures. Disseminated fusariosis occurs primarily in neutropenic patients with hematologic malignancies and in allogeneic hematopoietic cell transplant recipients, especially those with graft-versus-host disease. Disseminated fusariosis differs from disseminated aspergillosis in that skin lesions are extremely common with fusariosis; the lesions are nodular or necrotic, are usually painful, and appear over time in different locations **(Fig. 214-2)**.

FIGURE 214-2

Painful necrotic foot lesion that developed over a week in a woman who had acute leukemia and who had been neutropenic for 2 months. *Fusarium* species were grown from a punch biopsy. *(Courtesy of Dr. Nessrine Ktaich.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Diagnosis

The diagnostic approach usually includes both documentation of the growth of *Fusarium* species from involved tissue and demonstration of invasion by histopathologic techniques that show septate hyphae in tissues. The organism is difficult to differentiate from *Aspergillus* species in tissues; thus, identification with culture is imperative. An extremely helpful diagnostic clue is growth in blood cultures, which are positive in as many as 50% of patients with disseminated fusariosis.

Treatment and Prognosis

Fusarium species are resistant to many antifungal agents. A lipid formulation of AmB, voriconazole, or posaconazole is recommended. Many physicians use both a lipid formulation of AmB and either voriconazole or posaconazole because susceptibility information is not available when therapy must be initiated. Serum drug levels should be monitored with either azole to ensure that absorption is adequate and with voriconazole to avoid toxicity. Mortality rates for disseminated fusariosis have been as high as 85%. With the improved antifungal therapy now available, mortality rates have fallen to ~50%. However, if neutropenia persists, the mortality rate approaches 100%.

SCEDOSPORIOSIS

Etiologic Agent

The genus *Scedosporium* includes several pathogens. The major causes of human infections are the *Scedosporium apiospermum* complex (composed of several species) and *Scedosporium prolificans*, which has been renamed *Lomentospora prolificans*.

Epidemiology and Pathogenesis

Organisms of the *S. apiospermum* complex are found worldwide in temperate climates in tidal flats, swamps, ponds, manure, and soil. These organisms cause pneumonia, disseminated infection, and brain abscess and are common pathogens in near-drowning victims. *L. prolificans* is also found in soil but is more geographically restricted. Infection occurs predominantly through inhalation of conidia, but direct inoculation through the skin or into the eye also can occur.

Clinical Manifestations

Among immunocompetent persons, *Scedosporium* species are a prominent cause of eumycetoma. Keratitis as a result of accidental corneal inoculation is a sight-threatening infection. In patients who have hematologic malignancies (especially acute leukemia with neutropenia), recipients of solid organ or hematopoietic stem cell transplants, and patients receiving glucocorticoids, *Scedosporium* species are angioinvasive, causing pneumonia and widespread dissemination. Pulmonary infection mimics aspergillosis; nodules, cavities, and lobar infiltrates are common. Disseminated infection involves the skin, heart, brain, and many other organs. Skin lesions are not as common or as painful as those of fusariosis.

Diagnosis

Diagnosis depends on the growth of *Scedosporium* species from involved tissue and the demonstration of invasion by histopathologic techniques that show septate hyphae in tissues. Culture evidence is essential because *Scedosporium* species are difficult to differentiate from *Aspergillus* in tissues, and demonstration of tissue invasion is essential because these ubiquitous environmental molds can be mere contaminants or colonizers. *L. prolificans* can grow in blood cultures, but *S. apiospermum* usually does not.

Treatment and Prognosis

Scedosporium species are resistant to AmB, echinocandins, and some azoles. Voriconazole is the agent of choice for *S. apiospermum*, and posaconazole also can be used for this infection. *L. prolificans* is resistant in vitro to almost every available antifungal agent; the addition of agents such as terbinafine to a voriconazole regimen has been attempted because in vitro data suggest possible synergy against some strains of *L. prolificans*. Mortality rates for invasive *S. apiospermum* infection are ~50%, but those for invasive *L. prolificans* infection remain as high as 85–100%.

TRICHOSPORONOSIS

Etiologic Agent

The genus *Trichosporon* contains many species, some of which cause localized infection of hair and nails. The major pathogen responsible for invasive infection is *Trichosporon asahii. Trichosporon* species grow as yeast-like colonies in vitro; in vivo, however, hyphae, pseudohyphae, and arthroconidia, in addition to yeast forms, can be seen.

Epidemiology and Pathogenesis

These yeasts are commonly found in soil, sewage, and water and in rare instances can colonize human skin and the human gastrointestinal tract. Most infections follow inhalation or entry via central venous catheters. Systemic infection occurs almost exclusively in immunocompromised hosts, including those who have hematologic malignancies, are neutropenic, have received a solid organ transplant, or are receiving glucocorticoids.

Clinical Manifestations

Disseminated trichosporonosis resembles invasive candidiasis, and fungemia is often the initial manifestation of infection. Pneumonia, skin lesions, and sepsis are common. The skin lesions begin as papules or nodules surrounded by erythema and progress to central necrosis. A chronic form of infection mimics hepatosplenic candidiasis (chronic disseminated candidiasis).

Diagnosis

The diagnosis of systemic *Trichosporon* infection is established by growth of the organism from involved tissues or from blood. Histopathologic examination of a skin lesion showing a mixture of yeast forms, arthroconidia, and hyphae can lead to an early presumptive diagnosis of trichosporonosis. The serum cryptococcal antigen latex agglutination test may be positive in patients with disseminated trichosporonosis because *T. asahii* and *Cryptococcus neoformans* share polysaccharide antigens.

Treatment and Prognosis

Rates of response to AmB have been disappointing, and many *Trichosporon* isolates are resistant in vitro. Voriconazole is the antifungal agent of choice. The mortality rates for disseminated *Trichosporon* infection have been as high as 70% but are decreasing with the use of voriconazole; however, patients who remain neutropenic are likely to succumb to this infection.

SUPERFICIAL CUTANEOUS INFECTIONS

Fungal infections of the skin and skin structures are caused by molds and yeasts that do not invade deeper tissues but rather cause disease merely by inhabiting the superficial layers of skin, hair follicles, and nails. These agents are the most common fungal infections of humans but only rarely cause serious infections.

YEAST INFECTIONS

Etiologic Agents, Epidemiology, and Pathogenesis

The lipophilic yeast *Malassezia* is dimorphic in that it colonizes the skin in the yeast phase but transforms to the mold phase when it causes disease. *Malassezia* species are part of the indigenous human flora found in the stratum corneum of the back, chest, scalp, and face—areas rich in sebaceous glands. The organisms do not invade below the stratum corneum and generally elicit little if any inflammatory response.

Clinical Manifestations

Malassezia species cause tinea versicolor (also called *pityriasis versicolor*), folliculitis, and seborrheic dermatitis. Tinea versicolor presents as flat round scaly patches of hypo- or hyperpigmented skin on the neck, chest, or upper arms. The lesions are usually asymptomatic but can be pruritic. They can be mistaken for vitiligo, but the latter is not scaly. Folliculitis occurs on the back and chest and mimics bacterial folliculitis. Seborrheic dermatitis manifests as erythematous pruritic scaly lesions in the eyebrows, moustache, nasolabial folds, and scalp (*dandruff*). Seborrheic dermatitis can be severe in patients with advanced AIDS. Fungemia and disseminated infection occur rarely with *Malassezia* species—almost always in premature neonates receiving parenteral lipid preparations through a central venous catheter.

Diagnosis

Malassezia infections are diagnosed clinically in most cases. If scrapings are collected on a microscope slide on which a drop of potassium hydroxide has been placed, a mixture of budding yeasts and short septate hyphae is seen. In order to culture *M. furfur* from those patients in whom disseminated infection is suspected, sterile olive oil must be added to the medium.

Treatment and Prognosis

Topical creams and lotions, including selenium sulfide shampoo, ketoconazole shampoo or cream, and terbinafine cream, are effective in treating *Malassezia* infections and are usually given for 2 weeks. Other more expensive antifungal creams are rarely needed. Mild topical steroid creams are sometimes used to treat seborrheic dermatitis. For extensive disease, oral itraconazole or fluconazole (200 mg daily) can be used for 5–7 days. The rare cases of fungemia caused by *Malassezia* species are treated with AmB or fluconazole, prompt removal of the catheter, and discontinuance of parenteral lipid infusions. *Malassezia* skin infections are benign and self-limited, although recurrences are the rule. The outcome of systemic infection depends on the host's underlying conditions, but most infected infants do well.

DERMATOPHYTE (MOLD) INFECTIONS

Etiologic Agents

The molds that cause skin infections in humans include the genera *Trichophyton, Microsporum*, and *Epidermophyton*. These organisms, which are not components of the normal skin flora, can live within the keratinized structures of the skin—hence the term *dermatophytes*.

Epidemiology and Pathogenesis

Dermatophytes occur worldwide, and infections with these organisms are extremely common. Some organisms cause disease only in humans and can be transmitted by person-to-person contact and by fomites, such as hairbrushes or wet floors, that have been contaminated by infected individuals. Several species cause infections in cats and dogs and can readily be transmitted from these animals to humans. Finally, some dermatophytes are spread from contact with soil. The characteristic ring shape of cutaneous lesions is the result of the organisms' outward growth in a centrifugal pattern in the stratum corneum. Fungal invasion of the nail usually occurs through the lateral or superficial nail plates and then spreads throughout the nail; when hair shafts are invaded, the organisms can be found either within the shaft or surrounding it. Symptoms are caused by the inflammatory reaction elicited by fungal antigens and not by tissue invasion. Dermatophyte infections occur more commonly in males than in females, and progesterone has been shown to inhibit dermatophyte growth.

Clinical Manifestations

Dermatophyte infection of the skin is often called *ringworm*. This term is confusing because worms are not involved. *Tinea*, the Latin word for *worm*, describes the serpentine nature of the skin lesions and is a less confusing designation that is used in conjunction with the name of the body part affected—e.g., tinea capitis (head), tinea pedis (feet), tinea corporis (body), tinea cruris (crotch), and tinea unguium (nails, although infection at this site is more often termed *onychomycosis*).

Tinea capitis occurs most commonly in children 3–7 years old. Children with tinea capitis usually present with well-demarcated scaly patches in which hair shafts are broken off right above the skin;

alopecia can result. Tinea corporis is manifested by well-demarcated, annular, pruritic, scaly lesions that undergo central clearing. Usually one or several small lesions are present. However, in some patients, tinea corporis can involve much of the trunk. The rash should be differentiated from contact dermatitis, eczema, and psoriasis. Tinea cruris is seen almost exclusively in men. The perineal rash is erythematous and pustular, has a discrete scaly border, is without satellite lesions, and is usually pruritic. The rash must be differentiated from intertriginous candidiasis, erythrasma, and psoriasis.

Tinea pedis also is more common among men than among women. It usually starts in the web spaces of the toes; peeling, maceration, and pruritus are followed by development of a scaly pruritic rash along the lateral and plantar surfaces of the feet. Hyperkeratosis of the soles of the feet often ensues. Tinea pedis has been implicated in lower-extremity cellulitis, as streptococci and staphylococci can gain entrance to the tissues through fissures between the toes. Onychomycosis affects toenails more often than fingernails and is most common among persons who have tinea pedis. The nail becomes thickened and discolored and may crumble; onycholysis almost always occurs. Onychomycosis is more common in older adults and in persons with vascular disease, diabetes mellitus, and trauma to the nails. Fungal infection must be differentiated from psoriasis, which can mimic onychomycosis but usually has associated skin lesions.

Diagnosis

Many dermatophyte infections are diagnosed by their clinical appearance. If the diagnosis is in doubt, scrapings should be taken from the edge of a lesion with a scalpel blade, transferred to a slide to which a drop of potassium hydroxide is added, and examined under a microscope for the presence of hyphae. Cultures are indicated if an outbreak is suspected or the patient does not respond to therapy.

Treatment and Prognosis

Dermatophyte infections usually respond to topical therapy. Lotions or sprays are easier than creams to apply to large or hairy areas. Particularly for tinea cruris, the affected area should be kept as dry as possible. When patients have extensive skin lesions, oral itraconazole or terbinafine can hasten resolution **(Table 214-3)**. Terbinafine interacts with fewer drugs than itraconazole and is generally the first-line agent.

TABLE 214-3 Suggested Oral Treatment for Extensive Tinea Infections and Onychomycosis

Antifungal Agent	Suggested Dosage	Comments
Extensive Tinea Infection		
Terbinafine	250 mg/d for 1–2 weeks	Adverse reactions minimal with short treatment period
Itraconazole[a]	200 mg/d for 1–2 weeks	Adverse reactions minimal with short treatment period except for drug interactions
Onychomycosis		
Terbinafine	250 mg/d for 3 months	Slightly superior to itraconazole; monitor for hepatotoxicity
Itraconazole[a]	200 mg/d for 3 months *or* 200 mg bid for 1 week each month for 3 months	Drug interactions frequent; monitor for hepatotoxicity; rarely causes hypokalemia, hypertension, edema; use with caution in patients with congestive heart failure

[a]Itraconazole capsules require food and gastric acid for absorption, whereas itraconazole solution is taken on an empty stomach.

Onychomycosis does not respond to topical therapy although ciclopirox nail lacquer applied daily for a year is occasionally beneficial. Itraconazole and terbinafine both accumulate in the nail plate and can be used to treat onychomycosis (Table 214-3). The major decision to be made with regard to therapy

is whether the extent of nail involvement justifies the use of systemic antifungal agents that have adverse effects, may interact with other drugs, and are costly. Treating for cosmetic reasons alone is discouraged. Relapses of tinea cruris and tinea pedis are common and should be treated early with topical creams to avoid development of more extensive disease. Relapses of onychomycosis follow treatment in 25–30% of cases.

FURTHER READING

Chowdhary A et al: ESCMID and ECCM joint clinical guidelines for the diagnosis and management of systemic phaeohyphomycosis: Diseases caused by black fungi. Clin Microbiol Infect 20(Suppl 3):47, 2014. [PubMed: 24483780]
de Berker D: Clinical practice. Fungal nail disease. N Engl J Med 360:2108, 2009. [PubMed: 19439745]
Garnica M et al: Difficult mycoses of the skin: Advances in the epidemiology and management of eumycetoma, phaeohyphomycosis, and chromoblastomycosis. Curr Opin Infect Dis 22:559, 2009. [PubMed: 19773651]
Gupta AK et al: Update in antifungal therapy of dermatophytosis. Mycopathologia 166:353, 2008. [PubMed: 18478357]
Kauffman CA et al: Clinical practice guidelines for the management of sporotrichosis: 2007 update by the Infectious Diseases Society of America. Clin Infect Dis 45:1255, 2007. [PubMed: 17968818]
Nelson KE et al: Penicilliosis, in *Essentials of Clinical Mycology*, 2nd ed. CA Kauffman et al (eds). New York, Springer, 2011, pp 399–411.
Nucci F et al: Fusariosis. Sem Respir Crit Care Med 36:706, 2015.
Restrepo A et al: Pulmonary paracoccidioidomycosis. Semin Resp Crit Care Med 29:182, 2008.
Rodriguez-Tudela JL et al: Epidemiology and outcome of *Scedosporium prolificans* infection, a review of 162 cases. Med Mycol 47:359, 2009. [PubMed: 19031336]
Ruan SY et al: Invasive trichosporonosis caused by *Trichosporon asahii* and other unusual *Trichosporon* species at a medical center in Taiwan. Clin Infect Dis 49:1457, 2009. [PubMed: 19824853]

Chapter 215: *Pneumocystis* Infections

Alison Morris; Henry Masur

DEFINITION AND DESCRIPTION

Pneumocystis is an opportunistic pathogen that is an important cause of pneumonia in immunocompromised hosts, particularly those with HIV infection **(Chap. 197)**, organ transplants, or hematologic malignancies and those receiving high-dose glucocorticoids or certain immunosuppressive monoclonal antibodies. *Pneumocystis* was discovered in rodents in 1906 and was initially believed to be a protozoan. Because *Pneumocystis* cannot be cultured, our understanding of its biology has been limited, but molecular techniques have demonstrated that the organism is actually a fungus. Formerly known as *Pneumocystis carinii*, the species infecting humans has been renamed *Pneumocystis jirovecii*.

EPIDEMIOLOGY

P. jirovecii pneumonia (PCP) came to medical attention when cases were reported in malnourished orphans in Europe during World War II. The disease was later recognized in other immunosuppressed populations but was rare in the era before HIV/AIDS and before intensive immunosuppressive therapy for organ transplantation and autoimmune disorders. In 1981, PCP was first reported in men who had sex with men and in IV drug users who had no obvious cause of immunosuppression. These cases were subsequently recognized as the first cases of what came to be known as the acquired immunodeficiency syndrome (AIDS) **(Chap. 197)**.

The incidence of PCP increased dramatically as the AIDS epidemic grew: without chemoprophylaxis or antiretroviral therapy (ART), 80–90% of patients with HIV/AIDS in North America and Western Europe ultimately develop one or more episodes of PCP. While its incidence declined with the introduction of anti-*Pneumocystis* prophylaxis and combination ART, PCP has continued to be a leading cause of AIDS-associated morbidity in the United States and Western Europe, particularly in individuals who do not know they are infected with HIV until they are profoundly immunosuppressed and in HIV-infected patients with CD4+ T lymphocyte counts of <200/μL who are not receiving ART or PCP prophylaxis.

PCP also develops in HIV-uninfected patients who are immunocompromised secondary to hematologic or malignant neoplasms, stem cell or solid organ transplantation, and treatment with immunosuppressive medications. The incidence of PCP depends on the degree and duration of immunosuppression. PCP is increasingly reported among individuals receiving tumor necrosis factor α inhibitors and antilymphocyte monoclonal antibodies for rheumatologic or other diseases. While clinical PCP in immunocompetent hosts has not been clearly documented, studies have shown that *Pneumocystis* organisms can colonize the airways of children and adults who are not overtly immunocompromised. The relevance of these organisms to acute or chronic syndromes, such as chronic obstructive pulmonary disease (COPD), in immunocompetent patients is being investigated.

 In some developing countries, the incidence of PCP among HIV-infected individuals has been found to be lower than that in industrialized countries. This lower incidence may be due to competing mortality from infectious diseases such as tuberculosis and bacterial pneumonia, which typically occur before patients become immunosuppressed enough to develop PCP. Geographic variations in *Pneumocystis* exposure and underdiagnosis attributable to lack of diagnostic resources also may explain the apparent lower frequency of PCP in some countries.

PATHOGENESIS AND PATHOLOGY

Life Cycle and Transmission

The life cycle of *Pneumocystis* involves both sexual and asexual reproduction, and the organism exists as a trophic form, a cyst, and a precyst at various points. Serologic and molecular studies have demonstrated that most humans are exposed to *Pneumocystis* early in life. It was historically thought that *Pneumocystis* developed from reactivation of latent infection, but de novo infections from environmental sources and person-to-person transmission occur as well. Outbreaks of PCP suggest that nosocomial transmission can take place, and studies with rodents show that immunocompetent animals can serve as reservoirs for transmission of *P. carinii* (the infecting species in rodents) to immunocompetent and immunosuppressed animals. However, *Pneumocystis* organisms are species-specific. Thus, humans are infected only by other humans who transmit *P. jirovecii*; humans cannot be infected with species of *Pneumocystis* that infect other animals, such as *P. carinii* (rats), *P. murina* (mice), or *P. oryctolagi* (rabbits). The utility of respiratory isolation in preventing transmission from patients with PCP to other immunosuppressed individuals has been debated; no clear evidence exists, although it seems prudent to isolate patients with active PCP from other immunosuppressed patients.

Role of Immunity

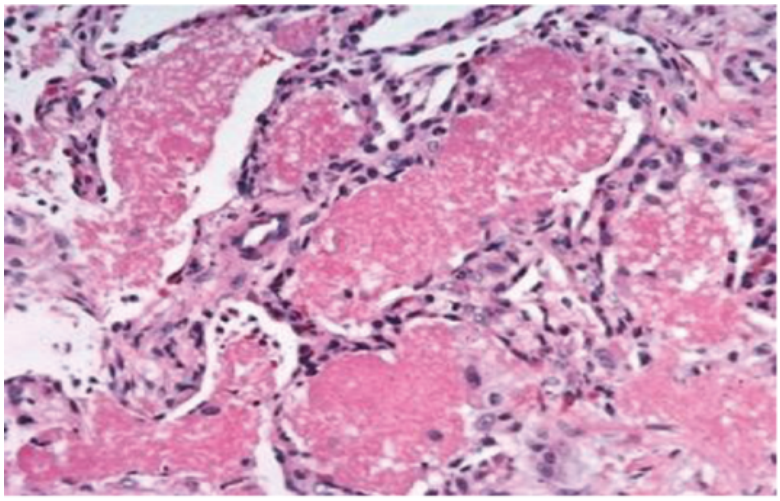
Defects in cellular and/or humoral immunity predispose to development of PCP. Such defects may be congenital, or they may be acquired as a result of HIV infection or of treatment with immunosuppressive drugs such as glucocorticoids, fludarabine, temozolomide, temsirolimus, cyclophosphamide, rituximab, or alemtuzumab. CD4+ T cells are critical in host defense against *Pneumocystis*. Among HIV-infected patients, the incidence is inversely related to the CD4+ T cell count: at least 80% of cases occur at counts of <200/μL, and most of these cases develop at counts of <100/μL. HIV load is another factor that predisposes patients to PCP. CD4+ T cell counts are less specific and thus less useful in predicting the risk of PCP in patients who are immunosuppressed for reasons other than HIV infection.

Lung Pathology

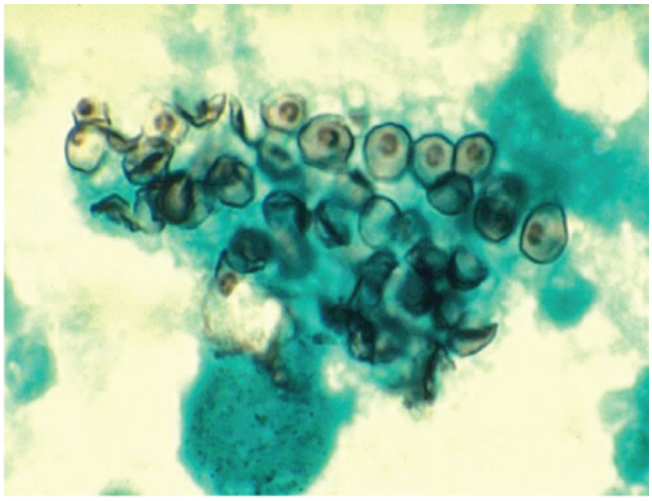
Pneumocystis has a unique tropism for the lung. It is presumably inhaled into the alveolar space. Clinically apparent pneumonia occurs only if an individual is immunocompromised. *Pneumocystis* proliferates in the lung, provoking a mononuclear cell response. The alveoli become filled with proteinaceous material, and alveolar damage results in increased alveolar-capillary injury and surfactant abnormalities. Stained lung sections typically show foamy, vacuolated alveolar exudates composed largely of viable and nonviable organisms **(Fig. 215-1A)**. Interstitial edema and fibrosis may develop, and organisms can be seen in the alveolar space with silver or other stains. Moreover, the organisms can be seen when tissue is subjected to colorimetric or immunofluorescent staining **(Fig. 215-1B–1D)**.

FIGURE 215-1

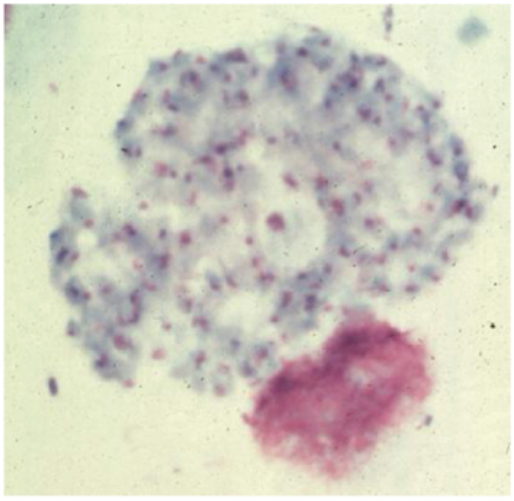
Direct microscopy of *Pneumocystis* pneumonia. *A.* Transbronchial lung biopsy stained with hematoxylin and eosin shows eosinophilic alveolar filling. ***B.*** Methenamine silver–stained bronchoalveolar lavage (BAL) fluid. ***C.*** Giemsa-stained BAL fluid. ***D.*** Immunofluorescent stain of BAL fluid.



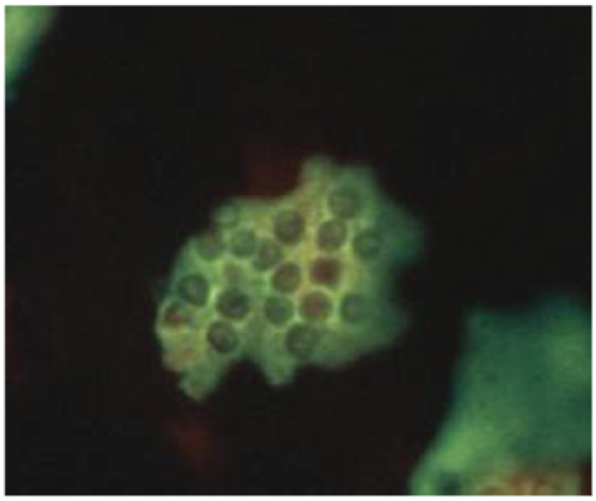
A



B



C



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CLINICAL FEATURES

Clinical Presentation

PCP presents as acute or subacute pneumonia that may initially be characterized by a vague sense of dyspnea alone but that subsequently manifests as fever and nonproductive cough with progressive shortness of breath, ultimately resulting in respiratory failure and death. Extrapulmonary manifestations of PCP are rare but can include involvement of almost any organ, most notably lymph nodes, spleen, and liver.

Physical Examination

The physical examination findings in PCP are nonspecific. Patients have decreased oxygen saturation—at rest or with exertion—that, without treatment, progresses to severe hypoxemia. Patients may initially have a normal chest examination and no adventitious sounds, but later develop diffuse rales and signs of consolidation.

Laboratory Findings

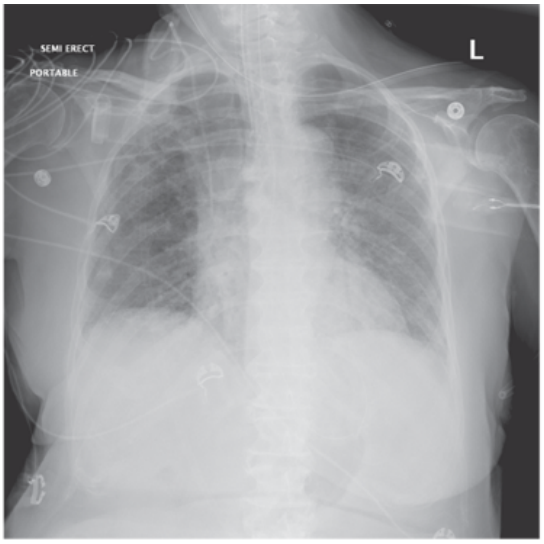
The results of routine laboratory tests are nonspecific in PCP. Serum levels of lactate dehydrogenase (LDH) are often elevated as a result of pulmonary damage; however, a normal LDH level does not rule out PCP, nor is an elevated LDH value specific for PCP. The peripheral white blood cell count may be elevated in relation to the patient's baseline values, but the increase is usually modest. Hepatic and renal function are typically normal.

Radiographic Findings

Although the initial chest radiograph may be normal when patients have mild symptoms, the classic radiographic appearance of symptomatic PCP consists of diffuse bilateral interstitial infiltrates that are perihilar and symmetric **(Fig. 215-2A)**—yet another finding that is not specific for PCP. The interstitial infiltrates can progress to alveolar filling **(Fig. 215-2B)**. High-resolution chest CT shows diffuse ground-glass opacities in virtually all patients with PCP **(Fig. 215-2C)**. A normal chest CT essentially rules out the diagnosis of PCP. Cysts and pneumothoraces are common chest radiographic findings, especially in patients with HIV infection **(Fig. 215-2D)**. A wide variety of atypical radiographic findings have been described, including asymmetric patterns, upper-lobe infiltrates, mediastinal adenopathy, nodules, cavities, and effusions.

FIGURE 215-2

Radiographs in *Pneumocystis* pneumonia. ***A.*** Posterior–anterior chest radiograph showing symmetric interstitial infiltrates. ***B.*** Posterior–anterior chest radiograph showing symmetric alveolar infiltrates (courtesy of Alison Morris). ***C.*** CT image demonstrating symmetric interstitial infiltrates and ground-glass opacities. ***D.*** CT image showing symmetric interstitial infiltrates, ground-glass opacities, and pneumatoceles.



A



B



C



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS

The optimal sample for diagnostic examination depends on how ill the patient is and what resources are available. Before the 1990s, diagnoses of PCP were usually established by open lung biopsy; later, transbronchial lung biopsy was employed. Hematoxylin and eosin staining of pulmonary tissue demonstrates a foamy alveolar infiltrate and a mononuclear interstitial infiltrate (Fig. 215-1A). This appearance is pathognomonic for PCP even though the organisms cannot be specifically identified with this stain. The diagnosis is typically established in lung tissue or pulmonary secretions by highly specific staining of the cyst—e.g., with methenamine silver (Fig. 215-1B), toluidine blue O, or Giemsa (Fig. 215-1C)—or by staining with a specific immunofluorescent antibody (Fig. 215-1D).

The demonstration of organisms in bronchoalveolar lavage (BAL) fluid is almost 100% sensitive and specific for PCP in patients with HIV infection and is almost as sensitive in patients with immunosuppression due to other processes. The organisms are identified with the specific stains indicated above for lung biopsy. While expectorated sputum or throat swabs have very low sensitivity, an induced sputum sample obtained and interpreted by an experienced provider can be highly sensitive and specific. The reported sensitivity of induced sputum for PCP is widely variable (55–90%), however, and is dependent on both the characteristics of the patient and the experience of the center conducting the test.

Many laboratories now offer polymerase chain reaction (PCR) testing of respiratory specimens for *Pneumocystis* in preference to direct microscopy of appropriately stained respiratory secretions. However, these PCR tests are so sensitive that it is difficult to distinguish patients with colonization (i.e., those whose acute lung disease is due to some other process but who have low levels of *Pneumocystis* DNA in the lungs) from those with acute pneumonia due to *Pneumocystis*. Such PCR tests on appropriate samples may be more useful for ruling out a diagnosis of PCP if they are negative than for definitively attributing the disease to *Pneumocystis*.

There has been considerable interest in serologic tests such as assays for (1→3)-β-D-glucan, levels of which are frequently elevated in patients with PCP. However, no serologic assays developed to date offer both substantial sensitivity and specificity.

COURSE AND PROGNOSIS

Untreated, PCP is invariably fatal. Patients with HIV infection often have an indolent course that presents as mild exercise intolerance or chest tightness without fever or cough and a normal or nearly normal posterior–anterior chest radiograph, with progression over days, weeks, or even a few months to fever, cough, diffuse alveolar infiltrates, and profound hypoxemia. Some patients with HIV infection and most patients with other types of immunosuppression have more acute disease that progresses over a few days to respiratory failure. Rare patients also develop distributive shock. A few unusual patients present with extrapulmonary manifestations in the skin or soft tissue, retina, brain, liver, kidney, or spleen. Extrapulmonary disease is nonspecific in presentation and can be diagnosed only by histology.

Factors that influence mortality risk include the patient's age and degree of immunosuppression as well as comorbidities, the presence of preexisting lung disease, a low serum albumin level, the need for mechanical ventilation, and the development of a pneumothorax. With advances in supportive critical care, the prognosis for patients with PCP who require intubation and respiratory support has improved and now depends to a large extent on comorbidities and the prognosis of the underlying disease. Since patients typically do not respond to therapy for 4–8 days, supportive care for a minimum of 10 days is a reasonable consideration if such support is compatible with the patient's wishes and the prognosis of comorbidities. Patients whose condition continues to deteriorate after 3 or 4 days or has not improved after 7–10 days should be reevaluated to determine whether other infectious processes are present (either having been missed on initial evaluation or having developed during treatment), whether initial anti-*Pneumocystis* treatment has failed, or whether noninfectious processes (e.g., congestive heart failure, pulmonary emboli, pulmonary hypertension, drug toxicity, or a neoplastic process) are causing pulmonary dysfunction.

TREATMENT

TREATMENT *P. jirovecii* Pneumonia

The treatment of choice for PCP is trimethoprim-sulfamethoxazole (TMP-SMX), given either IV or PO for 14 days to non-HIV-infected patients with mild disease and for 21 days to all other patients **(Table 215-1)**. TMP-SMX, which interferes with the organism's folate metabolism, is at least as effective as alternative agents and is better tolerated. TMP-SMX can cause leukopenia, hepatitis, rash, and fever as well as anaphylactic and anaphylactoid reactions, and patients with HIV infection have an unusually high incidence of hypersensitivity to TMP-SMX. Monitoring of serum drug levels is useful if renal function or toxicities are issues. Maintenance of a 2-h post-dose serum sulfamethoxazole level of 100–150 μg/mL has been associated with a successful outcome. Resistance to TMP-SMX cannot be measured by organism growth inhibition in the laboratory because *Pneumocystis* cannot be cultured. However, mutations in the target gene for sulfamethoxazole that confer in vitro sulfa resistance to other organisms have been found in *Pneumocystis*. The clinical relevance of these mutations for the response to therapy is unknown. Sulfadiazine plus pyrimethamine, an oral regimen more often used for treatment of toxoplasmosis, also is highly effective. Dapsone plus pyrimethamine or dapsone plus trimethoprim also can be used.

Intravenous pentamidine or the combination of clindamycin plus primaquine is an option for patients who cannot tolerate TMP-SMX and for patients in whose treatment TMP-SMX appears to be failing. Pentamidine must be given IV over at least 60 min to avoid potentially lethal hypotension. Adverse effects can be severe and irreversible and include renal dysfunction, dysglycemia (life-threatening hypoglycemia that can occur days or weeks after initial infusion and be followed by hyperglycemia), neutropenia, and torsades des pointes. Clindamycin plus primaquine is effective, but primaquine can be given only by the oral route—a disadvantage for patients who cannot ingest or absorb oral drugs. Oral atovaquone is also a reasonable option for patients with mild disease who have no impediments to absorbing an oral drug that requires a high-fat meal for optimal absorption.

A major advance in therapy for PCP was the recognition that glucocorticoids could improve survival rates among HIV-infected patients with moderate to severe disease (room air PO_2, <70 mmHg; or alveolar–arterial oxygen gradient, ≥35 mmHg). Glucocorticoids appear to reduce the pulmonary inflammation that occurs after specific therapy is started and organisms begin to die, eliciting inflammation. Therapy with glucocorticoids should be the standard of care for patients with HIV infection and probably is also effective for patients with other immunodeficiencies. This treatment should be started for moderate or severe disease when therapy for PCP is initiated, even if the diagnosis has not yet been confirmed. If HIV-infected or HIV-uninfected patients are receiving high-dose glucocorticoids when they develop PCP, there are theoretical advantages to either increasing the steroid dose (to reduce the inflammatory response to the dying organisms) or decreasing the steroid dose (to improve immune function), but there is no convincing evidence on which to base any specific strategy.

No definitive trials have defined the best therapeutic algorithm for patients in whom TMP-SMX treatment for PCP is failing. If no other treatable infectious or noninfectious processes are detected and pulmonary dysfunction appears to be due to PCP alone, many authorities would switch from TMP-SMX to either IV pentamidine or IV clindamycin plus oral primaquine. Some authorities would add the second drug or drug combination to TMP-SMX rather than switching regimens. If patients are not already receiving them, glucocorticoids should be added to the regimen; the dosage and regimen, which are usually chosen empirically, depend on what glucocorticoid regimen (if any) the patient was receiving when PCP therapy was begun.

For patients with HIV infection who present with PCP before the initiation of ART, ART should be started within the first 2 weeks of therapy for PCP in most situations. Immune reconstitution inflammatory syndrome (IRIS) can occur, however, and the decision to initiate ART thus requires considerable expertise in optimal timing relative to PCP recovery as well as in the other factors that are relevant when ART is initiated in any patient.

TABLE 215-1 Treatment of Pneumocystosis[a]

Drug(S)	Dose, Route	Adverse Effects
First-Choice Agent		
TMP-SMX	TMP (5 mg/kg) plus SMX (25 mg/kg) q6–8h PO or IV (i.e., 2 double-strength tablets tid or qid)	Fever, rash, cytopenias, hepatitis, hyperkalemia
Alternative Agents		
Atovaquone	750 mg bid PO	Rash, fever, hepatitis
Clindamycin	300–450 mg q6h PO or 600 mg q6–8h IV	
plus		Hemolysis (G6PD deficiency), methemoglobinemia, neutropenia, rash
Primaquine	15–30 mg qd PO	
Pentamidine	3–4 mg/kg qd IV	Hypotension, azotemia, cardiac arrhythmias (torsades des pointes), pancreatitis, dysglycemias, hypocalcemia, neutropenia, hepatitis
Adjunctive Agent		
Prednisone or methylprednisolone	40 mg bid × 5 d, 40 mg qd × 5 d, 20 mg qd × 11 d; PO or IV	Peptic ulcer disease, hyperglycemia, mood alteration, hypertension

[a]Treatment can be administered for 14 days to non-HIV-infected patients with mild disease and for 21 days to all other patients.

Abbreviations: G6PD, glucose-6-phosphate dehydrogenase; TMP-SMX, trimethoprim-sulfamethoxazole.

PREVENTION

The most effective method for preventing PCP is to eliminate the cause of immunosuppression by withdrawing immunosuppressive therapy or treating the underlying cause (e.g., HIV infection). Patients who are susceptible to PCP benefit from chemoprophylaxis during the period of susceptibility. For patients with HIV infection, CD4+ T cell counts are a reliable marker of susceptibility, and counts below 200/μL are an indication to start prophylaxis **(Table 215-2)**.

TABLE 215-2 Prophylaxis of Pneumocystosis

Drug(S)	Dose, Route	Comments
First-Choice Agent		
TMP-SMX	1 tablet (double- or single-strength) qd PO	Incidence of hypersensitivity is high. Rechallenge for non-life-threatening hypersensitivity; consider dose-escalation protocol.
Alternative Agents		
Dapsone	50 mg bid or 100 mg qd PO	Hemolysis is associated with G6PD deficiency.
Dapsone	50 mg qd PO	
plus		
Pyrimethamine	50 mg weekly PO	Leucovorin ameliorates cytopenias due to pyrimethamine.
plus		
Leucovorin	25 mg weekly PO	
Dapsone	200 mg weekly PO	
plus		
Pyrimethamine	75 mg weekly PO	Leucovorin ameliorates cytopenias due to pyrimethamine.
plus		
Leucovorin	25 mg weekly PO	
Pentamidine	300 mg monthly via Respirgard II nebulizer	Aerosol may cause bronchospasm. Pentamidine is probably less effective than TMP-SMX or dapsone regimens.
Atovaquone	1500 mg qd PO	Requires fatty meal for optimal absorption

Abbreviations: G6PD, glucose-6-phosphate dehydrogenase; TMP-SMX, trimethoprim-sulfamethoxazole.

For patients who are immunosuppressed as a result of factors other than HIV infection, there is no laboratory parameter, including the CD4+ T cell count, that predicts susceptibility to PCP with adequate positive and negative accuracy. The period of susceptibility is usually estimated on the basis of experience with the underlying disease and immunosuppressive regimen. Premature cessation of prophylaxis has been associated with clusters of cases in certain patient populations, such as solid-organ transplant recipients. Patients receiving a prolonged course of high-dose glucocorticoids appear to be particularly susceptible to PCP. The glucocorticoid exposure threshold that warrants chemoprophylaxis is controversial, but such preventive therapy should be strongly considered for any patient who is receiving more than the equivalent of 20 mg of prednisone daily for 30 days or who is receiving glucocorticoids in conjunction with other immunosuppressive agents. Clinical experience also suggests that chemoprophylaxis is useful for patients receiving certain immunosuppressive agents (e.g., tumor necrosis factor inhibitors, antithymocyte globulin, rituximab, and alemtuzumab). The duration of such chemoprophylaxis is empirically determined by clinical and laboratory indicators of immunologic vulnerability.

TMP-SMX is the most effective prophylactic drug; few patients experience a PCP breakthrough when they are reliably taking a recommended TMP-SMX chemoprophylactic regimen. Several TMP-SMX regimens have been used successfully. One double-strength tablet daily is the regimen with which there is the most experience, but either one single-strength tablet daily or one double-strength tablet two or three times weekly also has been recommended for various populations of patients.

For patients who cannot tolerate TMP-SMX (usually because of hypersensitivity or bone marrow suppression), alternative drugs include daily dapsone, weekly dapsone-pyrimethamine, atovaquone, and monthly aerosol pentamidine. Patients who develop hypersensitivity to TMP-SMX can sometimes tolerate the drug if a gradual dose-escalation protocol is used. Dapsone cross-reacts with sulfonamides in a substantial fraction of patients and therefore is rarely useful in patients with a history of life-threatening reactions to TMP-SMX. Aerosolized pentamidine is highly effective, but it is not as effective as TMP-SMX and may not provide protection in areas of the lung that are not well ventilated. Atovaquone is also effective and well tolerated; however, this drug is available only as an oral preparation, and gastrointestinal absorption is unpredictable in patients with abnormal gastrointestinal motility or function.

FURTHER READING

Akgun KM, Miller RF: Critical care in human immunodeficiency virus–infected patients. Semin Respir Crit Care Med 37:303, 2016. [PubMed: 26974306]
Bozzette SA et al: A controlled trial of early adjunctive treatment with corticosteroids for *Pneumocystis carinii* pneumonia in the acquired immunodeficiency syndrome. California Collaborative Treatment Group. N Engl J Med 323:1451, 1990. [PubMed: 2233917]
Buchacz K et al: Incidence of AIDS-defining opportunistic infections in a multicohort analysis of HIV-infected persons in the United States and Canada, 2000–2010. J Infect Dis 214:862, 2016. [PubMed: 27559122]
Drummond MB et al: Factors associated with abnormal spirometry among HIV-infected individuals. AIDS 29:1691, 2015. [PubMed: 26372280]
Kaur N, Mahl TC: *Pneumocystis jiroveci* (*carinii*) pneumonia after infliximab therapy: A review of 84 cases. Dig Dis Sci 52:1481, 2007. [PubMed: 17429728]
Ma L et al: Genome analysis of three *Pneumocystis* species reveals adaptation mechanisms to life exclusively in mammalian hosts. Nat Commun 7:10740, 2016. [PubMed: 26899007]
Panel on Opportunistic Infections in HIV-Infected Adults and Adolescents: Guidelines for the prevention and treatment of opportunistic infections in HIV-infected adults and adolescents: Recommendations from the Centers for Disease Control and Prevention, the National Institutes of Health, and the HIV Medicine Association of the Infectious Diseases Society of America. Available at *http://aidsinfo.nih.gov.ezproxy1.library.usyd.edu.au/contentfiles/lvguidelines/adult_oi.pdf*. Accessed January 5, 2018.
Zolopa A et al: Early antiretroviral therapy reduces AIDS progression/death in individuals with acute opportunistic infections: A multicenter randomized strategy trial. PLoS One 4:e5575, 2009. [PubMed: 19440326]

Chapter 216: Introduction to Parasitic Infections

Sharon L. Reed; Charles E. Davis

INTRODUCTION

The word *parasite* comes originally from the Greek *parasitos* (*para*, alongside of; and *sitos*, food), meaning someone who eats at another's table or lives at another's expense. Although the same is true of many bacteria and viruses, the designation *parasite* is reserved, by convention, for helminths and protozoa. These organisms are larger and more complex than bacteria, with a eukaryotic cell structure similar to that of human host cells. Historically, this similarity has made it difficult to find effective antiparasitic agents that do not cause unacceptable toxicity to human cells. Fortunately, intensive research and modern techniques have now provided suitable agents for safe and effective treatment of most parasitic infections. **See Chap. S14 for details on diagnostic procedures and Chap. 217 for details on treatment.**

Internal parasites of human beings are divided into two types: helminths (worms) and protozoa. *Helminths* are multicellular organisms that can be seen with the naked eye **(Chap. 225)**. There are two phyla: Platyhelminthes (flat worms) and Nemathelminthes (roundworms). Both phyla include some genera that mature in the gastrointestinal tract and others that migrate through the tissue after ingestion or skin penetration. **Tables S14-1 and S14-2** present the helminthic genera, their definitive and intermediate hosts, geographic distributions, and the parasitic stages in the human body.

The key to understanding which helminths use humans as definitive hosts is to remember that helminth ova develop into larvae, and larval stages develop into adults. Humans serve as the definitive host when they ingest helminth larvae, which develop into adults in the intestine and usually cause mild disease, often without any symptoms. (The exception is ingestion of the late-stage larvae of the somatic or tissue flukes, as shown in **Table S14-2**.) In contrast, if humans ingest helminth ova and serve as the intermediate host, the ova develop into larvae, which penetrate the intestine, migrate through the tissue, and invade organs where they mature into adults. Intermediate hosts with parasitic invasion of organs may experience severe disease.

Protozoa are microscopic single-celled organisms. Among the many differences between helminths and protozoans, the most important is the ability of protozoa (like bacteria) to multiply within the human body and cause overwhelming infections. A major mechanism promoting unrestrained growth is evasion of the host immune response either by antigenic variation (*Trypanosoma brucei*) or by survival inside host cells (e.g., *Plasmodium*, *Babesia*, *Cryptosporidium*, *Leishmania*, and *Toxoplasma*). In contrast, almost all helminths require stages in other hosts to complete their life cycles and multiply. As a result, except for *Strongyloides* and *Capillaria*, which can complete their life cycle in humans, increases in the burden of infection with helminths require repeated exogenous reinfections. Thus permanent residents of endemic countries, who are exposed repeatedly, may have heavy severe infections, while most travelers with one or two exposures are unlikely to experience the full spectrum of chronic helminthic infections.

In contrast to helminthic infections, naïve patients with their first protozoal infection usually are the most severely affected because partial immunity often limits the number of parasites during recurrent infections. Protozoan replication to large numbers in the host also promotes the development of drug-resistant forms, especially in malaria **(Chap. 217)**. Because protozoa belong to many different phyla, it is easier to understand the pathogenesis and management of protozoal infections when they are classified by the site of infection (intestinal protozoans, free-living amebae, and blood and tissue protozoans) **(Table S14-3)**. Immunocompromised hosts are at risk of disseminated infection with a number of protozoa, including *Leishmania*, *Toxoplasma*, *Cryptosporidium*, *and Trypanosoma cruzi*, which are AIDS-defining illnesses. In contrast, *Strongyloides* is the only helminth to disseminate.

HELMINTHIC INFECTIONS

The Platyhelminthes (flatworms) are categorized as tapeworms (cestodes) and flukes (trematodes). Tapeworms are composed of a head or scolex bearing the holdfast organs and segments, which become gravid as they mature. Some tapeworms can reach lengths of many yards; the longest tapeworms develop in the intestine, where they rarely cause serious disease. In contrast, flukes are small leaf-shaped organisms whose size is not a measure of disease severity.

FLATWORMS

Cestodes

Tapeworms cause either intestinal or somatic infection, depending on the species. *Intestinal* infections occur when the human host ingests larvae in the tissue of the intermediate host, whereas *somatic* infections occur when humans accidentally ingest ova excreted from the wild or domesticated definitive animal host.

INTESTINAL TAPEWORMS

As shown in **Table S14-1**, humans acquire most intestinal tapeworms by eating the insufficiently cooked flesh of the intermediate host. Thus *Taenia saginata* is commonly called the beef tapeworm, *Taenia solium* the pork tapeworm, and *Diphyllobothrium latum* the fish tapeworm. *Hymenolepis nana* is capable of completing its life cycle in the human intestine and is acquired by ingestion of infected grain beetles or of ova from infected humans or mice. None of these parasites causes significant damage, and infection is usually asymptomatic. There are two occasional exceptions. When people ingest *T. solium* ova from their own intestine or from another infected individual, it can cause somatic infection. *D. latum* avidly absorbs vitamin B_{12} in the intestine and can cause pernicious anemia in 1–2% of infected Scandinavians with a genetic predisposition.

SOMATIC TAPEWORMS

There are three major causes of somatic tapeworm infections. Two species of *Echinococcus* cause echinococcosis. *E. granulosus* is acquired by accidental ingestion of ova from dogs infected when fed the infected tissues of sheep or other animals by sheepherders or hunters. *E. multilocularis* is transmitted primarily in sub-Arctic areas when humans ingest ova from foxes, dogs, or cats that have been infected through consumption of the tissues of infected rodents. Both species cause hydatid cysts when the eggs hatch into larvae, penetrate the intestine, and migrate into the liver or lung. Ingested *T. solium* ova cause somatic disease (cysticercosis) when the larvae penetrate the intestine, migrate into tissue, and form cysts (*cysterci*), usually in the muscles or central nervous system (CNS).

Trematodes

Flukes also cause both intestinal and somatic infections **(Chap. 229 and Table S14-1)**. Most fluke infections are localized to Asia, Africa, Southeast Asia, or the Pacific islands. Infection with intestinal flukes is usually asymptomatic, although heavy infections sometimes cause abdominal discomfort and mucous diarrhea. Liver flukes and lung flukes cause somatic infections when humans ingest a larval form from an intermediate host. Adults develop in the intestine, migrate into adjacent tissues, and cause disease. The major liver flukes (*Clonorchis sinensis*, *Opisthorchis* spp., and *Fasciola hepatica*) are causes of recurrent bacterial cholangitis (due to obstruction) or portal hypertension and cirrhosis. Only *F. hepatica* can be acquired worldwide; it is especially common in sheep-raising areas, where the animals ingest water plants (e.g., watercress). The lung flukes (*Paragonimus* spp.) occur globally except in North America and Europe; most lesions occur as pulmonary cysts, although occasional lesions develop in the CNS or the abdominal cavity.

The blood flukes cause schistosomiasis, one of the most common and serious parasitic infections **(Chap. 229 and Table S14-1)**. The major species are *Schistosoma mansoni*, S. *haematobium*, and *S. japonicum*. All are transmitted to humans when free-swimming larvae exit an infected snail in freshwater and penetrate the skin. Swimmer's itch sometimes follows skin penetration but is usually of short duration. The larvae then wander in the skin until they find a blood vessel and migrate to the target organ. *S. mansoni* and *S. japonicum* migrate to the mesentery vessels and eventually make their way to the liver, while *S. haematobium* targets the veins around the ureter and bladder. Extensive egg deposition by *S. mansoni* and *S. japonicum* and the immune reactions to the ova cause granuloma formation and, with many repeated exposures, portal vein obstruction and cirrhosis. The same process in the ureters and bladders during infection with *S. haematobium* eventually interferes with urine flow and leads to repeated urinary tract infections and kidney damage.

ROUNDWORMS

Nematodes

Roundworms are nonsegmented bisexual organisms. The species that infect humans include intestinal and tissue groups. Humans may also acquire certain nonhuman mammalian roundworms that either can be limited to the skin or can migrate to tissues and cause serious disease (the larva migrans syndromes).

INTESTINAL ROUNDWORMS

The major intestinal roundworms are *Ascaris lumbricoides*, *Necator americanus* (New World hookworms), *Ancylostoma duodenale* (Old World hookworms), *Trichuris trichiura* (whipworms), *Enterobius vermicularis* (pinworms), and *Strongyloides stercoralis*. Taken together, infections caused by intestinal roundworms are the most common infections in the world. *Ascaris*, hookworms, and *Trichuris* each infect about 1 billion individuals, and at least 30–100 million have strongyloidiasis. These infections are most common in resource-poor developing countries, especially where people defecate outside and/or human feces is used as fertilizer ("night soil"). Infection is transmitted either by ingestion of ova (*A. lumbricoides*, *T. trichiura*, and *E. vermicularis*) or by active penetration of the skin by larvae (hookworms and *S. stercoralis* **(Table S14-2)**.

Intestinal roundworms cause serious health problems in residents of endemic regions with poor sanitation, but travelers are at low risk of developing significant disease from most of these parasites. Intestinal blockage and malnutrition from heavy *Ascaris* infections and anemia from heavy hookworm infections are now restricted to areas of heavy endemicity. Except in the case of *Strongyloides* and *Capillaria*, which can reproduce in the body, multiple exposures over time are necessary for the development of severe disease. S*trongyloides* infection persists over decades and can disseminate when the immune system is compromised. Although *Capillaria* remains localized to the intestine,

infections can become so heavy that protein-losing enteropathy and malnutrition cause serious disease.

The life cycles of *Ascaris* and the hookworms involve migration through the heart and lungs before development into adults in the intestine. In particular, *Ascaris* occasionally causes eosinophilic pneumonia (Loeffler's syndrome) during heavy infections. Pinworms are the most common causes of intestinal roundworm infection persisting in the United States and other developed countries. The anal and perineal itching caused by pinworm migration out of the anus and subsequent egg deposition is well known to families throughout the world.

TISSUE ROUNDWORMS

The major diseases caused by tissue roundworms are filariasis, angiostrongyliasis, gnathostomiasis, and trichinellosis. By far the most important globally is filariasis; the thread-like filarial worms infect more than 150 million individuals, and almost 1 billion are at risk in sub-Saharan Africa and other poor tropical countries. Four filarial species cause three distinct diseases: lymphatic filariasis (*Wuchereria bancrofti* and *Brugia malayi*), river blindness (*Onchocercus volvulus*), and loiasis (*Loa loa*, the African eye worm). Humans, the major reservoir, acquire these infections from bites of infected arthropods **(Table S14-2)**. The larvae develop into adults, which remain static in tissue: the lymphatics for lymphatic filariasis and subcutaneous tissue for *O. volvulus* and *L. loa*. After adults mate, next-stage larvae are produced, and their migration causes additional damage.

Repeated bouts of migrating larvae and blocking of the lymphatics by adults are necessary to establish the syndrome of lymphatic filariasis; thus it is unusual for the short-term traveler (<3 months' residence in an endemic region) to develop significant disease. In river blindness, the larvae produced by adult *O. volvulus* migrate through the skin and eye, causing skin damage and eventual blindness. Loiasis is a milder disease restricted to central and western Africa. Although both the adults and the larvae of *L. loa* migrate through the skin and eye, many infected individuals are asymptomatic, and the infection is often diagnosed only when an adult worm migrates across the subconjunctival tissue and is visible to the patient and the physician. Red lumps in the skin from heavy cutaneous migration are called *Calabar swellings*.

The other four major roundworm tissue infections are acquired by ingestion of larvae in undercooked food. The sources for trichinellosis are swine and other large mammals; for gnathostomiasis, freshwater fish and chicken; for ancylostomiasis, snails, fish, prawns, and crabs; and for Guinea worm, infected water fleas. Guinea worm infection (dracunculiasis, caused by *Dracunculus medinensis*) has been almost eradicated. *Trichinella spiralis* larvae penetrate the intestine and migrate widely, with a preference for skeletal tissue; the release of eosinophils and IgE causes muscle soreness and may cause palpebral swelling and other manifestations of generalized allergic reactions. *Angiostrongylus cantonensis* is the most common parasitic cause of eosinophilic meningitis. Ingested larvae penetrate the intestine and migrate to the brain and meninges, where they quickly die and attract massive numbers of eosinophils. Although complications can occur, most individuals recover spontaneously. *Gnathostoma spinigerum* larvae also penetrate the intestine and migrate, showing a preference for the skin, eyes, and meninges. Mechanical damage from the migration and inflammation produced by the resultant immune reaction can cause boil-like lesions on the skin, painful eye damage, and eosinophilic meningitis. Although eosinophilic meningitis caused by *G. spinigerum* is less common than that caused by *A. cantonensis*, it is often more severe and can result in paralysis or brain hemorrhage.

PROTOZOAL INFECTIONS

INTESTINAL PROTOZOA

Entamoeba histolytica is the one intestinal protozoan that causes invasive disease. This disease consists of dysentery or bloody diarrhea that must be differentiated from that due to bacteria such as *Salmonella*, *Campylobacter*, and *Shigella.* Although amebiasis usually has a slower onset with lower fever than these bacterial infections, *E. histolytica* can disseminate from the bloodstream to cause distant abscesses, particularly of the liver. The diagnosis cannot be made by identification of the characteristic cyst or trophozoites **(Chap. 218)** as they are identical to those of the noninvasive *E. dispar*, which is more common globally.

Cryptosporidium and *Giardia* are the most common water-borne protozoal infections. *Cryptosporidium* can cause major outbreaks because it is highly infectious and resistant to high levels of chlorine **(Chap. 224)**. Without immune reconstitution, immunosuppressed patients, particularly those with AIDS, can develop severe, even fatal watery diarrhea. Infections caused by the remaining intestinal protozoans—*Giardia*, *Isospora*, *Cyclospora*, and microsporidia **(Chap. 224)**—have a much more indolent course, with intermittent diarrhea. Microsporidia, unique intracellular protozoa that form infectious spores, may cause limited gastrointestinal infection in immunocompetent hosts, but patients with AIDS can develop chronic diarrhea and wasting or disseminated infection to the biliary or respiratory tract.

FREE-LIVING AMEBAS

The free-living amebas *Acanthamoeba* and *Naegleria* are found worldwide in freshwater and brackish water **(Chap. 218 and Table S14-3)**. Organisms of these two genera cause very different syndromes. In immunocompromised individuals, *Acanthamoeba* usually causes invasive infection, with brain masses and skin lesions. However, all humans are susceptible to *Acanthamoeba* keratitis after trauma to the eye and exposure to contaminated water. In contrast, naeglerial meningitis, acquired in warm lakes or hot springs, causes sudden pyogenic and usually fatal meningitis. *Balamuthia*, reported only from the Americas, causes indolent meningoencephalitis, with both cerebrospinal fluid pleocytosis and a space-occupying lesion, in immunocompetent patients. Despite the availability of miltefosine, which is active in vitro against *Naegleria*, infection of the CNS is almost universally fatal.

BLOOD AND TISSUE PROTOZOANS

Plasmodium and *Babesia*

Malaria, caused by six species of *Plasmodium,* carries higher mortality rates than any other parasitic infection **(Chap. 219)**. All species are transmitted in tropical and subtropical areas by female *Anopheles* mosquitoes. *Plasmodium falciparum* is most common in sub-Saharan Africa, where it causes more than 80% of malaria infections and 90% of malarial deaths. Infection with *P. falciparum* may be particularly severe because the organism can invade any erythrocyte, reaches very high parasite loads, damages organs by adhering to vascular epithelium, and is the most likely *Plasmodium* species to be resistant to antimalarial drugs. *Plasmodium vivax,* the dominant cause of malaria outside sub-Saharan Africa, reaches lower levels of parasitemia and exhibits less drug resistance because it invades only reticulocytes with Duffy antigen. Many Africans, especially in the western part of the continent, lack the Duffy blood group; consequently, *Plasmodium ovale*, another cause of milder malaria, can compete successfully with *P. vivax*. Both *P. vivax* and *P. ovale* produce persistent liver forms, which must be treated with primaquine **(Chap. 217)**. Because malaria can cause a variety of symptoms ranging from acute fever to coma, this diagnosis must be considered in any traveler or immigrant from a malarial area. *Babesia* also infects erythrocytes and may cause a nonspecific febrile illness or, in asplenic patients, severe infection. This parasite is carried by ixodid ticks and is geographically limited to the northeastern and midwestern United States, with only sporadic cases in Europe and other temperate areas.

Trypanosomes

The three species of trypanosomes all have flagellated bloodstream forms, but they cause very different diseases. *T. cruzi*, the cause of Chagas disease, is transmitted in South and Central America in the feces of blood-sucking reduviid bugs **(Chap. 222)**. After initial parasitemia, patients are often asymptomatic for years while the parasite multiplies intracellularly in muscle and ganglion cells. Although only a minority of patients go on to develop organ damage (megaesophagus and cardiomyopathy), all infected patients can spread the disease through transfusions, mother-to-child transmission, and organ transplants.

African trypanosomiasis is limited to sub-Saharan Africa, where it is transmitted by the bite of a tsetse fly. A history of a tsetse bite and the presence of a painful chancre are strong diagnostic clues **(Chap. 222)**. Although the parasites causing this disease in western Africa (*Trypanosoma brucei gambiense*) and eastern Africa (*T. brucei rhodesiense*) look identical, they are genetically and clinically distinct. *T. b. gambiense* causes low-level parasitemia with cyclical fevers over months or years before CNS invasion, while *T. b. rhodesiense* causes high-level parasitemia, invades the CNS early on, and can lead to death within weeks of onset.

Leishmania

Leishmaniasis is caused by more than 20 species of obligate intracellular protozoa transmitted by sandflies, which are present in almost 100 countries in tropical and temperate zones **(Chap. 221)**. A wide spectrum of clinical symptoms result, ranging from self-healing, painless skin ulcers to mucocutaneous disease with destruction of the nose and palate to disseminated visceral leishmaniasis with hepatic and splenic involvement. The resulting disease depends on the infecting strain and the host immune response. Visceral leishmaniasis can present as an acute febrile illness, with the later development of hepatosplenomegaly, and is an AIDS-defining illness in HIV-infected patients. More than 90% of cases of visceral leishmaniasis occur in India, Bangladesh, Ethiopia, Sudan, and Brazil.

Toxoplasma

Toxoplasma gondii is an obligate intracellular parasite that is found worldwide. Infection follows ingestion of oocysts in food or water contaminated by cat feces, ingestion of tissue cysts in undercooked meat, or transplacental transmission. After gastrointestinal invasion, tachyzoites can invade any nucleated cell and cause lifelong infection in most patients **(Chap. 223)**. Clinical manifestations depend on the host's age and immune status at the time of infection. Congenital toxoplasmosis results from primary maternal infection; outcomes are most severe early in pregnancy and include visual, hearing, and cognitive impairments. Babies infected later in pregnancy may appear normal but can develop chorioretinitis decades later. Primary infection in immunocompetent hosts may be asymptomatic, may present as an infectious mononucleosis–like syndrome, or may manifest as chorioretinitis during outbreaks. During immunosuppression by AIDS or organ transplantation, reactivation of latent cerebral infection can be fatal unless diagnosed and treated early.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Parasitic Infection

A thorough history and physical examination are the keys to diagnosis of any disease and particularly of parasitic infections. Because many of the more serious parasitic infections are uncommon in the United States, a travel history, particularly to developing nations, is a critical component. The longer the stay in an area endemic for significant parasitic infections, the greater the risk, even for healthy travelers. In addition, other factors increase the chance of acquiring these infections. Notably, immunocompromise greatly increases the likelihood of developing some of the more serious parasitic infections. Even healthy travelers with adventure itineraries, extensive travel to rural areas, or

involvement in war zones or refugee camps are at increased risk. Immigrants from developing countries may seek care for symptoms or signs associated with parasitic infections.

Information on the patient's immunization history and adherence to appropriate malarial chemoprophylaxis is critical. Although no vaccines against parasitic infections are commercially available, the likelihood of many viral and bacterial infections is much lower if the patient has been properly immunized. For example, typhoid fever is much less likely to be the cause of prolonged fever in an immunized individual. Similarly, hepatitis A or B is unlikely to be the cause of jaundice and fever in fully immunized patients. In this era of increasing drug resistance, even adherence to appropriate malarial chemoprophylaxis does not guarantee that fever is not malarial. Nevertheless, most travelers who acquire malaria have taken inadequate or no prophylaxis. Although these considerations do not prove that the symptoms are caused by parasites, they narrow the differential diagnosis.

There are many other important aspects of the history, including when symptoms began. Was the individual still in the endemic area at the time, or did the symptoms commence after return to the United States? If they started during travel, was any treatment received? If the patient was well upon return from travel, the timing of symptom onset is a critical point. For example, if the chief manifestation is fever that began >10–14 days after departure from the endemic region, many tropical diseases can be ruled out, including dengue fever, chikungunya fever, and Zika virus infection. On the other hand, fever beginning several months or later after return makes malaria a likely diagnosis. Travelers' diarrhea, the most common complaint of travelers, is usually caused by bacteria or viruses and resolves in a short time with or without treatment. Travelers' diarrhea that persists for weeks is much more likely to be parasitic in origin.

Most patients who consult physicians after international travel either have troublesome symptoms or have been referred for symptoms or signs whose source was unclear to a referring caregiver. After a careful travel history including the individual's symptoms and the exact geographic zones visited, a thorough physical examination must be conducted. The symptoms, signs, and physical findings should help to establish possible diagnoses. **Table 216-1** breaks down the symptoms of major parasitic infections by organ system and geographic distribution, with comments on clinical and epidemiologic associations.

Table 216-1 Parasitic Infections, by Organ System and Signs/Symptoms[a]

Organ System, Major Sign(s)/Symptom(s)	Parasite(s)	Geographic Distribution	Comments
Skin			
	Hookworm	Worldwide	Can cause anemia in heavy infections
Serpentine rash	*Strongyloides*	Moist tropics and subtropics	Disseminated infection in immunocompromise
	Toxocara (animal roundworm)	Tropical and temperate zones	Cutaneous or visceral larva migrans
Itchy skin rash	*Onchocerca*	Mexico, Central/South America, Africa	Larvae detectable in skin snips and nodules
Painless ulcers	*Leishmania*	Tropics and subtropics	Amastigotes detectable in biopsies; may cause destructive mucocutaneous infection; AIDS-defining infection
	Onchocerca	Mexico, South America, Africa	Large nodules of adult worms
Skin nodules	*Loa loa* (African eye worm)	Western and central Africa	Migratory nodules
	Gnathostoma	Southeast Asia and China	Migratory nodules with eosinophilia
Painful nodules, especially involving feet	*Dracunculus* (Guinea worm)	Africa	Nearly eradicated
Central Nervous System			
Somnolence, seizures, coma	*Plasmodium falciparum*	Subtropics and tropics	Cerebral malaria, especially in children
	Trypanosoma brucei rhodesiense	Sub-Saharan eastern Africa	Painful chancre from tsetse fly bite; death in weeks to months
Space-occupying lesions, seizures	*Acanthamoeba*	Worldwide	Immunocompromised individuals
	Balamuthia	Americas	Indolent meningoencephalitis with brain mass
	Toxoplasma	Worldwide	Reactivation disease in immunocompromise; ring-enhancing lesions; AIDS-defining infection
	Taenia solium	Mexico, Central/South America, Africa	Cysticercosis; variable sized or calcified larval cysts on CT
	Schistosoma japonicum	Far East	Aberrant eggs can form brain or spinal cord masses.

Organ System, Major Sign(s)/Symptom(s)	Parasite(s)	Geographic Distribution	Comments
	Schistosoma mansoni	Africa, Central/South America	Aberrant eggs can form brain or spinal cord masses.
Pyogenic meningitis	*Naegleria*	Worldwide	Motile trophozoites in fresh cerebrospinal fluid; rapid death
Eosinophilic meningitis	*Angiostrongylus* (rat lung worm)	Southeast Asia, Pacific, Caribbean	Most common cause globally; spontaneous resolution
	Gnathostoma	Southeast Asia and China	Migratory nodules
Eyes			
Painful corneal ulcers	*Acanthamoeba*	Worldwide	Freshwater and brackish water; corneal trauma; long-wear contact lenses
Corneal opacification	*Onchocerca*	Mexico, Central/South America, Africa	Immune response to microfilaria in cornea
Congenital or adult visual loss	*Toxoplasma*	Worldwide	Primary infection in pregnancy and subsequent primary or reactivation infection
Retinal mass	*Toxocara*	Worldwide	Ocular larva migrans
Visible roundworm in eye	*Onchocerca*	Mexico, Central/South America, Africa	Worms may cross eye during migration.
	L. loa	Western and central Africa	Worms may cross eye during migration.
Pain, possible vision loss	*Gnathostoma*	Southeast Asia and China	Migratory skin nodules, eosinophilia
Lungs			
Pulmonary nodule/abscess	*Paragonimus*	Far East , Africa, Americas	Ectopic migration to abdomen or central nervous system
Cough, transient infiltrates, eosinophilia	Migrating helminths	Worldwide	Loeffler’s syndrome from migrating *Ascaris*, hookworm, *Strongyloides*
Heart			
Pulmonary edema	*P. falciparum* (complication)	Tropics and subtropics	End-organ damage from severe malaria
Cardiomegaly, arrhythmias	*Trypanosoma cruzi*	Mexico, Central/South America	Late amastigote infection of myocardium; AIDS-defining infection
Gastrointestinal Tract			

Organ System, Major Sign(s)/Symptom(s)	Parasite(s)	Geographic Distribution	Comments
	Malaria (multiple episodes)	Tropics and subtropics	Splenomegaly with anemia and recurrent fever are hallmarks of malaria.
Hepatosplenomegaly	*S. mansoni*	Africa, Central/South America	Portal obstruction with cirrhosis and late varices
	Leishmania donovani complex	Tropics and subtropics	Visceral leishmaniasis; AIDS-defining infection
	Entamoeba histolytica	Tropics	Acute with fever, right-upper-quadrant pain; or chronic with enlarged liver; hypoechoic abscess(es) on ultrasound or CT
Hepatomegaly	*Echinococcus*	Sheep-raising areas	Characteristic cysts of liver > lung
	Fasciola	Sheep-raising areas	Eosinophilia
	Clonorchis	China, Southeast Asia	Recurrent cholangitis and late cholangiocarcinoma
Cholangitis	Microsporidia	Worldwide	AIDS
	Cryptosporidium	Worldwide	AIDS-defining infection
	E. histolytica	Tropics	Less fever than in diarrhea of bacterial etiology
Bloody diarrhea	*S. mansoni*	Africa, Central/South America	Only in heavy, acute infection with fever and eosinophilia
	S. japonicum	Far East	Only in heavy, acute infection
	Cryptosporidium	Worldwide	Severe in immunocompromised patients
	Giardia	Worldwide	Foul-smelling stool with steatorrhea
Watery diarrhea	*Isospora belli*	Worldwide	Fever, abdominal pain, chronic diarrhea
	Microsporidia	Worldwide	Chronic diarrhea with AIDS
	Capillaria	Southeast Asia, Egypt	Malabsorption, wasting
Passage of large roundworm (>6 cm)	*Ascaris*	Worldwide	Patients may confuse the roundworm with an earthworm.
Small roundworms visible around anus	Pinworm	Worldwide	Anal itching; eggs rarely detected by ova and parasite (O&P) exam
	Trichuris	Worldwide	Rectal prolapse with heavy infection in children
Passage of tapeworm segments	*T. solium* or *Taenia saginata*	Worldwide	Usual reason for seeking medical care
	Diphyllobothrium latum	Worldwide	Pernicious anemia in genetically predisposed Scandinavians
Genitourinary System			
Itchy discharge	*Trichomonas vaginalis*	Worldwide	Common sexually transmitted disease of both sexes

Organ System, Major Sign(s)/Symptom(s)	Parasite(s)	Geographic Distribution	Comments
Hematuria	*Schistosoma haematobium*	Africa	Hematuria with negative cultures, urinary tract infections, and late bladder cancer
Muscular System			
Myalgias, myositis	*Trichinella*	Worldwide	Palpebral swelling; high-level eosinophilia
Bloodstream			
	Plasmodium	Tropics and subtropics	Consider in any patient from a malarious area.
	Babesia	New England, United States	Geographically limited; worse with splenectomy
Fever without localizing symptoms	*T. brucei rhodesiense, T. brucei gambiense*	Sub-Saharan Africa	Limited to tsetse fly range; painful chancre; adenopathy and cyclical fevers; early (*rhodesiense*) or late (*gambiense*) central nervous system involvement
	Filariae	Asia, India	Periodic fever with eosinophilia, adenolymphangitis, chronic lymphangitis
	L. donovani complex	Tropics and subtropics	Hepatosplenomegaly, fever, wasting; AIDS-defining infection

[a]See also text and **Tables S14-1, S14-2, and S14-3** for vectors and routes of transmission.

FURTHER READING

Bern C: Chagas' disease. N Engl J Med 373:456, 2015. [PubMed: 26222561]
Cox FE: History of human parasitic diseases. Infect Dis Clin North Am 18:171, 2004. [PubMed: 15145374]
Hahn WO, Pottinger PS: Malaria in the traveler. Med Clin North Am 100:289, 2016. [PubMed: 26900114]
Leder K et al: GeoSentinel surveillance of illness in returned travelers, 2007–2011. Ann Intern Med 158:456, 2013. [PubMed: 23552375]
Ross AG et al: Enteropathogens and chronic illness in returning travelers. N Engl J Med 368:1817, 2013. [PubMed: 23656647]

Chapter 217: Agents Used to Treat Parasitic Infections

Thomas A. Moore

INTRODUCTION

Parasitic infections continue to afflict more than half of the world's population and impose a substantial health burden, particularly in underdeveloped nations, where they are most prevalent. The reach of some parasitic diseases, including malaria, has expanded over the past few decades as a result of factors such as deforestation, population shifts, global warming, and other climatic events. Although there have been significant advances in vaccine development and vector control, chemotherapy remains the single most effective means of controlling parasitic infections. Efforts to combat the spread of some diseases are hindered by the development and spread of drug resistance, the limited introduction of new antiparasitic agents, the proliferation of counterfeit medications, and, most recently, profiteering, which has dramatically increased the cost of once-affordable agents. However, there are good reasons to be optimistic. Ambitious global initiatives aimed at controlling or eliminating threats such as AIDS, tuberculosis, and malaria have demonstrated successes. The ongoing efforts of multinational partnerships to address the substantial burden imposed by neglected tropical diseases have generated mechanisms to develop and deploy effective antiparasitic agents. In addition, the development of vaccines against several tropical diseases, including malaria, continues.

This chapter deals exclusively with the agents used to treat infections due to parasites. Specific treatment recommendations for the parasitic diseases of humans are listed in subsequent chapters. Many of the agents discussed herein are approved by the U.S. Food and Drug Administration (FDA), but are considered investigational for the treatment of certain infections. Drugs marked in the text with an asterisk (*) are available through the Centers for Disease Control and Prevention (CDC) Drug Service (telephone: 404-639-3670; email: drugservice@cdc.gov; *www.cdc.gov/ncpdcid/dsr/*). Drugs marked with a dagger (†) are available only through their manufacturers; contact information for these manufacturers may be available from the CDC.

Table 217-1 presents a brief overview of each agent (including some drugs that are covered in other chapters), along with major toxicities, spectrum of activity, and safety for use during pregnancy and lactation.

TABLE 217-1 Overview of Agents Used for the Treatment of Parasitic Infections

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
4-Aminoquinolines					
Amodiaquine	Malaria[b]	Agranulocytosis, hepatotoxicity	No information	Not assigned	Yes[c]
Chloroquine	Malaria[b]	*Occasional:* pruritus, nausea, vomiting, headache, hair depigmentation, exfoliative dermatitis, reversible corneal opacity *Rare:* irreversible retinal injury, nail discoloration, blood dyscrasias	Antacids and kaolin: reduced absorption of chloroquine Ampicillin: bioavailability reduced by chloroquine Cimetidine: increased serum levels of chloroquine Cyclosporine: serum levels increased by chloroquine	Not assigned[d]	Yes[c]
Piperaquine	Malaria[b]	*Occasional:* GI disturbances	None reported	Not assigned	Yes
8-Aminoquinolines					
Primaquine	Malaria[b]	*Frequent:* hemolysis in patients with G6PD deficiency *Occasional:* methemoglobinemia, GI disturbances *Rare:* CNS symptoms	Quinacrine: potentiated toxicity of primaquine	Contraindicated	Yes
Tafenoquine	Malaria[b]	*Frequent:* hemolysis in patients with G6PD deficiency, mild GI upset *Occasional:* methemoglobinemia, headaches	No information	Not assigned	Yes
Aminoalcohols					

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
		Frequent: abdominal pain, diarrhea *Occasional:* ECG disturbances (dose-related prolongation of QTc and PR interval), nausea, pruritus; contraindicated in persons who have cardiac disease or who have taken mefloquine in the preceding 3 weeks	Concomitant use of agents that prolong QTc interval contraindicated	C	No information
Lumefantrine	Malaria[b]	*Occasional:* nausea, vomiting, diarrhea, abdominal pain, anorexia, headache, dizziness	Plasma levels increased by darunavir and nevirapine, decreased by etravirine	Not assigned	No information
Aminoglycosides					
Paromomycin	Amebiasis,[b] infection with *Dientamoeba fragilis*, giardiasis, cryptosporidiosis, leishmaniasis	*Frequent:* GI disturbances (oral dosing only) *Occasional:* nephrotoxicity, ototoxicity, vestibular toxicity (parenteral dosing only)	No major interactions	Oral: B Parenteral: not assigned[d]	No information
Amphotericin B					
Amphotericin B deoxycholate Amphotec (InterMune) Amphotericin B lipid complex, ABLC (Abelcet) Amphotericin B, liposomal (AmBisome)	Leishmaniasis,[e] amebic meningoencephalitis	*Frequent:* fever, chills, hypokalemia, hypomagnesemia, nephrotoxicity *Occasional:* vomiting, dyspnea, hypotension	Antineoplastic agents: renal toxicity, bronchospasm, hypotension Glucocorticoids, ACTH, digitalis: hypokalemia Zidovudine: increased myelo- and nephrotoxicity	B	No information
Antimonials	Leishmaniasis	*Frequent:* arthralgias/myalgias, pancreatitis, ECG changes (QT prolongation, T wave flattening or inversion)			
Pentavalent antimony[f]			No major interactions	Not assigned	Yes
Meglumine antimoniate			Antiarrhythmics and tricyclic antidepressants: increased risk of cardiotoxicity	Not assigned	No information

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
Artemisinin and derivatives	Malaria[g]	*Occasional:* neurotoxicity (ataxia, convulsions), nausea, vomiting, anorexia, contact dermatitis			
Arteether			No information	Not assigned	Yes[c]
Artemether			Artemether levels decreased by darunavir, etravirine, and nevirapine	C	Yes[c]
Artesunate[f]			Mefloquine: levels decreased and clearance accelerated by artesunate	C	Yes[c]
Dihydroartemisinin			Mefloquine: increased absorption	Not assigned	Yes[c]
Atovaquone	Malaria,[b] babesiosis	*Frequent:* nausea, vomiting *Occasional:* abdominal pain, headache	Plasma levels decreased by rifampin, tetracycline, atazanavir, efavirenz, lopinavir/ritonavir; bioavailability decreased by metoclopramide	C	No information

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
Azoles Fluconazole Itraconazole Ketoconazole	Leishmaniasis	*Serious:* hepatotoxicity *Rare:* exfoliative skin disorders, anaphylaxis	Warfarin, oral hypoglycemics, phenytoin, cyclosporine, theophylline, digoxin, dofetilide, quinidine, carbamazepine, rifabutin, busulfan, docetaxel, vinca alkaloids, pimozide, alprazolam, diazepam, midazolam, triazolam, verapamil, atorvastatin, cerivastatin, lovastatin, simvastatin, tacrolimus, sirolimus, indinavir, ritonavir, saquinavir, alfentanil, buspirone, methylprednisolone, trimetrexate: plasma levels increased by azoles Carbamazepine, phenobarbital, phenytoin, isoniazid, rifabutin, rifampin, antacids, H_2-receptor antagonists, proton pump inhibitors, nevirapine: decreased plasma levels of azoles Clarithromycin, erythromycin, indinavir, ritonavir: increased plasma levels of azoles	C	Yes
Benzimidazoles					

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
Albendazole	Ascariasis, capillariasis, clonorchiasis, cutaneous larva migrans, cysticercosis,[b] echinococcosis,[b] enterobiasis, eosinophilic enterocolitis, gnathostomiasis, hookworm, lymphatic filariasis, microsporidiosis, strongyloidiasis, trichinellosis, trichostrongyliasis, trichuriasis, visceral larva migrans	*Occasional:* nausea, vomiting, abdominal pain, headache, reversible alopecia, elevated aminotransferases *Rare:* leukopenia, rash	Dexamethasone, praziquantel: plasma level of albendazole sulfoxide increased by ~50%	C	Yes[c]
Mebendazole	Ascariasis,[b] capillariasis, eosinophilic enterocolitis, enterobiasis,[b] hookworm,[b] trichinellosis, trichostrongyliasis, trichuriasis,[b] visceral larva migrans	*Occasional:* diarrhea, abdominal pain, elevated aminotransferases *Rare:* agranulocytosis, thrombocytopenia, alopecia *Frequent:* anorexia, nausea, vomiting, diarrhea, headache, dizziness, asparagus-like urine odor	Cimetidine: inhibited mebendazole metabolism	C	No information
Thiabendazole	Strongyloidiasis,[b] cutaneous larva migrans,[b] visceral larva migrans[b]	*Occasional:* drowsiness, giddiness, crystalluria, elevated aminotransferases, psychosis *Rare:* hepatitis, seizures, angioneurotic edema, Stevens-Johnson syndrome, tinnitus	Theophylline: serum levels increased by thiabendazole	C	No information
Triclabendazole	Fascioliasis, paragonimiasis	*Occasional:* abdominal cramps, diarrhea, biliary colic, transient headache	No information	Not assigned	Yes

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
Benznidazole	Chagas disease	*Frequent:* rash, pruritus, nausea, leukopenia, paresthesias	No major interactions	Not assigned	No information
Clindamycin	Babesiosis, malaria, toxoplasmosis	*Occasional:* pseudomembranous colitis, abdominal pain, diarrhea, nausea/vomiting *Rare:* pruritus, skin rashes	No major interactions	B	Yes[c]
Diloxanide furoate	Amebiasis	*Frequent:* flatulence *Occasional:* nausea, vomiting, diarrhea *Rare:* pruritus	None reported	Contraindicated	No information
Eflornithine[h] (difluoromethylornithine, DFMO)	Trypanosomiasis	*Frequent:* pancytopenia *Occasional:* diarrhea, seizures *Rare:* transient hearing loss *Severe:* cardiotoxicity	No major interactions	Contraindicated	No information
Emetine and dehydroemetine[f]	Amebiasis, fascioliasis	*Frequent:* pain at injection site *Occasional:* dizziness, headache, GI symptoms	None reported	X	No information
Folate antagonists					
Dihydrofolate reductase inhibitors					
Pyrimethamine	Malaria,[b] isosporiasis, toxoplasmosis[b]	*Occasional:* folate deficiency *Rare:* rash, seizures, severe skin reactions (toxic epidermal necrolysis, erythema multiforme, Stevens-Johnson syndrome)	Sulfonamides, proguanil, zidovudine: increased risk of bone marrow suppression when used concomitantly	C	Yes
Proguanil and chlorproguanil	Malaria	*Occasional:* urticaria *Rare:* hematuria, GI disturbances	Atazanavir, efavirenz, lopinavir/ritonavir: plasma levels of proquanil decreased	C	Yes

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
			Methotrexate: reduced clearance		
Trimethoprim	Cyclosporiasis, isosporiasis	Hyperkalemia, GI upset, mild stomatitis	Warfarin: effect prolonged Phenytoin: hepatic metabolism increased	C	Yes
Dihydropteroate synthetase inhibitors: sulfonamides Sulfadiazine Sulfamethoxazole Sulfadoxine	Malaria,[b] toxoplasmosis[b]	*Frequent:* GI disturbances, allergic skin reactions, crystalluria *Rare:* severe skin reactions (toxic epidermal necrolysis, erythema multiforme, Stevens-Johnson syndrome), agranulocytosis, aplastic anemia, hypersensitivity of the respiratory tract, hepatitis, interstitial nephritis, hypoglycemia, aseptic meningitis	Thiazide diuretics: increased risk of thrombocytopenia in elderly patients Warfarin: effect prolonged by sulfonamides Methotrexate: levels increased by sulfonamides Phenytoin: metabolism impaired by sulfonamides Sulfonylureas: effect prolonged by sulfonamides	B	Yes
Dihydropteroate synthetase inhibitors: sulfones					
Dapsone	Leishmaniasis, malaria, toxoplasmosis	*Frequent:* rash, anorexia *Occasional:* hemolysis, methemoglobinemia, neuropathy, allergic dermatitis, anorexia, nausea, vomiting, tachycardia, headache, insomnia, psychosis, hepatitis *Rare:* agranulocytosis	Rifampin: lowered plasma levels of dapsone	C	Yes
Fumagillin	Microsporidiosis	*Rare:* neutropenia, thrombocytopenia	None reported	No information	No information

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
Furazolidone	Giardiasis	*Frequent:* nausea/vomiting, brown urine *Occasional:* rectal itching, headache *Rare:* hemolytic anemia, disulfiram-like reactions, MAO inhibitor interactions	Risk of hypertensive crisis when administered for >5 days with MAO inhibitors	C	No information
Iodoquinol	Amebiasis,[b] balantidiasis, *D. fragilis* infection	*Occasional:* headache, rash, pruritus, thyrotoxicosis, nausea, vomiting, abdominal pain, diarrhea *Rare:* optic neuritis, peripheral neuropathy, seizures, encephalopathy	No major interactions	C	No information
Ivermectin	Ascariasis, cutaneous larva migrans, gnathostomiasis, loiasis, lymphatic filariasis, onchocerciasis,[b] scabies, strongyloidiasis,[b] trichuriasis	*Occasional:* fever, pruritus, headache, myalgias *Rare:* hypotension	No major interactions	C	Yes[c]
Macrolides					
Azithromycin	Babesiosis	*Occasional:* nausea, vomiting, diarrhea, abdominal pain *Rare:* angioedema, cholestatic jaundice	Cyclosporine and digoxin: levels increased by azithromycin Nelfinavir: increased levels of azithromycin	B	Yes
Spiramycin[h]	Toxoplasmosis	*Occasional:* GI disturbances, transient skin eruptions *Rare:* thrombocytopenia, QT prolongation in an infant, cholestatic hepatitis	No major interactions	Not assigned[d]	Yes[c]

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
Mefloquine	Malaria[b]	*Frequent:* lightheadedness, nausea, headache *Occasional:* confusion; nightmares; insomnia; visual disturbance; transient and clinically silent ECG abnormalities, including sinus bradycardia, sinus arrhythmia, first-degree AV block, prolongation of QTc interval, and abnormal T waves *Rare:* psychosis, convulsions, hypotension	Administration of halofantrine <3 weeks after mefloquine use may produce fatal QTc prolongation. Mefloquine may lower plasma levels of anticonvulsants. Levels are decreased and clearance is accelerated by artesunate. Mefloquine decreases plasma levels of ritonavir and possibly other protease inhibitors.	C	Yes
Melarsoprol[f]	Trypanosomiasis	*Frequent:* myocardial injury, encephalopathy, peripheral neuropathy, hypertension *Occasional:* G6PD-induced hemolysis, erythema nodosum leprosum *Rare:* hypotension	No major interactions	Not assigned	No information
Metrifonate	Schistosomiasis	*Frequent:* abdominal pain, nausea, vomiting, diarrhea, headache, vertigo, bronchospasm *Rare:* cholinergic symptoms	No major interactions	B	No
Miltefosine	Leishmaniasis,[b] primary amebic meningoencephalitis	*Frequent:* mild and transient (1–2 days) GI disturbances within first 2 weeks of therapy (resolve after treatment completion); motion sickness *Occasional:* reversible elevations of creatinine and aminotransferases	No major interactions	Not assigned	No information

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
Niclosamide	Intestinal cestode infections[b]	*Occasional:* nausea, vomiting, dizziness, pruritus	No major interactions	B	No information
Nifurtimox[f]	Chagas disease	*Frequent:* nausea, vomiting, abdominal pain, insomnia, paresthesias, weakness, tremors *Rare:* seizures (all reversible and dose-related)	No major interactions	Not assigned	No information
Nitazoxanide	Cryptosporidiosis,[b] giardiasis[b]	*Occasional:* abdominal pain, diarrhea *Rare:* vomiting, headache	Increases plasma levels of highly protein-bound drugs (e.g., phenytoin, warfarin)	B	No information
Nitroimidazoles					
Metronidazole	Amebiasis,[b] balantidiasis, dracunculiasis, giardiasis, trichomoniasis,[b] *D. fragilis* infection	*Frequent:* nausea, headache, anorexia, metallic aftertaste *Occasional:* vomiting, insomnia, vertigo, paresthesias, disulfiram-like effects *Rare:* seizures, peripheral neuropathy	Warfarin: effect enhanced by metronidazole Disulfiram: psychotic reaction Phenobarbital, phenytoin: accelerate elimination of metronidazole Lithium: serum levels elevated by metronidazole Cimetidine: prolonged half-life of metronidazole Oral solutions of antiretrovirals containing alcohol: disulfiram effect due to alcohol	B	Yes
Tinidazole	Amebiasis,[b] giardiasis, trichomoniasis	*Occasional:* nausea, vomiting, metallic taste	See metronidazole	C	Yes
Oxamniquine	Schistosomiasis	*Occasional:* dizziness, drowsiness, headache, orange urine, elevated aminotransferases *Rare:* seizures	No major interactions	C	No information

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
Pentamidine isethionate	Leishmaniasis, trypanosomiasis	*Frequent:* hypotension, hypoglycemia, pancreatitis, sterile abscesses at IM injection sites, GI disturbances, reversible renal failure *Occasional:* hepatotoxicity, cardiotoxicity, delirium *Rare:* anaphylaxis	No major interactions	C	No information
Piperazine and derivatives					
Piperazine	Ascariasis, enterobiasis	*Occasional:* nausea, vomiting, diarrhea, abdominal pain, headache *Rare:* neurotoxicity, seizures	None reported	C	No information
Diethylcarbamazine[f]	Lymphatic filariasis, loiasis, tropical pulmonary eosinophilia	*Frequent:* dose-related nausea, vomiting *Rare:* fever, chills, arthralgias, headaches	None reported	Not assigned[d]	No information
Praziquantel	Clonorchiasis,[b] cysticercosis, diphyllobothriasis, hymenolepiasis, taeniasis, opisthorchiasis, intestinal trematodes, paragonimiasis, schistosomiasis[b]	*Frequent:* abdominal pain, diarrhea, dizziness, headache, malaise *Occasional:* fever, nausea *Rare:* pruritus, singultus	No major interactions	B	Yes
Pyrantel pamoate	Ascariasis, eosinophilic enterocolitis, enterobiasis,[b] hookworm, trichostrongyliasis	*Occasional:* GI disturbances, headache, dizziness, elevated aminotransferases	No major interactions	C	No information
Pyronaridine	Malaria	*Occasional:* headache, nausea	None reported to date	B	Yes

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
Quinacrine[h]	Giardiasis[b]	*Frequent:* headache, nausea, vomiting, bitter taste *Occasional:* yellow-orange discoloration of skin, sclerae, urine; begins after 1 week of treatment and lasts up to 4 months after drug discontinuation *Rare:* psychosis, exfoliative dermatitis, retinopathy, G6PD-induced hemolysis, exacerbation of psoriasis, disulfiram-like effects	Primaquine: toxicity potentiated by quinacrine	C	No information
Quinine and quinidine	Malaria, babesiosis	*Frequent:* cinchonism (tinnitus, high-tone deafness, headache, dysphoria, nausea, vomiting, abdominal pain, visual disturbances, postural hypotension), hyperinsulinemia resulting in life-threatening hypoglycemia *Occasional:* deafness, hemolytic anemia, arrhythmias, hypotension due to rapid IV infusion	Carbonic anhydrase inhibitors, thiazide diuretics: reduced renal elimination of quinidine Amiodarone, cimetidine: increased quinidine levels Nifedipine: decreased quinidine levels; quinidine slows metabolism of nifedipine Phenobarbital, phenytoin, rifampin: accelerated hepatic elimination of quinidine Verapamil: reduced hepatic clearance of quinidine Diltiazem: decreased clearance of quinidine	X	Yes[c]
Quinolones					

Drugs by Class	Parasitic Infection(S)	Adverse Effects	Major Drug–Drug Interactions	Pregnancy Class[a]	Breast Milk
Ciprofloxacin	Cyclosporiasis, isosporiasis	*Occasional:* nausea, diarrhea, vomiting, abdominal pain/discomfort, headache, restlessness, rash *Rare:* myalgias/arthralgias, tendon rupture, CNS symptoms (nervousness, agitation, insomnia, anxiety, nightmares or paranoia); convulsions	Probenecid: increased serum levels of ciprofloxacin Theophylline, warfarin: serum levels increased by ciprofloxacin	C	Yes
Suramin[f]	Trypanosomiasis	*Frequent:* immediate: fever, urticaria, nausea, vomiting, hypotension; delayed (up to 24 h): exfoliative dermatitis, stomatitis, paresthesias, photophobia, renal dysfunction *Occasional:* nephrotoxicity, adrenal toxicity, optic atrophy, anaphylaxis *Frequent:* GI disturbances	No major interactions	Not assigned	No information
Tetracyclines	Balantidiasis, *D. fragilis* infection, malaria; lymphatic filariasis (doxycycline)	*Occasional:* photosensitivity dermatitis *Rare:* exfoliative dermatitis, esophagitis, hepatotoxicity	Warfarin: effect prolonged by tetracyclines	D	Yes

[a]Based on U.S. Food and Drug Administration (FDA) pregnancy categories of A–D, X. [b]Approved by the FDA for this indication. [c]Not believed to be harmful. [d]Use in pregnancy is recommended by international organizations outside the United States. [e]Only AmBisome has been approved by the FDA for this indication. [f]Available through the CDC. [g]Only artemether (in combination with lumefantrine) and artesunate have been approved by the FDA for this indication. [h]Available through the manufacturer.

Abbreviations: ACTH, adrenocorticotropic hormone; AV, atrioventricular; CNS, central nervous system; ECG, electrocardiogram; G6PD, glucose 6-phosphate dehydrogenase; GI, gastrointestinal; MAO, monoamine oxidase.

Albendazole

Like all benzimidazoles, albendazole acts by selectively binding to free β-tubulin in nematodes, inhibiting the polymerization of tubulin and the microtubule-dependent uptake of glucose. Irreversible damage occurs in gastrointestinal (GI) cells of the nematodes, resulting in starvation, death, and expulsion by the host. This fundamental disruption of cellular metabolism offers treatment for a wide range of parasitic diseases.

Albendazole is poorly absorbed from the GI tract, a feature that is advantageous for the treatment of intestinal helminths but not for that of tissue helminth infections (e.g., hydatid disease and neurocysticercosis), which requires that a sufficient amount of active drug reach the site of infection. Administration with a high-fat meal (~40 g) increases the drug's absorption by up to fivefold. The metabolite albendazole sulfoxide is responsible for the drug's therapeutic effect outside the gut lumen. Albendazole sulfoxide crosses the blood–brain barrier, reaching a level significantly higher than that achieved in plasma. The high concentrations of albendazole sulfoxide attained in cerebrospinal fluid (CSF) may explain the efficacy of albendazole in the treatment of neurocysticercosis.

Albendazole is extensively metabolized in the liver, but there are few data regarding the drug's use in patients with hepatic disease. Single-dose albendazole therapy in humans is largely without side effects (overall frequency, ≤1%). More prolonged courses (e.g., as administered for cystic and alveolar echinococcal disease) have been associated with liver function abnormalities and bone marrow toxicity. Thus, when prolonged use is anticipated, the drug should be administered in treatment cycles of 28 days interrupted by 14-day intervals off therapy. Prolonged therapy with full-dose albendazole (800 mg/d) should be approached cautiously in patients also receiving drugs with known effects on the cytochrome P450 system.

Amodiaquine

Amodiaquine has been widely used in the treatment of malaria for >60 years. Like chloroquine (the other major 4-aminoquinoline), amodiaquine is now of limited use because of the spread of resistance. Amodiaquine interferes with hemozoin formation through complexation with heme. It is rapidly absorbed and acts as a prodrug after oral administration; the principal plasma metabolite, monodesethylamodiaquine, is the predominant antimalarial agent. Amodiaquine and its metabolites are all excreted in urine, but there are no recommendations concerning dosage adjustment in patients with impaired renal function. Agranulocytosis and hepatotoxicity can develop with repeated use; therefore, this drug should not be used for prophylaxis. Despite widespread resistance, amodiaquine is effective in some areas when combined with other antimalarial drugs (e.g., artesunate, sulfadoxine-pyrimethamine), particularly in children. Although on the World Health Organization's *List of Essential Medicines*, amodiaquine is not yet available in the United States.

Amphotericin B

See Table 217-1 and Chap. 206.

Antimonials*

Despite associated adverse reactions and the need for prolonged parenteral treatment, the pentavalent antimonial compounds (designated Sb^V) have remained the first-line therapy for all forms of leishmaniasis throughout the world, primarily because they are affordable and effective and have survived the test of time. Pentavalent antimonials are active only after bioreduction to the trivalent Sb(III) form, which inhibits trypanothione reductase, a critical enzyme involved in the oxidative stress management of *Leishmania* species. The fact that *Leishmania* species use trypanothione rather than glutathione (which is used by mammalian cells) may explain the parasite-specific activity of antimonials. The drugs are taken up by the reticuloendothelial system, and their activity against *Leishmania* species may be enhanced by this localization. Sodium stibogluconate is the only pentavalent antimonial available in the United States; meglumine antimoniate is used principally in francophone countries.

Resistance is a major problem in some areas. Although low-level unresponsiveness to Sb^V was identified in India in the 1970s, incremental increases in both the recommended daily dosage (to 20 mg/kg) and the duration of treatment (to 28 days) satisfactorily compensated for the growing resistance until around 1990. There has since been steady erosion in the capacity of Sb^V to induce long-term cure in patients with kala-azar who live in eastern India. Co-infection with HIV impairs the treatment response.

Sodium stibogluconate is available in aqueous solution and is administered parenterally. Antimony appears to have two elimination phases. When the drug is administered IV, the mean half-life of the first phase is <2 h; the mean half-life of the terminal elimination phase is nearly 36 h. This slower phase may be due to conversion of pentavalent antimony to a trivalent form that is the likely cause of the side effects often seen with prolonged therapy.

Artemisinin Derivatives*

Artesunate, artemether, artemotil, and the parent compound artemisinin are sesquiterpene lactones derived from the wormwood plant *Artemisia annua*. These agents are at least tenfold more potent in vivo than other antimalarial drugs and presently show no cross-resistance with known antimalarial drugs; thus, they have become first-line agents for the treatment of severe falciparum malaria. The artemisinin compounds are rapidly effective against the asexual blood forms of *Plasmodium* species but are not active against intrahepatic forms. With the exception of artesunate, artemisinin and its derivatives are highly lipid soluble and readily cross both host and parasite cell membranes. One factor that explains the drugs' highly selective toxicity against malaria is that parasitized erythrocytes concentrate artemisinin and its derivatives to concentrations 100-fold higher than those in uninfected erythrocytes. The antimalarial effect of these agents results primarily from the active metabolite dihydroartemisinin; in the presence of heme or molecular iron, the endoperoxide moiety of dihydroartemisinin decomposes, generating free radicals and other metabolites that damage parasite proteins. The compounds are available for oral, rectal, IV, or IM administration, depending on the derivative. In the United States, IV artesunate is available for the treatment of severe, quinidine-unresponsive malaria through the CDC malaria hotline (770-488-7788 or 855-856-4713 [toll-free], M–F, 0800–1630 EST; 770-488-7100 after hours). Artemisinin and its derivatives are cleared rapidly from the circulation. Their short half-lives limit their value for prophylaxis and monotherapy. Side effects appear to be minor, although sinus bradycardia and transient first-degree heart block have been reported. Although seen in animal models, embryotoxicity and neurotoxicity have not been identified in humans despite active investigation. These agents should be used only in combination with another, longer-acting agent (e.g., artesunate-mefloquine, dihydroartemisinin-piperaquine). A combined formulation of artemether and lumefantrine is available for the treatment of acute uncomplicated falciparum malaria acquired in areas where *Plasmodium falciparum* is resistant to chloroquine and antifolates.

Atovaquone

Atovaquone is a hydroxynaphthoquinone that exerts broad-spectrum antiprotozoal activity via selective inhibition of parasite mitochondrial electron transport. This agent exhibits potent activity against toxoplasmosis and babesiosis when used with pyrimethamine and azithromycin, respectively. Atovaquone possesses a novel mode of action against *Plasmodium* species, inhibiting the electron transport system at the level of the cytochrome bc1 complex. The drug is active against both the erythrocytic and the exoerythrocytic stages of *Plasmodium* species; however, because it does not eradicate hypnozoites from the liver, patients with *Plasmodium vivax* or *Plasmodium ovale* infections must be given radical prophylaxis.

Malarone® is a fixed-dose combination of atovaquone and proguanil used for malaria prophylaxis as well as for the treatment of acute, uncomplicated *P. falciparum* malaria. Malarone has been shown to be effective in regions with multidrug-resistant *P. falciparum*. Resistance to atovaquone develops rapidly via mutations in the parasite's mitochondrial cytochrome b complex. However, the mutations result in sterility of female parasites; thus, atovaquone-resistant parasites cannot be transmitted to another person. This situation may explain why clinical resistance has yet to be reported.

The bioavailability of atovaquone varies considerably. Absorption after a single oral dose is slow, increases two- to threefold with a fatty meal, and is dose-limited above 750 mg. The elimination half-life is increased in patients with moderate hepatic impairment. Because of the potential for drug accumulation, the use of atovaquone is generally contraindicated in persons with a creatinine clearance rate <30 mL/min. No dosage adjustments are needed in patients with mild to moderate renal impairment.

Azithromycin

See Table 217-1 and Chap. 139.

Azoles

See Table 217-1 and Chap. 206.

Benznidazole*

This oral nitroimidazole derivative is used to treat Chagas disease, with cure rates of 80–90% recorded in acute infections. Benznidazole is believed to exert its trypanocidal effects by generating oxygen radicals to which the

parasites are more sensitive than mammalian cells because of a relative deficiency in antioxidant enzymes. Benznidazole also appears to alter the balance between pro- and anti-inflammatory mediators by downregulating the synthesis of nitrite, interleukin (IL) 6, and IL-10 in macrophages. Benznidazole is highly lipophilic and readily absorbed. The drug is extensively metabolized; only 5% of the dose is excreted unchanged in the urine. Benznidazole is well tolerated; adverse effects are rare and usually manifest as GI upset or pruritic rash.

Chloroquine

This 4-aminoquinoline has marked, rapid schizonticidal and gametocidal activity against blood forms of *P. ovale* and *Plasmodium malariae* and against susceptible strains of *P. vivax* and *P. falciparum*. It is not active against intrahepatic forms (*P. vivax* and *P. ovale*). Parasitized erythrocytes accumulate chloroquine in significantly greater concentrations than do normal erythrocytes. Chloroquine, a weak base, concentrates in the food vacuoles of intraerythrocytic parasites because of a relative pH gradient between the extracellular space and the acidic food vacuole. Once it enters the acidic food vacuole, chloroquine is rapidly converted to a membrane-impermeable protonated form and is trapped. Continued accumulation of chloroquine in the parasite's acidic food vacuoles results in drug levels that are 600-fold higher at this site than in plasma. The high accumulation of chloroquine results in an increase in pH within the food vacuole to a level above that required for the acid proteases' optimal activity, inhibiting parasite heme polymerase; as a result, the parasite is effectively killed with its own metabolic waste. Compared with susceptible strains, chloroquine-resistant plasmodia transport chloroquine out of intraparasitic compartments more rapidly and maintain lower chloroquine concentrations in their acid vesicles. Hydroxychloroquine, a congener of chloroquine, is equivalent to chloroquine in its antimalarial efficacy but is preferred to chloroquine for the treatment of autoimmune disorders because it produces less ocular toxicity when used in high doses.

Chloroquine is well absorbed. However, because it exhibits extensive tissue binding, a loading dose is required to yield effective plasma concentrations. A therapeutic drug level in plasma is reached 2–3 h after oral administration (the preferred route). Chloroquine can be administered IV, but excessively rapid parenteral administration can result in seizures and death from cardiovascular collapse. The mean half-life of chloroquine is 4 days, but the rate of excretion decreases as plasma levels decline, making once-weekly administration possible for prophylaxis in areas with sensitive strains. About one-half of the parent drug is excreted in urine, but the dose should not be reduced for persons with acute malaria and renal insufficiency.

Ciprofloxacin

See Table 217-1 and Chap. 139.

Clindamycin

See Table 217-1 and Chap. 139.

Dapsone

See Table 217-1 and Chap. 176.

Dehydroemetine

Emetine is an alkaloid derived from ipecac; dehydroemetine is synthetically derived from emetine and is considered less toxic. Both agents are active against *Entamoeba histolytica* and appear to work by blocking peptide elongation and thus inhibiting protein synthesis. Emetine is rapidly absorbed after parenteral administration, rapidly distributed throughout the body, and slowly excreted in the urine in unchanged form. Both agents are contraindicated in patients with renal disease.

Diethylcarbamazine*

A derivative of the antihelminthic agent piperazine with a long history of successful use, diethylcarbamazine (DEC) remains the treatment of choice for lymphatic filariasis and loiasis and has also been used for visceral larva migrans. Although piperazine itself has no antifilarial activity, the piperazine ring of DEC is essential for the drug's activity. DEC's mechanism of action remains to be fully defined. Proposed mechanisms include immobilization due to inhibition of parasite cholinergic muscle receptors, disruption of microtubule formation, and alteration of

helminthic surface membranes resulting in enhanced killing by the host's immune system. DEC enhances adherence properties of eosinophils. The development of resistance under drug pressure (i.e., a progressive decrease in efficacy when the drug is used widely in human populations) has not been observed, although DEC has variable effects when administered to persons with filariasis. Monthly administration provides effective prophylaxis against both bancroftian filariasis and loiasis.

DEC is well absorbed after oral administration, with peak plasma concentrations reached within 1–2 h. No parenteral form is available. The drug is eliminated largely by renal excretion, with <5% found in feces. If more than one dose is to be administered to an individual with renal dysfunction, the dose should be reduced commensurate with the reduction in creatinine clearance rate. Alkalinization of the urine prevents renal excretion and increases the half-life of DEC. Use in patients with onchocerciasis can precipitate a Mazzotti reaction, with pruritus, fever, and arthralgias. Like other piperazines, DEC is active against *Ascaris* species. Patients co-infected with this nematode may expel live worms after treatment.

Diloxanide Furoate

Diloxanide furoate, a substituted acetanilide, is a luminally active agent used to eradicate the cysts of *E. histolytica*. After ingestion, diloxanide furoate is hydrolyzed by enzymes in the lumen or mucosa of the intestine, releasing furoic acid and the ester diloxanide; the latter acts directly as an amebicide.

Diloxanide furoate is given alone to asymptomatic cyst passers. For patients with active amebic infections, diloxanide is generally administered in combination with a 5-nitroimidazole such as metronidazole or tinidazole. Diloxanide furoate is rapidly absorbed after oral administration. When coadministered with a 5-nitroimidazole, diloxanide levels peak within 1 h and disappear within 6 h. About 90% of an oral dose is excreted in the urine within 48 h, chiefly as the glucuronide metabolite. Diloxanide furoate is contraindicated in pregnant and breast-feeding women and in children <2 years of age.

Eflornithine*

Eflornithine (difluoromethylornithine, or DFMO) is a fluorinated analogue of the amino acid ornithine. Although originally designed as an antineoplastic agent, eflornithine has proven effective against some trypanosomatids.

Eflornithine has specific activity against all stages of infection with *Trypanosoma brucei gambiense*; however, it is inactive against *Trypanosoma brucei rhodesiense*. The drug acts as an irreversible suicide inhibitor of ornithine decarboxylase, the first enzyme in the biosynthesis of the polyamines putrescine and spermidine. Polyamines are essential for the synthesis of trypanothione, an enzyme required for the maintenance of intracellular thiols in the correct redox state and for the removal of reactive oxygen metabolites. However, polyamines are also essential for cell division in eukaryotes, and ornithine decarboxylase is similar in trypanosomes and mammals. The selective antiparasitic activity of eflornithine is partly explained by the structure of the trypanosomal enzyme, which lacks a 36-amino-acid C-terminal sequence found on mammalian ornithine decarboxylase. This difference results in a lower turnover of ornithine decarboxylase and a more rapid decrease of polyamines in trypanosomes than in the mammalian host. The diminished effectiveness of eflornithine against *T. b. rhodesiense* appears to be due to the parasite's ability to replace the inhibited enzyme more rapidly than *T. b. gambiense*.

Eflornithine is less toxic but more costly than conventional therapy. It can be administered IV or PO. The dose should be reduced in renal failure. Eflornithine readily crosses the blood–brain barrier; CSF levels are highest in persons with the most severe central nervous system (CNS) involvement.

Fumagillin†

Originally discovered as an anti-angiogenic compound derived from the fungus *Aspergillus fumigatus*, fumagillin is a water-insoluble antibiotic that is active against microsporidia and is used topically to treat ocular infections due to *Encephalitozoon* species. When given systemically, fumagillin was effective but caused thrombocytopenia in all recipients in the second week of treatment; this side effect was readily reversed when administration of the drug was stopped. Fumagillin acts by binding to methionine aminopeptidase 2, thus inhibiting microsporidial replication by irreversibly blocking the active site.

Furazolidone

This nitrofuran derivative is an effective alternative agent for the treatment of giardiasis and also exhibits activity against *Isospora belli*. Because it is the only agent active against *Giardia* that is available in liquid form, it is most often used to treat young children. Furazolidone undergoes reductive activation in *Giardia lamblia* trophozoites—an event that, unlike the reductive activation of metronidazole, involves an NADH oxidase. The killing effect correlates with the toxicity of reduced products, which damage important cellular components, including DNA. Although furazolidone had been thought to be largely unabsorbed when administered orally, the occurrence of systemic adverse reactions indicates that this is not the case. More than 65% of the drug dose can be recovered from the urine as colored metabolites. Omeprazole reduces the oral bioavailability of furazolidone.

Furazolidone is a monoamine oxidase (MAO) inhibitor; thus, caution should be used in its concomitant administration with other drugs (especially indirectly acting sympathomimetic amines) and in the consumption of food and drink containing tyramine during treatment. However, hypertensive crises have not been reported in patients receiving furazolidone, and it has been suggested that—because furazolidone inhibits MAOs gradually over several days—the risks are small if treatment is limited to a 5-day course. Because hemolytic anemia can occur in patients with glucose-6-phosphate dehydrogenase (G6PD) deficiency and glutathione instability, furazolidone treatment is contraindicated in mothers who are breast-feeding and in neonates.

Halofantrine

This 9-phenanthrenemethanol is one of three classes of arylaminoalcohols first identified as potential antimalarial agents by the World War II Malaria Chemotherapy Program. Its activity is believed to be similar to that of chloroquine, although it is an oral alternative for the treatment of malaria due to chloroquine-resistant *P. falciparum*.

Halofantrine is thought to share one or more mechanisms with the 4-aminoquinolines, forming a complex with ferriprotoporphyrin IX and interfering with the degradation of hemoglobin. It has been shown to bind to plasmepsin, a hemoglobin-degrading enzyme unique to plasmodia.

Halofantrine exhibits erratic bioavailability, but its absorption is significantly enhanced when it is taken with a fatty meal. The elimination half-life of halofantrine is 1–2 days; it is excreted mainly in feces. Halofantrine is metabolized into *N*-debutyl-halofantrine by the cytochrome P450 enzyme CYP3A4. Grapefruit juice should be avoided during treatment because it increases both halofantrine's bioavailability and halofantrine-induced QT interval prolongation by inhibiting CYP3A4 at the enterocyte level.

Iodoquinol

Iodoquinol (diiodohydroxyquin), a hydroxyquinoline, is an effective luminal agent for the treatment of amebiasis, balantidiasis, and infection with *Dientamoeba fragilis*. Its mechanism of action is unknown. It is poorly absorbed. Because the drug contains 64% organically bound iodine, it should be used with caution in patients with thyroid disease. Iodine dermatitis occurs occasionally during iodoquinol treatment. Protein-bound serum iodine levels may be increased during treatment and can interfere with certain tests of thyroid function. These effects may persist for as long as 6 months after discontinuation of therapy. Iodoquinol is contraindicated in patients with liver disease. Most serious are the reactions related to prolonged high-dose therapy (optic neuritis, peripheral neuropathy), which should not occur if the recommended dosage regimens are followed.

Ivermectin

Ivermectin (22,23-dihydroavermectin) is a derivative of the macrocyclic lactone avermectin produced by the soil-dwelling actinomycete *Streptomyces avermitilis*. Ivermectin is active at low doses against a wide range of helminths and ectoparasites. It is the drug of choice for the treatment of onchocerciasis, strongyloidiasis, cutaneous larva migrans, and scabies. Ivermectin is highly active against microfilariae of the lymphatic filariases but has no macrofilaricidal activity. When ivermectin is used in combination with other agents such as DEC or albendazole for treatment of lymphatic filariasis, synergistic activity is seen. Although active against the intestinal helminths *Ascaris lumbricoides* and *Enterobius vermicularis*, ivermectin is only variably effective in trichuriasis and is ineffective against hookworms. Widespread use of ivermectin for treatment of intestinal nematode infections in sheep and goats has led to the emergence of drug resistance in veterinary practice; this development may portend problems in human medical use.

Data suggest that ivermectin acts by opening the neuromuscular membrane-associated, glutamate-dependent chloride channels. The influx of chloride ions results in hyperpolarization and muscle paralysis—particularly of the

nematode pharynx, with consequent blockage of the oral ingestion of nutrients. As these chloride channels are present only in invertebrates, paralysis is seen only in the parasite.

Ivermectin is available for administration to humans only as an oral formulation. The drug is highly protein bound; it is almost completely excreted in feces. Both food and beer increase the bioavailability of ivermectin significantly. Ivermectin is distributed widely throughout the body; animal studies indicate that it accumulates at the highest concentration in adipose tissue and liver, with little accumulation in the brain. Few data exist to guide therapy in hosts with conditions that may influence drug pharmacokinetics.

Ivermectin is generally administered as a single dose of 150–200 µg/kg. In the absence of parasitic infection, the adverse effects of ivermectin in therapeutic doses are minimal. Adverse effects in patients with filarial infections include fever, myalgia, malaise, lightheadedness, and (occasionally) postural hypotension. The severity of such side effects is related to the intensity of parasite infection, with more symptoms in individuals with a heavy parasite burden. In onchocerciasis, skin edema, pruritus, and mild eye irritation may also occur. The adverse effects are generally self-limiting and only occasionally require symptom-based treatment with antipyretics or antihistamines. More severe complications of ivermectin therapy for onchocerciasis include encephalopathy in patients heavily infected with *Loa loa*.

Lumefantrine

Lumefantrine (benflumetol), a fluorene arylaminoalcohol derivative synthesized in the 1970s by the Chinese Academy of Military Medical Sciences (Beijing), has marked blood schizonticidal activity against a wide range of plasmodia. This agent conforms structurally and in mode of action to other arylaminoalcohols (quinine, mefloquine, and halofantrine). Lumefantrine exerts its antimalarial effect as a consequence of its interaction with heme, a degradation product of hemoglobin metabolism. Although its antimalarial activity is slower than that of the artemisinin-based drugs, the recrudescence rate with the recommended lumefantrine regimen is lower. The pharmacokinetic properties of lumefantrine are reminiscent of those of halofantrine, with variable oral bioavailability, considerable augmentation of oral bioavailability by concomitant fat intake, and a terminal elimination half-life of ~4–5 days in patients with malaria.

Artemether and lumefantrine have synergistic activity, and the combined formulation of artemether and lumefantrine is effective for the treatment of falciparum malaria in areas where *P. falciparum* is resistant to chloroquine and antifolates.

Mebendazole

This benzimidazole is a broad-spectrum antiparasitic agent widely used to treat intestinal helminthiases. Its mechanism of action is similar to that of albendazole; however, it is a more potent inhibitor of parasite malic dehydrogenase and exhibits a more specific and selective effect against intestinal nematodes than the other benzimidazoles.

Mebendazole is available only in oral form but is poorly absorbed from the GI tract; only 5–10% of a standard dose is measurable in plasma. The proportion absorbed from the GI tract is extensively metabolized in the liver. Metabolites appear in the urine and bile; impaired liver or biliary function results in higher plasma mebendazole levels in treated patients. No dose reduction is warranted in patients with renal function impairment. Because mebendazole is poorly absorbed, its incidence of side effects is low. Transient abdominal pain and diarrhea sometimes occur, usually in persons with massive parasite burdens.

Mefloquine

Mefloquine is effective prophylaxis of chloroquine-resistant malaria; high doses can be used for treatment. Despite the development of drug-resistant strains of *P. falciparum* in parts of Africa and Southeast Asia, mefloquine remains an effective drug throughout most of the world. Cross-resistance of mefloquine with halofantrine and with quinine has been documented in limited areas. Like quinine and chloroquine, this quinoline is active only against the asexual erythrocytic stages of malarial parasites. Unlike quinine, however, mefloquine has a relatively poor affinity for DNA and, as a result, does not inhibit the synthesis of parasitic nucleic acids and proteins. Although both mefloquine and chloroquine inhibit hemozoin formation and heme degradation, mefloquine differs in that it forms a complex with heme that may be toxic to the parasite.

Mefloquine HCl is poorly water soluble and intensely irritating when given parenterally; thus it is available only in tablet form. Its absorption is adversely affected by vomiting and diarrhea but is significantly enhanced when the drug

is administered with or after food. About 98% of the drug binds to protein. Mefloquine is excreted mainly in the bile and feces; therefore, no dose adjustment is needed in persons with renal insufficiency. The drug and its main metabolite are not appreciably removed by hemodialysis. No special chemoprophylactic dosage adjustments are indicated for the achievement of plasma concentrations in dialysis patients that are similar to those in healthy persons. Pharmacokinetic differences have been detected among various ethnic populations; however, these distinctions are of minor importance compared with host immune status and parasite sensitivity. In patients with impaired liver function, the elimination of mefloquine may be prolonged, leading to higher plasma levels.

Mefloquine should be used with caution by individuals participating in activities requiring alertness and fine-motor coordination because dizziness, vertigo, or tinnitus can develop and persist. If the drug is to be administered for a prolonged period, periodic evaluations are recommended, including liver function tests and ophthalmic examinations. Sleep abnormalities (insomnia, abnormal dreams) have occasionally been reported. Psychosis and seizures occur rarely; mefloquine should not be prescribed to patients with neuropsychiatric conditions. The development of acute anxiety, depression, restlessness, or confusion may be considered prodromal to a more serious event, and the drug should be discontinued.

Concomitant use of quinine, quinidine, or drugs producing β-adrenergic blockade may cause significant electrocardiographic abnormalities or cardiac arrest. Halofantrine must not be given simultaneously with or <3 weeks after mefloquine because a potentially fatal prolongation of the QTc interval on electrocardiography may occur. No data exist on mefloquine use after halofantrine use. Administration of mefloquine with quinine or chloroquine may increase the risk of convulsions. Mefloquine may lower plasma levels of anticonvulsants. Caution should be exercised with regard to concomitant antiretroviral therapy, since mefloquine has been shown to exert variable effects on ritonavir pharmacokinetics that are not explained by hepatic CYP3A4 activity or ritonavir protein binding. Vaccinations with attenuated live bacteria should be completed at least 3 days before the first dose of mefloquine.

Women of childbearing age who are traveling to areas where malaria is endemic should be warned against becoming pregnant and encouraged to practice contraception during malaria prophylaxis with mefloquine and for up to 3 months thereafter. However, in the case of unplanned pregnancy, use of mefloquine is not considered an indication for pregnancy termination. Analysis of prospectively monitored cases demonstrates a prevalence of birth defects and fetal loss comparable to background rates.

Melarsoprol*

Melarsoprol has been used since 1949 for the treatment of human African trypanosomiasis. This trivalent arsenical compound is indicated for the treatment of African trypanosomiasis with neurologic involvement and for the treatment of early disease that is resistant to suramin or pentamidine. Melarsoprol, like other drugs containing heavy metals, interacts with thiol groups of several different proteins; however, its antiparasitic effects appear to be more specific. Trypanothione reductase is a key enzyme involved in the oxidative stress management of both *Trypanosoma* and *Leishmania* species, helping to maintain an intracellular reducing environment by reduction of disulfide trypanothione to its dithiol derivative dihydrotrypanothione. Melarsoprol sequesters dihydrotrypanothione, depriving the parasite of its main sulfhydryl antioxidant, and inhibits trypanothione reductase, depriving the parasite of the essential enzyme system that is responsible for keeping trypanothione reduced. These effects are synergistic. The selectivity of arsenical action against trypanosomes is due at least in part to the greater melarsoprol affinity of reduced trypanothione than of other monothiols (e.g., cysteine) on which the mammalian host depends for maintenance of high thiol levels. Melarsoprol enters the parasite via an adenosine transporter; drug-resistant strains lack this transport system.

Melarsoprol is always administered IV. A small but therapeutically significant amount of the drug enters the CSF. The compound is excreted rapidly, with ~80% of the arsenic found in feces.

Melarsoprol is highly toxic. The most serious adverse reaction is reactive encephalopathy, which affects 6% of treated individuals and usually develops within 4 days of the start of therapy, with an average case–fatality rate of 50%. Glucocorticoids are administered with melarsoprol to prevent this development. Because melarsoprol is intensely irritating, care must be taken to avoid infiltration of the drug.

Metrifonate

Metrifonate has selective activity against *Schistosoma haematobium*. This organophosphorous compound is a prodrug that is converted nonenzymatically to dichlorvos (2,2-dichlorovinyl dimethylphosphate, DDVP), a highly active chemical that irreversibly inhibits the acetylcholinesterase enzyme. Schistosomal cholinesterase is more

susceptible to dichlorvos than is the corresponding human enzyme. The exact mechanism of action of metrifonate is uncertain, but the drug is believed to inhibit tegumental acetylcholine receptors that mediate glucose transport.

Metrifonate is administered in a series of three doses at 2-week intervals. After a single oral dose, metrifonate produces a 95% decrease in plasma cholinesterase activity within 6 h, with a fairly rapid return to normal. However, 2.5 months are required for erythrocyte cholinesterase levels to return to normal. Treated persons should not be exposed to neuromuscular blocking agents or organophosphate insecticides for at least 48 h after treatment.

Metronidazole and Other Nitroimidazoles

See Table 217-1 and **Chap. 139**.

Miltefosine

In the early 1990s, miltefosine (hexadecylphosphocholine), originally developed as an antineoplastic agent, was discovered to have significant antiproliferative activity against *Leishmania* species, *Trypanosoma cruzi*, and *Trypanosoma brucei* parasites in vitro and in experimental animal models. Miltefosine is the first oral drug that has proved to be highly effective and comparable to amphotericin B against visceral leishmaniasis in India, where antimonial-resistant cases are prevalent. Miltefosine is also effective in previously untreated visceral infections. Cure rates in cutaneous leishmaniasis are comparable to those obtained with antimony. Miltefosine is also effective against the free-living ameba *Naegleria fowleri*.

The activity of miltefosine is attributed to interaction with cell signal transduction pathways and inhibition of phospholipid and sterol biosynthesis. Resistance to miltefosine has not been observed clinically. The drug is readily absorbed from the GI tract, is widely distributed, and accumulates in several tissues. The efficacy of a 28-day treatment course in Indian visceral leishmaniasis is equivalent to that of amphotericin B therapy; however, it appears that a shortened course of 21 days may be equally efficacious.

General recommendations for the use of miltefosine are limited by the exclusion of specific groups from the published clinical trials: persons <12 or >65 years of age, persons with the most advanced disease, breast-feeding women, HIV-infected patients, and individuals with significant renal or hepatic insufficiency.

Niclosamide†

Niclosamide is active against a wide variety of adult tapeworms but not against tissue cestodes. The drug uncouples oxidative phosphorylation in parasite mitochondria, thereby blocking the uptake of glucose by the intestinal tapeworm and resulting in the parasite's death. Niclosamide rapidly causes spastic paralysis of intestinal cestodes in vitro. Its use is limited by its side effects, the necessarily long duration of therapy, the recommended use of purgatives, and—most important—limited availability (i.e., on a named-patient basis from the manufacturer).

Niclosamide is poorly absorbed. Tablets are given on an empty stomach in the morning after a liquid meal the night before, and this dose is followed by another 1 h later. For treatment of hymenolepiasis, the drug is administered for 7 days. A second course is often prescribed. The scolex and proximal segments of the tapeworms are killed on contact with niclosamide and may be digested in the gut. However, disintegration of the adult tapeworm results in the release of viable ova, which theoretically can result in autoinfection. Although fears of the development of cysticercosis in patients with *Taenia solium* infections have proved unfounded, it is still recommended that a brisk purgative be given 2 h after the first dose.

Nifurtimox*

This nitrofuran compound is an inexpensive and effective oral agent for the treatment of acute Chagas disease. Trypanosomes lack catalase and have very low levels of peroxidase; as a result, they are very vulnerable to by-products of oxygen reduction. When nifurtimox is reduced in the trypanosome, a nitro anion radical is formed and undergoes autooxidation, resulting in the generation of the superoxide anion O_2^-, hydrogen peroxide (H_2O_2), hydroperoxyl radical (HO_2), and other highly reactive and cytotoxic molecules. Despite the abundance of catalases, peroxidases, and superoxide dismutases that neutralize these destructive radicals in mammalian cells, nifurtimox has a poor therapeutic index. Prolonged use is required, but the course may have to be interrupted because of drug toxicity, which develops in 40–70% of recipients. Nifurtimox is well absorbed and undergoes rapid and extensive biotransformation; <0.5% of the original drug is excreted in urine.

Nitazoxanide

Nitazoxanide is a 5-nitrothiazole compound used for the treatment of cryptosporidiosis and giardiasis; it is active against other intestinal protozoa as well. The drug is approved for use in children 1–11 years of age.

The antiprotozoal activity of nitazoxanide is believed to be due to interference with the pyruvate-ferredoxin oxidoreductase (PFOR) enzyme–dependent electron transfer reaction that is essential to anaerobic energy metabolism. Studies have shown that the PFOR enzyme from *G. lamblia* directly reduces nitazoxanide by transfer of electrons in the absence of ferredoxin. The DNA-derived PFOR protein sequence of *Cryptosporidium parvum* appears to be similar to that of *G. lamblia*. Interference with the PFOR enzyme–dependent electron transfer reaction may not be the only pathway by which nitazoxanide exerts antiprotozoal activity.

After oral administration, nitazoxanide is rapidly hydrolyzed to an active metabolite, tizoxanide (desacetyl-nitazoxanide). Tizoxanide then undergoes conjugation, primarily by glucuronidation. It is recommended that nitazoxanide be taken with food; however, no studies have been conducted to determine whether the pharmacokinetics of tizoxanide and tizoxanide glucuronide differ in fasted versus fed subjects. Tizoxanide is excreted in urine, bile, and feces, and tizoxanide glucuronide is excreted in urine and bile. The pharmacokinetics of nitazoxanide in patients with impaired hepatic and/or renal function have not been studied. Tizoxanide is highly bound to plasma protein (>99.9%). Therefore, caution should be used when administering this agent concurrently with other highly plasma protein–bound drugs that have narrow therapeutic indices, as competition for binding sites may occur.

Oxamniquine

This tetrahydroquinoline derivative is an effective alternative agent for the treatment of *Schistosoma mansoni*, although susceptibility to this drug exhibits regional variation. Oxamniquine exhibits anticholinergic properties, but its primary mode of action seems to rely on ATP-dependent enzymatic drug activation generating an intermediate that alkylates essential macromolecules, including DNA. In treated adult schistosomes, oxamniquine produces marked tegumental alterations that are similar to those seen with praziquantel but that develop less rapidly, becoming evident 4–8 days after treatment.

Oxamniquine is administered orally as a single dose and is well absorbed. Food retards absorption and reduces bioavailability. About 70% of an administered dose is excreted in urine as a mixture of pharmacologically inactive metabolites. Patients should be warned that their urine might have an intense orange–red color. Side effects are uncommon and usually mild, although hallucinations and seizures have been reported.

Paromomycin (Aminosidine)

First isolated in 1956, this aminoglycoside is an effective oral agent for the treatment of infections due to intestinal protozoa. Parenteral paromomycin appears to be effective against visceral leishmaniasis in India.

Paromomycin inhibits protozoan protein synthesis by binding to the 30S ribosomal RNA in the aminoacyl-tRNA site, causing misreading of mRNA codons. Paromomycin is less active against *G. lamblia* than standard agents; however, like other aminoglycosides, paromomycin is poorly absorbed from the intestinal lumen, and the high levels of drug in the gut compensate for this relatively weak activity. If absorbed or administered systemically, paromomycin can cause ototoxicity and nephrotoxicity. However, systemic absorption is very limited, and toxicity should not be a concern in persons with normal kidneys. Topical formulations are not generally available.

Pentamidine Isethionate

This diamidine is an effective alternative agent for some forms of leishmaniasis and trypanosomiasis. It is available for parenteral and aerosolized administration. Although its mechanism of action remains undefined, it is known to exert a wide range of effects, including interaction with trypanosomal kinetoplast DNA; interference with polyamine synthesis by a decrease in the activity of ornithine decarboxylase; and inhibition of RNA polymerase, topoisomerase, ribosomal function, and the synthesis of nucleic acids and proteins.

Pentamidine isethionate is well absorbed, highly tissue bound, and excreted slowly over several weeks, with an elimination half-life of 12 days. No steady-state plasma concentration is attained in persons given daily injections; the result is extensive accumulation of pentamidine in tissues, primarily the liver, kidney, adrenal gland, and spleen.

Pentamidine does not penetrate well into the CNS. Pulmonary concentrations of pentamidine are increased when the drug is delivered in aerosolized form, but not when it is delivered systemically.

Rapid (<1-h) infusion of intravenous pentamidine often results in hypotension. Because electrolyte disturbances and mild to moderate nephrotoxicity occur commonly, pentamidine should be used with caution with other nephrotoxic agents. Pancreatitis and QT prolongation may also occur; cumulative damage to pancreatic islet cells may result in drug-induced diabetes mellitus. Similarly, hypoglycemia can develop, although much less commonly when pentamidine is given by the inhaled route.

Piperaquine

This bisquinoline was synthesized in the 1960s and used widely for malaria control in China. The development of artemisinin-based combination therapy led to its evaluation as a partner drug, and it is now combined with dihydroartemisinin. Piperaquine is highly lipophilic and has a prolonged half-life (~20 days), thus providing a period of post-treatment prophylaxis. The drug's mechanisms of action and resistance have not been well studied but are presumed to be similar to those of the other 4-aminoquinolines.

Piperazine

The antihelminthic activity of piperazine is confined to ascariasis and enterobiasis. Piperazine acts as an agonist at extrasynaptic γ-aminobutyric acid (GABA) receptors, causing an influx of chloride ions in the nematode somatic musculature. Although the initial result is hyperpolarization of the muscle fibers, the ultimate effect is flaccid paralysis, leading to the expulsion of live worms. Patients should be warned, as this occurrence can be unsettling.

Praziquantel

This heterocyclic pyrazinoisoquinoline derivative is highly active against a broad spectrum of trematodes and cestodes. It is the mainstay of treatment for schistosomiasis and is a critical part of community-based control programs.

All of the effects of praziquantel can be attributed either directly or indirectly to an alteration of intracellular calcium concentrations. Although the exact mechanism of action remains unclear, the major mechanism is disruption of the parasite tegument, causing tetanic contractures with loss of adherence to host tissues and, ultimately, disintegration or expulsion. Praziquantel induces changes in the antigenicity of the parasite by causing the exposure of concealed antigens. Praziquantel also produces alterations in schistosomal glucose metabolism, including decreases in glucose uptake, lactate release, glycogen content, and ATP levels.

Praziquantel exerts its parasitic effects directly and does not need to be metabolized to be effective. It is well absorbed but undergoes extensive first-pass hepatic clearance. Levels of the drug are increased when it is taken with food, particularly carbohydrates, or with cimetidine. Serum levels are reduced by glucocorticoids, chloroquine, carbamazepine, and phenytoin. Praziquantel is completely metabolized in humans, with 80% of the dose recovered as metabolites in urine within 4 days. It is not known to what extent praziquantel crosses the placenta, but retrospective studies suggest that it is safe in pregnancy.

Patients with schistosomiasis who have heavy parasite burdens may develop abdominal discomfort, nausea, headache, dizziness, and drowsiness. Symptoms begin 30 min after ingestion, may require spasmolytics for relief, and usually disappear spontaneously after a few hours.

Primaquine Phosphate

Primaquine, an 8-aminoquinoline, has a broad spectrum of activity against all stages of plasmodial development in humans but has been used most effectively for eradication of the hepatic stage of these parasites. Despite its toxicity, it remains the drug of choice for radical cure of *P. vivax* infections. Primaquine must be metabolized by the host to be effective. It is, in fact, rapidly metabolized; only a small fraction of the dose of the parent drug is excreted unchanged. Although the parasiticidal activity of the three oxidative metabolites remains unclear, they are believed to affect both pyrimidine synthesis and the mitochondrial electron transport chain. The metabolites appear to have significantly less antimalarial activity than primaquine; however, their hemolytic activity is greater than that of the parent drug.

Primaquine causes marked hypotension after parenteral administration and therefore is given only by the oral route. It is rapidly and almost completely absorbed from the GI tract.

Patients should be tested for G6PD deficiency before they receive primaquine. The drug may induce the oxidation of hemoglobin into methemoglobin, regardless of the G6PD status of the patient. Primaquine is otherwise well tolerated.

Proguanil (Chloroguanide)

Proguanil inhibits plasmodial dihydrofolate reductase and is used with atovaquone for oral treatment of uncomplicated malaria or with chloroquine for malaria prophylaxis in parts of Africa without widespread chloroquine-resistant *P. falciparum*.

Proguanil exerts its effect primarily by means of the metabolite cycloguanil, whose inhibition of dihydrofolate reductase in the parasite disrupts deoxythymidylate synthesis, thus interfering with a key pathway involved in the biosynthesis of pyrimidines required for nucleic acid replication. There are no clinical data indicating that folate supplementation diminishes drug efficacy; women of childbearing age for whom atovaquone/proguanil is prescribed should continue taking folate supplements to prevent neural tube birth defects.

Proguanil is extensively absorbed regardless of food intake. The drug is 75% protein bound. The main routes of elimination are hepatic biotransformation and renal excretion; 40–60% of the proguanil dose is excreted by the kidneys. Drug levels are increased and elimination is impaired in patients with hepatic insufficiency.

Pyrantel Pamoate

Pyrantel is a tetrahydropyrimidine formulated as pamoate. This safe, well-tolerated, inexpensive drug is used to treat a variety of intestinal nematode infections but is ineffective in trichuriasis. Pyrantel pamoate is usually effective in a single dose. Its target is the nicotinic acetylcholine receptor on the surface of nematode somatic muscle. Pyrantel depolarizes the neuromuscular junction of the nematode, resulting in its irreversible paralysis and allowing the natural expulsion of the worm.

Pyrantel pamoate is poorly absorbed from the intestine; >85% of the dose is passed unaltered in feces. The absorbed portion is metabolized and excreted in urine. Piperazine is antagonistic to pyrantel pamoate and should not be used concomitantly.

Pyrantel pamoate has minimal toxicity at the oral doses used to treat intestinal helminthic infection. It is not recommended for pregnant women or for children <12 months old.

Pyrimethamine

When combined with short-acting sulfonamides, this diaminopyrimidine is effective in malaria, toxoplasmosis, and isosporiasis. Unlike mammalian cells, the parasites that cause these infections cannot use preformed pyrimidines obtained through salvage pathways but rather rely completely on de novo pyrimidine synthesis, for which folate derivatives are essential cofactors. The efficacy of pyrimethamine is increasingly limited by the development of resistant strains of *P. falciparum* and *P. vivax*. Single amino acid substitutions to parasite dihydrofolate reductase confer resistance to pyrimethamine by decreasing the enzyme’s binding affinity for the drug.

Pyrimethamine is well absorbed; the drug is 87% bound to human plasma proteins. In healthy volunteers, drug concentrations remain at therapeutic levels for up to 2 weeks; drug levels are lower in patients with malaria.

At the usual dosage, pyrimethamine alone causes little toxicity except for occasional skin rashes and, more rarely, blood dyscrasias. Bone marrow suppression sometimes occurs at the higher doses used for toxoplasmosis; at these doses, the drug should be administered with folinic acid.

Pyronaridine

This potent antimalarial is a benzonaphthyridine derivative first synthesized by Chinese researchers in 1970. Like chloroquine, pyronaridine targets hematin formation, inhibiting the production of β-hematin by forming complexes with it, with consequent enhancement of hematin-induced hemolysis. However, this drug is more potent than chloroquine: for complete lysis, pyronaridine is required at only 1/100th of the concentration needed with

chloroquine. It also inhibits glutathione-dependent heme degradation. Despite its similar mode of action, pyronaridine remains effective against chloroquine-resistant strains. When combined with artesunate, it is effective for the treatment of acute, uncomplicated infection caused by *P. falciparum* or *P. vivax* in areas of low transmission with evidence of artemisinin resistance.

Pyronaridine is readily absorbed, widely distributed throughout the body, metabolized by the liver, and excreted in urine and stool. Its use is contraindicated in patients with severe liver or kidney impairment. Pyronaridine inhibits both CYP2D6 and P-glycoprotein in vitro, and these effects may have clinical relevance for patients taking medications for cardiac disease (e.g., metoprolol and digoxin).

Quinacrine*

Quinacrine is the only drug approved by the FDA for the treatment of giardiasis. Although its production was discontinued in 1992, quinacrine can be obtained from alternative sources through the CDC Drug Service. The antiprotozoal mechanism of quinacrine has not been fully elucidated. The drug inhibits NADH oxidase—the same enzyme that activates furazolidone. The differing relative quinacrine uptake rate between human cells and *G. lamblia* may explain the selective toxicity of the drug. Resistance correlates with decreased drug uptake.

Quinacrine is rapidly absorbed from the intestinal tract and is widely distributed in body tissues. Alcohol is best avoided because of a disulfiram-like effect.

Quinine and Quinidine

When combined with another agent, the cinchona alkaloid quinine is effective for the oral treatment of both uncomplicated, chloroquine-resistant malaria and babesiosis. Quinine acts rapidly against the asexual blood stages of all forms of the human malaria parasites. For severe malaria, only quinidine (the dextroisomer of quinine) is available in the United States. Quinine concentrates in the acidic food vacuoles of *Plasmodium* species. The drug inhibits the nonenzymatic polymerization of the highly reactive, toxic heme molecule into the nontoxic polymer pigment hemozoin.

Quinine is readily absorbed when given orally. In patients with malaria, the elimination half-life of quinine increases according to the severity of the infection. However, toxicity is avoided by an increase in the concentration of plasma glycoproteins. The cinchona alkaloids are extensively metabolized, particularly by CYP3A4; only 20% of the dose is excreted unchanged in urine. The drug's metabolites are also excreted in urine and may be responsible for toxicity in patients with renal failure. Renal excretion of quinine is decreased when cimetidine is taken and increased when the urine is acidic. The drug readily crosses the placenta.

Quinidine is both more potent as an antimalarial and more toxic than quinine. Its use requires cardiac monitoring. Dose reduction is necessary in persons with severe renal impairment.

Spiramycin†

This macrolide is used to treat acute toxoplasmosis in pregnancy and congenital toxoplasmosis. While the mechanism of action is similar to that of other macrolides, the efficacy of spiramycin in toxoplasmosis appears to stem from its rapid and extensive intracellular penetration, which results in macrophage drug concentrations 10–20 times greater than serum concentrations.

Spiramycin is rapidly and widely distributed throughout the body and reaches concentrations in the placenta up to five times those in serum. This agent is excreted mainly in bile. Indeed, in humans, the urinary excretion of active compounds represents only 20% of the administered dose.

Serious reactions to spiramycin are rare. Of the available macrolides, spiramycin appears to have the lowest risk of drug interactions. Complications of treatment are rare but, in neonates, can include life-threatening ventricular arrhythmias that disappear with drug discontinuation.

Sulfonamides

See Table 217-1 and **Chap. 139**.

Suramin*

This derivative of urea is the drug of choice for the early stage of African trypanosomiasis. The drug is polyanionic and acts by forming stable complexes with proteins, thus inhibiting multiple enzymes essential to parasite energy metabolism. Suramin appears to inhibit all trypanosome glycolytic enzymes more effectively than it inhibits the corresponding host enzymes.

Suramin is parenterally administered. It binds to plasma proteins and persists at low levels for several weeks after infusion. Its metabolism is negligible. This drug does not penetrate the CNS.

Tafenoquine

Tafenoquine is an 8-aminoquinoline with causal prophylactic activity. Its prolonged half-life (2–3 weeks) allows longer dosing intervals when the drug is used for prophylaxis. Tafenoquine has been well tolerated in clinical trials. When tafenoquine is taken with food, its absorption is increased by 50% and the most commonly reported adverse event—mild GI upset—is diminished. Like primaquine, tafenoquine is a potent oxidizing agent, causing hemolysis in patients with G6PD deficiency as well as methemoglobinemia.

Tetracyclines

See Table 217-1 and **Chap. 139**.

Thiabendazole

Discovered in 1961, thiabendazole remains one of the most potent of the numerous benzimidazole derivatives. However, its use has declined significantly because of a higher frequency of adverse effects than is seen with other, equally effective agents.

Thiabendazole is active against most intestinal nematodes that infect humans. Although the exact mechanism of its antihelminthic activity has not been fully elucidated, it is likely to be similar to that of other benzimidazole drugs: namely, inhibition of polymerization of parasite β-tubulin. The drug also inhibits the helminth-specific enzyme fumarate reductase. In animals, thiabendazole has anti-inflammatory, antipyretic, and analgesic effects, which may explain its usefulness in dracunculiasis and trichinellosis. Thiabendazole also suppresses egg and/or larval production by some nematodes and may inhibit the subsequent development of eggs or larvae passed in feces. Despite the emergence and global spread of thiabendazole-resistant trichostrongyliasis among sheep, there have been no reports of drug resistance in humans.

Thiabendazole is available in tablet form and as an oral suspension. The drug is rapidly absorbed from the GI tract but can also be absorbed through the skin. Thiabendazole should be taken after meals. This agent is extensively metabolized in the liver before ultimately being excreted; most of the dose is excreted within the first 24 h. The usual dose of thiabendazole is determined by the patient's weight, but some treatment regimens are parasite specific. No particular adjustments are recommended in patients with renal or hepatic failure; only cautious use is advised.

Coadministration of thiabendazole to patients taking theophylline can result in an increase in theophylline levels by >50%. Therefore, serum levels of theophylline should be monitored closely in this situation.

Tinidazole

This nitroimidazole is effective for the treatment of amebiasis, giardiasis, and trichomoniasis. Like metronidazole, tinidazole must undergo reductive activation by the parasite's metabolic system before it can act on protozoal targets. Tinidazole inhibits the synthesis of new DNA in the parasite and causes degradation of existing DNA. The reduced free-radical derivatives alkylate DNA, with consequent cytotoxic damage to the parasite. This damage appears to be produced by short-lived reduction intermediates, resulting in helix destabilization and strain breakage of DNA. The mechanism of action and side effects of tinidazole are similar to those of metronidazole, but adverse events appear to be less frequent and severe with tinidazole. In addition, the significantly longer half-life of tinidazole (>12 h) offers potential cure with a single dose.

Triclabendazole

While most benzimidazoles have broad-spectrum antihelminthic activity, they exhibit minimal or no activity against *Fasciola hepatica*. In contrast, the antihelminthic activity of triclabendazole is highly specific for *Fasciola* and *Paragonimus* species, with little activity against nematodes, cestodes, and other trematodes. Triclabendazole is effective against all stages of *Fasciola* species. The active sulfoxide metabolite of triclabendazole binds to fluke tubulin by assuming a unique nonplanar configuration and disrupts microtubule-based processes. Resistance to triclabendazole in veterinary use has been reported in Australia and Europe; however, no resistance has been documented in humans.

Triclabendazole is rapidly absorbed after oral ingestion; administration with food enhances its absorption and shortens the elimination half-life of the active metabolite. Both the sulfoxide and the sulfone metabolites are highly protein bound (>99%). Treatment with triclabendazole is typically given in one or two doses. No clinical data are available regarding dose adjustment in renal or hepatic insufficiency; however, given the short course of therapy and extensive hepatic metabolism of triclabendazole, dose adjustment is unlikely to be necessary. No information exists on drug interactions.

Trimethoprim-Sulfamethoxazole

See Table 217-1 and **Chap. 139**.

FURTHER READING

Fehintola FA et al: Drug interactions in the treatment and chemoprophylaxis of malaria in HIV infected individuals in sub Saharan Africa. Curr Drug Metab 12:51, 2011. [PubMed: 21222586]
Keiser J et al: Antiparasitic drugs for paediatrics: Systematic review, formulations, pharmacokinetics, safety, efficacy and implications for control. Parasitology 138:1620, 2011. [PubMed: 21349223]
Kelesidis T, Falagas ME: Substandard/counterfeit antimicrobial drugs. Clin Microbiol Rev 28:443, 2015. [PubMed: 25788516]
Pink R et al: Opportunities and challenges in antiparasitic drug discovery. Nat Rev Drug Discov 4:727, 2005. [PubMed: 16138106]

Chapter 218: Amebiasis and Infection with Free-Living Amebae

Rosa M. Andrade; Sharon L. Reed

AMEBIASIS

DEFINITION

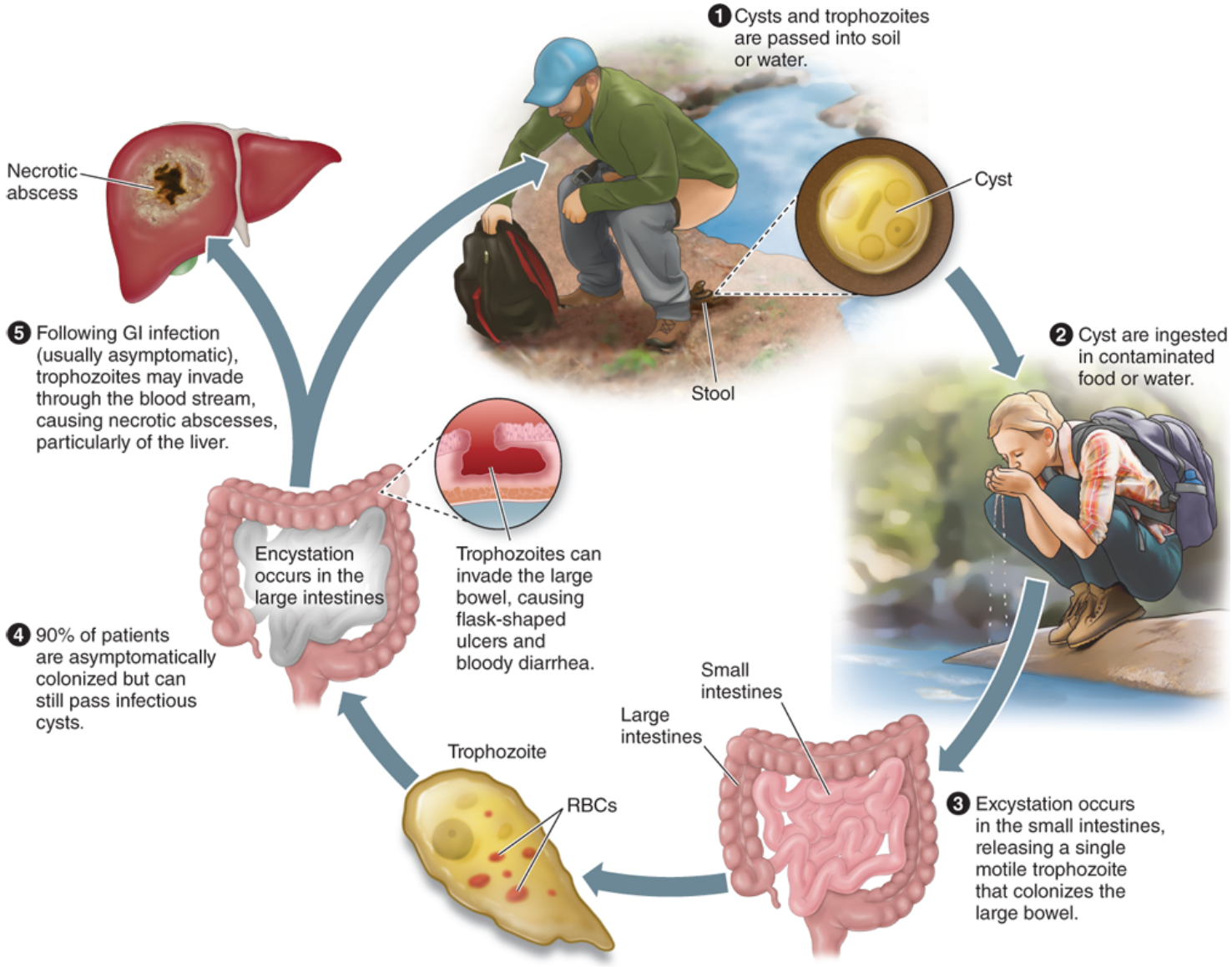
Amebiasis is an infection caused by *Entamoeba histolytica*, an intestinal protozoan. Its spectrum of clinical syndromes ranges from asymptomatic colonization (90% of cases) to invasive amebiasis, which accounts for 10% of affected individuals. Invasive amebiasis frequently presents as intestinal colitis (dysentery or diarrhea) or as extraintestinal amebiasis, in which abscesses of the liver are more commonly found than involvement of the lungs or brain.

LIFE CYCLE AND TRANSMISSION

E. histolytica is acquired by ingestion of viable cysts from fecally contaminated water, food, or hands **(Fig. 218-1)**. Food-borne exposure is most prevalent and is particularly likely when food handlers are shedding cysts or food is being grown with feces-contaminated soil, fertilizer, or water. Besides the drinking of contaminated water, less common means of transmission include oral and anal sexual practices and—in rare instances—direct rectal inoculation through colonic irrigation devices. Motile trophozoites are released from cysts in the small intestine and, in most patients, remain as harmless commensals in the large bowel. After encystation, infectious cysts are shed in the stool and can survive for several weeks in a moist environment. In some patients, the trophozoites invade either the bowel mucosa, causing symptomatic colitis, or the bloodstream, causing distant abscesses of the liver, lungs, or brain. The trophozoites may not encyst in patients with active dysentery, and motile hematophagous trophozoites are frequently present in fresh stools. Trophozoites are rapidly killed by exposure to air or stomach acid, however, and therefore cannot transmit infection.

Figure 218-1

Life cycle of *Entamoeba histolytica*. GI, gastrointestinal; RBCs, red blood cells.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPIDEMIOLOGY

E. histolytica infection typically affects tropical underdeveloped regions with poor sanitation systems and hygiene, occurring particularly often in children <5 years of age. This infection is widespread on the Indian subcontinent and in Africa, parts of East Asia (Thailand), and Central and South America (Mexico and Colombia). According to the Global Burden of Disease 2015 study, amebiasis accounts for 67,900 all-age deaths, including 15,500 children <5 years old.

In contrast, the main groups at risk for amebiasis in developed countries are returned travelers, recent immigrants, men who have sex with men (MSM), military personnel, and inmates of institutions. Data for 1997–2011 from the GeoSentinel Surveillance Network, which encompasses information from tropical medicine clinics on six continents, showed that, among long-term travelers (trip duration, >6 months), diarrhea due to *E. histolytica* was among the most common diagnoses. In fact, amebiasis may be considered an emerging infectious disease in developed countries such as Japan, where the number of reported cases among HIV-positive patients, and particularly among MSM, has increased.

Worldwide, *E. histolytica* is the second most common cause of death related to parasitic infection (after malaria). Invasive colitis and liver abscesses are tenfold more common among men than among women; this difference has been attributed to a disparity in complement-mediated killing and effects of testosterone on the secretion of interferon γ. The wide spectrum of clinical disease caused by *Entamoeba* is due in part to the differences between the two major infecting species, *E. histolytica* and *E. dispar. E. histolytica* has unique surface antigens, is genetically distinct, and possesses virulence properties that distinguish it from the morphologically identical *E. dispar*.

Most asymptomatic carriers, including MSM and patients with AIDS, harbor *E. dispar* and have self-limited infections. In this respect, *E. dispar* is dissimilar to other enteric pathogens such as *Cryptosporidium* and *Cystoisospora belli*, which can cause self-limited illnesses in immunocompetent hosts but devastating diarrhea in patients with AIDS. These observations indicate that *E. dispar* is incapable of causing invasive disease. Through genomic sequencing, new species of *Entamoeba* have been identified: *E. moshkowskii* and *E. bangladeshi*. These new species are microscopically indistinguishable from *E. histolytica*. Although *E. moshkovskii* causes diarrhea, weight loss, and colitis in

mice, a prospective evaluation of children from the Mirpur community of Dhaka, Bangladesh, found that most children who had diarrheal diseases associated with *E. moshkovskii* were simultaneously infected with at least one other enteric pathogen. In 2012, *E. bangladeshi nov. sp., Bangladesh* was first reported in this same Bangladeshi community, having been isolated from the stools of both asymptomatic children and those with diarrhea. Additional clinical and epidemiologic studies are needed to discern the true role of *E. bangladeshi* in the human host.

PATHOGENESIS AND PATHOLOGY

Both trophozoites and cysts are found in the intestinal lumen, but only trophozoites of *E. histolytica* invade tissue. The trophozoite is 20–60 µm in diameter and contains vacuoles and a nucleus with a characteristic central nucleolus. Trophozoites attach to colonic mucus and epithelial cells by Gal/GalNAc adherence lectin and release glycosidases and proteases that cause degradation of mucous polymers. Extracellular cysteine proteinases degrade collagen, elastin, IgA, IgG, and the anaphylatoxins C3a and C5a. After disruption of the mucous layer, trophozoites damage the mucosa by contact-dependent and contact-independent cytotoxicity. The contact-dependent cytotoxicity is attributable to induction of apoptotic cell death; trogocytosis-mediated cell death (ingestion of fragments of living cells); and lysis of inflammatory cells (neutrophils, monocytes, and lymphocytes), colonic cells, and hepatic cells through release of phospholipase A and pore-forming peptides. Contact-independent cytotoxicity follows production of inflammatory mediators, such as prostaglandin E2, by trophozoites, ultimately leading to increased ion permeability of intercellular tight junctions.

E. histolytica trophozoites are constantly exposed to reactive oxygen and nitrogen species arising from their own metabolism and from the host during tissue invasion. The ability to resist reactive oxygen species or reactive nitrogen species such as nitric oxide or *S*-nitrosothiols (e.g., *S*-nitrosoglutathione [GSNO] and *S*-nitrosocysteine [CySNO]) is also a virulence factor. Overexpression of hydrogen peroxide–regulatory motif-binding protein appears to increase *E. histolytica* cytotoxicity. Since *E. histolytica* lacks glutathione and glutathione reductase, it relies on its thioredoxin–thioredoxin reductase system to prevent, regulate, and repair the damage caused by oxidative stress. This antioxidant system is versatile: it has the ability to reduce reactive nitrogen species and use an alternative electron donor, such as nicotinamide adenine dinucleotide. Metronidazole, the current standard of therapy for amebiasis, seems to exert its antiparasitic effect through inhibition of this antioxidant system. Auranofin, a reprofiled drug approved by the U.S. Food and Drug Administration for rheumatoid arthritis, inhibits thioredoxin reductase and displays in vitro and in vivo efficacy against *E. histolytica* and *Giardia intestinalis*. Auranofin is currently undergoing clinical trials against *E. histolytica* and *Giardia* infections in Bangladesh.

Phagocytosis is a virulence factor that leads to defective proliferation of *E. histolytica* if inhibited. Trophozoites use membrane-associated carbohydrate-binding proteins to phagocytose intestinal bacteria, especially gram-negative Enterobacteriaceae, for their nutrients. Interactions with commensal bacteria, such as *Escherichia coli*, can attenuate the virulence of *E. histolytica* by decreasing the expression of Gal/GalNAc lectin. In contrast, ingestion of enteropathogenic bacteria, such as enteropathogenic *E. coli* and *Shigella dysenteriae*, increases expression of the Gal/GalNAc lectin and enhances *E. histolytica* cysteine protease activity.

E. histolytica is capable of altering the commensal gut microbiota. In a cohort in northern India, adult patients who had had amebic dysentery for 5–7 days had significant decreases in intestinal *Bacteroides*, the *Clostridium coccoides* subgroup, the *Clostridium leptum* subgroup, *Lactobacillus*, *Campylobacter*, and *Eubacterium* but displayed increases in *Bifidobacterium*. During the first 2 years of life, the gut immune system and the microbiota mature rapidly. In one study, ~80% of children from the Bangladeshi community of Dhaka were found to be infected with *E. histolytica* by 2 years of age. Fecal anti–Gal/GalNAc lectin IgA was associated with protection from reinfection, while a high parasite burden in the first year of life and expansion of *Prevotella copri* levels were associated with diarrhea.

Antimicrobial peptides, such as cathelicidins, are an important component of innate immunity and are induced by *E. histolytica* upon intestinal invasion in a mouse model. In this model, cecal cathelicidin-related antimicrobial peptide mRNA increased by >4-fold at 3 days and >100-fold at 7 days. However, *E. histolytica* remained resistant to cathelicidin-mediated killing, probably because the antimicrobial peptide was digested by amebic cysteine proteinases.

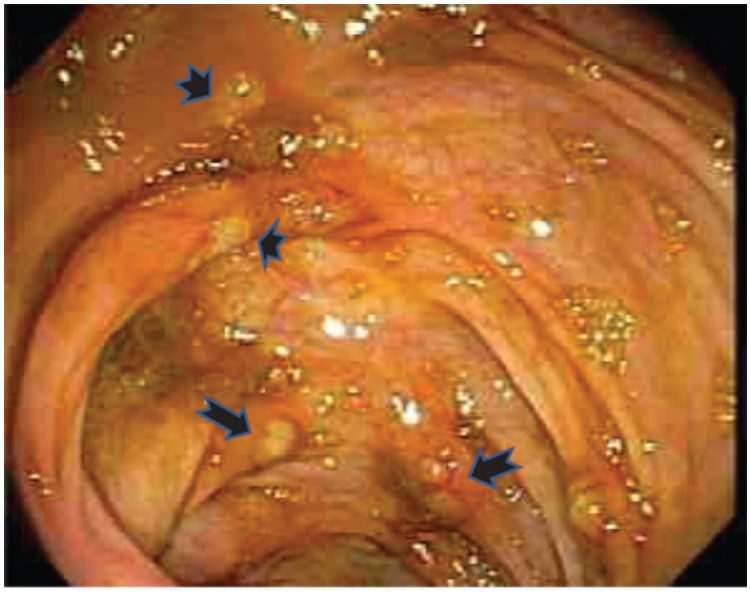
IgA plays a critical role in acquired immunity to *E. histolytica*. A study in Bangladeshi schoolchildren revealed that an intestinal IgA response to Gal/GalNAc reduced the risk of new *E. histolytica* infection by 64%. Serum IgG antibody is not protective; titers correlate with the duration of illness rather than with the severity of disease. Indeed, Bangladeshi children with a serum IgG response were more likely than those without such a response to develop new *E. histolytica* infection. In infants from this same Bangladeshi community, passive immunity conferred by maternal parasite-specific IgA via breastfeeding resulted in a 39% reduced risk of infection and a 64% reduced risk of diarrheal disease from *E. histolytica* during the first year of life. However, this protection appeared to be species-specific, with little or no protection conferred from infections with other species such as *E. dispar* or *E. bangladeshi*.

Genetic susceptibility to amebiasis in humans is associated with a polymorphism in the receptor for the adipocytokine leptin (LEPR). Children in a Bangladeshi cohort carrying arginine at LEPR position 223 (R223) were nearly twice as susceptible to *E. histolytica* as those carrying glutamine at this position (Q223). This mutant allele is overrepresented in many geographic areas with a high prevalence of amebiasis, such as Bangladesh and India. LEPRs are expressed on intestinal epithelial cells, where they prevent apoptosis, promote tissue repair, and may decrease neutrophil infiltration.

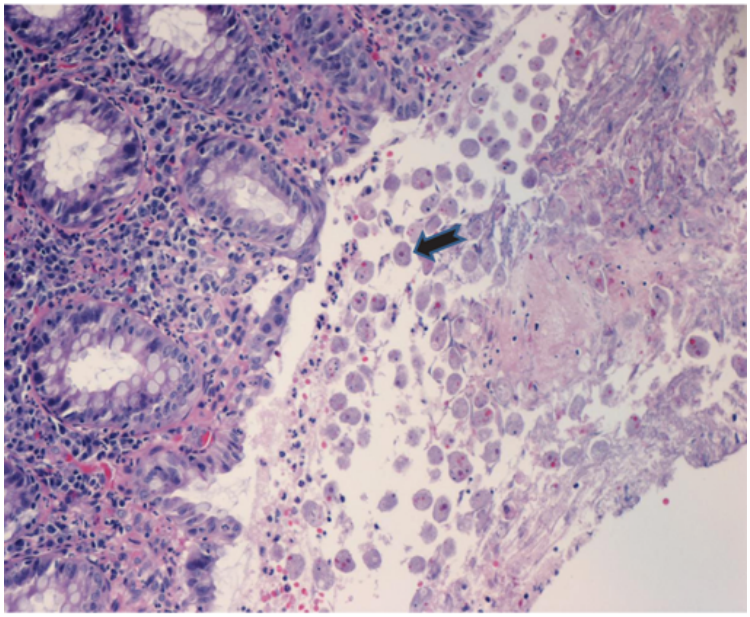
The earliest intestinal lesions are micro-ulcerations of the mucosa of the cecum, sigmoid colon, or rectum that release erythrocytes, inflammatory cells, and epithelial cells. Colonoscopy reveals small ulcers with heaped-up margins and normal intervening mucosa **(Fig. 218-2A)**. Submucosal extension of ulcerations under viable-appearing surface mucosa causes the classic “flask-shaped” ulcer containing trophozoites at the margins of dead and viable tissues. Although neutrophilic infiltrates may accompany early lesions in animals, human intestinal infection is marked by a paucity of inflammatory cells, probably in part because of the killing of neutrophils by trophozoites **(Fig. 218-2B)**. Treated ulcers characteristically heal with little or no scarring. Occasionally, however, full-thickness necrosis and perforation occur.

Figure 218-2

Endoscopic and histopathologic features of intestinal amebiasis. *A.* Appearance of ulcers on colonoscopy (*arrows*). ***B.*** Inflammatory infiltrate and *Entamoeba histolytica* trophozoites (*arrows*) in invasive amebic colitis (hematoxylin and eosin). (*Courtesy of the Department of Pathology and Gastroenterology, San Diego VA Medical Center.)*



A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Rarely, intestinal infection results in the formation of a mass lesion, or *ameboma*, in the bowel lumen. The overlying mucosa is usually thin and ulcerated, while other layers of the wall are thickened, edematous, and hemorrhagic; this condition results in exuberant formation of granulation tissue with little fibrous-tissue response.

Amebic liver abscesses are age- and gender-dependent. Men 30–60 years of age are most commonly infected at a rate 10–12 times higher than women in the same age group. Studies in animal models have demonstrated that testosterone may increase susceptibility to amebic liver abscess by modulating the secretion of interferon γ by natural killer T cells, which are activated through *E. histolytica* lipopeptidophosphoglycan present on the surface of ameba trophozoites. Liver abscesses are always preceded by intestinal colonization, which may be asymptomatic. Blood vessels may be compromised early by wall lysis and thrombus formation. Trophozoites invade veins to reach the liver through the portal venous system. *E. histolytica* is resistant to complement-mediated lysis—a property critical to survival in the bloodstream. Inoculation of amebae into the portal system of hamsters results in an acute cellular infiltrate consisting predominantly of neutrophils. Later, the neutrophils are lysed by contact with amebae, and the release of neutrophil toxins may contribute to necrosis of hepatocytes. The liver parenchyma is replaced by necrotic material that is surrounded by a thin rim of congested liver tissue. Although the necrotic contents of a liver abscess are classically described as “anchovy paste,” the fluid is variable in color; it is composed of bacteriologically sterile granular debris with few or no cells. Amebae, if seen, tend to be found near the capsule of the abscess.

CLINICAL SYNDROMES

Intestinal Amebiasis

The most common type of amebic infection is asymptomatic cyst passage. Even in highly endemic areas, most patients harbor *E. dispar*.

Symptomatic amebic colitis develops 2–6 weeks after the ingestion of infectious cysts. A gradual onset of lower abdominal pain and mild diarrhea is followed by malaise, weight loss, and diffuse lower abdominal or back pain. Cecal involvement may mimic acute appendicitis. Patients with full-blown dysentery may pass 10–12 stools per day. The stools contain little fecal material and consist mainly of blood and mucus. In contrast to those with bacterial diarrhea, fewer than 40% of patients with amebic dysentery are febrile. Virtually all patients have heme-positive stools.

More fulminant intestinal infection, with severe abdominal pain, high fever, and profuse diarrhea, is rare and occurs predominantly in children. Patients may develop toxic megacolon, in which there is severe bowel dilation with intramural air. Patients receiving glucocorticoids are at risk for severe amebiasis. The association between severe amebiasis complications and glucocorticoid therapy emphasizes the importance of excluding amebiasis when inflammatory bowel disease is suspected. An occasional patient presents with only an asymptomatic or tender abdominal mass caused by an ameboma, which is easily confused with cancer on barium studies. A positive serologic test or biopsy can prevent unnecessary surgery in this setting.

Environmental enteropathy (“impoverished gut”; blunted small-intestinal villi with lamina propria inflammation) is observed in tropical developing areas with endemic enteric infections, such as amebiasis. It is associated with functional gastrointestinal impairment causing malnutrition and stunted growth in children within the first 2 years of life. Bangladeshi children with symptomatic *E. histolytica* infections were 2.9 times more likely to be malnourished and 4.7 times more likely to be short for their age than were children without symptomatic infections. These factors affect their cognitive development and may be linked to loss of productivity in adulthood.

Amebic Liver Abscess

Extraintestinal infection by *E. histolytica* most often involves the liver. Of travelers who develop an amebic liver abscess after leaving an endemic area, 95% do so within 5 months. Young patients with an amebic liver abscess are more likely than older patients to present in the acute phase with prominent symptoms of <10 days’ duration. Most patients are febrile and have right-upper-quadrant pain, which may be dull or pleuritic in nature and may radiate to the shoulder. Point tenderness over the liver and right-sided pleural effusion are common. Jaundice is rare. Although the initial site of infection is the colon, fewer than one-third of patients with an amebic abscess have active diarrhea. Older patients from endemic areas are more likely to have a subacute course lasting 6 months, with weight loss and hepatomegaly. About one-third of patients with chronic presentations are febrile. Thus, the clinical diagnosis of an amebic liver abscess may be difficult to establish because the symptoms and signs are often nonspecific. Since 10–15% of patients present only with fever, amebic liver abscess must be considered in the differential diagnosis of fever of unknown origin **(Chap. 17)**.

Complications of Amebic Liver Abscess

Pleuropulmonary involvement, which is reported in 20–30% of patients, is the most frequent complication of amebic liver abscess. Manifestations include sterile effusions, contiguous spread from the liver, and rupture into the pleural space. Sterile effusions and contiguous spread usually resolve with medical therapy, but frank rupture into the pleural space requires drainage. A hepatobronchial fistula may cause cough productive of large amounts of necrotic material that may contain amebae. This dramatic complication carries a good prognosis. Abscesses that rupture into the peritoneum may present as an indolent leak or an acute abdomen and require both percutaneous catheter drainage and medical therapy. Rupture into the pericardium, usually from abscesses of the left lobe of the liver, carries the gravest prognosis; it can occur during medical therapy and requires surgical drainage.

Involvement of Other Extraintestinal Sites

The genitourinary tract may become involved by direct extension of amebiasis from the colon or by hematogenous spread of the infection. Painful genital ulcers, characterized by a punched-out appearance and profuse discharge, may develop secondary to extension from either the intestine or the liver. Both of these conditions respond well to medical therapy. Cerebral involvement has been reported in fewer than 0.1% of patients in large clinical series. Symptoms and prognosis depend on the size and location of the lesion.

DIAGNOSTIC TESTS

Laboratory Diagnosis

Stool examinations, serologic tests, and noninvasive imaging of the liver are the most important procedures in the diagnosis of amebiasis. Fecal findings suggestive of amebic colitis include a positive test for heme, a paucity of neutrophils, and amebic cysts or trophozoites. The definitive diagnosis of amebic colitis is made by the demonstration of hematophagous trophozoites of *E. histolytica*. Because trophozoites are killed rapidly by water, drying, or barium, it is important to examine at least three fresh stool specimens. Examination of a combination of wet mounts, iodine-stained concentrates, and trichrome-stained preparations of fresh stool and concentrates for cysts or trophozoites confirms the diagnosis in 75–95% of cases. Cultures of amebae are more sensitive but are not routinely available. If stool examinations are negative, sigmoidoscopy with biopsy of the edge of ulcers may increase the yield, but this procedure is dangerous during fulminant colitis because of the risk of perforation. Trophozoites in a biopsy specimen from a colonic mass confirm the diagnosis of ameboma, but trophozoites are rare in liver aspirates because they are found in the abscess capsule and not in the readily aspirated necrotic center. Accurate diagnosis requires experience, since the trophozoites may be confused with neutrophils and the cysts must be differentiated morphologically from those of *Entamoeba hartmanni*, *Entamoeba coli*, and *Endolimax nana*, which do not cause clinical disease and do not warrant therapy. Unfortunately, the cysts of *E. histolytica* cannot be distinguished microscopically from those of *E. dispar*, *E. moshkovskii*, or *E. bangladeshi*. Therefore, the microscopic diagnosis of *E. histolytica* can be made only by the detection of *Entamoeba* trophozoites that have ingested erythrocytes. More sensitive and specific tests in stool include enzyme immunoassay detection of the Gal/GalNAc lectin of *E. histolytica* and new multiplex polymerase chain reaction (PCR) stool panels that include *E. histolytica*.

Serology is an important addition to the methods used for parasitologic diagnosis of invasive amebiasis. Enzyme-linked immunosorbent assays and agar gel diffusion assays are positive in >90% of cases with colitis, ameboma, or liver abscess. Positive results in conjunction with the appropriate clinical syndrome suggest active disease because serologic findings usually revert to negative within 6–12 months. Even

in highly endemic areas such as South Africa, fewer than 10% of asymptomatic individuals have a positive amebic serology. The interpretation of the indirect hemagglutination test is difficult because titers may remain positive for as long as 10 years.

Up to 10% of patients with acute amebic liver abscess may have negative serologic findings; in suspected cases with an initially negative result, testing should be repeated in a week. In contrast to carriers of *E. dispar*, most asymptomatic carriers of *E. histolytica* develop antibodies. Thus, serologic tests are helpful in assessing the risk of invasive amebiasis in asymptomatic, cyst-passing individuals in nonendemic areas. Serologic tests also should be performed in patients with ulcerative colitis before the institution of glucocorticoid therapy to prevent the development of severe colitis or toxic megacolon owing to unsuspected amebiasis. Recently, a loop-mediated isothermal amplification (LAMP) assay was shown to be a potential alternative for direct detection of *E. histolytica* DNA in pus samples from amebic liver abscesses. LAMP is a relatively simple, rapid, and low-cost method of DNA amplification that could be a better alternative for diagnosis in developing countries. Routine hematology and chemistry tests usually are not very helpful in the diagnosis of invasive amebiasis. About three-fourths of patients with an amebic liver abscess have leukocytosis (>10,000 cells/μL); this condition is particularly likely if symptoms are acute or complications have developed. Invasive amebiasis does not elicit eosinophilia. Anemia, if present, is usually multifactorial. Even with large liver abscesses, liver enzyme levels are normal or minimally elevated. The alkaline phosphatase level is most often elevated and may remain so for months. Aminotransferase elevations suggest acute disease or a complication.

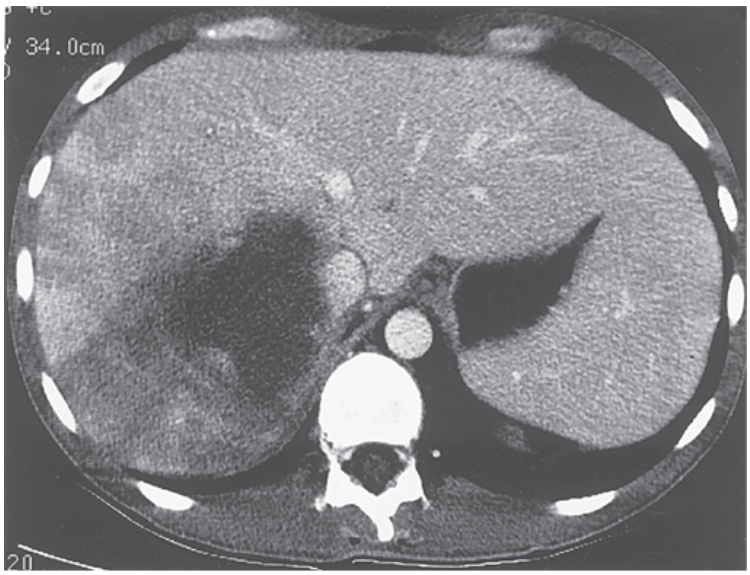
Radiographic Studies

Radiographic barium studies are potentially dangerous in acute amebic colitis. Amebomas are usually identified first by a barium enema, but biopsy is necessary for differentiation from carcinoma.

Radiographic techniques such as ultrasonography, CT, and MRI are all useful for detection of the round or oval hypoechoic cyst. More than 80% of patients who have had symptoms for >10 days have a single abscess of the right lobe of the liver **(Fig. 218-3)**. Approximately 50% of patients who have had symptoms for <10 days have multiple abscesses. Findings associated with complications include large abscesses (>10 cm) in the superior part of the right lobe, which may rupture into the pleural space; multiple lesions, which must be differentiated from pyogenic abscesses; and lesions of the left lobe, which may rupture into the pericardium. Because abscesses resolve slowly and may increase in size despite a clinical response to therapy, frequent follow-up ultrasonography may prove confusing. Complete resolution of a liver abscess within 6 months can be anticipated in two-thirds of patients, but 10% may have persistent abnormalities for a year.

Figure 218-3

Abdominal CT scan of a large amebic abscess of the right lobe of the liver. *(Courtesy of the Department of Radiology, UCSD Medical Center, San Diego; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Differential Diagnosis

The differential diagnosis of intestinal amebiasis includes bacterial diarrheas **(Chap. 128)** caused by *Campylobacter* **(Chap. 162)**; enteroinvasive *Escherichia coli* **(Chap. 156)**; and species of *Shigella* **(Chap. 161)**, *Salmonella* **(Chap. 160)**, and *Vibrio* **(Chap. 163)**. Because the typical patient with amebic colitis has less prominent fever than in these other conditions as well as heme-positive stools with few neutrophils, correct diagnosis requires bacterial cultures, microscopic examination of stools, and amebic serologic testing. As has been mentioned, amebiasis must be ruled out in any patient thought to have inflammatory bowel disease.

Because of the variety of presenting signs and symptoms, amebic liver abscess can easily be confused with pulmonary or gallbladder disease or with any febrile illness with few localizing signs, such as malaria **(Chap. 219)** or typhoid fever **(Chap. 160)**. The diagnosis should be considered in members of high-risk groups who have recently traveled outside the United States **(Chap. 119)** and in inmates of institutions. Once radiographic studies have identified an abscess in the liver, the most important differential diagnosis is between amebic and pyogenic abscess. Patients with pyogenic abscess typically are older and have a history of underlying bowel disease or recent surgery.

Amebic serology is helpful, but aspiration of the abscess, with Gram's staining and culture of the material, may be required for differentiation of the two diseases.

TREATMENT

TREATMENT Amebiasis Intestinal Disease

The drugs used to treat amebiasis can be classified according to their primary site of action **(Table 218-1)**. Luminal amebicides are poorly absorbed and reach high concentrations in the bowel, but their activity is limited to cysts and trophozoites close to the mucosa. Only two luminal drugs are available in the United States: iodoquinol and paromomycin. Indications for the use of luminal agents include eradication of cysts in patients with colitis or a liver abscess and treatment of asymptomatic carriers. The majority of asymptomatic individuals who pass cysts are colonized with *E. dispar*, which does not warrant specific therapy. However, it is prudent to treat asymptomatic individuals who pass cysts unless *E. dispar* colonization can be definitively demonstrated by specific antigen-detection tests.

Tissue amebicides reach high concentrations in the blood and tissue after oral or parenteral administration. The development of nitroimidazole compounds, especially metronidazole, was a major advance in the treatment of invasive amebiasis. Patients with amebic colitis should be treated with IV or oral metronidazole. Side effects include nausea, vomiting, abdominal discomfort, and a disulfiram-like reaction. Another, longer-acting imidazole compound, tinidazole, is likewise effective and is available in the United States. All patients should also receive a full course of therapy with a luminal agent, since metronidazole does not eradicate cysts. Resistance to metronidazole has been selected in the laboratory but has not been found in clinical isolates. Relapses are not uncommon and probably represent reinfection or failure to eradicate amebae from the bowel because of an inadequate dosage or duration of therapy.

Amebic Liver Abscess

Metronidazole is the drug of choice for amebic liver abscess. Longer-acting nitroimidazoles (tinidazole and ornidazole) have been effective as single-dose therapy in developing countries. With early diagnosis and therapy, mortality rates from uncomplicated amebic liver abscess are <1%. There is no evidence that combined therapy with two drugs is more effective than the single-drug regimen. Studies of South Africans with liver abscesses demonstrated that 72% of patients without intestinal symptoms had bowel infection with *E. histolytica*; thus, all treatment regimens should include a luminal agent to eradicate cysts and prevent further transmission. Amebic liver abscess recurs rarely.

More than 90% of patients respond dramatically to metronidazole therapy with decreases in both pain and fever within 72 h. Indications for aspiration of liver abscesses are (1) the need to rule out a pyogenic abscess, particularly in patients with multiple lesions; (2) the lack of a clinical response in 3–5 days; (3) the threat of imminent rupture; and (4) the need to prevent rupture of left-lobe abscesses into the pericardium. There is no evidence that aspiration, even of large abscesses (up to 10 cm), accelerates healing. Percutaneous drainage may be successful even if the liver abscess has already ruptured. Surgery should be reserved for instances of bowel perforation and rupture into the pericardium.

TABLE 218-1 Drug Therapy for Amebiasis

Indication	Therapy
Asymptomatic carriage	Luminal agent: iodoquinol (650-mg tablets), 650 mg tid for 20 days; *or* paromomycin (250-mg tablets), 500 mg tid for 10 days
	Metronidazole (250- or 500-mg tablets), 750 mg PO or IV tid for 5–10 days; *or* tinidazole, 2 g/d PO for 3 days
Acute colitis	***plus***
	Luminal agent as above
	Metronidazole, 750 mg PO or IV for 5–10 days; *or* tinidazole, 2 g PO once; *or* ornidazole,[a] 2 g PO once
Amebic liver abscess	***plus***
	Luminal agent as above

[a]Not available in the United States.

PREVENTION

Amebic infection is spread by ingestion of food or water contaminated with cysts. Since an asymptomatic carrier may excrete up to 15 million cysts per day, prevention of infection requires adequate sanitation and eradication of cyst carriage. In high-risk areas, infection can be minimized by the avoidance of unpeeled fruits and vegetables and the use of bottled water. Because cysts are resistant to readily attainable levels of chlorine, disinfection by iodination (tetraglycine hydroperiodide) is recommended. There is no effective prophylaxis.

INFECTION WITH FREE-LIVING AMEBAE

EPIDEMIOLOGY

There are multiple genera of free-living amebae, but the major human pathogens are *Acanthamoeba*, *Naegleria*, and *Balamuthia*. All of these parasites can cause serious central nervous system (CNS) infections, which are almost always fatal. *Acanthamoeba* and *Naegleria* are distributed throughout the world and have been isolated from a wide variety of fresh and brackish water, including water from taps, lakes, hot springs, swimming pools, heating and air-conditioning units, and hospital water networks, and even from the nasal passages of healthy children. Encystation may protect these protozoa from desiccation and food deprivation. The persistence of *Legionella pneumophila* in water supplies is attributable in part to chronic infection of free-living amebae, particularly *Acanthamoeba*. Recent in vitro studies have suggested that a number of pathogens that can resist phagosome-mediated killing may be able to survive within water systems in free-living amebae. These include *Pseudomonas aeruginosa*, mycobacteria (both slow-growing species—e.g., those in the *Mycobacterium avium* complex, *M. kansasii*, and *M. gordonae*—and rapid-growing species—e.g., *M. chelonae and M. abscessus*), and viruses such as adenoviruses and echoviruses.

In contrast, the environmental niche of free-living amebae of the genus *Balamuthia* appears to be soil. A soil sample from a flowerpot was linked to a fatal infection in a child. Cases have been reported from all continents except Africa, but the majority of cases are from warm, dry areas of the southwestern United States and Latin America.

With better recognition of these pathogens, additional risk factors have been identified. Since 2010, five cases of *Naegleria fowleri* infection have been reported in northern U.S. states and have been associated with exposure to piped water, which represents a new ecologic niche. Since 2009, three clusters of *Balamuthia mandrillaris* infections have been associated with organ transplantation. *Acanthamoeba* species have caused large outbreaks of microbial keratitis associated with contact lens wear.

NAEGLERIA INFECTIONS

Primary amebic meningoencephalitis (PAM) is a fulminant CNS infection caused by the free-living ameba *N. fowleri*, which thrives in warm freshwater of lakes and rivers. In the United States, 138 cases of PAM were reported from 1962 through 2015. Although the number of infections reported annually has remained stable (0–8), recent changes in the epidemiology of PAM are a cause of concern. In 2010–2015, 24 cases of PAM were reported and confirmed by the Centers for Disease Control and Prevention (CDC). In 2010, a PAM case was reported for the first time from the northern state of Minnesota; this case was followed by additional cases from Minnesota, Indiana, and Kansas in 2011 and 2012. With climate change, other areas may be at risk because of higher temperatures. The remaining cases were reported mostly from southern states. Sixty-three percent of cases affected female patients, and the median age of patients was 11 years (range, 4–56 years). The majority of patients (19, or 79%) were exposed to recreational freshwater from lakes, reservoirs, rivers, streams, or ditches. The remaining five cases (21%) were due to tap-water exposure through nasal irrigation with a neti pot, playing on a backyard waterslide, and swimming in a poorly maintained pool.

PAM follows the aspiration of water contaminated with trophozoites or cysts or the inhalation of contaminated dust leading to invasion of the olfactory neuroepithelium. Infection is most common in otherwise healthy children or young adults, who often report recent swimming in lakes or heated swimming pools. In rare instances, cases occur when contaminated water is used for nasal irrigation. After an incubation period of 2–15 days, severe headache, high fever, nausea, vomiting, and meningismus develop. Photophobia and palsies of the third, fourth, and sixth cranial nerves are common. Rapid progression to seizures and coma may follow. The prognosis is uniformly poor: most patients die within a week.

The diagnosis of *Naegleria* infection should be considered in any patient who has purulent meningitis without evidence of bacteria on Gram's staining, antigen detection assay, and culture. Other laboratory findings resemble those for fulminant bacterial meningitis, with elevated intracranial pressure, high white blood cell counts (up to 20,000/μL), and elevated protein concentrations and low glucose levels in cerebrospinal fluid (CSF). Diagnosis depends on the detection of motile trophozoites in wet mounts of fresh spinal fluid. Antibodies to *Naegleria* species have been detected in healthy adults; thus serologic testing is not useful in the diagnosis of acute infection. Diagnostic PCR and histochemical staining of biopsies are available through the CDC.

A number of antimicrobial agents have in vitro activity against *N. fowleri*, but the prognosis remains poor. The few survivors have been treated with amphotericin B and rifampin. The new antiparasitic agent miltefosine—an alkylphosphocholine compound used to treat visceral leishmaniasis—is active in vitro against *Naegleria*, *Acanthamoeba*, and *Balamuthia* and is available from the CDC. Of three patients who received miltefosine for *Naegleria* infection, one recovered completely, one survived with significant neurologic deficits, and one died. Since 2013, when miltefosine became available through the CDC, this drug has been administered to both of two surviving U.S. patients with PAM and to three (33%) of nine patients who died of PAM (CDC, unpublished data). Early diagnosis, prompt combination therapy including miltefosine, and aggressive management of neurologic complications are important factors in better outcomes. A clinician whose patient may have PAM should contact the CDC Emergency Operations Center at (770) 488-7100 for assistance in diagnosis by PCR and treatment recommendations (which should include miltefosine).

ACANTHAMOEBA INFECTIONS

Granulomatous Amebic Encephalitis

Infection with *Acanthamoeba* species follows a more indolent course than *Naegleria* infection and typically occurs in chronically ill or debilitated patients. Risk factors include lymphoproliferative disorders, chemotherapy, glucocorticoid therapy, lupus erythematosus, and AIDS. Infection usually reaches the CNS hematogenously from a primary focus in the sinuses, skin, or lungs. In the CNS, the onset is insidious, and the syndrome often mimics a space-occupying lesion. Altered mental status, headache, and stiff neck may be accompanied by focal findings such as cranial nerve palsies, ataxia, and hemiparesis. Cutaneous ulcers or hard nodules containing amebae are frequently detected in AIDS patients with disseminated *Acanthamoeba* infection.

Examination of the CSF for trophozoites may be diagnostically helpful, but lumbar puncture may be contraindicated because of increased intracerebral pressure. CT frequently reveals cortical and subcortical lesions of decreased density consistent with embolic infarcts. In other patients, multiple enhancing lesions with edema may mimic the CT appearance of toxoplasmosis **(Chap. 223)**. Demonstration of the trophozoites and cysts of *Acanthamoeba* on wet mounts or in biopsy specimens establishes the diagnosis. Culture on non-nutrient agar plates seeded with *Escherichia coli* also may be helpful. Fluorescein-labeled antiserum is available from the CDC for the detection of protozoa in biopsy specimens. Granulomatous amebic encephalitis in patients with AIDS may have an accelerated course (with survival for only 3–40 days) because of the difficulty these individuals have in forming granulomas. Various antimicrobial agents have been used to treat *Acanthamoeba* infection, but miltefosine from the CDC should be included in combination therapy.

Keratitis

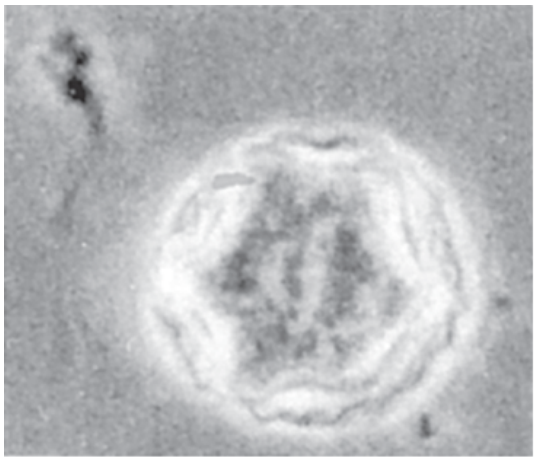
The incidence of keratitis caused by *Acanthamoeba* has increased in the past 20 years, in part as a result of improved diagnosis. Earlier infections were associated with trauma to the eye and exposure to contaminated water. At present, most infections are linked to extended-wear contact lenses, and rare cases are associated with laser-assisted in situ keratomileusis (LASIK). Risk factors include the use of homemade saline, the wearing of lenses while swimming, and inadequate disinfection. Since contact lenses presumably cause microscopic trauma, early corneal findings may be nonspecific. The first symptoms usually include tearing and the painful sensation of a foreign body.

Once infection is established, progression is rapid. The characteristic clinical sign is an annular, paracentral corneal ring representing a corneal abscess. Deeper corneal invasion and loss of vision may follow.

The differential diagnosis includes bacterial, mycobacterial, and herpetic infection. The irregular polygonal cysts of *Acanthamoeba* **(Fig. 218-4)** may be identified in corneal scrapings or biopsy material, and trophozoites can be grown on special media. Cysts are resistant to available drugs, and the results of medical therapy have been disappointing. Some reports have suggested partial responses to propamidine isethionate eyedrops. Severe infections usually require keratoplasty.

Figure 218-4

Double-walled cyst of *Acanthamoeba castellani*, as seen by phase-contrast microscopy. *(From DJ Krogstad et al, in A Balows et al [eds]: Manual of Clinical Microbiology, 5th ed. Washington, DC, American Society for Microbiology, 1991.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

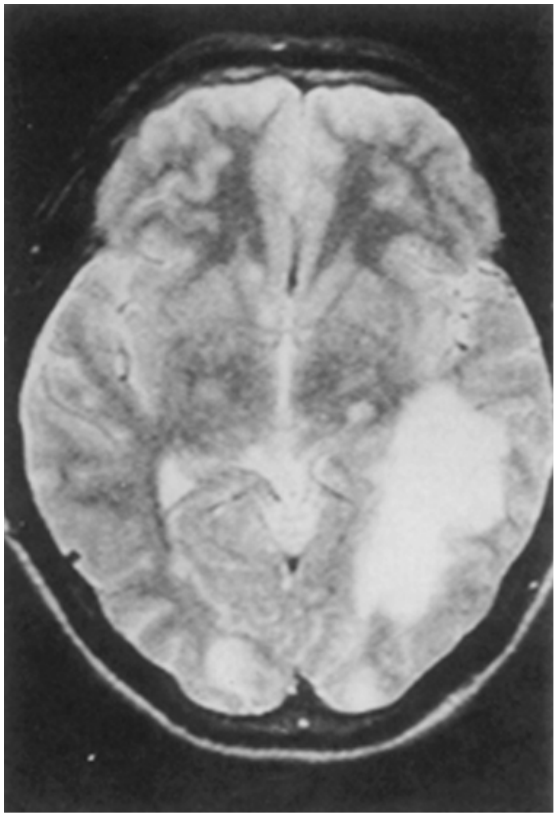
BALAMUTHIA INFECTIONS

Balamuthia mandrillaris is a free-living ameba that was first identified in 1986 as the cause of a fatal infection in a mandrill baboon at the Wild Animal Park in San Diego, California. The parasite has been isolated from soil and dust and is probably widespread in the environment. It is an important etiologic agent of granulomatous amebic encephalitis, cutaneous lesions, and sinus infections in humans. The potential risk factors for granulomatous amebic encephalitis identified by the California Encephalitis Project include young age, immunocompromising conditions, and Hispanic ethnicity. The infection likely starts with percutaneous or mucous membrane exposure and then spreads hematogenously to the brain and other organs—a pattern that explains the risk for transmission through organ transplantation. In 2009–2010, two clusters of organ transplant–transmitted *B. mandrillaris* infections were detected by recognition of severe unexpected illness in multiple recipients from the same donor after an incubation period of 17–24 days.

Frequently, *Balamuthia* affects immunocompetent individuals, in whom the course is typically subacute, with focal neurologic signs, fever, seizures, and headaches leading to death within 1 week to several months after onset. Skin lesions may occur on the face, trunk, or extremities. In addition to dust inhalation, inoculation of trophozoites or cysts from stagnant water may occur through open wounds or mucous membranes. Diagnosis relies on examination of CSF, which reveals mononuclear or neutrophilic pleocytosis, elevated protein levels, and normal to low glucose concentrations. Amebae are rarely isolated from CSF. Multiple hypodense lesions are usually detected with imaging studies **(Fig. 218-5)**. Fluorescent antibody and PCR assays are available from the CDC.

Figure 218-5

Brain MRI of amebic meningoencephalitis due to *Balamuthia mandrillaris*. A large lesion in the parieto-occipital lobe and other smaller lesions are seen. *(Courtesy of the Department of Radiology, UCSD Medical Center, San Diego.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The fewer than five surviving patients in the United States have been treated with a variety of drugs, including pentamidine, flucytosine, sulfadiazine, and macrolides. The CDC recommends that miltefosine now be included, as for treatment of other free-living amebae. The differential diagnosis includes tuberculomas **(Chap. 173)** and neurocysticercosis **(Chap. 230)**.

FURTHER READING

Amebiasis

Debnath A et al: A high-throughput drug screen for *Entamoeba histolytica* identifies a new lead and target. Nature Med 18:956, 2012. [PubMed: 22610278]
Gilchrist CA et al: Role of the gut microbiota of children in diarrhea due to the protozoan parasite *Entamoeba histolytica*. J Infect Dis 213:1579, 2016. [PubMed: 26712950]
Watanabe K, Petri WA Jr: Molecular biology research to benefit patients with *Entamoeba histolytica* infection. Molec Microbiol 98:208, 2015.

Free-Living Amebae

Capewell LG et al: Diagnosis, clinical course, and treatment of primary amoebic meningoencephalitis in the United States, 1937–2013. J Ped Infect Dis Soc 4:e68, 2015.
Farnon EC et al: Transmission of *Balamuthia mandrillaris* by organ transplantation. Clin Infect Dis 63:878, 2016. [PubMed: 27358357]

Chapter 219: Malaria

Nicholas J. White; Elizabeth A. Ashley

INTRODUCTION

Humanity has but three great enemies: Fever, famine, and war; of these by far the greatest, by far the most terrible, is fever.

—William Osler, 1896

Malaria is a protozoan disease transmitted by the bite of infected female *Anopheles* mosquitoes. The most important of the parasitic diseases of humans, malaria is transmitted in 91 countries containing 3 billion people and causes ~1200 deaths each day. Mortality rates have decreased dramatically over the past 15 years as a result of highly effective control programs in several countries. Malaria was eliminated from the United States, Canada, Europe, and Russia >50 years ago, but its prevalence rose in many parts of the tropics between 1970 and 2000. In response to this rise, there has been substantial investment aimed at increasing access to accurate diagnosis, effective treatments, and insecticide-treated bed nets. This investment has resulted in a decline in the global burden of malaria, although in the past few years progress has stalled. An increasing number of countries are now targeting malaria elimination. This ambitious goal is threatened by increasing resistance to antimalarial drugs and insecticides.

Malaria remains today, as it has been for centuries, a heavy burden on tropical communities, a threat to nonendemic countries, and a danger to travelers.

ETIOLOGY AND PATHOGENESIS

Six species of the genus *Plasmodium* cause nearly all malarial infections in humans. These are *P. falciparum*, *P. vivax*, two morphologically identical sympatric species of *P. ovale* (*curtisi* and *wallikeri*), *P. malariae*, and—in Southeast Asia—the monkey malaria parasite *P. knowlesi* **(Table 219-1)**. While almost all deaths are caused by falciparum malaria, *P. knowlesi* and occasionally *P. vivax* can also cause severe illness. Human infection begins when a female anopheline mosquito inoculates plasmodial *sporozoites* from its salivary glands during a blood meal **(Fig. 219-1)**. These microscopic motile forms of the malaria parasite are carried rapidly via the bloodstream to the liver, where they invade hepatic parenchymal cells and begin a period of asexual reproduction. By this amplification process (known as *intrahepatic* or *preerythrocytic schizogony*), a single sporozoite may produce from 10,000 to >30,000 daughter merozoites. The swollen infected liver cells eventually burst, discharging motile *merozoites* into the bloodstream. These merozoites then invade red blood cells (RBCs) to become *trophozoites* and multiply six- to twentyfold every 48 h (*P. knowlesi*, 24 h; *P. malariae*, 72 h). When the parasites reach densities of ~50/μL of blood (~100 million parasites in the blood of an adult), the symptomatic stage of the infection begins. In *P. vivax* and *P. ovale* infections, a proportion of the intrahepatic forms do not divide immediately but remain inert for a period ranging from 2 weeks to ≥1 year. These dormant forms, or *hypnozoites*, are the cause of the relapses that characterize infection with these species.

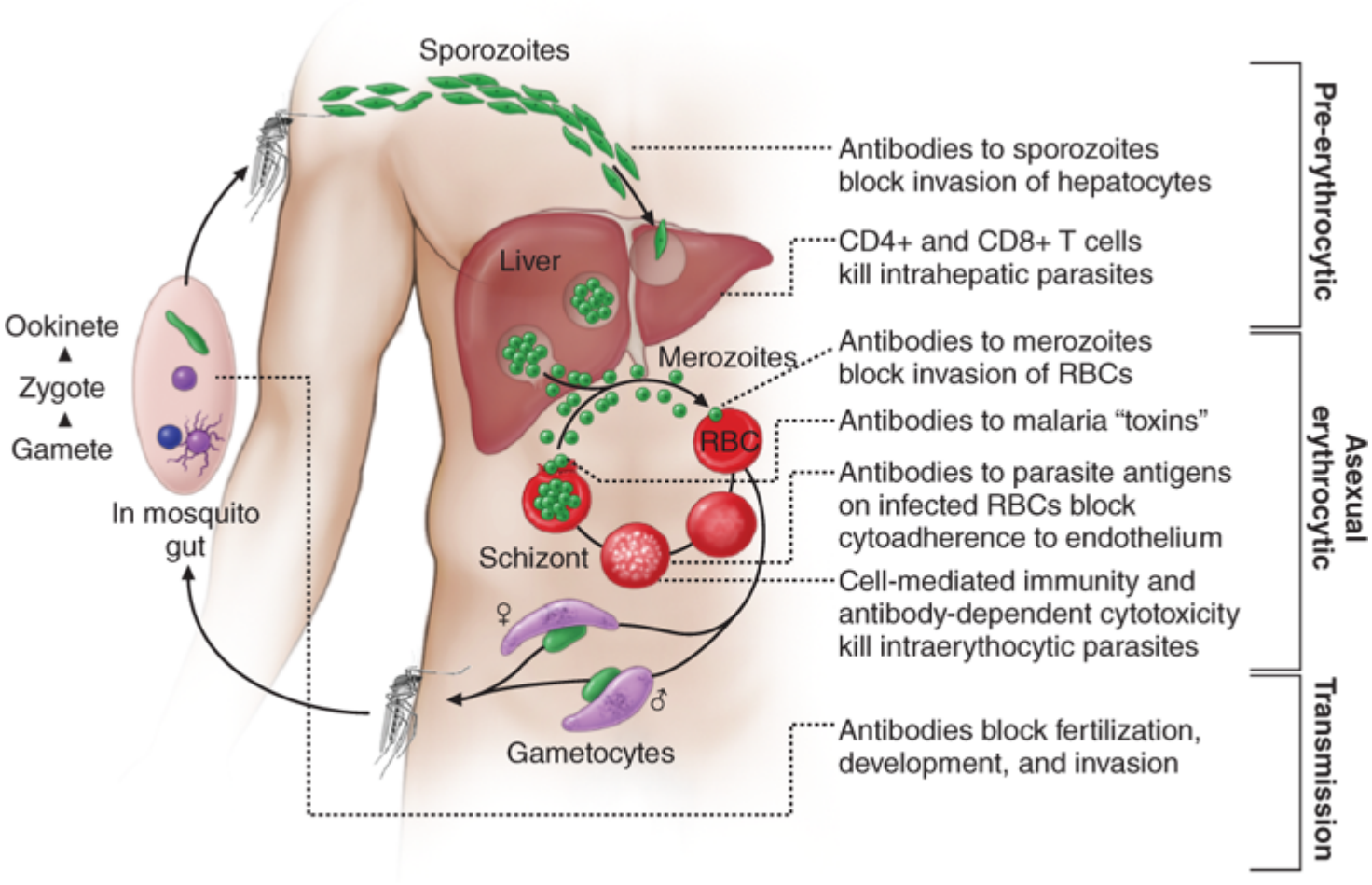
TABLE 219-1 Characteristics of *Plasmodium* Species Infecting Humans

Characteristic	Finding for Indicated Species				
	P. falciparum	***P. vivax***	***P. ovale*[a]**	***P. malariae***	***P. knowlesi***
Duration of intrahepatic phase (days)	5.5	8	9	15	5.5
Number of merozoites released per infected hepatocyte	30,000	10,000	15,000	15,000	20,000
Duration of erythrocytic cycle (hours)	48	48	50	72	24
Red cell preference	Younger cells (but can invade cells of all ages)	Reticulocytes and cells up to 2 weeks old	Reticulocytes	Older cells	Younger cells
Morphology	Usually only ring forms; banana-shaped gametocytes	Irregularly shaped large rings and trophozoites; enlarged erythrocytes; Schüffner's dots	Infected erythrocytes, enlarged and oval with tufted ends; Schüffner's dots	Band or rectangular forms of trophozoites common	Resembles *P. falciparum* (early trophozoites) or *P. malariae* (later trophozoites, including band forms)
Pigment color	Black	Yellow-brown	Dark brown	Brown-black	Dark brown
Ability to cause relapses	No	Yes	Yes	No	No

[a]Genomic studies have revealed *P. ovale* to be two sympatric species: *P. ovale curtisi* and *P. ovale wallikeri*.

FIGURE 219-1

The malaria transmission cycle from mosquito to human and targets of immunity. RBC, red blood cell.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Attachment of merozoites to erythrocytes is mediated via a complex interaction with several specific erythrocyte surface receptors. *P. falciparum* merozoites bind to erythrocyte binding antigen 175 and glycophorin A. The other glycophorins also contribute. The merozoite reticulocyte-binding protein homologue 5 (PfRh5) plays a critical role binding to red cell basigin (CD147, EMMPRIN). *P. vivax* binds to receptors on young red cells. The Duffy blood-group antigen Fy^a or Fy^b plays an important role in invasion. Most West Africans and people with origins in that region carry the Duffy-negative FyFy phenotype and are generally resistant to *P. vivax* malaria. *P. knowlesi* also invades Duffy-positive human RBCs preferentially. During the first few hours of intraerythrocytic development, the small "ring forms" of the different malaria species appear similar under light microscopy. As the trophozoites enlarge, species-specific characteristics become evident, malaria pigment (hemozoin) becomes visible, and the parasite assumes an irregular or ameboid shape. By the end of the intraerythrocytic life cycle, the parasite has consumed two-thirds of the RBC's hemoglobin and has grown to occupy most of the cell. It is now called a *schizont*. Multiple nuclear divisions have taken place (*schizogony* or *merogony*). The infected RBC then ruptures to release 6–30 daughter merozoites, each potentially capable of invading a new RBC and repeating the cycle. The disease in human beings is caused by the direct effects of the asexual parasite—RBC invasion and destruction—and by the host's reaction. Some of the blood-stage parasites develop into morphologically distinct, longer-lived sexual forms (*gametocytes*) that can transmit malaria. In falciparum malaria, a delay of several asexual cycles precedes this switch to gametocytogenesis. Female gametocytes typically outnumber males by 4:1.

After being ingested in the blood meal of a biting female anopheline mosquito, the male and female gametocytes fuse to form a zygote in the insect's midgut. This zygote matures into an ookinete, which penetrates and encysts in the mosquito's gut wall. The resulting oocyst expands by asexual division

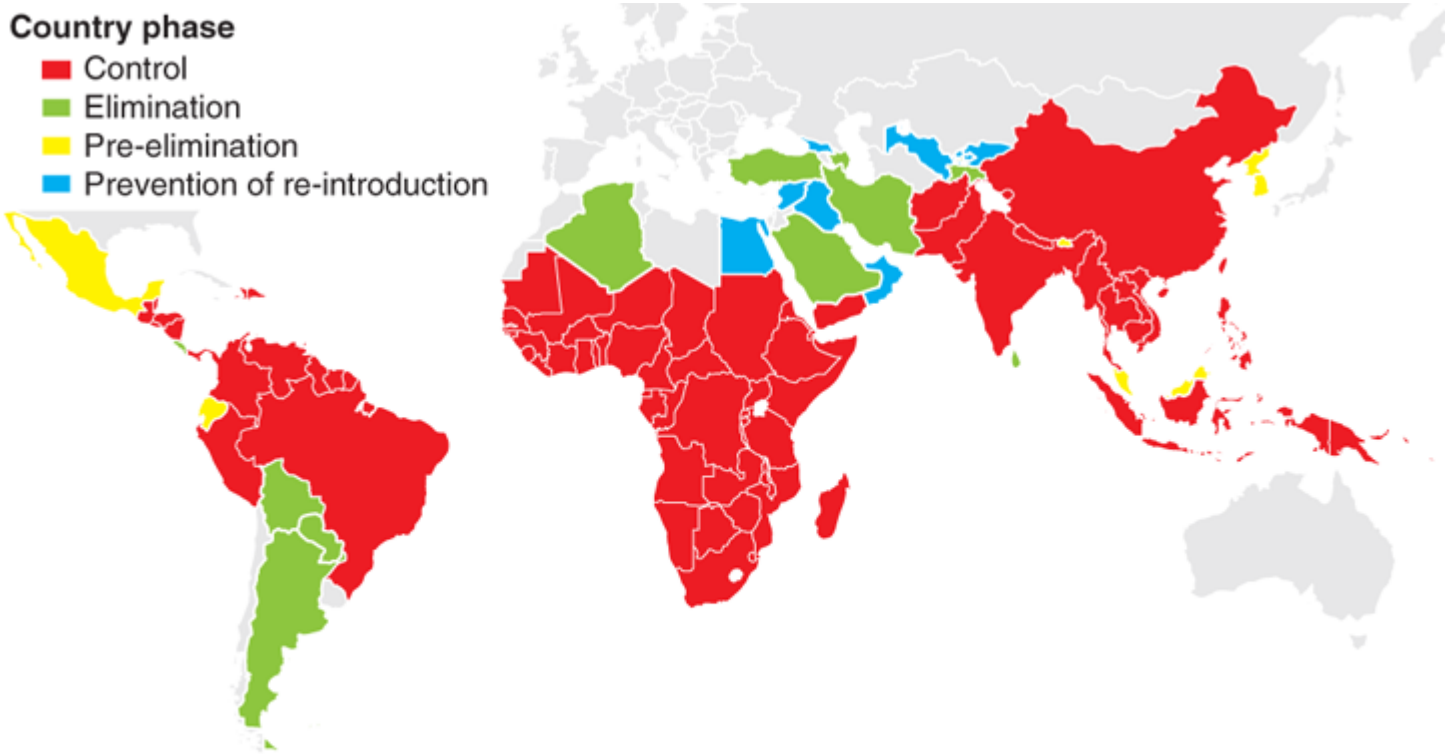
until it bursts to liberate myriad motile sporozoites, which then migrate in the hemolymph to the salivary gland of the mosquito to await inoculation into another human at the next feed, thus completing the life cycle.

EPIDEMIOLOGY

Malaria occurs throughout most of the tropical regions of the world **(Fig. 219-2)**. *P. falciparum* predominates in Africa, New Guinea, and Hispaniola (i.e., the Dominican Republic and Haiti); *P. vivax* is more common in Central and South America. The prevalence of these two species is approximately equal on the Indian subcontinent and in eastern Asia and Oceania. *P. malariae* is found in most endemic areas, especially throughout sub-Saharan Africa, but is much less common. *P. ovale* is relatively unusual outside of Africa and, where it is found, comprises <1% of isolates. *P. knowlesi* causes human infections commonly on the island of Borneo and, to a lesser extent, elsewhere in Southeast Asia, where the main hosts, long-tailed and pig-tailed macaques, are found.

FIGURE 219-2

Malaria-endemic countries showing progress towards elimination. (*Source:* worldmalariareport.org/.)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The epidemiology of malaria is complex and may vary considerably even within relatively small geographic areas. Endemicity traditionally has been defined in terms of rates of microscopy-detected parasitemia or palpable spleens in children 2–9 years of age and has been classified as hypoendemic (<10%), mesoendemic (11–50%), hyperendemic (51–75%), and holoendemic (>75%). In holo- and hyperendemic areas (e.g., certain regions of tropical Africa or coastal New Guinea) where there is intense *P. falciparum* transmission, people may sustain as much as one infectious mosquito bite per day and are infected repeatedly throughout their lives. In such settings, malaria morbidity and mortality are substantial during early childhood. Immunity against disease is hard won in these areas following

repeated symptomatic infections in childhood, but, if the child survives, infections become increasingly likely to be asymptomatic. These asymptomatic older children and adults are a major source of malaria transmission. As control measures progress and urbanization expands, environmental conditions become less conducive to malaria transmission, and all age groups may lose protective immunity and become susceptible to illness. Constant, frequent, year-round infection is termed *stable transmission*. In areas where transmission is low, erratic, or focal, full protective immunity is not acquired, and symptomatic disease may occur at all ages. This situation usually exists in hypoendemic areas and is termed *unstable transmission*. Even in stable-transmission areas, there is often an increased incidence of symptomatic malaria during the rainy season coinciding with increased mosquito breeding and transmission. Malaria can behave like an epidemic disease in some areas, particularly those with unstable malaria, such as northern India (the Punjab region), the Horn of Africa, Rwanda, Burundi, southern Africa, and Madagascar. Epidemics may occur when changes in environmental, economic, or social conditions (e.g., heavy rains following drought or migration—usually of refugees or workers—from a non-malarious region to an area of high transmission) are compounded by failure to invest in national programs or by a breakdown in malaria control and prevention services caused by war or civil disorder. Epidemics often result in high mortality rates among all age groups.

The principal determinants of the epidemiology of malaria are the number (density), the human-biting habits, and the longevity of the anopheline mosquito vectors. More than 100 of the >400 anopheline species can transmit malaria, but the ~40 species that do so commonly vary considerably in their efficiency as malaria vectors. More specifically, the transmission of malaria is directly proportional to the density of the vector, the square of the number of human bites per day per mosquito, and the tenth power of the probability of the mosquito's surviving for 1 day. Mosquito longevity is particularly important as a determinant of malaria transmissibility because the portion of the parasite's life cycle that takes place within the mosquito—from gametocyte ingestion to subsequent inoculation (*sporogony*)—lasts 8–30 days, depending on ambient temperature. In order to transmit malaria, the mosquito must therefore survive for >7 days. Sporogony is not completed at cooler temperatures—i.e., <16°C (<60.8°F) for *P. vivax* and <21°C (<69.8°F) for *P. falciparum*; thus transmission does not occur below these temperatures or at high altitudes, although malaria outbreaks and transmission have occurred in the highlands (>1500 m) of eastern Africa, which were previously free of vectors. The most effective mosquito vectors of malaria are those, such as the *Anopheles gambiae* species complex in Africa, that are long-lived, occur in high densities in tropical climates, breed readily, and bite humans in preference to other animals. The entomologic inoculation rate (i.e., the number of sporozoite-positive mosquito bites per person per year) is the most common measure of malaria transmission and varies from <1 in some parts of Latin America and Southeast Asia to >300 in parts of tropical Africa.

PATHOPHYSIOLOGY

ERYTHROCYTE CHANGES

After invading an erythrocyte, the growing malarial parasite progressively consumes and degrades intracellular proteins, principally hemoglobin. The potentially toxic heme is detoxified by lipid-mediated crystallization to biologically inert hemozoin (malaria pigment). The parasite also alters the RBC membrane by changing its transport properties, exposing cryptic surface antigens, and inserting new parasite-derived proteins. The RBC becomes more irregular in shape, more antigenic, and less deformable.

In *P. falciparum* infections, membrane protuberances appear on the erythrocyte's surface 12–15 h after the cell's invasion. These "knobs" extrude a high-molecular-weight, antigenically variant, strain-specific erythrocyte membrane adhesive protein (PfEMP1) that mediates attachment to receptors on venular and capillary endothelium (*cytoadherence*). Several vascular receptors have been identified;

intercellular adhesion molecule 1 and endothelial protein C receptor are important in the brain, chondroitin sulfate B predominates in the placenta, and CD36 binds parasitized RBCs in most other organs. Erythrocytes containing more mature parasites stick inside and eventually block capillaries and venules. These infected RBCs may also adhere to uninfected RBCs (to form rosettes) and to other parasitized erythrocytes (agglutination). The processes of cytoadherence, rosetting, and agglutination are central to the pathogenesis of falciparum malaria. They result in the sequestration of infected RBCs in vital organs (particularly the brain), where they interfere with microcirculatory flow and metabolism. Sequestered parasites continue to develop out of reach of the principal host defense mechanism: splenic processing and filtration. As a consequence, only the younger ring forms of the asexual parasites are seen circulating in the peripheral blood in falciparum malaria, and the level of peripheral parasitemia underestimates the true number of parasites within the body. Severe malaria is also associated with reduced deformability of uninfected erythrocytes, which compromises their passage through the partially obstructed capillaries and venules and shortens their survival.

In the other human malarias, significant sequestration does not occur, and all stages of the parasite's development are evident on peripheral-blood smears. *P. vivax* and *P. ovale* show a marked predilection for young RBCs and *P. malariae* for old cells; these species produce a level of parasitemia that seldom exceeds 2%. In contrast, *P. falciparum* can invade erythrocytes of all ages and may be associated with very high parasite densities. Dangerously high parasite densities may also occur in *P. knowlesi* infections, with rapid increases as a result of the shorter (24-h) asexual life cycle.

HOST RESPONSE

Initially, the host responds to malaria infection by activating nonspecific defense mechanisms. Splenic immunologic and filtrative clearance functions are augmented, and the removal of both parasitized and uninfected erythrocytes is accelerated. The spleen also removes damaged ring-form parasites (a process known as "pitting") and returns the once-infected erythrocytes to the circulation, where their survival is shortened. The parasitized cells escaping splenic removal are destroyed when the schizont ruptures. The material released induces monocyte/macrophage activation and the release of proinflammatory cytokines, which cause fever and other pathologic effects. Temperatures of ≥40°C (≥104°F) damage mature parasites; in untreated infections, the effect of such temperatures is to further synchronize the parasitic cycle, with eventual production of the regular fever spikes and rigors that originally characterized the different malarias. These regular fever patterns (*quotidian*, daily; *tertian*, every 2 days; *quartan*, every 3 days) are seldom seen today as patients receive prompt and effective antimalarial treatment.

The geographic distributions of the thalassemias, sickle cell disease, hemoglobins C and E, hereditary ovalocytosis, and glucose-6-phosphate dehydrogenase (G6PD) deficiency closely resemble that of falciparum malaria before the introduction of control measures. This similarity suggests that these genetic disorders confer protection against death from falciparum malaria. For example, HbA/S heterozygotes (sickle cell trait) have a sixfold reduction in the risk of dying from severe falciparum malaria and are correspondingly protected from bacterial infections that complicate malaria. Hemoglobin S–containing RBCs impair parasite growth at low oxygen tensions, and *P. falciparum*–infected RBCs containing hemoglobin S or C exhibit reduced cytoadherence because of reduced surface presentation of the adhesin PfEMP1. Parasite multiplication in HbA/E heterozygotes is reduced at high parasite densities. In Melanesia, children with α-thalassemia have more frequent malaria (both vivax and falciparum) in the early years of life, and this pattern of infection appears to protect them against severe disease. In Melanesian ovalocytosis, rigid erythrocytes resist merozoite invasion, and the intraerythrocytic milieu is hostile.

Nonspecific host defense mechanisms stop the infection's expansion, and the subsequent strain-specific immune response then controls the infection. Eventually, exposure to sufficient strains confers protection from high-level parasitemia and disease but not from infection. As a result of this state of infection without illness (*premunition*), asymptomatic parasitemia is very common among adults and

older children living in regions with stable and intense transmission (i.e., holo- or hyperendemic areas) and also in parts of low-transmission areas. Parasitemia in asymptomatic infections fluctuates in density but often averages ~5000/mL—just below the level of microscopy detection but sufficient to generate transmissible densities of gametocytes. Immunity is mainly specific for both the species and the strain of infecting malarial parasite. Both humoral immunity and cellular immunity are necessary for protection, but the mechanisms of each are incompletely understood (Fig. 219-1). Immune individuals have a polyclonal increase in serum levels of IgM, IgG, and IgA, although much of this antibody is unrelated to protection. Antibodies to a variety of parasite antigens presumably act in concert to limit in vivo replication of the parasite. In the case of falciparum malaria, the most important of these antigens is the surface adhesin—the variant protein family PfEMP1. Passively transferred IgG from immune adults has been shown to reduce levels of parasitemia in children. Passive transfer of maternal antibody contributes to the partial protection of infants from severe malaria in the first months of life. This complex immunity to disease declines when a person lives outside an endemic area for several months or longer.

Several factors retard the development of cellular immunity to malaria. These factors include the absence of major histocompatibility antigens on the surface of infected RBCs, which precludes direct T cell recognition; malaria antigen–specific immune unresponsiveness; and the enormous strain diversity of malarial parasites, along with the ability of the parasites to express variant immunodominant antigens on the erythrocyte surface that change during the course of infection. Parasites may persist in the blood for months or years (or, in the case of *P. malariae*, for decades) if treatment is not given. The complexity of the immune response in malaria, the sophistication of the parasites' evasion mechanisms, and the lack of a good in vitro correlate with clinical immunity have all slowed progress toward an effective vaccine.

CLINICAL FEATURES

Malaria is a common cause of fever in tropical countries. Clinical diagnosis is notoriously unreliable. The first symptoms of malaria are nonspecific; the lack of a sense of well-being, headache, fatigue, abdominal discomfort, and muscle aches followed by fever are all similar to the symptoms of a minor viral illness. In some instances, a prominence of headache, chest pain, abdominal pain, cough, arthralgia, myalgia, or diarrhea may suggest another diagnosis. Although headache may be severe in malaria, the neck stiffness and photophobia seen in meningitis do not occur. While myalgia may be prominent, it is not usually as severe as in dengue fever, and the muscles are not tender as in leptospirosis or typhus. Nausea, vomiting, and orthostatic hypotension are common. The classic malarial paroxysms, in which fever spikes, chills, and rigors occur at regular intervals, are relatively unusual and suggest infection (often relapse) with *P. vivax* or *P. ovale*. The fever is usually irregular at first (that of falciparum malaria may never become regular). The temperature of nonimmune individuals and children often rises above 40°C (104°F), with accompanying tachycardia and sometimes delirium. Although childhood febrile convulsions may occur with any of the malarias, generalized seizures are associated specifically with falciparum malaria and may herald the development of encephalopathy (cerebral malaria). Many clinical abnormalities have been described in acute malaria, but most patients with uncomplicated infections have few abnormal physical findings other than fever, malaise, mild anemia, and (in some cases) a palpable spleen. Anemia is common among young children living in areas with stable transmission (e.g., much of West Africa), particularly where resistance has compromised the efficacy of antimalarial drugs. Frequent vivax relapse is an important cause of anemia in young children in some areas (e.g., on the island of New Guinea). In nonimmune individuals with acute malaria, the spleen takes several days to become palpable, but splenic enlargement is found in a high proportion of otherwise healthy individuals in malaria-endemic areas and reflects repeated infections. Slight enlargement of the liver is also common, particularly among young children. Mild jaundice is common among adults; it may develop in patients with otherwise uncomplicated malaria and usually resolves over 1–3 weeks. Malaria is not associated with

a rash. Petechial hemorrhages in the skin or mucous membranes—features of viral hemorrhagic fevers and leptospirosis—develop only very rarely in severe falciparum malaria.

SEVERE FALCIPARUM MALARIA

Appropriately and promptly treated, uncomplicated falciparum malaria (i.e., that in which the patient can sit or stand unaided and can swallow medicines and food) carries a mortality rate of <0.1%. However, once vital-organ dysfunction occurs or the total proportion of erythrocytes infected increases to >2% (a level corresponding to $>10^{12}$ parasites in an adult), mortality risk rises steeply, depending on the immunity of the host. The major manifestations of severe falciparum malaria are shown in **Table 219-2**, and features indicating a poor prognosis are listed in **Table 219-3**.

TABLE 219-2 Manifestations of Severe Falciparum Malaria

Signs	Manifestations
Major	
Unarousable coma/cerebral malaria	Failure to localize or respond appropriately to noxious stimuli; coma persisting for >30 min after generalized convulsion
Acidemia/acidosis	Arterial pH of <7.25, base deficit >8 meq/L, or plasma bicarbonate level of <15 mmol/L; venous lactate level of >5 mmol/L; manifests as labored deep breathing, often termed “respiratory distress”
Severe normochromic, normocytic anemia	Hematocrit of <15% or hemoglobin level of <50 g/L (<5 g/dL) with parasitemia level of <10,000/μL
Renal failure	Serum or plasma creatinine level of >265 μmol/L (>3 mg/dL); urine output (24 h) of <400 mL for adults or <12 mL/kg for children; no improvement with rehydration
Pulmonary edema/adult respiratory distress syndrome	Noncardiogenic pulmonary edema, often aggravated by overhydration
Hypoglycemia	Plasma glucose level of <2.2 mmol/L (<40 mg/dL)
Hypotension/shock	Systolic blood pressure of <50 mmHg in children 1–5 years or <80 mmHg in adults; core/skin temperature difference of >10°C; capillary refill >2 s
Bleeding/disseminated intravascular coagulation	Significant bleeding and hemorrhage from the gums, nose, and gastrointestinal tract and/or evidence of disseminated intravascular coagulation
Convulsions	More than two generalized seizures in 24 h; signs of continued seizure activity, sometimes subtle (e.g., tonic-clonic eye movements without limb or face movement)
Other	
Hemoglobinuria[a]	Macroscopic black, brown, or red urine; not associated with effects of oxidant drugs and red blood cell enzyme defects (such as G6PD deficiency)
Extreme weakness	Prostration; inability to sit unaided[b]
Hyperparasitemia	Parasitemia level of >5% in nonimmune patients (>10% in any patient)
Jaundice	Serum bilirubin level of >50 mmol/L (>3 mg/dL) if combined with a parasite density of 100,000/μL or other evidence of vital-organ dysfunction

[a]Hemoglobinuria may also occur in uncomplicated malaria and in patients with G6PD deficiency who take primaquine. [b]In children who are normally able to sit.

Abbreviation: G6PD, glucose-6-phosphate dehydrogenase.

TABLE 219-3 Features Indicating a Poor Prognosis in Severe Falciparum Malaria

Clinical

Marked agitation

Hyperventilation (respiratory distress)

Low core temperature (<36.5°C; <97.7°F)

Bleeding

Deep coma

Repeated convulsions

Anuria

Shock

Laboratory

Biochemistry

- Hypoglycemia (<2.2 mmol/L)
- Hyperlactatemia (>5 mmol/L)
- Acidemia (arterial pH <7.25, base deficit >8 meq/L, or serum HCO_3 <15 mmol/L)
- Elevated serum creatinine (>265 μmol/L)
- Elevated total bilirubin (>50 μmol/L)
- Elevated liver enzymes (AST/ALT 3 times upper limit of normal)
- Elevated muscle enzymes (CPK ↑, myoglobin ↑)
- Elevated urate (>600 μmol/L)

Hematology

- Leukocytosis (>12,000/μL)
- Severe anemia (PCV <15%)
- Coagulopathy
  - Decreased platelet count (<50,000/μL)
  - Prolonged prothrombin time (>3 s)
  - Prolonged partial thromboplastin time
  - Decreased fibrinogen (<200 mg/dL)

Parasitology

- Hyperparasitemia
  - Increased mortality at >100,000/μL

High mortality at >500,000/μL

>20% of parasites identified as pigment-containing trophozoites and schizonts

>5% of neutrophils contain visible malaria pigment

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; CPK, creatine phosphokinase; PCV, packed cell volume.

Cerebral Malaria

Coma is a characteristic and ominous feature of falciparum malaria and, even with treatment, has been associated with death rates of ~20% among adults and 15% among children. Any obtundation, delirium, or abnormal behavior in falciparum malaria should be taken very seriously. The onset of coma may be gradual or sudden following a convulsion.

Cerebral malaria manifests as diffuse symmetric encephalopathy; focal neurologic signs are unusual. Although some passive resistance to head flexion may be detected, signs of meningeal irritation are absent. The eyes may be divergent, and bruxism and a pout reflex are common, but other primitive reflexes are usually absent. The corneal reflexes are preserved, except in deep coma. Muscle tone may be either increased or decreased. The tendon reflexes are variable, and the plantar reflexes may be flexor or extensor; the abdominal and cremasteric reflexes are absent. Flexor or extensor posturing may be seen. On routine funduscopy, ~15% of patients have retinal hemorrhages; with pupillary dilation and indirect ophthalmoscopy, this figure increases to 30–40%. Other funduscopic abnormalities **(Fig. 219-3)** include discrete spots of retinal opacification (30–60%), papilledema (8% among children, rare among adults), cotton wool spots (<5%), and decolorization of a retinal vessel or segment of vessel (occasional cases). Convulsions, which are usually generalized and often repeated, occur in ~10% of adults and up to 50% of children with cerebral malaria. More covert seizure activity is common, particularly among children, and may manifest as repetitive tonic–clonic eye movements or even hypersalivation. Whereas adults rarely (<3% of cases) suffer neurologic sequelae, ~10% of children surviving cerebral malaria—especially those with hypoglycemia, severe anemia, repeated seizures, and deep coma—have residual neurologic deficits when they regain consciousness; hemiplegia, cerebral palsy, cortical blindness, deafness, and impaired cognition may all occur. The majority of these deficits improve markedly or resolve completely within 6 months. However, the prevalence of some other deficits increases over time; ~10% of children surviving cerebral malaria have a persistent language deficit. There may also be deficits in learning, planning and executive functions, attention, memory, and nonverbal functioning. The incidence of epilepsy is increased and life expectancy decreased among these children.

FIGURE 219-3

The eye in cerebral malaria: perimacular whitening and pale-centered retinal hemorrhages. *(Courtesy of N. Beare, T. Taylor, S. Harding, S. Lewallen, and M. Molyneux; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Hypoglycemia

Hypoglycemia, an important and common complication of severe malaria, is associated with a poor prognosis and is particularly problematic in children and pregnant women. Hypoglycemia in malaria results from a failure of hepatic gluconeogenesis and an increase in the consumption of glucose by both the host and, to a much lesser extent, the malaria parasites. This abnormality may be compounded by quinine, a powerful stimulant of pancreatic insulin secretion, which is still widely used for the treatment of both severe and uncomplicated falciparum malaria. Hyperinsulinemic hypoglycemia is especially troublesome in pregnant women receiving quinine treatment. In severe disease, the clinical diagnosis of hypoglycemia is difficult: the usual physical signs (sweating, gooseflesh, tachycardia) are absent, and the neurologic impairment caused by hypoglycemia cannot be distinguished from that caused by malaria.

Acidosis

Acidosis is an important cause of death from severe malaria and results from accumulation of organic acids. Hyperlactatemia commonly coexists with hypoglycemia. In adults, coexisting renal impairment often compounds acidosis. In children, ketoacidosis also may contribute. Hydroxyphenyllactic acid, α-hydroxybutyric acid, and β-hydroxybutyric acid concentrations are elevated. Acidotic breathing, sometimes called “respiratory distress,” is a sign of poor prognosis. It is followed often by circulatory failure refractory to volume expansion or inotropic drug treatment and ultimately by respiratory arrest.

Plasma concentrations of bicarbonate or lactate are the best biochemical prognosticators in severe malaria. Hypovolemia is not a major contributor to acidosis. Lactic acidosis is caused by the combination of anaerobic glycolysis in tissues where sequestered parasites interfere with microcirculatory flow, lactate production by the parasites, and a failure of hepatic and renal lactate clearance.

Noncardiogenic Pulmonary Edema

Adults with severe falciparum malaria may develop noncardiogenic pulmonary edema even after several days of antimalarial therapy. The pathogenesis of this variant of the adult respiratory distress syndrome is unclear. The mortality rate is >80%. Pulmonary edema can be precipitated by overly vigorous administration of IV fluid. Noncardiogenic pulmonary edema can also develop in otherwise uncomplicated vivax malaria, where recovery is usual.

Renal Impairment

Acute kidney injury is common in severe falciparum malaria. The pathogenesis of renal failure is unclear but may be related to erythrocyte sequestration and agglutination interfering with renal microcirculatory flow and metabolism. Clinically and pathologically, this syndrome manifests as acute tubular necrosis. Acute renal failure may occur simultaneously with other vital-organ dysfunction (in which case the mortality risk is high) or may progress as other disease manifestations resolve. In survivors, urine flow resumes in a median of 4 days, and serum creatinine levels return to normal in a mean of 17 days **(Chap. 304)**. Early dialysis or hemofiltration considerably enhances the likelihood of a patient's survival, particularly in acute hypercatabolic renal failure. Oliguric renal failure is rare among children.

Hematologic Abnormalities

Anemia results from accelerated RBC removal by the spleen, obligatory RBC destruction at parasite schizogony, and ineffective erythropoiesis. In severe malaria, the deformability of both infected and uninfected RBCs is reduced. The degree of reduced deformability correlates with prognosis and with the development of anemia. Splenic clearance of all RBCs is increased. In nonimmune individuals and in areas with unstable transmission, anemia can develop rapidly and transfusion is often required. Acute hemolytic anemia with massive hemoglobinuria ("blackwater fever") may occur. Hemoglobinuria may contribute to renal injury. Some patients with blackwater fever have G6PD deficiency, but in the majority of cases it is unclear why massive hemolysis has occurred. Sudden hemolysis may follow many days after artesunate treatment of hyperparasitemia, usually as a result of relatively synchronous loss of once-parasitized "pitted" RBCs. As a consequence of repeated malarial infections, children in high-transmission areas may develop severe anemia resulting from both shortened survival of uninfected RBCs and marked dyserythropoiesis. Anemia is a common consequence of antimalarial drug resistance, which results in repeated or continued infection.

Slight coagulation abnormalities are common in falciparum malaria, and mild thrombocytopenia is usual (a normal platelet count should raise questions about the diagnosis of malaria). Fewer than 5% of patients with severe malaria have significant bleeding with evidence of disseminated intravascular coagulation. Hematemesis from stress ulceration or acute gastric erosions also may occur rarely.

Liver Dysfunction

Mild hemolytic jaundice is common in malaria. Severe jaundice is associated with *P. falciparum* infections; is more common among adults than among children; and results from hemolysis, hepatocyte injury, and cholestasis. When accompanied by other vital-organ dysfunction (often renal impairment), liver dysfunction carries a poor prognosis. Hepatic dysfunction contributes to

hypoglycemia, lactic acidosis, and impaired drug metabolism. Occasional patients with falciparum malaria may develop deep jaundice (with hemolytic, hepatic, and cholestatic components) without evidence of other vital-organ dysfunction, in which case the prognosis is good.

Other Complications

HIV/AIDS and malnutrition predispose to more severe malaria in nonimmune individuals. Malaria anemia is worsened by concurrent infections with intestinal helminths, hookworm in particular. Septicemia may complicate severe malaria, particularly in children. Differentiating severe malaria from sepsis with incidental parasitemia in childhood is very difficult. In endemic areas, *Salmonella* spp. bacteremia has been associated specifically with *P. falciparum* infections. Chest infections and catheter-induced urinary tract infections are common among patients who are unconscious for >3 days. Aspiration pneumonia may follow generalized convulsions. The frequencies of complications of severe falciparum malaria are summarized in **Table 219-4**.

TABLE 219-4 Relative Incidence of Severe Complications of Falciparum Malaria

Complication	Nonpregnant Adults	Pregnant Women	Children
Anemia	+	++	+++
Convulsions	+	+	+++
Hypoglycemia	+	+++	+++
Jaundice	+++	+++	+
Renal failure	+++	+++	–
Pulmonary edema	++	+++	+

Note: –, rare; +, infrequent; ++, frequent; +++, very frequent.

MALARIA IN PREGNANCY

Malaria in early pregnancy causes fetal loss. In areas of high malaria transmission, falciparum malaria in primi- and secundigravid women is associated with low birth weight (average reduction, ~170 g) and consequently increased infant mortality rates. In general, infected mothers in areas of stable transmission remain asymptomatic despite intense accumulation of parasitized erythrocytes in the placental microcirculation. Maternal HIV infection predisposes pregnant women to more frequent and higher-density malaria infections, predisposes their newborns to congenital malarial infection, and exacerbates the reduction in birth weight associated with malaria.

In areas with unstable transmission of malaria, pregnant women are prone to severe infections and are particularly likely to develop high parasitemias with anemia, hypoglycemia, and acute pulmonary edema. Fetal distress, premature labor, and stillbirth or low birth weight are common results. Fetal death is usual in severe malaria. Congenital malaria occurs in fewer than 5% of newborns whose mothers are infected; its frequency and the level of parasitemia are related directly to the timing of maternal infection and the parasite density in maternal blood and in the placenta. *P. vivax* malaria in pregnancy is also associated with a reduction in birth weight (average, 110 g) but, in contrast to observations in falciparum malaria, this effect is more pronounced in multigravid than in primigravid women. About 300,000 women die in childbirth yearly, with most deaths occurring in low-income countries; maternal death from hemorrhage at childbirth is correlated with malaria-induced anemia.

MALARIA IN CHILDREN

Most of the estimated 445,000 deaths from falciparum malaria each year are in young African children. Convulsions, coma, hypoglycemia, metabolic acidosis, and severe anemia are relatively common among children with severe malaria, whereas deep jaundice, oliguric acute kidney injury, and

acute pulmonary edema are unusual. Severely anemic children may present with labored deep breathing, which in the past has been attributed incorrectly to “anemic congestive cardiac failure” but in fact is usually caused by metabolic acidosis, sometimes compounded by hypovolemia. In general, children tolerate antimalarial drugs well and respond rapidly to treatment.

TRANSFUSION MALARIA

Malaria can be transmitted by blood transfusion, needlestick injury, or organ transplantation. The incubation period in these settings is often short because there is no preerythrocytic stage of development. The clinical features and management of these cases are the same as for naturally acquired infections. Radical chemotherapy with primaquine is unnecessary for transfusion-transmitted *P. vivax* and *P. ovale* infections.

CHRONIC COMPLICATIONS OF MALARIA

HYPERREACTIVE MALARIAL SPLENOMEGALY

Chronic or repeated malarial infections produce hypergammaglobulinemia; normochromic, normocytic anemia; and, in certain situations, splenomegaly. Some residents of malaria-endemic areas in tropical countries exhibit an abnormal immunologic response to repeated infections that is characterized by massive splenomegaly, hepatomegaly, marked elevations in serum IgM and malarial antibody titers, hepatic sinusoidal lymphocytosis, and (in Africa) peripheral B cell lymphocytosis. This syndrome has been associated with the production of cytotoxic IgM antibodies to CD8+ T lymphocytes, antibodies to CD5+ T lymphocytes, and an increase in the ratio of CD4+ to CD8+ T cells. These events may lead to uninhibited B cell production of IgM and the formation of cryoglobulins (IgM aggregates and immune complexes). This immunologic process stimulates lymphoid hyperplasia and clearance activity and eventually produces splenomegaly. Patients with hyperreactive malarial splenomegaly present with an abdominal mass or a dragging sensation in the abdomen and occasional sharp abdominal pains suggesting perisplenitis. There is usually anemia and some degree of pancytopenia (hypersplenism). In some cases, malaria parasites cannot be found in peripheral-blood smears by microscopy. Vulnerability to respiratory and skin infections is increased; many patients die of overwhelming sepsis. Persons with hyperreactive malarial splenomegaly living in endemic areas should receive antimalarial chemoprophylaxis; the results are usually good. In nonendemic areas, antimalarial treatment is advised. Some cases have been mistaken for hematologic malignancy. However, in other cases refractory to therapy, clonal lymphoproliferation may develop and can evolve into a malignant lymphoproliferative disorder.

QUARTAN MALARIAL NEPHROPATHY

Chronic or repeated infections with *P. malariae* (and possibly with other malarial species) may cause soluble immune complex injury to the renal glomeruli, resulting in the nephrotic syndrome. Other unidentified factors must contribute to this process since only a very small proportion of infected patients develop renal disease. The histologic appearance is that of focal or segmental glomerulonephritis with splitting of the capillary basement membrane. Subendothelial dense deposits are seen on electron microscopy, and immunofluorescence reveals deposits of complement and immunoglobulins; in samples of renal tissue from children, *P. malariae* antigens are often visible. A coarse-granular pattern of basement membrane immunofluorescent deposits (predominantly IgG3) with selective proteinuria carries a better prognosis than a fine-granular, predominantly IgG2 pattern with nonselective proteinuria. Quartan nephropathy is rarely reported nowadays. It usually responds poorly to treatment with either antimalarial agents or glucocorticoids and cytotoxic drugs.

BURKITT’S LYMPHOMA AND EPSTEIN-BARR VIRUS INFECTION

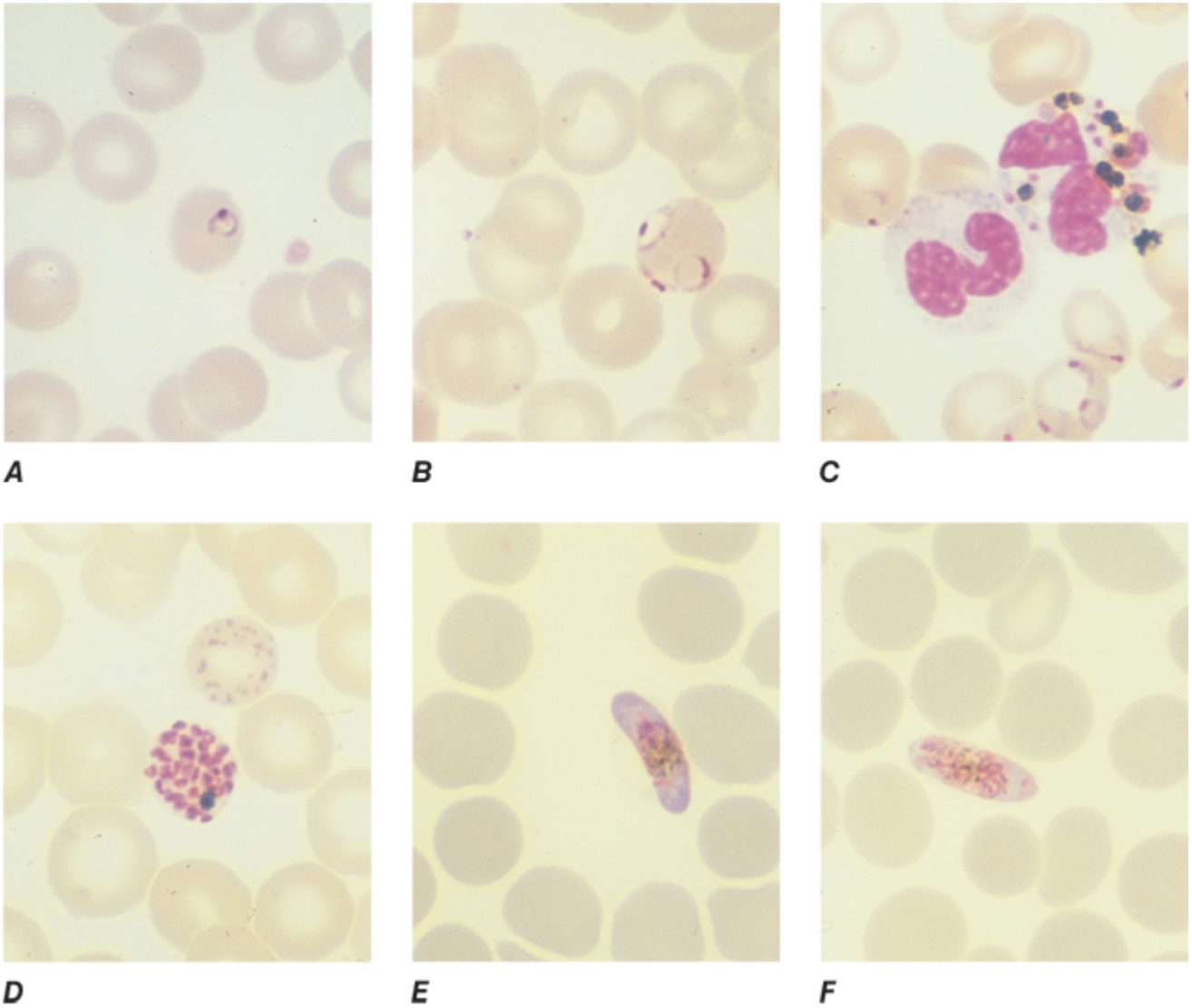
It is possible that malaria-related immune dysregulation provokes infection with lymphoma viruses. Burkitt's lymphoma is strongly associated with Epstein-Barr virus. The prevalence of this childhood tumor is high in high-malaria-transmission areas of Africa.

DIAGNOSIS OF MALARIA

When a patient in or from a malarious area presents with fever, thick and thin blood smears should be prepared and *examined immediately* to confirm the diagnosis and identify the species of infecting parasite **(Figs. 219-4, 219-5, 219-6, 219-7, 219-8, 219-9)**. In general, if the blood smear is negative when examined by an experienced microscopist, the patient does not have malaria. If reliable microscopy is not available, a rapid test should be performed.

FIGURE 219-4

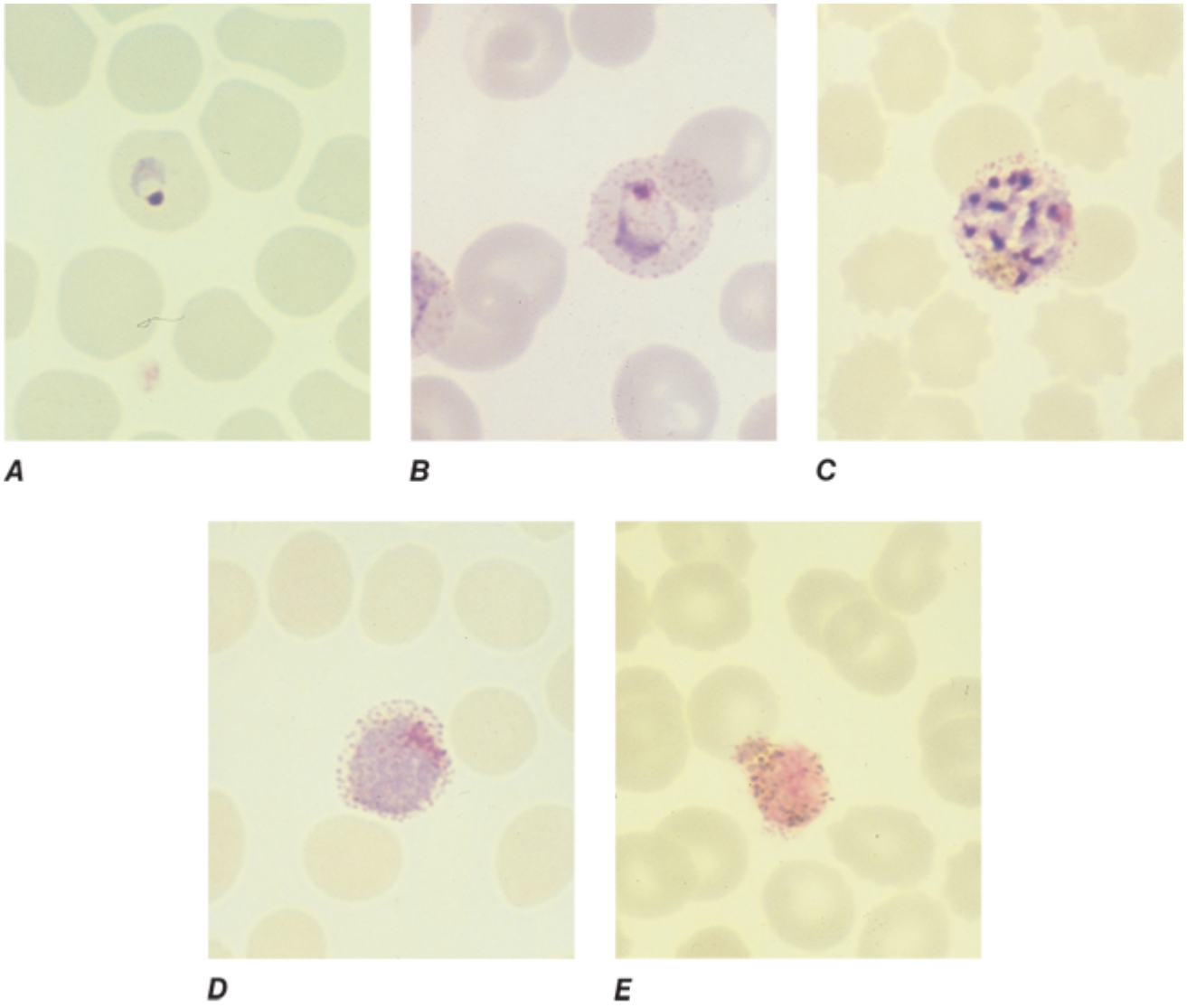
Thin blood films of *Plasmodium falciparum*. *A.* Young trophozoite. ***B.*** Old trophozoite. ***C.*** Trophozoites in erythrocytes and pigment in polymorphonuclear cells. ***D.*** Mature schizont. ***E.*** Female gametocyte. ***F.*** Male gametocyte. *(Reproduced from Bench Aids for the Diagnosis of Malaria Infections, 2nd ed, with the permission of the World Health Organization*.)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 219-5

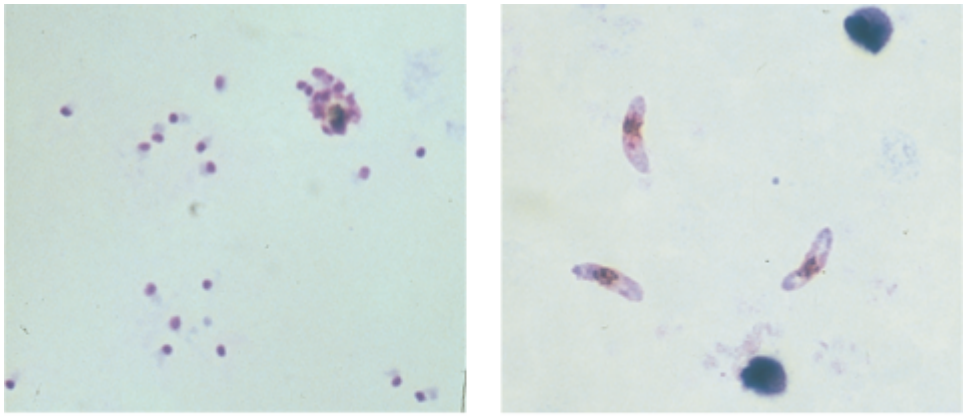
Thin blood films of *Plasmodium vivax*. *A.* Young trophozoite. ***B.*** Old trophozoite. ***C.*** Mature schizont. ***D.*** Female gametocyte. ***E.*** Male gametocyte. *(Reproduced from Bench Aids for the Diagnosis of Malaria Infections, 2nd ed, with the permission of the World Health Organization.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 219-6

Thick blood films of *Plasmodium falciparum*. *A.* Trophozoites. ***B.*** Gametocytes. *(Reproduced from Bench Aids for the Diagnosis of Malaria Infections, 2nd ed, with the permission of the World Health Organization.)*

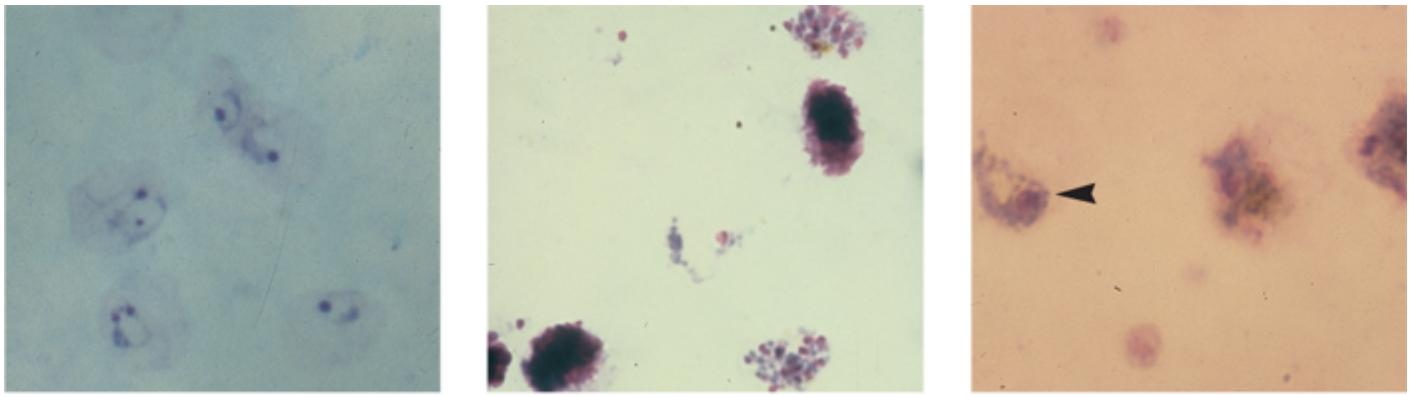


A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 219-7

Thick blood films of *Plasmodium vivax*. *A.* Trophozoites. ***B.*** Schizonts. ***C.*** Gametocytes. *(Reproduced from Bench Aids for the Diagnosis of Malaria Infections, 2nd ed, with the permission of the World Health Organization.)*

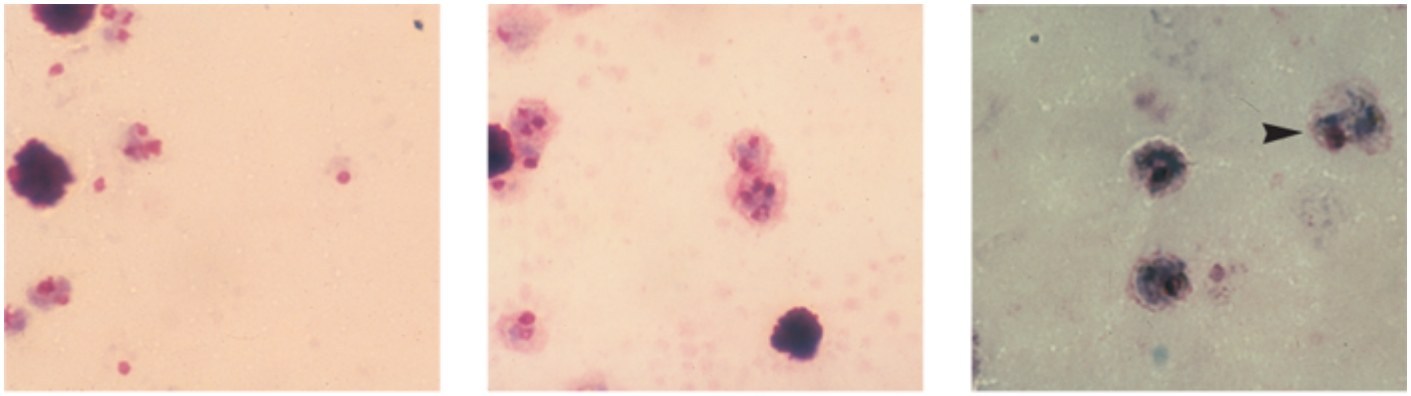


A B C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 219-8

Thick blood films of *Plasmodium ovale*. *A.* Trophozoites. ***B.*** Schizonts. ***C.*** Gametocytes. *(Reproduced from Bench Aids for the Diagnosis of Malaria Infections, 2nd ed, with the permission of the World Health Organization.)*

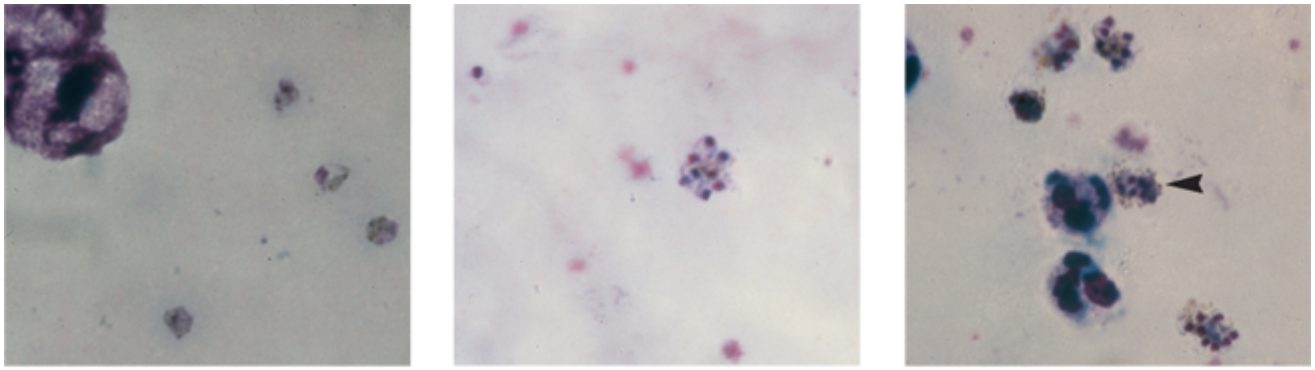


A B C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 219-9

Thick blood films of *Plasmodium malariae*. *A.* Trophozoites. ***B.*** Schizonts. ***C.*** Gametocytes. *(Reproduced from Bench Aids for the Diagnosis of Malaria Infections, 2nd ed, with the permission of the World Health Organization.)*



A B C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DEMONSTRATION OF THE PARASITE

The diagnosis of malaria rests on the demonstration of asexual forms of the parasite in stained peripheral-blood smears. Of the Romanowsky stains, Giemsa at pH 7.2 is preferred; Field's, Wright's, or Leishman's stain can also be used. Staining of parasites with the fluorescent dye acridine orange allows more rapid diagnosis of malaria (but not speciation of the infection) in patients with low-level parasitemia.

Both thin (Figs. 219-4 and 219-5) and thick (Figs. 219-6, 219-7, 219-8, and 219-9) blood smears should be examined. The thin blood smear should be air-dried, fixed in anhydrous methanol, and stained; the RBCs in the tail of the film should then be examined under oil immersion (×1000 magnification). The density of parasitemia is expressed as the number of parasitized erythrocytes per 1000 RBCs. The thick blood film should be of uneven thickness. The smear should be dried thoroughly and stained without fixing. As many layers of erythrocytes overlie one another and are lysed during the staining procedure, the thick film has the advantage of concentrating the parasites (by

40- to 100-fold compared with a thin blood film) and thus increasing diagnostic sensitivity. Both parasites and white blood cells (WBCs) are counted, and the number of parasites per unit volume is calculated from the total leukocyte count. Alternatively, a WBC count of 8000/μL is assumed. This figure is converted to the number of parasitized erythrocytes per microliter. A minimum of 200 WBCs should be counted under oil immersion. Interpretation of blood smears, particularly thick films, requires some experience because artifacts are common. Before a thick smear is judged to be negative, 100–200 fields should be examined. In high-transmission areas, the presence of up to 10,000 parasites/μL of blood may be tolerated without symptoms or signs in partially immune individuals. Thus, in these areas, the detection of low-density malaria parasitemia is sensitive but has low specificity in identifying malaria as the cause of illness. Because the prevalence of asymptomatic parasitemia is often high, low-density parasitemia is a common incidental finding in other conditions causing fever.

Rapid, simple, sensitive, and specific antibody-based diagnostic stick or card tests that detect *P. falciparum*–specific, histidine-rich protein 2 (PfHRP2), lactate dehydrogenase, or aldolase antigens in finger-prick blood samples are now being used widely in control programs **(Table 219-5)**. Some of these rapid diagnostic tests carry a second antibody (either pan-malaria or *P. vivax*–specific) and so distinguish falciparum malaria from the less dangerous malarias. PfHRP2-based tests may remain positive for several weeks after acute infection. This prolonged positivity is a disadvantage in high-transmission areas where infections are frequent, but it is of value in the diagnosis of severe malaria in patients who have taken antimalarial drugs and cleared peripheral parasitemia but who still have a strongly positive PfHRP2 test. A disadvantage of rapid tests is that they do not quantify parasitemia. Widespread use of PfHRP2 rapid tests has put strong selection pressure on *P. falciparum* populations in some areas, leading to an increased prevalence of mutant parasites that are not detected by the current generation of PfHRP2-based tests.

TABLE 219-5 Standard Methods for the Diagnosis of Malaria[a]

Method	Procedure	Advantages	Disadvantages
Thick blood film[b]	Blood should be uneven in thickness but thin enough that the hands of a watch can be read through part of the spot. Stain dried, unfixed blood spot with Giemsa, Field's, or another Romanowsky stain. Count number of asexual parasites per 200 WBCs (or per 500 at low densities). Count gametocytes separately.[c]	Sensitive (0.001% parasitemia); species specific; inexpensive	Requires experience (artifacts may be misinterpreted as low-level parasitemia); underestimates true count
Thin blood film[d]	Stain fixed smear with Giemsa, Field's, or another Romanowsky stain. Count number of RBCs containing asexual parasites per 1000 RBCs. In severe malaria, assess stage of parasite development and count neutrophils containing malaria pigment.[e] Count gametocytes separately.[c]	Rapid; species specific; inexpensive; in severe malaria, provides prognostic information[e]	Insensitive (<0.05% parasitemia); uneven distribution of *P. vivax*, as enlarged infected red cells concentrate at leading edge
PfHRP2 dipstick or card test	A drop of blood is placed on the stick or card, which is then immersed in washing solutions. Monoclonal antibody capture of parasitic antigens reads out as a colored band.	Robust and relatively inexpensive; rapid; sensitivity similar to or slightly lower than that of thick films (~0.001% parasitemia)	Detects only *Plasmodium falciparum*; remains positive for weeks after infection[f]; does not quantitate *P. falciparum* parasitemia; evasion of detection by certain strains due to polymorphisms in HRP2 gene
Plasmodium LDH dipstick or card test	A drop of blood is placed on the stick or card, which is then immersed in washing solutions. Monoclonal antibody capture of parasitic antigens reads out as two colored bands. One band is genus specific (all malarias), and the other is specific for *P. falciparum.*	Rapid; sensitivity similar to or slightly lower than that of thick films for *P. falciparum* (~0.001% parasitemia)	May miss low-level parasitemia with *P. vivax*, *P. ovale*, and *P. malariae* and may not speciate these organisms; does not quantitate *P. falciparum* parasitemia; lower sensitivity for detection of *P. knowlesi*, which may be misidentified as *P. falciparum*
Microtube concentration methods with acridine orange staining	Blood is collected in a specialized tube containing acridine orange, anticoagulant, and a float. After centrifugation, which concentrates the parasitized cells around the float, fluorescence microscopy is performed.	Sensitivity similar or superior to that of thick films (~0.001% parasitemia); ideal for processing large numbers of samples rapidly	Does not speciate or quantitate; requires fluorescence microscopy

[a]Malaria cannot be diagnosed clinically with accuracy, but treatment should be started on clinical grounds if laboratory confirmation is likely to be delayed. In areas of the world where malaria is endemic and transmission rates are high, low-level asymptomatic parasitemia is common in otherwise healthy people. Thus malaria may not be the cause of a fever, although in this context the presence of >10,000 parasites/μL (~0.2% parasitemia) does indicate that malaria is the cause. Antibody and polymerase chain reaction (PCR) tests have no role in the diagnosis of malaria except that PCR is increasingly used for genotyping and speciation in mixed infections and for detection of low-level parasitemia in asymptomatic residents of endemic areas. [b]Asexual parasites/200 WBCs × 40 = parasite count/μL (assumes a WBC count of 8000/μL). See Figs. 219-6, 219-7, 219-8, 219-9. [c]*P. falciparum* gametocytemia may persist for days or weeks after clearance of asexual parasites. Gametocytemia without asexual parasitemia does not indicate active infection. [d]Parasitized RBCs (%) × hematocrit × 1256 = parasite count/μL. See Figs. 219-4 and 219-5. [e]The presence of >100,000 parasites/μL (~2% parasitemia) is associated with an increased risk of severe malaria, but some patients have severe malaria with lower counts. At any level of parasitemia, the finding that >50% of parasites are tiny rings (cytoplasm thickness less than half of nucleus width) carries a relatively good prognosis. The presence of visible pigment in >20% of parasites or of phagocytosed pigment in >5% of polymorphonuclear leukocytes (indicating massive recent schizogony) carries a worse prognosis. [f]Persistence of PfHRP2 is a disadvantage in high-transmission settings, where many asymptomatic people have positive tests, but can be used to diagnostic advantage in low-transmission settings when a sick patient has previously received unknown treatment (which, in endemic areas, often consists of antimalarial drugs). In this situation, a positive PfHRP2 test indicates that the illness is falciparum malaria, even if the blood smear is negative.

Abbreviations: LDH, lactate dehydrogenase; PfHRP2, *P. falciparum* histidine-rich protein 2; RBCs, red blood cells; WBCs, white blood cells.

The relationship between parasite density and prognosis is complex; in general, patients with >10^5 parasites/μL are at increased risk of dying, but nonimmune patients may die with much lower counts, and partially immune persons may tolerate parasitemia levels many times higher with only minor symptoms. In severe malaria, a poor prognosis is indicated by a predominance of more mature *P. falciparum* parasites (i.e., >20% of parasites with visible pigment) in the peripheral-blood film or by the presence of phagocytosed malarial pigment in >5% of neutrophils (an indicator of recent schizogony). In *P. falciparum* infections, gametocytemia peaks 1 week after the peak of asexual parasite densities. Because the mature gametocytes of *P. falciparum* (unlike those of other plasmodia) are not affected by most antimalarial drugs, their persistence does not constitute evidence of drug resistance or a need to re-treat if a full course of appropriate antimalarial drugs has already been given. Phagocytosed malarial pigment seen inside peripheral-blood monocytes may provide a clue to recent infection if malaria parasites are not detectable. After parasite clearance, this intraphagocytic malarial pigment is often evident for several days in peripheral-blood films or for longer in bone marrow aspirates or smears of fluid expressed after intradermal puncture.

Molecular diagnosis by polymerase chain reaction (PCR) amplification of parasite nucleic acid is more sensitive than microscopy or rapid diagnostic tests for detecting malaria parasites and defining malarial species. While currently impractical in the standard clinical setting, PCR is used in reference centers in endemic areas. In epidemiologic surveys, ultrasensitive PCR detection may prove very useful in identifying asymptomatic infections as control and eradication programs drive parasite prevalences down to very low levels. Serologic diagnosis with either indirect fluorescent antibody or enzyme-linked immunosorbent assays is useful for screening of prospective blood donors and may prove useful as a measure of transmission intensity in future epidemiologic studies. It has no place in the diagnosis of acute illness.

LABORATORY FINDINGS IN ACUTE MALARIA

Normochromic, normocytic anemia is usual. The leukocyte count is generally normal, although it may be raised in very severe infections. There is slight monocytosis, lymphopenia, and eosinopenia, with reactive lymphocytosis and eosinophilia in the weeks after acute infection. The platelet count is usually reduced to ~10^5/μL. The erythrocyte sedimentation rate, plasma viscosity, and levels of C-reactive protein and other acute-phase proteins are elevated. Severe infections may be accompanied by prolonged prothrombin and partial thromboplastin times and by more severe thrombocytopenia. Antithrombin III levels are reduced even in mild infection. In uncomplicated malaria, plasma concentrations of electrolytes, blood urea nitrogen (BUN), and creatinine are usually normal. Findings in severe malaria may include metabolic acidosis, with low plasma concentrations of glucose, sodium, bicarbonate, phosphate, and albumin, together with elevations in lactate, BUN, creatinine, urate, muscle and liver enzymes, and conjugated and unconjugated bilirubin. Hypergammaglobulinemia is usual in immune and semi-immune subjects living in malaria-endemic areas. Urinalysis generally gives normal results. In adults and children with cerebral malaria, the mean cerebrospinal fluid (CSF) opening pressure at lumbar puncture is ~160 mm H_2O; usually the CSF content is normal or there is a slight elevation of total protein level (<1.0 g/L [<100 mg/dL]) and cell count (<20/μL).

TREATMENT

TREATMENT Malaria

Patients with severe malaria and those unable to take oral drugs should receive parenteral antimalarial therapy immediately **(Table 219-6)**. Antimalarial drug susceptibility testing can be performed but is rarely available, has poor predictive value in an individual case, and yields results too slowly to influence the choice of treatment. If there is any doubt about the resistance status of the infecting organism, it should be considered resistant.

The World Health Organization (WHO) recommends artemisinin-based combination therapy (ACT) as first-line treatment for uncomplicated falciparum malaria in malaria-endemic areas. ACT is also the recommended first-line treatment for *P. knowlesi* infections and is highly effective against the other malarias as well. The choice of an ACT partner drug depends on the likely sensitivity of the infecting parasites. Artemisinin-based combinations are sometimes unavailable in temperate countries, where treatment recommendations are limited to the registered available drugs. Despite increasing evidence of chloroquine resistance in *P. vivax* (from parts of Indonesia, Oceania, eastern and southern Asia, and Central and South America), chloroquine remains an effective treatment for *P. vivax* malaria in many areas and for *P. ovale* and *P. malariae* infections everywhere.

Artemisinin resistance in *P. falciparum* has emerged in Southeast Asia over the past decade and has been followed by piperaquine and mefloquine resistance. ACTs are starting to fail in Cambodia, Vietnam, and the border regions of Thailand. Significant artemisinin resistance is now prevalent throughout the Greater Mekong subregion but has not been reported from other malaria-endemic regions. Falsified or substandard antimalarial drugs are sold in many Asian and African countries and may be the cause of a failure to respond to therapy. Characteristics of antimalarial drugs are shown in **Table 219-7**.

SEVERE MALARIA

In large randomized controlled clinical trials, parenteral artesunate, a water-soluble artemisinin derivative, has reduced mortality rates in severe falciparum malaria among Asian adults and children by 35% and among African children by 22.5% compared with quinine treatment. Artesunate therefore is now the drug of choice for all patients with severe malaria everywhere. Artesunate is given by IV injection but is also absorbed rapidly following IM injection. Artemether and the closely related drug artemotil (arteether) are oil-based formulations given by IM injection; they are erratically absorbed and do not confer the same survival benefit as artesunate. A rectal formulation of artesunate has been developed as a community-based pre-referral treatment for patients in the rural tropics who cannot

take oral medications. Pre-referral administration of rectal artesunate has been shown to decrease mortality rates among severely ill children without access to immediate parenteral treatment. Although the artemisinin compounds are safer than quinine and considerably safer than quinidine, only one formulation is available in the United States. IV artesunate has been approved by the U.S. Food and Drug Administration for emergency use in severe malaria and can be obtained through the Centers for Disease Control and Prevention (CDC) Drug Service (see end of chapter for contact information). The antiarrhythmic quinidine gluconate was used to treat severe malaria in the United States previously but is now in short supply; artesunate is much more effective and safer. Parenteral quinidine is potentially dangerous and must be closely monitored if dysrhythmias and hypotension are to be avoided. If total plasma levels exceed 8 μg/mL, if the QT_c interval exceeds 0.6 s, or if the QRS complex widens by more than 25% over baseline, then infusion rates should be slowed or infusion stopped temporarily. If arrhythmia or saline-unresponsive hypotension develops, treatment with this drug should be discontinued. Quinine is safer than quinidine; cardiovascular monitoring is not required except when the recipient has cardiac disease. Although parenteral quinine is steadily being replaced by parenteral artesunate in endemic areas, it still has a role in the very few cases of artemisinin-resistant severe falciparum malaria from Southeast Asia, where both artesunate and quinine are given together in full doses.

Severe falciparum malaria constitutes a medical emergency requiring intensive nursing care and careful management. Frequent evaluation of the patient's condition is essential. Adjunctive treatments such as high-dose glucocorticoids, urea, heparin, dextran, desferrioxamine, antibody to tumor necrosis factor α, high-dose phenobarbital (20 mg/kg), mannitol, or large-volume fluid or albumin boluses have proved either ineffective or harmful in clinical trials and should not be used. In acute renal failure or severe metabolic acidosis, hemofiltration or hemodialysis should be started as early as possible.

In severe malaria, parenteral antimalarial treatment should be started immediately. Artesunate, given by either IV or IM injection, is the treatment of choice; it is simple to administer, very safe, and rapidly effective. It does not require dose adjustments in liver dysfunction or renal failure. It should be used in pregnant women with severe malaria. If artesunate is unavailable and artemether, quinine, or quinidine is used, an initial loading dose must be given so that therapeutic concentrations are reached as soon as possible. Both quinine and quinidine will cause dangerous hypotension if injected rapidly; when given IV, they must be administered carefully by rate-controlled infusion only. If this approach is not possible, quinine may be given by deep IM injections into the anterior thigh. The optimal therapeutic ranges for quinine and quinidine in severe malaria are not known with certainty, but total plasma concentrations of 8–15 mg/L for quinine and 3.5–8.0 mg/L for quinidine are effective and do not cause serious toxicity. The systemic clearance and apparent volume of distribution of these alkaloids are markedly reduced and plasma protein binding is increased in severe malaria, so that the blood concentrations attained with a given dose are higher. If the patient remains seriously ill or in acute renal failure for >2 days, maintenance doses of quinine or quinidine should be reduced by 30–50% to prevent toxic accumulation of the drug. The initial doses should never be reduced. If safe and feasible, exchange transfusion may be considered for patients with severe malaria, although the precise indications for this procedure have not been agreed upon and there is no clear evidence that this measure is beneficial, particularly with artesunate treatment. Convulsions should be treated promptly with IV (or rectal) benzodiazepines. The role of prophylactic anticonvulsants in children is uncertain. If respiratory support is not available, a full loading dose of phenobarbital (20 mg/kg) to prevent convulsions should *not* be given as it may cause respiratory arrest.

When the patient is unconscious, the blood glucose level should be measured every 4–6 h. All patients should receive a continuous infusion of dextrose, and blood concentrations ideally should be maintained above 4 mmol/L. Hypoglycemia (<2.2 mmol/L or 40 mg/dL) should be treated immediately with bolus glucose. The parasite count and hematocrit should be measured every 6–12 h. Anemia develops rapidly; if the hematocrit falls to <20%, whole blood (preferably fresh) or packed cells should be transfused slowly, with careful attention to circulatory status. In areas with higher malaria

transmission, where blood for transfusion is in short supply, a threshold of 15% is widely used. Renal function should be checked at least daily. Children presenting with severe anemia and acidotic breathing require immediate blood transfusion. Accurate assessment is vital. Management of fluid balance is difficult in severe malaria, particularly in adults, because of the thin dividing line between overhydration (leading to pulmonary edema) and underhydration (contributing to renal impairment). Fluid balance management is *different* from that in sepsis: fluid boluses are potentially dangerous in severe malaria. Nasogastric feeding should be delayed in non-intubated patients (for 60 h in adults and 36 h in children) to reduce the risk of aspiration pneumonia. As soon as the patient can take fluids, oral therapy should be substituted for parenteral treatment and a full 3-day course of ACT given. Mefloquine should be avoided as follow-on treatment for severe malaria because of the increased risk of post-malaria neurologic syndrome.

In areas of high transmission of both *P. falciparum* and *P. vivax* (the island of New Guinea), severe and potentially life-threatening anemia is common among children, and both species contribute. Elsewhere, severe vivax malaria may occur but is uncommon. Many patients have had comorbidities contributing to vital-organ dysfunction.

P. knowlesi can cause severe disease associated with high parasite densities. Acute kidney injury, respiratory distress, and shock have all been described, but cerebral malaria does not occur. Treatment for severe vivax and knowlesi malaria should follow the recommendations given for falciparum malaria.

UNCOMPLICATED MALARIA

P. falciparum and *P. knowlesi* infections should be treated with an artemisinin-based combination because of their propensity for high parasite densities and severe disease. Infections with sensitive strains of *P. vivax*, *P. malariae*, and *P. ovale* should be treated with an artemisinin-based combination or oral chloroquine (total dose, 25 mg of base/kg). The ACT regimens now recommended are safe and effective in adults, children, and pregnant women. The rapidly eliminated artemisinin component is usually an artemisinin derivative (artesunate, artemether, or dihydroartemisinin) given for 3 days, and the partner drug is usually a more slowly eliminated antimalarial to which *P. falciparum* in the area is sensitive. Five ACT regimens are currently recommended by the WHO: artemether-lumefantrine, artesunate-mefloquine, dihydroartemisinin-piperaquine, artesunate-sulfadoxine-pyrimethamine, and artesunate-amodiaquine. There is increasing evidence for both the efficacy and the safety of artesunate-pyronaridine. In areas of low malaria transmission, a single dose of primaquine (0.25 mg/kg) should be added to ACT as a *P. falciparum* gametocytocide to reduce the transmissibility of the infection. This low dose of primaquine is safe even in G6PD deficiency. Pregnant women should not be given primaquine. Atovaquone-proguanil is highly effective everywhere, although it is seldom used in endemic areas because of its high cost and the propensity for rapid emergence of resistance. Recovery is slower after atovaquone-proguanil treatment than after ACT. Of great concern is the emergence of artemisinin-resistant *P. falciparum* in the Greater Mekong subregion of Southeast Asia. Infections with these parasites are cleared slowly from the blood, with clearance times typically exceeding 3 days, and cure rates with ACT have fallen to unacceptably low levels in some areas. Extended treatment courses and triple antimalarial combinations are under evaluation.

The 3-day ACT regimens are all well tolerated, although mefloquine is associated with increased rates of vomiting and dizziness. As second-line treatments for recrudescence following first-line therapy, a different ACT regimen may be given; another alternative is a 7-day course of either artesunate or quinine plus tetracycline, doxycycline, or clindamycin. Tetracycline and doxycycline cannot be given to pregnant women after 15 weeks of gestation or to children <8 years of age. Oral quinine is extremely bitter and regularly produces cinchonism comprising tinnitus, high-tone deafness, nausea, vomiting, and dysphoria. Clinical responses are slower than those following ACT. Adherence is poor with the required 7-day regimens of quinine.

Patients should be monitored for vomiting for 1 h after the administration of any oral antimalarial drug. If there is vomiting, the dose should be repeated. Symptom-based treatment, with tepid sponging and acetaminophen (paracetamol) administration, lowers fever and thereby reduces the patient's propensity to vomit these drugs. Minor central nervous system reactions (nausea, dizziness, sleep disturbances) are common. The incidence of serious adverse neuropsychiatric reactions to mefloquine treatment is ~1 in 1000 in Asia but may be as high as 1 in 200 among Africans and Caucasians. All the antimalarial quinolines (chloroquine, mefloquine, and quinine) exacerbate the orthostatic hypotension associated with malaria, and all are tolerated better by children than by adults. Pregnant women, young children, patients unable to tolerate oral therapy, and nonimmune individuals (e.g., travelers) with suspected malaria should be evaluated carefully and hospitalization considered. If there is any doubt as to the identity of the infecting malarial species, treatment for falciparum malaria should be given. A negative blood smear read by an experienced microscopist makes malaria very unlikely but does not rule it out completely; thick blood films should be checked again 1 and 2 days later to exclude the diagnosis. Nonimmune patients receiving treatment for malaria should have daily parasite counts performed until the thick films are negative. If the level of parasitemia does not fall below 25% of the admission value in 72 h or if parasitemia has not cleared by 7 days (and adherence is assured), drug resistance is likely and the regimen should be changed.

To eradicate persistent liver stages and prevent relapse (radical treatment), primaquine (0.5 mg of base/kg in East Asia and Oceania and 0.25 mg/kg elsewhere) should be given once daily for 14 days to patients with *P. vivax* or *P. ovale* infection after laboratory tests for G6PD deficiency have proved negative. If the patient has a mild variant of G6PD deficiency, primaquine can be given in a dose of 0.75 mg of base/kg (maximum, 45 mg) once weekly for 8 weeks. Pregnant women with vivax or ovale malaria should not be given primaquine but should receive suppressive prophylaxis with chloroquine (5 mg of base/kg per week) until delivery, after which radical treatment can be given.

MANAGEMENT OF COMPLICATIONS Acute Renal Failure

If plasma levels of BUN or creatinine rise despite adequate rehydration, fluid administration should be restricted to prevent volume overload. As in other forms of hypercatabolic acute renal failure, renal replacement therapy is best performed early **(Chap. 304)**. Hemofiltration and hemodialysis are more effective than peritoneal dialysis and are associated with lower mortality risk. Some patients with renal impairment pass small volumes of urine sufficient to allow control of fluid balance; these cases can be managed conservatively if other indications for dialysis do not arise. Renal function usually improves within days, but full recovery may take weeks.

Acute Pulmonary Edema (Acute Respiratory Distress Syndrome)

This syndrome is caused by increased pulmonary capillary permeability. Patients should be positioned with the head of the bed at a 45° elevation and should be given oxygen and IV diuretics. Positive-pressure ventilation should be started early if the immediate measures fail **(Chap. 298)**. Rarely, patients may require extracorporeal membrane oxygenation.

Hypoglycemia

An initial slow injection of 20% dextrose (2 mL/kg over 10 min) should be followed by an infusion of 10% dextrose (0.10 g/kg per hour). The blood glucose level should be checked regularly thereafter as recurrent hypoglycemia is common, particularly among patients receiving quinine or quinidine. In severely ill patients, hypoglycemia commonly occurs together with metabolic (lactic) acidosis and carries a poor prognosis.

Sepsis

Hypoglycemia or gram-negative septicemia should be suspected when the condition of any patient suddenly deteriorates for no obvious reason during antimalarial treatment. In malaria-endemic areas

where a high proportion of children are parasitemic, it is usually impossible to distinguish severe malaria from bacterial sepsis with confidence. These children should be treated with both antimalarials and broad-spectrum antibiotics from the outset. Because infections with nontyphoidal *Salmonella* species are particularly common, empirical antibiotics should be selected to cover these organisms. Antibiotics should be considered for severely ill patients of any age who are not responding to antimalarial treatment.

Other Complications

Patients who develop spontaneous bleeding should be given fresh blood and IV vitamin K. Convulsions should be treated with IV or rectal benzodiazepines and, if necessary, respiratory support. Aspiration pneumonia should be suspected in any unconscious patient with convulsions, particularly with persistent hyperventilation; IV antimicrobial agents and oxygen should be administered, and pulmonary toilet should be undertaken.

TABLE 219-6 Regimens for the Treatment of Malaria[a]

Type of Disease or Treatment	Regimen(S)
Uncomplicated Malaria	
Known chloroquine-sensitive strains of *Plasmodium vivax*, *P. malariae*, *P. ovale*, *P. falciparum*[b]	Chloroquine (10 mg of base/kg stat followed by 5 mg/kg at 12, 24, and 36 h or by 10 mg/kg at 24 h and 5 mg/kg at 48 h) ***or*** Amodiaquine (10–12 mg of base/kg qd for 3 days)
Radical treatment for *P. vivax or P. ovale* infection	In addition to chloroquine or amodiaquine as detailed above or ACT as detailed below, primaquine (0.5 mg of base/kg qd in Southeast Asia and Oceania and 0.25 mg/kg elsewhere) should be given for 14 days to prevent relapse. In mild G6PD deficiency, 0.75 mg of base/kg should be given once weekly for 8 weeks. Primaquine should not be given in severe G6PD deficiency.
P. falciparum malaria[c]	Artesunate[d] (4 mg/kg qd for 3 days) *plus* sulfadoxine (25 mg/kg)/pyrimethamine (1.25 mg/kg) as a single dose ***or*** Artesunate[d] (4 mg/kg qd for 3 days) *plus* amodiaquine (10 mg of base/kg qd for 3 days)[e] ***or*** Artemether-lumefantrine[d] (1.5/9 mg/kg bid for 3 days with food) ***or*** Artesunate[d] (4 mg/kg qd for 3 days) *plus* mefloquine (24–25 mg of base/kg—either 8 mg/kg qd for 3 days or 15 mg/kg on day 2 and then 10 mg/kg on day 3)[e] ***or*** DHA-piperaquine[d] (target dose: 4/24 mg/kg qd for 3 days in children weighing <25 kg and 4/18 mg/kg qd for 3 days in persons weighing ≥25 kg)
Second-line treatment/treatment of imported malaria	Artesunate[d] (2 mg/kg qd for 7 days) *or* quinine (10 mg of salt/kg tid for 7 days) *plus 1 of the following 3:* 1. Tetracycline[f] (4 mg/kg qid for 7 days) 2. Doxycycline[f] (3 mg/kg qd for 7 days) 3. Clindamycin (10 mg/kg bid for 7 days) ***or*** Atovaquone-proguanil (20/8 mg/kg qd for 3 days with food)
Severe Falciparum Malaria[g,h]	

Type of Disease or Treatment	Regimen(S)
	Artesunate[d] (2.4 mg/kg stat IV followed by 2.4 mg/kg at 12 and 24 h and then daily if necessary; for children weighing <20 kg, give 3 mg/kg per dose)[h] ***or, if unavailable,*** Artemether[d] (3.2 mg/kg stat IM followed by 1.6 mg/kg qd) ***or, if unavailable,*** Quinine dihydrochloride (20 mg of salt/kg[i] infused over 4 h, followed by 10 mg of salt/kg infused over 2–8 h q8h[j]) ***or, if none of the above are available,*** Quinidine (10 mg of base/kg[i] infused over 1–2 h, followed by 1.2 mg of base/kg per hour[j] with electrocardiographic monitoring)

[a]In endemic areas where malaria transmission is low, except in pregnant women and infants, a single dose of primaquine (0.25 mg of base/kg) should be added as a gametocytocide to all falciparum malaria treatments to prevent transmission. This addition is considered safe, even in G6PD deficiency. [b]Very few areas now have chloroquine-sensitive *P. falciparum* malaria. [c]In areas where the partner drug to artesunate is known to be effective. [d]Artemisinin derivatives are not readily available in some temperate countries. [e]Fixed-dose co-formulated combinations are available. The World Health Organization now recommends artemisinin combination regimens as first-line therapy for falciparum malaria in all tropical countries and advocates use of fixed-dose combinations. [f]Tetracycline and doxycycline should not be given to pregnant women after 15 weeks of gestation or to children <8 years of age. [g]Oral treatment should be substituted as soon as the patient recovers sufficiently to take fluids by mouth. [h]Artesunate is the drug of choice when available. The data from large studies in Southeast Asia showed a 35% lower mortality rate than with quinine, and very large studies in Africa showed a 22.5% reduction in mortality rate compared with quinine. The doses of artesunate in children weighing <20 kg should be 3 mg/kg. [i]A loading dose should not be given if therapeutic doses of quinine or quinidine have definitely been administered in the previous 24 h. Some authorities recommend a lower dose of quinidine. [j]Infusions can be given in 0.9% saline and 5–10% dextrose in water. Infusion rates for quinine and quinidine should be carefully controlled.

Abbreviations: ACT, artemisinin combination therapy; DHA, dihydroartemisinin; G6PD, glucose-6-phosphate dehydrogenase.

TABLE 219-7 Properties of Antimalarial Drugs

Drug(S)	Pharmacokinetic Properties	Antimalarial Activity	Minor Toxicity	Major Toxicity
Quinine, quinidine	Good oral and IM absorption (quinine); *Cl* and V_d reduced, but plasma protein binding (principally to α1 acid glycoprotein) increased (90%) in malaria; quinine $t_{1/2}$: 16 h in malaria, 11 h in healthy persons; quinidine $t_{1/2}$: 13 h in malaria, 8 h in healthy persons	Acts mainly on trophozoite blood stage; kills gametocytes of *P. vivax*, *P. ovale*, and *P. malariae* (but not *P. falciparum*); no action on liver stages	*Common:* cinchonism (tinnitus, high-tone hearing loss, nausea, vomiting, dysphoria, postural hypotension); ECG QT interval prolongation (quinine usually by <10% but quinidine by up to 25%). *Rare:* diarrhea, visual disturbance, rashes. *Note:* very bitter taste	*Common:* hypoglycemia. *Rare:* hypotension, blindness, deafness, cardiac arrhythmias, thrombocytopenia, hemolysis, hemolytic-uremic syndrome, vasculitis, cholestatic hepatitis, neuromuscular paralysis. *Note:* quinidine more cardiotoxic
Chloroquine	Good oral absorption, very rapid IM and SC absorption; complex pharmacokinetics; enormous *Cl* and V_d (unaffected by malaria); blood concentration profile determined by distribution processes in malaria; $t_{1/2}$: 1–2 months	As for quinine, but acts slightly earlier in asexual cycle	*Common:* nausea, dysphoria, pruritus in dark-skinned patients, postural hypotension, ECG QT prolongation. *Rare:* accommodation difficulties, keratopathy, rash. *Note:* bitter taste but usually well tolerated	*Acute:* hypotensive shock (parenteral), cardiac arrhythmias, neuropsychiatric reactions. *Chronic:* retinopathy (cumulative dose, >100 g), skeletal and cardiac myopathy
Piperaquine	Adequate oral absorption, may be enhanced by fats; similar pharmacokinetics to chloroquine; $t_{1/2}$: 21–28 days	As for chloroquine; retains activity against multidrug-resistant *P. falciparum*, but resistance has emerged in Southeast Asia	Occasional epigastric pain, diarrhea, ECG QT_c prolongation	None identified
Amodiaquine	Good oral absorption; largely converted to active metabolite desethylamodiaquine; $t_{1/2}$: 4–5 days	As for chloroquine, but more active against chloroquine-resistant *P. falciparum*	Nausea (tastes better than chloroquine), dysphoria, headache, ECG QTc prolongation	Agranulocytosis; hepatitis, mainly with prophylactic use; should not be used with efavirenz

Drug(S)	Pharmacokinetic Properties	Antimalarial Activity	Minor Toxicity	Major Toxicity
Primaquine	Complete oral absorption; active metabolite produced via CYP2D6; $t_{1/2}$: 5–7 h	Radical cure; eradicates hepatic forms of *P. vivax* and *P. ovale;* kills all stages of *P. falciparum* gametocyte development; kills developing liver stages of all species	Nausea, vomiting, diarrhea, abdominal pain, hemolysis, methemoglobinemia	Serious hemolytic anemia, severe G6PD deficiency; hemoglobinuria
Mefloquine	Adequate oral absorption; no parenteral preparation; $t_{1/2}$: 14–20 days (shorter in malaria)	As for quinine	Nausea, giddiness, dysphoria, fuzzy thinking, sleeplessness, nightmares, sense of dissociation	Neuropsychiatric reactions, convulsions, encephalopathy
Lumefantrine	Highly variable absorption related to fat intake; $t_{1/2}$: 3–4 days	As for quinine	None identified	None identified
Artemisinin and derivatives (artemether, artesunate)	Good oral absorption; good absorption of IM artesunate but slow and variable absorption of IM artemether; artesunate and artemether biotransformed to active metabolite dihydroartemisinin; all drugs eliminated very rapidly; $t_{1/2}$: <1 h	Broader stage specificity and more rapid than other drugs; no action on liver stages; kills all but fully mature gametocytes of *P. falciparum*	Reduction in reticulocyte count (but not anemia); neutropenia at high doses; in some cases, delayed anemia after treatment of severe malaria with hyperparasitemia	Anaphylaxis, urticaria, fever
Pyrimethamine	Good oral absorption, variable IM absorption; $t_{1/2}$: 4 days	For blood stages, acts mainly on mature forms; causal prophylactic	Well tolerated	Megaloblastic anemia, pancytopenia, pulmonary infiltration
Proguanil[a] (chloroguanide)	Good oral absorption; biotransformed to active metabolite cycloguanil; $t_{1/2}$: 16 h; biotransformation reduced by oral contraceptive use and in pregnancy	Causal prophylactic; not used alone for treatment	Well tolerated; mouth ulcers and rare alopecia	Megaloblastic anemia in renal failure

Drug(S)	Pharmacokinetic Properties	Antimalarial Activity	Minor Toxicity	Major Toxicity
Atovaquone[a]	Highly variable absorption related to fat intake; $t_{1/2}$: 30–70 h	Acts mainly on trophozoite blood stage	None identified	None identified
Tetracycline, doxycycline[b]	Excellent absorption; $t_{1/2}$: 8 h for tetracycline, 18 h for doxycycline	Weak antimalarial activity; should not be used alone for treatment	Gastrointestinal intolerance, deposition in growing bones and teeth, photosensitivity, moniliasis, benign intracranial hypertension	Renal failure in patients with impaired renal function (tetracycline)
Pyronaridine	Rapid variable absorption, large V_d; $t_{1/2}$: 12–14 days	Acts mainly on trophozoite blood stage; kills gametocytes of *P. vivax*, *P. ovale*, and *P. malariae* (but not *P. falciparum*); no action on liver stages	Gastrointestinal intolerance, anemia, transient elevation of aminotransferases, hypoglycemia, headache	None identified
Arterolane	$t_{1/2}$: 3 h	Broad stage specificity; no action on liver stages; kills all but fully mature gametocytes of *P. falciparum*	Gastrointestinal intolerance, transient elevation of aminotransferases	None identified

[a]Atovaquone and proguanil are prescribed as a fixed-dose combination. This and proguanil alone should not be given if the estimated glomerular filtration rate is <30 mL/min. [b]Tetracycline and doxycycline should not be given to pregnant women after 15 weeks of gestation or to children <8 years of age.

Abbreviations: Cl, systemic clearance; ECG, electrocardiogram; G6PD, glucose-6-phosphate dehydrogenase; V_d, total apparent volume of distribution.

GLOBAL CONSIDERATIONS

In recent years, considerable progress has been made in malaria prevention and control. Distribution of insecticide-treated bed nets (ITNs) has been shown to reduce all-cause mortality in African children by 20%. New drugs have been discovered and are being developed, and one vaccine candidate (the RTS,S/AS01 vaccine) has been licensed for use. Highly effective drugs, long-lasting ITNs, and insecticides for anopheline vector control are being purchased for endemic countries by international donors. The WHO now calls for all countries to work toward a goal of malaria elimination, and many countries have set ambitious timelines to achieve this goal. Success will require strong

leadership, increased national commitment, and international support. The numerous challenges that lie ahead include the widespread distribution of *Anopheles* breeding sites, the enormous number of infected persons, the emergence and spread of resistance in *P. falciparum* to common artemisinin-based combinations in Southeast Asia, increasing insecticide resistance and behavioral changes (to avoid ITN contact) in anopheline mosquito vectors, and inadequacies in human and material resources, infrastructure, and control programs. Eliminating vivax malaria is further hindered by the lack of a simple, safe radical curative regimen.

MALARIA PREVENTION

Malaria may be contained by judicious use of insecticides to kill the mosquito vector, rapid diagnosis, patient management, and—where effective and feasible—administration of intermittent preventive treatments, seasonal malaria chemoprevention, or chemoprophylaxis to high-risk groups such as pregnant women and young children. Focal elimination of *P. falciparum* can be accelerated by mass treatment with slowly eliminated antimalarials such as dihydroartemisinin-piperaquine. Despite the enormous investment in efforts to develop a malaria vaccine, no safe, highly effective, long-lasting vaccine is likely to be available for general use in the near future **(Chap. 118)**. The licensed recombinant protein sporozoite-targeted adjuvanted vaccine RTS,S was only moderately efficacious in protecting African children from malaria in field trials, and protection of the very youngest recipients waned to 16% only 4 years after vaccination. The vaccine will be deployed in Ghana, Kenya, and Malawi as part of a large-scale pilot project before a decision on its more general use is taken. An irradiated live sporozoite vaccine is in late-stage development, and research is ongoing to develop a vaccine to protect against placental malaria (targeting VAR2CSA). While there is great promise for one or several malaria vaccines on the more distant horizon, prevention and control measures will continue to rely on antivector and drug-use strategies for the foreseeable future.

PERSONAL PROTECTION AGAINST MALARIA

Simple measures to reduce the frequency of bites by infected mosquitoes in malarious areas are very important. These measures include the avoidance of exposure to mosquitoes at their peak feeding times (usually dusk to dawn) and the use of insect repellents containing 10–35% DEET (or, if DEET is unacceptable, 7% picaridin), suitable clothing, and ITNs or other insecticide-impregnated materials. Widespread use of bed nets treated with residual pyrethroids reduces the incidence of malaria in areas where vectors bite indoors at night.

CHEMOPROPHYLAXIS

(Table 219-8; *https://wwwnc.cdc.gov/travel/yellowbook/2018/infectious-diseases-related-to-travel/malaria*) Recommendations for malaria prophylaxis depend on knowledge of local patterns of drug sensitivity in *Plasmodium* species and the likelihood of acquiring malarial infection. When there is uncertainty, drugs effective against resistant *P. falciparum* should be used (atovaquone-proguanil [Malarone], doxycycline, or mefloquine). Chemoprophylaxis is never entirely reliable, and malaria should always be considered in the differential diagnosis of fever in patients who have traveled to endemic areas, even if they are taking prophylactic antimalarial drugs.

TABLE 219-8 Drugs Used in the Prophylaxis of Malaria

Drug	Usage	Adult Dose	Pediatric Dose	Comments
Atovaquone-proguanil (Malarone)	Prophylaxis in areas with chloroquine- or mefloquine-resistant *Plasmodium falciparum*	1 adult tablet PO[a]	5–8 kg: ½ pediatric tablet[b] daily ≥8–10 kg: ¾ pediatric tablet daily ≥10–20 kg: 1 pediatric tablet daily ≥20–30 kg: 2 pediatric tablets daily ≥30–40 kg: 3 pediatric tablets daily ≥40 kg: 1 adult tablet daily	Begin 1–2 days before travel to malarious areas. Take daily at the same time each day while in the malarious areas and for 7 days after leaving such areas. Atovaquone-proguanil is contraindicated in persons with severe renal impairment (creatinine clearance rate, <30 mL/min). In the absence of data, it is not recommended for children weighing <5 kg, pregnant women, or women breast-feeding infants weighing <5 kg. Atovaquone-proguanil should be taken with food or a milky drink.
Chloroquine phosphate (Aralen and generic)	Prophylaxis only in areas with chloroquine-sensitive *P. falciparum*[c] or areas with *P. vivax* only	300 mg of base (500 mg of salt) PO once weekly	5 mg of base/kg (8.3 mg of salt/kg) PO once weekly, up to maximum adult dose of 300 mg of base	Begin 1–2 weeks before travel to malarious areas. Take weekly on the same day of the week while in the malarious areas and for 4 weeks after leaving such areas. Chloroquine phosphate may exacerbate psoriasis.

Drug	Usage	Adult Dose	Pediatric Dose	Comments
Doxycycline (many brand names and generic)	Prophylaxis in areas with chloroquine- or mefloquine-resistant *P. falciparum*[c]	100 mg PO qd (except in pregnant women; see Comments)	≥8 years of age: 2 mg/kg, up to adult dose	Begin 1–2 days before travel to malarious areas. Take daily at the same time each day while in the malarious areas and for 4 weeks after leaving such areas. Doxycycline is contraindicated in children aged <8 years and in pregnant women after 15 weeks of gestation.
Hydroxychloroquine sulfate (Plaquenil)	An alternative to chloroquine for primary prophylaxis only in areas with chloroquine-sensitive *P. falciparum*[c] or areas with *P. vivax* only	310 mg of base (400 mg of salt) PO once weekly	5 mg of base/kg (6.5 mg of salt/kg) PO once weekly, up to maximum adult dose of 310 mg of base	Begin 1–2 weeks before travel to malarious areas. Take weekly on the same day of the week while in the malarious areas and for 4 weeks after leaving such areas. Hydroxychloroquine may exacerbate psoriasis.

Drug	Usage	Adult Dose	Pediatric Dose	Comments
			≤9 kg: 4.6 mg of base/kg (5 mg of salt/kg) PO once weekly	
Mefloquine (Lariam and generic)	Prophylaxis in areas with chloroquine-resistant *P. falciparum*[c]	228 mg of base (250 mg of salt) PO once weekly	10–19 kg: ¼ tablet[d] once weekly 20–30 kg: ½ tablet once weekly 31–45 kg: ¾ tablet once weekly ≥46 kg: 1 tablet once weekly	Begin 1–2 weeks before travel to malarious areas. Take weekly on the same day of the week while in the malarious areas and for 4 weeks after leaving such areas. Mefloquine is contraindicated in persons allergic to this drug or related compounds (e.g., quinine and quinidine) and in persons with active or recent depression, generalized anxiety disorder, psychosis, schizophrenia, other major psychiatric disorders, or seizures. Use with caution in persons with psychiatric disturbances or a history of depression. Mefloquine is not recommended for persons with cardiac conduction abnormalities.
Primaquine	For prevention of malaria in areas with mainly *P. vivax*	30 mg of base (52.6 mg of salt) PO qd	0.5 mg of base/kg (0.8 mg of salt/kg) PO qd, up to adult dose; should be taken with food	Begin 1–2 days before travel to malarious areas. Take daily at the same time each day while in the malarious areas and for 7 days after leaving such areas. Primaquine is contraindicated in persons with G6PD deficiency. It is also contraindicated during pregnancy.

Drug	Usage	Adult Dose	Pediatric Dose	Comments
Primaquine	Used for presumptive anti-relapse therapy (terminal prophylaxis) to decrease risk of relapses of *P. vivax* and *P. ovale*	30 mg of base (52.6 mg of salt) PO qd for 14 days after departure from the malarious area	0.5 mg of base/kg (0.8 mg of salt/kg), up to adult dose, PO qd for 14 days after departure from the malarious area	This therapy is indicated for persons who have had prolonged exposure to *P. vivax* and/or *P. ovale*. It is contraindicated in persons with G6PD deficiency as well as during pregnancy.

[a]An adult tablet contains 250 mg of atovaquone and 100 mg of proguanil hydrochloride. [b]A pediatric tablet contains 62.5 mg of atovaquone and 25 mg of proguanil hydrochloride. [c]Very few areas now have chloroquine-sensitive malaria. [d]One tablet contains 228 mg of base (250 mg of salt).

Abbreviation: G6PD, glucose-6-phosphate dehydrogenase.

Source: CDC: *www.cdc.gov/malaria/travelers/drugs.html*.

Pregnant women planning to visit malarious areas should be warned about the potential risks and advised to avoid all nonessential travel. All pregnant women who live in endemic areas should be encouraged to attend regular antenatal clinics. Mefloquine is the only drug advised for pregnant women traveling to areas with drug-resistant malaria; this drug is generally considered safe in the second and third trimesters of pregnancy; the data on first-trimester exposure, although limited, are reassuring. Chloroquine and proguanil are regarded as safe, but there are now very few regions where these drugs can be recommended for protection. Doxycycline may be given until 15 weeks of pregnancy, at which point it should be discontinued. The safety of other prophylactic antimalarial agents in pregnancy has not been established. Antimalarial prophylaxis has been shown to reduce mortality rates among children between the ages of 3 months and 4 years in malaria-endemic areas; however, it is not a logistically or economically feasible option in many countries. The alternative—to give intermittent preventive treatment (IPT) to pregnant women, and in some areas to infants as well, or seasonal malaria chemoprevention (SMC) to young children—is being implemented. Other strategies are being evaluated, such as intermittent screening and treatment.

IPT in pregnancy (IPTp) involves giving treatment doses of sulfadoxine-pyrimethamine at each antenatal visit (maximum, once monthly) in the second and third trimesters of pregnancy. Women with HIV infection who are taking trimethoprim-sulfamethoxazole as prophylaxis should not be given concomitant sulfadoxine-pyrimethamine. IPT in infancy (IPTi) involves giving treatment doses of sulfadoxine-pyrimethamine along with the immunizations included in the WHO's Expanded Program on Immunization at 2, 3, and 9 months of life. Seasonal malaria chemoprevention involves giving monthly doses of amodiaquine and sulfadoxine-pyrimethamine to children 3 months to 5 years of age during the 3- to 4-month rainy season across the Sahel region of Africa. Children born to nonimmune mothers in malaria-endemic areas (usually expatriates moving to these areas) should receive prophylaxis from birth.

Travelers should start taking antimalarial drugs 2 days to 2 weeks before departure so that any untoward reactions can be detected before travel and so that therapeutic antimalarial blood concentrations will be present if and when any infections develop (Table 219-8). Antimalarial prophylaxis should continue for 4 weeks after the traveler has left the endemic area, except if atovaquone-proguanil or primaquine has been taken; these drugs have significant activities against the liver stage of the infection (causal prophylaxis) and can be discontinued 1 week after departure from the endemic area. If suspected malaria develops while a traveler is abroad, obtaining a reliable diagnosis and antimalarial treatment locally is a top priority. Presumptive self-treatment for malaria with atovaquone-proguanil (for 3 consecutive days) or one of the artemisinin-based combinations can be considered under special circumstances; medical advice on self-treatment should be sought before departure for malaria-endemic areas and as soon as possible after illness begins. Every effort should be made to confirm the diagnosis.

Atovaquone-proguanil (Malarone; 3.75/1.5 mg/kg or 250/100 mg, daily adult dose) is a fixed-combination, once-daily prophylactic agent that is very well tolerated by adults and children. This combination is effective against all types of malaria, including multidrug-resistant falciparum malaria. Atovaquone-proguanil is best taken with food or a milky drink to optimize absorption. It is not recommended if the estimated glomerular filtration rate is <30 mL/min. There are insufficient data on the safety of this regimen in pregnancy.

Mefloquine (250 mg of salt weekly, adult dose) has been widely used for malarial prophylaxis because it is usually effective against multidrug-resistant falciparum malaria and is reasonably well tolerated. Mefloquine has been associated with rare episodes of psychosis and seizures at prophylactic doses; these reactions are more frequent at the higher doses used for treatment. More common side effects with prophylactic doses of mefloquine include mild nausea, dizziness, fuzzy thinking, disturbed sleep patterns, vivid dreams, dysphoria, and malaise. Mefloquine is contraindicated for use by travelers with known hypersensitivity and by persons with active or recent depression, anxiety disorder, psychosis, schizophrenia, another major psychiatric disorder, or seizures; it is not recommended for persons with cardiac conduction abnormalities although the evidence that it is cardiotoxic is very weak. Confidence is increasing with regard to the safety of mefloquine prophylaxis during pregnancy; in studies in Africa, mefloquine prophylaxis was found to be effective and safe during pregnancy. Daily administration of doxycycline (100 mg daily, adult dose) is an effective alternative to atovaquone-proguanil or mefloquine. Doxycycline is generally well tolerated but may cause vulvovaginal thrush, diarrhea, and photosensitivity and is not recommended for prophylaxis in children <8 years old or pregnant women after 15 weeks of gestation.

Chloroquine can no longer be relied upon to prevent *P. falciparum* infections in most areas but is still used to prevent and treat malaria due to the other human *Plasmodium* species and for *P. falciparum* malaria in Central American countries west and north of the Panama Canal and in Caribbean countries. Chloroquine-resistant *P. vivax* has been reported from parts of eastern Asia, Oceania, and Central and South America. High-level resistance in *P. vivax* is prevalent in Oceania and Indonesia. Chloroquine is generally well tolerated, although some patients cannot take it because of malaise, headache, visual symptoms (due to reversible keratopathy), gastrointestinal intolerance, alopecia, or pruritus. Chloroquine is considered safe in pregnancy. With chronic administration for >5 years, a characteristic dose-related retinopathy may develop, but this condition is rare at the doses used for antimalarial prophylaxis. Idiosyncratic or allergic reactions are also rare. Skeletal and/or cardiac myopathy is a potential problem with protracted prophylactic use, although it is more likely to occur at the high doses used in the treatment of rheumatoid arthritis. Neuropsychiatric reactions and skin rashes are unusual. Amodiaquine should not be used for weekly prophylaxis because continuous weekly use is associated with a high risk of agranulocytosis (~1 person in 2000) and hepatotoxicity (~1 person in 16,000).

Primaquine (0.5 mg of base/kg or a daily adult dose of 30 mg taken with food), an 8-aminoquinoline compound, has proved safe and effective in the prevention of drug-resistant falciparum and vivax

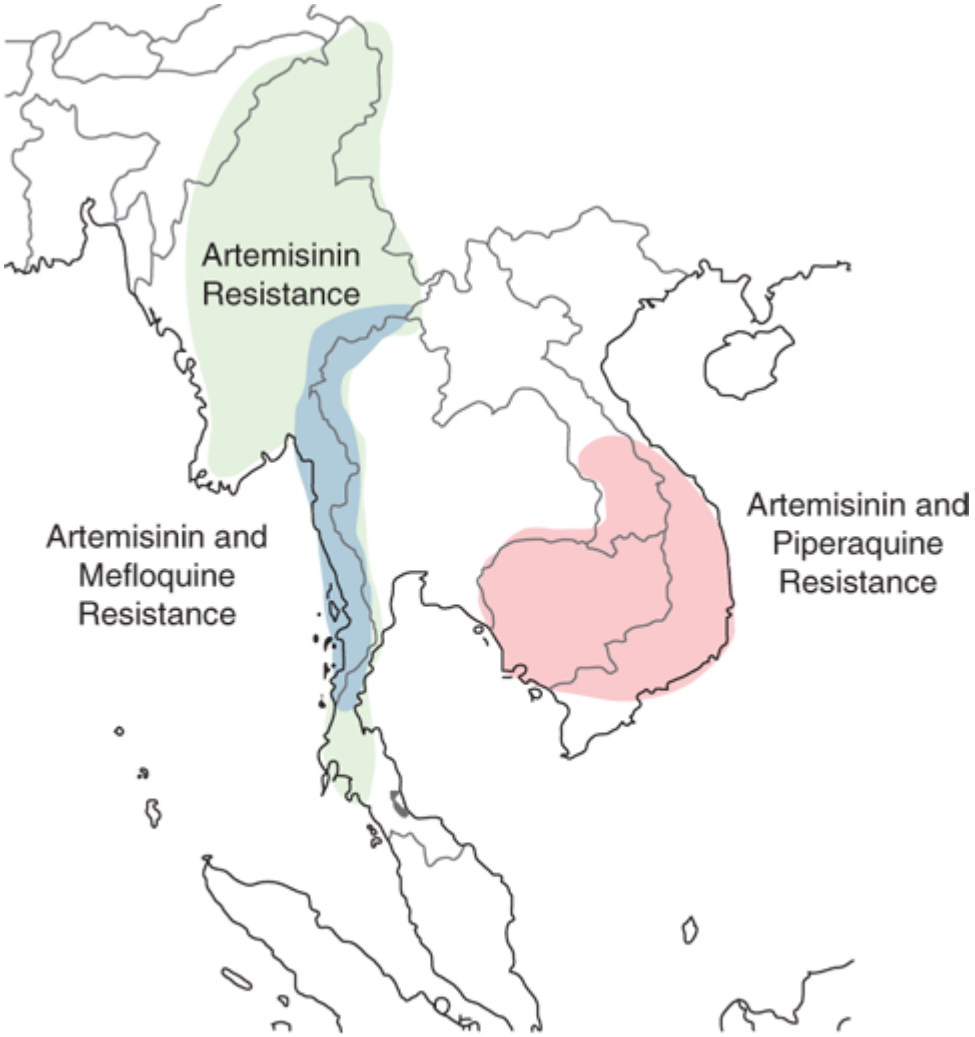
malaria in adults. Primaquine can be considered for persons who are intolerant to other recommended drugs. Abdominal pain can be prevented by taking primaquine with food. Primaquine should not be given to G6PD-deficient persons, in whom it can cause serious hemolysis; G6PD deficiency must therefore be excluded before primaquine is prescribed. Primaquine should not be given to pregnant women or infants <6 months old.

In the past, the dihydrofolate reductase inhibitors pyrimethamine and proguanil (chloroguanide) were administered widely, but the rapid selection of resistance in both *P. falciparum* and *P. vivax* has limited their use. Whereas antimalarial quinolines such as chloroquine (a 4-aminoquinoline) act only on the erythrocyte stage of parasitic development, the dihydrofolate reductase inhibitors (as well as atovaquone and primaquine) also inhibit preerythrocytic growth in the liver (causal prophylaxis) and development in the mosquito (sporontocidal activity). Proguanil is safe and well tolerated, although mouth ulceration occurs in ~8% of persons using this drug; it is considered safe for antimalarial prophylaxis in pregnancy. Prophylactic use of the combination of pyrimethamine and sulfadoxine is not recommended for weekly administration because of an unacceptable incidence of severe toxicity, principally exfoliative dermatitis and other skin rashes, agranulocytosis, hepatitis, and pulmonary eosinophilia (incidence, 1 in 7000; fatal reactions, 1 in 18,000).

Because of the increasing spread and intensity of antimalarial drug resistance **(Fig. 219-10)**), the CDC recommends that travelers and their providers consider their destination, type of travel, and current medications and health risks when choosing antimalarial chemoprophylaxis. There is an increasingly appreciated problem of falsified and substandard antimalarial drugs (and other medicines) on the shelves of pharmacies in Southeast Asia and sub-Saharan Africa; hence, travelers should purchase their preventive drugs from a reputable source before going to a malarious country. Consultation for the evaluation of prophylaxis failures or treatment of malaria can be obtained from state and local health departments and the CDC Malaria Hotline (855-856-4713) or the CDC Emergency Operations Center (770-488-7100).

FIGURE 219-10

Current geographic extent of artemisinin resistance and artemisinin-based combination therapy partner drug resistance in *Plasmodium falciparum* in the Greater Mekong subregion.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ACKNOWLEDGMENT

The authors gratefully acknowledge the substantial contributions of Joel G. Breman to this chapter in previous editions.

FURTHER READING

Dondorp AM et al: Artesunate versus quinine in the treatment of severe falciparum malaria in African children (AQUAMAT): An open-label, randomised trial. Lancet 376:1647, 2010. [PubMed: 21062666]
Singh B et al: A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. Lancet 363:1017, 2004. [PubMed: 15051281]
World Health Organization: Control and elimination of *Plasmodium vivax* malaria—A technical brief, July 2015. Available at *www.who.int/malaria/publications/atoz/9789241509244/en/*. Accessed December 8, 2017.
World Health Organization: *Guidelines for the Treatment of Malaria*, 3rd ed. Geneva, World Health Organization, 2015. Available at *apps.who.int/iris/bitstream/10665/162441/1/9789241549127_eng.pdf.* Accessed December 8, 2017.
World Health Organization: Severe malaria. Trop Med Int Health 19(S1):i–vii, 2014. Available at *dx.doi.org/10.1111/tmi.12313_1.* Accessed December 8, 2017.

Chapter 220: Babesiosis

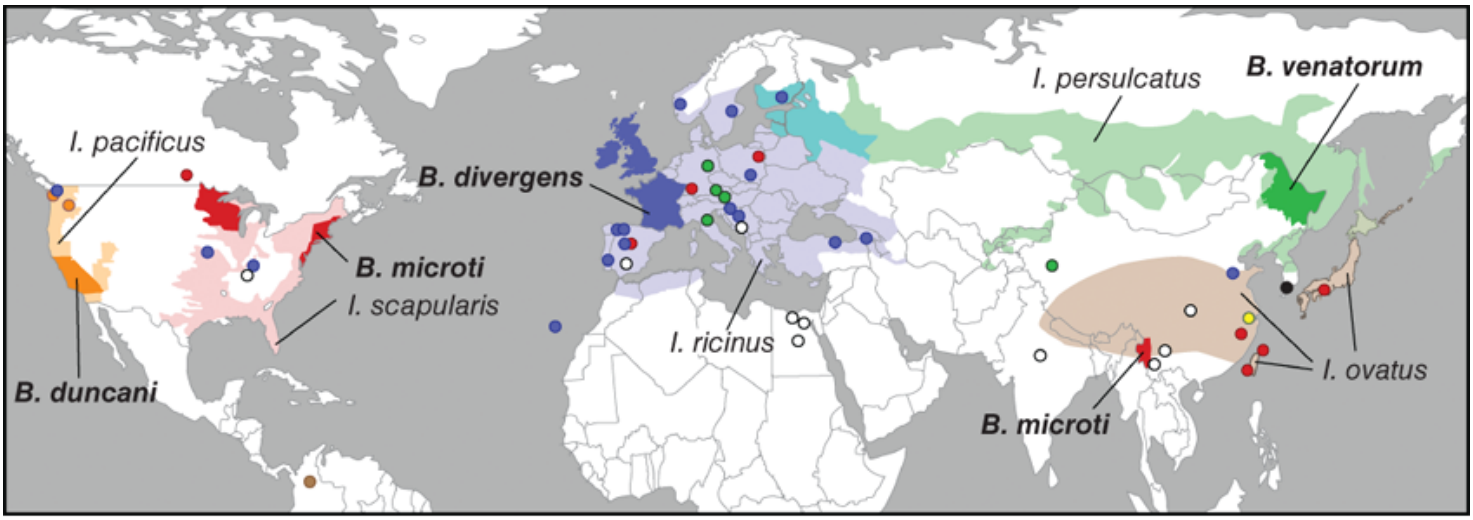
Edouard Vannier; Peter J. Krause

INTRODUCTION

Babesiosis is a worldwide **(Fig. 220-1)** emerging tick-borne infectious disease caused by protozoan parasites of the genus *Babesia* that invade and eventually lyse red blood cells (RBCs). More than 100 *Babesia* species infect a broad array of wild and domestic animals, but only a few of these species have been identified as etiologic agents of human babesiosis. Most cases are due to *Babesia microti* and occur in the United States. The infection typically is mild or asymptomatic in young and otherwise healthy individuals but can be severe and sometimes fatal in the elderly and the immunocompromised.

FIGURE 220-1

Worldwide distribution of human babesiosis. The geographic distribution of human babesiosis and its tick vectors is shown. Dark colors indicate areas where human babesiosis is endemic or sporadic (defined by ≥5 cases), whereas light colors indicate areas where tick vectors are present but human babesiosis is rare (<5 cases), undocumented, or absent. Circles depict single cases except in two locations (Montenegro and eastern Poland), where ≥5 cases occurred but all patients were diagnosed at the same hospital. Colors distinguish the etiologic agents: red for *B. microti,* orange for *B. duncani,* blue for *B. divergens,* green for *B. venatorum,* brown for *B. bovis,* black for KO-1, and yellow for *Babesia* species XXB. White circles depict cases caused by uncharacterized *Babesia* species. Asymptomatic infections are omitted. *(Adapted from E Vannier, PJ Krause: N Engl J Med 366:2397, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ETIOLOGY AND EPIDEMIOLOGY

United States

GEOGRAPHIC DISTRIBUTION

In the United States, human babesiosis caused by *B. microti* is endemic in the Northeast and upper Midwest; seven states in these two regions (Connecticut, Massachusetts, Minnesota, New Jersey, New York, Rhode Island, and Wisconsin) account for more than 90% of reported cases. Whole-genome analysis shows that isolates in the

continental United States began to diverge from those in Asia between 1400 and 14,000 years ago and are polyphyletic. Isolates in New England appear to have separated from those in the Midwest some 600 years ago; those on Nantucket Island form a separate subgroup. Other *Babesia* species causing sporadic disease in the United States include *B. duncani* and *B. duncani*–type organisms along the Pacific Coast and *B. divergens*–like organisms in Arkansas, Kentucky, Missouri, and Washington State.

INCIDENCE

National surveillance for human babesiosis was begun in the United States in January 2011. More than 1600 cases were reported in 2016—up from ~100 cases in 1996 and ~500 cases in 2006. The steady increase in the number of reported cases is due to the geographic expansion of *Babesia*-infected ticks and reservoir hosts as well as to a greater awareness of the disease among health care workers and improved reporting to state health departments and the Centers for Disease Control and Prevention (CDC). Babesiosis is reported from areas that have long been endemic for Lyme disease but rarely from areas to which *Borrelia burgdorferi* has recently spread. The delay in geographic expansion of *B. microti* is best explained by its poor ecologic fitness compared with that of *B. burgdorferi* and supports the hypothesis that *B. burgdorferi* promotes maintenance of *B. microti* in the enzootic cycle. Even in highly *Babesia*-endemic areas, the incidence of babesiosis is uneven, sometimes reaching one-third the incidence of Lyme disease. The incidence of babesiosis is underestimated because symptoms are nonspecific and because young healthy individuals typically experience mild or asymptomatic infection and may not seek medical attention.

MODES OF TRANSMISSION

In the United States, *B. microti* is transmitted to humans primarily by the nymphal stage of the deer tick (*Ixodes scapularis*), the same tick that transmits the causative agents of Lyme disease **(Chap. 181)** and human granulocytotropic anaplasmosis **(Chap. 182)**. Transmission generally occurs from May through October, with at least three-fourths of cases presenting from June through August. The vectors for transmission of *B. duncani*/*B. duncani*–type and *B. divergens*–like organisms are thought to be *Ixodes pacificus* and *Ixodes dentatus*, respectively.

Babesiosis occasionally is acquired through transfusion of blood products, mostly packed RBCs. More than 200 cases of transfusion-transmitted babesiosis due to *B. microti* have been reported but only three cases due to *B. duncani*. One-fifth of patients whose cases were caused by *B. microti* died, whereas all three patients infected with *B. duncani* survived. Like that of tick-transmitted babesiosis, the incidence of transfusion-transmitted babesiosis has steadily increased over the past 15 years. Transfusion-transmitted cases occur year-round but are most common from June through November. More than 85% of transfusion-transmitted cases occur in endemic areas. transfusion-transmitted babesiosis occurs in nonendemic areas when unrecognized *Babesia*-contaminated blood products are imported from endemic areas; when asymptomatically infected residents of endemic areas donate blood in nonendemic areas; or when residents of nonendemic areas travel to endemic areas, become infected, and donate blood after they return home.

B. microti can be transmitted through solid organ transplantation. Congenital *B. microti* transmission has been described but is rare. Other cases of neonatal babesiosis are acquired through transfusion or tick bite.

Global Considerations

 In Europe, the primary causative agent of human babesiosis is *B. divergens*, but *B. venatorum* and *B. microti* occasionally are reported. *Ixodes ricinus* is the tick vector for all three of these *Babesia* species. *B. microti* infection is indigenous to Europe but often is diagnosed in individuals who have recently returned from areas of the United States where babesiosis is endemic. Travel-associated babesiosis may become more frequent if, as anticipated, *Babesia* species continue to emerge worldwide. In Asia, cases due to *B. microti* were first documented in Japan and Taiwan. The case in Japan was acquired through blood transfusion, but *B. microti* organisms were found in *Ixodes ovatus* ticks in the region. Cases of *B. microti* infection have been identified in southwestern China along the border with Myanmar, where malaria is endemic. Two of the patients were co-infected with *Plasmodium* species. A recent case series established that *B. venatorum* is endemic in the northeastern province of Heilongjiang and that *Ixodes persulcatus* is the likely vector. *I. persulcatus* also can transmit *B. microti*. One case of *B. microti* infection has been reported in Australia and another in Canada, along the border with the upper midwestern United States. Sporadic cases due to uncharacterized *Babesia* species have been reported in mainland China, Egypt, India, Mexico, Montenegro, and South Africa.

CLINICAL MANIFESTATIONS

Asymptomatic *B. microti* Infection

Studies in highly endemic areas consistently indicate that 1–2% of individuals who donate blood are seropositive for *B. microti* without ever having been diagnosed with babesiosis. A carefully designed epidemiologic study revealed that 20% of adults and 40% of children do not experience symptoms following *B. microti* infection. If left untreated, asymptomatic infection may persist for >2 years. There is no evidence of long-term complications following asymptomatic infection; however, people who are asymptomatically infected may transmit the infection when they donate blood.

Mild to Moderate *B. microti* Illness

Symptoms typically develop 1–4 weeks after tick bite and 1–9 weeks (but as long as 6 months) after transfusion of contaminated blood products. Patients experience a gradual onset of fatigue, malaise, and weakness. Fever can reach 40.9°C (105.6°F) and often is accompanied by chills, sweats, headache, myalgia, and anorexia. Less frequent symptoms include arthralgia, nausea, vomiting, and dry cough. Sore throat, photophobia, abdominal pain, weight loss, shortness of breath, neck stiffness, and emotional lability have been reported. On physical examination, fever is the salient feature. Splenomegaly and hepatomegaly occasionally are noted, but lymphadenopathy is absent. Ecchymoses, petechiae, jaundice, slight pharyngeal erythema, and retinopathy with splinter hemorrhages and retinal infarcts rarely are observed. An erythema migrans rash **(Fig. A1-8)** signifies concurrent Lyme disease **(Chap. 181)**. Symptoms typically last 1–2 weeks, but fatigue may persist for several months. Patients who are co-infected with *B. burgdorferi* and *B. microti* experience a greater number of symptoms for a longer duration than patients with Lyme disease alone.

Severe *B. microti* Illness

Severe babesiosis requires hospital admission. The median length of hospital stay was 4 days (range, 1–39 days) among babesiosis patients reported to the CDC in 2011–2014. Severe babesiosis typically occurs in patients with one or more of the following characteristics: age >50 years, neonatal prematurity, asplenia/hyposplenism, HIV/AIDS, malignancy, and immunosuppressive therapy. More than one-third of hospitalized patients develop one or more complications, including acute respiratory distress syndrome, disseminated intravascular coagulation, congestive heart failure, renal failure, and hemophagocytic lymphohistiocytosis. Patients who develop these complications tend to have severe anemia (hemoglobin, ≤10 g/dL). Splenic infarcts and splenic rupture can occur, despite low-level parasitemia. In the absence of hemoperitoneum, splenic rupture should be managed without surgery, as removal of the spleen leaves these patients at risk for relapse of *Babesia* infection and for severe disease caused by other microorganisms. Production of autoantibodies can result in autoimmune hemolytic anemia, even after parasitemia has resolved. Babesiosis-associated immune thrombocytopenia has been reported. Death is not uncommon among hospitalized patients (3–9%), particularly those who are immunocompromised or who acquire the infection through blood transfusion (~20%). Laboratory prognostic factors for severe outcome, as defined by hospitalization for >14 days, intensive care unit stay of >2 days, or death, include an alkaline phosphatase level >125 U/L and a parasitemia level >4%.

Other *Babesia* Infections

B. duncani and *B. duncani*–type infections range in severity from asymptomatic to fatal. Although few cases have been documented, clinical manifestations have been similar to those reported for *B. microti* infections. All four patients infected with *B. divergens*–like organisms in the United States required hospitalization; two died. Most cases of *B. divergens* infection in Europe have occurred in people lacking a spleen. The incubation period is 1–3 weeks. Symptoms develop suddenly and consist of fever (>41°C [105.8°F]), shaking chills, drenching sweats, headache, myalgia, and lumbar and abdominal pain. Hemoglobinuria and jaundice are common, and mild hepatomegaly may occur. If the infection is not treated rapidly, patients may experience pulmonary edema and renal failure. All four patients infected with *B. venatorum* in Europe had been splenectomized; their illness ranged from mild to severe, and none died. Forty-eight cases of *B. venatorum* infection were reported in northeastern China in immunocompetent residents. Symptoms were similar to those of *B. microti* infection, although fever and chills were less common. Seven of the 48 patients were hospitalized, but all recovered despite receiving nonstandard antibiotic regimens, including clindamycin without quinine.

PATHOGENESIS

Anemia is a key feature of the pathogenesis of babesiosis. Hemolytic anemia caused by rupture of infected RBCs generates cell debris that may accumulate in the kidney and cause renal failure. Anemia also results from the

clearance of intact RBCs as they pass through the splenic red pulp and encounter resident macrophages. *Babesia* antigens expressed on the RBC membrane promote opsonization and facilitate uptake by splenic macrophages. In addition, RBCs are poorly deformable as a result of oxidation generated by the parasite and the host immune response and are filtered out as they attempt to squeeze across the venous vasculature. Bone marrow suppression due to cytokine production may also contribute to anemia.

An appropriate immune response is necessary for the control and clearance of *Babesia*. Studies in laboratory mice have established that $CD4^+$ T cells are critical for resistance to and resolution of *B. microti* infection. $CD4^+$ T cells are a major source of interferon γ (IFN-γ), and lack of this cytokine causes resistant mice to become highly susceptible to *B. microti*. IFN-γ is central to host resistance in *B. duncani* infection, but natural killer cells are its main source. Several lines of evidence suggest that an excessive immune response contributes to pathogenesis. Blockade of tumor necrosis factor receptor p55 (TNFRp55) accelerates resolution of *B. microti* parasitemia. *B. duncani* infection is more severe than *B. microti* infection in rodents and is characterized by pulmonary inflammation. Tumor necrosis factor α is expressed around alveolar septa, whereas IFN-γ is detected around pulmonary vessels. Blockade of either cytokine promotes the survival of mice infected with *B. duncani*.

DIAGNOSIS

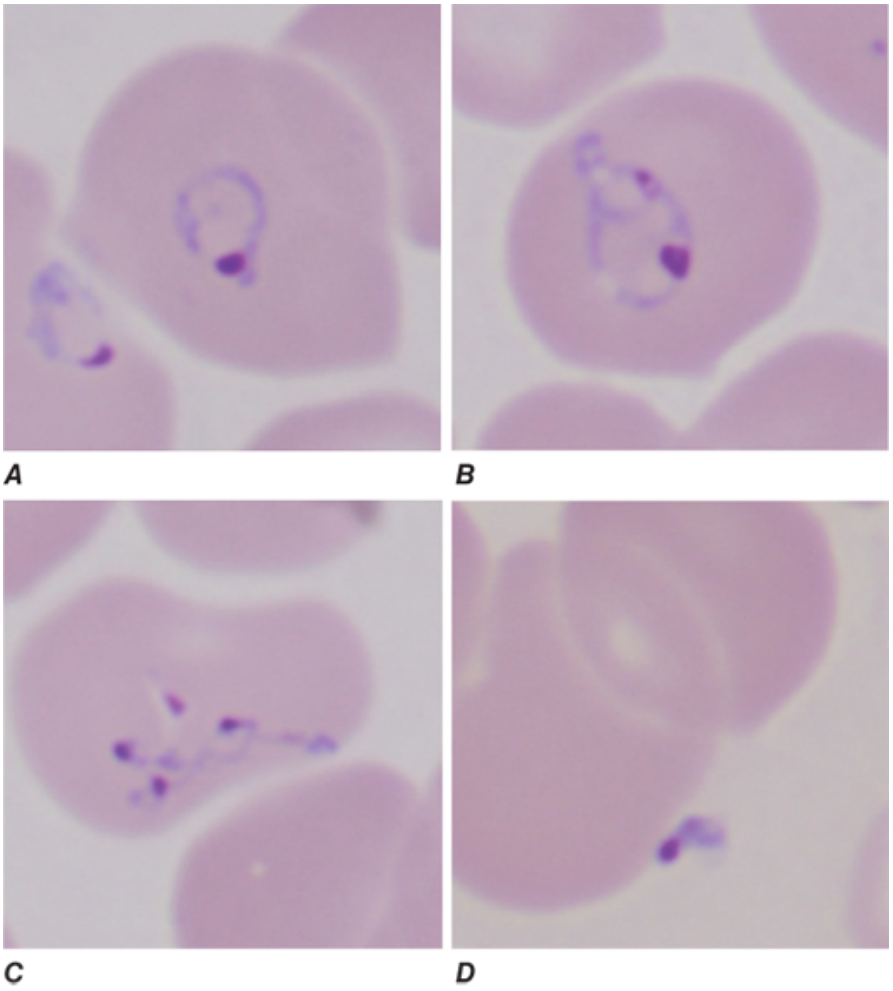
A diagnosis of babesiosis should be considered for any patient who lives or travels in a *Babesia*-endemic area and presents with a febrile illness in late spring, summer, or early autumn or within 6 months after a blood transfusion. Co-infection with *Babesia* should be considered when symptoms are more severe than expected or persist in patients who are diagnosed with and treated for Lyme disease and/or human granulocytotropic anaplasmosis.

Screening laboratory tests can help support the diagnosis of babesiosis. A complete blood count often shows anemia and thrombocytopenia. White blood cell counts can be elevated, reduced, or unchanged, but neutropenia occurs in approximately three-quarters of infants and one-third of adults. A low haptoglobin level or an elevated lactate dehydrogenase level is consistent with hemolytic anemia. Liver function tests often reveal elevated levels of alkaline phosphatase, aspartate and alanine aminotransferases, and bilirubin. Urinalysis may show hemoglobinuria, excess urobilinogen, and proteinuria. Elevated levels of blood urea nitrogen and serum creatinine indicate renal compromise.

A specific diagnosis usually is established by microscopic examination of Giemsa-stained thin blood smears **(Fig. 220-2)**. *Babesia* trophozoites appear round or ameboid. The ring form is most common and lacks the central brownish deposit (hemozoin) typical of *Plasmodium falciparum* trophozoites **(see Fig. A6-1*C*)**. Other distinguishing features are the absence of schizonts and gametocytes and the occasional presence of tetrads (“Maltese cross”). Tetrads are characteristic of *B. microti*, *B. duncani*, *B. venatorum,* and *B. divergens*–like organisms in human erythrocytes but rarely are observed. Because parasitemia may be as low as 0.01%, particularly at the onset of symptoms, identification of the parasite may require multiple blood smears over several days. Parasitemia generally ranges from 0.1 to 5% but has reached as high as 85% in an immunocompromised patient. If parasites cannot be identified by microscopy and the disease is still suspected, amplification of the *Babesia* 18S rRNA gene by polymerase chain reaction (PCR) or real-time PCR is recommended. Real-time PCR assays detect as few as 0.1–10 parasites/μL of blood, thereby increasing analytical sensitivity by 10- to 1000-fold over that of blood smear examination.

FIGURE 220-2

Giemsa-stained thin blood films showing *Babesia microti* parasites. *B. microti* are obligate parasites of erythrocytes. Trophozoites may appear as ring forms (***A***) or as ameboid forms (***B***). Merozoites can be arranged in tetrads that are pathognomonic (***C***). Extracellular parasites can be noted, particularly when parasitemia is high (***D***). *(Adapted from E Vannier, PJ Krause: N Engl J Med 366:2397, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Serology can suggest or confirm the diagnosis of babesiosis. An indirect immunofluorescent antibody test for *B. microti* is most commonly used. IgM titers of ≥1:64 and IgG titers of ≥1:1024 suggest active or recent infection. Titers typically decline over 6–12 months but may persist for >1 year. Antibodies to *B. microti* do not cross-react with *B. duncani* or *B. divergens* antigen. In *B. divergens* infection, serology is of limited use because symptoms often appear before antibodies can be detected. Sera from patients infected with *B. divergens*–like organisms or *B. venatorum* are reactive against *B. divergens* antigen. In the past decade, automated antibody assays that are standardized and suitable for screening the blood supply have been developed, including both an enzyme-linked immunosorbent assay that uses four *Babesia* peptides as antigen and an IgG arrayed fluorescence immunoassay that uses whole-cell sonicate as antigen.

TREATMENT

TREATMENT Babesiosis ASYMPTOMATIC *B. MICROTI* INFECTION

Asymptomatic *B. microti* infection is seldom diagnosed, and no evidence suggests that treatment is beneficial. Because of the increasing incidence and high mortality rate of transfusion-transmitted babesiosis, it is likely that screening of blood donations will be instituted. The combined use of assays to detect *B. microti* antibody and *B. microti* DNA can reliably identify blood donations that contain *B. microti* organisms, thereby effectively preventing transfusion-transmitted babesiosis once a screening algorithm approved by the U.S. Food and Drug Administration is implemented.

MILD TO MODERATE *B. MICROTI* ILLNESS

The first successful therapy for babesiosis consisted of clindamycin combined with quinine, but the combination of atovaquone plus azithromycin was subsequently shown in a prospective trial to be as effective in resolving symptoms and clearing parasitemia in adults with nonlife-threatening babesiosis. In that study, adverse effects were reported in

15% of patients who received atovaquone plus azithromycin but in 72% of those who received clindamycin plus quinine. The adverse reactions were so severe that treatment had to be stopped or the dosage reduced in one-third of participants taking clindamycin plus quinine, but in only 2% of those taking atovaquone plus azithromycin. Thus, atovaquone plus azithromycin, given orally for 7–10 days, is the recommended antibiotic regimen for mild to moderate babesiosis **(Table 220-1)**. Symptoms usually begin to resolve within 48 h of therapy initiation, but complete resolution of symptoms may take weeks or months. An atypical or poor response to therapy should raise suspicion of concurrent tick-borne disease or drug resistance.

SEVERE *B. MICROTI* ILLNESS

Azithromycin given intravenously plus atovaquone given orally for 7–10 days is recommended for treatment of severe babesiosis. Clindamycin given intravenously plus quinine given orally is an alternative regimen. Standard antimicrobial therapy sometimes is insufficient to resolve symptoms and clear parasitemia, especially in patients with marked immunosuppression—e.g., those who have received or are receiving rituximab or prednisone for B cell lymphomas or autoimmune disorders, who receive other immunosuppressive regimens for organ or bone marrow transplantation or malignancy, or who have HIV infection with low CD4+ T cell counts (AIDS). In such patients, antimicrobial therapy should be administered for at least 6 weeks, including 2 weeks after parasites are no longer observed on blood smear. In most instances, PCR should not be used to monitor the response to therapy because *B. microti* can persist in the blood at very low levels for weeks or months after symptoms have resolved and parasites are no longer detected on blood smears. A combination of high-dose azithromycin (600–1000 mg/d) plus atovaquone has been successfully used in immunocompromised patients. Failure to respond to atovaquone plus azithromycin has been documented in a few highly immunocompromised patients and has been attributed to accumulation of mutations in the *Babesia* genome, particularly in genes for which the encoded proteins are targets of atovaquone or azithromycin. Given that *B. microti* organisms circulate in a zoonotic cycle and that humans are dead-end hosts, use of atovaquone or azithromycin for the treatment of babesiosis does not increase the risk of overall antibiotic resistance. Patients who are unresponsive to atovaquone plus azithromycin or who do not tolerate clindamycin plus quinine can be successfully treated with alternative regimens (Table 220-1, footnote d).

Partial or complete RBC exchange transfusion is recommended in patients with high-grade parasitemia (≥10%); severe anemia (hemoglobin, <10 g/dL); or pulmonary, hepatic, or renal compromise. A desirable target of RBC exchange transfusion is a reduction in parasitemia by >90%. Parasitemia and hematocrit should be monitored daily until symptoms improve.

OTHER *BABESIA* INFECTIONS

B. duncani and *B. duncani*–type infections have been treated with IV clindamycin (600 mg three or four times daily or 1200 mg twice daily) plus oral quinine (600–650 mg three times daily) for 7–10 days. A regimen used for *B. divergens*–like infections consists of IV clindamycin (600 mg three or four times daily, 900 mg three times daily, or 1200 mg twice daily) plus oral quinine (650 mg three times daily). In Europe, *B. divergens* infection is considered a medical emergency. The recommended treatment is immediate, complete blood exchange transfusion combined with administration of IV clindamycin plus oral quinine. Some cases have been cured with exchange transfusion and clindamycin monotherapy. Anemia may persist for >1 month and require blood transfusion. The first line of therapy for *B. venatorum* infection in Europe has been the combination of clindamycin plus quinine. In an immunocompromised patient intolerant to quinine, cure was achieved by administration of atovaquone plus azithromycin. A pediatric case of *B. venatorum* in northwestern China was successfully treated by a standard course of atovaquone plus azithromycin.

TABLE 220-1 Treatment of Human Babesiosis

Adults	Children
***B. microti* Infection (Mild to Moderate Illness[a,b])**	
Atovaquone (750 mg q12h PO)	Atovaquone (20 mg/kg q12h PO; maximum, 750 mg/dose)
plus	***plus***
Azithromycin (500 mg/d PO on day 1, 250 mg/d PO thereafter)	Azithromycin (10 mg/kg qd PO on day 1 [maximum, 500 mg/dose], 5 mg/kg qd PO thereafter [maximum, 250 mg/dose])
***B. microti* Infection (Severe Illness[c,d])**	
Azithromycin (500 mg q24h IV)	Azithromycin (10 mg/kg up to adult dose q24h IV)
plus	***plus***
Atovaquone (750 mg q12h PO)	Atovaquone (20 mg/kg per dose up to adult dose q12h PO)
OR	***OR***
Clindamycin (300–600 mg q6h IV or 600 mg q8h PO)	Clindamycin (7–10 mg/kg q6–8h IV or 7–10 mg/kg q6–8h PO; maximum, 600 mg/dose)
plus	***plus***
Quinine (650 mg q6–8h PO)	Quinine (8 mg/kg q8h PO; maximum, 650 mg/dose)
plus	***plus***
Consider exchange transfusion	Consider exchange transfusion
***B. divergens* Infection**	
Immediate complete exchange transfusion	Immediate complete exchange transfusion
plus	***plus***
Clindamycin (600 mg q6–8h IV)	Clindamycin (7–10 mg/kg q6–8h IV; maximum, 600 mg/dose)
plus	***plus***
Quinine (650 mg q8h PO)	Quinine (8 mg/kg q8h PO; maximum, 650 mg/dose)

[a]Treatment duration, 7–10 days. [b]A high dose of azithromycin (600–1000 mg) combined with atovaquone has been recommended for immunocompromised hosts. [c]Treatment typically is given for 7–10 days, but its duration may vary. In severely immunocompromised patients, therapy should be continued for at least 6 weeks, including 2 weeks after parasites are no longer detected on blood smear. [d]Several alternative regimens have been used in a limited number of cases of *B. microti* infection, and their efficacy is uncertain. These regimens include atovaquone plus clindamycin (with or without azithromycin), azithromycin plus quinine, and atovaquone-proguanil added to atovaquone plus azithromycin and/or clindamycin plus quinine.

Sources: (1) ME Falagas, MS Klempner: Clin Infect Dis 22:809, 1996. (2) PJ Krause et al: N Engl J Med 343:1454, 2000. (3) PJ Krause et al: Clin Infect Dis 46:370, 2008. (4) CM Shih, CC Wang: Am J Trop Med Hyg 59:509, 1998. (5) CP Stowell et al: N Engl J Med 356:2313, 2007. (6) JM Vyas et al: Clin Infect Dis 45:1588, 2007. (7) GP Wormser et al: Clin Infect Dis 50:381, 2010.

PREVENTION

No vaccine is available for human use. There is no role for antibiotic prophylaxis. Individuals who reside in endemic areas, especially those at risk for severe babesiosis, should wear clothing that covers the lower and upper parts of the body, apply tick repellents (such as DEET) to the skin and permethrin to clothing, and limit outdoor activities where ticks may abound from May through October. The skin should be thoroughly examined after outdoor activities, and

ticks should be removed with tweezers. Individuals with a history of babesiosis or asymptomatic *Babesia* infection confirmed by laboratory testing are indefinitely deferred from donating blood.

FURTHER READING

Kletsova EA et al: Babesiosis in Long Island: Review of 62 cases focusing on treatment with azithromycin and atovaquone. Ann Clin Microbiol Antimicrob 16:26, 2017. [PubMed: 28399851]
Krause PJ et al: Atovaquone and azithromycin for the treatment of human babesiosis. N Engl J Med 343:1454, 2000. [PubMed: 11078770]
Krause PJ et al: Persistent and relapsing babesiosis in immunocompromised patients. Clin Infect Dis 46:370, 2008. [PubMed: 18181735]
Moritz ED et al: Screening for *Babesia microti* in the U.S. blood supply. N Engl J Med 375:2236, 2016. [PubMed: 27959685]
Sanchez E et al: Diagnosis, treatment, and prevention of Lyme disease, human granulocytic anaplasmosis, and babesiosis: A review. JAMA 315:1767, 2016. [PubMed: 27115378]
Vannier E, Krause PJ: Human babesiosis. N Engl J Med 366:2397, 2012. [PubMed: 22716978]
Wolley AE et al: Post-babesiosis warm autoimmune hemolytic anemia. N Engl J Med 376:939, 2017. [PubMed: 28273010]

Chapter 221: Leishmaniasis

Shyam Sundar

INTRODUCTION

Encompassing a complex group of disorders, leishmaniasis is caused by unicellular eukaryotic obligatory intracellular protozoa of the genus *Leishmania* and primarily affects the host's reticuloendothelial system. *Leishmania* species produce widely varying clinical syndromes ranging from self-healing cutaneous ulcers to fatal visceral disease. These syndromes fall into three broad categories: visceral leishmaniasis (VL), cutaneous leishmaniasis (CL), and mucosal leishmaniasis (ML).

ETIOLOGY AND LIFE CYCLE

Leishmaniasis is caused by ~20 species of the genus *Leishmania* in the order Kinetoplastida and the family Trypanosomatidae **(Table 221-1)**. Several clinically important species are of the subspecies *Viannia*. The organisms are transmitted by phlebotomine sandflies of the genus *Phlebotomus* in the "Old World" (Asia, Africa, and Europe) and the genus *Lutzomyia* in the "New World" (the Americas). Transmission may be anthroponotic (i.e., the vector transmits the infection from infected humans to healthy humans) or zoonotic (i.e., the vector transmits the infection from an animal reservoir to humans). Human-to-human transmission via shared infected needles has been documented in IV drug users in the Mediterranean region. In utero transmission to the fetus occurs rarely.

TABLE 221-1 Geographic Distribution and Characteristic Epidemiology of Leishmaniases

Organism, Endemic Region	Clinical Syndrome	Species	Vector	Reservoir	Transmission	Setting
***Leishmania donovani* Complex**						
South Asia	VL, PKDL	*L. donovani*	*Phlebotomus argentipes*	Humans	Anthroponotic	Rural, domestic
Sudan, South Sudan, Somalia, Ethiopia, Kenya, Uganda	VL, PKDL	*L. donovani*	*P. orientalis, P. martini*	Humans, rodents in Sudan, canines	Anthroponotic, occasionally zoonotic	Majority peridomestic, occasionally sylvatic
Mediterranean basin, Middle East, Central Asia, China	VL, CL	*L. infantum*	*P. perniciosus, P. ariasi*	Dogs, foxes, jackals	Zoonotic	Domestic, peridomestic
Middle East, Saudi Arabia, Yemen	VL	*L. donovani*	*P. perniciosus, P. ariasi*	Dogs, foxes, jackals	Zoonotic	Domestic, peridomestic
Central and South America	VL, CL	*L. infantum*[a]	*Lutzomyia longipalpis*	Foxes, dogs, opossums	Zoonotic	Domestic, peridomestic, periurban
Azerbaijan, Armenia, Georgia, Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan	VL	*L. infantum*	*P. turanicus*	Humans, dogs, foxes	Anthroponotic, zoonotic	Domestic
L. tropica						
Western India to Turkey, parts of North and East Africa	CL, leishmaniasis recidivans	*L. tropica*	*P. sergenti*	Humans	Anthroponotic	Urban domestic, peridomestic
L. major						
Western and Central Asia, North and sub-Saharan Africa	CL	*L. major*	*P. papatasi, P. duboscqi*	Nile rats, rodents	Zoonotic	Sylvatic, peridomestic
Kazakhstan, Turkmenistan, Uzbekistan	CL	*L. major*	*P. papatasi, P. duboscqi*	Gerbils	Zoonotic	Rural
L. aethiopica						
Ethiopia, Uganda, Kenya	CL, DCL	*L. aethiopica*	*P. longipes, P. pedifer*	Hyraxes	Zoonotic	Sylvatic, peridomestic
Subspecies *Viannia*						
Peru, Ecuador	CL, ML	*L. (V.) peruviana*	*Lutzomyia verrucarum, L. peruensis*	Wild rodents	Zoonotic	Andean Valleys
Guyana, Surinam, French Guyana, Ecuador, Brazil, Colombia, Bolivia	CL, ML	*L. (V.) guyanensis*	*L. umbratilis*	Sloths, arboreal anteaters, opossums	Zoonotic	Tropical forest
Central America, Ecuador, Colombia	CL, ML	*L. (V.) panamensis*	*L. trapidoi*	Sloths	Zoonotic	Tropical forest and deforested areas
South and Central America	CL, ML	*L. (V.) braziliensis*	*Lutzomyia* spp., *L. umbratilis, Psychodopygus wellcomei*	Forest rodents, peridomestic animals	Zoonotic	Tropical forest and deforested areas
***L. mexicana* Complex**						
Central America and northern parts of South America	CL, ML, DCL	*L. amazonensis*	*L. flaviscutellata*	Forest rodents	Zoonotic	Tropical forest and deforested areas
	CL, ML, DCL	*L. mexicana*	*L. olmeca*	Variety of forest rodents and marsupials	Zoonotic	Tropical forest and deforested areas

Organism, Endemic Region	Clinical Syndrome	Species	Vector	Reservoir	Transmission	Setting
	CL, DCL	*L. pifanoi*	*L. olmeca*	Variety of forest rodents and marsupials	Zoonotic	Tropical forest and deforested areas

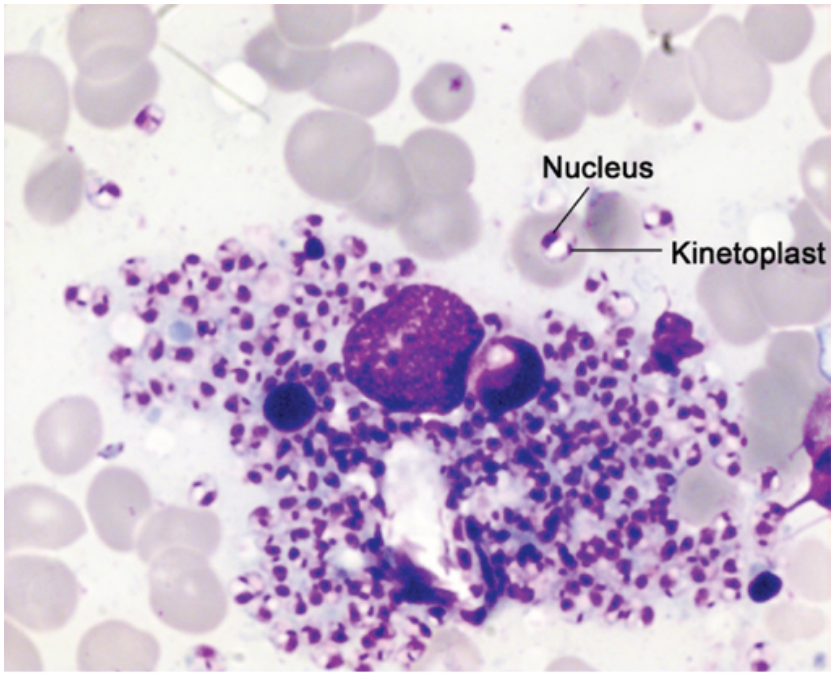
[a]*L. infantum* is designated *L. chagasi* in the New World.

Abbreviations: CL, cutaneous leishmaniasis; DCL, diffuse cutaneous leishmaniasis; ML, mucosal leishmaniasis; PKDL, post–kala-azar dermal leishmaniasis; VL, visceral leishmaniasis.

Leishmania organisms occur in two forms: extracellular, flagellate promastigotes (length, 10–20 μm) in the sandfly vector and intracellular, nonflagellate amastigotes (length, 2–4 μm; Fig. 221-1) in vertebrate hosts, including humans. Promastigotes are introduced through the proboscis of the female sandfly into the skin of the vertebrate host. Neutrophils predominate among the host cells that first encounter and take up promastigotes at the site of parasite delivery. The infected neutrophils may undergo apoptosis and release viable parasites that are taken up by macrophages, or the apoptotic cells may themselves be taken up by macrophages and dendritic cells. The parasites multiply as amastigotes inside macrophages, causing cell rupture with subsequent invasion of other macrophages. While feeding on infected hosts, sandflies pick up amastigotes, which transform into the flagellate form in the flies' posterior midgut and multiply by binary fission; the promastigotes then migrate to the anterior midgut and can infect a new host when flies take another blood meal.

FIGURE 221-1

A macrophage with numerous intracellular amastigotes (2–4 μm) in a Giemsa-stained splenic smear from a patient with visceral leishmaniasis. Each amastigote contains a nucleus and a characteristic kinetoplast consisting of multiple copies of mitochondrial DNA. A few extracellular parasites are also visible.



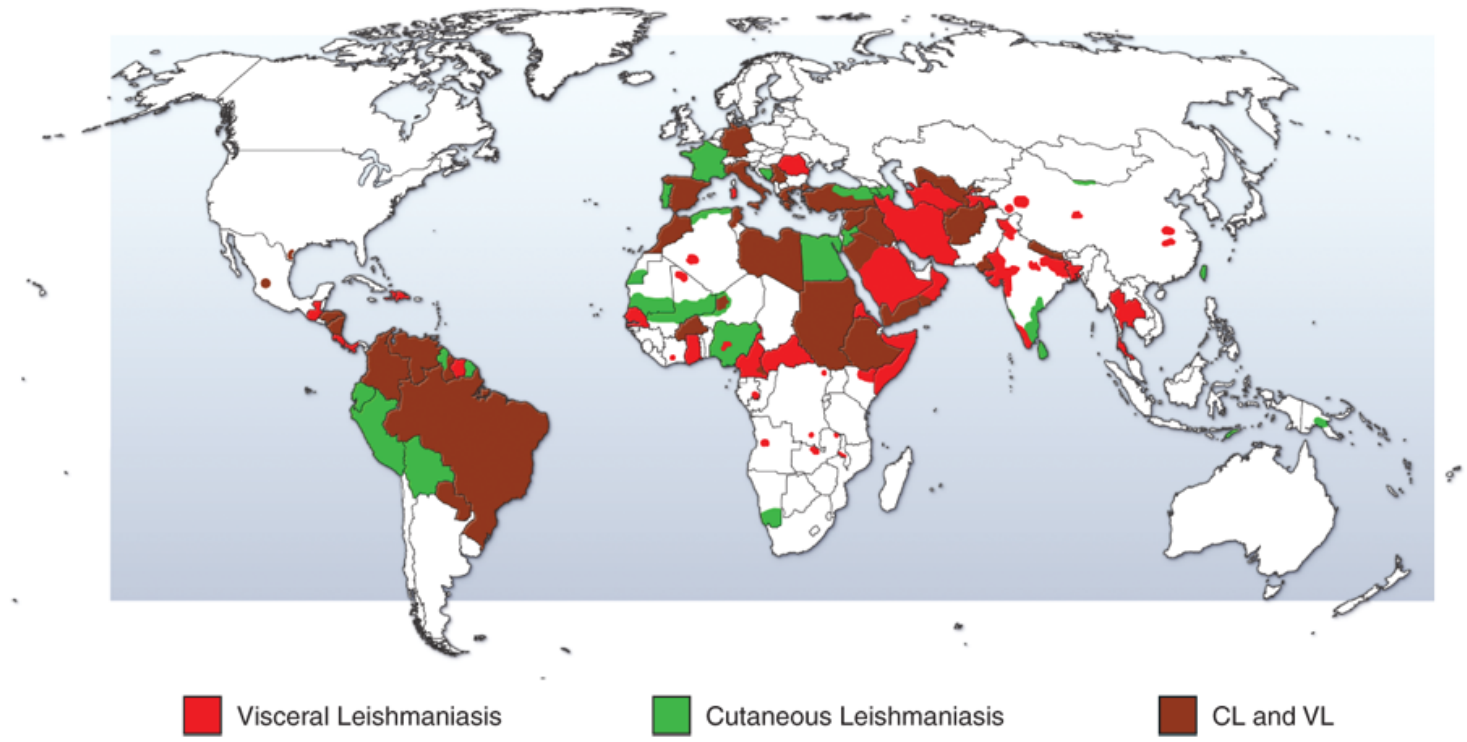
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPIDEMIOLOGY

Image not available. Leishmaniasis occurs in 98 countries—most of them developing—in tropical and temperate regions (Fig. 221-2). More than 1.5 million cases occur annually, of which 0.7–1.2 million are CL (and its variations) and 200,000–400,000 are VL. More than 350 million people are at risk, with an overall prevalence of 12 million. The distribution of *Leishmania* is limited by the distribution of sandfly vectors. Human leishmaniasis is on the increase worldwide except on the Indian subcontinent, where a VL elimination program has been implemented and VL incidence is markedly declining.

FIGURE 221-2

Worldwide distribution of human leishmaniasis. CL, cutaneous leishmaniasis; VL, visceral leishmaniasis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

VISCERAL LEISHMANIASIS

VL (also known as *kala-azar*, a Hindi term meaning “black fever”) is caused by the *Leishmania donovani* complex, which includes *L. donovani* and *Leishmania infantum* (the latter designated *Leishmania chagasi* in the New World); these species are responsible for anthroponotic and zoonotic transmission, respectively. India and neighboring Bangladesh, Sudan and neighboring South Sudan, Ethiopia, and Brazil are the four largest foci of VL and account for 90% of the world’s VL burden. Zoonotic VL is reported from all countries in the Middle East, Pakistan, and other countries from western Asia to China. Endemic foci also exist in the independent states of the former Soviet Union, mainly Georgia and Azerbaijan. In the Horn of Africa, Sudan, South Sudan, Ethiopia, Kenya, Uganda, and Somalia report VL. In Sudan and South Sudan, large outbreaks are thought to be anthroponotic, although zoonotic transmission also occurs. VL is rare in West and sub-Saharan Africa.

Mediterranean VL, long an established endemic disease due to *L. infantum*, has a large canine reservoir and was seen primarily in infants before the advent of HIV infection. In Mediterranean Europe, 70% of adult VL cases are associated with HIV co-infection. The combination is deadly because of the combined impact of the two infections on the immune system. IV drug users are at particular risk. Other forms of immunosuppression (e.g., that associated with organ transplantation) also predispose to VL. In the Americas, disease caused by *L. infantum* is endemic from Mexico to Argentina, but 90% of cases in the New World are reported from northeastern Brazil. After the introduction of highly active antiretroviral therapy, the incidence of HIV–VL co-infection declined significantly in Europe; however, ~30 and 5% of VL patients are co-infected with HIV in Ethiopia and India, respectively.

Immunopathogenesis

The majority of individuals infected by *L. donovani* or *L. infantum* mount a successful immune response and control the infection, never developing symptomatic disease. Forty-eight hours after intradermal injection of killed promastigotes, these individuals exhibit delayed-type hypersensitivity (DTH) to leishmanial antigens in the leishmanin skin test (also called the Montenegro skin test). Results in mouse models indicate that the development of acquired resistance to leishmanial infection is controlled by the production of interleukin (IL) 12 by antigen-presenting cells and the subsequent secretion of interferon (IFN) γ, tumor necrosis factor (TNF) α, and other proinflammatory cytokines by the T helper 1 (T_H1) subset of T lymphocytes. The immune response in patients developing active VL is complex; in addition to increased production of multiple proinflammatory cytokines and chemokines, patients with active disease have markedly elevated levels of IL-10 in serum as well as enhanced IL-10 mRNA expression in lesional tissues. A direct role for IL-10 in the pathology of VL in humans is supported by studies demonstrating that IL-10 blockade can enhance IFN-γ responses in whole blood from VL patients. The main disease-promoting activity of IL-10 in VL may be to condition host macrophages for enhanced survival and growth of the parasite. IL-10 can render macrophages unresponsive to activation signals and inhibit killing of amastigotes by downregulating

the production of TNF-α and nitric oxide. Multiple antigen-presentation functions of dendritic cells and macrophages are also suppressed by IL-10. Patients with such suppression do not have positive leishmanin skin tests, nor do their peripheral-blood mononuclear cells respond to leishmanial antigens in vitro. Organs of the reticuloendothelial system are predominantly affected, with remarkable enlargement of the spleen, liver, and lymph nodes in some regions. The tonsils and intestinal submucosa are also heavily infiltrated with parasites. Bone marrow dysfunction results in pancytopenia.

Clinical Features

On the Indian subcontinent and in the Horn of Africa, persons of all ages are affected by VL. In endemic areas of the Americas and the Mediterranean basin, immunocompetent infants and small children as well as immunodeficient adults are affected especially often. The most common presentation of VL is an abrupt onset of moderate- to high-grade fever associated with rigor and chills. Fever may continue for several weeks with decreasing intensity, and the patient may become afebrile for a short period before experiencing another bout of fever. The spleen may be palpable by the second week of illness and, depending on the duration of illness, may become hugely enlarged **(Fig. 221-3)**. Hepatomegaly (usually moderate in degree) soon follows. Lymphadenopathy is common in most endemic regions of the world except the Indian subcontinent, where it is rare. Patients lose weight and feel weak, and the skin gradually develops dark discoloration due to hyperpigmentation that is most easily seen in brown-skinned individuals. In advanced illness, hypoalbuminemia may manifest as pedal edema and ascites. Anemia appears early and may become severe enough to cause congestive heart failure. Epistaxis, retinal hemorrhages, and gastrointestinal bleeding are associated with thrombocytopenia. Secondary infections such as measles, pneumonia, tuberculosis, bacillary or amebic dysentery, and gastroenteritis are common. Herpes zoster, chickenpox, boils in the skin, and scabies may also occur. Untreated, the disease is fatal in most patients, including 100% of those with HIV co-infection.

FIGURE 221-3

A patient with visceral leishmaniasis has a hugely enlarged spleen visible through the surface of the abdomen. Splenomegaly is the most important feature of visceral leishmaniasis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Leukopenia and anemia occur early and are followed by thrombocytopenia. There is a marked polyclonal increase in serum immunoglobulins. Serum levels of hepatic aminotransferases are raised in a significant proportion of patients, and serum bilirubin levels are elevated occasionally. Renal dysfunction is uncommon.

Laboratory Diagnosis

Demonstration of amastigotes in smears of tissue aspirates is the gold standard for the diagnosis of VL (Fig. 221-1). The sensitivity of splenic smears is >95%, whereas smears of bone marrow (60–85%) and lymph node aspirates (50%) are less sensitive. Culture of tissue aspirates increases sensitivity. Splenic aspiration is invasive and may be dangerous in untrained hands. Several serologic techniques are currently used to detect antibodies to *Leishmania*. An enzyme-linked immunosorbent assay (ELISA) and the indirect immunofluorescent antibody test (IFAT) are used in sophisticated laboratories.

Image not available. In the field, however, a rapid immunochromatographic test based on the detection of antibodies to a recombinant antigen (rK39) consisting of 39 amino acids conserved in the kinesin region of *L. infantum* is used worldwide. The test requires only a drop of fingerprick blood or serum, and the result can be read within 15 min. Except in East Africa (where both its sensitivity and its specificity are lower), the sensitivity of the rK39 rapid diagnostic test (RDT) in immunocompetent individuals is ~98% and its specificity is ~90%. In Sudan, an RDT based on a new synthetic polyprotein, rK28, was more sensitive (96.8%) and specific (96.2%) than rK39-based RDTs. Since these antibody detection tests remain positive for years after cure, they cannot be used for measurement of cure or detection of relapse. Qualitative detection of leishmanial nucleic acid by polymerase chain reaction (PCR) or by loop-mediated isothermal amplification (LAMP) and quantitative detection by real-time PCR are highly sensitive; however, because the capacity to perform these tests is confined to specialized laboratories, they have yet to be used for routine diagnosis of VL in endemic areas. PCR can distinguish among the major species of *Leishmania* infecting humans.

Differential Diagnosis

VL is easily mistaken for malaria. Other febrile illnesses that may mimic VL include typhoid fever, tuberculosis, brucellosis, schistosomiasis, and histoplasmosis. Splenomegaly due to portal hypertension, chronic myeloid leukemia, tropical splenomegaly syndrome, and (in Africa) schistosomiasis may also be confused with VL. Fever with neutropenia or pancytopenia in patients from an endemic region strongly suggests a diagnosis of VL; hypergammaglobulinemia in patients with long-standing illness strengthens the diagnosis. In nonendemic countries, a careful travel history is essential when any patient presents with fever.

TREATMENT

TREATMENT Visceral Leishmaniasis GENERAL CONSIDERATIONS

Severe anemia should be corrected by blood transfusion, and other comorbid conditions should be managed promptly. Treatment of VL is complex because the optimal drug, dosage, and duration vary with the endemic region. Despite completing recommended treatment, some patients experience relapse (most often within 6 months), and prolonged follow-up is recommended. A pentavalent antimonial is the drug of choice in most endemic regions of the world, but there is widespread resistance to antimony in the Indian state of Bihar, where either amphotericin B (AmB)—deoxycholate or liposomal—or miltefosine is preferred. Dose requirements for AmB are lower in India than in the Americas, Africa, or the Mediterranean region. In Mediterranean countries, where cost is seldom an issue, liposomal AmB (LAmB) is the drug of choice. In immunocompetent patients, relapses are uncommon with AmB in its deoxycholate and lipid formulations. Antileishmanial therapy has recently evolved as new drugs and delivery systems have become available and resistance to antimonial compounds has emerged.

Except for AmB (deoxycholate and lipid formulations), antileishmanial drugs are available in the United States only from the Centers for Disease Control and Prevention.

PENTAVALENT ANTIMONIAL COMPOUNDS

Two pentavalent antimonial (Sb^V) preparations are available: sodium stibogluconate (100 mg of Sb^V/mL) and meglumine antimoniate (85 mg of Sb^V/mL). The daily dose is 20 mg/kg by IV infusion or IM injection, and therapy continues for 28–30 days. Cure rates exceed 90% in Africa, the Americas, and most of the Old World but are <50% in Bihar, India, as a result of resistance. Adverse reactions to Sb^V treatment are common and include arthralgia, myalgia, and elevated serum levels of aminotransferases. Electrocardiographic changes are common. Concave ST-segment elevation is not significant, but prolongation of QT_c to >0.5 s may herald ventricular arrhythmia and sudden death. Chemical pancreatitis is common but usually does not require discontinuation of treatment; severe clinical pancreatitis occurs in immunosuppressed patients.

AMPHOTERICIN B

AmB is currently used as a first-line drug in Bihar, India. In other parts of the world, it is used when initial antimonial treatment fails. Conventional AmB deoxycholate is administered in doses of 0.75–1.0 mg/kg on alternate days for a total of 15 infusions. Fever with chills is an almost universal adverse reaction to AmB infusions. Nausea and vomiting are also common, as is thrombophlebitis in the infused veins. Acute toxicities can be minimized by administration of antihistamines like chlorpheniramine and antipyretic agents like acetaminophen before each infusion. AmB can cause renal dysfunction and hypokalemia and, in rare instances, elicits hypersensitivity reactions, bone marrow suppression, and myocarditis, all of which can be fatal.

Several lipid formulations of AmB, developed to replace the deoxycholate formulation, are preferentially taken up by reticuloendothelial tissues. Because very little free drug is available to cause toxicity, a large amount of drug can be delivered over a short period. LAmB has been used extensively to treat VL in all parts of the world. With a terminal half-life of ~150 h, LAmB can be detected in the liver and spleen of animals for several weeks after a single dose. This is the only drug approved by the U.S. Food and Drug Administration (FDA) for the treatment of VL; the regimen is 3 mg/kg daily on days 1–5, 14, and 21 (total dose, 21 mg/kg). However, the total-dose requirement for different regions of the world varies widely. In Asia, it is 10–15 mg/kg; in Africa, ~18 mg/kg; and in Mediterranean/American regions, ≥20 mg/kg. The daily dose is flexible (1–10 mg/kg). In a study in India, a single dose of 10 mg/kg cured infection in 96% of patients. This single-dose regimen is the preferred treatment in India, Bangladesh, and Nepal. Adverse effects of LAmB are usually mild and include infusion reactions, backache, and occasional reversible nephrotoxicity.

PAROMOMYCIN

Paromomycin (aminosidine) is an aminocyclitol-aminoglycoside antibiotic with antileishmanial activity. Its mechanism of action against *Leishmania* has yet to be established. Paromomycin is approved in India for the treatment of VL at an IM dose of 11 mg of base/kg daily for 21 days; this regimen produces a cure rate of 94.6%. However, the optimal dose has not been established in other endemic regions. Paromomycin is a relatively safe drug, but some patients develop hepatotoxicity, reversible ototoxicity, and (in rare instances) nephrotoxicity and tetany. Paromomycin, in combination with Sb^V, is used in sub-Saharan Africa.

MILTEFOSINE

Miltefosine, an alkylphosphocholine, is the first oral compound approved for the treatment of leishmaniasis. This drug has a long half-life (150–200 h); its mechanism of action is not clearly understood. The recommended therapeutic regimens for patients on the Indian subcontinent are a daily dose of 50 mg for 28 days for patients weighing <25 kg, a twice-daily dose of 50 mg for 28 days for patients weighing ≥25 kg, and 2.5 mg/kg for 28 days for children 2–11 years of age. These regimens have resulted in a cure rate of 94% in India. However, recent studies from the Indian subcontinent indicate a decline in the cure rate. Doses in other regions remain to be established. Because of its long half-life, miltefosine is prone to induce resistance in *Leishmania*. Its adverse effects include mild to moderate vomiting and diarrhea in 40 and 20% of patients, respectively; these reactions usually clear spontaneously after a few days. Rare cases of severe allergic dermatitis, hepatotoxicity, and nephrotoxicity have been reported. Because miltefosine is expensive and is associated with significant adverse events, it is best administered as directly observed therapy to ensure completion of treatment and to minimize the risk of resistance induction. Because miltefosine is teratogenic in rats, its use is contraindicated during pregnancy and (unless contraceptive measures are strictly adhered to for at least 3 months after treatment) in women of childbearing age.

MULTIDRUG THERAPY

Multidrug therapy for leishmaniasis is likely to be preferred in the future. Its potential advantages in VL include (1) better compliance and lower costs associated with shorter treatment courses and decreased hospitalization, (2) less toxicity due to lower drug doses and/or shorter duration of treatment, and (3) a reduced likelihood that resistance to either agent will develop. In a study from India, one dose of LAmB (5 mg/kg) followed by miltefosine for 7 days, paromomycin for 10 days, or both miltefosine and paromomycin simultaneously for 10 days (in their usual daily doses) produced a cure rate of >97% (all three combinations). In Africa, a combination of Sb^V and paromomycin given for 17 days was as effective and safe as Sb^V given for 30 days.

Prognosis of Treated VI Patients

Recovery from VL is quick. Within a week after the start of treatment, defervescence, regression of splenomegaly, weight gain, and recovery of hematologic parameters are evident. With effective treatment, no parasites are recovered from tissue aspirates at the post-treatment evaluation. Continued clinical improvement over 6–12 months is suggestive of cure. A small percentage of patients (with the exact figure depending on the regimen used) relapse but respond well to treatment with AmB deoxycholate or lipid formulations.

VI in the Immunocompromised Host

HIV/VL co-infection has been reported from 35 countries. Where both infections are endemic, VL behaves as an opportunistic infection in HIV-1-infected patients. HIV infection can increase the risk of VL development by several-fold in endemic areas. Co-infected patients usually show the classic signs of VL, but they may present with atypical features due to loss of immunity and involvement of unusual anatomic locations—e.g., infiltration of the skin, oral mucosa, gastrointestinal tract, lungs, and other organs. Serodiagnostic tests may be negative in up to 50% of patients. Parasites can be recovered from unusual sites such as bronchoalveolar lavage fluid and buffy coat. LAmB is the drug of choice for HIV/VL co-infection—both for primary treatment and for treatment of relapses. A total dose of 40 mg/kg, administered as 4 mg/kg on days 1–5, 10, 17, 24, 31, and 38, is considered optimal and is approved by the FDA, but most patients experience a relapse within 1 year. Pentavalent antimonials and AmB deoxycholate can also be used where LAmB is not accessible. Reconstitution of patients' immunity by antiretroviral therapy has led to a dramatic decline in the incidence of co-infection in the Mediterranean basin. In contrast, HIV/VL co-infection is on the rise in African and Asian countries. Ethiopia is worst affected: up to 30% of VL patients are also infected with HIV. Because restoration of the CD4+ T cell count to >200/μL does decrease the frequency of relapse, antiretroviral therapy (in addition to antileishmanial therapy) is a cornerstone of the management of HIV/VL co-infection. Secondary prophylaxis with pentamidine or lipid AmB has been shown to delay relapses, but no regimen has been established as optimal.

Post–Kala-Azar Dermal Leishmaniasis

On the Indian subcontinent and in Sudan and other East African countries, 2–50% of patients develop skin lesions concurrent with or after the cure of VL. Most common are hypopigmented macules, papules, and/or nodules or diffuse infiltration of the skin and sometimes of the oral mucosa. The African and Indian diseases differ in several respects; important features of post–kala-azar dermal leishmaniasis (PKDL) in these two regions are listed in **Table 221-2**, and disease in an Indian patient is depicted in **Fig. 221-4**.

TABLE 221-2 Clinical, Epidemiologic, and Therapeutic Features of Post–Kala-Azar Dermal Leishmaniasis: East Africa and the Indian Subcontinent

Feature	East Africa	Indian Subcontinent
Most affected country	Sudan and South Sudan	Bangladesh
Incidence among patients with VL	~50%	~2–17%
Interval between VL and PKDL	During VL to 6 months	6 months to 3 years
Age distribution	Mainly children	Any age
History of prior VL	Yes	Not necessarily
Rashes of PKDL in presence of active VL	Yes	No
Treatment with sodium stibogluconate	2–3 months	2–4 months
Natural course	Spontaneous cure in majority of patients	Spontaneous cure in minority of patients

Abbreviations: PKDL, post–kala-azar dermal leishmaniasis; VL, visceral leishmaniasis.

FIGURE 221-4

Post–kala-azar dermal leishmaniasis in an Indian patient. Note nodules of varying size involving the entire face. The face is erythematous, and the surface of some of the large nodules is discolored.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In PKDL, parasites are scanty in hypopigmented macules but may be seen and cultured more easily from nodular lesions. Cellular infiltrates are heavier in nodules than in macules. Lymphocytes are the dominant cells; next most common are histiocytes and plasma cells. In about half of cases, epithelioid cells—scattered individually or forming compact granulomas—

are seen. The diagnosis is based on history and clinical findings, but rK39 and other serologic tests are positive in most cases. Indian PKDL was treated with prolonged courses (up to 120 days) of pentavalent antimonials. This prolonged course frequently led to noncompliance. The alternative—several courses of AmB spread over several months—is expensive and unacceptable for most patients. Oral miltefosine for 12 weeks, in the usual daily doses, cures most patients with Indian PKDL. The efficacy of LAmB is being tested on the Indian subcontinent. In East Africa, a majority of patients experience spontaneous healing. In those with persistent lesions, the response to 60 days of treatment with a pentavalent antimonial is good.

CUTANEOUS LEISHMANIASIS

CL can be broadly divided into Old World and New World forms. Old World CL caused by *Leishmania tropica* is anthroponotic and is confined to urban or suburban areas throughout its range. Zoonotic CL is most commonly due to *Leishmania major*, which naturally parasitizes several species of desert rodents that act as reservoirs over wide areas of the Middle East, Africa, and central Asia. Local outbreaks of human disease are common. Major outbreaks currently affect Afghanistan, Syria, Iraq, Lebanon, and Turkey in association with refugees and population movement. CL is increasingly seen in tourists and military personnel on mission in CL-endemic regions of countries and as a co-infection in HIV-infected patients. *Leishmania aethiopica* is restricted to the highlands of Ethiopia, Kenya, and Uganda, where it is a natural parasite of hyraxes. New World CL is mainly zoonotic and is most often caused by *Leishmania mexicana, Leishmania (Viannia) panamensis*, and *Leishmania amazonensis*. A wide range of forest animals act as reservoirs, and human infections with these species are predominantly rural. As a result of extensive urbanization and deforestation, *Leishmania (Viannia) braziliensis* has adapted to peridomestic and urban animals, and CL due to this organism is increasingly becoming an urban disease. In the United States, a few cases of CL have been acquired indigenously in Texas.

Immunopathogenesis

As in VL, the proinflammatory (T_H1) response in CL may result in either asymptomatic or subclinical infection. However, in some individuals, the immune response causes ulcerative skin lesions, the majority of which heal spontaneously, leaving a scar. Healing is usually followed by immunity to reinfection with that species of parasite.

Clinical Features

A few days or weeks after the bite of a sandfly, a papule develops and grows into a nodule that ulcerates over weeks or months. The base of the ulcer, which is usually painless, consists of necrotic tissue and crusted serum, but secondary bacterial infection sometimes occurs. The margins of the ulcer are raised and indurated. Lesions may be single or multiple and vary in size from 0.5 to >3 cm **(Fig. 221-5)**. Lymphatic spread and lymph gland involvement may be palpable and may precede the appearance of the skin lesion. There may be satellite lesions, especially in *L. major* and *L. tropica* infections. The lesions usually heal spontaneously after 2–15 months. Lesions due to *L. major* and *L. mexicana* tend to heal rapidly, whereas those due to *L. tropica* and parasites of subspecies *Viannia* heal more slowly. In CL caused by *L. tropica*, new lesions—usually scaly, erythematous papules and nodules—develop in the center or periphery of a healed sore, a condition known as *leishmaniasis recidivans*. Lesions of *L. mexicana* and *Leishmania (Viannia) peruviana* closely resemble those seen in the Old World; however, lesions on the pinna of the ear are common, chronic, and destructive in the former infections. *L. mexicana* is responsible for chiclero's ulcer, the so-called self-healing sore of Mexico. CL lesions on exposed body parts (e.g., the face and hands), permanent scar formation, and social stigmatization may cause anxiety and depression and may affect the quality of life of CL patients.

FIGURE 221-5

Cutaneous leishmaniasis in a Bolivian child. There are multiple ulcers resulting from several sandfly bites. The edges of the ulcers are raised. *(Courtesy of P. Desjeux, Retired Medical Officer, World Health Organization, Geneva, Switzerland.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Differential Diagnosis

A typical history (an insect bite followed by the events leading to ulceration) in a resident of or a traveler to an endemic focus strongly suggests CL. Cutaneous tuberculosis, fungal infections, leprosy, sarcoidosis, and malignant ulcers are sometimes mistaken for CL.

Laboratory Diagnosis

Demonstration of amastigotes in material obtained from a lesion remains the diagnostic gold standard. Microscopic examination of slit skin smears, aspirates, or biopsies of the lesion is used for detection of parasites. Culture of smear or biopsy material may yield *Leishmania*. PCR is more sensitive than microscopy and culture and allows identification of *Leishmania* to the species level. This information is important in decisions about therapy because responses to treatment can vary with the species. Isoenzyme profiling is used to determine species for research purposes.

TREATMENT

TREATMENT Cutaneous Leishmaniasis

Although lesions heal spontaneously in the majority of cases, their spread or persistence indicates that treatment may be needed. One or a few small lesions due to “self-healing species” can be treated with topical agents. Systemic treatment is required for lesions over the face, hands, or joints; multiple lesions; large ulcers; lymphatic spread; New World CL with the potential for development of ML; and CL in HIV-co-infected patients.

A pentavalent antimonial is the first-line drug for all forms of CL and is used in a dose of 20 mg/kg for 20 days. The exceptions to this rule are CL caused by *Leishmania (Viannia) guyanensis*, for which pentamidine isethionate is the drug of choice (two injections of 4 mg of salt/kg separated by a 48-h interval), and CL due to *L. aethiopica*, which responds to paromomycin (16 mg/kg daily) but not to antimonials. Relapses usually respond to a second course of treatment. In Peru, topical imiquimod (5–7.5%) plus parenteral antimonials have been shown to cure CL more rapidly than antimonials alone. Azoles and triazoles have been used with mixed responses in both Old and New World CL, but have not been adequately assessed for this indication in clinical trials. In *L. major* infection, oral fluconazole (200 mg/d for 6 weeks) resulted in a higher rate of cure than placebo (79% vs 34%) and also cured infection faster. Adverse effects include gastrointestinal symptoms and hepatotoxicity. Ketoconazole (600 mg/d for 28 days) is 76–90% effective in CL due to *L. (V.) panamensis* and *L. mexicana* in Panama and Guatemala. Miltefosine has been used in CL in doses of 2.5 mg/kg for 28 days. This agent is effective against *L. major* infections. In Colombia, where CL is due to *L. (V.) panamensis*, miltefosine was also effective, with a cure rate of 91%. For *L. (V.) braziliensis* infections, however, the results with miltefosine are less consistent. In Brazil, miltefosine cured 71% of patients with *L. (V.) guyanensis* infection. Other drugs, such as dapsone, allopurinol, rifampin, azithromycin, and pentoxifylline, have been used either alone or in combinations, but most of the relevant studies have had design limitations that preclude meaningful conclusions.

Small lesions (≤3 cm in diameter) may conveniently be treated weekly until cure with an intralesional injection of a pentavalent antimonial at a dose adequate to blanch the lesion (0.2–2.0 mL). An ointment containing 15% paromomycin sulfate, either

alone or with 0.5% gentamicin or 12% methylbenzonium chloride, cured 70–82% of lesions due to *L. major* in 20 days and may be suitable for lesions caused by other species. Heat therapy with an FDA-approved radiofrequency generator and cryotherapy with liquid nitrogen have also been used successfully.

Diffuse Cutaneous Leishmaniasis (DCL)

DCL is a rare form of leishmaniasis caused by *L. amazonensis* and *L. mexicana* in South and Central America and by *L. aethiopica* in Ethiopia and Kenya. DCL is characterized by the lack of a cell-mediated immune response to the parasite, the uncontrolled multiplication of which thus continues unabated. The DTH response does not develop, and lymphocytes do not respond to leishmanial antigens in vitro. DCL patients have a polarized immune response with high levels of immunosuppressive cytokines, including IL-10, transforming growth factor (TGF) β, and IL-4, and low concentrations of IFN-γ. Profound immunosuppression leads to widespread cutaneous disease. Lesions may initially be confined to the face or a limb but spread over months or years to other areas of the skin. They may be symmetrically or asymmetrically distributed and include papules, nodules, plaques, and areas of diffuse infiltration. These lesions do not ulcerate. The overlying skin is usually erythematous in pale-skinned patients. The lesions are teeming with parasites, which are therefore easy to recover. DCL does not heal spontaneously and is difficult to treat. If relapse and drug resistance are to be prevented, treatment should be continued for some time after lesions have healed and parasites can no longer be isolated. In the New World, repeated 20-day courses of pentavalent antimonials are given, with an intervening drug-free period of 10 days. Miltefosine has been used for several months with a good initial response. Combinations should be tried. In Ethiopia, a combination of paromomycin (14 mg/kg per day) and sodium stibogluconate (10 mg/kg per day) is effective.

MUCOSAL LEISHMANIASIS

The subgenus *Viannia* is widespread from the Amazon basin to Paraguay and Costa Rica and is responsible for deep sores and for ML (Table 221-1). In *L. (V.) braziliensis* infections, cutaneous lesions may be simultaneously accompanied by mucosal spread of the disease or followed by spread years later. ML is typically caused by *L. (V.) braziliensis* and rarely by *L. amazonensis*, *L. (V.) guyanensis*, and *L. (V.) panamensis*. Young men with chronic lesions of CL are at particular risk. Overall, ~3% of infected persons develop ML. Not every patient with ML has a history of prior CL. ML is almost entirely confined to the Americas. In rare cases, ML may also be caused by Old World species like *L. major, L. infantum* (*L. chagasi*), or *L. donovani*.

Immunopathogenesis and Clinical Features

The immune response is polarized toward a T_H1 response, with marked increases of IFN-γ and TNF-α and varying levels of T_H2 cytokines (IL-10 and TGF-β). Patients have a stronger DTH response with ML than with CL, and their peripheral-blood mononuclear cells respond strongly to leishmanial antigens. The parasite spreads via the lymphatics or the bloodstream to mucosal tissues of the upper respiratory tract. Intense inflammation leads to destruction, and severe disability ensues. Lesions in or around the nose or mouth (espundia; **Fig. 221-6**) are the typical presentation of ML. Patients usually provide a history of self-healed CL preceding ML by 1–5 years. Typically, ML presents as nasal stuffiness and bleeding followed by destruction of nasal cartilage, perforation of the nasal septum, and collapse of the nasal bridge. Subsequent involvement of the pharynx and larynx leads to difficulty in swallowing and phonation. The lips, cheeks, and soft palate may also be affected. Secondary bacterial infection is common, and aspiration pneumonia may be fatal. Despite the high degree of T_H1 immunity and the strong DTH response, ML does not heal spontaneously.

FIGURE 221-6

Mucosal leishmaniasis in a Brazilian patient. There is extensive inflammation around the nose and mouth, destruction of the nasal mucosa, ulceration of the upper lip and nose, and destruction of the nasal septum. *(Courtesy of R. Dietz, Universidade Federal do Espírito Santo, Vitória, Brazil.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Laboratory Diagnosis

Tissue biopsy is essential for identification of parasites, but the rate of detection is poor unless PCR techniques are used. The strongly positive DTH response fails to distinguish between past and present infection.

TREATMENT

TREATMENT Mucosal Leishmaniasis

The regimen of choice is a pentavalent antimonial agent administered at a dose of 20 mg of Sb^V/kg for 30 days. Patients with ML require long-term follow-up with repeated oropharyngeal and nasal examination. With failure of therapy or relapse, patients may receive another course of an antimonial but then become unresponsive, presumably because of resistance in the parasite. In this situation, AmB should be used. An AmB deoxycholate dose totaling 25–45 mg/kg is appropriate. There are no controlled trials of LAmB, but administration of 2–3 mg/kg for 20 days is considered adequate. Miltefosine (2.5 mg/kg for 28 days) cured 71% of ML patients in Bolivia. The more extensive the disease, the worse the prognosis; thus prompt, effective treatment and regular follow-up are essential.

PREVENTION OF LEISHMANIASIS

No vaccine is available for any form of leishmaniasis. Inoculation with live *L. major* ("leishmanization") is practiced in Iran; 80% of recipients were protected, according to one report. Anthroponotic leishmaniasis is controlled by case finding, treatment, and vector control with insecticide-impregnated bed nets and curtains and residual insecticide spraying. Control of zoonotic leishmaniasis is more difficult. Use of insecticide-impregnated collars for dogs, treatment of infected domestic dogs, and culling of street dogs are measures that have been used with uncertain efficacy to prevent transmission of *L. infantum*. In Brazil, a

canine vaccine has been found to promote a decrease in the human and canine incidence of zoonotic VL. Two vaccines, Leishmune® and Leish-Tec®, are licensed in Brazil; Leishmune provides significant protection to vaccinated dogs. CaniLeish® is the first licensed canine vaccine developed in Europe. Personal prophylaxis with bed nets and repellants may reduce the risk of CL infections in the New World.

FURTHER READING

Alvar J et al: Leishmaniasis worldwide and global estimates of its incidence. PLoS One 7:e35671, 2012. [PubMed: 22693548]
Boelaert M, Sundar S: Leishmaniasis, in *Manson's Tropical Diseases*, 23rd ed. J Farrar et al (eds). Philadelphia, Elsevier Saunders, 2014, pp 631–651.
Sundar S, Chakravarty J: An update on pharmacotherapy for leishmaniasis. Expert Opin Pharmacother 25:1, 2014.
World Health Organization: Control of the leishmaniases. WHO Tech Rep Ser 949. Available from *http://apps.who.int/iris/bitstream/10665/44412/1/WHO_TRS_949_eng.pdf*. Accessed April 16, 2018.

Chapter 222: Chagas Disease and African Trypanosomiasis

François Chappuis; Yves Jackson

INTRODUCTION

Myriads of protozoan parasites of the genus *Trypanosoma* infect plants and animals worldwide. Among these, three are of clinical significance for humans: *T. cruzi* causes Chagas disease, and *T. brucei gambiense* and *T. brucei rhodesiense* cause human African trypanosomiasis (HAT), which is also known as “sleeping sickness.” Despite obvious differences in their geographic distribution, parasitic life cycle, clinical presentation, treatment, and outcome, these vector-borne diseases are archetypal examples of neglected tropical diseases. More broadly, these infectious diseases affect neglected populations of the lowest socioeconomic class who have limited access to care and who live either in remote rural areas of low- or middle-income tropical/subtropical countries or in urban areas of both endemic and nonendemic countries. The drugs to treat these conditions are several decades old, their availability is fragile, and their efficacy and/or safety is suboptimal.

Other trypanosome species (e.g., *T. congolense* and *T. evansi*) predominantly cause nonhuman zoonoses and only occasionally cause illness in humans.

CHAGAS DISEASE (AMERICAN TRYPANOSOMIASIS)

DEFINITION

First described in 1909 by Carlos Chagas, Chagas disease (American trypanosomiasis) is a zoonosis caused by the flagellated protozoan *T. cruzi*. After a frequently asymptomatic acute phase, 30–40% of patients develop life-threatening chronic cardiomyopathy and/or digestive-tract dysfunction over the course of decades. Acute reactivation may occur in immunocompromised patients. Chagas disease imposes an important human and social burden in Latin America and has recently spread outside its natural boundaries to become a global public health problem. A vast majority of affected individuals are unaware of being infected and do not have access to appropriate clinical management and counseling.

TRANSMISSION

Vectorial Transmission

T. cruzi infection is primarily a zoonosis transmitted to a range of wild and domestic mammals by blood-sucking triatomine bugs. Sylvatic, peridomiciliary, and intradomiciliary vectorial cycles sometimes overlap. Over a large geographic area in the Americas (from northern Argentina to the southern United States), most human infections are intradomiciliary, arising from a triatomine bite during nighttime sleep. Feces released by triatomines during a blood meal contain the infective metacyclic form of *T. cruzi* that enters the human body through cutaneous breaks, mucosae, or conjunctivae. Despite recent laboratory research showing the potential for transmission by bedbugs, there is no evidence that bedbugs actually transmit *T. cruzi* to humans.

Nonvectorial Transmission

Other modes of transmission can cause infection in both endemic and nonendemic regions. *T. cruzi* can be transmitted congenitally from mother to newborn, by transfusion of blood products, by tissue or organ transplantation, or by ingestion of contaminated food or drink. Congenital infection occurs in 1–10% of newborns of infected mothers. The risk of infection from contaminated blood products is low (1.7% overall, 13% for platelet recipients, and close to 0 for recipients of red blood cells and plasma). Transmission by infected organ and tissue transplants mostly affects heart, liver, and kidney recipients. Oral transmission is increasingly reported after ingestion of contaminated food (berries) or drinks (fruit or sugar cane juice) and occasionally causes outbreaks.

EPIDEMIOLOGY

Image not available. An estimated 5.7 million people are infected by *T. cruzi*, including >1 million individuals with chronic cardiomyopathy. However, the true global burden of Chagas disease is in fact uncertain. The highest numbers of infected individuals reside in Argentina, Brazil, and Mexico; the prevalence is highest in Bolivia (6.1%), Argentina (3.6%), and Paraguay (2.1%). In highly endemic regions of these countries, the prevalence may exceed 40%. Formerly restricted to poor rural populations, the distribution of cases—and, to some extent, *T. cruzi* transmission—has progressively extended to cities in the context of rapid urbanization and rural migration. A recent history of migration from a rural area is the main risk factor in urban settings.

Overall, the prevalence and incidence of Chagas disease have sharply declined in recent decades because of improved housing and socioeconomic conditions as well as public health interventions, including regional vector-control initiatives, implementation of systematic screening of blood products, and improved detection of congenital transmission. Several countries have been declared free of domiciliary t[Loading [Contrib]/a11y/accessibility-menu.js]esidual insecticide-spraying campaigns. This progress is threatened by adaptation of the vector to the

periurban environment, its resurgence in areas where spraying has been discontinued, the development of resistance to pyrethroid insecticides, and the persistence of peridomiciliary transmission. A growing number of localized outbreaks are being reported in previously stable areas, with the Amazon basin particularly at risk.

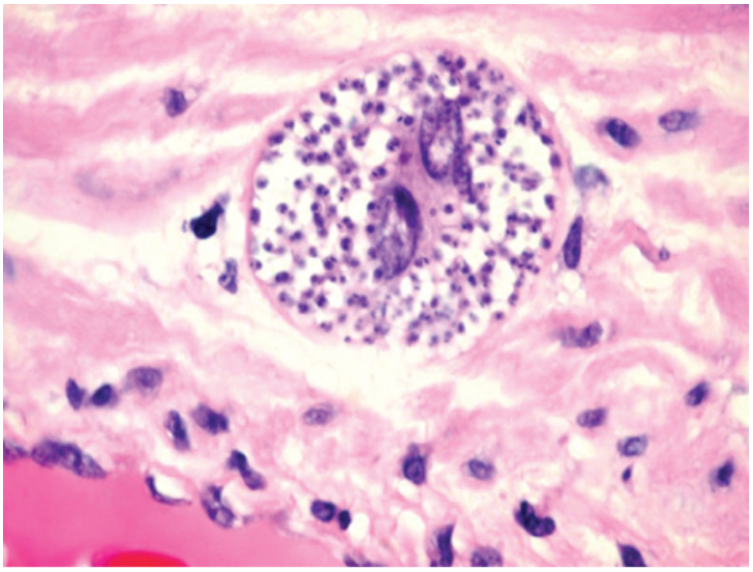
Chagas disease distribution has recently expanded to nonendemic countries in the context of increased global travel, with cases reported more frequently in North America, Western Europe, Australia, and Japan. The United States harbors up to 300,000 cases, mostly among immigrants from Central America. In addition, sporadic vector-borne infections occur in the southern states. Western Europe has 68,000–123,000 cases, and Japan and Australia report a few thousand cases. Despite the implementation of blood bank screening and of some dedicated medical programs, only a small proportion of cases have been identified and properly managed to date. A low level of awareness among health care professionals and difficulties experienced by some groups in accessing care appear to be major drivers. At-risk migrant communities are frequently subject to factors that render them socially, legally, or economically vulnerable. Moreover, the cultural perception of Chagas as a disease embedded in poverty can create a social stigma that complicates its management at the community level. In contrast to immigrants, international tourists visiting endemic countries are at very low risk of being infected, whether by reduviid bug bites or by other routes, and reports of Chagas disease in travelers are rare.

PATHOLOGY

Several *T. cruzi* strains have been identified. These strains have partially overlapping transmission cycles and geographic distributions, but no definitive evidence supports an association of certain strains with specific clinical manifestations or with variation in disease severity. The rarity of digestive tract involvement north of the Amazon basin suggests that specific parasitic and host genetic factors may influence the disease course. The pathogenesis of Chagas disease results from the complex interactions between the pathogen and the host immune response. Many questions about the relative importance of these interactions, including the role of autoimmune mechanisms, remain unanswered. After local penetration of trypomastigotes, parasites rapidly enter the bloodstream and disseminate through the body, infecting a wide range of nucleated cells in which they differentiate into amastigotes **(Fig. 222-1)**. The innate immune response triggered by parasite mucins and DNA leads to a predominantly T-helper 1 response. The production of various proinflammatory cytokines and the activation of $CD8^+$ T lymphocytes reduce parasitemia to a subpatent level within 4–8 weeks, a point marking the end of the acute phase.

FIGURE 222-1

A cluster of *Trypanosoma cruzi* amastigotes with an inflammatory infiltrate in the placenta of a congenitally infected newborn infant.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Immune evasion mechanisms allow persistent low-intensity proliferation of amastigotes and their release into the bloodstream, with subsequent infection of potentially all types of nucleated cells—notably cardiac, skeletal, and smooth-muscle cells. Mechanisms that have been postulated to determine the pathogenic evolution toward cardiomyopathy include the parasites' persistence and the host's inability to downregulate the initial immune response, resulting in cell-mediated damage and an imbalance of T-helper 1 and 2 responses with excessive production of proinflammatory cytokines. Secondary mechanisms, such as microcirculation abnormalities and dysautonomia, may also influence the progression of tissue damage.

In the myocardium, chronic inflammation results in cellular destruction and the development of fibrosis leading to a segmental loss of contractility and dilatation of the chambers, with the associated risk of left-ventricle apical aneurism. Focal hypoperfusion and tissue damage are sources of ventricular arrhythmias, while scarring lesions mostly affect the conduction system. Autonomic cell destruction leads to vagal and sympathetic denervation whose exact clinical significance remains to be clarified.

T. cruzi appears to have a direct toxic effect on digestive-tract intramural autonomic ganglion cells. Over time, the loss of neural cells affects muscular tone, leading to motility disorders and ultimately to organ dilatation (megaviscera syndrome). The esophagus and colon are primarily affected, but lesions may occur along the whole digestive tract. Inadequate relaxation of the lower esophageal sphincter causes s[illegible]age to the colon ultimately mimics Hirschsprung disease, with severe constipation and the risk of volvulus and toxic dilatation.

Loading [Contrib]/a11y/accessibility-menu.js

Factors reducing the cellular immune response, such as HIV infection, posttransplantation immunosuppressive therapies, or hematologic malignancies, may increase intracellular replication of amastigotes, with increased parasitemia (reactivation). Lesions develop predominantly in the central nervous system (CNS), the heart, and the skin. Among HIV patients, the risk of reactivation is ~20% in the absence of antiretroviral therapies and occurs when the $CD4^+$ T cell count falls <100/μL. Clinically manifest *T. cruzi* reactivation is an AIDS-defining opportunistic infection.

CLINICAL MANIFESTATIONS

The clinical manifestations of *T. cruzi* infection vary greatly among individuals. The infection course is divided into two phases that are associated with different clinical features, duration, and prognosis **(Table 222-1)**. The acute phase remains undetected and undiagnosed in most individuals. While 5–10% of these early infections spontaneously resolve without treatment, *T. cruzi* persists for life in the vast majority of individuals (the chronic phase); 60–70% of these individuals never develop apparent tissue damage (the indeterminate form), but the remaining 30–40% progress toward detectable organ damage of variable severity over decades (the determinate form). These chronic complications include cardiac (20–30%), digestive (5–20%), or mixed (5–10%) disorders. There is no predictor of evolution toward clinical manifestations during the chronic phase. In patients with cardiomyopathy, bundle branch blocks are usually the first signs and may cause no symptoms for years until more severe conduction-system disease, arrhythmias, and left ventricular dysfunction occur. Advanced cardiac damage entails a worse prognosis than other cardiomyopathies—notably, ischemic heart disease.

TABLE 222-1 Characteristics of the Stages of *Trypanosoma cruzi* Infection

Phase or Setting	Context	Onset of First Symptoms	Clinical Manifestations	Duration	Prognosis
Acute (congenital)	~5% risk of maternal transmission to newborn	At birth or weeks after delivery	>90% asymptomatic; rare lymphadenopathy, hepatosplenomegaly, jaundice, respiratory distress, growth retardation	2–8 weeks	Favorable when infant is born alive; unknown rate of in utero or neonatal death
Acute	Vector-borne transmission; oral transmission (ingestion of contaminated food/drinks); blood product transfusion; tissue/organ transplantation	1–2 weeks after vectorial transmission; may be sooner (days) after oral transmission or later (months) after transfusion/transplantation	>90% asymptomatic or mild febrile illness; local swelling at inoculation site (eyelid [Romaña sign] or skin [chagoma]); polyadenopathy; splenomegaly; myocarditis, hepatitis, and encephalitis more frequent with oral transmission	4–8 weeks	Mortality: 0.1–5% with oral transmission or myocarditis/encephalitis
Chronic (indeterminate form)	Balanced immune response after acute phase subsides	No symptoms	Normal clinical examination and ECG result	Lifelong or until determinate phase	No attributable mortality
Chronic (determinate form)	Predominant inflammatory response (in cardiomyopathy only)	Years to decades after initial infection	Dyspnea, chest pain, palpitation, syncope, sudden death, stroke, dysphagia, regurgitation, constipation, fecaloma, volvulus, peripheral neuropathy	Chronic	5-year mortality: 2–63%, depending on extent of cardiac damage; most important causes of death: cardiac failure and sudden death, followed by stroke
Acute (reactivation)	Severe immunosuppression	Variable	Myocarditis, erythema nodosum, panniculitis, *Toxoplasma*-like focal brain lesion, meningoencephalitis	Variable	Mortality depends on rapidity of diagnosis and treatment and on underlying conditions

Abbreviation: ECG, electrocardiography.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Chagas Disease (American Trypanosomiasis)

More than 90% of infections go undiagnosed, and cases are frequently identified at a late stage once chronic complications develop. The vast majority of *T. cruzi*–infected individuals are asymptomatic (i.e., in the indeterminate form of the chronic phase). An awareness of potential Chagas disease is important for general practitioners as well as for physicians from various specialties, including gastroenterologists, cardiologists, neurologists, obstetricians, pediatricians, and infectious disease specialists. Outside endemic areas, screening for Chagas disease should be proposed when any Latin American individual has evocative symptoms and signs, including abnormalities on electrocardiography (ECG) or increased risk of (1) *T. cruzi* infection (Chagas disease in the mother or other family members; origins in a highly endemic country or area; history of unscreened blood transfusion in Latin America); (2) transmission to others (e.g., via pregnancy or blood or organ donation); or (3) reactivation (current or pending immunosuppression). Screening of the relatives of an index case will probably identify additional cases.

DIAGNOSIS AND STAGING

Diagnostic Confirmation

Loading [Contrib]/a11y/accessibility-menu.js

Diagnostic strategies depend on the clinical phase **(Table 222-2)**. Detection of circulating parasites by microscopy of the blood with concentration (e.g., by the Strout method, microhematocrit) or by nucleic acid–based assay (polymerase chain reaction [PCR]) is the best diagnostic approach when the parasitemia level is high—i.e., during the acute phases, including reactivation. Once parasitemia becomes undetectable by microscopy (a point marking the end of the acute phase), diagnosis relies on immunologic tests that detect anti–*T. cruzi* IgG. The most common techniques include a conventional or recombinant enzyme-linked immunosorbent assay (ELISA) and immunofluorescence assays. Two positive serologic tests using different techniques and targeting different antigens confirm the diagnosis of Chagas disease during the chronic phase. In the presence of discordant serologic results, a third serologic test is warranted. Some of the immunochromatographic rapid diagnostic tests on the market have sufficient sensitivity and specificity to be used as first-line screening tests where laboratory facilities are not easily accessible. If the rapid diagnostic test result is positive, at least one conventional serologic assay is necessary to confirm infection.

TABLE 222-2 Diagnostic Procedures of Choice for Clinical Stages of *T. cruzi* Infection

Stage	Technique of Choice	Sample	Diagnostic Criteria
Acute	Microscopy after concentration, PCR	Peripheral blood, cerebrospinal or other body fluids	Positivity in one test
Acute (early congenital during first 9 months of life)	Microscopy after concentration, PCR	Cord or peripheral blood	Positivity in one test
Chronic (indeterminate and determinate forms)	Serology	Peripheral blood	Positivity in two tests with different techniques and antigens
Reactivation	Microscopy after concentration, PCR	Peripheral blood, cerebrospinal or other body fluids	Positivity with evidence of increasing parasitemia on serial samples or extremely high parasite load

Abbreviation: PCR, polymerase chain reaction.

Diagnosis of congenital infection relies on examination of cord and/or peripheral blood by microscopy or PCR during the first days or weeks of life. A test conducted after 4 weeks of age is most accurate: PCR earlier in life may be falsely positive, likely because of the passage of *T. cruzi* DNA fragments from the mother to the child. If results are negative, serologic tests should be performed at 9 months of age, once maternal antibodies have been cleared. During the chronic phase, the limited sensitivity (50–80%) of PCR restricts its usefulness for primary diagnosis; however, PCR can document therapeutic failure if it yields positive results after the completion of treatment. In the United States, the Centers for Disease Control and Prevention (CDC) provides reference laboratory testing (see contact information in the treatment section).

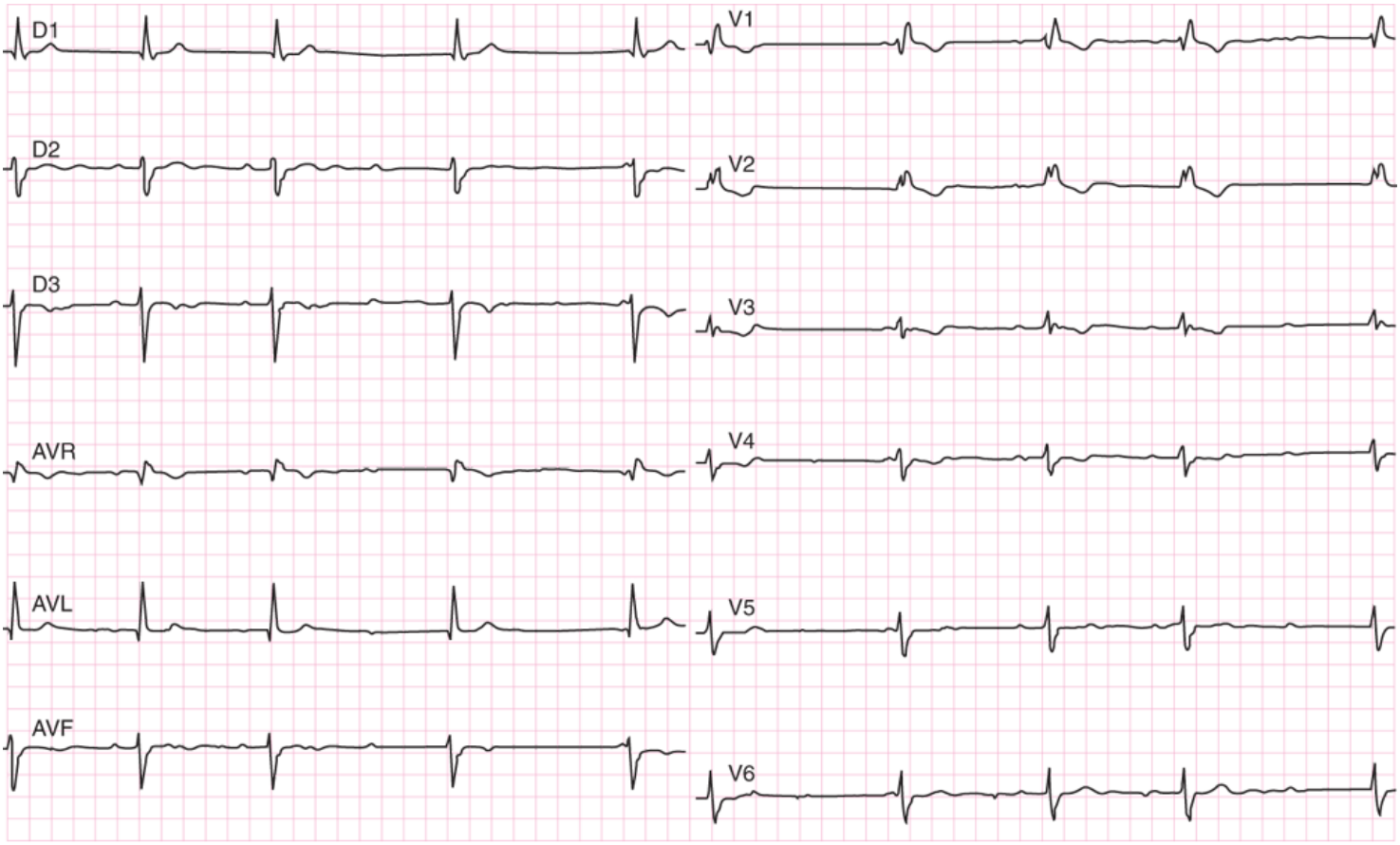
Disease Staging

Once *T. cruzi* infection is confirmed, clinicians should assess the presence of complications and concomitant factors that may influence the course of the disease. The initial evaluation includes a thorough cardiac, neurologic, and digestive history and a clinical examination. Twelve-lead ECG with a 30-s strip is a good screening test for Chagas-associated cardiomyopathy. The most frequently found abnormalities are right bundle branch block, left anterior fascicular block, ventricular premature beats, repolarization disorders, Q waves, and low QRS voltage (**Fig. 222-2**). An abnormal ECG result or the presence of suggestive cardiac symptoms warrants further investigation. Echocardiography and the 24-h Holter test are the preferred methods for assessment of chamber dilatation, apical aneurysm, ventricular dysfunction, and arrhythmias. Depending on the findings, the workup can be supplemented by MRI or electrophysiologic studies. Gastroenterologic investigations are performed in patients with suggestive symptoms, such as dysphagia and severe constipation. Barium esophagraphy and enema are first-line diagnostic procedures, which can be supplemented by esophageal manometry. Megacolon is diagnosed when the sigmoid or descending colon diameter is ≥6.5 cm.

FIGURE 222-2

Electrocardiogram of a 43-year-old patient shows bradycardia with alternating second- and third-degree atrioventricular (type 1), left anterior fascicular, and right bundle branch blocks.

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Comorbidities, including other cardiovascular risk factors, immunosuppressive conditions, and other chronic infections (e.g., with *Strongyloides stercoralis* or HIV) should be investigated.

TREATMENT

TREATMENT Chagas Disease (American Trypanosomiasis) ETIOLOGIC TREATMENT

Only two drugs, benznidazole and nifurtimox **(Table 222-3)**, have shown efficacy against *T. cruzi* infection when administered for ≥30 days. While these drugs have been used since the early 1970s, many questions remain about their mode of action and efficacy at the different stages of infection. The treatment goal depends on the clinical stage; the overall objectives are to cure patients who have recent infection or reactivation, to reduce morbidity, and to prevent transmission at later stages. Treatment is most effective during the acute (including congenital) phase and the early chronic phase (i.e., in patients <18 years of age), with a 60–100% cure rate. The efficacy of treatment during the indeterminate form of the chronic phase in patients >18 years old is not known; however, treatment may protect against the development of cardiac damage later in life and sharply reduces the risk of vertical transmission when given before conception. In adults with chronic cardiomyopathy, benznidazole has no impact on disease progression and mortality risk. Neither benznidazole nor nifurtimox is effective against digestive complications. Treatment is contraindicated during pregnancy and in advanced renal or hepatic failure. Preferred regimens and drug tolerance vary with age. Adverse events are more frequent among adults, who are therefore at increased risk of premature treatment discontinuation (Table 222-3). As benznidazole seems better tolerated than nifurtimox in adults, it is the recommended first-line drug in this age range. Close (e.g., weekly) clinical and biological monitoring is necessary during treatment. While treatment is usually prescribed for 60 days, the optimal duration remains a matter of debate, with a growing interest in shorter courses.

Treatment should be undertaken for all children, women of child-bearing age, patients in the acute phase, and patients with reactivation. Given the uncertainties about the impact of treatment, the decision to treat patients >18 years old who have the indeterminate form of the chronic phase should be made on an individual basis after discussing the pros and cons with the patient. A negative pregnancy test is mandatory before initiating treatment as the recommended drugs have not been proven to be safe in pregnancy. The efficacy of second-line treatment (e.g., nifurtimox after failure with benznidazole) has not been evaluated to date.

The limited efficacy of current regimens and the understanding that living parasites are a driver of immunopathologic processes have fueled interest in novel therapeutic approaches. These include the addition of immunomodulatory interventions to antiparasitic treatment and the use of combinations of antiparasitic drugs. Drugs can be obtained through the CDC (Parasitic Diseases Public Inquiries line [404-718-4745] or parasites@cdc.gov), the CDC Drug Service (404-639-3670), or the CDC Emergency Operations Center (770-488-7100). In 2017, benznidazole was approved by the U.S. Food and Drug Administration for treatment of children 2–12 years of age.

NONETIOLOGIC TREATMENT

Loading [Contrib]/a11y/accessibility-menu.js

The management of Chagas cardiomyopathy generally follows the management guidelines for heart failure, conduction disturbances, or ventricular arrhythmia of other etiologies. Given the high risk of sudden death, early initiation of treatment with amiodarone or implantation of a cardioverter defibrillator should be considered in the presence of pathologic electrophysiologic abnormalities. Anticoagulation is recommended for primary and secondary prevention of cardioembolic events in the presence of an intramural thrombus or apical aneurysm. Strict control of other cardiovascular risk factors is warranted. Chagas cardiomyopathy is a prominent indication for heart transplantation in Latin America; some evidence indicates that the results are better than in cardiomyopathy of other etiologies. Posttransplantation immunosuppression requires close monitoring, given the high risk of reactivation.

Treatment of digestive dysmotility includes dietary counseling and meals rich in fiber and hydration, with smaller portions eaten more frequently. Drugs releasing the lower esophageal sphincter (e.g., nifedipine or isosorbide dinitrate before meals), pneumatic balloon dilatation, or laparoscopic myotomy improves upper gastrointestinal symptoms in the early stage. Use of botulinum toxin is effective but requires repeated injections. Laxatives and enemas alleviate chronic constipation in most patients. Surgery is indicated in patients with distressing symptoms that are refractory to medical treatment.

CLINICAL FOLLOW-UP

Defining the optimal cure after treatment remains very challenging and is a crucial topic of research. While the search for biomarkers (including through proteomics) to identify early indicators of treatment response holds some promise, serologic follow-up remains the cornerstone of posttreatment monitoring in the acute phase. In the chronic phase, there is no assay of proven value for documentation of response. The time needed for negative seroconversion after treatment indeed depends on the duration of infection. The interval is short (usually months) when infection is treated during the acute (including congenital) phase. In contrast, decades are required in adults infected during childhood. A positive result in a posttreatment PCR indicates treatment failure, but a negative result cannot be interpreted because of the low sensitivity of PCR during the chronic phase. The status of patients with negative PCR results but persistent positive serology is therefore uncertain, but these patients should be considered potentially infective as long as serologic tests continue to yield positive results. All patients, treated or not, should be regularly monitored. The basic yearly assessment includes history-taking for detection of new symptoms, clinical examination, and 12-lead ECG.

TABLE 222-3 Chagas Treatment Regimens and Adverse Reactions to Benznidazole and Nifurtimox

Drug	Regimen	Duration	Adverse Events in Adults (Frequency)	Premature Discontinuation (Rate)
Benznidazole	Age <12 years: 5–7.5 mg/kg per day in 2 doses Age >12 years: 5 mg/kg per day in 2 doses Age <10 years: 15–20 mg/kg per day in 3 or 4 doses	30–60 days	Allergic dermatitis (29–50%), anorexia and weight loss (5–40%), paresthesia (0–30%), peripheral neuropathy (0–30%), nausea and vomiting (0–5%), leukopenia and thrombocytopenia (<1%)	7–20%
Nifurtimox	Age 11–16 years: 12.5–15 mg/kg per day in 3 or 4 doses Age >16 years: 15 mg/kg per day in 3 or 4 doses	60–90 days	Anorexia and weight loss (50–81%), nausea and vomiting (15–50%), abdominal discomfort (12–40%), headaches (13–70%), dizziness and vertigo (12–33%), anxiety and depression (10–49%), insomnia (10–54%), myalgia (13–30%), peripheral neuropathy (2–5%), memory loss (6–14%), leukopenia (<1%)	6–44%

Source: C Bern: N Engl J Med 373:456, 2015.

PREVENTION

In the absence of a vaccine, preventive measures—primary (prevention of *T. cruzi* transmission), secondary (avoidance of complications), and tertiary (reduction of morbidity and mortality)—are necessary. Screening of blood donations is being progressively implemented in endemic areas and in countries to which high-risk groups are immigrating, and screening should be extended to organ donation. When sustained over prolonged periods, vector control is an effective and cost-effective strategy to curb intradomiciliary transmission. Insecticide-impregnated bed nets (as used for malaria) provide individual protection against reduviid bug bites. Screening of child-bearing-age and pregnant Latin American migrant women has been highly cost-effective in Spain, although the cost per case detected varies with the prevalence of infection in the targeted population. Early identification of cases through passive and active screening of the population at risk, along with provision of treatment, may reduce the risk of complications and secondary transmission, particularly congenital transmission. Finally, identification and treatment of cardiac complications and prevention of cardioembolic events at an early stage positively influence the disease course.

GLOBAL CONSIDERATIONS

Image not available. With its geographic expansion, Chagas disease has become a global health issue, predominantly affecting vulnerable p other neglected tropical diseases, progress against Chagas is limited by a lack of research and

Loading [Contrib]/a11y/accessibility-menu.js

development and a lack of financial and political commitment. For example, the production and registration of existing drugs, and access to them, are still problematic in many countries, including the United States. Research on and development of new drugs are compounded by the lack of financial incentives. The future of Chagas disease is likely to be influenced by global phenomena. Climatic changes, population aging, increasing prevalences of noncommunicable comorbidities (e.g., diabetes, hypertension) in low- and middle-income countries, and increasing use of immunosuppressive drugs are likely to impact the epidemiology, clinical course, and burden of Chagas disease. To tackle these challenges, clinical, public health, and policy interventions need to be scaled up and improved in areas of high or hidden prevalence (e.g., in the Chaco Region of Argentina, Bolivia, and Paraguay and in Mexico, Western Europe, and the United States, respectively).

HUMAN AFRICAN TRYPANOSOMIASIS (SLEEPING SICKNESS)

DEFINITION

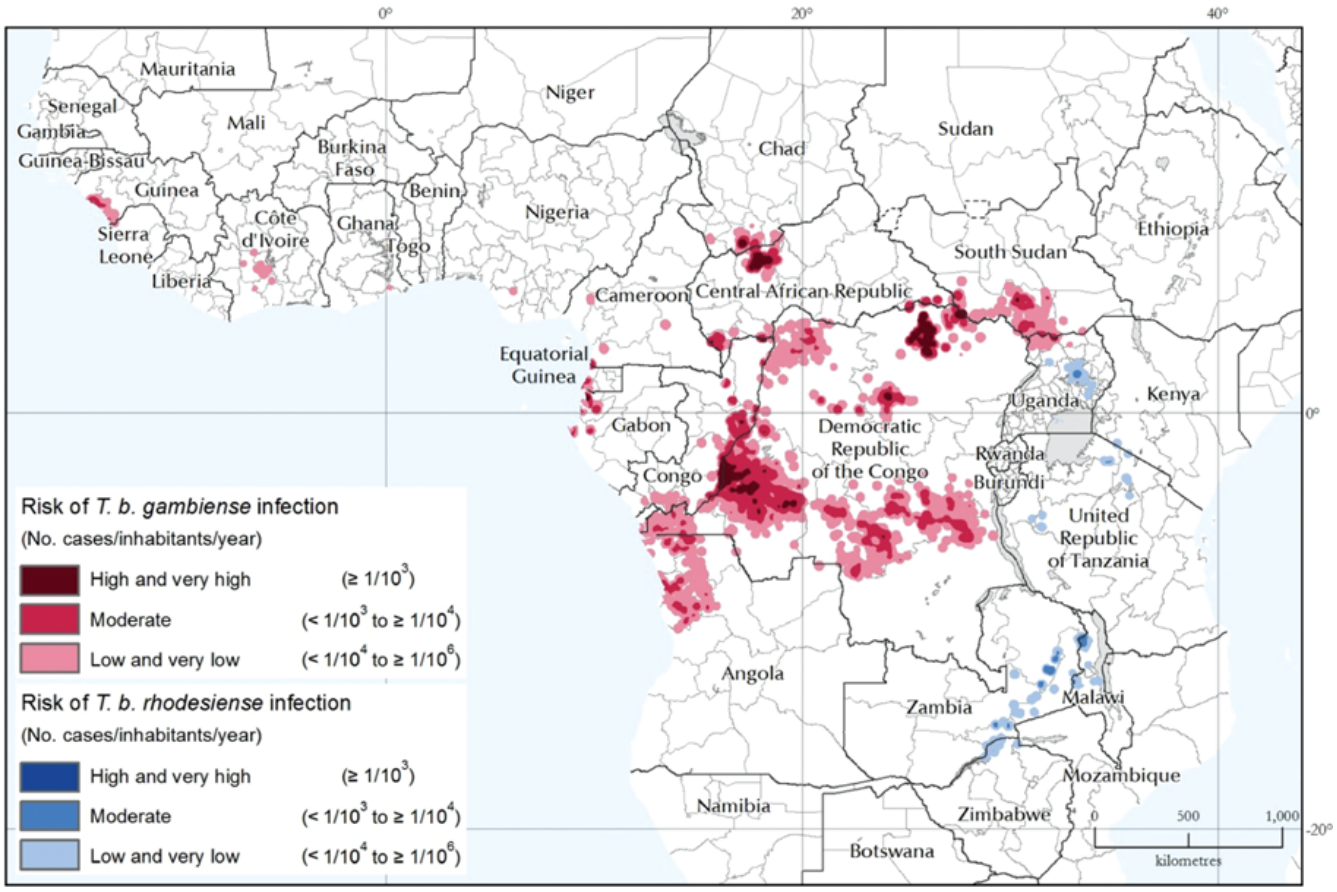
HAT is a life-threatening illness caused by infection with extracellular protozoan parasites that are transmitted by tsetse flies in sub-Saharan Africa. *T. b. gambiense* and *T. b. rhodesiense* are the two pathogenic subspecies affecting humans; their epidemiologic and clinical features largely differ.

EPIDEMIOLOGY

Image not available. The geographic range of HAT is restricted to sub-Saharan Africa in line with the distribution of its vector, the tsetse fly (*Glossina* species; **Fig. 222-3**). HAT due to *T. b. gambiense* is endemic in 24 countries of western and central Africa. Between 1999 and 2015, the number of reported cases fell by 90% (from 27,862 to 2733) as a result of successful control measures based on systematic screening of populations at risk, diagnostic confirmation, and treatment of infected individuals. During the same period, the number of reported cases of HAT due to *T. b. rhodesiense* fell by 89% (from 619 to 71) in the 13 disease-endemic countries of eastern and southeastern Africa. However, the ratio of reported to unreported cases remains uncertain for disease caused by both species. In 2015, most cases of *T. b. gambiense* HAT were reported by the Democratic Republic of the Congo (DRC; 86%), whereas Malawi and Uganda reported most of the cases caused by *T. b. rhodesiense* (42 and 39%, respectively). The geographic distributions of *T. b. gambiense* and *T. b. rhodesiense* do not overlap, but the two species are present in distinct regions of Uganda and the DRC. A roadmap for HAT elimination as a public health problem is being mapped out by the World Health Organization; the objective is to decrease the frequency of new cases to <1 per 10,000 people in endemic areas by 2020.

FIGURE 222-3

Areas at risk for human African trypanosomiasis, 2010–2014. *(Courtesy of Jose Ramon Franco, MD.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition

Copyright © McGraw-Hill Education. All rights reserved.

Loading [Contrib]/a11y/accessibility-menu.js

Humans are the predominant or exclusive reservoir of *T. b. gambiense*. Rare cases of vertical (in utero) or transfusional transmission have been reported, but almost all patients are infected by the bite of tsetse flies during their daily activities along or near rivers, where the flies live and reproduce. In contrast, *T. b. rhodesiense* causes zoonosis in a variety of wild and domesticated animals (e.g., antelopes and cattle, respectively), which act as reservoirs. Humans are infected by *T. b. rhodesiense* via tsetse bites in woodland savannah. Honey gatherers, game park rangers, poachers, and firewood collectors are particularly at risk. Imported cases of HAT are occasionally diagnosed among African immigrants and other travelers. While long-term travelers (>30 days) are at increased risk of *T. b. gambiense* HAT, most imported cases of *T. b. rhodesiense* HAT are seen in short-term travelers, typically following visits to game parks.

PATHOLOGY AND PATHOGENESIS

T. b. rhodesiense and *T. b. gambiense*, unlike other trypanosome species, can infect humans because they resist lytic factors in human serum—namely, apolipoprotein L-1. The serum resistance–associated protein is responsible for resistance in *T. b. rhodesiense*, whereas other mechanisms, notably involving the *T. b. gambiense*–specific glycoprotein (TgsGP) gene, are used by *T. b. gambiense*.

Trypanosomes are transmitted to humans by the tsetse bite, proliferate, and induce a local inflammatory reaction that is sometimes clinically apparent as a chancre. Trypanosomes then disseminate into the hematolymphatic system, with lymph nodes becoming enlarged after infiltration by mononuclear cells and lymphocytes. The degree of enlargement of the liver and spleen is usually mild to moderate, with infiltration by mononuclear cells as a prominent feature. Trypanosomes multiply in the blood, but their presence and density vary. This variation is mainly due to a cyclic immune-evasion process, whereby the parasite population can be decimated by the host's immune response until the reemergence of offspring parasites that express a different variant surface glycoprotein to which the immune system is temporarily blind. Each trypanosome genome encodes a repertoire of ~1000 variant surface glycoproteins between which the parasites can switch genetically. Trypanosomes also multiply in extravascular tissues during the first stage of illness. The skin, skeletal muscles, serous membranes (peritoneum, pleurae, and pericardium), and heart can be involved, with interstitial infiltration of mononuclear cells and vasculitis evident on microscopic examination. Myocarditis and pericarditis with myocardial degeneration and interstitial hemorrhage are common features of *T. b. rhodesiense* infection.

The CNS is invaded weeks to months (*T. b. rhodesiense*) or months to years (*T. b. gambiense*) after initial infection. This invasion corresponds to the second stage of HAT, which is defined by the presence of trypanosomes or mononuclear cells in the cerebrospinal fluid (CSF). The white matter is predominantly affected, with perivascular proliferation of astrocytes, microglial cells, and Mott's (morular) cells that contain IgM in intracellular vacuoles. The location of white-matter lesions in the brain correlates with the main neurologic clinical features. The cerebral cortex and neurons are spared until the terminal stages of illness. Because reversible inflammatory lesions predominate over the irreversible destruction of tissue, neuropsychiatric symptoms and signs resolve partially or completely during or after treatment of second-stage HAT.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Human African Trypanosomiasis

HAT is usually lethal in the absence of treatment, and treatment is simpler and safer during the first stage of illness. Therefore, early diagnosis is crucial; physicians should include HAT in the differential diagnosis of several clinical syndromes when a patient has traveled or lived in at-risk sub-Saharan African countries, and obtaining a thorough recent and remote travel history from the patient is a prerequisite for diagnosis. In particular, HAT due to *T. b. gambiense* should be suspected in patients with persistent and intermittent fever or headaches, progressive neuropsychiatric disorders, and biological signs of systemic inflammation, even if the last exposure occurred several years previously. HAT due to *T. b. rhodesiense* should be suspected in patients with an acute febrile illness and a recent exposure to tsetse flies in an eastern African country, especially if diagnostic tests for malaria are negative.

CLINICAL MANIFESTATIONS

The clinical presentations of *T. b. gambiense* and *T. b. rhodesiense* HAT usually differ. *T. b. gambiense* HAT is a slowly evolving illness with a long incubation period (months to years) and a prolonged disease course. In contrast, *T. b. rhodesiense* HAT is an acute febrile illness with a short (<3-week) incubation period and a shorter (weeks to months) disease course. There are exceptions to this classical pattern. Acute forms of *T. b. gambiense* HAT have been reported, especially among travelers, and chronic forms of *T. b. rhodesiense* HAT occur in the southern range of its geographic distribution (e.g., Zambia and Malawi). Trypanotolerance—i.e., the long-term persistence of parasites without clinical features of disease—is increasingly being reported for *T. b. gambiense*. Concomitant HIV co-infection does not seem to predispose individuals to an increased risk of HAT, and the virus's impact on the clinical presentation of HAT is not known.

T. b. gambiense

The occurrence of trypanosomal chancre is reported in a sizeable proportion of travelers, but very rarely in patients living in endemic areas, where the nonpurulent, painful, and itchy nodule can easily be confused with the bite of another arthropod. The chancre spontaneously disappears in 1–3 weeks.

SYSTEMIC FEATURES

After an asymptomatic incubation period that usually lasts for weeks or months but occasionally lasts for years, patients may present with irregular and remittent fever, sometimes accompanied by fatigue, malaise, and myalgia. Fever is more frequent among travelers than among natives, but the absence of fever in no way rules out the disease. Circinate or serpiginous rashes, commonly called *trypanids*, can occur on the trunk and on proximal parts of the extremities. Trypanids are almost impossible to detect on dark skin and have been reported only in Caucasians. Pruritus is a common but nonspecific symptom that affects up to half of patients during the second stage. Painless edema of the face and extremities occasionally occurs during the first phase.

Loading [Contrib]/a11y/accessibility-menu.jsn of HAT—are detected in 38–85% patients at both disease stages. Cervical palpation is essential in patients with suspected HAT. The lateroposterior cervical group (Winterbottom sign) and the supraclavicular group are most commonly

affected. Lymph nodes are movable, soft initially, harder later, and painless. A variable proportion of patients present with mild to moderate hepatomegaly and splenomegaly. Signs of myocarditis and pericarditis are occasionally detected by ECG and echocardiography but are usually clinically silent. Symptoms of HAT may mimic hypothyroidism or adrenal insufficiency, but thyroid and adrenal function tests yield normal results. Loss of libido, impotence, and amenorrhea, with decreased levels of testosterone and estradiol, are common in second-stage patients and are most likely caused by dysfunction of the hypothalamic–pituitary axis.

NEUROPSYCHIATRIC FEATURES

Most patients with second-stage illness have no or only mild specific neuropsychiatric symptoms and signs, which, when they develop, tend to do so late in the disease course. In contrast, some nonspecific features, such as headaches and mood and behavioral changes, are present in both disease stages but become more permanent and severe during the second stage. As mentioned earlier, HAT is commonly called “sleeping sickness” because of various sleep disturbances (daytime somnolence, nocturnal insomnia) that are more pronounced later in the second stage. Dysregulation of the daily sleep/wake cycle and fragmentation of sleeping patterns are characteristic. Depending on the area of the brain affected, various neurologic syndromes can also develop, including disorders that are pyramidal-related (e.g., motor weakness, rare instances of hemiplegia), extrapyramidal-related (e.g., rigidity, paratonia), and cerebellar-related (e.g., ataxia, abnormal gait). Fine tremor, resting myoclonus, and abnormal (athetoid or choreic) movements have also been reported. Mental disorder is a key feature of HAT and can easily be misdiagnosed as primary psychiatric illness. Common presentations are antisocial or aggressive behavior, mood disorders (e.g., irritability, indifference), apathy or hyperactivity, and depression or psychosis (e.g., delirium, hallucinations). In the final stage of illness, decreased consciousness, dementia, and sometimes epilepsy are present, leading to coma, bed sores, aspiration pneumonia, or other bacterial infections and ultimately to death.

T. b. rhodesiense

The clinical presentation of *T. b. rhodesiense* HAT can be similar to that of *T. b. gambiense* HAT in areas (e.g., Zambia, Malawi) that characteristically harbor specific parasite genotypes and host factors. The typical acute form with an incubation period of <3 weeks occurs in the northern range of the disease’s distribution (e.g., Tanzania, Uganda) and in travelers. The initial trypanosomal chancre is clinically similar to that seen in *T. b. gambiense* HAT but is more common, especially among travelers.

SYSTEMIC FEATURES

Fever can be high and occurs in both first- and second-stage patients, often in association with headaches and with diffuse myalgia and arthralgia. Pruritus and edema of the face and legs can be present. Lymphadenopathies have been reported in variable proportions in both disease stages and predominately affect the submandibular, axillary, and inguinal regions. Mild to moderate hepatomegaly and splenomegaly are documented in a minority of patients. Myocarditis and pericarditis appear to influence clinical course and outcome, even though clinical features of cardiac failure or arrhythmia have not been prominent findings in large case series. In contrast, conduction abnormalities, with various degrees of atrioventricular block, have been reported in travelers. Sepsis-like features, with disseminated intravascular coagulation and multiple-organ failure, can occur in the terminal stage.

NEUROPSYCHIATRIC FEATURES

Neuropsychiatric symptoms and signs in *T. b. rhodesiense* HAT are reported with varying frequency but overall are similar to those described above for *T. b. gambiense* HAT. The notable exception in *T. b. rhodesiense* disease is a more rapid evolution toward coma and death.

DIAGNOSIS

The clinical and biological features of *T. b. gambiense* and *T. b. rhodesiense* HAT—anemia, thrombocytopenia, elevated levels of C-reactive protein and IgM—are not sufficiently specific and current drug regimens are not sufficiently simple and safe to allow the initiation of treatment solely on the basis of suspicion. Diagnostic confirmation is therefore mandatory in all patients.

T. b. gambiense

The diagnosis of *T. b. gambiense* HAT is based on a three-step approach: screening, diagnostic confirmation, and staging.

SCREENING

Immunologic (serologic) methods constitute the preferred screening tool. The card agglutination test for trypanosomiasis (CATT) has been used in most endemic areas for several decades. The test reagent contains stained, freeze-dried trypanosomes of selected variable-antigen types. If specific antibodies are present in the patient’s blood or serum, agglutination can be seen with the naked eye. The sensitivity of the CATT on undiluted blood or serum is 69–100% (>90% in most studies), with some regional variation; its specificity is 84–99%. The CATT and associated equipment (e.g., a rotator) are manufactured and distributed by the Institute of Tropical Medicine in Antwerp, Belgium, but are not widely available outside endemic areas. In recent years, lateral flow tests have been developed and commercialized, first based on whole parasites and later on recombinant antigens. Their diagnostic performance appears similar to that of the CATT. Other serologic test formats (ELISA, immunofluorescence, indirect hemagglutination) are available in some reference laboratories in both endemic and nonendemic countries.

DIAGNOSTIC CONFIRMATION

The microscopic observation of trypanosomes in the lymph, blood, or CSF confirms the diagnosis. Direct observation of motile trypanosomes on a wet preparation of lymph obtained by cervical lymph-node puncture is simple and cheap but has limited sensitivity (50–65% in most studies). Trypanosomes are present in the blood but often occur at low densities. Therefore, stained thin and thick blood smears have very low sensitivity. Sensitivity is improved (to 40–60% in most studies) with the microhematocrit centrifugation technique, which is based on

Loading [Contrib]/a11y/accessibility-menu.js

microscopic examination of the buffy coat after centrifugation of four to six microhematocrit tubes. The most sensitive method (~90%) is the miniature anion-exchange centrifugation technique, which is based on the visualization of trypanosomes in eluate after the passage of a large volume (500 µL) of blood through an anion-exchange column and subsequent centrifugation.

STAGING

As long as treatment of first- and second-stage HAT differs, staging remains an obligatory diagnostic step and is based on the examination of CSF obtained by lumbar puncture. Second-stage HAT is defined by the presence in CSF of a raised leukocyte count (>5/µL) and/or of trypanosomes. The latter can be detected in the cell-counting chamber or, preferably, after centrifugation of the CSF.

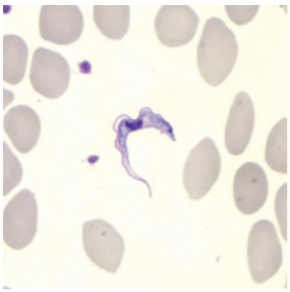
Several molecular methods based on PCR or loop-mediated isothermal amplification have been developed, mostly based on the detection of multiple-copy DNA targets of the Trypanozoon group (to which *T. brucei* belongs) or the single-copy TgsGP gene of *T. b. gambiense.* None of these methods have been fully validated for diagnostic purposes, and a positive result of their application to blood should be interpreted as suspected rather than confirmed HAT. Molecular methods applied to CSF (to detect biomarkers) have not proven more accurate than classical methods for staging and have yielded false-positive results in a substantial proportion of cases.

T. b. rhodesiense

The diagnosis of *T. b. rhodesiense* HAT is usually simpler because parasites are more numerous in body fluids. They can occasionally be visualized in a chancre aspirate. In light of the lack of available serologic tests and the high sensitivity of parasite detection methods in blood, wet mounts, and thin/thick smears **(Fig. 222-4)**, the microhematocrit or other concentration techniques are used for both screening and confirmation. For staging, the definition and methods used are the same as for *T. b. gambiense* HAT.

FIGURE 222-4

***Trypanosoma brucei rhodesiense* in blood** (thin smear, Giemsa stain). *(Credit to the DPDx team, U.S. Centers for Disease Control and Prevention, Atlanta.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT Human African Trypanosomiasis

The management of HAT is based on general supportive therapy (e.g., rehydration, pain management), treatment of concomitant infections (e.g., malaria, pneumonia), and antiparasitic treatment. The modalities of antitrypanosomal treatment depend on the *Trypanosoma* species, the stage of illness, and the presence of contraindications **(Table 222-4)**.

T. B. GAMBIENSE

Pentamidine isethionate is highly effective (>95%) against first-stage *T. b. gambiense* HAT. It is generally well tolerated and can therefore be administered in peripheral health care centers in endemic countries **(Fig. 222-5)**. Hypotension after injection is common but generally mild. Hypoglycemia or hyperglycemia occasionally occurs, but permanent diabetes is very rare. Pain at the injection site is common after intramuscular (IM) injections, but local sterile or bacterial abscesses are rare if basic aseptic precautions are taken. Severe adverse events, such as acute pancreatitis and anaphylaxis, occur extremely rarely.

Nifurtimox–eflornithine combination therapy is also extremely effective (>95% cure rate) and is safer than 14 days of eflornithine monotherapy for treatment of second-stage *T. b. gambiense* HAT. Common adverse reactions include gastrointestinal disturbances (nausea, vomiting, abdominal pain), headaches, anorexia, and reversible bone-marrow toxicity (anemia, leukopenia). Convulsions and psychosis are reported in fewer than 5% of patients.

T. B. RHODESIENSE

Suramin has been used for >90 years and remains the first-line treatment for first-stage *T. b. rhodesiense* HAT. Common adverse events are pyrexia and nephrotoxicity, which is usually mild and reversible but necessitates surveillance of albuminuria and renal function before each dose.

As eflornithine is ineffective against *T. b. rhodesiense*, melarsoprol, an arsenic-based derivative, remains the only existing treatment for second-stage *T. b. rhodesiense* HAT. Reactive encephalopathy is a life-threatening adverse event that occurs in 5–18% of patients, with an a[illegible]he efficacy of concomitant high-dose prednisolone to prevent reactive encephalopathy in patients with

Loading [Contrib]/a11y/accessibility-menu.js

T. b. rhodesiense HAT is not known. Other severe but less frequent adverse reactions to melarsoprol include exfoliative dermatitis, bloody diarrhea, peripheral neuropathy, renal dysfunction, and liver toxicity. Phlebitis is common, as is soft tissue necrosis if the drug is accidentally given paravenously.

TABLE 222-4 Treatment of Human African Trypanosomiasis (HAT)

Disease and Stage	First-Line Treatment		Alternative Treatment
	Drug(S) and Route	Dose and Duration	
***T. b. gambiense* HAT**[a]			
First stage	Pentamidine isethionate IM or IV[a]	4 mg/kg per day for 7 days	Suramin IV[b]
Second stage	Eflornithine IV + nifurtimox PO	Eflornithine: 200 mg/kg bid for 7 days Nifurtimox: 5 mg/kg tid for 10 days	Eflornithine IV: 100 mg/kg qid for 14 days[c]
***T. b. rhodesiense* HAT**			
First stage	Suramin IV	4–5 mg/kg on day 1 followed by 5 weekly injections of 20 mg/kg (e.g., days 3, 10, 17, 24, 31)[d]	Pentamidine isethionate IM or IV[e]
Second stage	Melarsoprol IV	2.2 mg/kg per day for 10 days	—

[a]For IV administration, slow infusion (60–120 min) should be used. [b]Use only if there is a strict contraindication to pentamidine and after exclusion of concomitant onchocerciasis (in which there is a risk of severe immunologic reaction to suramin). The dose and duration are the same as for first-stage *T. b. rhodesiense* HAT. [c]Use if there is a contraindication to nifurtimox, such as allergy, severe epilepsy, or psychosis. [d]The maximal dose is 1 g per injection; drug should be diluted in distilled water. [e]Use at the same dose and for the same duration as for first-stage *T. b. gambiense* HAT.

Source: Control and surveillance of human African trypanosomiasis: Report of a WHO Expert Committee. WHO Technical Report Series 984, 2013.

FIGURE 222-5

Intramuscular injection of pentamidine by a nurse in a village health center, Province Orientale, Democratic Republic of the Congo.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PROGNOSIS

More than 95% of patients with first-stage and second-stage *T. b. gambiense* HAT are definitively cured with pentamidine and nifurtimox–eflornithine combination therapy, respectively. The overall case–fatality rate is <1% except in very advanced cases. As relapses can occur long after completion of treatment, follow-up visits are advised every 6 months for at least 2 years. The relapse rate is very low (<2%) with c and CSF examinations during follow-up visits are no longer standard but can be restricted to

Loading [Contrib]/a11y/accessibility-menu.js

symptomatic patients. Patients with second-stage *T. b. rhodesiense* HAT are at a 5–10% risk of dying during or after melarsoprol treatment, but relapses are very rare.

GLOBAL CONSIDERATIONS

Image not available. The elimination of sleeping sickness as a public health problem is in sight, thanks to increased control activities run by national control programs and nongovernmental medical organizations, improved funding, and the end of several civil wars (e.g., in Angola) in the last 15 years. Funding for research, development, and implementation of improved diagnostic methods (e.g., rapid diagnostic tests) and therapeutic tools (e.g., oral drugs) remains crucial to sustain recent achievements.

FURTHER READING

Brun R et al: Human African trypanosomiasis. Lancet 375:148, 2010. [PubMed: 19833383]
Morillo CA et al: Randomized trial of benznidazole for chronic Chagas disease cardiomyopathy. N Engl J Med 373:1295, 2015. [PubMed: 26323937]
Pérez-molina JA et al: Old and new challenges in Chagas disease. Lancet Infect Dis 15:1347, 2015. [PubMed: 26231478]
Rassi A Jr et al: Chagas disease. Lancet 375:1388, 2010. [PubMed: 20399979]
Urech K et al: Sleeping sickness in travelers—Do they really sleep? PLoS Negl Trop Dis 5:e1358, 2011. [PubMed: 22069503]

Loading [Contrib]/a11y/accessibility-menu.js

Chapter 223: *Toxoplasma* Infections

Kami Kim

DEFINITION

Toxoplasmosis is caused by infection with the obligate intracellular parasite *Toxoplasma gondii*. Acute infection acquired after birth may be asymptomatic but is thought to result in the lifelong chronic persistence of cysts in the host's tissues. In both acute and chronic toxoplasmosis, the parasite is responsible for clinically evident disease, including lymphadenopathy, encephalitis, myocarditis, and pneumonitis. *Congenital* toxoplasmosis is an infection of newborns that results from the transplacental passage of parasites from an infected mother to the fetus. These infants may be asymptomatic at birth, but many later manifest a wide range of signs and symptoms, including chorioretinitis, strabismus, epilepsy, and psychomotor retardation. In immunocompetent individuals, toxoplasmosis can also present as acute disease (typically chorioretinitis) associated with food- or waterborne sources.

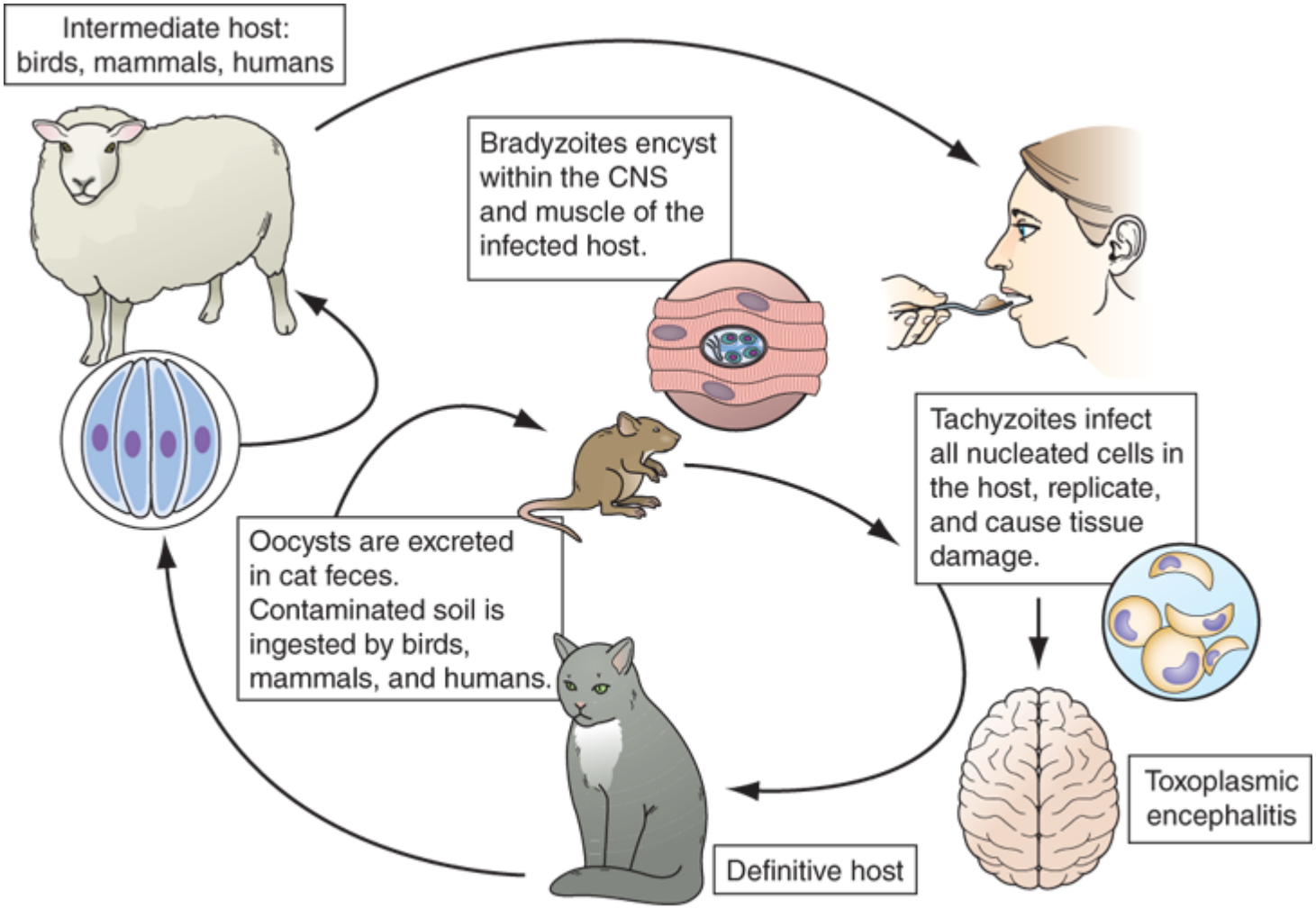
ETIOLOGY

T. gondii is an intracellular coccidian that infects both birds and mammals. There are two distinct stages in the life cycle that are transmissible to humans **(Fig. 223-1)**. Tissue cysts that contain bradyzoites are transmitted in undercooked meat. After an intermediate host (e.g., a human, mouse, sheep, pig, or bird) ingests the cyst, it is rapidly digested by the acidic-pH gastric secretions. Sporulated oocysts that contain sporozoites are products of the sexual cycle in feline intestines and are acquired by ingestion of food or water contaminated with infected cat feces. Bradyzoites or sporozoites are released, enter the intestinal epithelium, and transform into rapidly dividing tachyzoites. The tachyzoites can infect and replicate in all mammalian cells except red blood cells. The parasite actively penetrates the cell and forms a parasitophorous vacuole. Parasite replication continues within the vacuole. After the parasites reach a critical mass, intracellular signaling within the host and the parasite, including calcium fluxes, results in parasite egress from the vacuole. The host cell is destroyed, and the released tachyzoites infect adjoining cells. Parasites can disseminate throughout the body as free tachyzoites or within phagocytic cells in the bloodstream or via lymphatics. Tachyzoites actively invade host cells and can cross epithelial barriers.

FIGURE 223-1

Life cycle of *Toxoplasma gondii.* The cat is the definitive host in which the sexual phase of the cycle is completed. Oocysts shed in cat feces can infect a wide range of animals, including birds, rodents, grazing domestic animals, and humans. The bradyzoites found in the muscle of food animals may infect humans who eat insufficiently cooked meat products, particularly lamb and pork. Although human disease can take many forms, congenital infection and encephalitis from reactivation of latent infection in the brains of immunosuppressed persons are the most important manifestations. CNS, central nervous system. (*Courtesy of Dominique Buzoni-Gatel, Institut Pasteur, Paris; with permission.*)

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The tachyzoite replication cycle within an infected organ causes cytopathology and clinical symptoms. Most tachyzoites are eliminated by the host's humoral and cell-mediated immune responses. Tissue cysts containing bradyzoites develop 7–10 days after systemic tachyzoite infection. These tissue cysts occur in various host organs but persist principally within the central nervous system (CNS) and muscle. The development of this chronic stage completes the asexual portion of the life cycle. Active infection in the immunocompromised host is most likely to be due to the spontaneous release of encysted parasites that undergo rapid transformation into tachyzoites within the CNS and are not contained by the immune system.

The *sexual* stage in the life cycle takes place in the cat (the definitive host) and is defined by the formation of oocysts within the feline host. This enteroepithelial cycle begins with the ingestion of the bradyzoite tissue cysts and, after several intermediate stages, culminates in the production of gametes. Gamete fusion produces a zygote, which envelops itself in a rigid wall and is secreted in the feces as an unsporulated oocyst. After 2–3 days of exposure to air at ambient temperature, the noninfectious oocyst sporulates to produce eight sporozoite progeny. The sporulated oocyst can be ingested by an intermediate host, such as a person emptying a cat's litter box or a pig rummaging in a barnyard. It is in the intermediate host that *T. gondii* completes its life cycle.

Image not available. Sporulated oocysts are environmentally hardy and very infectious; they are thought to be sources of waterborne outbreaks such as those reported in Victoria (British Columbia, Canada) and in South America. Strains found in South America are more virulent than those from the Northern Hemisphere and are more likely to be associated with symptomatic disease—usually ocular

Loading [Contrib]/a11y/accessibility-menu.js

—in immunocompetent individuals; thus ocular toxoplasmosis should be considered in a person from South America with ocular symptoms and retinal abnormalities. Severe disease, including sepsis, fever of unknown origin, and pneumonia, have been reported and should be considered in a patient with a history of travel to South America.

EPIDEMIOLOGY

Image not available. *T. gondii* infects a wide range of mammals and birds. Its seroprevalence depends on the locale and the age of the population. Generally, hot arid climatic conditions are associated with a low prevalence of infection. In the United States and most European countries, seroprevalence increases with age and exposure. In the United States, 13.2% of individuals >6 years old had serologic evidence of exposure in a 2009–2010 survey, with foreign-born Americans having a higher rate of seroprevalence. In most other regions of the world, the seroprevalence is higher. Perhaps because of increased awareness of foodborne infections, the prevalence of seropositivity has decreased worldwide over the past two decades.

TRANSMISSION

Oral Transmission

Most cases of human *Toxoplasma* infection are thought to be acquired by the oral route. Transmission can be attributable to ingestion of sporulated oocysts from contaminated soil, food, or water. During acute feline infection, a cat may excrete as many as 100 million oocysts per day. These sporozoite-containing oocysts are highly infectious and may remain viable for many years in soil or water. Humans infected during an oocyst-transmitted infection develop stage-specific antibodies to the oocyst/sporozoite.

Children and adults also acquire infection from tissue cysts containing bradyzoites. Undercooking or insufficient freezing of meat is an important source of infection in the developed world. More recent epidemiologic studies have associated acute infections with ingestion of untreated water or shellfish (oysters, mussels, and clams).

Transmission via Blood or Organs

In addition to being transmitted orally, *T. gondii* can be transmitted directly from a seropositive donor to a seronegative recipient in a transplanted heart, heart–lung, kidney, liver, or pancreas. Viable parasites can be cultured from refrigerated anticoagulated blood, which may be a source of infection in individuals receiving blood transfusions. *T. gondii* reactivation has been reported in bone marrow, hematopoietic stem cell, and liver transplant recipients as well as in individuals with AIDS. Although antibody titers generally are not useful in monitoring *T. gondii* infection, individuals with higher antibody titers may be at relatively high risk for reactivation after hematopoietic stem cell transplantation; thus routine polymerase chain reaction (PCR) screening of blood from these patients may be in order. Screening of *Toxoplasma* serologies (donor and recipient) before transplantation may identify patients potentially at risk for reactivated toxoplasmosis. Finally, laboratory personnel can be infected after contact with contaminated needles or glassware or with infected tissue.

Transplacental Transmission

On average, about one-third of all women who acquire infection with *T. gondii* during pregnancy transmit the parasite to the fetus; the remainder give birth to normal, uninfected babies. Of the various fLoading [Contrib]/a11y/accessibility-menu.jsestational age at the time of infection is the most critical (see below). Recrudescent maternal infection is rarely the source of congenital disease, although rare

cases of transmission by immunocompromised women (e.g., those infected with HIV or those receiving high-dose glucocorticoids) have been reported. Thus women who are seropositive before pregnancy usually are protected against acute infection and do not give birth to congenitally infected neonates.

There is essentially no risk of congenital infection if the mother becomes infected ≥6 months before conception. If infection is acquired <6 months before conception, the likelihood of transplacental infection increases as the interval between infection and conception decreases. Women with documented acute toxoplasmosis should be counseled to use appropriate measures to prevent pregnancy for 6 months after infection. In pregnancy, if the mother becomes infected during the first trimester, the incidence of transplacental infection is lowest (~15%), but the disease in the neonate is most severe. If maternal infection occurs during the third trimester, the incidence of transplacental infection is greatest (65%), but the infant is usually asymptomatic at birth. Infected infants who are normal at birth may have a higher incidence of learning disabilities and chronic neurologic sequelae than uninfected children. Only a small proportion (20%) of women infected with *T. gondii* develop clinical signs of infection. Often the diagnosis is first appreciated when routine postconception serologic tests show evidence of specific antibody.

PATHOGENESIS

Upon the host's ingestion of either tissue cysts containing bradyzoites or oocysts containing sporozoites, the parasites are released from the cysts by the digestive process. Bradyzoites are resistant to the effect of pepsin and invade the host's gastrointestinal tract. Within enterocytes (or other gut-associated cells), the parasites undergo morphologic transformation, giving rise to invasive tachyzoites. From the gastrointestinal tract, parasites disseminate to a variety of organs, particularly lymphatic tissue, skeletal muscle, myocardium, retina, placenta, and the CNS. At these sites, the parasite infects host cells, replicates, and invades the adjoining cells. In this fashion, the hallmarks of the infection develop: cell death and focal necrosis surrounded by an acute inflammatory response.

In the immunocompetent host, both the humoral and the cellular immune responses control infection; parasite virulence and tissue tropism may be strain specific. Tachyzoites are sequestered by a variety of immune mechanisms, including induction of parasiticidal antibody, activation of macrophages with radical intermediates, production of interferon γ (IFN-γ), and stimulation of CD8+ cytotoxic T lymphocytes. These antigen-specific lymphocytes are capable of killing both extracellular parasites and target cells infected with parasites. As tachyzoites are cleared from the acutely infected host, tissue cysts containing bradyzoites begin to appear, usually within the CNS, the skeletal muscle, and the retina. Studies indicate that *Toxoplasma* secretes signaling molecules into infected host cells and that these molecules modulate host gene expression, host metabolism, and host immune response. While it was initially thought that cysts with bradyzoites are not eliminated by the immune system, recent studies in the murine model indicate that both CD8+ T cells and alternatively activated macrophages are able to kill cysts in vivo; some cysts persist, however, and the ability to eliminate cysts may depend on the genetic background of the infected host.

In the immunocompromised or fetal host, the immune factors necessary to control the spread of tachyzoite infection are lacking. This altered immune state allows the persistence of tachyzoites and gives rise to progressive focal destruction in affected organs (i.e., necrotizing encephalitis, pneumonia, and myocarditis).

It is thought that all infected individuals have persistent infection with cysts containing bradyzoites, but this lifelong infection usually remains subclinical. Although bradyzoites are in a slow metabolic phase, cysts do degenerate and rupture within the CNS. This degenerative process, with the development of new bradyzoite-containing cysts, is the most probable source of recrudescent infection in i[Loading [Contrib]/a11y/accessibility-menu.js]he most likely stimulus for the persistence of antibody titers in the immunocompetent host. Although the concept is controversial, the persistence of toxoplasmosis

has been hypothesized to be a contributing factor to a variety of neuropsychiatric conditions, including schizophrenia and bipolar disease. In rodents, chronic infection clearly has significant effects on behavior, increasing predation.

PATHOLOGY

Cell death and focal necrosis due to replicating tachyzoites induce an intense mononuclear inflammatory response in any tissue or cell type infected. Tachyzoites rarely can be visualized by routine histopathologic staining of these inflammatory lesions. However, immunofluorescent staining with parasitic antigen–specific antibodies can reveal the organism. In contrast to this inflammatory process caused by tachyzoites, bradyzoite-containing cysts cause inflammation only at the early stages of development. Once the cysts reach maturity, the inflammatory process can no longer be detected, and the cysts remain immunologically quiescent within the brain matrix until they rupture.

Lymph Nodes

During acute infection, lymph node biopsy demonstrates characteristic findings, including follicular hyperplasia and irregular clusters of tissue macrophages with eosinophilic cytoplasm. Granulomas rarely are evident in these specimens. Although tachyzoites are not usually visible, they can be sought either by subinoculation of infected tissue into mice, with resultant disease, or by PCR. PCR amplification of DNA fragments of *Toxoplasma* genes is effective and sensitive in establishing lymph node infection by tachyzoites.

Eyes

In the eye, infiltrates of monocytes, lymphocytes, and plasma cells may produce uni- or multifocal lesions. Granulomatous lesions and chorioretinitis can be observed in the posterior chamber after acute necrotizing retinitis. Other ocular complications include iridocyclitis, cataracts, and glaucoma.

Central Nervous System

During CNS involvement, both focal and diffuse meningoencephalitis can be documented, with evidence of necrosis and microglial nodules. Necrotizing encephalitis in patients without AIDS is characterized by small diffuse lesions with perivascular cuffing in contiguous areas. In the AIDS population, polymorphonuclear leukocytes may be present in addition to monocytes, lymphocytes, and plasma cells. Cysts containing bradyzoites frequently are found contiguous with the necrotic tissue border. As a consequence of combined antiretroviral therapy (cART) for AIDS, the incidence of toxoplasmosis has decreased in the developed world. Its incidence in under-resourced settings is not known but is likely to be high.

Lungs and Heart

Among patients with AIDS who die of toxoplasmosis, 40–70% have involvement of the lungs and heart. Interstitial pneumonitis can develop in neonates and immunocompromised patients. Thickened and edematous alveolar septa infiltrated with mononuclear and plasma cells are apparent. This inflammation may extend to the endothelial walls. Tachyzoites and bradyzoite-containing cysts have been observed within the alveolar membrane. Superimposed bronchopneumonia can be caused by other microbial agents. Cysts and aggregates of parasites in cardiac muscle tissue are evident in patients with AIDS who die of toxoplasmosis. Focal necrosis surrounded by inflammatory cells is associated with hyaline necrosis and disrupted myocardial cells. Pericarditis is associated with t

Loading [Contrib]/a11y/accessibility-menu.js

Gastrointestinal Tract

Rare cases of human gastrointestinal tract infection with *T. gondii* have presented as ulcerations in the mucosa. Acute infection in certain strains of inbred mice (C57BL/6) results in lethal ileitis within 7–9 days. This inflammatory bowel disease has been recognized in several other mammalian species, including pigs and nonhuman primates.

Other Sites

Pathologic changes during disseminated infection are similar to those described for the lymph nodes, eyes, and CNS. In patients with AIDS, the skeletal muscle, pancreas, stomach, and kidneys can be involved, with necrosis, invasion by inflammatory cells, and (rarely) tachyzoites detectable by routine staining. Large necrotic lesions may cause direct tissue destruction. In addition, secondary effects from acute infection of these various organs, including pancreatitis, myositis, and glomerulonephritis, have been reported.

HOST IMMUNE RESPONSE

Acute *Toxoplasma* infection evokes a cascade of protective immune responses in the immunocompetent host. *Toxoplasma* enters the host at the gut mucosal level and evokes a mucosal immune response that includes the production of antigen-specific secretory IgA. Titers of serum IgA antibody directed at the tachyzoite surface antigen p30/SAG1 are a useful marker for congenital and acute toxoplasmosis.

Within the host, *T. gondii* rapidly induces detectable levels of both IgM and IgG serum antibodies. Monoclonal gammopathy of the IgG class can occur in congenitally infected infants. IgM levels may be increased in newborns with congenital infection. The polyclonal IgG antibodies evoked by infection are parasiticidal in vitro in the presence of serum complement and are the basis for the Sabin–Feldman dye test. However, cell-mediated immunity is the major protective response evoked by the parasite during infection. Macrophages are activated after phagocytosis of antibody-opsonized parasites. This activation can lead to death of the parasite by either an oxygen-dependent or an oxygen-independent process. If the parasite is not phagocytosed and enters the macrophage by active penetration, it continues to replicate, and this replication may represent the mechanism for transport and dissemination to distant organs. *Toxoplasma* stimulates a robust interleukin (IL) 12 response by human dendritic cells. The CD4+ and CD8+ T cell responses are antigen-specific and further stimulate the production of a variety of important lymphokines that expand the T cell and natural killer cell repertoire. *T. gondii* is a potent inducer of a T_H1 phenotype, with IL-12 and IFN-γ playing an essential role in the control of the parasites' growth in the host. Regulation of the inflammatory response is at least partially under the control of a T_H2 response that includes the production of IL-4 and IL-10 in seropositive individuals. Human T cell clones of both the CD4+ and the CD8+ phenotypes are cytolytic against parasite-infected macrophages. These T cell clones produce cytokines that are "microbistatic." IL-18, IL-7, and IL-15 upregulate the production of IFN-γ and may be important during acute and chronic infection. The effect of IFN-γ may be paradoxical, with stimulation of a host downregulatory response as well.

Although *T. gondii* infection is thought to be recrudescent in patients with AIDS or other immunocompromised states, antibody titers are not useful in establishing reactivation or in following the activity of infection. An absence of positive serologic results suggests an alternative diagnosis, although AIDS patients may have borderline positive or low serologic values. T cells from AIDS patients with reactivation of toxoplasmosis fail to secrete both IFN-γ and IL-2. This alteration in the production of these critical immune cytokines contributes to the persistence of infection. *Toxoplasma* i[Loading [Contrib]/a11y/accessibility-menu.js]e course of AIDS (CD4+ T cell count, <100/μL), when the loss of T cell–dependent protective mechanisms, particularly CD8+ T cells, becomes most pronounced.

CLINICAL MANIFESTATIONS

In persons whose immune systems are intact, acute toxoplasmosis is usually asymptomatic and self-limited. This condition can go unrecognized in 80–90% of adults and children with acquired infection. The asymptomatic nature of this infection makes diagnosis difficult in mothers infected during pregnancy. In contrast, the wide range of clinical manifestations in congenitally infected children includes severe neurologic complications such as hydrocephalus, microcephaly, mental retardation, and chorioretinitis. If prenatal infection is severe, multiorgan failure and subsequent intrauterine fetal death can occur. In children and adults, chronic infection can persist throughout life, with little consequence to the immunocompetent host.

Toxoplasmosis in Immunocompetent Patients

The most common manifestation of acute toxoplasmosis is cervical lymphadenopathy. The nodes may be single or multiple, are usually nontender, are discrete, and vary in firmness. Lymphadenopathy also may be found in suboccipital, supraclavicular, inguinal, and mediastinal areas. Generalized lymphadenopathy occurs in 20–30% of symptomatic patients. Between 20 and 40% of patients with lymphadenopathy also have headache, malaise, fatigue, and fever (usually with a temperature of <40°C [<104°F]). A smaller proportion of symptomatic individuals have myalgia, sore throat, abdominal pain, maculopapular rash, meningoencephalitis, and confusion. Rare complications associated with infection in the normal immune host include pneumonia, myocarditis, encephalopathy, pericarditis, and polymyositis. Signs and symptoms associated with acute infection usually resolve within several weeks, although the lymphadenopathy may persist for some months. In one epidemic, toxoplasmosis was diagnosed correctly in only 3 of the 25 patients who consulted physicians. If toxoplasmosis is considered in the differential diagnosis, routine laboratory and serologic screening should precede node biopsy.

Image not available. In North America and Europe, there are three predominant genotypes of *T. gondii*, but strains are more genetically diverse in South America. Genotypes of *T. gondii* prevalent in South America are more virulent than those typically seen in North America or Europe. These genotypes may be associated with acute or recurrent ocular disease in immunocompetent individuals and have also been associated with pneumonitis and a fulminant sepsis picture in immunologically normal individuals. Thus a detailed history is critical for establishing a diagnosis.

The results of routine laboratory studies are usually unremarkable except for minimal lymphocytosis, an elevated erythrocyte sedimentation rate, and a nominal increase in serum aminotransferase levels. Evaluation of cerebrospinal fluid (CSF) in cases with evidence of encephalopathy or meningoencephalitis shows an elevation of intracranial pressure, mononuclear pleocytosis (10–50 cells/mL), a slight increase in protein concentration, and (occasionally) an increase in the gamma globulin level. PCR amplification of the *Toxoplasma* DNA target sequence in CSF is specific for active toxoplasmosis, but not sensitive. The CSF of chronically infected individuals is normal.

Infection of Immunocompromised Patients

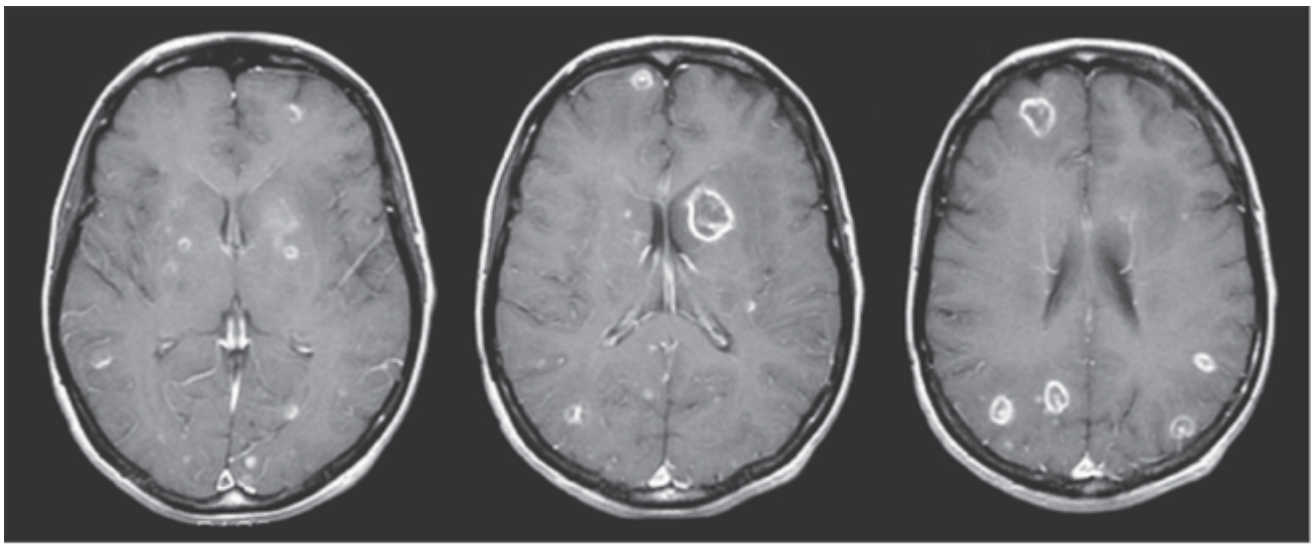
Patients with AIDS and those receiving immunosuppressive therapy for lymphoproliferative disorders are at greatest risk for developing acute toxoplasmosis. Toxoplasmosis has also been reported after treatment with antibodies to tumor necrosis factor. The infection may be due either to reactivation of latent infection or to acquisition of parasites from exogenous sources such as blood or transplanted organs. In individuals with AIDS, >95% of cases of *Toxoplasma* encephalitis (TE) are believed to be due to recrudescent infection. In most of these cases, encephalitis develops when the CD4+ T cell count falls below 100/μL. In immunocompromised hosts, the disease may be rapidly fatal if untreated. TLoading [Contrib]/a11y/accessibility-menu.js of appropriate therapy are necessary to prevent fulminant infection.

Toxoplasmosis is a principal opportunistic infection of the CNS in persons with AIDS. Although geographic origin may be related to frequency of infection, it has no correlation with the severity of disease in immunocompromised hosts. Individuals with AIDS who are seropositive for *T. gondii* are at high risk for encephalitis. Before the advent of current cART, about one-third of the 15–40% of adult AIDS patients in the United States who were latently infected with *T. gondii* developed TE. TE may still be a presenting infection in individuals who are unaware of their positive HIV status.

The signs and symptoms of acute toxoplasmosis in immunocompromised patients principally involve the CNS **(Fig. 223-2)**. More than 50% of patients with clinical manifestations have intracerebral involvement. Clinical findings at presentation range from nonfocal to focal dysfunction. CNS findings include encephalopathy, meningoencephalitis, and mass lesions. Patients may present with altered mental status (75%), fever (10–72%), seizures (33%), headaches (56%), and focal neurologic findings (60%), including motor deficits, cranial nerve palsies, movement disorders, dysmetria, visual-field loss, and aphasia. Patients who present with evidence of diffuse cortical dysfunction develop evidence of focal neurologic disease as infection progresses. This altered condition is due not only to the necrotizing encephalitis caused by direct invasion by the parasite but also to secondary effects, including vasculitis, edema, and hemorrhage. The onset of infection can range from an insidious process over several weeks to an acute presentation with fulminant focal deficits, including hemiparesis, hemiplegia, visual-field defects, localized headache, and focal seizures.

FIGURE 223-2

Toxoplasmic encephalitis in a 36-year-old patient with AIDS. The multiple lesions are demonstrated by MRI scanning (T1-weighted with gadolinium enhancement). (*Courtesy of Clifford Eskey, Dartmouth Hitchcock Medical Center, Hanover, NH; with permission*.)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although lesions can occur anywhere in the CNS, the areas most often involved appear to be the brainstem, basal ganglia, pituitary gland, and corticomedullary junction. Brainstem involvement gives rise to a variety of neurologic dysfunctions, including cranial nerve palsy, dysmetria, and ataxia. With basal ganglion infection, patients may develop hydrocephalus, choreiform movements, and choreoathetosis. *Toxoplasma* usually causes encephalitis, and meningeal involvement is uncommon. CSF findings may be unremarkable or may include a modest increase in cell count and in protein—but not glucose—concentration.

Loading [Contrib]/a11y/accessibility-menu.js

Cerebral toxoplasmosis must be differentiated from other opportunistic infections or tumors in the CNS of AIDS patients. The differential diagnosis includes herpes simplex encephalitis, cryptococcal meningitis, progressive multifocal leukoencephalopathy, and primary CNS lymphoma. Involvement of the pituitary gland can give rise to panhypopituitarism and hyponatremia from inappropriate secretion of vasopressin (antidiuretic hormone). HIV-associated neurocognitive disorder (HAND) may present as cognitive impairment, attention loss, and altered memory. Brain biopsy in patients who have been treated for TE but who continue to exhibit neurologic dysfunction often fails to identify organisms.

Autopsies of *Toxoplasma*-infected patients have demonstrated the involvement of multiple organs, including the lungs, gastrointestinal tract, pancreas, skin, eyes, heart, and liver. *Toxoplasma* pneumonia can be confused with *Pneumocystis* pneumonia (PcP). Respiratory involvement usually presents as dyspnea, fever, and a nonproductive cough and may rapidly progress to acute respiratory failure with hemoptysis, metabolic acidosis, hypotension, and (occasionally) disseminated intravascular coagulation. Histopathologic studies demonstrate necrosis and a mixed cellular infiltrate. The presence of organisms is a helpful diagnostic indicator, but organisms can also be found in healthy tissue. Infection of the heart is usually asymptomatic but can be associated with cardiac tamponade or biventricular failure. Infections of the gastrointestinal tract and the liver have been documented.

Congenital Toxoplasmosis

Between 400 and 4000 infants born each year in the United States are affected by congenital toxoplasmosis. Acute infection in mothers acquiring *T. gondii* during pregnancy is usually asymptomatic; most such women are diagnosed via prenatal serologic screening. Infection of the placenta leads to hematogenous infection of the fetus. As gestation proceeds, the proportion of fetuses that become infected increases, but the clinical severity of the infection declines. Although infected children may initially be asymptomatic, the persistence of *T. gondii* can result in reactivation and clinical disease—most frequently chorioretinitis—decades later. Factors associated with relatively severe disabilities include delays in diagnosis and in initiation of therapy, neonatal hypoxia and hypoglycemia, profound visual impairment (see “Ocular Infection,” below), uncorrected hydrocephalus, and increased intracranial pressure. If treated appropriately, upwards of 70% of children have normal developmental, neurologic, and ophthalmologic findings at follow-up evaluations. Treatment for 1 year with pyrimethamine, a sulfonamide, and folinic acid is tolerated with minimal toxicity (see “Treatment,” below).

Ocular Infection

Image not available. Infection with *T. gondii* is estimated to cause 35% of all cases of chorioretinitis in the United States and Europe. It was formerly thought that the majority of cases of ocular disease were due to congenital infection. New ocular toxoplasmosis in immunocompetent individuals occurs more commonly than was previously appreciated and has been associated with outbreaks in Victoria (British Columbia) and in South America. A variety of ocular manifestations are documented, including blurred vision, scotoma, photophobia, and eye pain. Macular involvement occurs, with loss of central vision, and nystagmus is secondary to poor fixation. Involvement of the extraocular muscles may lead to disorders of convergence and to strabismus. Ophthalmologic examination should be undertaken in newborns with suspected congenital infection. As the inflammation resolves, vision improves, but episodic flare-ups of chorioretinitis, which progressively destroy retinal tissue and lead to glaucoma, are common. The ophthalmologic examination reveals yellow-white, cotton-like patches with indistinct margins of hyperemia. As the lesions age, white plaques with distinct borders and black spots within the retinal pigment become more apparent. Lesions usually are located near the posterior pole of the retina; they may be single but are more commonly multiple. Congenital lesions may be unilateral or bilateral and show evidence of massive chorioretinal degeneration with extensive fibrosis. Surrounding tLoading [Contrib]/a11y/accessibility-menu.jsal retina and vasculature. In patients with AIDS, retinal lesions

are often large, with diffuse retinal necrosis, and include both free tachyzoites and cysts containing bradyzoites. Toxoplasmic chorioretinitis may be a prodrome to the development of encephalitis.

DIAGNOSIS

Tissues and Body Fluids

The differential diagnosis of acute toxoplasmosis can be made by appropriate culture, serologic testing, and PCR **(Table 223-1)**. Although performed only at specialized laboratories, the isolation of *T. gondii* from blood or other body fluids can be accomplished after subinoculation of the sample into the peritoneal cavity of mice. If no parasites are found in the mouse's peritoneal fluid 6–10 days after inoculation, its anti-*Toxoplasma* serum titer can be evaluated 4–6 weeks after inoculation. Isolation of *T. gondii* from the patient's body fluids reflects acute infection, whereas isolation from biopsied tissue is an indication only of the presence of tissue cysts and should not be misinterpreted as evidence of acute toxoplasmosis. Persistent parasitemia in patients with latent, asymptomatic infection is rare. Histologic examination of lymph nodes may suggest the characteristic changes described above. Demonstration of tachyzoites in lymph nodes establishes the diagnosis of acute toxoplasmosis. Like subinoculation into mice, histologic demonstration of cysts containing bradyzoites confirms prior infection with *T. gondii* but is nondiagnostic for acute infection.

Loading [Contrib]/a11y/accessibility-menu.js

TABLE 223-1 Differential Laboratory Diagnosis of Toxoplasmosis

Clinical Setting	Alternative Diagnosis	Distinguishing Characteristics
Mononucleosis syndrome	Epstein-Barr virus infection	Serology/PCR
	Cytomegalovirus infection	Serology/PCR or culture
	HIV infection	Serology/viral load
	Bartonella infection (cat-scratch disease)	Biopsy (PCR or culture)/serology
	Lymphoma	Biopsy
Congenital infection	Cytomegalovirus infection	Viral culture/PCR
	Herpes simplex virus infection	Viral culture/PCR
	Rubella virus infection	Serology
	Syphilis	Serology
	Listeriosis	Bacterial culture
Chorioretinitis in immunocompetent individual	Tuberculosis	Bacterial culture/PCR
	Syphilis	Serology
	Histoplasmosis	Serology/culture/antigen
Chorioretinitis in AIDS patient	Cytomegalovirus infection	Viral culture/PCR
	Syphilis	Serology
	Herpes simplex virus infection	Viral culture/PCR
	Varicella-zoster virus infection	Viral culture/PCR
	Fungal infection	Culture
	Lymphoma or metastatic tumor	Tissue biopsy
CNS lesions in AIDS patient	Brain abscess	Biopsy/culture
	Progressive multifocal leukoencephalopathy	PCR for JC virus
	Fungal infection	Biopsy/culture
	Mycobacterial infection	Biopsy/culture/PCR

Abbreviations: CNS, central nervous system; PCR, polymerase chain reaction.

Source: Adapted from JD Schwartzman: Toxoplasmosis, in *Principles and Practice of Clinical Parasitology*. Hoboken, Wiley, 2001.

Serology

Because some diagnostic tests are available only at specialty laboratories and are technically challenging, serologic testing has become the routine method of diagnosis. Diagnosis of acute infection with *T. gondii* can be established by detection of the simultaneous presence of IgG and IgM antibodies to *Toxoplasma* in serum. The presence of circulating IgA favors the diagnosis of an acute infection. The Sabin–Feldman dye test, the indirect fluorescent antibody test, and the enzyme-linked immunosorbent assay (ELISA) all satisfactorily measure circulating IgG antibody to *Toxoplasma*. Positive IgG titers (>1:10) can be detected as early as 2–3 weeks after infection. These titers usually peak at 6–8 weeks and decline slowly to a new baseline level that persists for life. Antibody avidity increases with time and can be useful in difficult cases during pregnancy for establishing when infection may have occurred. The serum IgM titer should be measured in concert with the IgG titer to better establish the time of infection; either the double-sandwich IgM-ELISA or the IgM-i[illegible]ould be used. Both assays are specific and sensitive, with

Loading [Contrib]/a11y/accessibility-menu.js

fewer false-positive results than other commercial tests. The double-sandwich IgA-ELISA is more sensitive than the IgM-ISAGA for detecting congenital infection in the fetus and newborn. Although a negative IgM result with a positive IgG titer indicates distant infection, IgM can persist for >1 year and should not necessarily be considered a reflection of acute disease. If acute toxoplasmosis is suspected, a more extensive panel of serologic tests can be performed. In the United States, testing is available at the *Toxoplasma* Serology Laboratory at Palo Alto Medical Foundation (*http://www.pamf.org/serology/clinicianguide.html*).

Molecular Diagnostics

Molecular approaches can directly detect *T. gondii* in biologic samples independent of the serologic response. Results obtained with PCR have suggested high sensitivity, specificity, and clinical utility in the diagnosis of TE, and PCR technology may be becoming more readily available in resource-poor settings. Real-time PCR is a promising technique that can provide quantitative results. Isolates can be genotyped and polymorphic sequences can be obtained, with consequent identification of the precise strain. Molecular epidemiologic studies with polymorphic markers have been useful in correlating clinical signs and symptoms of disease with different *T. gondii* genotypes.

The Immunocompetent Adult or Child

For the patient who presents with lymphadenopathy only, a positive IgM titer is an indication of acute infection—and an indication for therapy, if clinically warranted (see "Treatment," below). The serum IgM titer should be determined again in 3 weeks. An elevation in the IgG titer without an increase in the IgM titer suggests that infection is present but is not acute. If there is a borderline increase in either IgG or IgM, the titers should be reassessed in 3–4 weeks.

The Immunocompromised Host

A presumptive clinical diagnosis of TE in patients with AIDS is based on clinical presentation, history of exposure (as evidenced by positive serology), and radiologic evaluation. To detect latent infection with *T. gondii*, HIV-infected persons should be tested for IgG antibody to *Toxoplasma* soon after HIV infection is diagnosed. When these criteria are used, the predictive value is as high as 80%. More than 97% of patients with AIDS and toxoplasmosis have IgG antibody to *T. gondii* in serum. IgM serum antibody usually is not detectable. Although IgG titers do not correlate with active infection, serologic evidence of infection virtually always precedes the development of TE. It is therefore important to determine the *Toxoplasma* antibody status of all patients infected with HIV. Antibody titers may range from negative to 1:1024 in patients with AIDS and TE. Fewer than 3% of patients have no demonstrable antibody to *Toxoplasma* at diagnosis of TE.

Patients with TE have focal or multifocal abnormalities demonstrable by CT or MRI. Neuroradiologic evaluation should include double-dose contrast CT of the head. By this test, single and frequently multiple contrast-enhancing lesions (<2 cm) may be identified. MRI usually demonstrates multiple lesions located in both hemispheres, with the basal ganglia and corticomedullary junction most commonly involved; MRI provides a more sensitive evaluation of the efficacy of therapy than does CT (Fig. 223-2). These findings are not pathognomonic of *Toxoplasma* infection, because 40% of CNS lymphomas are multifocal and 50% are ring-enhancing. For both MRI and CT scans, the rate of false-negative results is ~10%. The finding of a single lesion on an MRI scan increases the likelihood of primary CNS lymphoma (in which solitary lesions are four times more likely than in TE) and strengthens the argument for the performance of a brain biopsy. A therapeutic trial of anti-*Toxoplasma* medications is frequently used to assess the diagnosis. Treatment of presumptive TE with pyrimethamine plus sulfadiazine or clindamycin results in quantifiable clinical improvement in >50% of [illegible] istered to prevent bone marrow toxicity. By day 7, >90% of treated patients show evidence of improvement. In contrast, if patients fail to respond or have

Loading [Contrib]/a11y/accessibility-menu.js

lymphoma, clinical signs and symptoms worsen by day 7. Patients in this category require brain biopsy with or without a change in therapy. This procedure can now be performed by a stereotactic CT-guided method that reduces the potential for complications. Brain biopsy for *T. gondii* identifies organisms in 50–75% of cases. PCR amplification of CSF may also confirm toxoplasmosis or suggest alternative diagnoses (Table 223-1), such as progressive multifocal leukoencephalopathy (JC virus positive) or primary CNS lymphoma (Epstein-Barr virus positive).

CT and MRI with contrast are currently the standard diagnostic imaging tests for TE. As in other conditions, the radiologic response may lag behind the clinical response. Resolution of lesions may take from 3 weeks to 6 months. Some patients show clinical improvement despite worsening radiographic findings.

Congenital Infection

The issue of concern when a pregnant woman has evidence of recent *T. gondii* infection is whether the fetus is infected. PCR analysis of the amniotic fluid for the B1 gene of *T. gondii* has replaced fetal blood sampling. Serologic diagnosis is based on the persistence of IgG antibody or a positive IgM titer after the first week of life (a time frame that excludes placental leak). The IgG determination should be repeated every 2 months. An increase in IgM beyond the first week of life is indicative of acute infection. Up to 25% of infected newborns may be seronegative and have normal routine physical examinations. Thus assessment of the eye and the brain, with ophthalmologic testing, CSF evaluation, and radiologic studies, is important in establishing the diagnosis.

Ocular Toxoplasmosis

The serum antibody titer may not correlate with the presence of active lesions in the fundus, particularly in cases of congenital toxoplasmosis. In general, a positive IgG titer (measured in undiluted serum if necessary) in conjunction with typical lesions establishes the diagnosis. If lesions are atypical and the serum antibody titer is in the low-positive range, the diagnosis is presumptive. The parasitic antigen–specific polyclonal IgG assay as well as parasite-specific PCR may facilitate the diagnosis. Accordingly, the clinical diagnosis of ocular toxoplasmosis can be supported in 60–90% of cases by laboratory tests, depending on the time of anterior chamber puncture and the panel of antibody analyses used. In the remaining cases, the possibility of a falsely negative laboratory diagnosis or of an incorrect clinical diagnosis cannot be clarified further.

TREATMENT

TREATMENT Toxoplasmosis CONGENITAL INFECTION

Congenitally infected neonates are treated with daily oral pyrimethamine (1 mg/kg) and sulfadiazine (100 mg/kg) along with folinic acid for 1 year. Depending on the signs and symptoms, prednisone (1 mg/kg per day) may be used for congenital infection. Some U.S. states and some countries routinely screen pregnant women (France, Austria) and/or newborns (Denmark, Massachusetts). Management and treatment regimens vary with the country and the treatment center. Most experts use spiramycin to treat pregnant women who have acute toxoplasmosis early in pregnancy and use pyrimethamine/sulfadiazine/folinic acid to treat women who seroconvert after 18 weeks of pregnancy or in cases of documented fetal infection. This treatment is somewhat controversial: clinical studies, which have included few untreated women, have not proven the efficacy of such therapy in preventing congenital toxoplasmosis. However, studies do suggest that treatment during pregnancy decreases the severity of infection. Many women who are infected in the first trimester elect termination of pregnancy. Those who do not terminate pregnancy are offered prenatal antibiotic therapy to reduce the frequency and severity of *Toxoplasma* infection in the infant. The optimal duration of treatment for a child with asymptomatic congenital toxoplasmosis is not clear, although most clinicians in the United

Loading [Contrib]/a11y/accessibility-menu.js

States would treat the child for 1 year in light of cohort investigations conducted by the National Collaborative Chicago-Based Congenital Toxoplasmosis Study.

INFECTION IN IMMUNOCOMPETENT PATIENTS

Immunologically competent adults and older children who have only lymphadenopathy do not require specific therapy unless they have persistent, severe symptoms. Patients with ocular toxoplasmosis are usually treated for 1 month with pyrimethamine plus either sulfadiazine or clindamycin and sometimes with prednisone. Treatment should be supervised by an ophthalmologist familiar with *Toxoplasma* disease. Ocular disease can be self-limited without treatment, but therapy is typically considered for lesions that are severe or close to the fovea or optic disc. Prolonged treatment may prevent recurrences of ocular toxoplasmosis, but whether treatment improves long-term visual outcomes is unclear.

INFECTION IN IMMUNOCOMPROMISED PATIENTS Primary Prophylaxis

Patients with AIDS should be treated for acute toxoplasmosis; in immunocompromised patients, toxoplasmosis is rapidly fatal if untreated. Despite their toxicity, the drugs used to treat TE were required for survival prior to cART. The incidence of TE has declined as the survival of patients with HIV infection has increased through the use of cART.

Image not available. In Africa, many patients are diagnosed with HIV infection only after developing opportunistic infections. Hence, the optimal management of these opportunistic infections is important if the benefits of subsequent cART are to be realized. The incidence of TE in under-resourced settings is unknown because serologic testing and imaging are not available. AIDS patients who are seropositive for *T. gondii* and who have a CD4+ T lymphocyte count of <100/μL should receive prophylaxis against TE.

Of the currently available agents, trimethoprim-sulfamethoxazole (TMP-SMX) appears to be an effective alternative for treatment of TE in resource-poor settings where the preferred combination of pyrimethamine plus sulfadiazine is not available. The daily dose of TMP-SMX (one double-strength tablet) recommended for prophylaxis of PcP is effective against TE. If patients cannot tolerate TMP-SMX, the recommended alternative is dapsone-pyrimethamine, which likewise is effective against PcP. Atovaquone with or without pyrimethamine also can be considered. Prophylactic monotherapy with dapsone, pyrimethamine, azithromycin, clarithromycin, or aerosolized pentamidine is probably insufficient. AIDS patients who are seronegative for *Toxoplasma* and are not receiving prophylaxis for PcP should be retested for IgG antibody to *Toxoplasma* if their CD4+ T cell count drops to <100/μL. If seroconversion has taken place, then the patient should be given prophylaxis as described above.

Discontinuing Primary Prophylaxis

Current studies indicate that prophylaxis against TE can be discontinued in patients who have responded to cART and whose CD4+ T lymphocyte count has been >200/μL for 3 months. Although patients with CD4+ T lymphocyte counts of <100/μL are at greatest risk for developing TE, the risk that this condition will develop when the count has increased to 100–200/μL has not been established. Thus, prophylaxis should be discontinued when the count has increased to >200/μL. Discontinuation of therapy reduces the pill burden; the potential for drug toxicity, drug interaction, or selection of drug-resistant pathogens; and cost. Prophylaxis should be recommenced if the CD4+ T lymphocyte count again decreases to <100–200/μL.

Individuals who have completed initial therapy for TE should receive treatment indefinitely unless immune reconstitution, with a CD4+ T cell count of >200/μL, occurs as a consequence of cART. Combination therapy with pyrimethamine plus sulfadiazine plus leucovorin is effective for this purpose. A[illegible]gimen is clindamycin.

Loading [Contrib]/a11y/accessibility-menu.js

Discontinuing Secondary Prophylaxis (Long-Term Maintenance Therapy)

Patients receiving secondary prophylaxis for TE are at low risk for recurrence when they have completed initial therapy for TE, remain asymptomatic, and have evidence of restored immune function. Individuals with HIV infection should have a CD4+ T lymphocyte count of >200/μL for at least 6 months after cART. This recommendation is consistent with more extensive data indicating the safety of discontinuing secondary prophylaxis for other opportunistic infections during advanced HIV disease. A repeat MRI brain scan is recommended. Secondary prophylaxis should be reintroduced if the CD4+ T lymphocyte count decreases to <200/μL.

PREVENTION

All HIV-infected persons should be counseled regarding sources of *Toxoplasma* infection. The chances of primary infection with *Toxoplasma* can be reduced by not eating undercooked meat and by avoiding oocyst-contaminated material (i.e., a cat's litter box). Specifically, lamb, beef, and pork should be cooked to an internal temperature of 165–170°F (74–77°C); from a more practical perspective, meat cooked until it is no longer pink inside usually satisfies this requirement. Hands should be washed thoroughly after work in the garden, and all fruits and vegetables should be washed. Ingestion of raw shellfish is a risk factor for toxoplasmosis, given that the filter-feeding mechanism of clams and mussels concentrates oocysts.

If the patient owns a cat, the litter box should be cleaned or changed daily, preferably by an HIV-negative, nonpregnant person; alternatively, patients should wash their hands thoroughly after changing the litter box. Litter boxes should be changed daily if possible, as freshly excreted oocysts will not have sporulated and will not be infectious. Patients should be encouraged to keep their cats inside and not to adopt or handle stray cats. Cats should be fed only canned or dried commercial food or well-cooked table food, not raw or undercooked meats. Patients need not be advised to part with their cats or to have their cats tested for toxoplasmosis. Blood intended for transfusion into *Toxoplasma*-seronegative immunocompromised individuals should be screened for antibody to *T. gondii*. Although such serologic screening is not routinely performed, seronegative women should be screened for evidence of infection several times during pregnancy if they are exposed to environmental conditions that put them at risk for infection with *T. gondii*. HIV-positive individuals should adhere closely to these preventive measures.

ACKNOWLEDGMENT

The author would like to acknowledge Dr. Lloyd Kasper for his numerous contributions to our understanding of the pathogenesis of toxoplasmosis and his essential role in preparation of this chapter for prior editions.

FURTHER READING

Aleixo AL et al: Toxoplasmic retinochoroiditis: Clinical characteristics and visual outcome in a prospective study. PLoS Negl Trop Dis 10:e0004685, 2016. [PubMed: 27136081]
Peyron F et al: Congenital toxoplasmosis in France and the United States: One parasite, two diverging approaches. PLoS Negl Trop Dis 11:e0005222, 2017. [PubMed: 28207736]
Wang ZD et al: Prevalence and burden of *Toxoplasma gondii* infection in HIV-infected people: A systematic review and meta-analysis. Lancet HIV 4:e177, 2017. [PubMed: 28159548]

Loading [Contrib]/a11y/accessibility-menu.js

Chapter 224: Protozoal Intestinal Infections and Trichomoniasis

Peter F. Weller

PROTOZOAL INFECTIONS

GIARDIASIS

Image not available. *Giardia intestinalis* (also known as *G. lamblia* or *G. duodenalis*) is a cosmopolitan protozoal parasite that inhabits the small intestines of humans and other mammals. Giardiasis is one of the most common parasitic diseases in both developed and developing countries worldwide, causing both endemic and epidemic intestinal disease and diarrhea.

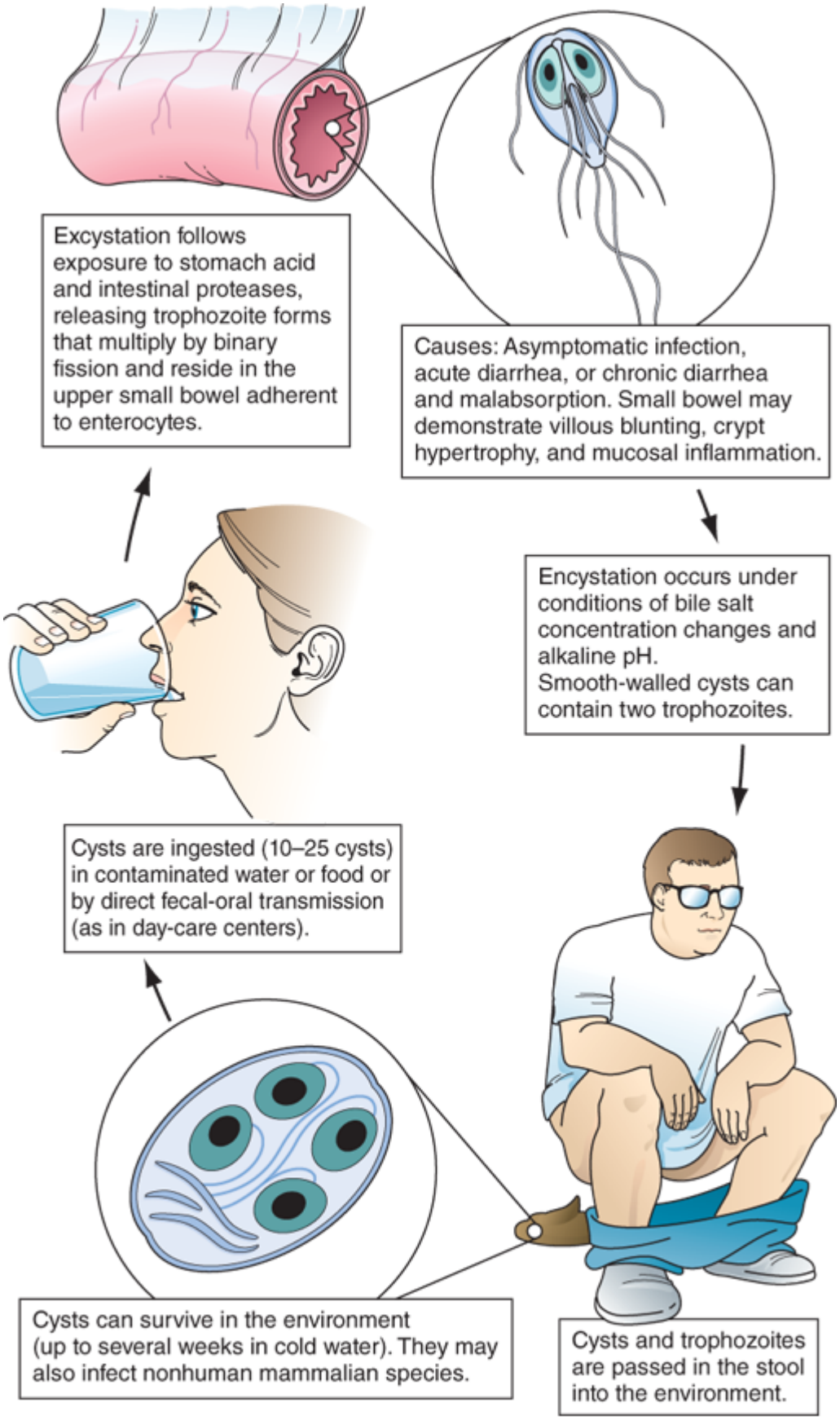
Life Cycle and Epidemiology

Infection follows the ingestion of environmentally hardy cysts, which excyst in the small intestine, releasing flagellated trophozoites **(Fig. 224-1)**, **(Fig. 224-2)** that multiply by binary fission. *Giardia* remains a pathogen of the proximal small bowel and does not disseminate hematogenously. Trophozoites remain free in the lumen or attach to the mucosal epithelium by means of a ventral sucking disk. As a trophozoite encounters altered conditions, it forms a morphologically distinct cyst, which is the stage of the parasite usually found in the feces. Trophozoites may be present and even predominate in loose or watery stools, but it is the resistant cyst that survives outside the body and is responsible for transmission. Cysts do not tolerate heating or desiccation, but they do remain viable for months in cold fresh water. The number of cysts excreted varies widely but can approach 10^7 per gram of stool.

FIGURE 224-1

Life cycle of *Giardia.* *(Reprinted with permission from RL Guerrant et al [eds]: Tropical Infectious Diseases: Principles, Pathogens and Practice, 2nd ed, p 987. © 2006, with permission from Elsevier Science.)*

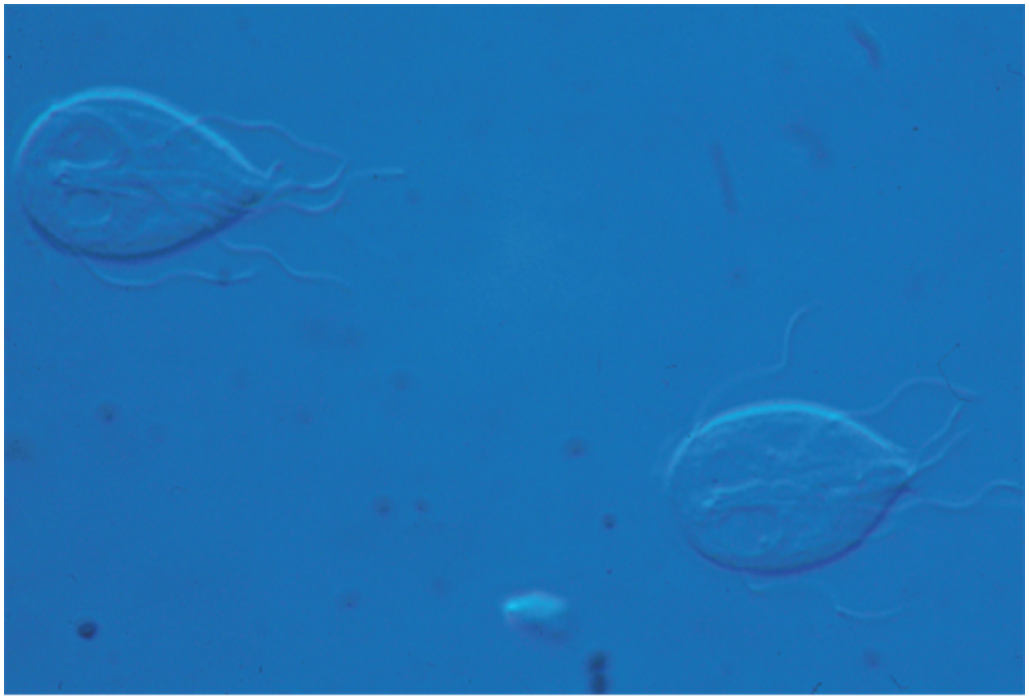
Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Loading [Contrib]/a11y/accessibility-menu.js

Flagellated, binucleate *Giardia* trophozoites.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ingestion of as few as 10 cysts is sufficient to cause infection in humans. Because cysts are infectious when excreted, person-to-person transmission occurs where fecal hygiene is poor. Giardiasis is especially prevalent in day-care centers; person-to-person spread also takes place in other institutional settings with poor fecal hygiene and during anal–oral contact. If food is contaminated with *Giardia* cysts after cooking or preparation, food-borne transmission can occur. Waterborne transmission accounts for episodic infections (e.g., in campers and travelers) and for major epidemics in metropolitan areas. Surface water, ranging from mountain streams to large municipal reservoirs, can become contaminated with fecally derived *Giardia* cysts. The efficacy of water as a means of transmission is enhanced by the small infectious inoculum of *Giardia*, the prolonged survival of cysts in cold water, and the resistance of cysts to killing by routine chlorination methods that are adequate for controlling bacteria. Viable cysts can be eradicated from water by either boiling or filtration.

Image not available. In the United States, *Giardia* (like *Cryptosporidium*; see below) is a common cause of waterborne epidemics of gastroenteritis. *Giardia* is common in developing countries, and infections may be acquired by travelers.

There are several recognized genotypes or assemblages of *G. intestinalis*. Human infections are due to assemblages A and B, whereas other assemblages are more common in other animals, including cats and dogs. Like beavers from reservoirs implicated in epidemics, dogs and cats have been found to be infected with assemblages A and B; this finding suggests both that these animals may have been infected from human sources and that they might be sources of further human infections.

Giardiasis, like cryptosporidiosis, creates a significant economic burden because of the costs incurred in the installation of water filtration systems required to prevent waterborne epidemics, in the management of epidemics that involve large communities, and in the evaluation and treatment of endemic infections.

Loading [Contrib]/a11y/accessibility-menu.js

The reasons that some, but not all, infected patients develop clinical manifestations and the mechanisms by which *Giardia* causes alterations in small-bowel function are largely unknown. Although trophozoites adhere to the epithelium, they are not invasive but may elicit apoptosis of enterocytes, epithelial barrier dysfunction, and epithelial cell malabsorption and secretion. Consequent lactose intolerance and, in a minority of infected adults and children, significant malabsorption are clinical signs of the loss of brush-border enzyme activities. In most infections, the morphology of the bowel is unaltered; however, in chronically infected, symptomatic patients, the histopathologic findings (including flattened villi) and the clinical manifestations at times resemble those of tropical sprue and gluten-sensitive enteropathy. The pathogenesis of diarrhea in giardiasis is not known.

The natural history of *Giardia* infection varies markedly. Infections may be aborted, transient, recurrent, or chronic. *G. intestinalis* parasites vary genotypically, and such variations might contribute to different courses of infection. Parasite as well as host factors may be important in determining the course of infection and disease. Both cellular and humoral responses develop in human infections, but their precise roles in disease pathogenesis and/or control of infection are unknown. Because patients with hypogammaglobulinemia suffer from prolonged, severe infections that are poorly responsive to treatment, humoral immune responses appear to be important. The greater susceptibilities of the young than of the old and of newly exposed persons than of chronically exposed populations suggest that at least partial protective immunity may develop.

Clinical Manifestations

Disease manifestations of giardiasis range from asymptomatic carriage to fulminant diarrhea and malabsorption. Most infected persons are asymptomatic, but in epidemics the proportion of symptomatic cases may be higher. Symptoms may develop suddenly or gradually. In persons with acute giardiasis, symptoms develop after an incubation period that lasts at least 5–6 days and usually 1–3 weeks. Prominent early symptoms include diarrhea, abdominal pain, bloating, belching, flatus, nausea, and vomiting. Although diarrhea is common, upper intestinal manifestations such as nausea, vomiting, bloating, and abdominal pain may predominate. The duration of acute giardiasis is usually >1 week, although diarrhea often subsides. Individuals with chronic giardiasis may present with or without having experienced an antecedent acute symptomatic episode. Diarrhea is not necessarily prominent, but increased flatus, loose stools, sulfurous belching, and (in some instances) weight loss occur. Symptoms may be continual or episodic and may persist for years. Some persons who have relatively mild symptoms for long periods recognize the extent of their discomfort only in retrospect. Fever, the presence of blood and/or mucus in the stools, and other signs and symptoms of colitis are uncommon and suggest a different diagnosis or a concomitant illness. Symptoms tend to be intermittent yet recurring and gradually debilitating, in contrast with the acute disabling symptoms associated with many enteric bacterial infections. Because of the less severe illness early on and the propensity for chronic infections, patients may seek medical advice late in the course of the illness; however, disease can be severe, resulting in malabsorption, weight loss, growth retardation, and dehydration. A number of extraintestinal manifestations have been described, such as urticaria, anterior uveitis, and arthritis; whether these are caused by giardiasis or concomitant processes is unclear.

Giardiasis can be severe in patients with hypogammaglobulinemia and can complicate other preexisting intestinal diseases, such as that occurring in cystic fibrosis. In patients with AIDS, *Giardia* can cause enteric illness that is refractory to treatment.

Diagnosis

Giardiasis is diagnosed by detection of parasite antigens in the feces, by identification of cysts in the feces or of trophozoites in the feces or small intestines, or by nucleic acid amplification tests (NAATs) (Loading [Contrib]/a11y/accessibility-menu.js e 8–12 μm × 7–10 μm, and characteristically contain four nuclei. Trophozoites are pear-shaped, dorsally convex, flattened parasites with two nuclei and four pairs of

flagella (Fig. 224-2). The diagnosis is sometimes difficult to establish. Direct examination of fresh or properly preserved stools as well as concentration methods should be used. Because cyst excretion is variable and may be undetectable at times, repeated examination of stool, sampling of duodenal fluid, and biopsy of the small intestine may be required to detect the parasite. Tests for parasitic antigens in stool are at least as sensitive and specific as good microscopic examinations and are easier to perform. Newer NAATs are highly sensitive but are not always available for clinical use at present.

TABLE 224-1 Diagnosis of Intestinal Protozoal Infections

Parasite	Stool O+P[a]	Fecal Acid-Fast Stain	Fecal Antigen Immunoassays	Fecal NAATs[b]	Other
Giardia	+		+	+	
Cryptosporidium	–	+	+	+	
Isospora	–	+		+	
Cyclospora	–	+		+	
Microsporidia	–			+	Special fecal stains, tissue biopsies

[a]O+P, ova and parasites. [b]Nucleic acid amplification tests.

TREATMENT

TREATMENT Giardiasis

Cure rates with metronidazole (250 mg thrice daily for 5 days) are usually >90%. Tinidazole (2 g once by mouth) may be more effective than metronidazole. Nitazoxanide (500 mg twice daily for 3 days) is an alternative agent for treatment of giardiasis. Paromomycin, an oral aminoglycoside that is not well absorbed, can be given to symptomatic pregnant patients, although information is limited on how effectively this agent eradicates infection.

Almost all patients respond to therapy and are cured, although some with chronic giardiasis experience delayed resolution of symptoms after eradication of *Giardia*. For many of the latter patients, residual symptoms probably reflect delayed regeneration of intestinal brush-border enzymes. Continued infection should be documented by stool examinations before treatment is repeated. Patients who remain infected after repeated treatments should be evaluated for reinfection through family members, close personal contacts, and environmental sources as well as for hypogammaglobulinemia. In cases refractory to multiple treatment courses, prolonged therapy with metronidazole (750 mg thrice daily for 21 days) or therapy with varied combinations of multiple agents has been successful.

Prevention

Giardiasis can be prevented by consumption of uncontaminated food and water and by personal hygiene during the provision of care for infected children. Boiling or filtering potentially contaminated water prevents infection.

CRYPTOSPORIDIOSIS

Image not available. The coccidian parasite *Cryptosporidium* causes diarrheal disease that is self-limited in immunocompetent human hosts but can be severe in persons with AIDS or other forms of immunodeficiency. Two species of *Cryptosporidium*, *C. hominis* and *C. parvum*, cause most human i

Loading [Contrib]/a11y/accessibility-menu.js

Life Cycle and Epidemiology

Cryptosporidium species are widely distributed in the world. Cryptosporidiosis is acquired by the consumption of oocysts (50% infectious dose: ~132 *C. parvum* oocysts in nonimmune individuals), which excyst to liberate sporozoites that in turn enter and infect intestinal epithelial cells. The parasite's further development involves both asexual and sexual cycles, which produce forms capable of infecting other epithelial cells and of generating oocysts that are passed in the feces. *Cryptosporidium* species infect a number of animals, and *C. parvum* can spread from infected animals to humans. Since oocysts are immediately infectious when passed in feces, person-to-person transmission takes place in day-care centers and among household contacts and medical providers. Waterborne transmission (especially that of *C. hominis*) accounts for infections in travelers and for common-source epidemics. Oocysts are quite hardy and resist killing by routine chlorination. Both drinking water and recreational water (e.g., pools, waterslides) have been increasingly recognized as sources of infection.

Pathophysiology

Although intestinal epithelial cells harbor cryptosporidia in an intracellular vacuole, the means by which secretory diarrhea is elicited remain uncertain. No characteristic pathologic changes are found by biopsy. The distribution of infection can be spotty within the principal site of infection, the small bowel. Cryptosporidia are found in the pharynx, stomach, and large bowel of some patients and at times in the respiratory tract. Especially in patients with AIDS, involvement of the biliary tract can cause papillary stenosis, sclerosing cholangitis, or cholecystitis.

Clinical Manifestations

Asymptomatic infections can occur in both immunocompetent and immunocompromised hosts. In immunocompetent persons, symptoms develop after an incubation period of ~1 week and consist principally of watery nonbloody diarrhea, sometimes in conjunction with abdominal pain, nausea, anorexia, fever, and/or weight loss. In these hosts, the illness usually subsides after 1–2 weeks. In contrast, in immunocompromised hosts (especially those with AIDS and CD4+ T cell counts <100/μL), diarrhea can be chronic, persistent, and remarkably profuse, causing clinically significant fluid and electrolyte depletion. Stool volumes may range from 1 to 25 L/d. Weight loss, wasting, and abdominal pain may be severe. Biliary tract involvement can manifest as mid-epigastric or right-upper-quadrant pain.

Diagnosis

Evaluation starts with fecal examination for small oocysts, which are smaller (4–5 μm in diameter) than the fecal stages of most other parasites (Table 224-1). Because conventional stool examination for ova and parasites (O+P) does not detect *Cryptosporidium*, specific testing must be requested. Detection is enhanced by evaluation of stools (obtained on multiple days) by several techniques, including modified acid-fast and direct immunofluorescent stains and enzyme immunoassays. Newer NAATs are being employed. Cryptosporidia can also be identified by light and electron microscopy at the apical surfaces of intestinal epithelium from biopsy specimens of the small bowel and, less frequently, the large bowel.

TREATMENT

TREATMENT Cryptosporidiosis

Loading [Contrib]/a11y/accessibility-menu.js

od and Drug Administration (FDA) for the treatment of cryptosporidiosis, is available in tablet form for adults (500 mg twice daily for 3 days) and as an elixir

for children. This agent has not been effective for the treatment of HIV-infected patients, in whom improved immune status due to antiretroviral therapy can lead to amelioration of cryptosporidiosis. Otherwise, treatment includes supportive care with replacement of fluids and electrolytes and administration of antidiarrheal agents. Biliary tract obstruction may require papillotomy or T-tube placement. Prevention requires minimizing exposure to infectious oocysts in human or animal feces. Use of submicron water filters may minimize acquisition of infection from drinking water.

CYSTOISOSPORIASIS

The coccidian parasite *Cystoisospora belli* causes human intestinal disease. Infection is acquired by the consumption of oocysts, after which the parasite invades intestinal epithelial cells and undergoes both sexual and asexual cycles of development. Oocysts excreted in stool are not immediately infectious but must undergo further maturation.

Although *C. belli* infects many animals, little is known about the epidemiology or prevalence of this parasite in humans. It is most common in tropical and subtropical countries. Acute infections can begin abruptly with fever, abdominal pain, and watery nonbloody diarrhea and can last for weeks or months. In patients who have AIDS or are immunocompromised for other reasons, infections often are not self-limited but rather resemble cryptosporidiosis, with chronic, profuse watery diarrhea. Eosinophilia, which is not found in other enteric protozoan infections, may be detectable. The diagnosis (Table 224-1) is usually made by detection of the large (~25 μm) oocysts in stool by modified acid-fast staining. Oocyst excretion may be low-level and intermittent; if repeated stool examinations are unrevealing, sampling of duodenal contents by aspiration or small-bowel biopsy (often with electron microscopic examination) may be necessary. NAATs are promising newer diagnostic tools.

TREATMENT

TREATMENT Cystoisosporiasis

Trimethoprim-sulfamethoxazole (TMP-SMX, 160/800 mg two times daily for 10 days; and, for HIV-infected patients, then continuing three times daily for 3 weeks) is effective. For patients intolerant of sulfonamides, pyrimethamine (50–75 mg/d) can be used. Relapses can occur in persons with AIDS and necessitate maintenance therapy with TMP-SMX (160/800 mg three times per week).

CYCLOSPORIASIS

Image not available. *Cyclospora cayetanensis*, a cause of diarrheal illness, is globally distributed: illness due to *C. cayetanensis* has been reported in the United States, Asia, Africa, Latin America, and Europe. The epidemiology of this parasite has not yet been fully defined, but waterborne transmission and food-borne transmission (e.g., by basil, sweet peas, and imported raspberries) have been recognized. The full spectrum of illness attributable to *Cyclospora* has not been delineated. Some infected patients may be without symptoms, but many have diarrhea, flulike symptoms, and flatulence and belching. The illness can be self-limited, can wax and wane, or, in many cases, can involve prolonged diarrhea, anorexia, and upper gastrointestinal symptoms, with sustained fatigue and weight loss in some instances. Diarrheal illness may persist for >1 month. *Cyclospora* can cause enteric illness in patients infected with HIV.

The parasite is detectable in epithelial cells of small-bowel biopsy samples and elicits secretory diarrhea by unknown means. The absence of fecal blood and leukocytes indicates that disease due to *Cyclospora* is not caused by destruction of the small-bowel mucosa. The diagnosis (Table 224-1) can be made by detection of spherical 8- to 10-μm oocysts in the stool, although routine stool O+P examinations are not sufficient. Specific fecal examinations must be requested to detect the oocysts,

Loading [Contrib]/a11y/accessibility-menu.js

which are variably acid-fast and are fluorescent when viewed with ultraviolet light microscopy. Newer NAATs are proving to be sensitive. Cyclosporiasis should be considered in the differential diagnosis of prolonged diarrhea, with or without a history of travel by the patient to other countries.

TREATMENT

TREATMENT Cyclosporiasis

Cyclosporiasis is treated with TMP-SMX (160/800 mg twice daily for 7–10 days). HIV-infected patients may experience relapses after such treatment and thus may require longer-term suppressive maintenance therapy.

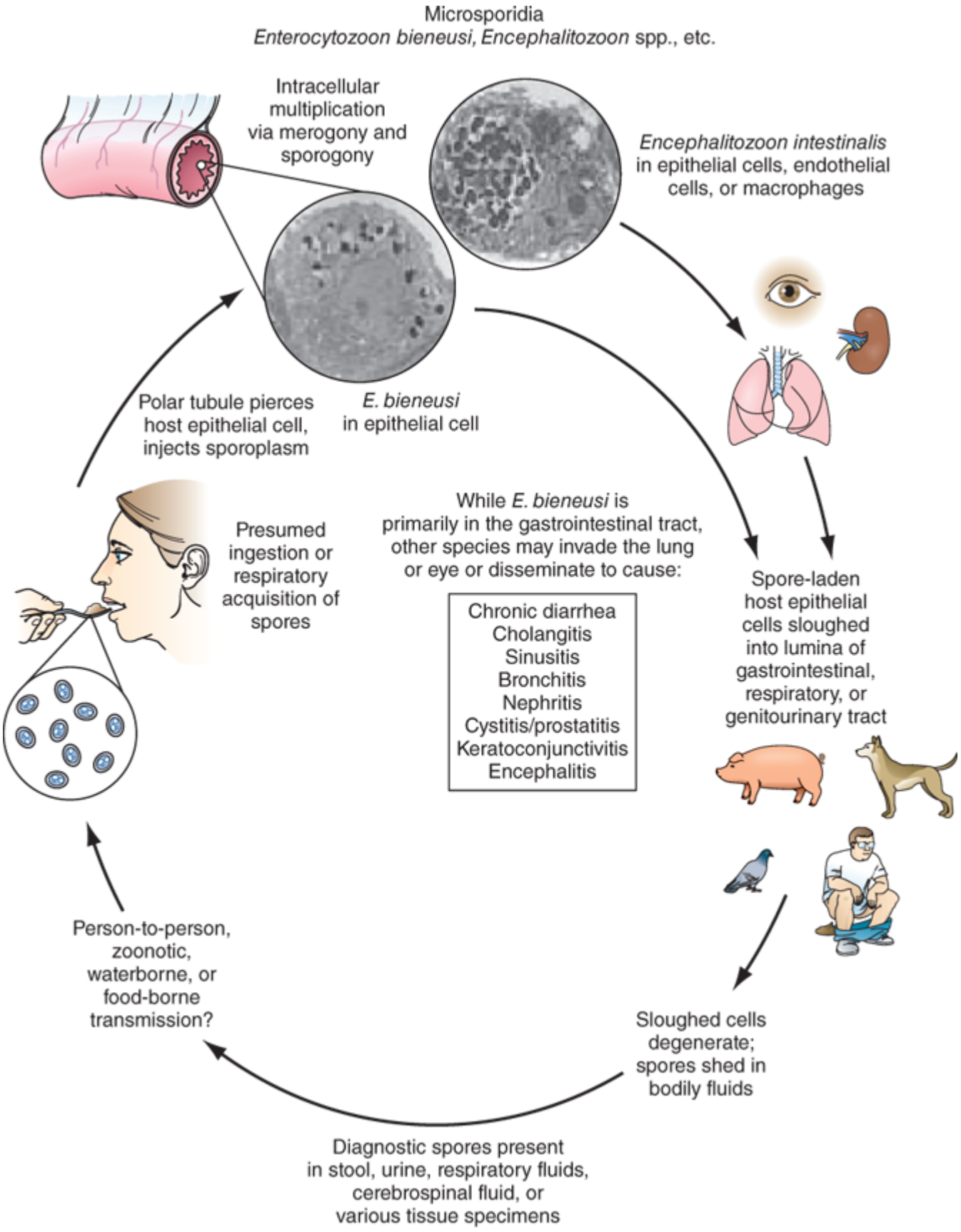
MICROSPORIDIOSIS

Microsporidia are obligate intracellular spore-forming protozoa that infect many animals and cause disease in humans, especially as opportunistic pathogens in AIDS. Microsporidia are members of a distinct phylum, Microspora, which contains dozens of genera and hundreds of species. The various microsporidia are differentiated by their developmental life cycles, ultrastructural features, and molecular taxonomy based on ribosomal RNA. The complex life cycles of the organisms result in the production of infectious spores **(Fig. 224-3)**. Currently, eight genera of microsporidia—*Encephalitozoon*, *Pleistophora*, *Nosema*, *Vittaforma*, *Trachipleistophora*, *Anncalia*, *Microsporidium*, and *Enterocytozoon*—are recognized as causes of human disease. Although some microsporidia are probably prevalent causes of self-limited or asymptomatic infections in immunocompetent patients, little is known about how microsporidiosis is acquired.

FIGURE 224-3

Life cycle of microsporidia. *(Reprinted with permission from RL Guerrant et al [eds]: Tropical Infectious Diseases: Principles, Pathogens and Practice, 2nd ed, p 1128. © 2006, with permission from Elsevier Science.)*

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Microsporidiosis is most common among patients with AIDS, less common among patients with other types of immunocompromise, and rare among immunocompetent hosts. In patients with AIDS,

Loading [Contrib]/a11y/accessibility-menu.js

intestinal infections with *Enterocytozoon bieneusi* and *Encephalitozoon* (formerly *Septata*) *intestinalis* are recognized to contribute to chronic diarrhea and wasting; these infections have been found in 10–40% of patients with chronic diarrhea. Both organisms have been found in the biliary tracts of patients with cholecystitis. *E. intestinalis* may also disseminate to cause fever, diarrhea, sinusitis, cholangitis, and bronchiolitis. In patients with AIDS, *Encephalitozoon hellem* has caused superficial keratoconjunctivitis as well as sinusitis, respiratory tract disease, and disseminated infection. Myositis due to *Pleistophora* has been documented. *Nosema*, *Vittaforma*, and *Microsporidium* have caused stromal keratitis associated with trauma in immunocompetent patients.

Microsporidia are small gram-positive organisms with mature spores measuring 0.5–2 μm × 1–4 μm. Diagnosis of microsporidial infections in tissue often requires electron microscopy, although intracellular spores can be visualized by light microscopy with hematoxylin and eosin, Giemsa, or tissue Gram's stain. For the diagnosis of intestinal microsporidiosis, modified trichrome or chromotrope 2R–based staining and Uvitex 2B or calcofluor fluorescent staining reveal spores in smears of feces or duodenal aspirates. Definitive therapies for microsporidial infections remain to be established. For superficial keratoconjunctivitis due to *E. hellem*, topical therapy with fumagillin suspension has shown promise **(Chap. 217)**. For enteric infections with *E. bieneusi* and *E. intestinalis* in HIV-infected patients, therapy with albendazole may be efficacious **(Chap. 217)**.

OTHER INTESTINAL PROTOZOA

Balantidiasis

Image not available. *Balantidium coli* is a large ciliated protozoal parasite that can produce a spectrum of large-intestinal disease analogous to amebiasis. The parasite is widely distributed in the world. Since it infects pigs, cases in humans are more common where pigs are raised. Infective cysts can be transmitted from person to person and through water, but many cases are due to the ingestion of cysts derived from porcine feces in association with slaughtering, with use of pig feces for fertilizer, or with contamination of water supplies by pig feces.

Ingested cysts liberate trophozoites, which reside and replicate in the large bowel. Many patients remain asymptomatic, but some have persisting intermittent diarrhea, and a few develop more fulminant dysentery. In symptomatic individuals, the pathology in the bowel—both gross and microscopic—is similar to that seen in amebiasis, with varying degrees of mucosal invasion, focal necrosis, and ulceration. Balantidiasis, unlike amebiasis, only rarely spreads hematogenously to other organs. The diagnosis is made by detection of the trophozoite stage in stool or sampled colonic tissue. Tetracycline (500 mg four times daily for 10 days) is an effective therapeutic agent.

Blastocystosis

Blastocystis hominis remains an organism of uncertain pathogenicity. Some patients who pass *B. hominis* in their stools are asymptomatic, whereas others have diarrhea and associated intestinal symptoms. Diligent evaluation reveals other potential bacterial, viral, or protozoal causes of diarrhea in some but not all patients with symptoms. Because the pathogenicity of *B. hominis* is uncertain and because therapy for *Blastocystis* infection is neither specific nor uniformly effective, patients with prominent intestinal symptoms should be fully evaluated for other infectious causes of diarrhea. If diarrheal symptoms associated with *Blastocystis* are prominent, either metronidazole (750 mg thrice daily for 10 days) or TMP-SMX (160 mg/800 mg twice daily for 7 days) can be used.

Dientamoebiasis

Dientamoeba fragilis is unique among intestinal protozoa in that it has a trophozoite stage but not a cyst stage. How trophozoites survive to transmit infection is not known. When symptoms develop in

Loading [Contrib]/a11y/accessibility-menu.js

patients with *D. fragilis* infection, they are generally mild and include intermittent diarrhea, abdominal pain, and anorexia. The diagnosis is made by the detection of trophozoites in stool; the lability of these forms accounts for the greater yield when fecal samples are preserved immediately after collection. Since fecal excretion rates vary, examination of several samples obtained on alternate days increases the rate of detection. Iodoquinol (650 mg three times daily for 20 days) or paromomycin (25–35 mg/kg per day in three doses for 7 days) is appropriate for treatment.

TRICHOMONIASIS

Various species of trichomonads can be found in the mouth (in association with periodontitis) and occasionally in the gastrointestinal tract. *Trichomonas vaginalis*—one of the most prevalent protozoal parasites in the United States—is a pathogen of the genitourinary tract and a major cause of symptomatic vaginitis **(Chap. 131)**.

Life Cycle and Epidemiology

T. vaginalis is a pear-shaped, actively motile organism that measures about 10 × 7 μm, replicates by binary fission, and inhabits the lower genital tract of females and the urethra and prostate of males. In the United States, it accounts for ~3 million infections per year in women. While the organism can survive for a few hours in moist environments and could be acquired by direct contact, person-to-person venereal transmission accounts for virtually all cases of trichomoniasis. Its prevalence is greatest among persons with multiple sexual partners and among those with other sexually transmitted diseases **(Chap. 131)**.

Clinical Manifestations

Many men infected with *T. vaginalis* are asymptomatic, although some develop urethritis and a few have epididymitis or prostatitis. In contrast, infection in women, which has an incubation period of 5–28 days, is usually symptomatic and manifests with malodorous vaginal discharge (often yellow), vulvar erythema and itching, dysuria or urinary frequency (in 30–50% of patients), and dyspareunia. These manifestations, however, do not clearly distinguish trichomoniasis from other types of infectious vaginitis.

Diagnosis

Detection of motile trichomonads by microscopic examination of wet mounts of vaginal or prostatic secretions has been the conventional means of diagnosis. Although this approach provides an immediate diagnosis, its sensitivity for the detection of *T. vaginalis* is only ~50–60% in routine evaluations of vaginal secretions. Direct immunofluorescent antibody staining is more sensitive (70–90%) than wet-mount examinations. *T. vaginalis* can be recovered from the urethra of both males and females and is detectable in males after prostatic massage. NAATs are FDA approved and are highly sensitive and specific for urine and for endocervical and vaginal swabs from women.

TREATMENT

TREATMENT Trichomoniasis

Metronidazole (either a single 2-g dose or 500-mg doses twice daily for 7 days) or tinidazole (a single 2-g dose) is effective. All sexual partners must be treated concurrently to prevent reinfection, especially from asymptomatic males. In males with persistent symptomatic urethritis after therapy for nongonococcal urethritis, metronidazole therapy should be considered for possible trichomoniasis. Alternatives to metronidazole for treatment during pregnancy are not readily available. Reinfection

Loading [Contrib]/a11y/accessibility-menu.js

often accounts for apparent treatment failures, but strains of *T. vaginalis* exhibiting high-level resistance to metronidazole have been encountered. Treatment of these resistant infections with higher oral doses, parenteral doses, or concurrent oral and vaginal doses of metronidazole or with tinidazole has been successful.

FURTHER READING

Adam EA et al: Giardiasis outbreaks in the United States, 1971–2011. Epidemiol Infect 144:2790, 2016. [PubMed: 26750152]
Cacciò SM, Chalmers RM: Human cryptosporidiosis in Europe. Clin Microbiol Infect 22:471, 2016. [PubMed: 27172805]
Cama VA, Mathison BA: Infections by intestinal coccidia and *Giardia duodenalis.* Clin Lab Med 35:423, 2015. [PubMed: 26004650]
Garcia LS: *Dientamoeba fragilis*, one of the neglected intestinal protozoa. J Clin Microbiol 54:2243, 2016. [PubMed: 27053676]
Heyworth MF: *Giardia duodenalis* genetic assemblages and hosts. Parasite 23:13, 2016. [PubMed: 26984116]
Hlavsa MC et al: Using molecular characterization to support investigations of aquatic facility–associated outbreaks of cryptosporidiosis—Alabama, Arizona, and Ohio, 2016. MMWR 66:493, 2017. [PubMed: 28520707]
McHardy IH et al: Detection of intestinal protozoa in the clinical laboratory. J Clin Microbiol 52:712, 2014. [PubMed: 24197877]
Ramanan P, Pritt BS: Extraintestinal microsporidiosis. J Clin Microbiol 52:3839, 2014. [PubMed: 24829239]
Ryan U et al: New technologies for detection of enteric parasites. Trends Parasitol 33:532, 2017.
Van Der Pol B: Clinical and laboratory testing for *Trichomonas vaginalis* infection. J Clin Microbiol 54:7, 2016. [PubMed: 26491181]

Chapter 225: Introduction to Helminthic Infections

Peter F. Weller

INTRODUCTION

The word *helminth* is derived from the Greek *helmins* (“parasitic worm”). Helminthic worms are highly prevalent and, depending on the species, may exist as free-living organisms or as parasites of plant or animal hosts. The parasitic helminths have co-evolved with specific mammalian and other host species. Accordingly, most helminthic infections are restricted to nonhuman hosts, and only rarely do these zoonotic helminths accidentally cause human infections.

Helminthic parasites of humans belong to two phyla: Nemathelminthes, which includes nematodes (roundworms), and Platyhelminthes, which includes cestodes (tapeworms) and trematodes (flukes). Helminthic parasites of humans reside within the human body and hence are the cause of true infections. In contrast, parasites of other genera that reside only on mucocutaneous surfaces of humans (e.g., the parasites causing myiasis and scabies) are considered to represent infestations rather than infections.

Helminthic parasites differ substantially from protozoan parasites in several respects. First, protozoan parasites are unicellular organisms, whereas helminthic parasites are multicellular worms that possess differentiated organ systems. Second, helminthic parasites have complex life cycles that require sequential stages of development outside the human host. Thus, most helminths do not complete their replication within the human host; rather, they develop to a certain stage within the mammalian host and, as part of their obligatory life cycle, must mature further outside that host. During the “extra-human” stages of their life cycle, helminths exist either as free-living organisms or as parasites within another host species and thereafter mature into new developmental stages capable of infecting humans. Thus, with only two exceptions (*Strongyloides stercoralis* and *Capillaria philippinensis*, which are capable of internal human reinfections), increases in the number of adult helminths (i.e., the “worm burden”) within the human host require repeated exogenous reinfections. In the case of protozoan parasites, a brief, even singular exposure (e.g., a single mosquito bite transmitting malaria) may lead rapidly to intense parasite loads and overwhelming infections; in contrast, for all but the two helminths noted above, increases in worm burden require multiple and usually ongoing exposures to infectious forms, such as ingestion of eggs of intestinal helminths or waterborne exposures to infectious cercariae of *Schistosoma mansoni*. This requirement is germane both to the consideration of helminthic infections in individuals and to ongoing global efforts to interrupt and/or minimize the acquisition of helminthic infections by humans.

Third, helminthic infections have a predilection toward stimulation of host immune responses that elicit eosinophilia within human tissues and blood. The many protozoan infections characteristically do not elicit eosinophilia in infected humans, with only three exceptions (two intestinal protozoan parasites, *Cystoisospora belli* and *Dientamoeba fragilis*, and tissue-borne *Sarcocystis* species). The magnitude of helminth-elicited eosinophilia tends to correlate with the extent of tissue invasion by larvae or adult h[Loading [Contrib]/a11y/accessibility-menu.js]minthic infections, including acute schistosomiasis (Katayama syndrome), paragonimiasis, and hookworm and *Ascaris* infections, eosinophilia is most pronounced

during the early phases of infection, when migrations of infecting larvae and progression of subsequent developmental stages through the tissues are greatest. In established infections, local eosinophilia is often present around helminths in tissues, but blood eosinophilia may be intermittent, mild, or absent. In helminthic infections in which parasites are well contained within tissues (e.g., echinococcal cysts) or confined within the lumen of the intestinal tract (e.g., adult *Ascaris* or tapeworms), eosinophilia is usually absent.

NEMATODES

Nematodes are nonsegmented roundworms. Species of nematodes are remarkably diverse and abundant in nature. Among the many thousands of nematode species, few are parasites of humans. Most nematodes are free-living, and these species have variably evolved to survive in diverse ecologic niches, including saltwater, freshwater, or soil. The well-studied organism *Caenorhabditis elegans* is a free-living nematode. Nematodes can be either beneficial or deleterious parasites of plants. Parasitic nematodes have co-evolved with specific mammalian hosts and have no capacity to live their full life cycles in other hosts. Uncommonly, humans are exposed to infectious stages of nonhuman nematode parasites, and the resultant zoonotic nematode infections can elicit inflammatory and immune responses as larval forms migrate and die in the unsuitable human host. Examples include pulmonary coin lesions due to mosquito-transmitted infections with the dog heartworm *Dirofilaria immitis*; eosinophilic meningoencephalitis due to ingested eggs of the raccoon ascarid *Baylisascaris procyonis*; and eosinophilic meningitis due to ingestion of larvae of the rat lungworm *Angiostrongylus cantonensis*.

Nematode parasites of humans include worms that reside in the intestinal tract or localize in extraintestinal vascular or tissue sites. Roundworms are bisexual, with separate male and female forms (except for *S. stercoralis*, whose adult females are hermaphroditic in the human intestinal tract). Depending on the species, fertilized females release either larvae or eggs containing larvae. Nematodes have five developmental stages: an adult stage and four sequential larval stages. These parasites characteristically are surrounded by a durable outer cuticular layer. Nematodes have a nervous system; a muscular system, including muscle cells under the cuticle; and a developed intestinal tract, including an oral cavity and an elongated gut that ends in an anal pore. Adults may range in size from minute to >1 meter in length (with *Dracunculus medinensis*, for example, at the long end of this spectrum).

Humans acquire infections with nematode parasites by various routes, depending on the parasitic species. Ingestion of eggs passed in human feces is a major global health problem with many of the intestinal helminths (e.g., *Ascaris lumbricoides).* In other species, infecting larvae penetrate skin exposed to fecally contaminated soil (e.g., *S. stercoralis*, hookworms) or traverse the skin after the bite of infected insect vectors (e.g., filariae). Some nematode infections are acquired by consumption of specific animal-derived foods (e.g., trichinellosis from raw or undercooked pork or wild carnivorous mammals). As noted above, only two nematodes, *S. stercoralis* and *C. philippinensis*, can internally reinfect humans; thus, for all other nematodes, any increases in worm burden must be due to continued exogenous reinfections.

CESTODES

Tapeworms are the cestode parasites of humans. Adult tapeworms are elongated, segmented, hermaphroditic flatworms that reside in the intestinal lumen or, in their larval forms, may live in extraintestinal tissues. Tapeworms include a head (*scolex*) and a number of attached segments (*proglottids*). The worms attach to the intestinal tract via their scolices, which may possess suckers, hooks, or grooves. The scolex is the site of formation of new proglottids. Tapeworms do not have a fLoading [Contrib]/a11y/accessibility-menu.jsorm segment passively and actively obtains nutrients through its specialized surface tegument. Mature proglottids possess both male and female sex organs, but

insemination usually occurs between adjacent proglottids. Fertilized proglottids release eggs that are passed in the feces. When ingested by an intermediate host, an egg releases an oncosphere that penetrates the gut and develops further in tissues as a cysticercus. Humans acquire infection by ingesting animal tissues that contain cysticerci, and the resultant tapeworms develop and reside in the proximal small bowel (e.g., *Taenia solium*, *T. saginata*). Alternatively, if humans ingest eggs of these cestodes that have been passed in human or animal feces, oncospheres develop and can cause space-occupying extraintestinal cystic lesions in tissues; examples include cysticercosis due to *T. solium* and hydatid disease due to species of *Echinococcus*.

TREMATODES

Trematodes of medical importance include blood flukes, intestinal flukes, and tissue flukes. Adult flukes are often leaf-shaped flatworms. Oral and/or ventral suckers help adult flukes maintain their positions in situ. Flukes have an oral cavity but no distal anal pore. Nutrients are obtained both through their integument and by ingestion into the blind intestinal tract. Flukes are hermaphroditic except for blood flukes (schistosomes), which are bisexual. Eggs are passed in human feces (*Fasciola*, *Fasciolopsis*, *Clonorchis*, *Schistosoma japonicum*, *S. mansoni*), urine (*Schistosoma haematobium*), or sputum and feces (*Paragonimus*). Expelled eggs release miracidia—usually in water—that infect specific snail species. Within snails, parasites multiply and cercariae are released. Depending on the species, cercariae can penetrate the skin (schistosomes) or can develop into metacercariae that can be ingested with plants (e.g., watercress for *Fasciola*) or with fish (*Clonorchis*) or crabs (*Paragonimus*).

CONCLUSION

Image not available. Many of the so-called neglected tropical diseases are due to helminthic infections. The health impacts of many helminthic infections are varied and are based on the frequent need for repeated exposures to increase the worm burdens in infected humans. In global regions where exposures to specific helminths occur even in childhood (e.g., fecally derived intestinal nematodes, mosquito-transmitted filariae, or waterborne snail-transmitted schistosomes), the morbidities in infected individuals can include nutritional, developmental, cognitive, and functional impairments. Ongoing global mass-treatment programs are currently aimed at diminishing the local prevalences of specific helminths and their consequent impacts on the health of local populations.

Loading [Contrib]/a11y/accessibility-menu.js

Chapter 226: Trichinellosis and Other Tissue Nematode Infections

Peter F. Weller

INTRODUCTION

Nematodes are elongated, symmetric roundworms. Parasitic nematodes of medical significance may be broadly classified as either predominantly intestinal or tissue nematodes. The intestinal nematodes are covered in **Chap. 227**. This chapter covers the tissue nematodes that cause trichinellosis, visceral and ocular larva migrans, cutaneous larva migrans, cerebral angiostrongyliasis, and gnathostomiasis. All of these zoonotic infections result from incidental exposure to infectious nematodes. The clinical symptoms of these infections are due largely to invasive larval stages that (except in the case of *Trichinella*) do not reach maturity in humans.

TRICHINELLOSIS

Trichinellosis develops after the ingestion of meat containing cysts of *Trichinella* (e.g., pork or other meat from a carnivore). Although most infections are mild and asymptomatic, heavy infections can cause severe enteritis, periorbital edema, myositis, and (infrequently) death.

Life Cycle and Epidemiology

Image not available. Eight species of *Trichinella* are recognized as causes of infection in humans. Two species are distributed worldwide: *T. spiralis*, which is found in a great variety of carnivorous and omnivorous animals, and *T. pseudospiralis*, which is found in mammals and birds. *T. nativa* is present in Arctic regions and infects bears; *T. nelsoni* is found in equatorial eastern Africa, where it is common among felid predators and scavengers such as hyenas and bush pigs; and *T. britovi* is found in Europe, western Africa, and western Asia among carnivores but not among domestic swine. *T. murrelli* is present in North American game animals.

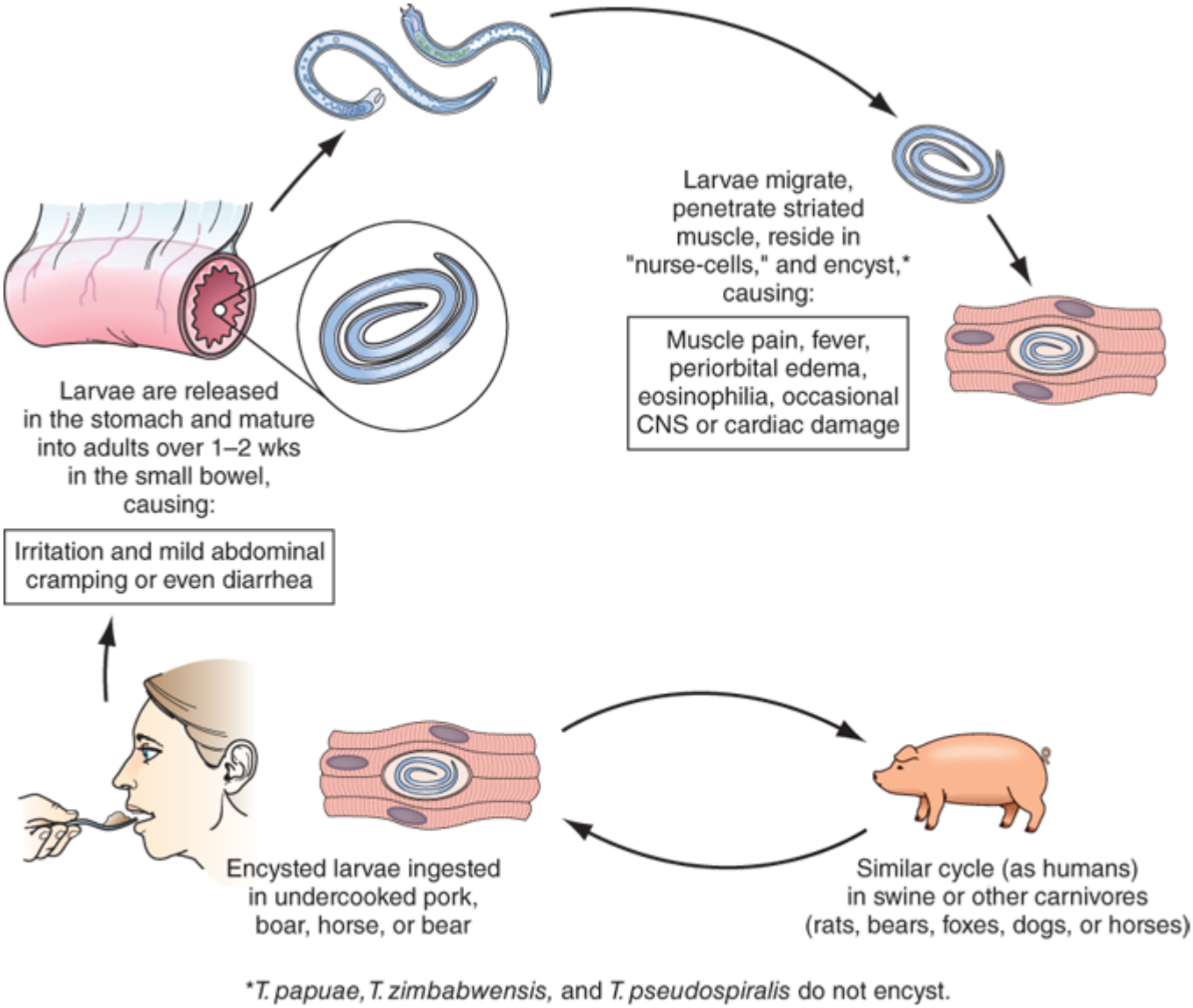
After human consumption of trichinous meat, encysted larvae are liberated by digestive acid and proteases **(Fig. 226-1)**. The larvae invade the small-bowel mucosa and mature into adult worms. After ~1 week, female worms release newborn larvae that migrate via the circulation to striated muscle. The larvae of all species except *T. pseudospiralis*, *T. papuae*, and *T. zimbabwensis* then encyst by inducing a radical transformation in the muscle cell architecture. Although host immune responses may help to expel intestinal adult worms, they have few deleterious effects on muscle-dwelling larvae.

FIGURE 226-1

Life cycle of *Trichinella* *spiralis* (cosmopolitan); *nelsoni* (equatorial Africa); *britovi* (Europe, western Africa, western Asia); *nativa* (Arctic); *murrelli* (North America); *papuae* (Papua New Guinea); *spiralis* (cosmopolitan). CNS, central nervous system.

Loading [Contrib]/a11y/accessibility-menu.js

(Reprinted from RL Guerrant et al [eds]: Tropical Infectious Diseases: Principles, Pathogens and Practice, 2nd ed, p 1218. © 2006, with permission from Elsevier Science.)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Human trichinellosis is often caused by the ingestion of infected pork products and thus can occur in almost any location where the meat of domestic or wild swine is eaten. Human trichinellosis may also be acquired from the meat of other animals, including dogs (in parts of Asia and Africa), horses (in Italy and France), and bears and walruses (in northern regions). Although cattle (being herbivores) are not natural hosts of *Trichinella*, beef has been implicated in outbreaks when contaminated or adulterated with trichinous pork. Laws that prohibit the feeding of uncooked garbage to pigs have greatly reduced the transmission of trichinellosis in the United States. About 12 cases of trichinellosis are reported annually in this country, but most mild cases probably remain undiagnosed. Recent U.S. and Canadian outbreaks have been attributable to consumption of wild game (especially bear meat) and, less frequently, of pork.

Pathogenesis and Clinical Features

Loading [Contrib]/a11y/accessibility-menu.js

C e from the successive phases of parasite enteric invasion, larval migration, and muscle encystment (Fig. 226-1). Most light infections (those with <10 larvae per gram

of muscle) are asymptomatic, whereas heavy infections (which can involve >50 larvae per gram of muscle) can be life-threatening. Invasion of the gut by large numbers of parasites occasionally provokes diarrhea during the first week after infection. Abdominal pain, constipation, nausea, or vomiting also may be prominent.

Symptoms due to larval migration and muscle invasion begin to appear in the second week after infection. The migrating *Trichinella* larvae provoke a marked local and systemic hypersensitivity reaction, with fever and hypereosinophilia. Periorbital and facial edema is common, as are hemorrhages in the subconjunctivae, retina, and nail beds (“splinter” hemorrhages). A maculopapular rash, headache, cough, dyspnea, or dysphagia sometimes develops. Myocarditis with tachyarrhythmias or heart failure—and, less commonly, encephalitis or pneumonitis—may develop and accounts for most deaths of patients with trichinellosis.

Upon onset of larval encystment in muscle 2–3 weeks after infection, symptoms of myositis with myalgias, muscle edema, and weakness develop, usually overlapping with the inflammatory reactions to migrating larvae. The most commonly involved muscle groups include the extraocular muscles; the biceps; and the muscles of the jaw, neck, lower back, and diaphragm. Peaking ~3 weeks after infection, symptoms subside only gradually during a prolonged convalescence. Uncommon infections with *T. pseudospiralis*, whose larvae do not encapsulate in muscles, elicit prolonged polymyositis-like illness.

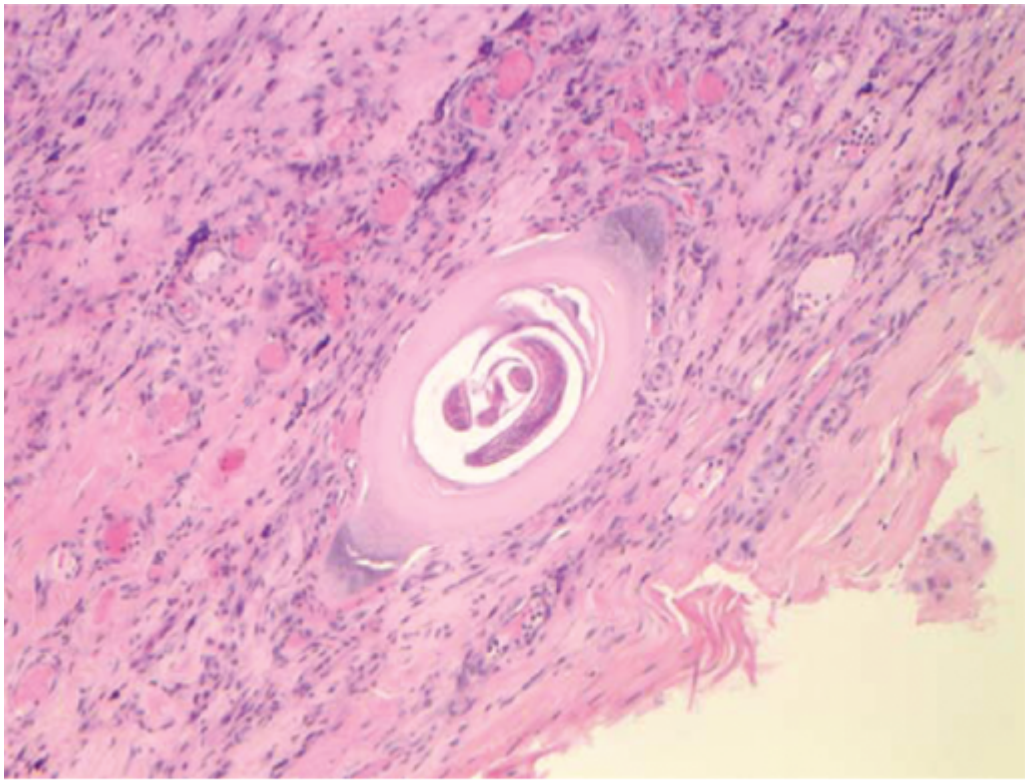
Laboratory Findings and Diagnosis

Blood eosinophilia develops in >90% of patients with symptomatic trichinellosis and may peak at a level of >50% 2–4 weeks after infection. Serum levels of muscle enzymes, including creatine phosphokinase, are elevated in most symptomatic patients. Patients should be questioned thoroughly about their consumption of pork or wild animal meat and about illness in other individuals who ate the same meat. A presumptive clinical diagnosis can be based on fevers, eosinophilia, periorbital edema, and myalgias after a suspect meal. A rise in the titer of parasite-specific antibody, which usually does not occur until after the third week of infection, confirms the diagnosis. Alternatively, a definitive diagnosis requires surgical biopsy of at least 1 g of involved muscle; the yields are highest near tendon insertions. The fresh muscle tissue should be compressed between glass slides and examined microscopically **(Fig. 226-2)** because larvae may be missed by examination of routine histopathologic sections alone.

FIGURE 226-2

***Trichinella* larva** encysted in a characteristic hyalinized capsule in striated muscle tissue. *(Photo/Wadsworth Center, New York State Department of Health. Reprinted from MMWR 53:606, 2004; public domain.)*

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT Trichinellosis

Most lightly infected patients recover uneventfully with bed rest, antipyretics, and analgesics. Glucocorticoids like prednisone **(Table 226-1)** are beneficial for severe myositis and myocarditis. Mebendazole and albendazole are active against enteric stages of the parasite, but their efficacy against encysted larvae has not been conclusively demonstrated.

Loading [Contrib]/a11y/accessibility-menu.js

TABLE 226-1 Therapy for Tissue Nematode Infections

Infection	Severity	Treatment
Trichinellosis	Mild	Supportive
	Moderate	Albendazole (400 mg bid × 8–14 days) *or* Mebendazole (200–400 mg tid × 3 days, then 400 mg tid × 8–14 days)
	Severe	Add glucocorticoids (e.g., prednisone, 1 mg/kg qd × 5 days)
Visceral larva migrans	Mild to moderate	Supportive
	Severe	Glucocorticoids (as above)
	Ocular	Not fully defined; albendazole (800 mg bid for adults, 400 mg bid for children) with glucocorticoids × 5–20 days has been effective
Cutaneous larva migrans		Ivermectin (single dose, 200 μg/kg) *or* Albendazole (200 mg bid × 3 days)
Angiostrongyliasis	Mild to moderate	Supportive
	Severe	Glucocorticoids (as above)
Gnathostomiasis		Ivermectin (200 μg/kg per day × 2 days) *or* Albendazole (400 mg bid × 21 days)

Prevention

Larvae are usually killed by cooking pork until it is no longer pink or by freezing it at −15°C for 3 weeks. However, Arctic *T. nativa* larvae in walrus or bear meat are relatively resistant and may remain viable despite freezing.

VISCERAL AND OCULAR LARVA MIGRANS

Visceral larva migrans is a syndrome caused by nematodes that are normally parasitic for nonhuman host species. In humans, these nematode larvae do not develop into adult worms but instead migrate through host tissues and elicit eosinophilic inflammation. The most common form of visceral larva migrans is toxocariasis due to larvae of the canine ascarid *Toxocara canis*; the syndrome is due less commonly to the feline ascarid *T. cati* and even less commonly to the pig ascarid *Ascaris suum*. Rare cases with eosinophilic meningoencephalitis have been caused by the raccoon ascarid *Baylisascaris procyonis*.

Life Cycle and Epidemiology

Image not available. The canine roundworm *T. canis* is distributed among dogs worldwide. Ingestion of infective eggs by dogs is followed by liberation of *Toxocara* larvae, which penetrate the gut wall and migrate intravascularly into canine tissues, where most remain in a developmentally arrested state. During pregnancy, some larvae resume migration in bitches and infect puppies prenatally (through transplacental transmission) or after birth (through suckling). Thus, in lactating bitches and puppies, larvae return to the intestinal tract and develop into adult worms, which produce eggs that are released in the feces. Eggs must undergo embryonation over several weeks to become infectious. Humans acquire toxocariasis mainly by eating soil contaminated by puppy feces that contains infective *T. canis* eggs. Visceral larva migrans is most common among children who habitually eat dirt.

Pathogenesis and Clinical Features

Loading [Contrib]/a11y/accessibility-menu.js

Clinical disease most commonly afflicts preschool children. After humans ingest *Toxocara* eggs, the larvae hatch and penetrate the intestinal mucosa, from which they are carried by the circulation to a wide variety of organs and tissues. The larvae invade the liver, lungs, central nervous system (CNS), and other sites, provoking intense local eosinophilic granulomatous responses. The degree of clinical illness depends on larval number and tissue distribution, reinfection, and host immune responses. Most light infections are asymptomatic and may be evidenced only by blood eosinophilia. Characteristic symptoms of visceral larva migrans include fever, malaise, anorexia and weight loss, cough, wheezing, and rashes. Hepatosplenomegaly is common. These features may be accompanied by extraordinary peripheral eosinophilia at levels that may approach 90%. Uncommonly, seizures or behavioral disorders develop. Rare deaths are due to severe neurologic, pneumonic, or myocardial involvement.

The ocular form of the larva migrans syndrome occurs when *Toxocara* larvae invade the eye. An eosinophilic granulomatous mass, most commonly in the posterior pole of the retina, develops around the entrapped larva. The retinal lesion can mimic retinoblastoma in appearance, and mistaken diagnosis of the latter condition can lead to unnecessary enucleation. The spectrum of eye involvement also includes endophthalmitis, uveitis, and chorioretinitis. Unilateral visual disturbances, strabismus, and eye pain are the most common presenting symptoms. In contrast to visceral larva migrans, ocular toxocariasis usually develops in older children or young adults with no history of pica; these patients seldom have eosinophilia or visceral manifestations.

Diagnosis

In addition to eosinophilia, leukocytosis and hypergammaglobulinemia may be evident. Transient pulmonary infiltrates are apparent on chest x-rays of about one-half of patients with symptoms of pneumonitis. The clinical diagnosis can be confirmed by an enzyme-linked immunosorbent assay for toxocaral antibodies. Stool examination for parasite eggs is worthless in toxocariasis, since the larvae do not develop into egg-producing adults in humans.

TREATMENT

TREATMENT Visceral and Ocular Larva Migrans

The vast majority of *Toxocara* infections are self-limited and resolve without specific therapy. In patients with severe myocardial, CNS, or pulmonary involvement, glucocorticoids may be employed to reduce inflammatory complications. Available anthelmintic drugs, including mebendazole and albendazole, have not been shown conclusively to alter the course of larva migrans. Control measures include prohibiting dog excreta in public parks and playgrounds, deworming dogs, and preventing pica in children. Treatment of ocular disease is not fully defined, but the administration of albendazole in conjunction with glucocorticoids has been effective (Table 226-1).

CUTANEOUS LARVA MIGRANS

Cutaneous larva migrans (“creeping eruption”) is a serpiginous skin eruption caused by burrowing larvae of animal hookworms, usually the dog and cat hookworm *Ancylostoma braziliense*. The larvae hatch from eggs passed in dog and cat feces and mature in the soil. Humans become infected after skin contact with soil in areas frequented by dogs and cats, such as areas underneath house porches. Cutaneous larva migrans is prevalent among children and travelers in regions with warm humid climates, including the southeastern United States.

After larvae penetrate the skin, erythematous lesions form along the tortuous tracks of their migration tLoading [Contrib]/a11y/accessibility-menu.js the larvae advance several centimeters in a day. The intensely pruritic lesions may occur anywhere on the body and can be numerous if the patient has lain on the

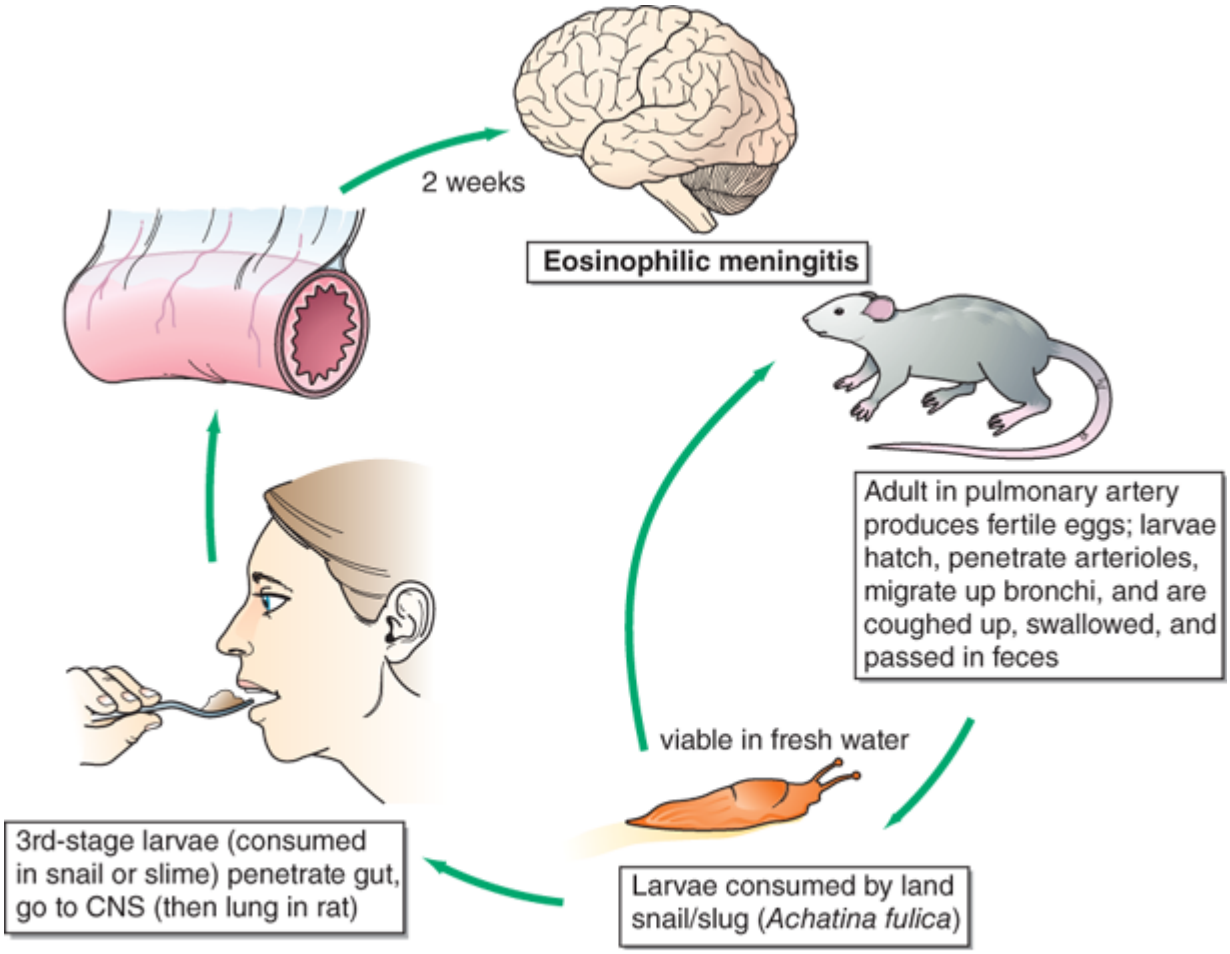
ground. Vesicles and bullae may form later. The animal hookworm larvae do not mature in humans and, without treatment, will die after an interval ranging from weeks to a couple of months, with resolution of skin lesions. The diagnosis is made on clinical grounds. Skin biopsies only rarely detect diagnostic larvae. Symptoms can be alleviated by ivermectin or albendazole (Table 226-1).

ANGIOSTRONGYLIASIS

Angiostrongylus cantonensis, the rat lungworm, is the most common cause of human eosinophilic meningitis **(Fig. 226-3)**.

FIGURE 226-3

Life cycle of *Angiostrongylus cantonensis* (rat lung worm) found in Southeast Asia and the Pacific Basin as well as on Caribbean islands, in countries of Central and South America, and in the southern United States. CNS, central nervous system. *(Reprinted from RL Guerrant et al [eds]: Tropical Infectious Diseases: Principles, Pathogens and Practice, 2nd ed, p 1225. © 2006, with permission from Elsevier Science.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Life Cycle and Epidemiology

Loading [Contrib]/a11y/accessibility-menu.js

ccurs principally in Southeast Asia and the Pacific Basin but has spread to other areas of the world, including the Caribbean islands, countries in Central and

South America, and the southern United States. *A. cantonensis* larvae produced by adult worms in the rat lung migrate to the gastrointestinal tract and are expelled with the feces. They develop into infective larvae in land snails and slugs. Humans acquire the infection by ingesting raw infected mollusks; vegetables contaminated by mollusk slime; or crabs, freshwater shrimp, and certain marine fish that have themselves eaten infected mollusks. The larvae then migrate to the brain.

Pathogenesis and Clinical Features

The parasites eventually die in the CNS, but not before initiating pathologic consequences that, in heavy infections, can result in permanent neurologic sequelae or death. Migrating larvae cause marked local eosinophilic inflammation and hemorrhage, with subsequent necrosis and granuloma formation around dying worms. Clinical symptoms develop 2–35 days after the ingestion of larvae. Patients usually present with an insidious or abrupt excruciating frontal, occipital, or bitemporal headache. Neck stiffness, nausea and vomiting, and paresthesias are also common. Fever, cranial and extraocular nerve palsies, seizures, paralysis, and lethargy are uncommon.

Laboratory Findings

Examination of cerebrospinal fluid (CSF) is mandatory in suspected cases and usually reveals an elevated opening pressure, a white blood cell count of 150–2000/μL, and an eosinophilic pleocytosis of >20%. The protein concentration is usually elevated and the glucose level normal. The larvae of *A. cantonensis* are only rarely seen in CSF. Peripheral-blood eosinophilia may be mild. The diagnosis is generally based on the clinical presentation of eosinophilic meningitis together with a compatible epidemiologic history.

TREATMENT

TREATMENT Angiostrongyliasis

Specific chemotherapy is not of benefit in angiostrongyliasis; larvicidal agents may exacerbate inflammatory brain lesions. Management consists of supportive measures, including the administration of analgesics, sedatives, and—in severe cases—glucocorticoids (Table 226-1). Repeated lumbar punctures with removal of CSF can relieve symptoms. In most patients, cerebral angiostrongyliasis has a self-limited course, and recovery is complete. The infection may be prevented by adequately cooking snails, crabs, and prawns and inspecting vegetables for mollusk infestation. Other parasitic or fungal causes of eosinophilic meningitis in endemic areas may include gnathostomiasis (see below), paragonimiasis **(Chap. 229)**, schistosomiasis **(Chap. 229)**, neurocysticercosis **(Chap. 230)**, and coccidioidomycosis **(Chap. 208)**.

GNATHOSTOMIASIS

Infection of human tissues with larvae of *Gnathostoma spinigerum* can cause eosinophilic meningoencephalitis, migratory cutaneous swellings, or invasive masses of the eye and visceral organs.

Life Cycle and Epidemiology

Image not available. Human gnathostomiasis occurs in many countries and is notably endemic in Southeast Asia and parts of China and Japan. In nature, the mature adult worms parasitize the gastrointestinal tract of dogs and cats. First-stage larvae hatch from eggs passed into water and are i as). Infective third-stage larvae develop in the flesh of many animal species (including fish, frogs, eels, snakes, chickens, and ducks) that have eaten either

Loading [Contrib]/a11y/accessibility-menu.js

infected *Cyclops* or another infected second intermediate host. Humans typically acquire the infection by eating raw or undercooked fish or poultry. Raw fish dishes, such as *som fak* in Thailand and *sashimi* in Japan, account for many cases of human gnathostomiasis. Some cases in Thailand result from the local practice of applying frog or snake flesh as a poultice.

Pathogenesis and Clinical Features

Clinical symptoms are due to the aberrant migration of a single larva into cutaneous, visceral, neural, or ocular tissues. After invasion, larval migration may cause local inflammation, with pain, cough, or hematuria accompanied by fever and eosinophilia. Painful, itchy, migratory swellings may develop in the skin, particularly in the distal extremities or periorbital area. Cutaneous swellings usually last ~1 week, but often recur intermittently over many years. Larval invasion of the eye can provoke a sight-threatening inflammatory response. Invasion of the CNS results in eosinophilic meningitis with myeloencephalitis, a serious complication due to ascending larval migration along a large nerve tract. Patients characteristically present with agonizing radicular pain and paresthesias in the trunk or a limb, which are followed shortly by paraplegia. Cerebral involvement, with focal hemorrhages and tissue destruction, is often fatal.

Diagnosis and Treatment

Cutaneous migratory swellings with marked peripheral eosinophilia, supported by an appropriate geographic and dietary history, generally constitute an adequate basis for a clinical diagnosis of gnathostomiasis. However, patients may present with ocular or cerebrospinal involvement without antecedent cutaneous swellings. In the latter case, eosinophilic pleocytosis is demonstrable (usually along with hemorrhagic or xanthochromic CSF), but worms are almost never recovered from CSF. Surgical removal of the parasite from subcutaneous or ocular tissue, though rarely feasible, is both diagnostic and therapeutic. Albendazole or ivermectin may be helpful (Table 226-1). At present, cerebrospinal involvement is managed with supportive measures and generally with a course of glucocorticoids. Gnathostomiasis can be prevented by adequate cooking of fish and poultry in endemic areas.

FURTHER READING

Feldmeier H, Schuster A: Mini review: Hookworm-related cutaneous larva migrans. Eur J Clin Microbiol Infect Dis 31:915, 2012. [PubMed: 21922198]
Lupi O et al: Mucocutaneous manifestations of helminth infections. Nematodes. J Am Acad Dermatol 73:929, 2015. [PubMed: 26568337]
Martins YC et al: Central nervous system manifestations of *Angiostrongylus cantonensis* infection. Acta Trop 141:46, 2015. [PubMed: 25312338]
Pozio E, Zarlenga DS: New pieces of the *Trichinella* puzzle. Int J Parasitol 43:983, 2013. [PubMed: 23816802]
Sitcar AD et al: Raccoon roundworm infection associated with central nervous system disease and ocular disease—six states, 2013–2015. MMWR 65:930, 2016. [PubMed: 27608169]
Wilson NO et al: Trichinellosis surveillance—United States, 2008–2012. MMWR 64:SS-1, 2015.

Loading [Contrib]/a11y/accessibility-menu.js

Chapter 227: Intestinal Nematode Infections

Peter F. Weller; Thomas B. Nutman

FIGURE 227-1

INTRODUCTION

More than a billion persons worldwide are infected with one or more species of intestinal nematodes. **Table 227-1** summarizes biologic and clinical features of infections due to the major intestinal parasitic nematodes. These parasites are most common in regions with poor fecal sanitation, particularly in resource-poor countries in the tropics and subtropics, but they have also been seen with increasing frequency among immigrants and refugees to resource-rich countries. Although nematode infections are not usually fatal, they contribute to malnutrition and diminished work capacity. It is interesting that these helminth infections may protect some individuals from allergic disease. Humans may on occasion be infected with nematode parasites that ordinarily infect animals; these zoonotic infections produce diseases such as trichostrongyliasis, anisakiasis, capillariasis, and abdominal angiostrongyliasis.

Loading [Contrib]/a11y/accessibility-menu.js

TABLE 227-1 Major Human Intestinal Parasitic Nematodes

	Parasitic Nematode				
Feature	***Ascaris lumbricoides* (Roundworm)**	***Necator americanus*, *Ancylostoma duodenale* (Hookworm)**	***Strongyloides stercoralis***	***Trichuris trichiura* (Whipworm)**	***Enterobius vermicularis* (Pinworm)**
Global prevalence in humans (millions)	807	576	100	604	209
Endemic areas	Worldwide	Hot, humid regions	Hot, humid regions	Worldwide	Worldwide
Infective stage	Egg	Filariform larva	Filariform larva	Egg	Egg
Route of infection	Oral	Percutaneous	Percutaneous or autoinfective	Oral	Oral
Gastrointestinal location of worms	Jejunal lumen	Jejunal mucosa	Small-bowel mucosa	Cecum, colonic mucosa	Cecum, appendix
Adult worm size	15–40 cm	7–12 mm	2 mm	30–50 mm	8–13 mm (female)
Pulmonary passage of larvae	Yes	Yes	Yes	No	No
Incubation period[a] (days)	60–75	40–100	17–28	70–90	35–45
Longevity	1 year	*N. americanus*: 2–5 years *A. duodenale:* 6–8 years	Decades (owing to autoinfection)	5 years	2 months
Fecundity (eggs/day/worm)	240,000	*N. americanus:* 4000–10,000 *A. duodenale:* 10,000–25,000	5000–10,000	3000–7000	2000
Principal symptoms	Rarely, biliary obstruction or, in heavy infections, gastrointestinal obstruction	Iron-deficiency anemia in heavy infection	Gastrointestinal symptoms; malabsorption or sepsis in hyperinfection	Gastrointestinal symptoms or anemia in heavy infection	Perianal pruritus
Diagnostic stage	Eggs in stool	Eggs in fresh stool, larvae in old stool	Larvae in stool or duodenal aspirate; sputum in hyperinfection	Eggs in stool	Eggs from perianal skin on cellulose acetate tape
Treatment	Mebendazole Albendazole	Mebendazole Albendazole	Ivermectin Albendazole	Mebendazole Albendazole Ivermectin	Mebendazole Albendazole

Loading [Contrib]/a11y/accessibility-menu.js

[a]Time from infection to egg production by mature female worm.

Intestinal nematodes are roundworms; they range in length from 1 mm to many centimeters when mature (Table 227-1). Their life cycles are complex and highly varied; some species, including *Strongyloides stercoralis* and *Enterobius vermicularis*, can be transmitted directly from person to person, while others, such as *Ascaris lumbricoides*, *Necator americanus*, and *Ancylostoma duodenale*, require a soil phase for development. Because most helminth parasites do not self-replicate, the acquisition of a heavy burden of adult worms requires repeated exposure to the parasite in its infectious stage, whether larva or egg. Hence, clinical disease, as opposed to asymptomatic (or subclinical) infection, generally develops only with prolonged exposure in an endemic area and is typically related to infection intensity. In persons with marginal nutrition, intestinal helminth infections may impair growth and development. Eosinophilia and elevated serum IgE levels are features of many helminth infections and, when unexplained, should always prompt a search for intestinal helminths. Significant protective immunity to intestinal nematodes appears not to develop in humans, although the host immune response to these infections has not been elucidated in detail.

ASCARIASIS

A. lumbricoides is the largest intestinal nematode parasite of humans, reaching up to 40 cm in length. Most infected individuals have low worm burdens and are asymptomatic. Clinical disease arises from larval migration in the lungs or effects of the adult worms in the intestines.

Life Cycle

Adult worms live in the lumen of the small intestine. Mature female *Ascaris* worms are extraordinarily fecund, each producing up to 240,000 eggs a day that pass with the feces. Ascarid eggs, which are remarkably resistant to environmental stresses, become infective after several weeks of maturation in the soil and can remain infective for years. After infective eggs are swallowed, larvae hatched in the intestine invade the mucosa, migrate through the circulation to the lungs, break into the alveoli, ascend the bronchial tree, and return—through swallowing—to the small intestine, where they develop into adult worms. Between 2 and 3 months elapse between initial infection and egg production. Adult worms live for 1–2 years.

Epidemiology

Image not available. *Ascaris* is widely distributed in tropical and subtropical regions as well as in other humid areas in more temperate regions of the world. Transmission typically occurs through fecally contaminated soil and is due either to a lack of sanitary facilities or to the use of human feces as fertilizer. With their propensity for hand-to-mouth fecal carriage, younger children are most often affected. Infection outside endemic areas, though uncommon, can occur when eggs on transported vegetables are ingested.

Clinical Features

During the lung phase of larval migration, ~9–12 days after egg ingestion, patients may develop an irritating nonproductive cough and burning substernal discomfort that is aggravated by coughing or deep inspiration. Dyspnea and blood-tinged sputum are less common. Fever can occur. Eosinophilia develops during this symptomatic phase and subsides slowly over weeks. Chest x-rays may reveal evidence of eosinophilic pneumonitis (Löffler's syndrome), with rounded infiltrates a few millimeters to several centimeters in size. These infiltrates may be transient and intermittent, clearing after several weeks. With the migration of the parasite, seasonal pneumonitis with eosinophilia may develop in previously infected and sensitized hosts.

Loading [Contrib]/a11y/accessibility-menu.js

In established infections, adult worms in the small intestine usually cause no symptoms. In heavy infections, particularly in children, a large bolus of entangled worms can cause pain and small-bowel obstruction, sometimes complicated by perforation, intussusception, or volvulus. Single worms may cause disease when they migrate into aberrant sites. A large worm can enter and occlude the biliary tree, causing biliary colic, cholecystitis, cholangitis, pancreatitis, or (rarely) intrahepatic abscesses. Migration of an adult worm up the esophagus can provoke coughing and oral expulsion of the worm. In highly endemic areas, intestinal and biliary ascariasis can rival acute appendicitis and gallstones as causes of surgical acute abdomen.

Laboratory Findings

Most cases of ascariasis can be diagnosed by microscopic detection of characteristic *Ascaris* eggs (65 by 45 µm) in fecal samples, although increasingly polymerase chain reaction (PCR) of DNA extracted from stool is being used in research and some clinical settings. Occasionally, patients present after passing an adult worm—identifiable by its large size and smooth cream-colored surface—in the stool or, much less commonly, through the mouth or nose. During the early transpulmonary migratory phase, when eosinophilic pneumonitis occurs, larvae can be found in sputum or gastric aspirates before diagnostic eggs appear in the stool. The eosinophilia that is prominent during this early stage usually decreases to minimal levels in established infection. Adult worms may be visualized, occasionally serendipitously, on contrast studies of the gastrointestinal tract. A plain abdominal film may reveal masses of worms in gas-filled loops of bowel in patients with intestinal obstruction. Pancreaticobiliary worms can be detected by ultrasound and endoscopic retrograde cholangiopancreatography; the latter method also has been used to extract biliary *Ascaris* worms.

TREATMENT

TREATMENT Ascariasis

Ascariasis should always be treated to prevent potentially serious complications. Albendazole (400 mg once), mebendazole (100 g twice daily for 3 days or 500 mg once), or ivermectin (150–200 µg/kg once) is effective. These medications are contraindicated in pregnancy, however. Mild diarrhea and abdominal pain are uncommon side effects of these agents. Partial intestinal obstruction should be managed with nasogastric suction, IV fluid administration, and instillation of piperazine through the nasogastric tube, but complete obstruction and its severe complications require immediate surgical intervention.

HOOKWORM

Image not available. Two species (*A. duodenale* and *N. americanus*) are responsible for most human hookworm infections, although *A. ceylanicum* is being recognized as a major hookworm pathogen in parts of Asia. Most infected individuals are asymptomatic. Hookworm disease develops from a combination of factors—a heavy worm burden, a prolonged duration of infection, and an inadequate iron intake—and results in iron-deficiency anemia and, on occasion, hypoproteinemia.

Life Cycle

Adult hookworms, which are ~1 cm long, use buccal teeth (*Ancylostoma*) or cutting plates (*Necator*) to attach to the small-bowel mucosa and suck blood (0.2 mL/d per *Ancylostoma* adult) and interstitial fluid. The adult hookworms produce thousands of eggs daily. The eggs are deposited with feces in soil, where rhabditiform larvae hatch and develop over a 1-week period into infectious filariform larvae. Infective larvae penetrate the skin and reach the lungs by way of the bloodstream. There they invade alveoli and ascend the airways before being swallowed and reaching the small intestine. The

Loading [Contrib]/a11y/accessibility-menu.js

prepatent period from skin invasion to appearance of eggs in the feces is ~6–8 weeks, but it may be longer with *A. duodenale*. Larvae of *A. duodenale*, if swallowed, can survive and develop directly in the intestinal mucosa. Adult hookworms may survive over a decade but usually live ~6–8 years for *A. duodenale* and 2–5 years for *N. americanus*.

Epidemiology

Image not available. *A. duodenale* is prevalent in southern Europe, North Africa, and northern Asia, and *N. americanus* is the predominant species in the Western Hemisphere and equatorial Africa. *A. ceylanicum* is most prevalent in Southeast Asia. The species can overlap geographically, particularly in Southeast Asia. Age prevalence studies have shown a constant increase in hookworm prevalence over time; older children have the greatest intensity of hookworm infection; however, in rural areas where fields are fertilized with human feces, older working adults also may be heavily infected.

Clinical Features

Most hookworm infections are clinically asymptomatic. Infective larvae may provoke pruritic maculopapular dermatitis (“ground itch”) at the site of skin penetration as well as serpiginous tracks of subcutaneous migration (similar to those of cutaneous larva migrans; **Chap. 226**) in previously sensitized hosts. Larvae migrating through the lungs occasionally cause mild transient pneumonitis, but this condition develops less frequently in hookworm infection than in ascariasis. In the early intestinal phase, infected persons may develop epigastric pain (often with postprandial accentuation), inflammatory diarrhea, or other abdominal symptoms accompanied by eosinophilia. The major consequence of chronic hookworm infection is iron deficiency. Symptoms are minimal if iron intake is adequate, but marginally nourished individuals develop symptoms of progressive iron-deficiency anemia and hypoproteinemia, including weakness and shortness of breath.

Laboratory Findings

The diagnosis is established by the finding of characteristic 40- by 60-μm oval hookworm eggs in the feces. Stool-concentration procedures may be required to detect light infections. Eggs of the three species are indistinguishable by light microscopy, whereas PCR has provided a significant improvement in species-specific diagnosis. In a stool sample that is not fresh, the eggs may have hatched to release rhabditiform larvae, which need to be differentiated from those of *S. stercoralis*. Hypochromic microcytic anemia, occasionally with eosinophilia or hypoalbuminemia, is characteristic of hookworm disease.

TREATMENT

TREATMENT Hookworm Infection

Hookworm infection can be treated with several safe and highly effective anthelmintic drugs, including albendazole (400 mg once) and mebendazole (500 mg once). Mild iron-deficiency anemia can often be treated with oral iron alone. Severe hookworm disease with protein loss and malabsorption necessitates nutritional support and oral iron replacement along with deworming. There is significant concern that the benzimidazoles (mebendazole and albendazole) are becoming much less effective against human hookworms.

Ancylostoma caninum and *Ancylostoma braziliense*

Image not available. *A. caninum*, the canine hookworm, has been identified as a cause of human eosinophilic enteritis, especially in northeastern Australia. In this zoonotic infection, adult hookworms

Loading [Contrib]/a11y/accessibility-menu.js

attach to the small intestine (where they may be visualized by endoscopy) and elicit abdominal pain and intense local eosinophilia. Treatment with mebendazole (100 mg twice daily for 3 days) or albendazole (400 mg once) or endoscopic removal is effective. Both of these animal hookworm species can cause cutaneous larva migrans (“creeping eruption”; **Chap. 226**).

STRONGYLOIDIASIS

S. stercoralis is distinguished by its ability—unique among helminths (except for *Capillaria*; see below)—to replicate in the human host. This capacity permits ongoing cycles of autoinfection as infective larvae are internally produced. Infection with *S. stercoralis* can thus persist for decades without further exposure of the host to exogenous infective larvae. In immunocompromised hosts, large numbers of invasive *Strongyloides* larvae can disseminate widely and can be fatal.

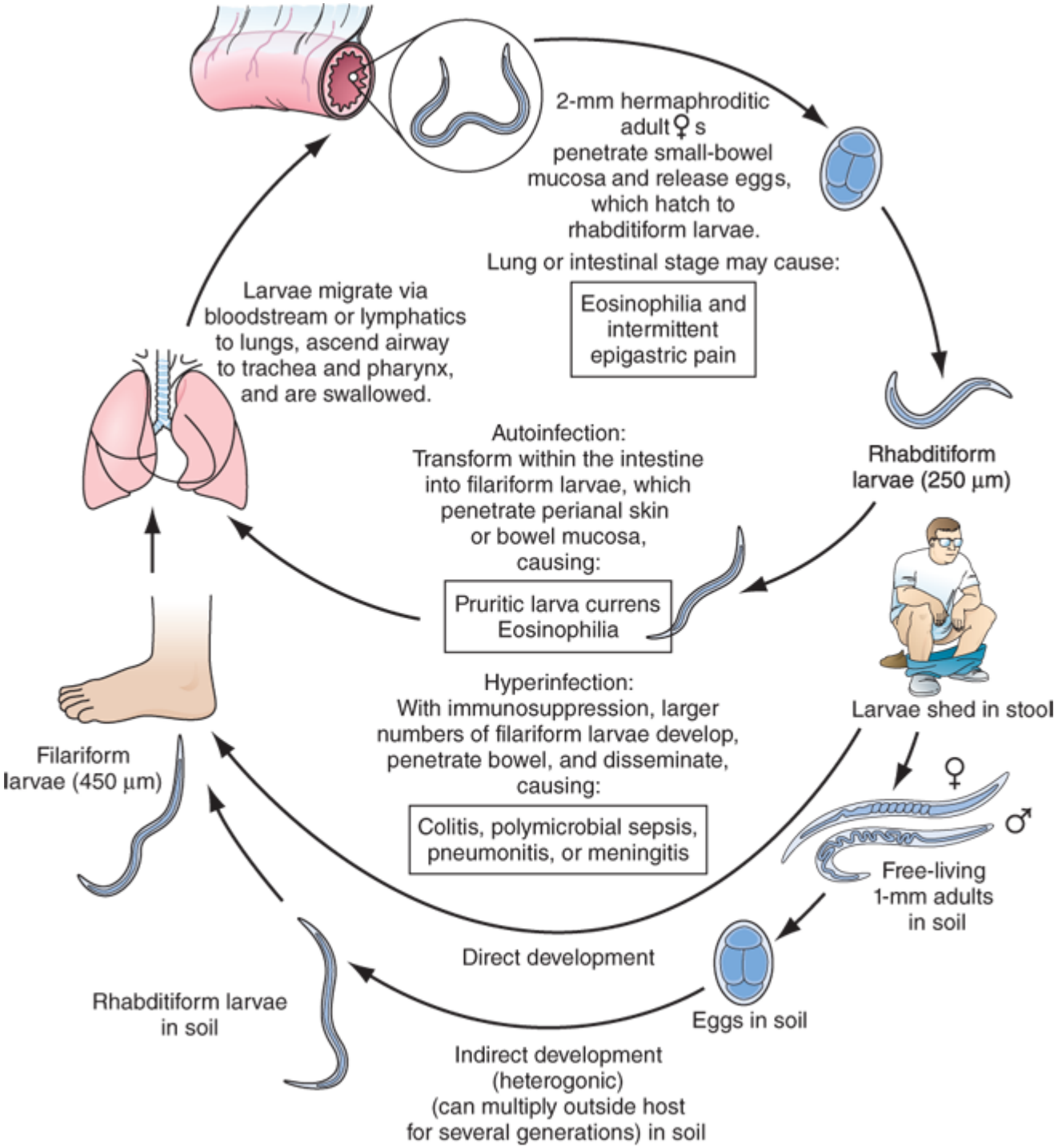
Life Cycle

In addition to a parasitic cycle of development, *Strongyloides* can undergo a free-living cycle of development in the soil **(Fig. 227-1)**. This adaptability facilitates the parasite’s survival in the absence of mammalian hosts. Rhabditiform larvae passed in feces can transform into infectious filariform larvae either directly or after a free-living phase of development. Humans acquire *S. stercoralis* when filariform larvae in fecally contaminated soil penetrate the skin or mucous membranes. The larvae then travel through the bloodstream to the lungs, where they break into the alveolar spaces, ascend the bronchial tree, are swallowed, and thereby reach the small intestine. There the larvae mature into adult worms that penetrate the mucosa of the proximal small bowel. The minute (2-mm-long) parasitic adult female worms reproduce by parthenogenesis; adult males do not exist. Eggs hatch in the intestinal mucosa, releasing rhabditiform larvae that migrate to the lumen and pass with the feces into soil. Alternatively, rhabditiform larvae in the bowel can develop directly into filariform larvae that penetrate the colonic wall or perianal skin and enter the circulation to repeat the migration that establishes ongoing internal reinfection. This autoinfection cycle allows strongyloidiasis to persist for decades.

FIGURE 227-1

Life cycle of *Strongyloides stercoralis.* *(Adapted from Guerrant RL et al [eds]: Tropical Infectious Diseases: Principles, Pathogens and Practice, 2nd ed, p 1276. © 2006, with permission from Elsevier Science.)*

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Epidemiology

Image not available. *S. stercoralis* is spottily distributed in tropical areas and other hot, humid regions and is particularly common in Southeast Asia, sub-Saharan Africa, and Brazil. In the United States, the parasite is endemic in parts of the Southeast and is found in immigrants, refugees, travelers, and military personnel who have lived in endemic areas.

Clinical Features

Loading [Contrib]/a11y/accessibility-menu.js

In uncomplicated strongyloidiasis, many patients are asymptomatic or have mild cutaneous and/or abdominal symptoms. Recurrent urticaria, often involving the buttocks and wrists, is the most common cutaneous manifestation. Migrating larvae can elicit a pathognomonic serpiginous eruption, *larva currens* ("running larva"). This pruritic, raised, erythematous lesion advances as rapidly as 10 cm/h along the course of larval migration. Adult parasites burrow into the duodenojejunal mucosa and can cause abdominal (usually midepigastric) pain, which resembles peptic ulcer pain except that it is aggravated by food ingestion. Nausea, diarrhea, gastrointestinal bleeding, mild chronic colitis, and weight loss can occur. Small-bowel obstruction may develop with early, heavy infection. Pulmonary symptoms are rare in uncomplicated strongyloidiasis. Eosinophilia is common, with levels fluctuating over time.

The ongoing autoinfection cycle of *S. stercoralis* is normally constrained by unknown factors of the host's immune system. Abrogation of host immunity, especially with glucocorticoid therapy and much less commonly with other immunosuppressive medications, leads to hyperinfection, with the generation of large numbers of filariform larvae. Colitis, enteritis, or malabsorption may develop. In disseminated strongyloidiasis, larvae may invade not only gastrointestinal tissues and the lungs but also the central nervous system, peritoneum, liver, and kidneys. Moreover, bacteremia may develop because of the passage of enteric flora through disrupted mucosal barriers. Gram-negative sepsis, pneumonia, or meningitis may complicate or dominate the clinical course. Eosinophilia is often absent in severely infected patients. Disseminated strongyloidiasis, particularly in patients with unsuspected infection who are given glucocorticoids, can be fatal. Strongyloidiasis is a frequent complication of infection with human T cell lymphotropic virus type 1 (HTLV-1), but disseminated strongyloidiasis is not common among patients infected with HIV-1.

Diagnosis

In uncomplicated strongyloidiasis, the finding of rhabditiform larvae in feces is diagnostic. Rhabditiform larvae are ~250 μm long, with a short buccal cavity that distinguishes them from hookworm larvae. In uncomplicated infections, few larvae are passed and single stool examinations detect only about one-third of cases. Serial examinations and the use of the agar plate detection method improve the sensitivity of stool diagnosis. Again, PCR has begun to be used more widely and provides increased diagnostic specificity. In uncomplicated strongyloidiasis (but not in hyperinfection), microscopy-based stool examinations may be repeatedly negative. *Strongyloides* larvae may also be found by sampling of the duodenojejunal contents by aspiration or biopsy. An enzyme-linked immunosorbent assay for serum antibodies to antigens of *Strongyloides* is a sensitive method for diagnosing uncomplicated infections. Such serologic testing should be performed for patients whose geographic histories indicate potential exposure, especially those who exhibit eosinophilia and/or are candidates for glucocorticoid treatment of other conditions. In disseminated strongyloidiasis, filariform larvae should be sought in stool as well as in samples obtained from sites of potential larval migration, including sputum, bronchoalveolar lavage fluid, or surgical drainage fluid.

TREATMENT

TREATMENT Strongyloidiasis

Even in the asymptomatic state, strongyloidiasis must be treated because of the potential for subsequent dissemination and fatal hyperinfection. Ivermectin (200 μg/kg daily for 2 days) is consistently more effective than albendazole (400 mg daily for 3 days). For disseminated strongyloidiasis, treatment with ivermectin should be extended for at least 5–7 days or until the parasites have been eradicated. In immunocompromised hosts, the course of ivermectin should be repeated 2 weeks after initial treatment.

Loading [Contrib]/a11y/accessibility-menu.js

TRICHURIASIS

Image not available. Most infections with *Trichuris trichiura* are asymptomatic, but heavy infections may cause gastrointestinal symptoms. Like the other soil-transmitted helminths, whipworm is distributed globally in the tropics and subtropics and is most common among poor children from resource-poor regions of the world.

Life Cycle

Adult *Trichuris* worms reside in the colon and cecum, the anterior portions threaded into the superficial mucosa. Thousands of eggs laid daily by adult female worms pass with the feces and mature in the soil. After ingestion, infective eggs hatch in the duodenum, releasing larvae that mature before migrating to the large bowel. The entire cycle takes ~3 months, and adult worms may live for several years.

Clinical Features

Tissue reactions to *Trichuris* are mild. Most infected individuals have no symptoms or eosinophilia. Heavy infections may result in anemia, abdominal pain, anorexia, and bloody or mucoid diarrhea resembling inflammatory bowel disease. Rectal prolapse can result from massive infections in children, who often suffer from malnourishment and other diarrheal illnesses. Moderately heavy *Trichuris* burdens also contribute to growth retardation.

Diagnosis and Treatment

The characteristic 50- by 20-μm lemon-shaped *Trichuris* eggs are readily detected on stool examination. Adult worms, which are 3–5 cm long, are occasionally seen on proctoscopy. Mebendazole (500 mg once) or albendazole (400 mg daily for 3 doses) is safe and moderately effective for treatment, with cure rates of 70–90%. Ivermectin (200 μg/kg daily for 3 doses) is also safe but is not quite as efficacious as the benzimidazoles.

ENTEROBIASIS (PINWORM)

Image not available. *E. vermicularis* is more common in temperate countries than in the tropics. In the United States, ~40 million persons are infected with pinworms, with a disproportionate number of cases among children.

Life Cycle and Epidemiology

Enterobius adult worms are ~1 cm long and dwell in the cecum. Gravid female worms migrate nocturnally into the perianal region and release up to 2000 immature eggs each. The eggs become infective within hours and are transmitted by hand-to-mouth passage. From ingested eggs, larvae hatch and mature into adults. This life cycle takes ~1 month, and adult worms survive for ~2 months. Self-infection results from perianal scratching and transport of infective eggs on the hands or under the nails to the mouth. Because of the ease of person-to-person spread, pinworm infections are common among family members.

Clinical Features

Most pinworm infections are asymptomatic. Perianal pruritus is the cardinal symptom. The itching, which is often worse at night as a result of the nocturnal migration of the female worms, may lead to excoriation and bacterial superinfection. Heavy infections have been alleged to cause abdominal pain

Loading [Contrib]/a11y/accessibility-menu.js

and weight loss. On rare occasions, pinworms invade the female genital tract, causing vulvovaginitis and pelvic or peritoneal granulomas. Eosinophilia is uncommon.

Diagnosis

Since pinworm eggs are not released in feces, the diagnosis cannot be made by conventional fecal ova and parasite tests. Instead, eggs are detected by the application of clear cellulose acetate tape to the perianal region in the morning. After the tape is transferred to a slide, microscopic examination will detect pinworm eggs, which are oval, measure 55 by 25 μm, and are flattened along one side.

TREATMENT

TREATMENT Enterobiasis

Infected children and adults should be treated with mebendazole (100 mg once) or albendazole (400 mg once), with the same treatment repeated after 2 weeks. Treatment of household members is advocated to eliminate asymptomatic reservoirs of potential reinfection.

TRICHOSTRONGYLIASIS

 Image not available. *Trichostrongylus* species, which are normally parasites of herbivorous animals, occasionally infect humans, particularly in Asia and Africa. Humans acquire the infection by accidentally ingesting *Trichostrongylus* larvae on contaminated leafy vegetables. The larvae do not migrate in humans but mature directly into adult worms in the small bowel. These worms ingest far less blood than hookworms; most infected persons are asymptomatic, but heavy infections may give rise to mild anemia and eosinophilia. In stool examinations, *Trichostrongylus* eggs resemble hookworm eggs but are larger (85 by 115 μm). Treatment consists of mebendazole or albendazole **(Chap. 217)**.

ANISAKIASIS

 Image not available. Anisakiasis is a gastrointestinal infection caused by the accidental ingestion in uncooked saltwater fish of nematode larvae belonging to the family Anisakidae. The incidence of anisakiasis in the United States has increased as a result of the growing popularity of raw fish dishes. Most cases occur in Japan, the Netherlands, and Chile, where raw fish—sashimi, pickled green herring, and ceviche, respectively—are national culinary staples. Anisakid nematodes parasitize large sea mammals such as whales, dolphins, and seals. As part of a complex parasitic life cycle involving marine food chains, infectious larvae migrate to the musculature of a variety of fish. Both *Anisakis simplex* and *Pseudoterranova decipiens* have been implicated in human anisakiasis, but an identical gastric syndrome may be caused by the red larvae of eustrongylid parasites of fish-eating birds.

When humans consume infected raw fish, live larvae may be coughed up within 48 h. Alternatively, larvae may immediately penetrate the mucosa of the stomach. Within hours, violent upper abdominal pain accompanied by nausea and occasionally vomiting ensues, mimicking an acute abdomen. The diagnosis can be established by direct visualization on upper endoscopy, outlining of the worm by contrast radiographic studies, or histopathologic examination of extracted tissue. Extraction of the burrowing larvae during endoscopy is curative. In addition, larvae may pass to the small bowel, where they penetrate the mucosa and provoke a vigorous eosinophilic granulomatous response. Symptoms may appear 1–2 weeks after the infective meal, with intermittent abdominal pain, diarrhea, nausea, and fever resembling the manifestations of Crohn's disease. Ingestion of *Anisakis*-derived proteins t[illegible]aining *Anisakis* parasites can elicit allergic gastrointestinal and even anaphylactic responses.

Loading [Contrib]/a11y/accessibility-menu.js

The diagnosis may be suggested by barium studies and confirmed by curative surgical resection of a granuloma in which the worm is embedded. Anisakid eggs are not found in the stool, since the larvae do not mature in humans. Serologic tests have been developed but are not widely available.

Anisakid larvae in saltwater fish are killed by cooking to 60°C, freezing at 20°C for 3 days, or commercial blast freezing, but usually not by salting, marinating, or cold smoking. No medical treatment is available; surgical or endoscopic removal should be undertaken.

CAPILLARIASIS

Image not available. Intestinal capillariasis is caused by ingestion of raw fish infected with *Capillaria philippinensis*. Subsequent autoinfection can lead to a severe wasting syndrome. The disease occurs in the Philippines and Thailand and, on occasion, elsewhere in Asia. The natural cycle of *C. philippinensis* involves fish from fresh and brackish water. When humans eat infected raw fish, the larvae mature in the intestine into adult worms, which produce invasive larvae that cause intestinal inflammation and villus loss. Capillariasis has an insidious onset with nonspecific abdominal pain and watery diarrhea. If untreated, progressive autoinfection can lead to protein-losing enteropathy, severe malabsorption, and ultimately death from cachexia, cardiac failure, or superinfection. The diagnosis is established by identification of the characteristic peanut-shaped (20- by 40-μm) eggs on stool examination. Severely ill patients require hospitalization and supportive therapy in addition to prolonged anthelmintic treatment with albendazole (200 mg twice daily for 10 days; **Chap. 217**).

ABDOMINAL ANGIOSTRONGYLIASIS

Abdominal angiostrongyliasis is found in Latin America and Africa. The zoonotic parasite *Angiostrongylus costaricensis* causes eosinophilic ileocolitis after the ingestion of contaminated vegetation. *A. costaricensis* normally parasitizes the cotton rat and other rodents, with slugs and snails serving as intermediate hosts. Humans become infected by accidentally ingesting infective larvae in mollusk slime deposited on fruits and vegetables; children are at highest risk. The larvae penetrate the gut wall and migrate to the mesenteric artery, where they develop into adult worms. Eggs deposited in the gut wall provoke an intense eosinophilic granulomatous reaction, and adult worms may cause mesenteric arteritis, thrombosis, or frank bowel infarction. Symptoms may mimic those of appendicitis, including abdominal pain and tenderness, fever, vomiting, and a palpable mass in the right iliac fossa. Leukocytosis and eosinophilia are prominent. CT with contrast medium typically shows inflamed bowel, often with concomitant obstruction, but a definitive diagnosis is usually made surgically with partial bowel resection. Pathologic study reveals a thickened bowel wall with eosinophilic granulomas surrounding the *Angiostrongylus* eggs. In nonsurgical cases, the diagnosis rests solely on clinical grounds because larvae and eggs cannot be detected in the stool. Medical therapy for abdominal angiostrongyliasis is of uncertain efficacy. Careful observation and surgical resection for severe symptoms are the mainstays of treatment.

FURTHER READING

Bethony J et al: Soil-transmitted helminth infections: Ascariasis, trichuriasis, and hookworm. Lancet 367:1521, 2006. [PubMed: 16679166]
Fox LM: Ivermectin: Uses and impact 20 years on. Curr Opin Infect Dis 19:588, 2006. [PubMed: 17075336]
Hochberg NS, Hamer DH: Anisakidosis: Perils of the deep. Clin Infect Dis 51:806, 2010. [PubMed: 20804423]
Horton J: Albendazole: A review of anthelmintic efficacy and safety in humans. Parasitology 1[illegible]86684]
Loukas A et al: Hookworm infection. Nat Rev Dis Primers 2:16088, 2016. [PubMed: 27929101]

Loading [Contrib]/a11y/accessibility-menu.js

Nutman TB: Human infection with *Strongyloides stercoralis* and other related *Strongyloides* species. Parasitology 144:263, 2017.
O'Connell EM, Nutman TB: Molecular diagnostics for soil-transmitted helminths. Am J Trop Med Hyg 95:508, 2016. [PubMed: 27481053]

Loading [Contrib]/a11y/accessibility-menu.js

Chapter 228: Filarial and Related Infections

Thomas B. Nutman; Peter F. Weller

INTRODUCTION

Filarial worms are nematodes that dwell in the subcutaneous tissues and the lymphatics. Eight filarial species infect humans **(Table 228-1)**; of these, four—*Wuchereria bancrofti*, *Brugia malayi*, *Onchocerca volvulus*, and *Loa loa*—are responsible for most serious filarial infections. Filarial parasites, which infect an estimated 170 million persons worldwide, are transmitted by specific species of mosquitoes or other arthropods and have a complex life cycle, including infective larval stages carried by insects and adult worms that reside in either lymphatic or subcutaneous tissues of humans. The offspring of adults are microfilariae, which, depending on their species, are 200–250 μm long and 5–7 μm wide, may or may not be enveloped in a loose sheath, and either circulate in the blood or migrate through the skin (Table 228-1). To complete the life cycle, microfilariae are ingested by the arthropod vector and develop over 1–2 weeks into new infective larvae. Adult worms live for many years, whereas microfilariae survive for 3–36 months. The bacterial endosymbiont *Wolbachia* has been found intracellularly in all stages of *Brugia*, *Wuchereria*, *Mansonella*, and *Onchocerca* species and has become a target for antifilarial chemotherapy.

Loading [Contrib]/a11y/accessibility-menu.js

TABLE 228-1 Characteristics of the Filariae

Organism	Periodicity	Distribution	Vector	Location of Adult	Microfilarial Location	Sheath
Wuchereria bancrofti	Nocturnal	Cosmopolitan areas worldwide, including South America, Africa, southern Asia, Papua New Guinea, China, Indonesia	*Culex*, *Anopheles* (mosquitoes)	Lymphatic tissue	Blood	+
	Subperiodic	Eastern Pacific	*Aedes* (mosquitoes)	Lymphatic tissue	Blood	+
Brugia malayi	Nocturnal	Southeast Asia, Indonesia, India	*Mansonia*, *Anopheles* (mosquitoes)	Lymphatic tissue	Blood	+
	Subperiodic	Indonesia, Southeast Asia	*Coquillettidia*, *Mansonia* (mosquitoes)	Lymphatic tissue	Blood	+
Brugia timori	Nocturnal	Indonesia	*Anopheles* (mosquitoes)	Lymphatic tissue	Blood	+
Loa loa	Diurnal	West and Central Africa	*Chrysops* (deerflies)	Subcutaneous tissue	Blood	+
Onchocerca volvulus	None	South and Central America, Africa	*Simulium* (blackflies)	Subcutaneous tissue	Skin, eye	–
Mansonella ozzardi	None	South and Central America	*Culicoides* (midges)	Undetermined site	Blood	–
	None	Caribbean	*Simulium* (blackflies)	Undetermined site	Blood	–
Mansonella perstans	None	South and Central America, Africa	*Culicoides* (midges)	Body cavities, mesentery, perirenal tissue	Blood	–
Mansonella streptocerca	None	West and Central Africa	*Culicoides* (midges)	Subcutaneous tissue	Skin	–

Usually, infection is established only with repeated, prolonged exposures to infective larvae. Since the clinical manifestations of filarial diseases develop relatively slowly, these infections should be considered to induce chronic infections with possible long-term debilitating effects. In terms of the nature, severity, and timing of clinical manifestations, patients with filarial infections who are native to endemic areas and have lifelong exposure may differ significantly from those who are travelers or who have recently moved to these areas. Characteristically, filarial disease is more acute and intense in newly exposed individuals than in natives of endemic areas.

LYMPHATIC FILARIASIS

Lymphatic filariasis is caused by *W. bancrofti*, *B. malayi*, or *Brugia timori*. The threadlike adult parasites reside in afferent lymphatics or lymph nodes, where they may remain viable for more than two decades.

EPIDEMIOLOGY

Loading [Contrib]/a11y/accessibility-menu.js

Image not available. *W. bancrofti*, the most widely distributed filarial parasite of humans, affects an estimated 110 million people and is found throughout the tropics and subtropics, including Asia and the Pacific Islands, Africa, areas of South America, and the Caribbean basin. Humans are the only definitive host for the parasite. Generally, the subperiodic form is found only in the Pacific Islands; elsewhere, *W. bancrofti* is nocturnally periodic. Nocturnally periodic forms of microfilariae are scarce in peripheral blood by day and increase at night, whereas subperiodic forms are present in peripheral blood at all times and reach maximal levels in the afternoon. Natural vectors for *W. bancrofti* are *Culex* mosquitoes in urban settings and *Anopheles* or *Aedes* mosquitoes in rural areas.

Brugian filariasis due to *B. malayi* occurs primarily in eastern India, Indonesia, Malaysia, and the Philippines. *B. malayi* also has two forms distinguished by the periodicity of microfilaremia. The more common nocturnal form is transmitted in areas of coastal rice fields, while the subperiodic form is found in forests. *B. malayi* naturally infects cats as well as humans. The distribution of *B. timori* is limited to the islands of southeastern Indonesia.

PATHOLOGY

The principal pathologic changes result from inflammatory damage to the lymphatics, which is typically caused by adult worms and not by microfilariae. Adult worms live in afferent lymphatics or sinuses of lymph nodes and cause lymphatic dilation and thickening of the vessel walls. The infiltration of plasma cells, eosinophils, and macrophages in and around the infected vessels, along with endothelial and connective tissue proliferation, leads to tortuosity of the lymphatics and damaged or incompetent lymph valves. Lymphedema and chronic stasis changes with hard or brawny edema develop in the overlying skin. These consequences of filarial infection are due both to the direct effects of the worms and to the host's inflammatory response to the parasite. Inflammatory responses are believed to cause the granulomatous and proliferative processes that precede total lymphatic obstruction. It is thought that the lymphatic vessel remains patent as long as the worm remains viable and that the death of the worm leads to enhanced granulomatous reactions and fibrosis. Lymphatic obstruction results, and, despite collateralization, lymphatic function is compromised.

CLINICAL FEATURES

The most common presentations of the lymphatic filariases are asymptomatic (or subclinical) microfilaremia, hydrocele **(Fig. 228-1)**, acute adenolymphangitis (ADL), and chronic lymphatic disease. In areas where *W. bancrofti* or *B. malayi* is endemic, the overwhelming majority of infected individuals have few overt clinical manifestations of filarial infection despite the presence of circulating microfilariae in the peripheral blood. Although they may be clinically asymptomatic, virtually all persons with *W. bancrofti* or *B. malayi* microfilaremia have some degree of subclinical disease that includes microscopic hematuria and/or proteinuria, dilated (and tortuous) lymphatics (visualized by imaging), and—in men with *W. bancrofti* infection—scrotal lymphangiectasia (detectable by ultrasound). Despite these findings, the majority of individuals appear to remain clinically asymptomatic for years; in relatively few does the infection progress to either acute or chronic disease.

FIGURE 228-1

Hydrocele associated with *Wuchereria bancrofti* infection.

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ADL is characterized by high fever, lymphatic inflammation (lymphangitis and lymphadenitis), and transient local edema. The lymphangitis is retrograde, extending peripherally from the lymph node draining the area where the adult parasites reside. Regional lymph nodes are often enlarged, and the entire lymphatic channel can become indurated and inflamed. Concomitant local thrombophlebitis can occur as well. In brugian filariasis, a single local abscess may form along the involved lymphatic tract and subsequently rupture to the surface. The lymphadenitis and lymphangitis can involve both the upper and lower extremities in both bancroftian and brugian filariasis, but involvement of the genital lymphatics occurs almost exclusively with *W. bancrofti* infection. This genital involvement can be manifested by funiculitis, epididymitis, and scrotal pain and tenderness. In endemic areas, another type of acute disease—dermatolymphangioadenitis (DLA)—is recognized as a syndrome that includes high fever, chills, myalgias, and headache. Edematous inflammatory plaques clearly demarcated from normal skin are seen. Vesicles, ulcers, and hyperpigmentation also may be noted. There is often a history of trauma, burns, irradiation, insect bites, punctiform lesions, or chemical injury. Entry lesions, especially in the interdigital area, are common. DLA is often diagnosed as cellulitis.

If lymphatic damage progresses, transient lymphedema can develop into lymphatic obstruction and the permanent changes associated with elephantiasis **(Fig. 228-2)**. Brawny edema follows early pitting edema, the subcutaneous tissues thicken, and hyperkeratosis occurs. Fissuring of the skin develops, as do hyperplastic changes. Superinfection of these poorly vascularized tissues becomes a problem. In bancroftian filariasis, in which genital involvement is common, hydroceles may develop (Fig. 228-1);

Loading [Contrib]/a11y/accessibility-menu.js

in advanced stages, this condition may evolve into scrotal lymphedema and scrotal elephantiasis. Furthermore, if there is obstruction of the retroperitoneal lymphatics, increased renal lymphatic pressure leads to rupture of the renal lymphatics and the development of chyluria, which is usually intermittent and most prominent in the morning.

FIGURE 228-2

Elephantiasis of the lower extremity associated with *Wuchereria bancrofti* infection.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The clinical manifestations of filarial infections in travelers or transmigrants who have recently entered an endemic region are distinctive. Given a sufficient number of bites by infected vectors, usually over a 3- to 6-month period, recently exposed patients can develop acute lymphatic or scrotal inflammation with or without urticaria and localized angioedema. Lymphadenitis of epitrochlear, axillary, femoral, or inguinal lymph nodes is often followed by evolving retrograde lymphangitis. Acute attacks are short-lived and are not usually accompanied by fever. With prolonged exposure to infected mosquitoes, these attacks, if untreated, become more severe and lead to permanent lymphatic inflammation and

Loading [Contrib]/a11y/accessibility-menu.js

DIAGNOSIS

A definitive diagnosis can be made only by detection of the parasites and hence can be difficult. Adult worms localized in lymphatic vessels or nodes are largely inaccessible. Microfilariae can be found in blood, in hydrocele fluid, or (occasionally) in other body fluids. Such fluids can be examined microscopically, either directly or—for greater sensitivity—after concentration of the parasites by the passage of fluid through a polycarbonate cylindrical-pore filter (pore size, 3 μm) or by the centrifugation of fluid fixed in 2% formalin (Knott's concentration technique). The timing of blood collection is critical and should be based on the periodicity of the microfilariae in the endemic region involved. Many infected individuals do not have microfilaremia, and definitive diagnosis in such cases can be difficult. Assays for circulating antigens of *W. bancrofti* permit the diagnosis of microfilaremic and cryptic (amicrofilaremic) infection. Two tests are commercially available: an enzyme-linked immunosorbent assay and a rapid-format immunochromatographic card test. Both assays have sensitivities of 93–100% and specificities approaching 100%. There are currently no tests for circulating antigens in brugian filariasis.

Polymerase chain reaction (PCR)–based assays for DNA of *W. bancrofti* and *B. malayi* in blood have been developed. A number of studies indicate that the sensitivity of this diagnostic method is equivalent to or greater than that of parasitologic methods.

In cases of suspected lymphatic filariasis, examination of the scrotum, the lymph nodes, or (in female patients) the breast by means of high-frequency ultrasound in conjunction with Doppler techniques may result in the identification of motile adult worms within dilated lymphatics. Worms may be visualized in the lymphatics of the spermatic cord in up to 80% of men infected with *W. bancrofti*. Live adult worms have a distinctive pattern of movement within the lymphatic vessels (termed the *filarial dance sign*). Radionuclide lymphoscintigraphic imaging of the limbs reliably demonstrates widespread lymphatic abnormalities in both subclinical microfilaremic persons and those with clinical manifestations of lymphatic pathology. Although of potential utility in the delineation of anatomic changes associated with infection, lymphoscintigraphy is unlikely to assume primacy in the diagnostic evaluation of individuals with suspected infection; it is principally a research tool, although it has been used more widely for assessment of lymphedema of any cause. Eosinophilia and elevated serum concentrations of IgE and antifilarial antibody support the diagnosis of lymphatic filariasis. There is, however, extensive cross-reactivity between filarial antigens and antigens of other helminths. Of note, *W. bancrofti*– and *B. malayi*–specific antigens have been identified and are now available for use in rapid diagnostic tests with specificities of >98%. However, seropositivity cannot be equated with active infection: residents of endemic areas can become sensitized to filarial antigens through exposure to infective mosquitoes without having patent filarial infections.

The ADL associated with lymphatic filariasis must be distinguished from thrombophlebitis, infection, and trauma. Retrograde evolution is a characteristic feature that helps distinguish filarial lymphangitis from ascending bacterial lymphangitis. Chronic filarial lymphedema must also be distinguished from the lymphedema of malignancy, postoperative scarring, trauma, chronic edematous states, and congenital lymphatic system abnormalities.

TREATMENT

TREATMENT Lymphatic Filariasis

With newer definitions of clinical syndromes in lymphatic filariasis and new tools to assess clinical status (e.g., ultrasound, lymphoscintigraphy, circulating filarial antigen assays, PCR), approaches to treatment based on infection status can be considered.

Diethylcarbamazine (DEC; 6 mg/kg daily for 12 days), which has both macro- and microfilaricidal properties, remains the drug of choice for the treatment of active lymphatic filariasis

Loading [Contrib]/a11y/accessibility-menu.js

(defined by microfilaremia, antigen positivity, or adult worms on ultrasound), although albendazole (400 mg twice daily by mouth for 21 days) also has demonstrated macrofilaricidal efficacy. A 4- to 6-week course of oral doxycycline (targeting the intracellular *Wolbachia*) also has significant macrofilaricidal activity, as does DEC/albendazole used daily for 7 days. The addition of DEC to a 3-week course of doxycycline is efficacious in lymphatic filariasis.

Regimens that combine single doses of albendazole (400 mg) with either DEC (6 mg/kg) or ivermectin (200 µg/kg) all have a sustained microfilaricidal effect and are the mainstay of programs for the eradication of lymphatic filariasis in Africa (albendazole/ivermectin) and elsewhere (albendazole/DEC) (see "Prevention and Control," below). Recently, a regimen using single doses of the three major antifilarial drugs (albendazole/DEC/ivermectin) has been shown to sustain microfilarial clearance out to at least 2 years.

As has already been mentioned, a growing body of evidence indicates that, although they may be asymptomatic, virtually all persons with *W. bancrofti* or *B. malayi* microfilaremia have some degree of subclinical disease (hematuria, proteinuria, abnormalities on lymphoscintigraphy). Thus, early treatment of asymptomatic persons who have microfilaremia is recommended to prevent further lymphatic damage. For ADL, supportive treatment (including the administration of antipyretics and analgesics) is recommended, as is antibiotic therapy if secondary bacterial infection is likely. Similarly, because lymphatic disease is associated with the presence of adult worms, treatment with DEC is recommended for microfilaria-negative carriers of adult worms.

In persons with chronic manifestations of lymphatic filariasis, treatment regimens that emphasize hygiene, prevention of secondary bacterial infections, and physiotherapy have gained wide acceptance for morbidity control. These regimens are similar to those recommended for lymphedema of most nonfilarial causes and are known by a variety of names, including *complex decongestive physiotherapy* and *complex lymphedema therapy*. Hydroceles (Fig. 228-1) can be managed surgically. With chronic manifestations of lymphatic filariasis, drug treatment should be reserved for individuals who have evidence of active infection; however, a 6-week course of doxycycline has been shown to provide improvement in filarial lymphedema irrespective of disease activity.

Side effects of DEC treatment include fever, chills, arthralgias, headaches, nausea, and vomiting. Both the development and the severity of these reactions are directly related to the number of microfilariae circulating in the bloodstream. The adverse reactions may represent either an acute hypersensitivity reaction to the antigens being released by dead and dying parasites or an inflammatory reaction induced by the intracellular *Wolbachia* endosymbionts freed from their intracellular niche.

Ivermectin has a side effect profile similar to that of DEC when used in lymphatic filariasis. In patients infected with *L. loa* who have high levels of microfilaremia, DEC—like ivermectin (see "Loiasis," below)—can elicit severe encephalopathic complications. When used in single-dose regimens for the treatment of lymphatic filariasis, albendazole is associated with relatively few side effects.

PREVENTION AND CONTROL

To protect themselves against filarial infection, individuals must avoid contact with infected mosquitoes by using personal protective measures, including bed nets, particularly those impregnated with insecticides such as permethrin. Mass drug administration (MDA) is the current approach to elimination of lymphatic filariasis as a public health problem. The underlying tenet of this approach is that mass annual distribution of antifilarial chemotherapy—albendazole with either DEC (for all areas except those where onchocerciasis is coendemic; see section on onchocerciasis treatment, below) or ivermectin or with both ivermectin and DEC (triple-drug therapy) will profoundly suppress microfilaremia. If the suppression is sustained, then transmission can be interrupted.

Loading [Contrib]/a11y/accessibility-menu.js

Image not available. Created by the World Health Organization in 1997, the Global Programme to Eliminate Lymphatic Filariasis is based on mass administration of single annual doses of DEC plus albendazole in non-African regions and of albendazole plus ivermectin in Africa. Available information from late 2013 indicated that more than 792 million persons in 53 countries had thus far participated. Not only has lymphatic filariasis been eliminated in some defined areas, but collateral benefits—avoidance of disability and treatment of intestinal helminths and other conditions (e.g., scabies and louse infestation)—also have been noted. The strategy of the global program is being refined, and attempts are being made to integrate this effort with other mass-treatment strategies (e.g., deworming programs, malaria control, and trachoma control) in an integrated control strategy.

TROPICAL PULMONARY EOSINOPHILIA

Image not available. Tropical pulmonary eosinophilia (TPE) is a distinct syndrome that develops in some individuals infected with the lymphatic-dwelling filarial species. The majority of cases have been reported from India, Pakistan, Sri Lanka, Brazil, Guyana, and Southeast Asia; the decreasing incidence of TPE in last decade probably reflects global MDA efforts.

CLINICAL FEATURES

The main features include a history of residence in filaria-endemic regions, paroxysmal cough and wheezing (usually nocturnal and probably related to the nocturnal periodicity of microfilariae), weight loss, low-grade fever, lymphadenopathy, and pronounced blood eosinophilia (>3000 eosinophils/μL). Chest x-rays or CT scans may be normal, but generally show increased bronchovascular markings. Diffuse miliary lesions or mottled opacities may be present in the middle and lower lung fields. Tests of pulmonary function show restrictive abnormalities in most cases and obstructive defects in half. Characteristically, total serum IgE levels (4–40 KIU/mL) and antifilarial antibody levels are markedly elevated.

PATHOLOGY

In TPE, microfilariae and parasite antigens are rapidly cleared from the bloodstream by the lungs. The clinical symptoms result from allergic and inflammatory reactions elicited by the cleared parasites. In some patients, trapping of microfilariae in other reticuloendothelial organs can cause hepatomegaly, splenomegaly, or lymphadenopathy. A prominent, eosinophil-enriched, intra-alveolar infiltrate is common, and with it comes the release of cytotoxic proinflammatory eosinophil granule proteins that may mediate some of the pathology seen in TPE. In the absence of successful treatment, interstitial fibrosis can lead to progressive pulmonary damage.

DIFFERENTIAL DIAGNOSIS

TPE must be distinguished from asthma, Löffler syndrome, allergic bronchopulmonary aspergillosis, allergic granulomatosis with polyangiitis (EGPA or Churg-Strauss syndrome), other systemic vasculitides (most notably, periarteritis nodosa), chronic eosinophilic pneumonia, and the hypereosinophilic syndromes (HESs).

TREATMENT

TREATMENT Tropical Pulmonary Eosinophilia

DEC is used at a daily dosage of 4–6 mg/kg for 14 days. Symptoms usually resolve within 3–7 days a[illegible] which occurs in ~12–25% of cases (sometimes after an interval of several years), requires re-treatment.

Loading [Contrib]/a11y/accessibility-menu.js

ONCHOCERCIASIS

EPIDEMIOLOGY

Image not available. Onchocerciasis (“river blindness”) is caused by the filarial nematode *O. volvulus*, which infects an estimated 37 million individuals in 31 countries worldwide. The majority of individuals infected with *O. volvulus* live in the equatorial region of Africa extending from the Atlantic coast to the Red Sea. In the Americas, the only remaining countries with isolated foci are Venezuela and Brazil. The infection is also found in Yemen.

ETIOLOGY

Infection in humans begins with the deposition of infective larvae on the skin by the bite of an infected blackfly. The larvae develop into adults, which are typically found in subcutaneous nodules. About 7 months to 3 years after infection, the gravid female releases microfilariae that migrate out of the nodule and throughout the tissues, concentrating in the dermis. Infection is transmitted to other persons when a female fly ingests microfilariae from the host’s skin and these microfilariae then develop into infective larvae. Adult *O. volvulus* females and males are ~40–60 cm and ~3–6 cm in length, respectively. The life span of adults can be as long as 18 years, with an average of ~9 years. Because the blackfly vector breeds along free-flowing rivers and streams (particularly in rapids) and generally restricts its flight to an area within several kilometers of these breeding sites, both biting and disease transmission are most intense in these locations.

PATHOLOGY

Onchocerciasis primarily affects the skin, eyes, and lymph nodes. In contrast to the pathology in lymphatic filariasis, the damage in onchocerciasis is elicited by microfilariae and not by adult parasites. In the skin, there are mild but chronic inflammatory changes that can result in loss of elastic fibers, atrophy, and fibrosis. The subcutaneous nodules (*onchocercomata*) consist primarily of fibrous tissues surrounding the adult worm, often with a peripheral ring of inflammatory cells surrounded by an endothelial layer (characterized as lymphatic in origin). In the eye, neovascularization and corneal scarring lead to corneal opacities and blindness. Inflammation in the anterior and posterior chambers frequently results in anterior uveitis, chorioretinitis, and optic atrophy. Although punctate opacities are due to an inflammatory reaction surrounding dead or dying microfilariae, the pathogenesis of most manifestations of onchocerciasis is still unclear.

CLINICAL FEATURES

Skin

Pruritus and rash are the most common manifestations of onchocerciasis. The pruritus can be incapacitating; the rash is typically a papular eruption **(Fig. 228-3)** that is generalized rather than localized to a particular region of the body. Long-term infection results in exaggerated and premature wrinkling of the skin, loss of elastic fibers, and epidermal atrophy that can lead to loose, redundant skin and hypo- or hyperpigmentation. Localized eczematoid dermatitis can cause hyperkeratosis, scaling, and pigmentary changes. In an immunologically hyperreactive form of onchodermatitis (commonly termed *sowdah* or *localized onchodermatitis*), the affected skin darkens as a consequence of the profound inflammation that occurs as microfilariae in the skin are cleared.

FIGURE 228-3

Loading [Contrib]/a11y/accessibility-menu.js

Papular eruption as a consequence of onchocerciasis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Onchocercomata

These subcutaneous nodules, which can be palpable and/or visible, contain the adult worm. They are most common over the coccyx and sacrum, the trochanter of the femur, the lateral anterior crest, and other bony prominences. Nodules vary in size and characteristically are firm and not tender. It has been estimated that, for every palpable nodule, there are four deeper nonpalpable ones.

Ocular Tissue

Visual impairment is the most serious complication of onchocerciasis and usually affects only those persons with moderate or heavy infections. Lesions may develop in all parts of the eye. The most common early finding is conjunctivitis with photophobia. Punctate keratitis—acute inflammatory reactions surrounding dying microfilariae and manifested as “snowflake” opacities—is common among younger patients and resolves without apparent complications. Sclerosing keratitis occurs in 1–5% of infected persons and is the leading cause of onchocercal blindness. Anterior uveitis and iridocyclitis develop in ~5% of infected persons. Characteristic chorioretinal lesions develop as a result of atrophy and hyperpigmentation of the retinal pigment epithelium. Constriction of the visual fields and overt optic atrophy may occur.

Loading [Contrib]/a11y/accessibility-menu.js

Lymph Nodes

Mild to moderate lymphadenopathy is common, particularly in the inguinal and femoral areas, where the enlarged nodes may hang down in response to gravity (“hanging groin”), sometimes predisposing to inguinal and femoral hernias.

Other Manifestations

Some heavily infected individuals develop cachexia with loss of adipose tissue and muscle mass. A form of dwarfism, Nakalanga dwarfism, has been attributed to pituitary involvement in this infection. An association between onchocerciasis and epilepsy (including an epidemic form termed nodding syndrome) has gained attention recently. Among adults who become blind, there is a three- to fourfold increase in mortality rate.

DIAGNOSIS

Definitive diagnosis depends on the detection of an adult worm in an excised nodule or, more commonly, of microfilariae in a skin snip. Skin snips are obtained with a corneal-scleral punch or by lifting of the skin with the tip of a needle and excision of a small (1- to 3-mm) piece with a sterile scalpel blade. Both methods collect a blood-free skin biopsy sample extending to just below the epidermis. The biopsy tissue can be incubated in tissue culture medium or in saline on a glass slide or flat-bottomed microtiter plate. After incubation for 2–4 h (or occasionally overnight in light infections), microfilariae emergent from the skin can be seen by low-power microscopy or can be detected by PCR.

Eosinophilia and elevated serum IgE levels are common but, because these features are seen in many parasitic infections, are not diagnostic in themselves. Immunoassays to detect antibodies to *Onchocerca*-specific antigens are being used both in specialized laboratories and at the point of contact in rapid-diagnostic formats.

TREATMENT

TREATMENT Onchocerciasis

The main goals of therapy are to prevent the development of irreversible lesions and to alleviate symptoms. Chemotherapy is the mainstay of management. Ivermectin, a semisynthetic macrocyclic lactone active against microfilariae, is the first-line agent for the treatment of onchocerciasis. It is given orally in a single dose of 150 μg/kg, either yearly or semiannually. More frequent ivermectin administration (every 3 months) has been suggested to ameliorate pruritus and skin disease.

Image not available. After treatment, most individuals have few or no reactions. Pruritus, cutaneous edema, and/or maculopapular rash occurs in ~1–10% of treated individuals. In areas of Africa coendemic for *O. volvulus* and *L. loa*, however, ivermectin is contraindicated (as it is for pregnant or breast-feeding women) because of severe post-treatment encephalopathy, especially in patients who are heavily microfilaremic for *L. loa* (>30,000 microfilariae/mL). Although ivermectin treatment results in a marked drop in microfilarial density, its effect can be short-lived (<3 months in some cases). Thus, it is occasionally necessary to give ivermectin more frequently for persistent symptoms.

A 6-week course of doxycycline is macrofilaristatic, rendering female adult worms sterile for long periods.

PREVENTION

Loading [Contrib]/a11y/accessibility-menu.js

Vector control has been beneficial in highly endemic areas in which breeding sites are vulnerable to insecticide spraying, but most areas endemic for onchocerciasis are not suited to this type of control. Community-based administration of ivermectin every 6–12 months is being used to interrupt transmission in endemic areas. This measure, in conjunction with vector control, has already helped eliminate the infection in most of Latin America and has reduced the prevalence of disease in many endemic foci in Africa. No drug has proved useful for prophylaxis of *O. volvulus* infection.

LOIASIS

ETIOLOGY AND EPIDEMIOLOGY

Loiasis is caused by *L. loa* (the African eye worm), which is present in the rainforests of West and Central Africa. Adult parasites (females, 50–70 mm long and 0.5 mm wide; males, 25–35 mm long and 0.25 mm wide) live in subcutaneous tissues. Microfilariae circulate in the blood with a diurnal periodicity that peaks between 10:00 A.M. and 2:00 P.M.

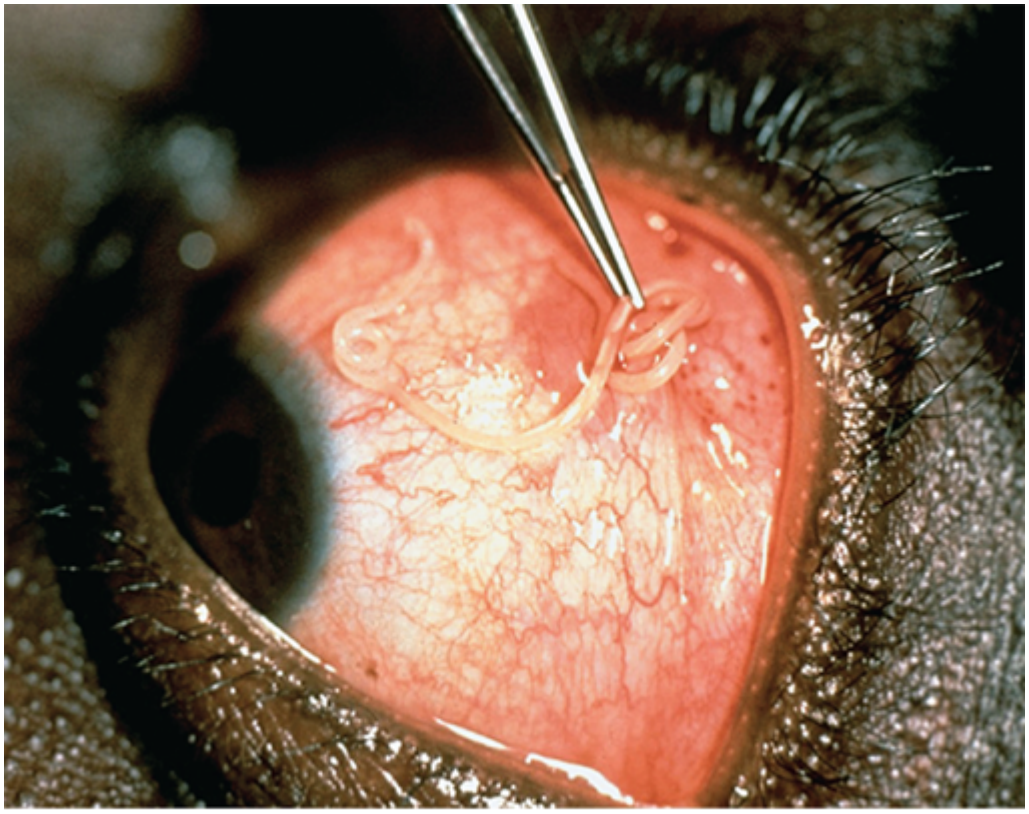
CLINICAL FEATURES

Manifestations of loiasis in natives of endemic areas may differ from those in temporary residents or visitors. Among the indigenous population, loiasis is often an asymptomatic infection with microfilaremia. Infection may be recognized only after subconjunctival migration of an adult worm **(Fig. 228-4)** or may be manifested by episodic *Calabar swellings*—evanescent localized areas of angioedema and erythema developing on the extremities and less frequently at other sites. Nephropathy, encephalopathy, and cardiomyopathy can occur but are rare. In patients who are not residents of endemic areas, allergic symptoms predominate, episodes of Calabar swelling tend to be more frequent, microfilaremia is less common, and eosinophilia and increased levels of antifilarial antibodies are characteristic.

FIGURE 228-4

Adult *Loa loa* worm being surgically removed after its subconjunctival migration.

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOLOGY

The pathogenesis of the manifestations of loiasis is poorly understood. Calabar swellings are thought to result from a hypersensitivity reaction to adult worm antigens.

DIAGNOSIS

Definitive diagnosis of loiasis requires the detection of microfilariae in the peripheral blood or the isolation of the adult worm from the eye (Fig. 228-4) or from a subcutaneous biopsy specimen collected from a site of swelling developing after treatment. PCR-based assays for the detection of *L. loa* DNA in blood are available in specialized laboratories and are highly sensitive and specific, as are some newer recombinant antigen–based serologic techniques. In practice, the diagnosis must often be based on a characteristic history and clinical presentation, blood eosinophilia, and elevated levels of antifilarial antibodies, particularly in travelers to an endemic region, who are often amicrofilaremic.

TREATMENT

TREATMENT Loiasis

DEC (8–10 mg/kg per day administered orally for 21 days) is effective against both the adult and the microfilarial forms of *L. loa*, but multiple courses are frequently necessary before loiasis resolves completely. In cases of heavy microfilaremia, allergic or other inflammatory reactions can take place during treatment, including central nervous system involvement with coma and encephalitis. Heavy infections can be treated initially with apheresis to remove the microfilariae and with glucocorticoids (40–60 mg of prednisone per day) followed by doses of DEC (0.5 mg/kg per day). If antifilarial

Loading [Contrib]/a11y/accessibility-menu.js

treatment has no adverse effects, the prednisone dose can be rapidly tapered and the dose of DEC gradually increased to 8–10 mg/kg per day.

Albendazole or ivermectin is effective in reducing microfilarial loads, although neither is approved for this purpose by the U.S. Food and Drug Administration. Moreover, ivermectin is contraindicated in patients with >30,000 microfilariae/mL because this drug has been associated with severe adverse events (including encephalopathy and death) in heavily infected patients with loiasis in West and Central Africa. DEC (300 mg weekly) is an effective prophylactic regimen for loiasis.

STREPTOCERCIASIS

Image not available. *Mansonella streptocerca*, found mainly in the tropical forest belt of Africa from Ghana to the Democratic Republic of the Congo, is transmitted by biting midges. The major clinical manifestations involve the skin and include pruritus, papular rashes, and pigmentation changes. Many infected individuals have inguinal adenopathy, although most are asymptomatic. The diagnosis is made by detection of the characteristic microfilariae in skin snips. Ivermectin at a single dose of 150 µg/kg leads to sustained suppression of microfilariae in the skin and is probably the treatment of choice for streptocerciasis.

MANSONELLA PERSTANS INFECTION

Image not available. *M. perstans*, distributed across the center of Africa and in northeastern South America, is transmitted by midges. Adult worms reside in serous cavities—pericardial, pleural, and peritoneal—as well as in the mesentery and the perirenal and retroperitoneal tissues. Microfilariae circulate in the blood without periodicity. The clinical and pathologic features of the infection are poorly defined. Most patients appear to be asymptomatic, but manifestations may include transient angioedema and pruritus of the arms, face, or other parts of the body (analogous to the Calabar swellings of loiasis); fever; headache; arthralgias; and right-upper-quadrant pain. Occasionally, pericarditis and hepatitis occur. The diagnosis is based on the demonstration of microfilariae in blood or serosal effusions. Perstans filariasis is often associated with peripheral-blood eosinophilia and antifilarial antibody elevations.

With the identification of a *Wolbachia* endosymbiont in *M. perstans*, doxycycline (200 mg twice a day) for 6 weeks has been established as the first effective treatment for this infection.

MANSONELLA OZZARDI INFECTION

Image not available. The distribution of *M. ozzardi* is restricted to Central and South America and certain Caribbean islands. Adult worms are rarely recovered from humans. Microfilariae circulate in the blood without periodicity. Although this organism has often been considered nonpathogenic, headache, articular pain, fever, pulmonary symptoms, adenopathy, hepatomegaly, pruritus, and eosinophilia have been ascribed to *M. ozzardi* infection. The diagnosis is made by detection of microfilariae in peripheral blood. Ivermectin is effective in treating this infection.

ZOONOTIC FILARIAL INFECTIONS

Dirofilariae that affect primarily dogs, cats, and raccoons occasionally infect humans incidentally, as do *Brugia* and *Onchocerca* parasites that affect small mammals. Because humans are an abnormal host, the parasites never develop fully. Pulmonary dirofilarial infection caused by the canine heartworm *Dirofilaria immitis* generally presents in humans as a solitary pulmonary nodule. Chest pain, hemoptysis, and cough are uncommon. Infections with *D. repens* (from dogs) or *D. tenuis* (from

Loading [Contrib]/a11y/accessibility-menu.js

raccoons) can cause local subcutaneous nodules in humans. Zoonotic *Brugia* infection can produce isolated lymph node enlargement, whereas zoonotic *Onchocerca* species (particularly *O. lupi*) can cause subconjunctival masses. Eosinophilia levels and antifilarial antibody titers are not commonly elevated. Excisional biopsy is both diagnostic and curative. These infections usually do not respond to antifilarial chemotherapy.

DRACUNCULIASIS (GUINEA WORM INFECTION)

ETIOLOGY AND EPIDEMIOLOGY

Image not available. The incidence of dracunculiasis, caused by *Dracunculus medinensis*, has declined dramatically because of global eradication efforts. In 2017, only 30 cases worldwide were identified. The infection appears to be endemic only in Chad and Ethiopia.

Humans acquire *D. medinensis* when they ingest water containing infective larvae derived from *Cyclops*, a crustacean that is the intermediate host. Larvae penetrate the stomach or intestinal wall, mate, and mature. The adult male probably dies; the female worm develops over a year and migrates to subcutaneous tissues, usually in the lower extremity. As the thin female worm, ranging in length from 30 cm to 1 m, approaches the skin, a blister forms that, over days, breaks down and forms an ulcer. When the blister opens, large numbers of motile, rhabditiform larvae can be released into stagnant water; ingestion by *Cyclops* completes the life cycle.

CLINICAL FEATURES

Few or no clinical manifestations of dracunculiasis are evident until just before the blister forms, when there is an onset of fever and generalized allergic symptoms, including periorbital edema, wheezing, and urticaria. The emergence of the worm is associated with local pain and swelling. When the blister ruptures (usually as a result of immersion in water) and the adult worm releases larva-rich fluid, symptoms are relieved. The shallow ulcer surrounding the emerging adult worm heals over weeks to months. Such ulcers, however, can become secondarily infected, the result being cellulitis, local inflammation, abscess formation, or (uncommonly) tetanus. Occasionally, the adult worm does not emerge but becomes encapsulated and calcified.

DIAGNOSIS

The diagnosis is based on the findings developing with the emergence of the adult worm, as described above.

TREATMENT

TREATMENT Dracunculiasis

Gradual extraction of the worm by winding of a few centimeters on a stick each day remains the common and effective practice. Worms may be excised surgically. No drug is effective in treating dracunculiasis.

PREVENTION

Prevention, which remains the only real control measure, depends on the provision of safe drinking water.

Loading [Contrib]/a11y/accessibility-menu.js

FURTHER READING

Herrick JA et al: Eosinophil-associated processes underlie differences in clinical presentation of loiasis between temporary residents and those indigenous to *Loa*-endemic areas. Clin Infect Dis 60:55, 2015. [PubMed: 25234520]
Hopkins DR et al: Progress toward global eradication of dracunculiasis—January 2016–June 2017. Morb Mortal Wkly Rep 66:1327, 2017.
Mand S et al: Doxycycline improves filarial lymphedema independent of active filarial infection: A randomized controlled trial. Clin Infect Dis 55:621, 2012. [PubMed: 22610930]
Ramaiah KD, Ottesen EA: Progress and impact of 13 years of the Global Programme to Eliminate Lymphatic Filariasis on reducing the burden of filarial disease. PLoS Negl Trop Dis 8:e3319, 2014. [PubMed: 25412180]
Steel C et al: Rapid point-of-contact tool for mapping and integrated surveillance of *Wuchereria bancrofti* and *Onchocerca volvulus* infection. Clin Vaccine Immunol 22:896, 2015. [PubMed: 26018537]
Taylor MJ et al: Lymphatic filariasis and onchocerciasis. Lancet 376:1175, 2010. [PubMed: 20739055]

Chapter 229: Schistosomiasis and Other Trematode Infections

Birgitte Jyding Vennervald

INTRODUCTION

Trematodes, or flatworms, are a group of helminths that belong to the phylum Platyhelminthes. The adult flatworms share some common characteristics, such as macroscopic size (from one to several centimeters); dorsoventrally flattened, bilaterally symmetric bodies; and two suckers—oral and ventral. Except for schistosomes, which have separate sexes, all human parasitic trematodes are hermaphroditic. Their life cycles involve a mammalian/human definitive host, in which sexual reproduction by adult worms takes place, and an intermediate host (snails), in which asexual multiplication occurs. Some species of trematodes have more than one intermediate host.

Humans are infected either by direct penetration of intact skin (schistosomiasis) or by ingestion of raw freshwater fish, crustaceans, or aquatic plants with metacercariae—the infective larval stage.

Significant trematode infections of humans may be divided according to the location of the adult worms: blood, liver (biliary tree), intestines, or lungs **(Table 229-1)**. Adult worms do not multiply within the mammalian host but can live for up to 30 years. Infections are often chronic.

TABLE 229–1 Major Human Trematode Infections

Trematode	Transmission Route	Geographic Distribution
Blood Flukes		
Intestinal schistosomiasis		
Schistosoma mansoni	Skin penetration by cercariae released from snails (*Biomphalaria* spp.)	Africa, Brazil, Venezuela, Surinam, the Caribbean (low risk)
Shistosoma japonicum	Skin penetration by cercariae released from snails (*Oncomelania* spp.)	China, Indonesia, Philippines
Schistosoma guineensis and *Schistosoma intercalatum*	Skin penetration by cercariae released from snails (*Bulinus* spp.)	Rain forest areas of Central Africa
Schistosoma mekongi	Skin penetration by cercariae released from snails (*Neotricula aperta*)	Several districts of Cambodia and Lao People's Democratic Republic (PDR)
Urogenital schistosomiasis		
Schistosoma haematobium	Skin penetration by cercariae released from snails (*Bulinus* spp.)	Africa, Middle East, Corsica (France)
Liver Flukes		
Clonorchis sinensis	Ingestion of metacercariae in freshwater fish	Asia, including Republic of Korea, China, Taiwan, Vietnam
Opisthorchis viverrini	Ingestion of metacercariae in freshwater fish	Northeast Thailand, Lao PDR, Cambodia, Vietnam
Opisthorchis felineus	Ingestion of metacercariae in freshwater fish	Former Soviet Union, Kazakhstan, Ukraine, Turkey
Fasciola hepatica	Ingestion of metacercariae on aquatic plants or in water	Worldwide
Fasciola gigantica	Ingestion of metacercariae on aquatic plants or in water	Africa, Asia
Intestinal Flukes		
Fasciolopsis buski	Ingestion of metacercariae on aquatic plants	Bangladesh, China, India, Indonesia, Lao PDR, Malaysia, Taiwan, Thailand, Vietnam
Echinostoma spp.	Ingestion of freshwater fish, frogs, mussels, snails	China, India, Indonesia, Japan, Malaysia, Russia, Republic of Korea, Philippines, Thailand
Heterophyes heterophyes, several other species	Ingestion of metacercariae in freshwater or brackish-water fish	Egypt, Greece, Islamic Republic of Iran, Italy, Japan, Republic of Korea, Sudan, Tunisia, Turkey
Lung Flukes		
Paragonimus westermani	Ingestion of metacercariae in crayfish or crabs	Tropical and subtropical areas of eastern and southern Asia and sub-Saharan Africa
Paragonimus kellicotti	Ingestion of metacercariae in crayfish or crabs	North America

Image not available. Although it is relatively rare to encounter patients with trematode infections in the United States, many millions of people are infected worldwide. Both schistosomiasis and food-borne trematode infections are poverty-related chronic diseases with high morbidity and a significant public health impact. Various factors may increase the spread of the infections globally. Increasing temperatures may render new areas suitable for the intermediate host snails, and an increase in travel and migration may increase the number of patients with trematode infections—for example, in the United States.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Trematode Infection

In the evaluation of a patient in whom trematode infection is suspected, certain questions are highly relevant and can assist in establishing a diagnosis: Where have you been? If you have traveled, when did you return? What activities have you been involved in (trekking, swimming, whitewater rafting)? What have you been eating (local dishes while traveling; raw, poorly cooked, or pickled freshwater fish or crustaceans)? Definitive diagnosis is based on detection of parasite eggs in stool, urine, sputum, and sometimes tissue samples or on serologic tests. The presence of eosinophilia and a history of travel to endemic areas should raise suspicion of trematode infection. The U.S. Centers for Disease Control and Prevention (CDC) can provide guidance with respect to diagnosis and treatment.

SCHISTOSOMIASIS

Human schistosomiasis is caused by five species of the parasitic genus *Schistosoma*: *S. mansoni, S. japonicum, S. mekongi*, and *S. intercalatum* cause intestinal disease, and *S. haematobium* causes urogenital disease (Table 229-1). The infection may cause considerable intestinal, hepatic, and genitourinary morbidity. Avian schistosomes may penetrate human skin, but they die in subcutaneous tissue, producing only cutaneous manifestations.

ETIOLOGY

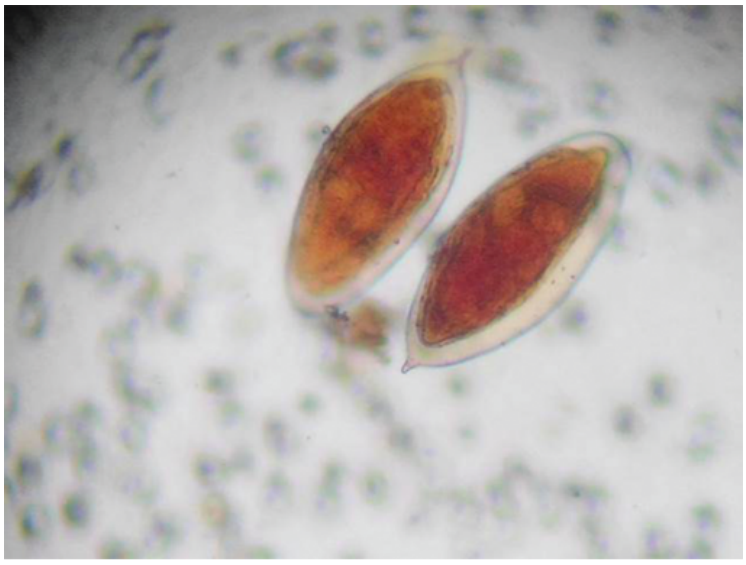
Schistosoma infection is contracted through contact with freshwater bodies harboring infected intermediate-host snails. Cercariae, the infective larval stage released from the snail, penetrate intact human skin within a few minutes after attaching to the skin. After penetration, the cercariae transform to schistosomula, which then enter a small vein or lymphatic vessel, circulate in the bloodstream through the lung capillaries, and are pumped via the heart to all parts of the body to reach the portal vein. There, the worms mature into adult males or females, pair, and migrate to their final location in the mesenteric or pelvic venous plexus.

The interval from cercarial penetration to sexual maturation and egg production, termed the *prepatent period*, lasts 5–7 weeks (up to 12 weeks for *S. haematobium*). The female worm then begins to produce eggs, which are excreted via feces or, for *S. haematobium*, urine. Approximately 50% of eggs are retained in tissue, where they are responsible for organ-specific morbidity (see "Pathogenesis," below). When excreted eggs reach water, they hatch and release a free-swimming larval stage (*miracidium*), which, after penetrating a host snail, undergoes several rounds of asexual multiplication. After ~4–6 weeks, infective cercariae are shed from the infected snails into the water. One snail, infected by one miracidium, can shed thousands of cercariae per day for several months; thus the transmission potential of schistosomes is enormous.

The schistosome egg **(Fig. 229-1)** is the only stage of the parasites' life cycle that can be detected in humans, either in excreta or in tissue biopsies. The eggs are large and can easily be distinguished morphologically from other helminth eggs. *S. haematobium* eggs are ~140 mm long, with a terminal spine; *S. mansoni* eggs are ~150 mm long, with a lateral spine; and *S. japonicum* eggs are smaller, rounder, and ~90 mm long, with a small lateral spine or knob.

Figure 229-1

***Schistosoma haematobium* eggs.**



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Adult schistosomes are ~1–2 cm long. The male worm is flat, and the body forms a groove or gynecophoric canal in which the mature adult female is held like a sausage in a hotdog roll. Females are longer, thinner, and rounded. The females produce hundreds (African species) to thousands (Asian species) of eggs per day. Each ovum contains a ciliated miracidium larva, which secretes proteolytic enzymes that help the eggs to migrate into the lumen of the bladder (*S. haematobium*) or the intestine (other species). The lifespan of an adult schistosome averages

3–5 years but can be as long as 30 years. Schistosome worms feed on red blood cells; the debris is regurgitated in the host's blood, where it can be detected as circulating antigens (see "Diagnosis," below).

Adult schistosomes persist in the bloodstream for years and have evolved strategies of evading attack using immune effector mechanisms. This immune evasion is a result of several processes, such as binding of host proteins to the schistosome surface, which renders the parasite invisible to the host immune system.

Image not available. The genome of schistosomes is relatively large (~270 Mb). Whole-genome sequences are available for *S. mansoni, S. japonicum*, and *S. haematobium*.

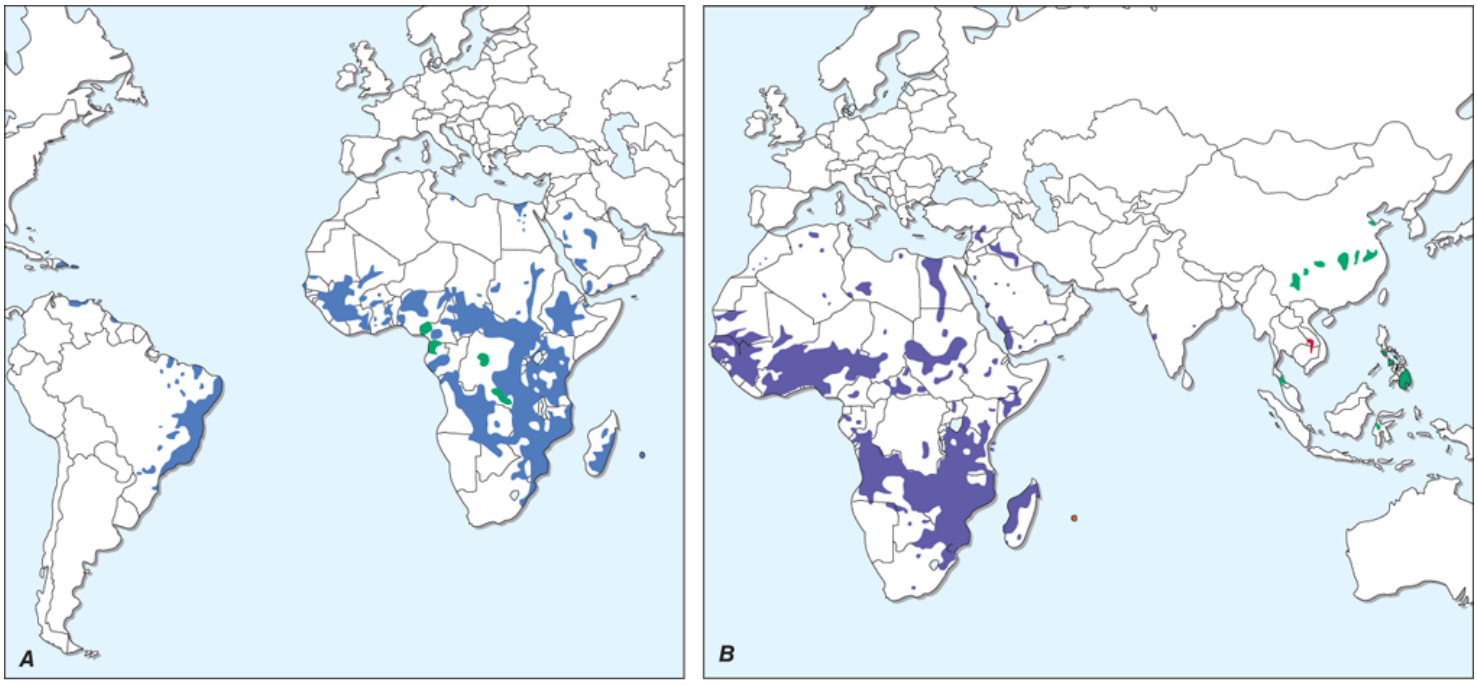
EPIDEMIOLOGY

Image not available. Because of the complex life cycle of schistosomes, with snails as an intermediate host and humans as the final host, transmission is dependent on freshwater habitats that are suitable for the snails, are areas of human activity, and have climatic conditions favoring the survival of the snails and the development of the parasites inside the snail host. These requirements are reflected in the global distribution of schistosomiasis as well as in its microgeographic distribution within an endemic area. For *S. mansoni, S. haematobium*, and *S. intercalatum*, humans are the most important definitive host. *S. japonicum* and *S. mekongi* are zoonotic parasites, with a wide range of definitive hosts such as pigs, water buffaloes, and various rodents.

It is estimated that 230 million people are infected globally, with ~800 million people living in areas where there is a risk of infection **(Fig. 229-2)**. More than 70% of infected people live in sub-Saharan Africa. Schistosomiasis is the most important of the neglected tropical diseases and is second only to malaria in public health impact. It is a poverty-related disease, and infection is prevalent in areas where adequate water supplies and sanitary facilities are lacking. In these areas, people come into contact with infested water through a variety of activities, including bathing, washing clothes, and collecting water for drinking or cooking. In some areas, adults have a high occupational risk of exposure; fishermen, canal cleaners, and workers in rice fields fall into this category. Among children, playing in water and swimming pose a risk. Large-scale irrigation and hydroelectric power operations can create suitable habitats for host snails and thus increase the risk of schistosomiasis transmission.

Figure 229-2

Global distribution of human schistosomiasis. ***A.*** *Schistosoma mansoni* infection (*dark blue*) is endemic in Africa, the Middle East, South America, and a few Caribbean countries. *S. intercalatum* infection (*green*) is endemic in sporadic foci in West and Central Africa. ***B.*** *Schistosoma haematobium* infection (*purple*) is endemic in Africa and the Middle East. The major endemic countries for *S. japonicum* infection (*green*) are China, the Philippines, and Indonesia. *Schistosoma mekongi* infection (*red*) is endemic in sporadic foci in Southeast Asia. *(Reprinted with permission from CH King, AAF Mahmoud: Schistosomiasis and Other Trematode Infections, in DL Kasper et al [eds], Harrison's Principles of Internal Medicine, 19th ed. New York, McGraw-Hill Education, 2015, pp 1423–1429.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In general, children living in endemic areas initially acquire infection at ~3–4 years of age—i.e., when they are old enough to walk and come into contact with infested water. However, infection does occur in much younger children. As children grow older, the prevalence and intensity of infection increase, peaking around puberty. A characteristic feature of schistosomiasis infection in human populations is a convex age–prevalence curve, with low prevalence in very young children, higher prevalence in older children with a peak at 10–15 years of age, and declining prevalence in adults. The same pattern is observed between age and intensity of infection and is attributable to various factors. Generally, children have more frequent, prolonged, and extensive water contact than adults through activities like playing and swimming. Furthermore, several studies have indicated that acquired immunity to schistosomiasis develops slowly over several years, so that adults are

reinfected to a much lesser extent than children. These factors, combined with progressive spontaneous death of adult worms from infections acquired during childhood, lead to lower levels of infection in the adult population.

PATHOGENESIS

Cercarial invasion may be associated with dermatitis arising from dermal and subdermal inflammatory reactions in response to dying cercariae that trigger innate immune responses. However, most manifestations of schistosomiasis—in the acute, established, and chronic phases of infection—are due to immunologic reactions to eggs retained in host tissues.

Around the time when oviposition commences, acute schistosomiasis (Katayama fever) may occur (see "Clinical Features," below). Antigen excess from eggs results in the formation of soluble immune complexes, which may be deposited in several tissues and initiate a serum sickness–like illness. All evidence suggests that schistosome eggs, and not adult worms, induce the organ-specific morbidity caused by schistosome infections. Approximately half of the eggs are not excreted via feces or urine but are trapped in intestinal or hepatic tissue (*S. mansoni, S. japonicum*, and *S. mekongi*) or in the bladder and urogenital system (*S. haematobium*). The eggs induce a granulomatous host immune response composed primarily of lymphocytes, eosinophils, and alternatively activated macrophages. The lymphocytes produce various T_H2 cytokines such as interleukins 4, 5, and 13. Later, in the chronic phase of infection, regulatory cytokines are responsible for immunomodulation or downregulation of host responses to schistosome eggs and play an important role in reducing the size of granulomas.

When *S. mansoni* or *S. japonicum* eggs are swept into the small portal branches of the liver via the portal vein, they lodge in the presinusoidal periportal tissues. The formation of granulomas around the eggs can cause significant enlargement of the spleen and liver. High-intensity infections in children are often accompanied by hepatosplenomegaly that generally decreases over time, partly because the number of eggs being deposited in the tissue gradually declines after the early teenage years as partial immunity to new infections develops and partly because of immunologic downregulation of the granulomatous response. However, in some infected individuals, egg-induced granulomatous responses lead to severe periportal fibrosis (*Symmers clay pipestem fibrosis*), with deposition of collagen around the portal vein, occlusion of the smaller portal branches, and severe, often irreversible, pathology. Occlusion of the portal branches may result in marked portal hypertension.

The signs and symptoms of *S. haematobium* infection relate to the worms' predilection for the veins of the urogenital plexus and result from deposition of eggs in the bladder, ureters, and genital organs. During established active infection, clusters of living eggs in the urogenital tissues can be found surrounded by intense inflammatory reactions and intense tissue eosinophilia. Movement of egg clusters into the lumen of the bladder is often followed by sloughing off of the epithelial surface, ulceration, and bleeding. Intense egg-induced tissue inflammation can result in bladder wall thickening and development of masses and pseudopolyps. Inflammation and granuloma formation around the ureteral ostia can lead to hydronephrosis.

Generally, late chronic-stage infections are characterized by accumulation of dead calcified eggs in tissue. Characteristic cervical lesions are found in *S. haematobium* infections, including active-stage lesions with intense tissue inflammation around live eggs and chronic-stage sandy patches with clusters of calcified eggs.

CLINICAL FEATURES

In general, disease manifestations of schistosomiasis occur in three stages—acute, active, and chronic—according to the duration and intensity of infection.

Cercarial Dermatitis ("Swimmer's Itch")

Cercarial penetration of the skin may result in a maculopapular rash called cercarial dermatitis or "swimmer's itch." Cercarial dermatitis can develop in people who have not previously been exposed to schistosomiasis (e.g., travelers), whereas it is rare among people living in endemic areas. A particularly severe form of cercarial dermatitis is commonly seen after exposure to cercariae from avian schistosomes. These cercariae cannot complete their development in humans and die in the skin, causing an inflammatory allergic reaction. This form of cercarial dermatitis can occur in people who have been in contact with water from lakes (e.g., in Europe or the United States) where various species of water birds, such as ducks, geese, and swans, are found. The rash may last for 1–2 weeks. This condition normally requires no treatment, but systemic antihistamines or topical antihistamines or glucocorticoids can be used to reduce symptoms.

Acute Schistosomiasis (Katayama Fever)

Symptomatic acute schistosomiasis, also known as Katayama fever or Katayama syndrome, is usually seen in travelers who have contracted the infection for the first time. The onset occurs between 2 weeks and 3 months after exposure to the parasite. The symptoms may appear suddenly and include fever, myalgia, general malaise and fatigue, headache, nonproductive cough, and intestinal symptoms such as abdominal tenderness or pain. Various combinations of these symptoms are often accompanied by eosinophilia and transient pulmonary infiltrates. Many patients recover spontaneously from acute schistosomiasis after 2–10 weeks, but the illness follows a more severe clinical course in some individuals, with weight loss, dyspnea, diarrhea, and hepatomegaly. Severe cerebral or spinal cord manifestations may occur, and even light infections may cause severe illness. The syndrome can, in rare cases, be fatal.

Differential diagnosis includes many other febrile infectious diseases with acute onset, including malaria, salmonellosis, and acute hepatitis. Fever and eosinophilia occur in trichinosis, tropical eosinophilia, invasive ankylostomiasis, strongyloidiasis, visceral larva migrans, and infections with *Opisthorchis* and *Clonorchis* species. Katayama fever is rare in people chronically exposed to infection in areas endemic for *S. mansoni* or *S. haematobium*.

Intestinal Schistosomiasis (*S. mansoni, S. japonicum, S. mekongi*)

In intestinal schistosomiasis, adult worms are located in the mesenteric veins, and disease manifestations are associated with parasite eggs passing through or becoming trapped in intestinal tissue. This event induces mucosal granulomatous inflammation with microulcerations, superficial bleeding, and sometimes pseudopolyposis. The symptoms tend to be more pronounced with a high intensity of infection and

include intermittent abdominal pain, loss of appetite, and sometimes bloody diarrhea. The clinical manifestations of *S. intercalatum* and *S. mekongi* infection are generally milder.

Hepatosplenic Schistosomiasis

Hepatosplenic schistosomiasis is caused by schistosome eggs trapped in liver tissue and occurs in *S. mansoni* and *S. japonicum* infections. There are two distinct clinical entities: early inflammatory hepatosplenomegaly and late hepatosplenic disease with periportal fibrosis.

Early inflammatory hepatosplenic schistosomiasis is the main entity seen in children and adolescents. The liver is enlarged, especially the left lobe, and is smooth and firm. The spleen is enlarged, often extending below the umbilicus, and is firm or hard. Generally, ultrasonography shows no hepatic fibrosis. This form of hepatosplenic schistosomiasis may be found in up to 80% of infected children. Its severity is closely associated with the intensity of infection and may also be associated with concomitant chronic exposure to malaria.

Late hepatosplenic schistosomiasis with periportal or Symmers fibrosis may develop in young and middle-aged adults with long-standing, high-level exposure to infection. Patients with periportal fibrosis may excrete very few or no eggs in feces. During the early stage, the liver is enlarged, especially the left lobe; it is smooth and firm or hard. The spleen is enlarged, often massively, and is firm or hard. The patient may report a left hypochondrial mass with discomfort and anorexia. Ultrasonography reveals typical periportal fibrosis and dilation of the portal vein. Other complications include delayed growth and puberty, especially in *S. japonicum* infections, and severe anemia. Severe hepatosplenic schistosomiasis may lead to portal hypertension, but hepatic function usually remains normal, even in cases with marked periportal fibrosis and portal hypertension.

Ascites, attributable both to portal hypertension and to hypoalbuminemia, may be seen, especially in *S. japonicum* infection. Patients with severe hepatosplenic disease and portal hypertension may develop esophageal varices detectable by endoscopy or ultrasound. These patients may experience repeated bouts of hematemesis, melena, or both. Hematemesis is the most severe complication of hepatosplenic schistosomiasis, and death may result from massive loss of blood.

Urogenital Schistosomiasis (*S. haematobium*)

The signs and symptoms of *S. haematobium* infection relate to the worms' predilection for the veins of the urogenital tract. Two stages of infection are recognized. An active stage occurring mainly in children, adolescents, and younger adults is characterized by egg excretion in the urine, with proteinuria and macroscopic or microscopic hematuria and deposition of eggs in the urinary tract. A chronic stage in older individuals is characterized by sparse or no urinary egg excretion despite urogenital tract pathology.

A characteristic sign in the active stage is painless, terminal hematuria. Dysuria and suprapubic discomfort or pain are associated with active urogenital schistosomiasis and may persist throughout the course of active infection. Eggs deposited in the bladder mucosa may give rise to an intense inflammatory response of the bladder wall, which may cause ureteric obstruction and lead to hydroureter and hydronephrosis. These early inflammatory lesions, including obstructive uropathy, can be visualized by ultrasonography.

As the infection progresses, the inflammatory component decreases and fibrosis becomes more prominent. The symptoms at this stage are nocturia, urine retention, dribbling, and incontinence. Cystoscopy reveals "sandy patches" composed of large numbers of calcified eggs surrounded by fibrous tissue and an atrophic mucosal surface. The ureters are less commonly involved, but ureteral fibrosis can cause irreversible obstructive uropathy that can progress to uremia.

Egg deposition may cause granulomas and lesions in the genital organs, most commonly in the cervix and vagina in women and the seminal vessels in men. The results may include dyspareunia, abnormal vaginal discharge, contact bleeding, and lower back pain in women and perineal pain, painful ejaculation, and hematospermia in men. Genital symptoms like bloody discharge and genital itch are associated with *S. haematobium* infection in school-aged girls living in schistosomiasis-endemic areas. Symptoms such as hematospermia and perineal discomfort have been described in travelers, and eggs have been demonstrated in seminal fluid. An association between female genital schistosomiasis and HIV infection has been demonstrated, but the impact of genital schistosomiasis on HIV transmission needs further elucidation.

S. haematobium has been classified by the International Agency for Research on Cancer (IARC) as definitely carcinogenic to humans (i.e., a group 1 carcinogen). Chronic *S. haematobium* infection is associated with squamous cell carcinoma of the urinary bladder.

Other Manifestations

Worms and eggs can sometimes be located in ectopic sites, causing site-specific manifestations and symptoms. Neuroschistosomiasis is one of the most severe clinical forms of schistosomiasis and is caused by the inflammatory response around eggs in the cerebral or spinal venous plexus. *S. mansoni* and *S. haematobium* worms can end up in the spinal venous plexus, where they may cause transverse myelitis—an acute complication sometimes seen in travelers returning home with schistosomiasis. *S. japonicum* is mainly associated with granulomatous lesions in the brain, causing epileptic seizures, encephalopathy with headache, visual impairment, motor deficit, and ataxia. Pulmonary schistosomiasis is caused by portacaval shunting of eggs into the lung capillaries, where they induce granulomas in the perialveolar area. The consequences may be fibrosis, pulmonary hypertension, and cor pulmonale.

DIAGNOSIS

Anamnestic information on recent travels to endemic areas and exposure to freshwater bodies through recreational or other activities is important in the diagnosis of schistosomiasis in travelers. Information about exact geographic locations can facilitate identification of the relevant species of *Schistosoma*. Eosinophilia is a common finding and is often associated with helminthic infections such as schistosomiasis.

Detection of schistosome eggs in stool or urine is indicative of active infection and is the standard diagnostic method. The diagnosis is often based on the detection of eggs in a fixed small amount of excreta—e.g., 50 mg of stool or filtration of 10 mL of urine. This method is widely used among populations in endemic areas and allows quantitation of the level of infection (eggs per gram of feces or per 10 mL of urine).

However, levels of egg excretion in people from nonendemic areas may be very low, in which case a larger sample and concentration methods (e.g., formol-ether concentration) may be needed.

Eggs can also be detected in rectal biopsies (both *S. mansoni* and *S. haematobium*) and occasionally in Pap smears and semen samples (*S. haematobium*). Polymerase chain reaction (PCR)–based detection of parasite DNA in stool or urine is more sensitive than parasitologic methods and is increasingly used. *Schistosoma* DNA can be detected in cerebrospinal fluid samples for diagnosis of neuroschistosomiasis.

Serology, with detection of specific antibodies to schistosomes, is useful in travelers but less so in people from endemic areas where transmission is ongoing. The serologic assays employed at the CDC are a Falcon assay screening test/enzyme-linked immunosorbent assay (FAST-ELISA) using *S. mansoni* adult microsomal antigen and a confirmatory species-specific immunoblot assay performed in light of the patient's travel history.

Schistosome proteoglycans—circulating anodic and cathodic antigens (CAAs and CCAs)—regurgitated into the bloodstream by the feeding worms can be detected in serum and urine by ELISA or monoclonal antibody–based lateral flow assays. The presence of CAA or CCA is an indication of active infection, and levels of these antigens correlate well with the intensity of infection. However, detection of CAAs and CCAs is not currently suitable for diagnosis in travelers, who are likely to have low levels of infection and very few worms. A commercially available point-of-care assay (Rapid Medical Diagnostics, Pretoria, South Africa) that detects CCA in urine is now widely used for screening of infected communities in relation to mass drug administration programs.

TREATMENT

TREATMENT Schistosomiasis

The drug of choice for treatment of schistosomiasis is praziquantel. It is administered orally, is available as 600-mg tablets, and is effective against all schistosome species infecting humans. The drug is safe and well tolerated. Standard regimens are shown in **Table 229-2**. In patients who are not cured by initial treatment, the same dose can be repeated at weekly intervals for 2 weeks. Since praziquantel does not affect the young migrating stages of the schistosomes, it may be necessary to repeat the dose 6–12 weeks later, especially if eosinophilia or symptoms persist despite treatment.

As a general principle, all patients with acute schistosomiasis should be treated with praziquantel. Glucocorticoids can be added in Katayama fever to suppress the hypersensitivity reaction. However, treatment for acute schistosomiasis or Katayama fever must be adjusted appropriately for each case, and in the most severe cases management in an acute-care setting is necessary.

Praziquantel is effective in cerebral *S. japonicum* infections, resulting in rapid dissipation of cerebral edema and resolution of cerebral masses. However, glucocorticoids and anticonvulsants are sometimes needed in neuroschistosomiasis.

The effect of antischistosomal treatment on disease manifestations depends on the stage and severity of the lesions. Early hepatosplenomegaly, mild or moderate fibrosis, and urinary bladder lesions seen during active infection resolve after chemotherapy. However, for late-stage manifestations (e.g., severe fibrosis with portal hypertension), praziquantel treatment is only one component of management, since the main complications are due to obstructive pathology. Management of portal hypertension and prevention of bleeding from esophageal varices should follow clinical guidelines for treatment of these conditions.

TABLE 229–2 Treatment of Schistosomiasis and Food-Borne Trematode Infections

Infection	Drug of Choice	Adult Dose[a]
Schistosoma mansoni, S. haematobium, S. intercalatum, S. guineensis	Praziquantel[b]	40 mg/kg PO in 2 divided doses for 1 day
S. japonicum, S. mekongi	Praziquantel	60 mg/kg PO in 3 divided doses for 1 day
Clonorchis sinensis, Opisthorchis viverrini, Opisthorchis felineus	Praziquantel	25 mg/kg PO tid for 2 consecutive days
Fasciola hepatica, Fasciola gigantica	Triclabendazole[c]	10 mg/kg PO as a single dose[d]
Fasciolopsis buski	Praziquantel	75 mg/kg PO in 3 divided doses for 1 day
Echinostoma spp., *Heterophyes heterophyes*, several other species	Praziquantel	25 mg/kg PO tid
Paragonimus westermani, Paragonimus kellicotti	Praziquantel	25 mg/kg PO tid for 2 consecutive days
	Triclabendazole[c]	10 mg/kg PO once (or twice, 12–24 h apart)

[a]The pediatric dose is the same as the adult dose in all instances. [b]The safety of praziquantel in children <4 years old has not been established, although many children in this age group have been treated with praziquantel during mass drug-administration programs. [c]Triclabendazole is not approved by the U.S. Food and Drug Administration and is not yet commercially available in the United States. It is available through the Centers for Disease Control and Prevention Drug Service (404-639-3670; *drugservice@cdc.gov*). [d]A second dose (10 mg/kg) can be administered 12–24 h after the first dose in severe fascioliasis.

PREVENTION AND CONTROL

Schistosomiasis is contracted through direct contact with infested freshwater. Travelers should be made aware of the risk of infection if they come into contact with freshwater sources in schistosomiasis-endemic areas. For people living in rural areas where schistosomiasis is endemic, it may be very difficult, if not impossible, to avoid water contact—for example, during occupational activities such as fishing and working in rice fields. Schistosomiasis is a poverty-related disease, and access to safe water and good sanitary facilities may rarely be available. Because *S. japonicum* is a zoonotic parasite, preventive measures should target not only the human population but also animals such as water buffalo, which act as reservoirs for infection.

Praziquantel treatment of infected people, often during mass drug-administration programs, is a cornerstone of the management and control of schistosomiasis. Regular treatment will reduce the level of schistosomiasis morbidity in affected populations. However, treatment should be

combined with other relevant strategies, such as control of the intermediate host snails, improved water-quality and sanitation facilities, and health education. Schistosomiasis control measures should be integrated into local health programs.

There have been intensive efforts to develop vaccines, but none is yet available. One vaccine candidate, *S. haematobium* 28GST, has been tested in a clinical phase 3 trial in populations living in an endemic area.

FOOD-BORNE TREMATODE INFECTIONS

Food-borne trematode infections are a group of zoonotic diseases caused by hepatic, intestinal, and pulmonary parasitic flukes. These infections are contracted by ingestion of infective parasites in undercooked aquatic food or water plants. In 2005, an estimated 56.2 million people were infected with food-borne trematodes and 7.9 million had severe sequelae of these infections.

LIVER FLUKES

The most important liver flukes causing human infections are the related species *Opisthorchis viverrini* and *Opisthorchis felineus*, which cause opisthorchiasis; *Clonorchis sinensis*, which causes clonorchiasis; and *Fasciola hepatica* and *Fasciola gigantica*, which cause fascioliasis (Table 229-1).

Opisthorchiasis and Clonorchiasis

Image not available. *O. viverrini* is found mainly in northeastern Thailand, Laos, and Cambodia; *O. felineus* mainly in Europe and Asia, including the former Soviet Union; and *C. sinensis* in Asia, including Korea, China, Taiwan, Vietnam, Japan, and Asian regions of Russia. Parasite eggs excreted from infected humans or animals are ingested by a host snail (the first intermediate host), where they undergo several developmental stages. Cercariae are then released from the snail and penetrate freshwater fish (the second intermediate host), encysting as metacercariae in the muscles or under the scales. Humans become infected by eating raw or undercooked fish from endemic countries. After ingestion, the metacercariae excyst in gastric juices and migrate via the duodenum, the ampulla of Vater, and the extrahepatic biliary system to the intrahepatic bile ducts.

The clinical manifestations of infection with *Opisthorchis* species and *C. sinensis* are similar. Pathologic changes are typically seen in the bile ducts, liver, and gallbladder **(Table 229-3)**. Tissue damage and intense inflammation is caused by mechanical and chemical irritation and immune responses to worms or worm products, and chronic inflammation may result in the development of cholangiocarcinoma. Both *O. viverrini* and *C. sinensis* are classified by the IARC as definitely carcinogenic (class 1). Acute and light infections are mostly asymptomatic, but hepatitis-like signs and symptoms, with high fever and chills, have been reported, especially in *O. felineus* infections. In general, only heavily infected people have symptoms and severe complications (Table 229-3).

TABLE 229–3 Clinical Features of Food-Borne Trematode Infections

	Symptoms or Signs		
Infection	**Early or Acute Stage**	**Established or Chronic Stage**	**Complications**
Liver Flukes			
Clonorchis sinensis, Opisthorchis viverrini, Opisthorchis felineus	Often asymptomatic; sometimes hepatitis-like symptoms and high fever (especially with *O. felineus*)	Biliary colic, cholestatic jaundice, recurrent cholangitis and cholelithiasis; hepatomegaly, gallbladder enlargement, periductal fibrosis. Light infections are often asymptomatic and remain so for years.	Pancreatitis, cholangiocarcinoma[a]
Fasciola hepatica, Fasciola gigantica	Acute onset (1–4 weeks after infection) with high fever, weight loss, sometimes urticaria and liver tenderness	Biliary colic, cholestatic jaundice, recurrent cholangitis and cholelithiasis; thickening, enlargement, and fibrosis of biliary ducts; sometimes repeated relapses of acute symptoms	Pancreatitis. In rare cases: ectopic infections in the central nervous system, orbital area, gastrointestinal tract, lungs, and other organs. Rarely, fascioliasis can be fatal.
Intestinal Flukes			
Fasciolopsis buski, Echinostoma spp., *Heterophyes heterophyes*, several other species	Often asymptomatic; sometimes nonspecific gastrointestinal symptoms	Heavy infection may lead to ulceration of intestinal mucosa and malabsorption. Mild infections are often asymptomatic.	Malnutrition, anemia; rarely, ectopic infection in the central nervous system
Lung Flukes			
Paragonimus westermani, Paragonimus kellicotti	Often asymptomatic; sometimes insidious onset with anorexia and weight loss	Bronchitis-, asthma-, and tuberculosis-like symptoms and signs such as chronic cough, dyspnea, bloody (“rusty”) sputum	Pulmonary cyst formation; ectopic infection in the central nervous system, eyes, skin, heart, abdominal and reproductive organs

[a]Carcinogenesis has not yet been established for *O. felineus*.

The diagnosis of these infections is based on microscopic identification of parasite eggs in stool specimens. The eggs of *Opisthorchis* are indistinguishable from those of *Clonorchis*.

Fascioliasis

Fascioliasis occurs in many areas of the world and usually is caused by *Fasciola hepatica*, a common liver fluke of sheep and cattle. *F. hepatica* is found in more than 50 countries on all continents except Antarctica; *F. gigantica* is less widespread. The areas with the highest

known rates of human *Fasciola* infection are in the Andean highlands of Bolivia and Peru. In other areas where fascioliasis is found, human cases are sporadic.

Unlike the other liver flukes, *Fasciola* species have no second intermediate host, as their infectious metacercariae adhere directly to aquatic plants. Humans usually acquire infection by ingesting aquatic plants, such as watercress, that contain viable metacercariae or by drinking water with free metacercariae.

After metacercariae have excysted in the duodenum, *Fasciola* species migrate through the intestinal wall into the body cavity, penetrate the liver capsule, and move through the liver into the bile ducts. This migration route is different from that of other liver flukes and gives rise to symptoms during the acute migratory phase; the parasites may cause tissue destruction, focal bleeding, and inflammation. Some migrating flukes may deviate from their usual route to cause ectopic infections. In the established latent stage of infection, the parasites may cause bile duct inflammation, resulting in thickening and expansion of the ducts, fibrosis, and ultimately biliary obstruction (Table 229-3). Although some infected people are asymptomatic in the latent phase, others may experience repeated relapses of acute manifestations.

The most widely used diagnostic approach is direct detection of *Fasciola* eggs by microscopic examination of stool or of duodenal or biliary aspirates. Eggs generally cannot be detected until 3–4 months after exposure, whereas antibodies to the parasite may become detectable 2–4 weeks after exposure. More than one stool specimen may be needed for diagnosis, especially in light infections.

INTESTINAL FLUKES

More than 70 species of intestinal flukes can cause human infection. These parasites are found in different geographic areas, with a relatively high prevalence in Southeast Asia. Humans are infected by ingestion of infective metacercariae attached to aquatic plants (*Fasciolopsis buski*) or encysted in freshwater fish. Flukes mature in the human intestines, and eggs are passed with feces. Mechanical irritation of the intestinal wall and inflammation may lead to nonspecific gastrointestinal symptoms such as diarrhea, constipation, and abdominal pain. Most individuals infected with intestinal flukes are asymptomatic, but heavy infections can be severe, with intestinal mucosal ulcerations and malabsorption (Table 229-3). The diagnosis is established by detection of eggs in stool samples. However, eggs from various intestinal trematodes are often morphologically similar, and it is very difficult to distinguish among species. A cautionary note: *Fasciola* eggs can be difficult to distinguish on the basis of morphologic criteria from the eggs of the intestinal fluke *F. buski*. The distinction has implications for therapy: infection with *F. buski* is treated with praziquantel, which is not effective against fascioliasis (Table 229-2).

LUNG FLUKES

Paragonimiasis is a parasitic lung infection caused by lung flukes of the genus *Paragonimus*. It is a food-borne parasitic zoonosis, with most cases reported from Asia and attributable to consumption of raw or undercooked freshwater crustaceans. *Paragonimus westermani* and related species (e.g., *Paragonimus africanus*) are endemic in West Africa, Central and South America, and Asia. The United States has one indigenous species of lung fluke, *Paragonimus kellicotti.*

Paragonimus species require two intermediate hosts: first, a freshwater snail; and second, a freshwater crustacean, such as a freshwater crab. Humans are infected by consuming raw or undercooked infected crustaceans containing *Paragonimus* metacercariae. *Paragonimus* infects other carnivores such as cats, dogs, foxes, rodents, and pigs in addition to humans. After ingestion, metacercariae quickly penetrate the duodenum and traverse the peritoneal cavity, diaphragm, and parietal pleura to mature into hermaphroditic worm pairs in the pleural spaces or lungs within 6–10 weeks. Adults cross-fertilize in cystic cavities in the pleural spaces or lungs within another 4–16 weeks and release unembryonated eggs into bronchioles. The eggs are then coughed up in bloody (“rusty”) sputum and either discharged in sputum or swallowed and later excreted in feces. Unembryonated eggs are passed from the mammalian host into freshwater ecosystems, where they infect intermediate host snails.

The symptoms and signs of paragonimiasis are fever, cough, hemoptysis, and peripheral eosinophilia. Some patients with paragonimiasis and low parasite burdens may remain relatively asymptomatic for prolonged periods or may have recurrent attacks of cough, sputum production, fever, and night sweats that mimic tuberculosis. Infective metacercariae may migrate to extrapulmonary sites such as the brain (cerebral paragonimiasis).

Pulmonary paragonimiasis is diagnosed by detection of parasite ova in sputum and/or feces. Serology can be helpful in egg-negative cases and in cerebral paragonimiasis. Anamnestic information about the consumption of raw or undercooked freshwater crabs by immigrants, expatriates, and returning travelers—and, in the United States, the consumption of raw or undercooked crayfish from freshwater river systems where *P. kellicotti* is endemic—is important in patients presenting with fever, cough, hemoptysis, pleural effusions, and peripheral eosinophilia.

TREATMENT

TREATMENT Food-Borne Trematode Infections

Praziquantel and triclabendazole are the two drugs of choice; Table 229-2 summarizes the dosages recommended for the various trematode infections. All confirmed cases of human paragonimiasis should be treated with praziquantel (Table 229-2) to avoid the complications of extrapulmonary disease. Surgical management may be needed for pulmonary or cerebral lesions.

CONTROL AND PREVENTION

Drugs are currently the main method of controlling the morbidity associated with food-borne trematode infections, but integrated programs (including improved sanitation; food inspections; and information, education, and communication campaigns) are important for sustainable disease control. Collaboration with other sectors (e.g., agricultural, environmental, and educational) is necessary to tackle highly complex situations in which human behavior, biological factors, and agricultural practices all play a role.

FURTHER READING

Andrade G et al: Decline in infection-related morbidities following drug-mediated reductions in the intensity of *Schistosoma* infection: A systematic review and meta-analysis. PLoS Negl Trop Dis 11:e0005372, 2017. [PubMed: 28212414]
Clerinx J, Van Gompel A: Schistosomiasis in travellers and migrants. Travel Med Infect Dis 9:6, 2011. [PubMed: 21216199]
Fried B, Abruzzi A: Food-borne trematode infections of humans in the United States of America. Parasitol Res 106:1263, 2010. [PubMed: 20352454]
Fürst T et al: Global burden of human food-borne trematodiasis: A systematic review and meta-analysis. Lancet Infect Dis 12:210, 2012. [PubMed: 22108757]
Gryseels B et al: Human schistosomiasis. Lancet 368:1106, 2006. [PubMed: 16997665]
Jordan P et al (eds): *Human Schistosomiasis*. CAB International, Wallingford, 1993.
Keiser J, Utzinger J: Food-borne trematodiases. Clin Microbiol Rev 22:466, 2009. [PubMed: 19597009]
Ross AG et al: Katayama syndrome. Lancet Infect Dis 7:218, 2007. [PubMed: 17317603]
Sripa B et al: Liver fluke induces cholangiocarcinoma. PLoS Med 4:e201, 2007. [PubMed: 17622191]
World Health Organization: *Female Genital Schistosomiasis: A Pocket Atlas for Clinical Health-Care Professionals*. Geneva, World Health Organization, 2015. (Available at *http://brightresearch.org/wp-content/uploads/2016/05/FGS-pocket-atlas_eng.pdf* WHO/HTM/NTD/2015.4, 2015. Accessed July 18, 2017.)

Chapter 230: Cestode Infections

A. Clinton White, Jr.; Peter F. Weller

INTRODUCTION

Cestodes, or tapeworms, are segmented worms. The adults reside in the gastrointestinal tract, but the larvae can be found in almost any organ. Human tapeworm infections can be divided into two major clinical groups. In one group, humans are the definitive hosts, with the adult tapeworms living in the gastrointestinal tract (*Taenia saginata*, *Diphyllobothrium*, and *Dipylidium caninum*). In the other, humans are intermediate hosts, with larval-stage parasites present in the tissues; diseases in this category include echinococcosis, sparganosis, and coenurosis. Humans may be the definitive and/or intermediate hosts for *Taenia solium*; both stages of *Hymenolepis nana* are found simultaneously in the human intestines.

The ribbon-shaped tapeworm attaches to the intestinal mucosa by means of sucking cups or hooks located on the scolex. Behind the scolex is a short, narrow neck from which proglottids (segments) form. As each proglottid matures, it is displaced further back from the neck by the formation of new, less mature segments. The progressively elongating chain of attached proglottids, called the *strobila*, constitutes the bulk of the tapeworm. The length varies among species. In some, the tapeworm may consist of more than 1000 proglottids and may be several meters long. The mature proglottids are hermaphroditic and produce eggs, which are subsequently released. Because eggs of the different *Taenia* species are morphologically identical, differences in the morphology of the scolex or proglottids provide the basis for diagnostic identification to the species level.

Most human tapeworms require at least one intermediate host for complete larval development. After ingestion of the eggs or proglottids by an intermediate host, the larval oncospheres are activated, escape the egg, and penetrate the intestinal mucosa. The oncosphere migrates to tissues and develops into an encysted form known as a *cysticercus* (single scolex), a *coenurus* (multiple scolices), or a *hydatid* (cyst with daughter cysts, each containing several protoscolices). The definitive host's ingestion of tissues containing a cyst enables a scolex to develop into a tapeworm.

TAENIASIS SAGINATA AND TAENIASIS ASIATICA

Image not available. The beef tapeworm *T. saginata* occurs in all countries where raw or undercooked beef is eaten. It is most prevalent in sub-Saharan African and Middle Eastern countries. *Taenia asiatica* is closely related to *T. saginata* and is found in Asia, with pigs as intermediate hosts. The clinical manifestations and morphology of these two species are very similar and are therefore discussed together.

Etiology and Pathogenesis

Humans are the only definitive host for the adult stage of *T. saginata* and *T. asiatica*. The tapeworms, which can reach 8 m in length with 1000–2000 proglottids, inhabit the upper jejunum. The scolex of *T. saginata* has four prominent suckers, whereas *T. asiatica* has an unarmed rostellum. Each gravid segment has 15–30 uterine branches (in contrast to 8–12 for *T. solium*). The eggs are indistinguishable from those of *T. solium*; they measure 30–40 μm, contain the oncosphere, and have a thick brown striated shell. Eggs deposited on vegetation can live for months or years until they are ingested by cattle or other herbivores (*T. saginata*) or pigs (*T. asiatica*). The embryo released after ingestion invades the intestinal wall and is carried to striated muscle or viscera, where it transforms into the cysticercus. When ingested in raw or undercooked meat, the cysticercus evaginates and forms a tapeworm in the human intestines. Over ~2 months, the adult worm matures and begins to produce eggs.

Clinical Manifestations

Patients become aware of the infection most commonly by noting passage of proglottids in their feces. The proglottids of *T. saginata* are motile, and patients may experience perianal discomfort when proglottids are discharged. Mild abdominal pain or discomfort, nausea, change in appetite, weakness, and weight loss can occur.

Diagnosis

The diagnosis is made by the detection of eggs or proglottids in the stool. Eggs may also be present in the perianal area; thus, if proglottids or eggs are not found in the stool, the perianal region should be examined with use of a cellophane-tape swab (as in pinworm infection; **Chap. 227**). Distinguishing *T. saginata* or *T. asiatica* from *T. solium* requires examination of mature proglottids. All three species can be distinguished by examining the scolex. Available serologic tests are not helpful diagnostically. Eosinophilia and elevated levels of serum IgE are usually absent.

TREATMENT

TREATMENT Taeniasis Saginata and Taeniasis Asiatica

A single dose of praziquantel (10 mg/kg) is highly effective.

Loading [Contrib]/a11y/accessibility-menu.js

Prevention

The major method of preventing infection is the adequate cooking of beef or pork viscera; exposure to temperatures as low as 56°C for 5 min will destroy cysticerci. Refrigeration or salting for long periods or freezing at –10°C for 9 days also kills cysticerci in beef. General preventive measures include inspection of beef and proper disposal of human feces.

TAENIASIS SOLIUM AND CYSTICERCOSIS

The pork tapeworm *T. solium* can cause two distinct forms of infection in humans: adult tapeworms in the intestine or larval forms in the tissues (cysticercosis). Humans are the only definitive hosts for *T. solium*; pigs are the usual intermediate hosts, although other animals may harbor the larval forms.

Image not available. *T. solium* is found worldwide in areas where pigs are raised and have access to human feces. However, it is most prevalent in Latin America, sub-Saharan Africa, China, India, and Southeast Asia. Cysticercosis occurs in industrialized nations largely as a result of the immigration of infected persons from endemic areas.

Etiology and Pathogenesis

The adult tapeworm generally resides in the upper jejunum. The scolex attaches by both sucking disks and two rows of hooklets. The adult worm usually lives for a few years. The mature tapeworm, usually ~3 m in length, may have as many as 1000 proglottids, each of which produces up to 50,000 eggs. Proglottids are released and excreted into the feces, and the eggs in these proglottids are infective for both humans and animals. After ingestion of eggs by the pig intermediate host, the larvae are activated, escape the egg, penetrate the intestinal wall, and are carried to many tissues; they are most frequently identified in striated muscle of the neck, tongue, and trunk. Within 60–90 days, the encysted larval stage develops. These cysticerci can survive for months to years. By ingesting undercooked pork containing cysticerci, humans acquire infections that lead to intestinal tapeworms. Infections that cause human cysticercosis follow the ingestion of *T. solium* eggs. The eggs are sticky and may be found under the fingernails of tapeworm carriers. Transmission is usually associated with close contact with a tapeworm carrier. Autoinfection may occur if an individual with an egg-producing tapeworm ingests eggs derived from his or her own feces.

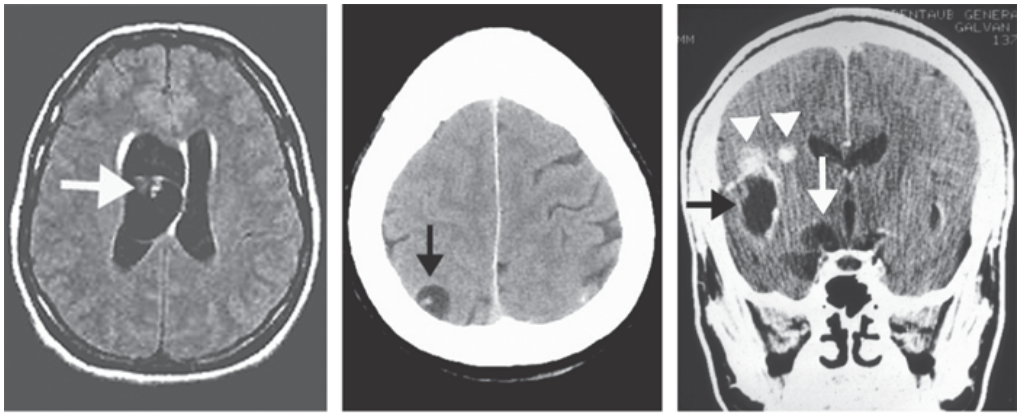
Clinical Manifestations

Intestinal infections with *T. solium* may be asymptomatic. Fecal passage of proglottids may be noted by patients. Other symptoms are infrequent.

In cysticercosis, the clinical manifestations are variable. Cysticerci can be found anywhere in the body but are most commonly detected in the brain, cerebrospinal fluid (CSF), skeletal muscle, subcutaneous tissue, or eye. The clinical presentation of cysticercosis depends on the number and location of cysticerci as well as on the extent of associated inflammatory responses or scarring. Neurologic manifestations are the most common **(Fig. 230-1)**. Seizures are associated with inflammation surrounding cysticerci in the brain parenchyma. These seizures may be generalized, focal, or Jacksonian. Hydrocephalus results from CSF flow obstruction by cysticerci and accompanying inflammation or by CSF outflow obstruction from arachnoiditis. Symptoms of increased intracranial pressure, including headache, nausea, vomiting, changes in vision, dizziness, ataxia, or confusion, are often evident. Patients with hydrocephalus may develop papilledema or display altered mental status. When cysticerci develop at the base of the brain or in the subarachnoid space, they may cause chronic meningitis or arachnoiditis, communicating hydrocephalus, hemorrhages, or strokes.

FIGURE 230-1

Neurocysticercosis is caused by *Taenia solium.* Neurologic infection can be classified on the basis of the location and viability of the parasites. When the parasites are in the ventricles, they often cause obstructive hydrocephalus. ***Left:*** Magnetic resonance imaging showing a cysticercus in the lateral ventricle, with resultant hydrocephalus. The *arrow* points to the scolex within the cystic parasite. ***Center:*** CT showing a parenchymal cysticercus, with enhancement of the cyst wall and an internal scolex (*arrow*). ***Right:*** Multiple cysticerci, including calcified lesions from prior infection (*arrowheads*), viable cysticerci in the basilar cisterns (*white arrow*), and a large degenerating cysticercus in the Sylvian fissure (*black arrow*). *(Modified with permission from JC Bandres et al: Clin Infect Dis 15:799, 1992. © The University of Chicago Press.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Loading [Contrib]/a11y/accessibility-menu.js

The diagnosis of intestinal *T. solium* infection is made by the detection of eggs or proglottids, as described for *T. saginata*. More sensitive methods, including antigen-capture enzyme-linked immunosorbent assay (ELISA), polymerase chain reaction (PCR), and serology for tapeworm stage-specific antigens, are currently available only as research techniques. In cysticercosis, diagnosis can be difficult. A panel of international experts recently proposed revised diagnostic criteria **(Table 230-1)**. Diagnostic certainty is possible only with definite demonstration of the parasite (absolute criteria). This task can be accomplished by histologic observation of the parasite in excised tissue, by funduscopic visualization of the parasite in the subretinal space of the eye, or by neuroimaging studies with definite evidence of a cystic lesion containing a characteristic scolex (Fig. 230-1). With improving resolution of neuroimaging studies, the scolex can now be identified in a large proportion of cases. In other instances, a clinical diagnosis is based on a combination of clinical presentation, radiographic studies, and exposure history (including serodiagnosis).

TABLE 230-1 Revised Diagnostic Criteria for Neurocysticercosis[a]

1. **Absolute criteria**
    a. Histologic demonstration of the parasite from biopsy of a brain or spinal cord lesion
    b. Visualization of subretinal cysticercus
    c. Conclusive demonstration of a scolex within a cystic lesion on neuroimaging studies
2. **Neuroimaging criteria**
    a. **Major neuroimaging criteria**

        Cystic lesions without a discernible scolex, typical small enhancing lesions, multilobulated cystic lesions in the subarachnoid space, typical parenchymal brain calcifications
    b. **Confirmative neuroimaging criteria**

        Resolution of cystic lesions spontaneously or after cysticidal drug therapy

        Migration of ventricular cysts documented on sequential neuroimaging studies
    c. **Minor neuroimaging criteria**

        Obstructive hydrocephalus or abnormal enhancement of basal leptomeninges
3. **Clinical/exposure criteria**
    a. **Major clinical/exposure criteria**

        Detection of specific anticysticercal antibodies (e.g., by enzyme-linked immunoelectrotransfer blot [EITB]) or cysticercal antigens by well-standardized immunodiagnostic tests

        Cysticercosis outside the central nervous system

        Evidence of a household contact with *T. solium* infection
    b. **Minor clinical/exposure criteria**

        Clinical manifestations suggestive of neurocysticercosis

        Individuals coming from or living in an area where cysticercosis is endemic

[a]Diagnosis is confirmed by one absolute criterion, by two major criteria or one major and one confirmatory neuroimaging criteria plus any clinical/exposure criterion, or by one major neuroimaging criterion plus two clinical/exposure criteria (including at least one major clinical/exposure criterion), together with the exclusion of other pathologies producing similar neuroimaging findings. A probable diagnosis is supported by one major neuroimaging criterion plus any two clinical/exposure criteria or by one minor neuroimaging criterion plus at least one major clinical/exposure criterion.

Source: Modified from OH Del Brutto et al: J Neurol Sci 372:202, 2017.

Neuroimaging findings constitute the primary major diagnostic criteria (Fig. 230-1). Major findings include cystic lesions with or without enhancement (e.g., ring enhancement), one or more nodular calcifications (which may also have associated enhancement), focal enhancing lesions, or multilobulated cystic lesions in the subarachnoid space. Cysticerci in the brain parenchyma are usually 5–20 mm in diameter and rounded. Cystic lesions in the subarachnoid space or fissures may enlarge up to 6 cm in diameter and may be lobulated. For cysticerci within the subarachnoid space or ventricles, the walls may be very thin and the cyst fluid is often isodense with CSF. Thus, obstructive hydrocephalus or enhancement of the basilar meninges may be the only finding on CT in extraparenchymal neurocysticercosis. However, since these findings are less specific, they are considered only minor criteria. Cysticerci in the ventricles or subarachnoid space are more readily identified by MRI, especially with three-dimensional views (e.g., fast imaging employing steady-state acquisition [FIESTA] or three-dimensional constructive interference in steady state [3D CISS]). CT is more sensitive than MRI in identifying calcified lesions, whereas MRI is better for identifying cystic lesions, scolices, and enhancement. Spontaneous resolution, resolution after therapy with albendazole, or mobile cystic lesions within the ventricles are findings that can confirm the diagnosis of neurocysticercosis.

Prior exposure significantly modifies the interpretation of neuroimaging studies. Detection of specific antibodies to or antigens of *T. solium* are rLoading [Contrib]/a11y/accessibility-menu.js using unfractionated antigens (e.g., ELISAs using crude parasite antigen) have high rates of false-positive and false-negative results and should be avoided. An immunoblot assay using lentil lectin–purified glycoproteins is >99% specific and

highly sensitive. However, patients with single intracranial lesions or with calcifications may be seronegative. With this assay, serum samples provide greater diagnostic sensitivity than CSF. All of the diagnostic antigens have been cloned, and assays using recombinant antigens are being developed. Antigen detection assays using monoclonal antibodies to detect parasite antigen in the blood or CSF may also facilitate diagnosis and patient follow-up. These assays are currently available commercially in Europe but not in the United States.

Other major clinical/exposure criteria for neurocysticercosis include the presence of cysticerci outside the central nervous system (CNS) (e.g., typical cigar-shaped calcifications in muscle) or exposure to a tapeworm carrier or a household member infected with *T. solium*. Minor clinical/exposure criteria include residence in an endemic area or clinical symptoms suggestive of neurocysticercosis (e.g., seizures or obstructive hydrocephalus).

Studies have demonstrated that clinical criteria may aid in diagnosis in selected cases. In patients from endemic areas who had single enhancing lesions presenting with seizures, a normal physical examination, and no evidence of systemic disease (e.g., no fever, adenopathy, or chest radiographic abnormalities), the constellation of rounded CT lesions 5–20 mm in diameter with no midline shift was almost always caused by neurocysticercosis.

A definite or probable diagnosis is made in accordance with the criteria and combinations of criteria listed in the footnote of Table 230-1. Patients may have CSF pleocytosis with a predominance of lymphocytes, neutrophils, or eosinophils. The protein level in CSF may be elevated; the glucose concentration is usually normal but may be depressed.

TREATMENT

TREATMENT Taeniasis Solium and Cysticercosis

Intestinal *T. solium* infection is treated with a single dose of praziquantel (10 mg/kg). However, praziquantel occasionally evokes an inflammatory response in the CNS if concomitant cryptic cysticercosis is present. Niclosamide (2 g) is also effective but is not widely available.

The initial management of neurocysticercosis should focus on symptom-based treatment of seizures or hydrocephalus. Seizures can usually be controlled with antiepileptic treatment. If parenchymal lesions resolve without development of calcifications and patients remain free of seizures, antiepileptic therapy can usually be discontinued after 1–2 years. Placebo-controlled trials are clarifying the clinical advantage of antiparasitic drugs for parenchymal neurocysticercosis. Trends toward faster resolution of neuroradiologic abnormalities have been observed in most studies. The clinical benefits are less dramatic and consist mainly of shortening the period during which recurrent seizures occur and decreasing the number of patients who have many recurrent seizures. For the treatment of patients with brain parenchymal cysticerci, most authorities favor antiparasitic drugs, including albendazole (15 mg/kg per day for 8–28 days) or praziquantel (50–100 mg/kg daily in three divided doses for 15–30 days). A combination of albendazole and praziquantel (50 mg/kg per day) is more effective in patients with more than two cystic lesions. A longer course or combination therapy is often needed in patients with multiple subarachnoid cysticerci. Both agents may exacerbate the inflammatory response around the dying parasite, thereby exacerbating seizures or hydrocephalus as well. Thus, patients receiving these drugs should be carefully monitored. High-dose glucocorticoids should be used during treatment. Because glucocorticoids induce first-pass metabolism of praziquantel and may decrease its antiparasitic effect, cimetidine should be co-administered to inhibit praziquantel metabolism.

For patients with hydrocephalus, the emergent reduction of intracranial pressure is the mainstay of therapy. In the case of obstructive hydrocephalus, the preferred approach is removal of the cysticercus via endoscopic surgery. However, this intervention is not always possible. An alternative approach is initially to perform a diverting procedure, such as ventriculoperitoneal shunting. Historically, shunts have usually failed, but failure rates have been lowered by administration of antiparasitic drugs and glucocorticoids. Open craniotomy to remove cysticerci is now required only infrequently but is effective for fourth-ventricular cysticerci. For patients with subarachnoid cysts or giant cysticerci, anti-inflammatory medications such as glucocorticoids are needed to reduce arachnoiditis and accompanying vasculitis. Most authorities recommend prolonged courses of antiparasitic drugs as well as shunting when hydrocephalus is present. Methotrexate should be used as a steroid-sparing agent in patients requiring prolonged therapy. In patients with diffuse cerebral edema and elevated intracranial pressure due to multiple inflamed lesions, glucocorticoids are the mainstay of therapy, and antiparasitic drugs should be avoided. For ocular and spinal medullary lesions, drug-induced inflammation may cause irreversible damage. Ocular disease should be managed surgically. Recent data suggest that either medical or surgical therapy can be used for spinal disease.

Prevention

Measures for the prevention of intestinal *T. solium* infection consist of the application to pork of precautions similar to those described above for beef with regard to *T. saginata* infection. The prevention of cysticercosis involves minimizing the opportunities for ingestion of fecally derived eggs by means of good personal hygiene, effective fecal disposal, and treatment and prevention of human intestinal infections. Mass chemotherapy has been administered to human and porcine populations in efforts at disease eradication. Finally, vaccines to prevent porcine cysticercosis have shown promise in studies and are under development.

ECHINOCOCCOSIS

Image not available. Echinococcosis is an infection caused in humans by the larval stage of *Echinococcus granulosus* sensu lato, *E. multilocularis*, or *E. vogeli*. *E. granulosus* sensu lato parasites produce cystic hydatid disease, with unilocular cystic lesions. These infections are prevalent in most areas where livestock is raised in association with dogs. Molecular evidence has demonstrated that *E. granulosus* strains belong to a range of genotypes and more than one species. Currently, human cystic hydatid disease is caused by organisms formerly termed *E. granulosus* that are now classified as *E. granulosus* sensu stricto (genotypes 1–3), *E. canadensis* (genotypes 6–8 and 10), and *E. ortleppi* (genotype 5). Other species—*E. equinus* (genotype 4) and *E. felidis* (lion strain)—have not been identified in human infections. *E. granulosus* sensu lato parasites are found on all continents, with areas of high prevalence in China, central Asia, the Middle East, the Mediterranean region, eastern Africa, and parts of South America. *E. multilocularis*, which causes multilocular alveolar lesions that are locally invasive, is found in Alpine, sub-Arctic, or Arctic regions, including central and northern Europe; western China and central Asia; and isolated areas in North America. *E. vogeli* causes polycystic hydatid disease and is found only in Central and South America.

Loading [Contrib]/a11y/accessibility-menu.js

Like other cestodes, echinococcal species have both intermediate and definitive hosts. The definitive hosts are canines that pass eggs in their feces. After the ingestion of eggs, cysts develop in the intermediate hosts—sheep, cattle, humans, goats, camels, and horses for the *E. granulosus* complex and mice and other rodents for *E. multilocularis*. When a dog (*E. granulosus*) or fox (*E. multilocularis*) ingests infected meat containing cysts, the life cycle is completed.

Etiology

The small (5-mm-long) adult *E. granulosus* sensu lato worms live for 5–20 months in the jejunum of dogs. They have three proglottids: one immature, one mature, and one gravid. The gravid segment splits to release eggs that are morphologically similar to *Taenia* eggs and are extremely hardy. After humans ingest the eggs, embryos escape from the eggs, penetrate the intestinal mucosa, enter the portal circulation, and are carried to various organs, most commonly the liver and lungs. Larvae of *E. granulosus* sensu lato develop into fluid-filled unilocular hydatid cysts that consist of an external membrane and an inner germinal layer. Daughter cysts develop from the inner aspect of the germinal layer, as do germinating cystic structures called *brood capsules*. New larvae, called *protoscolices*, develop in large numbers within the brood capsule. The cysts expand slowly over a period of years.

The life cycle of *E. multilocularis* is similar except that wild canines, such as foxes, serve as the definitive hosts, and small rodents serve as the intermediate hosts. The larval form of *E. multilocularis*, however, is quite different in that it remains in the proliferative phase, the parasite is always multilocular, and vesicles without brood capsules or protoscolices progressively invade the host tissue by peripheral extension of processes from the germinal layer.

Clinical Manifestations

Slowly enlarging echinococcal cysts generally remain asymptomatic until their expanding size or their space-occupying effect in an involved organ elicits symptoms. The liver and the lungs are the most common sites of these cysts. The liver is involved in about two-thirds of *E. granulosus* infections and in nearly all *E. multilocularis* infections. Because a period of years elapses before cysts enlarge sufficiently to cause symptoms, they may be discovered incidentally on a routine x-ray or ultrasound study.

Patients with hepatic echinococcosis who are symptomatic most often present with abdominal pain or a palpable mass in the right upper quadrant. Compression of a bile duct or leakage of cyst fluid into the biliary tree may mimic recurrent cholelithiasis, and biliary obstruction can result in jaundice. Rupture of or episodic leakage from a hydatid cyst may produce fever, pruritus, urticaria, eosinophilia, or anaphylaxis. Pulmonary hydatid cysts may rupture into the bronchial tree or pleural cavity and produce cough, salty phlegm, dyspnea, chest pain, or hemoptysis. Rupture of hydatid cysts, which can occur spontaneously or at surgery, may lead to multifocal dissemination of protoscolices, which can form additional cysts. Other presentations are due to the involvement of bone (invasion of the medullary cavity with slow bone erosion producing pathologic fractures), the CNS (space-occupying lesions), the heart (conduction defects, pericarditis), and the pelvis (pelvic mass).

The larval forms of *E. multilocularis* characteristically present as a slowly growing hepatic tumor, with progressive destruction of the liver and extension into vital structures. Patients commonly report upper-quadrant and epigastric pain. Liver enlargement and obstructive jaundice may be apparent. The lesions may infiltrate adjoining organs (e.g., diaphragm, kidneys, or lungs) or may metastasize to the spleen, lungs, or brain.

Diagnosis

Radiographic and related imaging studies are important in detecting and evaluating echinococcal cysts. Plain x-rays will define pulmonary cysts of *E. granulosus* sensu lato—usually as rounded masses of uniform density—but may miss cysts in other organs unless there is cyst wall calcification (as occurs in the liver). MRI, CT, and ultrasound reveal well-defined cysts with thick or thin walls. Imaging methods may reveal a fluid layer of different density, termed *hydatid sand*, that contains protoscolices. However, the most pathognomonic finding, if demonstrable, is that of daughter cysts within the larger cyst. This finding, like eggshell or mural calcification on CT, is indicative of *E. granulosus* infection and helps to distinguish the cyst from carcinomas, bacterial or amebic liver abscesses, or hemangiomas. In contrast, ultrasound or CT of alveolar hydatid cysts reveals indistinct solid masses with central necrosis and plaquelike calcifications.

A specific diagnosis of cystic hydatid disease can be made by the examination of aspirated fluids for protoscolices or hooklets, but diagnostic aspiration is not usually recommended because of the potential risk of fluid leakage resulting in either dissemination of infection or anaphylactic reactions. Serodiagnostic assays can be useful, although a negative test does not exclude the diagnosis of echinococcosis. Cysts in the liver elicit positive antibody responses in ~90% of cases, whereas up to 50% of individuals with cysts in the lungs are seronegative. Detection of antibody to specific echinococcal antigens by immunoblotting has the highest degree of specificity.

TREATMENT

TREATMENT Echinococcosis

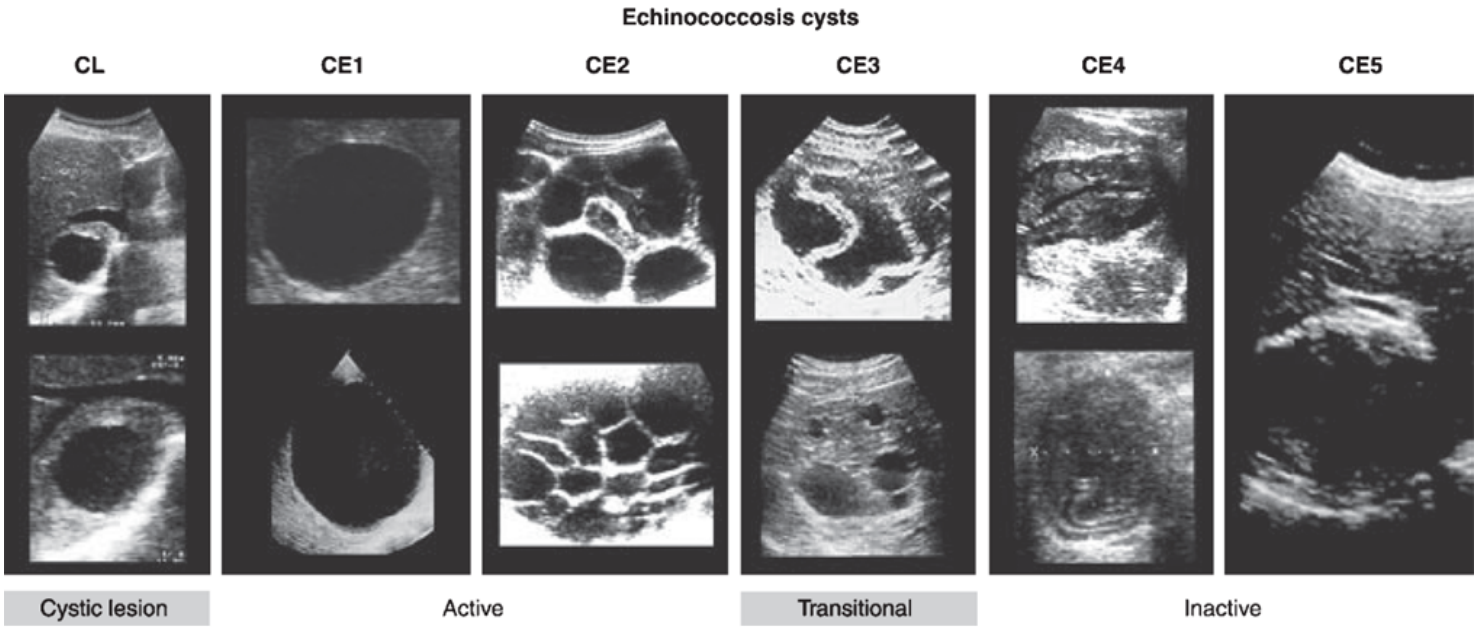
Therapy for cystic echinococcosis is based on considerations of the size, location, and manifestations of cysts and the overall health of the patient. Surgery has traditionally been the principal definitive method of treatment. Currently, ultrasound staging is recommended for cystic echinococcosis **(Fig. 230-2)**. Small CL, CE1, and CE3 lesions may respond to chemotherapy with albendazole. For CE1 lesions and uncomplicated CE3 lesions, PAIR (*p*ercutaneous *a*spiration, *i*nfusion of scolicidal agents, and *r*easpiration) is now recommended instead of surgery. PAIR is contraindicated for superficially located cysts (because of the risk of rupture), for cysts with multiple thick internal septal divisions (honeycombing pattern), and for cysts communicating with the biliary tree. For prophylaxis of secondary peritoneal echinococcosis due to inadvertent spillage of fluid during PAIR, the administration of albendazole (15 mg/kg daily in two divided doses) should be initiated at least 2 days before the procedure and continued for at least 4 weeks afterward. Ultrasound- or CT-guided aspiration allows confirmation of the diagnosis by demonstration of protoscolices in the aspirate. After aspiration, contrast material should be injected to detect occult communications with the biliary tract. Alternatively, the fluid should be checked for bile staining visually and by dipstick. If no bile is found and no communication is visualized, the contrast material is reaspirated, with subsequent infusion of scolicidal agents (usually 95% ethanol; alternatively, hypertonic saline). This approach, when implemented by a skilled practitioner, yields rates of cure and relapse equivalent to tLoading [Contrib]/a11y/accessibility-menu.jsperative morbidity and shorter hospitalization. In experienced hands, some CE2 lesions can be treated by modified catheter drainage. Daughter cysts within the primary cyst may need to be punctured separately.

Surgery remains the treatment of choice for complicated cystic echinococcosis (e.g., cysts communicating with the biliary tract), for most thoracic and intracranial cysts, and for areas where PAIR is not possible. For liver cysts, the preferred surgical approach is total cystectomy, in which the entire cyst and the surrounding fibrous tissue are removed. The risks posed by leakage of fluid during surgery or PAIR include anaphylaxis and dissemination of infectious protoscolices. The latter complication has been minimized by careful attention to the prevention of spillage of the cyst and by soaking of the drapes with hypertonic saline. Infusion of scolicidal agents is no longer recommended because of problems with hypernatremia, intoxication, or sclerosing cholangitis. Albendazole, which is active against *Echinococcus*, should be administered adjunctively, beginning several days before resection of the liver and continuing for several weeks for *E. granulosus*. Praziquantel (50 mg/kg daily for 2 weeks) may hasten the death of the protoscolices. Medical therapy with albendazole alone for 12 weeks to 6 months results in cure in ~30% of cases and in improvement in another 50%. In many instances of treatment failure, *E. granulosus* infections are subsequently treated successfully with PAIR or additional courses of medical therapy. Response to treatment is best assessed by serial imaging studies, with attention to cyst size and consistency. Some cysts may not demonstrate complete radiologic resolution even though no viable protoscolices are present. Some of these cysts with partial radiologic resolution (e.g., CE4 or CE5) can be managed with observation only.

Surgical resection remains the treatment of choice for *E. multilocularis* infection. Complete removal of the parasite continues to offer the best chance for cure. Ongoing therapy with albendazole for at least 2 years after presumptively curative surgery is recommended. Positron emission tomography can be used to follow disease activity. Most cases are diagnosed at a stage at which complete resection is not possible; in these cases, albendazole treatment should be continued indefinitely, with careful monitoring. In some cases, liver transplantation has been used because of the size of the necessary liver resection. However, continuous immunosuppression favors the proliferation of *E. multilocularis* larvae and reinfection of the transplant. Thus, indefinite treatment with albendazole is required.

FIGURE 230-2

Management of cystic hydatid disease caused by *Echinococcus granulosus* should be based on viability of the parasite, which can be estimated from radiographic appearance. The ultrasound appearance includes lesions classified as active, transitional, and inactive. *Active* cysts include types CL (with a cystic lesion and no visible cyst wall), CE1 (with a visible cyst wall and internal echoes [*snowflake sign*]), and CE2 (with a visible cyst wall and internal septation). *Transitional* cysts (CE3) may have detached laminar membranes or may be partially collapsed. *Inactive cysts* include types CE4 (a nonhomogeneous mass) and CE5 (a cyst with a thick calcified wall). *(Adapted from RL Guerrant et al [eds]: Tropical Infectious Diseases: Principles, Pathogens and Practice, 2nd ed, p 1312. © 2005, with permission from Elsevier Science.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prevention

Image not available. In endemic areas, echinococcosis can be prevented by administering praziquantel to infected dogs, by denying dogs access to viscera from infected animals, or by vaccinating sheep. Limiting the number of stray dogs is helpful in reducing the prevalence of infection among humans. In Europe, *E. multilocularis* infection has been associated with gardening; gloves should be used when working with soil.

HYMENOLEPIASIS NANA

Image not available. Infection with *H. nana*, the dwarf tapeworm, is the most common of all the cestode infections. *H. nana* is endemic in both temperate and tropical regions of the world. Infection is spread by fecal/oral contamination.

Etiology and Pathogenesis

Loading [Contrib]/a11y/accessibility-menu.js

H. nana is the only cestode of humans that does not require an intermediate host. Both the larval and adult phases of the life cycle take place in the human. The adult—the smallest tapeworm parasitizing humans—is ~2 cm long and dwells in the proximal ileum. Proglottids, which are small and rarely seen in the stool, release spherical eggs 30–44 µm in diameter, each of which contains an oncosphere with six hooklets. The eggs are immediately infective and are unable to survive for >10 days in the external environment. When the egg is ingested by a new host, the oncosphere is freed and penetrates the intestinal villi, becoming a cysticercoid larva. Larvae migrate back into the intestinal lumen, attach to the mucosa, and mature into adult worms over 10–12 days. Eggs may also hatch before passing into the stool, causing internal autoinfection with increasing numbers of intestinal worms. Although the life span of adult *H. nana* worms is only ~4–10 weeks, the autoinfection cycle perpetuates the infection.

Clinical Manifestations

H. nana infection, even with many intestinal worms, is usually asymptomatic. Heavy infection may be associated with diarrhea, abdominal pain, and weight loss.

Diagnosis

Infection is diagnosed by the finding of eggs in the stool.

TREATMENT

TREATMENT Hymenolepiasis Nana

Praziquantel (25 mg/kg once) is the treatment of choice, because it acts against both the adult worms and the cysticercoids in the intestinal villi. Nitazoxanide (500 mg bid for 3 days) may be used as an alternative.

Prevention

Good personal hygiene and improved sanitation can eradicate the disease. Epidemics have been controlled by mass chemotherapy coupled with improved hygiene.

HYMENOLEPIASIS DIMINUTA

Hymenolepis diminuta, a cestode of rodents, occasionally infects small children, who ingest the larvae in uncooked cereal foods contaminated by fleas and other insects in which larvae develop. Infection is usually asymptomatic and is diagnosed by the detection of eggs in the stool. Treatment with praziquantel results in cure in most cases.

DIPHYLLOBOTHRIASIS

Image not available. *Diphyllobothrium latum* and other diphyllobothriid parasites (including *Adenocephlus pacificus* and *Diplogonoporus* species) are found in the lakes, rivers, and deltas of the Northern Hemisphere, central Africa, and South America.

Etiology and Pathogenesis

The adult worm—the longest tapeworm (up to 25 m)—attaches to the ileal and occasionally to the jejunal mucosa by its suckers, which are located on its elongated scolex. The adult worm has 3000–4000 proglottids, which release ~1 million eggs daily into the feces. If an egg reaches water, it hatches and releases a free-swimming embryo that can be eaten by small freshwater crustaceans (*Cyclops* or *Diaptomus* species). After an infected crustacean containing a developed procercoid is swallowed by a fish, the larva migrates into the fish's flesh and grows into a sparganum, or plerocercoid larva. Humans acquire the infection by ingesting infected raw or smoked fish. Within 3–5 weeks, the tapeworm matures into an adult in the human intestine.

Clinical Manifestations

Most *D. latum* infections are asymptomatic, although manifestations may include transient abdominal discomfort, diarrhea, vomiting, weakness, and weight loss. Occasionally, infection can cause acute abdominal pain and intestinal obstruction; in rare cases, cholangitis or cholecystitis may be produced by migrating proglottids.

Image not available. Because the tapeworm absorbs large quantities of vitamin B_{12} and interferes with ileal B_{12} absorption, vitamin B_{12} deficiency can develop, but this effect has been noted only in Scandinavia, where up to 2% of infected patients, especially the elderly, have megaloblastic anemia resembling pernicious anemia and may exhibit neurologic sequelae of B_{12} deficiency.

Diagnosis

The diagnosis is made readily by the detection of the characteristic eggs in the stool. The eggs possess a single shell with an operculum at one end and a knob at the other. Mild to moderate eosinophilia may be detected.

TREATMENT

TREATMENT Diphyllobothriasis

Loading [Contrib]/a11y/accessibility-menu.js

Praziquantel (5–10 mg/kg once) is highly effective. Parenteral vitamin B_{12} should be given if B_{12} deficiency is manifest.

Prevention

Infection can be prevented by heating fish to 54°C for 5 min or by freezing it at −18°C for 24 h. Placing fish in brine with a high salt concentration for long periods kills the eggs.

DIPYLIDIASIS

Dipylidium caninum, a common tapeworm of dogs and cats, may accidentally infect humans. Dogs, cats, and occasionally humans become infected by ingesting fleas harboring cysticercoids. Children are more likely to become infected than adults. Most infections are asymptomatic, but passage of segments in the stool or vague abdominal symptoms may occur. The diagnosis is made by the detection of proglottids or ova in the stool. As in *D. latum* infection, therapy consists of praziquantel. Prevention requires anthelmintic treatment and flea control for pet dogs or cats.

SPARGANOSIS

Humans can be infected by the sparganum, or plerocercoid larva, of a diphyllobothriid tapeworm of the genus *Spirometra*. Infection can be acquired by the consumption of water containing infected *Cyclops*; by the ingestion of infected snakes, birds, or mammals; or by the application of infected flesh as poultices. The worm migrates slowly in tissues, and infection commonly presents as a subcutaneous swelling. Periorbital tissues can be involved, and ocular sparganosis may destroy the eye. Surgical excision is used to treat localized sparganosis.

COENUROSIS

This rare infection of humans by the larval stage (coenurus) of the dog tapeworm *Taenia multiceps* or *T. serialis* results in a space-occupying cystic lesion. As in cysticercosis, involvement of the CNS and subcutaneous tissue is most common. Both definitive diagnosis and treatment require surgical excision of the lesion. Chemotherapeutic agents generally are not effective.

FURTHER READING

Brunetti E et al: Expert consensus for the diagnosis and treatment of cystic and alveolar echinococcosis in humans. Acta Trop 114:1, 2010. [PubMed: 19931502]
Del Brutto OH et al: Revised diagnostic criteria for neurocysticercosis. J Neurol Sci 372:202, 2017. [PubMed: 28017213]
Ito A et al: Cystic echinococcosis: Future perspectives of molecular epidemiology. Acta Trop 165:3, 2017. [PubMed: 27237060]
Kern P et al: The echinococcoses: Diagnosis, clinical management and burden of disease. Adv Parasitol 96:259, 2017. [PubMed: 28212790]
O'Neal SE, Flecker RH: Hospitalization frequency and charges for neurocysticercosis, United States, 2003–2012. Emerg Infect Dis 21:969, 2015. [PubMed: 25988221]
Scholz T et al: Update on the human broad tapeworm (genus *Diphyllobothrium*), including clinical relevance. Clin Microbiol Rev 22:146, 2009. [PubMed: 19136438]
White AC Jr et al: Diagnosis and treatment of neurocysticercosis: 2017 clinical practice guidelines by the Infectious Diseases Society of America (IDSA) and the American Society of Tropical Medicine and Hygiene (ASTMH). Clin Infect Dis 66:1159, 2018.

Loading [Contrib]/a11y/accessibility-menu.js

Chapter 231: Approach to the Patient with Possible Cardiovascular Disease

Joseph Loscalzo

FIGURE 231-1

THE MAGNITUDE OF THE PROBLEM

Cardiovascular diseases comprise the most prevalent serious disorders in industrialized nations and are a rapidly growing problem in developing nations **(Chap. 233)**. Age-adjusted death rates for coronary heart disease have declined by two-thirds in the last four decades in the United States, reflecting the identification and reduction of risk factors as well as improved treatments and interventions for the management of coronary artery disease, arrhythmias, and heart failure. Nonetheless, cardiovascular diseases remain the most common causes of death, responsible for 35% of all deaths, almost 1 million deaths each year. Approximately one-fourth of these deaths are sudden. In addition, cardiovascular diseases are highly prevalent, diagnosed in 80 million adults, or ~35% of the adult population. The growing prevalence of obesity **(Chap. 395)**, type 2 diabetes mellitus **(Chap. 396)**, and metabolic syndrome **(Chap. 401)**, which are important risk factors for atherosclerosis, now threatens to reverse the progress that has been made in the age-adjusted reduction in the mortality rate of coronary heart disease.

For many years, cardiovascular disease was considered to be more common in men than in women. In fact, the percentage of all deaths secondary to cardiovascular disease is higher among women (43%) than among men (37%) **(Chap. 391)**. In addition, although the absolute number of deaths secondary to cardiovascular disease has declined over the past decades in men, this number has actually risen in women. Inflammation, obesity, type 2 diabetes mellitus, and the metabolic syndrome appear to play more prominent roles in the development of coronary atherosclerosis in women than in men. Coronary artery disease (CAD) is more frequently associated with dysfunction of the coronary microcirculation in women than in men. Exercise electrocardiography has a lower diagnostic accuracy in the prediction of epicardial obstruction in women than in men.

NATURAL HISTORY

Cardiovascular disorders often present acutely, as in a previously asymptomatic person who develops an acute myocardial infarction **(Chap. 269)**, or a previously asymptomatic patient with hypertrophic cardiomyopathy **(Chap. 254)** or with a prolonged QT interval **(Chap. 247)** whose first clinical manifestation is syncope or even sudden death. However, the alert physician may recognize the patient at risk for these complications long before they occur and often can take measures to prevent their occurrence. For example, a patient with acute myocardial infarction will often have had risk factors for atherosclerosis for many years. Had these risk factors been recognized, their elimination or

reduction might have delayed or even prevented the infarction. Similarly, a patient with hypertrophic cardiomyopathy may have had a heart murmur for years and a family history of this disorder. These findings could have led to an echocardiographic examination, recognition of the condition, and appropriate therapy long before the occurrence of a serious acute manifestation.

Patients with valvular heart disease or idiopathic dilated cardiomyopathy, by contrast, may have a prolonged course of gradually increasing dyspnea and other manifestations of chronic heart failure that is punctuated by episodes of acute deterioration only late in the course of the disease. Understanding the natural history of various cardiac disorders is essential for applying appropriate diagnostic and therapeutic measures to each stage of the condition, as well as for providing the patient and family with the likely prognosis.

CARDIAC SYMPTOMS

The symptoms caused by heart disease result most commonly from myocardial ischemia, disturbance of the contraction and/or relaxation of the myocardium, obstruction to blood flow, or an abnormal cardiac rhythm or rate. Ischemia, which is caused by an imbalance between the heart's oxygen supply and demand, is manifest most frequently as chest discomfort **(Chap. 11)**, whereas reduction of the pumping ability of the heart commonly leads to fatigue and elevated intravascular pressure upstream of the failing ventricle. The latter results in abnormal fluid accumulation, with peripheral edema **(Chap. 37)** or pulmonary congestion and dyspnea **(Chap. 33)**. Obstruction to blood flow, as occurs in valvular stenosis, can cause symptoms resembling those of myocardial failure **(Chap. 252)**. Cardiac arrhythmias often develop suddenly, and the resulting symptoms and signs—palpitations **(Chap. 39)**, dyspnea, hypotension, and syncope **(Chap. 18)**—generally occur abruptly and may disappear as rapidly as they develop.

Although dyspnea, chest discomfort, edema, and syncope are cardinal manifestations of cardiac disease, they occur in other conditions as well. Thus, dyspnea is observed in disorders as diverse as pulmonary disease, marked obesity, and anxiety **(Chap. 33)**. Similarly, chest discomfort may result from a variety of noncardiac and cardiac causes other than myocardial ischemia **(Chap. 11)**. Edema, an important finding in untreated or inadequately treated heart failure, also may occur with primary renal disease and in hepatic cirrhosis **(Chap. 37)**. Syncope occurs not only with serious cardiac arrhythmias but in a number of neurologic conditions as well **(Chap. 18)**. Whether heart disease is responsible for these symptoms frequently can be determined by carrying out a careful clinical examination **(Chap. 234)**, supplemented by noninvasive testing using electrocardiography at rest and during exercise **(Chap. 235)**, echocardiography, roentgenography, and other forms of myocardial imaging **(Chap. 236)**.

Myocardial or coronary function that may be adequate at rest may be insufficient during exertion. Thus, dyspnea and/or chest discomfort that appear during activity are characteristic of patients with heart disease, whereas the opposite pattern, that is, the appearance of these symptoms at rest and their remission during exertion, is rarely observed in such patients. It is important, therefore, to question the patient carefully about the relation of symptoms to exertion.

Many patients with cardiovascular disease may be asymptomatic both at rest and during exertion but may present with an abnormal physical finding such as a heart murmur, elevated arterial pressure, or an abnormality of the electrocardiogram (ECG) or imaging test. It is important to assess the global risk of CAD in asymptomatic individuals, using a combination of clinical assessment and measurement of cholesterol and its fractions, as well as other biomarkers, such as C-reactive protein, in some patients. Since the first clinical manifestation of CAD may be catastrophic—sudden cardiac death, acute myocardial infarction, or stroke in previous asymptomatic persons—it is mandatory to identify those at high risk of such events and institute further testing and preventive measures.

DIAGNOSIS

As outlined by the New York Heart Association (NYHA), the elements of a complete cardiac diagnosis include the systematic consideration of the following:

1. *The underlying etiology*. Is the disease congenital, hypertensive, ischemic, or inflammatory in origin?

2. *The anatomic abnormalities*. Which chambers are involved? Are they hypertrophied, dilated, or both? Which valves are affected? Are they regurgitant and/or stenotic? Is there pericardial involvement? Has there been a myocardial infarction?

3. *The physiologic disturbances*. Is an arrhythmia present? Is there evidence of congestive heart failure or myocardial ischemia?

4. *Functional disability*. How strenuous is the physical activity required to elicit symptoms? The classification provided by the NYHA has been found to be useful in describing functional disability **(Table 231-1)**.

TABLE 231-1 New York Heart Association Functional Classification

Class I	**Class III**
No limitation of physical activity	Marked limitation of physical activity
No symptoms with ordinary exertion	Less than ordinary activity causes symptoms
Class II	Asymptomatic at rest
Slight limitation of physical activity	**Class IV**
Ordinary activity causes symptoms	Inability to carry out any physical activity without discomfort
	Symptoms at rest

Source: Data from The Criteria Committee of the New York Heart Association.

One example may serve to illustrate the importance of establishing a complete diagnosis. In a patient who presents with exertional chest discomfort, the identification of myocardial ischemia as the etiology is of great clinical importance. However, the simple recognition of ischemia is insufficient to formulate a therapeutic strategy or prognosis until the underlying anatomic abnormalities responsible for the myocardial ischemia, for example, coronary atherosclerosis or aortic stenosis, are identified and a judgment is made about whether other physiologic disturbances that cause an imbalance between myocardial oxygen supply and demand, such as severe anemia, thyrotoxicosis, or supraventricular tachycardia, play contributory roles. Finally, the severity of the disability should govern the extent and tempo of the workup and strongly influence the therapeutic strategy that is selected.

The establishment of a correct and complete cardiac diagnosis usually commences with the history and physical examination **(Chap. 234)**. Indeed, the clinical examination remains the basis for the diagnosis of a wide variety of disorders. The clinical examination may then be supplemented by five types of laboratory tests: (1) ECG **(Chap. 235)**, (2) noninvasive imaging examinations (chest roentgenogram, echocardiogram, radionuclide imaging, computed tomographic imaging, positron emission tomography, and magnetic resonance imaging) **(Chap. 236)**, (3) blood tests to assess risk (e.g., lipid determinations, C-reactive protein) or cardiac function (e.g., brain natriuretic peptide [BNP] **[Chap. 252]**), (4) occasionally specialized invasive examinations (i.e., cardiac catheterization and coronary arteriography **[Chap. 237]**), and (5) genetic tests to identify monogenic cardiac diseases (e.g., hypertrophic cardiomyopathy **[Chap. 254]**, Marfan's syndrome **[Chap. 406]**, and abnormalities of cardiac ion channels that lead to prolongation of the QT interval and an increase in the risk of sudden death **[Chap. 241]**). These tests are becoming more widely available.

FAMILY HISTORY

In eliciting the history of a patient with known or suspected cardiovascular disease, particular attention should be directed to the family history. Familial clustering is common in many forms of heart disease. Mendelian transmission of single-gene defects may occur, as in hypertrophic cardiomyopathy **(Chap. 254)**, Marfan's syndrome **(Chap. 406)**, and sudden death associated with a prolonged QT syndrome **(Chap. 247)**. Premature coronary disease and essential hypertension, type 2 diabetes mellitus, and hyperlipidemia (the most important risk factors for CAD) are usually polygenic disorders. Although familial transmission may be less obvious than in the monogenic disorders, it is helpful in assessing risk and prognosis in polygenic disorders, as well. Familial clustering of cardiovascular diseases not only may occur on a genetic basis but also may be related to familial dietary or behavior patterns, such as excessive ingestion of salt or calories and cigarette smoking.

ASSESSMENT OF FUNCTIONAL IMPAIRMENT

When an attempt is made to determine the severity of functional impairment in a patient with heart disease, it is helpful to ascertain the level of activity and the rate at which it is performed before symptoms develop. Thus, it is not sufficient to state that the patient complains of dyspnea. The breathlessness that occurs after running up two long flights of stairs denotes far less functional impairment than do similar symptoms that occur after taking a few steps on level ground. In addition, the degree of customary physical activity at work and during recreation should be considered. The development of two-flight dyspnea in a well-conditioned marathon runner may be far more significant than the development of one-flight dyspnea in a previously sedentary person. The history should include a detailed consideration of the patient's therapeutic regimen. For example, the persistence or development of edema, breathlessness, and other manifestations of heart failure in a patient who is receiving optimal doses of diuretics and other therapies for heart failure **(Chap. 252)** is far graver than are similar manifestations in the absence of treatment. Similarly, the presence of angina pectoris despite treatment with optimal doses of multiple antianginal drugs **(Chap. 267)** is more serious than it is in a patient on no therapy. In an effort to determine the progression of symptoms, and thus the severity of the underlying illness, it may be useful to ascertain what, if any, specific tasks the patient could have carried out 6 months or 1 year earlier that he or she cannot carry out at present.

ELECTROCARDIOGRAM

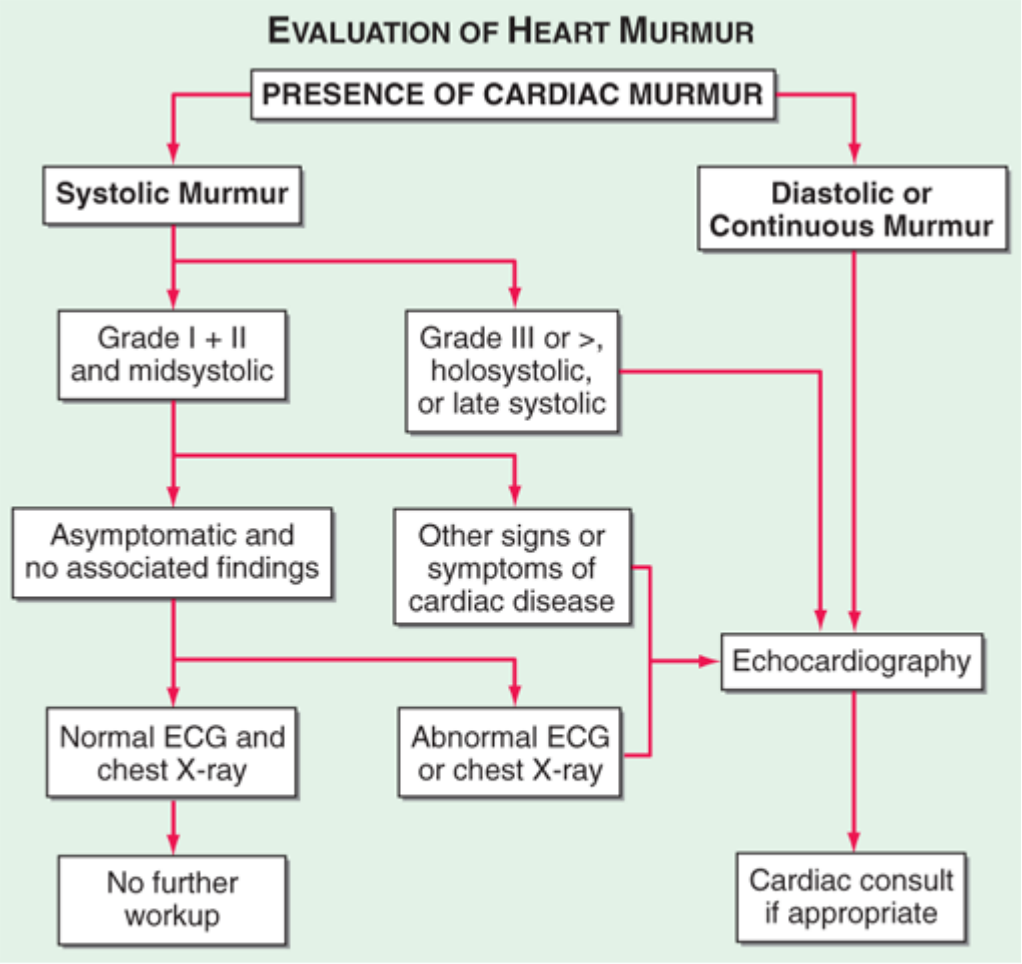
(See also Chap. 235) Although an ECG usually should be recorded in patients with known or suspected heart disease, with the exception of the identification of arrhythmias, conduction abnormalities, ventricular hypertrophy, and acute myocardial infarction, it generally does not establish a specific diagnosis. The range of normal electrocardiographic findings is wide, and the tracing can be affected significantly by many noncardiac factors, such as age, body habitus, and serum electrolyte concentrations. In general, electrocardiographic changes should be interpreted in the context of other abnormal cardiovascular findings.

ASSESSMENT OF THE PATIENT WITH A HEART MURMUR

(Fig. 231-1) The cause of a heart murmur can often be readily elucidated from a systematic evaluation of its major attributes: timing, duration, intensity, quality, frequency, configuration, location, and radiation when considered in the light of the history, general physical examination, and other features of the cardiac examination, as described in **Chap. 234**.

FIGURE 231-1

Approach to the evaluation of a heart murmur. ECG, electrocardiogram. *(Reproduced with permission from Primary Cardiology, 2nd ed, E Braunwald, L Goldman [eds]. Philadelphia, Saunders, 2003.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The majority of heart murmurs are midsystolic and soft (grades I–II/VI). When such a murmur occurs in an asymptomatic child or young adult *without* other evidence of heart disease on clinical examination, it is usually benign and echocardiography generally is not required. By contrast, two-dimensional and Doppler echocardiography **(Chap. 236)** are indicated in patients with loud systolic murmurs (grades ≥III/VI), especially those that are holosystolic or late systolic, and in most patients with diastolic or continuous murmurs.

PITFALLS IN CARDIOVASCULAR MEDICINE

Increasing subspecialization in internal medicine and the perfection of advanced diagnostic techniques in cardiology can lead to several undesirable consequences. Examples include the following:

1. Failure by the *noncardiologist* to recognize important cardiac manifestations of systemic illnesses. For example, the presence of mitral stenosis, patent foramen ovale, and/or transient atrial arrhythmia should be considered in a patient with stroke, or the presence of pulmonary hypertension and cor pulmonale should be considered in a patient with scleroderma or Raynaud's syndrome. A cardiovascular examination should be carried out to identify and

estimate the severity of the cardiovascular involvement that accompanies many noncardiac disorders.

2. Failure by the *cardiologist* to recognize underlying systemic disorders in patients with heart disease. For example, hyperthyroidism should be considered in an elderly patient with atrial fibrillation and unexplained heart failure, and Lyme disease should be considered in a patient with unexplained fluctuating atrioventricular block. A cardiovascular abnormality may provide the clue critical to the recognition of some systemic disorders. For example, an unexplained pericardial effusion may provide an early clue to the diagnosis of tuberculosis or a neoplasm.

3. Overreliance on and overutilization of laboratory tests, particularly invasive techniques, for the evaluation of the cardiovascular system. Cardiac catheterization and coronary arteriography **(Chap. 237)** provide precise diagnostic information that may be crucial in developing a therapeutic plan in patients with known or suspected CAD. Although a great deal of attention has been directed to these examinations, it is important to recognize that they serve to *supplement*, not *supplant*, a careful examination carried out with clinical and noninvasive techniques. A coronary arteriogram should not be performed in lieu of a careful history in patients with chest pain suspected of having ischemic heart disease. Although coronary arteriography may establish whether the coronary arteries are obstructed and to what extent, the results of the procedure by themselves often do not provide a definitive answer to the question of whether a patient's complaint of chest discomfort is attributable to coronary atherosclerosis and whether or not revascularization is indicated.

Despite the value of invasive tests in certain circumstances, they entail some small risk to the patient, involve discomfort and substantial cost, and place a strain on medical facilities. Therefore, they should be carried out only if the results can be expected to modify the patient's management.

DISEASE PREVENTION AND MANAGEMENT

The prevention of heart disease, especially of CAD, is one of the most important tasks of primary health care givers as well as cardiologists. Prevention begins with risk assessment, followed by attention to lifestyle, such as achieving optimal weight, physical activity, and smoking cessation, and then aggressive treatment of all abnormal risk factors, such as hypertension, hyperlipidemia, and diabetes mellitus **(Chap. 396)**.

After a complete diagnosis has been established in patients with known heart disease, a number of management options are usually available. Several examples may be used to demonstrate some of the principles of cardiovascular therapeutics:

1. In the absence of evidence of heart disease, the patient should be clearly informed of this assessment and *not* be asked to return at intervals for repeated examinations. If there is no evidence of disease, such continued attention may lead to the patient's developing inappropriate concern about the possibility of heart disease.

2. If there is no evidence of cardiovascular disease but the patient has one or more risk factors for the development of ischemic heart disease **(Chap. 267)**, a plan for their reduction should be developed and the patient should be retested at intervals to assess compliance and efficacy in risk reduction.

3. Asymptomatic or mildly symptomatic patients with valvular heart disease that is anatomically severe should be evaluated periodically, every 6 to 12 months, by clinical and noninvasive examinations. Early signs of deterioration of ventricular function may signify the need for surgical treatment before the development of disabling symptoms, irreversible myocardial damage, and excessive risk of surgical treatment **(Chap. 256)**.

4. In patients with CAD **(Chap. 267)**, available practice guidelines should be considered in the decision on the form of treatment (medical, percutaneous coronary intervention, or surgical revascularization). Mechanical revascularization may be employed too frequently in the United States and too infrequently in Eastern Europe and developing nations. The mere presence of angina pectoris and/or the demonstration of critical coronary arterial narrowing at angiography should not reflexively evoke a decision to treat the patient by revascularization. Instead, these interventions should be limited to patients with CAD whose angina has not responded adequately to medical treatment or in whom revascularization has been shown to improve the natural history (e.g., acute coronary syndrome or multivessel CAD with left ventricular dysfunction).

Chapter 232: Basic Biology of the Cardiovascular System

Joseph Loscalzo; Peter Libby; Calum A. MacRae

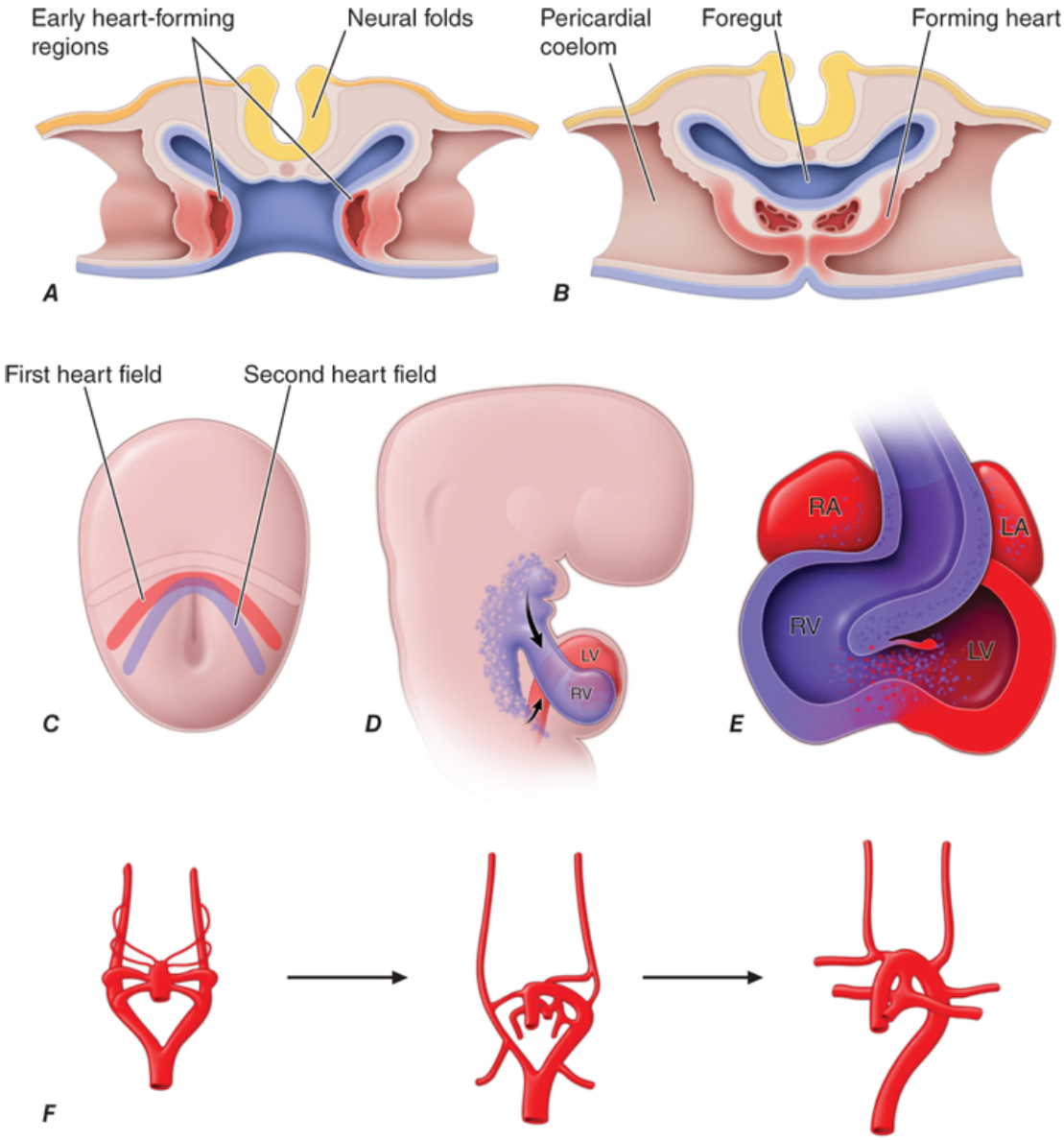
DEVELOPMENTAL BIOLOGY OF THE CARDIOVASCULAR SYSTEM

The heart forms early during embryogenesis **(Fig. 232-1)**, circulating blood, nutrients, and oxygen to the other developing organs while continuing to grow and undergo complex morphogenetic changes. Early cardiac progenitors arise within crescent-shaped fields of lateral splanchnic mesoderm under the influence of multiple signals and migrate to the midline to form the linear heart tube: a single layer of endocardium and a single layer of cardiomyocyte precursors.

FIGURE 232-1

A. Schematic depiction of a transverse section through an early embryo depicts the bilateral regions where early heart tubes form. ***B.*** The bilateral heart tubes subsequently migrate to the midline and fuse to form the linear heart tube. ***C.*** At the early cardiac crescent stage of embryonic development, cardiac precursors include a primary heart field fated to form the linear heart tube and a second heart field fated to add myocardium to the inflow and outflow poles of the heart. ***D.*** Second heart field cells populate the pharyngeal region before subsequently migrating to the maturing heart. ***E.*** Large portions of the right ventricle and outflow tract and some cells within the atria derive from the second heart field. ***F.*** The aortic arch arteries form as symmetric sets of vessels that then remodel under the influence of the neural crest to form the asymmetric mature vasculature. LA, left atrium; LV, left ventricle; RA, right atrium; RV, right ventricle.

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The linear heart tube undergoes asymmetric looping, that coordinates with chamber specification and multilayer growth of different regions of the heart tube to produce the presumptive atria and ventricles. Cells continue to migrate into the heart at both ends from later, or second, heart fields in pharyngeal mesoderm as looping and growth occur. These cells exhibit distinctive gene expression (e.g., Islet-1) and distinctive physiology (e.g., calcium handling), contributing to discrete areas of the adult heart, including the right atrium and the right ventricle. Different embryologic origins of cells within the right and left ventricles help explain why some forms of congenital and adult heart diseases affect regions of the heart to varying degrees.

After looping and chamber formation, a series of morphogenetic events divide the left and right sides of the heart, separate the atria from the ventricles, and form the aorta and pulmonary artery from the truncus arteriosus. Cardiac valves form between the atria and the ventricles and between the ventricles and the outflow vessels. Early in development, the single layer of myocardial cells secretes an extracellular matrix rich in hyaluronic acid, or “cardiac jelly,” which accumulates within the endocardial cushions, precursors of the cardiac valves. Signals from overlying myocardial cells trigger migration, invasion, and phenotypic changes in underlying endocardial cells, which undergo an epithelial-mesenchymal transformation to invade and populate the endocardial cushion matrix with cells. Mesenchymal cells then proliferate and form the mature valve leaflets.

The great vessels form as a series of bilaterally symmetric aortic arch arteries that remodel asymmetrically to define the mature central vasculature. Migrating neural crest cells from the dorsal neural tube orchestrate this process and are necessary for aortic arch remodeling and the septation of the truncus arteriosus. The smooth-muscle cells within the tunica media of the aortic arch, the ductus arteriosus, and the carotid arteries all derive from neural crest. By contrast, smooth-muscle within the descending aorta arises from lateral plate mesoderm, and smooth muscle of the proximal outflow tract arises from the second heart field. Neural crest cells are sensitive to both vitamin A and folic acid, and congenital heart disease involving abnormal remodeling of the aortic arch arteries can associate with maternal deficiencies of these vitamins. The shared embryologic origins of different cardiovascular cell types lead to syndromic associations between various congenital heart diseases and a range of extracardiac abnormalities.

Loading [Contrib]/a11y/accessibility-menu.js

Coronary artery formation requires the addition of yet another cell population to the embryonic heart. Epicardial cells arise in the proepicardial organ, a derivative of the septum transversum, which also contributes to the fibrous portion of the diaphragm and to the liver. Proepicardial cells contribute smooth-muscle to the coronary arteries and are required for proper coronary patterning. Other cell types within the heart, (e.g., fibroblasts) also can arise from the proepicardium.

The cardiac conduction system, which generates and propagates electrical impulses, differentiates from cardiomyocyte precursors. The conduction system is composed of slow-conducting (proximal) components, such as the sinoatrial (SA) and atrioventricular (AV) nodes, as well as fast-conducting (distal) components, including the His bundle, bundle branches, and Purkinje fibers. Precursors within the sinus venosus give rise to the SA node, whereas those within the AV canal mature into heterogeneous cell types that compose the AV node. So-called decremental conduction through the AV node delays the electrical impulses between atria and ventricles, whereas the distal conduction system rapidly delivers the impulse throughout the ventricles. Each compartment within the conduction system expresses distinct gap junction proteins and ion channels that characterize the discrete cell fates and electrical properties. Developmental defects in the conduction system can lead to clinical electrophysiologic disorders, such as congenital heart block or pre-excitation (Wolff-Parkinson-White syndrome) **(Chap. 241)**.

ORIGIN OF VASCULAR CELLS

As noted above, smooth-muscle cells in various types of artery derive from different sources. Some upper-body arterial smooth-muscle cells derive from the neural crest, whereas lower-body arteries generally recruit smooth-muscle cells from neighboring mesodermal structures during development. Bone marrow–derived endothelial progenitors may aid repair of damaged or aging arteries. In addition, multipotent vascular stem cells resident in vessel walls may give rise to the smooth-muscle cells that accumulate in injured or atheromatous arteries **(Chaps. 92 and 473)**.

THE BLOOD VESSEL

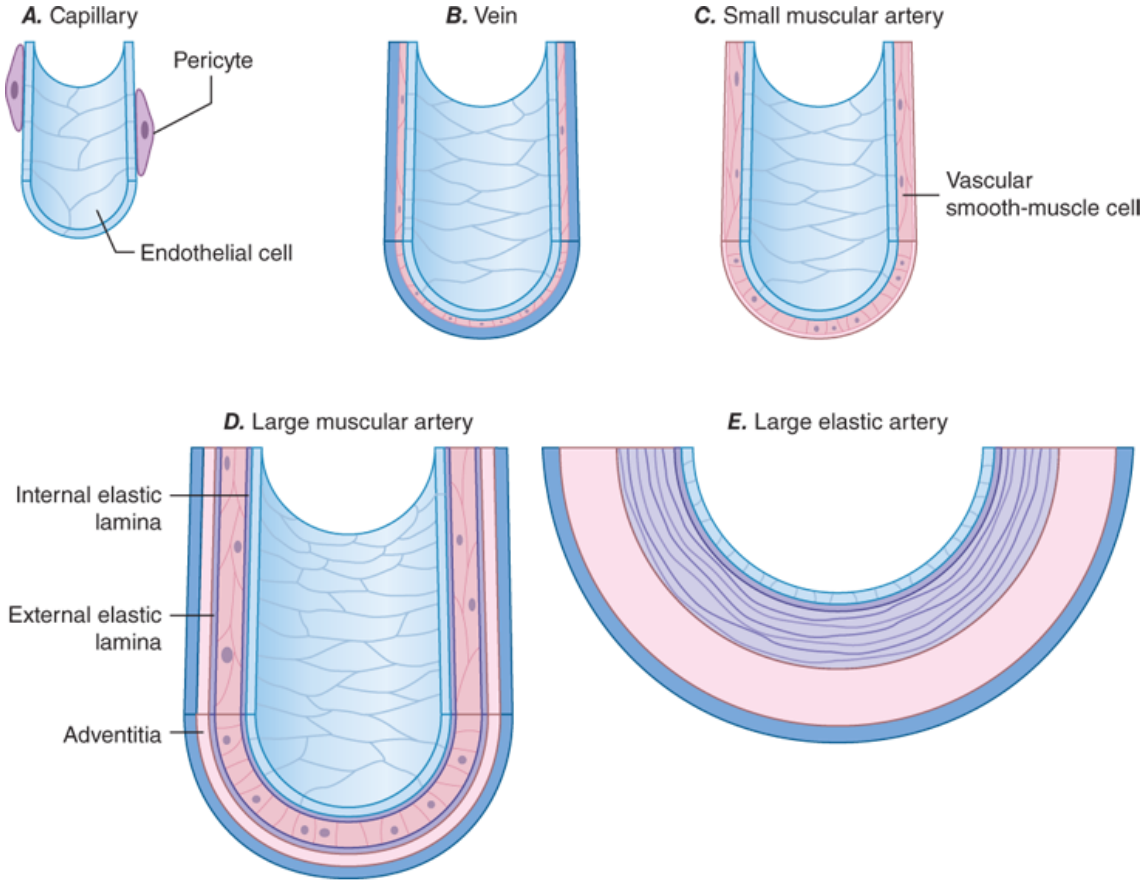
VASCULAR ULTRASTRUCTURE

Blood vessels participate in physiologic function and play roles in disease biology in virtually every organ system. The smallest blood vessels—capillaries—consist of a monolayer of endothelial cells on a basement membrane, adjacent to a discontinuous layer of smooth-muscle-like cells known as *pericytes* **(Fig. 232-2A)**. Arteries typically have a trilaminar structure **(Fig. 232-2B–E)**. The *intima* consists of a monolayer of endothelial cells continuous with those of the capillaries. The middle layer, or *tunica media*, consists of smooth-muscle cells, in veins, the media can contain just a few layers of smooth-muscle cells (Fig. 232-2B). The outer layer, the *adventitia*, consists of looser extracellular matrix with fibroblasts, mast cells, and nerve terminals. Larger arteries have their own vasculature, the *vasa vasorum*, which nourishes the tunica media.

FIGURE 232-2

Schematics of the structures of various types of blood vessels. *A*. Capillaries consist of an endothelial tube in contact with a discontinuous population of pericytes. ***B*.** Veins typically have thin medias and thicker adventitias. ***C*.** A small muscular artery features a prominent tunica media. ***D*.** Larger muscular arteries have a prominent media with smooth-muscle cells embedded in a complex extracellular matrix. ***E*.** Larger elastic arteries have cylindrical layers of elastic tissue alternating with concentric rings of smooth-muscle cells.

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Arteriolar muscle tone regulates blood pressure and flow through arterial beds (Fig. 232-2C). Medium-size muscular arteries also contain prominent smooth muscle layers (Fig. 232-2D) that participate in atherosclerosis. Larger elastic arteries have a highly structured tunica media with concentric bands of smooth-muscle cells, interspersed with strata of elastin-rich extracellular matrix (Fig. 232-2E). Larger arteries form an internal elastic lamina between intima and media while an external elastic lamina partitions media from surrounding adventitia.

VASCULAR CELL BIOLOGY

Endothelial Cell

The endothelium forms the interface between tissues and the blood compartment, regulating the passage of molecules and cells. The ability of endothelial cells to serve as a selectively permeable barrier fails in vascular diseases, including atherosclerosis, hypertension, and renal disease, as well as in pulmonary edema, sepsis and other situations of “capillary leak.”

The endothelium also participates in the local regulation of vascular tone and blood flow. Endogenous substances produced by endothelial cells such as prostacyclin, endothelium-derived hyperpolarizing factor, nitric oxide (NO), and hydrogen peroxide (H_2O_2) provide tonic vasodilatory stimuli under physiologic conditions in vivo **(Table 232-1)**. Impaired production or excess catabolism of NO impairs endothelium-dependent vasodilator function contributing to pathologic vasoconstriction. Measurement of flow-mediated dilatation can assess endothelial vasodilator function in humans **(Fig. 232-3)**. Endothelial cells also produce potent vasoconstrictor substances such as endothelin. Excessive production of reactive oxygen species, such as superoxide anion (O_2^-), by endothelial or smooth-muscle cells under pathologic conditions (e.g., excessive exposure to angiotensin II), can promote local oxidative stress and inactivate NO.

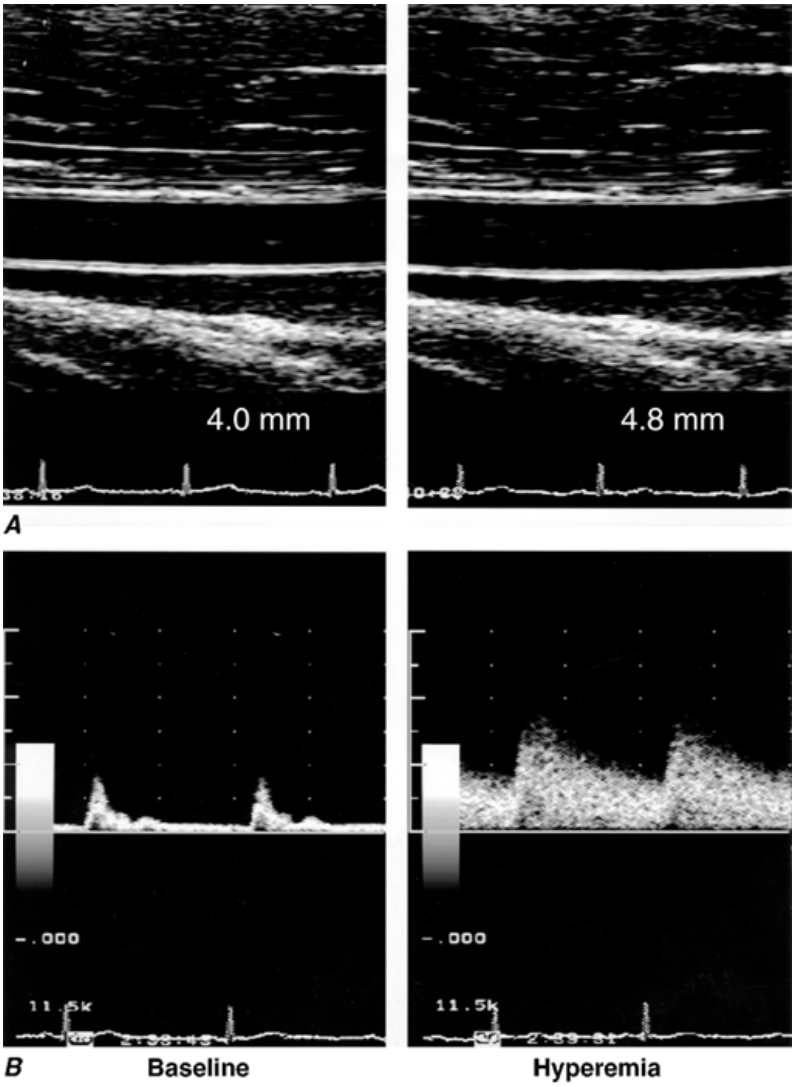
TABLE 232-1 Endothelial Functions in Health and Disease

Homeostatic Properties	Dysfunctional Properties
Optimize balance between vasodilation and vasoconstriction	Impaired dilation, vasoconstriction
Antithrombotic, profibrinolytic	Prothrombotic, antifibrinolytic
Anti-inflammatory	Proinflammatory
Antiproliferative	Proproliferative
Antioxidant	Prooxidant
	Impaired barrier function

Loading [Contrib]/a11y/accessibility-menu.js

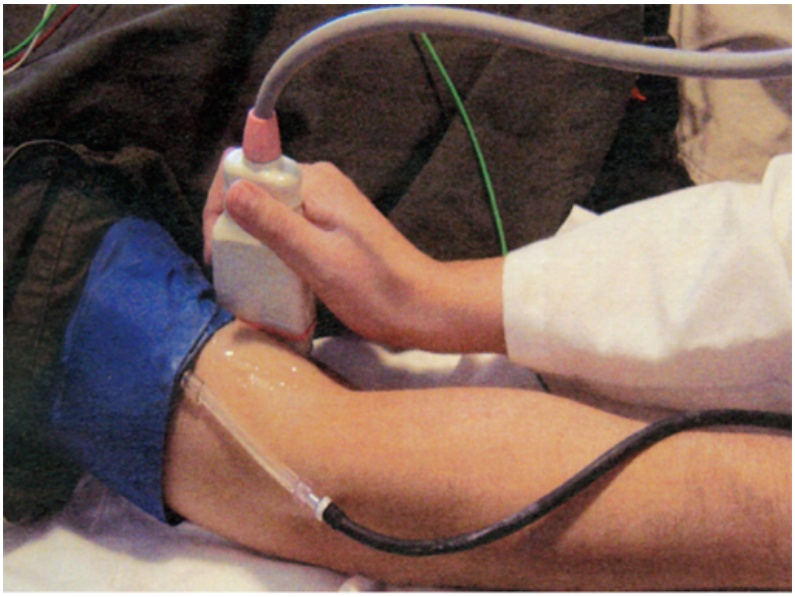
FIGURE 232-3

Assessment of endothelial function in vivo using blood pressure cuff occlusion and release. Upon deflation of the cuff, an ultrasound probe monitors changes in diameter (***A***) and blood flow (***B***) of the brachial artery (***C***). *(Reproduced with permission of J. Vita, MD.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Loading [Contrib]/a11y/accessibility-menu.js



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Normal endothelium exhibits limited interaction with circulating leukocytes, but when activated by bacterial products such as endotoxin or by proinflammatory cytokines released during infection or injury, endothelial cells express an array of adhesion molecules that selectively bind various classes of leukocytes in different pathologic conditions. The adhesion molecules and chemokines generated during acute bacterial infection tend to recruit granulocytes, while in chronic inflammatory diseases such as tuberculosis or atherosclerosis, the adhesion molecules expressed favor monocyte recruitment. Endothelial cells participate in the pathophysiology of many immune-mediated diseases. Complement-mediated lysis of endothelial cells is an example of immunologically mediated tissue injury. The presentation of foreign histocompatibility complex antigens by endothelial cells in solid-organ allografts can promote allograft arteriopathy, while immune-mediated endothelial injury also plays a role in thrombotic thrombocytopenic purpura or hemolytic-uremic syndrome.

The endothelium also regulates the balance between thrombosis and hemostasis through a highly tuned set of regulatory pathways. When activated by inflammatory cytokines, bacterial endotoxin, or angiotensin II, for example, endothelial cells can produce substantial quantities of the major inhibitor of fibrinolysis, plasminogen activator inhibitor 1 (PAI-1). Thus, in pathologic circumstances, the endothelial cell may promote local thrombus accumulation rather than combat it. Inflammatory stimuli also induce the expression of the potent procoagulant tissue factor, a contributor to disseminated intravascular coagulation in sepsis.

Endothelial cells regulate the growth of subjacent smooth-muscle cells. For example, heparan sulfate glycosaminoglycans elaborated by endothelial cells can inhibit smooth-muscle proliferation, and in the setting of injury endothelial cells produce growth factors and chemoattractants, such as platelet-derived growth factor, which cause the migration and proliferation of vascular smooth-muscle cells. Dysregulation of these growth-stimulatory molecules may promote smooth-muscle accumulation in atherosclerotic lesions.

Vascular Smooth-Muscle Cell

Contraction and relaxation of vascular smooth-muscle cells in muscular arteries determines blood pressure, regional flow and the afterload experienced by the left ventricle (see below). Venous tone regulates the capacitance of the venous tree and so influences ventricular preload. Smooth-muscle cells in the adult vessel seldom replicate in the absence of arterial injury or inflammatory activation, but proliferation and migration of arterial smooth-muscle cells contributes to arterial stenoses in atherosclerosis, arteriolar remodeling in hypertension, and the hyperplastic response of arteries injured by percutaneous intervention. In the pulmonary circulation, smooth-muscle migration and proliferation underlie the vascular disease that occurs in sustained high-flow states such as left-to-right shunts in congenital heart disease.

Smooth-muscle cells secrete the bulk of vascular extracellular matrix. Excessive production of collagen and glycosaminoglycans contributes to the remodeling, altered biomechanics and physiology of arteries affected by hypertension or atherosclerosis. In larger elastic arteries, such as the aorta, the ability to store the kinetic energy of systole promotes tissue perfusion during diastole. Arterial stiffness associated with aging or disease, evident in a widening pulse pressure, increases left ventricular afterload and portends a poor outcome.

Like endothelial cells, vascular smooth-muscle cells do not merely respond to vasomotor or inflammatory stimuli elaborated by other cell types, but can themselves serve as a source of such stimuli. For example, when exposed to proinflammatory stimuli, smooth-muscle cells elaborate cytokines and other mediators which drive thrombosis and fibrinolysis as well as proliferation.

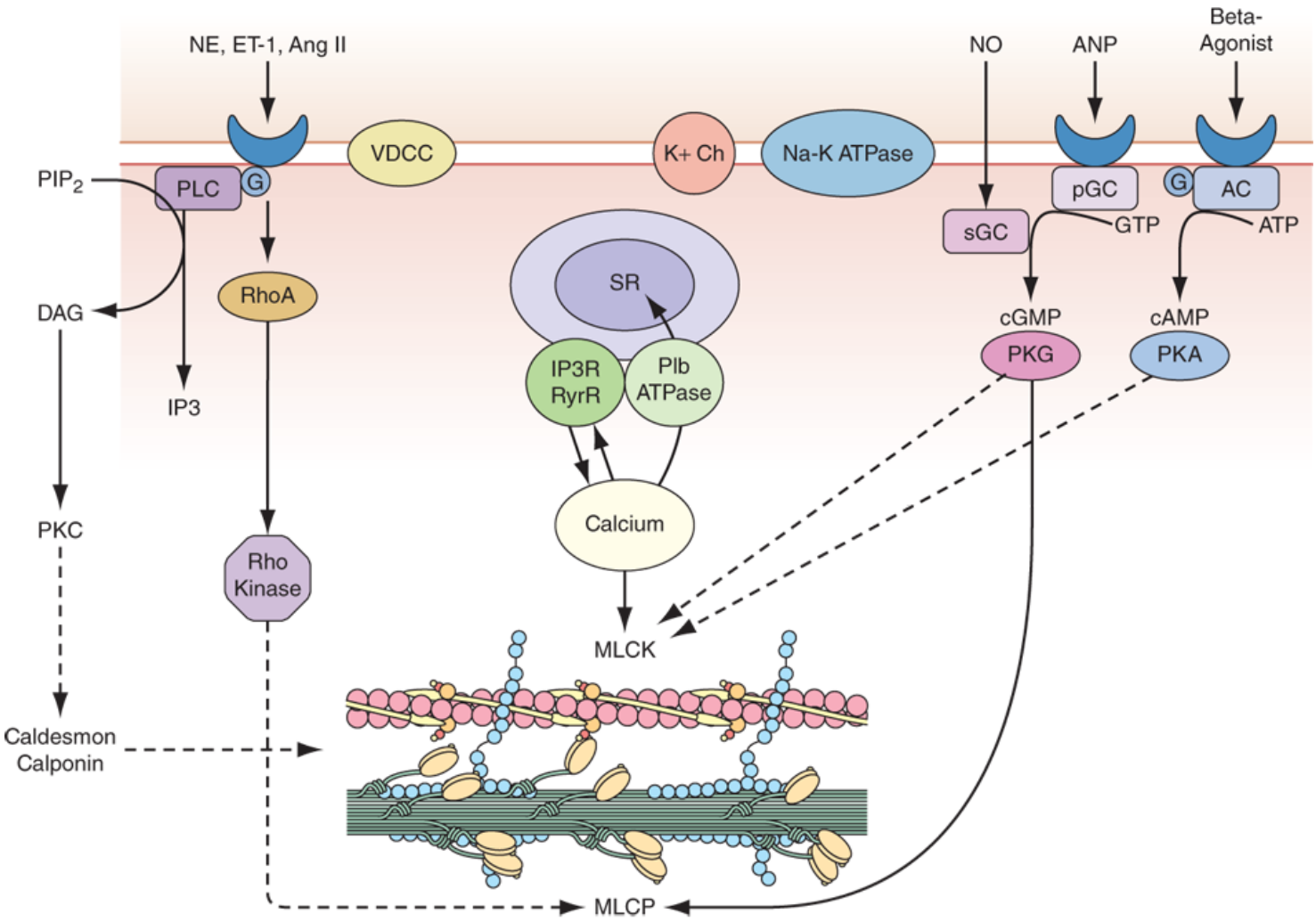
Vascular Smooth-Muscle Cell Contraction

Loading [Contrib]/a11y/accessibility-menu.js

Vascular smooth-muscle cells contract as cytoplasmic calcium concentration rises due to transmembrane influx and triggered release from intracellular calcium stores (**Fig. 232-4**). In vascular smooth-muscle cells, voltage-dependent L-type calcium channels open with membrane depolarization. Local changes in intracellular calcium concentration, termed *calcium sparks*, can trigger release from intracellular stores which results in more contraction and higher vessel tone (see below). Opposing currents balance the effects of individual ionic fluxes, promoting homeostasis which is tightly regulated by neural and metabolic influences.

FIGURE 232-4

Regulation of vascular smooth-muscle cell calcium concentration and actomyosin ATPase-dependent contraction. AC, adenylyl cyclase; Ang II, angiotensin II; ANP, atrial natriuretic peptide; DAG, diacylglycerol; ET-1, endothelin-1; G, G protein; IP_3, inositol 1,4,5-trisphosphate; MLCK, myosin light chain kinase; MLCP, myosin light chain phosphatase; NE, norepinephrine; NO, nitric oxide; pGC, particular guanylyl cyclase; PIP_2, phosphatidylinositol 4,5-bisphosphate; PKA, protein kinase A; PKC, protein kinase C; PKG, protein kinase G; PLC, phospholipase C; sGC, soluble guanylyl cyclase; SR, sarcoplasmic reticulum; VDCC, voltage-dependent calcium channel. *(Modified from B Berk, in Vascular Medicine, 3rd ed. Philadelphia, Saunders, Elsevier, 2006, p. 23; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Biochemical agonists also increase intracellular [Ca^{2+}] by various mechanisms including receptor-dependent phospholipase C activation with hydrolysis of phosphatidylinositol 4,5-bisphosphate to generate diacylglycerol (DAG) and inositol 1,4,5-trisphosphate (IP_3). These membrane lipid derivatives in turn activate protein kinase C and increase intracellular [Ca^{2+}]. In addition, IP_3 binds specific sarcoplasmic reticulum (SR) receptors to increase calcium efflux from this storage pool into the cytoplasm.

Vascular smooth-muscle cell contraction depends on myosin light chain phosphorylation, which in the steady state reflects the balance between the actions of the relevant kinases and phosphatases. Calcium activates myosin light chain kinase via calmodulin, augmenting myosin ATPase activity enhancing contraction. Myosin light chain phosphatase conversely reduces myosin ATPase activity and contractile force. Other kinase/phosphorylase combinations result in a complex regulatory network that refines vascular tone and links it to physiologic requirements.

Loading [Contrib]/a11y/accessibility-menu.js

Control of Vascular Smooth-Muscle Cell Tone

The autonomic nervous system and endothelial cells modulate vascular smooth-muscle cells through similar convergent pathways. Autonomic neurons enter vessel media and modulate vascular smooth-muscle cell tone in response to baroreceptors and chemoreceptors within the aortic arch or carotid bodies and to thermoreceptors in the skin. Rapidly acting reflex arcs modulated by central inputs respond to multiple sensory inputs as well as emotional stimuli through three neuronal classes: *sympathetic*, whose principal neurotransmitters are epinephrine and norepinephrine; *parasympathetic*, whose principal neurotransmitter is acetylcholine; and *nonadrenergic/noncholinergic*, which include two subgroups—nitrergic, whose principal neurotransmitter is NO, and peptidergic, whose principal neurotransmitters are substance P, vasoactive intestinal peptide, calcitonin gene-related peptide, as well as a non-peptide, adenosine triphosphate (ATP).

Each of these neurotransmitters acts through specific receptors on the vascular smooth-muscle cell to modulate intracellular Ca^{2+} and consequently, contractile tone. Norepinephrine activates α adrenergic receptors, and epinephrine activates both α and β receptors; in most blood vessels, norepinephrine activates postjunctional α_1 receptors in large arteries and α_2 receptors in small arteries and arterioles, leading to vasoconstriction. Most blood vessels express β_2-adrenergic receptors on their vascular smooth-muscle cells and respond to β agonists by cyclic AMP–dependent relaxation. Acetylcholine released from parasympathetic neurons binds to muscarinic receptors on vascular smooth-muscle cells causing vasorelaxation. Nitrergic neurons release NO, which relaxes vascular smooth-muscle cell via the cyclic GMP–dependent and –independent mechanisms outlined, and other peptidergic inputs that regulate vascular tone. **For the detailed molecular physiology of the autonomic nervous system, see Chap. 432.**

The release of endothelial effectors of vascular smooth-muscle cell tone (Figs. 232-2 and 232-3) integrates mechanical (shear stress, cyclic strain, etc.) and biochemical stimuli (purinergic agonists, muscarinic agonists, peptidergic agonists). In addition to these local paracrine modulators, a complex system of circulating modulators ranging from norepinephrine to the natriuretic peptides also modulate vascular smooth-muscle cell tone.

VASCULAR REGENERATION

Growth of new blood vessels can occur in response to conditions such as chronic hypoxemia and tissue ischemia. Growth factors, including vascular endothelial growth factor (VEGF) and forms of fibroblast growth factor (FGF), activate a signaling cascade that stimulates endothelial proliferation and tube formation, defined as *angiogenesis*. Guidance molecules, including members of the semaphorin family of secreted peptides, direct blood vessel patterning by attracting or repelling nascent endothelial tubes. The development of collateral vascular networks in the ischemic myocardium, an example of angiogenesis, can result from selective activation of local or circulating endothelial progenitor cells. True arteriogenesis, or the development of a new blood vessel that includes all three cell layers, normally does not occur in adult mammals, but recent scientific advances might help obviate such limitations **(Chaps. 92 and 473)**.

CELLULAR BASIS OF CARDIAC CONTRACTION

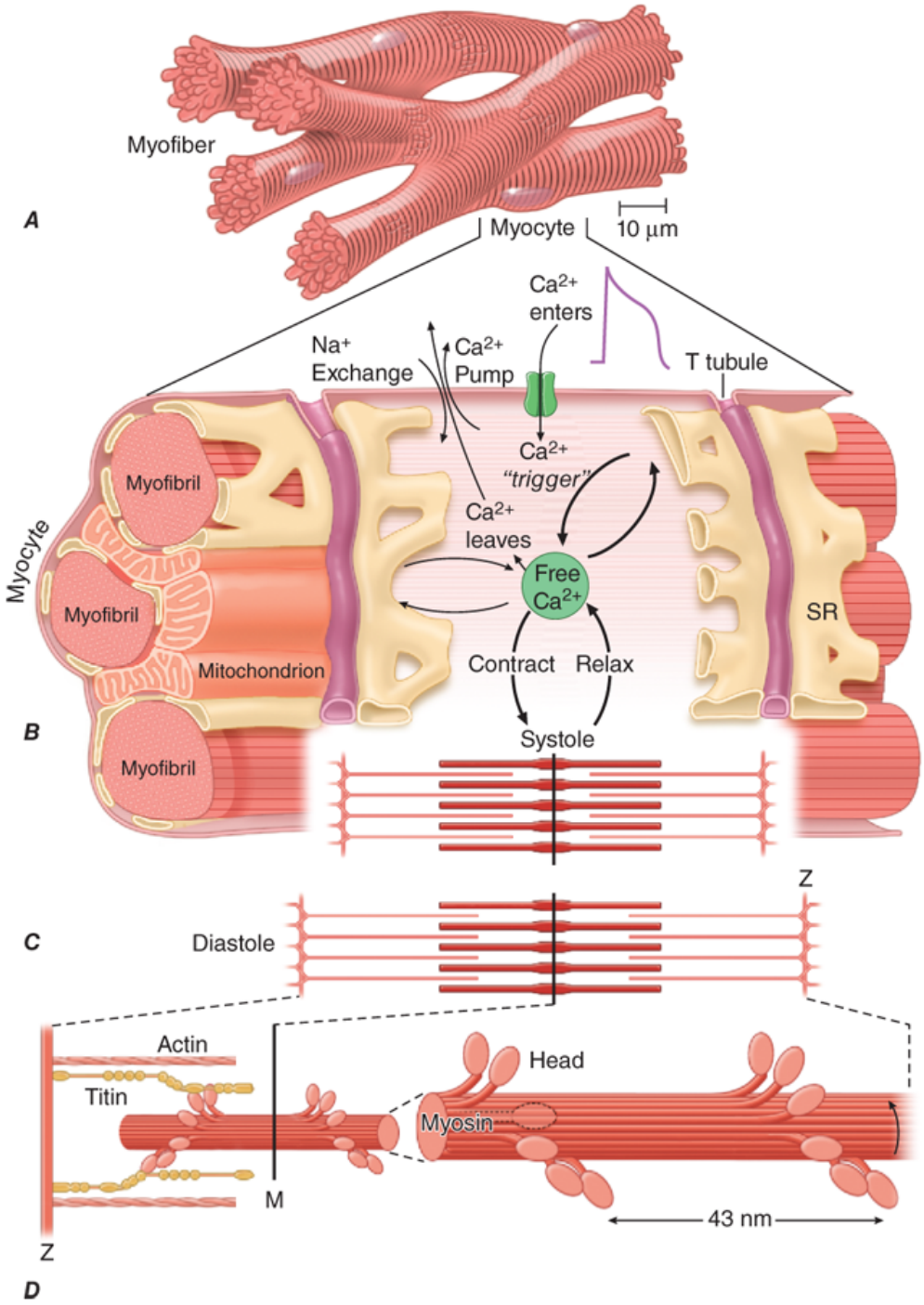
CARDIAC ULTRASTRUCTURE

Most of the ventricular mass is composed of cardiomyocytes, normally 60–140 μm in length and 17–25 μm in diameter **(Fig. 232-5A)**. Each cell contains multiple myofibrils that run the length of the cell and are composed of series of repeating sarcomeres. The cytoplasm between the myofibrils contains other cell constituents, including a single centrally located nucleus, mitochondria, and the intracellular membrane system, the SR.

FIGURE 232-5

A shows the branching myocytes making up the cardiac myofibers. ***B*** illustrates the critical role played by the changing $[Ca^{2+}]$ in the myocardial cytosol. Ca^{2+} ions are schematically shown as entering through the calcium channel that opens in response to the wave of depolarization that travels along the sarcolemma. These Ca^{2+} ions "trigger" the release of more calcium from the sarcoplasmic reticulum (SR) and thereby initiate a contraction-relaxation cycle. Eventually the small quantity of Ca^{2+} that has entered the cell leaves predominantly through an Na^+/Ca^{2+} exchanger, with a lesser role for the sarcolemmal Ca^{2+} pump. The varying actin-myosin overlap is shown for (***B***) systole, when $[Ca^{2+}]$ is maximal, and (***C***) diastole, when $[Ca^{2+}]$ is minimal. ***D.*** The myosin heads, attached to the thick filaments, interact with the thin actin filaments. *(From LH Opie: Heart Physiology: From Cell to Circulation, 4th ed. Philadelphia, Lippincott, Williams & Wilkins, 2004. Reprinted with permission. Copyright LH Opie, 2004.)*

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The *sarcomere*, the structural and functional unit of contraction, lies between adjacent Z lines, which are dark repeating bands apparent on transmission electron microscopy. The distance between Z lines varies with the degree of contraction or stretch of the muscle and ranges between 1.6 and 2.2 µm. At the center of the sarcomere is a dark band of constant length (1.5 µm), the A band, which is flanked by two lighter bands, the I bands, which are of variable length. The sarcomere of heart muscle, like that of skeletal muscle, consists of interdigitating thick and thin myofilaments. Thicker filaments, composed principally of the protein myosin, traverse the A band; they are about 10 nm (100 Å) in diameter, with tapered ends. Thinner filaments, composed primarily of actin, course from the Z lines through the I band into the A band; they are ~5 nm (50 Å) in diameter and 1.0 µm in length. Thus, thick and thin filaments overlap only within the (dark) A band, whereas the (light) I band contains only thin filaments. On electron-microscopic examination, bridges extend between the thick and thin filaments within the A band; these are myosin heads (see below) bound to actin filaments.

THE CONTRACTILE PROCESS

The sliding filament model for muscle contraction rests on the central observation that both the thick and the thin filaments are c[illegible]ion and relaxation. With activation, the actin filaments are propelled farther into the A band. In the process, the A band remains constant in length, whereas the I band shortens and the Z lines move toward one another.

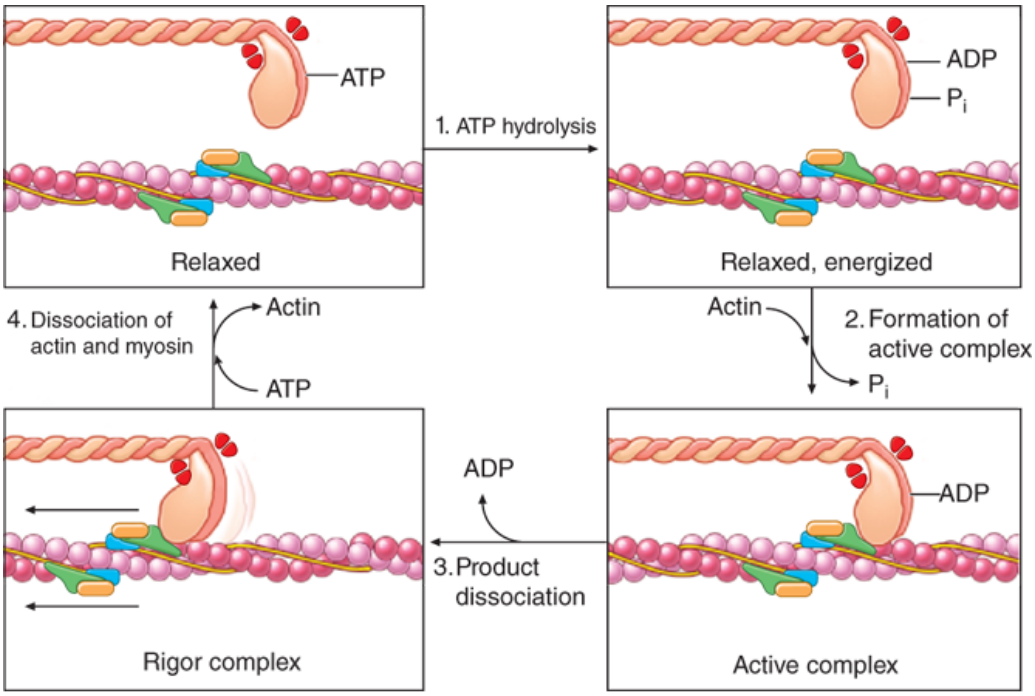
Loading [Contrib]/a11y/accessibility-menu.js

The *myosin* molecule is a complex, asymmetric protein with a molecular mass of about 500,000 Da; it has a rod-like portion that is about 150 nm (1500 Å) in length with a globular portion (head) at its end. The globular portions of myosin form the bridges to actin and are the site of ATPase activity. In thick myofilaments, composed of ~300 longitudinally stacked myosin molecules, the rod-like segments of myosin assume an orderly, polarized manner, with outwardly projecting globular heads interacting with actin to generate force and shorten **(Fig. 232-5B)**.

Actin has a molecular mass of about 47,000 Da. Thin filaments consist of a double helix of two chains of actin molecules wound about each other on a larger molecule, tropomyosin. A group of regulatory proteins—troponins C, I, and T—localize at regular intervals on this filament **(Fig. 232-6)**. In contrast to myosin, actin lacks intrinsic enzymatic activity, but combines reversibly with myosin in the presence of ATP and Ca^{2+}. Calcium activates the myosin ATPase, which breaks down ATP to supply the energy for contraction (Fig. 232-6). The activity of myosin ATPase determines the rate of actomyosin cross-bridge formation and breakdown, and ultimately determines contraction velocity. In relaxed muscle, tropomyosin inhibits this interaction. *Titin* (Fig. 232-5D) an enormous, flexible, myofibrillar protein, connects myosin to the Z line; its stretching contributes to the elasticity of the heart. Dystrophin, a long cytoskeletal protein that binds to the dystroglycan complex at adherens junctions on the cell membrane, tethers the sarcomere to the cell membrane at regions tightly coupled to adjacent contracting myocytes. Mutations in multiple sarcomeric and cytoskeletal proteins cause different forms of inherited disease involving the heart and skeletal muscle.

FIGURE 232-6

Four steps in cardiac muscle contraction and relaxation. In relaxed muscle (*upper left*), ATP bound to the myosin cross-bridge dissociates the thick and thin filaments. ***Step 1:*** Hydrolysis of myosin-bound ATP by the ATPase site on the myosin head transfers the chemical energy of the nucleotide to the activated cross-bridge (*upper right*). When cytosolic Ca^{2+} concentration is low, as in relaxed muscle, the reaction cannot proceed because tropomyosin and the troponin complex on the thin filament do not allow the active sites on actin to interact with the cross-bridges. Therefore, even though the cross-bridges are energized, they cannot interact with actin. ***Step 2:*** When Ca^{2+} binding to troponin C has exposed active sites on the thin filament, actin interacts with the myosin cross-bridges to form an active complex (*lower right*) in which the energy derived from ATP is retained in the actin-bound cross-bridge, whose orientation has not yet shifted. ***Step 3:*** The muscle contracts when ADP dissociates from the cross-bridge. This step leads to the formation of the low-energy rigor complex (*lower left*) in which the chemical energy derived from ATP hydrolysis has been expended to perform mechanical work (the "rowing" motion of the cross-bridge). ***Step 4:*** The muscle returns to its resting state, and the cycle ends when a new molecule of ATP binds to the rigor complex and dissociates the cross-bridge from the thin filament. This cycle continues until calcium is dissociated from troponin C in the thin filament, which causes the contractile proteins to return to the resting state with the cross-bridge in the energized state. ADP, adenosine diphosphate; ATP, adenosine triphosphate; ATPase, adenosine triphosphatase. *(From AM Katz: Heart failure: Cardiac function and dysfunction, in Atlas of Heart Diseases, 3rd ed, WS Colucci [ed]. Philadelphia, Current Medicine, 2002. Reprinted with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

During activation of the cardiac myocyte, Ca^{2+} binds the heterotrimer troponin C, resulting in conformational changes in the regulatory protein tropomyosin and exposing actin cross-bridge interaction sites (Fig. 232-6). Repetitive interaction between myosin heads and actin filaments is termed *cross-bridge cycling*, and results in sliding of the actin along the myosin filaments, with muscle s[illegible]ension. The splitting of ATP then dissociates the myosin cross-bridge from actin. In the presence of ATP (Fig. 232-6), actin and myosin filaments bind and dissociate cyclically if sufficient Ca^{2+} is present; these linkages

Loading [Contrib]/a11y/accessibility-menu.js

cease when [Ca^{2+}] falls below a critical level, and the troponin-tropomyosin complex once more inhibits actin-myosin interactions **(Fig. 232-7)**.

FIGURE 232-7

Signal systems involved in positive inotropic and lusitropic (enhanced relaxation) effects of β-adrenergic stimulation. When the β-adrenergic agonist interacts with the β receptor, a series of G protein–mediated changes leads to activation of adenylyl cyclase and the formation of cyclic adenosine monophosphate (cAMP). The latter acts via protein kinase A to stimulate metabolism (*left*) and phosphorylate the Ca^{2+} channel protein (*right*). The result is an enhanced opening probability of the Ca^{2+} channel, thereby increasing the inward movement of Ca^{2+} ions through the sarcolemma (SL) of the T tubule. These Ca^{2+} ions release more calcium from the sarcoplasmic reticulum (SR) to increase cytosolic Ca^{2+} and activate troponin C. Ca^{2+} ions also increase the rate of breakdown of adenosine triphosphate (ATP) to adenosine diphosphate (ADP) and inorganic phosphate (Pi). Enhanced myosin ATPase activity explains the increased rate of contraction, with increased activation of troponin C explaining increased peak force development. An increased rate of relaxation results from the ability of cAMP to activate as well the protein phospholamban, situated on the membrane of the SR, that controls the rate of uptake of calcium into the SR. The latter effect explains enhanced relaxation (lusitropic effect). P, phosphorylation; PL, phospholamban; TnI, troponin I. *(Modified from LH Opie: Heart Physiology: From Cell to Circulation, 4th ed. Philadelphia, Lippincott, Williams & Wilkins, 2004. Reprinted with permission. Copyright LH Opie, 2004.)*

Null

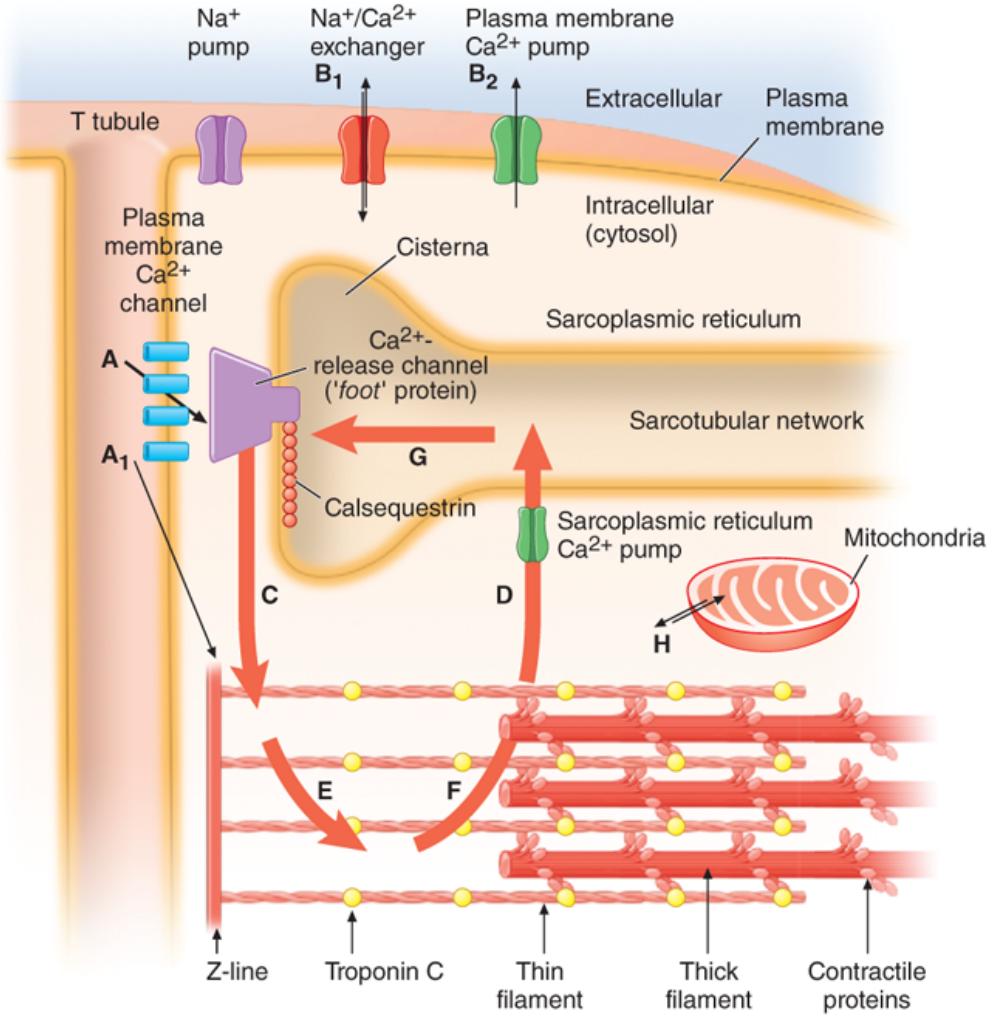
Intracytoplasmic [Ca^{2+}] is a principal determinant of the inotropic state of the heart. Most agents that stimulate myocardial contractility (positive inotropic stimuli), including digitalis glycosides and β-adrenergic agonists, increase cytoplasmic [Ca^{2+}], triggering cross-bridge cycling. Increased adrenergic neuronal activity stimulates myocardial contractility through norepinephrine release, activation of β adrenergic receptors and, via G_s-stimulated guanine nucleotide-binding proteins, activation of the adenylyl cyclase, which leads to the formation of the intracellular second messenger cyclic AMP from ATP (Fig. 232-7). Cyclic AMP in turn activates protein kinase A (PKA), which phosphorylates sarcolemmal Ca^{2+} channels, thereby enhancing the influx of Ca^{2+} into the myocyte.

The SR **(Fig. 232-8)**, a complex network of anastomosing intracellular channels, invests the myofibrils. The transverse tubules, or T system, closely related to the SR, both structurally and functionally, arise from invaginations of the sarcolemma that extend into the myocardial fiber along the Z lines, i.e., the ends of the sarcomeres.

FIGURE 232-8

The Ca^{2+} fluxes and key structures involved in cardiac excitation-contraction coupling. The *arrows* denote the direction of Ca^{2+} fluxes. The *thickness* of each arrow indicates the magnitude of the calcium flux. Two Ca^{2+} cycles regulate excitation-contraction coupling and relaxation. The larger cycle is entirely intracellular and involves Ca^{2+} fluxes into and out of the sarcoplasmic reticulum, as well as Ca^{2+} binding to and release from troponin C. The smaller extracellular Ca^{2+} cycle occurs when this cation moves into and out of the cell. The action potential opens plasma membrane Ca^{2+} channels to allow passive entry of Ca^{2+} into the cell from the extracellular fluid (*arrow A*). Only a small portion of the Ca^{2+} that enters the cell directly activates the contractile proteins (*arrow* A_1). The extracellular cycle is completed when Ca^{2+} is actively transported back out to the extracellular fluid by way of two plasma membrane fluxes mediated by the sodium-calcium exchanger (*arrow* B_1) and the plasma membrane calcium pump (*arrow* B_2). In the intracellular Ca^{2+} cycle, passive Ca^{2+} release occurs through channels in the cisternae (*arrow C*) and initiates contraction; active Ca^{2+} uptake by the Ca^{2+} pump of the sarcotubular network (*arrow D*) relaxes the heart. Diffusion of Ca^{2+} within the sarcoplasmic reticulum (*arrow G*) returns this activator cation to the cisternae, where it is stored in a complex with calsequestrin and other calcium-binding proteins. Ca^{2+} released from the sarcoplasmic reticulum initiates systole when it binds to troponin *C* (*arrow E*). Lowering of cytosolic [Ca^{2+}] by the sarcoplasmic reticulum (SR) causes this ion to dissociate from troponin (*arrow F*) and relaxes the heart. Ca^{2+} also may move between mitochondria and cytoplasm (H). *(Adapted from AM Katz: Physiology of the Heart, 4th ed. Philadelphia, Lippincott, Williams & Wilkins, 2005, with permission.)*

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CARDIAC ACTIVATION

In the inactive state, the cardiac cell is electrically polarized; i.e., the interior has a negative charge relative to the outside of the cell, with a transmembrane potential of –80 to –100 mV **(Chap. 238)**. The sarcolemma, which in the resting state is largely impermeable to Na^+, and a Na^+- and K^+- pump energized by ATP extrudes Na^+ from the cell, and maintains the resting potential. In the resting state, intracellular [K^+] is relatively high and [Na^+] is far lower; conversely, extracellular [Na^+] is high and [K^+] is low. At the same time, in the resting state, extracellular [Ca^{2+}] greatly exceeds free intracellular [Ca^{2+}].

The action potential has four phases **(see Fig. 238-1B)**. During the action potential plateau (phase 2), there is a slow inward current through sarcolemmal L-type Ca^{2+} channels (Fig. 232-8). Depolarizing current spreads across the cell membrane, penetrating deeply into the cell via the T tubular system. The absolute quantity of Ca^{2+} traversing sarcolemma and T tubules is modest and insufficient to fully activate contraction. However, this Ca^{2+} current, through *Ca^{2+}-induced Ca^{2+} release,* triggers substantial Ca^{2+} release from the SR, inducing contraction.

Ca^{2+} is released from the SR through a Ca^{2+} release channel, a cardiac isoform of the ryanodine receptor (RyR2). Several regulatory proteins, including *calstabin 2*, inhibit RyR2 and thus SR Ca^{2+} release. Inherited disorders or exogenous factors affecting the efficiency or stability of SR Ca^{2+} handling can impair contraction, leading to heart failure, or to ventricular arrhythmias.

The Ca^{2+} released from the SR diffuses to interact with myofibrillar troponin C (Fig. 232-7), repressing this protein's inhibition of contraction, and so activating myofilaments to shorten. During repolarization, the activity of the SR Ca^{2+} ATPase ($SERCA_{2A}$) leads to Ca^{2+} uptake against a concentration gradient into the SR where complexes with another specialized protein, *calsequestrin*. The uptake of Ca^{2+} is ATP (energy)-dependent and lowers cytoplasmic [Ca^{2+}] to a level where actomyosin interaction is inhibited and myocardial relaxation occurs. There is also a sarcolemmal exchange of Ca^{2+} for Na^+ (Fig. 232-8), reducing the cytoplasmic [Ca^{2+}]. Additional control of calcium compartmentalizaion results from cyclic AMP–dependent PKA phosphorylation of the SR protein

Loading [Contrib]/a11y/accessibility-menu.js

phospholamban, permitting $SERCA_{2A\ activation}$, increasing SR Ca^{2+} uptake, and so accelerating the relaxation rates, loading the SR with Ca^{2+} for subsequent release, and stimulating contraction.

Thus, the combination of the cell membrane, transverse tubules, and SR, with their ability to transmit the action potential and release and then reaccumulate Ca^{2+}, controls the cyclic contraction and relaxation of heart muscle. Genetic or pharmacologic alterations of any component, whatever its etiology, can disturb any of the functions of this finely tuned system.

CONTROL OF CARDIAC PERFORMANCE AND OUTPUT

The extent of shortening of heart muscle and, therefore, ventricular stroke volume in the intact heart, depends on three major influences: (1) the length of the muscle at the onset of contraction, i.e., the preload; (2) the tension that the muscle must develop during contraction, i.e., the afterload; and (3) muscle contractility, i.e., the extent and velocity of shortening at any given preload and afterload. **Table 232-2** lists the major determinants of preload, afterload, and contractility.

Loading [Contrib]/a11y/accessibility-menu.js

TABLE 232-2 Determinants of Stroke Volume

I. Ventricular Preload

A. Blood volume

B. Distribution of blood volume

1. Body position
2. Intrathoracic pressure
3. Intrapericardial pressure
4. Venous tone
5. Pumping action of skeletal muscles

C. Atrial contraction

II. Ventricular Afterload

A. Systemic vascular resistance

B. Elasticity of arterial tree

C. Arterial blood volume

D. Ventricular wall tension

1. Ventricular radius
2. Ventricular wall thickness

III. Myocardial Contractility[a]

A. Intramyocardial $[Ca^{2+}]$ ↑↓

B. Cardiac adrenergic nerve activity ↑↓[b]

C. Circulating catecholamines ↑↓[b]

D. Cardiac rate ↑↓[b]

E. Exogenous inotropic agents ↑

F. Myocardial ischemia ↓

G. Myocardial cell death (necrosis, apoptosis, autophagy) ↓

H. Alterations of sarcomeric and cytoskeletal proteins ↓

1. Genetic
2. Hemodynamic overload

I. Myocardial fibrosis ↓

J. Chronic overexpression of neurohormones ↓

K. Ventricular remodeling ↓

L. Chronic and/or excessive myocardial hypertrophy ↓

[a]Arrows indicate directional effects of determinants of contractility. [b]Contractility rises initially but later becomes depressed.

THE ROLE OF MUSCLE LENGTH (PRELOAD)

Preload determines sarcomere length at the onset of contraction. Contractile force is optimal at specific sarcomere lengths (~2.2 μm) where Ca^{2+} binding is maximal, and where myofilament interactions and activation of contraction are most efficient. The relationship between initial muscle fiber length and the developed force is the basis of Starling's law of the heart, which states

Loading [Contrib]/a11y/accessibility-menu.js

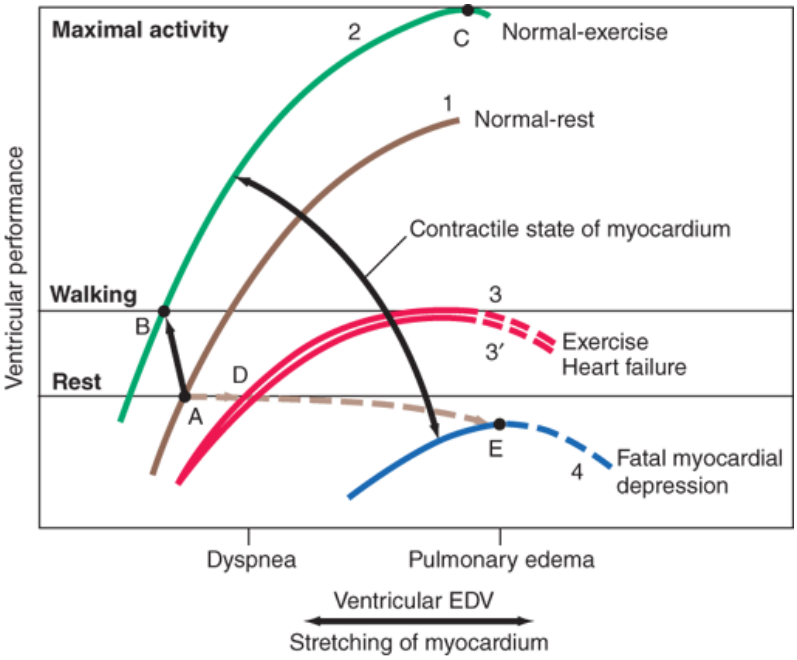
that within limits, the ventricular contraction force depends on the end-diastolic length of the cardiac muscle; which in vivo relates closely to the ventricular end-diastolic volume.

CARDIAC PERFORMANCE

Ventricular end-diastolic or “filling” pressure can serve as a surrogate for end-diastolic volume. In isolated heart and heart-lung preparations, stroke volume varies directly with the end-diastolic fiber length (preload) and inversely with the arterial resistance (afterload), and as the heart fails—i.e., as its contractility declines—it delivers a progressively smaller stroke volume from a normal or even elevated end-diastolic volume. The relation between ventricular end-diastolic pressure and the stroke work of the ventricle (the ventricular function curve) provides a working definition of cardiac contractility in the intact organism. An increase in contractility is accompanied by a shift of the ventricular function curve upward and to the left (greater stroke work at any level of ventricular end-diastolic pressure, or lower end-diastolic volume at any level of stroke work), whereas a shift downward and to the right characterizes depression of contractility (**Fig. 232-9**).

FIGURE 232-9

The interrelations among influences on ventricular end-diastolic volume (EDV) through stretching of the myocardium and the contractile state of the myocardium. Levels of ventricular EDV associated with filling pressures that result in dyspnea and pulmonary edema are shown on the abscissa. Levels of ventricular performance required when the subject is at rest, while walking, and during maximal activity are designated on the ordinate. The *broken lines* are the descending limbs of the ventricular-performance curves, which are rarely seen during life but show the level of ventricular performance if end-diastolic volume could be elevated to very high levels. For further explanation, see text. *(Modified from WS Colucci and E Braunwald: Pathophysiology of heart failure, in Braunwald’s Heart Disease, 7th ed, DP Zipes et al [eds]. Philadelphia: Elsevier, 2005, pp 509–538.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

VENTRICULAR AFTERLOAD

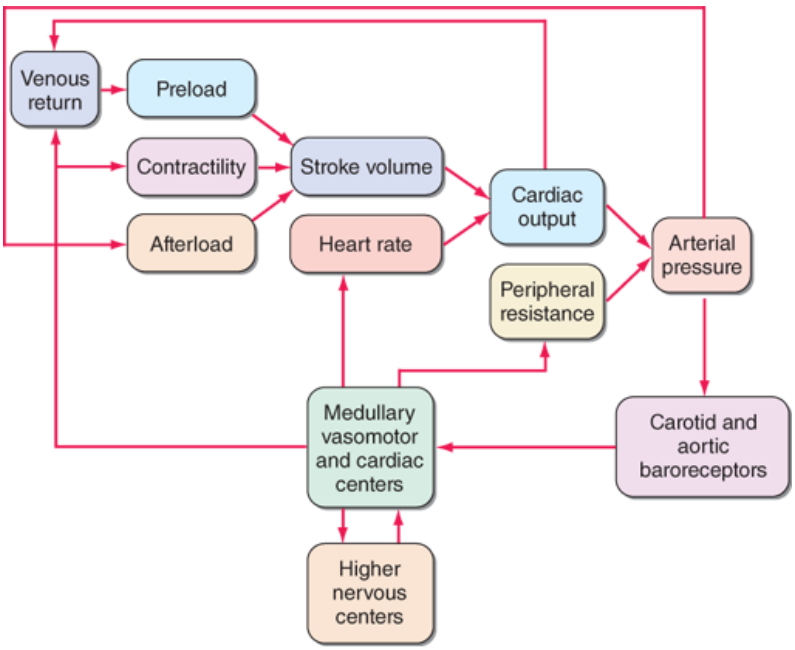
In the intact heart, as ex vivo, the extent and velocity of shortening of ventricular muscle fibers at any level of preload and of myocardial contractility relate inversely to the afterload, i.e., the instantaneous load opposing shortening. In the intact heart, the afterload may be defined as the tension developed in the ventricular wall during ejection. Afterload is determined by the aortic pressure as well as by the volume and thickness of the ventricular cavity. Laplace’s law specifies that the tension of the myocardial fiber is the product of the intra-cavitary ventricular pressure and ventricular radius divided by wall thickness. Therefore, at any given aortic pressure, the afterload on a dilated left ventricle exceeds that on a normal-sized ventricle. Conversely, at the same aortic pressure and ventricular diastolic volume, the afterload on a hypertrophied ventricle is lower than that on a normal chamber. Aortic pressure in turn depends on the peripheral vascular resistance, the biomechanics of the arterial tree, and the volume of blood it contains at the onset of ejection.

Ventricular afterload finely regulates cardiovascular performance (**Fig. 232-10**). As noted, elevations in both preload and contractility increase myocardial fiber shortening, whereas increases in afterload reduce it. The extent of myocardial fiber shortening and left ventricular size determine stroke volume. An increase in arterial pressure induced by vasoconstriction, for example, augments afterload, which opposes myocardial fiber shortening, reducing stroke volume.

Loading [Contrib]/a11y/accessibility-menu.js

FIGURE 232-10

Interactions in the intact circulation of preload, contractility, and afterload in producing stroke volume. Stroke volume combined with heart rate determines cardiac output, which, when combined with peripheral vascular resistance, determines arterial pressure for tissue perfusion. The characteristics of the arterial system also contribute to afterload, an increase that reduces stroke volume. The interaction of these components with carotid and aortic arch baroreceptors provides a feedback mechanism to higher medullary and vasomotor cardiac centers and to higher levels in the central nervous system to effect a modulating influence on heart rate, peripheral vascular resistance, venous return, and contractility. *(From MR Starling: Physiology of myocardial contraction, in Atlas of Heart Failure: Cardiac Function and Dysfunction, 3rd ed, WS Colucci and E Braunwald [eds]. Philadelphia: Current Medicine, 2002, pp 19–35.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

When myocardial contractility is impaired and the ventricle dilates, afterload rises (Laplace's law) and limits cardiac output. Increased afterload also may result from neural and humoral stimuli that occur in response to a fall in cardiac output. This increased afterload may reduce cardiac output further, thereby increasing ventricular volume and initiating a vicious circle, especially in patients with ischemic heart disease and limited myocardial O_2 supply. Treatment with vasodilators has the opposite effect; when afterload falls, cardiac output rises **(Chap. 252)**.

Under normal circumstances, the various influences acting on cardiac performance interact in a complex fashion to maintain cardiac output at a level responsive to the requirements of tissue metabolic demands (Fig. 232-10). Interference with a single mechanism may not influence the cardiac output due to homeostatic adjustments. For example, a moderate reduction of blood volume or the loss of the atrial contribution to ventricular contraction can be tolerated without a reduction in resting cardiac output. Under these circumstances, other factors, such as adrenergic neuronal impulses to the heart, heart rate, and venous tone, will serve as compensatory mechanisms and sustain cardiac output in a normal individual. Ultimately, understanding the complex interactions between these different variables requires rigorous models to predict relevant outcomes, and led to the early application of systems engineering principles in medicine.

EXERCISE

The integrated response to exercise illustrates the interactions among the three determinants of stroke volume: preload, afterload, and contractility (Fig. 232-9). Hyperventilation, the pumping action of the exercising muscles, and venoconstriction during exercise all augment venous return and hence ventricular filling and preload (Table 232-2). Simultaneously, the increase in the adrenergic neuronal stimulation of the myocardium, the increased concentration of circulating catecholamines, and the tachycardia that occur during exercise combine to augment the myocardial contractility (Fig. 232-9, curves 1 and 2), together elevating stroke volume and stroke work, without a change in or even a reduction of end-diastolic pressure and volume (Fig. 232-9, points A and B). Vasodilation occurs in the exercising muscles, thus limiting the increase in afterload that otherwise would occur as cardiac output rises to levels as high as five times greater than basal levels during maximal exercise. This vasodilation ultimately allows the achievement of elevated cardiac outputs during exercise at arterial pressures only moderately higher than the resting state.

ASSESSMENT OF CARDIAC FUNCTION

SLoading [Contrib]/a11y/accessibility-menu.jsd cardiac function in clinical practice. Cardiac output and stroke volume may decline in the presence of heart failure, but these variables are often within normal limits, especially at rest. A more sensitive index of cardiac

function is the ejection fraction, i.e., the ratio of stroke volume to end-diastolic volume (normal value = 67 ± 8%), which is frequently depressed in systolic heart failure even when stroke volume is normal. Alternatively, abnormally elevated ventricular end-diastolic volume (normal value = 75 ± 20 mL/m^2) or end-systolic volume (normal value = 25 ± 7 mL/m^2) signifies left ventricular systolic impairment.

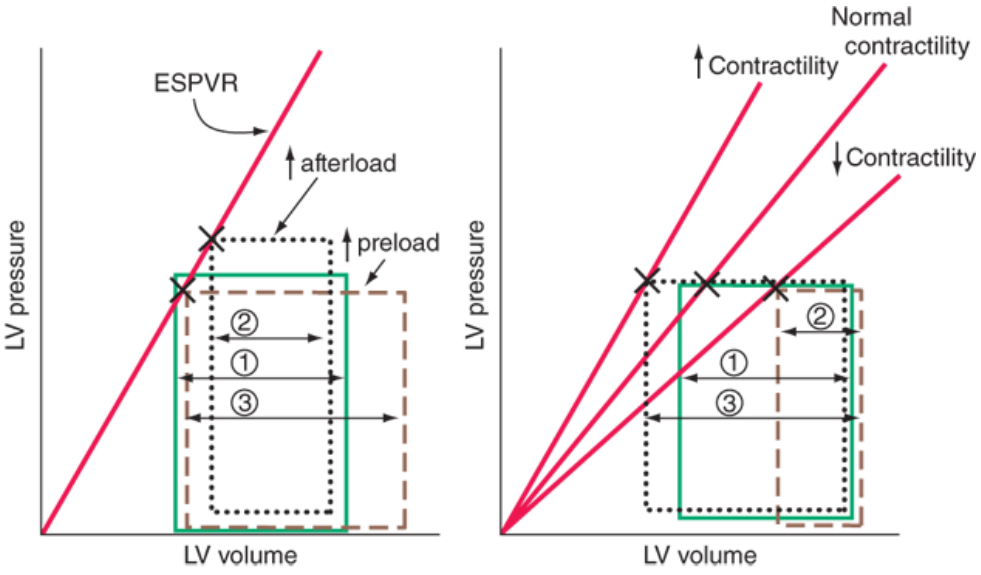
Noninvasive techniques, particularly echocardiography, radionuclide scintigraphy and cardiac magnetic resonance imaging (MRI) **(Chap. 236)** have great value in the clinical assessment of myocardial function. They provide measurements of end-diastolic and end-systolic volumes, ejection fraction, and systolic shortening rate, and they allow assessment of ventricular filling (see below) as well as regional contraction and relaxation. The latter measurements have particular importance in ischemic heart disease, as myocardial infarction causes regional myocardial damage.

Strong dependence on ventricular loading conditions influence the of measurements of cardiac output, ejection fraction, and ventricular volumes as indices of cardiac function. Thus, a depressed ejection fraction and lowered cardiac output may occur in patients with normal ventricular function but reduced preload, as occurs in hypovolemia, or with increased afterload, as occurs in acutely elevated arterial pressure.

The end-systolic left ventricular pressure-volume relationship has particular value as an index of ventricular performance as it does not depend on preload and afterload **(Fig. 232-11)**. At any level of myocardial contractility, left ventricular end-systolic volume varies inversely with end-systolic pressure; as contractility declines, end-systolic volume (at any level of end-systolic pressure) rises. Measurement of end-systolic left ventricular pressure-volume loops adds rigor to research studies of left ventricular function, but has not replaced the more readily assessed indices, such as ventricular volumes and ejection fraction, in clinical practice.

FIGURE 232-11

The responses of the left ventricle to increased afterload, increased preload, and increased and reduced contractility are shown in the pressure-volume plane. ***Left.*** Effects of increases in preload and afterload on the pressure-volume loop. Because there has been no change in contractility, the end-systolic pressure-volume relationship (ESPVR) is unchanged. With an increase in afterload, stroke volume falls (1 → 2); with an increase in preload, stroke volume rises (1 → 3). ***Right.*** With increased myocardial contractility and constant left ventricular end-diastolic volume, the ESPVR moves to the left of the normal line (lower end-systolic volume at any end-systolic pressure) and stroke volume rises (1 → 3). With reduced myocardial contractility, the ESPVR moves to the right; end-systolic volume is increased, and stroke volume falls (1 → 2).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIASTOLIC FUNCTION

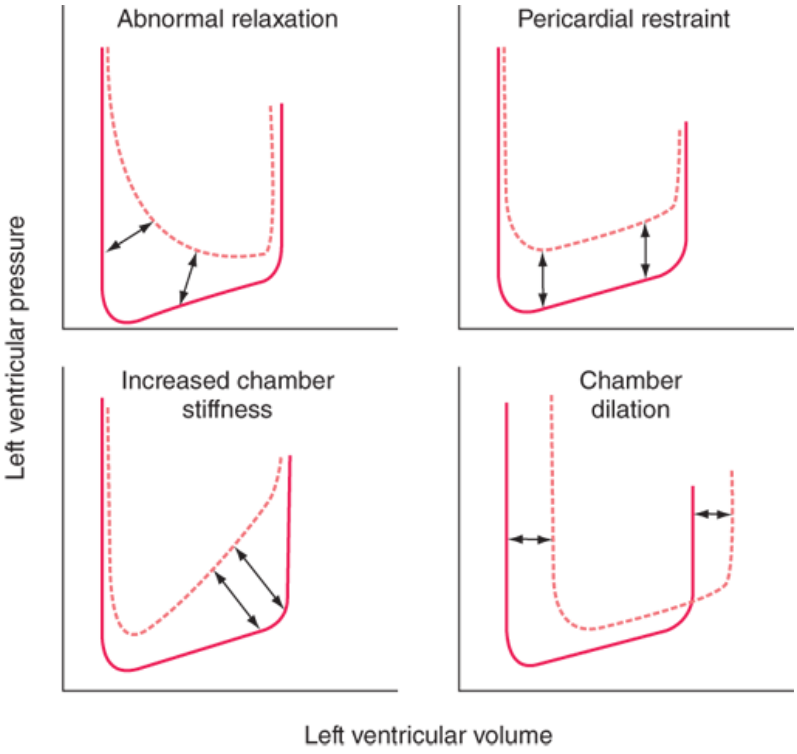
Ventricular filling is influenced by the extent and speed of myocardial relaxation, a function of the rate of uptake of Ca^{2+} by the SR; the latter may be enhanced by adrenergic activation and reduced by ischemia, which reduces the ATP available for pumping Ca^{2+} into the SR (see above). The passive stiffness of the ventricular wall also may impede filling. Ventricular stiffness increases with hypertrophy and conditions that infiltrate the ventricle, such as amyloid, or can result from an extrinsic constraint (e.g., pericardial compression) **(Fig. 232-12)**.

FIGURE 232-12

M… …sfunction reflected in the pressure-volume relation. The bottom half of the pressure-volume loop is depicted. *Solid lines* represent normal subjects; *broken lines* represent patients with diastolic dysfunction. *(From JD*

Loading [Contrib]/a11y/accessibility-menu.js

Carroll et al: The differential effects of positive inotropic and vasodilator therapy on diastolic properties in patients with congestive cardiomyopathy. Circulation 74:815, 1986; with permission.)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ventricular filling can be assessed by measuring flow velocity across the mitral valve using Doppler ultrasound. Normally, inflow velocity is more rapid in early diastole than during atrial systole; with mild to moderately impaired relaxation, the rate of early diastolic filling declines, as presystolic filling rates rise. With further stiffening, flow is “pseudo-normalized,” as early ventricular filling becomes more rapid with rising left atrial pressure upstream of the left ventricle.

CARDIAC METABOLISM

The heart requires a continuous supply of energy (ATP) not only to drive mechanical contraction, but also to maintain ionic and biochemical homeostasis. The development of tension, the frequency of contraction, and myocardial contractility levels are the principal determinants of the heart’s energy needs, rendering its O_2 requirements ~15% of that of the entire organism.

Most ATP production depends on oxidation of the substrates glucose and free fatty acids (FFAs). FFAs used by the myocardium derive from circulating FFAs, principally from lipolysis in adipose tissue, whereas the myocyte’s glucose derives from plasma as well as from the cell’s breakdown of its glycogen stores (glycogenolysis). These two principal sources of acetyl coenzyme A in cardiac muscle vary reciprocally. Glucose is broken down in the cytoplasm into a three-carbon product, pyruvate, which passes into mitochondria, where it is metabolized to the two-carbon fragment, acetyl-CoA, and undergoes oxidation. FFAs are converted to acyl-CoA in the cytoplasm and acetyl-CoA in the mitochondria. Acetyl-CoA enters the citric acid (Krebs) cycle to produce ATP by oxidative phosphorylation; ATP then enters the cytoplasm from the mitochondrial compartment. Intracellular adenosine diphosphate (ADP), resulting from ATP breakdown, enhances ATP production.

In the fasted, resting state, circulating FFA concentrations and their myocardial uptake are high, and they furnish most of the heart’s acetyl-CoA (~70%). In the fed state, with elevations of blood glucose and insulin, glucose oxidation increases and FFA oxidation subsides. Increased cardiac work, inotropic agents, hypoxia, and mild ischemia all enhance myocardial glucose uptake, glucose production resulting from glycogenolysis, and glucose metabolism to pyruvate (glycolysis). By contrast, β-adrenergic stimulation, possibly due to stress, raises the circulating levels and metabolism of FFAs in favor of glucose. Severe ischemia inhibits the cytoplasmic pyruvate dehydrogenase, and despite both glycogen and glucose breakdown, glucose undergoes incomplete metabolism to lactic acid (anaerobic glycolysis), which does not enter the citric acid cycle. Anaerobic glycolysis produces much less ATP than does aerobic glucose metabolism. High concentrations of circulating FFAs, which can occur when adrenergic stimulation is superimposed on severe ischemia, reduce oxidative phosphorylation and cause ATP wastage; the myocardial content of ATP declines impairing contraction. In addition, FFA breakdown products may exert toxic or arrhythmogenic effects on cardiac cell membranes.

Myocardial energy is stored as creatine phosphate (CP), which is in equilibrium with ATP, the immediate energy source. In states of reduced energy availability, the CP stores decline first. Cardiac hypertrophy, fibrosis, tachycardia, increased wall tension due to

Loading [Contrib]/a11y/accessibility-menu.js

cytoplasmic [Ca^{2+}] all contribute to increased myocardial energy needs. When coupled with

reduced coronary flow reserve, as occurs with obstruction of coronary arteries or abnormalities of the coronary microcirculation, an imbalance in myocardial ATP production relative to demand may occur, and the resulting ischemia can worsen or cause heart failure.

REGENERATING CARDIAC TISSUE

Until very recently, adult mammalian myocardial cells were viewed as fully differentiated and without regenerative potential. Evidence currently supports the existence of limited regenerative potential of the mature heart. Considerable current effort is being devoted to evaluating the utility of various putative stem cell populations and regenerative approaches to enhance cardiac repair after injury. The success of such approaches would offer the exciting possibility of reconstructing an infarcted or failing ventricle **(Chap. 473)**.

ACKNOWLEDGMENT

The authors wish to thank Jonathan Epstein for his contribution to the prior version of this chapter.

FURTHER READING

BAUTCH VL, CARON KM: Blood and lymphatic vessel formation. Cold Spring Harb Perspect Biol 7(3):a008268, 2015. [PubMed: 25731762]
DEJANA E et al: The molecular basis of endothelial cell plasticity. Nat Commun 8:14361, 2017. [PubMed: 28181491]
GREEN DJ et al: Vascular adaptation to exercise in humans: Role of hemodynamic stimuli. Physiol Rev 97:495, 2017. [PubMed: 28151424]
MACLEOD KT: Recent advances in understanding cardiac contractility in health and disease. F1000Res 5(F1000 Faculty Rev):1770, 2016.
MANN D, et al (eds): *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*, 10th ed. Philadelphia, Elsevier, 2015.
PAGE E et al (eds): *Handbook of Physiology: A Critical Comprehensive Presentation of Physiological Knowledge and Concepts. Section 2: The Cardiovascular System, Volume I: The Heart*. New York, Oxford University Press, 2002.
SPINALE FG: Assessment of cardiac function—Basic principles and approaches. Compr Physiol 5:1911, 2015. [PubMed: 26426471]
SRIVASTAVA D: Making or breaking the heart: From lineage determination to morphogenesis. Cell 126:1037, 2006. [PubMed: 16990131]

Loading [Contrib]/a11y/accessibility-menu.js

Chapter 233: Epidemiology of Cardiovascular Disease

Thomas A. Gaziano; J. Michael Gaziano

INTRODUCTION

Cardiovascular disease (CVD) is now the most common cause of death worldwide. Before 1900, infectious diseases and malnutrition were the most common causes, and CVD was responsible for <10% of all deaths. In 2015, CVD accounted for ~17.9 million deaths worldwide (32%), including nearly 34% of deaths in high-income countries and about 32% in low- and middle-income countries.

THE EPIDEMIOLOGIC TRANSITION

Image not available. The global rise in CVD is the result of an unprecedented transformation in the causes of morbidity and mortality during the twentieth century. Known as the epidemiologic transition, this shift is driven by industrialization, urbanization, and associated lifestyle changes and is taking place in every part of the world among all races, ethnic groups, and cultures. The transition is divided into four basic stages: pestilence and famine, receding pandemics, degenerative and man-made diseases, and delayed degenerative diseases. A fifth stage, characterized by an epidemic of inactivity and obesity, is emerging in some countries (**Table 233-1**).

TABLE 233-1 Five Stages of the Epidemiologic Transition

Stage	Description	Deaths Related to CVD, %	Predominant CVD Type
Pestilence and famine	Predominance of malnutrition and infectious diseases as causes of death; high rates of infant and child mortality; low mean life expectancy	<10	Rheumatic heart disease, cardiomyopathies caused by infection and malnutrition
Receding pandemics	Improvements in nutrition and public health lead to decrease in rates of deaths related to malnutrition and infection; precipitous decline in infant and child mortality rates	10–35	Rheumatic valvular disease, hypertension, CHD, and stroke (predominantly hemorrhagic)
Degenerative and man-made diseases	Increased fat and caloric intake and decrease in physical activity lead to emergence of hypertension and atherosclerosis; with increase in life expectancy, mortality from chronic, noncommunicable diseases exceeds mortality from malnutrition and infectious disease	35–65	CHD and stroke (ischemic and hemorrhagic)
Delayed degenerative diseases	CVD and cancer are the major causes of morbidity and mortality; better treatment and prevention efforts help avoid deaths among those with disease and delay primary events; age-adjusted CVD morality declines; CVD affecting older and older individuals	40–50	CHD, stroke, and congestive heart failure
Inactivity and obesity	Overweight and obesity increase at alarming rate; diabetes and hypertension increase; decline in smoking rates levels off; a minority of the population meets physical activity recommendations	38	CHD, stroke, and congestive heart failure, peripheral vascular disease

Abbreviations: CHD, coronary heart disease; CVD, cardiovascular disease.

Source: Data from AR Omran: The epidemiologic transition: A theory of the epidemiology of population change. Milbank Mem Fund Q 49:509, 1971; and SJ Olshansky, AB Ault: The fourth stage of the epidemiologic transition: The age of delayed degenerative diseases. Milbank Q 64:355, 1986.

The *age of pestilence and famine* is marked by malnutrition, infectious diseases, and high infant and child mortality that are offset by high fertility. Tuberculosis, dysentery, cholera, and influenza are often fatal, resulting in a mean life expectancy of about 30 years. CVD, which accounts for <10% of deaths, takes the form of rheumatic heart disease and cardiomyopathies due to infection and malnutrition. Approximately 10% of the world's population remains in the age of pestilence and famine.

Per capita income and life expectancy increase during the *age of receding pandemics* as the emergence of public health systems, cleaner water supplies, and improved nutrition combine to drive down deaths from infectious disease and malnutrition. Infant and childhood mortality also decline, but deaths due to CVD increase to between 10 and 35% of all deaths. Rheumatic valvular disease, hypertension, coronary heart disease (CHD), and stroke are the predominant forms of CVD. Almost 40% of the world's population is currently in this stage.

Loading [Contrib]/a11y/accessibility-menu.js

The *age of degenerative and man-made diseases* is distinguished by mortality from noncommunicable diseases—primarily CVD—surpassing mortality from malnutrition and infectious diseases. Caloric intake, particularly from animal fat, increases. CHD and stroke are prevalent, and between 35 and 65% of all deaths can be traced to CVD. Typically, the rate of CHD deaths exceeds that of stroke by a ratio of 2:1 to 3:1. During this period, average life expectancy surpasses the age of 50. Roughly 35% of the world's population falls into this category.

In the *age of delayed degenerative diseases*, CVD and cancer remain the major causes of morbidity and mortality, with CVD accounting for 40% of all deaths. However, age-adjusted CVD mortality declines, aided by preventive strategies (for example, smoking cessation programs and effective blood pressure control), acute hospital management, and technologic advances, such as the availability of bypass surgery. CHD, stroke, and congestive heart failure are the primary forms of CVD. About 15% of the world's population is now in the age of delayed degenerative diseases or is exiting this age and moving into the fifth stage of the epidemiologic transition.

In the industrialized world, physical activity continues to decline while total caloric intake increases. The resulting epidemic of overweight and obesity may signal the start of the *age of inactivity and obesity*. Rates of type 2 diabetes mellitus, hypertension, and lipid abnormalities are on the rise, trends that are particularly evident in children. If these risk factor trends continue, age-adjusted CVD mortality rates could increase in the coming years.

PATTERNS IN THE EPIDEMIOLOGIC TRANSITION

Unique regional features have modified aspects of the transition in various parts of the world. High-income countries experienced declines in CVD death rates by as much as 50–60% over the last 60 years, whereas CVD death rates increased by 15% over the past 20 years in the low- and middle-income range. However, given the large amount of available data, the United States serves as a useful reference point for comparisons. The age of pestilence and famine occurred before 1900, with a largely agrarian economy and population. Infectious diseases accounted for more deaths than any other cause. By the 1930s, the country proceeded through the age of receding pandemics. The establishment of public health infrastructures resulted in dramatic declines in infectious disease mortality rates. Lifestyle changes due to rapid urbanization resulted in a simultaneous increase in CVD mortality rates, reaching ~390 per 100,000. Between 1930 and 1965, the country entered the age of degenerative and man-made diseases. Infectious disease mortality rates fell to fewer than 50 per 100,000 per year, whereas CVD mortality rates reached peak levels with increasing urbanization and lifestyle changes in diet, physical activity, and tobacco consumption. The age of delayed degenerative diseases took place between 1965 and 2000. New therapeutic approaches, preventive measures, and exposure to public health campaigns promoting lifestyle modifications led to substantial declines in age-adjusted mortality rates and a steadily rising age at which a first CVD event occurs.

Currently, the United States is entering what appears to be a fifth phase. The decline in the age-adjusted CVD death rate of 3% per year through the 1970s and 1980s has tapered off in the 1990s to 2%. However, CVD death rates have declined by 3–5% per year during the first decade of the new millennium. Competing trends appear to be at play. On the one hand, an increase in the prevalence of diabetes and obesity, a slowing in the rate of decline in smoking, and a leveling off in the rate of detection and treatment for hypertension are in the negative column. On the other hand, cholesterol levels continue to decline in the face of increased statin use.

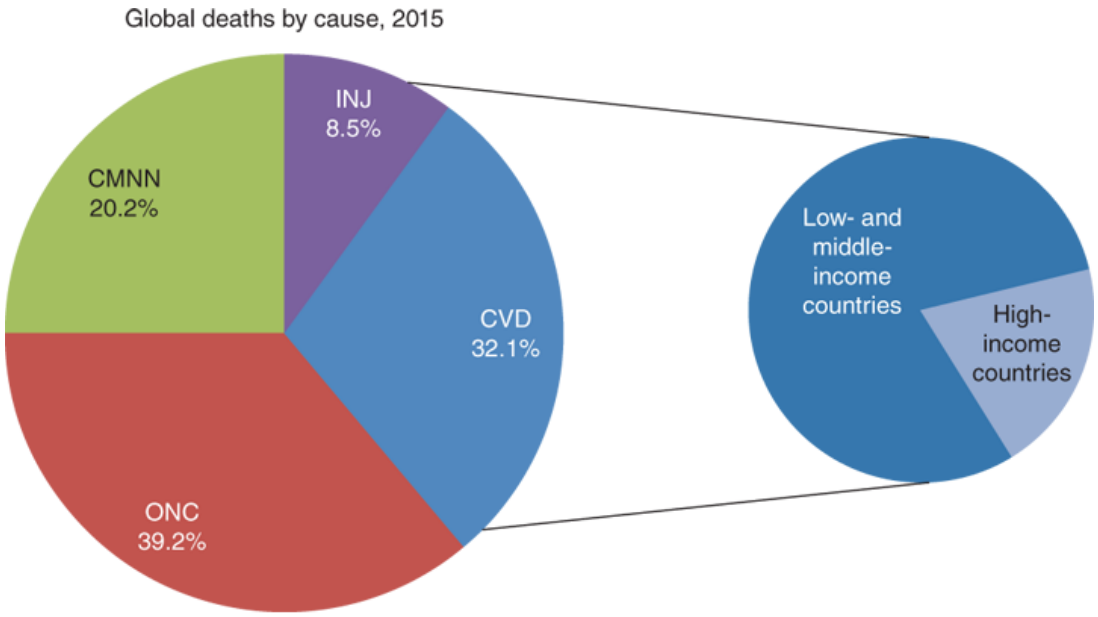
Many high-income countries (HICs)—which together account for 15% of the population—have proceeded through four stages of the epidemiologic transition in roughly the same pattern as the United States. CHD is the dominant form of CVD in these countries, with rates that tend to be two- to fivefold higher than stroke rates. However, variations exist. Whereas North America, Australia, and central northwestern European HICs experienced significant increases than rapid declines in CVD rates, southern and central European countries experienced a more gradual rise and fall in rates. More specifically, central European countries (i.e., Austria, Belgium, and Germany) declined at slower rates compared to their northern counterparts (i.e., Finland, Sweden, Denmark, and Norway). Countries such as Portugal, Spain, and Japan never reached the high mortality rates that the United States and other countries did, with CHD mortality rates at 200 per 100,000, or less. The countries of Western Europe also exhibit a clear north/south gradient in absolute rates of CVD, with rates highest in northern countries (i.e., Finland, Ireland, and Scotland) and lowest in Mediterranean countries (i.e., France, Spain, and Italy). Japan is unique among the HICs, most likely due to the unique dietary patterns of its population. Although stroke rates increased dramatically, CHD rates did not rise as sharply in Japan. However, Japanese dietary habits are undergoing substantial changes, reflected in an increase in cholesterol levels.

Patterns in low- and middle-income countries (LMICs; gross national income per capita <$12,736) depend, in part, on cultural differences, secular trends, and responses at the country level, with regard to both public health and treatment infrastructure. Although communicable diseases continue to be a major cause of death, CVD has emerged as a significant health concern in LMICs. With 85% of the world's population, LMICs are driving the rates of change in the global burden of CVD (**Fig. 233-1**). In most LMICs, an urban/rural gradient has emerged for CHD, stroke, and hypertension, with higher rates in urban centers.

FIGURE 233-1

Global deaths by cause, 2015. CMNN, communicable, maternal, neonatal, and nutritional disorders; CVD, cardiovascular diseases; INJ, injuries; ONC, other noncommunicable diseases. *(Based on data from Global Burden of Disease Study 2015. Global Burden of Disease Study 2015 [GBD 2015] Results. Seattle, United States: Institute for Health Metrics and Evaluation [IHME], 2016.)*

Loading [Contrib]/a11y/accessibility-menu.js



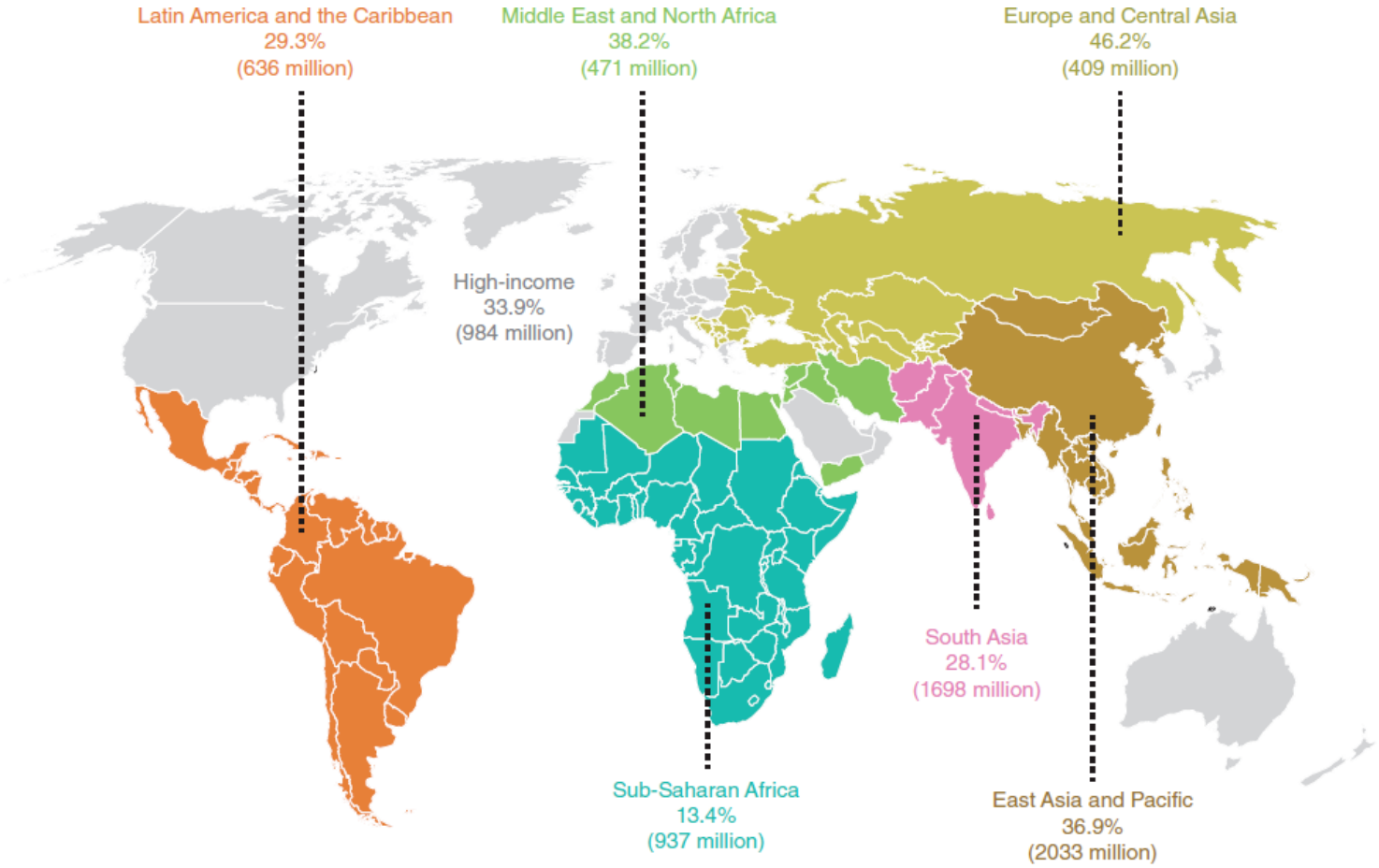
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

However, although CVD rates are rapidly rising, vast differences exist among the regions and countries, and even within the countries themselves **(Fig. 233-2)**. The East Asia and Pacific regions appear to be straddling the second and third phases of the epidemiologic transition. CVD is a major cause of death in China, but like Japan, stroke causes more deaths than CHD in a ratio of about three to one. Vietnam and Cambodia, on the other hand, are just emerging from the pestilence and famine transition. The Middle East and North Africa regions also appear to be entering the third phase of the epidemiologic transition, with increasing life expectancy and CVD death rates just below those of HICs. In general, Latin America appears to be in the third phase of the transition, although there is vast regional heterogeneity with some areas in the second phase of the transition and some in the fourth. The Eastern Europe and Central Asia regions, however, are firmly in the peak of the third phase, with the highest death rates due to CVD (~66%) in the world. Importantly, deaths due to CHD are not limited to the elderly in this region and have a significant effect on working-age populations. South Asia—and more specifically, India, which accounts for the greatest proportion of the region's population—is experiencing an alarming increase in heart disease. The transition appears to be in the Western style, with CHD as the dominant form of CVD. However, rheumatic heart disease continues to be a major cause of morbidity and mortality. As in South Asia, rheumatic heart disease is also an important cause of CVD morbidity and mortality in sub-Saharan Africa, which largely remains in the first phase of the epidemiologic transition.

FIGURE 233-2

Cardiovascular disease deaths as a percentage of total deaths and total population in seven economic regions of the world defined by the World Bank. *(Based on data from Global Burden of Disease Study 2015. Global Burden of Disease Study 2015 [GBD 2015] Results. Seattle, United States: Institute for Health Metrics and Evaluation [IHME], 2016.)*

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Many factors contribute to this heterogeneity among LMICs. First, the regions are in various stages of the epidemiologic transition. Second, vast differences in lifestyle and behavioral risk factors exist. Third, racial and ethnic differences may lead to altered susceptibilities to various forms of CVD. In addition, it should be noted that for most countries in these regions, accurate country-wide data on cause-specific mortality are not complete.

GLOBAL TRENDS IN CARDIOVASCULAR DISEASE

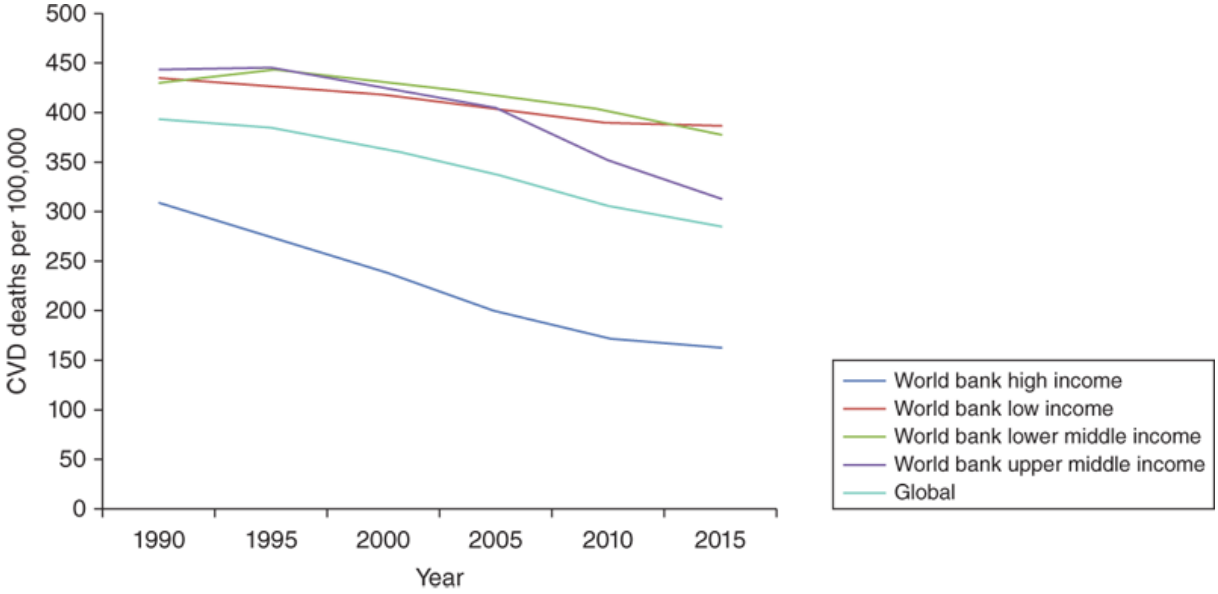
Image not available. CVD accounts for 32% of deaths worldwide, a number that is expected to increase. In 2015, CHD accounted for 16.7% of all deaths globally and the largest portion (10%) of global years of life lost (YLLs) and disability-adjusted life-years (DALYs) (7%). The third largest cause of death was stroke (11.9% of all deaths), which was also the third largest contributor to global YLLs and DALYs. Together, CHD and stroke accounted for nearly a quarter of all deaths worldwide. The burden of stroke is of growing concern among LMICs. The impact of stroke on DALYs and mortality rates is more than three times greater in LMICs as compared to HICs.

With nearly 81% of the world's population, LMICs largely drive global CVD rates and trends. More than 14 million (14.2) CVD deaths occurred in LMICs in 2015, compared to 3.7 million in HICs. Globally, there is evidence of significant delays in age of occurrence and/or improvements in case fatality rates; between 1990 and 2015, the number of CVD deaths increased by 42%, but age-adjusted death rates decreased by 27.3% in the same period. Age-standardized death rates, however, have declined faster in high-income countries than in middle-income and lower-income regions **(Fig. 233-3)**. Population growth has been greater in low- and middle-income countries compared to high-income countries. As a result of slower rates of population growth in high-income countries, overall CVD deaths remained steady. However, in the lower and middle-income countries, the population aging and growth outstripped gains in age-adjusted mortality reductions such that overall CVD deaths continued to climb over the last 25 years **(Fig. 233-4).**

FIGURE 233-3

Age-standardized cardiovascular diseases (CVD) death rate per 100,000 from 1990 to 2015, by World Bank income. *(Based on data from Global Burden of Disease Study 2015. Global Burden of Disease Study 2015 [GBD 2015] Results. Seattle, United States: Institute for Health Metrics and Evaluation [IHME], 2016.)*

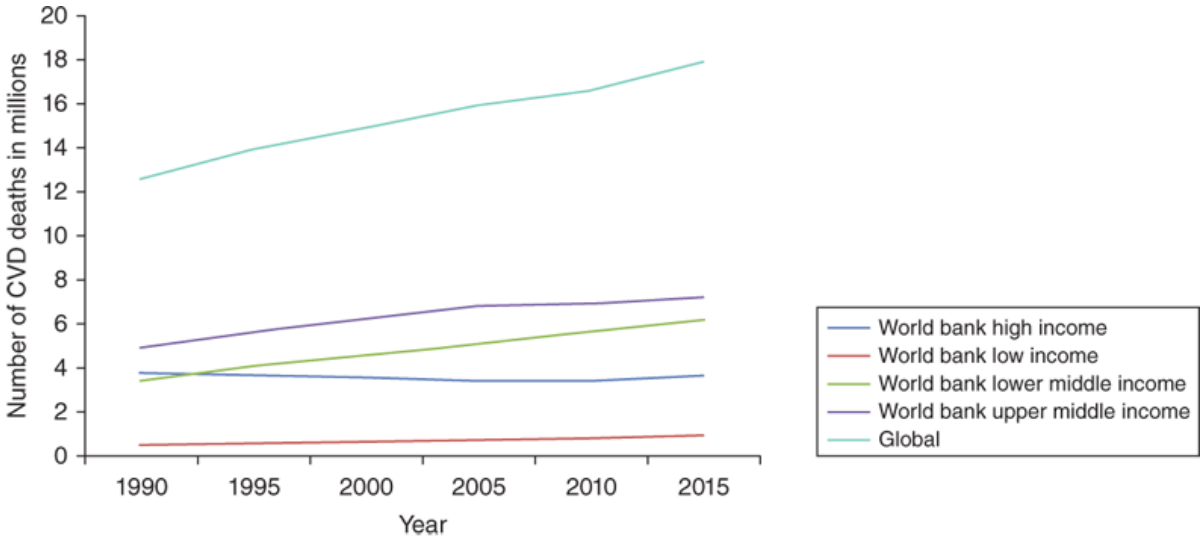
Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 233-4

Number of cardiovascular diseases (CVD) deaths from 1990 to 2015, by World Bank income. *(Based on data from Global Burden of Disease Study 2015. Global Burden of Disease Study 2015 [GBD 2015] Results. Seattle, United States: Institute for Health Metrics and Evaluation [IHME], 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although HIC population growth will be fueled by emigration from LMICs, the populations of HICs will shrink as a proportion of the world's population. The modest decline in CVD death rates that began in the HICs in the latter third of the twentieth century will continue, but the rate of decline appears to be slowing. However, these countries are expected to see an increase in the prevalence of CVD, as well as the absolute number of deaths as the population ages.

Significant portions of the population living in LMICs have entered the third phase of the epidemiologic transition, and some are entering the fourth stage. Changing demographics play a significant role in future predictions for CVD throughout the world. For example, the population growth rate in Eastern Europe and Central Asia was 0.7% in 2014, whereas it was 1.4% in South Asia.

CVD rates will also have an economic impact. Even assuming no increase in CVD risk factors, most countries, but especially India and South Africa, will see a large number of people between 35 and 64 die of CVD over the next 30 years, as well as an increasing level of morbidity among middle-aged people related to heart disease and stroke. In China, it is estimated that there will be 9 million deaths from CVD in 2030—up from 2.4 million in 2002—with half occurring in individuals between 35 and 64 years old.

Loading [Contrib]/a11y/accessibility-menu.js

The global variation in CVD rates is related to temporal and regional variations in known risk behaviors and factors. Ecological analyses of major CVD risk factors and mortality demonstrate high correlations between expected and observed mortality rates for the three main risk factors—smoking, serum cholesterol, and hypertension—and suggest that many of the regional variations are based on differences in conventional risk factors.

Behavioral Risk Factors

Tobacco

Over 1.3 billion people use tobacco worldwide, a number that is projected to increase to 1.6 billion by 2030. Tobacco use currently causes about 6.4 million deaths annually (11.5% of all deaths), ~2.4 million of which are CVD-related. If current smoking patterns continue, the global burden of disease attributable to tobacco will reach 10 million deaths by 2030. Although tobacco use has been greatest in HICs historically, consumption has shifted dramatically to LMICs in recent decades. Some of the highest tobacco use now occurs in the East Asia and Pacific region. A unique feature of LMICs is easy access to smoking during the early stages of the epidemiologic transition due to the availability of relatively inexpensive tobacco products. In South Asia, the prominence of other locally produced forms of tobacco besides manufactured cigarettes makes control of consumption more challenging. Secondhand smoke is another well-established cause of CHD, responsible for 886,000 deaths of nonsmokers in 2015. Although smoking bans have both immediate and long-term benefits, implementation varies greatly between countries.

Diet

Total caloric intake per capita increases as countries develop. With regard to CVD, a key element of dietary change is an increase in intake of saturated animal fats and hydrogenated vegetable fats, which contain atherogenic *trans* fatty acids, along with a decrease in intake of plant-based foods and an increase in simple carbohydrates. Fat contributes <20% of calories in rural China and India, <30% in Japan, and well above 30% in the United States. Caloric contributions from fat appear to be falling in the HICs. In the United States, between 1971 and 2010, the percentage of calories derived from saturated fat decreased from 13 to 11%.

Physical Inactivity

The increased mechanization that accompanies the economic transition leads to a shift from physically demanding, agriculture-based work to largely sedentary industry- and office-based work. In the United States, approximately one-quarter of the population does not participate in any leisure-time physical activity, and only 51.6% of adults report engaging in physical activity three or more times a week. Physical inactivity is similarly high in other regions of the world and is increasing in countries that are rapidly urbanizing as part of their economic transition. In urban China, for example, the proportion of adults who participate in moderate- or high-level activity has decreased significantly, whereas proportion of those who participate in low-level activity has increased.

METABOLIC RISK FACTORS

Examination of trends in metabolic risk factors provides insight into changes in the CVD burden globally. Here we describe four metabolic risk factors—lipid levels, hypertension, obesity, and diabetes mellitus—using data from the Global Burden of Disease, Injuries, and Risk Factors Study (GBD 2015). The GBD project identified and compiled mortality and morbidity data from 187 countries from 1990 to 2015.

Lipid Levels

Worldwide, high cholesterol levels are estimated to play a role in 56% of ischemic heart disease events and 18% of strokes, amounting to 4.3 million deaths annually. Although mean population plasma cholesterol levels tend to rise as countries move through the epidemiologic transition, mean serum total cholesterol levels have decreased globally between 1980 and 2008 by 0.08 mmol/L per decade in men and 0.07 mmol/L per decade in women. In 2008, age-standardized mean total cholesterol was 4.64 mmol/L (179.4 mg/dL) in men and 4.76 mmol/L (184.2 mg/dL) in women. Large declines occurred in Australasia, North America, and Western Europe (0.19–0.21 mmol/L). Countries in the East Asia and Pacific region experienced increases of >0.08 mmol/L in both men and women. Social and individual changes that accompany urbanization clearly play a role because plasma cholesterol levels tend to be higher among urban residents than among rural residents. This shift is largely driven by greater consumption of dietary fats—primarily from animal products and processed vegetable oils—and decreased physical activity. In HICs, in general, mean population cholesterol levels are falling, whereas wide variation is seen in the LMICs.

Hypertension

Elevated blood pressure is an early indicator of the epidemiologic transition. Worldwide, ~62% of strokes and 49% of CHD are attributable to suboptimal (>115 mmHg systolic) blood pressure, which is believed to account for >7 million deaths annually. Remarkably, nearly half of this burden occurs among those with systolic blood pressure <140 mmHg, even as this level is used at the arbitrary threshold for defining hypertension in many national guidelines. Between 1980 and 2008, the age-standardized prevalence of uncontrolled prevalence has decreased even as the number of people with uncontrolled hypertension has increased. This trend results largely from population growth and aging. Rising mean population blood pressure also occurs as populations industrialize and move from rural to urban settings. For example, the prevalence of hypertension in urban India is 25%, but varies between 10 and 15% in rural communities. One major concern in LMICs is the high rate of undetected, and therefore untreated, hypertension. This may explain, at least in part, the higher stroke rates in these countries in relation to CHD rates during the early stages of the transition. The high rates of hypertension throughout Asia, especially undiagnosed hypertension, likely contribute to the high prevalence of hemorrhagic stroke in

Loading [Contrib]/a11y/accessibility-menu.js

the region. Globally, however, mean systolic blood pressure has decreased among both genders (0.8 mmHg per decade among men; 1.0 mmHg per decade among women).

Obesity

Although clearly associated with increased risk of CHD, much of the risk posed by obesity may be mediated by other CVD risk factors, including hypertension, diabetes mellitus, and lipid profile imbalances. According to the latest GBD data, nearly 1.46 billion adults were overweight (body mass index [BMI] ≥25 kg/m^2) in 2008, and ~508 million were obese (BMI ≥30 kg/m^2). Obesity is increasing throughout the world, particularly in developing countries, where the trajectories are steeper than those experienced by the developed countries. In many of the LMICs, obesity appears to coexist with undernutrition and malnutrition. Adolescents are at particular risk. Currently, 1 in 10 children are estimated to be overweight, a number that is increasing worldwide. Women are also more affected than men, with the number of overweight women generally exceeding underweight women based on data from 36 LMICs.

Diabetes Mellitus

As a consequence of, or in addition to, increasing BMI and decreasing levels of physical activity, worldwide rates of diabetes—predominantly type 2 diabetes—are on the rise. According to the most recent data from the GBD project, mean fasting plasma glucose levels have increased globally between 1980 and 2008. An estimated 346 million people worldwide have diabetes. The International Diabetes Foundation predicts that this number will reach 522 million by 2030, a yearly rate of growth that is higher than that of the world's adult population. Nearly 50% of people with diabetes are undiagnosed, and 80% live in LMICs. The highest regional prevalence for diabetes occurs in the Middle East and North Africa, where an estimated 12.5% of the adult population has diabetes. Future growth will also largely occur in this region, along with other LMICs in South Asia and sub-Saharan Africa. There appear to be clear genetic susceptibilities to diabetes mellitus of various racial and ethnic groups. For example, migration studies suggest that South Asians and Indians tend to be at higher risk than those of European extraction.

SUMMARY

Although CVD rates are declining in the HICs, they are increasing in virtually every other region of the world. The consequences of this preventable epidemic will be substantial on many levels, including individual mortality and morbidity, family suffering, and staggering economic costs.

Three complementary strategies can be used to lessen the impact. First, the overall burden of CVD risk factors can be lowered through population-wide public health measures, such as national campaigns against cigarette smoking, unhealthy diets, and physical inactivity. Second, it is important to identify higher risk subgroups of the population who stand to benefit the most from specific, low-cost prevention interventions, including screening for and treatment of hypertension and elevated cholesterol. Simple, low-cost interventions, such as the "polypill," a regimen of aspirin, a statin, and an antihypertensive agent, also need to be explored. Third, resources should be allocated to acute as well as secondary prevention interventions. For countries with limited resources, a critical first step in developing a comprehensive plan is better assessment of cause-specific mortality and morbidity, as well as the prevalence of the major preventable risk factors.

In the meantime, the HICs must continue to bear the burden of research and development aimed at prevention and treatment, being mindful of the economic limitations of many countries. The concept of the epidemiologic transition provides insight into how to alter the course of the CVD epidemic. The efficient transfer of low-cost preventive and therapeutic strategies could alter the natural course of this epidemic and thereby reduce the excess global burden of preventable CVD.

FURTHER READING

DANAEI G et al: National, regional, and global trends in systolic blood pressure since 1980: Systematic analysis of health examination surveys and epidemiological studies with 786 country-years and 5.4 million participants. Lancet 377:568, 2011. [PubMed: 21295844]
GAZIANO T, GAZIANO JM: Global burden of cardiovascular disease, in *Heart Disease: A Textbook of Cardiovascular Medicine*, 11th ed, E Braunwald (ed). Philadelphia, Elsevier, 2018; in press.
LOZANO R et al: Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. Lancet 380:2095, 2012. [PubMed: 23245604]
ROTH G et al: Global and regional patterns in cardiovascular mortality from 1990 to 2013. Circulation 132:1667, 2015. [PubMed: 26503749]

Loading [Contrib]/a11y/accessibility-menu.js

Chapter 234: Physical Examination of the Cardiovascular System

Patrick T. O'Gara; Joseph Loscalzo

INTRODUCTION

The approach to a patient with known or suspected cardiovascular disease begins with the time-honored traditions of a directed history and a targeted physical examination. The scope of these activities depends on the clinical context at the time of presentation, ranging from an elective ambulatory follow-up visit to a more urgent emergency department encounter. There has been a gradual decline in physical examination skills over the last two decades at every level, from student to faculty specialist, a development of great concern to both clinicians and medical educators. Classic cardiac findings are recognized by only a minority of internal medicine and family practice residents. Despite popular perceptions, clinical performance does not improve predictably as a function of experience; instead, the acquisition of new examination skills may become more difficult for a busy individual practitioner. Less time is now devoted to mentored cardiovascular examinations during the training of students and residents. One widely recognized outcome of these trends is the progressive overutilization of noninvasive imaging studies to establish the presence and severity of cardiovascular disease even when the examination findings imply a low pretest probability of significant pathology. Proponents of the use of hand-held ultrasound devices to identify and characterize structural cardiac disease have called for its incorporation into educational curricula. Techniques to improve bedside examination skills include repetition, patient-centered teaching conferences, visual display feedback of auscultatory events using Doppler echocardiographic imaging and simulation-based training.

The evidence base that links the findings from the history and physical examination to the presence, severity, and prognosis of cardiovascular disease has been established most rigorously for coronary artery disease, heart failure, and valvular heart disease. For example, observations regarding heart rate, blood pressure, signs of pulmonary congestion, and the presence of mitral regurgitation (MR) contribute importantly to bedside risk assessment in patients with acute coronary syndromes. Observations from the physical examination in this setting can inform clinical decision-making before the results of cardiac biomarker testing are known. The prognosis of patients with systolic heart failure can be predicted on the basis of the jugular venous pressure (JVP) and the presence or absence of a third heart sound (S_3). Accurate characterization of cardiac murmurs provides important insight into the natural history of many valvular and congenital heart lesions. Finally, the important role played by the physical examination in enhancing the clinician-patient relationship cannot be overstated.

THE GENERAL PHYSICAL EXAMINATION

Any examination begins with an assessment of the general appearance of the patient, with notation of age, posture, demeanor, and overall health status. Is the patient in pain or resting quietly, dyspneic or diaphoretic? Does the patient choose to avoid certain body positions to reduce or eliminate pain, as might be the case with suspected acute pericarditis? Are there clues indicating that dyspnea may have a pulmonary cause, such as a barrel chest deformity with an increased anterior-posterior diameter, tachypnea, and pursed-lip breathing? Skin pallor, cyanosis, and jaundice can be appreciated readily and provide additional clues. A chronically ill-appearing emaciated patient may suggest the presence of long-standing heart failure or another systemic disorder, such as a malignancy. Various genetic syndromes, often with cardiovascular involvement, can also be recognized easily, such as trisomy 21, Marfan syndrome, and Holt-Oram syndrome. Height and weight should be measured routinely, and both body mass index and body surface area should be calculated. Knowledge of the waist circumference and the waist-to-hip ratio can be used to predict long-term cardiovascular risk. Mental status, level of alertness, and mood should be assessed continuously during the interview and examination.

Skin

Central cyanosis occurs with significant right-to-left shunting at the level of the heart or lungs, allowing deoxygenated blood to reach the systemic circulation. Peripheral cyanosis or acrocyanosis, in contrast, is usually related to reduced extremity blood flow due to small vessel constriction, as seen in patients with severe heart failure, shock, or peripheral vascular disease; it can be aggravated by the use of β-adrenergic blockers with unopposed α-mediated vasoconstriction. Differential cyanosis refers to isolated cyanosis affecting the lower but not the upper extremities in a patient with a large patent ductus arteriosus (PDA) and secondary pulmonary hypertension with right-to-left to shunting at the great vessel level. Hereditary telangiectasias on the lips, tongue, and mucous membranes, as part of the Osler-Weber-Rendu syndrome (hereditary hemorrhagic telangiectasia), resemble spider nevi and can be a source of right-to-left shunting when also present in the lung. Malar telangiectasias also are seen in patients with advanced mitral stenosis (MS) or scleroderma. An unusually tan or bronze discoloration of the skin may suggest hemochromatosis as the cause of the associated systolic heart failure. Jaundice, which may be visible first in the sclerae, has a broad differential diagnosis but, in the appropriate setting, can be consistent with advanced right heart failure and congestive hepatomegaly or anticoagulants or antiplatelet agents such as aspirin and P2Y12 receptor antagonists. Various hereditary lipid disorders sometimes are associated with subcutaneous xanthomas, particularly along the tendon sheaths or over the extensor surfaces of the extremities. Severe hypertriglyceridemia can be associated with eruptive xanthomatosis and lipemia retinalis. Palmar crease xanthomas are specific for type III hyperlipoproteinemia. Pseudoxanthoma elasticum, a disease associated with premature atherosclerosis, is manifested by a leathery, cobblestoned appearance of the skin in the axilla and neck creases and by angioid streaks on funduscopic examination. Extensive lentiginoses have been described in a variety of development delay—cardiovascular syndromes, including Carney's syndrome, which includes multiple atrial myxomas. Cutaneous manifestations of sarcoidosis such as lupus pernio and erythema nodosum may suggest this disease as a cause of an associated dilated cardiomyopathy, especially with heart block, intraventricular conduction delay, or ventricular tachycardia.

Head and Neck

Dentition and oral hygiene should be assessed in every patient both as a source of potential infection and as an index of general health. A high-arched palate is a feature of Marfan syndrome and other connective tissue disease syndromes. Bifid uvula has been described in patients with Loeys-Dietz syndrome, and orange tonsils are characteristic of Tangier disease. The ocular manifestations of hyperthyroidism have been well described. Many patients with congenital heart disease have associated hypertelorism, low-set ears, or micrognathia. Blue sclerae are a feature of osteogenesis imperfecta. An arcus senilis pattern lacks specificity as an index of coronary heart disease risk. The funduscopic examination is an often underused method by which to assess the microvasculature, especially among patients with established atherosclerosis, hypertension, or diabetes mellitus. A mydriatic agent may be necessary for optimal visualization. A funduscopic examination should be performed routinely in the assessment of patients with suspected endocarditis and those with a history of acute visual change. Branch retinal artery occlusion or visualization of a Hollenhorst plaque can narrow the differential diagnosis rapidly in the appropriate setting. Relapsing polychondritis may manifest as an inflamed pinna or, in its later stages, as a saddle-nose deformity because of destruction of nasal cartilage; granulomatosis with polyangiitis (Wegener's) can also lead to a saddle-nose deformity.

Chest

Midline sternotomy, left posterolateral thoracotomy, or infraclavicular scars at the site of pacemaker/defibrillator generator implantation should not be overlooked and may provide the first clue regarding an underlying cardiovascular disorder in patients unable to provide a relevant history. A prominent venous collateral pattern may suggest subclavian or vena caval obstruction. If the head and neck appear dusky and slightly cyanotic and the venous pressure is grossly elevated without visible pulsations, a diagnosis of superior vena cava syndrome should be entertained. Thoracic cage abnormalities have been well described among patients with connective tissue disease syndromes. They include pectus carinatum ("pigeon chest") and pectus excavatum ("funnel chest"). Obstructive lung disease is suggested by a barrel chest deformity, especially with tachypnea, pursed-lip breathing, and use of accessory muscles. The characteristically severe kyphosis and compensatory lumbar, pelvic, and knee flexion of ankylosing spondylitis should prompt careful auscultation for a murmur of aortic regurgitation (AR). Straight back syndrome refers to the loss of the normal kyphosis of the thoracic spine and has been described in patients with mitral valve prolapse (MVP) and its variants. In some patients with cyanotic congenital heart disease, the chest wall appears to be asymmetric, with anterior displacement of the left hemithorax. The respiratory rate and pattern should be noted during spontaneous breathing, with additional attention to depth, audible wheezing, and stridor. Lung examination can reveal adventitious sounds indicative of pulmonary edema, pneumonia, or pleuritis.

Abdomen

In some patients with advanced obstructive lung disease, the point of maximal cardiac impulse may be in the epigastrium. The liver is frequently enlarged and tender in patients with chronic heart failure. Systolic pulsations over the liver signify severe tricuspid regurgitation (TR). Splenomegaly may be a feature of infective endocarditis, particularly when symptoms have persisted for weeks or months. Ascites is a nonspecific finding but may be present with advanced chronic right heart failure, constrictive pericarditis, hepatic cirrhosis, or an intraperitoneal malignancy. The finding of an elevated JVP implies a cardiovascular etiology. In nonobese patients, the aorta typically is palpated between the epigastrium and the umbilicus. The sensitivity of palpation for the detection of an abdominal aortic aneurysm (pulsatile and expansile mass) decreases as a function of body size. Because palpation alone is not sufficiently accurate to establish this diagnosis, a screening ultrasound examination is advised when appropriate. The presence of an arterial bruit over the abdomen suggests high-grade atherosclerotic disease, although precise localization is difficult.

Extremities

The temperature and color of the extremities, the presence of clubbing, arachnodactyly, and pertinent nail findings can be surmised quickly during the examination. Clubbing implies the presence of central right-to-left shunting, although it has also been described in patients with endocarditis. Its appearance can range from cyanosis and softening of the root of the nail bed, to the classic loss of the normal angle between the base of the nail and the skin, to the skeletal and periosteal bony changes of hypertrophic osteoarthropathy, which is seen rarely in patients with advanced lung or liver disease. Patients with the Holt-Oram syndrome have an unopposable, "fingerized" thumb, whereas patients with Marfan syndrome may have arachnodactyly and a positive "wrist" (overlapping of the thumb and fifth finger around the wrist) or "thumb" (protrusion of the thumb beyond the ulnar aspect of the hand when the fingers are clenched over the thumb in a fist) sign. The Janeway lesions of endocarditis are nontender, slightly raised hemorrhages on the palms and soles, whereas Osler's nodes are tender, raised nodules on the pads of the fingers or toes. Splinter hemorrhages are classically identified as linear petechiae in the midposition of the nail bed and should be distinguished from the more common traumatic petechiae, which are seen closer to the distal edge.

Lower extremity or presacral edema in the setting of an elevated JVP defines volume overload and may be a feature of chronic heart failure or constrictive pericarditis. Lower extremity edema in the absence of jugular venous hypertension may be due to profound hypoalbuminemia as seen in nephrotic syndrome or liver failure. Other causes include lymphatic or venous obstruction or, more commonly, venous insufficiency, as would be further suggested by the appearance of varicosities, venous ulcers (typically medial in location), and brownish cutaneous discoloration from hemosiderin deposition (eburnation). Pitting edema can also be seen in patients who use dihydropyridine calcium channel blockers. A Homan's sign (posterior calf pain on active dorsiflexion of the foot against resistance) is neither specific nor sensitive for deep venous thrombosis. Muscular atrophy or the absence of hair along an extremity is consistent with severe arterial insufficiency or a primary neuromuscular disorder.

CARDIOVASCULAR EXAMINATION

Jugular Venous Pressure and Waveform

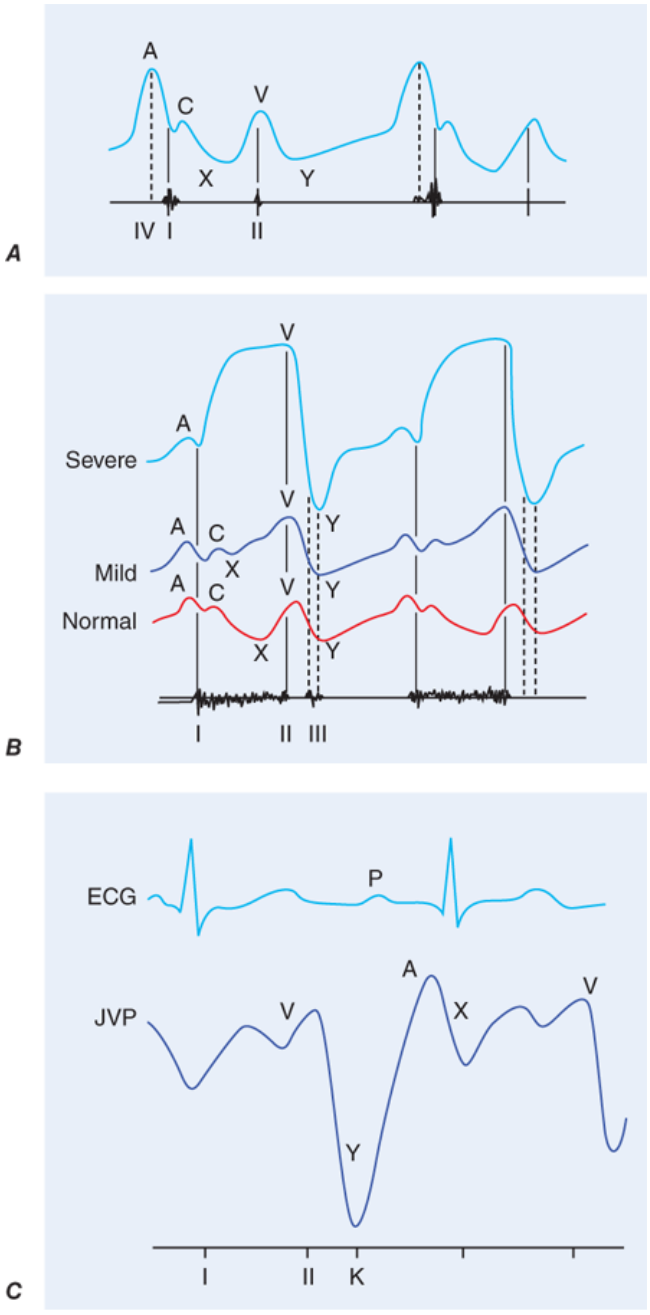
The JVP is the single most important bedside measurement from which to estimate the volume status. The internal jugular vein is preferred because the external jugular vein is valved and not directly in line with the superior vena cava and right atrium. Nevertheless, the external jugular vein has been used to discriminate between high and low central venous pressure (CVP) when tested among medical students, residents, and attending physicians. Precise estimation of the central venous or right atrial pressure from bedside assessment of the jugular venous waveform has proved difficult. Venous pressure traditionally has been measured as the vertical distance between the top of the jugular venous pulsation and the sternal inflection point (angle of Louis). A distance >4.5 cm at 30° elevation is considered abnormal. However, the actual distance between the mid-right atrium and the angle of Louis varies considerably as a function of both body size and the patient angle at which the assessment is made (30°, 45°, or 60°). The use of the sternal angle as a reference point leads to systematic underestimation of CVP,

and this method should be used less for semiquantification than to distinguish a normal from an abnormally elevated CVP. The use of the clavicle may provide an easier reference for standardization. Venous pulsations above this level in the sitting position are clearly abnormal, as the distance between the clavicle and the right atrium is at least 10 cm. The patient should always be placed in the sitting position, with the legs dangling below the bedside, when an elevated pressure is suspected in the semisupine position. It should also be noted that bedside estimates of CVP are made in centimeters of water, but must be converted to millimeters of mercury to provide correlation with accepted hemodynamic norms (1.36 cmH_2O = 1.0 mmHg).

The venous waveform sometimes can be difficult to distinguish from the carotid pulse, especially during casual inspection. Nevertheless, the venous waveform has several characteristic features, and its individual components can be appreciated in most patients (**Fig. 234-1**). The arterial pulsation is not easily obliterated with palpation; the venous waveform in patients with sinus rhythm is usually biphasic, while the carotid pulse is monophasic; and the jugular venous pulsation should change with changes in posture or inspiration (unless the venous pressure is quite elevated).

FIGURE 234-1

A. Jugular venous pulse wave tracing (*top*) with heart sounds (*bottom*). The A wave represents right atrial presystolic contraction and occurs just after the electrocardiographic P wave and just before the first heart sound (I). In this example, the A wave is accentuated and larger than normal due to decreased right ventricular compliance, as also suggested by the right-sided S_4 (IV). The C wave may reflect the carotid pulsation in the neck and/or an early systolic increase in right atrial pressure as the right ventricle pushes the closed tricuspid valve into the right atrium. The *x* descent follows the A wave just as atrial pressure continues to fall. The V wave represents atrial filling during ventricular systole and peaks at the second heart sound (II). The *y* descent corresponds to the fall in right atrial pressure after tricuspid valve opening. ***B.*** Jugular venous wave forms in mild (*middle*) and severe (*top*) tricuspid regurgitation, compared with normal, with phonocardiographic representation of the corresponding heart sounds below. With increasing degrees of tricuspid regurgitation, the waveform becomes "ventricularized." ***C.*** Electrocardiogram (ECG) (*top*), jugular venous waveform (JVP) (*middle*), and heart sounds (*bottom*) in pericardial constriction. Note the prominent and rapid *y* descent, corresponding in timing to the pericardial knock (K). *(From J Abrams: Synopsis of Cardiac Physical Diagnosis, 2nd ed. Boston, Butterworth Heinemann, 2001, pp 25–35.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The venous waveform is divided into several distinct peaks. The *a* wave reflects right atrial presystolic contraction and occurs just after the electrocardiographic P wave, preceding the first heart sound (S_1). A prominent *a* wave is seen in patients with reduced right ventricular compliance; a cannon *a* wave occurs with atrioventricular (AV) dissociation and right atrial contraction against a closed tricuspid valve. In a patient with a wide complex tachycardia, the appreciation of cannon *a* waves in the jugular venous waveform identifies the rhythm as ventricular in origin. The *a* wave is not present with atrial fibrillation. The *x* descent defines the fall in right atrial pressure after inscription of the *a* wave. The *c* wave, which occurs as the closed tricuspid valve is pushed into the right atrium during early ventricular systole, interrupts this *x* descent and is followed by a further descent. The *v* wave represents atrial filling (atrial diastole) and occurs during ventricular systole. The height of the *v* wave is determined by right atrial compliance as well as the volume of blood returning to the right atrium either antegrade from the cavae or retrograde through an incompetent tricuspid valve. In patients with TR, the *v* wave is accentuated and the subsequent fall in pressure (*y* descent) is rapid. With progressive degrees of TR, the *v* wave merges with the *c* wave, and the right atrial and jugular vein waveforms become “ventricularized.” The *y* descent, which follows the peak of the *v* wave, can become prolonged or blunted with obstruction to right ventricular inflow, as may occur with tricuspid stenosis or pericardial tamponade. Normally, the venous pressure should fall by at least 3 mmHg with inspiration. Kussmaul’s sign is defined by either a rise or a lack of fall of the JVP with inspiration and is classically associated with constrictive pericarditis, although it has been reported in patients with restrictive cardiomyopathy, massive pulmonary embolism, right ventricular infarction, and advanced left ventricular (LV) systolic heart failure. It is also a common, isolated finding in patients after cardiac surgery without other hemodynamic abnormalities.

Venous hypertension sometimes can be elicited by passive leg elevation or performance of the abdominojugular reflux maneuver. When these signs are positive, a volume-overloaded state with limited compliance of an overly distended or constricted venous system is present. Abdominojugular reflux is produced with firm and consistent pressure over the upper portion of the abdomen, preferably over the right upper quadrant, for >15 s. A positive response is defined by a sustained rise of >3 cm in the JVP during the application of firm abdominal pressure.

The response should be assessed after 10 s of continuous pressure to allow for respiratory artifacts and tensing of the abdominal muscles to subside. Patients must be coached to refrain from breath holding or a Valsalva-like maneuver during the procedure. Performance of the abdominojugular reflux maneuver is useful in predicting a pulmonary artery wedge pressure >15 mmHg in patients with heart failure.

Although the JVP estimates right ventricular filling pressure, it has a predictable relationship with the pulmonary artery wedge pressure. In a large study of patients with advanced heart failure, the presence of a right atrial pressure >10 mmHg (as predicted on bedside examination) had a positive value of 88% for the prediction of a pulmonary artery wedge pressure of >22 mmHg. In addition, an elevated JVP has prognostic significance in patients with both symptomatic heart failure and asymptomatic LV systolic dysfunction. The presence of an elevated JVP is associated with a higher risk of subsequent hospitalization for heart failure, death from heart failure, or both.

Assessment of Blood Pressure

Measurement of blood pressure usually is delegated to a medical assistant but should be repeated by the examining clinician. Accurate measurement depends on body position, arm size, time of measurement, place of measurement, device, device size, technique, and examiner. In general, physician-recorded blood pressures are higher than both nurse-recorded pressures and self-recorded pressures at home. Blood pressure is best measured in the seated position with the arm at the level of the heart and the feet on the floor with the back supported, using an appropriately sized cuff, after 5–10 min of relaxation. When it is measured in the supine position, the arm should be raised to bring it to the level of the mid-right atrium. The length and width of the blood pressure cuff bladder should be 80 and 40% of the arm's circumference, respectively. A common source of error in practice is to use an inappropriately small cuff, resulting in marked overestimation of true blood pressure, or an inappropriately large cuff, resulting in underestimation of true blood pressure. The cuff should be inflated to 30 mmHg above the expected systolic pressure and the pressure released at a rate of 2–3 mmHg/s. Systolic and diastolic pressures are defined by the first and fifth Korotkoff sounds, respectively. Very low (even 0 mmHg) diastolic blood pressures may be recorded in patients with chronic, severe AR or a large arteriovenous fistula because of enhanced diastolic "run-off." In these instances, both the phase IV and phase V Korotkoff sounds should be recorded. Blood pressure is best assessed at the brachial artery level, though it can be measured at the radial, popliteal, or pedal pulse level. In general, systolic pressure increases and diastolic pressure decreases when measured in more distal arteries. Blood pressure should be measured in both arms, and the difference should be <10 mmHg. A blood pressure differential that exceeds this threshold may be associated with atherosclerotic or inflammatory subclavian artery disease, supravalvular aortic stenosis, aortic coarctation, or aortic dissection. Systolic leg pressures are usually as much as 20 mmHg higher than systolic arm pressures. Greater leg–arm pressure differences are seen in patients with chronic severe AR as well as patients with extensive and calcified lower extremity peripheral arterial disease. The ankle-brachial index (systolic pressure in the dorsalis pedis and/or posterior tibial artery divided by the higher of the two brachial artery pressures) is a powerful predictor of long-term cardiovascular mortality.

The blood pressure measured in an office or hospital setting may not accurately reflect the pressure in other venues. "White coat hypertension" is defined by at least three separate clinic-based measurements >140/90 mmHg and at least two non-clinic-based measurements <140/90 mmHg in the absence of any evidence of target organ damage. Individuals with white coat hypertension may not benefit from drug therapy, although they may be more likely to develop sustained hypertension over time. Masked hypertension should be suspected when normal or even low blood pressures are recorded in patients with advanced atherosclerotic disease, especially when evidence of target organ damage is present or bruits are audible. Higher systolic blood pressures measured with a 24-h ambulatory blood pressure device are associated with a higher risk of cardiovascular disease and all-cause death independent of blood pressures measured in the outpatient setting.

Orthostatic hypotension is defined by a fall in systolic pressure >20 mmHg or in diastolic pressure >10 mmHg in response to assumption of the upright posture from a supine position within 3 min. There may also be a lack of a compensatory tachycardia, an abnormal response that suggests autonomic insufficiency, as may be seen in patients with diabetes or Parkinson's disease. Orthostatic hypotension is a common cause of postural lightheadedness/syncope and should be assessed routinely in patients for whom this diagnosis might pertain. It can be exacerbated by advanced age, dehydration, certain medications, food, deconditioning, and ambient temperature/humidity.

Arterial Pulse

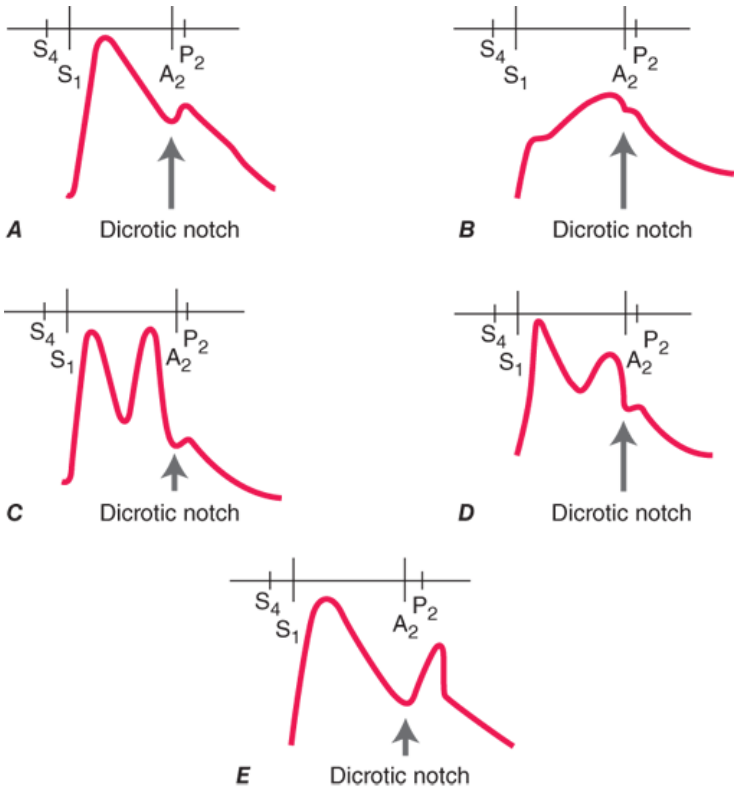
The carotid artery pulse occurs just after the ascending aortic pulse. The aortic pulse is best appreciated in the epigastrium, just above the level of the umbilicus. Peripheral arterial pulses that should be assessed routinely include the subclavian, brachial, radial, ulnar, femoral, popliteal, dorsalis pedis, and posterior tibial. In patients in whom the diagnosis of either temporal arteritis or polymyalgia rheumatica is suspected, the temporal arteries also should be examined. Although one of the two pedal pulses may not be palpable in up to 10% of normal subjects, the pair should be symmetric. The integrity of the arcuate system of the hand is assessed by Allen's test, which is performed routinely before instrumentation of the radial artery. The pulses should be examined for their symmetry, volume, timing, contour, amplitude, and duration. If necessary, simultaneous auscultation of the heart can help identify a delay in the arrival of an arterial pulse. Simultaneous palpation of the radial and femoral pulses may reveal a femoral delay in a patient with hypertension and suspected aortic coarctation. The carotid upstrokes should never be examined simultaneously or before listening for a bruit. Light pressure should always be used to avoid precipitation of carotid hypersensitivity syndrome and syncope in a susceptible elderly individual. The arterial pulse usually becomes more rapid and spiking as a function of its distance from the heart, a phenomenon that reflects the muscular status of the more peripheral arteries and the summation of the incident and reflected waves. In general, the character and contour of the arterial pulse depend on the stroke volume, ejection velocity, vascular compliance, and systemic vascular resistance. The pulse examination can be misleading in patients with reduced cardiac output and in those with stiffened arteries from aging, chronic hypertension, or peripheral arterial disease.

The character of the pulse is best appreciated at the carotid level (**Fig. 234-2**). A weak and delayed pulse (*pulsus parvus et tardus*) defines severe aortic stenosis (AS). Some patients with AS may also have a slow, notched, or interrupted upstroke (anacrotic pulse) with a thrill or shudder. With chronic severe AR, by contrast, the carotid upstroke has a sharp rise and rapid fall-off (Corrigan's or water-hammer pulse). Some patients with advanced AR may have a bifid or bisferiens pulse, in which two systolic peaks can be appreciated. A bifid pulse is also described in patients with hypertrophic obstructive cardiomyopathy (HOCM), with inscription of percussion and tidal waves. A bifid pulse is easily appreciated in patients on intraaortic balloon counterpulsation (IABP), in whom the second pulse is diastolic in timing.

FIGURE 234-2

Schematic diagrams of the configurational changes in carotid pulse and their differential diagnoses. Heart sounds are also illustrated. ***A.*** Normal. S_4, fourth heart sound; S_1, first heart sound; A_2, aortic component of second heart sound; P_2, pulmonic component of second heart

sound. ***B.*** Aortic stenosis. Anacrotic pulse with slow upstroke to a reduced peak. ***C.*** Bisferiens pulse with two peaks in systole. This pulse is rarely appreciated in patients with severe aortic regurgitation. ***D.*** Bisferiens pulse in hypertrophic obstructive cardiomyopathy. There is a rapid upstroke to the first peak (percussion wave) and a slower rise to the second peak (tidal wave). ***E.*** Dicrotic pulse with peaks in systole and diastole. This waveform may be seen in patients with sepsis or during intraaortic balloon counterpulsation with inflation just after the dicrotic notch. *(From K Chatterjee, W Parmley [eds]: Cardiology: An Illustrated Text/Reference. Philadelphia, Gower Medical Publishers, 1991.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pulsus paradoxus refers to a fall in systolic pressure >10 mmHg with inspiration that is seen in patients with pericardial tamponade but also is described in those with massive pulmonary embolism, hemorrhagic shock, severe obstructive lung disease, and tension pneumothorax. Pulsus paradoxus is measured by noting the difference between the systolic pressure at which the Korotkoff sounds are first heard (during expiration) and the systolic pressure at which the Korotkoff sounds are heard with each heartbeat, independent of the respiratory phase. Between these two pressures, the Korotkoff sounds are heard only intermittently and during expiration. The cuff pressure must be decreased slowly to appreciate the finding. It can be difficult to measure pulsus paradoxus in patients with tachycardia, atrial fibrillation, or tachypnea. A pulsus paradoxus may be palpable at the brachial artery or femoral artery level when the pressure difference exceeds 15 mmHg. This inspiratory fall in systolic pressure is an exaggerated consequence of interventricular dependence.

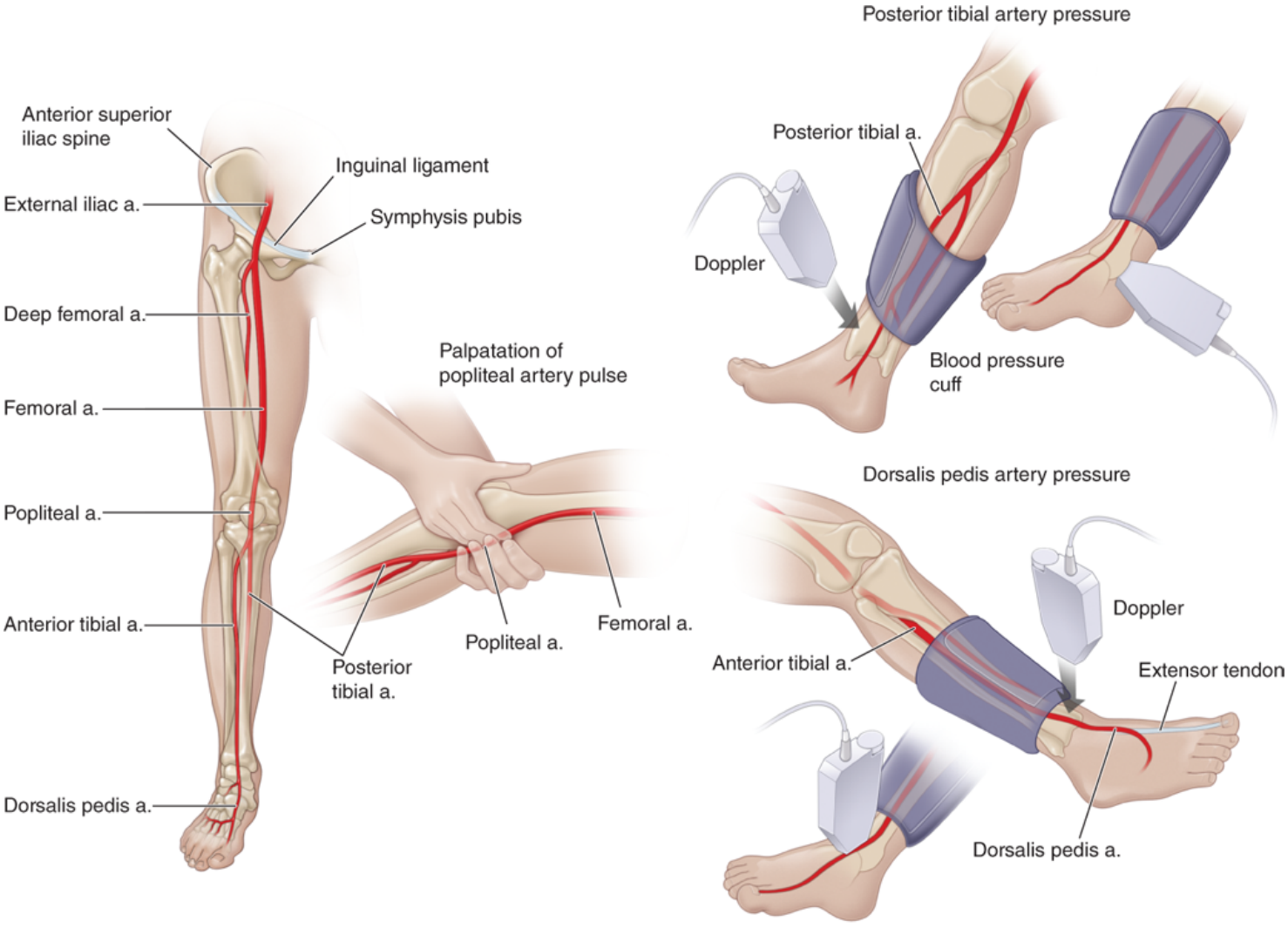
Pulsus alternans, in contrast, is defined by beat-to-beat variability of pulse amplitude. It is present only when every other phase I Korotkoff sound is audible as the cuff pressure is lowered slowly, typically in a patient with a regular heart rhythm and independent of the respiratory cycle. Pulsus alternans is seen in patients with severe LV systolic dysfunction and is thought to be due to cyclic changes in intracellular calcium and action potential duration. When pulsus alternans is associated with electrocardiographic T-wave alternans, the risk for an arrhythmic event appears to be increased.

Ascending aortic aneurysms can rarely be appreciated as a pulsatile mass in the right parasternal area. Appreciation of a prominent abdominal aortic pulse should prompt noninvasive imaging with ultrasound or computed tomography for better characterization. Femoral and/or popliteal artery aneurysms should be sought in patients with abdominal aortic aneurysm disease.

The level of a claudication-producing arterial obstruction can often be identified on physical examination (**Fig. 234-3**). For example, in a patient with calf claudication, a decrease in pulse amplitude between the common femoral and popliteal arteries will localize the obstruction to the level of the superficial femoral artery, although inflow obstruction above the level of the common femoral artery may coexist. Auscultation for carotid, subclavian, abdominal aortic, and femoral artery bruits should be routine. However, the correlation between the presence of a bruit and the degree of vascular obstruction is poor. A cervical bruit is a weak indicator of the degree of carotid artery stenosis; the absence of a bruit does not exclude the presence of significant luminal obstruction. If a bruit extends into diastole or if a thrill is present, the obstruction is usually severe. Another cause of an arterial bruit is an arteriovenous fistula with enhanced flow.

FIGURE 234-3

A. Anatomy of the major arteries of the leg. ***B.*** Measurement of the ankle systolic pressure. *(From NA Khan et al: JAMA 295:536, 2006.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The likelihood of significant lower extremity peripheral arterial disease increases with typical symptoms of claudication, cool skin, abnormalities on pulse examination, or the presence of a vascular bruit. Abnormal pulse oximetry (a >2% difference between finger and toe oxygen saturation) can be used to detect lower extremity peripheral arterial disease and is comparable in its performance characteristics to the ankle-brachial index.

Inspection and Palpation of the Heart

The LV apex beat may be visible in the midclavicular line at the fifth intercostal space in thin-chested adults. Visible pulsations anywhere other than this expected location are abnormal. The left anterior chest wall may heave in patients with an enlarged or hyperdynamic left or right ventricle. As noted previously, a visible right upper parasternal pulsation may be suggestive of ascending aortic aneurysm disease. In thin, tall patients and patients with advanced obstructive lung disease and flattened diaphragms, the cardiac impulse may be visible in the epigastrium and should be distinguished from a pulsatile liver edge.

Palpation of the heart begins with the patient in the supine position at 30° and can be enhanced by placing the patient in the left lateral decubitus position. The normal LV impulse is <2 cm in diameter and moves quickly away from the fingers; it is better appreciated at end expiration, with the heart closer to the anterior chest wall. Characteristics such as size, amplitude, and rate of force development should be noted.

Enlargement of the LV cavity is manifested by a leftward and downward displacement of an enlarged apex beat. A sustained apex beat is a sign of pressure overload, such as that which may be present in patients with AS or chronic hypertension. A palpable presystolic impulse corresponds to the fourth heart sound (S_4) and is indicative of reduced LV compliance and the forceful contribution of atrial contraction to ventricular filling. A palpable third sound (S_3), which is indicative of a rapid early filling wave in patients with heart failure, may be present even when the gallop itself is not audible. A large LV aneurysm may sometimes be palpable as an ectopic impulse, discrete from the apex beat. HOCM may very rarely cause a triple cadence beat at the apex with contributions from a palpable S_4 and the two components of the bisferiens systolic pulse.

Right ventricular pressure or volume overload may create a sternal lift. Signs of either TR (*cv* waves in the jugular venous pulse) and/or pulmonary arterial hypertension (a loud single or palpable P_2) would be confirmatory. The right ventricle can enlarge to the extent that left-sided

events cannot be appreciated. A zone of retraction between the right and LV impulses sometimes can be appreciated in patients with right ventricle pressure or volume overload when they are placed in the left lateral decubitus position. Systolic and diastolic thrills signify turbulent and high-velocity blood flow. Their locations help identify the origin of heart murmurs.

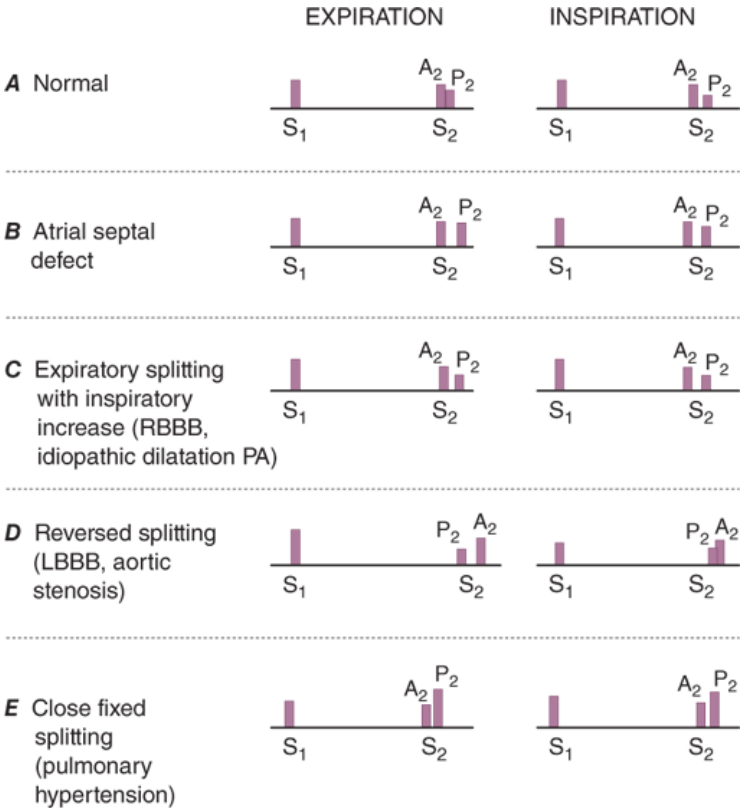
CARDIAC AUSCULTATION

Heart Sounds

Ventricular systole is defined by the interval between the first (S_1) and second (S_2) heart sounds **(Fig. 234-4)**. The first heart sound (S_1) includes mitral and tricuspid valve closure. Normal splitting can be appreciated in young patients and those with right bundle branch block, in whom tricuspid valve closure is relatively delayed. The intensity of S_1 is determined by the distance over which the anterior leaflet of the mitral valve must travel to return to its annular plane, leaflet mobility, LV contractility, and the PR interval. S_1 is classically loud in the early phases of rheumatic MS and in patients with hyperkinetic circulatory states or short PR intervals. S_1 becomes softer in the later stages of MS when the leaflets are rigid and calcified, after exposure to β-adrenergic receptor blockers, with long PR intervals, and with LV contractile dysfunction. The intensity of heart sounds, however, can be reduced by any process that increases the distance between the stethoscope and the responsible cardiac event, including mechanical ventilation, obstructive lung disease, obesity, pneumothorax, and a pericardial effusion.

FIGURE 234-4

Heart sounds. *A.* Normal. S_1, first heart sound; S_2, second heart sound; A_2, aortic component of the second heart sound; P_2, pulmonic component of the second heart sound. ***B.*** Atrial septal defect with fixed splitting of S_2. ***C.*** Physiologic but wide splitting of S_2 with right bundle branch block (RBBB). PA, pulmonary artery. ***D.*** Reversed or paradoxical splitting of S_2 with left bundle branch block (LBBB). ***E.*** Narrow splitting of S_2 with pulmonary hypertension. *(From NO Fowler: Diagnosis of Heart Disease. New York, Springer-Verlag, 1991, p 31.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Aortic and pulmonic valve closure constitutes the second heart sound (S_2). With normal or physiologic splitting, the A_2–P_2 interval increases with inspiration and narrows during expiration. This physiologic interval will widen with right bundle branch block because of the further delay in pulmonic valve closure and in patients with severe MR because of the premature closure of the aortic valve. An unusually narrowly split or even a singular S_2 is a feature of pulmonary arterial hypertension. Fixed splitting of S_2, in which the A_2–P_2 interval is wide and does not change during the respiratory cycle, occurs in patients with a secundum atrial septal defect (ASD). Reversed or paradoxical splitting refers to a pathologic delay in aortic valve closure, such as that which occurs in patients with left bundle branch block, right ventricular pacing, severe AS, HOCM, and acute myocardial ischemia. With reversed or paradoxical splitting, the individual components of S_2 are audible at end expiration, and their interval narrows with inspiration, the opposite of what would be expected under normal physiologic conditions. P_2 is considered loud when its intensity exceeds that of A_2 at the base, when it can be palpated in the area of the proximal main pulmonary artery (second left interspace), or when both components of S_2 can be appreciated at the lower left sternal border or apex. The intensity of A_2 and P_2 decreases with aortic and pulmonic stenosis (PS), respectively. In these conditions, a single S_2 may result.

Systolic Sounds

An ejection sound is a high-pitched early systolic sound that corresponds in timing to the upstroke of the carotid pulse. It usually is associated with congenital bicuspid aortic or pulmonic valve disease; however, ejection sounds are also sometimes audible in patients with isolated aortic or pulmonary root dilation and normal semilunar valves. The ejection sound that accompanies bicuspid aortic valve disease becomes softer and then inaudible as the valve calcifies and becomes more rigid. The ejection sound that accompanies PS moves closer to the first heart sound as the severity of the stenosis increases. In addition, the pulmonic ejection sound is the only right-sided acoustic event that decreases in intensity with inspiration. Ejection sounds are often heard more easily at the lower left sternal border than they are at the base. Nonejection sounds (clicks), which occur after the onset of the carotid upstroke, are related to MVP and may be single or multiple. The nonejection click may introduce a murmur. This click-murmur complex will move away from the first heart sound with maneuvers that increase ventricular preload, such as squatting. On standing, the click and murmur move closer to S_1.

Diastolic Sounds

The high-pitched opening snap (OS) of MS occurs after a very short interval after the second heart sound. The A_2–OS interval is inversely proportional to the height of the left atrial–left ventricular diastolic pressure gradient. The intensity of both S_1 and the OS of MS decreases with progressive calcification and rigidity of the anterior mitral leaflets. The pericardial knock (PK) is also high-pitched and occurs slightly later than the OS, corresponding in timing to the abrupt cessation of ventricular expansion after tricuspid valve opening and to an exaggerated *y* descent seen in the jugular venous waveform in patients with constrictive pericarditis. A tumor plop is a lower-pitched sound that rarely can be heard in patients with atrial myxoma. It may be appreciated only in certain positions and arises from the diastolic prolapse of the tumor across the mitral valve.

The third heart sound (S_3) occurs during the rapid filling phase of ventricular diastole. It can be a normal finding in children, adolescents, and young adults; however, in older patients, it signifies heart failure. A left-sided S_3 is a low-pitched sound best heard over the LV apex. A right-sided S_3 is usually better heard over the lower left sternal border and becomes louder with inspiration. A left-sided S_3 in patients with chronic heart failure is predictive of cardiovascular morbidity and mortality. Interestingly, an S_3 is equally prevalent among heart failure patients with and without LV systolic dysfunction.

The fourth heart sound (S_4) occurs during the atrial filling phase of ventricular diastole and indicates LV presystolic expansion. An S_4 is more common among patients who derive significant benefit from the atrial contribution to ventricular filling, such as those with chronic LV hypertrophy or active myocardial ischemia. An S_4 is not present with atrial fibrillation.

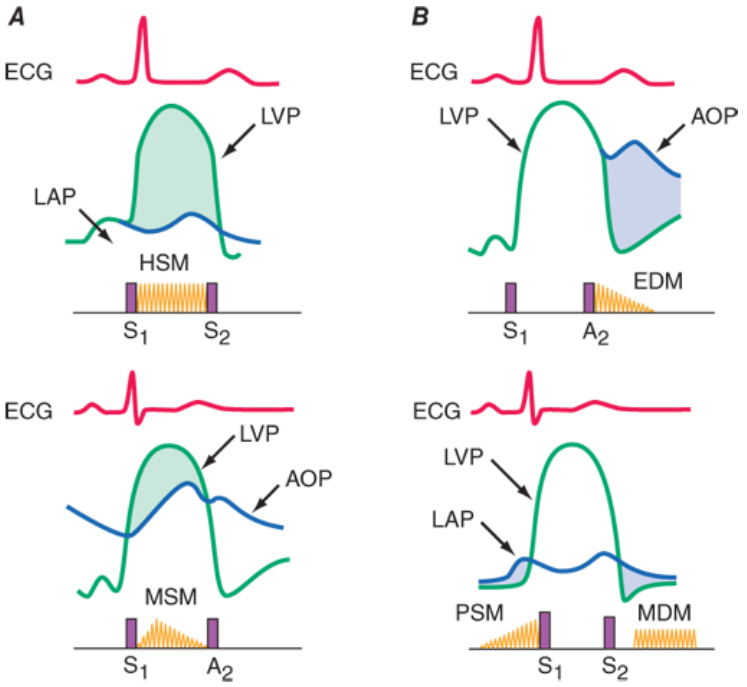
Cardiac Murmurs

Heart murmurs result from audible vibrations that are caused by increased turbulence and are defined by their timing within the cardiac cycle. Not all murmurs are indicative of structural heart disease, and the accurate identification of a benign or functional systolic murmur often can obviate the need for additional testing in healthy subjects. The duration, frequency, configuration, and intensity of a heart murmur are dictated by the magnitude, variability, and duration of the responsible pressure difference between two cardiac chambers, the two ventricles, or the ventricles and their respective great arteries. The intensity of a heart murmur is graded on a scale of 1 to 6; a thrill is present with murmurs of grade 4 or greater intensity. Other attributes of the murmur that aid in its accurate identification include its location, radiation, and response to bedside maneuvers. Although clinicians can detect and correctly identify heart murmurs with only fair reliability, a careful and complete bedside examination usually can identify individuals with valvular heart disease for whom transthoracic echocardiography and clinical follow-up are indicated and exclude subjects for whom no further evaluation is necessary.

Systolic murmurs can be early, mid, late, or holosystolic in timing **(Fig. 234-5)**. Acute severe MR results in a decrescendo early systolic murmur, the characteristics of which are related to the progressive attenuation of the LV to left atrial pressure gradient during systole because of the steep and rapid rise in left atrial pressure in this context. Severe MR associated with posterior leaflet prolapse or flail radiates anteriorly and to the base, where it can be confused with the murmur of AS. MR that is due to anterior leaflet involvement radiates posteriorly and to the axilla. With acute TR in patients with normal pulmonary artery pressures, an early systolic murmur that may increase in intensity with inspiration may be heard at the left lower sternal border, with regurgitant *cv* waves visible in the jugular venous pulse.

FIGURE 234-5

A. Top. Graphic representation of the systolic pressure difference (*green shaded area*) between left ventricle and left atrium with phonocardiographic recording of a holosystolic murmur (HSM) indicative of mitral regurgitation. ECG, electrocardiogram; LAP, left atrial pressure; LVP, left ventricular pressure; S_1, first heart sound; S_2, second heart sound. ***Bottom.*** Graphic representation of the systolic pressure gradient (*green shaded area*) between left ventricle and aorta in patient with aortic stenosis. A midsystolic murmur (MSM) with a crescendo-decrescendo configuration is recorded. AOP, aortic pressure. ***B. Top.*** Graphic representation of the diastolic pressure difference between the aorta and left ventricle (*blue shaded area*) in a patient with aortic regurgitation, resulting in a decrescendo, early diastolic murmur (EDM) beginning with A_2. ***Bottom.*** Graphic representation of the diastolic left atrial–left ventricular gradient (*blue areas*) in a patient with mitral stenosis with a mid-diastolic murmur (MDM) and late presystolic murmurs (PSM).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

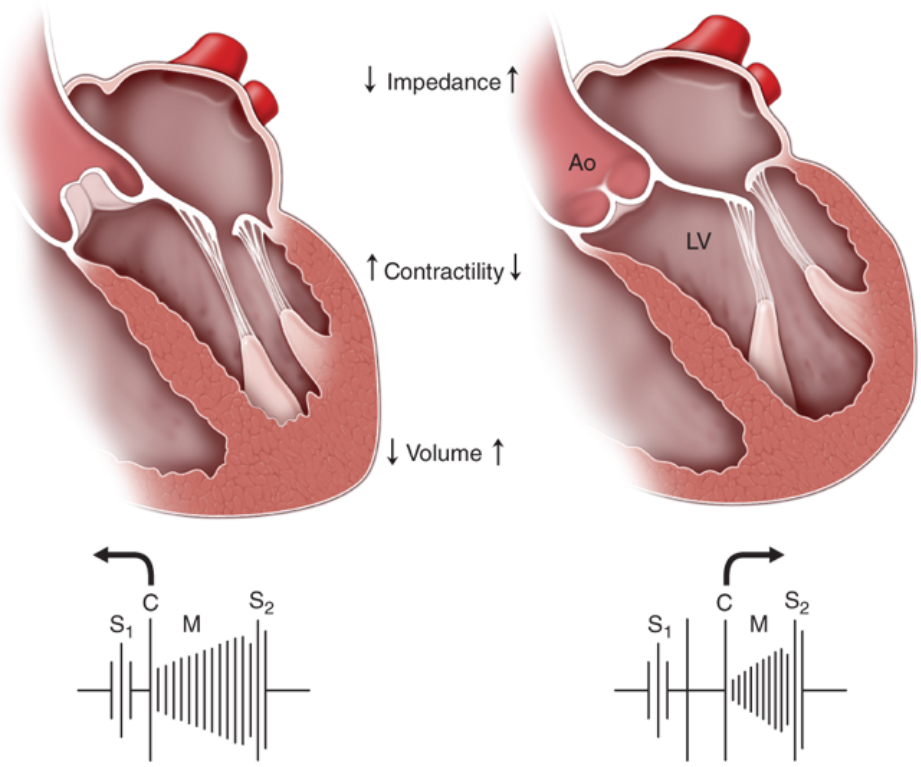
A midsystolic murmur begins after S_1 and ends before S_2; it is typically crescendo-decrescendo in configuration. AS is the most common cause of a midsystolic murmur in an adult. It is often difficult to estimate the severity of the valve lesion on the basis of the physical examination findings, especially in older hypertensive patients with stiffened carotid arteries or patients with low cardiac output in whom the intensity of the systolic heart murmur is misleadingly soft. Examination findings consistent with severe AS would include *parvus et tardus* carotid upstrokes, a late-peaking grade 3 or greater midsystolic murmur, a soft A_2, a sustained LV apical impulse, and an S_4. It is sometimes difficult to distinguish aortic sclerosis from more advanced degrees of valve stenosis. The former is defined by focal thickening and calcification of the aortic valve leaflets that is not severe enough to result in obstruction. These valve changes are associated with a Doppler jet velocity across the aortic valve of 2.5 m/s or less. Patients with aortic sclerosis can have grade 2 or 3 midsystolic murmurs identical in their acoustic characteristics to the murmurs heard in patients with more advanced degrees of AS. Other causes of a midsystolic heart murmur include pulmonic valve stenosis (with or without an ejection sound), HOCM, increased pulmonary blood flow in patients with a large ASD and left-to-right shunting, and several states associated with accelerated blood flow in the absence of structural heart disease, such as fever, thyrotoxicosis, pregnancy, anemia, and normal childhood/adolescence.

The murmur of HOCM has features of both obstruction to LV outflow and MR, as would be expected from knowledge of the pathophysiology of this condition. The systolic murmur of HOCM usually can be distinguished from other causes on the basis of its response to bedside maneuvers, including Valsalva, passive leg raising, and standing/squatting. In general, maneuvers that decrease LV preload (or increase LV contractility) will cause the murmur to intensify, whereas maneuvers that increase LV preload or afterload will cause a decrease in the intensity of the murmur. Accordingly, the systolic murmur of HOCM becomes louder during the strain phase of the Valsalva maneuver and after standing quickly from a squatting position. The murmur becomes softer with passive leg raising and when squatting. The murmur of AS is typically loudest in the second right interspace with radiation into the carotids, whereas the murmur of HOCM is best heard between the lower left sternal border and the apex. The murmur of PS is best heard in the second left interspace. The midsystolic murmur associated with enhanced pulmonic blood flow in the setting of a large ASD is usually loudest at the mid-left sternal border.

A late systolic murmur, heard best at the apex, indicates MVP. As previously noted, the murmur may or may not be introduced by a nonejection click. Differential radiation of the murmur, as previously described, may help identify the specific leaflet involved by the myxomatous process. The click-murmur complex behaves in a manner directionally similar to that demonstrated by the murmur of HOCM during the Valsalva and stand/squat maneuvers (**Fig. 234-6**). The murmur of MVP can be identified by the accompanying nonejection click.

FIGURE 234-6

Behavior of the click (C) and murmur (M) of mitral valve prolapse with changes in loading (volume, impedance) and contractility. S_1, first heart sound; S_2, second heart sound. With standing (left side of figure), volume and impedance decrease, as a result of which the click and murmur move closer to S_1. With squatting (right), the click and murmur move away from S_1 due to the increases in left ventricular volume and impedance (afterload). Ao, aorta; LV, left ventricle. *(Adapted from RA O'Rourke, MH Crawford: Curr Prob Cardiol 1:9, 1976.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Holosystolic murmurs are plateau in configuration and reflect a continuous and wide pressure gradient between the left ventricle and left atrium with chronic MR, the left ventricle and right ventricle with a ventricular septal defect (VSD), and the right ventricle and right atrium with TR. In contrast to acute MR, in chronic MR the left atrium is enlarged and its compliance is normal or increased to the extent that there is little if any further increase in left atrial pressure from any increase in regurgitant volume. The murmur of MR is best heard over the cardiac apex. The intensity of the murmur increases with maneuvers that increase LV afterload, such as sustained hand grip. The murmur of a VSD (without significant pulmonary hypertension) is holosystolic and loudest at the mid-left sternal border, where a thrill is usually present. The murmur of TR is loudest at the lower left sternal border, increases in intensity with inspiration (Carvallo's sign), and is accompanied by visible *cv* waves in the jugular venous wave form and, on occasion, by pulsatile hepatomegaly.

Diastolic Murmurs

In contrast to some systolic murmurs, diastolic heart murmurs always signify structural heart disease (Fig. 234-5). The murmur associated with acute, severe AR is relatively soft and of short duration because of the rapid rise in LV diastolic pressure and the progressive diminution of the aortic-LV diastolic pressure gradient. In contrast, the murmur of chronic severe AR is classically heard as a decrescendo, blowing diastolic murmur along the left sternal border in patients with primary valve pathology and sometimes along the right sternal border in patients with primary aortic root pathology. With chronic AR, the pulse pressure is wide and the arterial pulses are bounding in character. These signs of significant diastolic run-off are absent in the acute phase. The murmur of pulmonic regurgitation is also heard along the left sternal border. It is most commonly due to pulmonary hypertension and enlargement of the annulus of the pulmonic valve. S_2 is single and loud and may be palpable. There is a right ventricular/parasternal lift that is indicative of chronic right ventricular pressure overload. A less impressive murmur of PR is present after repair of tetralogy of Fallot or pulmonic valve atresia. In this postoperative setting, the murmur is softer and lower-pitched, and the severity of the accompanying pulmonic regurgitation can be underestimated significantly.

MS is the classic cause of a mid- to late diastolic murmur, which is best heard over the apex in the left lateral decubitus position, is low-pitched or rumbling, and is introduced by an OS in the early stages of the rheumatic disease process. Presystolic accentuation refers to an increase in the intensity of the murmur just before the first heart sound and occurs in patients with sinus rhythm. It is absent in patients with atrial fibrillation. The auscultatory findings in patients with rheumatic tricuspid stenosis typically are obscured by left-sided events, although they are similar in nature to those described in patients with MS. "Functional" mitral or tricuspid stenosis refers to the generation of mid-diastolic murmurs that are created by increased and accelerated transvalvular diastolic flow, even in the absence of valvular obstruction, in the setting of severe MR, severe TR, or a large ASD with left-to-right shunting. The Austin Flint murmur of chronic severe AR is a low-pitched mid- to late apical diastolic murmur that sometimes can be confused with MS. The Austin Flint murmur typically decreases in intensity after exposure to vasodilators, whereas the murmur of MS may be accompanied by an OS and also may increase in intensity after vasodilators because of the associated increase in cardiac output. Unusual causes of a mid-diastolic murmur include atrial myxoma, complete heart block, and acute rheumatic mitral valvulitis.

Continuous Murmur

A continuous murmur is predicated on a pressure gradient that persists between two cardiac chambers or blood vessels across systole and diastole. The murmurs typically begin in systole, envelop the second heart sound (S_2), and continue through some portion of diastole. They can often be difficult to distinguish from individual systolic and diastolic murmurs in patients with mixed valvular heart disease. The classic example of a continuous murmur is that associated with a PDA, which usually is heard in the second or third interspace at a slight distance from the sternal border. Other causes of a continuous murmur include a ruptured sinus of Valsalva aneurysm with creation of an aortic–right atrial or

right ventricular fistula, a coronary or great vessel arteriovenous fistula, and an arteriovenous fistula constructed to provide dialysis access. There are two types of benign continuous murmurs. The cervical venous hum is heard in children or adolescents in the supraclavicular fossa. It can be obliterated with firm pressure applied to the diaphragm of the stethoscope, especially when the subject turns his or her head toward the examiner. The mammary soufflé of pregnancy relates to enhanced arterial blood flow through engorged breasts. The diastolic component of the murmur can be obliterated with firm pressure over the stethoscope.

Dynamic Auscultation

Diagnostic accuracy can be enhanced by the performance of simple bedside maneuvers to identify heart murmurs and characterize their significance **(Table 234-1)**. Except for the pulmonic ejection sound, right-sided events increase in intensity with inspiration and decrease with expiration; left-sided events behave oppositely (100% sensitivity, 88% specificity). As previously noted, the intensity of the murmurs associated with MR, VSD, and AR will increase in response to maneuvers that increase LV afterload, such as hand grip and vasopressors. The intensity of these murmurs will decrease after exposure to vasodilating agents. Squatting is associated with an abrupt increase in LV preload and afterload, whereas rapid standing results in a sudden decrease in preload. In patients with MVP, the click and murmur move away from the first heart sound with squatting because of the delay in onset of leaflet prolapse at higher ventricular volumes. With rapid standing, however, the click and murmur move closer to the first heart sound as prolapse occurs earlier in systole at a smaller chamber dimension. The murmur of HOCM behaves similarly, becoming softer and shorter with squatting (95% sensitivity, 85% specificity) and longer and louder on rapid standing (95% sensitivity, 84% specificity). A change in the intensity of a systolic murmur in the first beat after a premature beat or in the beat after a long cycle length in patients with atrial fibrillation suggests valvular AS rather than MR, particularly in an older patient in whom the murmur of the AS may be well transmitted to the apex (Gallavardin effect). Of note, however, the systolic murmur of HOCM also increases in intensity in the beat after a premature beat. This increase in intensity of any LV outflow murmur in the beat after a premature beat relates to the combined effects of enhanced LV filling (from the longer diastolic period) and postextrasystolic potentiation of LV contractile function. In either instance, forward flow will accelerate, causing an increase in the gradient across the LV outflow tract (dynamic or fixed) and a louder systolic murmur. In contrast, the intensity of the murmur of MR does not change in a postpremature beat, because there is relatively little change in the nearly constant LV to left atrial pressure gradient or further alteration in mitral valve flow. Bedside exercise can sometimes be performed to increase cardiac output and, secondarily, the intensity of both systolic and diastolic heart murmurs. Most left-sided heart murmurs decrease in intensity and duration during the strain phase of the Valsalva maneuver. The murmurs associated with MVP and HOCM are the two notable exceptions. The Valsalva maneuver also can be used to assess the integrity of the heart and vasculature in the setting of advanced heart failure.

TABLE 234-1 Effects of Physiologic and Pharmacologic Interventions on the Intensity of Heart Murmurs and Sounds

Respiration Right-sided murmurs and sounds generally increase with inspiration, except for the PES. Left-sided murmurs and sounds are usually louder during expiration.

Valsalva maneuver Most murmurs decrease in length and intensity. Two exceptions are the systolic murmur of HOCM, which usually becomes much louder, and that of MVP, which becomes longer and often louder. After release of the Valsalva maneuver, right-sided murmurs tend to return to control intensity earlier than do left-sided murmurs.

After VPB or AF Murmurs originating at normal or stenotic semilunar valves increase in the cardiac cycle after a VPB or in the cycle after a long cycle length in AF. By contrast, systolic murmurs due to AV valve regurgitation do not change, diminish (papillary muscle dysfunction), or become shorter (MVP).

Positional changes With *standing*, most murmurs diminish, with two exceptions being the murmur of HOCM, which becomes louder, and that of MVP, which lengthens and often is intensified. With *squatting*, most murmurs become louder, but those of HOCM and MVP usually soften and may disappear. Passive leg raising usually produces the same results.

Exercise Murmurs due to blood flow across normal or obstructed valves (e.g., PS, MS) become louder with both isotonic and submaximal isometric (hand grip) exercise. Murmurs of MR, VSD, and AR also increase with hand grip exercise. However, the murmur of HOCM often decreases with nearly maximum hand grip exercise. Left-sided S_4 and S_3 sounds are often accentuated by exercise, particularly when due to ischemic heart disease.

Abbreviations: AF, atrial fibrillation; AR, aortic regurgitation; HOCM, hypertrophic obstructive cardiomyopathy; MR, mitral regurgitation; MS, mitral stenosis; MVP, mitral valve prolapse; PES, pulmonic ejection sound; PR, pulmonic regurgitation; PS, pulmonic stenosis; TR, tricuspid regurgitation; TS, tricuspid stenosis; VPB, ventricular premature beat; VSD, ventricular septal defect.

Prosthetic Heart Valves

The first clue that prosthetic valve dysfunction may contribute to recurrent symptoms is frequently a change in the quality of the heart sounds or the appearance of a new murmur. The heart sounds with a bioprosthetic valve resemble those generated by native valves. A mitral bioprosthesis usually is associated with a grade 2 or 3 midsystolic murmur along the left sternal border (created by turbulence across the valve struts as they project into the LV outflow tract) as well as by a soft mid-diastolic murmur that occurs with normal LV filling. This diastolic murmur often can be heard only in the left lateral decubitus position and after exercise. A high-pitched or holosystolic apical murmur is indicative of pathologic MR due to a paravalvular leak and/or intra-annular bioprosthetic regurgitation from leaflet degeneration, for which additional imaging is indicated. Clinical deterioration can occur rapidly after the first expression of mitral bioprosthetic valve failure. A tissue valve in the aortic position is always associated with a grade 2 to 3 midsystolic murmur at the base or just below the suprasternal notch. A diastolic murmur of AR is abnormal in any circumstance. Mechanical valve dysfunction may first be suggested by a decrease in the intensity of either the opening or the closing sound. A high-pitched apical systolic murmur in patients with a mechanical mitral prosthesis and a diastolic decrescendo murmur in patients with a mechanical aortic prosthesis indicate paravalvular regurgitation. Patients with prosthetic valve thrombosis may present clinically with signs of shock, muffled heart sounds, and soft murmurs.

Pericardial Disease

A pericardial friction rub is nearly 100% specific for the diagnosis of acute pericarditis, although the sensitivity of this finding is not nearly as high, because the rub may come and go over the course of an acute illness or be very difficult to elicit. The rub is heard as a leathery or scratchy three-component or two-component sound, although it may be monophasic. Classically, the three components are ventricular systole, rapid early diastolic filling, and late presystolic filling after atrial contraction in patients in sinus rhythm. It is necessary to listen to the heart in several positions. Additional clues may be present from the history and 12-lead electrocardiogram. The rub typically disappears as the volume of any pericardial effusion increases. Pericardial tamponade can be diagnosed with a sensitivity of 98%, a specificity of 83%, and a positive

likelihood ratio of 5.9 (95% confidence interval 2.4–14) by a pulsus paradoxus that exceeds 12 mmHg in a patient with a large pericardial effusion.

The findings on physical examination are integrated with the symptoms previously elicited with a careful history to construct an appropriate differential diagnosis and proceed with indicated imaging and laboratory assessment. The physical examination is an irreplaceable component of the diagnostic algorithm and in selected patients can inform prognosis. Educational efforts to improve clinician competence eventually may result in cost saving, particularly if the indications for imaging can be influenced by the examination findings.

FURTHER READING

DRAZNER MH et al: Value of clinician assessment of hemodynamics in advanced heart failure: The ESCAPE trial. Circ Heart Fail 1:170, 2008. [PubMed: 19675681]
FANAROFF AC et al: Does this patient with chest pain have acute coronary syndrome? The Rational Clinical Examination Systematic Review. JAMA 314:1955, 2015. [PubMed: 26547467]
FANG JC, O'GARA PT: The history and physical examination. An evidence-based approach, in *Braunwald's Heart Disease. A Textbook of Cardiovascular Medicine*, 10th ed, DL Mann et al (eds). Philadelphia, Elsevier/Saunders, 2015, pp 95–113.
MARCUS GM et al: Association between phonocardiographic third and fourth heart sounds and objective measures of left ventricular function. JAMA 293:2238, 2005. [PubMed: 15886379]

Chapter 235: Electrocardiography

Ary L. Goldberger

INTRODUCTION

An *electrocardiogram* (ECG or EKG) is a graphic representation of electrical activity generated by the heart. The signals, detected by means of metal electrodes attached to the extremities and chest wall, are amplified and recorded by the *electrocardiograph*. ECG leads (derivations) are configured to display the instantaneous *differences* in potential between specific pairs of electrodes. The utility of the ECG derives from its immediate availability as a noninvasive, inexpensive, and highly versatile test. In addition to its use in detecting arrhythmias and myocardial ischemia, it may reveal findings related to life-threatening metabolic disturbances or to increased susceptibility to sudden cardiac arrest (**see also Chaps. 299 and 401**).

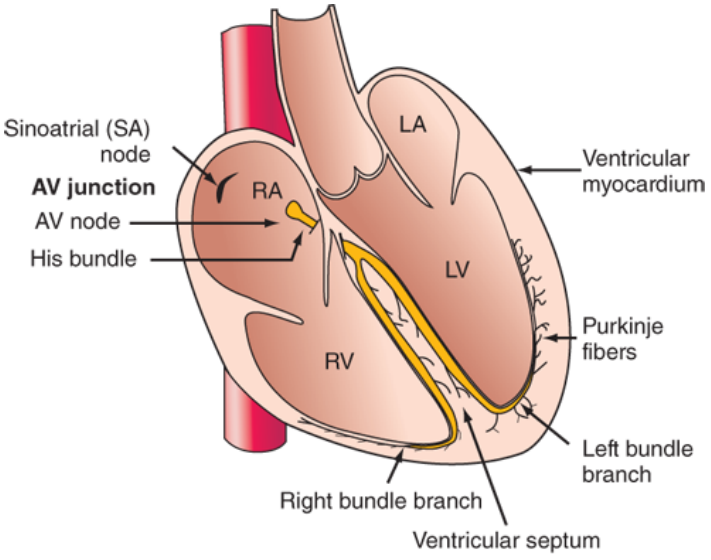
ELECTROPHYSIOLOGIC BACKGROUND

Depolarization of the heart is the initiating event for cardiac contraction. The electric currents that spread through the heart are produced by three components: cardiac pacemaker cells, specialized conduction tissue, and the heart muscle itself. The ECG records only the depolarization (stimulation) and repolarization (recovery) potentials generated by the "working" atrial and ventricular myocardium **(see also Chaps. 239 and 241)**.

The stimulus initiating the normal heartbeat originates in the *sinoatrial* (SA) *node* **(Fig. 235-1)**, which possesses spontaneous automaticity. Spread of the depolarization wave through the right and left atria induces contraction of these chambers. Next, the impulse stimulates specialized conduction tissues in the atrioventricular (AV) nodal and His-bundle areas; together, these two regions constitute the AV junction. The bundle of His fans bifurcates into two main branches, the right and left bundles, which rapidly transmit depolarization wavefronts in a synchronous way to the right and left ventricular myocardium by way of Purkinje fibers. The main left bundle bifurcates into left anterior and left posterior fascicle subdivisions. The depolarization wavefronts then spread through the ventricular wall, from endocardium to epicardium, triggering coordinated ventricular contraction. Since the cardiac depolarization and repolarization waves have directions and magnitudes, they can be represented by vectors.

FIGURE 235-1

Schematic of the cardiac conduction system.



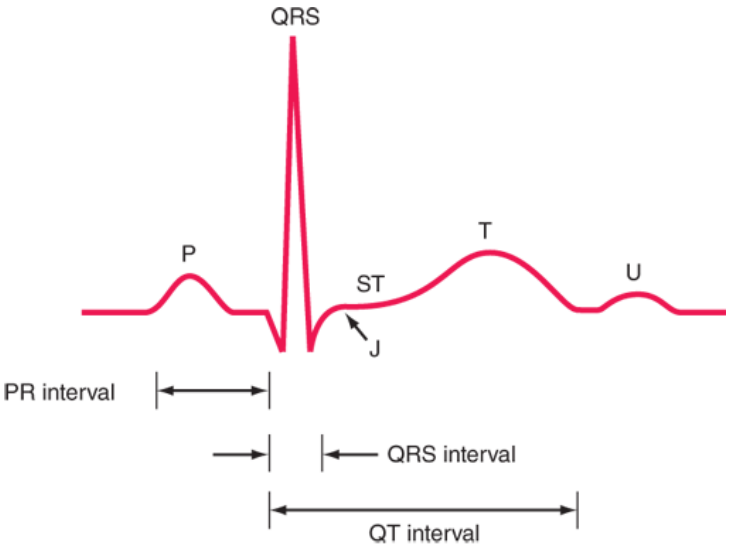
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ECG WAVEFORMS AND INTERVALS

The ECG waveforms are labeled alphabetically, beginning with the P wave, which represents atrial depolarization **(Fig. 235-2)**. The QRS complex represents ventricular depolarization, and the ST-T-U complex (ST segment, T wave, and U wave) represents ventricular repolarization. The J point is the junction between the end of the QRS complex and the beginning of the ST segment. Atrial repolarization waveforms ($ST\text{-}T_a$) are usually of too low in amplitude to be detected, it may become apparent in acute pericarditis, atrial infarction, and AV heart block.

Loading [Contrib]/a11y/accessibility-menu.js

Basic ECG waveforms and intervals. Not shown is the RR interval, the time between consecutive QRS complexes.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The QRS-T waveforms of the surface ECG correspond in a general way with the different phases of simultaneously obtained ventricular *action potentials*, the intracellular recordings from single myocardial fibers **(Chap. 239)**. The rapid upstroke (phase 0) of the action potential corresponds to the onset of QRS. The plateau (phase 2) corresponds to the isoelectric ST segment, and active repolarization (phase 3) corresponds to the inscription of the T wave. Factors that decrease the slope of phase 0 by impairing the influx of Na^+ (e.g., hyperkalemia and drugs such as flecainide) tend to increase QRS duration. Conditions that prolong phase 2 or 3 (amiodarone, hypocalcemia) increase the QT interval. In contrast, factors (e.g., hypercalcemia, digoxin) associated with shortening of ventricular repolarization duration shorten the QT.

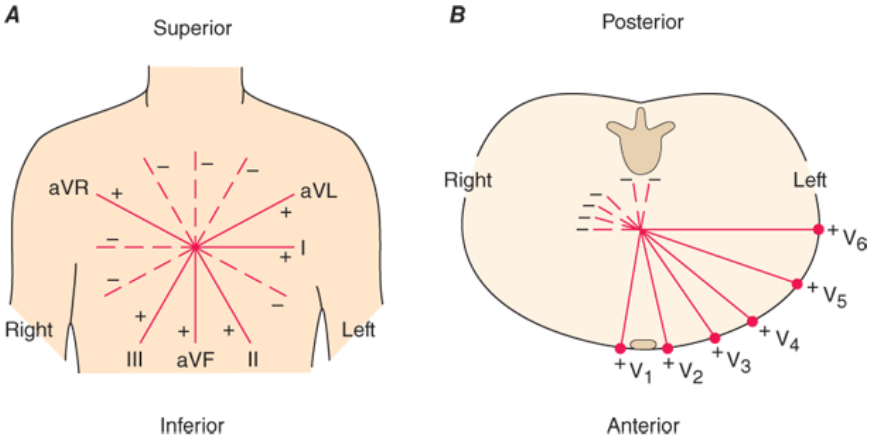
The ECG is usually recorded on graph paper divided into 1-mm^2 gridlike boxes. When the recording speed of 25 mm/s, the smallest (1 mm) horizontal divisions correspond to 0.04 (40 ms), with heavier lines at intervals of 0.20 s (200 ms). Vertically, the ECG graph measures the amplitude of a specific wave or deflection (1 mV = 10 mm with standard calibration; the voltage criteria for hypertrophy mentioned below are given in millimeters). There are four major sets of ECG intervals: RR, PR, QRS, and QT/QTc (Fig. 235-2). The instantaneous heart rate (beats per minute) can be computed from the interbeat (RR) interval by dividing the number of large (0.20 s) time units between consecutive R waves into 300 or the number of small (0.04 s) units into 1500. The PR interval measures the time (normally 120–200 ms) between atrial and ventricular depolarization, which includes the physiologic delay imposed by stimulation of cells in the AV junction area. The QRS interval (normally 100–110 ms or less) reflects the duration of ventricular depolarization. The QT interval subtends both ventricular depolarization and (primarily) repolarization times and varies inversely with the heart rate. A rate-related ("corrected") QT interval, QT_c, can be calculated as $QT/\sqrt{RR}$ and is normally ≤0.44 s. Some references give the QT_c upper normal limits as 0.45 s in men and 0.46 s in women. A number of other rate-correction formulas have been proposed, without consensus.

ECG LEADS

The 12 conventional ECG leads are divided into two groups: six limb (extremity) leads and six chest (precordial) leads. The limb leads record potentials transmitted onto the *frontal plane* **(Fig. 235-3A)**; the chest leads record potentials transmitted onto the *horizontal plane* **(Fig. 235-3B)**.

FIGURE 235-3

The six frontal plane (*A*) and six horizontal plane (*B*) leads provide a three-dimensional representation of cardiac electrical activity.

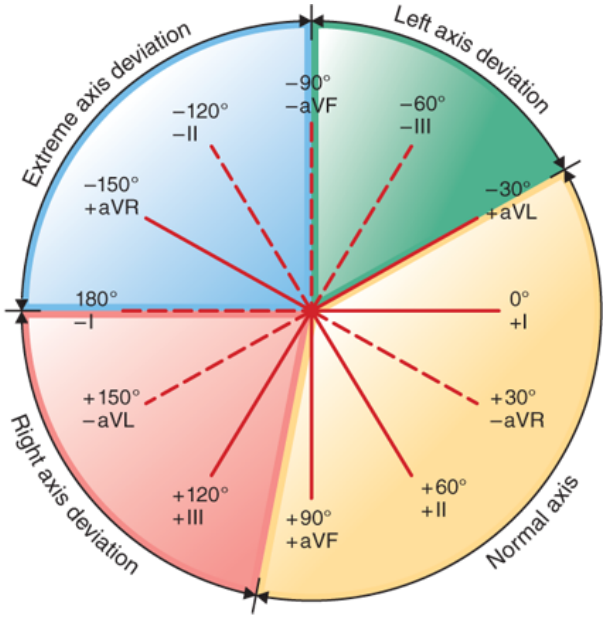


Loading [Contrib]/a11y/accessibility-menu.js
L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The orientation and polarity of the frontal plane leads are represented on a hexaxial diagram (**Fig. 235-4**). The six chest leads are obtained by exploring electrodes as shown in **Fig. 235-5**.

FIGURE 235-4

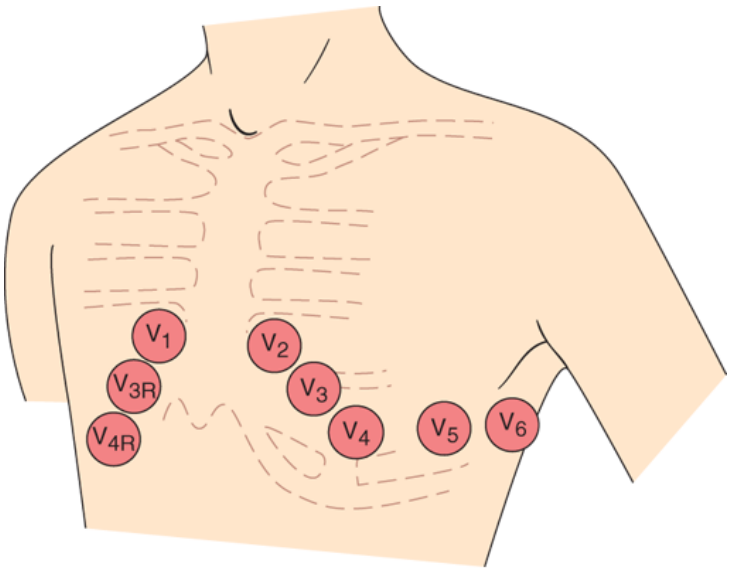
The frontal plane (limb or extremity) leads are represented on a hexaxial diagram. Each ECG lead has a specific spatial orientation and polarity. The positive pole of each lead axis (*solid line*) and the negative pole (*hatched line*) are designated by their angular position relative to the positive pole of lead I (0°). The mean electrical axis of the QRS complex is measured with respect to this display.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 235-5

The horizontal plane (chest or precordial) leads are obtained with electrodes in the locations shown. Additional posterior leads are sometimes placed on the same horizontal plane as V_4 to facilitate detection of acute posterolateral infarction (V_7, midaxillary line; V_8, posterior axillary line; and V_9, posterior scapular line). Right chest leads (V_3R–V_6R) may enhance detection of right ventricular involvement in the context of inferior infarction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Each lead is analogous to a different video camera angle “looking” at the same events—atrial and ventricular depolarization and repolarization—from different spatial orientations. The 12-lead ECG can be supplemented with additional leads in special circumstances. For example, right precordial leads V_6R to V_9R are useful in detecting evidence of acute right ventricular ischemia. Bedside monitors and ambulatory ECG (Holter and event recordings usually employ only one or two modified leads. The ECG leads are configured such that a positive (upright) deflection is

Loading [Contrib]/a11y/accessibility-menu.js

recorded in a lead if a wave of depolarization spreads toward the positive pole of that lead, and a negative deflection is recorded if the wave spreads toward the negative pole. If the *mean* orientation of the depolarization vector is at right angles to a particular lead axis, a biphasic (equally positive and negative) deflection will be recorded.

GENESIS OF THE NORMAL ECG

P WAVE

The normal atrial depolarization vector is oriented downward and toward the subject's left, reflecting the spread of depolarization from the sinus node to the right and then the left atrial myocardium. Since this vector points toward the positive pole of lead II and toward the negative pole of lead aVR, the normal P wave will be positive in lead II and negative in aVR. By contrast, activation of the atria from an ectopic pacemaker in the lower part of either atrium or in the AV junction region may produce retrograde P waves (negative in II, positive in aVR). The normal P wave in lead V_1 may be biphasic with a positive component reflecting right atrial depolarization, followed by a small (<1 mm^2) negative component reflecting left atrial depolarization.

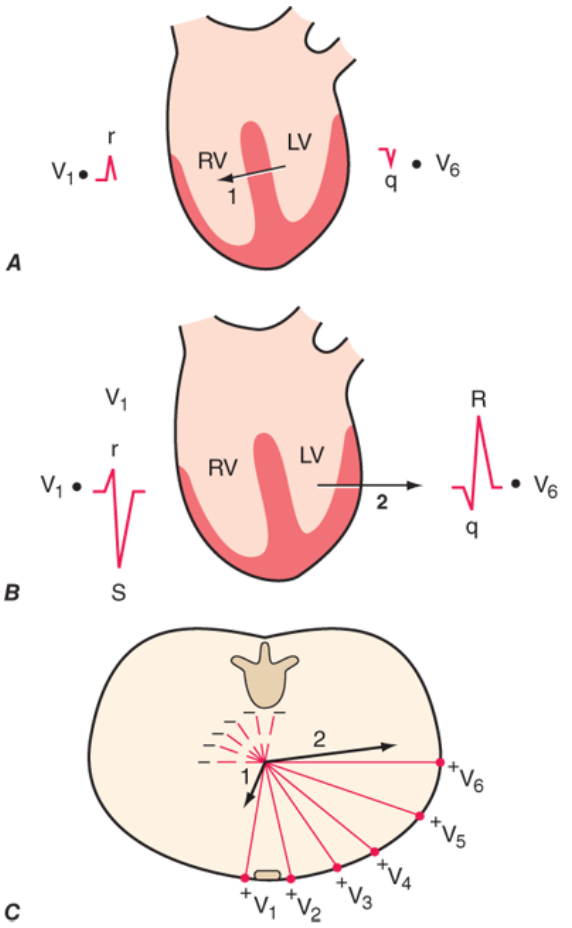
QRS COMPLEX

Normal ventricular depolarization proceeds as a rapid, continuous spread of activation wave fronts. This complex process can be divided into two major sequential phases, and each phase can be represented by a mean vector (**Fig. 235-6**). The first phase is depolarization of the interventricular septum from the left to the right and anteriorly (vector 1). The second results from the simultaneous depolarization of the right and left ventricles; it normally is dominated by the more massive left ventricle, so that vector 2 points leftward and posteriorly. Therefore, a right precordial lead (V_1) will record this biphasic depolarization process with a small positive deflection (septal r wave) followed by a larger negative deflection (S wave). A left precordial lead, for example, V_6, will record the same sequence with a small negative deflection (septal q wave) followed by a relatively tall positive deflection (R wave). Intermediate leads show a relative increase in R-wave amplitude (normal R-wave progression) and a decrease in S-wave amplitude progressing across the chest from right to left. The lead where the R and S waves are of about equal amplitude is referred to as the *transition zone* (usually V_3 or V_4) (**Fig. 235-7**).

FIGURE 235-6

Ventricular depolarization can be divided into two major phases, each represented by a vector. *A*. The first phase (*arrow 1*) denotes depolarization of the ventricular septum, beginning on the left side and spreading to the right. This process is represented by a small "septal" r wave in lead V_1 and a small septal q wave in lead V_6. ***B*.** Simultaneous depolarization of the left and right ventricles (LV and RV) constitutes the second phase. Vector 2 is oriented to the left and posteriorly, reflecting the electrical predominance of the LV. ***C*.** Vectors (*arrows*) representing these two phases are shown in reference to the horizontal plane leads. *(After AL Goldberger et al: Goldberger's Clinical Electrocardiography: A Simplified Approach, 9h ed. Philadelphia, Elsevier/Saunders, 2017.)*

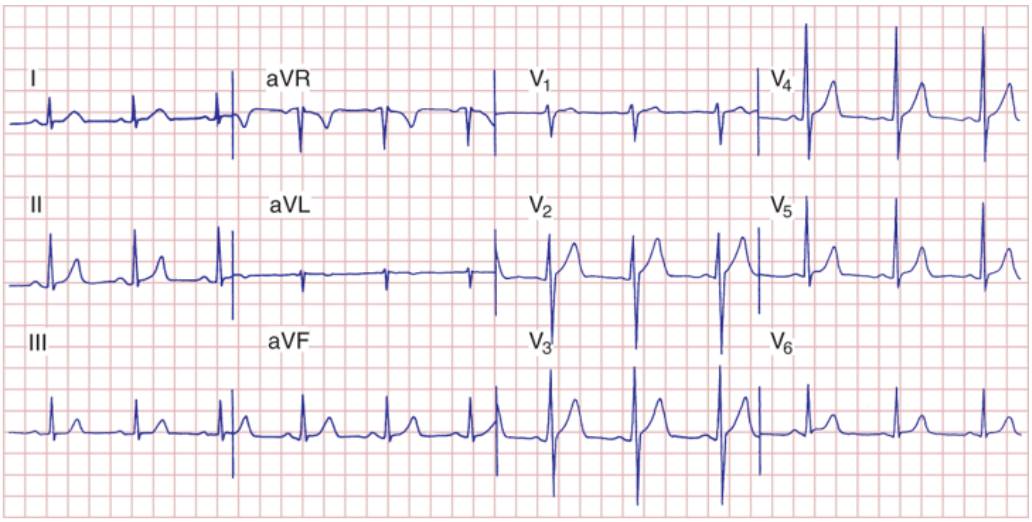
Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 235-7

Normal electrocardiogram from a healthy subject. Sinus rhythm is present with a heart rate of 75 beats per minute. PR interval is 0.16 s; QRS interval (duration) is 0.08 s; QT interval is 0.36 s; QT_c is 0.40 s; the mean QRS axis is about +70°. The precordial leads show normal R-wave progression with the transition zone (R wave = S wave) in lead V_3.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The QRS pattern in the extremity leads may vary considerably from one normal subject to another depending on the *electrical axis* of the QRS, which describes the mean orientation of the QRS vector with reference to the six frontal plane leads. Normally, the QRS axis ranges from −30° to +100° (Fig. 235-4). An axis more negative than –30° is referred to as *left axis deviation*, and an axis more positive than +90 to +100° is referred to as *right axis deviation*. Left axis deviation may occur as a normal variant but is more commonly associated with left ventricular hypertrophy, a block in the anterior fascicle of the left bundle system (left anterior fascicular block or hemiblock), or inferior myocardial infarction. Right axis

Loading [Contrib]/a11y/accessibility-menu.js

deviation also may occur as a normal variant (particularly in children and young adults), as a spurious finding due to reversal of the left and right arm electrodes, or in conditions such as right ventricular overload (acute or chronic), lateral infarction, dextrocardia, left pneumothorax, and left posterior fascicular block.

T WAVE AND U WAVE

Normally, the mean T-wave vector is oriented roughly concordant with the mean QRS vector (within about 45° in the frontal plane). Since depolarization and repolarization are electrically opposite processes, this normal QRS–T-wave vector concordance indicates that repolarization normally must proceed in the reverse direction from depolarization (i.e., from ventricular epicardium to endocardium). The normal U wave is a small, rounded deflection (≤1 mm) that follows the T wave and usually has the same polarity as the T wave. An abnormal increase in U-wave amplitude is most commonly due to drugs (e.g., dofetilide, amiodarone, sotalol, quinidine) or to hypokalemia. Very prominent U waves are a marker of increased susceptibility to *torsades de pointes* **(Chap. 241)**.

MAJOR ECG ABNORMALITIES

CARDIAC ENLARGEMENT AND HYPERTROPHY

Right atrial overload (acute or chronic) may lead to an increase in P-wave amplitude (≥2.5 mm) **(Fig. 235-8)**, previously referred to as "P-pulmonale." Left atrial overload typically produces a biphasic P wave in V_1 with a broad negative component or a broad (≥120 ms), often notched P wave in one or more limb leads (Fig. 235-8). This pattern, previously referred to as "P-mitrale," may also occur with left atrial conduction delays in the absence of actual atrial enlargement, leading to the more general designation of *left atrial abnormality*.

FIGURE 235-8

Right atrial (RA) overload may cause tall, peaked P waves in the limb or precordial leads. Left atrial (LA) abnormality may cause broad, often notched P waves in the limb leads and a biphasic P wave in lead V_1 with a prominent negative component representing delayed depolarization of the LA. *(After MK Park, WG Guntheroth: How to Read Pediatric ECGs, 4th ed. St. Louis, Mosby/Elsevier, 2006.)*

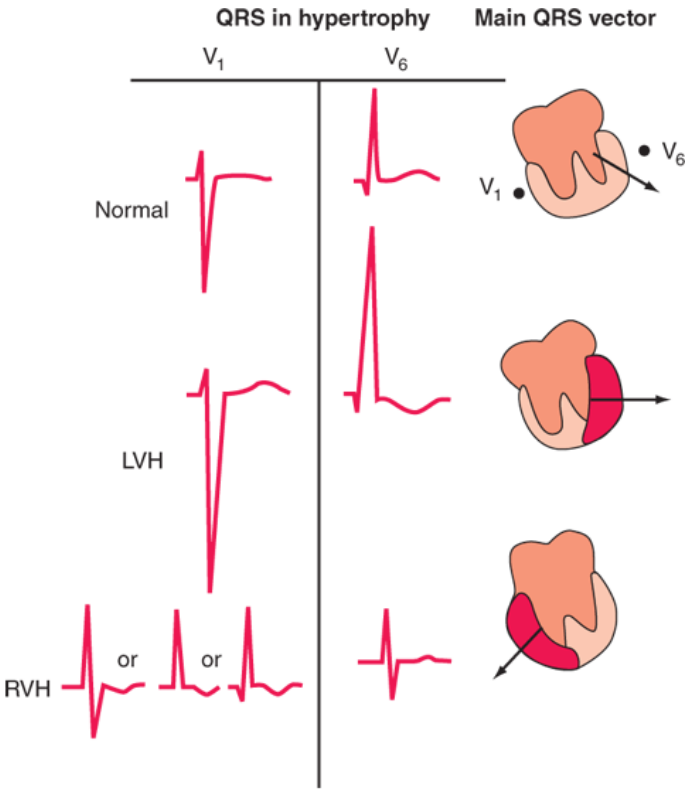
Image not available.

Right ventricular hypertrophy due to a sustained, severe pressure load (e.g., due to tight pulmonic valve stenosis or certain pulmonary artery hypertension syndromes) is characterized by a relatively tall R wave in lead V_1 (R ≥ S wave), usually with right axis deviation **(Fig. 235-9)**; alternatively, there may be a qR pattern in V_1 or V_3R. ST depression and T-wave inversion in the right-to-midprecordial leads are also often present. This pattern, formerly called right ventricular "strain," is attributed to repolarization abnormalities in acutely or chronically overloaded muscle. Prominent S waves may occur in the left lateral precordial leads. Right ventricular hypertrophy due to ostium secundum–type atrial septal defects, with the accompanying right ventricular volume overload, is commonly associated with an incomplete or complete right bundle branch block pattern with a rightward QRS axis.

FIGURE 235-9

Left ventricular hypertrophy (LVH) increases the amplitude of electrical forces directed to the left and posteriorly. In addition, repolarization abnormalities may cause ST-segment depression and T-wave inversion in leads with a prominent R wave. Right ventricular hypertrophy (RVH) may shift the QRS vector to the right; this effect usually is associated with an R, RS, or qR complex in lead V_1. T-wave inversions may be present in right precordial leads.

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Acute cor pulmonale due to pulmonary embolism **(Chap. 273)**, for example, may be associated with a normal ECG or a variety of abnormalities. Sinus tachycardia is the most common arrhythmia, although other tachyarrhythmias, such as atrial fibrillation or flutter, may occur. The QRS axis may shift to the right, sometimes in concert with the so-called $S_1Q_3T_3$ pattern (prominence of the S wave in lead I and the Q wave in lead III, with T-wave inversion in lead III). Acute right ventricular dilation also may be associated with slow R-wave progression and ST-T abnormalities in V_1 to V_4 simulating acute anterior infarction. A right ventricular conduction disturbance may appear.

Chronic cor pulmonale due to obstructive lung disease **(Chap. 252)** usually does not produce the classic ECG patterns of right ventricular hypertrophy noted above. Instead of tall right precordial R waves, chronic lung disease more typically is associated with small R waves in right-to-midprecordial leads (slow R-wave progression) due in part to downward displacement of the diaphragm and the heart. Low-voltage complexes are commonly present, owing to hyperaeration.

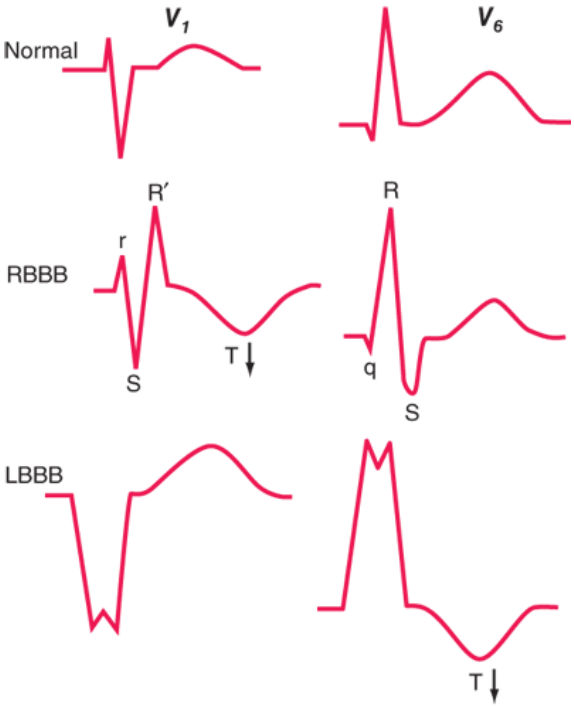
Multiple voltage criteria for *left ventricular hypertrophy* (Fig. 235-9) have been proposed on the basis of the presence of tall left precordial R waves and deep right precordial S waves (e.g., SV_1 + [RV_5 or RV_6] >35 mm). Repolarization abnormalities (ST depression with T-wave inversions, formerly called the left ventricular "strain" pattern) also may appear in leads with prominent R waves. However, prominent precordial voltages may occur as a normal variant, especially in athletic or young individuals. Left ventricular hypertrophy may increase limb lead voltage with or without increased precordial voltage (e.g., RaVL + SV_3 >20 mm in women and >28 mm in men). The presence of left atrial abnormality increases the likelihood of underlying left ventricular hypertrophy in cases with borderline voltage criteria. Left ventricular hypertrophy often progresses to incomplete or complete left bundle branch block. The sensitivity of conventional voltage criteria for left ventricular hypertrophy is decreased in obese persons and smokers. ECG evidence for left ventricular hypertrophy is a major noninvasive marker of increased risk of cardiovascular morbidity and mortality rates, including sudden cardiac death. However, because of false-positive and false-negative diagnoses, the ECG is of limited utility in diagnosing atrial or ventricular enlargement. More definitive information is provided by echocardiography **(Chap. 236)**.

BUNDLE BRANCH BLOCKS AND RELATED PATTERNS

Intrinsic impairment of conduction in either the right or the left bundle system (intraventricular conduction disturbances) leads to prolongation of the QRS interval. With complete bundle branch blocks, the widest QRS interval is ≥120 ms in duration; with incomplete blocks, the QRS interval is between about 110 and 120 ms. The QRS vector usually is oriented in the direction of the myocardial region where depolarization is delayed **(Fig. 235-10)**. Thus, with right bundle branch block, the terminal QRS vector is oriented to the right and anteriorly (rSR′ in V_1 and qRS in V_6, typically). Left bundle branch block alters both early and later phases of ventricular depolarization. The major QRS vector is directed to the left and posteriorly. In addition, the normal early left-to-right pattern of septal activation is disrupted such that septal depolarization proceeds from right to left as well. As a result, left bundle branch block generates wide, predominantly negative (QS) complexes in lead V_1 and entirely positive (R) complexes in V_6. A pattern identical to that of left bundle branch block, preceded by a sharp spike, is seen in most cases of electronic right ventricular pacing due to the relative delay in left ventricular activation.

FIGURE 235-10

Comparison of typical QRS-T patterns in right bundle branch block (RBBB) and left bundle branch block (LBBB) with the normal pattern in l
Loading [Contrib]/a11y/accessibility-menu.js
-wave inversions (*arrows*) in leads with an rSR′ complex with RBBB and in leads with a wide R wave with LBBB.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Bundle branch block may occur in a variety of conditions. In subjects without structural heart disease, right bundle branch block is seen more commonly than left bundle branch block. Right bundle branch block also occurs with heart disease, both congenital (e.g., atrial septal defect) and acquired (e.g., valvular, ischemic). Left bundle branch block is often a marker of one of four underlying conditions associated with increased risk of cardiovascular morbidity and mortality rates: coronary heart disease (frequently with impaired left ventricular function), hypertensive heart disease, aortic valve disease, and cardiomyopathy. Bundle branch blocks may be chronic or intermittent. A bundle branch block may be rate-related; for example, it often occurs when the heart rate exceeds some critical value.

Bundle branch blocks and depolarization abnormalities secondary to artificial pacemakers not only affect ventricular depolarization (QRS) but also are characteristically associated with *secondary repolarization* (ST-T) abnormalities. With bundle branch blocks, the T wave is typically opposite in polarity to the last deflection of the QRS (Fig. 235-10). This discordance of the QRS–T-wave vectors is caused by the altered sequence of repolarization that occurs secondary to altered depolarization. In contrast, *primary repolarization* abnormalities are independent of QRS changes and are related instead to actual alterations in the electrical properties of the myocardial fibers themselves (e.g., in the resting membrane potential or action potential duration), not just to changes in the sequence of repolarization. Ischemia, electrolyte imbalance, and drugs such as digitalis all cause such primary ST–T-wave changes. Primary and secondary T-wave changes may coexist. For example, T-wave inversions in the right precordial leads with left bundle branch block or in the left precordial leads with right bundle branch block may be important markers of underlying ischemia or other abnormalities. A distinctive abnormality simulating right bundle branch block with ST-segment elevations in the right chest leads is seen with the Brugada pattern **(Chap. 250)**.

Partial blocks (fascicular or "hemiblocks") in the left bundle system (left anterior or posterior fascicular blocks) generally do not prolong the QRS duration substantially but instead are associated with shifts in the frontal plane QRS axis (leftward or rightward, respectively). Left anterior fascicular block (QRS axis more negative than –45°) is probably the most common cause of marked left axis deviation in adults. In contrast, left posterior fascicular block (QRS axis more rightward than +110–120°) is extremely rare as an isolated finding and requires exclusion of other factors causing right axis deviation mentioned earlier. Intraventricular conduction delays also can be caused by extrinsic (toxic) factors that slow ventricular conduction, particularly hyperkalemia or drugs (e.g., class 1 antiarrhythmic agents, tricyclic antidepressants, phenothiazines). Prolongation of QRS duration does not necessarily indicate a conduction delay but may be due to *preexcitation* of the ventricles via a bypass tract, as in Wolff-Parkinson-White (WPW) patterns **(Chap. 244)** and related variants.

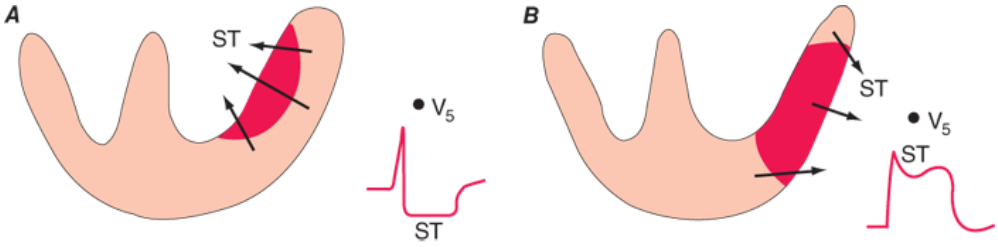
MYOCARDIAL ISCHEMIA AND INFARCTION

(See also Chap 269) The ECG is central to the diagnosis of acute and chronic ischemic heart disease. Ischemia exerts complex time-dependent effects on the electrical properties of myocardial cells. Severe, acute ischemia lowers the resting membrane potential and shortens the duration of the action potential. Such changes cause a voltage gradient between normal and ischemic zones. As a consequence, current flows between those regions. These currents of injury are represented on the surface ECG by deviation of the ST segment **(Fig. 235-11)**. When the acute ischemia is *transmural*, the ST vector usually is shifted in the direction of the outer (epicardial) layers, producing ST elevations and sometimes, in the earliest stages of ischemia, tall, positive so-called hyperacute T waves over the ischemic zone. With ischemia confined primarily to the *subendocardium*, the ST vector typically shifts toward the subendocardium and ventricular cavity, so that overlying (e.g., anterior precordial) leads show ST-segment depression (with ST elevation in lead aVR). Multiple factors affect the amplitude of acute ischemic ST deviations. Profound ST elevation or depression in multiple leads usually indicates very severe ischemia. From a clinical viewpoint, the division of acute myocardial infarction into ST-segment elevation and non-ST elevation types is useful since the consistent efficacy of emergency (minutes to hours) reperfusion therapy is limited to the former group; the evolving indications for acute reperfusion therapy in non-ST elevation MI are a focus of intensive investigation (**see Chap. 268**). Takostubo syndrome may exactly simulate the patterns of STEMI or non-STEMI **(Chap. 266)**.

FIGURE 235-11

Loading [Contrib]/a11y/accessibility-menu.js

Acute ischemia causes a current of injury. With predominant subendocardial ischemia (***A***), the resultant ST vector will be directed toward the inner layer of the affected ventricle and the ventricular cavity. Overlying leads therefore will record ST depression. With ischemia involving the outer ventricular layer (***B***) (transmural or epicardial injury), the ST vector will be directed outward. Overlying leads will record ST elevation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The ECG leads are usually more helpful in localizing regions of ST elevation than non-ST elevation ischemia. For example, acute transmural anterior (including apical and lateral) wall ischemia is reflected by ST elevations or increased T-wave positivity in one or more of the precordial leads (V_1–V_6) and leads I and aVL. Inferior wall ischemia produces changes in leads II, III, and aVF. "Posterior" wall ischemia (usually associated with lateral or inferior involvement) may be indirectly recognized by *reciprocal* ST depressions in leads V_1 to V_3 (thus constituting an ST elevation "equivalent" acute coronary syndrome). Right ventricular ischemia usually produces ST elevations in right-sided chest leads (Fig. 235-5). When ischemic ST elevations occur as the earliest sign of acute infarction, they typically are followed within a period ranging from hours to days by evolving T-wave inversions and often by Q waves occurring in the same lead distribution. Reversible transmural ischemia, for example, due to coronary vasospasm (Prinzmetal's angina) may cause transient ST-segment elevations without development of Q waves. Depending on the severity and duration of ischemia, the ST elevations may resolve completely in minutes or be followed by T-wave inversions that persist for hours or even days. Patients with ischemic chest pain who present with deep T-wave inversions in multiple precordial leads (e.g., V_1–V_4, I, and aVL) with or without cardiac enzyme elevations typically have severe obstruction in the left anterior descending coronary artery **(Fig. 235-12)**.

FIGURE 235-12

Severe anterior wall ischemia (with or without infarction) may cause prominent T-wave inversions in the precordial leads and in leads I and aVL. This pattern (sometimes referred to as Wellens T waves) is usually associated with a high-grade stenosis of the left anterior descending coronary artery.



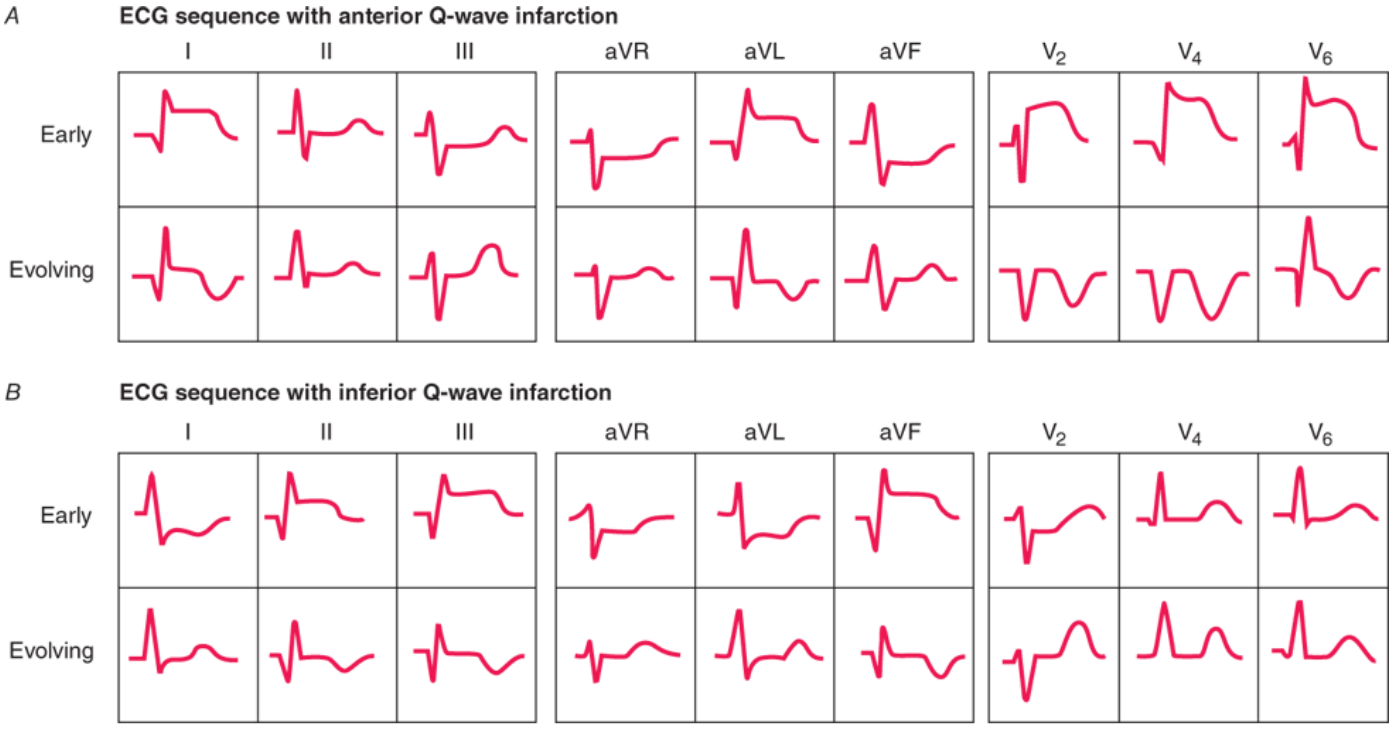
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

With infarction, depolarization (QRS) changes often accompany repolarization (ST-T) abnormalities. Necrosis of sufficient myocardial tissue may lead to decreased R-wave amplitude or abnormal Q waves (even in the absence of transmural ischemia) in the anterior or inferior leads **(Fig. 235-13)**. Abnormal Q waves were once considered markers of transmural myocardial infarction, whereas subendocardial infarcts were thought not to produce Q waves. However, careful ECG-pathology correlative studies have indicated that transmural infarcts may occur without Q waves and that subendocardial (nontransmural) infarcts sometimes may be associated with Q waves. Therefore, evolving or chronic infarcts are more appropriately classified as "Q-wave" or "non-Q-wave" **(Chap. A7)**. Loss of depolarization forces due to posterior or lateral infarction may cause reciprocal increases in R-wave amplitude in leads V_1 and V_2 without diagnostic Q waves in any of the conventional leads. (Additional leads V7-V9 may show acute changes.) In the weeks and months after infarction, these ECG changes may persist or begin to resolve. Complete normalization of the ECG after Q-wave infarction is uncommon but may occur, particularly with smaller infarcts. In contrast, ST-segment elevations that persist for several weeks or more after a Q-wave infarct usually correlate with a severe underlying wall motion disorder, although not necessarily a frank ventricular aneurysm.

FIGURE 235-13

Sequence of depolarization and repolarization changes with (***A***) acute anterior and (***B***) acute inferior wall Q-wave infarctions. With anterior infarcts, ST elevation in leads I and aVL and the precordial leads may be accompanied by reciprocal ST depressions in leads II, III, and aVF. Conversely, acute inferior (or posterolateral) infarcts may be associated with reciprocal ST depressions in leads V_1 to V_3. *(After AL Goldberger et al: Goldberger's Clinical Electrocardiography: A Simplified Approach, 9th ed. Philadelphia, Elsevier/Saunders, 2017.)*

Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The ECG has important limitations in both sensitivity and specificity in the diagnosis of ischemic heart disease. Although a single normal ECG does not exclude ischemia or even acute infarction, a normal ECG *throughout* the course of an acute infarct is distinctly uncommon. Prolonged chest pain without diagnostic ECG changes therefore should always prompt a careful search for other noncoronary causes of chest pain (**Chap. 11**). Furthermore, the diagnostic changes of acute or evolving ischemia are often masked by the presence of left bundle branch block, electronic ventricular pacemaker patterns, and Wolff-Parkinson-White preexcitation. However, clinicians continue to over-diagnose ischemia or infarction based on the presence of ST-segment elevations or depressions; T-wave inversions; tall, positive T waves; or Q waves *not* related to ischemic heart disease (pseudoinfarct patterns). For example, ST-segment elevations simulating ischemia may occur with acute pericarditis or myocarditis, as a normal variant (including the typical “early repolarization” pattern), or in a variety of other conditions (**Table 235-1**). Similarly, tall T waves do not invariably represent hyperacute ischemic changes but may also be caused by normal variants, hyperkalemia, cerebrovascular injury, among other causes.

Loading [Contrib]/a11y/accessibility-menu.js

TABLE 235-1 Differential Diagnosis of ST-Segment Elevations

Ischemia/myocardial infarction

- Noninfarction, transmural ischemia (Prinzmetal's angina, and probably Tako-tsubo syndrome, which may also exactly simulate classical acute infarction)
- Acute myocardial infarction
- Postmyocardial infarction (ventricular aneurysm pattern)

Acute pericarditis

Normal variants (including benign "early repolarization" patterns)

Left ventricular hypertrophy/left bundle branch block[a]

Other (rarer)

- Acute pulmonary embolism[a]
- Brugada patterns (right bundle branch block–like pattern with ST elevations in right precordial leads)[a]
- Class 1C antiarrhythmic drugs[a]
- DC cardioversion
- Hypercalcemia[a]
- Hyperkalemia[a]
- Hypothermia (J [Osborn] waves)
- Nonischemic myocardial injury
  - Myocarditis
  - Tumor invading left ventricle
  - Trauma to ventricles

[a]Usually localized to V_1–V_2 or V_3.

Source: Modified from AL Goldberger et al: *Goldberger's Clinical Electrocardiography: A Simplified Approach,* 9th ed. Philadelphia, Elsevier/Saunders, 2017.

ST-segment elevations and tall, positive T waves are common findings in leads V_1 and V_2 in left bundle branch block or left ventricular hypertrophy in the absence of ischemia. The differential diagnosis of Q waves includes physiologic or positional variants, ventricular hypertrophy, acute or chronic noncoronary myocardial injury, hypertrophic cardiomyopathy, and ventricular conduction disorders. Digoxin, ventricular hypertrophy, hypokalemia, and a variety of other factors may cause ST-segment depression mimicking subendocardial ischemia. Prominent T-wave inversion may occur with ventricular hypertrophy, cardiomyopathies, myocarditis, and cerebrovascular injury (particularly intracranial bleeds), among others causes.

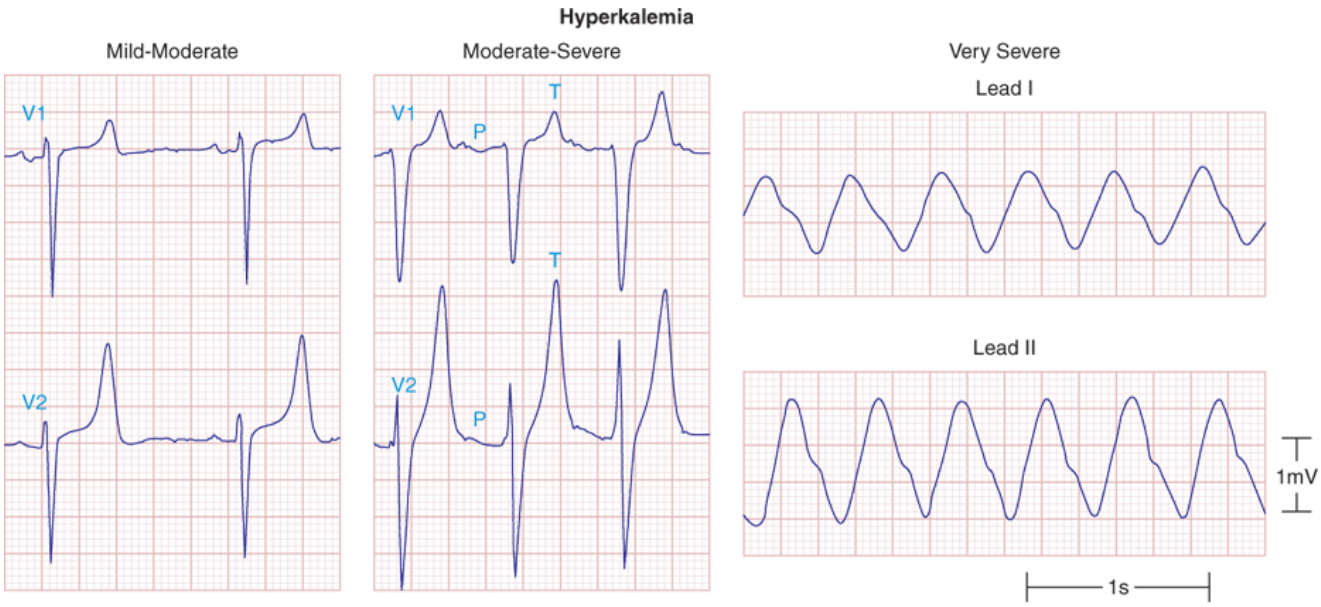
METABOLIC FACTORS AND DRUG EFFECTS

A variety of metabolic abnormalities and pharmacologic agents alter the ECG and, in particular, causing changes in repolarization (ST-T-U) and sometimes QRS prolongation. Certain life-threatening electrolyte disturbances may be diagnosed initially and monitored from the ECG. *Hyperkalemia* produces a sequence of changes **(Fig. 235-14)**, usually beginning with narrowing and peaking (tenting) of the T waves. Further elevation of extracellular K^+ leads to AV conduction disturbances, diminution in P-wave amplitude, and widening of the QRS interval. Severe hyperkalemia eventually causes cardiac arrest with a slow sinusoidal type of mechanism ("sine-wave" pattern) followed by asystole. *Hypokalemia* **(Fig. 235-15)** prolongs ventricular repolarization, often with prominent U waves. Prolongation of the QT interval is also seen with drugs that increase the duration of the ventricular action potential: class 1A antiarrhythmic agents and related drugs (e.g., quinidine, disopyramide, procainamide, tricyclic antidepressants, phenothiazines) and class III agents (e.g., amiodarone [Fig. 235-15], dofetilide, sotalol, ibutilide). Systemic *hypothermia* (Fig. 235-15) also prolongs repolarization, usually with a distinctive convex elevation of the J point (Osborn wave). Marked QT prolongation, sometimes with deep, wide T-wave inversions, may occur with intracranial bleeds, particularly subarachnoid hemorrhage ("CVA T-wave" pattern) (Fig. 235-15). *Hypocalcemia* typically prolongs the QT interval (ST portion), whereas *hypercalcemia* shortens it **(Fig. 235-16)**. Digitalis glycosides also shorten the QT interval, often with a characteristic "scooping" of the ST–T-wave complex (*digitalis effect*).

FIGURE 235-14

The earliest ECG change with hyperkalemia is usually peaking ("tenting") of the T waves. With further increases in the serum potassium concentration, the QRS complexes widen, the P waves decrease in amplitude and may disappear, and finally a sine-wave pattern leads to asystole unless emergency therapy is given. *(After AL Goldberger et al: Goldberger's Clinical Electrocardiography: A Simplified Approach, 9th ed. Philadelphia, Elsevier/Saunders, 2017.)*

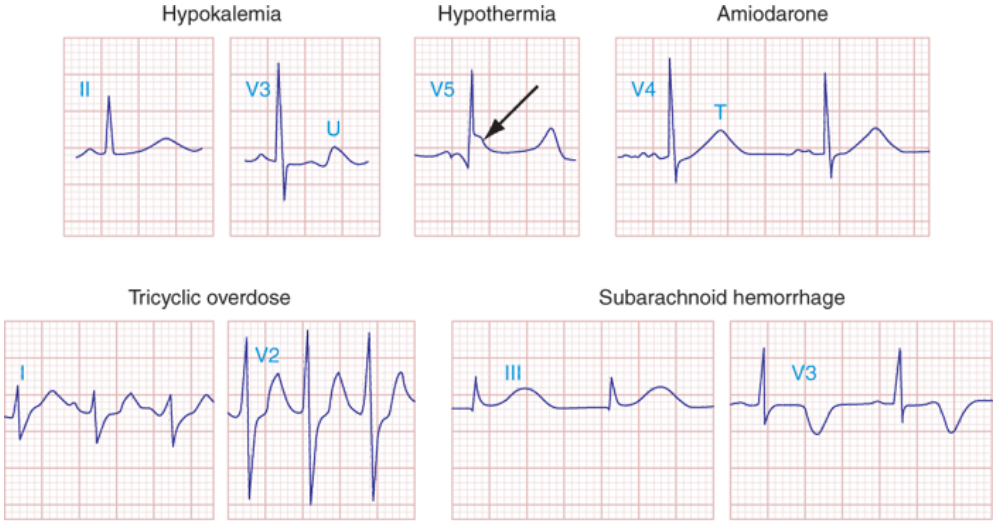
Loading [Contrib]/a11y/accessibility-menu.js



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 235-15

A variety of metabolic derangements, drug effects, and other factors may prolong ventricular repolarization with QT prolongation or prominent U waves. Prominent repolarization prolongation, particularly if due to hypokalemia, inherited “channelopathies,” or certain pharmacologic agents, indicates increased susceptibility to *torsades des pointes* ventricular tachycardia (**Chap. 249**). Marked systemic hypothermia is associated with a distinctive convex “hump” at the J point (Osborn wave, *arrow*) due to altered ventricular action potential characteristics. Note QRS and QT prolongation along with sinus tachycardia in the case of tricyclic antidepressant overdose.

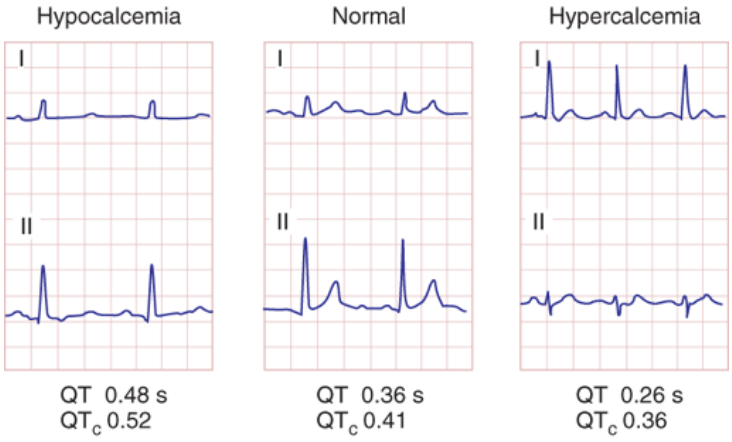


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 235-16

Prolongation of the Q-T interval (ST-segment portion) is typical of hypocalcemia. Hypercalcemia may cause abbreviation of the ST segment and shortening of the QT interval.

Loading [Contrib]/a11y/accessibility-menu.js



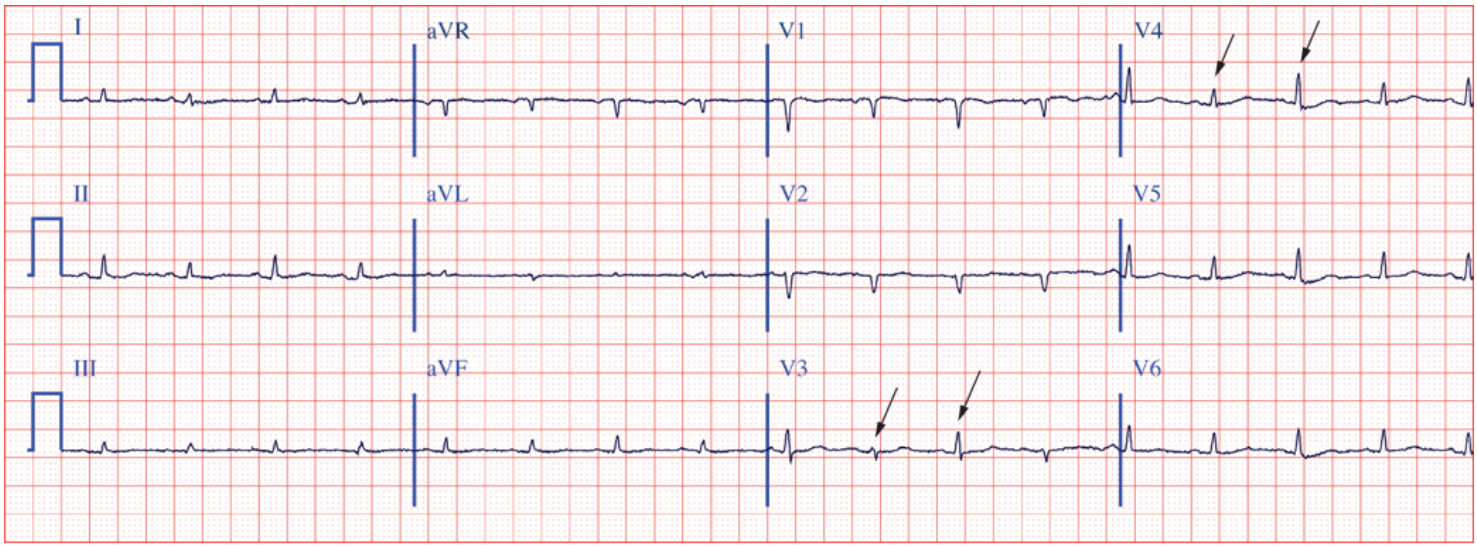
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

NON-SPECIFIC ST-T CHANGES AND LOW QRS VOLTAGE

Many other factors are associated with ECG changes, particularly alterations in ventricular repolarization. T-wave flattening, minimal T-wave inversions, or slight ST-segment depression ("nonspecific ST–T-wave changes") may occur with a variety of electrolyte and acid-base disturbances, infectious or inflammatory processes, central nervous system disorders, endocrine abnormalities, many drugs, ischemia, hypoxia, and virtually any type of cardiopulmonary abnormality, in addition to physiologic changes (e.g., with posture or with meals). Low QRS voltage is arbitrarily defined as peak-to-trough QRS amplitudes of ≤5 mm in the six limb leads and/or ≤10 mm in the chest leads. Multiple factors may be responsible. Among the most serious include pericardial **(Fig. 235-17)** or pleural effusions, chronic obstructive pulmonary disease, infiltrative cardiomyopathies, and anasarca.

FIGURE 235-17

Classic triad of findings for pericardial effusion with cardiac tamponade: (1) sinus tachycardia; (2) low QRS voltages; and (3) electrical alternans (best seen in leads V_3 and V_4 in this case; *arrows*). This triad is highly specific for pericardial effusion, usually with tamponade physiology, but of limited sensitivity. *(Adapted from LA Nathanson et al: ECG Wave-Maven. http://ecg.bidmc.harvard.edu.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ELECTRICAL ALTERNANS

Electrical alternans—a beat-to-beat alternation in one or more components of the ECG signal—is a common type of nonlinear cardiovascular response to a variety of hemodynamic and electrophysiologic perturbations. Total electrical alternans (P-QRS-T) with sinus tachycardia is a relatively specific sign of pericardial effusion, usually with cardiac tamponade **(Fig. 235-17)**. In contrast, pure repolarization (ST-T or U wave) alternans is a sign of electrical instability and may precede ventricular tachyarrhythmias.

CLINICAL INTEPRERATION OF THE ECG

Accurate analysis of ECGs requires thoroughness and care. The patient's age, gender, and clinical status should always be taken into account. Many mistakes in ECG interpretation are errors of omission. Therefore, a systematic approach is essential. The following 14 points should be analyzed carefully in every ECG: (1) standardization (calibration) and technical features (including lead placement and artifacts), (2) rhythm, (3)

Loading [Contrib]/a11y/accessibility-menu.js

heart rate, (4) PR interval/AV conduction, (5) QRS interval, (6) QT/QT_c intervals, (7) mean QRS electrical axis, (8) P waves, (9) QRS voltages, (10) precordial R-wave progression, (11) abnormal Q waves, (12) ST segments, (13) T waves, and (14) U waves. Comparison with any previous ECGs is invaluable.

COMPUTERIZED ELECTROCARDIOGRAPHY

Computerized systems are widely used for immediate retrieval of thousands of ECG records. Computer analysis of ECGs still has major limitations and, therefore, should not be accepted without careful clinician review.

FURTHER READING

GOLDBERGER AL, GOLDBERGER ZD, SHVILKIN A: *Goldberger's Clinical Electrocardiography: Simplified Approach*, 9th ed. Philadelphia, Elsevier, 2017.
KLIGFIELD P et al: Recommendations for the Standardization and Interpretation of the Electrocardiogram: Part I: The Electrocardiogram and Its Technology A Scientific Statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society Endorsed by the International Society for Computerized Electrocardiology. J Am Coll Cardiol 49:1109, 2007. [PubMed: 17349896]
MIRVIS DM, GOLDBERGER AL: Electrocardiography, in *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*, 11th ed, JL Jameson et al (eds). Philadelphia, Elsevier, 2017.
NATHANSON LA et al: ECG Wave-Maven. Self-Assessment Program for Students and Physicians. *https://ecg.bidmc.harvard.edu/maven/mavenmain.asp*. Last accessed May 8, 2017.
RAUTAHARJU PM et al: Recommendations for the standardization and interpretation of the electrocardiogram. Part IV: The ST segment, T and U waves. J Am Coll Cardiol 53:982, 2009. [PubMed: 19281931]
SURAWICZ B, KNILANS T: *Chou's Electrocardiography in Clinical Practice: Adult and Pediatric*, 6th ed. Philadelphia, Saunders, 2008.
SURAWICZ B et al: Recommendations for the standardization and interpretation of the electrocardiogram. Part III: Intraventricular conduction disturbances. Circulation 119:e235, 2009. [PubMed: 19228822]
WAGNER GS et al. AHA/ACCF/HRS Recommendations for the Standardization and Interpretation of the Electrocardiogram: Part VI: Acute Ischemia/Infarction: A Scientific Statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society Endorsed by the International Society for Computerized Electrocardiology. J Am Coll Cardiol 53:1003, 2009. [PubMed: 19281933]
ZIMETBAUM PJ, JOSEPHSON ME: Use of the electrocardiogram in acute myocardial infarction. N Engl J Med 348:933, 2003. [PubMed: 12621138]

Loading [Contrib]/a11y/accessibility-menu.js

Chapter 236: Noninvasive Cardiac Imaging: Echocardiography, Nuclear Cardiology, and Magnetic Resonance/Computed Tomography Imaging

Marcelo F. Di Carli; Raymond Y. Kwong; Scott D. Solomon

INTRODUCTION

The ability to image the heart and blood vessels noninvasively has been one of the greatest advances in cardiovascular medicine since the development of the electrocardiogram (ECG). Cardiac imaging complements history taking and physical examination, blood and laboratory testing, and exercise testing in the diagnosis and management of most diseases of the cardiovascular system. Modern cardiovascular imaging consists of echocardiography (cardiac ultrasound), nuclear scintigraphy including positron emission tomography (PET) imaging, magnetic resonance imaging (MRI), and computed tomography (CT). These studies, often used in conjunction with exercise testing, can be used independently or in concert depending on the specific diagnostic needs. In this chapter, we review the principles of each of these modalities and the utility and relative benefits of each for the most common cardiovascular diseases.

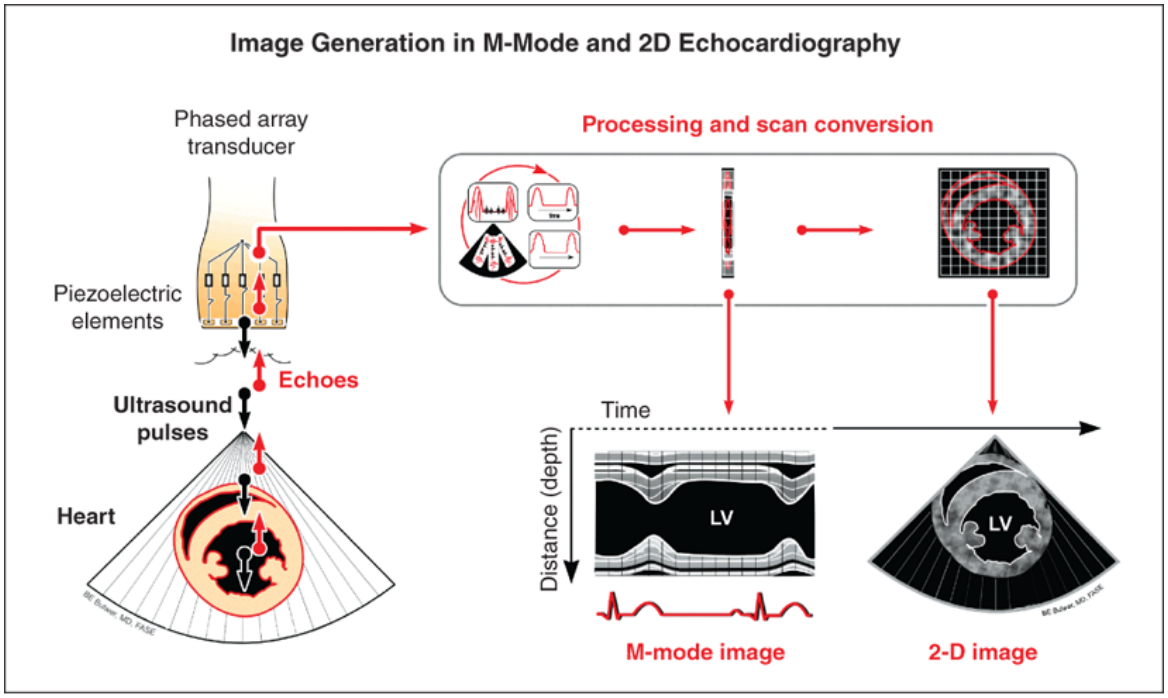
PRINCIPLES OF MULTIMODALITY CARDIAC IMAGING

ECHOCARDIOGRAPHY

Echocardiography uses high-frequency sound waves (ultrasound) to penetrate the body, reflect from relevant structures, and generate an image. The basic physical principles of echocardiography are identical to other types of ultrasound imaging, although the hardware and software are optimized for evaluation of cardiac structure and function. Early echocardiography machines displayed “M-mode” echocardiograms in which a single ultrasound beam was displayed over time on a moving sheet of paper **(Fig. 236-1)**, ***left panel***. Modern echocardiographic machinery uses phased array transducers that contain up to 512 elements and emit ultrasound in sequence. The reflected ultrasound is then sensed by the receiving elements. A “scan converter” uses information about the timing and magnitude of the reflected ultrasound to generate an image **(Fig. 236-1**, ***right panel***). This sequence happens repeatedly in “real time” to generate moving images with frame rates that are typically greater than 30 frames per second, but can exceed 100 frames per second. The gray scale of the image features indicates the intensity of the reflected ultrasound; fluid or blood appears black, and highly reflective structures, such as calcifications on cardiac valves or the pericardium, appear white. Tissues such as myocardium appear more gray, and tissues such as muscle display a unique speckle pattern. Although M-mode echocardiography has largely been supplanted by two-dimensional (2D) echocardiography, it is still used because of its high temporal resolution and accuracy for making linear measurements.

FIGURE 236-1

Principle of image generation in two-dimensional (2D) echocardiography. An electronically steerable phased-array transducer emits ultrasound from piezoelectric elements, and returning echoes are used to generate a 2D image (*right*) using a scan converter. Early echocardiography machines used a single ultrasound beam to generate an “M-mode” echocardiogram (see text), although modern equipment generates M-mode echocardiograms digitally from the 2D data. LV, left ventricle.



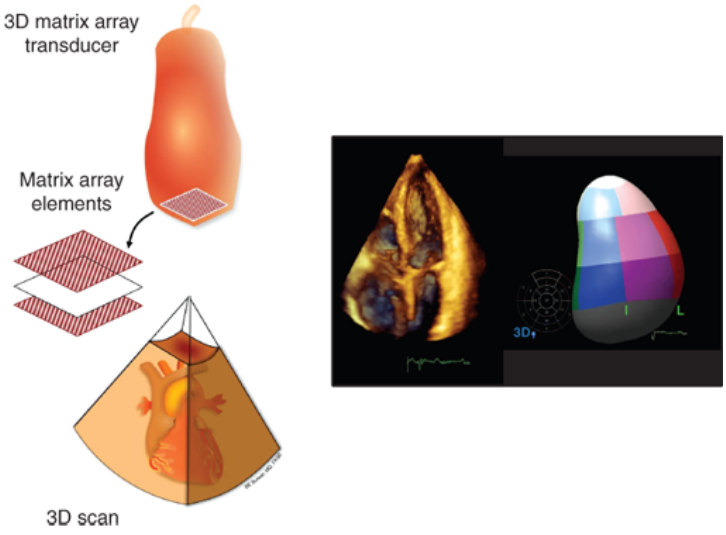
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The spatial resolution of ultrasound is dependent on the wavelength: the smaller the wavelength and the higher the frequency of the ultrasound beam, the greater are the spatial resolution and ability to discern small structures. Increasing the frequency of ultrasound will increase resolution but at the expense of reduced penetration. Higher frequencies can be used in pediatric imaging or transesophageal echocardiography where the transducer can be much closer to the structures being interrogated, and this is a rationale for using transesophageal echocardiography to obtain higher quality images.

Three-dimensional ultrasound transducers use a waffle-like matrix array transducer and receive a pyramidal data sector. Three-dimensional echocardiography is being increasingly used for assessment of congenital heart disease and valves, although current image quality lags behind 2D ultrasound **(Fig. 236-2)**.

FIGURE 236-2

Three-dimensional (3D) probe and 3D image.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In addition to the generation of 2D images that provide information about cardiac structure and function, echocardiography can be used to interrogate blood flow within the heart and blood vessels by using the Doppler principle to ascertain the velocity of blood flow. When ultrasound emitted from a transducer reflects off red blood cells that are moving toward the transducer, the reflected ultrasound will return at a slightly higher frequency than emitted; the opposite is true when flow moves away from the transducer. That frequency difference, termed the Doppler shift, is directly related to the velocity of the flow of the red blood cells. The velocity of blood flow between two chambers will be directly related to the pressure gradient between those chambers. A modified form of the Bernoulli equation,

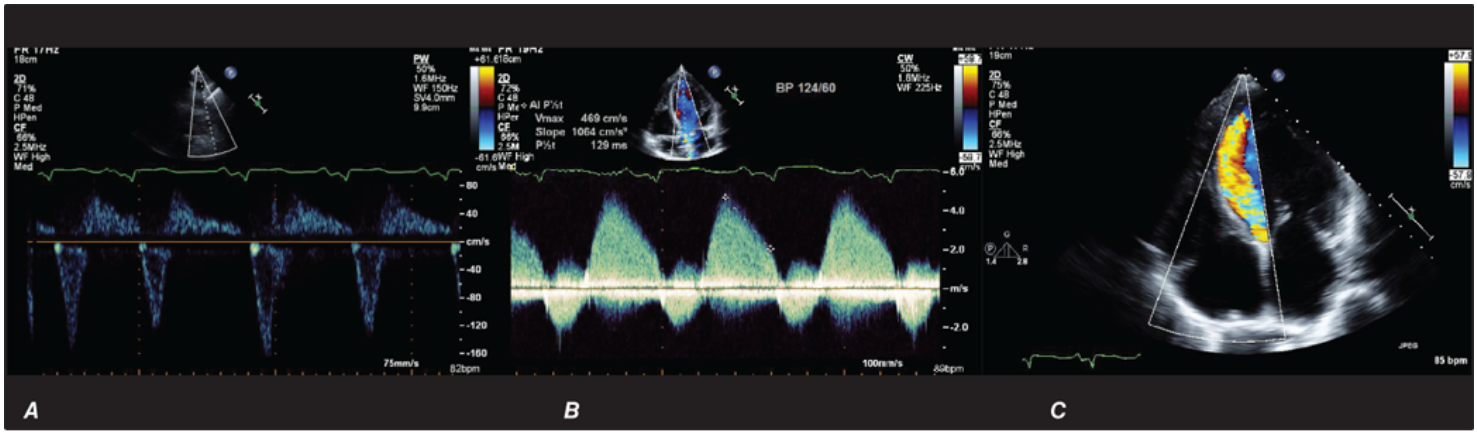
$p = 4v^2$

where p = the pressure gradient and v = the velocity of blood flow in meters per second, can be used to calculate this pressure gradient in the majority of clinical circumstances. This principle can be used to determine the pressure gradient between chambers and across valves and has become central to the quantitative assessment of valvular heart disease.

There are three types of Doppler ultrasound that are typically used in standard echocardiographic examinations: spectral Doppler, which consists of both pulsed wave Doppler and continuous wave Doppler, and color flow Doppler. Both types of spectral Doppler will display a waveform representing the velocity of blood flow, with time on the horizontal axis and velocity on the vertical axis. Pulsed wave Doppler is used to interrogate relatively low velocity flow and has the ability to determine blood flow velocity at a particular location within the heart. Continuous wave Doppler is used to assess high-velocity flow, but can only identify the highest velocity in a particular direction and cannot interrogate the velocity at a specific depth location. Both of these techniques can only accurately assess velocities that are in the direction of the ultrasound scan lines, and velocities that are at an angle to the direction of the ultrasound beam will be underestimated. Color flow Doppler is a form of pulsed wave Doppler in which the velocity of blood flow is color encoded according to a scale and superimposed on a 2D grayscale image in real time, giving the appearance of real-time flow within the heart. The Doppler principle can also be used to assess the velocity of myocardial motion, which is a sensitive way to assess myocardial function **(Fig. 236-3)**. A standard full transthoracic echocardiographic examination consists of a series of 2D views made up of different imaging planes from various scanning locations and spectral and color flow Doppler assessment.

FIGURE 236-3

Three types of Doppler ultrasound. ***A*** and ***B.*** Pulsed and continuous wave Doppler waveforms with time on horizontal axis and velocity of blood flow on vertical axis. ***C.*** Color flow Doppler, where velocities are encoded by colors according to scale on right side of screen and superimposed on a two-dimensional grayscale image.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Transesophageal echocardiography is a form of echocardiography in which the transducer is located on the tip of an endoscope that can be inserted into the esophagus. This procedure allows closer, less obstructed views of cardiac structures, without having to penetrate through chest wall, muscle, and ribs. Because less penetration is needed, a higher frequency probe can be used, and image quality and spatial resolution are generally higher than with standard transthoracic imaging, particularly for structures that are more posterior. Transesophageal echocardiography has become the test of choice for assessment of small lesions in the heart such as valvular vegetations, especially in the setting of a prosthetic valve disease, and intracardiac thrombi, including assessment of the left atrial appendage, which is difficult to visualize with standard transthoracic imaging, and for assessment of congenital abnormalities. Transesophageal echocardiography requires both topical and systemic anesthesia, generally conscious sedation, and carries additional risks such as potential damage to the esophagus, including the rare possibility of perforation, aspiration, and anesthesia-related complications. Patients generally need to give consent for transesophageal echocardiography and be monitored during and subsequent to the procedure. Transesophageal echocardiography can be carried out in intubated patients and is routinely used for intraoperative monitoring during cardiac surgery.

Stress echocardiography is routinely used to assess cardiac function during exercise and can be used to identify myocardial ischemia or to assess valvular function under exercise conditions. Stress echocardiography is typically performed in conjunction with treadmill or bicycle exercise testing, but can also be performed using pharmacologic stress most typically with an intravenous infusion of dobutamine (see section on stress imaging below).

Whereas typical echocardiographic equipment is large, bulky, and expensive, small hand-held ultrasound equipment developed over the last decade now offers diagnostic quality imaging in a package small enough to be carried on rounds **(Fig. 236-4)**. These relatively inexpensive point-of-care devices currently lack full diagnostic capabilities but represent an excellent screening tool if used by an experienced operator. As these units become even smaller and less expensive, they are being increasingly used not just by cardiologists, but also by emergency medicine physicians, intensivists, anesthesiologists, and internists.

FIGURE 236-4

Two examples of hand-held ultrasound equipment: V-Scan (General Electric, *left*) and Sonosite (*right*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

RADIONUCLIDE IMAGING

Radionuclide imaging techniques are commonly used for the evaluation of patients with known or suspected coronary artery disease (CAD), including for initial diagnosis and risk stratification as well as the assessment of myocardial viability. These techniques use small amounts of radiopharmaceuticals **(Table 236-1)**, which are injected intravenously and trapped in the heart and/or vascular cells. Radioactivity within the heart and vasculature decays by emitting gamma rays. The interaction between these gamma rays and the detectors in specialized scanners (single-photon emission computed tomography [SPECT] and PET) creates a scintillation event or light output, which can be captured by digital recording equipment to form an image of the heart and vasculature. Like CT and MRI, radionuclide images also generate tomographic (three-dimensional) views of the heart and vasculature.

TABLE 236-1 Radiopharmaceuticals for Clinical Nuclear Cardiology

Radiopharmaceutical	Imaging Technique	Physical Half-Life	Application
Technetium-99m sestamibi	SPECT	6 h	Myocardial perfusion imaging
Technetium-99m tetrofosmin	SPECT	6 h	Myocardial perfusion imaging
Thalium-201	SPECT	72 h	Myocardial perfusion imaging
Iodine-123 metaiodobenzylguanidine (MIBG)	SPECT	13 h	Cardiac sympathetic innervation
Rubidium-82	PET	76 s	Myocardial perfusion imaging
^{13}N-ammonia	PET	10 min	Myocardial perfusion imaging
^{18}F-fluorodeoxyglucose	PET	110 min	Myocardial viability and inflammation imaging

Abbreviations: PET, positron emission tomography; SPECT, single-photon emission computed tomography.

Radiopharmaceuticals Used in Clinical Imaging

Table 236-1 summarizes the most commonly used radiopharmaceuticals in clinical SPECT and PET imaging.

Protocols for Stress Myocardial Perfusion Imaging

Both exercise and pharmacologic stress can be used for myocardial perfusion imaging. Exercise stress is generally preferred because it is physiologic and provides additional clinically important information (i.e., clinical and hemodynamic responses, ST-segment changes, exercise duration, and functional status). However, submaximal effort will lower the sensitivity of the test and should be avoided, especially if the test is requested for initial diagnosis of CAD. In patients who are unable to exercise or who exercise submaximally, pharmacologic stress offers an adequate alternative to exercise stress testing. Pharmacologic stress can be accomplished either with coronary vasodilators, such as adenosine, dipyridamole, or regadenoson, or β_1-receptor agonists, such as dobutamine. For patients unable to exercise, vasodilators are the most commonly used stressors in combination with myocardial perfusion imaging. Dobutamine is a potent β_1-receptor agonist that increases myocardial oxygen demand by augmenting contractility, heart rate, and blood pressure similar to exercise. It is generally used as an alternative to vasodilator stress in patients with chronic pulmonary disease, in whom vasodilators may be contraindicated. Dobutamine is also commonly used as a pharmacologic alternative to stress testing in stress echocardiography.

Myocardial Perfusion and Viability Imaging Protocols

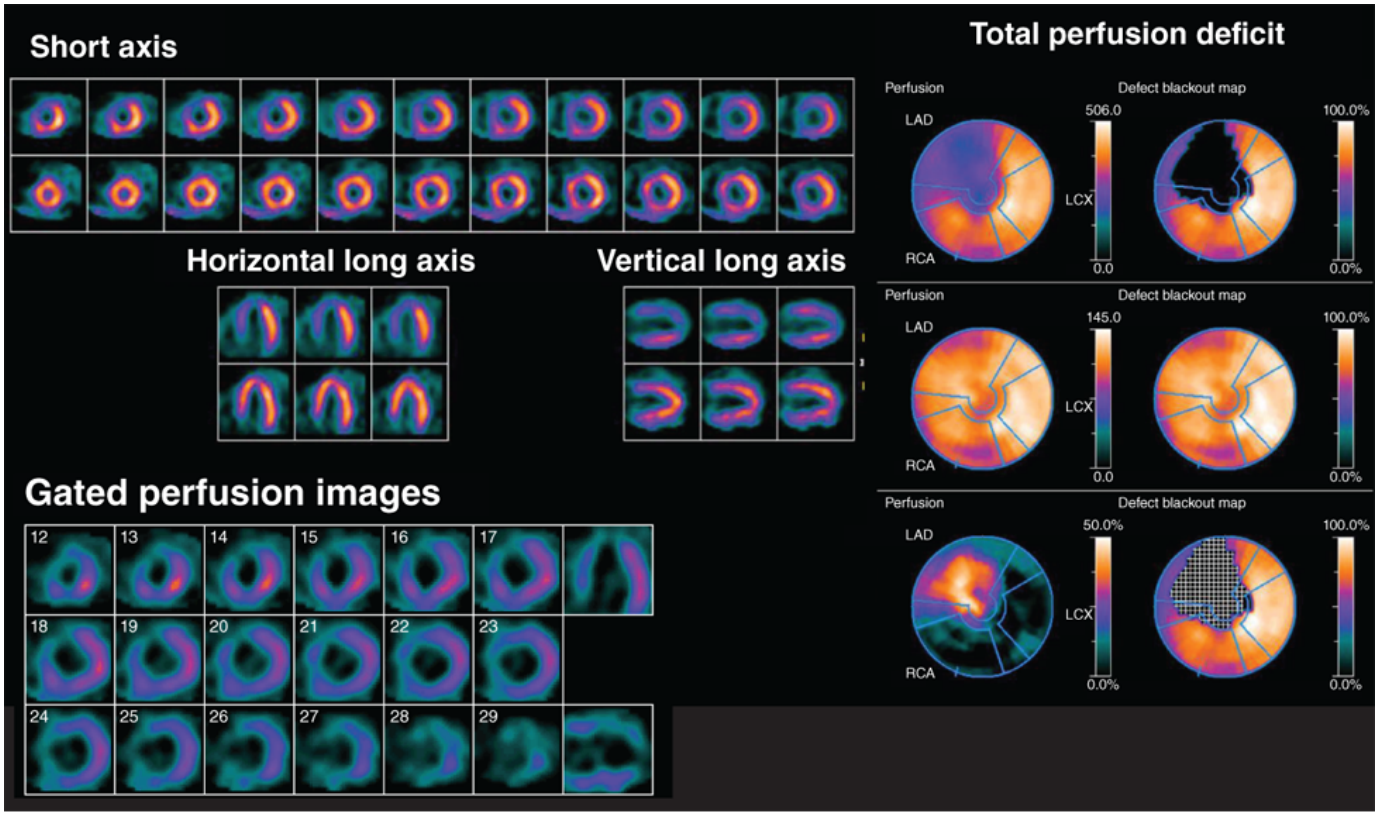
Imaging protocols are tailored to the individual patient based on the clinical question, patient's risk, ability to exercise, body mass index, and other factors.

For SPECT imaging, technetium-99m (^{99m}Tc)-labeled tracers are the most commonly used imaging agents because they are associated with the best image quality and the lowest radiation dose to the patient **(Fig. 236-5)**. Selection of the protocol (stress-only, single-day, or 2-day) depends on the patient and clinical question. After intravenous injection, myocardial uptake of ^{99m}Tc-labeled tracers is rapid (1–2 min). After uptake, these tracers become trapped intracellularly in mitochondria and show minimal change over time. This is why ^{99m}Tc tracers can be helpful in patients with chest pain of unclear etiology occurring at rest, because patients can be injected while having chest pain and imaged some time later after symptoms subside. Indeed, a normal myocardial perfusion study following a rest injection in a patient with active chest pain effectively excludes

myocardial ischemia as the cause of chest pain (high negative predictive value). While used commonly in the past for perfusion imaging, thallium-201 protocols are now rarely used because they are associated with a higher radiation dose to the patient.

FIGURE 236-5

Tomographic stress (top of each pair) and rest myocardial perfusion images with technetium-99m sestamibi single-photon emission computed tomography imaging demonstrating a large perfusion defect throughout the anterior and anteroseptal walls. The *right panel* demonstrates the quantitative extent of the perfusion abnormality at stress (*top bull's-eye*), at rest (*middle bull's-eye*), and the extent of defect reversibility (*lower bull's-eye*). The *lower left panel* demonstrates electrocardiogram-gated myocardial perfusion images from which one can determine the presence of regional wall motion abnormalities and calculate left ventricular volumes and ejection fraction.

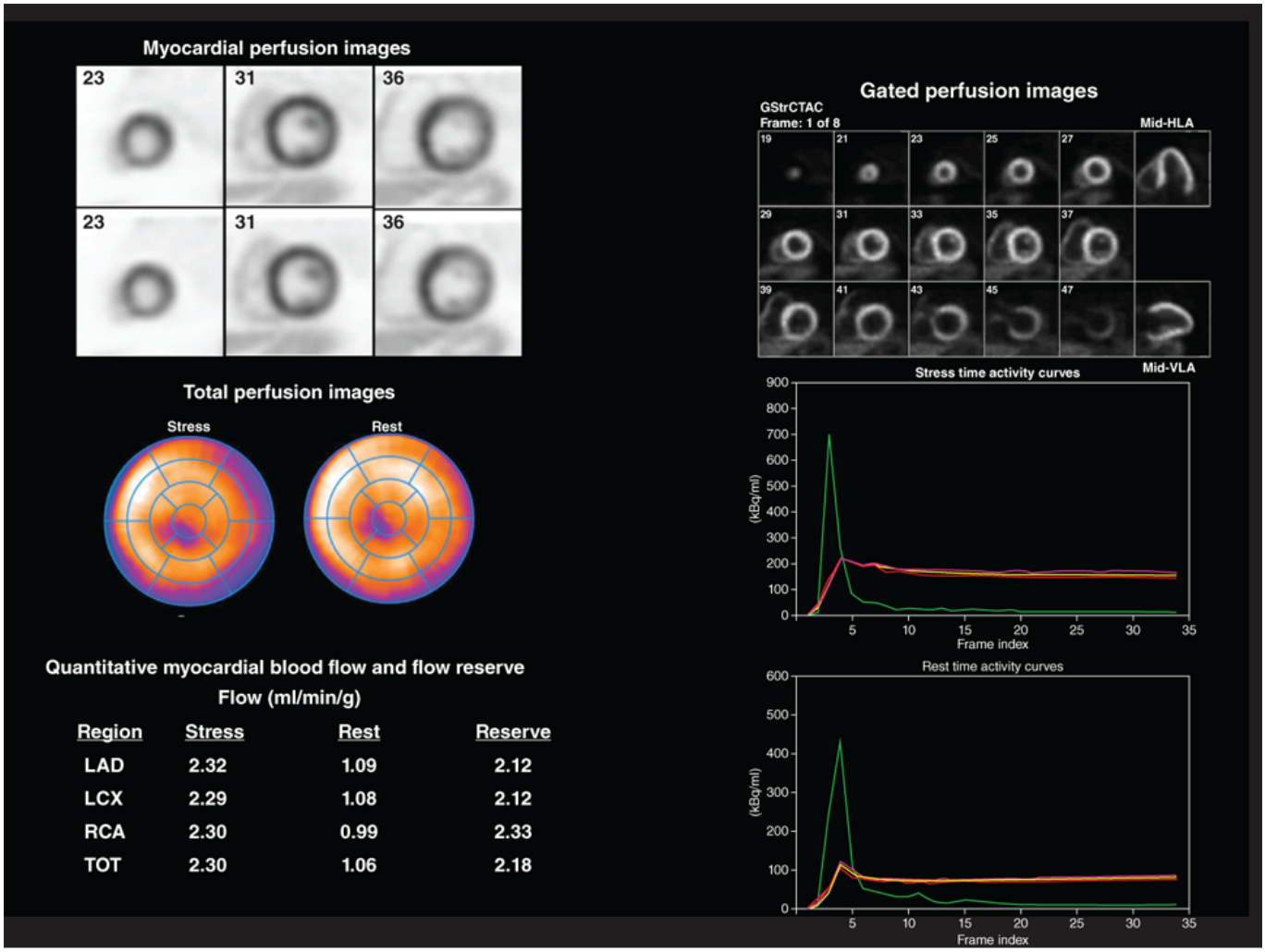


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PET myocardial perfusion imaging is an alternative to SPECT and is associated with improved diagnostic accuracy and lower radiation dose to patients (Table 236-1). The ultra-short half-life of some PET radiopharmaceuticals in clinical use (e.g., rubidium-82) is the primary reason why imaging is generally combined with pharmacologic stress, as opposed to exercise. However, exercise is possible for relatively longer lived radiotracers (e.g., ^{13}N-ammonia). PET imaging protocols are typically faster than SPECT, but more expensive. In comparison to SPECT, PET has improved spatial and contrast resolution and provides absolute measures of myocardial perfusion (in mL/min per gram of tissue), thereby providing the patients' regional and global coronary flow reserve. The latter helps improve diagnostic accuracy and risk stratification, especially in obese patients, women, and higher risk individuals (e.g., diabetes mellitus) **(Fig. 236-6)**. Contemporary PET and SPECT scanners are combined with a CT scanner (so-called *hybrid PET/CT* and *SPECT/CT*). CT is used primarily to guide patient positioning in the field of view and for correcting inhomogeneities in radiotracer distribution due to attenuation by soft tissues (so-called *attenuation correction*). However, it can also be used to obtain diagnostic data including coronary artery calcium (CAC) score and/or CT coronary angiography (discussed below).

FIGURE 236-6

Multidimensional cardiac imaging protocol with positron emission tomography. The *left upper panel* demonstrates stress and rest short-axis images of the left and right ventricles demonstrating normal regional myocardial perfusion. The *middle panel* demonstrates the quantitative bull's-eye display to evaluate the extent and severity of perfusion defects. The *lower right panel* illustrates the time-activity curves for quantification of myocardial blood flow. The *right upper panel* demonstrates electrocardiogram-gated myocardial perfusion images from which one can determine the presence of regional wall motion abnormalities and calculate left ventricular volumes and ejection fraction. LAD, left anterior descending artery; LCX, left circumflex artery; RCA, right coronary artery; TOT, total left ventricle.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

For the evaluation of myocardial viability in patients with ischemic cardiomyopathy, myocardial perfusion imaging (with SPECT or PET) is usually combined with metabolic imaging (i.e., fluorodeoxyglucose [FDG] PET). In hospital settings lacking access to PET scanning, thallium-201 SPECT imaging is an excellent alternative.

CARDIAC COMPUTED TOMOGRAPHY

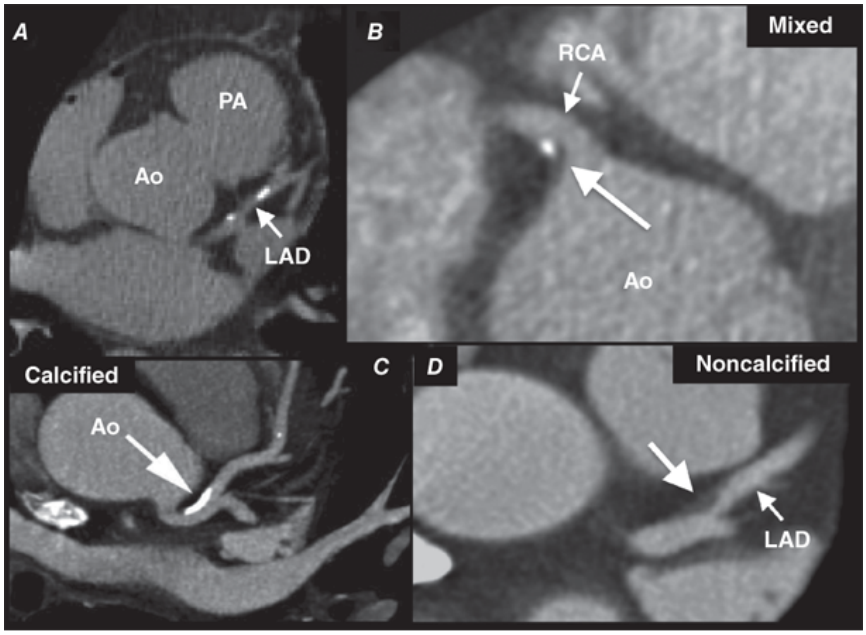
CT acquires images by passing a thin x-ray beam through the body at many angles to generate cross-sectional images. The x-ray transmission measurements are collected by a detector array and digitized into pixels that form an image. The grayscale information in individual pixels is determined by the attenuation of the x-ray beam along its path by tissues of different densities, referenced to the value for water in units known as Hounsfield units. In the resulting CT images, bone appears bright white, air is black, and blood and muscle show varying shades of gray. However, due to the limited contrast between cardiac chambers and vascular structures, iodinated contrast agents are necessary for most cardiovascular indications. Cardiac CT produces tomographic images of the heart and surrounding structures. With modern CT scanners, a three-dimensional dataset of the heart can be acquired in 5–15 s with submillimeter spatial resolution.

CT Calcium Scoring

CT calcium scoring is the simplest application of cardiac CT and does not require administration of iodinated contrast. The presence of coronary artery calcification has been associated with increased burden of atherosclerosis and cardiovascular mortality. Coronary calcium is then quantified (e.g., Agatston score) and categorized as minimal (0–10), mild (10–100), moderate (100–400), or severe (>400) **(Fig. 236-7)**. CAC scores are then normalized by age and gender and reported as percentile scores. Population-based studies in asymptomatic cohorts have reported high cardiac prognostic value of CT calcium score. With appropriate techniques, the radiation dose associated with CAC scanning is very low (~1–2 mSv).

FIGURE 236-7

Examples of non-contrast- and contrast-enhanced coronary imaging with computed tomography (CT). ***A.*** Calcified coronary plaques in the distal left main and proximal left anterior descending coronary artery (LAD) in a noncontrast cardiac CT scan. Calcium deposits are dense and present as bright white structures on CT, even without contrast enhancement. ***B, C,*** and ***D.*** Different types of atherosclerotic plaques on contrast-enhanced CT scans. Importantly, noncalcified plaques are evident only on contrast-enhanced CT scans. AO, aorta; PA, pulmonary artery; RCA, right coronary artery.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CT Coronary Angiography

Coronary CT angiography (CTA) is emerging as a viable alternative to invasive coronary angiography in selected patients. Imaging of the coronary arteries by CT is challenging because of their small luminal size and because of cardiac and respiratory motion. Respiratory motion can be reduced by breath-holding, and cardiac motion is best reduced by slowing the patient's heart rate, ideally to under 60 beats/min, using intravenous or oral beta blockade or other rate-lowering drugs. When performing a coronary CTA, image quality is further enhanced using sublingual nitroglycerin to enlarge the coronary lumen just prior to contrast injection. Imaging the whole-heart volume is synchronized to the administration of weight-based and appropriately timed intravenous iodinated contrast. Image acquisition is linked to the timing of the cardiac cycle through ECG triggering. The resulting images are then postprocessed using a three-dimensional workstation, which facilitates interpretation of the coronary anatomy and estimation of the severity of atherosclerosis (Fig. 236-7).

CARDIAC MAGNETIC RESONANCE

Cardiac magnetic resonance (CMR) imaging is based on imaging of protons in hydrogen, which is an advantage, given the abundance of water in the human body. When the body is placed inside a MRI scanner, protons in different tissues, such as in simple fluid or complex macromolecules such as fat or protein, interact with the magnetic field at their unique frequencies. A set of orthogonal gradient coils in the scanner is designed to locate protons spatially so that radiofrequency (RF) pulses of energy can be delivered to select imaging planes of interests. Once the RF pulses stop, the energy absorbed will be released, collected by the phased-array receiver coils placed on the patient's body surface, digitally recorded in a data matrix known as the K-space, then reconstructed into a magnetic resonance image. The large arrays of software methods of delivering RF pulses are known as pulse sequences which aim at extraction of different types of cardiac structural or physiologic information. In CMR, T1-weighted pulse sequences are most common and they assess cardiac structure and function, blood flow, and myocardial perfusion. T2-weighted and $T2^*$-weighted pulse sequences, on the other hand, evaluate myocardial edema and myocardial iron infiltration, respectively. Vector ECG-gating and repetitive patient breath-holding are used by convention to suppress cardiac and respiratory motions, respectively. However, with technical advent, rapid data collection algorithm and diaphragmatic position gating have eliminated the need for ECG-gating and breath-holding in challenging situations. A list of common pulse sequence used in CMR is shown in **Table 236-2**.

TABLE 236-2 Clinical Cardiac Magnetic Resonance Pulse Sequences and Their Application

Pulse Sequence	Key Imaging Interests
Cardiac Morphology	
Still frame imaging (black or bright blood)	Cardiac structures
Cardiac Function	
Cine imaging	Left ventricular volume and function
Cine myocardial tagging	Left ventricular deformation (strain)
Blood Flow Imaging	
Velocity-encoded phase contrast	Cardiac and great vessel flow
Stress Testing	
Myocardial perfusion imaging	Regional myocardial blood flow
Cine imaging	Regional wall motion
Myocardial Tissue Characterization	
Late gadolinium enhancement	Myocardial infarction and infiltrative disease
T2-weighted imaging	Myocardial edema
Iron content imaging	Myocardial iron infiltration
Magnetic Resonance Angiography	
Aorta, peripheral and coronary arteries	Luminal stenosis and vessel wall remodeling

ASSESSMENT OF CARDIAC STRUCTURE AND FUNCTION

Echocardiography, CMR, and cardiac CT are all capable of assessing cardiac structure and function, although echocardiography is generally considered the primary imaging method for these assessments. Radionuclide imaging can also be used to assess left ventricular regional and global systolic function. Echocardiography is most often used to assess the size of all four chambers and thickness of ventricular walls, which are affected by both cardiac and systemic diseases.

The structure of the left ventricle is generally assessed by determining its volume and mass. Left ventricular volumes can be easily estimated from 2D echocardiography by using methods incorporating geometric assumptions. The accuracy of these echocardiographic methods is reduced when foreshortening of the imaging plane leads to underestimation of volumes. Moreover, these methods require accurate delineation of the endocardial border. In this regard, high-resolution tomographic techniques such as CMR or cardiac CT are more accurate for volumetric assessment. Three-dimensional (3D) echocardiography does not require any geometric assumptions about the left ventricle for quantification of volumes and ejection fraction. However, 3D echocardiographic imaging requires substantial expertise and currently is not widely used in practice.

Left ventricular dilatation is common to a number of cardiac diseases. For example, regional dysfunction secondary to myocardial infarction can ultimately lead to progressive ventricular dilatation or remodeling. Although dilatation often begins in the region affected by the infarction, subsequent compensatory dilatation can occur in remote myocardial regions as well. The presence of regional wall motion abnormalities associated with ventricular thinning (reflecting scar) in a coronary distribution is strongly suggestive of an ischemic etiology. Direct assessment of infarcted myocardium is possible with both CMR (evident as areas of late gadolinium enhancement [LGE]) and radionuclide imaging (as assessed by regional perfusion or metabolic defects at rest). CMR can be particularly useful in determining etiology of cardiomegaly and ventricular dysfunction, with LGE in coronary distributions being nearly pathognomonic for infarction **(Video 236-1)**.

VIDEO 236-1: Cine steady-state free precession (SSFP) imaging (left) in short axis in a patient who had a large anterior myocardial infarction.

Only one cut of a stack of short axis is shown. This method allows quantification of left ventricular (LV) and right ventricular (RV) volumes in diastole and systole and calculation of the LV ejection fraction, stroke volumes, and cardiac output (a product of LV stroke volume and heart rate). Note that in this case there is anterior and anteroseptal akinesia (lack of systolic wall thickening, as shown by the left cine movie, *red arrows*) matching by a near-transmural myocardial infarction as seen by the matching late gadolinium enhancement (LGE) image (*right picture, white arrows*).

Image not available.
Play Video

More global ventricular dilatation is seen in cardiomyopathy and dilatation due to valvular heart disease. Idiopathic, nonischemic cardiomyopathies will typically result in global ventricular dilatation and dysfunction, with thinning of the walls. Patients with substantial ventricular dyssynchrony due to conduction abnormalities will have a typical pattern of contraction (e.g., delay of contraction of the lateral wall with left bundle branch block). As discussed later in this chapter, regurgitant lesions of either the mitral or aortic valves can lead to substantial ventricular dilatation, and assessment of ventricular size is integral in the evaluation and timing of surgical correction. Because changes in ventricular size are used clinically to determine which patients should undergo valve surgery, accurate assessment of changes in ventricular size is essential. Although serial echocardiography can provide these data, serial assessment by CMR may be more accurate when appreciation of subtle changes over time is important.

Left ventricular wall thickness and mass are also important measures of cardiac and systemic disease. The left ventricle will hypertrophy under any condition in which its afterload is increased, including conditions that obstruct outflow, such as aortic stenosis, hypertrophic cardiomyopathy, and subaortic membranes; in postcardiac aortic obstruction seen in coarctation; or in systemic conditions characterized by increased afterload, such as hypertension. The pattern of ventricular hypertrophy can change depending on the etiology. Aortic stenosis and hypertension are typically characterized by concentric hypertrophy, in which the ventricular walls thicken “concentrically” and cavity size is usually small. In volume overload conditions such as mitral or aortic regurgitation, there may be minimal increase in ventricular wall thickness, but substantial ventricular dilatation leads to marked increases in left ventricular mass.

Ventricular wall thickness can be measured and ventricular mass can be calculated by either echocardiography or CMR. Although radionuclide imaging and cardiac CT can also provide measures of left ventricular mass, they are not generally used for this purpose. Although measurement of wall thickness with echocardiography is relatively straightforward and accurate, determining left ventricular mass by echocardiography requires

using one of several formulas that takes into account both wall thickness and ventricular cavity dimensions. Assessment of left ventricular mass by CMR has the advantage of not requiring geometric assumptions and is thus more accurate than echocardiography.

ASSESSMENT OF LEFT VENTRICULAR SYSTOLIC FUNCTION

Assessment of ejection fraction, or the percentage of blood ejected with each beat, has been the primary method to assess systolic function and is generally calculated by subtracting end-systolic volume from end-diastolic volume and dividing by end-diastolic volume. All cardiac imaging modalities can provide direct measurements of left ventricular ejection fraction (LVEF). As discussed above, tomographic techniques (e.g., CMR, CT, and radionuclide imaging) are generally more accurate and reproducible than echocardiography because there are no geometric assumptions. A LVEF of 55% or greater is generally considered normal, and an LVEF of 50–55% is considered in the low-normal range.

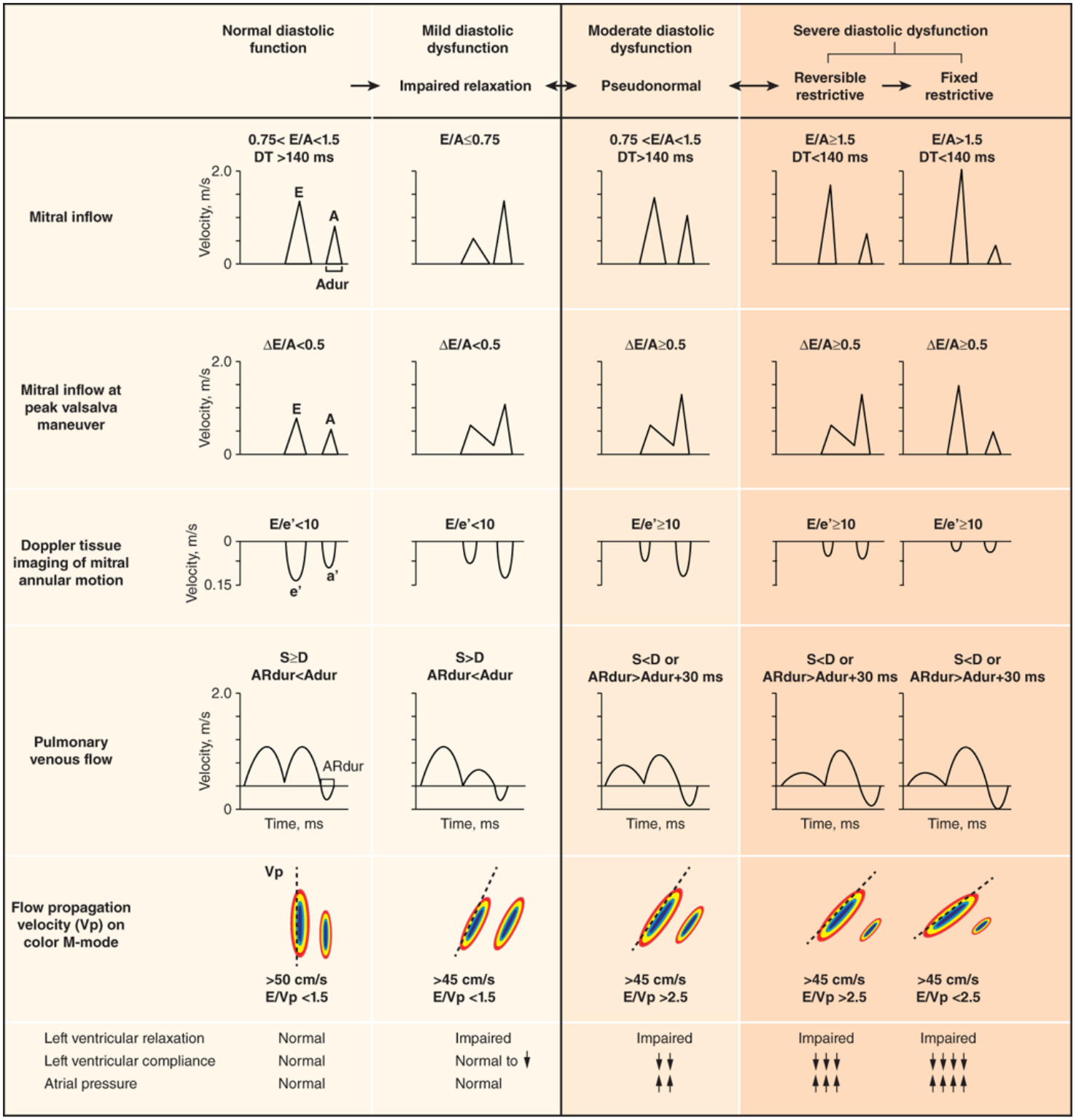
Newer methods to assess systolic function, such as myocardial strain or deformation imaging using speckle-tracking methods on echocardiography, or myocardial tagging, and more recently, feature tracking on CMR, can provide a more sensitive approach to detection of systolic dysfunction. Additional assessments based on these novel methods include assessment of myocardial twist and torsion. Although these techniques are not used routinely, they may be especially useful in certain conditions such as valvular heart disease and early detection of cardiotoxicity following chemotherapy and/or radiation therapy. In addition to estimation or calculation of ejection fraction, stroke volume can be assessed by any of the imaging methods, by subtracting the end-systolic volume from the end-diastolic volume, or by quantifying forward flows using echocardiographic Doppler methods or phase-contrast CMR imaging. They offer measures of systolic function other than LVEF.

ASSESSMENT OF LEFT VENTRICULAR DIASTOLIC FUNCTION

Echocardiography remains the primary method for clinical assessment of diastolic function. Recent advances in Doppler tissue imaging allow for accurate assessment of the velocity of myocardial wall motion by assessing the excursion of the mitral annulus in diastole. Mitral annular relaxation velocity, or E′, is inversely related to the time constant of relaxation, tau, and has been shown to have prognostic significance. Dividing the standard mitral inflow maximal velocity, E, by the mitral annular relaxation velocity yields E/E′, which has been shown to correlate with left ventricular filling pressures. The utility of standard E and A wave ratios for assessment of diastolic function has been questioned. Mitral deceleration time can be a useful measure if very short (<150 ms), suggesting restrictive physiology and severe diastolic dysfunction. Several grading methods for diastolic function have been proposed that take into account a number of diastolic parameters, including Doppler tissue-based relaxation velocities, pulmonary venous Doppler, and left atrial size **(Fig. 236-8)**. Diastolic function worsens with aging, and most diastolic parameters need to be adjusted for age.

FIGURE 236-8

Stages of diastolic function based on various parameters, including mitral inflow (with and without Valsalva maneuver), Doppler tissue imaging, pulmonary venous flow, and flow propagation. *(Adapted with permission from MM Redfield et al: JAMA 289:194, 2003.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ASSESSMENT OF RIGHT VENTRICULAR FUNCTION

Right ventricular size and function have been shown to be prognostically important in a variety of conditions, and can be assessed by echocardiography, CMR, CT, or radionuclide imaging methods. CMR is considered the most accurate noninvasive technique to evaluate the structure and ejection fraction of the right ventricle **(Video 236-2)**. Assessment of the right ventricle by echocardiography has generally been qualitative, owing in part to the unusual geometry of the right ventricle. However, several quantitative methods are available for assessment of right ventricular function, including fractional area change (FAC = [diastolic area – systolic area]/diastolic area), which has been shown to correlate with outcomes in heart failure and after myocardial infarction. Excursion of the tricuspid annulus (tricuspid annular plane systolic excursion, TAPSE) is another method to assess right ventricular function, although it is mostly used in research settings.

VIDEO 236-2: Cine cardiac magnetic resonance (CMR) imaging of a patient in the long-axis four-chamber view.

Note that the basal aspect of the right ventricular (RV) free wall is thickened, aneurysmal, and akinetic (*red arrows*). The global RV systolic function is mildly reduced, and the RV is dilated. CMR can image the RV using tomographic views and can quantify the RV volumes and ejection fraction volumetrically. This is a patient who presented with syncopal spells and inducible ventricular tachycardia on subsequent workup. He was diagnosed to have arrhythmogenic right ventricular dysplasia.


Play Video

Abnormalities of right ventricular size and function are generally secondary to either diseases that affect the right ventricle intrinsically or disease in which the right ventricle responds to abnormalities elsewhere in the heart or pulmonary vasculature. Intrinsic diseases that affect the RV include congenital abnormalities, including hypoplastic right ventricle and arrhythmogenic right ventricular dysplasia, and acquired conditions, such as right ventricular infarction and infiltrative diseases. Right ventricular dilatation can occur due to both chronic and acute processes. Long-standing pulmonary hypertension or pulmonary outflow tract obstruction leads to right ventricular hypertrophy and ultimately dilatation. An acute process that can cause profound right ventricular dilatation and dysfunction is acute pulmonary embolism. In the setting of acute occlusion of a pulmonary artery or branch, an acute rise in pulmonary vascular resistance causes a previously normal right ventricle to dilate and fail due to the increased afterload. In acute pulmonary embolism, right ventricular dilatation and dysfunction are signs of substantial hemodynamic compromise and are associated with a marked increased risk of death. In addition to right ventricular dilatation, acute pulmonary embolism is often associated with a specific pattern of regional right ventricular dysfunction, commonly referred to as the McConnell sign, characterized by preservation of right ventricular wall motion in the basal and apical regions and dyskinesis in the region of the mid right ventricular free wall. This abnormality is highly specific for acute pulmonary embolism and is likely secondary to acute increases in right ventricular load.

Any disease that causes increased pulmonary vascular resistance can lead to right ventricular dilatation and dysfunction. For example, long-standing chronic obstructive pulmonary disease increases pulmonary vascular resistance and results in cor pulmonale. Acute pneumonia can cause findings that are similar to acute pulmonary embolism. In patients with right ventricular dilatation without obvious pulmonary disease, intracardiac shunts should be considered. The increased flow through the pulmonary vasculature as a result of an atrial septal or ventricular septal defect can, over time, result in elevation in pulmonary vascular resistance with subsequent dilatation and hypertrophy of the right ventricle. Right ventricular dilatation and dysfunction also have prognostic significance in left-sided heart disease and have been shown to be important predictors of outcome in patients with heart failure or acute myocardial infarction.

In addition to assessment of left and right ventricular structure and function, assessment of the other cardiac chambers also provides important clues to intracardiac and systemic diseases. Enlargement of the left atrium is common in patients with hypertension and is also suggestive of increased left ventricular filling pressures; indeed, left atrial size is often termed the “hemoglobin A_{1c}” of diastolic function, because left atrial enlargement reflects long-standing increase in left-sided filling pressures. Right atrial dilatation and dilatation of the inferior vena cava are common in conditions in which central venous pressure is elevated.

PATIENT SAFETY CONSIDERATIONS

RADIATION EXPOSURE

Both cardiac CT and radionuclide imaging expose patients to ionizing radiation. Several recent publications have raised concern regarding the potential harmful effects of ionizing radiation associated with cardiac imaging. The *effective dose* is a measure used to estimate the biologic effects of radiation and is expressed in millisieverts (mSv). However, measuring the radiation effective dose associated with diagnostic imaging is complex and imprecise and often results in varying estimates, even among experts. The effective dose from a typical myocardial perfusion SPECT scan ranges between ~4 and 11 mSv, depending on the protocol and type of scanner used. The effective dose from a typical myocardial perfusion PET scan is lower, ~2.5–4 mSv. Radiation exposure associated with cardiac CT is variable and, as with radionuclide imaging, also depends on the imaging protocol and scanner used. Although historic radiation doses with cardiac CT have been quite high, the introduction of newer technologies (e.g., x-ray tube modulation, prospective ECG gating) has resulted in a significant dose reduction. The current average radiation dose for a coronary CTA ranges from 5 to 15 mSv and, in selected cases, can be as low as 1 mSv. Imaging laboratories follow the ALARA (as low as reasonably achievable) principle when balancing the clinical need and imaging approach. By comparison, the average dose for invasive coronary angiography is ~7 mSv, whereas exposure to radiation from natural sources in the United States amounts to ~3 mSv annually.

The risk of a fatal malignancy from medical imaging–related radiation is difficult to estimate precisely but is likely small and difficult to discern from the background risk of natural malignancies. The small but potential radiation risks from imaging mandate an assessment of the risk-versus-benefit ratio in the individual patient. In this context, one must not fail to take into account the risks of missing important diagnostic information by not performing a test (which could potentially influence near-term management and outcomes) for a theoretical concern of a small long-term risk of malignancy. Before ordering any test, especially one associated with ionizing radiation, we must ensure the appropriateness of the study and that the potential benefits outweigh the risks. The likelihood that the study being considered will affect clinical management of the patient should be addressed before testing is performed. It is also important that “routine” follow-up scans in asymptomatic individuals be avoided.

CONTRAST AGENTS

Contrast agents are commonly used in cardiac CT, CMR, and echocardiography. Although their use significantly enhances the diagnostic information of each of these tests, there are also potential risks from the administration of contrast agents that should be considered.

The risk of adverse reactions from iodinated contrast agents used in cardiac CT is well established. The precise pathogenesis of contrast reactions following intravascular administration of iodinated contrast media is not known. The overall incidence of contrast reactions is 0.4–3% with nonionic formulations and higher for ionic formulations. Most contrast adverse reactions are mild and self-limiting. The risk of contrast-induced nephropathy (CIN) in patients with relatively normal renal function (estimated glomerular filtration rate [eGFR] >60 mL/min) is low. In most patients, CIN is self-limited, and renal function usually returns to baseline within 7–10 days, without progressing to chronic renal failure. However, this risk increases in patients with GFR <60 mL/min, especially older diabetic subjects. In such patients, appropriate screening and pre- and postscan hydration are necessary.

The use of gadolinium-based contrast agents (GBCAs) enhances the versatility of CMR imaging. There are many commercially available GBCAs in the United States, but their use in cardiac imaging is off-label. Mild reactions from GBCAs occur in ~1% of patients, but severe or anaphylactic reactions are very rare. All GBCAs are chelated to make the compounds nontoxic and facilitate renal excretion. This chelation was less stable in some brands of GBCAs with a linear molecular structure. As a result the unchelated free form of gadolinium leads to a rare but serious condition

known as nephrogenic systemic fibrosis (NSF), which is an interstitial inflammatory reaction manifested as fibrosis of tissues or internal organs and even death. Risk factors to developing NSF include high-dose use in presence of severe renal dysfunction (eGFR <30 mL/min per 1.73 m^2), need for hemodialysis, an eGFR <15 mL/min per 1.73 m^2, acute renal deterioration, and concurrent proinflammatory/systemic illnesses. With widespread routine pretest screening and weight-based dosing, a near-zero incidence of NSF has been reported in the past decade. Most CMR centers use the newer, more stable macrocyclic GBCAs which further lower the risk of NSF.

Contrast agents can also be used in echocardiography. Injected agitated saline is used routinely to assess cardiac shunts, because these "bubbles" are too large to traverse the pulmonary circulation. After saline injection, the presence of bubbles in the left side of the heart is indicative of shunt, although the location can sometimes be difficult to determine. The current U.S. Food and Drug Administration (FDA) approved use of echocardiographic contrast agents is for opacification of left-sided chambers and to improve delineation of left ventricular endocardial border in patients with suboptimal echocardiograms. These agents are either albumin- or lipid-based microspheres filled with inert gases, typically perfluorocarbons. They are considered extremely safe, although they have, in extremely rare instances, been associated with allergic reactions and neurologic events.

SAFETY CONSIDERATIONS OF CMR IN PATIENTS WITH PACEMAKERS AND DEFIBRILLATORS

A presence of a pacemaker or defibrillator is a contraindication to MRI scanning, with patient risks include generation of electrical current from the metallic hardware (especially if wire loops exist), device movement induced by the magnetic field, inappropriate pacing and sensing, and heating because of the "antenna's effect." By contrast, experienced centers had reported success in performing CMR in highly selected patients and in a carefully monitored clinical setting. In past years, several models of permanent pacemakers and automatic implantable cardioverter defibrillator (AICD) have been designed and approved for clinical use to allow CMR imaging under strict imaging conditions and they have achieved FDA approval.

PATIENT-CENTERED APPLICATIONS OF CARDIAC IMAGING

CORONARY ARTERY DISEASE

The basis for the diagnostic application of imaging tests in patients with known or suspected CAD should be viewed considering the pretest probability of disease as well as the specific characteristics of imaging tests (i.e., sensitivity and specificity). In symptomatic patients, the prevalence or pretest probability of CAD differs based on the type of symptom (typical angina, atypical angina, noncardiac chest pain), as well as on age, gender, and coronary risk factors. In an individual patient, the results of the initial test inform the posttest likelihood of CAD. In patients undergoing sequential testing (e.g., ECG treadmill testing followed by stress imaging), the posttest probability of disease after the first test becomes the pretest likelihood of disease for the second test. Regardless of the sequence, the expectation is that a test will provide sufficient information to confirm or exclude the diagnosis of CAD and that such information will allow accurate risk stratification to be able to guide management decisions.

Table 236-3 summarizes the relative diagnostic accuracies of cardiac imaging modalities for the diagnosis of CAD.

TABLE 236-3 Comparative Diagnostic Accuracy of Cardiac Imaging Approaches to Coronary Artery Disease

Imaging Modality	Published Data	Sensitivity	Specificity
Exercise echocardiography	15 studies (n = 1849 patients)	84%	82%
Dobutamine echocardiography	28 studies (n = 2246 patients)	80%	84%
SPECT MPI	113 studies (n = 11,212 patients)	88%	76%
Myocardial perfusion PET	9 studies (n = 650 patients)	93%	81%
CMR perfusion	37 studies (n = 2841 patients)	91%	81%
CMR wall motion	14 studies (n = 754 patients)	83%	86%
Coronary CTA	18 studies (n = 1286 patients)	99%	89%

Note: In these studies, the diagnosis of coronary artery disease was based on the presence of a >50% or >70% stenosis on invasive coronary angiography.

Abbreviations: CMR, cardiac magnetic resonance; CTA, computed tomography angiography; MPI, myocardial perfusion imaging; PET, positron emission tomography; SPECT, single-photon emission computed tomography.

It is important to highlight that most studies included in meta-analyses of the diagnostic accuracy of cardiac imaging modalities for the diagnosis of CAD were retrospective, small, single-center studies, comprising predominantly male patients with a high prevalence of CAD (>50–60%). Multicenter studies assessing the performance of individual modalities or comparing different modalities have consistently resulted in more modest diagnostic accuracies, tracking more closely with how these tests perform in practice.

Stress Echocardiography

The hallmark of myocardial ischemia during stress echocardiography is the development of new regional wall motion abnormalities and reduced systolic wall thickening **(Video 236-3)**. Stress echocardiography can be performed in conjunction with exercise or dobutamine stress. Stress echocardiography is best at identifying inducible wall motion abnormalities in previously normally contracting segments. In a patient with wall motion abnormalities at rest, the specificity of stress echocardiography is reduced, and worsening regional function of a previously abnormal segment might reflect worsening contractile function in the setting of increased wall stress rather than new evidence of inducible ischemia.

VIDEO 236-3: Exercise echocardiogram showing rest images on left and poststress images on right, with parasternal long-axis, upper panel, and apical four-chamber, lower panel, end-systolic frames.

Following exercise, the distal septal/apical region becomes akinetic. A = upper left (UL); B = upper right (UR); C = lower left (LL); D = lower right (LR).



Play Video

The advantages of stress echocardiography over other stress imaging techniques include its relatively good diagnostic accuracy, widespread availability, no use of ionizing radiation, and relatively low cost. Limitations of stress echocardiography include (1) the technical challenges associated with image acquisition at peak exercise because of exertional hyperpnoea and cardiac excursion, (2) the fact that rapid recovery of wall motion abnormalities can be seen with mild ischemia (especially with one-vessel disease, which limits sensitivity), (3) difficulty detecting residual ischemia within an infarcted territory because of resting wall motion abnormality, (4) high operator dependence for acquisition of echocardiographic data and analysis of images, and (5) the fact that good-quality complete images viewing all myocardial segments occurs in only 85% of patients. Newer techniques including second harmonic imaging and the use of intravenous contrast agents improve image quality, but their effect on diagnostic accuracy has not been well documented.

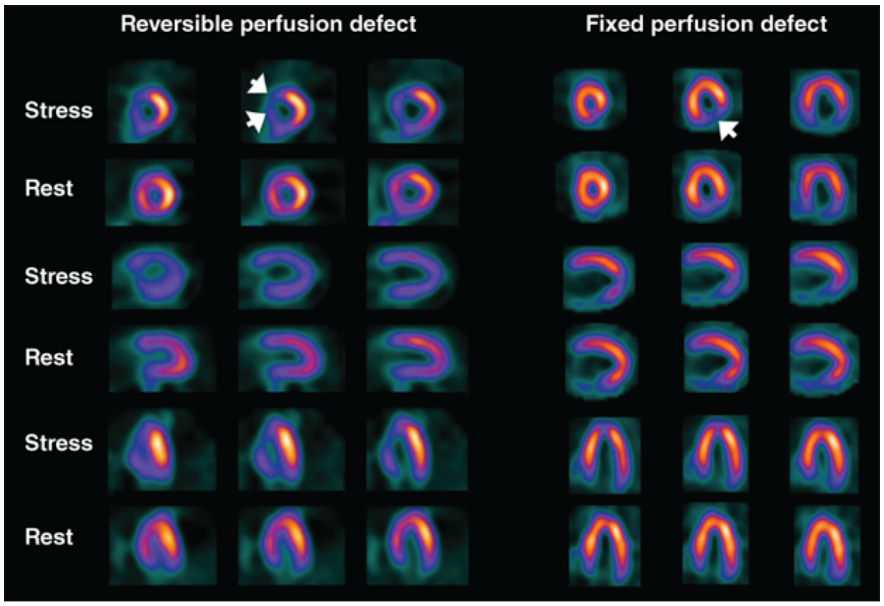
As with nuclear perfusion imaging, stress echocardiography is often used for risk stratification in patients with suspected or known CAD. A negative stress echocardiogram is associated with an excellent prognosis, allowing identification of patients at low risk. Conversely, the risk of adverse events increases with the extent and severity of wall motion abnormalities on stress echocardiography.

Stress Radionuclide Imaging

SPECT myocardial perfusion imaging is the most common form of stress imaging tests for CAD evaluation. The presence of a reversible myocardial perfusion defect is indicative of ischemia **(Fig. 236-9, left panel)**, whereas a fixed perfusion defect generally reflects prior myocardial infarction **(Fig. 236-9, right panel)**. As discussed above, PET has advantages compared to SPECT, but it is not widely available and is more expensive and, thus, considered an emerging technology in clinical practice.

FIGURE 236-9

Selected technetium-99m sestamibi myocardial perfusion single-photon emission computed tomography images of two different patients demonstrating a reversible perfusion defect involving the anterior and septal left ventricular wall, reflecting ischemia in the left anterior descending coronary territory (*arrows in left panel*), and a fixed perfusion defect involving the inferior and inferolateral walls consistent with myocardial scar in the right coronary territory (*arrow in right panel*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

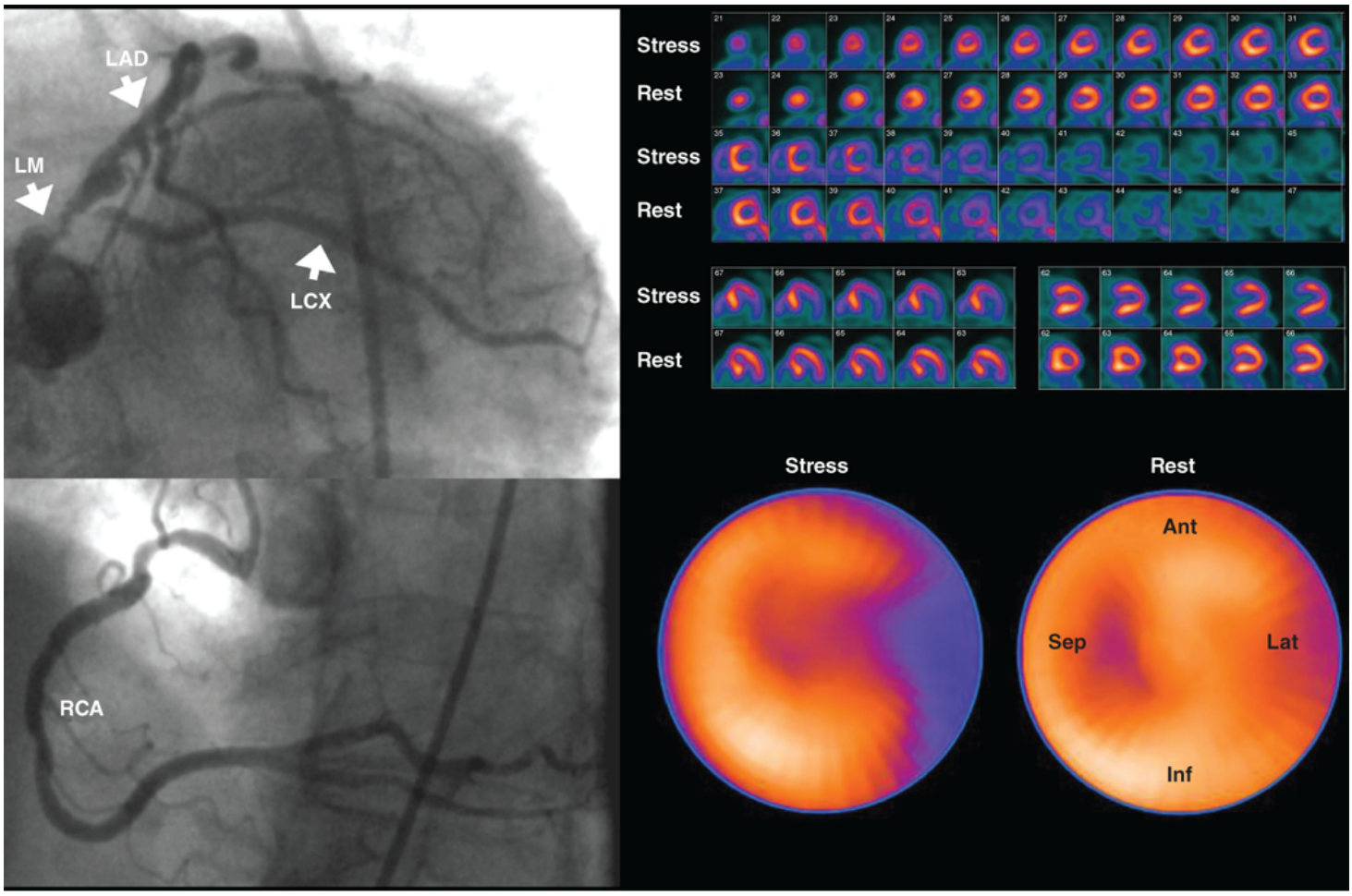
Nuclear perfusion imaging is another robust approach to diagnose obstructive CAD, quantify the magnitude of inducible myocardial ischemia, assess the extent of tissue viability, and guide therapeutic management (i.e., selection of patients for revascularization). One of the most valuable clinical applications of radionuclide perfusion imaging is for risk stratification. It is well established that patients with a normal SPECT or PET study exhibit a low rate of major adverse cardiac events of <1% annually. Importantly, the risks of death and myocardial infarction increase linearly with increasing magnitude of perfusion abnormalities, reflecting the extent and severity of CAD.

Despite the widespread use and clinical acceptance of radionuclide imaging in CAD evaluation, a recognized limitation of this approach is that it often uncovers only coronary territories supplied by the most severe stenoses. Consequently, it is relatively insensitive to accurately delineate the extent of obstructive angiographic CAD, especially in the setting of multivessel disease. The use of quantitative myocardial blood flow and coronary flow reserve with PET can help mitigate this limitation. In patients with so-called “balanced” ischemia or diffuse CAD, measurements of coronary flow reserve uncover areas of myocardium at risk that would generally be missed by performing only relative assessments of myocardial perfusion **(Fig. 236-10)**. Conversely, a normal coronary flow reserve is associated with a very high negative predictive value for excluding high-risk angiographic CAD. These measurements of coronary flow reserve also contribute to risk stratification across the spectrum of ischemic changes, including patients with visually normal myocardial perfusion.

FIGURE 236-10

Coronary angiographic (*left panel*) and rubidium-82 myocardial perfusion positron emission tomography images (*right panel*) in an 85-year-old female with diabetes presenting with chest pain. The coronary angiogram demonstrates significant stenoses of the left main and

circumflex coronary arteries. However, the perfusion images demonstrate only a reversible lateral wall defect. Quantification of stress and rest myocardial blood flow demonstrated a significant, global reduction on coronary flow reserve (estimated at 1.2, normal value >2.0), reflecting extensive myocardium risk that was underestimated by the semiquantitative estimates of myocardial perfusion. LAD, left anterior descending artery; LCX, left circumflex artery; LM, left main artery; RCA, right coronary artery.



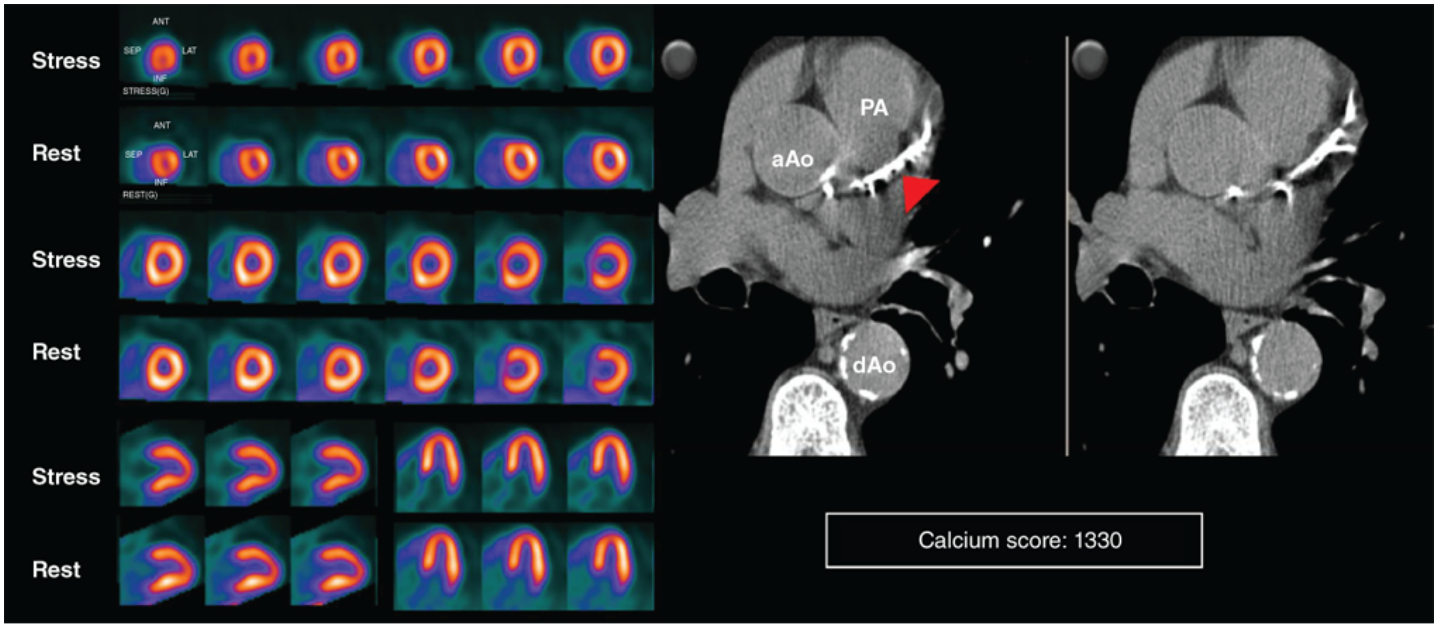
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HYBRID CT AND NUCLEAR PERFUSION IMAGING

Because many of the newer generation nuclear medicine scanners integrate CT and a gamma camera in the same acquisition gantry, it is now possible to acquire and quantify myocardial scar and ischemia and CAC scoring from a single dual-modality study (SPECT/CT or PET/CT) **(Fig. 236-11)**. The rationale for this integrated approach is predicated on the fact that the perfusion imaging approach is designed to uncover only obstructive atherosclerosis. Conversely, CAC scoring provides a quantitative measure of the anatomic extent of atherosclerosis. This provides an opportunity to improve the conventional models for risk assessment using nuclear imaging alone, especially in patients without known CAD.

FIGURE 236-11

Stress and rest rubidium-82 myocardial perfusion positron emission tomography (PET) images (*left*) and noncontrast gated computed tomography (CT) images (*right*) delineating the extent and severity of coronary artery calcifications obtained with integrated PET/CT imaging. The images demonstrate extensive atherosclerosis (Agatston coronary calcium score = 1330) without flow-limiting disease based on the normal perfusion study. aAo, ascending aorta; dAo, descending aorta; PA, pulmonary artery.



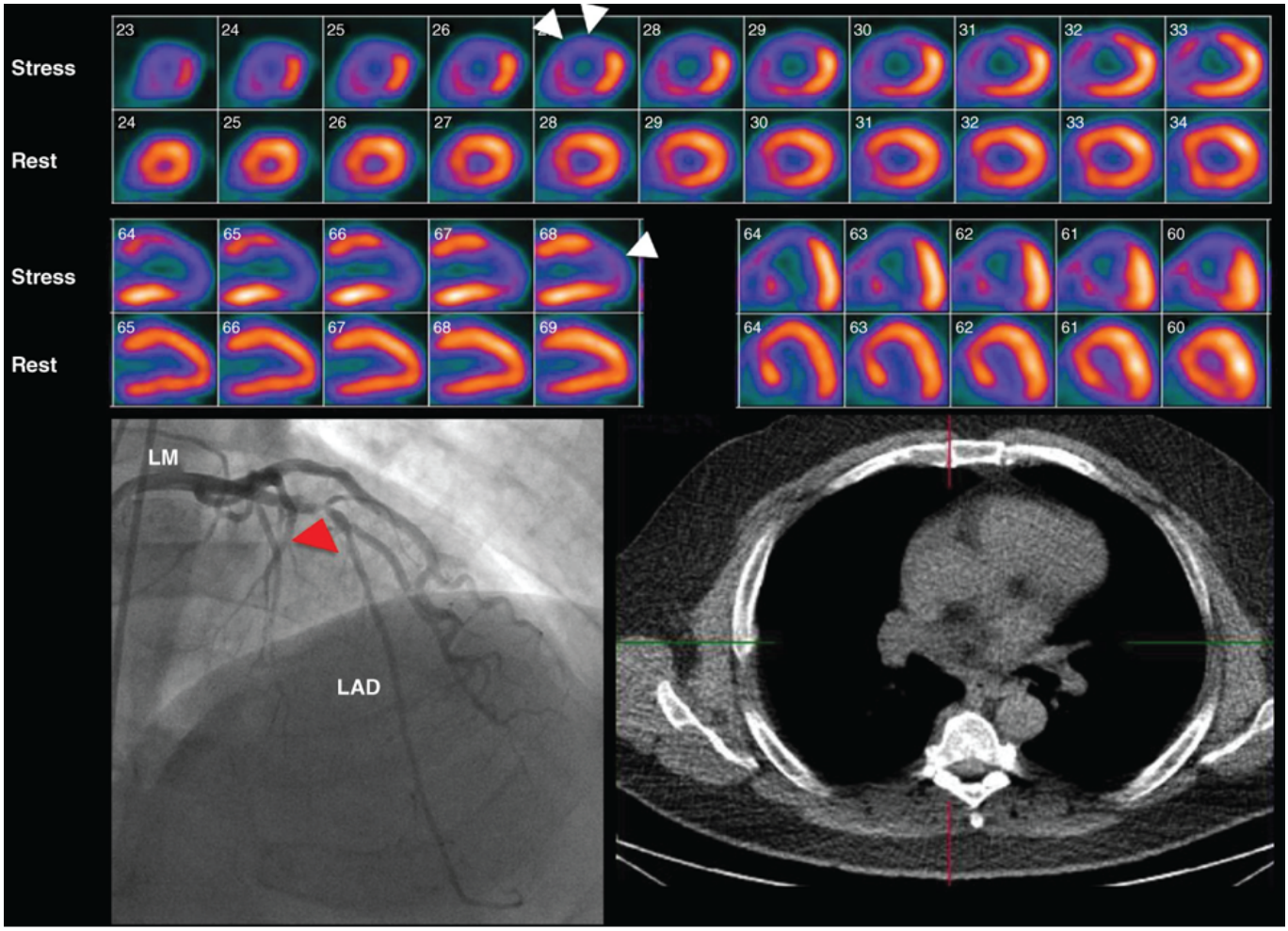
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Cardiac CT

Voluminous plaques are more prone to calcification, and stenotic lesions frequently contain large amounts of calcium. Indeed, there is evidence that high CAC scores are generally predictive of a higher likelihood of obstructive CAD, and the available data support the concept of a threshold phenomenon governing this relationship (i.e., Agatston score >400). However, given the fact that CAC scores are not specific markers of obstructive CAD, one should be cautious in using this information as the basis for referral of patients to coronary angiography, especially in symptomatic patients with low-risk stress tests. Conversely, CAC scores <400, especially in symptomatic patients with intermediate-high likelihood of CAD, as in those with typical angina, may be less effective in excluding CAD, especially in young symptomatic men and women who may have primarily noncalcified atherosclerosis **(Fig. 236-12)**.

FIGURE 236-12

Stress and rest rubidium-82 myocardial perfusion positron emission tomography images (*top*), noncontrast gated computed tomography images (*lower right*), and selected coronary angiographic images obtained on a 59-year-old male patient with atypical angina. Despite the absence of significant coronary calcifications (Agatston calcium score = 0), the perfusion images demonstrated a dense and reversible perfusion defect involving the anterior and anteroseptal walls (*arrows*), reflecting significant obstructive disease in the left anterior descending coronary artery (LAD), confirmed on angiography. LM, left main artery.

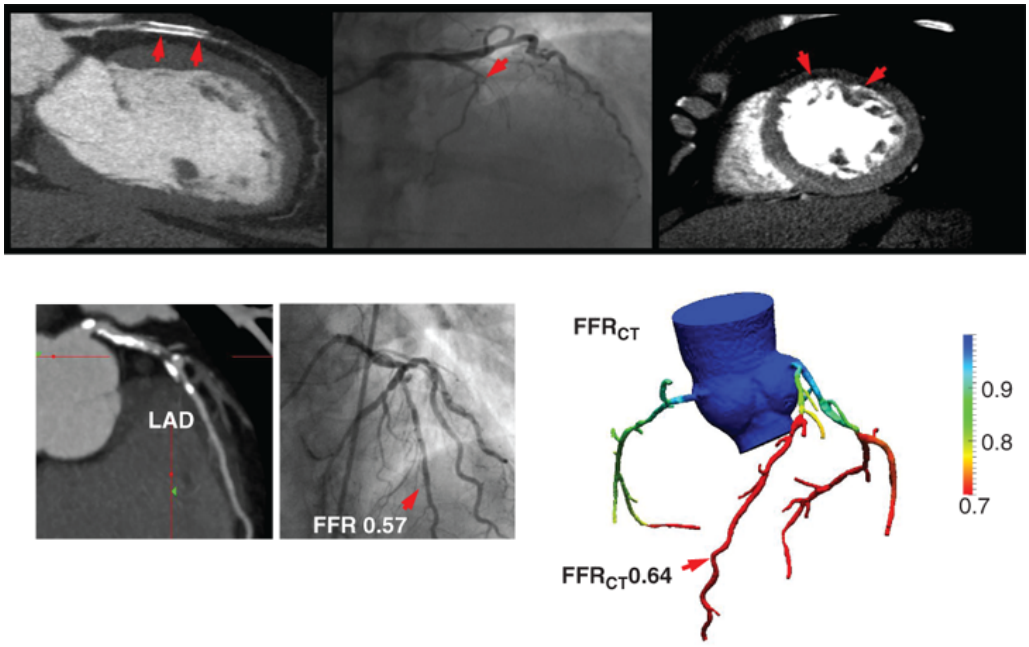


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

As discussed above, the improved temporal and spatial resolution of modern multidetector CT scanners offer a unique noninvasive approach to delineate the extent and severity of coronary atherosclerosis with coronary CTA. The extremely high sensitivity of this approach offers a very effective means for excluding the presence of CAD (high negative predictive value) (Table 236-3). In the setting of high coronary calcium scores (e.g., >400), however, specificity is reduced because the blooming artifact of calcium does not allow one to evaluate the vessel lumen accurately. Given the high negative predictive value of CTA, a normal scan result effectively excludes obstructive CAD and abolishes the need for further investigation. As discussed below, this may be quite useful in patients with low-intermediate clinical risk presenting to the emergency room for chest pain. However, the limited capability of this technique to determine which coronary plaques are flow limiting can make abnormal scan results more difficult to interpret, especially in terms of the possible need of revascularization. There are emerging data suggesting that by adding a stress myocardial perfusion CT evaluation (similar to stress perfusion CMR) **(Fig. 236-13, top panel)** or an estimated fractional flow reserve (so-called FFR_{CT}) **(Fig. 236-13, lower panel)**, one can define the hemodynamic significance of anatomic stenosis. However, these are not in routine clinical use and remain emerging technologies.

FIGURE 236-13

Examples of novel approaches to the assessment of flow-limiting coronary artery disease (CAD) with cardiac computed tomography (CT). In the *top panel*, representative views of a coronary CT angiogram (CTA; *left*), coronary angiogram (*middle*), and stress myocardial perfusion CT (*right*) images in a patient with CAD and prior stenting of the left anterior descending coronary artery (LAD) are presented. On the CTA, the stent (*arrows*) is totally occluded as evidenced by the loss of contrast enhancement distal to the stent. The coronary angiogram demonstrates a concordant total occlusion of the LAD. On the perfusion CT images, there is a black rim (*arrows*) involving the anterior and anterolateral walls, indicating the lack of contrast opacification during stress consistent with myocardial ischemia. *(Images courtesy of CORE 320 investigators.)* The *lower panel* illustrates an example of fractional flow reserve (FFR) estimates with coronary CTA (*left*) compared to the reference standard of invasive FFR. The FFR reflects the pressure differential between a coronary segment distal to a stenosis and the aorta. In normal coronary arteries, there is no gradient, and FFR is 1. An FFR <0.80 is consistent with a hemodynamically significant stenosis. *(Images courtesy of Dr. James Min, Cornell University, New York.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

As with invasive coronary angiography, assessments of the extent of CAD by CTA can also provide useful prognostic information. A low 1-year cardiac event rate has been reported for patients without coronary atherosclerosis on CTA. For patients with obstructive CAD, the risk of adverse cardiac events increases proportionally with the extent of angiographically obstructive CAD. There is new evidence that even the presence of non-obstructive atherosclerosis increases the risk of adverse cardiac events.

Although CTA can be helpful in assessing patency of bypass grafts, the assessment of stents is somewhat more challenging because the limited spatial resolution of CT and stent diameter (<3 mm being associated with the highest number of partial lumen visualization and nondiagnostic scans) both contribute to limited clinical results.

CMR Imaging

CMR evaluates for ischemia from CAD by assessing regional myocardial perfusion or regional wall motion at rest and during stress, the latter analogous to dobutamine echocardiography. Stress CMR studies use pharmacologic vasodilator agents or dobutamine. Myocardial perfusion is evaluated by injecting a GBCA bolus followed by imaging data acquisition as the contrast passes through the cardiac chambers and into the myocardium. Relative perfusion deficits are recognized as regions of low signal intensity (black) within the myocardium **(Video 236-4)**. Several minutes after GBCA injection, LGE imaging allows detection of bright areas of myocardial scar (white), which permits comparison of regions of hypoperfusion and infarction to quantify myocardial ischemia **(Fig. 236-14)**.

VIDEO 236-4: Cardiac magnetic resonance (CMR) myocardial perfusion imaging during vasodilating stress, in three parallel-short-axis views.

A bolus of gadolinium contrast was injected intravenously while rapid imaging acquisition occurred. The contrast enhances the right ventricle first, then travels through the pulmonary circulation, enters the left ventricle (LV), and then perfuses the LV myocardium. Myocardial perfusion defects with this technique show as black subendocardial rims, reflecting lack of contrast accumulation due to ischemia and/or scar. In this case, the anterior wall has a severe perfusion defect (*red arrow*). Figure 236-14 shows the late gadolinium enhancement (LGE) image of a mid short-axis view. There is no evidence of infarction in the anterior wall, which would be seen as bright white areas, indicating that the stress perfusion defect primarily represents myocardial ischemia. This patient had a significant stenosis of the left anterior descending coronary artery.

Image not available.
Play Video

FIGURE 236-14

The image shows the late gadolinium enhancement image of a mid short-axis view. There is no evidence of infarction in the anterior wall, which would be seen as bright white areas, indicating that the stress perfusion defect primarily represents myocardial ischemia. This patient had a significant stenosis of the left anterior descending coronary artery.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

With better delineation of the endocardial borders, dobutamine CMR has better diagnostic accuracy than dobutamine echocardiography for detection of CAD, especially in patients with poor acoustic window (Table 236-3). High-dose dobutamine carries the risk of serious ventricular arrhythmias (~1%), but most cases can be prevented with proper monitoring of vital signs and regional cine function. The advantages of stress perfusion CMR over SPECT include its higher spatial resolution which allows detection of subendocardial ischemia or infarction that may be missed by SPECT. As with other imaging modalities, stress CMR studies can provide robust prognostic value. A normal CMR study in patients with chest pain at intermediate pretest risk is associated with a <1% annual rate of death or myocardial infarction.

Selecting a Testing Strategy in Patients without Known CAD

As discussed above, there are many options for the evaluation of a patient with suspected CAD presenting with chest pain symptoms. The critical questions to be answered by a testing strategy include the following: (1) Does the chest pain reflect obstructive CAD? (2) What are the short- and long-term risks? (3) Does the patient need to be considered for revascularization?

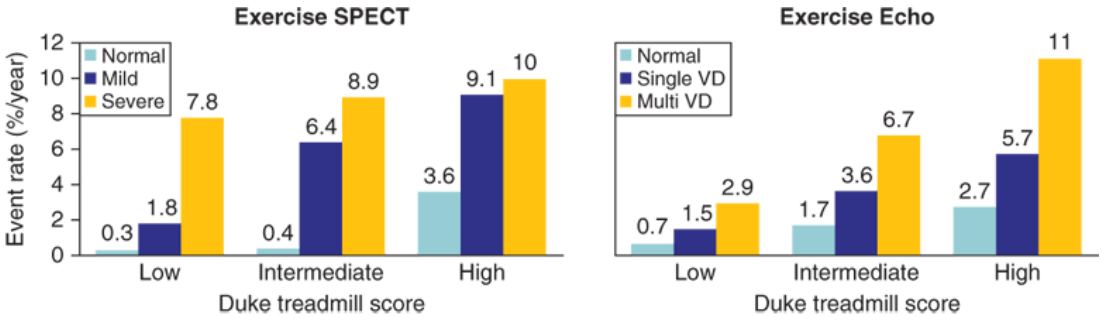
For symptomatic patients without a prior history of CAD and a normal or nearly normal resting ECG who are able to exercise, the American College of Cardiology/American Heart Association guidelines recommend standard exercise treadmill testing (ETT) as the initial testing strategy. The guidelines further suggest that patients who are categorized as low risk by ETT (e.g., those achieving >10 metabolic equivalents [METS] without chest pain or ECG changes) be treated initially with medical therapy, and those with high-risk ETT findings (i.e., typical angina with >2 mm ST-segment depression in multiple leads, ST elevation during exercise, drop in blood pressure, or sustained ventricular arrhtyhmias) be referred for coronary angiography.

The use of exercise testing in women presents difficulties that are not seen in men, reflecting the differences in the lower prevalence of obstructive CAD in women and the different accuracy of exercise testing in men and women. Compared with men, the lower pretest probability of disease in women means that more test results are false positive. In some of these patients, a positive ETT may reflect true myocardial ischemia caused my microvascular coronary artery dysfunction (so-called *microvascular disease*). In addition, the inability of many women to exercise to maximum aerobic capacity, the greater prevalence of mitral valve prolapse and microvascular disease, and possibly other reasons may contribute to the differences with men as well. The difficulties of using exercise testing for diagnosing obstructive CAD in women have led to speculation that stress imaging may be preferred over standard stress testing. However, recent data from the WOMEN study suggests that in symptomatic, low-risk women who can exercise, standard ETT is a very effective initial diagnostic strategy as compared to stress radionuclide imaging. Indeed, the 2-year outcomes were similar in both diagnostic strategies, and the ETT-first approach resulted in 48% lower costs compared to exercise radionuclide imaging.

Patients with intermediate-high risk after ETT (e.g., low exercise capacity, chest pain, and/or ST-segment depression without high-risk features) will often require additional testing, either stress imaging or coronary CTA, to more accurately characterize clinical risk. Most common stress imaging strategies in intermediate-risk patients include stress echocardiography and radionuclide imaging. In such patients, stress imaging with either SPECT or echocardiography has been shown to accurately reclassify patients who are initially classified as intermediate risk by ETT as low or high risk **(Fig. 236-15)**. Following this staged strategy of applying the low-cost ETT first and reserving more expensive imaging to refine risk stratification to patients initially classified as intermediate risk by ETT is more cost-effective than applying stress or anatomic imaging as the initial test routinely.

FIGURE 236-15

Incremental risk stratification of stress imaging over Duke treadmill score in patients with suspected coronary artery disease. Stress imaging is most valuable in the intermediate-risk group. SPECT, single-photon emission computed tomography; VD, vessel disease. *(Part A reproduced with permission from R Hachamovitch et al: Circulation 93:905, 1996. Part B data from TH Marwick et al: Circulation 103:2566, 2001.)*

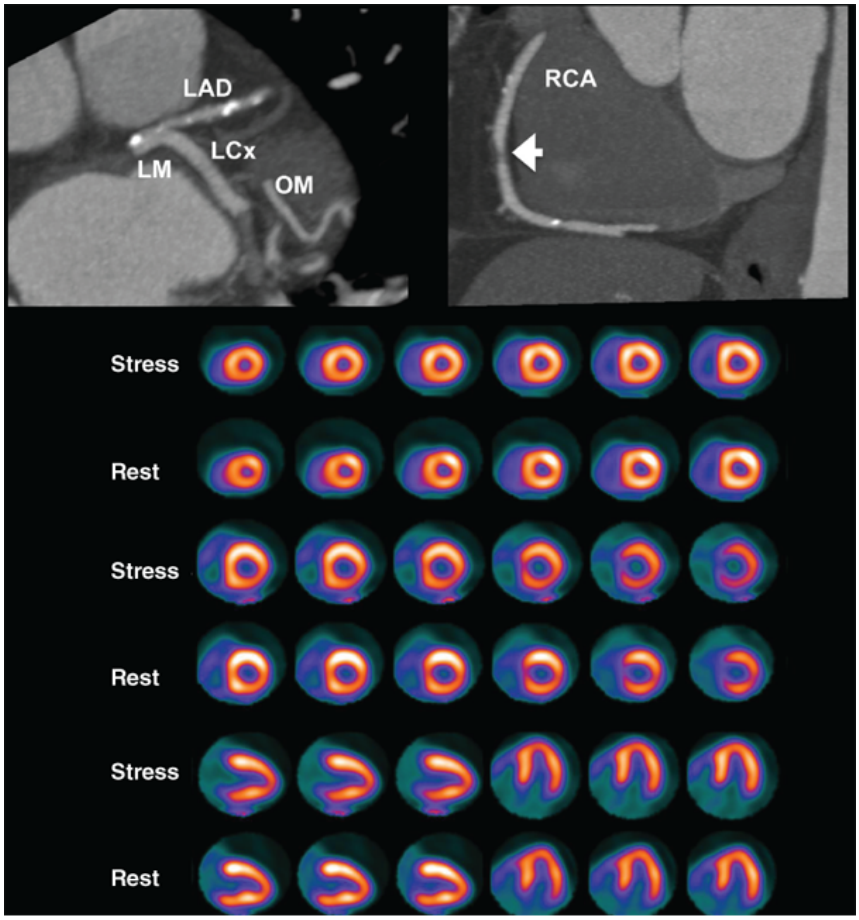


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

An imaging strategy is the recommended first step for patients who are unable to exercise to an adequate workload and/or those with abnormal resting ECGs (e.g., left ventricular hypertrophy with strain, left bundle branch block). Importantly, the most recent documents regarding appropriate use of imaging also considered that an imaging strategy may be an appropriate first step in patients with intermediate-high likelihood of CAD (e.g., diabetics, renal impairment) due to increased overall sensitivity for diagnosis of CAD and improved risk stratification. In considering an imaging strategy, the evidence supporting the role of ischemia assessment versus anatomy must be considered. From the discussion above, a normal coronary CTA is helpful because it effectively excludes the presence of obstructive CAD and the need for further testing, defines a low clinical risk, and makes management decisions regarding referral to coronary angiography straightforward. Because of its limited accuracy to define stenosis severity and predict ischemia, however, abnormal CTA results are more problematic to interpret and to use as the basis for defining the potential need of invasive coronary angiography and revascularization. In such patients, a follow-up stress test is usually required to determine the possible need of revascularization **(Fig. 236-16)**.

FIGURE 236-16

Selected views from coronary computed tomography angiographic (CTA) images (*top panel*) and stress and rest rubidium-82 myocardial perfusion positron emission tomography images (*lower panel*) obtained on a 64-year-old male patient with atypical angina. The CTA images demonstrate dense focal calcifications in the left main (LM) and left anterior descending (LAD) coronary arteries and a significant noncalcified plaque in the mid right coronary artery (RCA; *arrow*). The myocardial perfusion images demonstrated no evidence of flow-limiting stenosis. LCx, left circumflex artery; OM, obtuse marginal branch.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The justification of stress imaging in testing strategies has hinged on the identification of which patients may benefit from a revascularization strategy by means of noninvasive estimates of jeopardized myocardium rather than angiography-derived anatomic stenoses. Indeed, there is evidence that only the presence of moderate-severe ischemia identifies patients with apparent improved survival with revascularization. Patients with mild or no ischemia are better candidates for optimal medical therapy. The advantages of this approach include avoidance of excess catheterizations with their associated cost and risk and the potential for intervening unnecessarily. The acceptable diagnostic accuracy of stress imaging approaches, along with their robust risk stratification, and the ability of ischemia information to identify patients who would benefit from revascularization suggest a potential role as a first imaging strategy in patients with intermediate-high likelihood of CAD. While the available data suggest similar diagnostic accuracy for SPECT and echocardiography but higher for PET and CMR, the choice of strategy depends on availability and local expertise.

Selecting a Testing Strategy in Patients with Known CAD

Use and selection of testing strategies in symptomatic patients with established CAD (i.e., prior angiography, prior myocardial infarction, prior revascularization) differ from those in patients without prior CAD. Although standard ETT may help distinguish cardiac from non-cardiac chest pain, exercise ECG has several limitations following myocardial infarction and revascularization (especially coronary artery bypass grafting). These patients frequently have rest ECG abnormalities. In addition, there is a clinical need to document both the magnitude and localization of ischemia to be able to direct therapy, especially the potential need for targeted revascularization. Consequently, imaging tests are preferred for evaluating patients with known CAD.

There are also important differences in the effectiveness of imaging tests in these patients. As discussed above, coronary CTA is limited in patients with prior revascularization. While CTA provides excellent visualization of the bypass grafts, the native circulation tends to get heavily calcified and is generally not a good target for imaging with CTA. Likewise, blooming artifacts from metallic stents also limit the application of coronary CTA in patients with prior percutaneous coronary intervention. If an anatomic strategy is indicated, direct referral to invasive angiography is preferred.

Stress imaging approaches are especially useful and preferred in symptomatic patients with established CAD. As in patients without prior CAD, normal imaging studies in symptomatic patients with established CAD also identify a low-risk cohort. In those with abnormal stress imaging studies, the degree of abnormality relates to posttest risk. In addition, stress imaging approaches can localize and quantify the magnitude of ischemia, thereby assisting in planning targeted revascularization procedures. As in patients without prior CAD, the choice of stress imaging strategy depends on availability and local expertise.

Testing Strategy Considerations in Patients Presenting with Chest Pain to the Emergency Department

Although acute chest pain is a frequent reason for patient visits to the emergency department (ED), only a small minority of those presentations represent an acute coronary syndrome (ACS). Strategies used in the evaluation of these patients include novel cardiac biomarkers (e.g., serum

troponins), conventional stress testing (ETT), and noninvasive cardiac imaging. It is generally accepted that the primary goal of this evaluation is exclusion of ACS and other serious conditions rather than detection of CAD.

The routine evaluation of acute chest pain in most centers in the United States includes admission to a chest pain unit to rule out ACS with the use of serial ECGs and cardiac biomarkers. In selected patients, stress testing with or without imaging may be used for further risk stratification. Stress echocardiography and radionuclide imaging are among the most frequently used imaging approaches in these patients. Multiparametric CMR imaging has also been used successfully in patients with acute chest pain **(Video 236-5)**. Due to its ability to probe multiple aspects of myocardial physiology, cardiac anatomy, and tissue characterization with LGE imaging, CMR is useful in diagnosing conditions that mimic ACS (e.g., acute myocarditis, takotsubo cardiomyopathy, pericarditis) **(Fig. 236-17)**.

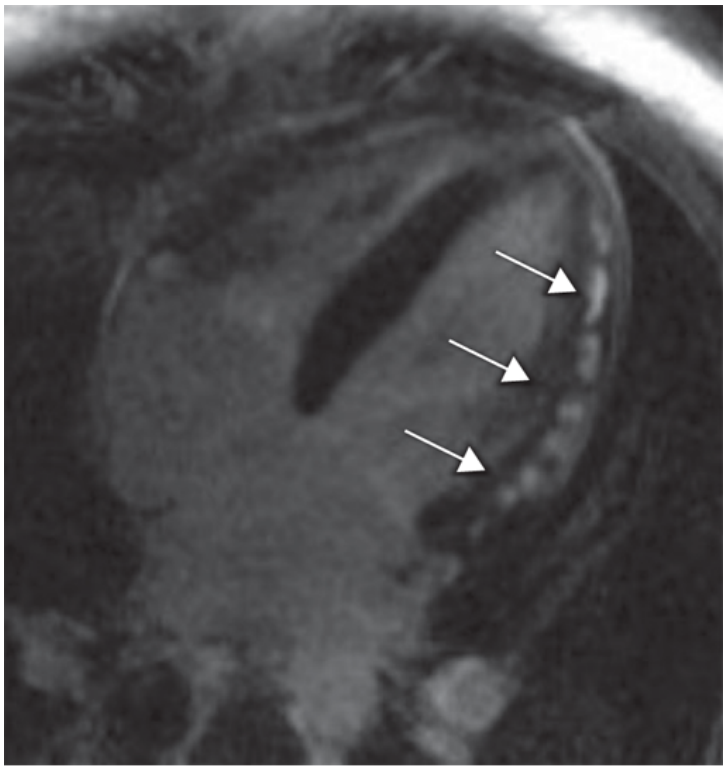
VIDEO 236-5: A 60-year-old female presented with intermittent chest pain of 3 days in duration but was pain free at the time of assessment in the emergency room.

Admission electrocardiogram (ECG) demonstrated T-wave inversion in the anterior precordial lead, but cardiac enzymes were normal. A resting cardiac magnetic resonance (CMR) study reviewed a large area of anteroseptal hypokinesia (*left picture*, region of hypokinesia shown by the *red arrows*), matching with a large resting perfusion defect (*middle picture*, perfusion defect shown by the *blue arrows*). Late gadolinium enhancement (LGE) imaging (*right picture*), however, did not show any enhancement to indicate any infarction in the anteroseptal wall, suggesting that the hypocontractile and hypoperfused anteroseptal wall was viable. Urgent coronary angiography demonstrated an acute thrombus in the mid left anterior descending coronary artery, which required coronary stenting. This case represents an example of acute coronary syndrome with hibernating but viable myocardium in the anteroseptal wall. The anteroseptal wall recovered contractile function when reassessed 6 months later.

Image not available.
Play Video

FIGURE 236-17

A four-chamber long-axis late gadolinium enhancement (LGE) image of a patient with acute myocarditis. Note that the LGE primarily involved the epicardial aspect of the myocardium (*arrows*), sparing the endocardium, which is a feature that distinguishes myocarditis from myocardial infarction, which affects the endocardium. Also note the multiple foci of LGE in this case affecting the lateral wall of the left ventricle. Viral myocarditis often presents with this pattern.

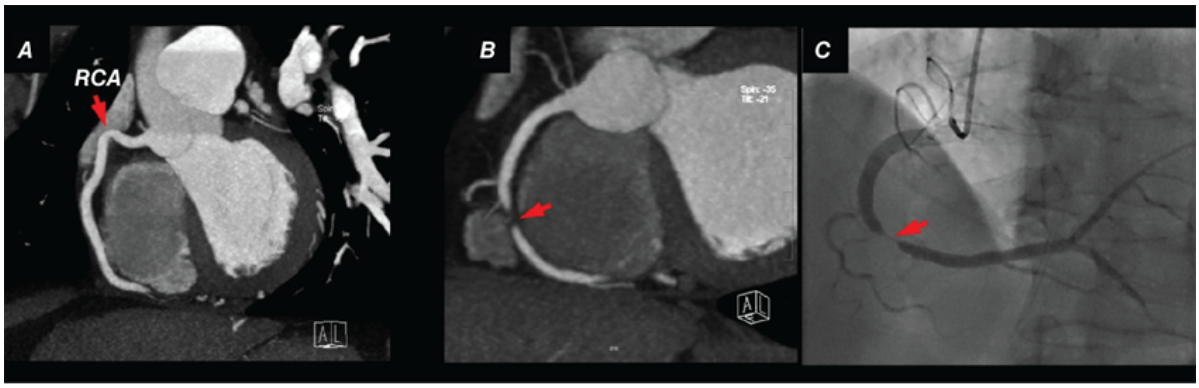


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

As discussed above, coronary CTA is a rapid and accurate imaging technique to exclude the presence of CAD and is well suited for the evaluation of patients with acute chest pain **(Fig. 236-18)**. Four randomized clinical trials have demonstrated the feasibility, safety, and accuracy of coronary CTA in the ED as compared to usual care (which typically includes stress imaging). Patients in these trials had a very low clinical risk. Overall, there were no deaths and very few myocardial infarctions without differences between the groups. Likewise, there were no differences in postdischarge ED visits or rehospitalizations. These studies showed decreased length of stay with coronary CTA, and most but not all reported cost savings. An observation from a recent meta-analysis was that, compared to usual care, more patients assigned to coronary CTA underwent cardiac catheterization (6.3% vs 8.4%, respectively) and revascularization (2.6% vs 4.6%, respectively). The relative increased frequency in the referral to cardiac catheterization and revascularization after coronary CTA compared to stress imaging testing strategies has also been observed in patients with stable chest pain syndromes.

FIGURE 236-18

Representative coronary computed tomography angiographic (CTA) images of two patients presenting to the emergency department with chest pain and negative biomarkers. The patient in ***A*** had angiographically normal coronary arteries; the panel shows a representative view of the right coronary artery (RCA). ***B*** and ***C*** show a corresponding significant stenosis in the mid portion of the RCA on both the CTA (***B***) and invasive angiographic view (***C***). *(Images used with permission from Dr. Quynh Truong, Massachusetts General Hospital, Boston, MA.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Taken together, the available data clearly suggest that not all patients presenting with acute chest pain require specialized imaging testing. Patients with very low clinical risk and negative biomarkers (especially high-sensitivity troponin assays) can be safely triaged. The use of imaging tests in patients with low-intermediate risk should be carefully considered, especially given the trade-offs discussed above.

VALVULAR HEART DISEASE

Abnormalities of any of the four valvular structures in the heart can lead to significant cardiac dysfunction, heart failure, or even death. Echocardiography, CMR, and cardiac CT can be used for the evaluation of valvular heart disease, although echocardiography has generally been considered the first imaging test for the assessment of valvular heart disease. In addition, echocardiography is the most cost-effective screening method for valvular heart disease. In some cases, CMR can complement echocardiography when echocardiographic acoustic window is inadequate, quantifying blood flow data more precisely, or providing complimentary assessment of adjacent vascular structures relevant to the valvular condition.

Echocardiography can be used to assess both regurgitant and stenotic lesions of any of the cardiac valves. Typical indications for echocardiography to assess valvular heart disease include cardiac murmurs identified on physical examination, symptoms of breathlessness that may represent valvular heart disease, syncope or presyncope, and preoperative exams in patients undergoing bypass surgery. A standard echocardiographic examination should include qualitative and quantitative assessment of all valves regardless of indication and should serve as an adequate screening test for significant valvular disease.

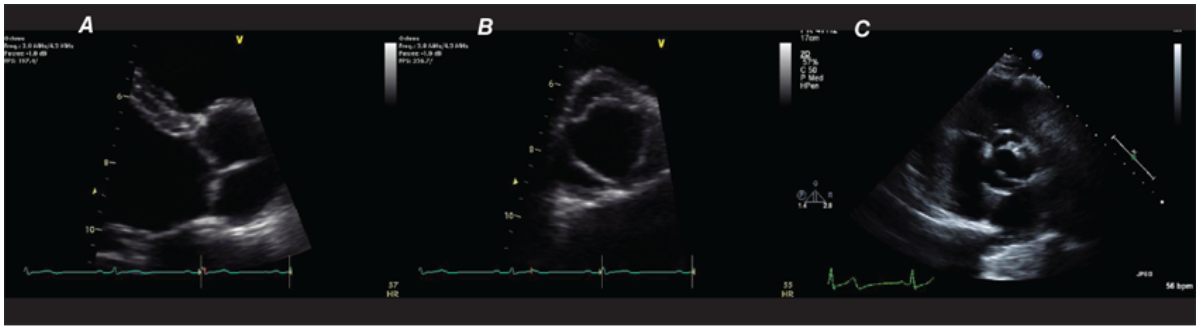
Assessment of Aortic Stenosis

Aortic stenosis, one of the most common forms of valvular heart disease, most often occurs because of gradual progression of valvular calcification in both normal and congenitally abnormal valves. Assessment of aortic stenosis is most commonly performed with echocardiography, although techniques for quantitative assessment of aortic stenosis with CMR have been developed and increasingly used over the past decade. Echocardiographic assessment generally begins with visual inspection of the valve. This allows for assessment of valvular morphology, whether it is tricuspid, bicuspid, or some variant; degree of leaflet calcification; and leaflet excursion.

The normal aortic valve consists of three leaflets or cusps: the right coronary, the left coronary, and the noncoronary cusps. Abnormalities of cusp development are some of the most common congenital heart anomalies, the most common of which is bicuspid aortic valve, with two opening leaflets rather than three **(Fig. 236-19)**. The aortic valve can be visualized on echocardiography, although sometimes it can be difficult to distinguish true bicuspid aortic valve from variants, including the presence of a vestigial commissure (raphe). Bicuspid aortic valve, one of the most common congenital anomalies, predisposes to both aortic stenosis and aortic insufficiency.

FIGURE 236-19

Normal aortic valve in the parasternal long-axis view (***A***) and short-axis view (***B***), and bicuspid aortic valve showing typical 10 o'clock to 4 o'clock leaflet orientation (***C***).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The degree of aortic stenosis is assessed by estimating both the pressure gradient across the valve and the valve area. Patients with moderate aortic stenosis or higher generally have peak instantaneous velocities of 3.0 m/s and higher, and often higher than 4.0 m/s, corresponding to pressure gradients of 36 and 64 mmHg, respectively. Because pressure gradients across the aortic valve can be underestimated in patients with severe left ventricular dysfunction, estimation of valve area by the continuity principle is the most accurate technique for assessing the severity of the stenosis. However, evaluation of the patient with so-called low-flow or low-gradient aortic stenosis can be challenging and can sometimes require provocative testing such as dobutamine echocardiography. In these cases, it is important to distinguish whether the valve is indeed capable of opening further or simply behaving like a stenotic valve because of the low-pressure gradient.

Aortic valve areas <1.0 cm^2 are generally considered severe, and valve areas <0.6 cm^2 are considered critical. Because patients with good left ventricular function can often tolerate severe aortic stenosis for a considerable period of time, valve areas or gradients alone should not be used to determine whether an individual patient should undergo aortic valve surgery, as this remains a clinical decision.

Some patients with apparent aortic stenosis have subvalvular or even supravalvular obstruction. Hypertrophic cardiomyopathy represents the classic form of subvalvular aortic stenosis, but this is usually easily distinguished from aortic stenosis on echocardiography as the valve leaflets can be seen opening during systole. Subaortic membranes can behave very similarly to leaflet aortic stenosis, and the membranes themselves can be very thin and difficult to visualize, although the presence of a murmur, a gradient across the valve with aortic leaflets that appear to open normally, is highly suggestive of a membrane. Supravalvular aortic stenosis, although exceedingly rare, also occurs.

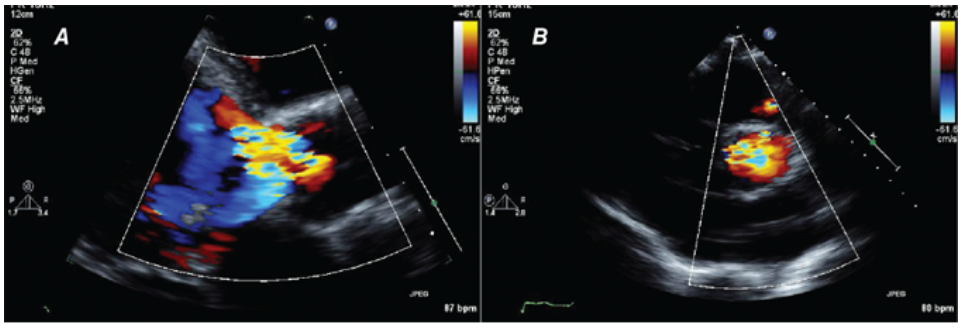
The emergence of transcatheter aortic valve intervention as a therapeutic option for patients with severe aortic stenosis who are not optimal candidates for surgical replacement has resulted in a very important clinical role for multimodality imaging. Imaging plays a critical role in preprocedural planning, intraprocedural implantation optimization, and follow-up of these patients. CT plays an important role in defining the eligibility of the proposed access site (CTA of the aorta and iliac arteries) and in defining the anatomic relationships between the aortic valve and aortic root, left ventricle, and coronary ostia. Cardiac CT and transesophageal echocardiography are also used to define the device size. Transesophageal echocardiography is used during the device implantation to ensure the best prosthesis–patient match, to assess prosthesis position and function after deployment, and to identify immediate complications (e.g., aortic insufficiency, paravalvular leak resulting from patient–prosthesis mismatch). Echocardiography is the imaging modality of choice for long-term surveillance.

Assessment of Aortic Regurgitation

Assessment of aortic regurgitation requires qualitative assessment of the aortic valve structure. Aortic regurgitation is common with congenital abnormalities of the aortic valve, the most common of which is bicuspid aortic valve. Aortic regurgitation often coexists with aortic stenosis, and it is not uncommon for patients to have both severe aortic stenosis and regurgitation. Congenital abnormalities of the aortic leaflets, such as bicuspid aortic valve, are common causes of aortic insufficiency. Dilatation of the aortic root, as occurs in patients with hypertension and other disorders in which aortic dilatation can occur, can also lead to aortic regurgitation even when the valve leaflets are intrinsically normal due to malcoaptation of the leaflets. Aortic root dilatation is common in patients with aortic regurgitation, both as a cause or coexisting lesion, and the aortic root and ascending aorta should be measured and followed in these patients **(Fig. 236-20)**.

FIGURE 236-20

Aortic regurgitation visualized by color flow Doppler in the parasternal long-axis view (***A***) and the parasternal short-axis view (***B***).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Because aortic regurgitation can result in dilatation of the left ventricle over time with ultimate reduction in ventricular function, caring for the patient with aortic regurgitation requires serial assessment of ventricular size and function. Patients whose ventricles dilate beyond an end-systolic

diameter of 5.5 cm or whose LVEF declines below normal are at significantly higher risk of death or heart failure, and these measures are often used to decide the need for valve surgery. Quantitation of regurgitation itself can be performed using a number of methods. Semiquantitative visual assessment of aortic regurgitant jet width and depth by color flow Doppler remains the most used. The jet diameter as a ratio of the left ventricular outflow tract diameter proximal to the valve represents one of the most reliable indices of severity and correlates well with angiographic assessment. Similarly, the vena contracta, which represents the smallest diameter of the regurgitant flow at the level of the valve, can be used to assess the severity of aortic regurgitation. Other Doppler-based methods include assessing the pressure half-time, or rate of decline of the pressure gradient between the aorta and left ventricle, a measure of acuity of aortic regurgitation, and assessing aortic flow reversal in the descending aorta. The regurgitant volume can be calculated by comparing the flow across the aortic and pulmonic valves, assuming the pulmonic valve is competent.

CMR offers several advantages over echocardiography in the assessment of aortic regurgitation. CMR is more accurate than echocardiography for assessing small changes in cardiac size or function longitudinally in patients with aortic insufficiency. In addition, CMR can accurately quantify aortic regurgitant volume secondary to aortic insufficiency better than echocardiography. CMR can also capture aortic size in 3D that may be helpful in determining the etiology of the aortic regurgitation or in monitoring progression of the condition **(Fig. 236-21 and Video 236-6)**.

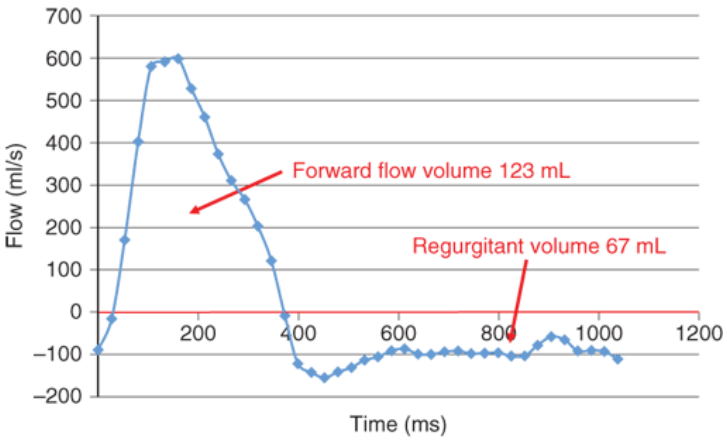
VIDEO 236-6: A patient with severe aortic regurgitation quantified by cardiac magnetic resonance (CMR).

Notice the dark flow jet during diastolic across the aortic valve. For quantitation of the aortic regurgitation severity, a cross-sectional cut was made just below the aortic valve, perpendicular to the aortic regurgitation jet, using phase contrast flow imaging. Apart from aortic regurgitation fraction and volume, CMR also can volumetrically quantify ventricular sizes and dimensions of the aorta, which are useful in monitoring patients with aortic valve diseases.

Image not available.
Play Video

FIGURE 236-21

The resultant flow curve generated from phase contrast imaging demonstrates a forward flow of 123 mL and a regurgitant volume of 67 mL, yielding a regurgitant fraction of 54% indicating severe aortic regurgitation.



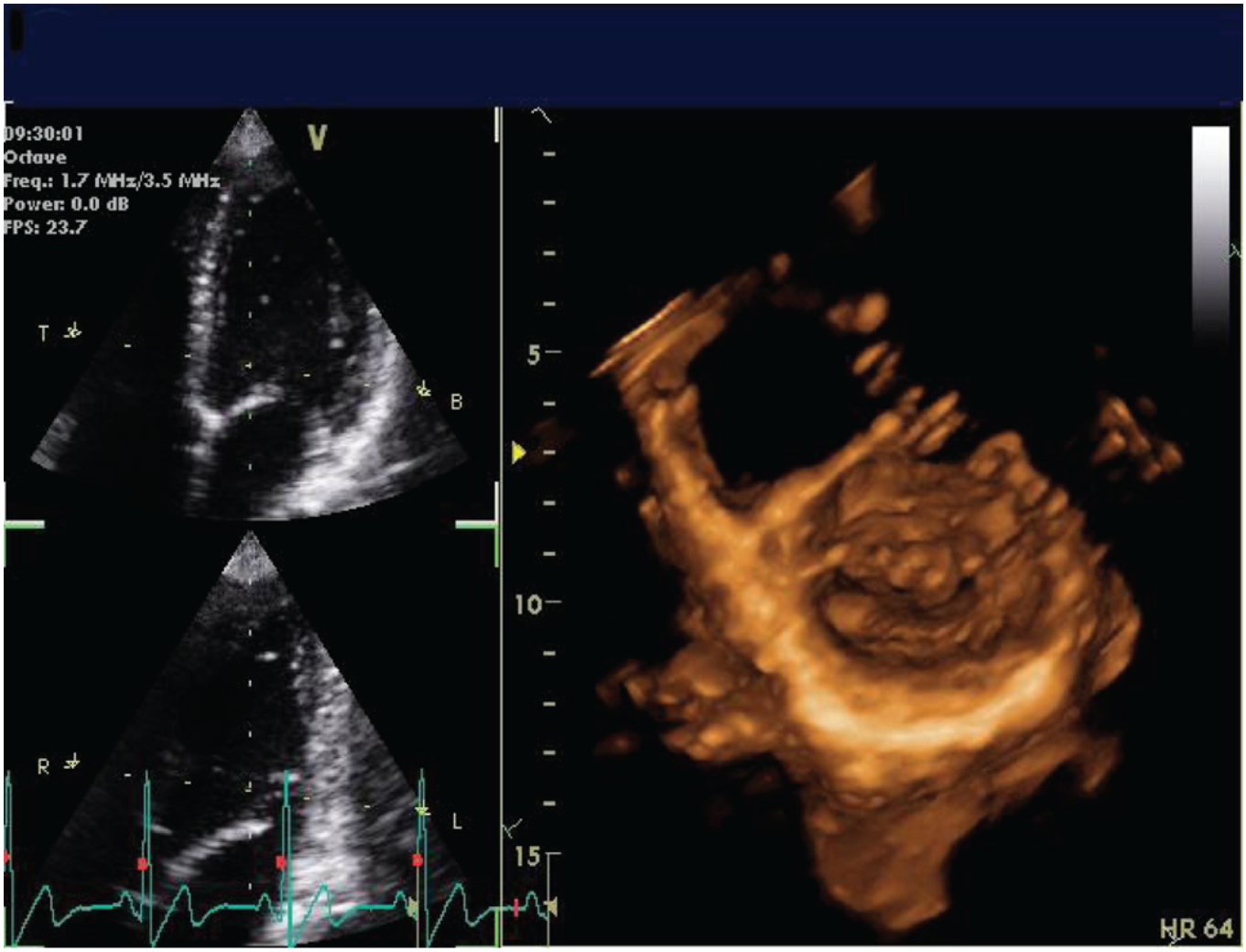
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Assessment of Mitral Regurgitation

The normal mitral valve consists of an anterior and posterior leaflet in a saddle shape configuration **(Fig. 236-22)**. The leaflets are attached to the papillary muscles via chordae tendineae that insert on the ventricular side of the leaflets. Mitral regurgitation can occur due to abnormalities of the leaflets, the chordal structures, or the ventricle, or any combination of these **(Fig. 236-23)**.

FIGURE 236-22

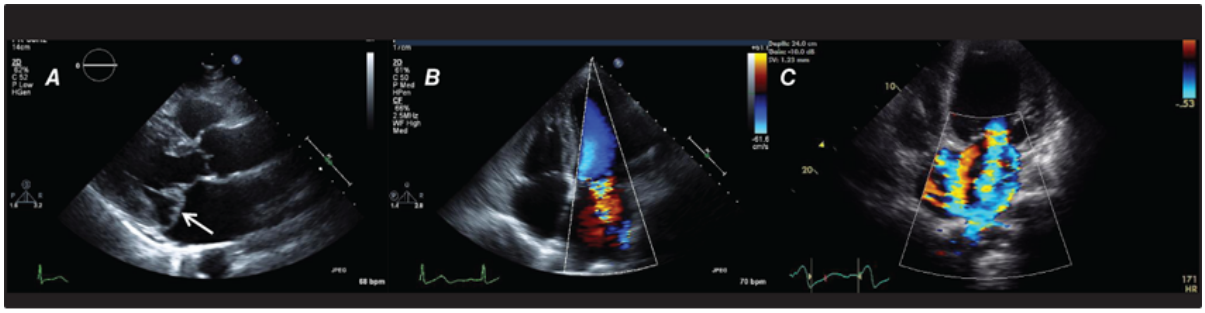
Normal mitral valve in two-dimensional views (*left*) and with three-dimensional imaging (*right*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 236-23

A. Mitral valve prolapse with posterior leaflet visualized prolapsing behind the plane of the anterior leaflet (*arrow*). ***B.*** Color flow Doppler showing mitral regurgitation in a patient with mitral valve prolapse. ***C.*** Severe functional mitral regurgitation in a patient with a dilated left ventricle.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Mitral valve prolapse, in which one leaflet moves behind the plane of the other leaflet, can be due to myxomatous degeneration of the valves and leaflet redundancy, disruption of chordal structures secondary to degenerative disease, or papillary muscle rupture or dysfunction following myocardial infarction. Regurgitant jets can be visualized using color flow Doppler. The velocity of regurgitant jets is driven by the pressure gradient between the two chambers. This velocity tends to be quite high for left-sided regurgitant lesions, including mitral regurgitation and aortic regurgitation, resulting in turbulent jets on color flow Doppler (Fig. 236-23). Visual estimation of color flow Doppler is generally sufficient for qualitative assessment of regurgitant severity but can dramatically under- or overestimate regurgitation severity, particularly when regurgitant jets are quite eccentric. For this reason, quantitative assessment is generally recommended, especially when making clinical decisions about surgical intervention. The proximal isovelocity surface area (PISA) method is generally used for quantitative assessment of severity of mitral regurgitation. This method relies on estimation of the velocity of flow acceleration at a specific distance proximal to the valve with the assumption that the flow accelerates in concentric hemispheres.

As with aortic insufficiency, assessment of ventricular structure and function is also integral in the evaluation of mitral regurgitation. Although some patients have mitral regurgitation due to intrinsic abnormalities of the valve itself, in others, the valve can be relatively normal but the mitral regurgitation can be secondary to dilatation and remodeling of the left ventricle. So-called functional mitral regurgitation is generally secondary to apical displacement of the papillary muscles in a dilated ventricle, resulting in the leaflets of the mitral valve being pulled toward the apex of the heart, resulting in poor coaptation during systole and resultant relatively central mitral regurgitation. This type of mitral regurgitation can generally be distinguished from intrinsic mitral valve disease, and the surgical or procedural treatment of these conditions can be different. Knowledge of the etiology of mitral regurgitation can be important for a surgeon planning mitral valve surgery. Moreover, new procedural approaches to mitral valve disease may be different depending on the etiology.

Ventricular dilatation is an important predictor of outcome in patients with mitral regurgitation of any cause. It is important to realize that in a patient with significant mitral regurgitation, a large portion of the blood being ejected from the left ventricle with every beat is regurgitant, thus artificially increasing the ejection fraction. Thus, an ejection fraction of 55% in a patient with severe mitral regurgitation may actually represent substantial reduction in myocardial systolic function.

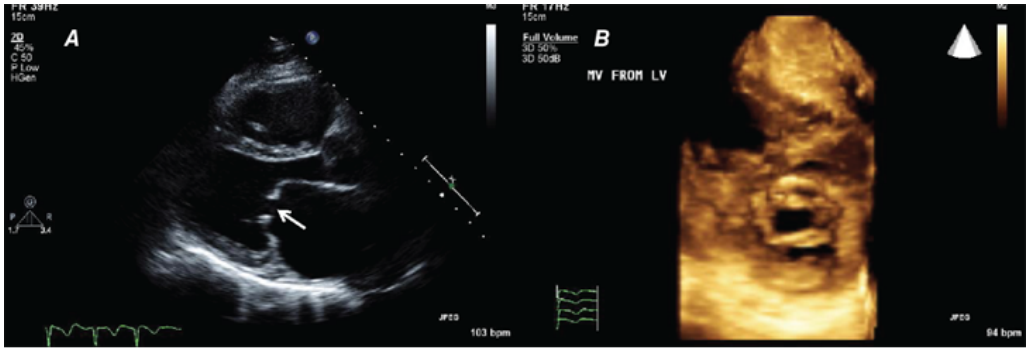
CMR can be helpful in evaluating mitral regurgitation in a subset of patients when echocardiographic assessment is inadequate. CMR can directly quantify regurgitant volume of the mitral regurgitant jet or indirectly quantify regurgitant volume by measuring the difference of left ventricular stroke volume and aortic forward flow.

Assessment of Mitral Stenosis

Rheumatic mitral disease remains the most common cause of mitral stenosis, although mitral stenosis can also result from severe calcification of the mitral leaflets. Rheumatic mitral stenosis has a distinct appearance characterized by tethering at the leaflet tips and relative pliability of the leaflets themselves, resulting in a hockey stick–type deformation particularly of the anterior leaflet **(Fig. 236-24).** Narrowing of the mitral orifice impedes flow from the left atrium to the left ventricle, resulting in increased pressures in the left atrium, which are then transmitted backward into the pulmonary vasculature and the right side of the heart. When mitral stenosis is suspected, echocardiography can be useful for determining etiology (specifically whether it is rheumatic or not), estimating the valve areas and gradients across the valve, assessing the left atrium, and assessing right ventricular size and function. Assessment of left atrial size and right ventricular size and function is particularly useful in helping determine the severity of the mitral stenosis.

FIGURE 236-24

A. Rheumatic mitral stenosis showing pliable leaflets tethered at the tips (*arrow*). Note the characteristically enlarged left atrium. ***B.*** Mitral stenosis visualized from a three-dimensional echocardiogram.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

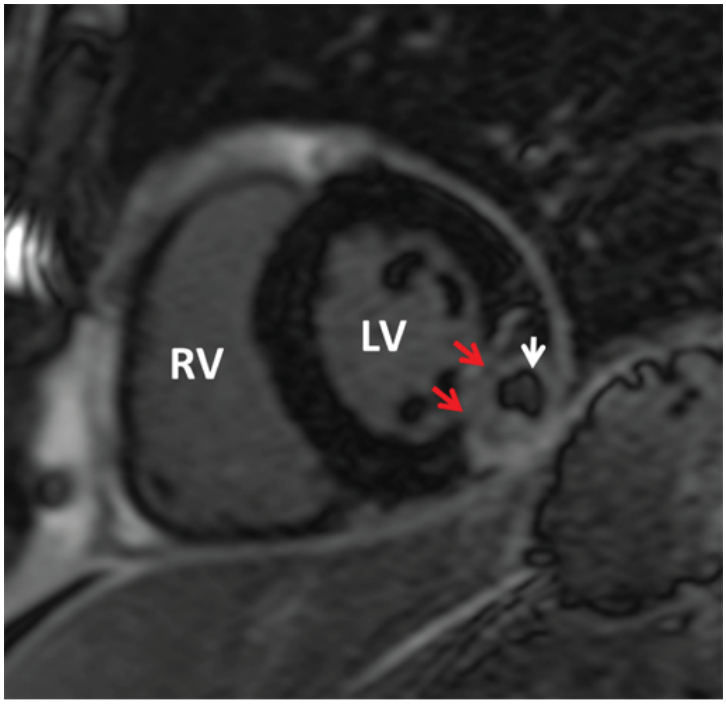
MYOCARDIAL INFARCTION AND HEART FAILURE

Role of Imaging after Myocardial Infarction

Imaging can be useful in the immediate and long-term follow-up of patients with myocardial infarction. As discussed earlier in the chapter, LGE imaging by CMR is the best technique for imaging for presence or the extent of infarcted myocardium. In a recent multicenter study, LGE imaging identified infarct location accurately and detected acute and chronic infarcts at a sensitivity of 99 and 94%, respectively. In addition, regions of microvascular obstruction (no-reflow) can be seen as dense hypoenhanced areas within the core of a bright region of infarction **(Fig. 236-25)**. Both the presence of LGE and microvascular obstruction are markers of increased clinical risk.

FIGURE 236-25

Example of a patient who presented with inferior ST segment elevation myocardial infarction after several days of intermittent chest pain. The MRI confirmed an inferior MI by the location of LGE (*red arrows*). In addition, there is a central area of microvascular obstruction (dark region surrounded by the bright LGE, *white arrow*). RV: right ventricle, LV: left ventricle.

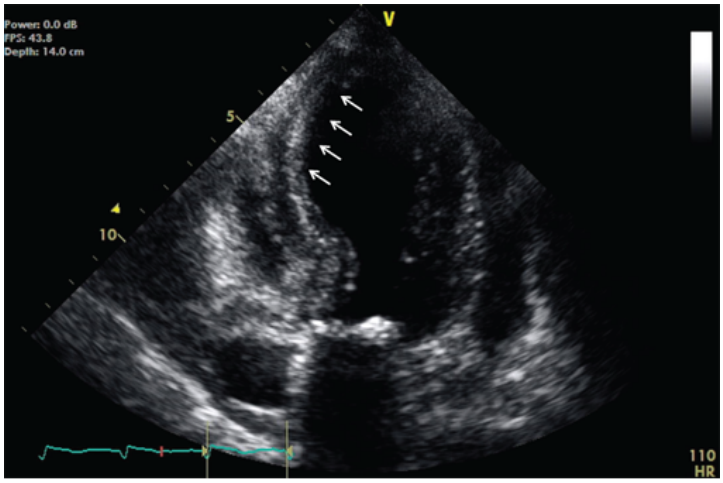


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

While echocardiography is often used to assess myocardial function immediately after myocardial infarction, myocardial stunning is common in the early post–myocardial infarction period, especially in patients who undergo reperfusion therapy. In these patients, either partial or complete recovery of ventricular function is common within several days, so that early estimation of ejection fraction may be misleading. In patients with uncomplicated myocardial infarction, imaging can generally be deferred for several days so that a more accurate assessment of cardiac function, including regional wall motion, can be assessed **(Fig. 236-26)**.

FIGURE 236-26

Acute left anterior descending artery distribution myocardial infarction at end systole showing akinetic region (*arrows*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

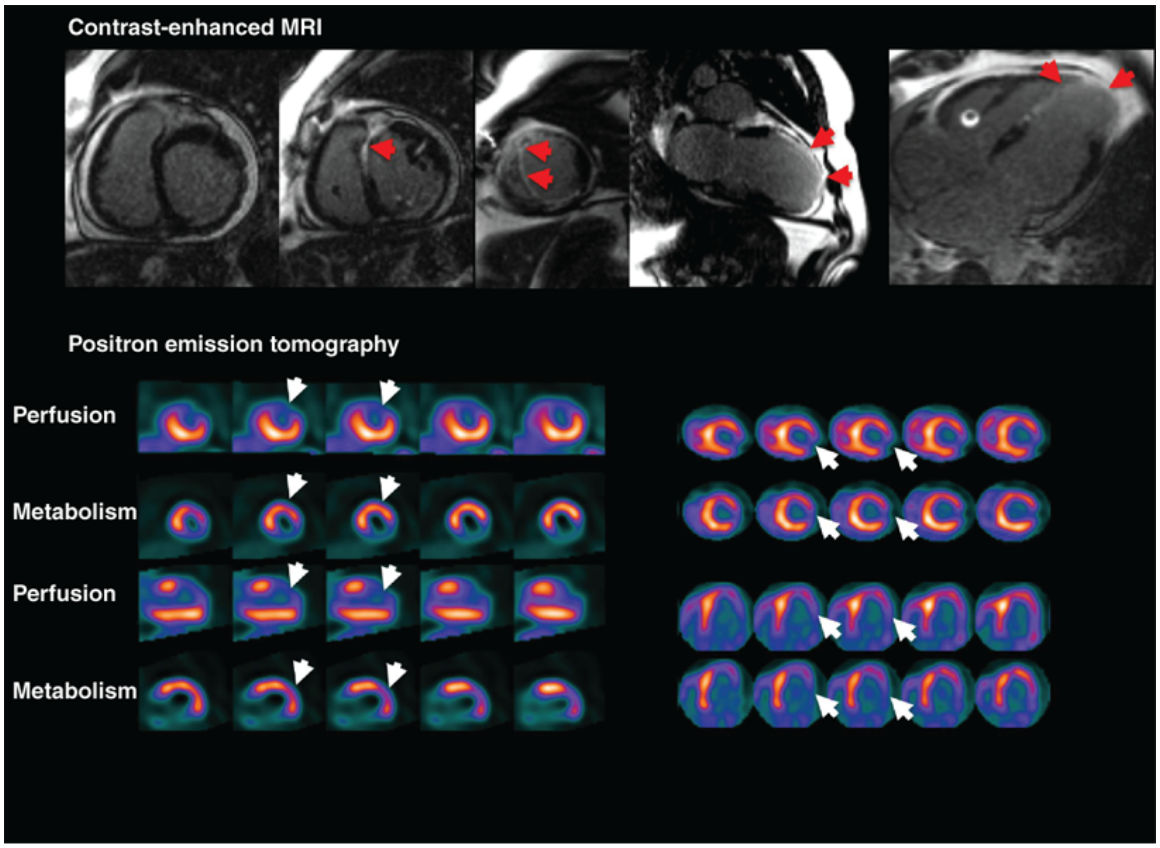
Echocardiography is the best method for assessment of patients with suspected mechanical complications after myocardial infarction. These include mitral regurgitation secondary to either papillary muscle dysfunction or rupture of papillary muscle head, ventricular septal defect, or even cardiac rupture. A new severe systolic murmur should raise suspicions for either severe mitral regurgitation or ventricular septal defect. While cardiac rupture is often catastrophic, contained ruptures, also known as pseudoaneurysms, can occur, and early diagnosis and surgical treatment are the best way to maximize survival. The presence of thrombus within the pericardial space following myocardial infarction should immediately raise suspicion of myocardial rupture and represents a surgical emergency.

Some patients demonstrate progressive left ventricular dilatation and dysfunction, known as cardiac remodeling, after myocardial infarction. Assessment of cardiac function and regional wall motion is useful in the follow-up period, generally between 1 and 6 months following infarction. The persistence of left ventricular systolic dysfunction following infarction is used to determine the type of therapy (e.g., angiotensin-converting enzyme inhibitors or angiotensin receptor blockers are typically used in patients with systolic dysfunction following myocardial infarction).

In patients with acute or subacute myocardial infarction, investigation of residual ischemia and/or viability is occasionally an important clinical question, especially among those with recurrent symptoms after myocardial infarction **(Fig. 236-27)**. All cardiac imaging techniques can provide information regarding myocardial viability and ischemia. The available data suggest that radionuclide imaging, especially PET, is highly sensitive, with higher negative predictive value than dobutamine echocardiography. In contrast, dobutamine echocardiography tends to be associated with higher specificity and positive predictive accuracy than the radionuclide imaging methods. The experience with CMR suggests that it offers similar predictive accuracies as those seen with dobutamine echocardiography.

FIGURE 236-27

Examples of myocardial viability patterns obtained with cardiac magnetic resonance imaging (MRI) and positron emission tomography (PET) in three different patients with coronary artery disease. The *top panel* demonstrates extensive late gadolinium enhancement (bright white areas) involving the anterior, anteroseptal, and apical left ventricular walls (*arrows*), consistent with myocardial scar and nonviable myocardium. The *lower left panel* demonstrates rubidium-82 myocardial perfusion and ^{18}F-fluorodeoxyglucose (FDG) images showing a large and severe perfusion defect in the anterior, anterolateral, and apical walls, indicating preserved glucose metabolism (so-called *perfusion-metabolic mismatch*) consistent with viable myocardium. The *right lower panel* shows similar PET images demonstrating concordant reduction in perfusion and metabolism (so-called *perfusion-metabolic match*) in the lateral wall, consistent with nonviable myocardium.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Role of Imaging in New-Onset Heart Failure

Echocardiography is usually a first-line test in patients presenting with new-onset heart failure. As discussed above, this test provides a direct assessment of ventricular function and can help distinguish patients with reduced from those with preserved ejection fraction. In addition, it provides additional structural information including an assessment of valves, myocardium, and pericardium.

Although coronary angiography is commonly performed in patients with reduced ejection fraction, the determination of heart failure etiology in an individual patient may be difficult even if angiographically obstructive CAD is present. Indeed, patients with heart failure and no angiographic CAD may have typical angina or regional wall motion abnormalities on noninvasive imaging, whereas patients with angiographically obstructive CAD may have no symptoms of angina or history of myocardial infarction. Thus, the appropriate classification for any given patient is not always clear, and it often requires the complementary information of coronary angiography and noninvasive imaging. Stress radionuclide imaging and echocardiography can be helpful in delineating the extent and severity of inducible myocardial ischemia and viability. Multiparametric CMR can be quite helpful in the differential diagnosis of heart failure etiologies. Apart from quantifying left and right ventricular volumes and function, CMR can provide information about myocardial ischemia and scar. The pattern of LGE helps differentiate infarction (typically starting in the subendocardium and involving a coronary territory) from other forms of infiltrative or inflammatory cardiomyopathies (typically involving the mid- or subepicardial layers without following a coronary distribution) **(Fig. 236-28)**. In addition, it can assess the presence of myocardial edema (e.g., myocarditis) and quantify myocardial iron deposition that can potentially lead to cardiac toxicity. Infiltrative cardiomyopathy such as amyloidosis typically has a restrictive cardiomyopathy pattern (bilateral atrial enlargement and biventricular increased wall thickness). CMR of patients with cardiac amyloidosis often also demonstrates a characteristic pattern of diffuse endocardial infiltration of the left ventricle and the atria (Fig. 236-28). Hypertrophic cardiomyopathy has variable degree of increased ventricular thickness, and often is seen to have outflow obstruction and intense LGE in regions with marked hypertrophy **(Fig. 236-29)**. CMR also can quantify myocardial iron content in patients at risk of iron-overload cardiomyopathy **(Video 236-7)**.

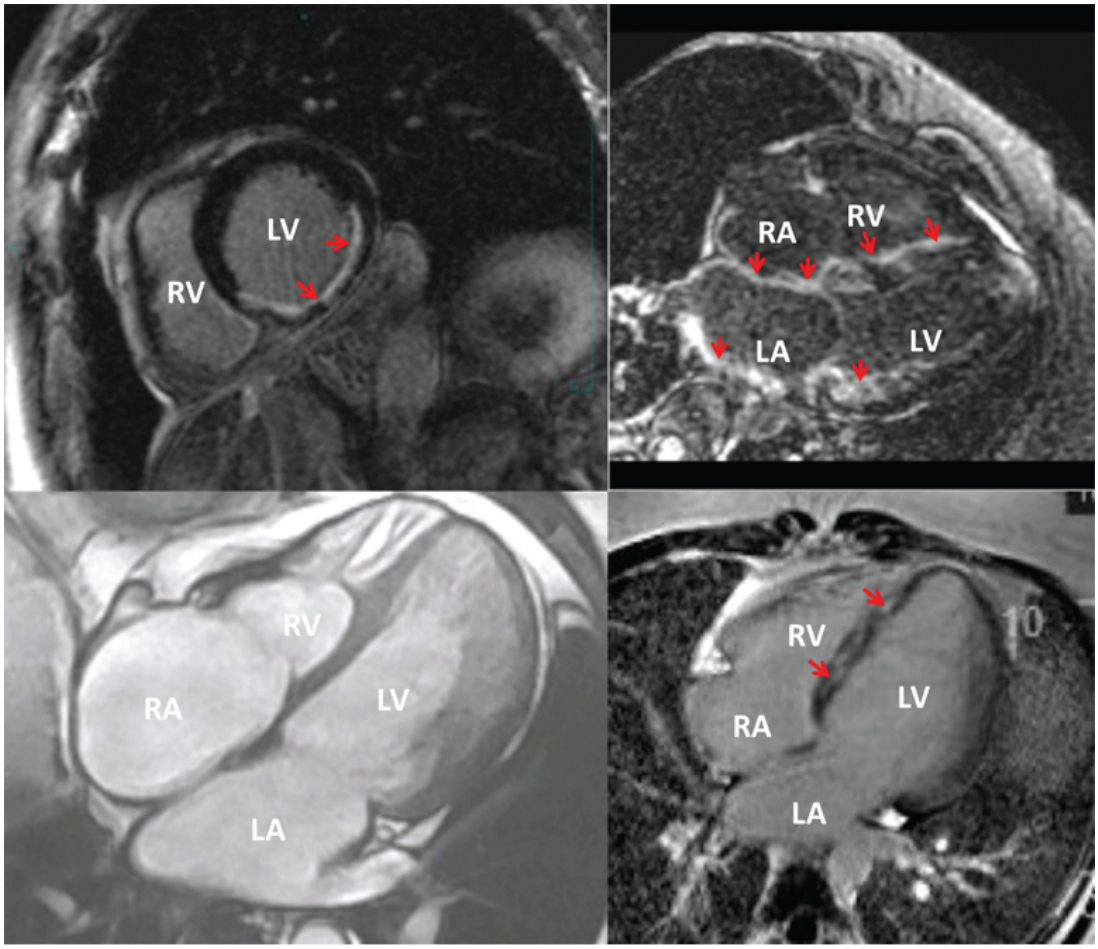
VIDEO 236-7: T2* images of the heart (left panel) and the liver (right panel) of a patient who has hemochromatosis.

Note that iron and the liver are markedly darkened in these movies, indicating high load of iron in the heart muscle and liver. The rate of signal reduction (decay) in the myocardium and liver can be calculated as T2* value expressed in milliseconds. In this case, the T2* was at 10 ms. T2* <20 ms in patients with cardiomyopathy has been shown to indicate iron toxicity as the etiology of the cardiomyopathy, and it carries prognostic valve for such patients at risk of cardiac iron toxicity.

Image not available.
Play Video

FIGURE 236-28

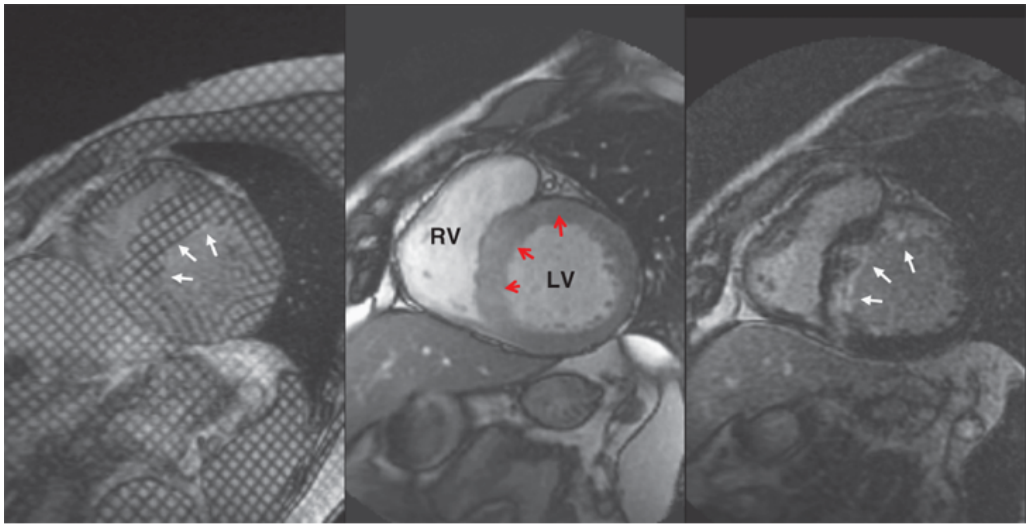
Differentiation of various cardiomyopathies by CMR. The left upper panel shows the short-axis late gadolinium enhancement (LGE) imaging of a patient who suffered an acute myocardial infarction. Note LGE of the endocardial myocardium in the inferior wall extending from the septum to the lateral wall associated with myocardial thinning (*arrows*). The right upper panel shows the long-axis LGE imaging of a patient who has cardiac amyloidosis. Note the diffuse LGE throughout left ventricular myocardium, the left atrium, and the interatrial septum (arrows). In addition, the blood pool is characteristically dark in signal indicating sequestration of gadolinium contrast out of the blood pool after injection due to a high burden of amyloidosis in other organs. The left lower panel shows a cine diastolic long-axis image of a patient with a non-ischemic dilated cardiomyopathy. Note that there is extensive sponge-like non-compacted myocardium of the LV as well as dilatation of all four cardiac chambers. This patient has a non-ischemic dilated cardiomyopathy secondary to LV non-compaction. The right lower panel shows a 22-year-old female patient with a recent episode of acute chest pain and troponins elevation. Note the multiple mid-wall foci of LGE which suggests acute myocarditis (*arrows*). LA, left atrium; LV, left ventricle; RA, right atrium; RV, right ventricle.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 236-29

This figure demonstrates three pulse sequence techniques by cardiac magnetic resonance that are often used to assess patients with hypertrophic cardiomyopathy, all displayed in the mid short-axis scan plane. The *center panel* demonstrates that the left ventricle (LV) was markedly thickened in its wall thickness especially in the LV septum (*red arrows*). This finding was matched by marked regions of late gadolinium enhancement (LGE), which was consistent with fibrosis in these segments (*right panel, white arrows*). The *left panel* was cine myocardial tagging in the same slice plane. Myocardial tagging is used to assess the normal intramyocardial strain by assessing distortion of the myocardial grids during systole. In this case, despite normal-appearing systolic radial wall thickening, the myocardial strain as assessed by the distortion of grids was markedly reduced (*left panel, white arrows*). This finding is consistent with substantial myofibril disarray in the anterior and anteroseptal segments in this patient. RV, right ventricle.

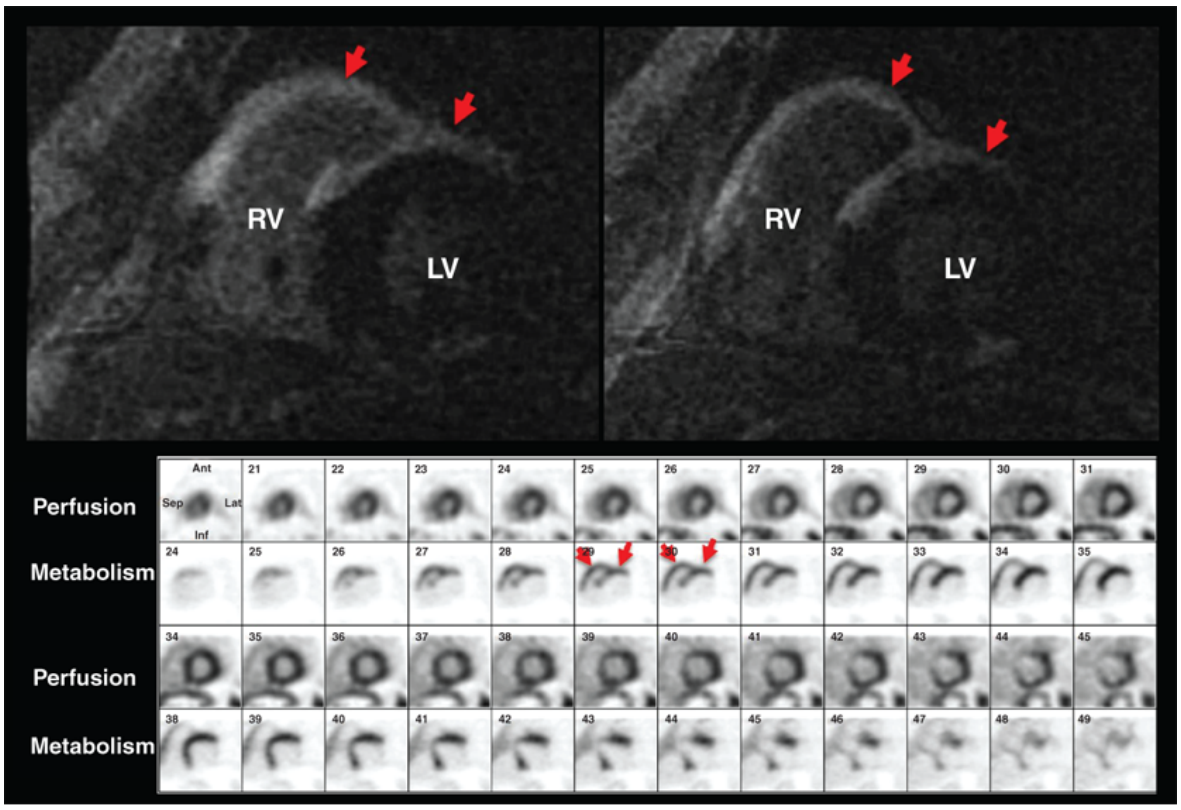


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PET metabolic imaging has a complementary role in the evaluation of inflammatory cardiomyopathies, especially sarcoidosis where the presence of focal and/or diffuse glucose uptake can help identify areas of active inflammation. In addition, for patients undergoing immunosuppressive therapy, PET is frequently used to monitor therapeutic response **(Fig. 236-30)**. In patients with ischemic cardiomyopathy, radionuclide imaging in general and PET in particular are frequently used to quantify the presence and extent of myocardial ischemia and viability to assist with clinical decision making related to myocardial revascularization (Fig. 236-26).

FIGURE 236-30

Representative cardiac magnetic resonance (CMR; *top panel*) and positron emission tomography (PET; *lower panel*) images from a 45-year-old male presenting with complete heart block. The CMR images demonstrate extensive late gadolinium enhancement in the subepicardial left ventricular (LV) anterior and anteroseptal walls and also in the right ventricular (RV) free wall (*arrows*). The PET images demonstrate extensive fluorodeoxyglucose uptake in the same areas, most consistent with active inflammation due to sarcoidosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ASSESSING CARDIAC FUNCTION IN PATIENTS UNDERGOING CANCER TREATMENT

Therapies used to treat cancer can adversely affect the cardiovascular system. As the efficacy of cancer treatment and survival improve, many patients are presenting with late adverse consequences from chemotherapy and/or radiation therapy on cardiovascular function. Thus, the morbidity and mortality from late cardiovascular complications threaten to offset the early gains in cancer survival, especially among children and young adults. Early recognition and treatment of cardiomyocyte injury are critical for successful application of preventative therapies, but difficult because the adverse effects on cardiac function are a relatively late manifestation after exposure to anticancer therapy.

The accepted standard for clinical diagnosis of cardiotoxicity is defined as a >5% reduction in LVEF to <55% with symptoms of heart failure, or a >10% drop in LVEF to <55% in patients who are asymptomatic. Thus, noninvasive imaging plays a major role in diagnosing and monitoring for cardiac toxicity in patients undergoing cancer treatment. Radionuclide angiography has been the technique of choice for quite some time. However, echocardiography now plays a major role in this application.

Recently, more novel imaging approaches have been advocated, including deformation imaging with echocardiography and fibrosis imaging with CMR. These techniques have shown promising results in experimental animal models and in humans. In addition, there are also proof-of-concept studies in animal models using molecular imaging approaches targeting the mechanisms of cardiac toxicity (e.g., apoptosis and oxidant stress), which can presumably provide the earliest signs of the off-target effects of these therapies. However, these techniques are currently considered experimental.

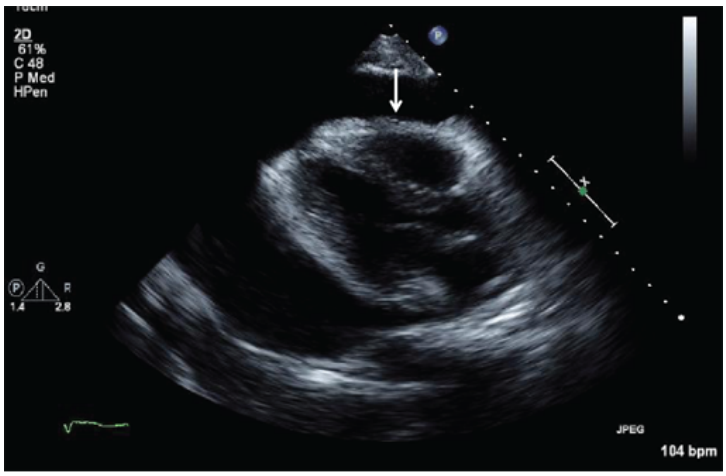
PERICARDIAL DISEASE

The fibroelastic pericardial sac surrounding the heart consists of a visceral, or epicardial, layer and a parietal layer, with a generally small amount of pericardial fluid in between layers. The pericardium is generally quite pliable and moves easily with the heart during contraction and relaxation. Abnormalities of the pericardium can affect cardiac function primarily by impairing the heart's ability to fill. Inflammation of the pericardium can lead to an accumulation of fluid between the two layers, or *pericardial effusion*, which can be visualized by echocardiography, CMR, or CT. Other reasons for accumulation of pericardial fluid include infection, malignancy, and bleeding into the pericardium. The latter can be the result of catastrophic processes such as trauma, cardiac rupture, perforation in the setting of a cardiac procedure, cardiac surgery, or dissection of the aorta with extension in the pericardium.

Echocardiography remains the initial test of choice for assessing pericardial disease, especially effusions **(Fig. 236-31)**. Moreover, echocardiography can be useful in evaluating for pericardial constrictive physiology, in which a thick noncompliant pericardium impairs cardiac filling. The location, size, and physiologic consequences of accumulated pericardial effusion can generally easily be determined by echocardiography. Pericardial tamponade occurs when enough pericardial fluid accumulates so that the intrapericardial pressure exceeds filling pressures of the heart, generally the right ventricle. The balance between intrapericardial pressure and ventricular pressure is more important than the extent of fluid accumulation. Conditions in which pericardial effusions accumulate over a long period of time, as can be the case in the setting of malignant effusions, can lead to large pericardial fluid accumulations without the classic hemodynamic findings associated with pericardial tamponade. In contrast, rapid accumulations of pericardial fluid, such as those that occur due to cardiac rupture or perforation, can lead to tamponade physiology without very large effusions. In patients with suspected pericardial effusion or tamponade, echocardiography can usually be performed rapidly, at the bedside, and even by operators with limited skill. The distance from the parietal to the visceral pericardial layer can be measured, and when this exceeds ~1 cm, an effusion is considered significant. Echocardiographic features suggestive of tamponade include diastolic collapse of the right ventricular free wall, suggestive of pericardial pressures that exceed right ventricular filling pressures, and Doppler evidence of respiratory flow variation, which is the Doppler equivalent of pulsus paradoxus. Despite the benefits of echocardiography in suspected pericardial tamponade, the diagnosis of tamponade remains a clinical diagnosis, and other important features, such as patient's blood pressure in the presence of pulsus paradoxus, needs to be taken into account when considering therapeutic options.

FIGURE 236-31

Pericardial effusion with tamponade physiology. The right ventricle (*arrow*) is small and collapsing in end diastole due to increased pericardial pressure.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Chronic inflammation of the pericardium leads to thickening and calcification of the parietal pericardium, resulting in pericardial constriction in which diastolic filling can be severely impaired. In these cases, filling of the ventricles comes to an abrupt halt when the volume of ventricular filling is impaired by the constricting pericardium. Assessment of pericardial thickness in these patients is important, but it is just as important to note that approximately one in five patients with severe pericardial constriction have no significant pericardial thickening by imaging or at surgery. Thus, a lack of thickened pericardium does not rule out pericardial constriction, and patients' signs and symptomatology and physiologic evidence of constriction should be assessed independently. Pericardial constriction typically demonstrates marked respiratory changes in diastolic flow on

Doppler echocardiography, in contrast to restrictive cardiomyopathy, but substantial overlap exists. CT and CMR offer tomographic, whole-heart assessment of pericardial thickening and other anatomy abnormalities in pericardial constriction (enlarged atria, vena cavas, pleural and pericardial effusions) **(Fig. 236-32 and Video 236-8)**. CMR offers the additional information of pericardial fibrosis and inflammation by LGE imaging and evidence of constrictive physiology (e.g., regional relaxation concordance due to myocardial adhesions, abnormal septal bounce with Valsalva maneuver).

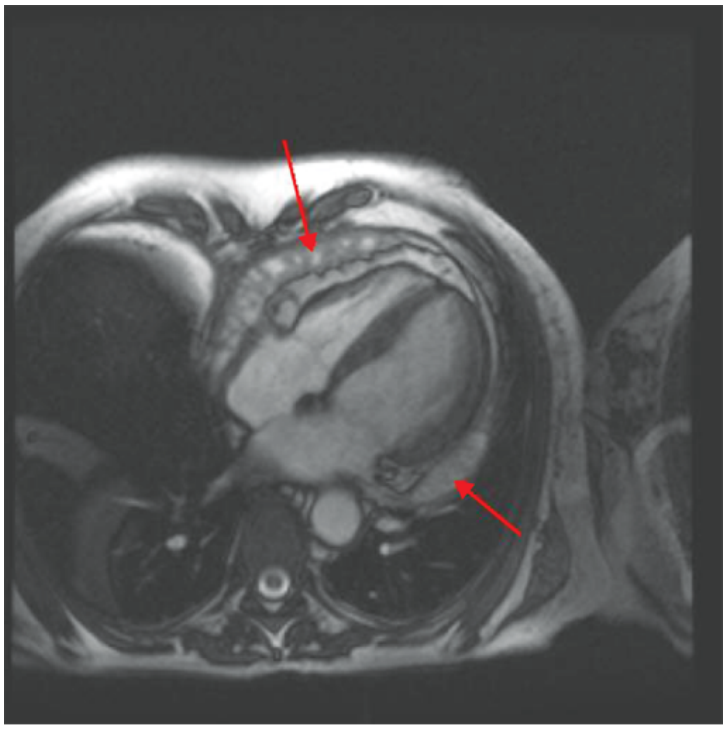
VIDEO 236-8: The heart in long and short axis.

Note the large atria, thickened pericardium, and extensive pericardial adhesions. Given the extensive pericardial adhesions, there is little shearing motion of the ventricles against the parietal pericardium.

Image not available.
Play Video

FIGURE 236-32

A female patient developed pericardial constriction and right heart failure, secondary to radiation therapy for breast cancer. Note the multiple pericardial adhesions (*red arrows*).



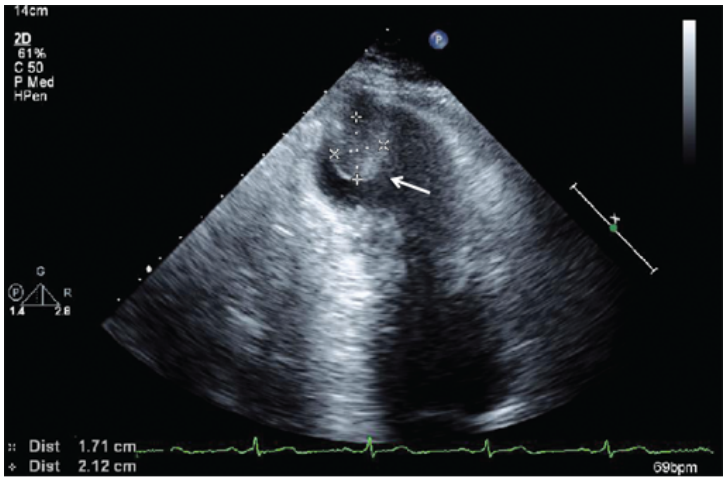
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CARDIAC THROMBUS AND MASS

Echocardiography is usually the modality that first detects a cardiac mass with differential diagnoses including thrombus, tumor, or vegetation. Given their unrestricted tomographic views and multiplanar three-dimensional imaging, CMR and CT can complement echocardiography by further characterizing the physical features of the cardiac mass. Compared to CT, CMR has the advantage of higher tissue contrast differentiation, more robust cine imaging, and the use of multifaceted techniques within the same imaging session to determine the physiologic characteristics of the mass. Gadolinium contrast enhancement patterns of increased capillary perfusion can detect vascularity within a mass which differentiates a tumor from a thrombus. Structures that are known to mimic a cardiac mass include (1) anatomic variants, such as the Eustachian valve, Chiari network, crista sagittalis or terminalis, and the right ventricular moderator band, and (2) “pseudotumors,” such as interatrial septal aneurysm, coronary or aortic aneurysm, lipomatous hypertrophy of interatrial septum, hiatal hernia, or a catheter/pacemaker lead. Coexisting abnormalities that raise the likelihood of a cardiac thrombus **(Fig. 236-33)** include regional wall motion abnormality from an infarction or ventricular aneurysm, atrial fibrillation leading to slow flow in the left atrial appendage, or presence of venous catheters or recent endovascular injury. CMR has the advantage of being able to assess regional wall motion and infarction or ventricular aneurysm in matching scan planes, adjacent to the cardiac thrombus, using cine and LGE imaging, respectively. For ventricular thrombus, gadolinium-enhanced LGE imaging can detect thrombus at a higher sensitivity than echocardiography by depicting high-contrast difference between the dark thrombus and its adjacent structures and by imaging in three dimensions. In addition, mural thrombus does not enhance on first-pass perfusion and often has a characteristic “etched” appearance (black border surrounding a bright center) on LGE imaging, thus providing higher diagnostic specificity than anatomic information alone **(Fig. 236-34)**. Comparing the signal intensities of a mass before and after contrast injection may confirm the lack of tissue vascularity (i.e., thrombus) by the lack of signal enhancement after contrast administration. Like intracardiac thrombus, regions of microvascular obstruction also appear dark, but microvascular obstruction is confined within the myocardium and surrounded by infarction and thus can be differentiated from intracardiac thrombus. Cardiac CT imaging is ideally suited for small thrombus in the left atrial appendage especially in cases where transesophageal echocardiography is suboptimal or not feasible.

FIGURE 236-33

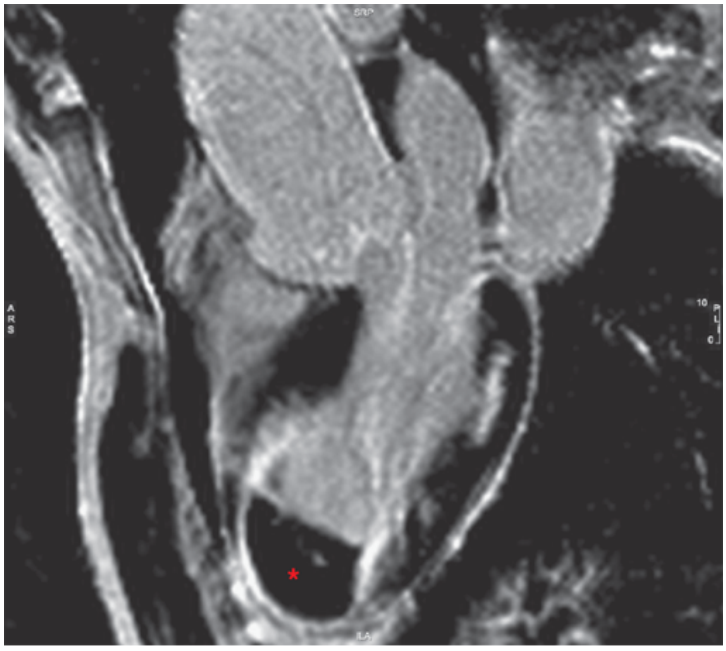
Cardiac thrombus (*arrow*) in an apical aneurysmal region following acute myocardial infarction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 236-34

Late gadolinium enhancement image of a massive anterior infarction complicated by a dyskinetic left ventricular LV aneurysm and intracavitary thrombus (*red asterisk*).

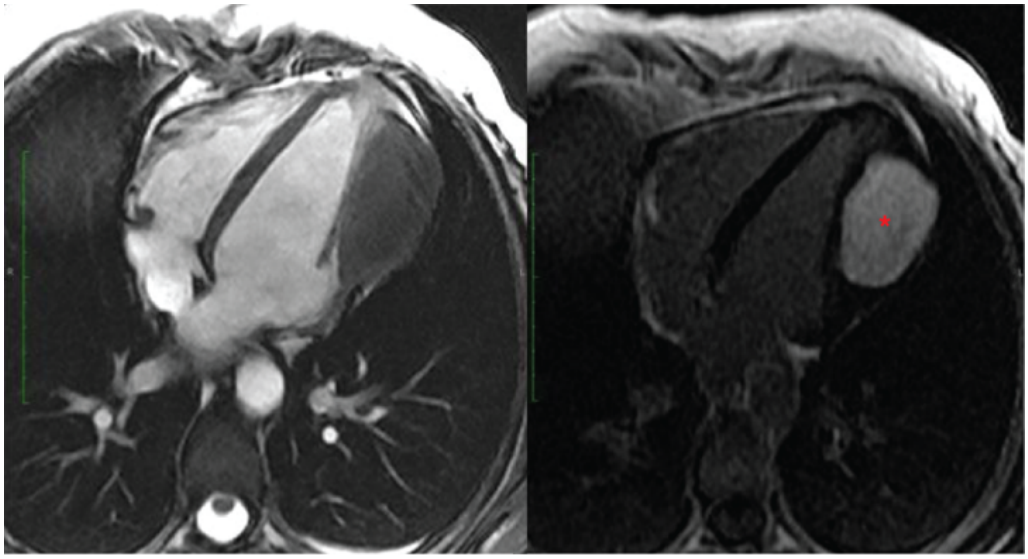


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The majority of cardiac malignancy is metastatic which is about twentyfold more common than primary cardiac malignancies. Metastasis to the heart can be the result of direct invasion (e.g., lung and breast), lymphatic spread (e.g., lymphomas and melanomas), or hematogenous spread (e.g., renal cell carcinoma). Primary benign cardiac tumors are seen mostly in children and young adults and include atrial myxoma, rhabdomyoma, fibroma, and endocardial fibroelastoma **(Fig. 236-35)**. Atrial myxomas are often seen as a round or multilobar mass in the left atrium (75%), right atrium (20%), or ventricles or mixed chambers (5%). They typically have inhomogeneous brightness in the center on cine steady-state free precession imaging due to their gelatinous contents and may have a pedunculated attachment to the fossa ovalis. Primary malignant cardiac tumors are rare and may include angiosarcoma, fibrosarcoma, rhabdomyosarcoma, and liposarcoma.

FIGURE 236-35

A case of a cardiac fibroma. A patient presented with shortness of breath and was found to have a large myocardial mass on echocardiography. Cine cardiac magnetic resonance imaging confirmed the large myocardial mass involving the anterolateral wall. Shortly after gadolinium contrast was injected, the myocardial mass demonstrated intense accumulation of contrast on LGE imaging (*right panel, asterisk*). This is a case of cardiac fibroma. The patient also has gingival hyperplasia and bifid thoracic ribs, a part of the rare Gorlin's syndrome.



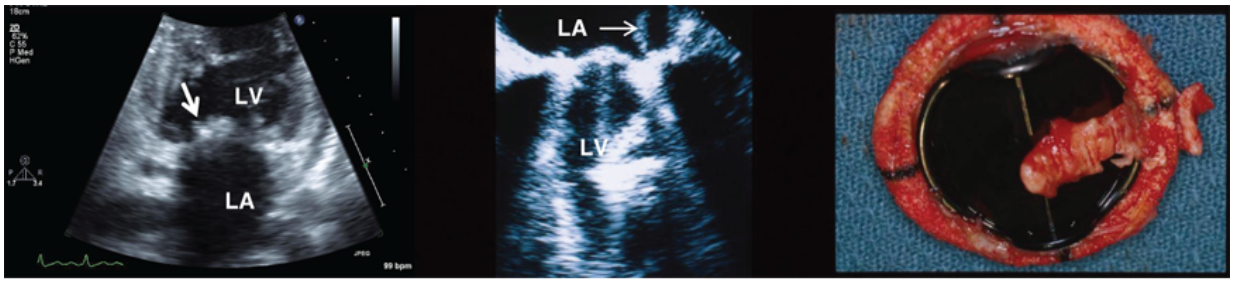
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ROLE OF IMAGING IN INFECTIOUS AND INFLAMMATORY DISEASE

Patients with suspected endocarditis often undergo echocardiography for the purpose of identifying vegetations or intramyocardial abscesses. Vegetations are generally highly mobile structures that most typically are attached to valves or present in areas of the heart with turbulent flow. The absence of a vegetation on echocardiography does not rule out endocarditis, because small vegetations below the resolution of the imaging techniques can be present. Echocardiography remains the best technique for assessment of vegetations because its high temporal resolution allows visualization of the typical oscillating motion, although large vegetations can be visualized with other techniques **(Fig. 236-36)**. The size and location of a vegetation do not necessarily provide any specific information about the type of infection. Abscesses, particularly around the aortic and mitral annuli, are particularly concerning in patients with endocarditis and should be suspected in patients with prolongation of cardiac intervals in the setting of endocarditis. Visualization of both vegetations and possible abscesses is best done with transesophageal echocardiography, particularly in patients with prosthetic valves. Indeed, transesophageal echocardiography is the first test of choice in a patient with a mechanical mitral or aortic valve and suspected endocarditis (Fig. 236-36). Vegetations should be measured because their size has prognostic importance and can be used to decide whether a patient should be taken to surgery.

FIGURE 236-36

Vegetation on native mitral valve (*left panel, arrow*). Left atrium (LA) and left ventricle (LV) are indicated. *Middle panel* shows a vegetation on a mechanical prosthesis (St. Jude) indicated by an *arrow*; *right panel* shows vegetation on prosthesis after excision.

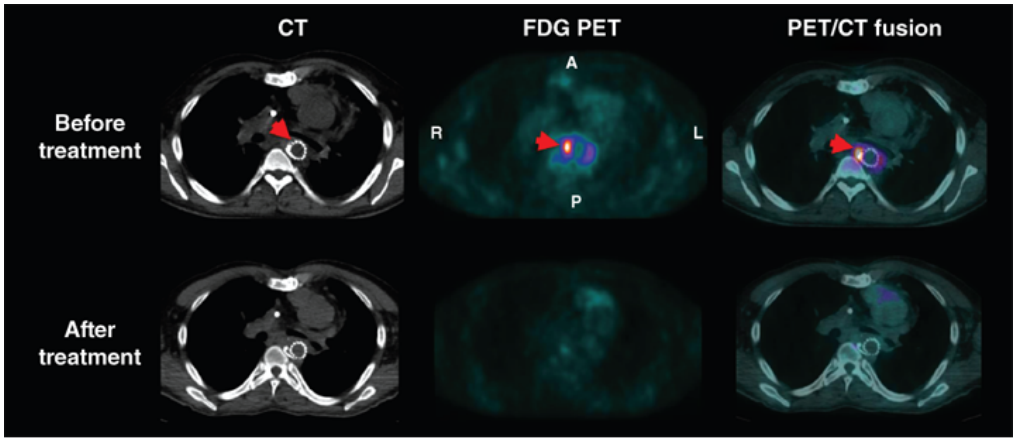


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PET metabolic imaging is emerging as a potentially useful imaging technique to identify the source of infection in patients with prosthetic valves, vascular grafts, and implantable pacemakers/defibrillators, especially in patients in whom echocardiography and/or blood cultures are negative. There is an emerging literature documenting the potential value of macrophage-targeted metabolic imaging with ^{18}F-FDG and PET **(Fig. 236-37)**. Likewise, FDG PET is also useful to identify vascular inflammation and monitor the response to immunosuppressive therapy **(Fig. 236-38)**.

FIGURE 236-37

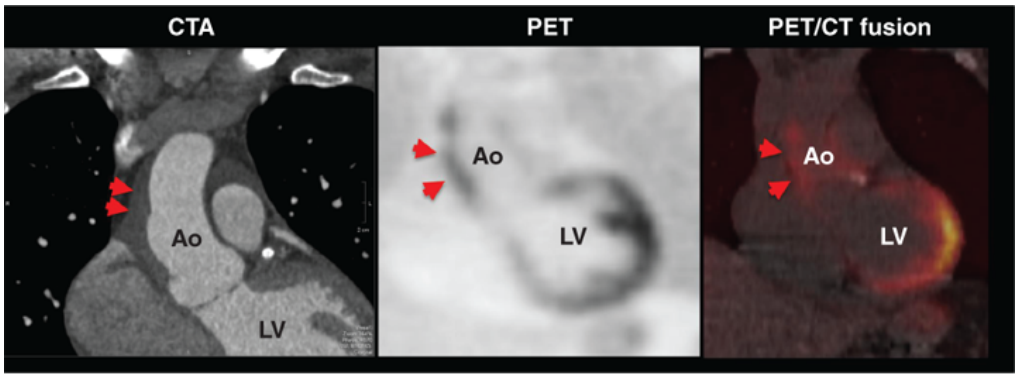
Representative cross-sectional computed tomography (CT; *left*), fluorodeoxyglucose (FDG) positron emission tomography (PET; *middle*), and fused CT and PET (*right*) images before and after antibiotic treatment in a patient with fever and suspected infection of the stent placed in the descending portion of the aortic arch (*arrow*) for treatment of aortic coarctation. The FDG images before treatment demonstrate intense glucose uptake within the stent, consistent with inflammation/infection. The *lower panel* demonstrates significant attenuation of the FDG signal after treatment. *(Images used with permission from Dr. Sharmila Dorbala, Brigham and Women's Hospital.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 236-38

Representative coronal computed tomography (CT) angiographic (CTA; *left panel*), fluorodeoxyglucose (FDG) positron emission tomography (PET; *middle panel*), and fused CT and PET (*right panel*) images in a patient with suspected aortitis. The CTA images demonstrate thickening of the ascending aorta (Ao), which correlates with intense, focal FDG uptake consistent with active inflammation. LV, left ventricle.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

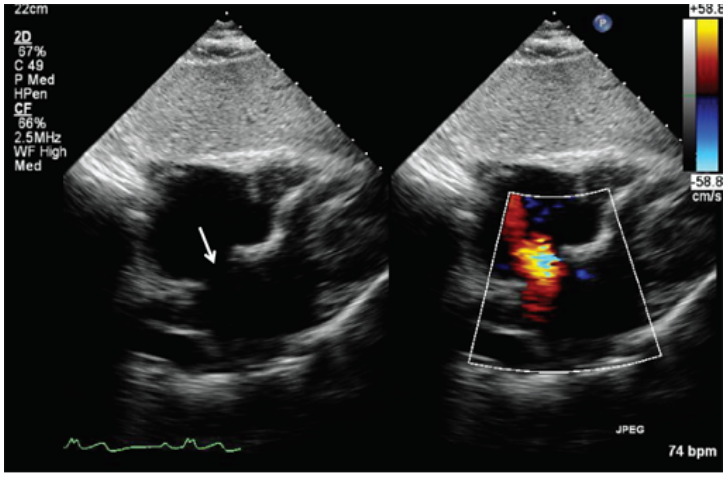
EVALUATION OF COMMON CONGENITAL ABNORMALITIES IN THE ADULT

While a discussion of complex congenital heart disease is beyond the scope of this chapter, several common congenital abnormalities are present in adults, and cardiac imaging is essential to diagnosing and managing these conditions. Abnormalities of the interatrial septum probably represent the most common adult congenital cardiac abnormalities. Patent foramen ovale (PFO) can be identified in almost 25% of patients. In patients with PFO, a one-way flap in the region of the fossa ovalis is normally kept close by the left atrial pressure, which is generally higher than right atrial pressure for much of the cardiac cycle. However, right-to-left flow through a PFO can occur any time the right atrial pressure exceeds the left atrial pressure, including with maneuvers or conditions in which intrathoracic pressure is increased. The presence of a PFO can increase the likelihood of the paradoxical embolus, and thus the presence of a PFO should be determined in patients with stroke or systemic embolus of unknown etiology. Because the one-way flap of the PFO will be closed during much of the cardiac cycle, color flow Doppler will usually not reveal a PFO. Instead, agitated saline (bubble study) is the best way to assess for PFO or atrial septal defect. Saline is agitated and injected peripherally and then enters the right atrium. If no shunt is present, only the right side of the heart will be pacified because the air bubbles will be too small to traverse the lungs. Because PFO is a one-way flap, maneuvers should be used to temporarily increase right atrial pressure. Either a Valsalva maneuver or sniff maneuver can be effective.

Atrial septal defects occur most commonly in the region of the fossa ovalis, referred to as secundum-type defects **(Fig. 236-39)**. Additional atrial septal defects include defects of the sinus venosus and atrium primum. Color flow Doppler echocardiography is usually sufficient for diagnosis of a secundum-type atrial septal defect, but agitated saline is generally needed for the diagnosis of other types of atrial septal defects.

FIGURE 236-39

Large secundum-type atrial septal defect (*arrow*) noted in the subcostal view with color flow Doppler showing flow through the defect (*right*).



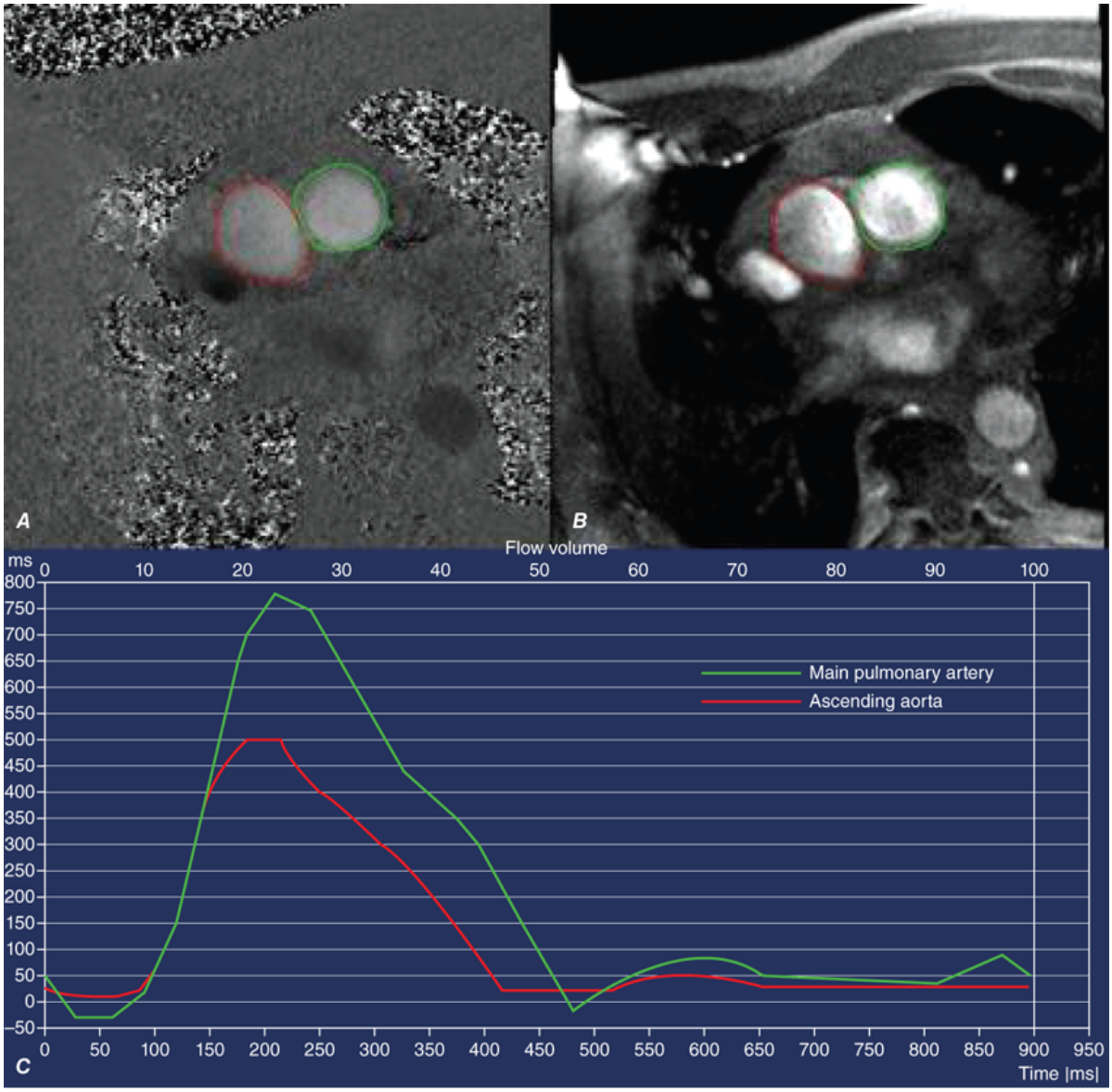
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ventricular septal defects can generally be visualized by color flow Doppler as turbulent high-velocity jets from the left to the right ventricle. In cases where the jet origin is unclear, continuous wave Doppler can estimate the velocities. These would be expected to be extremely high to reflect the pressure gradient between the left and right ventricles. Defects can occur in both the muscular and membranous portions of the ventricular septum.

In patients with either atrial or ventricular septal defects, estimation of the severity of the left-to-right shunt is essential and can be an important determinant in management decisions. Shunts are generally assessed by echocardiography by assessing the relationship between pulmonary flow and aortic flow, the Qp/Qs ratio. Shunts and cardiac anatomy of most congenital heart diseases can also be accurately evaluated by CMR **(Fig. 236-40)**.

FIGURE 236-40

A and ***B*** are phase contrast images that display blood flow (phase images on ***A***) and anatomy (structural images on ***B***) of the aorta (*red*) and pulmonary artery (*green*). ***C*** demonstrates the flow curves of the aorta (*red*) and the pulmonary artery (*green*). Note that the total flow (area under the curve) was substantially higher in the pulmonary artery than the aorta, indicative of a marked elevated pulmonary-to-systemic shunt ratio, as a result of the partial anomalous pulmonary venous return that drained into the superior vena cava.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

DI Carli MF et al: The future of cardiovascular imaging. Circulation 133:2640, 2016. [PubMed: 27324360]
Johnson NP et al: Invasive FFR and noninvasive CFR in the evaluation of ischemia: What is the future? J Am Coll Cardiol 67:2772, 2016. [PubMed: 27282899]
Naoum C et al: Cardiac computed tomography and magnetic resonance imaging in the evaluation of mitral and tricuspid valve disease: Implications for transcatheter interventions. Circ Cardiovasc Imaging 10:pii: e005331, 2017.
Solomon SD et al: *Essential Echocardiography, a Companion to Braunwald' Disease*, Elsevier, 2018.
Steel KE, Kwong RY: Application of cardiac magnetic resonance imaging in cardiomyopathy. Curr Heart Fail Rep 5:128, 2008. [PubMed: 18752762]
Vandoorne K, Nahrendorf M: Multiparametric imaging of organ system interfaces. Circ Cardiovasc Imaging 10:pii: e005613, 2017.

Chapter 237: Diagnostic Cardiac Catheterization and Coronary Angiography

Jane A. Leopold; David P. Faxon

INTRODUCTION

Diagnostic cardiac catheterization and coronary angiography are considered the gold standard in the assessment of the anatomy and physiology of the heart and its associated vasculature. In 1929, Forssmann demonstrated the feasibility of cardiac catheterization in humans when he passed a urological catheter from a vein in his arm to his right atrium and documented the catheter's position in the heart by x-ray. In the 1940s, Cournand and Richards applied this technique to patients with cardiovascular disease to evaluate cardiac function. These three physicians were awarded the Nobel Prize in 1956. In 1958, Sones inadvertently performed the first selective coronary angiography when a catheter in the left ventricle slipped back across the aortic valve, engaged the right coronary artery, and power-injected 40 mL of contrast down the vessel. The resulting angiogram provided superb anatomic detail of the artery, and the patient suffered no adverse effects. Sones went on to develop selective coronary catheters, which were modified further by Judkins, who developed preformed catheters and allowed coronary artery angiography to gain widespread use as a diagnostic tool. In the United States, cardiac catheterization is the second most common operative procedure, with more than 1.5 million procedures performed annually.

CARDIAC CATHETERIZATION

INDICATIONS, RISKS, AND PREPROCEDURE MANAGEMENT

Cardiac catheterization and coronary angiography are indicated to evaluate the extent and severity of cardiac disease in symptomatic patients and to determine if medical, surgical, or catheter-based interventions are warranted **(Table 237-1)**. They are also used to exclude severe disease in symptomatic patients with equivocal findings on noninvasive studies and in patients with chest-pain syndromes of unclear etiology for whom a definitive diagnosis is necessary for management. Cardiac catheterization is not mandatory prior to cardiac surgery in some younger patients who have congenital or valvular heart disease that is well defined by noninvasive imaging, and who do not have symptoms or risk factors that suggest concomitant coronary artery disease.

TABLE 237-1 Indications for Cardiac Catheterization and Coronary Angiography

CORONARY ARTERY DISEASE

Asymptomatic or Symptomatic

High risk for adverse outcome based on noninvasive testing

Sudden cardiac death

Sustained (>30 s) monomorphic ventricular tachycardia

Nonsustained (<30 s) polymorphic ventricular tachycardia

Symptomatic

Canadian Cardiology Society Class II, III, or IV stable angina on medical therapy

Acute coronary syndrome (unstable angina and non-ST-segment elevation myocardial infarction)

Chest-pain syndrome of unclear etiology and equivocal findings on noninvasive tests

ST-Segment Elevation Acute Myocardial Infarction

Reperfusion with primary percutaneous coronary intervention

Persistent or recurrent ischemia

Pulmonary edema and/or reduced ejection fraction

Cardiogenic shock or hemodynamic instability

Risk stratification or positive stress test after acute myocardial infarction

Mechanical complications—mitral regurgitation, ventricular septal defect

Valvular Heart Disease

Suspected severe valve disease in symptomatic patients—dyspnea, angina, heart failure, syncope

Infective endocarditis with need for cardiac surgery

Asymptomatic patients with aortic regurgitation and cardiac enlargement or ↓ ejection fraction

Prior to cardiac surgery in patients with suspected coronary artery disease

Congestive Heart Failure

New onset with angina or suspected undiagnosed coronary artery disease

New-onset cardiomyopathy of uncertain cause or suspected to be due to coronary artery disease

Congenital Heart Disease

Prior to surgical correction, when symptoms or noninvasive testing suggests coronary disease

Suspicion for congenital coronary anomalies

Pericardial Disease

Symptomatic patients with suspected cardiac tamponade or constrictive pericarditis

Cardiac Transplantation

Preoperative and postsurgical evaluation

Other Conditions

Hypertrophic cardiomyopathy with angina

Diseases of the aorta when knowledge of coronary artery involvement is necessary for management

The risks associated with elective cardiac catheterization are relatively low, with a reported risk of <0.1% for myocardial infarction, 0.01% for stroke, and 0.1% for death. These risks increase substantially if the catheterization is performed emergently, during acute myocardial infarction or in hemodynamically unstable patients. Additional risks of the procedure include tachy- or bradyarrhythmias that require countershock or pharmacologic therapy, acute renal failure leading to transient or permanent dialysis, vascular complications that necessitate surgical repair or percutaneous intervention, and significant access-site bleeding. Of these risks, vascular access-site bleeding is the most common complication, occurring in 1.5–2.0% of patients, with major bleeding events associated with a worse short- and long-term outcome.

In patients who understand and accept the risks associated with cardiac catheterization, there are no absolute contraindications when the procedure is performed in anticipation of a life-saving intervention. Relative contraindications do, however, exist; these include decompensated congestive heart failure; acute renal failure; severe chronic renal insufficiency, unless dialysis is planned; bacteremia; acute stroke; active gastrointestinal bleeding; excessive anticoagulation or recent lytic administration, severe, uncorrected electrolyte abnormalities; a history of an anaphylactic/anaphylactoid reaction to iodinated contrast agents; and a history of allergy/anaphylaxis/bronchospasm to aspirin in patients for whom progression to a percutaneous coronary intervention is likely and aspirin desensitization has not been performed.

Contrast allergy and contrast-induced acute kidney injury merit further consideration, because these adverse events may occur in otherwise healthy individuals and prophylactic measures exist to reduce risk. Allergic reactions to contrast agents occur in <5% of cases, with severe anaphylactoid (clinically indistinguishable from anaphylaxis, but not mediated by an IgE mechanism) reactions occurring in 0.1–0.2% of patients. Mild reactions manifest as nausea, vomiting, and urticaria, while severe anaphylactoid reactions lead to hypotensive shock, pulmonary edema, and cardiorespiratory arrest. Patients with a history of significant contrast allergy should be premedicated for at least 24 hours prior to planned coronary angiography with corticosteroids and antihistamines (H_1- and H_2-blockers) and studies performed with nonionic, low-osmolar contrast agents that have a lower reported rate of allergic reactions.

Contrast-induced acute kidney injury, defined as an increase in creatinine >0.5 mg/dL or 25% above baseline that occurs 48–72 h after contrast administration, occurs in ~2–7% of patients with rates of 20–30% reported in high-risk patients, including those with diabetes mellitus, congestive heart failure, chronic kidney disease, anemia, and older age. Dialysis is required in 0.3–0.7% of patients and is associated with a fivefold increase in in-hospital mortality. For all patients, adequate intravascular volume expansion with intravenous 0.9% saline (1.0–1.5 mL/kg per hour) for 3–12 h before and continued 6–24 h after the procedure limits the risk of contrast-induced acute kidney injury. Pretreatment with *N*-acetylcysteine (Mucomyst) has not reduced the risk of contrast-induced acute kidney injury consistently and, therefore, is no longer recommended routinely. Diabetic patients treated with metformin should stop the drug 24 hours prior to the procedure and not restart until 48 hours after contrast administration to limit the associated risk of lactic acidosis. Other strategies to decrease risk include the administration of sodium bicarbonate (3 mL/kg per hour) 1 hour before and 6 hours after the procedure; use of low- or iso-osmolar contrast agents; and limiting the volume of contrast to <50 mL per procedure.

Cardiac catheterization is performed after the patient has fasted for 6 h and has received intravenous conscious sedation to remain awake but sedated during the procedure. All patients with suspected coronary artery disease are pretreated with 325 mg aspirin. In patients in whom the procedure is likely to progress to a percutaneous coronary intervention, an additional antiplatelet agent should be started: clopidogrel (600-mg loading dose and 75 mg daily) or prasugrel (60-mg loading dose and 10 mg daily), or ticagrelor (180-mg loading and 90 mg twice daily). Prasugrel should not be selected for individuals with prior stroke or transient ischemic attack. Warfarin is held starting 2–3 days prior to the catheterization to allow the international normalized ratio (INR) to fall to <1.7 and limit access-site bleeding complications. The novel oral anticoagulants should be stopped 24–48 h prior to the test. Cardiac catheterization is a sterile procedure, so antibiotic prophylaxis is not required.

TECHNIQUE

Cardiac catheterization and coronary angiography provide a detailed hemodynamic and anatomic assessment of the heart and coronary arteries. The selection of procedures is dependent on the patient’s symptoms and clinical condition, with some direction provided by noninvasive studies.

Vascular Access

Cardiac catheterization procedures are performed using a percutaneous technique to enter the femoral or radial artery and femoral, brachial, or internal jugular vein as the access sites for left and right heart catheterization, respectively. A flexible sheath is inserted into the vessel over a guidewire, allowing diagnostic catheters to be introduced into the vessel and advanced toward the heart using fluoroscopic guidance. The radial artery (or rarely the brachial artery) access site is advantageous in patients with peripheral arterial disease that involves the abdominal aorta, iliac, or femoral vessels; severe iliac artery tortuosity; morbid obesity; or preference for early postprocedure ambulation. Use of radial-artery access is also gaining popularity due to a lower rate of access-site bleeding complications. A normal modified Allen’s test or Barbeau test confirming dual blood supply to the hand from the radial and ulnar arteries is recommended prior to access at this site. The internal jugular or antecubital veins serve as the preferred access sites to the right heart when the patient has an inferior vena cava filter in place or requires prolonged hemodynamic monitoring.

Right Heart Catheterization

This procedure measures pressures in the right heart and pulmonary artery. Right heart catheterization is no longer a routine part of diagnostic cardiac catheterization, but it is reasonable in patients with unexplained dyspnea, pulmonary hypertension, valvular heart disease, pericardial disease, right and/or left ventricular dysfunction, congenital heart disease, and suspected intracardiac shunts. Right heart catheterization most commonly uses a balloon-tipped flotation catheter that is advanced sequentially to the right atrium, right ventricle, pulmonary artery, and pulmonary wedge position (as a surrogate for left atrial pressure) using fluoroscopic guidance; in each cardiac

chamber, pressure is measured and blood samples are obtained for oxygen saturation analysis to screen for intracardiac shunts and calculate a cardiac output.

Left Heart Catheterization

This procedure measures pressures in the left heart as a determinant of left ventricular performance. With the aid of fluoroscopy, a catheter is guided to the ascending aorta and across the aortic valve into the left ventricle to provide a direct measure of left ventricular pressure. In patients with a tilting-disc prosthetic aortic valve, crossing the valve with a catheter is contraindicated, and the left heart may be accessed via a transseptal technique from the right atrium using a needle-tipped catheter to puncture the atrial septum at the fossa ovalis. Once the catheter crosses from the right to the left atrium, it can be advanced across the mitral valve to the left ventricle. This technique is also used for mitral valvuloplasty. Heparin is given for prolonged procedures to limit the risk of stroke from embolism of clots that may form on the catheter. For patients with heparin-induced thrombocytopenia, the direct thrombin inhibitors bivalirudin (0.75 mg/kg bolus, 1.75 mg/kg per hour for the duration of the procedure) or argatroban (350 μg/kg bolus, 15 μg/kg per min for the duration of the procedure) may be used.

HEMODYNAMICS

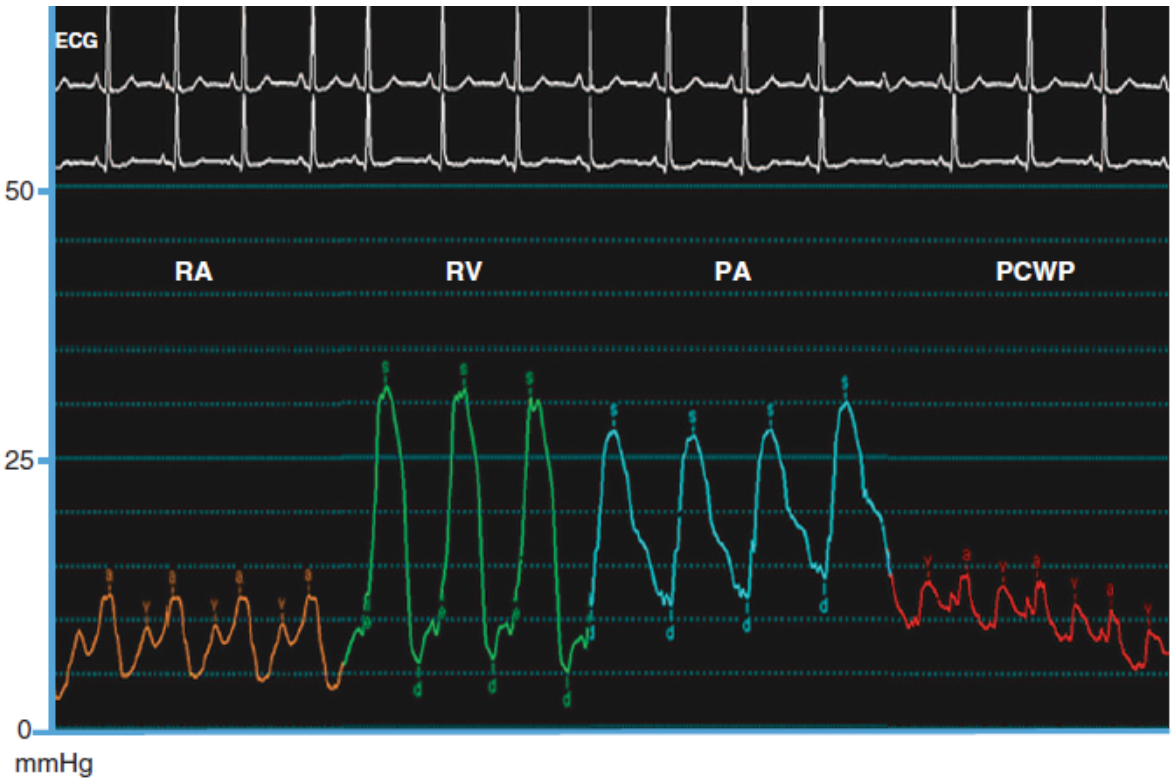
A comprehensive hemodynamic assessment involves obtaining pressure measurements in the right and left heart and peripheral arterial system and determining the cardiac output **(Table 237-2)**. The shape and magnitude of the pressure waveforms provide important diagnostic information; an example of normal pressure tracings is shown in **Fig. 237-1**. In the absence of valvular heart disease, the atria and ventricles are “one chamber” during diastole when the tricuspid and mitral valves are open while in systole, when the pulmonary and aortic valves are open, the ventricles and their respective outflow tracts are considered “one chamber.” These concepts form the basis by which hemodynamic measurements are used to assess valvular stenosis. When aortic stenosis is present, there is a systolic pressure gradient between the left ventricle and the aorta; when mitral stenosis is present, there is a diastolic pressure gradient between the pulmonary capillary wedge (left atrial) pressure and the left ventricle **(Fig. 237-2)**. Hemodynamic measurements also discriminate between aortic stenosis and hypertrophic obstructive cardiomyopathy where the asymmetrically hypertrophied septum creates a dynamic intraventricular pressure gradient during ventricular systole. The magnitude of this obstruction is measured using an end-hole catheter positioned at the left ventricular apex that is pulled back while recording pressure; once the catheter has passed the septal obstruction and is positioned in the apex of the left ventricle, a gradient can be measured between the left ventricular apex and the aorta. Hypertrophic obstructive cardiomyopathy is confirmed by the Brockenbrough-Braunwald sign: following a premature ventricular contraction, there is an increase in the left ventricular–aorta pressure gradient with a simultaneous decrease in the aortic pulse pressure. The finding of a decrease in pulse pressure is absent in aortic stenosis.

TABLE 237-2 Normal Values for Hemodynamic Measurements

Pressures (mmHg)	
Right atrium	
Mean	0–5
a wave	1–7
v wave	1–7
Right ventricle	
Peak systolic/end diastolic	17–32/1–7
Pulmonary artery	
Peak systolic/end diastolic	17–32/1–7
Mean	9–19
Pulmonary capillary wedge (mean)	4–12
Left atrium	
Mean	4–12
a wave	4–15
v wave	4–15
Left ventricle	
Peak systolic/end diastolic	90–130/5–12
Aorta	
Peak systolic/end diastolic	90–130/60–85
Mean	70–100
(Resistances [dyn-s]/cm^5)	
Systemic vascular resistance	900–1400
Pulmonary vascular resistance	40–120
Oxygen Consumption Index ([L-min]/m^2)	**115–140**
Arteriovenous oxygen difference (vol %)	3.5–4.8
Cardiac index ([L-min]/m^2)	2.8–4.2

FIGURE 237-1

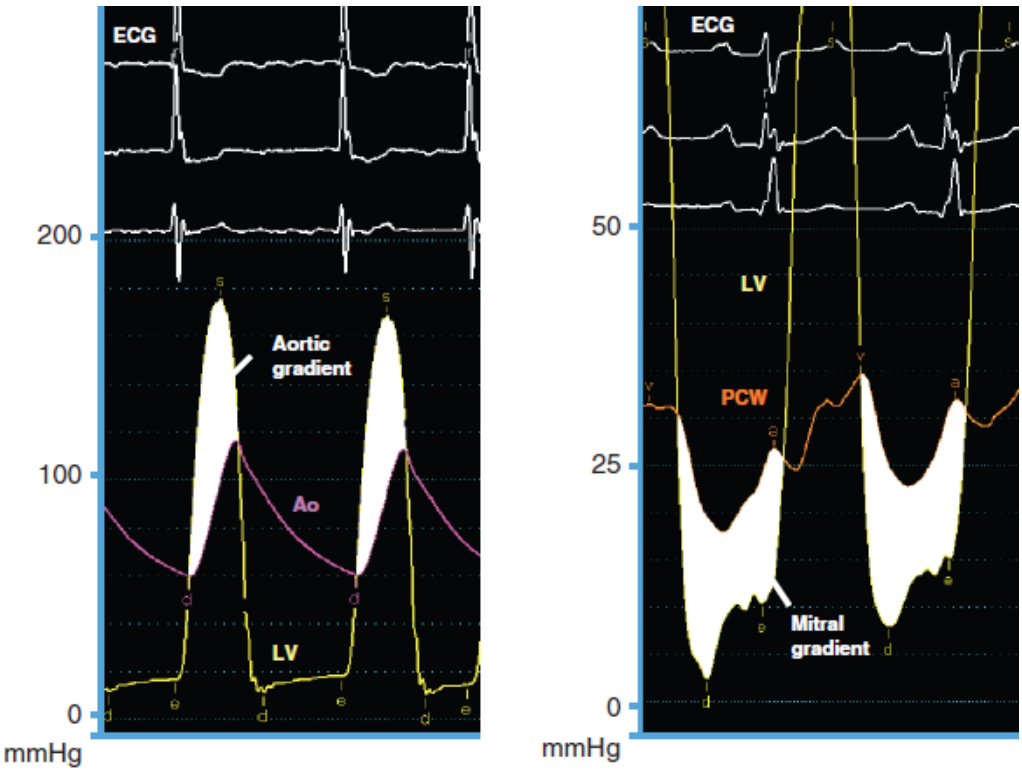
Normal hemodynamic waveforms recorded during right heart catheterization. Atrial pressure tracings have a characteristic "*a*" wave that reflects atrial contraction and a "*v*" wave that reflects pressure changes in the atrium during ventricular systole. Ventricular pressure tracings have a low-pressure diastolic filling period and a sharp rise in pressure that occurs during ventricular systole. d, diastole; PA, pulmonary artery; PCWP, pulmonary capillary wedge pressure; RA, right atrium; RV, right ventricle; s, systole.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 237-2

Severe aortic and mitral stenosis. Simultaneous recording of left ventricular (LV) and aortic (Ao) pressure tracings demonstrates a 62-mmHg mean systolic gradient (*shaded area*) that corresponds to an aortic valve area of 0.6 cm^2 (*left*). Simultaneous recording of LV and pulmonary capillary wedge (PCW) pressure tracings reveals a 14-mmHg mean diastolic gradient (*shaded area*) that is consistent with critical mitral stenosis (mitral valve area = 0.5 cm^2). d, diastole; e, end diastole; s, systole.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Regurgitant valvular lesions increase volume (and pressure) in the “receiving” cardiac chamber. In severe mitral and tricuspid regurgitation, the increase in blood flow to the atria takes place during ventricular systole, leading to an increase in the *v* wave (often two times greater than the mean pressure). Severe aortic regurgitation leads to a decrease in aortic diastolic pressure with a concomitant rise in left ventricular end-diastolic pressure, resulting in equalization of pressures between the two chambers at end-diastole.

Hemodynamic measurements are also used to differentiate between cardiac tamponade, constrictive pericarditis, and restrictive cardiomyopathy **(Table 237-3)**. In cardiac tamponade, right atrial pressure is increased with a decreased or absent “y” descent, indicative of impaired right atrial emptying in diastole, and there is diastolic equalization of pressures in all cardiac chambers. In constrictive pericarditis, right atrial pressure is elevated with a prominent “y” descent, indicating rapid filling of the right ventricle during early diastole. A diastolic dip and plateau or “square root sign,” in the ventricular waveforms due to an abrupt halt in ventricular filling during diastole; right ventricular and pulmonary artery pressures are elevated; and discordant pressure changes in the right and left ventricles with inspiration (right ventricular systolic pressure increases while left ventricular systolic pressure decreases) are observed. The latter hemodynamic phenomenon is the most specific for constriction. Restrictive cardiomyopathy may be distinguished from constrictive pericarditis by a marked increase in right ventricular and pulmonary artery systolic pressures (usually >60 mmHg), a separation of the left and right ventricular diastolic pressures by >5 mmHg (at baseline or with acute volume loading), and concordant changes in left and right ventricular diastolic filling pressures with inspiration (both increase).

TABLE 237-3 Hemodynamic Findings in Tamponade, Constrictive Pericarditis, and Restrictive Cardiomyopathy

	CARDIAC TAMPONADE	CONSTRICTIVE PERICARDITIS	EFFUSIVE-CONSTRICTIVE PERICARDITIS	RESTRICTIVE CARDIOMYOPATHY
Pericardial pressure	↑	↑	↑	Normal
Right atrium pressure	↑	↑	↑ (Fails to decrease by 50% or to <10 mmHg after pericardiocentesis)	↑
Right atrium pressure waveform	Prominent "x" descent Diminished or absent "y" descent	Prominent "x" descent Prominent "y" descent	Prominent "x" descent "y" descent less prominent than expected	Prominent "y" descent
Right ventricle systolic pressure	<50 mmHg	<50 mmHg	<50 mmHg	>60 mmHg
Right ventricle end-diastolic pressure		>1/3 right ventricular systolic pressure	>1/3 right ventricular systolic pressure	<1/3 right ventricular systolic pressure
	Equals left ventricular end-diastolic pressure within 5 mmHg	Equals left ventricular end-diastolic pressure within 5 mmHg	Equals left ventricular end-diastolic pressure within 5 mmHg	Less than left ventricular end-diastolic pressure by ≥5 mmHg
Right ventricle pressure waveform		Dip and plateau or "square root" sign	Dip and plateau or "square root" sign	Dip and plateau or "square root" sign
Right ventricle–left ventricle systolic pressure relationship with inspiration	Discordant	Discordant	Discordant	Concordant

Cardiac Output

Cardiac output is measured by the Fick method or the thermodilution technique. Typically, the Fick method and thermodilution technique are both performed during cardiac catheterization, although the Fick method is considered more reliable in the presence of tricuspid regurgitation and in low-output states. The Fick method uses oxygen as the indicator substance and is based on the principle that the amount of a substance taken up or released by an organ (oxygen consumption) is equal to the product of its blood flow (cardiac output) and the difference in the concentration of the substance in the arterial and venous circulation (arterial-venous oxygen difference). Thus, the formula for calculating the Fick cardiac output is:

$$\text{Cardiac output (L/min)} = (\text{oxygen consumption [mL/min]})/(\text{arterial-venous oxygen difference [mL/L]})$$

Oxygen consumption is estimated as 125 mL oxygen/minute × body surface area, and the arterial-venous oxygen difference is determined by first calculating the oxygen carrying capacity of blood (hemoglobin [g/100 mL] × 1.36 [mL oxygen/g hemoglobin] × 10) and multiplying this product by the fractional oxygen saturation. The thermodilution method measures a substance that is injected into and adequately mixes with blood. In contemporary practice, thermodilution cardiac outputs are measured using temperature as the indicator. Measurements are made with a thermistor-tipped catheter that detects temperature deviations in the pulmonary artery after the injection of 10 mL of room-temperature normal saline into the right atrium.

Vascular Resistance

Resistance across the systemic and pulmonary circulations is calculated by extrapolating from Ohm's law of electrical resistance and is equal to the mean pressure gradient divided by the mean flow (cardiac output). Therefore, systemic vascular resistance is ([mean aortic pressure − mean right atrial pressure]/cardiac output) multiplied by 80 to convert the resistance from Wood units to dyn-s-cm^{-5}. Similarly, the pulmonary vascular resistance is ([mean pulmonary artery − mean pulmonary capillary wedge pressure]/cardiac output) × 80. Pulmonary vascular resistance is lowered by oxygen, nitroprusside, calcium channel blockers, prostacyclin infusions, and inhaled nitric oxide; these therapies may be administered during catheterization to determine if increased pulmonary vascular resistance is fixed or reversible.

Valve Area

Hemodynamic data may also be used to calculate the valve area using the Gorlin formula that equates the area to the flow across the valve divided by the pressure gradient between the cardiac chambers surrounding the valve. The formula for the assessment of valve area is: Area = (cardiac output [cm^3/min]/[systolic ejection period or diastolic filling period][heart rate])/44.3 C × square root of the pressure gradient, where C = 1 for aortic valve and 0.85 for the mitral valve. A valve area of <1.0 cm^2 and a mean gradient of >40 mmHg indicate severe aortic stenosis, while a valve area of <1.5 cm^2 and a mean gradient >5–10 mmHg are consistent with moderate-to-severe mitral stenosis; in symptomatic patients with a mitral valve area >1.5 cm^2, a mean gradient >15 mmHg, pulmonary artery pressure >60 mmHg, or a pulmonary artery wedge pressure >25 mmHg after exercise is also considered significant and may warrant intervention. The modified Hakki formula has also been used to estimate aortic valve area. This formula calculates the valve area as the cardiac output (L/min) divided by the square root of the pressure gradient. Aortic valve area calculations based on the Gorlin formula are flow-dependent and, therefore, for patients with low cardiac outputs, it is imperative to determine if a decreased valve area actually reflects a fixed stenosis or is overestimated by a low cardiac output and stroke volume that is insufficient to open the valve leaflets fully. In these instances, cautious hemodynamic manipulation using dobutamine to increase the cardiac output and recalculation of the aortic valve area may be necessary.

Intracardiac Shunts

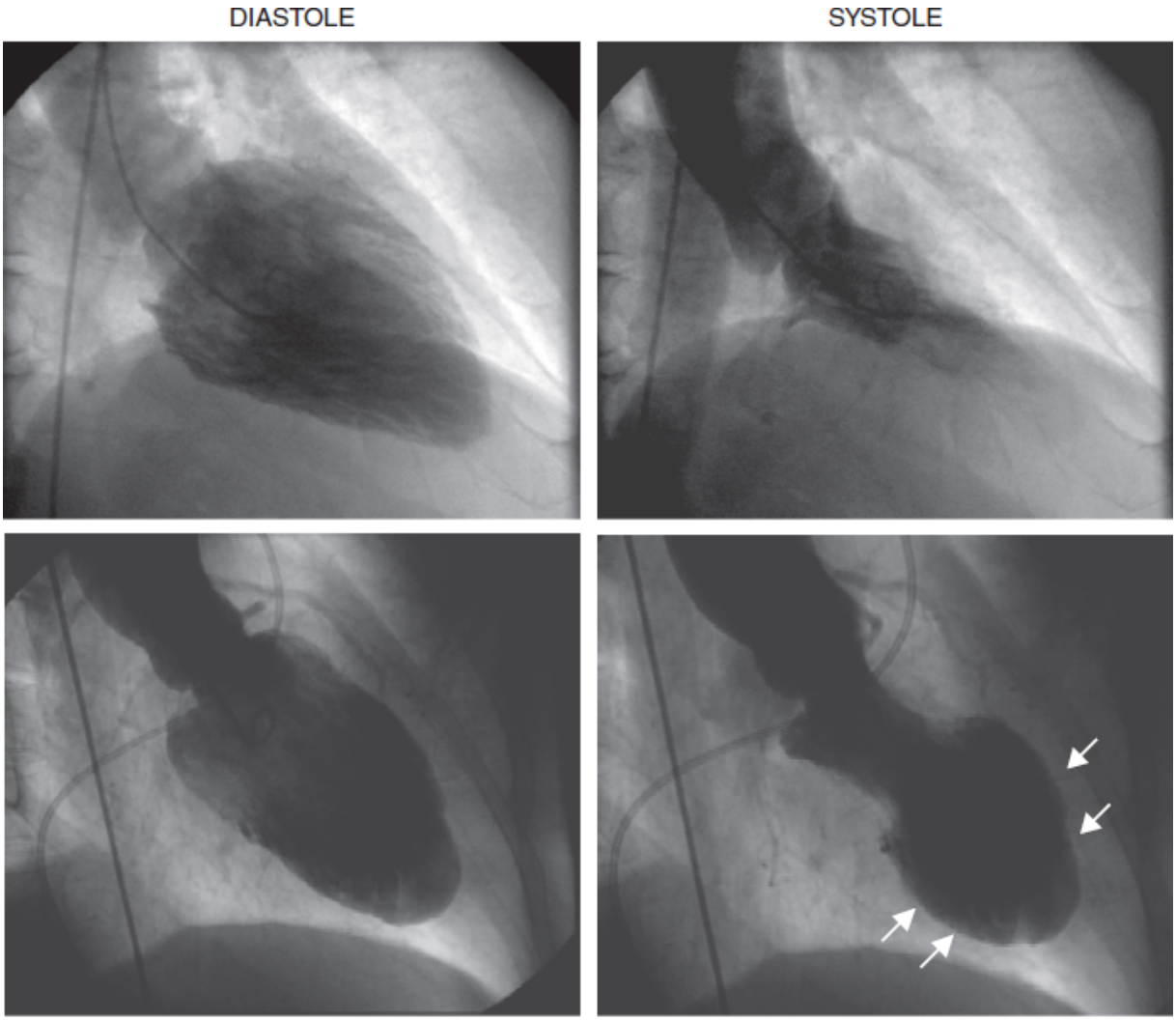
In patients with congenital heart disease, detection, localization, and quantification of the intracardiac shunt should be evaluated. A shunt should be suspected when there is unexplained arterial desaturation or increased oxygen saturation of venous blood. A “step up” or increase in oxygen content indicates the presence of a left-to-right shunt while a “step down” indicates a right-to-left shunt. The shunt is localized by detecting a difference in oxygen saturation levels of 5–7% between adjacent cardiac chambers. The severity of the shunt is determined by the ratio of pulmonary blood flow (Q_p) to the systemic blood flow (Q_s), or Q_p/Q_s = ([systemic arterial oxygen content − mixed venous oxygen content]/pulmonary vein oxygen content − pulmonary artery oxygen content). For an atrial septal defect, a shunt ratio of 1.5 is considered significant and factored with other clinical variables to determine the need for intervention. When a congenital ventricular septal defect is present, a shunt ratio of ≥2.0 with evidence of left ventricular volume overload is a strong indication for surgical correction.

VENTRICULOGRAPHY AND AORTOGRAPHY

Ventriculography to assess left ventricular function may be performed during cardiac catheterization. A pigtail catheter is advanced retrograde across the aortic valve into the left ventricle and 30–45 mL of contrast is power-injected to visualize the left ventricular chamber during the cardiac cycle. The ventriculogram is usually performed in the right anterior oblique projection to examine wall motion and mitral valve function. Normal wall motion is observed as symmetric contraction of all segments; hypokinetic segments have decreased contraction, akinetic segments do not contract, and dyskinetic segments appear to bulge paradoxically during systole **(Fig. 237-3)**. Ventriculography may also reveal a left ventricular aneurysm, pseudoaneurysm, or diverticulum and can be used to assess mitral valve prolapse and the severity of mitral regurgitation. The degree of mitral regurgitation is estimated by comparing the density of contrast opacification of the left atrium with that of the left ventricle. Minimal contrast reflux into the left atrium is considered 1+ mitral regurgitation, while contrast density in the left atrium that is greater than that in the left ventricle with reflux of contrast into the pulmonary veins within three beats defines 4+ mitral regurgitation. Ventriculography performed in the left anterior oblique projection can be used to identify a ventricular septal defect. Calculation of the ventricular volumes in systole and diastole allows calculation of stroke volume and cardiac output.

FIGURE 237-3

Left ventriculogram at end diastole (*left*) and end systole (*right*). In patients with normal left ventricular function, the ventriculogram reveals symmetric contraction of all walls (*top*). Patients with coronary artery disease may have wall motion abnormalities on ventriculography as seen in this 60-year-old male following a large anterior myocardial infarction. In systole, the anterior, apical, and inferior walls are akinetic (*white arrows*) (*bottom*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

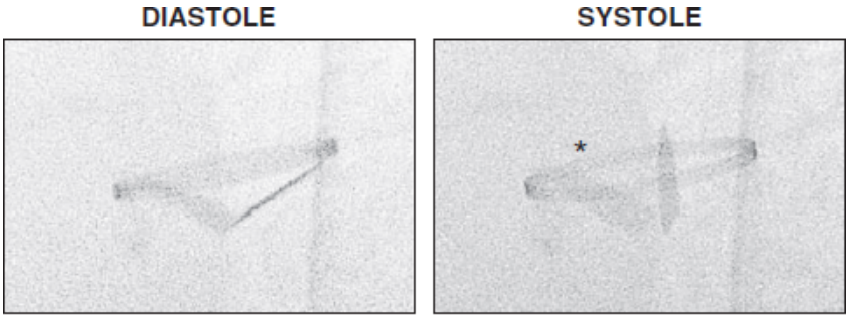
Aortography in the cardiac catheterization laboratory visualizes abnormalities of the ascending aorta, including aneurysmal dilation and involvement of the great vessels, as well as dissection with compression of the true lumen by an intimal flap that separates the true and false lumina. Aortography can also be used to identify patent saphenous vein grafts that elude selective cannulation, identify shunts that involve the aorta such as a patent ductus arteriosus, and provide a qualitative assessment of aortic regurgitation using a 1+−4+ scale similar to that used for mitral regurgitation.

CINEFLUOROSCOPY OF PROSTHETIC MECHANICAL VALVES

Prosthetic valve leaflet dysfunction may occur as a result of thrombus or obstruction of leaflet excursion by pannus (**Fig. 237-4**). The incidence of prosthetic valve thrombosis in left-sided valves is 0.1–6.0% per patient-year with differences in rates attributable to valve type, position, anticoagulation status, and left ventricular function. Prosthetic valve dysfunction should be suspected in patients with subtherapeutic anticoagulation with a low mean International Normalized Ratio (INR), a prothrombotic state, recent onset heart failure, cardiogenic shock, cardiac arrest, thromboembolic event or, in asymptomatic patients, an increasing gradient across the valve. Cinefluoroscopy visualizes the motion of mechanical valve leaflets, and is noninvasive, available in most centers, and can be performed rapidly with minimal radiation exposure. Prosthetic mechanical valves should be imaged *en face* and at a 90° angle over several cardiac cycles to document opening and closing of the valve leaflets as well as motion of the base ring. Each type of prosthetic valve has leaflet opening and closing angles that are reported by the manufacturer and can be used to determine if movement or closure of the valve leaflets is restricted suggestive of mechanical obstruction.

FIGURE 237-4

Cinefluoroscopic detection of mechanical valve leaflet dysfunction. Images of a bileaflet mechanical valve in the aortic position taken during diastole (*left*) and systole (*right*) show that one leaflet opens normally during systole while the other leaflet (below asterisk) remains immobile and fixed consistent with valve leaflet thrombosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

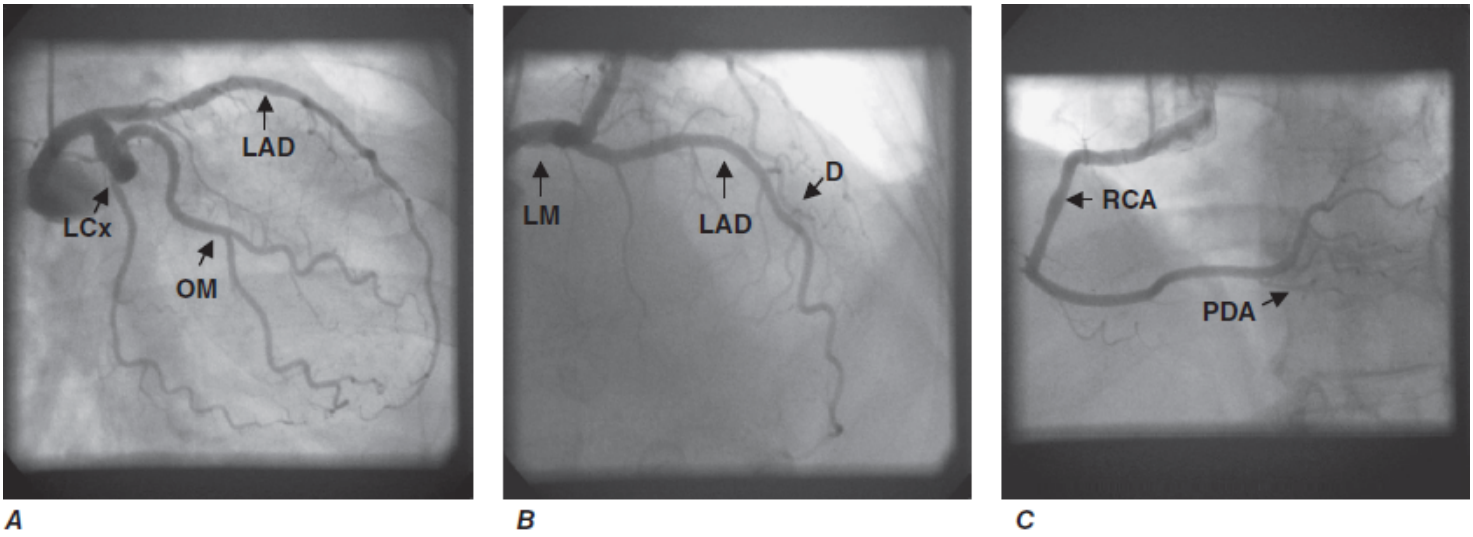
CORONARY ANGIOGRAPHY

Selective coronary angiography is almost always performed during cardiac catheterization and is used to define the coronary anatomy and determine the extent of epicardial coronary artery and coronary artery bypass graft disease. Specially shaped coronary catheters are used to engage the left and right coronary ostia. Hand injection of radiopaque contrast agents creates a coronary “luminogram” that is recorded as radiographic images (cine angiography). Because the coronary arteries are three-dimensional objects that are in motion with the cardiac cycle, angiograms of the vessels using several different orthogonal projections are taken to best visualize the vessels without overlap or foreshortening.

The normal coronary anatomy is highly variable between individuals, but, in general, there are two coronary ostia and three major coronary vessels—the left anterior descending, the left circumflex, and the right coronary arteries with the left anterior descending and left circumflex arteries arising from the left main coronary artery **(Fig. 237-5)**. When the right coronary artery is the origin of the atrioventricular nodal branch, the posterior descending artery, and the posterior lateral vessels, the circulation is defined as right dominant; this is found in ~85% of individuals. When these branches arise from the left circumflex artery as occurs in ~5% of individuals, the circulation is defined as left dominant. The remaining ~10% of patients have a codominant circulation with the posterior descending vessel arising from both the right coronary and the posterior lateral vessels from left coronary circulation. In some patients, a ramus intermedius branch arises directly from the left main coronary artery; this finding is a normal variant. Coronary artery anomalies occur in 1–2% of patients, with separate ostia for the left anterior descending and left circumflex arteries being the most common (0.41%).

FIGURE 237-5

Normal coronary artery anatomy. *A.* Coronary angiogram showing the left circumflex (LCx) artery and its obtuse marginal (OM) branches. The left anterior descending (LAD) artery is also seen but may be foreshortened in this view. ***B.*** The LAD and its diagonal (D) branches are best seen in cranial views. In this angiogram, the left main (LM) coronary artery is also seen. ***C.*** The right coronary artery (RCA) gives off the posterior descending artery (PDA), so this is a right dominant circulation.

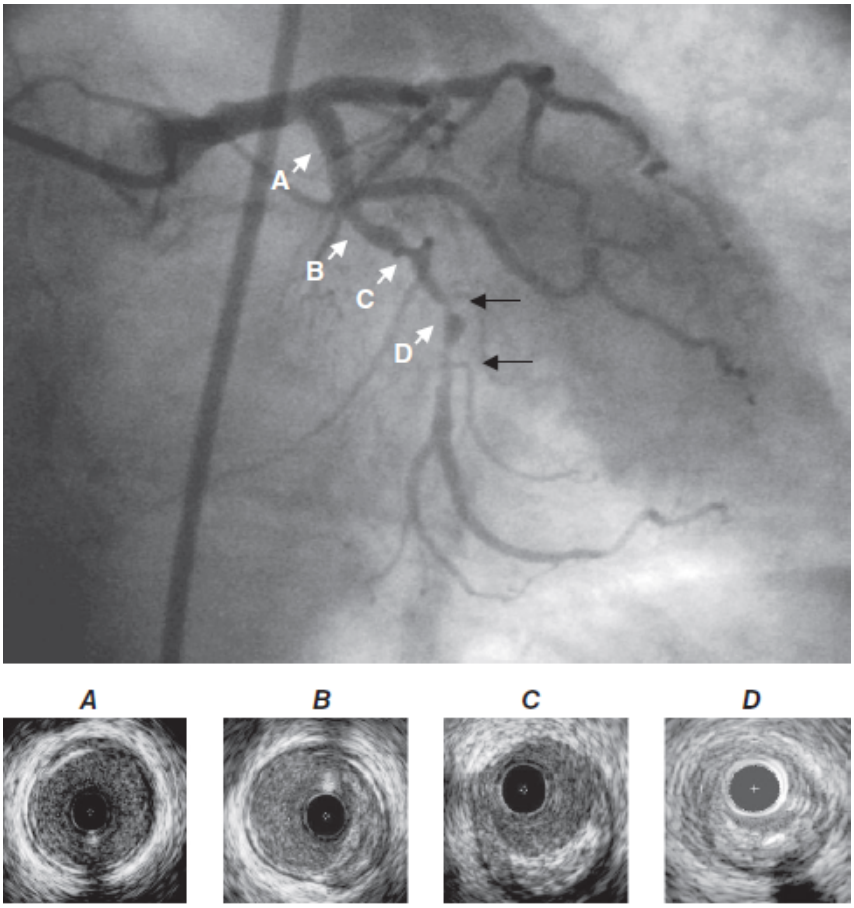


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Coronary angiography visualizes coronary artery stenoses as luminal narrowings on the cine angiogram. The degree of narrowing is referred to as the percent stenosis and is determined visually by comparing the most severely diseased segment with a proximal or distal “normal segment”; a stenosis >50% is considered significant **(Fig. 237-6)**. Online quantitative coronary angiography can provide a more accurate assessment of the percent stenosis and lessen the tendency to overestimate lesion severity visually. The presence of a myocardial bridge, which most commonly involves the left anterior descending artery, may be mistaken for a significant stenosis; this occurs when a portion of the vessel dips below the epicardial surface into the myocardium and is subject to compressive forces during ventricular systole. The key to differentiating a myocardial bridge from a fixed stenosis is that the “stenosed” part of the vessel returns to normal during diastole. Coronary calcification is also seen during angiography prior to the injection of contrast agents. Collateral blood vessels may be seen traversing from one vessel to the distal vasculature of a severely stenosed or totally occluded vessel. Thrombolysis in myocardial infarction (TIMI) flow grade, a measure of the relative duration of time that it takes for contrast to opacify the coronary artery fully, may provide an additional clue to the degree of lesion severity, and the presence of TIMI grade 1 (minimal filling) or 2 (delayed filling) suggests that a severe coronary artery stenosis is present.

FIGURE 237-6

Coronary stenoses on cine angiogram and intravascular ultrasound. Significant stenoses in the coronary artery are seen as narrowings (*black arrows*) of the vessel. Intravascular ultrasound shows a normal segment of artery (***A***), areas with eccentric plaque (***B, C***), and near total obliteration of the lumen at the site of the significant stenosis (***D***). Note that the intravascular ultrasound catheter is present in the images as a black circle.



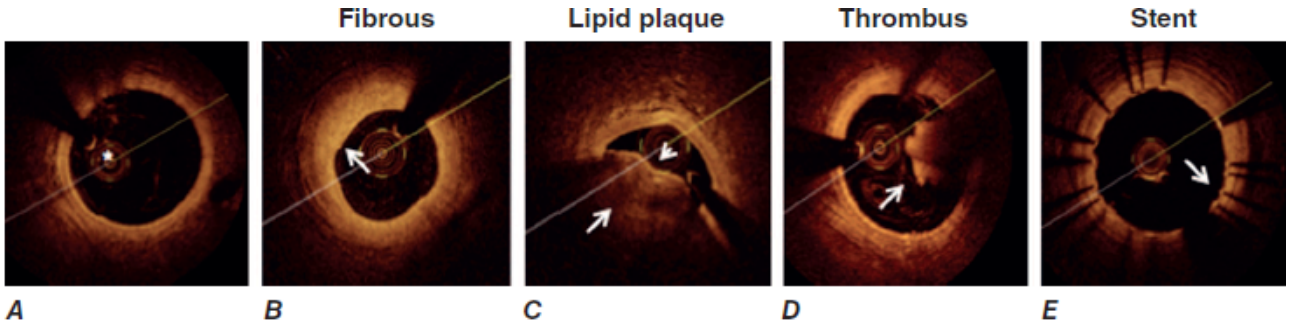
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

INTRAVASCULAR ULTRASOUND, OPTICAL COHERENCE TOMOGRAPHY, AND FRACTIONAL FLOW RESERVE

During coronary angiography, intermediate stenoses (40–70%), indeterminate findings, or anatomic findings that are incongruous with the patient's symptoms may require further interrogation. In these cases, intravascular ultrasound (IVUS) provides a more accurate anatomic assessment of the coronary artery and the degree of coronary atherosclerosis (Fig. 237-6). IVUS is performed using a small flexible catheter with a 40-mHz transducer at its tip that is advanced into the coronary artery over a guidewire. Data from IVUS studies may be used to image atherosclerotic plaque precisely, determine luminal cross-sectional area, and measure vessel size; it is also used during or following percutaneous coronary intervention to assess the stenosis and determine the adequacy of stent placement. Optical coherence tomography (OCT) is a catheter-based imaging technique that uses near-infrared light to generate images with better spatial resolution than IVUS (12–18 microns vs 150–200 microns); however, the depth of field is smaller. The advantage of OCT imaging over IVUS lies in its ability to image characteristics of the atherosclerotic plaque (lipid, fibrous cap) with high definition and to assess coronary stent placement, apposition, and patency **(Fig. 237-7)**.

FIGURE 237-7

Optical coherence tomography imaging. ***A.*** The optical coherence tomography (OCT) catheter (*) in the lumen of a coronary artery with limited neointima formation. The intima is seen with high definition, but unlike intravascular ultrasound imaging, the vessel media and adventitia are not well visualized. ***B.*** A fibrous plaque (*arrow*) is characterized by a bright signal. ***C.*** A large, eccentric, lipid-rich plaque obscures part of the vessel lumen. Because lipid in the plaque absorbs light, the lipid-rich plaque appears as a dark area with irregular borders (*arrow*). The plaque is covered by a thin fibrous cap (*arrowhead*) typical of a vulnerable plaque. ***D.*** A thrombus (*arrow*) adherent to a ruptured plaque that is protruding into the vessel lumen. ***E.*** A coronary stent is that is well opposed to the vessel wall. The stent struts appear as short bright lines with dropout behind the struts (*arrow*).

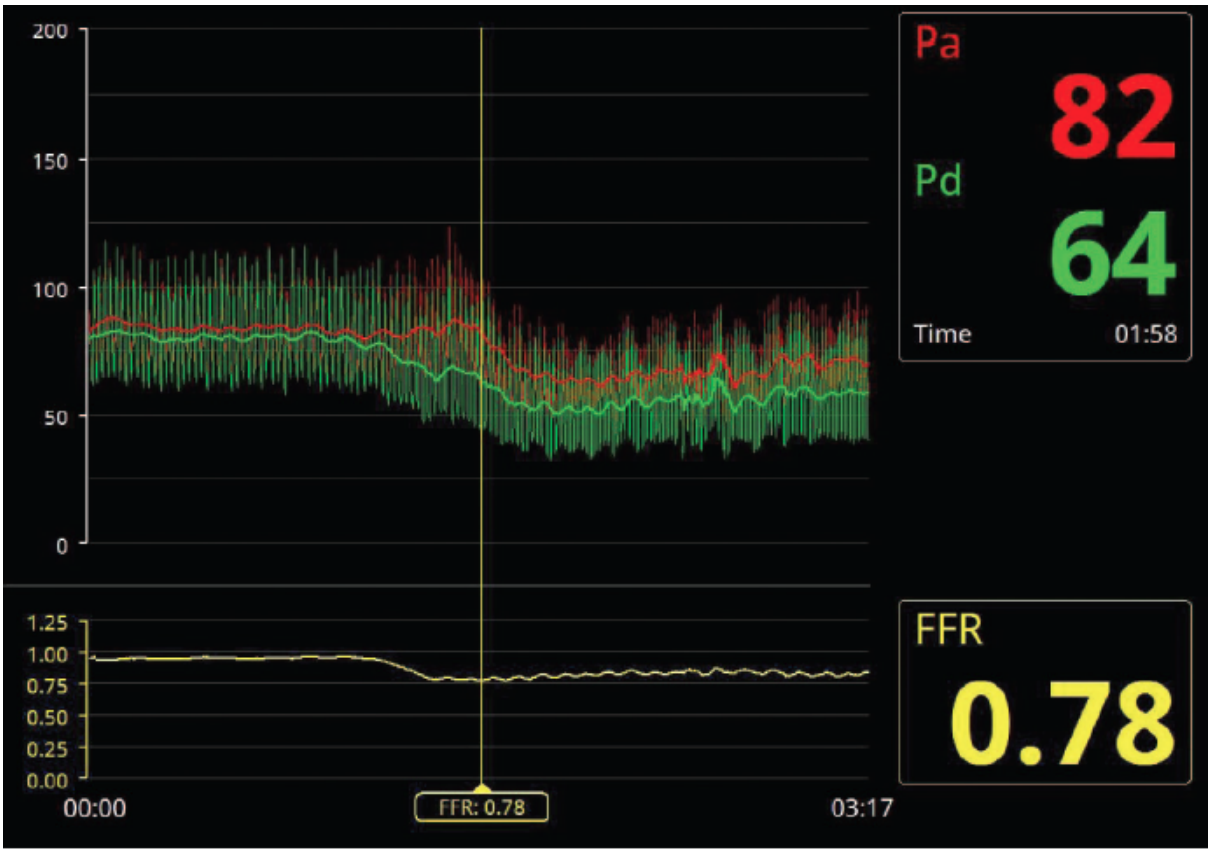


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Measurement of the fractional flow reserve provides a functional assessment of the stenosis and is more accurate in predicting long-term clinical outcome than imaging techniques. The fractional flow reserve is the ratio of the pressure in the coronary artery distal to the stenosis divided by the pressure in the artery proximal to the stenosis at maximal vasodilation. Fractional flow reserve is measured using a coronary pressure–sensor guidewire at rest and at maximal hyperemia following the injection of adenosine (**Fig. 237-8**). A fractional flow reserve of <0.80 indicates a hemodynamically significant stenosis that would benefit from intervention. Using both pressure and velocity, an index of myocardial resistance can also be calculated. Studies have shown this to be an important predictor of outcome as well.

FIGURE 237-8.

Fractional flow reserve. The fractional flow reserve is measured using a coronary pressure-sensor guidewire that measures the ratio of the pressure in the coronary artery distal to the stenosis (Pd, *green*) divided by the pressure in the artery proximal to the stenosis (Pa, *red*) at maximal hyperemia following the injection of adenosine. A fractional flow reserve of <0.80 indicates that revascularization would be beneficial.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

POSTPROCEDURE CARE

Once the procedure is completed, vascular access sheaths are removed. If the femoral approach is used, direct manual compression or vascular closure devices that immediately close the arteriotomy site with a staple/clip, collagen plug, or sutures are used to achieve hemostasis. These devices decrease the length of supine bed rest (from 6 hours to 2–4 hours) and improve patient satisfaction, but have not been shown definitively to be superior to manual compression with respect to access-site complications. With radial-artery access, bed rest is needed for only 2 h. When cardiac catheterization is performed as an elective outpatient procedure, the patient completes postprocedure bed rest in a monitored setting and is discharged home with instructions to liberalize fluids because contrast agents promote an osmotic diuresis, to avoid strenuous activity, and to observe the vascular access site for signs of complications. Overnight hospitalization may be required for high-risk patients with significant comorbidities, patients with complications occurring during the catheterization, or patients who have undergone a percutaneous coronary intervention. Hypotension early after the procedure may be due to inadequate fluid replacement or retroperitoneal bleeding from the access site. Patients who received >2 Gy of radiation during the procedure should be examined for signs of erythema. For patients who received higher doses (>5 Gy), clinical follow-up within 1 month to assess for skin injury is recommended.

FURTHER READING

Bashore TM et al: 2012 American College of Cardiology Foundation/Society for Cardiovascular Angiography and Interventions expert consensus document on cardiac catheterization laboratory standards update: A report of the American College of Cardiology Foundation Task Force on Expert Consensus documents developed in collaboration with the Society of Thoracic Surgeons and Society for Vascular Medicine. J Am Coll Cardiol 59:2221, 2012. [PubMed: 22575325]

Berry C et al: Fractional flow reserve-guided management in stable coronary disease and acute myocardial infarction: Recent developments. Eur Heart J 36:3155, 2015. [PubMed: 26038588]

Bezerra HG et al: Intracoronary optical coherence tomography: A comprehensive review. JACC Cardiovasc Interv 2:1035, 2009. [PubMed: 19926041]

Lim MJ et al (eds): *Hemodynamic Rounds: Interpretation of Cardiac Pathophysiology from Pressure Waveform Analysis*, 3rd ed. Hoboken, John Wiley and Sons, 2009.

Maehara A et al: Advances in intravascular imaging: Circulation. Cardiovasc Interv 2:482, 2009.

Moscucci M (ed): *Grossman & Baim's Cardiac Catheterization, Angiography, and Intervention,* 8th ed. Philadelphia, Lippincott Williams & Wilkins, 2014.

Chapter 238: Principles of Electrophysiology

David D. Spragg; Gordon F. Tomaselli

HISTORY AND INTRODUCTION

The field of cardiac electrophysiology was ushered in with the development of the electrocardiogram (ECG) by Einthoven at the turn of the twentieth century. Subsequent recording of cellular membrane currents demonstrated that the body surface ECG is the timed sum of the cellular action potentials in the atria and ventricles. In the late 1960s, the development of intracavitary recording, in particular, His bundle electrograms, marked the beginning of contemporary clinical electrophysiology. Adoption of radiofrequency (RF) technology to ablate cardiac tissue in the early 1990s heralded the birth of interventional cardiac electrophysiology.

The clinical problem of sudden death caused by ventricular arrhythmias, most commonly in the setting of coronary artery obstruction, was recognized as early as the late nineteenth century. The problem was vexing and led to the development of pharmacologic and nonpharmacologic therapies, including transthoracic defibrillators, cardiac massage, and, most recently, implantable intravascular and subcutaneous defibrillators. Over time the limitations of antiarrhythmic drug therapy have been highlighted repeatedly in clinical trials, and now ablation and devices are first-line therapy for a number of cardiac arrhythmias.

In the last two decades, the genetic basis of a number of heritable arrhythmias has been elucidated, revealing important insights into the mechanisms not only of these rare arrhythmias but also of similar rhythm disturbances observed in more common forms of heart disease.

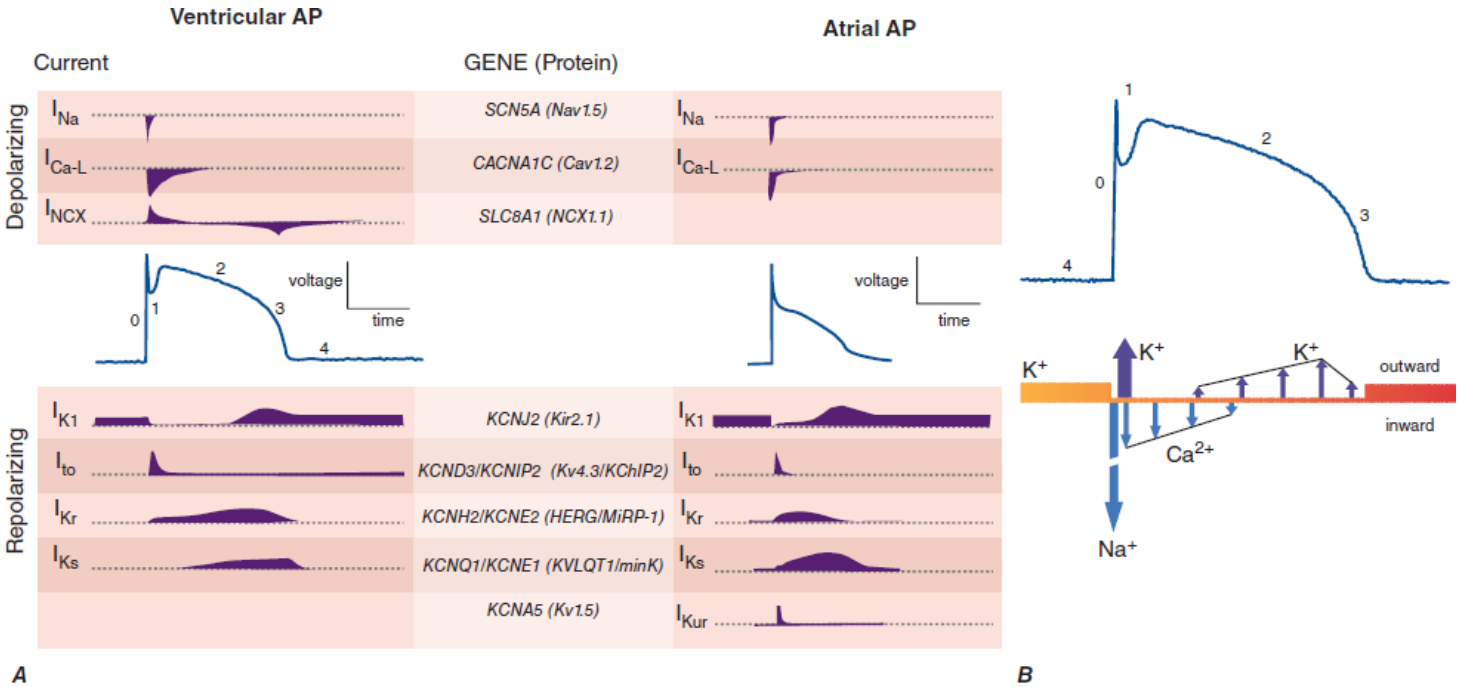
DESCRIPTIVE PHYSIOLOGY

The normal cardiac impulse is generated by pacemaker cells in the sinoatrial (SA) node situated at the junction of the right atrium and the superior vena cava **(see Fig. 235-1)**. This impulse is transmitted slowly through nodal tissue to the anatomically complex atria, where it is conducted more rapidly to the atrioventricular node (AVN), inscribing the P wave of the ECG **(see Fig. 235-2)**. There is a perceptible delay in conduction through the anatomically and functionally heterogeneous AV. The time needed for activation of the atria and the AVN delay is represented as the PR interval of the ECG. The AVN is the only electrical connection between the atria and the ventricles in the normal heart. The electrical impulse emerges from the AVN and is transmitted to the His-Purkinje system, specifically the common bundle of His, then the left and right bundle branches, and then to the Purkinje network, facilitating activation of ventricular muscle. In normal circumstances, the ventricles are activated rapidly in a well-defined fashion that is determined by the course of the Purkinje network, and this inscribes the QRS complex **(see** Fig. 235-2**)**. Recovery of electrical excitability occurs more slowly and is governed by the time of activation and duration of regional action potentials. The relative brevity of epicardial action potentials in the ventricle results in repolarization that occurs first on the epicardial surface and then proceeds to the endocardium, which inscribes a T wave normally of the same polarity as the QRS complex. The duration of ventricular activation and recovery is determined by the action potential duration and is represented on the body surface ECG by the QT interval **(see** Fig. 235-2**)**.

Cardiac myocytes exhibit a characteristically long action potential (200–400 ms) compared with neurons and skeletal muscle cells (1–5 ms). The action potential profile is sculpted by the orchestrated activity of multiple distinctive time- and voltage-dependent ionic currents **(Fig. 238-1A)**. The currents are carried by transmembrane proteins that passively conduct ions down their electrochemical gradients through selective pores (ion channels), actively transport ions against their electrochemical gradient (pumps, transporters), or electrogenically exchange ionic species (exchangers).

FIGURE 238-1

A. Cellular atrial and ventricular action potentials. Phases 0–4 are the rapid upstroke, early repolarization, plateau, late repolarization, and diastole, respectively. The ionic currents and their respective genes are shown above and below the action potentials. The major currents that underlie the action potentials vary in atrial and ventricular myocytes. ***B.*** A ventricular action potential with a schematic of the ionic currents flowing during the phases of the action potential. Potassium current (I_{K1}) is the principal current during phase 4 and determines the resting membrane potential of the myocyte. Sodium current generates the upstroke of the action potential (phase 0); activation of I_{to} with inactivation of the Na current inscribes early repolarization (phase 1). The plateau (phase 2) is generated by a balance of repolarizing potassium currents and depolarizing calcium current. Inactivation of the calcium current with persistent activation of potassium currents (predominantly I_{Kr} and I_{Ks}) causes phase 3 repolarization.



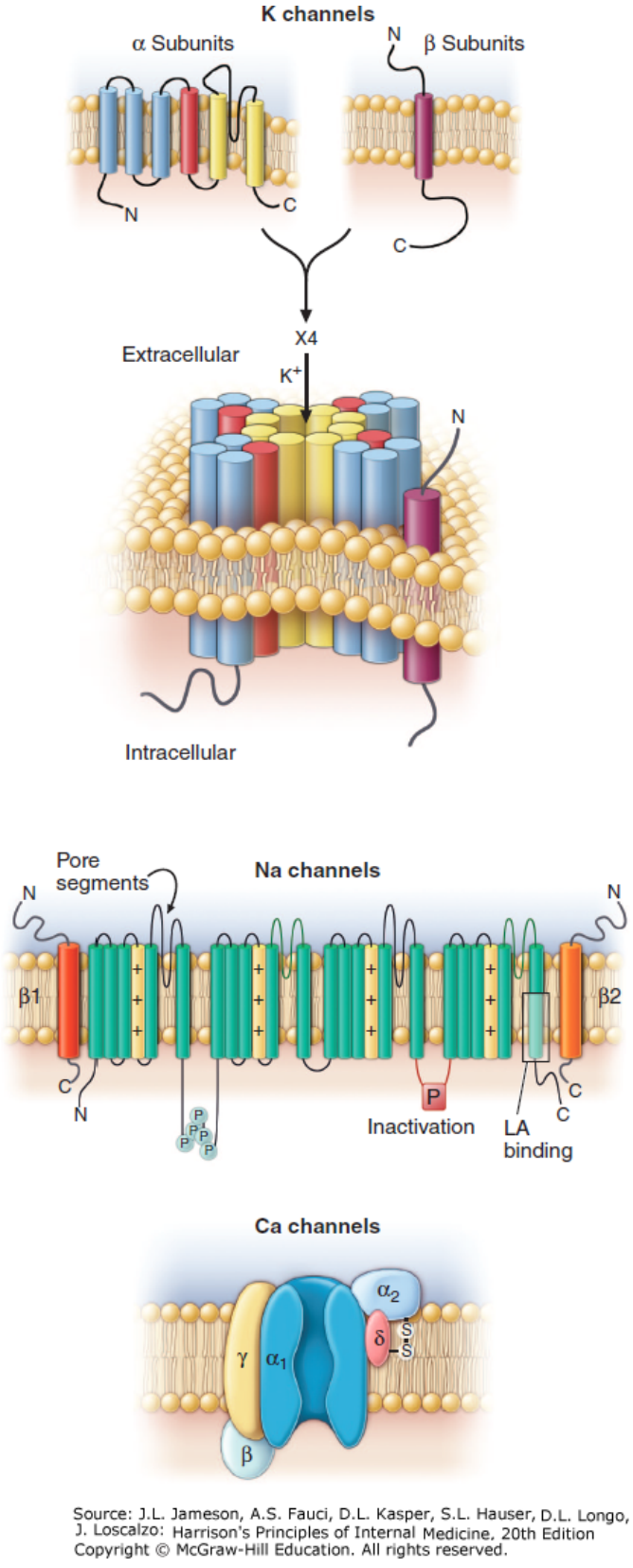
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Action potentials in the heart are regionally distinct. The regional variability in cardiac action potentials is a result of differences in the number and types of ion channel proteins expressed by different cell types in the heart. Further, unique sets of ionic currents are active in pacemaking and muscle cells, and the relative contributions of these currents may vary in the same cell type in different regions of the heart (Fig. 238-1A).

Ion channels are complex, multisubunit transmembrane glycoproteins that open and close in response to a number of biologic stimuli, including a change in membrane voltage, ligand binding (directly to the channel or to a G protein–coupled receptor), and mechanical deformation (**Fig. 238-2**). Other ion motive exchangers and transporters contribute importantly to cellular excitability in the heart. Ion pumps establish and maintain the ionic gradients across the cell membrane that serve as the driving force for current flow through ion channels. Transporters or exchangers that do not move ions in an electrically neutral manner (e.g., the sodium-calcium exchanger transports three Na^+ for one Ca^{2+}) are termed *electrogenic* and contribute directly to the action potential profile.

FIGURE 238-2

Topology and subunit composition of the voltage-dependent ion channels. Potassium channels are formed by the tetramerization of α or pore-forming subunits and one or more β subunits; only single β subunits are shown for clarity. Sodium and calcium channels are composed of α subunits with four homologous domains and one or more ancillary subunits. In all channel types, the loop of protein between the fifth and sixth membrane-spanning repeat in each subunit or domain forms the ion-selective pore. In the case of the sodium channel, the channel is a target for phosphorylation, the linker between the third and fourth homologous domain is critical to inactivation, and the sixth membrane-spanning repeat in the fourth domain is important in local anesthetic antiarrhythmic drug binding. The Ca channel is a multisubunit protein complex with the α_1 subunit containing the pore and major drug binding domain.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The most abundant superfamily of ion channels expressed in the heart is voltage gated. Several structural themes are common to all voltage-dependent ion channels. First, the architecture is modular, consisting either of four homologous subunits (e.g., K channels) or of four internally homologous domains (e.g., Na and Ca channels). Second, the proteins fold around a central pore lined by amino acids that exhibit exquisite conservation within a given channel family of like selectivity (e.g., all Na channels have very similar P segments). Third, the general strategy for activation gating (opening and closing in response to changes in membrane voltage) is highly conserved: the fourth transmembrane segment (S4), studded with positively charged residues, lies within the membrane field and moves in response to depolarization, opening the channel. Fourth, most ion channel complexes include not only the pore-forming proteins (α subunits) but also auxiliary subunits (e.g., β subunits) that modify channel function (Fig. 238-2).

Na and Ca channels are the primary carriers of depolarizing current in both the atria and the ventricles; inactivation of these currents and activation of repolarizing K currents hyperpolarize the heart cells, reestablishing the negative resting membrane potential (**Fig. 238-1B**).

The *plateau phase* is a time when little current is flowing, and relatively minor changes in depolarizing or repolarizing currents can have profound effects on the shape and duration of the action profile. Mutations in subunits of these channel proteins produce arrhythmogenic alterations in the action potentials that cause long and short QT syndrome, Brugada syndrome, idiopathic ventricular fibrillation, familial atrial fibrillation, and some forms of conduction system disease.

MECHANISMS OF CARDIAC ARRHYTHMIAS

Cardiac arrhythmias result from abnormalities of electrical impulse generation, conduction, or both. Bradyarrhythmias typically arise from disturbances in impulse formation at the level of the SA node or from disturbances in impulse propagation at any level, including exit block from the sinus node, conduction block in the AVN, and impaired conduction in the His-Purkinje system. Tachyarrhythmias can be classified according to mechanism, including enhanced automaticity (spontaneous depolarization of atrial, junctional, or ventricular pacemakers), triggered arrhythmias (initiated by afterdepolarizations occurring during or immediately after cardiac repolarization, during phase 3 or 4 of the action potential), or reentry (circus propagation of a depolarizing wavefront). A variety of mapping and pacing maneuvers typically performed during invasive electrophysiologic testing can often determine the underlying mechanism of a tachyarrhythmia (**Table 238-1**).

TABLE 238-1 Arrhythmia Mechanisms

ELECTROPHYSIOLOGIC PROPERTY	MOLECULAR COMPONENTS	MECHANISM	PROTOTYPIC ARRHYTHMIAS
Cellular			
Impulse Initiation			
Automaticity	I_f, $I_{Ca\text{-}L}$, $I_{Ca\text{-}T}$, I_K, I_{K1}	Suppression/acceleration of phase 4	Sinus bradycardia, sinus tachycardia
Triggered automaticity	Calcium overload, I_{TI}	DADs	Digitalis toxicity, reperfusion VT
	$I_{Ca\text{-}L}$, I_K, I_{Na}	EADs	Torsades des pointes, congenital and acquired
Excitation	I_{Na}	Suppression of phase 0	Ischemic VF
	$I_{K\text{-}ATP}$	AP shortening, inexcitability	
	$I_{Ca\text{-}L}$	Suppression	AV block
Repolarization	I_{Na}, $I_{Ca\text{-}L}$, I_K, I_{K1}, Ca^{2+} homeostasis	AP prolongation, EADs, DADs	Polymorphic VT (HF, LVH)
	$I_{Ca\text{-}L}$, K channels, Ca^{2+} homeostasis	AP shortening	Atrial fibrillation
Multicellular			
Cellular Coupling	Connexins (Cx43), I_{Na}, $I_{K\text{-}ATP}$	Decreased coupling	Ischemic VT/VF
Tissue Structure	Extracellular matrix, collagen	Excitable gap and functional reentry	Monomorphic VT, atrial fibrillation

Abbreviations: AP, action potential; AV, atrioventricular; DADs, delayed afterdepolarizations; EADs, early afterdepolarizations; HF, heart failure; LVH, left ventricular hypertrophy; VF, ventricular fibrillation; VT, ventricular tachyarrhythmia.

Alterations in Impulse Initiation: Automaticity

Spontaneous (phase 4) diastolic depolarization underlies the property of automaticity characteristic of pacemaking cells in the SA and atrioventricular (AV) nodes, His-Purkinje system, coronary sinus, and pulmonary veins. Phase 4 depolarization results from the concerted action of a number of ionic currents, including K^+ currents, Ca^{2+} currents, electrogenic Na, K-ATPase, the Na-Ca exchanger, and the so-called funny, or pacemaker, current (I_f); however, the relative importance of these currents remains controversial.

The rate of phase 4 depolarization and, therefore, the firing rates of pacemaker cells are dynamically regulated. Prominent among the factors that modulate phase 4 is autonomic nervous system tone. The negative chronotropic effect of activation of the parasympathetic nervous system is a result of the release of acetylcholine that binds to muscarinic receptors, releasing G protein βγ subunits that activate a potassium current (I_{KACh}) in nodal and atrial cells. The resulting increase in K^+ conductance opposes membrane depolarization, slowing the rate of rise of phase 4 of the action potential. Conversely, augmentation of sympathetic nervous system tone increases myocardial catecholamine concentrations, which activate both α- and β-adrenergic receptors. The effect of β_1-adrenergic stimulation predominates in pacemaking cells, augmenting both L-type Ca current ($I_{Ca\text{-}L}$) and I_f, thus increasing the slope of phase 4. Enhanced sympathetic nervous system activity can dramatically increase the rate of firing of SA nodal cells, producing sinus tachycardia with rates >200 beats/min. By contrast, the increased rate of firing of Purkinje cells is more limited, rarely producing ventricular tachyarrhythmias >120 beats/min.

Normal automaticity may be affected by a number of other factors associated with heart disease. Hypokalemia and ischemia may reduce the activity of Na, K-ATPase, thereby reducing the background repolarizing current and enhancing phase 4 diastolic depolarization. The end result would be an increase in the spontaneous firing rate of pacemaking cells. Modest increases in extracellular potassium may render the maximum diastolic potential more positive, thereby also increasing the firing rate of pacemaking cells. A more significant increase in $[K^+]_o$, however, renders the heart inexcitable by depolarizing the membrane potential.

Normal or enhanced automaticity of subsidiary latent pacemakers produces escape rhythms in the setting of failure of more dominant pacemakers. Suppression of a pacemaker cell by a faster rhythm leads to an increased intracellular Na^+ load ($[Na^+]_i$), and extrusion of Na^+ from the cell by Na, K-ATPase produces an increased background repolarizing current that slows phase 4 diastolic depolarization. At

slower rates, $[Na^+]_i$ is decreased, as is the activity of the Na, K-ATPase, resulting in progressively more rapid diastolic depolarization and warm-up of the tachycardia rate. Overdrive suppression and warm-up are characteristic of, but may not be observed in, all automatic tachycardias. Abnormal conduction into tissue with enhanced automaticity (*entrance block*) may blunt or eliminate the phenomena of overdrive suppression and warm-up of automatic tissue.

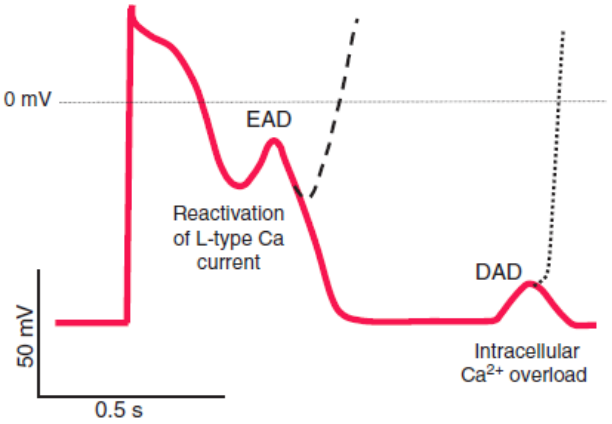
Abnormal automaticity may produce atrial tachycardia, accelerated idioventricular rhythms, and ventricular tachycardia, particularly associated with ischemia and reperfusion. It has also been suggested that injury currents at the borders of ischemic myocardium may depolarize adjacent nonischemic tissue, predisposing to automatic ventricular tachycardia.

Afterdepolarizations and Triggered Automaticity

Triggered automaticity or activity refers to impulse initiation that is dependent on afterdepolarizations **(Fig. 238-3)**. Afterdepolarizations are membrane voltage oscillations that occur during (early afterdepolarizations, EADs) or after (delayed afterdepolarizations, DADs) an action potential.

FIGURE 238-3

Schematic action potentials with early afterdepolarizations (EADs) and delayed afterdepolarizations (DADs). Afterdepolarizations are spontaneous depolarizations in cardiac myocytes. EADs occur before the end of the action potential (phases 2 and 3), interrupting repolarization. DADs occur during phase 4 of the action potential after completion of repolarization. The cellular mechanisms of EADs and DADs differ (see text).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The cellular feature common to the induction of DADs is the presence of an increased Ca^{2+} load in the cytosol and sarcoplasmic reticulum. Digitalis glycoside toxicity, catecholamines, and ischemia all can enhance Ca^{2+} loading sufficiently to produce DADs. Accumulation of lysophospholipids in ischemic myocardium with consequent Na^+ and Ca^{2+} overload has been suggested as a mechanism for DADs and triggered automaticity. Cells from damaged areas or cells that survive a myocardial infarction may display spontaneous release of calcium from the sarcoplasmic reticulum, and this may generate “waves” of intracellular calcium elevation and arrhythmias.

EADs occur during the action potential and interrupt the orderly repolarization of the myocyte. Traditionally, EADs have been thought to arise from action potential prolongation and reactivation of depolarizing currents, but more recent experimental evidence suggests a previously unappreciated interrelationship between intracellular calcium loading and EADs. Cytosolic calcium may increase when action potentials are prolonged. This, in turn, appears to enhance L-type Ca current, further prolonging action potential duration as well as providing the inward current driving EADs. Intracellular calcium loading by action potential prolongation may also enhance the likelihood of DADs. The interrelationship among intracellular $[Ca^{2+}]$, EADs, and DADs may be one explanation for the susceptibility of hearts that are calcium loaded (e.g., in ischemia or congestive heart failure [CHF]) to develop arrhythmias, particularly on exposure to action potential–prolonging drugs.

EAD-triggered arrhythmias exhibit rate dependence. In general, the amplitude of an EAD is augmented at slow rates when action potentials are longer. Indeed, a fundamental condition that underlies the development of EADs is action potential and QT prolongation. Hypokalemia, hypomagnesemia, bradycardia, and, most commonly, drugs can predispose to the generation of EADs, invariably in the context of prolonging the action potential. Antiarrhythmics with class IA and III action (see below) produce action potential and QT prolongation intended to be therapeutic but frequently causing arrhythmias. Noncardiac drugs such as phenothiazines, nonsedating antihistamines, and some antibiotics can also prolong the action potential duration and predispose to EAD-mediated triggered arrhythmias. Decreased $[K^+]_o$ paradoxically may decrease membrane potassium currents (particularly the delayed rectifier current, I_{Kr}) in the ventricular myocyte, explaining why hypokalemia causes action potential prolongation and EADs. In fact, potassium infusions in patients with the congenital long QT syndrome (LQTS) and in those with drug-induced acquired QT prolongation shorten the QT interval.

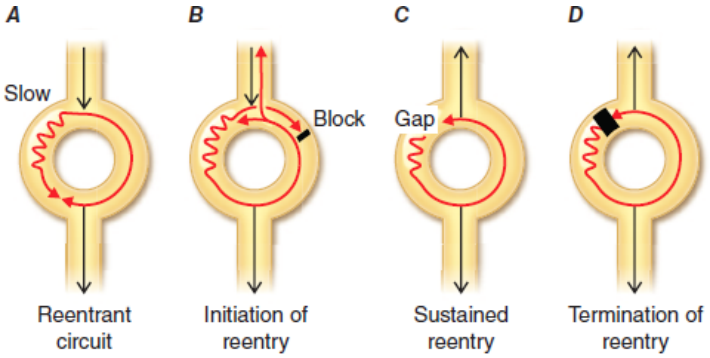
EAD-mediated triggered activity probably underlies initiation of the characteristic polymorphic ventricular tachycardia, torsades des pointes, seen in patients with congenital and acquired forms of LQTS. Structural heart disease, such as cardiac hypertrophy and heart failure, may also delay ventricular repolarization (so-called electrical remodeling) and predispose to arrhythmias related to abnormalities of repolarization. The abnormalities of repolarization in hypertrophy and heart failure are often magnified by concomitant drug therapy or electrolyte disturbances.

Abnormal Impulse Conduction: Reentry

The most common arrhythmia mechanism is reentry resulting from abnormal electrical impulse conduction and is defined as the circulation of an activation wave around an inexcitable obstacle. The requirements for reentry are two electrophysiologically dissimilar pathways for impulse propagation around an inexcitable region (**Fig. 238-4**). Reentry can occur around a fixed anatomic structure (e.g., myocardial scar), with a stable pattern of cardiac depolarization moving in series over the anterograde and retrograde limbs of the circuit. This form of reentry, referred to as *anatomic reentry* or *excitable gap reentry* (see below), is initiated when a depolarizing wavefront encounters an area of unidirectional conduction block in the retrograde limb of the circuit. Conduction across the anterograde limb occurs with a delay that, if of sufficient duration, allows for recovery of conduction in the retrograde limb with reentry of the depolarization wave into the retrograde limb of the circuit. Sustained reentry requires that the functional dimension of depolarized tissue or the tachycardia wavelength (λ = conduction velocity × refractory period) fits within the total anatomic length of the circuit, referred to as the path length. When the path length of the circuit exceeds the λ of the tachycardia, the region between the head of the activation wave and the refractory tail is referred to as the excitable gap. Anatomically determined, excitable gap reentry can explain several clinically important tachycardias, such as AV reentry, atrial flutter, bundle branch reentry ventricular tachycardia, and ventricular tachycardia in scarred myocardium.

FIGURE 238-4

Schematic diagram of reentry. *A.* The circuit contains two limbs, one with slow conduction. ***B.*** A premature impulse blocks in the fast pathway and conducts over the slow pathway, allowing the fast pathway to recover so that the activation wave can reenter the fast pathway from the retrograde direction. ***C.*** During sustained reentry utilizing such a circuit, a gap (excitable gap) exists between the activating head of the wave and the recovering tail. ***D.*** One mechanism of termination of reentry occurs when the conduction and recovery characteristics of the circuit change and the activating head of the wave collides with the tail, extinguishing the tachycardia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Reentrant arrhythmias may exist in the heart in the absence of an excitable gap and with a tachycardia wavelength nearly the same size as the path length. In this case, the wavefront propagates through partially refractory tissue without a fixed anatomic obstacle and no fully excitable gap; this is referred to as *leading circle reentry*, a form of functional reentry (reentry that depends on functional properties of the tissue). Unlike excitable gap reentry, there is no fixed anatomic circuit in leading circle reentry, and it may, therefore, not be possible to disrupt the tachycardia with pacing or destruction of a part of the circuit. Furthermore, the circuit in leading circle reentry tends to be less stable than that in excitable gap reentrant arrhythmias, with large variations in cycle length and a predilection to termination. There is strong evidence to suggest that less organized arrhythmias, such as atrial and ventricular fibrillation, are associated with more complex activation of the heart and are due to functional reentry.

Catheter-based and pharmacologic therapies for reentrant arrhythmias are designed to disrupt the anatomic circuit or alter the relationship between the wavelength and path length of the arrhythmia circuit, eliminating pathologic conduction. For example, antiarrhythmic drugs that prolong the action potential (Class III) are effective if they sufficiently prolong the λ such that it can no longer fit within the anatomic circuit. Catheter ablation is often undertaken with the goal of identifying and destroying a critical limb of the reentrant circuit (i.e., ablation of the cavotricuspid isthmus in the treatment of typical, right atrial flutter). Due to the less defined pathways of myocardial activation seen in functional reentry, ablation of these rhythms tends to target initiating triggers (e.g., pulmonary vein potentials in catheter ablation of atrial fibrillation) rather than the anatomic circuit.

Structural heart disease is associated with changes in conduction and refractoriness that increase the risk of reentrant arrhythmias. Chronically ischemic myocardium exhibits a downregulation of the gap junction channel protein (connexin 43) that carries intercellular ionic current. The border zones of infarcted and failing ventricular myocardium exhibit not only functional alterations of ionic currents but also remodeling of tissue and altered distribution of gap junctions. The changes in gap junction channel expression and distribution, in combination with macroscopic tissue alterations, support a role for slowed conduction in reentrant arrhythmias that complicate chronic coronary artery disease (CAD). Aged human atrial myocardium exhibits altered conduction, manifest as highly fractionated atrial electrograms, producing an ideal substrate for the reentry that may underlie the very common development of atrial fibrillation in the elderly.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Cardiac Arrhythmias

The evaluation of patients with suspected cardiac arrhythmias is highly individualized; however, two key features—the history and ECG—are pivotal in directing the diagnostic workup and therapy. Patients with cardiac arrhythmias exhibit a wide spectrum of clinical presentations that range from asymptomatic ECG abnormalities to survival from cardiac arrest. In general, the more severe the presenting

symptoms are, the more aggressive the evaluation and treatment are. Loss of consciousness that is believed to be of cardiac origin typically mandates an exhaustive search for the etiology and often requires invasive, device-based therapy. The presence of structural heart disease and prior myocardial infarction dictates a change in the approach to the management of syncope or ventricular arrhythmias. The presence of a family history of serious ventricular arrhythmias or premature sudden death will influence the evaluation of presumed heritable arrhythmias.

The physical examination is focused on determining whether there is cardiopulmonary disease that is associated with specific cardiac arrhythmias. The absence of significant cardiopulmonary disease often, but not always, suggests benignity of the rhythm disturbance. In contrast, palpitations, syncope, or near syncope in the setting of significant heart or lung disease have more ominous implications. In addition, the physical examination may reveal the presence of a persistent arrhythmia such as atrial fibrillation.

The judicious use of noninvasive diagnostic tests is an important element in the evaluation of patients with arrhythmias, and there is no test more important than the ECG, particularly if recorded at the time of symptoms. Uncommon but diagnostically important signatures of electrophysiologic disturbances may be unearthed on the resting ECG, such as delta waves in Wolff-Parkinson-White (WPW) syndrome, prolongation or shortening of the QT interval, right precordial ST-segment abnormalities in Brugada syndrome, and epsilon waves in arrhythmogenic right ventricular dysplasia. Variants of body surface ECG recording can provide important information about arrhythmia substrates and triggers. Holter monitoring and event recording, either continuous or intermittent, record the body surface ECG over longer periods, enhancing the possibility of observing the cardiac rhythm during symptoms. Holter monitoring is particularly useful in assessing daily symptoms thought to be attributable to arrhythmia or for quantifying a particular arrhythmia phenomenon (e.g., premature ventricular complex burden). Ambulatory event monitors are indicated when symptoms thought to be due to arrhythmia occur less frequently (i.e., several episodes per month), and, because the monitors are typically patient-activated, they are optimal for correlating symptoms with rhythm disturbances. Implantable long-term monitors permit prolonged telemetric monitoring both for diagnosis and to assess the efficacy of therapy. Implantable monitors are typically used for the evaluation of malignant symptoms that occur quite infrequently and that cannot be provoked at diagnostic electrophysiology study and increasingly assessing the presence or burden of atrial fibrillation in patients with complications of this arrhythmia, such as stroke.

Exercise electrocardiography is important in determining the presence of myocardial demand ischemia; more recently, analysis of the morphology of the QT interval with exercise has been used to assess the risk of serious ventricular arrhythmias. The exercise ECG may be particularly useful in patients with symptoms that occur during activity. Cardiac imaging plays an important role in the detection and characterization of myocardial structural abnormalities that may render the heart more susceptible to arrhythmia. Ventricular tachyarrhythmias, for instance, occur more frequently in patients with ventricular systolic dysfunction and chamber dilation, in hypertrophic cardiomyopathy, and in the setting of infiltrative diseases such as sarcoidosis. Supraventricular arrhythmias may be associated with particular congenital conditions, including AV reentry in the setting of Ebstein’s anomaly. Echocardiography is a frequently employed imaging technique to screen for disorders of cardiac structure and function. Increasingly, magnetic resonance imaging of the myocardium is being used to screen for scar burden, fibrofatty infiltration of the myocardium as seen in arrhythmogenic right ventricular cardiomyopathy, and other structural changes that affect arrhythmia susceptibility.

Head-up tilt (HUT) testing is useful in the evaluation of patients with syncope in whom there is a suspicion that exaggerated vagal tone or vasodepression may play a causal role. The physiologic response to HUT is incompletely understood; however, redistribution of blood volume and increased ventricular contractility occur consistently. Exaggerated activation of a central reflex in response to HUT produces a stereotypic response of an initial increase in heart rate, then a drop in blood pressure followed by a reduction in heart rate characteristic of neurally mediated hypotension. Other responses to HUT may be observed in patients with orthostatic hypotension and autonomic insufficiency. HUT is used most often in patients with recurrent syncope, although it may be useful in patients with single syncopal episodes with associated injury, particularly in the absence of structural heart disease. In patients with structural heart disease, HUT may be indicated in those with syncope, in whom other causes (e.g., asystole, ventricular tachyarrhythmias) have been excluded. HUT has been suggested as a useful tool in the diagnosis of and therapy for recurrent idiopathic vertigo, chronic fatigue syndrome, recurrent transient ischemic attacks, and repeated falls of unknown etiology in the elderly. Importantly, HUT is relatively contraindicated in the presence of severe CAD with proximal coronary stenoses, known severe cerebrovascular disease, severe mitral stenosis, and obstruction to left ventricular outflow (e.g., aortic stenosis).

Electrophysiologic testing is central to the understanding and treatment of many cardiac arrhythmias. Indeed, most frequently, electrophysiologic testing is interventional, providing both diagnosis and therapy. The indications for electrophysiologic testing fall into several categories: to define the mechanism of an arrhythmia; to deliver catheter-based ablative treatment; and to determine the etiology of symptoms that may be caused by an arrhythmia (e.g., syncope, palpitations). The components of the electrophysiologic test are baseline measurements of conduction under resting and stressed (rate or pharmacologic) conditions and maneuvers, both pacing and pharmacologic, to induce arrhythmias. A number of sophisticated electrical mapping and catheter-guidance techniques have been developed to facilitate catheter-based therapeutics in the electrophysiology laboratory.

TREATMENT

TREATMENT Cardiac Arrhythmias ANTIARRHYTHMIC DRUG THERAPY

The interaction of antiarrhythmic drugs with cardiac tissues and the resulting electrophysiologic changes are complex. An incomplete understanding of the effects of these drugs has produced serious missteps that have had adverse effects on patient outcomes and the development of newer pharmacologic agents. Currently, antiarrhythmic drugs have been relegated to an ancillary role in the treatment of most cardiac arrhythmias.

There are several explanations for the complexity of antiarrhythmic drug action: the structural similarity of target ion channels; regional differences in the levels of expression of channels and transporters, which change with disease; time and voltage dependence of drug action; and the effect of these drugs on targets other than ion channels. Because of the limitations of any scheme to classify antiarrhythmic agents, a shorthand that is useful in describing the major mechanisms of action is of some utility. Such a classification scheme was proposed in 1970 by Vaughan-Williams and later modified by Singh and Harrison. The classes of antiarrhythmic action are class I, local anesthetic effect due to blockade of Na^+ current; class II, interference with the action of catecholamines at the β-adrenergic receptor; class

III, delay of repolarization due to inhibition of K^+ current or activation of depolarizing current; and class IV, interference with calcium conductance **(Table 238-2)**. Class I antiarrhythmics have been further subdivided based on the kinetics and potency of Na^+ channel binding; class Ia agents (quinidine, procainamide) are those with moderate potency and intermediate kinetics; class Ib agents (lidocaine, mexiletine) are those with low potency and rapid kinetics; and class Ic drugs (flecainide, propafenone) are those with high potency and the slowest kinetics. The limitations of the Vaughan-Williams classification scheme include multiple actions of most drugs, overwhelming consideration of antagonism as a mechanism of action, and the fact that several agents have none of the four classes of action in the scheme.

CATHETER ABLATION

The use of catheter ablation is based on the principle that there is a critical anatomic region of impulse generation or propagation that is required for the initiation and maintenance of cardiac arrhythmias. Destruction of such a critical region results in the elimination of the arrhythmia. The use of RF energy in clinical medicine is nearly a century old. The first catheter ablation using a DC energy source was performed in the early 1980s by Scheinman and colleagues. By the early 1990s, RF had been adapted for use in catheter-based ablation in the heart **(Fig. 238-5)**.

The RF band (300–30,000 kHz) is used to generate energy for several biomedical applications, including coagulation and cauterization of tissues. Energy of this frequency will not stimulate skeletal muscle or the heart and heats tissue by a resistive mechanism, with the intensity of heating and tissue destruction being proportional to the delivered power. Alternative, less frequently used energy sources for catheter ablation of cardiac arrhythmias include microwaves (915 MHz or 2450 MHz), lasers, ultrasound, and freezing (cryoablation). Of these alternative ablation techniques, cryoablation is being used clinically with the most frequency, especially ablation in the region of the AVN. At temperatures just below 32°C, membrane ion transport is disrupted, producing depolarization of cells, decreased action potential amplitude and duration, and slowed conduction velocity (resulting in local conduction block)—all of which are reversible if the tissue is rewarmed in a timely fashion. Tissue cooling can be used for mapping and ablation. Cryomapping can be used to confirm the location of a desired ablation target, such as an accessory pathway in WPW syndrome, or can be used to determine the safety of ablation around the AVN by monitoring AV conduction during cooling. Another advantage of cryoablation is that once the catheter tip cools below freezing, it adheres to the tissue, increasing catheter stability independent of the rhythm or pacing.

DEVICE THERAPY

Bradyarrhythmias due either to primary sinus node dysfunction or to AV conduction defects are readily treated through implantation of a permanent pacemaker. Clinical indications for pacemaker implantation often depend on the presence either of symptomatic bradycardia or of an unreliable endogenous escape rhythm and are more fully reviewed in **Chaps. 239** and **240**.

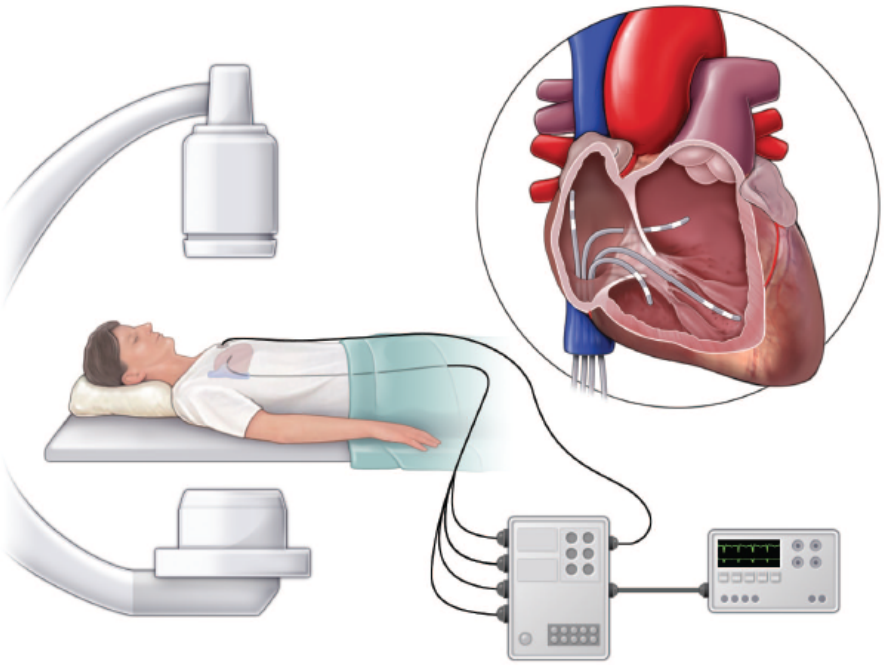
Ventricular tachyarrhythmias, particularly those occurring in the context of progressive structural heart diseases such as ischemic cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy, may recur despite therapy with antiarrhythmic drugs or catheter ablation. In appropriate candidates, implantation of an internal cardioverter-defibrillator (ICD) may reduce mortality rates from sudden cardiac death. In a subset of patients with CHF and ventricular mechanical dyssynchrony, ICD, or pacemaker platforms can be used to provide cardiac resynchronization therapy, typically through implantation of a left ventricular pacing lead. In patients with dyssynchronous CHF, such therapy has been shown to improve both morbidity and mortality rates. The use of a completely subcutaneous ICD may be most appropriate in patients at risk for arrhythmic sudden death without a need for pacing.

TABLE 238-2 Antiarrhythmic Drug Actions

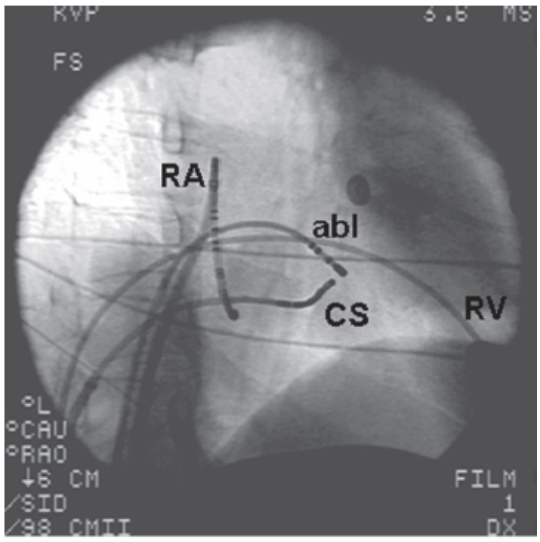
DRUG	CLASS ACTIONS				MISCELLANEOUS ACTION
	I	II	III	IV	
Quinidine	++		++		α-Adrenergic blockade
Procainamide	++		++		Ganglionic blockade
Flecainide	+++		+		
Propafenone	++	+			
Ranolazine	++		++		Late Na^+ current blockade
Eleclazine	++				Late Na^+ current blockade
Sotalol		++	+++		
Dofetilide			+++		
Amiodarone	++	++	+++	+	α-Adrenergic blockade
Dronedarone	++	+	+++	++	HCN4 blockade
Ibutilide			+++		Na^+ channel activator

FIGURE 238-5

Catheter ablation of cardiac arrhythmias. ***A.*** A schematic of the catheter system and generator in a patient undergoing radiofrequency catheter ablation (RFCA); the circuit involves the catheter in the heart and a dispersive patch placed on the body surface (usually the back). The inset shows a diagram of the heart with a catheter located at the AV valve ring for ablation of an accessory pathway. ***B.*** A right anterior oblique fluoroscopic image of the catheter position for ablation of a left-sided accessory pathway. A catheter is placed in the atrial side of the mitral valve ring (abl) via a transseptal puncture. Other catheters are placed in the coronary sinus (CS), in the right atrium (RA), and in the right ventricular (RV) apex to record local electrical activation. ***C.*** Body surface electrocardiogram recordings (I, II, V_1) and endocardial electrograms (HRA, high right atrium; HISp, proximal His bundle electrogram; CS 7, 8, recordings from poles 7 and 8 of a decapolar catheter placed in the coronary sinus) during RFCA of a left-sided accessory pathway in a patient with Wolff-Parkinson-White syndrome. The QRS narrows at the fourth complex; the *arrow* shows the His bundle electrogram, which becomes apparent with elimination of ventricular preexcitation over the accessory pathway.

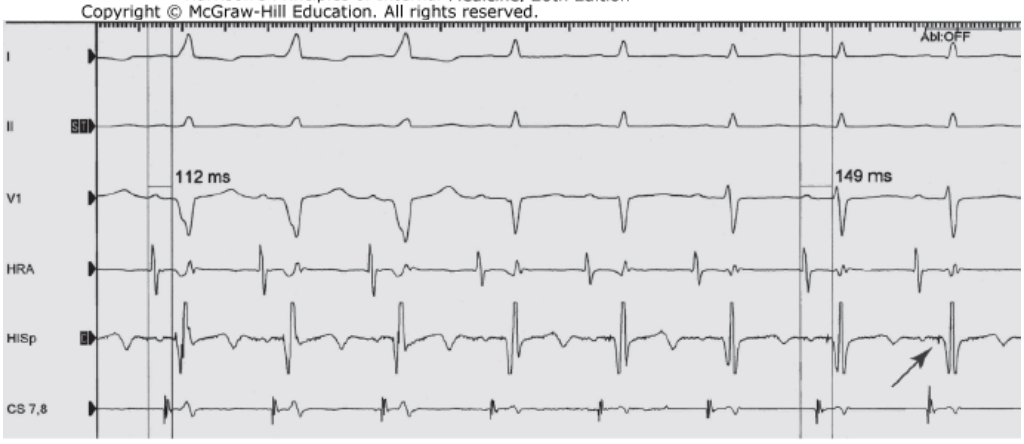


A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Josephson ME: *Clinical Cardiac Electrophysiology: Techniques and Interpretations*, 5th ed. Philadelphia, Wolters Kluwer, 2016.
Priori SG, Napolitano C: Genetics of channelopathies and clinical implications, in *Hurst's The Heart*, 13th ed, V Fuster et al (eds). New York, McGraw-Hill, 2011.
Saksena S, Camm (eds): *Electrophysiological Disorders of the Heart*, 2nd ed. Philadelphia, Elsevier Churchill Livingstone, 2012.
Zipes DP, Jalife J (eds): *Cardiac Electrophysiology: From Cell to Bedside*, 6th ed. Philadelphia, Elsevier, 2014.

Chapter 239: The Bradyarrhythmias: Disorders of the Sinoatrial Node

David D. Spragg; Gordon F. Tomaselli

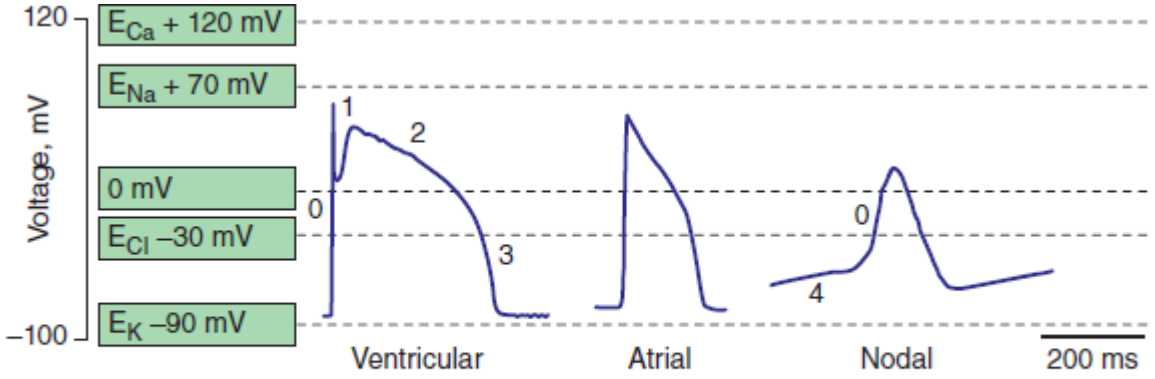
INTRODUCTION

Electrical activation of the heart normally originates in the sinoatrial (SA) node, the predominant pacemaker. Other subsidiary pacemakers in the atrioventricular (AV) node, specialized conducting system, and muscle may initiate electrical activation if the SA node is dysfunctional or suppressed. Typically, subsidiary pacemakers discharge at a slower rate and, in the absence of an appropriate increase in stroke volume, may result in tissue hypoperfusion.

Spontaneous activation and contraction of the heart are a consequence of the specialized pacemaking tissue in these anatomic locales. As described in **Chap. 238**, action potentials in the heart are regionally heterogeneous. The action potentials in cells isolated from nodal tissue are distinct from those recorded from atrial and ventricular myocytes **(Fig. 239-1)**. The complement of ionic currents present in nodal cells results in a less negative resting membrane potential compared with atrial or ventricular myocytes. Electrical diastole in nodal cells is characterized by slow diastolic depolarization (phase 4), which generates an action potential as the membrane voltage reaches threshold. The action potential upstrokes (phase 0) are slow compared with atrial or ventricular myocytes, being mediated by calcium rather than sodium current. Cells with properties of SA and AV nodal tissue are electrically connected to the remainder of the myocardium by cells with an electrophysiologic phenotype between that of nodal cells and that of atrial or ventricular myocytes. Cells in the SA node exhibit the most rapid phase 4 depolarization and thus are the dominant pacemakers in a normal heart.

FIGURE 239-1

Action potential profiles recorded in cells isolated from sinoatrial or atrioventricular nodal tissue compared with those of cells from atrial or ventricular myocardium. Nodal cell action potentials exhibit more depolarized resting membrane potentials, slower phase 0 upstrokes, and phase 4 diastolic depolarization.

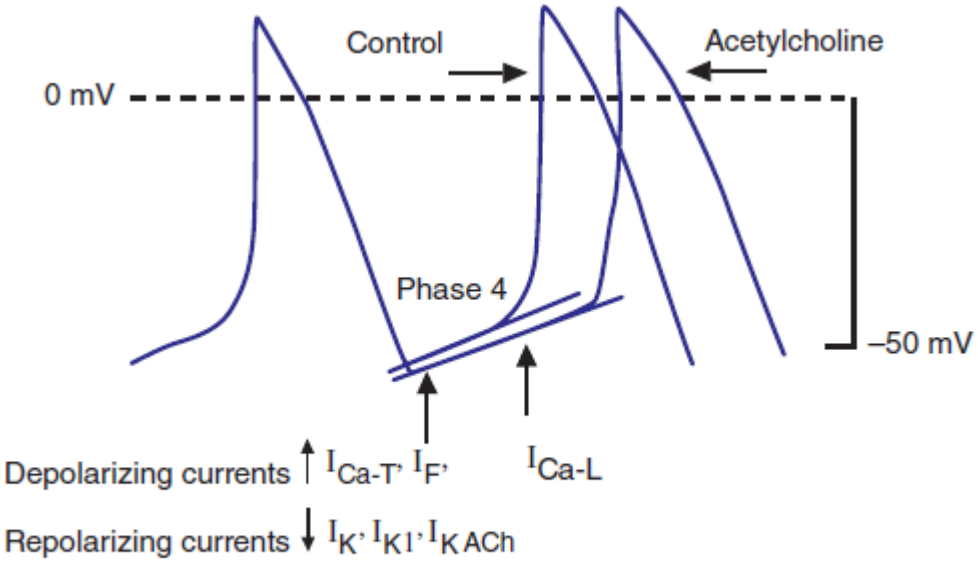


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Bradycardia results from a failure of either impulse initiation or impulse conduction. Failure of impulse initiation may be caused by depressed automaticity resulting from a slowing or failure of phase 4 diastolic depolarization **(Fig. 239-2)**, which may result from disease or exposure to drugs. Prominently, the autonomic nervous system modulates the rate of phase 4 diastolic depolarization and thus the firing rate of both primary (SA node) and subsidiary pacemakers. Failure of conduction of an impulse from nodal tissue to atrial or ventricular myocardium may produce bradycardia as a result of exit block. Conditions that alter the activation and connectivity of cells (e.g., fibrosis) in the heart may result in failure of impulse conduction.

FIGURE 239-2

Schematics of nodal action potentials and the currents that contribute to phase 4 depolarization. Relative increases in depolarizing L- ($I_{Ca\text{-}L}$) and T- ($I_{Ca\text{-}T}$) type calcium and pacemaker currents (I_f) along with a reduction in repolarizing inward rectifier (I_{K1}) and delayed rectifier (I_K) potassium currents result in depolarization. Activation of ACh-gated (I_{KACh}) potassium current and beta blockade slow the rate of phase 4 and decrease the pacing rate. *(Modified from J Jalife et al: Basic Cardiac Electrophysiology for the Clinician, Blackwell Publishing, 1999.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SA node dysfunction and AV conduction block are the most common causes of pathologic bradycardia. SA node dysfunction may be difficult to distinguish from physiologic sinus bradycardia,

particularly in the young. SA node dysfunction increases in frequency between the fifth and sixth decades of life and should be considered in patients with fatigue, exercise intolerance, or syncope and sinus bradycardia.

Permanent pacemaking is the only reliable therapy for symptomatic bradycardia in the absence of extrinsic and reversible etiologies such as increased vagal tone, hypoxia, hypothermia, and drugs **(Table 239-1)**. Approximately 50% of the 160,000 permanent pacemakers implanted in the United States and 20–30% of those in Europe were implanted for SA node disease.

TABLE 239-1 Etiologies of SA Node Dysfunction

EXTRINSIC	INTRINSIC
Autonomic	Sick-sinus syndrome (SSS)
Carotid sinus hypersensitivity	Coronary artery disease (chronic and acute MI)
Vasovagal (cardioinhibitory) stimulation	Inflammatory
Drugs	Pericarditis
Beta blockers	Myocarditis (including viral)
Calcium channel blockers	Rheumatic heart disease
Digoxin	Collagen vascular diseases
Ivabradine	Lyme disease
Antiarrhythmics (class I and III)	Senile amyloidosis
Adenosine	Congenital heart disease
Clonidine (other sympatholytics)	TGA/Mustard and Fontan repairs
Lithium carbonate	Iatrogenic
Cimetidine	Radiation therapy
Amitriptyline	Postsurgical
Phenothiazines	Chest trauma
Narcotics (methadone)	Familial
Pentamidine	SSS2, AD, OMIM #163800 (15q24-25)
Hypothyroidism	SSS1, AR OMIM #608567 (3p21)
Sleep apnea	SSS3, AD, OMIM #614090 (14q11.2)
Hypoxia	SA node disease with myopia, OMIM #182190
Endotracheal suctioning (vagal maneuvers)	Kearns-Sayre syndrome, OMIM #530000
Hypothermia	Myotonic dystrophy
Increased intracranial pressure	Type 1, OMIM #160900 (19q13.2-13.3)
	Type 2, OMIM #602668 (3q13.3-q24)
	Friedreich's ataxia, OMIM #229300 (9q13, 9p23-p11)

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; MI, myocardial infarction; OMIM, Online Mendelian Inheritance in Man (database); TGA, transposition of the great arteries.

STRUCTURE AND PHYSIOLOGY OF THE SA NODE

The SA node is composed of a cluster of small fusiform cells in the sulcus terminalis on the epicardial surface of the heart at the right atrial–superior vena caval junction, where they envelop the SA nodal artery. The SA node is structurally heterogeneous, but the central prototypic nodal cells have fewer distinct myofibrils than does the surrounding atrial myocardium, no intercalated disks visible on light microscopy, a poorly developed sarcoplasmic reticulum, and no T-tubules. Cells in the peripheral regions of the SA node are transitional in both structure and function. The SA nodal artery arises from the right coronary artery in 55–60% and the left circumflex artery in 40–45% of persons. The SA node is richly innervated by sympathetic and parasympathetic nerves and ganglia.

Irregular and slow propagation of impulses from the SA node can be explained by the electrophysiology of nodal cells and the structure of the SA node itself. The action potentials of SA nodal cells are characterized by a relatively depolarized membrane potential (Fig. 239-1) of −40 to −60 mV, slow phase 0 upstroke, and relatively rapid phase 4 diastolic depolarization compared with the action potentials recorded in cardiac muscle cells. The relative absence of inward rectifier potassium current (I_{K1}) accounts for the depolarized membrane potential; the slow upstroke of phase 0 results from the absence of available fast sodium current (I_{Na}) and is mediated by L-type calcium current ($I_{Ca\text{-}L}$); and phase 4 depolarization is a result of the aggregate activity of a number of ionic currents. Prominently, both L- and T-type ($I_{Ca\text{-}T}$) calcium currents, the pacemaker current (so-called funny current, or I_f) formed by hyperpolarization-activated cyclic nucleotide-gated channels, and the electrogenic sodium-calcium exchanger provide depolarizing current that is antagonized by delayed rectifier (I_{Kr}) and acetylcholine-gated (I_{KACh}) potassium currents. $I_{Ca\text{-}L}$, $I_{Ca\text{-}T}$, and I_f are modulated by β-adrenergic stimulation and I_{KACh} by vagal stimulation, explaining the exquisite sensitivity of diastolic depolarization to autonomic nervous system activity. The slow conduction within the SA node is explained by the absence of I_{Na} and poor electrical coupling of cells in the node, resulting from sizable amounts of interstitial tissue and a low abundance of gap junctions. The poor coupling allows for graded electrophysiologic properties within the node, with the peripheral transitional cells being silenced by electrotonic coupling to atrial myocardium.

ETIOLOGY OF SA NODAL DISEASE

SA nodal dysfunction has been classified as intrinsic or extrinsic. The distinction is important because extrinsic dysfunction is often reversible and generally should be corrected before pacemaker therapy is considered (Table 239-1). The most common causes of extrinsic SA node dysfunction are drugs and autonomic nervous system influences that suppress automaticity and/or compromise conduction. Other extrinsic causes include hypothyroidism, sleep apnea, and conditions likely to occur in critically ill patients such as hypothermia, hypoxia, increased intracranial pressure (Cushing's response), and endotracheal suctioning via activation of the vagus nerve.

Intrinsic sinus node dysfunction is degenerative and often is characterized pathologically by fibrous replacement of the SA node or its connections to the atrium. Acute and chronic coronary artery disease (CAD) may be associated with SA node dysfunction, although in the setting of acute myocardial infarction (MI; typically inferior), the abnormalities are transient. Inflammatory processes may alter SA node function, ultimately producing replacement fibrosis. Pericarditis, myocarditis, and rheumatic heart disease have been associated with SA nodal disease with sinus bradycardia, sinus arrest, and exit block. Carditis associated with systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), and mixed connective tissue disorders (MCTDs) may also affect SA node structure and function. Senile amyloidosis is an infiltrative disorder in patients typically in the ninth decade of life; deposition of amyloid protein in the atrial myocardium can impair SA node function. Some SA node disease is iatrogenic and results from direct injury to the SA node during cardiothoracic surgery.

Rare heritable forms of sinus node disease have been described, and several have been characterized genetically. Autosomal dominant sinus node dysfunction in conjunction with

supraventricular tachycardia (i.e., tachycardia-bradycardia variant of sick-sinus syndrome [SSS2]) has been linked to mutations in the pacemaker current (I_f) subunit gene *HCN4* on chromosome 15. An autosomal recessive form of SSS1 with the prominent feature of atrial inexcitability and absence of P waves on the electrocardiogram (ECG) is caused by mutations in the cardiac sodium channel gene, *SCN5A*, on chromosome 3. Variants in myosin heavy chain 6 (*MYH6*) increase the susceptibility to SSS (SSS3). SA node dysfunction associated with myopia has been described but not genetically characterized. There are several neuromuscular diseases, including Kearns-Sayre syndrome (ophthalmoplegia, pigmentary degeneration of the retina, and cardiomyopathy) and myotonic dystrophy that have a predilection for the conducting system and SA node.

SSS in both the young and the elderly is associated with an increase in fibrous tissue in the SA node. The onset of SSS may be hastened by coexisting disease, such as CAD, diabetes mellitus, hypertension, and valvular diseases and cardiomyopathies.

CLINICAL FEATURES OF SA NODE DISEASE

SA node dysfunction may be completely asymptomatic and manifest as an ECG anomaly such as sinus bradycardia; sinus arrest and exit block; or alternating supraventricular tachycardia, usually atrial fibrillation, and bradycardia. Symptoms associated with SA node dysfunction, in particular tachycardia-bradycardia syndrome, may be related to both slow and fast heart rates. For example, tachycardia may be associated with palpitations, angina pectoris, and heart failure, and bradycardia may be associated with hypotension, syncope, presyncope, fatigue, and weakness. In the setting of SSS, overdrive suppression of the SA node may result in prolonged pauses and syncope upon termination of the tachycardia. In many cases, symptoms associated with SA node dysfunction result from concomitant cardiovascular disease. A significant minority of patients with SSS develop signs and symptoms of heart failure that may be related to slow or fast heart rates.

One-third to one-half of patients with SA node dysfunction develop supraventricular tachycardia, usually atrial fibrillation or atrial flutter. The incidence of persistent atrial fibrillation in patients with SA node dysfunction increases with advanced age, hypertension, diabetes mellitus, left ventricular dilation, valvular heart disease, and ventricular pacing. Remarkably, some symptomatic patients may experience an improvement in symptoms with the development of atrial fibrillation, presumably from an increase in their average heart rate. Patients with the tachycardia-bradycardia variant of SSS, similar to patients with atrial fibrillation, are at risk for thromboembolism, and *those at greatest risk*, including patients aged ≥65 years and patients with a prior history of stroke, valvular heart disease, left ventricular dysfunction, or atrial enlargement, should be treated with anticoagulants. Up to one-quarter of patients with SA node disease will have concurrent AV conduction disease, although only a minority will require specific therapy for high-grade AV block.

The natural history of SA node dysfunction is one of varying intensity of symptoms even in patients who present with syncope. Symptoms related to SA node dysfunction may be significant, but overall longevity usually is not compromised in the absence of other significant comorbid conditions. These features of the natural history need to be taken into account in considering therapy for these patients.

ELECTROCARDIOGRAPHY OF SA NODE DISEASE

The electrocardiographic manifestations of SA node dysfunction include sinus bradycardia, sinus pauses, sinus arrest, sinus exit block, tachycardia (in SSS), and chronotropic incompetence. It is often difficult to distinguish pathologic from physiologic sinus bradycardia. By definition, sinus bradycardia is a rhythm driven by the SA node with a rate of <60 beats/min; sinus bradycardia is very common and typically benign. Resting heart rates <60 beats/min are very common in young healthy individuals and physically conditioned subjects. A sinus rate of <40 beats/min in the awake state in the absence of physical conditioning generally is considered abnormal. Sinus pauses and sinus arrest result from

failure of the SA node to discharge, producing a pause without P waves visible on the ECG **(Fig. 239-3)**. Sinus pauses of up to 3 s are common in awake athletes, and pauses of this duration or longer may be observed in asymptomatic elderly subjects. Intermittent failure of conduction from the SA node produces sinus exit block. The severity of sinus exit block may vary in a manner similar to that of AV block **(Chap. 240)**. Prolongation of conduction from the sinus node will not be apparent on the ECG; second-degree SA block will produce intermittent conduction from the SA node and a regularly irregular atrial rhythm.

FIGURE 239-3

Sinus slowing and pauses on the electrocardiogram (ECG). The ECG is recorded during sleep in a young patient without heart disease. The heart rate before the pause is slow, and the PR interval is prolonged, consistent with an increase in vagal tone. The P waves have a morphology consistent with sinus rhythm. The recording is from a two-lead telemetry system in which the tracing labeled II mimics frontal lead II and V represents Modified Central Lead 1, which mimics lead V_1 of the standard 12-lead ECG.

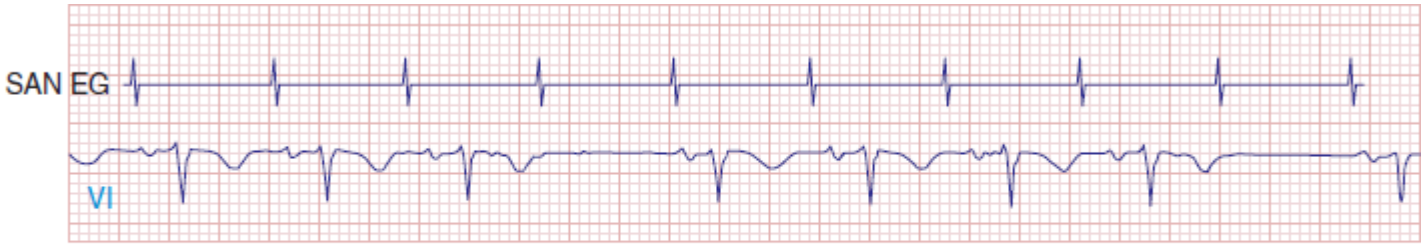


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Type I second-degree SA block results from progressive prolongation of SA node conduction with intermittent failure of the impulses originating in the sinus node to conduct to the surrounding atrial tissue. Second-degree SA block appears on the ECG as an intermittent absence of P waves **(Fig. 239-4)**. In type II second-degree SA block, there is no change in SA node conduction before the pause. Complete or third-degree SA block results in no P waves on the ECG. Tachycardia-bradycardia syndrome is manifest as alternating sinus bradycardia and atrial tachyarrhythmias. Although atrial tachycardia, atrial flutter, and atrial fibrillation may be observed, the latter is the most common tachycardia. Chronotropic incompetence is the inability to increase the heart rate in response to exercise or other stress appropriately and is defined in greater detail below.

FIGURE 239-4

Mobitz type I SA nodal exit block. A theoretical SA node electrogram (SAN EG) is shown. Note that there is grouped beating producing a regularly irregular heart rhythm. The SA node EG rate is constant with progressive delay in exit from the node and activation of the atria, inscribing the P wave. This produces subtly decreasing P-P intervals before the pause, and the pause is less than twice the cycle length of the last sinus interval.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSTIC TESTING

SA node dysfunction is most commonly a clinical or electrocardiographic diagnosis. Sinus bradycardia or pauses on the resting ECG are rarely sufficient to diagnose SA node disease, and longer-term recording and symptom correlation generally are required. Symptoms in the absence of sinus bradyarrhythmias may be sufficient to exclude a diagnosis of SA node dysfunction.

Electrocardiographic recording plays a central role in the diagnosis and management of SA node dysfunction. Despite the limitations of the resting ECG, longer-term recording employing Holter or event monitors may permit correlation of symptoms with the cardiac rhythm. Many contemporary event monitors may be automatically triggered to record the ECG when certain programmed heart rate criteria are met. Implantable ECG monitors permit long-term recording (12–18 months) in particularly challenging patients.

Failure to increase the heart rate with exercise is referred to as *chronotropic incompetence*. This is alternatively defined as failure to reach 85% of predicted maximal heart rate at peak exercise or failure to achieve a heart rate >100 beats/min with exercise or a maximal heart rate with exercise less than two standard deviations below that of an age-matched control population. Exercise testing may be useful in discriminating chronotropic incompetence from resting bradycardia and may aid in the identification of the mechanism of exercise intolerance.

Autonomic nervous system testing is useful in diagnosing carotid sinus hypersensitivity; pauses >3 s are consistent with the diagnosis but may be present in asymptomatic elderly subjects. Determining the intrinsic heart rate (IHR) may distinguish SA node dysfunction from slow heart rates that result from high vagal tone. The normal IHR after administration of 0.2 mg/kg propranolol and 0.04 mg/kg atropine is 117.2 − (0.53 × age) in beats/min; a low IHR is indicative of SA disease.

Electrophysiologic testing may play a role in the assessment of patients with presumed SA node dysfunction and in the evaluation of syncope, particularly in the setting of structural heart disease. In this circumstance, electrophysiologic testing is used to rule out more malignant etiologies of syncope, such as ventricular tachyarrhythmias and AV conduction block. There are several ways to assess SA node function invasively. They include the sinus node recovery time (SNRT), defined as the longest pause after cessation of overdrive pacing of the right atrium near the SA node (normal: <1500 ms or, corrected for sinus cycle length, <550 ms), and the sinoatrial conduction time (SACT), defined as one-half the difference between the intrinsic sinus cycle length and a noncompensatory pause after a premature atrial stimulus (normal <125 ms). The combination of an abnormal SNRT, an abnormal SACT, and a low IHR is a sensitive and specific indicator of intrinsic SA node disease.

TREATMENT

TREATMENT Sinoatrial Node Dysfunction

Since SA node dysfunction is not associated with increased mortality rates, the aim of therapy is alleviation of symptoms. Exclusion of extrinsic causes of SA node dysfunction and correlation of the cardiac rhythm with symptoms is an essential part of patient management. Pacemaker implantation is the primary therapeutic intervention in patients with symptomatic SA node dysfunction. Pharmacologic considerations are important in the evaluation and management of patients with SA nodal disease. A number of drugs modulate SA node function and are extrinsic causes of dysfunction (Table 239-1). Beta blockers and calcium channel blockers increase SNRT in patients with SA node dysfunction, and antiarrhythmic drugs with class I and III action may promote SA node exit block. In general, such agents should be discontinued before decisions regarding the need for permanent pacing in patients with SA node disease are made. Chronic pharmacologic therapy for sinus bradyarrhythmias is limited. Some pharmacologic agents may improve SA node function; digitalis, for example, has been shown to shorten SNRT in patients with SA node dysfunction. Isoproterenol or atropine administered IV may increase the sinus rate acutely. Theophylline has been used both acutely and chronically to increase heart rate but has liabilities when used in patients with tachycardia-bradycardia syndrome, increasing the frequency of supraventricular tachyarrhythmias, and in patients with structural heart disease, increasing the risk of potentially serious ventricular arrhythmias. Currently, there is only a single randomized study of therapy for SA node dysfunction. In patients with resting heart rates <50 and >30 beats/min on a Holter monitor, patients who received dual-chamber pacemakers experienced significantly fewer syncopal episodes and had symptomatic improvement compared with patients randomized to theophylline or no treatment.

In certain circumstances, sinus bradycardia requires no specific treatment or only temporary rate support. Sinus bradycardia is common in patients with acute inferior or posterior MI and can be exacerbated by vagal activation induced by pain or the use of drugs such as morphine. Ischemia of the SA nodal artery probably occurs in acute coronary syndromes more typically with involvement with the right coronary artery, and even with infarction, the effect on SA node function most often is transient.

Sinus bradycardia is a prominent feature of carotid sinus hypersensitivity and neurally mediated hypotension associated with vasovagal syncope that responds to pacemaker therapy. Carotid hypersensitivity with recurrent syncope or presyncope associated with a predominant cardioinhibitory component responds to pacemaker implantation. Several randomized trials have investigated the efficacy of permanent pacing in patients with drug-refractory vasovagal syncope, with mixed results. Although initial trials suggested that patients undergoing pacemaker implantation have fewer recurrences and a longer time to recurrence of symptoms, at least one follow-up study did not confirm these results.

PERMANENT PACEMAKERS Nomenclature and Complications

The main therapeutic intervention in SA node dysfunction is permanent pacing. Since the first implementation of permanent pacing in the 1950s, many advances in technology have resulted in miniaturization, increased longevity of pulse generators, improvement in leads, and increased functionality. To better understand pacemaker therapy for bradycardias, it is important to be familiar with the fundamentals of pacemaking. Pacemaker modes and function are named using a five-letter code. The first letter indicates the chamber(s) that is paced (O, none; A, atrium; V, ventricle; D, dual; S, single), the second is the chamber(s) in which sensing occurs (O, none; A, atrium; V, ventricle; D, dual; S, single), the third is the response to a sensed event (O, none; I, inhibition; T, triggered; D, inhibition + triggered), the fourth refers to the programmability or rate response (R, rate responsive), and the fifth refers to the existence of antitachycardia functions if present (O, none; P, antitachycardia pacing; S, shock; D, pace + shock). Almost all modern pacemakers are multiprogrammable and have the capability for rate responsiveness using one of several rate sensors: activity or motion, minute ventilation, or QT interval. The most commonly programmed modes of implanted single- and dual-chamber pacemakers are VVIR and DDDR, respectively, although multiple modes can be programmed in modern pacemakers.

Although pacemakers are highly reliable, they are subject to a number of complications related to implantation and electronic function. In adults, permanent pacemakers are most commonly implanted with access to the heart by way of the subclavian–superior vena cava venous system. Rare, but possible, acute complications of transvenous pacemaker implantation include infection, hematoma, pneumothorax, cardiac perforation, diaphragmatic/phrenic nerve stimulation, and lead dislodgment. Limitations of chronic pacemaker therapy include infection, erosion, lead failure, and abnormalities resulting from inappropriate programming or interaction with the patient's native electrical cardiac function. Rotation of the pacemaker pulse generator in its subcutaneous pocket, either intentionally or inadvertently, often referred to as "twiddler's syndrome," can wrap the leads around the generator and produce dislodgment with failure to sense or pace the heart. The small size and light weight of contemporary pacemakers make this a rare complication. Transvenous leads are considered the "Achilles heel" of permanent pacing systems. Enhancements in battery technology and component design have produced a pacing system small enough to be implanted in the heart without the need for a transvenous lead. These "leadless" pacemakers are appropriate for patients with indications for single chamber ventricular (right ventricle) pacing **(see Chap. 240)**.

Complications stemming from chronic cardiac pacing also result from disturbances in AV synchrony and/or left ventricular mechanical synchrony. Pacing modes that interrupt or fail to restore AV synchrony may lead to a constellation of signs and symptoms, collectively referred to as pacemaker syndrome, that include neck pulsation, fatigue, palpitations, cough, confusion, exertional dyspnea, dizziness, syncope, elevation in jugular venous pressure, canon A waves, and stigmata of congestive heart failure, including edema, rales, and a third heart sound. Right ventricular apical pacing can induce dyssynchronous activation of the left ventricle, leading to compromised left ventricular systolic function, mitral valve regurgitation, and the previously mentioned stigmata of congestive heart failure. Maintenance of AV synchrony can minimize the sequelae of pacemaker syndrome. Selection of pacing modes that minimize unnecessary ventricular pacing or implantation of a device capable of right and left ventricular pacing (biventricular pacing) can help minimize the deleterious consequences of pacing-induced mechanical dyssynchrony at the ventricular level.

Pacemaker Therapy in SA Node Dysfunction

Pacing in SA nodal disease is indicated to alleviate symptoms of bradycardia. Consensus guidelines published by the American Heart Association (AHA)/American College of Cardiology/Heart Rhythm Society (ACC/HRS) outline the indications for the use of pacemakers and categorize them by class based on levels of evidence. Class I conditions are those for which there is evidence or consensus of opinion that therapy is useful and effective. In class II conditions, there is conflicting evidence or a divergence of opinion about the efficacy of a procedure or treatment; in class IIa conditions, the weight of evidence or opinion favors treatment; and in class IIb conditions, efficacy is less well established by the evidence or opinion of experts. In class III conditions, the evidence or weight of opinion indicates that the therapy is not efficacious or useful and may be harmful.

Class I indications for pacing in SA node dysfunction include documented symptomatic bradycardia, sinus node dysfunction–associated long-term drug therapy for which there is no alternative, and symptomatic chronotropic incompetence. Class IIa indications include those outlined previously in which sinus node dysfunction is suspected but not documented and for syncope of unexplained origin in the presence of major abnormalities of SA node dysfunction. Mildly symptomatic individuals with heart rates consistently <40 beats/min constitute a class IIb indication for pacing. Pacing is not indicated in patients with SA node dysfunction who do not have symptoms and in those in whom bradycardia is associated with the use of nonessential drugs **(Table 239-2)**.

There is some controversy about the mode of pacing that should be employed in SA node disease. A number of randomized, single-blind trials of pacing mode have been performed. There are no trials that demonstrate an improvement in mortality rate with AV synchronous pacing compared with single-chamber pacing in SA node disease. In some of these studies, the incidence of atrial fibrillation and thromboembolic events was reduced with AV synchronous pacing. In trials of patients with dual-

chamber pacemakers designed to compare single-chamber with dual-chamber pacing by crossover design, the need for AV synchronous pacing due to pacemaker syndrome was common. Pacing modes that preserve AV synchrony appear to be associated with a reduction in the incidence of atrial fibrillation and improved quality of life. Because of the low but finite incidence of AV conduction disease, patients with SA node dysfunction usually undergo dual-chamber pacemaker implantation.

Pacemaker Therapy in Carotid Sinus Hypersensitivity and Vasovagal Syncope

Carotid sinus hypersensitivity, if accompanied by a significant cardioinhibitory component, responds well to pacing. In this circumstance, pacing is required only intermittently and single-chamber ventricular pacing is often sufficient. The mechanism of vasovagal syncope is incompletely understood but appears to involve activation of cardiac mechanoreceptors with consequent activation of neural centers that mediate vagal activation and withdrawal of sympathetic nervous system tone. Several randomized clinical trials have been performed in patients with drug-refractory vasovagal syncope, with some studies suggesting reduction in the frequency and the time to recurrent syncope in patients who were paced compared with those who were not. A recent follow-up study to one of those initial trials, however, found less convincing results, casting some doubt on the utility of pacing for vagally mediated syncope.

TABLE 239-2 Summary of Guidelines for Pacemaker Implantation in SA Node Dysfunction

Class I

1. SA node dysfunction with symptomatic bradycardia or sinus pause
2. Symptomatic SA node dysfunction as a result of essential long-term drug therapy with no acceptable alternatives
3. Symptomatic chronotropic incompetence
4. Atrial fibrillation with bradycardia and pauses >5 s

Class IIa

1. SA node dysfunction with heart rates <40 beats/min without a clear and consistent relationship between bradycardia and symptoms
2. SA node dysfunction with heart rates <40 beats/min on an essential long-term drug therapy with no acceptable alternatives, without a clear and consistent relationship between bradycardia and symptoms
3. Syncope of unknown origin when major abnormalities of SA node dysfunction are discovered or provoked by electrophysiologic testing

Class IIb

1. Mildly symptomatic patients with waking chronic heart rates <40 beats/min

Class III

1. SA node dysfunction in asymptomatic patients, even those with heart rates <40 beats/min
2. SA node dysfunction in which symptoms suggestive of bradycardia are not associated with a slow heart rate
3. SA node dysfunction with symptomatic bradycardia due to nonessential drug therapy

Source: Data from AE Epstein et al: J Am Coll Cardiol 51:e1, 2008 and CM Tracy et al: J Am Coll Cardiol 61:e6, 2013.

FURTHER READING

Epstein AE et al: 2012 ACCF/AHA/HRS Focused Update Incorporated into the ACCF/AHA/HRS 2008 Guidelines for Device-Based Therapy of Cardiac Rhythm Abnormalities: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society developed in collaboration with the American Association for Thoracic Surgery and the Society of Thoracic Surgeons. J Am Coll Cardiol 61:e6, 2013. [PubMed: 23265327]

Epstein AE et al: ACC/AHA/HRS 2008 Guidelines for Device-Based Therapy of Cardiac Rhythm Abnormalities: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (writing Committee to Revise the ACC/AHA/NASPE 2002 Guideline Update for Implantation of Cardiac Pacemakers and Antiarrhythmia Devices) developed in collaboration with the American Association for Thoracic Surgery and Society of Thoracic Surgeons. J Am Coll Cardiol 51:el, 2008.

Epstein AE et al: ACC/AHA/NASPE 2008 Guidelines for device-based therapy of cardiac rhythm abnormalities. J Am Coll Cardiol 51:21, 2008.

Goldschlager N et al: Sinus node dysfunction, in *Electrophysiological Disorders of the Heart*, 2nd ed, S Saksena, AJ Camm (eds). Philadelphia, Elsevier Churchill Livingstone, 2012.

Josephson ME: *Clinical Cardiac Electrophysiology: Techniques and Interpretations*, 5th ed. Philadelphia, Wolters Kluwer, 2016.

Vijayaraman P, Ellenbogen KA: Bradyarrhythmias and pacemakers, in *Hurst's The Heart*, 13th ed, V Fuster et al (eds). New York, McGraw-Hill, 2011.

Chapter 240: The Bradyarrhythmias: Disorders of the Atrioventricular Node

David D. Spragg; Gordon F. Tomaselli

INTRODUCTION

Impulses generated in the sinoatrial (SA) node or in ectopic atrial loci are conducted to the ventricles through the electrically and anatomically complex atrioventricular (AV) node. As described in **Chap. 239**, the electrophysiologic properties of nodal tissue are distinct from atrial and ventricular myocardium. Cells located in the AV node sit at a relatively higher resting membrane potential than surrounding atrial and ventricular myocytes, exhibit spontaneous depolarization during phase 4 of the action potential, and have slower phase 0 depolarization (mediated by calcium influx in nodal tissue) than that seen in ventricular tissue (mediated by sodium influx).

Bradycardia may occur when conduction across the AV node is compromised, resulting in slow ventricular rates, with the possibility of attendant symptoms, including fatigue, syncope, and (if subsidiary pacemaker activity is insufficient) even death. It is important to recognize that in the setting of disturbed AV conduction, SA activation and atrial systole may occur at normal or even accelerated rates, while ventricular activation is either slowed or nonexistent. Transient AV conduction block is common in the young and is most likely the result of high vagal tone found in up to 10% of young adults. Acquired and persistent failure of AV conduction is decidedly rare in healthy adult populations, with an estimated incidence of 200 per million population per year. In the setting of myocardial ischemia, aging and fibrosis, or cardiac infiltrative diseases, however, persistent AV block is much more common.

As with symptomatic bradycardia arising from SA node dysfunction, permanent pacing is the only reliable therapy for symptoms arising from AV conduction block. Approximately 50% of the 160,000 permanent pacemakers implanted in the United States and 70–80% of those in Europe are implanted for disorders of AV conduction.

STRUCTURE AND PHYSIOLOGY OF THE AV NODE

The AV conduction axis is structurally complex, involving the atria and ventricles as well as the AV node. Unlike the SA node, the AV node is a subendocardial structure originating in the transitional zone, which is composed of aggregates of cells in the posterior-inferior right atrium. Superior, medial, and posterior transitional atrionodal bundles converge on the compact AV node. The compact AV node (~1 × 3 × 5 mm) is situated at the apex of the triangle of Koch, which is defined by the coronary sinus ostium posteriorly, the septal tricuspid valve annulus anteriorly, and the tendon of Todaro superiorly. The compact AV node continues as the penetrating AV bundle where it immediately traverses the central fibrous body and is in close proximity to the aortic, mitral, and tricuspid valve annuli; thus, it is subject to injury in the setting of valvular heart disease or its surgical treatment. The penetrating AV bundle continues through the annulus fibrosis and emerges along the ventricular septum adjacent to the membranous septum as the bundle of His. The right bundle branch (RBB) emerges from the distal AV bundle in a band that traverses the right ventricle (moderator band). In

contrast, the left bundle branch (LBB) is a broad subendocardial sheet of tissue on the septal left ventricle. The Purkinje fiber network emerges from the RBB and LBB and extensively ramifies on the endocardial surfaces of the right and left ventricles, respectively.

The blood supply to the penetrating AV bundle is from the AV nodal artery and first septal perforator of the left anterior descending coronary artery. The bundle branches also have a dual blood supply from the septal perforators of the left anterior descending coronary artery and branches of the posterior descending coronary artery. The AV node is highly innervated with postganglionic sympathetic and parasympathetic nerves. The bundle of His and distal conducting system are minimally influenced by autonomic tone.

The cells that constitute the AV node complex are heterogeneous with a range of action potential profiles. In the transitional zones, the cells have an electrical phenotype between those of atrial myocytes and cells of the compact node **(see Fig. 239-1)**. Atrionodal transitional connections may exhibit *decremental conduction*, defined as slowing of conduction with increasingly rapid rates of stimulation. Fast and slow AV nodal pathways have been described, but it is controversial whether these two types of pathway are anatomically distinct or represent functional heterogeneities in different regions of the AV nodal complex. Myocytes that constitute the compact node are depolarized (resting membrane potential ~–60 mV) and exhibit action potentials with low amplitudes, slow upstrokes of phase 0 (<10 V/s), and phase 4 diastolic depolarization; high-input resistance; and relative insensitivity to external [K^+]. The action potential phenotype is explained by the complement of ionic currents expressed. AV nodal cells lack a robust inward rectifier potassium current (I_{K1}) and fast sodium current (I_{Na}); L-type calcium current ($I_{Ca\text{-}L}$) is responsible for phase 0; and phase 4 depolarization reflects the composite activity of the depolarizing currents—funny current (I_f), $I_{Ca\text{-}L}$, T-type calcium current ($I_{Ca\text{-}T}$), and sodium calcium exchanger current (I_{NCX})—and the repolarizing currents—delayed rectifier (I_{Kr}) and acetylcholine-gated (I_{KACh}) potassium currents. Electrical coupling between cells in the AV node is tenuous due to the relatively sparse expression of gap junction channels (predominantly connexin-40) and increased extracellular volume.

The His bundle and the bundle branches are insulated from ventricular myocardium. The most rapid conduction in the heart is observed in these tissues. The action potentials exhibit very rapid upstrokes (phase 0), prolonged plateaus (phase 2), and modest automaticity (phase 4 depolarization). Gap junctions, composed largely of connexin-40, are abundant, but bundles are poorly connected transversely to ventricular myocardium.

ETIOLOGY OF AV CONDUCTION DISEASE

Conduction block from the atrium to the ventricle can occur for a variety of reasons in a number of clinical situations, and AV conduction block may be classified in a number of ways. The etiologies may be functional or structural, in part analogous to extrinsic and intrinsic causes of SA nodal dysfunction. The block may be classified by its severity from first to third degree or complete AV block or by the location of block within the AV conduction system. **Table 240-1** summarizes the etiologies of AV conduction block. Those that are functional (autonomic, metabolic/endocrine, and drug-related) tend to be reversible. Most other etiologies produce structural changes, typically fibrosis, in segments of the AV conduction axis that are generally permanent. Heightened vagal tone during sleep or in well-conditioned individuals can be associated with all grades of AV block. Carotid sinus hypersensitivity, vasovagal syncope, and cough and micturition syncope may be associated with SA node slowing and AV conduction block. Transient metabolic and endocrinologic disturbances as well as a number of pharmacologic agents also may produce reversible AV conduction block.

TABLE 240-1 Etiologies of Atrioventricular Block

Autonomic	
Carotid sinus hypersensitivity	Vasovagal
Metabolic/Endocrine	
Hyperkalemia	Hypothyroidism
Hypermagnesemia	Adrenal insufficiency
Drug-Related	
Beta blockers	Adenosine
Calcium channel blockers	Antiarrhythmics (class I and III)
Digitalis	Lithium
Infectious	
Endocarditis	Tuberculosis
Lyme disease	Diphtheria
Chagas' disease	Toxoplasmosis
Syphilis	
Heritable/Congenital	
Congenital heart disease	
Maternal SLE	Facioscapulohumeral MD, OMIM #158900 (4q35)
Kearns-Sayre syndrome, OMIM #530000	Emery-Dreifuss MD, OMIM #310300 (Xq28)
	Progressive familial heart block, type IA OMIM #113900 (3p21)
Myotonic dystrophy	Progressive familial heart block, type IB, OMIM #604559 (19q13.32)
Type 1, OMIM #160900 (19q13.2-13.3)	Progressive familial heart block, type II, OMIM %140400 (1q32)
Type 2, OMIM #602668 (3q13.3-q24)	
Inflammatory	
SLE	MCTD
Rheumatoid arthritis	Scleroderma
Infiltrative	
Amyloidosis	Hemochromatosis
Sarcoidosis	
Neoplastic/Traumatic	
Lymphoma	Radiation
Mesothelioma	Catheter ablation
Melanoma	
Degenerative	
Lev's disease	Lenègre's disease
Coronary Artery Disease	
Acute MI	

Abbreviations: MCTD, mixed connective tissue disease; MI, myocardial infarction; OMIM, Online Mendelian Inheritance in Man (database; designations: #, phenotypic description, molecular basis known; %, phenotypic description); SLE, systemic lupus erythematosus.

Several infectious diseases have a predilection for the conducting system. Lyme disease may involve the heart in up to 50% of cases; 10% of patients with Lyme carditis develop AV conduction block, which is generally reversible but may require temporary pacing support. Chagas' disease, which is common in Latin America, and syphilis may produce more persistent AV conduction disturbances. Some autoimmune and infiltrative diseases may produce AV conduction block, including systemic lupus erythematosus (SLE), rheumatoid arthritis, mixed connective tissue disease, scleroderma, amyloidosis (primary and secondary), sarcoidosis, and hemochromatosis; rare malignancies also may impair AV conduction.

Idiopathic progressive fibrosis of the conduction system is one of the more common and degenerative causes of AV conduction block. Aging is associated with degenerative changes in the summit of the ventricular septum, central fibrous body, and aortic and mitral annuli and has been described as "sclerosis of the left cardiac skeleton." The process typically begins in the fourth decade of life and may be accelerated by atherosclerosis, hypertension, and diabetes mellitus. Accelerated forms of progressive familial heart block have been identified in families with mutations in the cardiac sodium channel gene (*SCN5A*) and other loci that have been mapped to chromosomes 1 and 19.

AV conduction block has been associated with heritable neuromuscular diseases, including the nucleotide repeat disease myotonic dystrophy, the mitochondrial myopathy Kearns-Sayre syndrome **(Chap. 441)**, and several of the monogenic muscular dystrophies. Congenital AV block may be observed in complex congenital cardiac anomalies **(Chap. 264)**, such as transposition of the great arteries, ostium primum atrial septal defects (ASDs), ventricular septal defects (VSDs), endocardial cushion defects, and some single-ventricle defects. Congenital AV block in the setting of a structurally normal heart has been seen in children born to mothers with SLE. Iatrogenic AV block may occur during mitral or aortic valve surgery, rarely in the setting of thoracic radiation, and as a consequence of catheter ablation. AV block is a decidedly rare complication of the surgical repair of VSDs or ASDs but may complicate repairs of transposition of the great arteries.

Coronary artery disease may produce transient or persistent AV block. In the setting of coronary spasm, ischemia, particularly in the right coronary artery distribution, may produce transient AV block. In acute myocardial infarction (MI), AV block transiently develops in 10–25% of patients; most commonly, this is first- or second-degree AV block, but complete heart block (CHB) may also occur. Second-degree and higher-grade AV block tends to occur more often in inferior than in anterior acute MI; however, the level of block in inferior MI tends to be in the AV node with more stable, narrow escape rhythms. In contrast, acute anterior MI is associated with block in the distal AV nodal complex, His bundle, or bundle branches and results in wide complex, unstable escape rhythms and a worse prognosis with high mortality rates.

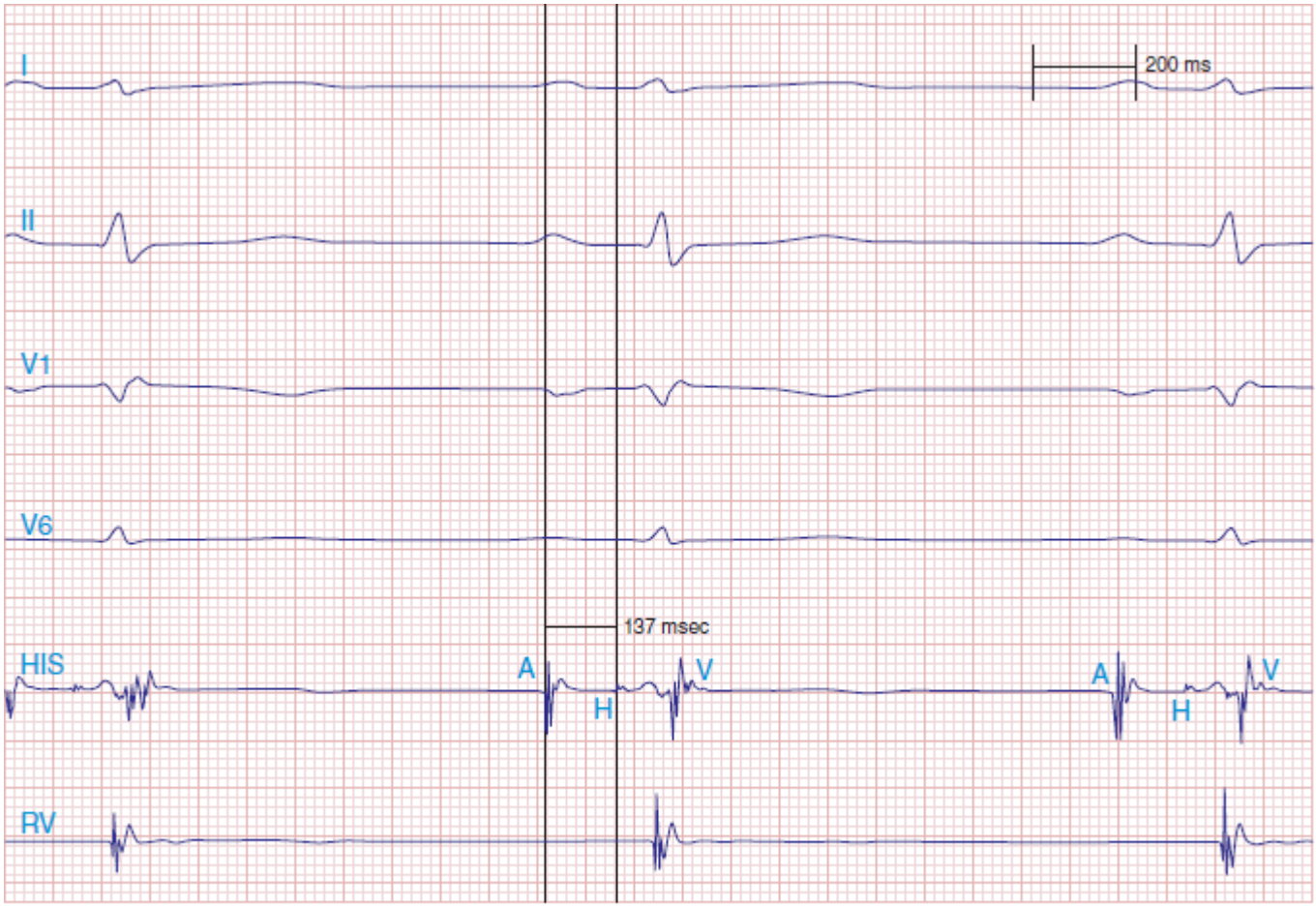
ELECTROCARDIOGRAPHY AND ELECTROPHYSIOLOGY OF AV CONDUCTION BLOCK

AV conduction block typically is diagnosed electrocardiographically, which characterizes the severity of the conduction disturbance and allows one to draw inferences about the location of the block. AV conduction block manifests as slow conduction in its mildest forms and failure to conduct, either intermittent or persistently, in more severe varieties. First-degree AV block (PR interval >200 ms) is a slowing of conduction through the AV junction **(Fig. 240-1)**. The site of delay is typically in the AV node but may be in the atria, bundle of His, or His-Purkinje system. A wide QRS is suggestive of delay in the distal conduction system, whereas a narrow QRS suggests delay in the AV node proper or, less commonly, in the bundle of His. In second-degree AV block there is an intermittent failure of electrical impulse conduction from atrium to ventricle. Second-degree AV block is subclassified as Mobitz type I (Wenckebach) or Mobitz type II. The periodic failure of conduction in Mobitz type I block is characterized by a progressively lengthening PR interval, shortening of the RR interval, and a pause that is less than two times the immediately preceding RR interval on the electrocardiogram (ECG).

The ECG complex after the pause exhibits a shorter PR interval than that immediately preceding the pause (**Fig. 240-2**). This ECG pattern most often arises because of decremental conduction of electrical impulses in the AV node.

FIGURE 240-1

First-degree AV block with slowing of conduction in the AV node as indicated by the prolonged atrial-to-His bundle electrogram (AH) interval, in this case 157 ms. The His bundle-to-earliest ventricular activation on the surface ECG (HV) interval is normal. The normal HV interval suggests normal conduction below the AV node to the ventricle. I and V_1 are surface ECG leads, and HIS is the recording of the endocavitary electrogram at the His bundle position. A, H, and V are labels for the atrial, His bundle, and right ventricular electrograms, respectively.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 240-2

Mobitz type I second-degree AV block. The PR interval prolongs before the pause, as shown in the ladder diagram. The ECG pattern results from slowing of conduction in the AV node.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

It is important to distinguish type I from type II second-degree AV nodal block because the latter has more serious prognostic implications. Type II second-degree AV block is characterized by intermittent failure of conduction of the P wave without changes in the preceding PR or RR intervals. When AV block is 2:1, it may be difficult to distinguish type I from type II block. Type II second-degree AV block typically occurs in the distal or infra-His conduction system, is often associated with intraventricular conduction delays (e.g., bundle branch block), and is more likely to proceed to higher grades of AV block than is type I second-degree AV block. Second-degree AV block (particularly type II) may be associated with a series of nonconducted P waves, referred to as *paroxysmal AV block* **(Fig. 240-3)**, and implies significant conduction system disease and is an indication for permanent pacing. Complete failure of conduction from atrium to ventricle is referred to as complete or third-degree AV block. AV block that is intermediate between second degree and third degree is referred to as high-grade AV block and, as with CHB, implies advanced AV conduction system disease. In both cases, the block is most often distal to the AV node, and the duration of the QRS complex can be helpful in determining the level of the block. In the absence of a preexisting bundle branch block, a wide QRS escape rhythm **(Fig. 240-4B)** implies a block in the distal His or bundle branches; in contrast, a narrow QRS rhythm implies a block in the AV node or proximal His and an escape rhythm originating in the AV junction **(Fig. 240-4A)**. Narrow QRS escape rhythms are typically faster and more stable than wide QRS escape rhythms and originate more proximally in the AV conduction system.

FIGURE 240-3

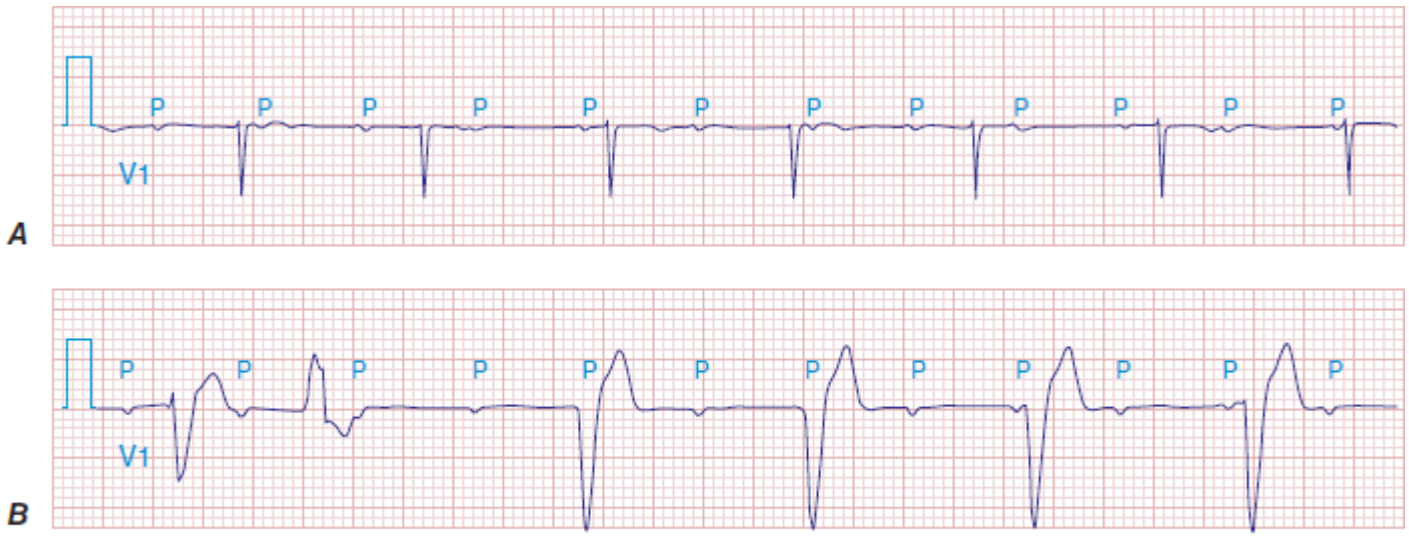
Paroxysmal AV block. Multiple nonconducted P waves after a period of sinus bradycardia with a normal PR interval. This implies significant conduction system disease, requiring permanent pacemaker implantation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 240-4

High-grade AV block. ***A.*** Multiple nonconducted P waves with a regular narrow complex QRS escape rhythm probably emanating from the AV junction. ***B.*** A wide complex QRS escape and a single premature ventricular contraction. In both cases, there is no consistent temporal relationship between the P waves and QRS complexes.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSTIC TESTING

Diagnostic testing in the evaluation of AV block is aimed at determining the level of conduction block, particularly in asymptomatic patients, since the prognosis and therapy depend on whether the block is

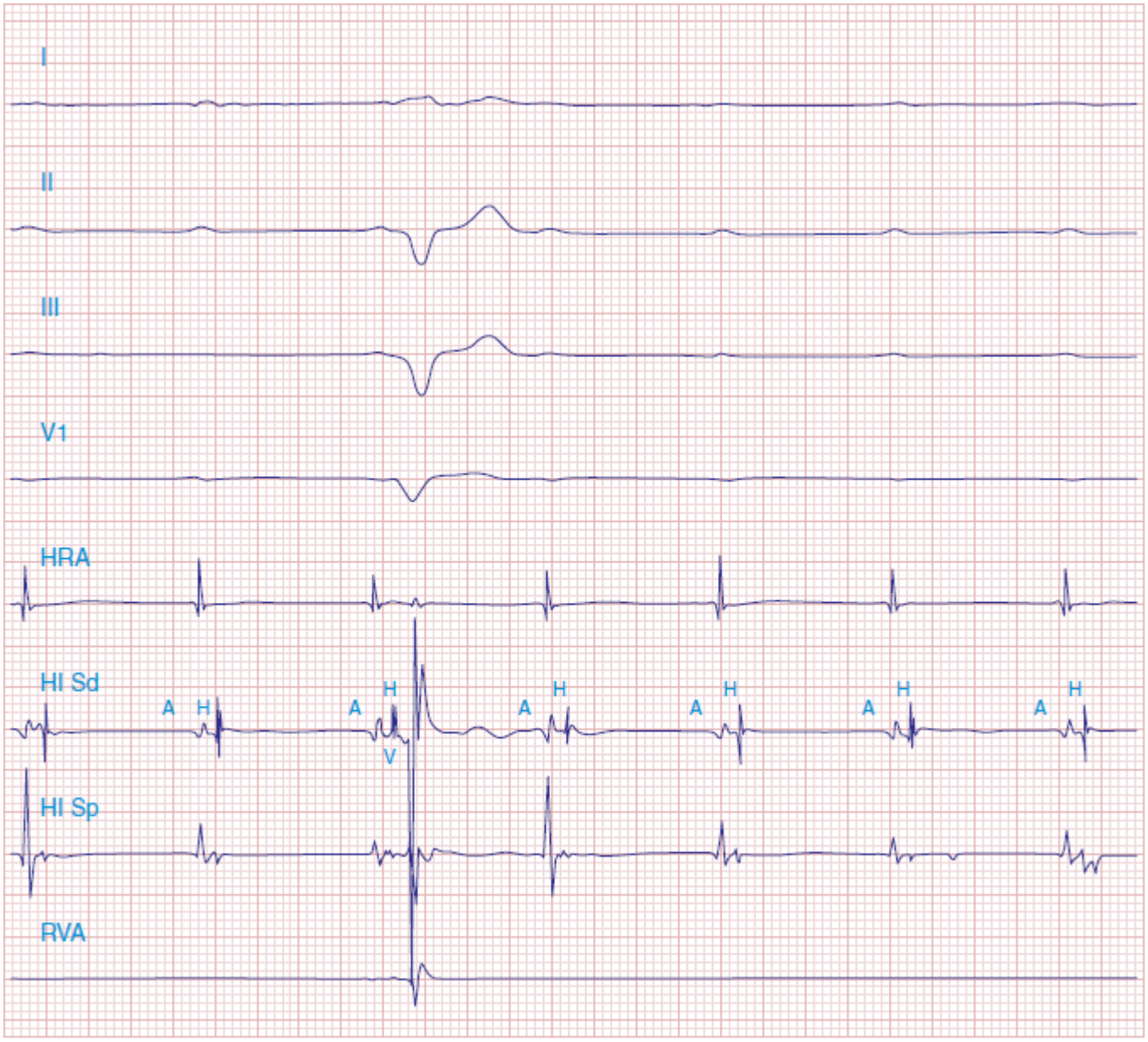
in or below the AV node. Vagal maneuvers, carotid sinus massage, exercise, and administration of drugs such as atropine and isoproterenol may be diagnostically informative. Owing to the differences in the innervation of the AV node and infranodal conduction system, vagal stimulation and carotid sinus massage slow conduction in the AV node but have less of an effect on infranodal tissue and may even improve conduction due to a reduced rate of activation of distal tissues. Conversely, atropine, isoproterenol, and exercise improve conduction through the AV node and impair infranodal conduction. In patients with congenital CHB and a narrow QRS complex, exercise typically increases heart rate; by contrast, those with acquired CHB, particularly with wide QRS, do not respond to exercise with an increase in heart rate.

Additional diagnostic evaluation, including electrophysiologic testing, may be indicated in patients with syncope and suspected high-grade AV block. This is particularly relevant if noninvasive testing does not reveal the cause of syncope or if the patient has structural heart disease with ventricular tachyarrhythmias as a cause of symptoms. Electrophysiologic testing provides more precise information regarding the location of AV conduction block and permits studies of AV conduction under conditions of pharmacologic stress and exercise. Recording of the His bundle electrogram by a catheter positioned at the superior margin of the tricuspid valve annulus provides information about conduction at all levels of the AV conduction axis. A properly recorded His bundle electrogram reveals local atrial activity, the His electrogram, and local ventricular activation; when it is monitored simultaneously with recorded body surface electrocardiographic traces, intraatrial, AV nodal, and infranodal conduction times can be assessed (Fig. 240-1). The time from the most rapid deflection of the atrial electrogram in the His bundle recording to the His electrogram (*AH interval*) represents conduction through the AV node and is normally <130 ms. The time from the His electrogram to the earliest onset of the QRS on the surface ECG (*HV interval*) represents the conduction time through the His-Purkinje system and is normally ≤55 ms.

Rate stress produced by pacing can unveil abnormal AV conduction. Mobitz I second-degree AV block at short atrial paced cycle lengths is a normal response. However, when it occurs at atrial cycle lengths >500 ms (<120 beats/min) in the absence of high vagal tone, it is abnormal. Typically, type I second-degree AV block is associated with prolongation of the AH interval, representing conduction slowing and block in the AV node. AH prolongation occasionally is due to the effect of drugs (beta blockers, calcium channel blockers, digitalis) or increased vagal tone. Atropine can be used to reverse high vagal tone; however, if AH prolongation and AV block at long pacing cycle lengths persist, intrinsic AV node disease is likely. Type II second-degree block is typically infranodal, often in the His-Purkinje system. Block below the node with prolongation of the HV interval or a His bundle electrogram with no ventricular activation **(Fig. 240-5)** is abnormal unless it is elicited at fast pacing rates or short coupling intervals with extra stimulation. It is often difficult to determine the type of second-degree AV block when 2:1 conduction is present; however, the finding of a His bundle electrogram after every atrial electrogram indicates that block is occurring in the distal conduction system.

FIGURE 240-5

High-grade AV block below the His. The AH interval is normal and is not changing before the block. Atrial and His bundle electrograms are recorded consistent with block below the distal AV junction. I, II, III, and V_1 are surface ECG leads. HISp, HISd, and RVA are the proximal HIS, distal HIS, and right ventricular apical electrical recordings, respectively. A, H, and V represent the atrial, His, and ventricular electrograms on the His bundle recording, respectively. *(Used with permission from Dr. Joseph Marine.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Intracardiac recording at electrophysiologic study that reveals prolongation of conduction through the His-Purkinje system (i.e., long HV interval) is associated with an increased risk of progression to higher grades of block and is generally an indication for pacing. In the setting of bundle branch block, the HV interval may reveal the condition of the unblocked bundle and the prognosis for developing more advanced AV conduction block. Prolongation of the HV interval in patients with asymptomatic bundle branch block is associated with an increased risk of developing higher-grade AV block. The risk increases with greater prolongation of the HV interval such that in patients with an HV interval >100 ms, the annual incidence of complete AV block approaches 10%, indicating a need for pacing. In patients with acquired CHB, even if intermittent, there is little role for electrophysiologic testing, and pacemaker implantation is almost always indicated.

TREATMENT

TREATMENT Management of AV Conduction Block

Temporary or permanent artificial pacing is the most reliable treatment for patients with symptomatic AV conduction system disease. However, exclusion of reversible causes of AV block and the need for temporary heart rate support based on the hemodynamic condition of the patient are essential considerations in each patient. Correction of electrolyte derangements and ischemia, inhibition of excessive vagal tone, and withholding of drugs with AV nodal blocking properties may increase the heart rate. Adjunctive pharmacologic treatment with atropine or isoproterenol may be useful if the block is in the AV node. Since most pharmacologic treatment may take some time to initiate and become effective, temporary pacing may be necessary. The most expeditious technique is the use of transcutaneous pacing, where pacing patches are placed anteriorly over the cardiac apex (cathode) and posteriorly between the spine and the scapula or above the right nipple (anode). Acutely, transcutaneous pacing is highly effective, but its duration is limited by patient discomfort and longer-term failure to capture the ventricle owing to changes in lead impedance. If a patient requires more than a few minutes of pacemaker support, transvenous temporary pacing should be instituted. Temporary pacing leads can be placed from the jugular or subclavian venous system and advanced to the right ventricle, permitting stable temporary pacing for many days, if necessary. In most circumstances, in the absence of prompt resolution, conduction block distal to the AV node requires permanent pacemaking.

PACEMAKERS IN AV CONDUCTION DISEASE

There are no randomized trials that evaluate the efficacy of pacing in patients with AV block, as there are no reliable therapeutic alternatives for AV block and untreated high-grade AV block is potentially lethal. The consensus guidelines for pacing in acquired AV conduction block in adults provide a general outline for situations in which pacing is indicated **(Table 240-2)**. Pacemaker implantation should be performed in any patient with symptomatic bradycardia and irreversible second- or third-degree AV block, regardless of the cause or level of block in the conducting system. Symptoms may include those directly related to bradycardia and low cardiac output or to worsening heart failure, angina, or intolerance to an essential medication. Pacing in patients with asymptomatic AV block should be individualized; situations in which pacing should be considered are patients with acquired CHB, particularly in the setting of cardiac enlargement; left ventricular dysfunction; and waking heart rates ≤40 beats/min. Patients who have asymptomatic second-degree AV block of either type should be considered for pacing if the block is demonstrated to be intra- or infra-His or is associated with a wide QRS complex. Pacing may be indicated in asymptomatic patients in special circumstances, in patients with profound first-degree AV block and left ventricular dysfunction in whom a shorter AV interval produces hemodynamic improvement, and in the setting of milder forms of AV conduction delay (first-degree AV block, intraventricular conduction delay) in patients with neuromuscular diseases that have a predilection for the conduction system, such as myotonic dystrophy and other muscular dystrophies, and Kearns-Sayre syndrome.

PACEMAKER THERAPY IN MYOCARDIAL INFARCTION

AV block in acute MI is often transient, particularly in inferior infarction. The circumstances in which pacing is indicated in acute MI are persistent second- or third-degree AV block, particularly if symptomatic, and transient second- or third-degree AV block associated with bundle branch block **(Table 240-3)**. Pacing is generally not indicated in the setting of transient AV block in the absence of intraventricular conduction delays or in the presence of fascicular block or first-degree AV block that develops in the setting of preexisting bundle branch block. Fascicular blocks that develop in acute MI in the absence of other forms of AV block also do not require pacing (Table 240-3 and **Table 240-4**).

PACEMAKER THERAPY IN BIFASCICULAR AND TRIFASCICULAR BLOCK

Distal forms of AV conduction block may require pacemaker implantation in certain clinical settings. Patients with bifascicular or trifascicular block and symptoms, particularly syncope that is not attributable to other causes, should undergo pacemaker implantation. Pacemaking is indicated in asymptomatic patients with bifascicular or trifascicular block who experience intermittent third-degree,

type II second-degree AV block or alternating bundle branch block. In patients with fascicular block who are undergoing electrophysiologic study, a markedly prolonged HV interval or block below the His at long cycle lengths also may constitute an indication for permanent pacing. Patients with fascicular block and the neuromuscular diseases previously described should also undergo pacemaker implantation (Table 240-4).

SELECTION OF PACING MODE

In general, a pacing mode that maintains AV synchrony reduces complications of pacing such as pacemaker syndrome and pacemaker-mediated tachycardia. This is particularly true in younger patients; the importance of dual-chamber pacing in the elderly, however, is not well established. The availability of leadless miniaturized pacing systems may be appropriate in patients with indications for single chamber ventricular pacing, such as patients with atrial fibrillation and AV conduction block. Several studies have failed to demonstrate a difference in mortality rate in older patients with AV block treated with a single-(VVI) compared with a dual-(DDD) chamber pacing mode. In some of the studies that randomized pacing mode, the risk of chronic atrial fibrillation and stroke risk decreased with physiologic pacing. In patients with sinus rhythm and AV block, the very modest increase in risk with dual-chamber pacemaker implantation appears to be justified to avoid the possible complications of single-chamber pacing.

TABLE 240-2 Guideline Summary for Pacemaker Implantation in Acquired AV Block

Class I

1. Third-degree or high-grade AV block at any anatomic level associated with:
    a. Symptomatic bradycardia
    b. Essential drug therapy that produces symptomatic bradycardia
    c. Periods of asystole >3 s or any escape rate <40 beats/min while awake, or an escape rhythm originating below the AV node
    d. Postoperative AV block not expected to resolve
    e. Catheter ablation of the AV junction
    f. Neuromuscular diseases such as myotonic dystrophy, Kearns-Sayre syndrome, Erb dystrophy, and peroneal muscular atrophy, regardless of the presence of symptoms
2. Second-degree AV block with symptomatic bradycardia
3. Type II second-degree AV block with a wide QRS complex with or without symptoms
4. Exercise-induced second- or third-degree AV block in the absence of ischemia
5. Atrial fibrillation with bradycardia and pauses >5 s

Class IIa

1. Asymptomatic third-degree AV block regardless of level
2. Asymptomatic type II second-degree AV block with a narrow QRS complex
3. Asymptomatic type II second-degree AV block with block within or below the His at electrophysiologic study
4. First- or second-degree AV block with symptoms similar to pacemaker syndrome

Class IIb

1. AV block in the setting of drug use/toxicity, when the block is expected to recur even with drug discontinuation
2. Neuromuscular diseases such as myotonic dystrophy, Kearns-Sayre syndrome, Erb dystrophy, and peroneal muscular atrophy with any degree of AV block regardless of the presence of symptoms

Class III

1. Asymptomatic first-degree AV block
2. Asymptomatic type I second-degree AV block at the AV node level
3. AV block that is expected to resolve or is unlikely to recur (Lyme disease, drug toxicity)

Source: Data from AE Epstein et al: J Am Coll Cardiol 51:e1, 2008.

TABLE 240-3 Guideline Summary for Pacemaker Implantation in AV Conduction Block in Acute Myocardial Infarction (AMI)

Class I

1. Persistent second-degree AV block in the His-Purkinje system with bilateral bundle branch block or third-degree block within or below the His after AMI
2. Transient advanced (second- or third-degree) infranodal AV block and associated bundle branch block. If the site of block is uncertain, an electrophysiologic study may be necessary
3. Persistent and symptomatic second- or third-degree AV block

Class IIb

1. Persistent second- or third-degree AV block at the AV node level

Class III

1. Transient AV block in the absence of intraventricular conduction defects
2. Transient AV block in the presence of isolated left anterior fascicular block
3. Acquired left anterior fascicular block in the absence of AV block
4. Persistent first-degree AV block in the presence of bundle branch block that is old or age-indeterminate

Source: Data from AE Epstein et al: J Am Coll Cardiol 51:e1, 2008.

TABLE 240-4 Indications for Pacemaker Implantation in Chronic Bifascicular and Trifascicular Block

Class I

1. Intermittent third-degree AV block
2. Type II second-degree AV block
3. Alternating bundle branch block

Class IIa

1. Syncope not demonstrated to be due to AV block when other likely causes (e.g., ventricular tachycardia) have been excluded
2. Incidental finding at electrophysiologic study of a markedly prolonged HV interval (>100 ms) in asymptomatic patients
3. Incidental finding at electrophysiologic study of pacing-induced infra-His block that is not physiologic

Class IIb

1. Neuromuscular diseases such as myotonic dystrophy, Kearns-Sayre syndrome, Erb dystrophy, and peroneal muscular atrophy with any degree of fascicular block regardless of the presence of symptoms, because there may be unpredictable progression of AV conduction disease

Class III

1. Fascicular block without AV block or symptoms
2. Fascicular block with first-degree AV block without symptoms

Source: Data from AE Epstein et al: J Am Coll Cardiol 51:e1, 2008.

FURTHER READING

Epstein AE et al: ACC/AHA/HRS 2008 Guidelines for Device-Based Therapy of Cardiac Rhythm Abnormalities: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (writing Committee to Revise the ACC/AHA/NASPE 2002 Guideline Update for Implantation of Cardiac Pacemakers and Antiarrhythmia Devices) developed in collaboration with the American Association for Thoracic Surgery and Society of Thoracic Surgeons. J Am Coll Cardiol 51:el, 2008.

Epstein AE et al: 2012 ACCF/AHA/HRS Focused Update Incorporated Into the ACCF/AHA/HRS 2008 Guidelines for Device-Based Therapy of Cardiac Rhythm Abnormalities: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society developed in collaboration with the American Association for Thoracic Surgery and the Society of Thoracic Surgeons. J Am Coll Cardiol 61:e6, 2013. [PubMed: 23265327]

Goldschlager N et al: Atrioventricular block, in *Electrophysiological Disorders of the Heart*, 2nd ed, S Saksena, AJ Camm (eds). Philadelphia, Elsevier Churchill Livingstone, 2012.

Chapter 241: Approach to Supraventricular Tachyarrhythmias

Gregory F. Michaud; William G. Stevenson

INTRODUCTION

Supraventricular tachyarrhythmias originate from or are dependent on conduction through the atrium or atrioventricular (AV) node to the ventricles. Most produce narrow QRS-complex tachycardia (QRS duration <120 ms) characteristic of ventricular activation over the Purkinje system. Conduction block in the left or right bundle branch or activation of the ventricles from an accessory pathway produces a wide QRS complex during supraventricular tachycardia that must be distinguished from ventricular tachycardia **(Chap. 249)**. Mechanisms of supraventricular tachyarrhythmia can be divided into physiologic sinus tachycardia and pathologic tachycardia **(Table 241-1)**. Pathologic tachycardia can be further sub-classified in terms of mechanism as reentrant arrhythmias dependendent on AV nodal conduction (e.g., AV reentry), large reentry circuits within the atrial tissue alone (e.g., atrial flutter) or focal atrial tachycardias that can be due to automaticity or small reentry circuits **(see Figs. 243-3 and 245-1**). The prognosis and treatment vary considerably depending on the mechanism and underlying heart disease.

TABLE 241-1

Supraventricular Tachycardia

I. Physiologic sinus tachycardia

Defining feature: normal sinus mechanism precipitated by exertion, stress, concurrent illness (Table 242-1)

II. Pathologic supraventricular tachycardia

A. Tachycardias originating from the atrium

Defining feature: tachycardia may continue despite beats that fail to conduct to the ventricles, indicating that the AV node is not participating in the tachycardia circuit

1. ***Inappropriate sinus tachycardia***

Defining feature: tachycardia from the normal sinus node area that occurs without an identifiable precipitating factor as a result of dysfunctional autonomic regulation

2. ***Focal atrial tachycardia (AT)***

Defining feature: Regular atrial tachycardia with defined p wave; may be sustained, nonsustained, paroxysmal, or incessant. Frequent sites of origin occur along the valve annuli of left or right atrium, pulmonary veins, coronary sinus musculature, superior vena cava

3. ***Atrial flutter***—macroreentrant atrial tachycardia

Defining feature: organized reentry creates organized atrial activity, commonly seen as sawtooth flutter waves at rates typically faster than 200 beats/min

a. Common atrial flutter

i. Right atrial reentry parallel to the tricuspid annulus and dependent on conduction through the isthmus between the inferior vena cava and tricuspid annulus

1. Counterclockwise (as viewed from the ventricular aspect)
2. Clockwise

b. Atypical atrial flutter

i. Usually due to reentry in left or right atrium associated with scars usually from prior surgery or catheter ablation for atrial fibrillation, but may be idiopathic

4. ***Atrial fibrillation***

Defining feature: chaotic rapid atrial electrical activity with variable ventricular rate; the most common sustained cardiac arrhythmia in older adults

5. ***Multifocal atrial tachycardia***

Defining feature: multiple discrete p waves often seen in patients with pulmonary disease during acute exacerbations of pulmonary insufficiency

B. AV nodal reentry tachycardia (AVNRT)

Defining feature: paroxysmal regular tachycardia with P waves visible at the end of the QRS complex or not visible at all; the most common paroxysmal sustained tachycardia in healthy young adults; more common in women

C. Tachycardias associated with accessory atrioventricular pathways

1. ***Orthodromic AV reentry tachycardia (AVRT)***

Defining feature: paroxysmal sustained tachycardia similar to AV nodal reentry; during sinus rhythm, evidence of ventricular preexcitation may be present (Wolff-Parkinson-White syndrome) or absent (concealed accessory pathway)

2. ***Preexcited tachycardia***

Defining feature: wide QRS tachycardia with QRS morphology similar to VT

a. Antidromic AV reentry—regular paroxysmal tachycardia

b. Atrial fibrillation with preexcitation—irregular wide complex, or intermittently wide complex tachycardia, some with dangerously rapid rates faster than 250/min

c. Atrial tachycardia or flutter with preexcitation

Abbreviations: AV, atrioventricular; VT, ventricular tachycardia.

Supraventricular tachycardia can be of brief duration, termed *nonsustained*, or can be sustained such that an intervention, such as cardioversion or drug administration, is required for termination. Episodes that occur with sudden onset and termination are referred to as paroxysmal. *Paroxysmal supraventricular tachycardia* (PSVT) refers to a family of tachycardias including AV node reentry, AV reentry using an accessory pathway, and atrial tachycardia.

CLINICAL PRESENTATION

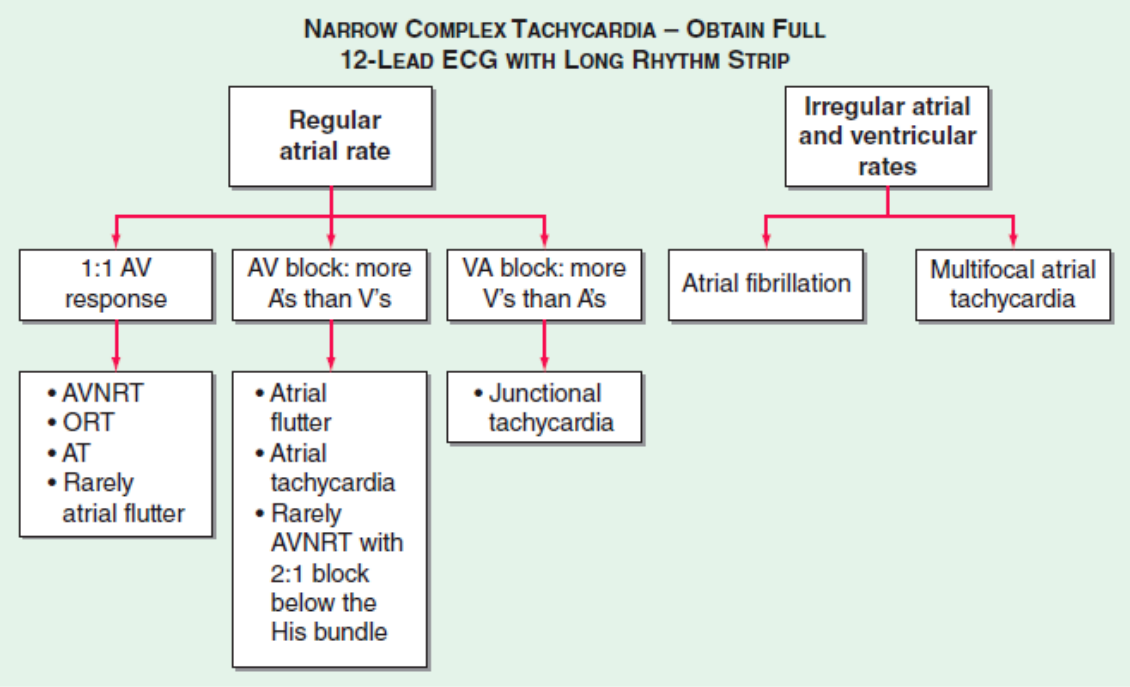
Symptoms of supraventricular arrhythmia vary depending on the rate, duration, associated heart disease, and comorbidities and include palpitations, chest pain, dyspnea, diminished exertional capacity, and occasionally syncope. Rarely, a supraventricular arrhythmia precipitates cardiac arrest in patients with the Wolff-Parkinson-White syndrome or severe heart disease, such as hypertrophic cardiomyopathy.

Initial Evaluation

Diagnosis requires obtaining an electrocardiogram (ECG) at the time of symptoms. When the arrhythmia is ongoing, the ECG usually establishes or suggests the diagnosis (Figs. 241-1, 241-2, 241-3 and Table 241-2). Treatment is determined by the type of arrhythmia and its hemodynamic effect. For transient arrhythmias, ambulatory ECG recording is warranted. Exercise testing is useful for assessing exercise-related symptoms. Occasionally an invasive electrophysiology study is warranted to provoke the arrhythmia with pacing, confirm the mechanism, and usually, perform catheter ablation.

FIGURE 241-1

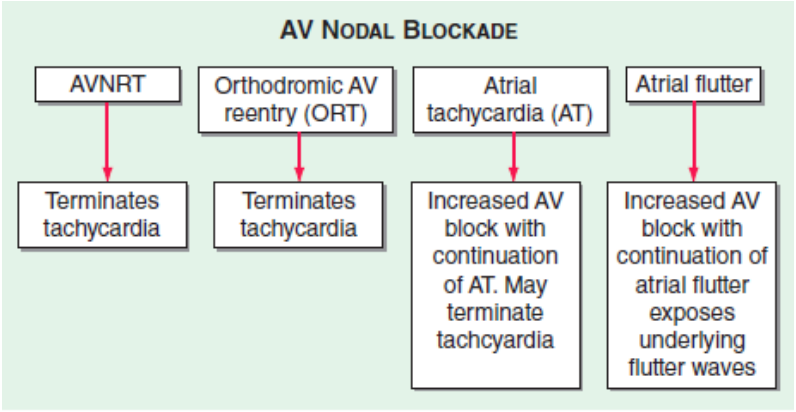
Diagnostic possibilities based on the appearance of the 12-lead ECG recorded during an episode of SVT. AVNRT, AV nodal reentry tachycardia; ORT, orthodromic AV reentry tachycardia; AT, focal atrial tachycardia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 241-2

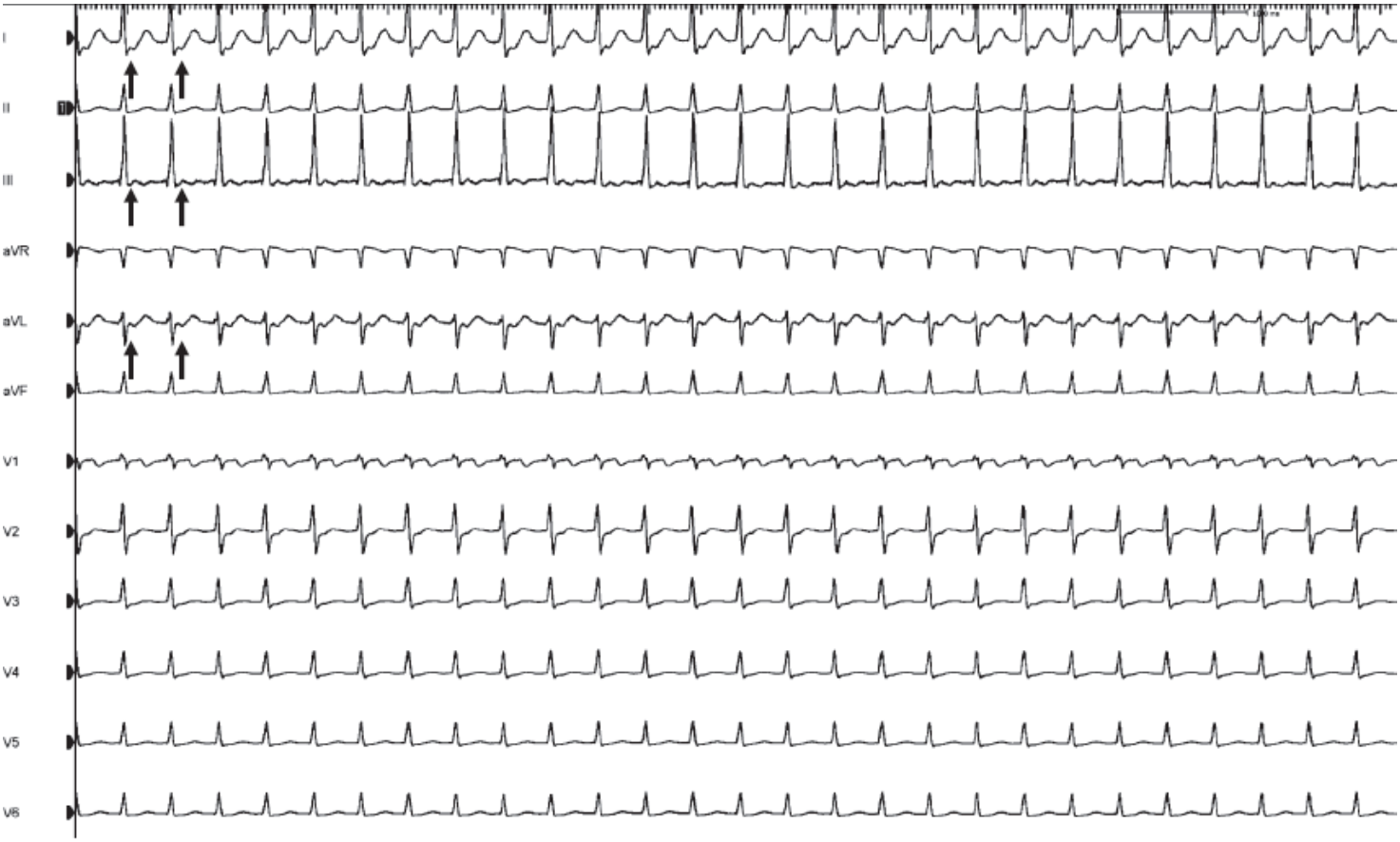
Diagnostic effect of increasing AV node blockade with vagal maneuvers, adenosine, verapamil or beta blockers.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 241-3

12-lead ECG of ORT due to an accessory pathway between the left ventricle and left atrium. The ECG is from an otherwise healthy young woman who had recurrent episodes of tachycardia, sometimes terminating with vagal maneuvers and always terminated by administration of adenosine. Termination with AV node blocking agents make atrial flutter and atrial tachycardia unlikely and are consistent with mechanisms dependent on AV nodal conduction, such as ORT or AVNRT. The 12-lead ECG shows a narrow complex tachycardia with a regular atrial rate and 1:1 atrioventricular response. Using the algorithm shown in Fig. 241-2, the most likely mechanisms are AVNRT or ORT. The P wave can be seen in the ST segment (arrows) and appears to be positive in lead III and negative in leads I and aVL, which suggests a left free wall origin. Ablation of the left free wall accessory pathway eliminated further episodes of SVT. **(See Chap. 244 for further discussion of ORT and accessory pathways).**



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 241-2

Usual Relation of P-wave to QRS in Paroxysmal Supraventricular Tachycardias (see also Fig. 241-3)

Regular tachycardia with 1:1 AV conduction:

– AVNRT either has no discernible p-waves because they are synchronous with the QRS, or p-waves that are negative in II, III, aVF immediately following the QRS (referred to as short R-P tachycardia). Atypical forms may have a longer R-P interval.

– ORT has p-waves following the QRS, although they may be difficult to define when simultaneous with the T-wave. (See example in Fig. 241-3.)

– AT typically has p-waves preceding the QRS (R-P) interval > P-R interval. P wave morphology depends on the focus location and is different compared to sinus rhythm unless the focus is near the sinus node.

Note: Further analysis of SVT with a regular ventricular rate may allow discrimination among the three most common forms: atrioventricular nodal reentry tachycardia (AVNRT), orthodromic reentrant tachycardia (ORT), or atrial tachycardia (AT).

Paroxysmal supraventricular tachycardia is most commonly encountered in patients who do not have structural heart disease. Other supraventricular arrhythmias, particularly atrial fibrillation, are associated with a variety of heart diseases. At initial evaluation history and examination should assess possible underlying heart disease. Any abnormal findings may warrant further cardiac evaluation.

The most common supraventricular tachycardia is sinus tachycardia in response to physiologic stress, such as exercise, but it can also be a manifestation acute illness. The first step in diagnosis of supraventricular tachycardia is to consider the possibility of sinus tachycardia. **(Chap. 242)**. Therapy is then determined by the clinical findings. If the arrhythmia is ongoing and is not due to sinus tachycardia, initial assessment determines whether immediate therapy is needed to terminate the arrhythmia or slow the rate. Arrhythmias causing hypotension, impaired consciousness, angina, or heart failure warrant immediate therapy, guided by the type of arrhythmia.

FURTHER READING

Appelboam A et al: Postural modification to the standard Valsalva manoeuvre for emergency treatment of supraventricular tachycardias (REVERT): A randomised controlled trial. Lancet 386:1747, 2015. [PubMed: 26314489]

Page RL et al: 2015 ACC/AHA/HRS Guideline for the Management of Adult Patients with Supraventricular Tachycardia: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. Circulation 133:e506, 2016. [PubMed: 26399663]

Chapter 242: Physiologic and Nonphysiologic Sinus Tachycardia

Gregory F. Michaud; William G. Stevenson

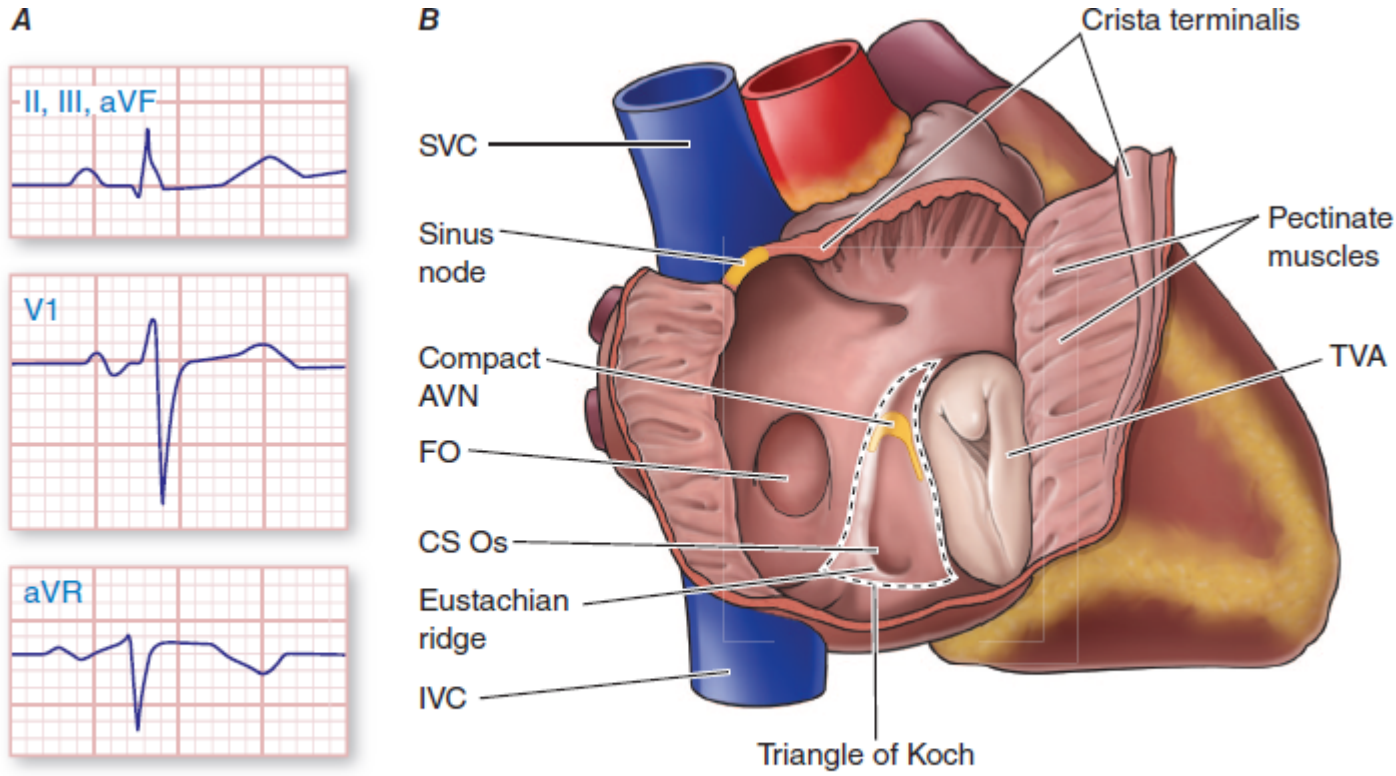
FIGURE 242-1

PHYSIOLOGIC SINUS TACHYCARDIA

The sinus node is comprised of a group of cells dispersed within the superior aspect of the thick ridge of muscle known as the crista terminalis where the posterior smooth atrial wall derived from the sinus venosus meets the trabeculated anterior portion of the right atrium (**Fig. 242-1**). Sinus p waves are characterized by a frontal plane axis directed inferiorly and leftward, with positive p waves in leads II, III, and aVF; a negative p wave in aVR; and an initially positive biphasic p wave in V1. Normal sinus rhythm has a range of rates between 60–100 beats/min. Sinus tachycardia (>100 beats/min) typically occurs in response to sympathetic stimulation and vagal withdrawal, whereby the rate of spontaneous depolarization of the sinus node increases and the focus of earliest activation within the node typically shifts more leftward and closer to the superior septal aspect of the crista terminalis, thus producing taller p waves in the inferior limb leads when compared to normal sinus rhythm. Sinus bradycardia is defined as rates less than 60 beats/min; however, bradycadia can be normal during sleep and in fit individuals.

FIGURE 242-1

Right atrial anatomy pertinent to normal sinus rhythm and supraventricular tachycardia. ***A.*** Typical P-wave morphology during normal sinus rhythm based on standard 12-lead electrocardiogram. There is a positive P wave in leads II, III, and aVF; biphasic, initially positive P wave in V_1; and negative P wave in aVR. ***B.*** Right atrial anatomy seen from a right lateral perspective with the lateral wall opened to view the septum. AVN, atrioventricular node; CS Os, coronary sinus ostium; FO, fossa ovalis; IVC, inferior vena cava; SVC, superior vena cava; TVA, tricuspid valve annulus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Sinus tachycardia is considered physiologic when it is an appropriate response to exercise, stress, or illness. Sinus tachycardia can be difficult to distinguish from focal atrial tachycardia (see below) that originates near the sinus node. A causative factor (such as exertion) and a gradual increase and decrease in rate favor sinus tachycardia, whereas abrupt onset and offset favor atrial tachycardia. The distinction can be difficult and occasionally requires extended ECG monitoring or even invasive electrophysiology study. Treatment for physiologic sinus tachycardia is aimed at the underlying condition (Table 242-1), but frequently no therapy is necessary.

TABLE 242-1

Common Causes of Physiologic Sinus Tachycardia

1. Exercise
2. Acute illness with fever, infection, pain
3. Hypovolemia, anemia
4. Hyperthyroidism
5. Pulmonary insufficiency
6. Drugs that have sympathomimetic, vagolytic, or vasodilator properties, e.g., albuterol, theophylline, tricyclic antidepressants, nifedipine, hydralazine
7. Pheochromocytoma

NONPHYSIOLOGIC SINUS TACHYCARDIA

Inappropriate sinus tachycardia is an uncommon condition in which the sinus rate increases spontaneously at rest or out of proportion to physiologic stress or exertion and is within a spectrum of ill-defined conditions associated with autonomic dysregulation. Affected individuals are often women in the third or fourth decade of life. Fatigue, dizziness, and even syncope may accompany palpitations, which can be disabling. Additional symptoms of chest pain, headaches, and gastrointestinal upset are common. It must be distinguished from appropriate sinus tachycardia and from focal atrial tachycardia. The distinction between physiologic sinus tachycardia due to an anxiety disorder and inappropriate sinus tachycardia can be difficult. Therapy is often ineffective or poorly tolerated. Careful titration of beta blockers and/or calcium channel blockers may reduce symptoms. Clonidine and serotonin reuptake inhibitors have also been used. Ivabradine, a drug that blocks the I_f current that causes sinus node depolarization, is now approved in the United States for use in heart failure, but it has also been effective in the treatment of inappropriate sinus tachycardia. Catheter ablation of the sinus node has been performed, but long-term control of symptoms is usually poor, and it often leaves young individuals with a permanent pacemaker.

Postural orthostatic tachycardia syndrome (POTS) is characterized by symptomatic sinus tachycardia that occurs with postural change from a supine position to standing. The sinus rate increases by 30 beats/min or to >120 beats/min within 10 min of standing and in the absence of hypotension. Symptoms are often similar to those in patients with inappropriate sinus tachycardia. POTS is sometimes due to autonomic dysfunction following a viral illness and may resolve spontaneously over 3–12 months. Volume expansion with salt supplementation, oral fludrocortisone, compression stockings, and the α-agonist midodrine, often in combination, can be helpful. Exercise training has also been purported to improve symptoms.

FURTHER READING

George SA et al: The international POTS registry: Evaluating the efficacy of an exercise training intervention in a community setting. Heart Rhythm 13:943, 2016.

Nwazue VC et al: Postural tachycardia syndrome and inappropriate sinus tachycardia: Role of autonomic modulation and sinus node automaticity. J Am Heart Assoc 3:e000700, 2014. [PubMed: 24721800]

Salazar Adum JP, Arora R: Treatment for inappropriate sinus tachycardia. Am J Ther 24:e574, 2017.

Sheldon RS et al: 2015 Heart Rhythm Society expert consensus statement on the diagnosis and treatment of postural tachycardia syndrome, inappropriate sinus tachycardia, and vasovagal syncope. Heart Rhythm 12:e41–63, 2015. [PubMed: 25980576]

Chapter 243: Focal Atrial Tachycardia

Gregory F. Michaud; William G. Stevenson

INTRODUCTION

Focal atrial tachycardia (AT) can be due to abnormal automaticity, triggered automaticity, or a small reentry circuit confined to the atrium or atrial tissue extending into a pulmonary vein, the coronary sinus, or vena cava. It can be sustained, nonsustained, paroxysmal, or incessant. Focal AT accounts for ~10% of PSVTs in patients referred for catheter ablation. Nonsustained AT is commonly observed on 24-h ambulatory ECG recordings, and the prevalence increases with age. In fact, frequent atrial ectopy and nonsustained AT is often a precursor to more significant arrhythmias such as atrial fibrillation and flutter. Though unsustained, frequent atrial ectopy or short bursts of AT may be symptomatic and require therapy similar to that required for focal AT.

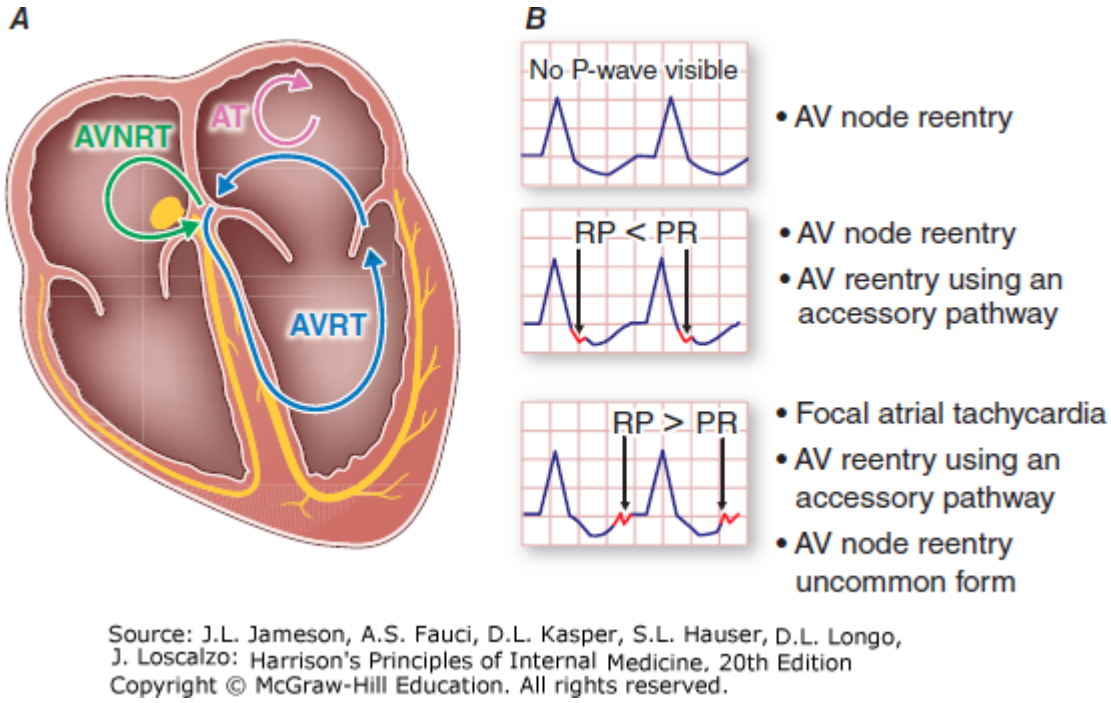
AT can occur in the absence of structural heart disease or may be associated with any condition that causes atrial fibrosis, including prior catheter ablation. Areas of fibrosis can be a nidus for abnormal automaticity from damaged cells or microreentry within zones of slow conduction within and on the border of fibrotic areas. Sympathetic stimulation is a promoting factor and the emergence of AT can be a sign of underlying illness. AT with AV block may occur in digitalis toxicity. Symptoms from AT are highly variable but similar to other supraventricular tachycardias (SVTs). Incessant AT can cause tachycardia-induced cardiomyopathy.

AT typically presents with 1:1 AV conduction or with AV block that can be Wenckebach type conduction or fixed (e.g., 2:1 or 3:1). Because it is not dependent on AV nodal conduction, AT will not terminate with AV block, and the atrial rate will not be affected, which distinguishes AT from most AV nodal–dependent SVTs, such as AV nodal reentry and AV reentry using an accessory pathway (see below). An accelerated warm-up phase after initiation or cool-down phase prior to termination also favors AT rather than AV nodal–dependent SVT, as this is a common observation with triggered automaticity. P waves are often discrete, with an intervening isoelectric segment, in contrast to atrial flutter and macroreentrant AT (see below) because atrial activation from a focal source occurs though a small portion of the tachycardia cycle. When 1:1 conduction to the ventricles is present, the arrhythmia can resemble sinus tachycardia typically with a P-R interval shorter than the R-P interval **(Fig. 243-1)**, ***particularly when sympathetic tone produces rapid AV nodal conduction.*** It can be distinguished from sinus tachycardia by the P-wave morphology, which usually differs from sinus p waves depending on the location of the focus. Focal AT tends to originate in areas of complex atrial anatomy, such as the crista terminalis, valve annuli, atrial septum, and atrial muscle extending along cardiac thoracic veins (superior vena cava, coronary sinus, and pulmonary veins) **(Fig. 243-2)**, and the location can often be

estimated by the P-wave morphology. AT from the atrial septum will frequently have a narrower P-wave duration than sinus rhythm. AT from the left atrium will usually have a monophasic, positive P wave in lead V_1 and negative P waves in I and aVL indicating movement away from the left atrial free wall. AT that originates from superior atrial locations, such as the superior vena cava or superior pulmonary veins, will be positive in the inferior limb leads II, III, and aVF, whereas AT from a more inferior location, such as the ostium of the coronary sinus, will inscribe negative P waves in these same leads. When the focus is in the superior aspect of the crista terminalis, close to the sinus node, however, the p wave will resemble that of sinus tachycardia. Abrupt onset and offset then favor AT rather than sinus tachycardia. Depending on the atrial rate, the P wave may fall on top of the T wave or, during 2:1 conduction, may fall coincident with the QRS. Maneuvers that increase AV block, such as carotid sinus massage, Valsalva maneuver, or administration of AV nodal–blocking agents, such as adenosine, are useful to create AV block that will expose the p wave (**Fig. 243-3**).

FIGURE 243-1

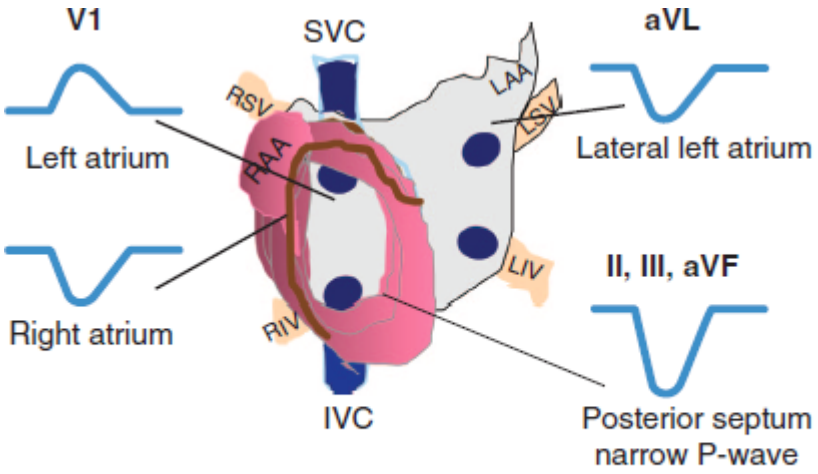
Common mechanisms underlying paroxysmal supraventricular tachycardia along with typical R-P relationships. ***A.*** Schematic showing a four-chamber view of the heart with atrioventricular node in *green* and an accessory pathway between the left atrium and left ventricle in *blue*. Atrial tachycardia (AT; *red circuit*) is confined completely to atrial tissue. Atrioventricular nodal reentry tachycardia (AVNRT; *blue circuit*) uses atrioventricular (AV) nodal and perinodal atrial tissue. Atrioventricular reentry tachycardia (AVRT; *black circuit*) uses atrial and ventricular tissue, accessory pathway, AV node, and specialized conduction fibers (His-Purkinje) as part of the reentry circuit. ***B.*** Typical relation of the P wave to QRS, commonly described as the R-P to P-R relationships for the different tachycardia mechanisms.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 243-2

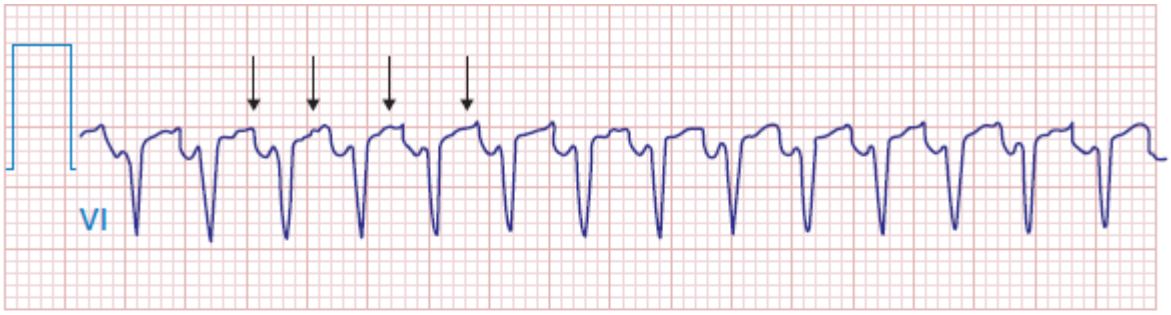
Location of focal atrial tachycardia focus estimated by P-wave morphology. LAA, left atrial appendage; LIV, left inferior pulmonary vein; LSV, left superior pulmonary vein; RAA, right atrial appendage; RIV, right inferior pulmonary vein; RSV, right superior pulmonary vein; SVC, superior vena cava.



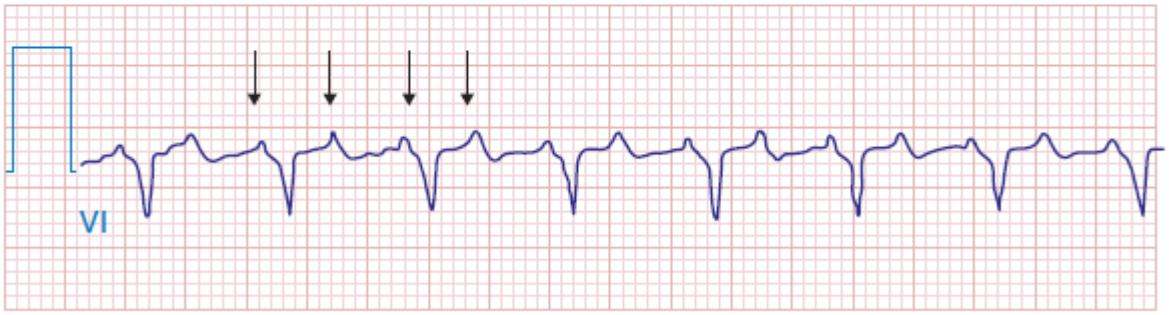
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 243-3

Atrial tachycardia (AT) with 1:1 and 2:1 atrioventricular (AV) conduction. *Arrows* indicate P waves. ***A.*** AT with 1:1 AV relationship and R-P > P-R. ***B.*** Same AT with 2:1 AV relationship after AV nodal–blocking agent administered. *(Adapted from F Marchlinski: The tachyarrhythmias. in DL Longo et al [eds]: Harrison's Principles of Internal Medicine, 18th ed. New York, McGraw-Hill, 2012, pp 1878–1900.)*



A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Acute management of sudden-onset, sustained AT is the same as for other forms of PSVT (**Chaps. 241 and 244**), but the response to pharmacologic therapy is variable, likely depending on the mechanism. For AT due to reentry, administration of adenosine or vagal maneuvers may transiently increase AV block without terminating tachycardia. Some ATs terminate with a sufficient dose of adenosine, consistent with triggered activity as the mechanism. Cardioversion can be effective in some, but fails in others because of immediate recurrence, suggesting automaticity as the mechanism in these cases. Beta blockers and calcium channel

blockers may slow the ventricular rate by increasing AV block, which can improve tolerance of the arrhythmias, but large doses are sometimes required. Potential precipitating factors and intercurrent illness should be sought and corrected. Underlying heart disease should be considered and excluded.

For patients with recurrent episodes, beta blockers, calcium channel blockers such as diltiazem or verapamil, and antiarrhythmic drugs such as flecainide, propafenone, disopyramide, sotalol, and amiodarone can be effective, but potential toxicities and adverse effects often warrant avoidance for long term use (**Tables 243-1, 243-2, and 243-3**). Catheter ablation targeting the AT focus is effective in more than 80% of patients and is recommended for recurrent symptomatic AT when drugs fail or are not desired or for incessant AT causing tachycardia-induced cardiomyopathy.

TABLE 243-1

Commonly Used Antiarrhythmic Agents—Intravenous Dose Range/Primary Indication

DRUG	LOADING	MAINTENANCE	PRIMARY INDICATION	CLASS[a]
Adenosine	6–18 mg (rapid bolus)	N/A	Terminate reentrant SVT involving AV node	—
Amiodarone	15 mg/min for 10 min, 1 mg/min for 6 h	0.5–1 mg/min	AF, AFL, SVT, VT/VF	III
Digoxin	0.25 mg q2h until 1 mg total	0.125–0.25 mg/d	AF/AFL rate control	—
Diltiazem	0.25 mg/kg over 3–5 min (max 20 mg)	5–15 mg/h	SVT, AF/AFL rate control	IV
Esmolol	500 µg/kg over 1 min	50 µg/kg per min	AF/AFL rate control	II
Ibutilide	1 mg over 10 min if over 60 kg	N/A	Terminate AF/AFL	III
Lidocaine	1–3 mg/kg at 20–50 mg/min	1–4 mg/min	VT	IB
Metoprolol	5 mg over 3–5 min × 3 doses	1.25–5 mg q6h	SVT, AF rate control; exercise-induced VT; long QT	II
Procainamide	15 mg/kg over 60 min	1–4 mg/min	Convert/prevent AF/VT	IA
Quinidine	6–10 mg/kg at 0.3–0.5 mg/kg per min	N/A	Convert/prevent AF/VT	IA
Verapamil	5–10 mg over 3–5 min	2.5–10 mg/h	SVT, AF rate control	IV

[a]Classification of antiarrhythmic drugs: class I—agents that primarily block inward sodium current; class IA agents also prolong action potential duration; class II—antisympathetic agents; class III—agents that primarily prolong action potential duration; class IV—calcium channel–blocking agents.

Abbreviations: AF, atrial fibrillation; AFL, atrial flutter; AV, atrioventricular; SVT, supraventricular tachycardia; VF, ventricular fibrillation; VT, ventricular tachycardia.

TABLE 243-2

Commonly Used Antiarrhythmic Agents: Chronic Oral Dosing/Primary Indications

DRUG	DOSING ORAL, mg, MAINTENANCE	HALF-LIFE, h	PRIMARY ROUTE(S) OF METABOLISM/ELIMINATION	MOST COMMON INDICATION	CLASS[a]
Acebutolol	200–400 q12h	6–7	Renal/hepatic	AF rate control/SVT	II
				Long QT/RVOT VT	
Amiodarone	100–400 qd	40–55 d	Hepatic	AF/VT prevention	III[b]
Atenolol	25–100 per d	6–9	Renal	AF rate control/SVT	II
				Long QT/RVOT VT	
Digoxin	0.125–0.25 qd	38–48	Renal	AF rate control	—
Diltiazem	30–60 q6h	3–4.5	Hepatic	AF rate control/SVT	IV
Disopyramide	100–300 q6–8h	4–10	Renal 50%/hepatic	AF/SVT prevention	Ia
Dofetilide	0.125–0.5 q12h	10	Renal	AF prevention	III
Dronedarone	400 q12h	13–19	Hepatic	AF prevention	IIIb
Flecainide	50–200 q12h	7–22	Hepatic 75%/renal	AF/SVT/VT prevention	Ic
Metoprolol	25–100 q6h	3–8	Hepatic	AF rate control/SVT	II

DRUG	DOSING ORAL, mg, MAINTENANCE	HALF-LIFE, h	PRIMARY ROUTE(S) OF METABOLISM/ELIMINATION	MOST COMMON INDICATION	CLASS[a]
				Long QT/RVOT VT	
Mexiletine	150–300 q8–12h	10–14	Hepatic	VT prevention	Ib
Nadolol	40–240 per d	10–24	Renal	Same as metoprolol	II
Propafenone	150–300 q8h	2–8	Hepatic	AF/SVT/VT prevention	Ic
Quinidine	300–600 q6h	6–8	Hepatic 75%/renal	AF/SVT/VT prevention	Ia
Sotalol	80–160 q12h	12	Renal	AF/VT prevention	III
Verapamil	80–120 q6–8h	4.5–12	Hepatic/renal	AF rate control/RVOT VT	IV
				Idiopathic LV VT	

[a]Classification of antiarrhythmic drugs: class I—agents that primarily block inward sodium current; class II—antisympathetic agents; class III—agents that primarily prolong action potential duration; class IV—calcium channel-blocking agents. [b]Amiodarone and dronedarone both are grouped in class III, but both also have class I, II, and IV properties.

Abbreviations: AF, atrial fibrillation; LV, left ventricular; RVOT, right ventricular outflow tract; SVT, supraventricular tachycardia; VT, ventricular tachycardia.

TABLE 243-3

Common and Proarrhythmic Toxicities of Antiarrhythmic Agents

DRUG	POTENTIAL PROARRHYTHMIC TOXICITIES	COMMON TOXICITIES
Amiodarone	Sinus bradycardia, AV block, increase in defibrillation threshold. Rare: long QT and torsades des pointes, incessant slow VT in heart disease	Tremor, peripheral neuropathy, pulmonary fibrosis or inflammation, hypo- and hyperthyroidism, hepatitis, photosensitivity
Adenosine	Transient profound pauses, atrial fibrillation	Cough, flushing, chest pain, anxiety
Digoxin	AV block, fascicular tachycardia, accelerated junctional rhythm, atrial tachycardia with AV block	Anorexia, nausea, vomiting, visual changes
Disopyramide	Long QT and torsades des pointes, 1:1 ventricular response to atrial flutter	Anticholinergic effects, acute urinary retention (males), negative inotropy
Dofetilide	Long QT and torsades des pointes	Nausea
Dronedarone	Bradyarrhythmias and AV block, long QT and torsades des pointes (rare)	Gastrointestinal intolerance, exacerbation of heart failure
Flecainide	1:1 Ventricular response to atrial flutter; increased risk of ventricular tachycardias in patients with structural heart disease; sinus bradycardia	Dizziness, nausea, headache, decreased myocardial contractility
Ibutilide	Long QT and torsades des pointes	Nausea
Lidocaine	Slow VT in some patients with structural heart disease	Dizziness, confusion, delirium, seizures, coma
Mexiletine	Slow VT in patients with structural heart disease	Ataxia, tremor, gait disturbances, rash, nausea
Procainamide	Long QT and torsades des pointes, accelerated ventricular rate in AF or flutter	Lupus erythematosus–like syndrome (more common in slow acetylators), anorexia, nausea, neutropenia

DRUG	POTENTIAL PROARRHYTHMIC TOXICITIES	COMMON TOXICITIES
Propafenone	1:1 Ventricular response to atrial flutter; increased risk ventricular tachycardias in patients with structural heart disease; sinus bradycardia	Taste disturbance, dyspepsia, nausea, vomiting
Quinidine	Long QT and torsades des pointes, accelerated ventricular rate in AF or flutter	Diarrhea, nausea, vomiting, cinchonism, thrombocytopenia
Sotalol	Long QT and torsades des pointes	Hypotension, bronchospasm from β-blocking effect

Abbreviations: AF, atrial fibrillation; AV, atrioventricular; VT, ventricular tachycardia.

FURTHER READING

Link MS: Clinical practice. Evaluation and initial treatment of supraventricular tachycardia. N Engl J Med 367:1438, 2012. [PubMed: 23050527]

Liu CF et al: Unifying algorithm for mechanistic diagnosis of atrial tachycardia. Circ Arrhythm Electrophysiol 9: pii: e004028, 2016. [PubMed: 27516463]

Page RL et al: 2015 ACC/AHA/HRS Guideline for the Management of Adult Patients with Supraventricular Tachycardia: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. Circulation 133:e506, 2016. [PubMed: 26399663]

Chapter 244: Paroxysmal Supraventricular Tachycardias

Gregory F. Michaud; William G. Stevenson

INTRODUCTION

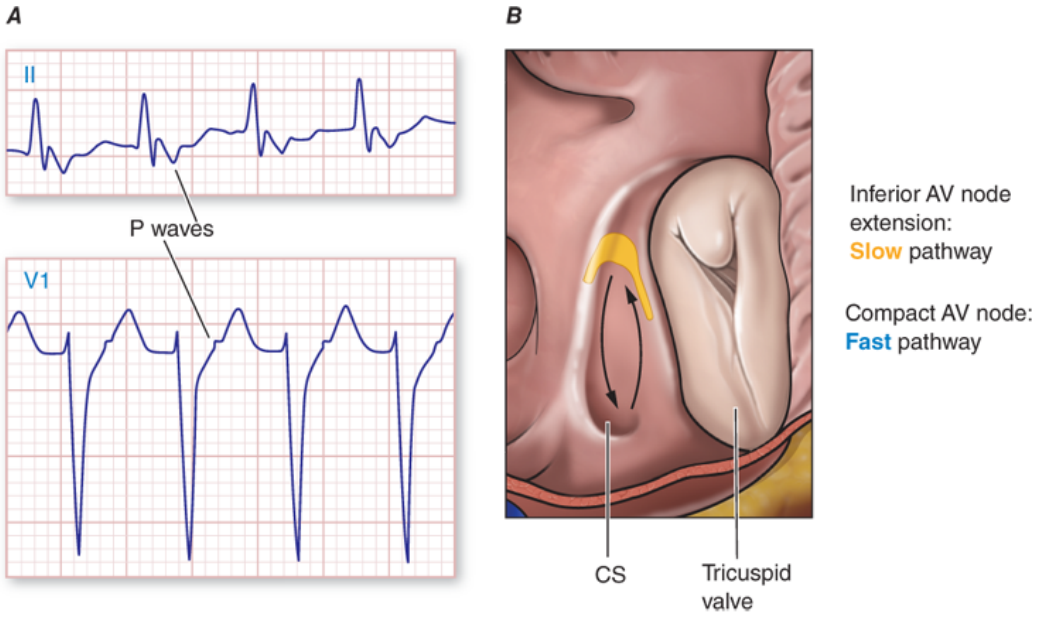
Atrioventricular Nodal Reentry Tachycardia

AV nodal reentry tachycardia (AVNRT) is the most common form of paroxysmal supraventricular tachycardia (PSVT), representing ~60% of cases referred for catheter ablation. It most commonly manifests in the second to fourth decades of life, often in women. It is often well tolerated, but rapid tachycardia, particularly in the elderly, may cause angina, pulmonary edema, hypotension, or syncope. It is not usually associated with structural heart disease.

The mechanism is reentry involving the AV node and the perinodal atrium, made possible by the existence of multiple pathways for conduction from the atrium into the AV node that are capable of conduction in two directions **(Fig. 244-1)**. Most forms of AVNRT utilize a slowly conducting AV nodal pathway (right inferior extension) that extends from the compact AV node near the His bundle, inferiorly along the tricuspid valve annulus to the floor of the coronary sinus. The reentry wavefront propagates up this slowly conducting pathway to the compact AV node and then exits from the fast pathway at the top of the AV node. The path back to the slow pathway probably involves the left atrial septum which has connections to the coronary sinus musculature. More unusual forms of AVNRT utilize a left inferior extension that connects to the compact AV node through the roof of the coronary sinus, or in extremely rare cases, directly from the mitral valve annulus avoiding the coronary sinus musculature altogether. In typical forms, the conduction time from the compact AV node region to the atrium is similar to that from the compact node to the His bundle and ventricles, such that atrial activation occurs at about the same time as ventricular activation. The p wave is therefore inscribed during, slightly before, or slightly after the QRS and can be difficult to discern. Often the P wave is seen at the end of the QRS complex as a pseudo-r' in lead V_1 and pseudo-S waves in leads II, III, and aVF **(Fig. 244-1A)**. More unusual forms of AVNRT have P waves falling later, anywhere between QRS complexes, in which case an inverted P wave is seen in the inferior limb leads as seen in **Fig. 244-2** where the inverted P wave is seen in the T wave. The rate can vary with sympathetic tone. Simultaneous atrial and ventricular contraction results in atrial contraction against a closed tricuspid valve producing cannon a wave visible in the jugular venous pulse often perceived as a fluttering sensation in the neck. Elevated venous pressures may also lead to release of natriuretic peptides that cause post-tachycardia diuresis. In contrast to ATs, maneuvers or medications that produce AV nodal block terminate the arrhythmia.

FIGURE 244-1

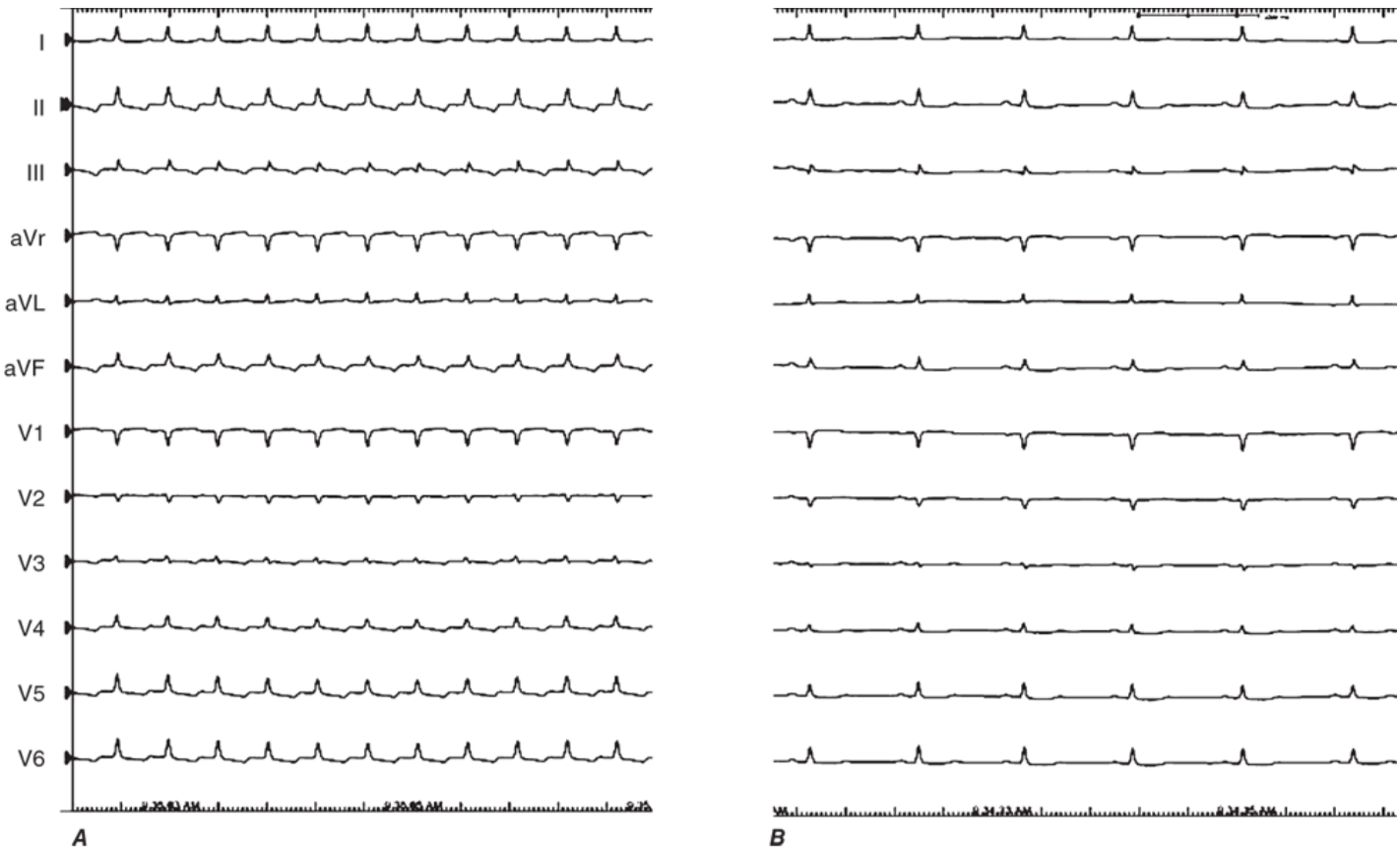
Atrioventricular (AV) node reentry. ***A.*** Leads II and V_1 are shown. P waves are visible at the end of the QRS complex and are negative in lead II, and may give the impression of S waves in the inferior limb leads II, III, and aVF and an R' in lead V_1. ***B.*** Stylized version of the AV nodal reentry circuit within the triangle of Koch (**see Fig. 242-1**) that involves AV node and its extensions along with perinodal atrial tissue.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 244-2

Comparison of 12-lead ECG tracings showing SVT (Panel A) and normal sinus rhythm (Panel B). The P wave is observed at the end of the T wave and morphology can be inferred from comparing to sinus rhythm. P waves are inverted in the inferior limb leads (II, III, and aVF), positive in V1, I, and aVL consistent with conduction retrogradely through the AV junction. In typical forms of AVNRT, the P wave is not visible or is seen at the end of the QRS complex.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Acute treatment is the same as for other forms of PSVT (discussed below). Whether ongoing therapy is warranted depends on the severity of symptoms and frequency of episodes. Reassurance and instruction as to how to perform the Valsalva maneuver to terminate episodes are sufficient for many patients. Administration of an oral beta blocker, verapamil, or diltiazem at the onset of an episode has been used to facilitate termination. Chronic therapy with these medications or flecainide is an option if prophylactic therapy is needed. Catheter ablation of the slow AV nodal pathway is recommended for patients with

recurrent or severe episodes or when drug therapy is ineffective, not tolerated, or not desired by the patient. Catheter ablation is curative in >95% of patients. The major risk is atrioventricular (AV) block requiring permanent pacemaker implantation, which occurs in <1% of patients.

Junctional Tachycardia

Junctional ectopic tachycardia (JET) is due to automaticity within the AV node. It is rare in adults and more frequently encountered as an incessant tachycardia in children, often in the perioperative period of surgery for congenital heart disease. It presents as a narrow QRS tachycardia, often with ventriculoatrial (VA) block, such that AV dissociation is present. JET can occur as a manifestation of increased adrenergic tone and may be seen after administration of isoproterenol, particularly after catheter ablation in the perinodal region. It may also occur for a short period of time after ablation for AVNRT.

Accelerated junctional rhythm is a junctional automatic rhythm between 50 and 100 beats/min. Initiation may occur with gradual acceleration in rate, suggesting an automatic focus, or after a premature ventricular contraction, suggesting a focus of triggered automaticity. VA conduction is usually present, with p-wave morphology and timing such that it resembles AVNRT at a slow rate. It can be related to increased sympathetic tone and may produce palpitations. It usually does not require specific therapy.

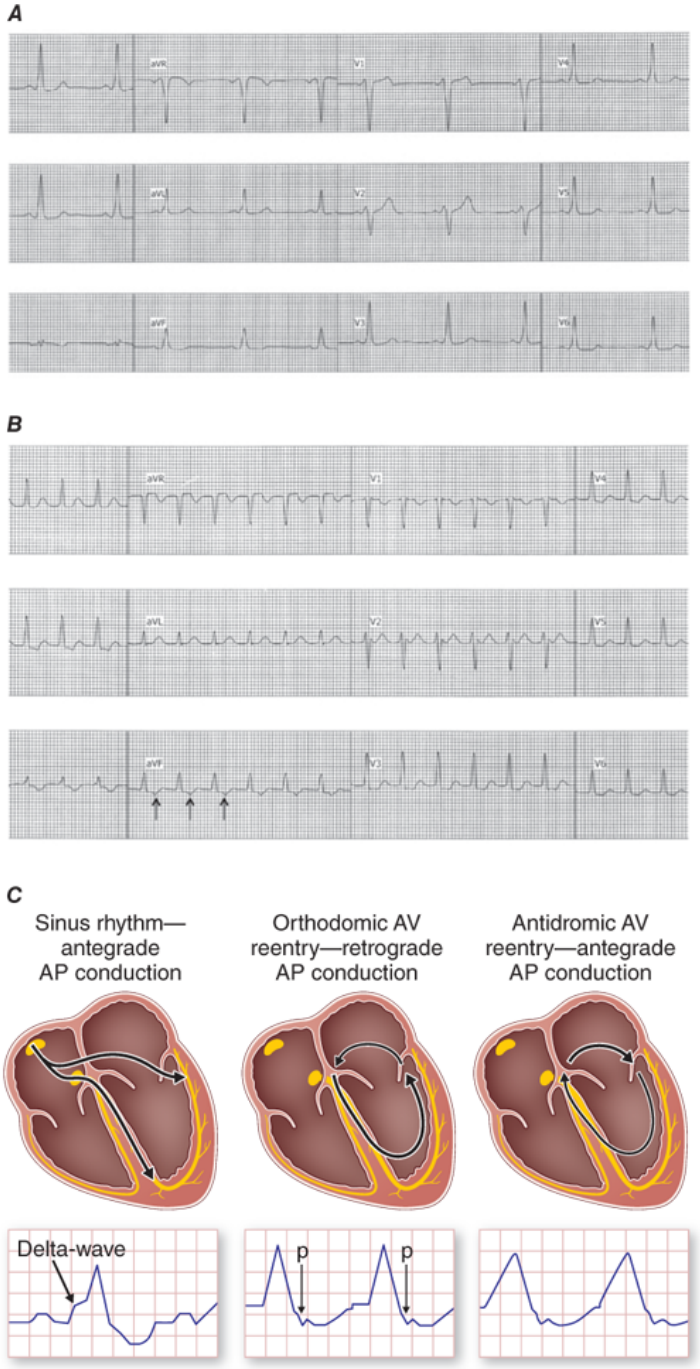
ACCESSORY PATHWAYS AND THE WOLFF-PARKINSON-WHITE SYNDROME

Accessory pathways (APs) occur in 1 in 1500–2000 people and are associated with a variety of arrhythmias including narrow-complex PSVT, wide-complex tachycardias, and, rarely, sudden death. Most patients have structurally normal hearts, but APs are associated with Ebstein's anomaly of the tricuspid valve and forms of hypertrophic cardiomyopathy including *PRKAG2* mutations, Danon's disease, and Fabry's disease.

APs are abnormal connections that allow conduction between the atrium and ventricles across the AV ring **(Fig. 244-3)**. They are present from birth and are due to failure of complete partitioning of atrium and ventricle by the fibrous AV rings. They occur across either an AV valve annulus or the septum, most frequently between the left atrium and free wall of the left ventricle, followed by posteroseptal, right free wall, and anteroseptal locations. If the impulse from the sinus node conducts through the AP to the ventricle (antegrade) before the impulse conducts through the AV node and His bundle, then the ventricles are preexcited during sinus rhythm, and the ECG shows a short P-R interval (<0.12 s), slurred initial portion of the QRS (delta wave), and prolonged QRS duration produced by slow conduction through direct activation of ventricular myocardium over the AP **(Fig. 244-3A)**. The morphology of the QRS and delta wave is determined by the AP location **(Fig. 244-4)** and the degree of fusion between the excitation wavefronts from conduction over the AV node and conduction over the AP. Right-sided pathways preexcite the right ventricle, producing a left bundle branch block–like configuration in lead V_1, and often create marked preexcitation because of relatively close proximity of the AP to the sinus node (Fig. 244-4). Left-sided pathways preexcite the left ventricle and may produce a right bundle branch–like configuration in lead V_1 and a negative delta wave in aVL, indicating initial depolarization of the lateral portion of the left ventricle that can mimic q waves of lateral wall infarction (Fig. 244-4). Because of the relatively large distance between the sinus node and left free wall APs, preexcitation may be minimal or absent on 12-lead ECG. Preexcitation due to an AP at the diaphragmatic surface of the heart, typically in the paraseptal region, produces delta waves that are negative in leads III and aVF, mimicking the q waves of inferior wall infarction (Fig. 244-4). Preexcitation can be intermittent and disappear during exercise as conduction over the AV node accelerates and may take over ventricular activation completely.

FIGURE 244-3

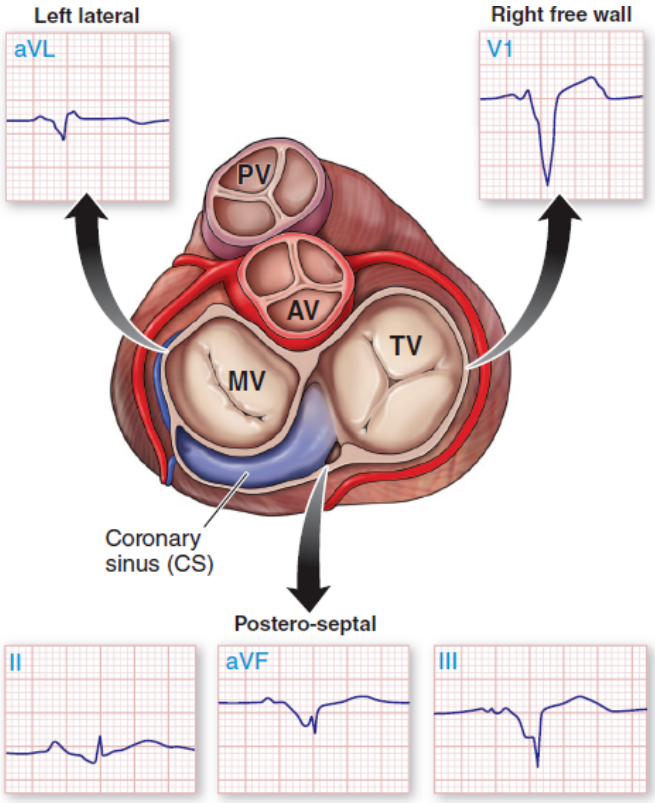
Wolff-Parkinson-White (WPW) syndrome. ***A.*** A 12-lead electro-cardiogram in sinus rhythm (SR) of a patient with WPW demonstrating short P-R interval, delta waves, and widened QRS complex. This patient had an anteroseptal location of the AP. ***B.*** Orthodromic AV reentry in a patient with WPW syndrome using a posteroseptal AP. Note the P waves in the ST segment (*arrows*) seen in lead III and normal appearance of QRS complex. ***C.*** Three most common rhythms associated with WPW syndrome: sinus rhythm demonstrating antegrade conduction over the AP and AV node; orthodromic AVRT using retrograde conduction over the AP and antegrade conduction over the AV node; and antidromic AVRT using retrograde conduction over the AV node and antegrade conduction over the AP. AP, accessory pathway; AV, atrioventricular; AVRT, atrioventricular reentry tachycardia; WPW, Wolff-Parkinson-White.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 244-4

Potential locations for accessory pathways in patients with Wolff-Parkinson-White Syndrome and typical QRS appearance of delta waves that can mimic underlying structural heart disease such as myocardial infraction of bundle branch block. AV, aortic valve; MV, mitral valve; PV, pulmonary valve; TV, tricuspid valve.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Wolff-Parkinson-White (WPW) syndrome is defined as a preexcited QRS during sinus rhythm and episodes of PSVT. There are a number of variations of APs, which may not cause preexcitation and/or arrhythmias. Concealed APs allow only retrograde conduction, from ventricle to atrium, so no preexcitation is present during sinus rhythm, but SVT can occur. Other unusual forms of APs occur. Fasciculoventricular connections between the His bundle and ventricular septum produce preexcitation but do not cause arrhythmia, probably because the circuit is too short to promote reentry. Atriofascicular pathways, also known as Mahaim fibers, probably represent a duplicate AV node and His-Purkinje system that connect the right atrium to fascicles of the right bundle branch and produce a wide complex tachycardia having a left bundle branch block configuration.

AV Reentry Tachycardia

The most common tachycardia caused by an AP is the PSVT designated *orthodromic AV reentry*. The circulating reentry wavefront propagates from the atrium anterogradely over the AV node and His-Purkinje system to the ventricles and then reenters the atria via retrograde conduction over the AP **(Fig. 244-3B)**. The QRS is narrow or may have typical right or left bundle branch block, but without preexcitation during tachycardia. Because excitation through the AV node and AP are necessary, AV or VA block results in tachycardia termination. During sinus rhythm, preexcitation is seen if the pathway also allows anterograde conduction (Fig. 244-3A). Most commonly, during tachycardia the R-P interval is shorter than the P-R interval and can resemble AVNRT **(see Fig. 242-1)**. Unlike typical AVNRT, P waves always follow the QRS and are never simultaneous with a narrow QRS complex because the ventricles must be activated before the reentry wavefront reaches the AP and conducts back to the atrium. The morphology of the P wave is determined by the pathway location, but can be difficult to assess because it is usually inscribed during the ST segment. The p wave in posteroseptal APs is negative in leads II, III, and aVF, similar to that of AV nodal reentry, but P-wave morphology differs from AV nodal reentry for pathways in other locations (Fig. 244-4). **Figure 241-3** is an example of a pathway that has clear negative P waves in leads I and aVL that was due to an AP inserting in the lateral left atrium.

Occasionally, an AP conducts extremely slowly in the retrograde direction, resulting in tachycardia with a long R-P interval, similar to most ATs. These pathways are usually located in the septal region and have negative p waves in leads II, III, and aVF. Slow AP conduction facilitates reentry, often leading to nearly incessant tachycardia, known as *permanent junctional reciprocating tachycardia* (PJRT). Tachycardia-induced cardiomyopathy can occur. Without an invasive electrophysiology study, it may be difficult to distinguish this form of orthodromic AV reentry from atypical AV nodal reentry or AT.

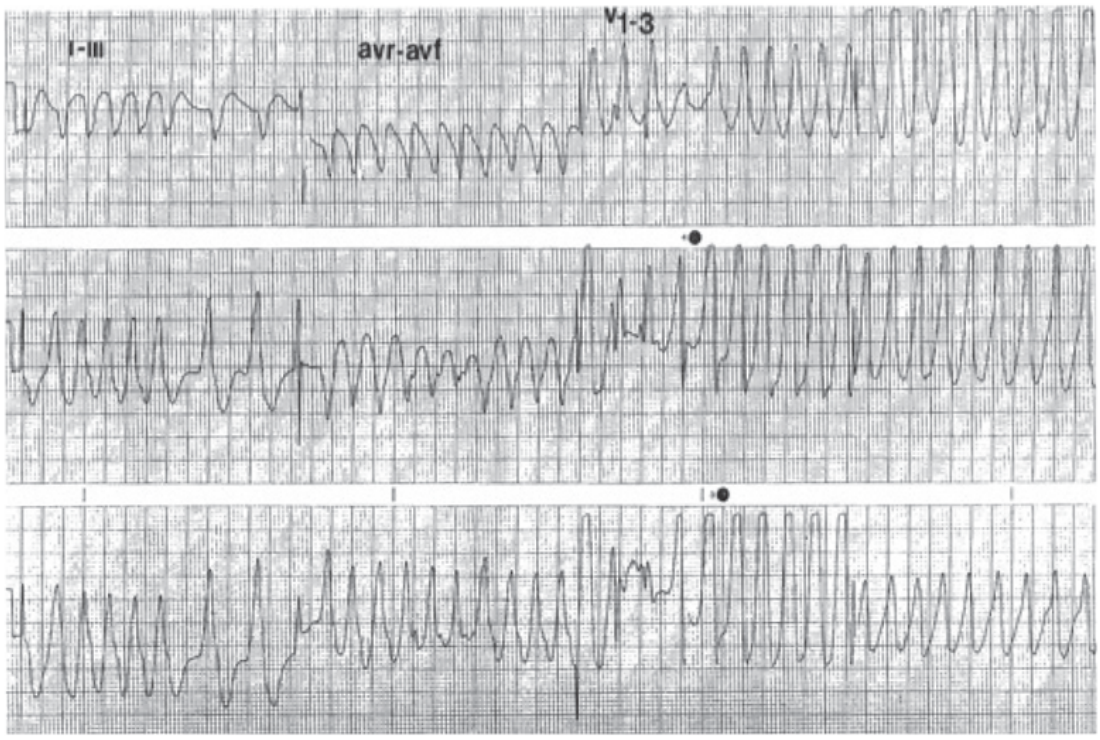
Preexcited Tachycardias

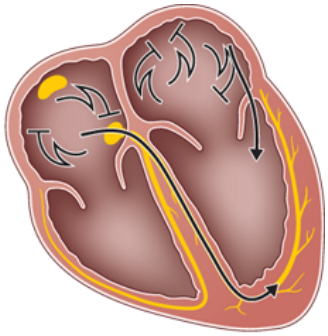
Preexcited tachycardia occurs when the ventricles are activated by antegrade conduction over the AP **(Fig. 244-3C)**. The most common mechanism is *antidromic AV reentry* in which activation propagates from atrium to ventricle via the AP and then conducts retrogradely to the atria via the His-Purkinje system and the AV node (or rarely a second AP). The wide QRS complex is produced entirely via ventricular excitation over the AP because there is no contribution of ventricular activation over more rapidly conducting specialized His-Purkinje fibers. This tachycardia is often indistinguishable from monomorphic ventricular tachycardia. The presence of preexcitation in sinus rhythm suggests the diagnosis.

Preexcited tachycardia also occurs if an AP allows antegrade conduction to the ventricles during AT, atrial flutter, atrial fibrillation (AF) **(Fig. 244-5)**, or AV nodal reentry, otherwise known as bystander AP conduction. AF and atrial flutter are potentially life threatening if the AP allows very rapid repetitive conduction. Approximately 25% of APs causing preexcitation allow minimum R-to-R intervals of <250 ms during AF and are associated with a higher risk of inducing ventricular fibrillation and sudden death. Preexcited AF presents as a wide-complex, very irregular rhythm. During AF, the ventricular rate is determined by the conduction properties of the AP and AV node. The QRS complex can appear quite bizarre and change on a beat-to-beat basis due to the variability in the degree of fusion from activation over the AV node and AP, or all beats may be due to conduction over the AP (Fig. 244-5). Ventricular activation from the Purkinje system may depolarize the ventricular end of the AP and prevent atrial wavefront conduction over the AP. Slowing AV nodal conduction without slowing AP conduction can thereby facilitate AP conduction and dangerously accelerate the ventricular rate. Administration of AV nodal–blocking agents including oral or intravenous verapamil, diltiazem, beta blockers, intravenous adenosine, and intravenous amiodarone are contraindicated during preexcited AF. Rapid preexcited tachycardia should be treated with electrical cardioversion or intravenous procainamide or ibutilide, which may terminate the arrhythmia or slow the ventricular rate.

FIGURE 244-5

Preexcited atrial fibrillation (AF) due to conduction over a left free wall accessory pathway (AP). The electrocardiogram shows rapid irregular QRS complexes that represent fusion between conduction over the atrioventricular node and left free wall AP. Shortest R-R intervals between preexcited QRS complexes of <250 ms, as in this case, indicate a risk of sudden death with this arrhythmia.





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Management of Patients with APs

Acute management of orthodromic AV reentry is discussed below for PSVT. Patients with WPW syndrome may have wide-complex tachycardia due to antidromic AV reentry, orthodromic AV with bundle branch block, or a preexcited tachycardia, and treatment depends on the underlying rhythm.

Initial patient evaluation should include assessment for aggravating factors, including intercurrent illness and factors that increase sympathetic tone. Examination should focus on excluding underlying heart disease. An echocardiogram is reasonable to exclude Ebstein's anomaly and forms of hypertrophic cardiomyopathy that can be associated with APs.

Patients with preexcitation who have symptoms of arrhythmia are at risk for developing AF and sudden death if they have an AP that allows rapid antegrade conduction. The risk of cardiac arrest is in the range of 2 per 1000 patients in adults but is likely greater in children. An invasive electrophysiology study is recommended to assess whether the pathway can support dangerously rapid heart rates if AF were to occur, and is usually combined with potentially curative catheter ablation. Catheter ablation is warranted for recurrent arrhythmias when drugs are ineffective, not tolerated, or not desired by the patient (Fig. 244-5). Efficacy is in the range of 95% depending on the location of the AP. Serious complications occur in fewer than 3% of patients, but can include AV block, cardiac tamponade, thromboemboli, coronary artery injury, and vascular access complications. Procedure mortality is <1 in 1000 patients. Alternatively attempts to gain reassurance that the AP is not high risk with ambulatory monitoring or exercise testing; abrupt loss of conduction

(preexcitation) at physiologic heart rates is consistent with a low risk pathway, but is not completely reliable. Gradual loss of AP conduction with increased sympathetic tone does not reliably indicate low risk since this can occur as AV nodal conduction time shortens.

For patients with concealed APs or known low-risk APs causing orthodromic AV reentry, chronic therapy is guided by symptoms and frequency of events. Vagal maneuvers may terminate episodes, as may a dose of beta blocker, verapamil, or diltiazem taken at the onset of an episode. Chronic therapy with these agents or flecainide can reduce the frequency of episodes in some patients.

Adults who have preexcitation but no arrhythmia symptoms have a risk of sudden death estimated to be 1 per 1000 patient-years. Electrophysiology study is usually advised for people in occupations for which an arrhythmia occurrence would place them or others at risk, such as police, military, and pilots, or for individuals who desire evaluation for risk. Routine follow-up without therapy is reasonable in others. Children are at greater risk of sudden death, ~2 per 1000 patient-years.

TREATMENT

TREATMENT

Paroxysmal Supraventricular Tachycardia

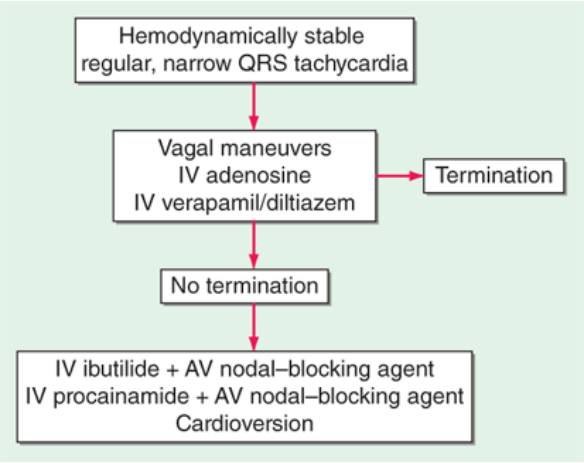
Acute management of narrow QRS PSVT is guided by the clinical presentation. Continuous ECG monitoring should be implemented and a 12-lead ECG should always be obtained when possible, since this may be useful in determining the mechanism. In the presence of hypotension with unconsciousness or respiratory distress, QRS-synchronous direct current cardioversion is warranted, but this is rarely needed, because intravenous adenosine works promptly in most situations (see below). For stable individuals, initial therapy takes advantage of the fact that most PSVTs are dependent on AV nodal conduction (AV nodal reentry or orthodromic AV reentry) and therefore likely to respond to sympatholytic and vagotonic maneuvers and drugs **(Fig. 244-6)**. As these are administered, the ECG should be continuously recorded, because the response can establish the diagnosis. AV block with only transient slowing of tachycardia may expose ongoing p waves, indicating AT or atrial flutter as the mechanism.

Carotid sinus massage is reasonable provided the risk of carotid vascular disease is low, as indicated by absence of carotid bruits and no prior history of stroke. A Valsalva maneuver should be attempted in cooperative individuals, and if effective, the patient can be taught to perform this maneuver as needed. If vagal maneuvers fail or cannot be performed, intravenous adenosine will terminate the vast majority of PSVT episodes by transiently blocking conduction in the AV node. Adenosine may produce transient chest pain, dyspnea, and anxiety. It is contraindicated in patients with prior cardiac transplantation due to potential hypersensitivity. It can theoretically aggravate bronchospasm. Adenosine precipitates AF, which is usually brief, in up to 15% of patients, so it should be used cautiously in patients with WPW syndrome in whom AF may produce hemodynamic instability. Intravenous beta blockers and calcium channel blockers (verapamil or diltiazem) are also effective but may cause hypotension before and after arrhythmia termination and have a longer duration of action. These agents can also be given orally and can be taken by the patient on an as-needed basis to slow ventricular rate and facilitate termination by Valsalva maneuver.

The differential diagnosis of wide-complex tachycardia includes ventricular tachycardia **(Chap. 247)**, PSVT with bundle branch block aberrancy, and preexcited tachycardia (see above). In general, these should be managed as ventricular tachycardia until proven otherwise. If the tachycardia is regular and the patient is stable, a trial of intravenous adenosine is reasonable. Very irregular wide-complex tachycardia is most likely preexcited AF or flutter (see above) and should be managed with cardioversion, intravenous procainamide, or ibutilide. If the diagnosis of PSVT with aberrancy is unequivocal, as may be the case in patients with prior episodes, treatment for PSVT with vagal maneuvers and adenosine is reasonable. In all cases, continuous ECG monitoring should be implemented, and emergency cardioversion and defibrillation should be available.

FIGURE 244-6

Treatment algorithm for patients presenting with hemodynamically stable paroxysmal supraventricular tachycardia. AV, atrioventricular.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Al-Khatib SM et al: Risk stratification for arrhythmic events in patients with asymptomatic pre-excitation: A systematic review for the 2015 ACC/AHA/HRS Guideline for the Management of Adult Patients with Supraventricular Tachycardia: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. Circulation 133:e575, 2016. [PubMed: 26399661]

Appelboam A et al: Postural modification to the standard Valsalva manoeuvre for emergency treatment of supraventricular tachycardias (REVERT): A randomised controlled trial. Lancet 386:1747, 2015. [PubMed: 26314489]

Gupta S, Figueredo VM: Tachycardia mediated cardiomyopathy: Pathophysiology, mechanisms, clinical features and management. Int J Cardiol 172:40, 2014. [PubMed: 24447747]

Link MS: Clinical practice. Evaluation and initial treatment of supraventricular tachycardia. N Engl J Med 367:1438, 2012. [PubMed: 23050527]

Page RL et al: 2015 ACC/AHA/HRS Guideline for the Management of Adult Patients with Supraventricular Tachycardia: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. Circulation 133:e506, 2016. [PubMed: 26399663]

Chapter 245: Common Atrial Flutter, Macroreentrant, and Multifocal Atrial Tachycardias

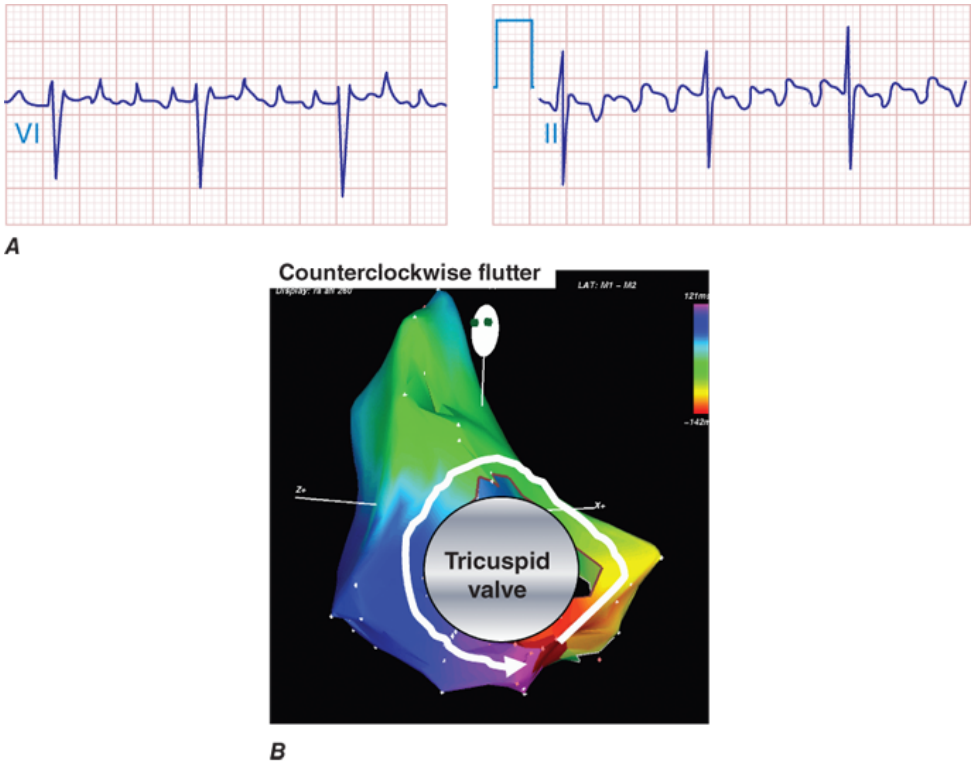
Gregory F. Michaud; William G. Stevenson

INTRODUCTION

Macroreentrant atrial tachycardia is due to a large reentry circuit, often associated with areas of scar in the atria. *Common or typical right atrial flutter* is due to a circuit that revolves around the tricuspid valve annulus, bounded anteriorly by the annulus and posteriorly by functional conduction block in the crista terminalis. The wavefront passes between the inferior vena cava and the tricuspid valve annulus, known as the sub-Eustachian or cavotricuspid isthmus, where it is susceptible to interruption by catheter ablation. Thus, common atrial flutter is also known as *cavotricuspid isthmus-dependent atrial flutter*. This circuit most commonly revolves in a counterclockwise direction (as viewed looking toward the tricuspid annulus from the ventricular apex), which produces the characteristic negative sawtooth flutter waves in leads II, III, and aVF and positive P waves in lead V_1 **(Fig. 245-1)**. When the direction is reversed, clockwise rotation produces the opposite P-wave vector in those leads. The atrial rate is typically 240–300 beats/min but may be slower in the presence of atrial disease or antiarrhythmic drugs. It often conducts to the ventricles with 2:1 AV block, creating a regular tachycardia at 150 beats/min, with p waves that may be difficult to discern. Maneuvers that increase AV nodal block will typically expose flutter waves, allowing diagnosis.

FIGURE 245-1

A. Common right atrial flutter, also known as cavotricuspid isthmus flutter, showing positive P waves in lead V_1 and negative "sawtooth" pattern in lead II typical of counterclockwise rotation relative to the tricuspid valve annulus. *(Adapted from F Marchlinski: The tachyarrhythmias. in DL Longo et al [eds]: Harrison's Principles of Internal Medicine, 18th ed. New York, McGraw-Hill, 2012, pp 1878–1900.)* ***B.*** A right atrial map of common counterclockwise flutter is shown. Colors indicate activation time, progressing from red to yellow to green, blue, and purple. The reentry path parallels the tricuspid annulus.



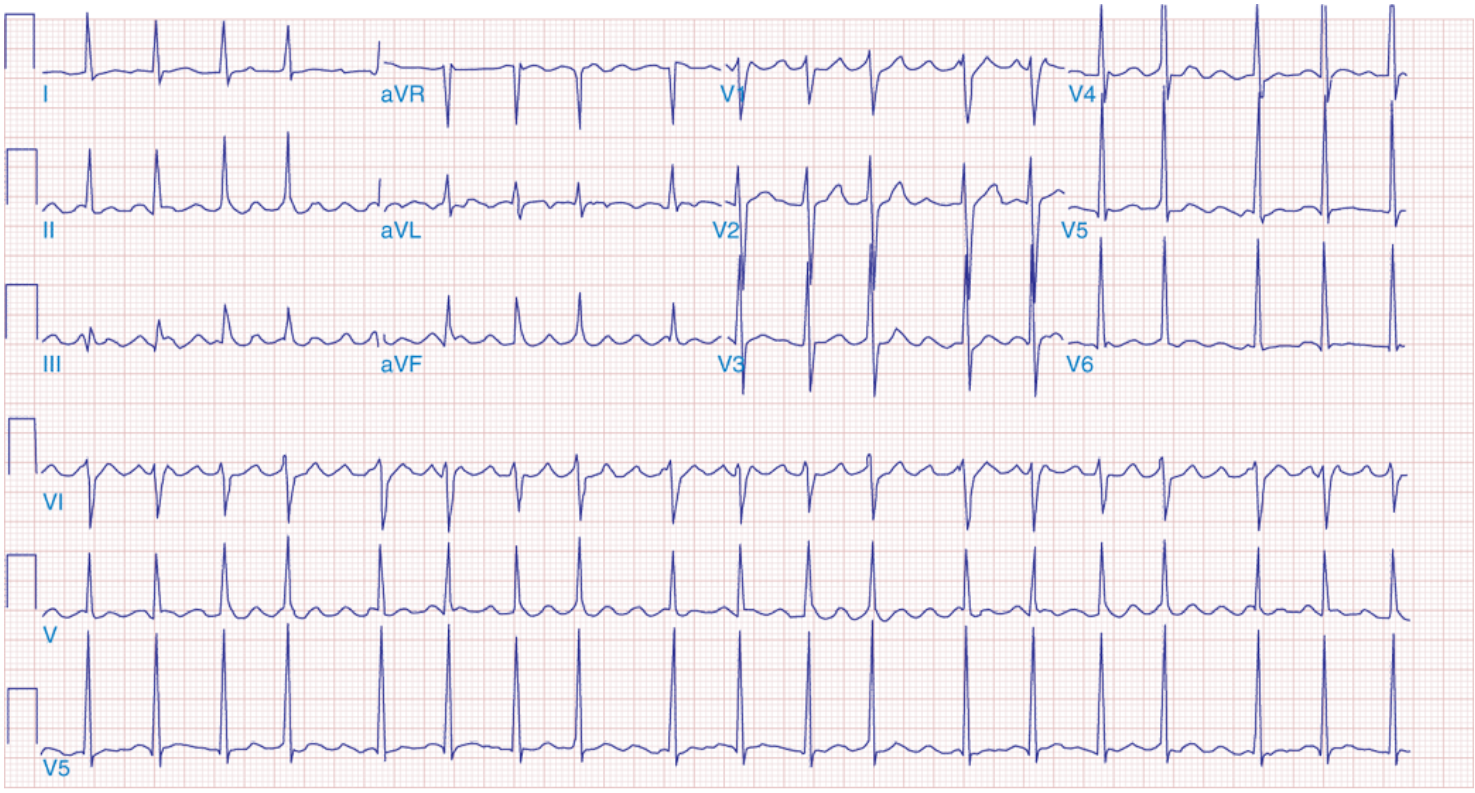
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Common right atrial flutter usually occurs in association with atrial fibrillation and often with atrial scar from senescence or prior cardiac surgery. Some patients with atrial fibrillation treated with an antiarrhythmic drug, particularly flecainide, propafenone, or amiodarone, will present with atrial flutter rather than fibrillation, since these agents slow atrial conduction velocity and can promote reentry.

Macroreentrant ATs that are not dependent on conduction through the cavotricuspid isthmus are referred to as *atypical atrial flutters*. They can occur in either atrium and are almost universally associated with areas of atrial scar. Left atrial flutter and perimitral left atrial flutter are commonly seen after extensive left atrial ablation for atrial fibrillation or atrial surgery. The clinical presentation is similar to common atrial flutter, but with different P-wave morphologies (Fig 245-2). They can be difficult to distinguish from focal AT, and in most cases, the mechanism can only be confirmed by an electrophysiology study.

FIGURE 245-2

Atrial flutter in a 52-year-old man that occurred one year after extensive left atrial ablation for persistent atrial fibrillation. In contrast to common flutter the P waves in V1 and inferior limb leads (II, III, and aVF) have the same polarity (positive in this case). Also, lead aVL shows a predominant negative P wave consistent with a left atrial focus, however P-wave morphology used to diagnose arrhythmia mechanism and location is unreliable in the setting of advanced atrial fibrosis, such as after extensive catheter ablation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Atrial Flutter

Initial management of atrial flutter is similar to that for atrial fibrillation, discussed in more detail below. Electrical cardioversion is warranted for hemodynamic instability or severe symptoms. Otherwise, rate control can be achieved with administration of AV nodal–blocking agents, but this is often more difficult than for atrial fibrillation. The risk of thromboembolic events is felt to be similar to that associated with atrial fibrillation. Anticoagulation is warranted prior to conversion for episodes more than 48 h in duration and chronically for patients at increased risk of thromboembolic stroke based on the CHA_2DS_2-VASc scoring system (Table 245-1).

For a first episode of atrial flutter, conversion to sinus rhythm with no antiarrhythmic drug therapy is reasonable. For recurrent episodes, antiarrhythmic drug therapy with sotalol, dofetilide, disopyramide, and amiodarone may be considered, but >70% of patients experience recurrences. For recurrent episodes of common atrial flutter, catheter ablation of the cavotricuspid isthmus abolishes the arrhythmia in >90% of patients with a low risk of complications that are largely related to vascular access, and rarely heart block. Approximately 50% of patients presenting with atrial flutter develop atrial fibrillation within 5 years after diagnosis, which is an important consideration in patients with a high risk profile for thromboembolism.

TABLE 245-1

CHA_2DS_2-VASc Risk Assessment and Oral Anticoagulants

Risk Factors	Points		CHA_2DS_2-VASc Score	Estimated Annual Stroke Rate[a]
C—congestive heart failure	1		0	0
H—hypertension	1		1	1.3%
A—age ≥75 y	2		2	2.2%
D—diabetes mellitus	1		3	3.2%
S—stroke or TIA, embolus	2		4	4.0%
V—vascular disease	1		5	6.7%
A—age 65–75 y	1		6–9	>9%
Sex—female	1			
Anticoagulants	**Mechanism**	**Excretion**	**Dosing Considerations**	**Risk/Benefit**
Warfarin	Vitamin K antagonist	Liver	Adjusted to INR 2–3 Days to therapeutic effect Multiple drug/food interactions (e.g., amiodarone)	Major hemorrhage: 1% per year Intracranial hemorrhage: 0.1–0.6% per year Risk of bleeding increases with INR >3.5 Inexpensive
Dabigatran[b]	Thrombin inhibitor	Kidney		
		CCr >30 mL/min CCr 15–30 mL/min	150 mg bid 75 mg bid P-glycoprotein substrate (inducers— rifampin, reduce concentration) (inhibitors—amiodarone, verapamil, dronedarone, quinidine) Proton pump inhibitors may reduce absorption	Onset of action within hours No reversal agent for bleeding
Rivaroxaban	Xa inhibitor	Kidney	P-glycoprotein substrate	No reversal agent for bleeding
		CCr ≥50 mL/min	20 mg daily	
		CCr 15–50 mL/min	15 mg daily	
Apixaban	Xa inhibitor	Kidney and liver	P-glycoprotein substrate	No reversal agent for bleeding
		Any 2 of: Cr >1.5 mg/dL, age >80 yrs, or wt <60 kg	5 mg bid 2.5 mg bid	

[a]Modified from GY Lip et al: Lancet 379:648, 2012. [b]U.S. Food and Drug Administration recommended dosing; other regimens are available outside the United States.

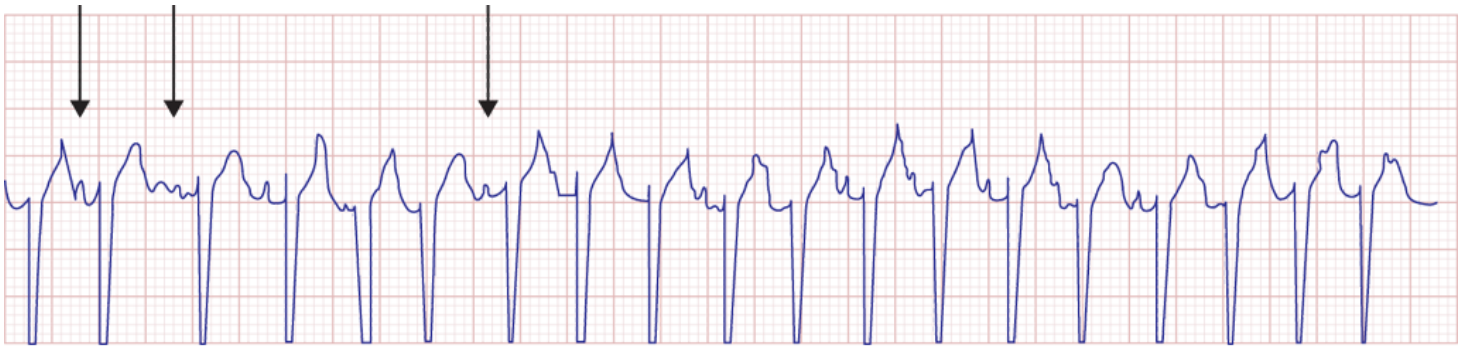
Abbreviations: CCr, creatinine clearance; Cr, creatinine; INR, international normalized ratio; TIA, transient ischemic attack; wt, weight.

MULTIFOCAL ATRIAL TACHYCARDIA

Multifocal AT (MAT) is characterized by a rhythm with at least three distinct P-wave morphologies with rates typically between 100 and 150 beats/min. Unlike atrial fibrillation, there are clear isoelectric intervals between P waves **(Fig. 245-3)** and the atrial rate is slower. The mechanism is likely triggered automaticity from multiple atrial foci. It is usually encountered in patients with chronic pulmonary disease and acute illness.

FIGURE 245-3

Multifocal atrial tachycardia. Rhythm strip obtained from a patient with severe pulmonary disease during an acute illness. *Arrows* note three distinct P-wave morphologies.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Therapy for MAT is directed at treating the underlying disease and correcting any metabolic abnormalities. Electrical cardioversion is ineffective. The calcium channel blockers verapamil or diltiazem may slow the atrial and ventricular rate. Patients with severe pulmonary disease often do not tolerate beta blocker therapy. MAT may respond to amiodarone, but long-term therapy with this agent is usually avoided due to its toxicities, particularly pulmonary fibrosis.

FURTHER READING

Page RL et al: 2015 ACC/AHA/HRS Guideline for the Management of Adult Patients with Supraventricular Tachycardia: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. Circulation 133:e506, 2016. [PubMed: 26399663]

Rahman F et al: Atrial flutter: Clinical risk factors and adverse outcomes in the Framingham Heart Study. Heart Rhythm 13:233, 2016. [PubMed: 26226213]

Chapter 246: Atrial Fibrillation

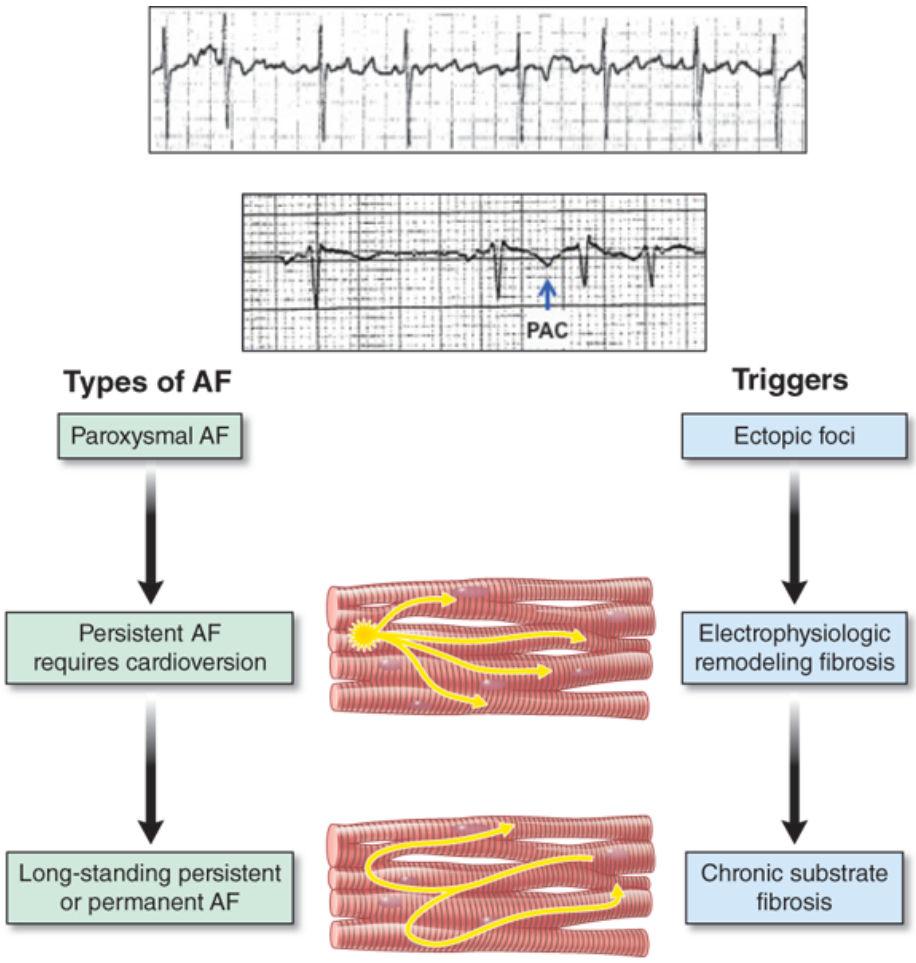
Gregory F. Michaud; William G. Stevenson

INTRODUCTION

Atrial fibrillation (AF) is characterized by disorganized, rapid, and irregular atrial activation with loss of atrial contraction and with an irregular ventricular rate that is determined by AV nodal conduction (Fig. 246-1). In an untreated patient, the ventricular rate also tends to be rapid and variable, between 120 and 160 beats/min, but in some patients, it may exceed 200 beats/min. Patients with high vagal tone or AV nodal conduction disease may have slow ventricular rates.

FIGURE 246-1

A rhythm strip of atrial fibrillation (AF) showing absence of distinct P-waves and an irregularly irregular ventricular response. Diagram depicts atrial fibrillation types. Paroxysmal AF is initiated by premature beats, as shown in the rhythm strip (*arrow*) after two sinus beats. Triggering foci are often an important cause of this arrhythmia. Persistent AF is associated with atrial structural and electrophysiologic remodeling, as well as with triggering foci in many patients. Long-standing persistent AF is associated with greater structural remodeling with atrial fibrosis and electrophysiologic remodeling.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

AF is the most common sustained arrhythmia and is a major public health problem. Prevalence increases with age, and >95% of AF patients are >60 years of age. The prevalence by age 80 is ~10%. The lifetime risk of developing AF for men 40 years old is ~25%. AF is

slightly more common in men than women and more common in whites than blacks. Risk factors for developing AF in addition to age and underlying cardiac disease include hypertension, diabetes mellitus, cardiac disease, obesity, and sleep apnea. AF is associated with a 1.5- to 1.9-fold increased risk of mortality after controlling for underlying heart disease. AF is also associated with a risk of developing heart failure and vice-versa—patients with heart failure have an increased risk of developing AF. AF increases the risk of stroke by fivefold and is estimated to be the cause of 25% of strokes. It also increases the risk of dementia and silent strokes detected by MRI. Since AF is a marker for other predictors of mortality and morbidity, such as the severity of heart disease, it is difficult to determine the extent to which AF itself contributes to associated increased mortality and morbidity.

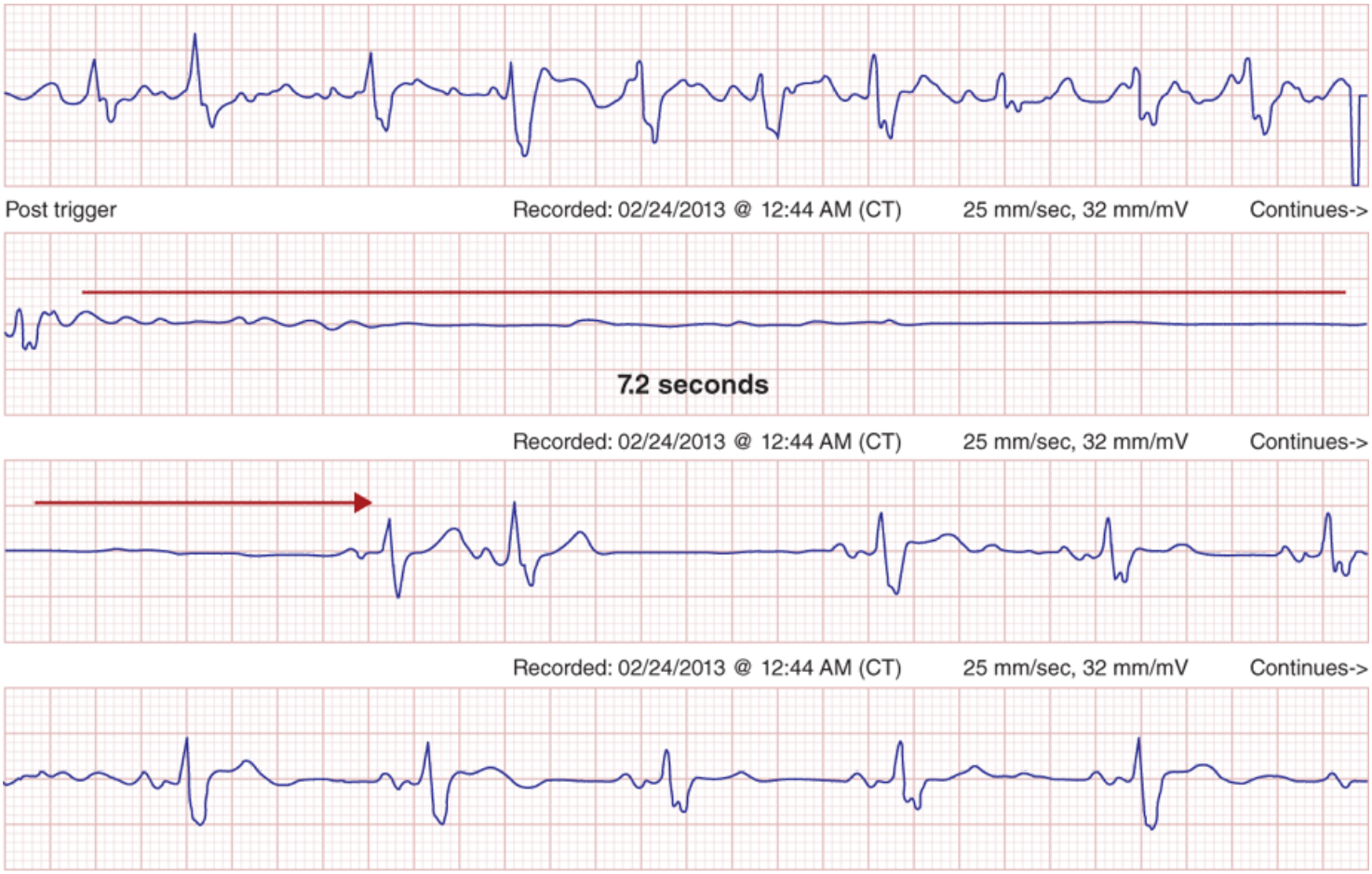
AF is occasionally associated with an acute precipitating factor such as hyperthyroidism, acute alcohol intoxication, or an acute illness such as myocardial infarction or pulmonary embolism. AF occurs in up to 30% of patients recovering from cardiac surgery, associated with inflammatory pericarditis.

The clinical pattern of AF suggests the underlying pathophysiology (Fig. 246-1). Paroxysmal AF is defined by episodes that start spontaneously and stop within 7 days of onset. Paroxysmal AF is often initiated by small reentrant or rapidly firing foci in sleeves of atrial muscle that extend into the pulmonary veins (PV). Catheter ablation that isolates these foci usually abolishes paroxysmal AF, although some patients also have initiating foci in other locations. These non-PV triggers tend to occur in older patients and those with more severe underlying cardiac disease. Persistent AF has a longer duration, exceeding 7 days, and, in many cases, will continue indefinitely unless cardioversion is performed. Cardioversion can be followed by prolonged periods of sinus rhythm. As for paroxysmal AF, episodes are often initiated by rapidly firing foci within PVs, but non-PV sites, including myocardial sleeves around the superior vena cava (SVC) or coronary sinus are encountered more often than when AF is paroxysmal. In addition, persistence of the AF is likely facilitated by structural and electrophysiologic atrial abnormalities, particularly fibrosis that uncouples atrial fibers, promoting reentry and focal automaticity. In patients with long-standing persistent AF (>1 year), significant fibrosis is usually present and it is difficult to restore and maintain sinus rhythm. Some patients progress over years from paroxysmal to persistent AF. Although fibrosis that develops with aging and atrial hypertrophy in response to hypertension and other cardiac disease appears to be an important promoting factor, electrophysiologic remodeling that affects conduction and refractoriness occur as well in response to chronic tachycardia. Thus, AF tends to promote AF.

Clinical consequences of AF are related to rapid ventricular rates, loss of atrial contribution to ventricular filling, and predisposition to thrombus formation in the left atrial appendage with potential embolization. Presentations vary with the ventricular rate and underlying heart disease and comorbidities. Rapid rates may cause hemodynamic collapse or heart failure exacerbation particularly in patients with impaired cardiac function, hypertrophic cardiomyopathy, and heart failure with preserved systolic function. Exercise intolerance and easy fatigability are common despite the absence of palpitations in many patients. Occasionally, dizziness or syncope occurs due to pauses when AF terminates to sinus rhythm **(Fig. 246-2)**. Depressed ventricular function with cardiomyopathy may develop in response to chronic tachycardia (rates persistently faster than 100–110 bpm) and is probably more common in patients who do not sense palpitations, since they may not seek medical care until heart failure symptoms develop. Tachycardia-related cardiomyopathy is usually reversible with control of ventricular rate.

FIGURE 246-2

A continuous rhythm strip is shown from a patient with tachy-brady syndrome. Atrial fibrillation is present at the top and abruptly terminates in the second tracing, with atrial and ventricular standstill for 7.2 s until resumption of sinus rhythm. The patient experienced syncope.



Treatment for AF is primarily guided by patients' symptoms, the hemodynamic effect of AF, the duration of AF, the risk of stroke and the underlying heart disease. *New-onset AF* that produces severe hypotension, pulmonary edema, or angina should be electrically cardioverted starting with a QRS synchronous shock of 200 J, ideally after sedation or anesthesia is achieved. Greater shock energy and different electrode placements may be tried if the shock fails to terminate AF. Administration of intravenous ibutilide lowers the energy requirement for atrial defibrillation and may be useful if AF terminates and reinitiates, but should not be used in patients with a prolonged QT interval or severe LV dysfunction because of a significant risk of torsades de Pointes. If the patient is stable, immediate management involves rate control to alleviate or prevent symptoms, and consideration of whether anticoagulation is warranted to reduce stroke risk. Consideration is then given to whether therapy is warranted to restore and maintain sinus rhythm, or whether the patient will be allowed to continue in AF, and managed with rate control and measures for stroke prevention. It is critical to consider the risk of stroke when attempting to restore sinus rhythm. If the duration of AF is unclear or is known to be >48 h, anticoagulation must be commenced before cardioversion. Anticoagulation strategies for new-onset AF are debated. In the absence of contraindications, it is usually appropriate to initiate systemic anticoagulation with heparin immediately or with an oral anticoagulant that has rapid onset of action, while evaluation and other therapies are implemented.

CARDIOVERSION AND ANTICOAGULATION

The major source of thromboembolism and stroke in AF is formation of thrombus in the left atrial appendage where flow is relatively stagnant, although thrombus occasionally forms in other locations as well. Following conversion from prolonged AF to sinus rhythm, atrial mechanical function can be delayed for weeks, such that thrombi can form even during sinus rhythm. When AF has been present for >48 h and in patients at high risk for thromboembolism, such as those with mitral stenosis or hypertrophic cardiomyopathy, conversion to sinus rhythm is associated with an increased risk of thromboembolism. Thromboembolism can occur soon, or several days after restoration of sinus rhythm if appropriate anticoagulation measures are not taken.

Cardioversion *within 48 h of the onset of AF* is common practice in patients who have not been anticoagulated, provided that they are not at high risk for stroke due to a prior history of embolic events, rheumatic mitral stenosis, or hypertrophic cardiomyopathy with marked left atrial enlargement. These low-risk patients with occasional episodes of AF can be instructed to notify their physician when AF occurs to arrange for cardioversion to be done within 48 h.

If the duration of AF exceeds 48 h or is unknown, there is greater concern for thromboembolism after cardioversion, even in patients considered low risk (CHA_2DS_2-VASc of 0 or 1 [see below]) for stroke. There are two approaches to mitigate the risk related to cardioversion. One option is to anticoagulate continuously for 3 weeks before and a minimum of 4 weeks after cardioversion. A second approach is to start anticoagulation and perform a transesophageal echocardiogram to determine if thrombus is present in the left atrial appendage. If thrombus is absent, cardioversion can be performed and anticoagulation continued for a minimum of 4 weeks to allow time for recovery of atrial mechanical function. In either case, cardioversion of AF is associated with a substantial risk of recurrence, which may not be symptomatic. Longer-term maintenance of anticoagulation is considered based on the patient's individual risk for stroke, commonly assessed from the CHA_2DS_2-VASc score.

RATE CONTROL

Acute rate control can be achieved with beta blockers and/or the calcium channel blockers verapamil and diltiazem administered either intravenously or orally, as warranted by the urgency of the clinical situation. Digoxin may be added, particularly in heart failure patients, if negative inotropic and other adverse effects of beta blockers and calcium channel blockers limit their use. Digoxin lacks negative inotropic effects, but is less effective in slowing the ventricular rate in AF, particularly when sympathetic tone is high. It is synergistic with the other AV nodal–blocking agents. Its use has been associated with increased mortality in some studies. Typically, the goal of acute rate control is to reduce the ventricular rate to less than 100/min, but the goal must be guided by the clinical situation and the adverse effects of rate control medications.

CHRONIC RATE CONTROL

For patients who remain in AF chronically, the goal of rate control is to alleviate and prevent deterioration of ventricular function from excessive rates. β-Adrenergic blockers and calcium channel blockers are often used in combination. Digoxin is added selectively when these are not sufficient. Exertion-related symptoms are often an indication of inadequate rate control. Rate should be assessed with exertion and medications adjusted accordingly. The initial goal is a resting heart rate of <80 beats/min that increases to <100 beats/min with light exertion, such as walking. If it is difficult to slow the ventricular rate to that degree, allowing a resting rate of up to 110 beats/min is acceptable provided it does not cause symptoms and ventricular functioniss normal, but periodic assessment of ventricular function is warranted because some patients develop tachycardia-induced cardiomyopathy.

If adequate rate control in AF is difficult to achieve, further consideration should be given to restoring sinus rhythm (see below). Catheter ablation of the AV junction to create heart block and implantation of a permanent pacemaker reliably achieves rate control without the need for AV nodal blocking agents, but mandates lifelong permanent pacing. Right ventricular apical pacing induces dyssynchronous ventricular activation that can depress ventricular function in some patients. Biventricular pacing or direct pacing of the His bundle may be used to minimize the degree of ventricular dyssynchrony.

STROKE PREVENTION IN ATRIAL FIBRILLATION

The majority of patients warrant chronic anticoagulation, but selection of therapy should be individualized based on patient profile and risks and benefits of individual agents. Anticoagulation is warranted for patients with mitral stenosis, hypertrophic cardiomyopathy, and those with a prior history of stroke. Patients without mitral stenosis are often referred to as having nonvalvular AF. The CHA_2DS_2-VASc score (Table 246-1) can be used to estimate stroke risk in these patients. Anticoagulation is recommended for a score of ≥2 and may be considered for a score of 1. The approach to patients with paroxysmal AF is the same as for persistent AF. It is recognized that many patients who appear to have infrequent AF episodes based on office visits often have asymptomatic episodes that put them at risk. Absence of AF during periodic monitoring is not sufficient to indicate low risk. The role of continuous monitoring with implanted recorders or pacemakers is not yet clear as a guide for anticoagulation in patients with a borderline risk profile.

TABLE 246-1

CHA_2DS_2-VASc Risk Assessment and Oral Anticoagulants

Risk Factors	Points		CHA_2DS_2-VASc Score	Estimated Annual Stroke Rate[a]
C—congestive heart failure	1		0	0
H—hypertension	1		1	1.3%
A—age ≥75 y	2		2	2.2%
D—diabetes mellitus	1		3	3.2%
S—stroke or TIA, embolus	2		4	4.0%
V—vascular disease	1		5	6.7%
A—age 65–75 y	1		6–9	>9%
Sex—female	1			
Anticoagulants	**Mechanism**	**Excretion**	**Dosing Considerations**	**Risk/Benefit**
Warfarin	Vitamin K antagonist	Liver	Adjusted to INR 2–3 Days to therapeutic effect Multiple drug/food interactions (e.g., amiodarone)	Major hemorrhage: 1% per year Intracranial hemorrhage: 0.1–0.6% per year Risk of bleeding increases with INR >3.5 Inexpensive
Dabigatran[b]	Thrombin inhibitor	Kidney		
		CCr >30 mL/min CCr 15–30 mL/min	150 mg bid 75 mg bid P-glycoprotein substrate (inducers—rifampin, reduce concentration) (inhibitors—amiodarone, verapamil, dronedarone, quinidine), Proton pump inhibitors may reduce absorption	Onset of action within hours Reversal agent available
Rivaroxaban	Xa inhibitor	Kidney	P-glycoprotein substrate	Reversal agent for bleeding in development

Risk Factors	Points		CHA_2DS_2-VASc Score	Estimated Annual Stroke Rate[a]
		CCr ≥50 mL/min	20 mg daily	
		CCr 15–50 mL/min	15 mg daily	
Apixaban	Xa inhibitor	Kidney and liver Any 2 of Cr>1.5 mg/dl, age > 80 yrs or wt < 60 kg.	P-glycoprotein substrate 5 mg bid 2.5 mg bid	Reversal agent for bleeding in development
Edoxaban	Xa inhibitor	Kidney and Liver CCr> 60 <95 mL/min CCr 15–60 mL/min	P-glycoprotein substrate 60 mg 30 mg	

[a]Modified from GY Lip et al: Lancet 379:648, 2012. [b]U.S. Food and Drug Administration recommended dosing; other regimens are available outside the United States.

Abbreviations: CCr, creatinine clearance; Cr, creatinine; INR, international normalized ratio; TIA, transient ischemic attack.

The major options for anticoagulation are the antithrombin inhibitor dabigitran, factor Xa inhibitors rivaroxaban, apixaban, and edoxaban, the vitamin K antagonist warfarin. Antiplatelet agents alone are generally not sufficient. In non-valvular AF, warfarin reduces the annual risk of stroke by 64% compared to placebo and by 37% compared to antiplatelet therapy. Patients with AF with an increased risk of stroke also have an increased risk of venous thromboembolism, which appears to be lower with oral anticoagulation. The direct acting anticoagulants, dabigatran, rivaroxaban, apixaban, and edoxaban were noninferior to warfarin in individual trials, and analysis of pooled data suggests superiority to warfarin by small absolute margins of 0.4–0.7% in reduction of mortality, stroke, major bleeding, and intracranial hemorrhage. Warfarin is the agent required for patients with rheumatic mitral stenosis or mechanical heart valves. The newer, direct acting anticoagulants have not been tested in rheumatic heart disease and a direct thrombin inhibitor did not prevent thromboemboli in patients with mechanical heart valves. Warfarin is an inconvenient agent that requires several days to achieve a therapeutic effect (prothrombin time [PT]/international normalized ratio [INR] >2), requires monitoring of PT/INR to adjust dose, and has many drug and food interactions, that can hinder patient compliance. The direct acting agents are easier to use and achieve reliable anticoagulation promptly without requiring dosage adjustment based on blood tests. Dabigatran, rivaroxaban, and apixaban have renal excretion, cannot be used with severe renal insufficiency (CrCl <15 mL/min), and require dose adjustment for modest renal impairment, which is of particular concern in the elderly, who are at increased bleeding risk. Excretion can also be influenced by P-glycoprotein inducers and inhibitors. Warfarin anticoagulation can be reversed by administration of fresh frozen plasma and vitamin K. A reversal agent (idarucizumab) is available for dabigatran and reversal agents for the Xa inhibitors are being evaluated (andexanet alfa and ciraparantag). For now, bleeding that occurs during Xa inhibitor use must be managed with supportive care, with the expectation that clotting will improve over 12 h as the anticoagulant is excreted, although reversal agents are anticipated.

The antiplatelet agents aspirin and clopidogrel are inferior to warfarin for stroke prevention in AF and do not have less risk of bleeding. Clopidogrel combined with aspirin is better than aspirin alone for stroke prevention, but is inferior to warfarin and has a greater bleeding risk than aspirin alone.

Bleeding is the major risk of anticoagulation. Major bleeding requiring transfusion or intracranial bleeding occurs in ~1% of patients per year with warfarin. Direct acting anticoagulants appear to have a lower risk of intracranial bleeding relative to warfarin without sacrificing protective effects against thromboembolism. Risk factors for bleeding include age >65–75 years, heart failure, renal insufficiency, prior bleeding, and excessive alcohol or nonsteroidal anti-inflammatory drug use. In patients who require dual antiplatelet therapy (e.g., aspirin and clopidogril) after coronary or peripheral arterial stenting, there is a substantially increased bleeding risk when standard oral anticoagulation with warfarin or a direct acting anticoagulant is added. The optimal combination of agents for patients with AF who also require antiplatelet therapy is unclear.

Chronic anticoagulation is contraindicated in some patients due to bleeding risks. Because most atrial thrombi likely originate in the left atrial appendage, surgical removal of the appendage, combined with atrial maze surgery, may be considered for patients undergoing surgery, although removal of the appendage has not been unequivocally shown to reduce the risk of thromboembolism. Percutaneously deployed devices that occlude or ligate the left atrial appendage are also available, appear to be non-inferior to warfarin in reducing stroke risk and are considered in patients who have a high risk of thromboembolism, but serious bleeding risk from chronic oral anticoagulation.

RHYTHM CONTROL

AF is associated with obesity, hypertension, excessive alcohol use and sleep apnea. Aggressive treatment of these risk factors can substantially reduce AF episodes in some patients and is warranted in all patients.

The decision to administer antiarrhythmic drugs or perform catheter ablation to attempt maintenance of sinus rhythm (commonly referred to as the “rhythm control strategy”) is mainly guided by patient symptoms and preferences regarding the benefits and risks of therapies. In general, patients who maintain sinus rhythm have better survival than those who continue to have AF. This may be because continued AF is a marker of disease severity. In randomized trials, administration of antiarrhythmic medications to maintain sinus rhythm did not improve survival or symptoms compared to a rate control strategy, and the drug therapy group had more hospitalizations. Disappointing efficacy and toxicities of available antiarrhythmic drugs and patient selection bias may be factors that influenced the results of these trials. The impact of catheter ablation on mortality is not known but the subject of ongoing randomized, controlled trials. A rhythm control strategy is usually selected for patients with symptomatic paroxysmal AF, recurrent episodes of symptomatic persistent AF, AF with difficult rate control, and AF that has resulted in depressed ventricular function or that aggravates heart failure. A rhythm control strategy is more likely to be favored in younger patients than in sedentary or elderly patients in whom rate control is more easily achieved. Even if sinus rhythm is apparently maintained, anticoagulation is recommended according to the CHA_2DS_2-VASc stroke risk profile because asymptomatic episodes of AF are common. Following a first episode of persistent AF, a strategy using AV nodal–blocking agents, cardioversion, and anticoagulation is reasonable, in addition to addressing possible aggravating factors. If recurrences are infrequent, periodic cardioversion is reasonable.

Pharmacologic Therapy for Maintaining Sinus Rhythm

The goal of pharmacologic therapy is to maintain sinus rhythm or reduce episodes of AF. Risks and side effects of antiarrhythmic drugs are a major consideration in selecting therapy. Drug therapy can be instituted once sinus rhythm has been established or in anticipation of cardioversion. β-Adrenergic blockers and calcium channel blockers help control ventricular rate, improve symptoms, and possess a low-risk profile, but have low efficacy for preventing AF episodes. Class I sodium channel–blocking agents (e.g., flecainide, propafenone, disopyramide) are options for subjects without significant structural heart disease, but negative inotropic and proarrhythmic effects warrant avoidance in patients with coronary artery disease or heart failure. The class III agents sotalol and dofetilide can be administered to patients with coronary artery disease or structural heart disease but have ~3% risk of inducing excessive QT prolongation and torsades des pointes. Dofetilide should be initiated only in a hospital with ECG monitoring, and many physicians take this approach with sotalol as well. Dronedarone increases mortality in patients with heart failure. All of these agents have modest efficacy in patients with paroxysmal AF, of whom ~30–50% will benefit. Amiodarone is more effective, maintaining sinus rhythm in approximately two-thirds of patients. It can be administered to patients with heart failure and coronary artery disease. Over 40% of patients experience amiodarone-related toxicities during long-term therapy.

CATHETER AND SURGICAL ABLATION FOR ATRIAL FIBRILLATION

Successful catheter ablation avoids antiarrhythmic drug toxicities but procedural risks and efficacy depend on operator experience. For patients with previously untreated but recurrent paroxysmal AF, catheter ablation has mildly better efficacy compared to antiarrhythmic drug therapy and is clearly superior to antiarrhythmic drugs for patients who have recurrent AF despite drug treatment. Long-term control of AF is more difficult to achieve in patients with persistent AF, likely because of more extensive atrial abnormality and associated greater co-morbidities in these patients.

Catheter ablation involves cardiac catheterization, trans(atrial) septal puncture, and radiofrequency ablation or cryoablation to electrically isolate the left atrial regions around the PV, abolishing the ability of triggering foci in these regions to initiate AF, and also likely

impacting the substrate for reentry in the left atrium. Extensive areas of ablation are required, and gaps in healed ablation areas or emergence of new trigger sites outside the PV necessitate a repeat procedure in 20–50% of patients.

In patients with paroxysmal AF, sinus rhythm is maintained for >1 year after one ablation procedure in ~60% of patients; and is achieved in 70–80% of patients after multiple procedures. Many patients become more responsive to antiarrhythmic drugs after a PV isolation procedure. Ablation is less effective in patients with persistent AF, and particularly long-standing persistent AF, particularly when associated with more extensive cardiac disease and co-morbidities. More extensive ablation is often required, targeting areas that likely support reentry in regions outside but adjacent to the pulmonary venous antra. There is no proven strategy for selecting ablation targets outside the PV antral regions and a variety of approaches have been pursued. Ablation of areas of rapid activity during AF or creation of ablation lines to block conduction across regions of the atria did not improve outcomes in some studies. Other ablation targets include foci that fire in response to isoproterenol, areas of atrial fibrosis, and regions with activation consistent with reentrant rotors during AF. More than one ablation procedure is often required to maintain sinus rhythm in patients with persistent and long-standing persistent AF.

Catheter ablation has a 2–7% risk of major procedure-related complications, including stroke (0.5–1%), cardiac tamponade (1%), phrenic nerve paralysis, bleeding from femoral access sites, and fluid overload with heart failure, that can emerge 1–3 days after the procedure. It is important to recognize the potential for delayed presentation of some complications. Ablation within the PV can lead to PV stenosis, presenting weeks to months after the procedure with dyspnea or hemoptysis. The esophagus abuts the posterior wall of the left atrium where it is subject to injury, and esophageal ulcers can form immediately after the procedure and may rarely lead to a fistula between the left atrium and esophagus (estimated incidence of < 0.1%) that presents as endocarditis and stroke 10 days to 3 weeks after the procedure. Early diagnosis of atrioesophageal fistula is important because delayed diagnosis leads to death. Diagnosis is made by chest CT scan with water soluble oral and IV contrast. Endoscopy should be avoided in patients with a suspected fistula because of the risk of air/esophageal fluid embolus. Definitive repair of the atrioesophageal fistula by emergent cardiothoracic surgery is required.

Surgical ablation of AF is most frequently performed concomitant with cardiac valve or coronary artery surgery and less commonly as a stand-alone procedure. However, for patients with persistent AF, surgical or hybrid procedures appear to have higher single-procedure efficacy than catheter ablation. Risks include sinus node injury requiring pacemaker implantation. Surgical removal of the left atrial appendage may reduce stroke risk, although thrombus can form in the remnant of the appendage or if the appendage is not completely ligated.

FURTHER READING

Calkins H et al: 2012 HRS/EHRA/ECAS expert consensus statement on catheter and surgical ablation of atrial fibrillation: Recommendations for patient selection, procedural techniques, patient management and follow-up, definitions, endpoints, and research trial design. Heart Rhythm 9:632.e21, 2012.

Cox JL: Effective practical management of patients with atrial fibrillation when using new oral anticoagulants. Ann Med 47:278, 2015. [PubMed: 25850527]

Douketis JD et al: Perioperative bridging anticoagulation in patients with atrial fibrillation. N Engl J Med 373:823, 2015. [PubMed: 26095867]

Freedman B et al: Stroke prevention in atrial fibrillation. Lancet 388:806, 2016. [PubMed: 27560276]

January CT et al: 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: Executive summary: A report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. Circulation 130:2071, 2014. [PubMed: 24682348]

Kirchhof P et al: 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Eur Heart J 37:2893, 2016. [PubMed: 27567408]

Masoudi FA et al: 2015 ACC/HRS/SCAI left atrial appendage occlusion device societal overview. Heart Rhythm 12:e122, 2015. [PubMed: 26134035]

Nair KK et al: Atrioesophageal fistula: A review. J Atr Fibrillation 8:1331, 2015. [PubMed: 27957213]

Nyong J et al: Efficacy and safety of ablation for people with non-paroxysmal atrial fibrillation. Cochrane Database Syst Rev 11:CD012088, 2016. [PubMed: 27871122]

Voskoboinik A et al: Alcohol and atrial fibrillation: A sobering review. J Am Coll Cardiol 68:2567, 2016. [PubMed: 27931615]

Chapter 247: Approach to Ventricular Arrhythmias

Roy M. John; William G. Stevenson

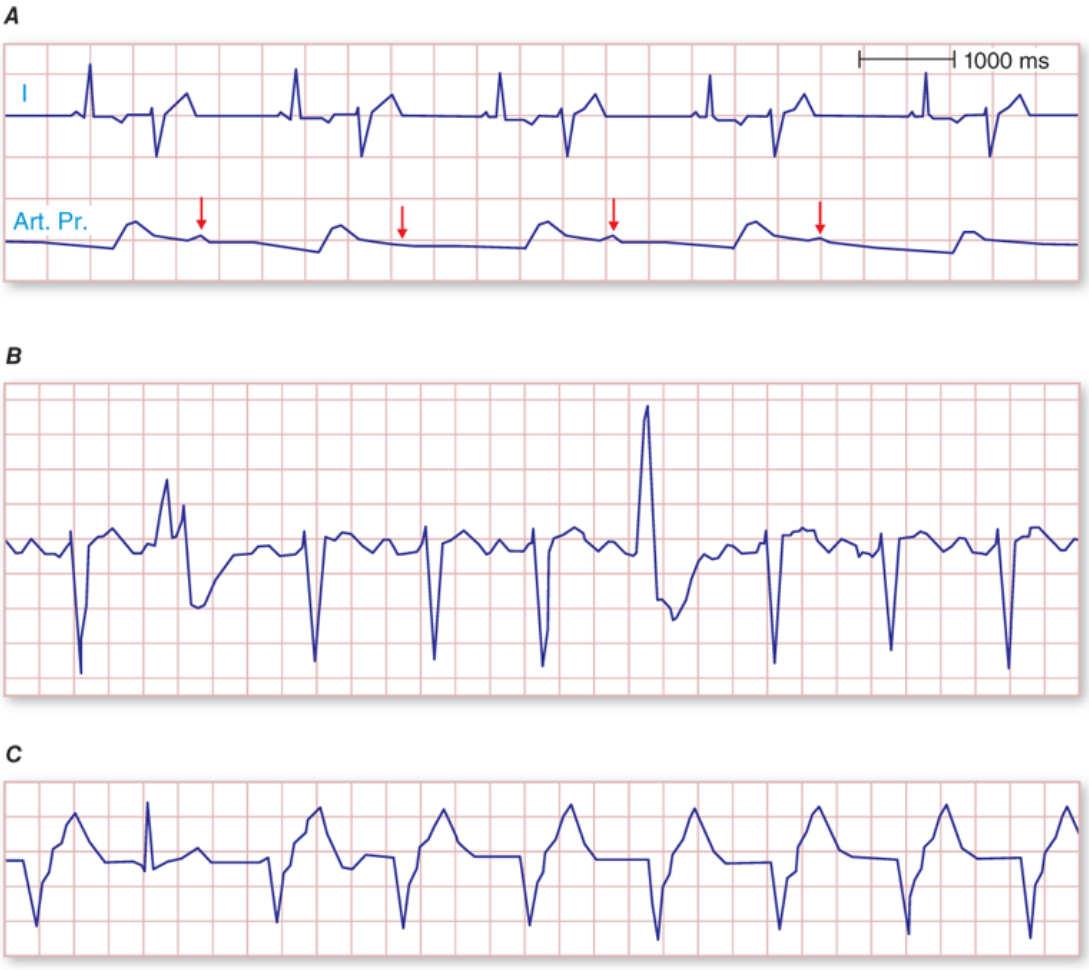
TYPES OF VENTRICULAR ARRHYTHMIAS

Ventricular arrhythmias originate from a focus of myocardial or Purkinje cells capable of automaticity, or triggered automaticity, or from reentry through areas of scar or a diseased Purkinje system. They are characterized by their electrocardiographic appearance and duration. Conduction away from the ventricular focus through the ventricular myocardium is slower than activation of the ventricles over the Purkinje system. Hence, the QRS complex during ventricular arrhythmias will be wide, typically >0.12 s.

Premature ventricular beats (also referred to as a premature ventricular *contraction or PVC*) are single ventricular beats that fall earlier than the next anticipated supraventricular beat **(Fig. 247-1)**. PVCs that originate from the same focus will have the same QRS morphology and are referred to as unifocal **(Fig. 247-1A)**. PVCs that originate from different ventricular sites have different QRS morphologies and are referred to as multifocal **(Fig. 247-1B)**. Two consecutive ventricular beats are ventricular couplets.

FIGURE 247-1

A. Unifocal premature ventricular contractions (PVCs) at bigeminal frequency. Trace shows ECG lead 1 and arterial pressure (Art. Pr.). Sinus rhythm beats are followed by normal arterial waveform. The arterial pressure following premature beats is attenuated (arrows) and imperceptible to palpation. The pulse in this patient is registered at half the heart rate. ***B.*** Multifocal PVCs. The two PVCs shown have different morphologies. ***C.*** Example of accelerated idio-ventricular rhythm (see text for details).

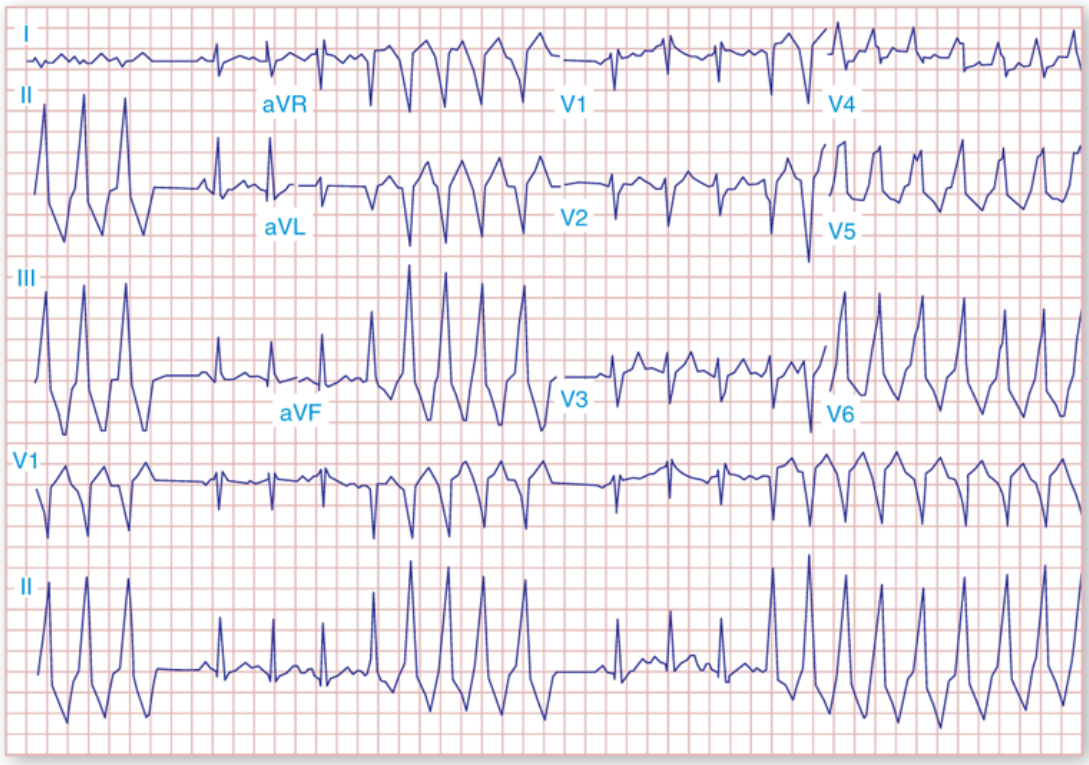


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ventricular tachycardia (*VT*) is three or more consecutive beats at a rate faster than 100 beats/min. Three or more consecutive beats at slower rates are designated an idioventricular rhythm **(Fig. 247-1C)**. VT that terminates spontaneously within 30 s is designated *non-sustained* **(Fig. 247-2)** whereas sustained VT persists >30 s or is terminated by an active intervention, such as administration of an intravenous medication, external cardioversion, or pacing or a shock from an implanted cardioverter defibrillator.

FIGURE 247-2

Repetitive monomorphic non-sustained ventricular tachycardia (VT) of right ventricular outflow tract origin. The VT has a left bundle branch block pattern with inferior axis with tall QRS complexes in the inferior leads.

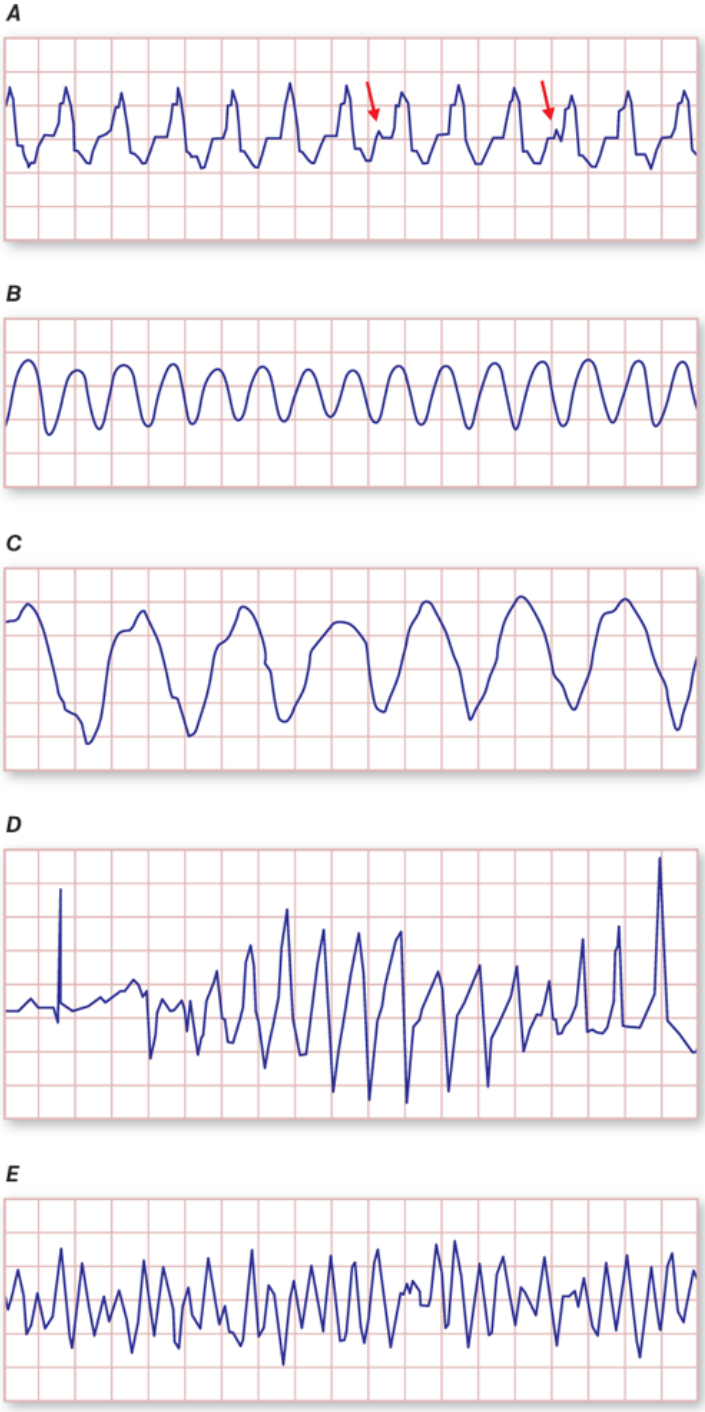


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Monomorphic VT has the same QRS complex from beat to beat, indicating that the activation sequence is the same from beat to beat, and that each beat likely originates from the same source **(Fig. 247-3A)**. The initial site of ventricular activation largely determines the sequence of ventricular activation. Therefore, the QRS morphology of PVCs and monomorphic VT provides an indication of the site of origin within the ventricles **(Fig. 247-4)**. The likely origin often suggests whether an arrhythmia is idiopathic or associated with structural disease. Arrhythmias that originate from the right ventricle or septum result in late activation of much of the left ventricle, thereby producing a prominent S-wave in V1 referred to as a left bundle branch block–like configuration. Arrhythmias that originate from the free wall of the left ventricle have a prominent positive deflection in V1, thereby producing a right bundle branch block–like morphology in V1. The frontal plane axis of the QRS is also useful. An axis that is directed inferiorly, as indicated by dominant R waves in lead II, III, and AVF, suggests initial activation of the cranial portion of the ventricle, whereas a frontal plane axis that is directed superiorly (dominant S waves in II, III, and AVF) suggests initial activation at the inferior wall.

FIGURE 247-3

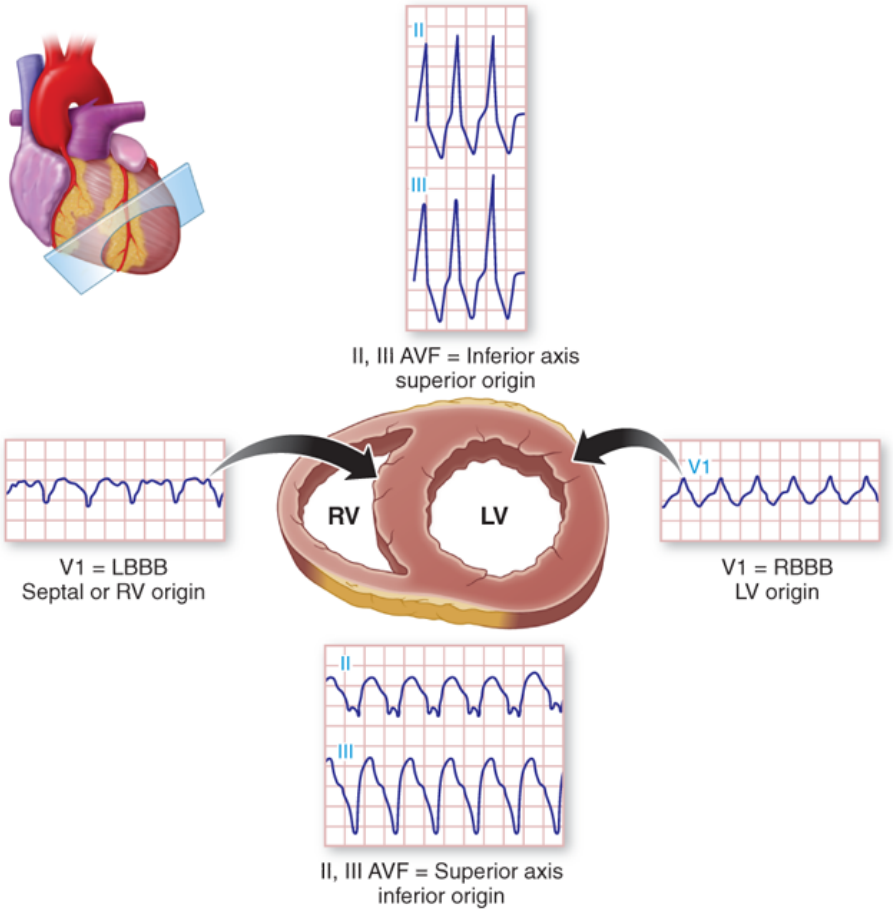
A. Monomorphic ventricular tachycardia (VT) with dissociated P waves (short arrows). ***B.*** Ventricular flutter. ***C.*** Sinusoidal VT due to electrolyte disturbance or drug effects. ***D.*** Polymorphic VT resulting from prolongation of QT interval (torsade de Pointes VT). ***E.*** Ventricular fibrillation (see text for details).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 247-4

Site of ventricular tachycardia (VT) origin based on QRS morphology (see text for details).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Very rapid monomorphic VT has a sinusoidal appearance, also called *ventricular flutter*, because it is not possible to distinguish the QRS complex from the T wave **(Fig. 247-3B)**. Relatively slow *sinusoidal* VTs have a wide QRS indicative of slowed ventricular conduction **(Fig. 247-3C)**. Hyperkalemia, toxicity from excessive effects of drugs that block sodium channels (e.g., flecainide, propafenone, or tricyclic antidepressants) and severe global myocardial ischemia are causes.

Polymorphic VT has a continually changing QRS morphology indicating a changing ventricular activation sequence. Polymorphic VT that occurs in the context of congenital or acquired prolongation of the QT interval often has a waxing and waning QRS amplitude creating a "twisting about the points" appearance referred to as Torsade de Pointes **(Fig. 247-3D)**.

Ventricular fibrillation (VF) has continuous irregular activation with no discrete QRS complexes **(Fig. 247-3E)**. Monomorphic or polymorphic VT may transition to VF in susceptible patients.

The term *idiopathic ventricular arrhythmias* generally refers to PVCs or VT that occurs in patients without structural heart disease and which is not associated with a genetic syndrome or risk of sudden death.

CLINICAL MANIFESTATIONS

Common symptoms of ventricular arrhythmias include palpitations, dizziness, exercise intolerance, episodes of lightheadedness, syncope or sudden cardiac arrest leading to sudden death. Ventricular arrhythmias can also be asymptomatic and encountered unexpectedly as an irregular pulse or heart sounds on examination, or seen on a routine ECG, exercise test or cardiac ECG monitoring.

Syncope is a concerning symptom, that can be due to an episode of VT that produces severe hypotension, which often indicates that there is a risk for cardiac arrest and sudden death with arrhythmia recurrence. Although benign processes, such as reflex mediated neuro-cardiogenic (vasovagal) syncope and orthostatic hypotension, are the most common causes, it is important to consider the possibility of heart disease or a genetic syndrome causing VT. When these are suspected, hospitalization for further evaluation and monitoring is often appropriate.

Sustained VT may present as a wide QRS complex tachycardia that must be distinguished from supraventricular tachycardia with aberrancy (see **Chap. 241**) causing symptoms that range from mild to severe impairment with hypotension with syncope and imminent cardiac arrest. Sustained VT may degenerate to VF, particularly if it is rapid and polymorphic. Many patients who are at risk for VT have known heart disease and many have an implantable

cardioverter defibrillator (ICD). In patients with an ICD, VT episodes may cause transient lightheadedness, palpitations or syncope that may be followed by a shock from the ICD (see below).

EVALUATION OF PATIENTS WITH DOCUMENTED OR SUSPECTED VENTRICULAR ARRHYTHMIAS

There are several important considerations that guide evaluation of patients with documented or suspected cardiac arrhythmias. First, establish whether a ventricular arrhythmia is the cause of the symptoms or clinical presentation. Second, determine whether the arrhythmia is associated with a cardiac disease and establish the prognostic significance of that disease, and in particular whether it is associated with a risk of sudden cardiac death. Finally, define the likelihood of arrhythmia recurrence and the symptoms and risk imposed by the recurrence. The risk of cardiac arrest and sudden cardiac death are largely determined by the cause of the arrhythmia and the associated underlying heart disease.

The diagnosis of ventricular arrhythmias is established by recording of the arrhythmia on an ECG, by an implanted rhythm management device such as a pacemaker or ICD, or in some cases, initiation of the arrhythmia during an electrophysiologic study **(Table 247-1)**. A 12 lead ECG of the arrhythmia should be obtained when possible and often provides clues to the potential site of origin and possible presence of underlying heart disease (see above). For patients with sustained wide complex tachycardia initial management is guided by the patient's hemodynamic stability. The approach to sustained wide complex tachycardia is discussed in **Chap. 249**. The management of VT that causes cardiac arrest is discussed in **Chap. 299**. Once hemodynamic stability is restored further management is guided by the possibility of a recurrence and the risk imposed by a recurrence.

Table 247-1

Diagnostic Tests for Ventricular Arrhythmias

12-Lead ECG

Event recorder

Non-looping, patient activated recorder

- Records only when the patient puts the device in contact with the chest wall and activates it
- Useful for episodes that are infrequent but last for more than several minutes

Looping event recorders

- Continuously recording, storing only segments triggered by the patient or with a heart rate outside set parameters
- Useful for intermittent, infrequent symptomatic and asymptomatic arrhythmias

Continuous ambulatory recording

Holter monitor—typically used for 24–48 h

- Records all arrhythmias during the recording period
- Useful for very frequent arrhythmias (daily) or when quantitation of an arrhythmia is needed (e.g., quantitation of premature ventricular contraction [PVCs] potentially contributing to depressed ventricular function)

Implanted loop recorders

- Allow continuous recording for >1 year
- Useful for capturing rare events such as rare syncopal episodes

Exercise Testing

- Useful for evaluation of exercise induced symptoms; arrhythmias usually emerge during the early recovery phase after exercise
- QT interval response to exercise may be abnormal in long QT syndrome

Electrophysiologic study

- Invasive test that attempts to initiate ventricular arrhythmias in a controlled setting
- Useful for assessing arrhythmia risk when there is concern for a risk of sudden death, but a sufficient diagnosis to guide therapy has not been achieved
- Useful for distinguishing between wide complex tachycardia due to ventricular tachycardia (VT) versus supraventricular tachycardia with aberrancy

Evaluation of the Patient with Arrhythmia Symptoms

When symptoms are intermittent, initial evaluation aims to establish symptom severity, provocative factors and presence of underlying heart disease. Syncope or near syncope raises concern that an arrhythmia is causing episodes of hypotension and that there may be a risk of cardiac arrest if that persists. Symptoms that occur with exertion suggest arrhythmias that are provoked by sympathetic stimulation, but can also be related to exertional ischemia in patients with coronary artery disease, although non-arrhythmia causes must also be considered. A past history of any cardiac disease is

important. A review of all medications is relevant. Medications that prolong the QT interval predispose to polymorphic VT (see **Chap. 250**). Adrenergic stimulants can provoke premature ventricular contractions.

Family history should determine the presence of premature coronary artery disease, cardiomyopathy, or cardiac arrhythmias, particularly a history of sudden death. Family history may also suggest that a possibility of a genetic cause of an arrhythmia warrants careful consideration. Details of premature deaths are relevant. Sudden death victims are often said to have died of a "massive heart attack" despite absence of definite confirmation of thrombotic myocardial infarction and when other causes such as arrhythmia may have been possible.

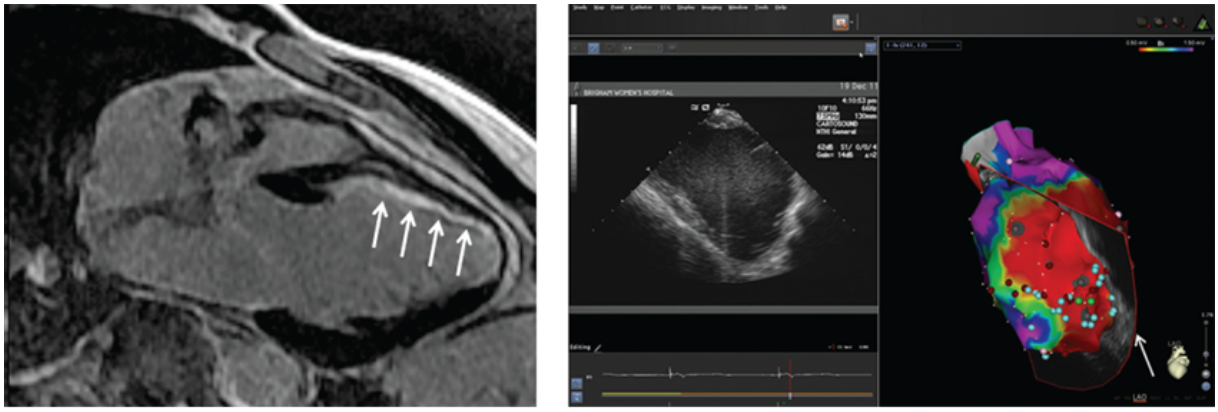
The physical examination focuses on evidence of structural heart disease with assessment of pulse, jugular venous pressure lung fields and cardiac auscultation. Stigmata of neuromuscular disease or dysmorphic features may suggest a genetic arrhythmia syndrome.

A 12-lead ECG should be obtained even if the patient is not having symptoms at the time of evaluation. Occasionally premature ventricular beats will be detected. Patients with benign idiopathic arrhythmias usually have a completely normal ECG during sinus rhythm. Any ECG abnormality warrants further evaluation. Particularly relevant findings include Q-waves that indicate prior myocardial infarction, which may have been silent, and ventricular hypertrophy, which may indicate hypertrophic cardiomyopathy or other ventricular disease. An ECG finding is the major diagnostic manifestation of several genetic arrhythmia syndromes in patients without structural heart disease, including the long QT syndrome, Brugada syndrome, and short QT syndrome (see **Chap. 250**).

If there is suspicion for structural heart disease, cardiac imaging is warranted to assess ventricular function and structure. Transthoracic echocardiography is most frequently employed for initial evaluation. Depressed ventricular function increases concern for a risk of sudden death and warrants further evaluation to establish the cause, which may be cardiomyopathy, coronary artery disease, or valvular heart disease. Ventricular thickening may indicate hypertrophic cardiomyopathy or infiltrative diseases such as amyloidosis. Cardiac MRI with gadolinium contrast imaging provides similar assessment, but also can detect areas of ventricular scar, evident as regions of delayed hyperenhancement, which are usually present in patients who have sustained monomorphic VT **(Fig. 247-5)**. The nature and location of abnormalities is helpful in assessing the type of heart disease. Evaluation to exclude atherosclerotic coronary artery disease should be performed in patients at risk, guided by age and other risk factors.

FIGURE 247-5

Imaging studies of the left ventricle (LV) used to assist ablation for VT. Left panel is an MRI image of a longitudinal section demonstrating thinning of the anterior wall and late gadolinium enhancement in a sub-endocardial scar (*white arrows*). The middle panel shows a 2-D image of the LV in long axis corresponding to the sector through the mid LV (*arrow in figure on right panel*) obtained by an intra-cardiac echo probe positioned in the RV. An electro-anatomic 3-D map of the LV in the left anterior oblique projection is displayed in the right panel. The purple areas depict areas of normal voltage (>1.5 mV). Blue, green, and yellow represent progressively lower voltages with the red areas indicating scar (<0.5 mV). Channel of viable myocardium with slow conduction within the scar are identified with the light blue dots. Areas of ablation delivered to regions involved in re-entrant VT are indicated by maroon dots.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT OPTIONS FOR VENTRICULAR ARRHYTHMIAS

Treatment of ventricular arrhythmias is guided by the severity and frequency of symptoms. For some, reassurance and removal of aggravating factors (e.g., caffeine) is all that is needed. For arrhythmias associated with a sudden death risk, ICD implantation is usually indicated and will provide a "safety-net" to terminate life-threatening VT or VF, preventing sudden death, but without preventing the arrhythmia. When suppression of the arrhythmia is required, antiarrhythmic drug therapy or catheter ablation are major considerations.

Antiarrhythmic Drugs

Use of antiarrhythmic drugs is based on consideration of the risks and potential benefit for the individual patient. Efficacy and side effects for the individual patient is not predictable and is assessed by individual therapeutic trial. Adverse effects are mostly non-cardiac and minor, but can sometimes be severe enough to limit their use. Cardiac side effects, however, include the potential for "pro-arrhythmia" whereby a drug can increase the frequency of arrhythmia or cause a new arrhythmia. Aggravation of bradyarrhythmias is also a common concern. Although anti-arrhythmic drugs are classified based on their actions on receptors or ion channels, most have multiple effects, affecting more than one channel.

Beta-adrenergic Blockers

Many ventricular arrhythmias are sensitive to sympathetic stimulation, and beta-adrenergic stimulation also diminishes the electrophysiologic effects of many membrane active anti-arrhythmic drugs. The safety of beta-blocking agents makes them the first choice of therapy for most ventricular arrhythmias. They are particularly useful for exercise-induced arrhythmias and idiopathic arrhythmias, but have limited efficacy for most arrhythmias associated with heart disease. Bradyarrhythmias and negative inotrophic effects are the major cardiac adverse effects.

Calcium Channel Blockers

The non-dihydropyridine calcium channel blockers diltiazem and verapamil can be effective for some idiopathic VTs. The risk of pro-arrhythmia is low, but they have negative inotropic and vasodilatory effects that can aggravate hypotension.

Sodium Channel Blocking Agents

Drugs whose major effect is mediated through sodium channel blockade include mexiletine, quinidine, disopyramide, flecainide, and propafenone, which are available for chronic oral therapy. Blockade of the fast inward sodium current has been referred to as a Class I antiarrhythmic drug effect. Antiarrhythmic actions are the result of depressing of cardiac conduction and membrane excitability. Conduction slowing can be manifest as a prolongation of QRS duration. Lidocaine, quinidine, and procainamide are available as intravenous formulations. Quinidine, disopyramide, and procainamide also have potassium channel blocking effects that prolong the QT interval (Class III antiarrhythmic drug action) that contributes to its antiarrhythmic effect. These agents have potential pro-arrhythmic effects and, with the possible exception of quinidine, also have negative inotropic effects that may contribute to increased mortality observed when some were administered chronically to patients with prior myocardial infarction. Long-term therapy is generally avoided in patients with structural heart disease, but may be used to reduce symptomatic arrhythmias in patients with ICDs.

Potassium Channel Blocking Agents

Sotalol and dofetilide block the delayed rectifier potassium channel IKr, thereby prolonging action potential duration (QT interval) and the cardiac refractory period, known as the Class III antiarrhythmic drug effect. Sotalol also has non-selective beta-adrenergic blocking activity. It has been shown to have a modest effect on reducing ICD shocks due to ventricular and atrial arrhythmias. Proarrhythmia due to the polymorphic VT *torsade de pointes* that is associated with QT prolongation occurs in 3–5% of patients. Both sotalol and dofetilide are excreted via the kidneys, necessitating dose adjustment or avoidance in renal insufficiency. These drugs must be avoided in patients with other risk factors for torsade de pointes, including QT prolongation, hypokalemia and significant bradycardia.

Amiodarone and Dronedarone

Amiodarone blocks multiple cardiac ionic currents and has sympatholytic activity. It is the most effective antiarrhythmic drug for suppressing ventricular arrhythmias. It is administered intravenously for life-threatening arrhythmias. During chronic oral therapy, electrophysiologic effects develop over several days. It is more effective than sotalol in reducing ICD shocks and is the preferred drug for ventricular arrhythmias in patients with heart disease who are not candidates for an ICD. Bradyarrhythmias are the major cardiac adverse effect. Ventricular pro-arrhythmia can occur, but *torsade de pointes* VT is rare. Non-cardiac toxicities are a major problem and contribute to drug discontinuation in approximately a third of patients during long-term therapy. Hyper or hypothyroidism are related to the iodine content of the drug. Pneumonitis or pulmonary fibrosis occurs in ~1% of patients. Photosensitivity is common, and neuropathy and ocular toxicity can occur. Systematic monitoring is recommended during chronic therapy including assessment for thyroid, liver, and pulmonary toxicity. Intravenous administration of amiodarone via a peripheral vein for >24 h can cause severe peripheral thrombophlebitis. Dronaedarone has structural similarities to amiodarone but without the iodine moiety. Efficacy for ventricular arrhythmias is poor and it increases mortality in patients with heart failure.

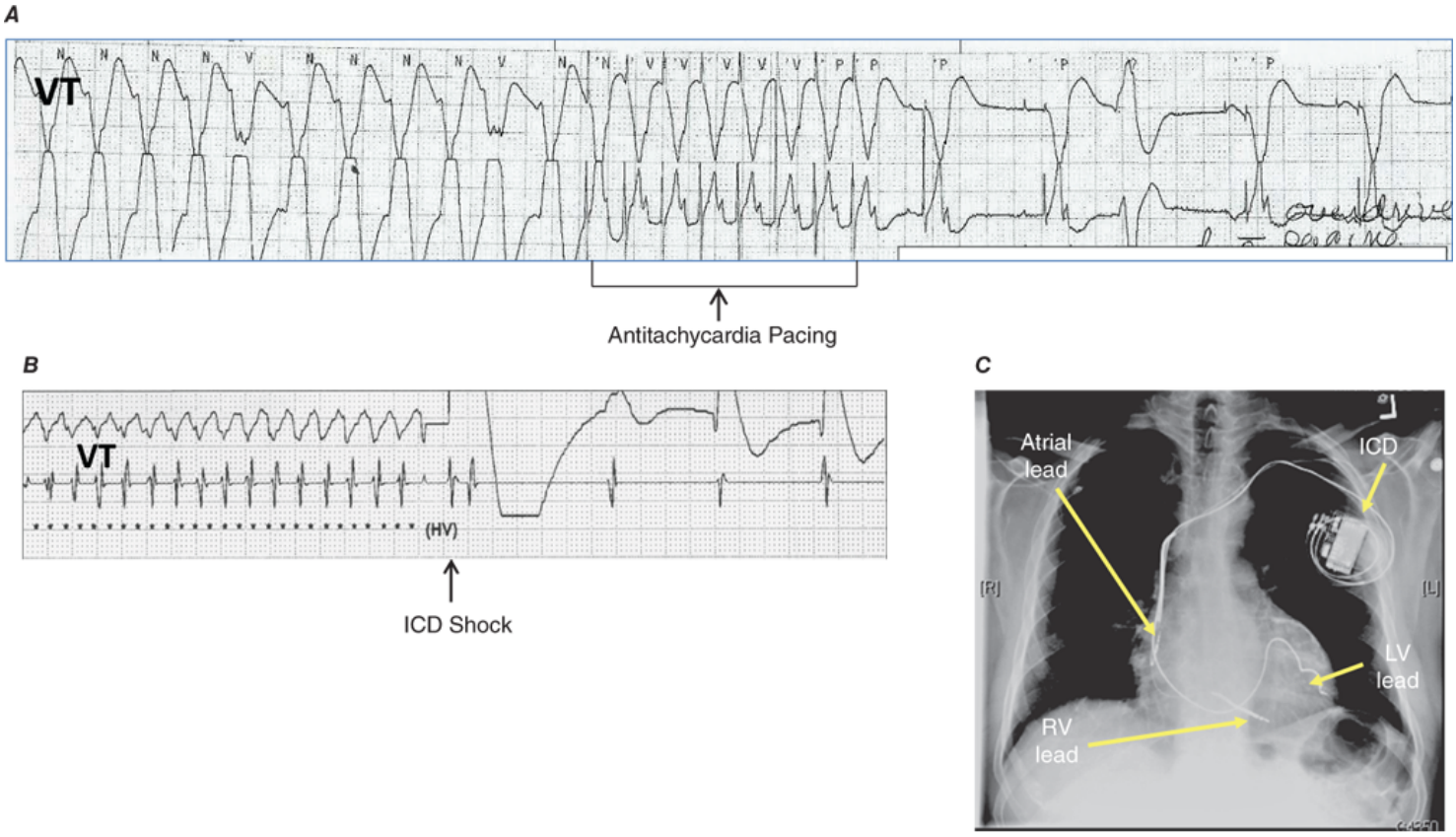
Implantable Cardioverter Defibrillators (ICD)

ICDs detect sustained VT, largely based on heart rate, and then terminate the arrhythmia. VF is terminated by a shock applied between a lead in the RV and the ICD pulse generator. Monomorphic VT can often be terminated by a burst of rapid pacing faster than the VT, known as anti-tachycardia pacing (ATP) **(Fig. 247-6A)**. If ATP fails or is not a programmed treatment, as is often the case for rapid VT or VF, a shock is delivered **(Fig. 247-6B)**. Shocks are painful if the patient is conscious. ICDs are highly effective for termination of VT and VF and also provide bradycardia pacing. The most common ICD complication is the delivery of unnecessary therapy (either ATP or shocks) in response to a rapid supraventricular tachycardia or electrical noise as a result of an ICD lead fracture. ICDs record and store electrograms from arrhythmia episodes which can be retrieved by interrogation of the ICD, which can

be performed remotely and communicated via internet. This assessment is critical after an ICD shock to determine the arrhythmia diagnosis and exclude an unnecessary therapy. Device infection occurs in ~1% of patients.

FIGURE 247-6

Implantable cardioverter defibrillator (ICD) and therapies for ventricular arrhythmias. ***A.*** A monomorphic ventricular tachycardia (VT) is terminated by a burst of pacing impulses at a rate faster than VT (anti-tachycardia pacing). ***B.*** A rapid VT is converted with a high voltage shock (*arrow*). The chest x-ray in Panel ***C*** shows the components of an ICD capable of biventricular pacing. ICD generator in the subcutaneous tissue of the left upper chest, pacing leads in the right atrium and the LV branch of the coronary sinus (LV lead) and a pacing/defibrillating lead in the right ventricle (RV lead) are shown.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ICDs decrease mortality in patients at risk for sudden death due to structural heart diseases. In all cases ICDs are recommended only if there is also expectation for survival of at least a year with acceptable functional capacity. The exception is in cases of patients with end-stage heart disease who are awaiting cardiac transplantation outside the hospital, or who have left bundle branch block QRS prolongation such that they are likely to have improvement in ventricular function with cardiac resynchronization therapy from a biventricular ICD **(Fig. 247-6C)**.

Despite prompt termination of VT or VF by an ICD, the occurrence of these arrhythmias predicts subsequent increased mortality and risk of heart failure. Occurrence of VT or VF should therefore prompt assessment for potential causes including worsening heart failure, electrolyte abnormalities, and ischemia. Repeated shocks, even if appropriate, often induce posttraumatic stress disorder. Antiarrhythmic drug therapy, most commonly amiodarone, or catheter ablation is often required for suppression of recurrent arrhythmias. Antiarrhythmic drug therapy can alter the VT rate and the energy required for defibrillation, thereby necessitating programming changes in the ICD's algorithms for detection and therapy.

The commonly used ICD system consists of endocardial leads to the right heart chambers with a pulse generator implanted in the pre-pectoral area (Fig. 247-6C). This transvenous form of ICD has the disadvantage of vascular occlusion, endocarditis in the event of infection, and difficulty with removal. A totally subcutaneous ICD system is now available. While it has the advantage of avoiding endovascular complications, the present iteration lacks the ability to pace the heart for tachycardia termination or for long-term pacing. A wearable ICD system with electrodes incorporated into a vest and an external battery pack is also available for short-term use in patients pending decision regarding a permanent implanted system.

Catheter Ablation for VT

Catheter ablation is usually performed by applying radiofrequency (RF) current to cause thermal injury by resistive heating of cardiac tissue responsible for the arrhythmia. An electrode catheter is used to map local electrical activity to identify the ventricular myocardium that is causing the arrhythmia, referred to as the arrhythmia substrate. The size and location of the arrhythmia substrate determines the ease and likely effectiveness of the procedure, as well as the potential complications. When the arrhythmia originates from the endocardium, as is most commonly the case, it can be reached from an

endovascular approach via a femoral vein or artery. Less commonly arrhythmias originate from the subepicardium, and percutaneous pericardial puncture, similar to pericardiocentesis, is required to insert a catheter into the pericardial space for mapping and ablation. In patients with scar-related VT due to prior infarction or cardiomyopathy, ablation targets abnormal regions in the scar. Because these scars often contain multiple reentry circuits over relatively large regions, extensive areas of ablation are required and these areas are often identified as regions of low voltage displayed on anatomic reconstructions of the ventricle (Fig. 247-5).

Catheter ablation is often performed in patients with recurrent ventricular arrhythmias associated with poor cardiac function, and the procedure-related mortality in this situation is 0.5–3%. Outcomes are better for patients with prior infarction and VT than for patients with nonischemic cardiomyopathies in which the scar locations are more variable and often intramural or sub-epicardial. Ablation can be lifesaving for patients with very frequent or incessant VT.

Idiopathic VTs and PVCs that occur in the absence of structural heart disease usually originate from a small focus, for which catheter ablation has a higher success rate for preventing recurrent arrhythmia (see **Chaps. 248** and **249**).

ARRHYTHMIA SURGERY

When antiarrhythmic drug therapy and catheter ablation fails, or is not an option, surgical cryoablation, often combined with aneurysmectomy, can be effective therapy for recurrent VT due to prior myocardial infarction and has also been used successfully in a few patients with nonischemic heart disease. Few centers now maintain the expertise for this therapy.

FURTHER READING

Pedersen CT et al: EHRA/HRS/APHRS expert consensus on ventricular arrhythmias. Europace 16:1257, 2014. [PubMed: 25172618]

Priori SG et al: 2015 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: The Task Force for the Management of Patients with Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC). Eur Heart J 36:2793, 2015. [PubMed: 26320108]

Chapter 248: Premature Ventricular Beats, Non-Sustained Ventricular Tachycardia, and Idioventricular Rhythm

Roy M. John,; William G. Stevenson

Figure 248-1

PREMATURE VENTRICULAR CONTRACTIONS AND NON-SUSTAINED VT

Premature ventricular contractions **(see Fig. 247-1A)** can be due to automaticity or reentry **(see Chap. A9)**. They are often sensitive to sympathetic stimulation and can be a sign of increased sympathetic tone, myocardial ischemia, hypoxia, electrolyte abnormalities, particularly hypokalemia, or underlying heart disease. During myocardial ischemia or in association with other heart disease, PVCs can be a harbinger of sustained VT or VF.

The ECG characteristics of the arrhythmia are often suggestive of whether structural heart disease is present. PVCs with smooth uninterrupted contours and sharp QRS deflections suggest an ectopic focus in relatively normal myocardium whereas broad notching and slurred QRS deflections suggest a diseased myocardial substrate. The QRS morphology also suggests the likely site of origin within the ventricle **(see Fig. 247-4)**. PVCs that have a dominant S-wave in V1, referred to as left bundle branch block–like configuration originate from the right ventricle or interventricular septum. Those with a dominant R-wave in V1 originate from the left ventricle. A superior frontal plane axis (negative in II, III, AVF) indicates initial depolarization of the inferior wall (diaphraghmatic aspect of the heart), while an inferior frontal plane axis (positive in II, III, AVF) indicates an origin in the cranial aspect of the heart. The location of arrhythmia origin often suggests the nature of underlying heart disease. Most ventricular arrhythmias that are not associated with structural heart disease have a left bundle branch block–like configuration. PVCs with RBBB configuration are more likely to be associated with structural heart disease. Multiple morphologies of PVCs (multifocal PVCs) are also more likely to indicate structural heart disease **(see Fig. 247-1B)**. In patients with heart disease, a greater frequency and complexity (couplets and non-sustained VT) of these arrhythmias are associated with more severe disease.

PVCs AND NON-SUSTAINED VT DURING ACUTE ILLNESS

These arrhythmias are often encountered in patients who are being evaluated in the emergency room, or who have been hospitalized and are on a cardiac monitor. When encountered during acute illness or as a new

finding, evaluation should focus on detection and correction of potential aggravating factors and causes, specifically myocardial ischemia, ventricular dysfunction and electrolyte abnormalities, most commonly hypokalemia. Underlying heart disease should be defined.

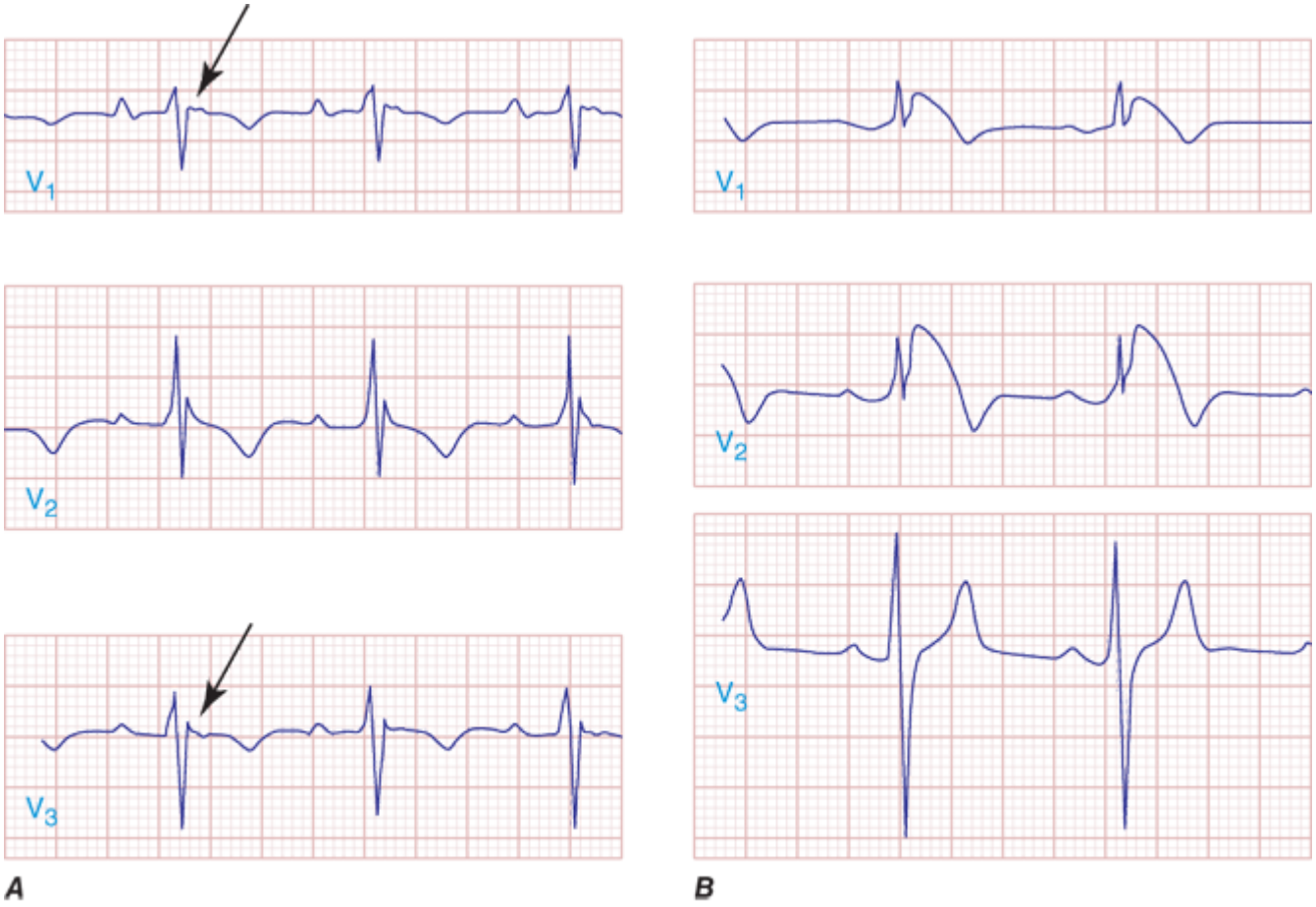
PVCs AND NON-SUSTAINED VT IN PATIENTS WITHOUT HEART DISEASE

The most frequent site of origin for idiopathic ventricular arrhythmias is the right ventricular outflow tract, giving rise to PVCs or VT that have a left bundle branch block–like configuration, with an inferiorly directed frontal plane axis as discussed below **(see Fig. 247-2)**. However, QRS morphology alone is not reliable as an indicator of disease or subsequent risk. Non-sustained VT is usually monomorphic with rates <200 beats/min and typically lasting <8 beats **(see Fig. 247-2)**. Non-sustained VT that is very rapid, polymorphic, or with a first beat that occurs prior to the peak of the T-wave ("short-coupled") is uncommon and should prompt careful evaluation for underlying disease or genetic syndromes associated with sudden death.

A family history of sudden death should prompt evaluation for genetic syndromes associated with sudden death, including cardiomyopathy, long QT syndrome and arrhythmogenic right ventricular cardiomyopathy (ARVC) (see below). Any abnormality on the 12-lead ECG warrants further evaluation **(Fig. 248-1)**. Repolarization abnormalities are seen in a number of genetically determined syndromes associated with sudden death, including the long QT syndrome, Brugada syndrome, ARVC, and hypertrophic cardiomyopathy (Fig. 248-1). An echocardiogram is often necessary to assess ventricular function, wall motion abnormalities, and valvular heart disease. Contrast enhanced cardiac magnetic resonance imaging is also useful for this purpose and for the detection of ventricular scarring that is the substrate for sustained VT **(see Fig. 247-5)**. Exercise stress testing should be performed in patients with effort-related symptoms and for those at risk for coronary artery disease.

Figure 248-1

Precordial chest leads V1–V3 showing typical abnormalities of arrhythmogenic right ventricular cardiomyopathy (ARVC) (*A*) and Brugada syndrome (*B*). In ARVC, there is T inversion and delayed ventricular activation manifest as Epsilon waves (*arrows*). Panel B shows ST elevation in V1 and V2 typical of the Brugada syndrome.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT OF IDIOPATHIC ARRHYTHMIAS

For PVC and non-sustained VT in the absence of structural heart disease, or a genetic sudden death syndrome, no specific therapy is needed unless the patient has significant symptoms or evidence that frequent PVCs are depressing ventricular function (see below). Reassurance that the arrhythmia is benign is often sufficient to allow the patient to cope with the symptoms, which will often wax and wane in frequency over years. Avoiding stimulants, such as caffeine and alcohol, is helpful in some patients. If symptoms require treatment, beta-adrenergic blockers and non-dihydropyridine calcium channel blockers (verapamil and diltiazem) are sometimes helpful. If these fail more membrane active antiarrhythmic drugs or catheter ablation are options. Antiarrhythmic agents flecainide, propafenone, mexiletine, and amiodarone can be effective but the potential for side effects, warrants careful consideration. Catheter ablation is effective at suppressing this arrhythmia in about 80% of patients. Failure of ablation is usually due to inability to provoke the arrhythmia for mapping in the electrophysiology laboratory, or an origin that is not accessible, such as deep within the myocardium.

PVCs AND NON-SUSTAINED VT ASSOCIATED WITH ACUTE CORONARY SYNDROMES

During and soon after acute myocardial infarction (MI), PVCs and non-sustained VT are common and can be an early manifestation of ischemia and a harbinger of subsequent VF. Treatment with beta-adrenergic blockers and correction of hypokalemia and hypomagnesemia reduce the risk of VF. Routine administration of antiarrhythmic drugs such as lidocaine does not reduce mortality and is not indicated for suppression of PVCs or asymptomatic non-sustained VT, but may be implemented transiently if an episode of sustained VT or VF occurs, with the goal of reducing the likelihood of a subsequent episode.

Following recovery from acute MI, frequent PVCs (typically >10 PVCs/h), repetitive PVCs with couplets, and non-sustained VT are markers for depressed ventricular function and increased mortality, but routine antiarrhythmic drug therapy to suppress these arrhythmias does not improve mortality and treatment with the sodium channel blocker flecainide increases mortality. Amiodarone therapy reduces sudden death, but does not improve total mortality. Therefore, amiodarone is an option for treatment of symptomatic arrhythmias in this population when the potential benefit outweighs its potential toxicities. Beta-adrenergic blockers reduce sudden death, but have limited effect on spontaneous arrhythmias.

For survivors of an acute MI, an implantable cardioverter defibrillator (ICD) reduces mortality in certain high-risk groups: patients who have survived >40 days after the acute MI and have a left ventricular ejection fraction of ≤0.30, or who have an ejection fraction <0.35 and have symptomatic heart failure (functional Class II or III); and patients >5 days after MI who have a reduced left ventricular ejection fraction, nonsustained VT, and inducible sustained VT or VF on electrophysiological testing. ICDs do not reduce mortality when routinely implanted soon after MI, and have not been demonstrated to improve mortality when implanted soon after coronary artery revascularization.

PVCs AND NON-SUSTAINED VT ASSOCIATED WITH DEPRESSED VENTRICULAR FUNCTION AND HEART FAILURE

Premature ventricular beats and non-sustained VT are common in patients with depressed ventricular function and heart failure, and are markers for disease severity and increased mortality, but antiarrhythmic drug therapy to suppress these arrhythmias has not been shown to improve survival. The use of anti-arrhythmic drugs whose major action is blockade of the cardiac sodium channel (flecainide, propafenone, mexiletine, quinidine, and disopyramide) is avoided in patients with structural heart disease because of a risk of pro-arrhythmia, negative inotropic effects and increased mortality. Therapy with the potassium channel blocker dofetilide, does not reduce mortality. Amiodarone suppresses ventricular ectopy and reduces sudden death but does not improve overall survival. ICDs are the major therapy to protect against sudden death in patients at high risk and are recommended for those with LV ejection fraction <0.35 and NYHA class II and III heart failure, in whom they reduce mortality from 36 to 29%, over 5 years.

PVC AND NON-SUSTAINED VT ASSOCIATED WITH OTHER CARDIAC DISEASES

Ventricular ectopy is associated with increased mortality in patients with *hypertrophic cardiomyopathy* (Chap. 254) or with *congenital heart disease* associated with right or left ventricular dysfunction. In these patients, management is similar to that for patients with ventricular dysfunction. Pharmacologic suppression of the arrhythmia has not been shown to improve mortality. ICDs are indicated for patients considered at high risk for sudden cardiac death.

PVC-INDUCED VENTRICULAR DYSFUNCTION

Very frequent ventricular ectopy and repetitive non-sustained VT **(see Fig. 247-2)** can depress ventricular function, possibly through an effect similar to chronic tachycardia or by inducing ventricular dysynchrony. Depression of ventricular function rarely occurs unless PVCs account for >15 to 20% of total beats over a 24-h period. Often the PVCs are idiopathic and unifocal, most commonly originating from the left ventricular papillary muscles or outflow tract regions where they can be targeted for ablation. The distinction between PVC- induced ventricular dysfunction as compared to a primary cardiomyopathic process causing ventricular dysfunction and arrhythmia is difficult and in some cases can be made only retrospectively by observing an improvement in ventricular function after the arrhythmia is suppressed with an antiarrhythmic drug, such as amiodarone, or by catheter ablation.

IDIOVENTRICULAR RHYTHMS

Three or more ventricular beats at a rate slower than 100 beats/min are termed an idioventricular rhythm **(see Fig. 247-1C)**. Automaticity is the likely mechanism. Idioventricular rhythms are common during acute MI and may emerge during sinus bradycardia. Often, they are not symptomatic, but hemodynamic compromise may occur with the loss of ventricular synchrony in susceptible patients. Atropine may be administered to increase the sinus rates if this is a concern. This rhythm is also common in patients with cardiomyopathies or sleep apnea. It can also be idiopathic, often emerging when the sinus rate slows during sleep. Therapy should target any underlying cause and correction of bradycardia. Specific antiarrhythmic therapy for asymptomatic idioventricular rhythm is not necessary.

FURTHER READING

Katritsis DG et al: Non-sustained ventricular tachycardia. J Am Coll Cardiol 60:1993, 2012. [PubMed: 23083773]

Pedersen CT et al: EHRA/HRS/APHRS expert consensus on ventricular arrhythmias. Europace 16:1257, 2014. [PubMed: 25172618]

Chapter 249: Sustained Ventricular Tachycardia

Roy M. John; William G. Stevenson

FIGURE 249-1

INTRODUCTION

Sustained monomorphic ventricular tachycardia (VT) presents as a wide QRS tachycardia that has the same QRS configuration from beat to beat indicating an identical sequence of ventricular depolarization for each beat **(see Fig. 247-3A)**. VT originates from a stable focus or reentry circuit. In structural heart disease, the substrate is often an area of patchy replacement fibrosis due to infarction, inflammation or prior cardiac surgery that creates anatomical or functional reentry pathways **(see Fig. 247-5)**. Less commonly, VT is related to reentry or automaticity in a diseased Purkinje system. Idiopathic VT occurs in the absence of structural heart disease and is due to a focal region of automaticity or reentry involving a portion of the Purkinje system.

The clinical presentation varies depending on the rate of the arrhythmia, underlying cardiac function, and autonomic adaptation in response to the arrhythmia. While rapid VT, >200 beats/min, usually causes hypotension that may present as syncope, patients with normal cardiac function might tolerate rapid VT, and those with severe left venricular (LV) dysfunction may experience symptoms of hypotension, even if VT is slower than 150 beats/min. Monomorphic VT that is rapid or associated with structural heart disease may deteriorate to ventricular fibrillation (VF), which may be the initial cardiac rhythm recorded at the time of resuscitation of a cardiac arrest.

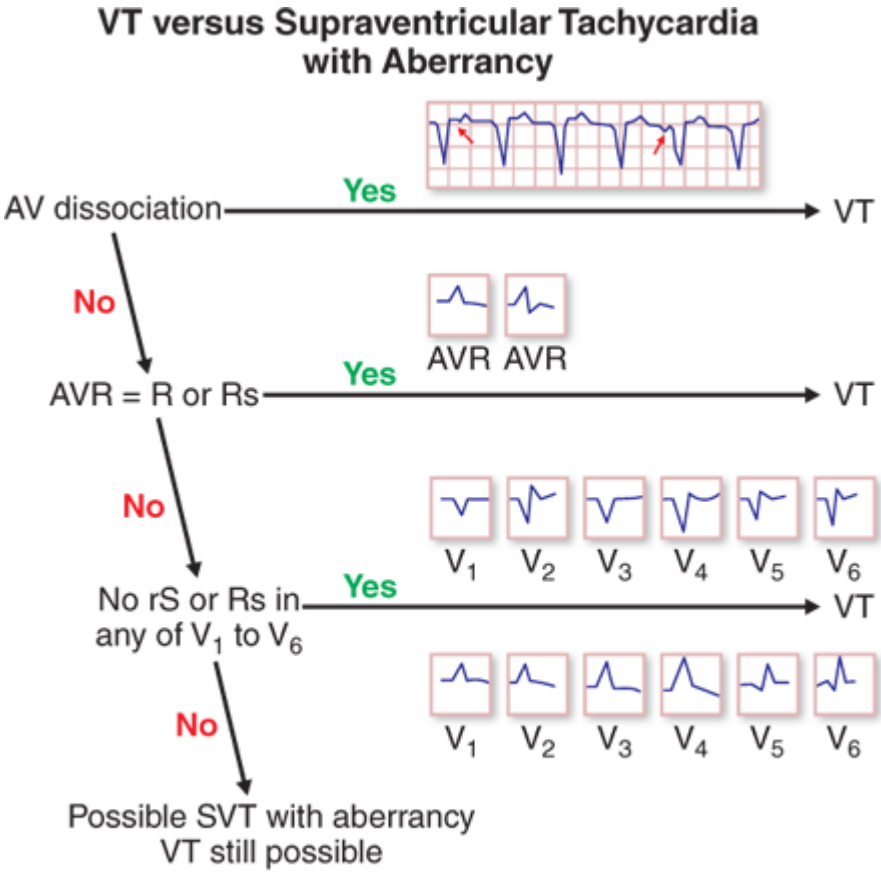
DIAGNOSIS

Sustained monomorphic VT has to be distinguished from other causes of uniform wide QRS tachycardia. These include supraventricular tachycardia with left or right bundle branch block aberrant conduction, supraventricular tachycardias conducted to the ventricles over an accessory pathway **(Chap. 241)**, and rapid cardiac pacing in a patient with a pacemaker or defibrillator. In the presence of known heart disease VT is the most likely diagnosis of a wide QRS tachycardia. Hemodynamic stability during the arrhythmia does not exclude VT. A number of ECG criteria have been evaluated to distinguish supraventricular tachycardia with aberrancy from VT. The presence of AV dissociation is usually a reliable marker for VT **(Fig. 249-1)**, but P-waves can be difficult to define. A P-wave following each QRS does not exclude VT because 1:1 conduction from ventricle to atrium can occur. A monophasic R wave or Rs complex in AVR or concordance from V1 to V6 of monophasic R or S waves are also relatively specific for VT (Fig. 249-1). A number of other QRS morphology

criteria have also been described, but all have limitations, and are not very reliable in patients with severe heart disease. In patients with known bundle branch block, the same QRS morphology during tachycardia as during sinus rhythm suggests supraventricular tachycardia rather than VT, but is not absolutely reliable. An electrophysiological study is sometimes required for definitive diagnosis. Occasionally, noise and movement artifacts on telemetry recordings can simulate VT; prompt recognition can avoid unnecessary tests and interventions.

FIGURE 249-1

Algorithm for differentiation of ventricular tachycardia from supraventricular tachycardia with aberration.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

When LV function is depressed or there is evidence of structural myocardial disease, scar-related reentry is the most likely cause of sustained monomorphic VT. Scars are suggested by pathologic Q-waves on the ECG, segmental left or right ventricular wall motion abnormalities on echocardiogram or nuclear imaging, and areas of delayed gadolinium enhancement during MR imaging **(see Fig. 247-5)**.

TREATMENT AND PROGNOSIS

Initial management follows Advanced Cardiac Life Support (ACLS) guidelines **(Chap. 299)**. If hypotension, impaired consciousness, or pulmonary edema are present, QRS synchronous electrical cardioversion should be performed, ideally after sedation if the patient is conscious. For stable tachycardia a trial of adenosine is reasonable, as this may clarify a supraventricular tachycardia with aberrancy **(Chap. 241)**. Intravenous

amiodarone is the drug of choice if heart disease is present. Following restoration of sinus rhythm, hospitalization and evaluation to define underlying heart disease is required. Assessment of cardiac biomarkers for evidence of myocardial infarction (MI) is appropriate, but acute MI is rarely a cause of sustained monomorphic VT, and elevations in troponin or CK-MB are more likely to indicate myocardial damage that is secondary to hypotension and ischemia from the VT. Subsequent management is determined by the underlying heart disease and frequency of VT. If VT recurs frequently or is incessant, administration of antiarrhythmic medications, or catheter ablation may be required to restore stability. More commonly sustained monomorphic VT occurs as an isolated episode, but with a risk of recurrence. Implantable cardioverter defibrillators (ICDs) are usually warranted for sustained VT associated with structural heart disease.

SUSTAINED MONOMORPHIC VT IN SPECIFIC DISEASES

Coronary Artery Disease

Patients who present with sustained monomorphic VT associated with coronary artery disease typically have a history of prior large MI and present years after the acute infarct with a remodeled ventricle and markedly depressed left ventricular function. Even when there is biomarker evidence of acute MI, a preexisting scar from previous MI should be suspected as the cause of the VT. Infarct scars provide a durable substrate for sustained VT and up to 70% of patients have a recurrence of the arrhythmia within 2 years. Scar-related reentry is not dependent on recurrent acute myocardial ischemia, so coronary revascularization cannot be anticipated to prevent recurrent VT, even when it may be appropriate for treatment of angina or other indications. Depressed ventricular function, which is a risk factor for sudden death, is usually present. Implantation of an ICD is indicated for most patients provided that there is a reasonable expectation of survival with acceptable functional status for the next year after recovery from the VT episode. Compared with antiarrhythmic drug therapy ICDs reduce annual mortality from 12.3 to 8.8% and lower arrhythmic deaths by 50% in patients with hemodynamically significant sustained VT or a history of cardiac arrest. Chronic amiodarone therapy may be considered for patients who are not candidates for, or who decline ICD placement.

Following ICD implantation, patients remain at risk for heart failure, recurrent ischemic events, and recurrent VT, with a 5-year mortality that exceeds 30%. Attention to therapies that benefit patients with depressed ventricular function, including beta-adrenergic blocking agents, angiotensin converting enzyme inhibitors, and statins is important. Patients with frequent symptomatic recurrences of VT require antiarrhythmic drug therapy or catheter ablation.

Nonischemic Dilated Cardiomyopathy

Sustained monomorphic VT associated with nonischemic cardiomyopathy is usually due to scar-related reentry. The etiology of scar is often unclear, but progressive replacement fibrosis is the likely cause. On cardiac MR imaging, scars are detectable as areas of delayed gadolinium enhancement and are more often

intramural or sub-epicardial in location as compared with patients with prior MI. Scars that cause VT are often located adjacent to a valve annulus and can occur in either ventricle. Any cardiomyopathic process can cause scars and VT, but cardiac sarcoidosis, Chagas disease, and cardiomyopathy due to Lamin A/C mutations are particularly associated with monomorphic VT **(Table 249-1)**. An ICD is usually indicated with additional drugs or catheter ablation for control of recurrent VT.

Table 249-1

Ventricular Arrhythmias Associated with Different Forms of Heart Disease

1. Idiopathic ventricular tachycardia (VT) without structural heart disease
   A. Outflow tract origin
      a. Right ventricular (RV) outflow tract: left bundle branch block pattern with inferior axis (tall QRS in inferior leads) and late transition in the precordial leads
      b. Left ventricular (LV) outflow tract: prominent "r" in V1
   B. Left posterior fascicular VT
      Right bundle branch block pattern with left axis deviation (most common)
2. Ischemic cardiomyopathy
- Monomorphic VT is common with prior large myocardial infarction
- Polymorphic VT and ventricular fibrillation (VF) should prompt ischemia evaluation
3. Nonischemic cardiomyopathy
- Polymorphic VT and VF more common but fibrotic scars can cause monomorphic VT especially with sarcoidosis, Chagas disease, Lamin A/C genetic cardiomyopathy
4. Arrhythmogenic Right Ventricular Cardiomyopathy
- Monomorphic VT usually of right ventricular origin (left bundle branch morphology)
- Polymorphic VT and VF can occur independently or through degeneration of monomorphic VT
5. Repaired Tetralogy of Fallot
- Monomorphic VT of right ventricular origin (usually left bundle branch morphology)
6. Hypertrophic Cardiomyopathy
- Polymorphic VT or ventricular fibrillation
- Less commonly, monomorphic VT associated with myocardial scars
7. Genetic Arrhythmia Syndromes:
   a. Long QT syndrome
   - *Torsade de Pointes* VT
   b. Brugada syndrome
   - Ventricular fibrillation
   c. Catecholaminergic polymorphic VT
   - Polymorphic VT or bidirectional VT
   d. Short QT syndrome
   - Ventricular fibrillation
   e. Early repolarization syndrome
8. Polymorphic VT or VF

MONOMORPHIC VT IN ARRHYTHMOGENIC Right Ventricular (RV) CARDIOMYOPATHY (ARVC)

ARVC **(Chap. 254)** is a rare genetic disorder most commonly due to mutations in genes encoding for cardiac desmosomal proteins. Approximately 50% have a familial transmission with autosomal dominant inheritance. A less common, autosomal recessive form is associated with cardio-cutaneous syndromes that include Naxos disease and Carvajal syndrome. Patients typically present between the second and fifth decade with palpitations, syncope or cardiac arrest owing to sustained monomorphic VT, although polymorphic VT can also occur. Fibrosis and fibro-fatty replacement most commonly involves the right ventricular myocardium, and provide the substrate for reentrant VT that usually has a left bundle branch block–like configuration, consistent with the right ventricular origin and can resemble idiopathic VT. The sinus rhythm ECG suggests the disease in >85% of patients, most often showing T-wave inversions in V1-V3 **(see Fig. 248-1)**. Delayed activation of the right ventricle may cause a widened QRS (>110 ms) in the right precordial leads and a prolonged S-wave upstroke in those leads, and occasionally a deflection at the end of the QRS known as an "Epsilon" wave **(see Fig. 248-1)**. Cardiac imaging may show right ventricular enlargement or areas of abnormal motion, or reveal areas of scar on contrast enhanced MRI.

Left ventricular involvement can occur and occasionally precede manifest right ventricular disease. Heart failure is rare except in late stages, and survival to advanced age can be anticipated provided that VT can be controlled. An ICD is recommended. When VT is exercise-induced, it may respond to beta-adrenergic blockers and limiting exercise. Sotalol and amiodarone have been used to reduce recurrences. Catheter ablation prevents or reduces VT episodes in 70% of patients, but epicardial mapping and ablation is often required.

Tetralogy of Fallot

Ventricular tachycardia occurs in 3–14% of patients late after repair of tetralogy of Fallot, and contributes to a 2% per decade risk of sudden death. Monomorphic VT is due to reentry around areas of surgically created scar in the RV (Table 249-1). Factors associated with VT risk include age >5 years at the time of repair, high-grade ventricular ectopy, inducible VT on an electrophysiologic study, abnormal RV hemodynamics, and sinus rhythm QRS duration >180 ms. An ICD is usually warranted for patients who have a spontaneous episode of VT, but criteria for a prophylactic ICD in other patients have not been established. Catheter ablation or anti-arrhythmic drug therapy is used to control recurrent episodes.

BUNDLE BRANCH REENTRY VT

Reentry through the Purkinje system occurs in ~5% of patients with monomorphic VT in the presence of structural heart disease. The reentry circuit typically revolves retrograde via the left bundle and anterograde down the right bundle, thereby producing VT that has a left bundle branch block configuration. Catheter ablation of the right bundle branch abolishes this VT. Bundle branch reentry is usually associated with severe underlying heart disease. Other scar-related VTs are often present and often require additional therapy or ICD implantation.

IDIOPATHIC MONOMORPHIC VT

Idiopathic VT in patients without structural heart disease usually presents with palpitations, lightheadedness, and occasionally syncope, often provoked by sympathetic stimulation during exercise or emotional upset. The QRS morphology of the arrhythmia suggests the diagnosis (see below). The sinus rhythm ECG is normal. Cardiac imaging shows normal ventricular function and no evidence of ventricular scar. Occasionally a patient with structural heart disease is found to have concomitant idiopathic VT, unrelated to the structural disease. Sudden death is rare.

Outflow *Tract* VTs originate from a focus, usually with features consistent with triggered automaticity. The arrhythmia may present with sustained VT, non-sustained VT or PVCs often provoked by exercise or emotional upset. Repeated bursts of non-sustained VT, which may occur incessantly, are known as repetitive monomorphic VT and can cause tachycardia—induced cardiomyopathy with depressed ventricular function that recovers after suppression of the arrhythmia. Most originate in the right ventricular outflow tract, which gives rise to VT that has a left bundle branch block configuration in V1 and an axis that is directed inferiorly, with tall R-waves in II, III, and AVF **(see Fig. 247-2)**. Idiopathic VT can also arise in the left ventricular outflow tract or in sleeves of myocardium that extend along the aortic root. LV origin is suspected when leads V1 or V2 have prominent r-waves (Table 249-1). Although this typical outflow tract QRS morphology favors idiopathic VT, some cardiomyopathies, notably ARVC, can cause PVCs or VT from this region. Excluding these diseases is an initial focus of evaluation.

***Left ventricular* intrafascicular VT** presents with sustained VT that has a right bundle branch block–like configuration. It is often exercise induced and occurs more often in men than women. The mechanism is reentry in or near the septal ramifications of the left ventricular Purkinje system.

MANAGEMENT OF IDIOPATHIC VT

Treatment is required for symptoms or when frequent or incessant arrhythmias depress ventricular function. Beta-adrenergic blockers are first-line therapy. Non-dihydropyridine calcium channel blockers (diltiazem and verapamil) are sometimes effective. Catheter ablation is warranted for severe symptoms or when beta-blockers or calcium channel blockers are not effective or not desired. Efficacy and risks of catheter ablation vary with the specific site of origin of the VT, being most favorable for arrhythmias originating in the right ventricular outflow tract. Failure of ablation is often due to inability to initiate the arrhythmia for mapping in the electrophysiology laboratory.

Left ventricular interfascicular VT can be terminated by intravenous administration of verapamil, although chronic therapy with oral verapamil is not always effective. Catheter ablation is recommended if beta-adrenergic blockers or calcium channel blockers are ineffective or not desired.

FURTHER READING

Alzand BS, Crijns HJ: Diagnostic criteria for broad QRS complex tachycardia: Decades of evolution. Europace 13:465, 2011. [PubMed: 21131372]

John RM, Stevenson WG: Outflow tract premature ventricular contractions and ventricular tachycardia: The typical and the challenging. Card Electrophysiol Clin 8:545, 2016. [PubMed: 27521088]

Pedersen CT et al: EHRA/HRS/APHRS Expert consensus on ventricular arrhythmias. Europace 16:1257, 2014. [PubMed: 25172618]

Chapter 250: Polymorphic Ventricular Tachycardia and Ventricular Fibrillation

Roy M. John; William G. Stevenson

FIGURE 250-1

POLYMORPHIC VENTRICULAR TACHYCARDIA (VT)

Sustained polymorphic VT can be seen with any form of structural heart disease. However, unlike sustained monomorphic VT, polymorphic VT does not always indicate a structural abnormality or focus of automaticity. Reentry with continually changing reentrant paths, spiral wave reentry and multiple automatic foci are potential mechanisms **(Chap. A9)**. Sustained polymorphic VT usually degenerates into ventricular fibrillation (VF). Polymorphic VT is typically seen in association with acute myocardial infarction or ischemia (MI), ventricular hypertrophy, and a number of genetic mutations that affect cardiac ion channels **(see Table 249-1)**.

Polymorphic VT Associated with Acute Myocardial Infarction/Ischemia

Acute MI or ischemia is a common cause of polymorphic VT and should be the initial consideration in management. Approximately 10% of patients with acute MI develop VT that degenerates to VF, likely related to reentry through the infarct border zone. The risk is greatest in the first hour of acute MI. Following defibrillation as per the Advanced Cardiac Life Support (ACLS) guidelines, management is as for acute MI **(Chap. 269)**. Beta-adrenergic blockers, correction of electrolyte abnormalities, and prompt myocardial reperfusion are required. Repeated episodes of polymorphic VT suggest ongoing MI and warrant assessment of adequacy of myocardial reperfusion. Polymorphic VT and VF that occur within the first 48 h of acute MI are associated with greater in-hospital mortality, but those that survive past hospital discharge are not at increased risk for arrhythmic sudden death. Long-term therapy for post-infarct ventricular arrhythmia is determined by residual left ventricular (LV) function with an implantable cardioverter defibrillator (ICD) being indicated for persistent severe left ventricular (LV) dysfunction (LV ejection fraction <0.35).

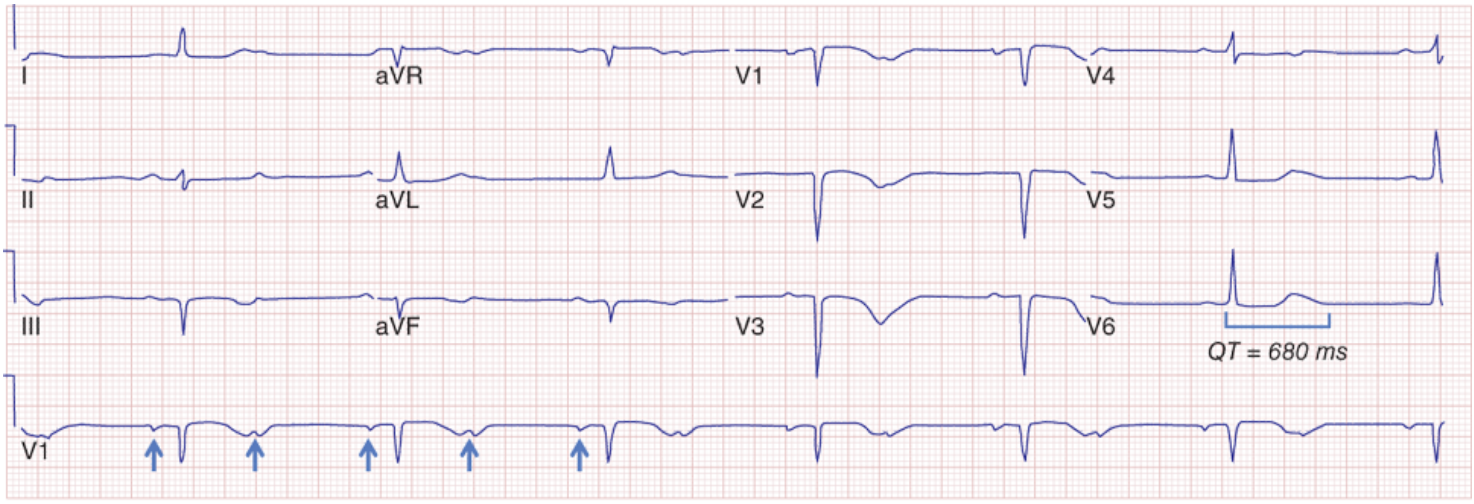
Repolarization Abnormalities and Genetic Arrhythmia Syndromes

ACQUIRED LONG QT

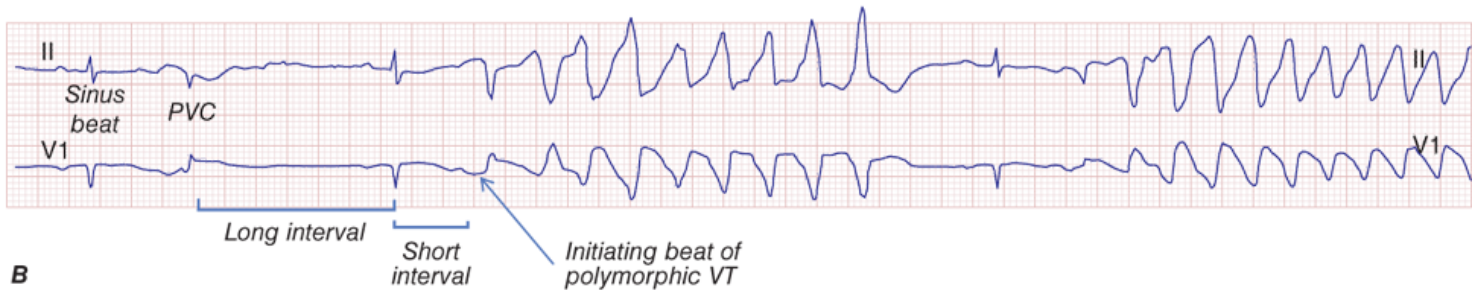
Abnormal prolongation of the QT interval is associated with the polymorphic VT *torsades des pointes* **(Fig. 250-1)**. The VT often has a characteristic initiation sequence of a premature ventricular beat that induces a pause, followed by a sinus beat that has a longer QT interval and interruption of the T-wave by the premature ventricular contraction (PVC) that is the first beat of the polymorphic VT. This characteristic initiation is termed “pause-dependent” (Fig. 250-1). Causes of QT prolongation include electrolyte abnormalities, bradycardia, and a number of medications that block repolarizing potassium currents, notably the antiarrhythmic drugs sotalol, dofetilide, and ibutilide, but also a number of other medications used for non-cardiac diseases, including erythromycin, pentamidine, haloperidol, phenothiazines, and methadone **(Table 250-1)**. Individual susceptibility may be related to genetic polymorphisms or mutations that influence repolarization.

FIGURE 250-1

Torsades des pointes **VT in patient with bradycardia and marked QT prolongation.** ***A.*** 12-lead ECG showing 2:1 AV block (*P waves marked by blue arrows*) with heart rate of 40 bpm and QT interval of 680 ms and corrected QT of 550 ms. ***B.*** The bottom panel shows a telemetry rhythm strip with periods of self-limiting *torsades des pointes* polymorphic VT. Following a normally conducted sinus beat, a premature ventricular contraction (PVC) causes a compensatory pause leading to a long RR interval. A PVC after the next sinus beat initiates VT. This is the classic “pause dependent” mode of initiation of *torsades des pointes* VT with long–short intervals. VT, ventricular tachycardia.



A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 250-1

Causes of QT Prolongation and *Torsade de Pointes* Ventricular Tachycardia (VT)

1. Congenital long QT syndromes (see text for details)
   - Long QT syndrome type 1: Reduced repolarizing current I_{Ks} due to mutation in *KCNQ1* gene
   - Long QT syndrome type 2: Reduced repolarizing current I_{Kr} due to mutation in *KCNH2* gene
   - Long QT syndrome type 3: Delayed inactivation of the I_{Na} due to mutations in *SCN5A* gene
   - Others: Several other types of Long QT syndromes have been described; long QT types 1, 2, and 3 account for 80–90% of cases

2. Acquired Prolongation of QT Interval
   - Electrolyte abnormalities:
     - Hypokalemia
     - Hypomagnesemia
     - Hypocalcemia
   - Drugs:
   - *Antiarrhythmic drugs*
     - Class IA: Quinidine, disopyramide, procainamide
     - Class III: Sotalol, amiodarone (QT prolongation common but torsade VT is rare), ibutilide, dofetilide, almokalant
   - *Antibiotics*
     - Macrolides: Erythromycin, clarithromycin, azithromycin
     - Fluroquinolones: Levofloxacin, moxifloxacin, gatifloxacin,
     - Trimethoprim-sulfamethoxazole
     - Clindamycin
     - Pentamidine
     - Chloroquine
     - Antifungals: Ketoconazole, itraconazole
     - Antivirals: Amantadine
   - *Antipsychotics*
     - Haloperidol, phenothiazines, thioridazine, trifluoperazine, sertindole, zimeldine, ziprasidone
     - Tricyclic and tetracyclic antidepressants
   - *Antihistamines (histamine 1-receptor antagonists)*
     - Terfenadine, astemizole, diphenhydramine, hydroxyzine
   - *Cholinergic antagonists:* Cisapride, organophosphates
   - *Citrate* (massive blood transfusions)
   - *Cocaine*
   - *Methadone*
   - *Fluoxetine* (in conjunction with other drugs that prolong QT)
   - *Cardiac conditions*
     - Myocardial ischemia and infarction
     - Myocarditis
     - Marked bradycardia
     - Stress cardiomyopathy
   - *Endocrine disorders*
     - Hypothyroidism
     - Hyperparathyroidism
     - Pheochromocytoma
     - Hyperaldosteronism
   - *Intracranial disorders*
     - Subarachnoid hemorrhage
     - Thalamic hematoma
     - Cerebrovascular accident
     - Encephalitis

Head injury

Nutritional disorders

Anorexia nervosa

Starvation

Liquid protein diets

Gastroplasty and ileojejunal bypass

Celiac disease

Patients typically present with near-syncope, syncope, or cardiac arrest. Sustained episodes degenerate to VF requiring defibrillation. PVCs and non-sustained VT often precede episodes of sustained VT. Intravenous administration of 1–2 g of magnesium sulphate, usually suppresses recurrent episodes. If magnesium alone is ineffective, increasing heart rate with isoproterenol infusion or pacing, to a rate of 100–120 depolarizations/min as required to suppress PVCs, usually suppresses VT recurrences. These maneuvers allow time for correction of associated electrolyte disturbance (hypokalemia and hypocalcemia) and bradycardia and removal of any causative drugs (Table 250-1). Drug interactions that elevate levels of the offending agent are often a precipitating factor. Patients who experience a polymorphic VT induced by QT prolongation should be considered to have a susceptibility to the arrhythmia, and should avoid all future exposure to medications known to prolong the QT interval.

CONGENITAL LONG QT SYNDROME

The congenital long QT syndrome (LQTS) is caused by mutations in genes coding for cardiac ion channels responsible for ventricular repolarization. The corrected QT (QTc) is typically prolonged to >440 ms in men and 460 ms in women. Symptoms are due to *torsades des pointes* VT (Fig. 250-1). Several forms of congenital LQTS have been identified, but three groups of mutations that lead to LQT-1, LQT-2 or LQT-3 syndromes account for 90% of cases. The most frequently encountered mutations, LQTS 1 and 2, are due to abnormalities of potassium channels, but mutations affecting the sodium channel (LQTS-3) and calcium channels have also been described (Table 250-1).

Typical presentation is with syncope or cardiac arrest, usually during childhood. In LQTS-1, episodes tend to occur during exertion, particularly swimming. In LQTS-2, sudden auditory stimuli or emotional upset predispose to events. In LQTS-3, sudden death tends to occur during sleep. Asymptomatic patients may be discovered in the course of family screening or on a routine ECG. Genotyping can be helpful for family screening and to provide reassurance regarding the diagnosis. Correlations of genotype with risk and response to therapy are beginning to emerge. In most patients with LQT-1 or LQT-2, adequate doses of beta-blocker therapy (the non-selective agents nadolol or propranolol are favored) are sufficient protection from arrhythmia episodes. Markers of increased risk include QTc interval exceeding 0.5 s, female gender, and a history of syncope or cardiac arrest. Recurrent syncope despite beta-blocker therapy or a high-risk profile merits consideration of an ICD. Avoidance of QT prolonging drugs is critical for all patients with the LQTS including those who are genotype positive, but have normal QT intervals.

SHORT QT SYNDROME

Short QT syndrome is very rare compared to the LQTS. The QTc is shorter than 0.36, and usually less than 0.3 s. The genetic abnormality causes a gain of function of the potassium channel (I_{Kr}) or reduced inward depolarizing currents. The abnormality is associated with atrial fibrillation, polymorphic VT, and sudden death.

BRUGADA SYNDROME

Brugada syndrome is a rare syndrome characterized by >0.2 mV of ST segment elevation with a coved ST segment and negative T-wave in more than one anterior precordial lead (V1–V3) **(see Fig. 248-1)** and episodes of syncope or cardiac arrest due to polymorphic VT in the absence of structural heart disease. Cardiac arrest may occur during sleep or be provoked by febrile illness. Males are more commonly affected than females. Mutations involving cardiac sodium channels are identified in ~25% of cases. Distinction from patients with similar ST elevation owing to left ventricular hypertrophy, pericarditis, myocardial ischemia or MI hyperkalemia, hypothermia, right bundle branch block and arrhythmogenic right ventricular cardiomyopathy (ARVC) is often difficult. Furthermore, the characteristic ST-segment elevation can wax and wane over time and may become pronounced during acute illness and fever. Administration of the sodium channel blocking drug flecainide, ajmaline, or procainamide can augment or unmask ST elevation in affected individuals. An ICD is indicated for individuals who have had unexplained syncope or been resuscitated from cardiac arrest. Quinidine and catheter ablation of abnormal regions in the epicardial right ventricular (RV) has been used successfully to suppress frequent episodes of VT.

EARLY REPOLARIZATION SYNDROME

Patients resuscitated from VF who have no structural heart disease or other identified abnormality have a higher prevalence of J-point elevation with notching in the terminal QRS. A family history of sudden death is present in some patients, suggesting a potential genetic basis. J-point elevation is also seen in some patients with the Brugada syndrome and is associated with a higher risk of arrhythmias. An ICD is recommended for those who have had prior cardiac arrest. It should be noted that J-point elevation is commonly seen as a normal variant in patients without arrhythmias and in the absence of specific symptoms, the clinical relevance is not known.

CATECHOLAMINERGIC POLYMORPHIC VT

This rare familial syndrome is due to mutations in the cardiac ryanodine receptor and less commonly, the sarcoplasmic calcium binding protein, calsequestin 2. These mutations result in abnormal sarcoplasmic calcium handling and polymorphic ventricular arrhythmias that resemble those seen with digitalis toxicity. The VT is polymorphic or has a characteristic alternating QRS morphology termed bidirectional VT. Patients usually present during childhood with exercise or emotion induced palpitations, syncope, or cardiac arrest. Beta-adrenergic blockers (e.g., nadolol and propranolol) and an implantable defibrillator are usually recommended. Verapamil or flecainide or surgical left cardiac sympathetic denervation reduces or prevents recurrent VT in some patients.

HYPERTROPHIC CARDIOMYOPATHY (HCM)

HCM is the most common genetic cardiovascular disorder occurring in 1 in 500 individuals and is a prominent cause of sudden death before the age of 35 years **(Chap. 254)**. Sudden death can be due to polymorphic VT/VF. Rarely sustained monomorphic VT occurs related to areas of ventricular scar. Risk factors for sudden death in this disease include young age, non-sustained VT, failure of blood pressure to increase during exercise, recent (within 6 months) syncope, ventricular wall thickness >3 cm, and possibly the severity of LV outflow obstruction. An ICD is generally indicated for high-risk subjects, but the specific risk profile warranting an ICD continues to be debated. Surgical myectomy, performed to relieve outflow obstruction has been associated with a sudden death rate of <1% per year. The reported annual rate of sustained VT or sudden death after transcoronary ethanol septal ablation done to relieve outflow obstruction has been reported to range between 1 and 5%.

GENETIC DILATED CARDIOMYOPATHIES

Genetic dilated cardiomyopathies account for 30–40% of cases of nonischemic dilated cardiomyopathies. Some are associated with muscular dystrophy. Autosomal dominant, recessive, X-linked, and mitochondrial inheritance patterns are recognized. Mutations in genes coding for structural proteins of the nuclear lamina (Lamin A/C) and the SCN5A gene are particularly associated with conduction system disease and ventricular arrhythmias. They can experience polymorphic VT and cardiac arrest or develop areas of scar causing sustained monomorphic VT. ICDs are recommended for those who have had a sustained VT or are at high risk due to significantly depressed ventricular function (LV ejection fraction ≤0.35 and associated with heart failure) or a malignant family history of sudden death.

VENTRICULAR FIBRILLATION

VF is characterized by disordered electrical ventricular activation without identifiable QRS complexes **(see Fig. 247-3E)**. Spiral wave reentry and multiple circulating reentry wavefronts are possible mechanisms. Sustained polymorphic or monomorphic VT that degenerates to VF is a common cause of out of hospital cardiac arrest. Treatment follows ACLS guidelines with defibrillation to restore sinus rhythm. If resuscitation is successful, further evaluation is performed to identify and treat underlying heart disease and potential causes of the arrhythmia, including the possibility that monomorphic or polymorphic VT could have initiated VF. If a transient reversible cause such as acute MI is not identified, therapy to reduce the risk of sudden death with an ICD is often warranted. Chronic amiodarone therapy may be considered for individuals who are not ICD candidates.

FURTHER READING

Adler A, Gollob M: A practical guide to early repolarization. Curr Opin Cardiol 30:8, 2015. [PubMed: 25389652]

Priori SG et al: HRS/EHRA/APHRS Expert consensus statement on the diagnosis and management of patients with inherited primary arrhythmia syndromes. Heart Rhythm 10:1932, 2013. [PubMed: 24011539]

Chapter 251: Electrical Storm and Incessant VT

Roy M. John; William G. Stevenson

FIGURE 251-1

ELECTRICAL STORM

Electrical storm or ventricular tachycardia (VT) storm refers to the occurrence of three or more episodes of VT or ventricular fibrillation (VF) within 24 h. This severity of electrical instability is associated with a high mortality and requires prompt therapeutic intervention. Electrical storms occur in 4% of patients with a primary prevention implantable cardioverter defibrillator (ICD) but in as many as 20% of patients with a history of known VT or resuscitated sudden death.

MANAGEMENT OF THE PATIENT WITH ELECTRICAL STORM

Patients should be adequately sedated to allay anxiety. Recurrent VT/VF is treated using standard advanced cardiac life support guidelines and include the use of medications such as beta blockers, amiodarone, lidocaine with correction of any metabolic abnormalities. Recordings from ECG monitoring or an implanted ICD are important to assess whether VT is monomorphic or polymorphic that suggest possible precipitating or aggravating factors. Ischemia should be considered especially if polymorphic VT or VF is identified as the primary arrhythmia. If QT prolongation causing *torsades des pointes* is possible intravenous magnesium should be administered and bradycardia treated. If the QT interval is not prolonged and Brugada syndrome is possible, administration of quinidine and/or isoproterenol may abolish recurrent polymorphic VT/VF episodes. If the above measures fail, general anesthesia should be considered for suppression of recurrent hemodynamically unstable ventricular arrhythmia. Left stellate ganglion block and upper thoracic epidural anesthesia may reduce cardiac sympathetic outflow and have been used to restore stability in some patients. Catheter ablation of PVCs that are observed to repeatedly initiate the arrhythmia can be effective. Rarely, mechanical ventricular support or transplantation may have to be considered.

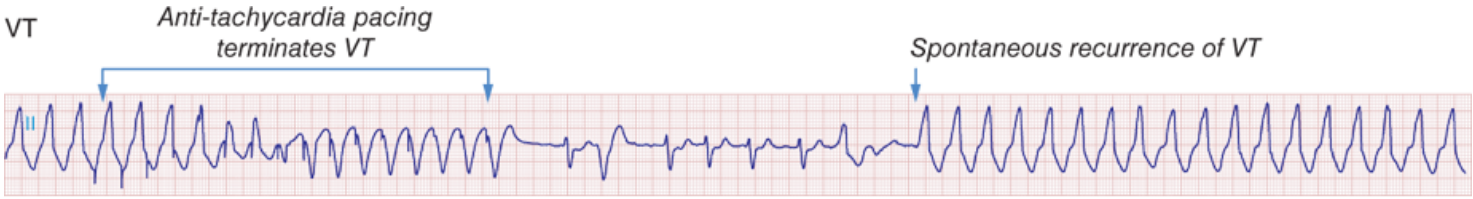
Once the acute episode is controlled, strategies to prevent recurrent VT or VF should be considered (see below).

INCESSANT VT

VT is designated incessant when VT continues to recur shortly after electrical, pharmacologic, or spontaneous conversion to sinus rhythm (**Fig. 251-1**). Typically, VT is monomorphic. Rarely, a slow incessant monomorphic VT will fail detection by the ICD because it falls outside of the programmed detection parameters. If the arrhythmia is hemodynamically stable acutely, patients can present with symptoms of gradual cardiac decompensation. VT may become incessant due to the pro-arrhythmic effect of an antiarrhythmic drugs such as amiodarone or a sodium channel blocker such as flecainide. Hemodynamic support may be required until the precipitating factors can be corrected. Urgent catheter ablation is often warranted.

FIGURE 251-1

Example of incessant monomorphic VT. In the initial portion of this ECG tracing, monomorphic VT is present. A train of antitachycardia pacing (*area bracketed by arrows*) that is initiated at the 4th VT complex results in ventricular capture with fusion by the 8th beat and termination of VT at cessation of pacing. The patient has underlying atrial fibrillation. Multifocal PVCs are present. VT similar in morphology to the initial VT restarts spontaneously toward the latter part of the trace (*arrow*). VT, ventricular tachycardia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MANAGEMENT OF PATIENTS PRESENTING WITH ICD SHOCKS

A substantial number of patients who receive an ICD can be expected to have an arrhythmia that is terminated by the ICD, either by a shock or antitachycardia pacing. Although this is an expected event, it can be a sign of impending instability, deterioration of cardiac function, emergence of a new arrhythmia or ICD malfunction, and therefore requires evaluation. Interrogation of the ICD is crucial after a patient reports a shock or symptoms of arrhythmia to confirm that the therapy was indeed delivered for a ventricular arrhythmia and not for lead malfunction or an atrial arrhythmia. After a single shock or two successive shocks occurring within a few seconds, and in the absence of other symptoms to suggest arrhythmia or ischemia, patients have the option of waiting until the next working day or using remote monitoring to transmit device interrogation data to their physician. However, occurrence of multiple ICD shocks constitutes a medical emergency and warrants immediate medical attention, usually by summoning emergency medical responders.

Spontaneous arrhythmias, particularly those that are converted with a shock, are associated with a subsequent increased risk of death and hospitalization in patients with depressed ventricular function. The occurrence of an arrhythmia, therefore, warrants a re-evaluation for possible decline in cardiac function, emergence of ischemia or inter-current illness.

If the ICD therapy is appropriate for VT or VF, consideration is given to whether therapy is warranted to reduce further episodes with either antiarrhythmic drug therapy or catheter ablation. Patients who have a rare episode of VT that is appropriately terminated and who have no other evidence of instability may not need any additional therapy, particularly if the VT is terminated by antitachycardia pacing rather than a shock (**see Fig. 247-6**). Shocks reduce quality of life and can lead to posttraumatic stress disorder. In many patients the possibility of a shock can be reduced with appropriate ICD programming. Studies have shown that antitachycardia pacing effectively terminates >70% of VT episodes, even when VT is very rapid. Most ICDs can be programmed to attempt overdrive pace-termination during capacitor charge. If the arrhythmia then terminates, the shock is aborted. Appropriate programming of antitachycardia pacing is therefore critical for reducing shocks. For patients implanted with ICDs as primary prevention, programming of VF detection zones >220 beats/min significantly reduce unnecessary and inappropriate shocks. Long detection times will also help avoid unnecessary therapies for VT episodes liable to terminate spontaneously.

Recurrent symptomatic episodes of VT or VF warrant specific therapy with antiarrhythmic drugs or ablation as discussed for the specific arrhythmia in **Chaps. 247**, **248**, **249**, **250**. Beta blockers sotalol and amiodarone are the most common pharmacological options. Amiodarone combined with beta blockers is more effective than sotalol or beta blockers alone. It is important to recognize that although VT/VF episodes may represent a deterioration of clinical status in these patients, interventions to control the arrhythmia itself may have adverse effects on outcome. Most antiarrhythmic drugs have the potential to induce bradycardia to the point of requiring pacing from the ICD that in itself, may have deleterious effects on ventricular function. Catheter ablation is an important option for patients with monomorphic VT.

FURTHER READING

Deneke T et al: Catheter ablation of electrical storm. Expert Rev Cardiovasc Ther 9:1051, 2011. [PubMed: 21878049]

John RM, Stevenson WG: Ventricular arrhythmias in patients with implanted cardioverter defibrillators. Trends Cardiovasc Med 22:169, 2012. [PubMed: 22902094]

Chapter 252: Heart Failure: Pathophysiology and Diagnosis

Douglas L. Mann; Murali Chakinala

HEART FAILURE

DEFINITION

Despite repeated attempts to develop a mechanistic definition that encompasses the heterogeneity and complexity of heart failure (HF), no single conceptual paradigm has withstood the test of time. The current American College of Cardiology Foundation (ACCF)/American Heart Association (AHA) guidelines define HF as a complex clinical syndrome that results from structural or functional impairment of ventricular filling or ejection of blood, which in turn leads to the cardinal clinical symptoms of dyspnea and fatigue and signs of HF, namely edema and rales. Because many patients present without signs or symptoms of volume overload, the term “heart failure” is preferred over the older term “congestive heart failure.”

EPIDEMIOLOGY


HF is a burgeoning problem worldwide, with >20 million people affected. The overall prevalence of HF in the adult population in developed countries is 2%. HF prevalence follows an exponential pattern, rising with age, and affects 6–10% of people aged >65. Although the relative incidence of HF is lower in women than in men, women constitute at least one-half the cases of HF because of their longer life expectancy. In North America and Europe, the lifetime risk of developing HF is approximately one in five for a 40-year-old. The overall prevalence of HF is thought to be increasing, in part because current therapies for cardiac disorders, such as myocardial infarction (MI), valvular heart disease, and arrhythmias, are allowing patients to survive longer. The prevalence of HF in emerging nations is uncertain because of the lack of population-based studies in those countries. HF was once thought to arise primarily in the setting of a depressed left ventricular (LV) ejection fraction (EF); however, epidemiologic studies have shown that approximately one-half of patients who develop HF have a normal or preserved EF (EF ≥50%). Accordingly, the historical terms “systolic” and “diastolic” HF have been abandoned, and HF patients are now broadly categorized into HF with a reduced EF (HFrEF; formerly *systolic failure*) or HF with a preserved EF (HRpEF; formerly *diastolic failure*). Patients with a LV EF between 40 and 50% have been considered as having a borderline or mid-range EF. At the time of this writing, the epidemiology of these patients is unclear.

ETIOLOGY

As shown in Table 252-1, any condition that leads to an alteration in LV structure or function can predispose a patient to developing HF. Although the etiology of HF in patients with a preserved EF differs from that of patients with depressed EF, there is considerable overlap between the etiologies of these two conditions. In industrialized countries, coronary artery disease (CAD) has become the predominant cause in men and women and is responsible for 60–75% of cases of HF. Hypertension contributes to the development of HF in 75% of patients, including most patients with CAD. Both CAD and hypertension interact to augment the risk of HF, as does diabetes mellitus.

TABLE 252-1

Etiologies of Heart Failure

Depressed Ejection Fraction (<40%)	
Coronary artery disease Myocardial infarction[a] Myocardial ischemia[a] Chronic pressure overload Hypertension[a] Obstructive valvular disease[a] Chronic volume overload Regurgitant valvular disease Intracardiac (left-to-right) shunting Extracardiac shunting Chronic lung disease Cor pulmonale Pulmonary vascular disorders	Nonischemic dilated cardiomyopathy Familial/genetic disorders Infiltrative disorders[a] Toxic/drug-induced damage Metabolic disorder[a] Viral Chagas' disease Disorders of rate and rhythm Chronic bradyarrhythmias Chronic tachyarrhythmias
Preserved Ejection Fraction (>40–50%)	
Pathologic hypertrophy Primary (hypertrophic cardiomyopathies) Secondary (hypertension) Aging Endomyocardial disorders	Restrictive cardiomyopathy Infiltrative disorders (amyloidosis, sarcoidosis) Storage diseases (hemochromatosis) Fibrosis
High-Output States	
Metabolic disorders Thyrotoxicosis Nutritional disorders (beriberi)	Excessive blood flow requirements Systemic arteriovenous shunting Chronic anemia

[a]Indicates conditions that can also lead to heart failure with a preserved ejection fraction.

In 20–30% of the cases of HF with a depressed EF, the exact etiologic basis is not known. These patients are referred to as having nonischemic, dilated, or idiopathic cardiomyopathy if the cause is unknown **(Chap. 254)**. Prior viral infection or toxin exposure (e.g., alcoholic or chemotherapeutic) also may lead to a dilated cardiomyopathy. Moreover, it is becoming increasingly clear that a large number of cases of dilated

cardiomyopathy are secondary to specific genetic defects, most notably those in the cytoskeleton. Most forms of familial dilated cardiomyopathy are inherited in an autosomal dominant fashion. Mutations of genes that encode cytoskeletal proteins (desmin, cardiac myosin, vinculin) and nuclear membrane proteins (laminin) have been identified thus far. Dilated cardiomyopathy also is associated with Duchenne's, Becker's, and limb-girdle muscular dystrophies. Conditions that lead to a high cardiac output (e.g., arteriovenous fistula, anemia) are seldom responsible for the development of HF in a normal heart; however, in the presence of underlying structural heart disease, these conditions can lead to overt HF.

GLOBAL CONSIDERATIONS


Rheumatic heart disease remains a major cause of HF in Africa and Asia, especially in the young. Hypertension is an important cause of HF in the African and African-American populations. Chagas' disease is still a major cause of HF in South America. Not surprisingly, anemia is a frequent concomitant factor in HF in many developing nations. As developing nations undergo socioeconomic development, the epidemiology of HF is becoming similar to that of Western Europe and North America, with CAD emerging as the single most common cause of HF. Although the contribution of diabetes mellitus to HF is not well understood, diabetes accelerates atherosclerosis and often is associated with hypertension.

PROGNOSIS

Despite recent advances in the management of HF, the development of symptomatic HF still carries a poor prognosis. Community-based studies indicate that 30–40% of patients die within 1 year of diagnosis and 60–70% die within 5 years, mainly from worsening HF or as a sudden event (probably because of a ventricular arrhythmia). Although it is difficult to predict prognosis in an individual, patients with symptoms at rest (New York Heart Association [NYHA] class IV) have a 30–70% annual mortality rate, whereas patients with symptoms with moderate activity (NYHA class II) have an annual mortality rate of 5–10%. Thus, functional status is an important predictor of patient outcome (**Table 252-2**).

TABLE 252-2

New York Heart Association Classification

Functional Capacity	Objective Assessment
Class I	Patients with cardiac disease but without resulting limitation of physical activity. Ordinary physical activity does not cause undue fatigue, palpitations, dyspnea, or anginal pain.
Class II	Patients with cardiac disease resulting in slight limitation of physical activity. They are comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnea, or anginal pain.
Class III	Patients with cardiac disease resulting in marked limitation of physical activity. They are comfortable at rest. Less than ordinary activity causes fatigue, palpitation, dyspnea, or anginal pain.
Class IV	Patients with cardiac disease resulting in inability to carry on any physical activity without discomfort. Symptoms of heart failure or the anginal syndrome may be present even at rest. If any physical activity is undertaken, discomfort is increased.

Source: Adapted from New York Heart Association, Inc., *Diseases of the Heart and Blood Vessels: Nomenclature and Criteria for Diagnosis,* 6th ed. Boston, Little Brown, 1964, p. 114.

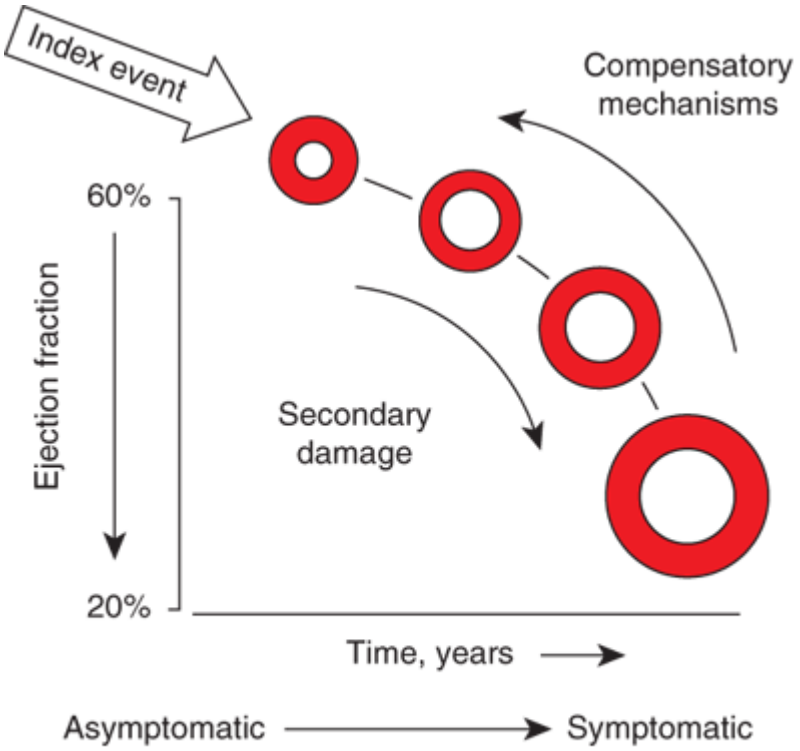
PATHOGENESIS

Figure 252-1 provides a conceptual framework for considering the development and progression of HFrEF. HF is a progressive disorder that is initiated after an *index event* either damages the heart muscle, with a resultant loss of functioning cardiac myocytes, or, alternatively, disrupts the ability of the myocardium to generate force, thereby preventing the heart from contracting normally. This index event may have an abrupt onset, as in the case of an MI; it may have a gradual or insidious onset, as in the case of hemodynamic pressure or volume overloading; or it may be hereditary, as in the case of many of the genetic cardiomyopathies. Regardless of the nature of the inciting event, the feature that is common to each of these index events is that they all in some manner produce a decline in the pumping capacity of the heart. In most instances, patients remain asymptomatic or minimally symptomatic after the initial decline in pumping capacity of the heart or develop symptoms only after the dysfunction has been present for some time.

FIGURE 252-1

Pathogenesis of heart failure with a depressed ejection fraction. Heart failure begins after an index event produces an initial decline in the heart's pumping capacity. After this initial decline in pumping capacity, a

variety of compensatory mechanisms are activated, including the adrenergic nervous system, the renin-angiotensin-aldosterone system, and the cytokine system. In the short term, these systems are able to restore cardiovascular function to a normal homeostatic range with the result that the patient remains asymptomatic. However, sustained activation of these systems leads to secondary end-organ damage within the ventricle, with worsening left ventricular remodeling and subsequent cardiac decompensation. *(From D Mann: Circulation 100:999, 1999.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

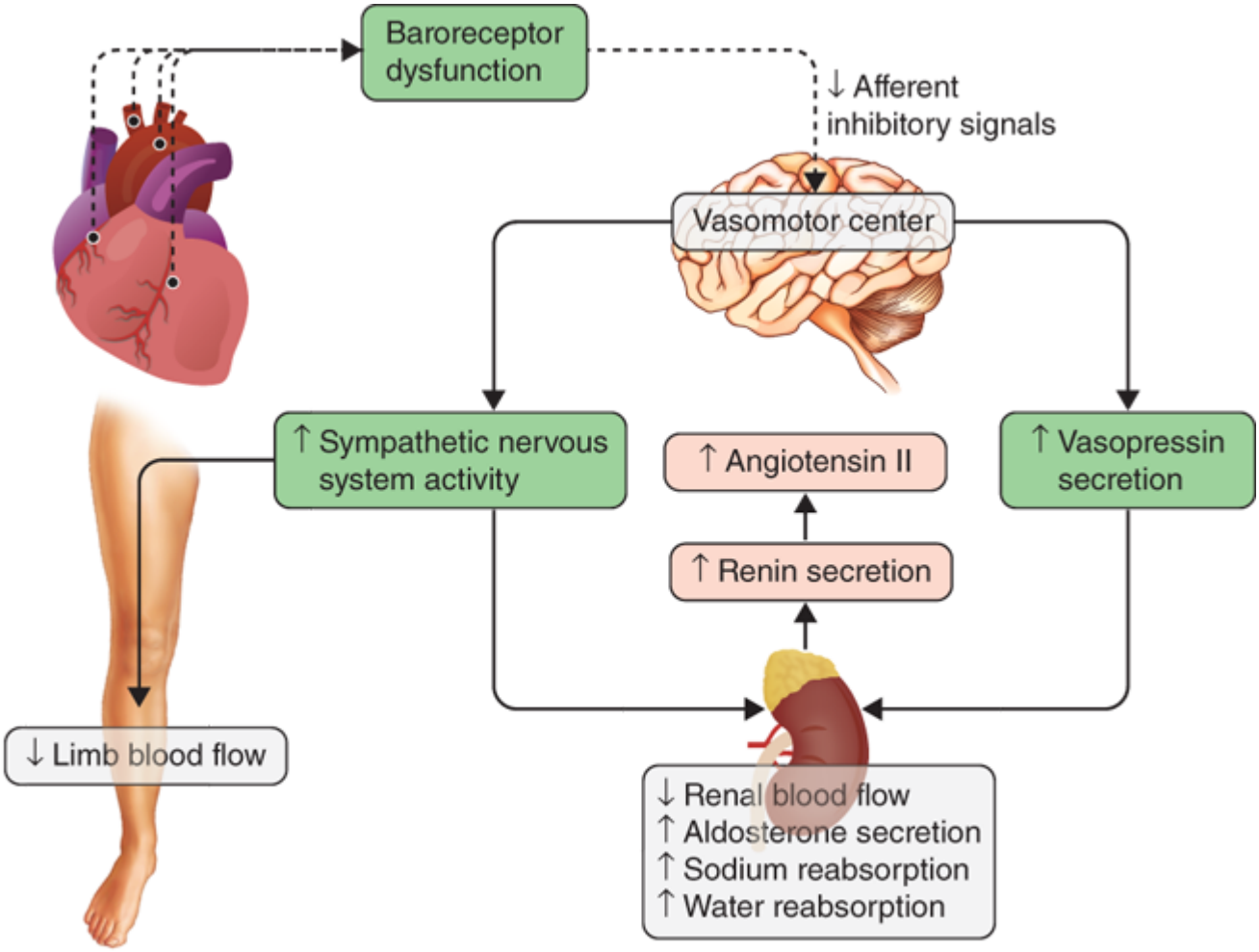
Although the precise reasons why patients with LV dysfunction may remain asymptomatic is not certain, one potential explanation is that a number of compensatory mechanisms become activated in the presence of cardiac injury and/or LV dysfunction allowing patients to sustain and modulate LV function for a period of months to years. The compensatory mechanisms that have been described thus far include (1) activation of the renin-angiotensin-aldosterone system (RAAS) and the adrenergic nervous system, which are responsible, respectively, for maintaining cardiac output through increased retention of salt and water **(Fig. 252-2)** and (2) increased myocardial contractility. In addition, a family of countervailing vasodilatory molecules are activated, including the atrial and brain natriuretic peptides (ANP and BNP), bradykinin, prostaglandins (PGE_2 and PGI_2), and nitric oxide (NO), that offset the excessive peripheral vascular vasoconstriction. Many of these vasodilatory peptides, including bradykinin and natriuretic peptides, are degraded by a neprilysin, which is a membrane-bound peptidase. These compensatory mechanisms are able to modulate LV function within a physiologic/homeostatic range so that the functional capacity of the patient is preserved or is minimally depressed. Thus, patients may remain asymptomatic or minimally symptomatic for a period of years; however, at some point patients become overtly symptomatic, with a resultant striking increase in morbidity and mortality rates. Although the exact mechanisms that are responsible for this transition are not known, as will be discussed below, the transition to symptomatic HF is accompanied by increasing activation

of neurohormonal, adrenergic, and cytokine systems that lead to a series of adaptive changes within the myocardium collectively referred to as *LV remodeling*.

FIGURE 252-2

Activation of neurohormonal systems in heart failure. The decreased cardiac output in heart failure (HF) patients results in an "unloading" of high-pressure baroreceptors (*circles*) in the left ventricle, carotid sinus, and aortic arch. This unloading of the peripheral baroreceptors leads to a loss of inhibitory parasympathetic tone to the central nervous system (CNS), with a resultant generalized increase in efferent sympathetic tone, and nonosmotic release of arginine vasopressin (AVP) from the pituitary. AVP (or antidiuretic hormone [ADH]) is a powerful vasoconstrictor that increases the permeability of the renal collecting ducts, leading to the reabsorption of free water. These afferent signals to the CNS also activate efferent sympathetic nervous system pathways that innervate the heart, kidney, peripheral vasculature, and skeletal muscles.

Sympathetic stimulation of the kidney leads to the release of renin, with a resultant increase in the circulating levels of angiotensin II and aldosterone. The activation of the renin-angiotensin-aldosterone system promotes salt and water retention and leads to vasoconstriction of the peripheral vasculature, myocyte hypertrophy, myocyte cell death, and myocardial fibrosis. Although these neurohormonal mechanisms facilitate short-term adaptation by maintaining blood pressure, these same neurohormonal mechanisms result in end-organ changes in the heart and the circulation, as well as to the excessive salt and water retention in advanced HF. *(Modified from A Nohria et al: Atlas of Heart Failure: Cardiac Function and Dysfunction, 4th ed, WS Colucci [ed]. Philadelphia, Current Medicine Group, 2002, p. 104 and J Hartupee, DL Mann: Nat Rev Cardiol 14:30, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In contrast to our understanding of the pathogenesis of HF with a depressed EF, our understanding of the mechanisms that contribute to the development of HF with a preserved EF is still evolving. That is, although diastolic dysfunction (see below) was thought to be the only mechanism responsible for the development of HF with a preserved EF, community-based studies suggest that additional extracardiac mechanisms may be important, such as increased vascular stiffness and impaired renal function.

BASIC MECHANISMS OF HF

HF with a Reduced Ejection Fraction

LV remodeling develops in response to a series of complex events that occur at the cellular and molecular levels **(Table 252-3)**. These changes include (1) myocyte hypertrophy; (2) alterations in the contractile properties of the myocyte; (3) progressive loss of myocytes through necrosis, apoptosis, and autophagic cell death; (4) β-adrenergic desensitization; (5) abnormal myocardial energetics and metabolism; and (6) reorganization of the extracellular matrix with dissolution of the organized structural collagen weave surrounding myocytes and subsequent replacement by an interstitial collagen matrix that does not provide structural support to the myocytes. The biologic stimuli for these profound changes include mechanical stretch of the myocyte, circulating neurohormones (e.g., norepinephrine, angiotensin II), inflammatory

cytokines (e.g., tumor necrosis factor [TNF]), other peptides and growth factors (e.g., endothelin), and reactive oxygen species (e.g., superoxide). The sustained overexpression of these biologically active molecules contributes to the progression of HF by virtue of the deleterious effects they exert on the heart and the circulation. Indeed, this insight forms the clinical rationale for using pharmacologic agents that antagonize these systems (e.g., angiotensin-converting enzyme [ACE] inhibitors, angiotensin receptor-neprilysin inhibitors [ARNIs] and beta blockers) in treating patients with HF **(Chap. 253)**.

TABLE 252-3

Overview of Left Ventricular Remodeling

Alterations in Myocyte Biology
Excitation-contraction coupling Myosin heavy chain (fetal) gene expression β-Adrenergic desensitization Hypertrophy Myocytolysis Cytoskeletal proteins
Myocardial Changes
Myocyte loss   Necrosis   Apoptosis   Autophagy Alterations in extracellular matrix   Matrix degradation   Myocardial fibrosis
Alterations in Left Ventricular Chamber Geometry
Left ventricular (LV) dilation Increased LV sphericity LV wall thinning Mitral valve incompetence

Source: Adapted from D. Mann: Pathophysiology of heart failure, in *Braunwald's Heart Disease,* 8th ed, PL Libby et al (eds). Philadelphia, Elsevier, 2008, p. 550.

To understand how the changes that occur in the failing cardiac myocyte contribute to depressed LV systolic function in HF, it is instructive first to review the biology of the cardiac muscle cell **(Chap. 232)**. Sustained neurohormonal activation and mechanical overload result in transcriptional and posttranscriptional changes in the genes and proteins that regulate excitation-contraction coupling and cross-bridge interaction **(see Figs. 232-6 and 232-7)**. The changes that regulate excitation-contraction include decreased function of sarcoplasmic reticulum (SR) Ca^{2+} adenosine triphosphatase (SERCA2A), resulting in decreased calcium uptake into the SR, and hyperphosphorylation of the ryanodine receptor, leading to calcium leakage from the SR. The changes that occur in the cross-bridges include decreased expression of α-myosin heavy chain and increased expression of β-myosin heavy chain, myocytolysis, and disruption of the cytoskeletal links between the sarcomeres and the extracellular matrix. Collectively, these changes impair the ability of the myocyte to contract and therefore contribute to the depressed LV systolic function observed in patients with HF.

Myocardial relaxation is an adenosine triphosphate (ATP)-dependent process that is regulated by uptake of cytoplasmic calcium into the SR by SERCA2A and extrusion of calcium by sarcolemmal pumps **(see Fig. 232-7)**. Accordingly, reductions in ATP concentration, as occurs in ischemia, may interfere with these processes and lead to slowed myocardial relaxation. Alternatively, if LV filling is delayed because LV compliance is reduced (e.g., from hypertrophy or fibrosis), LV filling pressures will similarly remain elevated at end diastole **(see Fig. 232-11)**. An increase in heart rate disproportionately shortens the time for diastolic filling, which may lead to elevated LV filling pressures, particularly in noncompliant ventricles. Elevated LV end-diastolic filling pressures result in increases in pulmonary capillary pressures, which can contribute to the dyspnea experienced by patients with diastolic dysfunction. In addition to impaired myocardial relaxation, increased myocardial stiffness secondary to cardiac hypertrophy and increased myocardial collagen content may contribute to diastolic failure. Importantly, diastolic dysfunction can occur alone or in combination with systolic dysfunction in patients with HF.

Left Ventricular Remodeling

Ventricular remodeling refers to the changes in LV mass, volume, and shape and the composition of the heart that occur after cardiac injury and/or abnormal hemodynamic loading conditions. LV remodeling contributes to the progression of HF by virtue of the mechanical burdens that are engendered by the changes in the geometry of the remodeled LV. In addition to the increase in LV end-diastolic volume, LV wall thinning occurs as the left ventricle begins to dilate. The increase in wall thinning, along with the increase in afterload created by LV dilation, leads to a functional *afterload mismatch* that may contribute further to a decrease in stroke volume. Moreover, the high end-diastolic wall stress might be expected to lead to (1) hypoperfusion of the subendocardium, with resultant worsening of LV function; (2) increased oxidative stress, with the resultant activation of families of genes that are sensitive to free radical generation (e.g., TNF and interleukin 1β); and (3) sustained expression of stretch activation of hypertrophic signaling pathways. Increasing LV dilation also results in tethering of the papillary muscles with resulting incompetence of the mitral valve apparatus and functional mitral regurgitation, which in turn leads to further hemodynamic overloading of the ventricle. Taken together, the mechanical burdens that are engendered by LV remodeling contribute to the progression

of HF. Recent studies have shown that LV remodeling can be reversed following medical and device therapy and that reverse LV remodeling is associated with improved clinical outcomes in patients with HFrEF. Indeed, one of the goals of therapy for HF is to prevent and/or reverse LV remodeling.

CLINICAL MANIFESTATIONS

Symptoms

The cardinal symptoms of HF are fatigue and shortness of breath. Although fatigue traditionally has been ascribed to the low cardiac output in HF, it is likely that skeletal-muscle abnormalities and other noncardiac comorbidities (e.g., anemia) also contribute to this symptom. In the early stages of HF, dyspnea is observed only during exertion; however, as the disease progresses, dyspnea occurs with less strenuous activity, and it ultimately may occur even at rest. The origin of dyspnea in HF is probably multifactorial **(Chap. 33)**. The most important mechanism is pulmonary congestion with accumulation of interstitial or intra-alveolar fluid, which activates juxtacapillary J receptors, which in turn stimulate the rapid, shallow breathing characteristic of cardiac dyspnea. Other factors that contribute to dyspnea on exertion include reductions in pulmonary compliance, increased airway resistance, respiratory muscle and/or diaphragm fatigue, and anemia. Dyspnea may become less frequent with the onset of right ventricular (RV) failure and tricuspid regurgitation.

ORTHOPNEA

Orthopnea, which is defined as dyspnea occurring in the recumbent position, is usually a later manifestation of HF than is exertional dyspnea. It results from redistribution of fluid from the splanchnic circulation and lower extremities into the central circulation during recumbency, with a resultant increase in pulmonary capillary pressure. Nocturnal cough is a common manifestation of this process and a frequently overlooked symptom of HF. Orthopnea generally is relieved by sitting upright or sleeping with additional pillows. Although orthopnea is a relatively specific symptom of HF, it may occur in patients with abdominal obesity or ascites and patients with pulmonary disease whose lung mechanics favor an upright posture.

PAROXYSMAL NOCTURNAL DYSPNEA (PND)

This term refers to acute episodes of severe shortness of breath and coughing that generally occur at night and awaken the patient from sleep, usually 1–3 h after the patient retires. PND may manifest as coughing or wheezing, possibly because of increased pressure in the bronchial arteries leading to airway compression, along with interstitial pulmonary edema that leads to increased airway resistance. Whereas orthopnea may be relieved by sitting upright at the side of the bed with the legs in a dependent position, patients with PND often have persistent coughing and wheezing even after they have assumed the upright position. *Cardiac asthma* is closely related to PND, is characterized by wheezing secondary to bronchospasm, and must be differentiated from primary asthma and pulmonary causes of wheezing.

CHEYNE-STOKES RESPIRATION

Also referred to as periodic respiration or cyclic respiration, Cheyne-Stokes respiration is present in 40% of patients with advanced HF and usually is associated with low cardiac output. Cheyne-Stokes respiration is caused by an increased sensitivity of the respiratory center to arterial P_{CO_2} and a lengthy circulatory time. There is an apneic phase, during which arterial P_{O_2} falls and arterial P_{CO_2} rises. These changes in the arterial

blood gas content stimulate the respiratory center, resulting in hyperventilation and hypocapnia, followed by recurrence of apnea.

ACUTE PULMONARY EDEMA

See Chap. 298.

Other Symptoms

Patients with HF also may present with gastrointestinal symptoms. Anorexia, nausea, and early satiety associated with abdominal pain and fullness are common complaints and may be related to edema of the bowel wall and/or a congested liver. Congestion of the liver and stretching of its capsule may lead to right upper-quadrant pain. Cerebral symptoms such as confusion, disorientation, and sleep and mood disturbances may be observed in patients with severe HF, particularly elderly patients with cerebral arteriosclerosis and reduced cerebral perfusion. Nocturia is common in HF and may contribute to insomnia.

PHYSICAL EXAMINATION

A careful physical examination is always warranted in the evaluation of patients with HF, in order to determine the cause of HF, as well as to assess the severity of the syndrome.

General Appearance and Vital Signs

In mild or moderately severe HF, the patient appears to be in no distress at rest except for feeling uncomfortable when lying flat for more than a few minutes. In more severe HF, the patient must sit upright, may have labored breathing, and may not be able to finish a sentence because of shortness of breath. Systolic blood pressure may be normal or high in early HF, but it generally is reduced in advanced HF because of severe LV dysfunction. The pulse pressure may be diminished, reflecting a reduction in stroke volume. Sinus tachycardia is a nonspecific sign caused by increased adrenergic activity. Peripheral vasoconstriction leading to cool peripheral extremities and cyanosis of the lips and nail beds is also caused by excessive adrenergic activity.

Jugular Veins

Examination of the jugular veins provides an estimation of right atrial pressure **(See also Chap. 234)**. The jugular venous pressure is best appreciated with the patient lying recumbent, with the head tilted at 45°. The jugular venous pressure should be quantified in centimeters of water (normal ≤8 cm) by estimating the height of the venous column of blood above the sternal angle in centimeters and then adding 5 cm. In the early stages of HF, the venous pressure may be normal at rest but may become abnormally elevated with sustained (~15 s) pressure on the abdomen (positive abdominojugular reflux). Giant *v* waves indicate the presence of tricuspid regurgitation.

Pulmonary Examination

Pulmonary crackles (rales or crepitations) result from the transudation of fluid from the intravascular space into the alveoli. In patients with pulmonary edema, rales may be heard widely over both lung fields and may be accompanied by expiratory wheezing (cardiac asthma). When present in patients without concomitant lung disease, rales are specific for HF. Importantly, rales are frequently absent in patients with chronic HF, even when LV filling pressures are elevated, because of increased lymphatic drainage of alveolar fluid. Pleural effusions result from the elevation of pleural capillary pressure and the resulting transudation of fluid into the pleural cavities. Since the pleural veins drain into both the systemic and the pulmonary veins, pleural effusions occur most commonly with biventricular failure. Although pleural effusions are often bilateral in HF, when they are unilateral, they occur more frequently in the right pleural space.

Cardiac Examination

Examination of the heart, although essential, frequently does not provide useful information about the severity of HF. If cardiomegaly is present, the point of maximal impulse (PMI) usually is displaced below the fifth intercostal space and/or lateral to the midclavicular line, and the impulse is palpable over two interspaces. Severe LV hypertrophy leads to a sustained PMI. In some patients, a third heart sound (S_3) is audible and palpable at the apex. Patients with enlarged or hypertrophied right ventricles may have a sustained and prolonged left parasternal impulse extending throughout systole. An S_3 (or *protodiastolic gallop*) is most commonly present in patients with volume overload who have tachycardia and tachypnea, and it often signifies severe hemodynamic compromise. A fourth heart sound (S_4) is not a specific indicator of HF but is usually present in patients with diastolic dysfunction. The murmurs of mitral and tricuspid regurgitation are frequently present in patients with advanced HF.

Abdomen and Extremities

Hepatomegaly is an important sign in patients with HF. When it is present, the enlarged liver is frequently tender and may pulsate during systole if tricuspid regurgitation is present. Ascites, a late sign, occurs as a consequence of increased pressure in the hepatic veins and the veins draining the peritoneum. Jaundice, also a late finding in HF, results from impairment of hepatic function secondary to hepatic congestion and hepatocellular hypoxemia and is associated with elevations of both direct and indirect bilirubin.

Peripheral edema is a cardinal manifestation of HF, but it is nonspecific and usually is absent in patients who have been treated adequately with diuretics. Peripheral edema is usually symmetric and dependent in HF and occurs predominantly in the ankles and the pretibial region in ambulatory patients. In bedridden patients, edema may be found in the sacral area (*presacral edema*) and the scrotum. Long-standing edema may be associated with indurated and pigmented skin.

Cardiac Cachexia

With severe chronic HF, there may be marked weight loss and cachexia. Although the mechanism of cachexia is not entirely understood, it is probably multifactorial. When present, cachexia augurs a poor overall

prognosis.

DIAGNOSIS

The diagnosis of HF is relatively straightforward when the patient presents with classic signs and symptoms of HF; however, the signs and symptoms of HF are neither specific nor sensitive. Accordingly, the key to making the diagnosis is to have a high index of suspicion, particularly for high-risk patients. When these patients present with signs or symptoms of HF, additional laboratory testing should be performed.

Routine Laboratory Testing

Patients with new-onset HF and those with chronic HF and acute decompensation should have a complete blood count, a panel of electrolytes, blood urea nitrogen, serum creatinine, hepatic enzymes, and a urinalysis. Selected patients should have assessment for diabetes mellitus (fasting serum glucose or oral glucose tolerance test), dyslipidemia (fasting lipid panel), and thyroid abnormalities (thyroid-stimulating hormone level).

Electrocardiogram (ECG)

A routine 12-lead ECG is recommended. The major importance of the ECG is to assess cardiac rhythm and determine the presence of LV hypertrophy or a prior MI (presence or absence of Q-waves) as well as to determine QRS width to ascertain whether the patient may benefit from resynchronization therapy (see below). A normal ECG virtually excludes LV systolic dysfunction.

Chest X-Ray

A chest x-ray provides useful information about cardiac size and shape, as well as the state of the pulmonary vasculature, and may identify noncardiac causes of the patient's symptoms. Although patients with acute HF have evidence of pulmonary hypertension, interstitial edema, and/or pulmonary edema, the majority of patients with chronic HF do not. The absence of these findings in patients with chronic HF reflects the increased capacity of the lymphatics to remove interstitial and/or pulmonary fluid.

Assessment of LV Function

Noninvasive cardiac imaging **(Chap. 236)** is essential for the diagnosis, evaluation, and management of HF. The most useful test is the two-dimensional (2-D) echocardiogram/Doppler, which can provide a semiquantitative assessment of LV size and function as well as the presence or absence of valvular and/or regional wall motion abnormalities (indicative of a prior MI). The presence of left atrial dilation and LV hypertrophy, together with abnormalities of LV diastolic filling provided by pulse-wave and tissue Doppler, is useful for the assessment of HF with a preserved EF. The 2-D echocardiogram/Doppler is also invaluable in assessing RV size and pulmonary pressures, which are critical in the evaluation and management of cor pulmonale (see below). Magnetic resonance imaging (MRI) also provides a comprehensive analysis of cardiac anatomy and function and is now the gold standard for assessing LV mass and volumes. MRI also is emerging

as a useful and accurate imaging modality for evaluating patients with HF, both in terms of assessing LV structure and for determining the cause of HF (e.g., amyloidosis, ischemic cardiomyopathy, hemochromatosis).

The most useful index of LV function is the EF (stroke volume divided by end-diastolic volume). Because the EF is easy to measure by noninvasive testing and easy to conceptualize, it has gained wide acceptance among clinicians. Unfortunately, the EF has a number of limitations as a true measure of contractility, since it is influenced by alterations in afterload and/or preload. Nonetheless, with the exceptions indicated above, when the EF is normal (≥50%), systolic function is usually adequate, and when the EF is significantly depressed (<30–40%), contractility is usually depressed. Myocardial strain rate imaging using speckle tracking has been shown to add incremental value to standard measurements of LV EF and to have prognostic value.

Biomarkers

Circulating levels of natriuretic peptides are useful and important adjunctive tools in the diagnosis of patients with HF. Both B-type natriuretic peptide (BNP) and N-terminal pro-BNP (NT-proBNP), which are released from the failing heart, are relatively sensitive markers for the presence of HF with depressed EF; they also are elevated in HF patients with a preserved EF, albeit to a lesser degree. In ambulatory patients with dyspnea, the measurement of BNP or NT-proBNP is useful to support clinical decision-making regarding the diagnosis of HF, especially in the setting of clinical uncertainty. Moreover, the measurement of BNP or NT-proBNP is useful for establishing prognosis or disease severity in chronic HF and can be useful to achieve optimal dosing of medical therapy in select clinically euvolemic patients. However, it is important to recognize that natriuretic peptide levels increase with age and renal impairment, are more elevated in women, and can be elevated in right HF from any cause. BNP levels may increase in patients taking ARNIs. Levels can be falsely low in obese patients. Other biomarkers, such as soluble ST-2 and galectin-3, are newer biomarkers that can be used for determining the prognosis of HF patients.

Exercise Testing

Treadmill or bicycle exercise testing is not routinely advocated for patients with HF, but either is useful for assessing the need for cardiac transplantation in patients with advanced HF **(Chap. 255)**. A peak oxygen uptake (vo_2) <14 mL/kg per min is associated with a relatively poor prognosis. Patients with a vo_2 <14 mL/kg per min have been shown, in general, to have better survival when transplanted than when treated medically.

DIFFERENTIAL DIAGNOSIS

HF resembles but should be distinguished from (1) conditions in which there is circulatory congestion secondary to abnormal salt and water retention but in which there is no disturbance of cardiac structure or function (e.g., renal failure), and (2) noncardiac causes of pulmonary edema (e.g., acute respiratory distress syndrome). In most patients who present with classic signs and symptoms of HF, the diagnosis is relatively

straightforward. However, even experienced clinicians have difficulty differentiating the dyspnea that arises from cardiac and pulmonary causes (Chap. 33). In this regard, noninvasive cardiac imaging, biomarkers, pulmonary function testing, and chest x-ray may be useful. A very low BNP or NT-proBNP may be helpful in excluding a cardiac cause of dyspnea in this setting. Ankle edema may arise secondary to varicose veins, obesity, renal disease, or gravitational effects. When HF develops in patients with a preserved EF, it may be difficult to determine the relative contribution of HF to the dyspnea that occurs in chronic lung disease and/or obesity.

COR PULMONALE

DEFINITION

Cor pulmonale, also referred to as *pulmonary heart disease*, is broadly defined by altered RV structure and/or function in the context of chronic lung disease and is triggered by the presence of pulmonary hypertension. Although RV dysfunction is an important sequela of HFpEF and HFrEF, this is not considered as cor pulmonale.

ETIOLOGY AND EPIDEMIOLOGY

Chronic cor pulmonale develops in response to chronic pulmonary hypertension resulting from parenchymal lung disorders, primary pulmonary vascular diseases, or conditions leading to alveolar hypoxia (Table 252-4). The true prevalence of cor pulmonale is difficult to ascertain. First, not all patients with chronic lung disease will develop cor pulmonale, which may be subclinical in compensated individuals. Second, the ability to detect pulmonary hypertension and cor pulmonale by routine physical examination and laboratory testing is relatively insensitive. However, advances in 2-D echo/Doppler imaging and biomarkers (BNP) can make it easier to identify.

TABLE 252-4

Etiology of Chronic Cor Pulmonale

Diseases of the Lung Parenchyma
Chronic obstructive pulmonary disease Emphysema Chronic bronchitis Interstitial lung diseases Idiopathic interstitial pneumonias (e.g., IPF, UIP) Secondary interstitial diseases Sarcoidosis Combined pulmonary fibrosis and emphysema Bronchiectasis Cystic fibrosis Pulmonary Langerhans cell histiocytosis Lymphangioleiomyomatosis Developmental lung disorders
Disorders of Chronic (Alveolar) Hypoxia
Alveolar hypoventilation syndromes Obesity hypoventilation syndrome Central hypoventilation syndrome Neuromuscular respiratory failure Chest wall disorders Kyphoscoliosis Chronic exposure to high altitude
Diseases of the Pulmonary Vasculature
Pulmonary arterial hypertension (PAH) Idiopathic PAH Heritable PAH Drug and toxin-induced Associated PAH Venoocclusive disease Chronic thromboembolic pulmonary hypertension Pulmonary tumor thrombotic microangiopathy Mediastinal disorders affecting central pulmonary vasculature

Abbreviations: IPF, idiopathic pulmonary fibrosis; UIP, usual interstitial pneumonitis.

Once susceptible patients develop cor pulmonale, their prognosis worsens, regardless of the underlying etiology. Although chronic obstructive pulmonary disease (COPD) and chronic bronchitis are responsible for ~50% of the cases of cor pulmonale in North America **(Chap. 286)**, any disease that affects the pulmonary vasculature **(Chap. 277)** or parenchyma can lead to cor pulmonale (Table 252-4). Primary pulmonary vascular disorders, such as pulmonary arterial hypertension or chronic thromboembolic pulmonary hypertension, are relatively rare causes of cor pulmonale, but cor pulmonale is extremely common with these conditions, given the magnitude of elevated pulmonary artery pressures and pulmonary vascular resistance.

PATHOPHYSIOLOGY AND BASIC MECHANISMS

Although many conditions can lead to cor pulmonale, the common pathophysiologic mechanism is pulmonary hypertension and increased RV afterload sufficient to alter RV structure (i.e., dilation with or without hypertrophy) and function. Normally, mean pulmonary artery pressure is only ~15 mmHg and does not increase significantly even with increasing multiples of cardiac output, because of pulmonary vasodilation and blood vessel recruitment in the pulmonary circulatory bed. But, in the setting of parenchymal lung diseases, primary pulmonary vascular disorders, or chronic (alveolar) hypoxia, the circulatory bed undergoes vascular remodeling, vasoconstriction, and destruction. As a result, pulmonary artery pressures and RV afterload increases, setting the stage for cor pulmonale (Table 252-4). The systemic consequences of cor pulmonale relate to alterations in cardiac output as well as salt and water homeostasis. Anatomically, the RV is a thin-walled, compliant chamber better suited to handle volume overload than pressure overload. Thus, the sustained pressure overload eventually leads to RV dysfunction and failure.

The response of the RV to pulmonary hypertension depends on the acuteness and severity of the pressure overload. Acute cor pulmonale occurs after a sudden and severe stimulus (e.g., massive pulmonary embolus), with RV dilatation and failure but no RV hypertrophy **(Chap. 273)**. Chronic cor pulmonale, however, evolves slowly and in conjunction with modest, compensatory RV hypertrophy that lowers wall tension and preserves RV function. Over time, RV dilation ensues leading to an increase in RV wall tension and overt dysfunction. Acute decompensation of compensated chronic cor pulmonale is a common clinical occurrence. Triggers include worsening hypoxia from any cause (e.g., pneumonia), acidemia (e.g., exacerbation of COPD), acute pulmonary embolus, atrial tachyarrhythmia, hypervolemia, and mechanical ventilation that compresses blood vessels associated with alveoli and further increasing RV afterload.

CLINICAL MANIFESTATIONS

Symptoms

The symptoms in chronic cor pulmonale generally are related to the underlying pulmonary disorder. Dyspnea, the most common symptom, is usually the result of the increased work of breathing secondary to changes in elastic recoil of the lung (fibrosing lung diseases), altered respiratory mechanics (e.g., over-

inflation with COPD), or inefficient ventilation (e.g., primary pulmonary vascular disease). Dyspnea can also be due to cardiovascular limitations with decreased oxygen delivery due to reduced cardiac output. Lower-extremity edema and even increased abdominal girth due to ascites formation occurs secondary to neurohormonal activation, elevated RV filling and right atrial pressures, or increased levels of carbon dioxide and hypoxemia, which can lead to peripheral vasodilation and edema formation.

Signs

Auscultation of the heart reveals the findings of pulmonary hypertension **(Chap. 33)**, while auscultation of the lungs can highlight the underlying parenchymal lung disorder. In chronic cor pulmonale, the murmur of tricuspid regurgitation, an S_3 gallop and a RV heave palpable along the left sternal border can be appreciated. But the most blatant findings are reflective of high right-sided filling pressures and hypervolemia such as elevated jugular venous pressures with prominent *v* waves indicative of tricuspid regurgitation, hepatomegaly, pulsatile liver, ascites, and especially lower-extremity edema. Cyanosis is a late finding in cor pulmonale and is secondary to a low cardiac output (i.e., cardiogenic shock), systemic vasoconstriction, and hypoxemia.

DIAGNOSIS

It is important to evaluate the patient for LV systolic and diastolic dysfunction as a cause of right-sided HF. The ECG in severe pulmonary hypertension shows P pulmonale, right axis deviation, and RV hypertrophy. Radiographic examination of the chest may show enlargement of the main central pulmonary arteries and hilar vessels. Spirometry and lung volumes can identify obstructive and/or restrictive defects indicative of parenchymal lung diseases and reduced diffusing capacity; arterial blood gases typically reveal hypoxemia with or without hypercapnia. A high-resolution computed tomography (CT) scan of the chest can identify interstitial lung disease and the extent of emphysema. Chest CT angiogram is useful in diagnosing acute pulmonary emboli; however, the ventilation-perfusion scan remains best suited for diagnosing *chronic thromboembolic disease* **(Chap. 273)**.

Two-dimensional echocardiography is used to measure RV wall thickness and chamber dimensions. The interventricular septum may move paradoxically during systole in the presence of RV pressure overload, highlighting a deleterious interaction between the RV and the LV. Doppler echocardiography can be used to assess pulmonary artery pressures. The location of the RV behind the sternum and its crescent shape can challenge assessment of RV function by echocardiography, especially when parenchymal lung disease is present. Calculated measures of RV function (e.g., tricuspid annular plane systolic excursion [TAPSE], systolic velocity of the RV free wall, strain of the RV free wall, or the Tei Index) supplement more subjective assessments. MRI is also useful for assessing RV structure and function, particularly in patients who are difficult to image with 2-D echocardiography because of severe lung disease. Cardiac catheterization confirms the diagnosis of pulmonary hypertension and can exclude elevated left-sided pressures (measured as the pulmonary capillary wedge pressure or the LV end-diastolic pressure) as a cause for right-sided HF.

BNP and N-terminal BNP levels are elevated in patients with cor pulmonale secondary to RV myocardial stretch.

FURTHER READING

Braunwald E: Heart failure. JACC Heart Fail 1:1, 2013. [PubMed: 24621794]

Hartupee J, Mann DL: Neurohormonal activation in heart failure with reduced ejection fraction. Nat Rev Cardiol 14:30, 2017. [PubMed: 27708278]

Pengo V et al: Incidence of chronic thromboembolic pulmonary hypertension after pulmonary embolism. N Engl J Med 350:2257, 2004. [PubMed: 15163775]

Reddy YN, Borlaug BA: Heart failure with preserved ejection fraction. Curr Probl Cardiol 41:145, 2016. [PubMed: 26952248]

Chapter 253: Heart Failure: Management

Mandeep R. Mehra

INTRODUCTION

Distinctive phenotypes of presentation with diverse management targets exemplify the extensive syndrome of heart failure. These range from chronic heart failure with reduced ejection fraction (HFrEF) or heart failure with preserved ejection fraction (HFpEF), acute decompensated heart failure (ADHF), and advanced heart failure. Early management evolved from symptom control to disease-modifying therapy in HFrEF with the advent of renin-angiotensin-aldosterone system (RAAS)–directed therapy, beta receptor antagonists, mineralocorticoid receptor antagonists, cardiac resynchronization therapy, and implantable cardio-defibrillators. However, similar advances have been elusive in the syndromes of HFpEF and ADHF, which have remained devoid of convincing therapeutic advances to alter their natural history. In advanced heart failure, a stage of disease typically encountered in HFrEF, the patient remains markedly symptomatic with demonstrated refractoriness or inability to tolerate full-dose neurohormonal antagonism, often requires escalating doses of diuretics, and exhibits persistent hyponatremia and renal insufficiency with frequent episodes of heart failure decompensation requiring recurrent hospitalizations. Such individuals are at the highest risk of sudden or progressive pump failure–related deaths **(Chap. 255)**. In contrast, early-stage asymptomatic left ventricular dysfunction is amenable to preventive care, and its natural history is modifiable by neurohormonal antagonism (not further discussed).

HEART FAILURE WITH PRESERVED EJECTION FRACTION

GENERAL PRINCIPLES

Therapeutic targets in HFpEF include control of congestion, stabilization of heart rate and blood pressure, and efforts at improving exercise tolerance. Addressing surrogate targets, such as regression of ventricular hypertrophy in hypertensive heart disease, and use of lusitropic agents, such as calcium channel blockers and beta receptor antagonists, have been disappointing. Experience has demonstrated that lowering blood pressure alleviates symptoms more effectively than targeted therapy with specific agents.

CLINICAL TRIALS IN HFpEF

The Candesartan in Heart Failure—Assessment of Mortality and Morbidity (CHARM) Preserved study showed a statistically significant reduction in hospitalizations but no difference in all-cause mortality in patients with HFpEF who were treated with the angiotensin receptor blocker (ARB), candesartan. Similarly, the Irbesartan in Heart Failure with Preserved Systolic Function (I-PRESERVE) trial demonstrated no differences in meaningful endpoints in such patients treated with irbesartan. An earlier analysis of a subset of the Digitalis Investigation Group (DIG) trial found no role for digoxin in the treatment of HFpEF. In the Study of the Effects of Nebivolol Intervention on Outcomes and Rehospitalization in Seniors with Heart Failure (SENIORS) trial of nebivolol, a vasodilating beta blocker, the subgroup of elderly patients with prior hospitalization and HFpEF did not appear to benefit in terms of all-cause or cardiovascular mortality. Much smaller mechanistic studies in the elderly with the angiotensin-converting enzyme inhibitor (ACEI) enalapril showed no effect on peak exercise oxygen consumption, 6-min walk distance, aortic distensibility, left ventricular mass, or peripheral neurohormone expression.

NOVEL TARGETS

A small trial demonstrated that the phosphodiesterase-5 inhibitor *sildenafil* improved filling pressures and right ventricular function in a cohort of HFpEF patients with pulmonary venous hypertension. This finding led to the phase II trial, Phosphodiesterase-5 Inhibition to Improve Clinical Status and Exercise Capacity in Diastolic Heart Failure (RELAX), in HFpEF patients (left ventricular ejection fraction [LVEF] >50%) with New York Heart Association (NYHA) functional class II or III symptoms, who received sildenafil at 20 mg three times daily for 3 months, followed by 60 mg three times daily for another 3 months, compared with a placebo. There was no improvement in functional capacity, quality of life (QOL), or other clinical and surrogate parameters. Conceptually targeting myocardial fibrosis in HFpEF, the large-scale Aldosterone Antagonist Therapy in Adults with Preserved Ejection Fraction Congestive Heart Failure (TOPCAT) trial has been completed. This trial demonstrated no improvement in the primary composite endpoint, but did show a secondary signal of benefit on HF hospitalizations, counterbalanced, however, by an increase in adverse effects, particularly hyperkalemia. However, pessimism has been generated by the negative outcome of the Aldosterone Receptor Blockade in Diastolic Heart Failure (ALDO-DHF) study wherein *spironolactone* improved echocardiographic indices of diastolic dysfunction but failed to improve exercise capacity, symptoms, or QOL measures. On the premise that nitrates, which are nitric oxide donors, might improve preload, coronary perfusion, endothelial function and improved exercise tolerance, the Nitrate's Effect on Activity Tolerance in Heart Failure with Preserved Ejection Fraction (NEAT-HFpEF) study was conducted. Isosorbide mononitrate did not improve QOL or sub-maximal exercise capacity, and decreased overall activity levels in treated patients. A unique molecule that hybridizes an *ARB with an*

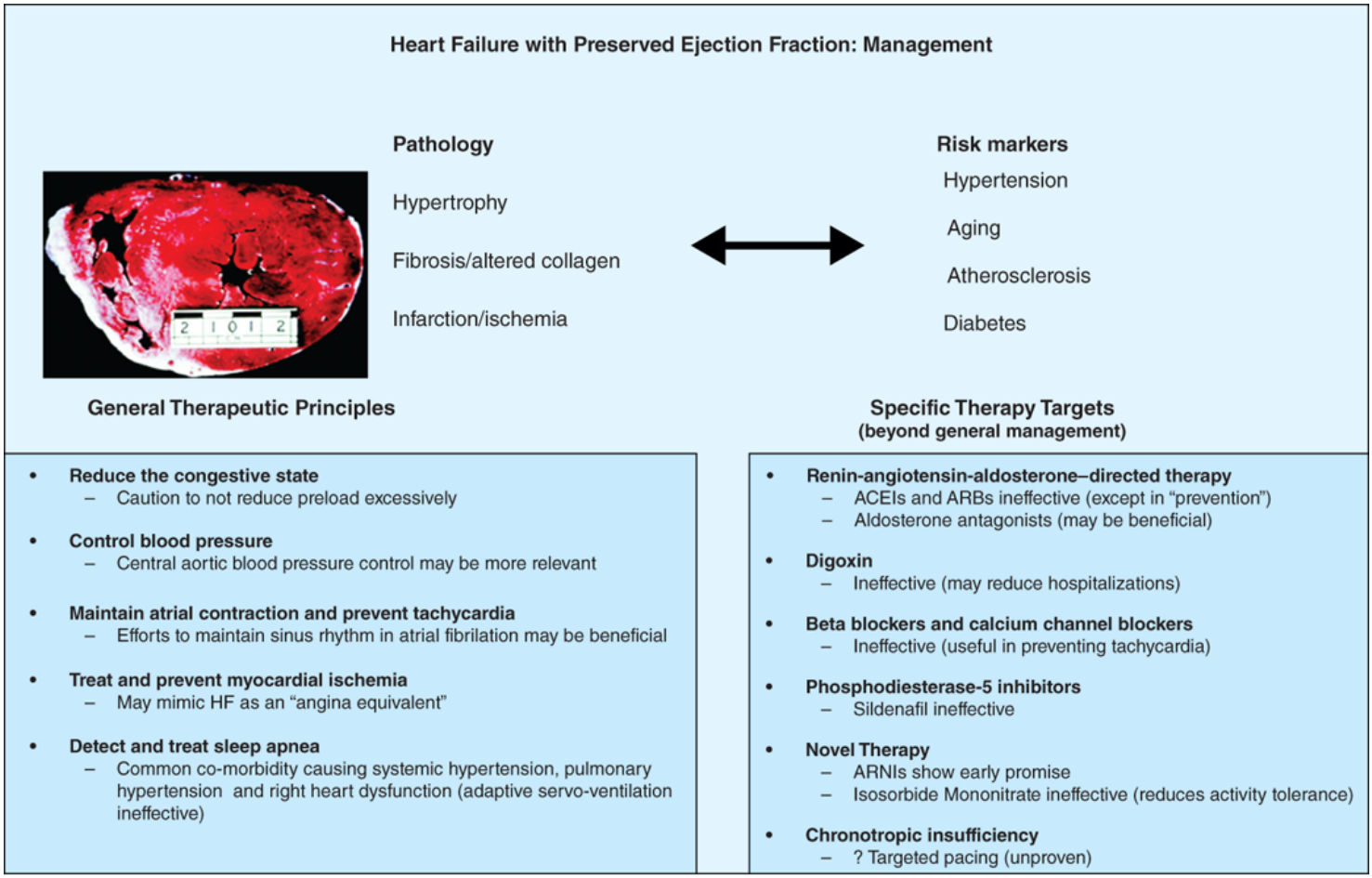
endopeptidase inhibitor, LCZ696, increases the generation of myocardial cyclic guanosine 3′,5′-monophosphate, enhances myocardial relaxation, and reduces ventricular hypertrophy. This dual blocker has been shown to reduce circulating natriuretic peptides and reduce left atrial size to a significantly greater extent than valsartan alone in patients with HFpEF. This molecule is currently being tested in a pivotal clinical trial (PARAGON-HF).

CLINICAL PEARLS

Even as efforts to control hypertension in HFpEF are critical, evaluation for and correction of underlying ischemia may be beneficial. Appropriate identification and treatment of sleep-disordered breathing should be strongly considered. Excessive decrease in preload with vasodilators may lead to underfilling the ventricle and subsequent hypotension and syncope. Some investigators have suggested that the exercise intolerance in HFpEF is a manifestation of chronotropic insufficiency and that such aberrations could be corrected with use of rate responsive pacemakers, but this remains an inadequately investigated contention (Fig. 253-1).

FIGURE 253-1

Pathophysiologic correlations, general therapeutic principles, and results of specific "directed" therapy in heart failure (HF) with preserved ejection fraction. ACEI, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; ARNI, angiotensin receptor neprilysin inhibitor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ACUTE DECOMPENSATED HEART FAILURE

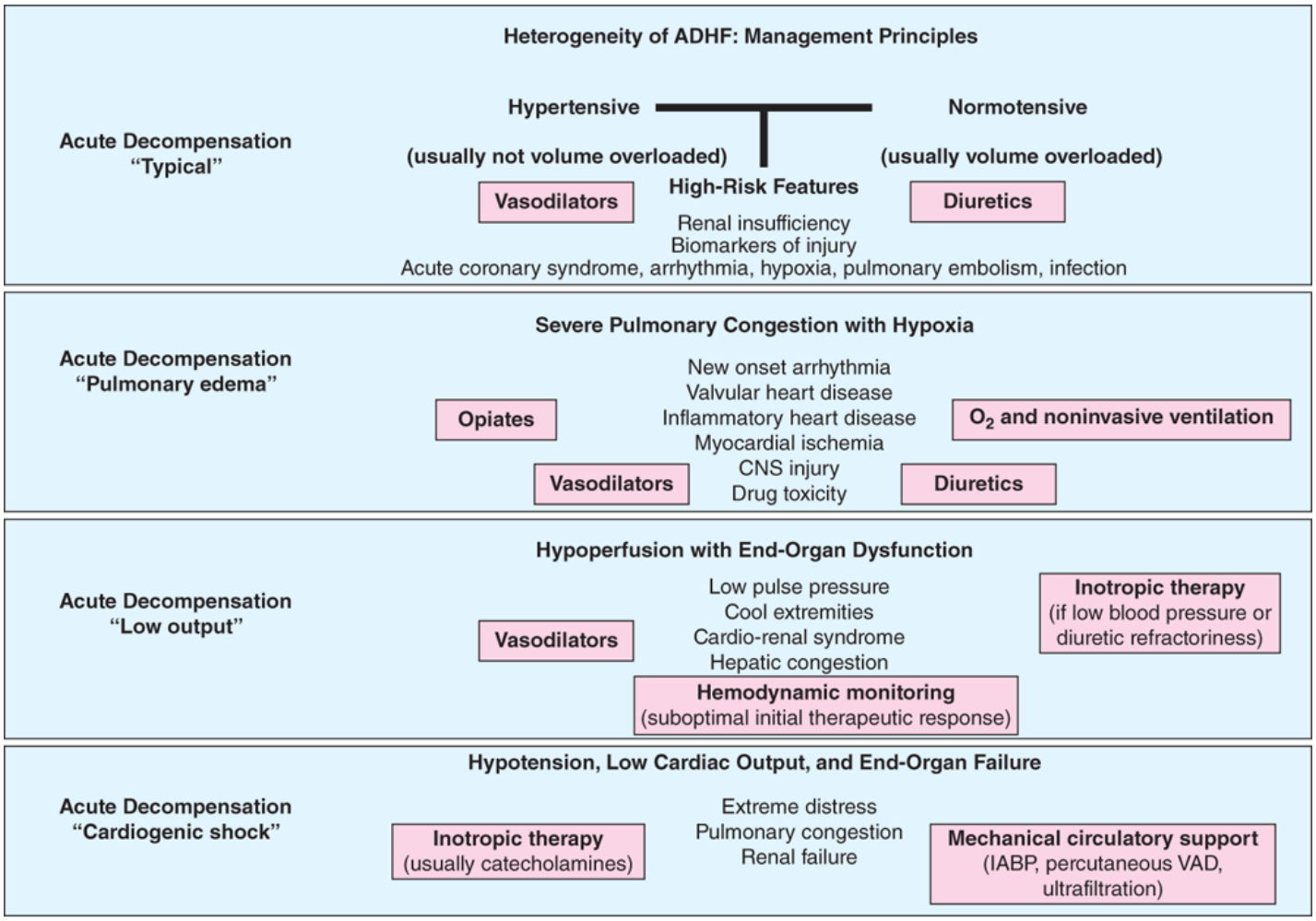
GENERAL PRINCIPLES

ADHF is a heterogeneous clinical syndrome most often resulting in need for hospitalization due to confluence of interrelated abnormalities of decreased cardiac performance, renal dysfunction, and alterations in vascular compliance. Admission with a diagnosis of ADHF is associated with excessive morbidity and mortality, with nearly half of these patients readmitted for management within 6 months, and a high short-term (5% in-hospital) and long-term cardiovascular mortality (20% at 1 year). Importantly, long-term aggregate outcomes remain poor, with a combined incidence of cardiovascular deaths, heart failure hospitalizations, myocardial infarction, strokes, or sudden death reaching 50% at 12 months after hospitalization. The management of these patients has remained difficult and principally revolves around volume control and decrease of vascular impedance while maintaining attention to end-organ perfusion (coronary and renal).

The first principle of management of these patients is to identify and tackle known precipitants of decompensation. Identification and management of medication nonadherence and use of prescribed medicines such as nonsteroidal anti-inflammatory drugs, cold and flu preparations with cardiac stimulants, and herbal preparations, including licorice, ginseng, and herbal forms of ephedrine (now banned in most places), are required. Active infection and overt or covert pulmonary thromboembolism should be sought, identified, and treated when clinical clues suggest such direction. When possible, arrhythmias should be corrected by controlling heart rate or restoring sinus rhythm in patients with poorly tolerated rapid atrial fibrillation and by correcting ongoing ischemia with coronary revascularization or by correcting offenders such as ongoing bleeding in demand-related ischemia. A parallel step in management involves stabilization of hemodynamics in those with instability. The routine use of a pulmonary artery catheter is not recommended and should be restricted to those who respond poorly to diuresis or experience hypotension or signs and symptoms suggestive of a low cardiac output where therapeutic targets are unclear. Analysis of in-hospital registries has identified several parameters associated with worse outcomes: a blood urea nitrogen level >43 mg/dL (to convert to mmol/L, multiply by 0.357), systolic blood pressure <115 mmHg, a serum creatinine level >2.75 mg/dL (to convert to µmol/L, multiply by 88.4), and an elevated troponin I level. A useful clinical schema to identify treatment targets for the various phenotypic presentations and management goals in ADHF is depicted in **Fig. 253-2**.

FIGURE 253-2

The distinctive phenotypes of acute decompensated heart failure (ADHF), their presentations, and suggested therapeutic routes. (Unique causes of ADHF, such as isolated right heart failure and pericardial disease, and rare causes, such as aortic and coronary dissection or ruptured valve structures or sinuses of Valsalva, are not delineated and are covered elsewhere.) IABP, intraaortic balloon pump; VAD, ventricular assist device.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

VOLUME MANAGEMENT

Intravenous Diuretic Agents

Intravenous diuretic agents rapidly and effectively relieve symptoms of congestion and are essential when oral drug absorption is impaired. When high doses of diuretic agents are required or when the effect is suboptimal, a continuous infusion may be needed to reduce toxicity and maintain stable serum drug levels. Randomized clinical trials of high- versus low-dose or bolus versus continuous infusion diuresis have not provided clear justification for the best diuretic strategy in ADHF, and as such, the use of diuretic regimens remains an art rather than science. Addition of a thiazide diuretic agent such as

metolazone in combination provides a synergistic effect and is often required in patients receiving long-term therapy with loop diuretic agents. Change in weight is often used as a surrogate for adequate diuresis, but this objective measure of volume status may be surprisingly difficult to interpret, and weight loss during hospitalization does not necessarily correlate closely with outcomes. It is generally advisable to continue diuresis until euvolemia has been achieved. Physical examination findings, specifically the jugular venous pressure coupled with biomarker trends, are useful in timing discharge planning.

The Cardiorenal Syndrome

The cardiorenal syndrome is being recognized increasingly as a complication of ADHF. Multiple definitions have been proposed for the cardiorenal syndrome, but at its simplest, it can be thought to reflect the interplay between abnormalities of heart and kidney function, with deteriorating function of one organ while therapy is administered to preserve the other. Approximately 30% of patients hospitalized with ADHF exhibit abnormal renal function at baseline, and this is associated with longer hospitalizations and increased mortality. However, mechanistic studies have been largely unable to find correlation between deterioration in renal function, cardiac output, left-sided filling pressures, and reduced renal perfusion; most patients with cardiorenal syndrome demonstrate a preserved cardiac output. It is hypothesized that in patients with established heart failure, this syndrome represents a complex interplay of neurohormonal factors, potentially exacerbated by “backward failure” resulting from increased intraabdominal pressure and impairment in return of renal venous blood flow. Continued use of diuretic therapy may be associated with a reduction in glomerular filtration rate and a worsening of the cardiorenal syndrome when right-sided filling pressures remain elevated. In patients in the late stages of disease characterized by profound low cardiac output state, inotropic therapy or mechanical circulatory support has been shown to preserve or improve renal function in selected individuals in the short term until more definitive therapy such as assisted circulation or cardiac transplantation is implemented.

Ultrafiltration

Ultrafiltration (UF) is an invasive fluid removal technique that may supplement the need for diuretic therapy. Proposed benefits of UF include controlled rates of fluid removal, neutral effects on serum electrolytes, and decreased neurohormonal activity. This technique has also been referred to as aquapheresis in recognition of its electrolyte depletion–sparing effects. In a pivotal study evaluating UF versus conventional therapy, fluid removal was improved and subsequent heart failure hospitalizations and urgent clinic visits were reduced with UF; however, no improvement in renal function and no subjective differences in dyspnea scores or adverse outcomes were noted. In the Cardiorenal Rescue Study in Acute Decompensated Heart Failure (CARRESS-HF) trial, 188 patients with ADHF and worsening renal failure were randomized to stepped pharmacologic care or UF. The primary endpoint was a change in serum creatinine and change in weight (reflecting fluid removal) at 96 h. Although similar weight loss occurred in both groups (~5.5 kg), there was worsening in creatinine in the UF group. Deaths and hospitalizations for heart failure were no different between groups, but there were more adverse events in the UF group, mainly due to kidney failure, bleeding complications, and intravenous catheter-related complications. This investigation argues against using UF as a primary strategy in patients with ADHF who are nonetheless responsive to diuretics. Whether UF is useful in states of diuretic unresponsiveness remains an open question, and this strategy continues to be employed judiciously in such situations.

VASCULAR THERAPY

Vasodilators including *intravenous nitrates, nitroprusside*, and *nesiritide* (a recombinant brain-type natriuretic peptide) have been advocated for upstream therapy in an effort to stabilize ADHF. The latter agent was introduced in a fixed dose for therapy after a comparison with intravenous nitrates suggested more rapid and greater reduction in pulmonary capillary wedge pressure. Enthusiasm for nesiritide waned due to concerns within the pivotal trials for development of renal insufficiency and an increase in mortality. To address these concerns, a large-scale morbidity and mortality trial, the Acute Study of Clinical Effectiveness of Nesiritide in Decompensated Heart Failure (ASCEND-HF) study was completed in 2011 and randomly enrolled 7141 patients with ADHF to nesiritide or placebo for 24–168 h in addition to standard care. Nesiritide was not associated with an increase or a decrease in the rates of death and rehospitalization and had a clinically insignificant benefit on dyspnea. Renal function did not worsen, but increased rates of hypotension were noted. Although this trial established the safety for this drug, the routine use cannot be advocated due to lack of significant efficacy. *Recombinant human relaxin-2*, or serelaxin, is a peptide upregulated in pregnancy and examined in ADHF patients with a normal or elevated blood pressure. In the Relaxin in Acute Heart Failure (RELAX-AHF) trial, serelaxin or placebo was added to a regimen of standard therapy in 1161 patients hospitalized with ADHF, evidence of congestion, and systolic pressure >125 mmHg. Serelaxin improved dyspnea, reduced signs and symptoms of congestion, and was associated with less early worsening of HF. Exploratory endpoints of hard outcomes at 6 months suggested positive signals in favor of mortality reduction. A subsequent larger study failed to demonstrate a benefit. Recently, the natriuretic peptide urodilatin was tested in a large trial (TRUE-AHF) in ADHF patients and while evidence for decongestion was forthcoming along with a reduction in net endogenous expression of natriuretic peptides, there was no improvement in clinical outcomes at 6 months. Urodilatin was associated with a higher rate of hypotension and worsening serum creatinine.

INOTROPIC THERAPY

Impairment of myocardial contractility often accompanies ADHF, and pharmacologic agents that increase intracellular concentration of cyclic adenosine monophosphate via direct or indirect pathways, such as sympathomimetic amines (dobutamine) and phosphodiesterase-3 inhibitors (milrinone), respectively, serve as positive inotropic agents. Their activity leads to an increase in cytoplasmic calcium. Inotropic therapy in those with a low-output state augments cardiac output, improves perfusion, and relieves congestion acutely. Although milrinone and dobutamine have similar hemodynamic profiles, milrinone is slower acting and is renally excreted and thus requires dose adjustments in the setting of kidney dysfunction. Since milrinone acts downstream from the β_1-adrenergic receptor, it may provide an advantage in patients receiving beta blockers when admitted to the hospital. Studies are in universal

agreement that long-term inotropic therapy increases mortality. However, the short-term use of inotropic agents in ADHF is also associated with increased arrhythmia, hypotension, and no beneficial effects on hard outcomes. Inotropic agents are currently indicated as bridge therapy (to either left ventricular assist device support or to transplant) or as selectively applied palliation in end-stage heart failure.

Novel inotropic agents that leverage the concept of myofilament calcium sensitization rather than increasing intracellular calcium levels have been introduced. *Levosimendan* is a calcium sensitizer that provides inotropic activity, but also possesses phosphodiesterase-3 inhibition properties that are vasodilators in action. This makes the drug unsuitable in states of low output in the setting of hypotension. Two trials, the second Randomized Multicenter Evaluation of Intravenous Levosimendan Efficacy (REVIVE II) and Survival of Patients with Acute Heart Failure in Need of Intravenous Inotropic Support (SURVIVE), have tested this agent in ADHF. SURVIVE compared levosimendan with dobutamine, and despite an initial reduction in circulating B-type natriuretic peptide levels in the levosimendan group compared with patients in the dobutamine group, this drug did not reduce all-cause mortality at 180 days or affect any secondary clinical outcomes. The second trial compared levosimendan against traditional non-inotropic therapy and found a modest improvement in symptoms with worsened short-term mortality and ventricular arrhythmias. Another drug that functions as a selective myosin activator, *omecamtiv mecarbil*, prolongs the ejection period and increases fractional shortening. Distinctively, the force of contraction is not increased, and as such, this agent does not increase myocardial oxygen demand. As a follow-up to early encouraging data, the COSMIC-HF (Chronic Oral Study of Myosin Activation to Increase Contractility in Heart Failure) study evaluated 448 patients with chronic heart failure and left ventricular systolic dysfunction, for 20 weeks of treatment, and observed improvements with this agent on cardiac function, left ventricular remodeling indices and natriuretic peptide expression. Other inotropic agents that increase myocardial calcium sensitivity through mechanisms that reduce cTnI phosphorylation or inhibit protein kinase A are being developed. (**Table 253-1** depicts typical inotropic, vasodilator, and diuretic drugs used in ADHF.)

TABLE 253-1

Intravenous Therapy in Acute Decompensated Heart Failure

Drug Class	Generic Drug	Usual Dosing	Special Caution	Comments
Inotropic therapy				Use in hypotension, end-organ hypoperfusion, or shock states
	Dobutamine	2–20 μg/kg per min	Increased myocardial oxygen demand, arrhythmia	Short acting, an advantage; variable efficacy in presence of beta blockers (requires higher doses); clinical tolerance to prolonged infusions; concerns with hypersensitivity carditis (rare)
	Milrinone	0.375–0.75 μg/kg per min	Hypotension, arrhythmia	Decrease dose in renal insufficiency; avoid initial bolus; effectiveness retained in presence of beta blockers
	Levosimendan	0.1 μg/kg per min, range, 0.05–0.2 μg/kg per min	Hypotension, arrhythmia	Long acting; should not be used in presence of low blood pressure; similar effectiveness as dobutamine but effectiveness retained in presence of beta blockers
	Omecamtiv Mecarbil	N/A	*In trials	Increases contractility without increasing myocardial oxygen demand; in confirmatory trials
Vasodilators				Use in presence of pulmonary congestion for rapid relief of dyspnea, in presence of a preserved blood pressure
	Nitroglycerine	10–20 μg/min, increase up to 200 μg/min	Headache, flushing, tolerance	Most common vasodilator but often underdosed; effective in higher doses
	Nesiritide	Bolus 2 μg/kg and infusion at 0.01 μg/kg per min	Hypotension	Decrease in blood pressure may reduce renal perfusion pressure; bolus may be avoided since it increases hypotension predilection
	Nitroprusside	0.3 μg/kg per min titrated to 5 μg/kg per min	Thiocyanate toxicity in renal insufficiency (>72 h)	Requires arterial line placement for titration for precise blood pressure management and prevention of hypotension
	Serelaxin	N/A (tested at 30 μg/kg per d)	Baseline blood pressure should be >125 mmHg	Not widely commercially available; ineffective in confirmatory trials
	Ularitide	15 ng/kg/min (48 h)	Baseline blood pressure >116 mmHg	Excess hypotension and increase serum creatinine
Diuretics				First line of therapy in volume overload with congestion; may use bolus or continuous dosing; initial low dose (1 × home dose) or high dose (2.5 × home dose) equally effective with higher risk of renal worsening with higher dose
	Furosemide	20–240 mg daily	Monitor for electrolyte loss	In severe congestion, use intravenously and consider continuous infusion (not trial supported)
	Torsemide	10–100 mg daily	Monitor for electrolyte loss	High bioavailability, can be given orally; anecdotally more effective in advanced heart failure states if furosemide less bioavailable (due to gut congestion)

Drug Class	Generic Drug	Usual Dosing	Special Caution	Comments
	Bumetanide	0.5–5 mg daily	Monitor for electrolyte loss	Can be used orally; intermediate bioavailability
	Adjuvant diuretics for augmentation	n/a	Metolazone, chlorthalidone, spironolactone, acetazolamide	Acetazolamide is useful in presence of alkalosis; metolazone given in 2.5- to 10-mg doses; causes severe electrolyte imbalance; spironolactone is useful in presence of severe hypokalemia and normal renal function

NEUROHORMONAL ANTAGONISTS

Other trials testing unique agents have yielded disappointing results in the situation of ADHF. The Placebo-Controlled Randomized Study of the Selective A1 Adenosine Receptor Antagonist Rolofylline for Patients Hospitalized with Acute Decompensated Heart Failure and Volume Overload to Assess Treatment Effect on Congestion and Renal Function (PROTECT) trial of selective adenosine antagonism and the Efficacy of Vasopressin Antagonism in Heart Failure Outcome Study with Tolvaptan (EVEREST) trial of an oral selective vasopressin-2 antagonist in ADHF were both negative with respect to hard outcomes.

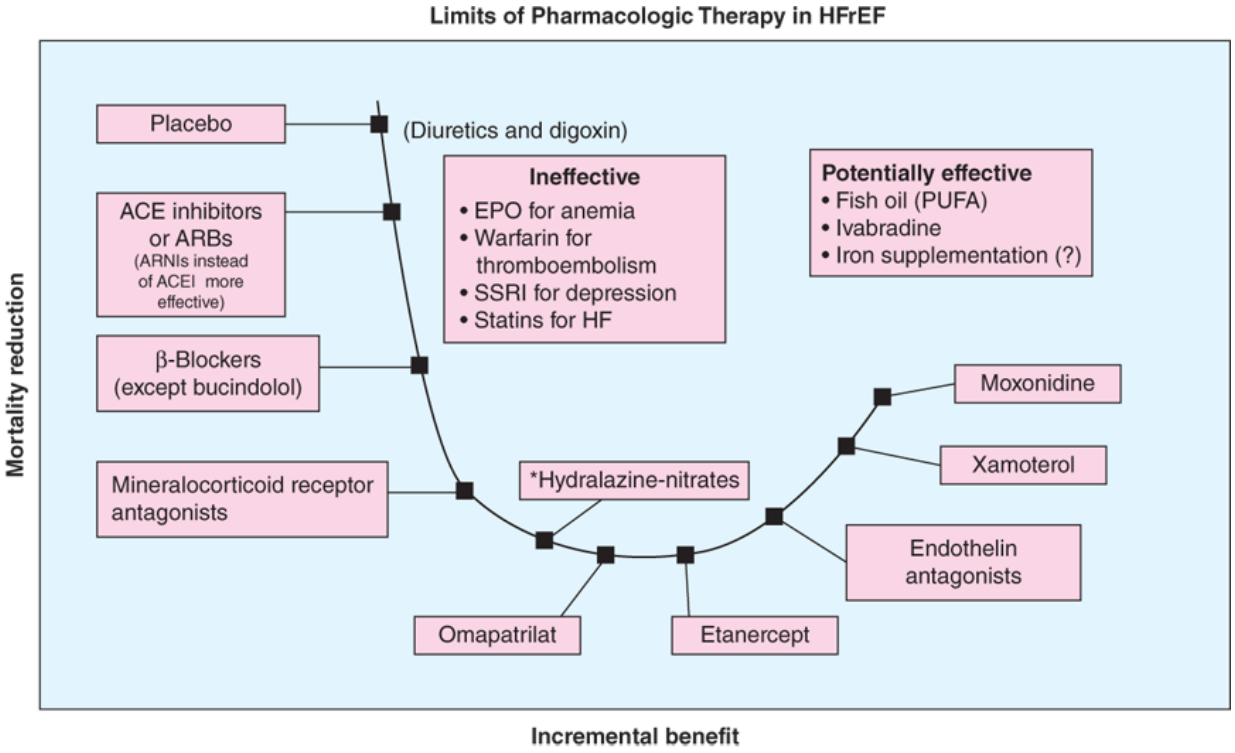
In patients who fail to respond adequately to medical therapy, mechanical assist devices may be required. This is covered in more detail in Chap. 255.

HEART FAILURE WITH REDUCED EJECTION FRACTION

The last 50 years have witnessed great strides in the management of HFrEF. The treatment of symptomatic heart failure that evolved from a renocentric (diuretics) and hemodynamic therapy model (digoxin, inotropic therapy) ushered in the era of disease-modifying therapy with neurohormonal antagonism. In this regard, RAAS blockers and beta blockers form the cornerstone of pharmacotherapy and lead to attenuation of decline and improvement in cardiac structure and function with consequent reduction in symptoms, improvement in QOL, decreased burden of hospitalizations, and a decline in mortality from both pump failure and arrhythmic deaths (**Fig. 253-3**).

FIGURE 253-3

Progressive decline in mortality with angiotensin-converting enzyme (ACE) inhibitors or angiotensin receptor blockers (ARBs) or angiotensin receptor neprilysin inhibitors (ARNIs), beta blockers, mineralocorticoid receptor antagonists, and balanced vasodilators (*selected populations such as African Americans); further stack-on neurohormonal therapy is ineffective or results in worse outcome; management of comorbidity is of unclear efficacy. EPO, erythropoietin; HF, heart failure; HFrEF, heart failure with reduced ejection fraction; PUFA, polyunsaturated fatty acid; SSRI, selective serotonin reuptake inhibitor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

NEUROHORMONAL ANTAGONISM

Meta-analyses suggest a 23% reduction in mortality and a 35% reduction in the combination endpoint of mortality and hospitalizations for heart failure in patients treated with ACEIs. Patients treated with beta blockers provide a further 35% reduction in mortality on top of the benefit provided by ACEIs alone. Increased experience with both agents in a broad range of patients with HFrEF has demonstrated the safety of ACEIs in treating patients with mild renal insufficiency and the tolerability of beta blockers in patients with moderately controlled diabetes, asthma, and obstructive lung disease. The benefits of ACEIs and beta blockers extend to advanced symptoms of disease (NYHA class IIIb–IV). However, a substantial number of patients with advanced heart failure may not be able to achieve optimal doses of neurohormonal inhibitors and require cautious reduction in dose exposure to maintain clinical stability. Such individuals with lower exposure to ACEIs and beta blockers represent a high-risk cohort with poor prognosis.

Class Effect and Sequence of Administration

ACEIs exert their beneficial effects in HFrEF as a class; however, the beneficial effects of beta blockers are thought to be limited to specific drugs. Beta blockers with intrinsic sympathomimetic activity (xamoterol) and other agents, including bucindolol, have not demonstrated a survival benefit. On the basis of investigations, beta blocker use in HFrEF should ideally be restricted to carvedilol, bisoprolol, and metoprolol succinate—agents tested and proven to improve survival in clinical trials. Whether beta blockers or ACEIs should be started first was answered by the Cardiac Insufficiency Bisoprolol Study (CIBIS) III, in which outcomes did not vary when either agent was initiated first. Thus, it matters little which agent is initiated first; what does matter is that optimally titrated doses of both ACEIs and beta blockers be established in a timely manner.

Dose and Outcome

A trial has indicated that higher tolerated doses of ACEIs achieve greater reduction in hospitalizations without materially improving survival. Beta blockers demonstrate a dose-dependent improvement in cardiac function and reductions in mortality and hospitalizations. Clinical experience suggests that, in the absence of symptoms to suggest hypotension (fatigue and dizziness), pharmacotherapy may be up-titrated every 2 weeks in stable ambulatory patients as tolerated.

MINERALOCORTICOID ANTAGONISTS

Aldosterone antagonism is associated with a reduction in mortality in all stages of symptomatic NYHA class II to IV HFrEF. Elevated aldosterone levels in HFrEF promote sodium retention, electrolyte imbalance, and endothelial dysfunction and may directly contribute to myocardial fibrosis. The selective agent eplerenone (tested in NYHA class II and post–myocardial infarction heart failure) and the nonselective antagonist spironolactone (tested in NYHA class III and IV heart failure) reduce mortality and hospitalizations, with significant reductions in sudden cardiac death (SCD). Hyperkalemia and worsening renal function are concerns, especially in patients with underlying chronic kidney disease, and renal function and serum potassium levels must be closely monitored.

RAAS THERAPY AND NEUROHORMONAL "ESCAPE"

Neurohormonal "escape" has been witnessed in patients with HFrEF by the finding that circulating levels of angiotensin II return to pretreatment levels with long-term ACEI therapy. ARBs blunt this phenomenon by binding competitively to the AT_1 receptor. Meta-analysis of 24 randomized trials demonstrated the superiority of ARBs to placebo in patients with intolerable adverse effects with ACEIs and their non-inferiority in all-cause mortality or hospitalizations when compared with ACEIs. The Valsartan Heart Failure Trial (Val-HeFT) suggested that addition of valsartan in patients already receiving treatment with ACEIs and beta blockers was associated with a trend toward worse outcomes. Similarly, adding valsartan to captopril in patients with heart failure after myocardial infarction who were receiving background beta blocker therapy was associated with an increase in adverse events without any added benefit compared with monotherapy for either group. Thus, the initial clinical strategy should be to use a two-drug combination first (ACEI and beta blocker; if beta blocker intolerant, then ACEI and ARB; if ACEI intolerant, then ARB and beta blocker). In symptomatic patients (NYHA class II–IV), an aldosterone antagonist should be strongly considered, but four-drug therapy should be avoided.

A recent trial called the Aliskiren Trial on Acute Heart Failure Outcomes (ASTRONAUT) tested a direct renin inhibitor, aliskiren, in addition to other heart failure medications, within a week after discharge from a hospitalization for decompensated HFrEF. No significant difference in cardiovascular death or hospitalization at 6 or 12 months was noted. Aliskiren was associated with a reduction in circulating natriuretic peptides, but any disease-modifying effect was overcome by excessive adverse events including hyperkalemia, hypotension, and renal dysfunction. These studies point to the limits achieved with RAAS modulation in this clinical syndrome.

ARTERIOVENOUS VASODILATION

The combination of hydralazine and nitrates has been demonstrated to improve survival in HFrEF. Hydralazine reduces systemic vascular resistance and induces arterial vasodilatation by affecting intracellular calcium kinetics; nitrates are transformed in smooth muscle cells into nitric oxide, which stimulates cyclic guanosine monophosphate production and consequent arterial-venous vasodilation. This combination improves survival, but not to the magnitude evidenced by ACEIs or ARBs. However, in individuals with HFrEF unable to tolerate RAAS-based therapy for reasons such as renal insufficiency or hyperkalemia, this combination is preferred as a disease-modifying approach. A trial conducted in self-identified African Americans, the African-American Heart Failure Trial (A-Heft), studied a fixed dose of isosorbide dinitrate with hydralazine in patients with advanced symptoms of HFrEF who were receiving standard background therapy. The study demonstrated benefit in survival and hospitalization recidivism in the treatment group. Adherence to this regimen is limited by the thrice-daily dosing schedule.

NOVEL NEUROHORMONAL ANTAGONISM

Despite an abundance of animal and clinical data demonstrating deleterious effects of activated neurohormonal pathways beyond the RAAS and sympathetic nervous system, targeting such pathways with incremental blockade has been largely unsuccessful. As an example, the endothelin antagonist bosentan is associated with worsening heart failure in HFrEF despite demonstrating benefits in right-sided heart failure due to pulmonary arterial hypertension. Similarly, the centrally acting sympatholytic agent moxonidine worsens outcomes in left heart failure. The combined drug omapatrilat hybridizes an ACEI with a neutral endopeptidase inhibitor, and this agent was tested in the Omapatrilat Versus Enalapril Randomized Trial of Utility in Reducing Events (OVERTURE) trial. This drug did not favorably influence the primary outcome measure of the combined risk of death or hospitalization for heart failure requiring intravenous treatment. The risk of angioedema was notably higher with omapatrilat than ACEIs alone.

More recently, the introduction of LCZ696, an ARB (valsartan) with an endopeptidase inhibitor (sacubitril), has shown a survival benefit in a large trial versus ARB alone. The drug, referred to as an angiotensin receptor–neprilysin inhibitor (ARNI) (and denoted Entrezto), was tested in the PARADIGM-HF trial as an alternate to optimally dosed ACEI and demonstrated an incremental improvement in survival when compared to ACEI alone. Most guidelines now advocate switching ACEI to this drug as a standard in patients with mild-moderate systolic heart failure when they remain symptomatic despite fully tolerated doses of conventional therapy.

Table 253-2 lists the common neurohormonal and vasodilator regimens for HFrEF.

TABLE 253-2

Pharmacologic Therapy and Target Doses in Heart Failure with Reduced Ejection Fraction

Drug Class	Generic Drug	Mean Daily Dose in Clinical Trials (mg)	Initiation (mg)	Target Dose (mg)
Angiotensin-Converting Enzyme Inhibitors				
	Lisinopril	4.5–33	2.5–5 qd	20–35 qd
	Enalapril	17	2.5 bid	10–20 bid
	Captopril	123	6.25 tid	50 tid
	Trandolapril	N/A	0.5–1 qd	4 qd
Angiotensin Receptor Blockers				
	Losartan	129	50 qd	150 qd
	Valsartan	254	40 bid	160 bid
	Candesartan	24	4–8 qd	32 qd
Aldosterone Antagonists				
	Eplerenone	42.6	25 qd	50 qd
	Spironolactone	26	12.5–25 qd	25–50 qd
Beta Blockers				
	Metoprolol succinate CR/XL	159	12.5–25 qd	200 qd
	Carvedilol	37	3.125 bid	25–50 bid
	Bisoprolol	8.6	1.25 qd	10 qd
Arteriovenous Vasodilators				
	Hydralazine isosorbide dinitrate	270/136	37.5/20 tid	75/40 tid
	Fixed-dose hydralazine/isosorbide dinitrate	143/76	37.5/20 qid	75/40 qid
Angiotensin Receptor Neprilysin Inhibitor				
	Sacubitril-valsartan	375	100 bid	200 bid

HEART RATE MODIFICATION

Ivabradine, an inhibitor of the I_f current in the sinoatrial node, slows the heart rate without a negative inotropic effect. The Systolic Heart Failure Treatment with Ivabradine Compared with Placebo Trial (SHIFT) was conducted in patients with class II or III HFrEF, a heart rate >70 beats/min, and history of hospitalization for heart failure during the previous year. Ivabradine reduced hospitalizations and the combined endpoint of cardiovascular-related death and heart failure hospitalization. The study population was not necessarily representative of North American patients with HFrEF since, with a few exceptions, most did not receive internal cardioverter-defibrillation or cardiac resynchronization therapy and 40% did not receive a mineralocorticoid receptor antagonist. Although 90% received beta blockers, only a quarter were on full doses. Whether this agent would have been effective in patients receiving robust, guideline-recommended therapy for heart failure remains unclear. In the 2012 European Society of Cardiology guidelines for the treatment of heart failure, clinically, Ivabradine should be considered in patients who remain symptomatic after guideline-based ACEIs, beta blockers, and

mineralocorticoid receptor antagonists and with residual heart rate >70 beats/min. Another group in whom potential benefit may be expected includes those unable to tolerate beta blockers.

DIGOXIN

Digitalis glycosides exert a mild inotropic effect, attenuate carotid sinus baroreceptor activity, and are sympatho-inhibitory. These effects decrease serum norepinephrine levels, plasma renin levels, and possibly aldosterone levels. The DIG trial demonstrated a reduction in heart failure hospitalizations in the treatment group (patients with heart failure and sinus rhythm) but no reduction in mortality or improvement in QOL. Importantly, treatment with digoxin resulted in a higher mortality rate and hospitalizations in women than men. It should be noted that low doses of digoxin are sufficient to achieve any potentially beneficial outcomes, and higher doses breach the therapeutic safety index. Although digoxin levels should be checked to minimize toxicity and although dose reductions are indicated for higher levels, no adjustment is made for low levels. Generally, digoxin is now relegated as therapy for patients who remain profoundly symptomatic despite optimal neurohormonal blockade and adequate volume control.

ORAL DIURETICS

Neurohormonal activation results in avid salt and water retention. Loop diuretic agents are often required because of their increased potency, and frequent dose adjustments may be necessary because of variable oral absorption and fluctuations in renal function. Importantly, clinical trial data confirming efficacy are limited, and no data suggest that these agents improve survival. Thus, diuretic agents should ideally be used in tailored dosing schedules to avoid excessive exposure. Indeed, diuretics are essential at the outset to achieve volume control before neurohormonal therapy is likely to be well tolerated or titrated.

CALCIUM CHANNEL ANTAGONISTS

Amlodipine and felodipine, second-generation calcium channel–blocking agents, safely and effectively reduce blood pressure in HFrEF but do not affect morbidity, mortality, or QOL. The first-generation agents, including verapamil and diltiazem, may exert negative inotropic effects and destabilize previously asymptomatic patients. Their use should be discouraged.

INFLAMMATION

Targeting inflammatory cytokines such as tumor necrosis factor α (TNF-α) by using anticytokine agents such as infliximab and etanercept has been unsuccessful and associated with worsening heart failure. Use of intravenous immunoglobulin therapy in nonischemic etiology of heart failure has not been shown to result in beneficial outcomes. Nonspecific immunomodulation has been tested in the Advanced Chronic Heart Failure Clinical Assessment of Immune Modulation Therapy (ACCLAIM-HF) trial where ex-vivo exposure of a blood sample from systolic heart failure patients to controlled oxidative stress was hypothesized to initiate apoptosis of leukocytes soon after intramuscular gluteal injection of the treated sample. The physiologic response to apoptotic cells results in a reduction in inflammatory cytokine production and upregulation of anti-inflammatory cytokines. This promising hypothesis was not proven, although certain subgroups (those with no history of previous myocardial infarction and those with mild heart failure) showed signals in favor of immunomodulation.

STATINS

Potent lipid-altering and pleiotropic effects of statins reduce major cardiovascular events and improve survival in non–heart failure populations. Once heart failure is well established, this therapy may not be as beneficial and theoretically could even be detrimental by depleting ubiquinone in the electron transport chain. Two trials, Controlled Rosuvastatin Multinational Trial in Heart Failure (CORONA) and Gruppo Italiano per lo Studio della Sopravvivenza nell'Insufficienza Cardiac (GISSI-HF), have tested low-dose rosuvastatin in patients with HFrEF and demonstrated no improvement in aggregate clinical outcomes. If statins are required to treat progressive coronary artery disease in the background setting of heart failure, then they should be employed. However, no rationale appears to exist for routine statin therapy in nonischemic heart failure.

ANTICOAGULATION AND ANTIPLATELET THERAPY

HFrEF is accompanied by a hypercoagulable state and therefore a high risk of thromboembolic events, including stroke, pulmonary embolism, and peripheral arterial embolism. Although long-term oral anticoagulation is established in certain groups, including patients with atrial fibrillation, the data are insufficient to support the use of warfarin in patients in normal sinus rhythm without a history of thromboembolic events or echocardiographic evidence of left ventricular thrombus. In the large Warfarin versus Aspirin in Reduced Cardiac Ejection Fraction (WARCEF) trial, full-dose aspirin or international normalized ratio–controlled warfarin was tested with follow-up for 6 years. Among patients with reduced LVEF in sinus rhythm, there was no significant overall difference in the primary outcome between treatment with warfarin and treatment with aspirin. A reduced risk of ischemic stroke with warfarin was offset by an increased risk of major hemorrhage. Aspirin blunts ACEI-mediated prostaglandin synthesis, but the clinical importance of this finding remains unclear. Current guidelines support the use of aspirin in patients with ischemic cardiomyopathy.

FISH OIL

Treatment with long-chain omega-3 polyunsaturated fatty acids (ω-3 PUFAs) has been shown to be associated with modestly improved clinical outcomes in patients with HFrEF. This observation from the GISSI-HF trial was extended to measurements of ω-3 PUFAs in plasma phospholipids at baseline and after 3 months. Three-month treatment with ω-3 PUFAs enriched circulating eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA). Low EPA levels are inversely related to total mortality in patients with HFrEF.

MICRONUTRIENTS

A growing body of evidence suggests an association between heart failure and micronutrient status. Reversible heart failure has been described as a consequence of severe thiamine and selenium deficiency. Thiamine deficiency has received attention in heart failure due to the fact that malnutrition and diuretics are prime risk factors for thiamine loss. Small exploratory randomized studies have suggested a benefit of supplementation of thiamine in HFrEF with evidence of improved cardiac function. This finding is restricted to chronic heart failure states and does not appear to be beneficial in the ADHF phenotype. Due to the exploratory nature of the evidence, no recommendations for routine supplementation or testing for thiamine deficiency can be made.

ENHANCED EXTERNAL COUNTERPULSATION (EECP)

Peripheral lower extremity therapy using graded external pneumatic compression at high pressure is administered in 1-h sessions for 35 treatments (7 weeks) and has been proposed to reduce angina symptoms and extend time to exercise-induced ischemia in patients with coronary artery disease. The Prospective Evaluation of Enhanced External Counterpulsation in Congestive Heart Failure (PEECH) study assessed the benefits of enhanced external counterpulsation in the treatment of patients with mild-to-moderate heart failure. This randomized trial improved exercise tolerance, QOL, and NYHA functional classification but without an accompanying increase in peak oxygen consumption. A placebo effect due to the nature of the intervention simply cannot be excluded.

EXERCISE

The Heart Failure: A Controlled Trial Investigating Outcomes of Exercise Training (HF-ACTION) study investigated short-term (3-month) and long-term (12-month) effects of a supervised exercise training program in patients with moderate HFrEF. Exercise was safe, improved patients' sense of well-being, and correlated with a trend toward mortality reduction. Maximal changes in 6-min walk distance were evident at 3 months with significant improvements in cardiopulmonary exercise time and peak oxygen consumption persisting at 12 months. Therefore, exercise training is recommended as an adjunctive treatment in patients with heart failure.

MANAGEMENT OF SELECTED COMORBIDITY

Sleep-disordered breathing is common in HF and particularly in HFrEF. A range of presentations exemplified by obstructive sleep apnea, central sleep apnea, and its extreme form of Cheyne-Stokes breathing are noted. Frequent periods of hypoxia and repeated micro- and macro-arousals trigger adrenergic surges, which can worsen hypertension and impair systolic and diastolic function. A high index of suspicion is required, especially in patients with difficult-to-control hypertension or with predominant symptoms of fatigue despite reverse remodeling in response to optimal medical therapy. Worsening of right heart function with improvement of left ventricular function noted on medical therapy should immediately trigger a search for underlying sleep-disordered breathing or pulmonary complications such as occult embolism or pulmonary hypertension. Treatment with nocturnal positive airway pressure improves oxygenation, LVEF, and 6-min walk distance. However, no conclusive data exist to support this therapy as a disease-modifying approach with reduction in mortality. A recent trial, using adaptive servo-ventilation in patients who had HFrEF and predominantly central sleep apnea increased all cause and cardiovascular mortality.

Anemia is common in heart failure patients, reduces functional status and QOL, and is associated with increased proclivity for hospital admissions and mortality. Anemia in heart failure is more common in the elderly, in those with advanced stages of HFrEF, in the presence of renal insufficiency, and in women and African Americans. The mechanisms include iron deficiency, dysregulation of iron metabolism, and occult gastrointestinal bleeding. Intravenous iron using either iron sucrose or carboxymaltose (Ferric Carboxymaltose Assessment in Patients with Iron Deficiency and Chronic Heart Failure [FAIR-HF] trial) has been shown to correct anemia and improve functional capacity. Another trial, CONFIRM-HF, enrolled similar patients with iron deficiency (ferritin <100 ng/mL or 100–300 ng/mL if transferrin saturation <20%) and demonstrated that use of ferric carboxymaltose in a simplified high dose schedule resulted in improvement in functional capacity, symptoms, and QOL. Oral iron supplementation does not appear to be effective in treating iron deficiency in heart failure. Erythropoiesis-regulating agents such as erythropoietin analogues have been studied with disappointing results. The Reduction of Events by Darbepoetin Alfa in Heart Failure (RED-HF) trial demonstrated that treatment with darbepoetin alfa did not improve clinical outcomes in patients with systolic heart failure.

Depression is common in HFrEF, with a reported prevalence of one in five patients, and is associated with a poor QOL, limited functional status, and increased risk of morbidity and mortality in this population. However, the largest randomized study of depression in HFrEF, the Sertraline Against Depression and Heart Disease in Chronic Heart Failure (SADHART-CHF) trial, showed that although sertraline was safe, it did not provide greater reduction in depression or improve cardiovascular status among patients with heart failure and depression compared with nurse-driven multidisciplinary management.

Atrial arrhythmias, especially atrial fibrillation, are common and serve as a harbinger of worse prognosis in patients with heart failure. When rate control is inadequate or symptoms persist, pursuing a rhythm control strategy is reasonable. Rhythm control may be achieved via pharmacotherapy or by percutaneous or surgical techniques, and referral to practitioners or centers experienced in these modalities is recommended. Antiarrhythmic drug therapy should be restricted to amiodarone and dofetilide, both of which have been shown to be safe and effective but do not alter the natural history of the underlying disease. The Antiarrhythmic Trial with Dronedarone in Moderate-to-Severe Congestive Heart Failure Evaluating Morbidity Decrease (ANDROMEDA) studied the effects of the novel antiarrhythmic agent dronedarone and found an increased mortality due to worsening heart failure. Catheter ablation and pulmonary vein isolation appear to be safe and effective in this high-risk cohort and compare favorably with the more established practice of atrioventricular node ablation and biventricular pacing.

Diabetes mellitus is a frequent co-morbidity in heart failure. Prior studies using thiazolidinediones (activators of peroxisome proliferator-activated receptors) have been associated with worsening heart failure. Glucagon-like peptide 1 (GLP-1) agonists such as liraglutide have also been tested and do not lead to greater post-hospitalization clinical stability or worsening in heart failure. Recently, the drug empagloflozin was tested in the EMPA-REG study and demonstrated a decrease in cardiovascular mortality as well as hospitalizations for heart failure. This drug, a sodium–glucose cotransporter 2 (SGLT2), induces an osmotic diuresis as well as ketosis. This drug class may represent a viable therapeutic avenue in diabetics with heart failure.

NEUROMODULATION USING DEVICE THERAPY

Autonomic dysfunction is common in heart failure and attempts at using devices to modulate the sympathetic and parasympathetic systems have been undertaken. Broadly, devices that achieve vagal nerve stimulation, baroreflex activation, renal sympathetic denervation, spinal cord stimulation, or left cardiac sympathetic denervation have been employed. While small preclinical and clinical studies have demonstrated benefits, large-sized randomized trials, when conducted, have failed. The INOVATE-HF study tested vagal nerve stimulation versus optimal medical therapy among individuals with stable HF. Vagus nerve stimulation did not reduce the rate of death or hospitalization for HF. However, functional capacity and QOL were favorably affected by vagus nerve stimulation.

CARDIAC RESYNCHRONIZATION THERAPY

Nonsynchronous contraction between the walls of the left ventricle (intraventricular) or between the ventricular chambers (interventricular) impairs systolic function, decreases mechanical efficiency of contraction, and adversely affects ventricular filling. Mechanical dyssynchrony results in an increase in wall stress and worsens functional mitral regurgitation. The single most important association of extent of dyssynchrony is a widened QRS interval on the surface electrocardiogram, particularly in the presence of a left bundle branch block pattern. With placement of a pacing lead via the coronary sinus to the lateral wall of the ventricle, cardiac resynchronization therapy (CRT) enables a more synchronous ventricular contraction by aligning the timing of activation of the opposing walls. Early studies showed improved exercise capacity, reduction in symptoms, and evidence of reverse remodeling. The Cardiac Resynchronization in Heart Failure Study (CARE-HF) trial was the first study to demonstrate a reduction in all-cause mortality with CRT placement in patients with HFrEF on optimal therapy with continued moderate-to-severe residual symptoms of NYHA class III or IV heart failure. More recent clinical trials have demonstrated disease-modifying properties of CRT in even minimally symptomatic patients with HFrEF, including the Resynchronization–Defibrillation for Ambulatory Heart Failure Trial (RAFT) and Multicenter Automatic Defibrillator Implantation Trial with Cardiac Resynchronization Therapy (MADIT-CRT), both of which sought to use CRT in combination with an implantable defibrillator. Most benefit in mildly symptomatic HFrEF patients accrues from applying this therapy in those with a QRS width of >149 ms and a left bundle branch block pattern. Attempts to further optimize risk stratification and expand indications for CRT using modalities other than electrocardiography have proven disappointing. In particular, echocardiographically derived measures of dyssynchrony vary tremendously, and narrow QRS dyssynchrony has not proven to be a good target for treatment. Uncertainty surrounds the benefits of CRT in those with ADHF, a predominant right bundle branch block pattern, atrial fibrillation, and evidence of scar in the lateral wall, which is the precise location where the CRT lead is positioned.

SUDDEN CARDIAC DEATH PREVENTION IN HEART FAILURE

SCD due to ventricular arrhythmias is the mode of death in approximately half of patients with heart failure and is particularly proportionally prevalent in HFrEF patients with early stages of the disease. Patients who survive an episode of SCD are considered to be at very high risk and qualify for placement of an implantable cardioverter-defibrillator (ICD). Although primary prevention is challenging, the degree of residual left ventricular dysfunction despite optimal medical therapy (≤35%) to allow for adequate remodeling and the underlying etiology (post–myocardial infarction or ischemic cardiomyopathy) are the two single most important risk markers for stratification of need and benefit. Currently, patients with NYHA class II or III symptoms of heart failure and an LVEF <35%, irrespective of etiology of heart failure, are appropriate candidates for ICD prophylactic therapy. In patients with a myocardial infarction and optimal medical therapy with residual LVEF ≤30% (even when asymptomatic), placement of an ICD is appropriate. A recent Danish trial suggested that prophylactic ICD implantation in patients with symptomatic systolic heart failure not caused by coronary artery disease was not associated with a significantly lower long-term rate of death from any cause than was usual clinical care. In this trial, benefits were noted in those aged <60 years. In patients with a terminal illness and a predicted life span of <6 months or in those with NYHA class IV symptoms who are refractory to medications and who are not candidates for transplant, the risks of multiple ICD shocks must be carefully weighed against the survival benefits. If a patient meets the QRS criteria for CRT, combined CRT with ICD is often employed (**Table 253-3**).

TABLE 253-3

Principles of ICD Implantation for Primary Prevention of Sudden Death

Principle	Comment
Arrhythmia–sudden death mismatch	Sudden death in heart failure patients is generally due to progressive LVD, not a focal arrhythmia substrate (except in patients with post-MI HF)
Diminishing returns with advanced disease	Intervention at early stages of HF most successful since sudden death diminishes as cause of death with advanced HF
Timing of benefits	LVEF should be evaluated on optimal medical therapy or after revascularization before ICD therapy is employed; no benefit to ICD implant within 40 days of an MI (unless for secondary prevention)
Estimation of benefits and prognosis	Patients and clinicians often overestimate benefits of ICDs; an ICD discharge is not equivalent to an episode of sudden death (some ventricular arrhythmias terminate spontaneously); appropriate ICD discharges are associated with a worse near-term prognosis

Abbreviations: HF, heart failure; ICD, implantable cardioverter-defibrillator; LVD, left ventricular disease; LVEF, left ventricular ejection fraction; MI, myocardial infarction.

SURGICAL THERAPY IN HEART FAILURE

Coronary artery bypass grafting (CABG) is considered in patients with ischemic cardiomyopathy with multivessel coronary artery disease. The recognition that hibernating myocardium, defined as myocardial tissue with abnormal function but maintained cellular function, could recover after revascularization led to the notion that revascularization with CABG would be useful in those with living myocardium. Revascularization is most robustly supported in individuals with ongoing angina and left ventricular failure. Revascularizing those with left ventricular failure in the absence of angina remains controversial. The Surgical Treatment for Ischemic Heart Failure (STICH) trial in patients with an ejection fraction of ≤35% and coronary artery disease amenable to CABG demonstrated no significant initial benefit compared to medical therapy. However, patients assigned to CABG had lower rates of death from cardiovascular causes and of death from any cause or hospitalization for cardiovascular causes over 10 years than among those who received medical therapy alone. An ancillary study of this trial also determined that the detection of hibernation pre-revascularization did not materially influence the efficacy of this approach, nor did it help to define a population unlikely to benefit if hibernation was not detected.

Surgical ventricular restoration (SVR), a technique characterized by infarct exclusion to remodel the left ventricle by reshaping it surgically in patients with ischemic cardiomyopathy and dominant anterior left ventricular dysfunction, has been proposed. However, in a 1000-patient trial in patients with HFrEF who underwent CABG alone or CABG plus SVR, the addition of SVR to CABG had no disease-modifying effect. However, left ventricular aneurysm surgery is still advocated in those with refractory heart failure, ventricular arrhythmias, or thromboembolism arising from an akinetic aneurysmal segment of the ventricle. Other remodeling procedures, such as use of an external mesh-like net attached around the heart to limit further enlargement, have not been shown to provide hard clinical benefits, although favorable cardiac remodeling was noted.

Mitral regurgitation (MR) occurs with varying degrees in patients with HFrEF and dilated ventricles. Annular dilatation and leaflet noncoaptation in the setting of anatomically normal papillary muscles, chordal structures, and valve leaflets characterize functional MR. In patients who are not candidates for surgical coronary revascularization, mitral valve repair remains controversial. Ischemic MR (or infarct-related MR) is typically associated with leaflet tethering and displacement related to abnormal left ventricular wall motion and geometry. No evidence to support the use of surgical or percutaneous valve correction for functional MR exists as disease-modifying therapy even though MR can be corrected.

CELLULAR AND GENE-BASED THERAPY

The cardiomyocyte possesses regenerative capacity and such renewal is accelerated under conditions of stress and injury, such as an ischemic event or heart failure. Investigations that use either bone marrow–derived precursor cells or autologous cardiac-derived cells have gained traction but have not generally improved clinical outcomes in a convincing manner. More promising, however, are cardiac-derived stem cells. Two preliminary pilot trials delivering cells via an intracoronary approach have been reported. In one, autologous c-kit–positive cells isolated from the atria obtained from patients undergoing CABG were cultured and reinfused. In another, cardiosphere-derived cells grown from endomyocardial biopsy specimens were used. These small trials demonstrated improvements in left ventricular function but require far more work to usher in a clinical therapeutic success. The appropriate route of administration, the quantity of cells to achieve a minimal therapeutic threshold, the constitution of these cells (single source or mixed), the mechanism by which benefit accrues, and short- and long-term safety remain to be elucidated.

Targeting molecular aberrations using gene transfer therapy, mostly with an adenoviral vector, has been tested in HFrEF. A cellular target includes calcium cycling proteins such as inhibitors of phospholamban such as SERCA2a which is deficient in patients with HFrEF. Primarily responsible for reincorporating calcium into the sarcoplasmic reticulum during diastole, this target was tested in the CUPID (Efficacy and Safety Study of Genetically Targeted Enzyme Replacement Therapy for Advanced Heart Failure) trial. This study used coronary arterial infusion of adeno-associated virus type 1 carrying the gene for SERCA2a and initially demonstrated that natriuretic peptides were decreased, reverse remodeling was noted, and symptomatic improvements were forthcoming. However, a confirmatory trial failed to meet its primary efficacy endpoint.

More advanced therapies for late-stage heart failure such as left ventricular assist devices and cardiac transplantation are covered in detail in Chap. 255.

DISEASE MANAGEMENT AND SUPPORTIVE CARE

Despite stellar outcomes with medical therapy, admission rates following heart failure hospitalization remain high, with nearly half of all patients readmitted to hospital within 6 months of discharge. Recurrent heart failure and related cardiovascular conditions account for only half of readmissions in patients with heart failure, whereas other comorbidity-related conditions account for the rest. The key to achieving enhanced outcomes must begin with the attention to transitional care at the index hospitalization with facilitated discharge through comprehensive discharge planning, patient and caregiver education, appropriate use of visiting nurses, and planned follow-up. Early postdischarge follow-up, whether by telephone or clinic-based, may be critical to ensuring stability because most heart failure–related readmissions tend to occur within the first 2 weeks after discharge. Although routinely advocated, intensive surveillance of weight and vital signs with use of telemonitoring has not decreased hospitalizations. Intrathoracic impedance measurements have been advocated for the identification of early rise in filling pressure and worsened hemodynamics so that preemptive management may be employed. However, this has not been successful and may worsen outcomes in the short term. Implantable pressure monitoring systems do tend to provide signals for early decompensation, and in patients with moderately advanced symptoms, such systems have been shown to provide information that can allow implementation of therapy to avoid hospitalizations by as much as 39% (in the CardioMEMS Heart Sensor Allows Monitoring of Pressure to Improve Outcomes in NYHA Class III Heart Failure Patients [CHAMPION] trial). Once heart failure becomes advanced, regularly scheduled review of the disease course and options with the patient and family is recommended including discussions surrounding end-of-life preferences when patients are comfortable in an outpatient setting. As the disease state advances further, integrating care with social workers, pharmacists, and community-based nursing may be critical in improving patient satisfaction with the therapy, enhancing QOL, and avoiding heart failure hospitalizations. Equally important is attention to seasonal influenza vaccinations and periodic pneumococcal vaccines that may obviate non–heart failure hospitalizations in these ill patients. When nearing end of life, facilitating a shift in priorities to outpatient and hospice palliation is key, as are discussions around advanced therapeutics and continued use of ICD prophylaxis, which may worsen QOL and prolong death.

GLOBAL CONSIDERATIONS


Substantial differences exist in the practice of heart failure therapeutics and outcomes by geographic location. The penetrance of CRT and ICD is higher in the United States than in Europe. Conversely, therapy unavailable in the United States, such as levosimendan, is designated as useful in Europe. Variation in the benefits of beta blockers based on world region remains an area of controversy. In oral pharmacologic therapy trials of HFrEF, patients from southwest Europe have a lower incidence of ischemic cardiomyopathy and those in North America tend to have more diabetes and prior coronary revascularization. There is also regional variation in medication use even after accounting for indication. In trials of heart failure, disparate effects are noted across populations. As a recent example, in TOPCAT, the drug spironolactone was effective when used in the US population while patients recruited from Russia and contiguous territories showed no difference. Whether this represents population differences or trial conduct disparity remains to be investigated. ADHF, patients in Eastern Europe tend to be younger, with higher ejection fractions and lower natriuretic peptide levels. Patients from South America tend to have the lowest rates of comorbidities, revascularization, and device use. In contrast, patients from North America have the highest comorbidity burden with high revascularization and device use rates. Given geographic differences in baseline characteristics and clinical outcomes, the generalizability of therapeutic outcomes in patients in the United States and Western Europe may require verification.

FURTHER READING

Braunwald E: Heart failure. JACC Heart Fail 1:1, 2013. [PubMed: 24621794]

Braunwald E: The war against heart failure: The Lancet lecture. Lancet 385:812, 2015. [PubMed: 25467564]

Cowie MR et al: Adaptive servo-ventilation for central sleep apnea in systolic heart failure. N Engl J Med 373:1095, 2015. [PubMed: 26323938]

Kusumoto FM et al: HRS/ACC/AHA expert consensus statement on the use of implantable cardioverter-defibrillator therapy in patients who are not included or not well represented in clinical trials. Circulation 130:94, 2014. [PubMed: 24815500]

McMurray JJ et al: PARADIGM-HF Investigators and Committees. Angiotensin-neprilysin inhibition versus enalapril in heart failure. N Engl J Med. 371:993, 2014. [PubMed: 25176015]

Ramani GV et al: Chronic heart failure: Contemporary diagnosis and management. Mayo Clin Proc 85:180, 2010. [PubMed: 20118395]

Redfield MM et al: NHLBI Heart Failure Clinical Research Network. Isosorbide mononitrate in heart failure with preserved ejection fraction. N Engl J Med 373:2314, 2015. [PubMed: 26549714]

Shah SJ et al: Phenotype-specific treatment of heart failure with preserved ejection fraction: A multiorgan roadmap. Circulation 134:73, 2016. [PubMed: 27358439]

Velazquez EJ et al: STICHES Investigators. Coronary-artery bypass surgery in patients with ischemic cardiomyopathy. N Engl J Med 374:1511, 2016. [PubMed: 27040723]

Zinman B et al: EMPA-REG OUTCOME investigators. Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. N Engl J Med 373:2117, 2015. [PubMed: 26378978]

Chapter 254: Cardiomyopathy and Myocarditis

Neal K. Lakdawala; Lynne Warner Stevenson; Joseph Loscalzo

DEFINITION AND CLASSIFICATION

Cardiomyopathy is disease of the heart muscle. It is estimated that cardiomyopathy accounts for 5–10% of the heart failure in the 5–6 million patients carrying that diagnosis in the United States. This term is intended to exclude cardiac dysfunction that results from other structural heart disease, such as coronary artery disease, primary valve disease, or severe hypertension; however, in general usage, the phrase *ischemic cardiomyopathy* is sometimes applied to describe diffuse dysfunction attributed to multivessel coronary artery disease, and *nonischemic cardiomyopathy* to describe cardiomyopathy from other causes. As of 2013, cardiomyopathies are defined as “disorders characterized by morphologically and functionally abnormal myocardium in the absence of any other disease that is sufficient, by itself, to cause the observed phenotype.” *It was further specified that many cardiomyopathies will be attributable to genetic disease.*[1]

The traditional classification of cardiomyopathies into a triad of dilated, restrictive, and hypertrophic was based initially on autopsy specimens and later on echocardiographic findings. Dilated and hypertrophic cardiomyopathies can be distinguished on the basis of left ventricular wall thickness and cavity dimension; however, restrictive cardiomyopathy can have variably increased wall thickness and chamber dimensions that range from reduced to slightly increased, with prominent atrial enlargement. Restrictive cardiomyopathy is now defined more on the basis of abnormal diastolic function, which is also present but initially less prominent in dilated and hypertrophic cardiomyopathy. Restrictive cardiomyopathy can overlap in presentation, gross morphology, and etiology with both hypertrophic and dilated cardiomyopathies **(Table 254-1)**.

TABLE 254-1

Presentation with Symptomatic Cardiomyopathy

	Dilated	Restrictive	Hypertrophic
Ejection fraction (normal >55%)	Usually <30% when symptoms severe	25–50%	>60%
Left ventricular diastolic dimension (normal <55 mm)	≥60 mm	<60 mm (may be decreased)	Often decreased
Left ventricular wall thickness	Normal or decreased	Normal or increased	Markedly increased
Atrial size	Increased, may also be primarily affected	Increased; may be massive	Increased; related to elevated filling pressures
Valvular regurgitation	Related to annular dilation; mitral appears earlier during decompensation; tricuspid regurgitation with right ventricular dysfunction	Related to endocardial involvement; frequent mitral and tricuspid regurgitation, rarely severe	Related to valve-septum interaction; mitral regurgitation
Common first symptoms	Exertional intolerance	Exertional intolerance, fluid retention early, may have dominant right-sided symptoms	Exertional intolerance; may have chest pain
Congestive symptoms[a]	Left before right, except right prominent in young adults	Right often dominates	Left-sided congestion at rest may develop late
Arrhythmias	Ventricular tachyarrhythmia; conduction block in Chagas' disease, and some families. Atrial fibrillation.	Ventricular uncommon except in sarcoidosis, conduction block in sarcoidosis and amyloidosis. Atrial fibrillation.	Ventricular tachyarrhythmias; atrial fibrillation

[a]Left-sided symptoms of pulmonary congestion: dyspnea on exertion, orthopnea, paroxysmal nocturnal dyspnea. Right-sided symptoms of systemic venous congestion: hepatic and abdominal distention, discomfort on bending, peripheral edema.

Expanding information renders this classification triad based on phenotype increasingly inadequate to define disease or therapy. Identification of more genetic determinants of cardiomyopathy has suggested a four-way classification scheme of etiology as primary (affecting primarily the heart) and secondary to other systemic disease. The primary causes are then divided into genetic, mixed genetic and acquired, and acquired. In practice however, genetic information is rarely available at initial presentation, the phenotypic expression of a given mutation varies widely, and genetic predisposition also influences acquired cardiomyopathies. Although the proposed genetic classification does not yet guide many current clinical strategies, it will become increasingly relevant as classification of disease moves beyond individual organ pathology to more integrated systems approaches.

[1]From E Arbustini et al: J Am Coll Cardiol 62:2046, 2013.

GENERAL PRESENTATION

For all cardiomyopathies, the early symptoms often relate to exertional intolerance with breathlessness or fatigue, usually from inadequate cardiac reserve during exercise. These symptoms may initially be unnoticed or attributed to other causes, commonly lung disease or "getting older." As fluid retention leads to elevation of resting filling pressures, shortness of breath may occur during routine daily activity such as dressing and may manifest as dyspnea or cough when lying down at night. Although often considered the hallmark of congestion, peripheral edema may be absent despite severe fluid retention, particularly in younger patients in whom ascites and abdominal discomfort may dominate. The nonspecific term *congestive heart failure* describes only the resulting syndrome of fluid retention, which is common to all three types of cardiomyopathy and also to cardiac structural diseases associated with elevated filling pressures. All three types of cardiomyopathy can be associated with atrioventricular (AV) valve regurgitation, typical and atypical chest pain, atrial and ventricular tachyarrhythmias, and embolic events (Table 254-1). Initial evaluation begins with a detailed clinical history and examination, looking for clues to cardiac, extracardiac, and familial disease (**Table 254-2**).

TABLE 254-2

Initial Evaluation of Cardiomyopathy

Clinical Evaluation
Thorough history and physical examination to identify cardiac and noncardiac disorders[a]
Detailed family history of heart failure, cardiomyopathy, skeletal myopathy, conduction disorders, tachyarrhythmias, and sudden death
History of alcohol, illicit drugs, chemotherapy or radiation therapy[a]
Assessment of ability to perform routine and desired activities[a]
Assessment of volume status, orthostatic blood pressure, body mass index[a]
Laboratory Evaluation
Electrocardiogram[a]
Chest radiograph[a]
Two-dimensional and Doppler echocardiogram[a]
Magnetic resonance imaging for evidence of myocardial inflammation and fibrosis
Chemistry:
Serum sodium,[a] potassium,[a] calcium,[a] magnesium[a]
Fasting glucose (glycohemoglobin in diabetes mellitus)
Creatinine,[a] blood urea nitrogen[a]
Albumin,[a] total protein,[a] liver function tests[a]
Lipid profile
Thyroid-stimulating hormone[a]
Serum iron, transferrin saturation
Urinalysis
Creatine kinase isoforms
Cardiac troponin levels
Hematology:
Hemoglobin/hematocrit[a]
White blood cell count with differential,[a] including eosinophils
Erythrocyte sedimentation rate
Initial Evaluation When Specific Diagnoses Are Suspected
DNA sequencing for genetic disease, panel selection based on phenotype

Titers for infection in the setting of clinical suspicion:
Acute viral (coxsackie, echovirus, influenza)
Human immunodeficiency virus
Chagas' (*Trypanosoma cruzi*), Lyme (*Borrelia burgdorferi*), toxoplasmosis
Catheterization with coronary angiography in patients with angina who are candidates for intervention[a]
Serologies for active rheumatologic disease
Endomyocardial biopsy including sample for electron microscopy when suspecting specific diagnosis with therapeutic implications
Screening for sleep-disordered breathing

[a]Level I recommendations from ACC/AHA Practice Guidelines for Chronic Heart Failure in the Adult.

GENETIC CAUSES OF CARDIOMYOPATHY



Estimates for the prevalence of genetic etiology of cardiomyopathy continue to rise, with increasing availability of genetic testing and attention to the family history. Well-recognized in hypertrophic cardiomyopathy, heritability is also present in at least 30% of dilated cardiomyopathy (DCM) without other clear etiology. Careful family history should elicit not only known cardiomyopathy and heart failure, but also family members who have had sudden death, often incorrectly attributed to "a massive heart attack," who have had atrial fibrillation or pacemaker implantation by middle age, or who have muscular dystrophy.

Most familial cardiomyopathies are inherited in an autosomal dominant pattern, with occasional autosomal recessive and X-linked inheritance **(Table 254-3)**. Missense mutations with amino acid substitutions are the most common in cardiomyopathy. Expressed mutant proteins may interfere with function of the normal allele through a dominant negative mechanism. Mutations introducing a premature stop codon (nonsense) or shift in the reading frame (frameshift) may create a truncated or unstable protein the lack of which causes cardiomyopathy (haploinsufficiency). Deletions or duplications of an entire exon or gene are uncommon causes of cardiomyopathy, except for the dystrophinopathies.

TABLE 254-3

Selected Genetic Defects Associated with Cardiomyopathy

	Gene Product	Inheritance	Cardiac Phenotype	Isolated Cardiac Phenotypea	Extracardiac Manifestations
Sarcomere	*MYH7* (β myosin heavy chain)	AD	HCM, DCM, LVNC	Yes	Skeletal myopathy
	MYBPC3 (myosin binding protein C)	AD	HCM	Yes	
	TNNT2 (cardiac troponin T)	AD	HCM, DCM, LVNC	Yes	
	TNNI3 (cardiac troponin I)	AD, AR	HCM, DCM, RCM	Yes	
	TTN (Titin)	AD	DCM	Yes	
	TPM1 (α-tropomyosin)	AD	HCM, DCM	Yes	
	TNNC1 (cardiac troponin C)	AD	DCM	Yes	
	MYL2 (myosin regulatory light chain)	AD	HCM	Yes	Skeletal myopathy
	MYL3 (myosin essential light chain)	AD	HCM	Yes	
Z-disk and Cytoskeleton	*DES* (Desmin)	AD	DCM, RCM	Yes	Skeletal myopathy
	ANKRD1 (CARP)	AD	HCM, (DCM)	Yes	
	CSRP3 (MLP)	AD	DCM, (HCM)	Yes	
	ACTN2 (α-actinin-2)	AD	DCM	Yes	
	CRYAB (αB-crystallin)	AD	DCM	Yes	
	FLNC (Filamin C)	AD	DCM	Yes	Skeletal myopathy
Nuclear Membrane	*LMNA* (Lamin A/C)	AD, AR	CDDC	Yes	Skeletal myopathy
	EMD (Emerin)	X-linked	CDDC	No	Skeletal myopathy, contractures
Excitation-Contraction Coupling	*PLN* (Phospholamban)	AD	DCM	Yes	
	SCN5A (NAV 1.5)	AD	CDDC	Yes	Note other mutations associated with Brugada syndrome
	RYR2 (cardiac ryanodine receptor)	AD	ARVC	Yes	
	CASQ2 (calsequestrin 2)	AR	ARVC	Yes	
Cellular Metabolism	*PRKAG2* (γ-subunit of AMP kinase)	AD	HCM+	Yes	

	Gene Product	Inheritance	Cardiac Phenotype	Isolated Cardiac Phenotypea	Extracardiac Manifestations
	LAMP2 (lysosomal associated membrane protein)	X-linked	HCM+	No[a]	Danon's disease: skeletal myopathy, cognitive impairment
	TAZ (Tafazzin)	X-linked	DCM, LVNC	No	Barth's syndrome: skeletal myopathy, cognitive impairment, neutropenia
	FXN (Frataxin)	AR	HCM	No	Friedreich's ataxia: ataxia, diabetes mellitus type 2
	TMEM43 (transmembrane protein 43)	AD	ARVC	Yes	
	GLA (α-galactosidase-A)	X-linked	HCM+	Yes	Fabry's disease: renal failure, angiokeratomas and painful neuropathy
Mitochondria	Mitochondrial DNA	Maternal transmission	DCM, HCM	No	MELAS, MERRF, Kearns-Sayre syndrome, ocular myopathy
Sarcolemmal Membrane	*DMD* (Dystrophin)	X-linked	DCM	No[a]	Duchenne's and Becker's muscular dystrophy
	DMPK (dystrophica myotonica protein kinase)	AD	DCM	No	Myotonic dystrophy type 1
	SGCD (δ-sarcoglycan)	AD	DCM	Yes	
Desmosome	*DSP* (Desmoplakin) *JUP* (Plakoglobin)	AD, AR	ARVC	Yes	Carvajal syndrome (AR), Naxos syndrome (AR), "woolly hair" and hyperkeratosis of palms and soles
	DSG2 (Desmoglein 2) *DSC2* (Desmocollin 2) *PKP2* (Plakophilin 2)	AD	ARVC	Yes	
Other Examples	*RBM20* (RNA binding motif 20)	AD	DCM	Yes	
	PSEN1 (Presenilin-1,2)	AD	DCM	Yes	Dementia
	BAG3 (BCL2-associated athanogene 3)	AD	DCM	Yes	
	ALPK3 (Alpha-kinase 3)	AR	HCM	Yes	

[a]Indicates that the usual clinical presentation is of isolated cardiomyopathy, however occasionally present extra cardiac manifestations are also provided. [a]Indicates that isolated cardiac phenotype can occur in women with the X-linked defects.

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; ARVC, arrhythmogenic right ventricular cardiomyopathy; CDDC, conduction disease with dilated cardiomyopathy; DCM, dilated cardiomyopathy; HCM, hypertrophic cardiomyopathy; HCM+, HCM with preexcitation; HCMc, HCM with conduction disease; LVNC, left ventricular noncompaction; MELAS, (mitochondrial) myopathy, encephalopathy, lactic acidosis, and strokelike episodes syndrome; MERRF, myoclonic epilepsy with ragged red fibers; RCM, restrictive cardiomyopathy.

Many different genes have been implicated in human cardiomyopathy (locus heterogeneity), and many mutations within those genes have been associated with disease (allelic heterogeneity). Although most identified mutations are "private" to individual families, several specific mutations are found repeatedly, either due to a founder effect or recurrent mutations at a common residue.

Genetic cardiomyopathy is characterized by age-dependent and incomplete penetrance. The defining phenotype of cardiomyopathy is rarely present at birth and, in some individuals, may never manifest. Related individuals who carry the *same* mutation may differ in the severity and rate of progression of cardiac dysfunction and associated rhythm disorders, indicating the important role of other genetic, epigenetic, and environmental modifiers in disease

expression. Sex appears to play a role, as penetrance and clinical severity may be greater in men for most cardiomyopathies. Clinical disease expression is generally more severe in the 3–5% of individuals who harbor two or more mutations linked to cardiomyopathy. However, the clinical course of a patient usually cannot be predicted based on which mutation is present; thus, current therapy is based on the phenotype rather than the genetic defect. Currently, the greatest utility of genetic testing for cardiomyopathy is to inform family evaluations. However, genetic testing occasionally enables the detection of a disease for which specific therapy is indicated, such as the replacements for defective metabolic enzymes in Fabry's disease and Gaucher disease.

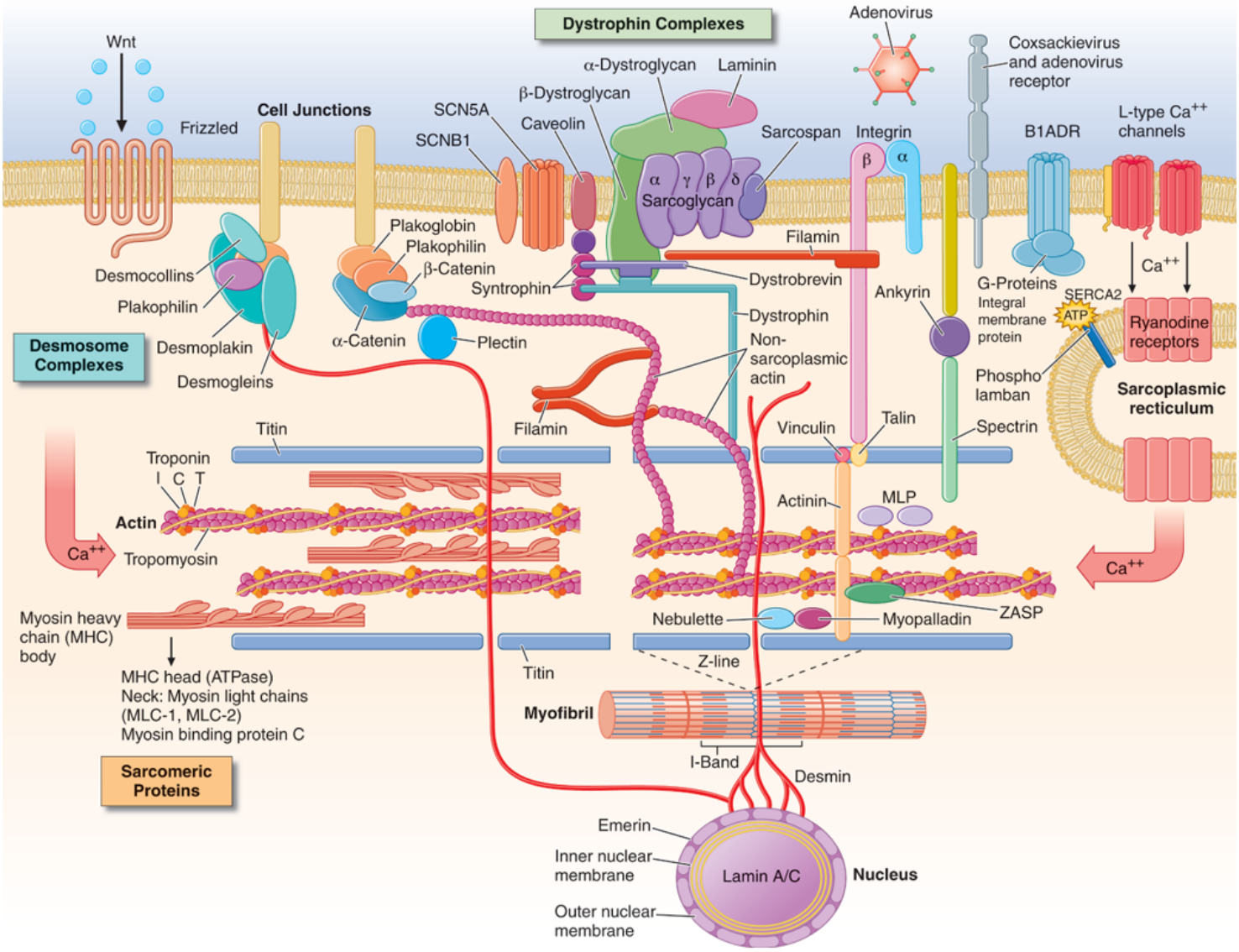
GENES AND PATHWAYS IN CARDIOMYOPATHY

Mutations in sarcomeric genes, encoding the thick and thin myofilament proteins, are the best characterized. While the majority are associated with hypertrophic cardiomyopathy, an increasing number of sarcomeric mutations have now been implicated in DCM, and some in left ventricular noncompaction. The most commonly recognized genetic causes of DCM are truncating mutations of the giant protein titin, encoded by *TTN*, which maintains sarcomere structure and acts as a key signaling molecule.

As cytoskeletal proteins play crucial roles in the structure, connection, and stability of the myocyte, multiple defects in these proteins can lead to cardiomyopathy, usually with a dilated phenotype (**Fig. 254-1**). For example, desmin forms intermediate filaments that connect the nuclear and plasma membranes, Z-lines, and the intercalated disks between muscle cells. Desmin mutations impair the transmission of force and signaling for both cardiac and skeletal muscle and may cause combined cardiac and skeletal myopathy.

FIGURE 254-1

Drawing of myocyte indicating multiple sites of abnormal gene products associated with cardiomyopathy. Major functional groups include the sarcomeric proteins (actin, myosin, tropomyosin, and the associated regulatory proteins), the dystrophin complex stabilizing and connecting the cell membrane to intracellular structures, the desmosome complexes associated with cell-cell connections and stability, and multiple cytoskeletal proteins that integrate and stabilize the myocyte. ATP, adenosine triphosphate. *(Figure adapted from Jeffrey A. Towbin, MD, University of Tennessee Health Science Center, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Defects in the sarcolemmal membrane proteins are associated with DCM. The best known is dystrophin, encoded by the X chromosome gene *DMD*, abnormalities of which cause Duchenne's and Becker's muscle dystrophy. (Interestingly, abnormal dystrophin can be acquired when the coxsackie virus cleaves dystrophin during viral myocarditis.) This protein provides a network that supports the sarcolemma and also connects to the sarcomere. The progressive functional defect in both cardiac and skeletal muscle reflects vulnerability to mechanical stress. Dystrophin is associated at the membrane with a complex of other proteins, such as metavinculin, abnormalities of which also cause DCM. Defects in the sarcolemmal channel proteins (*channelopathies*) are generally associated with primary arrhythmias, but mutations in *SCN5A*, distinct from those that cause the Brugada or long QT syndromes, have been implicated in DCM with conduction disease.

Nuclear membrane protein defects in cardiac and skeletal muscle occur in either autosomal (lamin A/C) or X-linked (emerin) patterns. These defects are associated with a high prevalence of atrial arrhythmias and conduction system disease, which can occur in some family members without or before detectable cardiomyopathy.

Intercalated disks contribute to intracellular connections, allowing mechanical and electrical coupling between cells and also connections to desmin filaments within the cell. Mutations in proteins of the desmosomal complex compromise attachment of the myocytes, which can become disconnected and die, to be replaced by fat and fibrous tissue. These areas are highly arrhythmogenic and may dilate to form aneurysms. Although more often noted in the right ventricle (arrhythmogenic right ventricular cardiomyopathy), this condition can affect both ventricles and has also been termed "arrhythmogenic cardiomyopathy."

As many signaling pathways are conserved over multiple systems, we anticipate discovering extracardiac manifestations of abnormal proteins initially considered restricted to the heart. In contrast, the monogenic disorders of metabolism that affect the heart are already clearly recognized to affect multiple organ systems. Currently, it is most important to diagnose defective enzymes for which specific enzyme replacement therapy can now ameliorate the course of disease, such as with alpha-galactosidase A deficiency (Fabry's disease). Abnormalities of mitochondrial DNA (maternally transmitted) impair

energy production with multiple clinical manifestations, including impaired cognitive function and skeletal myopathy. The phenotypic expression is highly variable depending on the distribution of the maternal mitochondria during embryonic development. Heritable systemic diseases, such as familial amyloidosis and hemochromatosis, can affect the heart without mutation of genes expressed in the heart.

For any patient with suspected or proven genetic disease, family members should be considered and evaluated in a longitudinal fashion. Screening includes an echocardiogram and electrocardiogram (ECG). The indications and implications for confirmatory specific genetic testing vary depending on the specific mutation. The profound questions raised by families about diseases shared and passed down merit serious and sensitive discussion, ideally provided by a trained genetic counselor.

DILATED CARDIOMYOPATHY

An enlarged left ventricle with reduced systolic function as measured by left ventricular ejection fraction characterizes DCM (**Figs. 254-2**, **254-3**, **and 254-4**). *Systolic failure* is more prominent than diastolic dysfunction. Although the syndrome of DCM has many disparate etiologies (**Table 254-4**), many converge to common pathways of secondary response and disease progression. When myocardial injury is acquired, some myocytes may die initially, whereas others survive only to have later programmed cell death (apoptosis), and remaining myocytes hypertrophy in response to increased wall stress. Local and circulating factors stimulate deleterious secondary responses that contribute to progression of disease. Dynamic remodeling of the interstitial scaffolding affects diastolic function and the amount of ventricular dilation. Mitral regurgitation commonly develops as the valvular apparatus is distorted and is usually substantial by the time heart failure is severe. Many cases that present “acutely” have progressed silently through these stages over months to years. Dilation and decreased function of the right ventricle may result directly from the initial injury, but more often develops later in response to elevated afterload presented by secondary pulmonary hypertension and in relation to mechanical interactions with the failing left ventricle.

FIGURE 254-2

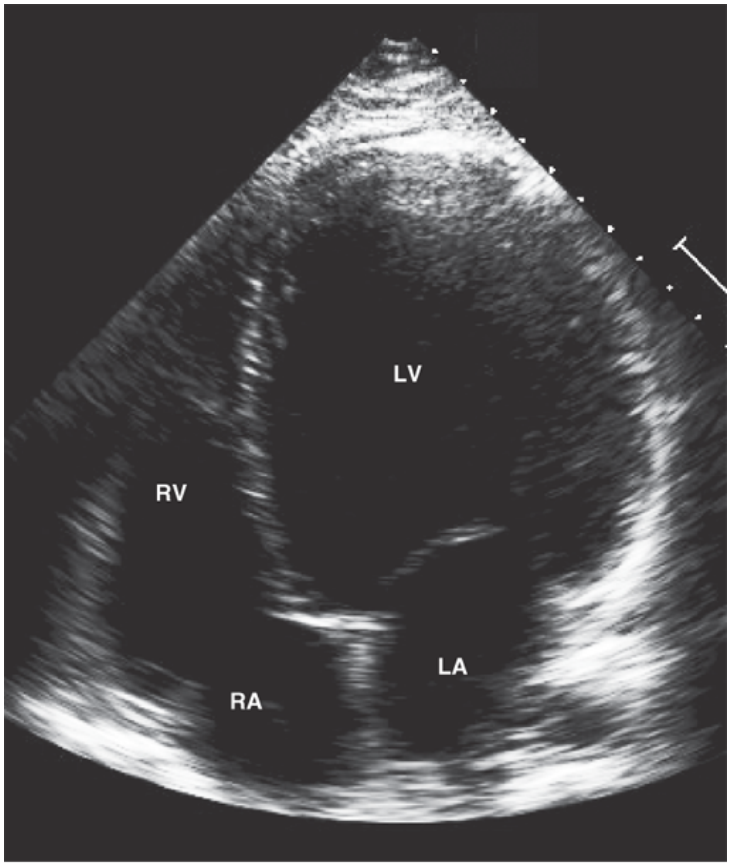
Dilated cardiomyopathy. This gross specimen of a heart removed at the time of transplantation shows massive left ventricular dilation and moderate right ventricular dilation. Although the left ventricular wall in particular appears thinned, there is significant hypertrophy of this heart, which weighs >800 g (upper limit of normal = 360 g). A defibrillator lead is seen traversing the tricuspid valve into the right ventricular apex. (*Image courtesy of Robert Padera, MD, PhD, Department of Pathology, Brigham and Women’s Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 254-3

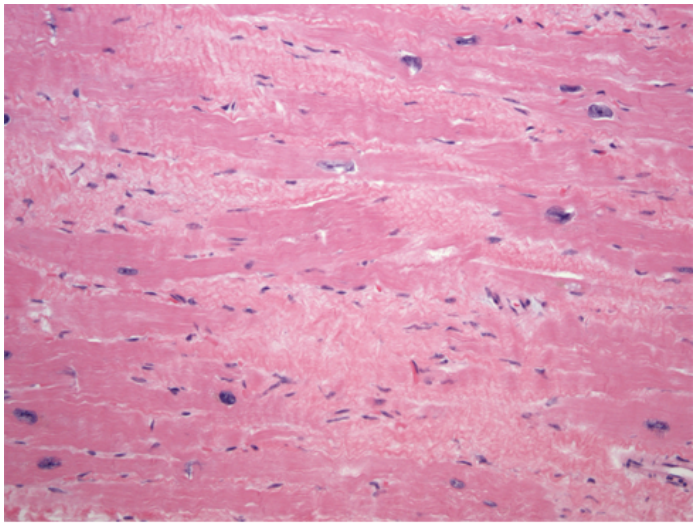
Dilated cardiomyopathy. This echocardiogram of a young man with dilated cardiomyopathy shows massive global dilation and thinning of the walls of the left ventricle (LV). The left atrium (LA) is also enlarged compared to normal. Note that the echocardiographic and pathologic images are vertically opposite, such that the LV is by convention on the top right in the echocardiographic image and bottom right in the pathologic images. RA, right atrium; RV, right ventricle. *(Image courtesy of Justina Wu, MD, Brigham and Women’s Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 254-4

Dilated cardiomyopathy. Microscopic specimen of a dilated cardiomyopathy showing the nonspecific changes of interstitial fibrosis and myocyte hypertrophy characterized by increased myocyte size and enlarged, irregular nuclei. Hematoxylin and eosin–stained section, 100× original magnification. *(Image courtesy of Robert Padera, MD, PhD, Department of Pathology, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 254-4

Major Causes of Dilated Cardiomyopathy (with Common Examples)

Inflammatory Myocarditis
Infective
Viral (coxsackie,[a] adenovirus,[a] HIV, hepatitis C)
Parasitic (*T. cruzi*—Chagas' disease, trypanosomiasis, toxoplasmosis)
Bacterial (diphtheria)
Spirochetal (*Borrelia burgdorferi*—Lyme disease)
Rickettsial (Q fever)
Fungal (with systemic infection)
Noninfective
Granulomatous inflammatory disease
Sarcoidosis
Giant cell myocarditis
Eosinophilic myocarditis
Polymyositis, dermatomyositis
Collagen vascular disease
Checkpoint inhibitor chemotherapy
Transplant rejection
Toxic
Alcohol
Catecholamines: amphetamines, cocaine
Chemotherapeutic agents (anthracyclines, trastuzumab)
Interferon
Other therapeutic agents (hydroxychloroquine, chloroquine)
Drugs of misuse (emetine, anabolic steroids)
Heavy metals: lead, mercury
Occupational exposure: hydrocarbons, arsenicals
Metabolic[a]
Nutritional deficiencies: thiamine, selenium, carnitine
Electrolyte deficiencies: calcium, phosphate, magnesium

Endocrinopathy
Thyroid disease
Pheochromocytoma
Diabetes
Obesity
Hemochromatosis
Inherited Metabolic Pathway Defects[a]
Familial[a] (See Table 254-3)
Skeletal and cardiac myopathy
Dystrophin-related dystrophy (Duchenne's, Becker's)
Mitochondrial myopathies (e.g., Kearns-Sayre syndrome)
Arrhythmogenic ventricular cardiomyopathy
Hemochromatosis
Associated with other systemic diseases
Susceptibility to immune-mediated myocarditis
Overlap with Nondilated Cardiomyopathy
"Minimally dilated cardiomyopathy"
Hemochromatosis[a]
Amyloidosis[a]
Hypertrophic cardiomyopathy[a] ("burned-out")
"Idiopathic"[a]
Miscellaneous (Shared Elements of Above Etiologies)
Peripartum cardiomyopathy
Left ventricular noncompaction[a]
Tachycardia-related cardiomyopathy
Supraventricular arrhythmias with uncontrolled rate
Very frequent nonsustained ventricular tachycardia or high premature ventricular complex burden

[a]Some specific cases can be linked now to specific genetic mutation in a familial cardiomyopathy; others with similar phenotypes that appear to be acquired or idiopathic may represent genetic factors not yet identified.

Regardless of the nature and degree of direct cell injury, the resulting functional impairment often reflects contribution from secondary responses that may be modifiable or reversible. Almost half of all patients with new-onset cardiomyopathy demonstrate substantial spontaneous recovery. Even with long-standing disease, some patients have dramatic improvement to near-normal ejection fractions during pharmacologic therapy, particularly notable with the β-adrenergic antagonists coupled with renin-angiotensin system inhibition. For patients in whom left bundle branch block precedes clinical heart failure by many years, cardiac resynchronization pacing may be particularly likely to improve ejection fraction and decrease ventricular size. Interest in the potential for recovery of cardiomyopathy has been further stimulated by occasional "recovery" of left ventricular function after prolonged mechanical circulatory support. The current evaluation and therapy for DCM is generally dictated by the stage of heart failure **(Chap. 252)**, with specific aspects discussed for relevant etiologies below.

MYOCARDITIS

Myocarditis (inflammation of the heart) can result from multiple causes but is most commonly attributed to infective agents that can injure the myocardium through direct invasion, production of cardiotoxic substances, or chronic inflammation with or without persistent infection. Myocarditis cannot be assumed from a presentation of decreased systolic function in the setting of an acute infection, as any severe infection causing systemic cytokine release can depress cardiac function transiently. Infectious myocarditis has been reported with almost all types of infective agents but is most commonly associated with viruses and the protozoan *Trypanosoma cruzi.*

INFECTIVE MYOCARDITIS

The pathogenesis of viral myocarditis has been extensively studied in murine models. After viruses gain entry through the respiratory or gastrointestinal tract, they can infect organs possessing specific receptors, such as the coxsackie-adenovirus receptor on the heart. Viral infection and replication can cause myocardial injury and lysis. For example, the enteroviral protease 2A facilitates viral replication and infection through degradation of the myocyte protein dystrophin, which is crucial for myocyte stability. Activation of viral receptor proteins can also activate host tyrosine kinases, which modify the cytoskeleton to facilitate further viral entry.

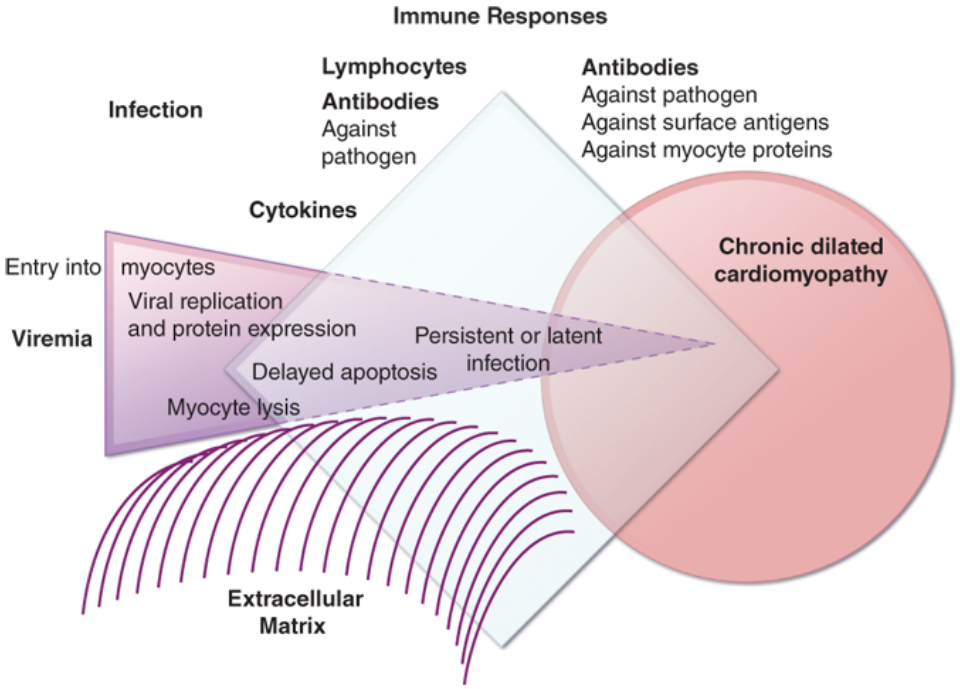
The first host response to infection is the nonspecific innate immune response, heavily dependent on Toll-like receptors that recognize common antigenic patterns. Cytokine release is rapid, followed by triggered activation and expansion of specific T- and B-cell populations. This initial response appears to be crucial, as early immunosuppression in animal models can increase viral replication and worsen cardiac injury. However, successful recovery from viral infection depends not only on the efficacy of the immune response to limit viral infection, but also on timely downregulation to prevent ongoing autoimmune injury to the host.

The secondary acquired immune response is specifically addressed against the viral proteins and can include both T-cell infiltration and antibodies to viral proteins. If unchecked, the acquired immune response can perpetuate secondary cardiac damage. Ongoing cytokine release activates matrix metalloproteinases that can disrupt the collagen and elastin scaffolding of the heart, potentiating ventricular dilation. Stimulation of pro-fibrotic factors leads to pathologic interstitial fibrosis. Some of the antibodies triggered through co-stimulation or molecular mimicry also recognize targets within the host myocyte, such as the β-adrenergic receptor, troponin, and Na^+/K^+ ATPase, but it remains unclear whether these antibodies contribute actively to cardiac dysfunction in humans or merely serve as markers of cardiac injury.

It is not known how long the viruses persist in the human heart, whether late persistence of the viral genome continues to be deleterious, or how often a dormant virus can again become pathogenic. Genomes of common viruses have frequently been detected in patients with clinical diagnoses of myocarditis or DCM, but there is little information on how often these are present in patients without cardiac disease (see below). Further information is needed to understand the relative timing and contribution of infection, immune responses, and secondary adaptations in the progression of heart failure after viral myocarditis **(Fig. 254-5)**.

FIGURE 254-5

Schematic diagram demonstrating the possible progression from infection through direct, secondary, and autoimmune responses to dilated cardiomyopathy. Most of the supporting evidence for this sequence is derived from animal models. It is not known to what degree persistent infection and/or ongoing immune responses contribute to ongoing myocardial injury in the chronic phase.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical Presentation of Viral Myocarditis

Acute viral myocarditis often presents with symptoms and signs of heart failure. Some patients present with chest pain suggestive of pericarditis or acute myocardial infarction. Occasionally, the presentation is dominated by atrial or ventricular tachyarrhythmias, or by pulmonary or systemic emboli from intracardiac thrombi. Electrocardiographic or echocardiographic abnormalities may also be detected incidentally during evaluation for other diagnoses. The typical patient with presumed viral myocarditis is a young to middle-aged adult who develops progressive dyspnea and weakness within a few days to weeks after a viral syndrome that was accompanied by fever and myalgias.

A small number of patients present with fulminant myocarditis, with rapid progression within hours from a severe febrile respiratory syndrome to cardiogenic shock that may involve multiple organ systems, leading to renal failure, hepatic failure, and coagulopathy. These patients are typically young adults who have recently been dismissed from urgent care settings with antibiotics for bronchitis or oseltamivir for viral syndromes, only to return within a few days in rapidly progressive cardiogenic shock. Prompt triage is vital to provide aggressive support with high-dose intravenous catecholamine therapy and sometimes with temporary mechanical circulatory support. Recognition of patients with this fulminant presentation is potentially life-saving as more than half can survive, with marked improvement demonstrable within the first few weeks. The ejection fraction function of these patients often recovers to near-normal, although residual diastolic dysfunction may limit vigorous exercise for some survivors.

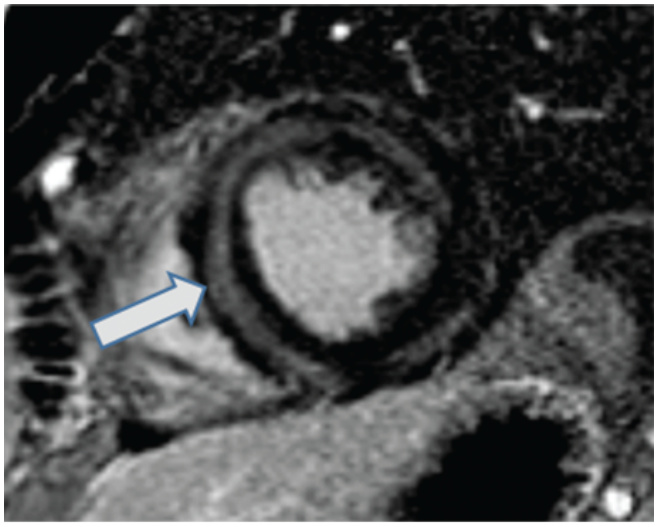
Chronic viral myocarditis is often invoked, but rarely proven, as a diagnosis when no other cause of DCM can be identified. However, many cases assumed to result from "silent" myocarditis will later be recognized as due to genetic causes or consumption of excess alcohol or illicit stimulant drugs. The proportion of chronic, DCM due to viral infection remains a subject of controversy.

Laboratory evaluation for myocarditis

The initial evaluation for suspected myocarditis includes an ECG, an echocardiogram, and serum levels of troponin and creatine phosphokinase fractions. Magnetic resonance imaging is increasingly used for the diagnosis of myocarditis, which is supported but not proven by evidence of increased tissue edema and gadolinium enhancement (**Fig. 254-6**), particularly in the mid-wall (as distinct from usual coronary artery territories).

FIGURE 254-6

Magnetic resonance image of myocarditis showing the typical mid-wall location (*arrow*) for late gadolinium enhancement from cardiac inflammation and scarring. *(Image courtesy of Ron Blankstein, MD, and Marcelo Di Carli, MD, Division of Nuclear Medicine, Brigham and Women's Hospital, Boston.)*

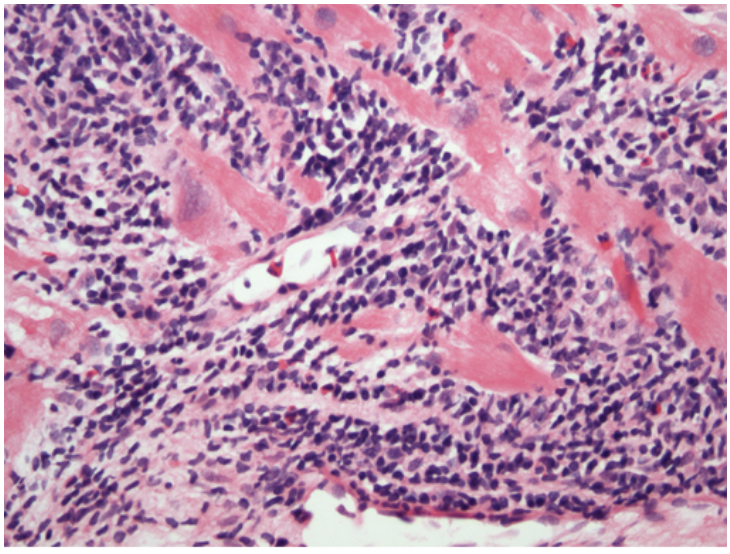


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Endomyocardial biopsy is not often indicated for the initial evaluation of suspected viral myocarditis unless ventricular tachyarrhythmias suggest possible etiologies of sarcoidosis or giant cell myocarditis. The indications, yield, and benefit of endomyocardial biopsy for evaluation of myocarditis or new-onset cardiomyopathy are not well-established. When biopsy is performed, the Dallas Criteria for myocarditis include lymphocytic infiltrate with evidence of myocyte necrosis **(Fig. 254-7)** and are negative in 80–90% of patients with clinical myocarditis. Negative Dallas Criteria can reflect sampling error or early resolution of lymphocytic infiltrates, but also the insensitivity of the test when inflammation results from cytokines and antibody-mediated injury. Routine histologic examination of endomyocardial biopsy rarely reveals a specific infective etiology, such as toxoplasmosis or *Cytomegalovirus*. Immunohistochemistry of myocardial biopsy samples is commonly used to identify active lymphocyte subtypes and may also detect upregulation of HLA antigens and the presence of complement components attributed to inflammation, but the specificity and significance of these findings are uncertain.

FIGURE 254-7

Acute myocarditis. Microscopic image of an endomyocardial biopsy showing massive infiltration with mononuclear cells and occasional eosinophils associated with clear myocyte damage. The myocyte nuclei are enlarged and reactive. Such extensive involvement of the myocardium would lead to extensive replacement fibrosis even if the inflammatory response could be suppressed. Hematoxylin and eosin–stained section, 200× original magnification. *(Image courtesy of Robert Padera, MD, PhD, Department of Pathology, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

An increase in circulating viral titers between acute and convalescent blood samples supports a diagnosis of acute viral myocarditis with potential spontaneous improvement. There is no established role for measuring circulating anti-heart antibodies, which may be the result, rather than a cause, of myocardial injury and have been found also in patients with coronary artery disease and genetic cardiomyopathy.

Patients with recent or ongoing viral syndromes have been classified into three levels of myocarditis diagnosis. (1) *Possible subclinical acute myocarditis* is diagnosed when a typical viral syndrome occurs without cardiac symptoms, but with elevated biomarkers of cardiac injury, ECG suggestive of acute injury, reduced left ventricular ejection fraction or regional wall motion abnormality. (2) *Probable acute myocarditis* is diagnosed when the above criteria are met

and accompanied by cardiac symptoms, such as shortness of breath or chest pain, which can result from pericarditis or myocarditis. When clinical findings of pericarditis are accompanied by elevated troponin or CK-MB or abnormal cardiac wall motion, the terms perimyocarditis or myopericarditis are sometimes used. (3) *Definite myocarditis* is diagnosed when there is histologic or immunohistologic evidence of inflammation on endomyocardial biopsy (see below) and does not require any other laboratory or clinical criteria. These have not been revised to include findings from MRI.

SPECIFIC VIRUSES IMPLICATED IN MYOCARDITIS

In humans, viruses are often suspected but rarely proven to be the direct cause of clinical myocarditis. First implicated was the picornavirus family of RNA viruses, principally the enteroviruses, coxsackie virus, echovirus, and poliovirus. Influenza, another RNA virus, is implicated with varying frequency every winter and spring as epitopes change. Of the DNA viruses, adenovirus, vaccinia (smallpox vaccine), and the herpesviruses (varicella zoster, cytomegalovirus, Epstein-Barr virus, and human herpesvirus 6 [HHV6]) are well-recognized to cause myocarditis but also occur commonly in the healthy population. Polymerase chain reaction (PCR) detects viral genomes in the majority of patients with DCM, but also in normal “control” hearts. Most often detected are parvovirus B19 and HHV6, which may affect the cardiovascular system, in part, through infection of vascular endothelial cells. However, their contribution to chronic cardiomyopathy is uncertain, as serologic evidence of exposure is present in many children and most adults.

Human immunodeficiency virus (HIV) was associated with an incidence of DCM of 1–2%; however, with the advent of highly active antiretroviral therapy (HAART), HIV has been associated with a significantly lower incidence of cardiac disease. Cardiomyopathy in HIV may result from cardiac involvement with other associated viruses, such as cytomegalovirus and hepatitis C, as well as by HIV directly. Antiviral drugs to treat chronic HIV can cause cardiomyopathy, both directly and through drug hypersensitivity. The clinical picture may be complicated by pericardial effusions and pulmonary hypertension. There is a high frequency of lymphocytic myocarditis found at autopsy, and viral particles have been demonstrated in the myocardium in some cases, consistent with direct causation.

Hepatitis C has been repeatedly implicated in cardiomyopathy, particularly in Germany and Asia. Cardiac dysfunction may improve after interferon therapy. As this cytokine itself often depresses cardiac function transiently, careful coordination of administration and ongoing clinical evaluation are critical. *The effect of new treatments for hepatitis C on cardiac function has not yet been well-studied*. Involvement of the heart with hepatitis B is uncommon, but can be seen when associated with systemic vasculitis (polyarteritis nodosa).

Additional viruses implicated specifically in myocarditis include *mumps, respiratory syncytial virus*, the *arboviruses (dengue fever and yellow fever),* and *arenaviruses (Lassa fever).* However, for any serious infection, the systemic inflammatory response can cause nonspecific depression of cardiac function, which is generally reversible if the patient survives.

THERAPY

There is currently no specific therapy recommended during any stage of viral myocarditis. During acute infection, therapy with anti-inflammatory or immunosuppressive medications is avoided, as their use has been shown to increase viral replication and myocardial injury in animal models. Therapy with specific antiviral agents (such as oseltamivir) has not been studied in relation to cardiac involvement. There is ongoing investigation into the impact of antiviral therapy to treat chronic viral persistence identified from endomyocardial biopsy. Large trials of immunosuppressive therapy for Dallas Criteria–positive myocarditis have been negative. There are some initial encouraging results and ongoing investigations with immunosuppressive therapy for immune-mediated myocarditis defined by immunohistologic criteria on biopsy or circulating anti-heart antibodies in the absence of myocardial viral genomes. However, neither antiviral nor anti-inflammatory therapies are currently recommended. Until we have a better understanding of the phases of viral myocarditis and the effects of targeted therapies, treatment will continue to be guided by general recommendations for DCM.

Parasitic Myocarditis

Chagas’ disease is the third most common parasitic infection in the world and the most common infective cause of cardiomyopathy. The protozoan *T. cruzi* is transmitted by the bite of the reduviid bug, endemic in the rural areas of South and Central America. Transmission can also occur through blood transfusion, organ donation, from mother to fetus, and occasionally orally. While programs to eradicate the insect vector have decreased the prevalence from about 16 million to <10 million in South America, cases are increasingly recognized in Western developed countries (see Global Perspectives below).

Multiple pathogenic mechanisms are implicated. The parasite itself can cause myocyte lysis and primary neuronal damage. Specific immune responses may recognize the parasites or related antigens and lead to chronic immune activation in the absence of detectable parasites. Molecular techniques have revealed persistent parasite DNA fragments in infected individuals. Further evidence for persistent infection is the eruption of parasitic skin lesions during immunosuppression after cardiac transplantation. As with viral myocarditis, the relative roles of persistent infection and of secondary autoimmune injury have not been resolved (Fig. 254-5). An additional factor in the progression of Chagas’ disease is the autonomic dysfunction and microvascular damage that may contribute to cardiac and gastrointestinal disease.

The acute phase of Chagas’ disease with parasitemia is usually unrecognized, but in fewer than 5% of cases, it presents clinically within a few weeks of infection, with nonspecific symptoms or occasionally with acute myocarditis and meningoencephalitis. In the absence of antiparasitic therapy, the silent stage progresses slowly for >10–30 years in almost half of patients to manifest chronically in the cardiac and gastrointestinal systems. Features typical of Chagas’ disease are conduction system abnormalities, particularly sinus node and AV node dysfunction and right bundle branch block. Atrial fibrillation

and ventricular tachyarrhythmias also occur. Small ventricular aneurysms are common, particularly at the ventricular apex. These dilated ventricles are particularly thrombogenic, giving rise to pulmonary and systemic emboli. Xenodiagnosis, detection of the parasite itself, is rarely performed. The serologic tests for specific IgG antibodies against the trypanosome lack sufficient specificity and sensitivity, requiring two separate positive tests required to make a diagnosis.

Treatment of the advanced stages focuses on clinical manifestations of the disease and includes heart failure medications, pacemaker-defibrillators, and anticoagulation. The most common antiparasitic therapies are benznidazole and nifurtimox which have been effective in children with chronic *T. cruzi* infection. Both drugs are associated with multiple severe reactions, including dermatitis, gastrointestinal distress, and neuropathy. Moreover, in a large trial of adults with established Chagas' cardiomyopathy, benznidazole did not prevent disease progression, leaving the role of antiparasitic therapy unclear. Survival is <30% at 5 years after the onset of overt clinical heart failure. Patients without major extracardiac disease have occasionally undergone transplantation, after which they require surveillance testing and recurrent antiparasitic therapy to suppress reactivation of infection.

African trypanosomiasis infection results from the tsetse fly bite and can occur in travelers exposed during trips to Africa. The West African form is caused by *Trypanosoma brucei gambiense* and progresses silently over years. The East African form caused by *T. brucei rhodesiense* can progress rapidly through perivascular infiltration to myocarditis and heart failure, with frequent arrhythmias. The diagnosis is made by identification of trypanosomes in blood, lymph nodes, or other affected sites. Antiparasitic therapy has limited efficacy and is determined by the specific type and the stage of infection (hemolymphatic or neurologic).

Toxoplasmosis is contracted through undercooked infected beef or pork, transmission from feline feces, organ transplantation, transfusion, or maternal-fetal transmission. Immunocompromised hosts are most likely to experience reactivation of latent infection from cysts, found in up to 40% of autopsies of patients dying from HIV infection. Toxoplasmosis may present with encephalitis or chorioretinitis and, in the heart, can cause myocarditis, pericardial effusion, constrictive pericarditis, and heart failure. The diagnosis in an immunocompetent patient is made when the IgM is positive and the IgG becomes positive later. Active toxoplasmosis may be suspected in an immunocompromised patient with myocarditis and a positive IgG titer for toxoplasmosis, particularly when avidity testing identifies high specificity of the antibody. Fortuitous sampling occasionally reveals the cysts in the myocardium. Combination therapy can include pyrimethamine and sulfadiazine or clindamycin.

Trichinellosis is caused by *Trichinella spiralis* larva ingested with undercooked meat. Larvae migrating into skeletal muscles cause myalgias, weakness, and fever. Periorbital and facial edema and conjunctival and retinal hemorrhage may also be seen. Although the larva may occasionally invade the myocardium, clinical heart failure is rare and, when observed, attributed to the eosinophilic inflammatory response. The diagnosis is made from the specific serum antibody and is further supported by the presence of eosinophilia. Treatment includes antihelminthic drugs (albendazole, mebendazole) and glucocorticoids if inflammation is severe.

Cardiac involvement with *Echinococcus* is rare, but cysts can form and rupture in the myocardium and pericardium.

Bacterial Infections

Most bacterial infections can involve the heart occasionally through direct invasion and abscess formation, but do so rarely. More commonly, systemic inflammatory responses depress contractility in severe infection and sepsis. *Diphtheria* specifically affects the heart in almost one-half of cases, and cardiac involvement is the most common cause of death in patients with this infection. The prevalence of vaccines has shifted the incidence of diphtheria from children worldwide to countries without routine immunization and to older populations who have lost their immunity. The bacillus releases a toxin that impairs protein synthesis and may particularly affect the conduction system. The specific antitoxin should be administered as soon as possible, with higher priority than antibiotic therapy. Other systemic bacterial infections that can involve the heart include *brucellosis, chlamydophila, legionella, meningococcus, mycoplasma, psittacosis*, and *salmonellosis,* for which specific treatment is directed at the systemic infection.

Clostridial infections cause myocardial damage from the released toxin. Gas bubbles can be detected in the myocardium, and occasionally abscesses can form in the myocardium and pericardium. *Streptococcal infection* with β-hemolytic streptococci is most commonly associated with acute rheumatic fever and is characterized by inflammation and fibrosis of cardiac valves and systemic connective tissue, but it can also lead to a myocarditis with focal or diffuse infiltrates of mononuclear cells.

Tuberculosis can involve the myocardium directly as well as through tuberculous pericarditis, but rarely does so when the disease is treated with antibiotics. *Whipple's disease* is caused by *Tropheryma whipplei.* The usual manifestations are in the gastrointestinal tract, but pericarditis, coronary arteritis, valvular lesions, and occasionally clinical heart failure may also occur. Multidrug antituberculous regimens are effective, but the disease tends to relapse even with appropriate treatment.

Other Infections

Spirochetal myocarditis has been diagnosed from myocardial biopsies containing *Borrelia burgdorferi* that causes *Lyme disease.* Lyme carditis most often presents with arthritis and conduction system disease that resolves within 1–2 weeks of antibiotic treatment, only rarely implicated in chronic heart failure. **Fungal myocarditis** can occur due to hematogenous or direct spread of infection from other sites, as has been described for aspergillosis, actinomycosis, blastomycosis, candidiasis, coccidioidomycosis, cryptococcosis, histoplasmosis, and mucormycosis. However, cardiac involvement is rarely the dominant

clinical feature of these infections. The **rickettsial infections,** *Q fever, Rocky Mountain spotted fever, and scrub typhus* are frequently accompanied by ECG changes, but most clinical manifestations relate to systemic vascular involvement.

NONINFECTIVE MYOCARDITIS

Myocardial inflammation can occur without apparent preceding infection. The paradigm of noninfective inflammatory myocarditis is cardiac transplant rejection, from which we have learned that myocardial depression can develop and reverse quickly, that noncellular mediators such as antibodies and cytokines play a major role in addition to lymphocytes, and that myocardial antigens are exposed by prior physical injury and viral infection.

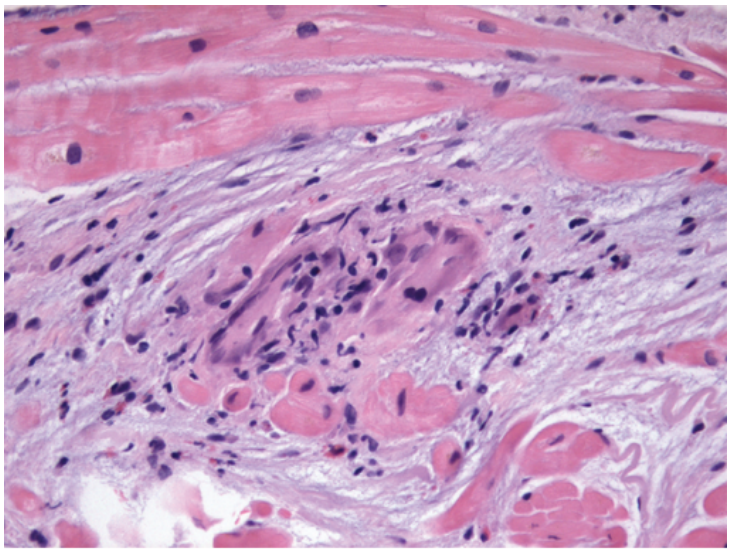
The most commonly diagnosed noninfective inflammation is granulomatous myocarditis, including both sarcoidosis and giant cell myocarditis. Sarcoidosis, as discussed in **Chap. 360**, is a multisystem disease most commonly affecting the lungs. Although classically presenting with higher prevalence in young African-American men, the epidemiology appears to be changing, with increasing recognition of sarcoidosis in Caucasian patients in nonurban areas. Patients with pulmonary sarcoid are at high risk for cardiac involvement, but cardiac sarcoidosis also occurs without clinical lung disease. Regional clustering of the disease supports the suspicion that the granulomatous reaction is triggered by an infectious or environmental allergen not yet identified.

The sites and density of cardiac granulomata, the time course, and the degree of extracardiac involvement are remarkably variable. Patients may present with rapid-onset heart failure and ventricular tachyarrhythmias, conduction block, chest pain syndromes, or minor cardiac findings in the setting of ocular involvement, an infiltrative skin rash, or a nonspecific febrile illness. They may also present less acutely after months to years of fluctuating cardiac symptoms. When ventricular tachycardia or conduction block dominates the initial presentation of heart failure without coronary artery disease, suspicion should be high for these granulomatous myocarditides.

Depending on the time course, the ventricles may appear restrictive or dilated. There is often right ventricular predominance of both dilation and ventricular arrhythmias, sometimes initially attributed to arrhythmogenic right ventricular cardiomyopathy. Small ventricular aneurysms are common. Computed tomography of the chest often reveals pulmonary lymphadenopathy even in the absence of clinical lung disease. Metabolic imaging (positron emission tomography [PET]) of the whole chest can highlight active sarcoid lesions that are avid for glucose. Magnetic resonance imaging (MRI) of the heart can identify areas likely to be inflammatory. To rule out chronic infections, such as tuberculosis or histoplasmosis as the cause of adenopathy, the diagnosis usually requires pathologic confirmation. Biopsy of enlarged mediastinal nodes may provide the highest yield. The scattered granulomata of sarcoidosis can easily be missed on cardiac biopsy **(Fig. 254-8)**.

FIGURE 254-8

Sarcoidosis. Microscopic image of an endomyocardial biopsy showing a noncaseating granuloma and associated interstitial fibrosis typical of sarcoidosis. No microorganisms were present on special stains, and no foreign material was identified. Hematoxylin and eosin–stained section, 200× original magnification. *(Image courtesy of Robert Padera, MD, PhD, Department of Pathology, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Immunosuppressive treatment for sarcoidosis is initiated with high-dose glucocorticoids, which are often more effective for arrhythmias than for the heart failure. Patients with sarcoid lesions that persist or recur during tapering of corticosteroids are considered candidates for other immunosuppressive therapies, frequently with agents also used for cardiac transplantation. Pacemakers and implantable defibrillators are generally indicated to prevent life-threatening heart block or ventricular tachycardia, respectively. Because the inflammation often resolves into extensive fibrosis that impairs cardiac function and provides pathways for reentrant arrhythmias, the prognosis for improvement is best when the granulomata are not extensive and the ejection fraction is not severely reduced.

Giant cell myocarditis is less common than sarcoidosis, but accounts for 10–20% of biopsy-positive cases of myocarditis. Giant cell myocarditis typically presents with rapidly progressive heart failure and tachyarrhythmias. Diffuse granulomatous lesions are surrounded by extensive inflammatory infiltrate unlikely to be missed on endomyocardial biopsy, often with eosinophilic infiltration. Associated conditions are thymomas, thyroiditis, pernicious anemia, other autoimmune diseases, and occasionally recent infections. Glucocorticoid therapy is less effective than for sarcoidosis and is sometimes combined with other immunosuppressive agents. The course is often of rapid deterioration requiring urgent mechanical support or transplantation. Although the severity of presentation and myocardial histology are more fulminant than with sarcoidosis, the occasional finding of giant cell myocarditis after sarcoidosis suggests that they may in some cases represent different stages of the same disease spectrum.

Eosinophilic myocarditis can be an important manifestation of the hypereosinophilic syndrome, which in Western countries is often considered idiopathic, although in Mediterranean and African countries, is associated with antecedent infection. It may also be seen with systemic eosinophilic syndromes such as Churg-Strauss syndrome or malignancies. *Hypersensitivity myocarditis* is often an unexpected diagnosis, made when the biopsy reveals infiltration with lymphocytes and mononuclear cells with a high proportion of eosinophils. Most commonly, the reaction is attributed to antibiotics, particularly those taken chronically, but thiazides, anticonvulsants, indomethacin, and methyldopa have also been implicated. Occasional associations with the smallpox vaccine have been reported. Although the circulating eosinophil count may be slightly elevated in hypersensitivity myocarditis, it does not reach the high levels of the hypereosinophilic syndrome. High-dose glucocorticoids and discontinuation of the trigger agent can be curative for hypersensitivity myocarditis. A severe lymphocytic myocarditis has been seen with combination of immune checkpoint inhibitors (see toxic cardiomyopathy below).

Myocarditis is often associated with systemic inflammatory diseases, such as *polymyositis* and *dermatomyositis*, which affect skeletal and cardiac muscle. Although noninfective inflammatory myocarditis is sometimes included in the differential diagnosis of cardiac findings in patients with connective tissue disease such as systemic lupus erythematosus, pericarditis, vasculitis, pulmonary hypertension, and accelerated coronary artery disease are more common cardiac manifestations of connective tissue disease.

PERIPARTUM CARDIOMYOPATHY

Peripartum cardiomyopathy (PPCM) develops during the last trimester or within the first 6 months after pregnancy, affecting between 1:2000 and 1:4000 deliveries in the United Sates. Risk factors are increased maternal age, increased parity, twin pregnancy, malnutrition, use of tocolytic therapy for premature labor, and preeclampsia or toxemia of pregnancy. Several of these risk factors contribute to anti-angiogenic signaling through secreted vascular endothelial growth factor (VEGF) inhibitors, such as soluble FLT1 (sFLT1). Recent animal and human studies have confirmed the role of decreased angiogenic reserve in the pathogenesis of PPCM, which may be rescued by correcting the angiogenic imbalance. Another recently proposed mechanism invokes an abnormal prolactin cleavage fragment, which is induced by oxidative stress and also affects angiogenesis; this observation has led to preliminary investigation of bromocriptine as possible therapy.

However, other processes also contribute to PPCM. Heart failure early after delivery was previously common in Nigeria, when the custom for new mothers included salt ingestion while reclining on a warm bed, which likely impaired mobilization of the excess circulating volume after delivery. In the Western world, lymphocytic myocarditis has sometimes been found on myocardial biopsy. This inflammation has been hypothesized to reflect increased susceptibility to viral myocarditis or an autoimmune myocarditis due to cross-reactivity of anti-uterine antibodies against cardiac muscle.

As the increased circulatory demand of pregnancy can aggravate other cardiac disease that was clinically unrecognized, it is crucial to the diagnosis of PPCM that there be no evidence for a preexisting cardiac disorder. By contrast, heart failure presenting earlier in pregnancy has been termed pregnancy-associated cardiomyopathy (PACM). Both PPCM and PACM have been found in some families with other presentations of DCM. As in familial and sporadic DCM, truncating mutations in *TTN* are present in 15% of patients with PPCM and are associated with systolic dysfunction that persists. Pregnancy may, thus, represent an environmental trigger for accelerated phenotypic expression of genetic and other cardiomyopathies.

TOXIC CARDIOMYOPATHY

Cardiotoxicity has been reported with multiple environmental and pharmacologic agents. Often these associations are seen only with very high levels of exposure or acute overdoses, in which acute electrocardiographic and hemodynamic abnormalities may reflect both direct drug effect and systemic toxicity.

Alcohol is the most common toxin implicated in chronic DCM. Excess consumption may contribute to more than 10% of cases of heart failure, including exacerbation of cases with other primary etiologies such as valvular disease or previous infarction. Toxicity is attributed both to alcohol and to its primary metabolite, acetaldehyde. Polymorphisms of the genes encoding alcohol dehydrogenase and the angiotensin-converting enzyme may influence the likelihood of alcoholic cardiomyopathy in an individual with excess consumption. Superimposed vitamin deficiencies and toxic alcohol additives are rarely implicated currently. The alcohol consumption necessary to produce cardiomyopathy in an otherwise normal heart has been estimated to be five to six drinks (about 4 ounces of pure ethanol) daily for 5–10 years, but frequent binge drinking may also be sufficient. Many patients with alcoholic cardiomyopathy are fully functional in their daily lives without apparent stigmata of alcoholism. The cardiac impairment in severe alcoholic cardiomyopathy is the sum of both permanent damage and a substantial component that is reversible after cessation of alcohol consumption. Atrial fibrillation occurs commonly both early in the disease (“holiday heart”) and in advanced stages. Medical therapy includes neurohormonal antagonists and diuretics as needed for fluid management. Withdrawal should be supervised to avoid exacerbations of heart failure or arrhythmias, and ongoing support arranged. Even with severe disease, marked improvement can occur within 3–6 months of abstinence. Implantable defibrillators are generally deferred

until an adequate period of abstinence, after which they may not be necessary if the ejection fraction has improved. With continued consumption, the prognosis is grim.

Cocaine, amphetamines, and related catecholaminergic stimulants can produce chronic cardiomyopathy as well as acute ischemia and tachyarrhythmias. Pathology reveals microinfarcts consistent with small vessel ischemia, similar to those seen with pheochromocytoma.

Chemotherapy agents are the most common drugs implicated in toxic cardiomyopathy. Judicious use of these drugs requires balancing the risks of the malignancy and the risks of cardiotoxicity, as many cancers have a chronic course with better prognosis than heart failure.

Anthracyclines (e.g., doxorubicin) cause characteristic histologic changes of vacuolar degeneration and myofibrillar loss. Generation of reactive oxygen species involving heme compounds is currently the favored explanation for myocyte injury and fibrosis. Risk for cardiotoxicity increases with higher doses, preexisting cardiac disease, extremes of age, concomitant chemotherapy, or chest irradiation and in women. Although cardiomyopathy has frequently been considered to occur late after exposure, a recent study shows that systolic dysfunction is usually evident within 1 year after anthracycline exposure among adult patients who develop cardiomyopathy. Doxorubicin cardiotoxicity generally results in minimal ventricular dilation, perhaps due to accompanying fibrosis. Thus, the stroke volume may be severely reduced with an ejection fraction of 30–40%, in contrast to the hemodynamic compensation possible in a dilated ventricle typical of other heart failure with reduced ejection fraction. Therapy includes angiotensin-converting enzyme inhibitors and β-adrenergic blocking agents, with careful suppression of “inappropriate” sinus tachycardia, and attention to postural hypotension that can occur in these patients. Once thought to have an inexorable downward course, many patients with doxorubicin cardiotoxicity improve with careful management to near-normal clinical function, particularly if additional insults such as hypertension or supraventricular tachycardias can be avoided. The course differs for patients receiving these drugs before puberty, in whom inadequate growth of the heart may lead to inexorable heart failure by the time the patient reaches the early twenties.

Trastuzumab (Herceptin) is a monoclonal antibody that interferes with human epidermal growth receptor 2 (HER2) crucial for some tumor growth and for cardiac adaptation. The incidence of cardiotoxicity is lower than for anthracyclines but enhanced by coadministration with them. Although considered to be more often reversible, trastuzumab cardiotoxicity does not always resolve, and some patients progress to clinical heart failure and death. As with anthracycline cardiotoxicity, therapy is as usual for heart failure, but it is not clear whether the spontaneous rate of improvement is enhanced by neurohormonal antagonists. The cardiotoxic effects of other recently introduced anti-HER2 therapies (e.g., pertuzumab) are similar to that caused by trastuzumab.

Cardiotoxicity with *cyclophosphamide and ifosfamide* generally occurs acutely and with very high doses. 5-Fluorouracil, cisplatin, and some other alkylating agents can cause recurrent coronary spasm that occasionally leads to depressed contractility. Acute administration of *interferon-α* can cause hypotension and arrhythmias. Clinical heart failure occurring during repeated chronic administration usually resolves after discontinuation.

Many small-molecule *tyrosine kinase inhibitors* that affect VEGF are under use for different malignancies. Although these agents are “targeted” at specific tumor receptors or pathways, the biologic conservation of signaling pathways can cause these inhibitors to have “off-target” effects that include the cardiovascular system and as a group are associated with a ~2.7-fold increased risk of heart failure. Recognition of cardiotoxicity during therapy with these agents is complicated because they occasionally cause peripheral fluid accumulation (ankle edema, periorbital swelling, pleural effusions) due to local factors rather than elevated central venous pressures. Therapeutic approaches include withdrawal of the tyrosine kinase inhibitor (when possible) and substitution with a congener (when available), as well as conventional treatment for heart failure.

Proteasome inhibitors used to treat multiple myeloma are associated with an increased risk of heart failure. The more potent agent, carfilzomib, appears more cardiotoxic than bortezomib.

Immune *checkpoint inhibitors*, such as ipilimumab and nivolumab, are associated with multisystem autoimmune inflammatory toxicities (e.g., thyroiditis, hypophysitis, pancreatitis, and pneumonitis) and rarely myocarditis. However, combination therapy with two checkpoint inhibitors can cause fulminant myocarditis with associated systolic dysfunction, AV block, and ventricular tachycardia within weeks after initial chemotherapy. This presentation has been accompanied by acute skeletal myocarditis and rapid progression to death.

Other therapeutic drugs that can cause cardiotoxicity during chronic use include hydroxychloroquine, chloroquine, emetine, and antiretroviral therapies.

Toxic exposures can cause arrhythmias or respiratory injury acutely during accidents. Chronic exposures implicated in cardiotoxicity include hydrocarbons, fluorocarbons, arsenicals, lead, and mercury.

METABOLIC CAUSES OF CARDIOMYOPATHY

Endocrine disorders affect multiple organ systems, including the heart. *Hyperthyroidism and hypothyroidism* do not often cause clinical heart failure in an otherwise normal heart, but commonly exacerbate heart failure. Clinical signs of thyroid disease may be masked, so tests of thyroid function are part of the routine evaluation of cardiomyopathy. Hyperthyroidism should always be considered with new-onset atrial fibrillation or ventricular tachycardia or atrial fibrillation in which the rapid ventricular response is difficult to control. The most common current reason for thyroid abnormalities in the cardiac population is the treatment of tachyarrhythmias with amiodarone, a drug with substantial iodine content. Hypothyroidism should be treated with very slow escalation of thyroid supplements to avoid exacerbating tachyarrhythmias and heart failure. Hyperthyroidism and heart failure create a dangerous

combination that merits very close supervision, often hospitalization, during titration of antithyroid medications, during which decompensation of heart failure may occur precipitously and fatally.

Pheochromocytoma is rare, but should be considered when a patient has heart failure and very labile blood pressure and heart rate, sometimes with episodic palpitations **(Chap. 380)**. Patients with pheochromocytoma often have postural hypotension. In addition to α-adrenergic receptor antagonists, definitive therapy requires surgical extirpation. Very high renin states, such as those caused by renal artery stenosis, can lead to modest depression in ejection fraction with little or no ventricular dilation and markedly labile symptoms with flash pulmonary edema, related to sudden shifts in vascular tone and intravascular volume.

Controversies remain regarding whether *diabetes* and *obesity* are sufficient to cause cardiomyopathy. Most heart failure in diabetes results from epicardial coronary disease, with further increase in coronary artery risk due to accompanying hypertension and renal dysfunction. Cardiomyopathy may result in part from insulin resistance and increased advanced-glycosylation end products, which impair both systolic and diastolic function. However, much of the dysfunction can be attributed to scattered focal ischemia resulting from distal coronary artery tapering and limited microvascular perfusion even without proximal focal stenoses. Diabetes is a typical factor in heart failure with "preserved" ejection fraction, along with hypertension, advanced age, and female gender.

The existence of a cardiomyopathy due to *obesity* is generally accepted. In addition to cardiac involvement from associated diabetes, hypertension, and vascular inflammation of the metabolic syndrome, obesity alone is associated with impaired excretion of excess volume load, which, over time, can lead to increased wall stress and secondary adaptive neurohumoral responses. Fluid retention may be aggravated by large fluid intake and the rapid clearance of natriuretic peptides by adipose tissue. In the absence of another obvious cause of cardiomyopathy in an obese patient with systolic dysfunction without marked ventricular dilation, effective weight reduction is often associated with major improvement in ejection fraction and clinical function. Improvement in cardiac function has been described after successful bariatric surgery, although all major surgical therapy poses increased risk for patients with heart failure. Postoperative malabsorption and nutritional deficiencies, such as calcium and phosphate deficiencies, may be particularly deleterious for patients with cardiomyopathy.

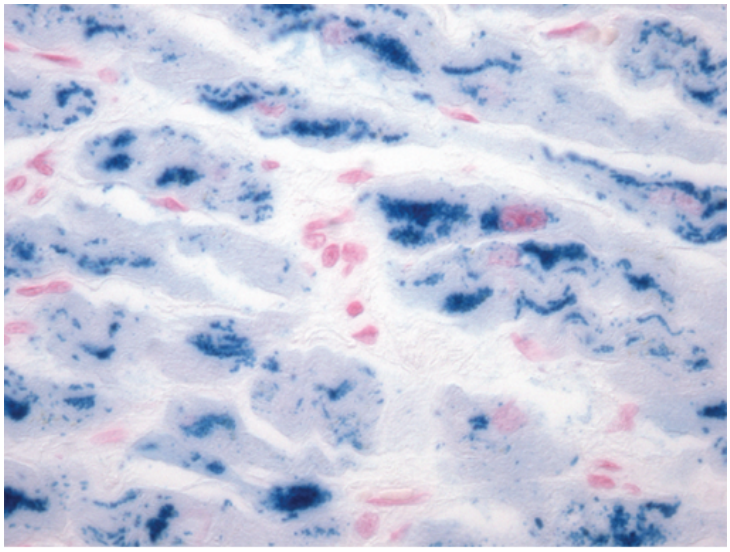
Nutritional deficiencies can occasionally cause DCM but are not commonly implicated in developed Western countries. *Beri-beri heart disease* due to thiamine deficiency can result from poor nutrition in undernourished populations and in patients deriving most of their calories from alcohol, and has been reported in teenagers subsisting only on highly processed foods. This disease is initially a vasodilated state with very high output heart failure that can later progress to a low output state; thiamine repletion can lead to prompt recovery of cardiovascular function. Abnormalities in *carnitine* metabolism can cause dilated or restrictive cardiomyopathies, usually in children. Deficiency of trace elements such as *selenium* can cause cardiomyopathy (Keshan's disease).

Calcium is essential for excitation-contraction coupling. Chronic deficiencies of calcium, such as can occur with hypoparathyroidism (particularly postsurgical) or intestinal dysfunction (from diarrheal syndromes and following extensive resection), can cause severe chronic heart failure that responds over days or weeks to vigorous calcium repletion. *Phosphate* is a component of high-energy compounds needed for efficient energy transfer and multiple signaling pathways. *Hypophosphatemia* can develop during starvation and early refeeding following a prolonged fast, and occasionally during hyperalimentation. *Magnesium* is a cofactor for thiamine-dependent reactions and for the sodium-potassium adenosine triphosphatase (ATPase), but hypomagnesemia rarely becomes sufficiently profound to cause clinical cardiomyopathy.

Hemochromatosis is variably classified as a metabolic or storage disease **(Chap. 407)**. It is included among the causes of restrictive cardiomyopathy, but the clinical presentation is often that of a DCM. The autosomal recessive form is related to the *HFE* gene. With up to 10% of the population heterozygous for one mutation, the clinical prevalence might be as high as 1 in 500. The lower observed rates highlight the limited penetrance of the disease, suggesting the role of additional genetic and environmental factors such as alcoholism affecting clinical expression. Hemochromatosis can also be acquired from iron overload due to hemolytic anemia and transfusions. Excess iron is deposited in the perinuclear compartment of cardiomyocytes, with resulting disruption of intracellular architecture and mitochondrial function. Diagnosis is easily made from measurement of serum iron and transferrin saturation, with a threshold of >60% for men and >45–50% for women. MRI can help to quantitate iron stores in the liver and heart, and endomyocardial biopsy tissue can be stained for iron **(Fig. 254-9)**, which is particularly important if the patient has another cause for cardiomyopathy. If diagnosed early, hemochromatosis can often be managed by repeated phlebotomy to remove iron. For more severe iron overload, iron chelation therapy with desferrioxamine (deferoxamine) or deferasirox can help to improve cardiac function if myocyte loss and replacement fibrosis are not too severe.

FIGURE 254-9

Hemochromatosis. Microscopic image of an endomyocardial biopsy showing extensive iron deposition within the cardiac myocytes with the Prussian blue stain (400× original magnification). *(Image courtesy of Robert Padera, MD, PhD, Department of Pathology, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Inborn disorders of metabolism occasionally present with DCM, although they are most often associated with restrictive cardiomyopathy (Table 254-4).

FAMILIAL DCM

The genetic basis for cardiomyopathy is discussed in the section "Genetic Etiologies of Cardiomyopathy." The recognized frequency of familial involvement in DCM has increased to over 30%. Mutations in *TTN*, encoding the giant sarcomeric protein titin, are the most common cause of DCM, accounting for up to 25% of familial disease. On average, men with *TTN* mutations develop cardiomyopathy a decade before women, without distinctive clinical features. Mutations in thick and thin filament genes account for ~8% of DCM and may manifest in early childhood.

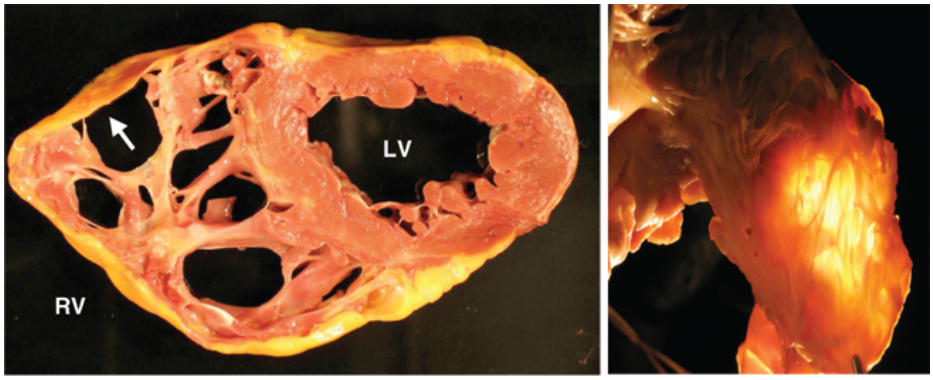
The most recognizable familial cardiomyopathy syndromes with extracardiac manifestations are the *muscular dystrophies.* Both Duchenne's and the milder Becker's dystrophies result from abnormalities in the X-linked dystrophin gene of the sarcolemmal membrane. Skeletal myopathy is present in multiple other genetic cardiomyopathies (Table 254-3), some of which are associated with creatine kinase elevations.

Patients and families with a history of arrhythmias and/or conduction system disease which precede or supersede cardiomyopathy may have abnormalities of the nuclear membrane lamin proteins. While all dilated cardiomyopathies carry a risk of sudden death, a family history of cardiomyopathy with sudden death raises suspicion for a particularly arrhythmogenic mutation; affected family members may be considered for implantable defibrillators even before meeting the reduced ejection fraction threshold for primary prevention of sudden death.

A prominent family history of sudden death or ventricular tachycardia before clinical cardiomyopathy suggests genetic defects in the desmosomal proteins (**Fig. 254-10**). Originally described as affecting the right ventricle (arrhythmogenic right ventricular cardiomyopathy [ARVC]), this disorder (arrhythmogenic cardiomyopathy) can affect either or both ventricles. Patients often present first with ventricular tachycardia. Genetic defects in proteins of the desmosomal complex disrupt myocyte junctions and adhesions, leading to replacement of myocardium by deposits of fat. Thin ventricular walls may be recognized on echocardiography but are better visualized on MRI. Because desmosomes are also important for elasticity of hair and skin, some of the defective desmosomal proteins are associated with striking "woolly hair" and thickened skin on the palms and soles. Implantable defibrillators are usually indicated to prevent sudden death. There is variable progression to right, left, or biventricular failure.

FIGURE 254-10

Arrhythmogenic right ventricular cardiomyopathy. *A.* Cross-sectional slice of a pathology specimen removed at transplantation, showing severe dilation and thinning of the right ventricle (RV) with extensive fatty replacement of right ventricular myocardium. ***B.*** The remarkably thin right ventricular free wall is revealed by transillumination. LV, left ventricle. *(Images courtesy of Gayle Winters, MD, and Richard Mitchell, MD, PhD, Division of Pathology, Brigham and Women's Hospital, Boston.)*



A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Left ventricular noncompaction is a condition of unknown prevalence that is increasingly revealed with the refinement of imaging techniques. The diagnostic criteria include the presence of multiple trabeculations in the left ventricle distal to the papillary muscles, creating a "spongy" appearance of the apex, but are increasingly recognized as non-specific findings in other cardiac diseases. Noncompaction has been associated with multiple genetic variants in the sarcomeric and other genes, such as *TAZ* (encoding tafazzin). The diagnosis may be made incidentally or in patients previously diagnosed with cardiomyopathy, in whom the criteria for noncompaction may appear and resolve with changing left ventricular size and function. The three cardinal clinical features of ventricular arrhythmias, embolic events, and heart failure are largely restricted to patients with concomitant systolic dysfunction. Treatment generally includes anticoagulation and early consideration for an implantable defibrillator, in addition to neurohormonal antagonists as indicated by stage of disease.

Some families inherit a susceptibility to viral-induced myocarditis. This propensity may relate to abnormalities in cell surface receptors, such as the coxsackie-adenovirus receptor, that bind viral proteins. Some may have partial homology with viral proteins such that an autoimmune response is triggered against the myocardium.

Prognosis and therapy of familial DCM are dictated primarily by the stage of clinical disease and the risk for sudden death. In some cases, the familial etiology facilitates prognostic decisions, particularly regarding the likelihood of recovery after a new diagnosis, which is unlikely for familial disease. The rate of progression of disease is to some extent heritable, although marked variation can be seen. However, there have been cases of remarkable clinical remission after acute presentation, likely after a reversible additional insult, such as prolonged tachycardia or infective myocarditis.

TAKOTSUBO CARDIOMYOPATHY

The apical ballooning syndrome, or stress-induced cardiomyopathy, occurs typically in older women after sudden intense emotional or physical stress. The ventricle shows global ventricular dilation with basal contraction, forming the shape of the narrow-necked jar (*takotsubo*) used in Japan to trap octopuses. Originally described in Japan, it is increasingly recognized elsewhere during emergency cardiac catheterization and intensive care unit admissions for noncardiac conditions. Presentations include pulmonary edema, hypotension, and chest pain with ECG changes mimicking an acute infarction. The left ventricular dysfunction extends beyond a specific coronary artery distribution and generally resolves within days to weeks. Animal models and ventricular biopsies suggest that this acute cardiomyopathy may result from intense sympathetic activation with heterogeneity of myocardial autonomic innervation, diffuse microvascular spasm, and/or direct catecholamine toxicity. Coronary angiography may be required to rule out acute coronary occlusion. No therapies have been proven beneficial, but reasonable strategies include nitrates for pulmonary edema, intraaortic balloon pump if needed for low output, combined alpha and beta blockers rather than selective beta blockade if hemodynamically stable, and magnesium for arrhythmias related to QT prolongation. Anticoagulation is generally withheld due to the occasional occurrence of ventricular rupture. While the prognosis is generally good, recurrences have been described in up to 10% of patients.

IDIOPATHIC DCM

Idiopathic DCM is a diagnosis of exclusion, when all other known factors have been excluded. Approximately two-thirds of dilated cardiomyopathies are still labeled as idiopathic; however, a substantial proportion of these may reflect unrecognized genetic disease. Continued reconsideration of etiology during chronic heart failure management often reveals specific causes later in a patient's course.

OVERLAPPING TYPES OF CARDIOMYOPATHY

The limitations of our phenotypic classification are revealed through the multiple overlaps between the etiologies and presentations of the three types. Cardiomyopathy with reduced systolic function but without severe dilation can represent early DCM, "minimally dilated cardiomyopathy," or restrictive diseases without marked increases in ventricular wall thickness. For example, sarcoidosis and hemochromatosis can present as dilated or restrictive disease. Early stages of amyloidosis are often mistaken for hypertrophic cardiomyopathy. Progression of hypertrophic cardiomyopathy into a "burned-out"

phase occurs occasionally, with decreased contractility and modest ventricular dilation. Overlaps are particularly common with the inherited metabolic disorders, which can present as any of the three major phenotypes (Fig. 254-4).

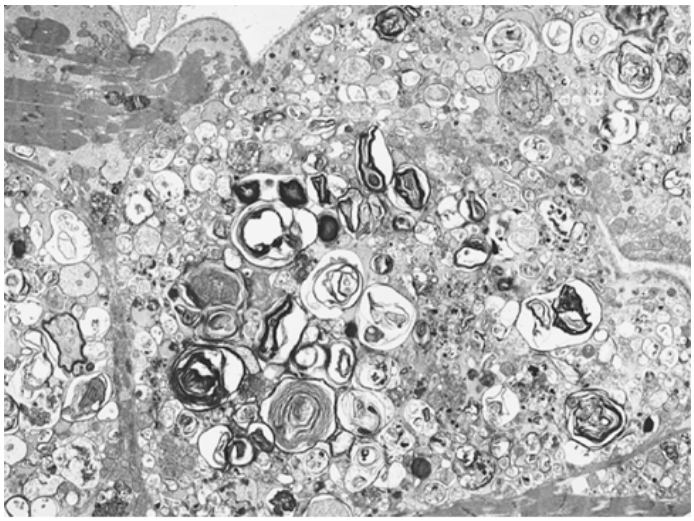
DISORDERS OF METABOLIC PATHWAYS

Multiple genetic disorders of metabolic pathways can cause myocardial disease, due to infiltration of abnormal products or cells containing them between the myocytes, and storage disease, due to their accumulation within cells **(see Tables 254-3, and 254-4)**. Hypertrophic cardiomyopathy may be mimicked by the myocardium thickened with these abnormal products causing "pseudohypertrophy," usually with an abnormally short PR interval. The pseudo-hypertrophic phenotype is most common, but restrictive and DCM may occur. Most of these diseases are diagnosed during childhood.

Fabry's disease results from a deficiency of the lysosomal enzyme alpha-galactosidase A caused by one of more than 160 mutations in *GLA.* This disorder of glycosphingolipid metabolism is an X-linked disorder that may also cause clinical disease in female carriers. Glycolipid accumulation may be limited to the cardiac tissues but usually also involves the skin, peripheral nerve, and kidney. Electron microscopy of endomyocardial biopsy tissue shows diagnostic vesicles containing concentric lamellar figures **(Fig. 254-11)**. Diagnosis can be made through assessment of enzyme activity and/or *GLA* sequencing and is crucial because enzyme replacement can reduce abnormal deposits and improve cardiac and clinical function. The magnitude of clinical impact has not been well-established for this therapy, which requires frequent infusions of the enzyme at a cost of >$100,000 a year. Enzyme replacement can also improve the course of Gaucher's disease, in which cerebroside-rich cells accumulate in multiple organs due to a deficiency of beta-glucosidase. Cerebroside-rich cells infiltrate the heart, which can also lead to a hemorrhagic pericardial effusion and valvular disease.

FIGURE 254-11

Fabry's disease. Transmission electron micrograph of a right ventricular endomyocardial biopsy specimen at high magnification showing the characteristic concentric lamellar inclusions of glycosphingolipids accumulating as a result of deficiency of the lysosomal enzyme alpha-galactosidase A. Image taken at 15,000× original magnification. *(Image courtesy of Robert Padera, MD, PhD, Department of Pathology, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Glycogen storage diseases lead to accumulation of lysosomal storage products and intracellular glycogen accumulation, particularly with *glycogen storage disease type III,* due to a defective debranching enzyme. There are >10 types of *mucopolysaccharidoses,* in which autosomal recessive or X-linked deficiencies of lysosomal enzymes lead to the accumulation of glycosaminoglycans in the skeleton, nervous system, and occasionally the heart. With characteristic facies, short stature, and frequent cognitive impairment, most individuals are diagnosed early in childhood and die before adulthood.

Carnitine is an essential cofactor in long-chain fatty acid metabolism. Multiple defects have been described that lead to carnitine deficiency, causing intracellular lipid inclusions and restrictive or DCM, often presenting in children. Fatty acid oxidation requires many metabolic steps with specific enzymes that can be deficient, with complex interactions with carnitine. Depending on the defect, cardiac and skeletal myopathy can be ameliorated with replacement of fatty acid intermediates and carnitine.

Two monogenic metabolic cardiomyopathies cause markedly increased ventricular wall thickness without an increase of muscle subunits or an increase in contractility. Mutations in the gamma-2 regulatory subunit of the adenosine monophosphate (AMP)-activated protein kinase important for glucose metabolism (*PRKAG2*) have been associated with a high prevalence of conduction abnormalities, such as AV block and ventricular preexcitation. Several defects have been reported in an X-linked lysosome-associated membrane protein (*LAMP2*). This defect can be maternally transmitted or sporadic and has occasionally been isolated to the heart, although it often leads to a syndrome of skeletal myopathy, mental retardation, and hepatic dysfunction referred to as *Danon's disease.* Extreme left ventricular hypertrophy appears early, often in childhood, and can progress rapidly to end-stage heart failure with low ejection fraction. Electron microscopy of these metabolic disorders shows that the myocytes are enlarged by multiple intracellular vacuoles of metabolic by-products.

RESTRICTIVE CARDIOMYOPATHY

Restrictive cardiomyopathy is dominated by abnormal diastolic function, often with mildly decreased contractility and ejection fraction (usually >30–50%). Both atria are enlarged, sometimes massively. Modest left ventricular dilation can be present, usually with an end-diastolic dimension <6 cm. End-diastolic pressures are elevated in both ventricles, with preservation of cardiac output until late in the disease. Subtle exercise intolerance is usually the first symptom but is often not recognized until after clinical presentation with congestive symptoms. The restrictive diseases often present with relatively more right-sided symptoms, such as edema, abdominal discomfort, and ascites, although filling pressures are elevated in both ventricles. The cardiac impulse is less displaced than in DCM and less dynamic than in hypertrophic cardiomyopathy. A fourth heart sound is more common than a third heart sound in sinus rhythm, but atrial fibrillation is common. Jugular venous pressures often show rapid Y descents and may increase during inspiration (positive Kussmaul's sign). Most restrictive cardiomyopathies are due to infiltration of abnormal substances between myocytes, storage of abnormal metabolic products within myocytes, or fibrotic injury (**Table 254-5**). The differential diagnosis should include constrictive pericardial disease, which may also be dominated by right-sided heart failure.

TABLE 254-5

Causes of Restrictive Cardiomyopathies

Infiltrative (Between Myocytes)
Amyloidosis
Primary (light chain amyloid)
Familial (abnormal transthyretin)[a]
Senile (normal transthyretin or atrial peptides)
Inherited metabolic defects[a]
Storage (Within Myocytes)
Hemochromatosis (iron)[a]
Inherited metabolic defects[a]
Fabry's disease
Glycogen storage disease (II, III)
Fibrotic
Radiation
Scleroderma
Endomyocardial
Possibly related fibrotic diseases
Tropical endomyocardial fibrosis
Hypereosinophilic syndrome (Löffler's endocarditis)
Carcinoid syndrome
Radiation
Drugs: e.g., serotonin, ergotamine
Overlap with Other Cardiomyopathies
Hypertrophic cardiomyopathy/"pseudohypertrophic"[a]
"Minimally dilated" cardiomyopathy
Early-stage dilated cardiomyopathy
Partial recovery from dilated cardiomyopathy
Sarcoidosis
Idiopathic[a]

[a]Can be familial.

INFILTRATIVE DISEASE

Amyloidosis is the major cause of restrictive cardiomyopathy **(Figs. 254-12, 254-13, and 254-14)**. Several proteins can self-assemble to form the beta-sheets of amyloid proteins, which deposit with different consequences depending on the type of protein. The systemic amyloidoses are discussed in **Chap. 108**. In addition to cardiac infiltration, neurologic involvement occurs commonly with primary amyloidosis (immunoglobulin light chains) and with familial amyloidosis (genetic abnormalities of transthyretin). There are >100 identified mutations in transthyretin on chromosome 13, among which the V122I transthyretin mutation has been identified in ~4% of African Americans in whom it is associated with a 50% increased risk of heart failure. However, penetrance of the V122I mutation is incomplete with most mutation carriers free of heart failure at 70 years of age.

FIGURE 254-12

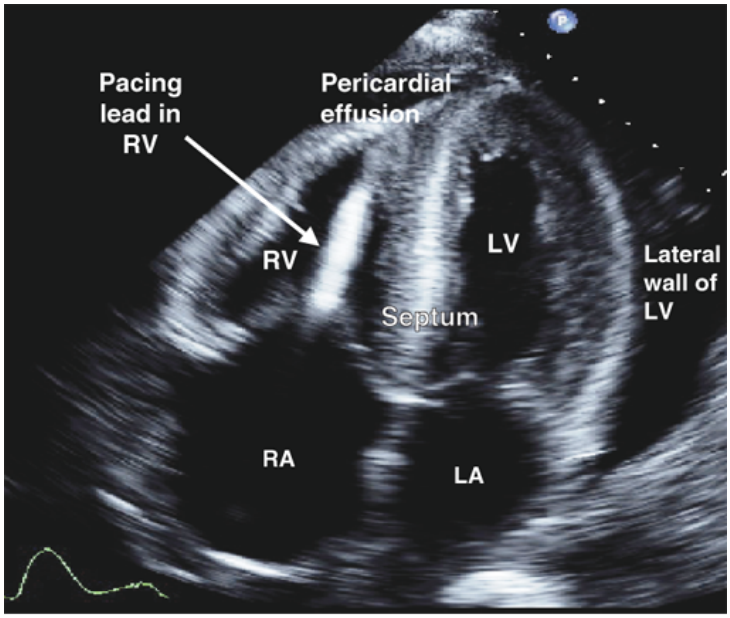
Restrictive cardiomyopathy—amyloidosis. Gross specimen of a heart with amyloidosis. The heart is firm and rubbery with a waxy cut surface. The atria are markedly dilated, and the left atrial endocardium, normally smooth, has yellow-brown amyloid deposits that give texture to the surface. *(Image courtesy of Robert Padera, MD, PhD, Department of Pathology, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 254-13

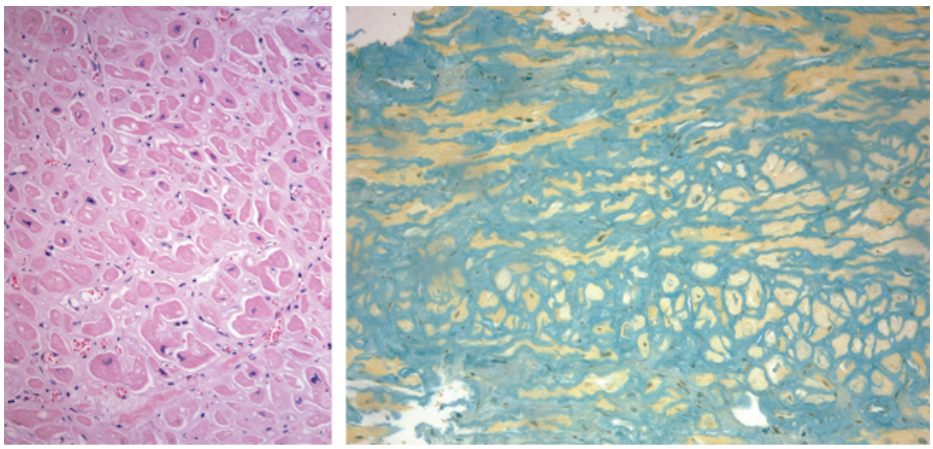
Restrictive cardiomyopathy—amyloidosis. Echocardiogram showing thickened walls of both ventricles without major chamber dilation. The atria are markedly dilated, consistent with chronically elevated ventricular filling pressures. In this example, there is a characteristic hyperrefractile "glittering" of the myocardium typical of amyloid infiltration, which is a non-specific finding with contemporary echocardiography. The mitral and tricuspid valves are thickened. A pacing lead is visible in the right ventricle (RV), and a pericardial effusion is evident. Note that the echocardiographic and pathologic images are vertically opposite, such that the left ventricle (LV) is by convention on the top right in the echocardiographic image and bottom right in the pathologic images. LA, left atrium; RA, right atrium. *(Image courtesy of Justina Wu, MD, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 254-14

Amyloidosis—microscopic images of amyloid involving the myocardium. The left panel (hematoxylin and eosin stain) shows glassy, grey-pink amorphous material infiltrating between cardiomyocytes, which stain a darker pink. The right panel shows a sulfated blue stain that highlights the amyloid green and stains the cardiac myocytes yellow. (The Congo red stain can also be used to highlight amyloid; under polarized light, amyloid will have an apple-green birefringence when stained with Congo red.) Images at 100× original magnification. *(Image courtesy of Robert Padera, MD, PhD, Department of Pathology, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Organ dysfunction in amyloidosis was once attributed solely to physical disruption from the infiltrating amyloid fibrils, but newer information suggests additional direct toxicity from the immunoglobulin light chain and abnormal transthyretin protein aggregates themselves. In senile amyloidosis, there is abnormal accumulation of normal transthyretin or natriuretic peptide folding, detected in 10% of people aged >80 years and half of those aged >90 years but often without apparent clinical disease. Men show a greater burden of amyloid deposition and twentyfold greater likelihood of clinical disease with senile amyloidosis. The aging of the population will soon render senile amyloidosis the most common of the amyloidoses.

Cardiac amyloid is classically suspected from thickened ventricular walls with an ECG that shows low voltage. However, low voltage is not always present and is less common in familial or senile amyloidosis than in primary AL amyloidosis. A characteristic refractile brightness in the septum on echocardiography is suggestive of the diagnosis, but neither sensitive nor specific. Both atria are dilated, often dramatically, and diastolic dysfunction may be more obvious than in left ventricular hypertrophy from other causes. Amyloid infiltration can also be detected with gadolinium enhancement in MRI. Technetium pyrophosphate imaging is sensitive and specific for *TTR* amyloidosis as opposed to AL amyloidosis. The diagnosis of primary or familial amyloidosis can sometimes be made from biopsies of an abdominal fat pad or the rectum, but cardiac amyloidosis is most reliably identified from a biopsy of the heart, in which amyloid fibrils infiltrate the myocardium diffusely, particularly around the conduction system and coronary vessels (Fig. 254-14). Diagnosis of the type of amyloid protein requires immunohistochemistry of biopsied tissue rather than serum or urine electrophoresis, which can lead to incorrect classification.

Therapy for all types of amyloid is predominantly for symptoms of fluid retention, which often requires high doses of loop diuretics. Digoxin bound to the amyloid fibrils can reach toxic levels, and should therefore be used only in very low doses, if at all. There is no evidence regarding use of neurohormonal antagonists in amyloid heart disease, where the possible theoretical benefit has to be balanced against the possibility of aggravating postural hypotension and diminishing the crucial heart rate reserve. The risk of intracardiac thrombi may warrant chronic anticoagulation.

The prognosis is worst for primary amyloid, with a median survival of 6–12 months after presentation, but that has improved substantially with the use of the proteasome inhibitor bortezomib. If present, multiple myeloma is treated with chemotherapy, the extent of which is often limited by the potential of worsening cardiac dysfunction. Immunoglobulin-associated amyloid has occasionally been treated with sequential heart transplantation and delayed bone marrow transplant, with frequent recurrence of amyloid in the transplanted heart. Abnormal transthyretin-associated cardiac amyloid has a somewhat better prognosis and can be treated in selected patients with heart and liver transplantation. Senile cardiac amyloid has the slowest progression and best overall prognosis. Novel therapies including RNA interference and small molecules are being studied in *TTR* amyloidosis.

FIBROTIC RESTRICTIVE CARDIOMYOPATHY

Progressive fibrosis can cause restrictive myocardial disease without ventricular dilation. Thoracic radiation, common for breast and lung cancer or mediastinal lymphoma, can produce early or late restrictive cardiomyopathy. Patients with *radiation cardiomyopathy* may present with a possible diagnosis of constrictive pericarditis, as the two conditions often coexist. Careful hemodynamic evaluation and, often, endomyocardial biopsy should be performed if considering pericardial stripping surgery, which is unlikely to be successful in the presence of underlying restrictive cardiomyopathy. *Scleroderma* causes small vessel spasm and ischemia that can lead to a small, stiff heart with reduced ejection fraction without dilation. The pulmonary hypertension associated with scleroderma may lead to more clinical right heart failure because of concomitant fibrotic disease of the right ventricle.

ENDOMYOCARDIAL DISEASE

The physiologic picture of elevated filling pressures with atrial enlargement and preserved ventricular contractility with normal or reduced ventricular volumes can result from extensive fibrosis of the endocardium, without transmural myocardial disease. For patients who have not lived in the equatorial regions, this picture is rare, and when seen is often associated with a history of chronic hypereosinophilic syndrome (*Löffler's endocarditis*), which is more common in men than women. In this disease, persistent hypereosinophilia of >1500 eos/μL for at least 6 months can cause an acute phase of eosinophilic injury in the endocardium (see earlier discussion of eosinophilic myocarditis), with systemic illness and injury to other organs. There is usually no obvious cause, but the hypereosinophilia can occasionally be explained by allergic, parasitic, or malignant disease. It is postulated to be followed by a period in which cardiac inflammation is replaced by evidence of fibrosis with superimposed thrombosis. In severe disease, the dense fibrotic layer can obliterate the ventricular apices and extend to thicken and tether the AV valve leaflets. The clinical disease may present with heart failure, embolic events, and atrial arrhythmias. While plausible, the sequence of transition from eosinophilic myocarditis or Löffler's endocarditis to endomyocardial fibrosis has not been clearly demonstrated.

In tropical countries, up to one-quarter of heart failure may be due to *endomyocardial fibrosis,* affecting either or both ventricles. This condition shares with the previous condition the partial obliteration of the ventricular apex with fibrosis extending into the valvular inflow tract and leaflets; however, it is not clear that the etiologies are the same for all cases. Pericardial effusions frequently accompany endomyocardial fibrosis but are not common in Löffler's endocarditis. For endomyocardial fibrosis, there is no gender difference, but a higher prevalence in African-American populations. While tropical endomyocardial fibrosis could represent the end-stage of previous hypereosinophilic disease triggered by endemic parasites, neither prior parasitic infection nor hypereosinophilia is usually documented. Geographic nutritional deficiencies have also been proposed as an etiology.

Medical treatment focuses on glucocorticoids and chemotherapy to suppress hypereosinophilia when present. Fluid retention may become increasingly resistant to diuretic therapy. Anticoagulation is recommended. Atrial fibrillation is associated with worse symptoms and prognosis, but may be difficult to suppress. Surgical resection of the apices and replacement of the fibrotic valves can improve symptoms, but surgical morbidity and mortality and later recurrence rates are high.

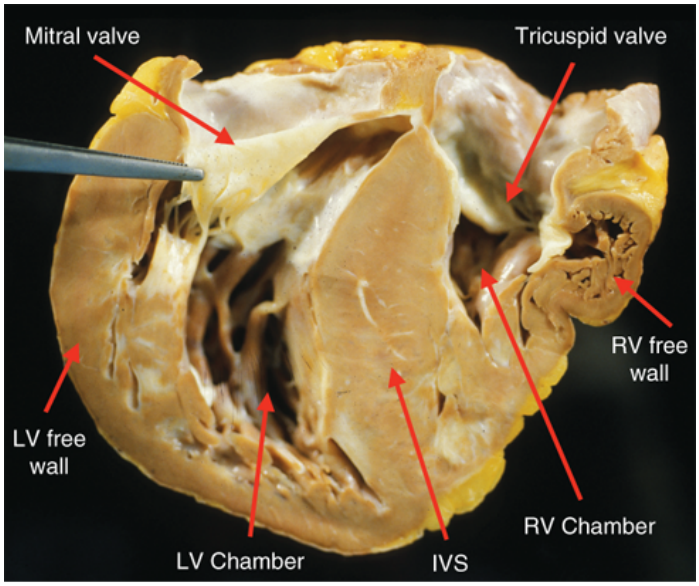
The serotonin secreted by *carcinoid* tumors can produce fibrous plaques in the endocardium and right-sided cardiac valves, occasionally affecting left-sided valves, as well. Valvular lesions may be stenotic or regurgitant. Systemic symptoms include flushing and diarrhea. Liver disease from hepatic metastases may play a role by limiting hepatic function and thereby allowing more serotonin to reach the venous circulation.

HYPERTROPHIC CARDIOMYOPATHY

Hypertrophic cardiomyopathy is defined as left ventricular hypertrophy that develops in the absence of causative hemodynamic factors, such as hypertension, aortic valve disease, or systemic infiltrative or storage diseases **(Figs. 254-15 and 254-16)**. It has previously been termed *hypertrophic obstructive cardiomyopathy* (HOCM), *asymmetric septal hypertrophy* (ASH), and *idiopathic hypertrophic subaortic stenosis* (IHSS). However, the accepted terminology is now hypertrophic cardiomyopathy with or without obstruction. Prevalence in North America, Africa, and Asia is about 1:500. It is the leading cause of sudden death in the young and is an important cause of heart failure. Although pediatric presentation is associated with increased early morbidity and mortality, the prognosis for patients diagnosed as adults is generally favorable.

FIGURE 254-15

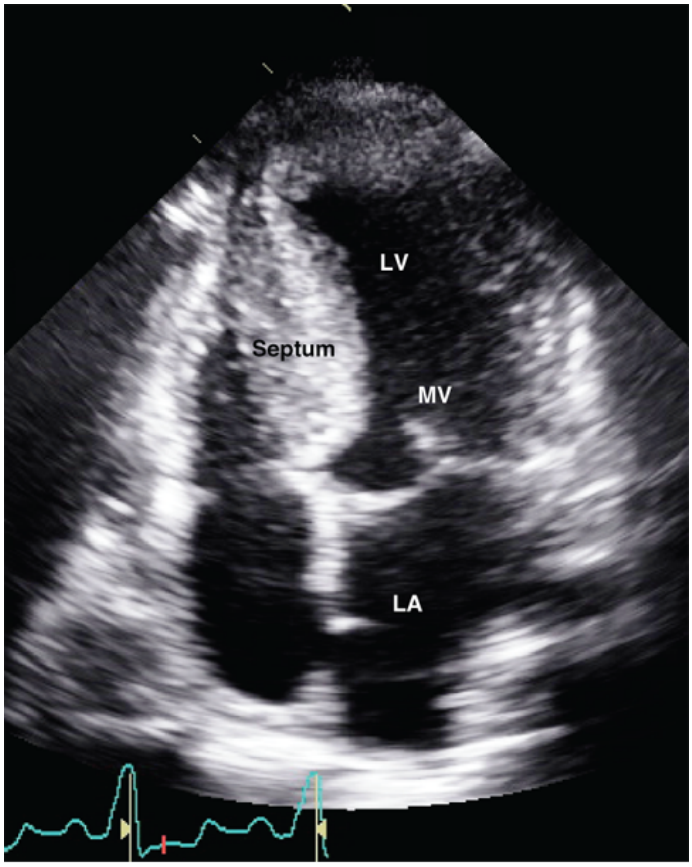
Hypertrophic cardiomyopathy. Gross specimen of a heart with hypertrophic cardiomyopathy removed at the time of transplantation, showing asymmetric septal hypertrophy (septum much thicker than left ventricular free wall) with the septum bulging into the left ventricular outflow tract causing obstruction. The forceps are retracting the anterior leaflet of the mitral valve, demonstrating the characteristic plaque of systolic anterior motion, manifest as endocardial fibrosis on the interventricular septum in a mirror-image pattern to the valve leaflet. There is patchy replacement fibrosis, and small thick-walled arterioles can be appreciated grossly, especially in the interventricular septum. IVS, interventricular septum; LV, left ventricle; RV, right ventricle. *(Image courtesy of Robert Padera, MD, PhD, Department of Pathology, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 254-16

Hypertrophic cardiomyopathy. This echocardiogram of hypertrophic cardiomyopathy shows asymmetric hypertrophy of the septum compared to the lateral wall of the left ventricle (LV). The mitral valve (MV) is moving anteriorly toward the hypertrophied septum in systole. The left atrium (LA) is enlarged. Note that the echocardiographic and pathologic images are vertically opposite, such that the LV is by convention on the top right in the echocardiographic image and bottom right in the pathologic images. *(Image courtesy of Justina Wu, MD, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

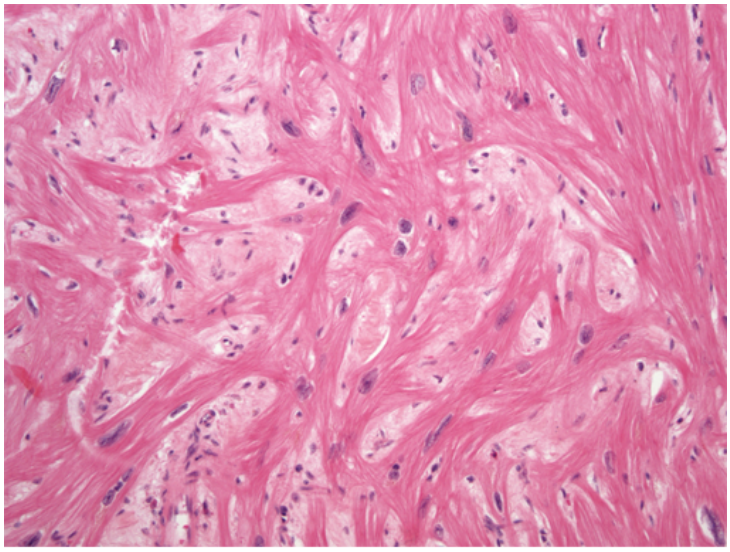
The clustering of hypertrophic cardiomyopathy within families has been appreciated since recognition of the disease ~55 years ago. Echocardiographic screening of families revealed an autosomal dominant pattern of inheritance. Initial genetic studies using linkage analysis in large families identified disease-causing mutations in sarcomeric genes. A sarcomere mutation is present in ~60% of patients with hypertrophic cardiomyopathy and is more common in those with familial disease and characteristic asymmetric septal hypertrophy. More than nine different sarcomere genes with >1400 mutations have been implicated, although ~80% of patients have a mutation in either *MYH7* or *MYBPC3* (Table 254-3).

Hypertrophic cardiomyopathy is characterized by age-dependent and incomplete penetrance. The defining phenotype of left ventricular hypertrophy is rarely present at birth and usually develops later in life. Accordingly, screening of family members should begin in adolescence and extend through adulthood. In *MYBPC3* mutation carriers, the average age of disease development is 40 years, while 30% remain free from hypertrophy after 70 years. Related individuals who carry the *same* mutation may have a different extent and pattern of hypertrophy (e.g., asymmetric versus concentric), occurrence of outflow tract obstruction, and associated clinical outcomes (e.g., sudden death, atrial fibrillation).

At the level of the sarcomere, hypertrophic cardiomyopathy mutations lead to enhanced calcium sensitivity, maximal force generation, and ATPase activity. Calcium handling is affected through modification of regulatory proteins. Sarcomere mutations lead to abnormal energetics and impaired relaxation, both directly and as a result of hypertrophy. Hypertrophic cardiomyopathy is characterized by misalignment and disarray of the enlarged myofibrils and myocytes (**Fig. 254-17**), which can also occur to a lesser extent in other cardiac diseases. Although hypertrophy is the defining feature of hypertrophic cardiomyopathy, fibrosis and microvascular disease are also present. Interstitial fibrosis is detectable before overt hypertrophy develops and likely results from early activation of profibrotic pathways. In the majority of patients with overt cardiomyopathy, focal areas of replacement fibrosis can be readily detected with MRI. These areas of "scar" may represent substrate for the development of ventricular arrhythmias. Increased thickness and decreased luminal area of the intramural vessels in hypertrophied myocardium contribute to microvascular ischemia and angina. Microinfarction of hypertrophied myocardium is a hypothesized mechanism for replacement scar formation.

FIGURE 254-17

Hypertrophic cardiomyopathy. Microscopic image of hypertrophic cardiomyopathy showing the characteristic disarrayed myocyte architecture with swirling and branching rather than the usual parallel arrangement of myocyte fibers. Myocyte nuclei vary markedly in size and interstitial fibrosis is present. *(Image courtesy of Robert Padera, MD, PhD, Department of Pathology, Brigham and Women's Hospital, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Macroscopically, hypertrophy is typically manifest as nonuniform ventricular thickening (Fig. 254-15). The interventricular septum is the typical location of maximal hypertrophy, although other patterns of hypertrophic remodeling include concentric and midventricular. Hypertrophy confined to the ventricular apex (apical hypertrophic cardiomyopathy) is less often familial and has a different genetic substrate, with sarcomere mutations present in only ~15%. Left ventricular outflow tract obstruction represents the most common focus of diagnosis and intervention, although diastolic dysfunction, myocardial fibrosis, and microvascular ischemia also contribute to contractile dysfunction and elevated intracardiac pressures. Obstruction is present in ~30% of patients at rest and can be provoked by exercise in another ~30%. Systolic obstruction is initiated by drag forces, which push an anteriorly displaced and enlarged anterior mitral leaflet into contact with the hypertrophied ventricular septum. Mitral leaflet coaptation may ensue, leading to posteriorly directed mitral regurgitation. In order to maintain stroke volume across outflow tract obstruction, the ventricle generates higher pressures, leading to higher wall stress and myocardial oxygen demand. Smaller chamber size and increased contractility exacerbate the severity of obstruction. Conditions of low preload, such as dehydration, and low afterload, such as arterial vasodilation, may lead to transient hypotension and near-syncope. The systolic ejection murmur of left ventricular outflow tract obstruction is harsh and late peaking and can be enhanced by bedside maneuvers that diminish ventricular volume and transiently worsen obstruction, such as standing from a squatting position or the Valsalva maneuver.

DIAGNOSIS

The substantial variability of hypertrophic cardiomyopathy pathology is reflected in the diversity of clinical presentations. Patients may be diagnosed after undergoing evaluations triggered by the abnormal physical findings (murmur) or symptoms of exertional dyspnea, angina, or syncope. Alternatively, diagnosis may follow evaluations prompted by the detection of disease in family members. Cardiac imaging (Fig. 254-16) is central to diagnosis due to the insensitivity of examination and ECG and the need to exclude other causes for hypertrophy. The identification of a disease-causing mutation in a proband can focus family evaluations on mutation carriers, but this strategy requires a high degree of certainty that the mutation is truly pathogenic and not a benign DNA variant. Biopsy is not needed to diagnose hypertrophic cardiomyopathy but can be used to exclude infiltrative and metabolic diseases. Rigorous athletic training (athlete's heart) may cause intermediate degrees of physiologic hypertrophy difficult to differentiate from mild hypertrophic cardiomyopathy. Unlike hypertrophic cardiomyopathy, hypertrophy in the athlete's heart regresses with cessation of training, and is accompanied by supernormal exercise capacity (VO_{2max} >50 mL/kg per min), mild ventricular dilation, and normal diastolic function.

TREATMENT

TREATMENT

Hypertrophic Cardiomyopathy

Management focuses on treatment of symptoms and prevention of sudden death and stroke **(Fig. 254-18)**. Left ventricular outflow tract obstruction can be controlled medically in the majority of patients. β-Adrenergic blocking agents and L-type calcium channel blockers (e.g., verapamil) are first-line agents that reduce the severity of obstruction by slowing heart rate, enhancing diastolic filling, and decreasing contractility. Persistent symptoms of exertional dyspnea or chest pain can sometimes be controlled with the addition of disopyramide, an antiarrhythmic agent with potent negative inotropic properties.

Patients with or without obstruction may develop heart failure symptoms due to fluid retention and require diuretic therapies for venous congestion. Severe medically refractory symptoms develop in ~5% of patients, for whom surgical myectomy or alcohol septal ablation may be effective. Developed over 50 years ago, surgical myectomy effectively relieves outflow tract obstruction by excising part of the septal myocardium involved in the dynamic obstruction. In selected patients, perioperative mortality is extremely low with excellent long-term survival free from recurrent obstruction and symptoms.

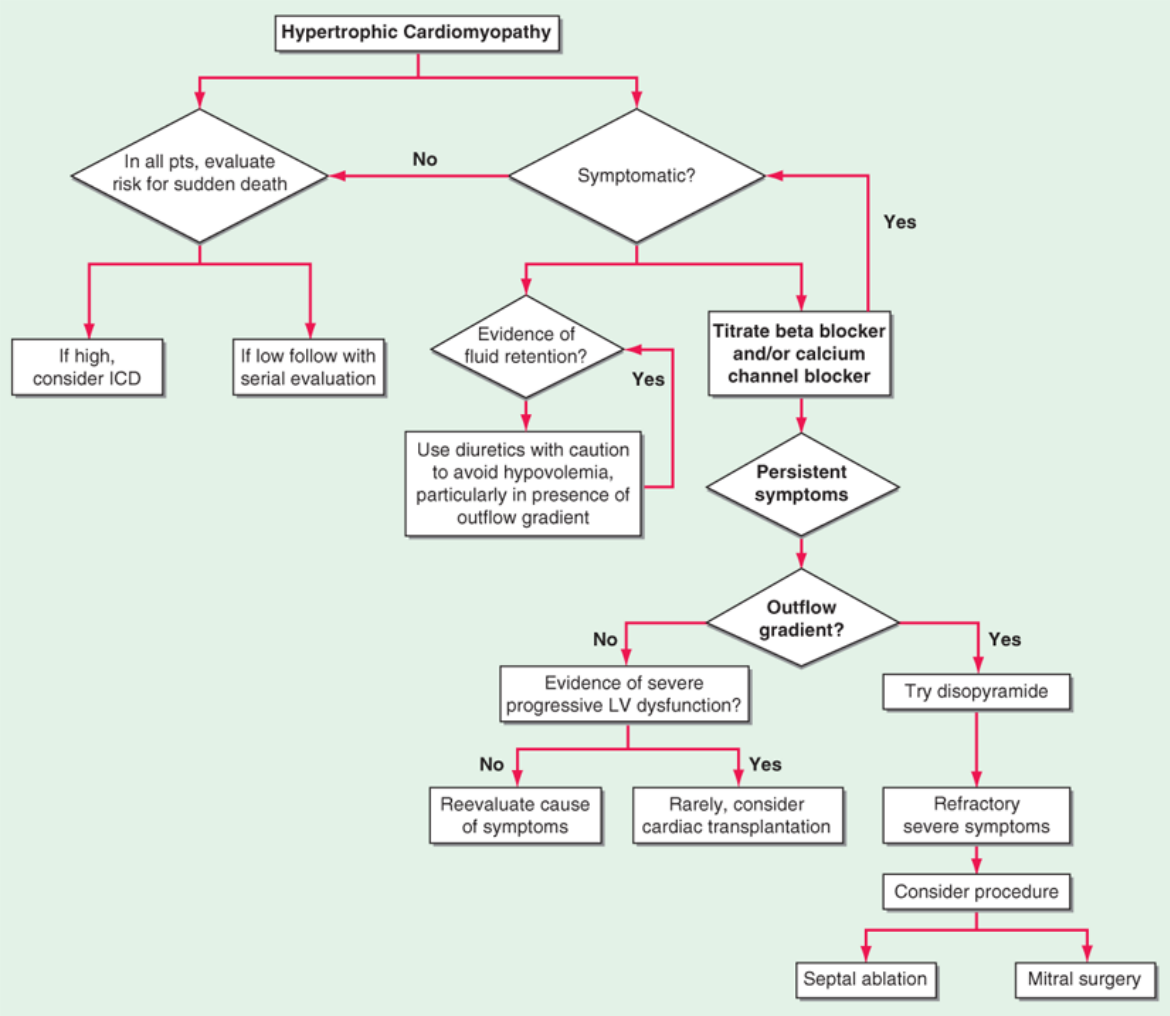
Mitral valve repair or replacement is usually unnecessary as associated eccentric mitral regurgitation resolves with myectomy alone. Alcohol septal ablation in patients with suitable coronary anatomy can relieve outflow tract obstruction via a controlled infarction of the proximal septum, which produces similar periprocedural outcomes and gradient reduction as surgical myomectomy. Until long-term outcomes are demonstrated for this procedure, it is relegated primarily to patients who wish to avoid surgery or who have limiting comorbidities. Neither procedure has been shown to improve outcomes other than symptoms. With both procedures, the most common complication is the development of complete heart block necessitating permanent pacing. However, ventricular pacing as a primary therapy for outflow tract obstruction is ineffective and not generally advised.

Patients with hypertrophic cardiomyopathy have an increased risk of sudden cardiac death from ventricular tachyarrhythmias. Vigorous physical activity and competitive sport are prohibited. Factors that increase the risk of sudden death from a baseline of 0.5% per year are presented in **Table 254-6**. As sudden death has not been reduced by medical or procedural interventions, an implantable cardioverter-defibrillator is advised for patients with two or more risk factors and is advised on a selected basis for patient with one risk factor. Nevertheless, the positive predictive value of most risk factors is low, and many patients receiving a defibrillator never receive an appropriate therapy. Long-term use of a defibrillator may be associated with serious device-related complications, particularly in young active patients. Refinement of sudden death risk through the application of contemporary technologies such as cardiac MRI is ongoing.

Atrial fibrillation is common in patients with hypertrophic cardiomyopathy and may lead to hemodynamic deterioration and embolic stroke. Rapid ventricular response is poorly tolerated and may worsen outflow tract obstruction. β-Adrenergic blocking agents and L-type calcium channel blockers slow AV nodal conduction and improve symptoms; cardiac glycosides should be avoided, as they may increase contractility and worsen obstruction. Symptoms exacerbated by atrial fibrillation may persist despite adequate rate control due to loss of AV synchrony and may require restoration of sinus rhythm. Disopyramide and amiodarone are the preferred antiarrhythmic agents, with radiofrequency ablation considered for medically refractory cases. Anticoagulation to prevent embolic stroke in atrial fibrillation is recommended.

FIGURE 254-18

Treatment algorithm for hypertrophic cardiomyopathy depending on the presence and severity of symptoms and the presence of an intraventricular gradient with obstruction to outflow. Note that all patients with hypertrophic cardiomyopathy should be evaluated for atrial fibrillation and risk of sudden death, whether or not they require treatment for symptoms. ICD, implantable cardioverter-defibrillator; LV, left ventricular.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 254-6

Risk Factors for Sudden Death in Hypertrophic Cardiomyopathy

Major Risk Factor		Screening Technique
History of cardiac arrest or spontaneous sustained ventricular tachycardia[a]		History
Syncope	Nonvagal, often with or after exertion	History
Family history of sudden cardiac death		Family history
Spontaneous nonsustained ventricular tachycardia[b]	>3 beats at rate >120	Exercise or 24- to 48-h ambulatory recording
LV thickness >30 mm	Present in <10% of patients	Echocardiography
Abnormal blood pressure response to exercise[b]	Systolic blood pressure fall or failure to increase at peak exercise	Maximal upright exercise testing

[a]Implantable cardioverter-defibrillator advised for patients with prior arrest or sustained ventricular tachycardia regardless of other risk factors. [b] Prognostic value most applicable to patients <40 years old.

Abbreviation: LV, left ventricle.

PROGNOSIS

The general prognosis for hypertrophic cardiomyopathy is better than in early studies of referral populations. For patients diagnosed as adults, survival is comparable to an age-matched population without cardiomyopathy. The sudden death risk is <1% per year; however, up to 1 in 20 patients will progress to overt systolic dysfunction with a reduced ejection fraction with or without dilated remodeling (“burned out” or end-stage hypertrophic cardiomyopathy). These patients suffer from low cardiac output and have a high risk of death from progressive heart failure and sudden death unless they undergo cardiac transplantation.

GLOBAL PERSPECTIVES

Comparison of myocardial diseases across eras and countries is complicated by differences in techniques for diagnosis, such as endomyocardial biopsy, testing for viral genomes, and specific antibodies. Deaths attributed to cardiomyopathy/myocarditis in the Global Burden of Disease study have increased by 51% between 1990 and 2013 while the age-adjusted mortality rates have declined by 12.6% and the disability-adjusted life years lost have declined by almost 4%. For comparison, the current mortality rates are comparable to those of rheumatic heart disease, which has declined overall by 26.5 and by 55% after adjustment for age. Deaths from Chagas’ cardiomyopathy worldwide have declined from 12.7 thousand to 10.6 thousand, with a reduction of 51.7% in the age-adjusted rates per 100,000 population to 0.2, attributable in major part to improved health conditions in rural areas of South and Central America. By contrast, there has been an increase in the prevalence of Chagas’ disease to an estimated 300,000 in the United States, detected largely through blood donation. It is no longer limited to patients from known endemic areas as de novo infection is increasingly recognized in warmer regions of the country.

Health care for other diseases affects myocarditis and cardiomyopathy. Developed nations will see a higher prevalence of cardiomyopathy due to chemotherapy. However, vaccination has reduced deaths from diphtheria myocarditis to <50 per 100 million population, currently most common in Russia. World regions providing highly active antiviral therapy for HIV have decreased not only transmission but also the rate of associated cardiomyopathy by several-fold. Increasing availability of clinical genetic testing is expected to shift the apparent epidemiology of cardiomyopathy away from acquired causes toward causative and facilitating genetic factors. For instance, heart failure with preserved ejection fraction attributed to hypertension and diabetes is increasingly recognized to represent amyloidosis from mutant transthyretin, with distinct recognized mutations in Portugal, Japan, and the African-Caribbean population.

FURTHER READING

Arbustini E et al: The MOGE(S) classification for a phenotype-genotype nomenclature of cardiomyopathy: Endorsed by the World Heart Federation. J Am Coll Cardiol 62:2046, 2013. [PubMed: 24263073]

Bozkurt BJ et al: Current diagnostic and treatment strategies for specific dilated cardiomyopathies: A scientific statement from the American Heart Association. Circulation 134:e579, 2016. [PubMed: 27832612]

Cooper LT et al: The role of endomyocardial biopsy in the management of cardiovascular disease. Circulation 116:2216, 2007. [PubMed: 17959655]

Gilmore JD et al: Nonbiopsy diagnosis of cardiac transthyretin amyloidosis. Circulation 133:2404, 2016. [PubMed: 27143678]

HO CY: Genetic considerations in hypertrophic cardiomyopathy. Prog Cardiovasc Dis 54:456, 2012. [PubMed: 22687586]

Morillo CA et al: Randomized trial of benznidazole for chronic Chagas' cardiomyopathy. N Engl J Med 373:1295, 2015. [PubMed: 26323937]

Sagar S et al: Myocarditis. Lancet 379:738, 2011. [PubMed: 22185868]

Ware JS et al: Shared genetic predisposition in peripartum and dilated cardiomyopathies. N Engl J Med 374:233, 2016. [PubMed: 26735901]

Zamorano JL et al: 2016 ESC Position Paper on cancer treatments and cardiovascular toxicity developed under the auspices of the ESC Committee for Practice Guidelines. Eur Heart J 37:2768, 2016. [PubMed: 27567406]

Chapter 255: Cardiac Transplantation and Prolonged Assisted Circulation

Mandeep R. Mehra

INTRODUCTION

Advanced heart failure, a distinct syndrome, is characterized by refractoriness to conventional therapy and represents a vexing clinical dilemma that is associated with an increased symptom burden, frequent hospitalization, a poor quality of life and high risk of death. Such individuals do not tolerate neurohormonal antagonists at recommended doses, exhibit cardiorenal syndrome, maintain markedly poor cardiac reserve on cardiopulmonary stress testing, and typically display a low cardiac output state with elevated pulmonary pressures. In general, therapeutic targets shift away from disease modifying neurohormonal therapy to surgical options that attend directly to the myocardial stress and strain relationship. Most often, prolonged circulatory assistance using mechanical pumps or cardiac transplantation is required to reliably improve quality of life and long-term survival.

CARDIAC TRANSPLANTATION

A decade after Norman Shumway had accomplished the technique of a successful heart transplant in canines, Christiaan Barnard successfully performed the first human to human transplant on December 3, 1967. Now, 5 decades later, this surgery has become entrenched in the standard armamentarium for treating patients with advanced heart failure who are otherwise healthy enough to receive such a life altering treatment. Globally, >150,000 patients have undergone cardiac transplantation with a 1 year survival >80% and median survival of nearly 11 years. These gains have been ushered in due to advances in immunosuppression and identification and management of allograft rejection, as well as a comprehensive appreciation for late complications including accelerated coronary artery disease, malignancy, and renal failure.

CANDIDATES FOR CARDIAC TRANSPLANTATION

The demand for cardiac transplantation outstrips the availability of organ donors. Hence, attention to the optimal utility, equitable allocation, and patient autonomy must dominate the decisions to identify and list candidates for transplantation. Simultaneously, attempts at expanding the donor pool have surfaced. However, vigilance to evaluating candidates most likely to have a successful outcome from transplantation takes pre-eminence. In 2006, the International Society for Heart and Lung Transplantation identified a set of criteria to guide listing of patients. These criteria were updated in 2016 and include additional attention to the growing epidemiology of candidates suffering from congenital heart disease, restrictive and infiltrative cardiomyopathy (such as amyloidosis), and chronic infections in recipients (such as Chagas' disease, tuberculosis and hepatitides). Selected general principles for listing candidates for cardiac transplantation are enumerated in Table 255-1.

TABLE 255-1

Principles for Listing Candidates for Cardiac Transplantation

Principle	Comment
Advanced Disease Severity	Refractory heart failure with a VO_2 of <14 mL/kg/min (<12, if on beta blockers) or percent predicted VO_2 <50%; combination of intolerance to disease modifying therapy, cardiorenal syndrome, use of inotropic therapy to maintain stability or need for a left ventricular assist system.
Co-Morbidity	Age is not an absolute contraindication, but frailty should be considered a relative contraindication; a BMI > 35 kg/m^2 should require weight loss; cancer should be dealt with on an individual basis (e.g., low-grade prostate cancer may not be a contraindication); poorly controlled diabetes mellitus or end-organ damage may be a contraindication; eGFR <30 mL/min/1.73 m^2 is a relative contraindication; severe cerebrovascular disease or peripheral vascular disease (which will limit rehabilitation or function) is also a relative contraindication.
Donor-Recipient Matching	Sensitized individuals with circulating antibodies should have a prospective or virtual cross match; pulmonary vascular resistance with a transpulmonary gradient >15, PVR >3 Wood Units and absolute PA systolic pressure >50 mmHg provided the systolic BP is >85 mmHg is a relative contraindication unless reactive.
Psychosocial Issues	Tobacco use in any form limits posttransplant survival and should be stopped for at least 6-months; substance abuse, including marijuana, should be a contraindication if the individual cannot demonstrate control and cessation; patients with severe cognitive-behavioral disabilities or dementia (inability to ever understand and cooperate with medical care) have the potential for self-harm and should not receive a transplant.

Abbreviations: BMI, body mass index; eGFR, estimated glomerular filtration rate; PA, pulmonary artery; PVR, pulmonary vascular resistance; VO_2, peak oxygen consumption.

PRINCIPLES OF DONOR RECOVERY AND ALLOCATION

Although listing criteria for candidates are typically adjudicated at a center level, organ allocation is handled by national regulatory processes in most countries. The allocation of donor hearts is based on (a) the urgency of the clinical situation, (b) the time spent on the waiting list, and (c) the distance from the recipient center. Thus, candidates who are hospitalized and require temporary mechanical cardiac support devices or daily invasive hemodynamic evaluation and intravenous inotropic therapy to maintain stability are given the highest urgency status, while those able to ambulate and live at home receive a lower urgency status. The geographical regional reach for allocation is based not only on territorial considerations but also on the time that a donor heart would be in transit and therefore in out of body "cold ischemia time," which is typically limited to 4 h. The final key feature that is included in the allocation offer relates to the ABO blood group. Donor organs are offered based on these initial characteristics and then a more detailed donor assessment ensues, resulting in acceptance or decline for any given donor heart. It is important to note that the time constraints imposed on the retrieval process make it difficult to invoke HLA matching of the donor and recipient. In cases where there is a high likelihood of sensitization in the recipient (preformed circulating antibodies against donor antigens), a prospective or

virtual cross match is entertained prior to acceptance. Other clinical criteria that are employed in the decision on accepting an offered donor include the donor-recipient size match, the age of the donor (typically restricted to under 55 years) and presence or absence of concomitant pathology such as coronary artery disease, left ventricular hypertrophy or severe injury to the allograft manifest by excess leak of injury markers (troponins) or poor contractile performance. In many cases, the prospective cardiac allograft can be reconditioned by use of hormonal therapy (including thyroid hormone supplementation) and used for transplantation even if the initial evaluation suggests poor function. In efforts to enhance the donor pool, systems that allow ex-vivo normothermic perfusion to evaluate and reanimate organs with a prolonged out of body time are being developed. The classic heart donor is derived from a donor with brain death; however, donors with circulatory death are being increasingly evaluated as candidates for cardiac reanimation using a variety of techniques including ex-vivo reanimation and subsequent transplantation.

SURGERY FOR CARDIAC TRANSPLANTATION

The most common contemporary operation is referred to as a “bicaval” orthotopic cardiac transplant that mimics the natural anatomic position. In this operation, the donor and recipient superior and inferior vena cava are connected as are the aortic and pulmonary great vessels. The left atrium of the recipient retains its roof including the draining pulmonary veins and the donor left atrium is then sutured to the retained atrial tissue. This technique maintains function of the donor right atrium, important for governing early postoperative right heart output, and may prevent atrial arrhythmias. The recipient is left with a surgical denervation and the allograft is not responsive to any direct sympathetic or parasympathetic stimuli. Therefore, early in the adaptive postoperative phase, high-dose catecholamines are required to maintain adequate function. Due to denervation, bradycardia in a cardiac allograft cannot be treated with atropine and the drug of choice is isoproterenol. Once the cardiac allograft adapts to its host circulation, the function is usually adequate at rest and with exercise to provide normal physical activity.

CARDIAC ALLOGRAFT REJECTION AND IMMUNOSUPPRESSION

The ability to perform endomyocardial biopsies, evaluate rejection pathologically and the introduction of the immunosuppression agent cyclosporine heralded cardiac transplantation as a viable clinical therapy. Triple drug immunosuppression, which includes a calcineurin inhibitor (cyclosporine or tacrolimus), corticosteroids and anti-proliferative immunosuppression (azathioprine, mycophenolate mofetil, sirolimus or everolimus) is now the standard cocktail used. The combination that is most commonly used and achieves the best standard outcomes includes the combination of tacrolimus, mycophenolate mofetil, and prednisone. In those at high risk for rejection (multiparous women, sensitized individuals) or in situations where use of calcineurin inhibitors is delayed (renal dysfunction), induction therapy using monoclonal (basiliximab) or polyclonal antibodies (anti-thymocyte globulin) to provide augmented immunosuppression is used. The typical management strategy includes gradual weaning of steroids over time as surveillance endomyocardial biopsies are performed and clinical as well as sub-clinical pathological quiescence is established. **Table 255-2** describes the immunosuppression drugs in common use.

TABLE 255-2

Immunoprophylaxis Drugs in Cardiac Transplantation

Drug Class	Generic Drug	Cellular Target	Major Side Effects
Calcineurin Inhibitors	Cyclosporine	Binds to cyclophilin which then inhibits calcineurin	Hypertension, dyslipidemia, gum hypertrophy, hypertrichosis
	Tacrolimus	Binds to immunophilin FK506 binding protein which inhibits calcineurin	Hypertension, dyslipidemia, alopecia, diabetes mellitus
Anti-Thymocyte Globulin (ATG)	Rabbit ATG	T-cell depletion in blood and peripheral lymphoid tissues through complement-dependent lysis and T-cell activation and apoptosis	Cytokine release syndrome, leukopenia, thrombocytopenia, serum sickness
	Horse ATG	Same as above	Same as above
Interleukin-2 receptor antagonists	Basiliximab	Inhibition of CD-25 of IL-2 receptor	Well tolerated; rare hypersensitivity; increased infection risk
Anti-metabolites	Azathioprine	Imidazolyl derivative and prodrug of 6-mercaptopurine (cell cycle inhibitor)	Bone marrow suppression, pancreatitis, hepatitis
	Mycophenolate Mofetil	Inhibits inosine monophosphate dehydrogenase, which controls guanine monophosphate in the de novo pathway of purine synthesis (inhibits T and B cell proliferation)	Leukopenia, gastrointestinal toxicity
Proliferation-Signal Inhibitors	Sirolimus	Binds with FKBP12 and complex inhibits the mechanistic Target of Rapamycin (mTOR)	Delayed wound healing, non-specific pneumonia, pericardial effusion, hyperlipidemia (hypertriglyceridemia)
	Everolimus	Binds to FKBP12, which inhibits mTORC1 (and not mTORC2)	Dyslipidemia, stomatitis, pericardial effusions and pancytopenia

Acute cellular rejection (ACR) and antibody-mediated rejection (AMR) are two separate forms of cardiac allograft rejection that are recognized and can sometimes coexist. ACR occurs early after transplantation and then declines in incidence after 6 months. This occurs due to a T cell–mediated assault on the donor allograft tissue and histologically is characterized by lymphocytic infiltrates. In mild cases these infiltrates are localized to the peri-venular regions, and in severe cases progresses diffusely into the cardiac interstitium. In late stages of severe ACR, most often associated with hemodynamic compromise, multi-clonal cells such as macrophages, neutrophils, and eosinophils are observed with intramyocardial hemorrhage, myocyte injury and myocyte necrosis. Subclinical ACR is typically treated with high doses of corticosteroid

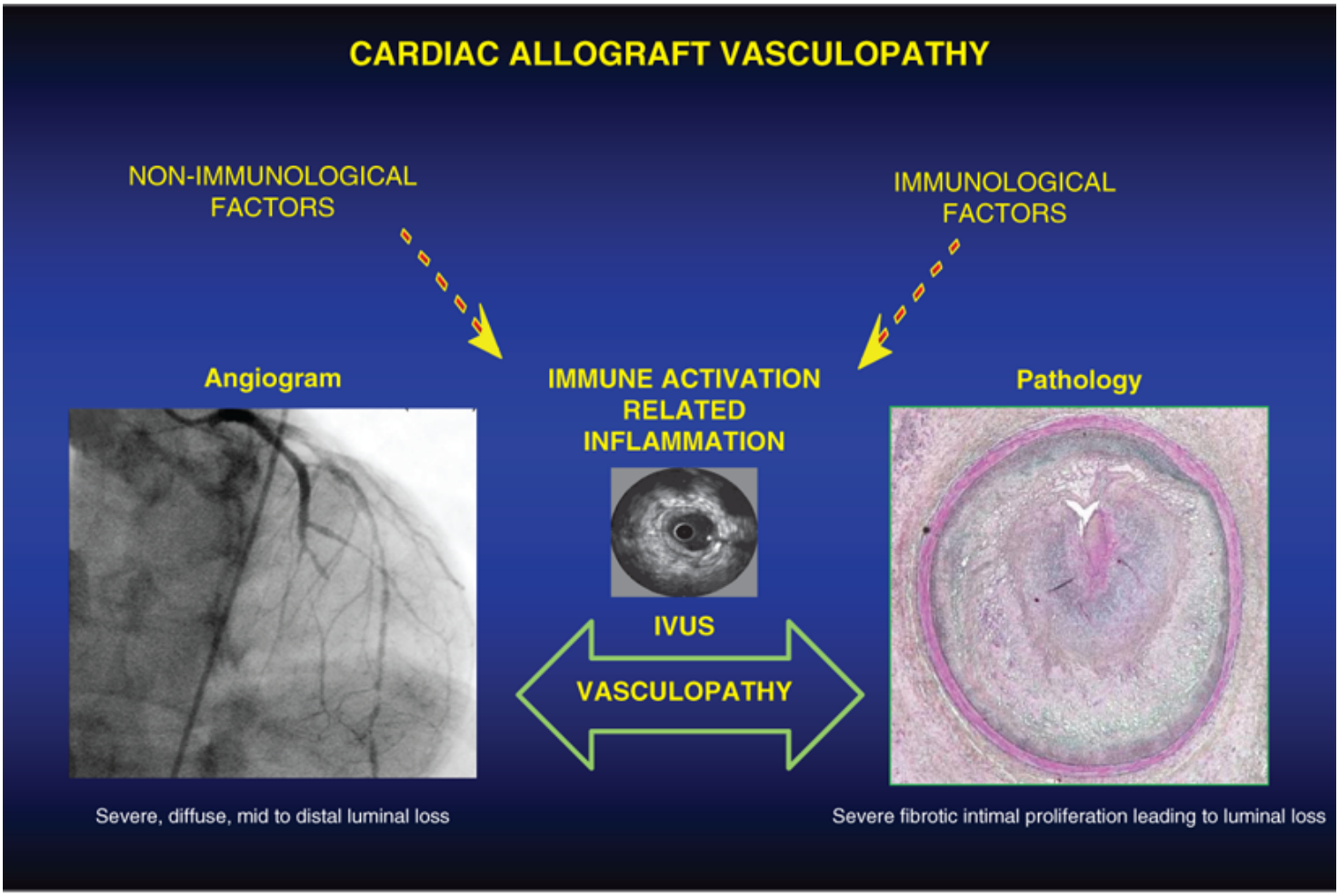
pulses although some centers choose to simply observe mild forms of infiltration since it is known that these recover longitudinally. If hemodynamic compromise occurs, rescue polyclonal antibodies are used in tandem with corticosteroids. Conversely, AMR is immunologically described as a non-cellular antibody-driven phenomenon associated with a pattern of immunopathologic findings of immunoglobulin deposition and complement fixation on immunofluorescence, along with histopathologic findings of endothelial swelling and interstitial edema and cardiac allograft arteriolar vasculitis. AMR is characterized by the emergence of circulating donor-specific antibodies that are thought to fix, complement, and bind to the allograft. Commonly, AMR leads to acute allograft dysfunction and increases the risk for cardiac allograft vasculopathy, and results in worsened cardiac allograft survival compared with ACR. In this form of rejection, therapy is directed towards suppression and removal of circulating antibodies using plasmapheresis and drugs such as rituximab (chimeric monoclonal antibody directed against the CD20 antigen) or in refractory cases, bortezomib (a proteasome inhibitor) or eculizumab (a terminal complement inhibitor). The treatment with immunosuppression requires prophylaxis for opportunistic infections and ongoing surveillance and expertise in recognizing the more common clinical presentations of infections caused by cytomegalovirus (CMV), aspergillus, and other opportunistic agents such as nocardia and toxoplasmosis.

LATE COMPLICATIONS AFTER CARDIAC TRANSPLANTATION

The long-term consequences of exposure to chronic immunosuppression result in a variety of non-immunological cardio-metabolic effects such as hypertension and hyperlipidemia as well as systemic disorders of bone loss and renal dysfunction. One aggressive complication that limits late survival of cardiac allografts includes the development of an accelerated form of coronary artery disease, referred to as cardiac allograft vasculopathy (CAV). This is characterized by a proliferative thickening of the vascular intima of the vasculature that is initiated as a diffuse endothelialitis in the setting of the confluence of the consequences of brain death, ischemia reperfusion injury during the transplant process and early immunological insults. Chronically, the metabolic consequences of hypertension, hyperlipidemia, and disordered glucose regulation result in a further worsening of vascular lesions that are diffuse and noted throughout the coronary tree. Early diagnosis and preventative therapy are critical since it is commonly silent in genesis. Statins, antihypertensive agents, and anti-CMV agents all have demonstrated benefits in reducing CAV. Anti-proliferative immunosuppressive therapy such as mycophenolate mofetil and sirolimus or everolimus prevent vascular intimal thickening compared with azathioprine-based regimens. However, retransplantation is the only definitive form of therapy for advanced allograft CAV **(Fig. 255-1)**.

FIGURE 255-1

Cardiac allograft vasculopathy is initiated and propagated by the combined influence of immunological and non-immunological insults on the allograft vasculature. An inflammatory milieu determines the development of diffuse, aggressive luminal blockages that in early forms exhibit intimal thickening and fibrosis. IVUS, Intravascular Ultrasound (can be used to diagnose early forms of intimal thickening).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Another consternation in cardiac transplantation is the development of malignancy with a greater frequency than in the normal population, suggesting that immunosuppression plays a sentinel role in its generation. Posttransplant lymphoproliferative disorders, typically driven by Epstein-Barr virus, occur most frequently and require a reduction in immunosuppression, administration of antiviral agents, and traditional chemo- and radiotherapy. Specific antilymphocyte (targeted against CD20) therapy has also shown promise. Solid cancers most often manifest as skin malignancies (both basal cell and squamous cell carcinomas), and use of sun-screens is advised. Future research is required to define strategies for immune modulation, immune suppression, and malignancy prevention, however the impact of decreasing immunosuppression in the treatment of these cancers is unclear.

PROLONGED ASSISTED CIRCULATION

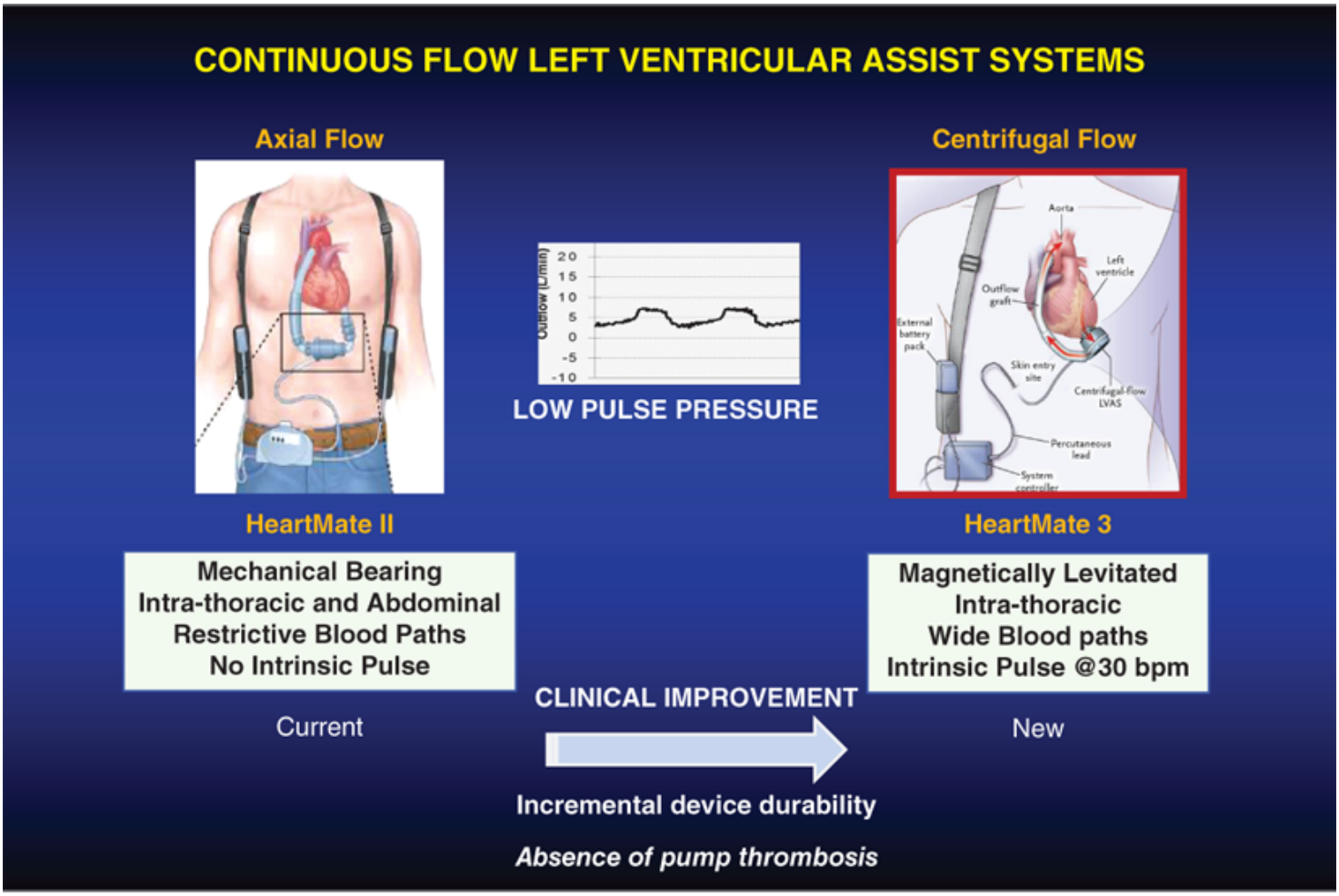
The quest for a prolonged implantable mechanical circulatory support device has led to the development of continuous flow left ventricular assist systems (LVAS). Initially designed for short-term support as a bridge to recovery or to cardiac transplantation, the most frequent use today entails permanent support for lifetime therapy (“destination therapy”). The decision to implant LVAS dichotomously as either a bridge to transplantation or for destination therapy is not always clear and in several instances, these devices are used as a “bridge to decision” (in those with potentially reversible underlying relative contraindications such as renal insufficiency or pulmonary hypertension, who may become future candidates for transplantation).

LEFT VENTRICULAR ASSIST SYSTEMS AND CLINICAL TRIALS

A pivotal trial, REMATCH, published in 2001, was the first study to reliably demonstrate that survival of transplant ineligible refractory, predominantly inotropic therapy supported heart failure is improved by implantation of a LVAS. This study used an early generation pulsatile flow device and demonstrated a 48% reduction in risk of death. However, the LVAS used was of limited durability and median meaningful "out of hospital" survival was prolonged by only 5 months. Furthermore, complications of strokes, multisystem organ failure, and infections reduced enthusiasm for widespread adoption. Over time, continuous flow systems that were small turbo-pumps with minimal moving parts and no valves were introduced, leading to a more generalized world-wide adoption. A landmark trial compared the older bulky pulsatile LVAS to the newer generation axial continuous flow LVAS, the HeartMate II, and demonstrated a marked improvement in short- and long-term survival, along with an improvement in functional capacity and meaningful quality of life enhancement. A centrifugal continuous flow LVAS, the HeartWare HVAD, is also in common use. A newer centrifugal device, with a fully magnetically levitated system, the HeartMate 3 LVAS is now available. Unlike the HeartMate II LVAS, which requires an abdominal pump pocket, this smaller device is fully implanted in the pericardial space in the thoracic cavity (**Fig. 255-2**). Real world experience from registry analyses has pointed to a >70% 2-year survival with currently available LVAS, nearly approaching the outcomes achieved with transplantation, however long-term durability beyond 5–10 years remains a question. The patients for whom LVAS should ideally be employed include those with severe persistent systolic heart failure symptoms who have failed to respond to optimal medical management. Commonly, these patients have marked functional limitation indicated by a peak oxygen consumption of <12 mL/kg/min; or the patient is bound to continuous intravenous inotropic therapy owing to symptomatic hypotension, decreasing renal function, or worsening congestion. Currently, the role of LVAS in "less sick" patients (those with moderate symptoms) is not advocated since sufficient equipoise does not exist due to the adverse risk benefit ratio from device-related complications.

FIGURE 255-2

Continuous flow left ventricular assist systems (LVAS), their types and mechanisms. The mechanical bearing axial flow HeartMate II pump is prone to thrombosis, while the frictionless magnetically levitated centrifugal flow HeartMate 3 does not induce hemolysis or pump thrombosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MANAGEMENT OF LVAS AND THEIR COMPLICATIONS

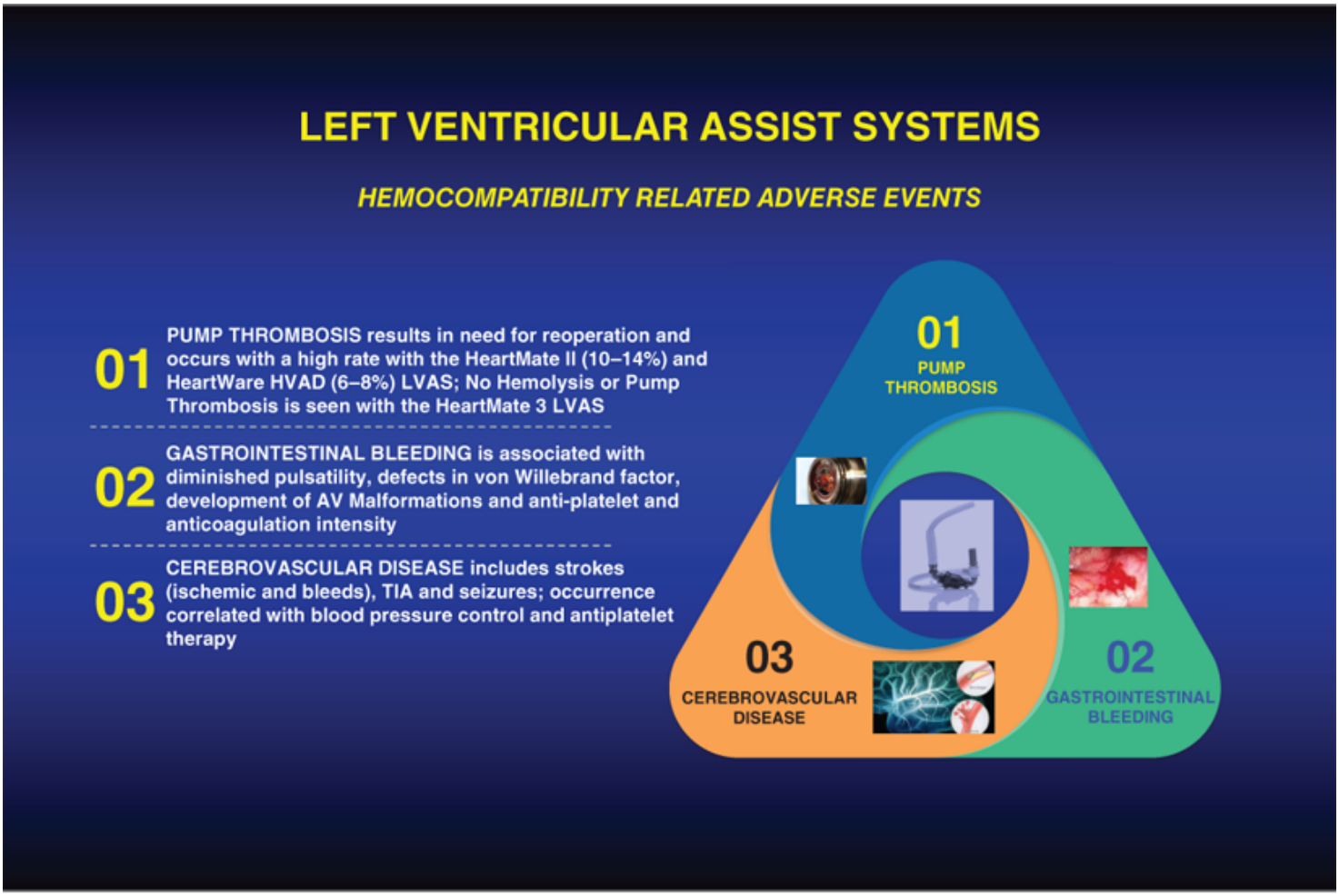
Continuous flow LVAS rely on pressure gradients between the left ventricular cavity and the aorta. As such, forward flow is critically dependent on management of systemic blood pressure. Due to the non-pulsatile nature of the blood flow, blood pressure is measured by using a Doppler ultrasound (mean or opening blood pressure, which is less than the systolic blood pressure) since a peripheral pulse is usually not detectable. The ideal mean arterial blood pressure should be kept to <90 mmHg and antihypertensive drug therapy prescribed using RAAS drugs or other vasodilators. A common complication encountered in patients is that of an anemia, often due to iron deficiency. The blood flow path through current devices results in increased shear stress which is manifested in the form of low grade hemolysis and the development of an acquired von Willebrand disease due to loss of high molecular weight multimers. This hematological aberration has been associated with a risk of gastrointestinal bleeding, particularly resulting from arteriovenous malformations in the intestines.

The unsupported right ventricle often demonstrates failure and results in congestion requiring diuretic therapy. While unloading of the left ventricle decreases right-sided afterload, increased device flow results in a greater right heart preload and effects of the LVAS on the septum reduce right ventricular contractile efficiency, leading to development of right ventricular dilatation and maladaptation between the right ventricle and pulmonary circuit. Cardiac arrhythmias are common in patients supported with LVAS and often require antiarrhythmic therapy for quiescence since such events can trigger low flow through the device.

Hemocompatibility-related adverse outcomes include neurological events (ischemic and hemorrhagic strokes), device-related thrombosis leading to pump malfunction, and non-surgical bleeding complications (**Fig. 255-3**). Antiplatelet therapy using aspirin in doses of 81–325 mg daily along with warfarin targeted to an INR of 2–3 are required for current LVAS to avoid the morbidity of hemocompatibility-related adverse events. On one hand, this therapy is protective for thrombotic complications while on another it predisposes the patient to bleeding complications. Strokes occur with a frequency ranging from 8% with the HeartMate II LVAS to as much as 29% with the HeartWare HVAD device, by 2 years of treatment. Optimal control of blood pressure is associated with improved rates of strokes; however, this complication is a critical reason for lack of adoption of device therapy to the less sick population. Another cause of morbidity is pump thrombosis requiring reoperation for device malfunction. This complication is noted in 6–12% of LVAS implants, occurs early (in the first 6 months), and is more common with the HeartMate II device than with other LVAS. The subclinical phase of LVAS thrombosis is characterized by increasing hemolysis and elevation in the device power. Progressively, inability to "unload" the left ventricle is manifest leading to decompensated heart failure and possibly hemodynamic compromise. Lactate dehydrogenase (LDH) is an excellent (although non-specific) biomarker of hemolysis and hence impending or established pump thrombosis. Patients who have suspected left ventricular assist device (LVAD) thrombosis and do not undergo LVAD exchange or cardiac transplantation have a 6-month mortality rate of 48%, inferring that medical therapy for VAD thrombosis may be inadequate (or cause harm in the case of thrombolytic use). Reoperation (pump exchange) carries a modest 6.5% perioperative mortality risk and a 65% 2-year survival following exchange.

FIGURE 255-3

Hemocompatibility-related adverse events with LVAS are often interrelated and typically result in cerebrovascular, gastrointestinal, or pump malfunction events.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Infection is common, most often involving the driveline (the conduit connecting the device to the external controller and batteries) and occurs in 1 in 5 patients following LVAS implant. Such an infection is treated with local internal exploration and requires long-term suppressive antibiotics unless the patient undergoes cardiac transplantation or the device is exchanged. Infection and its inflammatory sequelae predispose to thrombosis and heighten the risk of neurological complications, leading to a worsening milieu in hemocompatibility.

NOVEL DEVICES

The HeartMate 3 is a centrifugal, continuous flow pump that is placed in the thorax and is engineered to be a more hemocompatible LVAS. This device is constructed with a fully magnetically levitated motor, offers wider blood flow paths, and even exhibits a fixed intrinsic pulse (by the motor ramping its speed up and down at 2 s intervals). This pump has been tested in the large MOMENTUM 3 trial, the early 6 month results for which suggest that this LVAS incrementally improves outcome compared with the HeartMate II device. Importantly, this pump does not exhibit hemolysis or shear high molecular weight multimers of von Willebrand antigen as with other devices and is not associated with pump thrombosis. Long-term experience with this LVAS will be important in discerning whether these early short-term findings result in a reduction in hemocompatibility-related adverse events, improved quality of life, and survival.

TOTAL ARTIFICIAL HEART

Not all patients are candidates for a LVAS, particularly those with severe right-sided heart failure or conditions that do not allow placement of an LVAS (restrictive cardiomyopathy, massive anterior myocardial infarction, complex congenital heart

disease). In such patients, either a biventricular assist device approach or a total artificial heart pump can be considered. The SynCardia total artificial heart is a pulsatile, implantable pump that consists of two polyurethane ventricles with pneumatically driven diaphragms, and four tilting disc valves. This requires excision of the native ventricles and thus cannot be employed as a myocardial recovery strategy. There are specific clinical issues that are unique to the total artificial heart management. This device operates on a steep physiological curve and has little adaptability to tolerate either systemic blood pressure changes or large shifts in blood volume. As the ventricles are excised, most patients exhibit a sharp decline in renal function due to the loss of natriuretic peptide expression by the myocardium. Severe hemolysis is common due to the presence of four mechanical valves and aberrant erythropoiesis is noted leading to a severe refractory anemia. Newer artificial hearts using biocompatible surfaces are under development, as well as those that use continuous flow technology.

GLOBAL CONSIDERATIONS

 While LVAS are available worldwide, their use and indications vary from country to country. In the United States, payers require discrete discrimination of indication into either a bridge to transplant or destination therapy, whereas in most European countries this artificial segregation is not used. Cost effectiveness studies suggest improvement with newer devices, yet some countries only allow use of this technology as a bridge to transplantation (UK), awaiting more definitive long-term studies for lifetime use. Now, the use in moderately symptomatic ambulatory patients with chronic systolic heart failure is equally discouraged throughout the world, awaiting the availability of a more hemocompatible LVAS.

FURTHER READING

Mehra MR: Contemporary concepts in prevention and treatment of cardiac allograft vasculopathy. Am J Transplant 6:1248, 2006. [PubMed: 16686747]

Mehra MR et al: International Society for Heart and Lung Transplantation working formulation of a standardized nomenclature for cardiac allograft vasculopathy-2010. J Heart Lung Transplant 29:717, 2010. [PubMed: 20620917]

Mehra MR et al: The 2016 International Society for Heart Lung Transplantation listing criteria for heart transplantation: A 10-year update. J Heart Lung Transplant 35:1, 2016. [PubMed: 26776864]

Mehra MR et al: MOMENTUM 3 Investigators. A fully magnetically levitated circulatory pump for advanced heart failure. N Engl J Med 376:440, 2017. [PubMed: 27959709]

Nair N et al: Current and future challenges in therapy for antibody-mediated rejection. J Heart Lung Transplant 30:612, 2011. [PubMed: 21474341]

Nair N et al: Long-term immunosuppression and malignancy in thoracic transplantation: Where is the balance? J Heart Lung Transplant 33:461, 2014. [PubMed: 24742695]

Rogers JG et al: Intrapericardial left ventricular assist device for advanced heart failure. N Engl J Med 376:451, 2017. [PubMed: 28146651]

Shah SP, Mehra MR: Durable left ventricular assist device therapy in advanced heart failure: Patient selection and clinical outcomes. Indian Heart J 68:S45, 2016. [PubMed: 27056652]

Stehlik J et al: Organ allocation around the world: Insights from the ISHLT International Registry for Heart and Lung Transplantation. J Heart Lung Transplant 33:975, 2014. [PubMed: 25242122]

Stewart GC, Mehra MR: A history of devices as an alternative to heart transplantation. Heart Fail Clin 10:S1, 2014. [PubMed: 24262348]

Chapter 256: Aortic Valve Disease

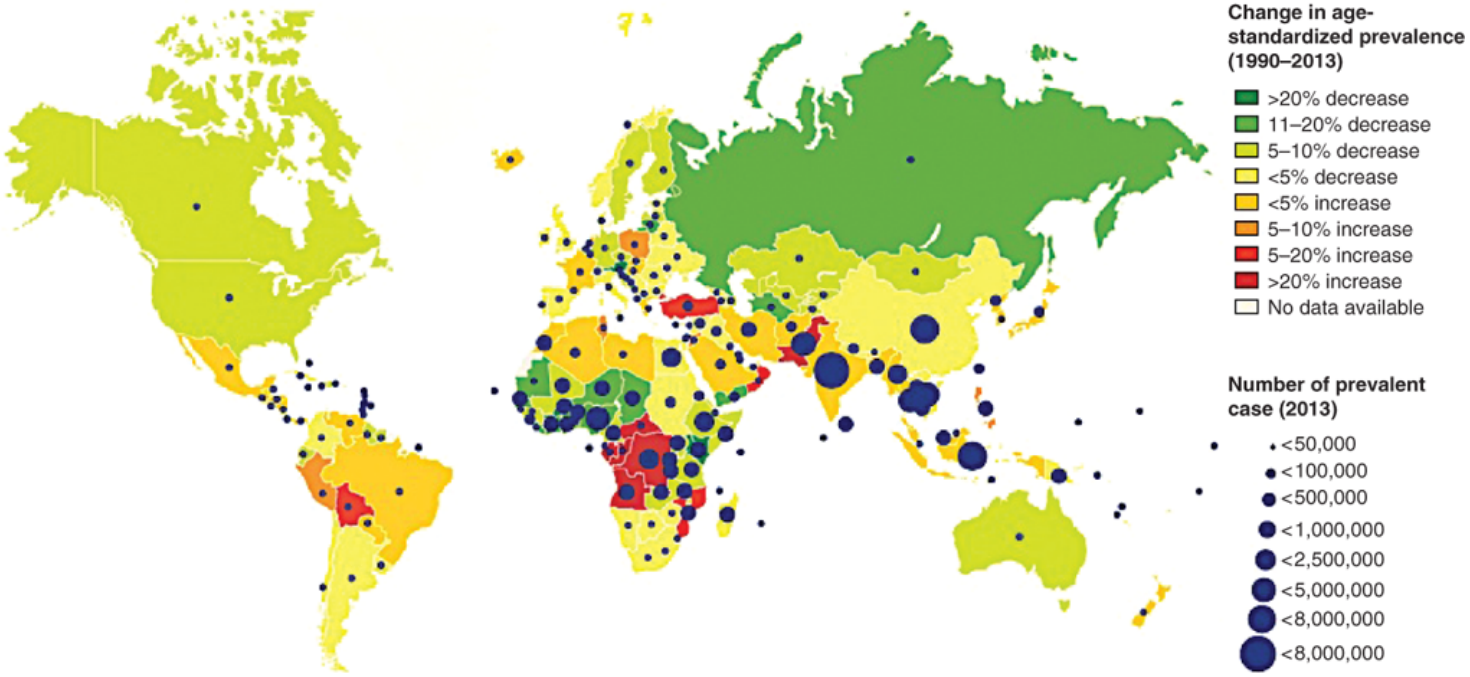
Patrick T. O'Gara; Joseph Loscalzo

GLOBAL BURDEN OF VALVULAR HEART DISEASE

Primary valvular heart disease ranks well below coronary heart disease, stroke, hypertension, obesity, and diabetes as a major threat to the public health. Nevertheless, it can cause significant morbidity and lead to premature death. Rheumatic fever (Chap. 352) is the dominant cause of valvular heart disease in developing and low-income countries. Its prevalence has been estimated to range from as low as 1 per 100,000 school-age children in Costa Rica to as high as 150 per 100,000 in China (Fig. 256-1). Rheumatic heart disease accounts for 12–65% of hospital admissions related to cardiovascular disease and 2–10% of hospital discharges in some developing countries. Prevalence and mortality rates vary among communities even within the same country as a function of overcrowding and the availability of medical resources and population-wide programs for detection and treatment of group A streptococcal pharyngitis. In economically deprived areas, tropical and subtropical climates (particularly on the Indian subcontinent and in Southeast Asia), Central America, and the Middle East, rheumatic valvular disease progresses more rapidly than in more-developed nations and frequently causes serious symptoms in patients aged <20 years. This accelerated natural history may be due to repeated infections with more virulent strains of rheumatogenic streptococci. Approximately 15–20 million people live with rheumatic heart disease worldwide, an estimated prevalence characterized by 300,000 new cases and 233,000 case fatalities per year, with the highest mortality rates reported from Southeast Asia (~7.6 per 100,000). In the United States, rheumatic heart disease accounted for 20,000 hospital admissions in 2010 and 3281 deaths in 2014.

FIGURE 256-1

The global burden of rheumatic heart disease. This world map provides a snapshot of both the change in prevalence of rheumatic heart disease cases between 1990 and 2013 (upper right legend) and the estimated number of rheumatic heart disease cases per country (lower right legend). Regions in which the disease is highly prevalent include sub-Saharan Africa, India, China, and Southeast Asia. *(From JR Carapetis et al: Nat Rev Dis Primers 2:15084, 2016.)*



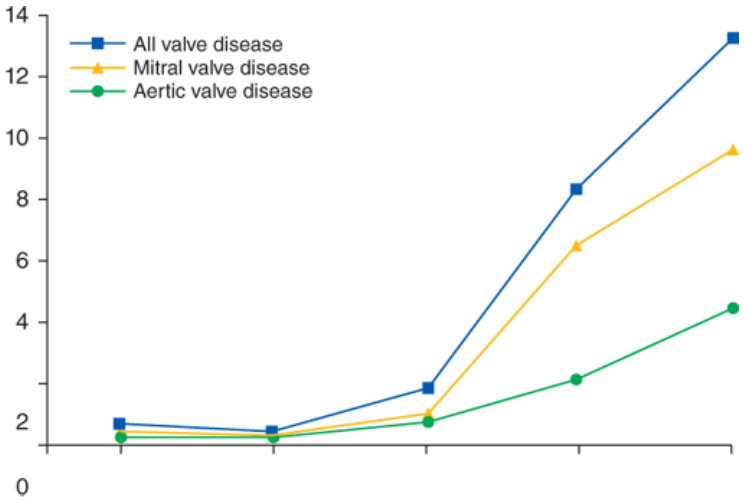
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although there have been recent reports of isolated outbreaks of streptococcal infection in North America, valve disease in high-income countries is dominated by degenerative or inflammatory processes that lead to valve thickening, calcification, and dysfunction. The prevalence of valvular heart disease increases significantly with age for both men and women. Important left-sided valve disease may affect as many as 12–13% of adults aged >75 years (Fig. 256-2). Severe aortic stenosis (AS) is estimated to affect 3.5% of the population aged >75 years. In the United States, there were 85,000

hospital discharges with valvular heart disease in 2010, and the vast majority of these were related to surgical procedures for heart valve disease (mostly involving the aortic and mitral valves).

FIGURE 256-2

The burden of moderate or severe mitral and aortic valve disease in the United States. Prevalence estimates are derived from three population-based studies comprising a total of 11,911 individuals: The Coronary Artery Risk Development in Young Adults (CARDIA), the Atherosclerosis Risk in Communities (ARIC), and the Cardiovascular Health Study (CHS). *(From VT Nkomo et al: Lancet 368:1005, 2006.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The incidence of infective endocarditis (Chap. 123) has increased with the aging of the population, the more widespread prevalence of vascular grafts and intracardiac devices, the emergence of more virulent multidrug-resistant microorganisms, and the growing epidemic of diabetes. The more restricted use of antibiotic prophylaxis since 2007 has thus far not been convincingly associated with an increase in incidence rates for infective endocarditis cases attributable to oropharyngeal pathogens. Infective endocarditis has become a relatively more frequent cause of acute valvular regurgitation. Rates of valve surgery during the acute phase of infective endocarditis increased significantly in U.S. hospitals over the decade from 2000 to 2010.

Bicuspid aortic valve (BAV) disease affects as many as 0.5–1.4% of the general population, with an associated incidence of aortopathy involving root or ascending aortic aneurysm disease or coarctation. An increasing number of childhood survivors of congenital heart disease present later in life with valvular dysfunction. The global burden of valvular heart disease is expected to progress.

As is true for many other chronic health conditions, disparities in access to and quality of care for patients with valvular heart disease have been well documented, especially for those patients with rheumatic heart disease in low- and middle-income countries. Management decisions and outcome differences based on age, gender, race, and geography require educational efforts across all levels of providers and prioritization of resources.

The role of the physical examination in the evaluation of patients with valvular heart disease is also considered in Chaps. 38 and 234; of electrocardiography (ECG) in Chap. 235; of echocardiography and other noninvasive imaging techniques in Chap. 236; and of cardiac catheterization and angiography in Chap. 237.

AORTIC STENOSIS

Aortic stenosis (AS) occurs in about one-fourth of all patients with chronic valvular heart disease; ~80% of adult patients with symptomatic, valvular AS are male.

ETIOLOGY AND PATHOGENESIS

AS in adults is due to degenerative calcification of the aortic cusps and occurs most commonly on a substrate of congenital disease (BAV), chronic (trileaflet) deterioration, or previous rheumatic inflammation (Table 256-1) . A pathologic study of specimens removed at the time of aortic valve replacement for AS in adults showed that 53% were bicuspid and 4% unicuspid. The process of aortic valve deterioration and calcification is not a passive one, but rather one that shares many features with vascular atherosclerosis, including endothelial dysfunction, lipid accumulation, inflammatory cell activation, cytokine release, and upregulation of several signaling pathways (Fig. 256-3). Eventually, a fibrocalcific response is established wherein collagen is deposited and valvular myofibroblasts differentiate phenotypically into osteoblasts and actively produce bone matrix proteins that allow for the deposition of calcium hydroxyapatite crystals. Genetic polymorphisms involving the vitamin D receptor, the estrogen receptor in postmenopausal women, interleukin 10, and apolipoprotein E4 have been linked to the development of calcific AS, and a strong familial clustering of

cases has been reported from western France. Several traditional atherosclerotic risk factors have also been associated with the development and progression of calcific AS, including low-density lipoprotein (LDL) cholesterol, lipoprotein a (Lp[a]), diabetes mellitus, smoking, chronic kidney disease, and the metabolic syndrome. The presence of aortic valve sclerosis (focal thickening and calcification of the leaflets not severe enough to cause obstruction) is associated with an excess risk of cardiovascular death and myocardial infarction (MI) among persons aged >65. Approximately 30% of persons aged >65 years exhibit some degree of aortic valve sclerosis. Rate and extent of progression to valve obstruction (stenosis) vary among individual patients.

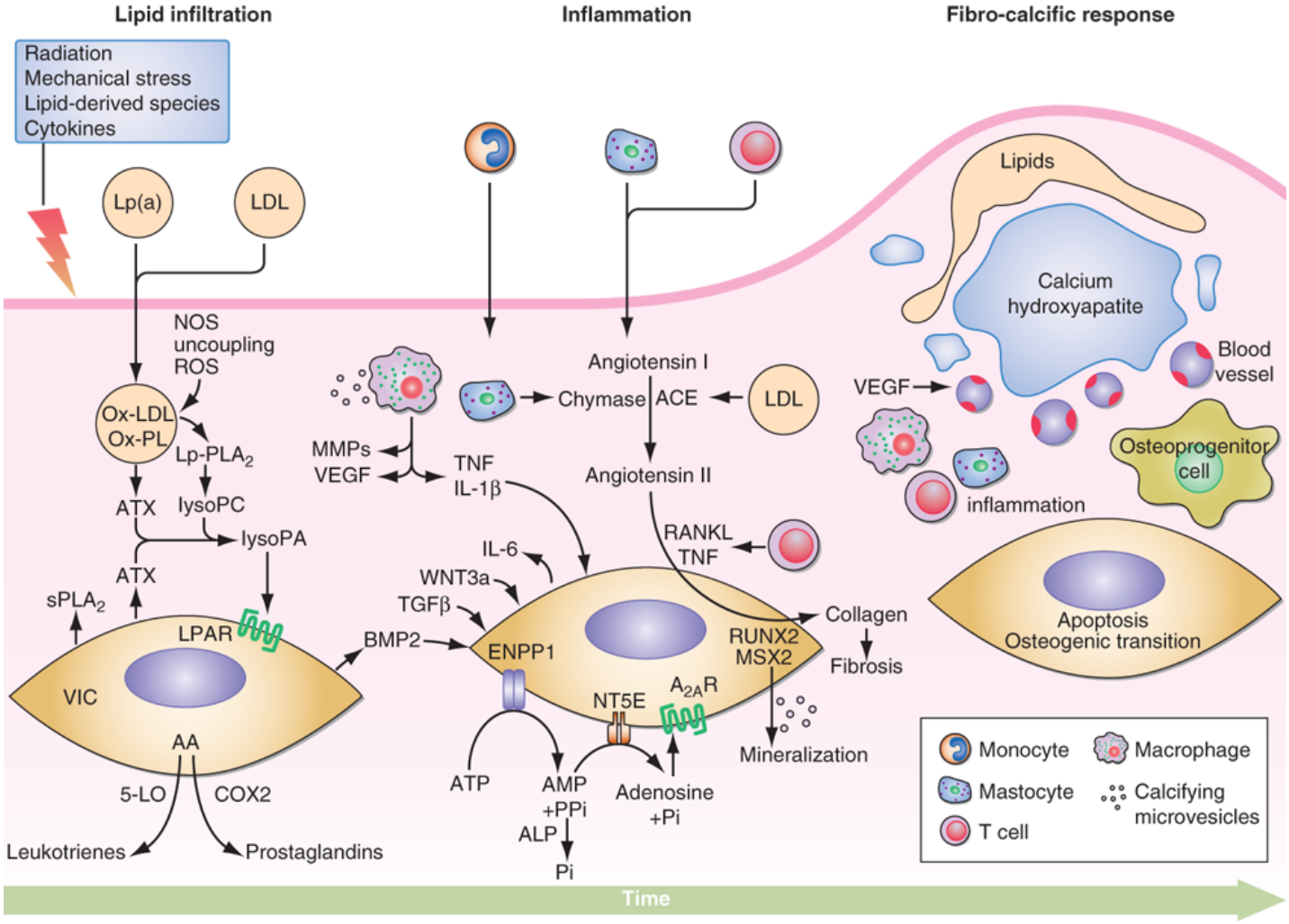
TABLE 256-1

Major Causes of Aortic Stenosis

Valve Lesion	Etiologies
Aortic stenosis	Congenital (bicuspid, unicuspid)
	Degenerative calcific
	Rheumatic fever
	Radiation

FIGURE 256-3

Pathogenesis of calcific aortic stenosis. Lipid and inflammatory cell infiltration occurs across damaged endothelium. A cascade of events follows that leads eventually to formation of disorganized collagen (fibrosis) and calcium hydroxyapatite (bone) deposition. Valvular interstitial cells (VIC) are critical participants in this active process. AA, arachidonic acid; ACE, angiotensin-converting enzyme; ALP, alkaline phosphatase; ApoB, apolipoprotein B; AMP, adenosine monophosphate; ATP, adenosine triphosphate; ATX, autotaxin; A2AR, adenosine A2A receptor; BMP, bone morphogenetic protein; COX2, cyclo-oxygenase2; ENPP, ectonucleotide pyrophosphatase/phosphodiesterase; IL, interleukin; 5-LO, 5-lipoxygenase; LDL, low-density lipoprotein; Lp(a), lipoprotein(a); LPAR, lysophosphatidic acid receptor; Lp-PLA2, lipoprotein-associated phospholipase A2; lysoPA, lysophosphatidic acid; lysoPC, lysophosphatidylcholine; MMP, matrix metalloproteinase; NOS, nitric oxide synthase; Ox-PL, oxidized phospholipid; Ox-LDL, oxidized LDL; RANKL, receptor activator of nuclear factor-κB ligand; ROS, reactive oxygen species; RUNX2, runt-related transcription factor 2; sPLA2, secreted PLA2; TGFβ, transforming growth factor-β; TNF, tumor necrosis factor; VEGF, vascular endothelial growth factor; VIC, valvular interstitial cell. *(From B Lindman et al: Nat Rev Dis Primers 2:16006, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Rheumatic disease of the aortic leaflets produces commissural fusion, sometimes resulting in a bicuspid-appearing valve. This condition, in turn, makes the leaflets more susceptible to trauma and ultimately leads to fibrosis, calcification, and further narrowing. By the time obstruction to left ventricular (LV) outflow causes serious clinical disability, the valve is usually a rigid calcified mass, and careful examination may make it difficult or even impossible to determine the etiology of the underlying process. Rheumatic AS is almost always associated with involvement of the mitral valve and with aortic regurgitation (AR). Mediastinal radiation can also result in late scarring, fibrosis, and calcification of the leaflets with AS.

BICUSPID AORTIC VALVE DISEASE

A bicuspid aortic valve (BAV) is the most common congenital heart valve defect and occurs in 0.5–1.4% of the population with a 2–4:1 male-to-female predominance. The inheritance pattern appears to be autosomal dominant with incomplete penetrance, although some have questioned an X-linked component as suggested by the prevalence of BAV disease among patients with Turner's syndrome. The prevalence of BAV disease among first-degree relatives of an affected individual is ~10%. A single gene defect to explain the majority of cases has not been identified, although a mutation in the *NOTCH1* gene has been described in some families. Abnormalities in endothelial nitric oxide synthase and NKX2.5 have been implicated as well. Medial degeneration with ascending aortic aneurysm formation occurs commonly among patients with BAV disease; aortic coarctation is less frequently encountered. Patients with BAV disease have larger aortas than patients with comparable tricuspid aortic valve disease. The aortopathy develops independently of the hemodynamic severity of the valve lesion, but directional shear forces dictated by the anatomic configuration of the valve may influence its expression. For example, enlargement of the ascending aorta along its greater curvature is most often associated with right-left cusp fusion, the most common bicuspid variant. Patients with BAV disease are at risk for aneurysm formation and/or dissection. A BAV can be a component of more complex congenital heart disease with or without other left heart obstructing lesions, as seen in Shone's complex.

OTHER FORMS OF OBSTRUCTION TO LEFT VENTRICULAR OUTFLOW

In addition to valvular AS, three other lesions may be responsible for obstruction to LV outflow: *hypertrophic obstructive cardiomyopathy* **(Chap. 254)**, *discrete fibromuscular/membranous subaortic stenosis*, and *supravalvular AS* **(Chap. 264)**. The causes of LV outflow obstruction can be differentiated on the basis of the cardiac examination and Doppler echocardiographic findings.

PATHOPHYSIOLOGY

The obstruction to LV outflow produces a systolic pressure gradient between the LV and aorta. When severe obstruction is suddenly produced experimentally, the LV responds by dilation and reduction of stroke volume. However, in some patients, the obstruction may be present at birth and/or increase gradually over the course of many years, and LV contractile performance is maintained by the presence of concentric LV hypertrophy. Initially, this serves as an adaptive mechanism because it reduces toward normal the systolic stress developed by the myocardium, as predicted by the Laplace relation ($S = Pr/h$, where S = systolic wall stress, P = pressure, r = radius, and h = wall thickness). A large transaortic valve pressure gradient may exist for many years without a reduction in cardiac output (CO) or the development of LV dilation. Ultimately, however, excessive hypertrophy becomes maladaptive, LV systolic function declines because of afterload mismatch, abnormalities of diastolic function progress, and irreversible myocardial fibrosis develops.

A mean systolic pressure gradient >40 mmHg with a normal CO or an effective aortic orifice area of ~<1 cm^2 (or ~<0.6 cm^2/m^2 body surface area in a normal-sized adult)—i.e., less than approximately one-third of the normal orifice area—is generally considered to represent severe obstruction to LV outflow. The elevated LV end-diastolic pressure observed in many patients with severe AS and preserved ejection fraction (EF) signifies the presence of diminished compliance of the hypertrophied LV. Although the CO at rest is within normal limits in most patients with severe AS, it usually fails to rise normally during exercise. Loss of an appropriately timed, vigorous atrial contraction, as occurs in atrial fibrillation (AF) or atrioventricular dissociation, may cause rapid progression of symptoms. Late in the course, contractile function deteriorates because of afterload excess, the CO and LV–aortic pressure gradient decline, and the mean left atrial (LA), pulmonary artery (PA), and right ventricular (RV) pressures rise. LV performance can be further compromised by superimposed epicardial coronary artery disease (CAD). Stroke volume (and thus CO) can also be reduced in patients with significant hypertrophy and a small LV cavity despite a normal EF. Low-flow, low-gradient AS (with either reduced or normal LV systolic function) is both a diagnostic and therapeutic challenge.

The hypertrophied LV causes an increase in myocardial oxygen requirements. In addition, even in the absence of obstructive CAD, coronary blood flow is impaired to the extent that ischemia can be precipitated under conditions of excess demand. Capillary density is reduced relative to wall thickness, compressive forces are increased, and the elevated LV end-diastolic pressure reduces the coronary driving pressure. The subendocardium is especially vulnerable to ischemia by this mechanism.

SYMPTOMS

AS is rarely of clinical importance until the valve orifice has narrowed to ~1 cm^2. Even severe AS may exist for many years without producing any symptoms because of the ability of the hypertrophied LV to generate the elevated intraventricular pressures required to maintain a normal stroke volume. Once symptoms occur, valve replacement is indicated.

Most patients with pure or predominant AS have gradually increasing obstruction over years but do not become symptomatic until the sixth to eighth decades. Adult patients with BAV disease, however, develop significant valve dysfunction and symptoms one to two decades sooner. Exertional dyspnea, angina pectoris, and syncope are the three cardinal symptoms. Often, there is a history of insidious progression of fatigue and dyspnea associated with gradual curtailment of activities and reduced effort tolerance. *Dyspnea* results primarily from elevation of the pulmonary capillary pressure caused by elevations of LV diastolic pressures secondary to impaired relaxation and reduced LV compliance. *Angina pectoris* usually develops somewhat later and reflects an imbalance between the augmented myocardial oxygen requirements and reduced oxygen availability. CAD may or may not be present, although its coexistence is common among AS patients age >65. *Exertional syncope* may result from a decline in arterial pressure caused by vasodilation in the exercising muscles and inadequate vasoconstriction in nonexercising muscles in the face of a fixed CO, or from a sudden fall in CO produced by an arrhythmia.

Because the CO at rest is usually well maintained until late in the course, marked fatigability, weakness, peripheral cyanosis, cachexia, and other clinical manifestations of a low CO are usually not prominent until this stage is reached. Orthopnea, paroxysmal nocturnal dyspnea, and pulmonary edema, i.e., symptoms of LV failure, also occur only in the advanced stages of the disease. Severe pulmonary hypertension leading to RV failure and systemic venous hypertension, hepatomegaly, AF, and tricuspid regurgitation (TR) are usually late findings in patients with isolated severe AS.

When AS and mitral stenosis (MS) coexist, the reduction in flow (CO) induced by MS lowers the pressure gradient across the aortic valve and, thereby, masks many of the clinical findings produced by AS. The transaortic pressure gradient can be increased in patients with concomitant AR due to higher aortic valve flow rates.

PHYSICAL FINDINGS

The rhythm is generally regular until late in the course; at other times, AF should suggest the possibility of associated mitral valve disease. Hypertension occurs commonly among older adults with AS. In the late stages, however, when stroke volume declines, the systolic pressure may fall and the pulse pressure narrow. The carotid arterial pulse rises slowly to a delayed peak (*pulsus parvus et tardus*). A thrill or anacrotic “shudder” may be palpable over the carotid arteries, more commonly the left. In the elderly, the stiffening of the arterial wall may mask this important physical sign. In many patients, the

a wave in the jugular venous pulse is accentuated. This results from the diminished distensibility of the RV cavity caused by the bulging, hypertrophied interventricular septum.

The LV impulse is sometimes displaced laterally in the later stages of the disease. A double apical impulse (with a palpable S_4) may be recognized, particularly with the patient in the left lateral recumbent position. A systolic thrill may be present at the base of the heart to the right of the sternum when leaning forward or in the suprasternal notch.

Auscultation

An early systolic ejection sound is frequently audible in children, adolescents, and young adults with congenital BAV disease. This sound usually disappears when the valve becomes calcified and rigid. As AS increases in severity, LV systole may become prolonged so that the aortic valve closure sound no longer precedes the pulmonic valve closure sound, and the two components may become synchronous, or aortic valve closure may even follow pulmonic valve closure, causing paradoxical splitting of S_2 **(Chap. 234)**. The sound of aortic valve closure can be heard most frequently in patients with AS who have pliable valves, and calcification diminishes the intensity of this sound. Frequently, an S_4 is audible at the apex and reflects the presence of LV hypertrophy and an elevated LV end-diastolic pressure; an S_3 generally occurs late in the course, when the LV dilates and its systolic function becomes severely compromised.

The murmur of AS is characteristically an ejection (mid) systolic murmur that commences shortly after the S_1, increases in intensity to reach a peak toward the middle of ejection, and ends just before aortic valve closure. It is characteristically low-pitched, rough and rasping in character, and loudest at the base of the heart, most commonly in the second right intercostal space. It is transmitted upward along the carotid arteries. Occasionally it is transmitted downward and to the apex, where it may be confused with the systolic murmur of mitral regurgitation (MR) (Gallavardin effect). In almost all patients with severe obstruction and preserved CO, the murmur is at least grade III/VI. In patients with mild degrees of obstruction or in those with severe stenosis with heart failure and low CO in whom the stroke volume and, therefore, the transvalvular flow rate are reduced, the murmur may be relatively soft and brief.

LABORATORY EXAMINATION

ECG

In most patients with severe AS, there is LV hypertrophy. In advanced cases, ST-segment depression and T-wave inversion (LV "strain") in standard leads I and aVL and in the left precordial leads are evident. However, there is no close correlation between the ECG and the hemodynamic severity of obstruction, and the absence of ECG signs of LV hypertrophy does not exclude severe obstruction. Systemic hypertension can coexist and also contribute to the development of hypertrophy.

Echocardiogram

The key findings on TTE are thickening, calcification, and reduced systolic opening of the valve leaflets and LV hypertrophy. Eccentric closure of the aortic valve cusps is characteristic of congenitally bicuspid valves. TEE imaging can display the obstructed orifice extremely well, but it is not routinely required for accurate characterization of AS. The valve gradient and aortic valve area can be estimated by Doppler measurement of the transaortic velocity. Severe AS is defined by a valve area <1 cm^2, whereas moderate AS is defined by a valve area of 1–1.5 cm^2 and mild AS by a valve area of 1.5–2 cm^2. Aortic valve sclerosis, conversely, is accompanied by a jet velocity of <2.5 m/s (peak gradient <25 mmHg). LV dilation and reduced systolic shortening reflect impairment of LV function. There is increasing experience with the use of longitudinal strain and strain rate to characterize earlier changes in LV systolic function, well before a decline in EF can be appreciated. Doppler indices of impaired diastolic function are frequently seen.

Echocardiography is useful for identifying coexisting valvular abnormalities, differentiating valvular AS from other forms of LV outflow obstruction, and measuring the aortic root and proximal ascending aortic dimensions. These aortic measurements are particularly important for patients with BAV disease. Dobutamine stress echocardiography is useful for the evaluation of patients with AS and severe LV systolic dysfunction (low-flow, low-gradient, severe AS with reduced EF), in whom the severity of the AS can often be difficult to judge. Patients with severe AS (i.e., valve area <1 cm^2) with a relatively low mean gradient (<40 mmHg) despite a normal EF (low-flow, low-gradient, severe AS with normal EF) are often hypertensive, and efforts to control their systemic blood pressure should be optimized before Doppler echocardiography is repeated. The use of dobutamine stress echocardiography in this setting is under investigation. When there is continued uncertainty regarding the severity of AS in patients with reduced CO, quantitative analysis of the amount of aortic valve calcium with chest computed tomography (CT) may be helpful. There is increasing use of chest CT to assess aortic valve morphology and function. It has become the imaging method of choice to plan for transcatheter aortic valve replacement (TAVR).

Chest X-Ray

The chest x-ray may show no or little overall cardiac enlargement for many years. Hypertrophy without dilation may produce some rounding of the cardiac apex in the frontal projection and slight backward displacement in the lateral view. A dilated proximal ascending aorta may be seen along the

upper right heart border in the frontal view. Aortic valve calcification may be discernible in the lateral view, but it is usually readily apparent on fluoroscopic examination or by echocardiography; the absence of valvular calcification on fluoroscopy in an adult suggests that severe valvular AS is *not* present. In later stages of the disease, as the LV dilates, there is increasing roentgenographic evidence of LV enlargement, pulmonary congestion, and enlargement of the LA, PA, and right-sided heart chambers.

Catheterization

Right- and left-sided heart catheterization for invasive assessment of AS is performed infrequently but can be useful when there is a discrepancy between the clinical and noninvasive findings. Concern has been raised that attempts to cross the aortic valve for measurement of LV pressures are associated with a risk of cerebral embolization. Catheterization is also useful in three distinct categories of patients: (1) *patients with multivalvular disease*, in whom the role played by each valvular deformity should be defined to aid in the planning of operative treatment; (2) *young, asymptomatic patients with noncalcific congenital AS*, to define the severity of obstruction to LV outflow, because operation or percutaneous aortic balloon valvuloplasty (PABV) may be indicated in these patients if severe AS is present, even in the absence of symptoms; and (3) *patients in whom it is suspected that the obstruction to LV outflow may not be at the level of the aortic valve* but rather at the sub- or supravalvular level.

Coronary angiography is indicated to screen for CAD in appropriate patients with severe AS who are being considered for surgical or transcatheter valve replacement. The incidence of significant CAD for which bypass grafting is indicated at the time of surgical aortic valve replacement (SAVR) exceeds 50% among adult patients.

NATURAL HISTORY

Death in patients with severe AS occurs most commonly in the seventh and eighth decades. Based on data obtained at postmortem examination in patients before surgical treatment became widely available, the average time to death after the onset of various symptoms was as follows: angina pectoris, 3 years; syncope, 3 years; dyspnea, 2 years; congestive heart failure, 1.5–2 years. Moreover, in >80% of patients who died with AS, symptoms had existed for <4 years. Among adults dying with valvular AS, sudden death, which presumably resulted from an arrhythmia, occurred in 10–20%; however, most sudden deaths occurred in patients who had previously been symptomatic. Sudden death as the first manifestation of severe AS is very uncommon (~1% per year) in asymptomatic adult patients. Calcific AS is a progressive disease, with an annual reduction in valve area averaging 0.1 cm^2 and annual increases in the peak jet velocity and mean valve gradient averaging 0.3 m/s and 7 mmHg, respectively **(Table 256-2)**.

TABLE 256-2

Mortality Rates after Aortic Valve Surgery[a]

Operation	Number	Unadjusted Operative Mortality (%)
AVR (isolated)	21,921	2.2
AVR + CAB	13,000	4.0
AVR + MVR	1349	8.0

[a]Data are for the first three quarters of calendar year 2015 during which 1033 sites reported a total of 210,421 procedures. Data (accessed March 9, 2017) are available from the Society of Thoracic Surgeons at *http://www.sts.org/sites/default/files/documents/2015Harvest4_ExecutiveSummary.pdf*.

Abbreviations: AVR, aortic valve replacement; CAB, coronary artery bypass; MVR, mitral valve replacement.

TREATMENT

TREATMENT

Aortic Stenosis

(Fig. 256-4)

MEDICAL TREATMENT

In patients with severe AS (valve area <1 cm^2), strenuous physical activity and competitive sports should be avoided, even in the asymptomatic stage. Care must be taken to avoid dehydration and hypovolemia to protect against a significant reduction in CO. Medications used for the treatment of hypertension or CAD, including beta blockers and angiotensin-converting enzyme (ACE) inhibitors, are generally safe for asymptomatic patients with preserved LV systolic function. Nitroglycerin is helpful in relieving angina pectoris in patients with CAD. Retrospective studies suggested that patients

with degenerative calcific AS who receive HMG-CoA reductase inhibitors ("statins") exhibit slower progression of leaflet calcification and aortic valve area reduction than those who do not. However, randomized prospective studies with either high-dose atorvastatin or combination simvastatin/ezetimibe have failed to show a measurable effect on valve-related outcomes. The use of statin medications should be driven by considerations regarding primary and secondary prevention of atherosclerotic cardiovascular disease (ASCVD) events. ACE inhibitors have not been studied prospectively for AS-related outcomes. The need for endocarditis prophylaxis is restricted to AS patients with a prior history of endocarditis.

SURGICAL TREATMENT

Asymptomatic patients with calcific AS and severe obstruction should be followed carefully for the development of symptoms and by serial echocardiograms for evidence of deteriorating LV function. Operation is indicated in patients with severe AS (valve area <1 cm^2 or 0.6 cm^2/m^2 body surface area) who are symptomatic, those who exhibit LV systolic dysfunction (EF <50%), and those with BAV disease and an aneurysmal root or ascending aorta (maximal dimension >5.5 cm). Operation for aneurysm disease is recommended at smaller aortic diameters (4.5–5.0 cm) for patients with a family history of an aortic catastrophe and for patients who exhibit rapid aneurysm growth (>0.5 cm/year). Patients with asymptomatic moderate or severe AS who are referred for coronary artery bypass grafting surgery should also have AVR. In patients without heart failure, the operative risk of SAVR (including patients with AS or AR) is ~2% (Table 256-2) but increases as a function of age and the need for concomitant aortic surgery or coronary revascularization with bypass grafting. The indications for SAVR in the asymptomatic patient have been the subject of intense debate, as surgical outcomes in selected patients have continued to improve. Relative indications for which surgery can be considered include an abnormal response to treadmill exercise; rapid progression of AS, especially when urgent access to medical care might be compromised; very severe AS, defined by an aortic valve jet velocity >5 m/s or mean gradient >60 mmHg and low operative risk; and excessive LV hypertrophy in the absence of systemic hypertension. Exercise testing can be safely performed in asymptomatic patients, as many as one-third of whom will show signs of functional impairment. A randomized controlled trial (RCT) of SAVR versus active surveillance in patients with asymptomatic severe AS is in the planning stages.

Operation should be carried out promptly (1–3 months) after symptom onset. In patients with low-flow, low-gradient severe AS with reduced LVEF, the perioperative mortality risk is high (15–20%), and evidence of myocardial disease may persist even when the operation is technically successful. Long-term postoperative survival correlates with preoperative LV function. Nonetheless, in view of the even worse prognosis of such patients when they are treated medically, there is usually little choice but to advise valve replacement, especially in patients in whom contractile reserve can be demonstrated by dobutamine stress echocardiography (defined by a ≥20% increase in stroke volume after dobutamine challenge). Patients in this high surgical risk group may benefit from TAVR, (see below), but data from RCTs in this population are lacking. The treatment of patients with low-flow, low-gradient severe AS with normal LVEF is also difficult. Outcomes appear to be better with surgery compared with conservative medical care for symptomatic patients with this type of "paradoxical" low-flow AS, but more research is needed to guide therapeutic decision making. In patients in whom severe AS and CAD coexist, relief of the AS and revascularization may sometimes result in striking clinical and hemodynamic improvement (Table 256-2).

Because many patients with calcific AS are elderly, particular attention must be directed to the adequacy of hepatic, renal, and pulmonary function before AVR is recommended. Age alone is not a contraindication to SAVR for AS. The perioperative mortality rate depends to a substantial extent on the patient's preoperative clinical and hemodynamic state. Assessment of frailty is a critical component of preprocedural evaluation. Treatment decisions for AS patients who are not at low operative risk should be made by a multidisciplinary heart team with representation from general cardiology, interventional cardiology, imaging, cardiac surgery, and other allied specialties as needed, including geriatrics. The 10-year survival rate of older adult patients with AVR is ~60%. Recommendations regarding the type of valve prosthesis (tissue or mechanical) must weigh the trade-offs between durability and thromboembolism/bleeding and are heavily influenced by patient age and preferences. Bioprostheses are favored for patients age >65 years. Shared decision making with younger patients must be individualized. Approximately 30% of bioprosthetic valves evidence primary valve failure in 10 years, requiring re-replacement, and an approximately equal percentage of patients with mechanical prostheses develop hemorrhagic complications as a consequence of treatment with vitamin K antagonists. Homograft AVR is usually reserved for patients with aortic valve endocarditis.

The Ross procedure involves replacement of the diseased aortic valve with the autologous pulmonic valve and implantation of a homograft in the native pulmonic position. Its use has declined considerably in the United States because of the technical complexity of the procedure and the incidence of late postoperative aortic root dilation and autograft failure with AR. There is also a low incidence of pulmonary homograft stenosis.

PERCUTANEOUS AORTIC BALLOON VALVULOPLASTY (PABV)

This procedure is preferable to operation in many children and young adults with congenital, noncalcific AS **(Chap. 264)**. It is not commonly used as definitive therapy in adults with severe calcific AS because of a very high restenosis rate (80% within 1 year) and the risk of procedural complications, but on occasion, it has been used successfully as a "bridge to operation" in patients with severe LV dysfunction and shock who are too ill to tolerate surgery. It is performed routinely as part of the TAVR procedure (see below).

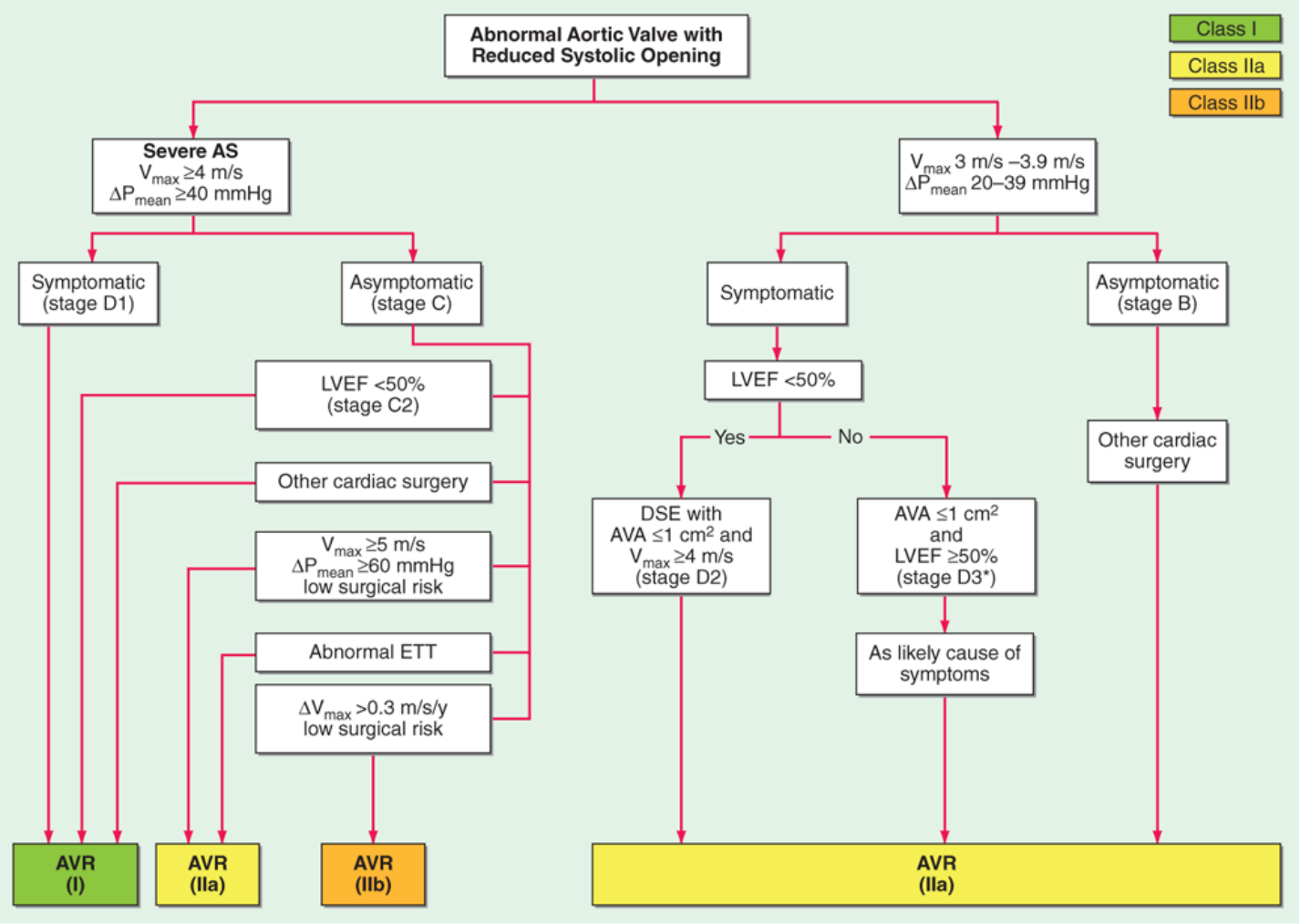
TRANSCATHETER AORTIC VALVE REPLACEMENT

TAVR for treatment of AS has been performed with increasing frequency in prohibitive-, high-, and intermediate surgical-risk adult patients worldwide using one of two available systems, a balloon-expandable valve and a self-expanding valve, both of which incorporate a pericardial prosthesis **(Fig. 256-5)**. Newer valve platforms are under investigation. Nearly 25,000 U.S. patients underwent TAVR in 2015 across more than 415 centers. TAVR is most frequently performed via the transfemoral route, although trans-LV apical, subclavian, carotid, and ascending aortic routes have been used. Aortic balloon valvuloplasty under rapid RV pacing is performed as a first step to create an orifice of sufficient size for the prosthesis. Procedural success rates

exceed 90%. Among elderly patients with severe AS who are considered inoperable (i.e., prohibitive surgical risk), 1- and 2-year survival rates are significantly higher with TAVR compared with medical therapy (including PABV) **(Fig. 256-6)**. One- and 2-year survival rates are essentially equal for high-surgical-risk patients treated with TAVR or SAVR **(Fig. 256-7)**. Last, the 2-year rate of death or disabling stroke is lower in intermediate surgical risk patients who undergo transfemoral TAVR compared with SAVR **(Fig. 256-8)**. TAVR is associated with an early hazard for stroke and a higher incidence of postprocedural, paravalvular AR, a risk factor for mortality over the next 2 years. The rate of postprocedural heart block requiring permanent pacemaker (~10%) is significantly more common after TAVR. Valve performance characteristics are excellent over 5 years; longer-term durability assessment is ongoing, but results at 5 years are acceptable. Overall outcomes with this transformative technology have been very favorable and have allowed the extension of AVR to groups of patients previously considered at high or prohibitive risk for conventional surgery. Nevertheless, some patients are not candidates for this procedure because their comorbidity profile and frailty would make its undertaking inappropriate. The heart team is specifically charged with making challenging decisions of this nature. The use of these devices for the treatment of patients with structural deterioration of bioprosthetic aortic valves ("valve-in-valve"), as an alternative to reoperative valve replacement, has been approved. Presently, patients with BAV are not candidates for TAVR.

FIGURE 256-4

Management strategy for patients with aortic stenosis. Preoperative coronary angiography should be performed routinely as determined by age, symptoms, and coronary risk factors. Cardiac catheterization and angiography may also be helpful when there is a discrepancy between clinical and noninvasive findings. Patients who do not meet criteria for intervention should be monitored periodically with clinical and echocardiographic follow-up. The class designations refer to the American Heart Association/American College of Cardiology methodology for treatment recommendations. Class I recommendations should be performed or are indicated; Class IIa recommendations are considered reasonable to perform; Class IIb recommendations may be considered. The stages refer to the stages of progression of the disease. At disease stage A, risk factors are present for the development of valve dysfunction; stage B refers to progressive, mild-moderate, asymptomatic valve disease; stage C disease is severe in nature but clinically asymptomatic; stage C1 characterizes asymptomatic patients with severe valve disease but compensated ventricular function; stage C2 refers to asymptomatic, severe disease with ventricular decompensation; stage D refers to severe, symptomatic valve disease. With aortic stenosis, stage D1 refers to symptomatic patients with severe aortic stenosis and a high valve gradient (>40 mmHg mean gradient); stage D2 comprises patients with symptomatic, severe, low-flow, low-gradient aortic stenosis and low left ventricular ejection fraction; and stage D3 characterizes patients with symptomatic, severe, low-flow, low-gradient aortic stenosis and preserved left ventricular ejection fraction (paradoxical, low-flow, low-gradient severe aortic stenosis). AS, aortic stenosis; AVA; aortic valve area; AVR, aortic valve replacement; DSE, dobutamine stress echocardiography; ETT, exercise treadmill test; LVEF, left ventricular ejection fraction; ΔP_{mean}, mean pressure gradient; V_{max}, maximum velocity. *(Adapted from RA Nishimura et al: J Am Coll Cardiol 70:252, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

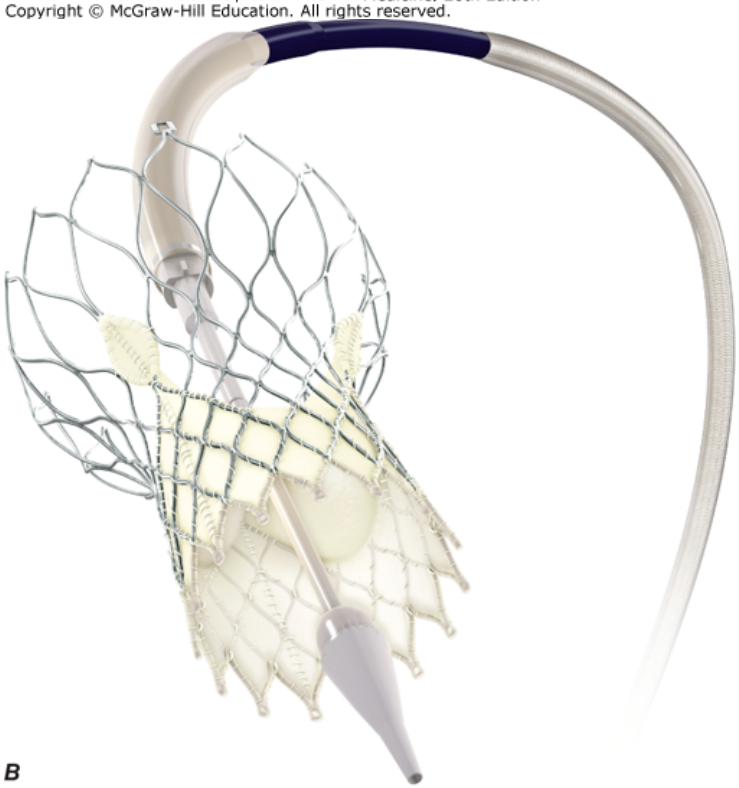
FIGURE 256-5

Balloon-expandable (***A***) and self-expanding (***B***) valves for transcatheter aortic valve replacement (TAVR). B, inflated balloon; N, nose cone; V, valve. *(Part A, courtesy of Edwards Lifesciences, Irvine, CA; with permission. NovaFlex+ is a trademark of Edwards Lifesciences Corporation. Part B, © Medtronic, Inc. 2015. Medtronic CoreValve Transcatheter Aortic Valve. CoreValve is a registered trademark of Medtronic, Inc.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

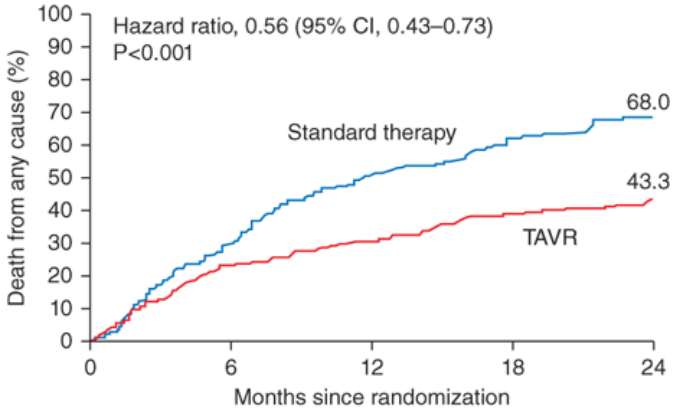


B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 256-6

Twenty-four-month outcomes following transcatheter aortic valve replacement (TAVR) for inoperable patients in the PARTNER I trial (cohort B). CI, confidence interval. *(Adapted from RR Makkar et al: N Engl J Med 366:1696, 2012; with permission.)*

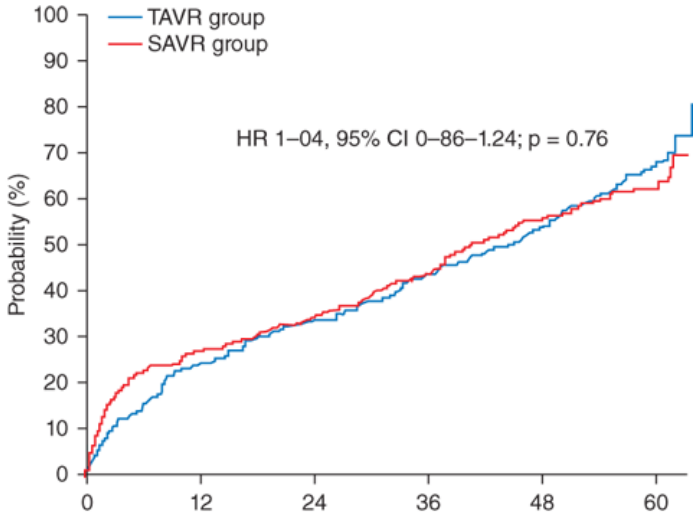


No. at Risk					
TAVR	179	138	124	110	83
Standard therapy	179	121	85	62	42

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 256-7

Five-year mortality rates following transcatheter (TAVR) or surgical aortic valve replacement (SAVR) for high-surgical-risk patients (cohort A) in the PARTNER I trial. Mortality rates in this AS patient cohort are similar for TAVR and SAVR. CI, confidence interval. *(Adapted from MJ Mack et al: Lancet 2015; 385:2477–2484.)*

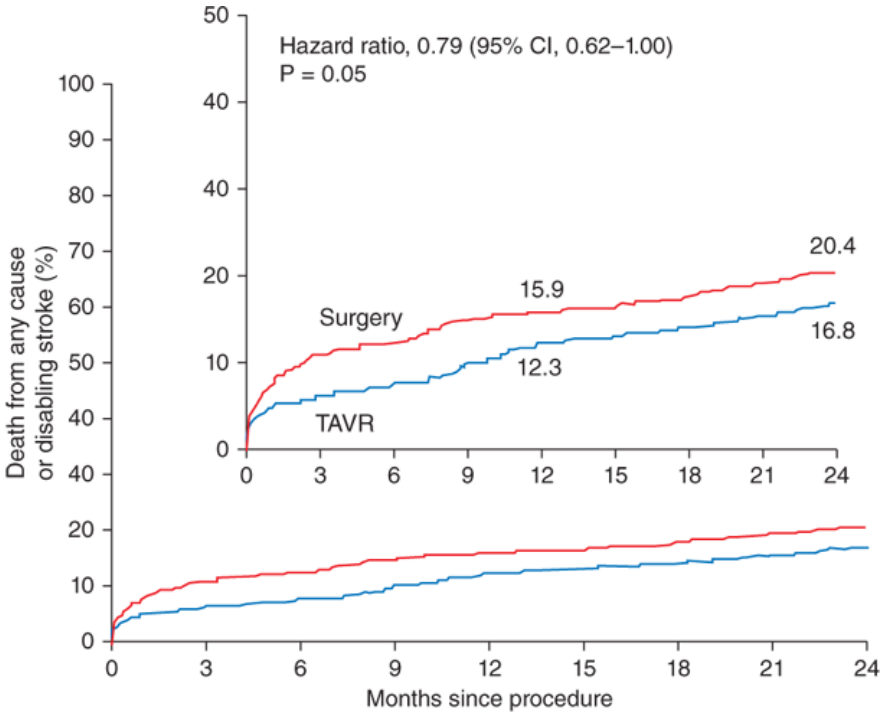


No. at Risk						
TAVR group	348	262	228	191	154	61
SAVR group	351	236	210	174	131	62

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 256-8

Rates of death from any cause or disabling stroke for intermediate surgical risk as patients undergoing surgical aortic valve replacement (SAVR) or transfemoral transcatheter aortic valve replacement (TAVR). In this intention-to-treat analysis, transfemoral TAVR proved superior to SAVR. *(Adapted from MB Leon et al: N Engl J Med 374:1609, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Carapetis JR et al: Acute rheumatic fever and rheumatic heart disease. Nat Rev Dis Primers 2:15084, 2016. [PubMed: 27188830]

Lindman B et al: Calcific aortic stenosis. Nat Rev Dis Primers 2:16006, 2016. [PubMed: 27188578]

Nishimura RA et al: 2014 AHA/ACC Guideline for management of patients with valvular heart disease: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 63:e57, 2014. [PubMed: 24603191]

Nishimura RA et al: 2017 AHA/ACC Focused Update of the 2014 AHA/ACC Guideline for the Management of Patients with Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 70:252, 2017.

Chapter 257: Aortic Regurgitation

Patrick T. O'Gara; Joseph Loscalzo

ETIOLOGY

Aortic regurgitation (AR) may be caused by primary valve disease, aortic root disease or their combination (Table 257-1).

TABLE 257-1

Major Causes of Aortic Valve Disease

Valve Lesion	Etiologies
Aortic regurgitation	**Valvular**
	Congenital (bicuspid)
	Endocarditis
	Rheumatic fever
	Myxomatous (prolapse)
	Traumatic
	Syphilis
	Ankylosing spondylitis
	Root disease
	Aortic dissection
	Cystic medial degeneration
	Marfan syndrome
	Bicuspid aortic valve
	Nonsyndromic familial aneurysm
	Aortitis
	Hypertension

Primary Valve Disease

Rheumatic disease results in thickening, deformity, and shortening of the individual aortic valve cusps, changes that prevent their proper opening during systole and closure during diastole. A rheumatic origin is much less common in patients with isolated AR who do not have associated rheumatic mitral valve disease. Patients with congenital bicuspid aortic valve (BAV) disease may develop predominant AR, and ~20% of these patients will require aortic valve surgery between 10 and 40 years of age. Congenital fenestrations of the aortic valve occasionally produce mild AR. Membranous subaortic

stenosis results in a high velocity systolic jet that often leads to thickening and scarring of the aortic valve leaflets and secondary AR. Prolapse of an aortic cusp, resulting in progressive chronic AR, occurs in ~15% of patients with ventricular septal defect **(Chap. 264)**, but may also occur as an isolated phenomenon or as a consequence of myxomatous degeneration sometimes associated with mitral and/or tricuspid valve involvement.

AR may result from infective endocarditis (IE), which can develop on a valve previously affected by rheumatic disease, a congenitally deformed valve, or on a normal aortic valve, and may lead to perforation or erosion of one or more leaflets. The aortic valve leaflets may become scarred and retracted during the course of syphilis or ankylosing spondylitis and contribute further to the AR that derives primarily from the associated root disease. Although traumatic rupture or avulsion of an aortic cusp is an uncommon cause of acute AR, it represents the most frequent serious lesion in patients surviving nonpenetrating cardiac injuries. The coexistence of hemodynamically significant AS with AR usually excludes all the rarer forms of AR because it occurs almost exclusively in patients with rheumatic or congenital AR. In patients with AR due to primary valvular disease, dilation of the aortic annulus may occur secondarily and lead to worsening regurgitation.

Primary Aortic Root Disease

AR also may be due entirely to marked aortic annular dilation, i.e., aortic root disease, without primary involvement of the valve leaflets; widening of the aortic annulus and lack of diastolic coaptation of the aortic leaflets are responsible for the AR **(Chap. 274)**. Medial degeneration of the ascending aorta, which may or may not be associated with other manifestations of Marfan syndrome; idiopathic dilation of the aorta; annuloaortic ectasia; osteogenesis imperfecta; and severe, chronic hypertension may all widen the aortic annulus and lead to progressive AR. Occasionally AR is caused by retrograde dissection of the aorta involving the aortic annulus. Syphilis and ankylosing spondylitis, both of which may also affect the aortic leaflets, may be associated with cellular infiltration and scarring of the media of the thoracic aorta, leading to aortic dilation, aneurysm formation, and severe regurgitation. In syphilis of the aorta **(Chap. 177)**, now a very rare condition, the involvement of the intima may narrow the coronary ostia, which in turn may be responsible for myocardial ischemia. Takayasu's aortitis and giant cell aortitis can also result in aneurysm formation and secondary AR.

PATHOPHYSIOLOGY

The total stroke volume ejected by the left ventricle (LV) (i.e., the sum of the effective forward stroke volume and the volume of blood that regurgitates back into the LV) is increased in patients with AR. In patients with severe AR, the volume of regurgitant flow may equal the effective forward stroke volume. In contrast to MR, in which a portion of the LV stroke volume is delivered into the low-pressure left atrium (LA), in AR the entire LV stroke volume is ejected into a high-pressure zone, the aorta. An increase in the LV end-diastolic volume (increased preload) constitutes the major hemodynamic compensation for AR. The dilation and eccentric hypertrophy of the LV allow this chamber to eject a larger stroke volume without requiring any increase in the relative shortening of each myofibril. Therefore, severe AR may occur with a normal effective forward stroke volume and a normal LV ejection fraction (LVEF, total [forward plus regurgitant] stroke volume/end-diastolic volume), together with an elevated LV end-diastolic pressure and volume. However, through the operation of Laplace's law, LV dilation increases the LV systolic tension required to develop any given level of systolic pressure. Chronic AR is, thus, a state in which LV preload and afterload are both increased. Ultimately, these adaptive measures fail. As LV function deteriorates, the end-diastolic volume rises further and the forward stroke volume and EF decline. Deterioration of LV function often precedes the development of symptoms. Considerable thickening of the LV wall also occurs with chronic AR, and at autopsy, the hearts of these patients may be among the largest encountered, sometimes weighing >1000 g.

The reverse pressure gradient from aorta to LV, which drives the AR flow, falls progressively during diastole, accounting for the decrescendo nature of the diastolic murmur. Equilibration between aortic and LV pressures may occur toward the end of diastole in patients with chronic severe AR, particularly when the heart rate is slow. In patients with acute severe AR, the LV is unprepared for the regurgitant volume load. LV compliance is normal or reduced, and LV diastolic pressures rise rapidly, occasionally to levels >40 mmHg. The LV pressure may exceed the LA pressure toward the end of diastole, and this reversed pressure gradient closes the mitral valve prematurely.

In patients with chronic severe AR, the effective forward cardiac output (CO) usually is normal or only slightly reduced at rest, but often it fails to rise normally during exercise. An early sign of LV dysfunction is a reduction in the EF. In advanced stages, there may be considerable elevation of the LA, pulmonary artery (PA) wedge, PA, and right ventricular (RV) pressures and lowering of the forward CO at rest.

Myocardial ischemia may occur in patients with AR because myocardial oxygen requirements are elevated by LV dilation, hypertrophy, and elevated LV systolic tension, and coronary blood flow may be compromised. A large fraction of coronary blood flow occurs during diastole, when arterial pressure is low, thereby reducing coronary perfusion or driving pressure. This combination of increased oxygen demand and reduced supply may cause myocardial ischemia, particularly of the subendocardium, even in the absence of epicardial coronary artery disease (CAD).

HISTORY

Approximately three-fourths of patients with pure or predominant valvular AR are men; women predominate among patients with primary valvular AR who have associated rheumatic mitral valve disease. A history compatible with IE may sometimes be elicited from patients with rheumatic or congenital involvement of the aortic valve, and the infection often precipitates or seriously aggravates preexisting symptoms.

In patients with *acute severe AR*, as may occur in IE, aortic dissection, or trauma, the LV cannot dilate sufficiently to maintain stroke volume, and LV diastolic pressure rises rapidly with associated marked elevations of LA and PA wedge pressures. Pulmonary edema and/or cardiogenic shock may develop rapidly.

Chronic severe AR may have a long latent period, and patients may remain relatively asymptomatic for as long as 10–15 years. Uncomfortable awareness of the heartbeat, especially on lying down, may be an early complaint. Sinus tachycardia, during exertion or with emotion, or premature ventricular contractions may produce particularly uncomfortable palpitations as well as head pounding. These complaints may persist for many years before the development of exertional dyspnea, usually the first symptom of diminished cardiac reserve. The dyspnea is followed by orthopnea, paroxysmal nocturnal dyspnea, and excessive diaphoresis. Anginal chest pain even in the absence of CAD may occur in patients with severe AR, even in younger patients. Anginal pain may develop at rest as well as during exertion. Nocturnal angina may be a particularly troublesome symptom, and it may be accompanied by marked diaphoresis. The anginal episodes can be prolonged and often do not respond satisfactorily to sublingual nitroglycerin. Systemic fluid accumulation, including congestive hepatomegaly and ankle edema, may develop late in the course of the disease.

PHYSICAL FINDINGS

In chronic severe AR, the jarring of the entire body and the bobbing motion of the head with each systole can be appreciated, and the abrupt distention and collapse of the larger arteries are easily visible. The examination should be directed toward the detection of conditions predisposing to AR, such as bicuspid valve, IE, Marfan syndrome, or ankylosing spondylitis.

Arterial Pulse

A rapidly rising “water-hammer” pulse, which collapses suddenly as arterial pressure falls rapidly during late systole and diastole (Corrigan’s pulse), and capillary pulsations, an alternate flushing and paling of the skin at the root of the nail while pressure is applied to the tip of the nail (Quincke’s pulse), are characteristic of chronic severe AR. A booming “pistol-shot” sound can be heard over the femoral arteries (Traube’s sign), and a to-and-fro murmur (Duroziez’s sign) is audible if the femoral artery is lightly compressed with a stethoscope.

The arterial pulse pressure is widened as a result of both systolic hypertension and a lowering of the diastolic pressure. The measurement of arterial diastolic pressure with a sphygmomanometer may be complicated by the fact that systolic sounds are frequently heard with the cuff completely deflated. However, the level of cuff pressure at the time of muffling of the Korotkoff sounds (phase IV) generally corresponds fairly closely to the true intraarterial diastolic pressure. As the disease progresses and the LV end-diastolic pressure rises, the arterial diastolic pressure may actually rise as well, because the aortic diastolic pressure cannot fall below the LV end-diastolic pressure. For the same reason, acute severe AR may also be accompanied by only a slight widening of the pulse pressure. Such patients are invariably tachycardic as the heart rate increases in an attempt to preserve the CO.

Palpation

In patients with chronic severe AR, the LV impulse is heaving and displaced laterally and inferiorly. The systolic expansion and diastolic retraction of the apex are prominent. A diastolic thrill may be palpable along the left sternal border in thin-chested individuals, and a prominent systolic thrill may be palpable in the suprasternal notch and transmitted upward along the carotid arteries. This systolic thrill and the accompanying murmur do not necessarily signify the coexistence of aortic stenosis (AS). In some patients with AR or with combined AS and AR, the carotid arterial pulse may be bisferiens, i.e., with two systolic waves separated by a trough **(see Fig. 234-2D)**.

Auscultation

In patients with severe AR, the aortic valve closure sound (A_2) is usually absent. A systolic ejection sound is audible in patients with BAV disease, and occasionally an S_4 also may be heard. The murmur of chronic AR is typically a high-pitched, blowing, decrescendo diastolic murmur, heard best in the third intercostal space along the left sternal border **(see Fig. 234-5B)**. In patients with mild AR, this murmur is brief, but as the severity increases, it generally becomes louder and longer, indeed holodiastolic. When the murmur is soft, it can be heard best with the diaphragm of the stethoscope and with the patient sitting up, leaning forward, and with the breath held in forced expiration. In patients in whom the AR is caused by primary valvular disease, the diastolic murmur is usually louder along the left than the right sternal border. However, when the murmur is louder along the right sternal border, it suggests that the AR is caused by aneurysmal dilation of the aortic root. “Cooing” or musical diastolic murmurs suggest eversion of an aortic cusp vibrating in the regurgitant stream.

A mid-systolic ejection murmur is frequently audible in isolated AR. It is generally heard best at the base of the heart and is transmitted along the carotid arteries. This murmur may be quite loud without signifying aortic obstruction. A third murmur sometimes heard in patients with severe AR is the *Austin Flint murmur*, a soft, low-pitched, rumbling mid-to-late diastolic murmur. It is probably produced by the diastolic displacement of the anterior leaflet of the mitral valve by the AR stream and is not associated with hemodynamically significant mitral obstruction. The auscultatory features of AR are intensified by strenuous and sustained handgrip, which augments systemic vascular resistance.

In acute severe AR, the elevation of LV end-diastolic pressure may lead to early closure of the mitral valve, a soft S_1, a pulse pressure that is not particularly wide, and a soft, short, early diastolic murmur of AR.

LABORATORY EXAMINATION

ECG

In patients with chronic severe AR, the ECG signs of LV hypertrophy are common **(Chap. 235)**. In addition, these patients frequently exhibit ST-segment depression and T-wave inversion in leads I, aVL, V_5, and V_6 ("LV strain"). Left-axis deviation and/or QRS prolongation denote diffuse myocardial disease, generally associated with patchy fibrosis, and usually signify a poor prognosis.

Echocardiogram

LV size is increased in chronic AR and systolic function is normal or even supernormal until myocardial contractility declines, as signaled by a decrease in ejection fraction (EF) or increase in the end-systolic dimension. A rapid, high-frequency diastolic fluttering of the anterior mitral leaflet produced by the impact of the regurgitant jet is a characteristic finding. The echocardiogram is also useful in determining the cause of AR, by detecting dilation of the aortic annulus and root, aortic dissection **(see Fig. 236-5**, or primary leaflet pathology). With severe AR, the central jet width assessed by color flow Doppler imaging exceeds 65% of the LV outflow tract, the regurgitant volume is ≥60 mL/beat, the regurgitant fraction is ≥50%, and there is diastolic flow reversal in the proximal descending thoracic aorta. The continuous-wave Doppler profile of the AR jet shows a rapid deceleration time in patients with acute severe AR, due to the rapid increase in LV diastolic pressure. Surveillance transthoracic echocardiography (TTE) forms the cornerstone of longitudinal follow-up and allows for the early detection of changes in LV size and/or function. For patients in whom TTE is limited by poor acoustical windows or inadequate characterization of LV function or the severity of the regurgitation, cardiac magnetic resonance (CMR) imaging can be performed. This modality also allows for accurate assessment of aortic size and contour. Transesophageal echocardiography (TEE) can also provide detailed anatomic assessment of the valve, root, and portions of the aorta. There is increasing experience with the use of 3D echocardiography to measure LV volumes.

Chest X-Ray

In chronic severe AR, the apex is displaced downward and to the left in the frontal projection. In the left anterior oblique and lateral projections, the LV is displaced posteriorly and encroaches on the spine. When AR is caused by primary disease of the aortic root, aneurysmal dilation of the aorta may be noted, and the aorta may fill the retrosternal space in the lateral view. Echocardiography, cardiac MRI, and chest CT angiography are more sensitive than the chest x-ray for the detection of root and ascending aortic enlargement.

Cardiac Catheterization and Angiography

When needed, right and left heart catheterization with contrast aortography can provide confirmation of the magnitude of regurgitation and the status of LV function. Coronary angiography is performed routinely in appropriate patients prior to surgery.

TREATMENT

TREATMENT

Aortic Regurgitation

ACUTE AORTIC REGURGITATION

Patients with acute severe AR may respond to intravenous diuretics and vasodilators (such as sodium nitroprusside), but stabilization is usually short-lived and operation is indicated urgently **(Fig. 257-1)**. Intraaortic balloon counterpulsation is contraindicated. Beta blockers are also best avoided so as not to reduce the CO further or slow the heart rate, thus allowing more time for diastolic filling of the LV. Surgery is the treatment of choice and is usually necessary within 24 h of diagnosis.

CHRONIC AORTIC REGURGITATION

Early symptoms of dyspnea and effort intolerance respond to treatment with diuretics; vasodilators (Angiotensin converting enzyme [ACE] inhibitors, dihydropyridine calcium channel blockers, or hydralazine) may be useful as well. Surgery can then be performed in a more controlled setting. The use of vasodilators to extend the compensated phase of chronic severe AR before the onset of symptoms or the development of LV dysfunction is more controversial and less well established. Systolic blood pressure should be controlled (goal <140 mmHg) in patients with chronic AR, and vasodilators are an excellent first choice as antihypertensive agents. It is often difficult to achieve adequate control because of the increased stroke volume that accompanies severe AR. Cardiac arrhythmias and systemic infections are poorly tolerated in patients with severe AR and must be treated promptly and vigorously. Although nitroglycerin and long-acting nitrates are not as helpful in relieving anginal pain as they are in patients with ischemic heart

disease, they are worth a trial. Patients with syphilitic aortitis should receive a full course of penicillin therapy **(Chap. 177)**. Beta blockers and the angiotensin receptor blocker (ARB) losartan may be useful to retard the rate of aortic root enlargement in young patients with Marfan syndrome and aortic root dilation. A randomized controlled trial showed no difference in efficacy between atenolol and losartan for this indication. Whether beta blockers or ARBs is useful in retarding the rate of growth of aortic aneurysms in other patient subsets (e.g., BAV disease with aortopathy, Takaysu's disease) have not been demonstrated. Beta blockers in patients with valvular AR were previously considered relatively contraindicated due to concerns that the resulting slowing of the heart rate would allow more time for diastolic regurgitation. Observational reports, however, have suggested that beta blockers may provide functional benefit in some patients with chronic AR. Beta blockers can sometimes provide incremental blood pressure lowering in patients with chronic AR and hypertension. They can also lessen the sense of forceful heart action that many patients find uncomfortable. Patients with severe AR, particularly those with an associated aortopathy, should avoid isometric exercises.

SURGICAL TREATMENT

In deciding on the advisability and proper timing of surgical treatment, two points should be kept in mind: (1) patients with chronic severe AR usually do not become symptomatic until *after* the development of myocardial dysfunction; and (2) when delayed too long (defined as >1 year from onset of symptoms or LV dysfunction), surgical treatment often does not restore normal LV size and function. Therefore, in patients with chronic severe AR, careful clinical follow-up and noninvasive testing with echocardiography at ~6- to 12-month intervals are necessary if operation is to be undertaken at the optimal time, i.e., *after* the onset of LV dysfunction but *prior to* the development of severe symptoms. Exercise testing may be helpful to assess effort tolerance more objectively. Operation can be deferred as long as the patient both remains asymptomatic and retains normal LV function without severe or progressive chamber dilation.

Aortic valve replacement (AVR) is indicated for the treatment of severe AR in symptomatic patients irrespective of LV function. In general, the operation should be carried out in asymptomatic patients with severe AR and progressive LV dysfunction defined by an LVEF <50%, an LV end-systolic dimension >50 mm, or an LV diastolic dimension >65 mm. Smaller dimensions may be appropriate thresholds in individuals of smaller stature or when there is evidence of progressively decreasing LV function or increasing LV size on serial studies. Patients with severe AR without indications for operation should be followed by clinical and echocardiographic examination every 6–12 months.

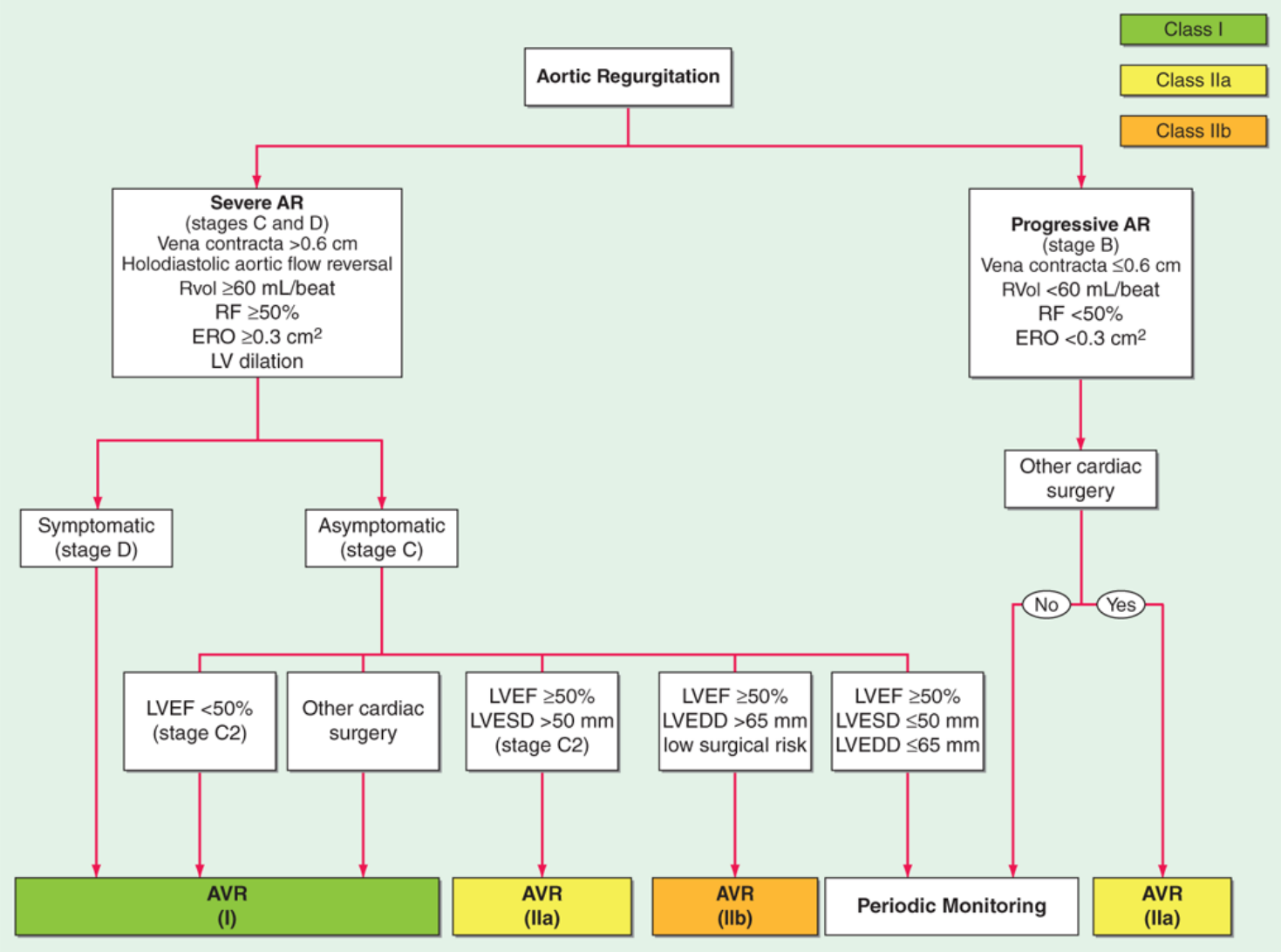
Surgical options for management of aortic valve and root disease have expanded considerably over the past decade. AVR with a suitable mechanical or tissue prosthesis is generally necessary in patients with rheumatic AR and in many patients with other forms of regurgitation. Rarely, when a leaflet has been perforated during IE or torn from its attachments to the aortic annulus by thoracic trauma, primary surgical repair may be possible. When AR is due to aneurysmal dilation of the root or proximal ascending aorta rather than to primary valve involvement, it may be possible to reduce or eliminate the regurgitation by narrowing the annulus or by excising a portion of the aortic root without replacing the valve. Elective, valve-sparing aortic root reconstruction generally involves reimplantation of the valve in a contoured graft with reattachment of the coronary artery buttons into the side of the graft and is best undertaken in specialized surgical centers **(Fig. 257-2)**. Resuspension of the native aortic valve leaflets is possible in ~50% of patients with acute AR in the setting of type A aortic dissection. In other conditions, however, regurgitation can be effectively eliminated only by replacing the aortic valve, the dilated or aneurysmal ascending aorta responsible for the regurgitation, and implanting a composite valve-graft conduit. This formidable procedure entails a higher risk than isolated AVR.

As in patients with other valvular abnormalities, both the operative risk and the late mortality rate are largely dependent on the stage of the disease and myocardial function at the time of operation. The overall operative mortality rate for isolated AVR (performed for either or both AS or AR) is ~2% **(Table 257-2)**. However, patients with AR, marked cardiac enlargement, and prolonged LV dysfunction experience an operative mortality rate of ~10% and a late mortality rate of ~5% per year due to LV failure despite a technically satisfactory operation. Nonetheless, because of the very poor prognosis with medical management, even patients with advanced LV systolic dysfunction should be considered for operation.

Patients with acute severe AR require prompt (24–48 h) surgical treatment, which may be lifesaving.

FIGURE 257-1

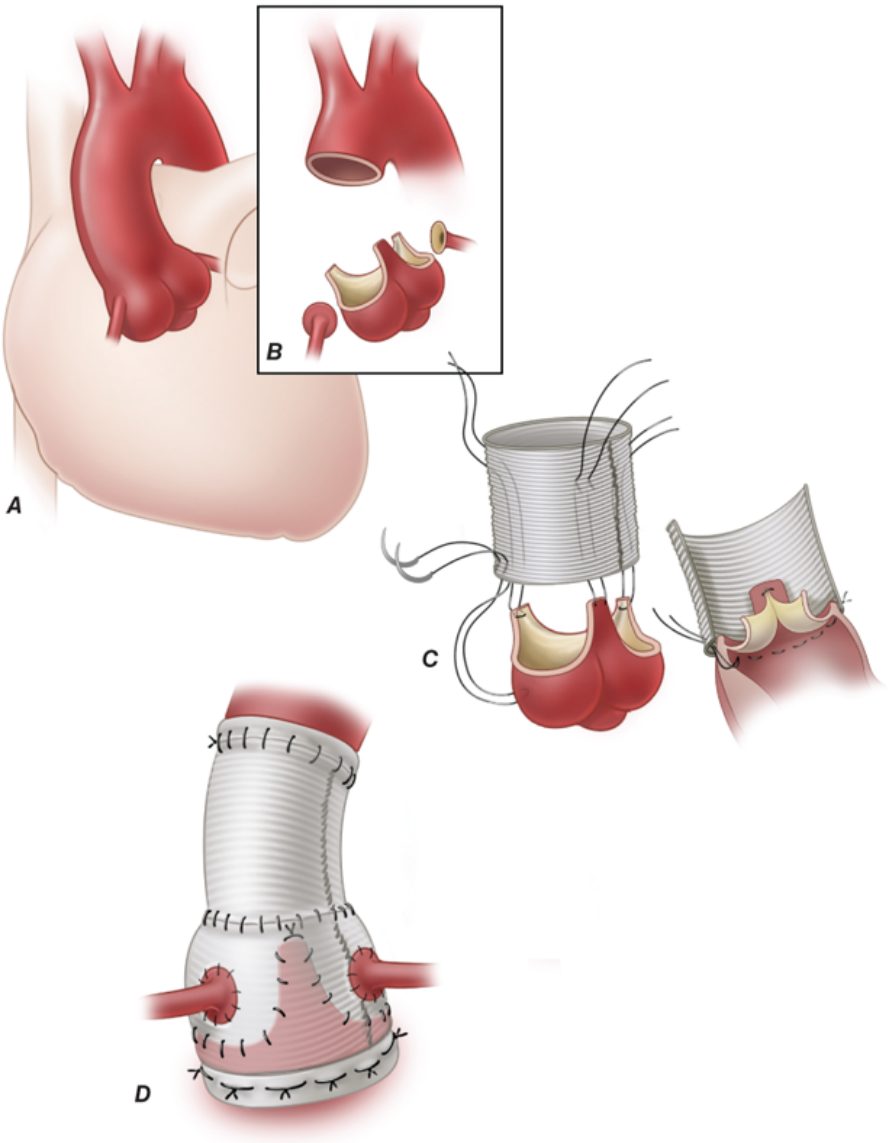
Management of patients with aortic regurgitation. See legend for **Fig. 256-4** for explanation of treatment recommendations (Class I, IIa, and IIb) and disease stages (B, C1, C2, and D). Preoperative coronary angiography should be performed routinely as determined by age, symptoms, and coronary risk factors. Cardiac catheterization and angiography may also be helpful when there is a discrepancy between clinical and noninvasive findings. Patients who do not meet criteria for intervention should be monitored periodically with clinical and echocardiographic follow-up. AR, aortic regurgitation; AVR, aortic valve replacement (valve repair may be appropriate in selected patients); ERO, effective regurgitant orifice; LV, left ventricular; LVEDD, left ventricular end-diastolic dimension; LVEF, left ventricular ejection fraction; LVESD, left ventricular end-systolic dimension; RF, regurgitant fraction; RVol, regurgitant volume. *(Adapted from RA Nishimura et al: 2014 AHA/ACC Guideline for the Management of Patients with Valvular Heart Disease. J Am Coll Cardiol 63:e57-185, 2014, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 257-2

Valve-sparing aortic root reconstruction (David procedure). Aortic root and proximal ascending aorta (***A***) are resected (***B***) with sinuses of Valsalva and mobilized coronary artery buttons remaining. Subannular sutures (***C***) are placed, commissural posts are drawn up inside the valve and the annular sutures are passed through the proximal end of the graft. The annular sutures are tied (***D***), the valve is re-implanted inside the graft, aortic continuity is re-established with another graft of appropriate size and the coronary buttons are attached to the side of the graft. *(From P Steltzer et al [eds]: Valvular Heart Disease: A Companion to Braunwald's Heart Disease, 3rd ed, Fig 12-27, p. 200.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 257-2

Mortality Rates After Aortic Valve Surgery[a]

Operation	Number	Unadjusted Operative Mortality (%)
AVR (isolated)	14,795	2.3
AVR + CAB	9158	4.2
AVR + MVR	876	8.8

Abbreviations: AVR, aortic valve replacement; CAB, coronary artery bypass; MVR, mitral valve replacement.

[a]Data are for the first two quarters of calendar year 2013, during which 1004 sites reported a total of 135,666 procedures. Data are available from the Society of Thoracic Surgeons at *http://www.sts.org/sites/default/files/documents/2013_3rdHarvestExecutiveSummary.pdf*.

FURTHER READING

Lacro RV et al: Atenolol versus losartan in children and young adults with Marfan's syndrome. N Engl J Med 371:2061, 2014. [PubMed: 25405392]

Malaisrie SC, McCarthy PM: Surgical approach to disease of the aortic valve and the aortic root, in *Valvular Heart Disease: A Companion to Braunwald's Heart Disease*, 4th ed. CM Otto, RO Bonow (eds). Philadelphia, Elsevier Saunders, 2014, pp 199–218.

Nishimura RA et al: 2014 AHA/ACC guidelines for management of patients with valvular heart disease. J Am Coll Cardiol 63:e57, 2014. [PubMed: 24603191]

Chapter 258: Mitral Stenosis

Patrick T. O'Gara; Joseph Loscalzo

INTRODUCTION

The role of the physical examination in the evaluation of patients with valvular heart disease is also considered in **Chaps. 38** and **234**; of electrocardiography (ECG) in **Chap. 235**; of echocardiography and other noninvasive imaging techniques in **Chap. 236**; and of cardiac catheterization and angiography in **Chap. 237**.

MITRAL STENOSIS

ETIOLOGY AND PATHOLOGY

Rheumatic fever is the leading cause of mitral stenosis (MS) **(Table 258-1; see also Chap. 352)**. Other less common etiologies of obstruction to left ventricular inflow include congenital mitral valve stenosis, cor triatriatum, mitral annular calcification with extension onto the leaflets, systemic lupus erythematosus, rheumatoid arthritis, left atrial myxoma, and infective endocarditis with large vegetations. Pure or predominant MS occurs in ~40% of all patients with rheumatic heart disease and a history of rheumatic fever **(Chap. 352)**. In other patients with rheumatic heart disease, lesser degrees of MS may accompany mitral regurgitation (MR) and aortic valve disease. With reductions in the incidence of acute rheumatic fever, particularly in temperate climates and developed countries, the incidence of MS has declined considerably over the past several decades. However, it remains a major problem in developing nations, especially in tropical and semitropical climates.

TABLE 258-1

Major Causes of Mitral Stenosis

Etiologies
Rheumatic fever
Congenital (parachute valve, cor triatriatum)
Severe mitral annular calcification with leaflet involvement
SLE, RA
Myxoma
IE with large vegetations

Abbreviations: IE, infective endocarditis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus.

In rheumatic MS, chronic inflammation leads to diffuse thickening of the valve leaflets with formation of fibrous tissue often with calcific deposits. The mitral commissures fuse, the chordae tendineae fuse and shorten, the valvular cusps become rigid, and the pathologic process eventually leads to narrowing at the apex of the funnel-shaped ("fish-mouth") valve. Although the initial insult to the mitral valve is rheumatic, later changes may be exacerbated by a nonspecific process resulting from trauma to the valve due to altered flow patterns. Calcification of the stenotic mitral valve immobilizes the leaflets and narrows the orifice further. Thrombus formation and arterial embolization may arise from the calcific valve itself, but in patients with atrial fibrillation (AF), thrombi arise more frequently from the dilated left atrium (LA), particularly from within the LA appendage.

PATHOPHYSIOLOGY

In normal adults, the area of the mitral valve orifice is 4–6 cm^2. In the presence of significant obstruction, i.e., when the orifice area is reduced to <~2 cm^2, blood can flow from the LA to the left ventricle (LV) only if propelled by an abnormally elevated left atrioventricular pressure gradient, the hemodynamic hallmark of MS. When the mitral valve opening is reduced to <1.5 cm^2, referred to as "severe" MS, an LA pressure of ~25 mmHg is required to maintain a normal cardiac output (CO). The elevated pulmonary venous and pulmonary arterial (PA) wedge pressures reduce pulmonary compliance, contributing to exertional dyspnea. The first bouts of dyspnea are usually precipitated by clinical events that increase the rate of blood flow across the mitral orifice, resulting in further elevation of the LA pressure (see below).

To assess the severity of obstruction hemodynamically, both the transvalvular pressure gradient and the flow rate must be measured **(Chap. 237)**. The latter depends not only on the CO but on the heart rate, as well. An increase in heart rate shortens diastole proportionately more than systole and diminishes the time available for flow across the mitral valve. Therefore, at any given level of CO, tachycardia, including that associated with rapid AF, augments the transvalvular pressure gradient and elevates further the LA pressure. Similar considerations apply to the pathophysiology of tricuspid stenosis (TS).

The LV diastolic pressure and ejection fraction (EF) are normal in isolated MS. In MS and sinus rhythm, the elevated LA and PA wedge pressures exhibit a prominent atrial contraction pattern (*a* wave) and a gradual pressure decline after the *v* wave and mitral valve opening (*y* descent). In severe MS and whenever pulmonary vascular resistance is significantly increased, the PA pressure (PAP) is elevated at rest and rises further during exercise, often causing secondary elevations of right ventricular (RV) end-diastolic pressure and volume.

Cardiac Output

In patients with severe MS (mitral valve orifice 1–1.5 cm^2), the CO is normal or almost so at rest, but rises subnormally during exertion. In patients with very severe MS (valve area <1 cm^2), particularly those in whom pulmonary vascular resistance is markedly elevated, the CO is subnormal at rest and may fail to rise or may even decline during activity.

Pulmonary Hypertension

The clinical and hemodynamic features of MS are influenced importantly by the level of the PAP. Pulmonary hypertension results from: (1) passive backward transmission of the elevated LA pressure; (2) pulmonary arteriolar constriction (the so-called "second stenosis"), which presumably is triggered by LA and pulmonary venous hypertension (reactive pulmonary hypertension); (3) interstitial edema in the walls of the small pulmonary vessels; and (4) at end stage, organic obliterative changes in the pulmonary vascular bed. Severe pulmonary hypertension results in RV enlargement, secondary tricuspid regurgitation (TR), and pulmonic regurgitation (PR), as well as right-sided heart failure.

SYMPTOMS

In temperate climates, the latent period between the initial attack of rheumatic carditis (in the increasingly rare circumstances in which a history of one can be elicited) and the development of symptoms due to MS is generally about two decades; most patients begin to experience disability in the fourth decade of life. Studies carried out before the development of surgical mitral valvotomy revealed that once a patient with MS became seriously symptomatic, the disease progressed inexorably to death within 2–5 years.

In patients whose mitral orifices are large enough to accommodate a normal blood flow with only mild elevations of LA pressure, marked elevations of this pressure leading to dyspnea and cough may be precipitated by sudden changes in the heart rate, volume status, or CO, as, for example, with severe exertion, excitement, fever, severe anemia, paroxysmal AF and other tachycardias, sexual intercourse, pregnancy, and thyrotoxicosis. As MS progresses, lesser degrees of stress precipitate dyspnea, the patient becomes limited in daily activities, and orthopnea and paroxysmal nocturnal dyspnea develop. The development of persistent AF often marks a turning point in the patient's course and is generally associated with acceleration of the rate at which symptoms progress. *Hemoptysis* **(Chap. 35)** results from rupture of pulmonary-bronchial venous connections secondary to pulmonary venous hypertension. It occurs most frequently in patients who have elevated LA pressures without markedly elevated pulmonary vascular resistances and is rarely fatal. *Recurrent pulmonary emboli* **(Chap. 273)**, sometimes with infarction, are an important cause of morbidity and mortality late in the course of MS. *Pulmonary infections*, i.e., bronchitis, bronchopneumonia, and lobar pneumonia, commonly complicate untreated MS, especially during the winter months.

Pulmonary Changes

In addition to the aforementioned changes in the pulmonary vascular bed, fibrous thickening of the walls of the alveoli and pulmonary capillaries occurs commonly in MS. The vital capacity, total lung capacity, maximal breathing capacity, and oxygen uptake per unit of ventilation are reduced **(Chap. 279)**. Pulmonary compliance falls further as pulmonary capillary pressure rises during exercise.

Thrombi and Emboli

Thrombi may form in the left atria, particularly within the enlarged atrial appendages of patients with MS. Systemic embolization, the incidence of which is 10–20%, occurs more frequently in patients with AF, in patients >65 years of age, and in those with a reduced CO. However, systemic embolization may be the presenting feature in otherwise asymptomatic patients with only mild MS.

PHYSICAL FINDINGS

(See also Chaps. 38 and 234)

Inspection and Palpation

In patients with severe MS, there may be a malar flush with pinched and blue facies. In patients with sinus rhythm and severe pulmonary hypertension or associated TS, the jugular venous pulse reveals prominent *a* waves due to vigorous right atrial systole. The systemic arterial pressure is usually normal or slightly low. A parasternal lift signifies an enlarged RV. A diastolic thrill may rarely be present at the cardiac apex, with the patient in the left lateral recumbent position.

Auscultation

The first heart sound (S_1) is usually accentuated in the early stages of the disease and slightly delayed. The pulmonic component of the second heart sound (P_2) also is often accentuated with elevated PAPs, and the two components of the second heart sound (S_2) are closely split. The opening snap (OS) of the mitral valve is most readily audible in expiration at, or just medial to, the cardiac apex. This sound generally follows the sound of aortic valve closure (A_2) by 0.05–0.12 s. The time interval between A_2 and OS varies inversely with the severity of the MS. The OS is followed by a low-pitched, rumbling, diastolic murmur, heard best at the apex with the patient in the left lateral recumbent position **(see Fig. 234-5)**; it is accentuated by mild exercise (e.g., a few rapid sit-ups) carried out just before auscultation. In general, the duration of this murmur correlates with the severity of the stenosis in patients with preserved CO. In patients with sinus rhythm, the murmur often reappears or becomes louder during atrial systole (presystolic accentuation). Soft, grade I or II/VI systolic murmurs may be heard at or medial to the apex and may signify mixed mitral valve disease with regurgitation. Hepatomegaly, ankle edema, ascites, and pleural effusion, particularly in the right pleural cavity, may occur in patients with MS and RV failure.

Associated Lesions

With severe pulmonary hypertension, a pansystolic murmur produced by functional TR may be audible along the left sternal border. This murmur is usually louder during inspiration and diminishes during forced expiration (Carvallo's sign). When the CO is markedly reduced in MS, the typical auscultatory findings, including the diastolic rumbling murmur, may not be detectable (silent MS), but they may reappear as compensation is restored. The *Graham Steell murmur* of PR, a high-pitched, diastolic, decrescendo blowing murmur along the left sternal border, results from dilation of the pulmonary valve ring and occurs in patients with mitral valve disease and severe pulmonary hypertension. This murmur may be indistinguishable from the more common murmur produced by aortic regurgitation (AR), although it may increase in intensity with inspiration and is accompanied by a loud and often palpable P_2.

LABORATORY EXAMINATION

ECG

In MS and sinus rhythm, the P wave usually suggests LA enlargement **(see Fig. 235-8)**. It may become tall and peaked in lead II and upright in lead V_1 when severe pulmonary hypertension or TS complicates MS and right atrial (RA) enlargement occurs. The QRS complex is usually normal. However, with severe pulmonary hypertension, right axis deviation and RV hypertrophy are often present.

Echocardiogram

Transthoracic echocardiography (TTE) with color flow and spectral Doppler imaging provides critical information, including measurements of mitral inflow velocity during early (E wave) and late (A wave in patients in sinus rhythm) diastolic filling, estimates of the transvalvular peak and mean gradients and mitral orifice area, the presence and severity of any associated MR, the extent of leaflet calcification and restriction, the degree of distortion of the subvalvular apparatus, and the anatomic suitability for percutaneous mitral balloon valvotomy (PMBV; see below) **(See also Chap. 236)**. In addition, TTE provides an assessment of LV and RV function, chamber sizes, an estimation of the PA systolic pressure based on the tricuspid regurgitant jet velocity, and an indication of the presence and severity of any associated valvular lesions, such as aortic stenosis (AS) and/or regurgitation. Transesophageal echocardiography (TEE) provides superior images and should be used when TTE is inadequate for guiding management decisions. TEE is especially indicated to exclude the presence of LA thrombus prior to PMBV. The performance of TTE with exercise to

evaluate the mean mitral diastolic gradient and PAPs can be very helpful in the evaluation of patients with MS when there is a discrepancy between the clinical findings and the resting hemodynamics.

Chest X-Ray

The earliest changes are straightening of the upper left border of the cardiac silhouette, prominence of the main PAs, dilation of the upper lobe pulmonary veins, and posterior displacement of the esophagus by an enlarged LA. Kerley B lines are fine, dense, opaque, horizontal lines that are most prominent in the lower and mid-lung fields that result from distention of interlobular septae and lymphatics with edema when the resting mean LA pressure exceeds ~20 mmHg.

DIFFERENTIAL DIAGNOSIS

Like MS, significant MR may also be associated with a prominent diastolic murmur at the apex due to increased antegrade transmitral flow, but in patients with isolated MR, this diastolic murmur commences slightly later than in patients with MS, and there is often clear-cut evidence of LV enlargement. An OS and increased P_2 are absent, and S_1 is soft or absent. An apical pansystolic murmur of at least grade III/VI intensity as well as an S_3 suggests significant MR. Similarly, the apical mid-diastolic murmur associated with severe AR (*Austin Flint murmur*) may be mistaken for MS, but can be differentiated from it because it is not intensified in presystole and becomes softer with administration of amyl nitrite or other arterial vasodilators. TS, which occurs rarely in the absence of MS, may mask many of the clinical features of MS or be clinically silent; when present, the diastolic murmur of TS increases with inspiration and the *y* descent in the jugular venous pulse is delayed.

Atrial septal defect (Chap. 264) may be mistaken for MS; in both conditions, there is often clinical, ECG, and chest x-ray evidence of RV enlargement and accentuation of pulmonary vascularity. However, the absence of LA enlargement and of Kerley B lines and the demonstration of fixed splitting of S_2 with a grade II or III mid-systolic murmur at the mid to upper left sternal border all favor atrial septal defect over MS. Atrial septal defects with large left-to-right shunts may result in functional TS because of the enhanced diastolic flow.

Left atrial myxoma (Chap. 266) may obstruct LA emptying, causing dyspnea, a diastolic murmur, and hemodynamic changes resembling those of MS. However, patients with an LA myxoma often have features suggestive of a systemic disease, such as weight loss, fever, anemia, systemic emboli, and elevated serum IgG and interleukin 6 (IL-6) concentrations. The auscultatory findings may change markedly with body position. The diagnosis can be established by the demonstration of a characteristic echo-producing mass in the LA with TTE.

CARDIAC CATHETERIZATION

Left and right heart catheterization can be useful when there is a discrepancy between the clinical and noninvasive findings, including those from TEE and exercise echocardiographic testing when appropriate. Catheterization is helpful in assessing associated lesions, such as AS and AR. Catheterization and coronary angiography are not usually necessary to aid in decision-making about surgery in patients <65 years of age with typical findings of severe mitral obstruction on physical examination and TTE. In men >40 years of age, women >45 years of age, and younger patients with coronary risk factors, especially those with positive noninvasive stress tests for myocardial ischemia, coronary angiography is advisable preoperatively to identify patients with critical coronary obstructions that should be bypassed at the time of operation. Computed tomographic coronary angiography (CTCA) is often used to screen preoperatively for the presence of coronary artery disease (CAD) in appropriate patients with valvular heart disease and low pretest likelihood of CAD. Catheterization and left ventriculography may be useful in patients who have undergone PMBV or previous mitral valve surgery for MS, and who have redeveloped limiting symptoms, especially if questions regarding the severity of the valve lesion(s) remain after noninvasive study.

TREATMENT

TREATMENT

Mitral Stenosis

Penicillin prophylaxis of group A β-hemolytic streptococcal infections (Chap. 352) for secondary prevention of rheumatic fever is important for at-risk patients with rheumatic MS (Fig. 258-1). Recommendations for infective endocarditis prophylaxis are similar to those for other valve lesions and are restricted to patients at high risk for complications from infection, including patients with a history of endocarditis. In symptomatic patients, some improvement usually occurs with restriction of sodium intake and small doses of oral diuretics. Beta blockers, nondihydropyridine calcium channel blockers (e.g., verapamil or diltiazem), and digitalis glycosides are useful in slowing the ventricular rate of patients with AF. Warfarin therapy targeted to an international normalized ratio (INR) of 2–3 should be administered indefinitely to patients with MS who have AF or a history of thromboembolism. The routine use of warfarin in patients in sinus rhythm with LA enlargement (maximal dimension >5.5 cm) with or without spontaneous echo contrast is more controversial. As of this writing, direct oral anticoagulants (e.g., apixaban, rivaroxaban) are not approved for use in patients with rheumatic MS.

If AF is of relatively recent onset in a patient whose MS is not severe enough to warrant PMBV or surgical intervention, reversion to sinus rhythm pharmacologically or by means of electrical countershock is indicated. Usually, cardioversion should be undertaken after the patient has had at least 3 consecutive weeks of anticoagulant treatment to a therapeutic INR. If cardioversion is indicated more urgently, then intravenous heparin should be provided and TEE performed to exclude the presence of LA thrombus before the procedure. Conversion to sinus rhythm is rarely successful or sustained in patients with severe MS, particularly those in whom the LA is especially enlarged or in whom AF has been present for more than 1 year.

MITRAL VALVOTOMY

Unless there is a contraindication, mitral valvotomy is indicated in symptomatic (New York Heart Association [NYHA] Functional Class II–IV) patients with isolated severe MS, whose effective orifice (valve area) is < ~1 cm^2/m^2 body surface area, or <1.5 cm^2 in normal-sized adults. Mitral valvotomy can be carried out either percutaneously or surgically. In PMBV (**Figs. 258-2 and 258-3**), a catheter is directed into the LA after transseptal puncture, and a single balloon is directed across the valve and inflated in the valvular orifice. Ideal patients have relatively pliable leaflets with little or no commissural calcium. In addition, the subvalvular structures should not be significantly scarred or thickened, and there should be no LA thrombus. The short- and long-term results of this procedure in appropriate patients are similar to those of surgical valvotomy, but with less morbidity and a lower periprocedural mortality rate. Event-free survival in younger (<45 years) patients with pliable valves is excellent, with rates as high as 80–90% over 3–7 years. Therefore, PMBV has become the procedure of choice for such patients when it can be performed by a skilled operator in a high-volume center.

TTE is helpful in identifying patients for the percutaneous procedure; TEE is performed routinely to exclude LA thrombus and to assess the degree of MR at the time of the scheduled procedure. An “echo score” has been developed to help guide decision-making. The score accounts for the degree of leaflet thickening, calcification, and mobility, and for the extent of subvalvular thickening. A lower score predicts a higher likelihood of successful PMBV.

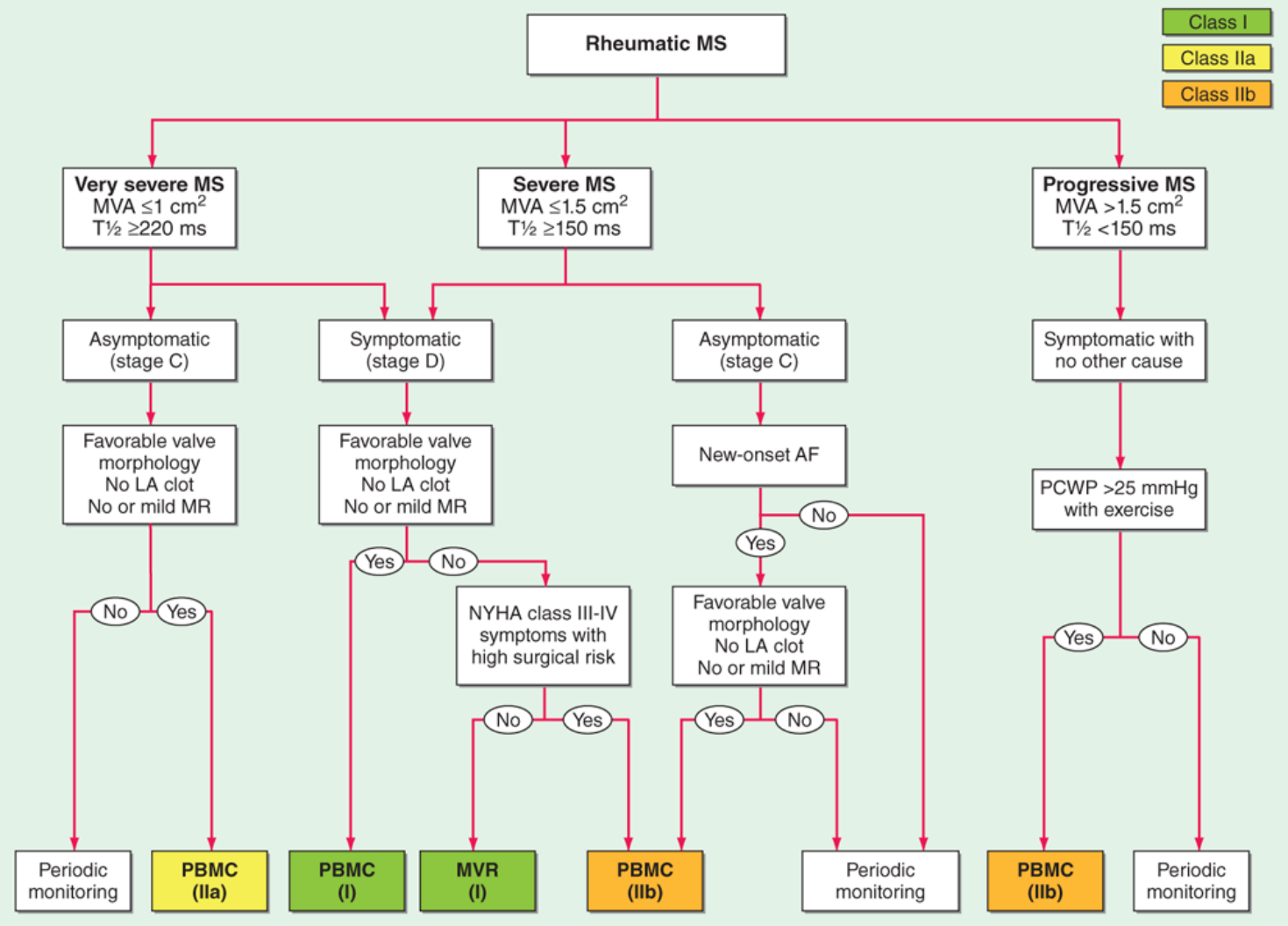
In patients in whom PMBV is not possible or unsuccessful, or in many patients with restenosis after previous surgery, an “open” surgical valvotomy using cardiopulmonary bypass is necessary. In addition to opening the valve commissures, it is important to loosen any subvalvular fusion of papillary muscles and chordae tendineae; to remove large deposits of calcium, thereby improving valvular function; and to remove atrial thrombi. The perioperative mortality rate is ~2%.

Successful valvotomy is defined by a 50% reduction in the mean mitral valve gradient and a doubling of the mitral valve area. Successful valvotomy, whether balloon or surgical, usually results in striking symptomatic and hemodynamic improvement and prolongs survival. However, there is no evidence that the procedure improves the prognosis of patients with slight or no functional impairment. Therefore, unless recurrent systemic embolization or severe pulmonary hypertension has occurred (PA systolic pressures >50 mmHg at rest or >60 mmHg with exercise), valvotomy is *not* recommended for patients who are entirely asymptomatic and/or who have mild or moderate stenosis (mitral valve area >1.5 cm^2). When there is little symptomatic improvement after valvotomy, it is likely that the procedure was ineffective, that it induced MR, or that associated valvular or myocardial disease was present. About half of all patients undergoing surgical mitral valvotomy require reoperation by 10 years. In the pregnant patient with MS, valvotomy should be carried out if pulmonary congestion occurs despite intensive medical treatment. PMBV is the preferred strategy in this setting and is performed with TEE and no or minimal x-ray exposure.

Mitral valve replacement (MVR) is necessary in patients with MS and significant associated MR, those in whom the valve has been severely distorted by previous transcatheter or operative manipulation, or those in whom the surgeon does not find it possible to improve valve function significantly with valvotomy. MVR is now routinely performed with preservation of the chordal attachments to optimize LV functional recovery. Perioperative mortality rates with MVR vary with age, LV function, the presence of CAD, and associated comorbidities. They average 5% overall but are lower in young patients and may be twice as high in patients >65 years of age with significant comorbidities (**Table 258-2**). Because there are also long-term complications of valve replacement, patients in whom preoperative evaluation suggests the possibility that MVR may be required should be operated on only if they have severe MS—i.e., an orifice area ≤1.5 cm^2—and are in NYHA Class III, i.e., symptomatic with ordinary activity despite optimal medical therapy. The overall 10-year survival of surgical survivors is ~70%. Long-term prognosis is worse in patients >65 years of age and those with marked disability and marked depression of the CO preoperatively. Pulmonary hypertension and RV dysfunction are additional risk factors for poor outcome.

FIGURE 258-1

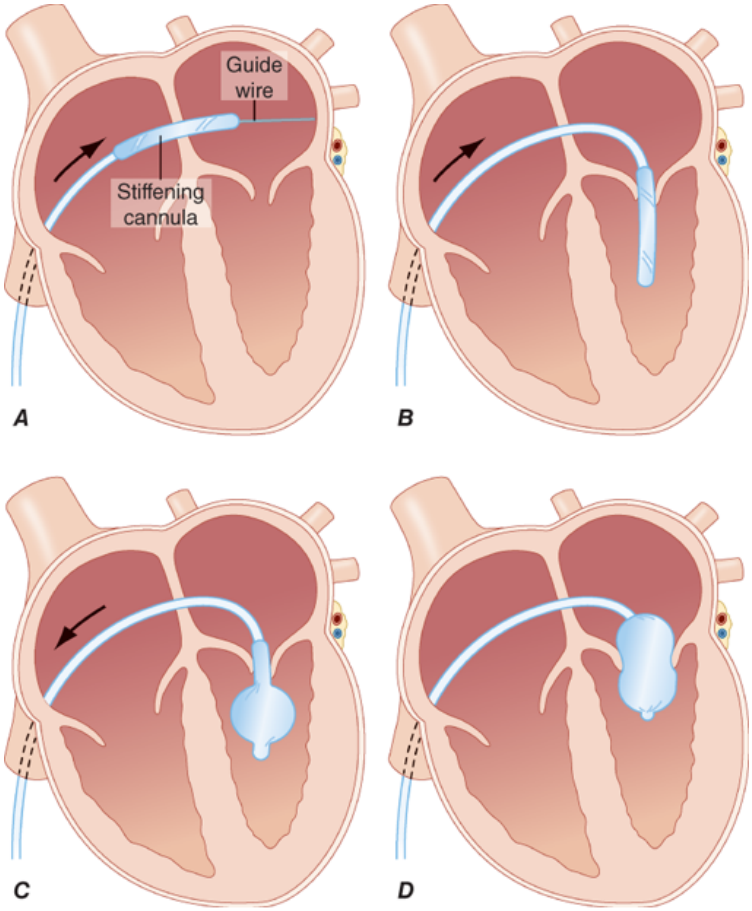
Management of rheumatic mitral stenosis. See legend for **Fig. 256-4** for explanation of treatment recommendations (class I, IIa, IIb) and disease stages (C, D). Preoperative coronary angiography should be performed routinely as determined by age, symptoms, and coronary risk factors. Cardiac catheterization and angiography may also be helpful when there is a discrepancy between clinical and noninvasive findings. AF, atrial fibrillation; LA, left atrial; MR, mitral regurgitation; MS, mitral stenosis; MVA, mitral valve area; MVR, mitral valve surgery (repair or replacement); NYHA, New York Heart Association; PCWP, pulmonary capillary wedge pressure; PMBC, percutaneous mitral balloon commissurotomy; and T ½, pressure half-time. *(Adapted from RA Nishimura et al: 2014 AHA/ACC Guideline for the Management of Patients with Valvular Heart Disease. J Am Coll Cardiol 63:e57-185, 2014, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 258-2

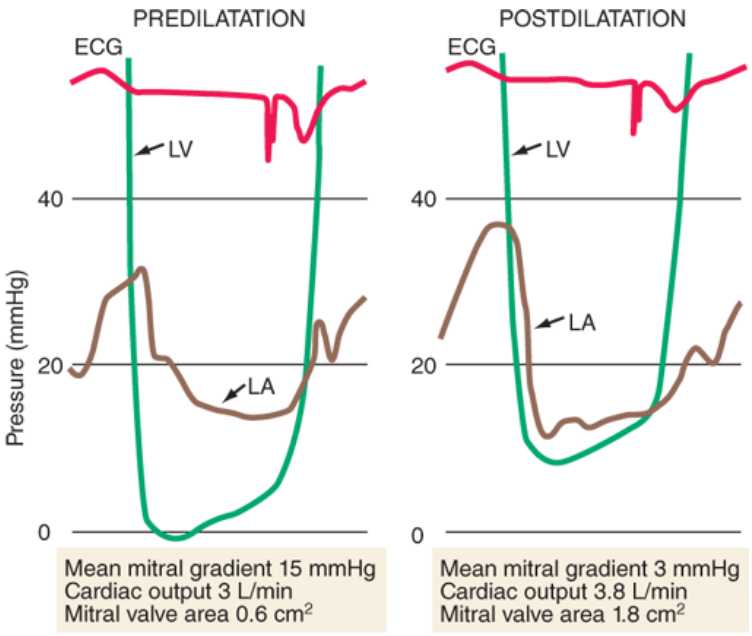
Inoue balloon technique for percutaneous mitral balloon valvotomy. ***A.*** After transseptal puncture, the deflated balloon catheter is advanced across the interatrial septum, then across the mitral valve and into the left ventricle. ***B–D.*** The balloon is inflated stepwise within the mitral orifice.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 258-3

Simultaneous left atrial (LA) and left ventricular (LV) pressure before and after percutaneous mitral balloon valvotomy (PMBV) in a patient with severe mitral stenosis. ECG, electrocardiogram. *(Courtesy of Raymond G. McKay, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 258-2

Mortality Rates after Mitral Valve Surgery[a]

Operation	Number	Unadjusted Operative Mortality (%)
MVR (isolated)	3448	4.6
MVR + CAB	1321	10.0
MVRp	4284	1.2
MVRp + CAB	2051	4.8

[a]Data are for the first two quarters of calendar year 2015, during which 1013 sites reported a total of 141,225 procedures. Data are available from the Society of Thoracic Surgeons at *http://www.sts.org/sites/default/files/documents/2015Harvest3_ExecutiveSummary.pdf.*

Abbreviations: CAB, coronary artery bypass; MVR, mitral valve replacement; MVRp, mitral valve repair.

FURTHER READING

Nishimura RA et al: 2014 AHA/ACC guidelines for management of patients with valvular heart disease. J Am Coll Cardiol 63:357, 2014.

Nishimura RA et al: Mitral valve disease–Current management and future challenges. Lancet 387:1324, 2016. [PubMed: 27025438]

Chapter 259: Mitral Regurgitation

Patrick T. O'Gara; Joseph Loscalzo

INTRODUCTION

The role of the physical examination in the evaluation of patients with valvular heart disease is also considered in **Chaps. 38** and **234**; of electrocardiography (ECG) in **Chap. 235**; of echocardiography and other noninvasive imaging techniques in **Chap. 236**; and of cardiac catheterization and angiography in **Chap. 237**.

ETIOLOGY

Mitral regurgitation (MR) may result from an abnormality or disease process that affects any one or more of the five functional components of the mitral valve apparatus (leaflets, annulus, chordae tendineae, papillary muscles, and subjacent myocardium) **(Table 259-1)**. Acute MR can occur in the setting of acute myocardial infarction (MI) with papillary muscle rupture **(Chap. 269)**, following blunt chest wall trauma, or during the course of infective endocarditis (IE) owing to leaflet perforation or destruction. With acute MI, the posteromedial papillary muscle is involved much more frequently than the anterolateral papillary muscle because of its singular blood supply. Transient, acute MR can occur during periods of active ischemia and bouts of angina pectoris. Rupture of chordae tendineae can result in "acute-on-chronic MR" in patients with myxomatous degeneration of the valve apparatus.

TABLE 259-1

Major Causes of Mitral Regurgitation

Etiologies
Acute
IE
Papillary muscle rupture (post-MI)
Chordal rupture/leaflet flail (MVP, IE)
Blunt trauma
Chronic
Primary (affecting leaflets, chordae)
Myxomatous (MVP, Barlow's, forme fruste)
Rheumatic fever
IE (healed)
Congenital (cleft, AV canal)
Radiation
Secondary (leaflets, chordae are "innocent bystanders")
Ischemic cardiomyopathy
Dilated cardiomyopathy
HOCM (with SAM)
Chronic AF with LA enlargement and annular dilatation
Mitral annular calcification[a]

[a]Mitral annular calcification may include elements of both primary and secondary MR as the disease process may encroach on the leaflets, impair the normal sphincteric function of the annulus, or both.

Abbreviations: AF, atrial fibrillation; AV, atrioventricular; IE, infective endocarditis; HOCM, hypertrophic obstructive cardiomyopathy; LA, left atrial; LV, left ventricular; MI, myocardial infarction; MVP, mitral valve prolapse; SAM, systolic anterior motion.

Chronic MR can result from several disease processes (Table 259-1). Distinction should be drawn between primary MR, in which the leaflets and/or chordae tendineae are primarily responsible for abnormal valve function, and secondary (functional) MR, in which the leaflets and chordae tendineae are usually normal but the regurgitation is caused by left ventricular (LV) remodeling with annular enlargement, papillary muscle displacement, leaflet tethering, or their combination. Patient assessment, treatment approach, and long-term prognosis differ significantly between primary and secondary MR. Mitral valve prolapse (MVP) is discussed more extensively in **Chap. 260**. The rheumatic process produces rigidity, deformity, and retraction of the valve cusps and commissural fusion, as well as shortening, contraction, and fusion of the chordae tendineae. MR can persist after resolution of the acute phase of infection and inflammation. MR may occur as a congenital anomaly **(Chap. 264)**, most commonly as a defect of the endocardial cushions (atrioventricular cushion defects). A cleft anterior mitral valve leaflet accompanies ostium primum atrial septal defect. Radiation can result in leaflet thickening, retraction, and calcification, often in association with annular and chordal involvement. Chronic MR is frequently secondary to ischemia and may occur as a consequence of ventricular remodeling, papillary muscle displacement, and leaflet tethering, or with fibrosis of a

papillary muscle, in patients with healed MI(s) and ischemic cardiomyopathy. Similar mechanisms of annular dilation and ventricular remodeling contribute to the MR that occurs among patients with nonischemic forms of dilated cardiomyopathy once the LV end-diastolic dimension reaches 6 cm. The MR associated with hypertrophic obstructive cardiomyopathy (HOCM) is usually dynamic in nature and dependent on systolic anterior motion of the anterior mitral valve leaflet into a narrowed LV outflow tract. Patients with long-standing, chronic atrial fibrillation (AF) develop atrial remodeling and annular dilatation that can result in MR. Annular calcification can result in MR when it encroaches on the leaflets or results in decreased sphincteric function and is especially prevalent among patients with advanced renal disease and is commonly observed in women >65 years of age with hypertension and diabetes. Irrespective of cause, chronic severe MR is often progressive, because enlargement of the left atrial (LA) places tension on the posterior mitral leaflet, pulling it away from the mitral orifice and thereby aggravating the valvular dysfunction. Similarly, LV dilation increases the regurgitation, which, in turn, enlarges the LA and LV further, resulting in a vicious circle; hence the aphorism, “MR begets MR.”

PATHOPHYSIOLOGY

The resistance to LV emptying (LV afterload) is reduced in patients with MR. As a consequence, the LV is decompressed into the LA during ejection, and with the reduction in LV size during systole, there is a rapid decline in LV tension. The initial compensation to MR is more complete LV emptying. However, LV volume increases progressively with time as the severity of the regurgitation increases and as LV contractile function deteriorates. This increase in LV volume is often accompanied by a reduced forward cardiac output (CO). LV compliance is often increased, and thus, LV diastolic pressure does not increase until late in the course. The regurgitant volume varies directly with the LV systolic pressure and the size of the regurgitant orifice; the latter, in turn, is influenced by the extent of LV and mitral annular dilation. Because ejection fraction (EF) rises in severe MR in the presence of normal LV function, even a modest reduction in this parameter (<60%) reflects significant contractile dysfunction.

During early diastole, as the distended LA empties, there is a particularly rapid *y* descent in the absence of accompanying MS. A brief, early diastolic LA-LV pressure gradient (often generating a rapid filling sound [S_3] and mid-diastolic murmur masquerading as MS) may occur in patients with pure, severe MR as a result of the very rapid flow of blood across a normal-sized mitral orifice.

Measurements of LV ejection fraction (LVEF), CO, pulmonary arterial (PA) systolic pressure, regurgitant volume, regurgitant fraction (RF), and the effective regurgitant orifice area can be obtained during a careful Doppler echocardiographic examination. These measurements can also be obtained accurately with cardiac magnetic resonance (CMR) imaging, although this technology is not widely available. Left and right heart catheterization with contrast ventriculography is used less frequently. Chronic, severe MR is defined by a regurgitant volume ≥60 mL/beat, RF ≥50%, and effective regurgitant orifice area ≥0.40 cm^2. In patients with secondary MR, in whom the severity of MR can be underappreciated, lesser degrees of regurgitation carry relatively greater prognostic weight.

LA Compliance

In acute severe MR, the regurgitant volume is delivered into a normal-sized LA having normal or reduced compliance. As a result, LA pressures rise markedly for any increase in LA volume. The *v* wave in the LA pressure pulse is usually prominent, LA and pulmonary venous pressures are markedly elevated, and pulmonary edema is common. Because of the rapid rise in LA pressures during ventricular systole, the murmur of acute MR is early in timing and decrescendo in configuration ending well before S_2, as a reflection of the progressive diminution in the LV-LA pressure gradient. LV systolic function in acute MR may be normal, hyperdynamic, or reduced, depending on the clinical context.

Patients with chronic severe MR, on the other hand, develop marked LA enlargement and *increased* LA compliance with little if any increase in LA and pulmonary venous pressures for any increase in LA volume. The LA *v* wave is relatively less prominent. The murmur of chronic MR is classically holosystolic in timing and plateau in configuration, as a reflection of the near-constant LV-LA pressure gradient. These patients usually complain of severe fatigue and exhaustion secondary to a low forward CO, whereas symptoms resulting from pulmonary congestion are less prominent initially; AF is almost invariably present once the LA dilates significantly.

SYMPTOMS

Patients with chronic mild-to-moderate, isolated MR are usually asymptomatic. This form of LV volume overload is well tolerated. Fatigue, exertional dyspnea, and orthopnea are the most prominent complaints in patients with chronic severe MR. Palpitations are common and may signify the onset of AF. Right-sided heart failure, with painful hepatic congestion, ankle edema, distended neck veins, ascites, and secondary tricuspid regurgitation (TR), occurs in patients with MR who have associated pulmonary vascular disease and pulmonary hypertension. Acute pulmonary edema is common in patients with acute severe MR.

PHYSICAL FINDINGS

In patients with chronic severe MR, the arterial pressure is usually normal, although the carotid arterial pulse may show a sharp, low-volume upstroke owing to the reduced forward CO. A systolic thrill is often palpable at the cardiac apex, the LV is hyperdynamic with a brisk systolic impulse and a palpable rapid-filling wave (S_3), and the apex beat is often displaced laterally.

In patients with acute severe MR, the arterial pressure may be reduced with a narrow pulse pressure, the jugular venous pressure and waveforms may be normal or increased and exaggerated, the apical impulse is not displaced, and signs of pulmonary congestion are prominent.

Auscultation

S_1 is generally absent, soft, or buried in the holosystolic murmur of chronic, severe MR. In patients with severe MR, the aortic valve may close prematurely, resulting in wide but physiologic splitting of S_2. A low-pitched S_3 occurring 0.12–0.17 s after the aortic valve closure sound, i.e., at the completion of the rapid-filling phase of the LV, is believed to be caused by the sudden tensing of the papillary muscles, chordae tendineae, and valve leaflets. It may be followed by a short, rumbling, mid-diastolic murmur, even in the absence of structural MS. In patients with ischemic or dilated cardiomyopathy, however, a third sound (S_3) may also signify ventricular dysfunction. A fourth heart sound is often audible in patients with *acute* severe MR who are in sinus rhythm. A presystolic murmur is not ordinarily heard with isolated MR.

A systolic murmur of at least grade III/VI intensity is the most characteristic auscultatory finding in chronic severe MR. It is usually holosystolic **(see Fig. 234-5A)**, but as previously noted, it is decrescendo and ceases in mid-to-late systole in patients with acute severe MR. The systolic murmur of chronic MR is usually most prominent at the apex and radiates to the axilla. However, in patients with ruptured chordae tendineae or primary involvement of the posterior mitral leaflet with prolapse or flail, the regurgitant jet is eccentric, directed anteriorly, and strikes the LA wall adjacent to the aortic root. In this situation, the systolic murmur is transmitted to the base of the heart and, therefore, may be confused with the murmur of AS. In patients with ruptured chordae tendineae, the systolic murmur may have a cooing or “seagull” quality, whereas a flail leaflet may produce a murmur with a musical quality. The systolic murmur of chronic MR not due to MVP is intensified by isometric exercise (handgrip) but is reduced during the strain phase of the Valsalva maneuver because of the associated decrease in LV preload.

LABORATORY EXAMINATION

ECG

In patients with sinus rhythm, there is evidence of LA enlargement, but right atrial (RA) enlargement also may be present when pulmonary hypertension is significant and affects RV function and size. Chronic severe MR is frequently associated with AF. In many patients, there is no clear-cut ECG evidence of enlargement of either ventricle. In others, the signs of eccentric LV hypertrophy are present.

Echocardiogram

Transthoracic echocardiography (TTE) is indicated to assess the mechanism of the MR and its hemodynamic severity. LV function can be assessed from LV end-diastolic and end-systolic volumes and EF. Observations can be made regarding leaflet structure and function, chordal integrity, LA and LV size, annular calcification, and regional and global LV systolic function. Doppler imaging should demonstrate the width or area of the color flow MR jet within the LA, the duration and intensity of the continuous wave Doppler signal, the pulmonary venous flow contour, the early peak mitral inflow velocity, and quantitative measures of regurgitant volume, RF, and effective regurgitant orifice area. In addition, the PA pressures (PAPs) can be estimated from the TR jet velocity. TTE is also indicated to follow the course of patients with chronic MR and to provide rapid assessment for any clinical change. Transesophageal echocardiography (TEE) provides greater anatomic detail than TTE **(see Fig. 236-5)**. Exercise testing with TTE can be useful to assess exercise capacity as well as any dynamic change in MR severity, PA systolic pressures, and biventricular function, for patients in whom there is a discrepancy between clinical findings and the results of other noninvasive testing.

Chest X-Ray

The LA and LV are the dominant chambers in chronic MR. Late in the course of the disease, the LA may be massively enlarged and forms the right border of the cardiac silhouette. Pulmonary venous congestion, interstitial edema, and Kerley B lines are sometimes noted. Marked calcification of the mitral leaflets occurs commonly in patients with long-standing, combined rheumatic MR and MS. Calcification of the mitral annulus may be visualized, particularly on the lateral view of the chest. Patients with acute severe MR may have asymmetric pulmonary edema if the regurgitant jet is directed predominantly to the orifice of an upper lobe pulmonary vein.

TREATMENT

TREATMENT

Mitral Regurgitation

MEDICAL TREATMENT

The management of chronic severe MR depends to some degree on its cause (**Fig. 259-1**). Anticoagulation with either warfarin or a direct oral agent (e.g., apixaban, rivaroxaban) should be provided if AF intervenes, as guided by the CHA2DS2-VASc risk score. The direct oral anticoagulants should not be used if rheumatic mitral stenosis is also present; they are also not approved for use in patients with mechanical prosthetic heart valves. Cardioversion should be considered depending on the clinical context, AF chronicity, LA size. In contrast to the acute setting, there are no large, long-term prospective studies to substantiate the use of vasodilators for the treatment of chronic, isolated severe MR with preserved LV systolic function *in the absence of systemic hypertension.* The severity of MR in the setting of an ischemic or dilated cardiomyopathy may diminish with aggressive guideline-directed treatment of heart failure including the use of diuretics, beta blockers, angiotensin-converting enzyme (ACE) inhibitors, digitalis, and biventricular pacing (cardiac resynchronization therapy [CRT]) when otherwise indicated. Antibiotic prophylaxis for prevention of IE is indicated for MR patients with a prior history of IE. Asymptomatic patients with severe MR in sinus rhythm with normal LV size and systolic function should avoid isometric forms of exercise.

Patients with acute severe MR require urgent stabilization and preparation for surgery. Diuretics, intravenous vasodilators (particularly sodium nitroprusside), and even mechanical support may be needed for patients with post-MI papillary muscle rupture or other forms of acute severe MR.

SURGICAL TREATMENT

In the selection of patients with chronic, severe, primary MR for surgical treatment, the often slowly progressive nature of the condition must be balanced against the immediate and long-term risks associated with operation. These risks are significantly lower for primary valve repair than for valve replacement **(Table 259-2)**. Repair usually consists of valve reconstruction using a variety of valvuloplasty techniques and insertion of an annuloplasty ring. Repair spares the patient the long-term adverse consequences of valve replacement, including thromboembolic and hemorrhagic complications in the case of mechanical prostheses and late valve failure necessitating repeat valve replacement in the case of bioprostheses. In addition, by preserving the integrity of the papillary muscles, subvalvular apparatus, and chordae tendineae, mitral repair and valvuloplasty maintain LV function to a relatively greater degree than does valve replacement.

Surgery for chronic severe primary MR is indicated once symptoms occur, especially if valve repair is feasible (Fig. 259-1). Surgery should also be recommended for asymptomatic patients with LV dysfunction characterized by an EF <60% or an LV end-systolic dimension (LV ESD) >40 mm. Other indications for early consideration of mitral valve repair in asymptomatic patients include (1) recent-onset AF (duration <3 months); (2) pulmonary hypertension (defined as a systolic PA pressure ≥50 mmHg at rest or ≥60 mmHg with exercise); and (3) a progressive decrease in LV EF or increase in LV ESD on serial imaging. These aggressive recommendations for surgery are predicated on the adverse long-term consequences of waiting for LV function to decline further as well as the outstanding results achievable with mitral valve repair by reference surgeons at high-volume centers. Indeed, repair of myxomatous MR (e.g., prolapse, flail) in patients <75 years with normal LV systolic function and no coronary artery disease (CAD) can now be performed by experienced surgeons with <1% perioperative mortality risk. The risk of stroke, however, is also ~1%. Repair is feasible in up to 95% of patients with myxomatous disease operated on by a high-volume surgeon in a referral center of excellence. Repair techniques include chordal transfer, creation of neochords, limited leaflet resection, and insertion of an annuloplasty band. Long-term durability is excellent; the incidence of reoperative surgery for failed primary repair is ~1% per year for the first 10 years after surgery. For patients with AF, left or biatrial maze surgery, or radiofrequency isolation of the pulmonary veins is often performed to reduce the risk of recurrent postoperative AF.

The surgical management of patients with secondary MR is more complicated. Surgery for patients with ischemic MR most often involves simultaneous coronary artery revascularization. Current surgical practice includes annuloplasty repair with an undersized, rigid ring or chord-sparing valve replacement for patients with moderate or greater degrees of MR. Valve repair for ischemic MR is associated with lower perioperative mortality rates than valve replacement but significantly higher rates of recurrent MR over time. In patients with ischemic MR and significantly impaired LV systolic function (EF <30%), the risk of surgery is higher, recovery of LV performance is incomplete, and long-term survival is reduced. Referral for surgery must be individualized and made only after aggressive attempts to improve symptoms with guideline-directed medical therapy and CRT, when indicated. The routine performance of valve repair in patients with significant secondary MR due to a dilated cardiomyopathy has not been shown to improve long-term survival compared with optimal medical therapy. Patients with acute severe MR can often be stabilized temporarily with appropriate medical therapy, but surgical correction will be necessary emergently in the case of papillary muscle rupture and within days to weeks in most other settings.

When surgical treatment is contemplated, left and right heart catheterization and left ventriculography *may* be helpful in confirming the presence of severe MR in patients in whom there is a discrepancy between the clinical and TTE findings that cannot be resolved with TEE or CMR. Coronary angiography identifies patients who require concomitant coronary revascularization.

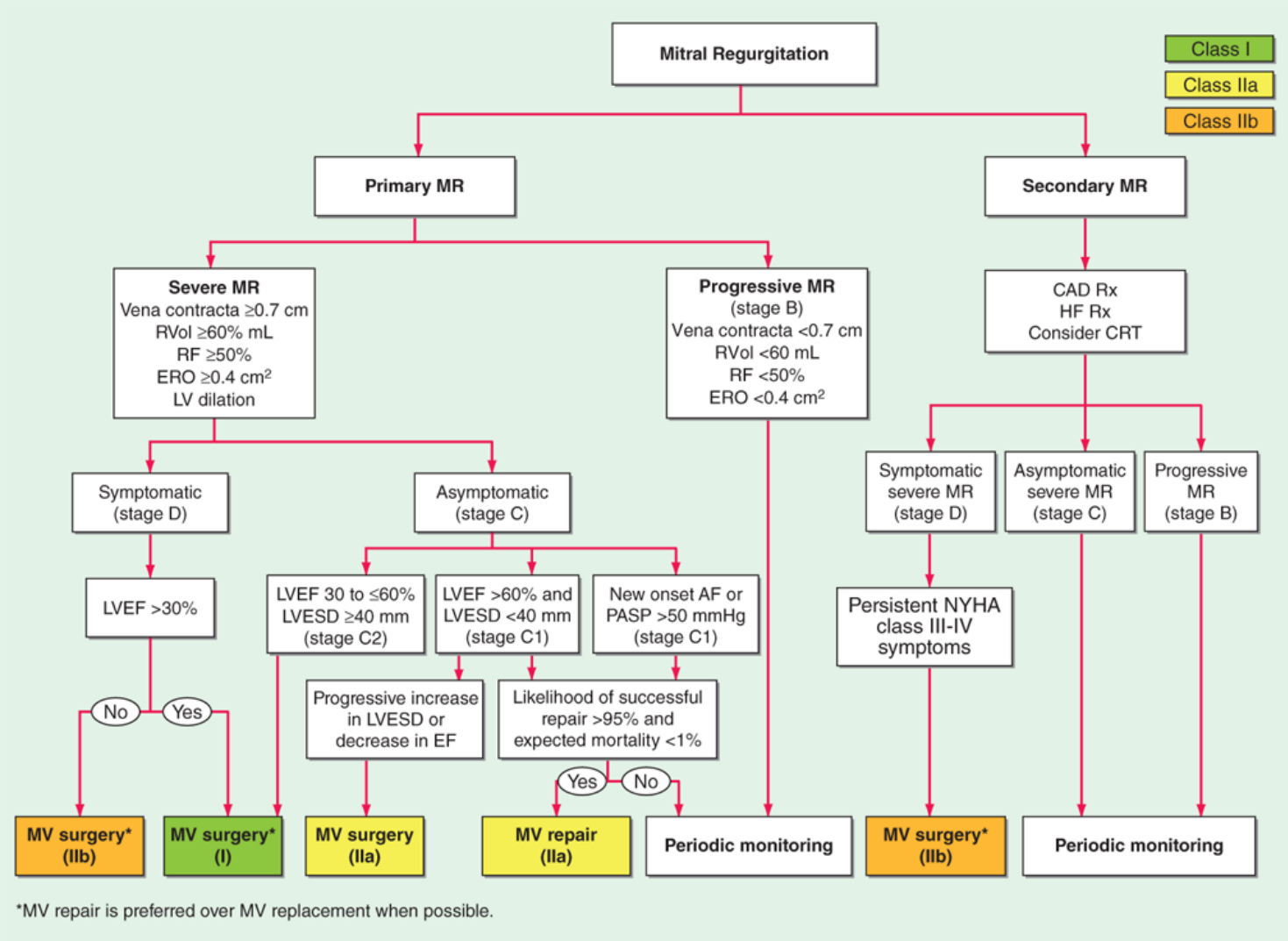
TRANSCATHETER MITRAL VALVE REPAIR AND REPLACEMENT

A transcatheter approach to the treatment of either primary or functional MR may be feasible in selected patients with appropriate anatomy. The proper role of currently available techniques remains under active investigation. One approach involves the deployment of a clip delivered via transseptal puncture that grasps the leading edges of the mitral leaflets in their mid-portion (anterior scallop to posterior scallop or A2-P2; **Fig. 259-2**). The length and width of the gap between these leading edges dictate patient eligibility. The device is commercially available. In the United States only

for the treatment of prohibitive- or high-surgical risk, symptomatic patients with severe, primary (myxomatous) MR. The edge-to-edge clip technique is undergoing study in the United States for treatment of patients with symptomatic heart failure, reduced LVEF, and severe, secondary MR despite guideline-directed medical therapy. Other approaches include the deployment of a device within the coronary sinus that can be adjusted to reduce mitral annular circumference and the effective orifice area of the valve much like a surgically implanted ring. Variations in the anatomic relationship of the coronary sinus to the mitral annulus and circumflex coronary artery have limited the applicability of this technique. Attempts to reduce the septal-lateral dimension of a dilated annulus using adjustable cords placed across the LV in a subvalvular location have also been investigated. Construction of neochords to the mitral leaflets under TEE guidance using a system delivered via the cardiac apex is also under study. Investigational experience to date with transcatheter mitral valve replacement systems is in early clinical stages, although the field is evolving rapidly.

FIGURE 259-1

Management of mitral regurgitation. See legend for Fig. 256-4 for explanation of treatment recommendations (class I, IIa, IIb) and disease stages (B, C1, C2, D). Preoperative coronary angiography should be performed routinely as determined by age, symptoms, and coronary risk factors. Cardiac catheterization and angiography may also be helpful when there is a discrepancy between clinical and noninvasive findings. AF, atrial fibrillation; CAD, coronary artery disease; CRT, cardiac resynchronization therapy; EF, ejection fraction; ERO, effective regurgitant orifice; HF, heart failure; LV, left ventricular; LVEF, left ventricular ejection fraction; LVESD, left ventricular end-systolic dimension; MR, mitral regurgitation, MV, mitral valve; MVR, mitral valve replacement; NYHA, New York Heart Association; PASP, pulmonary artery systolic pressure; RF, regurgitant fraction; RVol, regurgitant volume; and Rx, therapy. *(Adapted from RA Nishimura et al: 2017 Focused Update of the 2014 AHA/ACC Guideline for the Management of Patients with Valvular Heart Disease. J Am Coll Cardiol. Available at www.onlinejacc.org/lookup/doi/10.1016/j.jacc.2017.03.011.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 259-2

Mortality Rates after Mitral Valve Surgery[a]

Operation	Number	Unadjusted Operative Mortality (%)
MVR (isolated)	3448	4.6
MVR + CAB	1321	10.0
MVRp	4284	1.2
MVRp + CAB	2051	4.8

[a]Data are for the first two quarters of calendar year 2015, during which 1013 sites reported a total of 141,225 procedures. Data are available from the Society of Thoracic Surgeons at *http://www.sts.org/sites/default/files/documents/2015Harvest3_ExecutiveSummary.pdf.*

Abbreviations: CAB, coronary artery bypass; MVR, mitral valve replacement; MVRp, mitral valve repair.

FIGURE 259-2

Clip used to grasp the free edges of the anterior and posterior leaflets in their midsections during transcatheter repair of selected patients with mitral regurgitation. *(Courtesy of Abbott Vascular. © 2014 Abbott Laboratories. All rights reserved.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Carabello BA: The myocardium in mitral regurgitation. A tale of two ventricles. Circulation 127:1567, 2013. [PubMed: 23588959]

Maisano F et al: The future of transcatheter mitral valve interventions: Competitive or complementary role of repair vs. replacement. Eur Heart J 36:1651, 2015. [PubMed: 25870204]

Nishimura RA et al: Mitral valve disease. Current management and future challenges. Lancet 387:1324, 2016. [PubMed: 27025438]

Nishimura RA et al: 2017 Focused Update of the 2014 AHA/ACC Guideline for the Management of Patients with Valvular Heart Disease. J Am Coll Cardiol 2017. Available at *www.onlinejacc.org/lookup/doi/ 10.1016/j.jacc.2017.03.011*.

Otto CM, Verrier ED: Mitral regurgitation—What is best for my patient? N Engl J Med 364:1462, 2011. [PubMed: 21463152]

Rugueiro A et al: Transcatheter mitral valve replacement: insights from early clinical experience and future challenges. J Am Coll Cardiol 69:2175, 2017. [PubMed: 28449780]

Chapter 260: Mitral Valve Prolapse

Patrick T. O'Gara; Joseph Loscalzo

INTRODUCTION

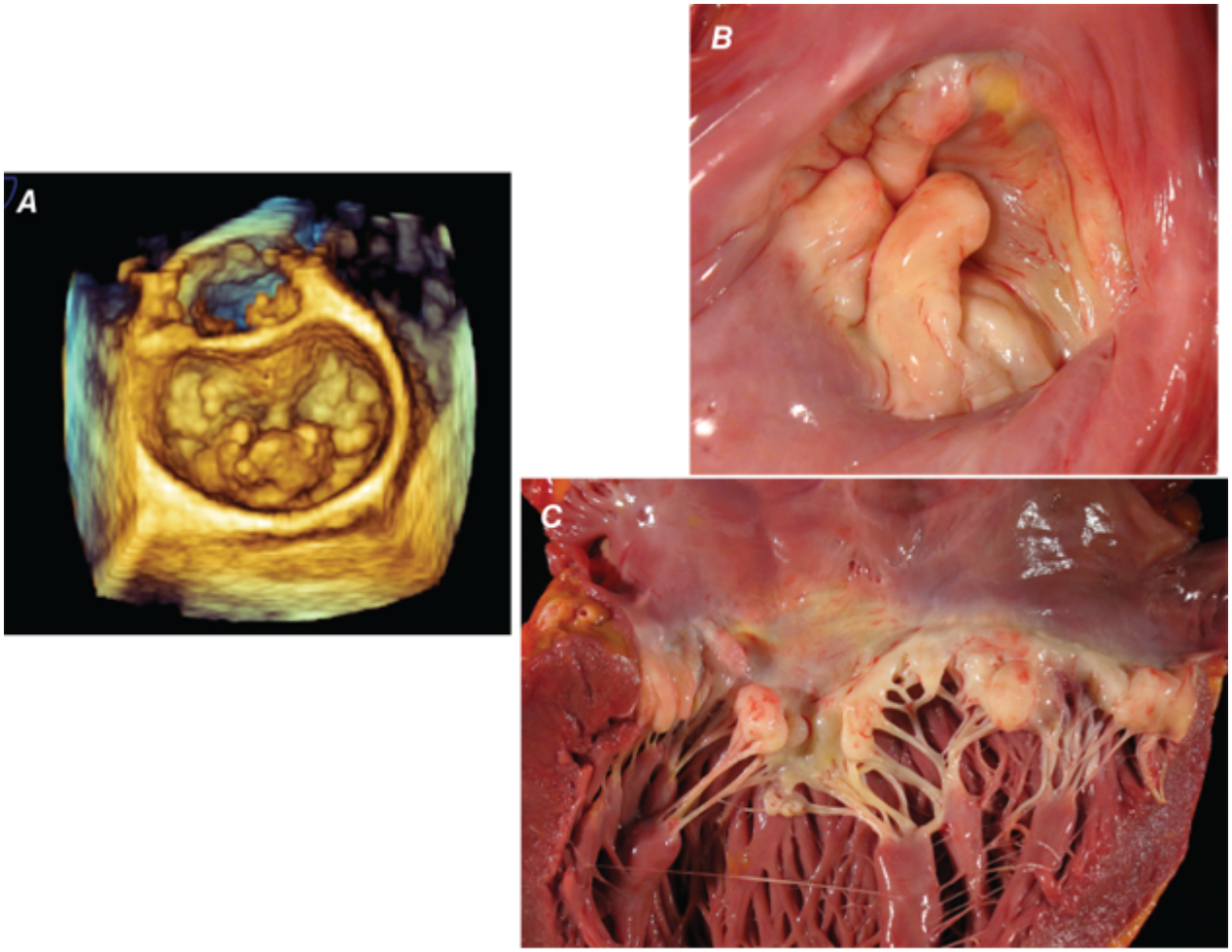
The role of the physical examination in the evaluation of patients with valvular heart disease is also considered in **Chaps. 38** and **234**; of electrocardiography (ECG) in **Chap. 235**; of echocardiography and other noninvasive imaging techniques in **Chap. 236**; and of cardiac catheterization and angiography in **Chap. 237**.

MITRAL VALVE PROLAPSE

Mitral valve prolapse (MVP), also variously termed the *systolic click-murmur syndrome, Barlow's syndrome* (**Fig. 260-1**), *floppy-valve syndrome, and billowing mitral leaflet syndrome*, is a relatively common but highly variable clinical syndrome resulting from diverse pathologic mechanisms of the mitral valve apparatus. Among these are excessive or redundant mitral leaflet tissue, which is commonly associated with myxomatous degeneration and greatly increased concentrations of certain glycosaminoglycans. MVP is the most common abnormality leading to primary mitral regurgitation (MR) (**see Chap. 259**).

FIGURE 260-1

Congenital or developmental mitral valve prolapse. Myxomatous thickening and prolapse of the mitral valve can occur in isolation in 2–3% of the general population, or may be associated with heritable collagen-vascular disorders and aortic root dilatation, such as in Marfan syndrome. Myxomatous degeneration of the valve predisposes to severe regurgitation and chordal rupture, and is a frequent indication for mitral valve repair or replacement. Prolapse can affect only one or both leaflets, to varying degrees. ***A.*** Three-dimensional transesophageal echocardiogram showing a myxomatous mitral valve from the left atrial *en face* aspect. There is billowing and prolapse of the entire middle scallop of the posterior leaflet. *(Figure courtesy of Douglas C. Shook, MD, Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital).* ***B.*** The posterior leaflet of the mitral valve demonstrates marked prolapse and hooding in all segments and severe redundancy in this photograph taken from the vantage point of the left atrium. ***C.*** Opening the left heart reveals prominent mitral leaflet hooding (*arrows*). The chordae are focally thickened, but are not fused as would be the case in rheumatic valve disease. *(Used with permission from JC Wu, RF Padera: Clinicopathologic correlates, in Atlas of Echocardiography, 2nd ed, SD Solomon [ed], E Braunwald [series ed]. Philadelphia, Current Medicine Group LLC, 2008. p 363.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In most patients with MVP, the cause is unknown, but in some, it appears to be genetically determined. A reduction in the production of type III collagen has been incriminated, and electron microscopy has revealed fragmentation of collagen fibrils.

MVP is a frequent finding in patients with heritable disorders of connective tissue, including Marfan syndrome **(Chap. 406)**, osteogenesis imperfecta, and Ehlers-Danlos syndrome. MVP may be associated with thoracic skeletal deformities similar to but not as severe as those in Marfan syndrome, such as a high-arched palate and alterations of the chest and thoracic spine, including the so-called straight back syndrome. Other associated features can include a history of inguinal hernias, joint dislocations, meniscal tears, and easy bruisability.

In most patients with MVP, myxomatous degeneration is confined to the mitral valve, although the tricuspid and aortic valves may also be affected. The posterior mitral leaflet is usually more affected than the anterior, and the mitral valve annulus is often dilated. In many patients, elongated, redundant, or ruptured chordae tendineae cause or contribute to the regurgitation.

MVP also may occur rarely as a sequel to acute rheumatic fever, in ischemic heart disease, and in various cardiomyopathies, as well as in 20% of patients with ostium secundum atrial septal defect.

MVP may lead to excessive stress on the papillary muscles, which, in turn, leads to dysfunction and ischemia of the papillary muscles and the subjacent ventricular myocardium. Rupture of chordae tendineae and progressive annular dilation and calcification contribute to valvular regurgitation, which then places more stress on the diseased mitral valve apparatus, thereby creating a vicious circle. ECG changes (see below) and ventricular arrhythmias described in some patients with MVP appear to result from regional ventricular dysfunction related to the increased stress placed on the papillary muscles.

CLINICAL FEATURES

MVP is more common in women and occurs most frequently between the ages of 15 and 30 years; the clinical course is most often benign. MVP may also be observed in older (>50 years) patients, often men, in whom MR is often more severe and requires surgical treatment. There is an increased familial incidence for some patients, suggesting an autosomal dominant form of inheritance with incomplete penetrance. MVP varies in its clinical expression, ranging from only a systolic click and murmur with mild prolapse of the posterior leaflet to severe MR due to chordal rupture and leaflet flail. The degree of myxomatous change of the leaflets can also vary widely. In many patients, the condition progresses over years or decades; in others, it worsens rapidly as a result of chordal rupture or endocarditis.

Most patients are asymptomatic and remain so for their entire lives. However, in North America, MVP is now the most common cause of isolated severe MR requiring surgical treatment. Arrhythmias, most commonly ventricular premature contractions and paroxysmal supraventricular and ventricular tachycardia, as well as atrial fibrillation (AF), have been reported and may cause palpitations, light-headedness, and syncope. Sudden death is a very rare complication and occurs most often in patients with severe MR and depressed left ventricle (LV) systolic function, although it can occur in individuals with normal LV size and function. There may be an excess risk of sudden death among patients with a flail leaflet. Many patients have chest pain that is difficult to evaluate; it is often substernal, prolonged, and not related to exertion, but may rarely resemble angina pectoris. Transient cerebral ischemic attacks secondary to emboli from the mitral valve due to endothelial disruption have been reported. Infective endocarditis may occur in patients with MR and/or leaflet thickening.

Auscultation

A frequent finding is the mid- or late- (nonejection) systolic click, which occurs 0.14 s or more after S_1 and is thought to be generated by the sudden tensing of slack, elongated chordae tendineae or by the prolapsing mitral leaflet when it reaches its maximal excursion. Systolic clicks may be multiple and may be followed by a high-pitched, mid-late systolic crescendo–decrescendo murmur, which occasionally is “whooping” or “honking” and is heard best at the apex. Radiation of the murmur will depend on the involved leaflet. With posterior leaflet prolapse, the jet of MR is directed anteriorly and the murmur will radiate to the base of the heart. With anterior leaflet involvement, the jet of MR is directed posteriorly and the murmur will radiate to the axilla and back. The click and murmur occur earlier with standing, during the strain phase of the Valsalva maneuver, and with any intervention that decreases LV volume (preload), exaggerating the propensity of

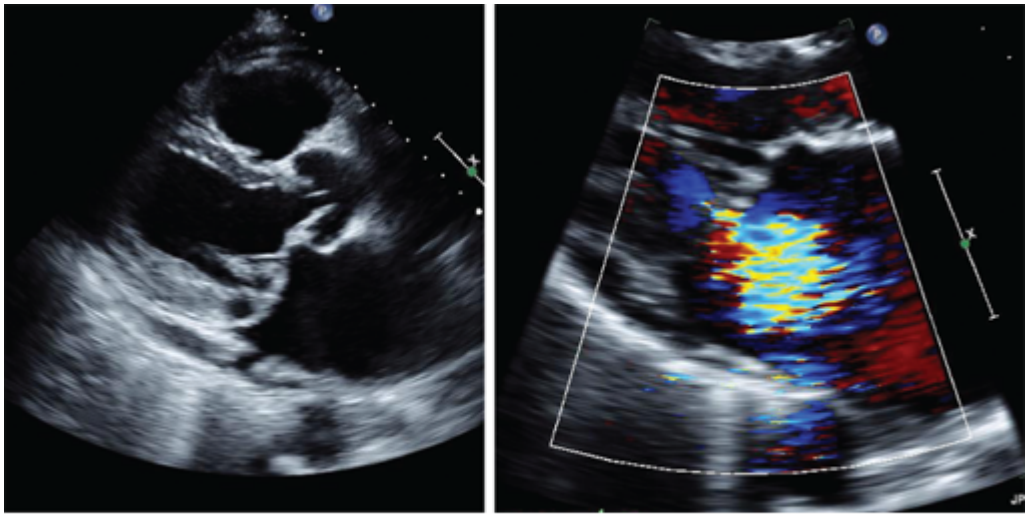
mitral leaflet prolapse. Conversely, squatting and isometric exercises, which increase LV volume, diminish MVP; the click-murmur complex is delayed, moves away from S_1, and may even disappear. Some patients have a mid-systolic click without a murmur; others have a murmur without a click. Still others have both sounds at different times.

LABORATORY EXAMINATION

The ECG most commonly is normal but may show biphasic or inverted T-waves in leads II, III, and aVF, and occasionally supraventricular or ventricular premature beats. Transthoracic echocardiography (TTE) is particularly effective in identifying the abnormal position and prolapse of the mitral valve leaflets. A useful echocardiographic definition of MVP is systolic displacement (in the parasternal long axis view) of the belly of the mitral valve leaflets by at least 2 mm into the left atrium (LA) superior to the plane of the mitral annulus. There can be prolapse of one or both leaflets **(Fig. 260-2)**. Color flow and continuous wave Doppler imaging is helpful to evaluate the associated MR and provide semiquantitative estimates of severity. The jet lesion of MR due to MVP is most often eccentric, and assessment of the effective regurgitant orifice area and regurgitant volume can be difficult. Both three-dimensional echocardiography and cardiac magnetic resonance can provide more precise determinations of LV volumes. Transesophageal echocardiography (TEE) is indicated when more accurate anatomic information is required and is performed routinely for intraoperative guidance during valve repair. Exercise testing can be performed when there is uncertainty regarding functional capacity. It is often combined with rest and immediate poststress TTE to assess LV and right ventricular (RV) function, and the dynamic nature of MR and pulmonary artery pressures. Invasive left ventriculography done at the time of right and left heart catheterization is rarely necessary but can also show prolapse of the posterior and sometimes of both mitral valve leaflets.

FIGURE 260-2.

Barlow's valve with classic mitral valve prolapse, as seen on transthoracic echocardiogram in parasternal long-axis windows. Left: parasternal long-axis window, showing both myxomatous leaflets billowing into the left atrium in late systole. Right: same window with color Doppler showing significant mitral regurgitation (*yellow jet*) in systole. *(Courtesy of Justina Wu, MD, PhD.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Mitral Valve Prolapse

Infective endocarditis prophylaxis is indicated for patients with a prior history of endocarditis. Beta blockers sometimes relieve chest pain and control palpitations. Decisions regarding anticoagulation for stroke prevention in AF should be based on the CHA2DS2-VASc score and an assessment of bleeding risk. If the patient is symptomatic from severe MR, mitral valve repair is indicated (**see Fig. 259-1**). Other indications for surgery for MVP with severe primary MR include signs of established or progressive LV systolic dysfunction, pulmonary artery hypertension, or recent onset AF. Mitral valve repair is preferred over replacement in patients with MVP or flail mitral leaflet (**see Table 258-2**); technical success is dependent not only on the anatomic findings, but also on the skill and experience of the surgeon. Repair of isolated posterior leaflet prolapse is usually straightforward, but increasingly more complex pathologies (e.g., anterior leaflet prolapse, bileaflet prolapse, Barlow's deformity) require advanced skills. Careful pre- and intraoperative TEE imaging is an important component of patient evaluation and surgical planning. Transcatheter edge-to-edge repair using a clip to grasp the anterior and posterior leaflets together can be considered for treatment of symptomatic patients at high surgical risk with severe primary MR due to MVP (**see Fig. 259-2**). Most often, the MR will be reduced in severity but not eliminated. Nevertheless, symptom status and indices of LV size and function can be improved with this approach, which is now offered at >200 specialized sites in the United States. Reported hospital mortality rates following the procedure are ~2%. Other transcatheter repair and replacement devices are not yet approved for clinical use in the United States (**see Chap. 259**).

FURTHER READING

Grover FL et al: 2016 Annual Report of the Society of Thoracic Surgeons/American College of Cardiology Transcatheter Valve Therapy Registry. J Am Coll Cardiol 69:1215, 2017. [PubMed: 27956264]

Nishimura RA et al: Mitral valve disease. Current management and future challenges. Lancet 387:1324, 2016. [PubMed: 27025438]

Nishimura RA et al: 2017 Focused Update of the 2014 AHA/ACC Guideline on Management of Patients with Valvular Heart Disease. J Am Coll Cardiol 70:252, 2017.

Chapter 261: Tricuspid Valve Disease

Patrick T. O'Gara; Joseph Loscalzo

FIGURE 261-1

TRICUSPID STENOSIS

Tricuspid stenosis (TS), which is much less prevalent than mitral stenosis (MS) in North America and Western Europe, is generally rheumatic in origin, and is more common in women than men **(Table 261-1)**. It does not occur as an isolated lesion and is usually associated with MS. Hemodynamically significant TS occurs in 5–10% of patients with severe MS; rheumatic TS is commonly associated with some degree of tricuspid regurgitation (TR). Nonrheumatic causes of TS are rare.

TABLE 261-1 Causes of Tricuspid Valve Diseases

Valve Lesion	Etiologies
Tricuspid stenosis	Rheumatic
	Congenital
Tricuspid regurgitation	Primary (organic)
	Rheumatic
	Endocarditis
	Myxomatous (TVP)
	Carcinoid
	Radiation
	Congenital (Ebstein's)
	Trauma
	Papillary muscle injury (post-MI)
	Secondary (functional)
	RV and tricuspid annular dilatation due to multiple causes of RV enlargement (e.g., long-standing pulmonary HTN, remodeling post-RV MI, left-sided valve disease, cardiomyopathy, AF)
	Chronic RV apical pacing

Abbreviations: AF, atrial fibrillation; HTN, hypertension; MI, myocardial infarction; RV, right ventricular; TVP , tricuspid valve prolapse.

PATHOPHYSIOLOGY

A diastolic pressure gradient between the right atrium (RA) and right ventricle (RV) defines TS. It is augmented when the transvalvular blood flow increases during inspiration and declines during expiration. A mean diastolic pressure gradient of 4 mmHg is usually sufficient to elevate the mean RA pressure to levels that result in systemic venous congestion. Unless sodium intake has been restricted and diuretics administered, this venous congestion is associated with hepatomegaly, ascites, and edema, sometimes severe. In patients with sinus rhythm, the RA *a* wave may be extremely tall and may even approach the level of the RV systolic pressure. The *y* descent is prolonged. The cardiac output (CO) at rest is usually depressed, and it fails to rise during exercise. The low CO is responsible for the normal or only slightly elevated left atrial (LA), pulmonary artery (PA), and RV systolic pressures despite the presence of MS. Thus, the presence of TS can mask the hemodynamic and clinical features of any associated MS.

SYMPTOMS

Because the development of MS generally precedes that of TS, many patients initially have symptoms of pulmonary congestion and fatigue. Characteristically, patients with severe TS complain of relatively little dyspnea for the degree of hepatomegaly, ascites, and edema that they have. However, fatigue secondary to a low CO and discomfort due to refractory edema, ascites, and marked hepatomegaly are common in patients with advanced TS and/or TR. In some patients, TS may be suspected for the first time when symptoms of right-sided failure persist after an adequate mitral valvotomy.

PHYSICAL FINDINGS

Because TS usually occurs in the presence of other obvious valvular disease, the diagnosis may be missed unless it is considered. Severe TS is associated with marked hepatic congestion, often resulting in cirrhosis, jaundice, serious malnutrition, anasarca, and ascites. Congestive hepatomegaly and, in cases of severe tricuspid valve disease, splenomegaly are present. The jugular veins are distended, and in patients with sinus rhythm, there may be giant *a* waves. The *v* waves are less conspicuous, and because tricuspid obstruction impedes RA emptying during diastole, there is a slow *y* descent. In patients with sinus rhythm, there may be prominent presystolic pulsations of the enlarged liver as well.

On auscultation, an opening snap (OS) of the tricuspid valve may rarely be heard ~0.06 s after pulmonic valve closure. The diastolic murmur of TS has many of the qualities of the diastolic murmur of MS, and because TS almost always occurs in the presence of MS, it may be missed. However, the tricuspid murmur is generally heard best along the left lower sternal border and over the xiphoid process, and is most prominent during presystole in patients with sinus rhythm. The murmur of TS is augmented during inspiration, and it is reduced during expiration and particularly during the strain phase of the Valsalva maneuver, when tricuspid transvalvular flow is reduced.

LABORATORY EXAMINATION

The electrocardiogram (ECG) features of RA enlargement **(see Fig. 235-8)** include tall, peaked P waves in lead II, as well as prominent, upright P waves in lead V_1. The *absence* of ECG evidence of RV hypertrophy (RVH) in a patient with right-sided heart failure who is believed to have MS should suggest associated tricuspid valve disease. The chest x-ray in patients with combined TS and MS shows particular prominence of the RA and superior vena cava without much enlargement of the PA and with less evidence of pulmonary vascular congestion than occurs in patients with isolated MS; engorgement of the azygos vein can often be appreciated. On transthoracic echocardiographic (TTE) examination, the tricuspid valve is usually thickened and domes in diastole; the transvalvular gradient can be estimated by continuous wave Doppler echocardiography. Severe TS is characterized by a valve area ≤1 cm^2 or pressure half-time of ≥190 ms. The RA and inferior vena cava (IVC) are enlarged. TTE provides additional information regarding the severity of any associated TR, mitral valve structure and function, left ventricle (LV) and RV size and function, and PA pressure. Cardiac catheterization is not routinely necessary for assessment of TS.

TREATMENT

TREATMENT Tricuspid Stenosis

Patients with TS generally exhibit marked systemic venous congestion; salt restriction, bed rest, and diuretic therapy are required during the preoperative period. Such a preparatory period may diminish hepatic congestion and thereby improve hepatic function sufficiently so that the risks of operation, particularly bleeding, are diminished. Surgical relief of the TS should be carried out, preferably at the time of surgical mitral valvotomy or mitral valve replacement (MVR) for mitral valve disease, in patients with moderate or severe TS who have mean diastolic pressure gradients exceeding ~4 mmHg and tricuspid orifice areas <1.5–2 cm^2. TS is almost always accompanied by significant TR. Operative repair may permit substantial improvement of tricuspid valve function. If repair cannot be accomplished, the tricuspid valve may have to be replaced. Meta-analysis has shown no difference in overall survival between mechanical and tissue valve replacement. Mechanical valves in the tricuspid position are more prone to thromboembolic complications than in other positions. Percutaneous tricuspid balloon valvotomy for isolated severe TS without significant TR is very rarely performed.

TRICUSPID REGURGITATION

More than 80% of TR cases encountered in clinical practice are secondary (functional) in nature and related to tricuspid annular dilatation and leaflet tethering in the setting of RV remodeling caused by pressure or volume overload (or both), myocardial infarction (MI) or trauma (Table 261-1). Secondary TR is commonly seen in the late stages of heart failure due to rheumatic or congenital heart disease with severe PA hypertension (PA systolic pressure >55 mmHg), as well as in other types of left-sided valvular (e.g., mitral regurgitation) or myocardial diseases (e.g., ischemic and idiopathic dilated cardiomyopathies). It is reversible in part if PA hypertension can be relieved. Secondary TR can also develop from chronic RV apical pacing and dyssynchronous contraction; in some patients, the RV leads may also perforate or entrap the TV leaflets. TR can often emerge in the setting of new onset atrial fibrillation (AF), particularly in elderly patients. Rheumatic fever may produce primary TR, often associated with TS. Tricuspid valve prolapse, carcinoid heart disease, endomyocardial fibrosis, radiation, infective endocarditis, and leaflet trauma can also produce primary TR. Less commonly, primary TR results from congenitally deformed tricuspid valves, and can occur with defects of the atrioventricular canal, as well as with Ebstein's malformation of the tricuspid valve **(Chap. 264)**.

PATHOPHYSIOLOGY

The incompetent tricuspid valve allows blood to flow backward from the RV into the RA, the volume of which is dependent on the driving pressure (i.e., RV systolic pressure) and the size of the regurgitant orifice. The severity and physical signs of TR can vary as a function of PA systolic pressure (in the absence of RV outflow tract stenosis), the dimension of the tricuspid valve annulus, the respiratory cycle-dependent changes in RV preload, and RA compliance. RV filling is increased during inspiration. Forward CO is reduced and does not augment with exercise. Significant degrees of TR will lead to RA enlargement and elevation of the RA and jugular venous pressures with prominent *c-v* waves in the pulse tracings. Progressively severe TR can lead to "ventricularization" of the RA wave form **(see Fig. 234-1B)**. Severe TR is also characterized by RV dilation (RV volume overload) and eventual systolic dysfunction, the progression of which can be accelerated by a concomitant pressure load from PA hypertension or by myocardial fibrosis from previous injury.

SYMPTOMS

Mild or moderate degrees of TR are usually well tolerated in the absence of other hemodynamic disturbances. Because TR most often coexists with left-sided valve lesions, LV dysfunction, and/or PA hypertension, symptoms related to these lesions may dominate the clinical picture. Fatigue and exertional dyspnea owing to reduced forward CO are early symptoms of isolated, severe TR. As the disease progresses and RV function declines, patients may report cervical pulsations, abdominal fullness/bloating, diminished appetite, and muscle wasting, although with progressive weight gain and painful swelling of the lower extremities.

PHYSICAL FINDINGS

The neck veins in patients with severe TR are distended with prominent *c-v* waves and rapid *y* descents (in the absence of TS). TR is more often diagnosed by examination of the neck veins than by auscultation of the heart sounds. Other findings may include marked hepatomegaly with systolic pulsations, ascites, pleural effusions, edema, and a positive hepatojugular reflux sign. A prominent RV pulsation along the left parasternal region and a blowing holosystolic murmur along the lower left sternal margin, which may be intensified during inspiration (Carvallo's sign) and reduced during expiration or the strain phase of the Valsalva maneuver, are characteristic findings. The murmur of TR may sometimes be confused with that of mitral regurgitation (MR) unless attention is paid to its variation during the respiratory cycle and the extent of RV enlargement is appreciated. AF is usually present in the chronic phase of the disease.

LABORATORY EXAMINATION

The ECG may show changes characteristic of the lesion responsible for the TR, e.g., an inferior Q-wave MI suggestive of a prior RV MI, RVH, or a bizarre right bundle branch block-type pattern with preexcitation in patients with Ebstein's anomaly. ECG signs of RA enlargement may be present in patients with sinus rhythm; AF is frequently noted. The chest x-ray may show RA and RV enlargement, depending on the chronicity and severity of TR. TTE is usually definitive with demonstration of RA dilation and RV volume overload and prolapsing, flail, scarred, or displaced/tethered tricuspid leaflets with annular dilatation; the diagnosis and assessment of TR can be made by color flow Doppler imaging **(see Fig. 236-8)**. Severe TR is accompanied by hepatic vein systolic flow reversal. Continuous wave Doppler of the TR velocity profile is useful in estimating PA systolic pressure, except when the TR is very severe and the jet velocity is blunted by rapidly increasing RA pressure. Accurate assessment of TR severity, PA pressures, and RV size and systolic function with TTE can be quite challenging in many patients. Real-time three-dimensional echocardiography and cardiac magnetic resonance (CMR) imaging provide alternative imaging modalities, although they are not widely available. In patients with severe TR, the CO is usually markedly reduced, and the RA pressure pulse may not exhibit an *x* descent during early systole but rather show a prominent *c*-*v* wave with a rapid *y* descent. The mean RA and RV end-diastolic pressures are often elevated. Exercise testing can be used to assess functional capacity in patients with asymptomatic severe TR. The prognostic significance of exercise-induced changes in TR severity and RV function has not been well studied.

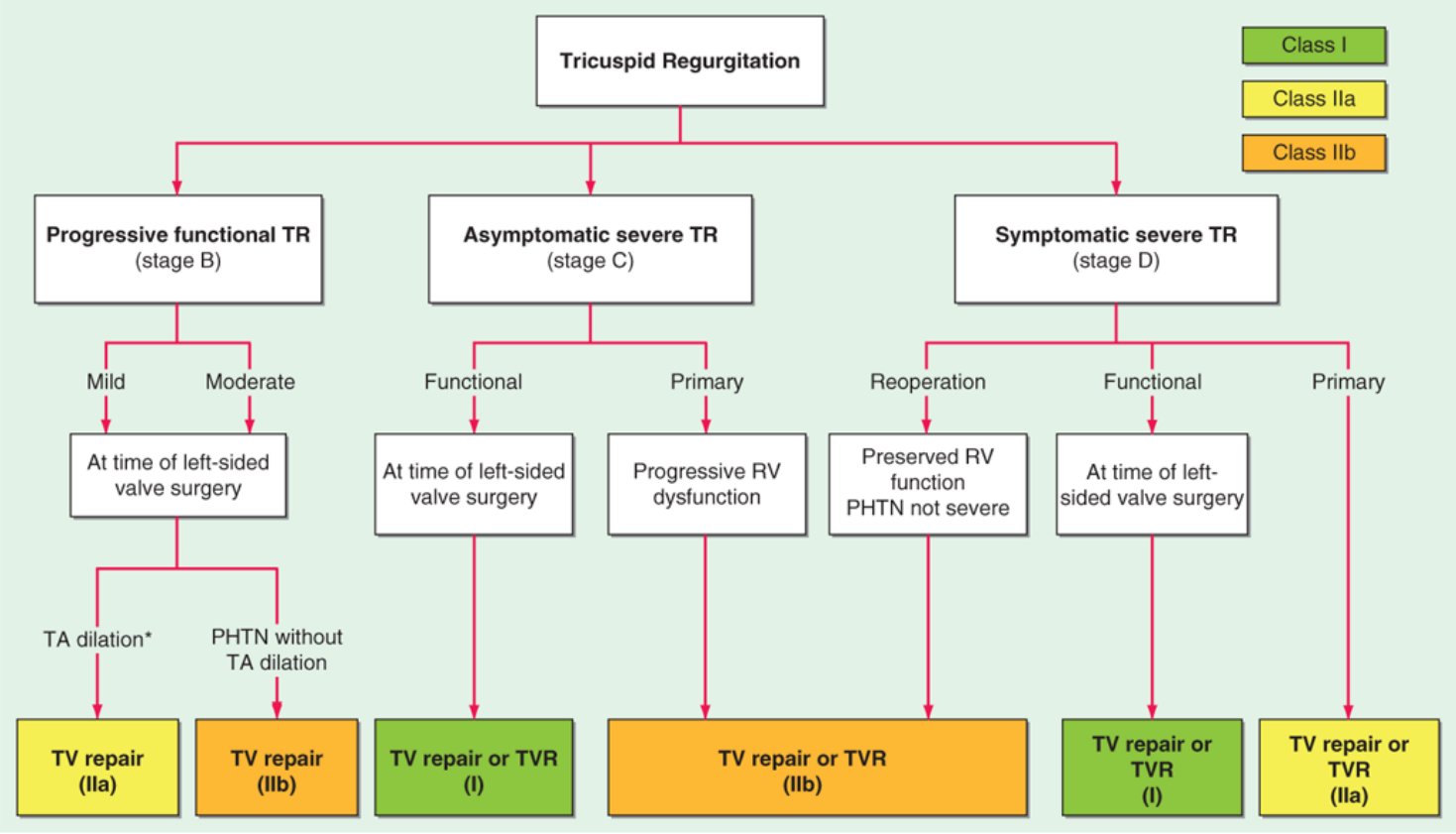
TREATMENT

TREATMENT Tricuspid Regurgitation

Diuretics can be useful for patients with severe TR and signs of right heart failure **(Fig. 261-1)**. An aldosterone antagonist may be particularly helpful because many patients have secondary hyperaldosteronism from marked hepatic congestion. Therapies to reduce elevated PA pressures and/or pulmonary vascular resistance, including those targeted at left-sided heart disease, can also be considered for patients with PA hypertension and severe secondary TR. Tricuspid valve surgery is recommended for patients with severe TR who are undergoing left-sided valve surgery and is also undertaken frequently for treatment of even moderate TR in patients undergoing left-sided valve surgery who have tricuspid annular dilation (>40 mm), a history of right heart failure, or PA hypertension. Operation most often comprises repair rather than replacement in these settings and has become routine in most major surgical centers. Surgery may also infrequently be required for treatment of severe, primary TR with right heart failure not responsive to standard medical therapy or because of progressively declining RV systolic function. Reported perioperative mortality rates for isolated tricuspid valve surgery (repair and replacement) are high (~8–9%) and likely are influenced by the hazards encountered during reoperation on patients who have undergone previous left-sided valve surgery and have reduced RV function. Indwelling pacemaker or defibrillator leads can also pose technical challenges. Investigation of transcatheter tricuspid valve repair and replacement systems is in its earliest clinical stages.

FIGURE 261-1

Management of tricuspid regurgitation. See legend for **Fig. 256-4** for explanation of treatment recommendations (Class I, IIa, IIb) and disease stages (B, C, D). Preoperative coronary angiography should be performed routinely as determined by age, symptoms, and coronary risk factors. Cardiac catheterization and angiography may also be helpful when there is a discrepancy between clinical and noninvasive findings. PHTN, pulmonary hypertension; RV, right ventricular; TA, tricuspid annular; TTE, transthoracic echocardiogram; TR, tricuspid regurgitation; TV, tricuspid valve; TVR, tricuspid valve replacement. TA dilation is defined by >40 mm on TTE (>21 mm/m^2) or >70 mm on direct intraoperative measurement. *(Adapted from RA Nishimura et al: 2014 AHA/ACC Guideline for the Management of Patients with Valvular Heart Disease. J Am Coll Cardiol 63:e57-185, 2014, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Dreyfus GD et al: Functional tricuspid regurgitation. J Am Coll Cardiol 65:2331, 2015. [PubMed: 26022823]
Hahn RT et al: Early feasibility study of transcatheter tricuspid valve annuloplasty. J Am Coll Cardiol 69:1795, 2017. [PubMed: 28385308]
Nishimura RA et al: 2014 AHA/ACC guidelines for management of patients with valvular heart disease. J Am Coll Cardiol 63:e59, 2014.
Rodés-Cabau J, Taramasso M, O'Gara PT: Diagnosis and treatment of tricuspid valve disease: Current and future perspectives. Lancet 388:2431, 2016. [PubMed: 27048553]
Rodés-Cabau J et al: Transcatheter therapies for treating tricuspid regurgitation. J Am Coll Cardiol 67:1829, 2016. [PubMed: 27081024]

Chapter 262: Pulmonic Valve Disease

Patrick T. O'Gara; Joseph Loscalzo

PULMONIC STENOSIS

Pulmonic valve stenosis (PS) is essentially a congenital disorder **(Table 262-1)**. With isolated PS, the valve is typically domed. Dysplastic pulmonic valves are seen as part of the Noonan syndrome **(Chap. 275)**, which maps to chromosome 12. Mutations in the *PTPN1* gene are associated with about half of all cases of Noonan syndrome. Much less common etiologies include carcinoid and obstructing tumors or bulky vegetations. The pulmonic valve is only very rarely affected by the rheumatic process.

TABLE 262-1 Causes of Pulmonic Valve Disease

Valve Lesion	Etiologies
Pulmonic stenosis	Congenital
	Carcinoid
	Tumor
	Endocarditis
Pulmonic regurgitation	Primary valve disease
	Congenital
	Postvalvotomy
	Endocarditis
	Carcinoid
	Annular enlargement
	Pulmonary hypertension
	Idiopathic dilation
	Marfan syndrome

PATHOPHYSIOLOGY

PS is defined hemodynamically by a systolic pressure gradient between the right ventricle (RV) and main pulmonary artery (PA). RV hypertrophy (RVH) develops as a consequence of sustained obstruction to RV outflow, and systolic ejection is prolonged. Compared with the ability of the LV to compensate for the pressure overload imposed by aortic stenosis (AS), RV dysfunction from afterload mismatch occurs earlier in the course of PS and at lower peak systolic pressures, because the RV adapts less well to this type of hemodynamic burden. With normal systolic function and cardiac output

(CO), severe PS is defined by a peak systolic gradient across the pulmonic valve of >50 mmHg; moderate PS correlates with a peak gradient of 30–50 mmHg. PS rarely progresses in patients with peak gradients <30 mmHg, but may worsen in those with moderate disease due to valve thickening and calcification with age. The RA *a* wave elevates in relation to the higher pressures needed to fill a noncompliant, hypertrophied RV. A prominent RA *v* wave signifies functional tricuspid regurgitation (TR) from RV and annular dilation. The CO is maintained until late in the course of the disease.

SYMPTOMS

Patients with mild or even moderate PS are usually asymptomatic and first come to medical attention because of a heart murmur (or early systolic click) that leads to echocardiography. With severe PS, patients may report exertional dyspnea or early-onset fatigue. Anginal chest pain from RV oxygen supply-demand mismatch and syncope may occur with very severe forms of obstruction, particularly in the presence of a destabilizing trigger such as atrial fibrillation, fever, infection, anemia, or pregnancy.

PHYSICAL FINDINGS

The murmur of mild or moderate PS is mid-systolic in timing, crescendo–decrescendo in configuration, heard best in the left second interspace, and usually introduced by an ejection sound (click) in younger adults whose valves are still pliable. The ejection sound is the only right-sided acoustic event that decreases in intensity with inspiration. This phenomenon reflects premature opening of the pulmonic valve by the elevated RV end-diastolic (postatrial *a* wave) pressure. The systolic murmur increases in intensity during inspiration. With progressively severe PS, the ejection sound moves closer to the first heart sound and eventually becomes inaudible. A right-sided fourth heart sound may emerge. The systolic murmur peaks later and may persist through the aortic component of the second heart sound (A_2). Pulmonic valve closure is delayed, and the pulmonic component of the second heart sound (P_2) is reduced or absent. A prominent *a* wave, indicative of the higher atrial pressure necessary to fill the noncompliant RV, may be seen in the jugular venous pulse. A parasternal or RV lift can be felt with significant pressure overload. Signs of right heart failure, such as hepatomegaly, ascites, and edema, are uncommon but may appear very late in the disease.

LABORATORY EXAMINATION

The electrocardiogram (ECG) will show right axis deviation, RVH, and RA enlargement in adult patients with severe PS. Chest x-ray findings include poststenotic dilation of the main PA in the frontal plane projection and filling of the retrosternal airspace due to RV enlargement on the lateral film. In some patients with RVH, the cardiac apex appears to be lifted off the left hemidiaphragm. The RA may also be enlarged. Transthoracic echocardiography (TTE) allows definitive diagnosis and characterization in most cases, with depiction of the valve and assessment of the gradient, RV function, PA pressures (which should be low), and any associated cardiac lesions. Transesophageal echocardiography (TEE) may be useful in some patients for improved delineation of the RV outflow tract (RVOT) and assessment of infundibular hypertrophy. Cardiac catheterization is not usually necessary for diagnostic purposes, but if performed, pressures should be obtained from just below and above the pulmonic valve with attention to the possibility that a dynamic component to the gradient may exist. The correlation between Doppler assessment of peak instantaneous gradient and catheterization-measured peak-to-peak gradient is weak. The latter may correlate better with the Doppler mean gradient.

TREATMENT

TREATMENT Pulmonic Stenosis

Diuretics can be used to treat symptoms and signs of right heart failure. Provided there is less than moderate pulmonic regurgitation (PR), percutaneous pulmonic balloon valvotomy is recommended for symptomatic patients with a domed valve and a peak gradient >50 mmHg (or mean gradient >30 mmHg) and for asymptomatic patients with a peak gradient >60 mmHg (or mean gradient >40 mmHg). Surgery may be required when the valve is dysplastic (as seen in patients with Noonan's syndrome and other disorders). A multidisciplinary heart team is best positioned to make treatment decisions of this nature.

PULMONIC REGURGITATION

PR may develop as a consequence of primary valve pathology, annular enlargement, or their combination; after surgical treatment of RVOT obstruction in children with such disorders as tetralogy of Fallot; or after percutaneous pulmonic balloon valvotomy (Table 262-1). Carcinoid usually causes mixed pulmonic valve disease with PR and PS. Long-standing severe PA hypertension from any cause can result in dilation of the pulmonic valve ring and PR.

PATHOPHYSIOLOGY

Severe PR results in RV chamber enlargement and eccentric hypertrophy. As is the case for aortic regurgitation (AR), PR is a state of increased preload and afterload. The reverse pressure gradient from the PA to the RV, which drives the PR, progressively decreases throughout diastole and accounts for the decrescendo nature of the diastolic murmur. As RV diastolic pressure increases, the murmur becomes shorter in duration. The forward CO is preserved during the early stages of the disease, but may not increase normally with exercise and declines over time. A reduction in RV ejection fraction may be an early indicator of hemodynamic compromise. In advanced stages, there is significant enlargement of the RV and RA with marked elevation of the jugular venous pressure.

SYMPTOMS

Mild or moderate degrees of PR do not, by themselves, result in symptoms. Other problems, such as PA hypertension, may dominant the clinical picture. With progressively severe PR and RV dysfunction, fatigue, exertional dyspnea, abdominal fullness/bloating, and lower extremity swelling may be reported.

PHYSICAL FINDINGS

The physical examination hallmark of PR is a high-pitched, decrescendo diastolic murmur (Graham Steell murmur) heard along the left sternal border that can be difficult to distinguish from the more frequently appreciated murmur of AR. The Graham Steell murmur may become louder with inspiration and is usually associated with a loud and sometimes palpable P_2 and an RV lift, as would be expected in patients with significant PA hypertension of any cause. Survivors of childhood surgery for tetralogy of Fallot or PS/pulmonary atresia may have an RV-PA conduit that is freely regurgitant because it does not contain a valve. PA pressures in these individuals are not elevated and the diastolic murmur can be misleadingly low pitched and of short duration despite significant degrees of PR and RV volume overload.

LABORATORY EXAMINATION

Depending on both the etiology and severity of PR, the ECG may show findings of RVH and RA enlargement. On chest x-ray, the RV and RA may be enlarged. Pulmonic valve morphology and function can be assessed with transthoracic Doppler echocardiography. PA pressures can be estimated from the tricuspid valve systolic jet velocity. Cardiac magnetic resonance (CMR) provides

greater anatomic detail, particularly in patients with repaired congenital heart disease, and more precise assessment of RV volumes. Cardiac catheterization is not routinely necessary but would be performed as part of a planned transcatheter procedure.

TREATMENT

TREATMENT Pulmonic Regurgitation

In patients with functional PR due to PA hypertension and annular dilation, efforts to reduce PA vascular resistance and pressure should be optimized. Such efforts may include pharmacologic/vasodilator and/or surgical/interventional strategies, depending on the cause of the PA hypertension (e.g., idiopathic PA hypertension, left-sided heart valve disease). Diuretics can be used to treat the manifestations of right heart failure. Surgical valve replacement for primary, severe, pulmonic valve disease, such as carcinoid or endocarditis, is rarely undertaken. Transcatheter pulmonic valve replacement has been successfully performed in many patients with severe PR after childhood repair of tetralogy of Fallot or pulmonic valve stenosis or atresia. This procedure was introduced clinically prior to transcatheter aortic valve replacement.

FURTHER READING

Ansari MM et al: Percutaneous pulmonary valve implantation. J Am Coll Cardiol 66:2246, 2015. [PubMed: 26564602]
Baumgartner H et al: ESC guidelines for the management of grown-up congenital heart disease. Eur Heart J 31:2915, 2010. [PubMed: 20801927]
Warnes CA et al: 2008 ACC/AHA guidelines for the management of adults with congenital heart disease. J Am Coll Cardiol 52:e143, 2008. [PubMed: 19038677]

Chapter 263: Multiple and Mixed Valvular Heart Disease

Patrick T. O'Gara; Joseph Loscalzo

INTRODUCTION

Many acquired and congenital cardiac lesions may result in stenosis and/or regurgitation of one or more heart valves. For example, rheumatic heart disease can involve the mitral (mitral stenosis [MS], mitral regurgitation [MR], or MS and MR), aortic (aortic stenosis [AS], aortic regurgitation [AR], or AS and AR), and tricuspid (tricuspid stenosis [TS], tricuspid regurgitation [TR], or TS and TR) valve, alone or in combination. The common association of functional TR with significant mitral valve disease is discussed in **Chap. 261**. Severe mitral annular calcification can result in regurgitation (due to decreased annular shortening during systole) and mild or moderate stenosis (caused by extension of the calcification onto the leaflets resulting in restricted valve opening). Patients with severe AS and LV remodeling may develop functional MR that may not improve after isolated aortic valve replacement (AVR). Chordal rupture has been described infrequently in patients with severe AS. Aortic valve infective endocarditis (IE) may secondarily involve the mitral apparatus either by abscess formation and contiguous spread via the intervalvular fibrosa or by "drop metastases" from the aortic leaflets onto the anterior leaflet of the mitral valve. Mediastinal radiation may result in aortic, mitral, and even tricuspid valve disease, most often with mixed stenosis and regurgitation. Carcinoid heart disease may cause mixed lesions of either or both the tricuspid and pulmonic valves. Ergotamines, and the previously used combination of fenfluramine and phentermine, can rarely result in mixed lesions of the aortic and/or mitral valve. Patients with Marfan syndrome may have both AR from aortic root dilation and MR due to mitral valve prolapse (MVP). Myxomatous degeneration causing prolapse of multiple valves (mitral, aortic, tricuspid) can also occur in the absence of an identifiable connective tissue disorder. Bicuspid aortic or pulmonic valve disease can result in mixed stenosis and regurgitation. The former is also associated with aortic aneurysm disease and a predisposition to aortic dissection.

PATHOPHYSIOLOGY

In patients with multivalvular heart disease, the pathophysiologic derangements associated with the more proximal valve disease can mask the full expression of the attributes of the more distal valve lesion. For example, in patients with rheumatic mitral and aortic valve disease, the reduction in cardiac output (CO) imposed by the mitral valve disease will decrease the magnitude of the hemodynamic derangements related to the severity of the aortic valve lesion (stenotic, regurgitant, or both). Alternatively, the development of atrial fibrillation (AF) during the course of MS can lead to sudden worsening in a patient whose aortic valve disease was not previously felt to be significant. The development of reactive pulmonary vascular disease, sometimes referred to as a "secondary obstructive lesion in series," can impose an additional challenge in these settings. As CO falls with progressive tricuspid valve disease, the severity of any associated mitral or aortic disease can be underestimated.

One of the most common examples of multivalve disease is that of functional TR in the setting of significant mitral valve disease. Functional TR occurs as a consequence of right ventricular and

annular dilation; pulmonary artery (PA) hypertension is often present. The tricuspid leaflets are morphologically normal. Progressive degrees of TR lead to right ventricular volume overload and continued chamber and annular dilation. The TR is usually central in origin; reflux into the right atrium (RA) is expressed as large, systolic *c-v* waves in the RA pressure pulse. The height of the *c-v* wave is dependent on RA compliance and the volume of regurgitant flow. The RA waveform may become "ventricularized" in advanced stages of chronic, severe TR with PA hypertension. CO falls and the severity of the associated mitral valve disease may become more difficult to appreciate. Findings related to advanced right heart failure (e.g., ascites, edema) predominate. Primary rheumatic tricuspid valve disease may occur with rheumatic mitral disease and cause hemodynamic changes reflective of TR, TS, or their combination. With TS, the *y* descent in the RA pressure pulse is prolonged.

Another example of rheumatic, multivalve disease involves the combination of mitral and aortic valve pathology, frequently characterized by MS and AR. In isolated MS, left ventricular (LV) preload and diastolic pressure are reduced as a function of the severity of inflow obstruction. With concomitant AR, however, LV filling is enhanced and diastolic pressure may rise depending on the compliance characteristics of the chamber. Because the CO falls with progressive degrees of MS, transaortic valve flows will decline, masking the potential severity of the aortic valve lesion (AR, AS, or its combination). As noted above, onset of AF in such patients can be especially deleterious. The loss of atrial systole with AF may result in a critical reduction in CO, a rise in LA and LV diastolic pressures, and a further deleterious increase in heart rate.

Secondary (functional) MR may complicate the course of some patients with severe AS. The mitral valve leaflets and chordae tendineae are usually normal. Incompetence is related to changes in LV geometry (remodeling) and abnormal systolic tethering of the leaflets in the context of markedly elevated LV systolic pressures. Relief of the excess afterload with surgical or transcatheter AVR often, but not always, results in reduction or elimination of the MR. Persistence of significant MR following AVR is associated with impaired functional outcomes and reduced survival. Identification of patients who would benefit from concomitant treatment of their functional MR at time of AVR is quite challenging. Most surgeons advocate for repair of moderate-to-severe or severe functional MR at time of surgical AVR. Significant primary MR is routinely managed with repair or replacement at the time of AVR.

In patients with mixed AS and AR, assessment of valve stenosis can be influenced by the magnitude of the regurgitant valve flow. Because transvalvular systolic flow velocities are augmented in patients with AR and preserved LV systolic function, the LV-aortic Doppler-derived pressure gradient and the intensity of the systolic murmur will be elevated to values higher than expected for the true systolic valve orifice size as delineated by planimetry. Uncorrected, the Gorlin formula, which relies on forward CO (systolic transvalvular flow) and the mean pressure gradient for calculation of valve area, is not accurate in the setting of mixed aortic valve disease. Similar considerations apply to patients with mixed mitral valve disease. The peak mitral valve Doppler E wave velocity (v_0) is increased in the setting of severe MR because of enhanced early diastolic flow and may not accurately reflect the contribution to left atrial (LA) hypertension from any associated MS. When either AR or MR is the dominant lesion in patients with mixed aortic or mitral valve disease, respectively, the LV is dilated. When AS or MS predominates, LV chamber size will be normal or small. It can sometimes be difficult to ascertain whether stenosis or regurgitation is the dominant lesion in patients with mixed valve disease, although an integrated clinical and noninvasive assessment can usually provide clarification for purposes of patient management and follow-up.

Patients with significant AS, a nondilated LV chamber, and concentric hypertrophy will poorly tolerate the abrupt development of aortic regurgitation, as may occur, for example, with IE or after surgical or transcatheter AVR (TAVR) complicated by paravalvular leakage. The noncompliant LV is not prepared to accommodate the sudden volume load, and as a result, LV diastolic pressure rises rapidly and severe heart failure develops. Indeed, paravalvular regurgitation is a significant risk factor for short- to intermediate-term death following transcatheter AVR. Conditions in which the LV may not be able to

dilate in response to chronic AR (or MR) include radiation heart disease and, in some patients, the cardiomyopathy associated with obesity and diabetes. Noncompliant ventricles of small chamber size predispose to earlier onset diastolic dysfunction and heart failure in response to any further perturbation in valve function.

SYMPTOMS

Compared with patients with isolated, single-lesion valve disease, patients with multiple or mixed valve disease may develop symptoms at a relatively earlier stage in the natural history of their disease. Symptoms such as exertional dyspnea and fatigue are usually related to elevated filling pressures, reduced CO, or their combination. Palpitations may signify AF and identify mitral valve disease as an important component of the clinical presentation, even when not previously suspected. Chest pain compatible with angina could reflect left or right ventricular oxygen supply/demand mismatch on a substrate of hypertrophy and pressure/volume overload with or without superimposed coronary artery disease. Symptoms related to right heart failure (abdominal fullness/bloating, edema) are late manifestations of advanced disease.

PHYSICAL FINDINGS

Mixed disease of a single valve is most often manifested by systolic and diastolic murmurs, each with the attributes expected for the valve in question. Thus, patients with AS and AR will have characteristic mid-systolic, crescendo-decrescendo and blowing, decrescendo diastolic murmurs at the base of the heart in the second right interspace and along the left sternal edge, respectively. Many patients with significant AR have mid-systolic outflow murmurs even in the absence of valve sclerosis/stenosis, and other findings of AS must be sought. The separate murmurs of AS and AR can occasionally be difficult to distinguish from the continuous murmurs associated with either a patent ductus arteriosus (PDA) or ruptured sinus of Valsalva aneurysm. With mixed aortic valve disease, the systolic murmur should end before, and not envelope or extend through, the second heart sound (S_2). The murmur associated with a PDA is heard best to the left of the upper sternum. The continuous murmur heard with a ruptured sinus of Valsalva aneurysm is often first appreciated after an episode of acute chest pain. An early ejection click, which usually defines bicuspid aortic valve disease in young adults, is often not present in patients with congenital, mixed AS and AR. As noted above, both the intensity and duration of these separate murmurs can be influenced by a reduction in CO and transvalvular flow due to coexistent mitral valve disease. In patients with isolated MS and MR, expected findings would include a blowing, holosystolic murmur and a mid-diastolic rumble (with or without an opening snap) best heard at the cardiac apex. An irregularly irregular heart rhythm in such patients would likely signify AF. Findings with TS and TR would mimic those of left-sided MS and MR, save for the expected changes in the murmurs with respiration. The murmurs of pulmonic stenosis and regurgitation behave in a fashion directionally similar to AS and AR; dynamic changes during respiration should be noted. Specific attributes of these cardiac murmurs are reviewed in **Chaps. 38 and 261**.

LABORATORY EXAMINATION

The electrocardiogram (ECG) may show evidence of ventricular hypertrophy and/or atrial enlargement. ECG signs indicative of right-sided cardiac abnormalities in patients with left-sided valve lesions should prompt additional assessment for PA hypertension and/or right-sided valve disease. The presence of AF in patients with aortic valve disease may be a clue to the presence of previously unsuspected mitral valve disease in the appropriate context. The chest x-ray can be reviewed for evidence of cardiac chamber enlargement, valve and/or annular calcification, and any abnormalities in the appearance of the pulmonary vasculature. The latter could include enlargement of the main and

proximal pulmonary arteries with PA hypertension and pulmonary venous redistribution/engorgement or Kerley B lines with increasing degrees of LA hypertension. An enlarged azygos vein in the frontal projection indicates RA hypertension. Roentgenographic findings not expected based on a single or mixed valve lesion may reflect other valve disease.

Transthoracic echocardiography (TTE) is the most commonly used imaging modality for the diagnosis and characterization of multiple and/or mixed valvular heart disease and may often demonstrate findings not clinically suspected. Transesophageal echocardiography (TEE) may sometimes be required for more accurate assessment of valve anatomy (specifically, the mitral valve) and when IE is considered responsible for the clinical presentation. TTE findings of particular interest include those related to valve morphology and function, calcification, chamber size, ventricular wall thickness, estimated PA systolic pressure, and the dimensions of the great vessels, including the root and ascending aorta, PA, and inferior vena cava. Exercise testing (with or without echocardiography) can be useful when the degree of functional limitation reported by the patient is not adequately explained by the findings on TTE performed at rest. An integrated assessment of the clinical and TTE findings is needed to help determine the dominant valve lesion(s) and establish an appropriate plan for treatment and follow-up. Natural history is usually influenced to a relatively greater degree by the dominant lesion.

Cardiac magnetic resonance (CMR) can be used to provide additional anatomic and physiologic information when echocardiography proves suboptimal, but is less well suited to the evaluation of valve morphology. Cardiac computed tomography (CT) has been used to assess intracardiac structures in patients with complicated IE. It is invaluable in planning for transcatheter valve replacement. Coronary CT angiography provides a noninvasive alternative for the assessment of coronary artery anatomy prior to surgery or transcatheter intervention.

Invasive hemodynamic evaluation with right and left heart catheterization may be required to characterize more completely the individual contributions of each lesion in patients with either multiple or mixed valvular heart disease. It is strongly recommended when there is a discrepancy between the clinical and non-invasive findings in a symptomatic patient. Measurement of PA pressures and calculation of pulmonary vascular resistance (PVR) can help inform clinical decision-making in certain patient subsets, such as those with advanced mitral and tricuspid valve disease. It is important to identify any potential contribution to the clinical picture from pulmonary vascular disease. Attention to the accurate assessment of CO is essential. Coronary angiography (if indicated) can be performed as part of the procedure. Contrast ventriculography and great vessel angiography are performed infrequently.

TREATMENT

TREATMENT Multiple and Mixed Valve Disease

Management of patients with multiple or mixed valve disease can be challenging. As noted above, it is helpful to determine the dominant valve lesion and proceed according to the treatment and follow-up recommendations for it **(Chaps. 256,257,258,259,260,261,262)**, being mindful of deviations from the expected course due to the contributions of more than one valve lesion. For example, AF that emerges in the course of moderate mitral valve disease may precipitate heart failure in patients with concomitant, severe aortic valve disease that was previously asymptomatic.

Medical therapies are limited and include diuretics when indicated for relief of congestion and anticoagulation to prevent stroke and thromboembolism in patients with AF. Blood pressure–lowering medications may be needed to treat systemic hypertension, which may aggravate left-sided regurgitant valve lesions, but should be initiated and titrated carefully. Pulmonary vasodilators to lower PVR are not generally effective in this context.

There is a paucity of evidence to inform practice guidelines for surgical and/or transcatheter valve intervention in patients with multiple or mixed valve disease. When there is a clear, dominant lesion, as for example in a patient with severe AS and mild AR, indications for intervention are straightforward and follow those recommended for patients with AS **(Chap. 256)**. In other patients, however, there is less clarity, and decisions regarding intervention should be based on several considerations, including those related to lesion severity, ventricular remodeling, functional capacity, and PA pressures. In this regard, it is important to realize that patients with multiple and/or mixed valve disease may develop limiting symptoms or signs of physiologic impairment even with moderate valve lesions.

Concomitant aortic and mitral valve replacement surgery is associated with a significantly higher perioperative mortality risk than replacement of either valve alone, and operation should be carefully considered. Double valve replacement surgery is usually performed for treatment of severe (unrepairable) valve disease at both locations and for the combination of severe disease at one location with moderate disease at the other, so as to avoid the hazards of reoperation in the intermediate to late term for progressive disease of the unoperated valve. In addition, the presence of a prosthesis in the aortic position significantly restricts surgical exposure of the native mitral valve. The need for double valve replacement may also impact the decision regarding the type of prosthesis (i.e., mechanical vs tissue).

Tricuspid valve repair for moderate or severe secondary (functional) TR at the time of left-sided valve surgery is now commonplace, particularly if there is dilation of the tricuspid annulus (>40 mm). The addition of tricuspid valve repair, consisting usually of insertion of an annuloplasty ring, adds little time or complexity to the procedure and is well tolerated. Reoperation for repair (or replacement) of progressive TR years after initial surgery for left-sided valve disease, on the other hand, is associated with a relatively high perioperative mortality risk. Repair of moderate or severe functional MR at time of AVR for AS can usually be undertaken with acceptable risk for perioperative death or major complication.

The presence of moderate or severe MR in patients with rheumatic MS is a contraindication to percutaneous mitral balloon valvotomy (PMBV). Likewise, the presence of significant AR in patients with AS disqualifies them from percutaneous aortic balloon valvotomy (PABV). TAVR is generally not undertaken for patients with severe, mixed AS and AR. Transcatheter management of both severe AS (with TAVR) and functional MR (with deployment of an edge-to-edge clip) has been reported. Further advances in transcatheter treatments for multiple and mixed valve disease are anticipated.

FURTHER READING

Bolling SF: Tricuspid regurgitation after left heart surgery. J Am Coll Cardiol 64:2643, 2014. [PubMed: 25524344]
Egbe AC et al: Outcomes in mixed aortic valve disease. J Am Coll Cardiol 167:2321, 2016.
Magne J et al: Pulmonary hypertension in valvular disease. JACC: Cardiovasc Imaging 8:83, 2015. [PubMed: 25592699]

Chapter 264: Congenital Heart Disease in the Adult

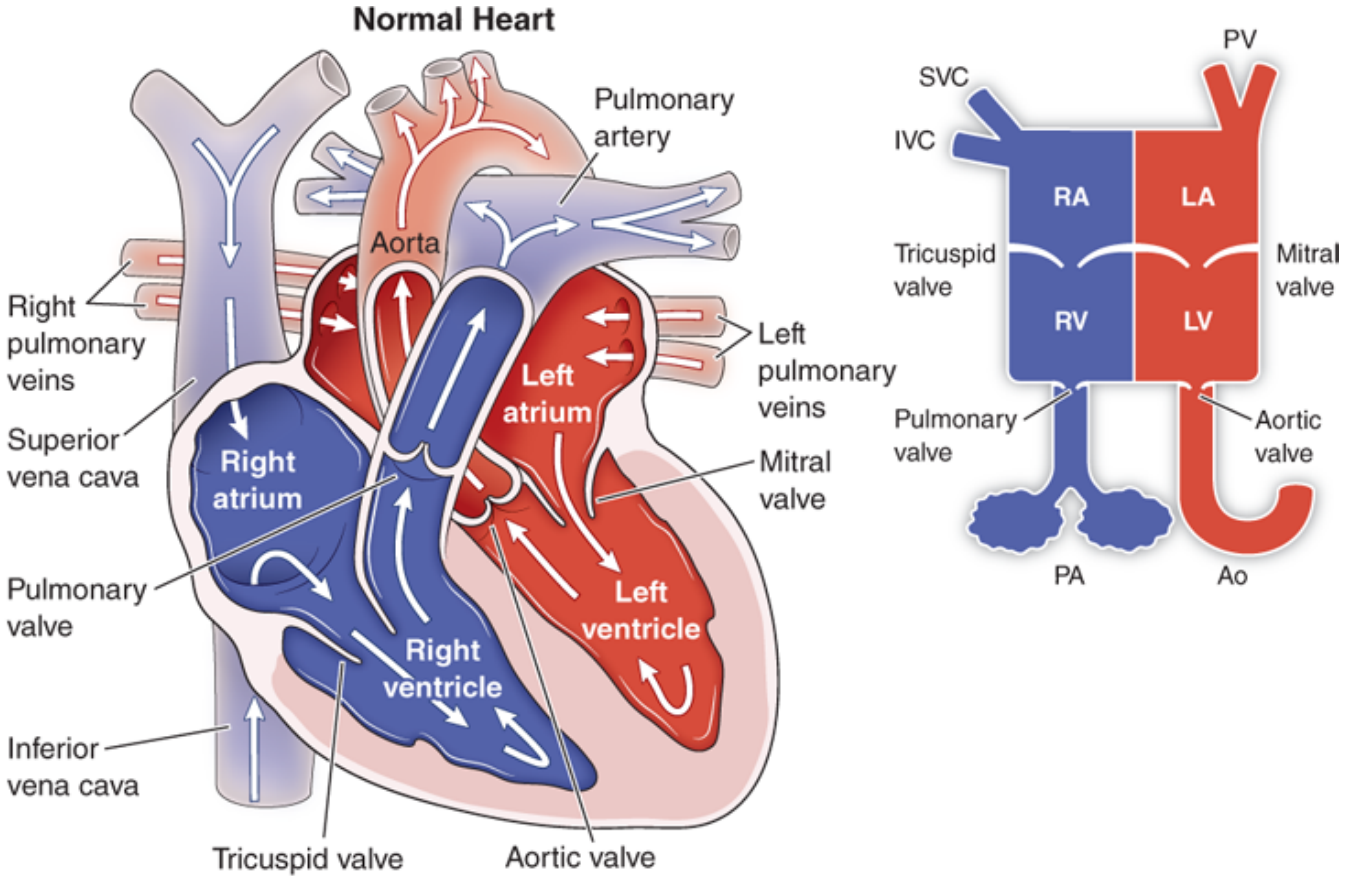
Anne Marie Valente; Michael J. Landzberg

PREVALENCE

The number of adults with congenital heart disease (CHD) living in the United States is estimated to be at least 1.4 million, with just over one in five having a complex form of CHD. The majority of adults with CHD were diagnosed in childhood, although a substantial percentage may have CHD first recognized as adults. Lifelong follow-up in coordination with, or directly by, clinicians with expertise in adult congenital heart disease (ACHD) is recommended. In this chapter, we will review the current field of ACHD, with an introduction to CHD nomenclature and cardiac development. This is followed by a summary of the more common CHD lesions that may be diagnosed in adulthood. Lastly, some of the common repaired CHD lesions that are encountered in adults are discussed. Throughout the chapter, to aid in the understanding of congenital cardiac anatomy and physiology, we include figures displaying the passage of blood flow between blood vessels and cardiac chambers in various disorders **(Fig. 264-1)**.

Figure 264-1

Normal heart. Understanding of congenital cardiac anatomy and physiology is facilitated by use of box diagrams, displaying passage of blood flow between blood vessels and cardiac chambers. Labeling (e.g., structure names, arrows to denote direction of flow, coloring to represent oxygen saturation, connections or obstructions, chamber or vascular pressures, oxygen saturations) can aid in representation. Ao, aorta; IVC, inferior vena cava; LA, left atrium; LV, left ventricle; PA, pulmonary artery; PV, pulmonary veins; RA, right atrium; RV, right ventricle; SVC, superior vena cava.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE CHANGING LANDSCAPE OF ADULT CHD

A Relatively New Subspecialty in Cardiovascular Disease

Over the past decade, the field of caring for adults with CHD (ACHD) has blossomed, and several nationwide initiatives have been initiated in an attempt to standardize care. The American College of Cardiology and American Heart Association developed guidelines for the care of adults with CHD, first published in 2008. These guidelines emphasize the need for collaboration among primary care practitioners, cardiologists, and ACHD subspecialty cardiologists. The body of medical knowledge and competencies attendant with ACHD combined with skill acquisition in coordination of complex care over a patient's medical lifetime led in 2015 to both ACHD board certification examinations by the American Board of Medical Subspecialties, as well as the establishment of requirements for 2-year subspecialty fellowship training in ACHD care, by the Accreditation Council for Graduate Medical Education. In temporal association, the Adult Congenital Heart Association (ACHA) developed a process for ACHD care program accreditation based upon standardization of infrastructural components felt requisite to achieve quality outcomes for ACHD.

SPECIAL CONSIDERATIONS FOR THE ACHD PATIENT

Adults with CHD may not recognize subtle changes in their exercise capacity, some of which are associated with worse survival; by the time symptoms are recognized, irreversible physiological changes may have occurred. ACHD patients are, therefore, advised to undergo regular evaluations for surveillance of anatomic, hemodynamic, and electrophysiologic sequelae that may be present. In addition, specific situations may arise in which it is prudent to review care in consultation with an ACHD specialist, several of which are outlined below.

Non-Cardiac Surgery

Nearly all adults with CHD can be classified with stage A (harboring risk) or greater degrees of heart failure. As such, adults with CHD may demonstrate limited hemodynamic reserve to altered myocardial perfusion or loading conditions, and may have subclinical organ dysfunction that is not recognized by standard laboratory assessment. Comprehensive, multi-specialty assessment and care strategy review are recommended in advance of invasive or operative procedures for adults with CHD. **Table 264-1** lists the multi-organ considerations that should be taken into account in adults with CHD during perioperative resuscitation and convalescence. Anesthetic management requires particular knowledge of anatomy, physiologic consequence of underlying defects, myocardial and vascular performance, presence and nature of previous palliative procedures and residual shunts, alteration of venous or arterial pathways within the circulation, and status of non-cardiovascular organ physiology.

TABLE 264-1 Multi-organ Considerations in ACHD Patients

Neurologic	Increased incidence of occult or clinically evident strokes
	Decreased level of executive functioning skills
	Anxiety, post-traumatic stress disorder, depression
	Psychosocial disorders
Lungs	Restrictive lung disease
	Pulmonary vascular disease
Renal	Decreased perfusion
Hepatic	Liver fibrosis
Peripheral Vasculature	Increased chronic venous insufficiency
Lymphatic	Impaired reabsorption
Orthopedic	Scoliosis
	Kyphosis
Hematologic	Anemia
	Coagulopathies

Pregnancy

Women with CHD should receive counseling regarding both maternal and fetal risks prior to conceiving a pregnancy and should be cared for in institutions with experience in treating CHD during pregnancy. Preconception evaluation includes detailed medical history, with a focus on the women's functional capacity, which is closely linked to maternal and fetal outcomes. **Table 264-2** lists the World Health Organization Classification of risk during pregnancy in women with heart disease; women at risk should be strongly counseled about the significant risks of morbidity and mortality during pregnancy and the postpartum period. Normal physiological hemodynamic changes of pregnancy are significant, occur over a relatively condensed period of time, and may be compounded in adults with CHD. Women with certain forms of CHD, particularly those complicated by elevated pulmonary artery (PA) pressures, decreased ventricular function, or symptomatic left-sided obstructive lesions, may not tolerate these dramatic changes.

TABLE 264-2 Modified WHO Classification of Heart Disease in Pregnancy

WHO I

- Uncomplicated, small or mild pulmonary stenosis, patent ductus arteriosus, mitral valve prolapse
- Successfully repaired simple lesions (atrial or ventricular septal defect, patent ductus arteriosus, partially anomalous pulmonary venous drainage).
- Isolated atrial or ventricular ectopic beats

WHO II (if otherwise well and uncomplicated)

- Unoperated atrial or ventricular septal defect
- Repaired tetralogy of Fallot
- Most arrhythmias

WHO II-III (depending on individual)

- Mild left ventricular impairment
- Hypertrophic cardiomyopathy
- Native or tissue valvular heart disease not considered WHO I or IV
- Marfan syndrome without aortic dissection
- Aorta < 45 mm in bicuspid aortic valve
- Repaired coarctation

WHO III

- Mechanical valve
- Systemic right ventricle
- Fontan circulation
- Cyanotic heart disease (unrepaired)
- Other complex congenital heart disease
- Aortic dilation 40–45 mm in Marfan syndrome
- Aortic dilation 45–50 mm in bicuspid aortic valve

WHO IV (pregnancy contraindicated)

- Pulmonary arterial hypertension
- Severe systemic ventricular dysfunction (LVEF <30%, NYHA class > II)
- Previous peripartum cardiomyopathy with any residual impairment of left ventricular function
- Severe mitral stenosis, severe symptomatic aortic stenosis
- Marfan syndrome with aorta dilated >45 mm
- Aortic dilation >50 mm in bicuspid aortic valve
- Native severe aortic coarctation

Pre-pregnancy medications should be reviewed to ensure their safety in pregnancy. Alternatives to angiotensin converting enzyme (ACE) inhibitors, angiotensin receptor blockers, and endothelin-receptor blockers should be considered, as these agents are teratogenic and contraindicated during pregnancy, and should be discontinued. Traditionally, the FDA utilized a classification system of five categories (A, B, C, D, X) to indicate the teratogenic potential of a drug; in December of 2014, the FDA promulgated the Pregnancy and Lactation Labeling Rule, requiring removal of these categories from all human prescription drugs and biological products, and replacement with three comprehensive subsections that provide details about the use of the drug in pregnancy and lactation in women and men of reproductive potential. These pregnancy subsections include potential risks to the developing fetus, known dosing alterations in pregnancy, effects of timing and duration of exposure during pregnancy, adverse maternal reactions, effects of the drug on labor and delivery, and information on pregnancy exposure registry for the drug, if such exists. Women requiring anticoagulation must be advised of the challenges of managing anticoagulation during pregnancy and individualized strategies should be developed. A fetal echocardiogram between 18 and 22 weeks of gestation is advised for parents with CHD. Additionally, both men and women with CHD should be counseled regarding the risk of CHD in their offspring.

CONGENITAL TERMINOLOGY, DEVELOPMENT, AND GENETICS

Congenital Nomenclature

One of the challenges in caring for adults with CHD is the inconsistent terminology used to describe the congenital heart lesions. Several classification systems have been proposed, from the initial descriptions by Maude Abbott, Maurice Lev, and Jesse Edwards, to the extensive characterizations by Stella and Richard Van Praagh and Robert Anderson. In this chapter, we follow a segmental approach. The heart is composed of several segments that are analyzed separately before formulating a comprehensive diagnosis. The principal segments are the atria, the ventricles, and the great arteries, which are joined together by the atrioventricular canal and the conus (infundibulum). In the normal heart, the right ventricle (RV) is right-sided and organized inflow-to-outflow from right to left, while the left ventricle (LV) is left-sided and organized inflow-to-outflow from left to right. It is important to determine the segmental alignments: that is, what drains into what. For example, in the normal heart the right atrium (RA) is aligned with the RV and the LV with the aorta. Finally, the segmental connections, the way in which adjacent segments are physically linked to each other, are described. For example, in the normal heart the PA is connected to the RV by a complete muscular conus (infundibulum), while the aorta is connected to the LV by aortic-mitral fibrous continuity (without a complete conus). Alignment and connection are different concepts and both are important, especially in complex defects.

Cardiac Development

The heart starts to form in the third week of gestation, and is nearly fully formed by 8 weeks' gestation. Mesodermal precardiac cells migrate to form the cardiac crescents (primary heart fields) in anterior lateral plate mesoderm, which are then brought together to form a primary linear heart tube by ventral closure of the embryo. Cells of the second heart field continue to proliferate outside the heart and are added to the heart tube over the course of embryogenesis, contributing to the atria, the RV, and outflow tract. Additionally, cardiac neural crest cells migrate into the developing heart in the 5th–6th weeks and are essential for septation of the outflow, formation of the semilunar valves, and patterning of the aortic arches. Once formed, the heart tube grows and elongates by addition of cells from the second heart field. The ends of the heart tube are relatively fixed by the pericardial sac so that as it elongates it must loop (bend), and in the vast majority of hearts the loop falls to the right (D-loop). Further elongation pushes the mid-portion of the tube (future ventricles) inferior or caudal to the inflow, resulting in the normal relationship between the atria and ventricles. Further growth pushes the outflow medially and is associated with outflow rotation, both processes essential for normal alignment of the outflow. Finally the proximal part of the outflow is incorporated in the RV, shortening the outflow in association with further rotation. While this remodeling is occurring, the outflow is undergoing septation under the influence of cardiac neural crest cells. Septation proceeds from distal to proximal, culminating in formation and muscularization of the infundibular, or muscular, outflow septum, which inserts onto the superior endocardial cushion at the rightward rim of the outflow foramen, walling the aorta into the LV via the outflow foramen and the PA directly into the RV.

Genetic Considerations

CHD is the most commonly occurring birth defect; etiological contributors are increasingly recognized, although often speculated to be multifactorial. Children born with Trisomy 21 have a 50% chance of having CHD, most commonly defects in the atrioventicular canal. Conotruncal defects are associated with a number of chromosomal abnormalities, most notably a deletion at chromosome 22q11 (DiGeorge syndrome). Echocardiographic clues to this association in patients with a conotruncal defect include an associated right aortic arch or aberrant subclavian artery. Many adults currently living with conotruncal defects may not have undergone testing for DiGeorge syndrome. This condition is important to recognize as a variety of psychiatric disorders and disabilities in cognitive function may be present and go untreated. Patients with Noonan syndrome commonly have a dysplastic pulmonary valve and have facial and lymphatic abnormalities. Several defects in specific genes have been associated with Noonan syndrome, most notably PTPN11. Adults with Williams syndrome (7q11.23 deletion) commonly have supravalvar aortic stenosis and diffuse arteriopathy, with a “cocktail-like” personality and hypercalcemia. There is a growing importance of genome-wide analyses in subjects with CHD.

SPECIFIC CHD LESIONS

Dilated Right Heart

There are many congenital etiologies for right heart dilation **(Table 264-3)**. These include congenital valvular anomalies (such as Ebstein anomaly or pulmonary regurgitation), intrinsic RV myocardial anomalies (arrhythromogenic RV dysplasia, Uhl’s anomaly), or shunt lesions occurring proximal to the tricuspid valve. Cardiac imaging is critical in determining the etiology of right heart dilation, and knowledge of the anatomy and physiology of various shunt lesions is essential.

TABLE 264-3 Congenital Etiologies of Right Heart Dilation

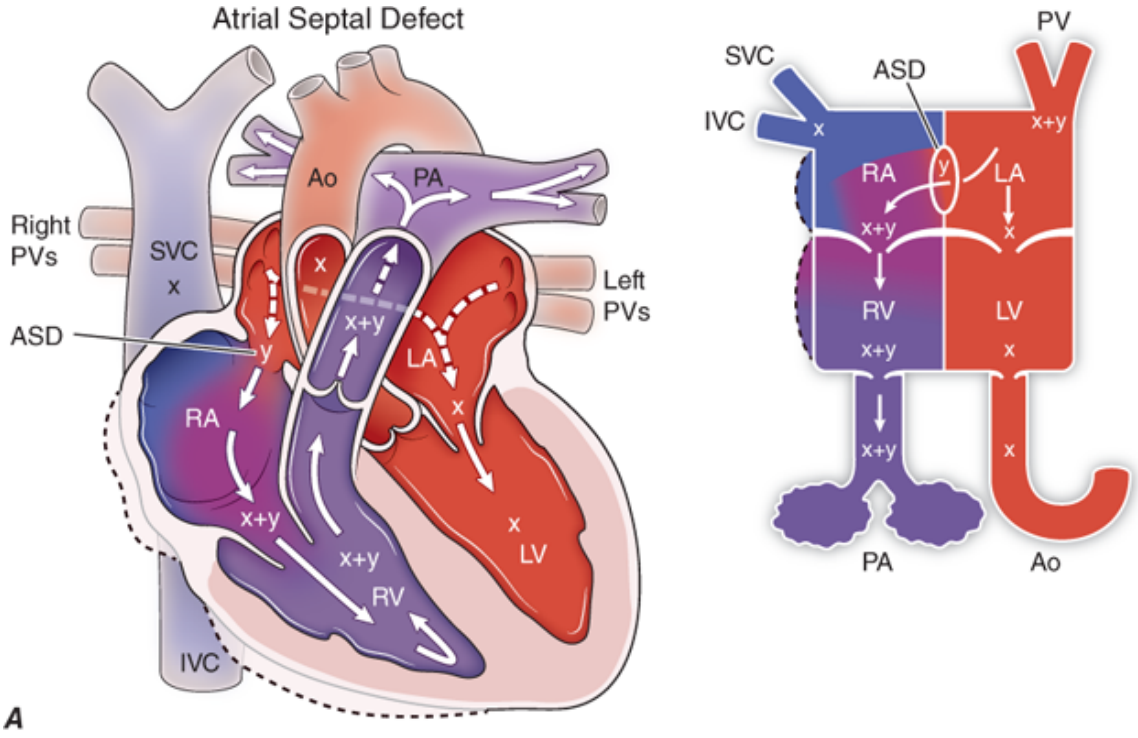
Congenital tricuspid valve disease
Tricuspid valve dysplasia with regurgitation
Ebstein anomaly
Congenital pulmonary valve regurgitation
Pulmonary arterial hypertension
Myocardial abnormalities
Arrhythmogenic RV cardiomyopathy
Uhl’s anomaly
Shunt lesions
Partial anomalous pulmonary venous return
Primum ASD
Secundum ASD
Sinus venosus defect
Coronary sinus septal defect
Gerbode defect (LV-RA shunt)
Coronary artery fistula to the RA, CS
Postoperative residual shunts

Atrial Septal Defect

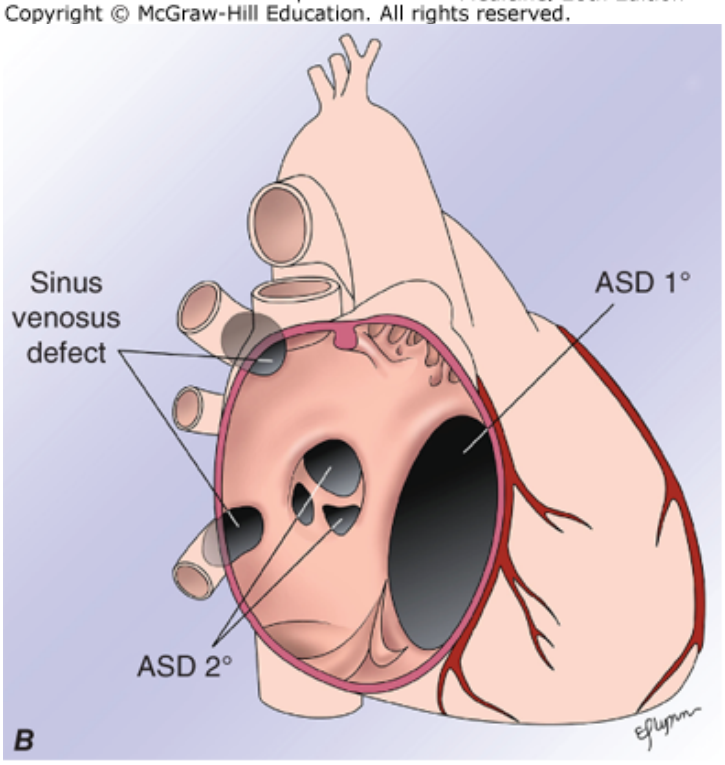
One of the most common etiologies of right heart dilation is presence of an atrial septal defect (ASD, **Fig. 264-2A**). Intracardiac holes allow blood transmission between chambers or spaces based upon relative resistance, propulsion, and flow patterns. Patients with large ASDs often present in childhood; however, many ASDs are not discovered until adult life. The physiology of an ASD is predominantly that of a “left-to-right” shunt (flow of pulmonary venous, or oxygenated, blood toward systemic venous, or deoxygenated, chambers or vessels). The degree of left-to-right shunting determines the amount of right heart volume loading and is dictated by the size of the defect as well as the diastolic properties of the heart. As patients age, several factors, such as diabetes mellitus, systemic hypertension, and atherosclerosis, may contribute to decreased compliance of the left-sided cardiac chambers and contribute to increased left-to-right shunting and symptomatology. The classic physical examination finding is a wide, fixed splitting of the second heart sound, which is due to prolonged RV ejection and increased PA capacitance, which, in turn, delay pulmonary valve closure. The surface electrocardiogram (ECG) commonly displays an incomplete right bundle branch block. Symptoms, when they occur, most commonly include exercise intolerance, arrhythmia, and dyspnea with exertion. It is not uncommon for adults to have incidentally noted asymptomatic ASD during evaluation of other comorbid issues. Right heart dilation, without additional etiology for such, in the setting of unrepaired ASD is considered a risk for progression toward symptomatic right heart failure, atrial arrhythmias, and potential development of pulmonary arterial hypertension (if such is not already present). Therefore, a patient with an ASD and right heart dilation, particularly with symptoms attributable to such, should be offered ASD closure. Pulmonary vascular disease leading to pulmonary hypertension develops in up to 10% of patients with unrepaired ASD, and Eisenmenger syndrome (ES) is a rare complication (see below). Management of patients with concomitant ASD and pulmonary hypertension should be coordinated with both ACHD and pulmonary hypertension experts.

Figure 264-2

A. **Atrial septal defect.** In the presence of an atrial septal defect, the difference in compliance between the (RA+ RV) as compared to the (LA+LV), combined with the size of the defect itself, allows for a “shunt” of flow (“y”) of “red” (oxygenated) blood from the left side of the heart to the right side (deoxygenated). Systemic venous return of pure deoxygenated blood (“x”) is increased by the oxygenated shunted blood (“y”) to increase volume of blood (“x + y”) in the RA, RV, and total blood flow to the lungs. If the volume or the sequelae of the shunted blood is sufficient, RA and RV can dilate (*hashed lines*), and arrhythmias or shortness of breath (and occasionally pulmonary hypertension) can ensue. Ao, aorta; ASD, atrial septal defect; IVC, inferior vena cava; LA, left atrium; LV, left ventricle; PA, pulmonary artery; PV, pulmonary veins; RA, right atrium; RV, right ventricle; SVC, superior vena cava. ***B.*** Diagrammatic representation of the location of various atrial septal defects. ASD 1, primum atrial septal defect; ASD 2, secundum atrial septal defect. *(Part B used with permission from Emily Flynn McIntosh, illustrator.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Figure 264-2B illustrates the locations of various ASDs. The most common type of an ASD is a secundum ASD, which is a defect, or true deficiency in the atrial septum, in the region of the fossa ovalis. This should be differentiated from a patent foramen ovale (PFO), which is persistence of patency of the flap valve of the fossa ovalis (not associated with right-sided cardiac dilation) and persists in up to 25% of adults. Secundum ASDs can often be closed with occluder devices placed percutaneously. However, certain anatomic determinants make percutaneous closure less favorable, including large defects, inadequate tissue rims surrounding the defect, and concomitance of anomalous draining pulmonary veins. A primum ASD is a deficiency of the AV canal portion of

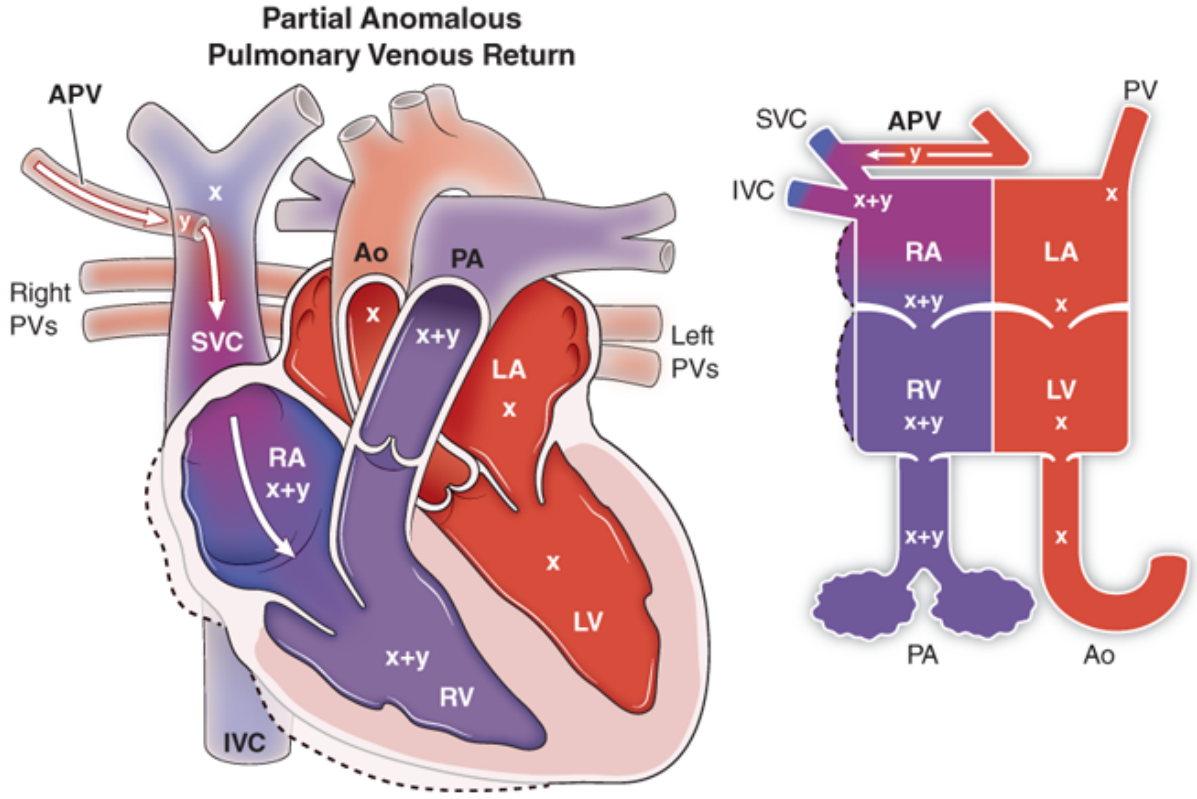
the atrial septum; primum ASD is always associated with abnormal development of the AV valves, most commonly resulting in a cleft in the mitral valve. A coronary sinus defect is rare and involves an opening between the coronary sinus and the left atrium. A sinus venosus defect is not a defect in the atrial septum, but, rather, a defect between either the right superior vena caval-atrial junction and the right upper pulmonary vein(s) or, less commonly, the inferior vena caval-atrial junction and the right lower pulmonary veins. Surgical closure is required for primum ASDs, sinus venosus defects, and coronary sinus septal defects.

Partial Anomalous Pulmonary Venous Return

Partial anomalous pulmonary venous return (PAPVR) is occasionally discovered in adults with right heart dilation, or incidentally on cross-sectional imaging **(Fig. 264-3)**. There are several possible anomalous connections, with the most common being a left upper pulmonary vein to an ascending vertical vein into the innominate vein or the right upper pulmonary vein draining to the superior vena cava. In the latter case, careful attention should be paid to ensure that there is not an associated sinus venosus defect. Concomitant pulmonary hypertension can occur, but is uncommon. Symptomatology may be absent, and decision to repair isolated PAPVR should take into account variance in anatomy, lung ventilation and perfusion, hemodynamic response to shunt, symptoms, and surgical experience.

Figure 264-3

Partial anomalous pulmonary venous return. In the presence of an anomalously draining pulmonary vein (typically to a systemic vein such as the left innominate vein, SVC, or rarely IVC), an obligate “shunt” of flow (“y”) of “red” (oxygenated) blood from the affected pulmonary vein to the right heart (deoxygenated) ensues. Systemic venous return of pure deoxygenated blood (“x”) is increased by the oxygenated shunted blood (“y”) to increase volume of blood (“x + y”) in the SVC, RA, RV, and total blood flow to the lungs. If the volume or the sequelae of the shunted blood is sufficient, RA and RV can dilate (*hashed lines*), or shortness of breath can ensue. Ao, aorta; APV, anomalous pulmonary vein; IVC, inferior vena cava; LA, left atrium; LV, left ventricle; PA, pulmonary arteries; PV, pulmonary veins; RA, right atrium; RV, right ventricle; SVC, superior vena cava.



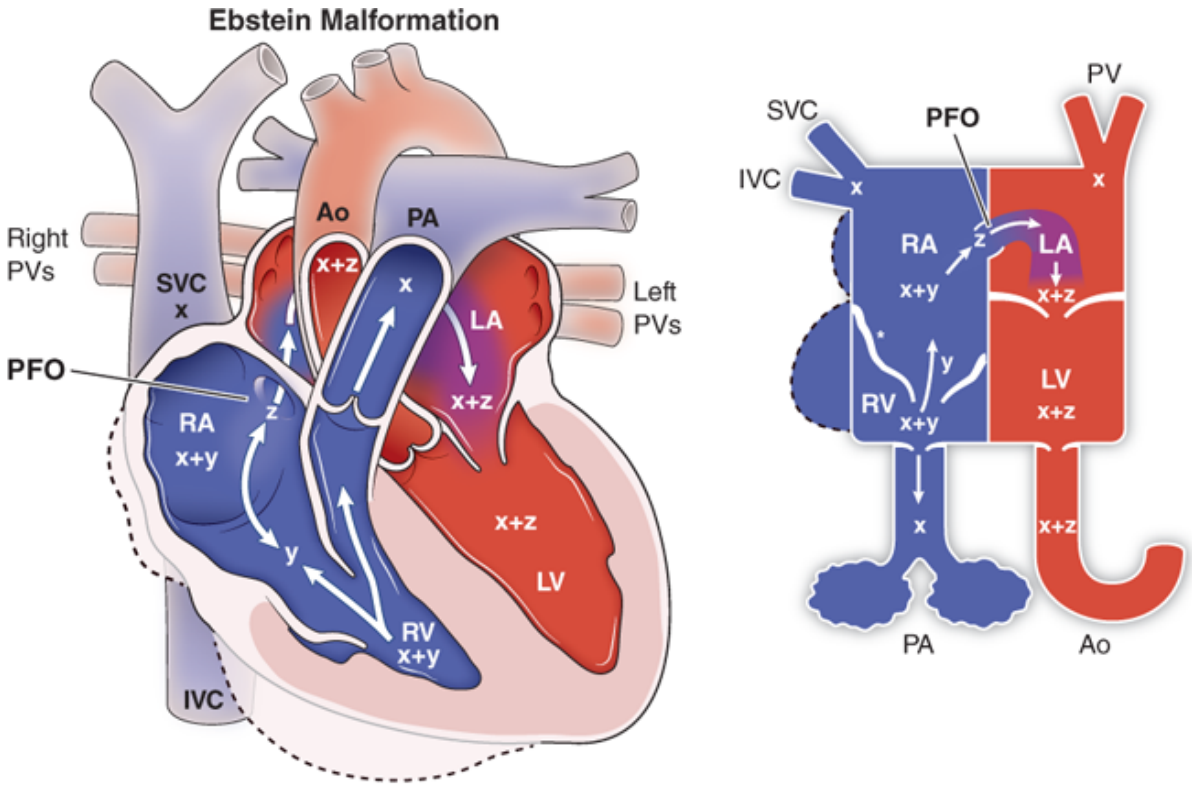
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ebstein Anomaly

Ebstein anomaly **(Fig. 264-4)** is the result of embryologic failure of delamination, or “peeling away,” of the tricuspid valve leaflets from the ventricular myocardium, resulting in adherence of the valve leaflets to the underlying myocardium. This results in a wide variety of abnormalities, including apical and posterior displacement of the dilated tricuspid valve annulus, dilation of the “atrialized” portion of the RV, and fenestrations, redundancy, and tethering typically of the anterior leaflet of the tricuspid valve. The malformed tricuspid valve is usually regurgitant, but may occasionally be stenotic. The clinical presentation of Ebstein anomaly in the adult depends on several factors, including the extent of tricuspid valve leaflet distortion, degree of tricuspid regurgitation (TR), right atrial pressure, and presence of an atrial level shunt. The physical examination of a patient with Ebstein anomaly may vary depending on the severity of disease. In more severe cases, the first heart sound may be split and the second component of the first heart sound may have a distinctive snapping quality (known as the sail sign, due to the redundancy of the anterior tricuspid valve leaflet). Patients with significant TR may have prominent “v” waves of the jugular venous pulsations; however, this finding is often absent due to abnormal right atrial compliance. The ECG is often abnormal, with right atrial and ventricular enlargement. Up to 20% of patients have evidence of ventricular pre-excitation (Wolff-Parkinson White pattern). Surgical treatment includes a tricuspid valve repair or replacement, closure of any atrial level defects, and arrhythmia ablative procedures.

Figure 264-4

Ebstein malformation. In the presence of Ebstein anomaly, the tricuspid valve leaflets can be redundant, fenestrated and sail-like (typically seen in the anterior leaflet *), or adherent to the underlying myocardium with apical displacement of the non-adherent components (typically the septal and posterior leaflets). Location and degree of leaflet coaptation are variable and account for varying degrees of tricuspid regurgitation, shift of the functional tricuspid valve anterior from the anatomic annulus into the right ventricle, “atrialization” of the right ventricle, and most commonly angulation of the tricuspid valve into the RV outflow tract. RA and RV dilation (*hashed lines*) can occur due to the effects of combined volume from systemic venous return (“x”) and tricuspid regurgitant flow (“y”). PFO is frequent; worsening compliance and elevation of pressure in the RA as compared to the LA can lead to increasing “right-to-left” (deoxygenated to oxygenated) shunt and cyanosis. RV myocardial function may be abnormal. Ao, aorta; IVC, inferior vena cava; LA, left atrium; LV, left ventricle; PA, pulmonary arteries; PFO, patent foramen ovale; PV, pulmonary veins; RA, right atrium; RV, right ventricle; SVC, superior vena cava; *, anterior tricuspid valve leaflet.



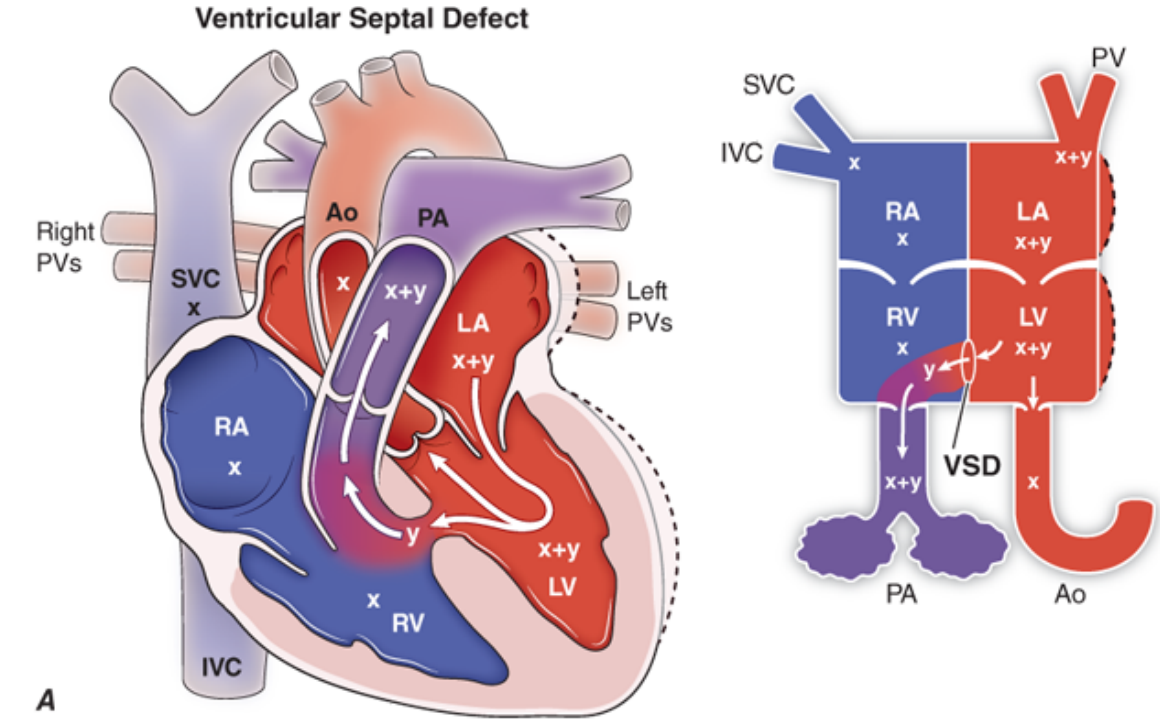
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Shunt Lesions Causing Left Heart Dilation

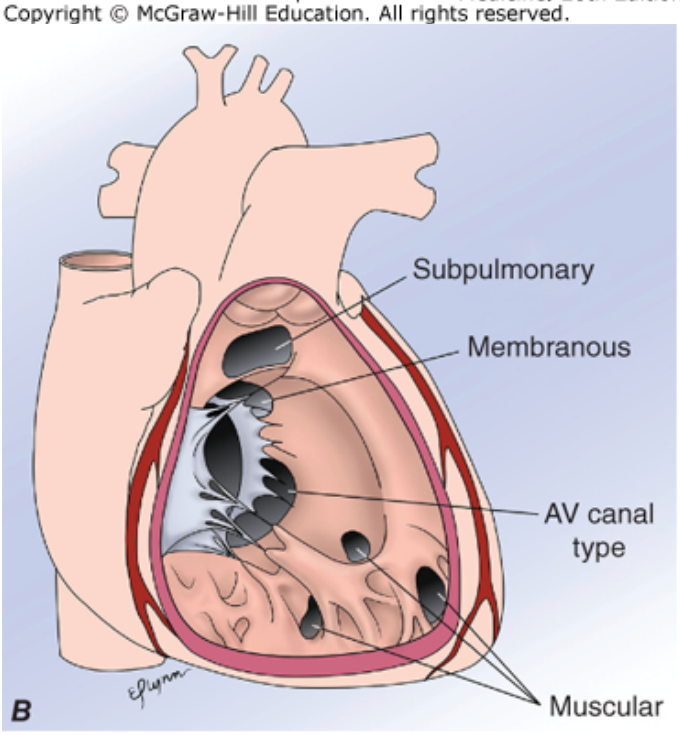
Intracardiac shunts or intravascular passages that occur below the level of the tricuspid valve result in left heart dilation. The two major types of congenital shunts that result in left heart dilation are a ventricular septal defect (VSD, Fig. 264-5A) and patent ductus arteriosus (PDA, Fig. 264-6).

Figure 264-5

A. Ventricular septal defect. In the presence of a ventricular septal defect, the difference in pressure and outflow resistance in systole (and the difference in compliance in diastole) between the RV and LV, combined with the size of the defect itself, allow for a "shunt" of flow ("y") of "red" (oxygenated) blood from the left side of the heart to the right side (deoxygenated). Systemic venous return of pure deoxygenated blood ("x") is increased by the oxygenated shunted blood ("y") to increase volume of blood ("x + y") through the outflow of the RV into the lungs, and in the left atrium and left ventricle. If the volume or the sequelae of the shunted blood is sufficient, LA and LV can dilate (*hashed lines*), and arrhythmias or shortness of breath (and occasionally pulmonary hypertension) can ensue. Ao, aorta; IVC, inferior vena cava; LA, left atrium; LV, left ventricle; PA, pulmonary arteries; PV, pulmonary veins; RA, right atrium; RV, right ventricle; SVC, superior vena cava; VSD, ventricular septal defect. ***B.*** Diagrammatic representation of the location of various ventricular septal defects. AV, atrioventricular. *(Part B used with permission from Emily Flynn McIntosh, illustrator.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

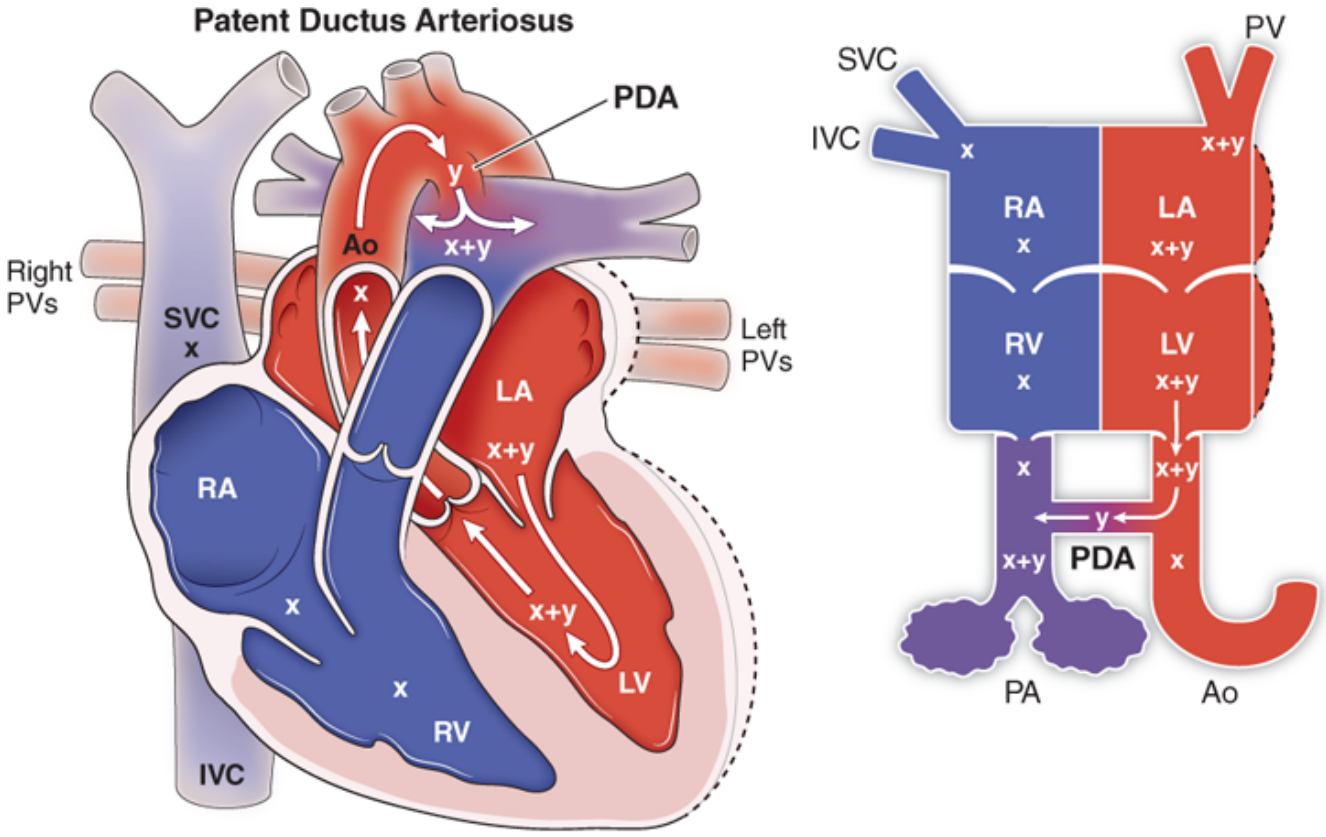


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Figure 264-6

Patent ductus arteriosus. In the presence of a patent ductus arteriosus, the difference in pressure and resistance in both systole and diastole between the pulmonary arteries and the aorta, combined with the size of the ductus itself, allow for a “shunt” of flow (“y”) of “red” (oxygenated) blood from the aorta to the pulmonary arteries (deoxygenated). Systemic venous return of pure deoxygenated blood (“x”) is increased by the oxygenated shunted blood (“y”) to increase volume of blood (“x + y”) in the lungs, the left atrium, the left ventricle, and out the aortic valve. If the volume or the sequelae of the shunted blood is sufficient, LA and LV

can dilate (*hashed lines*), and arrhythmias or shortness of breath (and occasionally pulmonary hypertension) can ensue. Ao, aorta; IVC, inferior vena cava; LA, left atrium; LV, left ventricle; PA, pulmonary arteries; PDA, patent ductus arteriosus; PV, pulmonary veins; RA, right atrium; RV, right ventricle; SVC, superior vena cava.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ventricular Septal Defects

VSD are the most common congenital anomaly recognized at birth, however, they account for only about 10% of CHD in the adult, due to the high rate of spontaneous closure of small VSDs during the early years of life. Large VSDs usually cause symptoms of heart failure and poor somatic growth, and are most often closed before adulthood. Several classification systems for VSDs exist. **Figure 264-5B** illustrates various locations of VSDs; the most common location is in the membranous septum (also referred to as perimembranous, or outlet defects). Muscular defects that persist into adult life are often pressure and flow restricted, resulting in no significant hemodynamic consequence. Atrioventricular canal defects, also referred to as inlet defects, are located in the crux of the heart and are associated with abnormalities of the atrioventricular valve leaflets. Subpulmonary defects, also known as conal septal defects, are commonly associated with prolapse of the right coronary cusp and aortic insufficiency. The outcome for adults with small VSDs without evidence of ventricular dilation or pulmonary hypertension is generally excellent.

Patent Ductus Arteriosus

A PDA courses between the aortic isthmus and the origin of one of the branch pulmonary arteries. Small PDAs are often silent to auscultation, and do not cause hemodynamic changes. The classic murmur is heard best just below the left clavicle and typically extends from systole past the second heart sound into diastole, reflecting flow turbulence and gradient between the aorta and the pulmonary arteries (resulting in left-to-right shunting). Large PDAs will lead to left heart dilation and may lead to chronically elevated pulmonary vascular resistance, including the potential for ES.

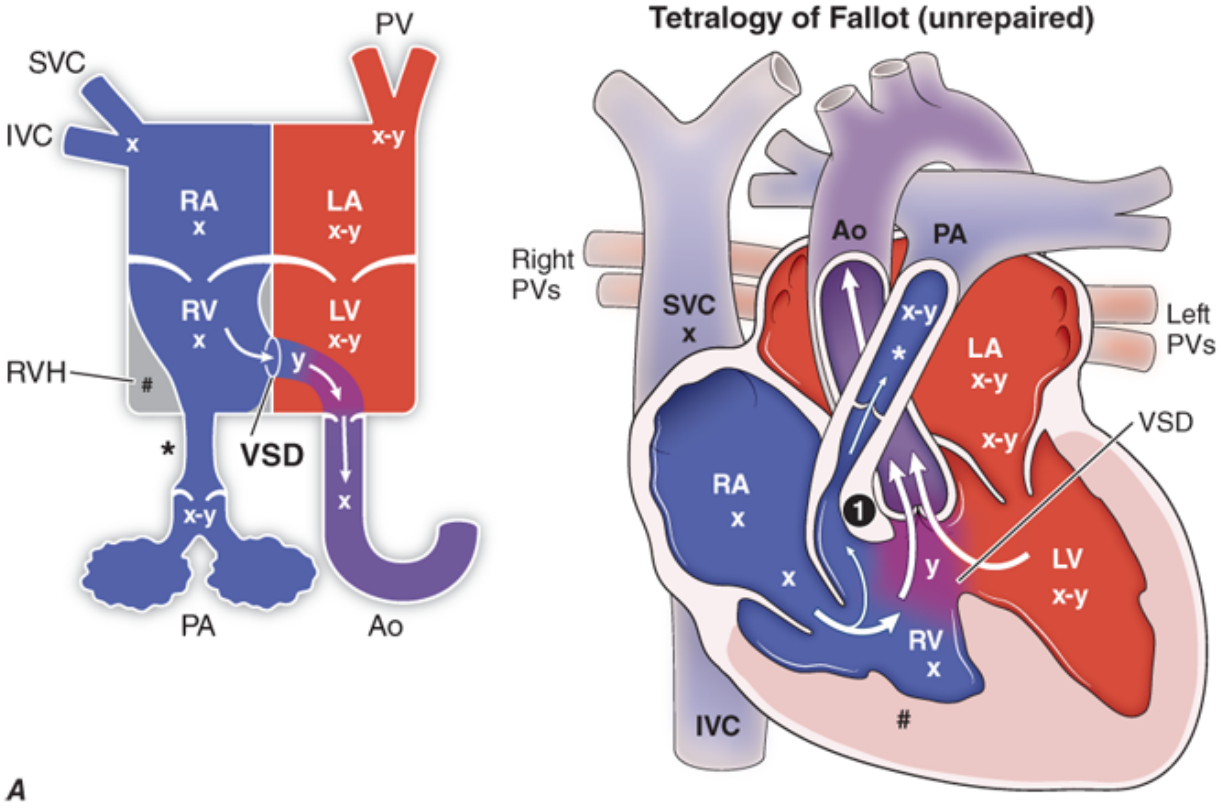
MODERATE AND COMPLEX CHD

Tetralogy of Fallot

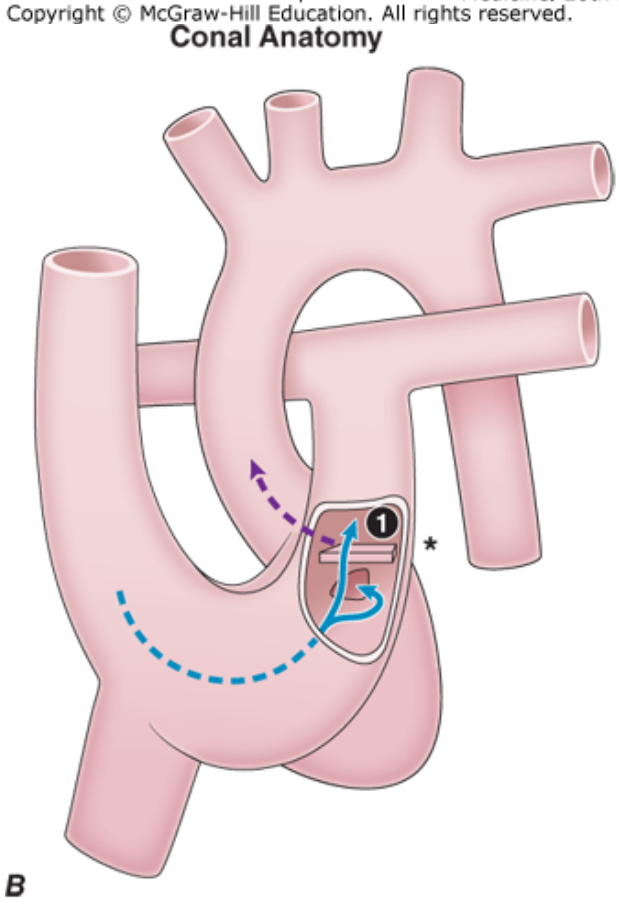
Tetralogy of Fallot (TOF) is the most common form of cyanotic CHD, occurring in 0.5 per 1000 live births. It involves what may be a singular deviation of the anterior conal septum, resulting in right ventricular outflow tract (RVOT) obstruction, a VSD, right ventricular hypertrophy and an overriding aorta **(Fig. 264-7A, B)**. There is a large spectrum of severity of disease in TOF, from patients who have only mild pulmonary stenosis to those with complete pulmonary atresia (TOF/PA). Current surgical strategies involve primary repair in infancy **(Fig. 264-7C)**; however, many adults may have first undergone palliative procedures (Blalock-Taussig, Potts, Waterston shunts) prior to a complete repair. The goal of surgical repair is to alleviate the pulmonary stenosis and close the VSD. Up to 7% of patients with TOF have an anomalous coronary artery, most commonly, an anomalous left anterior descending coronary artery from the right coronary cusp. Patients with an anomalous coronary as well as those with TOF/PA may require a RV-to-PA conduit.

Figure 264-7

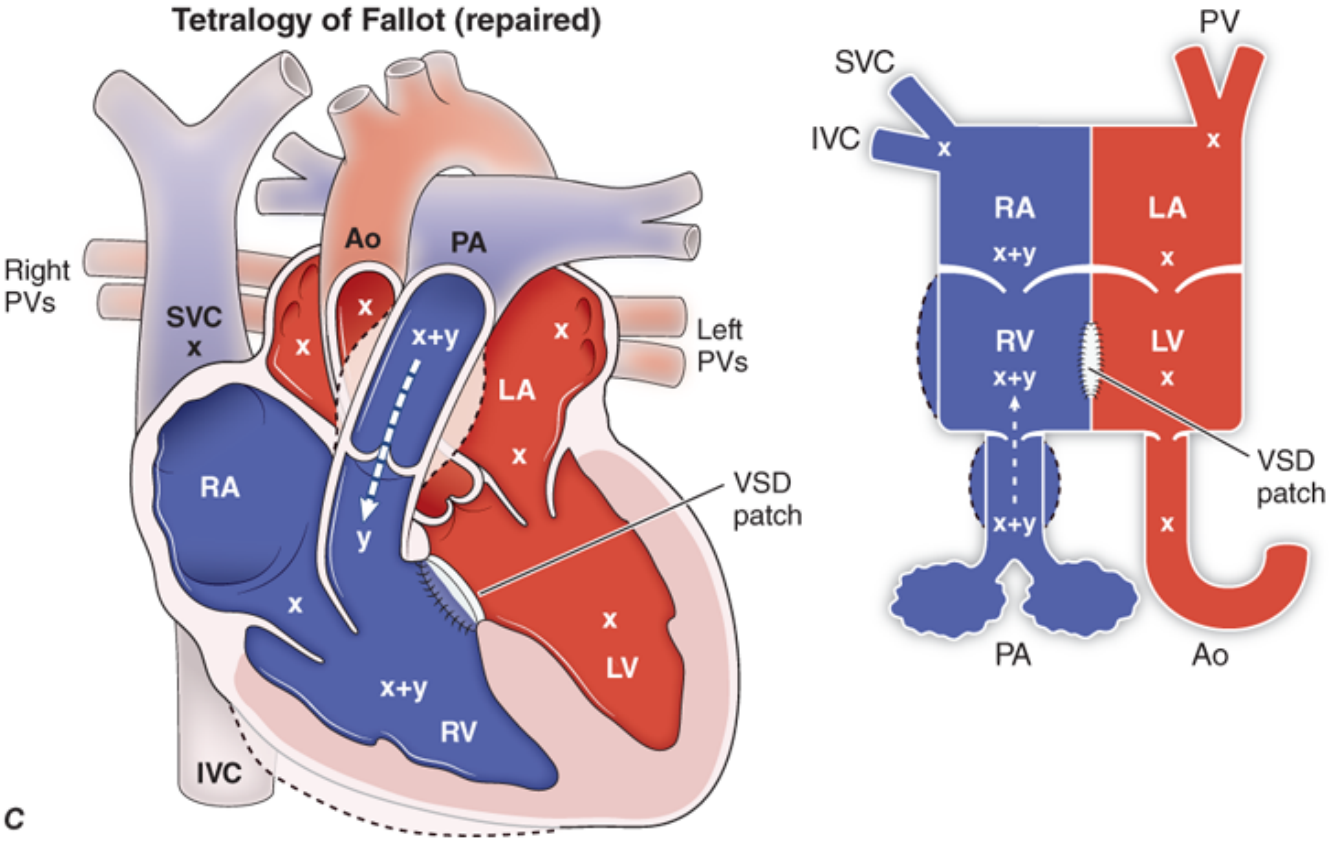
A. Tetralogy of Fallot involves anterior and superior malalignment of a bar of tissue (conal septum) (see * in Fig. 264-7B, which presents a cut-away view through the anterior surface of the RV, into the RV outflow), partially obstructing the right ventricular outflow (under the pulmonary valve, i.e., “subpulmonary stenosis”-labeled as 1), and leaving a gap in the interventricular septum (VSD). The pulmonary valve annulus is typically hypoplastic. Outflow obstruction prevents regression of right ventricular hypertrophy #, which was present in utero. The difference in pressure and outflow resistance in systole (and the difference in compliance in diastole) between the obstructed RV and the LV allow for a “shunt” of flow (“y”) of “blue” (deoxygenated) blood from the right side of the heart to the left side (oxygenated). Systemic venous return of pure deoxygenated blood (“x”) is decreased by the shunted blood (“y”) leading to a total decrease in the volume of blood (“x – y”) passing beyond into the lungs. The deoxygenated shunted blood (“y”) mixes with fully oxygenated blood in the LV, contributing to systemic arterial cyanosis. ***C.*** Tetralogy of Fallot—repaired. After modern repair of tetralogy of Fallot, VSD has been patched closed, and outflow tract obstruction has been surgically removed, frequently at the expense of a patch enlarging the pulmonary valve annulus at the expense of sacrificing the integrity of the pulmonary valve (causing pulmonary regurgitation). Volume of pulmonary regurgitant volume (“y”) is added to systemic venous return (“x”), contributing to RV chamber enlargement (*hashed lines*), and may be associated with tricuspid annular dilation and valve regurgitation resulting in RA enlargement. Ao, aorta; IVC, inferior vena cava; LA, left atrium; LV, left ventricle; PA, pulmonary arteries; PV, pulmonary veins; RA, right atrium; RV, right ventricle; RVH, right ventricular hypertrophy; SVC, superior vena cava; VSD, ventricular septal defect.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Adults with repaired TOF often have hemodynamic sequelae that may require re-intervention in adulthood **(Table 264-4)**. Pulmonary regurgitation is common following TOF repair and is usually associated with RV dilation. Accurate quantification of RV size, function, and mass is particularly important in adults after repair of TOF, as RV dilation, dysfunction, and hypertrophy are associated with adverse outcomes in these patients. Patients may also have residual RVOT obstruction, which may occur beneath the pulmonary valve, at the valve level, above the valve, or in the branch pulmonary arteries. Cardiac magnetic resonance imaging is routinely used in the surveillance of these patients. Left ventricular dysfunction is present in at least 20% of adults with repaired TOF, particularly those who were repaired later in life, had prior palliative shunts, or have concomitant RV dysfunction.

TABLE 264-4 Potential Sequelae of Repaired Tetralogy of Fallot

Right atrial dilation
Right ventricular dilation
Right ventricular dysfunction
Right ventricular outflow tract obstruction
Pulmonary regurgitation
Branch pulmonary artery stenosis
Tricuspid regurgitation
Residual ventricular septal defect
Left ventricular dysfunction
Aortic root dilation
Atrial arrhythmias
Ventricular arrhythmias
Sudden cardiac death

As patients age with repaired TOF, both atrial and ventricular arrhythmias occur with increasing frequency. A QRS duration on a resting ECG of 180 ms or more has been associated with increased risk of ventricular tachycardia and sudden death in this patient population. In one prospective follow-up study of 144 adults with repaired TOF, there was a 72% survival at 40 years, but only a 25% cumulative event free survival. These events include need for re-intervention (most commonly pulmonary valve replacement, PVR), symptomatic arrhythmias, and heart failure.

The most common re-intervention in a repaired TOF patient is a PVR. However, optimal timing of PVR in these patients remains unclear. Although PVR has been shown to decrease right ventricular volumes and subjectively improve symptoms, it has not been proven to result in an improved ejection fraction or less adverse outcomes, such as ventricular arrhythmias or death. Traditionally, PVR has been accomplished with a surgical procedure; however, percutaneous implantation of pulmonary valves is becoming increasingly utilized in clinical practice.

Patients with repaired TOF may also undergo interventions including closure of residual VSDs, dilation and/or stenting of the RVOT or branch pulmonary arteries, and tricuspid valve repair. Patients with clinically significant arrhythmias may benefit from catheter ablation.

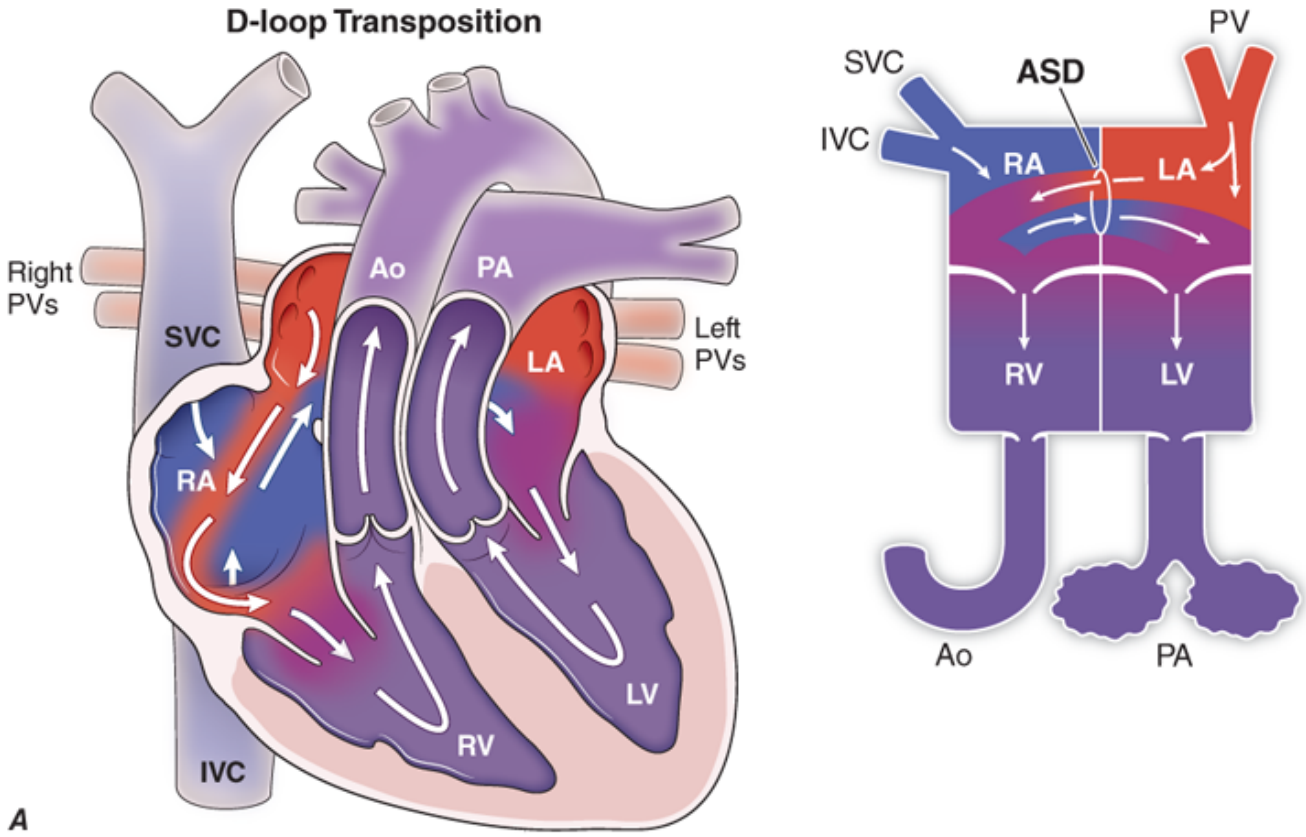
Transposition of the Great Arteries

Transposition of the great arteries (TGA) is defined by the great arteries arising from the opposite side of the ventricular septum than normal; as such, the aorta arises from the RV and the PA from the LV. The more common form of TGA, known as D-loop TGA, involves atrioventricular concordance and ventricular-arterial discordance, resulting in a physiology that allows two circuits to be in parallel rather than in series **(Fig. 264-8A)** and intense cyanosis shortly after birth. This physiology is not compatible with long-term survival without surgical intervention. Patients with TGA may be born with additional congenital defects (most commonly a VSD).

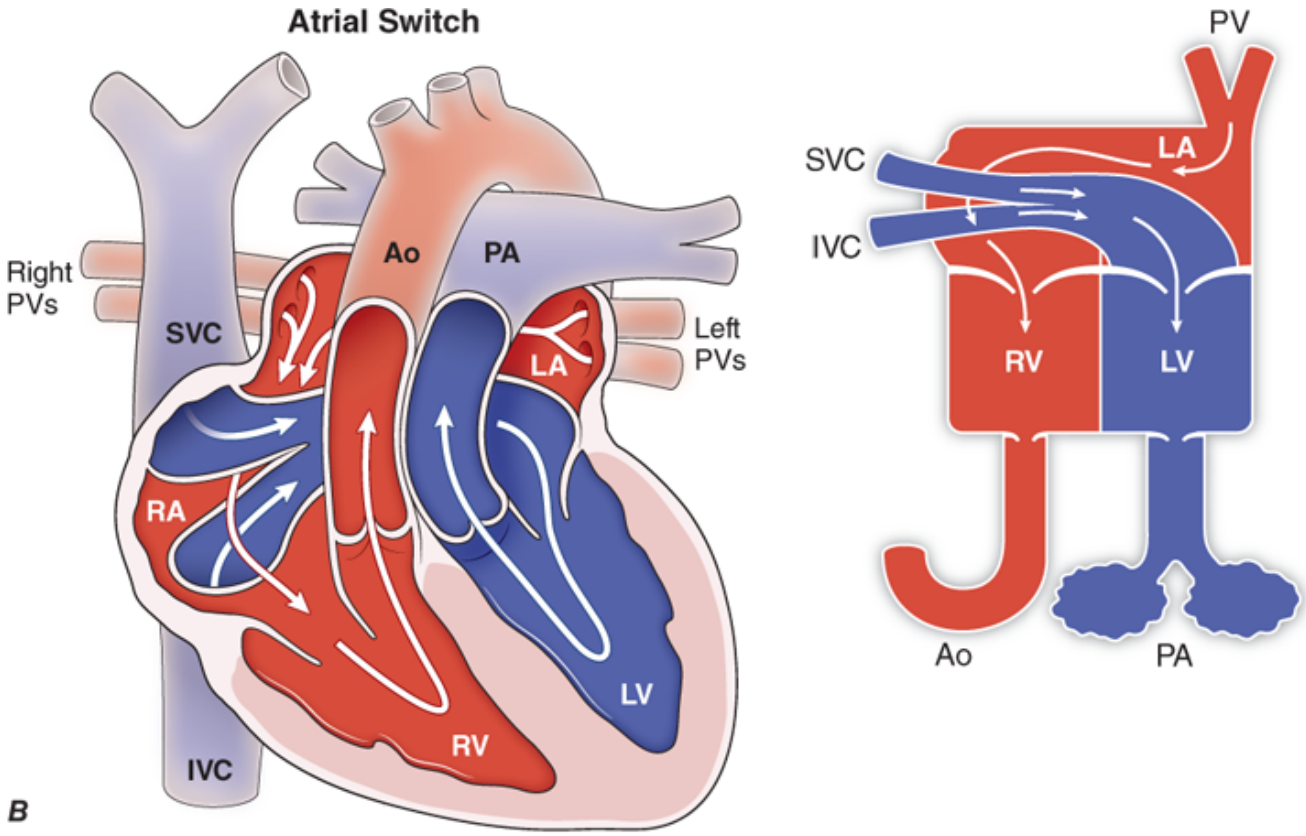
Figure 264-8

A. Transposition of the great arteries. When the great arteries are transposed, the aorta arises from the RV, and the pulmonary artery arises from the LV, leaving deoxygenated blood circulating from systemic veins to systemic arteries in separated fashion from oxygenated blood, which circulates from pulmonary veins to pulmonary arteries. Without interchamber or intravascular communications, this circulation is incompatible with life.

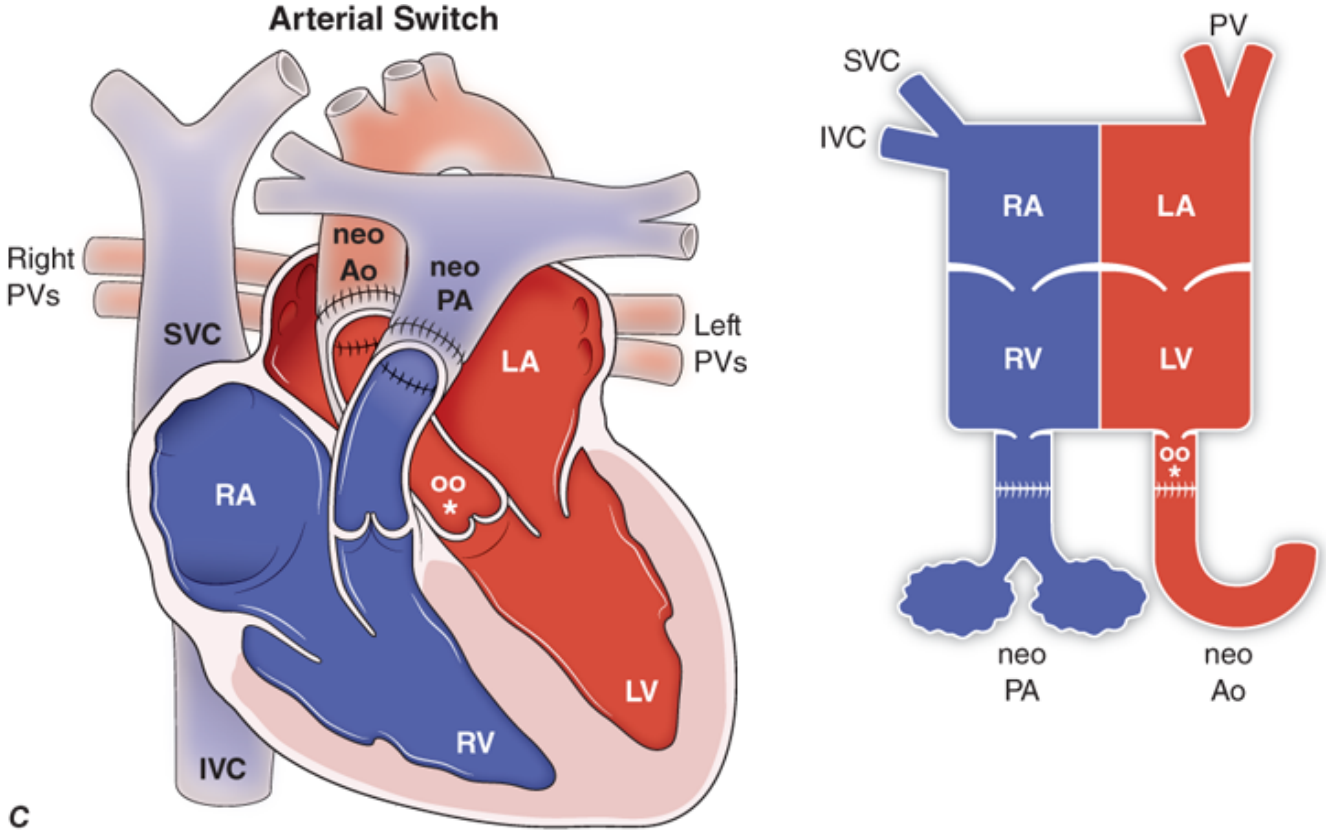
Presence of an atrial septal defect (ASD), depicted here, ventricular septal defect (VSD), or patent ductus arteriosus (PDA), allow for some interchamber or intravascular mixing, and, at best, partial relief of cyanosis and sustenance of life, at the expense of increased pulmonary blood flow. ***B.*** Atrial switch. Atrial level switch procedures ("Mustard" and "Senning") were the first standardized surgeries to alter the natural course of complex congenital heart disease, utilizing intracardiac re-routing via a "baffle" to re-direct blood flow. The atrial switch simulates inverted trousers, with each "pants-leg" * attaching to either the SVC or the IVC, transporting deoxygenated blood through the interior of the trousers to the "waist of the trousers" and directing blood through the mitral valve to the LV and out the PA. Surgical removal of the atrial septum allows pulmonary venous return to traverse from posterior left atrium through the space between the pants legs of the baffle, through the tricuspid valve to the RV (serving as the "systemic ventricle," i.e., that pumps to the systemic arterial circulation) and out the aorta. Non-infrequent sequelae include sinus node dysfunction, atrial arrhythmias, systolic dysfunction of the RV, tricuspid regurgitation (from RV to LA), leaks in the baffle material allowing shunting of blood, and obstruction of the systemic or pulmonary venous baffles. ***C.*** Arterial switch. The arterial switch operation allowed both anatomic and physiologic correction for D-loop transposition of the great arteries. Successful surgical switching of the PA and the Ao above the level of the native roots (*hashed lines*) necessitated ability to transfer coronary artery origins contained within a button of tissue * back to the neo-aorta (now supported by the LV). Deoxygenated blood flow from SVC and IVC pass from RA to RV to PA, and oxygenated blood passes from PV to LA to LV to Ao. Uncommon sequelae include obstruction at any of the surgical sites (supravalvar PA or Ao stenosis, coronary orifice obstruction), or more distal obstructions due to tension placed on the PA, Ao or coronary arteries. Ao, aorta; IVC, inferior vena cava; LA, left atrium; LV, left ventricle; PA, pulmonary arteries; PV, pulmonary veins; RA, right atrium; RV, right ventricle; SVC, superior vena cava.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The surgical repairs for D-loop TGA have evolved over time. In the late 1950s through the 1970s, the atrial switch procedure (Mustard, Senning procedures) was performed **(Fig. 264-8B)**. These atrial switch procedures relieved the cyanosis but left the patient with a systemic RV. Despite moderate-term survival over decades, there are multiple long-term sequelae that may present following the atrial switch procedure. The most worrisome complication is that of systemic right ventricular dysfunction. The prevalence of right ventricular dysfunction in this population is not well defined due to difficulties in quantifying systemic RV function. Limited study has failed to reveal medical therapies effective for systemic right ventricular dysfunction.

A subset of patients with D-loop TGA, VSD, and PS may have undergone a Rastelli procedure. This intervention involves placing a RV-to-PA conduit and routing the LV to the aorta through the VSD, which results in relief of cyanosis and the benefit of a systemic LV.

In the 1980s, the arterial switch procedure (ASO, **Fig. 264-8C**) became the surgical procedure of choice for D-loop TGA. This procedure involves transecting the great arteries above the sinuses, and placing the pulmonary arteries anteriorly to come into alignment with the RV, resulting in draping of the branch pulmonary arteries over the ascending aorta. A coronary artery translocation is performed. The arterial switch operation has resulted in substantial long-term survival.

The potential long-term sequelae of the various surgical procedures for D-loop TGA are listed in **Table 264-5**.

TABLE 264-5 Long-Term Sequelae of D-loop TGA Surgery

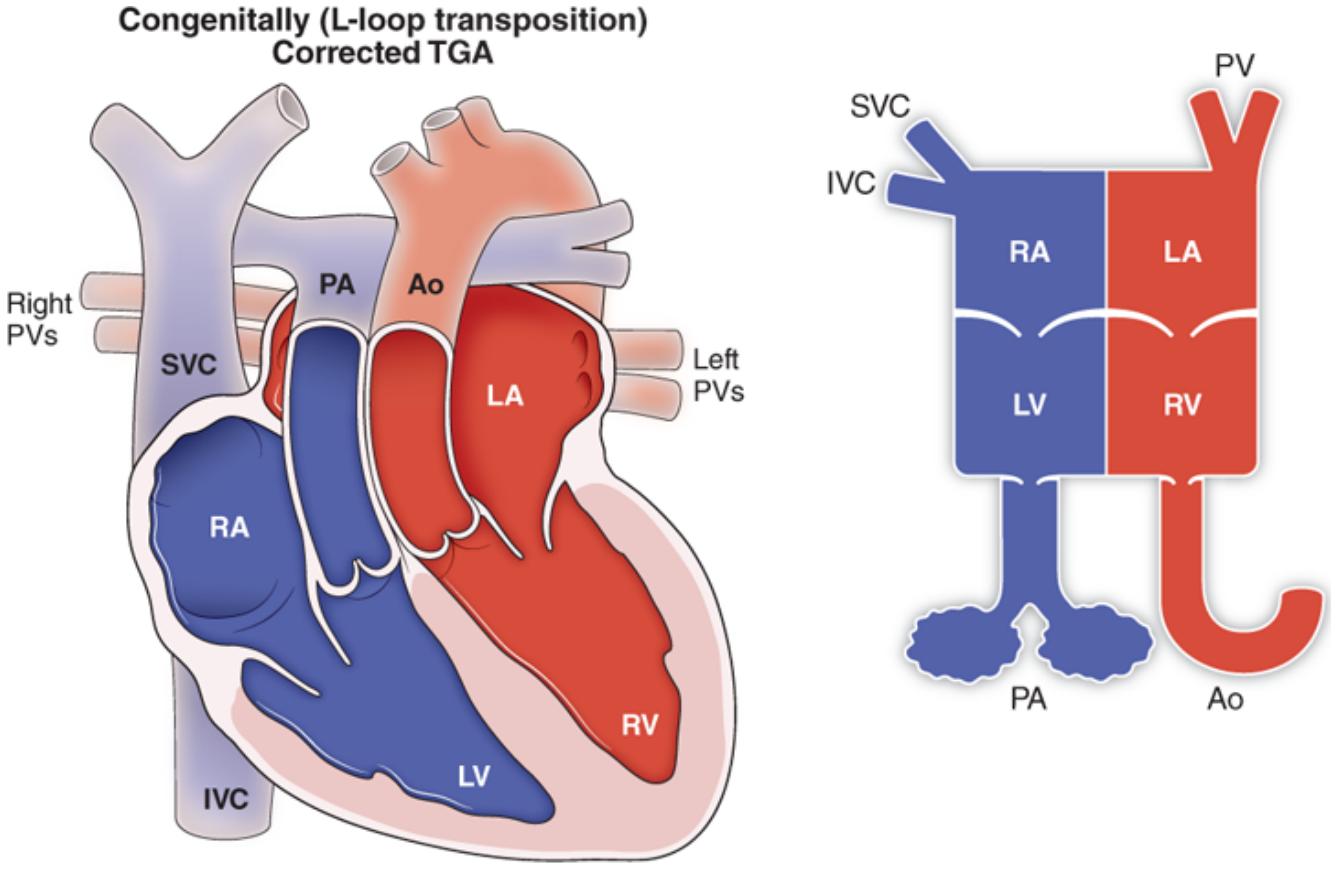
Atrial Switch	Arterial Switch	Rastelli Procedure
Systemic venous baffle	Arterial anastomosis stenosis	Subaortic stenosis
Pulmonary venous baffle	Branch PA stenosis	RV-PA conduit obstruction
RV (systemic) dysfunction	Neo-aortic root dilation	Pulmonary regurgitation
Tricuspid regurgitation	Neo-aortic regurgitation	Ventricular dysfunction
Baffle leaks	Coronary artery stenosis	
LVOT obstruction (PS)	LV dysfunction	

Abbreviations: LV, left ventricle; LVOT, left ventricular outflow tract; PA, pulmonary artery; RV, right ventricle.

The less common form of TGA, known as L-loop TGA (physiologically corrected TGA, **Fig. 264-9**), may not require surgical intervention, but is presented here in relation to other forms of TGA. L-loop TGA involves both atrioventricular discordance (RA allowing passage of deoxygenated systemic venous return to the LV, and conversely, the left atrium conducting oxygenated pulmonary venous blood to the RV) as well as ventriculo-arterial discordance (connections of LV to PA, RV to aorta). This results in normal arterial oxygen saturation, yet an RV associated with the aorta. Patients with L-loop TGA commonly have associated congenital anomalies, including dextrocardia, ASDs, a dysplastic tricuspid valve, and pulmonary stenosis. Conduction disturbances are common, and complete heart block occurs in up to 30% of patients. Those patients without associated defects may not present until later in life, most commonly with heart failure, tricuspid regurgitation, or newly recognized conduction disease.

Figure 264-9

Congenitally corrected transposition of the great arteries. Physiologically corrected transposition of the great arteries (also known as congenitally corrected transposition of the great arteries) is characterized by atrioventricular discordance and ventriculoarterial discordance. Systemic venous blood passes from the right atrium (RA) through the mitral valve into the morphologic left ventricle (LV) to the pulmonary artery (PA). Oxygenated blood then returns to the lungs to the left atrium (LA) through the tricuspid valve into the morphologic right ventricle (RV) and then out the aorta (Ao). IVC, inferior vena cava; PV, pulmonary veins; SVC, superior vena cava.



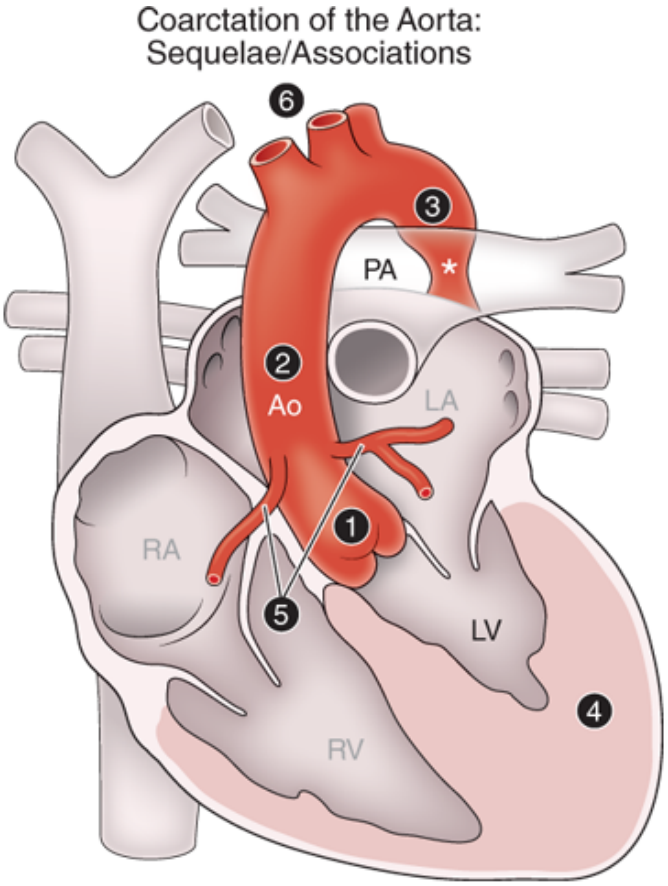
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Coarctation of the Aorta

Adults with coarctation of the aorta **(Fig. 264-10)** typically have a shelf-like obstruction at the level of the descending aorta that passes just posterior to the junction of the main and left PA; obstruction less commonly involves the transverse aortic arch. On physical examination, the lower extremity blood pressure and pulses are lower than (and delayed in timing, in contrast to) the upper extremity values, unless significant aortic collaterals have developed. A continuous murmur over the scapula may be present, due to the collateral blood flow. Significant coarctation increases afterload to all proximal structures in the path of oxygenated blood, from LV and coronary arteries, to ascending and transverse aorta, to cerebral and arm vessels and proximal descending aorta. Bicuspid aortic valve (typically with right-left commissural fusion) is a common association. In women with short stature, webbed neck, lymphedema, and primary amenorrhea, a concomitant diagnosis of Turner syndrome should be considered the presence of which indicates greater degree of, and risks from, sequelae from seemingly similar anatomy and physiology. Patients who have undergone surgical repair in general have a good prognosis; however, they remain at risk for systemic hypertension, premature atherosclerosis, LV failure, as well as aortic aneurysm, dissection, and recurrent coarctation.

Figure 264-10

Aortic coarctation (*). Bicuspid aortic valve (1) is most common concomitant lesion. Sequelae from aortic coarctation (unrepaired or repaired) include systemic arterial hypertension, ascending (2) or descending (3) aortic enlargement or aneurysm formation, left ventricular hypertrophy (4), LV diastolic and systolic heart failure, accelerated coronary (5) or cerebral (6) atherosclerosis, cerebral aneurysm formation, and recurrence of coarctation after repair. Ao, aorta; LV, left ventricle; PA, pulmonary arteries.



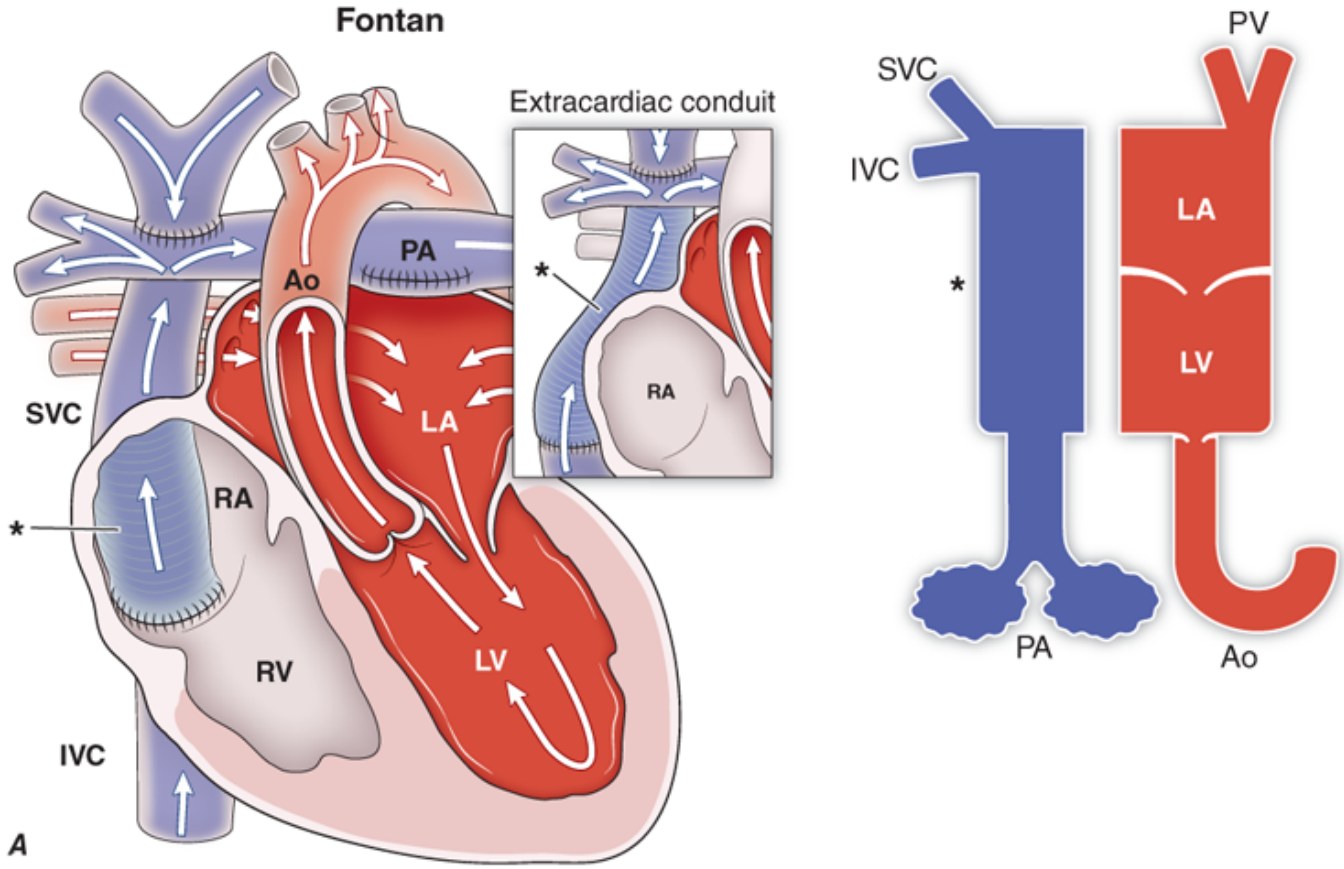
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Single Ventricle Physiology

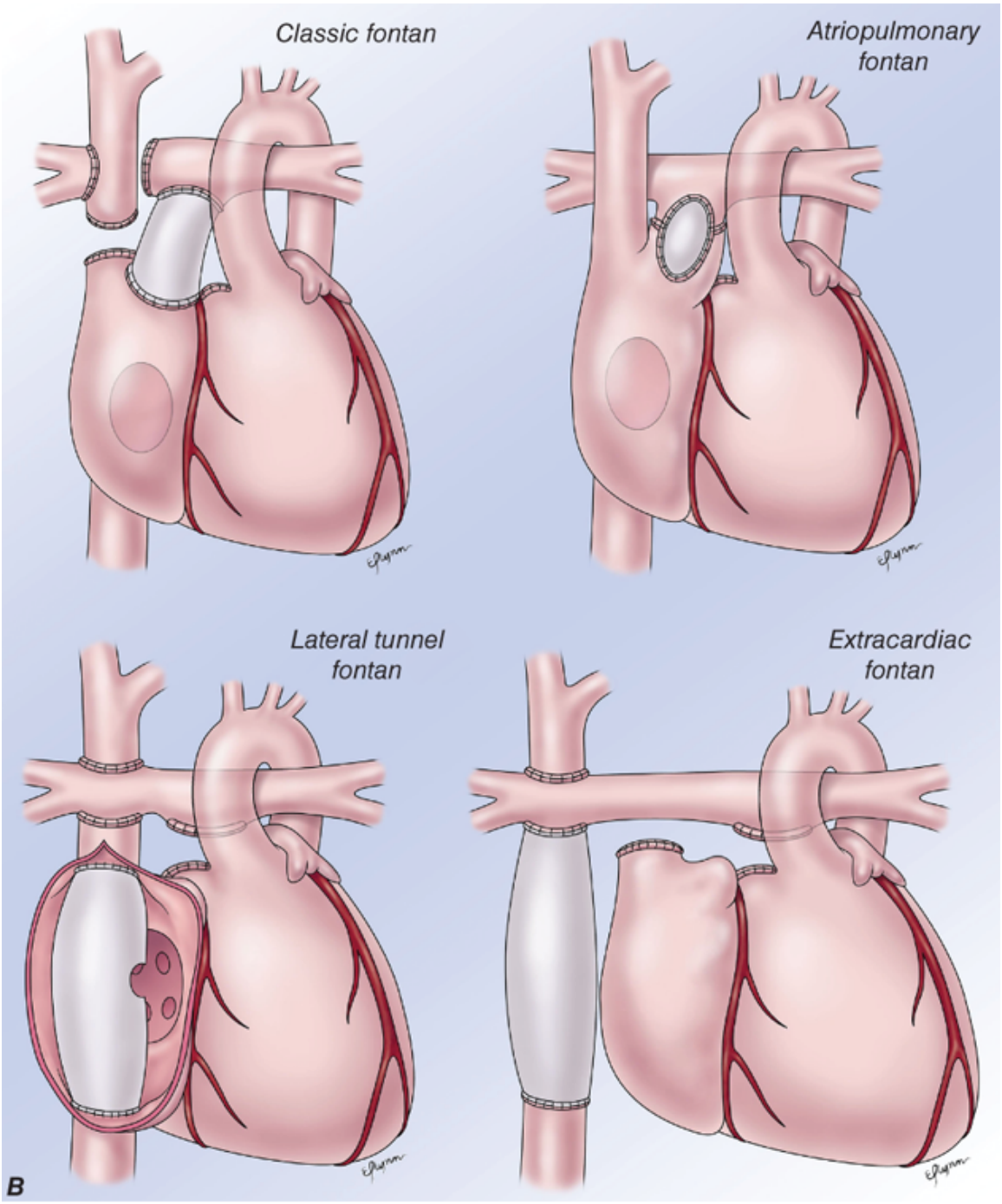
The term, “single ventricle heart disease,” is imprecise, but useful in some settings, as it refers to congenital heart conditions in which one ventricle or its valves preclude surgical creation of a biventricular circulation. Common congenital diagnoses in this category include tricuspid atresia, double inlet LV, and hypoplastic left heart syndrome. Most patients with single ventricle physiology undergo a series of surgeries culminating in a Fontan procedure **(Fig. 264-11A, B)**. Since its initial use for tricuspid valve atresia in 1971, multiple modifications of this procedure have occurred, with common features of near complete separation of the pulmonary and systemic circulations. The Fontan procedure utilizes the single ventricle to pump pulmonary venous (oxygenated) blood through the aorta to the body, and allows for “passive” flow of systemic venous return of deoxygenated blood through surgically created connections to the lungs. Patients who have undergone a Fontan procedure are at risk for multiple comorbidities in adulthood, including atrial arrhythmias, heart failure, renal and hepatic dysfunction, and both venous and arterial thrombosis and embolism.

Figure 264-11

A. Fontan surgery creates a unique circulation in which deoxygenated blood is directed to the PAs from the SVC and IVC in a fashion that bypasses any pumping chamber. The SVC and IVC are connected * via either an internal “tunnel” or an extracardiac conduit that guides flow to the PA. Pulmonary venous (oxygenated) return courses from PV to LA to LV to aorta. In contrast to physiology in normal adults (where pressure is generated by an RV to propel blood flow from a lower pressure RA to a higher pressure LA), in Fontan circulation, by definition, due to the absence of a pumping chamber to the PA, RA pressure is greater than LA pressure, permitting flow through the lungs. Ao, aorta; IVC, inferior vena cava; LA, left atrium; LV, left ventricle; PA, pulmonary arteries; PV, pulmonary veins; SVC, superior vena cava, * Fontan baffle. ***B.*** Diagrammatic representation of the location of various types of Fontan operations. *(Part B used with permission from Emily Flynn McIntosh, illustrator.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

UNREPAIRED CYANOTIC CHD

Eisenmenger Syndrome

ES is felt to be the consequence of a long-standing high volume or pressurized left-to-right shunt in which excessive blood flow to the pulmonary vasculature leads to severely increased pulmonary vascular resistance that eventually results in reversal of the shunt, creating bidirectional or right-to-left flow. ES is a multiple-organ condition and may occur with any CHD with an initial left-to-right shunt. The natural history of ES is variable, and although there is significant morbidity, in general, adults with ES appear to survive longer than those with other forms of pulmonary arterial hypertension. Medical care recommendations have included sustaining adequate hydration, avoiding and treating anemia including iron supplementation when appropriate, and anticoagulation (although this remains controversial due to predisposition to bleeding and occurrence of clinical

hemoptysis, which has frequently been associated with pulmonary vascular thrombosis). Elevation of hematocrit above that considered appropriate for the degree of cyanosis can be managed in symptomatic patients by hydration alone, or on occasion by performing phlebotomy with isovolumic replenishment. Routine phlebotomy in the asymptomatic adult with ES is contraindicated. Appropriate optimization of iron stores has been demonstrated to improve quality of life and functional performance in iron deficient adults with ES. Contraception for women with ES who are of childbearing age is strongly recommended, avoiding use of estrogen, which may be thrombogenic. Pregnancy is contraindicated in these women due to the high risk of maternal mortality.

Recent evidence suggests that the use of selective pulmonary vasodilators, such as bosentan or sildenafil, may be efficacious in ES. Select patients may be candidates for combined heart–lung transplantation or preferably lung transplantation with concomitant repair of the intracardiac defect, if feasible.

Global Considerations

As survival patterns improve for all medically complex patients, the internist and general practitioner are faced with particular challenges and dilemmas; foremost is accrual of sufficient knowledge and competency so as to be able both to engage in patient care provision as well as to seek greater expertise, guidance, and support, when such is appropriate. Across the globe, lifelong care for adults with CHD typifies this growing demand. Care for adults with CHD within medical care centers that contain an ACHD specialty care program has been associated with improved overall survival. However, current analyses suggest that the majority of adults with CHD seek and receive their medical care outside of such ACHD specialty care centers and within the hands of the general practitioner, internist and cardiologist. Under a surface of adaptability and determination, adults with CHD present a wide spectrum of cognitive and functional performance, multiple organ system comorbidities, abnormalities of systemic and pulmonary vasculature, and a near universal presence of heart failure of one stage or another, all over a lifetime. It appears incumbent on the ACHD specialist and ACHD specialty care centers to serve as a hub for partnering practitioners, encouraging engagement to the level of highest competencies, and providing education, oversight, and support, so as to achieve optimal outcomes.

FURTHER READING

Gilboa SM et al: Congenital Heart Defects in the United States: Estimating the magnitude of the affected population in 2010. Circulation 134:101, 2016. [PubMed: 27382105]
Gurvitz M et al: Emerging Research Directions in Adult Congenital Heart Disease: A report from an NHLBI/ACHA Working Group. J Am Coll Cardiol 67:1956, 2016. [PubMed: 27102511]
Regitz-Zagrosek V et al: ESC guidelines on the management of cardiovascular diseases during pregnancy: The Task Force on the Management of Cardiovascular Diseases during Pregnancy of the European Society of Cardiology (ESC). Eur Heart J 32:3147, 2011. [PubMed: 21873418]
Warnes CA et al: ACC/AHA 2008 guidelines for the management of adults with congenital heart disease: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (writing committee to develop guidelines on the management of adults with congenital heart disease). Circulation 118:e714, 2008. [PubMed: 18997169]
Webb G et al: Specialized adult congenital heart care saves lives. Circulation 129:1795, 2014. [PubMed: 24589850]
Webb G et al: The care of adults with congenital heart disease across the globe: Current assessment and future perspective: A position statement from the International Society for Adult Congenital Heart Disease (ISACHD). Int J Cardiol 195:326, 2015. [PubMed: 26056966]

Chapter 265: Pericardial Disease

Eugene Braunwald

NORMAL FUNCTIONS OF THE PERICARDIUM

The normal pericardium is a double-layered sac; the visceral pericardium is a serous membrane that is separated from the fibrous parietal pericardium by a small quantity (15–50 mL) of fluid, an ultrafiltrate of plasma. The normal pericardium, by exerting a restraining force, prevents sudden dilation of the cardiac chambers, especially the right atrium and ventricle, during exercise and with hypervolemia. It also restricts the anatomic position of the heart, and probably retards the spread of infections from the lungs and pleural cavities to the heart. Nevertheless, *total* absence of the pericardium, either congenital or after surgery, does not produce obvious clinical disease. In *partial* left pericardial defects, the main pulmonary artery and left atrium may bulge through the defect; very rarely, herniation and subsequent strangulation of the left atrium may cause sudden death.

ACUTE PERICARDITIS

Acute pericarditis, by far the most common pathologic process involving the pericardium **(Table 265-1)**, has four principal diagnostic features:

TABLE 265-1 Classification of Pericarditis

Clinical Classification

I. Acute pericarditis (<6 weeks)
  A. Fibrinous
  B. Effusive (serous or sanguineous)
II. Subacute pericarditis (6 weeks to 6 months)
  A. Effusive-constrictive
  B. Constrictive
III. Chronic pericarditis (>6 months)
  A. Constrictive
  B. Adhesive (nonconstrictive)

Etiologic Classification

I. Infectious pericarditis
  A. Viral (coxsackievirus A and B, echovirus, herpesviruses, mumps, adenovirus, hepatitis, HIV)
  B. Pyogenic (pneumococcus, *Streptococcus, Staphylococcus, Neisseria, Legionella, Chlamydia*)
  C. Tuberculous
  D. Fungal (histoplasmosis, coccidioidomycosis, *Candida*, blastomycosis)
  E. Other infections (syphilitic, protozoal, parasitic)
II. Noninfectious pericarditis
  A. Acute idiopathic
  B. Renal failure
  C. Neoplasia
    1. Primary tumors (benign or malignant, mesothelioma)
    2. Tumors metastatic to pericardium (lung and breast cancer, lymphoma, leukemia)
  D. Trauma (penetrating chest wall, nonpenetrating)
  E. Aortic dissection (with leakage into pericardial sac)
  F. Acute myocardial infarction
  G. Postirradiation
  H. Familial Mediterranean fever
  I. Familial pericarditis
    1. Mulibrey nanism[a]
  J. Metabolic (myxedema, cholesterol)
III. Pericarditis presumably related to hypersensitivity or autoimmunity
  A. Rheumatic fever
  B. Collagen vascular disease (systemic lupus erythematosus, rheumatoid arthritis, ankylosing spondylitis, scleroderma, acute rheumatic fever, granulomatosis with polyangiitis [Wegener's])
  C. Drug-induced (e.g., procainamide, hydralazine, phenytoin, isoniazid, minoxidil, anticoagulants, methysergide)
  D. Postcardiac injury
    1. Postpericardiotomy
    2. Posttraumatic
    3. Postmyocardial infarction (Dressler's syndrome)

[a]An autosomal recessive syndrome characterized by growth failure, muscle hypotonia, hepatomegaly, ocular changes, enlarged cerebral ventricles, mental retardation, ventricular hypertrophy, and chronic

constrictive pericarditis.

1. *Chest pain* is usually present in acute infectious pericarditis and in many of the forms presumed to be related to hypersensitivity, autoimmunity, or of unknown cause (idiopathic). The pain of acute pericarditis is often severe, retrosternal and/or left precordial, and referred to the neck, arms, or left shoulder. Frequently the pain is pleuritic, consequent to accompanying pleural inflammation (i.e., sharp and aggravated by inspiration and coughing), but sometimes it is steady, radiates to the trapezius ridge, or into either arm, and resembles that of myocardial ischemia; therefore, confusion with acute myocardial infarction (AMI) is common. Characteristically, pericardial pain may be intensified by lying supine, and relieved by sitting up and leaning forward **(Chap. 11)**. Pain is often absent in slowly developing tuberculous, postirradiation, neoplastic, and uremic pericarditis.

   The differentiation of AMI from acute pericarditis may be challenging when, with the latter, serum biomarkers of myocardial damage such as troponin and creatine kinase-MB rise, presumably because of concomitant involvement of the epicardium in the inflammatory process (an epi-myocarditis) with resulting myocyte necrosis. However, these elevations, if they occur, are quite modest compared to those in AMI, given the extensive electrocardiographic ST-segment elevation in pericarditis. This dissociation is useful in differentiating between these conditions.

2. A *pericardial friction rub* is audible at some point in the illness in about 85% of patients with acute pericarditis, it may have up to three components per cardiac cycle, is rasping, scratching, or grating **(Chap. 234)**. It is heard most frequently at end expiration with the patient upright and leaning forward.

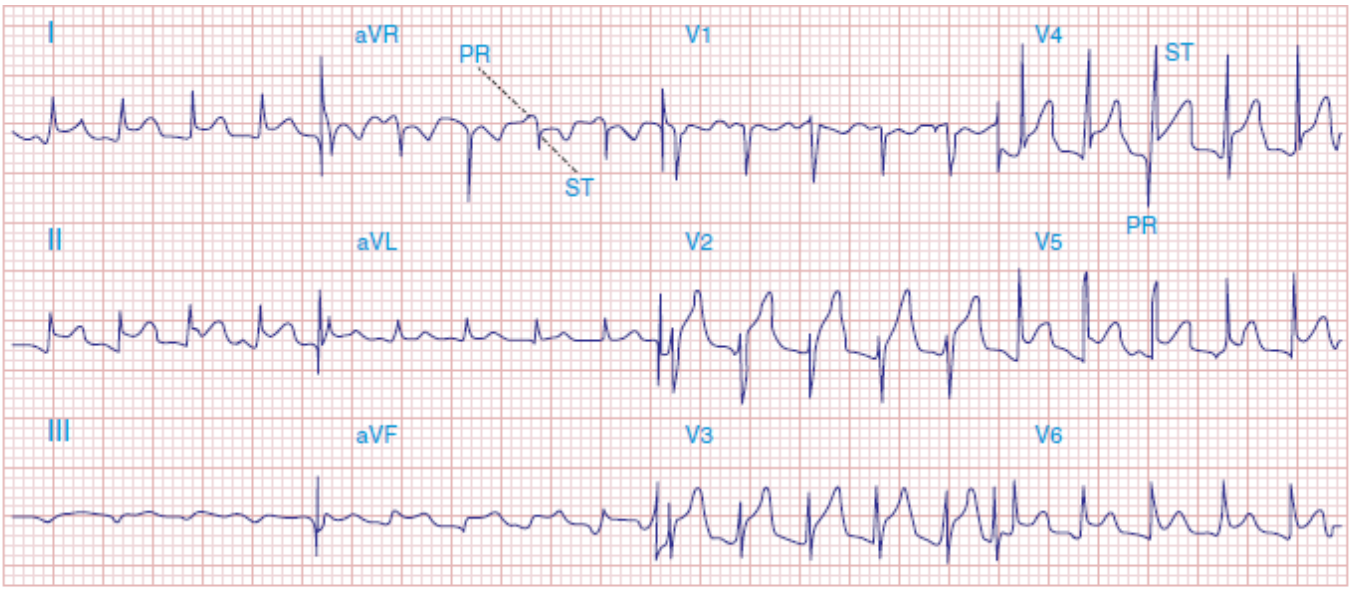
3. The *electrocardiogram* (ECG) in acute pericarditis without massive effusion usually displays changes secondary to acute subepicardial inflammation **(Fig. 265-1A)**. It typically evolves through four stages. In stage 1, there is widespread elevation of the ST segments, often with upward concavity, involving two or three standard limb leads and V_2–V_6, with reciprocal depressions only in aVR and sometimes V_1. Also, there is depression of the PR segment below the TP segment, reflecting atrial involvement. Usually there are no significant changes in QRS complexes, unless a large pericardial effusion develops (see below). After several days, the ST segments return to normal (stage 2), and only then, or even later, do the T waves become inverted (stage 3). Weeks or months after the onset of acute pericarditis, the ECG returns to normal (stage 4). In contrast, in AMI, ST elevations are upwardly convex, and reciprocal depression is usually more prominent; these changes may return to normal within a day or two. Q waves may develop, with loss of R-wave amplitude, and T-wave inversions; these changes are usually seen within hours *before* the ST segments have become isoelectric **(Chaps. 268 and 269)**.

4. *Pericardial effusion* is usually associated with pain and/or the ECG changes mentioned above, and if the effusion is large with electrical alternans **(Fig. 265-1B)**. Pericardial effusion is especially important clinically when it develops within a relatively short time because it may lead to cardiac tamponade (see below). Differentiation from cardiac enlargement on physical examination may be difficult, but heart sounds may be fainter with large pericardial effusion. The friction rub and the apex impulse may disappear. The base of the left lung may be compressed by pericardial fluid, producing *Ewart's sign*, a patch of dullness and increased fremitus beneath the angle of the left scapula. The chest roentgenogram may show enlargement of the cardiac silhouette, with a "water bottle" configuration, but may be normal in patients with small effusions.

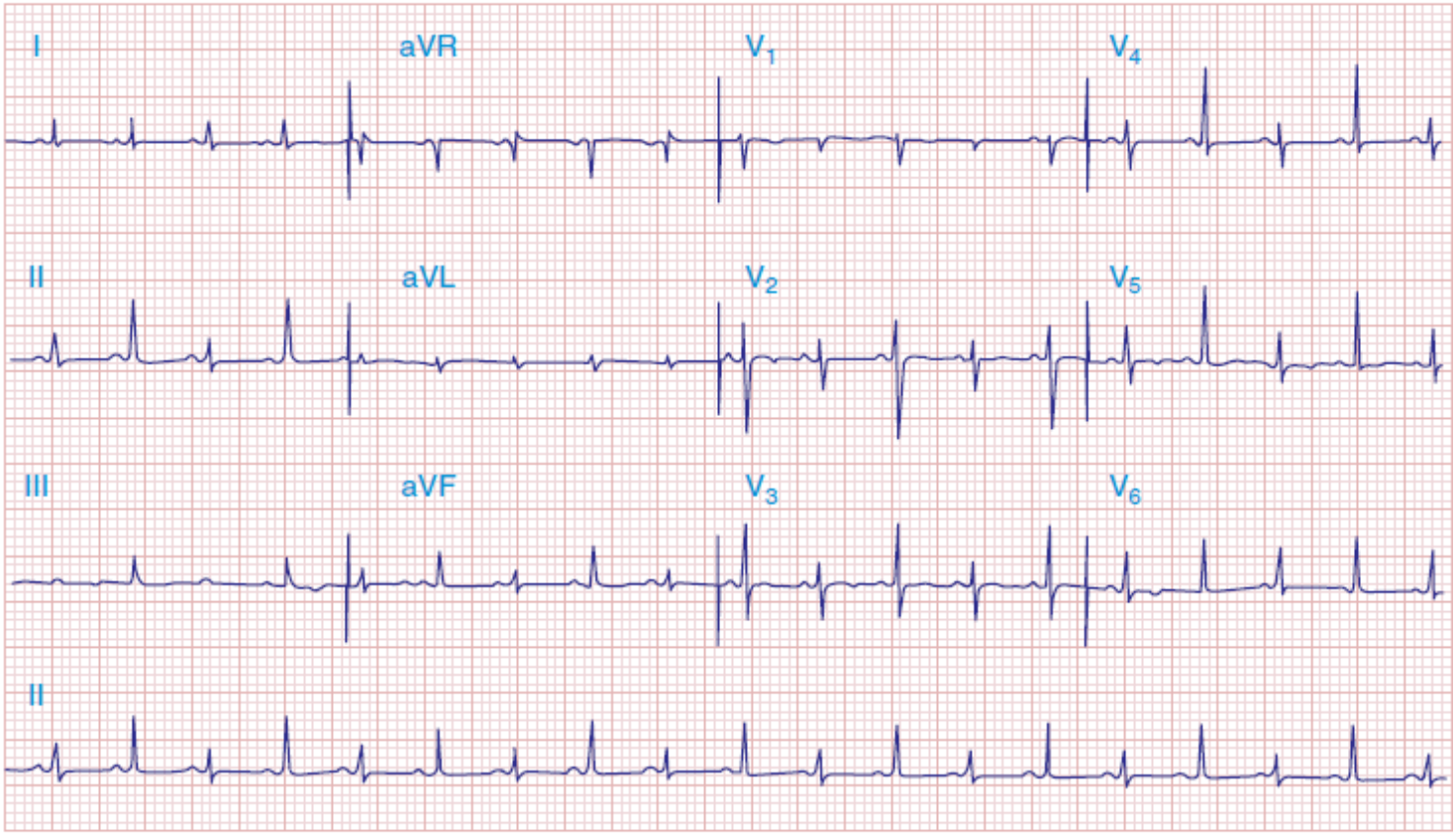
FIGURE 265-1

A. Acute pericarditis. There are diffuse ST-segment elevations in leads I, II, aVF, and V_2–V_6). There is PR-segment depression due to a concomitant atrial injury current. ***B.*** Electrical alternans. This tracing

was obtained from a patient with a large pericardial effusion with cardiac tamponade.



A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

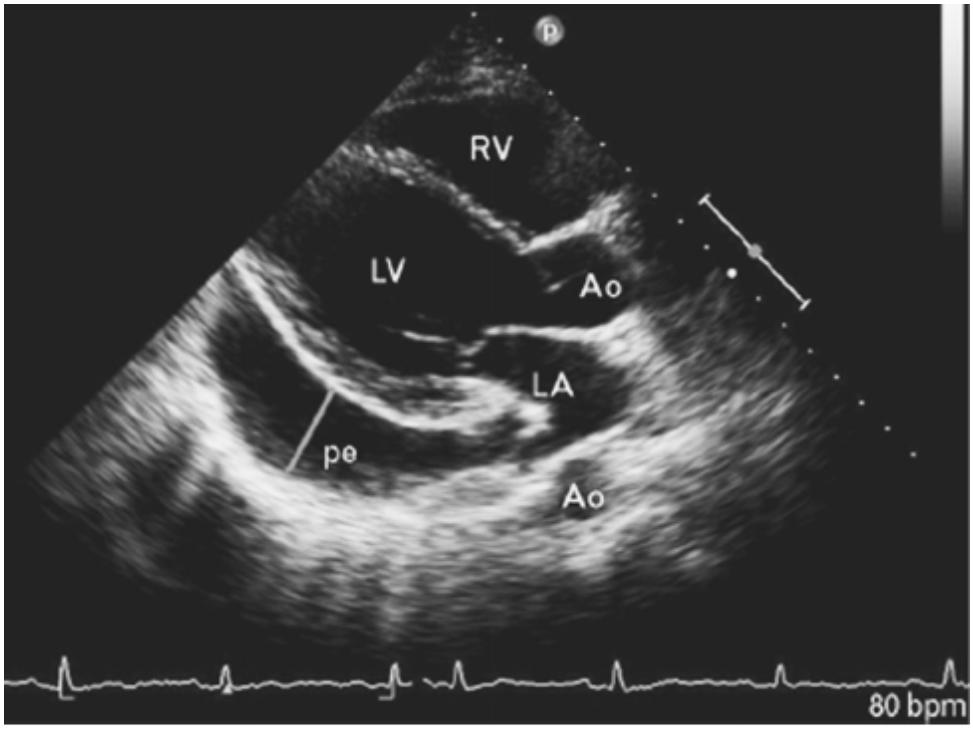
Diagnosis

Echocardiography **(Chap. 236)** is the most widely used imaging technique. It is sensitive, specific, simple, noninvasive, may be performed at the bedside, and allows localization and estimation of the quantity of pericardial fluid. The presence of pericardial fluid is recorded by two-dimensional

transthoracic echocardiography as a relatively echo-free space between the posterior pericardium and left ventricular epicardium and/or as a space between the anterior right ventricle and the parietal pericardium just beneath the anterior chest wall **(Fig. 265-2).**

FIGURE 265-2

Two-dimensional echocardiogram in lateral view in a patient with a large pericardial effusion. Ao, aorta; LV, left ventricle; pe, pericardial effusion; RV, right ventricle. *(From M Imazio: Curr Opin Cardiol 27:308, 2012.)*

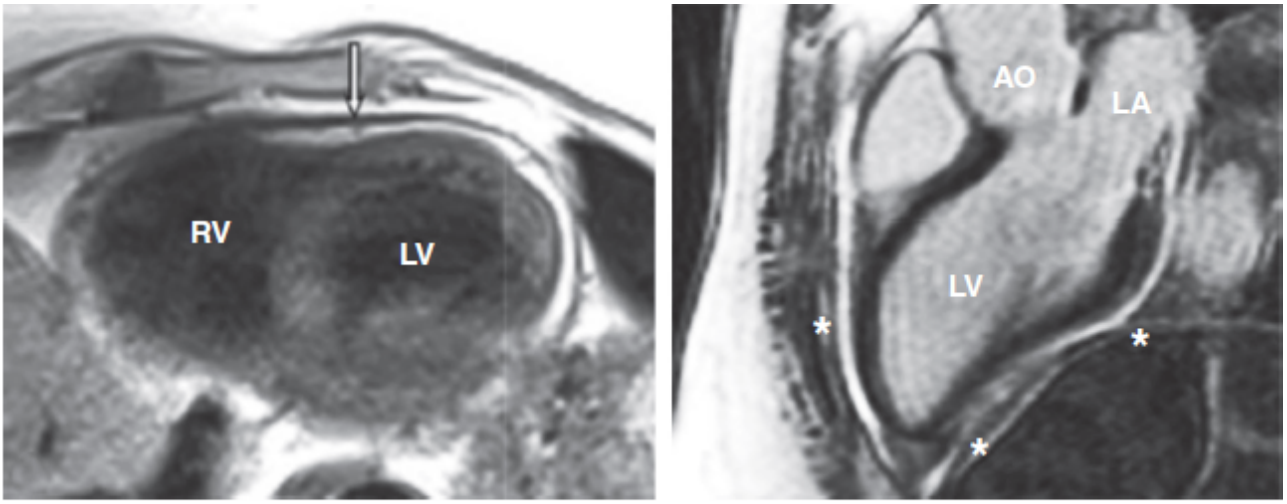


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The diagnosis of pericardial fluid or thickening may be confirmed by computed tomography (CT) or magnetic resonance imaging (MRI). These techniques may be superior to echocardiography in detecting loculated pericardial effusions, pericardial thickening, and the identification of pericardial masses. MRI is also helpful in detecting pericardial inflammation **(Fig. 265-3)**.

FIGURE 265-3

Pericardial inflammation by cardiac magnetic resonance imaging. *A.* Short axis view. The pericardium is thickened and enhanced on T2 magnetic images. Note thickened white line denoted by arrow. ***B.*** Long axis view. Late gadolinium enhancement of thickened, inflamed pericardium. AO, aorta; LA, left atrium; LV, left ventricle; RV, right ventricle. *(From RY Kwong: Cardiovascular magnetic resonance imaging, in Braunwald's Heart Disease, 10th ed, Mann DL et al [eds]. Philadelphia: Elsevier, 2015, pp 320–40.)*



A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT Acute Pericarditis

There is no specific therapy for acute idiopathic pericarditis, but bed rest and anti-inflammatory treatment with aspirin (2–4 g/d), with nonsteroidal anti-inflammatory drugs (NSAIDs), such as ibuprofen (600–800 mg tid) or indomethacin (25–50 mg tid), and should be administered along with gastric protection (e.g., omeprazole 20 mg/d). In responsive patients, these doses should be continued for 1–2 weeks and then tapered over several weeks. In addition, colchicine (0.5 mg qd [<70 kg] or 0.5 mg bid [>70 kg]), should be administered for 3 months. Colchicine enhances the response to NSAIDs and also aids in reducing the risk of recurrent pericarditis. This drug is concentrated in and interferes with the migration of neutrophils, may cause diarrhea and other gastrointestinal side effects, and is contraindicated in patients with hepatic or renal dysfunction. Glucocorticoids (e.g., prednisone 1 mg/kg per day) usually suppress the clinical manifestations of acute pericarditis in patients who have failed therapy with or do not tolerate NSAIDs and colchicine. However, since they increase the risk of subsequent recurrence, full-dose corticosteroids should be given for only 2–4 days and then tapered. Anticoagulants should be avoided because their use could cause bleeding into the pericardial cavity and tamponade.

In patients with multiple, frequent, and disabling recurrences that continue for more than 2 years, and are not prevented by continuing colchicine and other NSAIDs and are not controlled by glucocorticoids, azathioprine, or anakinra (an IL-1β receptor antagonist) have been reported to be of benefit. Rarely, pericardial stripping may be necessary but this procedure may not always terminate the recurrences.

The majority of patients with acute pericarditis can be managed as outpatients with careful follow-up. However, when specific causes (tuberculosis, neoplastic disease, bacterial infection) are suspected, or if any of the predictors of poor prognosis (fever >38°C, subacute onset, or large pericardial effusion) are present, hospitalization is advisable.

CARDIAC TAMPONADE

The accumulation of fluid in the pericardial space in a quantity sufficient to cause serious obstruction of the inflow of blood into the ventricles results in cardiac tamponade. This complication may be fatal if it is not recognized and treated promptly. The most common causes of tamponade are idiopathic

pericarditis and pericarditis secondary to neoplastic disease, tuberculosis, or bleeding into the pericardial space after leakage from an aortic dissection, cardiac operation, trauma, and treatment with anticoagulants.

The three principal features of tamponade (*Beck's triad*) are hypotension, soft or absent heart sounds, and jugular venous distention with a prominent *x* (early systolic) descent but an absent *y* (early diastolic) descent. The limitations to ventricular filling are responsible for reductions of cardiac output and arterial pressure. The quantity of fluid necessary to produce cardiac tamponade may be as small as 200 mL when the fluid develops rapidly to as much as >2000 mL in slowly developing effusions when the pericardium has had the opportunity to stretch and adapt to an increasing volume.

A high index of suspicion for cardiac tamponade is required because in many instances no obvious cause for pericardial disease is apparent, and this diagnosis should be considered in any patient with otherwise unexplained sudden enlargement of the cardiac silhouette, hypotension, and elevation of jugular venous pressure. There also may be reductions in amplitude of the QRS complexes, and *electrical alternans* of the P, QRS, or T waves should raise the suspicion of cardiac tamponade (Fig. 265-1).

Table 265-2 lists the features that distinguish acute cardiac tamponade from constrictive pericarditis.

TABLE 265-2 Features That Distinguish Cardiac Tamponade from Constrictive Pericarditis and Similar Clinical Disorders

Characteristic	Tamponade	Constrictive Pericarditis	Restrictive Cardiomyopathy	RVMI	Effusive Constrictive Pericarditis
Clinical					
Pulsus paradoxus	+++	+	+	+	+++
Jugular veins					
Prominent *y* descent	–	++	+	+	–
Prominent *x* descent	+++	++	+++	+	+++
Kussmaul's sign	–	+++	+	+++	++
Third heart sound	–	–	+	+	+
Pericardial knock	–	++	–	–	–
Electrocardiogram					
Low ECG voltage	++	++	+	–	+
Electrical alternans	++	–	–	–	+
Echocardiogram					
Thickened pericardium	–	+++	–	–	++
Pericardial calcification	–	++	–	–	–
Pericardial effusion	+++	–	–	–	++
RV size	Usually small	Usually normal	Usually normal	Enlarged	Usually normal
Exaggerated respiratory variation in flow velocity	+++	+++	–	+++	+
CT/MRI					
Thickened pericardium	–	+++	–		++
Equalization of diastolic pressures	+++	+++	–	++	++

Abbreviations: +++, always present; ++, usually present; +, rare; –, absent; DC, diastolic collapse; ECG, electrocardiogram; RV, right ventricle; RVMI, right ventricular myocardial infarction.

Source: Adapted from GM Brockington et al: Cardiol Clin 8:645, 1990, with permission.

Paradoxical Pulse

This important clue to the presence of cardiac tamponade consists of a greater than normal (10 mmHg) inspiratory decline in systolic arterial pressure. When severe it may be detected by palpating weakness or even disappearance of the arterial pulse during inspiration, but usually sphygmomanometric measurement of systolic pressure during slow respiration is required.

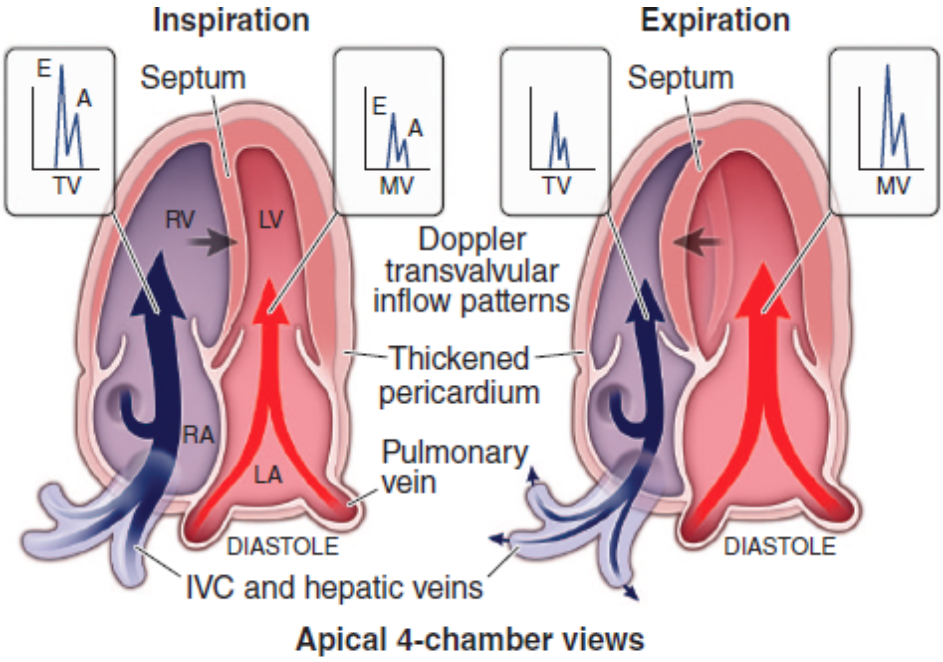
Because both ventricles share a tight incompressible covering, i.e., the pericardial sac, the inspiratory enlargement of the right ventricle causes leftward bulging of the interventricular septum, compresses and reduces left ventricular volume; stroke volume, and arterial systolic pressure. Paradoxical pulse also occurs in approximately one-third of patients with constrictive pericarditis (see below), and in some cases of hypovolemic shock, acute and chronic obstructive airway disease, and pulmonary embolism. Right ventricular infarction **(Chap. 269)** may resemble cardiac tamponade with hypotension, elevated jugular venous pressure, an absent *y* descent in the jugular venous pulse, and, occasionally, a paradoxical pulse (Table 265-2).

Diagnosis

Because immediate treatment of cardiac tamponade may be lifesaving, prompt establishment of the diagnosis, usually by echocardiography, should be undertaken. When pericardial effusion causes tamponade, Doppler ultrasound shows that tricuspid and pulmonic valve flow velocities increase markedly during inspiration, whereas pulmonic vein, mitral, and aortic flow velocities diminish (as in constrictive pericarditis, see below) **(Fig. 265-4)**. In tamponade, there is late diastolic inward motion (collapse) of the right ventricular free wall and the right atrium. Transesophageal echocardiography, CT, or cardiac MRI may be necessary to diagnose a loculated effusion responsible for cardiac tamponade.

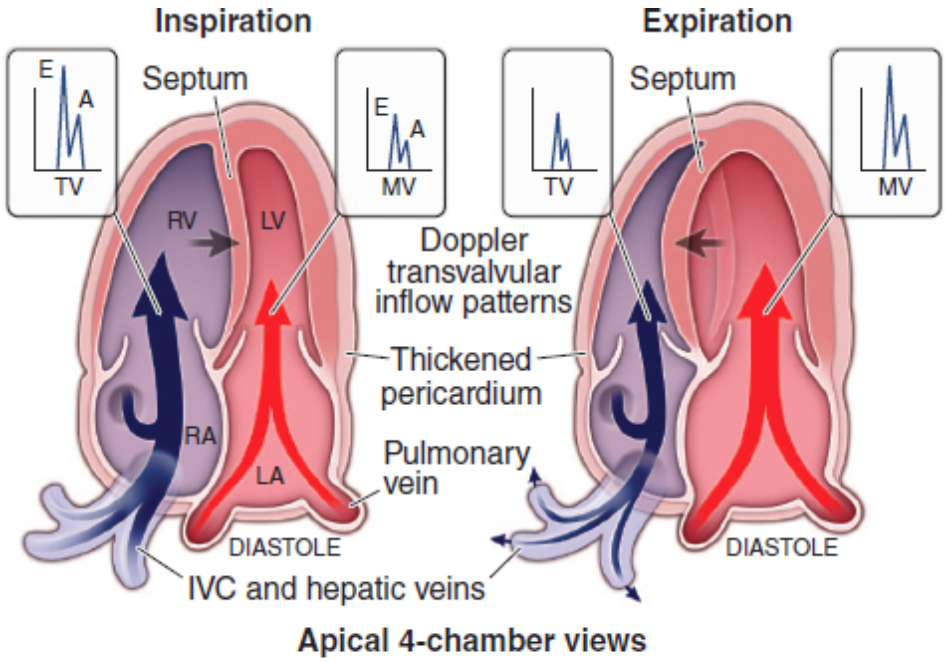
FIGURE 265-4

Constrictive pericarditis. Doppler schema of respirophasic changes in mitral and tricuspid inflow. Reciprocal patterns of ventricular filling are assessed on pulsed Doppler examination of mitral valve (MV) and tricuspid valve (TV) inflow. IVC, inferior vena cava; LA, left atrium; LV, left ventricle; RA, right atrium; RV, right ventricle. *(Courtesy of Bernard E. Bulwer, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

View Full Size|
Favorite Figure|Download Slide (.ppt)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT Cardiac Tamponade

Patients with acute pericarditis should be observed frequently for the development of an effusion. If a large effusion is present, pericardiocentesis should be carried out or the patient watched closely for signs of tamponade with serial echocardiography and monitoring of arterial and venous pressures.

PERICARDIOCENTESIS

If manifestations of tamponade appear, pericardiocentesis using an apical, parasternal, or, most commonly, subxiphoid approach must be carried out at once because if left untreated, tamponade may be rapidly fatal. Whenever possible, this procedure should be carried out under echocardiographic guidance. Intravenous saline may be administered as the patient is being readied for the procedure, but the pericardiocentesis must not be delayed. If possible, intrapericardial pressure should be measured before fluid is withdrawn, and the pericardial cavity should be drained as completely as possible. A small, multiholed catheter may be advanced over the needle inserted into the pericardial cavity and left in place to allow draining of the pericardial space if fluid reaccumulates. Surgical drainage through a limited (subxiphoid) thoracotomy may be required in recurrent tamponade to remove loculated effusions, and/or when it is necessary to obtain tissue for diagnosis.

Pericardial fluid obtained from an effusion may have the physical characteristics of an exudate. In developed nations, bloody fluid is most commonly due to neoplasm, renal failure, or after cardiac injury. In developing nations, tuberculosis, often associated with HIV infection, may also cause exudative and/or bloody effusion.

The pericardial fluid should be analyzed for red and white blood cells and cytology for neoplastic cells. Cultures should be obtained. The presence of DNA of *Mycobacterium tuberculosis* determined by the polymerase chain reaction strongly supports the diagnosis of tuberculous pericarditis **(Chap. 173)**.

VIRAL OR IDIOPATHIC ACUTE PERICARDITIS

In many instances, acute pericarditis occurs in association with or following illnesses of known or presumed viral origin and probably is caused by the same agent. There may be an antecedent infection of the respiratory tract, but viral isolation and serologic studies are usually negative. In some cases coxsackievirus A or B or the virus of influenza, echovirus, mumps, herpes simplex, chickenpox, adenovirus, or cytomegalovirus has been isolated from pericardial fluid and/or appropriate elevations in viral antibody titers have been observed. Frequently, a viral cause cannot be established, and the term *idiopathic acute pericarditis* is appropriate.

Viral or idiopathic acute pericarditis occurs at all ages but is most common in young adult males, and is often associated with pleural effusion and pneumonitis. The almost simultaneous development of fever and precordial pain, often 10–12 days after a presumed viral illness, constitutes an important feature in the differentiation of acute pericarditis from AMI, in which chest pain precedes fever. The constitutional symptoms are usually mild to moderate, and a pericardial friction rub is often audible. The disease ordinarily runs its course in a few days to 4 weeks. Elevations of C-reactive protein and of the white blood cell count are common. The ST-segment alterations in the ECG usually disappear after 1 or more weeks, but the abnormal T waves may persist for several years and be a source of confusion in persons without a clear history of pericarditis. Accumulation of some pericardial fluid is common, and both tamponade and constrictive pericarditis are possible, but infrequent, complications.

The most frequent complication is recurrent (relapsing) pericarditis, which occurs in about one-fourth of patients with acute idiopathic pericarditis. In a smaller number, there are multiple recurrences.

Postcardiac Injury Syndrome

Acute pericarditis may appear in a variety of circumstances that have one common feature—previous injury to the myocardium with blood in the pericardial cavity. The syndrome may develop after a cardiac operation (postpericardiotomy syndrome), after blunt or penetrating cardiac trauma **(Chap. S8)**, or after perforation of the heart with a catheter; rarely, it follows AMI.

The clinical picture mimics acute viral or idiopathic pericarditis. The principal symptom is the pain of acute pericarditis, which usually develops 1–4 weeks after the cardiac injury. Recurrences are common and may occur up to 2 years or more following the injury. Fever, pleuritis, and pneumonitis are accompanying features, and the illness usually subsides in 1 or 2 weeks. The pericarditis may be of the fibrinous variety, or it may be a pericardial effusion, which is often serosanguineous and rarely causes tamponade. ECG changes typical of acute pericarditis may also occur. This syndrome is probably the result of a hypersensitivity reaction to antigen(s) that originate from injured myocardial tissue and/or pericardium.

Often no treatment is necessary aside from aspirin and analgesics. When the illness is severe or followed by a series of disabling recurrences, therapy with another NSAID, colchicine, or a glucocorticoid, such as described for treatment of acute pericarditis, is usually effective.

DIFFERENTIAL DIAGNOSIS

Because there is no specific test for *acute idiopathic pericarditis*, the diagnosis is one of exclusion. Consequently, all other disorders that may be associated with acute fibrinous pericarditis must be considered. A common diagnostic error is mistaking acute viral or idiopathic pericarditis for AMI and vice versa.

Pericarditis secondary to postcardiac injury is differentiated from acute idiopathic pericarditis chiefly by timing. If it occurs within a few days or weeks of a chest blow, a cardiac perforation, a cardiac operation, or an AMI, the two are probably related.

It is important to distinguish *pericarditis due to collagen vascular disease* from acute idiopathic pericarditis. Most important in the differential diagnosis is the pericarditis due to systemic lupus erythematosus (SLE; **Chap. 349**) or drug-induced (procainamide or hydralazine) lupus. When pericarditis occurs in the absence of any obvious underlying disorder, the diagnosis of SLE may be suggested by a rise in the titer of antinuclear antibodies. Acute pericarditis is an occasional complication of *rheumatoid arthritis*, *scleroderma*, and *polyarteritis nodosa*, and other evidence of these diseases is usually obvious.

Pyogenic (purulent) pericarditis is usually secondary to cardiothoracic operations, by extension of infection from the lungs or pleural cavities, from rupture of the esophagus into the pericardial sac, or from rupture of a valvular ring abscess in a patient with infective endocarditis. It may also complicate the viral, bacterial, mycobacterial, and fungal infections that occur with HIV infection. It is generally accompanied by fever, chills, septicemia, and evidence of infection elsewhere and generally has a poor prognosis. The diagnosis is made by examination of the pericardial fluid. It requires immediate drainage as well as vigorous antibiotic treatment.

Pericarditis of renal failure (*uremic pericarditis)* occurs in up to one-third of patients with severe renal dysfunction, and is also seen in patients undergoing chronic dialysis who have normal levels of blood urea (*dialysis-associated pericarditis*). These two forms of pericarditis may be fibrinous and are generally associated with serosanguineous effusions. A pericardial friction rub is common, but pain is usually absent or mild. Treatment with an NSAID and intensification of dialysis are usually adequate. Occasionally, tamponade occurs and pericardiocentesis is required. When the pericarditis of renal failure is recurrent or persistent, a pericardial window should be created or pericardiectomy may be necessary.

Pericarditis due to *neoplastic diseases* results from extension or invasion of metastatic tumors (most commonly carcinoma of the lung and breast, malignant melanoma, lymphoma, and leukemia) to the pericardium. The pain of pericarditis, tamponade, and atrial arrhythmias are complications that occur occasionally. Diagnosis is made by pericardial fluid cytology or pericardial biopsy. *Mediastinal irradiation* for neoplasm may cause acute pericarditis and/or chronic constrictive pericarditis. Unusual causes of acute pericarditis include syphilis, fungal infection (histoplasmosis, blastomycosis, aspergillosis, and candidiasis), and parasitic infestation (amebiasis, toxoplasmosis, echinococcosis, and trichinosis) (Table 265-1).

CHRONIC PERICARDIAL EFFUSIONS

Chronic pericardial effusions are sometimes encountered in patients without an antecedent history of acute pericarditis. They may cause few symptoms *per se*, and their presence may be detected by finding an enlarged cardiac silhouette on a chest roentgenogram. *Tuberculosis* and myxedema may be causal. Neoplasms, SLE, rheumatoid arthritis, mycotic infections, radiation therapy to the chest, and chylopericardium may also cause chronic pericardial effusion and should be considered and specifically sought in such patients. Aspiration and analysis of the pericardial fluid are often helpful in diagnosis. Pericardial fluid should be analyzed as described under pericardiocentesis. Grossly sanguineous pericardial fluid results most commonly from a neoplasm, tuberculosis, renal failure, or slow leakage from an aortic dissection. Pericardiocentesis may resolve large effusions, but pericardiectomy may be required in patients with recurrence. Intrapericardial instillation of sclerosing agents may be used to prevent reaccumulation of fluid.

CHRONIC CONSTRICTIVE PERICARDITIS

This disorder results when the healing of an acute fibrinous or serofibrinous pericarditis or the resorption of a chronic pericardial effusion is followed by obliteration of the pericardial cavity with the formation of granulation tissue. The latter gradually contracts and forms a firm scar encasing the heart,

which may become calcified. In developing nations, a high percentage of cases are of tuberculous origin, but this is now an uncommon cause in North America or Western Europe. Chronic constrictive pericarditis may follow acute or relapsing viral or idiopathic pericarditis, trauma with organized blood clot, or cardiac surgery of any type, or results from mediastinal irradiation, purulent infection, histoplasmosis, neoplastic disease (especially breast cancer, lung cancer, and lymphoma), rheumatoid arthritis, SLE, or chronic renal failure treated by chronic dialysis. In many patients, the cause of the pericardial disease is undetermined, and in these patients an asymptomatic or forgotten bout of viral pericarditis, idiopathic or acute, may have been the inciting event.

The basic physiologic abnormality in patients with chronic constrictive pericarditis is the inability of the ventricles to fill because of the limitations imposed by the rigid, thickened pericardium. Ventricular filling is unimpeded during early diastole but is reduced abruptly when the elastic limit of the pericardium is reached, whereas in cardiac tamponade, ventricular filling is impeded throughout diastole. In both conditions, ventricular end-diastolic and stroke volumes are reduced and the end-diastolic pressures in both ventricles and the mean pressures in the atria, pulmonary veins, and systemic veins are all elevated to similar levels (i.e., within 5 mmHg of one another). Despite these hemodynamic changes, systolic function may be normal or only slightly impaired at rest. However, in advanced cases, the fibrotic process may extend into the myocardium and cause myocardial scarring and atrophy, and venous congestion may then be due to the combined effects of the pericardial and myocardial lesions.

In constrictive pericarditis, the right and left atrial pressure pulses display an M-shaped contour, with prominent *x* and *y* descents. The *y* descent, which is absent or diminished in cardiac tamponade, is the most prominent deflection in constrictive pericarditis; it reflects rapid early filling of the ventricles. The *y* descent is interrupted by a rapid rise in atrial pressure during early diastole, when ventricular filling is impeded by the constricting pericardium. These characteristic changes are transmitted to the jugular veins, where they may be recognized by inspection. In constrictive pericarditis, the ventricular pressure pulses in both ventricles exhibit characteristic “square root” signs during diastole. These hemodynamic changes, although characteristic, are not pathognomonic of constrictive pericarditis and may also be observed in restrictive cardiomyopathies (**Chap. 254**, **Table 254-2**).

CLINICAL AND LABORATORY FINDINGS

Weakness, fatigue, weight gain, increased abdominal girth, abdominal discomfort, and edema are common. The patient often appears chronically ill, and in advanced cases, anasarca, skeletal muscle wasting, and cachexia may be present. Exertional dyspnea is common, and orthopnea may occur, although it is usually not severe. The cervical veins are distended and may remain so even after intensive diuretic treatment, and venous pressure may fail to decline during inspiration (*Kussmaul’s sign*). The latter is common in chronic pericarditis but may also occur in tricuspid stenosis, right ventricular infarction, and restrictive cardiomyopathy.

The pulse pressure is normal or reduced. A paradoxical pulse can be detected in about one-third of cases. Congestive hepatomegaly is pronounced and may impair hepatic function and cause jaundice; ascites is common and is usually more prominent than dependent edema. Pleural effusions and splenomegaly may also be present. The apical pulse is reduced and may retract in systole (*Broadbent’s sign*). The heart sounds may be distant; an early third heart sound (i.e., a pericardial knock) occurring at the cardiac apex with the abrupt cessation of ventricular filling is often conspicuous.

The ECG frequently displays low voltage of the QRS complexes and diffuse flattening or inversion of the T waves. Atrial fibrillation is present in about one-third of patients. The *chest roentgenogram* shows a normal or slightly enlarged heart. Pericardial calcification is most common in tuberculous pericarditis. Pericardial calcification may, however, occur in the absence of constriction, and constriction may occur without calcification.

Inasmuch as the common physical signs of cardiac disease (murmurs, cardiac enlargement) may be inconspicuous or absent in chronic constrictive pericarditis, hepatic enlargement, and dysfunction associated with jaundice and intractable ascites may lead to a mistaken diagnosis of hepatic cirrhosis. This error can be avoided if the neck veins are inspected and found to be distended.

The transthoracic *echocardiogram* often shows pericardial thickening, dilation of the inferior vena cava and hepatic veins, and a sharp halt to rapid left ventricular filling in early diastole, with normal ventricular systolic function and flattening of the left ventricular posterior wall. There is a distinctive pattern of transvalvular flow velocity on Doppler echocardiography **(Fig. 265-4)**. During inspiration, there is an exaggerated reduction in blood flow velocity in the pulmonary veins and across the mitral valve and a leftward shift of the ventricular septum; the opposite occurs during expiration. Diastolic flow velocity in the inferior vena cava into the right atrium and across the tricuspid valve increases in an exaggerated manner during inspiration and declines during expiration. However, echocardiography cannot definitively establish or exclude the diagnosis of constrictive pericarditis; CT and MRI are more accurate, the latter is useful in evaluating myocardial involvement.

DIFFERENTIAL DIAGNOSIS

Like chronic constrictive pericarditis, cor pulmonale **(Chap. 252)** may be associated with marked systemic venous hypertension, little pulmonary congestion, a heart that is not enlarged, and a paradoxical pulse. However, in cor pulmonale, advanced parenchymal pulmonary disease is usually apparent and venous pressure *falls* during inspiration (i.e., Kussmaul's sign is negative). *Tricuspid stenosis* **(Chap. 261)** may also simulate chronic constrictive pericarditis with congestive hepatomegaly, splenomegaly, ascites, and venous distention. However, the characteristic murmur and that of accompanying mitral stenosis are usually present.

Because it can be corrected surgically, it is important to distinguish chronic constrictive pericarditis from restrictive cardiomyopathy **(Chap. 254)**, which has a similar physiologic abnormality (i.e., restriction of ventricular filling). The differentiating features are summarized in Table 265-2. When a patient has progressive, disabling, and unresponsive congestive heart failure and displays any of the features of constrictive heart disease, Doppler echocardiography to record respiratory effects on transvalvular flow (Fig. 265-4) and an MRI or CT scan should be obtained to detect or exclude constrictive pericarditis, because the latter is usually correctable.

TREATMENT

TREATMENT Constrictive Pericarditis

Pericardial resection is the only definitive treatment of constrictive pericarditis and should be as complete as possible. Dietary sodium restriction and diuretics are useful during preoperative preparation. Coronary arteriography should be carried out preoperatively in patients aged >50 years to exclude unsuspected accompanying coronary artery disease. The benefits derived from cardiac decortication are usually progressive over a period of months. The risk of this operation depends on the extent of penetration of the myocardium by the fibrotic and calcific process, the severity of myocardial atrophy, the extent of secondary impairment of hepatic and/or renal function, and the patient's general condition. Operative mortality is in the range of 5–10% even in experienced centers; the patients with the most severe disease, especially secondary to radiation therapy, are at highest risk. Therefore, surgical treatment should, if possible, be carried out as early as possible.

Subacute Effusive-Constrictive Pericarditis

This form of pericardial disease is characterized by the combination of a tense effusion in the pericardial space and constriction of the heart by thickened pericardium. As such, it shares a number of features with both chronic pericardial effusion producing cardiac compression and with pericardial

constriction. It may be caused by tuberculosis (see below), multiple attacks of acute idiopathic pericarditis, radiation, traumatic pericarditis, renal failure, scleroderma, and neoplasms. The heart is generally enlarged, and a paradoxical pulse is usually present. After pericardiocentesis, the physiologic findings may change from those of cardiac tamponade to those of pericardial constriction. Furthermore, the intrapericardial pressure and the central venous pressure may decline, but not to normal. The diagnosis can be established by pericardiocentesis followed by pericardial biopsy. Wide excision of both the visceral and parietal pericardium is usually effective therapy.

Tuberculous Pericardial Disease

This chronic infection is a common cause of chronic pericardial effusion, especially in the developing world where active tuberculosis and HIV are endemic. Tuberculous pericarditis may present as pericardial effusion, chronic constrictive pericarditis, or subacute effusive constrictive pericarditis (see above). The clinical picture is that of a chronic, systemic illness in a patient with pericardial effusion. It is important to consider this diagnosis in a patient with known tuberculosis, with HIV, and with fever, chest pain, weight loss, and enlargement of the cardiac silhouette of undetermined origin. If the etiology of chronic pericardial effusion remains obscure despite detailed analysis including culture of the pericardial fluid, a pericardial biopsy, preferably by a limited thoracotomy, should be performed. If definitive evidence is still lacking but the specimen shows granulomas with caseation, antituberculous chemotherapy **(Chap. 173)** is indicated.

If the biopsy specimen shows a thickened pericardium after 2–4 weeks of antituberculous therapy, pericardiectomy should be carried out to prevent the development of constriction. Tubercular cardiac constriction should be treated surgically while the patient is receiving antituberculous chemotherapy.

FURTHER READING

Adler Y et al: 2015 ESC guidelines for the diagnosis and management of pericardial diseases. Eur Heart J 36:2921, 2015. [PubMed: 26320112]
Alraies MC et al: Usefulness of cardiac magnetic resonance-guided management in patients with recurrent pericarditis. Am J Cardiol 115:542, 2015. [PubMed: 25547939]
Cremer PC et al: Complicated pericarditis: Understanding risk factors and pathophysiology to inform imaging and treatment. J Am Coll Cardiol 68:2311, 2016. [PubMed: 27884251]
Garcia MJ: Constrictive pericarditis versus restrictive cardiomyopathy? J Am Coll Cardiol 67:2061, 2016. [PubMed: 27126534]
Imazio M et al: Efficacy and safety of colchicine for treatment of multiple recurrences of pericarditis (CORP-2); a multicentre, double-blind, placebo-controlled, randomised trial. Lancet 383:2232, 2014. [PubMed: 24694983]
LeWinter MM: Acute pericarditis. N Engl J Med 371:2410, 2014. [PubMed: 25517707]
Lotan D et al: Usefulness of novel immunotherapeutic strategies for idiopathic recurrent pericarditis. Am J Cardiol 117:861, 2016. [PubMed: 26742476]
Vistarini N et al: Pericardiectomy for constrictive pericarditis. Ann Thorac Surg 100:107, 2015.

Chapter 266: Atrial Myxoma and Other Cardiac Tumors

Eric H. Awtry

INTRODUCTION

Cardiac tumors can be broadly classified into those that arise primarily in the heart and those that reflect metastatic disease from a distant primary source. Primary cardiac tumors can be further divided into those that are pathologically benign and those that are malignant. Overall, primary cardiac tumors are relatively uncommon, whereas secondary involvement of the heart or pericardium occurs in as many as 20% of patients with end-stage metastatic cancer. While patients with cardiac tumors may present with a variety of symptoms, many patients are asymptomatic at the time of diagnosis as the tumor may be identified incidentally on imaging studies performed for other reasons. Such findings need to be differentiated from other cardiac masses such as vegetation, thrombus, or myocardial hypertrophy. Echocardiography is usually the initial method of evaluation of cardiac tumors; however, a variety of imaging modalities are now available and a multimodality approach is often necessary for accurate diagnosis and clarification of treatment options **(Table 266-1)**.

TABLE 266-1 Imaging Modalities and Their Utility in the Evaluation of Cardiac Tumors

Modality	Utility in Cardiac Tumor Evaluation
Transthoracic echocardiography (TTE) (including 2-D, 3-D, and contrast)	Assessment of tumor location and size, and its impact on adjacent structures (e.g., valves, pericardium).
Transesophageal echocardiography (TEE)	Improved tumor characterization and spatial resolution compared with TTE. May aid in determining surgical approach.
Cardiac MRI with gadolinium contrast	Improved tissue characterization, definition of tumor size and identification of local invasion when compared with TTE or TEE. May differentiate tumor from thrombus.
Gated cardiac CT	Provides anatomic assessment and tissue characterization of the tumor. Useful when patients cannot tolerate MRI or when MRI is not feasible (e.g., patients with implantable cardiac devices). Allows for better assessment of calcified lesions and evaluation of extra-cardiac tumor involvement.
Nuclear Imaging (including ^{18}F-fluorodeoxyglucose positron emission tomography [FDG-PET])	Definition of extra-cardiac disease. May be useful in diagnosis of certain cardiac tumors (e.g., neuroendocrine tumors) but assessment of smaller tumors may be limited by surrounding myocardial FDG uptake.

PRIMARY TUMORS

Primary tumors of the heart are rare. Approximately three-quarters are histologically benign, and the majority of these tumors are myxomas. Malignant tumors, almost all of which are sarcomas, account for 25% of primary cardiac tumors. All cardiac tumors, regardless of pathologic type, have the potential to cause life-threatening complications. Many tumors are now surgically curable; thus, early diagnosis is imperative.

Clinical Presentation

Cardiac tumors may present with a wide array of cardiac and noncardiac manifestations. These manifestations, which depend in large part on the location and size of the tumor as well as its impact on surrounding cardiac structures, are often nonspecific features of more common forms of heart disease, such as chest pain, syncope, congestive heart failure (CHF), murmurs, arrhythmias, conduction disturbances, and pericardial effusion with or without tamponade. Additionally, embolic phenomena and constitutional symptoms may occur.

Myxoma

Myxomas are the most common type of primary cardiac tumor in adults, accounting for one-third to one-half of all cases at postmortem examination, and approximately three-quarters of the tumors treated surgically. They occur at all ages, most commonly in the third through sixth decades, with a female predilection. Approximately 90% of myxomas are sporadic; the remainder are familial with autosomal dominant transmission. The familial variety often occurs as part of a syndrome complex (Carney complex) that includes (1) myxomas (cardiac, skin, and/or breast), (2) lentigines and/or pigmented nevi, and (3) endocrine overactivity (primary nodular adrenal cortical disease with or without Cushing's syndrome, testicular tumors, and/or pituitary adenomas with gigantism or acromegaly). Certain constellations of findings have been referred to as the *NAME* syndrome (*n*evi, *a*trial myxoma, *m*yxoid neurofibroma, and *e*phelides) or the *LAMB* syndrome (*l*entigines, *a*trial *m*yxoma, and *b*lue nevi), although these syndromes probably represent subsets of the Carney complex. The genetic basis of this complex has not been elucidated completely; however, inactivating mutations in the tumor-suppressor gene *PRKAR1A*, which encodes the protein kinase A type I-α regulatory subunit, have been identified in ~70% of patients with Carney complex.

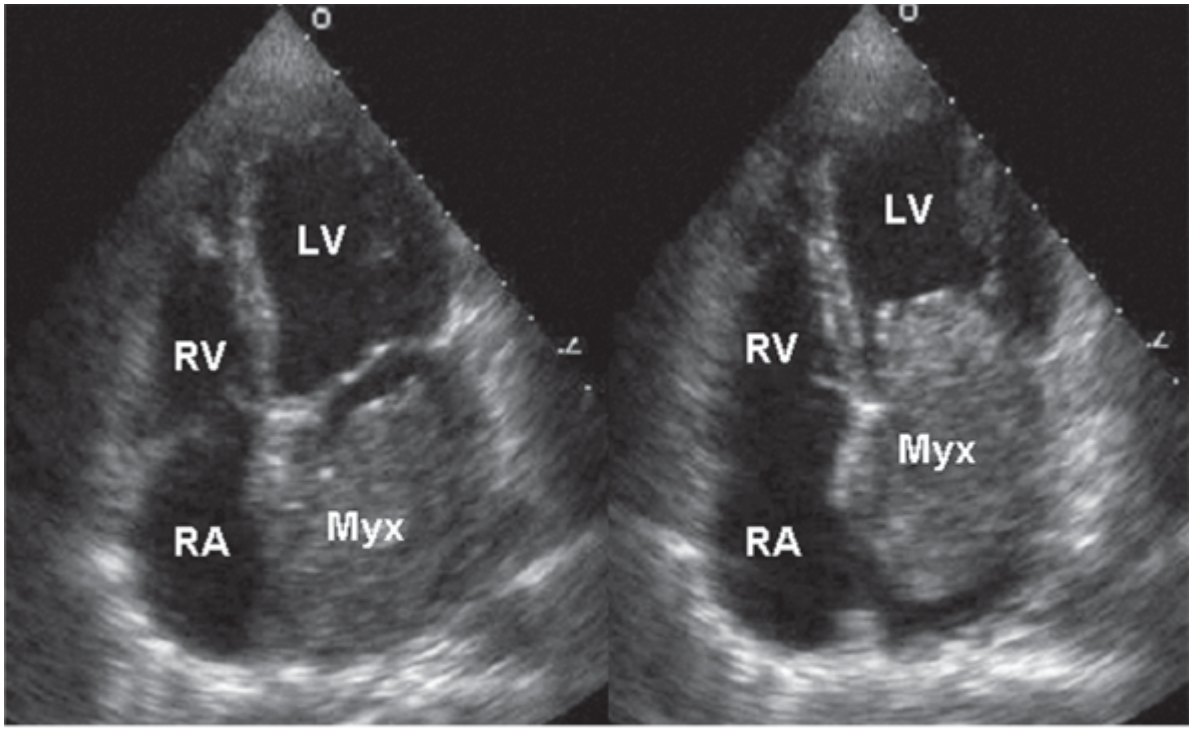
Pathologically, myxomas are gelatinous structures that consist of myxoma cells embedded in a stroma rich in glycosaminoglycans. Most sporadic tumors are solitary, arise from the interatrial septum in the vicinity of the fossa ovalis (particularly in the left atrium), and are often pedunculated on a fibrovascular stalk. In contrast, familial or syndromic tumors tend to occur in younger individuals, are often multiple, may be ventricular in location, and are more likely to recur after initial resection.

Myxomas commonly present with obstructive signs and symptoms. The most common clinical presentation mimics that of mitral valve disease: either stenosis owing to tumor prolapse into the mitral orifice or regurgitation resulting from tumor-induced valvular trauma or distortion. Ventricular myxomas may cause outflow tract obstruction similar to that caused by subaortic or subpulmonic stenosis. The symptoms and signs of myxoma may be sudden in onset or positional in nature, owing to the effects of gravity on tumor position. A characteristic low-pitched sound, a "tumor plop, " may be appreciated on auscultation during early or mid-diastole and is thought to result from the impact of the tumor against the mitral valve or ventricular wall. Myxomas also may present with peripheral or pulmonary embolic phenomenon (resulting from embolization of tumor fragments or tumor-associated thrombus) or with constitutional signs and symptoms, including fever, weight loss, cachexia, malaise, arthralgias, rash, digital clubbing, and Raynaud's phenomenon. Laboratory abnormalities, such as hypergammaglobulinemia, anemia, polycythemia, leukocytosis, elevated erythrocyte sedimentation rate, elevated C-reactive protein level, thrombocytopenia, and thrombocytosis are often present. These features account for the frequent misdiagnosis of patients with myxomas as having endocarditis, collagen vascular disease, or a paraneoplastic syndrome.

Two-dimensional and three-dimensional transthoracic and/or transesophageal echocardiography are useful in the diagnosis of cardiac myxoma and allow for assessment of tumor size and determination of the site of tumor attachment, both of which are important considerations in the planning of surgical excision **(Fig. 266-1)**. Computed tomography (CT) and magnetic resonance imaging (MRI) may provide important information regarding size, shape, composition, and surface characteristics of the tumor **(Fig. 266-2)**.

FIGURE 266-1

Transthoracic echocardiogram demonstrating a large atrial myxoma. The myxoma (Myx) fills the entire left atrium in systole (***A***) and prolapses across the mitral valve and into the left ventricle (LV) during diastole (***B***). RA, right atrium; RV, right ventricle. *(Courtesy of Dr. Michael Tsang; with permission.)*

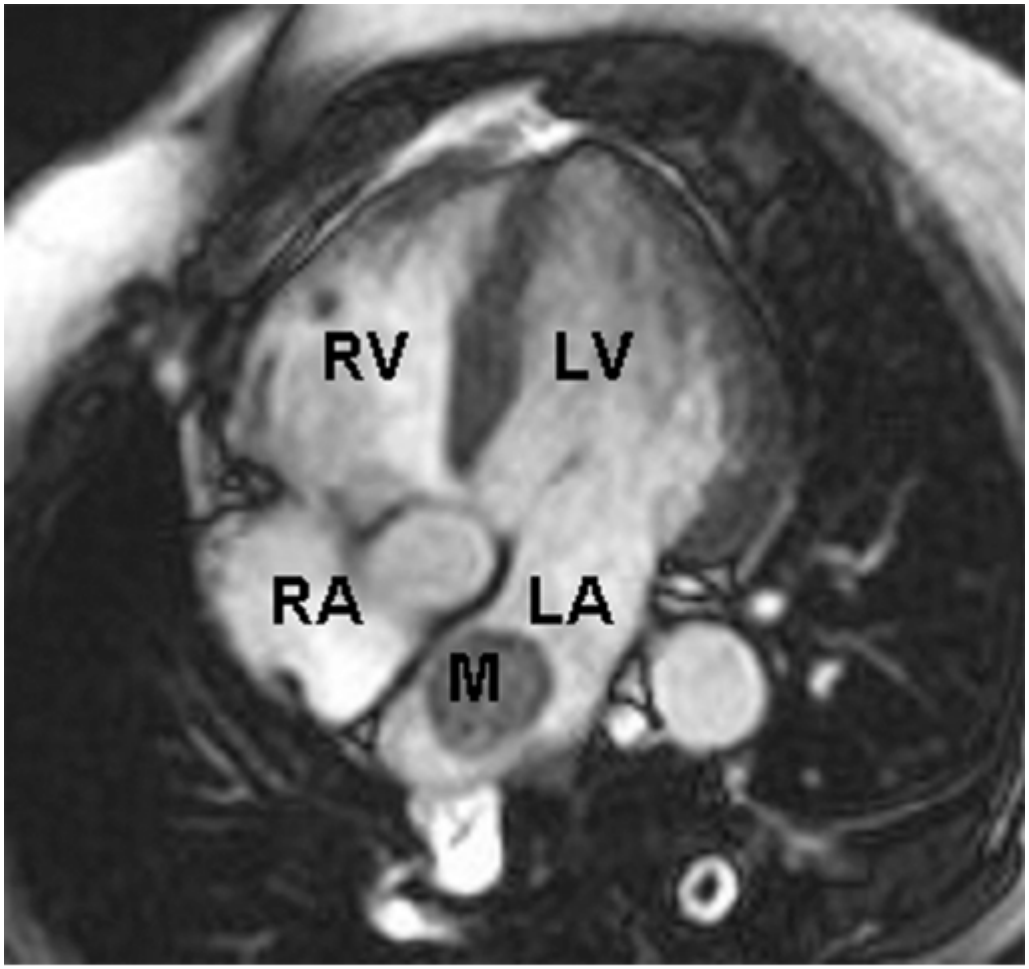


A ***B***

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 266-2

Cardiac magnetic resonance imaging demonstrating a rounded mass (M) within the left atrium (LA). Pathologic evaluation at the time of surgery revealed it to be an atrial myxoma. LV, left ventricle; RA, right atrium; RV, right ventricle.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although cardiac catheterization and angiography were previously performed routinely before tumor resection, they no longer are considered mandatory when adequate noninvasive information is available and other cardiac disorders (e.g., coronary artery disease) are not considered likely. Additionally, catheterization of the chamber from which the tumor arises carries the risk of tumor embolization. Because myxomas may be familial, echocardiographic screening of first-degree relatives is appropriate, particularly if the patient is young and has multiple tumors or features of a myxoma syndrome.

TREATMENT

TREATMENT Myxoma

Surgical excision using cardiopulmonary bypass is indicated regardless of tumor size, and is generally curative. Myxomas recur in 12–22% of familial cases but in only 1–2% of sporadic cases. Tumor recurrence most likely results from multifocal lesions in the former setting and incomplete tumor resection in the latter.

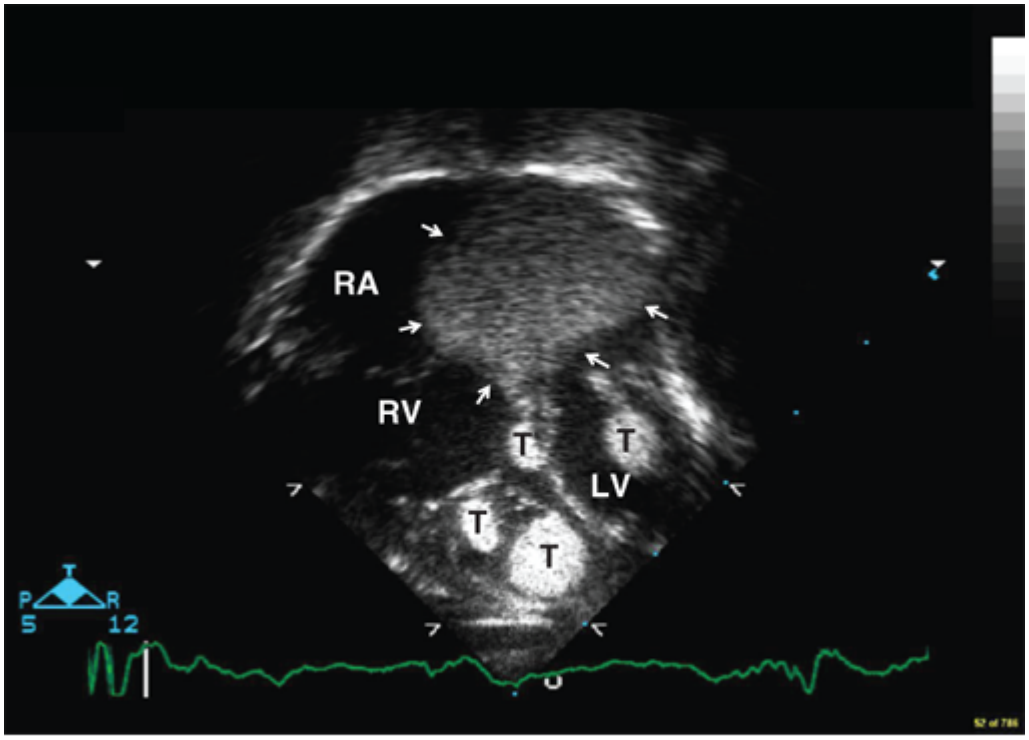
Other Benign Tumors

Cardiac *lipomas*, although relatively common, are usually incidental findings at postmortem examination; however, they may grow as large as 15 cm, may present as an abnormality of the cardiac silhouette on chest x-ray, and should be resected if they produce symptoms owing to

mechanical interference with cardiac function, arrhythmias, or conduction disturbances. *Papillary fibroelastomas* are friable tumors with frond-like projections that are usually solitary and are the most common tumors of the cardiac valves. Remnants of cytomegalovirus have been recovered from these tumors, raising the possibility that they arise as a result of chronic viral endocarditis. Although usually clinically silent, they can cause valve dysfunction and may embolize distally, resulting in transient ischemic attacks, stroke, or myocardial infarction. In general, these tumors should be resected even when asymptomatic, although a more conservative approach may be considered for small, right-sided lesions. *Rhabdomyomas* and *fibromas* are the most common cardiac tumors in infants and children and usually occur in the ventricles, where they may produce mechanical obstruction to blood flow, thereby mimicking valvular stenosis, CHF, restrictive or hypertrophic cardiomyopathy, or pericardial constriction. Rhabdomyomas are probably hamartomatous growths, are multiple in 90% of cases, and are strongly associated with tuberous sclerosis **(Fig. 266-3)**. These tumors have a tendency to regress completely or partially; only tumors that cause obstruction require surgical resection. Fibromas are usually single, universally ventricular in location, often calcified, tend to grow and cause arrhythmias and obstructive symptoms, and should be completely resected when possible. *Paragangliomas* are rare chromaffin cell tumors that represent extra-adrenal pheochromocytomas. Most are located in the roof of the left atrium and can be identified on cardiac CT or MRI, or with nuclear scanning using 131-I-metaiodobenzylguanidine. They are highly vascular and may be hormonally active, resulting in uncontrolled hypertension. Extensive surgical resection is usually required. *Hemangiomas* and *mesotheliomas* are generally small tumors, most often intramyocardial in location, and may cause atrioventricular (AV) conduction disturbances and even sudden death as a result of their propensity to develop in the region of the AV node. Other benign tumors arising from the heart include *teratoma*, *chemodectoma*, *neurilemoma*, *granular cell myoblastoma*, and *paraganglioma*.

FIGURE 266-3

Transthoracic echocardiogram revealing multiple tumors (T) consistent with rhabdomyomas in a 1-day-old infant. The largest tumor (*arrows*) was located in the left AV groove and measured 2 cm × 2 cm. LV; left ventricle; RA, right atrium; RV, right ventricle.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Sarcoma

Almost all malignant primary cardiac tumors are sarcomas; isolated cardiac lymphomas have been rarely described, but usually occur in the context of more systemic disease. Sarcomas may be of several histologic types; angiosarcomas are the most common type in adults while rhabdomyosarcomas are the most common type in children. In general, sarcomas are characterized by rapid progression that culminates in the patient's death within weeks to months from the time of presentation as a result of hemodynamic compromise, local invasion, or distant metastases. Almost one-third are metastatic at the time of initial diagnosis, usually involving the lungs. Sarcomas commonly involve the right side of the heart, are rapidly growing, frequently invade the pericardial space, and may obstruct the cardiac chambers or venae cavae. Sarcomas also may occur on the left side of the heart and may be mistaken for myxomas.

TREATMENT

TREATMENT Sarcoma

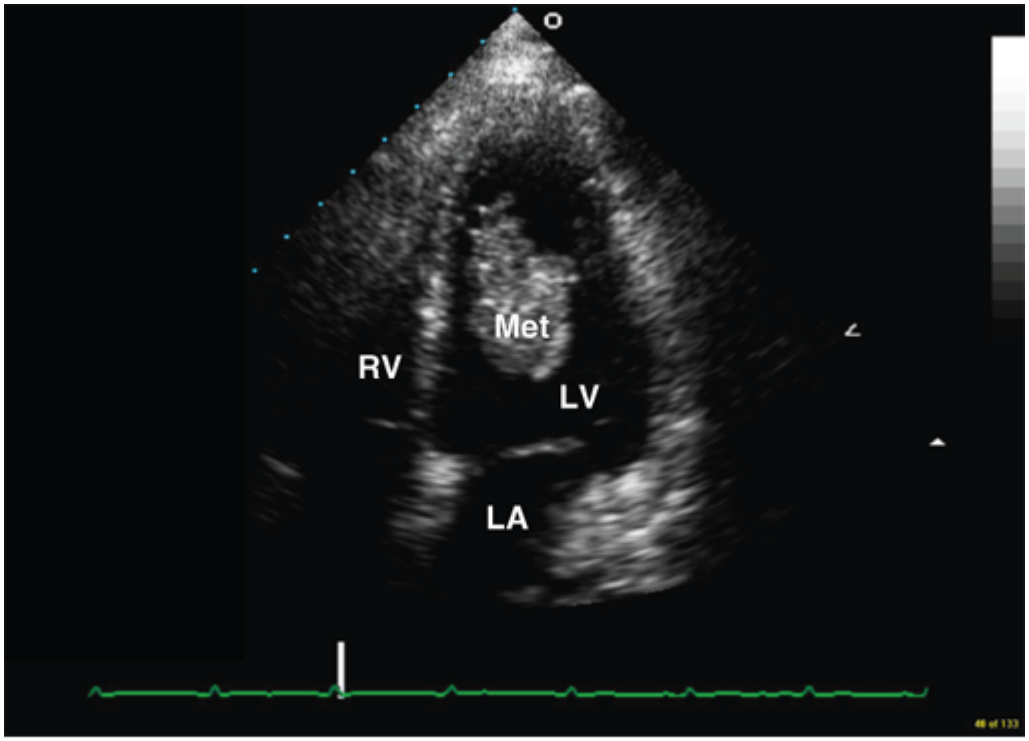
The optimal therapy for cardiac sarcoma is complete resection often with neoadjuvant and postoperative chemotherapy; however, at the time of presentation, many of these tumors have spread too extensively to allow for surgical excision. Although there are scattered reports of palliation with radiotherapy and/or chemotherapy, the response of cardiac sarcomas to these therapies is generally poor. The one exception appears to be cardiac lymphosarcomas, which may respond to a combination of chemo- and radiotherapy.

TUMORS METASTATIC TO THE HEART

Tumors metastatic to the heart are much more common than primary tumors, and their incidence is likely to increase as the life expectancy of patients with various forms of malignant neoplasms is extended by more effective therapy and improved imaging modalities allow earlier identification of metastatic disease. Although cardiac metastases may occur with any tumor type, the relative incidence is especially high in malignant melanoma and, to a somewhat lesser extent, leukemia and lymphoma **(Fig. 266-4)**. In absolute terms, the most common primary sites from which cardiac metastases originate are carcinoma of the breast and lung, reflecting the high incidence of those cancers. Cardiac metastases almost always occur in the setting of widespread primary disease, and most often there is either primary or metastatic disease elsewhere in the thoracic cavity. Nevertheless, cardiac metastasis occasionally may be the initial presentation of an extrathoracic tumor.

FIGURE 266-4

Large metastatic lesion (Met) in the left ventricle (LV) of a patient with diffusely metastatic bladder cancer. The mass arose from the interventricular septum and prolapsed into the aortic outflow tract during systole.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Cardiac metastases may occur via hematogenous or lymphangitic spread or by direct tumor invasion. They generally manifest as small, firm nodules; diffuse infiltration also may occur, especially with sarcomas or hematologic neoplasms. The pericardium is most often involved, followed by myocardial involvement of any chamber and, rarely, by involvement of the endocardium or cardiac valves.

Cardiac metastases are clinically apparent only ~10% of the time, are usually not the cause of the patient's presentation, and rarely are the cause of death. The vast majority occur in the setting of a previously recognized malignant neoplasm. As with primary cardiac tumors, the clinical presentation reflects more the location and size of the tumor than its histologic type. When symptomatic, cardiac metastases may result in a variety of clinical features, including dyspnea, acute pericarditis, cardiac tamponade, ectopic tachyarrhythmias, heart block, and CHF. Importantly, many of these signs and symptoms may also result from myocarditis, pericarditis, or cardiomyopathy induced by radiotherapy or chemotherapy, and a high index of suspicion for cardiac involvement should be maintained for patients with malignant disease who develop these symptoms.

Electrocardiographic (ECG) findings are nonspecific but may reveal features consistent with pericarditis or may demonstrate low QRS voltage and electrical alternans in the setting of a large pericardial effusion. On chest x-ray, the cardiac silhouette is most often normal but may be enlarged or exhibit a bizarre contour. Echocardiography is useful for identifying and assessing the significance of pericardial effusions and visualizing larger metastases, although CT and radionuclide imaging may define the tumor burden more clearly. Cardiac MRI offers superb image quality and plays a central role in the diagnostic evaluation of cardiac metastases and cardiac tumors in general. Pericardiocentesis may allow for a specific cytologic diagnosis in patients with malignant pericardial effusions. Angiography is rarely necessary but may help to delineate discrete myocardial lesions.

TREATMENT

TREATMENT Tumors Metastatic to the Heart

Most patients with cardiac metastases have advanced malignant disease; thus, therapy is generally palliative and consists of controlling symptoms and treatment of the primary tumor. Symptomatic malignant pericardial effusions should be drained by pericardiocentesis. Prolonged drainage (3–5 days) and concomitant instillation of a sclerosing agent (e.g., tetracycline or bleomycin) may delay or prevent reaccumulation of the effusion, and creation of a pericardial window allows drainage of the effusion to the adjacent pleural or peritoneal space. Given the overall poor prognosis of these patients, discussions regarding goals of care and involvement of palliative care services are often appropriate.

FURTHER READING

Buckley O et al: Cardiac masses, part 1: Imaging strategies and technical considerations. AJR Am J Roentgenol 197:W837, 2011. [PubMed: 22021530]
Buckley O et al: Cardiac masses, part 2: Key imaging features for diagnosis and surgical planning. AJR 1 Am J Roentgenol 97:W842, 2011.
Bussani R et al: Cardiac metastases. J Clin Path 60:27, 2007. [PubMed: 17098886]
Shapira O et al: Tumors of the heart, in *Sabiston and Spenser Surgery of the Chest*, 9th ed, FW Sellke et al (eds). Philadelphia, Elsevier, 2016, pp 1849–1857.
Tamin SS et al: Prognostic and bioepidemiologic implications of papillary fibroelastomas. J Am Coll Cardiol 65:2420, 2015. [PubMed: 26046736]

Chapter 267: Ischemic Heart Disease

Elliott M. Antman; Joseph Loscalzo

INTRODUCTION

Ischemic heart disease (IHD) is a condition in which there is an inadequate supply of blood and oxygen to a portion of the myocardium; it typically occurs when there is an imbalance between myocardial oxygen supply and demand. The most common cause of myocardial ischemia is atherosclerotic disease of an epicardial coronary artery (or arteries) sufficient to cause a regional reduction in myocardial blood flow and inadequate perfusion of the myocardium supplied by the involved coronary artery. **Chapter 291e** (from the 19th edition of *Harrison's*) deals with the development and treatment of atherosclerosis. This chapter focuses on the chronic manifestations and treatment of IHD. The subsequent chapters address the acute phases of IHD.

EPIDEMIOLOGY AND GLOBAL TRENDS

IHD causes more deaths and disability and incurs greater economic costs than any other illness in the developed world. IHD is the most common, serious, chronic, life-threatening illness in the United States, where 15.5 million persons have IHD, and 3.4 million people aged ≥40 years have angina pectoris. Although there is regional variation, about 4% of the population has sustained a myocardial infarction. Genetic factors, a high-fat and energy-rich diet, smoking, and a sedentary lifestyle are associated with the emergence of IHD. In the United States and Western Europe, IHD is growing among low-income groups, but primary prevention has delayed the disease to later in life across socioeconomic groups. Despite these sobering statistics, it is worth noting that epidemiologic data show a decline in the rate of deaths due to IHD, about half of which is attributable to treatments and half to prevention by risk factor modification.

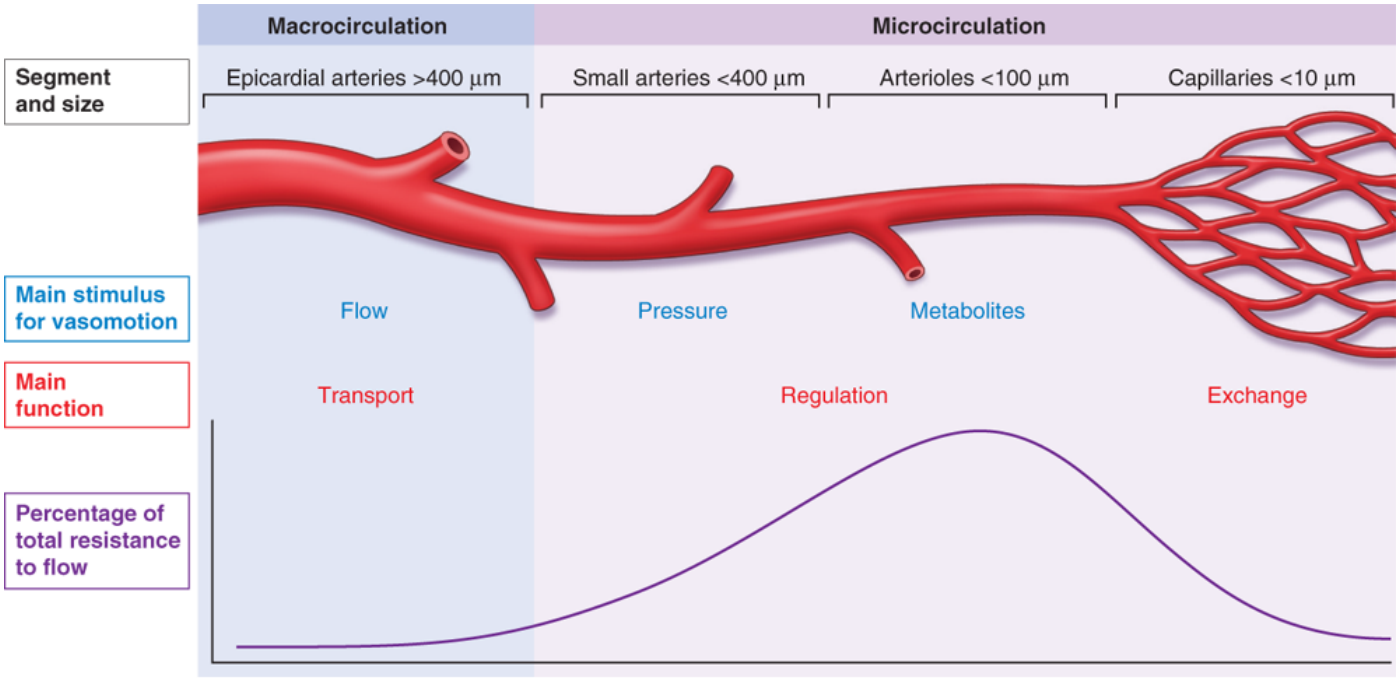
Obesity, insulin resistance, and type 2 diabetes mellitus are increasing and are powerful risk factors for IHD. These trends are occurring in the general context of population growth and as a result of the increase in the average age of the world's population. With urbanization in countries with emerging economies and a growing middle class, elements of the energy-rich Western diet are being adopted. As a result, the prevalence of risk factors for IHD and the prevalence of IHD itself are both increasing rapidly, so that in analyses of the global burden of disease, there is a shift from communicable to noncommunicable diseases. Population subgroups that appear to be particularly affected are men in South Asian countries, especially India and the Middle East. In light of the projection of large increases in IHD throughout the world, IHD is likely to become the most common cause of death worldwide by 2020.

PATHOPHYSIOLOGY

Central to an understanding of the pathophysiology of myocardial ischemia is the concept of myocardial supply and demand. In normal conditions, for any given level of a demand for oxygen, the myocardium will control the supply of oxygen-rich blood to prevent underperfusion of myocytes and the subsequent development of ischemia and infarction. The major determinants of myocardial oxygen demand (MVO_2) are heart rate, myocardial contractility, and myocardial wall tension (stress). An adequate supply of oxygen to the myocardium requires a satisfactory level of oxygen-carrying capacity of the blood (determined by the inspired level of oxygen, pulmonary function, and hemoglobin concentration and function) and an adequate level of coronary blood flow. Blood flows through the coronary arteries in a phasic fashion, with the majority occurring during diastole. About 75% of the total coronary resistance to flow occurs across three sets of arteries: (1) large epicardial arteries (Resistance 1 = R_1), (2) prearteriolar vessels (R_2), and (3) arteriolar and intramyocardial capillary vessels (R_3). In the absence of significant flow-limiting atherosclerotic obstructions, R_1 is trivial; the major determinant of coronary resistance is found in R_2 and R_3 **(Fig. 267-1)**. The normal coronary circulation is dominated and controlled by the heart's requirements for oxygen. This need is met by the ability of the coronary vascular bed to vary its resistance (and, therefore, blood flow) considerably while the myocardium extracts a high and relatively fixed percentage of oxygen. Normally, intramyocardial resistance vessels demonstrate a great capacity for dilation (R_2 and R_3 decrease). For example, the changing oxygen needs of the heart with exercise and emotional stress affect coronary vascular resistance and in this manner regulate the supply of oxygen and substrate to the myocardium (*metabolic regulation*). The coronary resistance vessels also adapt to physiologic alterations in blood pressure to maintain coronary blood flow at levels appropriate to myocardial needs (*autoregulation*).

FIGURE 267-1

Macrocirculation and microcirculation across segments and sizes of the arteries. The location and size of the arteries supplying blood to the heart is shown at the top. Vasomotion of the arterial segments occurs in response to the stimuli shown. The main function of each of the arterial segments is shown next, followed by a depiction of the relative resistance to antegrade flow. *(Modified from B De Bruyne et al: J Am Coll Cardiol 67:1170, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

By reducing the lumen of the coronary arteries, atherosclerosis limits appropriate increases in perfusion when the demand for flow is augmented, as occurs during exertion or excitement. When the luminal reduction is severe, myocardial perfusion in the basal state is reduced. Coronary blood flow also can be limited by spasm (see Prinzmetal's angina in **Chap. 268**), arterial thrombi, and, rarely, coronary emboli as well as by ostial narrowing due to aortitis. Congenital abnormalities such as the origin of the left anterior descending coronary artery from the pulmonary artery may cause myocardial ischemia and infarction in infancy, but this cause is very rare in adults.

Myocardial ischemia also can occur if myocardial oxygen demands are markedly increased and particularly when coronary blood flow may be limited, as occurs in severe left ventricular hypertrophy (LVH) due to aortic stenosis. The latter can present with angina that is indistinguishable from that caused by coronary atherosclerosis largely owing to subendocardial ischemia **(Chap. 256)**. A reduction in the oxygen-carrying capacity of the blood, as in extremely severe anemia or in the presence of carboxyhemoglobin, rarely causes myocardial ischemia by itself but may lower the threshold for ischemia in patients with moderate coronary obstruction.

Not infrequently, two or more causes of ischemia coexist in a patient, such as an increase in oxygen demand due to LVH secondary to hypertension and a reduction in oxygen supply secondary to coronary atherosclerosis and anemia. Abnormal constriction or failure of normal dilation of the coronary resistance vessels also can cause ischemia. When it causes angina, this condition is referred to as *microvascular angina*.

CORONARY ATHEROSCLEROSIS

Epicardial coronary arteries are the major site of atherosclerotic disease. The major risk factors for atherosclerosis (high levels of plasma low-density lipoprotein [LDL], low plasma high-density lipoprotein [HDL], cigarette smoking, hypertension, and diabetes mellitus vary in their relative impact on disturbing the normal functions of the vascular endothelium. These functions include local control of vascular tone, maintenance of an antithrombotic surface, and control of inflammatory cell adhesion and diapedesis. The loss of these defenses leads to inappropriate constriction, luminal thrombus formation, and abnormal interactions between blood cells, especially monocytes and platelets, and the activated vascular endothelium. Functional changes in the vascular milieu ultimately result in the subintimal collections of fat, smooth muscle cells, fibroblasts, and intercellular matrix that define the atherosclerotic plaque. Rather than viewing atherosclerosis strictly as a vascular problem, it is useful to consider it in the context of alterations in the nature of the circulating blood (hyperglycemia; increased concentrations of LDL cholesterol, tissue factor, fibrinogen, von Willebrand factor, coagulation factor VII, and platelet microparticles). The combination of a "vulnerable vessel" in a patient with "vulnerable blood" promotes a state of hypercoagulability and hypofibrinolysis. This is especially true in patients with diabetes mellitus.

Atherosclerosis develops at irregular rates in different segments of the epicardial coronary tree and leads eventually to segmental reductions in cross-sectional area, i.e., plaque formation. There is also a predilection for atherosclerotic plaques to develop at sites of increased turbulence in coronary flow, such as at branch points in the epicardial arteries. When a stenosis reduces the diameter of an epicardial artery by 50%, there is a limitation of the ability to increase flow to meet increased myocardial demand. When the diameter is reduced by ~80%, blood flow at rest may be reduced, and further minor decreases in the stenotic orifice area can reduce coronary flow dramatically to cause myocardial ischemia at rest or with minimal stress.

Segmental atherosclerotic narrowing of epicardial coronary arteries is caused most commonly by the formation of a plaque, which is subject to rupture or erosion of the cap separating the plaque from the bloodstream. Upon exposure of the plaque contents to blood, two important and interrelated processes are set in motion: (1) platelets are activated and aggregate, and (2) the coagulation cascade is activated, leading to deposition of fibrin strands. A thrombus composed of platelet aggregates and fibrin strands traps red blood cells and can reduce coronary blood flow, leading to the clinical manifestations of myocardial ischemia.

The location of the obstruction influences the quantity of myocardium rendered ischemic and determines the severity of the clinical manifestations. Thus, critical obstructions in vessels, such as the left main coronary artery and the proximal left anterior descending coronary artery, are particularly hazardous. Chronic severe coronary narrowing and myocardial ischemia frequently are accompanied by the development of collateral vessels,

especially when the narrowing develops gradually. When well developed, such vessels can by themselves provide sufficient blood flow to sustain the viability of the myocardium at rest but not during conditions of increased demand.

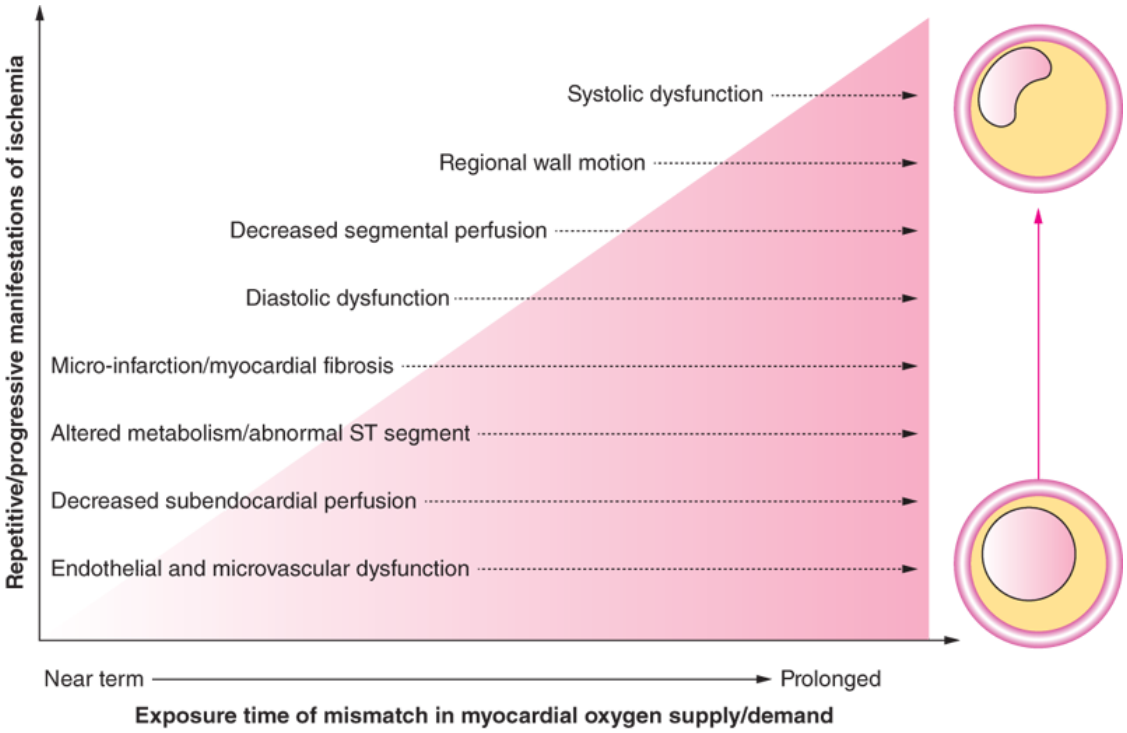
With progressive worsening of a stenosis in a proximal epicardial artery, the distal resistance vessels (when they function normally) dilate to reduce vascular resistance and maintain coronary blood flow. A pressure gradient develops across the proximal stenosis, and poststenotic pressure falls. When the resistance vessels are maximally dilated, myocardial blood flow becomes dependent on the pressure in the coronary artery distal to the obstruction. In these circumstances, ischemia, manifest clinically by angina or electrocardiographically by ST-segment deviation, can be precipitated by increases in myocardial oxygen demand caused by physical activity, emotional stress, and/or tachycardia. Changes in the caliber of the stenosed coronary artery resulting from physiologic vasomotion, loss of endothelial control of dilation (as occurs in atherosclerosis), pathologic spasm (Prinzmetal's angina), or small platelet-rich plugs also can upset the critical balance between oxygen supply and demand and thereby precipitate myocardial ischemia.

EFFECTS OF ISCHEMIA

During episodes of inadequate perfusion caused by coronary atherosclerosis, myocardial tissue oxygen tension falls and may cause transient disturbances of the mechanical, biochemical, and electrical functions of the myocardium **(Fig. 267-2)**. Coronary atherosclerosis is a focal process that usually causes nonuniform ischemia. During ischemia, regional disturbances of ventricular contractility cause segmental hypokinesia, akinesia, or, in severe cases, bulging (dyskinesia), which can reduce myocardial pump function.

FIGURE 267-2

Cascade of mechanisms and manifestations of ischemia. *(Modified from LJ Shaw et al: J Am Coll Cardiol 54:1561, 2009. Original figure illustration by Rob Flewell.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The abrupt development of severe ischemia, as occurs with total or subtotal coronary occlusion, is associated with almost instantaneous failure of normal muscle relaxation and then contraction. The relatively poor perfusion of the subendocardium causes more intense ischemia of this portion of the wall (compared with the subepicardial region). Ischemia of large portions of the ventricle causes transient left ventricular (LV) failure, and if the papillary muscle apparatus is involved, mitral regurgitation can occur. When ischemia is transient, it may be associated with angina pectoris; when it is prolonged, it can lead to myocardial necrosis and scarring with or without the clinical picture of acute myocardial infarction **(Chap. 269)**.

A wide range of abnormalities in cell metabolism, function, and structure underlie these mechanical disturbances during ischemia. The normal myocardium metabolizes fatty acids and glucose to carbon dioxide and water. With severe oxygen deprivation, fatty acids cannot be oxidized, and glucose is converted to lactate; intracellular pH is reduced, as are the myocardial stores of high-energy phosphates, i.e., ATP and creatine phosphate. Impaired cell membrane function leads to the leakage of potassium and the uptake of sodium by myocytes as well as an increase in cytosolic calcium. The severity and duration of the imbalance between myocardial oxygen supply and demand determine whether the damage is reversible (≤20 min for total occlusion in the absence of collaterals) or permanent, with subsequent myocardial necrosis (>20 min).

Ischemia also causes characteristic changes in the electrocardiogram (ECG) such as repolarization abnormalities, as evidenced by inversion of T waves and, when more severe, displacement of ST segments **(Chap. 235)**. Transient T-wave inversion probably reflects nontransmural, intramyocardial ischemia; transient ST-segment depression often reflects patchy subendocardial ischemia; and ST-segment elevation is thought to be caused by more severe transmural ischemia. Another important consequence of myocardial ischemia is electrical instability, which may lead to isolated ventricular premature beats or even ventricular tachycardia or ventricular fibrillation **(Chaps. 249 and 250)**. Most patients who die suddenly from IHD do so as a result of ischemia-induced ventricular tachyarrhythmias **(Chap. 299)**.

ASYMPTOMATIC VERSUS SYMPTOMATIC IHD

Although the prevalence is decreasing, postmortem studies of accident victims and military casualties in Western countries show that coronary atherosclerosis can begin before age 20 and is present even among adults who were asymptomatic during life. Exercise stress tests in asymptomatic persons may show evidence of silent myocardial ischemia, i.e., exercise-induced ECG changes not accompanied by angina pectoris; coronary angiographic studies of such persons may reveal coronary artery plaques and previously unrecognized obstructions **(Chap. 237)**. Coronary artery calcifications (CAC) may be seen on CT images of the heart, can be quantified in a CAC score, and may be used as adjunctive information to support a diagnosis of IHD. However, they should not be used as the primary screening modality or as the isolated basis on which to formulate therapeutic decisions. (See further discussion below.) Postmortem examination of patients with such obstructions without a history of clinical manifestations of myocardial ischemia often shows macroscopic scars secondary to myocardial infarction in regions supplied by diseased coronary arteries, with or without collateral circulation. According to population studies, ~25% of patients who survive acute myocardial infarction may not come to medical attention, and these patients have the same adverse prognosis as do those who present with the classic clinical picture of acute myocardial infarction **(Chap. 269)**. Sudden death may be unheralded and is a common presenting manifestation of IHD **(Chap. 299)**.

Patients with IHD also can present with cardiomegaly and heart failure secondary to ischemic damage of the LV myocardium that may have caused no symptoms before the development of heart failure; this condition is referred to as *ischemic cardiomyopathy*. In contrast to the asymptomatic phase of IHD, the symptomatic phase is characterized by chest discomfort due to either angina pectoris or acute myocardial infarction **(Chap. 269)**. Having entered the symptomatic phase, the patient may exhibit a stable or progressive course, revert to the asymptomatic stage, or die suddenly.

STABLE ANGINA PECTORIS

This episodic clinical syndrome is due to transient myocardial ischemia. Various diseases that cause myocardial ischemia and the numerous forms of discomfort with which it may be confused are discussed in **Chap. 11**. Males constitute ~70% of all patients with angina pectoris and an even greater proportion of those aged <50 years. It is, however, important to note that angina pectoris in women is often atypical in presentation (see below).

HISTORY

The typical patient with angina is a man >50 years or a woman >60 years of age who complains of episodes of chest discomfort, usually described as heaviness, pressure, squeezing, smothering, or choking and only rarely as frank pain. When the patient is asked to localize the sensation, he or she typically places a hand over the sternum, sometimes with a clenched fist, to indicate a squeezing, central, substernal discomfort (Levine's sign). Angina is usually crescendo-decrescendo in nature, typically lasts 2–5 min, and can radiate to either shoulder and to both arms (especially the ulnar surfaces of the forearm and hand). It also can arise in or radiate to the back, interscapular region, root of the neck, jaw, teeth, and epigastrium. Angina is rarely localized below the umbilicus or above the mandible. A useful finding in assessing a patient with chest discomfort is the fact that myocardial ischemic discomfort does not radiate to the trapezius muscles; that radiation pattern is more typical of pericarditis.

Although episodes of angina typically are caused by exertion (e.g., exercise, hurrying, or sexual activity) or emotion (e.g., stress, anger, fright, or frustration) and are relieved by rest, they also may occur at rest **(Chap. 268)** and while the patient is recumbent (angina decubitus). The patient may be awakened at night by typical chest discomfort and dyspnea. Nocturnal angina may be due to episodic tachycardia, diminished oxygenation as the respiratory pattern changes during sleep, or expansion of the intrathoracic blood volume that occurs with recumbency; the latter causes an increase in cardiac size (end-diastolic volume), wall tension, and myocardial oxygen demand that can lead to ischemia and transient LV failure.

The threshold for the development of angina pectoris may vary by time of day and emotional state. Many patients report a fixed threshold for angina, which occurs predictably at a certain level of activity, such as climbing two flights of stairs at a normal pace. In these patients, coronary stenosis and myocardial oxygen supply are fixed, and ischemia is precipitated by an increase in myocardial oxygen demand; they are said to have stable exertional angina. In other patients, the threshold for angina may vary considerably within any particular day and from day to day. In such patients, variations in myocardial oxygen supply, most likely due to changes in coronary vasomotor tone, may play an important role in defining the pattern of angina. A patient may report symptoms upon minor exertion in the morning (a short walk or shaving) yet by midday be capable of much greater effort without symptoms. Angina may also be precipitated by unfamiliar tasks, a heavy meal, exposure to cold, or a combination of these factors.

Exertional angina typically is relieved in 1–5 min by slowing or ceasing activities and even more rapidly by rest and sublingual nitroglycerin (see below). Indeed, the diagnosis of angina should be suspect if it does not respond to the combination of these measures. The severity of angina can be conveniently summarized by the Canadian Cardiac Society functional classification **(Table 267-1)**. Its impact on the patient's functional capacity can be described by using the New York Heart Association functional classification (Table 267-1).

TABLE 267-1 Cardiovascular Disease Classification Chart

Class	New York Heart Association Functional Classification	Canadian Cardiovascular Society Functional Classification
I	Patients have cardiac disease but *without* the resulting *limitations* of physical activity. Ordinary physical activity does not cause undue fatigue, palpitation, dyspnea, or anginal pain.	Ordinary physical activity, such as walking and climbing stairs, *does not cause angina*. Angina present with strenuous or rapid or prolonged exertion at work or recreation.
II	Patients have cardiac disease resulting in *slight limitation* of physical activity. They are comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnea, or anginal pain.	*Slight limitation* of ordinary activity. Walking or climbing stairs rapidly, walking uphill, walking or stair climbing after meals, in cold, or when under emotional stress or only during the few hours after awakening. Walking more than two blocks on the level and climbing more than one flight of stairs at a normal pace and in normal conditions.
III	Patients have cardiac disease resulting in *marked limitation* of physical activity. They are comfortable at rest. Less than ordinary physical activity causes fatigue, palpitation, dyspnea, or anginal pain.	*Marked limitation* of ordinary physical activity. Walking one to two blocks on the level and climbing more than one flight of stairs in normal conditions.
IV	Patients have cardiac disease resulting in *inability* to carry on any physical activity without discomfort. Symptoms of cardiac insufficiency or of the anginal syndrome may be present even at rest. If any physical activity is undertaken, discomfort is increased.	*Inability* to carry on any physical activity without discomfort—anginal syndrome *may* be present at rest.

Source: Modified from L Goldman et al: Circulation 64:1227, 1981.

Sharp, fleeting chest pain or a prolonged, dull ache localized to the left submammary area is rarely due to myocardial ischemia. However, especially in women and diabetic patients, angina pectoris may be atypical in location and not strictly related to provoking factors. In addition, this symptom may exacerbate and remit over days, weeks, or months. Its occurrence can be seasonal, occurring more frequently in the winter in temperate climates. Anginal “equivalents” are symptoms of myocardial ischemia other than angina. They include dyspnea, nausea, fatigue, and faintness and are more common in the elderly and in diabetic patients.

Systematic questioning of a patient with suspected IHD is important to uncover the features of an unstable syndrome associated with increased risk, such as angina occurring with less exertion than in the past, occurring at rest, or awakening the patient from sleep. Since coronary atherosclerosis often is accompanied by similar lesions in other arteries, a patient with angina should be questioned and examined for peripheral arterial disease (intermittent claudication **[Chap. 275]**), stroke, or transient ischemic attacks **(Chap. 419)**. It is also important to uncover a family history of premature IHD (<55 years in first-degree male relatives and <65 in female relatives) and the presence of diabetes mellitus, hyperlipidemia, hypertension, cigarette smoking, and other risk factors for coronary atherosclerosis.

The history of typical angina pectoris establishes the diagnosis of IHD until proven otherwise. The coexistence of advanced age, male sex, the postmenopausal state, and risk factors for atherosclerosis increase the likelihood of hemodynamically significant coronary disease. A particularly challenging problem is the evaluation and management of patients with persistent ischemic-type chest discomfort but no flow-limiting obstructions in their epicardial coronary arteries. This situation arises more often in women than in men. Potential etiologies include microvascular coronary disease (detectable on coronary reactivity testing in response to vasoactive agents such as intracoronary adenosine, acetylcholine, and nitroglycerin) and abnormal cardiac nociception. Treatment of microvascular coronary disease should focus on efforts to improve endothelial function, including nitrates, beta blockers, calcium antagonists, statins, and angiotensin-converting enzyme (ACE) inhibitors. Abnormal cardiac nociception is more difficult to manage and may be ameliorated in some cases by imipramine.

PHYSICAL EXAMINATION

The physical examination is often normal in patients with stable angina when they are asymptomatic. However, because of the increased likelihood of IHD in patients with diabetes and/or peripheral arterial disease, clinicians should search for evidence of atherosclerotic disease at other sites, such as an abdominal aortic aneurysm, carotid arterial bruits, and diminished arterial pulses in the lower extremities. The physical examination also should include a search for evidence of risk factors for atherosclerosis such as xanthelasmas and xanthomas. Evidence for peripheral arterial disease should be sought by evaluating the pulse contour at multiple locations and comparing the blood pressure between the arms and between the arms and the legs (ankle-brachial index). Examination of the fundi may reveal an increased light reflex and arteriovenous nicking as evidence of hypertension. There also may be signs of anemia, thyroid disease, and nicotine stains on the fingertips from cigarette smoking.

Palpation may reveal cardiac enlargement and abnormal contraction of the cardiac impulse (LV dyskinesia). Auscultation can uncover arterial bruits, a third and/or fourth heart sound, and, if acute ischemia or previous infarction has impaired papillary muscle function, an apical systolic murmur due to mitral regurgitation. These auscultatory signs are best appreciated with the patient in the left lateral decubitus position. Aortic stenosis, aortic regurgitation **(Chap. 256)**, pulmonary hypertension **(Chap. 277)**, and hypertrophic cardiomyopathy **(Chap. 254)** must be excluded, since these disorders may cause angina in the absence of coronary atherosclerosis. Examination during an anginal attack is useful, since ischemia can cause transient LV failure with the appearance of a third and/or fourth heart sound, a dyskinetic cardiac apex, mitral regurgitation, and even pulmonary edema. Tenderness of the chest wall, localization of the discomfort with a single fingertip on the chest, or reproduction of the pain with palpation of the chest makes it unlikely that the pain is caused by myocardial ischemia. A protuberant abdomen may indicate that the patient has the metabolic syndrome and is at increased risk for atherosclerosis.

LABORATORY EXAMINATION

Although the diagnosis of IHD can be made with a high degree of confidence from the history and physical examination, a number of simple laboratory tests can be helpful. The urine should be examined for evidence of diabetes mellitus and renal disease (including microalbuminuria) since these conditions accelerate atherosclerosis. Similarly, examination of the blood should include measurements of lipids (cholesterol—total, LDL, HDL—and triglycerides), glucose (hemoglobin A_{1C}), creatinine, hematocrit, and, if indicated based on the physical examination, thyroid function. A chest x-ray is important as it may show the consequences of IHD, i.e., cardiac enlargement, ventricular aneurysm, or signs of heart failure. These signs can support the diagnosis of IHD and are important in assessing the degree of cardiac damage. Evidence exists that an elevated level of high-sensitivity C-reactive protein (CRP) (specifically, between 0 and 3 mg/dL) is an independent risk factor for IHD and may be useful in therapeutic decision-making about the initiation of hypolipidemic treatment. The major benefit of high-sensitivity CRP is in reclassifying the risk of IHD in patients in the “intermediate” risk category on the basis of traditional risk factors.

ELECTROCARDIOGRAM

A 12-lead ECG recorded at rest may be normal in patients with typical angina pectoris, but there may also be signs of an old myocardial infarction **(Chap. 235)**. Although repolarization abnormalities, i.e., ST-segment and T-wave changes, as well as LVH and disturbances of cardiac rhythm or intraventricular conduction are suggestive of IHD, they are nonspecific, since they also can occur in pericardial, myocardial, and valvular heart disease or, in the case of the former, transiently with anxiety, changes in posture, drugs, or esophageal disease. The presence of LVH is a significant indication of increased risk of adverse outcomes from IHD. Of note, even though LVH and cardiac rhythm disturbances are nonspecific indicators of the development of IHD, they may be contributing factors to episodes of angina in patients in whom IHD has developed as a consequence of conventional risk factors. Dynamic ST-segment and T-wave changes that accompany episodes of angina pectoris and disappear thereafter are more specific.

STRESS TESTING

Electrocardiographic

The most widely used test for both the diagnosis of IHD and the estimation of risk and prognosis involves recording of the 12-lead ECG before, during, and after exercise, usually on a treadmill **(Fig. 267-3)**. The test consists of a standardized incremental increase in external workload **(Table 267-2)** while symptoms, the ECG, and arm blood pressure are monitored. Exercise duration is usually symptom-limited, and the test is discontinued upon evidence of chest discomfort, severe shortness of breath, dizziness, severe fatigue, ST-segment depression >0.2 mV (2 mm), a fall in systolic blood pressure >10 mmHg, or the development of a ventricular tachyarrhythmia. This test is used to discover any limitation in exercise performance, detect typical ECG signs of myocardial ischemia, and establish their relationship to chest discomfort. The ischemic ST-segment response generally is defined as flat or downsloping depression of the ST segment >0.1 mV below baseline (i.e., the PR segment) and lasting longer than 0.08 s (Fig. 267-2). Upsloping or junctional ST-segment changes are not considered characteristic of ischemia and do not constitute a positive test. Although T-wave abnormalities, conduction disturbances, and ventricular arrhythmias that develop during exercise should be noted, they are also not diagnostic. Negative exercise tests in which the target heart rate (85% of maximal predicted heart rate for age and sex) is not achieved are considered nondiagnostic.

TABLE 267-2 Relation of Metabolic Equivalent Tasks (METs) to Stages in Various Testing Protocols

FUNCTIONAL CLASS	CLINICAL STATUS				O_2 COST mL/kg/min	METs	TREADMILL PROTOCOLS			
							BRUCE Modified 3 min Stages		BRUCE 3 min Stages	
							MPH	%GR	MPH	%GR
							6.0	22	6.0	22
	HEALTHY, DEPENDENT ON AGE, ACTIVITY						5.5	20	5.2	20
							5.0	18	5.0	18
					56.0	16				
					52.5	15				
NORMAL AND I					49.0	14				
					45.5	13	4.2	16	4.2	16
					42.0	12				
					38.5	11	3.4	14	3.4	14
		SEDENTARY HEALTHY			35.0	10				
					31.5	9				
					28.0	8				
			LIMITED		24.5	7	2.5	12	2.5	12
II				SYMPTOMATIC	21.0	6				
					17.5	5	1.7	10	1.7	10
III					14.0	4				
					10.5	3	1.7	5		
					7.0	2	1.7	0		
IV					3.5	1				

Note: The standard Bruce treadmill protocol (right hand column) begins at 1.7 MPH and 10% gradient (GR) and progresses every 3 min to a higher speed and elevation. The corresponding oxygen consumption and clinical status of the patient are shown in the center and left hand columns.

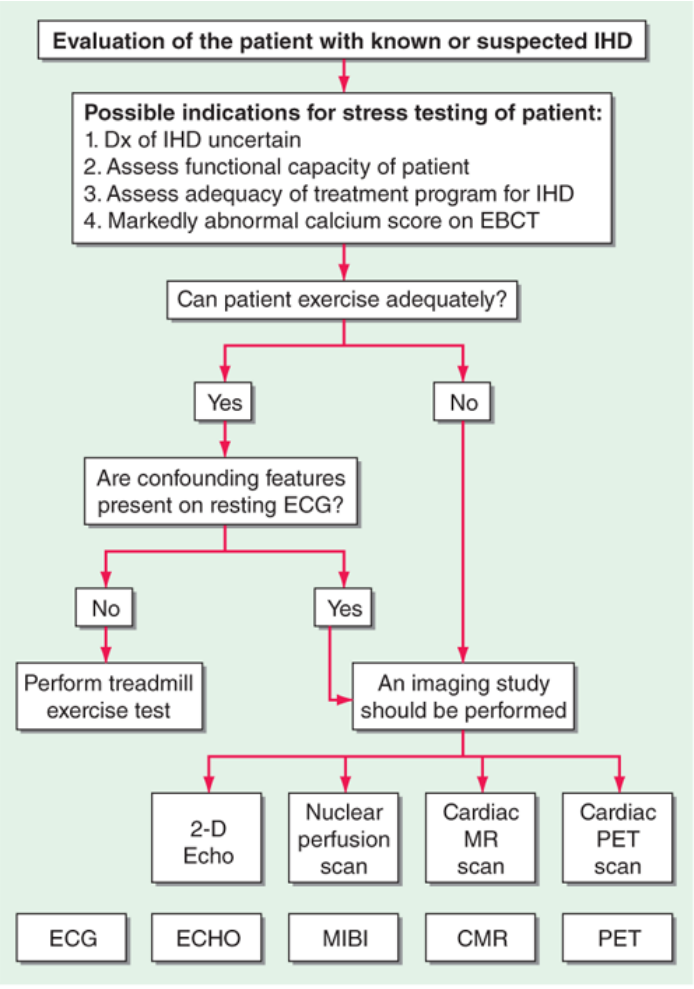
Abbreviations: GR, grade; MPH, miles per hour.

Source: Modified from GF Fletcher et al: Circulation 104:1694, 2001.

FIGURE 267-3

Evaluation of the patient with known or suspected ischemic heart disease. On the left of the figure is an algorithm for identifying patients who should be referred for stress testing and the decision pathway for determining whether a standard treadmill exercise with electrocardiogram (ECG) monitoring alone is adequate. A specialized imaging study is necessary if the patient cannot exercise adequately (pharmacologic challenge is given) or if there are confounding features on the resting ECG (symptom-limited treadmill exercise may be used to stress the coronary circulation). Panels B–E on the next page are examples of the data obtained with ECG monitoring and specialized imaging procedures. CMR, cardiac magnetic resonance; EBCT, electron beam computed tomography; ECHO, echocardiography; IHD, ischemic heart disease; MIBI, methoxyisobutyl isonitrite; MR, magnetic resonance; PET, positron emission tomography. ***A.*** Lead V_4 at rest (*top panel*) and after 4.5 min of exercise (*bottom panel*). There is 3 mm (0.3 mV) of horizontal ST-segment depression, indicating a positive test for ischemia. *(Modified from BR Chaitman, in E Braunwald et al [eds]: Heart Disease, 8th ed, Philadelphia, Saunders, 2008.)* ***B.*** A 45-year-old avid jogger who began experiencing classic substernal chest pressure underwent an exercise echo study. With exercise the patient's heart rate increased from 52 to 153 beats/min. The left ventricular chamber

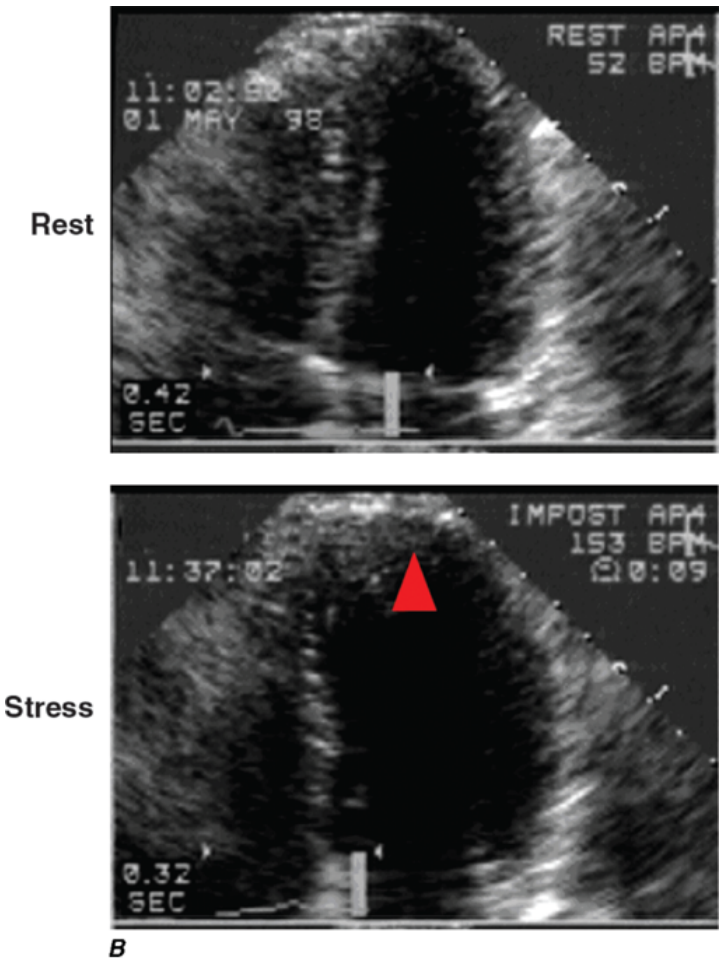
dilated with exercise, and the septal and apical portions became akinetic to dyskinetic (*red arrow*). These findings are strongly suggestive of a significant flow-limiting stenosis in the proximal left anterior descending artery, which was confirmed at coronary angiography. *(Modified from SD Solomon, in E. Braunwald et al [eds]: Primary Cardiology, 2nd ed, Philadelphia, Saunders, 2003.)* ***C.*** Stress and rest myocardial perfusion single-photon emission computed tomography images obtained with 99m-technetium sestamibi in a patient with chest pain and dyspnea on exertion. The images demonstrate a medium-size and severe stress perfusion defect involving the inferolateral and basal inferior walls, showing nearly complete reversibility, consistent with moderate ischemia in the right coronary artery territory (*red arrows*). *(Images provided by Dr. Marcello Di Carli, Nuclear Medicine Division, Brigham and Women's Hospital, Boston, MA.)* ***D.*** A patient with a prior myocardial infarction presented with recurrent chest discomfort. On cardiac magnetic resonance (CMR) cine imaging, a large area of anterior akinesia was noted (marked by the *arrows* in the top left and right images, systolic frame only). This area of akinesia was matched by a larger extent of late gadolinium-DTPA enhancements consistent with a large transmural myocardial infarction (marked by *arrows* in the middle left and right images). Resting (*bottom left*) and adenosine vasodilating stress (*bottom right*) first-pass perfusion images revealed reversible perfusion abnormality that extended to the inferior septum. This patient was found to have an occluded proximal left anterior descending coronary artery with extensive collateral formation. This case illustrates the utility of different modalities in a CMR examination in characterizing ischemic and infarcted myocardium. DTPA, diethylenetriamine penta-acetic acid. *(Images provided by Dr. Raymond Kwong, Cardiovascular Division, Brigham and Women's Hospital, Boston, MA.)* ***E.*** Stress and rest myocardial perfusion PET images obtained with rubidium-82 in a patient with chest pain on exertion. The images demonstrate a large and severe stress perfusion defect involving the mid and apical anterior, anterolateral, and anteroseptal walls and the left ventricular apex, showing complete reversibility, consistent with extensive and severe ischemia in the mid-left anterior descending coronary artery territory (*red arrows*). *(Images provided by Dr. Marcello Di Carli, Nuclear Medicine Division, Brigham and Women's Hospital, Boston, MA.)*





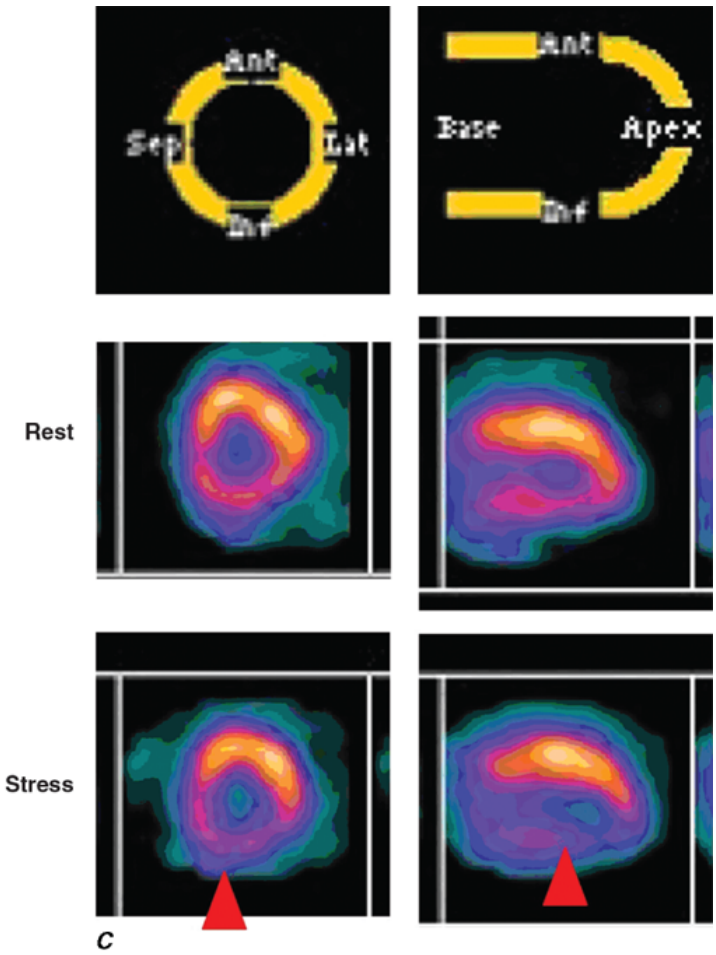
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



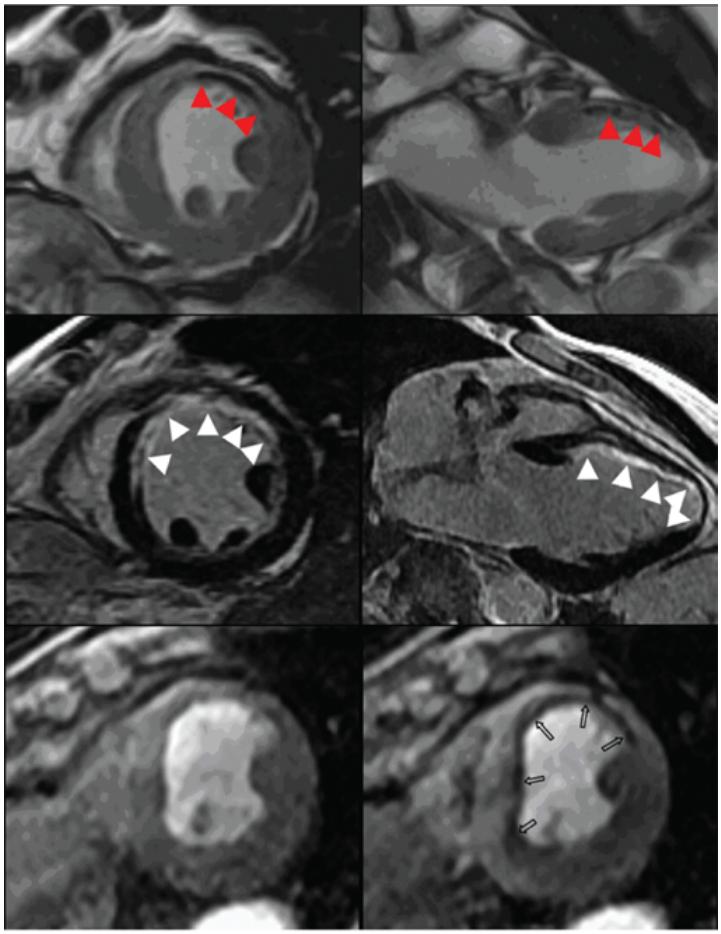
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



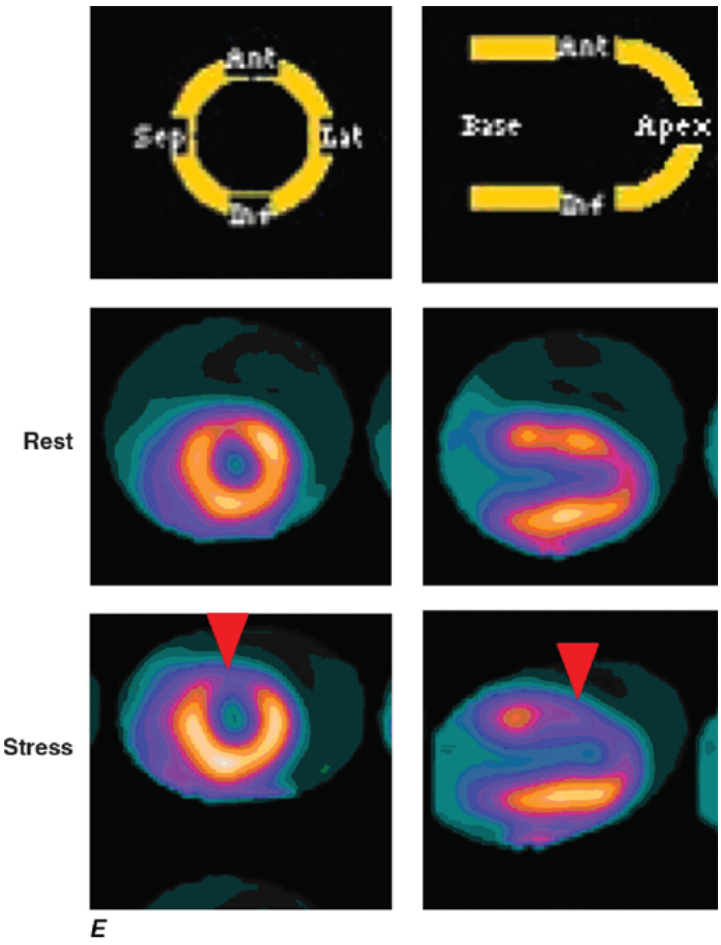
C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



E

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In interpreting ECG stress tests, the probability that coronary artery disease (CAD) exists in the patient or population under study (i.e., pretest probability) should be considered. Overall, false-positive or false-negative results occur in one-third of cases. However, a positive result on exercise indicates that the likelihood of CAD is 98% in males who are >50 years with a history of typical angina pectoris and who develop chest discomfort during the test. The likelihood decreases if the patient has atypical or no chest pain by history and/or during the test.

The incidence of false-positive tests is significantly increased in patients with low probabilities of IHD, such as asymptomatic men age <40 or premenopausal women with no risk factors for premature atherosclerosis. It is also increased in patients taking cardioactive drugs, such as digitalis and antiarrhythmic agents, and in those with intraventricular conduction disturbances, resting ST-segment and T-wave abnormalities, ventricular hypertrophy, or abnormal serum potassium levels. Obstructive disease limited to the circumflex coronary artery may result in a false-negative stress test since the lateral portion of the heart that this vessel supplies is not well represented on the surface 12-lead ECG. Since the overall sensitivity of exercise stress electrocardiography is only ~75%, a negative result does not exclude CAD, although it makes the likelihood of three-vessel or left main CAD extremely unlikely.

A medical professional should be present throughout the exercise test. It is important to measure total duration of exercise, the times to the onset of ischemic ST-segment change and chest discomfort, the external work performed (generally expressed as the stage of exercise), and the internal cardiac work performed, i.e., by the heart rate–blood pressure product. The depth of the ST-segment depression and the time needed for recovery of these ECG changes are also important. Because the risks of exercise testing are small but real—estimated at one fatality and two nonfatal complications per 10,000 tests—equipment for resuscitation should be available. Modified (heart rate–limited rather than symptom-limited) exercise tests can be performed safely in patients as early as 6 days after uncomplicated myocardial infarction (Table 267-2). Contraindications to exercise stress testing include rest angina within 48 h, unstable rhythm, severe aortic stenosis, acute myocarditis, uncontrolled heart failure, severe pulmonary hypertension, and active infective endocarditis.

The normal response to graded exercise includes progressive increases in heart rate and blood pressure. Failure of the blood pressure to increase or an actual decrease with signs of ischemia during the test is an important adverse prognostic sign, since it may reflect ischemia-induced global LV dysfunction. The development of angina and/or severe (>0.2 mV) ST-segment depression at a low workload, i.e., before completion of stage II of the Bruce protocol, and/or ST-segment depression that persists >5 min after the termination of exercise increases the specificity of the test and suggests severe IHD and a high risk of future adverse events.

Cardiac Imaging

(See also Chap. 236) When the resting ECG is abnormal (e.g., preexcitation syndrome, >1 mm of resting ST-segment depression, left bundle branch block, paced ventricular rhythm), information gained from an exercise test can be enhanced by stress myocardial radionuclide perfusion imaging after the intravenous administration of thallium-201 or 99m-technetium sestamibi during exercise (or with pharmacologic) stress. Contemporary data also suggest positron emission tomography (PET) imaging (with exercise or pharmacologic stress) using N-13 ammonia or rubidium-82 nuclide as another technique for assessing perfusion. Images obtained immediately after cessation of exercise to detect regional ischemia are compared with those obtained at rest to confirm reversible ischemia and regions of persistently absent uptake that signify infarction.

A sizable fraction of patients who need noninvasive stress testing to identify myocardial ischemia and increased risk of coronary events cannot exercise because of peripheral vascular or musculoskeletal disease, exertional dyspnea, or deconditioning. In these circumstances, an intravenous pharmacologic challenge is used in place of exercise. For example, dipyridamole or adenosine can be given to create a coronary "steal" by temporarily increasing flow in nondiseased segments of the coronary vasculature at the expense of diseased segments. Alternatively, a graded incremental infusion of dobutamine may be administered to increase MVO_2. A variety of imaging options are available to accompany these pharmacologic stressors (Fig. 267-3). The development of a transient perfusion defect with a tracer such as thallium-201 or 99m-technetium sestamibi is used to detect myocardial ischemia.

Echocardiography is used to assess LV function in patients with chronic stable angina and patients with a history of a prior myocardial infarction, pathologic Q waves, or clinical evidence of heart failure. Two-dimensional echocardiography can assess both global and regional wall motion abnormalities of the left ventricle that are transient when due to ischemia. Stress (exercise or dobutamine) echocardiography may cause the emergence of regions of akinesis or dyskinesis that are not present at rest. Stress echocardiography, like stress myocardial perfusion imaging, is more sensitive than exercise electrocardiography in the diagnosis of IHD. Cardiac magnetic resonance (CMR) stress testing is also evolving as an alternative to radionuclide, PET, or echocardiographic stress imaging. CMR stress testing performed with dobutamine infusion can be used to assess wall motion abnormalities accompanying ischemia, as well as myocardial perfusion. CMR can be used to provide more complete ventricular evaluation using multislice magnetic resonance imaging (MRI) studies.

Atherosclerotic plaques become progressively calcified over time, and coronary calcification in general increases with age. For this reason, methods for detecting coronary calcium have been developed as a measure of the presence of coronary atherosclerosis. These methods involve computed tomography (CT) applications that achieve rapid acquisition of images (electron beam [EBCT] and multidetector [MDCT] detection). Coronary calcium detected by these imaging techniques most commonly is quantified by using the Agatston score, which is based on the area and density of calcification. Although the diagnostic accuracy of this imaging method is high (sensitivity, 90–94%; specificity, 95–97%; negative predictive value, 93–99%), its prognostic utility has not been defined. Thus, its role in CT, EBCT, and MDCT scans for the detection and management of patients with IHD has not been clarified.

CORONARY ARTERIOGRAPHY

(See also Chap. 237) This diagnostic method outlines the lumina of the coronary arteries and can be used to detect or exclude serious coronary obstruction. However, coronary arteriography provides no information about the arterial wall, and severe atherosclerosis that does not encroach on the lumen may go undetected. Of note, atherosclerotic plaques characteristically are scattered throughout the coronary tree, tend to occur more frequently at branch points, and grow progressively in the intima and media of an epicardial coronary artery at first without encroaching on the lumen, causing an outward bulging of the artery—a process referred to as remodeling. Later in the course of the disease, further growth causes luminal narrowing.

Indications

Coronary arteriography is indicated in (1) patients with chronic stable angina pectoris who are severely symptomatic despite medical therapy and are being considered for revascularization, i.e., a percutaneous coronary intervention (PCI) or coronary artery bypass grafting (CABG); (2) patients with troublesome symptoms that present diagnostic difficulties in whom there is a need to confirm or rule out the diagnosis of IHD; (3) patients with known or possible angina pectoris who have survived cardiac arrest; (4) patients with angina or evidence of ischemia on noninvasive testing with clinical or laboratory evidence of ventricular dysfunction; and (5) patients judged to be at high risk of sustaining coronary events based on signs of severe ischemia on noninvasive testing, regardless of the presence or severity of symptoms (see below).

Examples of other indications for coronary arteriography include the following:

1. Patients with chest discomfort suggestive of angina pectoris but a negative or nondiagnostic stress test who require a definitive diagnosis for guiding medical management, alleviating psychological stress, career or family planning, or insurance purposes.
2. Patients who have been admitted repeatedly to the hospital for a suspected acute coronary syndrome **(Chaps. 268 and 269)**, but in whom this diagnosis has not been established and in whom the presence or absence of CAD should be determined.
3. Patients with careers that involve the safety of others (e.g., pilots, firefighters, police) who have questionable symptoms or suspicious or positive noninvasive tests and in whom there are reasonable doubts about the state of the coronary arteries.
4. Patients with aortic stenosis or hypertrophic cardiomyopathy and angina in whom the chest pain could be due to IHD.
5. Male patients >45 years and females >55 years who are to undergo a cardiac operation such as valve replacement or repair and who may or may not have clinical evidence of myocardial ischemia.
6. Patients after myocardial infarction, especially those who are at high risk after myocardial infarction because of the recurrence of angina or the presence of heart failure, frequent ventricular premature contractions, or signs of ischemia on the stress test.
7. Patients with angina pectoris, regardless of severity, in whom noninvasive testing indicates a high risk of coronary events (poor exercise performance or severe ischemia).
8. Patients in whom coronary spasm or another nonatherosclerotic cause of myocardial ischemia (e.g., coronary artery anomaly, Kawasaki disease) is suspected.

Noninvasive alternatives to diagnostic coronary arteriography include CT angiography and CMR angiography **(Chap. 236)**. Although these new imaging techniques can provide information about obstructive lesions in the epicardial coronary arteries, their exact role in clinical practice has not been rigorously defined. Important aspects of their use that should be noted include the substantially higher radiation exposure with CT angiography compared to conventional diagnostic arteriography and the limitations on CMR imposed by cardiac movement during the cardiac cycle, especially at high heart rates.

PROGNOSIS

The principal prognostic indicators in patients known to have IHD are age, the functional state of the left ventricle, the location(s) and severity of coronary artery narrowing, and the severity or activity of myocardial ischemia. Angina pectoris of recent onset, unstable angina **(Chap. 268)**, early

postmyocardial infarction angina, angina that is unresponsive or poorly responsive to medical therapy, and angina accompanied by symptoms of congestive heart failure all indicate an increased risk for adverse coronary events. The same is true for the physical signs of heart failure, episodes of pulmonary edema, transient third heart sounds, and mitral regurgitation and for echocardiographic or radioisotopic (or roentgenographic) evidence of cardiac enlargement and reduced (<0.40) ejection fraction.

Most important, any of the following signs during noninvasive testing indicates a high risk for coronary events: inability to exercise for 6 min, i.e., stage II (Bruce protocol) of the exercise test; a strongly positive exercise test showing onset of myocardial ischemia at low workloads (≥0.1 mV ST-segment depression before completion of stage II, ≥0.2 mV ST-segment depression at any stage, ST-segment depression for >5 min after the cessation of exercise, a decline in systolic pressure >10 mmHg during exercise, or the development of ventricular tachyarrhythmias during exercise); the development of large or multiple perfusion defects or increased lung uptake during stress radioisotope perfusion imaging; and a decrease in LV ejection fraction during exercise on radionuclide ventriculography or during stress echocardiography. Conversely, patients who can complete stage III of the Bruce exercise protocol and have a normal stress perfusion scan or negative stress echocardiographic evaluation are at very low risk for future coronary events. The finding of frequent episodes of ST-segment deviation on ambulatory ECG monitoring (even in the absence of symptoms) is also an adverse prognostic finding.

On cardiac catheterization, elevations of LV end-diastolic pressure and ventricular volume and reduced ejection fraction are the most important signs of LV dysfunction and are associated with a poor prognosis. Patients with chest discomfort but normal LV function and normal coronary arteries have an excellent prognosis. Obstructive lesions of the left main (>50% luminal diameter) or left anterior descending coronary artery proximal to the origin of the first septal artery are associated with a greater risk than are lesions of the right or left circumflex coronary artery because of the greater quantity of myocardium at risk. Atherosclerotic plaques in epicardial arteries with fissuring or filling defects indicate increased risk. These lesions go through phases of inflammatory cellular activity, degeneration, endothelial dysfunction, abnormal vasomotion, platelet aggregation, and fissuring or hemorrhage. These factors can temporarily worsen the stenosis and cause thrombosis and/or abnormal reactivity of the vessel wall, thus exacerbating the manifestations of ischemia. The recent onset of symptoms, the development of severe ischemia during stress testing (see above), and unstable angina pectoris **(Chap. 268)** all reflect episodes of rapid progression in coronary lesions.

With any degree of obstructive CAD, mortality is greatly increased when LV function is impaired; conversely, at any level of LV function, the prognosis is influenced importantly by the quantity of myocardium perfused by critically obstructed vessels. Therefore, it is essential to collect all the evidence substantiating past myocardial damage (evidence of myocardial infarction on ECG, echocardiography, radioisotope imaging, or left ventriculography), residual LV function (ejection fraction and wall motion), and risk of future damage from coronary events (extent of coronary disease and severity of ischemia defined by noninvasive stress testing). The larger the quantity of established myocardial necrosis is, the less the heart is able to withstand additional damage and the poorer the prognosis is. Risk estimation must include age, presenting symptoms, all risk factors, signs of arterial disease, existing cardiac damage, and signs of impending damage (i.e., ischemia).

The greater the number and severity of risk factors for coronary atherosclerosis (advanced age [>75 years], hypertension, dyslipidemia, diabetes, morbid obesity, accompanying peripheral and/or cerebrovascular disease, previous myocardial infarction), the worse the prognosis of an angina patient. Evidence exists that elevated levels of CRP in the plasma, extensive coronary calcification on electron beam CT (see above), and increased carotid intimal thickening on ultrasound examination also indicate an increased risk of coronary events.

TREATMENT

TREATMENT Stable Angina Pectoris

Once the diagnosis of IHD has been made, each patient must be evaluated individually with respect to his or her level of understanding, expectations and goals, control of symptoms, and prevention of adverse clinical outcomes such as myocardial infarction and premature death. The degree of disability and the physical and emotional stress that precipitates angina must be recorded carefully to set treatment goals. The management plan should include the following components: (1) explanation of the problem and reassurance about the ability to formulate a treatment plan, (2) identification and treatment of aggravating conditions, (3) recommendations for adaptation of activity as needed, (4) treatment of risk factors that will decrease the occurrence of adverse coronary outcomes, (5) drug therapy for angina, and (6) consideration of revascularization.

EXPLANATION AND REASSURANCE

Patients with IHD need to understand their condition and realize that a long and productive life is possible even though they have angina pectoris or have experienced and recovered from an acute myocardial infarction. Offering results of clinical trials showing improved outcomes can be of great value in encouraging patients to resume or maintain activity and return to work. A planned program of rehabilitation can encourage patients to lose weight, improve exercise tolerance, and control risk factors with more confidence.

IDENTIFICATION AND TREATMENT OF AGGRAVATING CONDITIONS

A number of conditions may increase oxygen demand or decrease oxygen supply to the myocardium and may precipitate or exacerbate angina in patients with IHD. LVH, aortic valve disease, and hypertrophic cardiomyopathy may cause or contribute to angina and should be excluded or treated. Obesity, hypertension, and hyperthyroidism should be treated aggressively to reduce the frequency and severity of anginal episodes. Decreased myocardial oxygen supply may be due to reduced oxygenation of the arterial blood (e.g., in pulmonary disease or, when carboxyhemoglobin is present, due to cigarette or cigar smoking) or decreased oxygen-carrying capacity (e.g., in anemia). Correction of these abnormalities, if present, may reduce or even eliminate angina pectoris.

ADAPTATION OF ACTIVITY

Myocardial ischemia is caused by a discrepancy between the demand of the heart muscle for oxygen and the ability of the coronary circulation to meet that demand. Most patients can be helped to understand this concept and utilize it in the rational programming of activity. Many tasks that ordinarily evoke angina may be accomplished without symptoms simply by reducing the speed at which they are performed. Patients must appreciate the diurnal variation in their tolerance of certain activities and should reduce their energy requirements in the morning, immediately after meals, and in cold or inclement weather. On occasion, it may be necessary to recommend a change in employment or residence to avoid physical stress.

Physical conditioning usually improves the exercise tolerance of patients with angina and has substantial psychological benefits. A regular program of isotonic exercise that is within the limits of the individual patient's threshold for the development of angina pectoris and that does not exceed 80% of the heart rate associated with ischemia on exercise testing should be strongly encouraged. Based on the results of an exercise test, the number

of metabolic equivalent tasks (METs) performed at the onset of ischemia can be estimated (Table 267-2) and a practical exercise prescription can be formulated to permit daily activities that will fall below the ischemic threshold **(Table 267-3)**.

TREATMENT OF RISK FACTORS

A *family history* of premature IHD is an important indicator of increased risk and should trigger a search for treatable risk factors such as hyperlipidemia, hypertension, and diabetes mellitus. *Obesity* impairs the treatment of other risk factors and increases the risk of adverse coronary events. In addition, obesity often is accompanied by three other risk factors: diabetes mellitus, hypertension, and hyperlipidemia. The treatment of obesity and these accompanying risk factors is an important component of any management plan. A diet low in saturated and *trans*-unsaturated fatty acids and a reduced caloric intake to achieve optimal body weight are a cornerstone in the management of chronic IHD. It is especially important to emphasize weight loss and regular exercise in patients with the metabolic syndrome or overt diabetes mellitus.

Cigarette smoking accelerates coronary atherosclerosis in both sexes and at all ages and increases the risk of thrombosis, plaque instability, myocardial infarction, and death. In addition, by increasing myocardial oxygen needs and reducing oxygen supply, it aggravates angina. Smoking cessation studies have demonstrated important benefits with a significant decline in the occurrence of these adverse outcomes. Noncombustible tobacco in the form of electronic cigarettes (nicotine delivery systems) may also increase the frequency of anginal episodes. The physician's message must be clear and strong and supported by programs that achieve and monitor abstinence of tobacco product use **(Chap. 448)**.

Hypertension **(Chap. 271)** is associated with an increased risk of adverse clinical events from coronary atherosclerosis as well as stroke. In addition, the LVH that results from sustained hypertension aggravates ischemia. There is evidence that long-term effective treatment of hypertension can decrease the occurrence of adverse coronary events **(Chap. 271)**.

Diabetes mellitus **(Chap. 396)** accelerates coronary and peripheral atherosclerosis and is frequently associated with dyslipidemias and increases in the risk of angina, myocardial infarction, and sudden coronary death. Aggressive control of the dyslipidemia (target LDL cholesterol <70 mg/dL) and hypertension (target blood pressure 120/80 mmHg) that are frequently found in diabetic patients is highly effective and therefore essential, as described below.

DYSLIPIDEMIA

The treatment of dyslipidemia is central in aiming for long-term relief from angina, reduced need for revascularization, and reduction in myocardial infarction and death. The control of lipids can be achieved by the combination of a diet low in saturated and *trans*-unsaturated fatty acids, exercise, and weight loss. Nearly always, HMG-CoA reductase inhibitors (statins) are required and can lower LDL cholesterol (25–50%), raise HDL cholesterol (5–9%), and lower triglycerides (5–30%). A powerful treatment effect of statins on atherosclerosis, IHD, and outcomes is seen regardless of the pretreatment LDL cholesterol level. Fibrates or niacin can be used to raise HDL cholesterol and lower triglycerides **(Chap. 400)**. Controlled trials with lipid-regulating regimens have shown equal proportional benefit for men, women, the elderly, diabetic patients, and smokers. Injectable monoclonal antibodies against PCSK9 are now available and are capable of producing dramatic lowering of LDL cholesterol beyond that achieved with a statin alone.

Compliance with the health-promoting behaviors listed above is generally very poor, and a conscientious physician must not underestimate the major effort required to meet this challenge. Many patients who are discharged from the hospital with proven coronary disease do not receive adequate treatment for dyslipidemia. In light of the proof that treating dyslipidemia brings major benefits, physicians need to establish treatment pathways, monitor compliance, and follow up regularly.

RISK REDUCTION IN WOMEN WITH IHD

The incidence of clinical IHD in premenopausal women is very low; however, after menopause, the atherogenic risk factors increase (e.g., increased LDL, reduced HDL) and the rate of clinical coronary events accelerates to the levels observed in men. Women have not given up cigarette smoking as effectively as have men. Diabetes mellitus, which is more common in women, greatly increases the occurrence of clinical IHD and amplifies the deleterious effects of hypertension, hyperlipidemia, and smoking. Cardiac catheterization and coronary revascularization are underused in women and are performed at a later and more severe stage of the disease than in men. When cholesterol lowering, beta blockers after myocardial infarction, and CABG are applied in the appropriate patient groups, women benefit to the same degree as men.

DRUG THERAPY

The commonly used drugs for the treatment of angina pectoris are summarized in **Tables 267-4**,**267-5**,**267-6**. Pharmacotherapy for IHD is designed to reduce the frequency of anginal episodes, myocardial infarction, and coronary death. Trial data emphasize how important medical management is when added to the health-promoting behaviors discussed above. To achieve maximum benefit from medical therapy for IHD, it is frequently necessary to combine agents from different classes and titrate the doses as guided by the individual profile of risk factors, symptoms, hemodynamic responses, and side effects.

NITRATES

The organic nitrates are a valuable class of drugs in the management of angina pectoris (Table 267-4). Their major mechanisms of action include systemic venodilation with concomitant reduction in LV end-diastolic volume and pressure, thereby reducing myocardial wall tension and oxygen requirements; dilation of epicardial coronary vessels; and increased blood flow in collateral vessels. When metabolized, organic nitrates release nitric oxide (NO) that binds to guanylyl cyclase in vascular smooth muscle cells, leading to an increase in cyclic guanosine monophosphate, which causes relaxation of vascular smooth muscle. Nitrates also exert antithrombotic activity by NO-dependent activation of platelet guanylyl cyclase, impairment of intraplatelet calcium flux, and platelet activation.

The absorption of these agents is rapid and complete through mucous membranes. For this reason, nitroglycerin is most commonly administered sublingually in tablets of 0.4 or 0.6 mg. Patients with angina should be instructed to take the medication both to relieve angina and also ~5 min before activities that are likely to induce an episode. The value of this prophylactic use of the drug cannot be overemphasized.

Nitrates improve exercise tolerance in patients with chronic angina and relieve ischemia in patients with unstable angina as well as patients with Prinzmetal's variant angina **(Chap. 268)**. A diary of angina and nitroglycerin use may be valuable for detecting changes in the frequency, severity, or threshold for discomfort that may signify the development of unstable angina pectoris and/or herald an impending myocardial infarction.

Long-Acting Nitrates

None of the long-acting nitrates is as effective as sublingual nitroglycerin for the acute relief of angina. These organic nitrate preparations can be swallowed, chewed, or administered as a patch or paste by the transdermal route (Table 267-4). They provide effective plasma levels for up to 24 h, but the therapeutic response is highly variable. Different preparations and/or administration during the daytime should be tried only to prevent discomfort while avoiding side effects such as headache and dizziness. Individual dose titration is important to prevent side effects. To minimize the effects of nitrate tolerance, the minimum effective dose should be used and a minimum of 8 h each day kept free of the drug to restore any useful response(s).

β-Adrenergic Blockers

These drugs represent an important component of the pharmacologic treatment of angina pectoris (Table 267-5). They reduce myocardial oxygen demand by inhibiting the increases in heart rate, arterial pressure, and myocardial contractility caused by adrenergic activation. Beta blockade reduces these variables most strikingly during exercise but causes only small reductions at rest. Long-acting beta-blocking drugs or sustained-release formulations offer the advantage of once-daily dosing (Table 267-5). The therapeutic aims include relief of angina and ischemia. These drugs also can reduce mortality and reinfarction rates in patients after myocardial infarction and are moderately effective antihypertensive agents.

Relative contraindications include asthma and reversible airway obstruction in patients with chronic lung disease, atrioventricular conduction disturbances, severe bradycardia, Raynaud's phenomenon, and a history of mental depression. Side effects include fatigue, reduced exercise tolerance, nightmares, impotence, cold extremities, intermittent claudication, bradycardia (sometimes severe), impaired atrioventricular conduction, LV failure, bronchial asthma, worsening claudication, and intensification of the hypoglycemia produced by oral hypoglycemic agents and insulin. Reducing the dose or even discontinuation may be necessary if these side effects develop and persist. Since sudden discontinuation can intensify ischemia, the doses should be tapered over 2 weeks. Beta blockers with relative β_1-receptor specificity such as metoprolol and atenolol may be preferable in patients with mild bronchial obstruction and insulin-requiring diabetes mellitus.

Calcium Channel Blockers

Calcium channel blockers (Table 267-6) are coronary vasodilators that produce variable and dose-dependent reductions in myocardial oxygen demand, contractility, and arterial pressure. These combined pharmacologic effects are advantageous and make these agents as effective as beta blockers in the treatment of angina pectoris. They are indicated when beta blockers are contraindicated, poorly tolerated, or ineffective. Because of differences in the dose-response relationship on cardiac electrical activity between the dihydropyridine and nondihydropyridine calcium channel blockers, verapamil and diltiazem may produce symptomatic disturbances in cardiac conduction and bradyarrhythmias. They also exert negative inotropic actions and are more likely to aggravate LV failure, particularly when used in patients with LV dysfunction, especially if the patients are also receiving beta blockers. Although useful effects usually are achieved when calcium channel blockers are combined with beta blockers and nitrates, individual titration of the doses is essential with these combinations. Variant (Prinzmetal's) angina responds particularly well to calcium channel blockers (especially members of the dihydropyridine class), supplemented when necessary by nitrates **(Chap. 268)**.

Verapamil ordinarily should not be combined with beta blockers because of the combined adverse effects on heart rate and contractility. Diltiazem can be combined with beta blockers in patients with normal ventricular function and no conduction disturbances. Amlodipine and beta blockers have complementary actions on coronary blood supply and myocardial oxygen demands. Whereas the former decreases blood pressure and dilates coronary arteries, the latter slows heart rate and decreases contractility. Amlodipine and the other second-generation dihydropyridine calcium antagonists (nicardipine, isradipine, long-acting nifedipine, and felodipine) are potent vasodilators and are useful in the simultaneous treatment of angina and hypertension. Short-acting dihydropyridines should be avoided because of the risk of precipitating infarction, particularly in the absence of concomitant beta blocker therapy.

Choice Between Beta Blockers and Calcium Channel Blockers for Initial Therapy

Since beta blockers have been shown to improve life expectancy after acute myocardial infarction **(Chaps. 268 and 269)** and calcium channel blockers have not, the former may also be preferable in patients with angina and a damaged left ventricle. However, calcium channel blockers are indicated in patients with the following: (1) inadequate responsiveness to the combination of beta blockers and nitrates; many of these patients do well with a combination of a beta blocker and a dihydropyridine calcium channel blocker; (2) adverse reactions to beta blockers such as depression, sexual disturbances, and fatigue; (3) angina and a history of asthma or chronic obstructive pulmonary disease; (4) sick-sinus syndrome or significant atrioventricular conduction disturbances; (5) Prinzmetal's angina; or (6) symptomatic peripheral arterial disease.

A comparison of the common side effects, contraindications, and potential drug interactions of many of the frequently presented antianginal agents is shown in **Table 267-7**.

Antiplatelet Drugs

Aspirin is an irreversible inhibitor of platelet cyclooxygenase and thereby interferes with platelet activation. Chronic administration of 75–325 mg orally per day has been shown to reduce coronary events in asymptomatic adult men over age 50, patients with chronic stable angina, and patients who have or have survived unstable angina and myocardial infarction. There is a dose-dependent increase in bleeding when aspirin is used chronically. It is preferable to use an enteric-coated formulation in the range of 81–162 mg/d. Administration of this drug should be considered in all patients with IHD in the absence of gastrointestinal bleeding, allergy, or dyspepsia. Clopidogrel (300–600 mg loading and 75 mg/d) is an oral agent that blocks P2Y12 ADP receptor–mediated platelet aggregation. It provides benefits similar to those of aspirin in patients with stable chronic IHD and may be substituted for aspirin if aspirin causes the side effects listed above. Clopidogrel combined with aspirin reduces death and coronary ischemic events in patients with an acute coronary syndrome **(Chap. 268)** and also reduces the risk of thrombus formation in patients undergoing implantation of a stent in a coronary artery **(Chap. 270)**. Alternative antiplatelet agents that block the P2Y12 platelet receptor such as prasugrel and ticagrelor have been shown to be more effective than clopidogrel for prevention of ischemic events after placement of a stent for an acute coronary syndrome, but are associated with an increased risk of bleeding. Although combined treatment with clopidogrel and aspirin for at least a year is recommended in patients with an acute coronary syndrome treated with implantation of a drug-eluting stent, studies have not shown any benefit from the routine addition of clopidogrel to aspirin in patients with chronic stable IHD.

OTHER THERAPIES

The ACE inhibitors are widely used in the treatment of survivors of myocardial infarction, patients with hypertension or chronic IHD including angina pectoris, and those at high risk of vascular diseases such as diabetes. The benefits of ACE inhibitors are most evident in IHD patients at increased risk, especially if diabetes mellitus or left ventricle dysfunction is present, and those who have not achieved adequate control of blood pressure and LDL cholesterol on beta blockers and statins. However, the routine administration of ACE inhibitors to IHD patients who have normal LV function and have achieved blood pressure and LDL goals on other therapies does not reduce the incidence of events and therefore is not cost-effective.

Despite treatment with nitrates, beta blockers, or calcium channel blockers, some patients with IHD continue to experience angina, and additional medical therapy is now available to alleviate their symptoms. Ranolazine, a piperazine derivative, may be useful for patients with chronic angina despite standard medical therapy (see Table 267-7). Its antianginal action is believed to occur via inhibition of the late inward sodium current (I_{Na}). The benefits of I_{Na} inhibition include limitation of the Na overload of ischemic myocytes and prevention of Ca^{2+} overload via the Na^{+}–Ca^{2+} exchanger. A dose of 500–1000 mg orally twice daily is usually well tolerated. Ranolazine is contraindicated in patients with hepatic impairment or with conditions or drugs associated with QT_c prolongation and when drugs that inhibit the CYP3A metabolic system (e.g., ketoconazole, diltiazem, verapamil, macrolide antibiotics, HIV protease inhibitors, and large quantities of grapefruit juice) are being used.

Nonsteroidal anti-inflammatory drug (NSAID) use in patients with IHD may be associated with a small but finite increased risk of myocardial infarction and mortality. For this reason, they generally should be avoided in IHD patients. If they are required for symptom relief, it is advisable to coadminister aspirin and strive to use an NSAID associated with the lowest risk of cardiovascular events, in the lowest dose required, and for the shortest period of time.

Another class of agents opens ATP-sensitive potassium channels in myocytes, leading to a reduction of free intracellular calcium ions. The major drug in this class is nicorandil, which typically is administered orally in a dose of 20 mg twice daily for prevention of angina. (Nicorandil is not available for use in the United States but is used in several other countries.)

Ivabradine (2.5–7.5 mg orally twice daily) is a specific sinus node inhibiting agent that may be helpful for preventing cardiovascular events in patients with IHD who have a resting heart rate ≥70 beats/min (alone or in combination with a beta blocker) *and* LV systolic dysfunction. It does not appear to offer any benefit in patients with IHD but who do *not* have clinical heart failure.

Angina and Heart Failure

Transient LV failure with angina can be controlled by the use of nitrates. For patients with established congestive heart failure, the increased LV wall tension raises myocardial oxygen demand. Treatment of congestive heart failure with an ACE inhibitor, a diuretic, and digoxin **(Chap. 252)** reduces heart size, wall tension, and myocardial oxygen demand, which helps control angina and ischemia. If the symptoms and signs of heart failure are controlled, an effort should be made to use beta blockers not only for angina but because trials in heart failure have shown significant improvement in survival. A trial of the intravenous ultra-short-acting beta blocker esmolol may be useful to establish the safety of beta blockade in selected patients. Nocturnal angina often can be relieved by the treatment of heart failure.

The combination of congestive heart failure and angina in patients with IHD usually indicates a poor prognosis and warrants serious consideration of cardiac catheterization and coronary revascularization.

TABLE 267-3 Energy Requirements for Some Common Activities

Less than 3 METs	3–5 METs	5–7 METs	7–9 METs	More than 9 METs
Self-Care				
Washing/shaving	Cleaning windows	Easy digging in garden		Carrying loads up stairs (objects more than 90 lb)
Dressing	Raking	Level hand lawn mowing	Heavy shoveling	Climbing stairs (quickly)
Light housekeeping	Power lawn mowing	Carrying objects (30–60 lb)	Carrying objects (60–90 lb)	Shoveling heavy snow
Desk work	Bed making/stripping			
Driving auto	Carrying objects (15-30 lb)			
Occupational				
Sitting (clerical/assembly)	Stocking shelves (light objects)	Carpentry (exterior)	Digging ditches (pick and shovel)	Heavy labor
Desk work	Light welding/carpentry	Shoveling dirt		
Standing (store clerk)		Sawing wood		
Recreational				
Golf (cart)	Dancing (social)	Tennis (singles)	Canoeing	Squash
Knitting	Golf (walking)	Snow skiing (downhill)	Mountain climbing	Ski touring
	Sailing	Light backpacking		Vigorous basketball
	Tennis (doubles)	Basketball		
		Stream fishing		
Physical Conditioning				
Walking (2 mph)	Level walking (3–4 mph)	Level walking (4.5–5.0 mph)	Level jogging (5 mph)	Running more than 6 mph
Stationary bike	Level biking (6–8 mph)	Bicycling (9–10 mph)	Swimming (crawl stroke)	Bicycling (more than 13 mph)
Very light calisthenics	Light calisthenics	Swimming, breast stroke	Rowing machine	Rope jumping
			Heavy calisthenics	Walking uphill (5 mph)
			Bicycling (12 mph)	

Abbreviation: METs, metabolic equivalent tasks.

Source: Modified from WL Haskell: Rehabilitation of the coronary patient, in NK Wenger, HK Hellerstein (eds): *Design and Implementation of Cardiac Conditioning Program*. New York, Churchill Livingstone, 1978.

TABLE 267-4 Nitrate Therapy in Patients with Ischemic Heart Disease

Preparation of Agent	Dose	Schedule
Nitroglycerin[a]		
Ointment	0.5–2 in.	Two or three times daily
Transdermal patch	0.2–0.8 mg/h	Every 24 h; remove at bedtime for 12–14 h
Sublingual tablet	0.3–0.6 mg	As needed, up to three doses 5 min apart
Spray	One or two sprays	As needed, up to three doses 5 min apart
Isosorbide dinitrate[a]		
Oral	10–40 mg	Two or three times daily
Oral sustained release	80–120 mg	Once or twice daily (eccentric schedules)
Isosorbide 5-mononitrate		
Oral	20 mg	Twice daily (given 7–8 h apart)
Oral sustained release	30–240 mg	Once daily

[a]A 10- to 12-h nitrate-free interval is recommended.

Source: Modified from DA Morrow, WE Boden: Stable ischemic heart disease. In RO Bonow et al (eds): *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*, 9th ed. Philadelphia, Saunders, 2012, p. 1224.

TABLE 267-5 Properties of Beta Blockers in Clinical Use for Ischemic Heart Disease

Drugs	Selectivity	Partial Agonist Activity	Usual Dose for Angina
Acebutolol	β_1	Yes	200–600 mg twice daily
Atenolol	β_1	No	50–200 mg/d
Betaxolol	β_1	No	10–20 mg/d
Bisoprolol	β_1	No	10 mg/d
Esmolol (intravenous)[a]	β_1	No	50–300 μg/kg/min
Labetalol[b]	None	Yes	200–600 mg twice daily
Metoprolol	β_1	No	50–200 mg twice daily
Nadolol	None	No	40–80 mg/d
Nebivolol	β_1 (at low doses)	No	5–40 mg/d
Pindolol	None	Yes	2.5–7.5 mg 3 times daily
Propranolol	None	No	80–120 mg twice daily
Timolol	None	No	10 mg twice daily

[a]Esmolol is an ultra-short-acting beta blocker that is administered as a continuous intravenous infusion. Its rapid offset of action makes esmolol an attractive agent to use in patients with relative contraindications to beta blockade. [b]Labetolol is a combined alpha and beta blocker.

Note: This list of beta blockers that may be used to treat patients with angina pectoris is arranged alphabetically. The agents for which there is the greatest clinical experience include atenolol, metoprolol, and propranolol. It is preferable to use a sustained-release formulation that may be taken once daily to improve the patient's compliance with the regimen.

Source: Data from RJ Gibbons et al: J Am Coll Cardiol 41:159, 2003.

TABLE 267-6 Calcium Channel Blockers in Clinical Use for Ischemic Heart Disease

Drugs	Usual Dose	Duration of Action	Side Effects
Dihydropyridines			
Amlodipine	5–10 mg qd	Long	Headache, edema
Felodipine	5–10 mg qd	Long	Headache, edema
Isradipine	2.5–10 mg bid	Medium	Headache, fatigue
Nicardipine	20–40 mg tid	Short	Headache, dizziness, flushing, edema
Nifedipine	Immediate release:[a] 30–90 mg daily orally Slow release: 30–180 mg orally	Short	Hypotension, dizziness, flushing, nausea, constipation, edema
Nisoldipine	20–40 mg qd	Short	Similar to nifedipine
Nondihydropyridines			
Diltiazem	Immediate release: 30–80 mg 4 times daily Slow release: 120–320 mg qd	Short Long	Hypotension, dizziness, flushing, bradycardia, edema
Verapamil	Immediate release: 80–160 mg tid Slow release: 120–480 mg qd	Short Long	Hypotension, myocardial depression, heart failure, edema, bradycardia

[a]May be associated with increased risk of mortality if administered during acute myocardial infarction.

Note: This list of calcium channel blockers that may be used to treat patients with angina pectoris is divided into two broad classes, dihydropyridines and nondihydropyridines, and arranged alphabetically within each class. Among the dihydropyridines, the greatest clinical experience has been obtained with amlodipine and nifedipine. After the initial period of dose titration with a short-acting formulation, it is preferable to switch to a sustained-release formulation that may be taken once daily to improve patient compliance with the regimen.

Source: Data from RJ Gibbons et al: J Am Coll Cardiol 41:159, 2003.

TABLE 267-7 Antianginal Agents

Agent	Common Side Effects	Contraindications	Potential Drug Interactions
Agents That Have a Physiological Effect			
Short-acting and long-acting nitrates	Headache, flushing, hypotension, syncope and postural hypotension, reflex tachycardia, methemoglobinemia	Hypertrophic obstructive cardiomyopathy	Phosphodiesterase type 5 inhibitors (sildenafil and similar agents), beta-adrenergic blockers, calcium-channel blockers
Beta blockers	Fatigue, depression, bradycardia, heart block, bronchospasm, peripheral vasoconstriction, postural hypotension, impotence, masked signs of hypoglycemia	Low heart rate or heart conduction disorder, cardiogenic shock, asthma, severe peripheral vascular disease, decompensated heart failure, vasospastic angina; use with caution in patients with COPD (cardio-selective beta blockers may be used if patient receives adequate treatment with long-acting beta agonists)	Heart-rate-lowering calcium-channel blockers, sinus-node or AV conduction depressors
Calcium-channel blockers			
Heart-rate-lowering agents	Bradycardia, heart conduction defect, low ejection fraction, constipation, gingival hyperplasia	Cardiogenic shock, severe aortic stenosis, obstructive cardiomyopathy	CYP3A4 substrates (digoxin, simvastatin, cyclosporine)
Dihydropyridine	Headache, ankle swelling fatigue, flushing, reflex tachycardia	Low heart rate or heart rhythm disorder, sick sinus syndrome, congestive heart failure, low blood pressure	Agents with cardiodepressant effects (beta-blockers, flecainide), CYP3A4 substrates
Agents That Affect Myocardial Metabolism			
Ranolazine	Dizziness, constipation, nausea, QT-interval prolongation	Liver cirrhosis	CYP3A4 substrates (digoxin, simvastatin, cyclosporine), drugs that prolong the corrected QT interval

Abbreviations: COPD, chronic obstructive pulmonary disease; CYP3A4, cytochrome P-450 3A4.

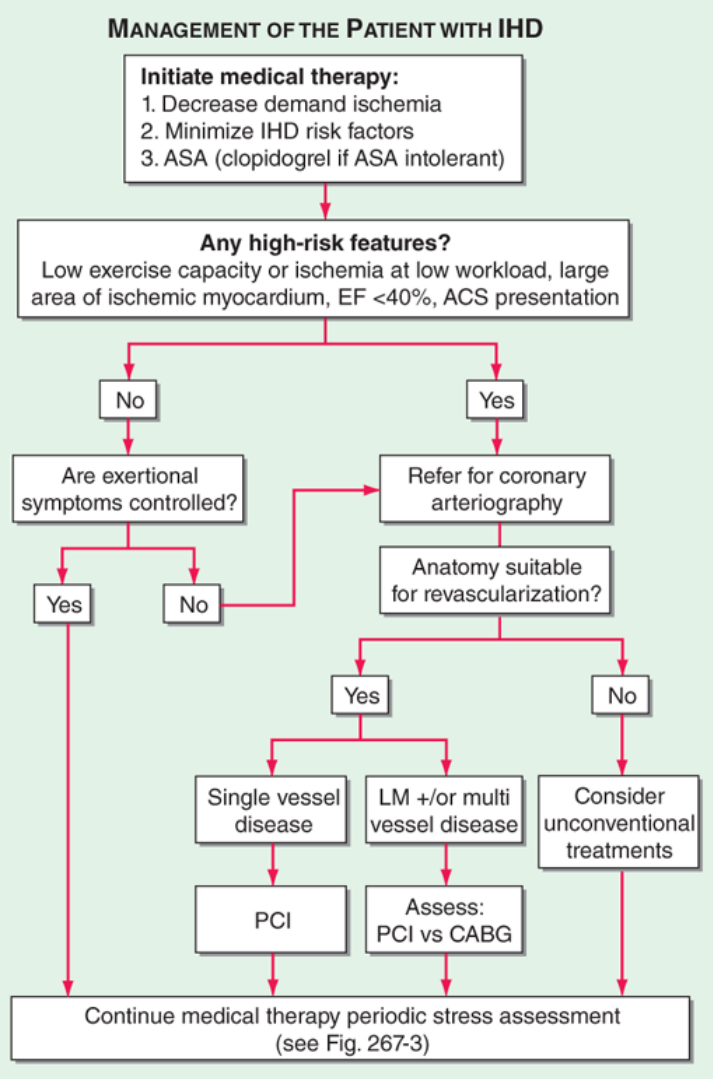
Source: Data from SE Husted: Lancet 386:691, 2015 and EM Ohman: N Engl J Med 374:1167, 2016.

CORONARY REVASCULARIZATION

Clinical trials have confirmed that with the initial diagnosis of stable IHD, it is first appropriate to initiate a medical regimen as described above. Revascularization should be considered in the presence of unstable phases of the disease, intractable symptoms, severe ischemia or high-risk coronary anatomy, diabetes, and impaired LV function. *Revascularization should be employed in conjunction with but not replace the continuing need to modify risk factors and assess medical therapy.* An algorithm for integrating medical therapy and revascularization options in patients with IHD is shown in **Fig. 267-4**.

FIGURE 267-4

Algorithm for management of a patient with ischemic heart disease. All patients should receive the core elements of medical therapy as shown at the top of the algorithm. If high-risk features are present, as established by the clinical history, exercise test data, and imaging studies, the patient should be referred for coronary arteriography. Based on the number and location of the diseased vessels and their suitability for revascularization, the patient is treated with a percutaneous coronary intervention (PCI) or coronary artery bypass graft (CABG) surgery or should be considered for unconventional treatments. See text for further discussion. ACS, acute coronary syndrome; ASA, aspirin; EF, ejection fraction; IHD, ischemic heart disease; LM, left main.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PERCUTANEOUS CORONARY INTERVENTION

(See also Chap. 270) PCI involving balloon dilatation usually accompanied by coronary stenting is widely used to achieve revascularization of the myocardium in patients with symptomatic IHD and suitable stenoses of epicardial coronary arteries. Whereas patients with stenosis of the left main coronary artery and those with three-vessel IHD (especially with diabetes and/or impaired LV function) who require revascularization are best treated with CABG, PCI is widely employed in patients with symptoms and evidence of ischemia due to stenoses of one or two vessels and even in selected patients with three-vessel disease (and, perhaps, in some patients with left main disease) and may offer many advantages over surgery.

Indications and Patient Selection

The most common clinical indication for PCI is symptom-limiting angina pectoris, despite medical therapy, accompanied by evidence of ischemia during a stress test. PCI is more effective than medical therapy for the relief of angina. PCI improves outcomes in patients with unstable angina or when used early in the course of myocardial infarction with and without cardiogenic shock. However, in patients with stable exertional angina, clinical trials have confirmed that PCI does not reduce the occurrence of death or myocardial infarction compared to optimum medical therapy. PCI can be used to treat stenoses in native coronary arteries as well as in bypass grafts in patients who have recurrent angina after CABG.

Risks

When coronary stenoses are discrete and symmetric, two and even three vessels can be treated in sequence. However, case selection is essential to avoid a prohibitive risk of complications, which are usually due to dissection or thrombosis with vessel occlusion, uncontrolled ischemia, and

ventricular failure **(Chap. 270)**. Oral aspirin, a $P2Y_{12}$ antagonist, and an antithrombin agent are given to reduce coronary thrombus formation. Left main coronary artery stenosis generally is regarded as a lesion that should be treated with CABG. In selected cases such as patients with prohibitive surgical risks, PCI of an unprotected left main can be considered, but such a procedure should be performed only by a highly skilled operator; importantly, there are regional differences in the use of this approach internationally.

Efficacy

Primary success, i.e., adequate dilation (an increase in luminal diameter >20% to a residual diameter obstruction <50%) with relief of angina, is achieved in >95% of cases. Recurrent stenosis of the dilated vessels occurs in ~20% of cases within 6 months of PCI with bare metal stents, and angina will recur within 6 months in 10% of cases. Restenosis is more common in patients with diabetes mellitus, arteries with small caliber, incomplete dilation of the stenosis, long stents, occluded vessels, obstructed vein grafts, dilation of the left anterior descending coronary artery, and stenoses containing thrombi. In diseased vein grafts, procedural success has been improved by the use of capture devices or filters that prevent embolization, ischemia, and infarction.

It is usual clinical practice to administer aspirin indefinitely and a $P2Y_{12}$ antagonist for 1–3 months after the implantation of a bare metal stent. Although aspirin in combination with a thienopyridine may help prevent coronary thrombosis during and shortly after PCI with stenting, there is no evidence that these medications reduce the incidence of restenosis.

The use of drug-eluting stents that locally deliver antiproliferative drugs can reduce restenosis to much less than 10%. Advances in PCI, especially the availability of drug-eluting stents, have vastly extended the use of this revascularization option in patients with IHD. Of note, however, the delayed endothelial healing in the region of a drug-eluting stent also extends the period during which the patient is at risk for subacute stent thrombosis. Aspirin administered indefinitely and a P2Y12 antagonist daily (dual antiplatelet therapy [DAPT]) for at least 1 year after implantation of a drug-eluting stent. Evidence exists of a benefit of continuing DAPT for up to 30 months, albeit at the cost of a higher risk of bleeding. When a situation arises in which temporary discontinuation of antiplatelet therapy is necessary, the clinical circumstances should be reviewed with the operator who performed the PCI and a coordinated plan should be established for minimizing the risk of late stent thrombus; central to this plan is the discontinuation of antiplatelet therapy for the shortest acceptable period. The risk of stent thrombosis is dependent on stent size and length, complexity of the lesions, age, diabetes, and technique. However, compliance with DAPT and individual responsiveness to platelet inhibition are very important factors as well.

Successful PCI produces effective relief of angina in >95% of cases. The majority of patients with symptomatic IHD who require revascularization can be treated initially by PCI. Successful PCI is less invasive and expensive than CABG and permits savings in the *initial* cost of care. Successful PCI avoids the risk of stroke associated with CABG surgery, allows earlier return to work, and allows the resumption of an active life. However, the early health-related and economic benefit of PCI is reduced over time because of the greater need for follow-up and the increased need for repeat procedures. When directly compared in patients with diabetes or three-vessel or left main CAD, CABG was superior to PCI in preventing major adverse cardiac or cerebrovascular events over a 12-month follow-up.

CORONARY ARTERY BYPASS GRAFTING

Anastomosis of one or both of the internal mammary arteries or a radial artery to the coronary artery distal to the obstructive lesion is the preferred procedure. For additional obstructions that cannot be bypassed by an artery, a section of a vein (usually the saphenous) is used to form a venous bypass conduit between the aorta and the coronary artery distal to the obstructive lesion.

Although some indications for CABG are controversial, certain areas of agreement exist:

1. The operation is relatively safe, with mortality rates <1% in patients without serious comorbid disease and normal LV function and when the procedure is performed by an experienced surgical team.

2. Intraoperative and postoperative mortality rates increase with the severity of ventricular dysfunction, comorbidities, age >80 years, and lack of surgical experience. The effectiveness and risk of CABG vary widely depending on case selection and the skill and experience of the surgical team.

3. Occlusion of *venous* grafts is observed in 10–20% of patients during the first postoperative year and in ~2% per year during 5- to 7-year follow-up and 4% per year thereafter. Long-term patency rates are considerably higher for internal mammary and radial artery implantations than for saphenous vein grafts. In patients with left anterior descending coronary artery obstruction, survival is better when coronary bypass involves the internal mammary artery rather than a saphenous vein. Graft patency and outcomes are improved by meticulous treatment of risk factors, particularly dyslipidemia.

4. Angina is abolished or greatly reduced in ~90% of patients after complete revascularization. Although this usually is associated with graft patency and restoration of blood flow, the pain may also have been alleviated as a result of infarction of the ischemic segment or a placebo effect. Within 3 years, angina recurs in about one-fourth of patients but is rarely severe.

5. Survival may be improved by operation in patients with stenosis of the left main coronary artery as well as in patients with three- or two-vessel disease with significant obstruction of the proximal left anterior descending coronary artery. The survival benefit is greater in patients with abnormal LV function (ejection fraction <50%). Survival *may* also be improved in the following patients: (a) patients with obstructive CAD who have survived sudden cardiac death or sustained ventricular tachycardia; (b) patients who have undergone previous CABG and have multiple saphenous vein graft stenoses, especially of a graft supplying the left anterior descending coronary artery; and (c) patients with recurrent stenosis after PCI and high-risk criteria on noninvasive testing.

6. Minimally invasive CABG through a small thoracotomy and/or off-pump surgery can reduce morbidity and shorten convalescence in suitable patients but does not appear to reduce significantly the risk of neurocognitive dysfunction postoperatively.

7. Among patients with type 2 diabetes mellitus and multivessel coronary disease, CABG surgery plus optimal medical therapy is superior to optimal medical therapy alone in preventing major cardiovascular events, a benefit mediated largely by a significant reduction in nonfatal myocardial infarction. The benefits of CABG are especially evident in diabetic patients treated with an insulin-sensitizing strategy as opposed to an insulin-providing strategy. CABG has also been shown to be superior to PCI (including the use of drug-eluting stents) in preventing death, myocardial infarction, and repeat revascularization in patients with diabetes mellitus and multivessel IHD.

Indications for CABG usually are based on the severity of symptoms, coronary anatomy, and ventricular function. The ideal candidate is male, <80 years of age, has no other complicating disease, and has troublesome or disabling angina that is not adequately controlled by medical therapy or does not tolerate medical therapy. Great symptomatic benefit can be anticipated if a patient wishes to lead a more active life and has severe stenoses of two or three epicardial coronary arteries with objective evidence of myocardial ischemia as a cause of the chest discomfort. Congestive heart failure and/or LV dysfunction, advanced age (>80 years), reoperation, urgent need for surgery, and the presence of diabetes mellitus are all associated with a higher perioperative mortality rate.

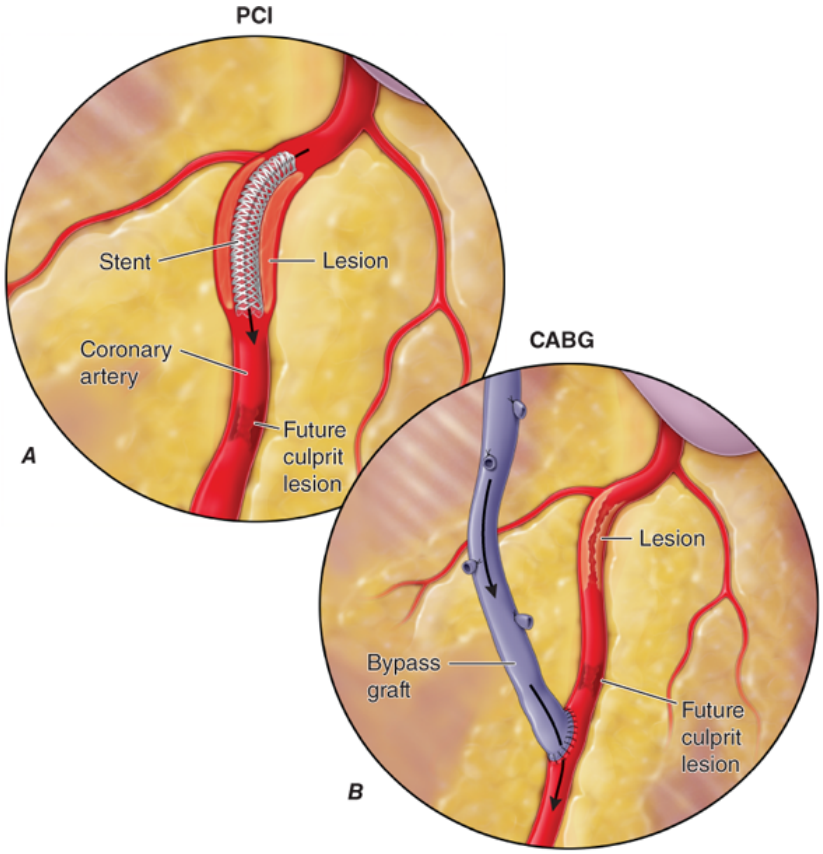
LV dysfunction can be due to noncontractile or hypocontractile segments that are viable but are chronically ischemic (hibernating myocardium). As a consequence of chronic reduction in myocardial blood flow, these segments downregulate their contractile function. They can be detected by using radionuclide scans of myocardial perfusion and metabolism, PET, cardiac MRI, or delayed scanning with thallium-201 or by improvement of regional functional impairment provoked by low-dose dobutamine. In such patients, revascularization improves myocardial blood flow, can return function, and can improve survival.

The Choice Between PCI and CABG

All the clinical characteristics of each individual patient must be used to decide on the method of revascularization (e.g., LV function, diabetes, lesion complexity). A number of randomized clinical trials have compared PCI and CABG in patients with multivessel CAD who were suitable technically for both procedures. The redevelopment of angina requiring repeat coronary angiography and repeat revascularization is higher with PCI. This is a result of restenosis in the stented segment (a problem largely solved with drug-eluting stents) and the development of new stenoses in unstented portions of the coronary vasculature. It has been argued that PCI with stenting focuses on culprit lesions, whereas a bypass graft to the target vessel also provides a conduit around future culprit lesions proximal to the anastomosis of the graft to the native vessel **(Fig. 267-5)**. By contrast, stroke rates are lower with PCI.

FIGURE 267-5

Difference in the approach to the lesion with percutaneous coronary intervention (PCI) and coronary artery bypass grafting (CABG). PCI is targeted at the “culprit” lesion or lesions, whereas CABG is directed at the epicardial vessel, including the culprit lesion or lesions and future culprits, proximal to the insertion of the vein graft, a difference that may account for the superiority of CABG, at least in the intermediate term, in patients with multivessel disease. *(Reproduced from BJ Gersh, RL Frye: N Engl J Med 352:2235, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Based on available evidence, it is now recommended that patients with an unacceptable level of angina despite optimal medical management be considered for coronary revascularization. Patients with single- or two-vessel disease with normal LV function and anatomically suitable lesions ordinarily are advised to undergo PCI **(Chap. 270)**. Patients with three-vessel disease (or two-vessel disease that includes the proximal left descending coronary artery) and impaired global LV function (LV ejection fraction <50%) or diabetes mellitus and those with left main CAD or other lesions unsuitable for catheter-based procedures should be considered for CABG as the initial method of revascularization. In light of the complexity of the decision-making, it is desirable to have a multidisciplinary team, including a cardiologist and a cardiac surgeon in conjunction with the patient’s primary care physician, provide input along with ascertaining the patient’s preferences before committing to a particular revascularization option.

UNCONVENTIONAL TREATMENTS FOR IHD

On occasion clinicians will encounter a patient who has persistent disabling angina despite maximally tolerated medical therapy and for whom revascularization is not an option (e.g., small diffusely diseased vessels not amenable to stent implantation or acceptable targets for bypass grafting). In such situations, unconventional treatments should be considered.

Enhanced external counterpulsation utilizes pneumatic cuffs on the lower extremities to provide diastolic augmentation and systolic unloading of blood pressure to decrease cardiac work and oxygen consumption while enhancing coronary blood flow. Clinical trials have shown that regular application improves angina, exercise capacity, and regional myocardial perfusion. Experimental approaches, such as stem cell therapies and cardiac repair with small non-coding RNA molecules (miRNA), are also under active study.

ASYMPTOMATIC (SILENT) ISCHEMIA

Obstructive CAD, acute myocardial infarction, and transient myocardial ischemia are frequently asymptomatic. During continuous ambulatory ECG monitoring, the majority of ambulatory patients with typical chronic stable angina are found to have objective evidence of myocardial ischemia (ST-segment depression) during episodes of chest discomfort while they are active outside the hospital. In addition, many of these patients also have more frequent episodes of asymptomatic ischemia. Frequent episodes of ischemia (symptomatic and asymptomatic) during daily life appear to be associated with an increased likelihood of adverse coronary events (death and myocardial infarction). In addition, patients with asymptomatic ischemia after a myocardial infarction are at greater risk for a second coronary event. The widespread use of exercise ECG during routine examinations has also identified some of these previously unrecognized patients with asymptomatic CAD. Longitudinal studies have demonstrated an increased incidence of coronary events in asymptomatic patients with positive exercise tests.

TREATMENT

TREATMENT Asymptomatic Ischemia

The management of patients with asymptomatic ischemia must be individualized. When coronary disease has been confirmed, the aggressive treatment of hypertension and dyslipidemia is essential and will decrease the risk of infarction and death. In addition, the physician should consider the following: (1) the degree of positivity of the stress test, particularly the stage of exercise at which ECG signs of ischemia appear; the magnitude and number of the ischemic zones of myocardium on imaging; and the change in LV ejection fraction that occurs on radionuclide ventriculography or echocardiography during ischemia and/or during exercise; (2) the ECG leads showing a positive response, with changes in the anterior precordial leads indicating a less favorable prognosis than changes in the inferior leads; and (3) the patient's age, occupation, and general medical condition.

Most would agree that an asymptomatic 45-year-old commercial airline pilot with significant (0.4-mV) ST-segment depression in leads V_1 to V_4 during mild exercise should undergo coronary arteriography, whereas an asymptomatic, sedentary 85-year-old retiree with 0.1-mV ST-segment depression in leads II and III during maximal activity need not. However, there is no consensus about the most appropriate approach in the large majority of patients for whom the situation is less extreme. Asymptomatic patients with silent ischemia, three-vessel CAD, and impaired LV function may be considered appropriate candidates for CABG.

The treatment of risk factors, particularly lipid lowering and blood pressure control as described above, and the use of aspirin, statins, and beta blockers after infarction have been shown to reduce events and improve outcomes in asymptomatic as well as symptomatic patients with ischemia and proven CAD. Although the incidence of asymptomatic ischemia can be reduced by treatment with beta blockers, calcium channel blockers, and long-acting nitrates, it is not clear whether this is necessary or desirable in patients who have not had a myocardial infarction.

FURTHER READING

De Bruyne B et al: Microvascular (dys)function and clinical outcome in stable coronary disease. J Am Coll Cardiol 67:1170, 2016. [PubMed: 26965537]
Fihn SD et al: 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease: A report of the American College of Cardiology Foundation/American Heart Association task force on practice guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. Circulation 126:e354, 2012. [PubMed: 23166211]
Levine GN et al: 2016 ACC/AHA guideline focused update on duration of dual antiplatelet therapy in patients with coronary artery disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines: An Update of the 2011 ACCF/AHA/SCAI Guideline for Percutaneous Coronary Intervention, 2011 ACCF/AHA Guideline for Coronary Artery Bypass Graft Surgery, 2012 ACC/AHA/ACP/AATS/PCNA/SCAI/STS Guideline for the Diagnosis and Management of Patients with Stable Ischemic Heart Disease, 2013 ACCF/AHA Guideline for the Management of ST-Elevation Myocardial Infarction, 2014 AHA/ACC Guideline for the Management of Patients with Non-ST-Elevation Acute Coronary Syndromes, and 2014 ACC/AHA Guideline on Perioperative Cardiovascular Evaluation and Management of Patients Undergoing Noncardiac Surgery. Circulation 134:e123, 2016. [PubMed: 27026020]
Mody P et al: Antianginal agents for the management of stable ischemic heart disease: A review. Cardiol Rev 24:177, 2016. [PubMed: 26274534]
Mozaffarian D et al: Heart disease and stroke statistics—2016 update: A report from the American Heart Association. Circulation 2016;133:e38, 2016. [PubMed: 26673558]
Omland T, White HD: State of the art: Blood biomarkers for risk stratification in patients with stable ischemic heart disease. Clin Chem 63:165, 2017.
Singh M, Arora R: Newer therapies for management of stable ischemic heart disease with focus on refractory angina. Am J Ther 23:e1842, 2016. [PubMed: 25590765]

Chapter 268: Non-ST-Segment Elevation Acute Coronary Syndrome (Non-ST-Segment Elevation Myocardial Infarction and Unstable Angina)

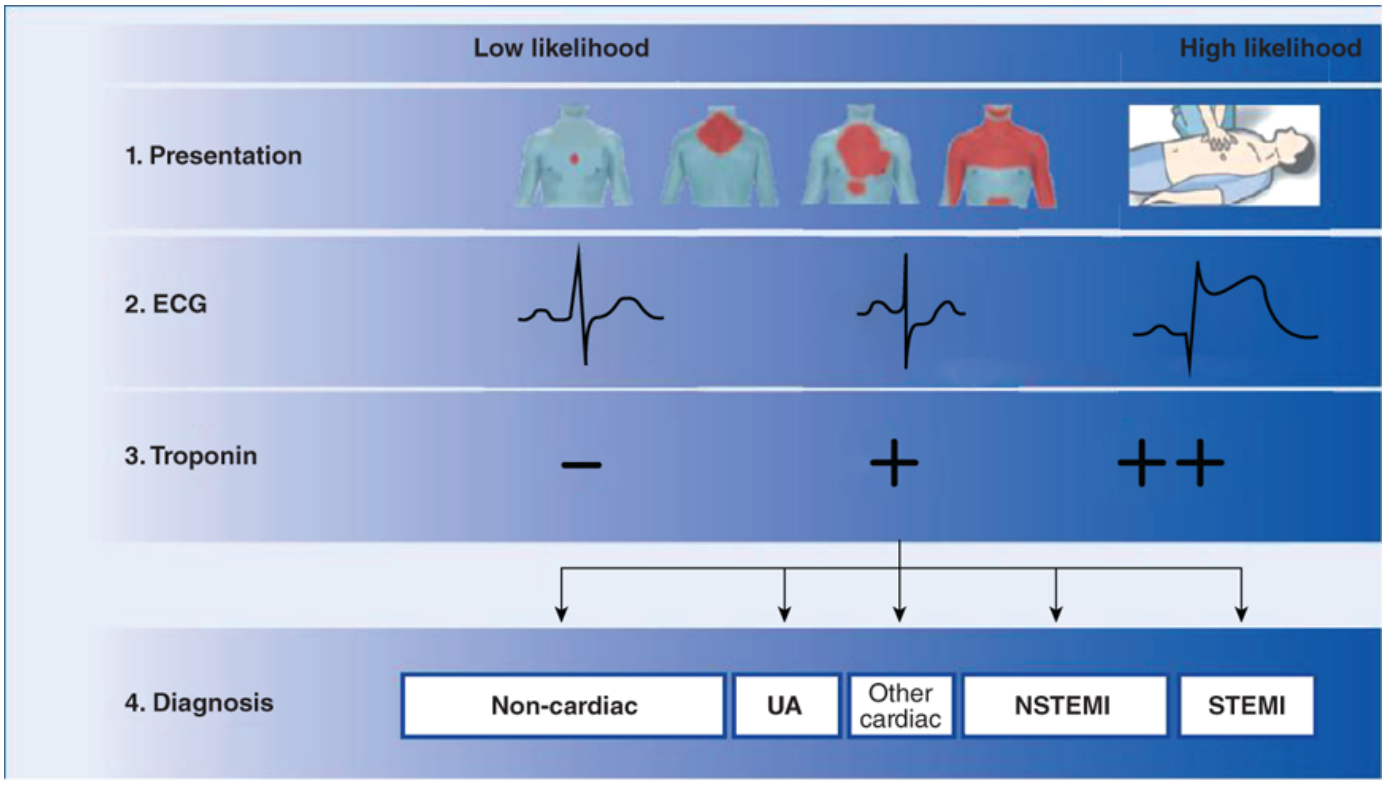
Robert P. Giugliano; Christopher P. Cannon; Eugene Braunwald

INTRODUCTION

Patients with acute coronary syndrome (ACS) commonly are classified into two groups to facilitate evaluation and management, namely patients with acute myocardial infarction with ST-segment elevation (STEMI) on their presenting electrocardiogram (ECG) **(Chap. 269)** and those with non-ST-segment elevation acute coronary syndrome (NSTE-ACS). The latter include patients with non-ST-segment elevation myocardial infarction (NSTEMI), who, by definition, have evidence of myocyte necrosis, and those with unstable angina (UA), who do not **(Fig. 268-1)**.

FIGURE 268-1

Assessment of patients with suspected acute coronary syndromes. The initial assessment is based on the integration of low-likelihood and/or high-likelihood features derived from clinical presentation (i.e., symptoms, vital signs), 12-lead electrocardiogram and cardiac troponin. The proportion of the final diagnoses derived from the integration of these parameters is visualized by the size of the respective boxes. (*From Roffi M et al: 2015 European Society of Cardiology Guidelines for the management of acute coronary syndromes. Eur Heart J 37:267, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The relative incidence of NSTEMI is rising due to the increasing burden of diabetes and chronic kidney disease in an aging population, while STEMI is declining due to greater use of aspirin, statins, and less smoking. Among patients with NSTE-ACS, the

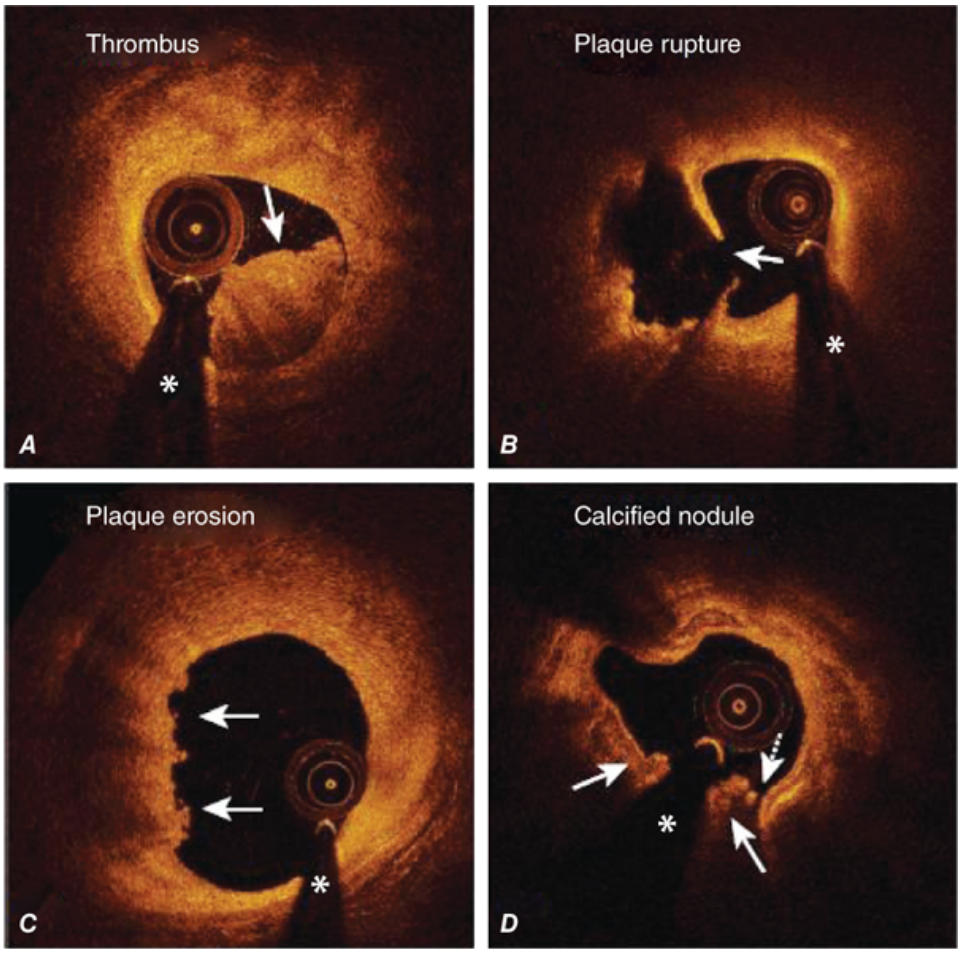
proportion with NSTEMI is rising while that with UA is falling because of the wider use of troponin assays with higher sensitivity to detect myocyte necrosis, thereby reclassifying UA as NSTEMI.

PATHOPHYSIOLOGY

NSTE-ACS is caused by an imbalance between myocardial oxygen supply and demand resulting from one or more of the following four processes that lead to thrombus formation: (1) disruption of an unstable coronary plaque due to plaque rupture, erosion, or a calcified protruding nodule that leads to intracoronary thrombus formation **(Fig. 268-2)** and an inflammatory response; (2) coronary arterial vasoconstriction; (3) gradual intraluminal narrowing; and (4) increased myocardial oxygen demand produced by conditions such as fever, tachycardia, and thyrotoxicosis in the presence of fixed epicardial coronary obstruction. While plaque rupture remains the most common etiology of coronary thrombosis, erosion of an intracoronary plaque is increasing in frequency, perhaps related to the above mentioned shifts in the underlying risk factors for ACS.

FIGURE 268-2

Intracoronary thrombosis and the three most common plaque morphologies resulting in acute coronary syndrome as visualized by optical coherence tomography. ***A.*** Thrombus (*arrow*) is identified as a protruding mass attached to the arterial wall. ***B.*** Plaque rupture is identified as lipid plaque with fibrous cap discontinuity (*arrow*) and cavity formation inside the plaque. ***C.*** Plaque erosion is confirmed by the presence of attached thrombus (*arrows*) overlying an intact and visualized plaque. ***D.*** Calcified nodule appears on optical coherence tomography as a site with fibrous cap disruption (*dotted arrow*) and underlying plaque characterized by protruding calcification, superficial calcium, and significant calcium adjacent to the lesion (*arrows*). The asterisks denote guidewire shadow artifact. *(Modified from H Jia et al: J Am Coll Cardiol 62:1748, 2013 and I Jang, D Ong: Optical coherence tomography and other emerging diagnostic procedures for vulnerable plaque, in D Morrow (ed): Myocardial Infarction: A Companion to Braunwald's Heart Disease. Philadelphia, Elsevier Health Sciences, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Among patients with NSTE-ACS studied at angiography, ~10% have stenosis of the left main coronary artery, 35% have three-vessel CAD, 20% have two-vessel disease, 20% have single-vessel disease, and 15% have no apparent critical epicardial coronary artery stenosis; some of the latter may have obstruction of the coronary microcirculation and/or spasm of the epicardial vessels. The so-called "vulnerable plaques" responsible for ischemia may show an eccentric stenosis with scalloped or

overhanging edges and a narrow neck on coronary angiography. Vulnerable plaques are composed of a lipid-rich core with a thin fibrous cap. Patients with NSTE-ACS frequently have multiple such plaques that are at risk of disruption.

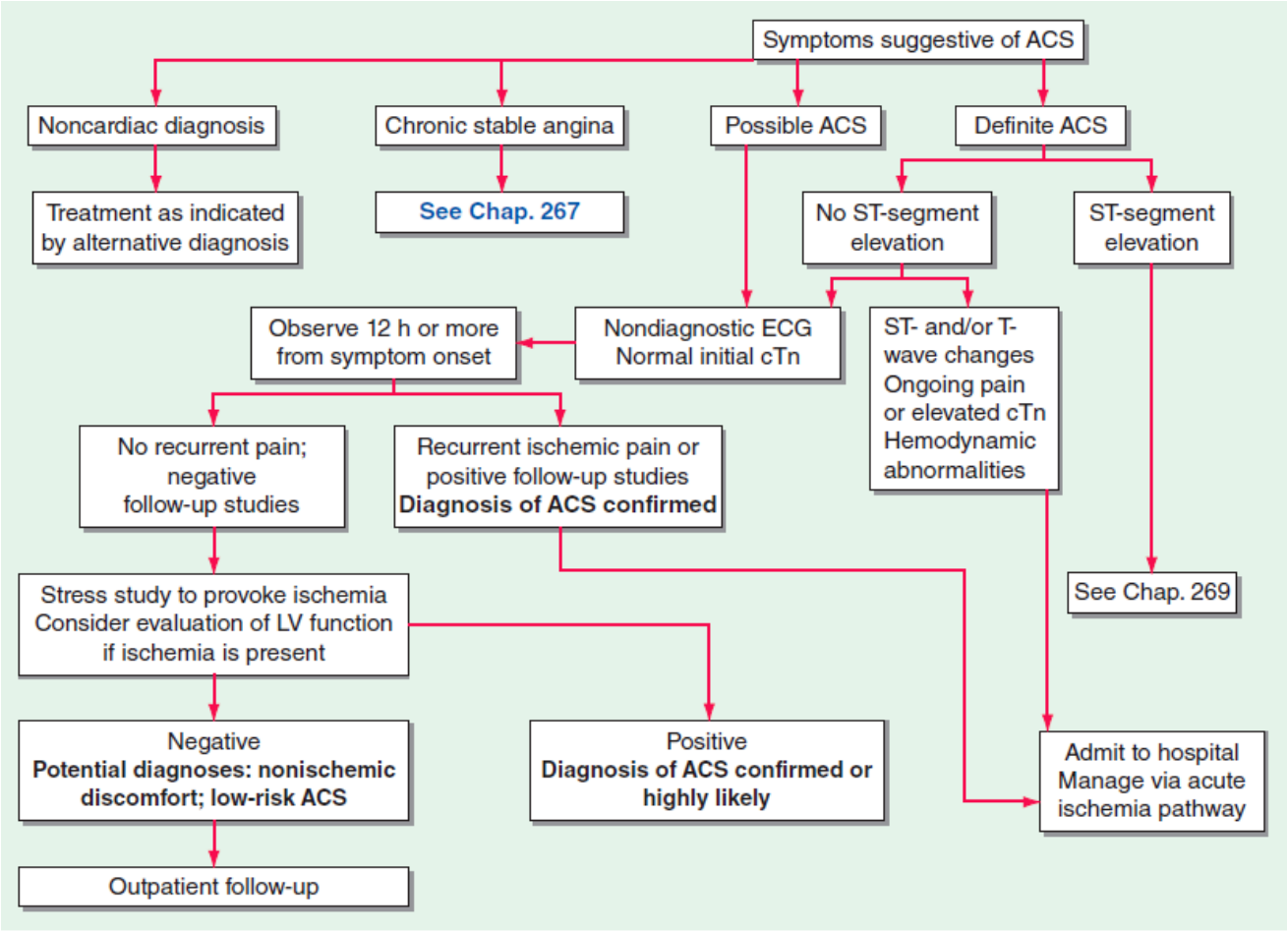
CLINICAL PRESENTATION

Diagnosis

The diagnosis of NSTE-ACS is based largely on the clinical presentation **(Fig. 268-3)**.

FIGURE 268-3

Algorithm for evaluation and management of patients with suspected acute coronary syndrome (ACS). Follow-up studies refer to ST deviation and elevation of troponin levels. cTn, cardiac troponin; ECG, electrocardiogram; LV, left ventricular. *(Modified from J Anderson et al: J Am Coll Cardiol 61:e179, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

History and Physical Examination

Typically, chest discomfort is severe and has at least one of three features: (1) occurrence at rest (or with minimal exertion), lasting >10 min; (2) of relatively recent onset (i.e., within the prior 2 weeks); and/or (3) a crescendo pattern, i.e., distinctly more severe, prolonged, or frequent than previous episodes. The diagnosis of NSTEMI is established if a patient with any of these features (without electrocardiographic ST segment elevations) develops evidence of myocardial necrosis, as reflected in abnormally elevated levels of biomarkers (see below). The chest discomfort is typically located in the substernal region and radiates to the left arm, left shoulder, and/or superiorly to the neck and jaw. Anginal equivalents such as dyspnea, epigastric discomfort, nausea, or weakness may occur instead of chest discomfort. They appear to be more frequent in women, the elderly, and patients with diabetes mellitus. The physical examination resembles that in patients with stable angina **(Chap. 267)** and may be unremarkable. However, if the patient has a large area of myocardial ischemia or a large NSTEMI, the physical findings can include diaphoresis; pale, cool skin; sinus tachycardia; a third and/or fourth heart sound; basilar rales; and, sometimes, hypotension.

Electrocardiogram

New ST-segment depression occurs in about one-third of patients with NSTE-ACS. It may be transient but may persist for several days following NSTEMI. T-wave changes are more common but are less specific signs of ischemia, unless they are new and deep T-wave inversions (≥0.3 mV).

Cardiac Biomarkers

Patients with NSTEMI have elevated biomarkers of necrosis, such as cardiac troponin (cTn) I or T, which are specific, sensitive, and the preferred markers of myocardial necrosis. The MB isoform of creatine kinase (CK-MB) is a less sensitive alternative. Elevated levels of any of these markers distinguish patients with NSTEMI from those with UA. There is a characteristic temporal rise and fall peaking 12–24 h post onset of symptoms of the plasma concentration of these markers and a direct relationship between the degree of elevation and mortality. However, in patients *without* a clear clinical history of myocardial ischemia, minor cTn elevations have been reported and can be caused by heart failure, myocarditis, or pulmonary embolism, or with high-sensitivity assays (hs cTn) may be observed in ostensibly normal subjects. Thus, in patients with an *unclear* history, small elevations of cTn, especially if they are persistent, may not be diagnostic of an ACS. In such cases, both cardiac and non-cardiac causes of an elevated cTn should be considered **(Table 268-1)**.

TABLE 268-1 Causes of Elevated Cardiac Troponin Reflecting Direct Myocardial Damage Other Than Spontaneous Myocardial Infarction (Type 1)

Cardiac	Non-Cardiac or Systemic
Tachyarrhythmias	Pulmonary embolism/pulmonary hypertension
Congestive heart failure	Trauma (e.g., electrical shock, burns, blunt chest wall)
Hypertensive emergencies	Hypo or hyperthyroidism
Infection/inflammation (e.g., myocarditis, pericarditis)	Toxicity (e.g., anthracyclines, snake venom)
Stress cardiomyopathy (Tako-Tsubo cardiomyopathy)	Renal failure
Structural heart disease (e.g., aortic stenosis)	Sepsis, shock
Aortic dissection	Stroke or other acute neurologic event
Coronary spasm	Extreme endurance efforts (e.g., ultra-marathon)
Cardiac procedures (endomyocardial biopsy, ablation, CABG, PCI)	Rhabdomyolysis
Infiltrative diseases (e.g., amyloidosis, hemochromatosis, malignancy)	

Source: Data from LK Newby et al: J Am Coll Cardiol 60:2427, 2012 and M Roffi: Eur Heart J 37:267, 2016.

DIAGNOSTIC EVALUATION

In addition to the clinical examination, three major noninvasive tools are used in the evaluation of NSTEMI-ACS: the ECG, cardiac biomarkers, and stress testing. In equivocal cases, coronary computed tomographic angiography (CCTA) may be useful to improve the accuracy and speed of the diagnostic evaluation. The goals are to: (1) recognize or exclude myocardial infarction (MI) using cardiac biomarkers, preferably cTn; (2) detect rest ischemia (using serial or continuous ECGs); and (3) detect significant coronary obstruction at rest with CCTA and/or myocardial ischemia using stress testing **(Chap. 236)**.

Patients with a low likelihood of ischemia are usually managed in an emergency department or a dedicated “chest pain unit” following a critical pathway. Evaluation of such patients includes clinical monitoring for recurrent ischemic discomfort and continuous monitoring of ECGs and cardiac markers, typically obtained at baseline and at 4–6 h and 12 h after presentation. If new elevations in cardiac markers or ST-T-wave changes on the ECG are noted, the patient should be admitted to the hospital. Patients who remain pain-free with negative markers may proceed to stress testing to determine the presence of ischemia or CCTA to detect coronary luminal obstruction (Fig. 268-3).

The hs cTn assays permit a more rapid (3-h and even 1-h) rule-out MI determination and have been adopted by the 2015 European Guidelines for the management of NSTE-ACS.

RISK STRATIFICATION

Patients with documented NSTE-ACS exhibit a wide spectrum of early (30 days) risk of death, ranging from 1 to 10%, and a recurrent ACS rate of 5–15% during the first year. Assessment of risk can be accomplished by clinical risk scoring systems such as that developed from the Thrombolysis in Myocardial Infarction (TIMI) Trials, which includes seven independent risk factors (age ≥ 65 years, 3 or more of the traditional risk factors for coronary heart disease, known history of coronary artery disease or coronary stenosis of at least 50%, daily aspirin use in the prior week, more than one anginal episode in the past 24 h, ST segment deviation of at least 0.5 mm, and an elevated cardiac specific biomarker above the upper limit of normal). Additional risk factors include diabetes mellitus, left ventricular dysfunction, renal dysfunction, and elevated levels of B-type natriuretic peptides. Multibiomarker strategies are now gaining favor, both to define more fully the pathophysiologic mechanisms underlying a given patient’s presentation and to stratify the patient’s risk further. Patients with ACS without elevated levels of cTn (infrequently

encountered with the new sensitive troponin assays) are considered to have UA and have a more favorable prognosis than those with cTn elevations (NSTEMI).

Early risk assessment is useful in identifying patients who would derive the greatest benefit from an early invasive strategy (see below). For example, in the TACTICS-TIMI 18 Trial, an early invasive strategy conferred a 40% reduction in recurrent cardiac events in patients with an elevated cTn level, whereas no benefit was observed in those without detectable troponin.

TREATMENT

TREATMENT Non-ST-Segment Elevation Acute Coronary Syndrome (Non-ST-Segment Elevation Myocardial Infarction and Unstable Angina) MEDICAL TREATMENT

Patients should be placed at bed rest with continuous ECG monitoring for ST-segment deviation and cardiac arrhythmias, preferably on a specialized cardiac unit. Ambulation is permitted if the patient shows no recurrence of ischemia (symptoms or ECG changes) and does not develop an elevation of a biomarker of necrosis for 12–24 h.

Medical therapy consists of an acute phase focused on the clinical symptoms and stabilization of the culprit lesion(s) and a longer-term phase that involves therapies directed at the prevention of disease progression and future plaque rupture/erosion.

ANTI-ISCHEMIC TREATMENT

To provide relief and prevention of recurrence of ischemic discomfort, initial treatment should include bed rest, nitrates, beta adrenergic blockers, and inhaled oxygen in patients with arterial O_2 saturation (<90%) and/or in those with heart failure and rales **(TABLE 268-2)**.

Nitrates

These should first be given sublingually or by buccal spray (0.3–0.6 mg) if the patient is experiencing ischemic discomfort. If symptoms persists after three doses given 5 min apart, intravenous nitroglycerin (5–10 μg/min using nonabsorbing tubing) is recommended. The rate of the infusion may be increased by 10 μg/min every 3–5 min until symptoms are relieved, systolic arterial pressure falls to <90 mmHg, or the dose reaches 200 μg/min. Topical or oral nitrates **(Chap. 267)** can be used when the pain has resolved, or they may replace intravenous nitroglycerin when the patient has been symptom-free for 12–24 h. The only absolute contraindications to the use of nitrates are hypotension or the recent use of a phosphodiesterase type 5 (PDE-5) inhibitor, sildenafil or vardenafil (within 24 h), or tadalafil (within 48 h).

Beta-Adrenergic Blockers and Other Agents

Beta blockers are the other mainstay of anti-ischemic treatment. They may be started by the intravenous route in patients with severe ischemia, but should be avoided in the presence of acute or severe heart failure, low cardiac output, hypotension, or contraindications to beta-blocker therapy (e.g., high-degree atrioventricular block, active bronchospasm). Ordinarily, oral beta blockade targeted to a heart rate of 50–60 beats/min is recommended. Heart rate–slowing calcium channel blockers, e.g., verapamil or diltiazem, are recommended for patients who have persistent symptoms or ECG signs of ischemia after treatment with full-dose nitrates and beta blockers and in patients with contraindications to either class of these agents. Additional medical therapy includes angiotensin-converting enzyme (ACE) inhibitors or angiotensin receptor blockers. Early administration of intensive HMG-CoA reductase inhibitors (statins), such as atorvastatin 80 mg/d, prior to percutaneous coronary intervention (PCI), and continued thereafter, has been shown to reduce peri-procedural MI and recurrences of ACS. In patients who do not have an adequate response to maximally tolerated statin (i.e., <50% decrease in LDL-C from untreated baseline or LDL-C on treatment >70 mg/dL), addition of ezetimibe 10 mg daily to reduce further the LDL-C has been shown to reduce future cardiovascular events.

ANTITHROMBOTIC THERAPY

Antithrombotic therapy consisting of antiplatelet and anticoagulant drugs represent the second major cornerstone of treatment **(FIG. 268-4 AND TABLE 268-3)**.

Antiplatelet Drugs

Initial treatment should begin with the cyclooxygenase inhibitor aspirin with a dose of at least 162 mg of a rapidly acting preparation (oral non-enteric coated or intravenous) (See **Chap. 114**). Lower doses (75–100 mg/d) are recommended thereafter, since they maintain efficacy while causing less bleeding. Contraindications are severe active bleeding or aspirin allergy.

In the absence of a high risk for bleeding, patients with NSTE-ACS, irrespective of whether an invasive or conservative strategy (see below) is selected, also should receive a platelet $P2Y_{12}$ receptor blocker to inhibit platelet activation. There are now four oral and one intravenous $P2Y_{12}$ inhibitors to choose from (although the first in class, ticlopidine, is rarely used due to poor tolerability); advantages for each of the others are noted below.

The thienopyridine clopidogrel is an inactive prodrug that is converted into an active metabolite that causes irreversible blockade of the platelet $P2Y_{12}$ receptor. The loading dose of clopidogrel is 600 or 300 mg while the maintenance dose is 75 mg daily. When

clopidogrel is added to aspirin, so-called dual antiplatelet therapy (DAPT), has been shown to confer a 20% relative reduction in cardiovascular death, MI, or stroke, compared to aspirin alone, but to be associated with a moderate (absolute 1%) increase in major bleeding.

Two newer $P2Y_{12}$ inhibitors (prasugrel, ticagrelor) have been found to be superior to clopidogrel in preventing recurrent cardiac ischemic events in randomized double-blind studies although both increase bleeding. Prasugrel, also a thienopyridine, achieves a more rapid onset and higher level of platelet inhibition than clopidogrel. It has been approved for ACS patients following angiography when PCI is planned. It should be administered at a loading dose of 60 mg followed by 10 mg/d. Compared to clopidogrel, prasugrel was shown to significantly reduce by 19% the combined risk of cardiovascular death, MI, or stroke, and reduced stent thrombosis by 50%. Prasugrel is contraindicated in patients with prior stroke or transient ischemic attack or at high risk for bleeding. It has not been found to be effective in patients treated by a conservative strategy prior to coronary angiography (see below).

Ticagrelor is a novel, potent, *reversible* platelet $P2Y_{12}$ inhibitor that was shown to reduce the risk of cardiovascular death, total mortality or MI compared to clopidogrel across a broad spectrum patients with ACS. After a loading dose of 180 mg, 90 mg bid is administered as maintenance. Unlike prasugrel, ticagrelor demonstrated benefit whether patients were managed conservatively or with an early invasive strategy. Some patients may develop dyspnea early after administration of ticagrelor, although the symptoms are most often transient and infrequently serious, and are not associated with clinical exacerbations of chronic obstructive pulmonary disease or congestive heart failure.

DAPT should continue for at least 1 year in patients with NSTE-ACS, especially those with a drug-eluting stent, to prevent stent thrombosis. Up to one-third of patients have an inadequate response to clopidogrel, and a substantial proportion of these cases are related to a genetic variant of the cytochrome P450 system involving the 2C19 gene that leads to reduced conversion of clopidogrel into its active metabolite. Thus, alternate $P2Y_{12}$ blockers should be considered in patients with NSTE-ACS who develop a coronary event while receiving clopidogrel and aspirin, who are hyporesponsive to clopidogrel, or are at high risk for ischemic complications. Clinicians should select the antiplatelet regimen that provides the best balance of efficacy and safety based on the individual patient characteristics and clinical scenario.

More recently, an intravenous, direct and rapidly acting, $P2Y_{12}$ inhibitor, cangrelor, was evaluated in three large outcome studies in >25,000 patients undergoing PCI across a broad spectrum of clinical presentations (stable angina, UA, NSTEMI, STEMI). Among the 14,282 patients who underwent PCI following a NSTE-ACS, cangrelor reduced the risk of the primary composite outcome of death, MI, ischemia-driven revascularization, and stent thrombosis at 48 h by 18% relative to control. There was an excess of 3 per 1000 major bleeding events with cangrelor. This drug is approved as an adjunct to PCI for reducing the risk of periprocedural MI, repeat coronary revascularization, and stent thrombosis in patients who have not been treated with a $P2Y_{12}$ platelet inhibitor and are not being given a GP IIb/III inhibitor.

In the 1990s and early 2000s several trials had shown the benefit of intravenous glycoprotein IIb/IIIa inhibitors in patients with NSTE-ACS, with the majority of studies performed without concomitant $P2Y_{12}$ inhibition. The benefits, however, were modest (i.e., an ~1% absolute reduction in death or MI at 30 day) and counterbalanced by a 1% absolute increase in the rate of major bleeding. Two recent studies failed to show a benefit of routine early initiation of a drug in this class compared with their use only in patients who undergo PCI. The addition of these agents to aspirin and a $P2Y_{12}$ inhibitor (i.e., triple antiplatelet therapy) should be reserved for unstable patients undergoing PCI. These include patients with recurrent rest pain, elevated cTn, and ECG changes, as well as those who have a coronary thrombus evident on angiography.

Anticoagulants

Four options are available for anticoagulant therapy to be added to antiplatelet agents: (1) unfractionated heparin (UFH), long the mainstay of therapy; (2) the low-molecular-weight heparin (LMWH), enoxaparin, which has been shown to be superior to UFH in reducing recurrent cardiac events, especially in patients managed by a conservative strategy (See **Chap. 114**). However, it is accompanied by a slight increase in bleeding compared to UFH; (3) bivalirudin, a direct thrombin inhibitor that is similar in efficacy to either UFH or LMWH but causes less bleeding and is used just prior to and/or during PCI; and (4) the indirect factor Xa inhibitor, fondaparinux, which is equivalent in efficacy to enoxaparin but has a lower risk of major bleeding. While UFH and enoxaparin have been widely studied in patients managed either with an early conservative or invasive strategy, the role of bivalirudin in conservatively managed patients is less clear, while fondaparinux requires supplemental UFH or bivalirudin during PCI to prevent procedure-related thrombosis.

Excessive bleeding is the most important adverse effect of all antithrombotic agents, including both antiplatelet agents and anticoagulants. Therefore, attention must be directed to the doses of antithrombotic agents, accounting for body weight, creatinine clearance, and a previous history of excessive bleeding, as a means of reducing the risk of bleeding. Patients who have experienced a stroke are at higher risk of intracranial bleeding with potent antiplatelet agents and combinations of antithrombotic drugs.

INVASIVE VERSUS CONSERVATIVE STRATEGY

In an invasive strategy, following initiation of anti-ischemic and antithrombotic agents, coronary arteriography is carried out within ~48 h of presentation, followed by coronary revascularization (PCI or coronary artery bypass grafting), depending on the coronary anatomy. Multiple clinical trials have demonstrated the benefit of this strategy in high-risk patients (i.e., patients with multiple clinical risk factors, ST-segment deviation, and/or positive biomarkers) **(Table 268-4)**. In patients at low risk, the outcomes from

an invasive strategy are similar to those obtained from a conservative strategy. The latter consists of anti-ischemic and antithrombotic therapy followed by a “selective invasive approach,” in which the patient is closely observed and coronary arteriography is carried out if rest pain or ST-segment changes recur, a biomarker of necrosis becomes positive, or there is evidence of severe ischemia on a stress test.

TABLE 268-2 Drugs Commonly Used in Intensive Medical Management of Patients with Unstable Angina and Non-ST-Segment Elevation Myocardial Infarction

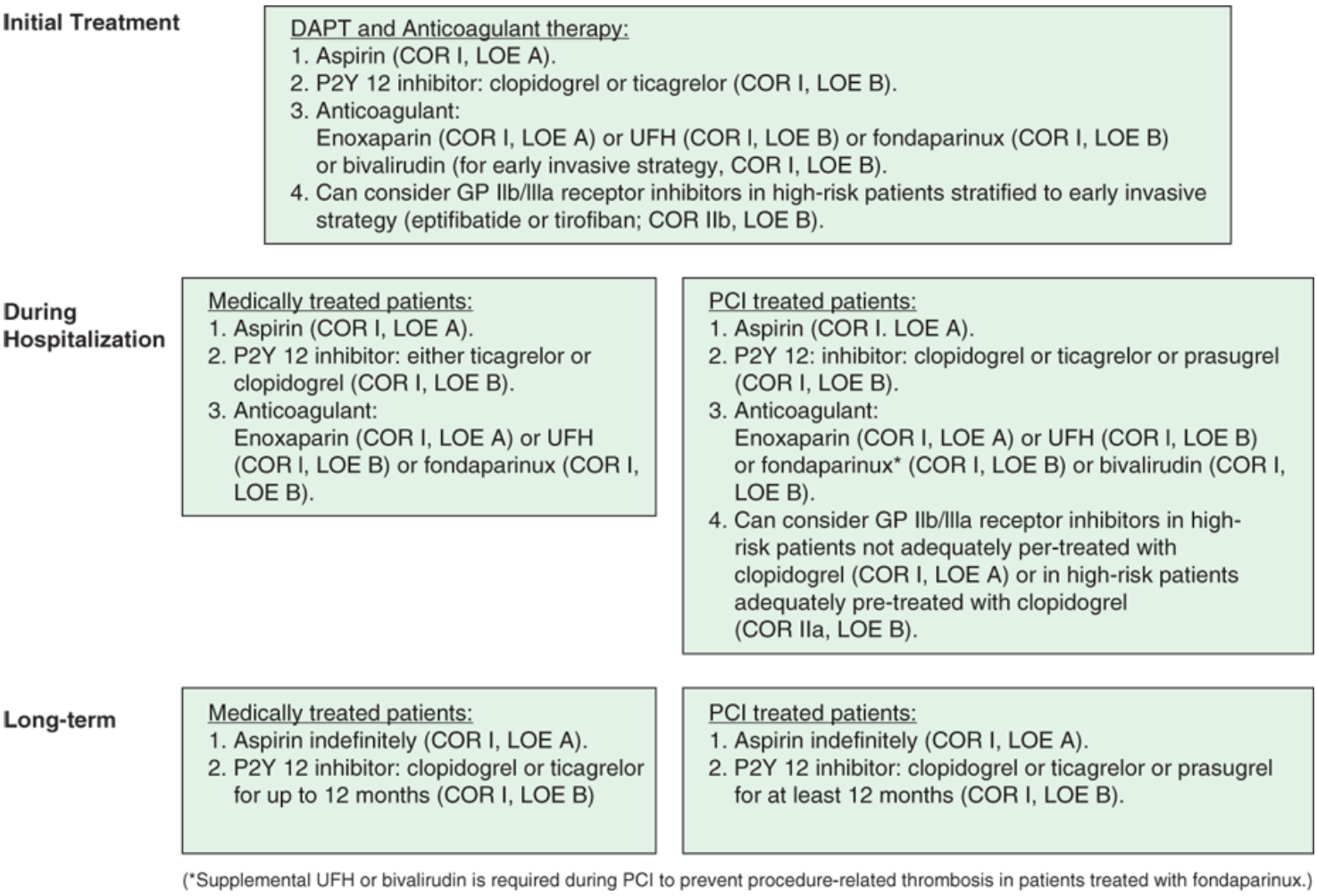
Drug Category	Clinical Condition	When to Avoid[a]	Dosage
Nitrates	Patients with ACS who have chest discomfort or an anginal equivalent	Hypotension Right ventricular infarction Severe aortic stenosis Patient receiving a PDE-5 inhibitor PR interval (ECG) >0.24 s 2° or 3° atrioventricular block Heart rate <50 beats/min	Initially administer via sublingual or buccal route, and, if symptoms persist, intravenously. Topical or oral nitrates are acceptable alternatives for patients without ongoing or refractory symptoms 5–10 μg/min by continuous infusion titrated up to 75–100 μg/min until relief of symptoms or limiting side effects (headache or hypotension with a systolic blood pressure <90 mmHg or >30% below starting mean arterial pressure levels if significant hypertension is present)
Beta blockers[b]	All patients with ACS	Systolic pressure <90 mmHg Shock Left ventricular failure Severe reactive airway disease Pulmonary edema	Metoprolol 25–50 mg by mouth every 6 h If needed, and no heart failure, 5-mg increments by slow (over 1–2 min) IV administration
Calcium channel blockers	Patients whose symptoms are not relieved by adequate doses of nitrates and beta blockers, or in patients unable to tolerate adequate doses of one or both of these agents, or in patients with variant angina	Evidence of left ventricular dysfunction (for diltiazem or verapamil) Hypotension	Dependent on specific agent
Morphine sulfate	Patients whose symptoms are not relieved after three serial sublingual nitroglycerin tablets or whose symptoms recur with adequate anti-ischemic therapy	Respiratory depression Confusion Obtundation	2–5 mg IV dose May be repeated every 5–30 min as needed to relieve symptoms and maintain patient comfort

[a]Allergy or prior intolerance is a contraindication for all categories of drugs listed in this chart. [b]Choice of the specific agent is not as important as ensuring that appropriate candidates receive this therapy.

Source: Modified from J Anderson et al: J Am Coll Cardiol 61:e179, 2013.

FIGURE 268-4

Antiplatelet and anticoagulation treatment summary for NSTE-ACS according to the 2014 American Heart Association/American College of Cardiology Practice Guideline. COR, classes of recommendation; DAPT, dual antiplatelet therapy; GP IIb/IIIa, glycoprotein IIb/IIIa; LOE, levels of evidence; NSTE-ACS, non-ST-segment elevation acute coronary syndrome; PCI, percutaneous coronary intervention; UFH, unfractionated heparin. *(From A Eisen, RP Giugliano: Cardiol Rev 24;170, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 268-3 Clinical Use of Antithrombotic Therapy

Oral Antiplatelet Therapy	
Aspirin	Initial dose of 325 mg nonenteric formulation followed by 75–100 mg/d of an enteric or a nonenteric formulation
Clopidogrel	Loading dose of 300–600 mg followed by 75 mg/d
Prasugrel	Pre-PCI: Loading dose 60 mg followed by 10 mg/d
Ticagrelor	Loading dose of 180 mg followed by 90 mg twice daily
Intravenous Antiplatelet Therapy	
Abciximab	0.25 mg/kg bolus followed by infusion of 0.125 μg/kg per min (maximum 10 μg/min) for 12–24 h
Eptifibatide	180 μg/kg bolus followed 10 min later by second bolus of 180 μg with infusion of 2.0 μg/kg per min for 72–96 h following first bolus
Tirofiban	25 μg/kg per min followed by infusion of 0.15 μg/kg per min for 48–96 h
Cangrelor	30 μg/kg bolus followed immediately by a 4 μg/kg per min infusion
Anticoagulants[a]	
Unfractionated heparin (UFH)	[b]Bolus 70–100 U/kg (maximum 5000 U) IV followed by infusion of 12–15 U/kg per h (initial maximum 1000 U/h) titrated to ACT 250–300 s
Enoxaparin	1 mg/kg SC every 12 h; the first dose may be preceded by a 30-mg IV bolus; renal adjustment to 1 mg/kg once daily if creatine clearance <30 mL/min
Fondaparinux	2.5 mg SC qd
Bivalirudin	Initial IV bolus of 0.75 mg/kg and an infusion of 1.75 mg/kg per h

[a]Other low-molecular-weight heparins have been studied other than enoxaparin; however there are less data to support their use.
[b]If no glycoprotein IIb/IIIa inhibitor planned.

Abbreviations: ACT, activated clotting time for HemoTec; IV, intravenous; SC, subcutaneous.

Source: Modified from J Anderson et al: J Am Coll Cardiol 61:e179, 2013.

TABLE 268-4 Factors Associated with Appropriate Selection of Early Invasive Strategy or Ischemia-Guided Strategy in Patients with NSTE-ACS

Immediate invasive (within 2 h)	Refractory angina
	Signs or symptoms of heart failure or new or worsening mitral regurgitation
	Hemodynamic instability
	Recurrent angina or ischemia at rest or with low-level activities despite intensive medical therapy
	Sustained ventricular tachycardia or ventricular fibrillation
Early invasive (within 24 h)	None of the above, but GRACE[a] risk score >140
	Temporal change in troponin
	New or presumably new ST segment depression
	None of the above but diabetes mellitus
Delayed invasive (within 25–72 h)	Renal insufficiency (eGFR <60 mL/min per 1.73 m^2)
	Reduced left ventricular systolic function (ejection fraction <0.40)
	Early postinfarction angina
	Percutaneous coronary intervention within 6 months prior
	Prior coronary artery bypass graft surgery
Ischemia-guided strategy	GRACE[a] risk score 109–140 or TIMI[b] risk score ≥2
	Low-risk score (e.g., TIMI[b] [0 or 1], GRACE[a] [<109])
	Low-risk, troponin-negative female patients
	Patient or clinician preference in the absence of high-risk features

[a]See CB Granger (Arch Intern Med 163:2345, 2003). [b]See EM Antman (JAMA 284:835, 2000).

Abbreviations: eGFR, estimated glomerular filtration rate; GRACE, Global Registry of Acute Coronary Events; TIMI, Thrombolysis in Myocardial Infarction.

Source: Modified from EA Amsterdam et al: J Am Coll Cardiol 64:e139, 2014.

LONG-TERM MANAGEMENT

The time of hospital discharge is a “teachable moment” for the patient with NSTE-ACS, when the physician can review and optimize the medical regimen. Risk-factor modification is key, and the caregiver should discuss with the patient the importance of smoking cessation, achieving optimal weight, daily exercise, blood-pressure control, following an appropriate diet, control of hyperglycemia (in diabetic patients), and lipid management as recommended for patients with chronic stable angina **(Chap. 267)**.

There is evidence of benefit with long-term therapy with five classes of drugs that are directed at different components of the atherothrombotic process. Beta blockers, lipid lowering therapy (statins at high dose, e.g., atorvastatin 80 mg/d, with ezetimibe if needed to achieve an LDL-C below 70 mg/dL), and ACE inhibitors or angiotensin receptor blockers are recommended. The recommended antiplatelet regimen consists of the combination of low-dose (75–100 mg/d) aspirin and a $P2Y_{12}$ inhibitor (clopidogrel, prasugrel, or ticagrelor) for 1 year, with aspirin continued thereafter. In selected patients at high ischemic risk (e.g., those with prior MI, diabetes mellitus, vein graft stent, congestive heart failure) who are also at low risk of bleeding, continuation of DAPT out to 3 years has been shown to be beneficial. These measures, taken together, reduce the incidence of recurrent ACS.

Registries have shown that women and racial minorities, as well as patients with NSTE-ACS at high risk, including the elderly and patients with diabetes or chronic kidney disease, are less likely to receive evidence-based pharmacologic and interventional therapies with resultant poorer clinical outcomes and quality of life. Special attention should be directed to these groups.

PRINZMETAL’S VARIANT ANGINA

In 1959, Prinzmetal et al. described a syndrome of severe ischemic pain that usually occurs at rest and is associated with transient ST-segment elevation. Prinzmetal’s variant angina (PVA) is caused by focal spasm of an epicardial coronary artery with resultant transmural ischemia and abnormalities in left ventricular function that may lead to acute MI, ventricular tachycardia or fibrillation, and sudden cardiac death. The cause of the spasm is not well defined, but it may be related to hypercontractility of vascular smooth muscle due to adrenergic vasoconstrictors, leukotrienes, or serotonin. For reasons that are not clear, the prevalence of PVA has decreased substantially during the past few decades, although it remains more frequent in Japan than in North America or Western Europe.

Clinical and Angiographic Manifestations

Patients with PVA are generally younger and, with the exception of cigarette smoking, have fewer coronary risk factors than do patients with NSTE-ACS. Cardiac examination is usually unremarkable in the absence of ischemia. However, a minority of patients have a generalized vasospastic disorder associated with migraine and/or Raynaud’s phenomenon. The clinical diagnosis of PVA is made by the detection of transient ST-segment *elevation* with rest pain, although many patients may also exhibit episodes of silent ischemia.

Coronary angiography demonstrates transient coronary spasm as the diagnostic hallmark of PVA. Atherosclerotic plaques in at least one proximal coronary artery occur in about half of patients. Hyperventilation or intracoronary acetylcholine has been used to provoke focal coronary stenosis on angiography or to provoke rest angina with ST-segment elevation to establish the diagnosis.

TREATMENT

TREATMENT Prinzmetal’s Variant Angina

Nitrates and calcium channel blockers are the main therapeutic agents. Aspirin may actually increase the severity of ischemic episodes, possibly as a result of the sensitivity of coronary tone to modest changes in the synthesis of prostacyclin. Statin therapy has been shown to reduce the risk of major adverse events, although the precise mechanism is not established. The response to beta blockers is variable. Coronary revascularization may be helpful in patients who also have discrete, flow-limiting, proximal fixed obstructive lesions. Patients who have had ischemia-associated ventricular fibrillation despite maximal medical therapy should receive an implantable cardioverter-defibrillator.

Prognosis

Many patients with PVA pass through an acute, active phase, with frequent episodes of angina and cardiac events during the first 6 months after presentation. Survival at 5 years is excellent (~90–95%), but as many as 20% of patients experience an MI. Patients with no or mild fixed coronary obstruction experience a low rate of cardiac death or MI compared to patients with associated severe obstructive lesions, although about half of the patients without obstructive CAD still experience frequent angina at rest. Patients with PVA who develop serious arrhythmias during spontaneous episodes of pain are at a higher risk for sudden cardiac death. In most patients who survive an infarction or the initial 3- to 6-month period of frequent episodes, there is a tendency for symptoms and cardiac events to diminish over time.

GLOBAL CONSIDERATIONS

 Ischemic heart disease (IHD), and its most dangerous manifestation, ACS, remains the most frequent cause of death and disability worldwide. In the mid-twentieth century these conditions were most common in high income countries. The elucidation

of risk factors leading to IHD and the development of therapies to reduce the deleterious consequences of ACS were responsible for dramatic reductions in these events, and in cardiovascular and all cause mortality. Although these achievements were most prominent in North America, Western Europe, and Japan, they have not affected all population groups equally. In Europe, there remains a northeast to southwest gradient, with higher prevalence in northern Russia and the Baltic nations, and considerably lower prevalence in France, Italy, and Spain.

Simultaneous with these important advances in the high income countries, the low and middle income countries have moved in the opposite direction. The improvements in agriculture, nutrition, sanitation, prevention and treatment of infections, management of maternal-early childhood disorders, as well as urbanization, and a reduction of physical labor have, in combination, led to marked increases in coronary risk factors—hypertension, cigarette smoking, obesity, diabetes mellitus, and elevations of circulating low density lipoprotein cholesterol. These, in turn, have been responsible for marked increases in ACS events and in premature mortality. The region in which these changes have been most prominent are central Asia, India, and Pakistan, as well as in the more developed regions of sub-Saharan Africa.

However, while there are many similarities, there are major differences between the rise of IHD which occurred in the high income countries in mid-twentieth century, and that which is now taking place in low and middle income countries. When the former occurred, the coronary risk factors had not yet been clearly defined and treatments of ACS were primitive by current standards. It was the successful application of prevention of IHD and therapy of ACS in high income countries that was responsible for the above-mentioned striking improvements in life expectancy. The current challenge is to apply what was learned in high-income countries to the vast populations in the low and middle income countries that are now at high risk. This will require large educational efforts directed at both the populations and their caregivers. An additional challenge will be to provide the trained specialized personnel, facilities, drugs, and devices to deal with these threats. The successful implementation of measures to reduce threats in the developing world is now principally a socio-politico-economic issue. One mitigating factor is that many of the important drugs to prevent and treat these disorders, such as statins, angiotensin converting enzyme inhibitors, diuretics, beta blockers, and calcium antagonists are off patent and are now inexpensive.

FURTHER READING

Amsterdam EA et al: 2014 AHA/ACC Guideline for the Management of Patients with Non-ST-Elevation Acute Coronary Syndromes: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 64:e139, 2014. [PubMed: 25260718]
Bohula EA et al: Atherothrombotic risk stratification and ezetimibe for secondary prevention. J Am Coll Cardiol 69:911, 2017.
Cannon CP et al: Ezetimibe added to statin therapy after acute coronary syndromes. N Engl J Med 372:2387, 2015.
de Luna AB et al: Prinzmetal angina: ECG changes and clinical considerations: A consensus paper. Ann Noninvasive Electrocardiol 19:442, 2014.
Eisen A et al: Updates on acute coronary syndrome: A review. JAMA Cardiol 1:718, 2016.
Fuster V, Kovacic JC (eds): Acute Coronary Syndrome Compendium. Circ Res 114:1847, 2014.
Kolte D et al: Trends in Coronary Angiography, Revascularization, and Outcomes of Cardiogenic Shock Complicating Non-ST-Elevation Myocardial Infarction. Am J Cardiol 117:1, 2016.
Lloyd-jones DM Decision Pathway on the Role of Non-Statin Therapies for LDL-Cholesterol Lowering in the Management of Atherosclerotic Cardiovascular Disease Risk. J Am Coll Cardiol 68:92, 2016.
Roffi M et al: 2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: Task Force for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-Segment Elevation of the European Society of Cardiology. Eur Heart J 37:267, 2016.
Steg PG et al: Effect of cangrelor on periprocedural outcomes in percutaneous coronary interventions: A pooled analysis of patient-level data. Lancet 382:1981, 2013.

Chapter 269: ST-Segment Elevation Myocardial Infarction

Elliott M. Antman; Joseph Loscalzo

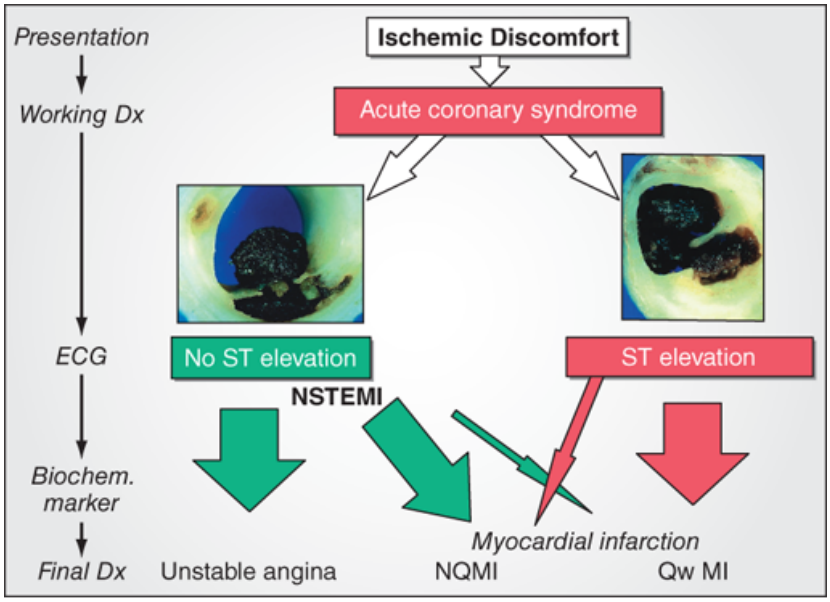
INTRODUCTION

Acute myocardial infarction (AMI) is a most common diagnosis in hospitalized patients in industrialized countries. In the United States, ~660,000 patients experience a new AMI, and 305,000 experience a recurrent AMI each year. About half of AMI-related deaths occur before the stricken individual reaches the hospital. Of note, the in-hospital mortality rate after admission for AMI has declined from 10 to about 5% over the past decade. The 1-year mortality rate after AMI is about 15%. Mortality is approximately fourfold higher in elderly patients (aged >75) as compared with younger patients.

When patients with prolonged ischemic discomfort at rest are first seen, the working clinical diagnosis is that they are suffering from an acute coronary syndrome **(Fig. 269-1)**. The 12-lead electrocardiogram (ECG) is a pivotal diagnostic and triage tool because it is at the center of the decision pathway for management; it permits distinction of those patients presenting with ST-segment elevation from those presenting without ST-segment elevation. Serum cardiac biomarkers are obtained to distinguish unstable angina (UA) from non-ST-segment elevation myocardial infarction (NSTEMI) and to assess the magnitude of an ST-segment elevation myocardial infarction (STEMI). Epidemiologic studies indicate there is a shift in the pattern of AMI over the last 15 years with more patients with NSTEMI than STEMI. This chapter focuses on the evaluation and management of patients with STEMI, while **Chap. 268** discusses UA/NSTEMI.

FIGURE 269-1

Acute coronary syndromes. Following disruption of a vulnerable plaque, patients experience ischemic discomfort resulting from a reduction of flow through the affected epicardial coronary artery. The flow reduction may be caused by a completely occlusive thrombus (*right*) or subtotally occlusive thrombus (*left*). Patients with ischemic discomfort may present with or without ST-segment elevation. Of patients with ST-segment elevation, the majority (*wide red arrow*) ultimately develop a Q wave on the ECG (Qw MI), while a minority (*thin red arrow*) do not develop Q wave and, in older literature, were said to have sustained a non-Q-wave MI (NQMI). Patients who present without ST-segment elevation are suffering from either unstable angina or a non-ST-segment elevation MI (NSTEMI) (*wide green arrows*), a distinction that is ultimately made based on the presence or absence of a serum cardiac biomarker such as CK-MB or a cardiac troponin detected in the blood. The majority of patients presenting with NSTEMI do not develop a Q wave on the ECG; a minority develop a Qw MI (*thin green arrow*). Dx, diagnosis; ECG, electrocardiogram; MI, myocardial infarction. *(Adapted from CW Hamm et al: Lancet 358:1533, 2001, and MJ Davies: Heart 83:361, 2000; with permission from the BMJ Publishing Group.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOPHYSIOLOGY: ROLE OF ACUTE PLAQUE RUPTURE

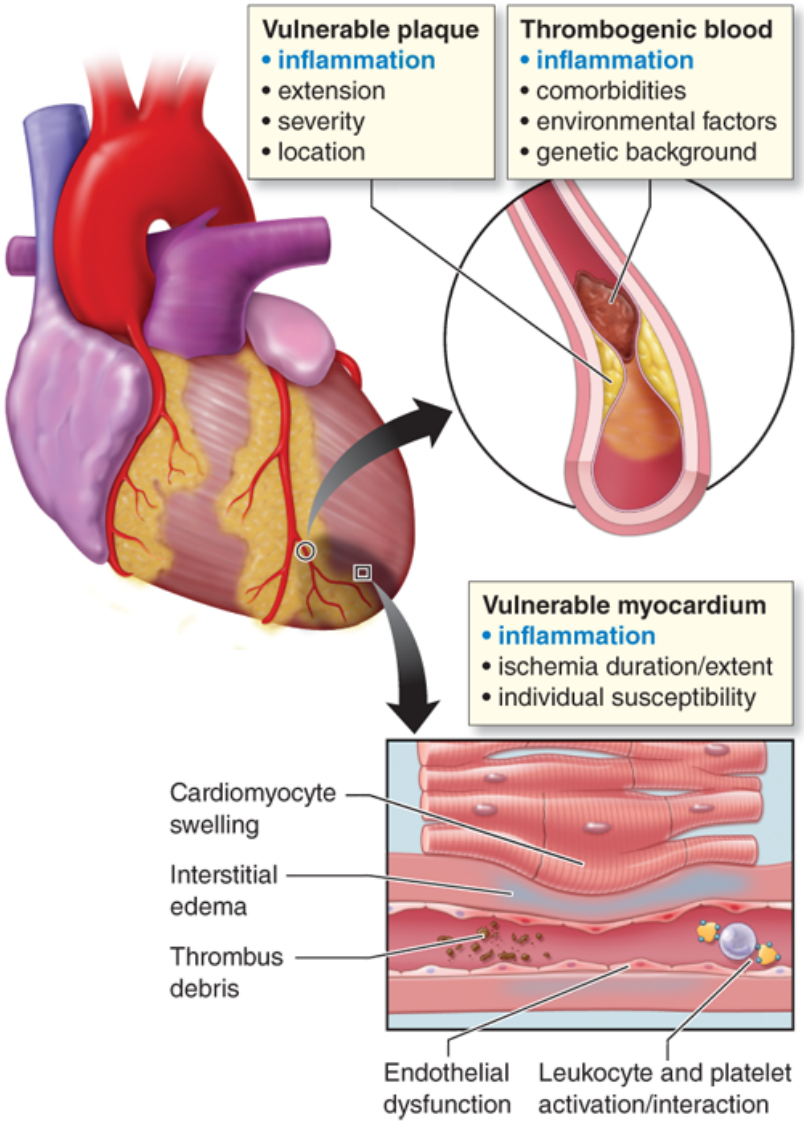
STEMI usually occurs when coronary blood flow decreases abruptly after a thrombotic occlusion of a coronary artery previously affected by atherosclerosis. Slowly developing, high-grade coronary artery stenoses do not typically precipitate STEMI because of the development of a rich collateral network over time. Instead, STEMI occurs when a coronary artery thrombus develops rapidly at a site of vascular injury. This injury is produced or facilitated by factors such as cigarette smoking, hypertension, and lipid accumulation. In most cases, STEMI occurs when the surface of an atherosclerotic plaque becomes disrupted (exposing its contents to the blood) and conditions (local or systemic) favor thrombogenesis. A mural thrombus forms at the site of plaque disruption, and the involved coronary artery becomes occluded. Histologic studies indicate that the coronary plaques prone to disruption are those with a rich lipid core and a thin fibrous cap **(Chap. 291e** from the 19th edition of *Harrison's***)**. After an initial platelet monolayer forms at the site of the disrupted plaque, various agonists (collagen, ADP, epinephrine, serotonin) promote platelet activation. After agonist stimulation of platelets, thromboxane A_2 (a potent local vasoconstrictor) is released, further platelet activation occurs, and potential resistance to fibrinolysis develops.

In addition to the generation of thromboxane A_2, activation of platelets by agonists promotes a conformational change in the glycoprotein IIb/IIIa receptor **(Chap. 111)**. Once converted to its functional state, this receptor develops a high affinity for soluble adhesive proteins (i.e., integrins) such as fibrinogen. Since fibrinogen is a multivalent molecule, it can bind to two different platelets simultaneously, resulting in platelet cross-linking and aggregation.

The coagulation cascade is activated on exposure of tissue factor in damaged endothelial cells at the site of the disrupted plaque. Factors VII and X are activated, ultimately leading to the conversion of prothrombin to thrombin, which then converts fibrinogen to fibrin **(Chap. 112)**. Fluid-phase and clot-bound thrombin participate in an autoamplification reaction leading to further activation of the coagulation cascade. The culprit coronary artery eventually becomes occluded by a thrombus containing platelet aggregates and fibrin strands **(Fig. 269-2)**.

FIGURE 269-2

Critical determinants of myocardial infarction injury. The overlapping of vulnerable plaque and thrombogenic blood are critical determinants for myocardial infarction occurrence and extension. In addition, myocardial vulnerability, which is largely due to coronary microvascular dysfunction, contributes to extension and severity of ischemic injury. In the most severe form (known as no-reflow), structural and functional impairment sustain vascular obstruction. Endothelial dysfunction triggers leukocyte and platelet activation/interaction, whereas thrombotic debris may worsen the obstruction. Furthermore, cardiomyocyte swelling, interstitial edema, and tissue inflammation promote extravascular compression. *(Reproduced from F Montecucco et al: Eur Heart J 37:1268, 2016. Published on behalf of the European Society of Cardiology. All rights reserved. © The Author 2015.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In rare cases, STEMI may be due to coronary artery occlusion caused by coronary emboli, congenital abnormalities, coronary spasm, and a wide variety of systemic—particularly inflammatory—diseases. The amount of myocardial damage caused by coronary occlusion depends on (1) the territory supplied by the affected vessel, (2) whether or not the vessel becomes totally occluded, (3) the duration of coronary occlusion, (4) the quantity of blood supplied by collateral vessels to the affected tissue, (5) the demand for oxygen of the myocardium whose blood supply has been suddenly limited, (6) endogenous factors that can produce early spontaneous lysis of the occlusive thrombus, and (7) the adequacy of myocardial perfusion in the infarct zone when flow is restored in the occluded epicardial coronary artery.

Patients at increased risk for developing STEMI include those with multiple coronary risk factors and those with UA **(Chap. 268)**. Less common underlying medical conditions predisposing patients to STEMI include hypercoagulability, collagen vascular disease, cocaine abuse, and intracardiac thrombi or masses that can produce coronary emboli.

There have been major advances in the management of STEMI with recognition that the “chain of survival” involves a highly integrated system starting with prehospital care and extending to early hospital management so as to provide expeditious implementation of a reperfusion strategy.

CLINICAL PRESENTATION

In up to one-half of cases, a precipitating factor appears to be present before STEMI, such as vigorous physical exercise, emotional stress, or a medical or surgical illness. Although STEMI may commence at any time of the day or night, circadian variations have been reported such that clusters are seen in the morning within a few hours of awakening.

Pain is the most common presenting complaint in patients with STEMI. The pain is deep and visceral; adjectives commonly used to describe it are *heavy*, *squeezing*, and *crushing*; although, occasionally, it is described as stabbing or burning **(Chap. 11)**. It is similar in character to the discomfort of angina pectoris **(Chap. 267)** but commonly occurs at rest, is usually more severe, and lasts longer. Typically, the pain involves the central portion of the chest and/or the epigastrium, and, on occasion, it radiates to the arms. Less common sites of radiation include the abdomen, back, lower jaw, and neck. The frequent location of the pain beneath the xiphoid and epigastrium and the patients' denial that they may be suffering a heart attack are chiefly responsible for the common mistaken impression of indigestion. The pain of STEMI may radiate as high as the occipital area but not below the umbilicus. It is often accompanied by weakness, sweating, nausea, vomiting, anxiety, and a sense of impending doom. The pain may commence when the patient is at rest, but when it begins during a period of exertion, it does not usually subside with cessation of activity, in contrast to angina pectoris.

The pain of STEMI can simulate pain from acute pericarditis **(Chap. 265)**, pulmonary embolism **(Chap. 273)**, acute aortic dissection **(Chap. 274)**, costochondritis, and gastrointestinal disorders. These conditions should therefore be considered in the differential diagnosis. Radiation of discomfort to the trapezius is not seen in patients with STEMI and may be a useful distinguishing feature that suggests pericarditis is the correct diagnosis. However, *pain is not uniformly present in patients with STEMI*. The proportion of painless STEMIs is greater in patients with diabetes mellitus, and it increases with age. In the elderly, STEMI may present as sudden-onset breathlessness, which may progress to pulmonary edema. Other less common presentations, with or without pain, include sudden loss of consciousness, a confusional state, a sensation of profound weakness, the appearance of an arrhythmia, evidence of peripheral embolism, or merely an unexplained drop in arterial pressure.

PHYSICAL FINDINGS

Most patients are anxious and restless, attempting unsuccessfully to relieve the pain by moving about in bed, altering their position, and stretching. Pallor associated with perspiration and coolness of the extremities occurs commonly. The combination of substernal chest pain persisting for >30 min and diaphoresis strongly suggests STEMI. Although many patients have a normal pulse rate and blood pressure within the first hour of STEMI, about one-fourth of patients with anterior infarction have manifestations of sympathetic nervous system hyperactivity (tachycardia and/or hypertension), and up to one-half with inferior infarction show evidence of parasympathetic hyperactivity (bradycardia and/or hypotension).

The precordium is usually quiet, and the apical impulse may be difficult to palpate. In patients with anterior wall infarction, an abnormal systolic pulsation caused by dyskinetic bulging of infarcted myocardium may develop in the periapical area within the first days of the illness and then may resolve. Other physical signs of ventricular dysfunction include fourth and third heart sounds, decreased intensity of the first heart sound, and paradoxical splitting of the second heart sound **(Chap. 234)**. A transient midsystolic or late systolic apical systolic murmur due to dysfunction of the mitral valve apparatus may be present. A pericardial friction rub may be heard in patients with transmural STEMI at some time in the course of the illness, if they are examined frequently. The carotid pulse is often decreased in volume, reflecting reduced stroke volume. Temperature elevations up to 38°C may be observed during the first week after STEMI. The arterial pressure is variable; in most patients with transmural infarction, systolic pressure declines by ~10–15 mmHg from the preinfarction state.

LABORATORY FINDINGS

STEMI progresses through the following temporal stages: (1) acute (first few hours–7 days), (2) healing (7–28 days), and (3) healed (≥29 days). When evaluating the results of diagnostic tests for STEMI, the temporal phase of the infarction must be considered. The laboratory tests of value in confirming the diagnosis may be divided into four groups: (1) ECG, (2) serum cardiac biomarkers, (3) cardiac imaging, and (4) nonspecific indices of tissue necrosis and inflammation.

ELECTROCARDIOGRAM

The electrocardiographic manifestations of STEMI are described in **Chap. 235**. During the initial stage, total occlusion of an epicardial coronary artery produces ST-segment elevation. Most patients initially presenting with ST-segment elevation ultimately evolve Q waves on the ECG. However, Q waves in the leads overlying the infarct zone may vary in magnitude and even appear only transiently, depending on the reperfusion status of the ischemic myocardium and restoration of transmembrane potentials over time. A small proportion of patients initially presenting with ST-segment elevation will not develop Q waves when the obstructing thrombus is not totally occlusive, obstruction is transient, or if a rich collateral network is present. Among patients presenting with ischemic discomfort but *without* ST-segment elevation, if a serum cardiac biomarker of necrosis (see below) is detected, the diagnosis of NSTEMI is ultimately made (Fig. 269-1). A minority of patients who present initially without ST-segment elevation may develop a Q-wave MI. Previously, it was believed that transmural myocardial infarction (MI) is present if the ECG demonstrates Q waves or loss of R waves, and nontransmural MI may be present if the ECG shows only transient ST-segment and T-wave changes. However, electrocardiographic-pathologic correlations are far from perfect and terms such as *Q-wave MI*, *non-Q-wave MI*, *transmural MI*, and *nontransmural MI* have been replaced by STEMI and NSTEMI (Fig. 269-1). Contemporary studies using magnetic resonance imaging (MRI) suggest that the development of a Q wave on the ECG is more dependent on the volume of infarcted tissue rather than the transmurality of infarction.

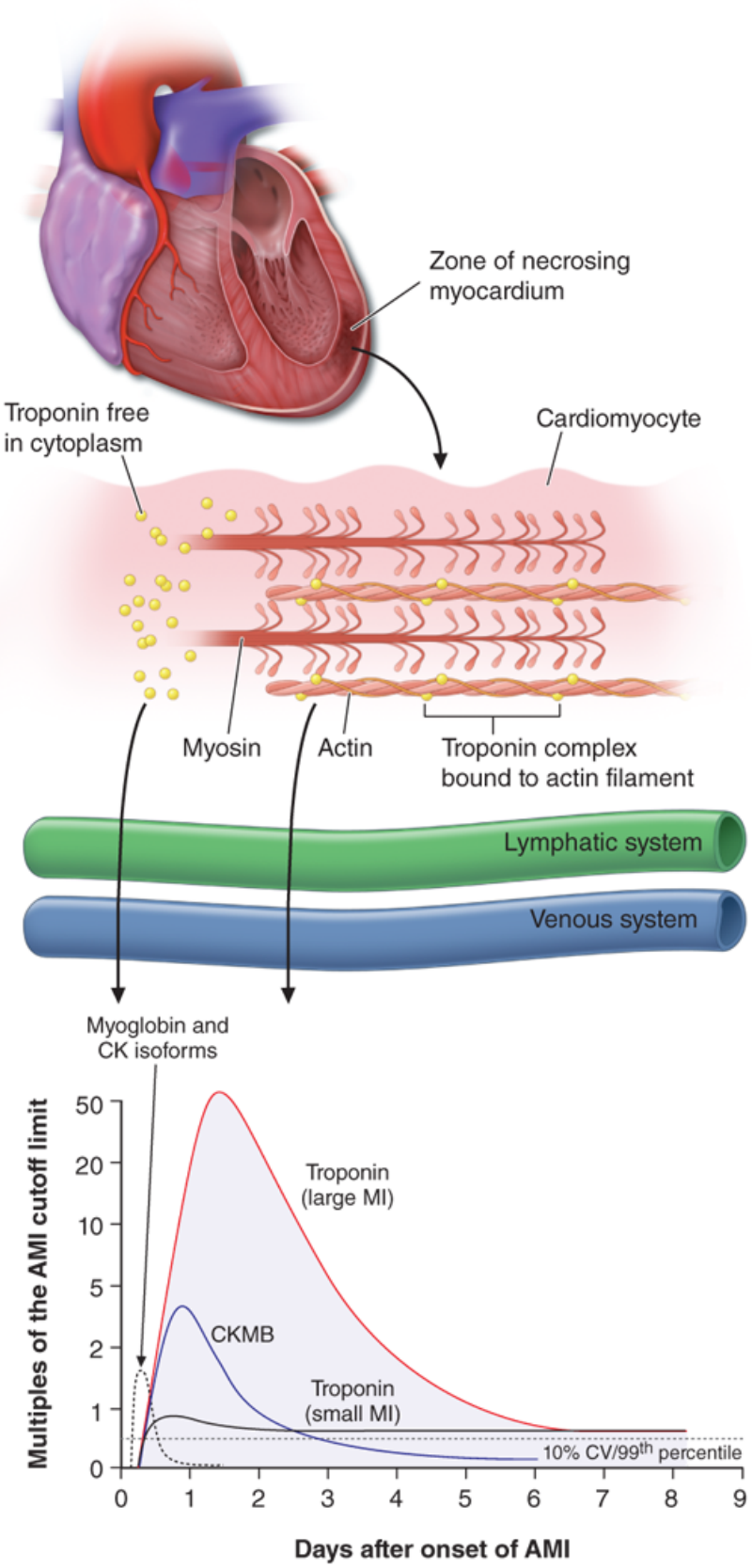
SERUM CARDIAC BIOMARKERS

Certain proteins, referred to as serum cardiac biomarkers, are released from necrotic heart muscle after STEMI. The rate of liberation of specific proteins differs depending on their intracellular location, their molecular weight, and the local blood and lymphatic flow. Cardiac biomarkers become detectable in the peripheral blood once the capacity of the cardiac lymphatics to clear the interstitium of the infarct zone is exceeded and spillover into the venous circulation occurs. The temporal pattern of protein release is of diagnostic importance. The criteria for AMI require a rise and/or fall in cardiac biomarker values with at least one value above the 99th percentile of the upper reference limit for normal individuals.

Cardiac-specific troponin T (cTnT) and *cardiac-specific troponin I* (cTnI) have amino-acid sequences different from those of the skeletal muscle forms of these proteins. These differences permitted the development of quantitative assays for cTnT and cTnI with highly specific monoclonal antibodies. cTnT and cTnI may increase after STEMI to levels many times higher than the upper reference limit (the highest value seen in 99% of a reference population not suffering from MI), the measurement of cTnT or cTnI is of considerable diagnostic usefulness, and they are now the preferred biochemical markers for MI **(Fig. 269-3)**. With improvements in the assays for the cardiac-specific troponins, it is now possible to detect concentrations <1 ng/L in patients without ischemic-type chest discomfort. The cardiac troponins are particularly valuable when there is clinical suspicion of either skeletal muscle injury or a small MI that may be below the detection limit for creatine phosphokinase (CK) and its MB isoenzyme (CK-MB) measurements, and they are, therefore, of particular value in distinguishing UA from NSTEMI. In practical terms, the high-sensitivity troponin assays are of less immediate value in patients with STEMI. Contemporary urgent reperfusion strategies necessitate making a decision (based largely on a combination of clinical and ECG findings) before the results of blood tests have returned from the laboratory. Levels of cTnI and cTnT may remain elevated for 7–10 days after STEMI.

FIGURE 269-3

The zone of necrosing myocardium is shown at the top of the figure, followed in the middle portion of the figure by a diagram of a cardiomyocyte that is in the process of releasing biomarkers. The biomarkers that are released into the interstitium are first cleared by lymphatics followed subsequently by spillover into the venous system. After disruption of the sarcolemmal membrane of the cardiomyocyte, the cytoplasmic pool of biomarkers is released first (*left-most arrow in bottom portion of figure*). Markers such as myoglobin and CK isoforms are rapidly released, and blood levels rise quickly above the cutoff limit; this is then followed by a more protracted release of biomarkers from the disintegrating myofilaments that may continue for several days. Cardiac troponin levels rise to about 20–50 times the upper reference limit (the 99th percentile of values in a reference control group) in patients who have a “classic” acute myocardial infarction (MI) and sustain sufficient myocardial necrosis to result in abnormally elevated levels of the MB fraction of creatine kinase (CK-MB). Clinicians can now diagnose episodes of microinfarction by sensitive assays that detect cardiac troponin elevations above the upper reference limit, even though CK-MB levels may still be in the normal reference range (not shown). CV, coefficient of variation. *(Modified from EM Antman: Decision making with cardiac troponin tests. N Engl J Med 346:2079, 2002 and AS Jaffe, L Babiun, FS Apple: Biomarkers in acute cardiac disease: The present and the future. J Am Coll Cardiol 48:1, 2006.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CK rises within 4–8 h and generally returns to normal by 48–72 h (Fig. 269-3). An important drawback of total CK measurement is its lack of specificity for STEMI, as CK may be elevated with skeletal muscle disease or trauma, including intramuscular injection. The MB isoenzyme of CK has the advantage over total CK that it is not present in significant concentrations in extracardiac tissue and, therefore, is considerably more specific. However, cardiac surgery, myocarditis, and electrical

cardioversion often result in elevated serum levels of the MB isoenzyme. A ratio (relative index) of CK-MB mass to CK activity ≥2.5 suggests but is not diagnostic of a myocardial rather than a skeletal muscle source for the CK-MB elevation.

Many hospitals are using cTnT or cTnI rather than CK-MB as the routine serum cardiac marker for diagnosis of STEMI, although any of these analytes remains clinically acceptable. It is *not* cost-effective to measure both a cardiac-specific troponin and CK-MB at all time points in every patient.

While it has long been recognized that the total quantity of protein released correlates with the size of the infarct, the peak protein concentration correlates only weakly with infarct size. Recanalization of a coronary artery occlusion (either spontaneously or by mechanical or pharmacologic means) in the early hours of STEMI causes earlier peaking of biomarker measurements (Fig. 269-3) because of a rapid washout from the interstitium of the infarct zone, quickly overwhelming lymphatic clearance of the proteins.

The *nonspecific reaction* to myocardial injury is associated with polymorphonuclear leukocytosis, which appears within a few hours after the onset of pain and persists for 3–7 days; the white blood cell count often reaches levels of 12,000–15,000/μL. The erythrocyte sedimentation rate rises more slowly than the white blood cell count, peaking during the first week and sometimes remaining elevated for 1 or 2 weeks.

CARDIAC IMAGING

Abnormalities of wall motion on *two-dimensional echocardiography* **(Chap. 236)** are almost universally present. Although acute STEMI cannot be distinguished from an old myocardial scar or from acute severe ischemia by echocardiography, the ease and safety of the procedure make its use appealing as a screening tool in the Emergency Department setting. When the ECG is not diagnostic of STEMI, early detection of the presence or absence of wall motion abnormalities by echocardiography can aid in management decisions, such as whether the patient should receive reperfusion therapy (e.g., fibrinolysis or a percutaneous coronary intervention [PCI]). Echocardiographic estimation of left ventricular (LV) function is useful prognostically; detection of reduced function serves as an indication for therapy with an inhibitor of the renin-angiotensin-aldosterone system. Echocardiography may also identify the presence of right ventricular (RV) infarction, ventricular aneurysm, pericardial effusion, and LV thrombus. In addition, Doppler echocardiography is useful in the detection and quantitation of a ventricular septal defect and mitral regurgitation, two serious complications of STEMI.

Several *radionuclide imaging techniques* **(Chap. 236)** are available for evaluating patients with suspected STEMI. However, these imaging modalities are used less often than echocardiography because they are more cumbersome and lack sensitivity and specificity in many clinical circumstances. Myocardial perfusion imaging with [^{201}Tl] or [^{99m}Tc]-sestamibi, which are distributed in proportion to myocardial blood flow and concentrated by viable myocardium **(Chap. 267)**, reveals a defect (“cold spot”) in most patients during the first few hours after development of a transmural infarct. Although perfusion scanning is extremely sensitive, it cannot distinguish acute infarcts from chronic scars and, thus, is not specific for the diagnosis of *acute* MI. Radionuclide ventriculography, carried out with [^{99m}Tc]-labeled red blood cells, frequently demonstrates wall motion disorders and reduction in the ventricular ejection fraction in patients with STEMI. While of value in assessing the hemodynamic consequences of infarction and in aiding in the diagnosis of RV infarction when the RV ejection fraction is depressed, this technique is nonspecific, as many cardiac abnormalities other than MI alter the radionuclide ventriculogram.

MI can be detected accurately with high-resolution cardiac MRI **(Chap. 236)** using a technique referred to as late enhancement. A standard imaging agent (gadolinium) is administered and images are obtained after a 10-min delay. Since little gadolinium enters normal myocardium, where there are tightly packed myocytes, but does percolate into the expanded intercellular region of the infarct zone, there is a bright signal in areas of infarction that appears in stark contrast to the dark areas of normal myocardium.

An Expert Consensus Task Force for the Universal Definition of Myocardial Infarction has provided a comprehensive set of criteria for the definition of MI that integrates the clinical and laboratory findings discussed earlier **(Table 269-1)** as well as a classification of MI into five types that reflect the clinical circumstances in which it may occur **(Table 269-2)**.

TABLE 269-1 Definition of Myocardial Infarction

Criteria for Acute Myocardial Infarction

The term acute myocardial infarction (MI) should be used when there is evidence of myocardial necrosis in a clinical setting consistent with acute myocardial ischemia. Under these conditions, any one of the following criteria meets the diagnosis for MI:

- Detection of a rise and/or fall of cardiac biomarker values (preferably cardiac troponin [cTn]) with at least one value above the 99th percentile upper reference limit (URL) and with at least one of the following:
    - – Symptoms of ischemia
    - – New or presumed new significant ST-segment T-wave (ST-T) changes or new left bundle branch block (LBBB)
    - – Development of pathologic Q waves in the electrocardiogram (ECG)
    - – Imaging evidence of new loss of viable myocardium or new regional wall motion abnormality
    - – Identification of an intracoronary thrombus by angiography or autopsy
- Cardiac death with symptoms suggestive of myocardial ischemia and presumed new ischemic ECG changes of new LBBB, but death occurred before cardiac biomarkers were obtained or before cardiac biomarker values would be increased.
- Percutaneous coronary intervention (PCI)–related MI is arbitrarily defined by elevation of cTn values (>5 × 99th percentile URL) in patients with normal baseline values (≤99th percentile URL) or a rise of cTn values >20% if the baseline values are elevated and are stable or falling. In addition, either (i) symptoms suggestive of myocardial ischemia, or (ii) new ischemic ECG changes, or (iii) angiographic findings consistent with a procedural complication, or (iv) imaging demonstration of new loss of viable myocardium or new regional wall motion abnormality are required.
- Stent thrombosis associated with MI when detected by coronary angiography or autopsy in the setting of myocardial ischemia and with a rise and/or fall of cardiac biomarker values with at least one value above the 99th percentile URL.
- Coronary artery bypass grafting (CABG)–related MI is arbitrarily defined by elevation of cardiac biomarker values (>10 × 99th percentile URL) in patients with normal baseline cTn values (≤99th percentile URL). In addition, either (i) new pathologic Q waves or new LBBB, or (ii) angiographic documented new graft or new native coronary artery occlusion, or (iii) imaging evidence of new loss of viable myocardium or new regional wall motion abnormality.

Criteria for Prior Myocardial Infarction

Any one of the following criteria meets the diagnosis for prior MI:

- Pathologic Q waves with or without symptoms in the absence of nonischemic causes.
- Imaging evidence of a region of loss of viable myocardium that is thinned and fails to contract, in the absence of a nonischemic cause.
- Pathologic findings of a prior MI.

Source: Data from K Thygesen: Eur Heart J 33:2551, 2012.

TABLE 269-2 Classification of Myocardial Infarction

Type I: Spontaneous Myocardial Infarction

Spontaneous myocardial infarction related to atherosclerotic plaque rupture, ulceration, fissuring, erosion, or dissection with resulting intraluminal thrombus in one or more of the coronary arteries leading to decreased myocardial blood flow or distal platelet emboli with ensuing myocyte necrosis. The patient may have underlying severe coronary artery disease (CAD) but on occasion nonobstructive or no CAD.

Type 2: Myocardial Infarction Secondary to an Ischemic Imbalance

In instances of myocardial injury with necrosis where a condition other than CAD contributes to an imbalance between myocardial oxygen supply and/or demand, e.g., coronary endothelial dysfunction, coronary artery spasm, coronary embolism, tachy-brady-arrhythmias, anemia, respiratory failure, hypotension, and hypertension with or without left ventricular hypertrophy.

Type 3: Myocardial Infarction Resulting in Death When Biomarker Values Are Unavailable

Cardiac death with symptoms suggestive of myocardial ischemia and presumed new ischemic electrocardiogram (ECG) changes or new left bundle branch block (LBBB), but death occurring before blood samples could be obtained or before cardiac biomarker could rise, or in rare cases, cardiac biomarkers were not collected.

Type 4a: Myocardial Infarction Related to Percutaneous Coronary Intervention (PCI)

Myocardial infarction associated with PCI is arbitrarily defined by elevation of cardiac troponin (cTn) values >5 × 99th percentile upper reference limit (URL) in patients with normal baseline values (≤99th percentile URL) or a rise of cTn values >20% if the baseline values are elevated and are stable or falling. In addition, either (i) symptoms suggestive of myocardial ischemia, or (ii) new ischemic ECG changes or new LBBB, or (iii) angiographic loss of patency of a major coronary artery or a side branch or persistent slow or no flow or embolization, or (iv) imaging demonstration of new loss of viable myocardium or new regional wall motion abnormality is required.

Type 4b: Myocardial Infarction Related to Stent Thrombosis

Myocardial infarction associated with stent thrombosis is detected by coronary angiography or autopsy in the setting of myocardial ischemia and with a rise and/or fall of cardiac biomarker values with at least one value above the 99th percentile URL.

Type 5: Myocardial Infarction Related to Coronary Artery Bypass Grafting (CABG)

Myocardial infarction associated with CABG is arbitrarily defined by elevation of cardiac biomarker values >10 × 99th percentile URL in patients with normal baseline cTn values (≤99th percentile URL). In addition, either (i) new pathologic Q waves or new LBBB, or (ii) angiographic documented new graft or new native coronary artery occlusion, or (iii) imaging evidence of new loss of viable myocardium or new regional wall motion abnormality.

Source: K Thygesen: Eur Heart J 33:2551, 2012.

INITIAL MANAGEMENT

PREHOSPITAL CARE

The prognosis in STEMI is largely related to the occurrence of two general classes of complications: (1) electrical complications (arrhythmias) and (2) mechanical complications (“pump failure”). Most out-of-hospital deaths from STEMI are due to the sudden development of ventricular fibrillation. The vast majority of deaths due to ventricular fibrillation occur within the first 24 h of the onset of symptoms, and of these, over half occur in the first hour. Therefore, the major elements of prehospital care of patients with suspected STEMI include (1) recognition of symptoms by the patient and prompt seeking of medical attention; (2) rapid deployment of an emergency medical team capable of performing resuscitative maneuvers, including defibrillation; (3) expeditious transportation of the patient to a hospital facility that is continuously staffed by physicians and nurses skilled in managing arrhythmias and providing advanced cardiac life support; and (4) expeditious implementation of reperfusion therapy. The greatest delay usually occurs not during transportation to the hospital but, rather, between the onset of pain and the patient’s decision to call for help. This delay can best be reduced by health care professionals educating the public concerning the significance of chest discomfort and the importance of seeking early medical attention. Regular office visits with patients having a history of, or who are at risk for ischemic heart disease are important “teachable moments” for clinicians to review the symptoms of STEMI and the appropriate action plan.

Increasingly, monitoring and treatment are carried out by trained personnel in the ambulance, further shortening the time between the onset of the infarction and appropriate treatment. General guidelines for initiation of fibrinolysis in the prehospital setting include the ability to transmit 12-lead ECGs to confirm the diagnosis, the presence of paramedics in the ambulance, training of paramedics in the interpretation of ECGs and management of STEMI, and online medical command and control that can authorize the initiation of treatment in the field.

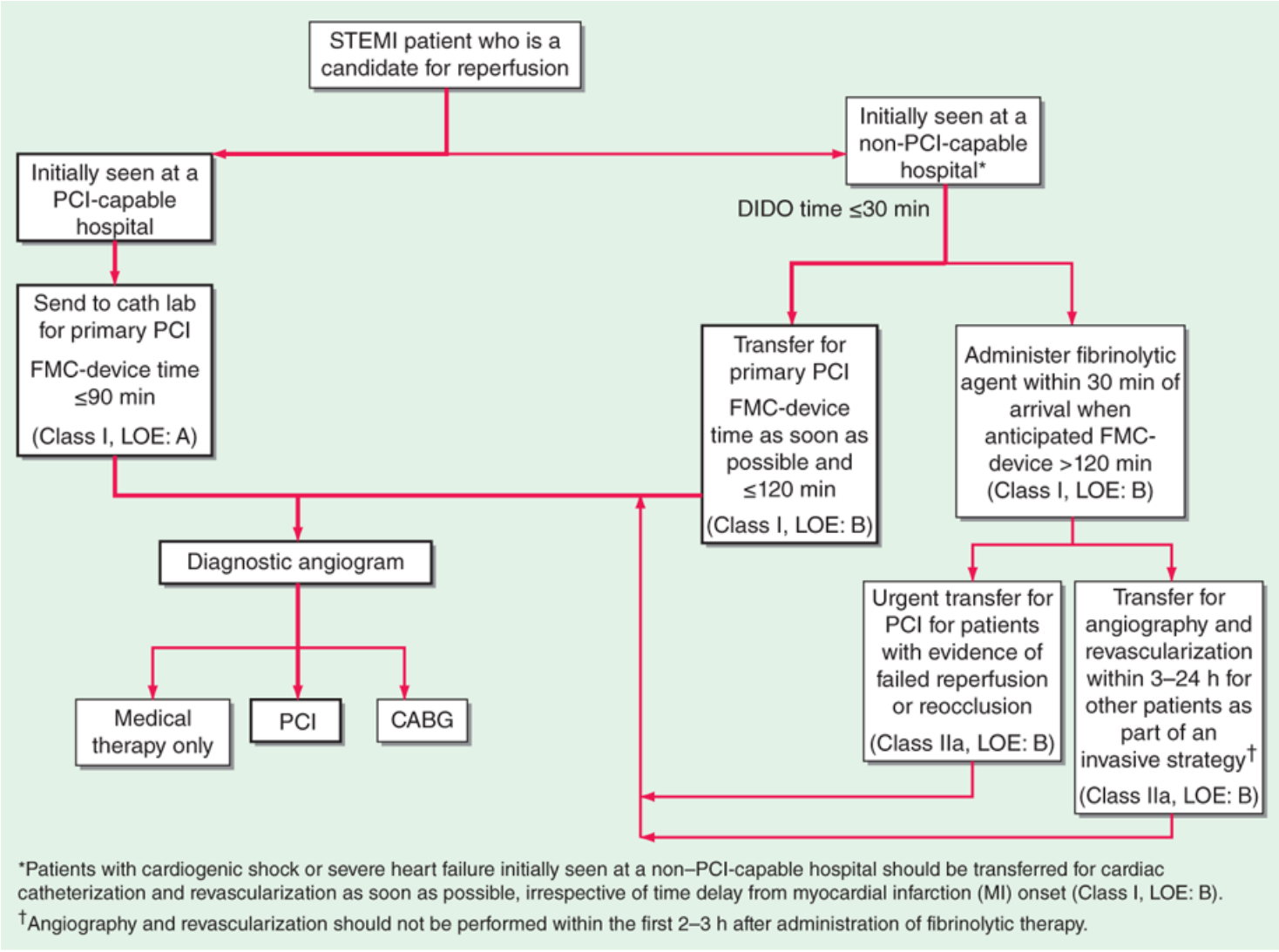
MANAGEMENT IN THE EMERGENCY DEPARTMENT

In the Emergency Department, the goals for the management of patients with suspected STEMI include control of cardiac discomfort, rapid identification of patients who are candidates for urgent reperfusion therapy, triage of lower-risk patients to the appropriate location in the hospital, and avoidance of inappropriate discharge of patients with STEMI. Many aspects of the treatment of STEMI are initiated in the Emergency Department and then continued during the in-hospital phase of management

(Fig. 269-4). The overarching goal is to minimize the time from first medical contact to initiation of reperfusion therapy. This may involve transfer from a non-PCI hospital to one that is PCI capable, with a goal of initiating PCI within 120 min of first medical contact (Fig. 269-4).

FIGURE 269-4

Reperfusion therapy for patients with ST-segment elevation myocardial infarction (STEMI). The *bold arrows* and *boxes* are the preferred strategies. Performance of percutaneous coronary intervention (PCI) is dictated by an anatomically appropriate culprit stenosis. CABG, coronary artery bypass graft; DIDO, door-in–door-out; FMC, first medical contact; LOE, level of evidence; STEMI, ST-elevation myocardial infarction. *(Adapted with permission from P O'Gara et al: Circulation 127:e362, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Aspirin is essential in the management of patients with suspected STEMI and is effective across the entire spectrum of acute coronary syndromes (Fig. 269-1). Rapid inhibition of cyclooxygenase-1 in platelets followed by a reduction of thromboxane A_2 levels is achieved by buccal absorption of a chewed 160–325-mg tablet in the Emergency Department. This measure should be followed by daily oral administration of aspirin in a dose of 75–162 mg.

In patients whose arterial O_2 saturation is normal, supplemental O_2 is of limited if any clinical benefit and therefore is not cost-effective. However, when hypoxemia is present, O_2 should be administered by nasal prongs or face mask (2–4 L/min) for the first 6–12 h after infarction; the patient should then be reassessed to determine if there is a continued need for such treatment.

CONTROL OF DISCOMFORT

Sublingual *nitroglycerin* can be given safely to most patients with STEMI. Up to three doses of 0.4 mg should be administered at about 5-min intervals. In addition to diminishing or abolishing chest discomfort, nitroglycerin may be capable of both decreasing myocardial oxygen demand (by lowering preload) and increasing myocardial oxygen supply (by dilating infarct-related coronary vessels or collateral vessels). In patients whose initially favorable response to sublingual nitroglycerin is followed by the return of chest discomfort, particularly if accompanied by other evidence of ongoing ischemia such as further ST-segment or T-wave shifts, the use of intravenous nitroglycerin should be considered. Therapy with nitrates should be avoided in patients who present with low systolic arterial pressure (<90 mmHg) or in whom there is clinical suspicion of RV infarction (inferior infarction on ECG, elevated jugular venous pressure, clear lungs, and hypotension). Nitrates should not be administered to patients who have taken a phosphodiesterase-5 inhibitor for erectile dysfunction within the preceding 24 h, because it may potentiate the hypotensive effects of nitrates. An idiosyncratic reaction to nitrates, consisting of sudden marked hypotension, sometimes occurs but can usually be reversed promptly by the rapid administration of intravenous atropine.

Morphine is a very effective analgesic for the pain associated with STEMI. However, it may reduce sympathetically mediated arteriolar and venous constriction, and the resulting venous pooling may reduce cardiac output and arterial pressure. These hemodynamic disturbances usually respond promptly to elevation of the legs, but in some patients, volume expansion with intravenous saline is required. The patient may experience diaphoresis and nausea, but these events usually pass and are replaced by a feeling of well-being associated with the relief of pain. Morphine also has a vagotonic effect and may cause bradycardia or advanced degrees of heart block, particularly in patients with inferior infarction. These side effects usually respond to atropine (0.5 mg intravenously). Morphine is routinely administered by repetitive (every 5 min) intravenous injection of small doses (2–4 mg), rather than by the subcutaneous administration of a larger quantity, because absorption may be unpredictable by the latter route.

Intravenous *beta blockers* are also useful in the control of the pain of STEMI. These drugs control pain effectively in some patients, presumably by diminishing myocardial O_2 demand and hence ischemia. More important, there is evidence that intravenous beta blockers reduce the risks of reinfarction and ventricular fibrillation (see "Beta-Adrenoceptor Blockers" below). A commonly employed regimen is metoprolol, 5 mg every 2–5 min for a total of three doses, provided the patient has a heart rate >60 beats/min, systolic pressure >100 mmHg, a PR interval <0.24 s, and rales that are no higher than 10 cm up from the diaphragm. Fifteen minutes after the last intravenous dose, an oral regimen is initiated of 50 mg every 6 h for 48 h, followed by 100 mg every 12 h.

Patient selection is important when considering beta blockers for STEMI. Oral beta blocker therapy should be initiated in the first 24 h for patients who do not have any of the following: (1) signs of heart failure, (2) evidence of a low-output state, (3) increased risk for cardiogenic shock, or (4) other relative contraindications to beta blockade (PR interval >0.24 s, second- or third-degree heart block, active asthma, or reactive airway disease).

Unlike beta blockers, calcium antagonists are of little value in the acute setting, and there is evidence that short-acting dihydropyridines may be associated with an increased mortality risk.

MANAGEMENT STRATEGIES

The primary tool for screening patients and making triage decisions is the initial 12-lead ECG. When ST-segment elevation of at least 2 mm in two contiguous precordial leads and 1 mm in two adjacent limb leads is present, a patient should be considered a candidate for *reperfusion therapy* (Figs. 269-1 and 269-4). The process of selecting patients for fibrinolysis versus primary PCI (angioplasty or stenting; **Chap. 270**) is discussed below. In the absence of ST-segment elevation, fibrinolysis is not helpful, and evidence exists suggesting that it may be harmful.

LIMITATION OF INFARCT SIZE

The quantity of myocardium that becomes necrotic as a consequence of a coronary artery occlusion is determined by factors other than just the site of occlusion. While the central zone of the infarct contains necrotic tissue that is irretrievably lost, the fate of the surrounding ischemic myocardium (ischemic penumbra) may be improved by timely restoration of coronary perfusion, reduction of myocardial O_2 demands, prevention of the accumulation of noxious metabolites, and blunting of the impact of mediators of reperfusion injury (e.g., calcium overload and oxygen-derived free radicals). Up to one-third of patients with STEMI may achieve *spontaneous* reperfusion of the infarct-related coronary artery within 24 h and experience improved healing of infarcted tissue. Reperfusion, either pharmacologically (by fibrinolysis) or by PCI, accelerates the opening of infarct-related arteries in those patients in whom spontaneous fibrinolysis ultimately would have occurred and also greatly increases the number of patients in whom restoration of flow in the infarct-related artery is accomplished. Timely restoration of flow in the epicardial infarct–related artery combined with improved perfusion of the downstream zone of infarcted myocardium results in a limitation of infarct size. Protection of the ischemic myocardium by the maintenance of an optimal balance between myocardial O_2 supply and demand through pain control, treatment of congestive heart failure (CHF), and minimization of tachycardia and hypertension extends the "window" of time for the salvage of myocardium by reperfusion strategies.

Glucocorticoids and nonsteroidal anti-inflammatory agents, with the exception of aspirin, should be avoided in patients with STEMI. They can impair infarct healing and increase the risk of myocardial rupture, and their use may result in a larger infarct scar. In addition, they can increase coronary vascular resistance, thereby potentially reducing flow to ischemic myocardium.

PRIMARY PERCUTANEOUS CORONARY INTERVENTION

(See also Chap. 270) PCI, usually angioplasty and/or stenting without preceding fibrinolysis, referred to as *primary PCI*, is effective in restoring perfusion in STEMI when carried out on an emergency basis in the first few hours of MI. It has the advantage of being applicable to patients who have contraindications to fibrinolytic therapy (see below) but otherwise are considered appropriate candidates for reperfusion. It appears to be more effective than fibrinolysis in opening occluded coronary arteries and, *when performed by experienced operators in dedicated medical centers*, is associated with better short-term and long-term clinical outcomes. Compared with fibrinolysis, primary PCI is generally preferred when the diagnosis is in doubt, cardiogenic shock is present, bleeding risk is increased, or symptoms have been present for at least 2–3 h when the clot is more mature and less easily lysed by fibrinolytic drugs. However, PCI is expensive in terms of personnel and facilities, and its applicability is limited by its availability, around the clock, in only a minority of hospitals (Fig. 269-4).

FIBRINOLYSIS

If no contraindications are present (see below), fibrinolytic therapy should ideally be initiated within 30 min of presentation (i.e., door-to-needle time ≤30 min). The principal goal of fibrinolysis is prompt restoration of full coronary arterial patency. The fibrinolytic agents tissue plasminogen activator (tPA), streptokinase, tenecteplase (TNK), and reteplase (rPA) have been approved by the U.S. Food and Drug Administration for intravenous use in patients with STEMI. These drugs all act by promoting the conversion of plasminogen to plasmin, which subsequently lyses fibrin thrombi. Although considerable emphasis was first placed on a distinction between more fibrin-specific agents, such as tPA, and non-fibrin-specific agents, such as streptokinase, it is now recognized that these differences are only relative, as some degree of systemic fibrinolysis occurs with the former agents. TNK and rPA are referred to as *bolus fibrinolytics* since their administration does not require a prolonged intravenous infusion.

When assessed angiographically, flow in the culprit coronary artery is described by a simple qualitative scale called the *Thrombolysis in Myocardial Infarction (TIMI) grading system*: grade 0 indicates complete occlusion of the infarct-related artery; grade 1 indicates some penetration of the contrast material beyond the point of obstruction, but without perfusion of the distal coronary bed; grade 2 indicates perfusion of the entire infarct vessel into the distal bed, but with flow that is delayed compared with that of a normal artery; and grade 3 indicates full perfusion of the infarct vessel with normal flow. The latter is the goal of reperfusion therapy, because full perfusion of the infarct-related coronary artery yields far better results in terms of limiting infarct size, maintenance of LV function, and reduction of both short- and long-term mortality rates. Additional methods of angiographic assessment of the efficacy of fibrinolysis include counting the number of frames on the cine film required for dye to flow from the origin of the infarct-related artery to a landmark in the distal vascular bed (*TIMI frame count*) and determining the rate of entry and exit of contrast dye from the microvasculature in the myocardial infarct zone (*TIMI myocardial perfusion grade*). These methods have an even tighter correlation with outcomes after STEMI than the more commonly employed TIMI flow grade.

tPA and the other relatively fibrin-specific plasminogen activators, rPA and TNK, are more effective than streptokinase at restoring full perfusion—i.e., TIMI grade 3 coronary flow—and have a small edge in improving survival as well. The current recommended regimen of tPA consists of a 15-mg bolus followed by 50 mg intravenously over the first 30 min, followed by 35 mg over the next 60 min. Streptokinase is administered as 1.5 million units (MU) intravenously over 1 h. rPA is administered in a double-bolus regimen consisting of a 10-MU bolus given over 2–3 min, followed by a second 10-MU bolus 30 min later. TNK is given as a single weight-based intravenous bolus of 0.53 mg/kg over 10 s. In addition to the fibrinolytic agents discussed earlier, pharmacologic reperfusion typically involves adjunctive antiplatelet and antithrombotic drugs, as discussed subsequently.

Clear contraindications to the use of fibrinolytic agents include a history of cerebrovascular hemorrhage at any time, a nonhemorrhagic stroke or other cerebrovascular event within the past year, marked hypertension (a reliably determined systolic arterial pressure >180 mmHg and/or a diastolic pressure >110 mmHg) at any time during the acute presentation, suspicion of aortic dissection, and active internal bleeding (excluding menses). While advanced age is associated with an increase in hemorrhagic complications, the benefit of fibrinolytic therapy in the elderly appears to justify its use if no other contraindications are present and the amount of myocardium in jeopardy appears to be substantial.

Relative contraindications to fibrinolytic therapy, which require assessment of the risk-to-benefit ratio, include current use of anti-coagulants (international normalized ratio ≥2), a recent (<2 weeks) invasive or surgical procedure or prolonged (>10 min) cardiopulmonary resuscitation, known bleeding diathesis, pregnancy, a hemorrhagic ophthalmic condition (e.g., hemorrhagic diabetic retinopathy), active peptic ulcer disease, and a history of severe hypertension that is currently adequately controlled. Because of the risk of an allergic reaction, patients should not receive streptokinase if that agent had been received within the preceding 5 days to 2 years.

Allergic reactions to streptokinase occur in ~2% of patients who receive it. While a minor degree of hypotension occurs in 4–10% of patients given this agent, marked hypotension occurs, although rarely, in association with severe allergic reactions.

Hemorrhage is the most frequent and potentially the most serious complication. Because bleeding episodes that require transfusion are more common when patients require invasive procedures, unnecessary venous or arterial interventions should be avoided in patients receiving fibrinolytic agents. Hemorrhagic stroke is the most serious complication and occurs in ~0.5–0.9% of patients being treated with these agents. This rate increases with advancing age, with patients >70 years experiencing

roughly twice the rate of intracranial hemorrhage as those <65 years. Large-scale trials have suggested that the rate of intracranial hemorrhage with tPA or rPA is slightly higher than with streptokinase.

INTEGRATED REPERFUSION STRATEGY

Evidence has emerged that suggests PCI plays an increasingly important role in the management of STEMI. Prior approaches that segregated the pharmacologic and catheter-based approaches to reperfusion have now been replaced with an integrated approach to triage and transfer of STEMI patients to receive PCI (Fig. 269-4). To achieve the degree of integration required to care for a patient with STEMI, all communities should create and maintain a regional system of STEMI care that includes assessment and continuous quality improvement of emergency medical services and hospital-based activities.

Cardiac catheterization and coronary angiography should be carried out after fibrinolytic therapy if there is evidence of either (1) failure of reperfusion (persistent chest pain and ST-segment elevation >90 min), in which case a *rescue PCI* should be considered; or (2) coronary artery reocclusion (re-elevation of ST segments and/or recurrent chest pain) or the development of recurrent ischemia (such as recurrent angina in the early hospital course or a positive exercise stress test before discharge), in which case an *urgent PCI* should be considered. Routine angiography and *elective PCI* even in asymptomatic patients following administration of fibrinolytic therapy are used with less frequency, given the numerous technologic advances that have occurred in the catheterization laboratory and the increasing number of skilled interventionalists. Coronary artery bypass surgery should be reserved for patients whose coronary anatomy is unsuited to PCI but in whom revascularization appears to be advisable because of extensive jeopardized myocardium or recurrent ischemia.

HOSPITAL PHASE MANAGEMENT

CORONARY CARE UNITS

These units are routinely equipped with a system that permits continuous monitoring of the cardiac rhythm of each patient and hemodynamic monitoring in selected patients. Defibrillators, respirators, noninvasive transthoracic pacemakers, and facilities for introducing pacing catheters and flow-directed balloon-tipped catheters are also usually available. Equally important is the organization of a highly trained team of nurses who can recognize arrhythmias; adjust the dosage of antiarrhythmic, vasoactive, and anticoagulant drugs; and perform cardiac resuscitation, including electroshock, when necessary.

Patients should be admitted to a coronary care unit early in their illness when it is expected that they will derive benefit from the sophisticated and expensive care provided. The availability of electrocardiographic monitoring and trained personnel outside the coronary care unit has made it possible to admit lower-risk patients (e.g., those not hemodynamically compromised and without active arrhythmias) to “intermediate care units.”

The duration of stay in the coronary care unit is dictated by the ongoing need for intensive care. If symptoms are controlled with oral therapy, patients may be transferred out of the coronary care unit. Also, patients who have a confirmed STEMI but who are considered to be at low risk (no prior infarction and no persistent chest discomfort, CHF, hypotension, or cardiac arrhythmias) may be safely transferred out of the coronary care unit within 24 h.

Activity

Factors that increase the work of the heart during the initial hours of infarction may increase the size of the infarct. Therefore, patients with STEMI should be kept at bed rest for the first 6–12 h. However, in the absence of complications, patients should be encouraged, under supervision, to resume an upright posture by dangling their feet over the side of the bed and sitting in a chair within the first 24 h. This practice is psychologically beneficial and usually results in a reduction in the pulmonary capillary wedge pressure. In the absence of hypotension and other complications, by the second or third day, patients typically are ambulating in their room with increasing duration and frequency, and they may shower or stand at the sink to bathe. By day 3 after infarction, patients should be increasing their ambulation progressively to a goal of 185 m (600 ft) at least three times a day.

Diet

Because of the risk of emesis and aspiration soon after STEMI, patients should receive either nothing or only clear liquids by mouth for the first 4–12 h. The typical coronary care unit diet should provide ≤30% of total calories as fat and have a cholesterol content of ≤300 mg/d. Complex carbohydrates should make up 50–55% of total calories. Portions should not be unusually large, and the menu should be enriched with foods that are high in potassium, magnesium, and fiber, but low in sodium. Diabetes mellitus and hypertriglyceridemia are managed by restriction of concentrated sweets in the diet.

Bowel Management

Bed rest and the effect of the narcotics used for the relief of pain often lead to constipation. A bedside commode rather than a bedpan, a diet rich in bulk, and the routine use of a stool softener such as dioctyl sodium sulfosuccinate (200 mg/d) are

recommended. If the patient remains constipated despite these measures, a laxative can be prescribed. Contrary to prior belief, it is safe to perform a gentle rectal examination on patients with STEMI.

Sedation

Many patients require sedation during hospitalization to withstand the period of enforced inactivity with tranquility. Diazepam (5 mg), oxazepam (15–30 mg), or lorazepam (0.5–2 mg), given three to four times daily, is usually effective. An additional dose of any of the above medications may be given at night to ensure adequate sleep. Attention to this problem is especially important during the first few days in the coronary care unit, where the atmosphere of 24-h vigilance may interfere with the patient's sleep. However, sedation is no substitute for reassuring, quiet surroundings. Many drugs used in the coronary care unit, such as atropine, H_2 blockers, and narcotics, can produce delirium, particularly in the elderly. This effect should not be confused with agitation, and it is wise to conduct a thorough review of the patient's medications before arbitrarily prescribing additional doses of anxiolytics.

PHARMACOTHERAPY

ANTITHROMBOTIC AGENTS

The use of antiplatelet and anticoagulant therapy during the initial phase of STEMI is based on extensive laboratory and clinical evidence that thrombosis plays an important role in the pathogenesis of this condition. The primary goal of treatment with antiplatelet and anticoagulant agents is to maintain patency of the infarct-related artery, in conjunction with reperfusion strategies. A secondary goal is to reduce the patient's tendency to thrombosis and, thus, the likelihood of mural thrombus formation or deep-venous thrombosis, either of which could result in pulmonary embolization. The degree to which antiplatelet and anticoagulant therapy achieves these goals partly determines how effectively it reduces the risk of mortality from STEMI.

As noted previously (see "Management in the Emergency Department" earlier), aspirin is the standard antiplatelet agent for patients with STEMI. The most compelling evidence for the benefits of antiplatelet therapy (mainly with aspirin) in STEMI is found in the comprehensive overview by the Antiplatelet Trialists' Collaboration. Data from nearly 20,000 patients with MI enrolled in 15 randomized trials were pooled and revealed a relative reduction of 27% in the mortality rate, from 14.2% in control patients to 10.4% in patients receiving antiplatelet agents.

Inhibitors of the $P2Y_{12}$ ADP receptor prevent activation and aggregation of platelets. The addition of the $P2Y_{12}$ inhibitor clopidogrel to background treatment with aspirin to STEMI patients reduces the risk of clinical events (death, reinfarction, stroke) and, in patients receiving fibrinolytic therapy, has been shown to prevent reocclusion of a successfully reperfused infarct artery. New $P2Y_{12}$ ADP receptor antagonists, such as prasugrel and ticagrelor, are more effective than clopidogrel in preventing ischemic complications in STEMI patients undergoing PCI, but are associated with an increased risk of bleeding. Glycoprotein IIb/IIIa receptor inhibitors appear useful for preventing thrombotic complications in patients with STEMI undergoing PCI.

The standard anticoagulant agent used in clinical practice is unfractionated heparin (UFH). The available data suggest that when UFH is added to a regimen of aspirin and a non-fibrin-specific thrombolytic agent such as streptokinase, additional mortality benefit occurs (about 5 lives saved per 1000 patients treated). It appears that the immediate administration of intravenous UFH, in addition to a regimen of aspirin and relatively fibrin-specific fibrinolytic agents (tPA, rPA, or TNK), helps to maintain patency of the infarct-related artery. This effect is achieved at the cost of a small increased risk of bleeding. The recommended dose of UFH is an initial bolus of 60 U/kg (maximum 4000 U) followed by an initial infusion of 12 U/kg per h (maximum 1000 U/h). The activated partial thromboplastin time during maintenance therapy should be 1.5–2 times the control value.

Alternatives to UFH for anticoagulation of patients with STEMI are the low-molecular-weight heparin (LMWH) preparations, a synthetic version of the critical pentasaccharide sequence (fondaparinux), and the direct antithrombin bivalirudin. Advantages of LMWHs include high bioavailability permitting administration subcutaneously, reliable anticoagulation without monitoring, and greater antiXa:IIa activity. Enoxaparin has been shown to reduce significantly the composite endpoints of death/nonfatal reinfarction and death/nonfatal reinfarction/urgent revascularization compared with UFH in STEMI patients who receive fibrinolysis. Treatment with enoxaparin is associated with higher rates of serious bleeding, but net clinical benefit—a composite endpoint that combines efficacy and safety—still favors enoxaparin over UFH. Interpretation of the data on fondaparinux is difficult because of the complex nature of the pivotal clinical trial evaluating it in STEMI (OASIS-6). Fondaparinux appears superior to placebo in STEMI patients not receiving reperfusion therapy, but its relative efficacy and safety compared with UFH is less certain. Owing to the risk of catheter thrombosis, fondaparinux should not be used alone at the time of coronary angiography and PCI but should be combined with another anticoagulant with antithrombin activity such as UFH or bivalirudin. Contemporary trials of bivalirudin used an open-label design to evaluate its efficacy and safety compared with UFH plus a glycoprotein IIb/IIIa inhibitor. Bivalirudin was associated with a lower rate of bleeding, largely driven by reductions in vascular access site hematomas ≥5 cm or the administration of blood transfusions.

Patients with an anterior location of the infarction, severe LV dysfunction, heart failure, a history of embolism, two-dimensional echocardiographic evidence of mural thrombus, or atrial fibrillation are at increased risk of systemic or pulmonary thromboembolism. Such individuals should receive full therapeutic levels of anticoagulant therapy (LMWH or UFH) while hospitalized, followed by at least 3 months of warfarin therapy.

BETA-ADRENOCEPTOR BLOCKERS

The benefits of beta blockers in patients with STEMI can be divided into those that occur immediately when the drug is given acutely and those that accrue over the long term when the drug is given for secondary prevention after an infarction. Acute intravenous beta blockade improves the myocardial O_2 supply-demand relationship, decreases pain, reduces infarct size, and decreases the incidence of serious ventricular arrhythmias. In patients who undergo fibrinolysis soon after the onset of chest pain, no incremental reduction in mortality rate is seen with beta blockers, but recurrent ischemia and reinfarction are reduced.

Thus, beta-blocker therapy after STEMI is useful for most patients (including those treated with an angiotensin-converting enzyme [ACE] inhibitor) except those in whom it is specifically contraindicated (patients with heart failure or severely compromised LV function, heart block, orthostatic hypotension, or a history of asthma) and perhaps those whose excellent long-term prognosis (defined as an expected mortality rate of <1% per year, patients <55 years, no previous MI, with normal ventricular function, no complex ventricular ectopy, and no angina) markedly diminishes any potential benefit.

INHIBITION OF THE RENIN-ANGIOTENSIN-ALDOSTERONE SYSTEM

ACE inhibitors reduce the mortality rate after STEMI, and the mortality benefits are additive to those achieved with aspirin and beta blockers. The maximum benefit is seen in high-risk patients (those who are elderly or who have an anterior infarction, a prior infarction, and/or globally depressed LV function), but evidence suggests that a short-term benefit occurs when ACE inhibitors are prescribed unselectively to all hemodynamically stable patients with STEMI (i.e., those with a systolic pressure >100 mmHg). The mechanism involves a reduction in ventricular remodeling after infarction (see "Ventricular Dysfunction" later) with a subsequent reduction in the risk of CHF. The rate of recurrent infarction may also be lower in patients treated chronically with ACE inhibitors after infarction.

Before hospital discharge, LV function should be assessed with an imaging study. ACE inhibitors should be continued indefinitely in patients who have clinically evident CHF, in patients in whom an imaging study shows a reduction in global LV function or a large regional wall motion abnormality, or in those who are hypertensive.

Angiotensin receptor blockers (ARBs) should be administered to STEMI patients who are intolerant of ACE inhibitors and who have either clinical or radiologic signs of heart failure. Long-term aldosterone blockade should be prescribed for STEMI patients without significant renal dysfunction (creatinine ≥2.5 mg/dL in men and ≥2.0 mg/dL in women) or hyperkalemia (potassium ≥5.0 mEq/L) who are already receiving therapeutic doses of an ACE inhibitor, have an LV ejection fraction ≤40%, and have either symptomatic heart failure or diabetes mellitus. A multidrug regimen for inhibiting the renin-angiotensin-aldosterone system has been shown to reduce both heart failure–related and sudden cardiac death–related cardiovascular mortality after STEMI, but has not been as thoroughly explored as ACE inhibitors in STEMI patients.

OTHER AGENTS

Favorable effects on the ischemic process and ventricular remodeling (see below) previously led many physicians to routinely use *intravenous nitroglycerin* (5–10 µg/min initial dose and up to 200 µg/min as long as hemodynamic stability is maintained) for the first 24–48 h after the onset of infarction. However, the benefits of routine use of intravenous nitroglycerin are less in the contemporary era where beta-adrenoceptor blockers and ACE inhibitors are routinely prescribed for patients with STEMI.

Results of multiple trials of different calcium antagonists have failed to establish a role for these agents in the treatment of most patients with STEMI. Therefore, the routine use of calcium antagonists cannot be recommended. Strict control of blood glucose in diabetic patients with STEMI has been shown to reduce the mortality rate. Serum magnesium should be measured in all patients on admission, and any demonstrated deficits should be corrected to minimize the risk of arrhythmias.

COMPLICATIONS AND THEIR MANAGEMENT

VENTRICULAR DYSFUNCTION

After STEMI, the left ventricle undergoes a series of changes in shape, size, and thickness in both the infarcted and noninfarcted segments. This process is referred to as *ventricular remodeling* and generally precedes the development of clinically evident CHF in the months to years after infarction. Soon after STEMI, the left ventricle begins to dilate. Acutely, this results from expansion of the infarct, i.e., slippage of muscle bundles, disruption of normal myocardial cells, and tissue loss within the necrotic zone, resulting in disproportionate thinning and elongation of the infarct zone. Later, lengthening of the noninfarcted segments occurs as well. The overall chamber enlargement that occurs is related to the size and location of the infarct, with greater dilation following infarction of the anterior wall and apex of the left ventricle and causing more marked hemodynamic impairment, more frequent heart failure, and a poorer prognosis. Progressive dilation and its clinical consequences may be ameliorated by therapy with ACE inhibitors and other vasodilators (e.g., nitrates). In patients with an ejection fraction <40%, regardless of whether or not heart failure is present, ACE inhibitors or ARBs should be prescribed (see "Inhibition of the Renin-Angiotensin-Aldosterone System" earlier).

HEMODYNAMIC ASSESSMENT

Pump failure is now the primary cause of in-hospital death from STEMI. The extent of infarction correlates well with the degree of pump failure and with mortality, both early (within 10 days of infarction) and later. The most common clinical signs are pulmonary rales and S_3 and S_4 gallop sounds. Pulmonary congestion is also frequently seen on the chest roentgenogram. Elevated LV filling pressure and elevated pulmonary artery pressure are the characteristic hemodynamic findings, but these findings may result from a reduction of ventricular compliance (diastolic failure) and/or a reduction of stroke volume with secondary cardiac dilation (systolic failure) **(Chap. 252)**.

A classification originally proposed by Killip divides patients into four groups: class I, no signs of pulmonary or venous congestion; class II, moderate heart failure as evidenced by rales at the lung bases, S_3 gallop, tachypnea, or signs of failure of the right side of the heart, including venous and hepatic congestion; class III, severe heart failure, pulmonary edema; and class IV, shock with systolic pressure <90 mmHg and evidence of peripheral vasoconstriction, peripheral cyanosis, mental confusion, and oliguria. When this classification was established in 1967, the expected hospital mortality rate of patients in these classes was as follows: class I, 0–5%; class II, 10–20%; class III, 35–45%; and class IV, 85–95%. With advances in management, the mortality rate in each class has fallen, perhaps by as much as one-third to one-half.

Hemodynamic evidence of abnormal global LV function appears when contraction is seriously impaired in 20–25% of the left ventricle. Infarction of ≥40% of the left ventricle usually results in cardiogenic shock **(Chap. 298)**. Positioning of a balloon flotation (Swan-Ganz) catheter in the pulmonary artery permits monitoring of LV filling pressure; this technique is useful in patients who exhibit hypotension and/or clinical evidence of CHF. Cardiac output can also be determined with a pulmonary artery catheter. With the addition of intraarterial pressure monitoring, systemic vascular resistance can be calculated as a guide to adjusting vasopressor and vasodilator therapy. Some patients with STEMI have markedly elevated LV filling pressures (>22 mmHg) and normal cardiac indices (2.6–3.6 L/[min/m^2]), while others have relatively low LV filling pressures (<15 mmHg) and reduced cardiac indices. The former patients usually benefit from diuresis, while the latter may respond to volume expansion.

HYPOVOLEMIA

This is an easily corrected condition that may contribute to the hypotension and vascular collapse associated with STEMI in some patients. It may be secondary to previous diuretic use, to reduced fluid intake during the early stages of the illness, and/or to vomiting associated with pain or medications. Consequently, hypovolemia should be identified and corrected in patients with STEMI and hypotension before more vigorous forms of therapy are begun. Central venous pressure reflects RV rather than LV filling pressure and is an inadequate guide for adjustment of blood volume, because LV function is almost always affected much more adversely than RV function in patients with STEMI. The optimal LV filling or pulmonary artery wedge pressure may vary considerably among patients. Each patient's ideal level (generally ~20 mmHg) is reached by cautious fluid administration during careful monitoring of oxygenation and cardiac output. Eventually, the cardiac output plateaus, and further increases in LV filling pressure only increase congestive symptoms and decrease systemic oxygenation without raising arterial pressure.

TREATMENT

TREATMENT Congestive Heart Failure

The management of CHF in association with STEMI is similar to that of acute heart failure secondary to other forms of heart disease (avoidance of hypoxemia, diuresis, afterload reduction, inotropic support) **(Chap. 252)**, except that the benefits of digitalis administration to patients with STEMI are unimpressive. By contrast, diuretic agents are extremely effective, as they diminish pulmonary congestion in the presence of systolic and/or diastolic heart failure. LV filling pressure falls and orthopnea and dyspnea improve after the intravenous administration of furosemide or other loop diuretics. These drugs should be used with caution, however, as they can result in a massive diuresis with associated decreases in plasma volume, cardiac output, systemic blood pressure, and, hence, coronary perfusion. Nitrates in various forms may be used to decrease preload and congestive symptoms. Oral isosorbide dinitrate, topical nitroglycerin ointment, and intravenous nitroglycerin all have the advantage over a diuretic of lowering preload through venodilation without decreasing the total plasma volume. In addition, nitrates may improve ventricular compliance if ischemia is present, as ischemia causes an elevation of LV filling pressure. Vasodilators must be used with caution to prevent serious hypotension. As noted earlier, ACE inhibitors are an ideal class of drugs for management of ventricular dysfunction after STEMI, especially for the long term. (See "Inhibition of the Renin-Angiotensin-Aldosterone System" earlier.)

CARDIOGENIC SHOCK

Prompt reperfusion, efforts to reduce infarct size and treatment of ongoing ischemia and other complications of MI appear to have reduced the incidence of cardiogenic shock from 20 to about 7%. Only 10% of patients with this condition present with it on admission, while 90% develop it during hospitalization. Typically, patients who develop cardiogenic shock have severe multivessel coronary artery disease with evidence of "piecemeal" necrosis extending outward from the original infarct zone. **The evaluation and management of cardiogenic shock and severe power failure after STEMI are discussed in detail in Chap. 298.**

RIGHT VENTRICULAR INFARCTION

Approximately one-third of patients with inferior infarction demonstrate at least a minor degree of RV necrosis. An occasional patient with inferoposterior LV infarction also has extensive RV infarction, and rare patients present with infarction limited primarily to the RV. Clinically significant RV infarction causes signs of severe RV failure (jugular venous distention, Kussmaul's sign, hepatomegaly **[Chap. 234]**) with or without hypotension. ST-segment elevations of right-sided precordial ECG leads, particularly lead V_4R, are frequently present in the first 24 h in patients with RV infarction. Two-dimensional echocardiography is helpful in determining the degree of RV dysfunction. Catheterization of the right side of the heart often reveals a distinctive hemodynamic pattern resembling constrictive pericarditis (steep right atrial "y" descent and an early diastolic dip and plateau in RV waveforms) **(Chap. 265)**. Therapy consists of volume expansion to maintain adequate RV preload and efforts to improve LV performance with attendant reduction in pulmonary capillary wedge and pulmonary arterial pressures.

ARRHYTHMIAS

(See also Chaps. 239 and 241) The incidence of arrhythmias after STEMI is higher in patients seen early after the onset of symptoms. The mechanisms responsible for infarction-related arrhythmias include autonomic nervous system imbalance, electrolyte disturbances, ischemia, and slowed conduction in zones of ischemic myocardium. An arrhythmia can usually be managed successfully if trained personnel and appropriate equipment are available when it develops. Since most deaths from arrhythmia occur during the first few hours after infarction, the effectiveness of treatment relates directly to the speed with which patients come under medical observation. The prompt management of arrhythmias constitutes a significant advance in the treatment of STEMI.

Ventricular Premature Beats

Infrequent, sporadic ventricular premature depolarizations occur in almost all patients with STEMI and do not require therapy. Whereas in the past, frequent, multifocal, or early diastolic ventricular extrasystoles (so-called warning arrhythmias) were routinely treated with antiarrhythmic drugs to reduce the risk of development of ventricular tachycardia and ventricular fibrillation, pharmacologic therapy is now reserved for patients with sustained ventricular arrhythmias. Prophylactic antiarrhythmic therapy (either intravenous lidocaine early or oral agents later) is contraindicated for ventricular premature beats in the absence of clinically important ventricular tachyarrhythmias, because such therapy may actually increase the mortality rate. Beta-adrenoceptor blocking agents are effective in abolishing ventricular ectopic activity in patients with STEMI and in the prevention of ventricular fibrillation. As described earlier (see "Beta-Adrenoceptor Blockers"), they should be used routinely in patients without contraindications. In addition, hypokalemia and hypomagnesemia are risk factors for ventricular fibrillation in patients with STEMI; to reduce the risk, the serum potassium concentration should be adjusted to ~4.5 mmol/L and magnesium to about 2.0 mmol/L.

Ventricular Tachycardia and Fibrillation

Within the first 24 h of STEMI, ventricular tachycardia and fibrillation can occur without prior warning arrhythmias. The occurrence of ventricular fibrillation can be reduced by prophylactic administration of intravenous lidocaine. However, prophylactic use of lidocaine has not been shown to reduce overall mortality from STEMI. In fact, in addition to causing possible noncardiac complications, lidocaine may predispose to an excess risk of bradycardia and asystole. For these reasons, and with earlier treatment of active ischemia, more frequent use of beta-blocking agents, and the nearly universal success of electrical cardioversion or defibrillation, routine prophylactic antiarrhythmic drug therapy *is no longer recommended*.

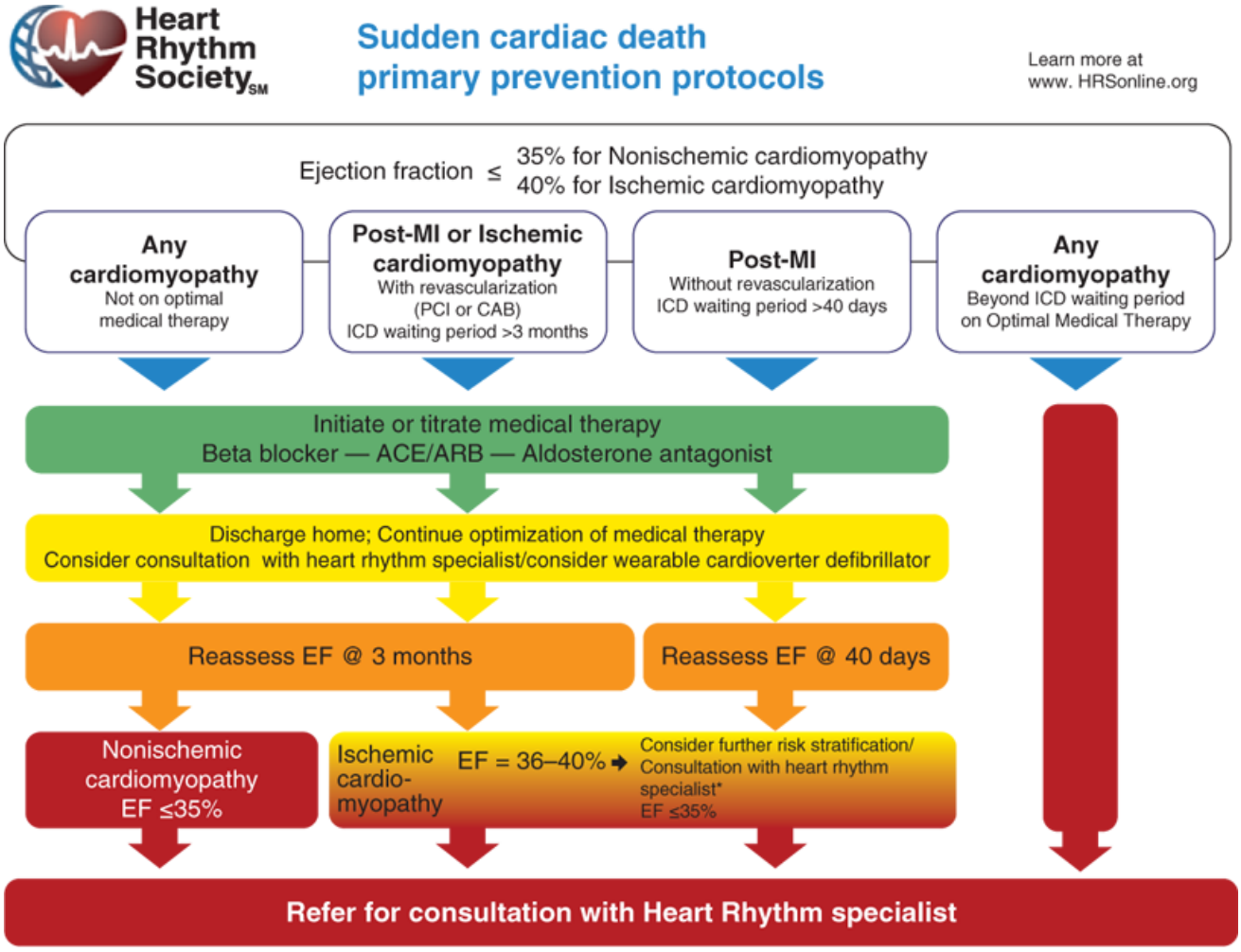
Sustained ventricular tachycardia that is well tolerated hemodynamically should be treated with an intravenous regimen of amiodarone (bolus of 150 mg over 10 min, followed by infusion of 1.0 mg/min for 6 h and then 0.5 mg/min). A less desirable but alternative regimen is procainamide (bolus of 15 mg/kg over 20–30 min; infusion of 1–4 mg/min). If ventricular tachycardia does not stop promptly, electroversion should be used **(Chap. 241)**. An unsynchronized discharge of 200–300 J (monophasic waveform; ~50% of these energies with biphasic waveforms) is used immediately in patients with ventricular fibrillation or when ventricular tachycardia causes hemodynamic deterioration. Ventricular tachycardia or fibrillation that is refractory to electroshock may be more responsive after the patient is treated with epinephrine (1 mg intravenously or 10 mL of a 1:10,000 solution via the intracardiac route) or amiodarone (a 75–150-mg bolus).

Ventricular arrhythmias, including the unusual form of ventricular tachycardia known as torsades des pointes **(Chaps. 247 and 249)**, may occur in patients with STEMI as a consequence of other concurrent problems (such as hypoxia, hypokalemia, or other electrolyte disturbances) or of the toxic effects of an agent being administered to the patient (such as digoxin or quinidine). A search for such secondary causes should always be undertaken.

Although the in-hospital mortality rate is increased, the long-term survival is excellent in patients who survive to hospital discharge after *primary* ventricular fibrillation; i.e., ventricular fibrillation that is a primary response to acute ischemia that occurs during the first 48 h and is not associated with predisposing factors such as CHF, shock, bundle branch block, or ventricular aneurysm. This result is in sharp contrast to the poor prognosis for patients who develop ventricular fibrillation *secondary* to severe pump failure. For patients who develop ventricular tachycardia or ventricular fibrillation late in their hospital course (i.e., after the first 48 h), the mortality rate is increased both in-hospital and during long-term follow-up. Such patients should be considered for electrophysiologic study and implantation of a cardioverter-defibrillator (ICD) **(Chap. 247)**. A more challenging issue is the prevention of sudden cardiac death from ventricular fibrillation late after STEMI in patients who have not exhibited sustained ventricular tachyarrhythmias during their index hospitalization. An algorithm for selection of patients who warrant prophylactic implantation of an ICD is shown in **Fig. 269-5**.

FIGURE 269-5

Algorithm for assessment of need for implantation of a cardioverter-defibrillator. The appropriate management is selected based on measurement of left ventricular ejection fraction, the timing following infarction, and whether revascularization has been performed. *(Reproduced from data at www.hrsonline.org.)*



* Buxton AE, Lee KL, Fisher JD, Josephson ME, Prystowsky EN, Hafley G.A randomized study of the prevention of sudden death in patients with coronary artery disease. Multicenter Unsustained Tachycardia Trial Investigators. N Engl J Med. December 16, 1999;341(25):1882–1890.

Recommended by SCA Prevention Protocols Working Group (Version 2; Revised; 9/10/2012; Review date: 9/10/2013) All rights reserved. Copyright ©2012 Heart Rhythm Society

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Accelerated Idioventricular Rhythm

Accelerated idioventricular rhythm (AIVR, “slow ventricular tachycardia”), a ventricular rhythm with a rate of 60–100 beats/min, often occurs transiently during fibrinolytic therapy at the time of reperfusion. For the most part, AIVR, whether it occurs in association with fibrinolytic therapy or spontaneously, is benign and does not presage the development of classic ventricular tachycardia. Most episodes of AIVR do not require treatment if the patient is monitored carefully, as degeneration into a more serious arrhythmia is rare.

Supraventricular Arrhythmias

Sinus tachycardia is the most common supraventricular arrhythmia. If it occurs secondary to another cause (such as anemia, fever, heart failure, or a metabolic derangement), the primary problem should be treated first. However, if it appears to be due to sympathetic overstimulation (e.g., as part of a hyperdynamic state), then treatment with a beta blocker is indicated. Other common arrhythmias in this group are atrial flutter and atrial fibrillation, which are often secondary to LV failure. Digoxin is usually the treatment of choice for supraventricular arrhythmias if heart failure is present. If heart failure is absent, beta blockers, verapamil, or diltiazem are suitable alternatives for controlling the ventricular rate, as they may also help to control ischemia. If the abnormal rhythm persists for >2 h with a ventricular rate >120 beats/min, or if tachycardia induces heart failure, shock, or ischemia (as manifested by recurrent pain or ECG changes), a synchronized electroshock (100–200 J monophasic waveform) should be used.

Accelerated junctional rhythms have diverse causes but may occur in patients with inferoposterior infarction. Digitalis excess must be ruled out. In some patients with severely compromised LV function, the loss of appropriately timed atrial systole results in a marked reduction of cardiac output. Right atrial or coronary sinus pacing is indicated in such instances.

Sinus Bradycardia

Treatment of sinus bradycardia is indicated if hemodynamic compromise results from the slow heart rate. Atropine is the most useful drug for increasing heart rate and should be given intravenously in doses of 0.5 mg initially. If the rate remains <50–60 beats/min, additional doses of 0.2 mg, up to a total of 2.0 mg, may be given. Persistent bradycardia (<40 beats/min) despite atropine may be treated with electrical pacing. Isoproterenol should be avoided.

Atrioventricular and Intraventricular Conduction Disturbances

(See also Chap. 239) Both the in-hospital mortality rate and the postdischarge mortality rate of patients who have complete atrioventricular (AV) block in association with anterior infarction are markedly higher than those of patients who develop AV block with inferior infarction. This difference is related to the fact that heart block in inferior infarction is commonly a result of increased vagal tone and/or the release of adenosine and therefore is transient. In anterior wall infarction, however, heart block is usually related to ischemic malfunction of the conduction system, which is commonly associated with extensive myocardial necrosis.

Temporary electrical pacing provides an effective means of increasing the heart rate of patients with bradycardia due to AV block. However, acceleration of the heart rate may have only a limited impact on prognosis in patients with anterior wall infarction and complete heart block in whom the large size of the infarct is the major factor determining outcome. It should be carried out if it improves hemodynamics. Pacing does appear to be beneficial in patients with inferoposterior infarction who have complete heart block associated with heart failure, hypotension, marked bradycardia, or significant ventricular ectopic activity. A subgroup of these patients, those with RV infarction, often respond poorly to ventricular pacing because of the loss of the atrial contribution to ventricular filling. In such patients, dual-chamber AV sequential pacing may be required.

External noninvasive pacing electrodes should be positioned in a “demand” mode for patients with sinus bradycardia (rate <50 beats/min) that is unresponsive to drug therapy, Mobitz II second-degree AV block, third-degree heart block, or bilateral bundle branch block (e.g., right bundle branch block plus left anterior fascicular block). Retrospective studies suggest that permanent pacing may reduce the long-term risk of sudden death due to bradyarrhythmias in the rare patient who develops combined persistent bifascicular and transient third-degree heart block during the acute phase of MI.

OTHER COMPLICATIONS

Recurrent Chest Discomfort

Because recurrent or persistent ischemia often heralds extension of the original infarct or reinfarction in a new myocardial zone and is associated with a near tripling of mortality after STEMI, patients with these symptoms should be referred for prompt coronary arteriography and mechanical revascularization. Administration of a fibrinolytic agent is an alternative to early mechanical revascularization.

Pericarditis

(See also Chap. 265) Pericardial friction rubs and/or pericardial pain are frequently encountered in patients with STEMI involving the epicardium. This complication can usually be managed with aspirin (650 mg four times daily). It is important to diagnose the chest pain of pericarditis accurately, because failure to recognize it may lead to the erroneous diagnosis of recurrent ischemic pain and/or infarct extension, with resulting inappropriate use of anticoagulants, nitrates, beta blockers, or coronary arteriography. When it occurs, complaints of pain radiating to either trapezius muscle is helpful, because such a pattern of discomfort is typical of pericarditis but rarely occurs with ischemic discomfort. Anticoagulants potentially could cause tamponade in the presence of acute pericarditis (as manifested by either pain or persistent rub) and therefore should not be used unless there is a compelling indication.

Thromboembolism

Clinically apparent thromboembolism complicates STEMI in ~10% of cases, but embolic lesions are found in 20% of patients in necropsy series, suggesting that thromboembolism is often clinically silent. Thromboembolism is considered to be an important contributing cause of death in 25% of patients with STEMI who die after admission to the hospital. Arterial emboli originate from LV mural thrombi, while most pulmonary emboli arise in the leg veins.

Thromboembolism typically occurs in association with large infarcts (especially anterior), CHF, and an LV thrombus detected by echocardiography. The incidence of arterial embolism from a clot originating in the ventricle at the site of an infarction is small but real. Two-dimensional echocardiography reveals LV thrombi in about one-third of patients with anterior wall infarction but in few patients with inferior or posterior infarction. Arterial embolism often presents as a major complication, such as hemiparesis when the cerebral circulation is involved or hypertension if the renal circulation is compromised. When a thrombus has been

clearly demonstrated by echocardiographic or other techniques or when a large area of regional wall motion abnormality is seen even in the absence of a detectable mural thrombus, systemic anticoagulation should be undertaken (in the absence of contraindications), as the incidence of embolic complications appears to be markedly lowered by such therapy. The appropriate duration of therapy is unknown, but 3–6 months is probably prudent.

Left Ventricular Aneurysm

The term *ventricular aneurysm* is usually used to describe *dyskinesis* or local expansile paradoxical wall motion. Normally functioning myocardial fibers must shorten more if stroke volume and cardiac output are to be maintained in patients with ventricular aneurysm; if they cannot, overall ventricular function is impaired. True aneurysms are composed of scar tissue and neither predispose to nor are associated with cardiac rupture.

The complications of LV aneurysm do not usually occur for weeks to months after STEMI; they include CHF, arterial embolism, and ventricular arrhythmias. Apical aneurysms are the most common and the most easily detected by clinical examination. The physical finding of greatest value is a double, diffuse, or displaced apical impulse. Ventricular aneurysms are readily detected by two-dimensional echocardiography, which may also reveal a mural thrombus in an aneurysm.

Rarely, myocardial rupture may be contained by a local area of pericardium, along with organizing thrombus and hematoma. Over time, this *pseudoaneurysm* enlarges, maintaining communication with the LV cavity through a narrow neck. Because a pseudoaneurysm often ruptures spontaneously, it should be surgically repaired if recognized.

POSTINFARCTION RISK STRATIFICATION AND MANAGEMENT

Many clinical and laboratory factors have been identified that are associated with an increase in cardiovascular risk after initial recovery from STEMI. Some of the most important factors include persistent ischemia (spontaneous or provoked), depressed LV ejection fraction (<40%), rales above the lung bases on physical examination or congestion on chest radiograph, and symptomatic ventricular arrhythmias. Other features associated with increased risk include a history of previous MI, age >75, diabetes mellitus, prolonged sinus tachycardia, hypotension, ST-segment changes at rest without angina (“silent ischemia”), an abnormal signal-averaged ECG, nonpatency of the infarct-related coronary artery (if angiography is undertaken), and persistent advanced heart block or a new intraventricular conduction abnormality on the ECG. Therapy must be individualized on the basis of the relative importance of the risk(s) present.

The goal of preventing reinfarction and death after recovery from STEMI has led to strategies to evaluate risk after infarction. In stable patients, submaximal exercise stress testing may be carried out before hospital discharge to detect residual ischemia and ventricular ectopy and to provide the patient with a guideline for exercise in the early recovery period. Alternatively, or in addition, a maximal (symptom-limited) exercise stress test may be carried out 4–6 weeks after infarction. Evaluation of LV function is usually warranted as well. Recognition of a depressed LV ejection fraction by echocardiography or radionuclide ventriculography identifies patients who should receive medications to inhibit the renin-angiotensin-aldosterone system. Patients in whom angina is induced at relatively low workloads, those who have a large reversible defect on perfusion imaging or a depressed ejection fraction, those with demonstrable ischemia, and those in whom exercise provokes symptomatic ventricular arrhythmias should be considered at high risk for recurrent MI or death from arrhythmia (Fig. 269-5). Cardiac catheterization with coronary angiography and/or invasive electrophysiologic evaluation is advised.

Exercise tests also aid in formulating an individualized exercise prescription, which can be much more vigorous in patients who tolerate exercise without any of the previously mentioned adverse signs. In addition, predischarge stress testing may provide an important psychological benefit, building the patient’s confidence by demonstrating a reasonable exercise tolerance.

In many hospitals, a cardiac rehabilitation program with progressive exercise is initiated in the hospital and continued after discharge. Ideally, such programs should include an educational component that informs patients about their disease and its risk factors.

The usual duration of hospitalization for an uncomplicated STEMI is about 3–5 days. The remainder of the convalescent phase may be accomplished at home. During the first 1–2 weeks, the patient should be encouraged to increase activity by walking about the house and outdoors in good weather. Normal sexual activity may be resumed during this period. After 2 weeks, the physician must regulate the patient’s activity on the basis of exercise tolerance. Most patients will be able to return to work within 2–4 weeks.

SECONDARY PREVENTION

Various secondary preventive measures are at least partly responsible for the improvement in the long-term mortality and morbidity rates after STEMI. Long-term treatment with an antiplatelet agent (usually aspirin) after STEMI is associated with a 25% reduction in the risk of recurrent infarction, stroke, or cardiovascular mortality (36 fewer events for every 1000 patients treated). An alternative antiplatelet agent that may be used for secondary prevention in patients intolerant of aspirin is clopidogrel (75 mg orally daily). ACE inhibitors or ARBs and, in appropriate patients, aldosterone antagonists should be used indefinitely by patients with clinically evident heart failure, a moderate decrease in global ejection fraction, or a large regional wall motion abnormality to prevent late ventricular remodeling and recurrent ischemic events.

The chronic routine use of oral beta-adrenoceptor blockers for at least 2 years after STEMI is supported by well-conducted, placebo-controlled trials.

Evidence suggests that warfarin lowers the risk of late mortality and the incidence of reinfarction after STEMI. Most physicians prescribe aspirin routinely for all patients without contraindications and add warfarin for patients at increased risk of embolism (see "Thromboembolism" earlier). Several studies suggest that in patients <75 years old a low dose of aspirin (75–81 mg/d) in combination with warfarin administered to achieve an international normalized ratio >2.0 is more effective than aspirin alone for preventing recurrent MI and embolic cerebrovascular accident. However, there is an increased risk of bleeding and a high rate of discontinuation of warfarin that has limited clinical acceptance of combination antithrombotic therapy. There is an increased risk of bleeding when warfarin is added to dual antiplatelet therapy **(see Chap. 267)**. However, patients who have had a stent implanted and have an indication for anticoagulation should receive dual antiplatelet therapies in combination with warfarin. Such patients should also receive a proton pump inhibitor to minimize the risk of gastrointestinal bleeding and should have regular monitoring of their hemoglobin levels and stool hematest while on combination antithrombotic therapy.

Finally, risk factors for *atherosclerosis* **(Chap. 232)** should be discussed with the patient and, when possible, favorably modified.

FURTHER READING

Al-Khatib SM et al: Defibrillators: Selecting the right device for the right patient. Circulation 134:1390, 2016. [PubMed: 27799257]
Claeys MJ: Is primary percutaneous coronary intervention still the superior reperfusion strategy? JAMA Intern Med 175:216, 2015. [PubMed: 25486596]
Hollander JE et al: State-of-the-art evaluation of emergency department patients presenting with potential acute coronary syndromes. Circulation 134:547, 2016. [PubMed: 27528647]
Kidambi A, Plein S: Risk stratification in acute myocardial infarction with multiparametric cardiac magnetic resonance imaging: Getting to the core of the matter. Eur Heart J 37:1060, 2016. [PubMed: 26486972]
Levine GN et al: 2015 ACC/AHA/SCAI focused update on primary percutaneous coronary intervention for patients with ST-elevation myocardial infarction: An update of the 2011 ACCF/AHA/SCAI guideline for percutaneous coronary intervention and the 2013 ACCF/AHA guideline for the management of ST-elevation myocardial infarction: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Society for Cardiovascular Angiography and Interventions. Circulation 133:1135, 2016. [PubMed: 26490017]
Mannsverk J et al: Trends in modifiable risk factors are associated with declining incidence of hospitalized and nonhospitalized acute coronary heart disease in a population. Circulation 133:7, 2016.
Montecucco F et al: Pathophysiology of ST-segment elevation myocardial infarction: Novel mechanisms and treatments. Eur Heart J 37:1268, 2016. [PubMed: 26543047]
Mozaffarian D et al: Heart disease and stroke statistics—2016 update: A report from the American Heart Association. Circulation 133:e3, 2016.
Sugiyama T et al: Differential time trends of outcomes and costs of care for acute myocardial infarction hospitalizations by ST elevation and type of intervention in the United States, 2001–2011. JAMA 4:e001445, 2015.

Chapter 270: Percutaneous Coronary Interventions and Other Interventional Procedures

David P. Faxon; Deepak L. Bhatt

INTRODUCTION

Percutaneous transluminal coronary angioplasty (PTCA) was first introduced by Andreas Gruentzig in 1977 as an alternative to coronary bypass surgery. The concept was initially demonstrated by Charles Dotter in 1964 in peripheral vessels. The development of a small inelastic balloon catheter by Gruentzig allowed expansion of the technique into smaller peripheral and coronary vessels. Initial coronary experience was limited to single-vessel coronary disease and discrete proximal lesions due to the technical limitations of the equipment. Advances in technology and greater operator experience allowed the procedure to grow rapidly with expanded use in patients with more complex lesions and multivessel disease. The introduction of coronary stents in 1994 was one of the major advances in the field. These devices reduced acute complications and reduced by half the significant problems of acute thrombosis and late restenosis (or recurrence of the stenosis). Further reductions in restenosis were achieved by the introduction of drug-eluting stents in 2003. These stents slowly release antiproliferative drugs directly into the plaque over a few months. Percutaneous coronary intervention (PCI) is the most common revascularization procedure in the United States and is performed more than twice as often as coronary artery bypass surgery: nearly 900,000 patients a year.

Interventional cardiology is a separate discipline in cardiology that requires a dedicated 1-year interventional cardiology fellowship following a 3-year general cardiology fellowship in order to obtain a separate board certification. The discipline has also expanded to include interventions for structural heart disease including treatment of congenital heart disease and valvular heart disease; it also includes interventions to treat peripheral vascular disease, including atherosclerotic and nonatherosclerotic lesions in the carotid, renal, aortic, and peripheral arterial and venous circulations.

TECHNIQUE

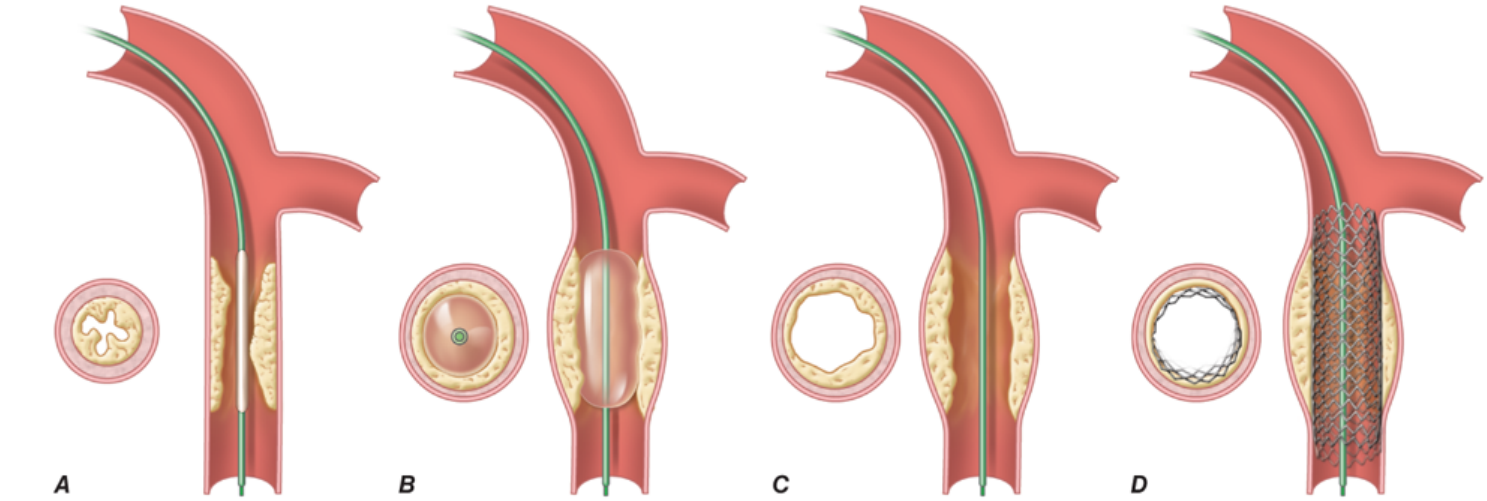
The initial procedure is performed in a similar manner as a diagnostic cardiac catheterization **(Chap. 237)**. Arterial access is obtained via the femoral or radial artery. To prevent thrombotic complications during the procedure, patients who are anticipated to need an angioplasty are given aspirin (325 mg) and may be given a platelet P2Y12 inhibitor such as clopidogrel (loading dose of 600 mg), prasugrel (loading dose of 60 mg), or ticagrelor (loading dose of 180 mg) before the procedure. Cangrelor, an IV P2Y12 inhibitor, is approved for use in patients who have not received an oral agent prior to the procedure. During the procedure, anticoagulation is achieved by administration of unfractionated heparin, enoxaparin (a low-molecular-weight heparin), or bivalirudin (a direct thrombin inhibitor). In patients with ST-elevation myocardial infarction, high-risk acute coronary syndrome, or a large thrombus in the coronary artery, an intravenous glycoprotein IIb/IIIa inhibitor (abciximab, tirofiban, or eptifibatide) may also be given, though cangrelor may be as effective with less bleeding risk.

Following placement of an introducing sheath into the artery, preformed guiding catheters are used to cannulate selectively the origins of the coronary arteries. Through the guiding catheter, a flexible, steerable guidewire is negotiated down the coronary artery lumen using fluoroscopic guidance; it is then advanced through the stenosis and into the vessel beyond. This guidewire then serves as a “rail” over which angioplasty balloons, stents, or other therapeutic devices can be advanced to enlarge the narrowed segment of coronary artery. The artery is usually dilated with a balloon catheter followed by placement of a stent. The catheters and introducing sheath are removed and the artery manually held or in the case of radial access use of an inflatable cuff. One of several femoral arterial closure devices can also be used to achieve hemostasis. Because PCI is performed under local anesthesia and mild sedation, it requires only a short (1-day) hospitalization or less.

Angioplasty works by stretching the artery and displacing the plaque away from the lumen, enlarging the entire vessel **(Figs. 270-1 and 270-2)**. The procedure rarely results in embolization of atherosclerotic material. Owing to inelastic elements in the plaque, the stretching of the vessel by the balloon results in small localized dissections that can protrude into the lumen and be a nidus for acute thrombus formation. If the dissections are severe, then they can obstruct the lumen or induce a thrombotic occlusion of the artery (acute closure). Stents have largely prevented this complication by holding the dissection flaps up against the vessel wall (Fig. 270-1).

FIGURE 270-1

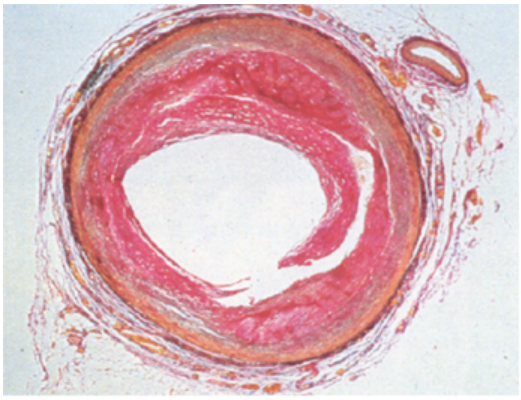
Schematic diagram of the primary mechanisms of balloon angioplasty and stenting. ***A.*** A balloon angioplasty catheter is positioned into the stenosis over a guidewire under fluoroscopic guidance. ***B.*** The balloon is inflated, temporarily occluding the vessel. ***C.*** The lumen is enlarged primarily by stretching the vessel, often resulting in small dissections in the neointima. ***D.*** A stent mounted on a deflated balloon is placed into the lesion and pressed against the vessel wall with balloon inflation (not shown). The balloon is deflated and removed, leaving the stent permanently against the wall acting as a scaffold to hold the dissections against the wall and prevent vessel recoil. *(Adapted from EJ Topol: Textbook of Cardiovascular Medicine, 2nd ed. Philadelphia, Lippincott Williams & Wilkins, 2002.)*



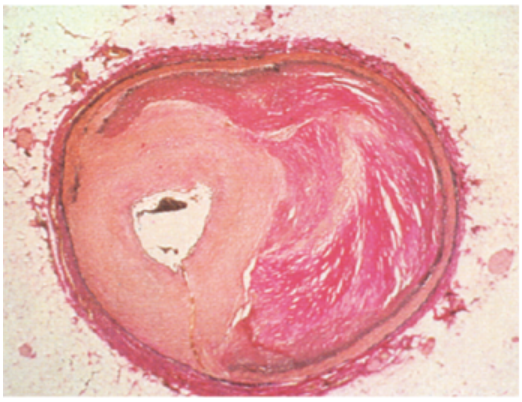
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 270-2

Pathology of acute effects of balloon angioplasty with intimal dissection and vessel stretching (***A***) and an example of neointimal hyperplasia and restenosis showing renarrowing of the vessel (***B***). *(Panel A from M Ueda et al: Eur Heart J 12:937, 1991; with permission. Panel B from CE Essed et al: Br Heart J 49:393, 1983; with permission.)*



A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Stents are currently used in >90% of coronary angioplasty procedures. Stents are wire meshes (usually made of stainless steel or other metals, such as cobalt chromium or nitinol) that are compressed over a deflated angioplasty balloon. When the balloon is inflated, the stent is enlarged to approximate the “normal” vessel lumen. The balloon is then deflated and removed, leaving the stent behind to provide a permanent scaffold in the artery. Owing to the design of the struts, these devices are flexible, allowing their passage through

diseased and tortuous coronary vessels. Stents are rigid enough to prevent elastic recoil of the vessel and have dramatically improved the success and safety of the procedure as a result.

Drug-eluting stents further enhanced the efficacy of PCI. An anti-proliferative agent is attached to the metal stent by use of a thin polymer coating. The antiproliferative drug elutes from the stent over a 1- to 3-month period or longer after implantation. Drug-eluting stents have been shown to reduce clinical restenosis by 50%, so that in uncomplicated lesions symptomatic restenosis occurs in 5–10% of patients. Not surprisingly, this led to the rapid acceptance of these devices; currently 80–90% of all stents implanted are drug-eluting. The first-generation devices were coated with either sirolimus or paclitaxel. Second-generation drug-eluting stents use newer agents such as everolimus, biolimus, and zotarolimus. These second-generation drug-eluting stents appear to be more effective with fewer complications, such as early or late stent thrombosis, than the first-generation devices and, therefore, have replaced the first-generation stents. Biodegradable polymers that are used to attach the drugs to the stents may be theoretically superior to permanent polymers in preventing late stent thrombosis. In addition, the everolimus-eluting biodegradable vascular scaffold (BVS) stent has been shown to be reasonably safe with gradual degradation over several years with improvement in vessel function. Additional bioresorbable stents are under investigation. Drug-coated balloons are covered with an antiproliferative drug that can also reduce restenosis, and are used primarily to treat in-stent restenosis.

Other interventional devices include atherectomy devices and thrombectomy catheters. These devices are designed to remove atherosclerotic plaque or thrombus and are used in conjunction with balloon dilatation and stent placement. Rotational atherectomy is the most commonly used adjunctive device and is modeled after a dentist's drill, with small round burrs of 1.25–2.5 mm at the tip of a flexible wire shaft. The burr is passed over the guidewire up to the stenosis and drills away atherosclerotic material. Because the atherosclerotic particles are ≤25 μm, they pass through the coronary microcirculation and rarely cause problems. The device is particularly useful in heavily calcified plaques that are resistant to balloon dilatation. Given the current advances in stents, rotational atherectomy is infrequently used. Orbital atherectomy is a newer approach to calcified lesions that also relies on a spinning burr. Directional atherectomy catheters that slice off the plaque and remove it are not used in the coronaries any longer but are used in peripheral artery disease. In acute ST-elevation myocardial infarction, specialized catheters without a balloon are used to aspirate thrombus in order to prevent embolization down the coronary vessel and to improve blood flow before angioplasty and stent placement. Current studies suggest that manual catheter thrombus aspiration should not be used routinely, but in certain cases of a large thrombus burden, can improve blood flow in primary PCI.

PCI of degenerated saphenous vein graft lesions has been associated with a significant incidence of distal embolization of atherosclerotic material, unlike PCI of native vessel disease. A number of distal protection devices have been shown to significantly reduce embolization and myocardial infarction in this setting. Most devices work by using a collapsible wire filter at the end of a guidewire that is expanded in the distal vessel before PCI. If atherosclerotic debris is dislodged, the basket captures the material, and at the end of the PCI, the basket is pulled into a delivery catheter and the debris safely removed from the patient.

SUCCESS AND COMPLICATIONS

A successful procedure (angiographic success), defined as a reduction of the stenosis to less than a 20% diameter narrowing, occurs in 95–99% of patients. Lower success rates are seen in patients with tortuous, small, or calcified vessels or chronic total occlusions. Chronic total occlusions have the lowest success rates and their recanalization is significantly better if the occlusion is recent (within 3 months) or there are favorable anatomic features. Improvements in equipment and complex antegrade and retrograde techniques have increased the success rates of recanalization of chronic total occlusions to 70–80%.

Serious complications are rare but include a mortality rate of 0.1–0.3% for elective cases, a large myocardial infarction in <3%, and stroke in <0.1%. Patients who are elderly (>65 years), undergoing an emergent or urgent procedure, have chronic kidney disease, present with an ST-segment elevation myocardial infarction (STEMI), or are in shock have significantly higher risk. Scoring systems can help to estimate the risk of the procedure. Myocardial infarction during PCI can occur for multiple reasons including an acute occluding thrombus, severe coronary dissection, embolization of thrombus or atherosclerotic material, or closure of a side branch vessel at the site of angioplasty or stent placement. Most myocardial infarctions are small and only detected by a rise in the creatine phosphokinase (CPK) or troponin level after the procedure. Only those with significant enzyme elevations (more than five times the upper limit of normal) are associated with a less favorable long-term outcome. Coronary stents have largely prevented coronary dissections due to the scaffolding effect of the stent.

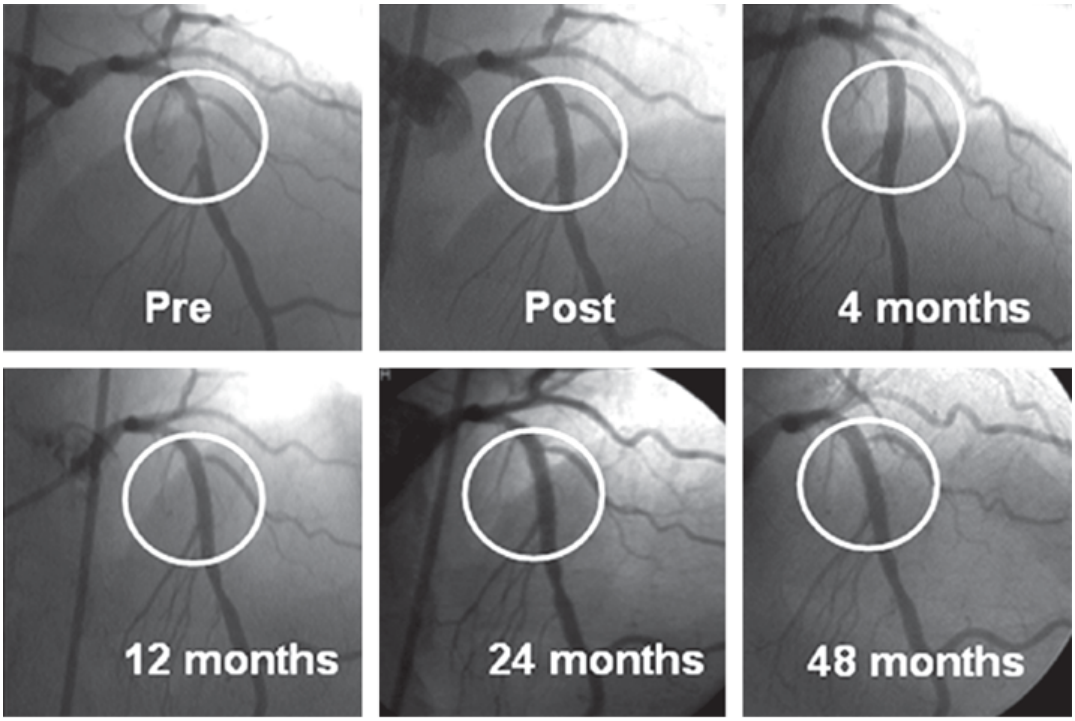
All types of stents are prone to stent thrombosis (1–3%), either acute (<24 h) or subacute (1–30 days), which can be ameliorated by greater attention to full initial stent deployment and the use of dual antiplatelet therapy (DAPT) (aspirin, plus a platelet $P2Y_{12}$ receptor blocker [clopidogrel, prasugrel, or ticagrelor]). Late (30 days–1 year) and very late stent thromboses (>1 year) occur very infrequently with stents but are slightly more common with first-generation drug-eluting stents, necessitating DAPT for up to 1 year or longer. Use of the second-generation stents is associated with lower rates of late and very late stent thromboses, and shorter durations of DAPT (6 months) are recommended. Premature discontinuation of DAPT, particularly in the first month after implantation, is associated with a significantly increased risk for stent thrombosis (three- to ninefold greater). Stent thrombosis results in death in 10–20% and myocardial infarction in 30–70% of patients. Elective surgery that requires discontinuation of antiplatelet therapy after drug-eluting stent implantation should be postponed until after 3 months and preferably after 6 months, if at all possible.

Restenosis, or renarrowing of the dilated coronary stenosis, is the most common complication of angioplasty and occurs in 20–50% of patients with balloon angioplasty alone, 10–30% of patients with bare metal stents, and 5–15% of patients with drug-eluting stents within the first year. The fact that stent placement provides a larger acute luminal area than balloon angioplasty alone reduces the incidence of subsequent restenosis. Drug-eluting stents further reduce restenosis through a reduction in excessive neointimal growth over the stent. If restenosis does not occur, the long-term outcome is excellent **(Fig. 270-3)**. Clinical restenosis is recognized by recurrence of angina or symptoms within 12 months of the procedure. Less frequently, patients with restenosis can present with non-ST-segment elevation myocardial infarction (NSTEMI) (10%) or STEMI (2%) as well. Very late stent thrombosis and restenosis after 1 year is more likely to be due to neoatherosclerosis than intimal hyperplasia seen within the first year. Clinical restenosis requires confirmation of a significant stenosis at the site of the prior PCI. *Target lesion revascularization* (TLR) or *target vessel revascularization* (TVR) is defined as

angiographic restenosis with repeat PCI or coronary artery bypass grafting (CABG). By angiography, the incidence of restenosis is significantly higher than clinical restenosis (TLR or TVR) because many patients have mild restenosis that does not result in a recurrence of symptoms. The management of clinical restenosis is usually to repeat the PCI with balloon dilatation and placement of another drug-eluting stent. Once a patient has had restenosis, the risk of a second restenosis is further increased. The risk factors for restenosis are diabetes, myocardial infarction, long lesions, small-diameter vessels, and suboptimal initial PCI result.

FIGURE 270-3

Long-term results from one of the first patients to receive a sirolimus-eluting stent from early Sao Paulo experience. *(From GW Stone, in D Baim [ed]: Cardiac Catheterization, Angiography and Intervention, 7th ed. Philadelphia, Lippincott Williams & Wilkins, 2006; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

INDICATIONS

The American College of Cardiology (ACC)/American Heart Association (AHA) guidelines extensively review the indications for PCI in patients with stable angina, unstable angina, NSTEMI, and STEMI and should be referred to for a comprehensive discussion of the indications. Briefly, the two principal indications for coronary revascularization in patients with *chronic stable angina* **(Chap. 267)** are (1) to improve angina symptoms in patients who remain symptomatic despite adequate medical therapy and (2) to reduce mortality rates in patients with severe and extensive coronary disease. In patients with stable angina who are well controlled on medical therapy, studies such as the Clinical Outcomes Utilizing Revascularization and Aggressive Drug Evaluation (COURAGE) and Bypass Angioplasty Revascularization Investigation 2 Diabetes (BARI 2D) trials have shown that initial revascularization does not lead to better outcomes (death or MI) and can be safely delayed until symptoms worsen or evidence of severe ischemia on noninvasive testing occurs. When revascularization is indicated, the choice of PCI or CABG depends on a number of clinical and anatomic factors. The Synergy between Percutaneous Coronary Intervention with Taxus and Cardiac Surgery (SYNTAX) trial compared PCI with the paclitaxel drug-eluting stent to CABG in 1800 patients with three-vessel coronary disease or left main disease. The study found no difference in death or myocardial infarction at 1 year, but repeat revascularization was significantly higher in the stent-treated group (13.5 vs 5.9%), while stroke was significantly higher in the surgical group (2.2 vs 0.6%). The primary endpoint of death, myocardial infarction, stroke, or revascularization was significantly better with CABG, particularly in those with the most extensive coronary artery disease such as three-vessel disease. The 5-year results confirm these findings. The Future Revascularization Evaluation in Patients With Diabetes Mellitus: Optimal Management of Multivessel Disease (FREEDOM) trial randomized 1900 patients with diabetes and multivessel disease and found a significantly lower primary endpoint of death, myocardial infarction, or stroke with CABG than PCI. Recent trials comparing PCI with CABG have shown similar outcomes for those with less extensive disease, but a better outcome when the coronary disease is severe and extensive. These studies support CABG for those with the most severe left main and three-vessel disease or those with diabetes. Lesser degrees of multivessel disease in patients with or without diabetes have an equal outcome with PCI, including left main disease with favorable angiographic characteristics.

The choice of PCI versus CABG is also related to the anticipated procedural success and complications of PCI and the risks of CABG. For PCI, the characteristics of the coronary anatomy are critically important. The location of the lesion in the vessel (proximal or distal), the degree of tortuosity, and the size of the vessel are considered. In addition, the lesion characteristics, including the degree of the stenosis, the presence of calcium, lesion length, and presence of thrombus, are assessed. The most common reason to decide not to do

PCI is that the lesion(s) felt to be responsible for the patient's symptoms are not treatable. This is most commonly due to the presence of a chronic total occlusion (>3 months in duration) with unfavorable characteristics. A lesion classification to characterize the likelihood of success or failure of PCI has been developed by the ACC/AHA. Lesions with the highest success are called type A lesions (such as proximal non-calcified subtotal lesions), and those with the lowest success or highest complication rate are type C lesions (such as chronic total occlusions). Intermediate lesions are classified as type B1 or B2 depending on the number of unfavorable characteristics. Approximately 25–30% of patients will not be candidates for PCI due to unfavorable anatomy, whereas only 5% of CABG patients will not be candidates for surgery due to coronary anatomy. The primary reason for being considered inoperable with CABG is the presence of severe comorbidities such as advanced age, frailty, severe chronic obstructive pulmonary disease (COPD), poor left ventricular function, or lack of suitable surgical conduits or poor distal targets for bypass.

Another consideration in choosing a revascularization strategy is the degree of revascularization. In patients with multivessel disease, bypass grafts can usually be placed to all vessels >2 mm with significant stenosis, whereas PCI may be able to treat only some of the lesions due to the presence of unfavorable anatomy. Assessment of the significance of intermediate lesions using fractional flow reserve (FFR) **(Chap. 237)** can assist in determining which lesions should be revascularized. The Fractional Flow Reserve versus Angiography for Multivessel Evaluation (FAME) trial showed a 30% reduction in adverse events when revascularization by PCI was restricted to those lesions that were hemodynamically significant (FFR ≤0.80) rather than when guided by angiography alone. Thus, complete revascularization of all functionally significant lesions should be favored and considered when choosing the optimal revascularization strategy. Given the multiple factors that need to be considered in choosing the best revascularization for an individual patient with multivessel disease, it is optimal to have a discussion among the cardiac surgeon, interventional cardiologist, and the physicians caring for the patient (so-called Heart Team) to weigh the choices properly.

Patients with acute coronary syndrome are at excess risk of short- and long-term mortality. Randomized clinical trials have shown that PCI is superior to intensive medical therapy in reducing mortality and myocardial infarction, with the benefit largely confined to those patients who are high risk. High-risk patients are defined as those with any one of the following: refractory ischemia, recurrent angina, positive cardiac-specific enzymes, new ST-segment depression, low ejection fraction, severe arrhythmias, or a recent PCI or CABG. PCI is preferred over surgical therapy in most high-risk patients with acute coronary syndromes unless they have severe multivessel disease or the culprit lesion responsible for the unstable presentation cannot be adequately treated. In STEMI, thrombolysis or PCI (primary PCI) are effective methods to restore coronary blood flow and salvage myocardium within the first 12 h after onset of chest pain. Because PCI is more effective in restoring flow than thrombolysis, it is preferred if readily available within 90 min of presentation to the hospital. PCI is also performed following thrombolysis to facilitate adequate reperfusion or as a rescue procedure in those who do not achieve reperfusion from thrombolysis, cannot be rapidly transferred to a hospital that can perform primary PCI, or in those who develop cardiogenic shock.

OTHER INTERVENTIONAL TECHNIQUES

STRUCTURAL HEART DISEASE

Interventional treatment for structural heart disease (adult congenital heart disease and valvular heart disease) is a significant and growing component of the field of interventional cardiology.

The most common adult congenital lesion to be treated with percutaneous techniques is closure of atrial septal defects **(Chap. 264)**. The procedure is done as in a diagnostic right heart catheterization with the passage of a catheter up the femoral vein into the right atrium. With echo and fluoroscopic guidance, the size and location of the defect can be accurately defined, and closure is accomplished using one of several approved devices. All devices use a left atrial and right atrial wire mesh or covered disk that are pulled together to capture the atrial septum around the defect and seal it off. The Amplatzer Septal Occluder device (AGA Medical, Minneapolis, Minnesota) is the most commonly used in the United States. The success rate in selected patients is 85–95%, and the device complications are rare and include device embolization, infection, or erosion. Closure of patent foramen ovale (PFO) is done in a similar way. PFO closure may be considered in patients who have had recurrent paradoxical stroke or transient ischemic attack (TIA) despite adequate medical therapy including anticoagulation or antiplatelet therapy. The CLOSURE I trial randomized 909 patients with cryptogenic stroke or TIA who had a PFO. Closure did not reduce the primary endpoint of death within 30 days or death following a neurologic cause during 2 years of follow-up or stroke/TIA within 2 years. Other trials have confirmed these findings. However, the 10-year follow up from the Randomized Evaluation of Recurrent Stroke Comparing PFO Closure to Established Current Standard of Care Treatment (RESPECT) trial did suggest a benefit of closure in reducing the risk for recurrent cryptogenic stroke. The use in the treatment of migraine is not supported by the current data.

Similar devices can also be used to close patent ductus arteriosus and ventricular septal defects. Other congenital diseases that can be treated percutaneously include coarctation of the aorta, pulmonic stenosis, peripheral pulmonary stenosis, and other abnormal communications between the cardiac chambers or vessels.

The treatment of valvular heart disease is the most rapidly growing area in interventional cardiology. Until recently, the only available techniques were balloon valvuloplasty for the treatment of aortic, mitral, or pulmonic stenosis **(Chap. 256)**. Mitral valvuloplasty is the preferred treatment for symptomatic patients with rheumatic mitral stenosis who have favorable anatomy. The outcome in these patients is equal to that of surgical commissurotomy. The success is highly related to the echocardiographic appearance of the valve. The most favorable setting is commissural fusion without calcification or subchordal fusion and the absence of significant mitral regurgitation. Access is obtained from the femoral vein using a transseptal technique in which a long metal catheter with a needle tip is advanced from the femoral vein through the right atrium and atrial septum at the level of the foramen ovale into the left atrium. A guidewire is advanced into the left ventricle, and a balloon-dilatation catheter is negotiated across the mitral valve and inflated to a predetermined size to enlarge the valve. The most commonly used dilatation catheter is the Inoue balloon. The technique splits the commissural fusion and commonly results in a doubling of the mitral valve area. The success of the procedure in favorable anatomy is 95% and severe complications are rare (1–2%). The most common complications are tamponade due to puncture into the pericardium during the transseptal puncture or the creation of severe mitral regurgitation due to damage to the valve leaflets.

Severe mitral regurgitation can be treated percutaneously using the MitraClip (Abbott, Abbott Park, Illinois) device. The procedure involves the passage of a catheter into the left atrium using the transseptal technique. A special catheter with a metallic clip on the end is passed through the mitral valve and retracted to catch and clip together the mid portion of the anterior and posterior mitral valve leaflets. The clip creates a double opening in the mitral valve and thereby reduces mitral regurgitation similar to the surgical Alfieri repair. In the Endovascular Valve Edge-to-Edge Repair Study (EVEREST II) trial, the device was less effective than surgical repair or replacement but was shown to be safe. Subsequent trials have shown it to be reasonably effective for patients who are not good candidates for surgical repair, particularly when the regurgitation is due to functional causes.

Severe aortic stenosis can be treated with balloon valvuloplasty as well. In this setting, the valvuloplasty balloon catheter is placed retrograde across the aortic valve from the femoral artery and briefly inflated to stretch open the valve. The success is much less favorable, with only 50% achieving an aortic valve area of >1 cm^2 and a restenosis rate of 25–50% after 6–12 months. This poor success rate has limited its use to patients who are not surgical candidates or as a bridge to surgery or transcatheter aortic valve replacement (TAVR). In this setting, the intermediate-term mortality rate of the procedure is high (10%). Repeat aortic valvuloplasty as a treatment for aortic valve restenosis has been reported.

Percutaneous TAVR has been shown to be an effective treatment for intermediate, high-risk, and inoperable patients with aortic stenosis. Currently, two valve models, the Edwards SAPIEN valve (Edwards Lifescience, Irvine, California) and the CoreValve ReValving system (Medtronic, Minneapolis, Minnesota) are available. In more than 50,000 cases worldwide since 2002, follow-up shows no evidence of restenosis or severe prosthetic valve dysfunction in the midterm, but long-term outcomes of 5–10 years are not yet available. The CoreValve is self-expanding, while the Edwards valve is balloon expanded. The cannulas are large (14–22 French), and retrograde access via the femoral artery is most commonly chosen, if possible. In patients with peripheral artery disease, access via the subclavian artery, aorta, or transapically through a surgical incision can be used. Following balloon valvuloplasty, the valve is positioned across the valve and deployed with post-deployment balloon inflation to ensure full contact with the aortic annulus. The success rate is >90%, and the 30-day mortality rate is 2–15% based on preoperative risk. The Placement of Aortic Transcatheter Valve (PARTNER) randomized trial of the Edwards valve showed a 55% reduction in 1-year mortality and major adverse events in the extreme-risk group randomized to TAVR compared with medical therapy. In separate randomized trials, moderate- and high-risk patients had similar outcomes to surgical valve replacement at 1 year. As a result, this valve is approved for both intermediate-risk, high-risk, and extreme-risk patients with severe aortic stenosis.

Aortic and mitral bioprosthetic valve degeneration can be treated with repeat surgery or, in high-risk patients, with a valve-in-valve procedure where a percutaneous valve is placed inside of the prior surgical valve. It has been shown to be effective for aortic and mitral valves.

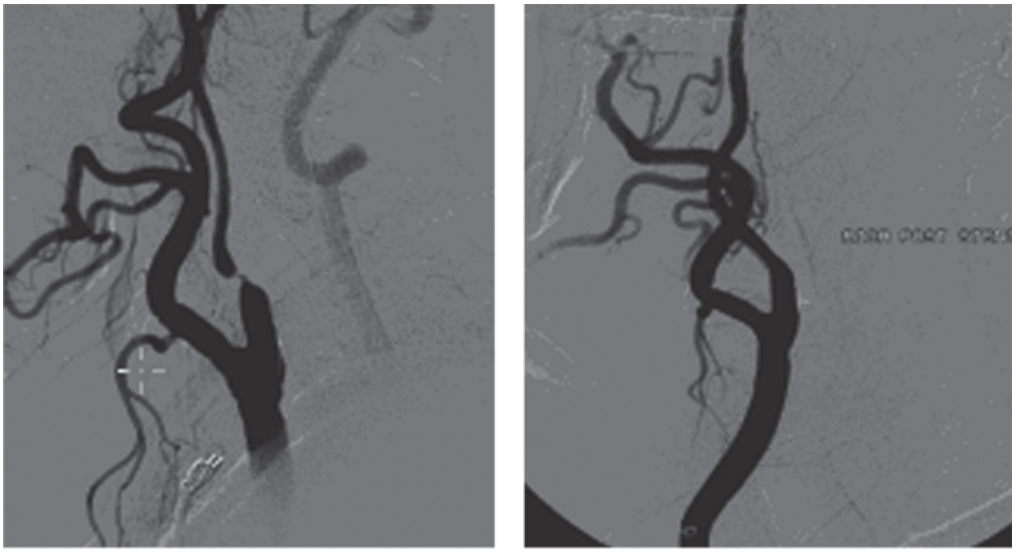
Pulmonic stenosis can also be effectively treated with balloon valvuloplasty and percutaneously replaced with the Melody valve (Medtronic). Tricuspid valve interventions remain experimental.

PERIPHERAL ARTERY INTERVENTIONS

The use of percutaneous interventions to treat symptomatic patients with arterial obstruction in the carotid, renal, aortic, and peripheral vessels is an effective alternative to vascular surgery. Randomized clinical trial data support the use of carotid stenting in patients at high risk of complications from carotid endarterectomy **(Fig. 270-4)**. Recent trials suggest similar outcomes with carotid stenting and carotid endarterectomy in patients at average risk, although depending on the patient's risk for periprocedural stroke or myocardial infarction, one procedure may be preferred over the other. The success rate of peripheral artery interventional procedures has been improving, including treatment for long segments of occlusive disease historically treated by peripheral bypass surgery **(Fig. 270-5)**. The use of drug-coated balloons and drug-eluting stents has shown to reduce restenosis when compared with balloon angioplasty alone. Peripheral intervention is increasingly part of the training of an interventional cardiologist, and most programs now require an additional year of training after the interventional cardiology training year. The techniques and outcomes are described in detail in the chapter on peripheral vascular disease **(Chap. 275)**.

FIGURE 270-4

A. An example of a high-risk patient who requires carotid revascularization, but who is not a candidate for carotid endarterectomy. **B.** Carotid artery stenting resulted in an excellent angiographic result. *(From M Belkin, DL Bhatt: Circulation 119:2302, 2009; with permission.)*

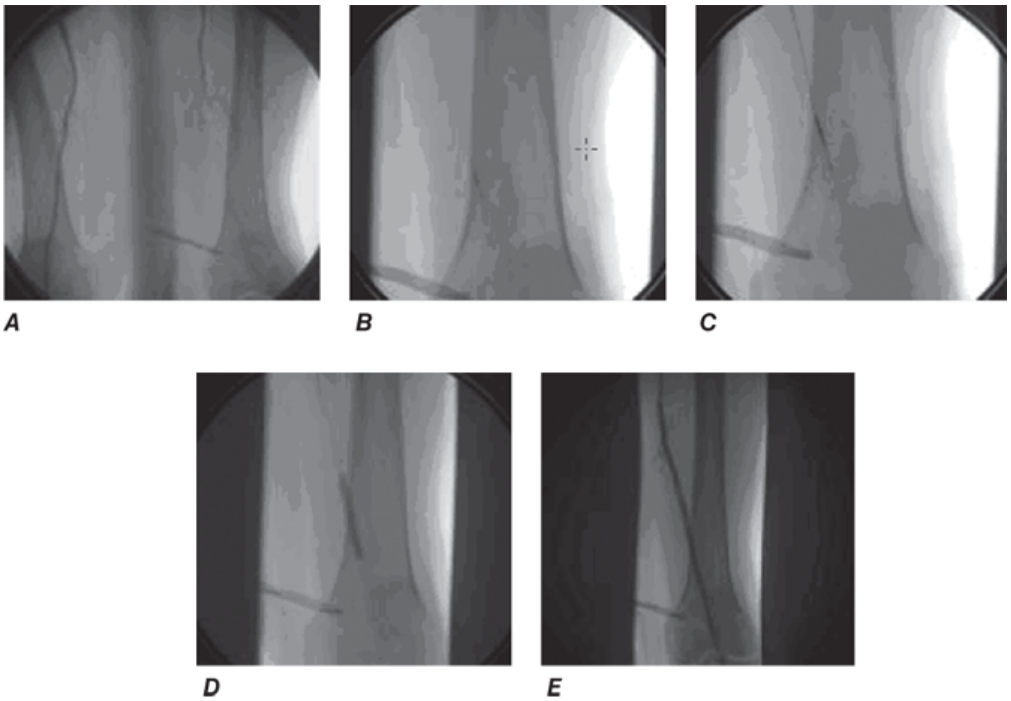


A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 270-5

Peripheral interventional procedures have become highly effective at treating anatomic lesions previously amenable only to bypass surgery. ***A.*** Complete occlusion of the left superficial femoral artery. ***B.*** Wire and catheter advanced into subintimal space. ***C.*** Intravascular ultrasound positioned in the subintimal space to guide retrograde wire placement through the occluded vessel. ***D.*** Balloon dilation of the occlusion. ***E.*** Stent placement with excellent angiographic result. *(From A Al Mahameed, DL Bhatt: Cleve Clin J Med 73:S45, 2006; with permission. Copyright © 2006 Cleveland Clinic Foundation. All rights reserved.)*



A B C

D E

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CIRCULATORY SUPPORT TECHNIQUES

The use of circulatory support techniques is indicated for the management of patients with shock or hemodynamic instability and occasionally is needed in order to safely perform PCI on hemodynamically unstable patients. It also can be useful in helping to stabilize patients before surgical interventions. The most commonly used device is the percutaneous intraaortic balloon pump developed in the early 1960s. A 7- to 10-French, 25- to 50-mL balloon catheter is placed retrograde from the femoral artery into the descending aorta between the aortic arch and the abdominal aortic bifurcation. It is connected to a helium gas inflation system that synchronizes the inflation to coincide with early diastole with deflation by mid-diastole. As a result, it increases early diastolic pressure, lowers systolic

pressure, and lowers late diastolic pressure through displacement of blood from the descending aorta (counterpulsation). This results in an increase in coronary blood flow and a decrease in afterload. It is contraindicated in patients with aortic regurgitation, aortic dissection, or severe peripheral artery disease. The major complications are vascular and thrombotic. Intravenous heparin is given in order to reduce thrombotic complications.

Another useful support device is the Impella (Abiomed, Danvers, Massachusetts). The Impella catheter is placed percutaneously from the femoral artery into the left ventricle. The catheter has a small microaxial pump at its tip that can pump up to 2.5–5 L/min from the left ventricle to the aorta. The smaller devices can be placed percutaneously but the larger devices need surgical access. Other support devices include TandemHeart (CardiacAssist, Pittsburgh, Pennsylvania), which involves placement of a large 21-French catheter from the femoral vein through the right atrium into the left atrium using the transseptal technique and a catheter in the femoral artery. A centrifugal pump can deliver 5 L of blood per minute. It may be useful in patients in shock or with STEMI or very-high-risk PCI. Patients can also be placed on peripheral extracorporeal membrane oxygenation (ECMO) using large cannulas placed in the femoral artery and vein. This technique can be performed in the catheterization laboratory and is useful for support of patients with acute respiratory failure or cardiac failure.

INTERVENTIONS FOR PULMONARY EMBOLISM

The treatment of deep vein thrombosis is intravenous anticoagulation, with placement of an inferior vena cava filter if recurrent pulmonary emboli (PE) occur or anticoagulation is not possible. Postphlebitic syndrome is a serious condition due to chronic venous obstruction that can lead to chronic leg edema and venous ulcers. Preliminary studies suggest that mechanical treatments may have a role in treatment, and a large trial is ongoing.

PE should be treated with fibrinolytic agents if massive and in some cases if submassive. Surgical pulmonary embolectomy is an option for the treatment of massive PE with hemodynamic instability in patients who have contraindications for systemic fibrinolysis or those in whom it has failed. Catheter-based therapies for submassive and massive PEs are still evolving, but studies have shown promise. The techniques employed include the use of aspiration of the clot with a large catheter (10 French), intraclot infusion of a thrombolytic agent followed by aspiration, ultrasound-assisted catheter-directed thrombolysis, and use of rheolytic thrombectomy. Success for these techniques has been reported to be 80–90%, with major complications occurring in 2–4% of patients.

INTERVENTIONS FOR REFRACTORY HYPERTENSION

The recent recognition of the importance of the renal sympathetic nerves in modulating blood pressure has led to a technique to selectively denervate renal sympathetic nerves in patients with refractory hypertension. The procedure involves applying low-power radiofrequency treatment via a catheter along the length of both renal arteries. In the randomized Symplicity HTN-2 trial, renal denervation significantly reduced blood pressure compared with medical therapy. The Symplicity device (Medtronic, Minnesota) is approved in Europe, though the randomized and blinded U.S. Symplicity HTN-3 trial showed no effect. Further optimization of the technique is needed with evidence from randomized trials of efficacy before approval in the United States.

CONCLUSION

Interventional cardiology continues to expand its borders. Treatment for coronary artery disease, including complex anatomic subsets, continues to advance. Technological advances such as drug-eluting stents, now already in their second generation are improving the results of PCI. PCI is the treatment of choice for patients with acute coronary syndromes. For patients with stable coronary disease, PCI is effective in symptom alleviation. Use of a Heart Team is the best way to make decisions concerning which revascularization—PCI or CABG—is best for an individual patient. Treatment of peripheral and cerebrovascular disease can be effective with percutaneous techniques. Structural heart disease is increasingly being treated with percutaneous options, with a high likelihood that interventional approaches will compete with open-heart surgery in a significant proportion of cases in years to come.

FURTHER READING

Faxon DP, Williams DO: Interventional cardiology: Current status and future directions in coronary disease and valvular heart disease. Circulation 133:2697, 2016. [PubMed: 27324364]

Levine GN et al: 2011 ACCF/AHA/SCAI Guideline for Percutaneous Coronary Intervention: Executive Summary: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Society for Cardiovascular Angiography and Interventions. Circulation 124:2574, 2011. [PubMed: 22064598]

Moscucci M (ed): *Grossman & Baim's Cardiac Catheterization, Angiography, and Intervention* 8th ed. Philadelphia, Lippincott Williams & Wilkins, 2014.

Vahl TP et al: Transcatheter aortic valve replacement 2016: A modern-day "Through the looking-glass" adventure. J Am Coll Cardiol 67:1472, 2016. [PubMed: 27012409]

Chapter 271: Hypertensive Vascular Disease

Theodore A. Kotchen

INTRODUCTION

Hypertension is one of the leading causes of the global burden of disease. Elevated blood pressure affects more than one billion individuals and causes an estimated 9.4 million deaths per year. Hypertension doubles the risk of cardiovascular diseases, including coronary heart disease (CHD), congestive heart failure (CHF), ischemic and hemorrhagic stroke, renal failure, and peripheral arterial disease (PAD). It often is associated with additional cardiovascular disease risk factors, and the risk of cardiovascular disease increases with the total burden of risk factors. Although antihypertensive therapy reduces the risks of cardiovascular and renal disease, large segments of the hypertensive population are either untreated or inadequately treated.

EPIDEMIOLOGY

Blood pressure levels, the rate of age-related increases in blood pressure, and the prevalence of hypertension vary among countries and among subpopulations within a country. Hypertension is present in all populations except for small numbers of individuals living in isolated societies. In industrialized societies, blood pressure increases steadily during the first two decades of life. In children and adolescents, blood pressure is associated with growth and maturation. Blood pressure “tracks” over time in children and between adolescence and young adulthood. In the United States, average systolic blood pressure is higher for men than for women during early adulthood, although among older individuals the age-related rate of rise is steeper for women. Consequently, among individuals aged ≥60 years, systolic blood pressures of women are higher than those of men. Among adults, diastolic blood pressure also increases progressively with age until ~55 years, after which it tends to decrease. The consequence is a widening of pulse pressure (the difference between systolic and diastolic blood pressure) beyond age 60.

In the United States, ~78 million adults have hypertension. Hypertension prevalence is 33.5% in non-Hispanic blacks, 28.9% in non-Hispanic whites, and 20.7% in Mexican Americans. The likelihood of hypertension increases with age, and among individuals aged ≥60 years, the prevalence is 65.4%. Recent evidence suggests that the prevalence of hypertension in the United States may be increasing, possibly as a consequence of increasing obesity. The prevalence of hypertension and stroke mortality rates is higher in the southeastern United States than in other regions. In African Americans, hypertension appears earlier, is generally more severe, and results in higher rates of morbidity and mortality from stroke, left ventricular hypertrophy, CHF, and end-stage renal disease (ESRD) than in white Americans. According to NHANES (National Health and Nutrition Examination Survey) data, in 2007–2010, 81.5% of those with hypertension were aware they had it, 74.9% were being treated, but only 52.5% were controlled.

Both environmental and genetic factors may contribute to regional and racial variations in hypertension prevalence. Studies of societies undergoing “acculturation” and studies of migrants from a less to a more urbanized setting indicate a profound environmental contribution to blood pressure. Obesity and weight gain are strong, independent risk factors for hypertension. It has been estimated that 60% of hypertensives are >20% overweight. Among populations, hypertension prevalence is related to dietary NaCl intake, and the age-related increase in blood pressure may be augmented by a high NaCl intake. Low dietary intakes of calcium and potassium also may contribute to the risk of hypertension. The urine sodium-to-potassium ratio (an index of both sodium and potassium intakes) is a stronger correlate of blood pressure than is either sodium or potassium alone. Alcohol consumption, psychosocial stress, and low levels of physical activity also may contribute to hypertension.

GENETIC CONSIDERATIONS

 Although specific genetic variants have been identified in rare Mendelian forms of hypertension (Table 271–5), these variants are not applicable to the vast majority (>98%) of patients with hypertension. For most individuals, it is likely that hypertension represents a polygenic disorder in which a combination of genes acts in concert with environmental exposures to make only a modest contribution to blood pressure. Furthermore, different subsets of genes may lead to different phenotypes associated with hypertension, e.g., obesity, dyslipidemia, insulin resistance.

Adoption, twin, and family studies document a significant heritable component to blood pressure levels and hypertension. Animal models (including selectively bred rats and congenic rat strains) have identified a number of genetic loci and genes associated with hypertension. Clinically, although replication has been a challenge, results of candidate gene studies and genome-wide association studies in large numbers of individuals have also identified a number of hypertension-related genes, several of which are involved in pathways that regulate arterial pressure, e.g., genes that encode components of the renin-angiotensin-aldosterone system, atrial natriuretic peptide, the beta-2 adrenoreceptor, and alpha adducin (associated with increased renal tubular reabsorption of sodium). Overall, identified genetic determinants account for ~1% of blood pressure variance, whereas based on family studies, heritability of hypertension is estimated to be in the range of 30–40%. One hypothesis to account for the “missing heritability” is that epigenetic modifications of DNA contribute to the heritability of blood pressure. Epigenetic processes are changes in gene expression that occur without changes in DNA sequence. In contrast to DNA sequence, the epigenome is relatively susceptible to modification by environmental exposures. Epigenetic dysregulation has emerged as a hallmark of several complex diseases, including hypertension. Several recent studies have described epigenetic modifications of specific genes associated with hypertension. However, current results of detailed genome-wide epigenetic modifications of DNA are limited and conflicting.

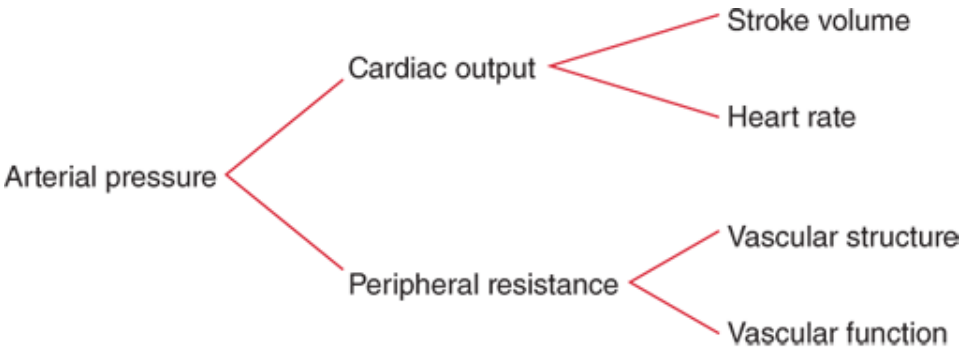
Preliminary evidence suggests that there may also be genetic determinants of target organ damage and vascular disease attributed to hypertension. Family studies indicate significant heritability of left ventricular mass, and there is considerable individual variation in the responses of the heart to hypertension. Family studies and variations in candidate genes associated with renal damage suggest that genetic factors also may contribute to hypertensive nephropathy. Specific genetic variants have been linked to CHD and stroke. In the future, it is possible that DNA and epigenetic analyses may predict individual risk for hypertension and target organ damage and will identify responders to specific classes of antihypertensive agents.

MECHANISMS OF HYPERTENSION

To provide a framework for understanding the pathogenesis and treatment options for hypertensive disorders, it is useful to understand factors involved in the regulation of both normal and elevated arterial pressure. Cardiac output and peripheral resistance are the two determinants of arterial pressure **(Fig. 271-1)**. Cardiac output is determined by stroke volume and heart rate; stroke volume is related to myocardial contractility and to the size of the vascular compartment. Peripheral resistance is determined by functional and anatomic changes in small arteries (lumen diameter 100–400 μm) and arterioles.

FIGURE 271-1

Determinants of arterial pressure.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

INTRAVASCULAR VOLUME

Sodium is predominantly an extracellular ion and is a primary determinant of the extracellular fluid volume. When NaCl intake exceeds the capacity of the kidney to excrete sodium, vascular volume may initially expand and cardiac output may increase. However, many vascular beds have the capacity to autoregulate blood flow, and if constant blood flow is to be maintained in the face of increased arterial pressure, resistance within that bed must increase, since

$$\text{Blood Flow} = \frac{\text{pressure across the vascular bed}}{\text{vascular resistance}}$$

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The initial elevation of blood pressure in response to vascular volume expansion may be related to an increase of cardiac output; however, over time, peripheral resistance increases and cardiac output reverts toward normal. Whether this hypothesized sequence of events occurs in the pathogenesis of hypertension is not clear. What is clear is that salt can activate a number of neural, endocrine/paracrine, and vascular mechanisms, all of which have the potential to increase arterial pressure. The effect of sodium on blood pressure is related to the provision of sodium with chloride; non-chloride salts of sodium have little or no effect on blood pressure. As arterial pressure increases in response to a high NaCl intake, urinary sodium excretion increases and sodium balance is maintained at the expense of an increase in arterial pressure. The mechanism for this “pressure-natriuresis” phenomenon may involve a subtle increase in the glomerular filtration rate, decreased absorbing capacity of the renal tubules, and possibly hormonal factors such as atrial natriuretic factor. In individuals with an impaired capacity to excrete sodium, greater increases in arterial pressure are required to achieve natriuresis and sodium balance.

NaCl-dependent hypertension may be a consequence of a decreased capacity of the kidney to excrete sodium, due either to intrinsic renal disease or to increased production of a salt-retaining hormone (mineralocorticoid) resulting in increased renal tubular reabsorption of sodium. Renal tubular sodium reabsorption also may be augmented by increased neural activity to the kidney. In each of these situations, a higher arterial pressure may be required to achieve sodium balance. Conversely, salt-wasting disorders are associated with low blood pressure levels. ESRD is an extreme example of volume-dependent hypertension. In ~80% of these patients, vascular volume and hypertension can be controlled with adequate dialysis; in the other 20%, the mechanism of hypertension is related to increased activity of the renin-angiotensin system and is likely to be responsive to pharmacologic blockade of renin-angiotensin.

AUTONOMIC NERVOUS SYSTEM

Adrenergic reflexes modulate blood pressure over the short term, and adrenergic function, in concert with hormonal and volume-related factors, contributes to the long-term regulation of arterial pressure.

Norepinephrine, epinephrine, and dopamine all play important roles in tonic and phasic cardiovascular regulation.

The activities of the adrenergic receptors are mediated by guanosine nucleotide-binding regulatory proteins (G proteins) and by intracellular concentrations of downstream second messengers. In addition to receptor affinity and density, physiologic responsiveness to catecholamines may be altered by the efficiency of receptor-effector coupling at a site “distal” to receptor binding. The receptor sites are relatively specific both for the transmitter substance and for the response that occupancy of the receptor site elicits. Based on their physiology and pharmacology, adrenergic receptors have been divided into two principal types: α and β. These types have been differentiated further into α_1, α_2, β_1, and β_2 receptors. Recent molecular cloning studies have identified several additional subtypes. α Receptors are occupied and activated more avidly by norepinephrine than by epinephrine, and the reverse is true for β receptors. α_1 Receptors are located on postsynaptic cells in smooth muscle and elicit vasoconstriction. α_2 Receptors are localized on presynaptic membranes of postganglionic nerve terminals that synthesize norepinephrine. When activated by catecholamines, α_2 receptors act as negative feedback controllers, inhibiting further norepinephrine release. In the kidney, activation of α_1-adrenergic receptors increases renal tubular reabsorption of sodium. Different classes of antihypertensive agents either inhibit α_1 receptors or act as agonists of α_2 receptors and reduce systemic sympathetic outflow. Activation of myocardial β_1 receptors stimulates the rate and strength of cardiac contraction and consequently increases cardiac output. β_1 Receptor activation also stimulates renin release from the kidney. Another class of antihypertensive agents acts by inhibiting β_1 receptors. Activation of β_2 receptors by epinephrine relaxes vascular smooth muscle and results in vasodilation.

Circulating catecholamine concentrations may affect the number of adrenoreceptors in various tissues. Downregulation of receptors may be a consequence of sustained high levels of catecholamines and provides an explanation for decreasing responsiveness, or tachyphylaxis, to catecholamines. For example, orthostatic hypotension frequently is observed in patients with pheochromocytoma, possibly due to the lack of norepinephrine-induced vasoconstriction with assumption of the upright posture. Conversely, with chronic reduction of neurotransmitter substances, adrenoreceptors may increase in number or be upregulated, resulting in increased responsiveness to the neurotransmitter. Chronic administration of agents that block adrenergic receptors may result in upregulation, and abrupt withdrawal of those agents may produce a condition of temporary hypersensitivity to sympathetic stimuli. For example, clonidine is an antihypertensive agent that is a centrally acting α_2 agonist that inhibits sympathetic outflow. Rebound hypertension may occur with the abrupt cessation of clonidine therapy, probably as a consequence of upregulation of α_1 receptors.

Several reflexes modulate blood pressure on a minute-to-minute basis. One arterial baroreflex is mediated by stretch-sensitive sensory nerve endings in the carotid sinuses and the aortic arch. The rate of firing of these baroreceptors increases with arterial pressure, and the net effect is a decrease in sympathetic outflow, resulting in decreases in arterial pressure and heart rate. This is a primary mechanism for rapid buffering of acute fluctuations of arterial pressure that may occur during postural changes, behavioral or physiologic stress, and changes in blood volume. However, the activity of the baroreflex declines or adapts to sustained increases in arterial pressure such that the baroreceptors are reset to higher pressures. Patients with autonomic neuropathy and impaired baroreflex function may have extremely labile blood pressures with difficult-to-control episodic blood pressure spikes associated with tachycardia.

In both normal-weight and obese individuals, hypertension often is associated with increased sympathetic outflow. Based on recordings of postganglionic muscle nerve activity (detected by a microelectrode inserted in a peroneal nerve in the leg), sympathetic outflow tends to be higher in hypertensive than in normotensive individuals. Sympathetic outflow is increased in obesity-related hypertension and in hypertension associated with obstructive sleep apnea. Baroreceptor activation via electrical stimulation of carotid sinus afferent nerves lowers blood pressure in patients with “resistant” hypertension. Drugs that block the sympathetic nervous system are potent antihypertensive agents, indicating that the sympathetic nervous system plays a permissive, although not necessarily a causative, role in the maintenance of increased arterial pressure.

Pheochromocytoma is the most blatant example of hypertension related to increased catecholamine production, in this instance by a tumor. Blood pressure can be reduced by surgical excision of the tumor or by

pharmacologic treatment with an α_1 receptor antagonist or with an inhibitor of tyrosine hydroxylase, the rate-limiting step in catecholamine biosynthesis.

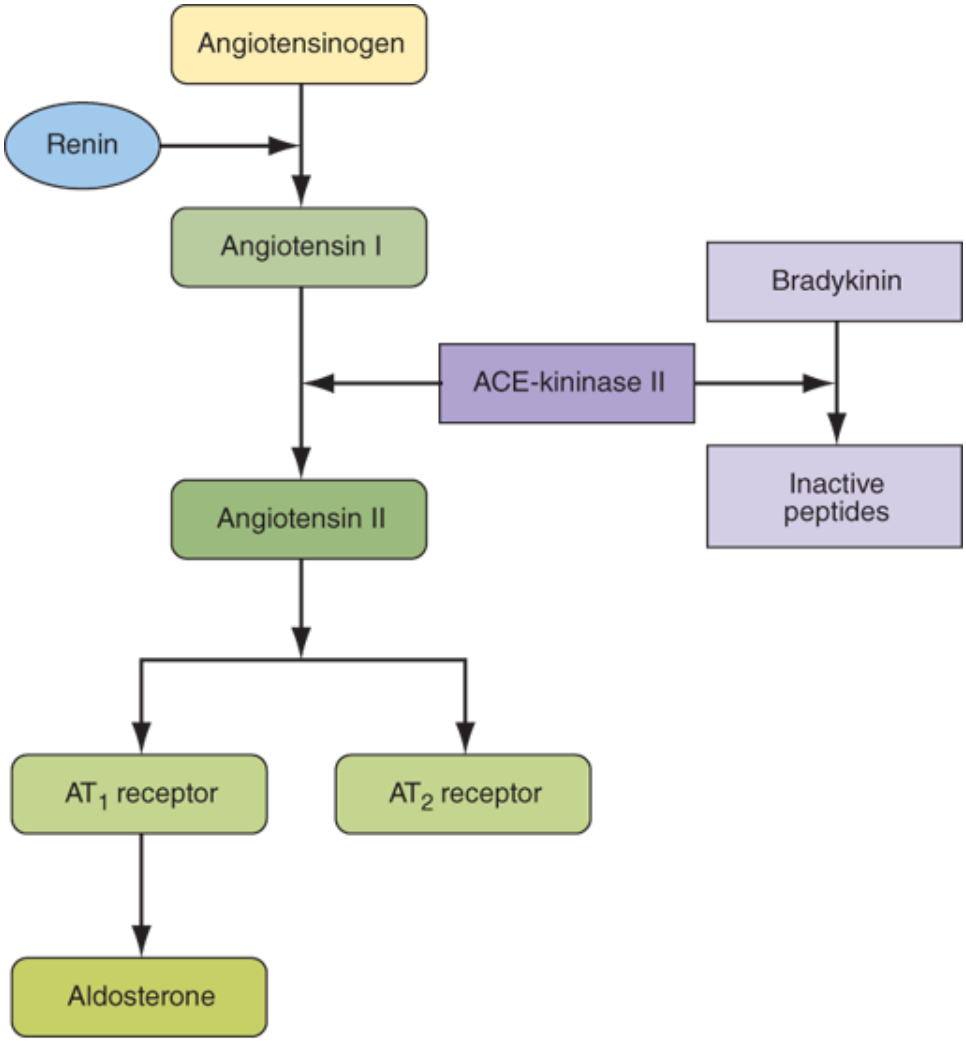
RENIN-ANGIOTENSIN-ALDOSTERONE

The renin-angiotensin-aldosterone system contributes to the regulation of arterial pressure primarily via the vasoconstrictor properties of angiotensin II and the sodium-retaining properties of aldosterone. Renin is an aspartyl protease that is synthesized as an enzymatically inactive precursor, prorenin. Most renin in the circulation is synthesized in the renal afferent renal arteriole. Prorenin may be secreted directly into the circulation or may be activated within secretory cells and released as active renin. Although human plasma contains two to five times more prorenin than renin, there is no evidence that prorenin contributes to the physiologic activity of this system. There are three primary stimuli for renin secretion: (1) decreased NaCl transport in the distal portion of the thick ascending limb of the loop of Henle that abuts the corresponding afferent arteriole (macula densa), (2) decreased pressure or stretch within the renal afferent arteriole (baroreceptor mechanism), and (3) sympathetic nervous system stimulation of renin-secreting cells via β_1 adrenoreceptors. Conversely, renin secretion is inhibited by increased NaCl transport in the thick ascending limb of the loop of Henle, by increased stretch within the renal afferent arteriole, and by β_1 receptor blockade. In addition, angiotensin II directly inhibits renin secretion due to angiotensin II type 1 receptors on juxtaglomerular cells, and renin secretion increases in response to pharmacologic blockade of either angiotensin-converting enzyme (ACE) or angiotensin II receptors.

Once released into the circulation, active renin cleaves a substrate, angiotensinogen, to form an inactive decapeptide, angiotensin I **(Fig. 271-2)**. A converting enzyme, located primarily but not exclusively in the pulmonary circulation, converts angiotensin I to the active octapeptide, angiotensin II, by releasing the C-terminal histidyl-leucine dipeptide. The same converting enzyme cleaves a number of other peptides, including and thereby inactivating the vasodilator bradykinin. Acting primarily through angiotensin II type 1 (AT_1) receptors on cell membranes, angiotensin II is a potent pressor substance, the primary tropic factor for the secretion of aldosterone by the adrenal zona glomerulosa. Independent of its hemodynamic effects, angiotensin II may play a role in the pathogenesis of atherosclerosis through a direct cellular action on the vessel wall. The angiotensin II type 2 (AT_2) receptor has the opposite functional effects of the AT_1 receptor. The AT_2 receptor induces vasodilation, sodium excretion, and inhibition of cell growth and matrix formation. Experimental evidence suggests that the AT_2 receptor improves vascular remodeling by stimulating smooth muscle cell apoptosis and contributes to the regulation of glomerular filtration rate. AT_1 receptor blockade induces an increase in AT_2 receptor activity.

FIGURE 271-2

Renin-angiotensin-aldosterone axis. ACE, angiotensin-converting enzyme.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Renin-secreting tumors are clear examples of renin-dependent hypertension. In the kidney, these tumors include benign hemangiopericytomas of the juxtaglomerular apparatus and, infrequently, renal carcinomas, including Wilms' tumors. Renin-producing carcinomas also have been described in lung, liver, pancreas, colon, and adrenals. Renovascular hypertension is another renin-mediated form of hypertension. Obstruction of the renal artery leads to decreased renal perfusion pressure, thereby stimulating renin secretion. Over time, possibly as a consequence of secondary renal damage, this form of hypertension may become less renin-dependent.

Angiotensinogen, renin, and angiotensin II are also synthesized locally in many tissues, including the brain, pituitary, aorta, arteries, heart, adrenal glands, kidneys, adipocytes, leukocytes, ovaries, testes, uterus, spleen, and skin. Angiotensin II in tissues may be formed by the enzymatic activity of renin or by other proteases, e.g., tonin, chymase, and cathepsins. In addition to regulating local blood flow, tissue angiotensin II is a mitogen that stimulates growth and contributes to modeling and repair. Excess tissue angiotensin II may contribute to atherosclerosis, cardiac hypertrophy, and renal failure, and consequently may be a target for pharmacologic therapy to prevent target organ damage.

Angiotensin II is the primary tropic factor regulating the synthesis and secretion of aldosterone by the zona glomerulosa of the adrenal cortex. Aldosterone synthesis is also dependent on potassium, and aldosterone secretion may be decreased in potassium-depleted individuals. Although acute elevations of adrenocorticotropic hormone (ACTH) levels also increase aldosterone secretion, ACTH is not an important tropic factor for the chronic regulation of aldosterone.

Aldosterone is a potent mineralocorticoid that increases sodium reabsorption by amiloride-sensitive epithelial sodium channels (ENaC) on the apical surface of the principal cells of the renal cortical collecting duct **(Chap. 303)**. Electric neutrality is maintained by exchanging sodium for potassium and hydrogen ions. Consequently, increased aldosterone secretion may result in hypokalemia and alkalosis. Beyond its renal effects, aldosterone can exert deleterious effects on the cardiovascular system, including fibrosis, endothelial dysfunction, inflammation, and oxidative stress, as well as an overall increase in cardiovascular morbidity and mortality.

Cortisol also binds to the mineralocorticoid receptor but normally functions as a less potent mineralocorticoid than aldosterone because cortisol is converted to cortisone by the enzyme 11 β-hydroxysteroid dehydrogenase type 2. Cortisone has no affinity for the mineralocorticoid receptor. Primary aldosteronism is a compelling example of mineralocorticoid-mediated hypertension. In this disorder, adrenal aldosterone synthesis and release are independent of renin-angiotensin, and renin release is suppressed by the resulting volume expansion.

Mineralocorticoid receptors are expressed in a number of tissues in addition to the kidney, and mineralocorticoid receptor activation induces structural and functional alterations in the heart, kidney, and blood vessels, leading to myocardial fibrosis, nephrosclerosis, and vascular inflammation and remodeling, perhaps as a consequence of oxidative stress. These effects are amplified by a high salt intake. In animal models, high circulating aldosterone levels stimulate cardiac fibrosis and left ventricular hypertrophy, and spironolactone (an aldosterone antagonist) prevents aldosterone-induced myocardial fibrosis. Pathologic patterns of left ventricular geometry also have been associated with elevations of plasma aldosterone concentration in hypertensive patients. In patients with CHF, low-dose spironolactone reduces the risk of progressive heart failure and sudden death from cardiac causes by 30%. Due to a renal hemodynamic effect, in patients with primary aldosteronism, high circulating levels of aldosterone also may cause glomerular hyperfiltration and albuminuria.

Increased activity of the renin-angiotensin-aldosterone axis is not invariably associated with hypertension. In response to a low-NaCl diet or to volume contraction, arterial pressure and volume homeostasis may be maintained by increased activity of the renin-angiotensin-aldosterone axis. Secondary aldosteronism (i.e., increased aldosterone secondary to increased renin-angiotensin), but not hypertension, also is observed in edematous states such as CHF and liver disease.

VASCULAR MECHANISMS

Vascular radius and compliance of resistance arteries are important determinants of arterial pressure. Resistance to flow varies inversely with the fourth power of the radius, and consequently, small decreases in lumen size significantly increase resistance. In hypertensive patients, structural, mechanical, or functional changes may reduce the lumen diameter of small arteries and arterioles. Remodeling refers to geometric alterations in the vessel wall without a change in vessel volume. Hypertrophic (increased cell size, and increased deposition of intercellular matrix) or eutrophic vascular remodeling results in decreased lumen size and, hence, increased peripheral resistance. Apoptosis, low-grade inflammation, and vascular fibrosis also contribute to remodeling. Lumen diameter also is related to elasticity of the vessel. Vessels with a high degree of elasticity can accommodate an increase of volume with relatively little change in pressure, whereas in a semi-rigid vascular system, a small increment in volume induces a relatively large increment of pressure.

An association between arterial stiffness and hypertension is well established. A stiffened vasculature is less able to buffer short-term alterations in flow. Although it has been assumed that arterial stiffness is a manifestation of hypertension, recent evidence suggests that vascular stiffness may also represent a cause of hypertension. Non-invasive determination of pulse wave velocity between the carotid and femoral arteries is often interpreted as an index of arterial stiffness. Due to arterial stiffness, central blood pressures (aortic, carotid) may not correspond to brachial artery pressures. Ejection of blood into the aorta elicits a pressure wave that is propagated at a given velocity. The forward traveling wave generates a reflected wave that travels backward toward the ascending aorta. Although mean arterial pressure is determined by cardiac output and peripheral resistance, pulse pressure is related to the functional properties of large arteries and the amplitude and timing of the incident and reflected waves. Increased arterial stiffness results in increased pulse wave velocity of both incident and reflected waves. Due to the timing of these waves, the consequence is augmentation of aortic systolic pressure and a reduction of aortic diastolic pressure, i.e., an increase in

pulse pressure. The aortic augmentation index, a surrogate index of arterial stiffening, is calculated as the ratio of central arterial pressure-to-pulse pressure. However, wave reflections are also influenced by left ventricular structure and function. Central blood pressure may be measured directly by placing a sensor in the aorta or noninvasively by radial tonometry using commercially available devices. Central blood pressure and the aortic augmentation index are strong, independent predictors of cardiovascular disease and all-cause mortality. Central blood pressure also appears to be more strongly associated with pre-clinical organ damage than brachial blood pressure.

Ion transport by vascular smooth muscle cells may contribute to hypertension-associated abnormalities of vascular tone and vascular growth, both of which are modulated by intracellular pH (pH_i). Three ion transport mechanisms participate in the regulation of pH_i: (1) Na^+-H^+ exchange, (2) Na^+-dependent HCO_3^--Cl^- exchange, and (3) cation-independent HCO_3^--Cl^- exchange. Based on measurements in cell types that are more accessible than vascular smooth muscle (e.g., leukocytes, erythrocytes, platelets, skeletal muscle), activity of the Na^+-H^+ exchanger is increased in hypertension, and this may result in increased vascular tone by two mechanisms. First, increased sodium entry may lead to increased vascular tone by activating Na^+-Ca^{2+} exchange and thereby increasing intracellular calcium. Second, increased pH_i enhances calcium sensitivity of the contractile apparatus, leading to an increase in contractility for a given intracellular calcium concentration. Additionally, increased Na^+-H^+ exchange may stimulate growth of vascular smooth muscle cells by enhancing sensitivity to mitogens.

Vascular endothelial function also modulates vascular tone. The vascular endothelium synthesizes and releases several vasoactive substances, including nitric oxide, a potent vasodilator. Endothelium-dependent vasodilation is impaired in hypertensive patients. This impairment often is assessed with high-resolution ultrasonography before and after the hyperemic phase of reperfusion that follows 5 min of forearm ischemia. Alternatively, endothelium-dependent vasodilation may be assessed in response to an intra-arterially infused endothelium-dependent vasodilator, e.g., acetylcholine. Endothelin is a vasoconstrictor peptide produced by the endothelium, and orally active endothelin antagonists may lower blood pressure in patients with resistant hypertension.

Currently, it is not known if the hypertension-related vascular abnormalities of ion transport and endothelial function are primary alterations or secondary consequences of elevated arterial pressure. Limited evidence suggests that vascular compliance and endothelium-dependent vasodilation may be improved by aerobic exercise, weight loss, and antihypertensive agents. It remains to be determined whether these interventions affect arterial structure and stiffness via a blood pressure–independent mechanism and whether different classes of antihypertensive agents preferentially affect vascular structure and function.

IMMUNE MECHANISMS, INFLAMMATION, AND OXIDATIVE STRESS

Inflammation and alterations of the immune response have been implicated in the pathogenesis of vascular injury and hypertension for at least four decades. Patients with primary hypertension have increased circulating levels of autoantibodies. Both hypertension and aortic stiffness are associated with activation of innate and adaptive immunity. Many forms of hypertension in experimental animals are associated with an inflammatory component requiring T lymphocytes. Inflammation and exudative injury are closely coupled. Inflammation, vascular stretch, angiotensin II, and salt have all been shown to result in the generation of reactive oxygen species (ROS), which modify T cell function and further enhance inflammation. ROS also attenuate the effects of endogenous small-molecule vasodilators. ROS within the renal medulla is a key determinant of the set point of the renal pressure-natriuresis curve. Increasing evidence suggests that infiltration of T cells into the renal interstitium contributes to inflammation and oxidative stress. Renal medullary oxidative stress disrupts pressure-natriuresis and contributes to the development of hypertension in experimental models. Clinically, markers of oxidative stress have been described in both hypertensive and pre-hypertensive patients.

PATHOLOGIC CONSEQUENCES OF HYPERTENSION

HEART

Heart disease is the most common cause of death in hypertensive patients. Hypertensive heart disease is the result of structural and functional adaptations leading to left ventricular hypertrophy, CHF, atherosclerotic coronary artery disease and microvascular disease, and cardiac arrhythmias, including atrial fibrillation. Individuals with left ventricular hypertrophy are at increased risk for CHD, stroke, CHF, and sudden death. Aggressive control of hypertension can regress or reverse left ventricular hypertrophy and reduce the risk of cardiovascular disease.

CHF may be related to systolic dysfunction, diastolic dysfunction, or a combination of the two. Abnormalities of diastolic function that range from asymptomatic heart disease to overt heart failure are common in hypertensive patients. Approximately one-third of patients with CHF have normal systolic function but abnormal diastolic function. Diastolic dysfunction is an early consequence of hypertension-related heart disease and is exacerbated by left ventricular hypertrophy and ischemia. Cardiac catheterization provides the most accurate assessment of diastolic function. Alternatively, diastolic function can be evaluated by several noninvasive methods, including echocardiography and radionuclide angiography.

BRAIN

Stroke is the second most frequent cause of death in the world; it accounts for 5 million deaths each year, with an additional 15 million persons having nonfatal strokes. Elevated blood pressure is the strongest risk factor for stroke. Approximately 85% of strokes are due to infarction, and the remainder are due to either intracerebral or subarachnoid hemorrhage. The incidence of stroke rises progressively with increasing blood pressure levels, particularly systolic blood pressure in individuals aged >65 years. Treatment of hypertension decreases the incidence of both ischemic and hemorrhagic strokes.

Hypertension also is associated with impaired cognition in an aging population, and longitudinal studies support an association between midlife hypertension and late-life cognitive decline. Hypertension is associated with beta amyloid deposition, a major pathologic factor in dementia. In addition to actual blood pressure level, arterial stiffness and visit-to-visit blood pressure variability may be independently related to subclinical small vessel disease and subsequent cognitive decline. Hypertension-related cognitive impairment and dementia may also be a consequence of a single infarct due to occlusion of a “strategic” larger vessel or multiple lacunar infarcts due to occlusive small vessel disease resulting in subcortical white matter ischemia. Several clinical trials suggest that antihypertensive therapy has a beneficial effect on cognitive function, although this remains an active area of investigation.

Cerebral blood flow remains unchanged over a wide range of arterial pressures (mean arterial pressure of 50–150 mmHg) through a process termed *autoregulation* of blood flow. In patients with the clinical syndrome of malignant hypertension, encephalopathy is related to failure of autoregulation of cerebral blood flow at the upper pressure limit, resulting in vasodilation and hyperperfusion. Signs and symptoms of hypertensive encephalopathy may include severe headache, nausea and vomiting (often of a projectile nature), focal neurologic signs, and alterations in mental status. Untreated, hypertensive encephalopathy may progress to stupor, coma, seizures, and death within hours. It is important to distinguish hypertensive encephalopathy from other neurologic syndromes that may be associated with hypertension, e.g., cerebral ischemia, hemorrhagic or thrombotic stroke, seizure disorder, mass lesions, pseudotumor cerebri, delirium tremens, meningitis, acute intermittent porphyria, traumatic or chemical injury to the brain, and uremic encephalopathy.

KIDNEY

The kidney is both a target and a cause of hypertension. Primary renal disease is the most common etiology of secondary hypertension. Mechanisms of kidney-related hypertension include a diminished capacity to excrete sodium, excessive renin secretion in relation to volume status, and sympathetic nervous system overactivity. Conversely, hypertension is a risk factor for renal injury and ESRD. The increased risk associated with high blood pressure is graded, continuous, and present throughout the distribution of blood pressure above optimal pressure. Renal risk appears to be more closely related to systolic than to diastolic blood pressure, and black men are at greater risk than white men for developing ESRD at every level of blood pressure.

Atherosclerotic, hypertension-related vascular lesions in the kidney primarily affect preglomerular arterioles, resulting in ischemic changes in the glomeruli and postglomerular structures. Glomerular injury also may be a consequence of direct damage to the glomerular capillaries due to glomerular hyperperfusion. Studies of hypertension-related renal damage, primarily in experimental animals, suggest that loss of autoregulation of renal blood flow at the afferent arteriole results in transmission of elevated pressures to an unprotected glomerulus with ensuing hyperfiltration, hypertrophy, and eventual focal segmental glomerular sclerosis. With progressive renal injury there is a loss of autoregulation of renal blood flow and glomerular filtration rate, resulting in a lower blood pressure threshold for renal damage and a steeper slope between blood pressure and renal damage. The result may be a vicious cycle of renal damage and nephron loss leading to more severe hypertension, glomerular hyperfiltration, and further renal damage. Glomerular pathology progresses to glomerulosclerosis, and eventually the renal tubules may also become ischemic and gradually atrophic. The renal lesion associated with malignant hypertension consists of fibrinoid necrosis of the afferent arterioles, sometimes extending into the glomerulus, and may result in focal necrosis of the glomerular tuft.

Clinically, macroalbuminuria (a random urine albumin/creatinine ratio >300 mg/g) or microalbuminuria (a random urine albumin/creatinine ratio 30–300 mg/g) are early markers of renal injury. These are also risk factors for renal disease progression and cardiovascular disease.

PERIPHERAL ARTERIES

In addition to contributing to the pathogenesis of hypertension, blood vessels are a target organ for atherosclerotic disease secondary to long-standing elevated blood pressure. In hypertensive patients, vascular disease is a major contributor to stroke, heart disease, and renal failure. Further, hypertensive patients with arterial disease of the lower extremities are at increased risk for future cardiovascular disease. Although patients with stenotic lesions of the lower extremities may be asymptomatic, intermittent claudication is the classic symptom of PAD. The ankle-brachial index is a useful approach for evaluating PAD and is defined as the ratio of noninvasively assessed ankle to brachial (arm) systolic blood pressure. An ankle-brachial index <0.90 is considered diagnostic of PAD and is associated with >50% stenosis in at least one major lower limb vessel. An ankle-brachial index <0.80 is associated with elevated blood pressure, particularly systolic blood pressure.

DEFINING HYPERTENSION

From an epidemiologic perspective, there is no obvious level of blood pressure that defines hypertension. In adults, there is a continuous, incremental risk of cardiovascular disease, stroke, and renal disease across levels of both systolic and diastolic blood pressure. The Multiple Risk Factor Intervention Trial (MRFIT), which included >350,000 male participants, demonstrated a continuous and graded influence of both systolic and diastolic blood pressure on CHD mortality, extending down to systolic blood pressures of 120 mmHg. Similarly, results of a meta-analysis involving almost 1 million participants indicate that ischemic heart disease mortality, stroke mortality, and mortality from other vascular causes are directly related to the height of the blood pressure, beginning at 115/75 mmHg, without evidence of a threshold. Cardiovascular disease risk doubles for every 20-mmHg increase in systolic and 10-mmHg increase in diastolic pressure. Among older individuals, systolic blood pressure and pulse pressure are more powerful predictors of cardiovascular disease than is diastolic blood pressure.

Clinically, hypertension may be defined as that level of blood pressure at which the institution of therapy reduces blood pressure–related morbidity and mortality. Clinical criteria for defining hypertension generally have been based on the average of two or more seated blood pressure readings during each of two or more outpatient visits. One classification recommends blood pressure criteria for defining normal blood pressure, prehypertension, hypertension (stages I and II), and isolated systolic hypertension, which is frequent among the elderly **(Table 271-1)**. In children and adolescents, hypertension generally is defined as systolic and/or diastolic blood pressure consistently >95th percentile for age, sex, and height. Blood pressures between the 90th and 95th percentiles are considered prehypertensive and are an indication for lifestyle interventions.

TABLE 271-1 Blood Pressure Classification

Blood Pressure Classification	Systolic, mmHg	Diastolic, mmHg
Normal	<120	*and* <80
Prehypertension	120–139	*or* 80–89
Stage 1 hypertension	140–159	*or* 90–99
Stage 2 hypertension	≥160	*or* ≥100
Isolated systolic hypertension	≥140	*and* <90

Source: Adapted from AV Chobanian et al: JAMA 289:2560, 2003.

Home blood pressure and average 24-h ambulatory blood pressure measurements are generally lower than clinic blood pressures. Because ambulatory blood pressure recordings yield multiple readings throughout the day and night, they provide a more comprehensive assessment of the vascular burden of hypertension than do a limited number of office readings. Increasing evidence suggests that home blood pressures, including 24-h blood pressure recordings, more reliably predict target organ damage than do office blood pressures. Blood pressure tends to be higher in the early morning hours, soon after waking, than at other times of day. Myocardial infarction and stroke are more common in the early morning hours. Nighttime blood pressures are generally 10–20% lower than daytime blood pressures, and an attenuated nighttime blood pressure “dip” may be associated with increased cardiovascular disease risk. Recommended criteria for a diagnosis of hypertension, based on 24-h blood pressure monitoring, are average awake blood pressure ≥135/85 mmHg and asleep blood pressure ≥120/75 mmHg. These levels approximate a clinic blood pressure of 140/90 mmHg.

Approximately 15–20% of patients with stage 1 hypertension based on office blood pressures have average ambulatory readings <135/85 mmHg, termed “white coat hypertension.” Long-term outcomes of individuals with white coat hypertension are more similar to normotensive individuals than to individuals with sustained hypertension (elevation of both office and out-of-office blood pressures). To confirm hypertension, some authorities recommend ambulatory blood pressure monitoring in all individuals with elevated clinic blood pressure, and postponing therapy with careful follow-up in those individuals with normal out-of-office blood pressures who are at low cardiovascular risk. In contrast, the prognosis of “masked hypertension” (normal office blood pressure and elevated out-of-office blood pressure) is nearly equivalent to that of sustained hypertension.

CLINICAL DISORDERS OF HYPERTENSION

Depending on methods of patient ascertainment, ~80–95% of hypertensive patients are diagnosed as having primary, or “essential,” hypertension. In the remaining 5–20% of hypertensive patients, a specific underlying disorder causing the elevation of blood pressure can be identified **(Tables 271-2 and 271-3)**. In individuals with “secondary” hypertension, a specific mechanism for the blood pressure elevation is often more apparent.

TABLE 271-2 Systolic Hypertension with Wide Pulse Pressure

1. Decreased vascular compliance (arteriosclerosis)
2. Increased cardiac output
    a. Aortic regurgitation
    b. Thyrotoxicosis
    c. Hyperkinetic heart syndrome
    d. Fever
    e. Arteriovenous fistula
    f. Patent ductus arteriosus

TABLE 271-3 Secondary Causes of Systolic and Diastolic Hypertension

Renal	Parenchymal diseases, renal cysts (including polycystic kidney disease), renal tumors (including renin-secreting tumors), obstructive uropathy
Renovascular	Arteriosclerotic, fibromuscular dysplasia
Adrenal	Primary aldosteronism, Cushing's syndrome, 17α-hydroxylase deficiency, 11β-hydroxylase deficiency, 11-hydroxysteroid dehydrogenase deficiency (licorice), pheochromocytoma
Aortic coarctation	
Obstructive sleep apnea	
Preeclampsia/eclampsia	
Neurogenic	Psychogenic, diencephalic syndrome, familial dysautonomia, polyneuritis (acute porphyria, lead poisoning), acute increased intracranial pressure, acute spinal cord section
Miscellaneous endocrine	Hypothyroidism, hyperthyroidism, hypercalcemia, acromegaly
Medications	High-dose estrogens, adrenal steroids, decongestants, appetite suppressants, cyclosporine, tricyclic antidepressants, monoamine oxidase inhibitors, erythropoietin, nonsteroidal anti-inflammatory agents, cocaine
Mendelian forms of hypertension	See Table 271-4

PRIMARY HYPERTENSION

Primary hypertension tends to be familial and is likely to be the consequence of an interaction between environmental and genetic factors. The prevalence of primary hypertension increases with age, and individuals with relatively high blood pressures at younger ages are at increased risk for the subsequent development of hypertension. It is likely that primary hypertension represents a spectrum of disorders with different underlying pathophysiologies. In the majority of patients with established hypertension, peripheral resistance is increased and cardiac output is normal or decreased; however, in younger patients with mild or labile hypertension, cardiac output may be increased and peripheral resistance may be normal.

When plasma renin activity (PRA) is plotted against 24-h sodium excretion, ~10–15% of hypertensive patients have high PRA and 25% have low PRA. High-renin patients may have a vasoconstrictor form of hypertension, whereas low-renin patients may have volume-dependent hypertension. Inconsistent associations between plasma aldosterone and blood pressure have been described in patients with primary hypertension. The association between aldosterone and blood pressure is more striking in African Americans, and PRA tends to be low in hypertensive African Americans. This raises the possibility that subtle increases in aldosterone may contribute to hypertension in at least some groups of patients who do not have overt

primary aldosteronism. Furthermore, spironolactone, an aldosterone antagonist, may be a particularly effective antihypertensive agent for some patients with primary hypertension, including some patients with “drug-resistant” hypertension.

OBESITY AND THE METABOLIC SYNDROME

(See also Chap. 401) There is a well-documented association between obesity (body mass index >30 kg/m^2) and hypertension. Further, cross-sectional studies indicate a direct linear correlation between body weight (or body mass index) and blood pressure. Centrally located body fat is a more important determinant of blood pressure elevation than is peripheral body fat. In longitudinal studies, a direct correlation exists between change in weight and change in blood pressure over time. Sixty percent of hypertensive adults are >20% overweight. It has been established that 60–70% of hypertension in adults may be directly attributable to adiposity.

Hypertension and dyslipidemia frequently occur together and in association with resistance to insulin-stimulated glucose uptake. This clustering of risk factors is often, but not invariably, associated with obesity, particularly abdominal obesity. Insulin resistance also is associated with an unfavorable imbalance in the endothelial production of mediators that regulate platelet aggregation, coagulation, fibrinolysis, and vessel tone. When these risk factors cluster, the risks for CHD, stroke, diabetes, and cardiovascular disease mortality are increased further.

Depending on the populations studied and the methodologies for defining insulin resistance, ~25–50% of nonobese, nondiabetic hypertensive persons are insulin resistant. The constellation of insulin resistance, abdominal obesity, hypertension, and dyslipidemia has been designated as the *metabolic syndrome*. As a group, first-degree relatives of patients with primary hypertension are also insulin resistant, and hyperinsulinemia (a surrogate marker of insulin resistance) may predict the eventual development of hypertension and cardiovascular disease. Although the metabolic syndrome may in part be heritable as a polygenic condition, the expression of the syndrome is modified by environmental factors, such as degree of physical activity and diet. Insulin sensitivity increases and blood pressure decreases in response to weight loss. The recognition that cardiovascular disease risk factors tend to cluster within individuals has important implications for the evaluation and treatment of hypertension. Evaluation of both hypertensive patients and individuals at risk for developing hypertension should include assessment of overall cardiovascular disease risk. Similarly, introduction of lifestyle modification strategies and drug therapies should address overall risk and not focus exclusively on hypertension.

RENAL PARENCHYMAL DISEASES

Virtually all disorders of the kidney may cause hypertension (Table 271-3), and renal disease is the most common cause of secondary hypertension. Hypertension is present in >80% of patients with chronic renal failure. In general, hypertension is more severe in glomerular diseases than in interstitial diseases such as chronic pyelonephritis. Conversely, hypertension may cause nephrosclerosis, and in some instances it may be difficult to determine whether hypertension or renal disease was the initial disorder. Proteinuria >1000 mg/d and an active urine sediment are indicative of primary renal disease. In either instance, the goals are to control blood pressure and retard the rate of progression of renal dysfunction.

RENOVASCULAR HYPERTENSION

Hypertension due to an occlusive lesion of a renal artery, renovascular hypertension, is a potentially curable form of hypertension. Two groups of patients are at risk for this disorder: older arteriosclerotic patients who have a plaque obstructing the renal artery, frequently at its origin, and patients with fibromuscular dysplasia. Atherosclerosis accounts for the large majority of patients with renovascular hypertension. Although fibromuscular dysplasia may occur at any age, it has a strong predilection for young white women. The lesions of fibromuscular dysplasia are frequently bilateral and, in contrast to atherosclerotic renovascular disease, tend to affect more distal portions of the renal artery.

Renovascular hypertension should be considered in patients with other evidence of atherosclerotic vascular disease. Severe or refractory hypertension, recent loss of hypertension control or recent onset of moderately

severe hypertension, and unexplained deterioration of renal function or deterioration of renal function associated with an ACE inhibitor should raise the possibility of renovascular hypertension. Approximately 50% of patients with renovascular hypertension have an abdominal or flank bruit, and the bruit is more likely to be hemodynamically significant if it lateralizes or extends throughout systole into diastole.

If renal artery stenosis is suspected and if the clinical condition warrants an intervention such as percutaneous transluminal renal angioplasty (PTRA), placement of a vascular endoprosthesis (stent), or surgical renal revascularization, imaging studies should be the next step in the evaluation. As a screening test, renal blood flow may be evaluated with a radionuclide [^{131}I]-orthoiodohippurate (OIH) scan, or glomerular filtration rate may be evaluated with a [^{99}mTc]-diethylenetriamine pentaacetic acid (DTPA) scan before and after a single dose of captopril (or another ACE inhibitor). In patients with normal, or nearly normal, renal function, a normal captopril renogram essentially excludes functionally significant renal artery stenosis; however, its usefulness is limited in patients with renal insufficiency (creatinine clearance <20 mL/min) or bilateral renal artery stenosis. Additional imaging studies are indicated if the scan is positive. Doppler ultrasound of the renal arteries produces reliable estimates of renal blood flow velocity and offers the opportunity to track a lesion over time. Positive studies usually are confirmed at angiography, whereas false-negative results occur frequently, particularly in obese patients. Gadolinium-contrast magnetic resonance angiography offers clear images of the proximal renal artery but may miss distal lesions. An advantage is the opportunity to image the renal arteries with an agent that is not nephrotoxic. Contrast arteriography remains the “gold standard” for evaluation and identification of renal artery lesions.

Some degree of renal artery obstruction may be observed in almost 50% of patients with atherosclerotic disease, and there are several approaches for evaluating the functional significance of such a lesion to predict the effect of vascular repair on blood pressure control and renal function. Each approach has varying degrees of sensitivity and specificity, and no single test is sufficiently reliable to determine a causal relationship between a renal artery lesion and hypertension. Functionally significant lesions generally occlude >70% of the lumen of the affected renal artery. On angiography, the presence of collateral vessels to the ischemic kidney suggests a functionally significant lesion. A lateralizing renal vein renin ratio (ratio >1.5 of affected side/contralateral side) has a 90% predictive value for a lesion that would respond to vascular repair; however, the false-negative rate for blood pressure control is 50–60%. Measurement of the pressure gradient across a renal artery lesion does not reliably predict the response to vascular repair.

In the final analysis, a decision concerning vascular repair vs medical therapy and the type of repair procedure should be individualized. If blood pressure is adequately controlled with medical therapy and renal function remains stable, there may be little impetus to pursue an evaluation for renal artery stenosis. Several recent randomized clinical trials have found that PTRA with stent placement in patients with arteriosclerotic renal artery stenosis offers no advantages to medical therapy in reducing cardiovascular events and mortality or in preserving kidney function. In addition, 5 of 7 trials found similar blood pressure control in the two groups of patients. These results suggest that laboratory evaluation for renal artery stenosis and stent placement should be considered only in those arteriosclerotic patients in whom medical therapy fails to control blood pressure or preserve renal function. Patients with long-standing hypertension, advanced renal insufficiency, or diabetes mellitus are less likely to benefit from renal vascular repair. The most effective medical therapies include an ACE inhibitor or an angiotensin II receptor blocker; however, these agents decrease glomerular filtration rate in a stenotic kidney owing to efferent renal arteriolar dilation. In the presence of bilateral renal artery stenosis or renal artery stenosis to a solitary kidney, progressive renal insufficiency may result from the use of these agents. Importantly, the renal insufficiency is generally reversible after discontinuation of the offending drug. Patients with fibromuscular disease have more favorable outcomes with vascular repair than do patients with atherosclerotic lesions, presumably owing to their younger age, shorter duration of hypertension, and less systemic disease.

PRIMARY ALDOSTERONISM

Excess aldosterone production due to primary aldosteronism is a potentially curable form of hypertension. In patients with primary aldosteronism, increased aldosterone production is independent of the renin-angiotensin system, and the consequences are sodium retention, hypertension, hypokalemia, and low PRA. The reported prevalence of this disorder varies from <2 to ~15% of hypertensive individuals. In part, this variation is related to the intensity of screening and the criteria for establishing the diagnosis.

History and physical examination provide little information about the diagnosis. The age at the time of diagnosis is generally the third through fifth decade. Hypertension is usually mild to moderate but occasionally may be severe; primary aldosteronism should be considered in all patients with refractory hypertension. Hypertension in these patients may be associated with glucose intolerance. Most patients are asymptomatic; however, infrequently, polyuria, polydipsia, paresthesias, or muscle weakness may be present as a consequence of hypokalemic alkalosis. Although aldosterone is a salt-retaining hormone, patients with primary aldosteronism rarely have edema. Renal dysfunction and cardiovascular disease are strikingly increased in patients with primary aldosteronism compared to those with primary hypertension.

In a hypertensive patient with unprovoked hypokalemia (i.e., unrelated to diuretics, vomiting, or diarrhea), the prevalence of primary aldosteronism approaches 40–50%. In patients on diuretics, serum potassium <3.1 mmol/L (<3.1 meq/L) also raises the possibility of primary aldosteronism; however, serum potassium is an insensitive and nonspecific screening test. Serum potassium is normal in ~25% of patients subsequently found to have an aldosterone-producing adenoma, and higher percentages of patients with other etiologies of primary aldosteronism are not hypokalemic.

The ratio of plasma aldosterone to PRA (PA/PRA) is a useful screening test. These measurements preferably are obtained in ambulatory patients in the morning. A ratio >30:1 in conjunction with a plasma aldosterone concentration >555 pmol/L (>20 ng/dL) reportedly has a sensitivity of 90% and a specificity of 91% for an aldosterone-producing adenoma. In a Mayo Clinic series, an aldosterone-producing adenoma subsequently was confirmed surgically in >90% of hypertensive patients with a PA/PRA ratio ≥20 and a plasma aldosterone concentration ≥415 pmol/L (≥15 ng/dL). There are, however, several caveats to interpreting the ratio. The cutoff for a “high” ratio is laboratory- and assay-dependent. Some anti-hypertensive agents may affect the ratio (e.g., aldosterone antagonists, angiotensin receptor antagonists, and ACE inhibitors may increase renin; aldosterone antagonists may increase aldosterone). Current recommendations are to withdraw aldosterone antagonists for at least 4–6 weeks before obtaining these measurements. Because aldosterone biosynthesis is potassium-dependent, hypokalemia should be corrected with oral potassium supplements prior to screening. With these caveats, the ratio has been reported to be useful as a screening test in measurements obtained with patients taking their usual antihypertensive medications except for aldosterone antagonists, which should be discontinued six weeks before testing. A high ratio in the absence of an elevated plasma aldosterone level is considerably less specific for primary aldosteronism since many patients with primary hypertension have low renin levels in this setting, particularly African Americans and elderly patients. In patients with renal insufficiency, the ratio may also be elevated because of decreased aldosterone clearance. In patients with an elevated PA/PRA ratio, the diagnosis of primary aldosteronism can be confirmed by demonstrating failure to suppress plasma aldosterone to <277 pmol/L (<10 ng/dL) after IV infusion of 2 L of isotonic saline over 4 h; post-saline infusion plasma aldosterone values between 138 and 277 pmol/L (5–10 ng/dL) are not determinant. Alternative confirmatory tests include failure to suppress aldosterone (based on test-specific criteria) in response to an oral NaCl load, fludrocortisone, or captopril.

Several sporadic and familial adrenal abnormalities may culminate in the syndrome of primary aldosteronism, and appropriate therapy depends on the specific etiology. The two most common causes of sporadic primary aldosteronism are an aldosterone-producing adenoma and bilateral adrenal hyperplasia. Together, they account for >90% of all patients with primary aldosteronism. The tumor is almost always unilateral, and most often measures <3 cm in diameter. Most of the remainder of these patients have bilateral adrenocortical hyperplasia (idiopathic hyperaldosteronism). Rarely, primary aldosteronism may be caused by an adrenal carcinoma or an ectopic malignancy, e.g., ovarian arrhenoblastoma. Most aldosterone-producing carcinomas, in contrast to adrenal adenomas and hyperplasia, produce excessive amounts of other adrenal steroids in addition to aldosterone. Functional differences in hormone secretion may assist in the diagnosis of adenoma vs hyperplasia. Aldosterone biosynthesis is more responsive to ACTH in patients with adenoma and more responsive to angiotensin in patients with hyperplasia. Consequently, patients with adenoma tend to have higher plasma aldosterone in the early morning that decreases during the day, reflecting the diurnal rhythm of ACTH, whereas plasma aldosterone tends to increase with upright posture in patients with hyperplasia, reflecting the normal postural response of the renin-angiotensin-aldosterone axis. However, there is overlap in the ability of these measurements to discriminate between adenoma and hyperplasia. Rare familial forms of primary aldosteronism include glucocorticoid-remediable primary aldosteronism and familial aldosteronism types II and III. Genetic testing may assist in the diagnosis of these familial disorders.

Adrenal computed tomography (CT) should be carried out in all patients diagnosed with primary aldosteronism. High-resolution CT may identify tumors as small as 0.3 cm and is positive for an adrenal tumor 90% of the time. If the CT is not diagnostic, an adenoma may be detected by adrenal scintigraphy with 6 β-[I^{131}] iodomethyl-19-norcholesterol after dexamethasone suppression (0.5 mg every 6 h for 7 days); however, this technique has decreased sensitivity for adenomas <1.5 cm.

When carried out by an experienced radiologist, bilateral adrenal venous sampling for measurement of plasma aldosterone is the most accurate means of differentiating unilateral from bilateral forms of primary aldosteronism. The sensitivity and specificity of adrenal venous sampling (95 and 100%, respectively) for detecting unilateral aldosterone hypersecretion are superior to those of adrenal CT; success rates are 90–96%, and complication rates are <2.5%. One frequently used protocol involves sampling for aldosterone and cortisol levels in response to ACTH stimulation. An ipsilateral/contralateral aldosterone ratio >4, with symmetric ACTH-stimulated cortisol levels, is indicative of unilateral aldosterone production.

Hypertension generally is responsive to surgery in patients with adenoma but not in patients with bilateral adrenal hyperplasia. Unilateral adrenalectomy, often done via a laparoscopic approach, is curative in 40–70% of patients with an adenoma. Transient hypoaldosteronism may occur up to 3 months postoperatively, resulting in hyperkalemia, which should be treated with potassium-wasting diuretics and with fludrocortisone, if needed. Patients with bilateral hyperplasia should be treated medically. The drug regimen for these patients, as well as for patients with an adenoma who are poor surgical candidates, should include an aldosterone antagonist and, if necessary, other potassium-sparing diuretics.

Glucocorticoid-remediable hyperaldosteronism is a rare, monogenic autosomal dominant disorder characterized by moderate to severe hypertension, often occurring at an early age. These patients may have a family history of hemorrhagic stroke at a young age. Hypokalemia is usually mild or absent. Normally, angiotensin II stimulates aldosterone production by the adrenal zona glomerulosa, whereas ACTH stimulates cortisol production in the zona fasciculata. Owing to a chimeric gene on chromosome 8, ACTH also regulates aldosterone secretion by the zona fasciculata in patients with glucocorticoid-remediable hyperaldosteronism. The consequence is overproduction in the zona fasciculata of both aldosterone and hybrid steroids (18-hydroxycortisol and 18-oxocortisol) due to oxidation of cortisol. The diagnosis may be established by urine excretion rates of these hybrid steroids that are 20–30 times normal or by direct genetic testing. Therapeutically, suppression of ACTH with low-dose glucocorticoids corrects the hyperaldosteronism, hypertension, and hypokalemia. Aldosterone antagonists are also therapeutic options. Patients with familial aldosteronism types II and III are treated with aldosterone antagonists or adrenalectomy.

CUSHING'S SYNDROME

(See also Chap. 379) Cushing's syndrome is related to excess cortisol production due either to excess ACTH secretion (from a pituitary tumor or an ectopic tumor) or to ACTH-independent adrenal production of cortisol. Hypertension occurs in 75–80% of patients with Cushing's syndrome. The mechanism of hypertension may be related to stimulation of mineralocorticoid receptors by cortisol and increased secretion of other adrenal steroids. If clinically suspected based on phenotypic characteristics, in patients not taking exogenous glucocorticoids, laboratory screening may be carried out with measurement of 24-h excretion rates of urine-free cortisol or an overnight dexamethasone-suppression test. Late night salivary cortisol is also a sensitive and convenient screening test. Further evaluation is required to confirm the diagnosis and identify the specific etiology of Cushing's syndrome. Appropriate therapy depends on the etiology.

PHEOCHROMOCYTOMA

(See also Chap. 380) Catecholamine-secreting tumors are located in the adrenal medulla (pheochromocytoma) or in extra-adrenal paraganglion tissue (paraganglioma) and account for hypertension in ~0.05% of patients. If unrecognized, pheochromocytoma may result in lethal cardiovascular consequences. Clinical manifestations, including hypertension, are primarily related to increased circulating catecholamines, although some of these tumors may secrete a number of other vasoactive substances. In a small percentage of patients, epinephrine is the predominant catecholamine secreted by the tumor, and these patients may present with hypotension rather than hypertension. The initial suspicion of the diagnosis is based on symptoms and/or the association of pheochromocytoma with other disorders **(Table 271-4)**.

Approximately 20% of pheochromocytomas are familial with autosomal dominant inheritance. Inherited pheochromocytomas may be associated with multiple endocrine neoplasia (MEN) type 2A and type 2B, von Hippel-Lindau disease, and neurofibromatosis (Table 271-4). Each of these syndromes is related to specific, identifiable germ-line mutations. Additionally, mutations of succinate dehydrogenase genes are associated with paraganglioma syndromes, generally characterized by head and neck paragangliomas. Laboratory testing consists of measuring catecholamines in either urine or plasma, e.g., 24-h urine metanephrine excretion or fractionated plasma-free metanephrines. The urine measurement is less sensitive but more specific. Genetic screening is available for evaluating patients and relatives suspected of harboring a pheochromocytoma associated with a familial syndrome. Surgical excision is the definitive treatment of pheochromocytoma and results in cure in ~90% of patients.

TABLE 271-4 Rare Mendelian Forms of Hypertension

Disease	Phenotype	Genetic Cause
Glucocorticoid-remediable hyperaldosteronism	Autosomal dominant Absent or mild hypokalemia	Chimeric 11β-hydroxylase/aldosterone gene on chromosome 8
17α-hydroxylase deficiency	Autosomal recessive Males: pseudohermaphroditism Females: primary amenorrhea, absent secondary sexual characteristics	Random mutations of the *CYP17* gene on chromosome 10
11β-hydroxylase deficiency	Autosomal recessive Masculinization	Mutations of the *CYP11B1* gene on chromosome 8q21-q22
11β-hydroxysteroid dehydrogenase deficiency (apparent mineralocorticoid excess syndrome)	Autosomal recessive Hypokalemia, low renin, low aldosterone	Mutations in the 11β-hydroxysteroid dehydrogenase gene
Liddle's syndrome	Autosomal dominant Hypokalemia, low renin, low aldosterone	Mutation subunits of the epithelial sodium channel *SCNN1B* and *SCNN1C* genes
Pseudohypoaldosteronism type II (Gordon's syndrome)	Autosomal dominant Hyperkalemia, normal glomerular filtration rate	Linkage to chromosomes 1q31-q42 and 17p11-q21
Hypertension exacerbated in pregnancy	Autosomal dominant Severe hypertension in early pregnancy	Missense mutation with substitution of leucine for serine at codon 810 (MR_{L810})
Polycystic kidney disease	Autosomal dominant Large cystic kidneys, renal failure, liver cysts, cerebral aneurysms, valvular heart disease	Mutations in the *PKD1* gene on chromosome 16 and *PKD2* gene on chromosome 4
Pheochromocytoma	Autosomal dominant (a) Multiple endocrine neoplasia, type 2A Medullary thyroid carcinoma, hyperparathyroidism	(a) Mutations in the *RET* protooncogene
	(b) Multiple endocrine neoplasia, type 2B Medullary thyroid carcinoma, mucosal neuromas, thickened corneal nerves, alimentary ganglioneuromatoses, marfanoid habitus	(b) Mutations in the *RET* protooncogene
	(c) von Hippel-Lindau disease Retinal angiomas, hemangioblastomas of the cerebellum and spinal cord, renal cell carcinoma	(c) Mutations in the *VHL* tumor-suppressor gene
	(d) Neurofibromatosis type 1 Multiple neurofibromas, café-au-lait spots	(d) Mutations in the *NF1* tumor-suppressor gene

MISCELLANEOUS CAUSES OF HYPERTENSION

Independent of obesity, hypertension occurs in >50% of individuals with *obstructive sleep apnea*. The severity of hypertension correlates with the severity of sleep apnea. Approximately 70% of patients with obstructive sleep apnea are obese. Hypertension related to obstructive sleep apnea also should be considered in patients with drug-resistant hypertension and patients with a history of snoring. The diagnosis can be confirmed by polysomnography. In obese patients, weight loss may alleviate or cure sleep apnea and related hypertension. Continuous positive airway pressure (CPAP) or bilevel positive airway pressure (BiPAP) administered during sleep is an effective therapy for obstructive sleep apnea. With CPAP or BiPAP, patients with apparently drug-resistant hypertension may be more responsive to antihypertensive agents.

Coarctation of the aorta is the most common congenital cardiovascular cause of hypertension **(Chap. 264)**. The incidence is 1–8 per 1000 live births. It is usually sporadic but occurs in 35% of children with Turner's syndrome. Even when the anatomic lesion is surgically corrected in infancy, up to 30% of patients develop subsequent hypertension and are at risk of accelerated coronary artery disease and cerebrovascular events. Patients with less severe lesions may not be diagnosed until young adulthood. Physical findings include diminished and delayed femoral pulses and a systolic pressure gradient between the right arm and the legs and, depending on the location of the coarctation, between the right and left arms. A blowing systolic murmur may be heard in the posterior left interscapular areas. The diagnosis may be confirmed by chest x-ray and transesophageal echocardiography. Therapeutic options include surgical repair and balloon angioplasty, with or without placement of an intravascular stent. Subsequently, many patients do not have a normal life expectancy but may have persistent hypertension, with death due to ischemic heart disease, cerebral hemorrhage, or aortic aneurysm.

Several additional endocrine disorders, including *thyroid diseases* and *acromegaly*, cause hypertension. Mild diastolic hypertension may be a consequence of hypothyroidism, whereas hyperthyroidism may result in systolic hypertension. *Hypercalcemia* of any etiology, the most common being primary hyperparathyroidism, may result in hypertension. Hypertension also may be related to a number of prescribed or over-the-counter *medications*.

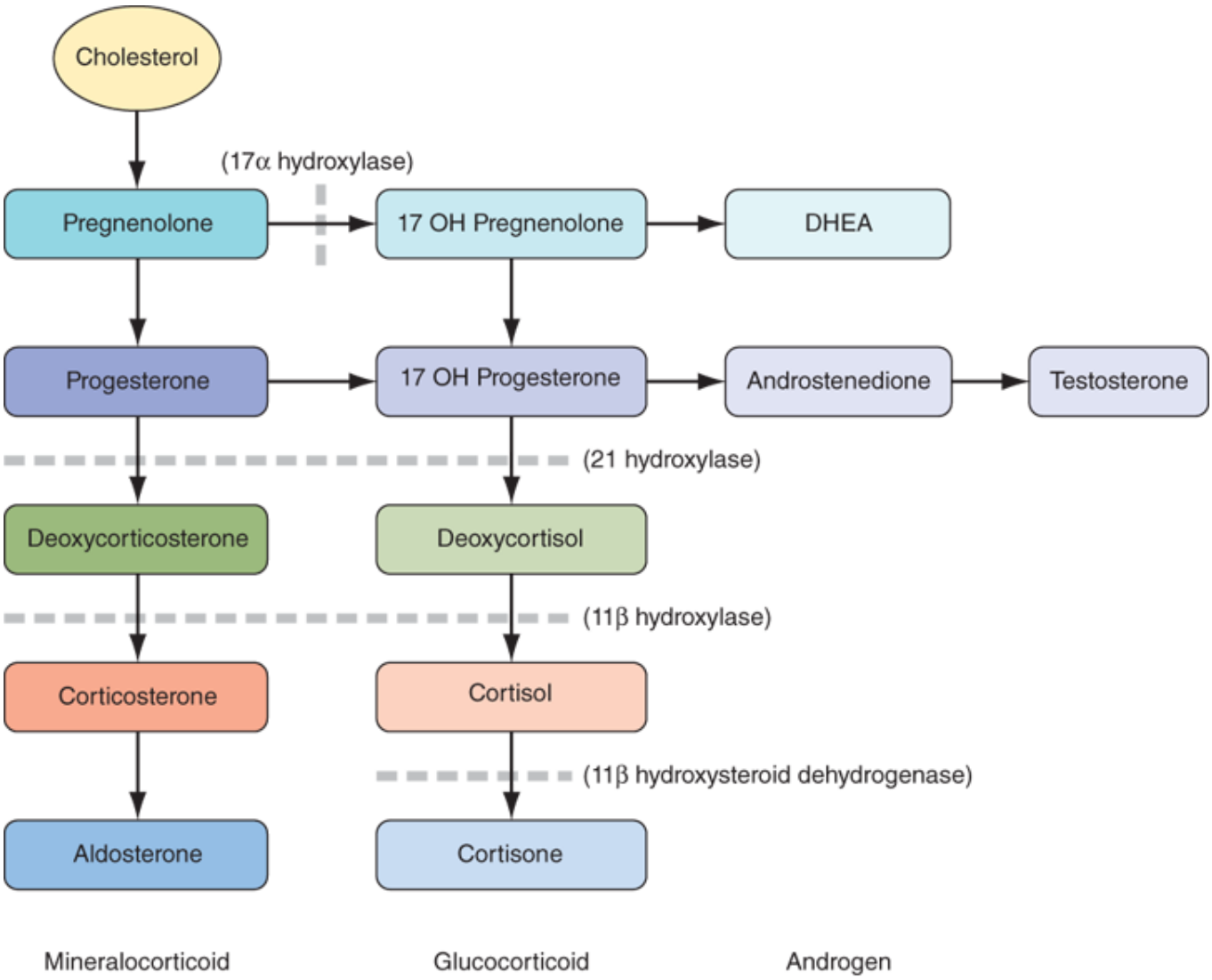
MONOGENIC HYPERTENSION

In addition to glucocorticoid-remediable primary aldosteronism, a number of rare forms of monogenic hypertension have been identified (Table 271–4). These disorders may be recognized by their characteristic phenotypes, and in many instances the diagnosis may be confirmed by genetic analysis. Several inherited defects in adrenal steroid biosynthesis and metabolism result in mineralocorticoid-induced hypertension and hypokalemia. In patients with a 17α-hydroxylase deficiency, synthesis of sex hormones and cortisol is decreased **(Fig. 271-3)**. Consequently, these individuals do not mature sexually; males may present with pseudohermaphroditism and females with primary amenorrhea and absent secondary sexual characteristics. Because cortisol-induced negative feedback on pituitary ACTH production is diminished, ACTH-stimulated adrenal steroid synthesis proximal to the enzymatic block is increased. Hypertension and hypokalemia are consequences of increased synthesis of mineralocorticoids proximal to the enzymatic block, particularly desoxycorticosterone. Increased steroid production and, hence, hypertension may be treated with low-dose glucocorticoids. An 11β-hydroxylase deficiency results in a salt-retaining adrenogenital syndrome that occurs in 1 in 100,000 live births. This enzymatic defect results in decreased cortisol synthesis, increased synthesis of mineralocorticoids (e.g., desoxycorticosterone), and shunting of steroid biosynthesis into the androgen pathway. In the severe form, the syndrome may present early in life, including the newborn period, with virilization and ambiguous genitalia in females and penile enlargement in males, or in older children as precocious puberty and short stature. Acne, hirsutism, and menstrual irregularities may be the presenting features when the disorder is first recognized in adolescence or early adulthood. Hypertension is less common in the late-onset forms. Patients with an 11β-hydroxysteroid dehydrogenase deficiency have an impaired capacity to metabolize cortisol to its inactive metabolite, cortisone, and hypertension is related to activation of mineralocorticoid receptors by cortisol. This defect may be inherited or acquired, due to licorice-containing glycyrrhizin acid. The same substance is present in the paste of several brands of chewing tobacco. The defect in Liddle's syndrome **(Chaps. 49 and 379)** results from constitutive activation of amiloride-sensitive ENaC on the distal renal tubule, resulting in excess sodium reabsorption; the syndrome is

ameliorated by amiloride. Hypertension exacerbated in pregnancy **(Chap. 466)** may be due to activation of the mineralocorticoid receptor by progesterone.

FIGURE 271-3

Adrenal enzymatic defects. DHEA, dehydroepiandrosterone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Hypertension HISTORY AND PHYSICAL

The initial assessment of a hypertensive patient should include a complete history and physical examination to confirm a diagnosis of hypertension, screen for other cardiovascular disease risk factors, screen for secondary causes of hypertension, identify cardiovascular consequences of hypertension and other comorbidities, assess blood pressure–related lifestyles, and determine the potential for intervention. Most patients with hypertension have no specific symptoms referable to their blood pressure elevation. **Table 271-5** lists salient features of the history and physical examination of the hypertensive patient.

Reliable measurements of blood pressure depend on attention to the details of the technique and conditions of the measurement. Proper training of observers, positioning of the patient, and selection of cuff size are essential. Owing to recent regulations preventing the use of mercury because of concerns about its potential

toxicity, most office measurements are made with aneroid sphygmomanometers or with oscillometric devices. These instruments should be calibrated periodically, and their accuracy confirmed.

LABORATORY TESTING

Table 271-6 lists recommended laboratory tests in the initial evaluation of hypertensive patients. Repeat measurements of renal function, serum electrolytes, fasting glucose, and lipids may be obtained after the introduction of a new antihypertensive agent and then annually or more frequently if clinically indicated. More extensive laboratory testing is appropriate for patients with apparent drug-resistant hypertension or when the clinical evaluation suggests a secondary form of hypertension.

TABLE 271-5 Relevant History and Physical

History
Duration of hypertension
Previous therapies: responses and side effects
Family history of hypertension and cardiovascular disease
Dietary and psychosocial history
Other risk factors: weight change, dyslipidemia, smoking, diabetes, physical inactivity
Evidence of secondary hypertension: history of renal disease; change in appearance; muscle weakness; spells of sweating, palpitations, tremor; erratic sleep, snoring, daytime somnolence; symptoms of hypo- or hyperthyroidism; use of agents that may increase blood pressure
Evidence of target organ damage: history of TIA, stroke, transient blindness; angina, myocardial infarction, congestive heart failure; sexual function
Other comorbidities
Physical
Body habitus
Blood pressure in both arms
Supine and standing blood pressures
Funduscopic examination of retina
Quality of femoral and pedal pulses
Vascular and abdominal bruits
Cardiac rate and rhythm
Signs of congestive heart failure
Signs of secondary hypertension

Abbreviation: TIA, transient ischemic attack.

TABLE 271-6 Basic Laboratory Tests for Initial Evaluation

System	Test
Renal	Microscopic urinalysis, albumin excretion, serum BUN and/or creatinine
Endocrine	Serum sodium, potassium, calcium, TSH
Metabolic	Fasting blood glucose, total cholesterol, HDL and LDL (often computed) cholesterol, triglycerides
Other	Hematocrit, electrocardiogram

Abbreviations: BUN, blood urea nitrogen; HDL, high-density lipoprotein; LDL, low-density lipoprotein; TSH, thyroid-stimulating hormone.

TREATMENT

TREATMENT Hypertension LIFESTYLE INTERVENTIONS

Implementation of lifestyles that favorably affect blood pressure has implications for both the prevention and the treatment of hypertension. Health-promoting lifestyle modifications are recommended for individuals with

prehypertension and as an adjunct to drug therapy in hypertensive individuals. These interventions should address overall cardiovascular disease risk. Although the impact of lifestyle interventions on blood pressure is more pronounced in persons with hypertension, in short-term trials, weight loss and reduction of dietary NaCl have been shown to prevent the development of hypertension. In hypertensive individuals, even if these interventions do not produce a sufficient reduction in blood pressure to avoid drug therapy, the number of medications or doses required for blood pressure control may be reduced. Dietary modifications that effectively lower blood pressure are weight loss, reduced NaCl intake, increased potassium intake, moderation of alcohol consumption, and an overall healthy dietary pattern **(Table 271-7)**.

Prevention and treatment of obesity are important for reducing blood pressure and cardiovascular disease risk. In short-term trials, even modest weight loss can lead to a reduction of blood pressure and an increase in insulin sensitivity. Average blood pressure reductions of 6.3/3.1 mmHg have been observed with a reduction in mean body weight of 9.2 kg. Regular physical activity facilitates weight loss, decreases blood pressure, and reduces the overall risk of cardiovascular disease. Blood pressure may be lowered by 30 min of moderately intense physical activity, such as brisk walking, 6–7 days a week, or by more intense, less frequent workouts.

There is individual variability in the sensitivity of blood pressure to NaCl, and this variability may have a genetic basis. Several genetic loci have been associated with NaCl sensitivity. Based on results of meta-analyses, lowering of blood pressure by limiting daily NaCl intake to 4.4–7.4 g (75–125 meq) results in blood pressure reductions of 3.7–4.9/0.9–2.9 mmHg in hypertensive individuals and lesser reductions in normotensive individuals. Results of randomized clinical trials on the impact of sodium reduction on the incidence of cardiovascular events are conflicting; however, for obvious practical reasons, such studies are challenging and often are not sufficiently powered to detect differences in cardiovascular endpoints. Although reduced salt intakes are generally recommended for both the prevention and treatment of hypertension, overly rigorous salt restriction may have adverse cardiovascular outcomes in diabetic patients and in patients with CHF aggressively treated with diuretics. Potassium and calcium supplementation have inconsistent, modest antihypertensive effects, and, independent of blood pressure, potassium supplementation may be associated with reduced stroke mortality. Consuming three or more alcoholic drinks per day (a standard drink contains ~14 g ethanol) is associated with higher blood pressures, and a reduction of alcohol consumption is associated with a reduction of blood pressure. In patients with advanced renal disease, dietary protein restriction may have a modest effect in mitigating renal damage by reducing the intrarenal transmission of systemic arterial pressure.

The DASH (Dietary Approaches to Stop Hypertension) trial convincingly demonstrated that over an 8-week period a diet high in fruits, vegetables, and low-fat dairy products lowers blood pressure in individuals with high-normal blood pressures or mild hypertension. Reduction of daily NaCl intake to <6 g (100 meq) augmented the effect of this diet on blood pressure. Fruits and vegetables are enriched sources of potassium, magnesium, and fiber, and dairy products are an important source of calcium.

PHARMACOLOGIC THERAPY

Lowering systolic blood pressure by 10–12 mmHg and diastolic blood pressure by 5–6 mmHg confers relative risk reductions of 35–40% for stroke and 12–16% for CHD within 5 years of the initiation of treatment. The risk of heart failure is reduced by >50%; although the benefit of blood pressure lowering on progression of renal failure is less apparent, hypertension control is the single most effective intervention for slowing the rate of progression of hypertension-related kidney disease.

There is considerable variation in individual responses to different classes of antihypertensive agents, and the magnitude of response to any single agent may be limited by activation of counter-regulatory mechanisms. Most available agents reduce systolic blood pressure by 7–13 mmHg and diastolic blood pressure by 4–8 mmHg when corrected for placebo effect. More often than not, combinations of agents, with complementary antihypertensive mechanisms, are required to achieve goal blood pressure reductions. Selection of antihypertensive agents and combinations of agents should be individualized, taking into account age, severity of hypertension, other cardiovascular disease risk factors, comorbid conditions, and practical considerations related to cost, side effects, and frequency of dosing **(Table 271-8)**.

Diuretics

Low-dose thiazide diuretics may be used alone or in combination with other antihypertensive drugs. Thiazides inhibit the Na^+/Cl^- pump in the distal convoluted tubule and hence increase sodium excretion. In the long term, they also may act as vasodilators. Thiazides are safe, efficacious, inexpensive, and reduce clinical events. They provide additive blood pressure–lowering effects when combined with beta blockers, angiotensin-converting enzyme inhibitors (ACEIs), or angiotensin receptor blockers (ARBs). In contrast, addition of a diuretic to a calcium channel blocker is less effective. Usual doses of hydrochlorothiazide range from 6.25 to 50 mg/d. Owing to an increased incidence of metabolic side effects (hypokalemia, insulin resistance, increased cholesterol), higher doses generally are not recommended. Chlorthalidone is a diuretic structurally similar to hydrochlorothiazide, and like hydrochlorothiazide, it blocks sodium-chloride cotransport in the early distal tubule. However, chlorthalidone has a longer half-life (40–60 h vs 9–15 h) and an antihypertensive potency ~1.5–2.0 times that of hydrochlorothiazide. Potassium loss is also greater with chlorthalidone. Two potassium-sparing diuretics, amiloride and triamterene, act by inhibiting ENaC in the distal nephron. These agents are weak antihypertensive agents but may be used in combination with a thiazide to protect against hypokalemia. The main pharmacologic target for loop diuretics is the Na^+-K^+-$2Cl^-$ cotransporter in the thick ascending limb of the loop of Henle. Loop diuretics generally are reserved for hypertensive patients with reduced glomerular filtration rates (reflected in serum creatinine >220 μmol/L [>2.5 mg/dL]), CHF, or sodium retention and edema for some other reason, such as treatment with a potent vasodilator, e.g., minoxidil.

Blockers of the Renin–Angiotensin System

ACEIs decrease the production of angiotensin II, increase bradykinin levels, and reduce sympathetic nervous system activity. ARBs provide selective blockade of AT_1 receptors, and the effect of angiotensin II on unblocked AT_2 receptors may augment their hypotensive effect. Both classes of agents are effective antihypertensive agents that may be used as monotherapy or in combination with diuretics, calcium antagonists, and alpha blocking agents. ACEIs and ARBs improve insulin action and ameliorate the adverse effects of diuretics on glucose metabolism. Although the overall impact on the incidence of diabetes is modest, compared with amlodipine (a calcium antagonist), valsartan (an ARB) has been shown to reduce the risk of developing diabetes in high-risk hypertensive patients. ACEI/ARB combinations are less effective in lowering blood pressure than is the case when either class of these agents is used in combination with other classes of agents. In patients with vascular disease or a high risk of diabetes, combination ACEI/ARB therapy has been associated with more adverse events (e.g., cardiovascular death, myocardial infarction, stroke, and hospitalization for heart failure) without increases in benefit.

Side effects of ACEIs and ARBs include functional renal insufficiency due to efferent renal arteriolar dilation in a kidney with a stenotic lesion of the renal artery. Additional predisposing conditions to renal insufficiency induced by these agents include dehydration, CHF, and use of nonsteroidal anti-inflammatory drugs. Dry cough occurs in ~15% of patients, and angioedema occurs in <1% of patients taking ACEIs. Angioedema occurs most commonly in individuals of Asian origin and more commonly in African Americans than in whites. Hyperkalemia due to hypoaldosteronism is an occasional side effect of both ACEIs and ARBs.

An alternative approach to blocking the renin-angiotensin system has recently been introduced into clinical practice for the treatment of hypertension: direct renin inhibitors. Blockade of the renin-angiotensin system is more complete with renin inhibitors than with ACEIs or ARBs. Aliskiren is the first of a class of oral, nonpeptide competitive inhibitors of the enzymatic activity of renin. Monotherapy with aliskiren seems to be as effective as an ACEI or ARB for lowering blood pressure, but not more effective. Further blood reductions may be achieved when aliskiren is used in combination with a thiazide diuretic or a calcium antagonist. Currently, aliskiren is not considered a first-line antihypertensive agent.

Aldosterone Antagonists

Spironolactone is a nonselective aldosterone antagonist that may be used alone or in combination with a thiazide diuretic. It may be a particularly effective agent in patients with low-renin primary hypertension, resistant hypertension, and primary aldosteronism. In patients with CHF, low-dose spironolactone reduces mortality and hospitalizations for heart failure when given in addition to conventional therapy with ACEIs, digoxin, and loop diuretics. Because spironolactone binds to progesterone and androgen receptors, side effects may include gynecomastia, impotence, and menstrual abnormalities. These side effects are circumvented by a newer agent, eplerenone, which is a selective aldosterone antagonist.

Beta Blockers

β-Adrenergic receptor blockers lower blood pressure by decreasing cardiac output owing to a reduction of heart rate and contractility. Other proposed mechanisms by which beta blockers lower blood pressure include a central nervous system effect and inhibition of renin release. Beta blockers are particularly effective in hypertensive patients with tachycardia, and their hypotensive potency is enhanced by co-administration with a diuretic. In lower doses, some beta blockers selectively inhibit cardiac β_1 receptors and have less influence on β_2 receptors on bronchial and vascular smooth muscle cells; however, there seems to be no difference in the antihypertensive potencies of cardioselective and nonselective beta blockers. Some beta blockers have intrinsic sympathomimetic activity, although it is uncertain whether this constitutes an overall advantage or disadvantage in cardiac therapy. Beta blockers without intrinsic sympathomimetic activity decrease the rate of sudden death, overall mortality, and recurrent myocardial infarction. In patients with CHF, beta blockers have been shown to reduce the risks of hospitalization and mortality. Overall, beta blockers may be less protective against cardiovascular and cerebrovascular endpoints, and some beta blockers may have less effect on central aortic pressure than other classes of antihypertensive agents. However, beta blockers remain appropriate therapy for hypertensive patients with concomitant heart disease and related comorbidities. Carvedilol and labetalol block both β receptors and peripheral α-adrenergic receptors. The potential advantages of combined β- and α-adrenergic blockade in treating hypertension remain to be determined. Nebivolol represents another class of cardioselective beta blockers that has additional vasodilator actions related to enhancement of nitric oxide activity. Whether this confers greater clinical effectiveness remains to be determined.

α-Adrenergic Blockers

Postsynaptic, selective α-adrenoreceptor antagonists lower blood pressure by decreasing peripheral vascular resistance. They are effective antihypertensive agents used either as monotherapy or in combination with other agents. However, in clinical trials of hypertensive patients, alpha blockade has not been shown to reduce cardiovascular morbidity and mortality or to provide as much protection against CHF as other classes of antihypertensive agents. These agents are also effective in treating lower urinary tract symptoms in men with prostatic hypertrophy. Nonselective α-adrenoreceptor antagonists bind to postsynaptic and presynaptic receptors and are used primarily for the management of patients with pheochromocytoma.

Sympatholytic Agents

Centrally acting α_2 sympathetic agonists decrease peripheral resistance by inhibiting sympathetic outflow. They may be particularly useful in patients with autonomic neuropathy who have wide variations in blood pressure due to baroreceptor denervation. Drawbacks include somnolence, dry mouth, and rebound hypertension on withdrawal. Peripheral sympatholytics decrease peripheral resistance and venous constriction by depleting nerve terminal norepinephrine. Although they are potentially effective antihypertensive agents, their usefulness is limited by orthostatic hypotension, sexual dysfunction, and numerous drug-drug interactions. Rebound hypertension is another concern with abrupt cessation of drugs with a short half-life.

Calcium Channel Blockers

Calcium antagonists reduce vascular resistance through L-channel blockade, which reduces intracellular calcium and blunts vasoconstriction. This is a heterogeneous group of agents that includes drugs in the following three classes: phenylalkylamines (verapamil), benzothiazepines (diltiazem), and 1,4-dihydropyridines (nifedipine-like). Used alone and in combination with other agents (ACEIs, beta blockers, α_1-adrenergic blockers), calcium antagonists effectively lower blood pressure; however, it is unclear if adding a diuretic to a calcium blocker results in a further lowering of blood pressure. Side effects of flushing, headache, and edema with dihydropyridine use are related to their potencies as arteriolar dilators; edema is due to an increase in transcapillary pressure gradients, not to net salt and water retention.

Direct Vasodilators

Direct vasodilators decrease peripheral resistance and concomitantly activate mechanisms that defend arterial pressure, notably the sympathetic nervous system, the renin-angiotensin-aldosterone system, and

sodium retention. Usually, they are not considered first-line agents but are most effective when added to a combination that includes a diuretic and a beta blocker. Hydralazine is a potent direct vasodilator that has antioxidant and nitric oxide–enhancing actions, and minoxidil is a particularly potent agent and is used most frequently in patients with renal insufficiency who are refractory to all other drugs. Hydralazine may induce a lupus-like syndrome, and side effects of minoxidil include hypertrichosis and pericardial effusion. Intravenous nitroprusside can be used to treat malignant hypertension and life-threatening left ventricular heart failure associated with elevated arterial pressure.

COMPARISONS OF ANTIHYPERTENSIVES

Based on pooling results from clinical trials, meta-analyses of the efficacy of different classes of antihypertensive agents suggest essentially equivalent blood pressure–lowering effects of the following six major classes of antihypertensive agents when used as monotherapy: thiazide diuretics, beta blockers, ACEIs, ARBs, calcium antagonists, and α_1 blockers. On average, standard doses of most antihypertensive agents reduce blood pressure by 8–10/4–7 mmHg; however, there may be subgroup differences in responsiveness. Younger patients may be more responsive to beta blockers and ACEIs, whereas patients aged >50 years may be more responsive to diuretics and calcium antagonists. There is a limited relationship between plasma renin and blood pressure response. Patients with high-renin hypertension may be more responsive to ACEIs and ARBs than to other classes of agents, whereas patients with low-renin hypertension are more responsive to diuretics and calcium antagonists. Hypertensive African Americans tend to have low renin and may require higher doses of ACEIs and ARBs than whites for optimal blood pressure control, although this difference is abolished when these agents are combined with a diuretic. Beta blockers also appear to be less effective than thiazide diuretics in African Americans than in non-African Americans. Early pharmacogenetic studies, utilizing a candidate gene approach, genome-wide scans, or integrated metabolomic and genetic profiles, have shown associations of gene polymorphisms with blood pressure responsiveness to specific antihypertensive drugs. However, the reported effects have generally been too small to affect clinical decisions, and associated polymorphisms remain to be confirmed. Currently, in practical terms, the presence of comorbidities often influences the selection of antihypertensive agents.

A meta-analysis of >30 randomized trials of blood pressure–lowering therapy indicates that for a given reduction in blood pressure, the major drug classes seem to produce similar overall net effects on total cardiovascular events. In both nondiabetic and diabetic hypertensive patients, most trials have failed to show significant differences in cardiovascular outcomes with different drug regimens as long as equivalent decreases in blood pressure were achieved. For example, the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) demonstrated that the occurrence of CHD and nonfatal myocardial infarction, as well as overall mortality, was virtually identical in hypertensive patients treated with either an ACEI (lisinopril), a diuretic (chlorthalidone), or a calcium antagonist (amlodipine). However, there is some evidence that beta blockers are inferior to other classes of agents for prevention of cardiovascular events, stroke, renal failure, and all-cause mortality, whereas calcium channel blockers may be inferior and diuretics but superior to other classes of agents for the prevention of heart failure.

However, in specific patient groups, ACEIs may have particular advantages, beyond that of blood pressure control, in reducing cardiovascular and renal outcomes. ACEIs and ARBs decrease intraglomerular pressure and proteinuria and may retard the rate of progression of renal insufficiency, not totally accounted for by their hypotensive effects, in both diabetic and nondiabetic renal diseases. In patients with type 2 diabetes, treatment with an ACEI, an ARB, or aliskiren decreases proteinuria and delays the progression of renal disease. In experimental models of hypertension and diabetes, renal protection with aliskiren is comparable to that with ACEIs and ARBs. However, in patients with type 2 diabetes, addition of aliskiren to an ACEI provides no additional protection against cardiovascular or renal disease and may be associated with more adverse outcomes. Among African Americans with hypertension-related renal disease, ACEIs appear to be more effective than beta blockers or dihydropyridine calcium channel blockers in slowing, although not preventing, the decline of glomerular filtration rate. The renoprotective effect of these renin-angiotensin blockers, compared with other antihypertensive drugs, is less obvious at lower blood pressures. In most patients with hypertension and heart failure due to systolic and/or diastolic dysfunction, the use of diuretics, ACEIs or ARBs, and beta blockers is recommended to improve survival. Independent of blood pressure, in both hypertensive and normotensive individuals, ACEIs attenuate the development of left ventricular hypertrophy, improve symptomatology and risk of death from CHF, and reduce morbidity and mortality rates in post-myocardial infarction patients. Similar benefits in cardiovascular morbidity and mortality rates in

patients with CHF have been observed with the use of ARBs. ACEIs provide better coronary protection than do calcium channel blockers, whereas calcium channel blockers provide more stroke protection than do either ACEIs or beta blockers. Results of a large, double-blind, prospective clinical trial (Avoiding Cardiovascular Events through Combination Therapy in Patients Living with Systolic Hypertension [ACCOMPLISH Trial]) indicated that combination treatment with an ACEI (benazepril) plus a calcium antagonist (amlodipine) was superior to treatment with the ACEI plus a diuretic (hydrochlorothiazide) in reducing the risk of cardiovascular events and death among high-risk patients with hypertension. However, the combination of an ACEI and a diuretic has recently been shown to produce major reductions in morbidity and mortality in the very elderly.

After a stroke, combination therapy with an ACEI and a diuretic, but not with an ARB, has been reported to reduce the rate of recurrent stroke. Some of these apparent differences may reflect differences in trial design and/or patient groups.

There has been a recent resurgence of interest in two non-pharmacologic, antihypertensive therapies that interrupt sympathetic outflow: (1) device-based carotid baroreflex activation by electrical stimulation of the carotid sinus; and (2) endovascular radiofrequency ablation of the renal sympathetic nerves. Both have been suggested as potential options for resistant hypertension. Whereas renal denervation is a minimally invasive procedure, carotid baroreceptor stimulation is a surgical procedure, usually performed under general anesthesia that currently involves implanting electrodes on both the right and left carotid arteries. Both interventions inhibit sympathetic drive and decrease blood pressure by increasing the capacity of the kidney to excrete sodium and by decreasing renin release. Sustained activation of the baroreflex most likely lowers blood pressure by other mechanisms as well; however, clinical experience with this intervention is limited. Enthusiasm for renal denervation has been questioned by the results of Simplicity 3, a randomized, prospective clinical trial, comparing bilateral renal denervation with a sham procedure in 535 patients with resistant hypertension. At the end of six months there was no benefit of renal artery denervation on both office and ambulatory systolic blood pressures, the trial’s primary endpoints. It remains to be seen whether these interventions will be adopted into clinical practice.

BLOOD PRESSURE GOALS OF ANTIHYPERTENSIVE THERAPY

Based on clinical trial data, the maximum protection against combined cardiovascular endpoints is achieved with pressures <135–140 mmHg for systolic blood pressure and <80–85 mmHg for diastolic blood pressure; however, treatment has not reduced cardiovascular disease risk to the level in non-hypertensive individuals. According to a recent meta-analysis, the magnitude of the proportional reduction of cardiovascular events is broadly consistent regardless of baseline co-morbidity, although the absolute benefit of blood pressure reduction is greater among individuals with the highest risk for cardiovascular events.

The degree of benefit derived from antihypertensive agents is related to the magnitude of the blood pressure reduction. Guidelines establishing blood pressure targets for hypertension control continue to evolve. An intensive blood pressure lowering strategy is superior to a less intensive strategy for prevention of stroke and myocardial infarction. For example, the SPRINT trial studied 9361 subjects aged >50 years at increased risk for cardiovascular evens. Intensive blood pressure control (systolic blood pressure <120 mmHg) reduced the risk of cardiovascular events and mortality by 25% compared with less intensive control (systolic blood pressure 135–139 mmHg). More intense control may also be associated with a higher incidence of adverse events (e.g., syncope, electrolyte abnormalities, deterioration of renal function), and recent studies suggest that the benefits of intensive blood-pressure lowering outweigh the risks. Nevertheless, the absolute impact of more intensive control is relatively small. In the final analysis, patients need to be carefully monitored, and clinical decision making should be individualized.

In diabetic patients, effective blood pressure control reduces the risk of cardiovascular events and death as well as the risk for microvascular disease (nephropathy, retinopathy). Various guidelines have been recommended for hypertension control in patients with type 2 diabetes (<140/90, <140/85, <130/80). One widely cited Action to Control Cardiovascular Risk in Diabetes clinical trial (ACCORD) failed to find superiority of intensive blood pressure lowering (<120 mmHg) over standard blood pressure control (<140 mmHg) in reducing the risk of the study’s primary outcome (a composite endpoint of myocardial infarction, stroke, and cardiovascular death) in diabetic patients. However, that trial did demonstrate a significant reduction of stroke and left ventricular hypertrophy with more intensive therapy.

In patients with chronic renal insufficiency, a small, non-progressive increase in the serum creatinine concentration may occur with intensive blood pressure lowering. This generally reflects a hemodynamic response, not structural renal injury, indicating that intraglomerular pressure has been reduced. Blood pressure control should not be allowed to deteriorate in order to prevent the modest creatinine rise. Among older patients with isolated systolic hypertension, further lowering of diastolic blood pressure does not result in harm. However, relatively little information is available concerning the risk-versus-benefit ratio of intensive antihypertensive therapy in individuals >80 years of age, and in this population, gradual blood pressure reduction to a less aggressive target level of control may be appropriate (e.g., 130–150 mmHg).

To achieve recommended blood pressure goals, the majority of individuals with hypertension will require treatment with more than one drug. Three or more drugs frequently are needed in patients with diabetes and renal insufficiency. For most agents, reduction of blood pressure at half-standard doses is only ~20% less than at standard doses. Appropriate combinations of agents at these lower doses may have additive or almost additive effects on blood pressure with a lower incidence of side effects.

The term *resistant hypertension* refers to patients with blood pressures persistently >140/90 mmHg despite taking three or more antihypertensive agents, including a diuretic. Resistant or difficult-to-control hypertension is more common in patients aged >60 years than in younger patients. Resistant hypertension may be related to “pseudoresistance” (high office blood pressures and lower home blood pressures), nonadherence to therapy, identifiable causes of hypertension (including obesity and excessive alcohol intake), and the use of any of a number of nonprescription and prescription drugs (Table 271-3). Rarely, in older patients, pseudohypertension may be related to the inability to measure blood pressure accurately in severely sclerotic arteries. This condition is suggested if the radial pulse remains palpable despite occlusion of the brachial artery by the cuff (Osler maneuver). The actual blood pressure can be determined by direct intra-arterial measurement. Evaluation of patients with resistant hypertension might include home blood pressure monitoring to determine if office blood pressures are representative of the usual blood pressure. A more extensive evaluation for a secondary form of hypertension should be undertaken if no other explanation for hypertension resistance becomes apparent.

HYPERTENSIVE EMERGENCIES

Probably due to the widespread availability of antihypertensive therapy, in the United States there has been a decline in the numbers of patients presenting with “crisis levels” of blood pressure. Most patients who present with severe hypertension are chronically hypertensive, and in the absence of acute end organ damage, precipitous lowering of blood pressure may result in significant morbidity and should be avoided. The key to successful management of severe hypertension is to differentiate hypertensive crises from hypertensive urgencies. The degree of target organ damage, rather than the level of blood pressure alone, determines the rapidity with which blood pressure should be lowered. **Tables 271-9 and 271-10** list a number of hypertension-related emergencies and recommended therapies.

Malignant hypertension is a syndrome associated with an abrupt increase of blood pressure in a patient with underlying hypertension or related to the sudden onset of hypertension in a previously normotensive individual. The absolute level of blood pressure is not as important as its rate of rise. Pathologically, the syndrome is associated with diffuse necrotizing vasculitis, arteriolar thrombi, and fibrin deposition in arteriolar walls. Fibrinoid necrosis has been observed in arterioles of kidney, brain, retina, and other organs. Clinically, the syndrome is recognized by progressive retinopathy (arteriolar spasm, hemorrhages, exudates, and papilledema), deteriorating renal function with proteinuria, microangiopathic hemolytic anemia, and encephalopathy. Historic inquiry should include questions about the use of monoamine oxidase inhibitors and recreational drugs (e.g., cocaine, amphetamines).

Although blood pressure should be lowered rapidly in patients with hypertensive encephalopathy, there are inherent risks of overly aggressive therapy. In hypertensive individuals, the upper and lower limits of autoregulation of cerebral blood flow are shifted to higher levels of arterial pressure, and rapid lowering of blood pressure to below the lower limit of autoregulation may precipitate cerebral ischemia or infarction as a consequence of decreased cerebral blood flow. Renal and coronary blood flows also may decrease with overly aggressive acute therapy. The initial goal of therapy is to reduce mean arterial blood pressure by no more than 25% within minutes to 2 h or to a blood pressure in the range of 160/100–110 mmHg. This may be accomplished with IV nitroprusside, a short-acting vasodilator with a rapid onset of action that allows for

minute-to-minute control of blood pressure. Parenteral labetalol and nicardipine are also effective agents for the treatment of hypertensive encephalopathy.

In patients with malignant hypertension without encephalopathy or another catastrophic event, it is preferable to reduce blood pressure over hours or longer rather than minutes. This goal may effectively be achieved initially with frequent dosing of short-acting oral agents such as captopril, clonidine, and labetalol.

Acute, transient blood pressure elevations that last days to weeks frequently occur after thrombotic and hemorrhagic strokes. Autoregulation of cerebral blood flow is impaired in ischemic cerebral tissue, and higher arterial pressures may be required to maintain cerebral blood flow. Although specific blood pressure targets have not been defined for patients with acute cerebrovascular events, aggressive reductions of blood pressure are to be avoided. With the increasing availability of improved methods for measuring cerebral blood flow (using CT technology), studies are in progress to evaluate the effects of different classes of antihypertensive agents on both blood pressure and cerebral blood flow after an acute stroke. Currently, in the absence of other indications for acute therapy, for patients with cerebral infarction who are not candidates for thrombolytic therapy, one recommended guideline is to institute antihypertensive therapy only for patients with a systolic blood pressure >220 mmHg or a diastolic blood pressure >130 mmHg. If thrombolytic therapy is to be used, the recommended goal blood pressure is <185 mmHg systolic pressure and <110 mmHg diastolic pressure. In patients with hemorrhagic stroke, there is no consistent evidence that acute reductions of systolic blood pressure to a more aggressive target than 140–179 mmHg improves functional outcome. The management of hypertension after subarachnoid hemorrhage is controversial. Cautious reduction of blood pressure is indicated if mean arterial pressure is >130 mmHg.

In addition to pheochromocytoma, an adrenergic crisis due to catecholamine excess may be related to cocaine or amphetamine overdose, clonidine withdrawal, acute spinal cord injuries, and an interaction of tyramine-containing compounds with monoamine oxidase inhibitors. These patients may be treated with phentolamine or nitroprusside.

Treatment of hypertension in patients with acute aortic dissection is discussed in Chap. 274, and treatment of hypertension in pregnancy is discussed in Chap. 466.

TABLE 271-7 Lifestyle Modifications to Manage Hypertension

Weight reduction	Attain and maintain BMI <25 kg/m^2
Dietary salt reduction	<6 g NaCl/d
Adapt DASH-type dietary plan	Diet rich in fruits, vegetables, and low-fat dairy products with reduced content of saturated and total fat
Moderation of alcohol consumption	For those who drink alcohol, consume ≤2 drinks/d in men and ≤1 drink/d in women
Physical activity	Regular aerobic activity, e.g., brisk walking for 30 min/d

Abbreviations: BMI, body mass index; DASH, Dietary Approaches to Stop Hypertension (trial).

TABLE 271-8 Examples of Oral Drugs Used in Treatment of Hypertension

Drug Class	Examples	Usual Total Daily Dose[a] (Dosing Frequency/Day)	Other Indications	Contraindications/Cautions
Diuretics				
Thiazides	Hydrochlorothiazide	6.25–50 mg (1–2)		Diabetes, dyslipidemia, hyperuricemia, gout, hypokalemia
	Chlorthalidone	25–50 mg (1)		
Loop diuretics	Furosemide	40–80 mg (2–3)	CHF due to systolic dysfunction, renal failure	Diabetes, dyslipidemia, hyperuricemia, gout, hypokalemia
	Ethacrynic acid	50–100 mg (2–3)		
Aldosterone antagonists	Spironolactone	25–100 mg (1–2)	CHF due to systolic dysfunction, primary aldosteronism	Renal failure, hyperkalemia
	Eplerenone	50–100 mg (1–2)		
K^+ retaining	Amiloride	5–10 mg (1–2)		Renal failure, hyperkalemia
	Triamterene	50–100 mg (1–2)		
Beta blockers				
Cardioselective	Atenolol	25–100 mg (1)	Angina, CHF due to systolic dysfunction, post-MI, sinus tachycardia, ventricular tachyarrhythmias	Asthma, COPD, 2nd- or 3rd-degree heart block, sick-sinus syndrome
	Metoprolol	25–100 mg (1–2)		
Nonselective	Propranolol	40–160 mg (2)		
	Propranolol LA	60–180 (1)		
Combined alpha/beta	Labetalol	200–800 mg (2)		
	Carvedilol	12.5–50 mg (2)		
Alpha antagonists				
Selective	Prazosin	2–20 mg (2–3)	Prostatism	
	Doxazosin	1–16 mg (1)		
	Terazosin	1–10 mg (1–2)		
Nonselective	Phenoxybenzamine	20–120 mg (2–3)	Pheochromocytoma	
Sympatholytics				
Central	Clonidine	0.1–0.6 mg (2)		
	Clonidine patch	0.1–0.3 mg (1/week)		
	Methyldopa	250–1000 mg (2)		
	Reserpine	0.05–0.25 mg (1)		
	Guanfacine	0.5–2 mg (1)		
ACE inhibitors	Captopril	25–200 mg (2)	Post-MI, coronary syndromes, CHF with low ejection fraction, nephropathy	Acute renal failure, bilateral renal artery stenosis, pregnancy, hyperkalemia
	Lisinopril	10–40 mg (1)		
	Ramipril	2.5–20 mg (1–2)		
Angiotensin II antagonists	Losartan	25–100 mg (1–2)	CHF with low ejection fraction,	Renal failure, bilateral renal artery stenosis, pregnancy, hyperkalemia

Drug Class	Examples	Usual Total Daily Dose[a] (Dosing Frequency/Day)	Other Indications	Contraindications/Cautions
	Valsartan	80–320 mg (1)	nephropathy, ACE inhibitor cough	
	Candesartan	2–32 mg (1–2)		
Renin inhibitors	Aliskiren	150–300 mg (1)	Diabetic nephropathy	Pregnancy
Calcium antagonists				
Dihydropyridines	Nifedipine (long-acting)	30–60 mg (1)		
Nondihydropyridines	Verapamil (long-acting)	120–360 mg (1–2)	Post-MI, supraventricular tachycardias, angina	2nd- or 3rd-degree heart block
	Diltiazem (long-acting)	180–420 mg (1)		
Direct vasodilators	Hydralazine	25–100 mg (2)		Severe coronary artery disease
	Minoxidil	2.5–80 mg (1–2)		

[a]At the initiation of therapy, lower doses may be preferable for elderly patients and for select combinations of antihypertensive agents.

Abbreviations: ACE, angiotensin-converting enzyme; CHF, congestive heart failure; COPD, chronic obstructive pulmonary disease; MI, myocardial infarction.

TABLE 271-9 Preferred Parenteral Drugs for Selected Hypertensive Emergencies

Hypertensive encephalopathy	Nitroprusside, nicardipine, labetalol
Malignant hypertension (when IV therapy is indicated)	Labetalol, nicardipine, nitroprusside, enalaprilat
Stroke	Nicardipine, labetalol, nitroprusside
Myocardial infarction/unstable angina	Nitroglycerin, nicardipine, labetalol, esmolol
Acute left ventricular failure	Nitroglycerin, enalaprilat, loop diuretics
Aortic dissection	Nitroprusside, esmolol, labetalol
Adrenergic crisis	Phentolamine, nitroprusside
Postoperative hypertension	Nitroglycerin, nitroprusside, labetalol, nicardipine
Preeclampsia/eclampsia of pregnancy	Hydralazine, labetalol, nicardipine

Source: Adapted from DG Vidt, in S Oparil, MA Weber (eds): *Hypertension*, 2nd ed. Philadelphia, Elsevier Saunders, 2005.

TABLE 271-10 Usual Intravenous Doses of Antihypertensive Agents Used in Hypertensive Emergencies[a]

Antihypertensive Agent	Intravenous Dose
Nitroprusside	Initial 0.3 (μg/kg)/min; usual 2–4 (μg/kg)/min; maximum 10 (μg/kg)/min for 10 min
Nicardipine	Initial 5 mg/h; titrate by 2.5 mg/h at 5–15 min intervals; max 15 mg/h
Labetalol	2 mg/min up to 300 mg *or* 20 mg over 2 min, then 40–80 mg at 10-min intervals up to 300 mg total
Enalaprilat	Usual 0.625–1.25 mg over 5 min every 6–8 h; maximum 5 mg/dose
Esmolol	Initial 80–500 μg/kg over 1 min, then 50–300 (μg/kg)/min
Phentolamine	5–15 mg bolus
Nitroglycerin	Initial 5 μg/min, then titrate by 5 μg/min at 3–5-min intervals; if no response is seen at 20 μg/min, incremental increases of 10–20 μg/min may be used
Hydralazine	10–50 mg at 30-min intervals

[a]Constant blood pressure monitoring is required. Start with the lowest dose. Subsequent doses and intervals of administration should be adjusted according to the blood pressure response and duration of action of the specific agent.

FURTHER READING

Ettehad D et al: Blood pressure lowering for prevention of cardiovascular disease and death: A systematic review and meta-analysis. Lancet 387:957, 2016. [PubMed: 26724178]
Feinberg AP, Fallin MD: Epigenetics at the crossroads of genes and the environment. JAMA 314:1129, 2015. [PubMed: 26372577]
Go AS et al: An effective approach to high blood pressure control: A science advisory from the American Heart Association, the American College of Cardiology, and the Centers for Disease Control and Prevention. Hypertension 63:878, 2014. [PubMed: 24243703]
Guo F et al: Trend in prevalence, awareness, management, and control of hypertension among United States adults, 1999 to 2010. J Am Coll Cardiol 60:599, 2012. [PubMed: 22796254]
Harrison DG et al: Inflammation, immunity and hypertension. Hypertension 57:132, 2011. [PubMed: 21149826]
Hughes TM, Sink TM: Hypertension and its role in cognitive function: Current evidence and challenges for the future. Am J Hypertension 29:149, 2016.
Oh YS et al: National Heart, Lung, and Blood Institute Working Group Report on Salt in Human Health and Sickness: Building on the current scientific evidence. Hypertension 68:281, 2016. [PubMed: 27324228]
Raman G et al: Comparative effectiveness of management strategies for renal artery stenosis: An updated systematic review. Ann Intern Med 165:635, 2016. [PubMed: 27536808]
Sprint Research Group: A randomized trial of intensive versus standard blood pressure control. N Engl J Med 373:2103, 2015. [PubMed: 26551272]

Chapter 272: Renovascular Disease

Stephen C. Textor

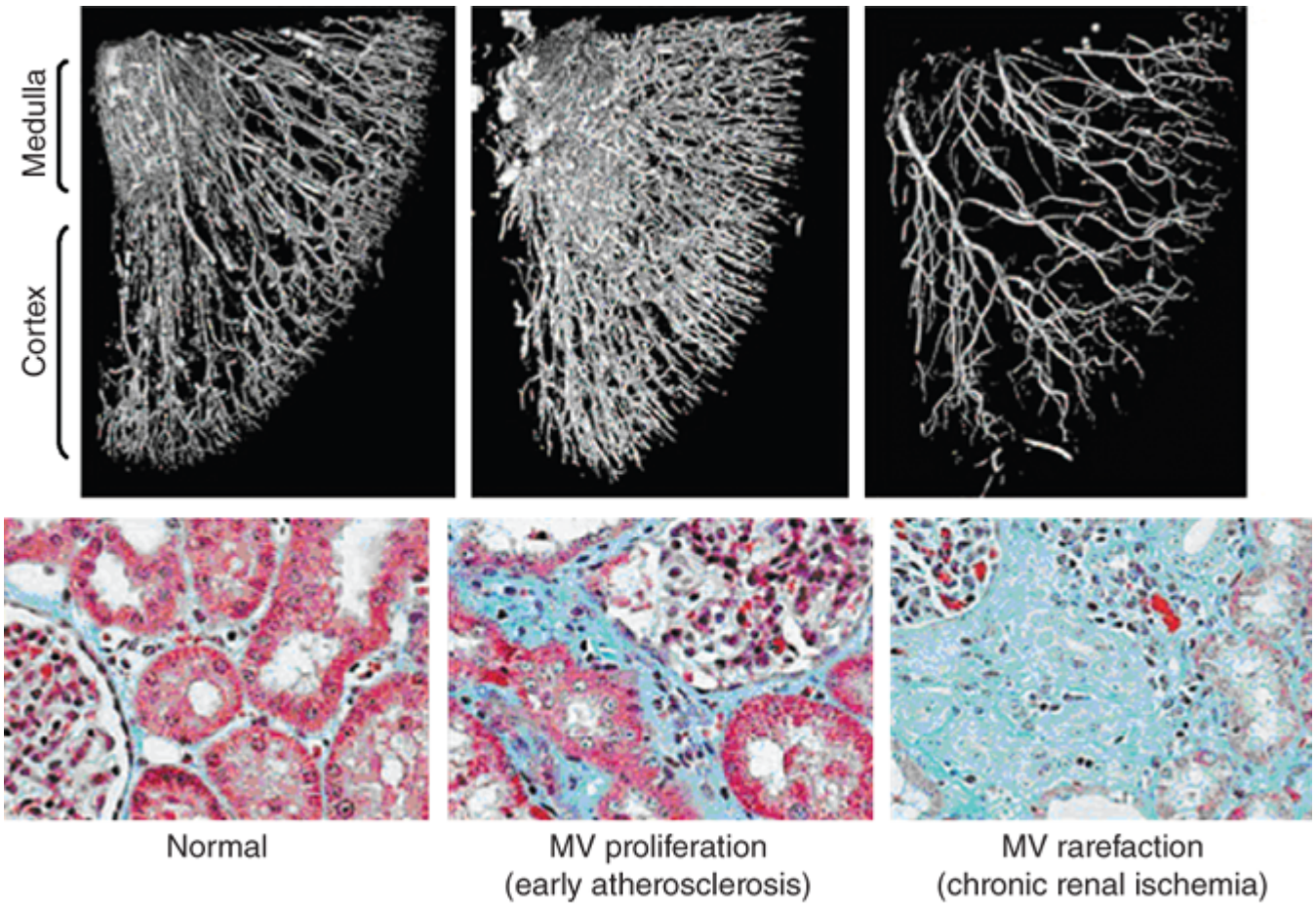
INTRODUCTION

The renal vasculature is unusually complex with rich arteriolar flow to the cortex in excess of metabolic requirements, consistent with its primary function as a filtering organ. After delivering blood to cortical glomeruli, the postglomerular circulation supplies deeper medullary segments that support energy-dependent solute transport at multiple levels of the renal tubule. These postglomerular vessels carry less blood, and high oxygen consumption leaves the deeper medullary regions at the margin of hypoxemia. Vascular disorders that commonly threaten the blood supply of the kidney include large-vessel atherosclerosis, fibromuscular diseases, and embolic disorders. **Microvascular injury, including inflammatory and primary hematologic disorders, is described in Chap. 311**.

The glomerular capillary endothelium shares susceptibility to oxidative stress, pressure injury, and inflammation with other vascular territories. Rates of urinary albumin excretion (UAE) are predictive of systemic atherosclerotic disease events. Increased UAE may develop years before cardiovascular events. UAE and the risk of cardiovascular events are both reduced with pharmacologic therapy such as statins. Experimental studies demonstrate functional changes and rarefaction of renal microvessels under conditions of accelerated atherosclerosis and/or compromise of proximal perfusion pressures with large-vessel disease **(Fig. 272-1)**.

FIGURE 272-1

Examples of micro-CT images from vessels defined by radiopaque casts injected into the renal vasculature. These illustrate the complex, dense cortical capillary network supplying the kidney cortex that can either proliferate or succumb to rarefaction under the influence of atherosclerosis and/or occlusive disease. Changes in blood supply are followed by tubulointerstitial fibrosis and loss of kidney function. MV, microvascular. *(From LO Lerman, AR Chade: Curr Opin Nephrol Hyper 18:160, 2009, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MACROVASCULAR DISEASE

Large-vessel renal artery occlusive disease can result from extrinsic compression of the vessel, intimal dissection, fibromuscular dysplasia (FMD), or, most commonly, atherosclerotic disease. Any disorder that reduces perfusion pressure to the kidney can activate mechanisms that tend to restore renal pressures at the expense of developing systemic hypertension. Because restoration of perfusion pressures can reverse these pathways, renal artery stenosis is considered a specifically treatable "secondary" cause of hypertension.

Renal artery stenosis is common and often has only minor hemodynamic effects. FMD is reported in 3–5% of normal subjects presenting as potential kidney donors without hypertension. It may present clinically with hypertension in younger individuals (between age 15 and 50), most often women. FMD does not often threaten kidney function, but sometimes produces total occlusion and can be associated with renal artery aneurysms. Atherosclerotic renal artery stenosis (ARAS) is common in the general population (6.8% of a community-based sample above age 65). The prevalence increases with age and for patients with other vascular conditions such as coronary artery disease (18–23%) and/or peripheral aortic or lower extremity disease (>30%). If untreated, ARAS progresses in nearly 50% of cases over a 5-year period, sometimes to total occlusion. Intensive treatment of arterial blood pressure and statin therapy appear to slow these rates and improve clinical outcomes.

Critical levels of stenosis lead to a reduction in perfusion pressure that activates the renin-angiotensin system, reduces sodium excretion, and activates sympathetic adrenergic pathways. These events lead to systemic hypertension characterized by angiotensin dependence in the early stages, widely varying pressures, loss of circadian blood pressure (BP) rhythms, and accelerated target organ injury,

including left ventricular hypertrophy and renal fibrosis. Renovascular hypertension can be treated with agents that block the renin-angiotensin system and other drugs that modify these pressor pathways. It can also be treated with restoration of renal blood flow by either endovascular or surgical revascularization. Most patients require continued antihypertensive drug therapy because revascularization alone rarely lowers BP to normal.

ARAS and systemic hypertension tend to affect both the post-stenotic and contralateral kidneys, reducing overall glomerular filtration rate (GFR) in ARAS. When kidney function is threatened by large-vessel disease primarily, it has been labeled ischemic nephropathy. Moderately reduced blood flow that develops gradually is associated with reduced GFR and limited oxygen consumption with preserved tissue oxygenation. Hence, kidney function often remains stable during medical therapy, sometimes for years. With more advanced disease, reductions in cortical perfusion and frank tissue hypoxia develop. Unlike FMD, ARAS develops in patients with other risk factors for atherosclerosis and is commonly superimposed upon preexisting small-vessel disease in the kidney resulting from hypertension, aging, and diabetes. Nearly 85% of patients considered for renal revascularization have stage 3–5 chronic kidney disease (CKD) with GFR <60 mL/min per 1.73 m^2. The presence of ARAS is a strong predictor of morbidity- and mortality-related cardiovascular events, independent of whether renal revascularization is undertaken.

Diagnostic approaches to renal artery stenosis depend partly on the specific clinical questions to be addressed. Noninvasive characterization of the renal vasculature may be achieved by several techniques, summarized in **Table 272-1**. Although activation of the renin-angiotensin system is a key step in developing renovascular hypertension, it is transient. Levels of renin activity are therefore subject to timing, the effects of drugs, and sodium intake, and do not reliably predict the response to vascular therapy. Renal artery velocities by Doppler ultrasound >200 cm/s generally predict hemodynamically important lesions (>60% vessel lumen occlusion), although some treatment trials require velocity >300 cm/s to avoid false positives. The renal resistive index has predictive value regarding the viability of the kidney. It remains operator- and institution-dependent, however. Captopril-enhanced renography has a strong negative predictive value when entirely normal. Magnetic resonance angiography (MRA) is now less often used, as gadolinium contrast has been associated with nephrogenic systemic fibrosis. Contrast-enhanced computed tomography (CT) with vascular reconstruction provides excellent vascular images and functional assessment, but carries a small risk of contrast toxicity.

TABLE 272-1 Summary of Imaging Modalities for Evaluating the Kidney Vasculature

Perfusion Studies to Assess Differential Renal Blood Flow			
Captopril renography with technetium ^{99m}Tc mertiatide (^{99m}Tc MAG3)	Captopril-mediated fall in filtration pressure amplifies differences in renal perfusion	Normal study excludes renovascular hypertension	Multiple limitations in patients with advanced atherosclerosis or creatinine >2.0 mg/dL (177 μmol/L)
Vascular Studies to Evaluate the Renal Arteries			
Duplex ultrasonography	Shows the renal arteries and measures flow velocity as a means of assessing the severity of stenosis	Inexpensive; widely available, suitable for follow-up studies	Heavily dependent on operator's experience; less useful than invasive angiography for the diagnosis of fibromuscular dysplasia and abnormalities in accessory renal arteries
Computed tomographic angiography	Shows the renal arteries and perirenal aorta	Provides excellent images; stents do not cause artifacts	Expensive, moderate volume of contrast required
Magnetic resonance angiography	Shows the renal arteries and perirenal aorta	Not nephrotoxic, but concerns for gadolinium toxicity exclude use in GFR <30 mL/min/1.73 m^2; provides excellent images	Expensive; gadolinium excluded in renal failure, unable to visualize stented vessels
Intraarterial angiography	Shows location and severity of vascular lesion	Considered "gold standard" for diagnosis of large-vessel disease, usually performed simultaneous with planned intervention	Expensive, associated hazard of atheroemboli, contrast toxicity, procedure-related complications, e.g., dissection

Abbreviation: GFR, glomerular filtration rate.

TREATMENT

TREATMENT Renal Artery Stenosis

While restoring renal blood flow and perfusion seems intuitively beneficial for high-grade occlusive lesions, revascularization procedures also pose hazards and expense. Patients with FMD are commonly younger females with otherwise normal vessels and a long life expectancy. These patients often respond well to percutaneous renal artery angioplasty. If BP can be controlled to goal levels and kidney function remains stable in patients with ARAS, it may be argued that medical therapy with follow-up for disease progression is equally effective. Multiple prospective randomized controlled trials have failed to identify compelling benefits for interventional procedures regarding short-term results of BP and renal function. Studies of cardiovascular outcomes including stroke, congestive heart failure, myocardial infarction, and end-stage renal failure, suggest a small mortality benefit for revascularized subjects without proteinuria. Medical therapy should include blockade of the renin-angiotensin system, attainment of goal BPs, cessation of tobacco, statins, and aspirin. Follow-up requires surveillance for progressive occlusion manifest by worsening renal function and/or loss of BP control. Renal revascularization is now often reserved for patients failing medical therapy or developing additional complications.

Techniques of renal revascularization are improving. With experienced operators, major complications occur in 5–9% of cases, including renal artery dissection, capsular perforation, hemorrhage, and occasional atheroembolic disease. Although not common, atheroembolic disease can be catastrophic and accelerate both hypertension and kidney failure, precisely the events that revascularization is intended to prevent. Although renal blood flow usually can be restored by endovascular stenting, recovery of renal function is limited to about 25% of cases, with no change in 50% and some deterioration evident in others. Patients with rapid loss of kidney function, sometimes associated with antihypertensive drug therapy, or with vascular disease affecting the entire functioning kidney mass are more likely to recover function after restoring blood flow. When hypertension is refractory to effective therapy, revascularization offers real benefits. **Table 272-2** summarizes currently accepted guidelines for considering renal revascularization.

TABLE 272-2 Clinical Factors That Determine the Role of Revascularization in Addition to Medical Therapy for Renal Artery Stenosis

Factors Favoring Medical Therapy and Revascularization for Renal Artery Stenosis

- Progressive decline in GFR during treatment of systemic hypertension
- Failure to achieve adequate blood pressure control with optimal medical therapy (medical failure)
- Rapid or recurrent decline in the GFR in association with a reduction in systemic pressure
- Decline in the GFR during therapy with ACE inhibitors or ARBs
- Recurrent congestive heart failure in a patient in whom the adequacy of left ventricular function does not explain a cause

Factors Favoring Medical Therapy and Surveillance of Renal Artery Disease

- Controlled blood pressure with stable renal function (e.g., stable renal insufficiency)
- Stable renal artery stenosis without progression on surveillance studies (e.g., serial duplex ultrasound)
- Very advanced age and/or limited life expectancy
- Extensive comorbidity that make revascularization too risky
- High risk for or previous experience with atheroembolic disease
- Other concomitant renal parenchymal diseases that cause progressive renal dysfunction (e.g., interstitial nephritis, diabetic nephropathy)

Abbreviations: ACE, angiotensin-converting enzyme; ARBs, angiotensin receptor blockers; GFR, glomerular filtration rate.

ATHEROEMBOLIC RENAL DISEASE

Emboli to the kidneys arise most frequently as a result of cholesterol crystals breaking free of atherosclerotic vascular plaque and lodging in downstream microvessels. Most clinical atheroembolic events follow angiographic procedures, often of the coronary vessels. It has been argued that nearly all vascular interventional procedures lead to plaque fracture and release of microemboli, but clinical manifestations develop only in a fraction of these. The incidence of clinical atheroemboli has been increasing with more vascular procedures and longer life spans. Atheroembolic renal disease is

suspected in >3% of elderly subjects with end-stage renal disease (ESRD) and is likely underdiagnosed. It is more frequent in males with a history of diabetes, hypertension, and ischemic cardiac disease. Atheroemboli in the kidney are strongly associated with aortic aneurysmal disease and renal artery stenosis. Most clinical cases can be linked to precipitating events, such as angiography, vascular surgery, anticoagulation with heparin, thrombolytic therapy, or trauma. Clinical manifestations of this syndrome commonly develop between 1 and 14 days after an inciting event and may continue to develop for weeks thereafter. Systemic embolic disease manifestations, such as fever, abdominal pain, and weight loss, are present in less than half of patients, although cutaneous manifestations including livedo reticularis and localized toe gangrene may be more common. Worsening hypertension and deteriorating kidney function are common, sometimes reaching a malignant phase. Progressive renal failure can occur and require dialytic support. These cases often develop after a stuttering onset over many weeks and have an ominous prognosis. Mortality rate after 1 year reaches 38%, and although some may eventually recover sufficiently to no longer require dialysis, many do not.

Beyond the clinical manifestations above, laboratory findings include rising creatinine, transient eosinophilia (60–80%), elevated sedimentation rate, and hypocomplementemia (15%). Establishing this diagnosis can be difficult and is often by exclusion. Definitive diagnosis depends on kidney biopsy demonstrating microvessel occlusion with cholesterol crystals that leave a “cleft” in the vessel. Biopsies obtained from patients undergoing surgical revascularization of the kidney indicate that silent cholesterol emboli are frequently present before any further manipulation is performed.

No effective therapy is available for atheroembolic disease once it has developed. Withdrawal of anticoagulation is recommended. Late recovery of kidney function after supportive measures sometimes occurs, and statin therapy may improve outcome. The role of embolic protection devices in the renal circulation is unclear, but a few prospective trials have failed to demonstrate major benefits. These devices are limited to distal protection during the endovascular procedure and offer no protection from embolic debris developing after removal.

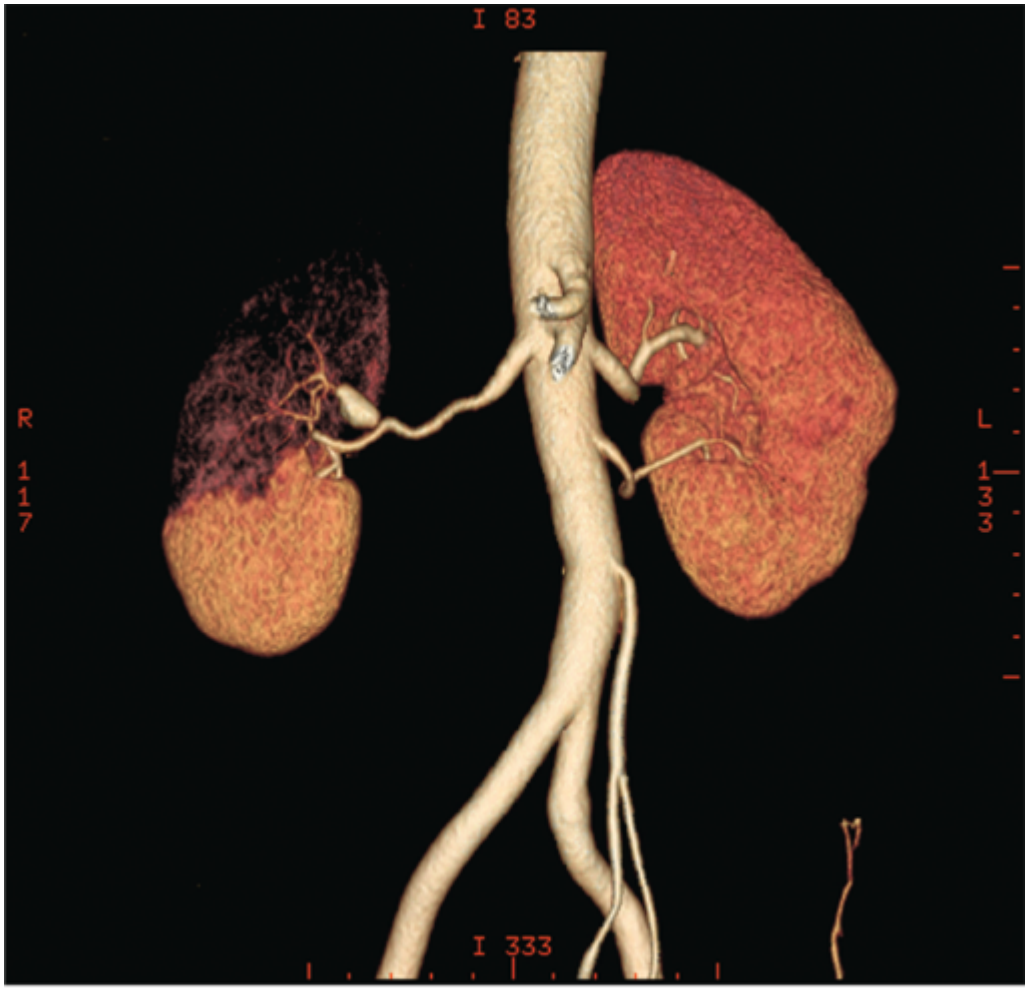
THROMBOEMBOLIC RENAL DISEASE

Thrombotic occlusion of renal vessels or branch arteries can lead to declining renal function and hypertension. It is difficult to diagnose and is often overlooked, especially in elderly patients. Thrombosis can develop as a result of local vessel abnormalities, such as local dissection, trauma, or inflammatory vasculitis. Local microdissections sometimes lead to patchy, transient areas of infarctions labeled “segmental arteriolar mediolysis.” Although hypercoagulability conditions sometimes present as renal artery thrombosis, this is rare. It can also derive from distant embolic events, e.g., the left atrium in patients with atrial fibrillation or from fat emboli originating from traumatized tissue, most commonly large bone fractures. Cardiac sources include vegetations from subacute bacterial endocarditis. Systemic emboli to the kidneys may also arise from the venous circulation if right-to-left shunting occurs, e.g., through a patent foramen ovale.

Clinical manifestations vary depending on the rapidity of onset and extent of occlusion. Acute arterial thrombosis may produce flank pain, fever, leukocytosis, nausea, and vomiting. If kidney infarction results, enzymes such as lactate dehydrogenase (LDH) rise to extreme levels. If both kidneys are affected, renal function will decline precipitously with a drop in urine output. If a single kidney is involved, renal functional changes may be minor. Hypertension related to sudden release of renin from ischemic tissue can develop rapidly, as long as some viable tissue in the “peri-infarct” border zone remains. If the infarct zone demarcates precisely, the rise in BP and renin activity may resolve. Diagnosis of renal infarction may be established by vascular imaging with MRI, CT angiography, or arteriography **(Fig. 272-2)**.

FIGURE 272-2

A. CT angiogram illustrating loss of circulation to the upper pole of the right kidney in a patient with fibromuscular disease and a renal artery aneurysm. Activation of the renin-angiotensin system produced rapidly developing hypertension. ***B.*** Angiogram illustrating high-grade renal artery stenosis affecting the left kidney. This lesion is often part of widespread atherosclerosis and sometimes is an extension of aortic plaque. This lesion develops in older individuals with preexisting atherosclerotic risk factors.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MANAGEMENT OF ARTERIAL THROMBOSIS OF THE KIDNEY

Options for interventions of newly detected arterial occlusion include surgical reconstruction, anticoagulation, thrombolytic therapy, endovascular procedures, and supportive care, particularly antihypertensive drug therapy. Application of these methods depends on the patient's overall condition, the precipitating factors (e.g., local trauma or systemic illness), the magnitude of renal tissue and function at risk, and the likelihood of recurrent events in the future. For unilateral disease, for example, arterial dissection with thrombosis, supportive care with anticoagulation may suffice. Acute, bilateral occlusion is potentially catastrophic, producing anuric renal failure. Depending on the precipitating event, surgical or thrombolytic therapies can sometimes restore kidney viability.

MICROVASCULAR INJURY ASSOCIATED WITH HYPERTENSION

ARTERIOLONEPHROSCLEROSIS

"Malignant" Hypertension

Although BP rises with age, it has long been recognized that some individuals develop rapidly progressive BP elevations with target organ injury including retinal hemorrhages, encephalopathy, and declining kidney function. Placebo arms during the controlled trials of hypertension therapy identified progression to severe levels in 20% of subjects over 5 years. If untreated, patients with target organ injury including papilledema and declining kidney function suffered mortality rates in excess of 50% over 6–12 months, hence the designation “malignant.” Postmortem studies of such patients identified vascular lesions, designated “fibrinoid necrosis,” with breakdown of the vessel wall, deposition of eosinophilic material including fibrin, and a perivascular cellular infiltrate. A separate lesion was identified in the larger interlobular arteries in many patients with hyperplastic proliferation of the vascular wall cellular elements, deposition of collagen, and separation of layers, designated the “onionskin” lesion. For many of these patients, fibrinoid necrosis led to obliteration of glomeruli and loss of tubular structures. Progressive kidney failure ensued and, without dialysis support, led to early mortality in untreated malignant-phase hypertension. These vascular changes could develop with pressure-related injury from a variety of hypertensive pathways, including but not limited to activation of the renin-angiotensin system and severe vasospasm associated with catecholamine release. Occasionally, endothelial injury is sufficient to induce microangiopathic hemolysis, as discussed below.

Antihypertensive therapy is the mainstay of therapy for malignant hypertension. With effective BP reduction, manifestations of vascular injury, including microangiopathic hemolysis and renal dysfunction, can improve over time. Whereas prior reports before the era of drug therapy suggested that 1-year mortality rates exceeded 90%, current survival over 5 years exceeds 50%.

Malignant hypertension is less common in Western countries, although it persists in parts of the world where medical care and antihypertensive drug therapy are less available. It most commonly develops in patients with treated hypertension who neglect to take medications or who may use vasospastic drugs, such as cocaine. Renal abnormalities typically include rising serum creatinine and occasionally hematuria and proteinuria. Biochemical findings may include evidence of hemolysis (anemia, schistocytes, and reticulocytosis) and changes associated with kidney failure. African-American males are more likely to develop rapidly progressive hypertension and kidney failure than are whites in the United States. Genetic polymorphisms for *APOL1* that are common in the African-American population predispose to subtle focal sclerosing glomerular disease, with severe hypertension developing at younger ages secondary to renal disease in this instance.

“Hypertensive Nephrosclerosis”

Based on experience with malignant hypertension and epidemiologic evidence linking BP with long-term risks of kidney failure, it has long been assumed that lesser degrees of hypertension induce less severe, but prevalent, changes in kidney vessels and loss of kidney function. As a result, a large portion of patients reaching ESRD without a specific etiologic diagnosis are assigned the designation “hypertensive nephrosclerosis.” Pathologic examination commonly identifies afferent arteriolar thickening with deposition of homogeneous eosinophilic material (hyaline arteriolosclerosis) associated with narrowing of vascular lumina. Clinical manifestations include retinal vessel changes associated with hypertension (arteriolar narrowing, arteriovenous crossing changes), left ventricular hypertrophy, and elevated BP. The role of these vascular changes in kidney function is unclear. Postmortem and biopsy samples from normotensive kidney donors demonstrate similar vessel changes associated with aging, dyslipidemia, and glucose intolerance. Although BP reduction does slow progression of proteinuric kidney diseases and is warranted to reduce the excessive cardiovascular risks associated with CKD, antihypertensive therapy does not alter the course of kidney dysfunction identified specifically as hypertensive nephrosclerosis.

FURTHER READING

de Mast Q, Beutler JJ: The prevalence of atherosclerotic renal artery stenosis in risk groups: A systemic literature review. J Hypertens 27:1333, 2009. [PubMed: 19365285]
Freedman BI, Cohen AH: Hypertension-attributed nephropathy: What's in a name? Nat Rev Nephrol 12:27, 2016. [PubMed: 26553514]
Herrmann SM et al: Management of atherosclerotic renovascular disease after Cardiovascular Outcomes in Renal Atherosclerotic Lesions (CORAL). Nephrol Dial Transplant 30:366, 2015. [PubMed: 24723543]
Modi KS, Rao VK: Atheroembolic renal disease. J Am Soc Nephrol 12:1781 2001. [PubMed: 11461954]
Parikh SA et al: SCAI expert consensus statement for renal artery stenting appropriate use. Catheter Cardiovasc Interv 84:1163, 2014. [PubMed: 25138644]
Persu A et al: European consensus on the diagnosis and management of fibromuscular dysplasia. J Hypertens 32:1367, 2014. [PubMed: 24842696]
Textor SC et al: Percutaneous revascularization for ischemic nephropathy: The past, present and future. Kidney Int 83:28, 2013. [PubMed: 23151953]

Chapter 273: Deep Venous Thrombosis and Pulmonary Thromboembolism

Samuel Z. Goldhaber

EPIDEMIOLOGY

Venous thromboembolism (VTE) encompasses deep venous thrombosis (DVT) and pulmonary embolism (PE) and causes cardiovascular death and disability as well as psychological illness and emotional distress. In the United States, the Surgeon General estimates that there are 100,000 to 180,000 deaths annually from PE and has declared that PE is the most common preventable cause of death among hospitalized patients. In a Canadian study, almost half of PE patients at 1 year had exercise limitation, decreased walking distance, or dyspnea, which lowered their quality of life. Survivors may suffer the complications of chronic thromboembolic pulmonary hypertension or postthrombotic syndrome. Chronic thromboembolic pulmonary hypertension causes breathlessness, especially with exertion. Postthrombotic syndrome (also known as *chronic venous insufficiency*) damages the venous valves of the leg and worsens the quality of life by causing ankle or calf swelling and leg aching, especially after prolonged standing. In its most severe form, postthrombotic syndrome causes skin ulceration **(Fig. 273-1)**.

FIGURE 273-1

Skin ulceration in the lateral malleolus from postthrombotic syndrome of the leg.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOPHYSIOLOGY

Inflammation and Platelet Activation

Virchow's triad of venous stasis, hypercoagulability, and endothelial injury leads to recruitment of activated platelets, which release microparticles. These microparticles contain proinflammatory mediators that bind neutrophils, stimulating them to release their nuclear material and form web-like extracellular networks called neutrophil extracellular traps. These prothrombotic networks contain histones that stimulate platelet aggregation and promote platelet-dependent thrombin generation. Venous thrombi form and flourish in an environment of stasis, low oxygen tension, and upregulation of proinflammatory genes.

Prothrombotic States

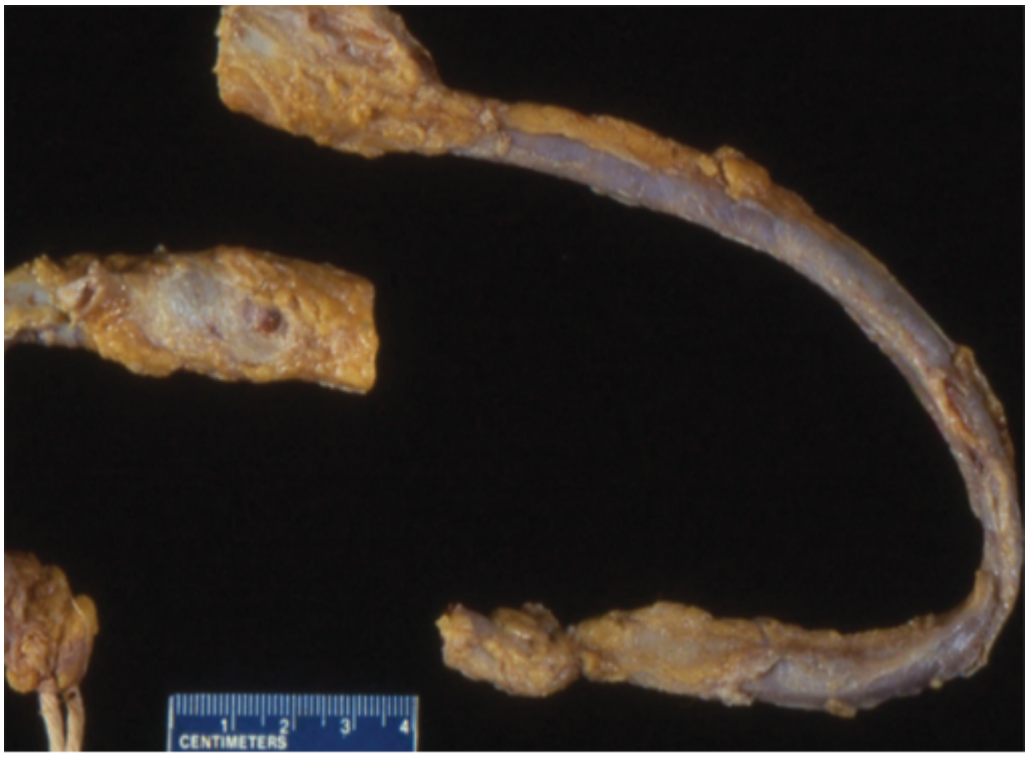
The two most common autosomal dominant genetic mutations are factor V Leiden, which causes resistance to the endogenous anticoagulant, activated protein C (which inactivates clotting factors V and VIII), and the prothrombin gene mutation, which increases the plasma prothrombin concentration **(Chaps. 61 and 113)**. Antithrombin, protein C, and protein S are naturally occurring coagulation inhibitors. Deficiencies of these inhibitors are associated with VTE but are rare. Antiphospholipid antibody syndrome is the most common acquired cause of thrombophilia and is associated with venous or arterial thrombosis. Other common predisposing factors include cancer, obesity, cigarette smoking, systemic arterial hypertension, chronic obstructive pulmonary disease, chronic kidney disease, blood transfusion, long-haul air travel, air pollution, estrogen-containing contraceptives, pregnancy, postmenopausal hormone replacement, surgery, and trauma. Inflammation predisposes to thrombosis, and conditions such as psoriasis and inflammatory bowel disease have become recognized risk factors of VTE. Sedentary lifestyle is an increasingly prevalent etiology of fatal PE. A Japanese study found that each 2 h per day increment of television watching is associated with a 40% increased likelihood of fatal PE.

Embolization

When deep venous thrombi **(Fig. 273-2)** detach from their site of formation, they embolize to the vena cava, right atrium, and right ventricle, and lodge in the pulmonary arterial circulation, thereby causing acute PE. Paradoxically, these thrombi occasionally embolize to the arterial circulation through a patent foramen ovale or atrial septal defect. Many patients with PE have no evidence of DVT because the clot has already embolized to the lungs.

FIGURE 273-2

Deep venous thrombosis at autopsy.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Physiology

The most common gas exchange abnormalities are arterial hypoxemia and an increased alveolar-arterial O_2 tension gradient, which represents the inefficiency of O_2 transfer across the lungs. Anatomic dead space increases because breathed gas does not enter gas exchange units of the lung. Physiologic dead space increases because ventilation to gas exchange units exceeds venous blood flow through the pulmonary capillaries.

Other pathophysiologic abnormalities include:

1. *Increased pulmonary vascular resistance* due to vascular obstruction or platelet secretion of vasoconstricting neurohumoral agents such as serotonin. Release of vasoactive mediators can produce ventilation-perfusion mismatching at sites remote from the embolus, thereby accounting for discordance between a small PE and a large alveolar-arterial O_2 gradient.

2. *Impaired gas exchange* due to increased alveolar dead space from vascular obstruction, hypoxemia from alveolar hypoventilation relative to perfusion in the nonobstructed lung, right-to-left shunting, or impaired carbon monoxide transfer due to loss of gas exchange surface.

3. *Alveolar hyperventilation* due to reflex stimulation of irritant receptors.

4. *Increased airway resistance* due to constriction of airways distal to the bronchi.

5. *Decreased pulmonary compliance* due to lung edema, lung hemorrhage, or loss of surfactant.

Pulmonary Hypertension, Right Ventricular (RV) Dysfunction, and RV Microinfarction

Pulmonary artery obstruction and neurohumoral mediators cause a rise in pulmonary artery pressure and in pulmonary vascular resistance. When RV wall tension rises, RV dilation and dysfunction ensue, with release of the cardiac biomarker, brain natriuretic peptide, due to abnormal RV stretch. The interventricular septum bulges into and compresses an intrinsically normal left ventricle (LV). Diastolic LV dysfunction reduces LV distensibility and impairs LV filling. Increased RV wall tension also compresses the right coronary artery, limits myocardial oxygen supply, and precipitates right coronary artery ischemia and RV microinfarction, with release of cardiac biomarkers such as troponin. Underfilling of the LV may lead to a fall in LV cardiac output and systemic arterial pressure, with consequent circulatory collapse and death.

CLASSIFICATION OF PULMONARY EMBOLISM AND DEEP VENOUS THROMBOSIS

Pulmonary Embolism

Massive PE accounts for 5–10% of cases, and is characterized by extensive thrombosis affecting at least half of the pulmonary vasculature. Dyspnea, syncope, hypotension, and cyanosis are hallmarks of massive PE. Patients with massive PE may present in cardiogenic shock and can die from multisystem organ failure. **Submassive PE** accounts for 20–25% of patients, and is characterized by RV dysfunction despite normal systemic arterial pressure. The combination of right heart failure and release of cardiac biomarkers indicates a high risk of clinical deterioration. **Low-risk PE** constitutes about 65–75% of cases. These patients have an excellent prognosis.

Deep Venous Thrombosis

Lower extremity DVT usually begins in the calf and propagates proximally to the popliteal vein, femoral vein, and iliac veins. Leg DVT is about 10 times more common than **upper extremity DVT,**

which is often precipitated by placement of pacemakers, internal cardiac defibrillators, or indwelling central venous catheters. The likelihood of upper extremity DVT increases as the catheter diameter and number of lumens increase. **Superficial venous thrombosis** usually presents with erythema, tenderness, and a “palpable cord.” Patients are at risk for extension of the thrombosis to the deep-venous system.

DIAGNOSIS

Clinical Evaluation

PE is known as “the Great Masquerader.” Diagnosis is difficult because symptoms and signs are nonspecific. The most common symptom is unexplained breathlessness. When occult PE occurs concomitantly with overt congestive heart failure or pneumonia, clinical improvement often fails to ensue despite standard medical treatment of the concomitant illness. This scenario presents a clinical clue to the possible coexistence of PE.

Hospitalization for syncope was associated with a 17% rate of newly diagnosed PE in an Italian multicenter study of 560 patients. Among those patients who had no alternative explanation for syncope, 25% had PE. Even when there was an alternative explanation for syncope, 13% had PE. When clinical suspicion was high according to the Wells Score or when the plasma D-dimer level was elevated, 42% had PE. PE in these patients was anatomically extensive, and 42% had thrombus in the main pulmonary artery.

With DVT, the most common symptom is a cramp or “charley horse” in the lower calf that persists and intensifies over several days. Wells Point Score criteria help estimate the clinical likelihood of DVT and PE **(Table 273-1)**. Patients with a low-to-moderate likelihood of DVT or PE should undergo initial diagnostic evaluation with D-dimer testing alone (see “Blood Tests”) without obligatory imaging tests **(Fig. 273-3)**. However, patients with a high clinical likelihood of VTE should skip D-dimer testing and undergo imaging as the next step in the diagnostic algorithm.

TABLE 273-1 Clinical Decision Rules

Low Clinical Likelihood of Deep Venous Thrombosis (DVT) If Point Score Is Zero or Less; Moderate Likelihood If Score Is 1 to 2; High Likelihood If Score Is 3 or Greater

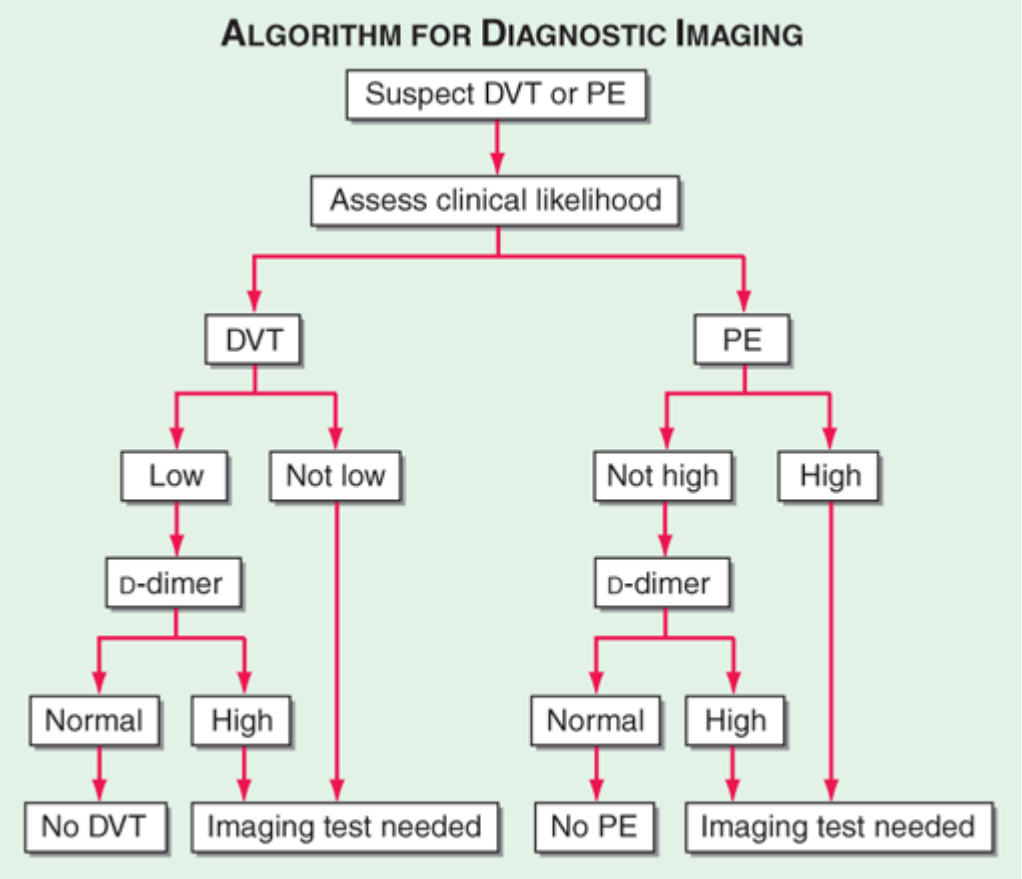
Clinical Variable	DVT Score
Active cancer	1
Paralysis, paresis, or recent cast	1
Bedridden for >3 days; major surgery <12 weeks	1
Tenderness along distribution of deep veins	1
Entire leg swelling	1
Unilateral calf swelling >3 cm	1
Pitting edema	1
Collateral superficial nonvaricose veins	1
Alternative diagnosis at least as likely as DVT	–2

High Clinical Likelihood of Pulmonary Embolism (PE) if Point Score Exceeds 4

Clinical Variable	PE Score
Signs and symptoms of DVT	3.0
Alternative diagnosis less likely than PE	3.0
Heart rate >100/min	1.5
Immobilization >3 days; surgery within 4 weeks	1.5
Prior PE or DVT	1.5
Hemoptysis	1.0
Cancer	1.0

FIGURE 273-3

How to decide whether diagnostic imaging is needed. For assessment of clinical likelihood, see Table 273-1.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical Pearls

Not all leg pain is due to DVT, and not all dyspnea is due to PE **(Table 273-2)**. Sudden, severe calf discomfort suggests a ruptured Baker's cyst. Fever and chills usually herald cellulitis rather than DVT. Physical findings, if present, may consist only of mild palpation discomfort in the lower calf. However, massive DVT often presents with marked thigh swelling, tenderness, and erythema. Recurrent left thigh edema especially in young women raises the possibility of May-Thurner Syndrome, with right proximal iliac artery compression of the left proximal iliac vein. However, if a leg is diffusely edematous, DVT is unlikely. More probable is an acute exacerbation of venous insufficiency due to postthrombotic syndrome. Upper extremity venous thrombosis may present with asymmetry in the supraclavicular fossa or in the circumference of the upper arms.

TABLE 273-2 Differential Diagnosis

Deep Venous Thrombosis (DVT)
Ruptured Baker's cyst
Muscle strain/injury
Cellulitis
Acute postthrombotic syndrome/venous insufficiency
Pulmonary Embolism (PE)
Pneumonia, asthma, chronic obstructive pulmonary disease
Congestive heart failure
Pericarditis
Pleurisy: "viral syndrome, " costochondritis, musculoskeletal discomfort
Rib fracture, pneumothorax
Acute coronary syndrome
Anxiety

Pulmonary infarction usually indicates a small PE. This condition is exquisitely painful because the thrombus lodges peripherally, near the innervation of pleural nerves. *Nonthrombotic PE* etiologies include fat embolism after pelvic or long bone fracture, tumor embolism, bone marrow, and air embolism. Cement embolism and bony fragment embolism can occur after total hip or knee replacement. Intravenous drug users may inject themselves with a wide array of substances that can embolize, such as hair, talc, and cotton. *Amniotic fluid embolism* occurs when fetal membranes leak or tear at the placental margin.

Nonimaging Diagnostic Modalities

BLOOD TESTS

The quantitative *plasma* D*-dimer enzyme-linked immunosorbent assay (ELISA)* rises in the presence of DVT or PE because of the breakdown of fibrin by plasmin. Elevation of D-dimer indicates endogenous although often clinically ineffective thrombolysis. The sensitivity of the D-dimer is >80% for DVT (including isolated calf DVT) and >95% for PE. The D-dimer is less sensitive for DVT than for PE because the DVT thrombus size is smaller. A normal D-dimer is a useful "rule out" test. However, the D-dimer assay is not specific. Levels increase in patients with myocardial infarction, pneumonia, sepsis, cancer, and the postoperative state and those in the second or third trimester of pregnancy. Therefore, D-dimer rarely has a useful role among hospitalized patients, because levels are frequently elevated due to systemic illness.

ELEVATED CARDIAC BIOMARKERS

Serum troponin and plasma heart-type fatty acid–binding protein levels increase because of RV microinfarction. Myocardial stretch causes release of brain natriuretic peptide or NT-pro-brain natriuretic peptide.

ELECTROCARDIOGRAM

The most frequently cited abnormality, in addition to sinus tachycardia, is the S1Q3T3 sign: an S wave in lead I, a Q wave in lead III, and an inverted T wave in lead III **(Chap. 235)**. This finding is relatively specific but insensitive. RV strain and ischemia cause the most common abnormality, T-wave inversion in leads V_1 to V_4.

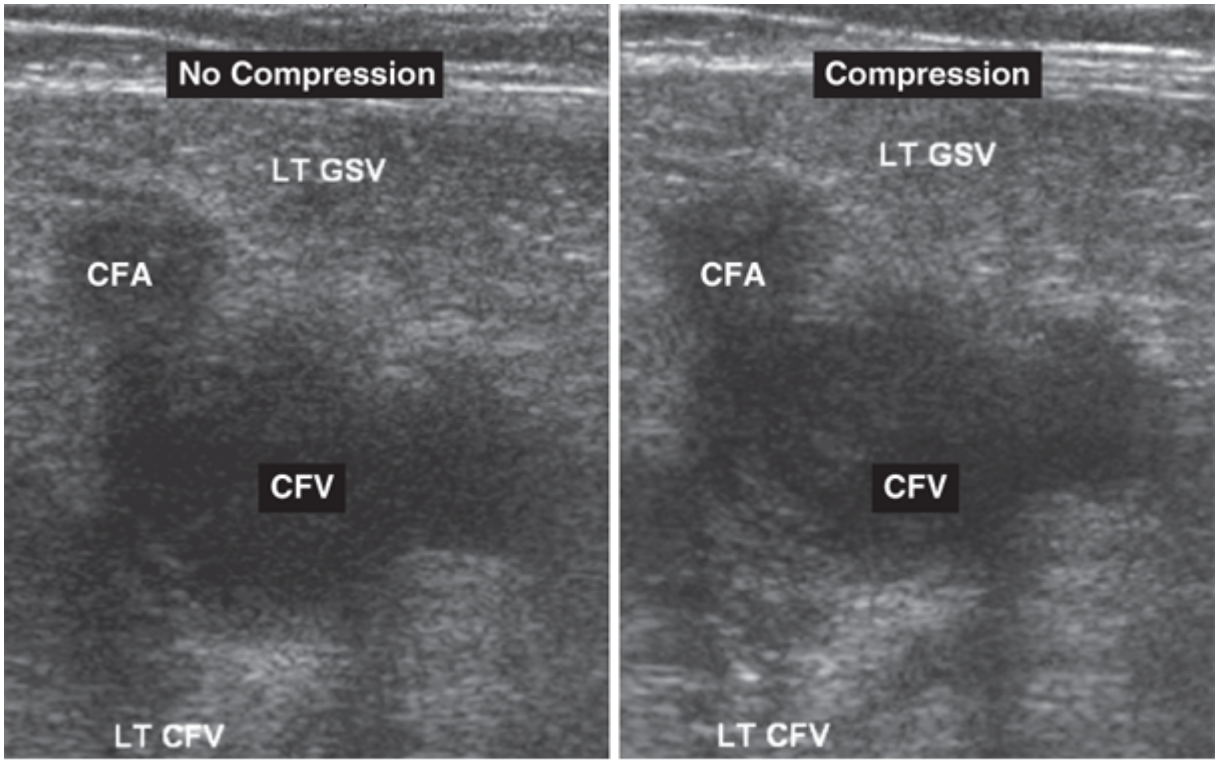
Noninvasive Imaging Modalities

VENOUS ULTRASONOGRAPHY

Ultrasonography of the deep-venous system relies on loss of vein compressibility as the primary diagnostic criterion for DVT. When a normal vein is imaged in cross-section, it readily collapses with gentle manual pressure on the ultrasound transducer. This creates the illusion of a “wink.” With acute DVT, the vein loses its compressibility because of passive distention by acute thrombus. The diagnosis of acute DVT is even more secure when thrombus is directly visualized. It appears homogeneous and has low echogenicity **(Fig. 273-4)**. The vein itself often appears mildly dilated, and collateral channels may be absent.

FIGURE 273-4

Venous ultrasound, with and without compression of the leg veins. CFA, common femoral artery; CFV, common femoral vein; GSV, great saphenous vein; LT, left.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Venous flow dynamics can be examined with Doppler imaging. Normally, manual calf compression causes augmentation of the Doppler flow pattern. Loss of normal respiratory variation is caused by an obstructing DVT or by any obstructive process within the pelvis. For patients with a technically poor or nondiagnostic venous ultrasound, one should consider alternative imaging modalities for DVT, such as computed tomography (CT) and magnetic resonance imaging.

CHEST ROENTGENOGRAPHY

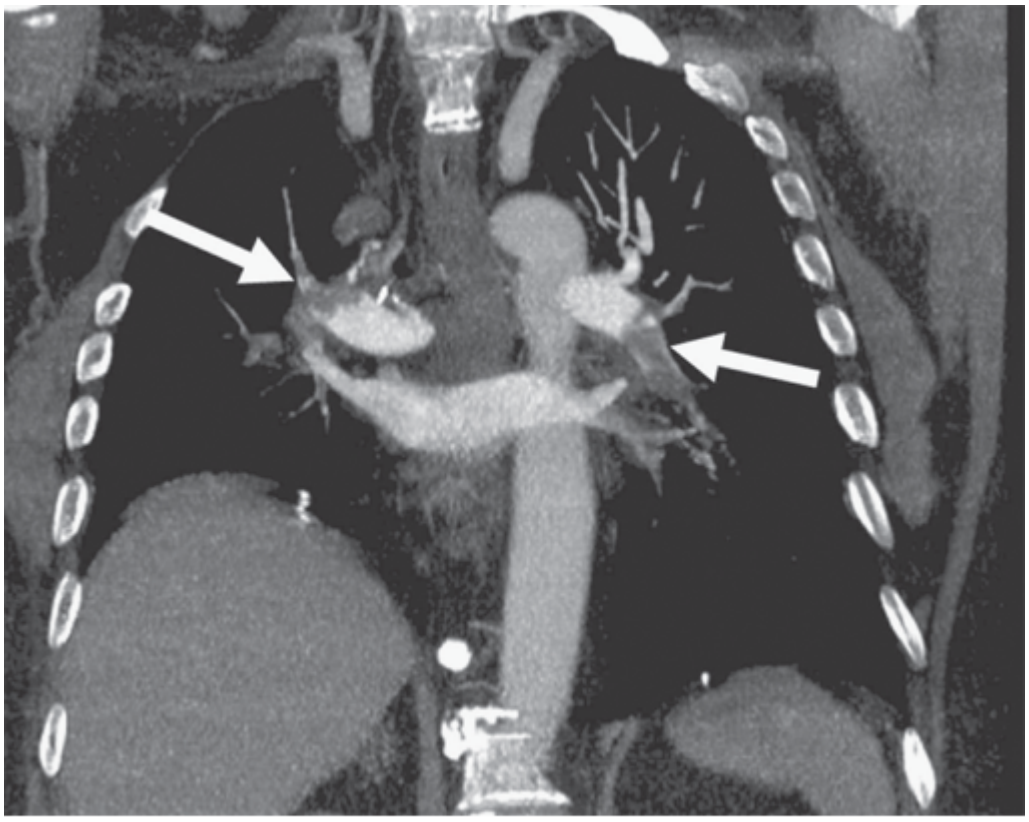
A normal or nearly normal chest x-ray often occurs in PE. Well-established abnormalities include focal oligemia (Westermark’s sign), a peripheral wedged-shaped density usually located at the pleural base (Hampton’s hump), and an enlarged right descending pulmonary artery (Palla’s sign).

CHEST CT

CT of the chest with intravenous contrast is the principal imaging test for the diagnosis of PE **(Fig. 273-5)**. “Thin-cut chest CT images” can provide exquisite detail, with ≤1 mm of resolution during a short breath hold. Sixth-order branches can be visualized with resolution superior to that of conventional invasive contrast pulmonary angiography. The CT scan also provides an excellent four-chamber view of the heart. RV enlargement on chest CT indicates an increased likelihood of death within the next 30 days compared with PE patients who have normal RV size. When imaging is extended distally below the chest to the knee, pelvic and proximal leg DVT also can be diagnosed by CT scanning. In patients without PE, the lung parenchymal images may establish alternative diagnoses not apparent on chest x-ray that explain the presenting symptoms and signs, such as pneumonia, emphysema, pulmonary fibrosis, pulmonary mass, and aortic pathology. Sometimes, asymptomatic early-stage lung cancer is diagnosed incidentally. Major efforts are underway to reduce radiation and contrast material requirements for chest CT. “Triple rule-out CT” utilizes ECG-synchronized acquisition, adjusts contrast material timing, and opacifies both the thoracic aorta and pulmonary artery circulation to exclude the three major causes of acute chest pain: PE, acute aortic syndrome, and acute coronary syndrome.

FIGURE 273-5

Large bilateral proximal PE on a coronal chest CT image in a 54-year-old man with lung cancer and brain metastases. He had developed sudden onset of chest heaviness and shortness of breath while at home. There are filling defects in the main and segmental pulmonary arteries bilaterally (*white arrows*). Only the left upper lobe segmental artery is free of thrombus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

LUNG SCANNING

Lung scanning has become a second-line diagnostic test for PE, used mostly for patients who cannot tolerate intravenous contrast. Small particulate aggregates of albumin labeled with a gamma-emitting radionuclide are injected intravenously and are trapped in the pulmonary capillary bed. The perfusion scan defect indicates absent or decreased blood flow, possibly due to PE. Ventilation scans, obtained with a radiolabeled inhaled gas such as xenon or krypton, improve the specificity of the perfusion scan. Abnormal ventilation scans indicate abnormal nonventilated lung, thereby providing possible explanations for perfusion defects other than acute PE, such as asthma and chronic obstructive pulmonary disease. A high-probability scan for PE is defined as two or more segmental perfusion defects in the presence of normal ventilation.

The diagnosis of PE is very unlikely in patients with normal and nearly normal scans and, in contrast, is about 90% certain in patients with high-probability scans. Unfortunately, most patients have nondiagnostic scans, and fewer than one-half of patients with angiographically confirmed PE have a high probability scan. As many as 40% of patients with high clinical suspicion for PE but “low-probability” scans do, in fact, have PE at angiography.

MAGNETIC RESONANCE (MR) (CONTRAST-ENHANCED) IMAGING

When ultrasound is equivocal, MR venography with gadolinium contrast is an excellent imaging modality to diagnose DVT. MR pulmonary angiography may detect large proximal PE, but is not reliable for smaller segmental and subsegmental PE.

ECHOCARDIOGRAPHY

Echocardiography is *not* a reliable diagnostic imaging tool for acute PE because most patients with PE have normal echocardiograms. However, echocardiography is a very useful diagnostic tool for detecting conditions that may mimic PE, such as acute myocardial infarction, pericardial tamponade, and aortic dissection. Transthoracic echocardiography rarely images thrombus directly. The best-known indirect sign of PE on transthoracic echocardiography is McConnell’s sign: hypokinesis of the RV free wall with normal or hyperkinetic motion of the RV apex. One should consider transesophageal echocardiography when CT scanning facilities are not available or when a patient has renal failure or severe contrast allergy that precludes administration of contrast despite premedication with high-dose steroids. This imaging modality can identify saddle, right main, or left main PE.

Invasive Diagnostic Modalities

PULMONARY ANGIOGRAPHY

Chest CT with contrast (see above) has virtually replaced invasive pulmonary angiography as a diagnostic test. Invasive catheter-based diagnostic testing is reserved for patients with technically unsatisfactory chest CTs and for those in whom an interventional procedure such as catheter-directed thrombolysis is planned. A definitive diagnosis of PE requires visualization of an intraluminal filling defect in more than one projection. Secondary signs of PE include abrupt occlusion (“cut-off”) of vessels, segmental oligemia or avascularity, and a prolonged arterial phase with slow filling, and tortuous, tapering peripheral vessels.

CONTRAST PHLEBOGRAPHY

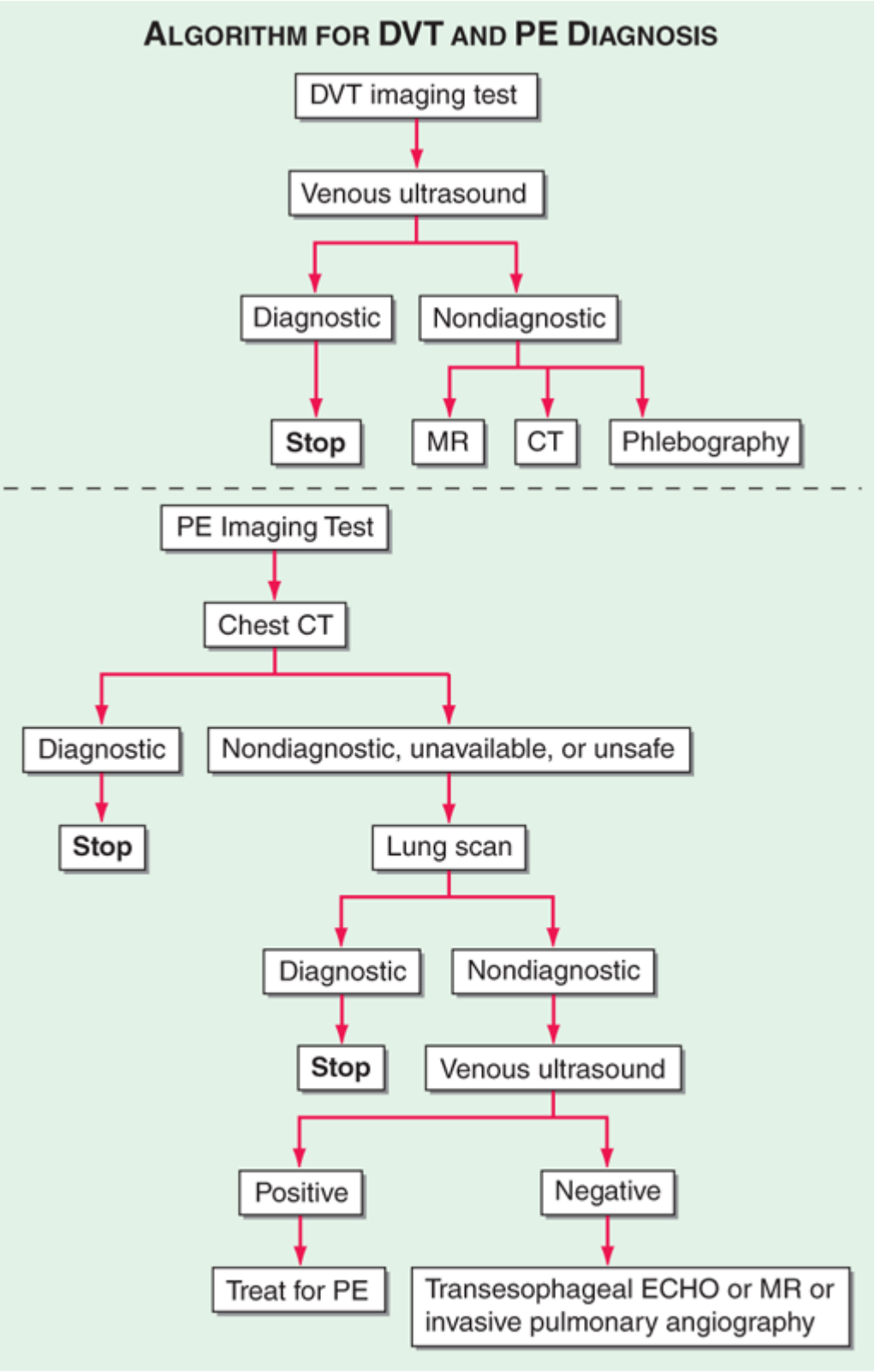
Venous ultrasonography has virtually replaced contrast phlebography as the principal diagnostic test for suspected DVT.

Integrated Diagnostic Approach

An integrated diagnostic approach (Fig. 273-3) streamlines the workup of suspected DVT and PE **(Fig. 273-6)**.

FIGURE 273-6

Imaging tests to diagnose DVT and PE. ECHO, echocardiography.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT Deep Venous Thrombosis PRIMARY THERAPY

Primary therapy consists of clot dissolution with pharmacomechanical therapy that usually includes low-dose catheter-directed thrombolysis. This approach is reserved for patients with extensive femoral, iliofemoral, or upper extremity DVT. The open vein hypothesis postulates that patients who receive primary therapy will sustain less long-term damage to venous valves, with consequent lower rates of postthrombotic syndrome. A National Heart, Lung, and Blood Institute–sponsored randomized controlled trial called ATTRACT (NCT00790335) is testing this hypothesis by randomizing femoral and iliofemoral DVT patients to conventional anticoagulation versus pharmacomechanical catheter-directed thrombolysis, and assessing the frequency of postthrombotic syndrome 2 years after randomization.

SECONDARY PREVENTION

Anticoagulation or placement of an inferior vena caval (IVC) filter constitutes *secondary prevention* of VTE. In 2016, the FDA approved a new retrievable IVC filter that is inserted at the bedside with ultrasound visualization of the femoral or internal jugular vein (Angel® Filter) but without the need for any fluoroscopic or other radiological imaging.

For patients with swelling of the legs when acute DVT is diagnosed, below-knee graduated compression stockings may be prescribed, usually 30–40 mmHg, to lessen patient discomfort. They should be replaced every 3 months because they lose their elasticity. However, prescription of vascular compression stockings in asymptomatic newly diagnosed acute DVT patients does not prevent the development of postthrombotic syndrome.

TREATMENT

TREATMENT Pulmonary Embolism RISK STRATIFICATION

Hemodynamic instability, RV dysfunction on echocardiography, RV enlargement on chest CT, or elevation of the troponin level due to RV microinfarction portend a high risk of an adverse clinical outcome despite anticoagulation. When RV function remains normal in a hemodynamically stable patient, a good clinical outcome is highly likely with anticoagulation alone **(Fig. 273-7)**.

ANTICOAGULATION

Effective anticoagulation is the foundation for successful treatment of DVT and PE. There are three major strategies: (1) the classical but waning strategy of parenteral anticoagulation with unfractionated heparin (UFH), low-molecular-weight heparin (LMWH), or fondaparinux “bridged” to warfarin, (2) parenteral therapy switched after 5 days to a novel oral anticoagulant such as dabigatran (a direct thrombin inhibitor) or edoxaban (an anti-Xa agent), or (3) oral anticoagulation monotherapy with rivaroxaban or apixaban (both are anti-Xa agents) with a 3-week or 1-week loading dose, respectively, followed by a maintenance dose without parenteral anticoagulation. For patients with VTE in the setting of suspected or proven heparin-induced thrombocytopenia, one can choose between two parenteral direct thrombin inhibitors: argatroban and bivalirudin **(Table 273-3)**.

Unfractionated Heparin

UFH anticoagulates by binding to and accelerating the activity of antithrombin, thus preventing additional thrombus formation. UFH is dosed to achieve a target activated partial thromboplastin time (aPTT) of 60–80 s. The most popular nomogram uses an initial bolus of 80 U/kg, followed by an initial infusion rate of 18 U/kg per h in patients with normal liver function. The major advantage of UFH is its short half-life, which is especially useful in patients in whom hour-to-hour control of the intensity of anticoagulation is desired. Heparin also has pleiotropic effects that may decrease systemic and local inflammation.

Low-Molecular-Weight Heparins

These fragments of UFH exhibit less binding to plasma proteins and endothelial cells and consequently have greater bioavailability, a more predictable dose response, and a longer half-life than does UFH. No monitoring or dose adjustment is needed unless the patient is markedly obese or has chronic kidney disease.

Fondaparinux

Fondaparinux, an anti-Xa pentasaccharide, is administered as a weight-based once-daily subcutaneous injection in a prefilled syringe. No laboratory monitoring is required. Fondaparinux is synthesized in a laboratory and, unlike LMWH or UFH, is not derived from animal products. It does not cause heparin-induced thrombocytopenia. The dose must be adjusted downward for patients with renal dysfunction.

Warfarin

This vitamin K antagonist prevents carboxylation activation of coagulation factors II, VII, IX, and X. The full effect of warfarin requires at least 5 days, even if the prothrombin time, used for monitoring, becomes elevated more rapidly. If warfarin is initiated as monotherapy during an acute thrombotic illness, a paradoxical exacerbation of hypercoagulability increases the likelihood of thrombosis. Overlapping UFH, LMWH, fondaparinux, or parenteral direct thrombin inhibitors with warfarin for at least 5 days will nullify the early procoagulant effect of warfarin.

Warfarin dosing

In an average-size adult, warfarin is often initiated in a dose of 5 mg. The prothrombin time is standardized by calculating the international normalized ratio (INR), which assesses the anticoagulant effect of warfarin **(Chap. 61)**. The target INR is usually 2.5, with a range of 2.0–3.0.

The warfarin dose is usually titrated empirically to achieve the target INR. Proper dosing is difficult because hundreds of drug-drug and drug-food interactions affect warfarin metabolism. Increasing age and systemic illness reduce the required warfarin dose. Pharmacogenomics may provide more precise initial dosing of warfarin. *CYP2C9* variant alleles impair the hydroxylation of S-warfarin, thereby lowering the dose requirement. Variants in the gene encoding the vitamin K epoxide reductase complex 1 (*VKORC1*) can predict whether patients require low, moderate, or high warfarin doses. However, genetic testing is not used clinically to dose patients with warfarin.

Centralized anticoagulation clinics have improved the efficacy and safety of warfarin dosing. Patients can self-monitor their INR with a home point-of-care fingerstick machine and can occasionally be taught to self-dose their warfarin.

Warfarin can cause major hemorrhage, including intracranial hemorrhage, even when the INR remains within the desired therapeutic range. Warfarin can cause “off target” side effects such as alopecia or arterial vascular calcification. Some patients complain that warfarin makes them feel cold or fatigued.

Novel Oral Anticoagulants

Novel oral anticoagulants (NOACs) are administered in a fixed dose, establish effective anticoagulation within hours of ingestion, require no laboratory coagulation monitoring, and have few of the drug-drug or drug-food interactions. Betrixaban, a direct factor Xa inhibitor, was approved by the FDA in 2017 for VTE prophylaxis in acutely ill medical patients during hospitalization and continuing for a total duration of 5 to 6 weeks. Rivaroxaban and apixaban, direct factor Xa inhibitors, are approved as monotherapy for acute and extended treatment of DVT and PE, without a parenteral

“bridging” anticoagulant. Dabigatran, a direct thrombin inhibitor, and edoxaban, a factor Xa inhibitor, are approved for treatment of VTE after an initial 5-day course of parenteral anticoagulation.

Complications of Anticoagulants

The most serious adverse effect of anticoagulation is hemorrhage. For life-threatening or intracranial hemorrhage due to heparin or LMWH, protamine sulfate can be administered. There is no specific reversal agent for bleeding caused by fondaparinux or factor Xa inhibitors. However, the dabigatran antibody, idarucizumab, is an effective and rapidly acting antidote for dabigatran that is now licensed for use. Andexanet is a universal anti-Xa antidote for betrixaban, rivaroxaban, apixaban, and edoxaban that is undergoing review by the FDA.

Major bleeding from warfarin is best managed with prothrombin complex concentrate. With less serious bleeding, fresh-frozen plasma or intravenous vitamin K can be used. Oral vitamin K is effective for managing minor bleeding or an excessively high INR in the absence of bleeding.

Duration of Anticoagulation

For DVT isolated to an upper extremity or calf that has been provoked by surgery, trauma, estrogen, or an indwelling central venous catheter or pacemaker, 3 months of anticoagulation usually suffice. For an initial episode of provoked proximal leg DVT or PE, 3–6 months of anticoagulation used to be the classic teaching. However, the EINSTEIN CHOICE study found that patients with provoked VTE derived as great a risk reduction in recurrent VTE with extended duration anticoagulation as patients with unprovoked VTE. For patients with cancer and VTE, prescribe LMWH as monotherapy without warfarin and continue anticoagulation indefinitely unless the patient is rendered cancer-free.

Among patients with idiopathic, unprovoked VTE, the recurrence rate is high after cessation of anticoagulation. VTE that occurs during long-haul air travel is considered unprovoked. Unprovoked VTE may be caused by an exacerbation of an underlying inflammatory state and can be conceptualized as a chronic illness, with latent periods between flares of recurrent episodes. American College of Chest Physicians (ACCP) guidelines recommend considering anticoagulation for an indefinite duration with a target INR between 2 and 3 for patients with idiopathic VTE and a low bleeding risk. An alternative approach after the first 6 months of anticoagulation is to reduce the intensity of anticoagulation and to lower the target INR range to between 1.5 and 2. Another approach for patients at lower risk of recurrence, especially if there is an important reason to avoid long-term anticoagulation, is to consider low-dose aspirin after completing the initial period of standard anticoagulation.

Counterintuitively, the presence of genetic mutations such as heterozygous factor V Leiden and prothrombin gene mutation does not appear to increase the risk of recurrent VTE. However, patients with antiphospholipid antibody syndrome may warrant indefinite-duration anticoagulation, even if the initial VTE was provoked by trauma or surgery.

INFERIOR VENA CAVA FILTERS

The two principal indications for insertion of an IVC filter are (1) active bleeding that precludes anticoagulation and (2) recurrent venous thrombosis despite intensive anticoagulation. Prevention of recurrent PE in patients with right heart failure who are not candidates for fibrinolysis and prophylaxis of extremely high-risk patients are “softer” indications for filter placement. The filter itself may fail by permitting the passage of small- to medium-size clots. Large thrombi may embolize to the pulmonary arteries via collateral veins that develop.

Paradoxically, by providing a nidus for clot formation, filters increase the DVT rate, even though they usually prevent PE. Therefore, a common complication is recurrent DVT or caval thrombosis with marked leg swelling. Retrievable filters can now be placed for patients with an anticipated temporary

bleeding disorder or for patients at temporary high risk of PE, such as individuals undergoing bariatric surgery who have a prior history of perioperative PE. The filters can be retrieved for months after insertion, unless thrombus forms and is trapped within the filter. The retrievable filter becomes permanent if it remains in place or if, for technical reasons such as rapid endothelialization, it cannot be removed.

MANAGEMENT OF MASSIVE PE

For patients with massive PE and hypotension, replete volume with 500 mL of normal saline. Additional fluid should be infused with extreme caution because excessive fluid administration exacerbates RV wall stress, causes more profound RV ischemia, and worsens LV compliance and filling by causing further interventricular septal shift toward the LV. Dopamine and dobutamine are first-line inotropic agents for treatment of PE-related shock. Maintain a low threshold for initiating these pressors. Often, a “trial-and-error” approach works best; other agents that may be effective include norepinephrine, vasopressin, or phenylephrine.

FIBRINOLYSIS

Successful fibrinolytic therapy rapidly reverses right heart failure and may result in a lower rate of death and recurrent PE by (1) dissolving much of the anatomically obstructing pulmonary arterial thrombus, (2) preventing the continued release of serotonin and other neurohumoral factors that exacerbate pulmonary hypertension, and (3) lysing much of the source of the thrombus in the pelvic or deep leg veins, thereby decreasing the likelihood of recurrent PE.

The preferred systemically administered fibrinolytic regimen is 100 mg of recombinant tissue plasminogen activator (tPA) prescribed as a continuous peripheral intravenous infusion over 2 h. The sooner thrombolysis is administered, the more effective it is. However, this approach can be used for at least 14 days after the PE has occurred. A popular off-label dosing regimen is 50 mg of TPA administered over 2 h. This lower dose is widely perceived to be associated with fewer bleeding complications.

Contraindications to fibrinolysis include intracranial disease, recent surgery, and trauma. The overall major bleeding rate is about 10%, including a 2–3% risk of intracranial hemorrhage. Careful screening of patients for contraindications to fibrinolytic therapy **(Chap. 269)** is the best way to minimize bleeding risk.

The only Food and Drug Administration–approved indication for PE fibrinolysis is massive PE. For patients with submassive PE, who have preserved systolic blood pressure but moderate or severe RV dysfunction, use of fibrinolysis remains controversial. Results of a 1006-patient European multicentered randomized trial of submassive PE, using the thrombolytic agent tenecteplase versus heparin alone, showed that death or hemodynamic collapse within 7 days of randomization was reduced by 56% in the tenecteplase group. However, hemorrhagic stroke occurred in 2% of tenecteplase patients versus 0.2% in patients who only received heparin.

PHARMACOMECHANICAL CATHETER-DIRECTED THERAPY

Many patients have relative contraindications to full-dose thrombolysis. Pharmacomechanical catheter-directed therapy usually combines physical fragmentation or pulverization of thrombus with catheter-directed low-dose thrombolysis. Mechanical techniques include catheter maceration and intentional embolization of clot more distally, suction thrombectomy, rheolytic hydrolysis, and low-energy ultrasound-facilitated thrombolysis. The dose of alteplase can be markedly reduced, usually to a range of 20–25 mg, instead of the peripheral intravenous systemic dose of 100 mg. In 2014, the FDA approved ultrasound-facilitated catheter-directed thrombolysis for acute massive and submassive PE. Using a total tPA dose of 24 mg, this approach decreased RV dilation, reduced pulmonary

hypertension, decreased anatomic thrombus burden, and minimized intracranial hemorrhage. Lower doses and durations of TPA are currently being studied.

PULMONARY EMBOLECTOMY

The risk of major hemorrhage with systemically administered fibrinolysis has prompted a renaissance of interest in surgical embolectomy, an operation that had almost become extinct. More rapid referral before the onset of irreversible multisystem organ failure and improved surgical technique have resulted in a high survival rate.

PULMONARY THROMBOENDARTERECTOMY

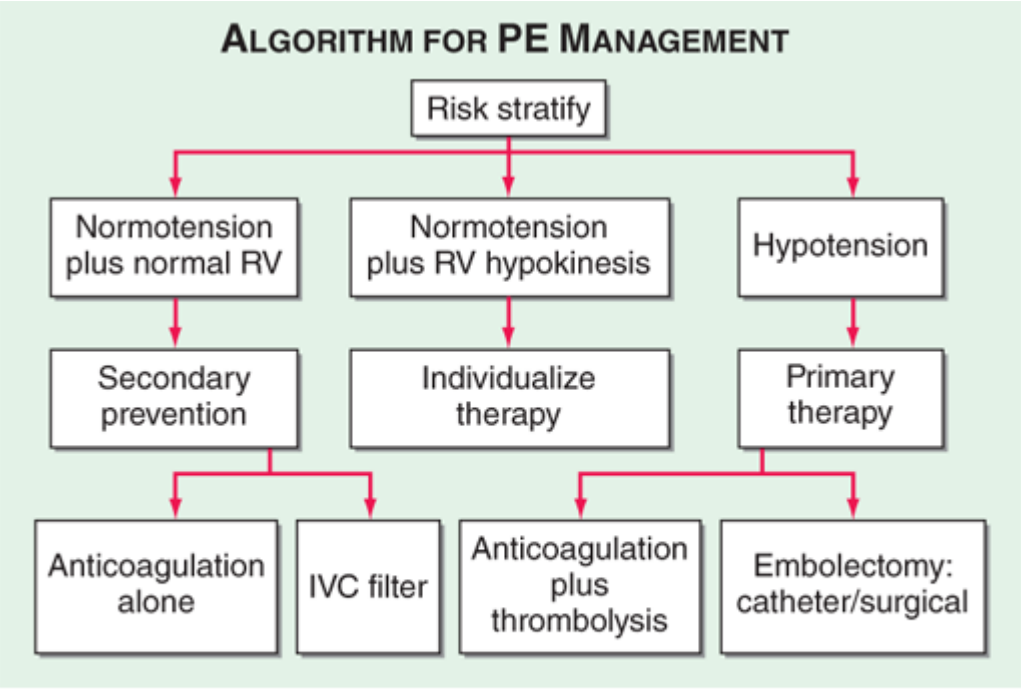
Chronic thromboembolic pulmonary hypertension develops in 2–4% of acute PE patients. Therefore, PE patients who have initial pulmonary hypertension (usually diagnosed with Doppler echocardiography) should be followed up at about 6 weeks with a repeat echocardiogram to determine whether pulmonary arterial pressure has normalized. Patients impaired by dyspnea due to chronic thromboembolic pulmonary hypertension should be considered for pulmonary thromboendarterectomy, which, when successful, can markedly reduce, and sometimes even cure, pulmonary hypertension **(Chap. 277)**. The operation requires median sternotomy, cardiopulmonary bypass, deep hypothermia, and periods of hypothermic circulatory arrest. The mortality rate at experienced centers is ~5%. Inoperable patients should be managed with pulmonary vasodilator therapy and balloon angioplasty of pulmonary arterial webs.

EMOTIONAL SUPPORT

Patients with VTE may feel overwhelmed when they learn that they are suffering from PE or DVT. Some have never previously encountered serious cardiovascular illness. They fear they will not be able to adapt to the new limitations imposed by anticoagulation. They worry about the health of their families and the genetic implications of their illness. Those who are advised to discontinue anticoagulation may feel especially vulnerable about the potential for suffering recurrent VTE. At Brigham and Women's Hospital, a physician-nurse–facilitated PE support group was initiated to address these concerns and has met monthly for >25 years. The nonprofit organization, North American Thrombosis Forum (*www.NATFonline.org*), has initiated other online support groups which garner worldwide participation.

FIGURE 273-7

Acute management of pulmonary thromboembolism. RV, right ventricular; IVC, inferior vena cava.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 273-3 Anticoagulation of Venous Thromboembolism (VTE)

Non-Warfarin Anticoagulation

Unfractionated heparin, bolus and continuous infusion, to achieve activated partial thromboplastin time (aPTT) 2–3 times the upper limit of the laboratory normal, *or*

Enoxaparin 1 mg/kg twice daily with normal renal function, *or*

Dalteparin 200 U/kg once daily or 100 U/kg twice daily, with normal renal function, *or*

Tinzaparin 175 U/kg once daily with normal renal function, *or*

Fondaparinux weight-based once daily; adjust for impaired renal function

Direct thrombin inhibitors: argatroban or bivalirudin (with suspected or proven heparin-induced thrombocytopenia)

Rivaroxaban 15 mg twice daily for 3 weeks, followed by 20 mg once daily with the dinner meal thereafter

Apixaban 10 mg twice daily for 1 week, followed by 5 mg twice daily thereafter

Dabigatran 5 days of unfractionated heparin, low-molecular-weight heparin (LMWH), or fondaparinux followed by dabigatran 150 mg twice daily

Edoxaban 5 days of unfractionated heparin, LMWH, or fondaparinux followed by edoxaban 60 mg once daily with normal renal function, weight >60 kg, in the absence of potent P-glycoprotein inhibitors

Warfarin Anticoagulation

Requires 5–10 days of administration to achieve effectiveness as monotherapy

(Unfractionated heparin, LMWH, and fondaparinux are the usual immediately effective “bridging agents” used when initiating warfarin)

Usual start dose is 5 mg

Titrate to international normalized ratio (INR), target 2.0–3.0

Continue parenteral anticoagulation for a minimum of 5 days and until two sequential INR values, at least 1 day apart, achieve the target INR range

PREVENTION OF VTE

Prevention of DVT and PE **(Table 273-4)** is of paramount importance because VTE is difficult to detect and poses a profound medical and economic burden. Low-dose UFH or LMWH is the most common form of in-hospital prophylaxis. Computerized reminder systems can increase the use of preventive measures and, at Brigham and Women's Hospital, have reduced the symptomatic VTE rate by >40%. Audits of hospitals to ensure that prophylaxis protocols are being used will also increase utilization of preventive measures. Duration of prophylaxis is an important consideration. Extended-duration prophylaxis with the novel anti-Xa agent, betrixaban, appears to be both effective and safe in medically ill patients during hospitalization, after hospital discharge, and is undergoing FDA review.

TABLE 273-4 Prevention of Venous Thromboembolism Among Hospitalized Patients

Condition	Prophylaxis Strategy
High-risk nonorthopedic surgery	Unfractionated heparin 5000 units SC bid or tid
	Enoxaparin 40 mg daily
	Dalteparin 2500 or 5000 units daily
Cancer surgery, including gynecologic cancer surgery	Enoxaparin 40 mg daily, consider 1 month of prophylaxis
	Warfarin (target INR 2.0–3.0)
Major orthopedic surgery	Enoxaparin 40 mg daily
	Enoxaparin 30 mg bid
	Dalteparin 2500 or 5000 units daily
	Fondaparinux 2.5 mg daily
	Rivaroxaban 10 mg daily, beginning 6-10 hours postoperatively
	Aspirin 81–325 mg daily
	Dabigatran 110 mg first day, then 220 mg daily
	Apixaban 2.5 mg bid, beginning 12–24 h postoperatively
	Intermittent pneumatic compression (with or without pharmacologic prophylaxis)
Medically ill patients, during hospitalization	Unfractionated heparin 5000 units bid or tid
	Enoxaparin 40 mg daily
	Dalteparin 2500 or 5000 units daily
	Fondaparinux 2.5 mg daily
Medically ill patients, during and after hospitalization	Betrixaban 80 mg daily for 35–42 days
Anticoagulation contraindicated	Intermittent pneumatic compression devices (but whether graduated compression stockings are effective in medical patients remains uncertain)

FURTHER READING

Becattini C, Giancarlo A: Treatment of venous thromboembolism with new anticoagulant agents. J Am Coll Cardiol 67:1941, 2016. [PubMed: 27102510]
Braekkan SK et al: Venous thromboembolism and subsequent work-related disability. J Thromb Haemost 14:1977, 2016.
Cohen AT et al: Extended thromboprophylaxis with betrixaban in acutely ill medical patients. N Engl J Med 375:534, 2016. [PubMed: 27232649]

Gibson CM et al: Extended-duration betrixaban reduces the risk of stroke versus standard-dose enoxaparin among hospitalized medically ill patients: An APEX trial substudy (acute medically ill venous thromboembolism prevention with extended duration betrixaban). Circulation 135:648, 2017. [PubMed: 27881569]
Kahn SR et al: Functional and exercise limitations after a first episode of pulmonary embolism: Results of the ELOPE prospective cohort study. Chest 151:1058, 2017. [PubMed: 27932051]
Kearon C et al: Antithrombotic therapy for VTE disease: CHEST guideline and expert panel report. CHEST 149:315, 2016. [PubMed: 26867832]
Konstantinides SV et al: Management of pulmonary embolism. J Am Coll Cardiol 67:975, 2015.
Piazza G et al: A prospective, single-arm, multicenter trial of ultrasound-facilitated, catheter-directed, low-dose fibrinolysis for acute massive and submassive pulmonary embolism. The SEATTLE II study. J Am Coll Cardiol Cardiovasc Interv 8:1382, 2015.
Poterucha TJ et al: More than an anticoagulant: Do heparins have direct anti-inflammatory effects? Thromb Haemost 117:437, 2017. [PubMed: 27975101]
Prandoni P et al: Prevalence of pulmonary embolism among patients hospitalized for syncope. N Engl J Med 375:1524, 2016. [PubMed: 27797317]
Raskob GE et al: The MARINER trial of rivaroxaban after hospital discharge for medical patients at high risk of VTE. Thromb Haemost 115:1240, 2016. [PubMed: 26842902]
Shirakawa T et al: Watching television and risk of mortality from pulmonary embolism among Japanese men and women. Circulation 134:355, 2016. [PubMed: 27462056]
Weitz JI et al: Rivaroxaban or aspirin for extended treatment of venous thromboembolism. N Engl J Med 376:1211, 2017. [PubMed: 28316279]

Chapter 274: Diseases of the Aorta

Mark A. Creager; Joseph Loscalzo

INTRODUCTION

The aorta is the conduit through which blood ejected from the left ventricle is delivered to the systemic arterial bed. In adults, its diameter is ~3 cm at the origin and in the ascending portion, 2.5 cm in the descending portion in the thorax, and 1.8–2 cm in the abdomen. The aortic wall consists of a thin intima composed of endothelium, subendothelial connective tissue, and an internal elastic lamina; a thick tunica media composed of smooth muscle cells and extracellular matrix; and an adventitia composed primarily of connective tissue enclosing the vasa vasorum and nervi vascularis. In addition to the conduit function of the aorta, its viscoelastic and compliant properties serve a buffering function. The aorta is distended during systole to allow a portion of the stroke volume and elastic energy to be stored, and it recoils during diastole so that blood continues to flow to the periphery. Owing to its continuous exposure to high pulsatile pressure and shear stress, the aorta is particularly prone to injury and disease resulting from mechanical trauma. The aorta is also more prone to rupture than is any other vessel, especially with the development of aneurysmal dilation, since its wall tension, as governed by Laplace's law (i.e., proportional to the product of pressure and radius), will be increased.

CONGENITAL ANOMALIES OF THE AORTA

Congenital anomalies of the aorta usually involve the aortic arch and its branches. Symptoms such as dysphagia, stridor, and cough may occur if an anomaly causes a ring around or otherwise compresses the esophagus or trachea. Anomalies associated with symptoms include double aortic arch, origin of the right subclavian artery distal to the left subclavian artery, and right-sided aortic arch with an aberrant left subclavian artery. A Kommerell's diverticulum is an anatomic remnant of a right aortic arch. Most congenital anomalies of the aorta do not cause symptoms and are detected during catheter-based procedures. The diagnosis of suspected congenital anomalies of the aorta typically is confirmed by computed tomographic (CT) or magnetic resonance (MR) angiography. Surgery is used to treat symptomatic anomalies.

Coarctation of the aorta **(Chap. 264)** typically occurs near the insertion of the ligamentum arteriosum, adjacent to the left subclavian artery. It may be associated with a bicuspid aortic valve, aortic arch hypoplasia, other congenital heart defects, and intracranial aneurysms. A pulse delay or pressure differential between the upper and lower extremities should raise suspicion of aortic coarctation. Imaging modalities, including echocardiography, CT and MR angiography are used to confirm the diagnosis. If untreated, hypertension develops in the arteries proximal to the coarctation. Treatment of hemodynamically significant aortic coarctation includes endovascular stent implantation if feasible or surgical repair.

AORTIC ANEURYSM

An *aneurysm* is defined as a pathologic dilation of a segment of a blood vessel. A *true aneurysm* involves all three layers of the vessel wall and is distinguished from a *pseudoaneurysm*, in which the

intimal and medial layers are disrupted and the dilated segment of the aorta is lined by adventitia only and, at times, by perivascular clot. Aneurysms also may be classified according to their gross appearance. A *fusiform aneurysm* affects the entire circumference of a segment of the vessel, resulting in a diffusely dilated artery. In contrast, a *saccular aneurysm* involves only a portion of the circumference, resulting in an outpouching of the vessel wall. Aortic aneurysms also are classified according to location, i.e., abdominal versus thoracic. Aneurysms of the descending thoracic aorta are usually contiguous with infradiaphragmatic aneurysms and are referred to as *thoracoabdominal aortic aneurysms*.

ETIOLOGY

Aortic aneurysms result from conditions that cause degradation or abnormal production of the structural components of the aortic wall: elastin and collagen. The causes of aortic aneurysms may be broadly categorized as degenerative disorders, genetic or developmental diseases, vasculitis, infections, and trauma **(Table 274-1)**. Inflammation, oxidative stress, proteolysis, and biomechanical wall stress contribute to the degenerative processes that characterize most aneurysms of the abdominal and descending thoracic aorta. These are mediated by B cell and T cell lymphocytes, macrophages, inflammatory cytokines, and matrix metalloproteinases that degrade elastin and collagen and alter the tensile strength and ability of the aorta to accommodate pulsatile stretch. The associated histopathology demonstrates destruction of elastin and collagen, decreased vascular smooth muscle, in-growth of new blood vessels, and inflammation. Factors associated with degenerative aortic aneurysms include aging, cigarette smoking, hypercholesterolemia, hypertension, and male sex.

TABLE 274-1 Diseases of the Aorta: Etiology and Associated Factors

- Aortic aneurysm
  - Degenerative
    - Aging
    - Cigarette smoking
    - Hypercholesterolemia
    - Hypertension
    - Atherosclerosis
  - Genetic or developmental
    - Marfan’s syndrome
    - Loeys-Dietz syndrome
    - Ehlers-Danlos syndrome type IV
    - Turner’s syndrome
    - Familial
    - Bicuspid aortic valve
  - Chronic aortic dissection
  - Aortitis (see below)
  - Infective (see below)
  - Trauma
- Acute aortic syndromes (aortic dissection, acute intramural hematoma, penetrating atherosclerotic ulcer)
  - Degenerative disorders (see above)
  - Genetic/developmental disorders (see above)
  - Hypertension
  - Aortitis (see below)
  - Pregnancy
  - Trauma
- Aortic occlusion
  - Atherosclerosis
  - Thromboembolism
- Aortitis
  - Vasculitis
    - Takayasu’s arteritis
    - Giant cell arteritis
  - Rheumatic
    - Rheumatoid aortitis
    - HLA-B27–associated spondyloarthropathies
    - Behçet’s syndrome
    - Cogan’s syndrome
    - IgG4-related systemic disease
    - Idiopathic aortitis
  - Infective
    - Syphilis
    - Tuberculosis

Mycotic (*Salmonella*, staphylococcal, streptococcal, fungal)

The most common pathologic condition associated with degenerative aortic aneurysms is *atherosclerosis*. Many patients with aortic aneurysms have coexisting risk factors for atherosclerosis, as well as atherosclerosis in other blood vessels.

Medial degeneration, previously designated *cystic medial necrosis*, is the histopathologic term used to describe the degeneration of collagen and elastic fibers in the tunica media of the aorta as well as the loss of medial cells that are replaced by multiple clefts of mucoid material, such as proteoglycans. Medial degeneration characteristically affects the proximal aorta, results in circumferential weakness and dilation, and leads to the development of fusiform aneurysms involving the ascending aorta and the sinuses of Valsalva. This pathologic condition occurs in patients with Marfan's syndrome, Loeys-Dietz syndrome, Ehlers-Danlos syndrome type IV **(Chap. 406)**, hypertension, congenital bicuspid aortic valves, Turner's syndrome, and familial thoracic aortic aneurysm syndromes; sometimes it appears as an isolated condition in patients without any other apparent disease.

Familial clusterings of aortic aneurysms occur in 20% of patients, suggesting a hereditary basis for the disease. Mutations of the gene that encodes fibrillin-1 are present in patients with Marfan's syndrome. Fibrillin-1 is an important component of extracellular microfibrils, which support the architecture of elastic fibers and other connective tissue. Deficiency of fibrillin-1 in the extracellular matrix leads to excessive signaling by transforming growth factor β (TGF-β). Loeys-Dietz syndrome is caused by mutations in the genes that encode TGF-β receptors 1 (*TGFBR1*) and 2 (*TGFBR2*). Increased signaling by TGF-β and mutations of *TGFBR1*, *TGFBR2*, *TGFBR3*, as well as *TGFB2* and *TGFB3*, may cause thoracic aortic aneurysms. Mutations of *SMAD3*, which encodes a downstream signaling protein involved with TGF binding to its receptors, have been described in a syndrome of thoracic aortic aneurysm; craniofacial, skeletal, and cutaneous anomalies; and osteoarthritis. Mutations of the genes encoding the smooth muscle–specific alpha-actin (*ACTA2*), smooth muscle cell–specific myosin heavy chain 11 (*MHC11*), and myosin light chain kinase (*MYLK*) and mutations of *TGFBR2* and *SMAD3* have been reported in some patients with nonsyndromic familial thoracic aortic aneurysms. Mutations of type III procollagen have been implicated in Ehlers-Danlos type IV syndrome.

The infectious causes of aortic aneurysms include syphilis, tuberculosis, and other bacterial infections. *Syphilis* **(Chap. 177)** is a relatively uncommon cause of aortic aneurysm. Syphilitic periaortitis and mesoaortitis damage elastic fibers, resulting in thickening and weakening of the aortic wall. Approximately 90% of syphilitic aneurysms are located in the ascending aorta or aortic arch. *Tuberculous aneurysms* **(Chap. 173)** typically affect the thoracic aorta and result from direct extension of infection from hilar lymph nodes or contiguous abscesses as well as from bacterial seeding. Loss of aortic wall elasticity results from granulomatous destruction of the medial layer. A *mycotic aneurysm* is a rare condition that develops as a result of staphylococcal, streptococcal, *Salmonella*, or other bacterial or fungal infections of the aorta, usually at an atherosclerotic plaque. These aneurysms are usually saccular. Blood cultures are often positive and reveal the nature of the infective agent.

Vasculitides associated with aortic aneurysm include Takayasu's arteritis and giant cell arteritis, which may cause aneurysms of the aortic arch and descending thoracic aorta. Spondyloarthropathies such as ankylosing spondylitis, rheumatoid arthritis, psoriatic arthritis, relapsing polychondritis, and reactive arthritis (formerly known as Reiter's syndrome) are associated with dilation of the ascending aorta. Aortic aneurysms occur in patients with Behçet's syndrome **(Chap. 357)**, Cogan's syndrome, and IgG4-related systemic disease. Aortic aneurysms also result from idiopathic aortitis. *Traumatic aneurysms* may occur after penetrating or nonpenetrating chest trauma and most commonly affect the descending thoracic aorta just beyond the site of insertion of the ligamentum arteriosum. Chronic aortic dissections are associated with weakening of the aortic wall that may lead to the development of aneurysmal dilatation.

THORACIC AORTIC ANEURYSMS

The clinical manifestations and natural history of thoracic aortic aneurysms depend on their location. Medial degeneration is the most common pathology associated with ascending aortic aneurysms, whereas atherosclerosis is the condition most frequently associated with aneurysms of the descending thoracic aorta. The average growth rate of thoracic aneurysms is 0.1–0.2 cm per year. Thoracic aortic aneurysms associated with Marfan's syndrome or aortic dissection may expand at a greater rate. The risk of rupture is related to the size of the aneurysm and the presence of symptoms, ranging approximately from 2–3% per year for thoracic aortic aneurysms <4.0 cm in diameter to 7% per year for those >6 cm in diameter. Most thoracic aortic aneurysms are asymptomatic; however, compression or erosion of adjacent tissue by aneurysms may cause symptoms such as chest pain, shortness of breath, cough, hoarseness, and dysphagia. Aneurysmal dilation of the ascending aorta may cause congestive heart failure as a consequence of aortic regurgitation, and compression of the superior vena cava may produce congestion of the head, neck, and upper extremities.

A chest x-ray may be the first test that suggests the diagnosis of a thoracic aortic aneurysm **(Fig. 274-1)**. Findings include widening of the mediastinal shadow and displacement or compression of the trachea or left main stem bronchus. Echocardiography, particularly transesophageal echocardiography, can be used to assess the proximal ascending aorta and descending thoracic aorta. Contrast-enhanced CT, magnetic resonance imaging (MRI), and conventional invasive aortography are sensitive and specific tests for assessment of aneurysms of the thoracic aorta and involvement of branch vessels **(Fig. 274-2)**. In asymptomatic patients whose aneurysms are too small to justify surgery, noninvasive testing with either contrast-enhanced CT or MRI should be performed at least every 6–12 months to monitor expansion.

FIGURE 274-1

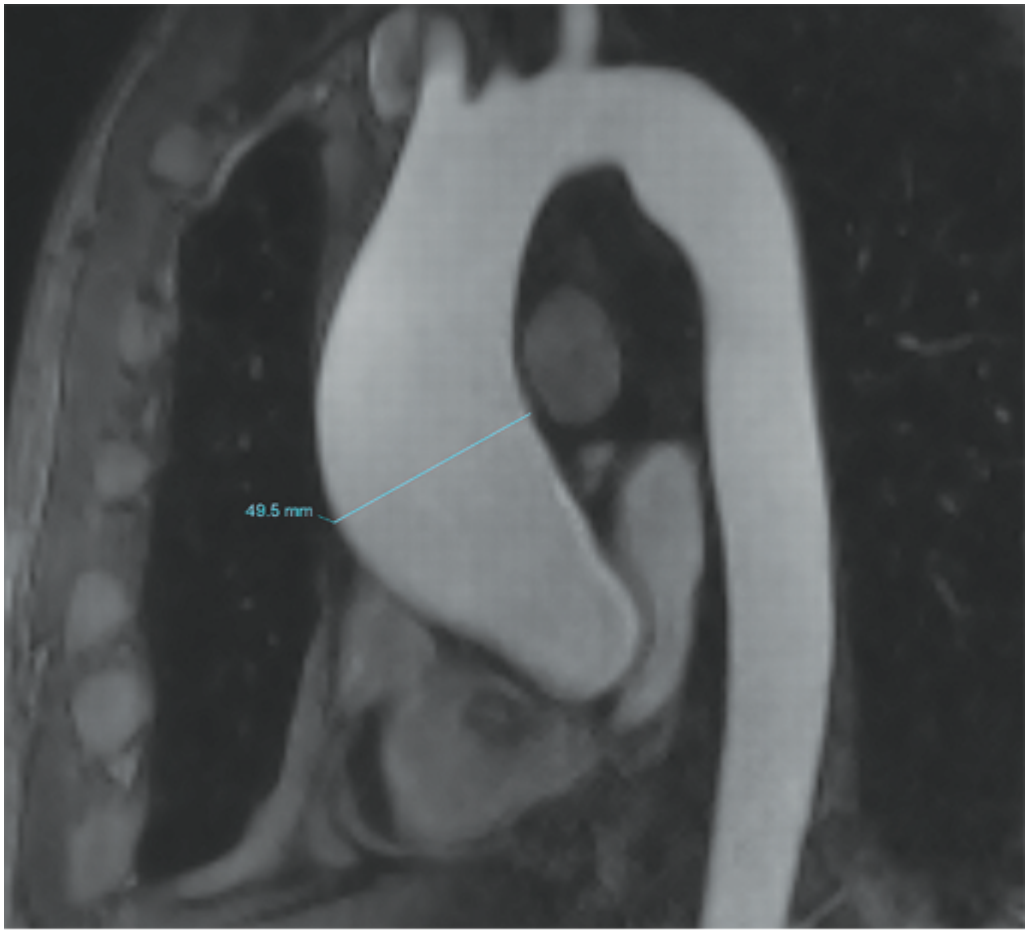
A chest x-ray of a patient with a thoracic aortic aneurysm.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 274-2

A magnetic resonance angiogram demonstrating a fusiform aneurysm of the ascending thoracic aorta. *(Courtesy of Dr. Michael Steigner, Brigham and Women's Hospital, Boston, MA, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT Thoracic Aortic Aneurysms

β-Adrenergic blockers currently are recommended for patients with thoracic aortic aneurysms, particularly those with Marfan's syndrome, who have evidence of aortic root dilatation to reduce the rate of further expansion. Additional medical therapy should be given as necessary to control hypertension. Angiotensin receptor antagonists may reduce the rate of aortic dilation in patients with Marfan's syndrome by blocking TGF-β signaling. Clinical outcome trials have found that the rate of aortic root enlargement in patients with Marfan's syndrome was similar with atenolol and losartan. Operative repair with placement of a prosthetic graft is indicated in patients with symptomatic ascending thoracic aortic aneurysms and for most asymptomatic aneurysms, including those associated with bicuspid aortic valves, when the aortic root or ascending aortic diameter is ≥5.5 cm, or when the growth rate is >0.5 cm per year. Replacement of the ascending aorta >4.5 cm is reasonable in patients with bicuspid aortic valves undergoing aortic valve replacement because of severe aortic stenosis or aortic regurgitation. In patients with Marfan's syndrome, ascending thoracic aortic aneurysms of 4–5 cm should be considered for surgery. Operative repair is indicated for patients with degenerative descending thoracic aortic aneurysms when the diameter is >6 cm, and endovascular repair should be considered if feasible when the diameter is >5.5 cm. Repair is also recommended when the diameter of a descending thoracic aortic aneurysm has increased >1 cm per year.

ABDOMINAL AORTIC ANEURYSMS

Abdominal aortic aneurysms occur more frequently in males than in females, and the incidence increases with age. Cigarette smoking is a potent modifiable risk factor. Abdominal aortic aneurysms ≥4.0 cm may affect 1–2% men aged >50 years. At least 90% of all abdominal aortic aneurysms >4.0 cm are related to atherosclerotic disease, and most of these aneurysms are below the level of the renal arteries. Prognosis is related to both the size of the aneurysm and the severity of coexisting coronary artery and cerebrovascular disease. The risk of rupture increases with the size of the aneurysm: the 5-year risk for aneurysms <5 cm is 1–2%, whereas it is 20–40% for aneurysms >5 cm in diameter. The formation of mural thrombi within aneurysms may predispose to peripheral embolization.

An abdominal aortic aneurysm commonly produces no symptoms. It usually is detected on routine examination as a palpable, pulsatile, expansile, and nontender mass, or it is an incidental finding observed on an abdominal imaging study performed for other reasons. As abdominal aortic aneurysms expand, however, they may become painful. Some patients complain of strong pulsations in the abdomen; others experience pain in the chest, lower back, or scrotum. Aneurysmal pain is usually a harbinger of rupture and represents a medical emergency. More often, acute rupture occurs without any prior warning, and this complication is always life-threatening. Rarely, there is leakage of the aneurysm with severe pain and tenderness. Acute pain and hypotension occur with rupture of the aneurysm, which requires an emergency operation.

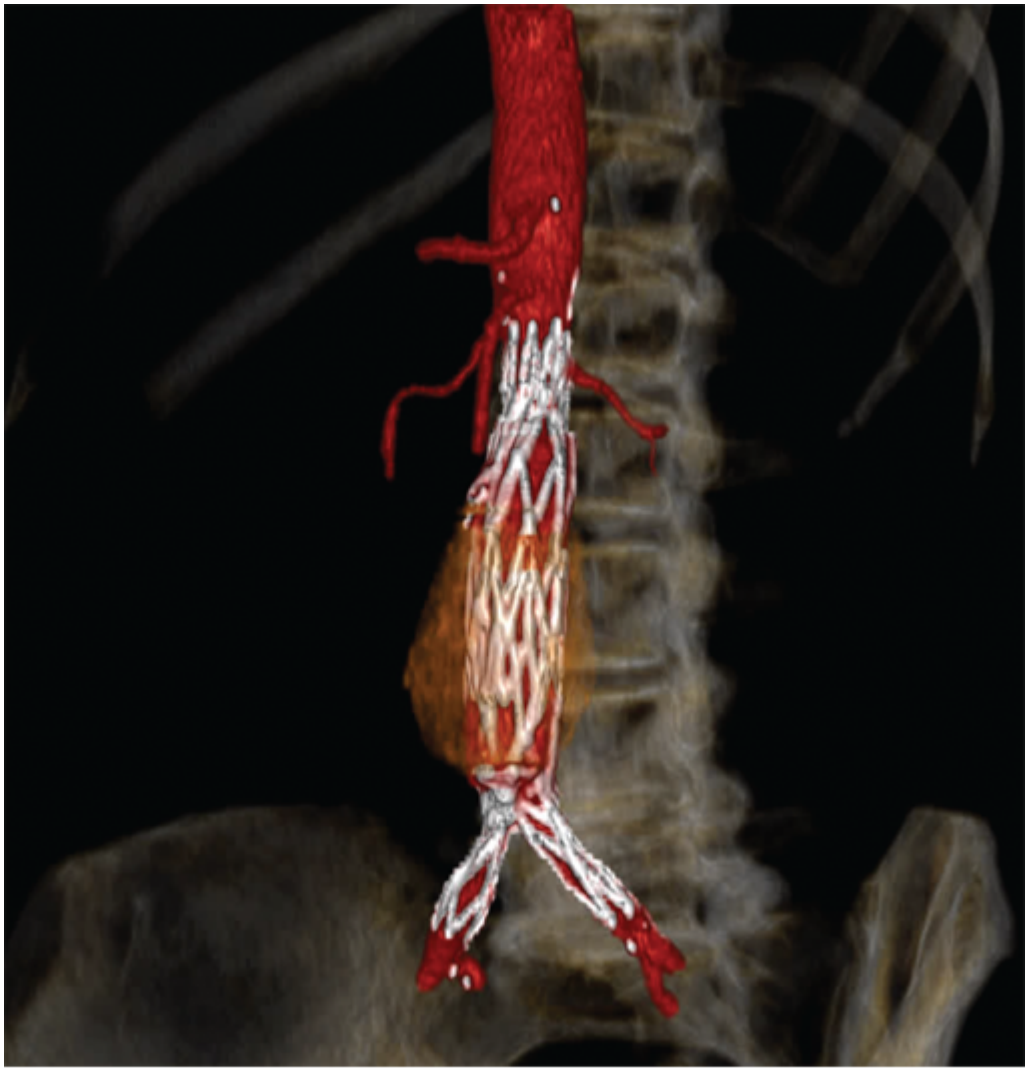
Abdominal radiography may demonstrate the calcified outline of the aneurysm; however, about 25% of aneurysms are not calcified and cannot be visualized by x-ray imaging. An abdominal ultrasound can delineate the transverse and longitudinal dimensions of an abdominal aortic aneurysm and may detect mural thrombus. Abdominal ultrasound is useful for serial documentation of aneurysm size and can be used to screen patients at risk for developing an aortic aneurysm. In one large study, ultrasound screening of men aged 65–74 years was associated with a risk reduction in aneurysm-related death of 42%. For this reason, screening by ultrasonography is recommended for men aged 65–75 years who have ever smoked. In addition, siblings or offspring of persons with abdominal aortic aneurysms, as well as individuals with thoracic aortic or peripheral arterial aneurysms, should be considered for screening for abdominal aortic aneurysms. CT with contrast and MRI are accurate noninvasive tests to determine the location and size of abdominal aortic aneurysms and to plan endovascular or open surgical repair **(Fig. 274-3)**. Contrast aortography may be used for the evaluation of patients with aneurysms, but the procedure carries a small risk of complications such as bleeding, allergic reactions, and atheroembolism. Since the presence of mural thrombi may reduce the luminal size, aortography may underestimate the diameter of an aneurysm.

FIGURE 274-3

A computed tomographic angiogram depicting a fusiform abdominal aortic aneurysm before (*left*) and after (*right*) treatment with a bifurcated stent graft. *(Courtesy of Drs. Elizabeth George and Frank Rybicki, Brigham and Women's Hospital, Boston, MA, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT Abdominal Aortic Aneurysms

Statins are indicated to reduce the risk of cardiovascular events related to atherosclerosis. Medical therapies, such as β-adrenergic blockers and renin-angiotensin inhibitors, have not proven effective in reducing the rate of aneurysm growth. Operative repair of the aneurysm with insertion of a prosthetic graft or endovascular placement of an aortic stent graft (Fig. 274-3) is indicated for abdominal aortic aneurysms of any size that are expanding rapidly or are associated with symptoms. For asymptomatic aneurysms, abdominal aortic aneurysm repair is indicated if the diameter is ≥5.5 cm. In randomized trials of patients with abdominal aortic aneurysms <5.5 cm, there was no difference in the long-term (5- to 8-year) mortality rate between those followed with ultrasound surveillance and those undergoing elective surgical repair. Thus, serial noninvasive follow-up of smaller aneurysms (<5.5 cm) is an alternative to immediate repair. The decision to perform an open surgical operation or endovascular repair is based in part on the vascular anatomy and comorbid conditions. Endovascular repair of abdominal aortic aneurysms has a lower short-term morbidity rate, but a comparable long-term mortality rate with open surgical reconstruction. Long-term surveillance with CT or MR aortography is indicated after endovascular repair to detect leaks and possible aneurysm expansion.

In surgical candidates, careful preoperative cardiac and general medical evaluations (followed by appropriate therapy for complicating conditions) are essential. Preexisting coronary artery disease, congestive heart failure, pulmonary disease, diabetes mellitus, and advanced age add to the risk of surgery. With careful preoperative cardiac evaluation and postoperative care, the operative mortality rate approximates 1–2%. After acute rupture, the mortality rate of emergent operation is 45–50%. Endovascular repair with stent placement is an alternative approach to treat ruptured aneurysms and may be associated with a lower mortality rate.

ACUTE AORTIC SYNDROMES

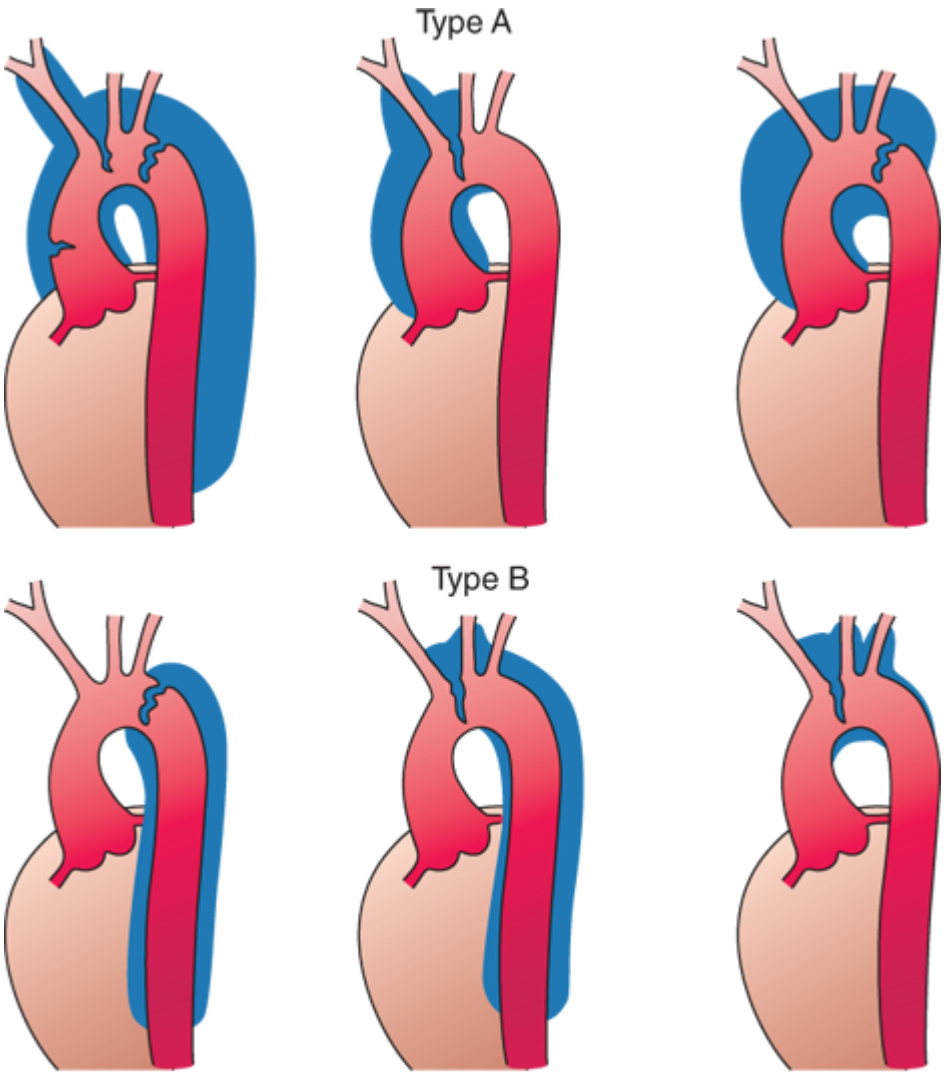
The four major acute aortic syndromes are aortic rupture (discussed earlier), aortic dissection, intramural hematoma, and penetrating atherosclerotic ulcer. Aortic dissection is caused by a circumferential or, less frequently, transverse tear of the intima. It often occurs along the right lateral wall of the ascending aorta where the hydraulic shear stress is high. Another common site is the descending thoracic aorta just below the ligamentum arteriosum. The initiating event is either a primary intimal tear with secondary dissection into the media or a medial hemorrhage that dissects into and disrupts the intima. The pulsatile aortic flow then dissects along the elastic lamellar plates of the aorta and creates a false lumen. The dissection usually propagates distally down the descending aorta and into its major branches, but it may propagate proximally. Distal propagation may be limited by atherosclerotic plaque. In some cases, a secondary distal intimal disruption occurs, resulting in the reentry of blood from the false to the true lumen.

There are at least two important pathologic and radiologic variants of aortic dissection: intramural hematoma without an intimal flap and penetrating atherosclerotic ulcer. Acute intramural hematoma is thought to result from rupture of the vasa vasorum with hemorrhage into the wall of the aorta. Most of these hematomas occur in the descending thoracic aorta. Acute intramural hematomas may progress to dissection and rupture. Penetrating atherosclerotic ulcers are caused by erosion of a plaque into the aortic media, are usually localized, and are not associated with extensive propagation. They are found primarily in the middle and distal portions of the descending thoracic aorta and are associated with extensive atherosclerotic disease. The ulcer can erode beyond the internal elastic lamina, leading to medial hematoma, and may progress to false aneurysm formation or rupture.

Several classification schemes have been developed for thoracic aortic dissections. DeBakey and colleagues initially classified aortic dissections as type I, in which an intimal tear occurs in the ascending aorta but involves the descending aorta as well; type II, in which the dissection is limited to the ascending aorta; and type III, in which the intimal tear is located in the descending aorta with distal propagation of the dissection **(Fig. 274-4)**. Another classification (Stanford) is that of type A, in which the dissection involves the ascending aorta (proximal dissection), and type B, in which it is limited to the arch and/or descending aorta (distal dissection). From a management standpoint, classification of aortic dissections and intramural hematomas into type A or B is more practical and useful, since DeBakey types I and II are managed in a similar manner.

FIGURE 274-4

Classification of aortic dissections. Stanford classification: Type A dissections (*top*) involve the ascending aorta independent of site of tear and distal extension; type B dissections (*bottom*) involve transverse and/or descending aorta without involvement of the ascending aorta. DeBakey classification: Type I dissection involves ascending to descending aorta (*top left*); type II dissection is limited to ascending or transverse aorta, without descending aorta (*top center* + *top right*); type III dissection involves descending aorta only (*bottom left*). *(From DC Miller, in RM Doroghazi, EE Slater [eds]: Aortic Dissection. New York, McGraw-Hill, 1983, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The factors that predispose to aortic dissection include those associated with medial degeneration and others that increase aortic wall stress (Table 274-1). Systemic hypertension is a coexisting condition in 70% of patients. Aortic dissection is the major cause of morbidity and mortality in patients with Marfan's syndrome **(Chap. 406)** or Loeys-Dietz syndrome, and similarly may affect patients with Ehlers-Danlos syndrome. The incidence also is increased in patients with inflammatory aortitis (i.e., Takayasu's arteritis, giant cell arteritis), congenital aortic valve anomalies (e.g., bicuspid valve), coarctation of the aorta, and a history of aortic trauma. In addition, the risk of dissection is increased in otherwise normal women during the third trimester of pregnancy. Aortic dissection also may occur as a consequence of weight lifting, cocaine use, or deceleration injury.

CLINICAL MANIFESTATIONS

The peak incidence of aortic dissection is in the sixth and seventh decades. Men are more affected than women by a ratio of 2:1. The presentations of aortic dissection and its variants are the consequences of intimal tear, dissecting hematoma, occlusion of involved arteries, and compression of adjacent tissues. Acute aortic dissection presents with the sudden onset of pain **(Chap. 11)**, which often is described as very severe and tearing and is associated with diaphoresis. The pain may be localized to the front or back of the chest, often the interscapular region, and typically migrates with propagation of the dissection. Other symptoms include syncope, dyspnea, and weakness. Physical findings may include hypertension or hypotension, loss of pulses, aortic regurgitation, pulmonary

edema, and neurologic findings due to carotid artery obstruction (hemiplegia, hemianesthesia) or spinal cord ischemia (paraplegia). Bowel ischemia, hematuria, and myocardial ischemia have all been observed. These clinical manifestations reflect complications resulting from the dissection occluding the major arteries. Furthermore, clinical manifestations may result from the compression of adjacent structures (e.g., superior cervical ganglia, superior vena cava, bronchus, esophagus) by the expanding dissection causing aneurysmal dilation, and include Horner's syndrome, superior vena cava syndrome, hoarseness, dysphagia, and airway compromise. Hemopericardium and cardiac tamponade may complicate a type A lesion with retrograde dissection. Acute aortic regurgitation is an important and common (>50%) complication of proximal dissection. It is the outcome of either a circumferential tear that widens the aortic root or a disruption of the annulus by a dissecting hematoma that tears a leaflet(s) or displaces it, inferior to the line of closure. Signs of aortic regurgitation include bounding pulses, a wide pulse pressure, a diastolic murmur often radiating along the right sternal border, and evidence of congestive heart failure. The clinical manifestations depend on the severity of the regurgitation.

In dissections involving the ascending aorta, the chest x-ray often reveals a widened superior mediastinum. A pleural effusion (usually left-sided) also may be present. This effusion is typically serosanguineous and not indicative of rupture unless accompanied by hypotension and falling hematocrit. In dissections of the descending thoracic aorta, a widened mediastinum may be observed on chest x-ray. In addition, the descending aorta may appear to be wider than the ascending portion. An electrocardiogram that shows no evidence of myocardial ischemia is helpful in distinguishing aortic dissection from myocardial infarction. Rarely, the dissection involves the right or, less commonly, left coronary ostium and causes acute myocardial infarction.

The diagnosis of aortic dissection can be established by noninvasive techniques such as echocardiography, CT, and MRI. Aortography is used less commonly because of the accuracy of these noninvasive techniques. Transthoracic echocardiography can be performed simply and rapidly and has an overall sensitivity of 60–85% for aortic dissection. For diagnosing proximal ascending aortic dissections, its sensitivity exceeds 80%; it is less useful for detecting dissection of the arch and descending thoracic aorta. Transesophageal echocardiography requires greater skill and patient cooperation, but is very accurate in identifying dissections of the ascending and descending thoracic aorta but not the arch, achieving 98% sensitivity and ~90% specificity. Echocardiography also provides important information regarding the presence and severity of aortic regurgitation and pericardial effusion. CT and MRI are both highly accurate in identifying the intimal flap and the extent of the dissection and involvement of major arteries; each has a sensitivity and specificity >90%. They are useful in recognizing intramural hemorrhage and penetrating ulcers. The relative utility of transesophageal echocardiography, CT, and MRI depends on the availability and expertise in individual institutions as well as on the hemodynamic stability of the patient, with CT and MRI obviously less suitable for unstable patients.

TREATMENT

TREATMENT Aortic Dissection

Medical therapy should be initiated as soon as the diagnosis is considered. The patient should be admitted to an intensive care unit for hemodynamic monitoring. Unless hypotension is present, therapy should be aimed at reducing cardiac contractility and systemic arterial pressure, and thus shear stress. For acute dissection, unless contraindicated, β-adrenergic blockers should be administered parenterally, using intravenous propranolol, metoprolol, or the short-acting esmolol to achieve a heart rate of ~60 beats/min. This should be accompanied by sodium nitroprusside infusion to lower systolic blood pressure to ≤120 mmHg. Labetalol **(Chap. 271)**, a drug with both β- and α-adrenergic blocking properties, also may be used as a parenteral agent in acute therapy for dissection.

The calcium channel antagonists verapamil and diltiazem may be used intravenously if nitroprusside or β-adrenergic blockers cannot be employed. The addition of a parenteral angiotensin-converting enzyme (ACE) inhibitor such as enalaprilat to a β-adrenergic blocker also may be considered. Isolated use of a direct vasodilator such as hydralazine is contraindicated because these agents can increase hydraulic shear and may propagate the dissection.

Emergent or urgent surgical correction is the preferred treatment for acute ascending aortic dissections and intramural hematomas (type A). Surgery involves excision of the intimal flap, obliteration of the false lumen, and placement of an interposition graft. Aortic valve repair or a composite valve-graft conduit is used if the aortic valve is disrupted. The overall in-hospital mortality rate after surgical treatment of patients with aortic dissection is reported to be 15–25%. The major causes of perioperative mortality and morbidity include myocardial infarction, paraplegia, renal failure, tamponade, hemorrhage, and sepsis. Thoracic endovascular aortic repair with an endoluminal stent graft is indicated for complicated type B dissections, including those characterized by propagation, compromise of major aortic branches, impending rupture, or continued pain. Other transcatheter techniques, such as fenestration of the intimal flaps and stenting of narrowed branch vessels to increase flow to compromised organs, are used in selected patients. Surgical correction is indicated for complicated type B dissections, particularly if endovascular repair is not feasible. Hybrid procedures consisting of both surgery and endovascular repair may be used when the dissection involves both the aortic arch and the descending thoracic aorta. For uncomplicated and stable distal dissections and intramural hematomas (type B), medical therapy is the preferred treatment. The in-hospital mortality rate of medically treated patients with type B dissection is ~10%. Long-term therapy for patients with aortic dissection and intramural hematomas (with or without surgery) consists of control of hypertension and reduction of cardiac contractility with the use of β-adrenergic blockers plus other antihypertensive agents, such as ACE inhibitors or calcium antagonists. Patients with chronic type B dissection and intramural hematomas should be followed on an outpatient basis every 6–12 months with contrast-enhanced CT or MRI to detect propagation or expansion. Patients with Marfan's syndrome are at high risk for postdissection complications. The long-term prognosis for patients with treated dissections is generally good with careful follow-up; the 10-year survival rate is ~60%.

CHRONIC ATHEROSCLEROTIC OCCLUSIVE DISEASE

Atherosclerosis may affect the thoracic and abdominal aorta. Occlusive aortic disease caused by atherosclerosis usually is confined to the distal abdominal aorta below the renal arteries. Frequently the disease extends to the iliac arteries **(Chap. 275)**. Claudication characteristically involves the buttocks, thighs, and calves and may be associated with impotence in males (Leriche's syndrome). The severity of the symptoms depends on the adequacy of collaterals. With sufficient collateral blood flow, a complete occlusion of the abdominal aorta may occur without the development of ischemic symptoms. The physical findings include the absence of femoral and other distal pulses bilaterally and the detection of an audible bruit over the abdomen (usually at or below the umbilicus) and the common femoral arteries. Atrophic skin, loss of hair, and coolness of the lower extremities usually are observed. In advanced ischemia, rubor on dependency and pallor on elevation can be seen.

The diagnosis usually is established by physical examination and noninvasive testing, including leg pressure measurements, Doppler velocity analysis, pulse volume recordings, and duplex ultrasonography. The anatomy may be defined by MRI, CT, or conventional aortography, typically performed when one is considering revascularization. Catheter-based endovascular or operative treatment is indicated in patients with lifestyle-limiting or debilitating symptoms of claudication and patients with critical limb ischemia.

ACUTE AORTIC OCCLUSION

Acute occlusion in the distal abdominal aorta constitutes a medical emergency because it threatens the viability of the lower extremities; it usually results from an occlusive (saddle) embolus that almost always originates from the heart. Rarely, acute occlusion may occur as the result of in situ thrombosis in a preexisting severely narrowed segment of the aorta.

The clinical picture is one of acute ischemia of the lower extremities. Severe rest pain, coolness, and pallor of the lower extremities and the absence of distal pulses bilaterally are the usual manifestations. Diagnosis should be established rapidly by MRI, CT, or aortography. Emergency thrombectomy or revascularization is indicated.

AORTITIS

Aortitis, a term referring to inflammatory disease of the aorta, may be caused by large vessel vasculitides such as Takayasu's arteritis and giant cell arteritis, rheumatic and HLA-B27–associated spondyloarthropathies, Behçet's syndrome, antineutrophil cytoplasmic antibody (ANCA)–associated vasculitides, Cogan's syndrome, Erdheim-Chester disease, IgG4-related systemic disease, and infections such as syphilis, tuberculosis, and *Salmonella*, or may be associated with retroperitoneal fibrosis. Aortitis may result in aneurysmal dilation and aortic regurgitation, occlusion of the aorta and its branch vessels, or acute aortic syndromes.

TAKAYASU'S ARTERITIS

(See also Chap. 356) This inflammatory disease often affects the ascending aorta and aortic arch, causing obstruction of the aorta and its major arteries. Takayasu's arteritis is also termed *pulseless disease* because of the frequent occlusion of the large arteries originating from the aorta. It also may involve the descending thoracic and abdominal aorta and occlude large branches such as the renal arteries. Aortic aneurysms also may occur. The pathology is a panarteritis characterized by mononuclear cells and occasionally giant cells, with marked intimal hyperplasia, medial and adventitial thickening, and, in the chronic form, fibrotic occlusion. The disease is most prevalent in young females of Asian descent but does occur in women of other geographic and ethnic origins and also in young men. During the acute stage, fever, malaise, weight loss, and other systemic symptoms may be evident. Elevations of the erythrocyte sedimentation rate and C-reactive protein are common. The chronic stages of the disease, which is intermittently active, present with symptoms related to large artery occlusion, such as upper extremity claudication, cerebral ischemia, and syncope. The process is progressive, and there is no definitive therapy. Glucocorticoids and immunosuppressive agents are effective in some patients during the acute phase. Biologically targeted agents are under investigation. Surgical bypass or endovascular intervention of a critically stenotic artery may be necessary.

GIANT CELL ARTERITIS

(See also Chap. 356) This vasculitis occurs in older individuals and affects women more often than men. Primarily large and medium-size arteries are affected. The pathology is that of focal granulomatous lesions involving the entire arterial wall; it frequently is associated with polymyalgia rheumatica. Obstruction of medium-size arteries (e.g., temporal and ophthalmic arteries) and major branches of the aorta and the development of aortitis and aortic regurgitation are important complications of the disease. High-dose glucocorticoid therapy should be administered early and then gradually tapered. Immunosuppressive therapy with methotrexate may allow reduction in steroid dosage and reduce the risk of relapse. Biologically targeted therapies are under investigation.

RHEUMATIC AORTITIS

Rheumatoid arthritis **(Chap. 351)**, ankylosing spondylitis **(Chap. 355)**, psoriatic arthritis **(Chap. 355)**, reactive arthritis (formerly known as Reiter's syndrome) **(Chap. 355)**, relapsing polychondritis, and inflammatory bowel disorders may all be associated with aortitis involving the ascending aorta. The inflammatory lesions usually involve the ascending aorta and may extend to the sinuses of Valsalva, the mitral valve leaflets, and adjacent myocardium. The clinical manifestations are aneurysm, aortic regurgitation, and involvement of the cardiac conduction system.

IDIOPATHIC AORTITIS

Idiopathic abdominal aortitis is characterized by adventitial and periaortic inflammation with thickening of the aortic wall. It is associated with abdominal aortic aneurysms and idiopathic retroperitoneal fibrosis. Affected individuals may present with vague constitutional symptoms, fever, and abdominal pain. Retroperitoneal fibrosis can cause ureteral obstruction and hydronephrosis. Glucocorticoids and immunosuppressive agents may reduce the inflammation.

INFECTIVE AORTITIS

Infective aortitis may result from direct invasion of the aortic wall by bacterial pathogens such as *Staphylococcus*, *Streptococcus*, and *Salmonella* or by fungi. These bacteria cause aortitis by infecting the aorta at sites of atherosclerotic plaque. Bacterial proteases lead to degradation of collagen, and the ensuing destruction of the aortic wall leads to the formation of a saccular aneurysm referred to as a mycotic aneurysm. Mycotic aneurysms have a predilection for the suprarenal abdominal aorta. The pathologic characteristics of the aortic wall include acute and chronic inflammation, abscesses, hemorrhage, and necrosis. Mycotic aneurysms typically affect the elderly and occur in men three times more frequently than in women. Patients may present with fever, sepsis, and chest, back, or abdominal pain; there may have been a preceding diarrheal illness. Blood cultures are positive in the majority of patients. Both CT and MRI are useful to diagnose mycotic aneurysms. Treatment includes antibiotic therapy and surgical removal of the affected part of the aorta and revascularization of the lower extremities with grafts placed in uninfected tissue.

Syphilitic aortitis is a late manifestation of luetic infection **(Chap. 177)** that usually affects the proximal ascending aorta, particularly the aortic root, resulting in aortic dilation and aneurysm formation. Syphilitic aortitis occasionally may involve the aortic arch or the descending aorta. The aneurysms may be saccular or fusiform and are usually asymptomatic, but compression of and erosion into adjacent structures may result in symptoms; rupture also may occur.

The initial lesion is an obliterative endarteritis of the vasa vasorum, especially in the adventitia. This is an inflammatory response to the invasion of the adventitia by the spirochetes. Destruction of the aortic media occurs as the spirochetes spread into this layer, usually via the lymphatics accompanying the vasa vasorum. Destruction of collagen and elastic tissues leads to dilation of the aorta, scar formation, and calcification. These changes account for the characteristic radiographic appearance of linear calcification of the ascending aorta.

The disease typically presents as an incidental chest radiographic finding 15–30 years after initial infection. Symptoms may result from aortic regurgitation, narrowing of coronary ostia due to syphilitic aortitis, compression of adjacent structures (e.g., esophagus), or rupture. Diagnosis is established by a positive serologic test, i.e., rapid plasmin reagin (RPR) or fluorescent treponemal antibody. Treatment includes penicillin and surgical excision and repair.

FURTHER READING

Buck DB et al: Endovascular treatment of abdominal aortic aneurysms. Nat Rev Cardiol 11:112, 2014. [PubMed: 24343568]

Erbel R et al: 2014 ESC Guidelines on the diagnosis and treatment of aortic diseases: Document covering acute and chronic aortic diseases of the thoracic and abdominal aorta of the adult. The Task Force for the Diagnosis and Treatment of Aortic Diseases of the European Society of Cardiology (ESC). Eur Heart J 35:2873, 2014. [PubMed: 25173340]
Gornik HL, Creager MA: Aortitis. Circulation 117:3039, 2008. [PubMed: 18541754]
Guirguis-Blake JM et al: Ultrasonography screening for abdominal aortic aneurysms: A systematic evidence review for the US Preventive Services Task Force. Ann Intern Med 160:321, 2014. [PubMed: 24473919]
Hiratzka LF et al: 2010 ACCF/AHA/AATS/ACR/ASA/SCA/SCAI/SIR/STS/SVM guidelines for the diagnosis and management of patients with Thoracic Aortic Disease: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, American Association for Thoracic Surgery, American College of Radiology, American Stroke Association, Society of Cardiovascular Anesthesiologists, Society for Cardiovascular Angiography and Interventions, Society of Interventional Radiology, Society of Thoracic Surgeons, and Society for Vascular Medicine. Circulation 121:e266, 2010. [PubMed: 20233780]
Isselbacher EM et al: Hereditary influence in thoracic aortic aneurysm and dissection. Circulation 133:2516, 2016. [PubMed: 27297344]
Kent KC: Clinical practice. Abdominal aortic aneurysms. N Engl J Med 371:2101, 2014. [PubMed: 25427112]
Nienaber CA, Clough RE: Management of acute aortic dissection. Lancet 385:800, 2015. [PubMed: 25662791]

Chapter 275: Arterial Diseases of the Extremities

Mark A. Creager; Joseph Loscalzo

PERIPHERAL ARTERY DISEASE

Peripheral artery disease (PAD) is defined as a clinical disorder in which there is a stenosis or occlusion in the aorta or the arteries of the limbs. Atherosclerosis is the leading cause of PAD in patients >40 years old. Other causes include thrombosis, embolism, vasculitis, fibromuscular dysplasia, entrapment, cystic adventitial disease, and trauma. The highest prevalence of atherosclerotic PAD occurs in the sixth and seventh decades of life. As in patients with atherosclerosis of the coronary and cerebral vasculature, there is an increased risk of developing PAD in cigarette smokers and in persons with diabetes mellitus, hypercholesterolemia, hypertension, or renal insufficiency.

Pathology

(See also Chap. 291e from HPIM 19e) Segmental lesions that cause stenosis or occlusion are usually localized to large and medium-size vessels. The pathology of the lesions includes atherosclerotic plaques with calcium deposition, thinning of the media, patchy destruction of muscle and elastic fibers, fragmentation of the internal elastic lamina, and thrombi composed of platelets and fibrin. The primary sites of involvement are the abdominal aorta and iliac arteries (30% of symptomatic patients), the femoral and popliteal arteries (80–90% of patients), and the more distal vessels, including the tibial and peroneal arteries (40–50% of patients). Atherosclerotic lesions occur preferentially at arterial branch points, which are sites of increased turbulence, altered shear stress, and intimal injury. Involvement of the distal vasculature is most common in elderly individuals and patients with diabetes mellitus.

Clinical Evaluation

Fewer than 50% of patients with PAD are symptomatic, although many have a slow or impaired gait. The most common *symptom* is intermittent claudication, which is defined as a pain, ache, cramp, numbness, or a sense of fatigue in the muscles; it occurs during exercise and is relieved by rest. The site of claudication is distal to the location of the occlusive lesion. For example, buttock, hip, thigh, and calf discomfort occurs in patients with aortoiliac disease, whereas calf claudication develops in patients with femoral-popliteal disease. Symptoms are far more common in the lower than in the upper extremities because of the higher incidence of obstructive lesions in the former region. In patients with severe arterial occlusive disease in whom resting blood flow cannot accommodate basal nutritional needs of the tissues, critical limb ischemia may develop. Patients complain of rest pain or a feeling of cold or numbness in the foot and toes. Frequently, these symptoms occur at night when the legs are horizontal and improve when the legs are in a dependent position. With severe ischemia, rest pain may be persistent.

Important *physical findings* of PAD include decreased or absent pulses distal to the obstruction, the presence of bruits over the narrowed artery, and muscle atrophy. With more severe disease, hair loss, thickened nails, smooth and shiny skin, reduced skin temperature, and pallor or cyanosis are common physical signs. In patients with critical limb ischemia, ulcers or gangrene may occur. Elevation of the legs and repeated flexing of the calf muscles produce pallor of the soles of the feet, whereas rubor, secondary to reactive hyperemia, may develop when the legs are dependent. The time required for rubor to develop or for the veins in the foot to fill when the patient’s legs are transferred from an elevated to a dependent position is related to the severity of the ischemia and the presence of collateral vessels. Patients with severe ischemia may develop peripheral edema because they keep their legs in a dependent position much of the time. Ischemic neuropathy can result in numbness and hyporeflexia.

Noninvasive Testing

The history and physical examination are often sufficient to establish the diagnosis of PAD. An objective assessment of the presence and severity of disease is obtained by noninvasive techniques. Arterial pressure can be recorded noninvasively in the legs by placement of sphygmomanometric cuffs at the ankles and the use of a Doppler device to auscultate or record blood flow from the dorsalis pedis and posterior tibial arteries. Normally, systolic blood pressure in the legs and arms is similar. Indeed, ankle pressure may be slightly higher than arm pressure due to pulse-wave

amplification. In the presence of hemodynamically significant stenoses, the systolic blood pressure in the leg is decreased. Thus, the ratio of the ankle and brachial artery pressures (termed the *ankle:brachial index*, or ABI) is 1.00–1.40 in normal individuals. ABI values of 0.91–0.99 are considered "borderline," and those <0.90 are abnormal and diagnostic of PAD. ABIs >1.40 indicate noncompressible arteries secondary to vascular calcification.

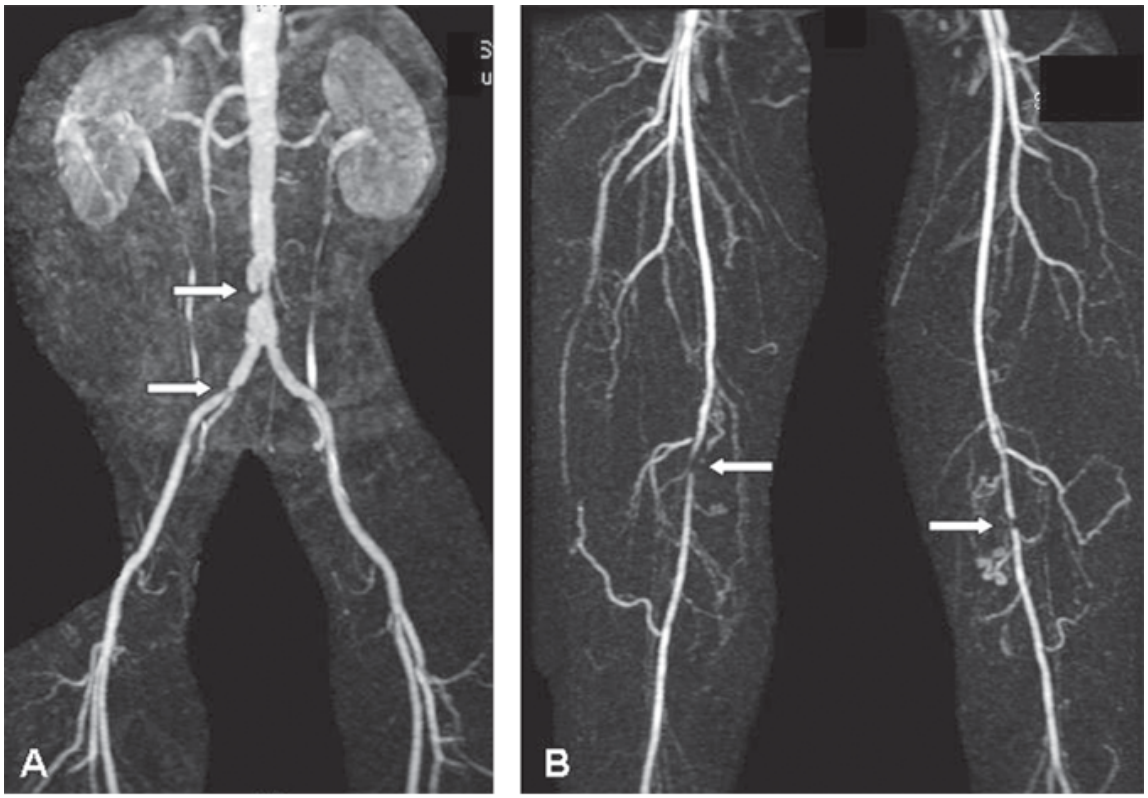
Other noninvasive tests include segmental pressure measurements, segmental pulse volume recordings, duplex ultrasonography (which combines B-mode imaging and Doppler flow velocity waveform analysis), transcutaneous oximetry, and stress testing (usually using a treadmill). Placement of pneumatic cuffs enables assessment of systolic pressure along the legs. The presence of pressure gradients between sequential cuffs provides evidence of the presence and location of hemodynamically significant stenoses. In addition, the amplitude of the pulse volume contour becomes blunted in the presence of significant PAD. Duplex ultrasonography is used to image and detect stenotic lesions in native arteries and bypass grafts.

Treadmill testing allows the physician to assess functional limitations objectively. Decline of the ABI immediately after exercise provides further support for the diagnosis of PAD in patients with equivocal symptoms and findings on examination.

Magnetic resonance angiography (MRA), computed tomographic angiography (CTA), and conventional catheter-based angiography should not be used for routine diagnostic testing, but are performed before potential revascularization **(Fig. 275-1)**. Each test is useful in defining the anatomy to assist planning for endovascular and surgical revascularization procedures.

FIGURE 275-1

Magnetic resonance angiography of a patient with intermittent claudication, showing stenoses of the distal abdominal aorta and right common iliac artery (***A***) and stenoses of the right and left superficial femoral arteries (***B***). *(Courtesy of Dr. Edwin Gravereaux, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prognosis

The natural history of patients with PAD is influenced primarily by the extent of coexisting coronary artery and cerebrovascular disease. Approximately one-third to one-half of patients with symptomatic PAD have evidence of coronary artery disease (CAD) based on clinical presentation and electrocardiogram, and over one-half have significant

CAD by coronary angiography. Patients with PAD have a 15–25% 5-year mortality rate and a two- to sixfold increased risk of death from coronary heart disease. Mortality rates are highest in those with the most severe PAD. Measurement of ABI is useful for detecting PAD and identifying persons at risk for future atherothrombotic events. The likelihood of symptomatic progression of PAD is lower than the chance of succumbing to CAD. Approximately 75–80% of nondiabetic patients who present with mild to moderate claudication remain symptomatically stable. Deterioration is likely to occur in the remainder, with ~1–2% of the group ultimately developing critical limb ischemia each year. Approximately 25–30% of patients with critical limb ischemia undergo amputation within 1 year. The prognosis is worse in patients who continue to smoke cigarettes or have diabetes mellitus.

TREATMENT

TREATMENT Peripheral Artery Disease

Patients with PAD should receive therapies to reduce the risk of associated cardiovascular events, such as myocardial infarction and death, and to improve limb symptoms, prevent progression to critical limb ischemia, and preserve limb viability. Risk factor modification and antiplatelet therapy should be initiated to improve cardiovascular outcomes. The importance of discontinuing cigarette smoking cannot be overemphasized. The physician must assume a major role in this lifestyle modification. Counseling and adjunctive drug therapy with the nicotine patch, bupropion, or varenicline increase smoking cessation rates and reduce recidivism. It is important to control blood pressure in hypertensive patients. Angiotensin-converting enzyme inhibitors and angiotensin receptor blockers may reduce the risk of cardiovascular events in patients with symptomatic PAD. β-Adrenergic blockers do not worsen claudication and may be used to treat hypertension, especially in patients with coexistent CAD. Treatment of hypercholesterolemia with statins is advocated to reduce the risk of myocardial infarction, stroke, and death. The 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults recommends high intensity statin treatment in patients with atherosclerotic disorders, including PAD. Platelet inhibitors, including aspirin and the ADP antagonist, clopidogrel, reduce the risk of adverse cardiovascular events in patients with atherosclerosis and are recommended for patients with symptomatic PAD, including those with intermittent claudication or critical limb ischemia or prior lower extremity revascularization. Outcomes with ticragrelor are similar to those with clopidogrel. The benefit of dual antiplatelet therapy with both aspirin and clopidogrel compared with aspirin alone in reducing cardiovascular morbidity and mortality rates in patients with PAD is uncertain. When added to other antiplatelet therapy, vorapaxar, a protease activated receptor-1 antagonist that inhibits thrombin-mediated platelet activation, decreases the risk of adverse cardiovascular events in patients with atherosclerosis, including PAD. It also reduces the risk of acute limb ischemia and peripheral revascularization; however, it is associated with an increased rate of moderate bleeding. The anticoagulant warfarin is as effective as antiplatelet therapy in preventing adverse cardiovascular events but causes more major bleeding; therefore, it is not indicated to improve outcomes in patients with chronic PAD. The combination of a low dose of the oral factor Xa inhibitor, rivaroxaban, and aspirin improves cardiovascular outcomes in patients with established atherosclerosis, including PAD, but is associated with increased risk of bleeding.

Therapies for intermittent claudication and critical limb ischemia include supportive measures, medications, exercise training, endovascular interventions, and surgery. Supportive measures include meticulous care of the feet, which should be kept clean and protected against excessive drying with moisturizing creams. Well-fitting and protective shoes are advised to reduce trauma. Elastic support hose should be avoided, as it reduces blood flow to the skin. In patients with critical limb ischemia, shock blocks under the head of the bed together with a canopy over the feet may improve perfusion pressure and ameliorate some of the rest pain.

Patients with claudication should be encouraged to exercise regularly and at progressively more strenuous levels. Supervised exercise training programs for 30- to 45-min sessions, three to five times per week for at least 12 weeks, prolong walking distance. The beneficial effect of supervised exercise training on walking performance in patients with claudication often is similar to or greater than that realized after a revascularization procedure. Structured home and community-based exercise programs are also effective. Pharmacologic treatment of PAD has not been as successful as the medical treatment of CAD (**Chap. 267**). In particular, vasodilators as a class have not proved to be beneficial. During exercise, peripheral vasodilation occurs distal to sites of significant arterial stenoses. As a result, perfusion pressure falls, often to levels lower than that generated in the interstitial tissue by the exercising muscle. Drugs such as α-adrenergic blocking agents, calcium channel antagonists, and other vasodilators have not been shown to be effective in patients with PAD.

Cilostazol, a phosphodiesterase inhibitor with vasodilator and antiplatelet properties, increases claudication distance by 40–60% and improves measures of quality of life. The mechanism of action accounting for its beneficial effects is not known. Pentoxifylline, a substituted xanthine derivative, increases blood flow to the microcirculation and enhances tissue oxygenation. Although several placebo-controlled studies have found that pentoxifylline modestly increases the duration of exercise, its efficacy has not been confirmed in other clinical trials. Statins appeared effective for treatment of intermittent claudication in initial clinical trials, but more studies are needed to confirm the efficacy of this class of drugs.

There is no definitive medical therapy for critical limb ischemia. Vasodilator prostaglandins are not effective in relieving symptoms or preventing limb loss. Enthusiasm for therapy with angiogenic growth factors abated when clinical trials of intramuscular gene transfer of DNA encoding vascular endothelial growth factor, fibroblast growth factor, hepatocyte growth factor, or hypoxia-inducible factor 1α failed to demonstrate improvement in symptoms or outcomes in patients with intermittent claudication or critical limb ischemia. Most clinical trials of bone marrow–derived vascular progenitor cells to promote angiogenesis and preserve limb viability in patients with critical limb ischemia have failed to demonstrate benefit, though a meta-analysis of these trials suggested a modest reduction in the risk of amputation. This remains an active area of preclinical and clinical investigation.

REVASCULARIZATION

Revascularization procedures, including catheter-based and surgical interventions, are usually indicated for patients with disabling, progressive, or severe symptoms of intermittent claudication despite medical therapy in order to improve walking distance and functional capacity. These are also indicated in patients with critical limb ischemia to relieve pain and prevent limb loss. MRA, CTA, or conventional angiography should be performed to assess vascular anatomy in patients who are being considered for revascularization. Endovascular interventions include percutaneous transluminal balloon angioplasty (PTA) (including drug coated balloons), stent placement (including drug eluting stents), stent grafts, and atherectomy **(Chap. 270)**. When endovascular intervention is performed in conjunction with a supervised exercise program, walking distance improves more than with exercise training alone.

PTA and stenting of the iliac artery are associated with higher success rates than are PTA and stenting of the femoral and popliteal arteries. Approximately 90–95% of iliac PTAs are initially successful, and the 3-year patency rate is >75%. Patency rates may be higher if a stent is placed in the iliac artery. The initial success rate for femoral-popliteal PTA and stenting approximate 90% with 60% 3-year patency rates. Outcomes following stenting of longer femoral-popliteal lesions (>5–10 cm) generally are better than after PTA. Several clinical trials have found lower restenosis rates with drug-coated balloons than with PTA, and with drug eluting stents compared with bare metal stents. Patency rates are influenced by the severity and length of pretreatment stenoses; the prognosis of occlusive lesions is worse than that of nonocclusive stenotic lesions. Endovascular interventions of the infrapoplital, tibial, and peroneal arteries, often in conjunction with treatment of more proximal lesions, can be used to treat critical limb ischemia and prevent limb loss.

Several operative procedures are available for treating patients with aortoiliac and femoral-popliteal artery disease. The preferred operative procedure depends on the location and extent of the obstruction(s) and the general medical condition of the patient. Operative procedures for aortoiliac disease include aortobifemoral bypass, axillofemoral bypass, femoro-femoral bypass, and aortoiliac endarterectomy. The most frequently used procedure is the aortobifemoral bypass using knitted Dacron grafts. Immediate graft patency approaches 99%, and 5- and 10-year graft patency rates in survivors are >90% and 80%, respectively. Operative complications include myocardial infarction and stroke, infection of the graft, peripheral embolization, and sexual dysfunction from interruption of autonomic nerves in the pelvis. The operative mortality rate ranges from 1 to 3%, mostly due to ischemic heart disease.

Operative therapy for femoral-popliteal artery disease includes in situ and reverse autogenous saphenous vein bypass grafts, placement of polytetrafluoroethylene (PTFE) or other synthetic grafts, and thromboendarterectomy. The operative mortality rate ranges from 1 to 3%. The long-term patency rate depends on the type of graft used, the location of the distal anastomosis, and the patency of runoff vessels beyond the anastomosis. Patency rates of femoral-popliteal saphenous vein bypass grafts approach 90% at 1 year and 70–80% at 5 years. Five-year patency rates of infrapopliteal saphenous vein bypass grafts are 60–70%. In contrast, 5-year patency rates of infrapopliteal PTFE grafts are <30%.

Preoperative cardiac risk assessment may identify individuals who are especially likely to experience an adverse cardiac event during the perioperative period. Patients with angina, prior myocardial infarction, heart failure, diabetes, or renal insufficiency are among those at increased risk. Stress testing with treadmill exercise (if feasible), radionuclide myocardial perfusion imaging, or echocardiography permits further stratification of risk in these patients, particularly those with poor or unknown functional capacity **(Chap. 270)**. Patients with abnormal test results require close supervision and adjunctive management with anti-ischemic medications. Coronary angiography and coronary artery revascularization compared with optimal medical therapy do not improve outcomes in most patients undergoing peripheral vascular surgery, but cardiac catheterization should be considered in patients with unstable angina and angina refractory to medical therapy as well as those suspected of having left main or three-vessel CAD.

FIBROMUSCULAR DYSPLASIA

Fibromuscular dysplasia is a hyperplastic disorder that affects medium-size and small arteries. It occurs predominantly in females and usually involves the renal and carotid arteries but can affect extremity vessels such as the iliac and subclavian arteries. The histologic classification includes intimal fibroplasia (also classified as focal), medial dysplasia (multifocal), and adventitial hyperplasia. Medial dysplasia is subdivided into medial fibroplasia, perimedial fibroplasia, and medial hyperplasia. Medial fibroplasia is the most common type and is characterized by alternating areas of thinned media and fibromuscular ridges. The internal elastic lamina usually is preserved. The iliac arteries are the limb arteries

most likely to be affected by fibromuscular dysplasia. It is identified angiographically by a “string of beads” appearance caused by thickened fibromuscular ridges contiguous with thin, less-involved portions of the arterial wall, which is typical of medial fibroplasia, or less commonly, as a focal tubular stenosis, and which is more typical of intimal fibroplasia. When limb vessels are involved, clinical manifestations are similar to those for atherosclerosis, including claudication and rest pain. PTA and surgical reconstruction have been beneficial in patients with debilitating symptoms or threatened limbs.

THROMBOANGIITIS OBLITERANS

Thromboangiitis obliterans (Buerger’s disease) is an inflammatory occlusive vascular disorder involving small and medium-size arteries and veins in the distal upper and lower extremities. Cerebral, visceral, and coronary vessels may be affected rarely. This disorder develops most frequently in men <40 years of age. The prevalence is higher in Asians and individuals of Eastern European descent. Although the cause of thromboangiitis obliterans is not known, there is a definite relationship to cigarette smoking in patients with this disorder.

In the initial stages of thromboangiitis obliterans, polymorphonuclear leukocytes infiltrate the walls of the small and medium-size arteries and veins. The internal elastic lamina is preserved, and a cellular, inflammatory thrombus develops in the vascular lumen. As the disease progresses, mononuclear cells, fibroblasts, and giant cells replace the neutrophils. Later stages are characterized by perivascular fibrosis, organized thrombus, and recanalization.

The clinical features of thromboangiitis obliterans often include a triad of claudication of the affected extremity, Raynaud’s phenomenon, and migratory superficial vein thrombophlebitis. Claudication usually is confined to the calves and feet or the forearms and hands because this disorder primarily affects distal vessels. In the presence of severe digital ischemia, trophic nail changes, painful ulcerations, and gangrene may develop at the tips of the fingers or toes. The physical examination shows normal brachial and popliteal pulses but reduced or absent radial, ulnar, and/or tibial pulses. MRA, CTA, and conventional arteriography are helpful in making the diagnosis. Smooth, tapering segmental lesions in the distal vessels are characteristic, as are collateral vessels at sites of vascular occlusion. Proximal atherosclerotic disease is usually absent. The diagnosis can be confirmed by excisional biopsy and pathologic examination of an involved vessel.

There is no specific treatment except abstention from tobacco. The prognosis is worse in individuals who continue to smoke, but results are discouraging even in those who stop smoking. Arterial bypass of the larger vessels may be used in selected instances, as well as local debridement, depending on the symptoms and severity of ischemia. Antibiotics may be useful; anticoagulants and glucocorticoids are not helpful. If these measures fail, amputation may be required.

VASCULITIS

Other vasculitides may affect the arteries that supply the upper and lower extremities. **Takayasu’s arteritis and giant cell (temporal) arteritis are discussed in Chap. 356.**

ACUTE LIMB ISCHEMIA

Acute limb ischemia occurs when arterial occlusion results in the sudden cessation of blood flow to an extremity. The severity of ischemia and the viability of the extremity depend on the location and extent of the occlusion and the presence and subsequent development of collateral blood vessels. Principal causes of acute arterial occlusion include embolism, thrombus in situ, arterial dissection, and trauma.

The most common sources of arterial emboli are the heart, aorta, and large arteries. Cardiac disorders that cause thromboembolism include atrial fibrillation; acute myocardial infarction; ventricular aneurysm; cardiomyopathy; infectious and marantic endocarditis; thrombi associated with prosthetic heart valves; and atrial myxoma. Emboli to the distal vessels may also originate from proximal sites of atherosclerosis and aneurysms of the aorta and large vessels. Less frequently, an arterial occlusion results paradoxically from a venous thrombus that has entered the systemic circulation via a patent foramen ovale or another septal defect. Arterial emboli tend to lodge at vessel bifurcations because the vessel caliber decreases at those sites; in the lower extremities, emboli lodge most frequently in the femoral artery, followed by the iliac artery, aorta, and popliteal and tibioperoneal arteries.

Acute arterial thrombosis in situ occurs most frequently in atherosclerotic vessels at the site of an atherosclerotic plaque or aneurysm and in arterial bypass grafts. Trauma to an artery may disrupt continuity of blood flow and cause acute limb ischemia via formation of an acute arterial thrombus or by disruption of an artery’s integrity and extravasation of blood. Arterial occlusion may complicate arterial punctures and placement of catheters; it also may result from arterial dissection if the intimal flap obstructs the artery. Less common causes include thoracic outlet compression syndrome, which causes subclavian artery occlusion, and entrapment of the popliteal artery by abnormal placement of the medial head of the gastrocnemius muscle. Polycythemia and hypercoagulable disorders **(Chaps. 99 and 112)** are also associated with acute arterial thrombosis.

CLINICAL FEATURES

The symptoms of an acute arterial occlusion depend on the location, duration, and severity of the obstruction. Often severe pain, paresthesia, numbness, and coldness develop in the involved extremity within 1 h. Paralysis may occur with severe and persistent ischemia. Physical findings include loss of pulses distal to the occlusion, cyanosis or pallor, mottling, decreased skin temperature, muscle stiffening, loss of sensation, weakness, and/or absent deep tendon reflexes. If acute arterial occlusion occurs in the presence of an adequate collateral circulation, as is often the case in acute graft occlusion, the symptoms and findings may be less severe. In this situation, the patient complains about an abrupt decrease in the distance walked before claudication occurs or of modest pain and paresthesia. Pallor and coolness are evident, but sensory and motor functions generally are preserved. The clinical evaluation includes Doppler assessment of peripheral blood flow. The diagnosis of acute limb ischemia is usually apparent from the clinical presentation. In most circumstances, MRA, CTA, or catheter-based arteriography is used to confirm the diagnosis and demonstrate the location and extent of arterial occlusion.

TREATMENT

TREATMENT Acute Limb Ischemia

Once the diagnosis is made, the patient should be anticoagulated with intravenous heparin to prevent propagation of the clot and recurrent embolism. In cases of severe ischemia of recent onset, particularly when limb viability is jeopardized, immediate intervention to ensure reperfusion is indicated. Catheter-directed thrombolysis/thrombectomy, surgical thromboembolectomy, and arterial bypass procedures are used to restore blood flow to the ischemic extremity promptly, particularly when a large proximal vessel is occluded.

Intraarterial thrombolytic therapy with recombinant tissue plasminogen activator, reteplase, or tenecteplase is most effective when acute arterial occlusion is recent (<2 weeks) and caused by a thrombus in an atherosclerotic vessel, arterial bypass graft, or occluded stent. Thrombolytic therapy is also indicated when the patient's overall condition contraindicates surgical intervention or when smaller distal vessels are occluded, thus preventing surgical access. Meticulous observation for hemorrhagic complications is required during intraarterial thrombolytic therapy. Another endovascular approach to thrombus removal is percutaneous mechanical thrombectomy using devices that employ hydrodynamic forces or rotating baskets to fragment and remove the clot. These treatments may be used alone but usually are used in conjunction with pharmacologic thrombolysis. Surgical revascularization is preferred when restoration of blood flow must occur within 24 h to prevent limb loss or when symptoms of occlusion have been present for >2 weeks. Amputation is performed when the limb is not viable, as characterized by loss of sensation, paralysis, and the absence of Doppler-detected blood flow in both arteries and veins.

Long-term anticoagulation is indicated when acute limb ischemia is caused by cardiac thromboembolism. Emboli resulting from infective endocarditis, the presence of prosthetic heart valves, or atrial myxoma often require surgical intervention to remove the cause.

ATHEROEMBOLISM

Atheroembolism is another cause of limb ischemia. In this condition, multiple small deposits of fibrin, platelets, and cholesterol debris embolize from proximal atherosclerotic lesions or aneurysmal sites. Large protruding aortic atheromas are a source of emboli that may lead to limb ischemia, as well as stroke and renal insufficiency. Atheroembolism may occur after intraarterial procedures. Since atheroemboli to limbs tend to lodge in the small vessels of the muscle and skin and may not occlude the large vessels, distal pulses usually remain palpable. Patients complain of acute pain and tenderness at the site of embolization. Digital vascular occlusion may result in ischemia and the "blue toe" syndrome; digital necrosis and gangrene may develop **(Fig. 275-2)**. Localized areas of tenderness, pallor, and livedo reticularis (see below) occur at sites of emboli. Skin or muscle biopsy may demonstrate cholesterol crystals.

FIGURE 275-2

Atheroembolism causing cyanotic discoloration and impending necrosis of the toes ("blue toe" syndrome).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ischemia resulting from atheroemboli is notoriously difficult to treat. Local foot care and occasionally amputation may be needed to treat necrotic areas. Analgesics are indicated for pain relief. Usually neither surgical revascularization procedures nor thrombolytic therapy is helpful because of the multiplicity, composition, and distal location of the emboli. Therapy with antiplatelet drugs and statins improves cardiovascular outcome in patients with atherosclerosis, but it is not established whether either class of drugs prevents recurrent atheroembolism. Similarly, it is not known whether anticoagulant therapy is effective. Surgical intervention to remove or bypass the atherosclerotic vessel or aneurysm that causes the recurrent atheroemboli may be necessary.

THORACIC OUTLET COMPRESSION SYNDROME

This is a symptom complex resulting from compression of the neurovascular bundle (artery, vein, or nerves) at the thoracic outlet as it courses through the neck and shoulder. Cervical ribs, abnormalities of the scalenus anticus muscle, proximity of the clavicle to the first rib, or abnormal insertion of the pectoralis minor muscle may compress the subclavian artery, subclavian vein, and brachial plexus as these structures pass from the thorax to the arm. Depending on the structures affected, thoracic outlet compression syndrome is divided into arterial, venous, and neurogenic forms. Patients with neurogenic thoracic outlet compression may develop shoulder and arm pain, weakness, and paresthesias. Patients with arterial compression may experience claudication, Raynaud's phenomenon, and even ischemic tissue loss and gangrene. Venous compression may cause thrombosis of the subclavian and axillary veins; this is often associated with effort and is referred to as *Paget-Schroetter syndrome.*

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Thoracic Outlet Compression Syndrome

Examination of a patient with arterial thoracic outlet compression syndrome is often normal unless provocative maneuvers are performed. Occasionally, distal pulses are decreased or absent and digital cyanosis and ischemia may be evident.

Several maneuvers that support the diagnosis of arterial thoracic outlet compression syndrome may be used to precipitate symptoms, cause a subclavian artery bruit, and diminish arm pulses. These maneuvers include the abduction and external rotation test, in which the affected arm is abducted by 90° and the shoulder is externally rotated; the scalene maneuver (extension of the neck and rotation of the head to the side of the symptoms); the costoclavicular maneuver (posterior rotation of shoulders); and the hyperabduction maneuver (raising the arm 180°). A chest x-ray will indicate the presence of cervical ribs. Duplex ultrasonography, MRA, and contrast angiography can be performed during provocative maneuvers to demonstrate thoracic outlet compression of the subclavian artery. Neurophysiologic tests such as the electromyogram, nerve conduction studies, and somatosensory evoked potentials may be abnormal if the brachial plexus is involved, but the diagnosis of neurogenic thoracic outlet syndrome is not necessarily excluded if these tests are normal owing to their low sensitivity.

Most patients can be managed conservatively. They should be advised to avoid the positions that cause symptoms. Many patients benefit from shoulder girdle exercises. Surgical procedures such as removal of the first rib and resection of the scalenus anticus muscle are necessary occasionally for relief of symptoms or treatment of ischemia.

POPLITEAL ARTERY ENTRAPMENT

Popliteal artery entrapment typically affects young athletic men and women when the gastrocnemius or popliteus muscle compresses the popliteal artery and causes intermittent claudication. Thrombosis, embolism, or popliteal artery aneurysm may occur. The pulse examination may be normal unless provocative maneuvers such as ankle dorsiflexion and plantar flexion are performed. The diagnosis is confirmed by duplex ultrasound, CTA, MRA, or conventional angiography. Treatment involves surgical release of the popliteal artery or vascular reconstruction.

POPLITEAL ARTERY ANEURYSM

Popliteal artery aneurysms are the most common peripheral artery aneurysms. Approximately 50% are bilateral. Patients with popliteal artery aneurysms often have aneurysms of other arteries, especially the aorta. The most common clinical presentation is limb ischemia secondary to thrombosis or embolism. Rupture occurs less frequently. Other complications include compression of the adjacent popliteal vein or peroneal nerve. Popliteal artery aneurysm can be detected by palpation and confirmed by duplex ultrasonography. Repair is indicated for symptomatic aneurysms or when the diameter exceeds 2–3 cm, owing to the risk of thrombosis, embolism, or rupture.

ARTERIOVENOUS FISTULA

Abnormal communications between an artery and a vein, bypassing the capillary bed, may be congenital or acquired. Congenital arteriovenous fistulas are a result of persistent embryonic vessels that fail to differentiate into arteries and veins; they may be associated with birthmarks, can be located in almost any organ of the body, and frequently occur in the extremities. Acquired arteriovenous fistulas either are created to provide vascular access for hemodialysis or occur as a result of a penetrating injury such as a gunshot or knife wound or as complications of arterial catheterization or surgical dissection. An uncommon cause of arteriovenous fistula is rupture of an arterial aneurysm into a vein.

The clinical features depend on the location and size of the fistula. Frequently, a pulsatile mass is palpable, and a thrill and a bruit lasting throughout systole and diastole are present over the fistula. With long-standing fistulas, clinical manifestations of chronic venous insufficiency, including peripheral edema; large, tortuous varicose veins; and stasis pigmentation become apparent because of the high venous pressure. Evidence of ischemia may occur in the distal portion of the extremity. Skin temperature is higher over the arteriovenous fistula. Large arteriovenous fistulas may result in an increased cardiac output with consequent cardiomegaly and high-output heart failure **(Chap. 252)**.

The diagnosis is often evident from the physical examination. Compression of a large arteriovenous fistula may cause reflex slowing of the heart rate (Nicoladoni-Branham sign). Duplex ultrasonography may detect an arteriovenous fistula, especially one that affects the femoral artery and vein at the site of catheter access. CTA and conventional angiography can confirm the diagnosis and are useful in demonstrating the site and size of the arteriovenous fistula.

Management of arteriovenous fistulas may involve surgery, radiotherapy, or embolization. Congenital arteriovenous fistulas are often difficult to treat because the communications may be numerous and extensive, and new communications frequently develop after ligation of the most obvious ones. Many of these lesions are best treated conservatively using elastic support hose to reduce the consequences of venous hypertension. Occasionally, embolization with autologous material, such as fat or muscle, or with hemostatic agents, such as gelatin sponges or silicon spheres, is used to obliterate the fistula. Acquired arteriovenous fistulas are usually amenable to surgical treatment that involves division or excision of the fistula. Occasionally, autogenous or synthetic grafting is necessary to reestablish continuity of the artery and vein.

RAYNAUD'S PHENOMENON

Raynaud's phenomenon is characterized by episodic digital ischemia, manifested clinically by the sequential development of digital blanching, cyanosis, and rubor of the fingers or toes after cold exposure and subsequent rewarming. Emotional stress may also precipitate Raynaud's phenomenon. The color changes are usually well demarcated and are confined to the fingers or toes. Typically, one or more digits will appear white when the patient is exposed to a cold environment or touches a cold object **(Fig. 275-3A)**. The blanching, or pallor, represents the ischemic phase of the phenomenon and results from vasospasm of digital arteries. During the ischemic phase, capillaries and venules dilate, and cyanosis results from the deoxygenated blood that is present in these vessels. A sensation of cold or numbness or paresthesia of the digits often accompanies the phases of pallor and cyanosis.

Figure 275-3

Vascular diseases associated with temperature: ***A.*** Raynaud's phenomenon; ***B.*** acrocyanosis; ***C.*** livedo reticularis; ***D.*** pernio; ***E.*** erythromelalgia; and ***F.*** frostbite.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

With rewarming, the digital vasospasm resolves, and blood flow into the dilated arterioles and capillaries increases dramatically. This "reactive hyperemia" imparts a bright red color to the digits. In addition to rubor and warmth, patients often experience a throbbing, painful sensation during the hyperemic phase. Although the triphasic color response is typical of Raynaud's phenomenon, some patients may develop only pallor and cyanosis; others may experience only cyanosis.

Raynaud's phenomenon is broadly separated into two categories: idiopathic, termed primary Raynaud's phenomenon, and secondary Raynaud's phenomenon, which is associated with other disease states or known causes of vasospasm **(Table 275-1)**.

TABLE 275-1 Classification of Raynaud's Phenomenon

Primary or idiopathic Raynaud's phenomenon

Secondary Raynaud's phenomenon

Collagen vascular diseases: scleroderma, systemic lupus erythematosus, rheumatoid arthritis, dermatomyositis, polymyositis, mixed connective tissue disease, Sjögren's syndrome

Arterial occlusive diseases: atherosclerosis of the extremities, thromboangiitis obliterans, acute arterial occlusion, thoracic outlet syndrome

Pulmonary hypertension

Neurologic disorders: intervertebral disk disease, syringomyelia, spinal cord tumors, stroke, poliomyelitis, carpal tunnel syndrome, complex regional pain syndrome

Blood dyscrasias: cold agglutinins, cryoglobulinemia, cryofibrinogenemia, myeloproliferative disorders, lymphoplasmacytic lymphoma

Trauma: vibration injury, hammer hand syndrome, electric shock, cold injury, typing, piano playing

Drugs and toxins: ergot derivatives, methysergide, β-adrenergic receptor blockers, bleomycin, vinblastine, cisplatin, gemcitabine, vinyl chloride

Primary Raynaud's Phenomenon

This appellation is applied when the secondary causes of Raynaud's phenomenon have been excluded. Over 50% of patients with Raynaud's phenomenon have the primary form. Women are affected about five times more often than men, and the age of presentation is usually between 20 and 40 years. The fingers are involved more frequently than the toes. Initial episodes may involve only one or two fingertips, but subsequent attacks may involve the entire finger and may include all the fingers. The toes are affected in 40% of patients. Although vasospasm of the toes usually occurs in patients with symptoms in the fingers, it may happen alone. Rarely, the earlobes, the tip of the nose, tongue, nipple, or penis are involved. Raynaud's phenomenon occurs frequently in patients who also have migraine headaches or variant angina. These associations suggest that there may be a common predisposing cause for the vasospasm.

Results of physical examination are often entirely normal; the radial, ulnar, and pedal pulses are normal. The fingers and toes may be cool between attacks and may perspire excessively. Nailfold capillaroscopy reveals normal superficial capillaries, which appear as regularly spaced hairpin loops. Thickening and tightening of the digital subcutaneous tissue (*sclerodactyly*) develop in 10% of patients. Angiography of the digits for diagnostic purposes is not indicated.

In general, patients with primary Raynaud's disease have milder clinical manifestations. Fewer than 1% of these patients lose a part of a digit. After the diagnosis is made, the disease improves spontaneously in ~15% of patients and progresses in about 30%.

Secondary Causes of Raynaud's Phenomenon

Raynaud's phenomenon occurs in 80–90% of patients with systemic sclerosis (scleroderma) and is the presenting symptom in 30% **(Chap. 353)**. It may be the only symptom of scleroderma for many years. Abnormalities of the digital vessels may contribute to the development of Raynaud's phenomenon in this disorder. Ischemic fingertip ulcers may develop and progress to gangrene and autoamputation. About 20% of patients with systemic lupus erythematosus (SLE) have Raynaud's phenomenon **(Chap. 349)**. Occasionally, persistent digital ischemia develops and may result in ulcers or gangrene. In most severe cases, the small vessels are occluded by a proliferative endarteritis. Raynaud's phenomenon occurs in about 30% of patients with dermatomyositis or polymyositis **(Chap. 358)**. It frequently develops in patients with rheumatoid arthritis and may be related to the intimal proliferation that occurs in the digital arteries.

Atherosclerosis of the extremities is a common cause of Raynaud's phenomenon in men aged >50 years. Thromboangiitis obliterans is an uncommon cause of Raynaud's phenomenon but should be considered in young men, particularly those who are cigarette smokers. The development of cold-induced pallor in these disorders may be confined to one or two digits of the involved extremity. Occasionally, Raynaud's phenomenon may follow acute occlusion of large and medium-sized arteries by a thrombus or embolus. Embolization of atheroembolic debris may cause digital ischemia. The latter situation often involves one or two digits and should not be confused with Raynaud's phenomenon. In patients with thoracic outlet compression syndrome, Raynaud's phenomenon may result from diminished intravascular pressure, stimulation of sympathetic fibers in the brachial plexus, or a combination of both. Raynaud's phenomenon occurs in patients with primary pulmonary hypertension **(Chap. 277)**; this is more than coincidental and may reflect a neurohumoral abnormality that affects both the pulmonary and digital circulations.

A variety of blood dyscrasias may be associated with Raynaud's phenomenon. Cold-induced precipitation of plasma proteins, hyperviscosity, and aggregation of red cells and platelets may occur in patients with cold agglutinins, cryoglobulinemia, or cryofibrinogenemia. Hyperviscosity syndromes that accompany myeloproliferative disorders and

lymphoplasmacytic lymphoma (Waldenström's macroglobulinemia) should also be considered in the initial evaluation of patients with Raynaud's phenomenon.

Raynaud's phenomenon occurs often in patients whose vocations require the use of vibrating hand tools, such as chain saws or jackhammers. The frequency of Raynaud's phenomenon also seems to be increased in pianists and keyboard operators. Electric shock injury to the hands or frostbite may lead to the later development of Raynaud's phenomenon.

Several drugs have been causally implicated in Raynaud's phenomenon. They include ergot preparations, methysergide, β-adrenergic receptor antagonists, and the chemotherapeutic agents bleomycin, vinblastine, cisplatin, and gemcitabine.

TREATMENT

TREATMENT Raynaud's Phenomenon

Most patients with Raynaud's phenomenon experience only mild and infrequent episodes. These patients need reassurance and should be instructed to dress warmly and avoid unnecessary cold exposure. In addition to gloves and mittens, patients should protect the trunk, head, and feet with warm clothing to prevent cold-induced reflex vasoconstriction. Tobacco use is contraindicated.

Drug treatment should be reserved for severe cases. Dihydropyridine calcium channel antagonists such as nifedipine, isradipine, felodipine, and amlodipine decrease the frequency and severity of Raynaud's phenomenon. Diltiazem may be considered but is less effective. The postsynaptic α_1-adrenergic antagonist prazosin has been used with favorable responses; doxazosin and terazosin may also be effective. Phosphodiesterase type 5 inhibitors such as sildenafil, tadalafil, and vardenafil may improve symptoms in patients with secondary Raynaud's phenomenon, as occurs with systemic sclerosis. There is limited evidence that topical nitroglycerin preparations are effective. Digital sympathectomy is helpful in some patients who are unresponsive to medical therapy. Injection of botulinum toxin into the perivascular tissue of the wrist or palm improved ischemic manifestations of severe Raynaud's phenomenon in case series, but controlled clinical trials are lacking.

ACROCYANOSIS

In this condition, there is arterial vasoconstriction and secondary dilation of the capillaries and venules with resulting persistent cyanosis of the hands and, less frequently, the feet. Cyanosis may be intensified by exposure to a cold environment. Acrocyanosis may be categorized as primary or secondary to an underlying condition. In primary acrocyanosis, women are affected much more frequently than men, and the age of onset is usually <30 years. Generally, patients are asymptomatic but seek medical attention because of the discoloration. The prognosis is favorable, and pain, ulcers, and gangrene do not occur. Examination reveals normal pulses, peripheral cyanosis, and moist palms **(Fig. 275-3B)**. Trophic skin changes and ulcerations do *not* occur. The disorder can be distinguished from Raynaud's phenomenon because it is persistent and not episodic, the discoloration extends proximally from the digits, and blanching does not occur. Ischemia secondary to arterial occlusive disease can usually be excluded by the presence of normal pulses. Central cyanosis and decreased arterial oxygen saturation are not present. Patients should be reassured and advised to dress warmly and avoid cold exposure. Pharmacologic intervention is not indicated.

Secondary acrocyanosis may result from hypoxemia, vasopressor medications, connective tissue diseases, atheroembolism, antiphospholipid antibodies, cold agglutinins, or cryoglobulins and is associated with anorexia nervosa and postural orthostatic tachycardia syndrome. Treatment should be directed at the underlying disorder.

LIVEDO RETICULARIS

In this condition, localized areas of the extremities develop a mottled or rete (netlike) appearance of reddish to blue discoloration **(Fig. 275-3C)**. The mottled appearance may be more prominent after cold exposure. There are primary and secondary forms of livedo reticularis. The primary, or idiopathic, form of this disorder may be benign or associated with ulcerations. The benign form occurs more frequently in women than in men, and the most common age of onset is the third decade. Patients with the benign form are usually asymptomatic and seek attention for cosmetic reasons. These patients should be reassured and advised to avoid cold environments. No drug treatment is indicated. Primary livedo reticularis with ulceration is also called *atrophie blanche en plaque*. The ulcers are painful and may take months to heal. Secondary livedo reticularis can occur with atheroembolism (see above), SLE and other vasculitides, antiphospholipid antibodies, hyperviscosity, cryoglobulinemia, and Sneddon's syndrome (ischemic stroke and livedo reticularis). Rarely, skin ulcerations develop.

PERNIO (CHILBLAINS)

Pernio is a vasculitic disorder associated with exposure to cold; acute forms have been described. Raised erythematous lesions develop on the lower part of the legs and feet in cold weather **(Fig. 275-3D)**. They are associated with pruritus and a burning sensation, and they may blister and ulcerate. Pathologic examination demonstrates angiitis characterized by intimal proliferation and perivascular infiltration of mononuclear and polymorphonuclear leukocytes. Giant cells may be present in the subcutaneous tissue. Patients should avoid exposure to cold, and ulcers should be kept clean and protected with sterile dressings. Sympatholytic drugs and dihydropyridine calcium channel antagonists may be effective in some patients.

ERYTHROMELALGIA

This disorder is characterized by burning pain and erythema of the extremities **(Fig. 275-3E)**. The feet are involved more frequently than the hands, and males are affected more frequently than females. Erythromelalgia may occur at any age but is most common in middle age. It may be primary (also termed erythermalgia) or secondary. Mutations in the *SCN9A* gene, which encodes the Nav1.7 voltage-gated sodium channel expressed in sensory and sympathetic nerves, has been described in inherited forms of erythromelalgia. The most common causes of secondary erythromelalgia are myeloproliferative disorders such as polycythemia vera and essential thrombocytosis. Less common causes include drugs, such as calcium channel blockers, bromocriptine, and pergolide; neuropathies; connective tissue diseases such as SLE; and paraneoplastic syndromes. Patients complain of burning in the extremities that is precipitated by exposure to a warm environment and aggravated by a dependent position. The symptoms are relieved by exposing the affected area to cool air or water or by elevation. Erythromelalgia can be distinguished from ischemia secondary to peripheral arterial disorders because the peripheral pulses are present. There is no specific treatment; aspirin may produce relief in patients with erythromelalgia secondary to myeloproliferative disease. Treatment of associated disorders in secondary erythromelalgia may be helpful.

FROSTBITE

In this condition, tissue damage results from severe environmental cold exposure or from direct contact with a very cold object. Tissue injury results from both freezing and vasoconstriction. Frostbite usually affects the distal aspects of the extremities or exposed parts of the face, such as the ears, nose, chin, and cheeks. Superficial frostbite involves the skin and subcutaneous tissue. Patients experience pain or paresthesia, and the skin appears white and waxy. After rewarming, there is cyanosis and erythema, wheal-and-flare formation, edema, and superficial blisters. Deep frostbite involves muscle, nerves, and deeper blood vessels. It may result in edema of the hand or foot, vesicles and bullae, tissue necrosis, and gangrene **(Fig. 275-3F)**.

Initial treatment is rewarming, performed in an environment where reexposure to freezing conditions will not occur. Rewarming is accomplished by immersion of the affected part in a water bath at temperatures of 40°–44°C (104°–111°F). Massage, application of ice water, and extreme heat are contraindicated. The injured area should be cleansed with soap or antiseptic, and sterile dressings should be applied. Analgesics are often required during rewarming. Antibiotics are used if there is evidence of infection. The efficacy of sympathetic blocking drugs is not established. After recovery, the affected extremity may exhibit increased sensitivity to cold.

FURTHER READING

Bonaca MP, Creager MA: Pharmacological treatment and current management of peripheral artery disease. Circ Res 116:1579, 2015. [PubMed: 25908730]
Creager MA et al: Acute limb ischemia. N Engl J Med 366:2198, 2012. [PubMed: 22670905]
Gerhard-Herman MD et al: 2016 AHA/ACC guideline on the management of patients with lower extremity peripheral artery disease: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation 135:e726, 2017. [PubMed: 27840333]
Kullo IJ, Rooke TW: Peripheral artery disease. N Engl J Med 374:861, 2016. [PubMed: 26962905]
Olin JW et al: Fibromuscular dysplasia: State of the science and critical unanswered questions: A Scientific Statement from the American Heart Association. Circulation 129:1048, 2014. [PubMed: 24548843]
Thukkani AK, Kinlay S: Endovascular intervention for peripheral artery disease. Circ Res 116:1599, 2015. [PubMed: 25908731]
Vartanian SM, Conte MS: Surgical intervention for peripheral arterial disease. Circ Res 116:1614, 2015. [PubMed: 25908732]
Wigley FM, Flavahan NA: Raynaud's Phenomenon. N Engl J Med 375:556, 2016. [PubMed: 27509103]

Chapter 276: Chronic Venous Disease and Lymphedema

Mark A. Creager; Joseph Loscalzo

CHRONIC VENOUS DISEASE

Chronic venous diseases range from telangiectasias and reticular veins, to varicose veins, to chronic venous insufficiency with edema, skin changes, and ulceration. This section of the chapter will focus on identification and treatment of varicose veins and chronic venous insufficiency, since these problems are encountered frequently by the internist. The estimated prevalence of varicose veins in the United States is ~15% in men and 30% in women. Chronic venous insufficiency with edema affects ~7.5% of men and 5% of women, and the prevalence increases with age ranging from 2% among those <50 years of age to 10% of those 70 years of age. Approximately 20% of patients with chronic venous insufficiency develop venous ulcers.

VENOUS ANATOMY

Veins in the extremities can be broadly classified as either superficial or deep. The superficial veins are located between the skin and deep fascia. In the legs, these include the great and small saphenous veins and their tributaries. The great saphenous vein is the longest vein in the body. It originates on the medial side of the foot and ascends anterior to the medial malleolus and then along the medial side of the calf and thigh, and drains into the common femoral vein. The small saphenous vein originates on the dorsolateral aspect of the foot, ascends posterior to the lateral malleolus and along the posterolateral aspect of the calf, and drains into the popliteal vein. The deep veins of the leg accompany the major arteries. There are usually paired peroneal, anterior tibial, and posterior tibial veins in the calf, which converge to form the popliteal vein. Soleal tributary veins drain into the posterior tibial or peroneal veins, and gastrocnemius tributary veins drain into the popliteal vein. The popliteal vein ascends in the thigh as the femoral vein. The confluence of the femoral vein and deep femoral vein form the common femoral vein, which ascends in the pelvis as the external iliac and then common iliac vein, which converges with the contralateral common iliac vein at the inferior vena cava. Perforating veins connect the superficial and deep systems in the legs at multiple locations, normally allowing blood to flow from the superficial to deep veins. In the arms, the superficial veins include the basilic, cephalic, and median cubital veins and their tributaries. The basilic and cephalic veins course along the medial and lateral aspects of the arm, respectively, and these are connected via the median cubital vein in the antecubital fossa. The deep veins of the arms accompany the major arteries and include the radial, ulnar, brachial, axillary, and subclavian veins. The subclavian vein converges with the internal jugular vein to form the brachiocephalic vein, which joins the contralateral brachiocephalic vein to form the superior vena cava. Bicuspid valves are present throughout the venous system to direct the flow of venous blood centrally.

Pathophysiology of Chronic Venous Disease

Varicose veins are dilated, bulging, tortuous superficial veins, measuring at least 3 mm in diameter. The smaller and less tortuous reticular veins are dilated intradermal veins, which appear blue-green, measure 1–3 mm in diameter, and do not protrude from the skin surface. Telangiectasias, or spider veins, are small, dilated veins, <1 mm in diameter, located near the skin surface, and form blue, purple, or red linear, branching, or spider-web patterns.

Varicose veins can be categorized as primary or secondary. Primary varicose veins originate in the superficial system and result from defective structure and function of the valves of the saphenous veins, intrinsic weakness of the vein wall, and high intraluminal pressure. Approximately one-half of these patients have a family history of varicose veins. Other factors associated with primary varicose veins include aging, pregnancy, hormonal therapy, obesity, and prolonged standing. Secondary varicose veins result from venous hypertension, associated with deep-venous insufficiency or deep-venous obstruction, and incompetent perforating veins that cause enlargement of superficial veins. Arteriovenous fistulas also cause varicose veins in the affected limb.

Chronic venous insufficiency is a consequence of incompetent veins in which there is venous hypertension and extravasation of fluid and blood elements into the tissue of the limb. It may occur in patients with varicose veins but usually is caused by disease in the deep veins. It also is categorized as primary or secondary. Primary deep-venous insufficiency is a consequence of an intrinsic structural or functional abnormality in the vein wall or venous valves leading to valvular reflux. Secondary deep-venous insufficiency is caused by obstruction and/or valvular incompetence from previous deep-vein thrombosis **(Chap. 273)**. Deep-venous insufficiency occurs following deep-vein thrombosis, as the delicate valve leaflets become thickened and contracted and can no longer prevent retrograde flow of blood and the vein itself becomes rigid and thick walled. Although most veins recanalize after an episode of thrombosis, the large proximal veins may remain occluded. Secondary incompetence develops in distal valves because high pressures distend the vein and separate the leaflets. Other causes of secondary deep-venous insufficiency include May-Thurner syndrome, where the left iliac vein is occluded or stenosed by extrinsic compression from the overlapping right common iliac artery; arteriovenous fistulas resulting in increased venous pressure; congenital deep-vein agenesis or hypoplasia; and venous malformations as may occur in Klippel-Trénaunay and Parkes-Weber syndromes.

Clinical Presentation

Patients with venous varicosities are often asymptomatic but still concerned about the cosmetic appearance of their legs. Superficial venous thrombosis may be a recurring problem, and, rarely, a varicosity ruptures and bleeds. Symptoms in patients with varicose veins or venous insufficiency, when they occur, include a dull ache, throbbing or heaviness, or pressure sensation in the legs typically after prolonged standing; these symptoms usually are relieved with leg elevation. Additional symptoms may include cramping, burning, pruritus, leg swelling, and skin ulceration.

The legs are examined in both the supine and standing positions. Visual inspection and palpation of the legs in the standing position confirm the presence of varicose veins. The location and extent of the varicose veins should be noted. Edema, stasis dermatitis, and skin ulceration near the ankle may be present if there is superficial venous insufficiency and venous hypertension. Findings of deep-venous insufficiency include increased leg circumference, venous varicosities, edema, and skin changes. The edema, which is usually pitting, may be confined to the ankles, extend above the ankles to the knees, or involve the thighs in severe cases. Over time, the edema may become less pitting and more indurated. Dermatologic findings associated with venous stasis include hyperpigmentation, erythema, eczema, lipodermatosclerosis, *atrophie blanche*, and a phlebectasia corona. Lipodermatosclerosis is the combination of induration, hemosiderin deposition, and inflammation, and typically occurs in the lower part of the leg just above the ankle. Atrophie blanche is a white patch of scar tissue, often with focal telangiectasias and a hyperpigmented border; it usually develops near the medial malleolus. A phlebectasia corona is a fan-shaped pattern of intradermal veins near the ankle or on the foot. Skin

ulceration may occur near the medial and lateral malleoli. A venous ulcer is often shallow and characterized by an irregular border, a base of granulation tissue, and the presence of exudate **(Fig. 276-1)**.

FIGURE 276-1

Venous insufficiency with active venous ulcer near the medial malleolus. *(Courtesy of Dr. Steven Dean, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Bedside maneuvers can be used to distinguish primary varicose veins from secondary varicose veins caused by deep-venous insufficiency. With the contemporary use of venous ultrasound (see below), however, these maneuvers are employed infrequently. The Brodie–Trendelenburg test is used to determine whether varicose veins are secondary to deep-venous insufficiency. As the patient is lying supine, the leg is elevated and the veins allowed to empty. Then, a tourniquet is placed on the proximal part of the thigh and the patient is asked to stand. Filling of the varicose veins within 30 s indicates that the varicose veins are caused by deep-venous insufficiency and incompetent perforating veins. Primary varicose veins with superficial venous insufficiency are the likely diagnosis if venous refilling occurs promptly after tourniquet removal. The Perthes test assesses the possibility of deep-venous obstruction. A tourniquet is placed on the midthigh after the patient has stood, and the varicose veins are filled. The patient is then instructed to walk for 5 min. A patent deep-venous system and competent perforating veins enable the superficial veins below the tourniquet to collapse. Deep-venous obstruction is likely to be present if the superficial veins distend further with walking.

Differential Diagnosis

The duration of leg edema helps to distinguish chronic venous insufficiency from acute deep-vein thrombosis. Lymphedema, as discussed later in this chapter, is often confused with chronic venous insufficiency, and both may occur together. Other disorders that cause leg swelling should be considered and excluded when evaluating a patient with presumed venous insufficiency. Bilateral leg swelling occurs in patients with congestive heart failure, hypoalbuminemia secondary to nephrotic syndrome or severe hepatic disease, myxedema caused by hypothyroidism or pretibial myxedema associated with Graves' disease, and with drugs such as dihydropyridine calcium channel blockers and thiazolidinediones. Unilateral causes of leg swelling also include ruptured leg muscles, hematomas secondary to trauma, and popliteal cysts. Cellulitis may cause erythema and swelling of the affected limb. Leg ulcers may be caused by severe peripheral artery disease and critical limb ischemia; neuropathies, particularly those associated with diabetes; and less commonly, skin cancer, vasculitis, or rarely as a complication of hydroxyurea. The location and characteristics of venous ulcers help to differentiate these from other causes.

Classification of Chronic Venous Disease

The CEAP (clinical, etiologic, anatomic, pathophysiologic) classification schema incorporates the range of symptoms and signs of chronic venous disease to characterize its severity. It also broadly categorizes the etiology as congenital, primary, or secondary; identifies the affected veins as superficial, deep, or perforating; and characterizes the pathophysiology as reflux, obstruction, both, or neither **(Table 276-1)**.

TABLE 276-1 CEAP (Clinical, Etiologic, Anatomic, Pathophysiologic) Classification

Clinical Classification

C0 No visible or palpable signs of venous disease

C1 Telangiectasias, reticular veins

C2 Varicose veins

C3 Edema without skin changes

C4 Skin changes, including pigmentation, eczema, lipodermatosclerosis, and atrophie blanche

C5 Healed venous ulcer

C6 Active venous ulcer

Etiologic Classification

Ec Congenital

Ep Primary

Es Secondary (postthrombotic)

En No venous etiology identified

Anatomic Classification

As Superficial veins

Ap Perforator veins

Ad Deep veins

An No venous location identified

Pathophysiologic Classification

Pr Reflux

Po Obstruction

Pr,o Reflux and obstruction

Pn No venous pathophysiology identifiable

Source: Data from B Eklöf et al: J Vasc Surg 40:1248, 2004.

Diagnostic Testing

The principal diagnostic test to evaluate patients with chronic venous disease is venous duplex ultrasonography. A venous duplex ultrasound examination uses a combination of B-mode imaging and spectral Doppler to detect the presence of venous obstruction and venous reflux in superficial and deep veins. Color-assisted Doppler ultrasound is useful to visualize venous flow patterns. Obstruction may be diagnosed by the absence of flow, the presence of an echogenic thrombus within the vein, or failure of the vein to collapse when a compression maneuver is applied by the sonographer, the last implicating the presence of an intraluminal thrombus. Venous reflux is detected by prolonged reversal of venous flow direction during a Valsalva maneuver, particularly for the common femoral vein or saphenofemoral junction, or after compression and release of a cuff placed on the limb distal to the area being interrogated.

Some vascular laboratories use air or strange gauge plethysmography to assess the severity of venous reflux and complement findings from the venous ultrasound examination. Venous volume and venous refilling time are measured when the legs are placed in a dependent position and after calf exercise to quantify the severity of venous reflux and the efficiency of the calf muscle pump to affect venous return.

Magnetic resonance, computed tomographic, and conventional venography are rarely required to determine the cause and plan treatment for chronic venous insufficiency unless there is suspicion for pathology that might warrant intervention. These modalities are used to identify obstruction or stenosis

of the inferior vena cava and iliofemoral veins, as may occur in patients with previous proximal deep-vein thrombosis; occlusion of inferior vena cava filters; extrinsic compression from tumors; and May-Thurner syndrome.

TREATMENT

TREATMENT Chronic Venous Disease SUPPORTIVE MEASURES

Varicose veins usually are treated with conservative measures. Symptoms often decrease when the legs are elevated periodically, prolonged standing is avoided, and elastic support hose are worn. External compression with elastic stockings or stretch bandages provides a counterbalance to the hydrostatic pressure in the veins. Although compression garments may improve symptoms, they do not prevent progression of varicose veins. Graduated compression stockings with pressures of 20–30 mmHg are suitable for most patients with simple varicose veins, although pressures of 30–40 mmHg may be required for patients with manifestations of venous insufficiency such as edema and ulcers.

Patients with chronic venous insufficiency also should be advised to avoid prolonged standing or sitting; frequent leg elevation is helpful. Graded compression therapy consisting of stockings or multilayered compression bandages is the standard of care for advanced chronic venous insufficiency characterized by edema, skin changes, or venous ulcers defined as CEAP clinical class C3–C6. Graduated compression stockings of 30–40 mmHg are more effective than lesser grades for healing venous ulcers. The length of stocking depends on the distribution of edema. Calf-length stockings are tolerated better by most patients, particularly elderly patients; for patients with varicose veins or edema extending to the thigh, thigh-length stockings or panty hose should be considered. Exercise training, including leg muscle strengthening, may improve calf muscle pump function and antegrade venous flow, and reduce the severity of chronic venous insufficiency. Overweight and obese patients should be advised to lose weight via caloric restriction and exercise.

In addition to a compression bandage or stocking, patients with venous ulcers also may be treated with low adherent absorbent dressings that take up exudates while maintaining a moist environment. Other types of dressings include hydrocolloid (an adhesive dressing composed of polymers such as carboxymethylcellulose that absorbs exudates by forming a gel), hydrogel (a nonabsorbent dressing comprising >80% water or glycerin that moisturizes wounds), foam (an absorbent dressing made with polymers such as polyurethane), and alginate (an absorbent, biodegradable dressing that is derived from seaweed), but there is little evidence that these are more effective than low-adherent absorbent dressings. The choice of specific dressing depends on the amount of drainage, presence of infection, and integrity of the skin surrounding the ulcer. Ulcers should be debrided of necrotic tissue. Antibiotics are not indicated unless the ulcer is infected. The multilayered compression bandage or graduated compression garment is then put over the dressing.

MEDICAL THERAPIES

There are no drugs approved by the U.S. Food and Drug Administration for the treatment of chronic venous insufficiency. Diuretics may reduce edema, but at the risk of volume depletion and compromise in renal function. Topical steroids may be used for a short period of time to treat inflammation associated with stasis dermatitis. Several herbal supplements, such as horse chestnut seed extract (aescin); flavonoids including diosmin, hesperidin, or the two combined as micronized purified flavonoid fraction; and French maritime pine bark extract, are touted to have venoconstrictive and anti-inflammatory properties. Although meta-analyses have suggested that aescin reduces edema, pruritus, and pain and that micronized purified flavonoid fraction in conjunction with compression therapy facilitates venous ulcer healing, there is insufficient evidence to recommend the general use of these substances in patients with chronic venous insufficiency.

INTERVENTIONAL AND SURGICAL THERAPIES

Ablative procedures, including endovenous thermal ablation, sclerotherapy, and surgery, are used to treat varicose veins in selected patients who have persistent symptoms, great saphenous vein incompetency, and complications of venous insufficiency including dermatitis, edema, and ulcers. Ablative therapy may also be indicated for cosmetic reasons.

Endovenous thermal ablation procedures of the saphenous veins include endovenous laser therapy and radiofrequency ablation. To ablate the great saphenous vein, a catheter is placed percutaneously and advanced from the level of the knee to just below the saphenofemoral junction via ultrasound guidance. Thermal energy is then delivered as the catheter is pulled back. The heat injures the endothelium and media and promotes thrombosis and fibrosis, resulting in venous occlusion. Average 1- and 5-year occlusion rates exceed 90% following endovenous laser therapy and are slightly less after radiofrequency ablation. Deep-vein thrombosis of the common femoral vein adjacent to the saphenofemoral junction is an uncommon but potential complication of endovenous thermal ablation. Other adverse effects of thermal ablation procedures include pain, paresthesias, bruising, hematoma, and hyperpigmentation.

Sclerotherapy involves the injection of a chemical into a vein to cause fibrosis and obstruction. Sclerosing agents approved by the U.S. Food and Drug Administration include sodium tetradecyl sulfate, polidocanol, sodium morrhuate, and glycerin. The sclerosing agent is administered as a liquid or mixed with air or CO_2/O_2 to create a foam. It first is injected into the great saphenous vein or its affected tributaries, often with ultrasound guidance. Thereafter, smaller more distal veins and incompetent perforating veins are injected. Following completion of the procedure, elastic bandages are applied, or 30–40 mmHg compression stockings are worn for 1–2 weeks. Average 1- and 5-year occlusion rates are 81 and 74%, respectively, following sclerotherapy. Complications are uncommon and include deep-vein thrombosis, hematomas, damage to adjacent saphenous or sural nerves, and infection. Anaphylaxis is a very rare but severe complication.

Surgical therapy usually involves ligation and stripping of the great and small saphenous veins. The procedure is performed under general anesthesia. Incisions are made at the groin and the upper calf. The great saphenous vein is ligated below the saphenofemoral junction, and a wire is inserted into the great saphenous vein and advanced distally. The proximal part of the great saphenous vein is secured to the wire and retrieved, i.e., stripped, via the calf incision. Stripping of the great saphenous vein below the knee and stripping of the small saphenous vein usually are not performed because of the respective risks of saphenous and sural nerve injury. Complications of great saphenous vein ligation and stripping include deep-vein thrombosis, bleeding, hematoma, infection, and nerve injury. Recurrent varicose veins occur in up to 50% patients by 5 years, due to technical failures, deep-venous insufficiency, and incompetent perforating veins.

Stab phlebectomy is another surgical treatment for of varicose veins. A small incision is made alongside the varicose vein, and it is avulsed by means of a forceps or hook. This procedure may be performed in conjunction with saphenous vein ligation and stripping or thermal ablation. Subfascial endoscopic perforator surgery (SEPS) uses endoscopy to identify and occlude incompetent perforating veins. It also may be performed along with other ablative procedures.

Endovascular interventions, surgical bypass, and reconstruction of the valves of the deep veins are performed when feasible to treat patients with advanced chronic venous insufficiency who have not responded to other therapies. Catheter-based interventions, usually involving placement of endovenous stents, may be considered to treat some patients with chronic occlusions of the iliac veins. Technical success rates exceed 85% in most series, and long-term patency is achieved in ~75% of these patients. Iliocaval bypass, femoroiliac venous bypass, and femorofemoral crossover venous bypass are procedures used occasionally to treat iliofemoral vein occlusion; saphenopopliteal vein bypass can be used to treat chronic femoropopliteal vein obstruction. Long-term patency rates for venous bypass procedures generally exceed 60% and are associated with improvement in symptoms. Surgical reconstruction of the valves of the deep veins and valve transfer procedures are used to treat

valvular incompetence. Valvuloplasty involves tightening the valve by commissural apposition. With valve transfer procedures, a segment of vein with a competent valve, such as a brachial or axillary vein, or adjacent saphenous or deep femoral vein, is inserted as an interposition graft in the incompetent vein. Both valvuloplasty and vein transfer operations result in ulcer healing in the majority of patients, although success rates are somewhat better with valvuloplasty.

Lymphedema

Lymphedema is a chronic condition caused by impaired transport of lymph and characterized by swelling of one or more limbs and occasionally the trunk and genitalia. Fluid accumulates in interstitial tissues when there is an imbalance between lymph production and lymph absorption, a process governed in large part by Starling forces. Deficiency, reflux, or obstruction of lymph vessels perturbs the ability of the lymphatic system to reabsorb proteins that had been filtered by blood vessels, and the tissue osmotic load promotes interstitial accumulation of water. Persistent lymphedema leads to inflammatory and immune responses characterized by infiltration of mononuclear cells, fibroblasts, and adipocytes, leading to adipose and collagen deposition in the skin and subcutaneous tissues.

Lymphatic Anatomy

Lymphatic capillaries are blind-ended tubes formed by a single layer of endothelial cells. The absent or widely fenestrated basement membrane of lymphatic capillaries allows access to interstitial proteins and particles. Lymphatic capillaries merge to form microlymphatic precollector vessels, which contain few smooth muscle cells. The precollector vessels drain into collecting lymphatic vessels, which comprise endothelial cells, a basement membrane, smooth muscle, and bileaflet valves. The collecting lymphatic vessels in term merge to form larger lymphatic conduits. Analogous to venous anatomy, there are superficial and deep lymphatic vessels in the legs, which communicate at the popliteal and inguinal lymph nodes. Pelvic lymphatic vessels drain into the thoracic duct, which ascends from the abdomen to the thorax and connects with the left brachiocephalic vein. Lymph is propelled centrally by the phasic contractile activity of lymphatic smooth muscle and facilitated by the contractions of contiguous skeletal muscle. The presence of lymphatic valves ensures unidirectional flow.

Etiology

Lymphedema may be categorized as primary or secondary **(Table 276-2)**. The prevalence of primary lymphedema is ~1.15 per 100,000 persons <20 years of age. Females are affected more frequently than males. Primary lymphedema may be caused by agenesis, hypoplasia, hyperplasia, or obstruction of the lymphatic vessels. There are three clinical subtypes: congenital lymphedema, which appears shortly after birth; lymphedema praecox, which has its onset at the time of puberty; and lymphedema tarda, which usually begins after age 35. Familial forms of congenital lymphedema (Milroy's disease) and lymphedema praecox (Meige's disease) may be inherited in an autosomal dominant manner with variable penetrance; autosomal or sex-linked recessive forms are less common. At least 19 genes are associated with inherited forms of lymphedema. Mutations in genes expressing vascular endothelial growth factor receptor 3 (*VEGFR3*), which is a determinant of lymphangiogenesis, cause Milroy's disease; and a mutation of the gene encoding *VEGF-C*, a ligand for *VEGFR3*, may cause a Milroy's disease-like phenotype. A mutation of the *LSC1* gene is associated with the cholestasis-lymphedema syndrome. Mutations in the *FOXC2* gene, which encodes a transcription factor that interacts with a signaling pathway involved in the development of lymphatic vessels, cause the lymphedema-distichiasis syndrome, in which lymphedema praecox occurs in patients who also have a double row of eyelashes. A mutation of *SOX18*, a transcription factor upstream of lymphatic endothelial cell differentiation, has been described in patients with lymphedema, alopecia, and telangiectasias (hypotrichosis, lymphedema, telangiectasia syndrome). Mutations of the *CCBE1* gene, which enhances the lymphangiogenic effects of VEGF-C, cause Hennekam lymphangiectasia-lymphedema

syndrome, and *KIF11* gene mutations are associated with microcephaly-lymphedema syndrome. Mutations of the *GATA2* gene, which is involved in the development of lymphatic valves, cause lymphedema and a predisposition to acute myeloid leukemia. Patients with a chromosomal aneuploidy, such as Turner's syndrome, Klinefelter's syndrome, or trisomy 18, 13, or 21, may develop lymphedema. Syndromic vascular anomalies associated with lymphedema also include Klippel-Trénaunay syndrome and Parkes-Weber syndrome. Other disorders associated with lymphedema include Noonan's syndrome, yellow nail syndrome, intestinal lymphangiectasia syndrome, lymphangiomyomatosis, and neurofibromatosis type 1.

TABLE 276-2 Causes of Lymphedema

Primary

Sporadic (no identified cause)

Genetic disorders

- Milroy's disease (*VEGFR3, VEGF-C*)
- Meige's disease (gene mutation not established)
- Lymphedema-distichiasis syndrome (*FOXC2*)
- Cholestasis-lymphedema (*LSC1*)
- Hennekam's lymphangiectasia-lymphedema syndrome (*LCCBE1*)
- Emberger's syndrome-lymphedema and predisposition to AML (*GATA2*)
- Microcephaly-lymphedema syndrome (*KIF11*)
- Hypotrichosis-lymphedema-telangiectasia (*SOX18*)

Chromosomal aneuploidies

- Turner's syndrome
- Klinefelter's syndrome
- Trisomy 13, 18, or 21

Other disorders associated with primary lymphedema

- Noonan's syndrome
- Klippel-Trénaunay syndrome
- Parkes-Weber syndrome
- Yellow nail syndrome
- Intestinal lymphangiectasia syndrome
- Lymphangiomyomatosis
- Neurofibromatosis type 1

Secondary

Infection

- Bacterial lymphangitis (*Streptococcus pyogenes, Staphylococcus aureus*)
- Lymphogranuloma venereum (*Chlamydia trachomatis*)
- Filariasis (*Wucheria bancrofti, Brugia malayi, B. timori*)
- Tuberculosis

Neoplastic infiltration of lymph nodes

Lymphoma

Prostate

Others

Surgery or irradiation of axillary or inguinal lymph nodes for treatment of cancer

Iatrogenic

Lymphatic division (during peripheral bypass surgery, varicose vein surgery, or harvesting of saphenous veins)

Miscellaneous

Contact dermatitis

Podoconiosis

Rheumatoid arthritis

Pregnancy

Factitious

Secondary lymphedema is an acquired condition that results from damage to or obstruction of previously normal lymphatic channels. Recurrent episodes of bacterial lymphangitis, usually caused by streptococci, are a very common cause of lymphedema. The most common etiology of secondary lymphedema worldwide is lymphatic filariasis and affecting >120 million children and adults and causing lymphedema and elephantiasis in 14 million of these affected individuals **(Chap. 228)**. Recurrent bacterial lymphangitis by *Streptococcus* may result in chronic lymphedema. Other infectious causes include lymphogranuloma venereum and tuberculosis. A common acquired cause of lymphedema in tropical countries is podoconiosis, which results from barefoot exposure and absorption of silicate particles in soil derived from volcanic rock. In developed countries, the most common secondary cause of lymphedema is surgical excision or irradiation of axillary and inguinal lymph nodes for treatment of cancers, such as breast, cervical, endometrial, and prostate cancer, sarcomas, and malignant melanoma. Lymphedema of the arm occurs in 13% of breast cancer patients after axillary node dissection and in 22% after both surgery and radiotherapy. Lymphedema of the leg affects ~15% of patients with cancer after inguinal lymph node dissection. Tumors, such as prostate cancer and lymphoma, also can infiltrate and obstruct lymphatic vessels. Less common causes include contact dermatitis, rheumatoid arthritis, pregnancy, and self-induced or factitious lymphedema after application of tourniquets.

Clinical Presentation

Lymphedema is generally a painless condition, but patients may experience a chronic dull, heavy sensation in the leg, and most often they are concerned about the appearance of the leg. Lymphedema of the lower extremity initially involves the foot and gradually progresses up the leg so that the entire limb becomes edematous **(Fig. 276-2)**. In the early stages, the edema is soft and pits easily with pressure. Over time, subcutaneous adipose tissue accumulates, the limb enlarges further and loses its normal contour, and the toes appear square. Thickening of the skin is detected by Stemmer's sign, which is the inability to tent the skin at the base of the toes. Peau d'orange is a term used to describe dimpling of the skin, resembling that of an orange peel, caused by lymphedema. In

the chronic stages, the edema no longer pits and the limb acquires a woody texture as the tissues become indurated and fibrotic. The International Society of Lymphology describes four clinical stages of lymphedema **(Table 276-3)**.

TABLE 276-3 Stages of Lymphedema

Stage 0 (or Ia)

A latent or subclinical condition where swelling is not evident despite impaired lymph transport. It may exist for months or years before overt edema occurs.

Stage I

Early accumulation of fluid relatively high in protein content that subsides with limb elevation. Pitting may occur. An increase in proliferating cells may also be seen.

Stage II

Limb elevation alone rarely reduces tissue swelling, and pitting is manifest. Late in stage II, the limb may or may not pit as excess fat and fibrosis supervene.

Stage III

Lymphostatic elephantiasis where pitting can be absent and trophic skin changes such as acanthosis, further deposition of fat and fibrosis, and warty overgrowths have developed.

Source: Adapted from The 2013 Consensus Document of the International Society of Lymphology: Lymphology 46:1, 2013.

FIGURE 276-2

A. Lymphedema characterized by swelling of the leg, nonpitting edema, and squaring of the toes. *(Courtesy of Dr. Marie Gerhard-Herman, with permission.)* ***B.*** Advanced chronic stage of lymphedema illustrating the woody appearance of the leg with acanthosis and verrucous overgrowths. *(Courtesy of Dr. Jeffrey Olin, with permission.)*



A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Differential Diagnosis

Lymphedema should be distinguished from other disorders that cause unilateral leg swelling, such as deep-vein thrombosis and chronic venous insufficiency. In the latter condition, the edema is softer, and there is often evidence of a stasis dermatitis, hyperpigmentation, and superficial venous varicosities, as described earlier. Other causes of leg swelling that resemble lymphedema are myxedema and lipedema. Lipedema usually occurs in women and is caused by accumulation of adipose tissue in the leg from the thigh to the ankle with sparing of the feet.

Diagnostic Testing

The evaluation of patients with lymphedema should include diagnostic studies to clarify the cause. Abdominal and pelvic ultrasound and computed tomography (CT) can be used to detect obstructing lesions such as neoplasms. Magnetic resonance imaging (MRI) of the affected limb may reveal a honeycomb pattern characteristic of lymphedema in the epifascial compartment and identify enlarged lymphatic channels and lymph nodes. MRI also is useful to distinguish lymphedema from lipedema. Lymphoscintigraphy and lymphangiography are rarely indicated, but either can be used to confirm the diagnosis or differentiate primary from secondary lymphedema. Lymphoscintigraphy involves the injection of radioactively labeled technetium-containing colloid into the distal subcutaneous tissue of the affected extremity, which is imaged with a scintigraphic camera to visualize lymphatic vessels and

lymph nodes. Findings indicative of primary lymphedema include absent or delayed filling of the lymphatic vessels or dermal back flow caused by lymphatic reflux. Findings of secondary lymphedema include dilated lymphatic vessels distal to an area of obstruction. In lymphangiography, iodinated radiocontrast material is injected into a distal lymphatic vessel that has been isolated and cannulated. In primary lymphedema, lymphatic channels are absent, hypoplastic, or ectatic. In secondary lymphedema, lymphatic channels often appear dilated beneath the level of obstruction. The complexities of lymphatic cannulation and the risk of lymphangitis associated with the contrast agent limit the utility of lymphangiography. A novel technique of optical imaging with a near-infrared fluorescence dye may enable quantitative imaging of lymph flow.

TREATMENT

TREATMENT Lymphedema

Patients with lymphedema of the lower extremities must be instructed to take meticulous care of their feet and legs to prevent cellulitis and lymphangitis. Skin hygiene is important, and emollients can be used to prevent drying. Prophylactic antibiotics are often helpful, and fungal infection should be treated aggressively. Patients should be encouraged to participate in physical activity; frequent leg elevation can reduce the amount of edema. Psychosocial support is indicated to assist patients cope with anxiety or depression related to body image, self-esteem, functional disability, and fear of limb loss.

Physical therapy, including massage to facilitate lymphatic drainage, may be helpful. The type of massage used in decongestive physiotherapy for lymphedema involves mild compression of the skin of the affected extremity to dilate the lymphatic channels and enhance lymphatic motility. Multilayered, compressive bandages are applied after each massage session to reduce recurrent edema. After optimal reduction in limb volume by decongestive physiotherapy, patients can be fitted with graduated compression hose. Occasionally, intermittent pneumatic compression devices can be applied at home to facilitate reduction of the edema. Diuretics are contraindicated and may cause depletion of intravascular volume and metabolic abnormalities.

Liposuction in conjunction with decongestive physiotherapy may be considered to treat lymphedema, particularly postmastectomy lymphedema. Other surgical interventions are rarely used and often not successful in ameliorating lymphedema. Microsurgical lymphaticovenous anastomotic procedures have been performed to rechannel lymph flow from obstructed lymphatic vessels into the venous system. Limb reduction procedures to resect subcutaneous tissue and excessive skin are performed occasionally in severe cases of lymphedema to improve mobility.

Therapeutic lymphangiogenesis has been studied in rodent models of lymphedema. Overexpression of VEGF-C generates new lymphatic vessels and improves lymphedema in a murine model of primary lymphedema, and administration of recombinant VEGF-C or VEGF-D stimulated lymphatic growth in preclinical models of postsurgical lymphedema. There may be additional benefit when administered in conjunction with lymph node transfer. Clinical trials in patients with lymphedema are required to determine efficacy of gene transfer and cell-based therapies for lymphedema.

FURTHER READING

Aspelund A et al: Lymphatic system in cardiovascular medicine. Circ Res 118:515, 2016. [PubMed: 26846644]
Brouillard P et al: Genetics of lymphatic anomalies. J Clin Invest 124:898, 2014. [PubMed: 24590274]
Eberhardt RT, Raffetto JD: Chronic venous insufficiency. Circulation 130:333, 2014. [PubMed: 25047584]

Garg N, Gloviczki P: Chronic venous insufficiency, in *Vascular Medicine*, Creager MA, Beckman JA, Loscalzo J (eds). Philadelphia, Elsevier, 2013, pp 652–666.

Gloviczki P and associated chronic venous diseases: Clinical practice guidelines of the Society for Vascular Surgery and the American Venous Forum. J Vasc Surg 53:2S, 2011. [PubMed: 21536172]

Kahn SR et al: The postthrombotic syndrome: Evidence-based prevention, diagnosis, and treatment strategies: A scientific statement from the American Heart Association. Circulation 130:1636, 2014. [PubMed: 25246013]

Karaca-Mandic P et al: The cutaneous, net clinical and health economic benefits of advanced pneumatic compression devices in patients with lymphedema. JAMA Dermatol 151:1187, 2015. [PubMed: 26444458]

Lee BB et al: Diagnosis and treatment of primary lymphedema. Consensus document of the International Union of Phlebology. Int Angiol 32:541, 2013. [PubMed: 24212289]

Mortimer PS, Rockson SG: New developments in clinical aspects of lymphatic disease. J Clin Invest 124:915, 2014. [PubMed: 24590276]

Tafur AJ, Rathbun S: Varicose Veins, in *Vascular Medicine*. Creager MA, Beckman JA, Loscalzo J (eds). Philadelphia, Elsevier, 2013, pp 639–651.

Chapter 277: Pulmonary Hypertension

Aaron B. Waxman; Joseph Loscalzo

INTRODUCTION

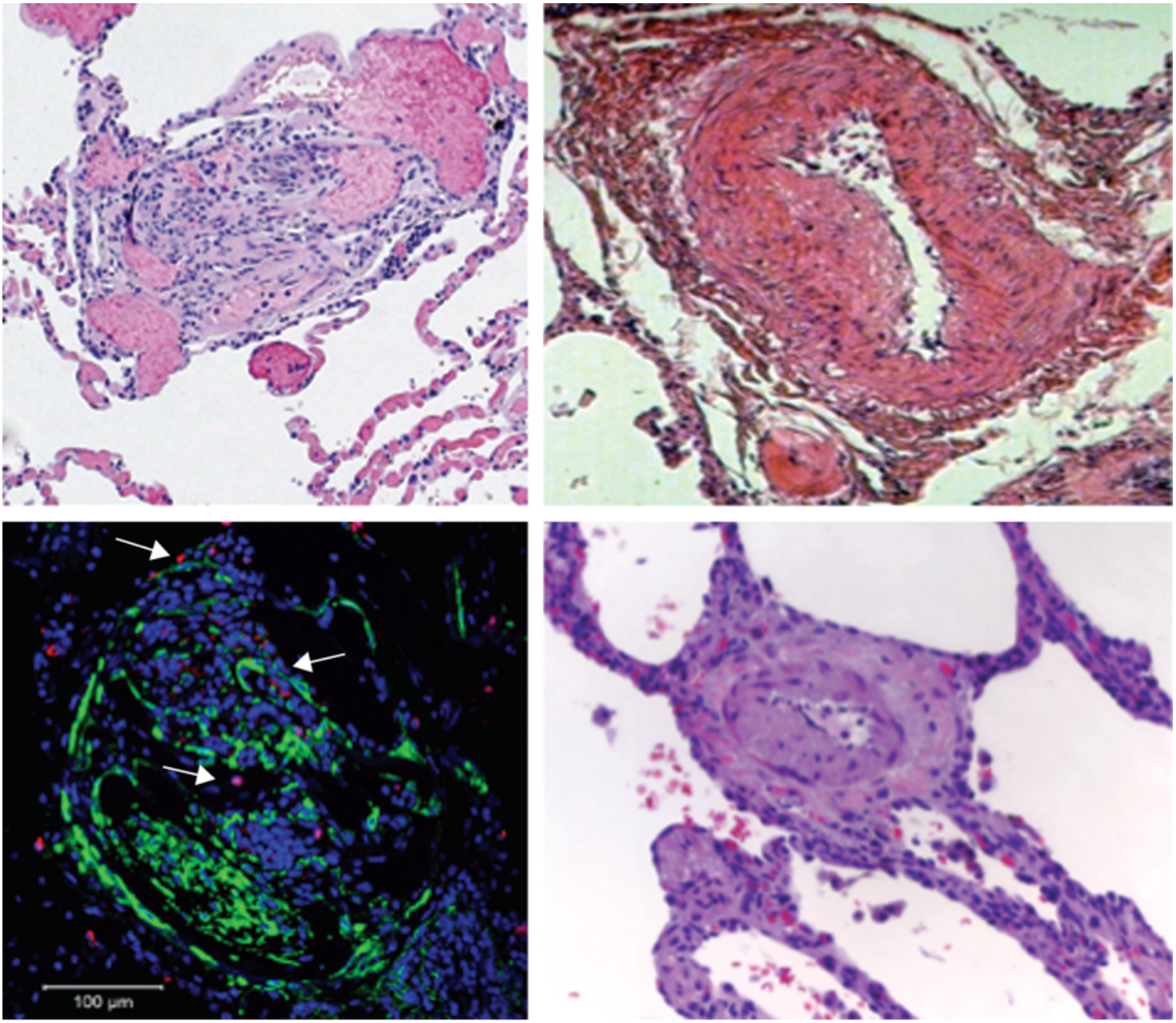
Pulmonary hypertension (PH) is a spectrum of diseases involving the pulmonary vasculature, and defined as an elevation in pulmonary arterial pressures (mean pulmonary artery pressure [PAP] >22 mmHg or an estimated systolic PAP >36 mmHg). Pulmonary arterial hypertension (PAH) is a relatively rare form of PH and is characterized by symptoms of dyspnea, chest pain, and syncope. If left untreated, the disease carries a high mortality rate, with the most common cause of death being decompensated right heart failure. There have been significant advances in this field in regard to understanding the pathogenesis, diagnosis, and classification of PAH. Despite these significant advances, there is still a substantial delay in diagnosis of up to 2 years. In many cases, patients whose primary complaint is exertional intolerance are frequently misdiagnosed with more common diseases such as asthma or chronic obstructive pulmonary disease. The availability of newer drugs has resulted in a radical change in the management of this disease with significant improvement in both quality of life and mortality. A delay in diagnosis results in an obvious delay in the initiation of appropriate treatment. Clinicians should be able to recognize the signs and symptoms of PH and to complete a systematic workup in at-risk patients. In this way, early diagnosis, prompt treatment, and improved outcomes for patients become achievable.

PATHOBIOLOGY

Vasoconstriction, vascular proliferation, thrombosis, and inflammation appear to underlie the development of PAH **(Fig. 277-1)**. Intimal proliferation and fibrosis, medial hypertrophy, and *in situ* thrombosis characterize the pathological findings in the pulmonary vasculature. Vascular remodeling at earlier stages may be confined to the small distal pulmonary arteries. As the disease advances, intimal proliferation and pathologic remodeling progress resulting in decreased compliance of the pulmonary vasculature. The outcome is a progressive increase in the right ventricular afterload or total pulmonary vascular resistance (PVR), and, thus, right ventricular work. In subjects with moderate to severe pulmonary vascular disease, as the resting PVR increases, there will be a corresponding increase in mean PAP until the cardiac output (CO) is compromised and starts to fall. With a decline in CO, the PAP will fall. As CO declines as a result of increased afterload and decreased contractility, tachycardia is a compensatory response. Tachycardia decreases filling time and, thus, preload, and results in a reduced fraction of stroke volume available to distend the pulmonary vascular tree.

Figure 277-1.

Panels on the left show examples of plexogenic pulmonary arteriopathy. These are obstructive and proliferative lesions of the small muscular pulmonary arteries, composed primarily of endothelial cells with intermixed inflammatory cells, myofibroblasts, and connective tissue components. The lower left panel demonstrates proliferating cells (red PCNA stain—*white arrows*). Panels on the right demonstrate medial hypertrophy of muscular pulmonary arteries. *(Photograph on the lower left is courtesy of Dr. Stephen Archer, Queen's University School of Medicine, Kingston, Ontario, Canada.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Abnormalities in multiple molecular pathways and genes that regulate the pulmonary vascular endothelial and smooth muscle cells have been identified **(Table 277-1)**. These abnormalities include decreased expression of the voltage-regulated potassium channel, mutations in the bone morphogenetic protein receptor-2, increased tissue factor expression, overactivation of the serotonin transporter, hypoxia-induced activation of hypoxia-inducible factor-1α, and activation of nuclear factor of activated T cells. As a result, there is a decrease in apoptosis of smooth muscle cells and the emergence of apoptosis-resistant endothelial cells that promote their accumulation and can obliterate the vascular lumen. In addition, thrombin deposition in the pulmonary vasculature from the prothrombotic state that develops as an independent abnormality or as a result of endothelial dysfunction may amplify the obliterative arteriopathy.

TABLE 277-1 Molecular Determinants of the Pathogenesis of Pulmonary Arterial Hypertension

Alterations in regulators of proliferation

- Growth factors
  - PDGF
  - FGF
  - VEGF
  - EGF
- TGF-β
- BMP
- Transcription factors
- MMPs
- Cytokines
- Chemokines
- Mitochondria

Alterations in inflammatory mediators

- Altered T-cell subsets
- Monocytes and macrophages
- IL-1β
- IL-6
- MCP-1
- RANTES
- Fractalkine

Alterations in vascular tone

- Endothelin
- Nitric oxide
- Serotonin
- Prostaglandin
- K^+ channels
- Ca^{2+} channels

Hypoxia induced remodeling

- HIF-1α
- ROS
- Mitochondria

Alterations in TGF-β signaling pathways

- BMPR2
- ALK1
- Endoglin
- Smad9
- TGF-β1

Abbreviations: ALK, anaplastic lymphoma kinase; BMP, bone morphogenic protein; EGF, epidermal-derived growth factor; FGF, fetal-derived growth factor; HIF-1α, hypoxia-inducible factor-1; IL, interleukin; MCP-1, monocyte chemoattractant protein-1; MMP, mucous membrane pemphigoid; PDGF, platelet-derived growth factor; ROS, reactive oxygen specie; TGFβ, transforming growth factor β; VEGF, vascular endothelial-derived growth factor.

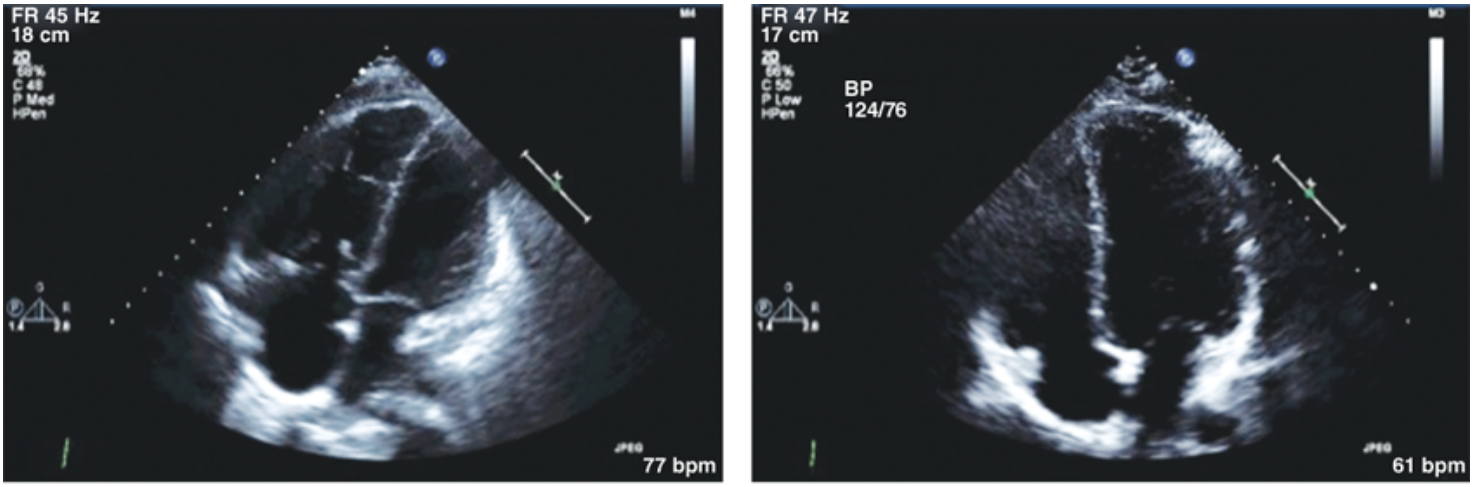
DIAGNOSIS AND CLASSIFICATION

The diagnosis of PH can be missed without a reasonable index of suspicion. PH symptoms are nonspecific, insidious, and overlap considerably with many common conditions, including asthma and other lung disease, and cardiac disease. Most patients will present with dyspnea and/or fatigue, whereas edema, chest pain, presyncope, and syncope are less common and associated with more advanced disease. In early phases of PAH, the physical examination is often unrevealing. As the disease progresses there may be evidence of right ventricular failure with elevated jugular venous pressure, lower extremity edema, and ascites. Additionally, the cardiovascular examination may reveal an accentuated P2 component of the second heart sound, a right-sided S3 or S4, and a holosystolic tricuspid regurgitant murmur. It is also important to seek signs of the diseases that are often concurrent with PH: clubbing may be seen in some chronic lung diseases, sclerodactyly and telangiectasia may signify scleroderma, and crackles on examination of the lungs and systemic hypertension may be clues to left-sided systolic or diastolic heart failure.

Once clinical suspicion is raised, a systematic approach to diagnosis and assessment is essential. An echocardiogram with *bubble study* is the most important initial screening test. Echocardiography is also important for determining specific causes. All forms of PH may demonstrate a hypertrophied and dilated right ventricle **(Fig. 277-2)** with elevated estimated pulmonary artery systolic pressure. Important additional information can be gleaned about specific etiologies of PH, such as valvular disease, left ventricular systolic and diastolic function, intra-cardiac shunts, and other cardiac diseases.

Figure 277-2

Panel (***A***) is a representative echocardiogram showing the apical 4-chamber view from a patient with pulmonary hypertension demonstrating an enlarged right atrium and ventricle with some compression of the left side of the heart. Panel (***B***) is the same echocardiographic view showing a normal echocardiogram.



A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

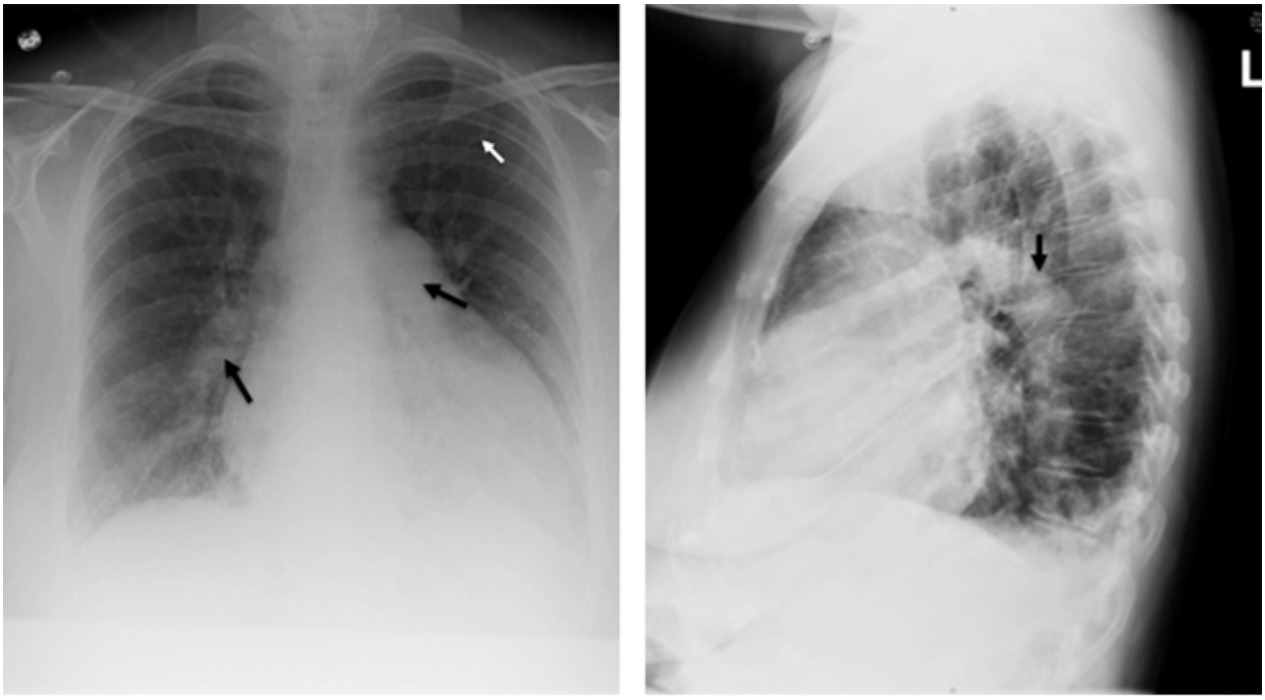
Although the accuracy of Doppler echocardiography is often debated, a high quality echocardiogram that is absolutely normal may obviate the need for further evaluation for PH. An echocardiogram is a screening test, whereas invasive hemodynamic monitoring is the gold standard for diagnosis and assessment of disease severity. With a normal echocardiogram, there may still be some concern for PH; this is particularly true if there is unexplained dyspnea or hypoxemia. In this setting, it is reasonable to proceed to right heart catheterization (RHC) for definitive diagnosis. Alternatively, if the patient has a reasonable functional capacity, a cardiopulmonary exercise test may help to identify a true physiologic limitation as well as differentiate between cardiac and pulmonary causes of dyspnea. If this test is normal, there is no indication for a RHC.

If the echocardiogram or cardiopulmonary exercise test (CPET) suggests PH and the diagnosis is confirmed by catheterization, a reasonable effort must be made to establish the etiology because this will largely determine the therapeutic approach. A stepwise approach to evaluation is outlined below.

Chest imaging and lung function tests are essential because lung disease is an important cause of PH. Signs of PH that may be evident on chest radiograph include enlargement of the central pulmonary arteries, “vascular pruning,” and cardiomegaly **(Fig. 277-3)**. High-resolution computed tomography (CT) may provide additional useful information. Classic findings of PH on CT include those found on chest radiograph: enlarged pulmonary arteries **(Fig. 277-4)**, peripheral pruning of the small vessels, and enlarged right ventricle and atrium. However, high-resolution CT may also reveal signs of venous congestion, including centrilobular ground glass infiltrate and thickened septal lines. In the absence of left heart disease, these findings suggest pulmonary venous disease, a rare cause of PAH that can be quite challenging to diagnose. CT is also critical for distinguishing co-morbid interstitial lung disease or emphysema.

Figure 277-3

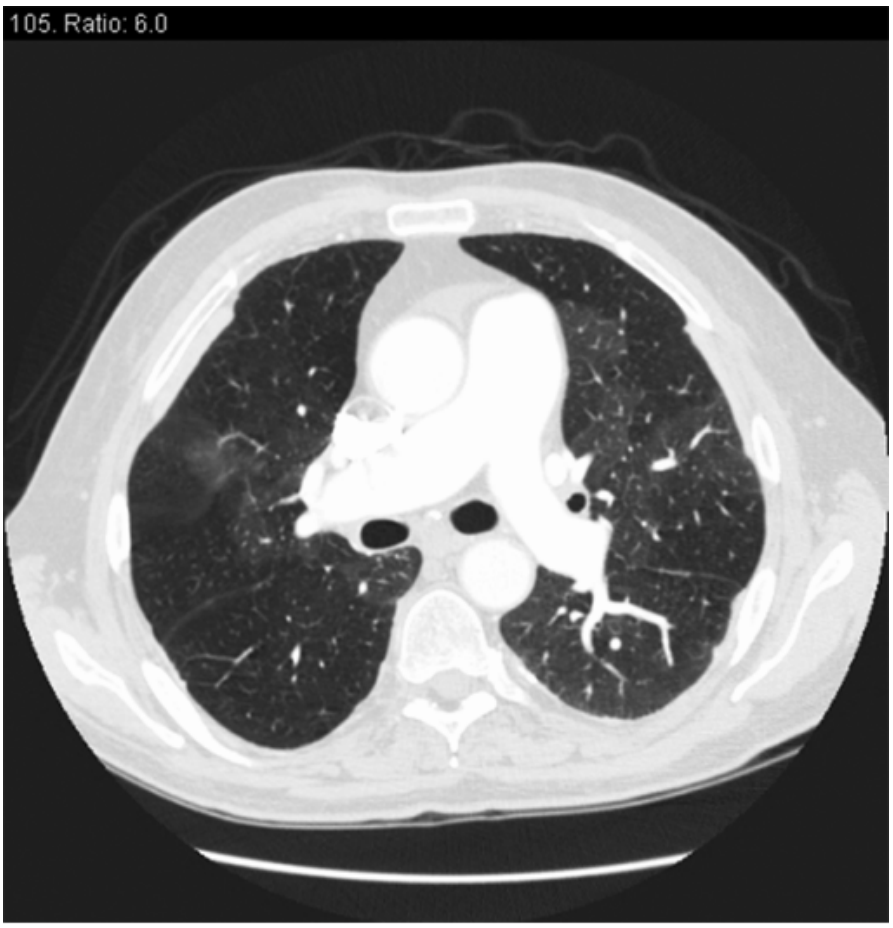
Postero-anterior (*left*) and lateral (*right*) chest radiograph showing enlarged pulmonary arteries (*black arrows*) and pruning of the distal pulmonary vasculature (*white arrows*) commonly seen with advanced pulmonary arterial hypertension.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Figure 277-4

Representative computed tomographic scan of the chest demonstrating enlarged main pulmonary arteries. There is also a mosaic pattern evident in both lungs.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CT angiograms are commonly used to evaluate acute thromboembolic disease and have demonstrated excellent sensitivity and specificity for that purpose. Ventilation-perfusion ($\dot{V}/\dot{Q}$) scanning has been used for screening because of its high sensitivity and its role in qualifying patients for surgical intervention. The role of CT angiograms in the diagnosis of chronic thromboembolic pulmonary hypertension (CTEPH) remains controversial, even with the advent of spiral CT. Although a negative $\dot{V}/\dot{Q}$ virtually rules out CTEPH, rare cases may be missed through the use of CT angiograms.

Pulmonary function tests are an important component of the evaluation. While an isolated reduction in diffusing capacity of the lungs for carbon monoxide (DLCO) is the classic finding in PAH, results of pulmonary function tests may also suggest restrictive or obstructive lung diseases as the cause of dyspnea or PH. The 6-minute walk test is also important to evaluate the degree of exertional hypoxemia and limitation, and to monitor progression and response to therapy. Increasingly, sub-maximal and maximal exercise testing are being utilized for screening and characterization of disease because it provides a more objective measure of breathing efficiency (V_E/VCO_2 slope).

Sleep-disordered breathing is another important cause of mild PH, but a sleep study is generally necessary only when indicated by the patient's history. Nocturnal desaturation is a common finding in PH, even in the absence of sleep-disordered breathing. Thus, all patients should undergo nocturnal oximetry screening, regardless of whether classic symptoms of obstructive sleep apnea or obesity-hypoventilation syndrome are observed. Laboratory tests that are important for screening include an HIV test when clinically indicated. In addition, all patients should have antinuclear antibodies, rheumatoid factor, and anti-scl-70 antibodies assessed to screen for the most common rheumatologic diseases associated with PH. Liver function and hepatitis serology tests are important to screen for underlying liver disease. Finally, there is an increasing role for brain natriuretic peptide testing in the diagnosis and management of PH. Brain naturetic peptide (BNP) and the N-terminus of its pro-peptide (NT-proBNP) correlate with right ventricular (dys)function, hemodynamic severity, and functional status in PAH.

RHC with pulmonary vasodilator testing remains the gold standard both to establish the diagnosis of PH and to guide selection of appropriate medical therapy. The definition of precapillary PH or PAH requires (1) an increased mean PAP (mPAP >25 mmHg); (2) a pulmonary capillary wedge pressure (PCWP), left atrial pressure, or left ventricular end-diastolic pressure (LVEDP) ≤15 mmHg; and (3) PVR >3 Wood units. Post capillary PH is differentiated from

precapillary PH by the PCWP being ≥15 mmHg; this is further differentiated into passive, based on a transpulmonary gradient <12 mmHg, or reactive, based on a transpulmonary gradient >12 mmHg and an increased PVR. In either case, the CO may be normal or reduced.

Vasodilators with a short duration of action, such as inhaled nitric oxide (NO), or inhaled epoprostenol are preferred for vasodilator testing. A decrease in mPAP by ≥10 mmHg to an absolute level ≤40 mmHg without a decrease in CO is defined as a positive pulmonary vasodilator response, and responders are considered for long-term treatment with calcium channel blockers (CCB). Less than 15% of patients are deemed vasoreactive during testing, and even fewer exhibit long-term responsiveness to CCB. An acute vasodilator-induced reduction in PVR and mPAP predict better long-term survival even among those patients not treated with CCB. The diagnosis of PH is increasing in the older population, in part, because of increased awareness of this disease in the elderly and increased use of screening echocardiograms. Furthermore, the increased availability of oral and less complicated therapeutic options has encouraged the referral of older patients for evaluation and treatment.

PULMONARY HYPERTENSION AS A COMORBID DISEASE

PAH is but one of a number of disease classifications that affect the pulmonary vascular bed. As understanding of the various contributing diseases has increased, classification systems have attempted to group these diseases by clinical features to aid in diagnosis. The World Health Organization (WHO) formulated a clinical classification of the various manifestations of PH, of which PAH is a subgroup, according to similarities in pathophysiologic mechanisms and clinical presentation. PH is a diverse mix of pathologies in which the only unifying theme is elevated PAP relative to left atrial pressure. The categorization of PH was designed by convenience for the purpose of facilitating novel treatments to be tested across different presentations. Efforts are underway to define pulmonary vascular diseases based on molecular phenotyping that, in the future, may offer a guide for improved management decisions as precision medicine strategies continue to evolve.

The current classification system, last revised in 2013 during the Fifth World Symposium on Pulmonary Hypertension, recognizes five categories of PH, including PAH, PH due to left heart disease, PH due to chronic lung disease, PH associated with chronic thromboemboli, and a group of miscellaneous diseases that only rarely cause PH.

Pulmonary Arterial Hypertension

WHO Group I PH, PAH, is a relatively rare cause of PH. PAH includes a group of diseases that result in pulmonary arterial precapillary remodeling marked by intimal fibrosis, increased medial thickness, pulmonary arteriolar occlusion, and classic plexiform lesions. PAH is defined as a sustained elevation in resting mean pulmonary arterial pressure (mPAP) ≥25 mmHg, PVR > 240 dyne-s/cm^5, and PCWP or LVEDP of ≤15 mmHg based on a RHC. With a normal PCWP and an elevated mPAP, these diseases demonstrate an increased transpulmonary gradient (mPAP– PCWP) and increased PVR.

Idiopathic PAH (IPAH) is a progressive disease that leads to right heart failure and death. The National Institutes of Health registry, the first large registry of patients with PAH, reported that the average age at diagnosis was 36 years, with only 9% of patients with IPAH over the age of 60. However, more current clinical data suggest that the patient demographics are changing. The Pulmonary Hypertension Connection registry found that the average age of diagnosis for IPAH was 45 years, with 8.5% of patients older than 70 years at diagnosis. This finding is supported by data from the Registry to Evaluate Early And Long-term PAH Disease Management (REVEAL) study, the largest cohort of PAH to date, which reported that the average age at diagnosis of IPAH was 44.9+/-0.6 years.

Other forms of PAH that deserve specific consideration are those associated with HIV, connective tissue disease, and portal hypertension. Although HIV is a rare cause of PAH, this form of PAH is indistinguishable from IPAH and is an important cause of mortality in the HIV-infected population. Importantly, there is no correlation between the stage of HIV infection and the development of PAH.

Among connective tissue diseases, the prevalence of PAH has been established only for systemic sclerosis, especially in those with limited cutaneous scleroderma. Although the average age of scleroderma onset is from 30 to 50 years old, patients who eventually develop scleroderma-associated PAH tend to be older at the time of scleroderma diagnosis. Outcomes of scleroderma are closely linked to the development of PAH and are associated with a poor prognosis, although modern therapies have improved outcomes.

Portopulmonary hypertension occurs in 2–10% of patients with established portal hypertension. Its occurrence appears to be independent of the cause of liver disease and is observed in patients with non-hepatic causes of portal hypertension. A hyperdynamic circulatory state is common, as in most patients with advanced liver disease; however, the same pulmonary vascular remodeling observed in other forms of PAH is seen in the pulmonary vascular bed in

portopulmonary hypertension. It is important to distinguish this process from hepatopulmonary syndrome, which can also manifest with dyspnea and hypoxemia but is pathophysiologically distinct from portopulmonary hypertension in that abnormal vasodilation of the pulmonary vasculature leads to intrapulmonary shunting.

Pulmonary Hypertension Associated with Left Heart Disease

WHO Group II PH includes patients with left heart systolic dysfunction, aortic and mitral valve disease, and heart failure with preserved ejection fraction (HFpEF). PH can develop as a result of each of these conditions. The hallmark of Group II PH is elevated left atrial pressure with resulting pulmonary venous hypertension. In general, the transpulmonary gradient and PVR remain normal. Although this phenomenon is well described in both left-sided valvular disease and left-sided systolic heart failure, studies suggest that HFpEF may carry a higher overall risk of PH.

Whatever the cause of elevated left atrial pressure (i.e., systolic or diastolic heart failure or valvular disease), the increased pulmonary venous pressure indirectly leads to a rise in pulmonary arterial pressure. The presence of PH portends a poor prognosis in all forms of heart failure. In particular, chronic pulmonary venous hypertension may lead to a reactive pulmonary arterial vasculopathy, seen as an elevated transpulmonary gradient (>12 mm Hg) and elevated PVR (>3 Wood units). Pathologically, this process is marked by pulmonary arteriolar remodeling with intimal fibrosis and medial hyperplasia akin to that seen in PAH.

Pulmonary Hypertension Associated with Lung Disease

Intrinsic lung disease is the second most common cause of PH, although its actual prevalence is difficult to ascertain. PH has been observed in both chronic obstructive lung disease and interstitial lung disease. It can also be seen in diseases with mixed obstructive/restrictive physiology: bronchiectasis, cystic fibrosis, mixed obstructive restrictive disease marked by fibrosis in the lower lung zones, and emphysema predominantly in the upper lung zones. As in patients with left heart disease, PH associated with chronic lung disease is usually modest; however, some of these patients appear to have PH “out of proportion” to their parenchymal lung disease, suggesting intrinsic pulmonary arterial disease. These patients typically have more severe PH, with results of pulmonary function tests demonstrating a very low DLCO.

Although PH is described in most forms of interstitial lung disease, it has been most extensively studied in idiopathic pulmonary fibrosis; however, the individual studies have been small. Early echocardiographic data suggested that the prevalence of PH in interstitial lung diseases is high, but invasive hemodynamic monitoring suggests that the incidence is considerably lower than originally believed. The diagnosis of PH portends poor outcome in pulmonary fibrosis.

Also included in Group III PH is sleep-disordered breathing. Sleep apnea has long been associated with PH. However, PH associated with sleep-disordered breathing is generally mild.

PH Associated with Chronic Thromboembolic Disease

The development of PH after chronic thromboembolic obstruction of the pulmonary arteries is well described, but its incidence is not known. The incidence of PH after a single pulmonary embolic event is thought to be quite low, and likely increases following recurrent embolism. The risk factors for developing CTEPH are unclear. Many patients have no history of clinical venous thromboembolism. The pathogenesis of CTEPH is poorly understood. Obstruction of the proximal pulmonary vasculature is important and often the dominant factor; however, additional pulmonary vascular remodeling occurs. Approximately 10–15% of patients will develop a disease very similar clinically and pathologically to PAH after resection of the proximal thrombus.

OTHER DISORDERS AFFECTING THE PULMONARY VASCULATURE

Sarcoidosis

Patients with sarcoidosis can develop PH as a result of lung involvement, and those who present with progressive dyspnea and PH require a thorough evaluation. While many sarcoidosis patients with PH do not respond to therapy for PAH, there is a subset of patients with sarcoidosis and severe PH who do have a beneficial response to therapy.

Sickle Cell Disease

Cardiovascular system abnormalities are prominent in the clinical spectrum of sickle cell disease, including PH. The etiology is multifactorial, including hemolysis, hypoxemia, thromboembolism, chronic high CO, and chronic liver disease. The presence of PH in patients with sickle cell disease is rare.

Schistosomiasis

Globally, schistosomiasis is one of the most common causes of PH. The development of PH occurs in the setting of hepatosplenic disease and portal hypertension. Studies suggest that inflammation from the infection triggers the pulmonary vascular changes that occur. The diagnosis is confirmed by finding the parasite ova in the urine or stool of patients with symptoms, which can be difficult. The efficacy of therapies directed toward PH in these patients is unknown.

PHARMACOLOGIC TREATMENT OF PAH

Without treatment PAH is invariably fatal. There are a rapidly growing number of approved agents for PAH, including prostacyclin and prostacyclin analogues and agonists, NO pathway enhancers, and endothelin receptor antagonists (ERAs) that have improved the outlook dramatically. While there is no cure for PAH, current pharmacologic therapies improve morbidity, and in some cases, mortality.

PROSTANOIDS

In PAH, endothelial dysfunction and platelet activation cause an imbalance of arachidonic acid metabolites with reduced prostacyclin levels and increased thromboxane A_2 production. Prostacyclin (PGI_2) activates cyclic adenosine monophosphate (cAMP)-dependent pathways that mediate vasodilation. PGI_2 also has antiproliferative effects on vascular smooth muscle and inhibits platelet aggregation. Protein levels of prostacyclin synthase are decreased in pulmonary arteries of patients with PAH. This imbalance of mediators is addressed by the exogenous administration of prostanoids as therapy in advanced PAH.

Epoprostenol was the first prostanoid available for the management of PAH. Epoprostenol delivered as a continuous intravenous infusion improves functional capacity and survival in PAH. The efficacy of epoprostenol in WHO FC class 3 and 4 PAH patient was demonstrated in a clinical trial that showed improved quality of life, mPAP, PVR, 6-minute walk distance (MWD) and mortality. Treprostinil has a longer half-life than epoprostenol (~4 hours vs ~6 minutes), which allows for subcutaneous and continuous intravenous administration. Treprostinil has been shown to improve pulmonary hemodynamics, symptoms, exercise capacity, and, survival in PAH.

Inhaled prostacyclins provide the beneficial effects of infused prostacyclin therapy without the inconvenience and side effects (risk of infection and infusion site reactions) of infusion catheters. Both inhaled iloprost and treprostinil have been approved for patients with WHO FC class 3 and 4 PAH. The main advantage of treprostinil is less frequent administration. Inhaled formulations can be efficacious in moderately symptomatic patients with PAH and may be appropriate when used in combination with an oral medication. Phosphodiesterase-5 (PDE5) inhibitors (e.g., sildenafil) increase cyclic guanosine monophosphate (cGMP) levels and activate cGMP-dependent signaling pathways that also mediate vasodilation and platelet inhibition. The addition of a phosphodiesterase-5 (PDE5) inhibitor, therefore, augments the pulmonary hemodynamic and functional capacity benefits of prostanoids in PAH.

Oral treprostinil (an extended release formulation) has been assessed in one randomized controlled trial in treatment-naïve and two-combination therapy randomized controlled trials. Oral treprostinil significantly improved 6MWD in comparison with placebo (+23 m in comparison with baseline, P=0.0125) but had no effect on clinical worsening. Both combination therapy trials, in which treprostinil was added on to background therapy with either a PDE5 inhibitor or an ETRA failed to meet their primary end point in 6MWD. As a result of improved exercise tolerance in the monotherapy trial, along with the established clinical efficacy of parenteral and inhaled treprostinil, however, the FDA-approved oral treprostinil for the treatment of WHO group 1 PAH. Oral treprostinil is dosed three times daily, and is slowly titrated to the (maximally) effective dose.

Selexipag is an oral nonprostanoid diphenylpyrazine derivative that binds the prostaglandin I_2 (IP) receptor with high affinity. The active metabolite of selexipag has a prolonged half-life in comparison with prostanoid analogues and permits twice daily dosing. The efficacy of selexipag has been evaluated in a phase 3 randomized controlled trial in patients with PAH in New York Heart Association (NYHA) FC II to III on background therapy with either an endothelin-1 (ET-1) receptor antagonist, sildenafil, or both. This trial represents the largest randomized placebo-controlled trial among patients with PAH ever completed, enrolling more than 1100 patients treated for a median of 1.4 years. Selexipeg significantly reduced the risk of hospitalization and the risk of disease progression by 43% (P<0.0001)

compared to those who received placebo. There were no significant differences in mortality between the two study groups. The side effect profile was similar to the prostacyclins.

Endothelin Receptor Antagonists

ERAs target ET-1, a potent endogenous vasoconstrictor and vascular smooth muscle mitogen that is elevated in PAH patients. Endothelin levels are increased coincident with increased PVR and mean PAP, and decreased CO and 6MWD.

ERAs block the binding of ET-1 to either endothelin receptor A (ET-A) and/or B (ET-B). ET-A receptors found on pulmonary artery smooth muscle cells (PASMC) mediate vasoconstriction. In the normal pulmonary vasculature, ET-B receptors are found on endothelial cells and mediate ET-1 clearance, as well as vasodilation via production of prostacyclin and NO. The three ERAs approved for use in the United States are bosentan and macitentan, non-selective receptor antagonists; and ambrisentan, a selective ET-A receptor antagonist.

Studies have shown that bosentan improves hemodynamics and exercise capacity and delays clinical worsening. The randomized, placebo controlled, phase III, Bosentan Randomized trial of Endothelin Antagonist THErapy (BREATHE)-1 trial comparing bosentan or placebo demonstrated improved symptoms, 6MWD, and WHO functional class. The Endothelin Antagonist Trial in Mildly Symptomatic Pulmonary Arterial Hypertension Patients (EARLY) study comparing bosentan to placebo demonstrated improved PVR and 6MWD.

Several studies, including the phase III placebo-controlled Ambrisentan in Pulmonary Arterial Hypertension, (ARIES)-1 trial, demonstrate that ambrisentan improves exercise tolerance, WHO functional class, hemodynamics, and quality of life in patients with PAH. There are no trial data to evaluate if the selective ET-A receptor antagonism of ambrisentan has any advantage over the non-selective ET receptor antagonism of bosentan or macitentan.

Nitric Oxide Pathway

Nitric oxide (NO) derived from endothelial cells activate guanylyl cyclase that, in turn, generates cGMP in vascular smooth muscle cells and platelets. cGMP is a second messenger that induces vasodilation through relaxation of the arterial smooth muscle cells and inhibits platelet activation. Phosphodiesterase type 5 enzymes metabolize cGMP. Therefore, cGMP phosphodiesterase type 5 (PDE5) inhibitors prolong the vasodilatory effect of NO, especially within the pulmonary arterial bed where high concentrations of cGMP are found. There are currently two PDE5 inhibitors used for the treatment of PAH, sildenafil and tadalafil. Both agents have been shown to improve hemodynamics and 6MWD.

Riociguate is a soluble guanylyl cyclase stimulator acting synergistically with endogenous NO, and also directly stimulating soluble guanylyl cyclase independent of NO availability. Riociguat significantly improved exercise capacity, pulmonary hemodynamics, WHO functional class, and time to clinical worsening in patients with PAH and CTEPH.

Combination Therapy

Current guidelines recommend add-on therapy targeting a different pathway when there is an inadequate clinical response or clinical deterioration with monotherapy. Combination therapy has a number of hypothetical advantages. As multiple pathogenic pathways are identified, and the neoplastic nature of PAH is increasingly recognized, this approach may provide increased benefit. This same approach has been used successfully in other complex diseases such as HIV, cancer, and heart failure. Using combination therapy could provide synergy and added benefits, while at the same time allowing for lower doses and potentially decreased side effects. Data are increasingly supportive of initial combination therapy at the time of diagnosis.

A retrospective study from the French PAH network reported benefits of upfront triple combination therapy targeting the three currently available therapeutic pathways. They reviewed 19 patients with severe PAH who were initiated on bosentan, sildenafil, and intravenous epoprostenol therapy simultaneously at the time of diagnosis. Of the 19 patients, 18 had a significant improvement in walking distance and hemodynamics at 4 months compared with baseline. The beneficial effects were sustained to final follow-up evaluation for more than 2 years. More importantly, the 1-, 2-, and 3-year survival rates were 100%. Although these results are encouraging, the findings from this single-center, retrospective analysis need further validation.

The AMBITION trial compared de novo combination therapy to monotherapy in newly diagnosed patients. Patients were randomized to a combination of ambrisentan and tadalafil, ambrisentan monotherapy, or tadalafil monotherapy. Upfront combination therapy with ambrisentan and tadalafil was associated with a 50% lower risk of clinical worsening (composite of death, lung transplantation, hospitalization for PAH worsening, and worsening PAH) when compared

with the monotherapy groups. This difference was driven primarily by the delay in time to first hospitalization. Importantly, initial combination therapy was not associated with an increase in adverse events.

Unmet and Future Research Needs in Pulmonary Hypertension

Presently there are only three classes of therapy for patients with PAH and, even with therapy, the median survival for a person with PAH is only 5–6 years **(Table 277-2)**. While there are five subtypes of PH, currently approved therapies only address one subtype. Not only do we need to expand the treatment options for patients with PAH, but we also need to develop effective therapies for all patients with PH. Limited survival is, in part, a result of delay in diagnosis. Improved awareness among clinicians and patients could lead to more timely diagnosis that will affect the response to therapy and survival. PH needs to be diagnosed in a timely manner so that therapy can be initiated as soon as possible. Patients should also have the option of referral to a specialty center that focuses on treatment of patients with pulmonary vascular disease, which will ensure their access to state-of-the-art care and a multidisciplinary approach to care. Finally, there needs to be continued efforts at developing new therapies that target the increasingly complex and overlapping pathways involved in the various forms of PH.

TABLE 277-2 FDA-Approved Therapies for the Treatment of Pulmonary Arterial Hypertension (PAH)

Generic Name	Route of Administration	Drug Class	Indication
Epoprostenol	IV	Prostacyclin derivative	Treatment of PAH to improve exercise capacity
Iloprost	Inhaled	Prostacyclin derivative	Treatment of PAH to improve a composite endpoint consisting of exercise tolerance, symptoms (NYHA Class), and lack of deterioration
Treprostinil	IV or SC	Prostacyclin derivative	Treatment of PAH to diminish symptoms associated with exercise
Treprostinil	Inhaled	Prostacyclin derivative	Treatment of PAH to improve exercise ability
Treprostinil	Oral	Prostacyclin derivative	Treatment of PAH to improve exercise ability
Selexipeg	Oral	Selective IP receptor agonist	Treatment of PAH to improve a composite endpoint lack of clinical deterioration
Bosentan	Oral	Endothelin receptor antagonist	Treatment of PAH to improve exercise capacity and to decrease clinical worsening
Ambrisentan	Oral	Endothelin receptor antagonist	Treatment of PAH to improve exercise capacity and delay clinical worsening
Macitentan	Oral	Endothelin receptor antagonist	Treatment of PAH to improve a composite endpoint of delay of clinical worsening
Sildenafil	Oral	PDE5 inhibitor	Treatment of PAH to improve exercise capacity and delay clinical worsening
Tadalafil	Oral	PDE5 inhibitor	Treatment of PAH to improve exercise ability
Riociguat	Oral	Soluble guanylyl cyclase stimulator	Treatment of PAH to improve exercise ability

Abbreviations: IV, intravenous; NYHA, New York Heart Association; PAH, pulmonary arterial hypertension; PDE5, phosphodiesterase-5; SC, subcutaneous.

FURTHER READING

Bossone E et al: Echocardiography in pulmonary arterial hypertension: From diagnosis to prognosis. J Am Soc Echocardiogr 26:1, 2013. [PubMed: 23140849]
Galie N et al: Initial use of ambrisentan plus tadalafil in pulmonary arterial hypertension. N Engl J Med 373:834, 2015. [PubMed: 26308684]
Ghofrani HA et al: Riociguat for the treatment of pulmonary arterial hypertension. N Engl J Med 369:330, 2013. [PubMed: 23883378]

Huetsch JC et al: Update on novel targets and potential treatment avenues in pulmonary hypertension. Am J Physiol Lung Cell Mol Physiol 311:L811, 2016. [PubMed: 27591245]
Pietra GG et al: Pathologic assessment of vasculopathies in pulmonary hypertension. J Am Coll Cardiol 43(12 Suppl S):25S, 2004. [PubMed: 15194175]
Rich JD, Rich S: Clinical diagnosis of pulmonary hypertension. Circulation 130:1820, 2014. [PubMed: 25385937]
Simonneau G et al: Updated clinical classification of pulmonary hypertension. J Am Coll Cardiol 62(25 Suppl):D34, 2013. [PubMed: 24355639]
Sitbon O et al: Selexipag for the treatment of pulmonary arterial hypertension. N Engl J Med 373:2522, 2015. [PubMed: 26699168]
Soubrier F et al: Genetics and genomics of pulmonary arterial hypertension. J Am Coll Cardiol 62(25 Suppl):D13, 2013. [PubMed: 24355637]
Waxman AB, Farber HW: Using clinical trial end points to risk stratify patients with pulmonary arterial hypertension. Circulation 132:2152, 2015. [PubMed: 26621638]

Chapter 278: Approach to the Patient with Disease of the Respiratory System

Patricia A. Kritek; Bruce D. Levy

INTRODUCTION

The majority of diseases of the respiratory system present with cough and/or dyspnea and fall into one of three major categories: (1) obstructive lung diseases; (2) restrictive disorders; and (3) abnormalities of the vasculature. Obstructive lung diseases are most common and primarily disorders of the airways, such as asthma, chronic obstructive pulmonary disease (COPD), bronchiectasis, and bronchiolitis. Diseases resulting in restrictive pathophysiology include parenchymal lung diseases, abnormalities of the chest wall and pleura, and neuromuscular disease. Pulmonary embolism, pulmonary hypertension, and pulmonary venoocclusive disease are all disorders of the pulmonary vasculature. Although many specific diseases fall into these major categories, both infective and neoplastic processes can affect the respiratory system and result in myriad pathologic findings, including those listed in the three categories above **(Table 278-1)**.

TABLE 278-1 Categories of Respiratory Disease

Category	Examples
Obstructive lung disease	Asthma
	Chronic obstructive pulmonary disease (COPD)
	Bronchiectasis
	Bronchiolitis
Restrictive pathophysiology—parenchymal disease	Idiopathic pulmonary fibrosis (IPF)
	Asbestosis
	Desquamative interstitial pneumonitis (DIP)
	Sarcoidosis
Restrictive pathophysiology—neuromuscular weakness	Amyotrophic lateral sclerosis (ALS)
	Guillain-Barré syndrome
	Myasthenia gravis
Restrictive pathophysiology—chest wall/pleural disease	Kyphoscoliosis
	Ankylosing spondylitis
	Chronic pleural effusions
Pulmonary vascular disease	Pulmonary embolism
	Pulmonary arterial hypertension (PAH)
	Pulmonary venoocclusive disease
	Vasculitis
Malignancy	Bronchogenic carcinoma (non-small-cell and small-cell lung cancer)
	Metastatic disease
Infectious diseases	Pneumonia
	Bronchitis
	Tracheitis

Disorders can also be grouped according to gas exchange abnormalities, including hypoxemic, hypercarbic, or combined impairment; however, many respiratory disorders do not manifest as gas exchange abnormalities.

As with the evaluation of most patients, the approach to a patient with a respiratory system disorder begins with a thorough history and a focused physical examination. Many patients will subsequently undergo pulmonary function testing, chest imaging, blood and sputum analysis, a variety of serologic or microbiologic studies, and diagnostic procedures, such as bronchoscopy. This stepwise approach is discussed in detail below.

HISTORY

Dyspnea and Cough

The cardinal symptoms of respiratory disease are dyspnea and cough **(Chaps. 33 and 34)**. Dyspnea has many causes, some of which are not predominantly due to lung pathology. The words a patient uses to describe shortness of breath can suggest certain etiologies for dyspnea. Patients with obstructive lung disease often complain of “chest tightness” or “inability to get a deep breath,” whereas patients with congestive heart failure more commonly report “air hunger” or a sense of suffocation.

The tempo of onset and the duration of a patient’s dyspnea are likewise helpful in determining the etiology. Acute shortness of breath is usually associated with sudden physiological changes, such as laryngeal edema, bronchospasm, myocardial infarction, pulmonary embolism, or pneumothorax. Patients with COPD and idiopathic pulmonary fibrosis (IPF) experience a gradual progression of dyspnea on exertion, punctuated by acute exacerbations of shortness of breath. In contrast, most asthmatics do not have daily symptoms, but experience intermittent episodes of dyspnea, cough, and chest tightness that are usually associated with specific triggers, such as an upper respiratory tract infection or exposure to allergens.

Specific questioning should focus on factors that incite dyspnea as well as on any intervention that helps resolve the patient’s shortness of breath. Asthma is commonly exacerbated by specific triggers, although this can also be true of COPD. Many patients with lung disease report dyspnea on exertion. Determining the degree of activity that results in shortness of breath gives the clinician a gauge of the patient’s degree of disability. Many patients adapt their level of activity to accommodate progressive limitation. For this reason, it is important, particularly in older patients, to delineate the activities in which they engage and how these activities have changed over time. Dyspnea on exertion is often an early symptom of underlying lung or heart disease and warrants a thorough evaluation.

For cough, the clinician should inquire about the duration of the cough, whether or not it is associated with sputum production, and any specific triggers that induce it. Acute cough productive of phlegm is often a symptom of infection of the respiratory system, including processes affecting the upper airway (e.g., sinusitis, tracheitis), the lower airways (e.g., bronchitis, bronchiectasis), and the lung parenchyma (e.g., pneumonia). Both the quantity and quality of the sputum, including whether it is blood-streaked or frankly bloody, should be determined. Hemoptysis warrants urgent evaluation as delineated in **Chap. 35**.

Chronic cough (defined as that persisting for >8 weeks) is commonly associated with obstructive lung diseases, particularly asthma, COPD and chronic bronchiectasis, as well as “nonrespiratory” diseases, such as gastroesophageal reflux and postnasal drip. Diffuse parenchymal lung diseases, including IPF, frequently present as a persistent, nonproductive cough. All causes of cough are not respiratory in origin, and assessment should encompass a broad differential, including cardiac and gastrointestinal diseases as well as psychogenic causes.

Additional Symptoms

Patients with respiratory disease may report wheezing, which is suggestive of airways disease, particularly asthma. Hemoptysis can be a symptom of a variety of lung diseases, including infections of the respiratory tract, bronchogenic carcinoma, and pulmonary embolism. In addition, chest pain or discomfort can be respiratory in origin. As the lung parenchyma is not innervated with pain fibers, pain in the chest from respiratory disorders usually results from either diseases of the parietal pleura (e.g., pneumothorax) or pulmonary vascular diseases (e.g., pulmonary hypertension). As many diseases of the lung can result in strain on the right side of the heart, patients may also present with symptoms of cor pulmonale, including abdominal bloating or distention and pedal edema **(Chap. 252)**.

Additional History

A thorough social history is an essential component of the evaluation of patients with respiratory disease. All patients should be asked about current or previous cigarette smoking, as this exposure is associated with many diseases of the respiratory system, including COPD, bronchogenic lung cancer, and select parenchymal lung diseases (e.g., desquamative interstitial pneumonitis and pulmonary Langerhans cell histiocytosis). For most of these disorders, increased cigarette smoke exposure (i.e., cigarette pack-years) increases the risk of disease. “Secondhand smoke” also increases risk for some respiratory disorders, so patients should also be asked about parents, spouses, or housemates who smoke. Possible inhalational exposures at work (e.g., asbestos, silica) or home (e.g., wood smoke, excrement from pet birds) should be explored **(Chap. 283)**. Travel predisposes to certain infections of the respiratory tract, most notably tuberculosis. Potential exposure to fungi is increased in specific geographic regions or climates (e.g., *Histoplasma capsulatum*), so exposures to these regions should be determined.

Associated symptoms of fever and chills should raise the suspicion of infective etiologies, both pulmonary and systemic. A comprehensive review of systems may suggest rheumatologic or autoimmune disease presenting with respiratory tract manifestations. Questions should focus on joint pain or swelling, rashes, dry eyes, dry mouth, or constitutional symptoms. In addition, carcinomas from a variety of primary sources commonly metastasize to the lung and cause respiratory symptoms. Finally, therapy for other conditions, including both irradiation and medications, can result in diseases of the chest.

Physical Examination

The clinician’s suspicion of respiratory disease often begins with patient’s vital signs. The respiratory rate is informative, whether elevated (tachypnea) or depressed (hypopnea). In addition, pulse oximetry should be measured, as many patients with respiratory disease have hypoxemia, either at rest or with exertion.

The first step of the physical examination is inspection. Patients with respiratory disease may be in distress, using accessory muscles of respiration to breathe. Severe kyphoscoliosis can result in restrictive pathophysiology. Inability to complete a sentence in conversation is generally a sign of severe impairment and should result in an expedited evaluation of the patient.

Percussion of the chest is used to establish diaphragm excursion and lung size. In the setting of decreased breath sounds, percussion is used to distinguish between pleural effusions (dull to percussion) and pneumothorax (hyper-resonant note).

The role of palpation is limited in the respiratory examination. Palpation can demonstrate subcutaneous air in the setting of barotrauma. It can also be used as an adjunctive assessment to determine whether an area of decreased breath sounds is due to consolidation (increased tactile fremitus) or a pleural effusion (decreased tactile fremitus). To detect unilateral disorders of ventilation, the symmetry and degree of chest wall expansion can be assessed during a deep inspiration by placing one’s thumbs together at the midline over the lower posterior chest while grasping the lateral rib cage.

The majority of the manifestations of respiratory disease present as abnormalities of auscultation. Wheezes are a manifestation of airway obstruction. While most commonly a sign of asthma, peribronchial edema in the setting of congestive heart failure can also result in diffuse wheezes, as can any other process that causes narrowing of small airways. Wheezes can be polyphonic, involving multiple different size airways (e.g., asthma) or monophonic, involving one size airway (e.g., bronchogenic carcinoma). For these reasons, clinicians must take care not to attribute all wheezing to asthma.

Rhonchi are a manifestation of obstruction of medium-sized airways, most often with secretions. In the acute setting, this manifestation may be a sign of viral or bacterial bronchitis. Chronic rhonchi suggest bronchiectasis or COPD. In contrast to expiratory wheezes and rhonchi, stridor is a high-pitched, focal inspiratory wheeze, usually heard over the neck as a manifestation of upper airway obstruction.

Crackles, or rales, are commonly a sign of alveolar disease. Processes that fill the alveoli with fluid may result in crackles, including pulmonary edema and pneumonia. Crackles in pulmonary edema are generally more prominent at the bases. Interestingly, diseases that result in fibrosis of the interstitium (e.g., IPF) also result in crackles that sound like Velcro being ripped apart. Although some clinicians make a distinction between “wet” and “dry” crackles, this distinction has not been shown to be a reliable way to differentiate among etiologies of respiratory disease.

One way to help distinguish between crackles associated with alveolar fluid and those associated with interstitial fibrosis is to assess for egophony. *Egophony* is the auscultation of the sound “AH” instead of “EEE” when a patient phonates “EEE.” This change in note is due to abnormal sound transmission through consolidated parenchyma and is present in pneumonia but not in IPF. Similarly, areas of alveolar filling have increased whispered *pectoriloquy* as well as transmission of larger-airway sounds (i.e., bronchial breath sounds in a lung zone where vesicular breath sounds are expected).

The lack or diminution of breath sounds can also help determine the etiology of respiratory disease. Patients with emphysema often have a quiet chest with diffusely decreased breath sounds. A pneumothorax or pleural effusion may present with an area of absent breath sounds.

Other Systems

Pedal edema, if symmetric, may suggest cor pulmonale; if asymmetric, it may be due to deep venous thrombosis and associated pulmonary embolism. Jugular venous distention may also be a sign of volume overload associated with right heart failure. *Pulsus paradoxus* is an ominous sign in a patient with obstructive lung disease, as it is associated with significant negative intrathoracic (pleural) pressures required for ventilation and impending respiratory failure.

As stated earlier, rheumatologic disease may manifest primarily as lung disease. Owing to this association, particular attention should be paid to joint and skin examination. Clubbing can be found in many lung diseases, including cystic fibrosis, IPF, and lung cancer. Cyanosis is seen in hypoxemic respiratory disorders that result in >5 g of deoxygenated hemoglobin/dL.

DIAGNOSTIC EVALUATION

The sequence of studies is dictated by the clinician’s differential diagnosis, as determined by the history and physical examination. Acute respiratory symptoms are often evaluated with multiple tests performed at the same time in order to diagnose any life-threatening diseases rapidly (e.g., pulmonary embolism or multilobar pneumonia). In contrast, chronic dyspnea and cough can be evaluated in a more protracted, stepwise fashion.

Pulmonary Function Testing

The initial pulmonary function test obtained is spirometry (See also Chap. 280). This study is an effort-dependent test used to assess for obstructive pathophysiology as seen in asthma, COPD, and bronchiectasis. A diminished-forced expiratory volume in 1 s (FEV_1)/forced vital capacity (FVC) (often defined as <70%) is diagnostic of obstruction. In addition to measuring FEV_1 and FVC, the clinician should examine the flow-volume loop (which is less effort-dependent). A plateau of the inspiratory and expiratory curves suggests large-airway obstruction in extrathoracic and intrathoracic locations, respectively.

Spirometry with symmetric decreases in FEV_1 and FVC warrants further testing, including measurement of lung volumes and the diffusion capacity of the lung for carbon monoxide (DL_{CO}). A total lung capacity <80% of the patient's predicted value defines restrictive pathophysiology. Restriction can result from parenchymal disease, neuromuscular weakness, or chest wall or pleural diseases (Table 278-1). Restriction with impaired gas exchange, as indicated by a decreased DL_{CO}, suggests parenchymal lung disease. Additional testing, such as measurements of maximal inspiratory and expiratory pressures, can help diagnose neuromuscular weakness. Normal spirometry, normal lung volumes, and a low DL_{CO} should prompt further evaluation for pulmonary vascular disease.

Arterial blood gas testing is often helpful in assessing respiratory disease. Hypoxemia, while usually apparent with pulse oximetry, can be further evaluated with the measurement of arterial PO_2 and the calculation of an alveolar gas and arterial blood oxygen tension difference ($[A–a]DO_2$). Patients with diseases that cause ventilation-perfusion mismatch or shunt physiology have an increased $(A–a)DO_2$ at rest. Arterial blood gas testing also allows the measurement of arterial PCO_2. Hypercarbia can accompany disorders of ventilation, as seen in severe airway obstruction (e.g., COPD) or progressive restrictive physiology.

Chest Imaging

Most patients with disease of the respiratory system undergo imaging of the chest as part of the initial evaluation (See Chap. A12). Clinicians should generally begin with ultrasound of the chest or a plain chest radiograph, preferably posterior-anterior and lateral films. Ultrasound is often readily available and can help rapidly diagnose pneumothorax, pleural effusion, and consolidation of lung parenchyma. Chest radiographs give additional detail and can reveal findings including opacities of the parenchyma, blunting of the costophrenic angles, mass lesions, and volume loss. However, many diseases of the respiratory system, particularly those of the airways and pulmonary vasculature, are associated with a normal chest radiograph.

CT scan of the chest can also be useful to delineate parenchymal processes, pleural disease, masses or nodules, and large airways. If the test includes administration of contrast, the pulmonary vasculature can be assessed with particular utility for determination of pulmonary emboli. Intravenous contrast also allows lymph nodes to be examined in greater detail. When coupled with positron emission testing (PET), lesions of the chest can be assessed for metabolic activity; helping differentiate between malignancy and scar.

FURTHER STUDIES

Depending on the clinician's suspicion, a variety of other studies may be done. Concern about large-airway lesions may warrant bronchoscopy. This procedure may also be used to sample the alveolar space with bronchoalveolar lavage or to obtain nonsurgical lung biopsies. Blood testing may include assessment for hypercoagulable states in the setting of pulmonary vascular disease, serologic testing for infectious or rheumatologic disease, or assessment of inflammatory markers or leukocyte counts (e.g., eosinophils). Genetic testing is increasingly used for heritable lung diseases such as cystic fibrosis. Sputum evaluation for malignant cells or microorganisms may be appropriate. An echocardiogram to assess right- and left-sided heart function is often obtained. Finally, at times, a surgical lung biopsy is needed to diagnose certain diseases of the respiratory system. All of these studies will be guided by the preceding history, physical examination, pulmonary function testing, and chest imaging.

FURTHER READING

Achilleos A: Evidence-based evaluation and management of chronic cough. Med Clin North Am 100:1033, 2016. [PubMed: 27542423]
Bohadana A et al: Fundamentals of lung auscultation. N Engl J Med 370:744, 2014. [PubMed: 24552321]
Koenig SJ et al: Thoracic ultrasonography for the pulmonary specialist. Chest 140:1332, 2011. [PubMed: 22045878]
Parshall MB et al: An official American Thoracic Society statement: Update on the mechanisms, assessment, and management of dyspnea. Am J Respir Crit Care Med 185:435, 2012. [PubMed: 22336677]
Pellegrino R et al: Interpretive strategies for lung function tests. Eur Respir J 26:948, 2005. [PubMed: 16264058]

Chapter 279: Disturbances of Respiratory Function

Edward T. Naureckas; Julian Solway

INTRODUCTION

The primary functions of the respiratory system—to oxygenate blood and eliminate carbon dioxide—require virtual contact between blood and fresh air, which facilitates diffusion of respiratory gases between blood and gas. This process occurs in the lung alveoli, where blood flowing through alveolar wall capillaries is separated from alveolar gas by an extremely thin membrane of flattened endothelial and epithelial cells, across which respiratory gases diffuse and equilibrate. Blood flow through the lung is unidirectional via a continuous vascular path along which venous blood absorbs oxygen from and loses CO_2 to inspired gas. The path for airflow, in contrast, reaches a dead end at the alveolar walls; thus the alveolar space must be ventilated tidally, with inflow of fresh gas and outflow of alveolar gas alternating periodically at the respiratory rate (RR). To provide an enormous alveolar surface area (typically 70 m^2) for blood-gas diffusion within the modest volume of a thoracic cavity (typically 7 L), nature has distributed both blood flow and ventilation among millions of tiny alveoli through multigenerational branching of both pulmonary arteries and bronchial airways. As a consequence of variations in tube lengths and calibers along these pathways as well as the effects of gravity, tidal pressure fluctuations, and anatomic constraints from the chest wall, the alveoli vary in their relative ventilations and perfusions. Not surprisingly, for the lung to be most efficient in exchanging gas, the fresh gas ventilation of a given alveolus must be matched to its perfusion.

For the respiratory system to succeed in oxygenating blood and eliminating CO_2, it must be able to ventilate the lung tidally and thus to freshen alveolar gas; it must provide for perfusion of the individual alveolus in a manner proportional to its ventilation; and it must allow adequate diffusion of respiratory gases between alveolar gas and capillary blood. Furthermore, it must accommodate several-fold increases in the demand for oxygen uptake or CO_2 elimination imposed by metabolic needs or acid-base derangement. Given these multiple requirements for normal operation, it is not surprising that many diseases disturb respiratory function. This chapter considers in some detail the physiologic determinants of lung ventilation and perfusion, elucidates how the matching distributions of these processes and rapid gas diffusion allow normal gas exchange, and discusses how common diseases derange these normal functions, thereby impairing gas exchange—or at least increasing the work required by the respiratory muscles or heart to maintain adequate respiratory function.

VENTILATION

It is useful to conceptualize the respiratory system as three independently functioning components: the lung, including its airways; the neuromuscular system; and the chest wall, which includes everything that is not lung or active neuromuscular system. Accordingly, the mass of the respiratory muscles is part of the chest wall, while the force these muscles generate is part of the neuromuscular system; the abdomen (especially an obese abdomen) and the heart (especially an enlarged heart) are, for these purposes, part of the chest wall. Each of these three components has mechanical properties that relate to its enclosed volume (or—in the case of the neuromuscular system—the respiratory system volume at which it is operating) and to the rate of change of its volume (i.e., flow).

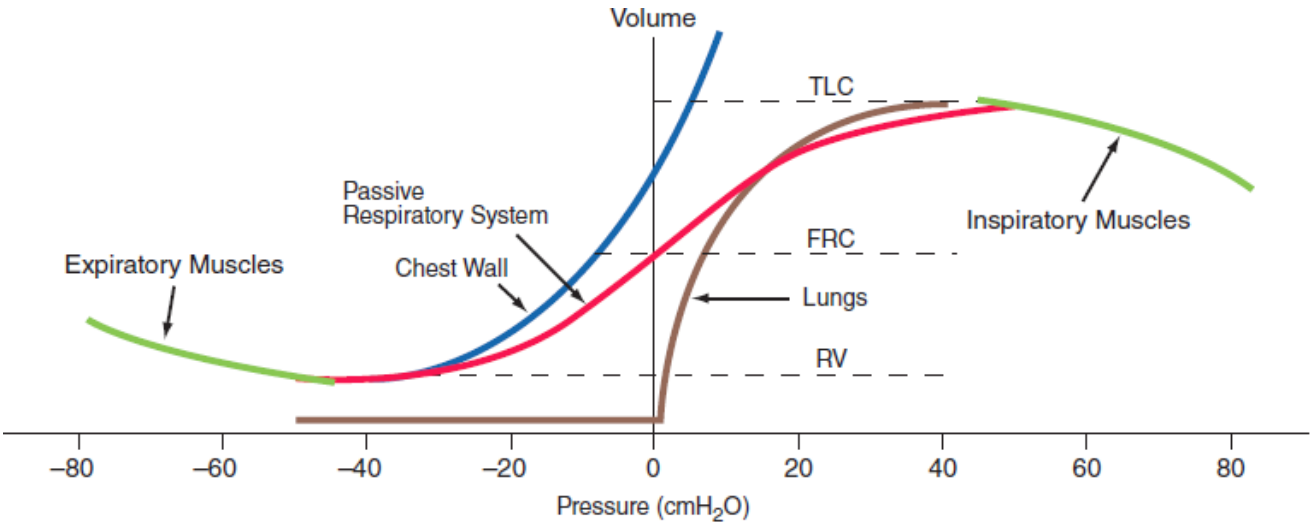
Volume-Related Mechanical Properties—Statics

Figure 279-1 shows the volume-related properties of each component of the respiratory system. Because of both surface tension at the air-liquid interface between alveolar wall lining fluid and alveolar gas and elastic recoil of the lung tissue itself, the lung requires a positive transmural pressure difference between alveolar gas and its pleural surface to stay inflated; this difference is called the *elastic recoil pressure* of the lung, and it increases with lung volume. The lung becomes rather stiff at high volumes, so that relatively small volume changes are accompanied by large changes in transpulmonary pressure; in contrast, the lung is compliant at lower volumes, including those at which tidal breathing normally occurs. At zero inflation pressure, even normal lungs retain some air in the alveoli. Because the small peripheral airways are tethered open by outward radial pull from inflated lung parenchyma attached to adventitia, as the

lung deflates during exhalation, those small airways are pulled open progressively less, and eventually close, trapping some gas in the alveoli. This effect can be exaggerated with age and especially with obstructive airway diseases, resulting in gas trapping at quite large lung volumes.

FIGURE 279-1

Pressure-volume curves of the isolated lung, isolated chest wall, combined respiratory system, inspiratory muscles, and expiratory muscles. FRC, functional residual capacity; RV, residual volume; TLC, total lung capacity.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

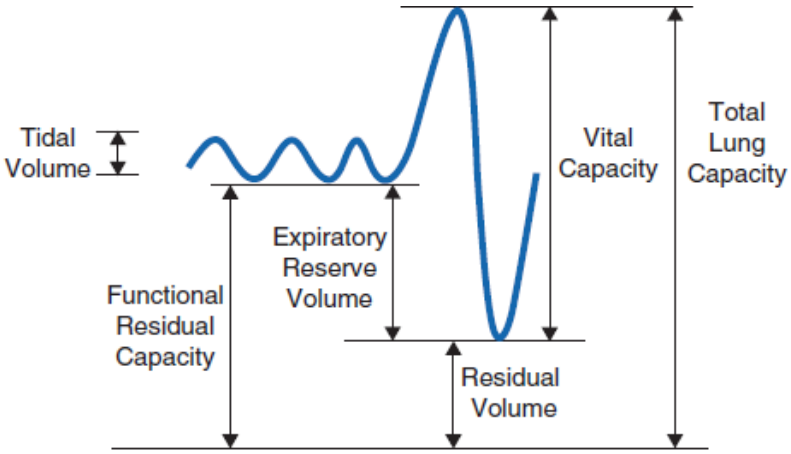
The elastic behavior of the passive chest wall (i.e., in the absence of neuromuscular activation) differs markedly from that of the lung. Whereas the lung tends toward full deflation with no distending (transmural) pressure, the chest wall encloses a large volume when pleural pressure equals body surface (atmospheric) pressure. Furthermore, the chest wall is compliant at high enclosed volumes, readily expanding even further in response to increases in transmural pressure. The chest wall also remains compliant at small negative transmural pressures (i.e., when pleural pressure falls slightly below atmospheric pressure), but as the volume enclosed by the chest wall becomes quite small in response to large negative transmural pressures, the passive chest wall becomes stiff due to squeezing together of ribs and intercostal muscles, diaphragm stretch, displacement of abdominal contents, and straining of ligaments and bony articulations. Under normal circumstances, the lung and the passive chest wall enclose essentially the same volume, the only difference being the volumes of the pleural fluid and of the lung parenchyma (normally both quite small in the absence of disease). For this reason and because the lung and chest wall function in mechanical series, the pressure required to displace the passive respiratory system (lungs plus chest wall) at any volume is simply the sum of the elastic recoil pressure of the lungs and the transmural pressure across the chest wall. When plotted against respiratory system volume, this relationship assumes a sigmoid shape, exhibiting stiffness at high lung volumes (imparted by the lung), stiffness at low lung volumes (imparted by the chest wall or sometimes by airway closure), and compliance in the middle range of lung volumes where normal tidal breathing occurs. In addition, a passive resting point of the respiratory system is attained when alveolar gas pressure equals body surface pressure (i.e., when the transrespiratory system pressure is zero). At this volume (called the *functional residual capacity* [FRC]), the outward recoil of the chest wall is balanced exactly by the inward recoil of the lung. As these recoils are transmitted through the pleural fluid, the lung is pulled both outward and inward simultaneously at FRC, and thus its pressure falls below atmospheric pressure (typically, −5 cmH_2O).

The normal passive respiratory system would equilibrate at the FRC and remain there were it not for the actions of the respiratory muscles. The inspiratory muscles act on the chest wall to generate the equivalent of positive pressure across the lungs and passive chest wall, while the expiratory muscles generate the equivalent of negative transrespiratory pressure. The maximal pressures these sets of muscles can generate vary with the lung volume at which they operate. This variation is due to length-tension relationships in striated muscle sarcomeres and to changes in mechanical advantage as the angles of insertion change with lung volume (Fig. 279-1). Nonetheless, under normal conditions, the respiratory muscles are substantially "overpowered" for their roles and generate more than adequate force to drive the respiratory system to its stiffness extremes, as determined by the lung (total lung capacity [TLC]) or by chest wall or airway closure (residual volume [RV]); the airway closure always prevents the adult lung from emptying completely under normal circumstances. The excursion between full and minimal lung inflation is called *vital capacity* (VC; **Fig. 279-2**) and is readily seen to be the difference between volumes at two unrelated stiffness extremes—one

determined by the lung (TLC) and the other by the chest wall or airways (RV). Thus, although VC is easy to measure (see below), it provides little information about the intrinsic properties of the respiratory system. As will become clear, it is much more useful for the clinician to consider TLC and RV individually.

FIGURE 279-2

Spirogram demonstrating a slow vital capacity maneuver and various lung volumes.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Flow-Related Mechanical Properties—Dynamics

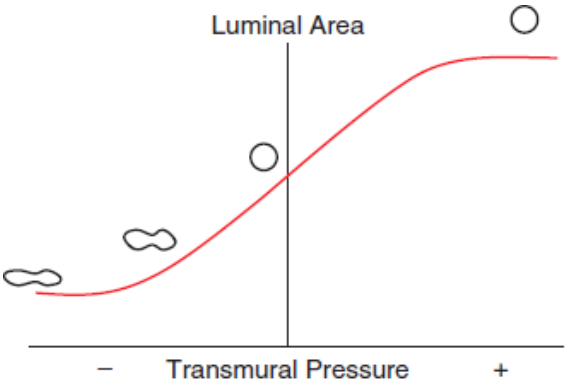
The passive chest wall and active neuromuscular system both exhibit mechanical behaviors related to the rate of change of volume, but these behaviors become quantitatively important only at markedly supraphysiologic breathing frequencies (e.g., during high-frequency mechanical ventilation), and thus will not be addressed here. In contrast, the dynamic airflow properties of the lung substantially affect its ability to ventilate and contribute importantly to the work of breathing, and these properties are often deranged by disease. Understanding dynamic airflow properties is, therefore, worthwhile.

As with the flow of any fluid (gas or liquid) in any tube, maintenance of airflow within the pulmonary airways requires a pressure gradient that falls along the direction of flow, the magnitude of which is determined by the flow rate and the frictional resistance to flow. During quiet tidal breathing, the pressure gradients driving inspiratory or expiratory flow are small owing to the very low frictional resistance of normal pulmonary airways (R_{aw}, normally <2 $cmH_2O/L/s$). However, during rapid exhalation, another phenomenon reduces flow below that which would have been expected if frictional resistance were the only impediment to flow. This phenomenon is called *dynamic airflow limitation*, and it occurs because the bronchial airways through which air is exhaled are collapsible rather than rigid **(Fig. 279-3)**. An important anatomic feature of the pulmonary airways is its treelike branching structure. While the individual airways in each successive generation, from most proximal (trachea) to most distal (respiratory bronchioles), are smaller than those of the parent generation, their number increases exponentially such that the summed cross-sectional area of the airways becomes very large toward the lung periphery. Because flow (volume/time) is constant along the airway tree, the velocity of airflow (flow/summed cross-sectional area) is much greater in the central airways than in the peripheral airways. During exhalation, gas leaving the alveoli must, therefore, gain velocity as it proceeds toward the mouth. The energy required for this “convective” acceleration is drawn from the component of gas energy manifested as its local pressure, which reduces intraluminal gas pressure, airway transmural pressure, airway size (Fig. 279-3), and flow. This phenomenon is the Bernoulli effect, the same effect that keeps an airplane airborne, generating a lifting force by decreasing pressure above the curved upper surface of the wing due to acceleration of air flowing over the wing. If an individual tries to exhale more forcefully, the local velocity increases further and reduces airway size further, resulting in no net increase in flow. Under these circumstances, flow has reached its maximum possible value, or its *flow limit*. Lungs normally exhibit such dynamic airflow limitation. This limitation can be assessed by spirometry, in which an individual inhales fully to TLC and then forcibly exhales to RV. One useful spirometric measure is the volume of air exhaled during the forced expiratory volume 1 s (FEV_1), as discussed later. Maximal expiratory flow at any lung volume is determined by gas density, airway cross-section and distensibility, elastic recoil pressure of the lung, and frictional pressure loss to the flow-limiting airway site. Under normal conditions, maximal expiratory flow falls with lung volume **(Fig. 279-4)**, primarily because of the dependence of lung recoil pressure on lung volume (Fig. 279-1). In pulmonary

fibrosis, lung recoil pressure is increased at any lung volume, and thus the maximal expiratory flow is elevated when considered in relation to lung volume. Conversely, in emphysema, lung recoil pressure is reduced; this reduction is a principal mechanism by which maximal expiratory flows fall. Diseases that narrow the airway lumen at any transmural pressure (e.g., asthma or chronic bronchitis) or that cause excessive airway collapsibility (e.g., tracheomalacia) also reduce maximal expiratory flow.

FIGURE 279-3

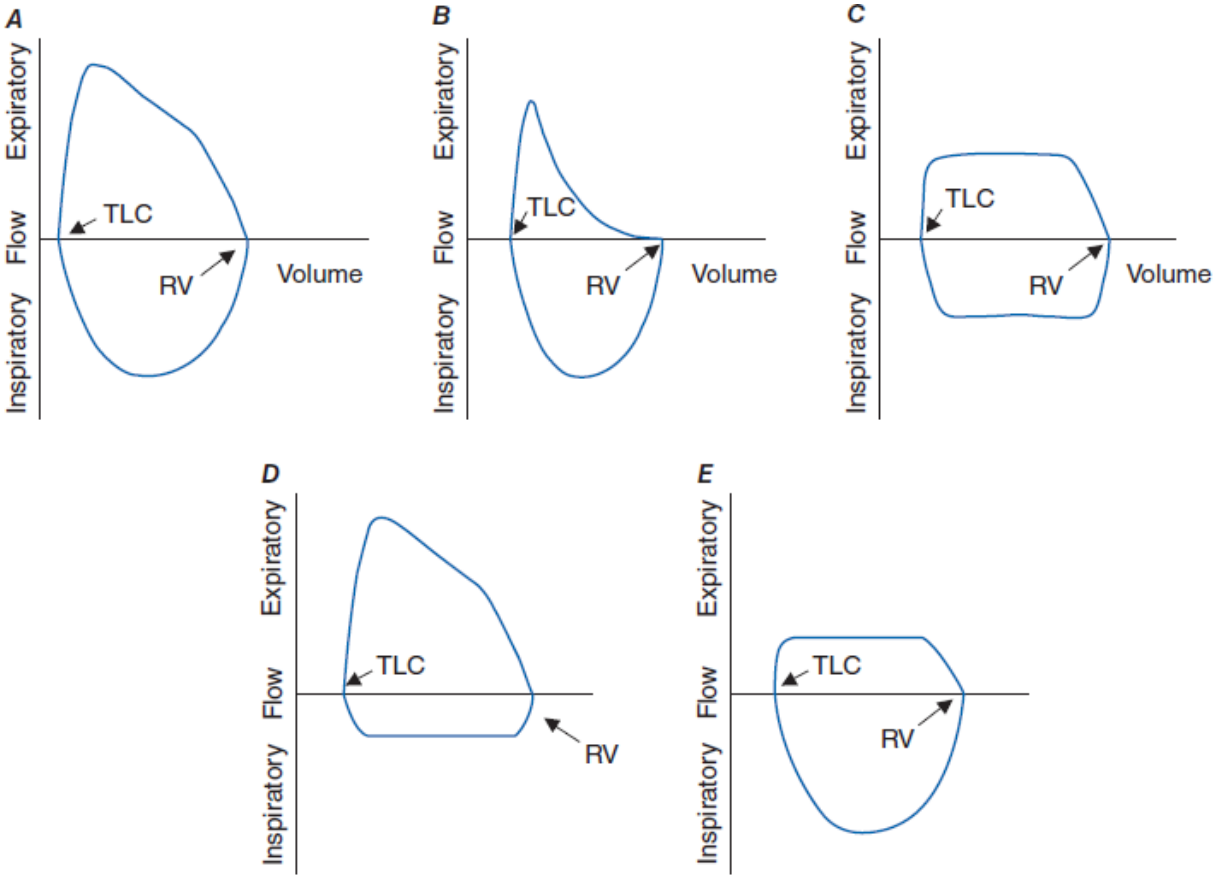
Luminal area versus transmural pressure relationship. Transmural pressure represents the pressure difference across the airway wall from inside to outside.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 279-4

Flow-volume loops. ***A.*** Normal. ***B.*** Airflow obstruction. ***C.*** Fixed central airway obstruction (either above or below the thoracic inlet). ***D.*** Variable upper airway obstruction (above the thoracic inlet) ***E.*** Variable lower airway obstruction (below the thoracic inlet); RV, residual volume; TLC, total lung capacity.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The Bernoulli effect also applies during inspiration, but the more negative pleural pressures during inspiration lower the pressure outside of the airways, thereby increasing transmural pressure and promoting airway expansion. Thus, inspiratory airflow limitation seldom occurs due to diffuse pulmonary airway disease. Conversely, extrathoracic airway narrowing (e.g., due to a tracheal adenoma or post-tracheostomy stricture) can lead to inspiratory airflow limitation (Fig. 279-4).

The Work of Breathing

In health, the elastic (volume change-related) and dynamic (flow-related) loads that must be overcome to ventilate the lungs at rest are small, and the work required of the respiratory muscles is minimal. However, the work of breathing can increase considerably due to a metabolic requirement for substantially increased ventilation, an abnormally increased mechanical load, or both. As discussed below, the rate of ventilation is primarily set by the need to eliminate carbon dioxide, and thus ventilation increases during exercise (sometimes by >20-fold) and during metabolic acidosis as a compensatory response. Naturally, the work rate required to overcome the elasticity of the respiratory system increases with both the depth and the frequency of tidal breaths, while the work required to overcome the dynamic load increases with total ventilation. A modest increase of ventilation is most efficiently achieved by increasing tidal volume but not RR, which is the normal ventilatory response to lower-level exercise. At higher levels of exercise, deep breathing persists, but RR also increases.

The work of breathing also increases when disease reduces the compliance of the respiratory system or increases the resistance to airflow. The former occurs commonly in diseases of the lung parenchyma (interstitial processes or fibrosis, alveolar filling diseases such as pulmonary edema or pneumonia, or substantial lung resection), and the latter occurs in obstructive airway diseases such as asthma, chronic bronchitis, emphysema, and cystic fibrosis. Furthermore, severe airflow obstruction can functionally reduce the compliance of the respiratory system by leading to dynamic hyperinflation. In this scenario, expiratory flows slowed by the obstructive airways disease may be insufficient to allow complete exhalation during the expiratory phase of tidal breathing; as a result, the "functional residual capacity (FRC)" from which the next breath is inhaled is greater than the static FRC. With repetition of incomplete exhalations of each tidal breath, the operating FRC becomes dynamically elevated, sometimes to a level that approaches TLC. At these high lung volumes, the respiratory system is much less compliant than at normal breathing volumes, and thus the elastic work of each tidal breath is also increased. The dynamic pulmonary hyperinflation that accompanies severe airflow

obstruction causes patients to sense difficulty in inhaling—even though the root cause of this pathophysiologic abnormality is expiratory airflow obstruction.

Adequacy of Ventilation

As noted above, the respiratory control system that sets the rate of ventilation responds to chemical signals, including arterial CO_2 and oxygen tensions and blood pH, and to volitional needs, such as the need to inhale deeply before playing a long phrase on the trumpet. Disturbances in ventilation are discussed in **Chap. 290**. The focus of this chapter is on the relationship between ventilation of the lung and CO_2 elimination.

At the end of each tidal exhalation, the conducting airways are filled with alveolar gas that did not reach the mouth when expiratory flow stopped. During the ensuing inhalation, fresh gas immediately enters the airway tree at the mouth, but the gas first entering the alveoli at the start of inhalation is that same alveolar gas in the conducting airways that had just left the alveoli. Accordingly, fresh gas does not enter the alveoli until the volume of the conducting airways has been inspired. This volume is called the *anatomic dead space* (V_D). Quiet breathing with tidal volumes smaller than the anatomic dead space introduces no fresh gas into the alveoli at all; only that part of the inspired tidal volume (V_T) that is greater than the V_D introduces fresh gas into the alveoli. The dead space can be further increased functionally if some of the inspired tidal volume is delivered to a part of the lung that receives no pulmonary blood flow and thus cannot contribute to gas exchange (e.g., the portion of the lung distal to a large pulmonary embolus). In this situation, exhaled minute ventilation ($\dot{V}_E = V_T \times RR$) includes a component of dead space ventilation ($\dot{V}_D = V_D \times RR$) and a component of fresh gas alveolar ventilation ($\dot{V}_A = [V_T - V_D] \times RR$). CO_2 elimination from the alveoli is equal to $\dot{V}_A$ times the difference in CO_2 fraction between inspired air (essentially zero) and alveolar gas (typically ~5.6% after correction for humidification of inspired air, corresponding to 40 mmHg). In the steady state, the alveolar fraction of CO_2 is equal to metabolic CO_2 production divided by alveolar ventilation. Because, as discussed below, alveolar and arterial CO_2 tensions are equal, and because the respiratory controller normally strives to maintain arterial P_{CO_2} (Pa_{CO_2}) at ~40 mmHg, the adequacy of alveolar ventilation is reflected in Pa_{CO_2}. If the Pa_{CO_2} falls much below 40 mmHg, alveolar hyperventilation is present; if the Pa_{CO_2} exceeds 40 mmHg, alveolar hypoventilation is present. Ventilatory failure is characterized by extreme alveolar hypoventilation.

As a consequence of oxygen uptake of alveolar gas into capillary blood, alveolar oxygen tension falls below that of inspired gas. The rate of oxygen uptake (determined by the body's metabolic oxygen consumption) is related to the average rate of metabolic CO_2 production, and their ratio—the "respiratory quotient" ($R = \dot{V}_{CO_2}/\dot{V}_{O_2}$)—depends largely on the fuel being metabolized. For a typical American diet, R is usually around 0.85. Together, these phenomena allow the estimation of alveolar oxygen tension, according to the following relationship, known as the *alveolar gas equation*:

$$Pa_{O_2} = FI_{O_2} \times (P_{bar} - P_{H_2O}) - Pa_{CO_2}/R$$

The alveolar gas equation also highlights the influences of inspired oxygen fraction (FI_{O_2}), barometric pressure (P_{bar}), and vapor pressure of water (P_{H_2O} = 47 mmHg at 37°C) in addition to alveolar ventilation (which sets Pa_{CO_2}) in determining Pa_{O_2}. An implication of the alveolar gas equation is that severe arterial hypoxemia rarely occurs as a pure consequence of alveolar hypoventilation at sea level while an individual is breathing air. The potential for alveolar hypoventilation to induce severe hypoxemia with otherwise normal lungs increases as P_{bar} falls with increasing altitude.

GAS EXCHANGE

Diffusion

For oxygen to be delivered to the peripheral tissues, it must pass from alveolar gas into alveolar capillary blood by diffusing through alveolar membrane. The aggregate alveolar membrane is highly optimized for this process, with a very large surface area and minimal thickness. Diffusion through the alveolar membrane is so efficient in the human lung that in most circumstances hemoglobin of a red blood cell becomes fully oxygen saturated by the time the cell has traveled just one-third the length of the alveolar capillary. Thus, the uptake of alveolar oxygen is ordinarily limited by the amount of blood transiting the alveolar capillaries rather than by the rapidity with which oxygen can diffuse across the membrane; consequently, oxygen uptake from the lung is said to be "perfusion limited" rather than diffusion limited. CO_2 also equilibrates rapidly across the alveolar membrane. Therefore, the oxygen and CO_2 tensions in capillary blood leaving a normal alveolus are essentially equal to those in alveolar gas. Only in rare circumstances (e.g., at high altitude

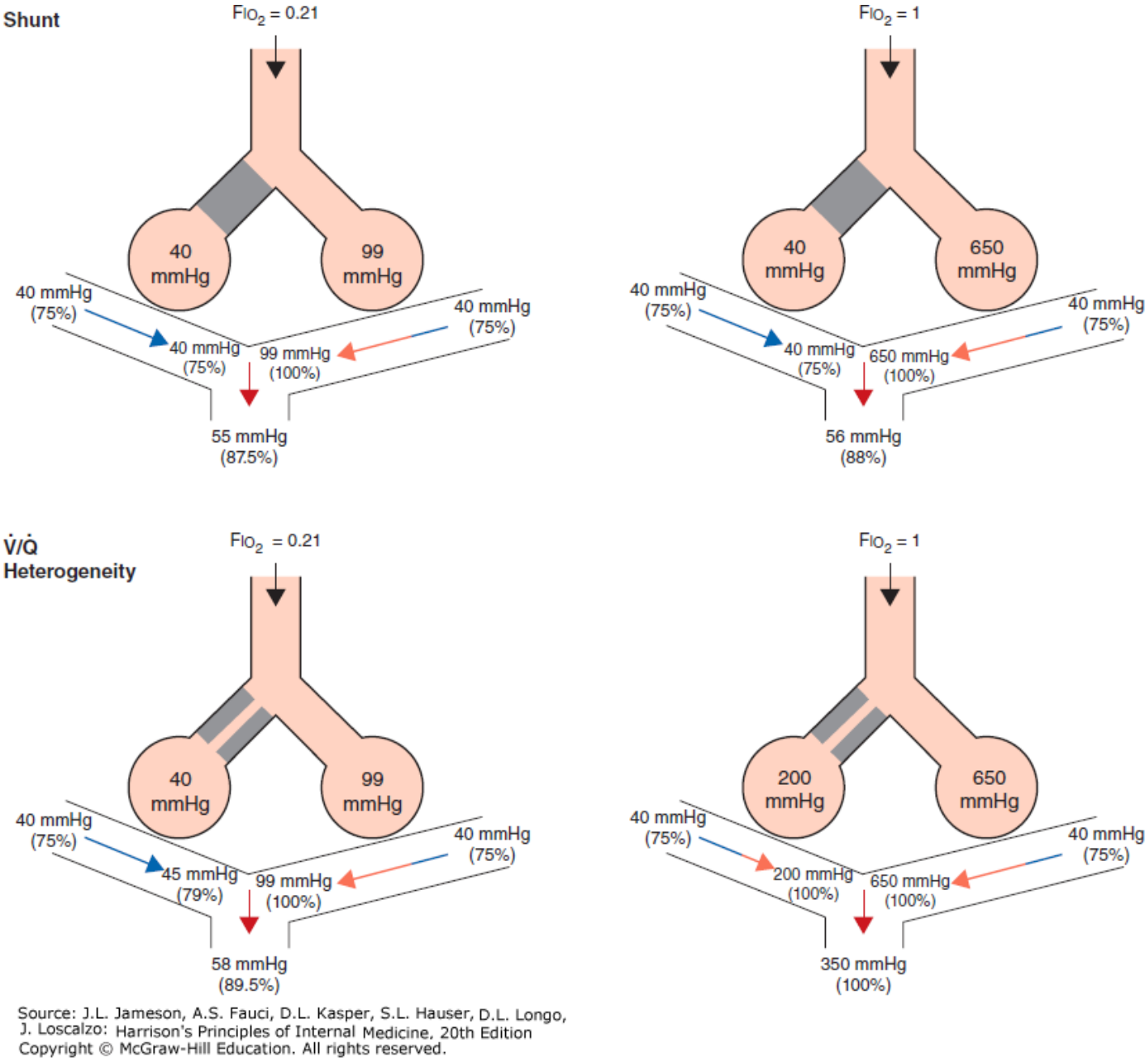
or in high-performance athletes exerting maximal effort) is oxygen uptake from normal lungs diffusion limited. Diffusion limitation can also occur in interstitial lung disease if substantially thickened alveolar walls remain perfused.

Ventilation/Perfusion Heterogeneity

As noted above, for gas exchange to be most efficient, ventilation to each individual alveolus (among the millions of alveoli) should match perfusion to its accompanying capillaries. Because of the differential effects of gravity on lung mechanics and blood flow throughout the lung and because of differences in airway and vascular architecture among various respiratory paths, there is minor ventilation/perfusion heterogeneity even in the normal lung; however, $\dot{V}/\dot{Q}$ heterogeneity can be particularly marked in disease. Two extreme examples are (1) ventilation of unperfused lung distal to a pulmonary embolus, in which ventilation of the physiologic dead space is “wasted” in the sense that it does not contribute to gas exchange; and (2) perfusion of nonventilated lung (a “shunt”), which allows venous blood to pass through the lung unaltered. When mixed with fully oxygenated blood leaving other well-ventilated lung units, shunted venous blood disproportionately lowers the mixed arterial Pa_{O_2} as a result of the nonlinear oxygen content versus Po_2 relationship of hemoglobin **(Fig. 279-5)**. Furthermore, the resulting arterial hypoxemia is refractory to supplemental inspired oxygen. The reason is that (1) raising the inspired FI_{O_2} has no effect on alveolar gas tensions in nonventilated alveoli and (2) while raising inspired FI_{O_2} increases Pa_{CO_2} in ventilated alveoli, the oxygen content of blood exiting ventilated units increases only slightly, as hemoglobin will already have been nearly fully saturated and the solubility of oxygen in plasma is quite small.

FIGURE 279-5

Influence of air versus oxygen breathing on mixed arterial oxygenation in shunt and ventilation/perfusion heterogeneity. Partial pressure of oxygen (mmHg) and oxygen saturations are shown for mixed venous blood, for end capillary blood (normal vs affected alveoli), and for mixed arterial blood. FI_{O_2}, fraction of inspired oxygen; $\dot{V}/\dot{Q}$, ventilation/perfusion.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A more common occurrence than the two extreme examples given above is a widening of the distribution of ventilation/perfusion ratios; such V̇/Q̇ heterogeneity is a common consequence of lung disease. In this circumstance, perfusion of relatively underventilated alveoli results in the incomplete oxygenation of exiting blood. When mixed with well-oxygenated blood leaving higher V̇/Q̇ regions, this partially reoxygenated blood disproportionately lowers arterial Pa_{O_2}, although to a lesser extent than does a similar perfusion fraction of blood leaving regions of pure shunt. In addition, in contrast to shunt regions, inhalation of supplemental oxygen raises the Pa_{O_2}, even in relatively underventilated low V̇/Q̇ regions, and so the arterial hypoxemia induced by V̇/Q̇ heterogeneity is typically responsive to oxygen therapy (Fig. 279-5).

In sum, arterial hypoxemia can be caused by substantial reduction of inspired oxygen tension, severe alveolar hypoventilation, perfusion of relatively underventilated (low V̇/Q̇) or completely unventilated (shunt) lung regions, and, in very unusual circumstances, by limitation of gas diffusion.

PATHOPHYSIOLOGY

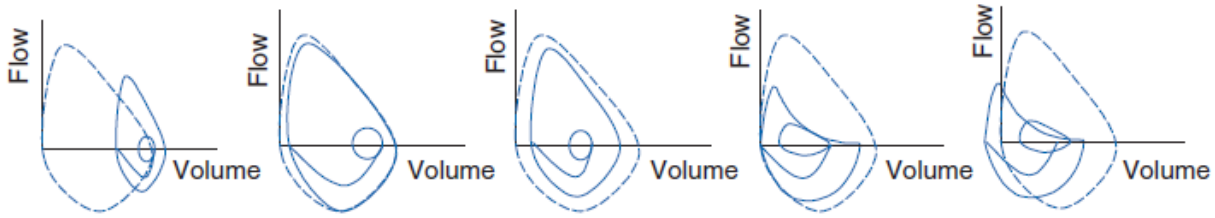
Although many diseases injure the respiratory system, this system responds to injury in relatively few ways. For this reason, the pattern of physiologic abnormalities may or may not provide sufficient information by which to discriminate among conditions.

Figure 279-6 lists abnormalities in pulmonary function testing that are typically found in a number of common respiratory disorders and highlight the simultaneous occurrence of multiple physiologic abnormalities. The coexistence of some of these respiratory disorders results in more complex superposition of these abnormalities. Methods to measure respiratory system function clinically are described later in this chapter.

FIGURE 279-6

Common abnormalities of pulmonary function (see text). Pulmonary function values are expressed as a percentage of normal predicted values, except for R_{aw}, which is expressed as $cmH_2O/L/s$ (normal, <2 $cmH_2O/L/s$). The figures at the bottom of each column show the typical configuration of flow-volume loops in each condition, including the flow-volume relationship during tidal breathing. b.d., bronchodilator; DL_{CO}, diffusion capacity of lung for carbon monoxide; FEV_1, forced expiratory volume in 1 s; FRC, functional residual capacity; FVC, forced vital capacity; R_{aw}, airways resistance; RV, residual volume; TLC, total lung capacity.

	Restriction due to increased lung elastic recoil (pulmonary fibrosis)	Restriction due to chest wall abnormality (moderate obesity)	Restriction due to respiratory muscle weakness (myasthenia gravis)	Obstruction due to airway narrowing (acute asthma)	Obstruction due to decreased elastic recoil (severe emphysema)
TLC	60%	95%	75%	100%	130%
FRC	60%	65%	100%	104%	220%
RV	60%	100%	120%	120%	310%
FVC	60%	92%	60%	90%	60%
FEV_1	75%	92%	60%	35% pre-b.d. 75% post-b.d.	35% pre-b.d. 38% post-b.d.
R_{aw}	1.0	1.0	1.0	2.5	1.5
DL_{CO}	60%	95%	80%	120%	40%



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ventilatory Restriction due to Increased Elastic Recoil—Example: Idiopathic Pulmonary Fibrosis

Idiopathic pulmonary fibrosis raises lung recoil at all lung volumes, thereby lowering TLC, FRC, and RV as well as forced vital capacity (FVC). Maximal expiratory flows are also reduced from normal values but are elevated when considered in relation to lung volumes. Increased flow occurs both because the increased lung recoil drives greater maximal flow at any lung volume and because airway diameters are relatively increased due to greater radially outward traction exerted on bronchi by the stiff lung parenchyma. For the same reason, airway resistance is also normal. Destruction of the pulmonary capillaries by the fibrotic process results in a marked reduction in diffusing capacity (see below). Oxygenation is often severely reduced by persistent perfusion of alveolar units that are relatively underventilated due to fibrosis of nearby (and mechanically linked) lung. The flow-volume loop (see below) looks like a miniature version of a normal loop but is shifted toward lower absolute lung volumes and displays maximal expiratory flows that are increased for any given volume over the normal tracing.

Ventilatory Restriction due to Chest Wall Abnormality—Example: Moderate Obesity

As the size of the average American continues to increase, this pattern may become the most common of pulmonary function abnormalities. In moderate obesity, the outward recoil of the chest wall is blunted by the weight of chest wall fat

and the space occupied by intra-abdominal fat. In this situation, preserved inward recoil of the lung overbalances the reduced outward recoil of the chest wall, and FRC falls. Because respiratory muscle strength and lung recoil remain normal, TLC is typically unchanged (although it may fall in massive obesity) and RV is normal (but may be reduced in massive obesity). Mild hypoxemia may be present due to perfusion of alveolar units that are poorly ventilated because of airway closure in dependent portions of the lung during breathing near the reduced FRC. Flows remain normal, as does the diffusion capacity of the lung for carbon monoxide ($D_{L_{CO}}$), unless obstructive sleep apnea (which often accompanies obesity) and associated chronic intermittent hypoxemia have induced pulmonary arterial hypertension, in which case $D_{L_{CO}}$ may be low.

Ventilatory Restriction due to Reduced Muscle Strength—Example: Myasthenia Gravis

In this circumstance, FRC remains normal, as both lung recoil and passive chest wall recoil are normal. However, TLC is low and RV is elevated because respiratory muscle strength is insufficient to push the passive respiratory system fully toward either volume extreme. Caught between the low TLC and the elevated RV, FVC, and FEV_1 are reduced as "innocent bystanders." As airway size and lung vasculature are unaffected, both R_{aw} and $D_{L_{CO}}$ are normal. Oxygenation is normal unless weakness becomes so severe that the patient has insufficient strength to reopen collapsed alveoli during sighs, with resulting atelectasis.

Airflow Obstruction due to Decreased Airway Diameter—Example: Acute Asthma

During an episode of acute asthma, luminal narrowing due to smooth muscle constriction as well as inflammation and thickening within the small- and medium-sized bronchi raise frictional resistance and reduce airflow. "Scooping" of the flow-volume loop is caused by reduction of airflow, especially at lower lung volumes. Often, airflow obstruction can be reversed by inhalation of β_2-adrenergic agonists acutely or by treatment with inhaled steroids chronically. TLC usually remains normal (although elevated TLC is sometimes seen in long-standing asthma), but FRC may be dynamically elevated. RV is often increased due to exaggerated airway closure at low lung volumes, and this elevation of RV reduces FVC. Because central airways are narrowed, R_{aw} is usually elevated. Mild arterial hypoxemia is often present due to perfusion of relatively underventilated alveoli distal to obstructed airways (and is responsive to oxygen supplementation), but $D_{L_{CO}}$ is normal or mildly elevated.

Airflow Obstruction due to Decreased Elastic Recoil—Example: Severe Emphysema

Loss of lung elastic recoil in severe emphysema results in pulmonary hyperinflation, of which elevated TLC is the hallmark. FRC is more severely elevated due to both loss of lung elastic recoil and dynamic hyperinflation—the same phenomenon as auto-PEEP (auto–positive end-expiratory pressure), which is the positive end-expiratory alveolar pressure that occurs when a new breath is initiated before the lung volume is allowed to return to FRC. RV is very severely elevated because of airway closure and because exhalation toward RV may take so long that RV cannot be reached before the patient must inhale again. Both FVC and FEV_1 are markedly decreased, the former because of the severe elevation of RV and the latter because loss of lung elastic recoil reduces the pressure driving maximal expiratory flow and also reduces tethering open of small intrapulmonary airways. The flow-volume loop demonstrates marked scooping, with an initial transient spike of flow attributable largely to expulsion of air from collapsing central airways at the onset of forced exhalation. Otherwise, the central airways remain relatively unaffected, so R_{aw} is normal in "pure" emphysema. Loss of alveolar surface and capillaries in the alveolar walls reduces $D_{L_{CO}}$; however, because poorly ventilated emphysematous acini are also poorly perfused (due to loss of their capillaries), arterial hypoxemia usually is not seen at rest until emphysema becomes very severe. However, during exercise, Pa_{O_2} may fall precipitously if extensive destruction of the pulmonary vasculature prevents a sufficient increase in cardiac output and mixed venous oxygen content falls substantially. Under these circumstances, any venous admixture through low $\dot{V}/\dot{Q}$ units has a particularly marked effect in lowering mixed arterial oxygen tension.

FUNCTIONAL MEASUREMENTS

Measurement of Ventilatory Function

LUNG VOLUMES

Figure 279-2 demonstrates a spirometry tracing in which the volume of air entering or exiting the lung is plotted over time. In a slow vital capacity maneuver, the patient inhales from FRC, fully inflating the lungs to TLC, and then exhales slowly to RV; VC, the difference between TLC and RV, represents the maximal excursion of the respiratory system. Spirometry discloses relative volume changes during these maneuvers but cannot reveal the absolute volumes at which

they occur. To determine absolute lung volumes, two approaches are commonly used: inert gas dilution and body plethysmography. In the former, a known amount of a nonabsorbable inert gas (usually helium or neon) is inhaled in a single large breath or is rebreathed from a closed circuit; the inert gas is diluted by the gas resident in the lung at the time of inhalation, and its final concentration reveals the volume of resident gas contributing to the dilution. A drawback of this method is that regions of the lung that ventilate poorly (e.g., due to airflow obstruction) may not receive much inspired inert gas and so do not contribute to its dilution. Therefore, inert gas dilution (especially in the single-breath method) often underestimates true lung volumes.

In the second approach, FRC is determined by measuring the compressibility of gas within the chest, which is proportional to the volume of gas being compressed. The patient sits in a body plethysmograph (a chamber usually made of transparent plastic to minimize claustrophobia) and, at the end of a normal tidal breath (i.e., when lung volume is at FRC), is instructed to pant against a closed shutter, thus periodically compressing air within the lung slightly. Pressure fluctuations at the mouth and volume fluctuations within the body box (equal but opposite to those in the chest) are determined, and from these measurements, the thoracic gas volume is calculated by means of Boyle's law. Once FRC is obtained, TLC and RV are calculated by adding the value for inspiratory capacity and subtracting the value for expiratory reserve volume, respectively (both values having been obtained during spirometry) (Fig. 279-2). The most important determinants of healthy individuals' lung volumes are height, age, and sex, but there is considerable additional normal variation beyond that accounted for by these parameters. In addition, race influences lung volumes; on average, TLC values are ~12% lower in African Americans and 6% lower in Asian Americans than in Caucasian Americans. In practice, a mean "normal" value is predicted by multivariate regression equations using height, age, and sex, and the patient's value is divided by the predicted value (often with "race correction" applied) to determine "percent predicted." For most measures of lung function, 85–115% of the predicted value can be normal; however, in health, the various lung volumes tend to scale together. For example, if one is "normal big" with a TLC 110% of the predicted value, all other lung volumes and spirometry values will also approximate 110% of their respective predicted values. This pattern is particularly helpful in evaluating airflow, as discussed below.

AIR FLOW

As noted above, spirometry plays a key role in lung volume determination. Even more often, spirometry is used to measure airflow, which reflects the dynamic properties of the lung. During an FVC maneuver, the patient inhales to TLC and then exhales rapidly and forcefully to RV; this method ensures that flow limitation has been achieved, so that the precise effort made has little influence on actual flow. The total amount of air exhaled is the FVC, and the amount of air exhaled in the first second is the FEV_1; the FEV_1 is a flow rate, revealing volume change per time. Like lung volumes, an individual's maximal expiratory flows should be compared with predicted values based on height, age, and sex. While the FEV_1/FVC ratio is typically reduced in airflow obstruction, this condition can also reduce FVC by raising RV, sometimes rendering the FEV_1/FVC ratio "artifactually normal" with the erroneous implication that airflow obstruction is absent. To circumvent this problem, it is useful to compare FEV_1 as a fraction of its predicted value with TLC as a fraction of its predicted value. In health, the results are usually similar. In contrast, even an FEV_1 value that is 95% of its predicted value may actually be relatively low if TLC is 110% of its respective predicted value. In this case, airflow obstruction may be present, despite the "normal" value for FEV_1.

The relationships among volume, flow, and time during spirometry are best displayed in two plots—the spirogram (volume vs time) and the flow-volume loop (flow vs volume) (Fig. 279-4). In conditions that cause airflow obstruction, the site of obstruction is sometimes correlated with the shape of the flow-volume loop. In diseases that cause lower airway obstruction, such as asthma and emphysema, flows decrease more rapidly with declining lung volumes, leading to a characteristic scooping of the flow-volume loop. In contrast, fixed upper-airway obstruction typically leads to inspiratory and/or expiratory flow plateaus (Fig. 279-4).

AIRWAYS RESISTANCE

The total resistance of the pulmonary and upper airways is measured in the same body plethysmograph used to measure FRC. The patient is asked once again to pant, but this time against a closed and then opened shutter. Panting against the closed shutter reveals the thoracic gas volume as described above. When the shutter is opened, flow is directed to and from the body box, so that volume fluctuations in the box reveal the extent of thoracic gas compression, which in turn reveals the pressure fluctuations driving flow. Simultaneous measurement of flow allows the calculation of lung resistance (as flow divided by pressure). In health, R_{aw} is very low (<2 cmH_2O/L/s), and half of the detected resistance resides within the upper airway. In the lung, most resistance originates in the central airways. For this reason, airways resistance measurement tends to be insensitive to peripheral airflow obstruction.

RESPIRATORY MUSCLE STRENGTH

To measure respiratory muscle strength, the patient is instructed to exhale or inhale with maximal effort against a closed shutter while pressure is monitored at the mouth. Pressures >±60 cmH_2O at FRC are considered adequate and make it unlikely that respiratory muscle weakness accounts for any other resting ventilatory dysfunction that is identified.

Measurement of Gas Exchange

DIFFUSING CAPACITY (DL_{CO})

This test uses a small (and safe) amount of carbon monoxide (CO) to measure gas exchange across the alveolar membrane during a 10-s breath hold. CO in exhaled breath is analyzed to determine the quantity of CO crossing the alveolar membrane and combining with hemoglobin in red blood cells. This “single-breath diffusing capacity” (DL_{CO}), value increases with the surface area available for diffusion and the amount of hemoglobin within the capillaries, and it varies inversely with alveolar membrane thickness. Thus, DL_{CO} decreases in diseases that thicken or destroy alveolar membranes (e.g., pulmonary fibrosis, emphysema), curtail the pulmonary vasculature (e.g., pulmonary hypertension), or reduce alveolar capillary hemoglobin (e.g., anemia). Single-breath diffusing capacity may be elevated in acute congestive heart failure, asthma, polycythemia, and pulmonary hemorrhage.

Arterial Blood Gases

The effectiveness of gas exchange can be assessed by measuring the partial pressures of oxygen and CO_2 in a sample of blood obtained by arterial puncture. The oxygen content of blood (Ca_{O_2}) depends on arterial saturation ($\%O_2Sat$), which is set by Pa_{O_2}, pH, and Pa_{CO_2} according to the oxyhemoglobin dissociation curve. Ca_{O_2} can also be measured by oximetry (see below):

$$Ca_{O_2}\,(mL/dL) = 1.39\,(mL/dL) \times [hemoglobin](g) \times \%\,O_2\,Sat + 0.003\,(mL/dL/mmHg) \times Pa_{O_2}\,(mmHg)$$

If hemoglobin saturation alone needs to be determined, this task can be accomplished noninvasively with pulse oximetry.

ACKNOWLEDGMENT

The authors wish to acknowledge the contributions of Drs. Steven E. Weinberger and Irene M. Rosen to this chapter in previous editions.

FURTHER READING

Bates JH: Systems physiology of the airways in health and obstructive pulmonary disease. Wiley Interdiscip Rev Sys Biol Med 8:423, 2016.
Hughes JM et al: Effect of lung volume on the distribution of pulmonary blood flow in man. Respir Physiol 4:58, 1968. [PubMed: 5639524]
Levitsky MG et al: *Pulmonary Physiology*, 8th ed. New York, McGraw Hill, 2013; *http://accessmedicine.mhmedical.com.ezproxy1.library.usyd.edu.au/book.aspx?bookid=575*. Accessed June 6, 2017.
Macklem PT, Murphy BR: The forces applied to the lung in health and disease. Am J Med 57:371, 1974. [PubMed: 4416277]
Pederson OF, Ingram RH: Configuration of maximal expiratory flow-volume curve: Model experiments with physiologic implications. J Appl Physiol 58:1305, 1985. [PubMed: 3988684]
Prange HD: *Respiratory Physiology: Understanding Gas Exchange*. New York, Chapman and Hill, 1996.
Weibel ER: Morphometric estimation of pulmonary diffusion capacity, I. Model and method. Respir Physiol 11:54, 1970. [PubMed: 4992513]
West JB: *Respiratory Physiology, The Essentials*, 9th ed. Philadelphia, Lippincott Williams & Wilkins, 2012.
Wiley Online Library: Comprehensive Physiology: The Respiratory System. Available from *http://www.comprehensivephysiology.com/WileyCDA/Section/id-420557.html*. Accessed August 12, 2016.

Chapter 280: Diagnostic Procedures in Respiratory Disease

Anne L. Fuhlbrigge; Augustine M.K. Choi

INTRODUCTION

The diagnostic modalities available for assessing the patient with suspected or known respiratory system disease include imaging studies and techniques for acquiring biologic specimens, some of which involve direct visualization of part of the respiratory system. Methods to characterize the functional changes developing as a result of disease, including pulmonary function tests and measurements of gas exchange, are discussed in **Chap. 279**.

IMAGING STUDIES

ROUTINE RADIOGRAPHY

Routine chest radiography, including both posteroanterior (PA) and lateral views, is an integral part of the diagnostic evaluation of diseases involving the pulmonary parenchyma, the pleura, and, to a lesser extent, the airways and the mediastinum **(see Chaps. 278 and A12)**. Lateral decubitus views are useful for determining whether pleural abnormalities represent freely flowing fluid, whereas apical lordotic views can visualize disease at the lung apices better than the standard PA view. Portable equipment is often used for acutely ill patients who cannot be transported to a radiology suite but are more difficult to interpret owing to several limitations: (1) the single anteroposterior (AP) projection obtained; (2) variability in over- and underexposure of film; (3) a shorter focal spot-film distance leading to lack of edge sharpness and loss of fine detail; and (4) magnification of the cardiac silhouette and other anterior structures by the AP projection. Common radiographic patterns and their clinical correlates are reviewed in **Chap. A12**.

Advances in computer technology have allowed the development of digital or computed radiography, which has several benefits: (1) immediate availability of the images; (2) significant postprocessing analysis of images to improve diagnostic information; and (3) ability to store images electronically and to transfer them within or between health care systems.

ULTRASOUND

Diagnostic ultrasound (US) produces images using echoes or reflection of the US beam from interfaces between tissues with differing acoustic properties. US is nonionizing and safe to perform on pregnant patients and children. It can detect and localize pleural abnormalities, guide percutaneous needle biopsy of peripheral lung, pleural, or chest wall lesions and identify septations within loculated pleural collections (i.e., for thoracentesis), improving the yield and safety of the procedure. Real-time imaging can be used to assess the movement of the diaphragm and can demonstrate changes in clinical condition. Availability of portable machines has allowed point of care (POC) ultrasound to provide rapid and accurate bedside diagnosis and monitoring of several common respiratory

conditions, including pneumothorax, and pleural effusions. In experienced hands, POC ultrasound has higher sensitivity and specificity than chest radiography in detecting pleural effusions and pneumothorax, with an accuracy approaching computed tomography (CT). Pulmonary congestion may be quantified using lung US monitoring pulmonary congestion in heart failure patients in response to therapy.

NUCLEAR MEDICINE TECHNIQUES

Nuclear imaging depends on the selective uptake of various compounds by organs of the body. In thoracic imaging, these compounds are concentrated by one of three mechanisms: blood pool or compartmentalization (e.g., within the heart), physiologic incorporation (e.g., bone or thyroid) and capillary blockage (e.g., lung scan). Radioactive isotopes can be administered by either the IV or inhaled routes or both. When injected intravenously, albumin macroaggregates labeled with technetium-99m (^{99m}Tc) become lodged in pulmonary capillaries; the distribution of the trapped radioisotope follows the distribution of blood flow. When inhaled, radiolabeled xenon gas can be used to demonstrate the distribution of ventilation. Using these techniques, ventilation-perfusion lung scanning was a commonly used technique for the evaluation of pulmonary embolism. Pulmonary thromboembolism produces one or more regions of ventilation-perfusion mismatch (i.e., regions in which there is a defect in perfusion that follows the distribution of a vessel and that is not accompanied by a corresponding defect in ventilation **[Chap. 273]**). However, with advances in CT scanning, scintigraphic imaging has been largely replaced by CT angiography in patients with suspected pulmonary embolism.

Another common use of ventilation-perfusion scans is in patients with impaired lung function, who are being considered for lung resection. Many patients with bronchogenic carcinoma have coexisting chronic obstructive pulmonary disease (COPD), and the question arises as to whether or not a patient can tolerate lung resection. The distribution of the isotope(s) can be used to assess the regional distribution of blood flow and ventilation, allowing the physician to estimate the level of postoperative lung function.

COMPUTED TOMOGRAPHY

CT offers several advantages over routine chest radiography **(Figs. 280-1A, B and 280-2A, B)**. First, the use of cross-sectional images allows distinction between densities that would be superimposed on plain radiographs. Second, CT is far better than routine radiographic studies at characterizing tissue density and providing accurate size assessment of lesions.

FIGURE 280-1

Chest x-ray (*A*) and computed tomography (CT) scan (*B*) from a patient with emphysema. The extent and distribution of emphysema are not well appreciated on plain film but clearly evident on the CT scan obtained.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

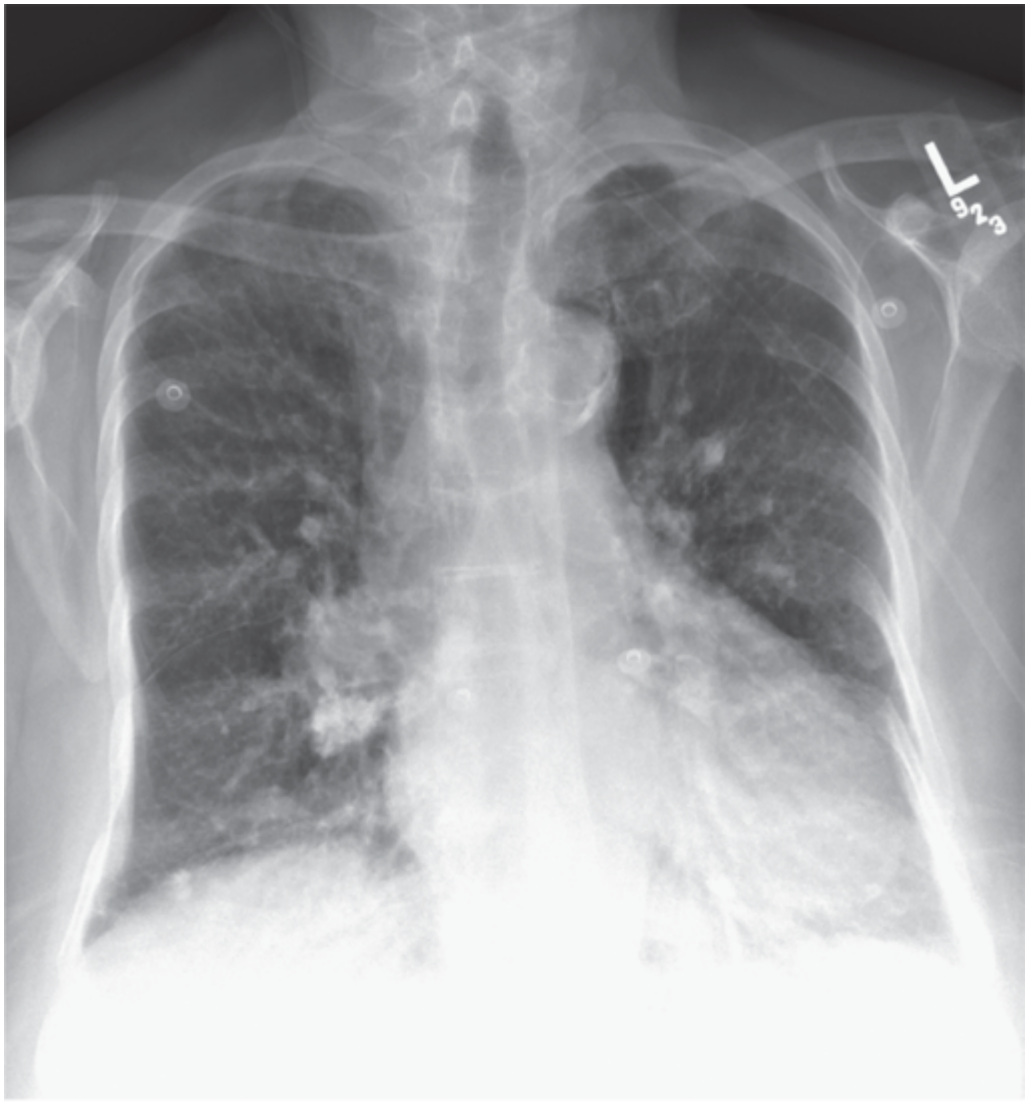


B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 280-2

Chest x-ray (*A*) and computed tomography (CT) scan (*B*) demonstrating a right lower-lobe mass. The mass is not well appreciated on the plain film because of the hilar structures and known calcified adenopathy. CT is superior to plain radiography for the detection of abnormal mediastinal densities and the distinction of masses from adjacent vascular structures.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

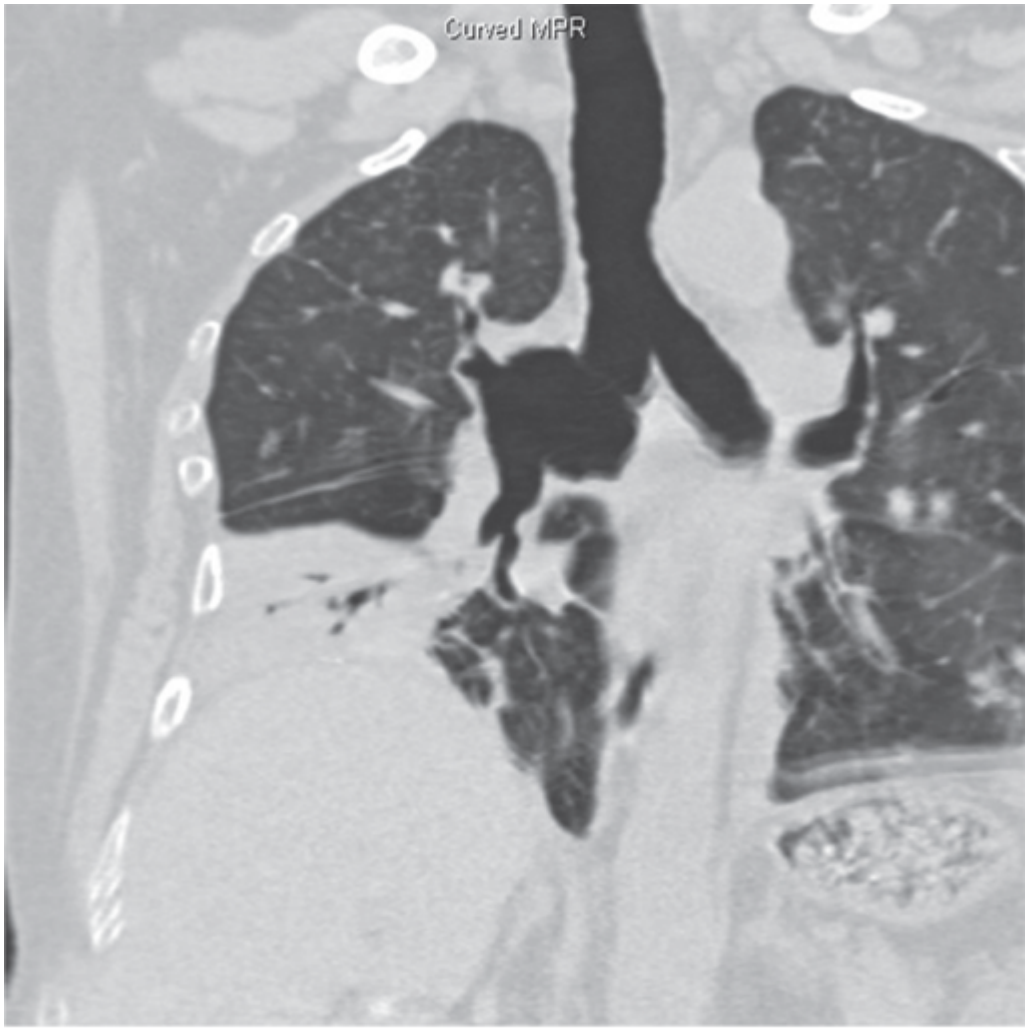
CT is particularly valuable in assessing hilar and mediastinal disease (often poorly characterized by plain radiography), in identifying and characterizing disease adjacent to the chest wall or spine (including pleural disease), and in identifying areas of fat density or calcification in pulmonary nodules **(Fig. 280-2)**. Its utility in the assessment of mediastinal disease has made CT an important tool in the staging of lung cancer **(Chap. 74)**. With the additional use of contrast material, CT also makes it possible to distinguish vascular from nonvascular structures, which is particularly important in distinguishing lymph nodes and masses from vascular structures primarily in the mediastinum, and vascular disorders such as pulmonary embolism.

Helical CT and Multidetector CT

Helical scanning is currently the standard method for thoracic CT. Helical CT technology results in faster scans with improved contrast enhancement and thinner collimation. Images are obtained during a single breath-holding maneuver that allows less motion artifact and collection of continuous data over a larger volume of lung than is possible with conventional CT. Data from the imaging procedure can be reconstructed in coronal or sagittal planes **(Fig. 280-3A)**, as well as the traditional cross-sectional (axial) view.

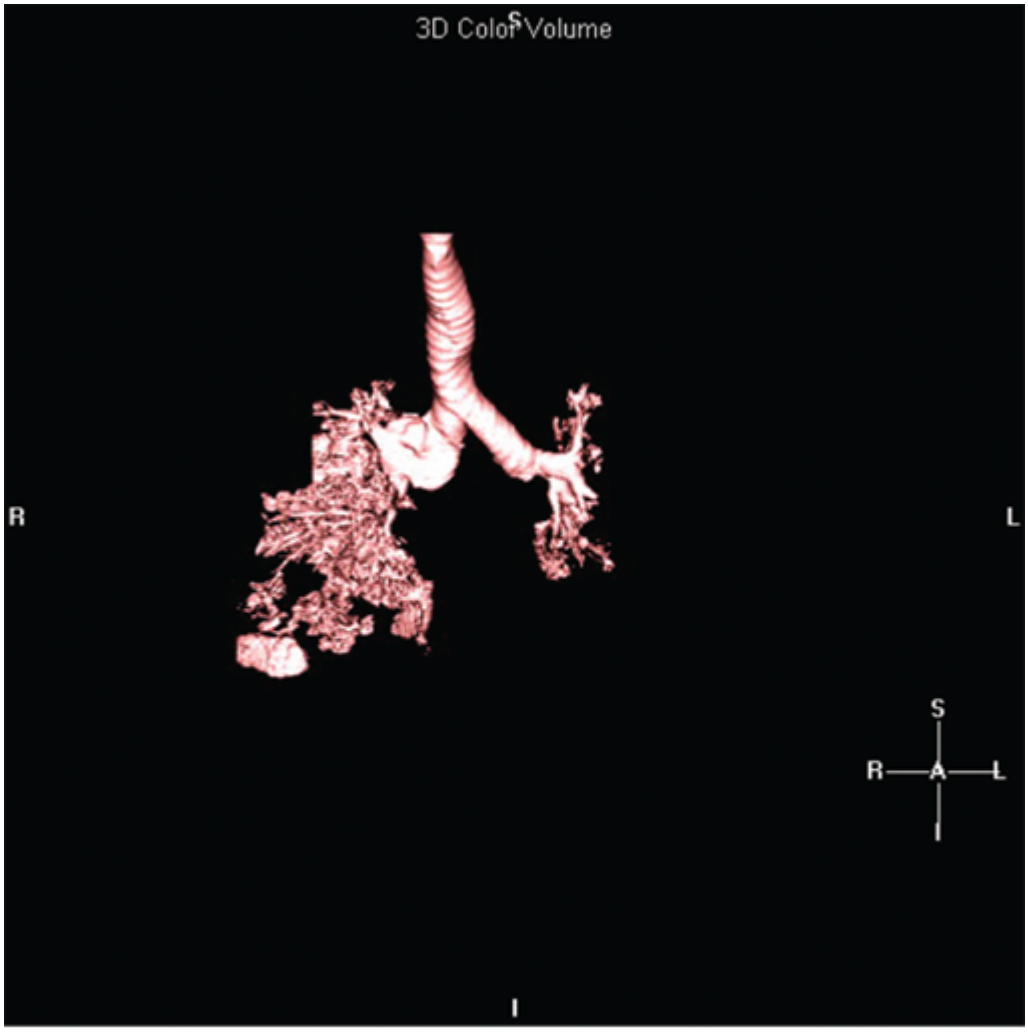
FIGURE 280-3

Spiral computed tomography (CT) with reconstruction of images in planes other than axial view. Spiral CT in a lung transplant patient with a dehiscence and subsequent aneurysm of the anastomosis. CT images were reconstructed in the sagittal view (***A***) and using digital subtraction to view images of the airways only (***B***), which demonstrate the exact location and extent of the abnormality.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Further refinements in detector technology have allowed production of scanners with additional detectors along the scanning axis (*z*-axis). These *multidetector CT* (MDCT) scanners can obtain multiple slices in a single rotation that are thinner and can be acquired in a shorter period of time. This results in enhanced resolution and increased image reconstruction ability. As the technology has progressed, higher numbers (currently up to 64) of detectors allow submillimeter spatial resolution to produce clearer final images, allowing this technique to essentially replace high-resolution CT (HRCT) in the evaluation of lung disease. The pattern of usual interstitial pneumonia (UIP) seen on MDCT (peripheral, basilar predominant honeycomb structure, and traction bronchiectasis), together with typical clinical presentation, and other causes are ruled out, the diagnosis of idiopathic pulmonary fibrosis (IPF) can be reliably diagnosed without histological confirmation. MDCT allows for even shorter breath holds, which are beneficial for all patients but especially children, the elderly, and the critically ill. It should be noted that despite the advantages of MDCT, there is an increase in radiation dose compared to single-detector CT to consider. However, using iterative reconstruction techniques, there is continued progress in reducing the radiation dose reported for CT scans of the thorax. Low dose MDCT is now a recommended screening procedure for lung cancer among persons who are aged 55–80 years with 30 pack year smoking history and currently smoke or quit within the past 15 years.

In MDCT, the additional detectors along the *z*-axis result in improved use of the contrast bolus. This and the faster scanning times and increased resolution have all led to improved imaging of the

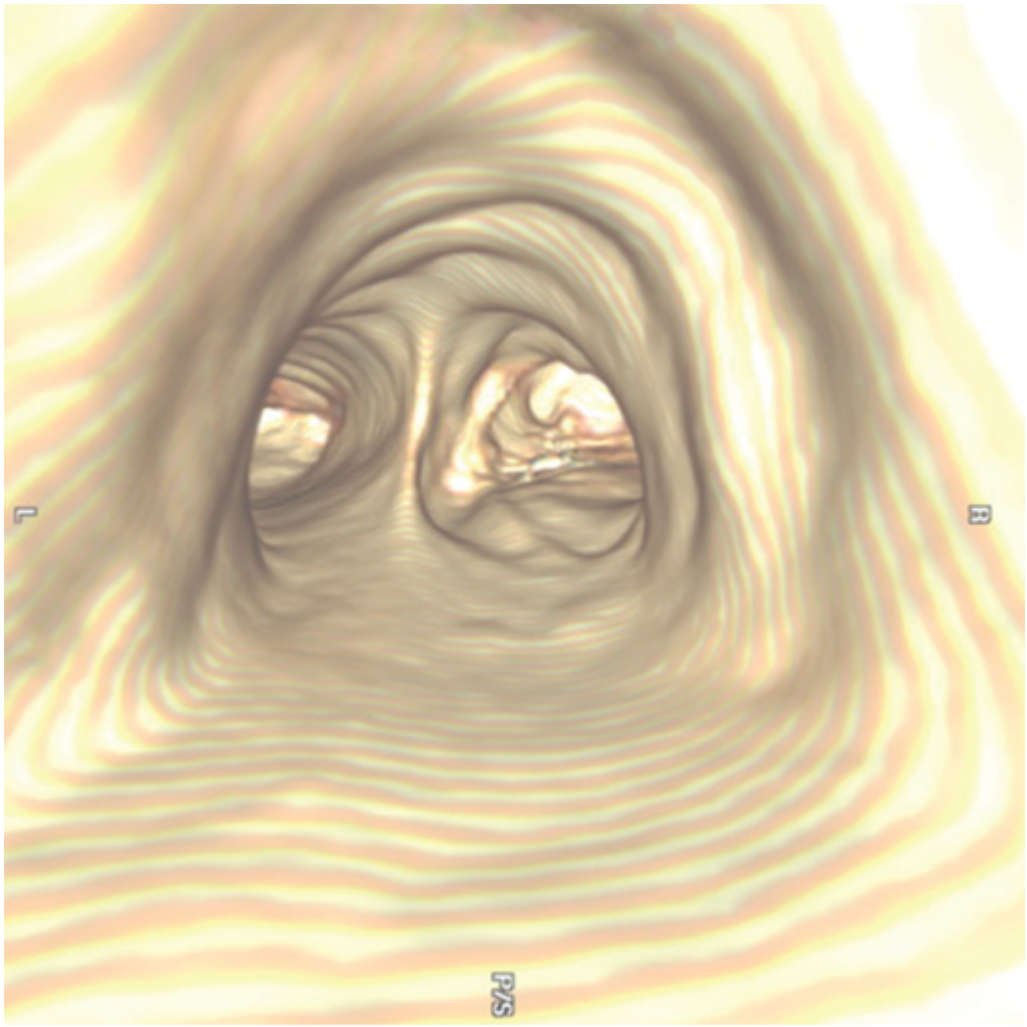
pulmonary vasculature and the ability to detect segmental and subsegmental emboli. CT pulmonary angiography (CTPA) also allows simultaneous detection of parenchymal abnormalities that may be contributing to a patient's clinical presentation. Secondary to these advantages and increasing availability, CTPA has rapidly become the test of choice for many clinicians in the evaluation of pulmonary embolism; compared with pulmonary angiography, it is considered equal in terms of accuracy and with less associated risks. A further development is the dual-source CT (DSCT), which uses two x-ray tubes and their corresponding detectors offset by 90°. These scanners can emphasize particular tissue characteristics and combine functional and morphological information, which may allow better detection of perfusion defects in the lung parenchyma. In addition, the newer generation DSCT systems allow high resolution scans of the thorax to be performed in <1 s, of particular interest for dyspneic patients who are unable to comply with breath hold instructions.

VIRTUAL BRONCHOSCOPY

The three-dimensional (3D) image of the thorax obtained by MDCT can be digitally stored, reanalyzed, and displayed as 3D reconstructions of the airways down to the sixth to seventh generation. Using these reconstructions, a "virtual" bronchoscopy can be performed **(Fig. 280-4)**. Virtual bronchoscopy has been proposed as an adjunct to conventional bronchoscopy in several clinical situations: It can allow accurate assessment of the extent and length of an airway stenosis, including the airway distal to the narrowing; it can provide useful information about the relationship of the airway abnormality to adjacent mediastinal structures; and it allows preprocedure planning for therapeutic bronchoscopy to help ensure the appropriate equipment is available for the procedure.

FIGURE 280-4

Virtual bronchoscopic image of the trachea. The view projected is one that would be obtained from the trachea looking down to the carina. The left and right main stem airways are seen bifurcating from the carina.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Electromagnetic navigational bronchoscopy systems (EMN or ENB), using virtual bronchoscopy, have been developed to allow accurate navigation to peripheral pulmonary target lesions. Electromagnetic navigation bronchoscopy (ENB) uses technology similar to a car global positioning system (GPS) unit, which allows precise tracking of both position and orientation through the use of electromagnetic fields.

POSITRON EMISSION TOMOGRAPHIC SCANNING

Positron emission tomographic (PET) scanning involves injection of a radiolabeled glucose analogue, [^{18}F]-fluoro-2-deoxyglucose (FDG), which is taken up by metabolically active malignant cells. This technique has been used in the evaluation of solitary pulmonary nodules and in staging lung cancer. Detection or exclusion of mediastinal lymph node involvement and identification of extrathoracic disease can be achieved. The development of hybrid imaging allows the superimposition of PET and CT images, a technique known as functional–anatomical mapping. Hybrid PET/CT scans provide images that help pinpoint the abnormal metabolic activity to anatomical structures seen on CT and provide more accurate diagnoses than the two scans performed separately. FDG-PET can differentiate benign from malignant lesions as small as 1 cm and can be very useful in detection of distant metastases. However, false-negative findings can occur in lesions with low metabolic activity such as carcinoid tumors and bronchioloalveolar cell carcinomas, or in lesions <1 cm in which the required threshold of metabolically active malignant cells is not present for PET diagnosis. False-

positive results can be seen due to FDG uptake in inflammatory conditions such as pneumonia and granulomatous diseases.

MAGNETIC RESONANCE IMAGING

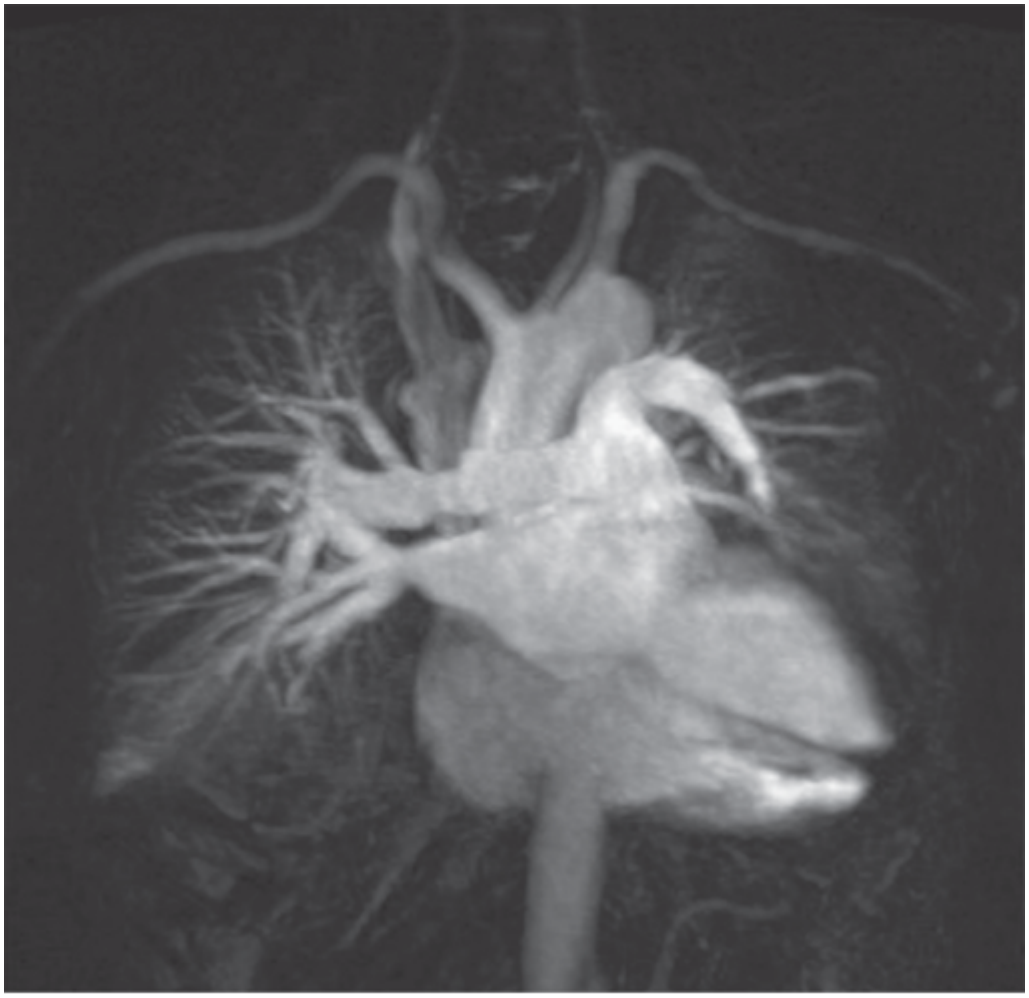
Magnetic resonance (MR) provides poorer spatial resolution and less detail of the pulmonary parenchyma and, for these reasons, is not currently considered a substitute for CT in imaging the thorax. However, because of the high soft tissue contract available with MRI, this technology may be used to distinguish tumor from post-stenotic atelectasis and assess infiltration of the chest wall and/or mediastinum. In addition, for superior sulcus tumors, MRI can be valuable in preoperative planning to better visualize if/where the tumor is in contact with the spine. Further, “diffusion-weighted” MRI is an emerging technique that has been used to differentiate metastatic lymph nodes from healthy lymph nodes with sensitivity, specificity, and positive predictive values higher than PET/CT or CT alone. Finally, the use of hyperpolarized gas in conjunction with MR has led to the investigational use of MR for imaging the lungs, particularly in obstructive lung disease. Imaging performed during an inhalation and exhalation can provide dynamic information on lung function.

An advantage of MR is the use of nonionizing electromagnetic radiation. Additionally, MR is well suited to distinguish vascular from nonvascular structures without the need for contrast. Blood vessels appear as hollow tubular structures because flowing blood does not produce a signal on MRI. Therefore, MR can be useful in demonstrating pulmonary emboli, defining aortic lesions such as aneurysms or dissection, or other vascular abnormalities **(Fig. 280-5)** if radiation and IV contrast medium cannot be used. Gadolinium can be used as an intravascular contrast agent for MR angiography (MRA); however, synchronization of data acquisition with the peak arterial bolus is one of the major challenges of MRA. The flow of contrast medium from the peripheral injection site to the vessel of interest is affected by a number of factors including heart rate, stroke volume, and the presence of proximal stenotic lesions.

Disadvantages of MRI include less spatial resolution and longer study acquisition times compared with CT. MR examinations are difficult to obtain among patients who cannot lie still or who cannot lay on their backs. MRI is generally avoided in unstable and/or ventilated patients and those with severe trauma because of the hazards of the MR environment and the difficulties in monitoring patients within the MR room. The presence of metallic foreign bodies, pacemakers, and intracranial aneurysm clips also preclude use of MRI.

FIGURE 280-5

Magnetic resonance angiography image of the vasculature of a patient after lung transplant. The image demonstrates the detailed view of the vasculature that can be obtained using digital subtraction techniques. Images from a patient after lung transplant show the venous and arterial anastomosis on the right; a slight narrowing is seen at the site of the anastomosis, which is considered within normal limits and not suggestive of obstruction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PULMONARY ANGIOGRAPHY

The pulmonary arterial system can be visualized by pulmonary angiography, in which radiopaque contrast medium is injected through a catheter placed in the pulmonary artery. When performed in cases of pulmonary embolism, pulmonary angiography demonstrates the consequences of an intravascular thrombus—either a defect in the lumen of a vessel (a filling defect) or an abrupt termination (cutoff) of the vessel. Other, less common indications for pulmonary angiography include visualization of a suspected pulmonary arteriovenous malformation and assessment of pulmonary arterial invasion by a neoplasm. The risks associated with modern arteriography are small, generally of greatest concern in patients with severe pulmonary hypertension or chronic kidney disease. With advances in CT scanning, MDCT angiography (MDCTA) is replacing conventional angiography for the diagnosis of pulmonary embolism.

MEDICAL TECHNIQUES FOR OBTAINING BIOLOGIC SPECIMENS

COLLECTION OF BLOOD AND SERUM

Testing of blood and/or serum can be useful in situations where respiratory diseases are secondary to systemic illness. Collagen vascular disease is a frequent cause of diffuse interstitial lung disease

(DILD) and serologic tests for autoantibodies may be helpful in determining if an autoimmune disorder is affecting the lungs. In addition, blood tests for inherited respiratory diseases are available. In patients presenting with COPD, a low level of α_1-antitrypsin (α_1AT) confirms α_1AT deficiency and further assessment with α_1AT protein phenotyping and/or α_1AT genotyping can be carried if needed. Beyond emphysema, next-generation sequencing has allowed development of respiratory gene panels that identify genes implicated in several different lung syndromes including cystic, fibrotic, and bronchiectatic diseases.

COLLECTION OF SPUTUM

Sputum can be collected either by spontaneous expectoration or induced (after inhalation of an irritating aerosol such as hypertonic saline). *Sputum induction* is used either because sputum is not spontaneously being produced or because of an expected higher yield of certain types of findings. Because sputum consists mainly of secretions from the tracheobronchial tree rather than the upper airway, the finding of alveolar macrophages and other inflammatory cells is consistent with a lower respiratory tract origin of the sample, whereas the presence of squamous epithelial cells in a “sputum” sample indicates contamination by secretions from the upper airways.

In addition to processing for routine bacterial pathogens by Gram’s method and culture, sputum can be processed for a variety of other pathogens, including staining and culture for mycobacteria or fungi, culture for viruses, and staining for *Pneumocystis jiroveci.* In the specific case of sputum obtained for evaluation of *P. jiroveci* pneumonia, for example, sputum should be collected by induction rather than spontaneous expectoration, and an immunofluorescent stain should be used to detect the organisms. Traditional stains and cultures are now also being supplemented in some cases by immunologic techniques and by molecular biologic methods, including the use of polymerase chain reaction (PCR) amplification and DNA probes. Cytologic staining of sputum for malignant cells, using the traditional Papanicolaou method, allows noninvasive evaluation for suspected lung cancer.

PERCUTANEOUS NEEDLE ASPIRATION (TRANSTHORACIC)

A needle can be inserted through the chest wall into a pulmonary lesion to obtain an aspirate or tissue core for cytologic/histologic or microbiologic analysis. Aspiration can be performed to obtain a diagnosis or to decompress and/or drain a fluid collection. The procedure is usually carried out under CT or US guidance to assist positioning of the needle and assure localization in the lesion. The low potential risk of this procedure (intrapulmonary bleeding or creation of a pneumothorax with collapse of the underlying lung) in experienced hands is usually acceptable compared with the information obtained. However, a limitation of the technique is sampling error due to the small size of the tissue sample. Thus, findings other than a specific cytologic or microbiologic diagnosis are of limited clinical value.

THORACENTESIS

Sampling of pleural liquid by thoracentesis is commonly performed for diagnostic purposes or, in the case of a large effusion, for palliation of dyspnea. Diagnostic sampling, either by blind needle aspiration or after localization by US, allows the collection of liquid for microbiologic and cytologic studies. Analysis of the fluid obtained for its cellular composition and chemical constituents allows classification of the effusion and can help with diagnosis and treatment **(Chap. 288)**.

BRONCHOSCOPY

Bronchoscopy is the process of direct visualization of the tracheobronchial tree. Although bronchoscopy is now performed almost exclusively with flexible fiberoptic instruments, rigid

bronchoscopy, generally performed in an operating room on a patient under general anesthesia, still has a role in selected circumstances, primarily because of a larger suction channel and the fact that the patient can be ventilated through the bronchoscope channel. These situations include the retrieval of a foreign body and the suctioning of a massive hemorrhage, for which the small suction channel of the bronchoscope may be insufficient.

FLEXIBLE FIBEROPTIC BRONCHOSCOPY

This outpatient procedure is usually performed in an awake but sedated patient (conscious sedation). The bronchoscope is passed through either the mouth or the nose, between the vocal cords, and into the trachea. The ability to flex the scope makes it possible to visualize virtually all airways to the level of subsegmental bronchi. The bronchoscopist is able to identify endobronchial pathology, including tumors, granulomas, bronchitis, foreign bodies, and sites of bleeding. Samples from airway lesions can be taken by several methods, including washing, brushing, and biopsy. Washing involves instillation of sterile saline through a channel of the bronchoscope and onto the surface of a lesion. A portion of the liquid is collected by suctioning through the bronchoscope, and the recovered material can be analyzed for cells (cytology) or organisms (by standard stains and cultures). Brushing or biopsy of the surface of the lesion, using a small brush or biopsy forceps at the end of a long cable inserted through a channel of the bronchoscope, allows recovery of cellular material or tissue for analysis by standard cytologic and histopathologic methods.

The bronchoscope can be used to sample material not only from the regions that can be directly visualized (i.e., the airways), but also from the more distal pulmonary parenchyma. With the bronchoscope wedged into a subsegmental airway, aliquots of sterile saline can be instilled through the scope, allowing sampling of cells and organisms from alveolar spaces. This procedure, called *bronchoalveolar lavage* (*BAL*), has been particularly useful for the recovery of fluid for culture. In addition, immunofluorescent staining with antibodies and/or nucleic acid analysis via PCR can facilitate more rapid diagnosis than culture techniques for some organisms. Cytology, cellular analysis, and examination of acellular components such as cytokines, viral particles, and microbial signatures are commonly performed.

Brushing and biopsy of the distal lung parenchyma can also be performed with the same instruments that are used for endobronchial sampling. These instruments can be passed through the scope into small airways. When biopsies are performed, the forceps penetrate the airway wall, allowing biopsy of peribronchial alveolar tissue. This procedure, called *transbronchial biopsy*, is used when there is either relatively diffuse disease or a localized lesion of adequate size. With the aid of fluoroscopic imaging, the bronchoscopist is able to determine not only whether and when the instrument is in the area of abnormality, but also the proximity of the instrument to the pleural surface. If the forceps are too close to the pleural surface, there is a risk of violating the visceral pleura and creating a pneumothorax; the other potential complication of transbronchial biopsy is pulmonary hemorrhage. The incidence of these complications is less than several percent.

TRANSBRONCHIAL NEEDLE ASPIRATION

Another procedure involves use of a hollow-bore needle passed through the bronchoscope for sampling of tissue adjacent to the trachea or a large bronchus. The needle is passed through the airway wall (transbronchial), and cellular material can be aspirated from mass lesions or enlarged lymph nodes, generally in a search for malignant cells. Mediastinoscopy has been considered the gold standard for mediastinal staging; however, transbronchial needle aspiration (TBNA) allows sampling from the lungs and surrounding lymph nodes without the need for surgery or general anesthesia.

ENDOBRONCHIAL ULTRASOUND (EBUS)–TRANSBRONCHIAL NEEDLE ASPIRATION (TBNA)

Further advances in needle aspiration techniques have been accomplished with the development of endobronchial ultrasound (EBUS). The technology uses an ultrasonic bronchoscope fitted with a probe that allows for needle aspiration of mediastinal and hilar lymph nodes guided by real-time US images. EBUS allows sampling of mediastinal lymph nodes and masses under direct vision to better identify and localize peribronchial and mediastinal pathology and offers access to more difficult-to-reach areas and smaller lymph nodes in the staging of malignancies. EBUS-TBNA has the potential to access the same paratracheal and subcarinal lymph node stations as mediastinoscopy, but also extends out to the hilar lymph nodes (levels 10 and 11).

Radial probe endobronchial ultrasound (RP-EBUS) produces a 360-degree ultrasound image of the surrounding lung parenchyma and has significantly improved the bronchoscopic diagnostic yield for peripheral pulmonary nodules, particularly for larger lesions (>2 cm). RP-EBUS can be combined with ENB (described above), to provide accurate navigational assistance to localize peripheral nodules and increase the diagnostic yield.

RP-EBUS has a superior safety profile compared with transthoracic approach, but limitations include a poor ultrasound signal for evaluating nodules that have a ground glass appearance on CT scan and a high reliance on the bronchoscopist’s ability to navigate the branching architecture of the airways to position the probe near the nodule.

EMERGING BRONCHOSCOPIC TECHNIQUES

Additional techniques that can be performed using bronchoscopy include video/autofluorescence bronchoscopy (AFB), narrow band imaging (NBI), optical coherence tomography (OCT), and endomicroscopy using confocal fluorescent laser microscopy (CFM). AFB uses bronchoscopy with an additional light source to screen high-risk individuals and identify premalignant lesions (airway dysplasia) and carcinoma in situ. NBI capitalizes on the increased absorption of blue and green wavelengths of light by hemoglobin to enhance the visibility of vessels of the mucosa and differentiate between inflammatory versus malignant mucosal lesions. CFM uses a blue laser to induce fluorescence, and its high degree of resolution provides a real-time view of living tissue at an almost histologic resolution. OCT uses near-infrared light source and has spatial resolution advantages over CT and MRI. It can penetrate the airway wall up to three times deeper than CFM and is less susceptible to motion artifacts from cardiac pulsation and respiratory movements. However, careful assessment is required before these methods find a place in the evaluation strategy of early lung cancer and other lung diseases.

MEDICAL THORACOSCOPY

Medical thoracoscopy (or pleuroscopy) focuses on the diagnosis of pleural-based problems. The procedure is performed with a conventional rigid or a semi-rigid pleuroscope (similar in design to a bronchoscope and enabling the operator to inspect the pleural surface, sample and/or drain pleural fluid, or perform targeted biopsies of the parietal pleura). Medical thoracoscopy can be performed in the endoscopy suite or operating room with the patient under conscious sedation and local anesthesia. In contrast, video-assisted thoracoscopic surgery (VATS) requires general anesthesia and is only performed in the operating room. A common diagnostic indication for medical thoracoscopy is the evaluation of a pleural effusion or biopsy of presumed parietal pleural carcinomatosis. It can also be used to place a chest tube under visual guidance, or perform chemical or talc pleurodesis, a therapeutic intervention to prevent a recurrent pleural effusion (usually malignant) or recurrent pneumothorax.

The increasing availability of advanced bronchoscopic and pleuroscopic techniques has motivated the development of IP programs. IP can be defined as “the art and science of medicine as related to the performance of diagnostic and invasive therapeutic procedures, that which require additional training

and expertise beyond that which required in a standard pulmonary medicine training program.” IP physicians provide alternatives to surgery for patients with a wide variety of thoracic disorders and problems, including therapeutic interventions (see further reading).

SURGICAL TECHNIQUES FOR OBTAINING BIOLOGIC SPECIMENS

Evaluation and diagnosis of disorders of the chest commonly involve collaboration between pulmonologists and thoracic surgeons. Although procedures such as mediastinoscopy, VATS, and thoracotomy are performed by thoracic surgeons, there is overlap in many minimally invasive techniques that can be performed by a pulmonologist, an interventional pulmonologist, or a thoracic surgeon.

MEDIASTINOSCOPY AND MEDIASTINOTOMY

Proper staging of lung cancer is of paramount concern when determining a treatment regimen. Although CT and PET scanning are useful for determining the size and nature of mediastinal lymph nodes as part of the staging of lung cancer, tissue biopsy and histopathologic examination are often critical for the diagnosis of mediastinal masses or enlarged mediastinal lymph nodes. The two major surgical procedures used to obtain specimens from masses or nodes in the mediastinum are mediastinoscopy (via a suprasternal approach) and mediastinotomy (via a parasternal approach). Both procedures are performed under general anesthesia by a qualified surgeon. In the case of suprasternal mediastinoscopy, a rigid mediastinoscope is inserted at the suprasternal notch and passed into the mediastinum along a pathway just anterior to the trachea. Tissue can be obtained with biopsy forceps passed through the scope, sampling masses or nodes that are in a paratracheal or pretracheal position (levels 2R, 2L, 3, 4R, 4L). Aortopulmonary lymph nodes (levels 5, 6) are not accessible by this route and thus are commonly sampled by parasternal mediastinotomy (the Chamberlain procedure). This approach involves a parasternal incision and dissection directly down to a mass or node that requires biopsy.

As an alternative to surgery, a bronchoscope can be used to perform TBNA to obtain tissue from the mediastinum, and, when combined with EBUS, can allow access to the same lymph node stations associated with mediastinoscopy, but also extend access out to the hilar lymph nodes (levels 10, 11). Finally, endoscopic ultrasound (EUS)–fine-needle aspiration (FNA) is a second procedure that complements EBUS-FNA in the staging of lung cancer. EUS-FNA is performed via the esophagus and is ideally suited for sampling lymph nodes in the posterior mediastinum (levels 7, 8, 9). Because US imaging cannot penetrate air-filled spaces, the area directly anterior to the trachea cannot accurately be assessed and is a “blind spot” for EUS-FNA. However, EBUS-FNA can visualize the anterior lymph nodes and can complement EUS-FNA. The combination of EUS-FNA and EBUS-FNA with the use of radial probes has made these techniques a clear nonoperative alternative for staging the mediastinum in thoracic malignancies.

VIDEO-ASSISTED THORACOSCOPIC SURGERY

VATS is the operative technique for the diagnosis and management of pleural as well as parenchymal lung disease. This procedure is performed in the operating room using single-lung ventilation with double-lumen endotracheal intubation and involves the passage of a rigid scope with a distal lens through a trocar inserted into the pleura. A high-quality image is shown on a monitor screen, allowing the operator to manipulate instruments passed into the pleural space through separate small intercostal incisions. With these instruments the operator can biopsy lesions of the pleura under direct visualization. In addition, this procedure is now used commonly to biopsy peripheral lung tissue or to remove peripheral nodules for both diagnostic and therapeutic purposes. This much less invasive

procedure has largely supplanted the traditional “open lung biopsy” performed via thoracotomy. The decision to use medical thorascopy versus VATS technique is based on the clinical scenario with input from the consulting pulmonary and thoracic surgery providers. If a surgical technique is preferred, the decision to use a VATS technique versus performing an open thoracotomy is made by the thoracic surgeon and based on whether a patient can tolerate the single-lung ventilation that is required to allow adequate visualization of the lung. With further advances in instrumentation and experience, VATS can be used to perform procedures previously requiring thoracotomy, including stapled lung biopsy, resection of pulmonary nodules, lobectomy, pneumonectomy, pericardial window, or other standard thoracic surgical procedures, but allows them to be performed in a minimally invasive manner.

THORACOTOMY

Although frequently replaced by VATS, thoracotomy remains an option for the diagnostic sampling of lung tissue. It provides the largest amount of material, and it can be used to biopsy and/or excise lesions that are too deep or too close to vital structures for removal by VATS. The choice between VATS and thoracotomy needs to be made on a case-by-case basis.

TECHNIQUES ON BIOLOGIC SPECIMENS

Histopathologic examination of tissue samples and cytologic examination of aspirates or fluid are critical components in the diagnosis of many respiratory disorders. In the area of lung cancer, improved understanding of molecular changes and genetic mutations that drive cancer has allowed development of specific molecular tests that guide therapy (e.g., epidermal growth factor receptor [EGFR] mutations and anaplastic lymphoma kinase [ALK] fusions).

FURTHER READING

BAHMER T et al: The use of auto-antibody testing in the evaluation of interstitial lung disease (ILD)—A practical approach for the pulmonologist. Respir Med 113:80, 2016. [PubMed: 26921132]
FLOHR TG, SCHMIDT B: CT technology for imaging the thorax: State of the art, in *Multidetector-Row CT of the Thorax*, *Medical Radiology*, Schoepf UJ, Meinel FG (eds). Springer International Publishing, 2016, pp. 3–28.
HIRSCH FR et al: New and emerging targeted treatments in advanced non-small-cell lung cancer. Lancet 388(10048):1012, 2016. [PubMed: 27598681]
HOFFMAN EA et al: For the IWPFI Investigators. Pulmonary CT and MRI phenotypes that help explain chronic pulmonary obstruction disease pathophysiology and outcomes J Magn Reson Imaging 43:544, 2016. [PubMed: 26199216]
IRWIN Z, COOK JO: Advances in point-of-care thoracic ultrasound. Emerg Med Clin North Am 34:151, 2016. [PubMed: 26614246]
MEYER KC et al: American Thoracic Society Committee on BAL in Interstitial Lung Disease. An official American Thoracic Society clinical practice guideline: The clinical utility of bronchoalveolar lavage cellular analysis in interstitial lung disease. Am J Respir Crit Care Med 185:1004, 2012. [PubMed: 22550210]
MIRSADRAEE S, VAN BEEK EJ: Functional imaging: Computed tomography and MRI. Clin Chest Med 36:349, 2015. [PubMed: 26024609]
THE NATIONAL LUNG SCREENING TRIAL RESEARCH TEAM: Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365:395, 2011. [PubMed: 21714641]
WALTERS DM, WOOD DE: Operative endoscopy of the airway. J Thorac Dis 8(Suppl 2):S130, 2016. [PubMed: 26981263]

Chapter 281: Asthma

Peter J. Barnes

INTRODUCTION

Asthma is a syndrome characterized by airflow obstruction that varies markedly, both spontaneously and with treatment. Asthmatics harbor a special type of inflammation in the airways that makes them more responsive than nonasthmatics to a wide range of triggers, leading to excessive narrowing with consequent reduced airflow and symptomatic wheezing and dyspnea. Narrowing of the airways is usually reversible, but in some patients with chronic asthma there may be an element of irreversible airflow obstruction. Asthma is a heterogeneous disease with several phenotypes recognized, but thus far these do not correspond well to specific pathogenic mechanisms (endotypes) or responses to therapy. The increasing global prevalence of asthma, the large burden it now imposes on patients, and the high health care costs have led to extensive research into its mechanisms and treatment.

PREVALENCE

Asthma is one of the most common chronic diseases globally and currently affects ~300 million people worldwide, with ~250,000 deaths annually. The prevalence of asthma has risen in affluent countries over the last 30 years but now appears to have stabilized, with ~10–12% of adults and 15% of children affected by the disease. In developing countries where the prevalence of asthma had been much lower, there is a rising prevalence, which is associated with increased urbanization. The prevalence of atopy and other allergic diseases has also increased over the same time, suggesting that the reasons for the increase are likely to be systemic rather than confined to the lungs. Most patients with asthma in affluent countries are atopic, with allergic sensitization to the house dust mite *Dermatophagoides pteronyssinus* and other environmental allergens, such as animal fur and pollens.

Asthma can present at any age, with a peak age of 3 years. In childhood, twice as many males as females are asthmatic, but by adulthood the sex ratio has equalized. Long-term studies that have followed children until they reach the age of 40 years suggest that many with asthma become asymptomatic during adolescence but that asthma returns in some during adult life, particularly in those with persistent symptoms and severe asthma. Adults with asthma, including those with onset during adulthood, rarely become permanently asymptomatic. The severity of asthma does not vary significantly within a given patient; those with mild asthma rarely progress to more severe disease, whereas those with severe asthma usually have severe disease at the onset.

Deaths from asthma are relatively uncommon, and in many affluent countries have been steadily declining over the last decade. A rise in asthma mortality seen in several countries during the 1960s was associated with increased use of short-acting inhaled β_2-adrenergic agonists (as rescue therapy), but there is now compelling evidence that the more widespread use of inhaled corticosteroids (ICS) in patients with persistent asthma is responsible for the decrease in mortality in recent years. Major risk factors for asthma deaths are poorly controlled disease with frequent use of bronchodilator inhalers, lack of or poor compliance with ICS therapy, and previous admissions to hospital with near-fatal asthma.

It has proved difficult to agree on a definition of asthma, but there is good agreement on the description of the clinical syndrome and disease pathology. Until the etiologic mechanisms of the disease are better

understood, it will be difficult to provide an accurate definition.

RISK FACTORS AND TRIGGERS

Asthma is a heterogeneous disease with interplay between genetic and environmental factors. Several risk factors that predispose to asthma have been identified **(Table 281-1)**. These should be distinguished from triggers, which are environmental factors that worsen asthma in a patient with established asthma.

TABLE 281-1Risk Factors and Triggers Involved in Asthma

ENDOGENOUS FACTORS	ENVIRONMENTAL FACTORS
	Indoor allergens
Genetic predisposition	Outdoor allergens
Atopy	Occupational sensitizers
Airway hyperresponsiveness	Passive smoking
Gender	Respiratory infections
Ethnicity	Air pollution (diesel particulates, nitrogen oxides)
Obesity	Diet
Early viral infections	Dampness and mold exposure
	Acetaminophen (paracetamol)
Triggers	
Allergens	
Upper respiratory tract viral infections	
Exercise and hyperventilation	
Cold air	
Sulfur dioxide and irritant gases	
Drugs (β-blockers, aspirin)	
Stress	
Irritants (household sprays, paint fumes)	

Atopy

Atopy is the major risk factor for asthma, and non-atopic individuals have a very low risk of developing asthma. Patients with asthma commonly suffer from other atopic diseases, particularly allergic rhinitis, which may be found in >80% of asthmatic patients, and atopic dermatitis (eczema). Atopy may be found in 40–50% of the population in affluent countries, but only a proportion of atopic individuals becoming asthmatic. This observation suggests that some other environmental or genetic factor(s) predispose to the development of asthma in atopic individuals. The allergens that lead to sensitization are usually proteins that have protease activity, and the most common allergens are derived from house dust mites, cat and dog fur, cockroaches (in inner cities), grass and tree pollens, and rodents (in laboratory workers).

Atopy is due to the genetically determined production of specific IgE antibody, with many patients showing a family history of allergic diseases.

Genetic Predisposition



The familial association of asthma and a high degree of concordance for asthma in identical twins indicate a genetic predisposition to the disease; however, whether or not the genes predisposing to asthma are similar or in addition to those predisposing to atopy is not yet clear. It now seems likely that different genes may also contribute to asthma specifically, and there is increasing evidence that the severity of asthma is also genetically determined. Genetic screens with classical linkage analysis and single-nucleotide polymorphisms of various candidate genes indicate that asthma is polygenic, with each gene identified having a small effect that is often not replicated in different populations. This observation suggests that the interaction of many genes is important, and these may differ in different populations. The most consistent findings have been associations with polymorphisms of genes on chromosome 5q, including the T helper 2 (T_H2) cells interleukin (IL)-4, IL-5, IL-9, and IL-13, which are associated with atopy. There is increasing evidence for a complex interaction between genetic polymorphisms and environmental factors that will require very large population studies to unravel. Novel genes that have been associated with asthma, including *ADAM-33*, *DPP-10*, and *ORMDL3*, have also been identified by positional cloning, but their function in disease pathogenesis is not yet clear. Recent genome-wide association studies have identified further novel genes, such as ORMDL3, although their functional role is not yet clear. Genetic polymorphisms may also be important in determining the response to asthma therapy. For example, the Arg-Gly-16 variant in the β_2-receptor has been associated with reduced response to β_2-agonists, and repeats of an Sp1 recognition sequence in the promoter region of 5-lipoxygenase may affect the response to antileukotrienes. However, these effects are small and inconsistent and do not yet have any implications for asthma therapy.

It is likely that environmental factors in early life determine which atopic individuals become asthmatic. The increasing prevalence of asthma, particularly in developing countries, over the last few decades also indicates the importance of environmental mechanisms interacting with a genetic predisposition.

Epigenetic Mechanisms

There is increasing evidence that epigenetic mechanisms may be important, particularly in the early development of asthma. DNA methylation and histone modification patterns may be influenced by diet, cigarette smoke exposure, and air pollution, and may affect genes involved in the pathogenesis of asthma. These epigenetic changes may occur in the fetus as a result of maternal environmental exposure.

Infections

Although viral infections (especially Rhinovirus) are common as triggers of asthma exacerbations, it is uncertain whether they play a role in etiology. There is some association between respiratory syncytial virus infection in infancy and the development of asthma, but the specific pathogenesis is difficult to elucidate, as this infection is very common in children. Atypical bacteria, such as *Mycoplasma* and *Chlamydophila*, have been implicated in the mechanism of severe asthma, but thus far, the evidence is not very convincing of a true association. Living in damp houses with exposure to mold spores is now recognized to be a risk factor, and removal of these factors may improve asthma.

The observation that allergic sensitization and asthma were less common in children with older siblings first suggested that lower levels of infection may be a factor in affluent societies that increase the risks of asthma. This “hygiene hypothesis” proposes that lack of infections in early childhood preserves the T_H2 cell bias at birth, whereas exposure to infections and endotoxin results in a shift toward a predominant

protective T_H1 immune response. Children brought up on farms who are exposed to a high level of endotoxin are less likely to develop allergic sensitization than children raised on dairy farms. Intestinal parasite infection, such as hookworm, may also be associated with a reduced risk of asthma. While there is considerable epidemiologic support for the hygiene hypothesis, it cannot account for the parallel increase in T_H1-driven diseases such as diabetes mellitus over the same period.

Diet

The role of dietary factors is controversial. Observational studies have shown that diets low in antioxidants such as vitamin C and vitamin A, magnesium, selenium, and omega-3 polyunsaturated fats (fish oil) or high in sodium and omega-6 polyunsaturates are associated with an increased risk of asthma. Vitamin D deficiency may also predispose to the development of asthma. However, interventional studies with supplementary diets have not supported an important role for these dietary factors. Obesity is also an independent risk factor for asthma, particularly in women, but the mechanisms are not yet clear.

Air Pollution

Air pollutants such as sulfur dioxide, ozone, and diesel particulates may trigger asthma symptoms, but the role of different air pollutants in the etiology of the disease is not yet clear. Asthma had a much lower prevalence in East Germany compared to West Germany despite a much higher level of air pollution, but since reunification these differences have decreased as Eastern Germany has become more affluent. There is increasing evidence that exposure to road traffic pollution is associated with increased asthma symptoms, with the main culprits being diesel particulates and nitrogen dioxide. Indoor air pollution is also important with exposure to nitrogen oxides from cooking stoves and exposure to passive cigarette smoke. There is some evidence that maternal smoking is a risk factor for asthma, but it is difficult to dissociate this association from an increased risk of respiratory infections.

Allergens

Inhaled allergens are common triggers of asthma symptoms and have also been implicated in allergic sensitization. Exposure to house dust mites in early childhood is a risk factor for allergic sensitization and asthma, but rigorous allergen avoidance has not shown any evidence for a reduced risk of developing asthma. The increase in house dust mites in centrally heated poorly ventilated homes with fitted carpets has been implicated in the increasing prevalence of asthma in affluent countries. Domestic pets, particularly cats, have also been associated with allergic sensitization, but early exposure to cats in the home may be protective through the induction of tolerance.

Occupational Exposure

Occupational asthma is relatively common and may affect up to 10% of young adults. Over 300 sensitizing agents have been identified. Chemicals such as toluene diisocyanate and trimellitic anhydride, may lead to sensitization independent of atopy. Individuals may also be exposed to allergens in the workplace such as small animal allergens in laboratory workers and fungal amylase in wheat flour in bakers. Cleaners commonly develop occupational asthma owing to exposure to aerosols of cleaning liquids. Occupational asthma may be suspected when symptoms improve during weekends and holidays.

Obesity

Asthma occurs more frequently in obese people (BMI >30 kg/m^2) and is often more difficult to control. Although mechanical factors may contribute, it may also be linked to the pro-inflammatory adipokines and reduced anti-inflammatory adipokines that are released from fat cells.

Other Factors

Several other factors have been implicated in the etiology of asthma, including lower maternal age, duration of breast-feeding, prematurity and low birthweight, and inactivity, but are unlikely to contribute to the recent global increase in asthma prevalence. There is also an association with acetaminophen (paracetamol) consumption in childhood, which may be linked to increased oxidative stress.

Intrinsic Asthma

A minority of asthmatic patients (~10%) have negative skin tests to common inhalant allergens and normal serum concentrations of IgE. These patients, with non-atopic or intrinsic asthma, usually show later onset of disease (adult-onset asthma), commonly have concomitant nasal polyps, and may be aspirin-sensitive. They usually have more severe, persistent asthma. Little is understood about mechanism, but the immunopathology in bronchial biopsies and sputum appears to be identical to that found in atopic asthma. There is recent evidence for increased local production of IgE in the airways, suggesting that there may be common IgE-mediated mechanisms; staphylococcal enterotoxins, which serve as “superantigens,” have been implicated. Type-2 innate lymphoid cells (ILC2) may drive the eosinophilic inflammation in these non-allergic patients.

Asthma Triggers

Several stimuli trigger airway narrowing, wheezing, and dyspnea in asthmatic patients. While the previous view held that these should be avoided, it is now seen as evidence for poor control and an indicator of the need to increase controller (preventive) therapy.

Allergens

Inhaled allergens activate mast cells with bound IgE directly leading to the immediate release of bronchoconstrictor mediators, resulting in the early response that is reversed by bronchodilators. Often, an experimental allergen challenge is followed by a late response when there is airway edema and an acute inflammatory response with increased eosinophils and neutrophils that are not very reversible with bronchodilators. The most common allergens to trigger asthma are *Dermatophagoides* species, and environmental exposure leads to low-grade chronic symptoms that are perennial. Other perennial allergens are derived from cats and other domestic pets, as well as cockroaches. Other allergens, including grass pollen, ragweed, tree pollen, and fungal spores, are seasonal. Pollens usually cause allergic rhinitis rather than asthma, but in thunderstorms the pollen grains are disrupted and the particles that may be released can trigger severe asthma exacerbations (thunderstorm asthma).

Virus Infections

Upper respiratory tract virus infections such as rhinovirus, respiratory syncytial virus, and coronavirus are the most common triggers of acute severe exacerbations and may invade epithelial cells of the lower as well as the upper airways. The mechanism whereby these viruses cause exacerbations is poorly understood, but there is an increase in airway inflammation with increased numbers of eosinophils and neutrophils. There is evidence for reduced production of type I interferons by epithelial cells from asthmatic patients, resulting in increased susceptibility to these viral infections and a greater inflammatory response.

Pharmacologic Agents

Several drugs may trigger asthma. Beta-adrenergic blockers commonly acutely worsen asthma, and their use may be fatal. The mechanisms are not clear but are likely mediated through increased cholinergic bronchoconstriction. All beta blockers need to be avoided and even selective β, β_2 blockers, or topical application (e.g., timolol eye drops) may be dangerous. Angiotensin-converting enzyme inhibitors are theoretically detrimental as they inhibit breakdown of kinins, which are bronchoconstrictors; however, they rarely worsen asthma, and the characteristic cough is no more frequent in asthmatics than in non-

asthmatics. Aspirin may worsen asthma in some patients (aspirin-sensitive asthma is discussed under “Special Considerations”).

Exercise

Exercise is a common trigger of asthma, particularly in children. The mechanism is linked to hyperventilation, which results in increased osmolality in airway lining fluid and triggers mast cell mediator release, resulting in bronchoconstriction. Exercise-induced asthma (EIA) typically begins after exercise has ended, and recovers spontaneously within about 30 min. EIA is worse in cold, dry climates than in hot, humid conditions. It is, therefore, more common in sports such as cross-country running in cold weather, overland skiing, and ice hockey than in swimming. It may be prevented by prior administration of β_2-agonists and antileukotrienes, but is best prevented by regular treatment with ICS, which reduce the population of surface mast cells required for this response.

Physical Factors

Cold air and hyperventilation may trigger asthma through the same mechanisms as exercise. Laughter may also be a trigger. Many patients report worsening of asthma in hot weather and when the weather changes. Some asthmatics become worse when exposed to strong smells or perfumes, but the mechanism of this response is uncertain.

Food and Diet

There is little evidence that allergic reactions to food lead to increased asthma symptoms, despite the belief of many patients that their symptoms are triggered by particular food constituents. Exclusion diets are usually unsuccessful at reducing the frequency of episodes. Some foods such as shellfish and nuts may induce anaphylactic reactions that may include wheezing. Patients with aspirin-induced asthma may benefit from a salicylate-free diet, but these are difficult to maintain. Certain food additives may trigger asthma. Metabisulfite, which is used as a food preservative, may trigger asthma through the release of sulfur dioxide gas in the stomach. Tartrazine, a yellow food-coloring agent, was believed to be a trigger for asthma, but there is little convincing evidence for this.

Air Pollution

Increased ambient levels of sulfur dioxide, ozone, diesel particulates and nitrogen oxides are associated with increased asthma symptoms.

Occupational Factors

Several substances found in the workplace may act as sensitizing agents, as discussed above, but may also act as triggers of asthma symptoms. Occupational asthma is characteristically associated with symptoms at work with relief on weekends and holidays. If removed from exposure within the first 6 months of symptoms, there is usually complete recovery. More persistent symptoms lead to irreversible airway changes, and, thus, early detection and avoidance are important.

Hormones

Some women show premenstrual worsening of asthma, which can occasionally be very severe. The mechanisms are not completely understood, but are related to a fall in progesterone and in severe cases may be improved by treatment with high doses of progesterone or gonadotropin-releasing factors. Thyrotoxicosis and hypothyroidism can both worsen asthma, although the mechanisms are uncertain.

Gastroesophageal Reflux

Gastroesophageal reflux is common in asthmatic patients as it is increased by bronchodilators. Although acid reflux might trigger reflex bronchoconstriction, it rarely causes asthma symptoms, and antireflux therapy usually fails to reduce asthma symptoms in most patients.

Stress

Many asthmatics report worsening of symptoms with stress. Psychological factors can induce bronchoconstriction through cholinergic reflex pathways. Paradoxically, very severe stress such as bereavement usually does not worsen, and may even improve, asthma symptoms.

PATHOPHYSIOLOGY

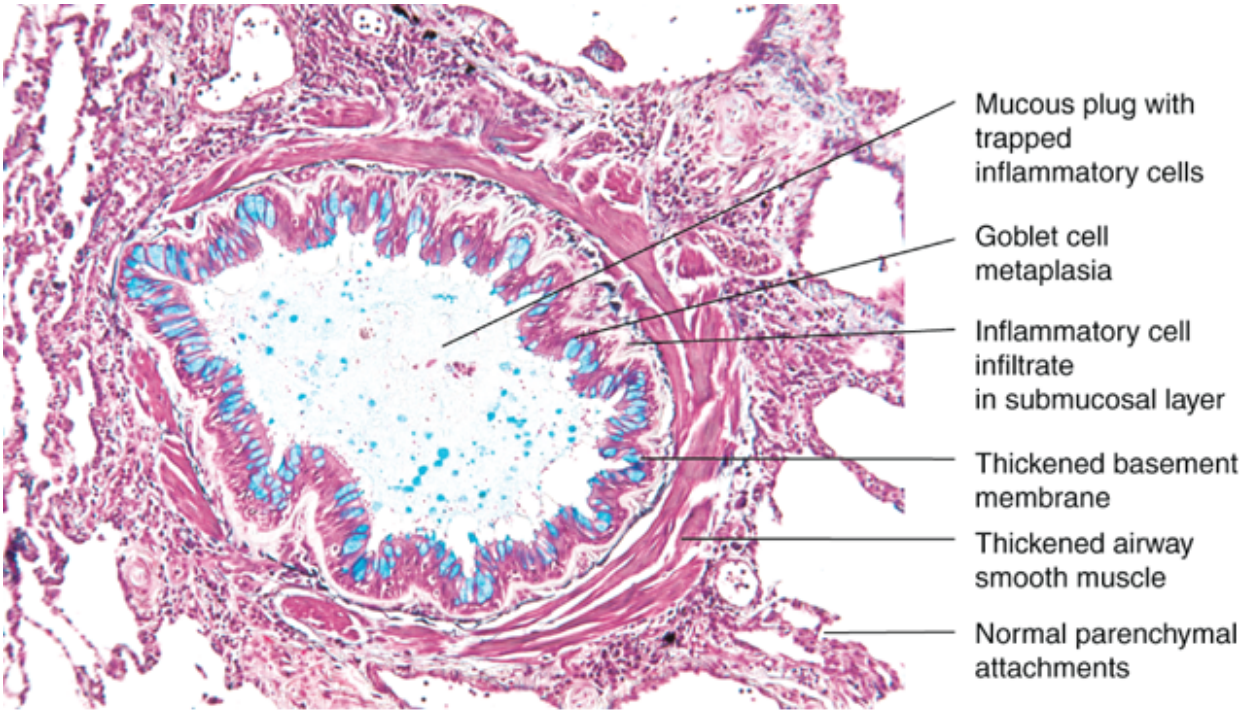
Asthma is associated with a specific chronic inflammation of the mucosa of the lower airways. One of the main aims of treatment is to reduce this inflammation.

Pathology

The pathology of asthma has been revealed through examining the lungs of patients who have died of asthma and from bronchial biopsies. The airway mucosa is infiltrated with activated eosinophils and T lymphocytes, and there is activation of mucosal mast cells. The degree of inflammation is poorly related to disease severity and may even be found in atopic patients without asthma symptoms. This inflammation is usually reduced by treatment with ICS. There are also structural changes in the airways (often termed remodeling). A characteristic finding is thickening of the basement membrane due to subepithelial collagen deposition. This feature is also found in patients with eosinophilic bronchitis presenting as cough who do not have asthma and is, therefore, likely to be a marker of eosinophilic inflammation in the airway as eosinophils release fibrogenic mediators. The epithelium is often shed or friable, with reduced attachments to the airway wall and increased numbers of epithelial cells in the lumen. The airway wall itself may be thickened and edematous, particularly in fatal asthma. Another common finding in fatal asthma is occlusion of the airway lumen by a mucous plug, which is comprised of mucous glycoproteins secreted from goblet cells and plasma proteins from leaky bronchial vessels **(Fig. 281-1)**. There is also vasodilation and increased numbers of blood vessels (angiogenesis). Direct observation by bronchoscopy indicates that the airways may be narrowed, erythematous, and edematous. The pathology of asthma is remarkably uniform in different phenotypes of asthma, including atopic (extrinsic), non-atopic (intrinsic), occupational, aspirin-sensitive, and pediatric asthma. These pathologic changes are found in all airways, but do not extend to the lung parenchyma; peripheral airway inflammation is found particularly in patients with severe asthma. The involvement of airways may be patchy and this is consistent with bronchographic findings of uneven narrowing of the airways.

FIGURE 281-1

Histopathology of a small airway in fatal asthma. The lumen is occluded with a mucous plug, there is goblet cell metaplasia, and the airway wall is thickened, with an increase in basement membrane thickness and airway smooth muscle. *(Courtesy of Dr. J. Hogg, University of British Colombia.)*



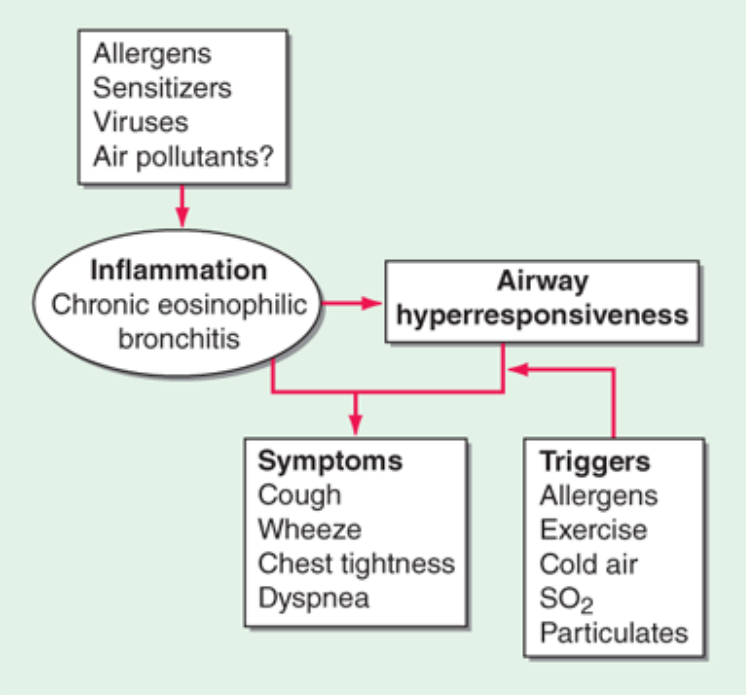
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Airway Inflammation

There is inflammation in the respiratory mucosa from the trachea to terminal bronchioles, but with a predominance in the bronchi (cartilaginous airways), but it is still uncertain how inflammatory cells interact and how inflammation translates into the symptoms of asthma **(Fig. 281-2)**. There is good evidence that the specific pattern of airway inflammation in asthma is associated with airway hyperresponsiveness (AHR), the physiologic abnormality of asthma, which is correlated with variable airflow obstruction. The pattern of inflammation in asthma is characteristic of allergic diseases, with similar inflammatory cells seen in the nasal mucosa in rhinitis. However, an indistinguishable pattern of inflammation is found in intrinsic asthma, and this may reflect local rather than systemic IgE production. Although most attention has focused on the acute inflammatory changes seen in asthma, this is a chronic condition, with inflammation persisting over many years in most patients. The mechanisms involved in persistence of inflammation in asthma are still poorly understood. Superimposed on this chronic inflammatory state are acute inflammatory episodes, which correspond to exacerbations of asthma. Although the common pattern of inflammation in asthma is characterized by eosinophil infiltration, some patients with severe asthma show a neutrophilic pattern of inflammation that is less sensitive to corticosteroids. However, many inflammatory cells are involved in asthma with no key cell that is predominant **(Fig. 281-3)**.

FIGURE 281-2

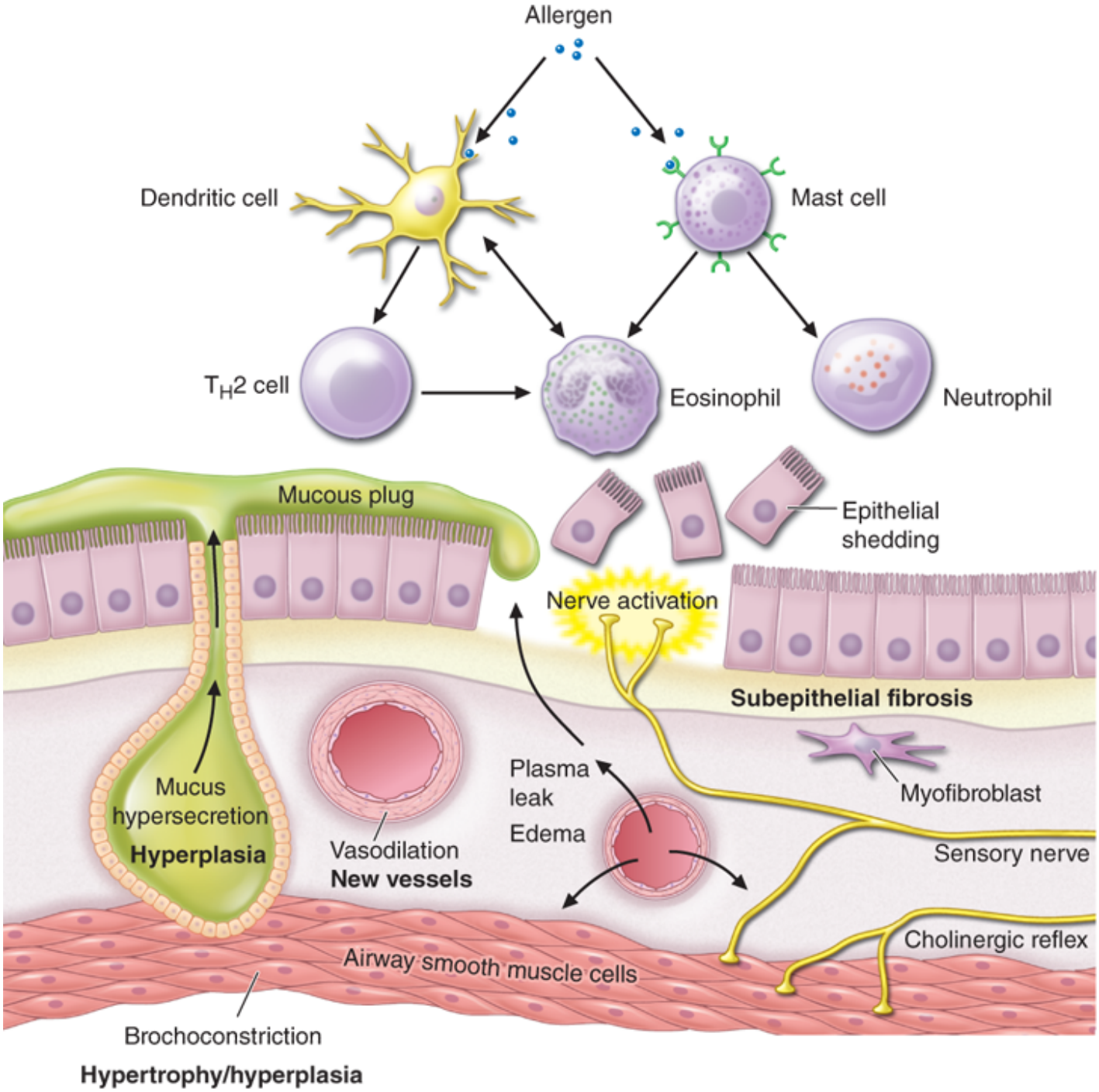
Inflammation in the airways of asthmatic patients leads to airway hyperresponsiveness and symptoms. SO_2, sulfur dioxide.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 281-3

The pathophysiology of asthma is complex with participation of several interacting inflammatory cells, which result in acute and chronic inflammatory effects on the airway.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Mast Cells

Mast cells are important in initiating the acute bronchoconstrictor responses to allergens and several other indirectly acting stimuli, such as exercise and hyperventilation (via osmolality changes), as well as fog. Activated mucosal mast cells are found at the airway surface in asthma patients and also in the airway smooth-muscle layer, whereas this is not seen in normal subjects or patients with eosinophilic bronchitis. Mast cells are activated by allergens through an IgE-dependent mechanism, and binding of specific IgE to mast cells renders them more sensitive to activation by physical stimuli such as osmolality. The importance of IgE in the pathophysiology of asthma has been highlighted by clinical studies with humanized anti-IgE antibodies, which inhibit IgE-mediated effects, reduce asthma symptoms, and reduce exacerbations. There are, however, uncertainties about the role of mast cells in more chronic allergic

inflammatory events. Mast cells release several bronchoconstrictor mediators, including histamine, prostaglandin D_2, and cysteinyl-leukotrienes, but also several cytokines, chemokines, growth factors, and neurotrophins.

Macrophages and Dendritic Cells

Macrophages, which are derived from blood monocytes, may traffic into the airways in asthma and may be activated by allergens via low-affinity IgE receptors ($Fc_{\varepsilon}RII$). Macrophages have the capacity to initiate a type of inflammatory response via the release of a certain pattern of cytokines, but these cells also release anti-inflammatory mediators (e.g., IL-10) and, thus, their roles in asthma are uncertain. Dendritic cells are specialized macrophage-like cells in the airway epithelium, which are the major antigen-presenting cells. Dendritic cells take up allergens, process them to peptides, and migrate to local lymph nodes where they present the allergenic peptides to uncommitted T lymphocytes to program the production of allergen-specific T cells. Immature dendritic cells in the respiratory tract promote T_H2 cell differentiation and require cytokines such as IL-12 and tumor necrosis factor α (TNF-α), to promote the normally preponderant T_H1 response. The cytokine thymic stromal lymphopoietin (TSLP) released from epithelial cells in asthmatic patients instructs dendritic cells to release chemokines that attract T_H2 cells into the airways.

Eosinophils

Eosinophil infiltration is a characteristic feature of asthmatic airways. Allergen inhalation results in a marked increase in activated eosinophils in the airways at the time of the late reaction. Eosinophils are linked to the development of AHR through the release of basic proteins and oxygen-derived free radicals. Eosinophil recruitment involves adhesion of eosinophils to vascular endothelial cells in the airway circulation due to interaction between adhesion molecules, migration into the submucosa under the direction of chemokines, and their subsequent activation and prolonged survival. Blocking antibodies to IL-5 causes a profound and prolonged reduction in circulating and sputum eosinophils, but is not associated with reduced AHR or asthma symptoms, although in selected patients with steroid-resistant airway eosinophils, there is a reduction in exacerbations. Eosinophils may be important in release of growth factors involved in airway remodeling and in exacerbations but probably not in AHR.

Neutrophils

Increased numbers of activated neutrophils are found in sputum and airways of some patients with severe asthma and during exacerbations, although there is a proportion of patients even with mild or moderate asthma who have a predominance of neutrophils. The roles of neutrophils in asthma that are resistant to the anti-inflammatory effects of corticosteroids are currently unknown.

Lymphocytes

T lymphocytes play a very important role in coordinating the inflammatory response in asthma through the release of specific patterns of cytokines, resulting in the recruitment and survival of eosinophils and in the maintenance of a mast cell population in the airways. The naïve immune system and the immune system of asthmatics are skewed to express the T_H2 phenotype, whereas in normal airways T_H1 cells predominate. T_H2 cells, through the release of IL-5, are associated with eosinophilic inflammation and, through the release of IL-4 and IL-13, are associated with increased IgE formation. Natural killer $CD4^+$ T lymphocytes that express high levels of IL-4 have been described in some studies. Regulatory T cells (Treg) play an important role in determining the expression of other T cells, and there is evidence for a reduction in a certain subset of Tregs ($CD4^+CD25^+$) that express the transcription factor FOXP3 in asthma that is associated with increased T_H2 cells. Recently innate T cells (ILC2) without T cell receptors have been identified that release T_H2 cytokines and are regulated by epithelial cytokines such as IL-25 and IL-33 and may be predominant in non-allergic asthma.

Structural Cells

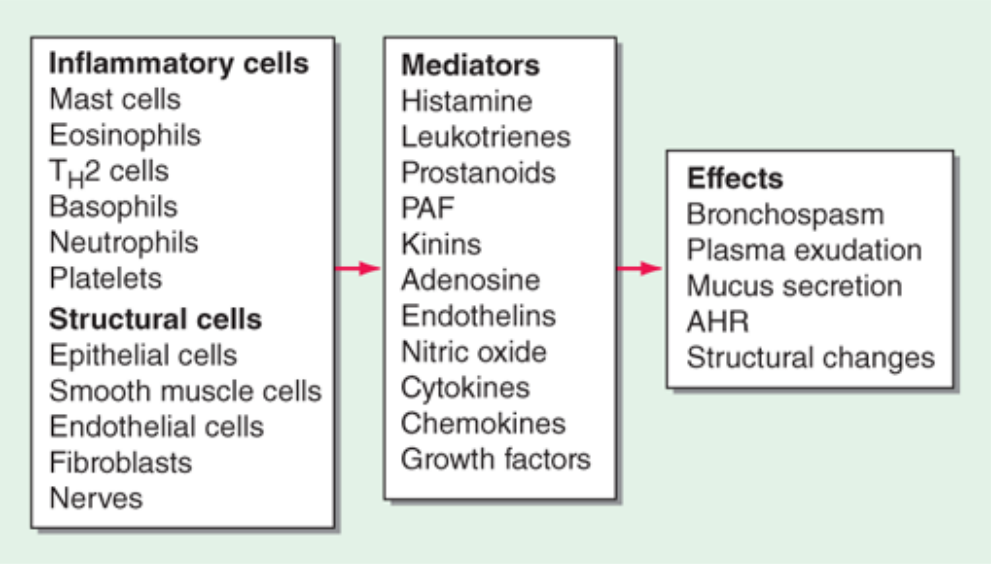
Structural cells of the airways, including epithelial cells, fibroblasts, and airway smooth-muscle cells, are also important sources of inflammatory mediators such as cytokines and lipid mediators, in asthma. Indeed, because structural cells far outnumber inflammatory cells, they may become the major sources of mediators driving chronic inflammation in asthmatic airways. In addition, epithelial cells may have key roles in translating inhaled environmental signals into an airway inflammatory response, and are probably major target cells for ICS.

Inflammatory Mediators

Multiple inflammatory mediators have been implicated in asthma, and they may have a variety of effects on the airways that account for the pathologic features of asthma (**Fig. 281-4**). Mast cell-derived mediators, such as histamine, prostaglandin D_2, and cysteinyl-leukotrienes, contract airway smooth muscle, increase microvascular leakage, increase airway mucus secretion, and attract other inflammatory cells. Because each mediator has many effects, the role of individual mediators in the pathophysiology of asthma is not yet clear. Although the multiplicity of mediators makes it unlikely that preventing the synthesis or action of a single mediator will have a major impact in clinical asthma, recent clinical studies with antileukotrienes suggest that cysteinyl-leukotrienes have clinically important effects.

FIGURE 281-4

Many cells and mediators are involved in asthma and lead to several effects on the airways. AHR, airway hyperresponsiveness; PAF, platelet-activating factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Cytokines

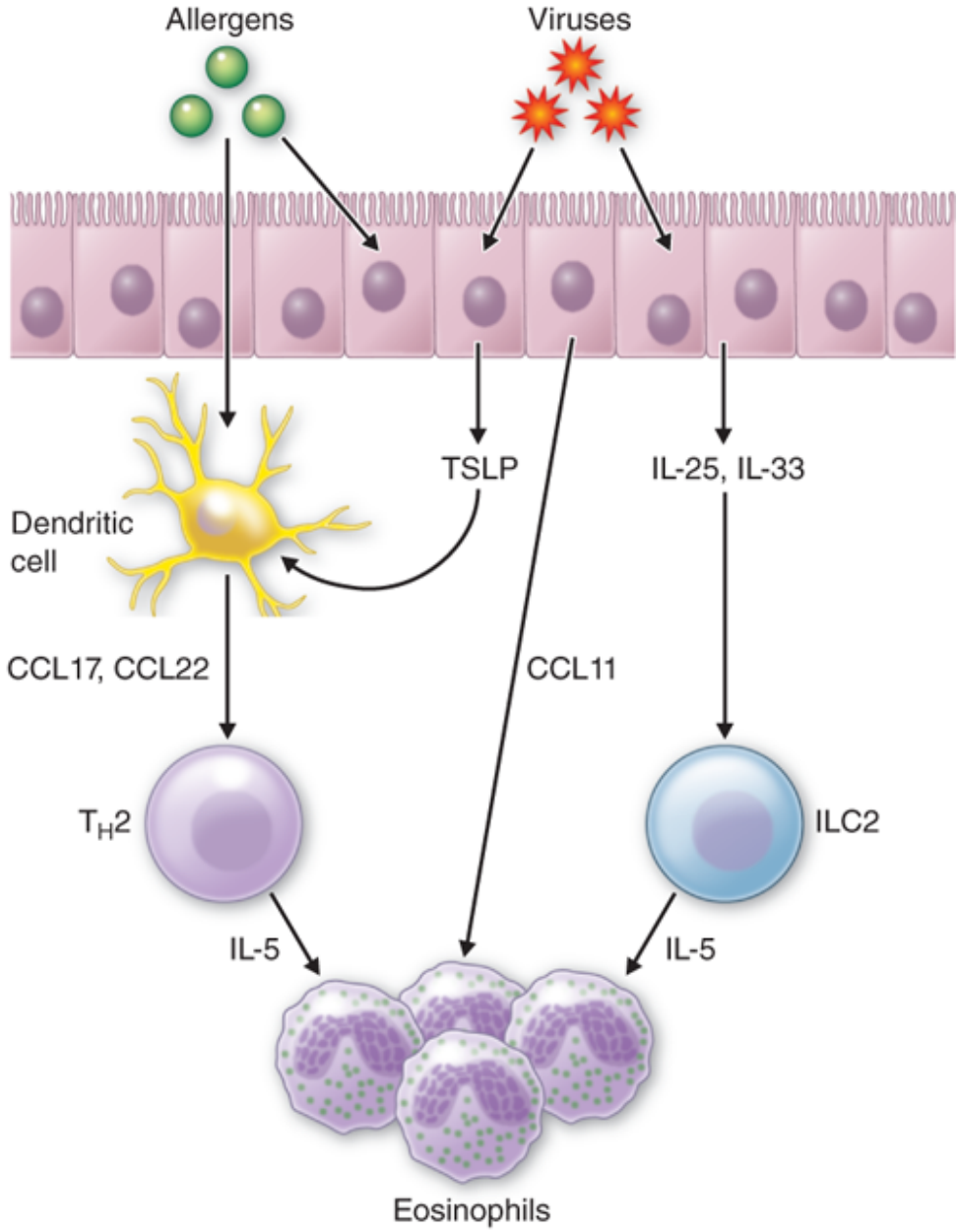
Multiple cytokines regulate the chronic inflammation of asthma. The T_H2 cytokines IL-4, IL-5, IL-9, and IL-13 mediate allergic inflammation, whereas proinflammatory cytokines such as TNF-α and IL-1β amplify the inflammatory response and play a role in more severe disease. TSLP is an upstream cytokine released from epithelial cells of asthmatics that orchestrates the release of chemokines that selectively attract T_H2 cells. Some cytokines such as IL-10 and IL-12 are anti-inflammatory and may be deficient in asthma.

Chemokines

Chemokines are involved in attracting inflammatory cells from the bronchial circulation into the airways. Eotaxin (CCL11) is selectively attractant to eosinophils via CCR3 and is expressed by epithelial cells of asthmatics, whereas CCL17 (TARC) and CCL22 (MDC) from epithelial cells attract T_H2 cells via CCR4 **(Fig. 281-5)**.

FIGURE 281-5

T lymphocytes in asthma. Allergen interacts with dendritic cells and releases thymus stimulated lymphopoietin (TSLP) which stimulate activated dendritic cells to release the chemokines CCL17 and CCL22, which attract T helper 2 (T_H2) lymphocytes. Allergens and viral infection may release interleukins (IL)-25 and -33 which recruit and activate type 2 innate lymphoid cells (ILC2). Both T_H2 and ILC2 cells release IL-5 and epithelial cells CCL11 (eotaxin), which together lead to recruitment of eosinophils into the airways.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Oxidative Stress

Activated inflammatory cells such as macrophages, eosinophils, and neutrophils produce reactive oxygen species. Evidence for increased oxidative stress in asthma is provided by the increased concentrations of 8-isoprostane (a product of oxidized arachidonic acid) in exhaled breath condensates and increased ethane (a product of lipid peroxidation) in the expired air of asthmatic patients. Increased oxidative stress is related to disease severity, it may amplify the inflammatory response, and may reduce responsiveness to corticosteroids.

Nitric Oxide

Nitric oxide (NO) is produced by NO synthases in several cells in the airway, particularly airway epithelial cells and macrophages. The level of NO in the expired air of patients with asthma is higher than normal and is related to the eosinophilic inflammation. Increased NO may contribute to the bronchial vasodilation

observed in asthma. Fractional exhaled NO (F_ENO) is increasingly used in the diagnosis and monitoring of asthmatic inflammation, although it is not yet used routinely in clinical practice.

Transcription Factors

Proinflammatory transcription factors such as nuclear factor-κB (NF-κB) and activator protein-1, are activated in asthmatic airways and orchestrate the expression of multiple inflammatory genes. More specific transcription factors that are involved include nuclear factor of activated T cells and GATA-3, which regulate the expression of T_H2 cytokines in T_H2 and ILC2 cells.

Effects of Inflammation

The chronic inflammatory response has several effects on the target cells of the airways, resulting in the characteristic pathophysiologic and remodeling changes associated with asthma. Asthma may be regarded as a disease with continuous inflammation and repair proceeding simultaneously, although the relationship between chronic inflammatory processes and asthma symptoms is often obscure.

Airway Epithelium

Airway epithelial shedding may be important in contributing to AHR and may explain how several mechanisms, such as ozone exposure, virus infections, chemical sensitizers, and allergens (usually proteases), can lead to its development, as all of these stimuli may lead to epithelial disruption. Epithelial damage may contribute to AHR in a number of ways, including loss of its barrier function to allow penetration of allergens; loss of enzymes (such as neutral endopeptidase/neprilysin) that degrade certain peptide inflammatory mediators like bradykinin; loss of a relaxant factor (so-called epithelial-derived relaxant factor); and exposure of sensory nerves, which may lead to reflex neural effects on the airway.

Fibrosis

In all asthmatic patients, the basement membrane is apparently thickened due to subepithelial fibrosis with deposition of types III and V collagen below the true basement membrane and is associated with eosinophil infiltration, presumably through the release of profibrotic mediators such as transforming growth factor-β. Mechanical manipulations can alter the phenotype of airway epithelial cells in a profibrotic fashion. In more severe patients, there is also fibrosis within the airway wall, which may contribute to irreversible narrowing of the airways.

Airway Smooth Muscle

In vitro airway smooth muscle from asthmatic patients usually shows no increased responsiveness to constrictors. Reduced responsiveness to β-agonists has also been reported in postmortem or surgically removed bronchi from asthmatics, although the number of β-receptors is not reduced, suggesting that β-receptors have been uncoupled. These abnormalities of airway smooth muscle may be secondary to the chronic inflammatory process. Inflammatory mediators may modulate the ion channels that serve to regulate the resting membrane potential of airway smooth-muscle cells, thus altering the level of excitability of these cells. In asthmatic airways there is also a characteristic hypertrophy and hyperplasia of airway smooth muscle, which is presumably the result of stimulation of airway smooth-muscle cells by various growth factors such as platelet-derived growth factor (PDGF) or endothelin-1 released from inflammatory or epithelial cells. Airway smooth-muscle cells from asthmatic patients also release multiple inflammatory mediators, particularly cytokines and chemokines.

Vascular Responses

There is increased airway mucosal blood flow in asthma, which may contribute to airway narrowing. There is an increase in the number of blood vessels in asthmatic airways as a result of angiogenesis in response to growth factors, particularly vascular-endothelial growth factor. Microvascular leakage from postcapillary venules in response to inflammatory mediators is observed in asthma, resulting in airway edema and plasma exudation into the airway lumen.

Mucus Hypersecretion

Increased mucus secretion contributes to the viscid mucous plugs that occlude asthmatic airways, particularly in fatal asthma. There is hyperplasia of submucosal glands that are confined to large airways and of increased numbers of epithelial goblet cells. IL-13 induces mucus hypersecretion in experimental models of asthma.

Neural Regulation

Various defects in autonomic neural control may contribute to AHR in asthma, but these are likely to be secondary to the disease, rather than primary defects. Cholinergic pathways, through the release of acetylcholine acting on muscarinic receptors, cause bronchoconstriction and may be activated reflexly in asthma. Inflammatory mediators may activate sensory nerves, resulting in reflex cholinergic bronchoconstriction or release of inflammatory neuropeptides. Inflammatory products may also sensitize sensory nerve endings in the airway epithelium such that the nerves become hyperalgesic. Neurotrophins, which may be released from various cell types in airways, including epithelial cells and mast cells, may cause proliferation and sensitization of airway sensory nerves. Airway nerves may also release neurotransmitters, such as substance P, which may have inflammatory effects.

Airway Remodeling

Several changes in the structure of the airway are characteristically found in asthma, and these may lead to irreversible narrowing of the airways. Population studies have shown a greater decline in lung function over time than in normal subjects; however, most patients with asthma preserve normal or near-normal lung function throughout life if appropriately treated. This suggests that the accelerated decline in lung function occurs in a smaller proportion of asthmatics, and these are usually patients with more severe disease. There is some evidence that the early use of ICS may reduce the decline in lung function. The characteristic structural changes are increased airway smooth muscle, fibrosis, angiogenesis, and mucus hyperplasia.

Physiology

Limitation of airflow is due mainly to bronchoconstriction (from mast cell mediators), but airway edema, vascular congestion, and luminal occlusion with exudate may contribute. This results in a reduction in forced expiratory volume in 1 second (FEV1), FEV1/forced vital capacity (FVC) ratio, and peak expiratory flow (PEF), as well as an increase in airway resistance. Early closure of peripheral airway results in lung hyperinflation (air trapping) and increased residual volume, particularly during acute exacerbations and in severe persistent asthma. In more severe asthma, reduced ventilation and increased pulmonary blood flow result in mismatching of ventilation and perfusion and in bronchial hyperemia. Ventilatory failure is very uncommon, even in patients with severe asthma, and arterial P_{co_2} tends to be low due to increased ventilation.

Airway Hyperresponsiveness

AHR is the characteristic physiologic abnormality of asthma and describes the excessive bronchoconstrictor response to multiple inhaled triggers that would have no effect on normal airways. The increase in AHR is linked to the frequency of asthma symptoms, and, thus, an important aim of therapy is

to reduce AHR. Increased bronchoconstrictor responsiveness is seen with *direct* bronchoconstrictors such as histamine and methacholine, which contract airway smooth muscle, but is characteristically also seen with many *indirect* stimuli, which release bronchoconstrictors from mast cells or activate sensory nerves. Most of the triggers for asthma symptoms appear to act indirectly, including allergens, exercise, hyperventilation, fog (via mast cell activation), irritant dusts, and sulfur dioxide (via a cholinergic reflex).

CLINICAL FEATURES AND DIAGNOSIS

The characteristic symptoms of asthma are wheezing, dyspnea, and coughing, which are variable, both spontaneously and with therapy. Symptoms may be worse at night and patients typically awake in the early morning hours. Patients may report difficulty in filling their lungs with air. There is increased mucus production in some patients, with typically tenacious mucus that is difficult to expectorate. There may be increased ventilation and use of accessory muscles of ventilation. Prodromal symptoms may precede an attack, with itching under the chin, discomfort between the scapulae, or inexplicable fear (impending doom).

Typical physical signs are inspiratory, and to a greater extent expiratory, rhonchi throughout the chest, and there may be hyperinflation. Some patients, particularly children, may present with a predominant nonproductive cough (“cough-variant asthma”). There may be no abnormal physical findings when asthma is under control.

DIAGNOSIS

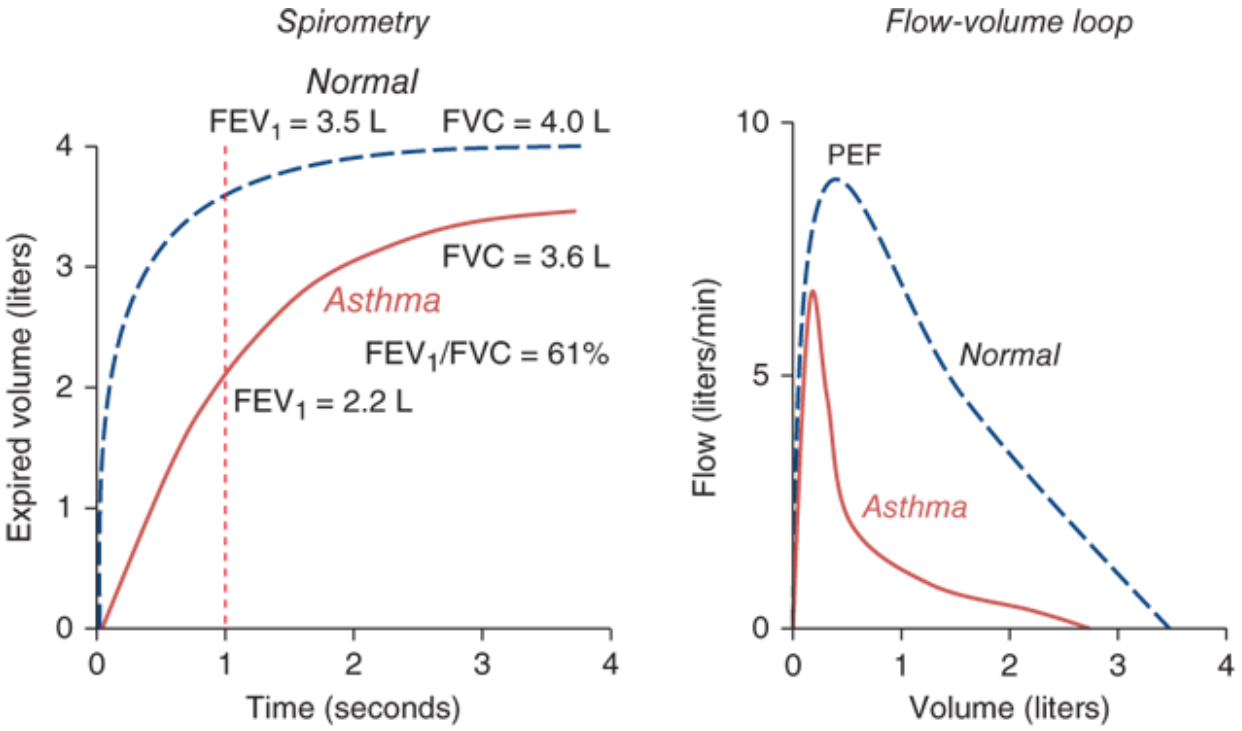
The diagnosis of asthma is usually apparent from the symptoms of variable and intermittent airways obstruction, but must be confirmed by objective measurements of lung function.

Lung Function Tests

Simple spirometry confirms airflow limitation with a reduced FEV_1, FEV_1/FVC ratio, and PEF **(Fig. 281-6)**. Reversibility is demonstrated by a >12% and 200-mL increase in FEV_1 15 min after an inhaled short-acting β_2-agonist (SABA; such as inhaled albuterol 400 μg) or in some patients by a 2–4 week trial of oral corticosteroids (OCS) (prednisone or prednisolone 30–40 mg daily). Measurements of PEF twice daily may confirm the diurnal variations in airflow obstruction. Flow-volume loops show reduced peak flow and reduced maximum expiratory flow. Further lung function tests are rarely necessary, but whole body plethysmography shows increased airway resistance and may show increased total lung capacity and residual volume. Gas diffusion, measured by carbon monoxide transfer, is usually normal, but there may be a small increase in some patients.

FIGURE 281-6

Spirometry and flow-volume loop in asthmatic compared to normal subject. There is a reduction in forced expiratory volume in 1 second (FEV_1) but less reduction in forced vital capacity (FVC), giving a reduced FEV_1/FVC ratio (<70%). The flow-volume loop shows reduced peak expiratory flow and a typical scalloped appearance indicating widespread airflow obstruction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Airway Responsiveness

The increased AHR is normally measured by methacholine or histamine challenge with calculation of the provocative concentration that reduces FEV_1 by 20% (PC20). This is rarely useful in clinical practice, but can be used in the differential diagnosis of chronic cough and when the diagnosis is in doubt in the setting of normal pulmonary function tests. Occasionally exercise testing is done to demonstrate the post-exercise bronchoconstriction if there is a predominant history of EIA. Allergen challenge is rarely necessary and should only be undertaken by a specialist if specific occupational agents are to be identified.

Hematologic Tests

Blood tests are not usually helpful. Total serum IgE and specific IgE to inhaled allergens (radioallergosorbent test [RAST]) may be measured in some patients.

Imaging

Chest roentgenography is usually normal but in more severe patients may show hyperinflated lungs. In exacerbations, there may be evidence of a pneumothorax. Lung shadowing usually indicates pneumonia or eosinophilic infiltrates in patients with bronchopulmonary aspergillosis (BPA). High-resolution CT may show areas of bronchiectasis in patients with severe asthma, and there may be thickening of the bronchial walls, but these changes are not diagnostic of asthma.

Skin Tests

Skin prick tests to common inhalant allergens (house dust mite, cat fur, grass pollen) are positive in allergic asthma and negative in intrinsic asthma, but are not helpful in diagnosis. Positive skin responses may be useful in persuading patients to undertake allergen avoidance measures.

Exhaled NO

Fractional exhaled nitric oxide (F_ENO) is now being used as a noninvasive test to measure eosinophilic airway inflammation. The typically elevated levels in asthma are reduced by ICS, so this may be a test of compliance with therapy. It may also be useful in demonstrating insufficient anti-inflammatory therapy and may be useful in down-titrating ICS. However, studies in unselected patients have not convincingly demonstrated improved clinical outcomes and it may be necessary to select patients who are poorly controlled.

Differential Diagnosis

It is usually not difficult to differentiate asthma from other conditions that cause wheezing and dyspnea. Upper airway obstruction by a tumor or laryngeal edema can mimic severe asthma, but patients typically present with stridor localized to large airways. The diagnosis is confirmed by a flow-volume loop that shows a reduction in inspiratory as well as expiratory flow, and bronchoscopy to demonstrate the site of upper airway narrowing. Persistent wheezing in a specific area of the chest may indicate endobronchial obstruction with a foreign body. Left ventricular failure may mimic the wheezing of asthma but basilar crackles are present in contrast to asthma. Vocal cord dysfunction may mimic asthma and is thought to be a hysterical conversion syndrome.

Eosinophilic pneumonias and systemic vasculitis, including Churg-Strauss syndrome (eosinophilic granulomatosis with polyangiitis) and polyarteritis nodosa, may be associated with wheezing. Chronic obstructive pulmonary disease (COPD) is usually easy to differentiate from asthma as symptoms show less variability, never completely remit, and show much less (or no) reversibility to bronchodilators. Approximately 15% of COPD patients have features of asthma, with increased sputum eosinophils and a response to OCS; these patients probably have both diseases concomitantly.

TREATMENT

TREATMENT Asthma

The treatment of asthma is straightforward, with the majority of patients now managed by internists and family doctors with effective and safe therapies. There are several aims of therapy **(Table 281-2)**. Most of the emphasis has been placed on drug therapy, but several non-pharmacologic approaches have also been used. The main drugs for asthma can be divided into bronchodilators, which give rapid relief of symptoms mainly through relaxation of airway smooth muscle, and controllers, which inhibit the underlying inflammatory process.

BRONCHODILATOR THERAPIES

Bronchodilators act primarily on airway smooth muscle to reverse the bronchoconstriction of asthma. This gives rapid relief of symptoms but has little or no effect on the underlying inflammatory process. Thus, bronchodilators are not sufficient to control asthma in patients with persistent symptoms. There are three classes of bronchodilator in current use: β_2-adrenergic agonists, anticholinergics, and theophylline; of these, β_2-agonists are by far the most effective.

β_2-Agonists

β_2-Agonists activate β_2-adrenergic receptors, which are widely expressed in the airways. β_2-Receptors are coupled through a stimulatory G protein to adenylyl cyclase, resulting in increased intracellular cyclic adenosine monophosphate (AMP), which relaxes smooth muscle cells and inhibits certain inflammatory cells, particularly mast cells.

Mode of Action

The primary action of β_2-agonists is to relax airway smooth-muscle cells of all airways, where they act as functional antagonists, reversing and preventing contraction of airway smooth-muscle cells by all known bronchoconstrictors. This generalized action is likely to account for their great efficacy as bronchodilators in asthma. There are also additional non-bronchodilator effects that may be clinically useful, including inhibition of mast cell mediator release, reduction in plasma exudation, and inhibition of sensory nerve activation. Inflammatory cells express small numbers of β_2-receptors, but these are rapidly down-regulated with β_2-agonist activation so that, in contrast to corticosteroids, there are no effects on inflammatory cells in the airways and there is no reduction in AHR.

Clinical Use

β_2-Agonists are usually given by inhalation to reduce side effects. SABA, such as albuterol and terbutaline, have a duration of action of 3–6 h. They have a rapid onset of bronchodilatation and are, therefore, used as needed for symptom relief (relievers). Increased use of SABA indicates that asthma is not controlled. They are also useful in preventing EIA if taken prior to exercise. SABA are used in high doses by nebulizer or via a metered-dose inhaler (MDI) with a spacer. Long-acting β_2-agonists (LABA) include salmeterol and formoterol, both of which have a duration of action over 12 h and are given twice daily by inhalation; and indacaterol, olodaterol, and vilanterol, which are given once daily. LABA have replaced the regular use of SABA, but LABA should not be given in the absence of ICS therapy as they do not control the underlying inflammation. They do, however, improve asthma control and reduce exacerbations when added to ICS, which allows asthma to be managed with lower doses of corticosteroids. This observation has led to the widespread use of fixed combination inhalers that contain a corticosteroid and a LABA, which have proved to be highly effective in the control of asthma and prevention of exacerbations.

Side Effects

Adverse effects are not usually a problem with β_2-agonists when given by inhalation. The most common side effects are muscle tremor and palpitations, which are seen more commonly in elderly patients. There is a small fall in plasma potassium due to increased uptake by skeletal muscle cells, but this effect does not usually cause any clinical problem.

Tolerance

Tolerance is a potential problem with any agonist given chronically, but while there is down-regulation of β_2-receptors, this does not reduce the bronchodilator response as there is a large receptor reserve in airway smooth-muscle cells. By contrast, mast cells become rapidly tolerant, but their tolerance may be prevented by concomitant administration of ICS.

Safety

The safety of β_2-agonists has been an important issue. There is an association between asthma mortality and the amount of SABA used, but careful analysis demonstrates that the increased use of rescue SABA reflects poor asthma control, which is a risk factor for asthma death. The slight excess in mortality that has been associated with the use of LABA is related to the lack of use of concomitant ICS, as the LABA therapy fails to suppress the underlying inflammation. This highlights the importance of always using an ICS when LABAs are given, which is most conveniently achieved by using a combination inhaler. Recent large safety studies have shown no adverse effects of LABA in adults or children.

Anticholinergics

Muscarinic receptor antagonists, such as ipratropium bromide, prevent cholinergic nerve-induced bronchoconstriction and mucus secretion. They are less effective than β_2-agonists in asthma therapy as they inhibit only the cholinergic reflex component of bronchoconstriction, whereas β_2-agonists prevent all bronchoconstrictor mechanisms. Long-acting muscarinic antagonists (LAMA), including tiotropium

bromide or glycopyrronium bromide, may be used as an additional bronchodilator in patients with asthma that is not controlled by maximal doses of ICS-LABA combinations, and improve lung function and further reduce exacerbations. High doses of short-acting anticholinergics may be given by nebulizer in treating acute severe asthma but should only be given following β_2-agonists, as they have a slower onset of bronchodilation.

Side effects are not usually a problem as there is little or no systemic absorption. The most common side effect is dry mouth; in elderly patients, urinary retention and glaucoma may also be observed.

Theophylline

Theophylline was widely prescribed as an oral bronchodilator several years ago, especially as it was inexpensive. It has now fallen out of favor as side effects are common, and inhaled β_2-agonists are much more effective as bronchodilators. The bronchodilator effect is due to inhibition of phosphodiesterases in airway smooth-muscle cells, which increases cyclic AMP, but doses required for bronchodilatation commonly cause side effects that are mediated mainly by phosphodiesterase inhibition. There is increasing evidence that theophylline at lower doses has anti-inflammatory effects, and these are likely to be mediated through different molecular mechanisms. Theophylline activates the key nuclear enzyme histone deacetylase-2 (HDAC2), which is a critical mechanism for switching off activated inflammatory genes and may therefore reduce corticosteroid insensitivity in severe asthma.

Clinical Use

Oral theophylline is usually given as a slow-release preparation once or twice daily as this gives more stable plasma concentrations than normal theophylline tablets. It may be used as an additional bronchodilator in patients with severe asthma when plasma concentrations of 10–20 mg/L are required, although these concentrations are often associated with side effects. Low doses of theophylline, giving plasma concentrations of 5–10 mg/L, have additive effects to ICS and are particularly useful in patients with severe asthma. Indeed, withdrawal of theophylline from these patients may result in marked deterioration in asthma control. At low doses, the drug is well tolerated. IV aminophylline (a soluble salt of theophylline) was used for the treatment of severe asthma but has now been largely replaced by high doses of inhaled SABA, which are more effective and have fewer side effects. Aminophylline is occasionally used (via slow IV infusion) in patients with severe exacerbations that are refractory to SABA.

Side Effects

Oral theophylline is well absorbed and is largely inactivated in the liver. Side effects are related to plasma concentrations; measurement of plasma theophylline may be useful in determining the correct dose. The most common side effects are nausea, vomiting, and headaches and are due to phosphodiesterase inhibition. Diuresis and palpitations may also occur, and at high concentrations cardiac arrhythmias, epileptic seizures, and death may occur due to adenosine A_1-receptor antagonism. Theophylline side effects are related to plasma concentration and are rarely observed at plasma concentrations <10 mg/L. Theophylline is metabolized by CYP450 (CYP1A2) in the liver, and, thus, plasma concentrations may be elevated by drugs that block CYP450 such as erythromycin and allopurinol. Other drugs may also reduce clearance by other mechanisms leading to increased plasma concentrations **(Table 281-3)**.

CONTROLLER THERAPIES Inhaled Corticosteroids

ICS are by far the most effective controllers for asthma, and their early use has revolutionized asthma therapy.

Mode of Action

ICS are the most effective anti-inflammatory agents used in asthma therapy, reducing inflammatory cell numbers and their activation in the airways. ICS reduce eosinophils in the airways and sputum, and

numbers of activated T lymphocytes and surface mast cells in the airway mucosa. These effects may account for the reduction in AHR that is seen with chronic ICS therapy.

The molecular mechanism of action of corticosteroids involves several effects on the inflammatory process. The major effect of corticosteroids is to switch off the transcription of multiple activated genes that encode inflammatory proteins such as cytokines, chemokines, adhesion molecules, and inflammatory enzymes. This effect involves several mechanisms, including inhibition of the transcription factors NF-κB, but an important mechanism is recruitment of HDAC2 to the inflammatory gene complex, which reverses the histone acetylation associated with increased gene transcription. Corticosteroids also activate anti-inflammatory genes such as mitogen-activated protein (MAP) kinase phosphatase-1, and increase the expression of β_2-receptors. Most of the metabolic and endocrine side effects of corticosteroids are also mediated through transcriptional activation.

Clinical Use

ICS are by far the most effective controllers in the management of asthma and are beneficial in treating asthma of any severity and age. ICS are usually given twice daily, but some may be effective once daily in mildly symptomatic patients. ICS rapidly improve the symptoms of asthma, and lung function improves over several days. They are effective in preventing asthma symptoms, such as EIA and nocturnal exacerbations, but also prevent severe exacerbations. ICS reduce AHR, but maximal improvement may take several months of therapy. Early treatment with ICS appears to prevent irreversible changes in airway function that occur with chronic asthma. Withdrawal of ICS results in slow deterioration of asthma control, indicating that they suppress inflammation and symptoms, but do not cure the underlying condition. ICS are now given as first-line therapy for patients with persistent asthma, but if they do not control symptoms at low doses, it is usual to add a LABA as the next step.

Side Effects

Local side effects include hoarseness (dysphonia) and oral candidiasis, which may be reduced with the use of a large-volume spacer device. There has been concern about systemic side effects from lung absorption, but many studies have demonstrated that ICS have minimal systemic effects **(Fig. 281-7)**. At the highest recommended doses, there may be some suppression of plasma and urinary cortisol concentrations, but there is no convincing evidence that long-term treatment leads to impaired growth in children or to osteoporosis in adults. Indeed effective control of asthma with ICS reduces the number of courses of OCS that are needed and, thus, reduces systemic exposure to ICS.

Systemic Corticosteroids

Corticosteroids are used intravenously (hydrocortisone or methylprednisolone) for the treatment of acute severe asthma, although several studies now show that OCS are as effective and easier to administer. A course of OCS (usually prednisone or prednisolone 30–45 mg once daily for 5–10 days) is used to treat acute exacerbations of asthma; no tapering of the dose is needed. Approximately 1% of asthma patients may require maintenance treatment with OCS; the lowest dose necessary to maintain control needs to be determined. Systemic side effects, including truncal obesity, bruising, osteoporosis, diabetes, hypertension, gastric ulceration, proximal myopathy, depression, and cataracts, may be a major problem, and steroid-sparing therapies may be considered if side effects are a significant problem. If patients require maintenance treatment with OCS, it is important to monitor bone density so that preventive treatment with bisphosphonates or estrogen in postmenopausal women may be initiated if bone density is low. Intramuscular triamcinolone acetonide is a depot preparation that is occasionally used in noncompliant patients, but proximal myopathy is a major problem with this therapy.

Antileukotrienes

Cysteinyl-leukotrienes are potent bronchoconstrictors; they cause microvascular leakage and increase eosinophilic inflammation through the activation of cys-LT_1-receptors. These inflammatory mediators are produced predominantly by mast cells and, to a lesser extent, eosinophils in asthma. Antileukotrienes,

such as montelukast and zafirlukast, block cys-LT_1-receptors and provide modest clinical benefit in asthma. They are less effective than ICS in controlling asthma and have less effect on airway inflammation, but are useful as an add-on therapy in some patients not controlled with low doses of ICS, although less effective than a LABA. They are given orally once or twice daily and are well tolerated. Some patients show a better response than others to antileukotrienes, but this has not been convincingly linked to any genomic differences in the leukotriene pathway.

Cromones

Cromolyn sodium and nedocromil sodium are asthma controller drugs that appear to inhibit mast cell and sensory nerve activation and are, therefore, effective in blocking trigger-induced asthma such as EIA and allergen- and sulfur dioxide-induced symptoms. Cromones have relatively little benefit in the long-term control of asthma due to their short duration of action (at least four times daily by inhalation). They are very safe and were popular in the treatment of childhood asthma, although now low doses of ICS are preferred as they are far more effective and have a proven safety profile.

Steroid-Sparing Therapies

Various immunomodulatory treatments have been used to reduce the requirement for OCS in patients with severe asthma, who have serious side effects with this therapy. Methotrexate, cyclosporin A, azathioprine, gold, and IV gamma globulin have all been used as steroid-sparing therapies, but none of these treatments has any long-term benefit and each is associated with a relatively high risk of side effects.

Anti-IgE

Omalizumab is a blocking antibody that neutralizes circulating IgE without binding to cell-bound IgE and, thus, inhibits IgE-mediated reactions. This treatment has been shown to reduce the number of exacerbations in patients with severe asthma and may improve asthma control. However, the treatment is very expensive and is only suitable for highly selected patients who are not controlled on maximal doses of inhaler therapy and have a circulating IgE within a specified range. Patients should be given a 3- to 4-month trial of therapy to show objective benefit. Omalizumab is usually given as a subcutaneous injection every 2–4 weeks and appears not to have significant side effects, although anaphylaxis is very occasionally seen.

Anti-IL-5

Antibodies that block IL-5 (mepolizumab, reslizumab) or its receptor (benralizumab) markedly reduce blood and tissue eosinophils and reduce exacerbations in patients who have persistently increased sputum eosinophils despite maximal ICS therapy.

Immunotherapy

Specific immunotherapy using injected extracts of pollens or house dust mites has not been very effective in controlling asthma and may cause anaphylaxis. Side effects may be reduced by sublingual dosing. It is not recommended in most asthma treatment guidelines because of lack of evidence of clinical efficacy and potential anaphylaxis.

Alternative Therapies

Nonpharmacologic treatments, including hypnosis, acupuncture, chiropraxis, breathing control, yoga, and speleotherapy, may be popular with some patients. However, placebo-controlled studies have shown that each of these treatments lacks efficacy and cannot be recommended. However, they are not detrimental and may be used as long as conventional pharmacologic therapy is continued.

Bronchial Thermoplasty

Bronchial thermoplasty is a bronchoscopic treatment using thermal energy to ablate airway smooth muscle in accessible bronchi. It may reduce exacerbations and improve asthma control in highly selected patients not controlled on maximal inhaler therapy, particularly when there is no increase in inflammation.

Future Therapies

It has proved very difficult to discover novel pharmaceutical therapies, particularly as current therapy with corticosteroids and β_2-agonists is so effective in the majority of patients. There is, however, a need for the development of new therapies for patients with refractory asthma who have side effects with systemic corticosteroids. Antagonists of specific mediators have little or no benefit in asthma, apart from antileukotrienes, which have rather weak effects, presumably reflecting the fact that multiple mediators are involved. Anti-TNF-α antibodies are not effective in severe asthma. Anti-IL-13 blocking antibodies have little clinical effect, but an antibody (dupilumab) against the common receptor for IL-4 and IL-13 (IL-4Rα) is more promising in reducing exacerbations and improving asthma control in severe asthma. Novel anti-inflammatory treatments that are in clinical development include inhibitors of phosphodiesterase-4, NF-κB, and p38 MAP kinase. However, these drugs, which act on signal transduction pathways common to many cells, have troublesome side effects, which may necessitate their delivery by inhalation. Safer and more effective immunotherapy using T cell peptide fragments of allergens or DNA vaccination are also being investigated. Bacterial products, such as CpG oligonucleotides that stimulate T_H1 immunity or Treg, are also currently under evaluation.

MANAGEMENT OF CHRONIC ASTHMA

There are several aims of chronic therapy in asthma (Table 281-2). It is important to establish the diagnosis objectively using spirometry or PEF measurements at home. Triggers that worsen asthma control, such as allergens or occupational agents, should be avoided, whereas triggers, such as exercise and fog, which result in transient symptoms, provide an indication that more controller therapy is needed. It is important to assess asthma control, assessed by symptoms, night awakening, need for reliever inhalers, limitation of activity and lung function **(Table 281-4)**. Avoidance of side effects and expense of medications are also important. There are several validated questionnaires for quantifying asthma control, such as the Asthma Quality of Life Questionnaire (AQLQ) and Asthma Control Test (ACT).

Stepwise Therapy

For patients with mild, intermittent asthma, a SABA is all that is required **(Fig. 281-8)**. However, use of a reliever medication more than twice a week indicates the need for regular controller therapy. The treatment of choice for all patients is an ICS given twice daily. It is usual to start with an intermediate dose (e.g., 200 [μg] bid of [beclomethasone dipropionate] BDP) or equivalent and to decrease the dose if symptoms are controlled after three months. If symptoms are not controlled, a LABA should be added, which is most conveniently given by switching to a combination inhaler. The dose of controller should be adjusted accordingly, as judged by the need for a rescue inhaler. Low doses of theophylline or an antileukotriene may also be considered as an add-on therapy, but these are less effective than LABA. In patients with severe asthma, low-dose oral theophylline is also helpful, and when there is irreversible airway narrowing, the long-acting anticholinergic may be tried. If asthma is not controlled despite the maximal recommended dose of inhaled therapy, it is important to check adherence and inhaler technique. In these patients, maintenance treatment with an OCS may be needed and the lowest dose that maintains control should be used. Occasionally omalizumab and anti-IL-5 may be tried in steroid-dependent asthmatics who are not well controlled. Once asthma is controlled, it is important to slowly decrease therapy in order to find the optimal dose to control symptoms.

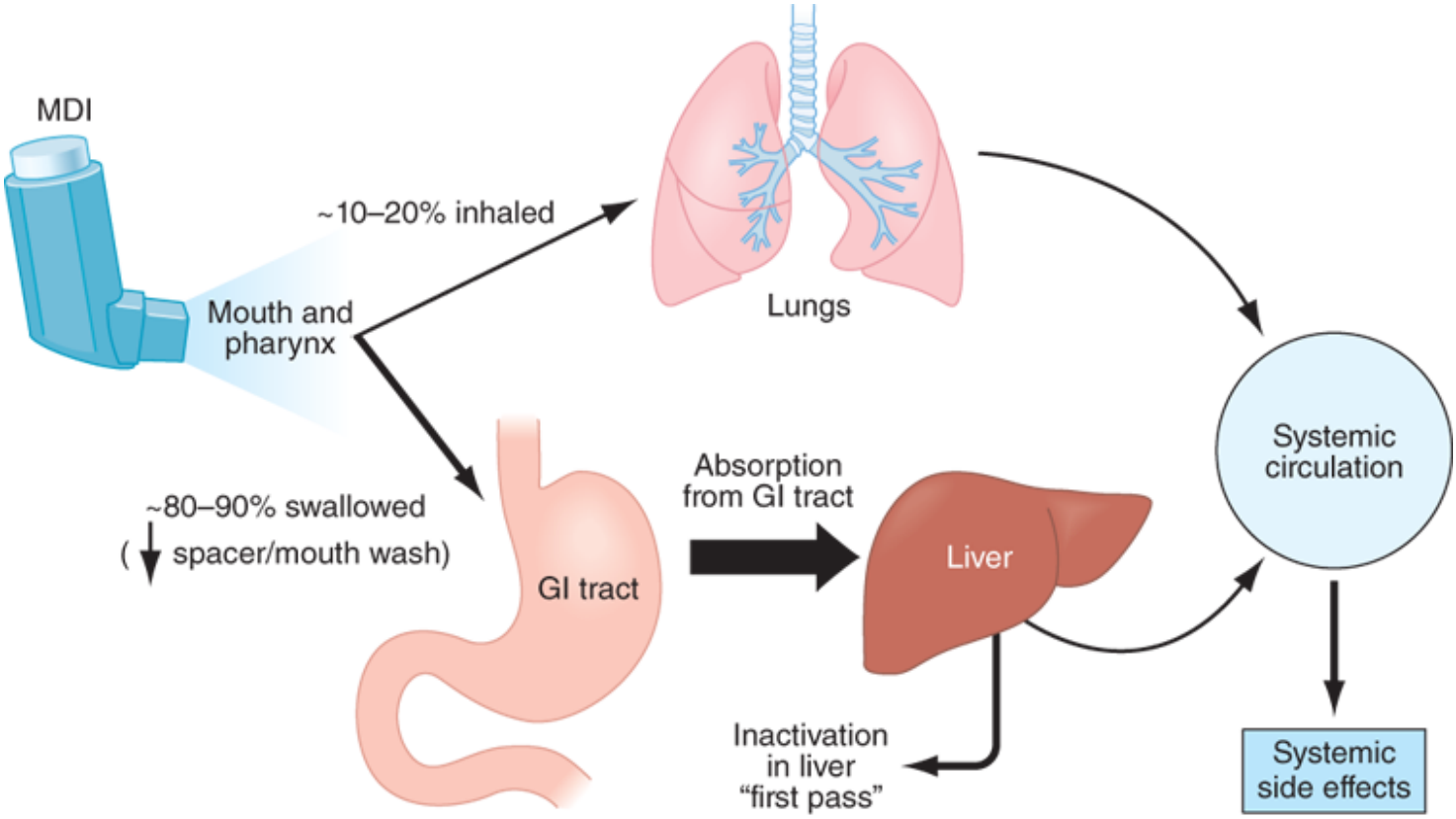
Education

Patients with asthma need to understand how to use their medications and the difference between reliever and controller therapies. Education may improve adherence, particularly with ICS. All patients should be taught how to use their inhalers correctly. In particular, they need to understand how to recognize worsening of asthma and how to step up therapy accordingly. Written action plans have been

shown to reduce hospital admissions and morbidity rates in adults and children, and are recommended particularly in patients with unstable disease who have frequent exacerbations.

FIGURE 281-7

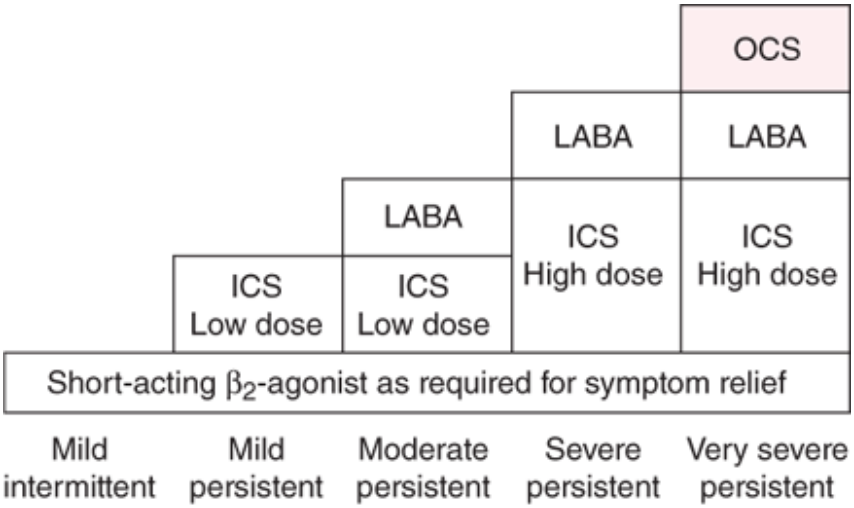
Pharmacokinetics of inhaled corticosteroids.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 281-8

Stepwise approach to asthma therapy according to the severity of asthma and ability to control symptoms. ICS, inhaled corticosteroids; LABA, long-acting β_2-agonist; OCS, oral corticosteroid.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 281-2Aims of Asthma Therapy

- Minimal (ideally no) chronic symptoms, including nocturnal
- Minimal (infrequent) exacerbations
- No emergency visits
- Minimal (ideally no) use of a required β_2-agonist
- No limitations on activities, including exercise
- Peak expiratory flow circadian variation <20%
- (Near) normal PEF
- Minimal (or no) adverse effects from medicine

Abbreviation: PEF, peak expiratory flow.

TABLE 281-3Factors Affecting Clearance of Theophylline

Increased Clearance

- Enzyme induction (rifampicin, phenobarbitone, ethanol)
- Smoking (tobacco, marijuana)
- High-protein, low-carbohydrate diet
- Barbecued meat
- Childhood

Decreased Clearance

- Enzyme inhibition (cimetidine, erythromycin, ciprofloxacin, allopurinol, zafirlukast)
- Congestive heart failure
- Liver disease
- Pneumonia
- Viral infection and vaccination
- High carbohydrate diet
- Old age

TABLE 281-4Asthma Control

CHARACTERISTIC	CONTROLLED (ALL OF FOLLOWING)	PARTLY CONTROLLED	UNCONTROLLED
Daytime symptoms	None (≤2/week)	>2/week	Three of more features of partly controlled
Limitation of activities	None	Any	
Nocturnal symptoms/awakening	None	Any	
Need for reliever/rescue treatment	None (≤2/week)	>2/week	
Lung function (PEF or FEV_1)	Normal	<80% predicted or personal best (if known)	

ACUTE SEVERE ASTHMA

Exacerbations of asthma are feared by patients and may be life threatening. One of the main aims of controller therapy is to prevent exacerbations; in this respect, ICS and combination inhalers are very effective.

Clinical Features

Patients are aware of increasing chest tightness, wheezing, and dyspnea that are often not or poorly relieved by their usual reliever inhaler. In severe exacerbations patients may be so breathless that they are unable to complete sentences and may become cyanotic. Examination usually shows increased

ventilation, hyperinflation, and tachycardia. Pulsus paradoxus may be present, but this is rarely a useful clinical sign. There is a marked fall in spirometric values and PEF. Arterial blood gases on air show hypoxemia, and P_{CO_2} is usually low due to hyperventilation. A normal or rising P_{CO_2} is an indication of impending respiratory failure and requires immediate monitoring and therapy. A chest roentgenogram is not usually informative, but may show pneumonia or pneumothorax.

TREATMENT

TREATMENT Acute Severe Asthma

A high concentration of oxygen should be given by face mask to achieve oxygen saturation of >90%. The mainstay of treatment are high doses of SABA given either by nebulizer or via a MDI with a spacer. In severely ill patients with impending respiratory failure, IV β_2-agonists may be given. A nebulized anticholinergic may be added if there is not a satisfactory response to β_2-agonists alone, as there are additive effects. In patients who are refractory to inhaled therapies, a slow infusion of aminophylline may be effective, but it is important to monitor blood levels, especially if patients have already been treated with oral theophylline. Magnesium sulfate given intravenously or by nebulizer is effective when added to inhaled β_2-agonists, and is relatively well tolerated but is not routinely recommended. Prophylactic intubation may be indicated for impending respiratory failure, when the P_{CO_2} is normal or rises. For patients with respiratory failure, it is necessary to intubate and institute ventilation. These patients may benefit from a general anesthetic, such as halothane, if they have not responded to conventional bronchodilators. Sedatives should never be given as they may depress ventilation. Antibiotics should not be used routinely unless there are signs of pneumonia.

SPECIAL CONSIDERATIONS

Refractory Asthma

Although most patients with asthma are easily controlled with appropriate medication, a small proportion of patients (~5%) are difficult to control despite maximal inhaled therapy. It is important to check adherence to therapy and inhaler technique. Some of these patients will require maintenance treatment with OCS. In managing these patients, it is important to investigate and correct any mechanisms that may be aggravating asthma. There are two major patterns of difficult asthma: some patients have persistent symptoms and poor lung function, despite appropriate therapy, whereas others may have normal or near normal lung function but intermittent, severe (sometimes life-threatening) exacerbations.

Mechanisms

The most common reason for poor control of asthma is poor adherence with medication, particularly ICS. Compliance with ICS may be low because patients do not feel any immediate clinical benefit or may be concerned about side effects. Adherence with ICS is difficult to monitor as there are no useful plasma measurements that can be made but measuring F_ENO may identify the problem. Compliance may be improved by giving the ICS as a combination with a LABA that gives symptom relief. Adherence with OCS may be measured by suppression of plasma cortisol and the expected concentration of prednisone/prednisolone in the plasma. There are several factors that may make asthma more difficult to control, including exposure to high, ambient levels of allergens or unidentified occupational agents. Severe rhinosinusitis may make asthma more difficult to control; upper airway disease should be vigorously treated. Drugs such as beta-adrenergic blockers, aspirin, and other cyclooxygenase (COX) inhibitors may worsen asthma. Some women develop severe premenstrual worsening of asthma, which is unresponsive to corticosteroids and requires treatment with progesterone or gonadotropin-releasing factors. Few systemic diseases make asthma more difficult to control, but hyper- and hypothyroidism may increase asthma symptoms and should be investigated if suspected.

Bronchial biopsy studies in refractory asthma may show the typical eosinophilic pattern of inflammation, whereas others have a predominantly neutrophilic pattern. There may be an increase in T_H1 cells, TH17 cells, and CD8 lymphocytes compared to mild asthma and increased expression of TNF-α. Structural changes in the airway, including fibrosis, angiogenesis, and airway smooth muscle thickening, are more commonly seen in these patients.

Corticosteroid-Resistant Asthma

A few patients with asthma show a poor response to corticosteroid therapy and may have various molecular abnormalities that impair the anti-inflammatory action of corticosteroids. Complete resistance to corticosteroids is extremely uncommon and affects <1 in 1000 patients. It is defined by a failure to respond to a high dose of oral prednisone/prednisolone (40 mg once daily over 2 weeks), ideally with a 2-week run-in with matched placebo. More common is reduced responsiveness to corticosteroids where control of asthma requires OCS (corticosteroid-dependent asthma). In patients with poor responsiveness to corticosteroids, there is a reduction in the response of circulating monocytes and lymphocytes to the anti-inflammatory effects of corticosteroids in vitro and reduced skin blanching in response to topical corticosteroids. There are several mechanisms that have been described, including an increase in the alternatively spliced form of the glucocorticoid receptor (GR)-β, an abnormal pattern of histone acetylation in response to corticosteroids, a defect in IL-10 production, and a reduction in HDAC2 activity (as in COPD). These observations suggest that there are likely to be heterogeneous mechanisms for corticosteroid resistance; whether these mechanisms are genetically determined has yet to be decided.

Brittle Asthma

Some patients show chaotic variations in lung function despite taking appropriate therapy. Some show a persistent pattern of variability and may require OCS or, at times, continuous infusion of β_2-agonists (type I brittle asthma), whereas others have generally normal or near-normal lung function but precipitous, unpredictable falls in lung function that may result in death (type 2 brittle asthma). These latter patients are difficult to manage as they do not respond well to corticosteroids, and the worsening of asthma does not reverse well with inhaled bronchodilators. The most effective therapy is subcutaneous epinephrine, which suggests that the worsening is likely to be a localized airway anaphylactic reaction with edema. In some of these patients, there may be allergy to specific foods. These patients should be taught to self-administer epinephrine and should carry a medical warning accordingly.

TREATMENT

TREATMENT Refractory Asthma

Refractory asthma is difficult to control, by definition. It is important to check adherence and the correct use of inhalers and to identify and eliminate any underlying triggers. Low doses of theophylline may be helpful in some patients, and theophylline withdrawal has been found to worsen in many patients. Many of these patients will require maintenance treatment with OCS, and the minimal dose that achieves satisfactory control should be determined by careful dose titration. Steroid-sparing therapies are rarely effective. In some patients with allergic asthma, omalizumab is effective, particularly when there are frequent exacerbations. Anti-IL-5 may be useful if sputum eosinophils persist despite maximal ICS or OCS therapy. Anti-TNF therapy is not effective in severe asthma and should not be used. A few patients may benefit from infusions of β_2-agonists. New therapies are needed for these patients, who currently consume a disproportionate amount of health care spending.

Aspirin-Sensitive Asthma

A small proportion (1–5%) of asthmatics become worse with aspirin and other COX inhibitors, although this is much more commonly seen in severe cases and in those patients with frequent hospital

admission. Aspirin-sensitive asthma is a well-defined phenotype of asthma that is usually preceded by perennial rhinitis and nasal polyps in nonatopic patients with a late onset of the disease. Aspirin, even in small doses, characteristically provokes rhinorrhea, conjunctival injection, facial flushing, and wheezing. There is a genetic predisposition to increased production of cysteinyl-leukotrienes with functional polymorphism of *cys*-leukotriene C4 synthase. Asthma is triggered by COX inhibitors, but is persistent even in their absence. All nonselective COX inhibitors should be avoided, but selective COX2 inhibitors are safe to use when an anti-inflammatory analgesic is needed. Aspirin-sensitive asthma responds to usual therapy with ICS. Although antileukotrienes should be effective in these patients, they are no more effective than in allergic asthma. Occasionally, aspirin desensitization is necessary, but this should only be undertaken in specialized centers.

Asthma in the Elderly

Asthma may start at any age, including in elderly patients. The principles of management are the same as in other asthmatics, but side effects of therapy may be a problem, including muscle tremor with β_2-agonists and more systemic side effects with ICS. Comorbidities are more frequent in this age group, and interactions with drugs such as β_2-blockers, COX inhibitors, and agents that may affect theophylline metabolism need to be considered. COPD is more likely in elderly patients and may coexist with asthma. A trial of OCS may be very useful in documenting the steroid responsiveness of asthma.

Pregnancy

Approximately one-third of asthmatic patients who are pregnant improve during the course of a pregnancy, one-third deteriorate, and one-third are unchanged. It is important to maintain good control of asthma as poor control may have adverse effects on fetal development. Adherence may be a problem as there is often concern about the effects of antiasthma medications on fetal development. The drugs that have been used for many years in asthma therapy have now been shown to be safe and without teratogenic potential. These drugs include SABA, ICS, and theophylline; there is less safety information about newer classes of drugs such as LABA, antileukotrienes, and anti-IgE. If an OCS is needed, it is better to use prednisone rather than prednisolone as it cannot be converted to the active prednisolone by the fetal liver, thus protecting the fetus from systemic effects of the corticosteroid. There is no contraindication to breast-feeding when patients are using these drugs.

Cigarette Smoking

Approximately 20% of asthmatics smoke, which may adversely affect asthma in several ways. Smoking asthmatics have more severe disease, more frequent hospital admissions, a faster decline in lung function, and a higher risk of death from asthma than nonsmoking asthmatics. There is evidence that smoking interferes with the anti-inflammatory actions of corticosteroids by reducing HDAC2, necessitating higher doses for asthma control. Smoking cessation improves lung function and reduces the steroid resistance, and, thus, vigorous smoking cessation strategies should be used. LABA and theophylline appear to overcome some of the steroid resistance; so, ICS-LABA combination therapy and low dose theophylline should be used. Some patients report a temporary worsening of asthma when they first stop smoking, possibly due to the loss of the bronchodilating effect of NO in cigarette smoke.

Surgery

If asthma is well controlled, there is no contraindication to general anesthesia and intubation. Patients who are treated with OCS will have adrenal suppression and should be treated with an increased dose of OCS immediately prior to surgery. Patients with FEV1 <80% of their normal levels should also be given a boost of OCS prior to surgery. High-maintenance doses of corticosteroids may be a contraindication to surgery because of increased risks of infection and delayed wound healing.

Bronchopulmonary Aspergillosis

BPA is uncommon and results from an allergic pulmonary reaction to inhaled spores of *Aspergillus fumigatus* and, occasionally, other *Aspergillus* species. A skin prick test to *A. fumigatus* is always positive, whereas serum *Aspergillus* precipitins are low or undetectable. Characteristically, there are fleeting eosinophilic infiltrates in the lungs, particularly in the upper lobes. Airways become blocked with mucoid plugs rich in eosinophils, and patients may cough up brown plugs and have hemoptysis. BPA may result in bronchiectasis, particularly affecting central airways, if not suppressed by corticosteroids. Asthma is controlled in the usual way by ICS, but it is necessary to give a course of OCS if any sign of worsening or pulmonary shadowing is found. Treatment with the oral antifungal itraconazole is beneficial in preventing exacerbations. Anti-IgE therapy may also be useful to reduce the need for OCS.

ASTHMA-COPD OVERLAP (ACO)

Although asthma and COPD are distinct syndromes with different clinical presentations and underlying inflammatory mechanisms, some patients with asthma have features of COPD (for example, asthmatics who smoke and severe asthmatics with irreversible airflow limitation) and some patients with COPD have features of asthma with more reversibility and increased airway and blood eosinophils. This may represent the coincidence of two common diseases, or these may be distinct phenotypes. ACO patients tend to have more symptoms and exacerbations. They may benefit from triple therapy with ICS, LABA, and LAMA.

FURTHER READING

BARNES PJ: Therapeutic approaches to asthma-chronic obstructive pulmonary disease overlap syndromes. J Allergy Clin Immunol 136:531, 2015. [PubMed: 26343937]
BARNES PJ: Asthma mechanisms. Medicine 44:265, 2016.
BEL EH: Clinical Practice. Mild asthma. N Engl J Med 369:549, 2013. [PubMed: 23924005]
GINA REPORT 2016: *http://ginasthma.org/2016-gina-report-global-strategy-for-asthma-management-and-prevention/*.
GROSS NJ, BARNES PJ: New therapies for asthma and COPD. Am J Respir Crit Care Med 195:159, 2017. [PubMed: 27922751]
LAMBRECHT B, HAMMID H: The immunology of asthma. Nature Immunol 16:45–56, 2016.
POSTMA DS, RABE K: The Asthma-COPD overlap syndrome. N Engl J Med 373:1241, 2015. [PubMed: 26398072]
RAY A et al: Current concepts of severe asthma. J Clin Invest 126:2394, 2016. [PubMed: 27367183]
TARLO SM, LEMIRE C: Occupational asthma. N Engl J Med 370:640, 2014. [PubMed: 24521110]

Chapter 282: Hypersensitivity Pneumonitis and Pulmonary Infiltrates with Eosinophilia

Praveen Akuthota; Michael E. Wechsler

HYPERSENSITIVITY PNEUMONITIS

INTRODUCTION AND DEFINITION

Hypersensitivity pneumonitis (HP), also referred to as extrinsic allergic alveolitis, is a pulmonary disease that occurs due to inhalational exposure to a variety of antigens leading to an inflammatory response of the alveoli and small airways. Systemic manifestations such as fever and fatigue can accompany respiratory symptoms. Although sensitization to an inhaled antigen as manifested by specific circulating IgG antibodies is necessary for the development of HP, sensitization alone is not sufficient as a defining characteristic, because many sensitized individuals do not develop HP. The incidence and prevalence of HP are variable, depending on geography, occupation, avocation, and environment of the cohort being studied. As yet unexplained is the decreased risk of developing HP in smokers.

OFFENDING ANTIGENS

HP can be caused by any of a large list of potential offending inhaled antigens **(Table 282-1)**. The various antigens and environmental conditions described to be associated with HP give rise to an expansive list of monikers given to specific forms of HP. Antigens derived from fungal, bacterial, mycobacterial, bird-derived, and chemical sources have all been implicated in causing HP.

TABLE 282-1Examples of Hypersensitivity Pneumonitis

DISEASE	ANTIGEN	SOURCE
Farming/Food Processing		
Farmer's lung	Thermophilic actinomycetes (e.g., *Saccharopolyspora rectivirgula*); fungus	Grain, moldy hay, silage
Bagassosis	Thermophilic actinomycetes	Sugarcane
Cheese washer's lung	*Penicillium casei; Aspergillus clavatus*	Cheese
Coffee worker's lung	Coffee bean dust	Coffee beans
Malt worker's lung	*Aspergillus* species	Barley
Miller's lung	*Sitophilus granarius* (wheat weevil)	Wheat flour
Mushroom worker's lung	Thermophilic actinomycetes; mushroom spores	Mushrooms
Potato riddler's lung	Thermophilic actinomycetes; *Aspergillus* species	Moldy hay around potatoes
Tobacco grower's lung	*Aspergillus* species	Tobacco
Wine maker's lung	*Botrytis cinerea*	Grapes
Birds and Other Animals		
Bird fancier's lung (also named by specific bird exposures)	Proteins derived by parakeets, pigeons, budgerigars	Bird feathers, droppings, serum proteins
Duck fever	Duck feathers, serum proteins	Ducks
Fish meal worker's lung	Fish meal dust	Fish meal
Furrier's lung	Dust from animal furs	Animal furs
Laboratory worker's lung	Rat urine, serum, fur	Laboratory rats
Pituitary snuff taker's lung	Animal proteins	Pituitary snuff from bovine and porcine sources
Poultry worker's lung	Chicken serum proteins	Chickens
Turkey handling disease	Turkey serum proteins	Turkeys
Other Occupational and Environmental Exposures		
Chemical worker's lung	Isocyanates	Polyurethane foam, varnish, lacquer
Detergent worker's lung	*Bacillus subtilis* enzymes	Detergent
Hot tub lung	*Cladosporium* species; *Mycobacterium avium* complex	Contaminated water, mold on ceiling

DISEASE	ANTIGEN	SOURCE
Humidifier fever (and air conditioner lung)	Several microorganisms including: *Aureobasidium pullulans; Candida albicans*; thermophilic actinomycetes; *Mycobacterium* species; *Klebsiella oxytoca*; *Naegleria gruberi*	Humidifiers and air conditioners (contaminated water)
Machine operator's lung	*Pseudomonas* species; *Mycobacteria* species	Metal working fluid
Sauna taker's lung	*Aureobasidium* species; other antigens	Sauna water
Suberosis	*Penicillium glabrum; Chrysonilia sitophila*	Cork dust
Summer-type pneumonitis	*Trichosporon cutaneum*	House dust mites, bird droppings
Woodworker's lung	*Alternaria* species; *Bacillus subtilis*	Oak, cedar, pine, mahogany dusts

Categories of individuals at particular risk in the United States include farmers, bird owners, industrial workers, and hot tub users. Farmer's lung occurs as a result of exposure to one of several possible sources of bacterial or fungal antigens such as grain, moldy hay, or silage. Potential offending antigens include thermophilic actinomycetes or *Aspergillus* species. Bird fancier's lung (also referred to by names corresponding to specific birds) must be considered in patients who give a history of keeping birds in their home and is precipitated by exposure to antigens derived from feathers, droppings, and serum proteins. Occupational exposure to birds may also cause HP, as is seen in poultry worker's lung. Chemical worker's lung is provoked by exposure to occupational chemical antigens such as diphenylmethane diisocyanate and toluene diisocyanate. Mycobacteria may cause HP rather than frank infection, a phenomenon observed in hot tub lung and in HP due to metalworking fluid.

PATHOPHYSIOLOGY

While much remains to be learned regarding the pathophysiology of HP, it has been established that HP is an immune-mediated condition that occurs in response to inhaled antigens that are small enough to deposit in distal airways and alveoli. From a lymphocyte perspective, HP has been categorized as a condition with a T_H1 inflammatory pattern. However, emerging evidence suggests that T_H17 lymphocyte subsets may be involved in the pathogenesis of the disease as well. Although the presence of precipitating IgG antibodies against specific antigens in HP suggests a prominent role for adaptive immunity in the pathophysiology of HP, innate immune mechanisms likely also make an important contribution. This is highlighted by the observation that Toll-like receptors and downstream signaling proteins such as MyD88 are activated in HP, leading to neutrophil recruitment. Although no clear genetic basis for HP has been established, in specific cohorts, polymorphisms in genes involved in antigen processing and presentation, including TAP1 and major histocompatibility complex type II, have been observed. In chronic HP, bone marrow-derived fibrocytes may contribute to lung inflammation and fibrosis.

CLINICAL PRESENTATION

Given the heterogeneity among patients, variability in offending antigens, and differences in the intensity and duration of exposure to antigen, the presentation of HP is accordingly variable. Although these categories are not fully satisfactory in capturing this variability, HP has been traditionally categorized as having *acute, subacute*, and *chronic* forms. Acute HP usually manifests itself 4–8 h following exposure to the inciting antigen, often intense in nature. Systemic symptoms, including fevers, chills, and malaise, are prominent and are accompanied by dyspnea. Symptoms resolve within hours to days if no further exposure to the offending antigen occurs. In subacute HP resulting from

ongoing antigen exposure, the onset of respiratory and systemic symptoms is typically more gradual over the course of weeks. A similar presentation may occur as a culmination of intermittent episodes of acute HP. Although respiratory impairment may be quite severe, antigen avoidance generally results in resolution of the symptoms, but with a slower time course, on the order of weeks to months, than that seen with acute HP. Chronic HP can present with an even more gradual onset of symptoms than subacute HP, with progressive dyspnea, cough, fatigue, weight loss, and clubbing of the digits. The insidious onset of symptoms and frequent lack of an anteceding episode of acute HP make diagnosing chronic HP a challenge. Unlike with the other forms of HP, there can be an irreversible component to the respiratory impairment that is not responsive to removal of the responsible antigen from the patient's environment. The disease progression of chronic HP to lung fibrosis and hypoxemic respiratory failure can mirror that seen in idiopathic pulmonary fibrosis (IPF). Diagnostic uncertainty between these two entities is not uncommon. Fibrotic lung disease is a potential feature of chronic HP due to exposure to bird antigens, whereas an emphysematous phenotype may be seen in farmer's lung.

The categories of acute, subacute, and chronic HP are not completely sufficient in classifying HP. The HP Study Group found on cluster analysis that a cohort of HP patients is best described in bipartite fashion, with one group featuring recurrent systemic signs and symptoms and the other featuring more severe respiratory findings.

Concordant with the variability in the presentation of HP is the observed variability in outcome. HP that has not progressed to chronic lung disease has a more favorable outcome with likely resolution if antigen avoidance can be achieved. However, chronic HP resulting in lung fibrosis has a poorer prognosis, with patients with chronic pigeon breeder's lung having demonstrated a similar mortality as seen in IPF.

DIAGNOSIS

Although there is no set of universally accepted criteria for arriving at a diagnosis of HP, diagnosis depends foremost on establishing a history of exposure to an offending antigen that correlates with respiratory and systemic symptoms. A careful occupational and home exposure history should be taken and may be supplemented if necessary by a clinician visit to the work or home environment. Specific inquiries will be influenced by geography and the occupation of the patient. When HP is suspected by history, the additional workup is aimed at establishing an immunologic and physiologic response to inhalational antigen exposure with chest imaging, pulmonary function testing (PFT), serologic studies, bronchoscopy, and, on occasion, lung biopsy. Re-exposure to the offending environment may be performed to aid in confirming the diagnosis of HP.

Chest Imaging

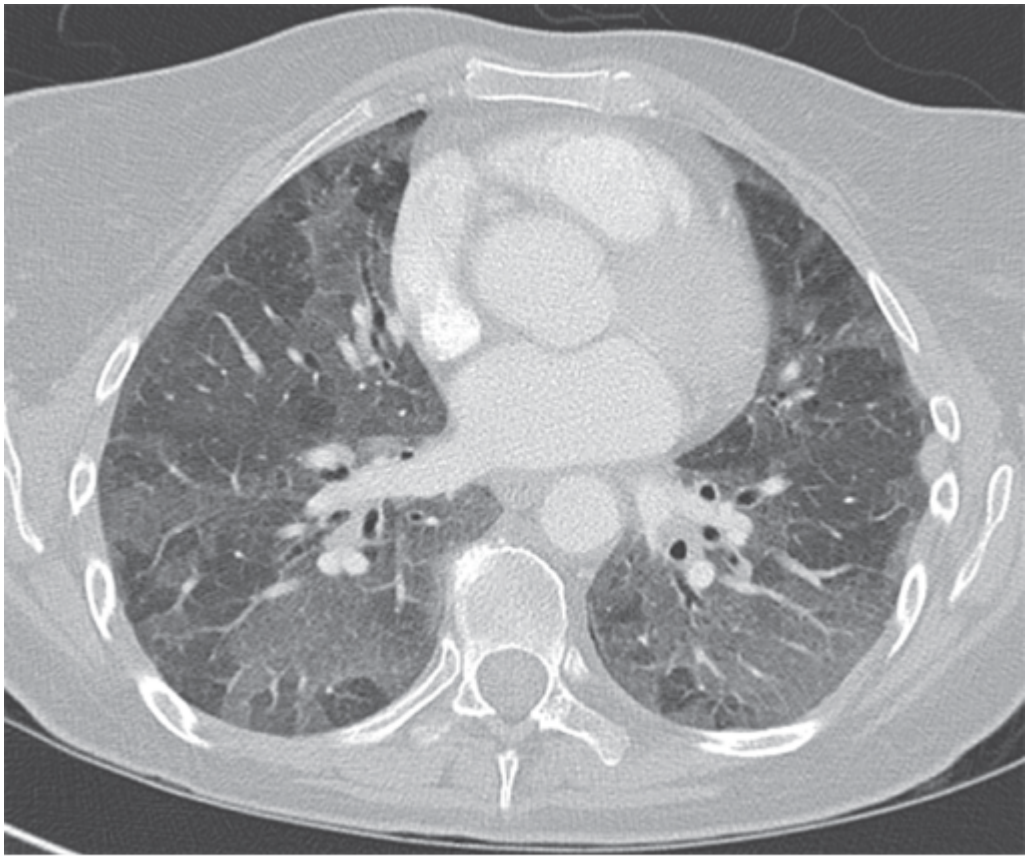
Chest x-ray findings in HP are nonspecific and can even lack any discernible abnormalities. In cases of acute and subacute HP, findings may be transient and can include ill-defined micronodular opacities or hazy ground-glass airspace opacities. Findings on chest x-ray will often resolve with removal from the offending antigen, although the time course of resolution may vary. With chronic HP, the abnormalities seen on the chest radiograph are frequently more fibrotic in nature and may be difficult to distinguish from IPF.

With the wide availability of high-resolution computed tomography (HRCT), this modality has become a common component in the diagnostic workup for HP. Although the HRCT may be normal in acute forms of HP, this may be due to lack of temporal correlation between exposure to the offending antigen and obtaining the imaging. Additionally, because of the transient nature of acute HP, HRCT is not always performed. In subacute forms of the disease, ground-glass airspace opacities are characteristic, as is the presence of centrilobular nodules. Expiratory images may show areas of air trapping that are likely caused by involvement of the small airways **(Fig. 282-1)**. Reticular changes

and traction bronchiectasis can be observed in chronic HP. Subpleural honeycombing similar to that seen in IPF may be present in advanced cases, although unlike in IPF, the lung bases are frequently spared.

FIGURE 282-1

Chest computed tomography scan of a patient with subacute hypersensitivity pneumonitis in which scattered regions of ground-glass infiltrates in a mosaic pattern consistent with air trapping are seen bilaterally. This patient had bird fancier's lung. *(Courtesy of TJ Gross; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pulmonary Function Testing

Either restrictive or obstructive PFTs can be present in HP, so the pattern of PFT change is not useful in establishing the diagnosis of HP. However, obtaining PFTs is of use in characterizing the physiologic impairment of an individual patient and in gauging the response to antigen avoidance and/or corticosteroid therapy. Diffusion capacity for carbon monoxide may be significantly impaired, particularly in cases of chronic HP with fibrotic pulmonary parenchymal changes.

Serum Precipitins

Assaying for precipitating IgG antibodies against specific antigens can be a useful adjunct in the diagnosis of HP. However, the presence of an immunologic response alone is not sufficient for establishing the diagnosis, because many asymptomatic individuals with high levels of exposure to antigen may display serum precipitins, as has been observed in farmers and in pigeon breeders. It should also be noted that panels that test for several specific serum precipitins often provide false-

negative results, because they represent an extremely limited proportion of the universe of potential offending environmental antigens.

Bronchoscopy

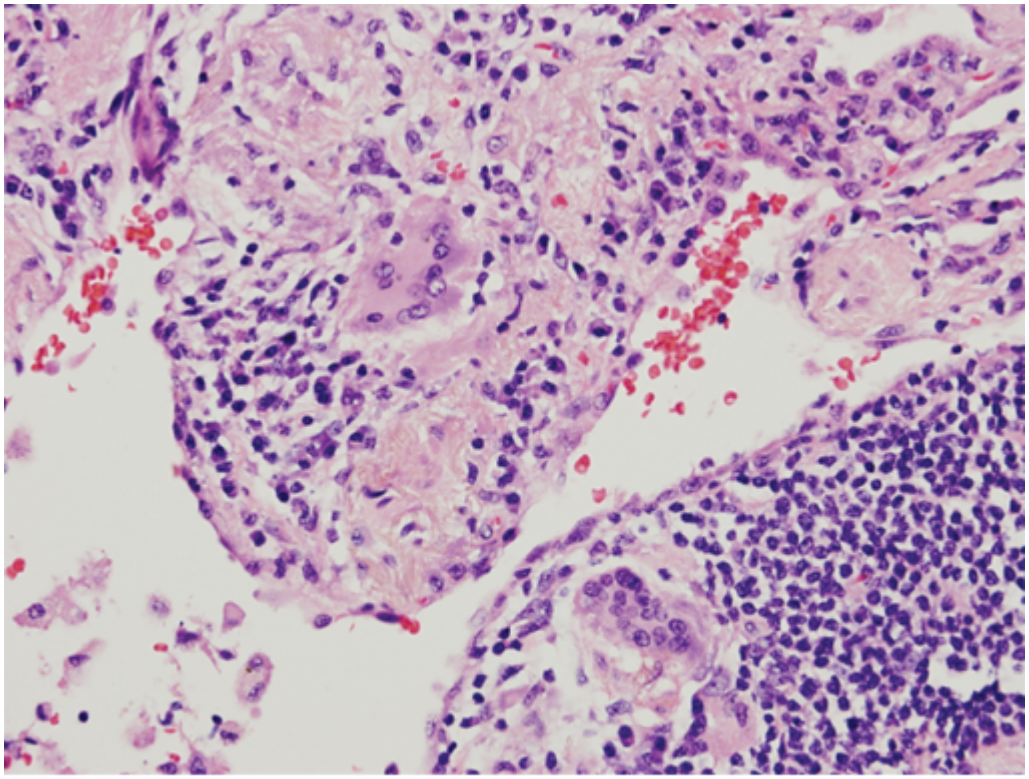
Bronchoscopy with bronchoalveolar lavage (BAL) may be used in the evaluation of HP. Although not a specific finding, BAL lymphocytosis is characteristic of HP. However, in active smokers, a lower threshold should be used to establish BAL lymphocytosis, because smoking will result in lower lymphocyte percentages. Most cases of HP have a CD4+/CD8+ lymphocyte ratio of <1, but again, this is not a specific finding and has limited utility in the diagnosis of HP.

Lung Biopsy

Tissue samples may be obtained by a bronchoscopic approach using transbronchial biopsy, or more architecturally preserved specimens may be obtained by a surgical approach (video-assisted thoracoscopy or open approach). As is the case with BAL, histologic specimens are not absolutely necessary to establish the diagnosis of HP, but they can be useful in the correct clinical context. A common histologic feature in HP is the presence of noncaseating granulomas in the vicinity of small airways **(Fig. 282-2)**. As opposed to pulmonary sarcoidosis, in which noncaseating granulomas are well defined, the granulomas seen in HP are loose and poorly defined in nature. Within the alveolar spaces and in the interstitium, a mixed cellular infiltrate with a lymphocytic predominance is observed that is frequently patchy in distribution. Bronchiolitis with the presence of organizing exudate is also often observed. Fibrosis may be present as well, particularly in chronic HP. Fibrotic changes may be focal but can be diffuse and severe with honeycombing in advanced cases, similar to findings in IPF.

FIGURE 282-2

Open-lung biopsy from a patient with subacute hypersensitivity pneumonitis demonstrating a loose, nonnecrotizing granuloma made up of histiocytes and multinucleated giant cells. Peribronchial inflammatory infiltrate made up of lymphocytes and plasma cells is also seen. *(Courtesy of TJ Gross; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical Prediction Rule

Although not meant as a set of validated diagnostic criteria, a clinical prediction rule for predicting the presence of HP has been published by the HP Study Group. They identified six statistically significant predictors for HP, the strongest of which was exposure to an antigen known to cause HP. Other predictive criteria were the presence of serum precipitins, recurrent symptoms, symptoms occurring 4–8 h after antigen exposure, crackles on inspiration, and weight loss.

DIFFERENTIAL DIAGNOSIS

Differentiating HP from other conditions that cause a similar constellation of respiratory and systemic symptoms requires an increased index of suspicion based on obtaining a history of possible exposure to an offending antigen. Presentations of acute or subacute HP can be mistaken for respiratory infection. In cases of chronic disease, HP must be differentiated from interstitial lung disease, such as IPF or nonspecific interstitial pneumonitis (NSIP); this can be a difficult task even with lung biopsy. Given the presence of pulmonary infiltrates and noncaseating granulomas on biopsy, sarcoidosis is also a consideration in the differential diagnosis of HP. Unlike in HP, however, hilar adenopathy may be prominent on chest x-ray, organs other than the lung may be involved, and noncaseating granulomas in pathologic specimens tend to be well formed. Other inhalational syndromes, such as organic toxic dust syndrome (OTDS), can be misdiagnosed as HP. OTDS occurs with exposure to organic dusts, including those produced by grains or mold silage, but neither requires prior antigen sensitization nor is characterized by positive serum precipitins.

TREATMENT

TREATMENT Hypersensitivity Pneumonitis

The mainstay of treatment for HP is antigen avoidance, if possible. A careful exposure history must be obtained to attempt to identify the potential offending antigen and to identify the location where a patient is exposed. Once a potential antigen and location are identified, efforts should be made to modify the environment to minimize patient exposure. This may be accomplished with measures such as removal of birds, removal of molds, and improved ventilation. Personal protective equipment including respirators and ventilated helmets can be used but may not provide adequate protection for sensitized individuals. In some cases, fully avoiding specific environments may be necessary, although such a recommendation must be balanced against the effects to an individual's lifestyle or occupation. It is not uncommon for patients with HP due to exposure to household birds to be unwilling to remove them from the home.

Because acute HP is generally a self-limited disease after a discrete exposure to an offending antigen, pharmacologic therapy is generally not necessary. However, in so-called subacute and chronic forms of the disease, there is a role for glucocorticoid therapy. In patients with particularly severe symptoms as a result of subacute HP, antigen avoidance may be insufficient after establishing the diagnosis. Although glucocorticoids do not change the long-term outcome in these patients, they can accelerate the resolution of symptoms. While there is significant variability in the approach to glucocorticoid therapy by individual clinicians, prednisone therapy can be initiated at 0.5–1 mg/kg of ideal body weight per day (not to exceed 60 mg/d or alternative glucocorticoid equivalent) over a duration of 1–2 weeks, followed by a taper over the next 2–6 weeks. In chronic HP, a similar trial of corticosteroids may be used, although a variable component of fibrotic disease may be irreversible. In advanced cases of chronic HP with extensive lung fibrosis, lung transplantation may be necessary.

GLOBAL CONSIDERATIONS

 As the ever-expanding list of antigens and exposures associated with the development of HP suggests, populations at risk for HP will vary globally based on specifics of local occupational, avocational, and environmental factors. Specific examples of geographically limited HP include summer-type pneumonitis seen in Japan and suberosis seen in cork workers in Portugal and Spain.

PULMONARY INFILTRATES WITH EOSINOPHILIA

Although eosinophils are normal constituents of the lungs, there are several pulmonary eosinophilic syndromes that are characterized by pulmonary infiltrates on imaging along with an increased number of eosinophils in lung tissue, in sputum, and/or in BAL fluid, with resultant increased respiratory symptoms and the potential for systemic manifestations. Because the eosinophil plays such an important role in each of these syndromes, it is often difficult to distinguish between them, but there are important clinical and pathologic differences as well as differences in prognosis and treatment paradigms.

CLASSIFYING PULMONARY INFILTRATES WITH EOSINOPHILIA AND GENERAL APPROACH

Because there are so many different diagnoses associated with pulmonary infiltrates with eosinophilia, the first step in classifying pulmonary eosinophilic syndromes is distinguishing between primary pulmonary eosinophilic lung disorders and those with eosinophilia that are secondary to a specific cause such as a drug reaction, an infection, a malignancy, or another pulmonary condition such as asthma. **Table 282-2** lists primary and secondary pulmonary eosinophilic disorders.

TABLE 282-2Pulmonary Infiltrates with Eosinophilia

Primary Pulmonary Eosinophilic Disorders

Acute eosinophilic pneumonia

Chronic eosinophilic pneumonia

Eosinophilic granulomatosis with polyangiitis (Churg-Strauss syndrome)

Hypereosinophilic syndrome

Pulmonary Disorders of Known Cause Associated with Eosinophilia

Asthma and eosinophilic bronchitis

Allergic bronchopulmonary aspergillosis

Bronchocentric granulomatosis

Drug/toxin reaction

Infection (Table 282-4)

- Parasitic/helminthic disease
- Nonparasitic infection

Lung Diseases Associated with Eosinophilia

Cryptogenic organizing pneumonia

Hypersensitivity pneumonitis

Idiopathic pulmonary fibrosis

Pulmonary Langerhans cell granulomatosis

Malignant Neoplasms Associated with Eosinophilia

Leukemia

Lymphoma

Lung cancer

Adenocarcinoma of various organs

Squamous cell carcinoma of various organs

Systemic Disease Associated with Eosinophilia

Postradiation pneumonitis

Rheumatoid arthritis

Sarcoidosis

Sjögren’s syndrome

For each patient, a detailed history is of utmost importance and can help elucidate what the underlying disease is. Details regarding onset, timing, and precipitants of specific symptoms can help discern one diagnosis from another. History regarding pharmacologic, occupational, and environmental exposures is instructive, and family and travel history are crucial. In addition to details about the sinuses and lungs, it is important to inquire about systemic manifestations and assess for physical findings of cardiac, gastrointestinal (GI), neurologic, dermatologic, and genitourinary involvement, all of which may give clues to specific diagnoses. Once the details from history and physical are teased out, laboratory testing (including measurements of blood eosinophils, cultures, and markers of inflammation), spirometry and radiographic imaging can help distinguish between different diseases. Often, however, BAL, transbronchial, or open lung biopsies are required. In many cases, biopsies or noninvasive diagnostic studies of other organs (e.g., echocardiogram, electromyogram, or bone marrow biopsy) can be helpful.

PATHOPHYSIOLOGY

Pathologically, the pulmonary eosinophilic syndromes are characterized by tissue infiltration by eosinophils (Fig. 282-2). In eosinophilic granulomatosis with polyangiitis (EGPA), extravascular granulomas and necrotizing vasculitis may occur in the lungs, as well as in the heart, skin, muscle, liver, spleen, and kidneys, and may be associated with fibrinoid necrosis and thrombosis.

The exact etiology of the various pulmonary eosinophilic syndromes is unknown; however, it is felt that these syndromes result from dysregulated eosinophilopoiesis or an autoimmune process because of the prominence of allergic features and the presence of immune complexes, heightened T cell immunity, and altered humoral immunity as evidenced by elevated IgE and rheumatoid factor. Because of its integral involvement in eosinophilopoiesis, interleukin 5 (IL-5) has been hypothesized to play an etiologic role. Monoclonal antibodies against IL-5 are now in clinical use for the treatment of eosinophilic asthma and are under investigation for conditions characterized by pulmonary infiltrates with eosinophilia. Antineutrophil cytoplasmic antibodies (ANCAs) are present in about half of patients with EGPA; binding of ANCAs to vascular walls likely contributes to vascular inflammation and injury as well as chemotaxis of inflammatory cells.

ACUTE EOSINOPHILIC PNEUMONIA

Acute eosinophilic pneumonia is a syndrome characterized by fevers, acute respiratory failure that often requires mechanical ventilation, diffuse pulmonary infiltrates, and pulmonary eosinophilia in a previously healthy individual **(Table 282-3)**.

TABLE 282-3Diagnostic Criteria of Acute Eosinophilic Pneumonia

Acute febrile illness with respiratory manifestations of <1 month in duration
Hypoxemic respiratory failure
Diffuse pulmonary infiltrates on chest x-ray
Bronchoalveolar lavage eosinophilia >25%
Absence of parasitic, fungal, or other infection
Absence of drugs known to cause pulmonary eosinophilia
Quick clinical response to corticosteroids
Failure to relapse after discontinuation of corticosteroids

Clinical Features and Etiology

At presentation, acute eosinophilic pneumonia is often mistaken for acute lung injury or acute respiratory distress syndrome (ARDS), until a BAL is performed and reveals >25% eosinophils. Although the predominant symptoms of acute eosinophilic pneumonia are cough, dyspnea, malaise, myalgias, night sweats, and pleuritic chest pain, physical examination findings include high fevers, basilar rales, and rhonchi on forced expiration. Acute eosinophilic pneumonia most often affects males between age 20 and 40 with no history of asthma. Although no clear etiology has been identified, several case reports have linked acute eosinophilic pneumonia to recent initiation of tobacco smoking or exposure to other environmental stimuli including dust from indoor renovations.

In addition to a suggestive history, the key to establishing a diagnosis of acute eosinophilic pneumonia is the presence of >25% eosinophilia on BAL fluid. While lung biopsies show eosinophilic infiltration with acute and organizing diffuse alveolar damage, it is generally not necessary to proceed to biopsy to establish a diagnosis. Although patients present with an elevated white blood cell count, in contrast to other pulmonary eosinophilic syndromes, acute eosinophilic pneumonia is often not associated with peripheral eosinophilia upon presentation. However, between 7 and 30 days of disease onset, peripheral eosinophilia often occurs with mean eosinophil counts of 1700. Erythrocyte sedimentation rate (ESR), C-reactive protein, and IgE levels are high but nonspecific, whereas HRCT is always abnormal with bilateral random patchy ground-glass or reticular opacities, and small pleural effusions in as many as two-thirds of patients. Pleural fluid is characterized by a high pH with marked eosinophilia.

Clinical Course and Response to Therapy

Although some patients improve spontaneously, most patients require admission to an intensive care unit and respiratory support with either invasive (intubation) or noninvasive mechanical ventilation. However, what distinguishes acute eosinophilic pneumonia from both other cases of acute lung injury as well as some of the other pulmonary eosinophilic syndromes is the absence of organ dysfunction or multisystem organ failure other than respiratory failure. One of the characteristic features of acute eosinophilic pneumonia is the high degree of corticosteroid responsiveness and the excellent prognosis. Another distinguishing feature of acute eosinophilic pneumonia is that complete clinical and radiographic recovery without recurrence or residual sequelae occurs in almost all patients within several weeks of initiation of therapy.

CHRONIC EOSINOPHILIC PNEUMONIA

In contrast to acute eosinophilic pneumonia, chronic eosinophilic pneumonia is a more indolent syndrome that is characterized by pulmonary infiltrates and eosinophilia in both the tissue and blood. Most patients are female nonsmokers with a mean age of 45, and patients do not usually develop the acute respiratory failure and significant hypoxemia appreciated in acute eosinophilic pneumonia. Similar to EGPA, a majority have asthma, with many having a history of allergies.

Patients present with a subacute illness over weeks to months, with cough, low-grade fevers, progressive dyspnea, weight loss, wheezing, malaise, and night sweats, and a chest x-ray with migratory bilateral peripheral or pleural-based opacities. Although this “photographic negative pulmonary edema” appearance on chest x-ray and chest CT is pathognomonic of chronic eosinophilic pneumonia, <25% of patients present with this finding. Other radiographic findings include atelectasis, pleural effusions, lymphadenopathy, and septal line thickening.

Almost 90% of patients have peripheral eosinophilia, with mean eosinophil counts of over 30% of total white blood cell count. BAL eosinophilia is also an important distinguishing feature with mean BAL eosinophil counts of ~60%. Both peripheral and BAL eosinophilia are very responsive to treatment with corticosteroids. Other laboratory features of chronic eosinophilic pneumonia include increased

ESR, C-reactive protein, platelets, and IgE. Lung biopsy is also often not required to establish a diagnosis, but may show accumulation of eosinophils and histiocytes in the lung parenchyma and interstitium, as well as cryptogenic organizing pneumonia, but with minimal fibrosis. Nonrespiratory manifestations are uncommon, but arthralgias, neuropathy, and skin and GI symptoms have all been reported; their presence may suggest EGPA or a hypereosinophilic syndrome. Another similarity is the rapid response to corticosteroids with quick resolution of peripheral and BAL eosinophilia and improvement in symptoms. In contrast to acute eosinophilic pneumonia, though, >50% of patients relapse, and many require prolonged courses of corticosteroids for months to years.

EOSINOPHILIC GRANULOMATOSIS WITH POLYANGIITIS (EGPA)

Previously known as allergic angiitis granulomatosis or Churg-Strauss syndrome, this complex syndrome is characterized by eosinophilic vasculitis that may involve multiple organ systems including the lungs, heart, skin, GI tract, and nervous system. Although EGPA is characterized by peripheral and pulmonary eosinophilia with infiltrates on chest x-ray, the primary features that distinguish EGPA from other pulmonary eosinophilic syndromes are the presence of eosinophilic vasculitis in the setting of asthma and involvement of multiple end organs (a feature it shares with hypereosinophilic syndrome). Although perceived to be quite rare, in the last few years, there has appeared to be an increased incidence of this disease, particularly in association with various asthma therapies, including leukotriene modifiers and anti-IGE therapy with omalizumab, possibly due to concurrent systemic corticosteroid withdrawal (*forme fruste* EGPA).

The primary features of EGPA include asthma, peripheral eosinophilia, neuropathy, pulmonary infiltrates, paranasal sinus abnormality, and presence of eosinophilic vasculitis. The mean age at diagnosis is 48 years, with a range of 14–74 years; the average length of time between diagnosis of asthma and vasculitis is 9 years. EGPA typically occurs in several phases. The prodromal phase is characterized by asthma and allergic rhinitis, and usually begins when the individual is in his or her twenties or thirties, typically persisting for many years. The eosinophilic infiltrative phase is characterized by peripheral eosinophilia and eosinophilic tissue infiltration of various organs including the lungs and GI tract. The third phase is the vasculitic phase and may be associated with constitutional signs and symptoms including fever, weight loss, malaise, and fatigue. This phasic progression supports the hypothesis that there is a pathophysiologic continuum between eosinophilic asthma, chronic eosinophilic asthma, and EGPA.

Similar to other pulmonary eosinophilic syndromes, constitutional symptoms are very common in EGPA and include weight loss of 10–20 lb, fevers, and diffuse myalgias and migratory polyarthralgias. Myositis may be present with evidence of vasculitis on muscle biopsies. In contrast to the eosinophilic pneumonias, EGPA involves many organ systems including the lungs, skin, nerves, heart, GI tract, and kidneys.

Symptoms and Clinical Manifestations

Respiratory

Most EGPA patients have asthma that arises later in life and in individuals who have no family history of atopy. The asthma can often be severe, and oral corticosteroids are often required to control symptoms but may lead to suppression of vasculitic symptoms. In addition to the more common symptoms of cough, dyspnea, sinusitis, and allergic rhinitis, alveolar hemorrhage and hemoptysis may also occur.

Neurologic

Over three-fourths of EGPA patients have neurologic manifestations. Mononeuritis multiplex most commonly involves the peroneal nerve, but also involves the ulnar, radial, internal popliteal, and occasionally, cranial nerves. Cerebral hemorrhage and infarction may also occur and are important causes of death. Despite treatment, neurologic sequelae often do not completely resolve.

Dermatologic

Approximately half of EGPA patients develop dermatologic manifestations. These include palpable purpura, skin nodules, urticarial rashes, and livedo.

Cardiovascular

Granulomas, vasculitis, and widespread myocardial damage may be found on biopsy or at autopsy, and cardiomyopathy and heart failure may be seen in up to half of all patients but are often at least partially reversible. Acute pericarditis, constrictive pericarditis, myocardial infarction, and other electrocardiographic changes all may occur. The heart is a primary target organ in EGPA, and cardiac involvement often portends a worse prognosis.

GI

GI symptoms are common in EGPA and likely represent an eosinophilic gastroenteritis characterized by abdominal pain, diarrhea, GI bleeding, and colitis. Ischemic bowel, pancreatitis, and cholecystitis have also been reported in association with EGPA and usually portend a worse prognosis.

Renal

Renal involvement is more common than once thought, and ~25% of patients have some degree of renal involvement. This may include proteinuria, glomerulonephritis, renal insufficiency, and rarely, renal infarct.

Lab Abnormalities

Systemic eosinophilia is the hallmark laboratory finding in patients with EGPA and reflects the likely pathogenic role that the eosinophil plays in this disease. Eosinophilia >10% is one of the defining features of this illness and may be as high as 75% of the peripheral white blood cell count. It is present at the time of diagnosis in >80% of patients, but may respond quickly (often within 24 h) to initiation of systemic corticosteroid therapy. Even in the absence of systemic eosinophilia, tissue eosinophilia may be present.

Although not specific to EGPA, ANCAs are present in up to two-thirds of patients, mostly with a perinuclear staining pattern. Nonspecific lab abnormalities that may be present in patients with EGPA include a marked elevation in ESR, a normochromic normocytic anemia, an elevated IgE, hypergammaglobulinemia, and positive rheumatoid factor and antinuclear antibodies (ANA). Although BAL often reveals significant eosinophilia, this may be seen in other eosinophilic lung diseases. Similarly, PFT often reveals an obstructive defect similar to asthma.

Radiographic Features

Chest x-ray abnormalities are extremely common in EGPA and consist of bilateral, nonsegmental, patchy infiltrates that often migrate and may be interstitial or alveolar in appearance. Reticulonodular and nodular disease without cavitation can be seen, as can pleural effusions and hilar adenopathy. The most common CT findings include bilateral ground-glass opacity and airspace consolidation that

is predominantly subpleural. Other CT findings include bronchial wall thickening, hyperinflation, interlobular septal thickening, lymph node enlargement, and pericardial and pleural effusions. Angiography may be used diagnostically and may show signs of vasculitis in the coronary, central nervous system, and peripheral vasculature.

Treatment and Prognosis of EGPA

Most patients diagnosed with EGPA have previously been diagnosed with asthma, rhinitis, and sinusitis, and have received treatment with inhaled or systemic corticosteroids. Because these agents are also the initial treatment of choice for EGPA patients, institution of these therapies in patients with EGPA who are perceived to have severe asthma may delay the diagnosis of EGPA because signs of vasculitis may be masked. Corticosteroids dramatically alter the course of EGPA: up to 50% of those who are untreated die within 3 months of diagnosis, whereas treated patients have a 6-year survival of >70%. Common causes of death include heart failure, cerebral hemorrhage, renal failure, and GI bleeding. Recent data suggest that clinical remission may be obtained in >90% of patients treated; ~25% of those patients may relapse, often due to corticosteroid tapering, with a rising eosinophil count heralding the relapse. Myocardial, GI, and renal involvement most often portend a poor prognosis. In such cases, treatment with higher doses of corticosteroids or the addition of cytotoxic agents such as cyclophosphamide is often warranted. Although survival does not differ between those treated or untreated with cyclophosphamide, cyclophosphamide is associated with a reduced incidence of relapse and an improved clinical response to treatment. Other therapies that have been used successfully in the management of EGPA include azathioprine, methotrexate, rituximab, omalizumab, intravenous gamma globulin, and interferon α. Plasma exchange has not been shown to provide any additional benefit. Recent studies examining the efficacy of anti-IL-5 therapy compared with placebo have shown promise as safe and effective corticosteroid sparing agents that can reduce exacerbations.

HYPEREOSINOPHILIC SYNDROMES

Hypereosinophilic syndromes (HES) constitute a heterogeneous group of disease entities manifest by persistent eosinophilia >1500 eosinophils/μL in association with end organ damage or dysfunction, in the absence of secondary causes of eosinophilia. In addition to familial, undefined, and overlap syndromes with incomplete criteria, the predominant HES subtypes are the myeloproliferative and lymphocytic variants. The myeloproliferative variants may have acquired genetic abnormalities, including to platelet-derived growth factor receptor α (PDGFRα), attributed to a constitutively activated tyrosine kinase fusion protein (Fip1L1-PDGFRα) due to a chromosomal deletion on 4q12; this variant is often responsive to imatinib. Myeloproliferative HES may also be associated with mutations involving platelet-derived growth factor β (PDGFRβ), Janus kinase 2 (JAK2), and fibroblast growth factor receptor 1 (FGFR1). Chronic eosinophilic leukemia with demonstrable cytogenetic abnormalities and/or blasts on peripheral smear is often categorized with the myeloproliferative HES. Clinical and laboratory findings in myeloproliferative HES may include dysplastic peripheral eosinophils, increased serum vitamin B_{12}, increased tryptase, anemia, thrombocytopenia, splenomegaly, bone marrow cellularity >80%, spindle-shaped mast cells, and myelofibrosis. The evaluation for lymphocytic HES includes searching for abnormal T cell clonal populations.

Extrapulmonary Manifestations of HES

More common in men than in women, HES occurs between the ages of 20 and 50 and is characterized by significant extrapulmonary involvement, including infiltration of the heart, GI tract, kidney, liver, joints, and skin. Cardiac involvement includes myocarditis and/or endomyocardial fibrosis, as well as a restrictive cardiomyopathy.

Pulmonary Manifestations of HES

Similar to the other pulmonary eosinophilic syndromes, these HES are manifested by high levels of blood, BAL, and tissue eosinophilia. Lung involvement occurs in 40% of these patients and is characterized by cough and dyspnea, as well as pulmonary infiltrates. Although it is often difficult to discern the pulmonary infiltrates and effusions seen on chest x-ray from pulmonary edema resulting from cardiac involvement, CT scan findings include interstitial infiltrates, ground-glass opacities, and small nodules. HES are typically not associated with ANCA. IgE may be elevated in lymphocytic HES variants.

Course and Response to Therapy

Unlike the other pulmonary eosinophilic syndromes, less than half of patients with these HES respond to corticosteroids as first-line therapy. Although other treatment options include hydroxyurea, cyclosporine, and interferon, the tyrosine kinase inhibitor imatinib has emerged as an important therapeutic option for patients with the myeloproliferative variant, particularly in individuals with the Fip1L1-PDGFRA gene fusion. Anti-IL-5 therapy with mepolizumab also holds promise for these patients and is currently being investigated.

ALLERGIC BRONCHOPULMONARY ASPERGILLOSIS

Allergic bronchopulmonary aspergillosis (ABPA) is an eosinophilic pulmonary disorder that occurs in response to allergic sensitization to antigens from *Aspergillus* species fungi. The predominant clinical presentation of ABPA is an asthmatic phenotype, often accompanied by cough with production of brownish plugs of mucus. ABPA has also been well described as a complication of cystic fibrosis. A workup for ABPA may be beneficial in patients who carry a diagnosis of asthma but have proven refractory to usual therapy. ABPA is a distinct diagnosis from simple asthma, characterized by prominent peripheral eosinophilia and elevated circulating levels of IgE (>417 IU/mL). Establishing a diagnosis of ABPA also requires establishing sensitivity to *Aspergillus* antigens by skin test reactivity, positive serum precipitins for *Aspergillus*, and/or direct measurement of circulating specific IgG and IgE to *Aspergillus*. Central bronchiectasis is described as a classic finding on chest imaging in ABPA but is not necessary for making a diagnosis. Other possible findings on chest imaging include patchy infiltrates and evidence of mucus impaction.

Systemic glucocorticoids may be used in the treatment of ABPA that is persistently symptomatic despite the use of inhaled therapies for asthma. Courses of glucocorticoids should be tapered over 3–6 months, and their use must be balanced against the risks of prolonged steroid therapy. Antifungal agents such as fluconazole and voriconazole given over a 4-month course reduce the antigenic stimulus in ABPA and may therefore modulate disease activity in selected patients. The use of monoclonal antibody against IgE (omalizumab) has been described in treating severe ABPA, particularly in individuals with ABPA as a complication of cystic fibrosis.

ABPA-like syndromes have been reported as a result to sensitization to several non-*Aspergillus* species fungi. However, these conditions are substantially rarer than ABPA, which may be present in a significant proportion of patients with refractory asthma.

INFECTIOUS PROCESSES

Infectious etiologies of pulmonary eosinophilia are largely due to helminths and are of particular importance in the evaluation of pulmonary eosinophilia in tropical environments and in the developing world (**Table 282-4**). These infectious conditions may also be considered in recent travelers to endemic regions. Loffler syndrome refers to transient pulmonary infiltrates with eosinophilia that occurs in response to passage of helminthic larvae through the lungs, most commonly larvae of *Ascaris* species (roundworm). Symptoms are generally self-limited and may include dyspnea, cough, wheeze, and hemoptysis. Loffler syndrome may also occur in response to hookworm infection with

Ancylostoma duodenale or *Necator americanus*. Chronic *Strongyloides stercoralis* infection can lead to recurrent respiratory symptoms with peripheral eosinophilia between flares. In immunocompromised hosts, including patients on glucocorticoids, a severe, potentially fatal, hyperinfection syndrome can result from *Strongyloides* infection. Paragonimiasis, filariasis, and visceral larval migrans can all cause pulmonary eosinophilia as well.

TABLE 282-4Infectious Causes of Pulmonary Eosinophilia

Löffler Syndrome

Ascaris

Hookworm

Schistosomiasis

Heavy Parasite Burden

Strongyloidiasis

Direct Pulmonary Penetration

Paragonimiasis

Visceral larval migrans

Immunologic Response to Organisms in Lungs

Filariasis

Dirofilariasis

Cystic Disease

Echinococcus

Cysticercosis

Other Nonparasitic

Coccidioidomycosis

Basidiobolomycosis

Paracoccidioidomycosis

Tuberculosis

Source: Adapted from P Akuthota, PF Weller: Clin Microbiol Rev 25:649, 2012.

DRUGS AND TOXINS

A host of medications are associated with the development of pulmonary infiltrates with peripheral eosinophilia. Therefore, drug reaction must always be included in the differential diagnosis of pulmonary eosinophilia. Although the list of medications associated with pulmonary eosinophilia is ever expanding, common culprits include nonsteroidal anti-inflammatory medications and systemic antibiotics, most specifically nitrofurantoin. Additionally, various and diverse environmental exposures such as particulate metals, scorpion stings, and inhalational drugs of abuse may also cause pulmonary eosinophilia. Radiation therapy for breast cancer has been linked with eosinophilic pulmonary infiltration as well. The mainstay of treatment is removal of the offending exposure, although glucocorticoids may be necessary if respiratory symptoms are severe.

GLOBAL CONSIDERATIONS

In the United States, drug-induced eosinophilic pneumonias are the most common cause of eosinophilic pulmonary infiltrates. A travel history or evidence of recent immigration should prompt the consideration of parasite-associated disorders. Tropical eosinophilia is usually caused by filarial infection; however, eosinophilic pneumonias also occur with other parasites such as *Ascaris* spp., *Ancylostoma* spp., *Toxocara* spp., and *Strongyloides stercoralis*. Tropical eosinophilia due to *Wuchereria bancrofti* or *Wuchereria malayi* occurs most commonly in southern Asia, Africa, and South America and is treated successfully with diethylcarbamazine. In the United States, *Strongyloides* is endemic to the southeastern and Appalachian regions.

FURTHER READING

AKUTHOTA P, WELLER PF: Eosinophilic pneumonias. Clin Microbiol Rev 25:649, 2012. [PubMed: 23034324]
COTTIN V: Eosinophilic Lung Diseases. Clin Chest Med 37:535, 2016. [PubMed: 27514599]
SELMAN M et al: Hypersensitivity pneumonitis: Insights in diagnosis and pathobiology. Am J Respir Crit Care Med 186:314, 2012. [PubMed: 22679012]

Chapter 283: Occupational and Environmental Lung Disease

John R. Balmes

INTRODUCTION

Occupational and environmental lung diseases are difficult to distinguish from those of nonenvironmental origin. Virtually all major categories of pulmonary disease can be caused by environmental agents, and environmentally related disease usually presents clinically in a manner indistinguishable from that of disease not caused by such agents. In addition, the etiology of many diseases may be multifactorial; occupational and environmental factors may interact with other factors (such as smoking and genetic risk). It is often only after a careful exposure history is taken that the underlying workplace or general environmental exposure is uncovered.

Why is knowledge of occupational or environmental etiology so important? Patient management and prognosis are affected significantly by such knowledge. For example, patients with occupational asthma or hypersensitivity pneumonitis often cannot be managed adequately without cessation of exposure to the offending agent. Establishment of cause may have significant legal and financial implications for a patient who no longer can work in his or her usual job. Other exposed people may be identified as having the disease or prevented from getting it. In addition, new associations between exposure and disease may be identified (e.g., nylon flock worker’s lung disease and diacetyl-induced bronchiolitis obliterans).

Although the exact proportion of lung disease due to occupational and environmental factors is unknown, a large number of individuals are at risk. For example, 15–20% of the burden of adult asthma and chronic obstructive pulmonary disease (COPD) has been estimated to be due to occupational factors.

HISTORY AND EXPOSURE ASSESSMENT

The patient’s history is of paramount importance in assessing any potential occupational or environmental exposure. Inquiry into specific work practices should include questions about the specific contaminants involved, the presence of visible dusts, chemical odors, the size and ventilation of workspaces, the use of respiratory protective equipment, and whether co-workers have similar complaints. The temporal association of exposure at work and symptoms may provide clues to occupation-related disease. In addition, the patient must be questioned about alternative sources of exposure to potentially toxic agents, including hobbies, home characteristics, exposure to secondhand smoke, and proximity to traffic or industrial facilities. Short-term and long-term exposures to potential toxic agents in the distant past also must be considered.

Workers in the United States have the right to know about potential hazards in their workplaces under federal Occupational Safety and Health Administration (OSHA) regulations. Employers must provide specific information about potential hazardous agents in products being used through Safety Data Sheets as well as training in personal protective equipment and environmental control procedures.

However, the introduction of new processes and/or new chemical compounds may change exposure significantly, and often only the employee on the production line is aware of the change. For the physician caring for a patient with a suspected work-related illness, a visit to the work site can be very instructive. Alternatively, an affected worker can request an inspection by OSHA. If reliable environmental sampling data are available, that information should be used in assessing a patient's exposure. Because chronic diseases may result from exposure over many years, current environmental measurements should be combined with work histories to arrive at estimates of past exposure.

LABORATORY TESTS

Exposures to inorganic and organic dusts can cause interstitial lung disease that presents with a restrictive pattern and a decreased diffusing capacity **(Chap. 279)**. Similarly, exposures to a number of dusts or chemical agents may result in occupational asthma or COPD that is characterized by airway obstruction. Measurement of change in forced expiratory volume (FEV_1) before and after a working shift can be used to detect an acute bronchoconstrictive response.

The chest radiograph is useful in detecting and monitoring the pulmonary response to mineral dusts, certain metals, and organic dusts capable of inducing hypersensitivity pneumonitis. The International Labour Organisation (ILO) International Classification of Radiographs of Pneumoconioses classifies chest radiographs by the nature and size of opacities seen and the extent of involvement of the parenchyma. In general, small rounded opacities are seen in silicosis or coal worker's pneumoconiosis, and small linear opacities are seen in asbestosis. Although useful for epidemiologic studies and screening large numbers of workers, the ILO system can be problematic when applied to an individual worker's chest radiograph. With dusts causing rounded opacities, the degree of involvement on the chest radiograph may be extensive, whereas pulmonary function may be only minimally impaired. In contrast, in pneumoconiosis causing linear, irregular opacities like those seen in asbestosis, the radiograph may lead to underestimation of the severity of the impairment until relatively late in the disease. For patients with a history of asbestos exposure, conventional computed tomography (CT) is more sensitive for the detection of pleural thickening, and high-resolution CT (HRCT) improves the detection of asbestosis.

Other procedures that may be of use in identifying the role of environmental exposures in causing lung disease include skin prick testing or specific IgE antibody titers for evidence of immediate hypersensitivity to agents capable of inducing occupational asthma (flour antigens in bakers), specific IgG precipitating antibody titers for agents capable of causing hypersensitivity pneumonitis (pigeon antigen in bird handlers), and assays for specific cell-mediated immune responses (beryllium lymphocyte proliferation testing in nuclear workers or tuberculin skin testing in health care workers). Sometimes a bronchoscopy to obtain transbronchial biopsies of lung tissue may be required for histologic diagnosis (chronic beryllium disease [CBD]). Rarely, video-assisted thoracoscopic surgery to obtain a larger sample of lung tissue may be required to determine the specific diagnosis of environmentally induced lung disease (hypersensitivity pneumonitis or giant cell interstitial pneumonitis due to cobalt exposure).

DETERMINANTS OF INHALATIONAL EXPOSURE

The chemical and physical characteristics of inhaled agents affect both the dose and the site of deposition in the respiratory tract. Water-soluble gases such as ammonia and sulfur dioxide are absorbed in the lining fluid of the upper and proximal airways and thus tend to produce irritative and bronchoconstrictive responses. In contrast, nitrogen dioxide and phosgene, which are less soluble, may penetrate to the bronchioles and alveoli in sufficient quantities to produce acute chemical pneumonitis.

Particle size of air contaminants must also be considered. Because of their settling velocities in air, particles >10–15 μm in diameter do not penetrate beyond the nose and throat. Particles <10 μm in size are deposited below the larynx. These particles are divided into three size fractions on the basis of their size characteristics and sources. Particles ~2.5–10 μm (coarse-mode fraction) contain crustal elements such as silica, aluminum, and iron. These particles mostly deposit relatively high in the tracheobronchial tree. Although the total mass of an ambient sample is dominated by these larger respirable particles, the number of particles, and therefore the surface area on which potential toxic agents can deposit and be carried to the lower airways, is dominated by particles <2.5 μm (fine-mode fraction). These fine particles are created primarily by the burning of fossil fuels or high-temperature industrial processes resulting in condensation products from gases, fumes, or vapors. The smallest particles, those <0.1 μm in size, represent the ultrafine fraction and make up the largest number of particles; they tend to remain in the airstream and deposit in the lung only on a random basis as they come into contact with the alveolar walls. If they do deposit, however, particles of this size range may penetrate into the circulation and be carried to extrapulmonary sites. New technologies create particles of this size (“nanoparticles”) for use in many commercial applications. Besides the size characteristics of particles and the solubility of gases, the actual chemical composition, mechanical properties, and immunogenicity or infectivity of inhaled material determine in large part the nature of the diseases found among exposed persons.

OCCUPATIONAL EXPOSURES AND PULMONARY DISEASE

Table 283-1 provides broad categories of exposure in the workplace and diseases associated with chronic exposure in those industries.

TABLE 283-1 Categories of Occupational Exposure and Associated Respiratory Conditions

Occupational Exposures	Nature of Respiratory Responses	Comment
Inorganic Dusts		
Asbestos: mining, processing, construction, ship repair	Fibrosis (asbestosis), pleural disease, cancer, mesothelioma	Virtually all new mining and construction with asbestos done in developing countries
Silica: mining, stone cutting, sandblasting, quarrying	Fibrosis (silicosis), progressive massive fibrosis (PMF), cancer, tuberculosis, chronic obstructive pulmonary disease (COPD)	Improved protection in United States; persistent risk in developing countries
Coal dust: mining	Fibrosis (coal worker's pneumoconiosis), PMF, COPD	Risk persists in certain areas of United States, increasing in countries where new mines open
Beryllium: processing alloys for high-tech industries	Acute pneumonitis (rare), chronic granulomatous disease, lung cancer (highly suspect)	Risk in high-tech industries persists
Other metals: aluminum, chromium, cobalt, nickel, titanium, tungsten carbide, or "hard metal" (contains cobalt)	Wide variety of conditions from acute pneumonitis to lung cancer and asthma	New diseases appear with new process development
Organic Dusts		
Cotton dust: milling, processing	Byssinosis (an asthma-like syndrome), chronic bronchitis, COPD	Increasing risk in developing countries with drop in United States as jobs shift overseas
Grain dust: elevator agents, dock workers, milling, bakers	Asthma, chronic bronchitis, COPD	Risk shifting more to migrant labor pool
Other agricultural dusts: fungal spores, vegetable products, insect fragments, animal dander, bird and rodent feces, endotoxins, microorganisms, pollens	Hypersensitivity pneumonitis (farmer's lung), asthma, chronic bronchitis	Important in migrant labor pool but also resulting from in-home exposures
Toxic chemicals: wide variety of industries; see Table 283-2	Asthma, chronic bronchitis, COPD, hypersensitivity pneumonitis, pneumoconiosis, and cancer	Reduced risk with recognized hazards; increasing risk for developing countries where controlled labor practices are less stringent
Other Environmental Agents		
Uranium and radon daughters, secondhand tobacco smoke, polycyclic aromatic hydrocarbons (PAHs), biomass smoke, diesel exhaust, welding fumes, wood finishing	Occupational exposures estimated to contribute to up to 10% of all lung cancers; chronic bronchitis, COPD, and fibrosis	In-home exposures important; in developing countries, biomass smoke is a major risk factor for COPD among women in these countries

ASBESTOS-RELATED DISEASES

Asbestos is a generic term for several different mineral silicates, including chrysolite, amosite, anthophyllite, and crocidolite. In addition to workers involved in the production of asbestos products (mining, milling, and manufacturing), many workers in the shipbuilding and construction trades, including pipe fitters and boilermakers, were occupationally exposed because asbestos was widely used during the twentieth century for its thermal and electrical insulation properties. Asbestos also was used in the manufacture of fire-resistant textiles, in cement and floor tiles, and in friction materials such as brake and clutch linings.

Exposure to asbestos is not limited to persons who directly handle the material. Cases of asbestos-related diseases have been encountered in individuals with only bystander exposure, such as painters and electricians who worked alongside insulation workers in a shipyard. Community exposure resulted from the use of asbestos-containing mine and mill tailings as landfill, road surface, and playground material (e.g., Libby, MT, the site of a vermiculite mine in which the ore was contaminated with asbestos). Finally, exposure can occur from the disturbance of naturally occurring asbestos (e.g., from increasing residential development in the foothills of the Sierra Mountains in California).

Asbestos has largely been replaced in the developed world with synthetic mineral fibers such as fiberglass and refractory ceramic fibers, but it continues to be used in the developing world. The major health effects from exposure to asbestos are pleural and pulmonary fibrosis, cancers of the respiratory tract, and pleural and peritoneal mesothelioma.

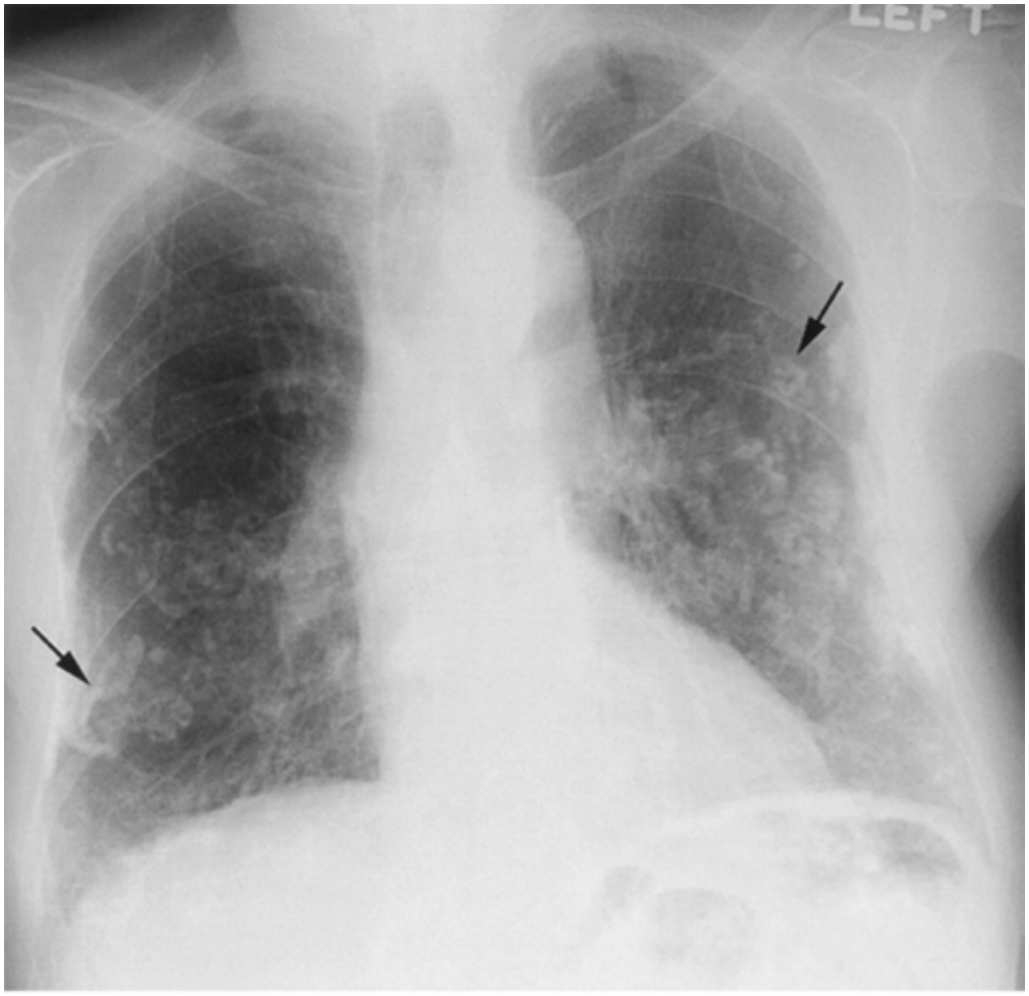
Asbestosis is a diffuse interstitial fibrosing disease of the lung that is directly related to the intensity and duration of exposure. The disease resembles other forms of diffuse interstitial fibrosis **(Chap. 287)**. Usually, exposure has taken place for at least 10 years before the disease becomes manifest. The mechanisms by which asbestos fibers induce lung fibrosis are not completely understood but are known to involve oxidative injury due to the generation of reactive oxygen species by the transition metals on the surface of the fibers as well as from cells engaged in phagocytosis.

Past exposure to asbestos is specifically indicated by pleural plaques on chest radiographs, which are characterized by either thickening or calcification along the parietal pleura, particularly along the lower lung fields, the diaphragm, and the cardiac border. Without additional manifestations, pleural plaques imply only exposure, not pulmonary impairment. Benign pleural effusions also may occur. The fluid is typically a serous or bloody exudate. The effusion may be slowly progressive or may resolve spontaneously.

Irregular or linear opacities that usually are first noted in the lower lung fields are the chest radiographic hallmark of asbestosis. An indistinct heart border or a “ground-glass” appearance in the lung fields may be seen. HRCT may show distinct changes of subpleural curvilinear lines 5–10 mm in length that appear to be parallel to the pleural surface **(Fig. 283-1)**.

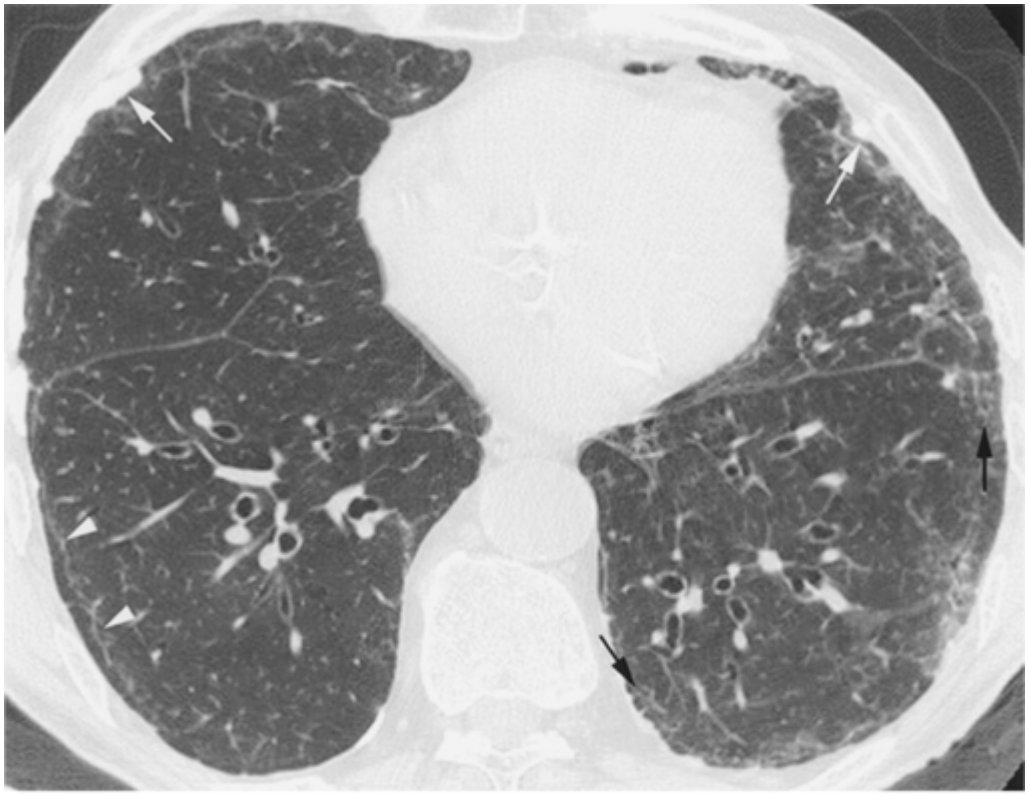
FIGURE 283-1

Asbestosis. ***A.*** Frontal chest radiograph shows bilateral calcified pleural plaques consistent with asbestos-related pleural disease. Poorly defined linear and reticular abnormalities are seen in the lower lobes bilaterally. ***B.*** Axial high-resolution computed tomography of the thorax obtained through the lung bases shows bilateral, subpleural reticulation (*black arrows*), representing fibrotic lung disease due to asbestosis. Subpleural lines are also present (*arrowheads*), characteristic of, though not specific for, asbestosis. Calcified pleural plaques representing asbestos-related pleural disease (*white arrows*) are also evident.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pulmonary function testing in asbestosis reveals a restrictive pattern with a decrease in both lung volumes and diffusing capacity. There may also be evidence of mild airflow obstruction (due to peribronchiolar fibrosis).

Because no specific therapy is available for asbestosis, supportive care is the same as that given to any patient with diffuse interstitial fibrosis of any cause. In general, newly diagnosed cases will have resulted from exposures that occurred many years before.

Lung cancer **(Chap. 74)** is the most common cancer associated with asbestos exposure. The excess frequency of lung cancer (all histologic types) in asbestos workers is associated with a minimum latency of 15–19 years between first exposure and development of the disease. Persons with more exposure are at greater risk of disease. In addition, there is a significant interactive effect of smoking and asbestos exposure that results in greater risk than what would be expected from the additive effect of each factor.

Mesotheliomas **(Chap. 288)**, both pleural and peritoneal, are also associated with asbestos exposure. In contrast to lung cancers, these tumors do not appear to be associated with smoking. Relatively short-term asbestos exposures of ≤1–2 years, occurring up to 40 years in the past, have been associated with the development of mesotheliomas (an observation that emphasizes the importance of obtaining a complete environmental exposure history). Although the risk of mesothelioma is much less than that of lung cancer among asbestos-exposed workers, >2000 cases were reported in the United States per year at the start of the twenty-first century.

Because epidemiologic studies have shown that >80% of mesotheliomas may be associated with asbestos exposure, documented mesothelioma in a patient with occupational or environmental exposure to asbestos may be compensable.

SILICOSIS

Despite being one of the oldest known occupational pulmonary hazards, *free silica* (SiO_2), or crystalline quartz, is still a major cause of disease. The major occupational exposures include mining; stonecutting; sand blasting; glass and cement manufacturing; foundry work; packing of silica flour; and quarrying, particularly of granite. Most often, pulmonary fibrosis due to silica exposure (silicosis) occurs in a dose-response fashion after many years of exposure.

Workers heavily exposed through sandblasting in confined spaces, tunneling through rock with a high quartz content (15–25%), or the manufacture of abrasive soaps may develop acute silicosis with as little as 10 months of exposure. The clinical and pathologic features of acute silicosis are similar to those of pulmonary alveolar proteinosis **(Chap. 287)**. The chest radiograph may show profuse miliary infiltration or consolidation, and there is a characteristic HRCT pattern known as “crazy paving” **(Fig. 283-2)**. The disease may be quite severe and progressive despite the discontinuation of exposure. Whole-lung lavage may provide symptomatic relief and slow the progression.

FIGURE 283-2

Acute silicosis. This high-resolution computed tomography scan shows multiple small nodules consistent with silicosis but also diffuse ground-glass densities with thickened intralobular and interlobular septa producing polygonal shapes. This has been referred to as “crazy paving.”

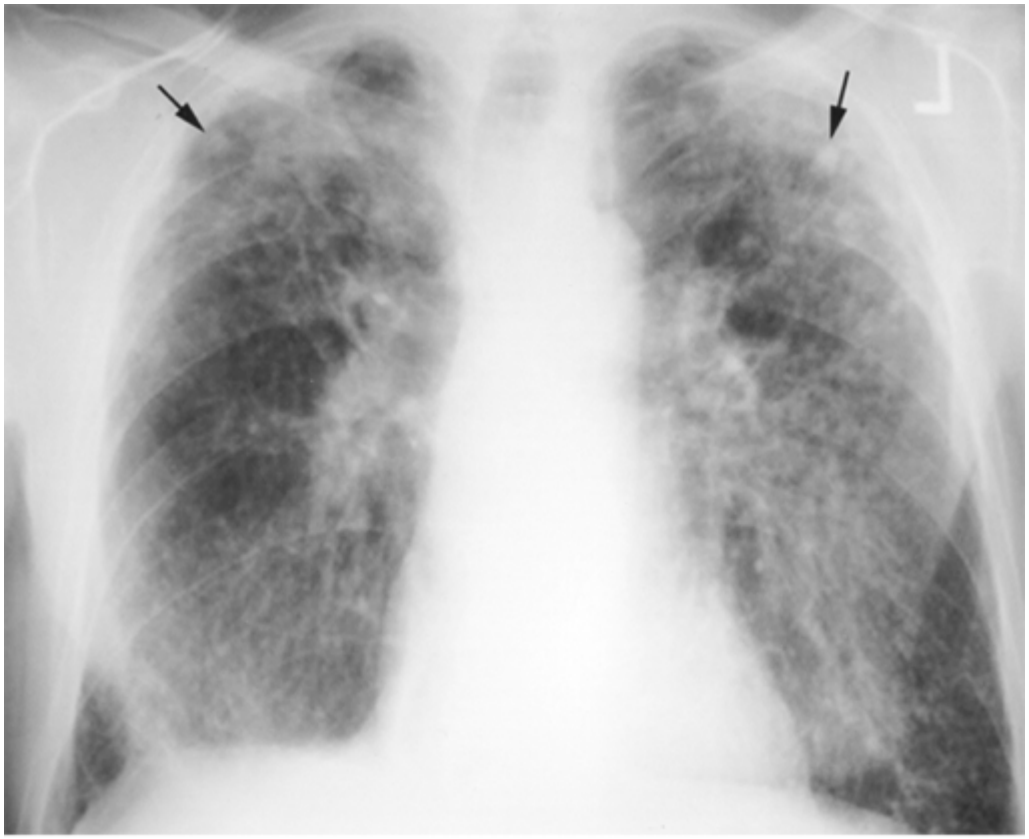


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

With long-term, less intense exposure, small rounded opacities in the upper lobes may appear on the chest radiograph after 15–20 years of exposure, usually without associated impairment of lung function (*simple silicosis*). Calcification of hilar nodes may occur in as many as 20% of cases and produces a characteristic “eggshell” pattern. Silicotic nodules may be identified more readily by HRCT **(Fig. 283-3)**. The nodular fibrosis may be progressive in the absence of further exposure, with coalescence and formation of nonsegmental conglomerates of irregular masses >1 cm in diameter (*complicated silicosis*). These masses can become quite large, and when this occurs, the term *progressive massive fibrosis* (PMF) is applied. Significant functional impairment with both restrictive and obstructive components may be associated with PMF.

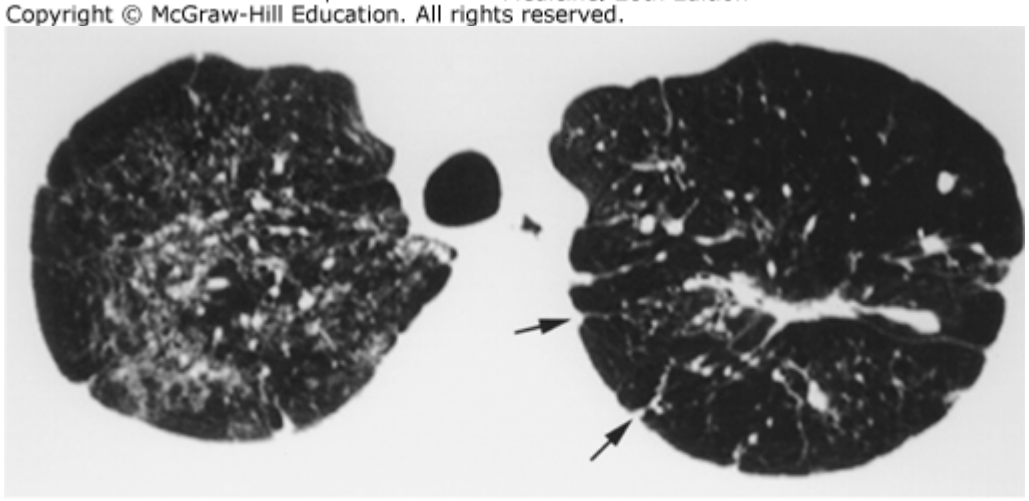
FIGURE 283-3

Chronic silicosis. ***A.*** Frontal chest radiograph in a patient with silicosis shows variably sized, poorly defined nodules (*arrows*) predominating in the upper lobes. ***B.*** Axial thoracic computed tomography image through the lung apices shows numerous small nodules, more pronounced in the right upper lobe. A number of the nodules are subpleural in location (*arrows*).



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Because silica causes alveolar macrophage dysfunction, patients with silicosis are at greater risk of acquiring lung infections that involve these cells as a primary defense (*Mycobacterium tuberculosis*, atypical mycobacteria and fungi). Because of the increased risk of active tuberculosis, the recommended treatment of latent tuberculosis in these patients is longer. Silica has immunoadjuvant properties and another potential clinical complication of silicosis is autoimmune connective tissue disorders such as rheumatoid arthritis and scleroderma. In addition, there are sufficient epidemiologic data that the International Agency for Research on Cancer lists silica as a probable lung carcinogen.

Other, less hazardous silicates include fuller's earth, kaolin, mica, diatomaceous earths, silica gel, soapstone, carbonate dusts, and cement dusts. The production of fibrosis in workers exposed to these agents is believed to be related either to the free silica content of these dusts or, for substances that contain no free silica, to the potentially large dust loads to which these workers may be exposed. Some silicates, including *talc* and *vermiculite*, may be contaminated with asbestos. Fibrosis of lung or pleura, lung cancer, and mesothelioma have been associated with chronic exposure to talc and vermiculite dusts.

COAL WORKER'S PNEUMOCONIOSIS (CWP)

Occupational exposure to *coal dust* can lead to CWP, which has enormous social, economic, and medical significance in every nation in which coal mining is an important industry. Simple radiographically identified CWP is seen in ~10% of all coal miners and in as many as 50% of anthracite miners with >20 years of work on the coal face. The prevalence of disease is lower in workers in bituminous coal mines.

With prolonged exposure to coal dust (i.e., 15–20 years), small, rounded opacities similar to those of silicosis may develop. As in silicosis, the presence of these nodules (*simple CWP*) usually is not associated with pulmonary impairment. In addition to CWP, coal dust can cause chronic bronchitis and COPD **(Chap. 286)**. The effects of coal dust are additive to those of cigarette smoking.

Complicated CWP is manifested by the appearance on the chest radiograph of nodules ≥1 cm in diameter generally confined to the upper half of the lungs. As in silicosis, this condition can progress to PMF that is accompanied by severe lung function deficits and associated with premature mortality. Despite improvements in technology to protect coal miners, cases of PMF still occur in the United States at a disturbing rate.

Caplan syndrome **(Chap. 351)**, first described in coal miners but subsequently in patients with silicosis, is the combination of pneumoconiotic nodules and seropositive rheumatoid arthritis. Silica is often present in anthracitic coal dust and its presence may contribute to risk of PMF.

CHRONIC BERYLLIUM DISEASE

Beryllium is a lightweight metal with tensile strength, good electrical conductivity, and value in the control of nuclear reactions through its ability to quench neutrons. Although beryllium may produce an acute pneumonitis, it is far more commonly associated with a chronic granulomatous inflammatory disease that is similar to sarcoidosis **(Chap. 360)**. Unless one inquires specifically about occupational exposures to beryllium in the manufacture of alloys, ceramics, or high-technology electronics in a patient with sarcoidosis, one may miss entirely the etiologic relationship to the occupational exposure. What distinguishes CBD from sarcoidosis is evidence of a specific cell-mediated immune response (i.e., delayed hypersensitivity) to beryllium.

The test that usually provides this evidence is the beryllium lymphocyte proliferation test (BeLPT). The BeLPT compares the in vitro proliferation of lymphocytes from blood or bronchoalveolar lavage in the presence of beryllium salts with that of unstimulated cells. Proliferation is usually measured by lymphocyte uptake of radiolabeled thymidine.

Chest imaging findings are similar to those of sarcoidosis (nodules along septal lines) except that hilar adenopathy is somewhat less common. As with sarcoidosis, pulmonary function test results may show restrictive and/or obstructive ventilatory deficits and decreased diffusing capacity. With early disease, both chest imaging studies and pulmonary function tests may be normal. Fiberoptic bronchoscopy with transbronchial lung biopsy usually is required to make the diagnosis of CBD. In a beryllium-sensitized individual, the presence of noncaseating granulomas or monocytic infiltration in lung tissue establishes the diagnosis. Accumulation of beryllium-specific CD4+ T cells occurs in the

granulomatous inflammation seen on lung biopsy. Susceptibility to CBD is highly associated with human leukocyte antigen DP (HLA-DP) alleles that have a glutamic acid in position 69 of the β chain.

OTHER METALS

Aluminum and titanium dioxide have been rarely associated with a sarcoid-like reaction in lung tissue. Exposure to dust containing tungsten carbide, also known as “hard metal,” may produce giant cell interstitial pneumonitis. Cobalt is a constituent of tungsten carbide and is the likely etiologic agent of both the interstitial pneumonitis and the occupational asthma that may occur. The most common exposures to tungsten carbide occur in tool and dye, saw blade, and drill bit manufacture. Diamond polishing may also involve exposure to cobalt dust. In patients with interstitial lung disease, one should always inquire about exposure to metal fumes and/or dusts. Especially when sarcoidosis appears to be the diagnosis, one should always consider possible CBD.

OTHER INORGANIC DUSTS

Most of the inorganic dusts discussed thus far are associated with the production of either dust macules or interstitial fibrotic changes in the lung. Other inorganic and organic dusts (see categories in Table 283-1), along with some of the dusts previously discussed, are associated with chronic mucus hypersecretion (chronic bronchitis), with or without reduction of expiratory flow rates. Cigarette smoking is the major cause of these conditions, and any effort to attribute some component of the disease to occupational and environmental exposures must take cigarette smoking into account. Most studies suggest an additive effect of dust exposure and smoking. The pattern of the irritant dust effect is similar to that of cigarette smoking, suggesting that small airway inflammation may be the initial site of pathologic response in those cases and continued exposure may lead to chronic bronchitis and COPD.

ORGANIC DUSTS

Some of the specific diseases associated with organic dusts are discussed in detail in the chapters on asthma **(Chap. 281)** and hypersensitivity pneumonitis **(Chap. 282)**. Many of these diseases are named for the specific setting in which they are found, e.g., farmer’s lung, malt worker’s disease, and mushroom worker’s disease. Often the temporal relation of symptoms to exposure furnishes the best evidence for the diagnosis. Three occupational exposures are singled out for discussion here because they affect the largest proportions of workers.

Cotton Dust (Byssinosis)

Workers occupationally exposed to cotton dust (but also to flax, hemp, or jute dust) in the production of yarns for textiles and rope making are at risk for an asthma-like syndrome known as byssinosis. The risk of byssinosis is associated with both cotton dust and endotoxin levels in the workplace environment.

Byssinosis is characterized clinically as occasional (early-stage) and then regular (late-stage) chest tightness toward the end of the first day of the workweek (“Monday chest tightness”). Exposed workers may show a significant drop in FEV_1 over the course of a Monday workshift. Initially the symptoms do not recur on subsequent days of the week, but in a subset of workers, chest tightness may recur or persist throughout the workweek. After >10 years of exposure, workers with recurrent symptoms are more likely to have an obstructive pattern on pulmonary function testing.

Dust exposure can be reduced by the use of exhaust hoods, general increases in ventilation, and wetting procedures, but respiratory protective equipment may be required during certain operations. Regular surveillance of pulmonary function in cotton dust–exposed workers using spirometry before

and after the workshift is required by OSHA. All workers with persistent symptoms or significantly reduced levels of pulmonary function should be moved to areas of lower risk of exposure.

Grain Dust

Worldwide, many farmers and workers in grain storage facilities are exposed to grain dust. The presentation of obstructive airway disease in grain dust–exposed workers is virtually identical to the characteristic findings in cigarette smokers, i.e., persistent cough, mucus hypersecretion, wheeze and dyspnea on exertion, and reduced FEV_1 and FEV_1/FVC (forced vital capacity) ratio **(Chap. 279)**.

Dust concentrations in grain elevators vary greatly but can be >10,000 $\mu g/m^3$ with many particles in the respirable size range. The effect of grain dust exposure is additive to that of cigarette smoking, with ~50% of workers who smoke having symptoms. Smoking grain dust–exposed workers are more likely to have obstructive ventilatory deficits on pulmonary function testing. As in byssinosis, endotoxin may play a role in grain dust–induced chronic bronchitis and COPD.

Farmer's Lung

This condition results from exposure to moldy hay containing spores of thermophilic actinomycetes that produce a hypersensitivity pneumonitis **(Chap. 282)**. A patient with acute farmer's lung presents 4–8 h after exposure with fever, chills, malaise, cough, and dyspnea without wheezing. The history of exposure is obviously essential to distinguish this disease from influenza or pneumonia with similar symptoms. In the chronic form of the disease, the history of repeated attacks after similar exposure is important in differentiating this syndrome from other causes of patchy fibrosis (e.g., sarcoidosis).

A wide variety of other organic dusts are associated with the occurrence of hypersensitivity pneumonitis **(Chap. 282)**. For patients who present with hypersensitivity pneumonitis, specific and careful inquiry about occupations, hobbies, and other home environmental exposures is necessary to uncover the source of the etiologic agent.

TOXIC CHEMICALS

Exposure to toxic chemicals affecting the lung generally involves gases and vapors. A common accident is one in which the victim is trapped in a confined space where the chemicals have accumulated to harmful levels. In addition to the specific toxic effects of the chemical, the victim often sustains considerable anoxia, which can play a dominant role in determining whether the individual survives.

Table 283-2 lists a variety of toxic agents that can produce acute and sometimes life-threatening reactions in the lung. All these agents in sufficient concentrations have been demonstrated, at least in animal studies, to affect the lower airways and disrupt alveolar architecture, either acutely or as a result of chronic exposure. Some of these agents may be generated acutely in the environment (see below).

TABLE 283-2 Selected Common Toxic Chemical Agents That Affect the Lung

Agent(s)	Selected Exposures	Acute Effects from High or Accidental Exposure	Chronic Effects from Relatively Low Exposure
Acid anhydrides	Manufacture of resin esters, polyester resins, thermoactivated adhesives	Nasal irritation, cough	Asthma, chronic bronchitis, hypersensitivity pneumonitis
Acid fumes: H_2SO_4, HNO_3	Manufacture of fertilizers, chlorinated organic compounds, dyes, explosives, rubber products, metal etching, plastics	Mucous membrane irritation, followed by chemical pneumonitis 2–3 days later	Bronchitis and suggestion of mildly reduced pulmonary function in children with lifelong residential exposure to high levels
Acrolein and other aldehydes	By-product of burning plastics, woods, tobacco smoke	Mucous membrane irritant, decrease in lung function	Upper respiratory tract irritation
Ammonia	Refrigeration; petroleum refining; manufacture of fertilizers, explosives, plastics, and other chemicals	Same as for acid fumes, but bronchiectasis also has been reported	Upper respiratory tract irritation, chronic bronchitis
Cadmium fumes	Smelting, soldering, battery production	Mucous membrane irritant, acute respiratory distress syndrome (ARDS)	Chronic obstructive pulmonary disease (COPD)
Formaldehyde	Manufacture of resins, leathers, rubber, metals, and woods; laboratory workers, embalmers; emission from urethane foam insulation	Same as for acid fumes	Nasopharyngeal cancer
Halides and acid salts (Cl, Br, F)	Bleaching in pulp, paper, textile industry; manufacture of chemical compounds; synthetic rubber, plastics, disinfectant, rocket fuel, gasoline	Mucous membrane irritation, pulmonary edema; possible reduced forced vital capacity (FVC) 1–2 years after exposure	Upper respiratory tract irritation, epistaxis, tracheobronchitis
Hydrogen sulfide	By-product of many industrial processes, oil, other petroleum processes and storage	Increase in respiratory rate followed by respiratory arrest, lactic acidosis, pulmonary edema, death	Conjunctival irritation, chronic bronchitis, recurrent pneumonitis
Isocyanates (TDI, HDI, MDI)	Production of polyurethane foams, plastics, adhesives, surface coatings	Mucous membrane irritation, dyspnea, cough, wheeze, pulmonary edema	Upper respiratory tract irritation, cough, asthma, hypersensitivity pneumonitis, reduced lung function
Nitrogen dioxide	Silage, metal etching, explosives, rocket fuels, welding, by-product of burning fossil fuels	Cough, dyspnea, pulmonary edema may be delayed 4–12 h; possible result from acute exposure: bronchiolitis obliterans in 2–6 weeks	Emphysema in animals, chronic bronchitis, associated with reduced lung function growth in children with lifelong residential exposure

Agent(s)	Selected Exposures	Acute Effects from High or Accidental Exposure	Chronic Effects from Relatively Low Exposure
Ozone	Arc welding, flour bleaching, deodorizing, emissions from copying equipment, photochemical air pollutant	Mucous membrane irritant, reduced pulmonary function transiently in children and adults, asthma exacerbation	Excess cardiopulmonary mortality rates, increased risk for new-onset asthma in children
Phosgene	Organic compound, metallurgy, volatilization of chlorine-containing compounds	Delayed onset of bronchiolitis and pulmonary edema	Chronic bronchitis
Sulfur dioxide	Manufacture of sulfuric acid, bleaches, coating of nonferrous metals, food processing, refrigerant, burning of fossil fuels, wood pulp industry	Mucous membrane irritant, epistaxis, bronchospasm (especially in people with asthma)	Chronic bronchitis

Abbreviations: HDI, hexamethylene diisocyanate; MDI, methylene diphenyl diisocyanate; TDI, toluene diisocyanate.

Firefighters and fire victims are at risk of *smoke inhalation*, an important cause of acute cardiorespiratory failure. Smoke inhalation kills more fire victims than does thermal injury. Carbon monoxide poisoning with resulting significant hypoxemia can be life-threatening **(Chap. 450)**. Synthetic materials (plastic, polyurethanes), when burned, may release a variety of other toxic agents (such as cyanide and hydrochloric acid), and this must be considered in evaluating smoke inhalation victims. Exposed victims may have some degree of lower respiratory tract inflammation and/or pulmonary edema.

Exposure to certain highly reactive, low-molecular-weight agents used in the manufacture of synthetic polymers, paints, and coatings (*diisocyanates* in polyurethanes, *aromatic amines* and *acid anhydrides* in epoxies) is associated with a high risk of occupational asthma. Although this occupational asthma manifests clinically as if sensitization has occurred, an IgE antibody–mediated mechanism is not necessarily involved. Hypersensitivity pneumonitis–like reactions also have been described in diisocyanate and acid anhydride–exposed workers.

Fluoropolymers such as Teflon, which at normal temperatures produce no reaction, become volatilized upon heating. The inhaled agents cause a characteristic syndrome of fever, chills, malaise, and occasionally mild wheezing, leading to the diagnosis of *polymer fume fever*. A similar self-limited, influenza-like syndrome—*metal fume fever*—results from acute exposure to fumes containing zinc oxide, typically from welding of galvanized steel. These inhalational fever syndromes may begin several hours after work and resolve within 24 h, only to return on repeated exposure.

Two other agents have been associated with potentially severe lung disease. Occupational exposure to nylon flock has been shown to induce a lymphocytic bronchiolitis, and workers exposed to diacetyl, which is used to provide “butter” flavor in the manufacture of microwave popcorn and other foods, have developed bronchiolitis obliterans **(Chap. 287)**.

World Trade Center Disaster

A consequence of the attack on the World Trade Center (WTC) on September 11, 2001, was relatively heavy exposure of a large number of firefighters and other rescue workers to the dust generated by the collapse of the buildings. Environmental monitoring and chemical characterization of WTC dust

has revealed a wide variety of potentially toxic constituents, although much of the dust was pulverized cement. Possibly because of the high alkalinity of WTC dust, significant cough, wheeze, and phlegm production occurred among firefighters and cleanup crews. New cough and wheeze syndromes also occurred among local residents. Heavier exposure to WTC dust among New York City firefighters was associated with accelerated decline of lung function over the first year after the disaster. More recently, concerns have been raised about risk of interstitial lung disease, especially of a granulomatous nature.

OCCUPATIONAL RESPIRATORY CARCINOGENS

Exposures at work have been estimated to contribute to 10% of all lung cancer cases. In addition to asbestos, other agents either proven or suspected to be respiratory carcinogens include acrylonitrile, arsenic compounds, beryllium, bis(chloromethyl) ether, chromium (hexavalent), formaldehyde (nasal), isopropanol (nasal sinuses), mustard gas, nickel carbonyl (nickel smelting), polycyclic aromatic hydrocarbons (coke oven emissions and diesel exhaust), secondhand tobacco smoke, silica (both mining and processing), talc (possible asbestos contamination in both mining and milling), vinyl chloride (sarcomas), wood (nasal), and uranium. Workers at risk of radiation-related lung cancer include not only those involved in mining or processing uranium but also those exposed in underground mining operations of other ores where radon daughters may be emitted from rock formations.

ASSESSMENT OF DISABILITY

Disability is the term used to describe the decreased ability to work due to the effects of a medical condition. Physicians are generally able to assess physiologic dysfunction, or *impairment*, but the rating of disability for compensation of loss of income also involves nonmedical factors such as the education and employability of the individual. The disability rating scheme differs with the compensation-granting agency. For example, the U.S. Social Security Administration requires that an individual be unable to do any work (i.e., *total* disability) before he or she will receive income replacement payments. Many state workers' compensation systems allow for payments for *partial* disability. In the Social Security scheme, no determination of cause is done, whereas work-relatedness must be established in workers' compensation systems.

For respiratory impairment rating, resting pulmonary function tests (spirometry and diffusing capacity) are used as the initial assessment tool, with cardiopulmonary exercise testing (to assess maximal oxygen consumption) used if the results of the resting tests do not correlate with the patient's symptoms. Methacholine challenge (to assess airway reactivity) can also be useful in patients with asthma who have normal spirometry when evaluated. Some compensation agencies (e.g., Social Security) have proscribed disability classification schemes based on pulmonary function test results. When no specific scheme is proscribed, the *Guidelines of the American Medical Association* should be used.

GENERAL ENVIRONMENTAL EXPOSURES

OUTDOOR AIR POLLUTION

In 1971, the U.S. government established national air quality standards for several pollutants believed to be responsible for excess cardiorespiratory diseases. Primary standards regulated by the U.S. Environmental Protection Agency (EPA) designed to protect the public health with an adequate margin of safety exist for sulfur dioxide, particulate matter, nitrogen dioxide, ozone, lead, and carbon monoxide. Standards for each of these pollutants are updated regularly through an extensive review

process conducted by the EPA. (For details on current standards, go to *https://www.epa.gov/criteria-air-pollutants/naaqs-table%20*.

Pollutants are generated from both stationary sources (power plants and industrial complexes) and mobile sources (motor vehicles), and none of the regulated pollutants occurs in isolation. Furthermore, pollutants may be changed by chemical reactions after being emitted. For example, sulfur dioxide and particulate matter emissions from a coal-fired power plant may react in air to produce acid sulfates and aerosols, which can be transported long distances in the atmosphere. Oxides of nitrogen and volatile organic compounds from automobile exhaust react with sunlight to produce ozone. Although originally thought to be confined to Los Angeles, photochemically derived pollution (“smog”) is now known to be a problem throughout the United States and in many other countries. Both acute and chronic effects of these exposures have been documented in large population studies.

The symptoms and diseases associated with air pollution are the same as conditions commonly associated with cigarette smoking. In addition, decreased growth of lung function and asthma have been associated with chronic exposure to only modestly elevated levels of traffic-related gases and respirable particles. Multiple population-based time-series studies within cities have demonstrated excess health care utilization for asthma and other cardiopulmonary conditions as well as increased mortality rates. Cohort studies comparing cities that have relatively high levels of particulate exposures with less polluted communities suggest excess morbidity and mortality rates from cardiopulmonary conditions in long-term residents of the former. The strong epidemiologic evidence that fine particulate matter is a risk factor for cardiovascular morbidity and mortality has prompted toxicologic investigations into the underlying mechanisms. The inhalation of fine particles from combustion sources probably generates oxidative stress followed by local injury and inflammation in the lungs that in turn lead to autonomic and systemic inflammatory responses that can induce endothelial dysfunction and/or injury. Recent research findings on the health effects of air pollutants has led to stricter U.S. ambient air quality standards for ozone, oxides of nitrogen, and particulate matter as well as greater emphasis on publicizing pollution alerts to encourage individuals with significant cardiopulmonary impairment to stay indoors during high-pollution episodes.

INDOOR EXPOSURES

Secondhand tobacco smoke **(Chap. 448)**, radon gas, wood smoke, and other biologic agents generated indoors must be considered. Several studies have shown that the respirable particulate load in any household is directly proportional to the number of cigarette smokers living in that home. Increases in prevalence of respiratory illnesses, especially asthma, and reduced levels of pulmonary function measured with simple spirometry have been found in the children of smoking parents in a number of studies. Recent meta-analyses for lung cancer and cardiopulmonary diseases, combining data from multiple secondhand tobacco smoke epidemiologic studies, suggest an ~25% increase in relative risk for each condition, even after adjustment for major potential confounders.

Exposure to *radon gas* in homes is a risk factor for lung cancer. The main radon product (radon-222) is a gas that results from the decay series of uranium-238, with the immediate precursor being radium-226. The amount of radium in earth materials determines how much radon gas will be emitted. Levels associated with excess lung cancer risk may be present in as many as 10% of the houses in the United States. When smokers reside in the home, the problem is potentially greater, because the molecular size of radon particles allows them to attach readily to smoke particles that are inhaled. Fortunately, technology is available for assessing and reducing the level of exposure.

Other indoor exposures of concern are bioaerosols that contain antigenic material (fungi, cockroaches, dust mites, and pet danders) associated with an increased risk of atopy and asthma. Indoor chemical agents include strong cleaning agents (bleach, ammonia), formaldehyde, perfumes, pesticides, and oxides of nitrogen from gas appliances. Nonspecific responses associated with “tight-building syndrome,” perhaps better termed “building-associated illness,” in which no particular agent

has been implicated, have included a wide variety of complaints, among them respiratory symptoms that are relieved only by avoiding exposure in the building in question. The degree to which “smells” and other sensory stimuli are involved in the triggering of potentially incapacitating psychological or physical responses has yet to be determined, and the long-term consequences of such environmental exposures are unknown.

GLOBAL CONSIDERATIONS

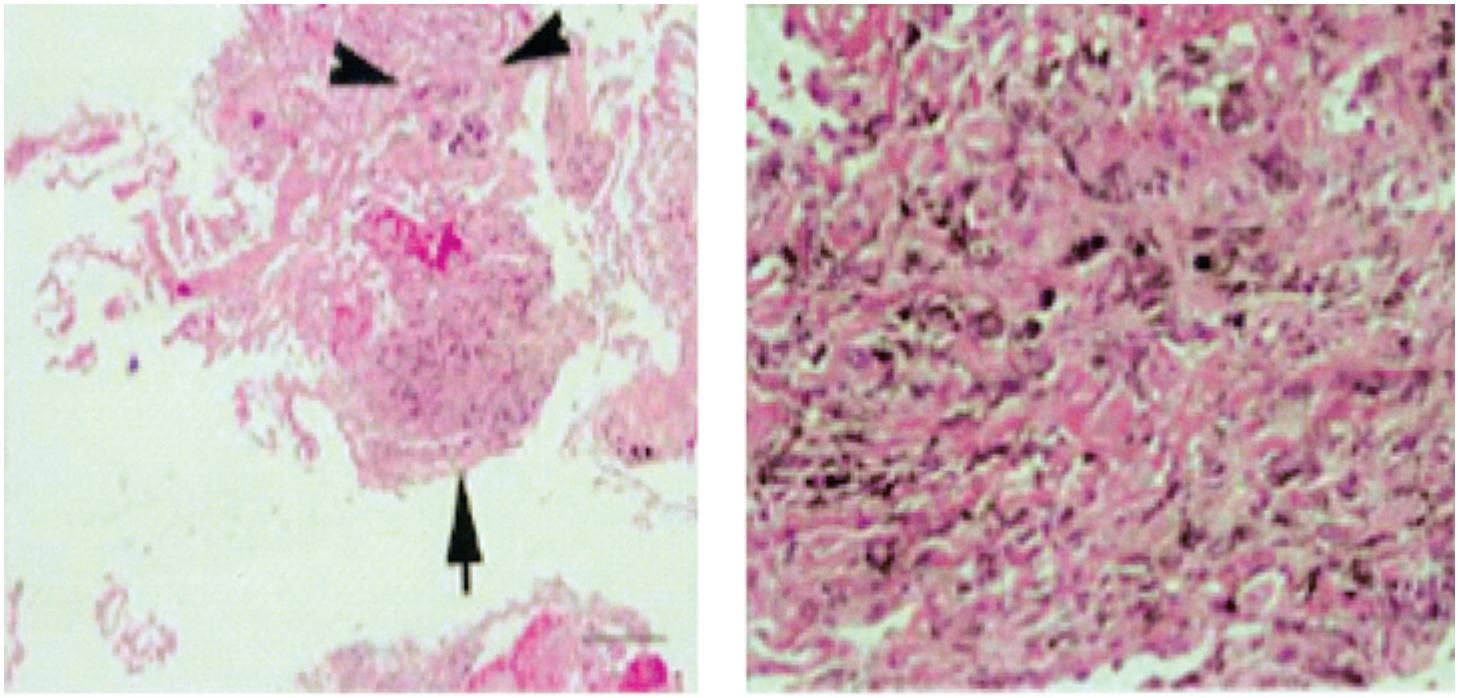
Indoor exposure to *household air pollution* from cooking or heating with solid fuels (wood, dung, crop residues, charcoal, coal) is estimated to be responsible for >4% of worldwide disability-adjusted life-years (DALYs) lost, due to acute lower respiratory infections in children, COPD and lung cancer in women, and cardiovascular disease among men. This burden of disease places exposure to household air pollution as the leading environmental hazard for poor health on a global scale.

Forty percent of the world’s population uses solid fuel for cooking, heating, or baking. Kerosene (similar to diesel fuel) is often used for lighting and sometimes cooking. This occurs predominantly in the rural areas of developing countries. Because many families burn coal or biomass fuels in open stoves, which are highly inefficient, and inside homes with poor ventilation, women and young children are exposed on a daily basis to high levels of smoke. In these homes, 24-h mean levels of fine particulate matter have been reported to be 2–30 times higher than the National Ambient Air Quality Standard set by the U.S. EPA.

Epidemiologic studies have consistently shown associations between exposure to biomass smoke and both chronic bronchitis and COPD. Because of increased migration to the United States from developing countries, clinicians need to be aware of the chronic respiratory effects of exposure to biomass smoke, which can include interstitial lung disease **(Fig. 283-4)**. Evidence is beginning to emerge that improved stoves that reduce biomass smoke exposure can reduce risk of respiratory illness in both children and adults.

FIGURE 283-4

Histopathologic features of biomass smoke–induced interstitial lung disease. ***A.*** Anthracitic pigment is seen accumulating along alveolar septae (*arrowheads*) and within a pigmented dust macule (*single arrow*). ***B.*** A high-power photomicrograph contains a mixture of fibroblasts and carbon-laden macrophages.



A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Household air pollution (HAP) from domestic use of solid fuels also contributes substantially to outdoor air pollution. Contributions from HAP, coal-fired power plants without emission scrubbers, and increased traffic congestion involving motor vehicles without pollution controls can lead to high concentrations of outdoor air pollution, especially fine particulate matter, in mega-cities in developing countries (e.g., Delhi).

ACKNOWLEDGMENT

The author acknowledges the contribution of Dr. Frank Speizer to the prior version of this chapter.

FURTHER READING

Banks DE: Clinical aspects of asbestos-related diseases—What are the unresolved topics? J Occup Environ Med 56 Suppl 10:S8, 2014. [PubMed: 25285978]
Blanc PD, Torén K: COPD and occupation: Resetting the agenda. Occup Environ Med 73:357, 2016. [PubMed: 27084077]
Gauderman WJ et al: Association of improved air quality with lung development in children. N Engl J Med 372:905, 2015. [PubMed: 25738666]
Gordon SB et al: Respiratory risks from household air pollution in low and middle income countries. Lancet Respir Med 2:823, 2014. [PubMed: 25193349]
Petsonk EL et al: Coal mine dust lung disease. New lessons from old exposure. Am J Respir Crit Care Med 187:1178, 2013. [PubMed: 23590267]

Chapter 284: Bronchiectasis

Rebecca M. Baron; Miriam Baron Barshak

FIGURE 284-1

INTRODUCTION

Bronchiectasis refers to an irreversible airway dilation that involves the lung in either a focal or a diffuse manner and that classically has been categorized as cylindrical or tubular (the most common form), varicose, or cystic.

ETIOLOGY

Bronchiectasis can arise from infectious or noninfectious causes **(Table 284-1)**. Clues to the underlying etiology are often provided by the pattern of lung involvement. *Focal bronchiectasis* refers to bronchiectatic changes in a localized area of the lung and can be a consequence of obstruction of the airway—either extrinsic (e.g., due to compression by adjacent lymphadenopathy or parenchymal tumor mass) or intrinsic (e.g., due to an airway tumor or aspirated foreign body, a scarred/stenotic airway, or bronchial atresia from congenital underdevelopment of the airway). *Diffuse bronchiectasis* is characterized by widespread bronchiectatic changes throughout the lung and often arises from an underlying systemic or infectious disease process.

TABLE 284-1 Major Etiologies of Bronchiectasis and Proposed Workup

Pattern of Lung Involvement	Etiology by Category (Examples)	Workup
Focal	Obstruction (aspirated foreign body, tumor mass)	Chest imaging (chest x-ray and/or chest CT); bronchoscopy
Diffuse	Infection (bacterial, nontuberculous mycobacterial)	Sputum Gram's stain/culture; stains/cultures for acid-fast bacilli and fungi. If no pathogen is identified, consider bronchoscopy with bronchoalveolar lavage.
	Immunodeficiency (hypogammaglobulinemia, HIV infection, bronchiolitis obliterans after lung transplantation)	Complete blood count with differential; immunoglobulin measurement; HIV testing
	Genetic causes (cystic fibrosis, Kartagener's syndrome, α_1 antitrypsin deficiency)	Measurement of chloride levels in sweat (for cystic fibrosis), α_1 antitrypsin levels; nasal or respiratory tract brush/biopsy (for dyskinetic/immotile cilia syndrome); genetic testing
	Autoimmune or rheumatologic causes (rheumatoid arthritis, Sjögren's syndrome, inflammatory bowel disease); immune-mediated disease (allergic bronchopulmonary aspergillosis)	Clinical examination with careful joint exam, serologic testing (e.g., for rheumatoid factor). Consider workup for allergic bronchopulmonary aspergillosis, especially in patients with refractory asthma.[a]
	Recurrent aspiration	Test of swallowing function and general neuromuscular strength
	Miscellaneous (yellow nail syndrome, traction bronchiectasis from postradiation fibrosis or idiopathic pulmonary fibrosis)	Guided by clinical condition
	Idiopathic	Exclusion of other causes

[a]Skin testing for *Aspergillus* reactivity; measurement of serum precipitins for *Aspergillus*, serum IgE levels, serum eosinophils, etc.

More pronounced involvement of the upper lung fields is most common in cystic fibrosis (CF) and is also observed in postradiation fibrosis, corresponding to the lung region encompassed by the radiation port. Bronchiectasis with predominant involvement of the lower lung fields usually has its source in chronic recurrent aspiration (e.g., due to esophageal motility disorders like those in scleroderma), end-stage fibrotic lung disease (e.g., traction bronchiectasis from idiopathic pulmonary fibrosis), or recurrent immunodeficiency-associated infections (e.g., hypogammaglobulinemia). Bronchiectasis resulting from infection by nontuberculous mycobacteria (NTM), most commonly the *Mycobacterium avium-intracellulare* complex (MAC), often preferentially affects the midlung fields. Congenital causes of bronchiectasis with predominant midlung field involvement include the dyskinetic/immotile cilia syndrome. Finally, predominant involvement of the central airways is reported in association with allergic bronchopulmonary aspergillosis (ABPA), in which an immune-mediated reaction to *Aspergillus* damages the bronchial wall. Congenital causes of central airway–predominant bronchiectasis resulting from cartilage deficiency include tracheobronchomegaly (Mounier-Kuhn syndrome) and Williams-Campbell syndrome.

In many cases, the etiology of bronchiectasis is not determined. In case series, as many as 25–50% of patients referred for bronchiectasis have idiopathic disease.

EPIDEMIOLOGY

The overall reported prevalence of bronchiectasis in the United States has recently increased, but the epidemiology of bronchiectasis varies greatly with the underlying etiology. For example, patients born with CF often develop significant clinical bronchiectasis in late adolescence or early adulthood, although atypical presentations of CF in adults in their thirties and forties are also possible. In contrast, bronchiectasis resulting from MAC infection classically affects nonsmoking women >50 years of age. In general, the incidence of bronchiectasis increases with age. Bronchiectasis is more common among women than among men.

In areas where tuberculosis is prevalent, bronchiectasis more frequently occurs as a sequela of granulomatous infection. Focal bronchiectasis can arise from extrinsic compression of the airway by enlarged granulomatous lymph nodes and/or from development of intrinsic obstruction as a result of erosion of a calcified lymph node through the airway wall (e.g., broncholithiasis). Especially in reactivated tuberculosis, parenchymal destruction from infection can result in areas of more diffuse bronchiectasis. Apart from cases associated with tuberculosis, an increased incidence of non-CF bronchiectasis with an unclear underlying mechanism has been reported as a significant problem in developing nations. It has been suggested that the high incidence of malnutrition in certain areas may predispose to immune dysfunction and development of bronchiectasis.

PATHOGENESIS AND PATHOLOGY

The most widely cited mechanism of infectious bronchiectasis is the “vicious cycle hypothesis,” in which susceptibility to infection and poor mucociliary clearance result in microbial colonization of the bronchial tree. Some organisms, such as *Pseudomonas aeruginosa*, exhibit a particular propensity for colonizing damaged airways and evading host defense mechanisms. Impaired mucociliary clearance can result from inherited conditions such as CF or dyskinetic cilia syndrome, and it has been proposed that a single severe infection (e.g., pneumonia caused by *Bordetella pertussis* or *Mycoplasma pneumoniae*) can result in significant airway damage and poor secretion clearance. The presence of the microbes incites continued chronic inflammation, with consequent damage to the airway wall, continued impairment of secretions and microbial clearance, and ongoing propagation of the infectious/inflammatory cycle. Moreover, it has been proposed that mediators released directly from bacteria can interfere with mucociliary clearance.

Classic studies of the pathology of bronchiectasis from the 1950s demonstrated significant small-airway wall inflammation and larger-airway wall destruction as well as dilation, with loss of elastin, smooth muscle, and cartilage. It has been proposed that inflammatory cells in the small airways release proteases and other mediators, such as reactive oxygen species and proinflammatory cytokines, that damage the larger-airway walls. Furthermore, the ongoing inflammatory process in the smaller airways results in airflow obstruction. It is thought that antiproteases, such as α_1 antitrypsin, play an important role in neutralizing the damaging effects of neutrophil elastase and in enhancing bacterial killing. Bronchiectasis and emphysema have been observed in patients with α_1 antitrypsin deficiency.

Proposed mechanisms for noninfectious bronchiectasis include immune-mediated reactions that damage the bronchial wall (e.g., those associated with systemic autoimmune conditions such as Sjögren’s syndrome and rheumatoid arthritis). *Traction bronchiectasis* refers to dilated airways arising from parenchymal distortion as a result of lung fibrosis (e.g., postradiation fibrosis or idiopathic pulmonary fibrosis).

CLINICAL MANIFESTATIONS

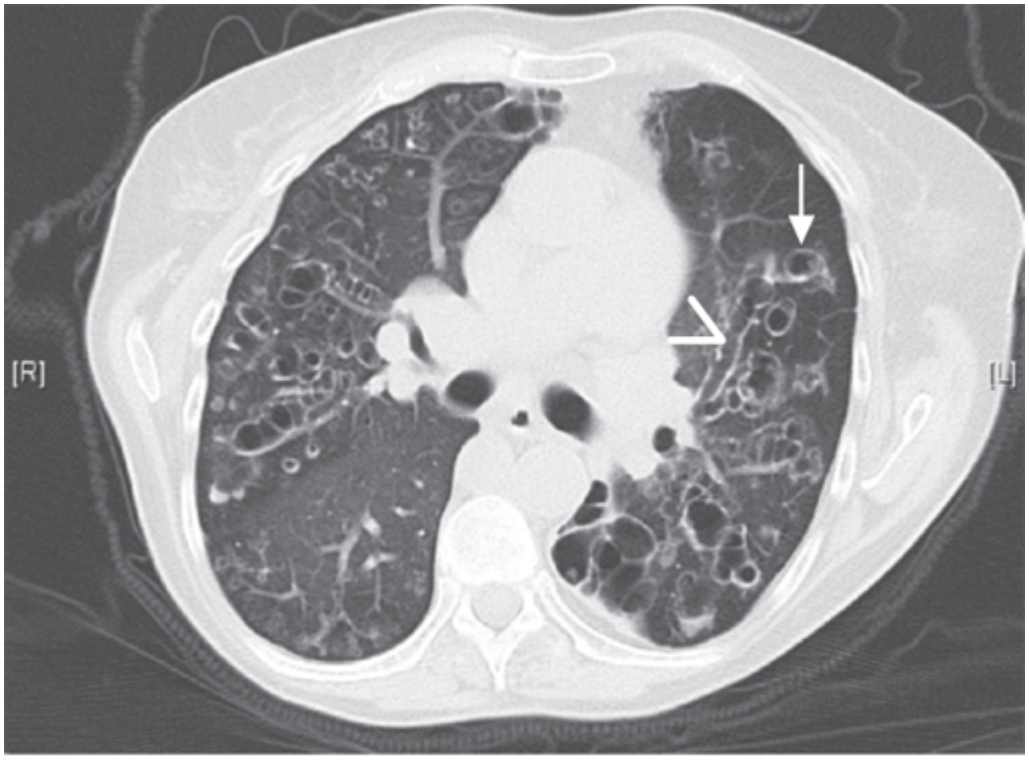
The most common clinical presentation is a persistent productive cough with ongoing production of thick, tenacious sputum. Physical findings often include crackles and wheezing on lung auscultation, and some patients with bronchiectasis exhibit clubbing of the digits. Mild to moderate airflow obstruction is often detected on pulmonary function tests, overlapping with that seen at presentation with other conditions, such as chronic obstructive pulmonary disease (COPD). Acute exacerbations of bronchiectasis are usually characterized by changes in the nature of sputum production, with increased volume and purulence. However, typical signs and symptoms of lung infection, such as fever and new infiltrates, may not be present.

DIAGNOSIS

The diagnosis is usually based on presentation with a persistent chronic cough and sputum production accompanied by consistent radiographic features. Although chest radiographs lack sensitivity, the presence of “tram tracks” indicating dilated airways is consistent with bronchiectasis. Chest CT is more specific for bronchiectasis and is the imaging modality of choice for confirming the diagnosis. CT findings include airway dilation (detected as parallel “tram tracks” or as the “signet-ring sign”—a cross-sectional area of the airway with a diameter at least 1.5 times that of the adjacent vessel), lack of bronchial tapering (including the presence of tubular structures within 1 cm from the pleural surface), bronchial wall thickening in dilated airways, inspissated secretions (e.g., the “tree-in-bud” pattern), or cysts emanating from the bronchial wall (especially pronounced in cystic bronchiectasis) **(Fig. 284-1)**.

FIGURE 284-1

Representative chest CT image of severe bronchiectasis. This patient’s CT demonstrates many severely dilated airways, seen both longitudinally (*arrowhead*) and in cross-section (*arrow*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Bronchiectasis

The evaluation of a patient with bronchiectasis entails elicitation of a clinical history, chest imaging, and a workup to determine the underlying etiology. Evaluation of focal bronchiectasis almost always requires bronchoscopy to exclude airway obstruction by an underlying mass or foreign body. A workup for diffuse bronchiectasis includes analysis for the major etiologies (Table 284-1), with an initial focus on excluding CF. Pulmonary function testing is an important component of a functional assessment of the patient.

TREATMENT

TREATMENT Bronchiectasis

Treatment of infectious bronchiectasis is directed at the control of active infection and improvements in secretion clearance and bronchial hygiene so as to decrease the microbial load within the airways and minimize the risk of repeated infections.

ANTIBIOTIC TREATMENT

Antibiotics targeting the causative or presumptive pathogen (with *Haemophilus influenzae* and *P. aeruginosa* isolated commonly) should be administered in acute exacerbations, usually for a minimum of 7–10 days and perhaps for as long as 14 days. Decisions about treatment of NTM infection can be difficult, given that these organisms can be colonizers as well as pathogens and the prolonged treatment course often is not well tolerated. Consensus guidelines have advised that diagnostic criteria for true clinical infection with NTM should be considered in patients with symptoms and radiographic findings of lung disease who have at least two sputum samples positive on culture; at least one bronchoalveolar lavage (BAL) fluid sample positive on culture; a biopsy sample displaying histopathologic features of NTM infection (e.g., granuloma or a positive stain for acid-fast bacilli) along with one positive sputum culture; or a pleural fluid sample (or a sample from another sterile extrapulmonary site) positive on culture. MAC strains are the most common NTM pathogens, and the recommended regimen for HIV-negative patients infected with macrolide-sensitive MAC includes a macrolide combined with rifampin and ethambutol. Consensus guidelines recommend macrolide susceptibility testing for clinically significant MAC isolates.

BRONCHIAL HYGIENE

The numerous approaches used to enhance secretion clearance in bronchiectasis include hydration and mucolytic administration, aerosolization of bronchodilators and hyperosmolar agents (e.g., hypertonic saline), and chest physiotherapy (e.g., postural drainage, traditional mechanical chest percussion via hand clapping to the chest, or use of devices such as an oscillatory positive expiratory pressure flutter valve or a high-frequency chest wall oscillation vest). Pulmonary rehabilitation and a regular exercise program may assist with secretion clearance as well as with other aspects of bronchiectasis, including improved exercise capacity and quality of life. The mucolytic dornase (DNase) is recommended routinely in CF-related bronchiectasis but not in non-CF bronchiectasis, given concerns about lack of efficacy and potential harm in the non-CF population.

ANTI-INFLAMMATORY THERAPY

It has been proposed that control of the inflammatory response may be of benefit in bronchiectasis, and relatively small-scale trials have yielded evidence of alleviated dyspnea, decreased need for inhaled β-agonists, and reduced sputum production with inhaled glucocorticoids. However, no significant differences in lung function or bronchiectasis exacerbation rates have been observed. Risks of immunosuppression and adrenal suppression must be carefully considered with use of anti-inflammatory therapy in infectious bronchiectasis. Nevertheless, administration of oral/systemic

glucocorticoids may be important in treatment of bronchiectasis due to certain etiologies, such as ABPA, or of noninfectious bronchiectasis due to underlying conditions, especially that in which an autoimmune condition is believed to be active (e.g., rheumatoid arthritis or Sjögren's syndrome). Patients with ABPA may also benefit from a prolonged course of treatment with the oral antifungal agent itraconazole.

REFRACTORY CASES

In select cases, surgery can be considered, with resection of a focal area of suppuration. In advanced cases, lung transplantation can be considered.

COMPLICATIONS

In more severe cases of infectious bronchiectasis, recurrent infections and repeated courses of antibiotics can lead to microbial resistance to antibiotics. In certain cases, combinations of antibiotics that have independent toxicity profiles may be necessary to treat resistant organisms.

Recurrent infections can result in injury to superficial mucosal vessels, with bleeding and, in severe cases, life-threatening hemoptysis. Management of massive hemoptysis usually requires intubation to stabilize the patient, identification of the source of bleeding, and protection of the nonbleeding lung. Control of bleeding often necessitates bronchial artery embolization and, in severe cases, surgery.

PROGNOSIS

Outcomes of bronchiectasis can vary widely with the underlying etiology and comorbid conditions and may also be influenced by the frequency of exacerbations and (in infectious cases) the specific pathogens involved (with worse outcomes associated with *P. aeruginosa* colonization). Increasing attention is being given to defining clinical phenotypes of bronchiectasis in light of clinical, radiographic, and microbial features and to developing screening tools for the assessment of quality of life and disease severity. In one study, the decline of lung function in patients with non-CF bronchiectasis was similar to that in patients with COPD, with the forced expiratory volume in 1 s (FEV_1) declining by 50–55 mL per year as opposed to 20–30 mL per year for healthy controls.

PREVENTION

Reversal of an underlying immunodeficient state (e.g., by administration of gamma globulin for immunoglobulin-deficient patients) and vaccination of patients with chronic respiratory conditions (e.g., influenza and pneumococcal vaccines) can decrease the risk of recurrent infections. Patients who smoke should be counseled about smoking cessation.

After resolution of an acute infection in patients with recurrences (e.g., ≥3 episodes per year), the use of suppressive antibiotics to minimize the microbial load and reduce the frequency of exacerbations has been proposed. Although there is less consensus about this approach in non-CF-associated bronchiectasis than in CF-related bronchiectasis, small studies have supported benefits of selected therapies. Possible suppressive treatments include (1) administration of an oral antibiotic (e.g., ciprofloxacin) daily for 1–2 weeks per month; (2) use of a rotating schedule of oral antibiotics (to minimize the risk of development of drug resistance); (3) administration of a macrolide antibiotic (see below) daily or three times per week (with mechanisms of possible benefit related to non-antimicrobial properties, such as anti-inflammatory effects and reduction of gram-negative bacillary biofilms); (4) inhalation of aerosolized antibiotics (e.g., tobramycin inhalation solution) for select patients on a rotating schedule (e.g., 30 days on, 30 days off), with the goal of decreasing the microbial load without eliciting the side effects of systemic drug administration; and (5) intermittent administration of IV

antibiotics (e.g., “clean-outs”) for patients with more severe bronchiectasis and/or resistant pathogens. In relation to macrolide therapy (point 3 above), a number of double-blind, placebo-controlled, randomized trials have been published in non-CF bronchiectasis and support a benefit of long-term macrolides (6–12 months of azithromycin or erythromycin) in decreasing rates of bronchiectasis exacerbation, mucus production, and decline in lung function. However, two of these studies also reported increased macrolide resistance in commensal pathogens, dampening enthusiasm for universal use of macrolides in this setting and raising the question of whether there might be select non-CF bronchiectasis patients with higher morbidity for whom benefits of long-term macrolides might outweigh the risks of emergence of antibiotic resistance. In particular, development of macrolide-resistant NTM is a potential concern, making treatment of those pathogens much more difficult. Furthermore, patients with different patterns of microbial colonization may not all experience similar benefits with macrolide therapy. Therefore, before chronic macrolide therapy is considered, it is advisable to rule out NTM infection and carefully consider each patient’s scenario closely, obtaining an electrocardiogram to rule out a prolonged QT interval that might place the patient at increased risk of arrhythmias.

In addition, ongoing consistent attention to bronchial hygiene can promote secretion clearance and decrease the microbial load in the airways.

FURTHER READING

Aliberti S et al: Clinical phenotypes in adult patients with bronchiectasis. Eur Respir J 47:1113, 2016. [PubMed: 26846833]
Khoo JK et al: Bronchiectasis in the last five years: New developments. J Clin Med 5:115, 2016.
McGrath EE et al: Guidelines on the diagnosis and treatment of pulmonary non-tuberculous mycobacteria infection. In J Clin Pract 62:1947, 2008.
Rogers GB et al: The effect of long-term macrolide treatment on respiratory microbiota composition in non–cystic fibrosis bronchiectasis: An analysis from the randomized, double-blind, placebo-controlled Bless trial. Lancet Respir Med 2:988, 2014. [PubMed: 25458200]

Chapter 285: Cystic Fibrosis

Eric J. Sorscher

CLINICAL FEATURES

Cystic fibrosis (CF) is an autosomal recessive exocrinopathy affecting multiple epithelial tissues. The gene product responsible for CF (the cystic fibrosis transmembrane conductance regulator [CFTR]) serves as an anion channel in the apical (luminal) plasma membranes of epithelial cells and regulates volume and composition of exocrine secretion. An increasingly sophisticated understanding of CFTR molecular genetics and membrane protein biochemistry has facilitated CF drug discovery, with a number of new agents recently approved or advancing through the clinical testing phase.

Respiratory Manifestations

The major morbidity and mortality associated with CF is attributable to respiratory compromise, characterized by copious hyperviscous and adherent pulmonary secretions that obstruct small and medium-sized airways. CF airway secretions are exceedingly difficult to clear, and a complex bacterial flora that includes *Staphylococcus aureus, Haemophilus influenzae*, and *Pseudomonas aeruginosa* (among other pathogens) is routinely cultured from CF sputum. Microbiome analysis has identified hundreds of other bacterial species in CF lungs, although their relationship to pulmonary failure remains to be determined. Robust pulmonary inflammation in the setting of inspissated mucus and chronic bacterial infection leads to collateral tissue injury and further aggravates respiratory decline. Organisms such as *P. aeruginosa* exhibit a stereotypic mode of pathogenesis; a sentinel and early colonization event often engenders lifelong pulmonary infection by the same genetic strain. Over a period of many years, *P. aeruginosa* evolves in CF lungs to adopt a mucoid phenotype (attributable to release of alginate exoproduct) that confers selective advantage for the pathogen and poor prognosis for the host. Strategies to eradicate *P. aeruginosa* early in the course of disease have been successful and are thought to improve prognosis significantly if sustained.

Pancreatic Findings

The complete name of the disease, *cystic fibrosis of the pancreas*, refers to profound tissue destruction of the exocrine pancreas, with fibrotic scarring and/or fatty replacement, cyst proliferation, loss of acinar tissue, and ablation of normal pancreatic architecture. As in the lung, tenacious exocrine secretions (sometimes termed *concretions*) obstruct pancreatic ducts and impair production and flow of digestive enzymes to the duodenum. The sequelae of exocrine pancreatic insufficiency include chronic malabsorption, poor growth, fat-soluble vitamin insufficiency, high levels of serum immunoreactive trypsinogen (a diagnostic test used in newborn screening), and loss of pancreatic islet cell mass. CF-related diabetes mellitus is a manifestation in over 30% of adults with the disease and is likely multifactorial in nature (attributable to progressive destruction of the endocrine pancreas, insulin resistance due to stress hormones, and additional factors).

Other Organ System Damage

As in CF lung and pancreas, thick and tenacious secretions compromise numerous other exocrine tissues. Obstruction of intrahepatic bile ducts and parenchymal fibrosis are commonly observed in pathologic specimens, with multilobular cirrhosis in 4–15% of patients with CF and significant hepatic insufficiency as a resulting manifestation among adults. Contents of the intestinal lumen are often difficult to excrete, leading to meconium ileus (a presentation in 10–20% of newborns with CF) or distal intestinal obstructive syndrome in older individuals. Men typically exhibit complete involution of the vas deferens and infertility (despite functioning spermatogenesis), and ~99% of males with CF are infertile. The etiology of this dramatic anatomic defect in the male genitourinary system is not understood but may represent a developmental abnormality secondary to improper secretion by the vas or associated structures. Abnormalities of female reproductive tract secretions are likely contributors to an increased incidence of infertility among women with the disease. Radiographic evidence of sinusitis occurs in most CF patients and is associated with pathogens similar to those recovered from lower airways, suggesting that the sinus may serve as a reservoir for bacterial seeding.

PATHOGENESIS

Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)

CFTR is an integral membrane protein that functions as an epithelial anion channel. The ~1480-amino-acid molecule encodes a passive conduit for chloride and bicarbonate transport across plasma membranes of epithelial tissues, with direction of ion flow dependent on the electrochemical driving force. Gating of CFTR involves conformational cycling between an open and closed configuration and is augmented by hydrolysis of adenosine triphosphate (ATP). Anion flux mediated by CFTR does not involve active transport against a concentration gradient but utilizes the energy provided from ATP hydrolysis as a central feature of ion channel mechanochemistry and gating.

CFTR is situated in the apical plasma membranes of acinar and other epithelial cells where it regulates the amount and composition of secretion by exocrine glands. In numerous epithelia, chloride and bicarbonate release is followed passively by the flow of water, allowing for mobilization and clearance of exocrine products. Along respiratory mucosa, CFTR is necessary to provide sufficient depth of the periciliary fluid layer (PCL), allowing normal ciliary extension and mucociliary transport. CFTR-deficient airway cells exhibit depleted PCL, causing ciliary collapse and failure to clear overlying mucus **(Video 285-1)**. In airway submucosal glands, CFTR is highly expressed in acini and may participate both in the formation of mucus and extrusion of glandular secretion onto the airway surface **(Fig. 285-1)**. In other exocrine glands characterized by abrogated mucus transport (e.g., pancreatic acini and ducts, as well as bile canaliculi, intestinal lumen), similar pathogenic mechanisms have been implicated. In these tissues, a driving force for apical chloride and/or bicarbonate secretion is believed to promote CFTR-mediated fluid and electrolyte release into the lumen, which confers proper rheology of mucins and other exocrine products. Failure of this mechanism disrupts normal hydration and transport of glandular secretion and is widely viewed as a proximate cause of obstruction, with concomitant tissue injury.

VIDEO 285-1: Cystic Fibrosis

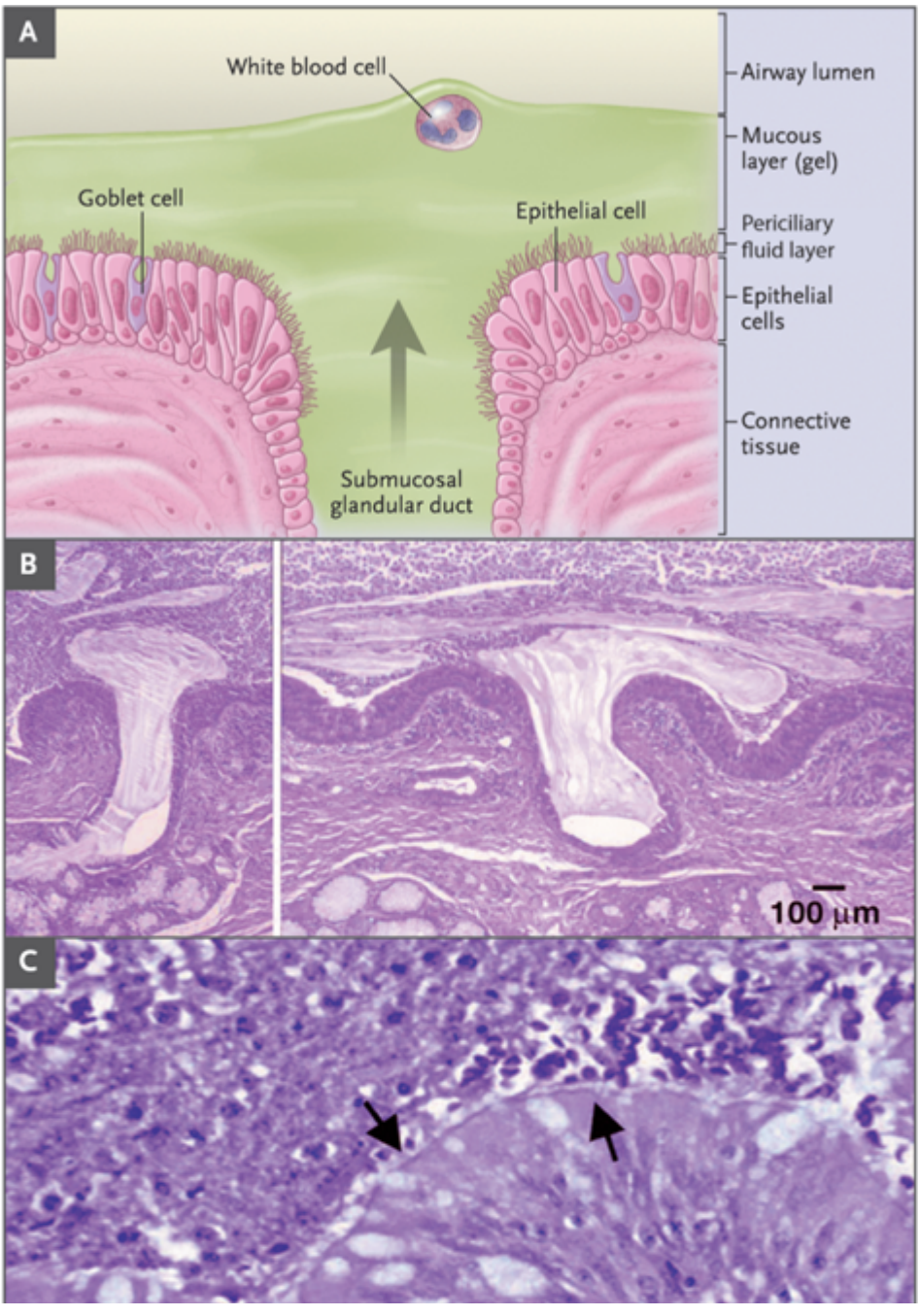
Initial video sequences describe establishment of the normal periciliary fluid layer bathing the surface airway epithelium, with spheres representing chloride and bicarbonate ions secreted through CFTR and across the apical (mucosal) respiratory surface. Later video sequences depict failure of CFTR anion transport and resulting depletion of the periciliary layer, “plastering” of cilia against the mucosal surface, and accumulation of mucus in the airway with resulting bacterial infection. *(Video courtesy of the Cystic Fibrosis Foundation.)*

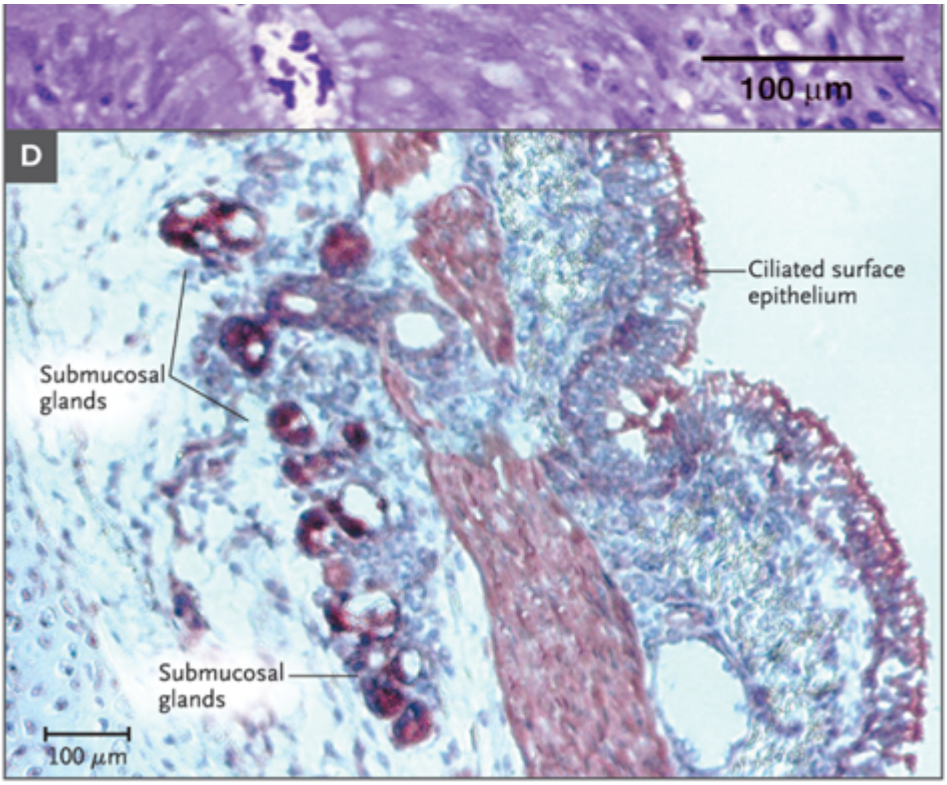
Image not available.

Play Video

FIGURE 285-1

Extrusion of mucus secretion onto the epithelial surface of airways in cystic fibrosis (CF). ***A.*** Schematic of the surface epithelium and supporting glandular structure of the human airway. ***B.*** The submucosal glands of a patient with CF are filled with mucus, and mucopurulent debris overlies the airway surfaces, essentially burying the epithelium. ***C.*** A higher magnification view of a mucus plug tightly adhering to the airway surface, with *arrows* indicating the interface between infected and inflamed secretions and the underlying epithelium to which the secretions adhere. (Both ***B*** and ***C*** were stained with hematoxylin and eosin, with the colors modified to highlight structures.) Infected secretions obstruct airways and, over time, dramatically disrupt the normal architecture of the lung. ***D.*** CFTR is expressed in surface epithelium and serous cells at the base of submucosal glands in a porcine lung sample, as shown by the dark staining, signifying binding by CFTR antibodies to epithelial structures (aminoethylcarbazole detection of horseradish peroxidase with hematoxylin counterstain). *(From SM Rowe, S Miller, EJ Sorscher: N Engl J Med 352:1992, 2005.)*





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pulmonary Inflammation and Remodeling

The CF airway is characterized by an aggressive, unrelenting, neutrophilic inflammatory response with release of proteases and oxidants leading to airway remodeling and bronchiectasis. Intense pulmonary inflammation is largely driven by chronic respiratory infection. Macrophages and other cells resident in CF lungs augment elaboration of proinflammatory cytokines, which contribute to innate and adaptive immune reactivity. CFTR-dependent abnormalities of airway surface fluid composition (e.g., pH) have been reported as contributors to impaired bacterial killing in CF lungs. The role of CFTR as a direct mediator of inflammatory responsiveness and/or pulmonary remodeling represents an important and topical area of investigation.

MOLECULAR GENETICS

DNA sequencing of *CFTR* from patients (and others) worldwide has revealed almost 2000 allelic variants; several hundred of these have been well-characterized as disease-causing mutations. Distinguishing the single nucleotide transversions or other polymorphisms with causal relevance can sometimes present a significant challenge. The CFTR2 resource (*www.cftr2.org/*) delineates gene variants with a clear etiologic role.

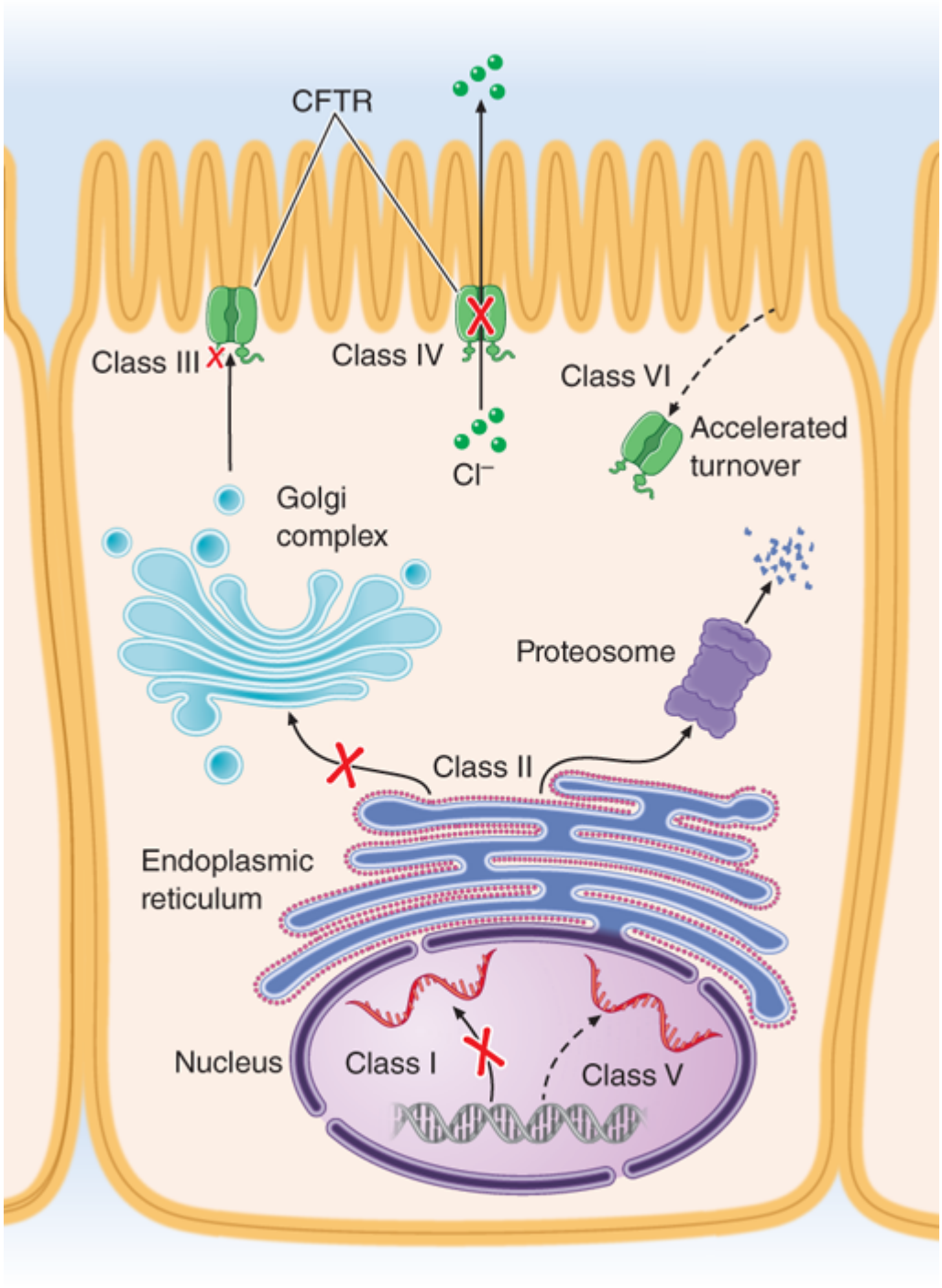
CFTR defects known to elicit disease are often categorized based on molecular mechanism. For example, the common F508del mutation (nomenclature denotes omission of a single phenylalanine residue [F] at CFTR position 508) leads to a folding abnormality recognized by cellular quality control pathways. CFTR encoding F508del retains partial ion channel function, but protein maturation is arrested in the endoplasmic reticulum, and CFTR fails to arrive at the plasma membrane. Instead, F508del CFTR is misrouted and undergoes endoplasmic reticulum–associated degradation via the proteasome. CFTR mutations that disrupt protein maturation are termed class II defects and are by far

the most common genetic abnormalities. F508del alone accounts for ~70% of defective *CFTR* alleles in the United States, where ~90% of individuals with CF carry at least one F508del mutation.

Other gene defects include CFTR ion channels properly trafficked to the apical cell surface but unable to open and/or gate. Such channel proteins include G551D (a glycine to aspartic acid replacement at CFTR position 551), which leads to an inability to transport Cl^- or HCO_3^- in the presence of ATP (a class III abnormality). Individuals with at least one G551D allele represent 4–5% of CF patients in North America. *CFTR* nonsense alleles such as G542X, R553X, and W1282X (premature termination codon replaces glycine, arginine, or tryptophan at positions 542, 553, or 1282, respectively) are among the common class I defects, in addition to large deletions or other major disruptions of the gene. The W1282X mutation, for example, is prevalent among individuals of Ashkenazi descent and is a predominant CF genotype in Israel. Additional categories of *CFTR* mutation include defects in the ion channel pore (class IV), RNA splicing (class V), and increased plasma membrane turnover (class VI) **(Fig. 285-2)**.

FIGURE 285-2

Categories of *CFTR* mutations. Classes of defects in the *CFTR* gene include the absence of synthesis (class I); defective protein maturation and premature degradation (class II); disordered gating/regulation, such as diminished adenosine triphosphate (ATP) binding and hydrolysis (class III); defective conductance through the ion channel pore (class IV); a reduced number of *CFTR* transcripts due to a promoter or splicing abnormality (class V); and accelerated turnover from the cell surface (class VI). *(From SM Rowe, S Miller, EJ Sorscher: N Engl J Med 352:1992, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS

During the past decade, newborn screening has led to most CF diagnoses, with confirmation through *CFTR* mutation analysis and sweat electrolyte measurements as cardinal diagnostic tests. DNA-based evaluation typically surveys numerous disease-associated mutations; panels that identify on the order of 20–140 CFTR variants are available through a variety of public health laboratories and commercial sources. For difficult cases, complete *CFTR* exonic sequencing together with analysis of splice junctions and key regulatory elements can be obtained.

Sweat electrolytes following pilocarpine iontophoresis continue to comprise an essential diagnostic element, with levels of chloride markedly elevated in CF compared to non-CF individuals. The sweat test result is highly specific and served as a mainstay of diagnosis for many decades prior to availability of *CFTR* genotyping. Notably, hyperviscosity of eccrine sweat is not a clinical feature of the disease. Sweat ducts function to reabsorb chloride from a primary sweat secretion produced by the glandular coil. Malfunction of CFTR leads to diminished chloride uptake from the ductular lumen, and sweat emerges on the skin with markedly elevated levels of chloride. For the unusual situation in which both *CFTR* genotype and sweat electrolytes are inconclusive, in vivo measurement of ion transport across the nasal airways can serve as a specific test for CF and is used by a number of referral centers. For example, elevated (sodium-dependent) transepithelial charge separation across airway epithelial tissue and persistent failure of isoproterenol-dependent chloride secretion (via CFTR) represent bioelectric findings specific for the disease. Measurements of CFTR activity in excised rectal mucosal biopsies can also be obtained.

COMPLEXITY OF A CF PHENOTYPE

CF classically presents in childhood with chronic productive cough, malabsorption including steatorrhea, and failure to thrive. The disease is most common among whites (~1 in 3300 live births) and much less frequent among African-American (~1 in 10,000) or Asian populations (~1 in 33,000). Several “severe” defects that impair CFTR activity (including F508del, G551D, and truncation alleles) are predictive of pancreatic insufficiency, which is clinically evident in 80–90% of individuals with CF. These few genotype-phenotype correlations notwithstanding, genotype is, in general, a poor predictor of overall respiratory prognosis.

A spectrum of CFTR-related diseases with features resembling classic CF has been well described. In addition to multiorgan involvement, forme frustes, such as isolated congenital bilateral absence of the vas deferens or pancreatitis (without other organ system findings), are strongly associated with *CFTR* mutations in at least one allele. Although CF is a classic monogenic disease, the importance of non-*CFTR* gene modifiers and proteins that regulate ion flux, inflammatory pathways, and airway remodeling has been increasingly appreciated as influencing clinical course. For example, the magnitude of transepithelial sodium reabsorption in CF airways, which helps control periciliary fluid depth and composition, is strongly influenced by CFTR and represents a molecular target for disease intervention.

THERAPEUTICS DIRECTED TOWARD CF SEQUELAE

Chronic Management

Standard care for outpatients with CF is intensive, with regimens that include exogenous pancreatic enzymes taken with meals, nutritional supplementation, anti-inflammatory medication, bronchodilators, and chronic or periodic administration of oral or aerosolized antibiotics (e.g., as maintenance therapy for patients with *P. aeruginosa*). Recombinant DNAse aerosols (degraded DNA strands that contribute to mucus viscosity) and nebulized hypertonic saline (serves to augment PCL depth, activate mucociliary clearance, and mobilize inspissated airway secretions) are administered routinely. Chest physiotherapy several times each day is a standard means to promote clearance of airway mucus. Among adults with CF, malabsorption, chronic inflammation, and endocrine abnormalities can lead to poor bone mineralization, requiring treatment with vitamin D, calcium, and other measures. The time, complexity, and expense of home care are considerable and take a significant toll on patients and their families. Improved treatments directed toward nutritional deficits, pulmonary inflammation, mucostasis, and other sequelae therefore remain a high priority in the field.

Pulmonary Exacerbation

Severe respiratory exacerbation is commonly managed by hospital admission for frequent chest physiotherapy and parenteral antibiotics directed against serious (and often multiply resistant) bacterial pathogens. Aggressive intervention in this setting can restore a large component of lung function, but ongoing and cumulative loss of pulmonary reserve reflects the natural history of the disease. Poor prognostic indicators such as sputum culture containing *Burkholderia cepacia* complex, mucoid *P. aeruginosa*, or atypical mycobacteria are rigorously monitored in the CF patient population. An increasing incidence of methicillin-resistant *S. aureus* has also been observed, although the clinical significance of this finding has not been fully elucidated. Typical inpatient antibiotic coverage includes combination drug therapy with an aminoglycoside and β-lactam for at least 14 days. Maximal improvement in lung function is often achieved by 8–10 days in this setting. Many families elect parenteral antibiotic treatment at home, and additional studies are needed to evaluate specific drug combinations, duration of therapy, and home versus inpatient management. Other CF respiratory sequelae that may require hospitalization include hemoptysis and pneumothorax. Hypersensitivity to *Aspergillus* (allergic bronchopulmonary aspergillosis) occurs in ~5% of individuals with the disease and should be suspected in the absence of a response to conventional treatment.

Considerations Regarding Lung Transplantation

In the setting of end-stage CF pulmonary failure, lung transplantation is a viable therapeutic option with 5-year survival rates on the order of 60% and median survival >8 years. Determining the optimal timing for surgery presents a substantial challenge, particularly because overall prognosis for individuals with severe lung disease is sometimes difficult to predict, and mortality associated with transplantation can be significant. Forced expiratory volume in 1s (FEV_1) measurements <30% predicted, together with an assortment of other clinical parameters (hospitalization frequency, need for supplemental oxygen, etc.), are often used as thresholds for entry onto transplantation lists, although waiting periods for healthy donor lungs can be quite protracted. Based on clinical outcome and other features, CF patients and their families sometimes do not pursue this option. The decision is best approached through consultation with health care providers specializing in both CF clinical management and transplantation.

CFTR MODULATION

Potentiation of Mutant CFTR Gating

A massive effort directed toward high-throughput drug analysis of large compound libraries (containing millions of individual agents) has identified novel and promising approaches to CF therapy. The approved compound ivacaftor, for example, robustly potentiates CFTR channel opening and stimulates ion transport. Ivacaftor overcomes the G551D CFTR gating defect, and individuals carrying this mutation exhibit pronounced improvement in lung function, weight gain, and other clinical benefit after only a few weeks of oral therapy. Remarkably, sweat chloride values are significantly reduced. Prior to ivacaftor, no clinical intervention of any sort had been shown to normalize the CF sweat abnormality. Partial function CFTR variants (in addition to G551D) for which ivacaftor has recently been approved include numerous other genotypes, with the list of registered indications expected to increase. Chronic administration studies of the drug are ongoing, and indicate significant benefit in terms of respiratory function and other clinical parameters. Ivacaftor has been viewed as the harbinger of a new era for CF therapeutics directed at treating the most fundamental causes of the disease.

Correction of the F508del Processing Abnormality

Advancement of new drugs that address specific CFTR defects in protein folding and maturation has been bolstered by clinical studies of F508del rescue in combination with ivacaftor. Lumacaftor, the first FDA approved “corrector” molecule (as distinct from a CFTR gating “potentiator” such as ivacaftor)

partially overcomes defective F508del CFTR biogenesis, and was discovered through compound library screening. The drug promotes cell surface localization of F508del protein. A dual formulation with ivacaftor confers improvement in pulmonary function among F508del homozygous individuals (~45% of the U.S. CF population). The combination of lumacaftor with ivacaftor has been associated with several important pharmacologic interactions, including those mediated by CYP3A, and diminished activity of oral contraceptives. A chest discomfort syndrome and dyspnea are also well-described.

Personalized Molecular Therapies

The advent of CFTR modulators with robust clinical impact has engendered new optimism regarding care of patients with CF. It is clear that future interventions will be tailored to specific genotypic abnormalities. Drug screening campaigns and other research programs have identified agents capable of suppressing CFTR nonsense alleles, augmenting potentiator activity, and further promoting F508del correction. Efforts to apply emerging compounds in a fashion that will benefit the ~90% of CF subjects carrying at least one copy of F508del (i.e., with F508del or a different *CFTR* mutation on the second allele) comprise an essential priority for the future. Several new molecules in combination with ivacaftor or other CFTR modulators are under evaluation as part of multi-center efficacy trials for this purpose.

Progress in CF drug discovery is emblematic of what might be accomplished in numerous refractory inherited diseases using an approach grounded in molecular mechanism and unbiased compound library screening. Genetic manipulation (CFTR gene transfer, genome editing, etc.) represents an alternative strategy less dependent on the specific defect, and potentially applicable to diverse CF variants. Such an approach in CF will require efficient and safe in vivo delivery strategies, particularly those addressing clinically prominent lung disease.

Challenges to Precision CF Therapeutics

Because hundreds of CF defects are very rare (or even “private,” e.g., reported in only one individual or family), detailed molecular profiling (ion transport behavior, protein folding/biogenesis, response to emerging CFTR modulators, etc.) represents an essential scientific objective. In contrast to the classic and invaluable paradigm of tailored therapeutics (Fig. 285-2, and above), a sizable majority of CF variants do not fall into a single diagnostic category. F508del CFTR, for example, predominately displays abnormalities of maturational processing (class II), but also exhibits more subtle defects related to channel activity/gating (class III) and instability at the plasma membrane (class VI). Comprehensive molecular analysis for hundreds of rare CFTR variants will help guide therapeutics, and may be especially important given the high cost of drugs such as ivacaftor and lumacaftor. Expense of these compounds has often restricted third-party reimbursement to include only the specific genotypes for which FDA approval has been obtained. As a consequence, patient access to potentially efficacious agents (and off-label prescribing) is largely precluded. Moreover, clinical trials intended to expand drug label are difficult based on the small numbers of patients carrying ultra-rare alleles. Similar challenges to drug access have been noted in numerous other settings for which precision medicine has become a therapeutic priority.

CF QUALITY IMPROVEMENT

As a direct result of advances in basic research, new therapies have transformed CF from a disease typically leading to death in early childhood to a condition with frequent survival well into the fourth decade of life. It has also become increasingly clear that carefully specified approaches to patient management can have an impact on overall prognosis. For example, standardization of clinical intervention throughout the United States has led to remarkable benefit among the CF population. Well-defined measures for outpatient care are now established, including thresholds for hospital

admission, antibiotic regimens, nutritional guidelines, periodicity of diagnostic tests, and other clinical parameters. These recommendations have become standard throughout specialized CF care centers and other accredited programs. The initiative has improved endpoints such as weight gain, body mass index, and pulmonary function. Information regarding standardized protocols for CF therapy can be accessed at *www.cff.org/treatments/cfcareguidelines/* or through a number of excellent reviews.

GLOBAL CONSIDERATIONS

 Newborn screening for CF is universal throughout the United States, most of the Canadian provinces, Australia, New Zealand, and much of Europe, and will facilitate early CF intervention. Based on data indicating that early nutritional and other therapies can be beneficial, newborn diagnosis is expected to significantly promote health among those with the disease. Implementation of quality improvement measures and novel therapeutics worldwide has become an increasing imperative. For example, median survival among individuals with CF is ~20 years in much of Latin America (compared to >40 years in the United States). The less favorable prognosis is attributable in part to lack of widespread diagnostic testing (newborn screening, sweat and genetic evaluation) and insufficient access to leading-edge, interdisciplinary CF care. Efforts to apply state-of-the-art management to underdiagnosed and underserved CF patient populations are expected to improve outcomes and mitigate CF health disparities in the future.

FURTHER READING

Farrell PM et al: Diagnosis of cystic fibrosis. J Pediatr 181S:S4, 2017. [PubMed: 28129811]
Huang YJ, LIPuma JJ: The microbiome in cystic fibrosis. Clin Chest Med 37:59, 2016. [PubMed: 26857768]
Morrell MR, Pilewski JM: Lung transplantation for cystic fibrosis. Clin Chest Med 37:127, 2016. [PubMed: 26857774]
Sabusap C et al: Analysis of cystic fibrosis-associated P67L CFTR illustrates barriers to personalized therapeutics for orphan diseases. JCI Insight 1:e386581, 2016.
Sawicki GS et al: Sustained benefit from ivacaftor demonstrated by combining clinical trial and cystic fibrosis patient registry data. Am J Respir Crit Care Med 192:836, 2015. [PubMed: 26132840]
Sosnay PR et al: Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. Nat Genet 45:1160, 2013. [PubMed: 23974870]
Stevens DP, Marshall BC: A decade of healthcare improvement in cystic fibrosis: Lessons for other chronic diseases. BMJ Qual Saf 23:i1, 2014. [PubMed: 24608544]
Stoltz DA et al: Origins of cystic fibrosis lung disease. N Engl J Med 372:351, 2015. [PubMed: 25607428]
Veit G et al: From CFTR biology towards combinatorial pharmacotherapy: Expanded classification of cystic fibrosis mutations. Mol Biol Cell 27:424, 2016. [PubMed: 26823392]
Wainwright CE et al: Lumacaftor-ivacaftor in patients with cystic fibrosis homozygous for Phe508del CFTR. N Engl J Med 373:220, 2015. [PubMed: 25981758]

Chapter 286: Chronic Obstructive Pulmonary Disease

Edwin K. Silverman; James D. Crapo; Barry J. Make

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is defined as a disease state characterized by persistent respiratory symptoms and airflow limitation that is not fully reversible (*http://www.goldcopd.com/*). COPD includes *emphysema*, an anatomically defined condition characterized by destruction of the lung alveoli with air space enlargement; *chronic bronchitis*, a clinically defined condition with chronic cough and phlegm; and *small airway disease*, a condition in which small bronchioles are narrowed and reduced in number. The classic definition of COPD requires the presence of chronic airflow obstruction, determined by spirometry, that usually occurs in the setting of noxious environmental exposures—most commonly cigarette smoking. Emphysema, chronic bronchitis, and small airway disease are present in varying degrees in different COPD patients. Patients with a history of cigarette smoking without chronic airflow obstruction may have chronic bronchitis, emphysema, and dyspnea. Although these patients are not included within the classic definition of COPD, they may have similar disease processes. Respiratory symptoms and other features of COPD can occur in subjects who do not meet a definition of COPD based only on airflow obstruction determined by spirometric thresholds of normality.

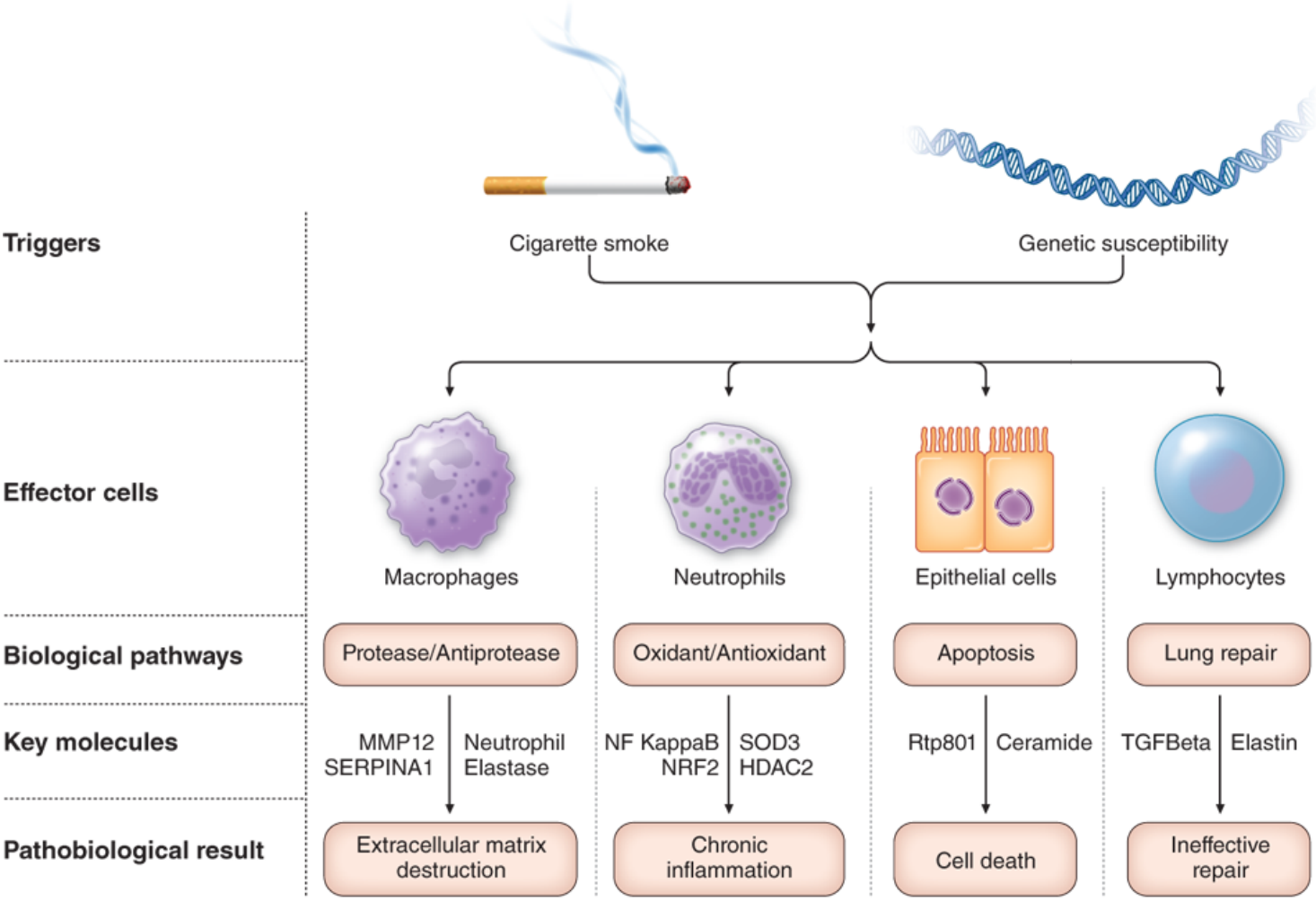
COPD is the third leading cause of death and affects >10 million persons in the United States. COPD is also a disease of increasing public health importance around the world. Estimates suggest that COPD will rise to the third most common cause of death worldwide by 2020.

PATHOGENESIS

Airflow limitation, a major physiologic change in COPD, can result from small airway disease and/or emphysema. Small airways may become narrowed by cells (hyperplasia and accumulation), mucus, and fibrosis, and extensive small airway destruction has been demonstrated to be a hallmark of advanced COPD. Although the precise biological mechanisms leading to COPD have not been determined, a number of key cell types, molecules, and pathways have been identified from cell-based and animal model studies. The pathogenesis of emphysema (shown in **Fig. 286-1**) is more clearly defined than the pathogenesis of small airway disease. Pulmonary vascular destruction occurs in concert with small airway disease and emphysema.

FIGURE 286-1

Pathogenesis of emphysema. Upon long-term exposure to cigarette smoke in genetically susceptible individuals, lung epithelial cells and T and B lymphocytes recruit inflammatory cells to the lung. Biological pathways of protease-antiprotease imbalance, oxidant/antioxidant imbalance, apoptosis, and lung repair lead to extracellular matrix destruction, cell death, chronic inflammation, and ineffective repair. Although most of these biological pathways influence multiple pathobiological results, only a single relationship between pathways and results is shown. A subset of key molecules related to these biological pathways is listed.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The dominant current paradigm for the pathogenesis of emphysema comprises a series of four interrelated events: (1) Chronic exposure to cigarette smoke in genetically susceptible individuals triggers inflammatory and immune cell recruitment within large and small airways and in the terminal air spaces of the lung. (2) Inflammatory cells release proteinases that damage the extracellular matrix supporting airways, vasculature, and gas exchange surfaces of the lung. (3) Structural cell death occurs through oxidant-induced damage, cellular senescence, and proteolytic loss of cellular-matrix attachments leading to extensive loss of smaller airways, vascular pruning, and alveolar destruction. (4) Disordered repair of elastin and other extracellular matrix components contributes to air space enlargement and emphysema.

INFLAMMATION AND EXTRACELLULAR MATRIX PROTEOLYSIS

Elastin, the principal component of elastic fibers, is a highly stable component of the extracellular matrix that is critical to the integrity of the lung. The elastase:antielastase hypothesis, proposed in the mid-1960s, postulated that the balance of elastin-degrading enzymes and their inhibitors determines the susceptibility of the lung to destruction resulting in air space enlargement. This hypothesis was based on the clinical observation that patients with genetic deficiency in α_1 antitrypsin (α_1AT), the inhibitor of the serine proteinase neutrophil elastase, were at increased risk of emphysema, and that instillation of elastases, including neutrophil elastase, into experimental animals, results in emphysema. The elastase:antielastase hypothesis remains a prevailing mechanism for the development of emphysema. However, a complex network of immune and inflammatory cells and additional proteinases that contribute to emphysema has subsequently been identified. Upon exposure to oxidants from cigarette smoke, lung macrophages and epithelial cells become activated, producing proteinases and chemokines that attract other inflammatory and immune cells. Oxidative stress is a key component of COPD pathobiology; the transcription factor NRF2, a major regulator of oxidant-antioxidant balance, and SOD3, a potent antioxidant, have been implicated in emphysema pathogenesis by animal models. Mitochondrial dysfunction in COPD may worsen oxidative stress. One mechanism of macrophage activation occurs via oxidant-induced inactivation of histone deacetylase-2 (HDAC2), shifting the balance toward acetylated or loose chromatin, exposing nuclear factor-kappaB sites, and resulting in transcription of matrix metalloproteinases and proinflammatory cytokines such as interleukin 8 (IL-8) and tumor necrosis factor α (TNF-α); this leads to neutrophil recruitment. CD8+ T cells are also recruited in response to cigarette smoke and release interferon-inducible protein-10 (IP-10, CXCL-7), which in turn leads to macrophage production of macrophage elastase (matrix metalloproteinase-12 [MMP-12]).

Matrix metalloproteinases and serine proteinases, most notably neutrophil elastase, work together by degrading the inhibitor of the other, leading to lung destruction. Proteolytic cleavage products of elastin serve as a macrophage chemokine, and proline-glycine-proline (generated by proteolytic cleavage of collagen) is a neutrophil chemokine—fueling this destructive positive feedback loop. Elastin degradation and disordered repair are thought to be primary mechanisms in the development of emphysema.

There is some evidence that autoimmune mechanisms may promote the progression of disease. Increased B cells and lymphoid follicles are present around the airways of COPD patients, particularly those with advanced disease. Antibodies have been found against elastin fragments as well; IgG autoantibodies with avidity for pulmonary epithelium and the potential to mediate cytotoxicity have been detected.

Concomitant cigarette smoke–induced loss of cilia in the airway epithelium and impaired macrophage phagocytosis predispose to bacterial infection with neutrophilia. In end-stage lung disease, long after smoking cessation, there remains an exuberant inflammatory response, suggesting that cigarette smoke–induced inflammation both initiates the disease and, in susceptible individuals, establishes a chronic process that can continue disease progression even after smoking cessation.

Cell Death

Cigarette smoke oxidant-mediated structural cell death occurs via a variety of mechanisms including excessive ceramide production and Rtp801 inhibition of mammalian target of rapamycin (mTOR), leading to cell death as well as inflammation and proteolysis. Involvement of mTOR and other senescence markers has led to the concept that emphysema resembles premature aging of the lung. Heterozygous gene-targeting of one of the leading genetic determinants of COPD identified by genome-wide association studies (GWAS), hedgehog interacting protein (*HHIP*), in a murine model leads to aging-related emphysema.

Ineffective Repair

The ability of the adult lung to replace lost smaller airways and microvasculature and to repair damaged alveoli appears limited. Uptake of apoptotic cells by macrophages normally results in production of growth factors and dampens inflammation, promoting lung repair. Cigarette smoke impairs macrophage uptake of apoptotic cells, limiting repair. It is unlikely that the intricate and dynamic process of septation that is responsible for alveologenesis during lung development can be reinitiated in the adult human lung.

PATHOLOGY

Cigarette smoke exposure may affect the large airways, small airways (≤2 mm diameter), and alveoli. Changes in large airways cause cough and sputum production, while changes in small airways and alveoli are responsible for physiologic alterations. Airway inflammation, destruction, and the development of emphysema are present in most persons with COPD; however, they appear to be relatively independent processes, and their relative contributions to obstruction vary from one person to another. The early stages of COPD, based on the severity of airflow obstruction **(Table 286-1)**, appear to be primarily associated with medium and small airway disease with the majority of Global Initiative for Chronic Obstructive Lung Disease (GOLD) 1 and GOLD 2 subjects demonstrating little or no emphysema. The early development of chronic airflow obstruction is driven by small airway disease. Advanced stages of COPD (GOLD 3 and 4) are typically characterized by extensive emphysema, although there are a small number of subjects with very severe (GOLD 4) obstruction with virtually no emphysema. The subjects at greatest risk of progression in COPD are those with both aggressive airway disease and emphysema. Thus, finding emphysema (by chest CT) either early or late in the disease process suggests enhanced risk for disease progression.

LARGE AIRWAYS

Cigarette smoking often results in mucus gland enlargement and goblet cell hyperplasia, leading to cough and mucus production that define chronic bronchitis, but these abnormalities are not related to airflow limitation. In response to cigarette smoking, goblet cells not only increase in number but in extent through the bronchial tree. Bronchi also undergo squamous metaplasia, predisposing to carcinogenesis and disrupting mucociliary clearance. Although not as prominent as in asthma, patients may have smooth-muscle hypertrophy and bronchial hyperreactivity leading to airflow limitation. Neutrophil influx has been associated with purulent sputum during respiratory tract infections. Independent of its proteolytic activity, neutrophil elastase is among the most potent secretagogues identified.

SMALL AIRWAYS

The major site of increased resistance in most individuals with COPD is in airways ≤2 mm diameter. Characteristic cellular changes include goblet cell metaplasia, with these mucus-secreting cells replacing surfactant-secreting Club cells. Smooth-muscle hypertrophy may also be present. Luminal narrowing can occur by fibrosis, excess mucus, edema, and cellular infiltration. Reduced surfactant may increase surface tension at the air-tissue interface, predisposing to airway narrowing or collapse. Respiratory bronchiolitis with mononuclear inflammatory cells collecting in distal airway tissues may cause proteolytic destruction of elastic fibers in the respiratory bronchioles and alveolar ducts where the fibers are concentrated as rings around alveolar entrances. Narrowing and drop-out of small airways precede the onset of emphysematous destruction. Advanced COPD has been shown to be associated with a loss of many of the smaller airways and a similar significant loss of the lung microvasculature.

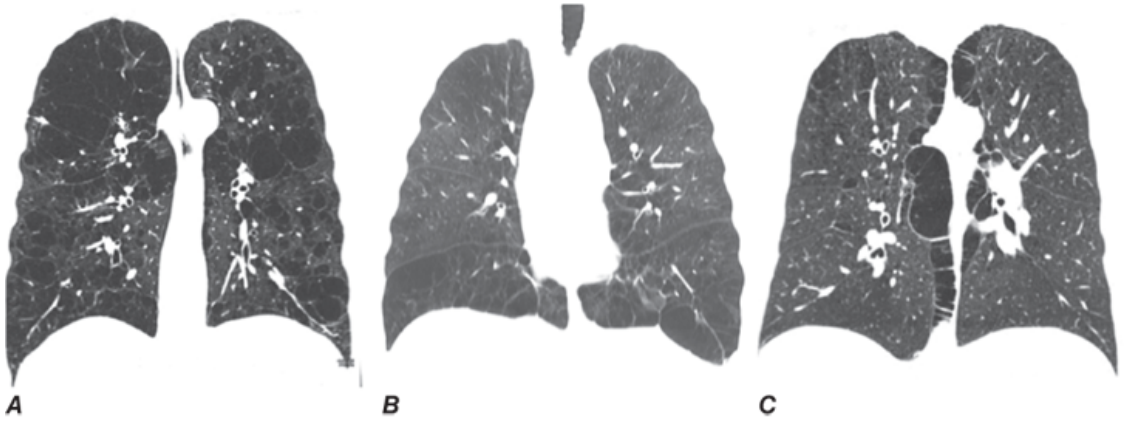
LUNG PARENCHYMA

Emphysema is characterized by destruction of gas-exchanging air spaces, i.e., the respiratory bronchioles, alveolar ducts, and alveoli. Their walls become perforated and later obliterated with coalescence of the delicate alveolar structure into large emphysematous air spaces. Large numbers of macrophages accumulate in respiratory bronchioles of essentially all smokers. Bronchoalveolar lavage fluid from such individuals contains roughly five times as many macrophages as lavage from nonsmokers. Neutrophils and T lymphocytes, particularly CD8+ cells, are also increased in the alveolar space of smokers.

Emphysema is classified into distinct pathologic types, which include centrilobular, panlobular, and paraseptal **(Fig 286-2)**. *Centrilobular emphysema*, the type most frequently associated with cigarette smoking, is characterized by enlarged air spaces found (initially) in association with respiratory bronchioles. Centrilobular emphysema is usually most prominent in the upper lobes and superior segments of lower lobes and is often quite focal. *Panlobular emphysema* refers to abnormally large air spaces evenly distributed within and across acinar units. Panlobular emphysema is commonly observed in patients with α_1AT deficiency, which has a predilection for the lower lobes. Paraseptal emphysema occurs in 10–15% of cases and is distributed along the pleural margins with relative sparing of the lung core or central regions. It is commonly associated with significant airway inflammation and with centrilobular emphysema.

FIGURE 286-2

CT patterns of emphysema. *A.* Centrilobular emphysema with severe upper lobe involvement in a 68-year-old man with a 70 pack-year smoking history but forced expiratory volume (FEV_1) 81% predicted (GOLD spirometry grade 1); ***B.*** Panlobular emphysema with diffuse loss of lung parenchymal detail predominantly in the lower lobes in a 64-year-old man with severe α_1AT deficiency; and ***C***. Paraseptal emphysema with marked airway inflammation in a 52-year-old woman with a 37 pack-year smoking history and FEV_1 40% predicted.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOPHYSIOLOGY

Persistent reduction in forced expiratory flow rates is the most typical finding in COPD. Increases in the residual volume and the residual volume/total lung capacity ratio, non-uniform distribution of ventilation, and ventilation-perfusion mismatching also occur.

AIRFLOW OBSTRUCTION

Airflow limitation, also known as airflow obstruction, is typically determined for clinical purposes by spirometry, which involves forced expiratory maneuvers after the subject has inhaled to total lung capacity. Key parameters obtained from spirometry include the volume of air exhaled within the first second of the forced expiratory maneuver (FEV_1) and the total volume of air exhaled during the entire spirometric maneuver (forced vital capacity [FVC]). Patients with airflow obstruction related to COPD have a chronically reduced ratio of FEV_1/FVC. In contrast to asthma, the reduced FEV_1 in COPD seldom shows large responses to inhaled bronchodilators, although improvements up to 15% are common.

HYPERINFLATION

Lung volumes are also routinely assessed in pulmonary function testing. In COPD there is often “air trapping” (increased residual volume and increased ratio of residual volume to total lung capacity) and progressive hyperinflation (increased total lung capacity) late in the disease. Hyperinflation of the thorax during tidal breathing preserves maximum expiratory airflow, because as lung volume increases, elastic recoil pressure increases, and airways enlarge so that airway resistance decreases.

Despite compensating for airway obstruction, hyperinflation can push the diaphragm into a flattened position with a number of adverse effects. First, by decreasing the zone of apposition between the diaphragm and the abdominal wall, positive abdominal pressure during inspiration is not applied as effectively to the chest wall, hindering rib cage movement and impairing inspiration. Second, because the muscle fibers of the flattened diaphragm are shorter than those of a more normally curved diaphragm, they are

less capable of generating inspiratory pressures than normal. Third, the flattened diaphragm must generate greater tension to develop the transpulmonary pressure required to produce tidal breathing. Fourth, the thoracic cage is distended beyond its normal resting volume and during tidal breathing the inspiratory muscles must do work to overcome the resistance of the thoracic cage to further inflation instead of gaining the normal assistance from the chest wall recoiling outward toward its resting volume.

GAS EXCHANGE

Although there is considerable variability in the relationships between the FEV_1 and other physiologic abnormalities in COPD, certain generalizations may be made. The partial pressure of oxygen in arterial blood Pa_{O_2} usually remains near normal until the FEV_1 is decreased to ~50% of predicted, and even much lower FEV_1 values can be associated with a normal Pa_{O_2}, at least at rest. An elevation of arterial level of carbon dioxide (Pa_{CO_2}) is not expected until the FEV_1 is <25% of predicted and even then may not occur. Pulmonary hypertension severe enough to cause cor pulmonale and right ventricular failure due to COPD typically occurs in individuals who have marked decreases in FEV_1 (<25% of predicted) and chronic hypoxemia (Pa_{O_2} <55 mmHg); however, recent evidence suggests that some patients will develop significant pulmonary hypertension independent of COPD severity **(Chap. 277)**.

Non-uniform ventilation and ventilation-perfusion mismatching are characteristic of COPD, reflecting the heterogeneous nature of the disease process within the airways and lung parenchyma. Physiologic studies are consistent with multiple parenchymal compartments having different rates of ventilation due to regional differences in compliance and airway resistance. Ventilation-perfusion mismatching accounts for essentially all of the reduction in Pa_{O_2} that occurs in COPD; shunting is minimal. This finding explains the effectiveness of modest elevations of inspired oxygen in treating hypoxemia due to COPD and therefore the need to consider problems other than COPD when hypoxemia is difficult to correct with modest levels of supplemental oxygen.

RISK FACTORS

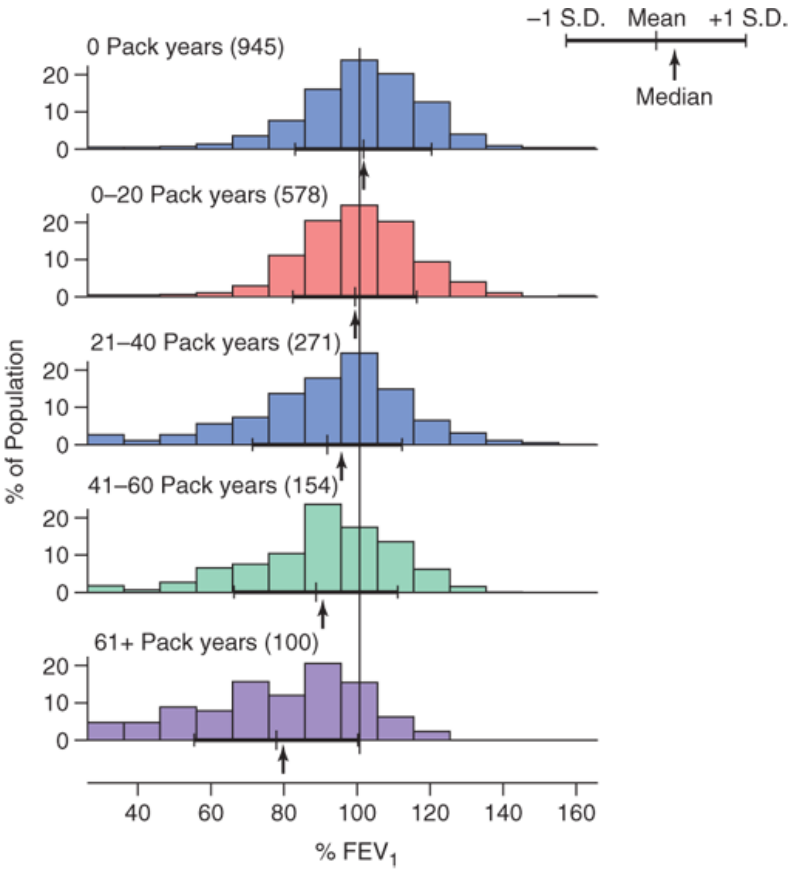
CIGARETTE SMOKING

By 1964, the Advisory Committee to the Surgeon General of the United States had concluded that cigarette smoking was a major risk factor for mortality from chronic bronchitis and emphysema. Subsequent longitudinal studies have shown accelerated decline in FEV_1 in a dose-response relationship to the intensity of cigarette smoking, which is typically expressed as pack-years (average number of packs of cigarettes smoked per day multiplied by the total number of years of smoking). This dose-response relationship between reduced pulmonary function and cigarette smoking intensity accounts, at least in part, for the higher prevalence rates of COPD with increasing age. The historically higher rate of smoking among males is the likely explanation for the higher prevalence of COPD among males; however, the prevalence of COPD among females is increasing as the gender gap in smoking rates has diminished in the past 50 years.

Although the causal relationship between cigarette smoking and the development of COPD has been absolutely proved, there is considerable variability in the response to smoking. Pack-years of cigarette smoking is the most highly significant predictor of FEV_1 **(Fig. 286-3)**, but only 15% of the variability in FEV_1 is explained by pack-years. This finding suggests that additional environmental and/or genetic factors contribute to the impact of smoking on the development of chronic airflow obstruction. Nonetheless, many patients with a history of cigarette smoking with normal spirometry have evidence for worse health-related quality of life, reduced exercise capacity, and emphysema and/or airway disease on chest CT evaluation; thus, they have not escaped the harmful effects of cigarette smoking. While they do not meet the classic definition of COPD based on population normals for FEV_1 and FEV_1/FVC, studies have shown that these subjects overall have a shift toward lower FEV_1 values, which is consistent with obstruction on an individual level.

FIGURE 286-3

Distributions of forced expiratory volume in 1 s (FEV_1) values in a general population sample, stratified by pack-years of smoking. Means, medians, and ±1 standard deviation of percent predicted FEV_1 are shown for each smoking group. Although a dose-response relationship between smoking intensity and FEV_1 was found, marked variability in pulmonary function was observed among subjects with similar smoking histories. *(From B Burrows et al: Am Rev Respir Dis 115:95, 1977; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although cigar and pipe smoking may also be associated with the development of COPD, the evidence supporting such associations is less compelling, likely related to the lower dose of inhaled tobacco by-products during cigar and pipe smoking. The impact of electronic cigarettes (e-cigarettes) on the development and progression of COPD has not yet been determined.

AIRWAY RESPONSIVENESS AND COPD

A tendency for increased bronchoconstriction in response to a variety of exogenous stimuli, including methacholine and histamine, is one of the defining features of asthma **(Chap. 281)**. However, many patients with COPD also share this feature of airway hyperresponsiveness. In older subjects, there is considerable overlap between persons with a history of chronic asthma and smokers with COPD in terms of airway responsiveness, airflow obstruction, and pulmonary symptoms. The origin of asthma is viewed as an allergic disease while COPD is thought to primarily result from smoking-related inflammation and damage; however, they likely share common environmental and genetic factors and the chronic form in older subjects can present similarly. This is particularly true for childhood asthmatic subjects who become chronic smokers.

Longitudinal studies that compared airway responsiveness at the beginning of the study to subsequent decline in pulmonary function have demonstrated that increased airway responsiveness is clearly a significant predictor of subsequent decline in pulmonary function. A recent study from the Childhood Asthma Management Program identified four lung function trajectories in children with persistent asthma. Asthmatics with reduced lung function early in life were more likely to meet spirometric criteria for COPD in early adulthood. Patients with features of both asthma and COPD have been described as the asthma-COPD overlap syndrome. Both asthma and airway hyperresponsiveness are risk factors for COPD.

RESPIRATORY INFECTIONS

The impact of adult respiratory infections on decline in pulmonary function is controversial, but significant long-term reductions in pulmonary function are not typically seen following an individual episode of acute bronchitis or pneumonia. However, respiratory infections are important causes of COPD exacerbations, and recent results from the COPDGene and ECLIPSE studies suggest that COPD exacerbations are associated with increased loss of lung function longitudinally, particularly among those individuals with better baseline lung function levels. The impact of the effects of childhood respiratory illnesses on the subsequent development of COPD has been difficult to assess due to a lack of adequate longitudinal data, but recent studies have suggested that childhood pneumonia may lead to increased risk for COPD later in life.

OCCUPATIONAL EXPOSURES

Increased respiratory symptoms and airflow obstruction have been suggested to result from exposure to dust and fumes at work. Several specific occupational exposures, including coal mining, gold mining, and cotton textile dust, have been implicated as risk factors for chronic airflow obstruction. Although nonsmokers in these occupations can develop some reductions in FEV_1, the importance of dust exposure as a risk factor for COPD, independent of cigarette smoking, is not certain for most of these exposures. However, among coal miners, coal mine dust exposure was a significant risk factor for emphysema in both smokers and nonsmokers. In most cases, the magnitude of these occupational exposures on COPD risk is likely substantially less important than the effect of cigarette smoking.

AMBIENT AIR POLLUTION

Some investigators have reported increased respiratory symptoms in those living in urban compared to rural areas, which may relate to increased pollution in the urban settings. However, the relationship of air pollution to chronic airflow obstruction remains unproved. Prolonged exposure to smoke produced by biomass combustion—a common mode of cooking in some countries—also appears to be a significant risk factor for COPD among women in those countries. However, in most populations, ambient air pollution is a much less important risk factor for COPD than cigarette smoking.

PASSIVE, OR SECOND-HAND, SMOKING EXPOSURE

Exposure of children to maternal smoking results in significantly reduced lung growth. In utero, tobacco smoke exposure also contributes to significant reductions in postnatal pulmonary function. Although passive smoke exposure has been associated with reductions in pulmonary function, the importance of this risk factor in the development of the severe pulmonary function reductions often observed in COPD remains uncertain.

GENETIC CONSIDERATIONS



Although cigarette smoking is the major environmental risk factor for the development of COPD, the development of airflow obstruction in smokers is highly variable. Severe α_1AT deficiency is a proven genetic risk factor for COPD; there is increasing evidence that other genetic determinants also exist.

α_1 Antitrypsin Deficiency

Many variants of the protease inhibitor (PI or *SERPINA1*) locus that encodes α_1AT have been described. The common M allele is associated with normal α_1AT levels. The S allele, associated with slightly reduced α_1AT levels, and the Z allele, associated with markedly reduced α_1AT levels, also occur with frequencies of >1% in most white populations. Rare individuals inherit null alleles, which lead to the absence of any α_1AT production through a heterogeneous collection of mutations. Individuals with two Z alleles or one Z and one null allele are referred to as Pi^Z, which is the most common form of severe α_1AT deficiency.

Although only ~1% of COPD patients are found to have severe α_1AT deficiency as a contributing cause of COPD, these patients demonstrate that genetic factors can have a profound influence on the susceptibility for developing COPD. Pi^Z individuals often develop early-onset COPD, but the ascertainment bias in the published series of Pi^Z individuals—which have usually included many Pi^Z subjects who were tested for α_1AT deficiency because they had COPD—means that the fraction of Pi^Z individuals who will develop COPD and the age-of-onset distribution for the development of COPD in Pi^Z subjects remain unknown. Approximately 1 in 3000 individuals in the United States inherits severe α_1AT deficiency, but only a small minority of these individuals has been identified. The clinical laboratory test used most frequently to screen for α_1AT deficiency is measurement of the immunologic level of α_1AT in serum (see “Laboratory Findings”).

A significant percentage of the variability in pulmonary function among Pi^Z individuals is explained by cigarette smoking; cigarette smokers with severe α_1AT deficiency are more likely to develop COPD at early ages. However, the development of COPD in Pi^Z subjects, even among current or ex-smokers, is not absolute. Among Pi^Z nonsmokers, impressive variability has been noted in the development of airflow obstruction. Asthma and male gender also appear to increase the risk of COPD in Pi^Z subjects. Other genetic and/or environmental factors likely contribute to this variability.

Specific treatment in the form of α_1AT augmentation therapy is available for severe α_1AT deficiency as a weekly IV infusion (see “Treatment,” below).

The risk of lung disease in heterozygous Pi^{MZ} individuals, who have intermediate serum levels of α_1AT (~60% of Pi^{MM} levels), has been controversial. Several recent large studies have demonstrated that Pi^{MZ} subjects who smoke are likely at increased risk for the development of COPD. However, alpha-1 antitrypsin augmentation therapy is not recommended for use in Pi^{MZ} subjects.

Other Genetic Risk Factors

Studies of pulmonary function measurements performed in general population samples have suggested that genetic factors other than PI type influence variation in pulmonary function. Familial aggregation of airflow obstruction within families of COPD patients has also been demonstrated.

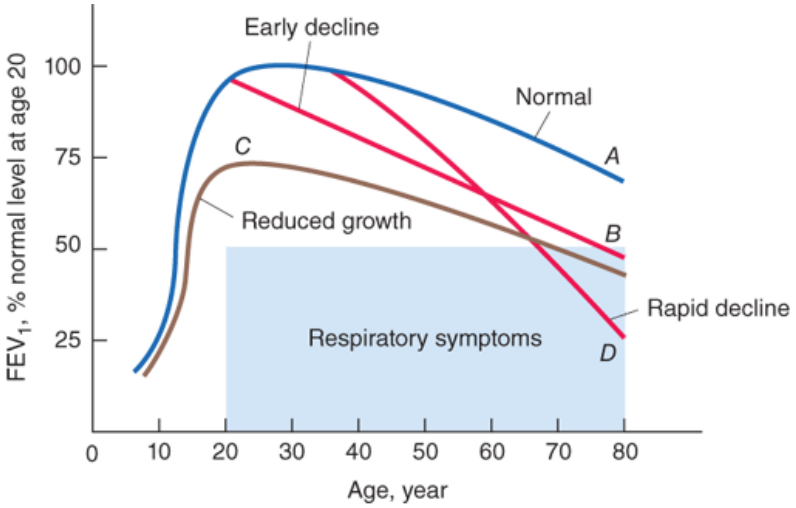
GWAS have identified >20 regions of the genome that contain COPD susceptibility loci, including a region near the *HHIP* gene on chromosome 4, a cluster of genes on chromosome 15 (including components of the nicotinic acetylcholine receptor and another gene, *IREB2*, related to mitochondrial iron regulation), and a region within a gene of unknown function (*FAM13A*). As with most other complex diseases, the risk associated with individual GWAS loci is modest, but these genetic determinants may identify important biological pathways related to COPD. Gene-targeted murine models for *HHIP*, *FAM13A*, and *IREB2* exposed to chronic cigarette smoke had altered emphysema susceptibility, suggesting that those genes are likely to be involved in COPD pathogenesis. A regulatory single nucleotide polymorphisms (SNP) upstream from the *HHIP* gene has been identified as one potential functional variant; the specific genetic determinants in the other COPD GWAS genomic regions have yet to be definitively identified.

NATURAL HISTORY

The effects of cigarette smoking on pulmonary function appear to depend on the intensity of smoking exposure, the timing of smoking exposure during growth, and the baseline lung function of the individual; other environmental factors may have similar effects. Most individuals follow a steady trajectory of increasing pulmonary function with growth during childhood and adolescence, followed by a plateau in early adulthood, and then gradual decline with aging. Individuals appear to track in their quantile of pulmonary function based on environmental and genetic factors that put them on different tracks. The risk of eventual mortality from COPD is closely associated with reduced levels of FEV_1. A graphic depiction of the natural history of COPD is shown as a function of the influences on tracking curves of FEV_1 in **Fig. 286-4**. Death or disability from COPD can result from a normal rate of decline after a reduced growth phase (curve *C*), an early initiation of pulmonary function decline after normal growth (curve *B*), or an accelerated decline after normal growth (curve *D*). Although accelerated rates of lung function decline have classically been associated with COPD, recent analyses of several population-based cohorts demonstrated that many subjects meeting the spirometric criteria for COPD had reduced growth but normal rates of lung function decline. The rate of decline in pulmonary function can be modified by changing environmental exposures (i.e., quitting smoking), with smoking cessation at an earlier age providing a more beneficial effect than smoking cessation after marked reductions in pulmonary function have already developed. The absolute annual loss in FEV_1 tends to be highest in mild COPD and lowest in very severe COPD. Multiple genetic factors influence the level of pulmonary function achieved during growth; genetic determinants likely also influence the rate of decline in response to smoking and potentially to other environmental factors as well.

FIGURE 286-4

Hypothetical tracking curves of forced expiratory volume in 1 s (FEV_1) for individuals throughout their life spans. The normal pattern of growth and decline with age is shown by curve *A*. Significantly reduced FEV_1 (<65% of predicted value at age 20) can develop from a normal rate of decline after a reduced pulmonary function growth phase (curve *C*), early initiation of pulmonary function decline after normal growth (curve *B*), or accelerated decline after normal growth (curve *D*). *(From B Rijcken: Doctoral dissertation, p 133, University of Groningen, 1991; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CLINICAL PRESENTATION

HISTORY

The three most common symptoms in COPD are cough, sputum production, and exertional dyspnea. Many patients have such symptoms for months or years before seeking medical attention. Although the development of airflow obstruction is a gradual process, many patients date the onset of their disease to an acute illness or exacerbation. A careful history, however, usually reveals the presence of symptoms prior to the acute exacerbation. The development of exertional dyspnea, often described as increased effort to breathe, heaviness, air hunger, or gasping, can be insidious. It is best elicited by a careful history focused on typical physical activities and how the patient's ability to perform them has changed. Activities involving significant arm work, particularly at or above shoulder level, are particularly difficult for many patients with COPD. Conversely, activities that allow the patient to brace the arms and use accessory muscles of respiration are better tolerated. Examples of such activities include pushing a shopping cart or walking on a treadmill. As COPD advances, the principal feature is worsening dyspnea on exertion with increasing intrusion on the ability to perform vocational or avocational activities. In the most advanced stages, patients are breathless doing simple activities of daily living.

Accompanying worsening airflow obstruction is an increased frequency of exacerbations (described below). Patients may also develop resting hypoxemia and require institution of supplemental oxygen.

PHYSICAL FINDINGS

In the early stages of COPD, patients usually have an entirely normal physical examination. Current smokers may have signs of active smoking, including an odor of smoke or nicotine staining of fingernails. In patients with more severe disease, the physical examination of the lungs is notable for a prolonged expiratory phase and may include expiratory wheezing. In addition, signs of hyperinflation include a barrel chest and enlarged lung volumes with poor diaphragmatic excursion as assessed by percussion. Patients with severe airflow obstruction may also exhibit use of accessory muscles of respiration, sitting in the characteristic "tripod" position to facilitate the actions of the sternocleidomastoid, scalene, and intercostal muscles. Patients may develop cyanosis, visible in the lips and nail beds.

Although traditional teaching is that patients with predominant emphysema, termed "pink puffers," are thin and noncyanotic at rest and have prominent use of accessory muscles, and patients with chronic bronchitis are more likely to be heavy and cyanotic ("blue bloaters"), current evidence demonstrates that most patients have elements of both chronic bronchitis and emphysema and that the physical examination does not reliably differentiate the two entities.

Advanced disease may be accompanied by cachexia, with significant weight loss, bitemporal wasting, and diffuse loss of subcutaneous adipose tissue. This syndrome has been associated with both inadequate oral intake and elevated levels of inflammatory cytokines (TNF-α). Such wasting is an independent poor prognostic factor in COPD. Some patients with advanced disease have paradoxical inward movement of the rib cage with inspiration (Hoover's sign), the result of alteration of the vector of diaphragmatic contraction on the rib cage as a result of chronic hyperinflation.

Signs of overt right heart failure, termed *cor pulmonale*, are relatively infrequent since the advent of supplemental oxygen therapy.

Clubbing of the digits is not a sign of COPD, and its presence should alert the clinician to initiate an investigation for causes of clubbing. In this population, the development of lung cancer is the most likely explanation for newly developed clubbing.

LABORATORY FINDINGS

The hallmark of COPD is airflow obstruction (discussed above). Pulmonary function testing shows airflow obstruction with a reduction in FEV_1 and FEV_1/FVC **(Chap. 279)**. With worsening disease severity, lung volumes may increase, resulting in an increase in total lung capacity, functional residual capacity, and residual volume. In patients with emphysema, the diffusing capacity may be reduced, reflecting the lung parenchymal destruction characteristic of the disease. The degree of airflow obstruction is an important prognostic factor in COPD and is the basis for the GOLD spirometric severity classification (Table 286-1). Although the degree of airflow obstruction *generally* correlates with the presence and severity of respiratory symptoms, exacerbations, emphysema, and hypoxemia, the correlations are far from perfect. Thus, clinical features should be carefully assessed in each individual patient with COPD to determine the most appropriate therapies. It has been shown that a multifactorial index (BODE) incorporating airflow obstruction, exercise performance, dyspnea, and body mass index is a better predictor of mortality rate than pulmonary function alone. Recently, the GOLD added additional elements to their classification system incorporating respiratory symptoms and exacerbation history; these metrics are used to guide COPD treatment (see below).

TABLE 286-1 GOLD Criteria for Severity of Airflow Obstruction in COPD

GOLD Stage	Severity	Spirometry
I	Mild	FEV_1/FVC <0.7 and FEV_1 ≥80% predicted
II	Moderate	FEV_1/FVC <0.7 and FEV_1 ≥50% but <80% predicted
III	Severe	FEV_1/FVC <0.7 and FEV_1 ≥30% but <50% predicted
IV	Very severe	FEV_1/FVC <0.7 and FEV_1 <30% predicted

Abbreviations: COPD, chronic obstructive pulmonary disease; GOLD, Global Initiative for Chronic Obstructive Lung Disease.

Source: Reproduced with permission from the Global Strategy for Diagnosis, Management and Prevention of COPD 2014, © Global Initiative for Chronic Obstructive Lung Disease (GOLD), all rights reserved. Available from *http://www.goldcopd.org*.

Arterial blood gases and oximetry may demonstrate resting or exertional hypoxemia. Arterial blood gases provide additional information about alveolar ventilation and acid-base status by measuring arterial P_{CO_2} and pH. The change in pH with P_{CO_2} is 0.08 units/10 mmHg acutely and 0.03 units/10 mmHg in the chronic state. Knowledge of the arterial pH therefore allows the classification of ventilatory failure, defined as P_{CO_2} >45 mmHg, into acute or chronic conditions with acute respiratory failure being associated with acidemia. The arterial blood gas is an important component of the evaluation of patients presenting with symptoms of an exacerbation. An elevated hematocrit suggests the presence of chronic hypoxemia, as does the presence of signs of right ventricular hypertrophy.

Radiographic studies may assist in the classification of the type of COPD. Obvious bullae, paucity of parenchymal markings, or hyperlucency on chest x-ray suggests the presence of emphysema. Increased lung volumes and flattening of the diaphragm suggest hyperinflation but do not provide information about chronicity of the changes. Chest computed tomography (CT) scan is the current definitive test for establishing the presence or absence of emphysema, the pattern of emphysema, and the presence of significant disease involving medium and large airways (Fig. 286-2). It also enables the discovery of coexisting interstitial lung disease and bronchiectasis, which are common complications in COPD. Smokers with COPD are at high risk for development of lung cancer, which can be identified on a chest CT scan. In advanced COPD, CT scans can help determine the possible value of surgical therapy (described below).

Recent guidelines have suggested testing for α_1AT deficiency in all subjects with COPD or asthma with chronic airflow obstruction. Measurement of the serum α_1AT level is a reasonable initial test. For subjects with low α_1AT levels, the definitive diagnosis of α_1AT deficiency requires PI type determination. This is typically performed by isoelectric focusing of serum or plasma, which reflects the genotype at the PI locus for the common alleles and many of the rare PI alleles as well. Molecular genotyping of DNA can be performed for the common PI alleles (M, S, and Z).

TREATMENT

TREATMENT Chronic Obstructive Pulmonary Disease STABLE PHASE COPD

The two main goals of therapy are to provide symptomatic relief (reduce respiratory symptoms, improve exercise tolerance, improve health status) and reduce future risk (prevent disease progression, prevent and treat exacerbations, and reduce mortality). The institution of therapies should be based on symptom assessment, benefits of therapy, potential risks, and costs. **Figure 286-5** provides the currently suggested categories of COPD patients based on respiratory symptoms and risk for exacerbations. Response to therapy should be assessed, and decisions should be made whether or not to continue or alter treatment.

Only three interventions—smoking cessation, oxygen therapy in chronically hypoxemic patients, and lung volume reduction surgery (LVRS) in selected patients with emphysema—have been demonstrated to improve survival of patients with COPD. There is suggestive, but not definitive, evidence that the use of inhaled corticosteroids (ICS) and muscarinic antagonists may reduce the mortality rate.

PHARMACOTHERAPY Smoking Cessation

It has been shown that middle-aged smokers who were able to successfully stop smoking experienced a significant improvement in the rate of decline in pulmonary function, often returning to annual changes similar to that of nonsmoking patients (See also **Chap. 448**). In addition, smoking cessation improves survival. Thus, all patients with COPD should be strongly urged to quit smoking and educated about the benefits of quitting. An emerging body of evidence demonstrates that combining pharmacotherapy with traditional supportive approaches considerably enhances the chances of successful smoking cessation. There are three principal pharmacologic approaches to the problem: nicotine replacement therapy available as gum, transdermal patch, lozenge, inhaler, and nasal spray; bupropion; and varenicline, a nicotinic acid receptor agonist/antagonist. Current recommendations from the U.S. Surgeon General are that all adult, nonpregnant smokers considering quitting be offered pharmacotherapy, in the absence of any contraindication to treatment. Smoking cessation counseling is also recommended and free counseling is available through state Smoking QuitLines.

Bronchodilators

In general, bronchodilators are the primary treatment for almost all patients with COPD and are used for symptomatic benefit and to reduce exacerbations. The inhaled route is preferred for medication delivery, because side effects are less than with systemic medication delivery. In symptomatic patients, both regularly scheduled use of long-acting agents and as-needed short-acting medications are indicated. **Figure 286-6** provides suggestions for prescribing inhaled medication therapy based on grouping patients by severity of symptoms and risk of exacerbations.

Anticholinergic Muscarinic Antagonists

Short-acting ipratropium bromide improves symptoms with acute improvement in FEV_1. Long-acting muscarinic antagonists (LAMA, including aclidinium, glycopyrrolate, tiotropium, and umeclidinium) improve symptoms and reduce exacerbations. In a large randomized clinical trial, there was a trend toward reduced mortality rate in tiotropium-treated patients that approached statistical significance. Side effects are minor; dry mouth is the most frequent side effect.

Beta Agonists

Short-acting beta agonists ease symptoms with acute improvements in lung function. Long-acting agents (LABA) provide symptomatic benefit and reduce exacerbations, though to a lesser extent than a LAMA. Currently available long-acting inhaled β

agonists are arformoterol, formoterol, indacaterol, olodaterol, salmeterol, and vilanterol. The main side effects are tremor and tachycardia.

Combinations of Beta Agonist — Muscarinic Antagonist

The combination inhaled β agonist and muscarinic antagonist therapy has been demonstrated to provide improvement in lung function that is greater than either agent alone and reduces exacerbations.

Inhaled Corticosteroids

The main role of ICS is to reduce exacerbations. Although one large trial and a meta-analysis demonstrated an apparent benefit from the regular use of inhaled glucocorticoids on the rate of decline of lung function, a number of other well-designed randomized trials have not. A meta-analysis and retrospective studies suggest a mortality benefit, but in a large randomized trial, differences in mortality rate approached, but did not reach, conventional criteria for statistical significance. Their use has been associated with increased rates of oropharyngeal candidiasis and pneumonia and in some studies an increased rate of loss of bone density. A trial of ICS should be considered in patients with frequent exacerbations, defined as two or more per year, and in patients with features of asthma, such as eosinophilia. In stable patients, ICS withdrawal may be considered. Although ICS withdrawal does not lead to an increase in exacerbations, there may be a small decline in lung function.

Oral Glucocorticoids

The chronic use of oral glucocorticoids for treatment of COPD is not recommended because of an unfavorable benefit/risk ratio. The chronic use of oral glucocorticoids is associated with significant side effects, including osteoporosis, weight gain, cataracts, glucose intolerance, and increased risk of infection. A recent study demonstrated that patients tapered off chronic low-dose prednisone (~10 mg/d) did not experience any adverse effect on the frequency of exacerbations, health-related quality of life, or lung function.

Theophylline

Theophylline produces modest improvements in airflow and vital capacity, but is not first-line therapy due to side effects and drug interactions. Nausea is a common side effect; tachycardia and tremor have also been reported. Monitoring of blood theophylline levels is required to minimize toxicity.

PDE4 Inhibitors

The selective phosphodiesterase 4 (PDE4) inhibitor roflumilast has been demonstrated to reduce exacerbation frequency in patients with severe COPD, chronic bronchitis, and a prior history of exacerbations; its effects on airflow obstruction and symptoms are modest.

Antibiotics

There are strong data implicating bacterial infection as a precipitant of a substantial portion of exacerbations. A randomized clinical trial of azithromycin, chosen for both its anti-inflammatory and antimicrobial properties, administered daily to subjects with a history of exacerbation in the past 6 months demonstrated a reduced exacerbation frequency and longer time to first exacerbation in the macrolide-treated cohort (hazard ratio, 0.73).

Oxygen

Supplemental O_2 is the only pharmacologic therapy demonstrated to unequivocally decrease mortality rates in patients with COPD. For patients with resting hypoxemia (resting O_2 saturation ≤88% in any patient or ≤89% with signs of pulmonary hypertension or right heart failure), the use of O_2 has been demonstrated to have a significant impact on mortality. Patients meeting these criteria should be on continuous oxygen supplementation because the mortality benefit is proportional to the number of hours per day oxygen is used. Various delivery systems are available, including portable systems that patients may carry to allow mobility outside the home.

A recent study failed to demonstrate significant benefits to COPD patients with moderate hypoxemia at rest or with hypoxemia only with activity.

α_1AT Augmentation Therapy

Specific treatment in the form of IV α_1AT augmentation therapy is available for individuals with severe α_1AT deficiency. Despite sterilization procedures for these blood-derived products and the absence of reported cases of viral infection from therapy, some physicians recommend hepatitis B vaccination prior to starting augmentation therapy. Although biochemical efficacy of α_1AT augmentation therapy has been shown, the benefits of α_1AT augmentation therapy are controversial. A recent randomized study suggested a reduction in emphysema progression in patients receiving α_1AT augmentation therapy. Eligibility for α_1AT augmentation therapy requires a serum α_1AT level <11 μ*M* (~50 mg/dL). Typically, Pi^Z individuals will qualify, although other rare types associated with severe deficiency (e.g., null-null) are also eligible. Because only a fraction of individuals with severe α_1AT deficiency will develop COPD, α_1AT augmentation therapy is not recommended for severely α_1AT-deficient persons with normal pulmonary function and a normal chest CT scan.

NONPHARMACOLOGIC THERAPIES

Patients with COPD should receive the influenza vaccine annually. Pneumococcal vaccines and vaccination for *Bordetella pertussis* are recommended.

Pulmonary Rehabilitation

This refers to a comprehensive treatment program that incorporates exercise, education, and psychosocial and nutritional counseling. In COPD, pulmonary rehabilitation has been demonstrated to improve health-related quality of life, dyspnea, and exercise capacity. It has also been shown to reduce rates of hospitalization over a 6- to 12-month period.

Lung Volume Reduction Surgery

In carefully selected patients with emphysema, surgery to remove the most emphysematous portions of lung improves exercise, lung function, and survival. The anatomic distribution of emphysema and post-rehabilitation exercise capacity are important prognostic characteristics. Patients with upper lobe–predominant emphysema and a low post-rehabilitation exercise capacity are most likely to benefit from LVRS.

Patients with an FEV_1 <20% of predicted and either diffusely distributed emphysema on CT scan or diffusing capacity of lung for carbon monoxide (DL_{CO}) <20% of predicted have increased mortality after the procedure, and thus are not candidates for LVRS.

Methods of achieving lung volume reduction by using bronchoscopic techniques are under investigation.

Lung Transplantation

COPD is currently the second leading indication for lung transplantation (See also **Chap. 292**). Current recommendations are that candidates for lung transplantation should have very severe airflow limitation, severe disability despite maximal medical therapy, and be free of significant comorbid conditions such as liver, renal, or cardiac disease.

EXACERBATIONS OF COPD

Exacerbations are a prominent feature of the natural history of COPD. Exacerbations are episodic acute worsening of respiratory symptoms, including increased dyspnea, cough, wheezing, and/or change in the amount and character of sputum. They may or may not be accompanied by other signs of illness, including fever, myalgias, and sore throat. The strongest single predictor of exacerbations is a history of a previous exacerbation. The frequency of exacerbations increases as airflow obstruction worsens; patients with severe (FEV_1 <50% predicted) or very severe airflow obstruction (FEV_1 <30% predicted) on average have 1–3 episodes per year. However, some individuals with very severe airflow obstruction do not have frequent exacerbations. Other factors, such as an elevated ratio of the diameter of the pulmonary artery to aorta on chest CT, and gastroesophageal reflux, are also associated with increased risk of COPD exacerbations. Economic analyses have shown that >70% of COPD-related health care expenditures are due to emergency department visits and hospital care for COPD exacerbations; this translates to over $10 billion annually in the United States.

Precipitating Causes and Strategies to Reduce Frequency of Exacerbations

A variety of stimuli may result in the final common pathway of airway inflammation and increased respiratory symptoms that are characteristic of COPD exacerbations. Studies suggest that acquiring a new strain of bacteria is associated with increased near-term risk of exacerbation and that bacterial infection/superinfection is involved in >50% of exacerbations. Viral respiratory infections are present in approximately one-third of COPD exacerbations. In a significant minority of instances (20–35%), no specific precipitant can be identified.

Patient Assessment

An attempt should be made to establish the severity of the exacerbation as well as the severity of preexisting COPD. The more severe either of these two components, the more likely that the patient will require hospital admission. The history should include quantification of the degree and change in dyspnea by asking about breathlessness during activities of daily living and typical activities for the patient. The patient should be asked about fever; change in character of sputum; and associated symptoms such as wheezing, nausea, vomiting, diarrhea, myalgias, and chills. Inquiring about the frequency and severity of prior exacerbations can provide important information; the single greatest risk factor for hospitalization with an exacerbation is a history of previous hospitalization.

The physical examination should incorporate an assessment of the degree of distress of the patient. Specific attention should be focused on tachycardia, tachypnea, use of accessory muscles, signs of perioral or peripheral cyanosis, the ability to speak in complete sentences, and the patient's mental status. The chest examination should establish the presence or absence of focal findings, degree of air movement, presence or absence of wheezing, asymmetry in the chest examination (suggesting large airway obstruction or pneumothorax mimicking an exacerbation), and the presence or absence of paradoxical motion of the abdominal wall.

Patients with severe underlying COPD, who are in moderate or severe distress, or those with focal findings should have a chest x-ray or chest CT scan. Approximately 25% of x-rays in this clinical situation will be abnormal, with the most frequent findings being pneumonia and congestive heart failure. Patients with advanced COPD, a history of hypercarbia, mental status changes (confusion, sleepiness), or those in significant distress should have an arterial blood-gas measurement. The presence of hypercarbia, defined as a P_{CO_2} >45 mmHg, has important implications for treatment (discussed below). In contrast to its utility in the management of exacerbations of asthma, measurement of pulmonary function has not been demonstrated to be helpful in the diagnosis or

management of exacerbations of COPD. Pulmonary embolus (PE) should also be considered, as the incidence of PE is increased in COPD exacerbations.

The need for inpatient treatment of exacerbations is suggested by the presence of respiratory acidosis and hypercarbia, new or worsening hypoxemia, severe underlying disease and those whose living situation is not conducive to careful observation and the delivery of prescribed treatment.

TREATMENT OF ACUTE EXACERBATIONS Bronchodilators

Typically, patients are treated with inhaled β agonists and muscarinic antagonists. These may be administered separately or together, and the frequency of administration depends on the severity of the exacerbation. Patients are often treated initially with nebulized therapy, as such treatment is often easier to administer in those in respiratory distress. It has been shown, however, that conversion to metered-dose inhalers is effective when accompanied by education and training of patients and staff. This approach has significant economic benefits and also allows an easier transition to outpatient care. The addition of methylxanthines (theophylline) to this regimen can be considered, although convincing proof of its efficacy is lacking. If added, serum levels should be monitored in an attempt to minimize toxicity.

Antibiotics

Patients with COPD are frequently colonized with potential respiratory pathogens, and it is often difficult to identify conclusively a specific species of bacteria responsible for a particular clinical event. Bacteria frequently implicated in COPD exacerbations include *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Moraxella catarrhalis.* In addition, *Mycoplasma pneumoniae* or *Chlamydia pneumoniae* are found in 5–10% of exacerbations. The choice of antibiotic should be based on local patterns of antibiotic susceptibility of the above pathogens as well as the patient's clinical condition. Patients with moderate or severe exacerbations are usually treated with antibiotics, even in the absence of data implicating a specific pathogen.

In patients admitted to the hospital, the use of systemic glucocorticoids reduces the length of stay, hastens recovery, and reduces the chance of subsequent exacerbation or relapse. One study demonstrated that 2 weeks of glucocorticoid therapy produced benefit indistinguishable from 8 weeks of therapy. Current recommendations suggest 30–40 mg of oral prednisolone or its equivalent typically for a period of 5–10 days in outpatients. Hyperglycemia, particularly in patients with preexisting diagnosis of diabetes, is the most frequently reported acute complication of glucocorticoid treatment.

Oxygen

Supplemental O_2 should be supplied to maintain oxygen saturation ≥90%. Studies have demonstrated that in patients with both acute and chronic hypercarbia, the administration of supplemental O_2 does not reduce minute ventilation. It does, in some patients, result in modest increases in arterial P_{CO_2}, chiefly by altering ventilation-perfusion relationships within the lung. This should not deter practitioners from providing the oxygen needed to correct hypoxemia.

Mechanical Ventilatory Support

The initiation of noninvasive positive-pressure ventilation (NIPPV) in patients with respiratory failure, defined as Pa_{CO_2} >45 mmHg, results in a significant reduction in mortality rate, need for intubation, complications of therapy, and hospital length of stay. Contraindications to NIPPV include cardiovascular instability, impaired mental status, inability to cooperate, copious secretions or the inability to clear secretions, craniofacial abnormalities or trauma precluding effective fitting of mask, extreme obesity, or significant burns.

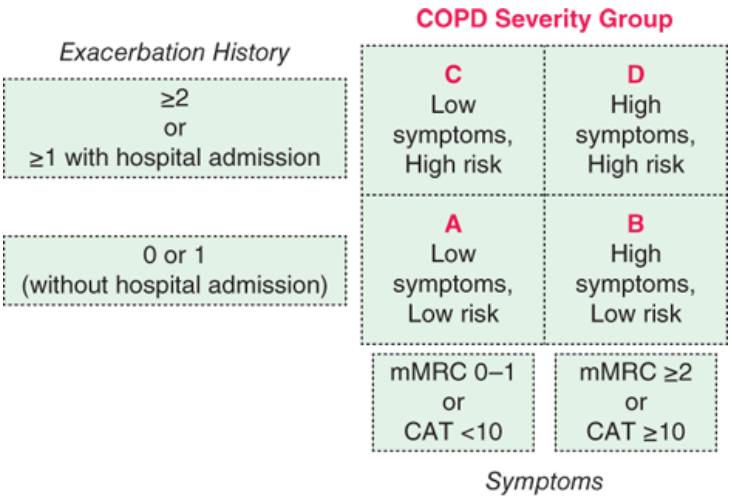
Invasive (conventional) mechanical ventilation via an endotracheal tube is indicated for patients with severe respiratory distress despite initial therapy, life-threatening hypoxemia, severe hypercarbia and/or acidosis, markedly impaired mental status, respiratory arrest, hemodynamic instability, or other complications. The goal of mechanical ventilation is to correct the aforementioned conditions. Factors to consider during mechanical ventilatory support include the need to provide sufficient expiratory time in patients with severe airflow obstruction and the presence of auto-PEEP (positive end-expiratory pressure), which can result in patients having to generate significant respiratory effort to trigger a breath during a demand mode of ventilation. The mortality rate of patients requiring mechanical ventilatory support is 17–30% for that particular hospitalization. For patients aged >65 admitted to the intensive care unit for treatment, the mortality rate doubles over the next year to 60%, regardless of whether mechanical ventilation was required.

Following a hospitalization for COPD, about 20% of patients are re-hospitalized in the subsequent 30 days and 45% are hospitalized in the next year. Mortality following hospital discharge is about 20% in the following year.

FIGURE 286-5

COPD severity assessment. COPD severity categories are based on respiratory symptoms (based on the mMRC or CAT scales) and annual frequency of COPD exacerbations. mMRC—Modified Medical Research Council Dyspnea Scale. Provides a single number for degree of breathlessness: 0—only with strenuous activity; 1—hurrying on level ground or walking up a slight hill; 2—walk slower than peers or stop walking at their own pace; 3—walking about 100 yards or after a few minutes on level ground; 4—too breathless to leave the house or when dressing. CAT—COPD Assessment Test. An 8-item COPD health status measure with Likert scale responses for questions about cough, phlegm, chest tightness, dyspnea on one flight of stairs, limitation in home activities, confidence in leaving the home, sleep and energy. Range of total score is 0–40. Both mMRC and CAT are available from Global

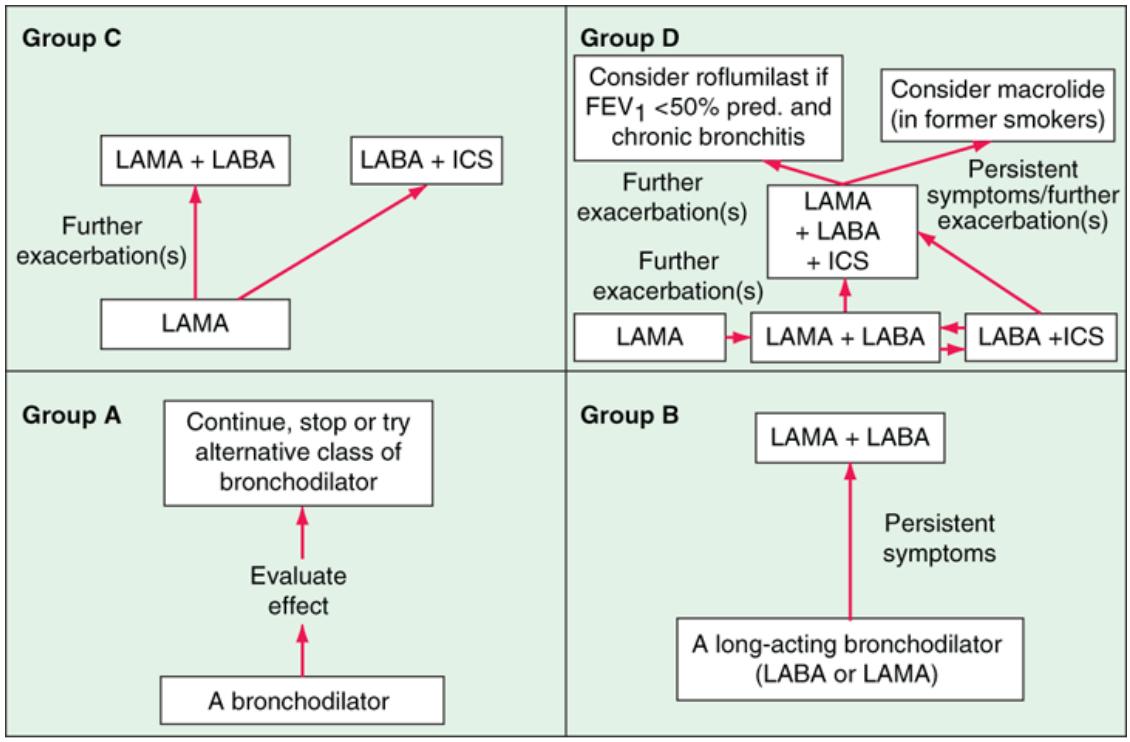
Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2017. With permission from http://goldcopd.org.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 286-6

Medication therapy for stable COPD. Recommended pharmacologic treatment of stable COPD is based on respiratory symptoms and exacerbation frequency. Preferred treatment Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2017. Reproduced with permission from http://goldcopd.org.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2017. Available from: http://goldcopd.org.
Hobbs B et al: Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. Nat Genet 49:426, 2017. [PubMed: 28166215]
Lange P et al: Lung-function trajectories leading to chronic obstructive pulmonary disease. N Engl J Med 373:111, 2015. [PubMed: 26154786]

The Long-Term Oxygen Treatment Trial research group: A randomized trial of long-term oxygen for COPD with moderate desaturation. N Engl J Med 375:1617, 2016. [PubMed: 27783918]
Lynch D et al: CT definable subtypes of COPD: A statement of the Fleischner Society. Radiology 277:192, 2015. [PubMed: 25961632]
Martinez FD: Early-life origins of chronic obstructive pulmonary disease. N Engl J Med 375:871, 2016. [PubMed: 27579637]
McDonough JE et al: Small-airway obstruction and emphysema in chronic obstructive pulmonary disease. N Engl J Med 365:1567, 2011. [PubMed: 22029978]
Regan E et al: Clinical and radiologic disease in smokers with normal spirometry. JAMA Intern Med 175:1539, 2015. [PubMed: 26098755]
Sandhaus RA et al: The diagnosis and management of alpha-1 antitrypsin deficiency in the adult. The diagnosis and management of alpha-1 antitrypsin deficiency in the adult. Chronic Obstr Pulm Dis (Miami). 3:668, 2016. doi: http://doi.org.ezproxy1.library.usyd.edu.au/10.15326/jcopdf.3.3.2015.0182
Spruit MA et al: An official American Thoracic Society/European Respiratory Society statement: Key concept and advances in pulmonary rehabilitation. Am J Resp Crit Care Med 188:e13, 2013. [PubMed: 24127811]

Chapter 287: Interstitial Lung Disease

Gary M. Hunninghake; Ivan O. Rosas

INTRODUCTION

Diffuse parenchymal lung diseases include a large number (>200) of heterogeneous conditions that affect the lung parenchyma with varying degrees of inflammation and fibrosis. While remodeling of the interstitial space, the region between the epithelium and endothelium, tends to be the dominant site of involvement for most of the interstitial lung diseases (ILDs), it is important to recognize the prominent role of the alveolar epithelium and endothelial cells (including both airways and vessels) in the pathogenesis of these interstitial lung disorders.

Despite the diverse array of conditions, most patients ultimately diagnosed with an ILD will come to medical attention with reports of progressive exertional dyspnea or a persistent dry cough. However, because some ILDs are part of multisystem disorders, some patients will be identified based on non-respiratory symptomatology (e.g., skin thickening in the setting of systemic sclerosis, **Chap. 353**) or physical examination findings (e.g., ulnar deviation of the fingers in the setting of rheumatoid arthritis [RA], **Chap. 351**). Additionally, ILDs can also be identified incidentally based on the results of abnormal pulmonary function tests, chest x-rays (CXRs), computed tomography (CT) studies of both the chest and abdomen (which can both visualize, at least a portion, of the lung parenchyma), and positron emission tomography (PET) scans. It is important to remember that ILDs can be associated with high rates of morbidity and mortality, and although prognosis depends on both disease extent and specificity, this fact makes these important disorders to recognize in a timely manner.

Owing to a variety of clinical presentations, as well as overlapping imaging and histopathologic findings **(Table 287-1)**, ILDs can be difficult to diagnose. A generally accepted central tenet of ILD diagnosis is that the combined weight of clinical data, laboratory studies, pulmonary function testing, imaging findings, and histopathology (if obtained) are jointly required to make a confident diagnosis. No single piece of data confers a diagnosis alone. For example, a lung biopsy demonstrating a usual interstitial pneumonia (UIP) pattern is helpful in diagnosing a patient with idiopathic pulmonary fibrosis (IPF) but can also be present in some connective tissue diseases (CTDs) (e.g., RA-associated ILD, **Chap. 351**). In light of this challenge, most ILD centers recommend a multidisciplinary approach to the diagnosis (and in some cases the management) of ILDs. An example of a multidisciplinary approach might include a conference attended by pulmonologists, rheumatologists, radiologists, and pathologists where all of the data generated on a patient can be discussed and reviewed jointly by those with unique sets of expertise in the care of patients with ILD.

TABLE 287-1 Common Interstitial Lung Disease Findings

	IPF	Nonspecific Interstitial Pneumonia	Respiratory Bronchiolitis Associated ILD	Systemic Sclerosis Associated ILD	Sarcoidosis
Clinical symptoms	Gradual onset of SOB, dry cough. Unusual in younger adults.	Subacute onset of SOB, dry cough. Frequently associated with other conditions.	Can be asymptomatic, or have SOB, and cough.	Gradual onset of SOB, dry cough. Fatigue, tightening of skin, exaggerated cold response, reflux, and difficulty swallowing.	Can be asymptomatic, or have SOB, and cough. Can also have fatigue, palpitations, eye, skin, and joint findings.
Physical examination findings	Frequent rales at lung bases, digital clubbing is common.	Frequent rales. Clubbing is less common.	Rales common. Clubbing is rare.	Can have rales in isolation. Also skin thickening, joint swelling, and telangiectasias.	Can be normal, rales may be present. Can have skin findings, joint pain, and enlarged lymph nodes.
Exposures	Idiopathic but many exposed to smoke. Genetic findings may explain >1/3 of the risk of the disease.	Can be idiopathic but should prompt consideration for associated conditions.	Strong association with smoking.	Mostly unknown, some debate about solvent and silicate exposures.	Mostly unknown, although silicate dusts thought to play a role in some cases.
HRCT findings	Bilateral subpleural reticular changes most prominent in lower, posterior lung zones. Traction bronchiectasis and honeycombing common. Classic UIP pattern is considered diagnostic.	Peripheral subpleural ground glass and reticular patterns. Traction bronchiectasis is common but honeycombing is rare. HRCT not diagnostic.	Diffuse patchy centrilobular ground glass nodules.	Can have UIP or NSIP patterns, also dilated esophagus, occasional mediastinal calcifications, and pulmonary vascular enlargement.	Can have mediastinal and hilar lymphadenopathy. Peribronchovascular reticular-nodular findings.
Histopathology	UIP pattern including fibroblastic foci, temporal and spatial heterogeneity, honeycombing.	Cellular or fibrotic pattern of NSIP. More uniform than a UIP pattern.	Respiratory bronchiolitis with adjacent inflammatory and fibrosing changes. Pigment laden macrophages.	Both UIP or NSIP patterns can occur.	Non-caseating granulomas.
Clinical course	50% 3–5 year mortality.	18% 5-year mortality.	25% 7-year mortality.	20–30% 10-year mortality.	Generally low but varies by state.

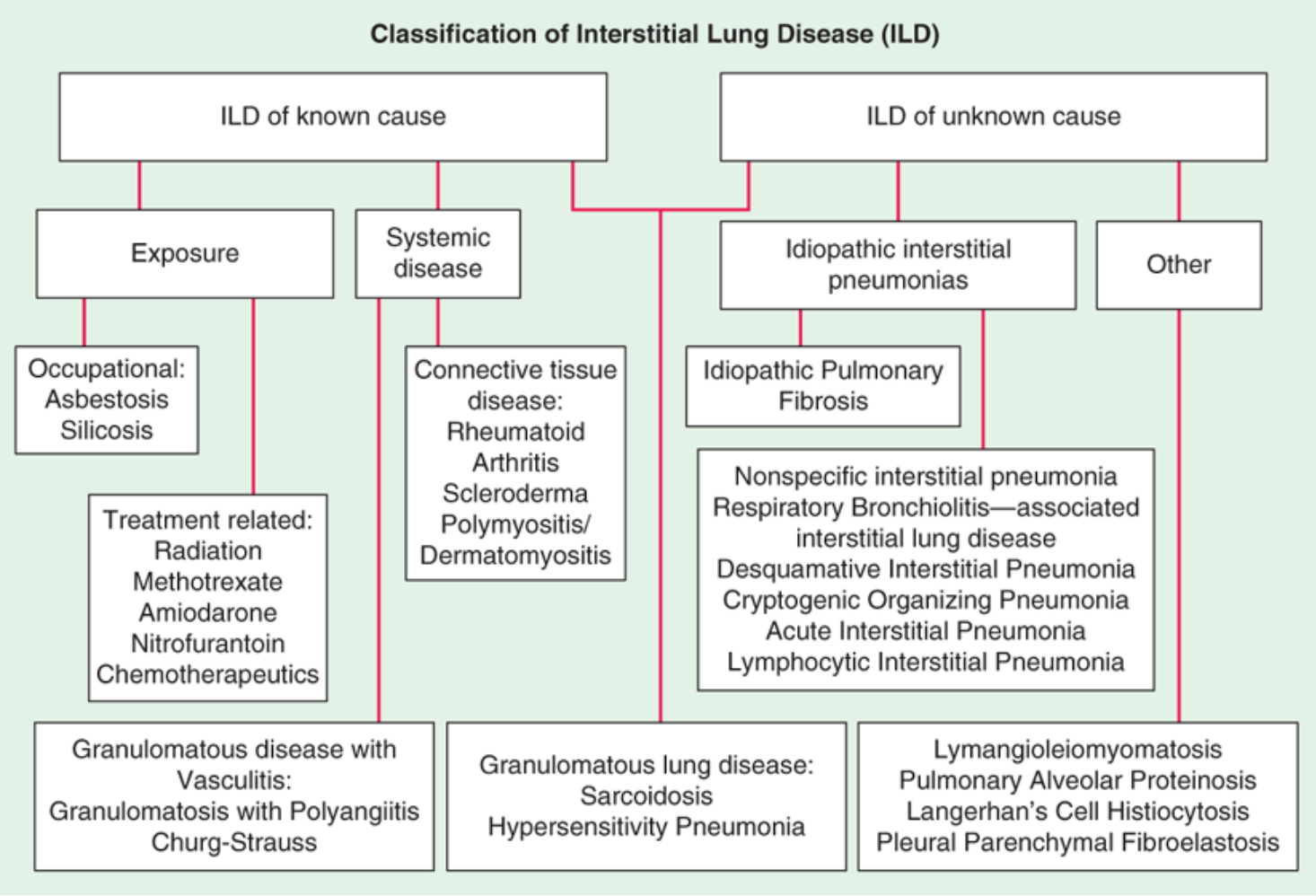
Abbreviations: HRCT, high resolution chest CT; ILD, interstitial lung disease; NSIP, non-specific interstitial pneumonia; SOB, shortness of breath; UIP, usual interstitial pneumonia

While there are numerous ways to categorize the ILDs, one classic approach is to divide the ILDs into those of known and unknown causes **(Fig. 287-1)**. Although even this approach has limitations (e.g., genetic studies demonstrate that a significant portion of familial and IPF [classically described as a diseases of unknown cause] may be explained, in part, by genetic factors), it is a useful place to start. Known causes of ILD include occupational exposures (e.g., asbestosis), medications (e.g., nitrofurantoin), and those related to an underlying systemic disease (e.g., cryptogenic organizing pneumonia [COP] in the setting of polymyositis). Unknown causes of ILD include groups of rare disorders often with classic presentations (e.g., a spontaneous pneumothorax in a young female with diffuse cystic changes on a chest CT might suggest lymphangioleiomyomatosis [LAM]) and the most common group of ILDs, the idiopathic interstitial pneumonias (IIPs). Granulomatous lung diseases straddle both known (e.g., hypersensitivity pneumonitis (HP) due to chronic bird exposure, **Chap. 282**), and unknown (e.g., sarcoidosis, **Chap. 360**) causes and are often separated due to their unique presentations, imaging findings, and

diagnostic evaluation. Equally important to knowledge of disease classification is knowledge of disease prevalence. Although there is variability within different demographic groups, most studies demonstrate that IPF, sarcoidosis **(Chap. 360)**, and ILDs related to CTDs **(Chap. 406)** as a group are among the most common forms of ILD.

Figure 287-1

Classification of interstitial lung disease. This algorithm represents a common approach to sub-classifying the interstitial lung diseases. It is typical to divide the interstitial lung diseases into those of known and unknown causes (although it is important to note that genetic studies demonstrate that a significant portion of familial and idiopathic pulmonary fibrosis [classically described as diseases of unknown cause] may be explained, in part, by genetic factors). The idiopathic interstitial pneumonias were more precisely defined by a 2002 study as described in Am J Respir Crit Care Med 165:277, 2002, referenced in the Further Reading list.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSTIC APPROACH

The initial diagnostic approach to diffuse parenchymal lung disease is often broader than a focus on ILD and should include an evaluation for alternate causes including cardiovascular disease (e.g., heart failure, **Chap. 253**), diffuse infections (e.g., *pneumocystis* pneumonia, **Chap. 215**), and malignancy (e.g., bronchoalveolar cell carcinoma, **Chap. 315 in HPIM 19e**). This chapter will focus on the diagnostic evaluation that helps to distinguish among the various forms of ILD.

HISTORY

Age

Age at presentation has a strong influence on the pretest probability that IPF, in particular, is present. For example, IPF occurs most commonly in patients aged >60 and is quite rare among patients aged <50. In fact, in patients aged >65 without strong evidence for an alternate diagnosis, atypical chest CT findings are still more likely to result in a histopathologic diagnosis of UIP (a pathologic hallmark of IPF) than they are to result in an alternate IIP diagnosis. Other common ILDs such as sarcoidosis, CTD associated ILD, and less common ILDs such as LAM, pulmonary Langerhans cell histiocytosis (PLCH) tend to present between the ages of 20 and 40.

Sex

Although less influential than age, sex has some influence on likelihood of various ILDs. LAM (and related disorder tuberous sclerosis) **(see Chap. 315 in HPIM 19e**) is a disorder that is frequently diagnosed in young women. Many CTD-associated ILDs are more common among women, with the exception of RA associated ILD which is more common among men. IPF and occupational/exposure-related ILDs (likely due to work related exposures that tend to differ between men and women) are more common among men.

Duration of Symptoms

Acute presentations (*days to weeks*) of ILD are unusual and are commonly misdiagnosed as more common diseases such as pneumonia, a COPD exacerbation, or heart failure. ILDs that can present acutely include eosinophilic pneumonia, acute interstitial pneumonia (AIP), HP, and granulomatosis with polyangiitis (GPA). An acute exacerbation of IPF as the initial presentation of this disease should also be a consideration given its prevalence. ILDs most commonly have a chronic indolent presentation (*months to years*) typified by IPF. However subacute presentations (*weeks to months*) can occur in most of the ILDs, but in the right context could suggest sarcoidosis, CTD associated ILD, drug-induced ILD, or COP.

Respiratory Symptoms

Progressive dyspnea, most frequently noted with exertion, is the most common complaint in patients presenting with an ILD. Despite this fact, both research studies of general population samples and clinical experiences of asymptomatic patient referrals with abnormal chest CT imaging patterns have also demonstrated that some patients, even those with more extensive disease, may not report dyspnea. Cough, particularly a dry cough, is also common, and can be the most prominent symptom in patients with IPF. Cough is often reported in other ILDs, particularly those that have prominent airway involvement including sarcoidosis and HP. Cough with hemoptysis is rare and could suggest an ILD associated with diffuse alveolar hemorrhage (DAH) (e.g., Goodpasture's syndrome), GPA, or LAM. Cough with hemoptysis could also suggest a secondary pulmonary infection that can be seen in patients with traction bronchiectasis and in those receiving immunosuppressive therapy. Chest pain is rare in most of the ILDs with the exception of sarcoidosis where chest discomfort is not uncommon. Fatigue is common to all of the ILDs.

Past Medical History

The most pertinent history includes a personal history of a CTD or a history of symptoms commonly associated with a CTD (e.g., Raynaud's phenomena). It is also important to remember that ILD associated with a CTD can be the initial presenting symptom of the disease and can precede the development of additional symptomatology by many years. A history of malignancy is important; as some malignancies can be associated with dermatomyositis associated COP and sarcoid-like reactions. A history of asthma and allergic rhinitis might suggest a diagnosis of eosinophilic GPA.

Medications

Many medications have been associated with ILD and to complicate matters further, many medications commonly used to treat inflammatory and granulomatous lung disease are also associated with ILD development (e.g., methotrexate, azathioprine, rituximab, and the tumor-necrosis factor-alpha blocking agents). Specific medications in many classes are also known to cause ILD, including antibiotics (e.g., nitrofurantoin), anti-arrhythmics (e.g., amiodarone) and many of the anti-neoplastic agents (e.g., bleomycin).

Family History

A family history of ILD (of almost any type) is important to ascertain. The percentage of pulmonary fibrosis that is familial, as opposed to idiopathic, varies by study, with estimates ranging from <5% to as high as 20%. Despite this variability, most agree that the presence of a close relative with an IIP is among the strongest risk factors for IPF. Family studies have consistently noted familial aggregation of diverse forms of IIP (such as IPF, non-specific interstitial pneumonia [NSIP], and DIP running in the same family) and in some cases other forms of ILD. To date, the most well replicated genetic factors for pulmonary fibrosis (a promoter variant of a mucin gene [*MUC5B*]) and various genetic determinants known to influence telomere length (e.g., variants in the telomerase reverse transcriptase gene [*TERT*]) appear to be associated with both familial and idiopathic forms of pulmonary fibrosis similarly.

Social History

A history of smoking is nearly always present in some forms of ILD (e.g., respiratory bronchiolitis and desquamative interstitial pneumonia [DIP]—sometimes referred by pathologists jointly as smoking related—ILD) where it is felt to be causative. A history of smoking is also noted in approximately three-quarters of IPF patients. Occupational and environmental exposure histories are also important to obtain as they might identify exposures known to cause pulmonary fibrosis (e.g., significant asbestos exposure) or HP (pigeon breeder's lung).

PHYSICAL EXAMINATION

End-inspiratory fine crackles, or rales, noted at the lung bases are found in most patients with IPF and may be one of the earliest signs of the disease. However, rales are nonspecific and can be found in many forms of ILD and other disorders. Wheezing is uncommon in most forms of ILD but can be present in some disorders, such as sarcoidosis, HP, and eosinophilic GPA. Signs of advanced disease include cyanosis, digital clubbing, and cor pulmonale.

LABORATORY STUDIES

Laboratory studies can be particularly helpful in the workup for an underlying CTD-associated ILD. As noted previously, these tests can reveal the presence of an underlying CTD as the cause of an ILD (e.g., a positive anti-cyclic citrullinated peptide [anti-CCP] antibody for RA) even when no other symptomatology or physical examination findings suggestive of the disorder are present. However, the cost-effectiveness and the extent of laboratory testing that should be ordered in various clinical contexts have yet to be determined (as there is a relatively long list of auto-antibody tests that could be ordered).

PULMONARY FUNCTION TESTS

Most forms of ILD will eventually result in a restrictive deficit on pulmonary function testing. A restrictive deficit is typified by a reduced total lung capacity (TLC), and symmetrically reduced measures of forced expiratory volume in 1 s (FEV1) and forced vital capacity (FVC). A reduction in the diffusing capacity of the lung for carbon monoxide (DLCO) is also common and may precede a reduction in lung volumes; however, there is more measurement variability in DLCO measurement and the test is less specific for ILD. A reduced FEV1 to FVC ratio, which is diagnostic of airway obstruction, is unusual in many forms of ILD but can be present as an isolated finding, or in conjunction with an additional restrictive deficit, in ILDs involving the airways such as sarcoidosis, HP, and LAM. Although pulmonary function testing is rarely diagnostic, reductions in lung function help to characterize the extent of disease, and evidence for decline in repeated measures of pulmonary function (e.g., FVC) have been correlated with an elevated rate of mortality.

CHEST IMAGING STUDIES

Chest X-Ray

Findings on CXR can be the first clinical indication that an ILD might be present. For example, enlarged hilar lymph nodes and a pattern of central nodular opacities in the mid to upper lung zones can suggest sarcoidosis. A basilar reticular pattern, with small cystic spaces, in the absence of clinical evidence for heart failure, might suggest IPF. With a few exceptions, CXR alone rarely leads to a specific diagnosis.

Chest CT

High resolution chest CT (HRCT) imaging is now considered to be standard of care in the initial evaluation of a patient with a suspected ILD. HRCT can be diagnostic for some ILDs (e.g., IPF) in right clinical context and may preclude the need for, and spare the patient the risk of, a lung biopsy. HRCT also helps to define the extent of the ILD, the presence of more concerning features suggestive of advanced disease (e.g., honeycombing), can provide information on coexisting diseases (e.g., emphysema and lung cancer), and when not diagnostic, can help to provide the most useful locations for obtaining lung biopsy specimens.

LUNG BIOPSY

Fiberoptic Bronchscopy

Bronchoscopy can be helpful in establishing a specific ILD diagnosis, and can help to establish an alternate diagnosis, in select cases. Examination of serial lavage fluid can be helpful in establishing DAH which can be present in ILDs with vasculitis (e.g., GPA), and in some cases, cellular examination can suggest a specific diagnosis (eosinophilia >25% in chronic eosinophilic pneumonia or fat globules in macrophages in lipoid pneumonia). Trans-bronchial lung biopsies and lymph node biopsies (in sarcoidosis in particular) can lead to a confident diagnosis in patients with likely granulomatous lung disease (e.g., sarcoidosis and HP). However, in general, bronchoscopically obtained tissue samples are often felt to be insufficient to diagnose most of the IIPs. There is some preliminary evidence that bronchoscopically obtained cryobiopsies, which can result in yields larger than those obtained by transbronchial forceps biopsies, could improve the diagnostic yield of bronchoscopy; however, the precise role cryobiopsies in the diagnostic workup of ILD has yet to be clarified.

Surgical Lung Biopsy

A surgically obtained lung biopsy specimen can help solidify the diagnosis of ILD. In many cases these are now obtained through a video-assisted thoracoscopic (VATS) approach (as compared to an open thoracotomy), which tends to reduce the length of operative times and hospital stays. The diagnostic yield of biopsies tends to be higher if obtained prior to treatment. The desire to obtain a surgical lung biopsy should be weighed against the risks which can include a short-term mortality rate of as high as 5%. These risks are reported to be higher in biopsies of patients ultimately diagnosed with IPF, and in those presenting acutely.

INDIVIDUAL FORMS OF ILD

The ILDs include a diverse group of lung pathologies that can be subclassified into those disorders of unknown cause (e.g., IIPs), and those of known cause (e.g., sometimes referred to as secondary interstitial pneumonias [connective tissue disease-associated ILDs]) (see Fig. 287-1). Although this remains a useful approach to classifying this diverse group of disorders it is important to recognize that genetic studies are challenging this classic categorization. For example, numerous ILDs commonly listed as having an “unknown cause” have been determined to have significant genetic underpinnings (e.g., IPF and LAM), while the pathophysiologic processes that result in ILDs of “known cause” (e.g., connective tissue disease) remain incompletely understood. Diagnosis is based on combined information obtained from a patient’s clinical presentation, measures of pulmonary function, imaging, immune serologies, and histopathology. It is important to remember that prognosis and treatment vary widely by disorder (and disease extent). In some cases, medical therapy that is felt to be effective for some ILDs has been proven to be harmful for others. Medical treatments range from immune modulators to anti-fibrotic medications while lung transplantation remains the standard of care for those with advanced and rapidly progressive ILDs.

IDIOPATHIC INTERSTITIAL PNEUMONIAS

IDIOPATHIC PULMONARY FIBROSIS

Clinical Manifestations

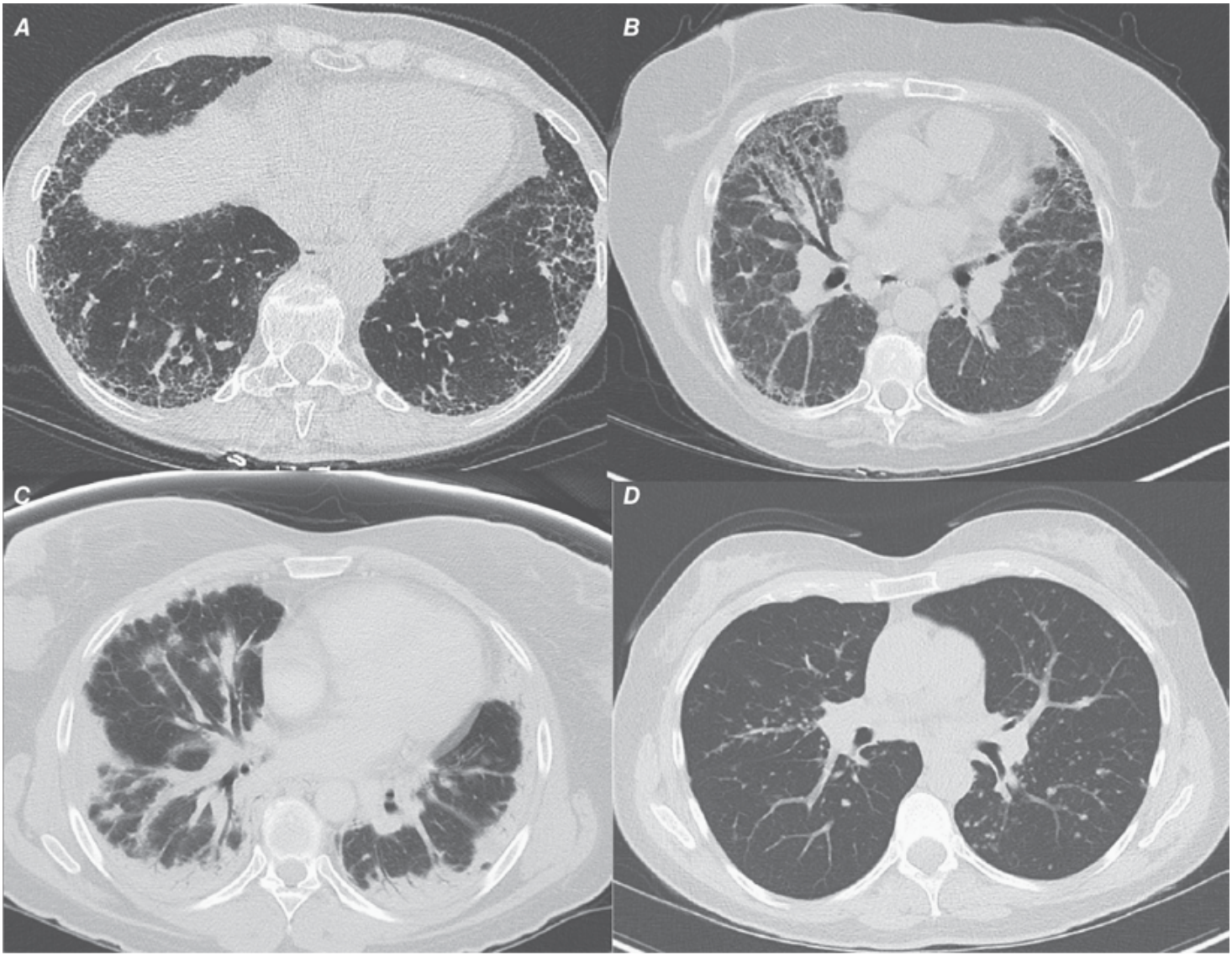
IPF is the most common ILD of unknown cause. Prevalence increases with age and is estimated at 50–200:100,000. IPF is commonly diagnosed in the fifth or sixth decade in life, affects men more than women, and is frequently associated with a history of smoking or other environmental exposures. IPF is a variably progressive disease that carries a poor prognosis with an estimated 50% 3–5-year survival.

HRCT Image Findings

Chest CT findings include subpleural reticulation with a posterior basal predominance usually including more advanced fibrotic features, such as honeycombing and traction bronchiectasis. Collectively these imaging findings are referred to as a UIP pattern. The presence of extensive ground glass opacities, bronchovascular changes, micronodules, mosaic attenuation, or an upper lung predominance should raise suspicion for an alternative diagnosis **(Fig. 287-2)**.

Figure 287-2

Chest CT imaging and interstitial lung disease. ***A***. Idiopathic pulmonary fibrosis (IPF): Classic findings of IPF (apparent on this image) include a posterior, basilar predominance of subpleural reticular markings and more advanced features of pulmonary fibrosis including traction bronchiectasis and honeycombing. This constellation of findings is often referred to as a usual interstitial pneumonia (UIP) pattern. ***B***. Non-specific interstitial pneumonia (NSIP): Chest CT findings of NSIP can overlap with those of a UIP pattern but tend to include a bilateral, symmetric pattern that presents with a greater percentage of ground-glass opacities than is apparent in a UIP pattern. Additional unique findings include more diffuse imaging abnormalities with a predominance not limited to the lung bases, imaging abnormalities that spare the subpleural regions, and thickening of the bronchovascular bundles (as is apparent in the right mid lung zone on this image). ***C***. Cryptogenic organizing pneumonia: Chest CT findings include patchy, sometimes migratory, subpleural consolidative opacities (as is apparent on this image) often with associated ground-glass opacities. Peribronchiolar, or perilobar opacities can be present and sometimes a rim of subpleural sparing (often referred to as a reversed halo or atoll sign) can be seen which can help to aid in the diagnosis. ***D***. Sarcoidosis: Sarcoidosis can present with varied imaging abnormalities but a pattern of mediastinal and hilar lymphadenopathy with a pattern of reticular-nodular opacities involving the bronchovascular bundles (apparent in this image) are common features. Additional findings can include diffuse small nodules in a miliary pattern, larger nodular opacities, extensive ground glass infiltrates and, mosaic attenuation suggestive of small airways involvement, and in more advanced cases, signs of pulmonary fibrosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

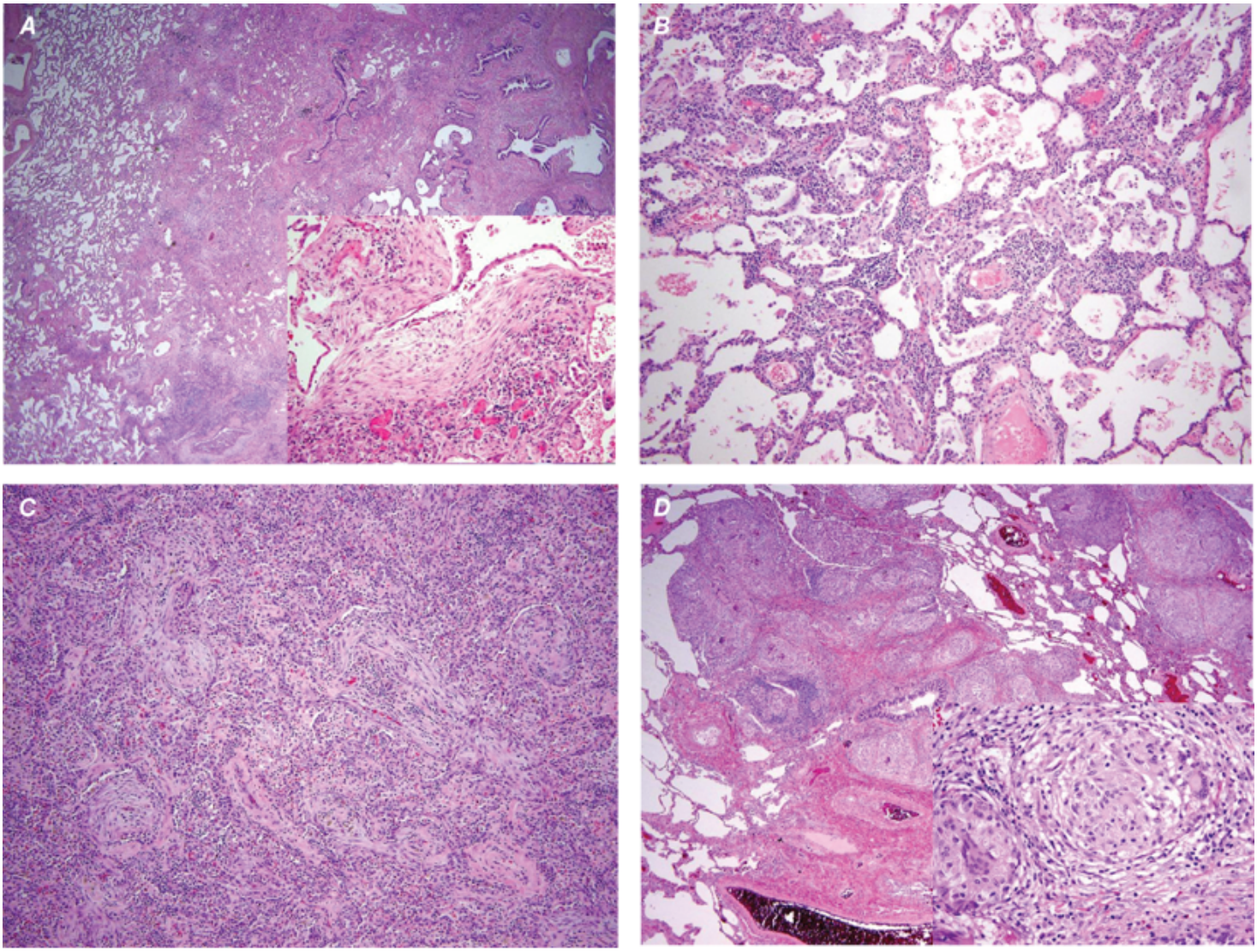
Histopathology

Diagnostic VATS biopsy findings include subpleural reticulation associated with honeycomb changes and fibroblast foci (subepithelial collections of myofibroblasts and collagen). These fibrotic changes alternate with areas of preserved normal alveolar architecture consistent with temporal and spatial heterogeneity **(Fig. 287-3)**. Collectively, these pathologic findings are referred to as UIP.

Figure 287-3

Histopathology of interstitial lung disease. ***A***. Idiopathic pulmonary fibrosis (IPF): Histopathologic findings include subpleural reticulation associated with honeycomb changes alternating with areas of preserved normal lung architecture referred to as temporal and spatial heterogeneity (as is apparent in the low power image above). Additional important diagnostic findings include fibroblast foci, which are subepithelial collections of myofibroblasts and collagen (as is apparent in the higher powered inset of this image). Collectively these pathologic findings are referred to as usual interstitial pneumonia (UIP). ***B***. Non-specific interstitial pneumonia (NSIP): Histopathologic findings of NSIP include varying amounts of interstitial inflammation and fibrosis with a uniform appearance (as is apparent in this image). Honeycomb changes are usually absent and fibroblast foci are rare. NSIP is often referred to histopathologically as being either predominantly cellular or fibrotic. ***C***. Cryptogenic organizing pneumonia (COP): Histopathologic findings of COP include patchy regions of organizing pneumonia with granulation tissue that commonly involves the small airways, alveolar ducts, and alveoli with surrounding inflammation that can

involve the alveolar walls (as is apparent in this image). ***D***. Sarcoidosis: The hallmark histopathologic feature of sarcoidosis is presence of granulomas (as are apparent numerously in the low powered image and more closely visualized in the higher powered inset image). Typically these are referred to as non-caseating which suggests the absence of necrosis. Caseating granulomas are rare in sarcoid and should prompt additional evaluation for an underlying infection. Because malignancy can result in a granulomatous reaction it is important to closely survey biopsy specimens with granulomatous involvement for additional signs of malignancy.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Treatment

Historically, IPF was felt to be refractory to medical therapy with lung transplantation the only viable therapeutic option. This dogma changed in 2014 with large clinical trials that demonstrated that antifibrotic therapy (pirfenidone and nintedanib) can slow decline of lung function in IPF patients. Further meta-analyses have suggested that anti-fibrotic therapy may also improve survival. In contrast, treatment with immunosuppression, which had been commonly prescribed to many IPF patients, has now been demonstrated (in some cases) to be associated with increased morbidity and mortality. Physical therapy and supplemental oxygen, when indicated, can improve exercise tolerance and reduce likelihood of developing pulmonary hypertension. Lung transplantation can extend survival and improve the quality of life in a subset of IPF patients who meet criteria to undergo transplant.

NON-SPECIFIC INTERSTITIAL PNEUMONIA

Clinical Manifestations

Idiopathic NSIP is a distinct clinical entity with characteristic clinical, radiologic, and pathologic features; however, NSIP is also commonly observed in patients with connective tissue disease and less frequently with familial interstitial pneumonia, drug toxicity, and infection. Although the prevalence of NSIP is not well established, it is commonly diagnosed in non-smoking females in their fifth decade of life. Positive serologic tests for connective tissue disease are frequently observed. Idiopathic NSIP has a relatively good prognosis, with a 5-year survival >80%; patients with a predominant cellular NSIP pattern have a more favorable prognosis than those with a fibrosing NSIP pattern.

HRCT Image Findings

Diffuse subpleural, symmetric, ground glass, and reticular opacities are common. Volume loss and traction bronchiectasis involving the lower lung zones can also be found. Occasionally subpleural sparing is noted, while peribronchiolar thickening and honeycombing are uncommon.

Histopathology

Diagnostic lung biopsy findings include varying amounts of interstitial inflammation and fibrosis with a uniform appearance. Honeycomb changes are usually absent and fibroblast foci are rare. NSIP is often referred to histopathologically as being either predominantly cellular (and potentially more responsive to medical therapy) or fibrotic (and potentially less likely to resolve with medical therapy).

Treatment

Pulmonary fibrosis associated with connective tissue disease is commonly treated with immunosuppression despite the paucity of randomized clinical trials to demonstrate efficacy. Idiopathic NSIP is often treated with oral steroids (prednisone), cytotoxic agents (mycophenolate, azathioprine, and cyclophosphamide), or biologics (rituximab). Oxygen therapy, pulmonary rehabilitation, and lung transplantation may be required in patients with progressive disease.

SMOKING-RELATED ILD

Although smoking-related ILDs including respiratory bronchiolitis with interstitial lung disease (RB-ILD), and DIP are frequently subclassified with the IIPs, these disorders (along with PLCH, an ILD with unique clinical, imaging and histopathologic manifestations) are commonly felt to be the result of active or prior tobacco smoke exposure. DIP has also been known to occur in children with familial pulmonary fibrosis (FPF). Smokers, particularly elderly smokers, frequently have radiologic (centrilobular) interstitial abnormalities. These interstitial abnormalities are often incidentally found on routine CXR or chest CT studies in asymptomatic, or minimally symptomatic individuals. Respiratory bronchiolitis is felt to correlate histopathologically with these imaging findings. However, in some cases these imaging findings can progress to more advanced radiologic changes where more diffuse signs of interstitial pneumonia tend to be present.

Clinical Manifestations

These disorders predominantly occur in active, and in many cases heavy, smokers who are typically between 40 and 50 years of age. In those ultimately diagnosed with RB-ILD or DIP, dyspnea and cough are relatively common and symptomatic wheezing is not rare. The prevalence of smoking-related ILDs is not well understood, but they are generally felt to account for <10% of the IIPs. While there is minimal data on the natural histories and prognoses of these conditions, prolonged survival can be expected in most patients with RB-ILD and death secondary to progressive ILD is felt to be rare.

HRCT Image Findings

Prominent and common findings in RB-ILD include central bronchial wall thickening, peripheral bronchial wall thickening, centrilobular nodules, and ground-glass opacities. Septal lines and a reticular pattern are also not uncommon. Honeycombing is generally felt to be rare (and indicates a worse prognosis). Similar findings are noted in patients with DIP where diffuse (or patchy) bilateral symmetric ground-glass opacities tend to be even more prominent.

Histopathology

Common features of RB-ILD include the accumulation of pigmented macrophages within the lumens of respiratory bronchioles and alveolar ducts, accompanied by chronic inflammation of the respiratory bronchiolar walls and both bronchiolar and peribronchiolar alveolar fibrosis causing architectural distortion. These features are patchy and confined to the peribronchiolar region. DIP tends to include similar changes but they have a more diffuse pattern characterized by pigmented macrophage accumulation, pneumocyte hyperplasia, and prominent interstitial thickening.

Treatment

All patients with smoking-related ILD should be counseled to discontinue smoking and/or encouraged to enroll in a formal smoking cessation program. Small studies have evaluated, and patients are often treated with immunosuppressive (e.g., prednisone) and cytotoxic (e.g., azathioprine, and cyclophosphamide) agents and in some cases with bronchodilators. To date there is no strong evidence that these therapies result in significant improvements symptoms, measures of pulmonary function, or if they prevent clinical deterioration.

CRYPTOGENIC ORGANIZING PNEUMONIA

Clinical Manifestations

COP typically involves patients in their 50–60s and often presents as a subacute flu-like illness, with cough, dyspnea, fever, and fatigue. Inspiratory rales are often present on examination and most patients are noted to have restrictive lung deficits on pulmonary function testing with hypoxemia. It is commonly mistaken for pneumonia. It is important to note that this syndrome can occur in isolation or can be secondary to an underlying connective tissue disease (e.g., polymyositis), medications, or can result from an underlying malignancy. Laboratory testing for various connective tissue diseases is helpful as they can both be diagnostic and suggest the need for prolonged medical therapy.

HRCT Image Findings

The most common imaging findings include patchy, sometimes migratory, subpleural consolidative opacities often with associated ground-glass opacities. Peribronchiolar, or perilobar opacities can be present and sometimes a rim of subpleural sparing (often referred to as a reversed halo or atoll sign) can be seen which can aid in the diagnosis.

Histopathology

Surgical lung biopsy specimens tend to reveal patchy regions of organizing pneumonia with granulation tissue that commonly involves the small airways, alveolar ducts, and alveoli with surrounding inflammation that can involve the alveolar walls (Fig. 287-2).

Treatment

Corticosteroids can result in substantial clinical improvement in many patients but usually need to be continued for at least 6 months as relapse rates are high. Evidence is growing that alternate cytotoxic (e.g., mycophenolate, cyclophosphamide) or biologic (e.g., rituximab) therapies can be helpful in both treating the disease and reducing the need for steroids. In some patients with secondary forms of the disease, long-term therapy may be needed.

ACUTE OR SUBACUTE IIPS

ACUTE INTERSTITIAL PNEUMONIA (HAMMAN-RICH SYNDROME)

Clinical Manifestations

AIP is a rare and often fatal lung disorder that is characterized by an acute onset of respiratory distress and hypoxemia. A prodromal period of symptoms consistent with an acute upper respiratory infection is common. The mortality rate within 6 months of presentation can be quite high (>50%) and recurrences are common. In those that recover, lung function improvement can be substantial. AIP can be difficult to distinguish from acute respiratory distress syndrome (ARDS) and an acute exacerbation of an unsuspected underlying pulmonary fibrotic process.

HRCT Image Findings

The most common imaging findings are patchy bilateral ground-glass opacities. Dependent regions of air-space consolidation are also common.

Histopathology

Similar to ARDS and acute exacerbations of underlying pulmonary fibrosis, AIP presents histopathologically as diffuse alveolar damage (DAD) demonstrated on a surgical lung biopsy.

Treatment

Treatment is mostly supportive and often includes mechanical ventilation. There is no proven drug therapy for AIP. Glucocorticoids are often given but they are not clearly effective and have been demonstrated not to be beneficial in other forms of DAD (e.g., ARDS).

ACUTE EXACERBATIONS OF IIPS

Clinical Manifestations

Acute exacerbations are not separate disorders, but rather an accelerated phase of lung injury that can occur in any ILD resulting in pulmonary fibrosis. Acute exacerbations are most commonly described, and most severe in, patients with known IPF. Acute exacerbations are characterized by an acute onset (<30 days) of respiratory distress and hypoxemia occurring in a patient with underlying pulmonary fibrosis not explained by an alternate cause (e.g., pneumonia, left heart failure). Reported mortality rates are very high (>85%) and mean survival periods range from as little as days to months.

HRCT Image Findings

The most common imaging findings include patchy bilateral ground-glass opacities and dependent regions of air-space consolidation. Sometimes these new changes can be appreciated on the background of the imaging findings typified by the underlying IIP, although sometimes they obscure the preceding imaging findings.

Histopathology

Acute exacerbations of underlying pulmonary fibrosis present histopathologically as DAD, although sometimes organizing pneumonia can also be demonstrated on a surgical lung biopsy.

Treatment

Treatment is mostly supportive. Mechanical ventilation, when not being used as a bridge to lung transplantation, is controversial as the survival rate in these patients tends to be poor. There is some evidence that drug therapy (e.g., Nintedanib) may reduce the rate of acute exacerbations in patients with IPF. Drug therapy, in the context of an acute exacerbation is also controversial. Immunosuppressive (e.g., prednisone) and cytotoxic (e.g., cyclophosphamide) therapies are commonly used without proven benefit.

ILD ASSOCIATED WITH CONNECTIVE TISSUE DISEASE

ILD is a common disease manifestation of many connective tissue diseases. Disease progression, response to therapy and survival is variable and associated with specific radiologic and histopathologic patterns. ILD occurs most commonly in patients with scleroderma (systemic sclerosis form, or SSc), RA, polymyositis/dermatomyositis, and less frequently Sjögren syndrome and systemic lupus erythematosus (SLE). ILD may precede the development of extrapulmonary manifestations of a specific connective tissue disease or may present as part of a poorly defined connective tissue disease. In rare cases, lung manifestations may be the sole feature of the patient's clinical presentation.

SYSTEMIC SCLEROSIS

Clinical Manifestations

ILD is the most common pulmonary manifestation of SSc **(Chap. 353)**. ILD occurs in about 50% of SSc patients with diffuse disease and in about 30% of patients with limited disease. Pulmonary hypertension can occur separately or concomitantly with ILD and is more frequent in patients with limited SSc.

HRCT Image Findings

Similar imaging findings noted in both patients with NSIP and IPF can be present, although findings consistent with COP and DAD may also be present. Additional HRCT findings may include a dilated esophagus and pulmonary artery enlargement.

Histopathology

Comparable to the imaging overlap, histopathologic changes commonly noted in patients with NSIP and IPF are frequently identified. Additionally, aspiration related to esophageal dysmotility is common in SSc, in these patients histopathologic findings consistent with COP and DAD may be observed.

Treatment

Cyclophosphamide has a modest benefit in preservation of lung function and is associated with significant toxicity. Mycophenolate has recently been shown to have similar efficacy and improved tolerability. Clinical trials testing antifibrotic therapies (pirfenidone and nintedanib) are presently being conducted. Minimizing the risk of reflux by using high-dose proton pump inhibitors or antireflux surgery should be considered in SSc with progressive ILD. Lung transplantation can potentially be offered to select patients without significant aspiration or chest wall restriction.

RHEUMATOID ARTHRITIS

Clinical Manifestations

A common extraarticular complication of RA is ILD **(Chap. 351)**. Although RA is more common in females, RA-ILD is more frequent in males and in patients with a history of tobacco exposure. In a small subset of patients, ILD is the first disease manifestation of RA. Clinically evident disease RA-ILD occurs in nearly 10% of the RA population; however, up to 40–50% of subjects have radiologic abnormalities on chest CT suggesting ILD in the context of RA may be under-diagnosed.

HRCT Image Findings

The most common imaging pattern of ILD in patients with RA is a UIP pattern, although NSIP patterns are not uncommon. There is evidence that survival in patients with RA is decreased in those with a UIP pattern and among those with more extensive fibrosis in general.

Histopathology

Histopathologic findings of UIP and NSIP are most common. Some studies suggest that UIP in the context of RA (as compared to IPF) may present with a reduced number of fibroblastic foci and an increased amount of germinal centers. Comparable to the imaging findings, UIP (and DAD) patterns in patients with RA are associated with reduced survival.

Treatment

In contrast with SSc, there are no randomized clinical trials testing the role of immune suppression in RA-ILD. Extrapolating from the scleroderma experience, immunosuppresive (e.g., prednisone) and cytotoxic (e.g., mycophenolate, azathioprine, cyclophosphamide, and calcineurin inhibitors) agents have been used with variable success. Clinical trials testing antifibrotic therapies (pirfenidone and nintedanib) are presently being conducted. Lung transplantation is a viable therapeutic approach for eligible patients with progressive disease that is not responsive to medical therapy.

DERMATOMYOSITIS/POLYMYOSITIS

Clinical Manifestations

The idiopathic inflammatory myopathies are disorders characterized by immune-mediated destruction and dysfunction of muscle, however this disorder can affect the skin, joints, cardiovascular system and lung **(Chap. 358)**. The prevalence of ILD associated with inflammatory myopathy varies by report, however ILD is present in up to 45% of patients with positive anti-synthetase antibodies. The anti-synthetase syndrome is characterized by positive anti-synthetase antibodies, myositis, fever, Raynaud phenomenon, mechanic's hands, arthritis, and progressive ILD. There is a subset of anti–Jo-1 antibody–positive individuals who can develop a rapidly progressive form of ILD consistent with an acute exacerbation. Some studies have suggested that ILD may be even more common in those with other antibodies (e.g., anti-PL-12). Dermatomyositis/polymyositis can occur as an isolated connective tissue disease or as a process associated with an underlying malignancy.

HRCT Image Findings

Common imaging patterns of ILD in patients with dermatomyositis/polymyositis include those consistent with NSIP with or without evidence for COP. A UIP pattern can also occur. Some studies have suggested that a UIP pattern may be more common among those with anti-PL-12 antibodies.

Histopathology

The antisynthetase syndrome is associated with multiple histopathologic subtypes including NSIP, COP, and UIP. DAD, a histopathologic pattern observed in AIP and acute exacerbations, is associated with rapidly progressive ILD in myositis patients.

Treatment

Immunosuppresive (e.g., prednisone) and cytotoxic (e.g., mycophenolate, azathioprine, cyclophosphamide, and calcineurin inhibitors) agents are often used in patients with progressive ILD. Some patients (particularly those with less fibrosis) have been noted to improve or resolve their ILD in response to medical therapy. In small studies relapses have been more common in patients treated with prednisone alone. Patients who fail immune suppressive therapy can benefit from lung transplantation.

GRANULOMATOUS ILDS

The most common granulomatous ILD is sarcoidosis, a multisystem disorder of unknown cause where lung involvement is often the most dominant feature, will be discussed in **Chap. 360**. HP, a granulomatous reaction due to inhalation of organic (e.g., bird fancier's lung secondary to exposure to bird feathers) and inorganic (e.g., coal worker's pneumoconiosis secondary to exposure to coal dusts) dusts, is also an important and common cause of ILD and is discussed in **Chap. 282**.

Granulomatous Vasculitides

These disorders are characterized by blood vessels with inflammatory infiltrates associated granulomatous lesions with or without the presence of tissue necrosis (**See Chap. 60**). The lungs are commonly involved and a unique feature of these disorders is that hemoptysis can be a presenting symptom. Although laboratory testing is often helpful and can provide specific information, biopsies of involved tissue can be essential for making the diagnosis. Many of these disorders include additional systemic manifestations. GPA, also referred to as Wegener's disease, is an example of a granulomatous vasculitis that commonly affects the lung (including inflammatory infiltrates in small to medium sized vessels), the ears, nose, throat, and kidney (resulting in glomerulonephritis). Common imaging abnormalities of GPA include nodules, patchy ground glass, and consolidative opacities that can be migratory, and hilar lymphadenopathy. Eosinophilic GPA (EG, also referred to as Churg-Strauss syndrome) is another example of a granulomatous vasculitis that affects the lung (including eosinophilic infiltrates in small to medium sized vessels) that can result in numerous clinical manifestations but frequently includes chronic sinusitis, asthma, and peripheral blood eosinophilia. Common imaging abnormalities of EG include peripheral consolidative opacities that can be migratory and small pleural effusions.

GENETICS AND ILD

Studies of genetic epidemiology have led to important insights in our understanding of ILD. First, studies of families with FPF have demonstrated that unique IIPs can cosegregate with specific genetic variants known to be associated with IPF. This suggests that many genetic variants appear to predispose to interstitial lung injury patterns more broadly than to unique diagnoses specifically. Second, most of the genetic variants known to be associated with FPF are also associated with more sporadic forms of the disease. Third, at least one of the genetic factors most strongly associated with FPF and IPF is both common and confers a large increase in the risk of these diseases. At least one copy of a mucin 5B (*MUC5B*) promoter variant is present in ~20% of Caucasian populations and 35–45% of patients with IPF and confers and approximate sixfold increase in the risk of this disease. Fourth, studies of general population samples demonstrate that imaging abnormalities suggestive of an early stage of pulmonary fibrosis in research participants without known ILD are not uncommon (occurring in ~7–9% of adults) and are also associated with the same genetic variants known to be associated with IPF (e.g., the *MUC5B* promoter variant). This latter finding suggests a path forward towards an early detection of IPF. Additional genetic findings demonstrating replicable associations with pulmonary fibrosis include numerous genetic variants in, and adjacent to, genes known to be involved in the regulation of telomere length (e.g., the *TERT* gene, the telomerase RNA component [*TERC*] gene, and the regulator of telomere elongation helicase 1 [*RTEL1*] gene) and surfactant protein genes (e.g., surfactant protein A2 [*SFTPA2*] gene).

Genetic studies have also provided some insights into other forms of ILD. Genome-wide association studies of sarcoidosis have demonstrated numerous variants in genes, and in genomic regions, that are associated with the disease. Some of these disease associated variants in sarcoidosis fall in human leukocyte antigen (*HLA*) regions, in regions of genes involved in immune regulation (e.g., interleukin 12B [*IL12B*]) in regions of genes that are less well understood (butroyrophilin-like 2 [*BTNL2*]) but also appear to be involved in T-cell activation. LAM is often associated with genetic variants in the tuberous sclerosis complex genes (e.g., *TSC1* and *TSC2*), consistent with the known evidence that this disease can occur in isolation but also in patients with known tuberous sclerosis. Many genetic factors for rare diseases such as Hermansky-Pudlak syndrome (a rare autosomal recessive disorder that results in pulmonary fibrosis but also includes oculocutaneous albinism, bleeding diatheses, and horizontal nystagmus) have also been discovered (e.g., *HSP1*, and *HSP3-7*).

GLOBAL CONSIDERATIONS

The prevalence, clinical presentation, and natural history of most ILDs in European countries resemble that described in the United States. However, as expected, there is growing evidence for racial differences in clinical (rate of acute exacerbations) or genetic (*MUC5B*) attributes between Caucasian and Asian populations. To date there are limited data on the prevalence of ILD in Hispanics, subjects of African descent and many other ethnic groups.

ACKNOWLEDGMENT

The authors gratefully acknowledge Talmadge King, Jr. for his contribution in the prior version of this chapter.

FURTHER READING

American Thoracic Society/European respiratory society consensus classification of the idiopathic interstitial pneumonias: Am J Respir Crit Care Med 165:277, 2002. [PubMed: 11790668]
Raghu G et al and ATS/ERS/JRS/ALAT Committee on Idiopathic Pulmonary Fibrosis: An official ATS/ERS/JRS/ALAT statement: Idiopathic pulmonary fibrosis: Evidence-based guidelines for the diagnosis and management. Am J Respir Crit Care Med 183:788, 2011. [PubMed: 21471066]
Travis WD et al: Idiopathic nonspecific interstitial pneumonia: Report of an American Thoracic Society project. Am J Respir Crit Care Med 177:1338, 2008. [PubMed: 18388353]
Travis WD et al: An official American Thoracic Society/European Respiratory Society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. Am J Respir Crit Care Med. 188:733, 2013. [PubMed: 24032382]

Chapter 288: Disorders of the Pleura

Richard W. Light

FIGURE 288-1

PLEURAL EFFUSION

The pleural space lies between the lung and the chest wall and normally contains a very thin layer of fluid, which serves as a coupling system. A pleural effusion is present when there is an excess quantity of fluid in the pleural space.

Etiology

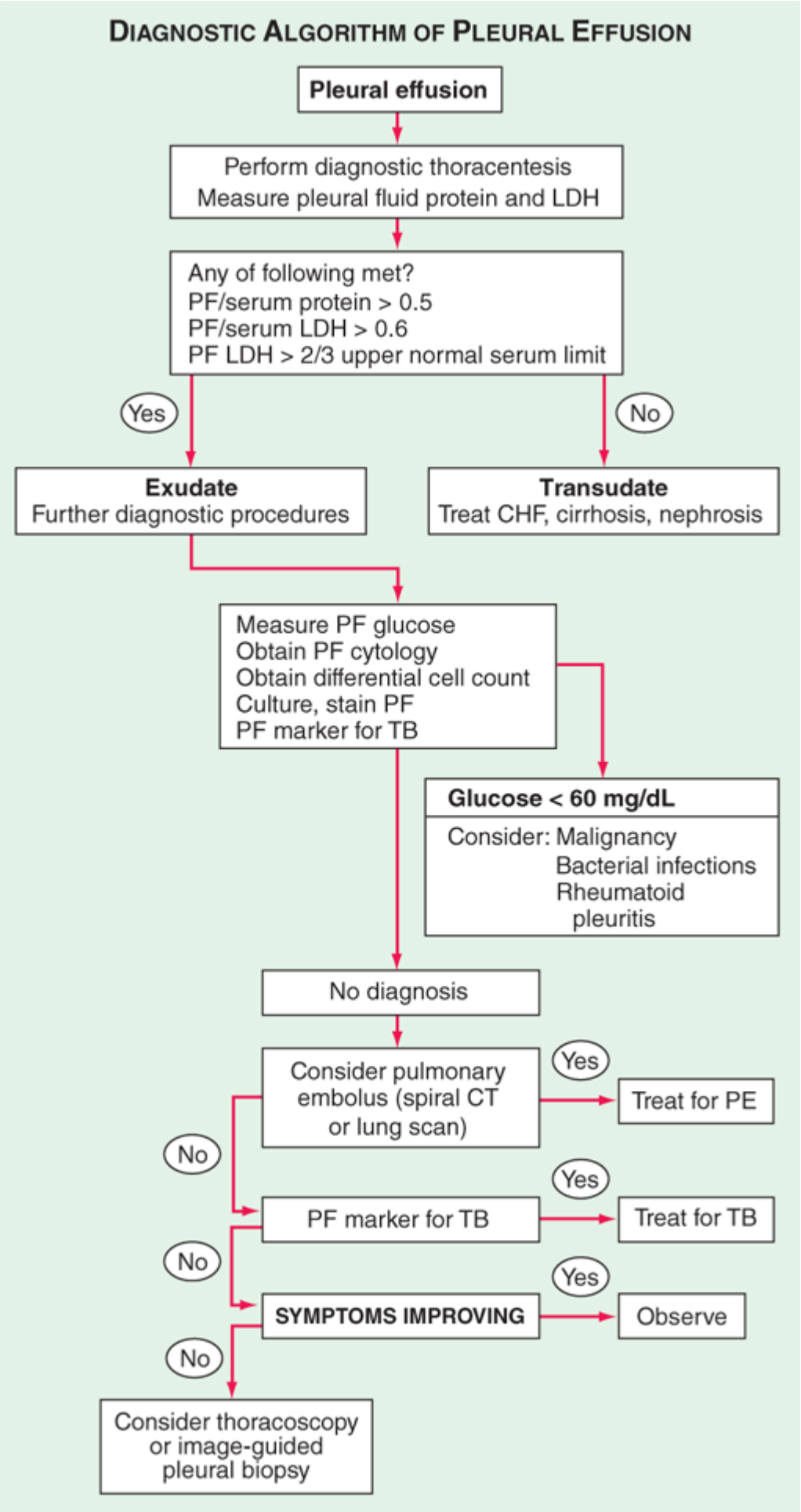
Pleural fluid accumulates when pleural fluid formation exceeds pleural fluid absorption. Normally, fluid enters the pleural space from the capillaries in the parietal pleura and is removed via the lymphatics in the parietal pleura. Fluid also can enter the pleural space from the interstitial spaces of the lung via the visceral pleura or from the peritoneal cavity via small holes in the diaphragm. The lymphatics have the capacity to absorb 20 times more fluid than is formed normally. Accordingly, a pleural effusion may develop when there is excess pleural fluid formation (from the interstitial spaces of the lung, the parietal pleura, or the peritoneal cavity) or when there is decreased fluid removal by the lymphatics.

Diagnostic Approach

Patients suspected of having a pleural effusion should undergo chest imaging to diagnose its extent. Chest ultrasound has replaced the lateral decubitus x-ray in the evaluation of suspected pleural effusions and as a guide to thoracentesis. When a patient is found to have a pleural effusion, an effort should be made to determine the cause **(Fig. 288-1)**. The first step is to determine whether the effusion is a transudate or an exudate. A *transudative pleural effusion* occurs when *systemic factors* that influence the formation and absorption of pleural fluid are altered. The leading causes of transudative pleural effusions in the United States are left ventricular failure and cirrhosis. An *exudative pleural effusion* occurs when *local factors* that influence the formation and absorption of pleural fluid are altered. The leading causes of exudative pleural effusions are bacterial pneumonia, malignancy, viral infection, and pulmonary embolism. The primary reason for making this differentiation is that additional diagnostic procedures are indicated with exudative effusions to define the cause of the local disease.

FIGURE 288-1

Approach to the diagnosis of pleural effusions. CHF, congestive heart failure; CT, computed tomography; LDH, lactate dehydrogenase; PE, pulmonary embolism; PF, pleural fluid; TB, tuberculosis.


DIAGNOSTIC ALGORITHM OF PLEURAL EFFUSION
Pleural effusion
Perform diagnostic thoracentesis
Measure pleural fluid protein and LDH
Any of following met?
PF/serum protein > 0.5
PF/serum LDH > 0.6
PF LDH > 2/3 upper normal serum limit
Yes
No
Exudate
Further diagnostic procedures
Transudate
Treat CHF, cirrhosis, nephrosis
Measure PF glucose
Obtain PF cytology
Obtain differential cell count
Culture, stain PF
PF marker for TB
Glucose < 60 mg/dL
Consider: Malignancy
Bacterial infections
Rheumatoid pleuritis
No diagnosis
Consider pulmonary embolus (spiral CT or lung scan)
Yes
Treat for PE
No
PF marker for TB
Yes
Treat for TB
No
SYMPTOMS IMPROVING
Yes
Observe
No
Consider thoracoscopy or image-guided pleural biopsy

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Transudative and exudative pleural effusions are distinguished by measuring the lactate dehydrogenase (LDH) and protein levels in the pleural fluid. Exudative pleural effusions meet at least one of the following criteria, whereas transudative pleural effusions meet none:

1. Pleural fluid protein/serum protein >0.5
2. Pleural fluid LDH/serum LDH >0.6
3. Pleural fluid LDH more than two-thirds the normal upper limit for serum

These criteria misidentify ~25% of transudates as exudates. If one or more of the exudative criteria are met and the patient is clinically thought to have a condition producing a transudative effusion, the difference between the protein levels in the serum and the pleural fluid should be measured. If this gradient is >31 g/L (3.1 g/dL), the exudative categorization by these criteria can be ignored because almost all such patients have a transudative pleural effusion.

If a patient has an exudative pleural effusion, the following tests on the pleural fluid should be obtained: description of the appearance of the fluid, glucose level, differential cell count, microbiologic studies, and cytology.

Effusion Due to Heart Failure

The most common cause of pleural effusion is left ventricular failure. The effusion occurs because the increased amounts of fluid in the lung interstitial spaces exit in part across the visceral pleura; this overwhelms the capacity of the lymphatics in the parietal pleura to remove fluid. In patients with heart failure, a diagnostic thoracentesis should be performed if the effusions are not bilateral and comparable in size, if the patient is febrile, or if the patient has pleuritic chest pain to verify that the patient has a transudative effusion. Otherwise the patient's heart failure is treated. If the effusion persists despite therapy, a diagnostic thoracentesis should be performed. A pleural fluid N-terminal pro-brain natriuretic peptide (NT-proBNP) >1500 pg/mL is virtually diagnostic that the effusion is secondary to congestive heart failure.

Hepatic Hydrothorax

Pleural effusions occur in ~5% of patients with cirrhosis and ascites. The predominant mechanism is the direct movement of peritoneal fluid through small openings in the diaphragm into the pleural space. The effusion is usually right-sided and frequently is large enough to produce severe dyspnea.

Parapneumonic Effusion

Parapneumonic effusions are associated with bacterial pneumonia, lung abscess, or bronchiectasis and are probably the most common cause of exudative pleural effusion in the United States. *Empyema* refers to a grossly purulent effusion.

Patients with aerobic bacterial pneumonia and pleural effusion present with an acute febrile illness consisting of chest pain, sputum production, and leukocytosis. Patients with anaerobic infections present with a subacute illness with weight loss, a brisk leukocytosis, mild anemia, and a history of some factor that predisposes them to aspiration.

The possibility of a parapneumonic effusion should be considered whenever a patient with bacterial pneumonia is initially evaluated. The presence of free pleural fluid can be demonstrated with a lateral decubitus radiograph, computed tomography (CT) of the chest, or ultrasound. If the free fluid separates the lung from the chest wall by >10 mm, a therapeutic thoracentesis should be performed. Factors indicating the likely need for a procedure more invasive than a thoracentesis (in increasing order of importance) include the following:

1. Loculated pleural fluid
2. Pleural fluid pH <7.20
3. Pleural fluid glucose <3.3 mmol/L (<60 mg/dL)
4. Positive Gram stain or culture of the pleural fluid
5. Presence of gross pus in the pleural space

If the fluid recurs after the initial therapeutic thoracentesis and if any of these characteristics is present, a repeat thoracentesis should be performed. If the fluid cannot be completely removed with the therapeutic thoracentesis, consideration should be given to inserting a chest tube and instilling the combination of a fibrinolytic agent (e.g., tissue plasminogen activator, 10 mg) and deoxyribonuclease (5 mg) or performing a thoracoscopy with the breakdown of adhesions. Decortication should be considered when these measures are ineffective.

Effusion Secondary to Malignancy

Malignant pleural effusions secondary to metastatic disease are the second most common type of exudative pleural effusion. The three tumors that cause ~75% of all malignant pleural effusions are lung carcinoma, breast carcinoma, and lymphoma. Most patients complain of dyspnea, which is frequently out of proportion to the size of the effusion. The pleural fluid is an exudate, and its glucose level may be reduced if the tumor burden in the pleural space is high.

The diagnosis usually is made via cytology of the pleural fluid. If the initial cytologic examination is negative, thoracoscopy is the best next procedure if malignancy is strongly suspected. At the time of thoracoscopy, a procedure such as pleural abrasion should be performed to effect a pleurodesis. An alternative to thoracoscopy is CT- or ultrasound-guided needle biopsy of pleural thickening or nodules. Patients with a malignant pleural effusion are treated symptomatically for the most part, since the presence of the effusion indicates disseminated disease and most malignancies associated with pleural effusion are not curable with chemotherapy. The only symptom that can be attributed to the effusion itself is dyspnea. If the patient’s lifestyle is compromised by dyspnea and if the dyspnea is relieved with a therapeutic thoracentesis, one of the following procedures should be considered: (1) insertion of a small indwelling catheter or (2) tube thoracostomy with the instillation of a sclerosing agent such as doxycycline (500 mg).

Mesothelioma

Malignant mesotheliomas are primary tumors that arise from the mesothelial cells that line the pleural cavities; most are related to asbestos exposure. Patients with mesothelioma present with chest pain and shortness of breath. The chest radiograph reveals a pleural effusion, generalized pleural thickening, and a shrunken hemithorax. The diagnosis is usually established with image-guided needle biopsy or thoracoscopy.

Effusion Secondary to Pulmonary Embolization

The diagnosis most commonly overlooked in the differential diagnosis of a patient with an undiagnosed pleural effusion is pulmonary embolism. Dyspnea is the most common symptom. The pleural fluid is almost always an exudate. The diagnosis is established by spiral CT scan or pulmonary arteriography **(Chap. 273)**. Treatment of a patient with a pleural effusion secondary to pulmonary embolism is the same as it is for any patient with pulmonary emboli. If the pleural effusion increases in size after anticoagulation, the patient probably has recurrent emboli or another complication, such as a hemothorax or a pleural infection.

Tuberculous Pleuritis

In many parts of the world, the most common cause of an exudative pleural effusion is tuberculosis (TB), but tuberculous effusions are relatively uncommon in the United States (**See also Chap. 173**) . Tuberculous pleural effusions usually are associated with primary TB and are thought to be due primarily to a hypersensitivity reaction to tuberculous protein in the pleural space. Patients with tuberculous pleuritis present with fever, weight loss, dyspnea, and/or pleuritic chest pain. The pleural fluid is an exudate with predominantly small lymphocytes. The diagnosis is established by demonstrating high levels of TB markers in the pleural fluid (adenosine deaminase >40 IU/L or interferon γ >140 pg/mL). Alternatively, the diagnosis can be established by culture of the pleural fluid, needle biopsy of the pleura, or thoracoscopy. The recommended treatments of pleural and pulmonary TB are identical **(Chap. 173)**.

Effusion Secondary to Viral Infection

Viral infections are probably responsible for a sizable percentage of undiagnosed exudative pleural effusions. In many series, no diagnosis is established for ~20% of exudative effusions, and these effusions resolve spontaneously with no long-term residua. The importance of these effusions is that one should not be too aggressive in trying to establish a diagnosis for the undiagnosed effusion, particularly if the patient is improving clinically.

Chylothorax

A chylothorax occurs when the thoracic duct is disrupted and chyle accumulates in the pleural space. The most common cause of chylothorax is trauma (most frequently thoracic surgery), but it also may result from tumors in the mediastinum. Patients with chylothorax present with dyspnea, and a large pleural effusion is present on the chest radiograph. Thoracentesis reveals milky fluid, and biochemical analysis reveals a triglyceride level that exceeds 1.2 mmol/L (110 mg/dL). Patients with chylothorax and no obvious trauma should have a lymphangiogram and a mediastinal CT scan to assess the mediastinum for lymph nodes. The treatment of choice for most chylothoraxes is insertion of a chest tube plus the administration of octreotide. If these modalities fail, percutaneous transabdominal thoracic duct blockage effectively controls most chylothoraces. An alternative treatment is ligation of the thoracic duct. Patients with chylothoraxes should not undergo prolonged tube thoracostomy with chest tube drainage because this will lead to malnutrition and immunologic incompetence.

Hemothorax

When a diagnostic thoracentesis reveals bloody pleural fluid, a hematocrit should be obtained on the pleural fluid. If the hematocrit is more than one-half of that in the peripheral blood, the patient is considered to have a hemothorax. Most hemothoraxes are the result of trauma; other causes include rupture of a blood vessel or tumor. Most patients with hemothorax should be treated with tube thoracostomy, which allows continuous quantification of bleeding. If the bleeding emanates from a laceration of the pleura, apposition of the two pleural surfaces is likely to stop the bleeding. If the

pleural hemorrhage exceeds 200 mL/h, consideration should be given to angiographic coil embolization, thoracoscopy or thoracotomy.

Miscellaneous Causes of Pleural Effusion

There are many other causes of pleural effusion **(Table 288-1)**. Key features of some of these conditions are as follows: If the pleural fluid amylase level is elevated, the diagnosis of esophageal rupture or pancreatic disease is likely. If the patient is febrile, has predominantly polymorphonuclear cells in the pleural fluid, and has no pulmonary parenchymal abnormalities, an intraabdominal abscess should be considered.

TABLE 288-1 Differential Diagnoses of Pleural Effusions

Transudative Pleural Effusions

1. Congestive heart failure
2. Cirrhosis
3. Nephrotic syndrome
4. Peritoneal dialysis
5. Superior vena cava obstruction
6. Myxedema
7. Urinothorax

Exudative Pleural Effusions

1. Neoplastic diseases
    a. Metastatic disease
    b. Mesothelioma
2. Infectious diseases
    a. Bacterial infections
    b. Tuberculosis
    c. Fungal infections
    d. Viral infections
    e. Parasitic infections
3. Pulmonary embolization
4. Gastrointestinal disease
    a. Esophageal perforation
    b. Pancreatic disease
    c. Intraabdominal abscesses
    d. Diaphragmatic hernia
    e. After abdominal surgery
    f. Endoscopic variceal sclerotherapy
    g. After liver transplant
5. Collagen vascular diseases
    a. Rheumatoid pleuritis

   b. Systemic lupus erythematosus
   c. Drug-induced lupus
   d. Sjögren syndrome
   e. Granulomatosis with polyangiitis (Wegener)
   f. Churg-Strauss syndrome
6. Post-coronary artery bypass surgery
7. Asbestos exposure
8. Sarcoidosis
9. Uremia
10. Meigs' syndrome
11. Yellow nail syndrome
12. Drug-induced pleural disease
    a. Nitrofurantoin
    b. Dantrolene
    c. Methysergide
    d. Bromocriptine
    e. Procarbazine
    f. Amiodarone
    g. Dasatinib
13. Trapped lung
14. Radiation therapy
15. Post-cardiac injury syndrome
16. Hemothorax
17. Iatrogenic injury
18. Ovarian hyperstimulation syndrome
19. Pericardial disease
20. Chylothorax

The diagnosis of an asbestos pleural effusion is one of exclusions. Benign ovarian tumors can produce ascites and a pleural effusion (Meigs' syndrome), as can the ovarian hyperstimulation

syndrome. Several drugs can cause pleural effusion; the associated fluid is usually eosinophilic. Pleural effusions commonly occur after coronary artery bypass surgery. Effusions occurring within the first weeks are typically left-sided and bloody, with large numbers of eosinophils, and respond to one or two therapeutic thoracenteses. Effusions occurring after the first few weeks are typically left-sided and clear yellow, with predominantly small lymphocytes, and tend to recur. Other medical manipulations that induce pleural effusions include abdominal surgery; radiation therapy; liver, lung, or heart transplantation; and the intravascular insertion of central lines.

PNEUMOTHORAX

Pneumothorax is the presence of gas in the pleural space. A *spontaneous pneumothorax* is one that occurs without antecedent trauma to the thorax. A *primary spontaneous pneumothorax* occurs in the absence of underlying lung disease, whereas a *secondary pneumothorax* occurs in its presence. A *traumatic pneumothorax* results from penetrating or nonpenetrating chest injuries. A *tension pneumothorax* is a pneumothorax in which the pressure in the pleural space is positive throughout the respiratory cycle.

Primary Spontaneous Pneumothorax

Primary spontaneous pneumothoraxes are usually due to rupture of apical pleural blebs, small cystic spaces that lie within or immediately under the visceral pleura. Primary spontaneous pneumothoraxes occur almost exclusively in smokers; this suggests that these patients have subclinical lung disease. Approximately one-half of patients with an initial primary spontaneous pneumothorax will have a recurrence. The initial recommended treatment for primary spontaneous pneumothorax is simple aspiration. If the lung does not expand with aspiration or if the patient has a recurrent pneumothorax, thoracoscopy with stapling of blebs and pleural abrasion is indicated. Thoracoscopy or thoracotomy with pleural abrasion is almost 100% successful in preventing recurrences.

Secondary Pneumothorax

Most secondary pneumothoraxes are due to chronic obstructive pulmonary disease, but pneumothoraxes have been reported with virtually every lung disease. Pneumothorax in patients with lung disease is more life-threatening than it is in normal individuals because of the lack of pulmonary reserve in these patients. Nearly all patients with secondary pneumothorax should be treated with tube thoracostomy. Most should also be treated with thoracoscopy or thoracotomy with the stapling of blebs and pleural abrasion. If the patient is not a good operative candidate or refuses surgery, pleurodesis should be attempted by the intrapleural injection of a sclerosing agent such as doxycycline.

Traumatic Pneumothorax

Traumatic pneumothoraxes can result from both penetrating and nonpenetrating chest trauma. Traumatic pneumothoraxes should be treated with tube thoracostomy unless they are very small. If a hemopneumothorax is present, one chest tube should be placed in the superior part of the hemithorax to evacuate the air and another should be placed in the inferior part of the hemithorax to remove the blood. Iatrogenic pneumothorax is a type of traumatic pneumothorax that is becoming more common. The leading causes are transthoracic needle aspiration, thoracentesis, and the insertion of central intravenous catheters. Most can be managed with supplemental oxygen or aspiration, but if these measures are unsuccessful, a tube thoracostomy should be performed.

Tension Pneumothorax

This condition usually occurs during mechanical ventilation or resuscitative efforts. The positive pleural pressure is life-threatening both because ventilation is severely compromised and because the positive pressure is transmitted to the mediastinum, resulting in decreased venous return to the heart and reduced cardiac output.

Difficulty in ventilation during resuscitation or high peak inspiratory pressures during mechanical ventilation strongly suggest the diagnosis. The diagnosis is made by physical examination showing an enlarged hemithorax with no breath sounds, hyperresonance to percussion, and shift of the mediastinum to the contralateral side. Tension pneumothorax must be treated as a medical emergency. If the tension in the pleural space is not relieved, the patient is likely to die from inadequate cardiac output or marked hypoxemia. A large-bore needle should be inserted into the pleural space through the second anterior intercostal space. If large amounts of gas escape from the needle after insertion, the diagnosis is confirmed. The needle should be left in place until a thoracostomy tube can be inserted.

FURTHER READING

Bhatnagar R et al: Advanced medical interventions in pleural disease. Eur Respir Rev 25:199, 2016. [PubMed: 27246597]
Light RW: *Pleural Diseases*, 6th ed. Lippincott, Williams and Wilkins, Baltimore, 2013.
Rahman NM et al: Intrapleural use of tissue plasminogen activator and DNase in pleural infection. N Engl J Med 365:518, 2011. [PubMed: 21830966]

Chapter 289: Disorders of the Mediastinum

Richard W. Light

INTRODUCTION

The mediastinum is the region between the pleural sacs. It is separated into three compartments **(Table 289-1)**. The *anterior mediastinum* extends from the sternum anteriorly to the pericardium and brachiocephalic vessels posteriorly. It contains the thymus gland, the anterior mediastinal lymph nodes, and the internal mammary arteries and veins. The *middle mediastinum* lies between the anterior and posterior mediastina and contains the heart; the ascending and transverse arches of the aorta; the venae cavae; the brachiocephalic arteries and veins; the phrenic nerves; the trachea, the main bronchi, and their contiguous lymph nodes; and the pulmonary arteries and veins. The *posterior mediastinum* is bounded by the pericardium and trachea anteriorly and the vertebral column posteriorly. It contains the descending thoracic aorta, the esophagus, the thoracic duct, the azygos and hemiazygos veins, and the posterior group of mediastinal lymph nodes.

TABLE 289-1 The Three Compartments of the Mediastinum

	Anterior Compartment	Middle Compartment	Posterior Compartment
Anatomical boundaries	Manubrium and sternum anteriorly, pericardium, aorta, and brachiocephalic vessels posteriorly	Anterior mediastinum anteriorly, posterior mediastinum posteriorly	Pericardium and trachea anteriorly; vertebral column posteriorly
Contents	Thymus gland, anterior mediastinal lymph nodes, internal mammary arteries, and veins	Pericardium, heart, ascending and transverse arch of aorta, superior and inferior vena cavae, brachiocephalic arteries and veins, phrenic nerves, trachea, and main bronchi and their contiguous lymph nodes, pulmonary arteries, and veins	Descending thoracic aorta, esophagus, thoracic duct, azygos and hemiazygos veins, sympathetic chains, and the posterior group of mediastinal lymph nodes
Common abnormalities	Thymoma, lymphomas, teratomatous neoplasms, thyroid masses, parathyroid masses, mesenchymal tumors, giant lymph node hyperplasia, hernia through foramen of Morgagni	Metastatic lymph node enlargement, granulomatous lymph node enlargement, pleuropericardial cysts, bronchogenic cysts, masses of vascular origin	Neurogenic tumors, meningocele, meningomyelocele, gastroenteric cysts, esophageal diverticula, hernia through foramen of Bochdalek, extramedullary hematopoiesis

MEDIASTINAL MASSES

The first step in evaluating a mediastinal mass is to place it in one of the three mediastinal compartments, since each has different characteristic lesions (Table 289-1).

Computed tomography (CT) scanning is the most valuable imaging technique for evaluating mediastinal masses and is the only imaging technique that should be done in most instances. Barium studies of the gastrointestinal tract are indicated in many patients with posterior mediastinal lesions, because hernias, diverticula, and achalasia are readily diagnosed in this manner. An iodine-131 scan can efficiently establish the diagnosis of intrathoracic goiter.

A definite diagnosis can be obtained with mediastinoscopy or anterior mediastinotomy in many patients with masses in the anterior or middle mediastinal compartments. A diagnosis can be established without thoracotomy via percutaneous fine-needle aspiration biopsy or endoscopic transesophageal or endobronchial ultrasound-guided biopsy of mediastinal masses in most cases. An alternative way to establish the diagnosis is video-assisted thoracoscopy. In many cases, the diagnosis can be established and the mediastinal mass removed with video-assisted thoracoscopy.

ACUTE MEDIASTINITIS

Cases of acute mediastinitis are usually due to esophageal perforation, occur after median sternotomy for cardiac surgery, or are infections descending from the neck, oral cavity, or facial area. Patients with esophageal rupture are acutely ill with chest pain and dyspnea due to the mediastinal infection. The esophageal rupture can occur spontaneously or as a complication of esophagoscopy or the insertion of a Blakemore tube. Appropriate treatment consists of exploration of the mediastinum with primary repair of the esophageal tear and drainage of the pleural space and the mediastinum.

The incidence of mediastinitis after median sternotomy is 0.4–5.0%. Patients most commonly present with wound drainage. Other presentations include sepsis and a widened mediastinum. The diagnosis usually is established with mediastinal needle aspiration. Treatment includes immediate drainage, debridement, and parenteral antibiotic therapy, but the mortality rate still exceeds 20%.

CHRONIC MEDIASTINITIS

The spectrum of chronic mediastinitis ranges from granulomatous inflammation of the lymph nodes in the mediastinum to fibrosing mediastinitis. Most cases are due to histoplasmosis or tuberculosis, but sarcoidosis, silicosis, and other fungal diseases are at times causative. Patients with granulomatous mediastinitis are usually asymptomatic. Those with fibrosing mediastinitis usually have signs of compression of a mediastinal structure such as the superior vena cava or large airways, phrenic or recurrent laryngeal nerve paralysis, or obstruction of the pulmonary artery or proximal pulmonary veins. If veins or arteries are involved, the placement of stents has relieved the symptoms in many patients.

PNEUMOMEDIASTINUM

In this condition, there is gas in the interstices of the mediastinum. The three main causes are (1) alveolar rupture with dissection of air into the mediastinum; (2) perforation or rupture of the esophagus, trachea, or main bronchi; and (3) dissection of air from the neck or the abdomen into the mediastinum. Typically, there is severe substernal chest pain with or without radiation into the neck and arms. The physical examination usually reveals subcutaneous emphysema in the suprasternal notch and *Hamman's sign*, which is a crunching or clicking noise synchronous with the heartbeat and is best heard in the left lateral decubitus position. The diagnosis is confirmed with the chest radiograph. Usually no treatment is required, but the mediastinal air will be absorbed faster if the

patient inspires high concentrations of oxygen. If mediastinal structures are compressed, the compression can be relieved with needle aspiration.

FURTHER READING

Jabłoński S et al: Acute mediastinitis: evaluation of clinical risk factors for death in surgically treated patients. ANZ J Surg 83:657, 2013. [PubMed: 22989305]
Ponamgi SP et al: Catheter-based intervention for pulmonary vein stenosis due to fibrosing mediastinitis: The Mayo Clinic experience. Int J Cardiol Heart Vasc 8:103, 2015. [PubMed: 26925456]

Chapter 290: Disorders of Ventilation

John F. McConville; Julian Solway; Babak Mokhlesi

FIGURE 290-1

DEFINITION AND PHYSIOLOGY

In health the arterial level of carbon dioxide ($Paco_2$) is maintained between 37 and 43 mmHg at sea level. All disorders of ventilation result in abnormal measurements of $Paco_2$. This chapter reviews chronic ventilatory disorders.

The continuous production of CO_2 by cellular metabolism necessitates its efficient elimination by the respiratory system. The relationship between CO_2 production and $Paco_2$ is described by the equation: $Paco_2 = (k)\ (\dot{V}co_2)/\dot{V}A$, where $\dot{V}co_2$ represents the carbon dioxide production, k is a constant and $\dot{V}A$ is fresh gas alveolar ventilation (see **Chap. 279**). $\dot{V}A$ can be calculated as minute ventilation × (1 – Vd/Vt), where the dead space fraction Vd/Vt represents the portion of a tidal breath that remains within the conducting airways at the conclusion of inspiration and so does not contribute to alveolar ventilation. As such, all disturbances of $Paco_2$ must reflect altered CO_2 production, minute ventilation, or dead space fraction.

Diseases that alter $\dot{V}co_2$ are often acute (sepsis, burns, or pyrexia, for example) and their contribution to ventilatory abnormalities and/or respiratory failure is reviewed elsewhere. Chronic ventilatory disorders typically involve inappropriate levels of minute ventilation or increased dead space fraction. Characterization of these disorders requires a review of the normal respiratory cycle.

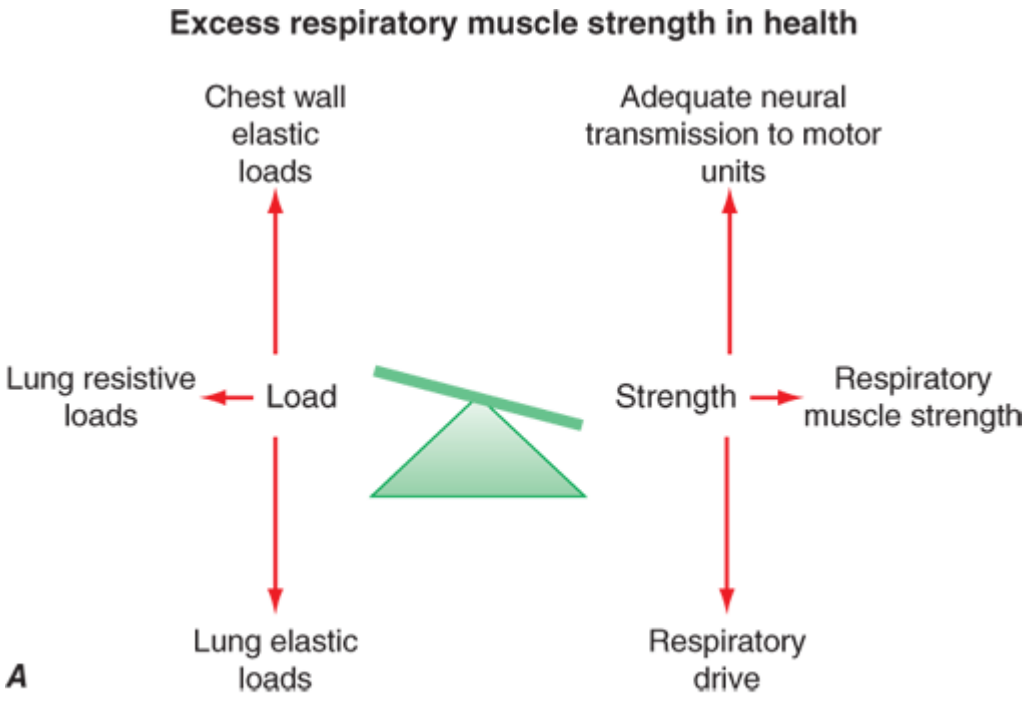
The spontaneous cycle of inspiration and expiration is automatically generated in the brainstem. Two groups of neurons located within the medulla are particularly important: the dorsal respiratory group (DRG) and the ventral respiratory column (VRC). These neurons have widespread projections including the descending projections into the contralateral spinal cord where they perform many functions. They initiate activity in the phrenic nerve/diaphragm, project to the upper airway muscle groups and spinal respiratory neurons, and innervate the intercostal and abdominal muscles that participate in normal respiration. The DRG acts as the initial integration site for many of the afferent nerves relaying information about Pao_2, $Paco_2$, pH, and blood pressure from the carotid and aortic chemoreceptors and baroreceptors to the central nervous system (CNS). In addition, the vagus nerve relays information from stretch receptors and juxtapulmonary-capillary receptors in the lung parenchyma and chest wall to the DRG. The respiratory rhythm is generated within the VRC as well as the more rostrally located parafacial respiratory group (pFRG), which is particularly important for the generation of active expiration. One particularly important area within the VRC is the so called pre-Bötzinger complex. This area is responsible for the generation of various forms of inspiratory activity, and lesioning of the pre-Bötzinger complex leads to the complete cessation of breathing. The neural output of these medullary respiratory networks can be voluntarily suppressed or augmented by input

from higher brain centers and the autonomic nervous system. During normal sleep there is an attenuated response to hypercapnia and hypoxemia resulting in mild nocturnal hypoventilation that corrects upon awakening.

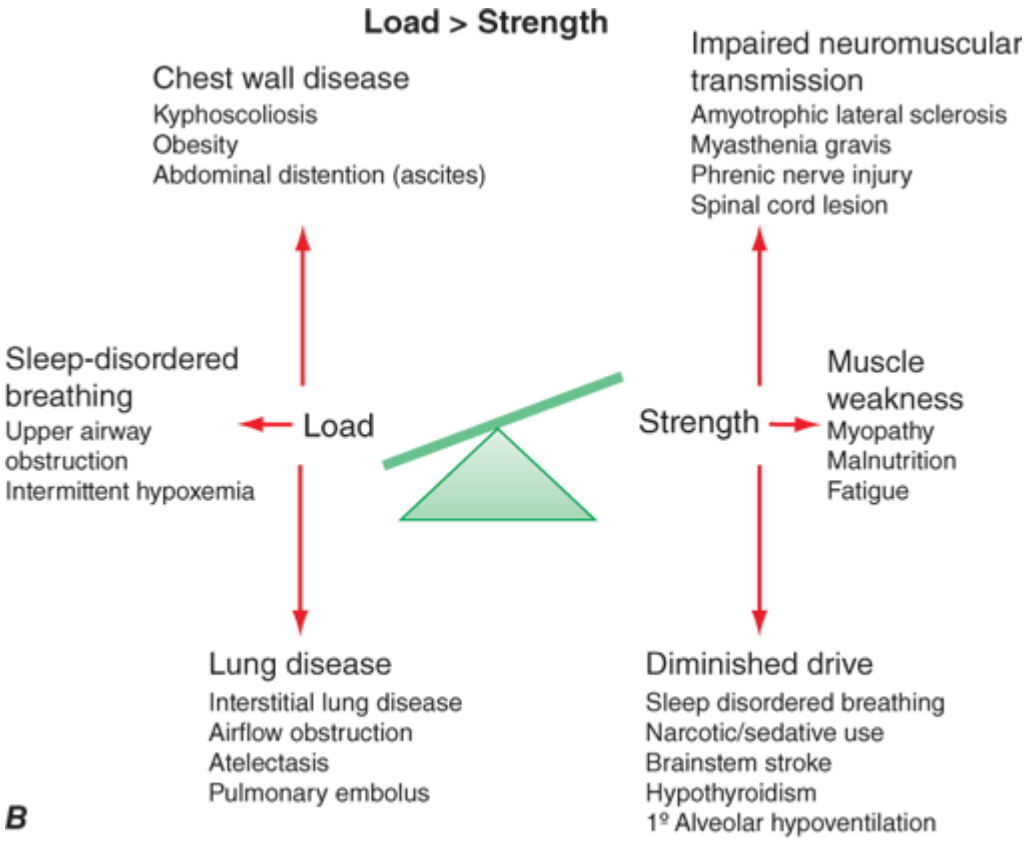
Once neural input has been delivered to the respiratory pump muscles, normal gas exchange requires an adequate amount of respiratory muscle strength to overcome the elastic and resistive loads of the respiratory system **(Fig. 290-1A)** **(also see Chap. 279)**. In health, the strength of the respiratory muscles readily accomplishes this and normal respiration continues indefinitely. Reduction in respiratory drive or neuromuscular competence or substantial increase in respiratory load can diminish minute ventilation, resulting in hypercapnia **(Fig. 290-1B)**. Alternatively, if normal respiratory muscle strength is coupled with excessive respiratory drive, then alveolar hyperventilation ensues and leads to hypocapnia **(Fig. 290-1C)**.

FIGURE 290-1

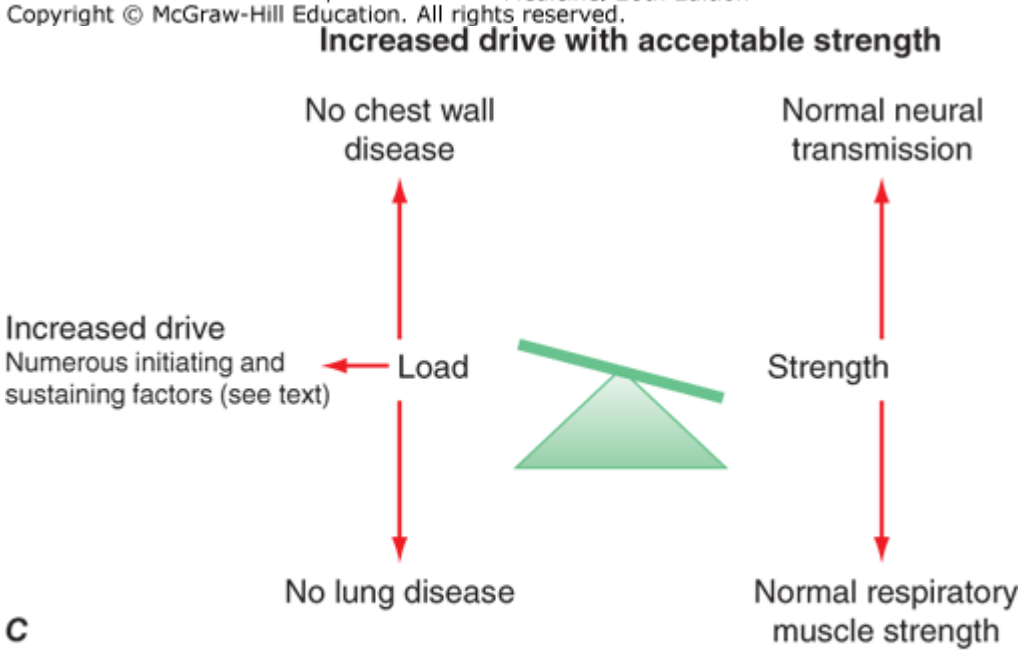
Examples of balance between respiratory system strength and load. ***A.*** Excess respiratory muscle strength in health. ***B.*** Load greater than strength. ***C.*** Increased drive with acceptable strength.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HYPOVENTILATION

CLINICAL FEATURES

Diseases that reduce minute ventilation or increase dead space fall into four major categories: parenchymal lung and chest wall disease, sleep disordered breathing, neuromuscular disease, and respiratory drive disorders (Fig. 290-1B). The clinical manifestations of hypoventilation syndromes are nonspecific **(Table 290-1)** and vary depending on the severity of hypoventilation, the rate at which hypercapnia develops, the degree of compensation for respiratory acidosis, and the underlying disorder. Patients with parenchymal lung or chest wall disease typically present with shortness of breath and diminished exercise tolerance. Episodes of increased dyspnea and sputum production are hallmarks of obstructive lung diseases such as COPD, whereas progressive dyspnea and cough are common in interstitial lung diseases. Excessive daytime somnolence, poor-quality sleep and snoring are common among patients with sleep-disordered breathing. Sleep disturbance and orthopnea are also described in neuromuscular disorders. As neuromuscular weakness progresses, the respiratory muscles, including the diaphragm, are placed at a mechanical disadvantage in the supine position due to the upward movement of the abdominal contents. New onset orthopnea is frequently a sign of reduced respiratory muscle force generation. More commonly however, extremity weakness or bulbar symptoms develop prior to sleep disturbance in neuromuscular diseases such as amyotrophic lateral sclerosis (ALS) or muscular dystrophy. Patients with respiratory drive disorders do not have symptoms distinguishable from other causes of chronic hypoventilation.

TABLE 290-1 Signs and Symptoms of Hypoventilation

Dyspnea during activities of daily living
Orthopnea in diseases affecting diaphragm function
Poor-quality sleep
Daytime hypersomnolence
Early morning headaches
Anxiety
Impaired cough in neuromuscular diseases

The clinical course of patients with chronic hypoventilation from neuromuscular or chest wall disease follows a characteristic sequence: An asymptomatic stage where daytime Pao_2 and $Paco_2$ are normal followed by nocturnal hypoventilation, initially during rapid eye movement (REM) sleep and later in non-REM sleep. Finally, if vital capacity drops further, daytime hypercapnia develops. Symptoms can develop at any point along this time-course and often depend on the pace of respiratory muscle functional decline. Regardless of cause, the hallmark of all alveolar hypoventilation syndromes is an increase in alveolar Pco_2 ($PAco_2$) and therefore in $Paco_2$. The resulting respiratory acidosis eventually leads to a compensatory increase in plasma bicarbonate concentration. The increase in $Paco_2$ results in an obligatory decrease in PAo_2, often resulting in hypoxemia. If severe, the hypoxemia manifests clinically as cyanosis and can stimulate erythropoiesis and so induce secondary erythrocytosis. The combination of chronic hypoxemia and hypercapnia may also induce pulmonary vasoconstriction, leading eventually to pulmonary hypertension, right ventricular hypertrophy, and right heart failure.

DIAGNOSIS

Elevated plasma bicarbonate in the absence of volume depletion is suggestive of hypoventilation. An arterial blood gas demonstrating elevated $Paco_2$ with a normal pH confirms chronic alveolar hypoventilation. The subsequent evaluation to identify an etiology should initially focus on whether the patient has lung disease or chest wall abnormalities. Physical examination, imaging studies (chest x-ray and/or CT scan) and pulmonary function tests are sufficient to identify most lung/chest wall

disorders leading to hypercapnia. If these evaluations are unrevealing then the clinician should screen for obesity hypoventilation syndrome (OHS), the most frequent sleep disorder leading to chronic hypoventilation, which is typically accompanied by obstructive sleep apnea (OSA). Several screening tools have been developed to identify patients at risk for OSA. The Berlin Questionnaire has been validated in a primary care setting and identifies patients likely to have OSA. The Epworth Sleepiness Scale (ESS) and the STOP-Bang questionnaires have not been validated in outpatient primary care settings but are quick and easy to use. The ESS measures daytime sleepiness, with a score of ≥10 identifying individuals who warrant additional investigation. The STOP-Bang survey has been used in preoperative anesthesia clinics to identify patients at risk of having OSA. In this population it has 93% sensitivity and 90% negative predictive value. Additionally, the STOP-Bang questionnaire has been validated as a screening tool for OSA in sleep and surgical clinics. The probability of moderate and severe OSA steadily increases with higher STOP-Bang scores.

If the ventilatory apparatus (lung, airways, chest wall) is not responsible for chronic hypercapnia then the focus should shift to respiratory drive and neuromuscular disorders. There is an attenuated increase in minute ventilation in response to elevated CO_2 and/or low O_2 in respiratory drive disorders. These diseases are difficult to diagnose and should be suspected when patients with hypercapnia are found to have normal respiratory muscle strength, normal pulmonary function, and normal alveolar-arterial Po_2 difference. Hypoventilation is more marked during sleep in patients with respiratory drive defects and polysomnography often reveals central apneas, hypopneas, or hypoventilation. Brain imaging (CT scan or MRI) can sometimes identify structural abnormalities in the pons or medulla that result in hypoventilation. Chronic narcotic use or significant hypothyroidism can depress the central respiratory drive and lead to chronic hypercapnia as well.

Respiratory muscle weakness has to be profound before lung volumes are compromised and hypercapnia develops. Typically physical examination reveals decreased strength in major muscle groups prior to the development of hypercapnia. Measurement of maximum inspiratory and expiratory pressures or forced vital capacity (FVC) can be used to monitor for respiratory muscle involvement in diseases with progressive muscle weakness. These patients also have increased risk for sleep-disordered breathing, including hypopneas, central and obstructive apneas, and hypoxemia. Nighttime oximetry and capnometry during polysomnography are helpful in better characterizing sleep disturbances in this patient population.

TREATMENT

TREATMENT Hypoventilation

Nocturnal non-invasive positive-pressure ventilation (NIPPV) has been used successfully in the treatment of hypoventilation and apneas, both central and obstructive, in patients with neuromuscular and chest wall disorders. Nocturnal NIPPV has been shown to improve daytime hypercapnia, prolong survival, and improve health-related quality of life when daytime hypercapnia is documented. ALS guidelines recommend consideration of nocturnal NIPPV if symptoms of hypoventilation exist and one of the following criteria is present: $Paco_2$ ≥45 mmHg; nocturnal oximetry demonstrates oxygen saturation ≤88% for 5 consecutive min; maximal inspiratory pressure <60 cmH_2O; FVC <50% predicted; sniff nasal pressure <40 cmH_2O. However, at present there is inconclusive evidence to support pre-emptive nocturnal NIPPV use in all patients with neuromuscular and chest wall disorders who demonstrate nocturnal but not daytime hypercapnia. Nevertheless, at some point, the institution of full-time ventilatory support with either pressure or volume-preset modes is required in progressive neuromuscular disorders. There is less evidence to direct the timing of this decision, but ventilatory failure requiring mechanical ventilation and chest infections related to ineffective cough are frequent triggers for the institution of full-time ventilatory support.

Treatment of chronic hypoventilation from lung or neuromuscular diseases should be directed at the underlying disorder. Pharmacologic agents that stimulate respiration, such as medroxyprogesterone and acetazolamide, have been poorly studied in chronic hypoventilation and should not replace treatment of the underlying disease process. Regardless of the cause, excessive metabolic alkalosis should be corrected, as plasma bicarbonate levels elevated out of proportion for the degree of chronic respiratory acidosis can result in additional hypoventilation. When indicated, administration of supplemental oxygen is effective in attenuating hypoxemia, polycythemia, and pulmonary hypertension. However, in some patients supplemental oxygen can worsen hypercapnia.

Phrenic nerve or diaphragm pacing is a potential therapy for patients with hypoventilation from high cervical spinal cord lesions or respiratory drive disorders. Prior to surgical implantation patients should have nerve conduction studies to ensure normal bilateral phrenic nerve function. Small case series suggest that effective diaphragmatic pacing can improve quality of life in these patients.

HYPOVENTILATION SYNDROMES

OBESITY HYPOVENTILATION SYNDROME

The diagnosis of OHS requires: BMI ≥30 kg/m^2 and chronic daytime alveolar hypoventilation, defined as $Paco_2$ ≥45 mmHg at sea level in the absence of other known causes of hypercapnia. In almost 90% of cases the sleep disordered breathing is in the form of OSA. Several international studies in different populations confirm that the overall prevalence of OSA syndrome, defined by an apnea hypopnea index ≥5 AND daytime sleepiness, is ~3–4% in middle-aged men and 2% in middle-aged women. Thus, the population at risk for the development of OHS continues to rise as the world-wide obesity epidemic persists. Although no population-based prevalence studies of OHS have been performed, some estimates suggest there may be as many as 500,000 individuals with OHS in the United States.

Some, but not all, studies suggest that severe obesity (BMI >40 kg/m^2) and severe OSA (AHI >30 events per h) are risk factors for the development of OHS. The pathogenesis of hypoventilation in these patients is the result of multiple physiologic variables and conditions including OSA, increased work of breathing, respiratory muscle impairment, ventilation-perfusion mismatching, and depressed central ventilatory responsiveness to hypoxemia and hypercapnia. These defects in central respiratory drive often improve with treatment which suggest that decreased ventilatory responsiveness is a consequence rather than a primary cause of OHS. The treatment of OHS is similar to that for OSA: weight reduction and nocturnal non-invasive positive pressure ventilation (NIPPV). There is evidence that weight loss alone lowers $Paco_2$ in patients with OHS. However, treatment with NIPPV should never be delayed while the patient attempts to lose weight. Continuous positive airway pressure (CPAP) improves daytime hypercapnia and hypoxemia in more than half of patients with OHS and concomitant OSA. Bi-level positive airway pressure should be reserved for patients not able to tolerate high levels of CPAP support or patients that remain hypoxemic despite resolution of obstructive respiratory events. NIPPV with bi-level PAP should be strongly considered if hypercapnia persists after several weeks of CPAP therapy with objectively proven adherence. Patients with OHS and no evidence of OSA are typically started on bi-level positive airway pressure, as are patients presenting with acute decompensated OHS. Finally, comorbid conditions that impair ventilation, such as chronic obstructive pulmonary disease, should be aggressively treated in conjunction with co-existing OHS.

CENTRAL HYPOVENTILATION SYNDROME

This syndrome can present later in life or in the neonatal period where it is often called Ondine's curse or congenital central hypoventilation syndrome (CCHS). Abnormalities in the gene encoding PHOX2b, a transcription factor with a role in neuronal development, have been implicated in the pathogenesis of

CCHS. Regardless of the age of onset, these patients have absent respiratory response to hypoxia or hypercapnia, mildly elevated $Paco_2$ while awake, and markedly elevated $Paco_2$ during non-REM sleep. Interestingly these patients are able to augment their ventilation and "normalize" $Paco_2$ during exercise and during REM sleep. These patients typically require NIPPV or mechanical ventilation as therapy and should be considered for phrenic nerve or diaphragmatic pacing at centers with experience performing these procedures.

HYPERVENTILATION

CLINICAL FEATURES

Hyperventilation is defined as ventilation in excess of metabolic requirements (CO_2 production) leading to a reduction in $Paco_2$. The physiology of patients with chronic hyperventilation is poorly understood and there is no typical clinical presentation. Symptoms can include dyspnea, paresthesias, tetany, headache, dizziness, visual disturbances, and atypical chest pain. Because symptoms can be so diverse, patients with chronic hyperventilation present to a variety of health care providers, including internists, neurologists, psychologists, psychiatrists, and pulmonologists.

It is helpful to think of hyperventilation as having initiating and sustaining factors. Some investigators believe that an initial event leads to increased alveolar ventilation and a drop in $Paco_2$ to ~20 mmHg. The ensuing onset of chest pain, breathlessness, paresthesia, or altered consciousness can be alarming. The resultant increase in minute volume to relieve these acute symptoms only serves to exacerbate symptoms that are often misattributed by the patient and health care workers to cardiopulmonary disorders. An unrevealing evaluation for causes of these symptoms often results in patients being anxious and fearful of additional attacks. It is important to note that **anxiety disorders and panic attacks are NOT synonymous with hyperventilation**. Anxiety disorders can be both an initiating and sustaining factor in the pathogenesis of chronic hyperventilation, but these are not necessary for the development of chronic hypocapnia.

DIAGNOSIS

Respiratory symptoms associated with acute hyperventilation can be the initial manifestation of systemic illnesses such as diabetic ketoacidosis. Causes of acute hyperventilation need to be excluded before a diagnosis of chronic hyperventilation is considered. Arterial blood gas sampling that demonstrates a compensated respiratory alkalosis with a near normal pH, low $Paco_2$, and low calculated bicarbonate are necessary to confirm chronic hyperventilation. Other causes of respiratory alkalosis, such as mild asthma, need to be diagnosed and treated before chronic hyperventilation can be considered. A high index of suspicion is required as increased minute ventilation can be difficult to detect on physical examination. Once chronic hyperventilation is established, a sustained 10% increase in alveolar ventilation is enough to perpetuate hypocapnia. This increase can be accomplished with subtle changes in the respiratory pattern, such as occasional sigh breaths or yawning 2–3 times per min.

TREATMENT

TREATMENT Hyperventilation

There are few well-controlled treatment studies of chronic hyperventilation owing to its diverse features and the lack of a universally accepted diagnostic process. Clinicians often spend considerable time identifying initiating factors, excluding alternative diagnoses, and discussing the patient's concerns and fears. In some patients, reassurance and frank discussion about

hyperventilation can be liberating. Identifying and eliminating habits that perpetuate hypocapnia, such as frequent yawning or sigh breathing, can be helpful. Some evidence suggests that breathing exercises and diaphragmatic retraining may be beneficial for some patients. The evidence for using medications to treat hyperventilation is scant. Beta-blockers may be helpful in patients with sympathetically mediated symptoms such as palpitations and tremors.

ACKNOWLEDGMENT

We acknowledge Jan-Marino Ramirez for his careful critique and helpful suggestions.

FURTHER READING

Anderson PM et al: EFNS guidelines on the clinical management if amyotrophic lateral sclerosis-revised report of the EFNS task force. Eur J Neurol 19:360, 2012. [PubMed: 21914052]
Chung F et al: STOP Questionnaire: A tool to screen patients for obstructive sleep apnea. Anesthesiology 108:812, 2008. [PubMed: 18431116]
Douglas IS: Acute-on-chronic respiratory failure, in *Principles of Critical Care*, 4th ed. Hall JB, Schmidt GS, Kress JP (eds). New York, McGraw-Hill, 2015, pp 482–495.
Gardner WN: The pathophysiology of hyperventilation disorders. Chest 109:516, 1996. [PubMed: 8620731]
Mokhlesi B: Obesity hypoventilation syndrome: A state-of-the-art review. Respir Care 55:1347. Discussion 1363, 2010. [PubMed: 20875161]
Nagappa M et al: Validation of the STOP-Bang questionnaire as a screening tool for obstructive sleep apnea among different populations: A systematic review and meta-analysis. PLoS One 10:e0143697, 2015. [PubMed: 26658438]
Piper AJ, Grunstein RR: Obesity hypoventilation syndrome, mechanisms and management. Am J Respir Crit Care Med 183:292, 2011. [PubMed: 21037018]
Simonds AK: Recent advances in respiratory care for neuromuscular disease. Chest 130:1879, 2006. [PubMed: 17167012]

Chapter 291: Sleep Apnea

Andrew Wellman; Susan Redline

INTRODUCTION

Obstructive sleep apnea/hypopnea syndrome (OSAHS) and central sleep apnea (CSA) are both classified as sleep-related breathing disorders. OSAHS and CSA share some risk factors and physiological bases but also have unique features. Each disorder is associated with impaired ventilation during sleep and disruption of sleep, and each diagnosis requires careful elicitation of the patient's history, physical examination, and physiological testing. OSAHS, the more common disorder, causes daytime sleepiness, impairs daily function, and is a major contributor to cardiovascular disease in adults and to behavioral problems in children. CSA is less common and may occur in combination with obstructive sleep apnea, as a primary condition, or secondary to a medical condition (such as heart failure) or medication. Patients with CSA often report frequent awakenings and daytime fatigue and are at increased risk for heart failure and atrial fibrillation.

OBSTRUCTIVE SLEEP APNEA/HYPOPNEA SYNDROME

Definition

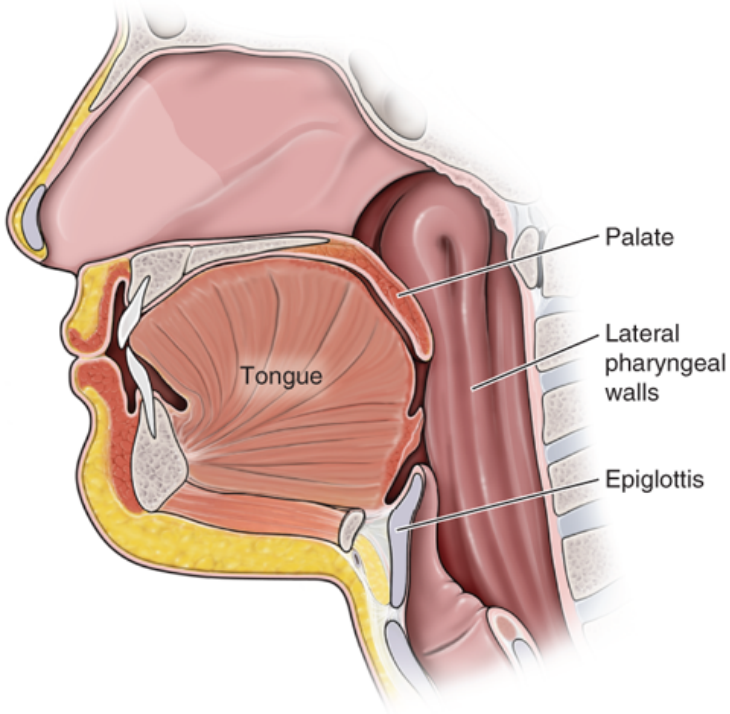
OSAHS is defined on the basis of nocturnal and daytime symptoms as well as sleep study findings. Diagnosis requires the patient to have (1) either symptoms of nocturnal breathing disturbances (snoring, snorting, gasping, or breathing pauses during sleep) or daytime sleepiness or fatigue that occurs despite sufficient opportunities to sleep and is unexplained by other medical problems; and (2) five or more episodes of obstructive apnea or hypopnea per hour of sleep (the apnea-hypopnea index [AHI], calculated as the number of episodes divided by the number of hours of sleep) documented during a sleep study. OSAHS also may be diagnosed in the absence of symptoms if the AHI is >15 episodes/h. Each episode of apnea or hypopnea represents a reduction in breathing for at least 10 s and commonly results in a ≥3% drop in oxygen saturation and/or a brain cortical arousal. OSAHS severity is based on the frequency of breathing disturbances (AHI), the amount of oxyhemoglobin desaturation with respiratory events, the duration of apneas and hypopneas, the degree of sleep fragmentation, and the level of daytime sleepiness or functional impairment.

Pathophysiology

During inspiration, intraluminal pharyngeal pressure becomes increasingly negative, creating a "suctioning" force. Because the pharyngeal airway has no bone or cartilage, airway patency is dependent on the stabilizing influence of the pharyngeal dilator muscles. Although these muscles are continuously activated during wakefulness, neuromuscular output declines with sleep onset. In patients with a collapsible airway, the reduction in neuromuscular output results in transient episodes of pharyngeal collapse (manifesting as an "apnea") or near collapse (manifesting as a "hypopnea"). The episodes of collapse are terminated when ventilatory reflexes are activated and cause arousal, thus stimulating an increase in neuromuscular activity and opening of the airway. The airway may collapse at different sites, such as the soft palate (most common), tongue base, lateral pharyngeal walls, and/or epiglottis **(Fig. 291-1)**. OSAHS may be most severe during REM (rapid eye movement) sleep, when neuromuscular output to the skeletal muscles is particularly low, and in the supine position due to gravitational forces.

FIGURE 291-1

The structures causing airway collapse in OSAHS include the palate, the tongue, and/or the epiglottis. In addition, collapse can also occur due to the lateral pharyngeal walls.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Individuals with a small pharyngeal lumen require relatively high levels of neuromuscular innervation to maintain patency during wakefulness and thus are predisposed to excessive airway collapsibility during sleep. The airway lumen may be narrowed with enlargement of soft tissue structures (tongue, palate, and uvula) due to fat deposition, increased lymphoid tissue, or genetic variation. Craniofacial factors such as mandibular retroposition or micrognathia, reflecting genetic variation or developmental influences, also can reduce lumen dimensions. In addition, lung volumes influence the caudal traction on the pharynx and consequently the stiffness of the pharyngeal wall. Accordingly, low lung volume in the recumbent position, which is particularly pronounced in the obese, contributes to collapse. A high degree of nasal resistance (e.g., due to nasal septal deviation or polyps) can contribute to airway collapse by increasing the negative intraluminal suction pressure. High-level nasal resistance also may trigger mouth opening during sleep, which breaks the seal between the tongue and the teeth and allows the tongue to fall posteriorly and occlude the airway.

Pharyngeal muscle activation is integrally linked to ventilatory drive. Thus, factors related to ventilatory control, particularly ventilatory sensitivity, arousal threshold, and neuromuscular responses to CO_2, contribute to the pathogenesis of OSAHS. A buildup in CO_2 during sleep activates both the diaphragm and the pharyngeal muscles, which stiffen the upper airway and can counteract inspiratory suction pressures and maintain airway patency to an extent that depends on the anatomic predisposition to collapse. However, pharyngeal collapse can occur when the ventilatory control system is overly sensitive to CO_2, with resultant wide fluctuations in ventilation and ventilatory drive and in upper airway instability. Moreover, increasing levels of CO_2 during sleep result in central nervous system arousal, causing the individual to move from a deeper to a lighter level of sleep or to awaken. A low arousal threshold (i.e., awaken to a low level of CO_2 or ventilatory drive) can preempt the CO_2-mediated process of pharyngeal muscle compensation and prevent airway stabilization. A high arousal threshold, conversely, may prevent appropriate termination of apneas, prolonging apnea duration, and exacerbating oxyhemoglobin desaturation severity. Finally, any impairment in the ability of the muscles to compensate during sleep can contribute to collapse of the pharynx. The relative contributions of risk factors vary among individuals. Approaches to the measurement of these factors in clinical settings, with consequent enhancement of “personalized” therapeutic interventions, are being actively investigated.

Risk Factors and Prevalence

The major risk factors for OSAHS are obesity and male sex. Additional risk factors include mandibular retrognathia and micrognathia, a positive family history of OSAHS, genetic syndromes that reduce upper airway patency (e.g., Down syndrome, Treacher-Collins syndrome), adenotonsillar hypertrophy (especially in children), menopause (in women), and various endocrine syndromes (e.g., acromegaly, hypothyroidism).

Approximately 40–60% of cases of OSAHS are attributable to excess weight. Obesity predisposes to OSAHS through the narrowing effects of upper airway fat on the pharyngeal lumen. Obesity also reduces chest wall compliance and decreases lung volumes, resulting in a loss of caudal traction on upper airway structures. Obese individuals are at a fourfold or greater risk for OSAHS than their normal-weight counterparts. A 10% weight gain is associated with a >30% increase in AHI. Even modest weight loss or weight gain can influence the risk and severity of OSAHS. However, the absence of obesity does not exclude this diagnosis.

The prevalence of OSAHS is two- to fourfold higher among men than among women. Factors that predispose men to OSAHS include android patterns of obesity (resulting in upper-airway and abdominal fat deposition) and relatively greater pharyngeal length, which

exacerbates collapsibility. Premenopausal women are relatively protected from OSAHS by the influence of sex hormones on ventilatory drive. The decline in sex differences in older age is associated with an increased OSAHS prevalence in women after menopause.

Variations in craniofacial morphology that reduce the size of the posterior airway space increase OSAHS risk. The contribution of hard-tissue structural features to OSAHS is most evident in nonobese patients. Identification of features such as retrognathia can influence therapeutic decision making.

OSAHS has a strong genetic basis, as evidenced by its significant familial aggregation and heritability. For a first-degree relative of a patient with OSAHS, the odds ratio of having OSAHS is approximately twofold higher than that for someone without an affected relative. Several genetic variants have been associated with prevalence of OSAHS or with related traits, such as duration of apneas and hypopneas and overnight levels of hypoxemia.

OSAHS prevalence varies with age, from 2 to 15% among middle-aged adults to >20% among elderly individuals. There is a peak due to lymphoid hypertrophy among children between the ages of 3 and 8 years; with airway growth and lymphoid tissue regression during later childhood, prevalence declines. Then, as obesity prevalence increases in middle life and women enter menopause, OSAHS again increases.

The prevalence of OSAHS is especially high among patients with diabetes or hypertension. Individuals of Asian ancestry appear to be at increased risk of OSAHS at relatively low levels of body mass index, possibly because of the influence of craniofacial risk factors that narrow the nasopharynx. In the United States, African Americans, especially children and young adults, are at higher risk for OSAHS than their Caucasian counterparts. In a majority of adults with OSAHS, the disorder is undiagnosed.

Course of the Disorder

The precise onset of OSAHS is usually hard to identify. A person may snore for many years, often beginning in childhood, before OSAHS is identified. Weight gain may precipitate an increase in symptoms, which in turn may lead the patient to pursue an evaluation. OSAHS may become less severe with weight loss, particularly after bariatric surgery. Marked increases and decreases in the AHI are uncommon unless accompanied by weight change.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Obstructive Sleep Apnea/Hypopnea Syndrome

An evaluation for OSAHS should be considered in patients with symptoms of OSAHS and one or more risk factors. Screening also should be considered in patients who report symptoms consistent with OSAHS and who are at high risk for OSAHS-related morbidities, such as hypertension, diabetes mellitus, and cardiac and cerebrovascular diseases.

SYMPTOMS AND HISTORY

When possible, a sleep history should be obtained with assistance from a bed partner or household member. Snoring is the most common complaint; however, its absence does not exclude the diagnosis, as pharyngeal collapse may occur without tissue vibration. Gasping or snorting during sleep may also be reported, reflecting termination of individual apneas with abrupt airway opening. Dyspnea is unusual, and its absence generally distinguishes OSAHS from paroxysmal nocturnal dyspnea, nocturnal asthma, and acid reflux with laryngospasm. Patients also may describe frequent awakening or sleep disruption, which is more common among women and older adults. The most common daytime symptom is excessive sleepiness, identified by a history of difficulty maintaining alertness or involuntary periods of dozing. However, many women preferentially report fatigue rather than sleepiness. Other symptoms include a dry mouth, nocturnal heartburn, diaphoresis of the chest and neck, nocturia, morning headaches, trouble concentrating, irritability, and mood disturbances. Although difficulty falling sleep and maintaining sleep are characteristics of insomnia disorders, they also may occur with OSAHS, especially in women. Several questionnaires that evaluate snoring frequency, self-reported apneas, and daytime sleepiness can facilitate OSAHS screening. The predictive ability of a questionnaire can be enhanced by a consideration of whether the patient is male or has risk factors such as obesity or hypertension.

PHYSICAL FINDINGS

Physical findings often reflect the etiologic factors for the disorder as well as comorbid conditions, particularly vascular disease. On examination, patients may exhibit hypertension and regional (central) obesity, as indicated by a large waist and neck circumference. The oropharynx may reveal a small orifice with crowding due to an enlarged tongue, a low-lying soft palate with a bulky uvula, large tonsils, a high-arched palate, and/or micro/retrognathia. Since nasal resistance can increase the propensity to pharyngeal collapse, the nasal cavity should be inspected for polyps, septal deviation, and other signs of obstruction. Because patients with heart failure are at increased risk for both OSAHS and CSA, a careful cardiac examination should be conducted to detect possible left- or right-sided cardiac dysfunction. Evidence of cor pulmonale suggests a comorbid cardiopulmonary condition; OSAHS alone is not thought to cause right-heart failure. A neurologic evaluation is needed to evaluate for conditions such as neuromuscular and cerebrovascular diseases, which increase OSAHS risk.

LABORATORY FINDINGS Diagnostic Findings

Since symptoms and signs do not accurately predict the severity of sleep-related breathing disturbances, specific diagnosis and categorization of OSAHS severity requires objective measurement of breathing during sleep. The gold standard for diagnosis of OSAHS is an overnight polysomnogram (PSG). A negative in-laboratory PSG usually rules out OSAHS. However, false-negative studies can result if the study did not collect representative information on the patient’s usual sleep, particularly if there was insufficient REM sleep or inadequate supine sleep during testing. Home sleep tests that record only a few respiratory and cardiac channels

commonly are used as a cost-effective means for diagnosing patients without significant comorbidity who have a high pretest probability of OSAHS. However, a home study may yield a false-negative result if sleep time is not accurately estimated or in individuals experiencing hypopneas with arousals rather than oxyhemoglobin desaturation. Further evaluation may therefore be required.

The key physiological information collected during a sleep study for OSAHS assessment includes measurement of breathing (changes in airflow, respiratory excursion), oxygenation (hemoglobin oxygen saturation), body position, and cardiac rhythm. In addition, PSGs and some home sleep studies measure sleep continuity and sleep stages (by electroencephalography, chin electromyography, electro-oculography, and actigraphy), limb movements (by leg sensors), and snoring intensity. This information is used to quantify the frequency and subtypes of abnormal respiratory events during sleep as well as associated changes in oxygen hemoglobin saturation, arousals, and sleep stage distributions. **Tables 291-1** and **291-2** define the respiratory events scored and the severity guidelines employed during a sleep study. **Fig. 291-2** shows examples of sleep-related respiratory events. A typical sleep study report provides quantitative data such as the AHI and the profile of oxygen saturation over the night (mean, nadir, time at low levels). Reports may also include the respiratory disturbance index, which includes the number of respiratory effort-related arousals in addition to the number of apneas plus hypopneas. In-laboratory, PSG also quantifies sleep latency (time from “lights off” to first sleep onset), sleep efficiency (percentage of time asleep relative to time in bed), arousal index (number of cortical arousals per hour of sleep), time in each sleep stage, and periodic limb movement index. OSAHS severity can be further characterized according to the degree of sleep fragmentation associated with respiratory disturbances. Relevant metrics include the frequency of cortical micro-arousals or awakenings per sleep hour (arousal index), reduction in sleep continuity (low sleep efficiency), reduction of time in deeper stages of sleep (stage N3 and REM sleep), and increases in light sleep (stage N1). The detection of autonomic arousals, such as surges in blood pressure, changes in heart rate, and abnormalities in cardiac rhythm, also provides relevant information on OSAHS severity.

Other Laboratory Findings

Various imaging studies, including cephalometric radiography, MRI, CT, and fiberoptic endoscopy, can be used to identify anatomic risk factors for OSAHS. Cardiac testing may yield evidence of impaired systolic or diastolic ventricular function or abnormal cardiac structure. Overnight blood pressure monitoring often displays a “non-dipping” pattern (absence of the typical 10-mmHg fall of blood pressure during sleep compared to wakefulness). Arterial blood gas measurements made during wakefulness are usually normal. Waking hypoxemia or hypercarbia suggests coexisting cardiopulmonary disease or hypoventilation syndromes. Patients with severe nocturnal hypoxemia may have elevated hemoglobin values. A multiple sleep latency test or a maintenance of wakefulness test can be useful in quantifying sleepiness and helping to distinguish OSAHS from narcolepsy.

TABLE 291-1 Respiratory Event Definitions

- *Apnea:* Cessation of airflow for ≥10 s during sleep, accompanied by:
  - Persistent respiratory effort (obstructive apneas, **Fig. 291-2A**), or
  - Absence of respiratory effort (central apneas, **Fig. 291-2B**)
- *Hypopnea:* A ≥30% reduction in airflow for at least 10 s during sleep that is accompanied by either a ≥3% desaturation or an arousal **(Fig. 291-2C)**
- *Respiratory effort-related arousal (RERA):* Partial obstruction that does not meet the criteria for hypopnea but provides evidence of increasing inspiratory effort (usually through pleural pressure monitoring) punctuated by an arousal **(Fig. 291-2D)**
- *Flow-limited breath:* A partially obstructed breath, typically within a hypopnea or RERA, identified by a flattened or “scooped-out” inspiratory flow shape **(Fig. 291-3)**

TABLE 291-2 Obstructive Sleep Apnea/Hypopnea Syndrome (OSAHS): Quantification and Severity Scale

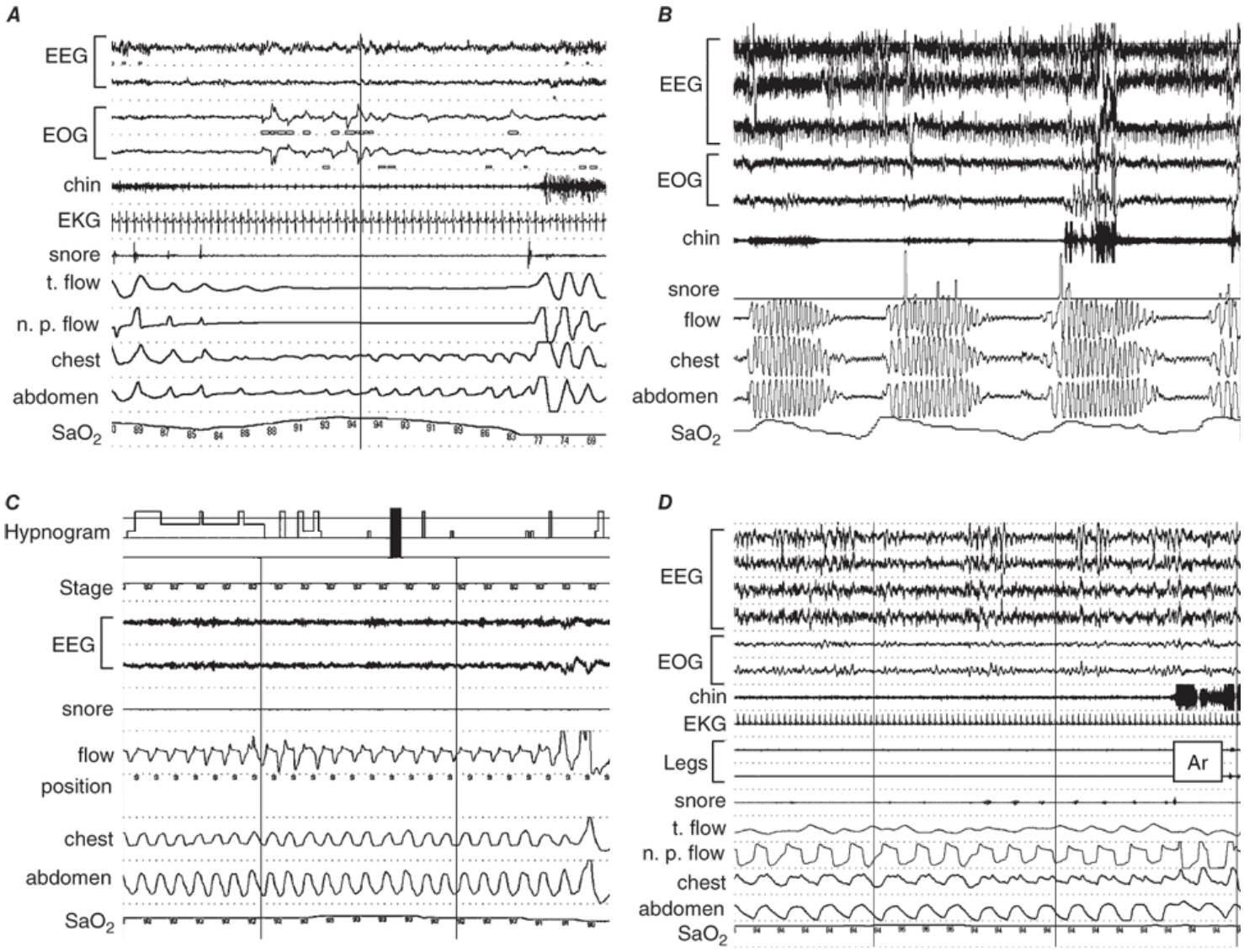
- *Apnea-hypopnea index (AHI)*[a]*:* Number of apneas plus hypopneas per hour of sleep
- *Respiratory disturbance index (RDI):* Number of apneas plus hypopneas plus RERAs per hour of sleep
- *Mild OSAHS:* AHI of 5–14 events/h
- *Moderate OSAHS:* AHI of 15–29 events/h
- *Severe OSAHS:* AHI of ≥30 events/h

[a]Each level of AHI can be further quantified by level of sleepiness and associated hypoxemia.

FIGURE 291-2

Obstructive apnea. ***A.*** There are 30 s of no airflow, as shown in the nasal pressure (n. p. flow) and thermistor-measured flow (t. flow). Note the presence of chest-abdomen paradox, indicating respiratory effort against an occluded airway. ***B.*** Central apnea in a patient with Cheyne-Stokes respiration due to congestive heart failure. The flat chest-abdomen tracings indicate the absence of inspiratory effort during the central apneas. ***C.*** Hypopnea. Partial obstruction of the pharyngeal airway can limit ventilation, leading to desaturation (a mild decrease in this patient, from 93 to 90%) and arousal. ***D.*** Respiratory effort-related arousal (RERA). Minimal flow reduction

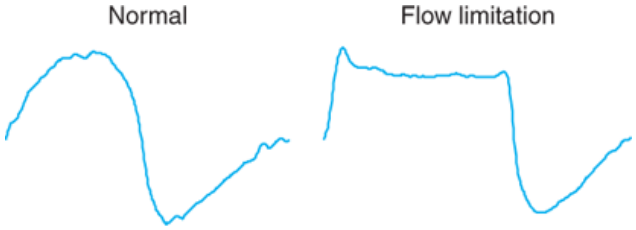
terminated by an arousal (Ar) without desaturation constitutes a RERA. EEG, electroencephalogram; EOG, electro-oculogram; EKG, electrocardiogram.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 291-3

Example of flow limitation. The inspiratory flow pattern in a patent airway is rounded and peaks in the middle. In contrast, a partially obstructed airway exhibits an early peak followed by mid-inspiratory flattening, yielding a scooped-out appearance.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Health Consequences and Comorbidities

OSAHS is a major contributor to cardiac, cerebrovascular, and metabolic disorders as well as to premature death. It is the most common medical cause of daytime sleepiness and negatively influences quality of life. This broad range of health effects is attributable to the impact of sleep fragmentation, cortical arousal, and intermittent hypoxemia on vascular, cardiac, metabolic, and neurologic

functions. OSAHS-related respiratory events stimulate sympathetic overactivity, leading to acute blood pressure surges during sleep, endothelial damage, and nocturnal as well as daytime hypertension. OSAHS-related hypoxemia also stimulates release of acute-phase proteins and reactive oxygen species that exacerbate insulin resistance and lipolysis and cause an augmented prothrombotic and proinflammatory state. Inspiratory effort against an occluded airway causes large intrathoracic negative pressure swings, altering cardiac preload and afterload and resulting in cardiac remodeling and reduced cardiac function. Hypoxemia and sympathetic-parasympathetic imbalance also may cause electrical remodeling of the heart and myocyte injury.

HYPERTENSION

OSAHS can raise blood pressure to pre-hypertensive and hypertensive ranges, increase the prevalence of a non-dipping overnight blood pressure pattern, and increase the risk of resistant hypertension. Elevations in blood pressure are due to augmented sympathetic nervous system activation as well as alterations in the renin-angiotensin–aldosterone system and fluid balance. Treatment of OSAHS with nocturnal continuous positive airway pressure (CPAP) has been shown to reduce 24-h ambulatory blood pressure. Although the overall impact of CPAP on blood pressure levels is relatively modest (averaging 2–4 mmHg), larger improvements are observed among patients who have a high AHI, report daytime sleepiness, or who have resistant hypertension.

CARDIOVASCULAR, CEREBROVASCULAR, AND METABOLIC DISEASES

Among the most serious health consequences of OSAHS is its impact on cardiac and metabolic functions. Strong epidemiologic evidence indicates that OSAHS significantly increases the risk of coronary artery disease, heart failure with and without reduced ejection fraction, atrial and ventricular arrhythmias, atherosclerosis and coronary artery disease, stroke, and diabetes. Treatment of OSAHS has been shown to reduce several markers of cardiovascular risk, improve insulin resistance, decrease the recurrence rate of atrial fibrillation, and improve various outcomes in patients with active cardiovascular disease. Large-scale trials have not yet, however, demonstrated that OSAHS treatment with CPAP reduces cardiac event rates and prolongs survival, perhaps due to limited adherence with treatment among trial participants.

SLEEPINESS

More than 50% of patients with moderate to severe OSAHS report daytime sleepiness. Patients with OSAHS symptoms have a twofold increased risk of occupational accidents. Individuals with elevated AHIs are involved in motor vehicle crashes as much as seven times more often than persons with normal AHIs. Randomized controlled trials have shown that treatment of OSAHS with nasal CPAP therapy alleviates sleepiness as measured by either questionnaire or objective testing in patients with both mild and more severe disease. However, the degree of improvement varies widely. Residual sleepiness may be due to several factors, including suboptimal treatment adherence, insufficient sleep time, other sleep disorders, or prior hypoxic-mediated damage in brain areas involved in alertness. Visceral adipose tissue, whose amounts are increased in patients with OSAHS, releases somnogenic cytokines that may contribute to sleepiness. Thus, even after treatment, it is important to assess and monitor patients for residual sleepiness and to evaluate the necessity of optimizing treatment adherence, improving sleep patterns, and identifying other disorders contributing to sleepiness. Careful and supervised use of alerting agents may be administered as adjunctive treatment in patients in whom sleepiness does not respond to CPAP alone.

QUALITY OF LIFE AND MOOD

Reductions in health-related quality of life are common in patients with OSAHS, with the largest decrements on the physical and vitality subscales. Numerous studies, including a large-scale trial of minimally symptomatic patients, have shown that treatment with CPAP can improve these patient-reported outcomes. Depression, in particular symptoms of somatic depression (irritability, fatigue, lack of energy), is commonly reported in OSAHS and improves with CPAP.

TREATMENT

TREATMENT Obstructive Sleep Apnea/Hypopnea Syndrome

A comprehensive approach to the management of OSAHS is needed to reduce risk factors and comorbidities. The clinician should seek to identify and address lifestyle and behavioral factors as well as comorbidities that may be exacerbating OSAHS. As appropriate, treatment should aim to reduce weight; optimize sleep duration (7–9 h); regulate sleep schedules (with similar bedtimes and wake times across the week); encourage the patient to avoid sleeping in the supine position; treat nasal allergies; increase physical activity; eliminate alcohol ingestion (which impairs pharyngeal muscle activity) within 3 h of bedtime; and minimize use of sedating medications. Patients should be counseled to avoid drowsy driving.

CPAP is the standard medical therapy with the highest level of evidence for efficacy. Delivered through a nasal or nasal-oral mask, CPAP works as a mechanical splint to hold the airway open, thus maintaining airway patency during sleep. An overnight CPAP titration study, performed either in a laboratory or with a home “auto-titrating” device, is required to determine the optimal pressure setting that reduces the number of apneas/hypopneas during sleep, improves gas exchange, and reduces arousals. Rates of adherence to CPAP treatment are highly variable (average, 50–80%) and may be improved with support by a skilled health care team who can address side effects, help the patient “problem solve,” and provide motivational education **(Table 291-3)**. Despite the limitations of CPAP, controlled studies have demonstrated its beneficial effect on blood pressure, alertness, mood, quality of life, and insulin sensitivity. Uncontrolled studies also indicate a favorable effect on cardiovascular outcomes, cardiac ejection fraction, atrial fibrillation recurrence, and mortality risk.

Oral appliances for OSAHS work by advancing the mandible, thus opening the airway by repositioning the lower jaw and pulling the tongue forward. These devices generally work better when customized for patient use; maximal adaptation can take several weeks. Efficacy studies show that these devices can reduce the AHI by ≥50% in two-thirds of individuals, although these data are based largely on patients with mild OSAHS. Some patients with moderate or severe OSAHS respond to oral appliances as well, although no consistent predictors of success have been identified in these groups and thus follow-up PSG testing is recommended. Side effects of oral appliances include temporomandibular joint pain and tooth movement. Oral appliances are most often used for treating patients with mild OSAHS or patients who do not tolerate CPAP. However, since adherence to the use of oral appliances sometimes exceeds CPAP adherence, these devices are under investigation for treatment of more severe disease.

Upper airway surgery for OSAHS is less effective than CPAP and is mostly reserved for the treatment of patients who snore, have mild OSAHS, or cannot tolerate CPAP. Uvulopalatopharyngoplasty (removal of the uvula and the margin of the soft palate) is the most common surgery and, although results vary greatly, is generally less successful than treatment with oral appliances. Upper airway surgery is less effective in severe OSAHS and in obese patients. Success rates may be higher for multilevel surgery (involving more than one site/structure) performed by an experienced surgeon, but the selection of patients is an important factor and relies on careful targeting of culprit areas for surgical resection. Bariatric surgery is an option for obese patients with OSAHS and can improve not only OSAHS but also other obesity-associated health conditions. Other procedures that can decrease snoring but have minimal effects on OSAHS include injection of the soft palate (resulting in stiffening), radiofrequency ablation, laser-assisted uvulopalatoplasty, and palatal implants.

Upper airway neuro-stimulation is a recently tested alternative treatment for OSAHS. Unilateral stimulation of the hypoglossal nerve through a surgically implanted device was shown to significantly decrease the AHI and improve a number of patient-reported outcomes, such as sleepiness and quality of life, for a duration of at least 18 months after treatment. Initial studies enrolled patients with a BMI ≤32 kg/m^2, moderate OSAHS, absence of complete concentric pharyngeal collapse (considered to decrease surgical efficacy) and were unable to be treated successfully with CPAP. Additional research is underway to further evaluate longer-term effectiveness and potential utility of this treatment in other patient groups.

Supplemental oxygen can improve oxygen saturation, but there is little evidence that it improves OSAHS symptoms or the AHI in unselected patients.

CENTRAL SLEEP APNEA

CSA, which is less common than OSAHS, may occur in isolation or, more often, in combination with obstructive events in the form of “mixed” apneas. CSA is often caused by an increased sensitivity to p_{CO2}, which leads to an unstable breathing pattern that manifests as hyperventilation alternating with apnea. A prolonged circulation delay between the pulmonary capillaries and carotid chemoreceptors is also a contributing cause; thus individuals with congestive heart failure are at risk for CSA. With prolonged circulation delay, there is a crescendo-decrescendo breathing pattern known as *Cheyne-Stokes respiration* **(Fig. 291–2B)**. Other risk factors for CSA include opioid medications (which appear to have a dose-dependent effect on CSA) and hypoxia (e.g., breathing at high altitude). In some individuals, CPAP—particularly at high pressures—seems to induce central apnea; this condition is referred to as *complex sleep apnea*. Rarely, CSA may be caused by blunted chemosensitivity due to congenital disorders (congenital central hypoventilation syndrome) or acquired factors. CSA is an independent risk factor for the development of both heart failure and atrial fibrillation, possibly related to elevations in sympathetic nervous system activity that accompany this disorder. Patients with CSA may report symptoms of frequent awakenings as well as daytime fatigue. Treatment of CSA is difficult and depends on the underlying cause. Limited data suggest that supplemental oxygen can reduce the frequency of central apneas, particularly in patients with hypoxemia. Cheyne-Stokes respiration is treated by optimizing therapy for heart failure. At this time, there is no good evidence that CPAP, including *adaptive servoventilation* (a form of ventilatory support that attempts to regularize the breathing pattern) improves health outcomes in patients with Cheyne-Stokes respiration without OSAHS.

TABLE 291-3 Side Effects of Continuous Positive Airway Pressure (CPAP) and Their Treatments

Side Effect	Treatment
Nasal congestion	Provide heated humidification, administer saline/steroid nasal sprays
Claustrophobia	Change mask interface (e.g., to nasal prongs), promote habituation (i.e., practice breathing on CPAP while awake)
Difficulty exhaling	Temporarily reduce pressure, provide bilevel positive airway pressure
Bruised nasal ridge	Change mask interface, provide protective padding
Aerophagia	Administer antacids

FURTHER READING

Berry R, Wagner M: *Sleep Medicine Pearls*, 3rd ed. Philadelphia, Elsevier, 2015.
Javaheri S et al: Sleep apnea: Types, mechanisms, and clinical cardiovascular consequences. J Am Coll Cardiol 769:841, 2017.
Kapur VK et al: Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline. J Clin Sleep Med 13:479, 2017. [PubMed: 28162150]
McEvoy RD et al: CPAP for prevention of cardiovascular events in obstructive sleep apnea. N Engl J Med 375:919, 2016. [PubMed: 27571048]
Strollo PJ Jr et al: Upper-airway stimulation for obstructive sleep apnea. N Engl J Med 370:139, 2014. [PubMed: 24401051]

Chapter 292: Lung Transplantation

Elbert P. Trulock, III

INTRODUCTION

Lung transplantation is a therapeutic consideration for many patients with nonmalignant end-stage lung disease, and it prolongs survival and improves quality of life in appropriately selected recipients. Since 1985 more than 51,000 adult lung transplants have been recorded worldwide, and annual volume has reached ~4000 transplants per year.

INDICATIONS

The indications for lung transplantation span the gamut of lung diseases, and the distribution reflects both the prevalence and prognosis of the diseases and the applicable organ allocation policies. According to international registry data, the most common indications in recent years have been idiopathic pulmonary fibrosis (IPF), ~30%; chronic obstructive pulmonary disease (COPD), ~27%; cystic fibrosis (CF), ~15%; α_1-antitrypsin deficiency emphysema, ~3%; and idiopathic pulmonary arterial hypertension (IPAH), ~2.5%. Other lung diseases have comprised the balance of primary indications, and retransplantation has accounted for ~3% of procedures. Since 2001, IPF has increased from ~15 to ~30%, and COPD has decreased from ~40 to ~27% among the indications.

REFERRAL AND RECIPIENT SELECTION

Transplantation should be considered when other therapeutic options have been exhausted and when the patient's prognosis is expected to improve as a result of the procedure. Survival rates after transplantation can be compared with predictive indices for the patient's disease, but each patient's individual clinical circumstances must be incorporated into the assessment. Moreover, quality of life is a primary motive for transplantation for many patients, and the prospect of improved quality-adjusted survival is often attractive even if the survival advantage itself is questionable.

Disease-specific consensus guidelines for referring patients for evaluation and for listing them for transplantation are summarized in **Table 292-1**. Candidates for lung transplantation are also thoroughly screened for comorbidities that might affect the outcome adversely. Conditions such as systemic hypertension, diabetes mellitus, gastroesophageal reflux, and osteoporosis are not unusual, but if uncomplicated and adequately managed, they do not disqualify patients from transplantation. The upper age limit is ~70–75 years at most centers, and the proportion of older recipients has been increasing. In 2014, 29% of adult recipients in the United States were ≥65 years old.

TABLE 292-1 Disease-Specific Guidelines for Referral and Transplantation

Chronic Obstructive Pulmonary Disease

Referral for Evaluation

- Progressive despite medications, oxygen, and pulmonary rehabilitation
- FEV_1 <25%
- PaO_2 <60 mmHg or Paco2 >50 mmHg
- BODE index 5–6

Listing for Transplantation

- BODE index ≥7
- FEV_1 <15–20%
- Moderate to severe pulmonary hypertension
- Three or more severe exacerbations in preceding year
- One severe exacerbation with acute hypercapnic respiratory failure

Cystic Fibrosis/Bronchiectasis

Referral for Evaluation

- FEV_1 <30% or rapidly declining despite optimal therapy
- Pulmonary hypertension (in absence of hypoxemic exacerbation)
- 6-min walk distance <400 m
- Clinical deterioration with increasing frequency of exacerbations, with
  - – An episode of acute respiratory failure requiring ventilatory support
  - – Increasing antibiotic resistance and poor recovery from exacerbations
  - – Worsening nutritional status despite adequate supplementation
  - – Refractory or recurrent pneumothorax
  - – Life-threatening hemoptysis despite bronchial artery embolization

Listing for Transplantation

- Chronic hypoxemic or hypercapnic respiratory failure
- Pulmonary hypertension
- Rapid decline in lung function
- Long-term noninvasive ventilator support
- Frequent hospitalization

WHO functional class IV

Idiopathic Pulmonary Fibrosis

Referral for Evaluation

Pathologic or radiographic evidence of UIP or NSIP regardless of lung function

FVC <80% or DL_{CO} <40%

Dyspnea or functional limitation attributable to lung disease

Any oxygen requirement (rest or exercise)

Listing for Transplantation

Decrement in FVC ≥10% or in DL_{CO} >15% during 6 months of follow-up

Pulmonary hypertension

Desaturation to Sp_{O2} <88% during 6-min walk test

6-min walk test distance <250 m or decrement >50 m over 6 months

Hospitalization for acute exacerbation

Idiopathic Pulmonary Arterial Hypertension

Referral for Evaluation

NYHA functional class III or IV during escalating therapy

Use of parenteral therapy regardless of NYHA functional class

Rapidly progressive disease

Listing for Transplantation

NYHA functional class III or IV despite combination therapy with a prostanoid

Cardiac index <2 L/min/m^2 or right atrial pressure >15 mmHg

6-minute walk test distance <350 m

Progressive right heart failure or significant pericardial effusion or hemoptysis

Abbreviations: BODE—body-mass index (B), airflow obstruction (O), dyspnea (D), exercise capacity (E); FVC, forced vital capacity; FEV_1, forced expiratory volume in 1 s; DL_{CO}, diffusing capacity for carbon monoxide; HRCT, high resolution computed tomography; ICU, intensive care unit; NSIP, nonspecific interstitial pneumonitis; NYHA, New York Heart Association; $PaCO_2$ and PaO_2, partial pressures of carbon dioxide and oxygen, respectively, in arterial blood; Sp_{O2}, arterial oxygen saturation by pulse oximetry; UIP, usual interstitial pneumonitis; WHO, World Health Organization

Source: Summarized from D. Weill et al: J Heart Lung Transplant 34:1, 2015. For BODE index, BR Celli et al: N Engl J Med 350:1005, 2004.

Standard exclusions include HIV infection, chronic active hepatitis B or C infection, uncontrolled or untreatable pulmonary or extra-pulmonary infection, uncured malignancy, active cigarette smoking,

drug or alcohol dependency, irreversible physical deconditioning, chronic nonadherence with medical care, significant disease of another vital organ (e.g., heart, liver, or kidney), and psychiatric or psychosocial situations that could substantially interfere with post-transplantation management. Other problems that may compromise the outcome constitute relative contraindications. Some typical issues are ventilator-dependent respiratory failure, extracorporeal life support, obesity, coronary artery disease, and previous thoracic surgical procedures. Chronic infection with antibiotic-resistant *Pseudomonas species, Burkholderia species, Aspergillus species,* or nontuberculous mycobacteria is a unique concern in some patients with CF. The potential impact of these and other factors has to be judged in clinical context to determine an individual candidate's suitability for transplantation.

WAITING LIST AND ORGAN ALLOCATION

Organ allocation policies are influenced by medical, ethical, geographical, and political factors, with systems varying from country to country. Regardless of the system, potential recipients are placed on a waiting list and must be matched for blood group compatibility and, with some latitude, for lung size with an acceptable donor. If the potential recipient is allosensitized with antibodies to any human leukocyte antigen (HLA), the donor also has to be HLA compatible.

Most lungs are procured from deceased donors after brain death, but only ~20% of brain-death organ donors yield either one or two lungs suitable for transplantation. Lungs from donors after circulatory death have been utilized to a limited extent (~2% of lung donors in the United States in 2014). Ex vivo lung perfusion is being used by some centers to assess donor lungs that are marginal for implantation by standard criteria; if the results of ex vivo testing are satisfactory, these lungs have been transplanted successfully.

In the United States, a lung allocation score (LAS) system has been used to prioritize patients on the waiting list since 2005. For the purposes of the LAS, patients are divided into four groups by diagnosis: A—COPD and emphysema; B—IPAH and other forms of pulmonary hypertension; C—cystic fibrosis and other forms of bronchiectasis; D—IPF and other interstitial diseases. The LAS is based on the patient's risk of death during 1 year on the waiting list and the patient's likelihood of survival for 1 year after transplantation. It can range from 0 to 100, and priority for transplantation is ranked from highest to lowest scores. Both the lung disease and its severity affect a patient's LAS; parameters in the LAS must be updated biannually, but can be submitted for recalculation whenever the patient's condition changes. Patients in group D usually have the highest scores, and those in group A, the lowest.

In recent years, the U.S. national waiting list typically has contained ~1500 patients with a median LAS of ~35–36. In 2014, 50% of the patients on the waiting list were in group D, and the median waiting time to transplantation was 3.7 months. The overall death rate (deaths per 100 waitlist years) on the waiting list was ~10, but the rate varied considerably by diagnostic group (e.g., A, ~4; B, ~11; C, ~12; D, ~18) and by LAS (e.g., 35–39, ~10; 40–49, ~22; ≥50, ~125). Because the LAS system prioritizes patients with the highest risk of death on the waiting list, many patients are now critically ill at the time of transplantation. In 2014, 25% of recipients were hospitalized (15% in intensive care), and 9% were being supported by mechanical ventilation, extracorporeal life support or both at the time of transplantation.

TRANSPLANT PROCEDURE

Bilateral transplantation is mandatory for CF and other forms of bronchiectasis because the risk of spillover infection from a remaining native lung precludes single-lung transplantation. Heart-lung transplantation is obligatory for Eisenmenger syndrome with complex anomalies that cannot be readily repaired in conjunction with lung transplantation and for concomitant end-stage lung and heart

disease. However, cardiac replacement is not necessary for cor pulmonale because right ventricular function will recover when pulmonary vascular afterload is normalized by lung transplantation.

Either bilateral or single-lung transplantation is an option for other diseases unless there is a special consideration, but for all indications, bilateral transplantation is performed most often. In 2014, 68% of all transplants in the United States were bilateral (68% of transplants for COPD; 54% for IPF).

Living donor lobar transplantation has played a limited role in adult lung transplantation, but is rarely performed anymore; only three cases were recorded in the U.S. registry from 2010 through 2015. When performed, it usually has been reserved for teenagers or young adults with CF who were unlikely to survive the wait for a deceased organ donor.

POSTTRANSPLANTATION MANAGEMENT

Immunosuppression

Induction therapy is increasingly utilized, and in 2014, ~65% of recipients in both the U.S. and international registries received an induction agent. The interleukin-2 receptor antagonist basiliximab has been the most widely used drug (~50% of recipients), but the antilymphocyte globulins and alemtuzumab have been used, as well. A three-drug maintenance immunosuppressive regimen that includes a calcineurin inhibitor (cyclosporine or tacrolimus), a purine synthesis antagonist (azathioprine or a mycophenolic acid precursor), and prednisone is traditional; the triad of tacrolimus, mycophenolate, and prednisone is the most commonly prescribed regimen. Subsequently, other drugs such as sirolimus or everolimus may be substituted for various reasons. Prophylaxis for *Pneumocystis jiroveci* pneumonia is standard, and prophylaxis against cytomegalovirus (CMV) infection and fungal infection is part of many protocols. The dose of cyclosporine, tacrolimus, sirolimus, or everolimus is adjusted by blood-level monitoring. All of these agents are metabolized by the hepatic cytochrome P450 system, and interactions with medications that affect this pathway can significantly alter their clearance and blood level.

Spirometry and Bronchoscopy

Routine management focuses on monitoring of the allograft, regulating immunosuppressive therapy, and detecting problems or complications expeditiously. Regular contact with a nurse coordinator, physician follow-up, chest radiography, blood tests, and spirometry are customary. Surveillance bronchoscopy with bronchoalveolar lavage and transbronchial biopsies is employed by some programs to screen for occult acute cellular rejection (ACR) or infection, and bronchoscopy is the standard invasive procedure to investigate problems with the allograft. If recovery is uncomplicated, lung function rapidly improves and then stabilizes by 3–6 months after transplantation. Subsequently, the variation in spirometric measurements is small, and a sustained decline of ≥10–15% signals a potentially significant problem.

OUTCOMES

Survival

Major registries publish survival rates **(Table 292-2)** and other outcomes annually (*www.ishlt.org; www.srtr.org*). In the international registry, for the cohort from 1990 to 2013, median survival for recipients with IPF was 4.7 years; IPAH, 5.7 years; COPD, 5.5 years; CF, 8.6 years; and sarcoidosis, 6.1 years. Transplant procedure and recipient age also have a significant impact on outcome. Long-term survival has been significantly better after bilateral transplantation than after unilateral transplantation for all of the major indications. For recipients 18–49 years of age, the median survival

is ~7.2 years, but it decreases to ~5.5 years for those 50–59 years old and to ~4.5 years for those ≥60 years old. In the U.S. registry, survival rates have been lowest for recipients who were in group D, were ≥65 years old, or had an LAS ≥60.

TABLE 292-2 Recipient Survival, by Pretransplantation Diagnosis

Diagnosis; Transplant Type	Survival Rate, %					
	3 Months	1 Year	3 Years	5 Years	10 Years	15 Years
Chronic obstructive pulmonary disease[a]						
Bilateral	93	86	71	59	37	16
Single	92	84	66	51	24	6
α_1-antitrypsin deficiency emphysema						
Bilateral	89	81	69	61	40	24
Single	88	78	62	52	29	13
Cystic fibrosis[a]	93	88	76	63	48	27
Idiopathic pulmonary fibrosis[a]						
Bilateral	89	82	66	55	34	16
Single	89	79	58	45	21	7
Idiopathic pulmonary arterial hypertension						
Bilateral	80	74	62	55	40	28
Single	71	62	52	41	24	15
Sarcoidosis	86	76	62	55	35	22

[a]Survival cohorts: 2009–2015 for 3 months, 1 year and 3 years; 1999–2008 for 5 years and 10 years; 1990–1998 for 15 years. For other diagnoses, cohort is 1990–2013 for all survival rates.

Source: Data from *www.ishlt.org/registries/slides.asp?slides=heartLungRegistry*. Accessed March 10, 2016.

The causes of death depend on the time period after transplantation. In the first 30 days, the major causes have been infection (~19%), graft failure (~24%), cardiovascular events (~11%) and technical problems (~11%), and for the remainder of the first year, the main contributors have been infection (~37%) and graft failure (~17%). After the first year, bronchiolitis and other forms of late graft failure have accounted for ~40–45% of deaths and infection for ~16–20%.

Risk factors for mortality have been analyzed in the international and U.S. registries. In these analyses, factors associated with an increased risk of death in the first year after transplantation have included the following: recipients hospitalized at the time of transplantation; recipients supported by mechanical ventilation, extracorporeal membrane oxygenation, or dialysis at the time of transplantation; and recipients undergoing retransplantation; however, other factors have contributed, too. The mortality risk has also been higher at centers with an annual volume below ~30 transplants/year.

Function

Regardless of the disease, successful transplantation impressively restores cardiopulmonary function. After bilateral transplantation, pulmonary function tests are typically normal; after unilateral transplantation, a mild abnormality characteristic of the remaining diseased lung is still apparent. Formal exercise testing usually demonstrates some impairment in maximum work rate and maximum oxygen uptake, but few recipients report any limitation to activities of daily living.

Quality of Life

Both overall and health-related quality of life measurements have improved after transplantation. With multidimensional profiles, improvements have extended across most domains and have been sustained longitudinally unless chronic rejection or some other complication develops.

Cost

The cost of transplantation depends on the health care system and economic factors that vary from country to country. In the United States in 2014, lung transplantation was covered by private insurance for 49% of recipients, Medicare for 36%, and other government-sponsored programs for 8.5%. In 2014, the average total of billed charges for a bilateral transplant for the period from 30 days before transplantation until 180 days after discharge from the transplant admission was $1,037,700. The total charge included the following components: all care during 30 days before transplantation, $30,700; organ procurement, $129,700; hospital transplant admission, $566,900; physician fees during transplant admission, $59,100; all inpatient and outpatient care for 180 days after discharge, $219,800; and all outpatient drugs, including immunosuppressants, for 180 days after discharge, $31,500. However, Medicare does not fully reimburse billed charges, and in the era from 2008 to 2012, the average Medicare cost from transplantation through the first posttransplant year was ~$240,000.

COMPLICATIONS

Lung transplantation can be complicated by a variety of problems **(Table 292-3)**. The average length of stay after bilateral transplantation in the United States in 2014 was 29.5 days, and the rehospitalization rate in the first year has been ~50%, higher than after any other solid organ transplant except intestine.

TABLE 292-3 Major Potential Complications of Lung Transplantation and Immunosuppression

Category	Complication
Allograft	Primary graft dysfunction; anastomotic dehiscence or stenosis; ischemic airway injury with bronchostenosis or bronchomalacia; rejection; infection; recurrence of primary disease (sarcoidosis, lymphangioleiomyomatosis, giant cell interstitial pneumonitis, diffuse panbronchiolitis, pulmonary alveolar proteinosis, Langerhans cell histiocytosis)
Thoracic	Phrenic nerve injury/diaphragmatic dysfunction; recurrent laryngeal nerve injury/vocal cord dysfunction; cervical ganglia injury/Horner's syndrome; pneumothorax; pleural effusion; chylothorax; empyema
Cardiovascular	Intraoperative or perioperative air embolism; postoperative pericarditis; perioperative myocardial injury/infarction; venous thromboembolism; supraventricular dysrhythmias; systemic hypertension
Gastrointestinal	Esophagitis (especially *Candida*, herpes or cytomegalovirus [CMV]); gastroparesis; gastroesophageal reflux; diarrhea (*C. difficile*; medications, especially mycophenolate mofetil and sirolimus); colitis (*C. difficile*; CMV)
Hepatobiliary	Hepatitis (especially CMV or medications); acalculous cholecystitis
Renal	Calcineurin inhibitor nephropathy; hemolytic-uremic syndrome (thrombotic microangiopathy)
Neurologic	Perioperative stroke; tremors; seizures; reversible posterior leukoencephalopathy; headaches
Musculoskeletal	Steroid myopathy; rhabdomyolysis (cyclosporine + HMG co-A reductase inhibitor treatment); osteoporosis; avascular necrosis
Metabolic	Obesity; diabetes mellitus; hyperlipidemia; idiopathic hyperammonemia
Hematologic	Anemia; leukopenia; thrombocytopenia; thrombotic microangiopathy
Oncologic	Lymphoproliferative disease and lymphoma; skin cancers; other malignancies

Primary Graft Dysfunction

Primary graft dysfunction (PGD), an acute lung injury, is a manifestation of multiple potential insults to the donor organ inherent in harvesting, preserving, and implanting it in the recipient. The principal clinical features are diffuse pulmonary infiltrates and hypoxemia within 72 h of transplantation; however, the presentation can be mimicked by pulmonary venous obstruction, hyperacute rejection, pulmonary edema, and pneumonia.

The severity is graded by a standardized system that is based on an edema pattern on chest radiograph and the Pa_{O2}/FiO_2 ratio (>300, grade 1; 200–300, grade 2; <200, grade 3). Up to 50% of recipients may have some degree of PGD, and ~10–20% have grade 3 PGD. The treatment follows the conventional, supportive paradigm for acute lung injury. Inhaled nitric oxide, inhaled epoprostenol, and extracorporeal membrane oxygenation have been used in severe cases. Retransplantation has also been performed, but when undertaken in the first 30 days, the 1-year survival rate has been only ~30%. Most recipients with mild PGD recover, but the mortality rate for severe PGD has been ~40–60%. PGD is also associated with longer postoperative ventilator support, longer intensive care unit and hospital stays, higher costs, and excess morbidity. Finally, severe (grade 3) PGD is a risk factor for the later development of chronic lung allograft dysfunction (CLAD).

Airway Complications

The bronchial blood supply to the donor lung is disrupted during procurement. Bronchial revascularization during transplantation is technically feasible in some cases, but it is not widely

practiced. Consequently, after implantation, the donor bronchus is dependent on retrograde bronchial blood flow from the pulmonary circulation and is vulnerable to ischemia.

The spectrum of airway problems includes anastomotic necrosis and dehiscence, occlusive granulation tissue, anastomotic or bronchial stenosis, and bronchomalacia. The incidence has been in the range of 7–18%, but the associated mortality rate has been low. These problems usually can be managed bronchoscopically with techniques such as simple endoscopic debridement, laser photoresection, balloon dilation, and bronchial stenting.

Lung Allograft Dysfunction

The transplanted lung is susceptible to a variety of conditions that can compromise graft function. Some of these, such as the various forms of rejection, are unique to transplantation, but others are not. Acute lung allograft dysfunction is most often caused by rejection or infection and may be completely reversible after diagnosis and treatment. CLAD can be the result of residual damage from an episode of acute allograft dysfunction, or it can develop separately in response to alloimmune and non-alloimmune injuries to the graft.

Acute Cellular Rejection

ACR is caused by T lymphocyte interactions with donor alloantigens, mainly in the major histocompatibility complex (MHC), and its incidence is highest in the first 6–12 months after transplantation. In the years 2008–2013, ~18% of recipients in the U.S. registry had an episode of ACR during the first year.

ACR can be clinically silent or can be manifested by nonspecific symptoms or signs that may include cough, low-grade fever, dyspnea, hypoxemia, inspiratory crackles, interstitial infiltrates, and declining lung function; however, clinical impressions are not reliable. The diagnosis is confirmed by transbronchial biopsies showing the characteristic lymphocytic infiltrates around arterioles or bronchioles, and a standardized pathologic scheme is used to grade the biopsies (grades A0–4 and B0–4 for the arteriolar and bronchiolar components, respectively).

Minimal ACR (grade A1) on a surveillance biopsy in a clinically stable recipient is not treated always, but higher grades (≥A2) generally are treated regardless of the clinical situation. Treatment usually includes a short course of high-dose steroid therapy and adjustment of the maintenance immunosuppressive regimen. Most episodes respond to this approach; however, more intensive therapy is sometimes necessary for persistent or recurrent episodes.

Chronic Lung Allograft Dysfunction

CLAD is the preferred term when lung function never reaches expected values because of an early complication or, more often, when there is a sustained decrement in lung function below previously normal baseline measurements. In the latter situation, two main forms of CLAD are recognized—bronchiolitis obliterans syndrome (BOS) and restrictive allograft syndrome (RAS), and both alloimmune and nonalloimmune fibroproliferative reactions can contribute to the pathogenesis. CLAD is the principal impediment to better long-term survival rates, and it is the source of substantial morbidity because of its impact on performance status and quality of life.

The distinguishing features of BOS and RAS are contrasted in **Table 292-4**. By definition, the decrement in lung function must persist for ≥3 weeks, and other causes of graft dysfunction must be excluded by an appropriate evaluation. Bronchoscopy with bronchoalveolar lavage and transbronchial biopsies is usually performed to exclude bronchostenosis, ACR, and infection. Transbronchial

biopsies, however, are insensitive for detecting obliterative bronchiolitis in BOS and are nonspecific in RAS, and pathologic confirmation is not required for diagnosis.

TABLE 292-4 Chronic Lung Allograft Dysfunction: Clinical Features of Bronchiolitis Obliterans Syndrome (BOS) and Restrictive Allograft Syndrome (RAS)

Syndrome	Pulmonary Function	HRCT Pattern	Pathology
BOS	Obstructive	Air trapping usually present	Obliterative bronchiolitis
	FEV_1 <80% of baseline	Minimal, if any, infiltrates ± bronchiectasis	Transbronchial biopsies insensitive
RAS	Restrictive	Infiltrates usually present	Parenchymal/pleural fibrosis
	TLC <90% of baseline, or FVC and FEV_1 <80% of baseline	± air trapping ± bronchiectasis	± obliterative bronchiolitis

Abbreviations: BOS, bronchiolitis obliterans syndrome; FEV_1, forced expiratory volume in 1 s; FVC, forced vital capacity; HRCT, high-resolution computed tomography; TLC, total lung capacity.

Source: Modified from GM Verleden et al: J Heart Lung Transplant 33:127, 2014.

BOS is the classic form of chronic rejection, and the prevalence approaches 50% by 5 years after transplantation. The severity is categorized by the decrement in FEV_1 from the average of the two best posttransplant values (20–35%, stage 1; 35–50%, stage 2; >50%, stage 3). Risk factors include PGD, ACR, humoral rejection and anti-HLA antibodies, viral infections (CMV pneumonia; community-acquired respiratory viral infections), airway colonization by *Pseudomonas aeruginosa* or *Aspergillus fumigatus*, and gastroesophageal reflux (GER).

BOS usually is treated with augmented immunosuppression, but there is no consensus about therapy. Strategies include adjustments to the maintenance drug regimen, the addition of azithromycin, and treatment with antilymphocyte globulin, photopheresis, or total lymphoid irradiation; antireflux surgery should be considered if GER is present. Although therapy may stabilize lung function, the overall results of treatment have been disappointing; median survival period after onset has been ~3–4 years. Retransplantation is a consideration if clinical circumstances and other comorbidities are not prohibitive, but survival rates have been inferior to that with primary transplantation.

RAS is less common than BOS, and occasionally the two forms coexist. Risk factors for RAS have not been delineated yet. Treatment usually includes a trial of steroid therapy and, in some cases, the same strategies that are used for BOS. Prognosis is worse than BOS, with a median survival ~1.5 years.

Humoral Rejection

The role of antibody-mediated rejection is still evolving. Hyperacute rejection is caused by preformed HLA antibodies in the recipient, but it is minimized by pretransplantation antibody screening coupled with virtual or direct cross-matching with any potential donor. Donor-specific HLA antibodies develop after transplantation in ~35–50% of recipients, and their presence has been associated with an increased risk of both ACR and BOS and with poorer overall survival. Criteria for antibody-mediated rejection include graft dysfunction, serologic detection of donor-specific antibodies, and a pathologic pattern of graft injury with evidence of antibody deposition; however, few cases in lung transplantation fulfill all of these criteria. Nonetheless, episodes of acute lung allograft dysfunction occasionally have been attributed directly to antibody-mediated rejection. If treatment is indicated, potential therapies include plasmapheresis and administration of intravenous immune globulin, rituximab, bortezomib, and eculizumab.

Infection

The lung allograft is especially susceptible to infection. In addition to a blunted immune response from the immunosuppressive drugs, other normal defenses are compromised: the cough reflex is diminished, and mucociliary clearance is impaired in the transplanted lung. The spectrum of infections includes both opportunistic and non-opportunistic pathogens.

Bacterial bronchitis or pneumonia can occur at any time, but it is very common in the perioperative period. Later, bronchitis occurs frequently in recipients with BOS, and *Pseudomonas aeruginosa* or methicillin-resistant *Staphylococcus aureus* is often the culprit.

Community-acquired respiratory viruses (influenza, parainfluenza, respiratory syncytial virus, metapneumovirus and others) are the most common viral infections and are easily identified with viral multiplex PCR testing of a nasopharyngeal swab or washing. CMV infection has become less problematic with highly sensitive blood CMV PCR monitoring and with widespread prophylactic and preemptive protocols using valganciclovir. However, CMV viremia, pneumonia, hepatitis, and gastroenteritis/colitis still occur occasionally, and treatment with ganciclovir is generally effective unless resistance has developed. The most problematic fungal infections are caused by *Aspergillus* species. The spectrum encompasses simple pulmonary colonization, tracheobronchitis, invasive pulmonary aspergillosis, and disseminated aspergillosis, and the clinical scenario dictates treatment.

Other Complications

Other potential complications are listed in Table 292-3. Many of them are related to side effects or toxicities of the immunosuppressive drugs, and the prevalence of some of them is high (hypertension—39% and 61%; renal dysfunction—16% and 45%; hyperlipidemia—24% and 47%; diabetes mellitus—13% and 34%; malignancy—4% and 22%, at 1 year and 5 years, respectively, in the U.S. registry). Management of these general medical problems is guided by standard practices, but the complex milieu of transplantation requires close collaboration and good communication among health care providers.

FURTHER READING

Levine DJ et al: Antibody-mediated rejection of the lung: A consensus report of the International Society for Heart and Lung Transplantation. J Heart Lung Transplant 35:397, 2016. [PubMed: 27044531]

Meyer KC et al: An international ISHLT/ATS/ERS practice guideline: Diagnosis and management of bronchiolitis obliterans syndrome. Eur Respir J 44:1479, 2014.

Valapour M et al: OPTN/SRTR 2014 annual data report: Lung. Am J Transplant 16:141, 2016. [PubMed: 26755267]

Verleden GM: A new classification system for chronic lung allograft dysfunction. J Heart Lung Transplant 33:127, 2014. [PubMed: 24374027]

Weil D et al: A consensus document for the selection of lung transplant candidates: 2014—an update from the Pulmonary Transplant Council of the International Society for Heart and Lung Transplantation. J Heart Lung Transplant 34:1, 2015. [PubMed: 25085497]

Yusen RD et al: The Registry of the International Society for Heart and Lung Transplantation: Thirty-second official adult lung and heart-lung transplantation report—2015; focus theme: early graft failure. J Heart Lung Transplant 34:1264, 2015. [PubMed: 26454740]

Chapter 293: Approach to the Patient with Critical Illness

John P. Kress; Jesse B. Hall

INTRODUCTION

The care of critically ill patients requires a thorough understanding of pathophysiology and centers initially on the resuscitation of patients at the extremes of physiologic deterioration. This resuscitation is often fast-paced and occurs early, without a detailed awareness of the patient's chronic medical problems. While physiologic stabilization is taking place, intensivists attempt to gather important background medical information to supplement the real-time assessment of the patient's current physiologic conditions. Numerous tools are available to assist intensivists in the accurate assessment of pathophysiology and management of incipient organ failure, offering a window of opportunity for diagnosing and treating underlying disease(s) in a stabilized patient. Indeed, the use of invasive interventions such as mechanical ventilation and renal replacement therapy is commonplace in the intensive care unit (ICU). An appreciation of the risks and benefits of such aggressive and often invasive interventions is vital to ensure an optimal outcome. Nonetheless, intensivists must recognize when a patient's chances for recovery are remote or nonexistent and must counsel and comfort dying patients and their significant others. Critical care physicians often must redirect the goals of care from resuscitation and cure to comfort when the resolution of an underlying illness is not possible.

ASSESSMENT OF ILLNESS SEVERITY

In the ICU, illnesses are frequently categorized by degree of severity. Numerous severity-of-illness (SOI) scoring systems have been developed and validated over the past three decades. Although these scoring systems have been validated as tools to assess populations of critically ill patients, their utility in predicting individual patient outcomes is not clear. SOI scoring systems are important for defining populations of critically ill patients. Such systematic scoring allows effective comparison of groups of patients enrolled in clinical trials. In verifying a purported benefit of therapy, investigators must be confident that different groups involved in a clinical trial have similar illness severities. SOI scores are also useful in guiding hospital administrative policies, directing the allocation of resources such as nursing and ancillary care and assisting in assessments of quality of ICU care over time. Scoring system validations are based on the premise that age, chronic medical illnesses, and derangements from normal physiology are associated with increased mortality rates. All existing SOI scoring systems are derived from patients who have already been admitted to the ICU.

SOI scoring systems cannot be used to predict survival in individual patients. No established scoring systems that purport to direct clinicians' decision-making regarding criteria for admission to an ICU are available. Thus the use of SOI scoring systems to direct therapy and clinical decision-making cannot be recommended. Instead, these tools should be used as a source of important data to complement clinical bedside decision-making.

The most commonly utilized scoring systems are the SOFA (Sequential Organ Failure Assessment), the APACHE (Acute Physiology and Chronic Health Evaluation), and the SAPS (Simplified Acute

Physiology Score) systems.

THE SOFA SCORING SYSTEM

The SOFA scoring system is composed of scores from six organ systems, graded from 0 to 4 according to the degree of dysfunction (**Table 293-1**). The score accounts for clinical interventions; it can be measured repeatedly (i.e., each day), and rising scores correlate well with increasing mortality. Patients with suspected infection can be predicted to have poor outcomes typical of sepsis if they have at least two of the following clinical criteria: respiratory rate >22, altered mental status, or systolic blood pressure <100 mmHg. Recently, a new bedside clinical score using two or more of the above clinical criteria has emerged and is termed quickSOFA (qSOFA). qSOFA is intended to screen patients for ICU admission from out-of-hospital, emergency department, and hospital ward settings.

TABLE 293-1 Calculation of SOFA Score[a]

	SCORE				
SYSTEM	**0**	**1**	**2**	**3**	**4**
Respiration					
Pa_{O_2}/Fi_{O_2}, mmHg (kPa)	≥400 (53.3)	<400 (53.3)	<300 (40)	<200 (26.7) with respiratory support	<100 (13.3) with respiratory support
Coagulation					
Platelets, × 10^3/μL	>150	<150	<100	<50	<20
Liver					
Bilirubin, mg/dL (μmol/L)	<1.2 (20)	1.2–1.9 (20–32)	2.0–5.9 (33–101)	6.0–11.9 (102–204)	>12.0 (204)
Cardiovascular	MAP >70 mmHg	MAP <70 mmHg	Dopamine < 5 or dobutamine (any dose)[b]	Dopamine 5.1–15 or epinephrine <0.1 or norepinephrine <0.1[b]	Dopamine >15 or epinephrine >0.1 or norepinephrine >0.1[b]
Central Nervous System					
Glasgow Coma Scale[c]	15	13–14	10–12	6–9	<6
Renal					
Creatinine, mg/dL (μmol/L)	<1.2 (110)	1.2–1.9 (110–170)	2.0–3.4 (171–299)	3.5–4.9 (300–440)	>5.0 (440)
Urine output, mL/dL				<500	<200

[a]Adapted from JL Vincent, R Moreno, J Takala, et al: Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med 22(7):707, 1996.
[b]Catecholamine doses are given as μg/kg per min for at least 1 h. [c] Glasgow Coma Scale scores range from 3 to 15; higher score indicates better neurological function.

Abbreviations: Fi_{O_2}, fraction of inspired oxygen; MAP, mean arterial pressure; Pa_{O_2}, partial pressure of oxygen.

THE APACHE II SCORING SYSTEM

The APACHE II system is the most commonly used SOI scoring system in North America. Age, type of ICU admission (after elective surgery vs nonsurgical or after emergency surgery), chronic health problems, and 12 physiologic variables (the worst values for each in the first 24 h after ICU admission) are used to derive a score. The predicted hospital mortality rate is derived from a formula that takes into account the APACHE II score, the need for emergency surgery, and a weighted, disease-specific diagnostic category **(Table 293-2)**. The relationship between APACHE II score and mortality risk is illustrated in **Fig. 293-1**. Updated versions of the APACHE scoring system (APACHE III and APACHE IV) have been published.

TABLE 293-2 Calculation of Acute Physiology and Chronic Health Evaluation II (APACHE II) Score[a]

Acute Physiology Score

SCORE	4	3	2	1	0	1	2	3	4
Rectal temperature (°C)	≥41	39.0–40.9		38.5–38.9	36.0–38.4	34.0–35.9	32.0–33.9	30.0–31.9	≤29.9
Mean blood pressure (mmHg)	≥160	130–159	110–129		70–109		50–69		≤49
Heart rate (beats/min)	≥180	140–179	110–139		70–109		55–69	40–54	≤39
Respiratory rate (breaths/min)	≥50	35–49		25–34	12–24	10–11	6–9		≤5
Arterial pH	≥7.70	7.60–7.69		7.50–7.59	7.33–7.49		7.25–7.32	7.15–7.24	<7.15
Oxygenation									
If FIo_2 >0.5, use (A – a) Do_2	≥500	350–499	200–349		<200				
If FIo_2 ≤ 0.5, use Pao_2					>70	61–70		55–60	<55
Serum sodium (meq/L)	≥180	160–179	155–159	150–154	130–149		120–129	111–119	≤110
Serum potassium (meq/L)	≥7.0	6.0–6.9		5.5–5.9	3.5–5.4	3.0–3.4	2.5–2.9		<2.5
Serum creatinine (mg/dL)	≥3.5	2.0–3.4	1.5–1.9		0.6–1.4		<0.6		
Hematocrit (%)	≥60		50–59.9	46–49.9	30–45.9		20–29.9		<20
WBC count (10^3/mL)	≥40		20–39.9	15–19.9	3–14.9		1–2.9		<1

Glasgow Coma Score[b,c]

Eye Opening	*Verbal (Nonintubated)*	*Verbal (Intubated)*	*Motor Activity*
4—Spontaneous	5—Oriented and talks	5—Seems able to talk	6—Verbal command
3—Verbal stimuli	4—Disoriented and talks	3—Questionable ability to talk	5—Localizes to pain
2—Painful stimuli	3—Inappropriate words	1—Generally unresponsive	4—Withdraws from pain
1—No response	2—Incomprehensible sounds		3—Decorticate
	1—No response		2—Decerebrate
			1—No response

Points Assigned to Age and Chronic Disease

Age, Years	*Score*
<45	0
45–54	2
55–64	3
65–74	5
≥75	6

Chronic Health (History of Chronic Conditions)[d]	*Score*

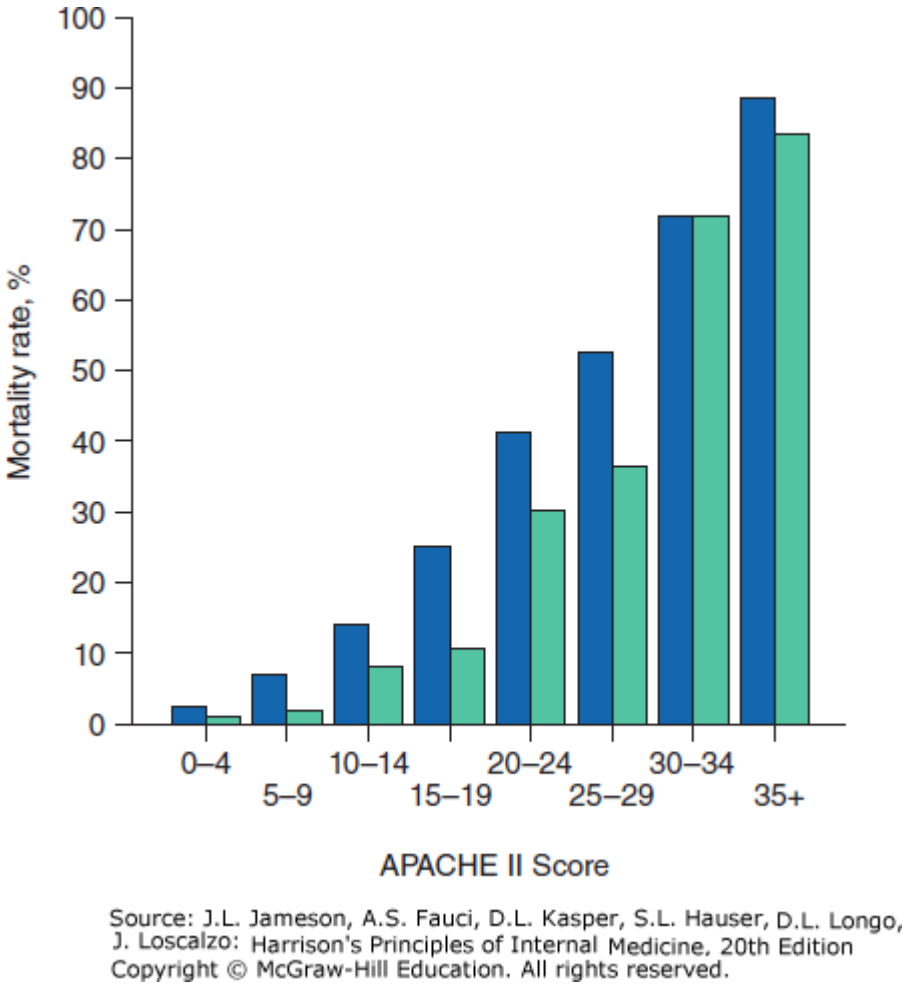
Acute Physiology Score									
SCORE	**4**	**3**	**2**	**1**	**0**	**1**	**2**	**3**	**4**
None					0				
If patient is admitted after elective surgery					2				
If patient is admitted after emergency surgery or for reason other than after elective surgery					5				

[a]The APACHE II score is the sum of the acute physiology score (vital signs, oxygenation, laboratory values), the Glasgow coma score, age, and chronic health points. The worst values during the first 24 h in the ICU should be used. [b]Glasgow coma score (GCS) = eye-opening score + verbal (intubated or nonintubated) score + motor score. [c]For GCS component of acute physiology score, subtract GCS from 15 to obtain points assigned. [d]Hepatic: cirrhosis with portal hypertension or encephalopathy; cardiovascular: class IV angina (at rest or with minimal self-care activities); pulmonary: chronic hypoxemia or hypercapnia, polycythemia, ventilator dependence; renal: chronic peritoneal or hemodialysis; immune: immunocompromised host.

Abbreviations: (A – a) Do_2, alveolar-arterial oxygen difference; FIo_2, fraction of inspired oxygen; Pao_2, partial pressure of oxygen; WBC, white blood cell count.

FIGURE 293-1

APACHE II survival curve. Blue, nonoperative; green, postoperative.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE SAPS SCORING SYSTEM

The SAPS II score, used more frequently in Europe than in the United States, was derived in a manner similar to the APACHE score. This score is not disease-specific but rather incorporates three underlying disease variables: AIDS, metastatic cancer, and hematologic malignancy. SAPS 3, which utilizes a 1-h rather than a 24-h window for measuring physiologic derangement scores, was developed in 2005.

SHOCK

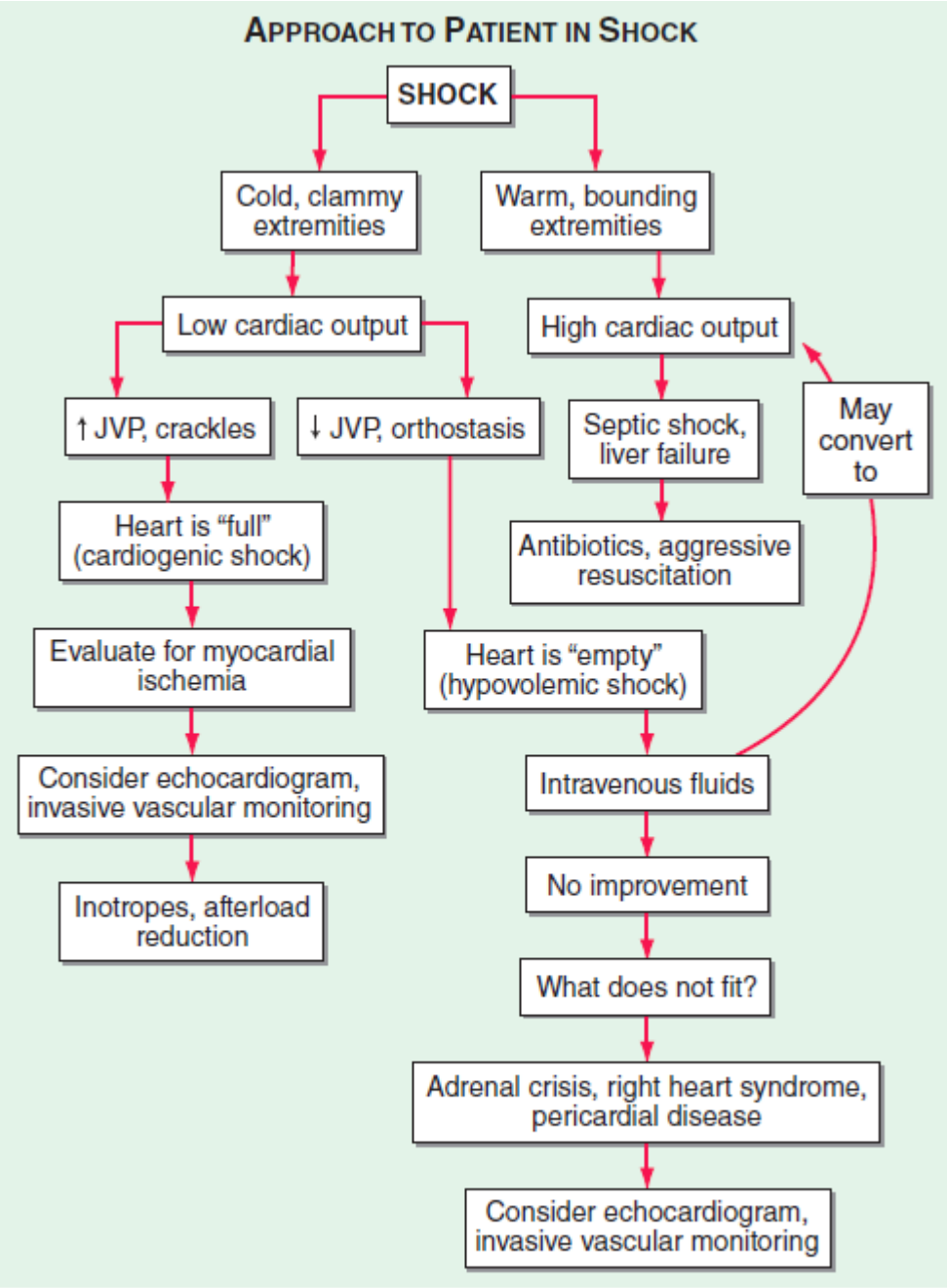
See also Chap. 296.

INITIAL EVALUATION

Shock, a common condition necessitating ICU admission or occurring in the course of critical care, is defined by the presence of multisystem end-organ hypoperfusion. Clinical indicators include reduced mean arterial pressure (MAP), tachycardia, tachypnea, cool skin and extremities, acute altered mental status, and oliguria. Hypotension is usually, though not always, present. The end result of multiorgan hypoperfusion is tissue hypoxia, often with accompanying lactic acidosis. Since the MAP is the product of cardiac output and systemic vascular resistance (SVR), reductions in blood pressure can be caused by decreases in cardiac output and/or SVR. Accordingly, once shock is contemplated, the initial evaluation of a hypotensive patient should include an early bedside assessment of the adequacy of cardiac output **(Fig. 293-2)**. Clinical evidence of *diminished* cardiac output includes a narrow pulse pressure (systolic BP minus diastolic BP)—a marker that correlates with stroke volume—and cool extremities with delayed capillary refill. Signs of *increased* cardiac output include a widened pulse pressure (particularly with a reduced diastolic pressure), warm extremities with bounding pulses, and rapid capillary refill. If a hypotensive patient has clinical signs of increased cardiac output, it can be inferred that the reduced blood pressure is from decreased SVR.

FIGURE 293-2

Approach to the patient in shock. JVP, jugular venous pressure.

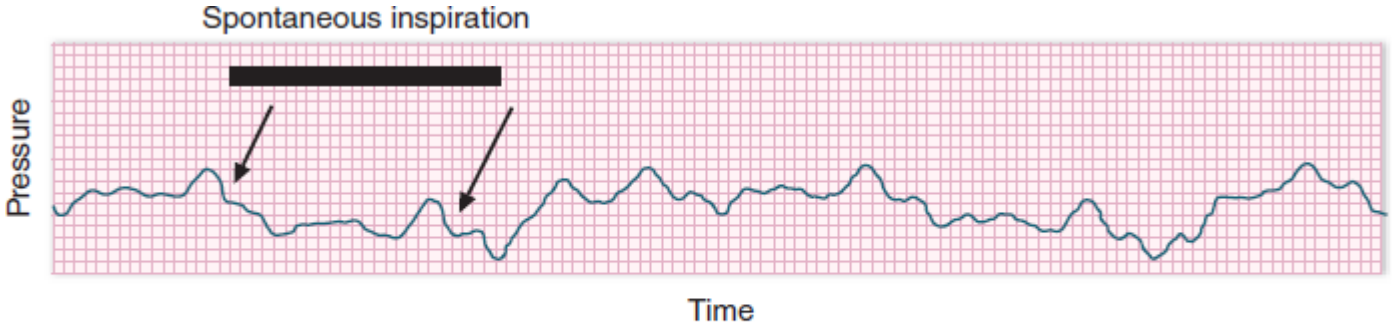


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In hypotensive patients with signs of reduced cardiac output, an assessment of intravascular volume status is appropriate. A hypotensive patient with decreased intravascular volume status may have a history suggesting hemorrhage or other volume losses (e.g., vomiting, diarrhea, polyuria). Although evidence of a reduced jugular venous pressure (JVP) is often sought, static measures of right atrial pressure do not predict fluid responsiveness reliably; the *change* in right atrial pressure as a function of spontaneous respiration is a better predictor of fluid responsiveness **(Fig. 293-3)**. Patients with fluid-responsive (i.e., hypovolemic) shock also may manifest large changes in pulse pressure as a function of respiration *during* mechanical ventilation **(Fig. 293-4)**. A hypotensive patient with increased intravascular volume and cardiac dysfunction may have S_3 and/or S_4 gallops on examination, increased JVP, extremity edema, and crackles on lung auscultation. The chest x-ray may show cardiomegaly, widening of the vascular pedicle, Kerley B lines, and pulmonary edema. Chest pain and electrocardiographic changes consistent with ischemia may be noted **(Chap. 298)**.

FIGURE 293-3

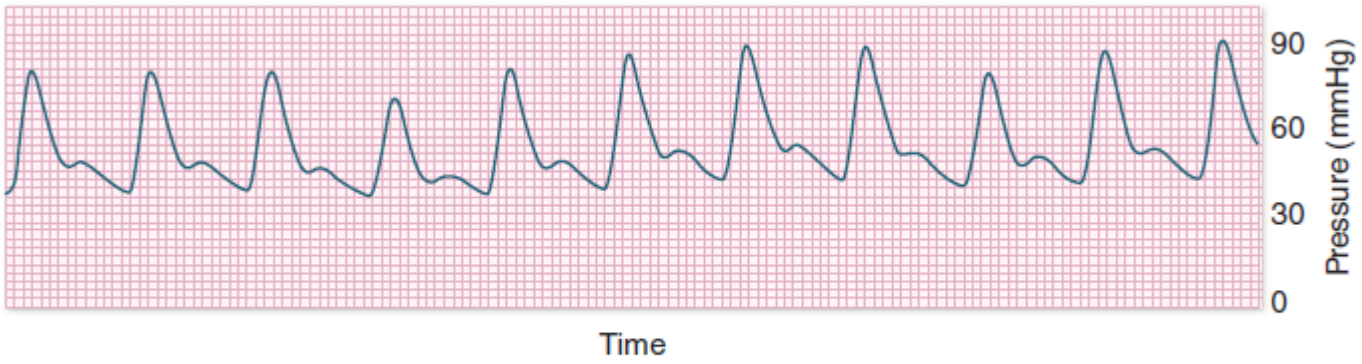
Right atrial pressure change during spontaneous respiration in a patient with shock whose cardiac output will increase in response to intravenous fluid administration. The right atrial pressure decreases from 7 mmHg to 4 mmHg. The horizontal bar marks the time of spontaneous inspiration.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 293-4

Pulse pressure change during mechanical ventilation in a patient with shock whose cardiac output will increase in response to intravenous fluid administration. The pulse pressure (systolic minus diastolic blood pressure) changes during mechanical ventilation in a patient with septic shock.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In hypotensive patients with clinical evidence of increased cardiac output, a search for causes of decreased SVR is appropriate. The most common cause of high-cardiac-output hypotension is sepsis (**Chap. 297**). Other causes include liver failure, severe pancreatitis, burns, trauma, anaphylaxis, thyrotoxicosis, and peripheral arteriovenous shunts.

In summary, the most common categories of shock are hypovolemic, cardiogenic, and high-cardiac-output with decreased SVR (high-output hypotension). Certainly more than one category can occur simultaneously (e.g., hypovolemic and septic shock).

The initial assessment of a patient in shock should take only a few minutes. It is important that aggressive resuscitation is instituted on the basis of the initial assessment, particularly since early resuscitation from septic and cardiogenic shock may improve survival (see below). If the initial bedside assessment yields equivocal or confounding data, more objective assessments such as ultrasound/echocardiography may be useful. In spontaneously breathing patients, inferior vena cava collapse seen on ultrasound predicts a fluid responsive state. Increasingly, ultrasound of the thorax and abdomen is used by intensivists as an extension of the physical examination to assess rapidly

imputed filling volumes, adequacy of cardiac performance, and for indices of other specific conditions (e.g., pericardial tamponade, pulmonary embolus, pulmonary edema, pneumothorax). The goal of aggressive resuscitation is to reestablish adequate tissue perfusion and thus to prevent or minimize end-organ injury.

MECHANICAL VENTILATORY SUPPORT

(See also Chap. 295) During the initial resuscitation of patients in shock, principles of advanced cardiac life support should be followed. As such patients may be obtunded and unable to protect the airway, an early assessment of the airway is mandatory. Early intubation and mechanical ventilation often are required. Reasons for the institution of endotracheal intubation and mechanical ventilation include acute hypoxemic respiratory failure and ventilatory failure, which frequently accompany shock. Acute hypoxemic respiratory failure may occur in patients with cardiogenic shock and pulmonary edema **(Chap. 298)** as well as in those who are in septic shock with pneumonia or acute respiratory distress syndrome (ARDS) **(Chaps. 294 and 297)**. Ventilatory failure often occurs as a consequence of an increased load on the respiratory system in the form of acute metabolic (often lactic) acidosis or decreased lung compliance due to pulmonary edema. Inadequate perfusion to respiratory muscles in the setting of shock may be another reason for early intubation and mechanical ventilation. Normally, the respiratory muscles receive a very small percentage of the cardiac output. However, in patients who are in shock with respiratory distress, the percentage of cardiac output dedicated to respiratory muscles may increase by tenfold or more. Lactic acid production from inefficient respiratory muscle activity presents an additional ventilatory load.

Mechanical ventilation may relieve the work of breathing and allow redistribution of a limited cardiac output to other vital organs. Patients demonstrate respiratory distress by an inability to speak full sentences, accessory use of respiratory muscles, paradoxical abdominal muscle activity, extreme tachypnea (>40 breaths/min), and decreasing respiratory rate despite an increasing drive to breathe due to exhaustion. When patients with shock are treated with mechanical ventilation, a major goal is for the ventilator to assume all or the majority of the work of breathing, facilitating a state of minimal respiratory muscle work. With the institution of mechanical ventilation for shock, further declines in MAP are frequently seen. The reasons include impeded venous return from positive-pressure ventilation, reduced endogenous catecholamine secretion once the stress associated with respiratory failure abates, and the actions of drugs used to facilitate endotracheal intubation (e.g., propofol, opiates). Patients with right heart dysfunction or preexiting pulmonary hypertension may also have diminished cardiac output related to the increases in right ventricular afterload resulting from positive pressure ventilation. Accordingly, hypotension should be anticipated during and following endotracheal intubation. Because many of these patients may be fluid responsive, IV volume administration should be considered. Figure 293-2 summarizes the diagnosis and treatment of different types of shock. **For further discussion of individual forms of shock, see Chaps. 296, 297, and 298.**

RESPIRATORY FAILURE

Respiratory failure is one of the most common reasons for ICU admission. In some ICUs, ≥75% of patients require mechanical ventilation during their stay. Respiratory failure can be categorized mechanistically on the basis of pathophysiologic derangements in respiratory function.

TYPE I: ACUTE HYPOXEMIC RESPIRATORY FAILURE

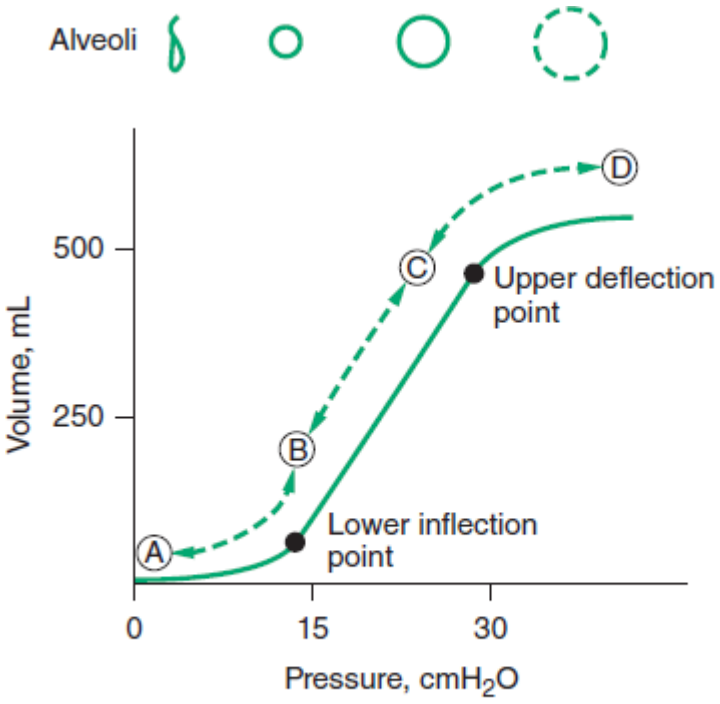
This type of respiratory failure occurs with alveolar flooding and subsequent intrapulmonary shunt physiology. Alveolar flooding may be a consequence of pulmonary edema, lung injury, pneumonia, or alveolar hemorrhage. Pulmonary edema can be further categorized as occurring due to elevated pulmonary microvascular pressures, as seen in heart failure and intravascular volume overload or ARDS (“low-pressure pulmonary edema,” **Chap. 294**). This syndrome is defined by acute onset (≤1

week) of bilateral opacities on chest imaging that are not fully explained by cardiac failure or fluid overload and of shunt physiology requiring positive end-expiratory pressure (PEEP). Type I respiratory failure occurs in clinical settings such as sepsis, gastric aspiration, pneumonia, near-drowning, multiple blood transfusions, and pancreatitis. The mortality rate among patients with ARDS was traditionally very high (50–70%), although changes in patient care have led to mortality rates closer to 30% (see below).

It is well established that mechanical ventilation of patients with ARDS may propagate lung injury. As seen in **Fig. 293-5**, the pressure-volume relationship of the lung in ARDS is not linear. Alveoli may collapse at very low lung volumes. Animal studies have suggested that stretching and overdistention of injured alveoli during mechanical ventilation can further injure the lung. Concern over this alveolar overdistention, termed *ventilator-induced* "*volutrauma*," led to a multicenter, randomized, prospective trial comparing traditional ventilator strategies for ARDS (large tidal volume: 12 mL/kg of ideal body weight) with a low tidal volume (6 mL/kg of ideal body weight). This study showed a dramatic reduction in mortality rate in the low-tidal-volume group from that in the high-tidal-volume group (31 versus 39.8%). Other studies have shown that large tidal volumes may lead to ARDS in patients who initially do not have this problem. Neuromuscular blockade and prone positioning have been shown to improve survival in those with severe ARDS. In addition, a "fluid-conservative" management strategy (maintaining a low central venous pressure [CVP] or pulmonary capillary wedge pressure [PCWP]) is associated with fewer days of mechanical ventilation than a "fluid-liberal" strategy (maintaining a relatively high CVP or PCWP) in ARDS. There is growing interest in avoiding intubation in patients with ARDS by the use of a variety of devices, such as masks, high flow oxygen delivery systems, and helmets for respiratory support; however, this is tempered by concern that higher tidal volumes during spontaneous breathing with these devices could result in progression of preexisting lung injury.

FIGURE 293-5

Pressure-volume relationship in the lungs of a patient with acute respiratory distress syndrome (ARDS). At the lower inflection point, collapsed alveoli begin to open and lung compliance changes. At the upper deflection point, alveoli become overdistended. The shape and size of alveoli are illustrated at the top of the figure.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TYPE II RESPIRATORY FAILURE

This type of respiratory failure is a consequence of alveolar hypoventilation and results from the inability to eliminate carbon dioxide effectively. Mechanisms are categorized by impaired central nervous system (CNS) drive to breathe, impaired strength with failure of neuromuscular function in the respiratory system, and increased load(s) on the respiratory system. Reasons for diminished CNS drive to breathe include drug overdose, brainstem injury, sleep-disordered breathing, and severe hypothyroidism. Reduced strength can be due to impaired neuromuscular transmission (e.g., myasthenia gravis, Guillain-Barré syndrome, amyotrophic lateral sclerosis) or respiratory muscle weakness (e.g., myopathy, electrolyte derangements, fatigue).

The overall load on the respiratory system can be subclassified into resistive loads (e.g., bronchospasm), loads due to reduced lung compliance (e.g., alveolar edema, atelectasis, intrinsic positive end-expiratory pressure [auto-PEEP]—see below), loads due to reduced chest wall compliance (e.g., pneumothorax, pleural effusion, abdominal distention), and loads due to increased minute ventilation requirements (e.g., pulmonary embolus with increased dead-space fraction, sepsis).

The mainstays of therapy for type II respiratory failure are directed at reversing the underlying cause(s) of ventilatory failure. Noninvasive positive-pressure ventilation with a tight-fitting facial or nasal mask, with avoidance of endotracheal intubation, often stabilizes these patients. This approach has been shown to be beneficial in treating patients with exacerbations of chronic obstructive pulmonary disease; it has been tested less extensively in other kinds of respiratory failure but may be attempted nonetheless in the absence of contraindications (hemodynamic instability, inability to protect the airway, respiratory arrest).

TYPE III RESPIRATORY FAILURE

This form of respiratory failure results from lung atelectasis. Because atelectasis occurs so commonly in the perioperative period, this form is also called *perioperative respiratory failure*. After general anesthesia, decreases in functional residual capacity lead to collapse of dependent lung units. Such atelectasis can be treated by frequent changes in position, chest physiotherapy, upright positioning, and control of incisional and/or abdominal pain. Noninvasive positive-pressure ventilation may also be used to reverse regional atelectasis.

TYPE IV RESPIRATORY FAILURE

This form results from hypoperfusion of respiratory muscles in patients in shock. Normally, respiratory muscles consume <5% of total cardiac output and oxygen delivery. Patients in shock often experience respiratory distress due to pulmonary edema (e.g., in cardiogenic shock), lactic acidosis, and anemia. In this setting, up to 40% of cardiac output may be distributed to the respiratory muscles. Intubation and mechanical ventilation can allow redistribution of the cardiac output away from the respiratory muscles and back to vital organs while the shock is treated.

CARE OF THE MECHANICALLY VENTILATED PATIENT

(See also Chap. 295) Whereas a thorough understanding of the pathophysiology of respiratory failure is essential for optimal patient care, recognition of a patient's readiness to be liberated from mechanical ventilation is likewise important. Several studies have shown that daily spontaneous breathing trials can identify patients who are ready for extubation. Accordingly, all intubated, mechanically ventilated patients should undergo daily screening of respiratory function. If oxygenation is stable (i.e., Pa_{O_2}/FI_{O_2} [partial pressure of oxygen/fraction of inspired oxygen] >200 and PEEP ≤5 cmH_2O), cough and airway reflexes are intact, and no vasopressor agents or sedatives are being

administered, the patient has passed the screening test and should undergo a spontaneous breathing trial. This trial consists of a period of breathing through the endotracheal tube without ventilator support (continuous positive airway pressure [CPAP] of 5 cmH_2O with or without low level pressure support [e.g., 5 cmH_2O] and an open T-piece breathing system have all been validated) for 30–120 min. The spontaneous breathing trial is declared a failure and stopped if *any* of the following occur: (1) respiratory rate >35/min for >5 min, (2) O_2 saturation <90%, (3) heart rate >140/min or a 20% increase or decrease from baseline, (4) systolic blood pressure <90 mmHg or >180 mmHg, or (5) increased anxiety or diaphoresis. If, at the end of the spontaneous breathing trial, none of the above events has occurred and the ratio of the respiratory rate and tidal volume in liters (f/V_T) is <105, the patient can be extubated. Such protocol-driven approaches to patient care can have an important impact on the duration of mechanical ventilation and ICU stay. In spite of such a careful approach to liberation from mechanical ventilation, up to 10% of patients develop respiratory distress after extubation and may require resumption of mechanical ventilation. Many of these patients will require reintubation. The use of noninvasive ventilation in patients in whom extubation fails may be associated with worse outcomes than are obtained with immediate reintubation.

Mechanically ventilated patients frequently require sedatives and analgesics. Opiates are the mainstay of therapy for analgesia in mechanically ventilated patients. After adequate pain control has been ensured, additional indications for sedation include anxiolysis; treatment of subjective dyspnea; reduction of autonomic hyperactivity, which may precipitate myocardial ischemia; and reduction of total O_2 consumption (V_{O_2}). Non-benzodiazepine sedatives are preferred since benzodiazepines are associated with worse patient outcomes.

The neuromuscular blocking agent cisatracurium is occasionally used to facilitate mechanical ventilation in patients with profound ventilator dyssynchrony despite optimal sedation, particularly in the setting of severe ARDS. Use of these agents may result in prolonged weakness—a myopathy known as the *postparalytic syndrome*. For this reason, neuromuscular blocking agents typically are used as a last resort when aggressive sedation fails to achieve patient-ventilator synchrony. Because neuromuscular blocking agents result in pharmacologic paralysis without altering mental status, sedative-induced amnesia is mandatory when these agents are administered.

Amnesia can be achieved reliably with propofol and benzodiazepines such as lorazepam and midazolam. Outside the setting of pharmacologic paralysis, few data support the idea that amnesia is mandatory in all patients who require intubation and mechanical ventilation. Since many of these critical patients have impaired hepatic and renal function, sedatives and opiates may accumulate when given for prolonged periods. A nursing protocol–driven approach to sedation of mechanically ventilated patients or daily interruption of sedative infusions paired with daily spontaneous breathing trials has been shown to prevent excessive drug accumulation and shorten the duration of both mechanical ventilation and ICU stay.

MULTIORGAN SYSTEM FAILURE

Multiorgan system failure, which is commonly associated with critical illness, is defined by the simultaneous presence of physiologic dysfunction and/or failure of two or more organs. Typically, this syndrome occurs in the setting of severe sepsis, shock of any kind, severe inflammatory conditions such as pancreatitis, and trauma. The fact that multiorgan system failure occurs commonly in the ICU is a testament to our current ability to stabilize and support single-organ failure. The ability to support single-organ failure aggressively (e.g., by mechanical ventilation or by renal replacement therapy) has reduced rates of early mortality in critical illness. As a result, it is uncommon for critically ill patients to die in the initial stages of resuscitation. Instead, many patients succumb to critical illness later in the ICU stay, after the initial presenting problem has been stabilized.

Although there is debate regarding specific definitions of organ failure, several general principles governing the syndrome of multiorgan system failure apply. First, organ failure, no matter how it is defined, must persist beyond 24 h. Second, mortality risk increases with the accrual of failing organs. Third, the prognosis worsens with increased duration of organ failure. These observations remain true across various critical care settings (e.g., medical versus surgical).

MONITORING IN THE ICU

Because respiratory failure and circulatory failure are common in critically ill patients, monitoring of the respiratory and cardiovascular systems is undertaken frequently. Evaluation of respiratory gas exchange is routine in critical illness. The “gold standard” remains arterial blood-gas analysis, in which pH, Pa_{O_2}, partial pressure of carbon dioxide (P_{CO_2}), and O_2 saturation are measured directly. With arterial blood-gas analysis, the two main functions of the lung—oxygenation of arterial blood and elimination of CO_2—can be assessed directly. In fact, the blood pH, which has a profound effect on the drive to breathe, can be assessed only by such sampling. Although sampling of arterial blood is generally safe, it may be painful and cannot provide continuous information. In light of these limitations, noninvasive monitoring of respiratory function is often employed.

PULSE OXIMETRY

The most commonly utilized noninvasive technique for monitoring respiratory function, pulse oximetry takes advantage of differences in the absorptive properties of oxygenated and deoxygenated hemoglobin. At wavelengths of 660 nm, oxyhemoglobin reflects light more effectively than does deoxyhemoglobin, whereas the reverse is true in the infrared spectrum (940 nm). A pulse oximeter passes both wavelengths of light through a perfused digit such as a finger, and the relative intensity of light transmission at these two wavelengths is recorded. From this information, the relative percentage of oxyhemoglobin is derived. Since arterial pulsations produce phasic changes in the intensity of transmitted light, the pulse oximeter is designed to detect only light of alternating intensity. This feature allows distinction of arterial and venous blood O_2 saturations.

RESPIRATORY SYSTEM MECHANICS

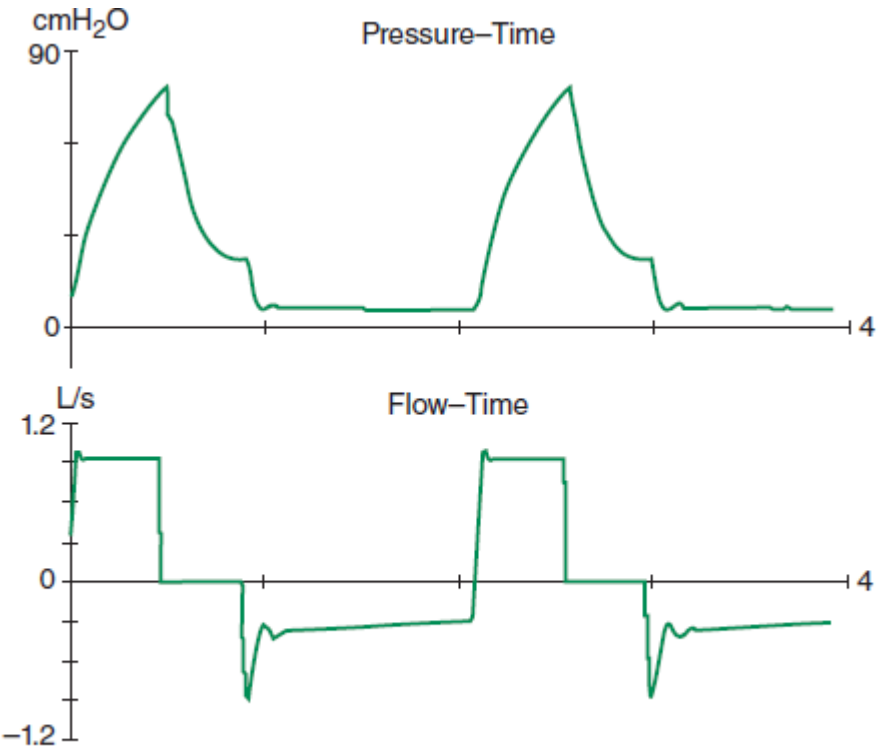
Respiratory system mechanics can be measured in patients during mechanical ventilation **(Chap. 295)**. When volume-controlled modes of mechanical ventilation are used, accompanying airway pressures can easily be measured as long as the patient is passive. The peak airway pressure is determined by two variables: airway resistance and respiratory system compliance. At the end of inspiration, inspiratory flow can be stopped transiently. This end-inspiratory pause (*plateau pressure*) is a static measurement, affected only by respiratory system compliance and not by airway resistance. Therefore, during volume-controlled ventilation, the difference between the peak (airway resistance + respiratory system compliance) and plateau (respiratory system compliance only) airway pressures provides a quantitative assessment of airway resistance. Accordingly, during volume-controlled ventilation, patients with increases in airway resistance typically have increased peak airway pressures as well as abnormally high gradients between peak and plateau airway pressures (typically >15 cmH_2O) at a constant inspiratory flow rate of 1 L/sec. The compliance of the respiratory system is defined by the change in volume of the respiratory system per unit change in pressure.

The respiratory system can be divided into two components: the lungs and the chest wall. Normally, respiratory system compliance is ~100 mL/cmH_2O. Pathophysiologic processes such as pleural effusions, pneumothorax, and increased abdominal girth all reduce chest wall compliance. Lung compliance may be reduced by pneumonia, pulmonary edema, interstitial lung disease, or auto-PEEP. Accordingly, patients with abnormalities in compliance of the respiratory system (lungs and/or chest wall) typically have elevated peak *and* plateau airway pressures but a normal gradient between these

two pressures. Auto-PEEP occurs when there is insufficient time for emptying of alveoli before the next inspiratory cycle. Since the alveoli have not decompressed completely, alveolar pressure remains positive at the end of exhalation (*functional residual capacity*). This phenomenon results most commonly from obstruction of distal airways in disease processes such as asthma and COPD. Auto-PEEP with resulting alveolar overdistention may result in diminished lung compliance, reflected by abnormally increased plateau airway pressures. Modern mechanical ventilators allow breath-to-breath display of pressure and flow, permitting detection of problems such as patient-ventilator dyssynchrony, airflow obstruction, and auto-PEEP **(Fig. 293-6)**.

FIGURE 293-6

Increased airway resistance with auto-PEEP. The top waveform (airway pressure vs. time) shows a large difference between the peak airway pressure (80 cmH_2O) and the plateau airway pressure (20 cmH_2O). The bottom waveform (flow vs. time) demonstrates airflow throughout expiration (reflected by the flow tracing on the negative portion of the abscissa) that persists up to the next inspiratory effort.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CIRCULATORY STATUS

Oxygen delivery (Qo_2) is a function of cardiac output and the content of O_2 in the arterial blood (Cao_2). The Cao_2 is determined by the hemoglobin concentration, the arterial hemoglobin saturation, and dissolved O_2 not bound to hemoglobin. For normal adults:

$$\begin{aligned} Qo_2 &= 50\ \text{dL/min} \times (1.39 \times 15\ \text{g/dL [hemoglobin concentration]} \\ &\quad \times 1.0\ \text{[hemoglobin \% saturation]} + 0.0031 \times 100\ [Pao_2]) \\ &= 50\ \text{dL/min (cardiac output)} \times 21.6\ \text{mL } O_2\ \text{per dL blood } (Cao_2) \\ &= 1058\ \text{mL } O_2\ \text{per min} \end{aligned}$$

It is apparent that nearly all of the O_2 delivered to tissues is bound to hemoglobin and that the dissolved O_2 (Pa_{O_2}) contributes very little to O_2 content in arterial blood or to O_2 delivery. Normally, the content of O_2 in mixed venous blood ($C^{-}v_{O_2}$) is 15.76 mL/dL since the mixed venous blood is 75% saturated. Therefore, the normal tissue extraction ratio for O_2 is $Ca_{O_2} - C^{-}v_{O_2}/Ca_{O_2}$ ([21.16 – 15.76]/21.16) or ~25%. A pulmonary artery catheter allows measurements of O_2 delivery and the O_2 extraction ratio.

Information on the venous O_2 saturation allows assessment of global tissue perfusion. A reduced venous O_2 saturation may be caused by inadequate cardiac output, reduced hemoglobin concentration, and/or reduced arterial O_2 saturation. An abnormally high V_{O_2} may also lead to a reduced venous O_2 saturation if O_2 delivery is not concomitantly increased. Abnormally increased V_{O_2} in peripheral tissues may be caused by problems such as fever, agitation, shivering, and thyrotoxicosis.

The pulmonary artery catheter originally was designed as a tool to guide therapy for acute myocardial infarction but has been used in the ICU for evaluation and treatment of a variety of other conditions, such as ARDS, septic shock, congestive heart failure, and acute renal failure. This device has never been validated as a tool associated with reduction in morbidity and mortality rates. Indeed, despite numerous prospective studies, mortality or morbidity rate benefits associated with use of the pulmonary artery catheter have never been reported in any setting. Accordingly, it appears that routine pulmonary artery catheterization is not indicated as a means of monitoring and characterizing circulatory status in most critically ill patients.

Static measurements of circulatory parameters (e.g., *CVP,* PCWP) do not provide reliable information on the circulatory status of critically ill patients. In contrast, dynamic assessments measuring the impact of breathing on the circulation are more reliable predictors of responsiveness to IV fluid administration. A decrease in CVP of >1 mmHg during inspiration in a spontaneously breathing patient may predict an increase in cardiac output after IV fluid administration. Similarly, a changing pulse pressure during mechanical ventilation of a passive patient has been shown to predict an increase in cardiac output after IV fluid administration, assuming the R-R interval is stable.

PREVENTION OF COMPLICATIONS OF CRITICAL ILLNESS

SEPSIS IN THE CRITICAL CARE UNIT

(See also Chap. 297) Sepsis, is defined as life-threatening organ dysfunction (i.e., an increase in Sequential Organ Failure Assessment [SOFA] of 2 points or more) caused by a dysregulated response to infection. Poor outcomes can be anticipated in patients with 2 or more of the following: respiratory rate >22 per min, altered mentation, systolic blood pressure <100 mmHg. Sepsis is a leading cause of death in noncoronary ICUs in the United States, with case rates expected to increase as the population ages and a higher percentage of people are vulnerable to infection.

NOSOCOMIAL INFECTIONS IN THE ICU

Many therapeutic interventions in the ICU are invasive and predispose patients to infectious complications. These interventions include endotracheal intubation, indwelling vascular catheters, transurethral bladder catheters, and other catheters placed into sterile body cavities (e.g., tube thoracostomy, percutaneous intraabdominal drainage catheterization). The longer such devices remain in place, the more prone to these infections patients become. For example, ventilator-associated events such as ventilator-associated pneumonia correlate strongly with the duration of intubation and mechanical ventilation. Therefore, an important aspect of preventive care is the timely

removal of invasive devices as soon as they are no longer needed. Moreover, multidrug-resistant organisms are commonplace in the ICU.

Infection control is critical in the ICU. Care bundles, which include measures such as frequent hand washing, are effective but underutilized strategies. Other components of care bundles, such as protective isolation of patients colonized or infected by drug-resistant organisms, are also commonly used. Silver-coated endotracheal tubes reportedly reduce the incidence of ventilator-associated pneumonia. Studies evaluating multifaceted, evidence-based strategies to decrease catheter-related bloodstream infections have shown improved outcomes with strict adherence to measures such as hand washing, full-barrier precautions during catheter insertion, chlorhexidine skin preparation, avoidance of the femoral site, and timely catheter removal.

DEEP-VENOUS THROMBOSIS (DVT)

(See also Chap. 273) All ICU patients are at high risk for this complication because of their predilection for immobility. Therefore, all should receive some form of prophylaxis against DVT. The most commonly employed forms of prophylaxis are subcutaneous low-dose heparin injections and sequential compression devices for the lower extremities. Observational studies report an alarming incidence of DVTs despite the use of these standard prophylactic regimens. Furthermore, heparin prophylaxis may result in heparin-induced thrombocytopenia, another nosocomial complication in critically ill patients.

Low-molecular-weight heparins such as enoxaparin are more effective than unfractionated heparin for DVT prophylaxis in high-risk patients (e.g., those undergoing orthopedic surgery) and are associated with a lower incidence of heparin-induced thrombocytopenia. Fondaparinux, a selective factor Xa inhibitor, is even more effective than enoxaparin in high-risk orthopedic patients.

STRESS ULCERS

Prophylaxis against stress ulcers is not necessary for all ICU patients. It should only be administered to high-risk patients, such as those with coagulopathy or respiratory failure. Histamine receptor-2 antagonists are preferred over proton pump inhibitors because the latter are associated with increased incidence of *C. difficile* colitis and pneumonia.

NUTRITION AND GLYCEMIC CONTROL

These are important issues that may be associated with respiratory failure, impaired wound healing, and dysfunctional immune response in critically ill patients. Early enteral feeding is reasonable, with some data suggesting that permissive underfeeding of nonprotein calories is not inferior to full goal feeding. Certainly, enteral feeding, if possible, is preferred over parenteral nutrition, which is associated with numerous complications, including hyperglycemia, fatty liver, cholestasis, and sepsis. When parenteral feeding is necessary to supplement enteral nutrition, delaying this intervention until day 8 in the ICU results in better recovery and fewer ICU-related complications. Tight glucose control is an area of controversy in critical care. Although one study showed a significant mortality benefit when glucose levels were aggressively normalized in a large group of surgical ICU patients, other studies of both medical and surgical ICU patients suggested that tight glucose control resulted in increased rates of mortality.

ICU-ACQUIRED WEAKNESS

ICU-acquired weakness occurs frequently in patients who survive critical illness, particularly those with SIRS and/or sepsis. Both neuropathies and myopathies have been described, most commonly after ~1 week in the ICU. The mechanisms behind ICU-acquired weakness syndromes are poorly

understood, they are known to present with heterogeneous muscle pathophysiology. Intensive insulin therapy may reduce polyneuropathy in critical illness. Very early physical and occupational therapy in mechanically ventilated patients reportedly results in significant improvements in functional independence at hospital discharge as well as in reduced durations of mechanical ventilation and delirium.

ANEMIA

Studies have shown that most ICU patients are anemic as a result of chronic inflammation. Phlebotomy also contributes to ICU anemia. A large multicenter study involving patients in many different ICU settings challenged the conventional notion that a hemoglobin level of 100 g/L (10 g/dL) is needed in critically ill patients, with similar outcomes noted in those whose transfusion trigger was 7 g/dL. Red blood cell transfusion is associated with impairment of immune function and increased risk of infections as well as of ARDS and volume overload, all of which may explain the findings in this study. A conservative transfusion strategy has shown similar outcomes in septic shock, post-cardiac surgery, and post-hip surgery patients. A conservative transfusion strategy has been shown to enhance survival among patients with active upper gastrointestinal hemorrhage.

ACUTE KIDNEY FAILURE

(See also Chap. 304) Acute kidney failure occurs in a significant percentage of critically ill patients. The most common underlying etiology is acute tubular necrosis, usually precipitated by hypoperfusion and/or nephrotoxic agents. Currently, no pharmacologic agents are available for prevention of kidney injury in critical illness. Studies have shown convincingly that neither low-dose dopamine, fenoldapam nor vasopressin are *not* effective in protecting the kidneys from acute injury.

NEUROLOGIC DYSFUNCTION IN CRITICALLY ILL PATIENTS

DELIRIUM

(See also Chaps. 24 and 300) This state is defined by (1) an acute onset of changes or fluctuations in mental status, (2) inattention, (3) disorganized thinking, and (4) an altered level of consciousness (i.e., a state other than alertness). Delirium is reported to occur in a wide range of mechanically ventilated ICU patients and can be detected by the Confusion Assessment Method (CAM)-ICU or the Intensive Care Delirium Screening Checklist. These tools are used to ask patients to answer simple questions and perform simple tasks and can be used readily at the bedside. The differential diagnosis of delirium in ICU patients is broad and includes infectious etiologies (including sepsis), medications (particularly sedatives and analgesics), drug withdrawal, metabolic/electrolyte derangements, intracranial pathology (e.g., stroke, intracranial hemorrhage), seizures, hypoxia, hypertensive crisis, shock, and vitamin deficiencies (particularly thiamine). The etiology of a patient's ICU delirium impacts the prognosis. Those with persistent ICU delirium not related to sedatives have increases in length of hospital stay, time on mechanical ventilation, cognitive impairment at hospital discharge, and 6-month mortality rate. Interventions to reduce ICU delirium are limited. The sedative dexmedetomidine has been less strongly associated with ICU delirium than midazolam. In addition, very early physical and occupational therapy in mechanically ventilated patients has been demonstrated to reduce delirium.

ANOXIC CEREBRAL INJURY

(See also Chap. 301) This condition is common after cardiac arrest and often results in severe and permanent brain injury in survivors. Active cooling of patients after cardiac arrest is controversial, with some studies showing improved neurologic outcomes and others showing no such improvement. Certainly patients suffering cardiac arrest should have a temperature targeted to no higher than 36°C.

STROKE

(See also Chap. 419) Stroke is a common cause of neurologic critical illness. Hypertension must be managed carefully, since abrupt reductions in blood pressure may be associated with further brain ischemia and injury. Acute ischemic stroke treated with tissue plasminogen activator (tPA) has an improved neurologic outcome when treatment is given within 4.5 h of onset of symptoms. The mortality rate is not reduced when tPA is compared with placebo, despite the improved neurologic outcome. The risk of cerebral hemorrhage is significantly higher in patients given tPA. No benefit is seen when tPA therapy is given beyond 4.5 h after symptom onset. Heparin has not been convincingly shown to improve outcomes in patients with acute ischemic stroke. Decompressive craniectomy is a surgical procedure that relieves increased intracranial pressure in the setting of space-occupying brain lesions or brain swelling from stroke; available evidence suggests that this procedure may improve survival among select patients (≤55 years or age), albeit at a cost of increased disability for some.

SUBARACHNOID HEMORRHAGE

(See also Chap. 419) Subarachnoid hemorrhage may occur secondary to aneurysm rupture and is often complicated by cerebral vasospasm, re-bleeding, and hydrocephalus. Vasospasm can be detected by either transcranial Doppler assessment or cerebral angiography; it is typically treated with the calcium channel blocker nimodipine, aggressive IV fluid administration, and therapy aimed at increasing blood pressure, typically with vasoactive drugs such as phenylephrine. The IV fluids and vasoactive drugs (hypertensive hypervolemic therapy) are used to overcome the cerebral vasospasm. Early surgical clipping or endovascular coiling of aneurysms is advocated to prevent complications related to re-bleeding. Hydrocephalus, typically heralded by a decreased level of consciousness, may require ventriculostomy drainage.

STATUS EPILEPTICUS

(See also Chap. 418) Recurrent or relentless seizure activity is a medical emergency. Cessation of seizure activity is required to prevent irreversible neurologic injury. Lorazepam is the most effective benzodiazepine for treating status epilepticus and is the treatment of choice for controlling seizures acutely. Phenytoin or fosphenytoin should be given concomitantly since lorazepam has a short half-life. Other drugs, such as gabapentin, carbamazepine, and phenobarbital, should be reserved for patients with contraindications to phenytoin (e.g., allergy or pregnancy) or ongoing seizures despite phenytoin.

BRAIN DEATH

(See also Chap. 301) Although deaths of critically ill patients usually are attributable to irreversible cessation of circulatory and respiratory function, a diagnosis of death also may be established by irreversible cessation of all functions of the entire brain, including the brainstem, even if circulatory and respiratory functions remain intact on artificial life support. Such a diagnosis requires demonstration of the absence of cerebral function (no response to any external stimulus) and brainstem functions (e.g., unreactive pupils, lack of ocular movement in response to head turning or ice-water irrigation of ear canals, positive apnea test [no drive to breathe]). Absence of brain function must have an established cause and be permanent without possibility of recovery; a sedative effect, hypothermia, hypoxemia, neuromuscular paralysis, and severe hypotension must be ruled out. If there is uncertainty about the cause of coma, studies of cerebral blood flow and electroencephalography should be performed.

WITHHOLDING OR WITHDRAWING CARE

(See also Chap. 9) Withholding or withdrawal of care occurs commonly in the ICU setting. The Task Force on Ethics of the Society of Critical Care Medicine reported that it is ethically sound to withhold or withdraw care if a patient or the patient's surrogate makes such a request or if the physician judges that the goals of therapy are not achievable. Since all medical treatments are justified by their expected benefits, the loss of such an expectation justifies the act of withdrawing or withholding such treatment; these two actions are judged to be fundamentally similar. An underlying stipulation derived from this report is that an informed patient should have his or her wishes respected with regard to life-sustaining therapy. Implicit in this stipulation is the need to ensure that patients are thoroughly and accurately informed regarding the plausibility and expected results of various therapies.

The act of informing patients and/or surrogate decision-makers is the responsibility of the physician and other health care providers. If a patient or surrogate desires therapy deemed futile by the treating physician, the physician is not obligated ethically to provide such treatment. Rather, arrangements may be made to transfer the patient's care to another care provider. Whether the decision to withdraw life support should be initiated by the physician or left to surrogate decision-makers alone is not clear. One study reported that slightly more than half of surrogate decision-makers preferred to receive such a recommendation, whereas the rest did not. Critical care providers should meet regularly with patients and/or surrogates to discuss prognosis when the withholding or withdrawal of care is being considered. After a consensus among caregivers has been reached, this information should be relayed to the patient and/or surrogate decision-maker. If a decision to withhold or withdraw life-sustaining care for a patient has been made, aggressive attention to analgesia and anxiolysis is needed.

FURTHER READING

Barr J et al: Clinical practice guidelines for the management of pain, agitation, and delirium in adult patients in the intensive care unit. Crit Care Med. 41:263, 2013. [PubMed: 23269131]
Girard TD et al: An Official American Thoracic Society/American College of Chest Physicians Clinical Practice Guideline: Liberation from Mechanical Ventilation in Critically Ill Adults. Rehabilitation Protocols, Ventilator Liberation Protocols, and Cuff Leak Tests. Am J Respir Crit Care Med 195:120, 2017. [PubMed: 27762595]
Guerin C et al: Prone positioning in severe acute respiratory distress syndrome. N Engl J Med 368:2159, 2013. [PubMed: 23688302]
MacLaren R et al: Histamine-2 receptor antagonists vs proton pump inhibitors on gastrointestinal tract hemorrhage and infectious complications in the intensive care unit. JAMA Intern Med 174:564, 2014. [PubMed: 24535015]
Moore CL, Copel JA: Point-of-care ultrasonography. N Engl J Med 364:749, 2011. [PubMed: 21345104]
Papazian L et al: Neuromuscular blockers in early acute respiratory distress syndrome. N Engl J Med 363:1107, 2010. [PubMed: 20843245]
Singer M et al: The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 315:801, 2016. [PubMed: 26903338]

Chapter 294: Acute Respiratory Distress Syndrome

Rebecca M. Baron; Bruce D. Levy

INTRODUCTION

Acute respiratory distress syndrome (ARDS) is a clinical syndrome of severe dyspnea of rapid onset, hypoxemia, and diffuse pulmonary infiltrates leading to respiratory failure. ARDS is caused by diffuse lung injury from many underlying medical and surgical disorders. The lung injury may be direct, as occurs in toxic inhalation, or indirect, as occurs in sepsis **(Table 294-1)**. The clinical features of ARDS are listed in **Table 294-2**. By expert consensus, ARDS is defined by three categories based on the degrees of hypoxemia (Table 294-2). These stages of mild, moderate, and severe ARDS are associated with mortality risk and with the duration of mechanical ventilation in survivors.

Table 294-1 Clinical Disorders Commonly Associated with ARDS

Direct Lung Injury	Indirect Lung Injury
Pneumonia	Sepsis
Aspiration of gastric contents	Severe trauma
Pulmonary contusion	Multiple bone fractures
Near-drowning	Flail chest
Toxic inhalation injury	Head trauma
	Burns
	Multiple transfusions
	Drug overdose
	Pancreatitis
	Postcardiopulmonary bypass

TABLE 294-2 Diagnostic Criteria for ARDS

Severity: Oxygenation[a]	Onset	Chest Radiograph	Absence of Left Atrial Hypertension
Mild: 200 mmHg < PaO_2/FIO_2 ≤ 300 mmHg *Moderate:* 100 mmHg < PaO_2/FIO_2 ≤ 200 mmHg *Severe:* PaO_2/FIO_2 ≤ 100 mmHg	Acute: Within 1 week of a clinical insult or new or worsening respiratory symptoms.	Bilateral opacities consistent with pulmonary edema not fully explained by effusions, lobar/lung collapse, or nodules	Hydrostatic edema is not the primary cause of respiratory failure. If no ARDS risk factor is present, then some objective evaluation is required (e.g., echocardiography) to rule out hydrostatic edema

[a]As assessed on at least 5 cm H_2O of positive end expiratory pressure (PEEP).

Abbreviations: ARDS, acute respiratory distress syndrome; FIO_2, inspired O_2 percentage; PaO_2, arterial partial pressure of O_2; PCWP, pulmonary capillary wedge pressure.

The annual incidence of ARDS is estimated to be as high as 60 cases/100,000 population. Approximately 10% of all intensive care unit (ICU) admissions involve patients with ARDS.

ETIOLOGY

While many medical and surgical illnesses have been associated with the development of ARDS, most cases (>80%) are caused by a relatively small number of clinical disorders: pneumonia and sepsis (~40–60%), followed in incidence by aspiration of gastric contents, trauma, multiple transfusions, and drug overdose. Among patients with trauma, the most frequently reported surgical conditions in ARDS are pulmonary contusion, multiple bone fractures, and chest wall trauma/flail chest, whereas head trauma, near-drowning, toxic inhalation, and burns are rare causes. The risks of developing ARDS are increased in patients with more than one predisposing medical or surgical condition.

Several other clinical variables have been associated with the development of ARDS. These include older age, chronic alcohol abuse, metabolic acidosis, pancreatitis, and severity of critical illness. Trauma patients with an Acute Physiology and Chronic Health Evaluation (APACHE) II score ≥16 **(Chap. 293)** have a 2.5-fold increased risk of developing ARDS.

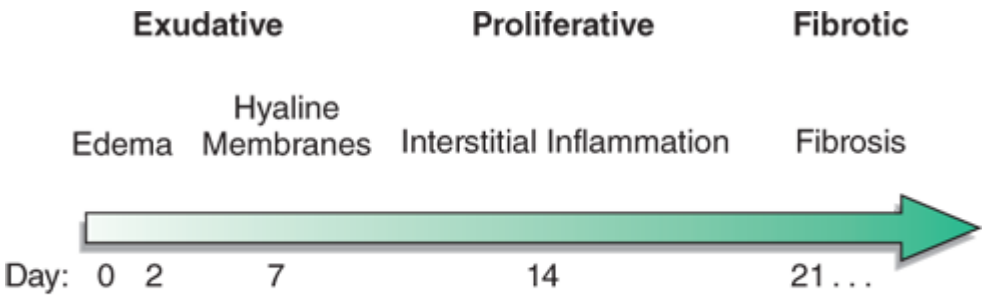
CLINICAL COURSE AND PATHOPHYSIOLOGY

The natural history of ARDS is marked by three phases—exudative, proliferative, and fibrotic—that each have characteristic clinical and pathologic features **(Fig. 294-1)**.

FIGURE 294-1

Diagram illustrating the time course for the development and resolution of ARDS. The exudative phase is notable for early alveolar edema and neutrophil-rich leukocytic infiltration of the lungs, with subsequent formation of hyaline membranes from diffuse alveolar damage. Within 7 days, a

proliferative phase ensues with prominent interstitial inflammation and early fibrotic changes. Approximately 3 weeks after the initial pulmonary injury, most patients recover. However, some patients enter the fibrotic phase, with substantial fibrosis and bullae formation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Exudative Phase

In this phase, alveolar capillary endothelial cells and type I pneumocytes (alveolar epithelial cells) are injured, with consequent loss of the normally tight alveolar barrier to fluid and macromolecules. Edema fluid that is rich in protein accumulates in the interstitial and alveolar spaces **(Fig. 294-2)**. Pro-inflammatory cytokines (e.g., interleukin 1, interleukin 8, and tumor necrosis factor α [TNF-α]) and lipid mediators (e.g., leukotriene B_4) are increased in this acute phase, leading to the recruitment of leukocytes (especially neutrophils) into the pulmonary interstitium and alveoli. In addition, condensed plasma proteins aggregate in the air spaces with cellular debris and dysfunctional pulmonary surfactant to form hyaline membrane whorls. Pulmonary vascular injury also occurs early in ARDS, with vascular obliteration by microthrombi and fibrocellular proliferation **(Fig. 294-3)**.

FIGURE 294-2

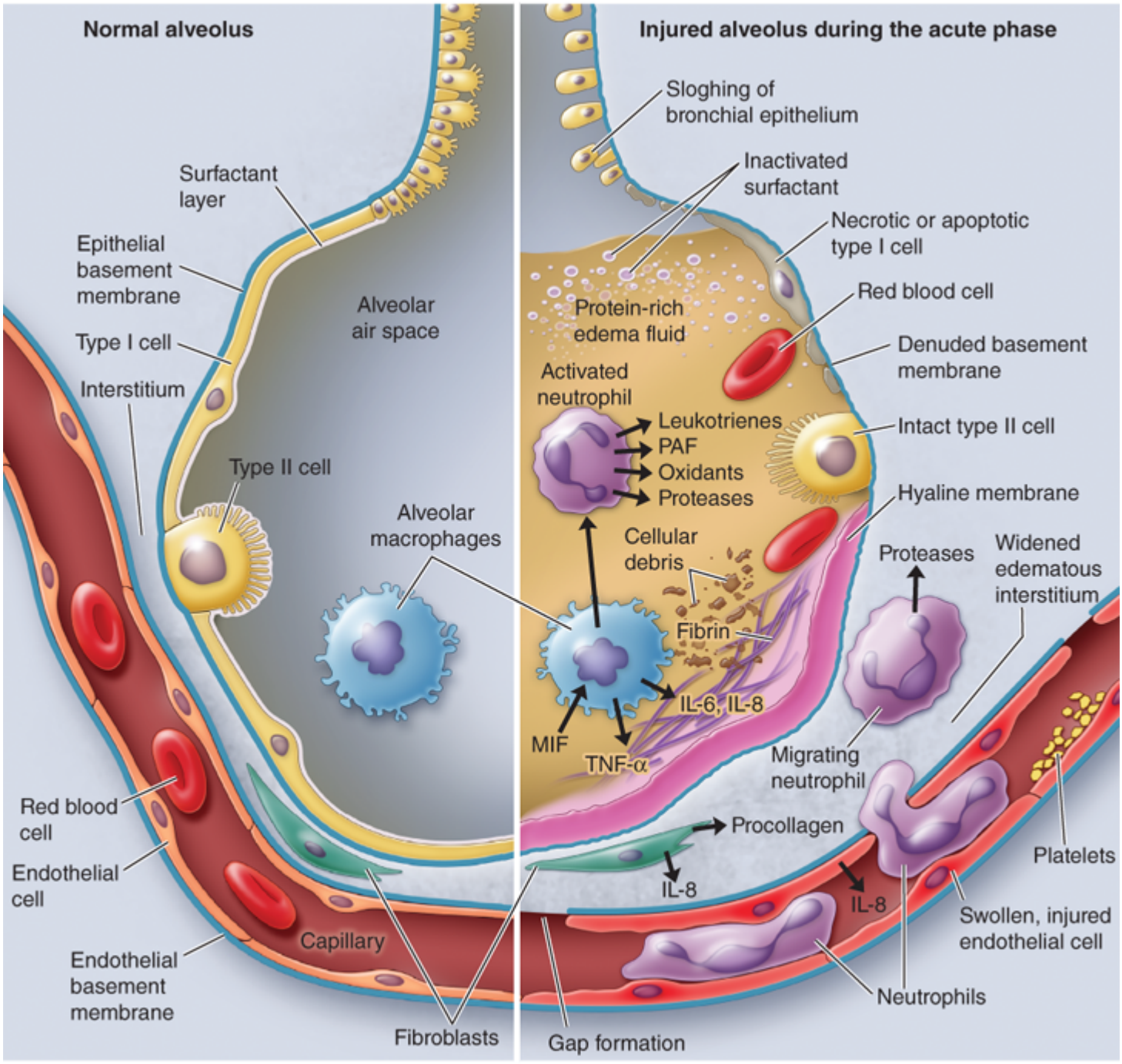
A representative anteroposterior chest x-ray in the exudative phase of ARDS shows bilateral opacities consistent with pulmonary edema that can be difficult to distinguish from left ventricular failure.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 294-3

The normal alveolus (*left*) and the injured alveolus in the acute phase of acute lung injury and the acute respiratory distress syndrome (*right*). In the acute phase of the syndrome (*right*), there is sloughing of both the bronchial and alveolar epithelial cells, with the formation of protein-rich hyaline membranes on the denuded basement membrane. Neutrophils are shown adhering to the injured capillary endothelium and transmigrating through the interstitium into the air space, which is filled with protein-rich edema fluid. In the air space, an alveolar macrophage is secreting pro-inflammatory cytokines—i.e., interleukins 1, 6, 8 (IL-1, 6, 8) and tumor necrosis factor α (TNF-α)—that act locally to stimulate chemotaxis and activate neutrophils. IL-1 can also stimulate the production of extracellular matrix by fibroblasts. Neutrophils can release oxidants, proteases, leukotrienes, and other proinflammatory molecules, such as platelet-activating factor (PAF). A number of anti-inflammatory mediators are also present in the alveolar milieu, including the IL-1-receptor antagonist, soluble TNF-α receptor, autoantibodies to IL-8, and cytokines such as IL-10 and IL-11 (not shown). The influx of protein-rich edema fluid into the alveolus can lead to the inactivation of surfactant. MIF, macrophage inhibitory factor. *(Adapted from LB Ware, MA Matthay: N Engl J Med 342:1334, 2000, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

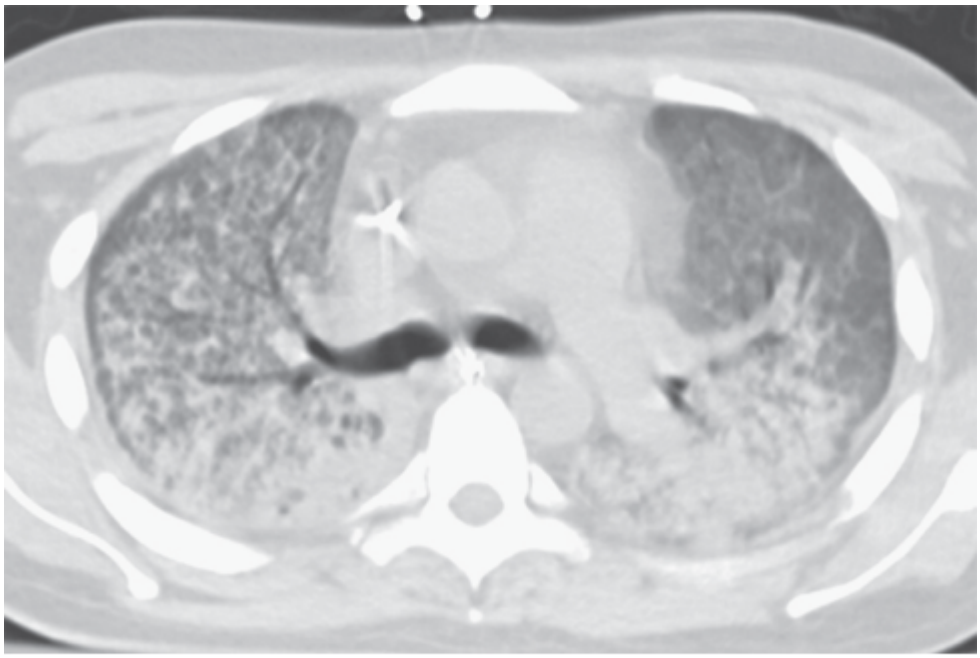
Alveolar edema predominantly involves *dependent* portions of the lung with diminished aeration. Collapse of large sections of dependent lung can contribute to decreased lung compliance. Consequently, intrapulmonary shunting and hypoxemia develop and the work of breathing increases, leading to dyspnea. The pathophysiologic alterations in alveolar spaces are exacerbated by microvascular occlusion that results in reductions in pulmonary arterial blood flow to ventilated portions of the lung (and thus in increased dead space) and in pulmonary hypertension. Thus, in addition to severe hypoxemia, hypercapnia secondary to an increase in pulmonary dead space can be prominent in early ARDS.

The exudative phase encompasses the first 7 days of illness after exposure to a precipitating ARDS risk factor, with the patient experiencing the onset of respiratory symptoms. Although usually

presenting within 12–36 h after the initial insult, symptoms can be delayed by 5–7 days. Dyspnea develops, with a sensation of rapid shallow breathing and an inability to get enough air. Tachypnea and increased work of breathing result frequently in respiratory fatigue and ultimately in respiratory failure. Laboratory values are generally nonspecific and are primarily indicative of underlying clinical disorders. The chest radiograph usually reveals opacities consistent with pulmonary edema and often involves at least three-quarters of the lung fields (Fig. 294-2). While characteristic for ARDS, these radiographic findings are not specific and can be indistinguishable from cardiogenic pulmonary edema **(Chap. 298)**. Unlike the latter, however, the chest x-ray in ARDS may not demonstrate cardiomegaly, pleural effusions, or pulmonary vascular redistribution as is often present in pure cardiogenic pulmonary edema. If no ARDS risk factor is present, then some objective evaluation is required (e.g., echocardiography) to exclude a cardiac etiology for hydrostatic edema. Chest computed tomography (CT) in ARDS also reveals the presence of bilateral pulmonary infiltrates and demonstrates extensive heterogeneity of lung involvement **(Fig. 294-4)**.

FIGURE 294-4

A representative CT scan of the chest during the exudative phase of acute respiratory distress syndrome (ARDS), in which *dependent* alveolar edema and atelectasis predominate.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Because the early features of ARDS are nonspecific, alternative diagnoses must be considered. In the differential diagnosis of ARDS, the most common disorders are cardiogenic pulmonary edema, bilateral pneumonia, and alveolar hemorrhage. Less common diagnoses to consider include acute interstitial lung diseases (e.g., acute interstitial pneumonitis; **Chap. 287**), acute immunologic injury (e.g., hypersensitivity pneumonitis; **Chap. 282**), toxin injury (e.g., radiation pneumonitis; **Chap. 71**), and neurogenic pulmonary edema **(Chap. 33)**.

Proliferative Phase

This phase of ARDS usually lasts from day 7 to day 21. Most patients recover rapidly and are liberated from mechanical ventilation during this phase. Despite this improvement, many patients still experience dyspnea, tachypnea, and hypoxemia. Some patients develop progressive lung injury and

early changes of pulmonary fibrosis during the proliferative phase. Histologically, the first signs of resolution are often evident in this phase, with the initiation of lung repair, the organization of alveolar exudates, and a shift from neutrophil- to lymphocyte-predominant pulmonary infiltrates. As part of the reparative process, type II pneumocytes proliferate along alveolar basement membranes. These specialized epithelial cells synthesize new pulmonary surfactant and differentiate into type I pneumocytes.

Fibrotic Phase

While many patients with ARDS recover lung function 3–4 weeks after the initial pulmonary injury, some enter a fibrotic phase that may require long-term support on mechanical ventilators and/or supplemental oxygen. Histologically, the alveolar edema and inflammatory exudates of earlier phases convert to extensive alveolar-duct and interstitial fibrosis. Marked disruption of acinar architecture leads to emphysema-like changes, with large bullae. Intimal fibroproliferation in the pulmonary microcirculation causes progressive vascular occlusion and pulmonary hypertension. The physiologic consequences include an increased risk of pneumothorax, reductions in lung compliance, and increased pulmonary dead space. Patients in this late phase experience a substantial burden of excess morbidity. Lung biopsy evidence for pulmonary fibrosis in any phase of ARDS is associated with increased mortality risk.

TREATMENT

TREATMENT Acute Respiratory Distress Syndrome GENERAL PRINCIPLES

Recent reductions in ARDS mortality rates are largely the result of general advances in the care of critically ill patients **(Chap. 293)**. Thus, caring for these patients requires close attention to (1) the recognition and treatment of underlying medical and surgical disorders (e.g., pneumonia, sepsis, aspiration, trauma); (2) the minimization of unnecessary procedures and their complications; (3) standardized “bundled care” approaches for ICU patients, including prophylaxis against venous thromboembolism, gastrointestinal bleeding, aspiration, excessive sedation, prolonged mechanical ventilation, and central venous catheter infections; (4) prompt recognition of nosocomial infections; and (5) provision of adequate nutrition via the enteral route when feasible.

MANAGEMENT OF MECHANICAL VENTILATION

(See also Chap. 295) Patients meeting clinical criteria for ARDS frequently become fatigued from increased work of breathing and progressive hypoxemia, requiring mechanical ventilation for support.

Minimizing Ventilator-Induced Lung Injury

Despite its life-saving potential, mechanical ventilation can aggravate lung injury. Experimental models have demonstrated that ventilator-induced lung injury can arise from at least two principal mechanisms: “volutrauma” from repeated alveolar overdistention from excess tidal volume and “atelectrauma” from recurrent alveolar collapse. As is evident from chest CT (Fig. 294-4), ARDS is a heterogeneous disorder, principally involving dependent portions of the lung with relative sparing of other regions. Because compliance differs in affected versus more “normal” areas of the lung, attempts to fully inflate the consolidated lung may lead to overdistention of and injury to the more normal areas. Ventilator-induced injury can be demonstrated in experimental models of acute lung injury, in particular with high-tidal-volume (V_T) ventilation.

A large-scale, randomized controlled trial sponsored by the National Institutes of Health and conducted by the ARDS Network compared low V_T ventilation (6 mL/kg of predicted body weight) to conventional V_T ventilation (12 mL/kg predicted body weight). Lower airway pressures were also

targeted in the low tidal volume group (i.e., plateau pressure measured on the ventilator after a 0.5-s pause after inspiration) ≤30 cm H_2O versus ≤50 cm H_2O in the high tidal volume group. The mortality rate was significantly lower in the low V_T patients (31%) than in the conventional V_T patients (40%). This improvement in survival represents a substantial ARDS-mortality benefit.

Minimizing Atelectrauma by Prevention of Alveolar Collapse

In ARDS, the presence of alveolar and interstitial fluid and the loss of surfactant can lead to a marked reduction of lung compliance. Without an increase in end-expiratory pressure, significant alveolar collapse can occur at end-expiration, with consequent impairment of oxygenation. In most clinical settings, positive end-expiratory pressure (PEEP) is adjusted to minimize F_{IO_2} (inspired O_2 percentage) and provide adequate Pa_{O_2} (arterial partial pressure of O_2) without causing alveolar overdistention. Currently, there is no consensus on the optimal method to set PEEP, because numerous trials have proved inconclusive. Possible approaches include using the table of PEEP-F_{IO_2} combinations from the ARDS Network trial group, generating a static pressure-volume curve for the respiratory system and setting PEEP at the lower inflection point on this curve to maximize respiratory system compliance, and measuring esophageal pressures to estimate transpulmonary pressure (which may be particularly helpful in patients with a stiff chest wall). Until more data become available on how best to optimize PEEP settings in ARDS, clinicians can use these options or a practical approach to empirically measure “best PEEP” at the bed side to determine the optimal settings that best promotes alveolar recruitment, minimizes alveolar overdistention and hemodynamic instability, and provides adequate Pa_{O_2} while minimizing F_{IO_2} **(Chap. 295)**.

Prone Positioning

While several prior trials demonstrated that mechanical ventilation in the prone position improved arterial oxygenation without a mortality benefit, a recent trial demonstrated a significant reduction in 28-day mortality with prone positioning (32.8 to 16%) for patients with severe ARDS (Pa_{O_2}/F_{IO_2} <150 mm Hg). Thus, many centers are increasing the use of prone positioning in severe ARDS, with the understanding that this maneuver requires a critical-care team that is experienced in “proning,” as repositioning critically ill patients can be hazardous, leading to accidental endotracheal extubation, loss of central venous catheters, and orthopedic injury.

OTHER STRATEGIES IN MECHANICAL VENTILATION

Recruitment maneuvers that transiently increase PEEP to high levels to “recruit” atelectatic lung can increase oxygenation, but a mortality benefit has not been established. *Alternate modes of mechanical ventilation*, such as airway pressure release ventilation and high frequency oscillatory ventilation, have not been proven beneficial over standard modes of ventilation in ARDS management and in many cases require specialized expertise at the bedside. *Lung-replacement therapy with extracorporeal membrane oxygenation* (ECMO) was shown to improve mortality for patients with ARDS in the United Kingdom who were referred to an ECMO center (though only 75% of referred patients received ECMO) and thus may have utility in select adult patients with severe ARDS as a rescue therapy.

FLUID MANAGEMENT

(See also Chap. 293) Increased pulmonary vascular permeability leading to interstitial and alveolar edema fluid rich in protein is a central feature of ARDS. In addition, impaired vascular integrity augments the normal increase in extravascular lung water that occurs with increasing left atrial pressure. Maintaining a low left atrial filling pressure minimizes pulmonary edema and prevents further decrements in arterial oxygenation and lung compliance; improves pulmonary mechanics; shortens ICU stay and the duration of mechanical ventilation. Thus, aggressive attempts to reduce left atrial filling pressures with fluid restriction and diuretics should be an important aspect of ARDS management, limited only by hypotension and hypoperfusion of critical organs such as the kidneys.

NEUROMUSCULAR BLOCKADE

In severe ARDS, sedation alone can be inadequate for the patient-ventilator synchrony required for lung-protective ventilation. In a multicenter, randomized, placebo-controlled trial of early neuromuscular blockade (with cisatracurium besylate) for 48 h, patients with severe ARDS had increased survival and ventilator-free days without increasing ICU-acquired paresis. These promising findings support the early administration of neuromuscular blockade if needed to facilitate mechanical ventilation in severe ARDS.

GLUCOCORTICOIDS

Many attempts have been made to treat both early and late ARDS with glucocorticoids, with the goal of reducing potentially deleterious pulmonary inflammation. Few studies have shown any significant mortality benefit. Current evidence does *not* support the routine use of glucocorticoids in the care of ARDS patients.

OTHER THERAPIES

Clinical trials of surfactant replacement and multiple other medical therapies have proved disappointing. Pulmonary vasodilators such as inhaled nitric oxide and inhaled epoprostenol sodium can transiently improve oxygenation but have not been shown to improve survival or decrease time on mechanical ventilation.

RECOMMENDATIONS

Many clinical trials have been undertaken to improve the outcome of patients with ARDS; most have been unsuccessful in modifying the natural history. While results of large clinical trials must be judiciously applied to *individual* patients, evidence-based recommendations are summarized in **Table 294-3**, and an algorithm for the initial therapeutic goals and limits in ARDS management is provided in **Fig. 294-5**.

TABLE 294-3 Evidence-Based Recommendations for ARDS Therapies

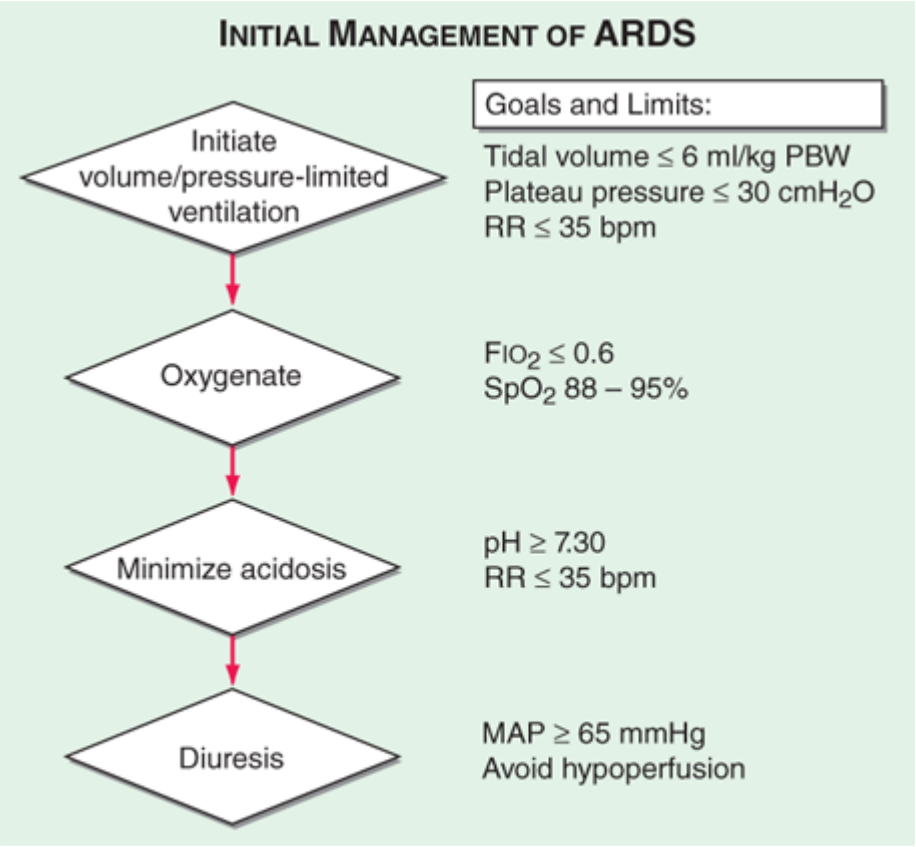
Treatment	Recommendation[a]
Mechanical ventilation	
Low tidal volume	A
Minimized left atrial filling pressures	B
High-PEEP or “open lung”	B[b]
Prone position	B[b]
Recruitment maneuvers	C
High-frequency ventilation	D
ECMO	B[b]
Early neuromuscular blockade	B[b]
Glucocorticoid treatment	D
Inhaled vasodilators (e.g., inhaled NO, inhaled epoprosteol)	C
Surfactant replacement, and other anti-inflammatory therapy (e.g., ketoconazole, PGE_1, NSAIDs)	D

[a]Key: A, recommended therapy based on strong clinical evidence from randomized clinical trials; B, recommended therapy based on supportive but limited clinical data; C, recommended only as alternative therapy on the basis of indeterminate evidence; D, not recommended on the basis of clinical evidence against efficacy of therapy. [b]As described in the text, there is no consensus on optimal PEEP setting in ARDS, but general consensus supports an open lung strategy that minimizes alveolar distention; prone positioning was shown to improve mortality in severe ARDS in one randomized controlled clinical trial; ECMO may be beneficial in select patients with severe ARDS; early neuromuscular blockade demonstrated a mortality benefit in one randomized controlled trial in patients with severe ARDS

Abbreviations: ARDS, acute respiratory distress syndrome; ECMO, extracorporeal membrane oxygenation; NO, nitric oxide; NSAIDs, nonsteroidal anti-inflammatory drugs; PEEP, positive end-expiratory pressure; PGE1, prostaglandin E_1.

FIGURE 294-5

Algorithm for the initial management of ARDS. Clinical trials have provided evidence-based therapeutic goals for a stepwise approach to the early mechanical ventilation, oxygenation, and correction of acidosis and diuresis of critically ill patients with ARDS. FIO_2, inspired O_2 percentage; MAP, mean arterial pressure; PBW, predicted body weight; RR, respiratory rate; SpO_2, arterial oxyhemoglobin saturation measured by pulse oximetry.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PROGNOSIS

Mortality

In the recent report from the Large Observational Study to Understand the Global Impact of Severe Acute Respiratory Failure (LUNG SAFE) trial, hospital mortality estimates for ARDS range from 34.9% for mild ARDS, 40.3% for moderate ARDS, and 46.1% with severe ARDS. There is substantial variability, but a trend toward improved ARDS outcomes over time appears evident. Of interest, mortality in ARDS is largely attributable to nonpulmonary causes, with sepsis and nonpulmonary organ failure accounting for >80% of deaths. Thus, improvement in survival is likely secondary to advances in the care of septic/infected patients and those with multiple organ failure **(Chap. 293)**.

The major risk factors for ARDS mortality are nonpulmonary. Advanced age is an important risk factor. Patients aged >75 years have a substantially higher mortality risk (~60%) than those <45 (~20%). Moreover, patients >60 years of age with ARDS and sepsis have a threefold higher mortality risk than those <60. Other risk factors include preexisting organ dysfunction from chronic medical illness—in particular, chronic liver disease, cirrhosis, chronic alcohol abuse, chronic immunosuppression **(Chap. 293)**. Patients with ARDS arising from direct lung injury (including pneumonia, pulmonary contusion, and aspiration; Table 294-1) are nearly twice as likely to die as those with indirect causes of lung injury, while surgical and trauma patients with ARDS—especially those without direct lung injury—generally have a higher survival rate than other ARDS patients.

Increasing severity of ARDS, as defined by the consensus Berlin definition, predicts increased mortality. Surprisingly, there is little additional value in predicting ARDS mortality from other parameters of lung injury, including the level of PEEP (≥10 cm H_2O), respiratory system compliance

(≤40 mL/cm H_2O), the extent of alveolar infiltrates on chest radiography, and the corrected expired volume per minute (≥10 L/min) (as a surrogate measure of dead space).

Functional Recovery in ARDS Survivors

While it is common for patients with ARDS to experience prolonged respiratory failure and remain dependent on mechanical ventilation for survival, it is a testament to the resolving powers of the lung that the majority of patients who survive regain nearly normal lung function. Patients usually recover maximal lung function within 6 months. One year after endotracheal extubation, more than one-third of ARDS survivors have normal spirometry values and diffusion capacity. Most of the remaining patients have only mild abnormalities in pulmonary function. Unlike mortality risk, recovery of lung function is strongly associated with the extent of lung injury in early ARDS. Low static respiratory compliance, high levels of required PEEP, longer durations of mechanical ventilation, and high lung injury scores are all associated with less recovery of pulmonary function. Of note, when physical function is assessed 5 years after ARDS, exercise limitation and decreased physical quality of life are often documented despite normal or nearly normal pulmonary function. When caring for ARDS survivors, it is important to be aware of the potential for a substantial burden of psychological problems in patients and family caregivers, including significant rates of depression and posttraumatic stress disorder.

ACKNOWLEDGMENT

The authors acknowledge the contributions to this chapter by the previous authors, Drs. Augustine Choi and Steven D. Shapiro.

FURTHER READING

ARDS definition task force: Acute respiratory distress syndrome: The Berlin Definition. JAMA 307:2526, 2012. [PubMed: 22797452]
ARDS network: Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N Engl J Med 342:1301, 2000. [PubMed: 10793162]
bellani G et al: Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. JAMA 315:788, 2016. [PubMed: 26903337]
tomashefski jf jr: Pulmonary pathology of acute respiratory distress syndrome. Clin Chest Med 21:435, 2000. [PubMed: 11019719]
ware lb, matthay ma: The acute respiratory distress syndrome. N Engl J Med 342:1334, 2000. [PubMed: 10793167]

WEBSITES

ARDS Support Center for patient-oriented education: *www.ards.org*
NHLBI ARDS Clinical Trials information: *www.ardsnet.org*
ARDS Foundation: *www.ardsusa.org*

Chapter 295: Mechanical Ventilatory Support

Bartolome R. Celli

MECHANICAL VENTILATORY SUPPORT

Mechanical ventilation (MV) is used to assist or replace spontaneous breathing. It is implemented with special devices that can support ventilatory function and improve oxygenation through the application of high-oxygen-content gas and positive pressure. The primary indication for initiation of MV is respiratory failure, of which there are two basic types: (1) *hypoxemic*, which is present when arterial O_2 saturation (SaO_2) <90% occurs despite an increased inspired O_2 fraction and usually results from ventilation-perfusion mismatch or shunt; and (2) *hypercarbic*, which is characterized by elevated arterial carbon dioxide partial pressure (PCO_2) values (usually >50 mmHg) resulting from conditions that decrease minute ventilation or increase physiologic dead space such that alveolar ventilation is inadequate to meet metabolic demands. When respiratory failure is chronic, neither of the two types is obligatorily treated with MV, but when it is acute, MV may be lifesaving.

INDICATIONS

The most common reasons for instituting MV are acute respiratory failure with hypoxemia (acute respiratory distress syndrome, heart failure with pulmonary edema, pneumonia, sepsis, complications of surgery and trauma), which accounts for ~65% of all ventilated cases, and hypercarbic ventilatory failure—e.g., due to coma (15%), exacerbations of chronic obstructive pulmonary disease (COPD; 13%), and neuromuscular diseases (5%). The primary objectives of MV are to decrease the work of breathing, thus avoiding respiratory muscle fatigue, and to reverse life-threatening hypoxemia and progressive respiratory acidosis.

In some cases, MV is used as an adjunct to other forms of therapy. For example, it is used to reduce cerebral blood flow in patients with increased intracranial pressure. MV also is used frequently in conjunction with endotracheal intubation for airway protection to prevent aspiration of gastric contents in otherwise unstable patients during gastric lavage for suspected drug overdose or during gastrointestinal endoscopy. In critically ill patients, intubation and MV may be indicated before the performance of essential diagnostic or therapeutic studies if it appears that respiratory failure may occur during those maneuvers.

TYPES OF MECHANICAL VENTILATION

There are two basic methods of MV: noninvasive ventilation (NIV) and invasive (or conventional mechanical) ventilation (MV).

Noninvasive Ventilation

NIV has gained acceptance because it is effective in certain conditions, such as acute or chronic respiratory failure, and is associated with fewer complications—namely, pneumonia and tracheolaryngeal trauma. NIV usually is provided with a tight-fitting face mask, a nasal mask similar to

that used for treatment of sleep apnea and in some cases with the use of a helmet or a hood. NIV has proved highly effective in patients with respiratory failure arising from exacerbations of COPD. It is most frequently implemented as bilevel positive airway pressure ventilation or pressure-support ventilation (PSV). Both modes, which apply a preset positive pressure during inspiration and a lower pressure during expiration, are well tolerated by a conscious patient and optimize patient-ventilator synchrony. The major limitation to the widespread application of NIV has been patient intolerance: the interface required for NIV can cause both physical and psychological discomfort. In addition, NIV has had limited success in patients with acute hypoxemic respiratory failure, for whom endotracheal intubation and conventional MV remain the ventilatory method of choice.

The most important group of patients who benefit from a trial of NIV are those with COPD exacerbations and respiratory acidosis (pH <7.35). Several randomized trials have shown that, in patients with ventilatory failure characterized by blood pH levels between 7.25 and 7.35, NIV is associated with low failure rates (15–20%) and good outcomes (as judged by intubation rate, length of stay in intensive care, and—in some series—mortality rates). In more severely ill patients with a blood pH <7.25, the rate of NIV failure is inversely related to the severity of respiratory acidosis, with higher failure rates as the pH decreases. In patients with milder acidosis (pH >7.35), NIV is not better than conventional treatment that includes controlled oxygen delivery and pharmacotherapy for exacerbations of COPD (systemic glucocorticoids, bronchodilators, and, if needed, antibiotics).

NIV is not useful in the majority of cases of respiratory failure and is contraindicated in patients with the conditions listed in **Table 295-1**. NIV can delay lifesaving ventilatory support in those cases and, in fact, can actually result in aspiration or hypoventilation. Once NIV is initiated, patients should be monitored; a reduction in respiratory frequency and a decrease in the use of accessory muscles (scalene, sternomastoid, and intercostals) are good clinical indicators of therapeutic benefit. Arterial blood gases should be determined at least within hours of the initiation of therapy to ensure that NIV is having the desired effect. Lack of benefit within that time frame should alert the physician to the possible need for conventional MV.

TABLE 295-1 Contraindications for Noninvasive Ventilation

Cardiac or respiratory arrest
Severe encephalopathy
Severe gastrointestinal bleed
Hemodynamic instability
Unstable angina and myocardial infarction
Facial surgery or trauma
Upper airway obstruction
High-risk aspiration and/or inability to protect airways
Inability to clear secretions

Conventional MV

Conventional MV is implemented once a cuffed tube is inserted into the trachea to allow conditioned gas (warmed, oxygenated, and humidified) to be delivered to the airways and lungs at pressures above atmospheric pressure. Care should be taken during intubation to avoid brain-damaging hypoxia. In most cases, the administration of mild sedation may facilitate the procedure. Opiates and benzodiazepines are good choices but can have a deleterious effect on hemodynamics in patients with depressed cardiac function or low systemic vascular resistance. Morphine can promote histamine release from tissue mast cells and may worsen bronchospasm in patients with asthma; fentanyl, sufentanil, and alfentanil are acceptable alternatives. Ketamine may increase systemic arterial pressure and has been associated with hallucinatory responses. The shorter-acting agents—etomidate and propofol—have been used for both induction and maintenance of anesthesia in

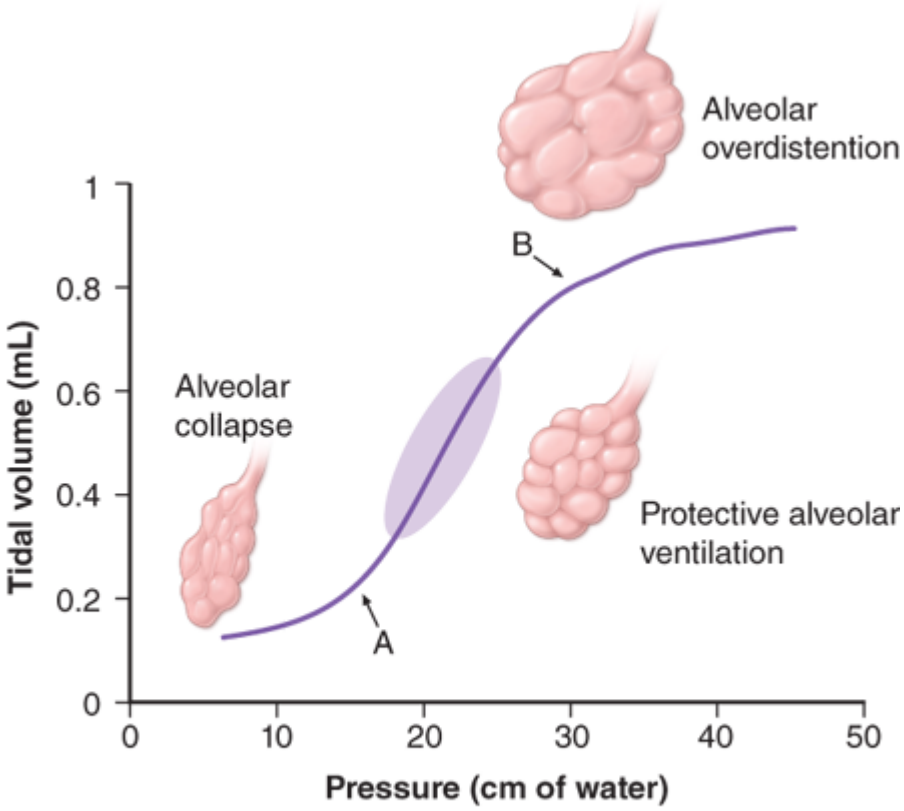
ventilated patients because they have fewer adverse hemodynamic effects, but both are significantly more expensive than older agents. Great care must be taken to avoid the use of neuromuscular paralysis during intubation of patients with renal failure, tumor lysis syndrome, crush injuries, medical conditions associated with elevated serum potassium levels, and muscular dystrophy syndromes; in particular, the use of agents whose mechanism of action includes depolarization at the neuromuscular junction, such as succinylcholine chloride, must be avoided.

PRINCIPLES OF MV

Once the patient has been intubated, the basic goals of MV are *to optimize oxygenation while avoiding ventilator-induced lung injury due to overstretch and collapse/re-recruitment*. This concept, known as the “protective ventilatory strategy” (see below and **Fig. 295-1**) is supported by evidence linking high airway pressures and volumes and overstretching of the lung as well as collapse/re-recruitment to poor clinical outcomes (barotrauma and volume trauma). Although normalization of pH through elimination of CO_2 is desirable, the risk of lung damage associated with the large volume and high pressures needed to achieve this goal has led to the acceptance of permissive hypercapnia. This condition is well tolerated when care is taken to avoid excess acidosis by pH buffering.

FIGURE 295-1

Hypothetical pressure-volume curve of the lung in a patient undergoing mechanical ventilation. Alveoli tend to close if the distending pressure falls below the lower inflection point A, whereas they overstretch if the pressure within them is higher than that of the upper inflection point B. Collapse and opening of ventilated alveoli are associated with poor outcomes in patients with acute respiratory failure. Protective ventilation (*purple shaded area*), using a lower tidal volume (6 mL/kg of ideal body weight) and maintaining positive end-expiratory pressure to prevent overstretching and collapse/opening of alveoli, has resulted in improved survival rates among patients receiving mechanical ventilatory support.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Modes of Ventilation

Mode refers to the manner in which ventilator breaths are triggered, cycled, and limited. The *trigger*, either an inspiratory effort or a time-based signal, defines what the ventilator senses to initiate an assisted breath. *Cycle* refers to the factors that determine the end of inspiration. For example, in volume-cycled ventilation, inspiration ends when a specific tidal volume is delivered. Other types of cycling include pressure cycling and time cycling. The *limiting factors* are operator-specified values, such as airway pressure, that are monitored by transducers internal to the ventilator circuit throughout the respiratory cycle; if the specified values are exceeded, inspiratory flow is terminated, and the ventilator circuit is vented to atmospheric pressure or the specified pressure at the end of expiration (positive end-expiratory pressure, or PEEP). Most patients are ventilated with assist-control ventilation (ACMV), intermittent mandatory ventilation (IMV), or PSV, with the latter two modes often used simultaneously **(Table 295-2)**.

TABLE 295-2 Characteristics of the Most Commonly Used Forms of Mechanical Ventilation

Ventilatory Mode	Variables Set by User (Independent)	Variables Monitored by User (Dependent)	Trigger	Cycle	Limit	Advantages	Disadvantages
ACMV (assist-control ventilation)	Tidal volume Ventilator rate F_{IO_2} PEEP level Pressure limit	Peak, mean, and plateau airway pressures VE ABG I/E ratio	Patient effort Timer	Pressure limit		Patient control Guaranteed ventilation	Potential hyperventilation Barotrauma and volume trauma Every effective breath generates a ventilator volume
IMV (intermittent mandatory ventilation)	Tidal volume Mandatory ventilator rate F_{IO_2} PEEP level Pressure limit Spontaneous breaths between assisted breaths	Peak, mean, and plateau airway pressures VE ABG I/E ratio	Patient effort Timer	Pressure limit		Patient control Comfort from spontaneous breaths Guaranteed ventilation	Potential dysynchrony Potential hypoventilation
PSV (pressure-support ventilation)	Inspiratory pressure level F_{IO_2} PEEP Pressure limit	Tidal volume Respiratory rate VE ABG	Pressure limit Inspiratory flow			Patient control Comfort Assures synchrony	No timer backup Potential hypoventilation
NIV (noninvasive ventilation)	Inspiratory and expiratory pressure level F_{IO_2}	Tidal volume Respiratory rate VE ABG	Pressure limit Inspiratory flow			Patient control	Mask interface may cause discomfort and facial bruising Leaks are common Hypoventilation

Abbreviations: ABG, arterial blood gases; F_{IO_2}, fraction of inspired oxygen; PEEP, positive end-expiratory pressure; I/E, inspiratory to expiratory time ratio; VE, minute ventilation.

Assist-Control Ventilation

ACMV is the most widely used mode of ventilation. In this mode, an inspiratory cycle is initiated either by the patient's inspiratory effort or, if none is detected within a specified time window, by a timer signal within the ventilator. Every breath delivered, whether patient- or timer-triggered, consists of the operator-specified tidal volume. Ventilatory rate is determined either by the patient or by the operator-

specified backup rate, whichever is of higher frequency. ACMV is commonly used for initiation of MV because it ensures a backup minute ventilation in the absence of an intact respiratory drive and allows for synchronization of the ventilator cycle with the patient's inspiratory effort.

Problems can arise when ACMV is used in patients with tachypnea due to nonrespiratory or nonmetabolic factors, such as anxiety, pain, and airway irritation. Respiratory alkalemia may develop and trigger myoclonus or seizures. Dynamic hyperinflation leading to increased intrathoracic pressures (so-called auto-PEEP) may occur if the patient's respiratory mechanics are such that inadequate time is available for complete exhalation between inspiratory cycles. Auto-PEEP can limit venous return, decrease cardiac output, and increase airway pressures, predisposing to barotrauma.

Intermittent Mandatory Ventilation

With this mode, the operator sets the number of mandatory breaths of fixed volume to be delivered by the ventilator; between those breaths, the patient can breathe spontaneously. In the most frequently used synchronized mode (SIMV), mandatory breaths are delivered in synchrony with the patient's inspiratory efforts at a frequency determined by the operator. If the patient fails to initiate a breath, the ventilator delivers a fixed-tidal-volume breath and resets the internal timer for the next inspiratory cycle. SIMV differs from ACMV in that only a preset number of breaths are ventilator-assisted.

SIMV allows patients with an intact respiratory drive to exercise inspiratory muscles between assisted breaths; thus it is useful for both supporting and weaning intubated patients. SIMV may be difficult to use in patients with tachypnea because they may attempt to exhale during the ventilator-programmed inspiratory cycle. Consequently, the airway pressure may exceed the inspiratory pressure limit, the ventilator-assisted breath will be aborted, and minute volume may drop below that programmed by the operator. In this setting, if the tachypnea represents a response to respiratory or metabolic acidosis, a change in ACMV will increase minute ventilation and help normalize the pH while the underlying process is further evaluated and treated.

Pressure-Support Ventilation

This form of ventilation is patient-triggered, flow-cycled, and pressure-limited. It provides graded assistance and differs from the other two modes in that the operator sets the pressure level (rather than the volume) to augment every spontaneous respiratory effort. The level of pressure is adjusted by observing the patient's respiratory frequency. During PSV, the inspiration is terminated when inspiratory airflow falls below a certain level; in most ventilators, this flow rate cannot be adjusted by the operator. With PSV, patients receive ventilator assistance only when the ventilator detects an inspiratory effort. PSV is often used in combination with SIMV to ensure volume-cycled backup for patients whose respiratory drive is depressed. PSV is well tolerated by most patients who are being weaned from MV; PSV parameters can be set to provide full ventilatory support and can be withdrawn to load the respiratory muscles gradually.

Other Modes of Ventilation

There are other modes of ventilation, each with its own acronym and each with specific modifications of the manner and duration in which pressure is applied to the airway and lungs and of the interaction between the mechanical assistance provided by the ventilator and the patient's respiratory effort. Although their use in acute respiratory failure is limited, the following modes have been used with varying levels of enthusiasm and adoption.

PRESSURE-CONTROL VENTILATION (PCV)

This form of ventilation is time-triggered, time-cycled, and pressure-limited. A specified pressure is imposed at the airway opening throughout inspiration. Since the inspiratory pressure is specified by the operator, tidal volume and inspiratory flow rate are *dependent*, rather than *independent*, variables and are not operator-specified. PCV is the preferred mode of ventilation for patients in whom it is desirable to regulate peak airway pressures, such as those with preexisting barotrauma, and for post–thoracic surgery patients, in whom the shear forces across a fresh suture line should be limited. When PCV is used, minute ventilation is altered through changes in rate or in the pressure-control value, with consequent changes in tidal volume.

INVERSE-RATIO VENTILATION (IRV)

This mode is a variant of PCV that incorporates the use of a prolonged inspiratory time with the appropriate shortening of the expiratory time. IRV has been used in patients with severe hypoxemic respiratory failure. This approach increases mean distending pressures without increasing peak airway pressures. It is thought to work in conjunction with PEEP to open collapsed alveoli and improve oxygenation. However, no clinical-trial data have shown that IRV improves outcomes.

CONTINUOUS POSITIVE AIRWAY PRESSURE (CPAP)

CPAP is not a true support mode of ventilation because all ventilation occurs through the patient's spontaneous efforts. The ventilator provides fresh gas to the breathing circuit with each inspiration and sets the circuit to a constant, operator-specified pressure. CPAP is used to assess extubation potential in patients who have been effectively weaned and who require little ventilatory support and in patients with intact respiratory system function who require an endotracheal tube for airway protection.

Nonconventional Ventilatory Strategies

Several nonconventional strategies have been evaluated for their ability to improve gas exchange and survival rates in severe hypoxemic respiratory failure. These strategies include high-frequency oscillatory ventilation (HFOV), airway pressure release ventilation (APRV), partial liquid ventilation (PLV) using perfluorocarbons and the administration of nitric oxide gas delivered through the airways. Although case reports and small uncontrolled cohort studies have shown benefit, randomized controlled trials have failed to demonstrate consistent improvements in outcome with these strategies.

Some "salvage" techniques have gained acceptance given recent positive clinical outcomes. A randomized trial of extracorporeal membrane oxygenation (ECMO) documented positive outcomes, although older studies had failed to document positive results. However, with the popularity of venous-venous access, and encouraging reports in small series and uncontrolled trials, the use of ECMO has increased world-wide. Guidelines for ECMO centers have been published, and if considered in patients with severe respiratory failure refractory to conventional therapy, the patient should be referred to a center familiar with the procedure. The second of these techniques, prone positioning, should be available in all units caring for these patients. Several multi-center randomized trials in patients with acute lung injury and refractory hypoxemia have shown that prone positioning improves ventilation-perfusion matching and provides short- and long-term survival advantage.

The design of new ventilator modes reflects attempts to improve patient-ventilator synchrony—a major practical issue during MV—by allowing patients to trigger the ventilator with their own effort while also incorporating flow algorithms that terminate the cycles once certain preset criteria are reached; this approach has greatly improved patient comfort. New modes of ventilation that synchronize not only the timing but also the levels of assistance to match the patient's effort have been developed. Proportional assist ventilation (PAV) and neurally adjusted ventilatory-assist ventilation (NAV) are two modes that are designed to deliver assisted breaths through algorithms incorporating not only pressure, volume, and time but also overall respiratory resistance as well as compliance (in the case

of PAV) and neural activation of the diaphragm (in the case of NAV). Although these modes enhance patient-ventilator synchrony, their practical use in the everyday management of patients undergoing MV needs further study.

PROTECTIVE VENTILATORY STRATEGY

Whichever mode of MV is used in acute respiratory failure, the evidence from several important controlled trials indicates that a protective ventilation approach guided by the following principles (and summarized in Fig. 295-1) is safe and offers the best chance of a good outcome: (1) Set a target tidal volume close to 6 mL/kg of ideal body weight. (2) Prevent plateau pressure (static pressure in the airway at the end of inspiration) exceeding 30 cm H_2O. (3) Use the lowest possible fraction of inspired oxygen (FIO_2) to keep the SaO_2 at ≥90%. (4) Adjust the PEEP to maintain alveolar patency while preventing overdistention and closure/reopening. With the application of these techniques, the mortality rate among patients with acute hypoxemic respiratory failure has decreased to ~30% from close to 50% a decade ago.

PATIENT MANAGEMENT

Once the patient's gas exchange has been stabilized, definitive therapy for the underlying process responsible for respiratory failure is continued. Subsequent modifications in ventilator therapy must be provided in parallel with changes in the patient's clinical status. As improvement in respiratory function is noted, the first priority is to reduce the level of mechanical ventilatory support. Patients on full ventilatory support should be monitored frequently, with the goal of switching to a mode that allows for weaning as soon as possible. Protocols and guidelines that can be applied by paramedical personnel when physicians are not readily available have proved to be of value in shortening ventilator and intensive care unit (ICU) time, with very good outcomes. Patients whose condition continues to deteriorate after ventilatory support is initiated may require increased O_2, PEEP, or one of the alternative modes of ventilation.

GENERAL SUPPORT DURING VENTILATION

Patients for whom MV has been initiated usually require sedation and analgesia to maintain an acceptable level of comfort. Often, this treatment consists of a combination of a benzodiazepine and an opiate administered intravenously. Medications commonly used for this purpose include lorazepam, midazolam, diazepam, morphine, and fentanyl. Oversedation must be avoided in the ICU because most studies show that daily interruption of sedation in patients with improved ventilatory status results in a shorter time on the ventilator and a shorter ICU stay.

Immobilized patients receiving mechanical ventilatory support are at risk for deep venous thrombosis and decubitus ulcers. Venous thrombosis should be prevented with the use of subcutaneous heparin and/or pneumatic compression boots. Fractionated low-molecular-weight heparin appears to be equally effective for this purpose. To help prevent decubitus ulcers, frequent changes in body position and the use of soft mattress overlays and air mattresses are employed. Early mobilization is recommended for patients on MV, since this approach is associated with better outcomes. Prophylaxis against diffuse gastrointestinal mucosal injury is indicated for patients undergoing MV. Histamine-receptor (H_2-receptor) antagonists, antacids, and cytoprotective agents such as sucralfate have all been used and appear to be effective. Nutritional support by enteral feeding through either a nasogastric or an orogastric tube should be initiated and maintained whenever possible. Delayed gastric emptying is common in critically ill patients taking sedative medications but often responds to promotility agents such as metoclopramide. Parenteral nutrition is an alternative to enteral nutrition in patients with severe gastrointestinal pathology who need prolonged MV.

COMPLICATIONS OF MECHANICAL VENTILATION

Endotracheal intubation and MV have direct and indirect effects on the lung and upper airways, the cardiovascular and the gastrointestinal system. Pulmonary complications include barotrauma, nosocomial pneumonia, oxygen toxicity, tracheal stenosis, and deconditioning of respiratory muscles. Barotrauma and volutrauma overdistend and disrupt lung tissue; may be clinically manifest by pneumomediastinum, interstitial and subcutaneous emphysema, or pneumothorax; and can result in the liberation of cytokines from overdistended tissues, further promoting tissue injury. Clinically significant pneumothorax requires tube thoracostomy. Intubated patients are at high risk for ventilator-associated pneumonia as a result of aspiration from the upper airways through small leaks around the endotracheal tube cuff; the most common organisms responsible for this condition are *Pseudomonas aeruginosa*, enteric gram-negative rods, and *Staphylococcus aureus*. Given the high associated mortality rates, when suspected, early initiation of empirical antibiotics directed against likely pathogens is recommended. *Hypotension* resulting from elevated intrathoracic pressures with decreased venous return is almost always responsive to intravascular volume repletion. In patients who are judged to have respiratory failure on the basis of alveolar edema but in whom the cardiac or pulmonary origin of the edema is unclear, hemodynamic monitoring with a pulmonary arterial catheter may be of value in helping to clarify the cause of the edema. Gastrointestinal effects of positive-pressure ventilation include stress ulceration and mild to moderate cholestasis.

WEANING FROM MECHANICAL VENTILATION

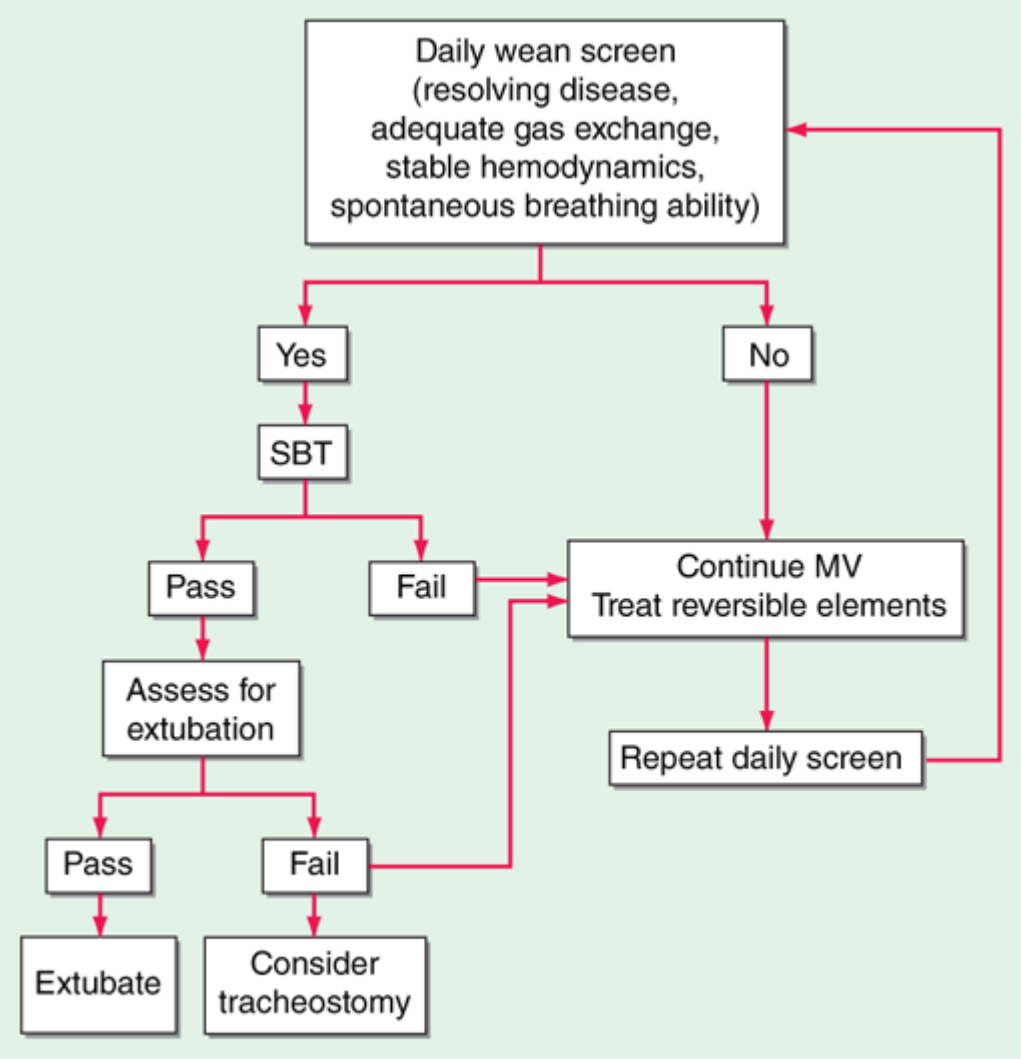
The Decision to Wean

It is important to consider discontinuation of MV once the underlying respiratory disease begins to reverse. Although the predictive capacities of multiple clinical and physiologic variables have been explored, the consensus from a ventilatory weaning task force cites the following conditions as indicating amenability to weaning: (1) Lung injury is stable or resolving; (2) gas exchange is adequate, with low PEEP (<8 cmH_2O) and F_{IO_2} (<0.5); (3) hemodynamic variables are stable, and the patient is no longer receiving vasopressors; and (4) the patient is capable of initiating spontaneous breaths. A "wean screen" based on these variables should be done at least daily. If the patient is deemed capable of beginning to wean, the recommendation is to perform a spontaneous breathing trial (SBT), whose value is supported by several randomized trials **(Fig. 295-2)**. The SBT involves an integrated patient assessment during spontaneous breathing with little or no ventilatory support. The SBT is usually implemented with a T-piece using 1–5 cmH_2O CPAP with 5–7 cmH_2O or PSV from the ventilator to offset resistance from the endotracheal tube. Once it is determined that the patient can breathe spontaneously, a decision must be made about the removal of the artificial airway, which should be undertaken only when it is concluded that the patient has the ability to protect the airway, is able to cough and clear secretions, and is alert enough to follow commands. In addition, other factors must be taken into account, such as the possible difficulty of replacing the tube if that maneuver is required. If upper airway difficulty is suspected, an evaluation using a "cuff-leak" test (assessing the presence of air movement around a deflated endotracheal tube cuff) is supported by current evidence. If the "cuff-leak test" suggests a risk of post-extubation stridor, the administration of systemic corticosteroids should be considered prior to extubation. Despite all precautions, ~10–15% of extubated patients require reintubation. Several studies suggest that NIV can be used to obviate reintubation, particularly in patients with ventilatory failure secondary to COPD exacerbation or congestive heart failure; in this setting, earlier extubation with the use of prophylactic NIV has yielded good results.

FIGURE 295-2

Flowchart to guide the daily approach to management of patients being considered for weaning off mechanical ventilation (MV). If attempts at extubation fail, a tracheostomy should be

considered. SBT, spontaneous breathing trial.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prolonged MV and Tracheostomy

From 5 to 13% of patients undergoing MV will go on to require prolonged MV (>21 days). In these instances, critical care personnel must decide whether and when to perform a tracheostomy. This decision is individualized and is based on the risk and benefits of tracheostomy and prolonged intubation as well as the patient's preferences and expected outcomes. A tracheostomy is thought to be more comfortable, to require less sedation, and to provide a more secure airway and may also reduce weaning time. However, tracheostomy carries the risk of complications, which occur in 5–40% of these procedures and include bleeding, cardiopulmonary arrest, hypoxia, structural damage, pneumothorax, pneumomediastinum, and wound infection. In patients with long-term tracheostomy, complex complications include tracheal stenosis, granulation, and erosion of the innominate artery. In general, if a patient needs MV for >10–14 days, a tracheostomy, planned under optimal conditions, is indicated. Whether it is completed at the bedside or as an operative procedure depends on local resources and experience. Some 5–10% of patients are deemed unable to wean in the ICU. These patients may benefit from transfer to special units where a multidisciplinary approach, including nutrition optimization, physical therapy with rehabilitation, and slower weaning methods (including SIMV with PSV), results in successful weaning rates of up to 30%. Unfortunately, close to 2% of ventilated patients may ultimately become dependent on ventilatory support to maintain life. Most of

these patients remain in chronic care institutions, although some with strong social, economic, and family support may live a relatively fulfilling life with at-home ventilation.

FURTHER READING

The Acute Respiratory Distress Syndrome Network: Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N Engl J Med 342:1301, 2000. [PubMed: 10793162]
Guérin C et al: Prone positioning in severe Acute Respiratory Distress Syndrome. N Engl J Med 368:2159, 2013. [PubMed: 23688302]
Mercat A et al for the Expiratory Pressure Study Group: Positive end-expiratory pressure settings in adults with acute lung injury and ARDS: A randomized controlled trial. JAMA 299:646, 2008. [PubMed: 18270353]
Schmidt GA et al: Liberation from mechanical ventilation in critically ill adults: Executive Summary of an Official American College of Chest Physicians/American Thoracic Society Clinical Practice Guideline. Chest 151;166, 2017. [PubMed: 27818331]
Squiers J et al: Contemporary extracorporeal membrane oxygenation therapy in adults: Fundamental principles and systematic review of the evidence. J Thorac Cardiovasc Surg 152:20, 2016. [PubMed: 27060027]

Chapter 296: Approach to the Patient with Shock

Anthony F. Massaro

INTRODUCTION

Shock is the clinical condition of organ dysfunction resulting from an imbalance between cellular oxygen supply and demand. This life-threatening condition is common in the intensive care unit (ICU). There are a multitude of heterogeneous disease processes that can lead to shock. The organ dysfunction seen in early shock is reversible with restoration of adequate oxygen supply. Left untreated, shock transitions from this reversible phase to an irreversible phase and death from multisystem organ dysfunction (MSOF). The clinician is required to identify the patient with shock promptly, make a preliminary assessment of the type of shock present, and initiate therapy to prevent irreversible organ dysfunction and death. In this chapter, we review a commonly used classification system that organizes shock into four major types based on the underlying physiologic derangement. We discuss the initial assessment utilizing the history, physical examination, and initial diagnostic testing to confirm the presence of shock and determine the type of shock causing the organ dysfunction. Finally, we will discuss key principles of initial therapy with the aim of reducing the high morbidity and mortality associated with shock.

PATHOPHYSIOLOGY OF SHOCK

The cellular oxygen imbalance of shock is most commonly related to impaired oxygen delivery in the setting of circulatory failure. Shock can also develop during states of increased oxygen consumption or impaired oxygen utilization. An example of the impaired oxygen utilization is cyanide poisoning, which causes uncoupling of oxidative phosphorylation. This chapter will focus on the approach to the patient with shock related to inadequate oxygen delivery.

In the setting of insufficient oxygen supply, the cell is no longer able to support aerobic metabolism. With adequate oxygen, the cell metabolizes glucose to pyruvate, which then enters the mitochondria where ATP is generated via oxidative phosphorylation. Without sufficient oxygen supply, the cell is forced into anaerobic metabolism, in which pyruvate is metabolized to lactate with much less ATP generation (per mole of glucose). Maintenance of the homeostatic environment of the cell is dependent on an adequate supply of ATP. ATP-dependent ion pumping systems, such as the Na+/K+ ATPase, consume 20–80% of the cell's energy. Inadequate oxygen delivery and subsequent decreased ATP disrupt the cell's ability to maintain osmotic, ionic, and intracellular pH homeostasis. Influx of calcium can lead to activation of calcium-dependent phospholipases and proteases, causing cellular swelling and death. In addition to direct cell death, cellular hypoxia can cause damage at the organ system level via leakage of the intracellular contents into the extracellular space activating inflammatory cascades and altering the microvascular circulation.

DETERMINANTS OF OXYGEN DELIVERY

Since shock is the clinical manifestation of inadequate oxygen delivery compared to cellular needs, we will review determinants of oxygen delivery (DO_2). Disease processes affecting any of the components of oxygen delivery have the potential to lead to the development of shock. Disturbances to key determinants of oxygen delivery form the basis of the four major shock types described below.

The two major components of DO_2 are cardiac output (CO) and arterial oxygen content (CaO_2):

$$DO_2 = CO \times CaO_2$$

The two components of CO are heart rate (HR) and stroke volume (SV), which can be substituted in the above equation as

$$DO_2 = (HR \times SV) \times CaO_2$$

The major determinants of SV are preload, afterload (systemic vascular resistance, SVR), and cardiac contractility. The relationship can be represented as

$$SV \alpha (Preload \times contractility)/SVR$$

In this equation, preload refers to the myocardial fiber length before contraction (the ventricular end-diastolic volume). Contractility refers to the ability of the ventricle to contract independent of preload and afterload. The SVR represents the afterload, or the force against which the ventricle must contract.

The CaO_2 is composed of oxygen carried by convection with hemoglobin and oxygen dissolved in blood, given as

$$CaO_2 = (Hb \times 1.39 \times SaO_2) + (PaO_2 \times 0.03)$$

A disease process that affects these variables (HR, preload, contractility, SVR, SaO_2, or Hb) has the potential to reduce oxygen delivery and cause cellular hypoxia. Each of the shock types described below has a distinctive physiologic hemodynamic profile corresponding with alterations in one of the variables affecting oxygen delivery described above.

CLASSIFICATION OF SHOCK

While there is a heterogeneous list of specific conditions that can cause shock, it is helpful to categorize these processes into four major shock types based on the primary physiologic derangement leading to reduced oxygen delivery and cellular hypoxia. The four major shock types are distributive, cardiogenic, hypovolemic, and obstructive. **Table 296-1** outlines these major shock types as well as specific disease processes that can result in that physiologic derangement. Each shock type has a distinct hemodynamic profile (**Table 296-2**). Familiarity with the major shock types and their unique hemodynamic profile is essential so that when evaluating a patient presenting with shock, the clinician can use the history, physical examination, and laboratory testing to determine the type of shock present and promptly begin appropriate initial therapy to restore oxygen delivery.

Table 296-1 Pathophysiologic Classification of Shock

1. Distributive
    a. Septic shock
    b. Pancreatitis
    c. Severe burns
    d. Anaphylactic shock
    e. Neurogenic shock
    f. Endocrine shock
    g. Adrenal crisis
2. Cardiogenic
    a. Myocardial infarction
    b. Myocarditis
    c. Arrhythmia
    d. Valvular
        i. Severe aortic valve insufficiency
        ii. Severe mitral valve insufficiency
3. Obstructive
    a. Tension pneumothorax
    b. Cardiac tamponade
    c. Restrictive pericarditis
    d. Pulmonary embolism
    e. Aortic dissection
4. Hypovolemic
    a. Hemorrhagic
    b. GI losses
    c. Burns
    d. Polyuria
        i. Diabetic ketoacidosis
        ii. Diabetes insipidus

Table 296-2 Hemodynamic Characteristics of the Major Types of Shock

Type of Shock	CVP	PCWP	Cardiac Output	Systemic Vascular Resistance
Distributive	↓	↓	↑	↓
Cardiogenic	↑	↑	↓	↑
Obstructive	↑	↓↑	↓	↑
Hypovolemic	↓	↓	↓	↑

Abbreviations: CVP, central venous pressure; PCWP, pulmonary capillary wedge pressure.

Distributive Shock

Distributive shock is the condition of reduced oxygen delivery where the primary physiologic disturbance is a reduction in SVR. It is unique among the types of shock in that there is a compensatory increase in CO (Table 296-2). The central venous pressure (CVP) and pulmonary capillary wedge pressure (PCWP) are usually reduced. The most common cause of distributive shock is sepsis. Sepsis has recently been redefined as the dysregulated host response to infection resulting in life-threatening organ dysfunction. When this process is accompanied by persistent hypotension requiring vasopressor support, it is classified as septic shock. Other processes that are manifest as cellular hypoxia related to a primary reduction of SVR include pancreatitis, severe burns, and liver failure. Anaphylaxis is predominantly an IgE-mediated allergic reaction that can rapidly develop after exposure to an allergen (food, medication, or insect bite), in which there is a profound distributive type of shock possibly mediated through histamine release. In this setting, there is evidence of both venous and arterial vasodilation. Studies have demonstrated extravasation of up to 35% of the circulating blood volume within 10 min. Patients with severe brain or spinal cord injury may have a reduction of SVR related to disruption of the autonomic pathways that regulate vascular tone. In these patients, there is pooling of blood in the venous system with a resulting decreased venous return and decreased CO. A final category of patients who present with distributive shock are those with adrenal insufficiency. Adrenal insufficiency may be related to chronic steroid use, metastatic malignancy, adrenal hemorrhage, infection (tuberculosis, HIV), autoimmune adrenalitis, or amyloidosis. In conditions of stress (such as infection or surgery), the deficit may become apparent with an inability to increase cortisol leading to vasodilation as well as aldosterone deficiency-mediated hypovolemia.

Cardiogenic Shock

Cardiogenic shock is characterized by reduced oxygen delivery related to a reduction in CO owing to a primary cardiac problem. There is usually a compensatory increase in SVR in cardiogenic shock. When the cardiac process (e.g., myocardial infarction) affects the left ventricle (LV), there will be elevation of the PCWP and when it affects the right ventricle (RV), the CVP will be elevated. As detailed above, the CO (and accordingly the DO_2) can be reduced by alterations in the SV or HR. In cardiogenic shock, the SV may be reduced by processes that affect myocardial contractility (myocardial infarction, ischemic cardiomyopathies, and primary myocarditis) or mechanical valvular disease (acute mitral insufficiency or aortic insufficiency). Both bradyarrhythmias and tachyarrhythmias (from either an atrial or ventricular source) may have associated hemodynamic consequence with a reduction in CO.

Hypovolemic Shock

Hypovolemic shock encompasses disease processes that reduce CO (and oxygen delivery) via a reduction in preload. In addition to the reduced CO, this shock type is characterized by an elevated SVR and low CVP and PCWP related to decreased intravascular volume. Any process causing a reduction in intravascular volume can cause shock of this type. Hypovolemic shock most commonly is related to hemorrhage, that may be external (secondary to trauma) or internal (most commonly upper

or lower gastrointestinal [GI]) bleeding. Hypovolemic shock can also be seen with nonhemorrhagic processes. Examples include GI illnesses causing profound emesis or diarrhea, renal losses (osmotic diuresis associated with diabetic ketoacidosis or diabetes insipidus), or skin loss (severe burns, inflammatory conditions such as Stevens-Johnson).

Obstructive Shock

Obstructive shock is also characterized by a reduction in oxygen delivery related to reduced CO, but in this case the etiology of the reduced CO is an extracardiac processes impairing blood flow. Processes that can impede venous return to the heart and reduce CO include tension pneumothorax (PTX), cardiac tamponade, and restrictive pericarditis. Similarly processes that obstruct cardiac outflow, such as pulmonary embolism (right heart) or aortic dissection (left heart), are included in this shock type category.

Mixed Shock

The types of shock outlined in this classification scheme are not mutually exclusive; not uncommonly, a patient will present with more than one type of shock. The initial physiologic disturbance leading to reduced perfusion and cellular hypoxia in sepsis is distributive shock. In this setting, a sepsis-induced cardiomyopathy can develop, which reduces myocardial contractility, thus producing a cardiogenic component to what now would be described as a mixed type of shock.

Undifferentiated Shock

Upon initial presentation, many patients have undifferentiated shock in which the shock type and specific disease process are not apparent. Using the history, physical examination, and initial diagnostic testing (including hemodynamic monitoring), the clinician attempts to classify a patient with one of the types of shock outlined above so that proper therapy can be initiated to restore tissue perfusion and oxygen delivery.

The type of shock seen most commonly is dependent upon the clinical area of practice. In the medical ICU, the largest number of patients have distributive shock related to sepsis. A cardiac ICU will have a population weighted toward cardiogenic or obstructive types of shock. The emergency department will see more of a mix of patients with trauma patients presenting with hypovolemic shock and septic patients having a distributive pathophysiology.

STAGES OF SHOCK

Regardless of type, shock progresses through a continuum of three stages. These stages are compensated shock (preshock), shock (decompensated shock), and irreversible shock. During compensated shock, the body utilizes a variety of physiologic responses to counteract the initial insult and attempts to reestablish the adequate perfusion and oxygen delivery. At this point, there are no overt signs of organ dysfunction. Laboratory evaluation may demonstrate mild organ dysfunction (i.e., elevated creatinine or troponin) or a mild elevation of lactate. The specific compensatory response is determined by the initial pathophysiologic defect. In early sepsis with reduction in SVR, there is a compensatory rise in HR (and CO). With early hemorrhagic volume loss, there will be a compensatory increase in SVR. As the host compensatory responses are overwhelmed, the patient transitions into true shock with evidence organ dysfunction. Appropriate interventions to restore perfusion and oxygen delivery during these initial two phases of shock can reverse the organ dysfunction. If untreated the patient will progress to the third phase of irreversible shock. At this point, the organ dysfunction is permanent and often the patient progresses to MSOF.

EVALUATION OF THE PATIENT WITH SHOCK

The evaluation of the patient with shock utilizes the history, physical examination, and diagnostic testing toward two specific aims. The first aim is confirmation of the presence of shock. Given the reversible nature of the organ dysfunction in early shock, it is important that the clinician has a high clinical suspicion for this condition. The possibility of shock should be considered all patients presenting with new organ dysfunction. This early recognition of the presence of shock is an essential tenet of shock care **(Table 296-3)**. A second aim of the initial assessment (history, physical examination, and diagnostic testing) is to identify either a specific shock etiology or to determine the type of shock present. We will discuss the role of the history, physical examination, and diagnostic testing toward these specific aims. While the assessment of shock etiology is ongoing, the initiation of therapy should not be delayed until the final diagnosis is determined. Evaluation of shock etiology and initiation of therapy should be simultaneous.

Table 296-3 Key Principles in the Treatment of Shock

1. Recognize shock early
2. Assess for type of shock present
3. Initiate therapy simultaneous with the evaluation into the etiology of shock
4. Restoration of oxygen delivery is the aim of therapy
5. Identify etiologies of shock which require additional lifesaving interventions

History

Obtaining a concise, focused history is essential. If the patient is unable to provide a history, ancillary information from anyone accompanying the patient should be obtained, and a brief chart review should be performed. As the history is being obtained, the clinician must be attentive to any details indicating new organ dysfunction. The most easily identified new organ dysfunction from the history is the presence of a newly altered mental status or decrease in renal function (oliguria). In some cases, the type of shock (and the specific disease process) is apparent from the history. Patients with distributive shock from sepsis may present with fever and a history revealing of a focal site of infection. Anaphylactic distributive shock may be suggested by the onset of hives, dyspnea, and new facial edema after exposure to common allergens. Cardiogenic shock may be identified by the onset of exertional chest discomfort. The patient with significant arrhythmia may have an initial complaint of palpitations with syncope or presyncope. Hypovolemic shock may be identified in patients who present with a history of trauma (blunt or penetrating) or GI bleed (hematemesis, melena, or bright red blood per rectum). A patient with hypertension and tearing chest or back pain may be presenting with acute aortic dissection and obstructive type shock. Acute onset chest pain with dyspnea in the setting of immobility and or underlying malignancy raises concern for obstructive shock due to pulmonary embolism.

For most patients, the specific etiology will be less clear but the history can be helpful in raising the likelihood of a particular type of shock. As an example, a patient with a preexisting immune dysfunction or medication-induced neutropenia may present with hypoperfusion and new organ dysfunction, in which the clinician must have a high suspicion for septic shock. Similarly, a patient with extensive cardiac disease requires a higher suspicion for cardiogenic shock.

Physical Examination

The physical examination should be conducted with the aim of answering two questions. Is shock present (either in compensated stage prior to overt evidence of organ dysfunction or decompensated indicated by the presence of new organ dysfunction)? Secondly, what type of shock is present (distributive, cardiogenic, hypovolemic, or obstructive)?

The physical examination findings present during the compensated phase of shock tend to be nonspecific. These include an elevation of the HR (with the body's attempt to increase CO) or tachypnea (to compensate for the developing metabolic acidosis). While nonspecific, the clinician should recognize these findings early as they may herald the development of end-organ dysfunction if perfusion and oxygen delivery are not restored. Shock is most commonly seen in the setting of circulatory failure. In most cases, this is manifest as hypotension (a mean arterial pressure [MAP] of <60 mmHg), but this finding is not always present. Many patients may have underlying conditions that cause longstanding low blood pressure without any evidence of organ dysfunction. Alternatively, patients with underlying hypertension may develop organ dysfunction at higher blood pressures.

The physical examination can confirm the presence of shock prior to the return of laboratory testing. The central nervous system (CNS), kidney, and skin are the organ systems most easily assessed for evidence of organ dysfunction. These organ systems are considered the "windows" through which we can identify organ dysfunction. Decreased oxygen delivery to the brain is manifest as confusion and encephalopathy. In the early stage of shock, the body will redirect blood flow to the CNS to maintain adequate perfusion. In the patient with shock and altered mental status, all the usual compensatory mechanisms have been outstripped by the magnitude of shock pathophysiology. New encephalopathy represents decompensated shock. To assess renal function during the physical examination, one should evaluate the patient's urine output since the time of presentation. If not already present, a urinary catheter should be placed for accurate hourly assessment of urine output. In patients with normal baseline renal function, oliguria (<0.5 mL/kg per h) may indicate shock. Finally, decreased capillary refill and cold and clammy skin are signs of hypoperfusion and shock.

Many components of the examination provide insight into hemodynamics and assist in elucidating the type of shock present. Evaluation of jugular venous pressure (JVP) and peripheral edema can provide insight into right-sided cardiac pressures. Pulmonary auscultation can identify signs of left-sided cardiac dysfunction. The physical examination may be used to differentiate shock with high CO (distributive) from that with low CO (cardiogenic shock, hypovolemic shock, and obstructive shock). Examination findings suggestive of high output shock (distributive) include warm peripheral extremities, brisk capillary refill (<2 s), and bounding pulses. Alternatively, cool extremities, delayed poor capillary refill, or weak pulses would indicate low CO forms of shock. Among those with evidence of low CO, the examination can be used to distinguish between conditions with increased intravascular filling pressure (cardiogenic shock) and intravascular volume depletion (hypovolemic shock). The JVP may be elevated cardiogenic shock (with right-sided failure) and reduced (JVP <8 cm) in hypovolemic shock. The presence of cardiogenic shock would be further supported by an S3 gallop. One must remember, however, that it is well established that patients with chronic heart failure do not present with the classical findings of acute heart failure.

At times, the physical examination may identify the specific etiology of shock. This is particularly helpful in the patient who cannot provide a detailed history. The examination may demonstrate the site of an untreated infection (cellulitis, abscess, infected pressure injury, or focal). The examination may reveal a brady- or tachyarrhythmia leading to development of shock. Similarly, large ecchymosis may indicate a significant bleed related to trauma or spontaneous retroperitoneal bleeding. The rectal examination may reveal GI hemorrhage. Pulsus paradox and elevated JVP may suggest the presence of cardiac tamponade. Patients with a tension PTX may have a paucity of breath sounds over the affected side, deviation of the trachea away from the affected side, or subcutaneous emphysema.

Combinations of easily assessed examination components have been combined to create a scoring system to identify high risk patient populations. The shock index (SI) is defined as the HR/systolic blood pressure (SBP) with a normal SI being 0.5–0.7. An elevated SI (>0.9) has been proposed to be

a more sensitive indicator of transfusion requirement and of patients with critical bleeding among those with hypovolemic (hemorrhagic) shock than either HR or BP alone. The SI may also identify patients at risk for postintubation hypotension. This concept of use of a clinical score to identify at-risk patients has been extended to patients with distributive shock from sepsis. The quick Sequential Organ Failure Assessment (qSOFA) score is a rapid assessment scale that assigns a point for SBP <100, respiratory rate >22, or altered mental status (Glasgow Coma Scale <15). A qSOFA ≥2 (with a concern for infection) is associated with a significantly greater risk of death or prolonged ICU stay. The Third International Consensus Definition of Sepsis has recommended the use of the qSOFA to identify the most acutely ill subset of patients with sepsis (longer length of stay, increased need for ICU admission, and higher in-hospital mortality).

Diagnostic Testing

Laboratory evaluation should be initiated promptly in all patients with suspected shock. The laboratory evaluation is directed toward the dual aim of assessing the extent of end-organ dysfunction and of gaining insight into the possible etiology of shock. **Table 296-4** outlines the recommended initial laboratory evaluation of the patient with undifferentiated shock.

Table 296-4 Initial Laboratory Evaluation of Undifferentiated Shock

1. Lactate
2. Renal function tests
3. Liver function tests
4. Cardiac enzymes
5. Complete blood count (with differential)
6. PT, PTT, and INR
7. Urinalysis and urine sediment
8. Arterial blood gas
9. ECG

Abbreviations: INR, International normalized ratio; PT, prothrombin time; PTT, partial thromboplastin time.

BLOOD TESTS

Evaluation of blood urea nitrogen (BUN), creatinine, and transaminases provide an assessment of the extent of end-organ dysfunction related to shock. Urine electrolytes with subsequent calculation of the fractional excretion of sodium (FENa) or fractional excretion of urea (FEUrea) may indicate states of hypovolemia or decreased effective circulating volume. Elevation of alkaline phosphatase may suggest biliary obstruction and may thereby identify a source of infection in patients with distributive shock. Elevation of cardiac enzymes can indicate a primary cardiac problem with myocyte damage related to ischemia, myocarditis, or a pulmonary embolism. An elevation of the white blood cell count may raise suspicion for an infective process, but this is certainly not diagnostic; an accompanying left shift may improve the sensitivity of this measure. While the extent of acidosis may be determined with a venous blood gas (VBG), if there is accompanying hypoxemia an arterial blood gas should be obtained. For patients with undifferentiated shock, there should always be a high index of suspicion for

possible infection. Urinalysis and urine sediment should be sent to evaluate for pyuria. Blood cultures, urine cultures, and sputum cultures should be obtained. Radiographic evaluation should be directed to seek sources of infection suggested by the history and physical examination.

Lactate measurement has a role in the diagnosis, risk stratification, and, potentially, the treatment of shock. Increased lactate (hyperlactemia) and lactic acidosis (hyperlactemia and pH <7.35) are common in shock. Lactate is a product of anaerobic glucose metabolism. In glycolysis, the enzyme phosphofuctokinase metabolizes glucose to pyruvate. Under aerobic conditions, the pyruvate is then converted (in the mitochondria) to acetyl CoA and enters the Krebs cycle with resulting ATP generation through oxidative phosphorylation. In the setting of cellular hypoxia, the Krebs (tricarboxylic acid) cycle cannot oxidize the pyruvate, and thus, the pyruvate is converted to lactate by the enzyme lactate dehydrogenase. Under normal conditions, lactate is produced from skeletal muscle, brain, skin, and intestine. In the setting of reduced oxygen delivery and cellular hypoxia, the amount of lactate produced from these tissues increases (and other tissue can begin to produce lactate). While most of the studies have been performed in patients with septic shock, there is evidence that elevated lactate correlates with a worse outcome. A recent systematic literature review evaluating the role of lactate measurement in a variety of critically ill populations supported the value of serial lactate measurements in the evaluation of critically ill patients and their response to therapy.

ECG

The electrocardiogram (ECG) is an essential part of the evaluation of the patient with shock. There may be a bradycardia or tachycardic arrhythmia causing a reduction in CO. ST segment elevation myocardial infarction may be identified. The presence of the S1 Q3 T3 pattern would raise concerns for pulmonary embolism. Reduced voltage in the presence of electrical alternans raises the possibility of pericardial tamponade.

Echocardiography

Echocardiography is increasingly used as an essential tool to help categorize shock, and it provides an assessment that is both rapid and noninvasive. Familiarity with basic echocardiographic techniques and interpretation is now expected in the critical care setting. Accordingly, competency standards have been proposed for critical care providers in both basic and advanced echocardiographic techniques. The bedside echocardiogam performed by the ICU team does not replace a formal examination performed by the echocardiography service.

The basic echocardiographic assessment for the shock patient is transthoracic echocardiography (TTE) utilizing both the two-dimensional (2D) and M mode. Standardized, focused echocardiography protocols such as the RACE protocol (rapid assessment for cardiac echocardiography) have been introduced to facilitate the assessment of cardiac function. It focuses the examination on LV function, RV function, and pericardium. It also can assess volume, but the use of echocardiography for volume assessment will be discussed in the section below.

The 2D mode can evaluate LV size, wall thickness, and ventricular function. Ventricular size and thickness can suggest longer standing cardiac processes. Evaluation of LV function through estimation of left ventricular ejection fraction (LVEF), and can identify shock with globally reduced LV function or regional wall motion abnormalities. Similarly, the assessment of RV function also examines RV size and wall thickness (to identify conditions such as elevated pulmonary pressures or suggest pulmonary embolism), and also evaluate the patient for pericardial tamponade. Two-dimensional echocardiography can also be used to assess valve function, including acute processes, such as mitral valve rupture. Assessment of valvular function is often a process that requires a higher skilled practitioner. The performance of the bedside echocardiogram by the critical care practitioner does not replace formal assessment by a cardiologist.

INITIAL TREATMENT of SHOCK

Since shock can progress rapidly to an irreversible stage, a key principle in shock management is to initiate treatment for circulatory shock simultaneous with efforts to elucidate shock etiology (Table 296-3). If the initial history, physical examination, and laboratory evaluation have identified the shock type or the specific etiology, then therapy is directed to reverse the underlying physiologic abnormality causing the hypoperfusion and reduced oxygen delivery. Details of the optimal care for the specific disease processes leading to shock may be found in other chapters of this text. As many patients will present with undifferentiated shock, in this section we will discuss treatment directed at the patient with undifferentiated shock. At the conclusion of this section, we will highlight etiologies of shock that require initiation of lifesaving specific therapy.

The development of shock is a medical emergency, and optimal therapy involves the involvement of a multidisciplinary team to allow the evaluation and initiation of therapy to begin simultaneously. Patients must be treated in a setting where adequate resources are available to support frequent reassessments and invasive monitoring. Most patients with shock should be cared for in an ICU setting.

A key early consideration is to ensure adequate intravenous access. Placement of a peripheral venous catheter (16G or 18G) will provide initial access for the aggressive volume resuscitation that is required for patients with distributive or hypovolemic shock. If there is concern for distributive shock with sepsis, this IV access will also permit prompt antibiotic administration. For patients with ongoing hypotension despite adequate volume resuscitation, placement of a central venous catheter (CVC) is indicated to provide therapy with vasopressors and inotropes. The CVC will provide a mechanism for hemodynamic monitoring (CVP) as well as a means to obtain central venous oxygen saturations ($ScvO_2$). The $ScvO_2$ is a surrogate of mixed venous oxygen saturation, and, thus, can provide insight into the adequacy of oxygen delivery. Central venous access using a sheath will provide an access point for placement of a Swan Ganz catheter if more detailed assessment of hemodynamic measurements are required (PCWP, CO, and SVR). If the patient presents critically ill or in the midst of cardiopulmonary arrest, the quickest method of obtaining central access will be through the use of an intraosseous device. Placement of an arterial line allows for intravascular measurement of blood pressure and continuous determination of MAP. In addition, it can provide insight into the adequacy of volume resuscitation through the measurement of systolic or pulse pressure variation. The arterial line will provide access for determination of arterial oxygen tension, which is helpful since peripheral oximetry measurements (SpO_2) can be unreliable in states of tissue hypoperfusion. The arterial line facilitates repeated measures of acid base status or lactate to assess the impact of treatment. All patients with shock should have a urinary catheter placed to permit hourly assessment of renal function as another potential indication of the adequacy of resuscitation.

Volume Resuscitation

Initial volume resuscitation has the aim of restoring tissue perfusion and is crucial to optimal shock therapy. Assessment of current intravascular volume status and determination of the optimal amount of volume resuscitation are challenging. The physiologic goal of volume resuscitation is to move the patient to the nonpreload-dependent portion of the Starling curve. Most patients with any of the four shock types will benefit from an increase in intravascular volume. For patients with distributive shock, the need for early aggressive volume replacement is well established. In the past, the use of early goal-directed therapy (EGDT) in septic shock targeted specific measures of CVP, MAP, and SvO_2 to guide volume resuscitation (and initiation of vasopressors and inotropes). More recent studies have demonstrated that targeted resuscitation using invasive monitoring is not required, but in all of these studies patients in the “usual care” arms of the study received early initial volume resuscitation. For patients with suspected septic shock, a minimum of 30 mL/kg is recommended by the Surviving Sepsis Campaign. While the need for volume resuscitation is most apparent for patients with

distributive or hypovolemic shock, even patients with cardiogenic shock may benefit by cautious volume replacement. In these patients, there should be a careful assessment of volume status prior to volume administration.

In general, volume replacement therapy should be given as a bolus with a predefined endpoint to assess the effect of the volume resuscitation. Most commonly, the volume resuscitation will begin with crystalloid. In patients with hypovolemic shock due to ongoing hemorrhage, volume replacement with packed red blood cells is warranted. In cases of massive transfusion, platelets and fresh frozen plasma should be provided to offset the dilution of these components during volume replacement. Since hemoglobin is a key determinant of $CaCO_2$ red cell administration may be a part of volume replacement even without hemorrhage if hemoglobin content is <7 g/dL in order to optimize oxygen delivery.

Assessment of intravascular volume status (and the adequacy of volume resuscitation) begins with the physical examination (described above). The passive leg raise (PLR) test can predict responsiveness to additional intravenous fluid (IVF) by providing the patient with an endogenous volume bolus. While the patient is resting in a semi-recumbent position at a 45-degree angle, the bed is placed in trendelenburg such that the patient's head becomes horizontal and the legs are extended at a 45-degree angle. There is then an immediate (within 1 min) assessment of changes in CO (or pulse pressure variation as a surrogate). It is important to emphasize that one does not merely look for changes in blood pressure; if the shock patient is mechanically ventilated there is the option of looking at changes in SV variation (or pulse pressure variation) during the respiratory cycle to assess volume responsiveness. A >12% SV variation suggests a volume-responsive state. This measurement requires that the patient be in a volume cycle mode of ventilation, without breath-to-breath variations in intrathoracic pressure and without arrhythmias. A final caveat to the use of these parameters to assess volume status is that these studies are performed on patients being ventilated with tidal volumes larger than currently used to minimize ventilator-induced lung injury.

There is also increased use of echocardiography to assist in determination of intravascular fluid status, with a variety of static and dynamic variables that the trained operator can assess. The most commonly used parameters to assess adequacy of volume resuscitation are inferior vena cava (IVC) diameter and IVC collapse. Alternatively, serial assessments of LV function can be performed while volume is being administered. Placement of a pulmonary artery catheter (PAC) is another tool for assessment of volume status. This more invasive measure involves placement of the PAC into the central venous circulation and through the right heart. Ports in the PAC (Swan Ganz catheter) allow for direct measurement of CVP, pulmonary artery (PA), and PCWPs. The PCWP is used as a surrogate for LA pressure. While studies have not identified a mortality or length-of-stay benefit with routine use of PA catheterization, there are cases where it may be beneficial. Patients with mixed shock (distributive and cardiogenic) or those with ongoing shock of unclear etiology are examples of situations in which it should be considered.

The need for continued volume replacement must be frequently reassessed. As the patient continues to receive treatment for shock, the initial proper strategy regarding volume management may change in light of development of processes that independently require a different volume management strategy. For patients who initially present with shock but then develop failure related to acute respiratory distress syndrome (ARDS) or renal failure, it may be reasonable to begin volume removal.

Vasopressor and Inotropic Support

If intravascular volume status has been optimized with volume resuscitation but hypotension and inadequate tissue perfusion persist, then vasopressor and inotropic support should be initiated. The use of vasopressors and inotropes must be tailored to the primary physiologic disturbance. The clinician must understand the receptor selectivity of various agents and that for some agents the selectivity may be dose-dependent. In patients with distributive shock, the aim is to increase the SVR.

Norepinephrine is the first choice vasopressor: with potent α_1 and β_1 adrenergic effects. The α_1 causes vasoconstriction while β_1 has positive inotropic and chronotropic effects. At high doses, epinephrine has a similar profile (at lower doses the β effects predominate), but is associated with tachyarrhythmia, myocardial ischemia, decreased splanchnic blood flow, pulmonary hypertension, and acidosis. In distributive shock, vasopressin deficiency may be present. Vasopressin acts on the vasopressin receptor to reverse vasodilation and redistribute flow to the splanchnic circulation. In a randomized trial in patients with septic shock, the addition of low-dose vasopressin did not reduce all-cause 28-day mortality compared to norepinephrine. Vasopressin is safe and has a role as a second agent for hypotension in septic shock. Dopamine does not have a role as a first line agent in distributive shock. A randomized control study in patients with all cause circulatory shock did not show a survival benefit, but did reveal an increase in adverse events (arrhythmia). In this study, the subgroup of patients with cardiogenic shock had increased mortality. For patients with cardiogenic shock, dobutamine is the first line agent; it is a synthetic catecholamine with primarily β-mediated effects and minimal α adrenergic effects. The β_1 effect is manifest in increased inotropy and the β_2 effect leads to vasodilation with decreased afterload; it can be used with norepinephrine in patients with mixed distributive and cardiogenic shock.

OXYGENATION AND VENTILATION SUPPORT

In addition to the cellular hypoxia caused by the circulatory failure, patients with shock may present with hypoxemia. For patients with distributive shock, this may be related to a primary pulmonary process (pneumonia in a patient with septic shock). For patients with cardiogenic or obstructive shock, the hypoxemia may be related to LV dysfunction and elevations of PCWP. For patients with all types of shock, there can be development of ARDS and subsequent $\dot{V}/\dot{Q}$ mismatch and shunt. Supplemental oxygen should be initiated and titrated to maintain SpO_2 of 92–95%. This may require intubation and initiation of mechanical ventilation. If the patient requires intubation and initiation of mechanical ventilation, this should be provided promptly so as to minimize the duration of tissue hypoxia. Patients with shock may have high minute ventilatory needs to compensate for metabolic acidosis. As shock progresses, they may not be able to maintain adequate respiratory compensation, which may be a second indication to initiate mechanical ventilator support. If mechanical support is initiated, it is important to provide ventilation with lung-protective strategies focused on low tidal volume ventilation and optimization of positive end-expiratory pressure to minimize ventilator-induced lung injury. In addition, there should be daily sedation cessation to assess underlying neurologic function and minimize time on mechanical ventilation. There are currently little data to support the use of noninvasive ventilation in the setting of shock.

Antibiotic Administration

Sepsis and septic shock are the most common cause of shock. For patients presenting with undifferentiated shock, if the diagnosis of septic shock is being entertained then broad spectrum antibiotics should be administered after obtaining appropriate cultures. For patients with sepsis, every hour delay in antibiotic administration is associated with an increase in mortality. While it is ideal to initiate antibiotics after appropriate cultures, the inability to obtain cultures should not delay the start of treatment. When sepsis is excluded as a cause of shock, an important aspect of antibiotic stewardship is to stop all antibiotics.

Specific Causes of Shock Requiring Tailored Intervention

The initial evaluation (history, physical examination, and diagnostic testing) may have identified an etiology of shock that requires urgent lifesaving intervention in addition to the initial treatment steps outlined above. Patients with distributive shock secondary to anaphylaxis require removal of the inciting allergen, administration of epinephrine, and vascular support with intravenous fluid

resuscitation and vasopressors. Adrenal insufficiency requires replacement with intravenous stress dose steroids. Cardiogenic shock patients with arrhythmia may require treatment as outlined in advanced cardiac life support algorithms or placement of an artificial pacemaker. In cases of acute ischemic events, consideration must be given to revascularization and temporary mechanical supportive measures. In the case of valve dysfunction, emergency surgery may be considered. Patients with hypovolemic shock due to hemorrhage may require surgical intervention in the case of trauma or endoscopic or interventional radiology procedures in the case of a GI source of blood loss. Among patients with obstructive shock, a tension PTX would necessitate immediate decompression. Proximal pulmonary embolism requires evaluation for thrombolytic therapy or surgical removal of the clot. Dissection of the ascending aorta may require surgical intervention.

FURTHER READING

Mebazaa A et al: Acute heart failure and cardiogenic shock: A multidisciplinary practical guidance. Intensive Care Med 42:147, 2016. [PubMed: 26370690]
Monnet X et al: Passive leg raising for predicting fluid responsiveness: A systematic review and meta-analysis. Intensive Care Med 42:1935, 2016. [PubMed: 26825952]
Pro CI et al: A randomized trial of protocol-based care for early septic shock. N Engl J Med 370:1683, 2014. [PubMed: 24635773]
Rhodes A et al: Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016. Intensive Care Med 43:304, 2017. [PubMed: 28101605]
Vincent JL, De Backer D: Circulatory shock. N Engl J Med 369:1726, 2013. [PubMed: 24171518]
Vincent JL et al: The value of blood lactate kinetics in critically ill patients: A systematic review. Crit Care 20:257, 2016. [PubMed: 27520452]

Chapter 297: Sepsis and Septic Shock

Christopher W. Seymour; Derek C. Angus

INTRODUCTION AND DEFINITIONS

Sepsis is a common and deadly disease. More than two millennia ago, Hippocrates wrote that sepsis was characterized by rotting flesh and festering wounds. Several centuries later, Galen described sepsis as a laudable event required for wound healing. Once the germ theory was proposed by Semmelweis, Pasteur, and others in the nineteenth century, sepsis was recast as a systemic infection referred to as “blood poisoning” and was thought to be due to pathogen invasion and spread in the bloodstream of the host. However, germ theory did not fully explain sepsis: many septic patients died despite successful removal of the inciting pathogen. In 1992, Bone and colleagues proposed that the host, not the germ, was responsible for the pathogenesis of sepsis. Specifically, they defined sepsis as a systemic inflammatory response to infection. Yet sepsis arose in response to many different pathogens, and septicemia was neither a necessary condition nor a helpful term. Thus, these investigators instead proposed the term *severe sepsis* to describe cases where sepsis was complicated by acute organ dysfunction and the term *septic shock* for a subset of sepsis cases that were complicated by hypotension despite adequate fluid resuscitation along with perfusion abnormalities.

In the past 20 years, research has revealed that many patients develop acute organ dysfunction in response to infection but without a measurable inflammatory excess (i.e., without the systemic inflammatory response syndrome [SIRS]). In fact, both pro- and anti-inflammatory responses are present along with significant changes in other pathways. To clarify terminology and reflect the current understanding of the pathobiology of sepsis, the Sepsis Definitions Task Force in 2016 proposed the Third International Consensus Definitions specifying that *sepsis* is a dysregulated host response to infection that leads to acute organ dysfunction. This definition distinguishes sepsis from uncomplicated infection that does not lead to organ dysfunction, a poor course, or death. In light of the wide variation in the ways that septic shock is identified in research, clinical, or surveillance settings, the Third International Consensus Definitions further specified that *septic shock* be defined as a subset of sepsis cases in which underlying circulatory and cellular/metabolic abnormalities are profound enough to substantially increase mortality risk.

To aid clinicians in identifying sepsis and septic shock at the bedside, new “Sepsis-3” clinical criteria for sepsis include (1) a suspected infection and (2) acute organ dysfunction, defined as an increase by two or more points from baseline (if known) on the sequential (or sepsis-related) organ failure assessment (SOFA) score **(Table 297-1)**. Criteria for septic shock include sepsis plus the need for vasopressor therapy to elevate mean arterial pressure to ≥65 mmHg with a serum lactate concentration >2.0 mmol/L despite adequate fluid resuscitation.

TABLE 297-1 Definitions and Criteria for Sepsis and Septic Shock

Condition	Definition	Common Clinical Features	Criteria in 1991/2003 (“Sepsis-1”/”Sepsis-2”)	Criteria in 2016 (“Sepsis-3”)
Sepsis	A life-threatening organ dysfunction caused by a dysregulated host response to infection	Include signs of infection, with organ dysfunction, plus altered mentation; tachypnea; hypotension; hepatic, renal, or hematologic dysfunction	Suspected (or documented) infection plus ≥2 systemic inflammatory response syndrome (SIRS) criteria[a]	Suspected (or documented) infection and an acute increase in ≥2 sepsis-related organ failure assessment (SOFA) points[b]
Septic shock	A subset of sepsis in which underlying circulatory and cellular/metabolic abnormalities lead to substantially increased mortality risk	Signs of infection, plus altered mentation, oliguria, cool peripheries, hyperlactemia	Suspected (or documented) infection plus persistent arterial hypotension (systolic arterial pressure, <90 mmHg; mean arterial pressure, <60 mmHg; or change in systolic by >40 mmHg from baseline	Suspected (or documented) infection plus vasopressor therapy needed to maintain mean arterial pressure at ≥65 mmHg and serum lactate >2.0 mmol/L despite adequate fluid resuscitation

[a]SIRS criteria include 1 point for each of the following (score range, 0–4): fever >38°C (>100.4°F) or <36°C (<96.8°F); tachypnea with >20 breaths per min; tachycardia with heart rate >90 beats per min; leukocytosis with white blood cell count >12,000/μL; leukopenia (<4000/μL) or >10% bands. [b]SOFA score is a 24-point measure of organ dysfunction that uses six organ systems (renal, cardiovascular, pulmonary, hepatic, neurologic, hematologic), where 0–4 points are assigned per organ system.

ETIOLOGY

Sepsis can arise from both community-acquired and hospital-acquired infections. Of these infections, pneumonia is the most common source, accounting for about half of cases; next most common are intraabdominal and genitourinary infections. Blood cultures are typically positive in only one-third of cases, while many cases are culture negative at all sites. *Staphylococcus aureus* and *Streptococcus pneumoniae* are the most common gram-positive isolates, while *Escherichia coli*, *Klebsiella* species, and *Pseudomonas aeruginosa* are the most common gram-negative isolates. In recent years, gram-positive infections have been reported more often than gram-negative infections, yet a 75-country point-prevalence study of 14,000 patients on intensive care units (ICUs) found that 62% of positive isolates were gram-negative bacteria, 47% were gram-positive bacteria, and 19% were fungi.

The many risk factors for sepsis are related to both the predisposition to develop an infection and, once infection develops, the likelihood of developing acute organ dysfunction. Common risk factors for increased risk of infection include chronic diseases (e.g., HIV infection, chronic

obstructive pulmonary disease, cancers) and immunosuppression. Risk factors for progression from infection to organ dysfunction are less well understood but may include underlying health status, preexisting organ function, and timeliness of treatment. Age, sex, and race/ethnicity all influence the incidence of sepsis, which is highest at the extremes of age, higher in males than in females, and higher in blacks than in whites. The differences in risk of sepsis by race are not fully explained by socioeconomic factors or access to care, raising the possibility that other factors, such as genetic differences in susceptibility to infection or in the expression of proteins critical to the host response, may play a role.

EPIDEMIOLOGY

The incidences of sepsis and septic shock depend on how acute organ dysfunction and infection are defined as well as on which data sources are studied. Disparate estimates come from administrative data, prospective cohorts with manual case identification, and large electronic health-record databases. Organ dysfunction is often defined by the provision of supportive therapy, in which case epidemiologic studies count the "treated," rather than the actual, incidence. In the United States, recent cohort studies using administrative data suggest that upwards of 2 million cases of sepsis occur annually. Shock is present in ~30% of cases, resulting in an estimated 230,000 cases in a recent systematic review. An analysis of data (both clinical and administrative) from 300 hospitals in the United Healthcare Consortium estimated that septic shock occurred in 19 per 1000 hospitalized encounters. The incidences of sepsis and septic shock are also reported to be increasing (according to ICD9-CM diagnosis and procedure codes), with a rise of almost 50% in the past decade. However, the stability of objective clinical markers (e.g., provision of organ support, detection of bacteremia) over this period in a two-center validation study suggests that new ICD-9 coding rules, confusion over semantics (e.g., *septicemia* versus *severe sepsis*), rising capacity to provide intensive care, and increased case-finding confound the interpretation of serial trends. Studies from other high-income countries report rates of sepsis in the ICU similar to those in the United States.

While the data demonstrate that sepsis is a significant public-health burden in high-income countries, its impact on the populations of low- and middle-income countries is probably even more substantial because of the increased incidence of infectious diseases and the high prevalence of HIV in some parts of the developing world. Although there are fewer high-quality studies on sepsis in these countries, the available data support sepsis as a major public-health problem. For example, a study of one cohort in rural Uganda found an incidence of laboratory-confirmed sepsis tenfold that of current global sepsis estimates; as only a minority of patients with sepsis develop bacteremia, the incidence of sepsis in the cohort was probably even higher. Case–fatality rates in low- and middle-income countries are also higher than those in high-income countries, as exemplified by two observational cohorts in Brazil with mortality rates >40%.

PATHOGENESIS

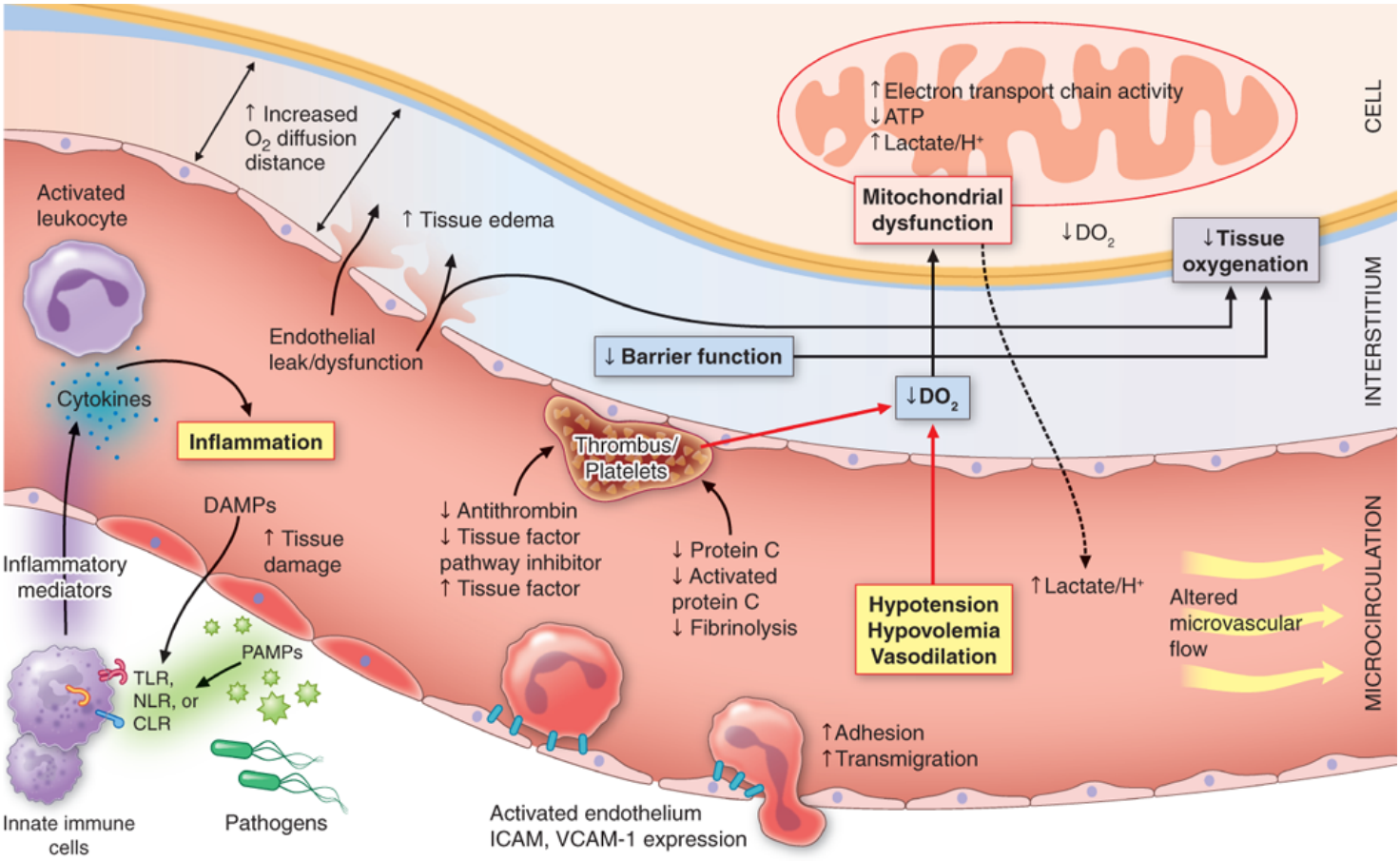
For many years, the clinical features of sepsis were considered the result of an excessive inflammatory host response (SIRS). More recently, it has become apparent that infection triggers a much more complex, variable, and prolonged host response than was previously thought. The specific response of each patient depends on the pathogen (load and virulence) and the host (genetic composition and comorbidity), with different responses at local and systemic levels. The host response evolves over time with the patient's clinical course. Generally, proinflammatory reactions (directed at eliminating pathogens) are responsible for "collateral" tissue damage in sepsis, whereas anti-inflammatory responses are implicated in the enhanced susceptibility to secondary infections that occurs later in the course. These mechanisms can be characterized as an interplay between two "fitness costs": direct damage to organs by the pathogen and damage to organs stemming from the host's immune response. The host's ability to resist as well as tolerate both direct and immunopathologic damage will determine whether uncomplicated infection becomes sepsis.

Initiation of Inflammation

Over the past decade, our knowledge of pathogen recognition has increased tremendously. Pathogens activate immune cells by an interaction with pattern recognition receptors **(Fig. 297-1)**, of which four main classes are prominent: Toll-like receptors (TLRs), RIG-I-like receptors, C-type lectin receptors, and NOD-like receptors; the activity of the last group occurs partially in protein complexes called *inflammasomes*. The recognition of structures conserved across microbial species—so-called pathogen-associated molecular patterns (PAMPs)—by all these receptors results in upregulation of inflammatory gene transcription and initiation of innate immunity. A common PAMP is the lipid A moiety of lipopolysaccharide (LPS or endotoxin), which attaches to the LPS-binding protein on the surface of monocytes, macrophages, and neutrophils. LPS is transferred to and signals via TLR4 to produce and release cytokines such as tumor necrosis factor that grow the signal and alert other cells and tissues. Up to 10 TLRs have been identified in humans.

FIGURE 297-1

Select mechanisms implicated in the pathogenesis of sepsis-induced organ and cellular dysfunction. The host response to sepsis involves multiple mechanisms that lead to decreased oxygen delivery (DO_2) at the tissue level. The duration, extent, and direction of these interactions are modified by the organ under threat, host factors (e.g., age, genetic characteristics, medications), and pathogen factors (e.g., microbial load and virulence). The inflammatory response is typically initiated by an interaction between pathogen-associated molecular patterns (PAMPs) expressed by pathogens and pattern recognition receptors expressed by innate immune cells on the cell surface (Toll-like receptors [TLRs] and C-type lectin receptors [CLRs]), in the endosome (TLRs), or in the cytoplasm (retinoic acid inducible gene 1–like receptors and nucleotide-binding oligomerization domain–like receptors [NLRs]). The resulting tissue damage and necrotic cell death lead to release of damage-associated molecular patterns (DAMPs) such as uric acid, high-mobility group protein B1, S100 proteins, and extracellular RNA, DNA, and histones. These molecules promote the activation of leukocytes, leading to greater endothelial dysfunction, expression of intercellular adhesion molecule (ICAM) and vascular cell adhesion molecule 1 (VCAM-1) on the activated endothelium, coagulation activation, and complement activation. This cascade is compounded by macrovascular changes such as vasodilation and hypotension, which are exacerbated by greater endothelial leak tissue edema, and relative intravascular hypovolemia. Subsequent alterations in cellular bioenergetics lead to greater glycolysis (e.g., lactate production), mitochondrial injury, release of reactive oxygen species, and greater organ dysfunction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

At the same time, these receptors also sense endogenous molecules released from injured cells—so-called damage-associated molecular patterns (DAMPs), such as high-mobility group protein B1, S100 proteins, and extracellular RNA, DNA, and histones. The release of DAMPs during sterile injuries such as those incurred during trauma gives rise to the concept that the pathogenesis of multiple-organ failure may be similar in sepsis and noninfectious critical illness. In addition to activating the proinflammatory cytokines, the inflammatory responses implicated in the pathogenesis of sepsis also activate the complement system, platelet-activating factor, arachidonic acid metabolites, and nitric oxide.

Coagulation Abnormalities

Sepsis is commonly associated with coagulation disorders and frequently leads to disseminated intravascular coagulation. Abnormalities in coagulation are thought to isolate invading microorganisms and/or to prevent the spread of infection and inflammation to other tissues and organs. Excess fibrin deposition is driven by coagulation via tissue factor, a transmembrane glycoprotein expressed by various cell types; by impaired anticoagulant mechanisms, including the protein C system and antithrombin; and by compromised fibrin removal due to depression of the fibrinolytic system. Coagulation (and other) proteases further enhance inflammation via protease-activated receptors. In infections with endothelial predominance (e.g., meningococcemia), these mechanisms can be common and deadly.

Organ Dysfunction

Although the mechanisms that underlie organ failure in sepsis are only partially known, impaired tissue oxygenation plays a key role. Several factors contribute to reduced oxygen delivery in sepsis and septic shock, including hypotension, reduced red-cell deformability, and microvascular thrombosis. Inflammation can cause dysfunction of the vascular endothelium, accompanied by cell death and loss of barrier integrity, giving rise to subcutaneous and body-cavity edema. An excessive and uncontrolled release of nitric oxide causes vasomotor collapse, opening of arteriovenous shunts, and pathologic shunting of oxygenated blood from susceptible tissues. In addition, mitochondrial damage due to oxidative stress and other mechanisms impairs cellular oxygen utilization. The slowing of oxidative metabolism, in parallel with impaired oxygen delivery, reduces cellular O_2 extraction. Yet energy (i.e., ATP) is still needed to support basal, vital cellular function, which derives from glycolysis and fermentation and thus yields H^+ and lactate. With severe or prolonged insult, ATP levels fall beneath a critical threshold, bioenergetic failure ensues, toxic reactive oxygen species are released, and apoptosis leads to irreversible cell death and organ failure. The actual morphologic changes in sepsis-induced organ failure are also complex. Generally, organs such as the lung undergo extensive microscopic changes, while other organs may undergo rather few histologic changes. In fact, some organs (e.g., the kidney) may lack significant structural damage while still having significant tubular-cell changes that impair function.

Anti-Inflammatory Mechanisms

The immune system harbors humoral, cellular, and neural mechanisms that may exacerbate the potentially harmful effects of the proinflammatory response. Phagocytes can switch to an anti-inflammatory phenotype that promotes tissue repair, while regulatory T cells and myeloid-derived suppressor cells further reduce inflammation. The so-called neuroinflammatory reflex may also contribute: sensory input is relayed through the afferent vagus nerve to the brainstem, from which the efferent vagus nerve activates the splenic nerve in the celiac plexus, with consequent norepinephrine release in the spleen and acetylcholine secretion by a subset of CD4+ T cells. The acetylcholine release targets α7 cholinergic receptors on macrophages, reducing proinflammatory cytokine release. Disruption of this neural-based system by vagotomy renders animals more vulnerable to endotoxin shock, while stimulation of the efferent vagus nerve or α7 cholinergic receptors attenuates systemic inflammation in experimental sepsis.

Immune Suppression

Patients who survive early sepsis but remain dependent on intensive care occasionally demonstrate evidence of a suppressed immune system. These patients may have ongoing infectious foci despite antimicrobial therapy or may experience the reactivation of latent viruses. Multiple investigations have documented reduced responsiveness of blood leukocytes to pathogens in patients with sepsis; these findings were recently corroborated by post-mortem studies revealing strong functional impairments of splenocytes harvested from ICU patients who died of sepsis. Immune suppression was evident in the lungs as well as the spleen; in both organs, the expression of ligands for T cell–inhibitory receptors on parenchymal cells was increased. Enhanced apoptotic cell death, especially of B cells, CD4+ T cells, and follicular dendritic cells, has been implicated in sepsis-associated immune suppression and death. In a cohort of >1000 ICU admissions for sepsis, secondary infections developed in 14% of patients, and the associated genomic response at the time of infection was consistent with immune suppression, including impaired glycolysis and cellular gluconeogenesis. The most common secondary infections included catheter-related bloodstream infections, ventilator-associated infections, and abdominal infections. What is not yet understood is the optimal way to identify those sepsis patients who have hyperinflamed rather than immunosuppressed phenotypes. Similarly, it is unknown whether the dysfunctional immune system is driving organ dysfunction and secondary infections or whether the immune system itself is just another dysfunctional organ.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Sepsis and Septic Shock

At the bedside, a clinician begins by asking, “Is this patient septic?” Consensus criteria for sepsis and septic shock agree on core diagnostic elements, including suspected or documented infection accompanied by acute, life-threatening organ dysfunction. If infection is documented, the clinician must determine the inciting cause and the severity of organ dysfunction, usually by asking: “What just happened?” Severe infection can be evident, but it is often quite difficult to recognize. Many infection-specific biomarkers and molecular diagnostics are under study to help discriminate sterile inflammation from infection, but these tools are not commonly used. The clinician’s acumen is still crucial to the diagnosis of infection. Next, the primary physiologic manifestations of organ dysfunction can be assessed quickly at the bedside with a six-organ framework, yielding the SOFA score. Particular focus should then be placed on the presence or absence of shock, which constitutes a clinical emergency. The general manifestations of shock include arterial hypotension with evidence of tissue hypoperfusion (e.g., oliguria, altered mental status, poor peripheral perfusion, or hyperlactemia).

CLINICAL MANIFESTATIONS

The specific clinical manifestations of sepsis are quite variable, depending on the initial site of infection, the offending pathogen, the pattern of acute organ dysfunction, the underlying health of the patient, and the delay before initiation of treatment. The signs of both infection and organ dysfunction may be subtle. Guidelines provide a long list of potential warning signs of incipient sepsis (Table 297-1). Once sepsis has been established and the inciting infection is assumed to be under control, the temperature and white blood cell (WBC) count often return to normal. However, organ dysfunction typically persists.

Cardiorespiratory Failure

Two of the most commonly affected organ systems in sepsis are the respiratory and cardiovascular systems. Respiratory compromise classically manifests as acute respiratory distress syndrome (ARDS), defined as hypoxemia and bilateral infiltrates of noncardiac origin that arise within 7 days of the suspected infection. ARDS can be classified by Berlin criteria as mild (PaO_2/FiO_2, 201–300 mmHg), moderate (101–200 mmHg), or severe (≤100 mmHg). A common competing diagnosis is hydrostatic edema secondary to cardiac failure or volume overload. Although traditionally identified by elevated pulmonary capillary wedge measurements from a pulmonary artery catheter (>18 mmHg), cardiac failure can be objectively evaluated on the basis of clinical judgment or focused echocardiography.

Cardiovascular compromise typically presents as hypotension. The cause can be frank hypovolemia, maldistribution of blood flow and intravascular volume due to diffuse capillary leakage, reduced systemic vascular resistance, or depressed myocardial function. After adequate volume expansion, hypotension frequently persists, requiring the use of vasopressors. In early shock, when volume status is reduced, systemic vascular resistance may be quite high with low cardiac output; after volume repletion, however, this picture may rapidly change to low systemic vascular resistance and high cardiac output.

Kidney Injury

Acute kidney injury (AKI) is documented in >50% of septic patients, increasing the risk of in-hospital death by six- to eightfold. AKI manifests as oliguria, azotemia, and rising serum creatinine levels and frequently requires dialysis. The mechanisms of sepsis-induced AKI are incompletely understood. AKI may occur in up to 25% of patients in the absence of overt hypotension. Current mechanistic work suggests that a combination of diffuse microcirculatory blood-flow abnormalities, inflammation, and cellular bioenergetic responses to injury contribute to sepsis-induced AKI beyond just organ ischemia.

Neurologic Complications

Typical central nervous system dysfunction presents as coma or delirium. Imaging studies typically show no focal lesions, and electroencephalographic findings are usually consistent with nonfocal encephalopathy. Sepsis-associated delirium is considered a diffuse cerebral dysfunction caused by the inflammatory response to infection without evidence of a primary central nervous system infection. Consensus guidelines recommend delirium screening with valid and reliable tools such as the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) and the Intensive Care Delirium Screening Checklist (ICDSC). Critical-illness polyneuropathy and myopathy are also common, especially in patients with a prolonged course. For survivors of sepsis, neurologic complications can be severe. In a national (U.S.) representative prospective cohort of >1000 elderly patients with severe sepsis, moderate to severe cognitive impairment increased by 10.6 percentage points among patients who survived severe sepsis (odds ratio, 3.34; 95% confidence interval [CI], 1.53–7.25) over that among survivors of nonsepsis hospitalizations. Many of these limitations persisted for up to 8 years.

Additional Manifestations

Many other abnormalities occur in sepsis, including ileus, elevated aminotransferase levels, altered glycemic control, thrombocytopenia and disseminated intravascular coagulation, adrenal dysfunction, and sick euthyroid syndrome. Adrenal dysfunction in sepsis is widely studied and is thought to be related more to reversible dysfunction of the hypothalamic–pituitary axis or tissue glucocorticoid resistance than to direct damage to the adrenal gland. The diagnosis is difficult to establish. Recent clinical practice guidelines do not recommend use of the adrenocorticotropic hormone stimulation test or determination of the plasma cortisol level to detect relative glucocorticoid insufficiency.

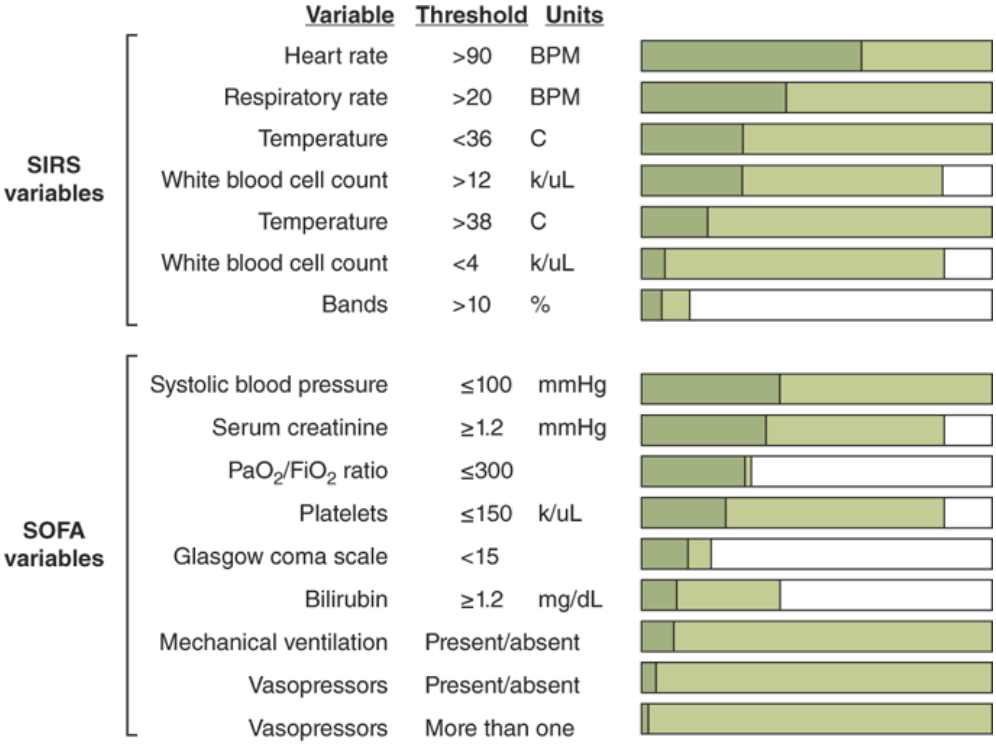
DIAGNOSIS

Laboratory and Physiologic Findings

A variety of laboratory and physiologic changes are found in patients with suspected infection who are at risk for sepsis. In a 12-hospital cohort of electronic health records related to >70,000 encounters **(Fig. 297-2)**, only tachycardia (heart rate, >90 beats per min) was present in >50% of encounters; the most common accompanying abnormalities were tachypnea (respiratory rate, >20 breaths per min), hypotension (systolic blood pressure, ≤100 mmHg), and hypoxia (SaO_2, ≤90%). Leukocytosis (WBC count, >12,000/μL) was present in fewer than one-third of patients and leukopenia (WBC count, <4000/μL) in fewer than 5%. Notably, many features that may identify acute organ dysfunction, such as platelet count, total bilirubin, or serum lactate level, are measured in only a small minority of at-risk encounters. If measured, metabolic acidosis with anion gap may be detected, as respiratory muscle fatigue occurs in sepsis-associated respiratory failure. Other, less common findings include serum hypoalbuminemia, troponin elevation, hypoglycemia, and hypofibrinogenemia.

FIGURE 297-2

Distribution of SIRS and SOFA variables among infected patients at risk for sepsis, as documented in the electronic health record. Dark green bars represent the proportion of such patients with abnormal findings; light green bars, the proportion with normal findings; and white bars, the proportion with missing data. *(Adapted from CW Seymour et al: Assessment of clinical criteria for sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock [Sepsis-3]. JAMA 315:762, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Diagnostic Criteria

There is no specific test for sepsis, nor is there a gold-standard method for determining whether a patient is septic. In fact, the definition of sepsis can be written as a logic statement:

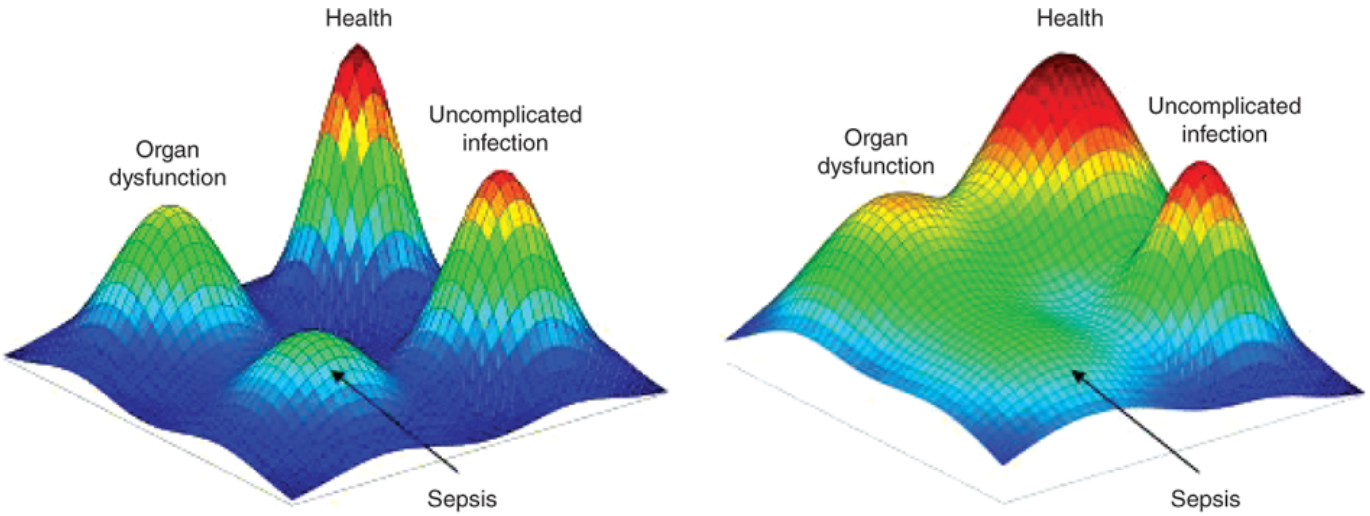
sepsis = f (*threat to life* | *organ dysfunction* | *dysregulated host response* | *infection*),

where sepsis is the dependent variable, which in turn is a function of four independent variables linked in a causal pathway, with—from left to right—one conditional upon the other. There may be uncertainty about whether each variable exists, whether it can be measured, and whether the causal and conditional relationships hold. If we assume that organ dysfunction exists and can be measured, then attributing the marginal degradation in function to a dysregulated host response is not simple and requires the ability to determine preexisting dysfunction, other noninfectious contributions to organ dysfunction, and—ideally—the mechanism by which the host response to an infection causes organ dysfunction.

In order to sort through these complex details, clinicians need simple bedside criteria to operationalize the logic statement **(Fig. 297-3)**. With this mandate, the Sepsis Definitions Task Force recommended that, once infection is suspected, clinicians consider whether it has caused organ dysfunction by determining a SOFA score. The SOFA score ranges from 0 to 24 points, with up to 4 points accrued across six organ systems. The SOFA score is widely studied in the ICU among patients with infection, sepsis, and shock. With ≥2 new SOFA points, the infected patient is considered septic and may be at ≥10% risk of in-hospital death.

Figure 297-3

Schematic of the importance of accurate, easy-to-use criteria for sepsis and its components, infection and organ dysfunction. In the ideal case (*left*), criteria clearly distinguish sepsis patients from other patients with uncomplicated infection or organ dysfunction. The reality (*right*), however, is that existing criteria fail to make clear distinctions, leaving a significant proportion of patients in areas of uncertainty. *(Adapted from DC Angus et al: A framework for the development and interpretation of different sepsis definitions and clinical criteria. Crit Care Med 44:e113, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Because the SOFA score requires multiple laboratory tests and may be costly to measure repeatedly, the quick SOFA (qSOFA) score was proposed as a clinical prompt to identify patients at high risk of sepsis outside the ICU, whether on the medical ward or in the emergency department. The qSOFA score ranges from 0 to 3 points, with 1 point each for systolic hypotension (≤100 mmHg), tachypnea (≥22 breaths/min), or altered mentation. A qSOFA score of ≥2 points has a predictive value for sepsis similar to that of more complicated measures of organ dysfunction. The qSOFA score is undergoing broader evaluation in other cohorts, in low-and middle-income settings, and in algorithms linked to clinical decision-making. Recent work has also shown that, although SIRS criteria may be fulfilled in sepsis, they sometimes are not and do not meaningfully contribute to the identification of patients with suspected infection who are at greater risk of a poor course, ICU admission, or death—outcomes more common among patients with sepsis than among those without.

As stated above, recent definitions have specified that septic shock is a subset of sepsis in which circulatory and cellular/metabolic abnormalities are profound enough to substantially increase mortality risk, but the application of this definition as a criterion for enrollment of patients varies significantly in clinical trials, observational studies, and quality improvement work. For clarity, criteria are proposed for septic shock that include (1) sepsis plus (2) the need for vasopressor therapy to elevate mean arterial pressure to ≥65 mmHg, with (3) a serum lactate concentration >2.0 mmol/L after adequate fluid resuscitation.

The new definitions and diagnostic criteria were externally validated in >1 million encounters stored in electronic health records. Nevertheless, given the uncertainty around the diagnosis of sepsis, Sepsis-3 is undergoing both validation in prospective studies and incorporation into clinical practice and quality improvement initiatives.

Arterial lactate is a long-studied marker of tissue hypoperfusion, and hyperlactemia and delayed lactate clearance are associated with a greater incidence of organ failure and death in sepsis. In a study of >1200 patients with suspected infection, 262 (24%) of 1081 patients exhibited an elevated lactate concentration (≥2.5 mmol/L) even in the setting of normal systolic blood pressure (>90 mmHg) and were at elevated risk of 28-day in-hospital mortality. However, lactic acidosis may occur in the presence of alcohol intoxication, liver disease, diabetes mellitus, administration of total parenteral nutrition, or antiretroviral treatment, among other conditions. Furthermore, in sepsis, an elevated lactate concentration may simply be the manifestation of impaired clearance. These factors may confound the use of lactate as a stand-alone biomarker for the diagnosis of sepsis; thus it should be used in the context of other markers of infection and organ dysfunction.

TREATMENT

TREATMENT Sepsis and Septic Shock Early Treatment of Sepsis and Septic Shock

Recommendations for sepsis care begin with prompt diagnosis. Recognition of septic shock by a clinician constitutes an emergency in which immediate treatment can be life-saving. Up-to-date guidelines for treatment are derived from international clinical practice guidelines provided by the Surviving Sepsis Campaign. This consortium of critical care, infectious disease, and emergency medicine professional societies has issued three iterations of clinical guidelines for the management of patients with sepsis and septic shock **(Table 297-2)**.

The initial management of infection requires several steps: forming a probable diagnosis, obtaining samples for culture, initiating empirical antimicrobial therapy, and achieving source control. More than 30% of patients with severe sepsis require source control, mainly for abdominal, urinary, and soft-tissue infections. The mortality rate is lower among patients with source control than among those without, although the timing of intervention is debated. For empirical antibiotic therapy (**Table 297-3**), the appropriate choice depends on the suspected site of infection, the location of infection onset (i.e., the community, a nursing home, or a hospital), the patient's medical history, and local microbial susceptibility patterns. In a single-center study of >2000 patients with bacteremia, the number of patients who needed to receive appropriate antimicrobial therapy in order to prevent one patient death was 4.0 (95% CI, 3.7–4.3).

Antibiotic delays may be deadly. For every 1-h delay among patients with sepsis, a 3–7% increase in the odds of in-hospital death is reported. Although meta-analyses report conflicting results, international clinical practice guidelines recommend the administration of appropriate broad-spectrum antibiotics within 1 h of recognition of severe sepsis or septic shock. Empirical antifungal therapy should be administered only to septic patients at high risk for invasive candidiasis.

The treatment elements listed above form the basis for two "bundles" of care: an initial management bundle to be completed within 3 h of presentation and a management bundle to be completed within 6 h. The initial management bundle includes (1) early administration of appropriate broad-spectrum antibiotics, (2) collection of blood for culture before antibiotic administration, and (3) measurement of serum lactate levels. The management bundle includes (1) an intravenous fluid bolus, (2) treatment with vasopressors for persistent hypotension or shock, and (3) re-measurement of serum lactate levels. Implementation of these two bundles has been associated with improved outcome in large multinational studies.

Other elements of the initial management bundle are cardiorespiratory resuscitation and mitigation of the immediate threats of uncontrolled infection. Early resuscitation requires a structured approach including the administration of IV fluids and vasopressors, with oxygen therapy and mechanical ventilation to support injured organs. The exact components required to optimize resuscitation, such as choice and amount of fluid, appropriate type and intensity of hemodynamic monitoring, and role of adjunctive vasoactive agents, all remain controversial, even after the completion and reporting of recent large randomized trials.

Evidence from an older study suggests that protocol-based, early goal-directed therapy (EGDT) may confer a greater survival advantage than clinical assessments of organ perfusion and management without a protocol. EGDT included an aggressive resuscitation protocol with specific hemodynamic thresholds for fluid administration, blood transfusion, and use of ionotropes. Given the many controversial features of this older single-center trial, the recent ProCESS trial compared protocol-based standard care with protocol-based EGDT and usual care in >31 emergency departments in the United States. Among 1341 patients, the 60-day in-hospital mortality rate for protocol-based standard care (18.2%) was similar to that for usual care (18.9%) and protocol-based EGDT (21%). The ARISE trial confirmed this finding, showing that, among 1600 patients with early septic shock at 51 centers in Australia and New Zealand, 90-day mortality was similar for EGDT and usual care. Finally, the ProMISe trial, which enrolled 1260 patients in 56 hospitals in England, found that EGDT offered no mortality benefit in early septic shock but did increase treatment intensity and cost. Multiple subsequent meta-analyses of the ProCESS, ARISE, and ProMISe trials confirmed that EGDT offers no mortality benefit while increasing health care utilization and ICU admission in well-resourced countries. Modified versions of EGDT were also tested in lower-resourced settings, with no change in outcome. Thus EGDT is no longer recommended as the primary strategy for early resuscitation in septic shock. Nonetheless, some form of resuscitation is considered essential, and a standardized approach, akin to the use of "trauma teams," has been advocated to ensure prompt care. The patient should be moved to an appropriate setting, such as the ICU, for ongoing care.

SUBSEQUENT TREATMENT OF SEPSIS AND SEPTIC SHOCK

After initial resuscitation, attention is focused on monitoring and support of organ function, avoidance of complications, and de-escalation of care when possible.

Monitoring

Hemodynamic monitoring devices may clarify the primary physiologic manifestations in sepsis and septic shock. The clinical usefulness of these monitoring devices can be attributable to the device itself, the algorithm linked to the device, or the static/dynamic target of the algorithm. Decades ago, the standard care of shock patients included invasive devices like the pulmonary artery catheter (PAC), also known as the continuous $ScvO_2$ catheter. The PAC can estimate cardiac output and measure mixed venous oxygen saturation, among other parameters, to refine the etiology of shock and potentially influence patient outcomes. Recently, a Cochrane review of 2923 general-ICU patients (among whom the proportion of patients in shock was not reported) found no difference in mortality with or without PAC management, and the PAC therefore is no longer recommended for routine use. Instead, a variety of noninvasive monitoring tools, such as arterial pulse contour analysis (PCA) or focused echocardiography, can provide continuous estimates of parameters such as cardiac output, beat-to-beat stroke volume, and pulse pressure variation. These tools, along with passive leg-raise maneuvers or inferior vena cava collapsibility on ultrasound, can help determine a patient's volume responsiveness but require that a variety of clinical conditions be met (e.g., patient on mechanical ventilation, sinus rhythm); in addition, more evidence from larger randomized trials on the impact of these tools in daily management is needed.

Support Of Organ Function

The primary goal of organ support is to improve delivery of oxygen to the tissues as quickly as possible. Depending on the underlying physiologic disturbance, this step may require administration of IV fluids or vasopressors, blood transfusions, or ventilatory support.

Many crystalloids can be used in septic shock, including 0.9% normal saline, Ringer's lactate, Hartmann's solution, and Plasma-Lyte. Because crystalloid solutions vary in tonicity and inorganic/organic anions, few of these preparations closely resemble plasma. Normal saline is widely used in the United States. Colloid solutions (e.g., albumin, dextran, gelatins, or hydroxyethyl starch) are the most widely used fluids in critically ill patients, with variability across ICUs and countries. A clinician's choice among colloids is influenced by availability, cost, and the desire to minimize interstitial edema. Many think that a greater intravascular volume is gained by use of colloids in shock, but the effects of colloids are modified by molecular weight and concentration as well as by vascular endothelial changes during inflammation. A network meta-analysis using direct and indirect comparisons in sepsis found evidence of higher mortality with starch than with crystalloids (relative risk [RR], 1.13; 95% CI, 0.99–1.30 [high confidence]) and no difference between albumin (RR, 0.83; 95% CI, 0.65, 1.04 [moderate confidence]) or gelatin (RR, 1.24; 95%CI, 0.61, 2.55 [very low confidence]) and crystalloids. In general, crystalloids are recommended on the basis of strong evidence as first-line fluids for sepsis resuscitation, with specific caveats; their use is guided by resolution of hypotension, oliguria, altered mentation, and hyperlactemia. Only weak evidence supports the use of balanced crystalloids, and guidelines recommend against using hydroxyethyl starches for intravascular volume replacement.

When circulating fluid volume is adequate, vasopressors are recommended to maintain perfusion of vital organs. Vasopressors such as norepinephrine, epinephrine, dopamine, and phenylephrine differ in terms of half-life, β- and α-adrenergic stimulation, and dosing regimens. Recent evidence comes from the SOAP II trial, a double-blind randomized clinical trial at eight centers comparing norepinephrine with dopamine in 1679 undifferentiated ICU patients with shock, of whom 63% were septic. Although no difference was observed in 28-day mortality or in predefined septic-shock subgroup, arrhythmias were significantly greater with dopamine. These findings were confirmed in a subsequent meta-analysis. As a result, expert opinion and consensus guidelines recommend norepinephrine as the first-choice vasopressor in septic shock. Levels of the endogenous hormone vasopressin may be low in septic shock, and the administration of vasopressin can reduce the norepinephrine dose. Consensus guidelines suggest adding vasopressin (up to 0.03 U/min) in patients without a contraindication to norepinephrine, with the intent of raising mean arterial pressure or decreasing the norepinephrine dose. There may be select indications for use of alternative vasopressors—e.g., when tachyarrhythmias from dopamine or norepinephrine, limb ischemia from vasopressin, or other adverse effects dictate.

The transfusion of red blood cells to high thresholds (>10 g/dL) had been suggested as part of EGDT in septic shock. However, the recent Scandinavian TRISS trial in 1005 septic shock patients demonstrated that a lower threshold (7 g/dL) resulted in 90-day mortality rates similar to those with a higher threshold (9 g/dL) and reduced transfusions by almost 50%.

Significant hypoxemia (PaO_2, <60 mmHg; or SaO_2, <90%), hypoventilation (rising $PaCO_2$), increased work of breathing, and inadequate or unsustainable compensation for metabolic acidosis (pH <7.20) are common indications for mechanical ventilatory support. Endotracheal intubation protects the airway, and positive-pressure breathing allows oxygen delivery to metabolically active organs in favor of inspiratory muscles of breathing and the diaphragm. An experiment in dogs showed that the relative proportion of cardiac output delivered to respiratory muscles in endotoxic shock decreased by fourfold with spontaneous ventilation over that with mechanical ventilation. During intubation, patients in shock should be closely monitored for vasodilatory effects of sedating medications or compromised cardiac output due to increased intrathoracic pressure, both of which may cause hemodynamic collapse. With hemodynamic instability, noninvasive mask ventilation may be less suitable in patients experiencing sepsis-associated acute respiratory failure.

Adjuncts

One of the great disappointments in sepsis management over the past 30 years has been the failure to convert advances in our understanding of the underlying biology into new therapies. Researchers have tested both highly specific agents and those with more pleotropic effects. The specific agents can be divided into those designed to interrupt the initial cytokine cascade (e.g., anti-LPS or anti-proinflammatory cytokine strategies) and those that interfere with dysregulated coagulation (e.g., antithrombin or activated protein C). Recombinant activated protein C (aPC) was one of the first agents approved by the U.S. Food and Drug Administration and was the most widely used. A large, randomized, double-blind, placebo-controlled, multicenter trial of aPC in severe sepsis (the PROWESS trial) was reported in 2001; the data suggested an absolute risk reduction of up to 6% among aPC-treated patients with severe sepsis. However, subsequent phase 3 trials failed to confirm this effect, and the drug was withdrawn from the market. It is no longer recommended in the care of sepsis or septic shock.

Many adjunctive treatments in sepsis and septic shock target changes in the innate immune response and coagulation cascade. Specific adjuncts like glucocorticoids in septic shock have continued to be widely used. A large negative clinical trial and a conflicting systematic review in 2009 extended the debate about whether glucocorticoids lower 28-day mortality or improve shock reversal. Most meta-analyses report no change in mortality but an increase in shock reversal with glucocorticoid treatment. The recent HYPRESS trial found no difference between patients with severe sepsis who were treated with glucocorticoids and control patients in terms of the development of shock or the mortality rate. These data and others led to a suggestion in international clinical practice guidelines against using IV hydrocortisone to treat septic shock if adequate fluid resuscitation and vasopressor therapy are able to restore hemodynamic stability. If not, the guidelines suggest the administration of IV hydrocortisone at a dose of 200 mg per day (weak recommendation, low quality of evidence).

Among other adjuncts, IV immunoglobulin may be associated with potential benefit, but significant questions remain and such treatment is not part of routine practice. Despite a large number of observational studies suggesting that statin use mitigates the incidence or outcome of sepsis and severe infection, there are no confirmatory randomized controlled trials, and statins are not an element in routine sepsis care.

De-Escalation of Care

Once patients with sepsis and septic shock are stabilized, it is important to consider which therapies are no longer required and how care can be minimized. The de-escalation of initial broad-spectrum therapy, which observational evidence indicates is safe, may reduce the emergence of resistant organisms as well as potential drug toxicity and costs. The added value of combination antimicrobial therapy over that of adequate single-agent antibiotic therapy in severe sepsis has not been established. Current guidelines recommend combination antimicrobial therapy only for neutropenic sepsis and sepsis caused by *Pseudomonas*. Large trials are under way in the United States to determine how serum biomarkers like procalcitonin can assist clinicians in minimizing antibiotic exposure, while European trials are indicating that this biomarker may lead to a reduction in the duration of treatment and in daily defined doses in critically ill patients with a presumed bacterial infection.

TABLE 297-2 Elements of Care in Sepsis and Septic Shock: Recommendations Adapted from International Consensus Guidelines

Resuscitation
Sepsis and septic shock constitute an emergency, and treatment should begin right away.
Resuscitation with IV crystalloid fluid (30 mL/kg) should begin within the first 3 h.
Saline or balanced crystalloids are suggested for resuscitation.
If the clinical examination does not clearly identify the diagnosis, hemodynamic assessments (e.g., with focused cardiac ultrasound) can be considered.
In patients with elevated serum lactate levels, resuscitation should be guided towards normalizing these levels when possible.
In patients with septic shock requiring vasopressors, the recommended target mean arterial pressure is 65 mmHg.
Hydroxyethyl starches and gelatins are not recommended.
Norepinephrine is recommended as the first-choice vasopressor.
Vasopressin should be used with the intent of reducing the norepinephrine dose.
The use of dopamine should be avoided except in specific situations—e.g., in those patients at highest risk of tachyarrhythmias or relative bradycardia.
Dobutamine use is suggested when patients show persistent evidence of hypoperfusion despite adequate fluid loading and use of vasopressors.
Red blood cell transfusion is recommended only when the hemoglobin concentration decreases to <7.0 g/dL in the absence of acute myocardial infarction, severe hypoxemia, or acute hemorrhage.
Infection Control
So long as no substantial delay is incurred, appropriate samples for microbiologic cultures should be obtained before antimicrobial therapy is started.
IV antibiotics should be initiated as soon as possible (within 1 h); specifically, empirical broad-spectrum therapy should be used to cover all likely pathogens.
Antibiotic therapy should be narrowed once pathogens are identified and their sensitivities determined and/or once clinical improvement is evident.
If needed, source control should be undertaken as soon as is medically and logistically possible.
Daily assessment for de-esclation of antimicrobial therapy should be conducted.
Respiratory Support
A target tidal volume of 6 mL/kg of predicted body weight (compared with 12 mL/kg in adult patients) is recommended in sepsis-induced ARDS.
A higher PEEP rather than a lower PEEP is used in moderate to severe sepsis-induced ARDS.
In severe ARDS (PaO_2/FIO_2, <150 mmHg), prone positioning is recommended, and recruitment maneuvers and/or neuromuscular blocking agents for ≤48 h are suggested.
A conservative fluid strategy should be used in sepsis-induced ARDS if there is no evidence of tissue hypoperfusion.
Routine use of a pulmonary artery catheter is not recommended.
Spontaneous breathing trials should be used in mechanically ventilated patients who are ready for weaning.
General Supportive Care
Patients requiring a vasopressor should have an arterial catheter placed as soon as is practical.
Hydrocortisone is not suggested in septic shock if adequate fluids and vasopressor therapy can restore hemodynamic stability.
Continuous or intermittent sedation should be minimized in mechanically ventilated sepsis patients, with titration targets used whenever possible.
A protocol-based approach to blood glucose management should be used in ICU patients with sepsis, with insulin dosing initiated when two consecutive blood glucose levels are >180 mg/dL.
Continuous or intermittent renal replacement therapy should be used in patients with sepsis and acute kidney injury.
Pharmacologic prophylaxis (unfractionated heparin or low-molecular-weight heparin) against venous thromboembolism should be used in the absence of contraindications.
Stress ulcer prophylaxis should be given to patients with risk factors for gastrointestinal bleeding.
The goals of care and prognosis should be discussed with patients and their families.

Abbreviations: ARDS, acute respiratory distress syndrome; ICU, intensive care unit; PEEP, positive end-expiratory pressure.

Source: Adapted from A Rhodes et al: Surviving Sepsis Campaign: International guidelines for management of sepsis and septic shock: 2016. Crit Care Med 45:486, 2017.

TABLE 297-3 Initial Antimicrobial Therapy for Severe Sepsis with No Obvious Source in Adults with Normal Renal Function

CLINICAL CONDITION	ANTIMICROBIAL REGIMENS[a]
Septic shock (immunocompetent adult)	The many acceptable regimens include (1) piperacillin-tazobactam (3.375–4.5 g q6h), (2) cefepime (2 g q12h), or (3) meropenem (1 g q8h) or imipenem-cilastatin (0.5 g q6h). If the patient is allergic to β-lactam antibiotics, use (1) aztreonam (2 g q8h) or (2) ciprofloxacin (400 mg q12h) or levofloxacin (750 mg q24h). Add vancomycin (loading dose of 25–30 mg/kg, then 15–20 mg/kg q8–12h) to each of the above regimens.
Neutropenia (<500 neutrophils/μL)	Regimens include (1) cefepime (2 g q8h), (2) meropenem (1 g q8h) or imipenem-cilastatin (0.5 g q6h) or doripenem (500 mg q8h), or (3) piperacillin-tazobactam (3.375 g q4h). Add vancomycin (as above) if the patient has a suspected central line–associated bloodstream infection, severe mucositis, skin/soft tissue infection, or hypotension. Add tobramycin (5–7 mg/kg q24h) plus vancomycin (as above) plus caspofungin (one dose of 70 mg, then 50 mg q24h) if the patient has severe sepsis/septic shock.
Splenectomy	Use ceftriaxone (2 g q24h, or—in meningitis—2 g q12h). If the local prevalence of cephalosporin-resistant pneumococci is high, add vancomycin (as above). If the patient is allergic to β-lactam antibiotics, use levofloxacin (750 mg q24h) or moxifloxacin (400 mg q24h) plus vancomycin (as above).

[a]All agents are administered by the intravenous route.

Source: Adapted in part from DN Gilbert et al: *The Sanford Guide to Antimicrobial Therapy*, 47th ed, 2017; and from RS Munford: Sepsis and septic shock, in DL Kasper et al (eds). *Harrison's Principles of Internal Medicine*, 19th ed. New York, McGraw-Hill, 2015, p 1757.

PROGNOSIS

Before modern intensive care, sepsis and septic shock were highly lethal, with infection leading to compromise of vital organs. Even with intensive care, nosocomial mortality rates for septic shock often exceeded 80% as recently as 30 years ago. Now, the U.S. Burden of Disease Collaborators report that the primary risk factor for sepsis and septic shock—i.e., infection—is the fifth leading cause of years of productive life lost because of premature death. More than half of sepsis cases require ICU admission, representing 10% of all ICU admissions. However, with advances in training, surveillance, monitoring, and prompt initiation of supportive care for organ dysfunction, the mortality rate from sepsis and septic shock is now closer to 20% in many series. Although some data suggest that mortality trends are even lower, attention has been focused on the trajectory of recovery among survivors. Patients who survive to hospital discharge after sepsis remain at increased risk of death in the following months and years. Those who survive often suffer from impaired physical or neurocognitive dysfunction, mood disorders, and low quality of life. In many studies, it is difficult to determine the causal role of sepsis. However, an analysis of the Health and Retirement Study—a large longitudinal cohort study of aging Americans—suggested that severe sepsis significantly accelerated physical and neurocognitive decline. Among survivors, the rate of hospital readmission within 90 days after sepsis exceeds 40%.

PREVENTION

In light of the persistently high mortality risk in sepsis and septic shock, prevention may be the best approach to reducing avoidable deaths, but preventing sepsis is a challenge. The aging of the population, the overuse of inappropriate antibiotics, the rising incidence of resistant microorganisms, and the use of indwelling devices and catheters contribute to a steady burden of sepsis cases. The number of cases could be reduced by avoiding unnecessary antibiotic use, limiting use of indwelling devices and catheters, minimizing immune suppression when it is not needed, and increasing adherence to infection control programs at hospitals and clinics. To facilitate earlier treatment, such pragmatic work could be complemented by research into the earliest pathophysiology of infection, even when symptoms of sepsis are nascent. In parallel, the field of implementation science could inform how best to increase adoption of infection control in high-risk settings and could guide appropriate care.

FURTHER READING

Angus DC et al: Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care. Crit Care Med 29:1303, 2001. [PubMed: 11445675]
Boomer JS et al: Immunosuppression in patients who die of sepsis and multiple organ failure. JAMA 306:2594, 2011. [PubMed: 22187279]
De Backer D et al: Comparison of dopamine and norepinephrine in the treatment of shock. N Engl J Med 362:779, 2010. [PubMed: 20200382]
Fleischmann C et al: Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations. Am J Respir Crit Care Med 193:259, 2016. [PubMed: 26414292]
Medzhitov R et al: Disease tolerance as a defense strategy. Science 335:936, 2012. [PubMed: 22363001]
Rhodes A et al: Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016. Crit Care Med 45:486, 2017. [PubMed: 28098591]
Rochwerg B et al: Fluid resuscitation in sepsis: A systematic review and network meta-analysis. Ann Intern Med 161:347, 2014. [PubMed: 25047428]
Seymour CW et al: Assessment of clinical criteria for sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 315:762, 2016. [PubMed: 26903335]
Vincent JL et al: The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med 22:707, 1996. [PubMed: 8844239]
Yealy DM et al: A randomized trial of protocol-based care for early septic shock. N Engl J Med 370:1683, 2014. [PubMed: 24635773]

Chapter 298: Cardiogenic Shock and Pulmonary Edema

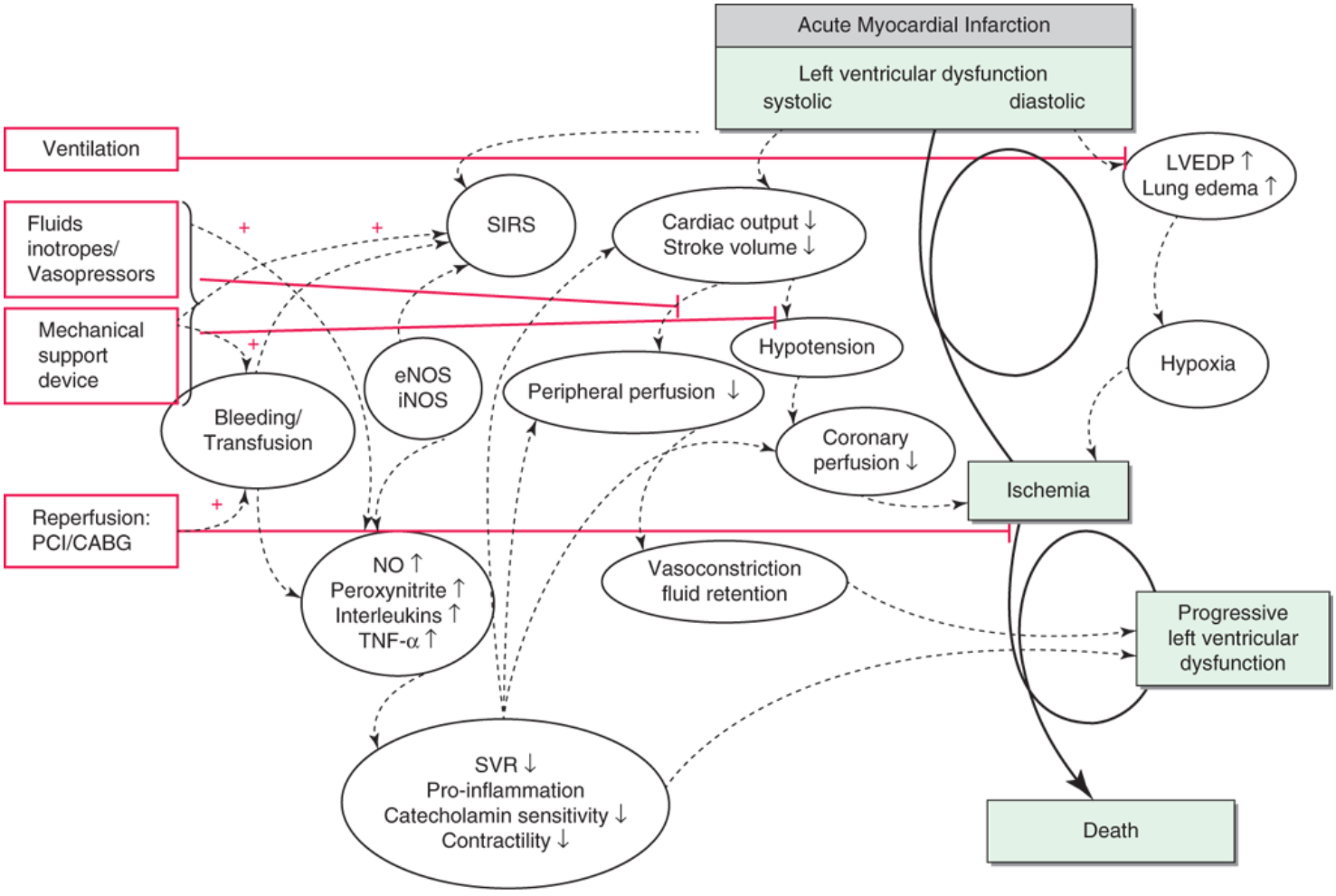
David H. Ingbar; Holger Thiele

INTRODUCTION

Cardiogenic shock (CS) and pulmonary edema are life-threatening high acuity conditions that require treatment as medical emergencies, usually in an intensive care unit (ICU) or cardiac intensive care unit (CICU). The most common joint etiology is severe left ventricular (LV) dysfunction from myocardial infarction (MI) that leads to pulmonary congestion and/or systemic hypoperfusion **(Fig. 298-1)**. **The pathophysiology of pulmonary edema and shock are discussed in Chaps. 33 and 296, respectively.**

FIGURE 298-1

Pathophysiology of cardiogenic shock and potential treatment targets. The pathophysiological concept of the expanded cardiogenic shock spiral and treatment targets. CABG, coronary artery bypass grafting; eNOS, endothelial nitric oxide synthase; iNOS, inducible nitric oxide synthase; LVEDP, left ventricular end diastolic pressure; NO, nitric oxide; PCI, percutaneous coronary intervention; SIRS, systemic inflammatory response syndrome; SVR, systemic vascular resistance; TNF, tumor necrosis factor. *(Reproduced and adapted, with permission, from H Thiele et al: Shock in acute myocardial infarction: The Cape Horn for trials? Eur Heart J 31: 1828, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CARDIOGENIC SHOCK

CS is a low cardiac output state resulting in life-threatening end-organ hypoperfusion and hypoxia. The clinical presentation is typically characterized by persistent hypotension (<90 mmHg systolic blood pressure [BP]) unresponsive to volume replacement and is accompanied by clinical features of peripheral hypoperfusion, such as elevated arterial lactate (>2 mmol/L). Objective hemodynamic parameters such as cardiac index or pulmonary capillary wedge pressure can help confirm the diagnosis, but are not mandatory. The in-hospital mortality rates range from 40 to 60%, depending on shock severity and the associated underlying cause. Acute MI with LV dysfunction remains the most frequent cause of CS with other causes listed in **Table 298-1**. Circulatory failure based on cardiac dysfunction may be caused by primary myocardial failure, most commonly secondary to acute MI **(Chap. 269)**, and less frequently by cardiomyopathy or myocarditis **(Chap. 254)**, cardiac tamponade **(Chap. 265)**, arrhythmias **(Chap. 249)**, or critical valvular heart disease **(Chap. 256)**.

TABLE 298-1 Etiologies of Cardiogenic Shock (CS)[a] and Cardiogenic Pulmonary Edema

Etiologies of Cardiogenic Shock or Pulmonary Edema
Acute myocardial infarction/ischemia
Left ventricular failure
Ventricular septal rupture
Papillary muscle/chordal rupture–severe mitral regurgitation
Ventricular free wall rupture
Other conditions complicating large myocardial infarctions
Excess negative inotropic or vasodilator medications
Post-cardiac arrest
Post-cardiotomy
Refractory sustained supra or ventricular tachyarrhythmias
Refractory sustained bradyarrhythmias
Acute fulminant myocarditis
End-stage cardiomyopathy
Takotsubo syndrome/Apical ballooning syndrome
Hypertrophic cardiomyopathy with severe outflow obstruction
Aortic dissection with aortic insufficiency or tamponade
Severe valvular heart disease
Critical aortic or mitral stenosis
Acute severe aortic regurgitation or mitral regurgitation
Toxic/metabolic
β-blocker or calcium channel antagonist overdose
Hypertensive crisis
Post-cardiac arrest stunning
Myocardial depression in setting of septic shock or SIRS
Myocardial contusion
Other Etiologies of Cardiogenic Shock[b]
Right ventricular failure due to:
Acute myocardial infarction
Acute or decompensated chronic cor pulmonale
Pericardial tamponade
Toxic/metabolic
Severe acidosis, severe hypoxemia

[a]The etiologies of CS are listed. Most of these can cause pulmonary edema instead of shock or pulmonary edema with CS. [b]These cause CS but not pulmonary edema.

Incidence

The incidence of CS complicating acute MI has decreased to 5–10%, largely due to increasing use of early mechanical reperfusion therapy for acute MI. Shock is more common with ST-elevation MI (STEMI) than with non-STEMI **(Chap. 269)**.

LV failure accounts for ~80% of cases of CS complicating acute MI. Acute severe mitral regurgitation (MR), ventricular septal rupture (VSR), predominant right ventricular (RV) failure, and free wall rupture or tamponade account for the remainder. A recently recognized uncommon cause of transient CS is the Takotsubo syndrome.

Pathophysiology

The understanding of the complex pathophysiology of CS has evolved over the past decades. In general, a profound depression of myocardial contractility results in a deleterious spiral of reduced cardiac output, low blood pressure, and ongoing myocardial ischemia, followed by further contractility reduction (Fig. 298-1). This vicious cycle usually leads to death if not interrupted. CS can result in both acute and subacute derangements to the entire circulatory system. Hypoperfusion of vital organs and extremities remains a clinical hallmark. Although ineffective stroke volume is the inciting event, inadequate circulatory compensation also may contribute to shock. Initial peripheral vasoconstriction may improve coronary and peripheral perfusion at the cost of increased afterload. However, over the course of CS systemic inflammation response triggered by acute cardiac injury often induces pathologic vasodilatation. Inflammatory cytokines, endothelial and inducible nitric oxide synthase may augment NO production, accompanied by peroxynitrite, which has a negative inotropic effect and is cardiotoxic. Lactic acidosis and

hypoxemia contribute to the vicious circle, as severe acidosis reduces the efficacy of endogenous and exogenous catecholamines. During ICU support bleeding and/or transfusions may trigger inflammation and are usually associated with higher mortality (Fig. 298-1).

Patient Profile

In patients with MI, older age, prior MI, diabetes mellitus, anterior MI location, and multivessel coronary artery disease with extensive coronary artery stenoses are associated with an increased risk of CS. Shock associated with a first inferior MI should prompt a search for a mechanical cause or RV involvement. CS may rarely occur in the absence of significant stenosis, as seen in Takotsubo syndrome or fulminant myocarditis.

Timing

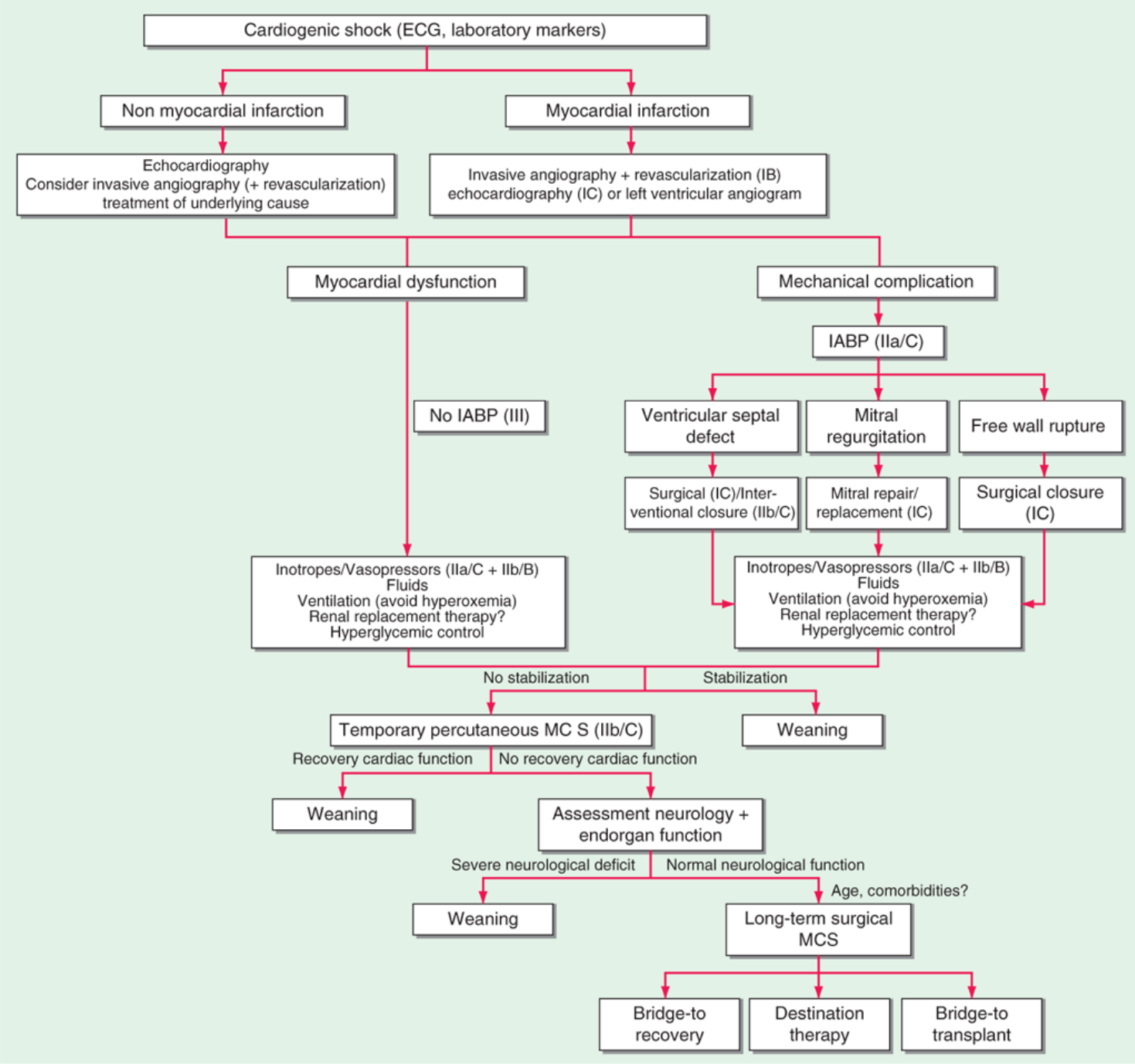
Shock is present on admission in approximately one-quarter of MI patients who develop CS; one-quarter develop it rapidly thereafter, within 6 h of MI onset, and another quarter develop shock later on the first day. Later onset of CS may be due to reinfarction, marked infarct expansion, or mechanical complications.

Diagnosis

For these unstable patients, supportive therapy must be initiated simultaneously with diagnostic evaluation **(Fig. 298-2)**. A focused history and physical examination should be performed along with an electrocardiogram (ECG), chest X-ray, arterial blood gas (ABG) analysis, lactate measurement, and blood specimens to the laboratory. Initial echocardiography is an invaluable tool to elucidate the underlying cause of CS.

FIGURE 298-2

Emergency management of patients with cardiogenic shock. Treatment algorithm for patients with CS. The class of recommendation and level of evidence according to European Society of Cardiology guidelines is provided (see Further Reading citations Authors/Task Force members, S Windecker et al: Eur Heart J 35:2541, 2014, and P Ponikowski et al: Eur Heart J 37:2129, 2016). ECG, electrocardiogram; IABP, intraaortic balloon pump; MCS, mechanical circulatory support.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CLINICAL FINDINGS

Most patients initially are dyspneic, appear pale, apprehensive, and diaphoretic, and mental status may be altered. The pulse is typically weak and rapid or occasionally severe bradycardia due to high-grade heart block may be present. BP is typically reduced (<90 mmHg; or catecholamines required to maintain blood pressure >90 mmHg), but occasionally BP may be maintained by very high systemic vascular resistance. Tachypnea and jugular venous distention may be present. Typically there is a weak apical pulse and soft S_1, and an S_3 gallop may be audible. Acute, severe MR and VSR usually are associated with characteristic systolic murmurs **(Chap. 269)**. Crackles are audible in most patients with LV failure. Oliguria/anuria is common. Often CS patients require early mechanical ventilation (~80%) for management of acute hypoxemia, increased work of breathing, and hemodynamic instability; catecholamines often are required to maintain adequate blood pressure.

LABORATORY FINDINGS

The white blood cell count and C-reactive protein typically are elevated. Renal function often is progressively impaired. Newer renal function markers such as Cystatin C or Neutrophil gelatinase-associated lipocalin (NGAL) do not add prognostic information over creatinine. Hepatic transaminases are elevated due to liver hypoperfusion in ~20% of patients and may be very high. The arterial lactate level is usually elevated to >2 mmol/L. ABGs usually demonstrate hypoxemia and anion gap metabolic acidosis. Glucose levels at admission are often elevated, a

strong independent predictor for mortality. Cardiac markers, creatine kinase and its MB fraction, and troponins I and T are typically markedly elevated in acute MI.

ELECTROCARDIOGRAM

In acute MI with CS, Q waves and/or ST elevation in multiple leads or left bundle branch block are usually present. Approximately one-half of MIs with CS are anterior infarctions. Global ischemia due to severe left main stenosis usually is accompanied by aVR lead ST segment elevation and ST depressions in multiple leads.

CHEST ROENTGENOGRAM

The chest x-ray typically shows pulmonary vascular congestion and often pulmonary edema, but may be normal in up to a third of patients. The heart size is usually normal when CS results from a first MI, but may be enlarged when it occurs in a patient with a previous MI.

ECHOCARDIOGRAM

An echocardiogram **(Chap. 236)** should be obtained promptly in patients with suspected/confirmed CS to help define its etiology. Echocardiography is able to delineate the extent of infarction/myocardium in jeopardy and the presence of mechanical complications such as VSR, MR, or cardiac tamponade. Furthermore, valvular obstruction or insufficiency, dynamic LV outflow tract obstruction, proximal aortic dissection with aortic regurgitation or tamponade may be seen, or indirect evidence for pulmonary embolism may be obtained **(Chap. 273)** **(Table 298-2)**.

TABLE 298-2 Utility of the Echocardiogram in Cardiogenic Shock or Pulmonary Edema

Clinical Question	Information
Ventricular Function	Predominantly left, right or biventricular involvement
Etiology	**Acute Myocardial Infarction** • Extent of infarction/myocardium in jeopardy • Status of the non-culprit infarct zone • Presence of mechanical complications **Acute/Chronic Valvular Insufficiency/Obstruction/Stenosis (Native/Prosthetic)** • Etiology: endocarditis; degenerative valve disease • Location and hemodynamic consequences **Dynamic Left Ventricular Tract Obstruction** **Takotsubo Syndrome** **Cardiac Tamponade** • Circumferential versus localized effusion • Route of pericardiocentesis if indicated **Acute Pulmonary Embolism** • Right ventricular function • Pulmonary artery pressure • Presence of clot in transition/Patent foramen ovale **Acute Aortic Syndrome** • Nature and extent of dissection • Degree of aortic insufficiency • Presence of pericardial effusion
Hemodynamics	Volume assessment by inferior vena cava diameter and inspiratory collapse Estimated pulmonary artery systolic pressure Estimated left atrial pressure
Therapeutic guidance	Guide vasoactive support Monitor response to therapy Mechanical circulatory support decisions Catheter position and guidance
Pulmonary	Pleural effusion Lung edema Pneumothorax Pulmonary infiltration

PULMONARY ARTERY CATHETERIZATION

The use of pulmonary artery catheter (PAC) hemodynamic monitoring is declining because clinical trials have shown no mortality benefit. However, hemodynamic data provided by a PAC can confirm the presence and severity of CS, involvement of the right ventricle, left-to-right shunting, pulmonary artery pressures and trans-pulmonary gradient, and the pulmonary and systemic vascular resistance. It can help in recognition of acute MR, decreased left atrial filling pressure, and secondary occult sepsis and to exclude left-to-right shunts. Equalization of diastolic pressures suggests cardiac tamponade, but echocardiogram is more definitive. The detailed hemodynamic profile can be used to individualize and monitor therapy and to provide prognostic information, such as cardiac index and cardiac power, can be obtained. The use of a PAC is currently recommended by the American Heart Association for potential utilization in cases of diagnostic or CS management uncertainty or in patients with severe CS who are unresponsive to initial therapy.

ADVANCED HEMODYNAMIC MONITORING

Recently new central venous catheter systems linked to computer-based algorithms provide continuous monitoring of a variety of derived hemodynamic parameters, including cardiac output, stroke volume, stroke volume variation, and systemic vascular resistance. When combined with a femoral arterial catheter, calculated extravascular lung water and pulmonary permeability index can be monitored. The information allows for more rational therapy and assessment, but has not yet shown improved clinical outcomes in patients with shock or pulmonary edema **(Table 298-3)**.

TABLE 298-3 Hemodynamic Patterns[a]

	RA, mmHg	RVS, mmHg	RVD, mmHg	PAS, mmHg	PAD, mmHg	PCW, mmHg	CI, (L/min)/m^2	SVR, (dyn · s)/cm^5
Normal values	<6	<25	0–12	<25	0–12	<6–12	≥2.5	(800–1600)
MI without pulmonary edema[b]	–	–	–	–	–	~13 (5–18)	~2.7 (2.2–4.3)	–
Pulmonary edema	↔↑	↔↑	↔↑	↑	↑	↑	↔↓	↑
Cardiogenic shock								
LV failure	↔↑	↔↑	↔↑	↔↑	↑	↑	↓	↔↑
RV failure[c]	↑	↓↔↑[d]	↑	↓↔↑[d]	↔↓↑[d]	↓↔↑[d]	↓	↑
Cardiac tamponade	↑	↔↑	↑	↔↑	↔↑	↔↑	↓	↑
Acute mitral regurgitation	↔↑	↑	↔↑	↑	↑	↑	↔↓	↔↑
Ventricular septal rupture	↑	↔↑	↑	↔↑	↔↑	↔↑	↑PBF ↓SBF	↔↑
Hypovolemic shock	↓	↔↓	↔↓	↓	↓	↓	↓	↑
Septic shock	↓	↔↓	↔↓	↓	↓	↓	↑	↓

[a]There is significant patient-to-patient variation. Pressure may be normalized if cardiac output is low. [b]Forrester et al classified nonreperfused MI patients into four hemodynamic subsets. (From JS Forrester et al: N Engl J Med 295:1356, 1976.) PCW pressure and CI in clinically stable subset 1 patients are shown. Values in parentheses represent range. [c]"Isolated" or predominant RV failure. [d]PCW and pulmonary artery pressures may rise in RV failure after volume loading due to RV dilation and right-to-left shift of the interventricular septum, resulting in impaired LV filling. When biventricular failure is present, the patterns are similar to those shown for LV failure.

Abbreviations: CI, cardiac index; MI, myocardial infarction; P/SBF, pulmonary/systemic blood flow; PAS/D, pulmonary artery systolic/diastolic; PCW, pulmonary capillary wedge; RA, right atrium; RVS/D, right ventricular systolic/diastolic; SVR, systemic vascular resistance.

Source: Table prepared with the assistance of Krishnan Ramanathan, MD.

CARDIAC CATHETERIZATION AND CORONARY ANGIOGRAPHY

The definition of the coronary anatomy provides useful information and is immediately indicated in all patients with CS complicating MI for further reperfusion treatment. Furthermore, cardiac catheterization should also be considered for resuscitated cardiac arrest survivors without ST-segment elevation because ~70% of these patients have relevant coronary artery disease.

TREATMENT

TREATMENT Acute Myocardial Infarction GENERAL MEASURES

In addition to the usual treatment of acute MI **(Chap. 269)**, initial therapy is aimed at maintaining adequate systemic and coronary perfusion by raising the blood pressure with vasopressors and adjusting volume status to a level that ensures optimum LV filling pressure (Fig. 298-2). There is some interpatient variability, but generally adequate perfusion occurs with a mean arterial BP of 60–65 mmHg or a systolic BP ~90 mmHg. Hypoxemia and acidosis need to be corrected; up to 90% of patients require ventilatory support, decreasing the stress from increased work of breathing (see "Pulmonary Edema," below) (Fig. 298-2). Moderate glucose control (≤180 mg/dL or 10.0 mmol/L) should be a goal and hypoglycemia must be avoided. Negative ionotropic agents should be discontinued. Bradyarrhythmias may require transvenous pacing. Recurrent ventricular tachycardia or rapid atrial fibrillation may require immediate treatment **(Chap. 241)**.

REPERFUSION-REVASCULARIZATION

Rapid revascularization of the infarct-related artery is the only evidence-based treatment strategy for mortality reduction in CS and forms the mainstay therapeutic intervention for CS due to MI (Fig. 298-2). In the SHOCK Trial 132 lives were saved per 1000 patients treated with early revascularization with percutaneous coronary intervention (PCI) or coronary artery bypass graft (CABG) compared with initial medical therapy. Outcome benefit correlates strongly with the time between symptom onset and reperfusion. In general, PCI with drug-eluting stents of the infarct-related artery is the preferred reperfusion strategy. Approximately 80% of CS patients present with multivessel coronary artery disease. The recent CULPRIT-SHOCK randomized trial showed that culprit lesion only PCI with possible staged revascularization led to a reduction in 30-day mortality or renal replacement therapy in comparison to immediate multivessel PCI. This reduction in the primary study endpoint was mainly driven by a 30-day mortality reduction. Currently, vascular access for diagnostic angiography and PIC via the radial artery is preferred when feasible over femoral arterial access due to its greater safety. CABG is currently performed in only 5% of cases mainly if coronary anatomy is not amenable to PCI.

VASOPRESSORS AND INOTROPES

Inotropic agents are theoretically appealing in CS treatment. However, current evidence is scarce. Vasoactive medications are often used in the management of patients with CS and all have important disadvantages, including increase in myocardial O_2 consumption, afterload, lethal arrhythmias, and possible myocardial cell death. As a consequence, catecholamines should be used in the lowest possible doses for the shortest possible time. Despite their frequent use, little clinical outcome data proves their benefit or is available to guide the initial selection of vasoactive therapies in patients with CS. No vasopressor has been demonstrated to change outcome in large clinical trials. Norepinephrine is

reasonable as the first line vasopressor based on randomized trials compared to dopamine. Norepinephrine was associated with fewer adverse events, including arrhythmias, compared to dopamine in a randomized trial of patients with several etiologies of circulatory shock and with improved survival in a pre-specified subgroup of CS patients. Norepinephrine dosing is usually begun at 2 to 4 μg/min and titrated upward based on blood pressure. *Dopamine*'s hemodynamic effects vary depending upon dose and there is interpatient variability in responses. Low doses stimulate renal dopaminergic receptors and with increasing dosage there is stimulation of first β-adrenergic receptors and then α adrenergic receptors. Dopamine should be avoided as first-line therapy for MI with CS based on hemodynamic and proarrhythmogenic effects.

Dobutamine is a synthetic sympathomimetic amine with positive inotropic action and minimal positive chronotropic activity at low doses (2.5 μg/kg per min), but moderate chronotropic activity at higher doses. Its vasodilating activity often precludes its use when a vasoconstrictor effect is required. Levosimendan may also be appealing despite a lack of randomized data, but was not beneficial for organ dysfunction in sepsis.

MECHANICAL CIRCULATORY SUPPORT

The most commonly used mechanical circulatory support (MCS) device has been the intraaortic balloon pump (IABP), which is inserted into the aorta via the femoral artery and provides passive hemodynamic support. However, routine IABP use in conjunction with early revascularization (predominantly with PCI) did not reduce either 30-day or 12-month mortality in the IABP-SHOCK II trial. IABP also had no benefit on secondary endpoints (arterial lactate, catecholamine doses, renal function, or intensive care severity of illness unit scores). IABP is no longer recommended for CS with LV failure.

Active MCS devices to support the left, right, or both ventricles can be placed percutaneously or surgically. Temporary percutaneous MCS can be used as bridge to recovery, to surgically implanted devices, to heart transplantation, or as a temporizing measure when the neurologic status is uncertain. Percutaneous MCS including the TandemHeart, Impella devices, and also venoarterial extracorporeal membrane oxygenation (VA-ECMO) have been used in patients not responding to standard treatment (catecholamines, fluids, and IABP) and also as a first-line treatment. Active percutaneous MCS results in better hemodynamic support compared to IABP. However, the appropriate role of MCS is uncertain as a positive impact on clinical outcomes or mortality has not yet been demonstrated in trials or metaanalyses.

Surgically implanted devices can support the circulation as bridging therapy for cardiac transplant candidates or as destination therapy **(Chap. 255)**. Assist devices should be used selectively in suitable patients based on decisions by a multidisciplinary team with expertise in the selection, implantation, and management of MCS devices.

Prognosis

The expected death rates for patients with MI complicated by CS range widely based on age, severity of hemodynamic abnormalities, severity of clinical hypoperfusion (arterial lactate, renal function), and performance of early revascularization. The recently introduced IABP-SHOCK II score predicts prognosis based on six readily available variables: age >73 years; prior stroke; glucose at admission >10.6 mmol/L (191 mg/dL); creatinine at admission >132.6 μmol/L (1.5 mg/dL); thrombolysis in myocardial infarction flow grade after PCI <3; and arterial blood lactate at admission >5 mmol/L. It also may help guide treatment strategies.

SHOCK SECONDARY TO RIGHT VENTRICULAR INFARCTION

Persistent CS due to predominant RV failure accounts for only 5% of CS complicating MI. It often results from proximal right coronary artery occlusion. The salient features are relatively high right atrial pressures, RV dilation and dysfunction, and only mildly or moderately depressed LV function. High right-sided pressures may be absent without volume loading. However, CS often has overlap combinations of both RV and LV ischemia, given a shared septum and the effect of ventricular interdependence on RV function. Management of isolated RV CS includes fluid administration to optimize right atrial pressure (10–15 mmHg); avoidance of excess fluids, which shift the interventricular septum into the LV; catecholamines; early reestablishment of infarct-artery flow; and MCS.

MITRAL REGURGITATION

(See also Chap. 269) Acute severe MR due to papillary muscle dysfunction and/or rupture may complicate MI and result in CS and/or pulmonary edema. This complication most often occurs on the first day, with a second peak several days later. The diagnosis is confirmed by echocardiography (Table 298-2). Afterload reduction with IABP and, if tolerated, vasodilators to reduce pulmonary edema, is recommended as a bridge to surgery or interventional treatment. Mitral valve repair or reconstruction is the definitive therapy and should be performed early in the course in suitable candidates. Other options include percutaneous edge-to-edge repair which has been successful in small case series.

VENTRICULAR SEPTAL RUPTURE

(See also Chap. 269) VSR complicating MI is a relatively rare event associated with very high mortality if CS is present (>80%). The incidence of infarct-related VSR without reperfusion was 1–2% but has decreased to 0.2% in the era of reperfusion. VSR occurs a median of 24 h after infarction, but may occur up to 2 weeks later. Echocardiography demonstrates shunting of blood from the left to the right ventricle and may visualize the opening in the interventricular septum. Current guidelines recommend immediate surgical VSR closure, irrespective of the patient's hemodynamic status, to avoid further hemodynamic deterioration. IABP support as bridge to surgery is recommended. Given high mortality, suboptimal surgical results and many patients not being eligible for surgery, interventional percutaneous VSR umbrella device closure has been developed. Results of interventional VSR closure suggest a similar outcome as surgery. How to close the VSR should be based on a heart team decision.

FREE WALL RUPTURE

Myocardial rupture is a dramatic complication of MI that is most likely to occur during the first week after the onset of symptoms. The clinical presentation typically is a sudden loss of pulse, blood pressure, and consciousness but sinus rhythm on ECG (pulseless electrical activity) due to cardiac tamponade **(Chap. 265)**. Free wall rupture may also result in CS due to subacute tamponade when the pericardium temporarily seals the rupture sites. Definitive surgical repair is required.

ACUTE FULMINANT MYOCARDITIS

(See also Chap. 254) Myocarditis can mimic acute MI with ST abnormalities or bundle branch block on the ECG and marked elevation of cardiac markers. Acute myocarditis causes CS in a small proportion of cases. These patients are typically younger than those with CS due to acute MI and often do not have typical ischemic chest pain. Echocardiography usually shows global LV dysfunction. Initial management is the same as for CS complicating acute MI but does not involve revascularization. Endomyocardial biopsy is recommended to determine the diagnosis and need for immunosuppressives for entities such as giant cell myocarditis. Refractory CS can be managed with MCS.

PULMONARY EDEMA

The etiologies and pathophysiology of pulmonary edema are discussed in Chap. 33.

Diagnosis

Acute pulmonary edema usually presents with the rapid onset of dyspnea at rest, tachypnea, tachycardia, and severe hypoxemia. Crackles and wheezing due to alveolar flooding and airway compression from peribronchial cuffing may be audible. Release of endogenous catecholamines often causes hypertension.

It is often difficult to distinguish between cardiogenic and non-cardiogenic causes of acute pulmonary edema. *Echocardiography* may identify systolic and diastolic ventricular dysfunction and valvular lesions. Electrocardiographic ST elevation and evolving Q waves are usually diagnostic of acute MI and should prompt immediate institution of MI protocols and coronary artery revascularization therapy **(Chap. 269)**. Brain natriuretic peptide levels, when substantially elevated, support heart failure as the etiology of acute dyspnea with pulmonary edema **(Chap. 252)**.

The use of a *Swan-Ganz catheter* permits measurement of pulmonary capillary wedge pressure (PCWP) and helps differentiate high-pressure (cardiogenic) from normal-pressure (non-cardiogenic) causes of pulmonary edema. PAC is indicated when the etiology of the pulmonary edema is uncertain, when edema is refractory to therapy, or when it is accompanied by hypotension. Data derived from use of a catheter often alter the treatment plan, but no impact on mortality rates has been demonstrated.

TREATMENT

TREATMENT Pulmonary Edema

The treatment of pulmonary edema depends on the specific etiology. As an acute, life-threatening condition, a number of measures must be applied immediately to support the circulation, gas exchange, and lung mechanics. Simultaneously, conditions that frequently complicate pulmonary edema, such as infection, acidemia, anemia, and acute kidney dysfunction, must be corrected.

SUPPORT OF OXYGENATION AND VENTILATION

Patients with acute cardiogenic pulmonary edema generally have an identifiable cause of acute LV failure—such as arrhythmia, ischemia/infarction, or myocardial decompensation **(Chap. 252)**—that may be rapidly treated, with improvement in gas exchange. In contrast, non-cardiogenic edema usually resolves much less quickly, and most patients require mechanical ventilation.

Oxygen Therapy

Support of oxygenation is essential to ensure adequate O_2 delivery to peripheral tissues, including the heart. Generally the goal is O_2 saturation of ≥92%, but very high saturation (>98%) may be detrimental.

Positive-Pressure Ventilation

Pulmonary edema increases the work of breathing and the O_2 requirements of this work, imposing a significant physiologic stress on the heart. When oxygenation or ventilation is not adequate in spite of supplemental O_2, positive-pressure ventilation by face or nasal mask or by endotracheal intubation should be initiated. Noninvasive ventilation **(Chap. 295)** can rest the respiratory muscles, improve oxygenation and cardiac function, and reduce the need for intubation. In refractory cases, mechanical ventilation can relieve the work of breathing more completely than can noninvasive ventilation. Mechanical ventilation with positive end-expiratory pressure can have multiple beneficial effects on pulmonary edema, as it: (1) decreases both preload and afterload, thereby improving cardiac function; (2) redistributes lung water from the intraalveolar to the extraalveolar space, where the fluid interferes less with gas exchange; and (3) increases lung volume to avoid atelectasis.

Renal Replacement Therapy

For pulmonary edema patients with refractory volume overload, metabolic acidosis (pH <7.15–7.25), hypoxemia, and/or persistent hyperkalemia, renal replacement therapy should be considered. For patients who are hypotensive or requiring ionotropic support, continuous renal replacement therapy usually is better tolerated than intermittent hemodialysis.

REDUCTION OF PRELOAD

In most forms of pulmonary edema, the quantity of extravascular lung water is determined by a combination of the pulmonary capillary pressures (PCWP), the pulmonary vascular permeability, and the intravascular volume status.

Diuretics

The “loop diuretics” furosemide, bumetanide, and torasemide are effective in most forms of pulmonary edema, even in the presence of hypoalbuminemia, hyponatremia, or hypochloremia. Furosemide is also a venodilator that rapidly reduces preload before any diuresis occurs, and is the diuretic of choice. The initial dose of furosemide should be ≤0.5 mg/kg, but a higher dose (1 mg/kg) is required in patients with renal insufficiency, chronic diuretic use, or hypervolemia or after failure of a lower dose. Combinations of diuretics and/or continuous infusion are helpful to achieve the desired degree of diuresis in selected patients.

Nitrates

Nitroglycerin and isosorbide dinitrate act predominantly as venodilators but have coronary vasodilating properties as well. Their onset is rapid and they are effectively administered by a variety of routes. Sublingual nitroglycerin (0.4 mg × 3 every 5 min) is first-line therapy for acute cardiogenic pulmonary edema. If pulmonary edema persists in the absence of hypotension, sublingual may be followed by IV nitroglycerin, commencing at 5–10 μg/min. IV nitroprusside (0.1–5 μg/kg per min) is a potent venous and arterial vasodilator. It is useful for patients with pulmonary edema and hypertension, but is not recommended in states of reduced coronary artery perfusion. It requires close monitoring and titration using an arterial catheter for continuous blood pressure measurement.

Morphine

Given in 2- to 4-mg IV boluses, morphine is a transient venodilator that reduces preload while relieving dyspnea and anxiety. These effects can diminish stress, catecholamine levels, tachycardia, and ventricular afterload in patients with pulmonary edema and systemic hypertension. However, some registry trials showed increased mortality by use of morphine.

Angiotensin-Converting Enzyme (ACE) Inhibitors

ACE inhibitors reduce both afterload and preload and are recommended for hypertensive patients. A low dose of a short-acting agent may be initiated and followed by increasing oral doses. In acute MI with heart failure, ACE inhibitors reduce short- and long-term mortality rates. The optimal starting point of ACE inhibitors has not been tested so far.

Other Preload-Reducing Agents

IV recombinant brain natriuretic peptide (nesiritide) is a potent vasodilator with diuretic properties and is effective in the treatment of cardiogenic pulmonary edema. It should be reserved for refractory patients and is not recommended in the setting of ischemia or MI.

Physical Methods

In nonhypotensive patients, venous return can be reduced by use of the sitting position with the legs dangling along the side of the bed.

Inotropic and Inodilator Drugs

The sympathomimetic amines dopamine and dobutamine (see above) are potent inotropic agents. The bipyridine phosphodiesterase-3 inhibitors (inodilators), such as milrinone (50 μg/kg followed by 0.25–0.75 μg/kg per min), stimulate myocardial contractility while promoting peripheral and pulmonary vasodilation. Inodilators may be helpful in selected patients with cardiogenic pulmonary edema and severe LV dysfunction, but there is little published clinical data.

Digitalis Glycosides

Once a mainstay of treatment because of their positive inotropic action **(Chap. 252)**, digitalis glycosides are rarely used at present. However, they may be useful for control of ventricular rate in patients with rapid ventricular response to atrial fibrillation or flutter and LV dysfunction with pulmonary edema, because they do not have the negative inotropic effects of other drugs that inhibit atrioventricular nodal conduction.

Intraaortic Balloon Counterpulsation IABP **(Chap. 255)** may be helpful in rare instances of acute MR from infective endocarditis, but is not typically used for pulmonary edema with CS.

Treatment of Tachyarrhythmias and Atrioventricular Resynchronization

Sinus tachycardia or atrial fibrillation can result from elevated left atrial pressure and sympathetic stimulation **(See also Chap. 247)** . Tachycardia itself can limit LV filling time and raise left atrial pressure further. Although relief of pulmonary congestion will slow the sinus rate or ventricular response in atrial fibrillation, a primary tachyarrhythmia may require cardioversion. In patients with reduced LV function and without atrial contraction or with lack of synchronized atrioventricular contraction, placement of an atrioventricular sequential pacemaker should be considered **(Chap. 239)**.

Reduction in Pulmonary Vascular Permeability

At present, no clinical therapies have been demonstrated as clinically effective to reduce the “leakiness” of the pulmonary capillaries.

Stimulation of Alveolar Fluid Clearance

A variety of drugs and cellular therapies can stimulate alveolar epithelial ion transport and upregulate the clearance of alveolar solute and water, but this strategy has not been proven beneficial in clinical trials thus far.

SPECIAL CONSIDERATIONS Risk of Iatrogenic Cardiogenic Shock

In the treatment of pulmonary edema, vasodilators lower blood pressure, and their use, particularly in combination, may lead to hypotension, coronary artery hypoperfusion, and shock (Fig. 298-1). In general, patients with a *hypertensive* response to pulmonary edema tolerate and benefit from these medications. In normotensive patients, low doses of single agents should be instituted sequentially, as needed.

Acute Coronary Syndromes

Acute STEMI complicated by pulmonary edema is associated with in-hospital mortality rates of 20–40% **(See also Chap. 269)** . After immediate stabilization, coronary artery blood flow must be reestablished rapidly. Early primary PCI is the method of choice; alternatively, a fibrinolytic agent should be administered. Early coronary angiography and revascularization by PCI or CABG also are indicated for patients with non-ST elevation acute coronary syndrome.

Extracorporeal Membrane Oxygenation (ECMO)

For patients with acute, severe non-cardiogenic edema with a potential rapidly reversible cause, ECMO may be considered in highly selected patients as a temporizing supportive measure to achieve adequate gas exchange with current survival to discharge rates of 50–60%. Usually

venovenous ECMO is used in this setting. ECMO can function as a bridge to transplantation or other interventions.

Unusual Types of Edema

Specific etiologies of pulmonary edema may require particular therapy. Re-expansion pulmonary edema can develop after removal of long-standing pleural space air or fluid. These patients may develop hypotension or oliguria with pulmonary edema resulting from rapid fluid shifts into the lung. Diuretics and preload reduction are contraindicated, and intravascular volume repletion often is needed while supporting oxygenation and gas exchange.

High-altitude pulmonary edema often can be prevented by use of dexamethasone, calcium channel–blocking drugs, or long-acting inhaled β_2-adrenergic agonists. Treatment includes descent from altitude, bed rest, oxygen, and, if feasible, inhaled nitric oxide; nifedipine may also be effective.

For pulmonary edema resulting from upper airway obstruction, recognition of the obstructing cause is key, because treatment then is to relieve or bypass the obstruction.

FURTHER READING

Authors/Task Force members, Windecker S et al: 2014 ESC/EACTS Guidelines on myocardial revascularization: The Task Force on Myocardial Revascularization of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J. 35:2541, 2014. [PubMed: 25173339]

Hochman JS: Cardiogenic shock complicating acute myocardial infarction: Expanding the paradigm. Circulation 107:2998, 2003. [PubMed: 12821585]

Hochman JS et al: Early revascularization in acute myocardial infarction complicated by cardiogenic shock. SHOCK Investigators. Should we emergently revascularize occluded coronaries for cardiogenic shock? N Engl J Med 341:625, 1999. [PubMed: 10460813]

Ouweneel DM et al: Impella CP versus intra-aortic balloon pump support in acute myocardial infarction complicated by cardiogenic shock. The IMPRESS in Severe Shock trial. J Am Coll Card 69:278, 2017.

Ponikowski P et al: 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J 37:2129, 2016. [PubMed: 27206819]

Thiele H et al: Intraaortic balloon support for myocardial infarction with cardiogenic shock. N Engl J Med 367:1287, 2012. [PubMed: 22920912]

Thiele H et al: Management of cardiogenic shock. Eur Heart J 36:1223, 2015. [PubMed: 25732762]

Thiele H et al: Percutaneous short-term active mechanical support devices in cardiogenic shock: A systematic review and collaborative meta-analysis of randomized trials. Eur Heart J 2017; epub ahead of print: doi.org/10.1093/eurheartj/ehx1363.

Thiele H et al: PCI strategies in patients with acute myocardial infarction and cardiogenic shock. N Engl J Med 2017; epub: DOI: 10.1056/NEJMoa1710261.

van Diepen S Contemporary management of cardiogenic shock: A scientific statement. Circulation 136:e232, 2017.

Chapter 299: Cardiovascular Collapse, Cardiac Arrest, and Sudden Cardiac Death

Christine M. Albert; William G. Stevenson

OVERVIEW AND DEFINITIONS

Cardiovascular collapse is severe hypotension from acute dysfunction of the heart or peripheral vasculature causing hypotension with resulting cerebral hypoperfusion and loss of consciousness that can be the result of a cardiac arrhythmia, severe myocardial or valvular dysfunction, loss of vascular tone, and/or acute disruption of venous return **(see Table 299-1)**. When an effective circulation is restored spontaneously, patients present with syncope **(see Chap. 18)**. If spontaneous resolution does not occur, then cardiac arrest occurs, ultimately resulting in death if resuscitation attempts are unsuccessful or not initiated. Underlying etiologies for cardiovascular collapse include benign conditions such as vasovagal syncope, but also life-threatening conditions, including: ventricular tachyarrhythmias, severe bradycardia, severely depressed myocardial contractility, as with massive acute myocardial infarction (MI) or pulmonary embolus, and other catastrophic events interfering with cardiac function such as myocardial rupture with cardiac tamponade or papillary muscle rupture with torrential mitral regurgitation.

Table 299-1 Distinction between Cardiovascular Collapse, Cardiac Arrest, and Death

Term	Definition	Qualifiers	Mechanisms
Cardiovascular collapse	Sudden loss of effective circulation due to cardiac and/or peripheral vascular factors that may reverse spontaneously (e.g., neurocardiogenic syncope, vasovagal syncope) or require interventions (e.g., cardiac arrest).	Broad term that includes cardiac arrest and transient events that characteristically revert spontaneously presenting as syncope.	Same as “Cardiac Arrest,” plus vasodepressor syncope or other causes of transient loss of blood flow.
Cardiac arrest	Abrupt cessation of cardiac function resulting in loss of effective circulation which may be reversible by prompt emergency medical intervention, but will lead to death in its absence.	Rare spontaneous reversions; likelihood of successful intervention relates to mechanism of arrest, clinical setting, availability of emergency medical services, and prompt return of circulation.	Ventricular fibrillation, ventricular tachycardia, asystole, bradycardia, pulseless electrical activity, noncardiac mechanical factors (e.g., pulmonary embolism).
Sudden cardiac death	Sudden unexpected death attributed to cardiac arrest, which if witnessed occurs within one hour of symptom onset.	In unwitnessed cases, the definition is often expanded to include unexpected deaths where the subject was documented to be well within the preceding 24 h.	Same as Cardiac Arrest.

Source: Modified from RJ Myerburg, A Castellanos: Cardiovascular collapse, cardiac arrest, and sudden cardiac death, in *Harrison’s Principles of Internal Medicine*, 19th ed, DL Kasper et al (eds). New York, McGraw-Hill Education, 2015, pp 1764–1771, Table 327-1.

Sudden cardiac arrest (SCA) refers to an *abrupt* loss of cardiac function resulting in complete cardiovascular collapse due either to an acute life-threatening cardiac arrhythmia or abrupt loss of myocardial pump function that requires emergency medical intervention for restoration of effective circulation. Most SCAs occur outside the hospital, and fewer than 10% of these victims survive to be discharged from the hospital despite undergoing attempted resuscitation by emergency medical services (EMS). For those that die prior to hospital admission, a cardiovascular cause for the arrest is often presumed based upon the absence of evidence for a traumatic or other non-cardiac cause at the time of the arrest. If the patient does not survive an SCA, the death is classified as a sudden cardiac death (SCD). Deaths that occur during hospitalization or within 30 days after resuscitated cardiac arrest are usually counted as SCDs in epidemiologic studies.

SCD also includes a broader category of unexplained rapid deaths thought to be due to cardiac causes where resuscitation was not attempted. In epidemiologic studies, SCD is usually defined as an unexpected death without obvious extra-cardiac cause that occurs in association with a witnessed rapid collapse or within 1 h of the onset of symptoms. This definition is based on the presumption that rapid deaths are often due to an arrhythmia, an assumption that cannot always be validated. Approximately half of all SCDs are not witnessed, and in the United States, few deaths undergo autopsies, and non-cardiac conditions that evolve rapidly such as acute cerebral hemorrhage, aortic rupture, and pulmonary embolism cannot be excluded without an autopsy. Therefore, definitive information necessary to establish the cause of death is usually not available. In unwitnessed cases, the definition is often further expanded to include unexpected deaths where the subject was documented to be well when last observed within the preceding 24 h. This expanded definition further decreases the certainty that the death was due to an arrhythmia or cardiac causes. The majority of countries, including the United States, do not have national surveillance systems or reporting requirements for SCD; thus the true incidence and frequency of SCD and its different mechanisms can only be estimated.

EPIDEMIOLOGY

DEMOGRAPHICS

SCA and SCD are major public health problems that account for 15% of all deaths and comprise 50% of all cardiac deaths. In the United States alone, there are an estimated 350,000 EMS-attended out-of-hospital cardiac arrests and 210,000 SCDs in the adult population. The estimated societal burden of premature death due to SCD is 2 million years of potential life lost for men and 1.3 million years of potential life lost for women, which is greater than most other leading causes of death. Although cardiac pathology, particularly coronary heart disease (CHD), underlies the majority of SCD, up to two-thirds of all SCD occur as the first clinical expression of previously undiagnosed heart disease. SCD rates have declined but not as steeply as rates for CHD in general. Age, gender, race, and geographic region all influence the incidence of SCD. Rates of out-of-hospital cardiac arrest are lower in Asia (52.5 per 100,000 person-years) than Europe (86.4 per 100,000 person-years), North America (98.1 per 100,000 person-years), or Australia (111.9 per 100,000 person-years); and also vary within geographic regions of the United States. SCD is rare in individuals of younger than 35 years of age (1–3 per 100,000 per year), and increases markedly with age as the incidence of coronary artery disease (CAD), heart failure (HF), and other predisposing conditions also increase. Although absolute SCD rates increase with age, the proportion of deaths that are due to SCD decreases markedly as other causes of death increase.

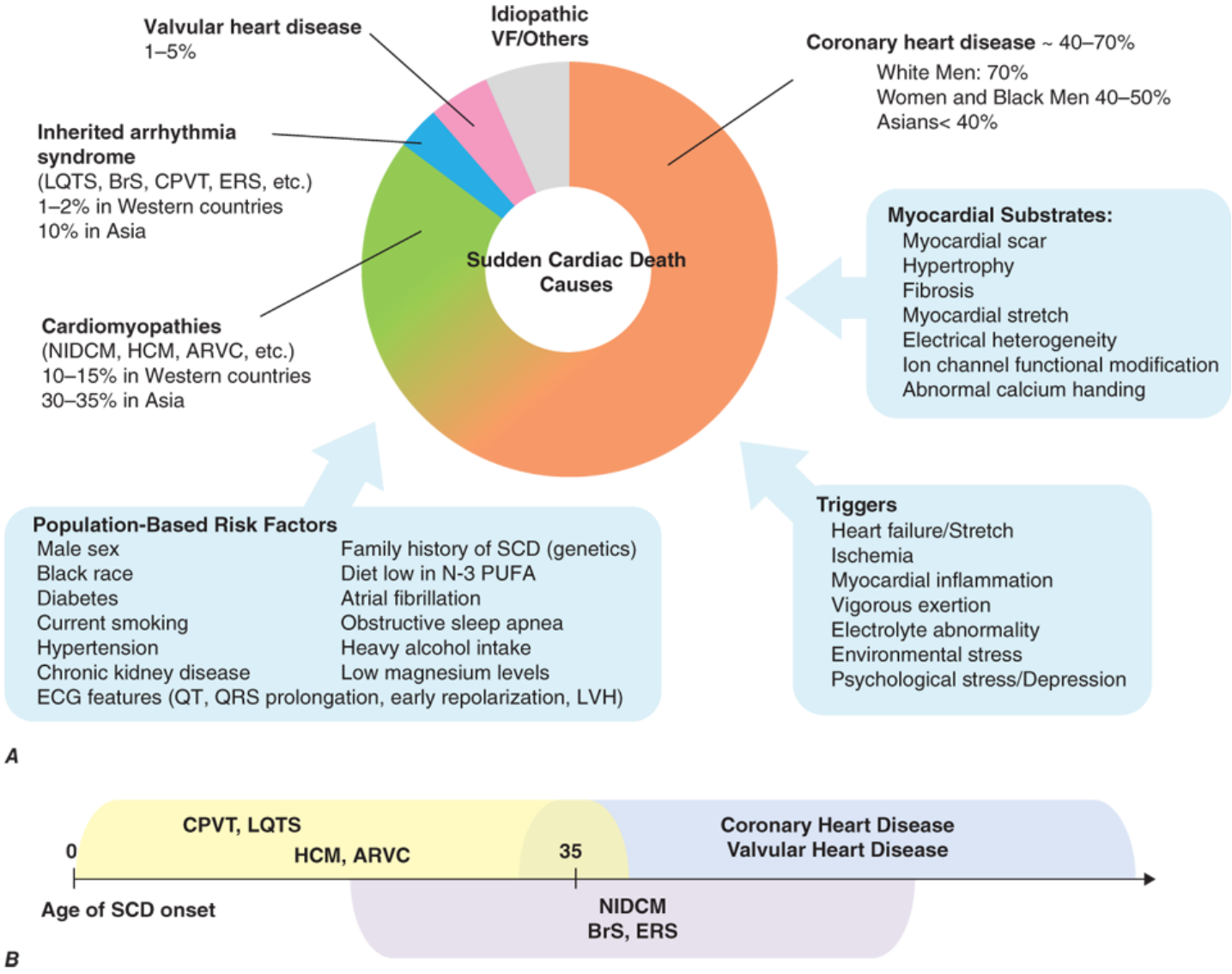
Women have a lower incidence of SCD and SCA than men, and women are more likely to present with pulseless electrical activity (PEA) and to have their SCD occur at home as compared to men. Possibly related to these factors, the SCD rate has not declined as much for younger women compared to men in recent years. Black as opposed to white Americans have higher rates of SCD, are more likely to have unwitnessed arrests, to be found with PEA, and have worse rates of survival. Socioeconomic disparities, with resuscitation being less likely in low income neighborhoods, is likely a contributing factor, but does not appear to account for the entirety of the elevated SCD rate in blacks. Alternatively, individuals of Hispanic ethnicity appear to have lower rates of SCD, despite having a higher prevalence of cardiac risk factors. It also appears that the incidence of SCD may be relatively low among Asian populations as well, both within the United States and globally. These gender and racial differences in SCD/SCA incidence and survival are poorly understood and warrant further research.

RISK FACTORS

The presence of overt structural heart disease and/or a certain types of inherited arrhythmia syndromes markedly elevates SCD risk **(see Chaps. 249 and 250) (See Fig. 299-1)**. Preexisting CHD and HF are the most prevalent predisposing cardiac conditions and are associated with four- to tenfold increases in SCD risk. Correspondingly, SCD shares many of the same risk factors with CHD and HF, including: hypertension, diabetes, hypercholesterolemia, obesity, and smoking. Diabetes is a particularly strong risk factor for SCD even in patients with established CHD. Hypertension and resultant left ventricular hypertrophy (LVH) appear to be particularly important markers of SCD risk in blacks, in whom the prevalence of these conditions is greater. Smoking markedly elevates risk, and smoking cessation lowers risk particularly among individuals who have not yet developed overt CHD. Serum cholesterol appears to be more strongly related to SCD at younger ages, and the benefits of cholesterol lowering on SCD incidence have not been firmly established. There also appears to be a genetic component to SCD risk that is distinct from that associated with other manifestations of atherosclerosis. A history of SCD among a first-degree relative is associated with an increased risk for SCD, and with the occurrence of ventricular fibrillation (VF) during acute MI, but is not associated with an increased risk for acute MI. These data suggest that genetic factors may predispose to fatal ventricular arrhythmia in the setting of ischemia, rather than to CHD in general.

FIGURE 299-1

A. Proportionate causes, substrates, risk factors, and triggers of sudden cardiac death (SCD); and ***B.*** variation of causes by age of onset. *(Modified from M Hayashi et al: The spectrum of epidemiology underlying sudden cardiac death. Circ Res 116:1887, 2015.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Obstructive sleep apnea and seizure disorders are also associated with increased SCD risk, and the underlying mechanism is not clear, but may be due to hypoxia and/or suffocation-induced cardiac arrest. Atrial fibrillation also appears to be associated with an increased risk of SCD, which is partly, but not entirely, accounted for by its association with underlying heart disease. Patients with chronic kidney disease are also at higher SCD risk with annualized SCD rates approaching 5.5% in patients undergoing dialysis. Electrolyte shifts and LVH, which are common in this population, have been suggested to play a role. There are also potential dietary influences on SCD risk. Individuals with higher intakes of polyunsaturated fatty acids, particularly n-3 fatty acids, and other components of a Mediterranean-style diet have lower SCD risks in observational studies, possibly due to antiarrhythmic effects of dietary components. Low levels of alcohol intake may be beneficial, but heavy intake (>3 drinks/day) appears to elevate risk.

PRECIPITATING FACTORS

SCD/SCA occurs with higher frequency at certain times, locations, and in association with certain activities and exposures. There are circadian variations in the incidence of SCD and cardiac arrest, with peaks in incidence in the morning hours and again in the later afternoon. There is also seasonal variability in SCD rates, which may be related to temperature and light exposure. Rates are highest during winter in the northern hemisphere and summer in the southern hemisphere. SCD rates also acutely peak during disasters such as earthquakes and terrorist attacks. SCA arrests are more likely to occur in certain locations as well, with notable clustering around train stations, airports, and other public places where there is significant population transit. SCD rates tend to be higher in urban areas and individuals that live near major roadways are at elevated SCD risk. There is also a well-recognized acute elevation in SCD risk that occurs during or shortly after bouts of vigorous exertion, and men appear to be more susceptible. Habitual exercise and training lowers this acute risk, but does not appear to eliminate it entirely. Exertion-associated SCDs are particularly tragic and highly publicized when they occur in highly trained athletes; however, the majority of such deaths actually occur in the general “non-athlete” population. The common thread amongst these precipitating factors is likely heightened autonomic tone, which can promote ischemia and has direct proarrhythmia and electrophysiologic actions that lower the threshold for VF.

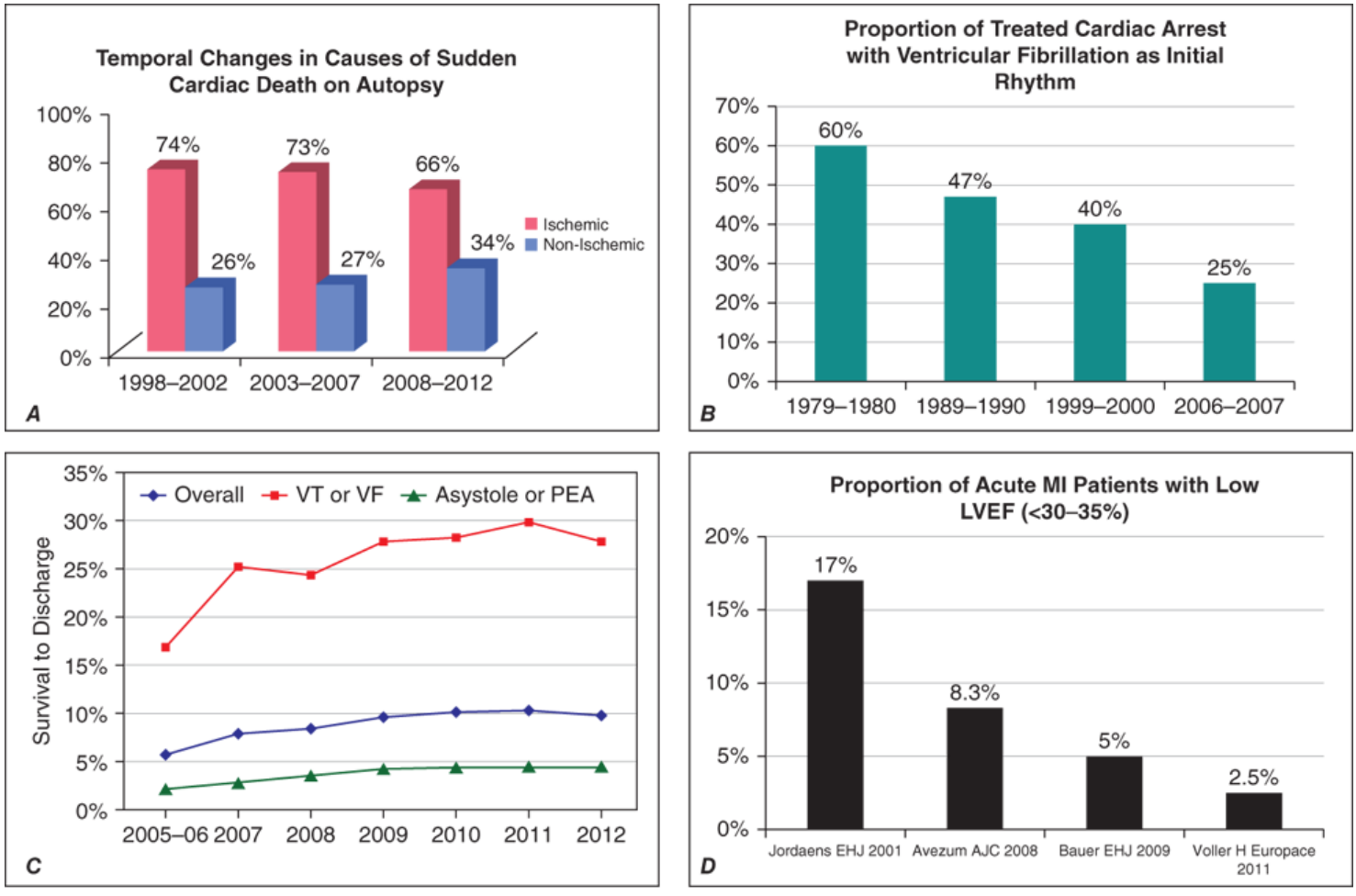
CAUSES OF SUDDEN CARDIAC DEATH

UNDERLYING HEART DISEASE

Our understanding regarding the diseases which contribute to SCD is derived primarily from autopsy series and cardiac evaluations in cardiac arrest survivors, which are highly variable in level of detail **(Fig. 299-1)**. Despite the limitations of these data, it is generally accepted that SCD is most commonly associated with underlying CAD, although the proportion due to CAD varies markedly by age, race, and sex. It is estimated that ~70–75% of SCDs in white men are due to CAD, as compared to only 40–50% in women and blacks. The proportion of SCDs with underlying CAD may be even lower in Asian ethnicities. Recent data suggest that the proportion of SCDs with CAD may be declining in some parts of Europe **(Fig. 299-2A)** and the United States, and, at the same time, increasing in parts of Japan and other parts of Asia. Beyond CAD, non-ischemic cardiomyopathies (hypertrophic, dilated, and infiltrative) are the second most frequent cause of SCD in the United States and European countries. Other less common causes include valvular heart disease, myocarditis, myocardial hypertrophy (often from hypertension), and rare primary electrical heart diseases such as the long QT and Brugada syndrome. On average, 5–10% of SCA victims do not have a significant cardiac abnormality at the time of autopsy or after extensive premortem cardiac evaluation, and this also varies by gender and race. Before 35 years of age, atherosclerotic CAD accounts for a much smaller proportion of deaths, with hypertrophic cardiomyopathy (HCM), coronary artery anomalies, myocarditis, arrhythmogenic right ventricular cardiomyopathy, and primary ion channelopathies accounting for a significant number of these deaths.

FIGURE 299-2

Changing epidemiology of sudden cardiac death/arrest. *A.* The proportion of sudden cardiac deaths attributable to coronary artery disease among individuals without a history of heart disease in Finland over time. Postmortem examinations are mandatory in Finland, which has the highest autopsy rate in Western World (*J Junttila et al: Circ Arrhythm Electrophysiol 2016*). ***B.*** Proportion of treated cardiac arrest with ventricular fibrillation as first recorded rhythm in Seattle, Washington, U.S. over time. *(Data from L Cobb et al: JAMA 288:3008, 2002, and G Nichol et al: JAMA 300:1423, 2008.)* ***C.*** Rates of overall survival and survival from shockable and nonshockable rhythms to hospital discharge among 70,027 out-of-hospital cardiac arrests across the United States from 2005 to 2012 (Cardiac Arrest to Enhance Survival Registry). (*From P Chan et al: Circulation 1876:1882, 2014.*) ***D.*** Proportion of myocardial infarction patients with left ventricular ejection fractions <30–35% in myocardial infarction registries over time.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CARDIAC RHYTHMS AND SUDDEN DEATH

The initial rhythm found when EMS arrive at the scene of an out-of- hospital cardiac arrest is an important indication of the potential cause of the arrest and of the prognosis. In the early days of EMS systems, over half of victims were found in VF, giving rise to the hypothesis

that ischmic VF or ventricular tachycardia (VT) degenerating to VF was the most common event. The proportion of cardiac arrests found in VF has decreased markedly since the 1970s, to only 20–25%, and PEA or asystole are now the most common scenarios **(Fig. 299-2B)**. However, the vast majority of cardiac arrests are not monitored at the time of collapse, and since arrhythmias are inherently unstable once hemodynamic collapse occurs, the rhythm at the time of EMS arrival may not reflect the rhythm that initially precipitated the SCA. VF and primary bradycardias can degenerate quickly into asystole. VF as an initial rhythm still predominates in public locations or in other situations when there is a short time between collapse and arrival of EMS, suggesting that VF remains a common initial rhythm. However, there are also data to support an absolute decrease in VF incidence. Proposed explanations include decreases in underlying CHD incidence, increased used of beta blockers in CHD, and implantable cardioverter defibrillators (ICD) in high-risk patients. There also appears to be an increase of PEA incidence over the past several years, suggesting that the proportion of SCD due to abrupt hemodynamic collapse in the absence of preceding fatal arrhythmia may be increasing. Proposed explanations for these proportional changes in PEA versus VF include the aging of the population and the increased prevalence of end-stage cardiovascular disease and other severe comorbidities. These older, sicker patients may be more likely to have arrests in the home and to have acute precipitants leading to PEA (i.e., respiratory, metabolic, vascular), and/or be less likely to sustain VF up to the point of EMS arrival.

DISEASE SPECIFIC MECHANISMS

Coronary artery disease can cause SCD through a number of mechanisms **(Table 299-2)**. The most common cause is acute MI or transient ischemia that leads to polymorphic VT and VF (**see Chap. 250**). Other primary mechanisms include severe bradyarrhythmias such as heart block with a slow escape rhythm, or PEA due to a massive MI or associated myocardial rupture. Areas of ventricular scar from prior infarcts increase the predisposition to reentrant VT, which often degenerates to VF. Once patients have suffered a MI, their risk of SCD elevates up to tenfold, with the highest absolute rates in the first 30 days after MI. The mechanisms underlying SCD vary at different time points after MI, with non-arrhythmic causes such as myocardial rupture and/or extensive reinfarction predominating early, within the first 1–2 months, and ischemic polymorphic VT and/or scar-related ventricular arrhythmias prevailing later. VT and sudden death can, and often does, occur years after an initial MI.

Table 299-2 Causes of Cardiovascular Collapse and Sudden Cardiac Arrest

Cause	Pathophysiologic Substrate	Rhythm Presentation
Cardiac Causes		
Coronary artery disease Atherosclerotic, coronary spasm, congenital anomalies	Acute myocardial ischemia /Infarction, ventricular rupture, tamponade Ventricular scar from healed infarction	Polymorphic VT/VF Bradyarrhythmia Pulseless electrical activity Ventricular tachycardia Ventricular fibrillation
Cardiomyopathies Dilated, hypertrophic, ARVC, infiltrative disease, valvular disease with LV failure	Ventricular scar Ventricular hypertrophy Pump failure	Ventricular tachycardia Polymorphic VT/VF Pulseless electrical activity Bradyarrhythmia
Congenital heart disease (*tetralogy of Fallot*, VSD, others)	Ventricular scar from surgical repair Hypertrophy	Ventricular tachycardia Bradyarrhythmias Polymorphic VT/VF Bradyarrhythmia
Aortic stenosis	Obstruction to outflow Ventricular hypertrophy	Pulseless electrical activity Bradyarrhythmia Polymorphic VT/VF
Mitral valve prolapse/Mitral regurgitation	Pump failure Ventricular scar	Ventricular tachycardia Polymorphic VT/VF
Arrhythmia syndromes without structural heart disease: Genetic: Long QT Brugada CPVT Idiopathic VF, early repolarization Drug toxicities (acquired long QT, others) Electrolyte abnormalities (severe hypokalemia)	Abnormal cellular electrophysiology	Polymorphic VT/VF
Wolff-Parkinson-White Syndrome	Accessory atrioventricular connection	Preexcited AF/VF
Non-Cardiac Causes of Cardiovascular Collapse		
Pulmonary embolism		PEA
Stroke		PEA, bradyarrhythmia
Aortic dissection		PEA, VF
Exsanguination		PEA
Tension pneumothorax		PEA
Sepsis		PEA
Neurogenic		PEA, bradyarrhythmia
Drug overdose		PEA, bradyarrhythmia

Abbreviations: AF, atrial fibrillation; ARVC, arrhythmogenic right ventricular cardiomyopathy; CPVT, catecholaminergic polymorphic ventricular tachycardia; LV, left ventricle; PEA, pulseless electrical activity; VF, ventricular fibrillation; VSD, ventricular septal defect; VT, ventricular tachycardia.

Cardiomyopathies and Other Forms of Structural Heart Disease

Scar-mediated reentrant VT can also occur in a host of nonischemic cardiomyopathies in which replacement fibrosis and/or inflammatory ventricular infiltrates occur **Chap. 249**. In congenital heart disease, surgical scars created during corrective surgery, such as those performed to correct ventricular septal defects in *tetralogy of Fallot*, can also serve as the substrate for ventricular re-entry. Other common predisposing processes such as LVH, ventricular stretch due to fluid overload, and cardiomyocyte dysfunction can result in electrical heterogeneity and other electrophysiologic changes that predispose to ventricular arrhythmias, including ion channel alterations that prolong action potential duration, impair cellular calcium handling, and diminished cellular coupling. These processes occur in a wide variety of diseases associated with depressed ventricular function and/or hypertrophy, including CAD, valvular heart disease, myocarditis, and non-ischemic cardiomyopathies.

Absence of Structural Heart Disease

In the absence of structural heart disease, VF can be due to an inherited ion channel abnormality, as in the long QT and Brugada syndromes **(Chap. 250)**, rapid atrial fibrillation associated with the Wolff-Parkinson-White syndrome **(Chap. 244)**, or drug toxicities, such as polymorphic VT due to drugs that prolong the QT interval **(Chap. 250)**. PEA can result from pulmonary emboli, exsanguination, or the terminal phase of respiratory arrest.

MANAGEMENT OF CARDIAC ARREST

As the ability to predict SCA in the population is very limited, community approaches to reduce death focus on the rapid identification of victims and implementation of resuscitation measures by those who first encounter the victim, most likely the lay public, who ideally summon EMS and initiate basic life support measures with chest compressions. The approach is codified in the “out-of-hospital chain of survival” which includes (1) initial evaluation and recognition of the SCA; (2) rapid initiation of cardiopulmonary resuscitation (CPR) with an emphasis on chest compressions; (3) defibrillation as quickly as possible usually with an automatic external defibrillation applied by the lay rescuer or EMT; (4) basic and advanced EMS; and (5) advanced life support and postcardiac arrest care. There have been major advances in each of these areas and survival rates to hospital discharge have increased from about 6% in 2005 to 10% in 2012, but much more progress is needed **(Fig. 299-2C)**.

The initial goal of resuscitation is to achieve the return of spontaneous circulation. Success is related to the time between collapse and initiation of resuscitation, decreasing markedly after 5 min, and the rhythm at the time of EMT arrival, being best for VT (25–30%), worse for VF and poor for PEA and asystole (<5%). Outcomes are also determined by the clinical state and comorbidities of the victim prior to the arrest, being worse for those with severe disease, such as cardiogenic shock, prior to arrest.

INITIAL EVALUATION AND INITIATION OF CPR

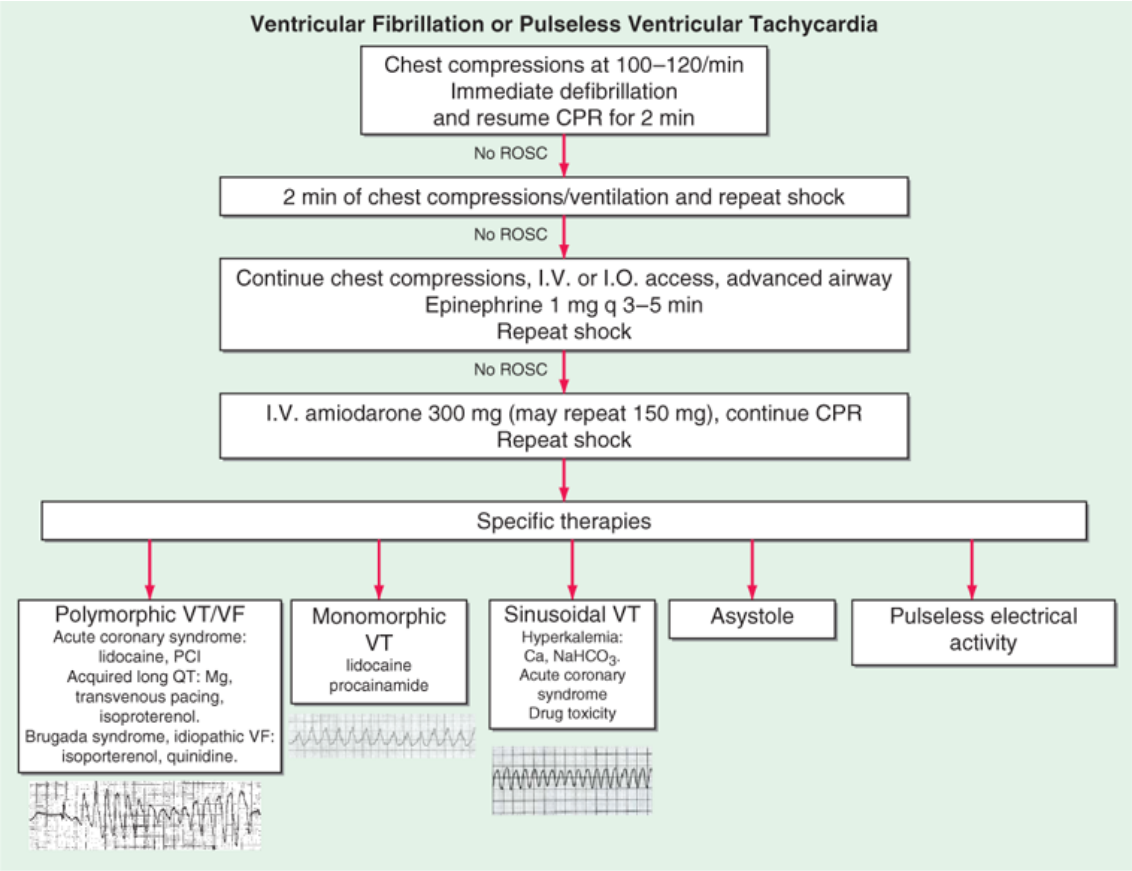
The rescuer should check for a response from the victim, shout for help, and call or ask someone else to call their local emergency number (e.g., 911), ideally on a cell phone that can be placed on speaker mode at the patient’s side such that the responding dispatcher can provide instructions and queries to the rescuer. Consideration of aspiration or airway obstruction is important and if suspected a Heimlich maneuver may dislodge the obstructing body. A trained healthcare provider would also check for a pulse (taking no longer than 10 s so as not to delay initiation of chest compressions) and assess breathing. Gasping respirations and brief seizure activity are common during SCA and may be misinterpreted as breathing and responsiveness. Chest compressions should be initiated without delay and administered at a rate of 100–120/min depressing the sternum by 5 cm (2 in.) and allowing full chest recoil between compressions. Chest compressions generate forward cardiac output with sequential filling and emptying of the cardiac chambers, with competent valves maintaining forward direction of flow. Interruption of chest compressions should be minimized to reduce end organ ischemia. Ventilation may be administered with two breaths for every 30 compressions if a trained rescuer is present, but for lay rescuers without training, chest compressions alone (“hands only CPR”) are more likely to be effectively applied and of similar benefit. If a second rescuer is present, they should be sent to seek out an automatic external defibrillator (AED), which are now widely available in many public areas.

RHYTHM-BASED MANAGEMENT

The rapidity with which defibrillation/cardioversion is achieved is an important predictor of outcome **(Fig. 299-3)**. A defibrillator, most often an AED, should be applied as soon as available. AEDs are easily used by lay rescuers and trained first responders, such as police officers and trained security guards. When the arrest is witnessed, the use of AEDs by lay responders can improve cardiac arrest survival rates. Once patches are applied to the chest, a brief pause in chest compressions is required to allow the AED to record the rhythm. An AED will advise a shock if the recorded rhythm meets criteria for VF or VT. Chest compressions are continued while the defibrillator is being charged. As soon as a diagnosis of VF or VT is established, a biphasic waveform shock of 150–200 J should be delivered. Chest compressions are resumed immediately and continue for 2 min until the next rhythm check. If VT/VF is still present, a second maximal energy shock is delivered. This sequence is continued until personnel to administer advanced life support are available, or return of spontaneous circulation is achieved. Electrocardiogram (ECG) rhythm strips produced by the AED should be retrieved, as the initial rhythm can be an important consideration in determining the cause of the arrest and to guide further therapy and evaluation if resuscitation is successful.

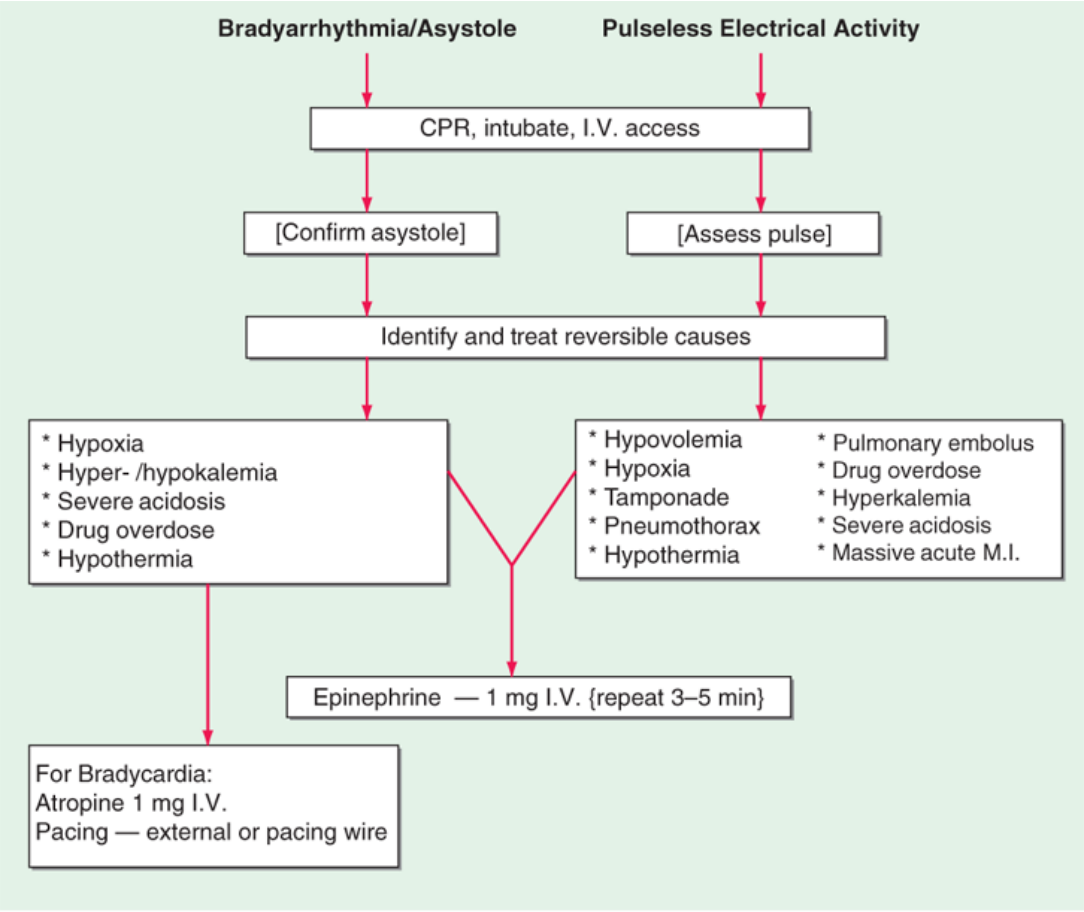
FIGURE 299-3

Algorithm for approach to cardiac arrest due to VT or VF (shockable rhythm). ***A.*** Chest compressions with ventilation and defibrillation or cardioversion should be initiated as soon as possible. Defibrillation should be repeated with minimal interruption of chest compressions. Once an intravenous or intraosseous access is established, administration of epinephrine defibrillation and amiodarone and defibrillation are performed. Further therapy can be guided by possible causes as suggested by the initial or recurrent cardiac rhythm as shown. i.v., intravenous; i.o., intraosseous; CPR, cardiopulmonary resuscitation; PCI, percutaneous coronary intervention. ***B.*** Algorithm for approach to cardiac arrest due to bradyarrhythmias/asystole and pulseless electrical activity. Chest compressions with ventilation (and intubation) should be initiated as soon as possible, and i.v. access should be obtained. Once an intravenous or intraosseous access is established, administration of epinephrine is performed. At the same time, an investigation for potential reversible causes should be made and any such causes should be treated if present. For bradycardic rhythms, atropine 1 mg IV and external subcutaneous or transvenous pacing are also performed. Defibrillation should be repeated with minimal interruption of chest compressions. Once an intravenous or intraosseous access is established, administration of epinephrine is performed. Further therapy can be guided by possible causes. CPR, cardiopulmonary resuscitation; I.O., intraosseous; I.V., intravenous.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

When advanced cardiac life support is available, an intravenous or intraosseous line is established for administration of medication and consideration given to placement of an advanced airway (endotracheal tube or supraglottic airway device). Epinephrine 1 mg every 3–5 min may be administered intravenously or intraosseously. If circulation is not restored or the patient is less than fully conscious despite return of circulation, confirmation that acidosis and hypoxia are adequately addressed should be assessed with arterial blood gas analysis. If metabolic acidosis persists after successful defibrillation and with adequate ventilation, 1 meq/kg $NaHCO_3$ may be administered.

The cardiac rhythm guides resuscitation when monitoring is available. VT is treated with external shocks synchronized to the QRS when VT is monomorphic, and asynchronous shocks for polymorphic VT or VF. If VT/VF recurs after one or more shocks, amiodarone 300 mg can be administered as a bolus via intravenous or intraosseous route in the hope that arrhythmia recurrence will be prevented after the next shock, followed by a 150 mg bolus if the arrhythmia recurs. If amiodarone fails, lidocaine can be administered.

Consideration of etiology should also guide therapy **(Chaps. 249 and 250)**. Commonly encountered causes of recurrent VT/VF may be due to ongoing myocardial ischemia or infarction that would benefit from emergent coronary angiography and revascularization, or QT prolongation causing the polymorphic VT *torsades des pointes* that may respond to administration of magnesium. Hyperkalemia should respond to administration of calcium, while other measures are implemented to reduce serum K.

PEA/asystole should be managed with CPR, ventilation, and administration of epinephrine. Causes of PEA/asystole that require specific therapy should be considered including airway obstruction, hypoxia, hypovolemia, acidosis, hyperkalemia, hypothermia, toxins, cardiac tamponade, tension pneumothorax, pulmonary embolism, and MI. Naloxone should be administered if opiate overdose is suspected.

POSTCARDIAC ARREST ACUTE MANAGEMENT

Following restoration of effective circulation, the possibility of acute MI should be immediately assessed. More than 90% of patients who have ST elevation consistent with acute MI will be found to have a culprit coronary stenosis/occlusion and likely benefit from emergent coronary angiography with percutaneous angioplasty and stenting. Angiography should also be considered if an acute coronary syndrome is suspected, even if ST segment elevation is absent, as more than half of selected patients undergoing angiography for this concern are found to have a coronary lesion as a potential cause of the ACA. Decisions regarding which patients without ST segment elevation should undergo urgent angiography are complex and factors such as hemodynamic or electrical instability, evidence of ongoing ischemia, comorbidities, and overall prognosis are taken into consideration.

Hemodynamic instability is often present following resuscitation and further ischemic end organ damage is a major consideration. Optimizing ventilation with consideration of acidosis, hypoxemia, and electrolyte abnormalities is important. Maintaining systolic BP at >90

mmHg, mean BP >65 mmHg is desirable and may require administration of vasopressors and adjustment of volume status. Potentially treatable reversible causes including hyperkalemia, severe hypokalemia, and drug toxicity with QT prolongation causing *torsades des pointes* should be identified and treated **Chap. 250**.

After stable spontaneous circulation is achieved, brain injury due to ischemia and reperfusion is a major determinant of survival and accounts for over two-thirds of deaths. The probability of successful neurologic recovery decreases rapidly with time between collapse and restoration of circulation and is <30% at 5 min in the absence of bystander CPR. The time between collapse and restoration of circulation is generally imprecise and some patients have a period of hypotensive VT prior to complete collapse, such that a reported long period prior to arrival of rescuers does not always preclude a good recovery. Therapeutic hypothermia (targeted temperature management) has been shown to improve the likelihood of survival and neurologic recovery in patients who present with shockable (VT or VF) rhythms and is recommended for all cardiac arrest patients who remain comatose, regardless of presenting rhythm, who have lack of purposeful response to verbal commands following return of spontaneous circulation. A constant target temperature of 32–36°C for at least 24 h is recommended. Shivering suppression with analgesics and sedatives may be needed. Induction of hypothermia should be started in-hospital, as no benefit was shown for implementation before hospital arrival, and administration of large volumes of cold saline for this purpose increased the risk of pulmonary edema. Brain injury is often accompanied by seizures and status epilepticus that may have further deleterious effects, warranting periodic or continuous electroencephalography (EEG) monitoring and treatment. A number of other therapies hoped to improve postarrest outcomes have been assessed, but have not been shown to be beneficial, including administration of corticosteroids, hemofiltration, and efforts to tightly control blood glucose.

Hypothermia and sedation preclude reliable prognostication for neurologic recovery. Functional neurologic assessment for neurologic recovery is generally deferred for at least 72 h after return to normothermia, typically 4–5 days after the cardiac arrest. Features that predict poor outcome include absence of pupillary reflex to light, status myoclonus, absence of EEG reactivity to external stimuli, and persistent burst suppression on EEG.

LONG-TERM MANAGEMENT AFTER SURVIVAL OF OUT-OF-HOSPITAL CARDIAC ARREST

For patients who survive cardiac arrest and have neurologic recovery, the likely underlying cause of the arrest guides further treatment. For arrests not due to an obvious non-cardiac cause, a full evaluation for the forms of structural heart disease outlined in **Fig. 299-1** and **Table 299-2** should be performed including an assessment for underlying CAD and ischemia as well as echocardiography and/or cardiac MRI to look for evidence of prior MI, valvular disease, nonischemic cardiomyopathies and to provide an assessment of left ventricular ejection fraction (LVEF). If the initial evaluation is not definitive or is suggestive of an inflammatory cardiomyopathy (i.e., sarcoidosis, myocarditis), a cardiac PET-scan and/or endomyocardial biopsy may also be performed. Patients without obvious structural abnormalities should undergo an evaluation for primary electrical disease (long-QT syndrome [LQTS], Brugada syndrome, early repolarization syndrome, or WPW). In cases where a heritable syndrome is suspected, further genetic evaluation should be considered. Diagnostic electrophysiology studies are warranted in selected patients to assess inducible arrhythmias, or perform provocative testing, such as with epinephrine challenge for LQTS, or sodium channel blocker (e.g., procainamide) challenge for Brugada syndrome.

Patients with shockable rhythms at arrest (VF and VT) that are not deemed to have been due to a transient reversible cause and have reasonable life expectancy should undergo insertion of an ICD for secondary prevention of SCA/SCD. Most of these patients will be found to have CAD. Patients with a VF arrest that occurs within the first 48 h of a documented acute MI generally do not require an ICD since they have a similar risk of sudden death over the next 5 years as infarct survivors who did not have a cardiac arrest. However, patients who have a large infarction with acutely depressed LV ejection fraction (e.g., <35%) have an increased risk for future development of life-threatening ventricular arrhythmias related to reentry in the infarct scar **Chap. 247**. The percentage of patients with such large infarcts has been declining due to improved treatment strategies for acute MI **(Fig. 299-2D)**. Implantation of an ICD early after MI in these patients does not, however, improve survival, in part because a significant number of sudden deaths in the first three months are due to recurrent myocardial ischemia or myocardial rupture, rather than arrhythmias. For patients with large infarcts a wearable defibrillator that will treat VT/VF if it occurs may be used, while left ventricular remodeling is taking place, followed by reevaluation of arrhythmia risk after the infarct is healed to determine if an ICD is warranted. Patients who experience VF in-hospital >48 h after MI or in the setting of myocardial ischemia without infarction may be at risk for recurrent VT/VF. These patients should be evaluated and optimally treated for ischemia. If there is evidence that clearly implicates ischemia immediately preceding the onset of VF without evidence of a prior MI, coronary revascularization may be adequate therapy. Others may warrant an ICD. When the cardiac arrest is due sustained monomorphic VT, a prior infarct scar is often present and the recurrence rate is significant regardless of whether the arrest occurred in association with elevated serum troponin. An ICD is usually warranted even if revascularization is also needed.

Patients who have cardiac arrest due to a treatable reversible cause, such as hyperkalemia, or drug toxicity with QT prolongation causing *torsades des pointes* **Chap. 250**, that can be adequately addressed and prevented by other means do not usually need an ICD. An ICD is usually recommended for cardiac arrest due to VT or VF without a clearly reversible cause, particularly when structural heart disease, such as hypertrophic or dilated cardiomyopathy, arrhythmogenic cardiomyopathy, cardiac sarcoidosis, or a cardiac syndrome associated with sudden death, including Brugada syndrome, or LQTS are present (**Chaps. 249 and 250**). In patients with structural heart disease, it is important to recognize that life-threatening arrhythmias can be an indication of terminal, end-stage heart disease with minimal prospect for meaningful survival despite successful resuscitation, and ICDs will not alter the course of these patients and should not be implanted in this situation, unless there is a prospect for cardiac replacement therapy with future cardiac transplantation or a ventricular assist device.

PREVENTION OF SCD

Although advances in CPR and postresuscitation care have improved survival rates after cardiac arrest, 90% of patients will not survive to be discharged from the hospital. Of those that do survive, a proportion (~20%) are left with severe neurologic and/or physical disability. The majority of cardiac arrests do not occur in public places where AEDs and rapid defibrillation have the greatest impact. Patients who suffer an arrest at home also have longer EMS response times and are much less likely to be found in VF. Finally, 50% of cardiac arrests are not witnessed precluding effective resuscitation efforts. Thus, preventive efforts are critical to reducing mortality from cardiac arrest.

SCD RISK STRATIFICATION

The presence of overt structural heart disease and/or primary electrical heart disease is associated with an increased risk of SCD that varies with the severity and type of disease. For patients with structural heart disease, depressed left ventricular function is the best validated marker for risk, and clinical HF elevates risk further. After MI, SCD risk increases gradually as the LVEF decreases to 40% and then exponentially thereafter. In addition to LVEF and CHF, other potential markers of increased SCD risk in the setting of structural heart disease include unexplained syncope, sustained VT induced at electrophysiology study (EP study), left ventricular scar size and heterogeneity on cardiac magnetic resonance, markers of altered autonomic function and altered repolarization, and QRS prolongation. The majority of these tests, with the exception of the EP study in post-MI patients, broadly predict death from cardiovascular causes and are not able to discriminate patients who will die suddenly from those who will die of other cardiac causes. For instance, patients with the greatest degree of systolic HF and/or lowest LVEF, although at elevated risk for SCD, are more likely to die from HF. Although sustained VT at EP study does identify individuals at a higher risk of SCA versus non-SCA in certain subsets of patients, the sensitivity of the test is generally inadequate when LV function is significantly reduced.

PREVENTIVE THERAPIES FOR SCD IN HIGH-RISK POPULATIONS

Therapy with beta-adrenergic blockers has been demonstrated to reduce SCD risk in a multitude of settings including after MI, among patients with ischemic and nonischemic cardiomyopathy, and in LQTS. Angiotensin-converting enzyme inhibitors, aldosterone antagonists, and most recently angiotensin-receptor/neprilysin inhibitors have been associated with reductions in SCD in subsets of patients with structural heart disease, primarily ischemic and non-ischemic cardiomyopathy accompanied by HF. Coronary artery bypass grafting has also been associated with reductions in SCD risk, and revascularization in general may lower SCD risk through reduction in ischemic events and resultant improvements in left ventricular systolic function by reducing areas of hibernating myocardium.

For patients whose disease continues to confer substantial risk of sustained VT or VF on optimal medical therapy, an ICD is recommended **(Table 299-3)**. The ICD indication in these patients is referred to as “primary prevention of sudden death.” The indications for primary prevention ICDs vary depending on the type of underlying structural heart disease and its severity, and variable strength of evidence. In patients with a history of MI more than 40 days ago, primary prevention ICDs are indicated for those with Class II-III NYHA HF and LVEF <35%, and those who are NYHA functional Class I with LVEF <30%. Although, ICDs have not been found to be beneficial when implanted within 40 days of a MI, those with recent or old MI, nonsustained VT, LVEF <40% and inducible sustained VT at EP study also warrant an ICD. In general, these criteria are not applied to patients who are within 90 days of myocardial revascularization, since some will experience improvement in ventricular function and older trial data suggested there was no benefit with ICDs in these patents. High-risk patients with low LVEFs may be considered for a wearable defibrillator with later reassessment of ventricular function and ICD placement.

Table 299-3 Implantable Cardioverter Defibrillator (ICD) Indications

	Indication	Class of Recommendation*	Level of Evidence**
Secondary Prevention			
All disease states	ICD therapy is indicated in patients who are survivors of cardiac arrest due to VF or hemodynamically unstable sustained VT after evaluation to define the cause of the event and to exclude any completely reversible causes.	CLASS I	A
VT or VF	ICD therapy is indicated in patients with structural heart disease and spontaneous sustained VT, whether hemodynamically stable or unstable.	CLASS I	B
	ICD implantation is reasonable for patients with sustained VT and normal or near-normal ventricular function.	CLASS IIa	C
Syncope	ICD therapy is indicated in patients with syncope of undetermined origin with clinically relevant, hemodynamically significant sustained VT or VF induced at electrophysiological study.	CLASS I	B
	ICD therapy may be considered in patients with syncope and advanced structural heart disease in whom invasive and noninvasive have failed to determine a cause.	CLASS IIb	C
Primary Prevention			
Ischemic cardiomyopathy	ICD therapy ***is indicated*** in patients with LVEF ≤ 35% due to prior MI who are at least 40 days post-MI and are rare in NYHA functional Class II or III.	CLASS I	B
	ICD therapy ***is indicated*** in patients with LV dysfunction due to prior MI who are at least 40 days post-MI, have an LVEF ≤ 30%, and are in NYHA functional Class I.	CLASS I	A
	ICD therapy ***is indicated*** in patients with nonsustained VT due to prior MI, LVEF ≤ 40%, and inducible VF or sustained VT at electrophysiological study.	CLASS I	B
	ICD therapy is ***indicated*** in patients with nonischemic DCM who have an LVEF ≤35% and who are in NYHA functional Class II or III.	CLASS I	B
Non-ischemic cardiomyopathy	ICD implantation ***is reasonable*** for patients with unexplained syncope, significant LV dysfunction, and nonischemic DCM.	CLASS IIa	C
	ICD therapy ***can be considered*** in patients with non-ischemic heart disease and NYHA functional Class I.	CLASS IIb	C
Hypertrophic cardiomyopathy	ICD implantation ***is reasonable*** for patients with HCM who have one or more major risk factors for SCD.	CLASS IIa	C
Arrhythmogenic right ventricular dysplasia	ICD implantation ***is reasonable*** for the prevention of SCD in patients with ARVC who have one or more risk factors for SCD.	CLASS IIa	C
Brugada syndrome	ICD implantation ***is reasonable*** for patients with Brugada syndrome who have had syncope.	CLASS IIa	C
	ICD implantation ***is reasonable*** for patients with Brugada syndrome who have documented VT that has not resulted in cardiac arrest.	CLASS IIa	C

	Indication	Class of Recommendation*	Level of Evidence**
Long-QT syndrome	ICD implantation ***is reasonable*** to reduce SCD in patients with long-QT syndrome who are experiencing syncope and/or VT while receiving beta blockers.	CLASS IIa	B
	ICD ***may be considered*** as primary therapy in patients long-QT syndrome who are deemed to be at very high risk, especially those with a contraindication to beta-blocker therapy.	CLASS IIb	B
Catecholaminergic polymorphic VT (CPVT)	ICD implantation ***is reasonable*** for patients with catecholaminergic polymorphic VT who have syncope and/or documented sustained VT while receiving beta blockers.	CLASS IIa	C
Familial cardiomyopathy	ICD therapy ***may be considered*** in patients with a familial cardiomyopathy associated with SCD.	CLASS IIa	C
LV noncompaction	ICD therapy ***may be considered***.	CLASS IIa	C

Class of Recommendation*			**Level of Evidence****	
			LEVEL B	LEVEL C
CLASS I Procedure/treatment SHOULD be performed/administered	CLASS IIa Additional studies with focused objectives needed. IT IS REASONABLE to perform procedure/administer treatment.	LEVEL A Multiple populations evaluated. Data derived from multiple randomized clinical trials or meta-analyses.	Limited populations evaluated. Data derived from a single randomized trial or nonrandomized studies.	Very limited populations evaluated. Only consensus opinion of experts, case studies, or standard of care.

Abbreviations: VF, ventricular fibrillation; VT, ventricular tachycardia.

ICDs for primary prevention of sudden death are also recommended for patients with diseases other than CAD, that put them at risk for SCD. ICDs are recommended for those with nonischemic DCM who have an LVEF ≤ 35% and who are in NYHA functional Class II or III and receiving medical therapy. Benefit is more likely in patients aged <60 years, as non-arrhythmia causes of death increase with age. For patients with LVEF <35%, HF, and left bundle branch block (LBBB) with QRS duration >150 ms cardiac resynchronization therapy also offers protection against SCD, particularly in patients with nonischemic cardiomyopathy. Primary prevention ICDs are also recommended in select high-risk patients with HCM, arrhythmogenic right ventricular dysplasia, cardiac sarcoidosis, and Brugada syndrome and some patients with congenital LQTS with high-risk features or that have failed therapy with beta-adrenergic blocking agents. There are circumstances where an ICD is not indicated even if there is a significant sudden death risk **(Table 299-4)**. Most notably, patients who do not have a reasonable expectation of survival with an acceptable functional status for at least 1 year, should not undergo ICD placement.

Table 299-4 Implantable Cardioverter Defibrillator (ICD) Not Indicated

Patients who do not have a reasonable expectation of survival with an acceptable functional status for at least one year, even if they meet ICD implantation criteria.
Patients with incessant VT or VF.
Patients with significant psychiatric illnesses that may be aggravated by device implantation or that may preclude systematic follow-up.
Patients with drug-refractory New York Heart Association class IV congestive heart failure who are not candidates for cardiac transplantation or cardiac resynchronization therapy.
Syncope of undetermined cause in a patient without inducible ventricular tachyarrhythmias and without structural heart disease.
VF or VT is amenable to surgical or catheter ablation in patients without other disease predisposing to SCA (e.g., atrial arrhythmias associated with the Wolff-Parkinson-White syndrome, RV or LV outflow tract VT, idiopathic VT, or fascicular VT in the absence of structural heart disease).
Patients with ventricular tachyarrhythmias due to a completely reversible disorder in the absence of structural heart disease (e.g., electrolyte imbalance, drugs, or trauma).

Adapted from: AE Epstein, JP DiMarco, KA Ellenbogen, et al: 2012 ACCF/AHA/HRS focused update incorporated into the ACCF/AHA/HRS 2008 guidelines for device-based therapy of cardiac rhythm abnormalities: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. Circulation 127:e283–352, 2013.

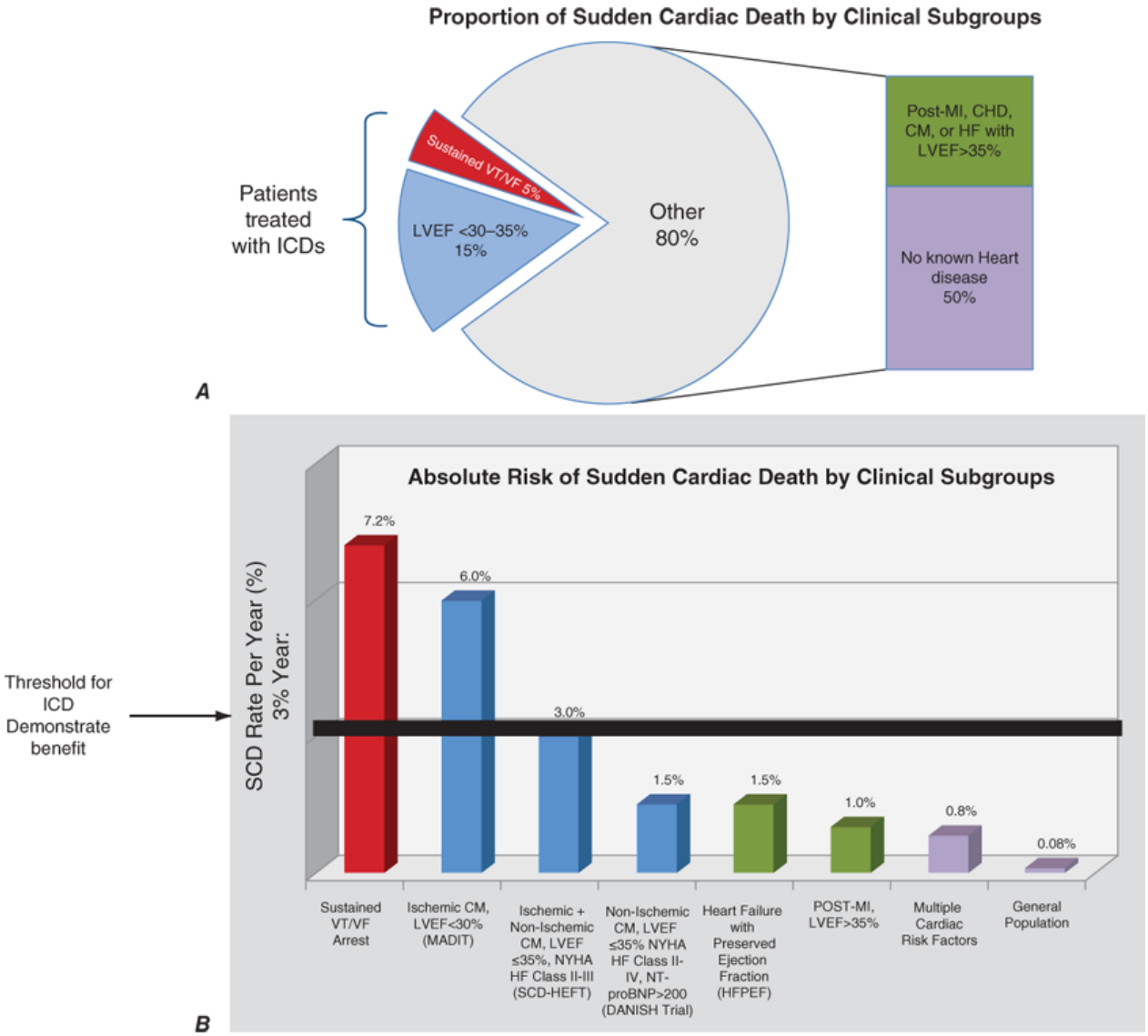
Abbreviations: LV, left ventricular; RV, right ventricular; VF, ventricular fibrillation; VT, ventricular tachycardia.

THE CHALLENGE OF SCD PREVENTION

(Fig. 299-4)

FIGURE 299-4

A. Proportion of sudden cardiac deaths that occur in clinical subgroups of the population treated and not treated with ICDs. ***B.*** Absolute risk on sudden cardiac death within clinical subgroups in comparison to the threshold of risk where ICDs demonstrated benefit.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The Greatest Number of Sudden Deaths Occur in "Low Risk" Patients

While patients with reduced left ventricular function and HF are at substantially elevated SCD risk, only ~20% of all SCDs occur in patients with poor left ventricular function. Most SCDs occur in individuals with preserved ventricular function who would not qualify for a primary prevention ICD. Although SCD rates are elevated compared to the general population, the absolute SCD risk in patients with CHD or HF who have an LVEF >35% is not high enough to warrant consideration of ICD therapy. While the incidence of SCD is lower in patients with preserved LVEF, SCD accounts for a greater proportion of cardiac deaths, and active efforts are being made to advance SCD risk stratification in this segment of the population. However, at the present time, SCD prevention primarily involves cardiac risk factor modification and standard medical therapy for the underlying condition.

Preventing Sudden Death in the General Population

Only about half of men and a third of women who suffer SCA are recognized to have heart disease prior to the event, and only half have warning symptoms prior to the event. SCD often occurs without warning as the first manifestation of cardiac disease. In order to prevent these SCDs, preventive interventions would need to be employed broadly to the general population. Although several risk scores have recently been developed with the intent to stratify SCD risk in low-risk populations, the clinical utility to date is limited by the low absolute incidence of SCD which is estimated to be only 50–90 per 100,000 in the general adult population. Therefore, current efforts aimed at preventing SCD in general populations primarily focus on modification of the SCD risk factors outlined previously. Individuals who adhere to a low risk, healthy lifestyle that includes avoidance of smoking, maintaining a healthy body weight, participating in moderate exercise,

and a Mediterranean-type dietary pattern have markedly lower rates of SCD. A substantial number of SCDs are likely to be preventable thorough lifestyle modifications and treatment of risk factors.

ACKNOWLEDGMENT

The authors gratefully acknowledge the prior contributions of Agustin Castellanos and Robert J. Myerburg.

FURTHER READING

Al-Khatib SM et al: 2017 AHA/ACC/HRS Guideline for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. Heart Rhythm Oct 26. pii: S1547-5271(17)31250-X. doi: 10.1016/j.hrthm.2017.10.036. [Epub ahead of print], 2017.
Callaway CW et al: Part 8: Post-Cardiac Arrest Care: 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation 132:S465, 2015. [PubMed: 26472996]
Deo R, Albert CM: Epidemiology and genetics of sudden cardiac death. Circulation 125:620, 2012. [PubMed: 22294707]
Fishman GI et al: Sudden cardiac death prediction and prevention report from a National Heart, Lung, and Blood Institute and Heart Rhythm Society workshop. Circulation 122:2335, 2010. [PubMed: 21147730]
Hayashi M et al: The spectrum of epidemiology underlying sudden cardiac death. Circ Res 116:1887, 2015. [PubMed: 26044246]
Link MS et al: Part 7: Adult advanced cardiovascular life support: 2015 American Heart Association guidelines update for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation 132:S444, 2015. [PubMed: 26472995]
Myerburg RJ et al: Pulseless electric activity: Definition, causes, mechanisms, management, and research priorities for the next decade: Report from a National Heart, Lung, and Blood Institute Workshop. Circulation 128:2532, 2013. [PubMed: 24297818]
Neumar RW et al: Part 1: Executive summary. 2015 American Heart Association guidelines update for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation 132:S315, 2015. [PubMed: 26472989]
Stecker EC et al: Public health burden of sudden cardiac death in the united states. Circ Arrhythm Electrophysiol 7:212, 2014. [PubMed: 24610738]
Zipes DP et al: ACC/AHA/ESC 2006 Guidelines for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: A report of the American College of Cardiology/American Heart Association Task Force and the European Society of Cardiology Committee for Practice Guidelines (writing committee to develop guidelines for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death): Developed in collaboration with the European Heart Rhythm Association and the Heart Rhythm Society. Circulation 114:e385, 2006. [PubMed: 16935995]

Chapter 300: Coma

S. Andrew Josephson; Allan H. Ropper; Stephen L. Hauser

INTRODUCTION

Coma is among the most common neurologic emergencies encountered general medicine and requires an organized approach. It accounts for a substantial portion of admissions to emergency wards and occurs on all hospital services.

There exists a continuum of states of reduced alertness, the most severe form being coma, defined as a deep sleeplike state with eyes closed from which the patient cannot be aroused. Stupor refers to a higher degree of arousability in which the patient can be transiently awakened by vigorous stimuli, accompanied by motor behavior that leads to avoidance or withdrawal from uncomfortable or aggravating stimuli. Drowsiness simulates light sleep and is characterized by easy arousal and the persistence of alertness for brief periods. Stupor and drowsiness are usually accompanied by some degree of confusion **(Chap. 24)**. A precise narrative description of the level of arousal and of the type of responses evoked by various stimuli as observed at the bedside is preferable to use of ambiguous terms such as lethargy, semicoma, or obtundation.

Several conditions that render patients unresponsive and simulate coma are considered separately because of their special significance. The *vegetative state* signifies an awake-appearing but nonresponsive state often in a patient who has emerged from coma. In the vegetative state, the eyelids may open periodically, giving the appearance of wakefulness. Respiratory and autonomic functions are retained. Yawning, coughing, swallowing, and limb and head movements persist, but there are few, if any, meaningful responses to the external and internal environment. There are always accompanying signs that indicate extensive damage in both cerebral hemispheres, e.g., decerebrate or decorticate limb posturing and absent responses to visual stimuli (see below). In the closely related but less severe *minimally conscious state*, the patient displays rudimentary vocal or motor behaviors, often spontaneous, but some in response to touch, visual stimuli, or command. Cardiac arrest with cerebral hypoperfusion and head trauma are the most common causes of the vegetative and minimally conscious states **(Chap. 301)**.

The prognosis for regaining mental faculties once the vegetative state has supervened for several months is very poor, and after a year, almost nil; hence the term *persistent vegetative state*. Most reports of dramatic recovery, when investigated carefully, are found to yield to the usual rules for prognosis, but there have been rare instances in which recovery has occurred to a severely disabled condition and, in rare childhood cases, to an even better state. Patients in the minimally conscious state carry a better prognosis for some recovery compared to those in a persistent vegetative state, but even in these patients, dramatic recovery after 12 months is unusual.

The possibility of incorrectly attributing meaningful behavior to patients in the vegetative and minimally conscious states creates inordinate problems and anguish for families and physicians. On the other hand, the question of whether these patients lack any capability for cognition has been reopened by functional MRI studies that have demonstrated, in a small proportion of usually posttraumatic cases, meaningful cerebral activation in response to verbal and other stimuli as discussed in more detail below. This finding suggests at a minimum that some of these patients could in the future be able to

communicate their needs using technological advances and that further research could shed light on treatment approaches targeting areas of the brain and their connections that seem to be preserved in individual patients.

Apart from the above conditions, several syndromes that affect alertness are prone to be misinterpreted as stupor or coma. Clinicians should be aware of these pitfalls when diagnosing coma at the bedside. Akinetic mutism refers to a partially or fully awake state in which the patient is able to form impressions and think, as demonstrated by later recounting of events, but remains virtually immobile and mute. The condition results from damage in the regions of the medial thalamic nuclei or the frontal lobes (particularly lesions situated deeply or on the orbitofrontal surfaces) or from extreme hydrocephalus. The term abulia describes a milder form of akinetic mutism characterized by mental and physical slowness and diminished ability to initiate activity. It is also usually the result of damage to the medial frontal lobes and their connections **(Chap. 26)**.

Catatonia is a hypomobile and mute syndrome that occurs usually as part of a major psychosis, typically schizophrenia or major depression. Catatonic patients make few voluntary or responsive movements, although they blink, swallow, and may not appear distressed. There are nonetheless signs that the patient is responsive, although it may take a careful examination to demonstrate them. For example, eyelid elevation is actively resisted, blinking occurs in response to a visual threat, and the eyes move concomitantly with head rotation, all of which are inconsistent with the presence of a brain lesion causing unresponsiveness. It is characteristic but not invariable in catatonia for the limbs to retain the postures in which they have been placed by the examiner (“waxy flexibility,” or catalepsy). With recovery, patients often have some memory of events that occurred during their catatonic stupor. Catatonia is superficially similar to akinetic mutism, but clinical evidence of cerebral damage such as hyperreflexia and hypertonicity of the limbs is lacking. The special problem of coma in brain death is discussed below.

The locked-in state describes an important type of pseudocoma in which an awake patient has no means of producing speech or volitional limb movement but retains voluntary vertical eye movements and lid elevation, thus allowing the patient to signal with a clear mind. The pupils are normally reactive. The usual cause is an infarction (e.g., basilar artery thrombosis) or hemorrhage of the ventral pons that transects all descending motor (corticospinal and corticobulbar) pathways. Another awake but de-efferented state occurs as a result of total paralysis of the musculature in severe cases of neuromuscular weakness such as in Guillain-Barré syndrome **(Chap. 439)**, critical illness neuropathy **(Chap. 301)**, and pharmacologic neuromuscular blockade.

THE ANATOMY AND PHYSIOLOGY OF COMA

Almost all instances of coma can be traced to either (1) widespread abnormalities of the cerebral hemispheres or to (2) reduced activity of a special thalamocortical alerting system termed the reticular activating system (RAS) which is diffusely located in the brainstem. The proper functioning of this system, its ascending projections to the cortex, and the cortex itself are required to maintain alertness and coherence of thought. In addition to structural damage of these two systems, suppression of reticulocerebral function can occur by drugs, toxins, or metabolic derangements such as hypoglycemia, anoxia, uremia, and hepatic failure; these types of metabolic causes of coma are far more common than structural injuries.

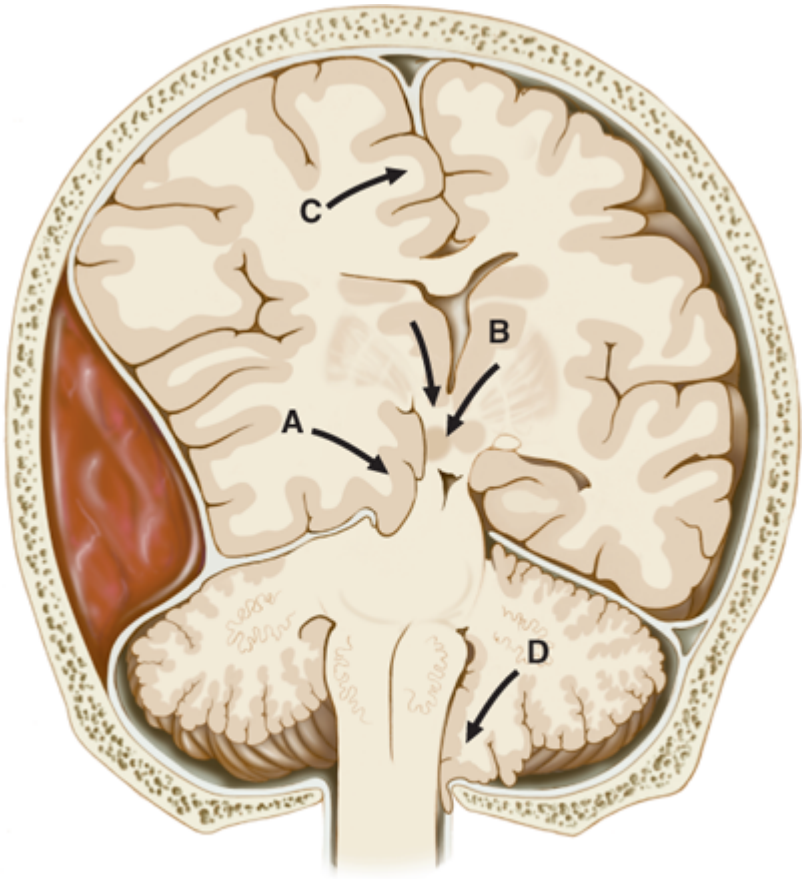
Coma Due to Cerebral Mass Lesions and Herniation Syndromes

In addition to the fixed restriction of the skull, the cranial cavity is separated into compartments by infoldings of the dura. The two cerebral hemispheres are separated by the falx, and the anterior and posterior fossae by the tentorium. Herniation refers to displacement of brain tissue by an overlying or adjacent mass into a contiguous compartment that it normally does not occupy. Coma and many of its

associated signs can be attributed to these tissue shifts, and certain clinical features are characteristic of specific configurations of herniation (Fig. 300-1). They are in essence “false localizing” signs because they derive from compression of brain structures at a distance from the mass lesion that is the direct cause of coma.

FIGURE 300-1

Types of cerebral herniation: (***A***) uncal; (***B***) central; (***C***) transfalcial; and (***D***) foraminal.

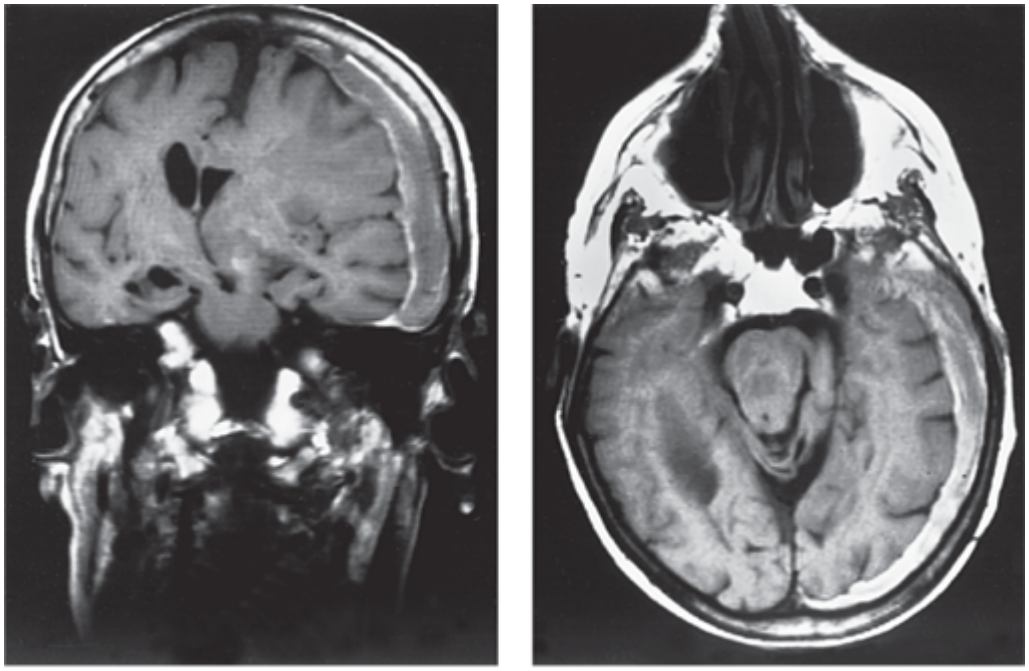


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In the most common form of herniation, brain tissue is displaced from the supratentorial to the infratentorial compartment through the tentorial opening; this is referred to as transtentorial herniation. Uncal transtentorial herniation refers specifically to impaction of the anterior medial temporal gyrus (the uncus) into the tentorial opening just anterior to and adjacent to the midbrain (Fig. 300-1A). The uncus compresses the third nerve as the nerve traverses the subarachnoid space, causing enlargement of the ipsilateral pupil as the first sign (the fibers subserving parasympathetic pupillary function are located peripherally in the nerve). The coma that follows is due to compression of the midbrain (and therefore the RAS) against the opposite tentorial edge by the displaced parahippocampal gyrus (Fig. 300-2). Lateral displacement of the midbrain may compress the opposite cerebral peduncle against the tentorial edge, producing a Babinski sign and hemiparesis contralateral to the hemiparesis that resulted from the mass (the Kernohan-Woltman sign). Herniation may also compress the anterior and posterior cerebral arteries as they pass over the tentorial reflections, with resultant brain infarction. These distortions may also entrap portions of the ventricular system, resulting in hydrocephalus.

FIGURE 300-2

Coronal (*A*) and axial (*B*) magnetic resonance images from a stuporous patient with a left third nerve palsy as a result of a large left-sided subdural hematoma (seen as a gray-white rim). The upper midbrain and lower thalamic regions are compressed and displaced horizontally away from the mass, and there is transtentorial herniation of the medial temporal lobe structures, including the uncus anteriorly. The lateral ventricle opposite to the hematoma has become enlarged as a result of compression of the third ventricle.



A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Central transtentorial herniation denotes a symmetric downward movement of the thalamic structures through the tentorial opening with compression of the upper midbrain **(Fig. 300-1B)**. Miotic pupils and drowsiness are the heralding signs, in contrast to a unilaterally enlarged pupil of the uncal syndrome. Both uncal and central transtentorial herniations cause progressive compression of the brainstem and RAS, with initial damage to the midbrain, then the pons, and finally the medulla. The result is an approximate sequence of neurologic signs that corresponds to each affected level, with respiratory centers in the brainstem often spared until late in the herniation syndrome. Other forms of herniation include transfalcial herniation (displacement of the cingulate gyrus under the falx and across the midline, **Fig. 300-1C**) and foraminal herniation (downward forcing of the cerebellar tonsils into the foramen magnum, **Fig. 300-1D**), which causes early compression of the medulla, respiratory arrest, and death.

A direct relationship between the various configurations of transtentorial herniation and coma is not always found. Drowsiness and stupor can occur with moderate horizontal displacement of the diencephalon (thalamus), before transtentorial herniation is evident. This lateral shift may be quantified on axial images of computed tomography (CT) and magnetic resonance imaging (MRI) scans (Fig. 300-2). In cases of acutely enlarging masses, horizontal displacement of the pineal gland (often calcified in adults) of 3–5 mm is generally associated with drowsiness, 6–8 mm with stupor, and >9 mm with coma. Intrusion of the medial temporal lobe into the tentorial opening is also apparent on MRI and CT scans as obliteration of the cisterna that surrounds the upper brainstem.

Coma Due to Metabolic Disorders and Toxins (including Drug-induced)

Many systemic metabolic abnormalities cause coma by interrupting the delivery of energy substrates (e.g., oxygen, glucose) or by altering neuronal excitability (drugs and alcohol, anesthesia, and epilepsy). These are some of the most common causes of coma in large case series. The metabolic abnormalities that produce coma may, in milder forms, induce an acute confusional state. Thus, in metabolic encephalopathies, clouded consciousness and coma are in a continuum.

Cerebral neurons are fully dependent on cerebral blood flow (CBF) and the delivery of oxygen and glucose. CBF is ~75 mL per 100 g/min in gray matter and 30 mL per 100 g/min in white matter (mean ~55 mL per 100 g/min); oxygen consumption is 3.5 mL per 100 g/min, and glucose utilization is 5 mg per 100 g/min. Brain stores of glucose are able to provide energy for ~2 min after blood flow is interrupted, and oxygen stores last 8–10 s after the cessation of blood flow. Simultaneous hypoxia and ischemia exhaust glucose more rapidly. The electroencephalogram (EEG) rhythm in these circumstances becomes diffusely slowed, typical of metabolic encephalopathies, and as substrate delivery worsens, eventually brain electrical activity ceases.

Unlike hypoxia-ischemia, which causes neuronal destruction, most metabolic disorders such as hypoglycemia, hyponatremia, hyperosmolarity, hypercapnia, hypercalcemia, and hepatic and renal failure cause only minor neuropathologic changes. The reversible effects of these conditions on the brain are not fully understood but may result from impaired energy supplies, changes in ion fluxes across neuronal membranes, and neurotransmitter abnormalities. In hepatic encephalopathy (HE), high ammonia concentrations lead to increased synthesis of glutamine in astrocytes with osmotic swelling, mitochondrial energy failure, production of reactive nitrogen and oxygen species, increases in the inhibitory neurotransmitter GABA, and synthesis of putative “false” neurotransmitters. Other factors, including coexisting inflammation and metabolic abnormalities, also contribute to the coma in some patients. Over time, development of a diffuse astrocytosis is typical of chronic HE. The mechanism of the encephalopathy of renal failure is also multifactorial. Unlike ammonia, urea does not produce central nervous system (CNS) toxicity, and contributors to uremic encephalopathy may include accumulation of neurotoxic substances such as creatinine, guanidine, and related compounds, depletion of catecholamines, altered glutamate and GABA tone, increases in brain calcium, inflammation with disruption of the blood brain barrier, and frequent coexisting vascular disease.

Coma and seizures are common accompaniments of large shifts in sodium and water balance in the brain. These changes in osmolarity arise from systemic medical disorders, including diabetic ketoacidosis, the nonketotic hyperosmolar state, and hyponatremia from any cause (e.g., water intoxication, excessive secretion of antidiuretic hormone, or atrial natriuretic peptides). Sodium levels <125 mmol/L induce confusion, and levels <119 mmol/L are typically associated with coma and convulsions, especially when these levels are achieved quickly. In hyperosmolar coma, the serum osmolarity is generally >350 mosmol/L. Hypercapnia depresses the level of consciousness in proportion to the rise in carbon dioxide (CO_2) in the blood. In all of these metabolic encephalopathies, the degree of neurologic change depends to a large extent on the rapidity with which the serum changes occur. The pathophysiology of other metabolic encephalopathies such as those due to hypercalcemia, hypothyroidism, vitamin B_{12} deficiency, and hypothermia are incompletely understood but must reflect derangements of CNS biochemistry, membrane function, or neurotransmitters.

Coma due to drugs and toxins are typically in large measure reversible and leave no residual damage provided there has not been cardiorespiratory failure. Many drugs and toxins are capable of depressing nervous system function. Some produce coma by affecting both the RAS and the cerebral cortex. The combination of cortical and brainstem signs, which occurs in certain drug overdoses, may lead to an incorrect diagnosis of structural brainstem disease. Overdose of medications that have atropinic actions produces signs such as dilated pupils, tachycardia, and dry skin; opiate overdose produces pinpoint pupils <1 mm in diameter. Some drug intoxications, such as with barbiturates, can mimic all of the signs of brain death, thus toxic etiologies must always be excluded prior to making a diagnosis of brain death.

Epileptic Coma

Generalized electrical seizures are associated with coma, even in the absence of motor convulsions (nonconvulsive status epilepticus). As a result, consideration of EEG monitoring is essential in the workup of coma to exclude this treatable etiology. The self-limited coma that follows a seizure, the postictal state, may be due to exhaustion of energy reserves or effects of locally toxic molecules that are the by-product of seizures. The postictal state produces continuous, generalized slowing of the background EEG activity similar to that of metabolic encephalopathies. It typically lasts for a few minutes, but in some cases can be prolonged for hours or even rarely for days.

Coma Due to Widespread Damage to the Cerebral Hemispheres

This category, comprising a number of unrelated disorders, results from extensive bilateral structural cerebral damage that simulates a metabolic disorder. Hypoxia-ischemia is perhaps the best characterized and one in which it is not possible initially to distinguish the acute reversible effects of oxygen deprivation of the brain from the subsequent effects of anoxic neuronal damage. Similar cerebral damage may be produced by disorders that occlude widespread small blood vessels throughout the brain; examples include cerebral malaria, thrombotic thrombocytopenic purpura, and hyperviscosity. Diffuse white matter damage from cranial trauma or inflammatory demyelinating diseases can cause a similar coma syndrome.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT Coma

A video examination of the comatose patient is shown in Chap. V4. Acute respiratory and cardiovascular problems should be attended to prior to neurologic assessment. In most instances, a complete medical evaluation, except for vital signs, funduscopy, and examination for nuchal rigidity, may be deferred until the neurologic evaluation has established the severity and nature of coma. **The approach to the patient with coma from cranial trauma is discussed in Chap. 435.**

HISTORY

The cause of coma may be immediately evident as in cases of trauma, cardiac arrest, or observed drug ingestion. In the remainder, certain points are useful: (1) the circumstances and rapidity with which neurologic symptoms developed; (2) antecedent symptoms (confusion, weakness, headache, fever, seizures, dizziness, double vision, or vomiting); (3) the use of medications, drugs, or alcohol; and (4) chronic liver, kidney, lung, heart, or other medical disease. Direct interrogation of family, observers, and ambulance technicians on the scene, in person or by telephone, is an important part of the evaluation when possible.

GENERAL PHYSICAL EXAMINATION

Fever suggests a systemic infection, bacterial meningitis, encephalitis, heat stroke, neuroleptic malignant syndrome, malignant hyperthermia due to anesthetics, or anticholinergic drug intoxication. Only rarely is fever attributable to a lesion that has disturbed hypothalamic temperature-regulating centers (“central fever”) and this diagnosis should only be considered after an exhaustive search for other causes fails to reveal an explanation for fever. A slight elevation in temperature may follow vigorous convulsions. Hypothermia is observed with alcohol, barbiturate, sedative, or phenothiazine intoxication; hypoglycemia; peripheral circulatory failure; or extreme hypothyroidism. Hypothermia itself causes coma when the temperature is <31°C (87.8°F) regardless of the underlying etiology. Tachypnea may indicate systemic acidosis or pneumonia. Aberrant respiratory patterns that reflect brainstem disorders are discussed below. Marked hypertension suggests hypertensive encephalopathy, cerebral hemorrhage, large cerebral infarction, or head injury. Hypotension is

characteristic of coma from alcohol or barbiturate intoxication, internal hemorrhage or myocardial infarction causing poor delivery of blood to the brain, sepsis, profound hypothyroidism, or Addisonian crisis. The funduscopic examination can detect increased intracranial pressure (ICP) (papilledema), subarachnoid hemorrhage (subhyaloid hemorrhages), and hypertensive encephalopathy (exudates, hemorrhages, vessel-crossing changes, papilledema). Cutaneous petechiae suggest thrombotic thrombocytopenic purpura, meningococcemia, or a bleeding diathesis associated with an intracerebral hemorrhage. Cyanosis and reddish or anemic skin coloration are other indications of an underlying systemic disease or carbon monoxide as responsible for the coma.

NEUROLOGIC EXAMINATION

The patient should first be observed without intervention by the examiner. Tossing about in the bed, reaching up toward the face, crossing legs, yawning, swallowing, coughing, or moaning reflect a drowsy state that is close to normal awakeness. Lack of restless movements on one side or an outturned leg suggests hemiplegia. Subtle, intermittent twitching movements of a foot, finger, or facial muscle may be the only sign of seizures. Multifocal myoclonus almost always indicates a metabolic disorder, particularly uremia, anoxia, drug intoxication, or rarely a prion disease **(Chap. 430)**. In a drowsy and confused patient, bilateral asterixis is a sign of metabolic encephalopathy or drug intoxication.

Decorticate rigidity and decerebrate rigidity, or “posturing,” describe stereotyped arm and leg movements occurring spontaneously or elicited by sensory stimulation. Flexion of the elbows and wrists and supination of the arm (decorticate posturing) suggests bilateral damage rostral to the midbrain, whereas extension of the elbows and wrists with pronation (decerebrate posturing) indicates damage to motor tracts caudal to the midbrain. These localizations have been adapted from animal work and cannot be applied with precision to coma in humans. In fact, acute and widespread disorders of any type, regardless of location, frequently cause limb extension.

LEVEL OF AROUSAL

A sequence of increasingly intense stimuli is first used to determine the threshold for arousal and the motor response of each side of the body. The results of testing may vary from minute to minute, and serial examinations are useful. Tickling the nostrils with a cotton wisp is a moderate stimulus to arousal—all but deeply stuporous and comatose patients will move the head away and arouse to some degree. An even greater degree of responsiveness is present if the patient uses his hand to remove an offending stimulus. Pressure on the knuckles or bony prominences and pinprick stimulation are humane forms of noxious stimuli; pinching the skin causes unsightly ecchymoses and is generally not necessary but may be useful in eliciting abduction withdrawal movements of the limbs. Posturing in response to noxious stimuli indicates severe damage to the corticospinal system, whereas abduction-avoidance movement of a limb is usually purposeful and denotes an intact corticospinal system. Posturing may also be unilateral and coexist with purposeful limb movements, reflecting incomplete damage to the motor system.

BRAINSTEM REFLEXES

Given that the nuclei of the cranial nerves and the RAS are both located in the brainstem, assessment of brainstem function is essential to localization of the lesion in coma **(Fig. 300-3)**. Patients with preserved brainstem reflexes typically have a bihemispheric localization to coma, including toxic or drug intoxication, whereas patients with abnormal brainstem reflexes either have an RAS localization to their coma or are suffering from a herniation syndrome impacting the brainstem remotely from a cerebral mass lesion. The most important brainstem reflexes that are examined are pupillary size and reaction to light, spontaneous and elicited eye movements, corneal responses, and the respiratory pattern.

Pupillary Signs

Pupillary reactions are examined with a bright, diffuse light. Reactive and round pupils of midsize (2.5–5 mm) essentially exclude upper midbrain damage, either primary or secondary to compression. A response to light may be difficult to appreciate in pupils <2 mm in diameter, and bright room lighting mutes pupillary reactivity. One enlarged and poorly reactive pupil (>6 mm) signifies compression or stretching of the third nerve from the effects of a cerebral mass above. Enlargement of the pupil contralateral to a hemispheral mass may occur but is infrequent. An oval and slightly eccentric pupil is a transitional sign that accompanies early midbrain–third nerve compression. The most extreme pupillary sign, bilaterally dilated and unreactive pupils, indicates severe midbrain damage, usually from compression by a supratentorial mass. Ingestion of drugs with anticholinergic activity, the use of mydriatic eye drops, nebulizer treatments, and direct ocular trauma are among the causes of misleading pupillary enlargement.

Reactive and bilaterally small (1–2.5 mm) but not pinpoint pupils are seen in metabolic encephalopathies or in deep bilateral hemispheral lesions such as hydrocephalus or thalamic hemorrhage. Even smaller reactive pupils (<1 mm) characterize narcotic or barbiturate overdoses but also occur with extensive pontine hemorrhage. The response to naloxone and the presence of reflex eye movements (see below) assist in distinguishing between these. Unilateral miosis in coma has been attributed to dysfunction of sympathetic efferents originating in the posterior hypothalamus and descending in the tegmentum of the brainstem to the cervical cord. It is an occasional finding in patients with a large cerebral hemorrhage that affects the thalamus.

Ocular Movements

The eyes are first observed by elevating the lids and observing the resting position and spontaneous movements of the globes. Horizontal divergence of the eyes at rest is normal in drowsiness. As coma deepens, the ocular axes may become parallel again.

Spontaneous eye movements in coma often take the form of conjugate horizontal roving. This finding alone exonerates extensive damage in the midbrain and pons and has the same significance as normal reflex eye movements (see below). Conjugate horizontal ocular deviation to one side indicates damage to the frontal lobe on the same side or less commonly the pons on the opposite side. This phenomenon is summarized by the following maxim: *The eyes look toward a hemispheral lesion and away from a brainstem lesion*. Seizures involving the frontal lobe drive the eyes to the opposite side, simulating a pontine destructive lesion. The eyes may occasionally turn paradoxically away from the side of a deep hemispheral lesion (“wrong-way eyes”). The eyes turn down and inward with thalamic and upper midbrain lesions, typically thalamic hemorrhage. “Ocular bobbing” describes brisk downward and slow upward movements of the eyes associated with loss of horizontal eye movements and is diagnostic of bilateral pontine damage, usually from thrombosis of the basilar artery. “Ocular dipping” is a slower, arrhythmic downward movement followed by a faster upward movement in patients with normal reflex horizontal gaze; it usually indicates diffuse cortical anoxic damage.

The oculocephalic reflexes, elicited by moving the head from side to side or vertically and observing eye movements in the direction opposite to the head movement, depend on the integrity of the ocular motor nuclei and their interconnecting tracts that extend from the midbrain to the pons and medulla (Fig. 300-3). The movements, called somewhat inappropriately “doll’s eyes,” are normally suppressed in the awake patient with intact frontal lobes. The ability to elicit them therefore reflects both reduced cortical influence on the brainstem and intact brainstem pathways. The opposite, an absence of reflex eye movements, usually signifies damage within the brainstem but can result from overdoses of certain drugs. In this circumstance, normal pupillary size and light reaction distinguishes most drug-induced comas from structural brainstem damage. Oculocephalic reflexes should never be elicited in patients with possible head or neck trauma, as vigorous head movements can precipitate or worsen a spinal cord injury.

Thermal, or “caloric,” stimulation of the vestibular apparatus (oculovestibular response) provides a more intense stimulus for the oculocephalic reflex but provides essentially the same information. The

test is performed by irrigating the external auditory canal with cold water in order to induce convection currents in the labyrinths. After a brief latency, the result is tonic deviation of both eyes to the side of cold-water irrigation. In comatose patients, nystagmus in the opposite direction may not occur. The acronym “COWS” has been used to remind generations of medical students of the direction of nystagmus—cold water opposite, warm water same—but since nystagmus is often absent in the opposite direction due to frontal lobe dysfunction in coma, this mnemonic does not often hold true.

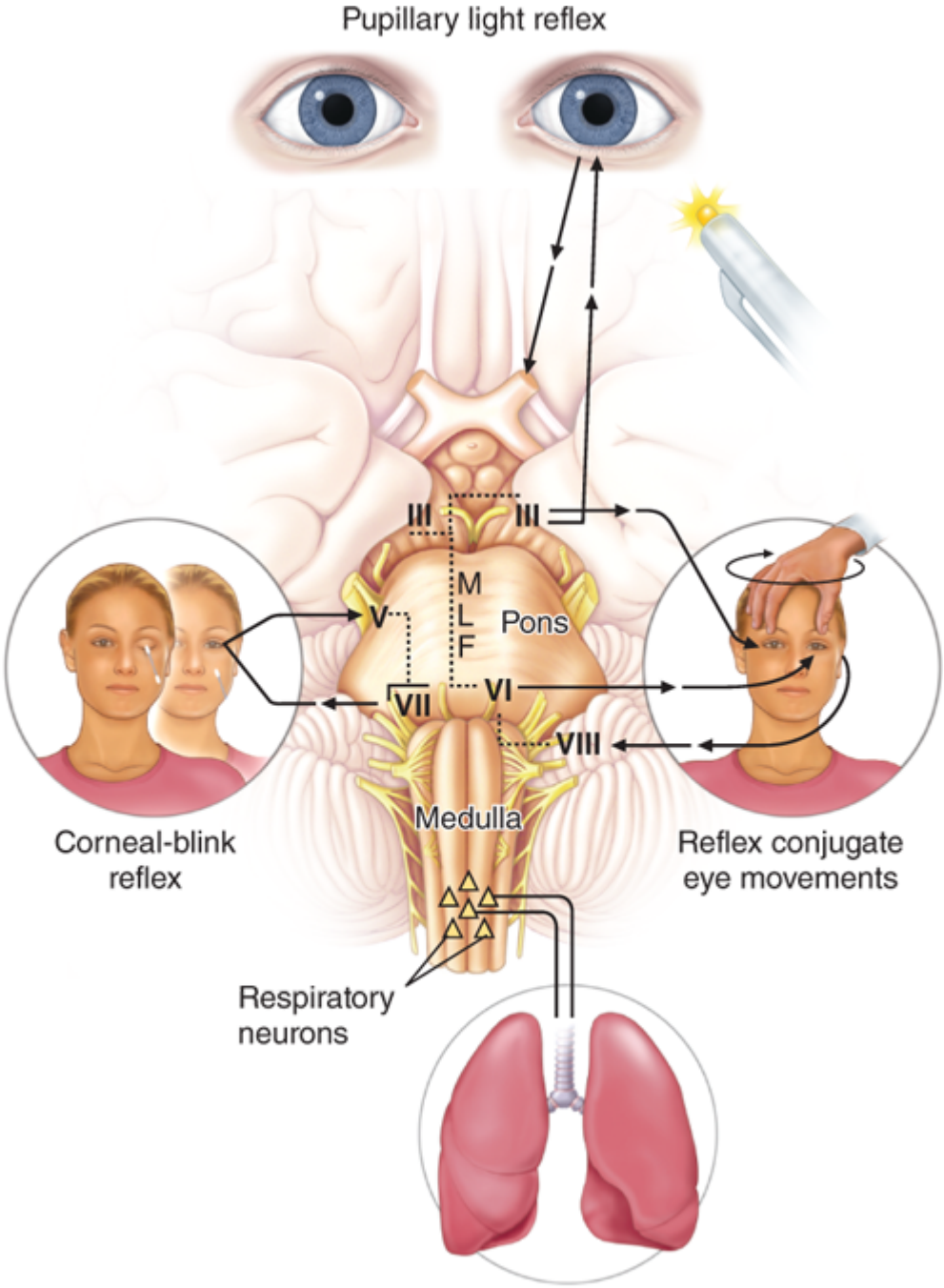
When touching the cornea with a wisp of cotton, a response consisting of brief bilateral lid closure is normally observed. The corneal reflex depends on the integrity of pontine pathways between the fifth (afferent) and both seventh (efferent) cranial nerves; in conjunction with reflex eye movements, it is a useful test of pontine function. CNS-depressant drugs diminish or eliminate the corneal responses soon after reflex eye movements are paralyzed but before the pupils become unreactive to light. The corneal response may be lost for a time on the side of an acute hemiplegia.

Respiratory Patterns

These are of less localizing value in comparison to other brainstem signs. Shallow, slow, but regular breathing suggests metabolic or drug depression. Cheyne-Stokes respiration in its typical cyclic form, ending with a brief apneic period, signifies bihemispheral damage or metabolic suppression and commonly accompanies light coma. Rapid, deep (Kussmaul) breathing usually implies metabolic acidosis but may also occur with pontomesencephalic lesions. Agonal gasps are the result of lower brainstem (medullary) damage and are recognized as the terminal respiratory pattern of severe brain damage. A number of other cyclic breathing variations have been described but are of lesser significance.

FIGURE 300-3

Examination of brainstem reflexes in coma. Midbrain and third nerve function are tested by pupillary reaction to light, pontine function by spontaneous and reflex eye movements and corneal responses, and medullary function by respiratory and pharyngeal responses. Reflex conjugate, horizontal eye movements are dependent on the medial longitudinal fasciculus (MLF) interconnecting the sixth and contralateral third nerve nuclei. Head rotation (oculocephalic reflex) or caloric stimulation of the labyrinths (oculovestibular reflex) elicits contraversive eye movements (for details see text).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

LABORATORY STUDIES AND IMAGING

The studies that are most useful in the diagnosis of coma are chemical-toxicologic analysis of blood and urine, cranial CT or MRI, EEG, and CSF examination. Arterial blood gas analysis is helpful in patients with lung disease and acid-base disorders. The metabolic aberrations commonly encountered in clinical practice are usually revealed by measurement of electrolytes, glucose, calcium, magnesium, osmolarity, and renal (blood urea nitrogen) and hepatic (NH_3) function. Toxicologic analysis may be necessary in any case of acute coma where the diagnosis is not immediately clear. However, the presence of exogenous drugs or toxins, especially alcohol, does not exclude the possibility that other factors, particularly head trauma, are also contributing to the clinical state. An ethanol level of 43 mmol/L (0.2 g/dL) in nonhabituated patients generally causes impaired mental activity; a level of >65

mmol/L (0.3 g/dL) is associated with stupor. The development of tolerance may allow some chronic alcoholics to remain awake at levels >87 mmol/L (0.4 g/dL).

The availability of CT and MRI has focused attention on causes of coma that are detectable by imaging (e.g., hemorrhage, tumor, or hydrocephalus). Resorting primarily to this approach, although at times expedient, is imprudent because most cases of coma (and confusion) are metabolic or toxic in origin. A normal CT scan does not exclude an anatomic lesion as the cause of coma; early bilateral hemisphere infarction, acute brainstem infarction, encephalitis, meningitis, mechanical shearing of axons as a result of closed head trauma, sagittal sinus thrombosis, hypoxic injury and subdural hematoma isodense to adjacent brain are some of the disorders that may not be detected. Sometimes imaging results can be misleading such as when small subdural hematomas or old strokes are found, but the patient's coma is due to intoxication.

The EEG **(Chap. 418)** provides clues in metabolic or drug-induced states but is rarely diagnostic. However, it is the essential test to reveal coma due to nonconvulsive seizures, and shows fairly characteristic patterns in herpesvirus encephalitis and prion (Creutzfeldt-Jakob) disease. The EEG may be further helpful in disclosing generalized slowing of the background activity, a reflection of the severity of an encephalopathy. Predominant high-voltage slowing (δ or triphasic waves) in the frontal regions is typical of metabolic coma, as from hepatic failure, and widespread fast (β) activity implicates sedative drugs (e.g., benzodiazepines). A special pattern of "alpha coma," defined by widespread, variable 8- to 12-Hz activity, superficially resembles the normal α rhythm of waking but, unlike normal α activity, is not altered by environmental stimuli. Alpha coma results from pontine or diffuse cortical damage and is associated with a poor prognosis. Normal α activity on the EEG, which is suppressed by stimulating the patient, also alerts the clinician to the locked-in syndrome or to hysteria or catatonia.

Lumbar puncture should be performed if no cause is readily apparent, as examination of the CSF remains indispensable in the diagnosis of various forms of meningitis and encephalitis. An imaging study should be performed prior to lumbar puncture to exclude a large intracranial mass lesion which could lead to herniation with lumbar puncture. Blood cultures and administration of antibiotics should precede the imaging study if infectious meningitis is suspected **(Chap. 133)**.

DIFFERENTIAL DIAGNOSIS OF COMA

(Table 300-1) The causes of coma can be divided into three broad categories: cases without focal neurologic signs (e.g., metabolic and toxic encephalopathies); cases with prominent focal signs (e.g., stroke, cerebral hemorrhage); and meningitis syndromes, characterized by fever or stiff neck and an excess of cells in the spinal fluid (e.g., bacterial meningitis, subarachnoid hemorrhage, encephalitis). Causes of sudden coma include drug ingestion, cerebral hemorrhage, trauma, cardiac arrest, epilepsy, and basilar artery occlusion. Coma that appears subacutely is usually related to a preexisting medical or neurologic problem or, less often, to secondary brain swelling surrounding a mass such as tumor or cerebral infarction.

TABLE 300-1 Differential Diagnosis of Coma

1. Diseases that cause no focal brainstem or lateralizing neurologic signs (CT scan is often normal)
    a. Intoxications: alcohol, sedative drugs, opiates, etc.
    b. Metabolic disturbances: anoxia, hyponatremia, hypernatremia, hypercalcemia, diabetic acidosis, nonketotic hyperosmolar hyperglycemia, hypoglycemia, uremia, hepatic coma, hypercarbia, Addisonian crisis, hypo- and hyperthyroid states, profound nutritional deficiency
    c. Severe systemic infections: pneumonia, septicemia, typhoid fever, malaria, Waterhouse-Friderichsen syndrome
    d. Shock from any cause
    e. Status epilepticus, nonconvulsive status epilepticus, postictal states
    f. Hyperperfusion syndromes including hypertensive encephalopathy, eclampsia, posterior reversible encephalopathy syndrome (PRES)
    g. Severe hyperthermia, hypothermia
    h. Concussion
    i. Acute hydrocephalus
2. Diseases that cause focal brainstem or lateralizing cerebral signs (CT scan is typically abnormal)
    a. Hemispheral hemorrhage (basal ganglionic, thalamic) or infarction (large middle cerebral artery territory) with secondary brainstem compression
    b. Brainstem infarction due to basilar artery thrombosis or embolism
    c. Brain abscess, subdural empyema
    d. Epidural and subdural hemorrhage, brain contusion
    e. Brain tumor with surrounding edema
    f. Cerebellar and pontine hemorrhage and infarction
    g. Widespread traumatic brain injury
    h. Metabolic coma (see above) in the setting of preexisting focal damage
3. Diseases that cause meningeal irritation with or without fever, and with an excess of WBCs or RBCs in the CSF
    a. Subarachnoid hemorrhage from ruptured aneurysm, arteriovenous malformation, trauma
    b. Infectious meningitis and meningoencephalitis
    c. Paraneoplastic and autoimmune meningitis

d. Carcinomatous and lymphomatous meningitis

The diagnosis of coma due to cerebrovascular disease can be difficult **(Chap. 419)**. The most common diseases are (1) basal ganglia and thalamic hemorrhage (acute but not instantaneous onset, vomiting, headache, hemiplegia, and characteristic eye signs); (2) pontine hemorrhage (sudden onset, pinpoint pupils, loss of reflex eye movements and corneal responses, ocular bobbing, posturing, and hyperventilation); (3) cerebellar hemorrhage (occipital headache, vomiting, gaze paresis, and inability to stand and walk); (4) basilar artery thrombosis (neurologic prodrome or warning spells, diplopia, dysarthria, vomiting, eye movement and corneal response abnormalities, and asymmetric limb paresis); and (5) subarachnoid hemorrhage (precipitous coma after sudden severe headache and vomiting). The most common stroke, infarction in the territory of the middle cerebral artery, does not cause coma, but edema surrounding large infarctions may expand over several days and cause coma from mass effect.

The syndrome of acute hydrocephalus accompanies many intracranial diseases, particularly subarachnoid hemorrhage. It is characterized by headache and sometimes vomiting that may progress quickly to coma with extensor posturing of the limbs, bilateral Babinski signs, small unreactive pupils, and impaired oculocephalic movements in the vertical direction. At times, the coma may be featureless without lateralizing signs, although papilledema is often present.

BRAIN DEATH

This is a state of irreversible cessation of all cerebral and brainstem function with preservation of cardiac activity and maintenance of respiratory and somatic function by artificial means. Brain death is the only type of brain damage recognized as morally, ethically, and legally equivalent to death. Criteria have been advanced for the diagnosis of brain death, and it is essential to adhere to consensus standards as multiple studies have shown variability in local practice. Given the implications of such a diagnosis, clinicians must be thorough and precise in determining brain death. Established criteria are simple, can be assessed at the bedside, and allow no chance of diagnostic error. They contain two essential elements, after assuring that no confounding factors (e.g., hypothermia, drug intoxication) are present: (1) widespread cortical destruction that is reflected by deep coma and unresponsiveness to all forms of stimulation; (2) global brainstem damage demonstrated by absent pupillary light reaction, absent corneal reflexes, loss of oculovestibular reflexes, and destruction of the medulla, manifested by complete and irreversible apnea. Diabetes insipidus is usually present, but may only develop hours or days after the other clinical signs of brain death appear. The pupils are usually midsized but may be enlarged. Loss of deep tendon reflexes is not required because the spinal cord remains functional. Occasionally other reflexes that originate from the spine may be present and should not preclude a diagnosis of brain death.

Demonstration that apnea is due to medullary damage requires that the P_{CO_2} be high enough to stimulate respiration during a test of spontaneous breathing. Apnea testing can be done safely by the use of preoxygenation with 100% oxygen prior to and following removal of the ventilator. CO_2 tension increases ~0.3–0.4 kPa/min (2–3 mmHg/min) during apnea. Apnea is confirmed if no respiratory effort has been observed in the presence of a sufficiently elevated P_{CO_2}. The apnea test is usually stopped if there is serious cardiovascular instability.

An isoelectric EEG may be used as an optional confirmatory test for total cerebral damage. Radionuclide brain scanning, cerebral angiography, or transcranial Doppler measurements may also be included to demonstrate the absence of blood flow when a confirmatory study is desired.

Some period of observation, usually 6–24 h, is recommended, during which the clinical signs of brain death are sustained. It is advisable to delay clinical testing for at least 24 h if a cardiac arrest has

caused brain death or if the inciting disease is not known.

It is largely accepted in Western society that the ventilator can be disconnected from a brain-dead patient and that organ donation is subsequently possible. Good communication between the physician and the family is important with appropriate preparation of the family for brain death testing and diagnosis.

TREATMENT

TREATMENT Coma

The immediate goal in a comatose patient is prevention of further nervous system damage. Hypotension, hypoglycemia, hypercalcemia, hypoxia, hypercapnia, and hyperthermia should be corrected rapidly. An oropharyngeal airway is adequate to keep the pharynx open in a drowsy patient who is breathing normally. Tracheal intubation is indicated if there is apnea, upper airway obstruction, hypoventilation, or emesis, or if the patient is at risk for aspiration. Mechanical ventilation is required if there is hypoventilation or a need to induce hypocapnia in order to lower ICP. IV access is established, and naloxone and dextrose are administered if narcotic overdose or hypoglycemia is a possibility; thiamine is given along with glucose to avoid provoking Wernicke's encephalopathy in malnourished patients. In cases of suspected ischemic stroke including basilar thrombosis with brainstem ischemia, IV tissue plasminogen activator or mechanical embolectomy is often used after cerebral hemorrhage has been excluded and when the patient presents within established time windows for these interventions **(Chap. 420)**. Physostigmine may awaken patients with anticholinergic-type drug overdose but should be used only with careful monitoring; many physicians believe that it should only be used to treat anticholinergic overdose–associated cardiac arrhythmias. The use of benzodiazepine antagonists offers some prospect of improvement after overdose; however, these drugs are not commonly used empirically in part due to their tendency to provoke seizures. Certain other toxic and drug-induced comas have specific treatments such as fomepizole for ethylene glycol ingestion.

Administration of hypotonic intravenous solutions should be monitored carefully in any serious acute brain illness because of the potential for exacerbating brain swelling. Cervical spine injuries must not be overlooked, particularly before attempting intubation or evaluation of oculocephalic responses. Fever and meningismus indicate an urgent need for examination of the CSF to diagnose meningitis. Whenever acute bacterial meningitis is suspected, antibiotics including vancomycin and a third-generation cephalosporin should be administered along with dexamethasone, preferably after obtaining blood cultures **(see Chap. 133)**. **The management of raised ICP is discussed in Chap. 301.**

PROGNOSIS

Some patients, especially children and young adults, may have ominous early clinical findings such as abnormal brainstem reflexes and yet recover; ultra-early prognostication outside of brain death therefore is unwise. Metabolic comas have a far better prognosis than traumatic ones. All systems for estimating prognosis in adults should be taken as approximations, and medical judgments must be tempered by factors such as age, underlying systemic disease, and general medical condition. In an attempt to collect prognostic information from large numbers of patients with head injury, the Glasgow Coma Scale was devised; empirically, it has predictive value in cases of brain trauma **(see Chap. 435)**. For anoxic coma, clinical signs such as the pupillary and motor responses after 1 day, 3 days, and 1 week have been shown to have predictive value; however, some of these prediction rules are less reliable in the setting of therapeutic hypothermia and therefore serial examinations are advised in this setting. The absence of the cortical responses of the somatosensory evoked potentials has also been shown to be a strong indicator of poor outcome following hypoxic injury.

The uniformly poor outcome of persistent vegetative state has already been mentioned, but recent reports of a number of such patients displaying consistent cortical activation on functional MRI in response to salient stimuli have begun to alter the perception of such individuals. In one series, about 10% of vegetative patients (mainly following traumatic brain injury) could activate their frontal or temporal lobes in response to requests by an examiner to imagine certain visuospatial tasks. In one case, a rudimentary form of communication could be established. There are also reports in exceptional patients of improvement in cognitive function with the implantation of thalamic-stimulating electrodes or the use of novel activating agents including zolpidem. It is prudent to avoid generalizations from these findings, but the need for future studies of novel techniques to help communication and possibly recovery is needed.

FURTHER READING

Edlow JA et al: Diagnosis of reversible causes of coma. Lancet 384:2064, 2014. [PubMed: 24767707]
Greer DM et al: Variability of brain death policies in the United States. JAMA Neurol 73:213, 2016. [PubMed: 26719912]
Monti MM et al: Willful modulation of brain activity in disorders of consciousness. N Engl J Med 362:579, 2010. [PubMed: 20130250]
Posner JB et al: *Plum and Posner's Diagnosis of Stupor and Coma*, 4th ed. New York, Oxford University Press, 2007.
Rossetti AO et al: Neurologic prognostication of outcome in patients in coma after cardiac arrest. Lancet Neurol 15:597, 2016. [PubMed: 27017468]

Chapter 301: Severe Acute Encephalopathies and Critical Care Weakness

J. Claude Hemphill, III; Wade S. Smith; S. Andrew Josephson; Daryl R. Gress

INTRODUCTION

Life-threatening neurologic illness may be caused by a primary disorder affecting any region of the neuraxis or may occur as a consequence of a systemic disorder such as hepatic failure, multisystem organ failure, or cardiac arrest (Table 301-1). Neurologic critical care focuses on preservation of neurologic tissue and prevention of secondary brain injury caused by ischemia, hemorrhage, edema, herniation, and elevated intracranial pressure (ICP). Encephalopathy is a general term describing brain dysfunction that is diffuse, global, or multi-focal. Severe acute encephalopathies represent a group of various disorders due to different neurological or systemic etiologies, but that share the common themes of primary and secondary brain injury.

TABLE 301-1

Neurologic Disorders in Critical Illness

Localization Along Neuroaxis	Syndrome
Central Nervous System	
Brain: Cerebral hemispheres	Global encephalopathy
	Delirium
	Sepsis
	Organ failure—hepatic, renal
	Medication related—sedatives, hypnotics, analgesics, H_2 blockers, antihypertensives
	Drug overdose
	Electrolyte disturbance—hyponatremia, hypoglycemia
	Hypotension/hypoperfusion
	Hypoxia
	Meningitis
	Subarachnoid hemorrhage
	Wernicke's disease
	Seizure—postictal or nonconvulsive status epilepticus
	Hypertensive encephalopathy
	Hypothyroidism—myxedema
	Focal deficits
	Ischemic stroke
	Tumor
	Abscess, subdural empyema
	Intraparenchymal hemorrhage
	Subdural/epidural hematoma
Brainstem/cerebellum	Mass effect and compression
	Basilar artery thrombosis
	Intraparenchymal hemorrhage
	Central pontine myelinolysis
Spinal cord	Mass effect and compression

Localization Along Neuroaxis	Syndrome
	Disk herniation
	Epidural hematoma
	Ischemia—hypotension/embolic
	Epidural abscess
	Trauma
	Myelitis
Peripheral Nervous System	
Peripheral nerve	
Axonal	Critical illness polyneuropathy
	Neuromuscular blocking agent complications
	Metabolic disturbances, uremia, hyperglycemia
	Medication effects—chemotherapeutic, antiretroviral
Demyelinating	Guillain-Barré syndrome
	Chronic inflammatory demyelinating polyneuropathy
Neuromuscular junction	Prolonged effect of neuromuscular blockade
	Medication effects—aminoglycosides
	Myasthenia gravis, Lambert-Eaton syndrome, botulism
Muscle	Critical illness myopathy
	Cachectic myopathy
	Acute necrotizing myopathy
	Thick-filament myopathy
	Electrolyte disturbances—hypokalemia/hyperkalemia, hypophosphatemia
	Rhabdomyolysis

PATHOPHYSIOLOGY

Brain Edema

Swelling, or edema, of brain tissue occurs with many types of brain injury. The two principal types of edema are vasogenic and cytotoxic. *Vasogenic edema* refers to the influx of fluid and solutes into the brain through an incompetent blood-brain barrier (BBB). In the normal cerebral vasculature, endothelial tight junctions associated with astrocytes create an impermeable barrier (the BBB), through which access into the brain interstitium is dependent upon specific transport mechanisms. The BBB may be compromised in ischemia, trauma, infection, and metabolic derangements. *Vasogenic edema* results from abnormal permeability of the BBB, and typically develops rapidly following injury. *Cytotoxic edema* results from cellular

swelling, membrane breakdown, and ultimately cell death. Clinically significant brain edema usually represents a combination of vasogenic and cytotoxic components. Edema can lead to increased ICP as well as tissue shifts and brain displacement or herniation from focal processes **(Chap. 300)**. These tissue shifts can cause injury by mechanical distention and compression in addition to the ischemia of impaired perfusion consequent to the elevated ICP.

Ischemic Cascade and Cellular Injury

When delivery of substrates, principally oxygen and glucose, is inadequate to sustain cellular function, a series of interrelated biochemical reactions known as the *ischemic cascade* is initiated **(see Fig. 419-2)**. The release of excitatory amino acids, especially glutamate, leads to influx of calcium and sodium ions, which disrupt cellular homeostasis. An increased intracellular calcium concentration may activate proteases and lipases, which then lead to lipid peroxidation and free radical–mediated cell membrane injury. Cytotoxic edema ensues, and ultimately necrotic cell death and tissue infarction occur. This pathway to irreversible cell death is common to ischemic stroke, global cerebral ischemia, and traumatic brain injury.

Penumbra refers to areas of ischemic brain tissue that have not yet undergone irreversible infarction, implying that these regions are potentially salvageable if ischemia can be reversed. Factors that may exacerbate ischemic brain injury include systemic hypotension and hypoxia, which further reduce substrate delivery to vulnerable brain tissue, and fever, seizures, and hyperglycemia, which can increase cellular metabolism, outstripping compensatory processes. Clinically, these events are known as *secondary brain insults* because they lead to exacerbation of the primary brain injury. Prevention, identification, and treatment of secondary brain insults are fundamental goals of management.

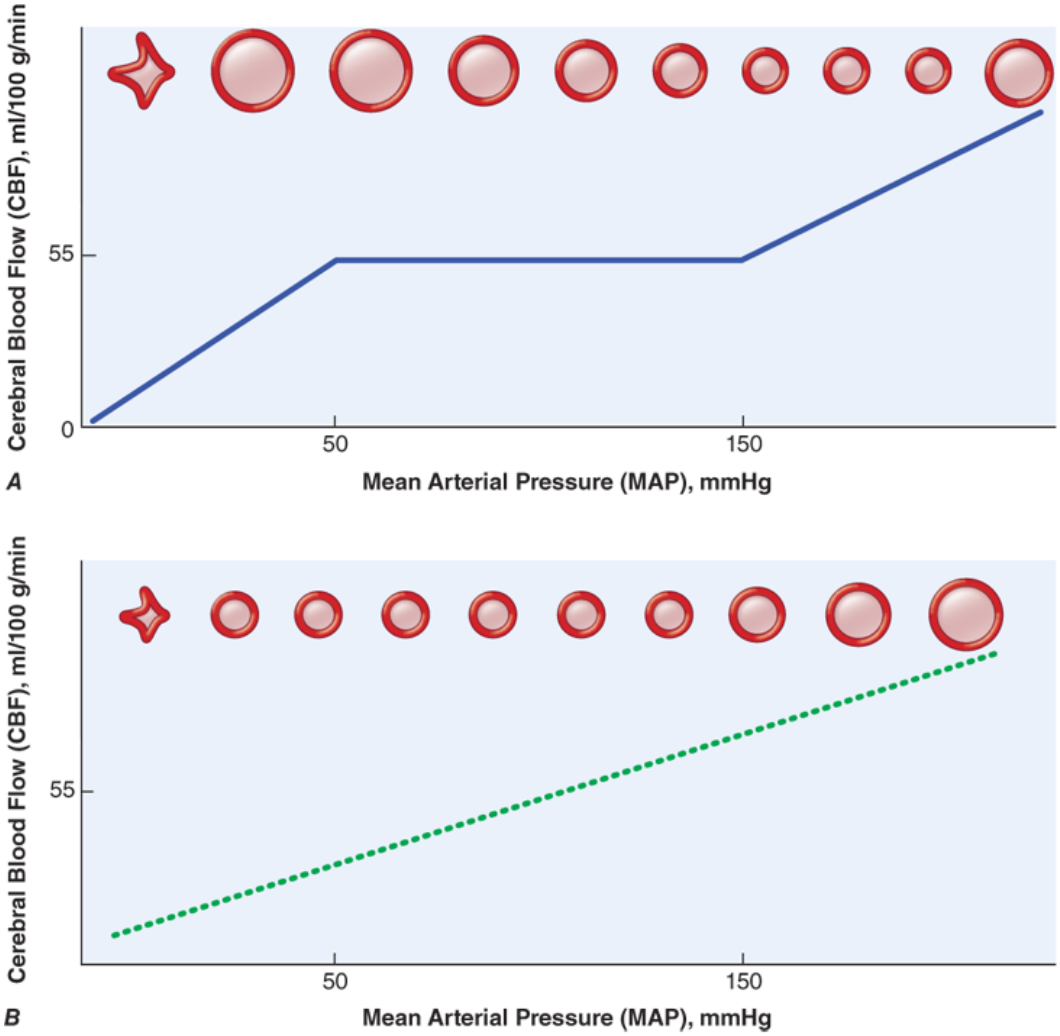
An alternative pathway of cellular injury is *apoptosis*. This process implies programmed cell death, which may occur in the setting of ischemic stroke, global cerebral ischemia, traumatic brain injury, and possibly intracerebral hemorrhage. Apoptotic cell death can be distinguished histologically from the necrotic cell death of ischemia and is mediated through a different set of biochemical pathways; apoptotic cell death occurs without cerebral edema and therefore is often not seen on brain imaging. At present, interventions for prevention and treatment of apoptotic cell death remain less well defined than those for ischemia.

Cerebral Perfusion and Autoregulation

Brain tissue requires constant perfusion in order to ensure adequate delivery of substrate. The hemodynamic response of the brain has the capacity to preserve perfusion across a wide range of systemic blood pressures. Cerebral perfusion pressure (CPP), defined as the mean systemic arterial pressure (MAP) minus the ICP, provides the driving force for circulation across the capillary beds of the brain. *Autoregulation* refers to the physiologic response whereby cerebral blood flow (CBF) is regulated via alterations in cerebrovascular resistance in order to maintain perfusion over wide physiologic changes such as neuronal activation or changes in hemodynamic function. If systemic blood pressure drops, cerebral perfusion is preserved through vasodilation of arterioles in the brain; likewise, arteriolar vasoconstriction occurs at high systemic pressures to prevent hyperperfusion, resulting in fairly constant perfusion across a wide range of systemic blood pressures **(Fig. 301-1)**. At the extreme limits of MAP or CPP (high or low), flow becomes directly related to perfusion pressure. These autoregulatory changes occur in the microcirculation and are mediated by vessels below the resolution of those seen on angiography. CBF is also strongly influenced by pH and $Paco_2$. CBF increases with hypercapnia and acidosis and decreases with hypocapnia and alkalosis because of pH related changes in cerebral vascular resistance. This forms the basis for the use of hyperventilation to lower ICP, and this effect on ICP is mediated through a decrease in both CBF and intracranial blood volume. Cerebral autoregulation is a complex process critical to the normal homeostatic functioning of the brain, and this process may be disordered focally and unpredictably in disease states such as traumatic brain injury and severe focal cerebral ischemia.

FIGURE 301-1

Pressure autoregulation of cerebral blood flow. In the normal state where autoregulation is intact ***A***, cerebral perfusion is constant over a wide range of systemic blood pressures (BP). This is mediated by dilation and constriction of small cerebral arterioles (round circles). Below the blood pressure threshold for maximal dilation, cerebral blood flow becomes pressure-dependent and decreases whereas above the threshold for maximum constriction cerebral blood flow increases with increasing systemic blood pressure. In severe brain injury, autoregulatory mechanisms may be impaired and cerebral blood flow becomes pressure-dependent throughout (**B**). At the extremes of BP there may be vascular collapse (very low BP) or forced vasodilation (very high BP).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Cerebrospinal Fluid (CSF) and ICP

The cranial contents consist essentially of brain, CSF, and blood. CSF is produced principally in the choroid plexus of each lateral ventricle, exits the brain via the foramens of Luschka and Magendie, and flows over the cortex to be absorbed into the venous system along the superior sagittal sinus. In adults, ~150 mL of CSF are contained within the ventricles and surrounding the brain and spinal cord; the cerebral blood volume is also ~150 mL. The bony skull offers excellent protection for the brain but allows little tolerance for additional volume. Significant increases in volume eventually result in increased ICP. Obstruction of CSF outflow, edema of cerebral tissue, or increases in volume from tumor or hematoma may increase ICP. Elevated ICP diminishes cerebral perfusion and can lead to tissue ischemia. Ischemia in turn may lead to vasodilation via autoregulatory mechanisms designed to restore cerebral perfusion. However, vasodilation also increases cerebral blood volume, which in turn then increases ICP, lowers CPP, and provokes further ischemia. This vicious cycle is commonly seen in traumatic brain injury, massive intracerebral hemorrhage, and large hemispheric infarcts with significant tissue shifts.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Severe Brain Dysfunction

Critically ill patients with severe central nervous system (CNS) dysfunction require rapid evaluation and intervention in order to limit primary and secondary brain injury. Initial neurologic evaluation should be performed concurrent with stabilization of basic respiratory, cardiac, and hemodynamic parameters. Significant barriers may exist to neurologic assessment in the critical care unit, including endotracheal intubation and the use of sedative or paralytic agents to facilitate procedures.

An impaired level of consciousness is common in critically ill patients. The essential first task in assessment is to determine whether the cause of dysfunction is related to a diffuse, usually metabolic, process or whether a focal, usually structural, process is implicated. Examples of diffuse processes

include metabolic encephalopathies related to organ failure, drug overdose, or hypoxia-ischemia. Focal processes include ischemic and hemorrhagic stroke and traumatic brain injury, especially with intracranial hematomas. Because these two categories of disorders have fundamentally different causes, treatments, and prognoses, the initial focus is on making this distinction rapidly and accurately. **The approach to the comatose patient is discussed in Chap. 300; etiologies are listed in Table 300-1.**

Minor focal deficits may be present on the neurologic examination in patients with metabolic encephalopathies. However, the finding of prominent focal signs such as pupillary asymmetry, hemiparesis, gaze palsy, or visual field deficit should suggest the possibility of a structural lesion. All patients with a decreased level of consciousness associated with focal findings should undergo an urgent neuroimaging procedure, as should all patients with coma of unknown etiology. Computed tomography (CT) scanning is usually the most appropriate initial study because it can be performed quickly in critically ill patients and demonstrates hemorrhage, hydrocephalus, and intracranial tissue shifts well. Magnetic resonance imaging (MRI) may provide more specific information in some situations, such as acute ischemic stroke (diffusion-weighted imaging [DWI]). Any suggestion of trauma from the history or examination should alert the examiner to the possibility of cervical spine injury and prompt an imaging evaluation using CT or MRI. Neurovascular imaging using CT or MRI angiography or venography is increasingly available and may suggest arterial occlusion or cerebral venous thrombosis.

Acute brainstem ischemia due to basilar artery thrombosis may cause brief episodes of spontaneous extensor posturing superficially resembling generalized seizures. Coma of sudden onset, accompanied by these movements and cranial nerve abnormalities, necessitates emergency imaging. A noncontrast CT scan of the brain may reveal a hyperdense basilar artery indicating thrombus in the vessel, and subsequent CT or MR angiography can assess basilar artery patency.

Other diagnostic studies are best used in specific circumstances, usually when neuroimaging studies fail to reveal a structural lesion and the etiology of the altered mental state remains uncertain. Electroencephalography (EEG) can be important in the evaluation of critically ill patients with severe brain dysfunction. The EEG of metabolic encephalopathy typically reveals generalized slowing. One of the most important uses of EEG is to help exclude inapparent seizures, especially nonconvulsive status epilepticus. Untreated continuous or frequently recurrent seizures may cause neuronal injury, making the diagnosis and treatment of seizures crucial in this patient group. Lumbar puncture (LP) may be necessary to exclude infectious or inflammatory processes, and an elevated opening pressure may be an important clue to cerebral venous sinus thrombosis. In patients with coma or profound encephalopathy, it is preferable to perform a neuroimaging study prior to LP. If bacterial meningitis is suspected, an LP may be performed first or antibiotics may be empirically administered before the diagnostic studies are completed. Standard laboratory evaluation of critically ill patients should include assessment of serum electrolytes (especially sodium and calcium), glucose, renal and hepatic function, complete blood count, and coagulation. Serum or urine toxicology screens should be performed in patients with encephalopathy of unknown cause. EEG and LP are most useful when the mechanism of the altered level of consciousness is uncertain; they are not routinely performed for diagnosis in clear-cut cases of stroke or traumatic brain injury.

Monitoring of ICP can be an important tool in selected patients. In general, patients who should be considered for ICP monitoring are those with primary neurologic disorders, such as stroke or traumatic brain injury, who are at significant risk for secondary brain injury due to elevated ICP and decreased CPP. Included are patients with the following: severe traumatic brain injury (Glasgow Coma Scale [GCS] score ≤8 [Table 435-2]); large tissue shifts from supratentorial ischemic or hemorrhagic stroke; or hydrocephalus from subarachnoid hemorrhage (SAH), intraventricular hemorrhage, or posterior fossa stroke. An additional disorder in which ICP monitoring can add important information is fulminant hepatic failure, in which elevated ICP may be treated with barbiturates or, eventually, liver transplantation. In general, ventriculostomy is preferable to ICP monitoring devices that are placed in the brain parenchyma, because ventriculostomy allows CSF drainage as a method of treating elevated ICP. However, parenchymal ICP monitoring is most appropriate for patients with diffuse edema and small ventricles (which may make ventriculostomy placement more difficult) or any degree of coagulopathy (in which ventriculostomy carries a higher risk of hemorrhagic complications) (Fig 301-2).

TREATMENT OF ELEVATED ICP

Elevated ICP may occur in a wide range of disorders, including head trauma, intracerebral hemorrhage, SAH with hydrocephalus, and fulminant hepatic failure. Because CSF and blood volume can be redistributed initially, by the time elevated ICP occurs, intracranial compliance is severely impaired. At this point, any small increase in the volume of CSF, intravascular blood, edema, or a mass lesion may result in a significant increase in ICP and a decrease in cerebral perfusion. This is a fundamental mechanism of secondary ischemic brain injury and constitutes an emergency that requires immediate attention. In general, ICP should be maintained at <20 mmHg and CPP should be maintained at ≥60 mmHg.

Interventions to lower ICP are ideally based on the underlying mechanism responsible for the elevated ICP (Table 301-2). For example, in hydrocephalus from SAH, the principal cause of elevated ICP is impairment of CSF drainage. In this setting, ventricular drainage of CSF is likely to be sufficient and most appropriate. In head trauma and stroke, cytotoxic edema may be most responsible, and the use of osmotic agents such as mannitol or hypertonic saline becomes an appropriate early step. As described above, elevated ICP may cause tissue ischemia, and, if cerebral autoregulation is intact, the resulting vasodilation can lead to a cycle of worsening ischemia. Paradoxically, administration of vasopressor agents to increase mean arterial pressure may actually lower ICP by improving perfusion, thereby allowing autoregulatory vasoconstriction as ischemia is relieved and ultimately decreasing intracranial blood volume.

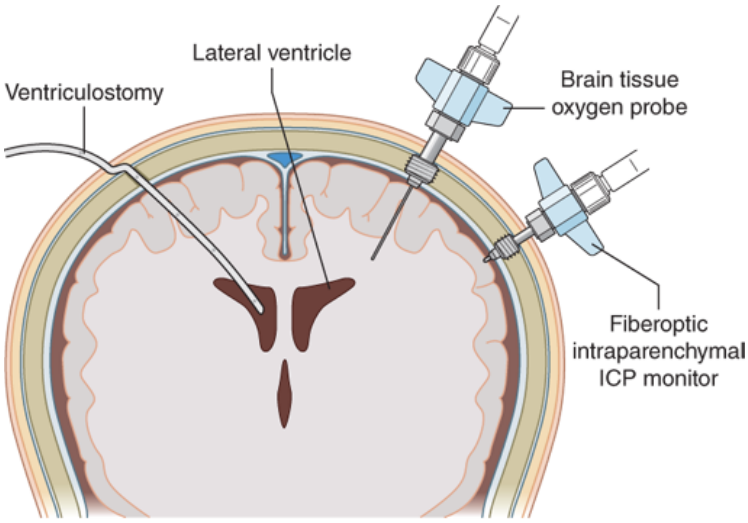
Early signs of elevated ICP include drowsiness and a diminished level of consciousness. Neuroimaging studies may reveal evidence of edema and mass effect. Hypotonic IV fluids should be avoided, and elevation of the head of the bed is recommended. Patients must be carefully observed for risk of aspiration and compromise of the airway as the level of alertness declines. Coma and unilateral pupillary changes are late signs and require immediate intervention. Emergent treatment of elevated ICP is most quickly achieved by intubation and hyperventilation, which causes vasoconstriction and reduces cerebral blood volume. To avoid provoking or worsening cerebral ischemia, hyperventilation, if used at all, is best administered only for short periods of time until a more definitive treatment can be instituted. Furthermore, the effects of hyperventilation on ICP are short-lived, often lasting only for several hours because of the buffering capacity of the cerebral interstitium, and rebound elevations of ICP may accompany abrupt discontinuation of hyperventilation. As the level of consciousness declines to coma, the ability to follow the neurologic status of the patient by examination lessens and measurement of ICP assumes greater importance. If a ventriculostomy device is in place, direct drainage of CSF to reduce ICP is possible. Finally, high-dose barbiturates, decompressive hemicraniectomy, and hypothermia are sometimes used for refractory elevations of ICP, although these have significant side effects and only decompressive hemicraniectomy has been shown to improve outcome in select patients.

SECONDARY BRAIN INSULTS

Patients with primary brain injuries, whether due to trauma or stroke, are at risk for ongoing secondary ischemic brain injury. Because secondary brain injury can be a major determinant of a poor outcome, strategies for minimizing secondary brain insults are an integral part of the critical care of all patients. Although elevated ICP may lead to secondary ischemia, most secondary brain injury is mediated through other clinical events that exacerbate the ischemic cascade already initiated by the primary brain injury. Episodes of secondary brain insults are usually not associated with apparent neurologic worsening. Rather, they lead to cumulative injury limiting eventual recovery, which manifests as a higher mortality rate or worsened long-term functional outcome. Thus, close monitoring of vital signs is important, as is early intervention to prevent secondary ischemia. Avoiding hypotension and hypoxia is critical, as significant hypotensive events (systolic blood pressure <90 mmHg) as short as 10 min in duration have been shown to adversely influence outcome after traumatic brain injury. Even in patients with stroke or head trauma who do not require ICP monitoring, close attention to adequate cerebral perfusion is warranted. Hypoxia (pulse oximetry saturation <90%), particularly in combination with hypotension, also leads to secondary brain injury. Likewise, fever and hyperglycemia both worsen experimental ischemia and have been associated with worsened clinical outcome after stroke and head trauma. Aggressive control of fever with a goal of normothermia is warranted but may be difficult to achieve with antipyretic medications and cooling blankets. The value of newer surface or intravascular temperature control devices for the management of refractory fever is under investigation. The use of IV insulin infusion is encouraged for control of hyperglycemia because this allows better regulation of serum glucose levels than SC insulin. A reasonable goal is to maintain the serum glucose level at <10.0 mmol/L (<180 mg/dL), although episodes of hypoglycemia appear equally detrimental and the optimal targets remain uncertain. New cerebral monitoring tools that allow continuous evaluation of brain tissue oxygen tension, CBF, and metabolism (via microdialysis) may further improve the management of secondary brain injury.

FIGURE 301-2

Intracranial pressure (ICP) and brain tissue oxygen monitoring. A ventriculostomy allows for drainage of cerebrospinal fluid to treat elevated ICP. Fiberoptic ICP and brain tissue oxygen monitors are usually secured using a screw-like skull bolt. Cerebral blood flow and microdialysis probes (not shown) may be placed in a manner similar to the brain tissue oxygen probe.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 301-2

Stepwise Approach to Treatment of Elevated Intracranial Pressure (ICP)[a]

Insert ICP monitor—ventriculostomy versus parenchymal device

General goals: maintain ICP <20 mmHg and CPP ≥60 mmHg. For ICP >20–25 mmHg for >5 min:

1. Elevate head of the bed; midline head position
2. Drain CSF via ventriculostomy (if in place)
3. Osmotherapy—mannitol 25–100 g q4h as needed (maintain serum osmolality <320 mosmol) or hypertonic saline (30 mL, 23.4% NaCl bolus)
4. Glucocorticoids—dexamethasone 4 mg q6h for vasogenic edema from tumor, abscess (avoid glucocorticoids in head trauma, ischemic and hemorrhagic stroke)
5. Sedation (e.g., morphine, propofol, or midazolam); add neuromuscular paralysis if necessary (patient will require endotracheal intubation and mechanical ventilation at this point, if not before)
6. Hyperventilation—to $Paco_2$ 30–35 mmHg (short-term use or skip this step)
7. Pressor therapy—phenylephrine, dopamine, or norepinephrine to maintain adequate MAP to ensure CPP ≥60 mmHg (maintain euvolemia to minimize deleterious systemic effects of pressors). May adjust target CPP in individual patients based on autoregulation status.
8. Consider second-tier therapies for refractory elevated ICP
    a. Decompressive craniectomy
    b. High-dose barbiturate therapy (“pentobarb coma”)
    c. Hypothermia to 33°C

[a]Throughout ICP treatment algorithm, consider repeat head computed tomography to identify mass lesions amenable to surgical evacuation. May alter order of steps based on directed treatment to specific cause of elevated ICP.

Abbreviations: CPP, cerebral perfusion pressure; CSF, cerebrospinal fluid; MAP, mean arterial pressure; $PaCO_2$, arterial partial pressure of carbon dioxide.

CRITICAL CARE DISORDERS OF THE CNS

HYPOXIC-ISCHEMIC ENCEPHALOPATHY

This occurs from lack of delivery of oxygen to the brain because of extreme hypotension (hypoxia-ischemia) or hypoxia due to respiratory failure. Causes include myocardial infarction, cardiac arrest, shock, asphyxiation, paralysis of respiration, and carbon monoxide or cyanide poisoning. In some circumstances, hypoxia may predominate. Carbon monoxide and cyanide poisoning are sometimes termed *histotoxic hypoxia* because they cause a direct impairment of the respiratory chain.

Clinical Manifestations

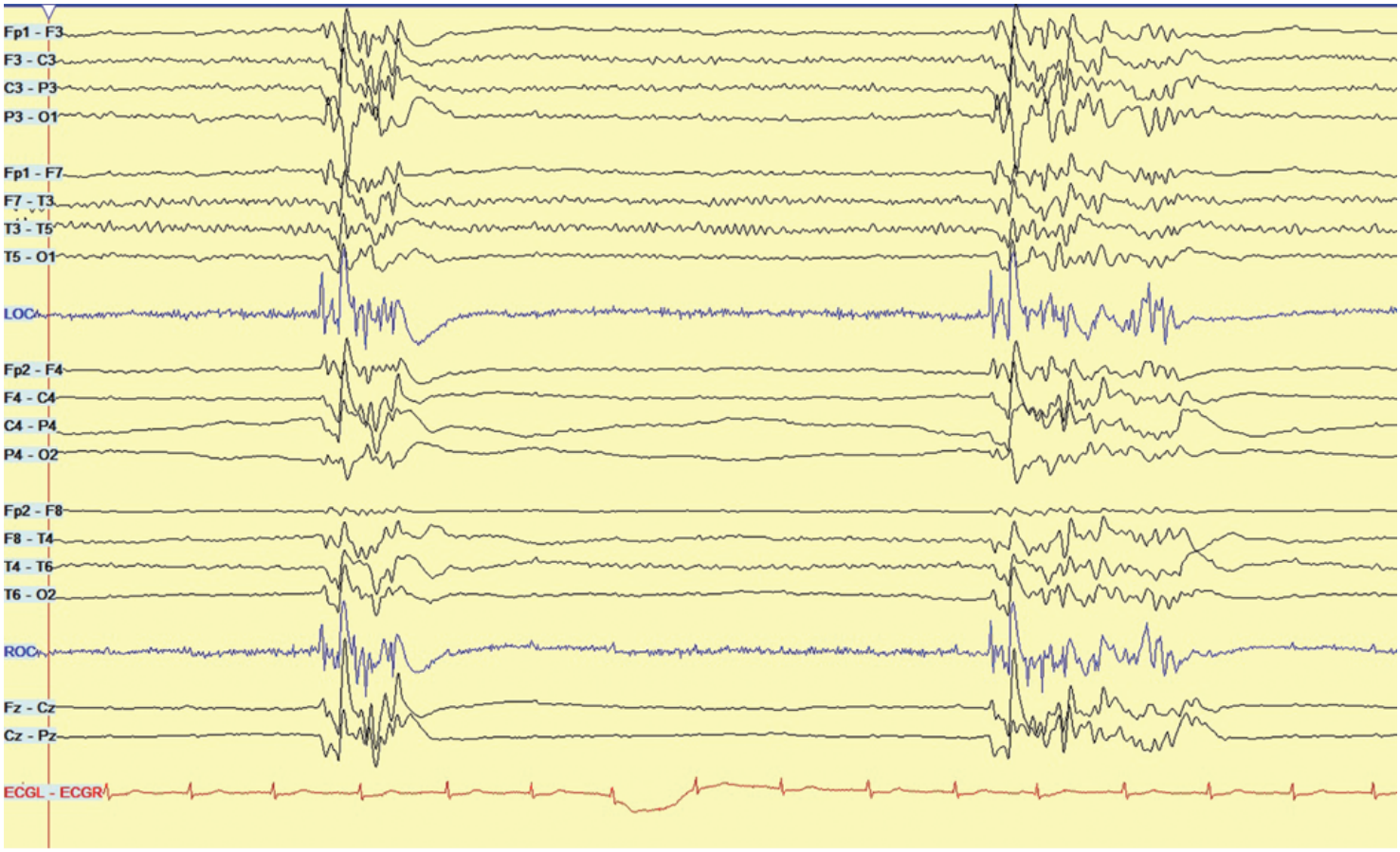
Mild degrees of pure hypoxia, such as occur at high altitudes, cause impaired judgment, inattentiveness, motor incoordination, and, at times, euphoria. However, with hypoxia-ischemia, such as occurs with circulatory arrest, consciousness is lost within seconds. If circulation is restored within 3–5 min, full recovery may occur, but if hypoxia-ischemia lasts beyond 3–5 min, some degree of permanent cerebral damage often results. Except in extreme cases, it may be difficult to judge the precise degree of hypoxia-ischemia, and some patients make a relatively full recovery after even 8–10 min of global cerebral ischemia. The brain is more tolerant to pure hypoxia than it is to hypoxia-ischemia. For example, a Pao_2 as low as 20 mmHg (2.7 kPa) can be well tolerated if it develops gradually, and normal blood pressure is maintained, whereas short durations of very low or absent cerebral circulation may result in permanent impairment.

Clinical examination at different time points after a hypoxic-ischemic insult (especially cardiac arrest) is useful in assessing prognosis for long-term neurologic outcome. The prognosis is better for patients with intact brainstem function, as indicated by normal pupillary light responses and intact oculocephalic (doll’s eyes), oculovestibular (caloric), and corneal reflexes. Absence of these reflexes and the presence of persistently dilated pupils that do not react to light are grave prognostic signs. A low likelihood of a favorable outcome from hypoxic-ischemic coma is strongly suggested by an absent pupillary light reflex or extensor or absent motor response to pain on day 3 following the injury, excluding patients with metabolic disturbances and those treated with high-dose barbiturates or hypothermia, which confound interpretation of these signs. Electrophysiologically, the bilateral absence of the N20 component of the somatosensory evoked potential (SSEP) in the first several days also conveys a poor prognosis. Also, the presence of a burst-suppression pattern of myoclonic status epilepticus on EEG **(Fig 301-3)** or a nonreactive EEG is associated with a low likelihood of good functional outcome. A very elevated serum level (>33 µg/L) of the biochemical marker neuron-specific enolase (NSE) within the first 3 days is indicative of brain damage after resuscitation from cardiac arrest and predicts a poor outcome. Current approaches to prognostication after cardiac arrest encourage the

use of a multimodal approach that includes these diagnostic tests, along with CT or MRI neuroimaging, in conjunction with clinical neurological assessment. Recent studies suggest that the administration of mild hypothermia after cardiac arrest (see "Treatment") may affect the time points when these clinical and electrophysiologic predictors become reliable in identifying patients with a very low likelihood of clinically meaningful recovery. For example, the false-positive rate for incorrect prediction of poor neurologic outcome may be as high as 21% (95% confidence interval [CI] 8–43%) for patients treated with mild hypothermia who exhibit 3-day motor function no better than extensor posturing. Thus, sufficient time from injury is important to ensure accuracy of prognostic assessment. Long-term consequences of hypoxic-ischemic encephalopathy include persistent coma or a vegetative state **(Chap. 300)**, dementia **(Chap. 25)**, visual agnosia **(Chap. 26)**, parkinsonism, choreoathetosis, cerebellar ataxia, myoclonus, seizures, and an amnestic state, which may be a consequence of selective damage to the hippocampus.

FIGURE 301-3

Electroencephalography (EEG) after cardiac arrest. A burst-suppression pattern is seen in a comatose patient with severe hypoxic-ischemic encephalopathy after cardiac arrest. In this patient, each burst on EEG was associated with a whole-body jerking movement leading to the clinical and electrophysiological diagnosis of myoclonic status epilepticus.



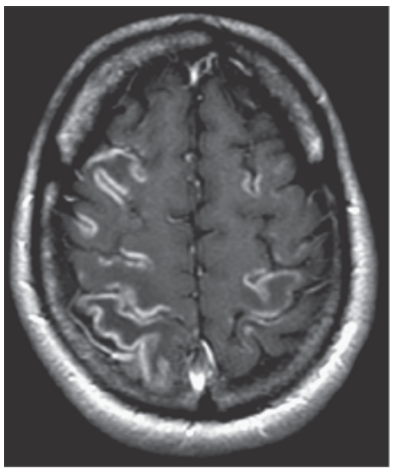
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pathology

Principal histologic findings are extensive multifocal or diffuse laminar cortical necrosis **(Fig. 301-4)**, with frequent involvement of the hippocampus. The hippocampal CA1 neurons are vulnerable to even brief episodes of hypoxia-ischemia, perhaps explaining why selective persistent memory deficits may occur after brief cardiac arrest. Scattered small areas of infarction or neuronal loss may be present in the basal ganglia, hypothalamus, or brainstem. In some cases, extensive bilateral thalamic scarring may affect pathways that mediate arousal, and this pathology may be responsible for the persistent vegetative state. A specific form of hypoxic-ischemic encephalopathy, so-called watershed infarcts, occurs at the distal territories between the major cerebral arteries and can cause cognitive deficits, including visual agnosia, and weakness that is greater in proximal than in distal muscle groups.

FIGURE 301-4

Cortical laminar necrosis in hypoxic-ischemic encephalopathy. T1-weighted postcontrast magnetic resonance imaging shows cortical enhancement in a watershed distribution consistent with laminar necrosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Diagnosis

Diagnosis is based on the history of a hypoxic-ischemic event such as cardiac arrest. Blood pressure <70 mmHg systolic or Pa_{O_2} <40 mmHg is usually necessary, although both absolute levels and duration of exposure are important determinants of cellular injury. Carbon monoxide intoxication can be confirmed by measurement of carboxyhemoglobin and is suggested by a cherry red color of the venous blood and skin, although the latter is an inconsistent clinical finding.

TREATMENT

TREATMENT

Hypoxic-Ischemic Encephalopathy

Treatment should be directed at restoration of normal cardiorespiratory function. This includes securing a clear airway, ensuring adequate oxygenation and ventilation, and restoring cerebral perfusion, whether by cardiopulmonary resuscitation, fluid, pressors, or cardiac pacing. Hypothermia may target the neuronal cell injury cascade and has substantial neuroprotective properties in experimental models of brain injury. In two trials, mild hypothermia (33°C) improved functional outcome in patients who remained comatose after resuscitation from a cardiac arrest. Treatment was initiated within minutes of cardiac resuscitation and continued for 12 h in one study and 24 h in the other. In a more recent study, targeted temperature management (TTM) to 33 or 36°C resulted in similar outcomes. Potential complications of hypothermia include coagulopathy and an increased risk of infection. Current guidelines recommend TTM for cardiac arrest patients who have no meaningful response to verbal commands after return of spontaneous circulation, with temperature maintained constant between 32 and 36°C for at least 24 h.

Severe carbon monoxide intoxication may be treated with hyperbaric oxygen. Anticonvulsants may be needed to control seizures, although these are not usually given prophylactically. Posthypoxic myoclonus may respond to oral administration of clonazepam at doses of 1.5–10 mg daily or valproate at doses of 300–1200 mg daily in divided doses. Myoclonic status epilepticus within 24 h after a primary circulatory arrest generally portends a very poor prognosis, even if seizures are controlled.

Carbon monoxide and cyanide intoxication can also cause a delayed encephalopathy. Little clinical impairment is evident when the patient first regains consciousness, but a parkinsonian syndrome characterized by akinesia and rigidity without tremor may develop. Symptoms can worsen over months, accompanied by increasing evidence of damage in the basal ganglia as seen on both CT and MRI.

POSTCARDIAC BYPASS BRAIN INJURY

CNS injuries following open heart or coronary artery bypass grafting (CABG) surgery are common and include acute encephalopathy, stroke, and a chronic syndrome of cognitive impairment. Hypoperfusion and embolic disease are frequently involved in the pathogenesis of these syndromes, although multiple mechanisms may be involved in these critically ill patients who are at risk for various metabolic and polypharmaceutical complications.

The frequency of hypoxic injury secondary to inadequate blood flow intraoperatively has been markedly decreased by the use of modern surgical and anesthetic techniques. Despite these advances, some patients still experience neurologic complications from cerebral hypoperfusion or may suffer focal ischemia from carotid or focal intracranial stenoses in the setting of regional hypoperfusion. Postoperative infarcts in the border zones between vascular territories commonly are blamed on systemic hypotension, although these infarcts can also result from embolic disease.

Embolic disease is likely the predominant mechanism of cerebral injury during cardiac surgery as evidenced by diffusion-weighted MRI and intraoperative transcranial Doppler ultrasound studies. Thrombus in the heart itself as well as atheromas in the aortic arch can become dislodged during cardiac surgeries, releasing a shower of particulate matter into the cerebral circulation. Cross-clamping of the aorta, manipulation of the heart, extracorporeal circulation techniques ("bypass"), arrhythmias such as atrial fibrillation, and introduction of air through suctioning have all been implicated as potential sources of emboli.

This shower of microemboli results in a number of clinical syndromes. Occasionally, a single large embolus leads to an isolated large-vessel stroke that presents with obvious clinical focal deficits. When there is a high burden of very small emboli, an acute encephalopathy can occur postoperatively, presenting as either a hyperactive or hypoactive confusional state, the latter of which is frequently and incorrectly ascribed to depression or a sedative-induced delirium. When the burden of microemboli is lower, no acute syndrome is recognized, but the patient may suffer a chronic cognitive deficit.

METABOLIC ENCEPHALOPATHIES

Altered mental states, variously described as confusion, delirium, disorientation, and encephalopathy, are present in many patients with severe illness in an intensive care unit (ICU). Older patients are particularly vulnerable to delirium, a confusional state characterized by disordered perception, frequent hallucinations, delusions, and sleep disturbance. This is often attributed to medication effects, sleep deprivation, pain, and anxiety. The presence of delirium is associated with worsened outcome in critically ill patients, even in those without an identifiable CNS pathology such as stroke or brain trauma. In these patients, the cause of delirium is often multifactorial, resulting from organ dysfunction, sepsis, and especially the use of medications given to treat pain, agitation, or anxiety. Critically ill patients are often treated with a variety of sedative and analgesic medications, including opiates, benzodiazepines, neuroleptics, and sedative-anesthetic medications, such as propofol. In critically ill patients requiring sedation, use of the centrally acting α_2 agonist dexmedetomidine may reduce delirium and shorten the duration of mechanical ventilation compared to the use of benzodiazepines such as lorazepam or midazolam. The presence of family members in the ICU may also help to calm and orient agitated patients, and in severe cases, low doses of neuroleptics (e.g., haloperidol 0.5–1 mg) can be useful. Current strategies focus on limiting the use of sedative medications when this can be done safely.

In the ICU setting, several metabolic causes of an altered level of consciousness predominate. Hypercarbic encephalopathy can present with headache, confusion, stupor, or coma. Hypoventilation syndrome occurs most frequently in patients with a history of chronic CO_2 retention who are receiving oxygen therapy for emphysema or chronic pulmonary disease **(Chap. 290)**. The elevated $Paco_2$ leading to CO_2 narcosis may have a direct anesthetic effect, and cerebral vasodilation from increased $Paco_2$ can lead to increased ICP. Hepatic encephalopathy is suggested by asterixis and can occur in chronic liver failure or acute fulminant hepatic failure. Both hyperglycemia and hypoglycemia can cause encephalopathy, as can hypernatremia and hyponatremia. Confusion, impairment of eye movements, and gait ataxia are the hallmarks of acute Wernicke's disease (see below).

SEPSIS-ASSOCIATED ENCEPHALOPATHY

Pathogenesis

In patients with sepsis, the systemic response to infectious agents leads to the release of circulating inflammatory mediators that appear to contribute to encephalopathy. Critical illness, in association with the systemic inflammatory response syndrome (SIRS), can lead to multisystem organ failure. This syndrome can occur in the setting of apparent sepsis, severe burns, or trauma, even without clear identification of an infectious agent. Many patients with critical illness, sepsis, or SIRS develop encephalopathy without obvious explanation. This condition is broadly termed *sepsis-associated encephalopathy*. Although the specific mediators leading to neurologic dysfunction remain uncertain, it is clear that the encephalopathy is not simply the result of metabolic derangements of multiorgan failure. The cytokines tumor necrosis factor, interleukin (IL)-1, IL-2, and IL-6 are thought to play a role in this syndrome.

Diagnosis

Sepsis-associated encephalopathy presents clinically as a diffuse dysfunction of the brain without prominent focal findings. Confusion, disorientation, agitation, and fluctuations in level of alertness are typical. In more profound cases, especially with hemodynamic compromise, the decrease in level of alertness can be more prominent, at times resulting in coma. Hyperreflexia and frontal release signs such as a grasp or snout reflex **(Chap. 26)** can be seen. Abnormal movements such as myoclonus, tremor, or asterixis can occur. Sepsis-associated encephalopathy is quite common, occurring in the majority of patients with sepsis and multisystem organ failure. Diagnosis is often difficult because of the multiple potential causes of neurologic dysfunction in critically ill patients and requires exclusion of structural, metabolic, toxic, and infectious (e.g., meningitis or encephalitis) causes. The mortality rate of patients with sepsis-associated encephalopathy severe enough to produce coma approaches 50%, although this principally reflects the severity of the underlying critical illness and is not a direct result of the encephalopathy. Patients dying from severe sepsis or septic shock may have elevated levels of the serum brain injury biomarker S-100β and neuropathologic findings of neuronal apoptosis and cerebral ischemic injury. Successful treatment of the underlying critical illness almost always results in substantial improvement of the encephalopathy. However, although severe

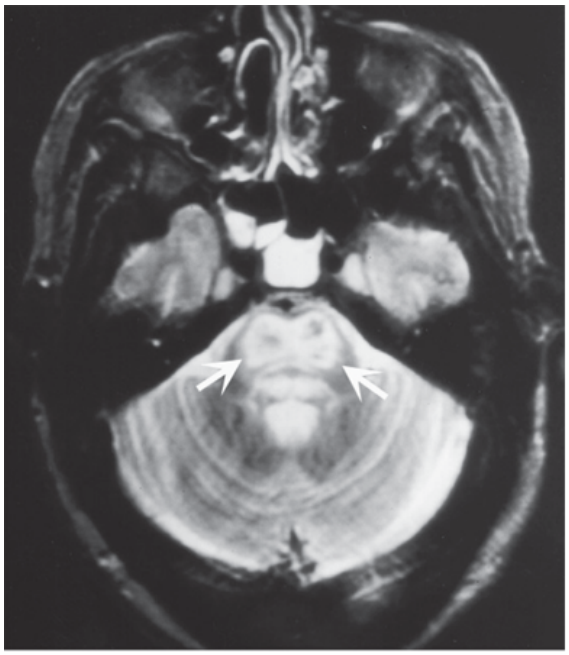
disability to the level of chronic vegetative or minimally conscious states is uncommon, long-term cognitive dysfunction clinically similar to dementia is being increasingly recognized in some survivors, especially in older patients.

OSMOTIC DEMYELINATION SYNDROME (CENTRAL PONTINE MYELINOLYSIS)

This disorder often presents in a devastating fashion as quadriplegia and pseudobulbar palsy although less severe presentations may occur. Predisposing factors include severe underlying medical illness or nutritional deficiency; most cases are associated with rapid correction of hyponatremia or with hyperosmolar states and clinical symptoms are usually identified a few days after sodium correction. Previously termed *central pontine myelinolysis*, the more accurate term *osmotic demyelination syndrome* is now preferred. The pathology consists of demyelination without inflammation in the base of the pons, with relative sparing of axons and nerve cells. MRI is useful in establishing the diagnosis **(Fig. 301-5)** and may also identify partial forms that present as confusion, dysarthria, and/or disturbances of conjugate gaze without quadriplegia. Occasional cases present with lesions outside of the brainstem. Therapeutic guidelines for the restoration of severe hyponatremia should aim for gradual correction, i.e., by ≤10 mmol/L (10 meq/L) within 24 h and 20 mmol/L (20 meq/L) within 48 h.

FIGURE 301-5

Osmotic demyelination syndrome. Axial T2-weighted magnetic resonance scan through the pons reveals a symmetric area of abnormal high signal intensity within the basis pontis (*arrows*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

WERNICKE'S DISEASE

Wernicke's disease is a common and preventable disorder due to a deficiency of thiamine **(Chap. 326)**. In the United States, alcoholics account for most cases, but patients with malnutrition due to hyperemesis, starvation, renal dialysis, cancer, HIV/AIDS, or rarely gastric surgery are also at risk. The characteristic clinical triad is ophthalmoplegia, ataxia, and global confusion. However, only one-third of patients with acute Wernicke's disease present with the classic clinical triad. Most patients are profoundly disoriented, indifferent, and inattentive, although rarely they have an agitated delirium related to ethanol withdrawal. If the disease is not treated, stupor, coma, and death may ensue. Ocular motor abnormalities include horizontal nystagmus on lateral gaze, lateral rectus palsy (usually bilateral), conjugate gaze palsies, and rarely ptosis. Gait ataxia probably results from a combination of polyneuropathy, cerebellar involvement, and vestibular paresis. The pupils are usually spared, but they may become miotic with advanced disease.

Wernicke's disease is usually associated with other manifestations of nutritional disease, such as polyneuropathy. Rarely, amblyopia or myelopathy occurs. Tachycardia and postural hypotension may be related to impaired function of the autonomic nervous system or to the coexistence of cardiovascular beriberi. Patients who recover show improvement in ocular palsies within hours after the administration of thiamine, but horizontal nystagmus may persist. Ataxia improves more slowly than the ocular motor abnormalities. Approximately half recover incompletely and are left with a slow, shuffling, wide-based gait and an inability to tandem walk. Apathy, drowsiness, and confusion improve more gradually. As these symptoms

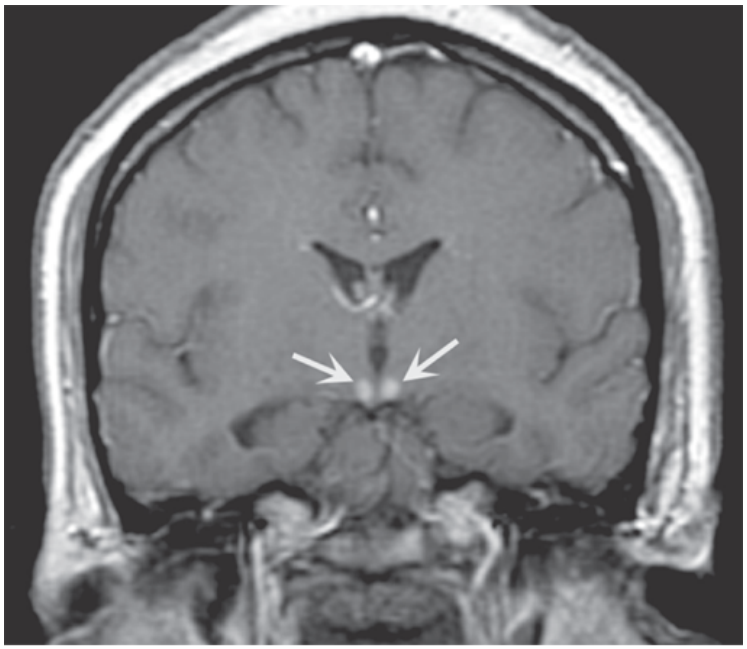
recede, an amnestic state with impairment in recent memory and learning may become more apparent (*Korsakoff's psychosis*). Korsakoff's psychosis is frequently persistent; the residual mental state is characterized by gaps in memory, confabulation, and disordered temporal sequencing.

Pathology

Periventricular lesions surround the third ventricle, aqueduct, and fourth ventricle, with petechial hemorrhages in occasional acute cases and atrophy of the mammillary bodies in most chronic cases. There is frequently endothelial proliferation, demyelination, and some neuronal loss. These changes may be detected by MRI **(Fig. 301-6)**. The amnestic defect is related to lesions in the dorsal medial nuclei of the thalamus.

FIGURE 301-6

Wernicke's disease. Coronal T1-weighted postcontrast magnetic resonance imaging reveals abnormal enhancement of the mammillary bodies (*arrows*), typical of acute Wernicke's encephalopathy.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pathogenesis

Thiamine is a cofactor of several enzymes, including transketolase, pyruvate dehydrogenase, and α-ketoglutarate dehydrogenase. Thiamine deficiency produces a diffuse decrease in cerebral glucose utilization and results in mitochondrial damage. Glutamate accumulates due to impairment of α-ketoglutarate dehydrogenase activity and, in combination with the energy deficiency, may result in excitotoxic cell damage.

TREATMENT

TREATMENT

Wernicke's Disease

Wernicke's disease is a medical emergency and requires immediate administration of thiamine, in a dose of 100 mg either IV or IM. The dose should be given daily until the patient resumes a normal diet and should be begun prior to treatment with IV glucose solutions. Larger doses, 100 mg four times a day or more, have been advocated by some. Glucose infusions may precipitate Wernicke's disease in a previously unaffected patient or cause a rapid worsening of an early form of the disease. For this reason, thiamine should be administered to all alcoholic patients requiring parenteral glucose.

HYPERPERFUSION DISORDERS (POSTERIOR REVERSIBLE ENCEPHALOPATHY SYNDROME)

Several seemingly diverse syndromes including hypertensive encephalopathy, eclampsia, postcarotid endarterectomy syndrome, and toxicity from calcineurin-inhibitor and other medications share the common pathogenesis of hyperperfusion likely due to endothelial dysfunction. Vasogenic edema is typically the primary process leading to neurologic dysfunction and this is thought to result from one of two mechanisms: exceeding the cerebral autoregulatory threshold leading to increased CBF and capillary leakage into the interstitium, or direct impairment of the BBB itself. The predilection of

all of the hyperperfusion disorders to affect the posterior rather than anterior portions of the brain may be due to a lower threshold for autoregulatory breakthrough in the posterior circulation or a vasculopathy that is more common in these blood vessels.

These disorders of hyperperfusion can be divided into those caused primarily by increased pressure and those due to endothelial dysfunction from a toxic or autoimmune etiology **(Table 301-3)**. In reality, both of these processes likely play some role in each of these disorders. The clinical presentation of all of the hyperperfusion syndromes is similar with prominent headaches, seizures, or focal neurologic deficits. Headaches have no specific characteristics, range from mild to severe, and may be accompanied by alterations in consciousness ranging from confusion to coma. Seizures may be present, and these can be of multiple types depending on the severity and location of the edema. Nonconvulsive seizures have been described in hyperperfusion states; therefore, a low threshold for obtaining an electroencephalogram (EEG) in these patients should be maintained. The typical focal deficit in hyperperfusion states is cortical visual loss, given the tendency of the process to involve the occipital lobes. However, any focal deficit can occur depending on the area affected, as evidenced by patients who, after carotid endarterectomy, exhibit neurologic dysfunction referable to the ipsilateral newly reperfused hemisphere. It appears as if the rapidity of rise, rather than the absolute value of pressure, is the most important risk factor.

TABLE 301-3

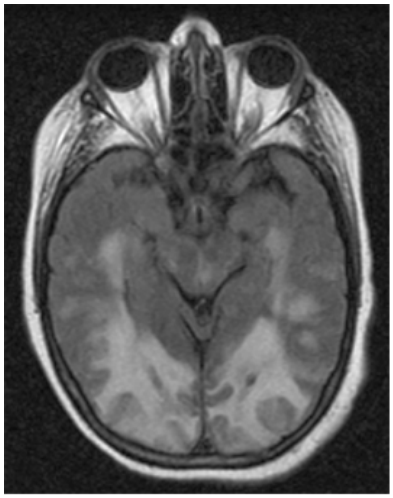
Common Etiologies of Posterior Reversible Encephalopathy Syndrome

Disorders in which increased capillary pressure dominates the pathophysiology
Hypertensive encephalopathy, including secondary causes such as renovascular hypertension, pheochromocytoma, cocaine use, etc.
Postcarotid endarterectomy syndrome
Preeclampsia/eclampsia
Disorders in which endothelial dysfunction dominates the pathophysiology
Calcineurin-inhibitor toxicity (e.g. cyclosporine, tacrolimus)
Chemotherapeutic agent toxicity (e.g., cytarabine, azathioprine, 5-fluorouracil, cisplatin, methotrexate, tumor necrosis factor α antagonists)
HELLP syndrome (hemolysis, elevated liver enzyme levels, low platelet count)
Hemolytic-uremic syndrome (HUS)

MRI classically exhibits the high T2 signal of edema primarily in the posterior occipital lobes, not respecting any single vascular territory **(Figure 301-7)**. CT is less sensitive but may show a pattern of patchy hypodensity in the involved territory. The term *posterior reversible encephalopathy syndrome* (PRES) is often used to describe these conditions; however, the clinical syndrome is not always reversible or limited just to the posterior brain regions. Vessel imaging may demonstrate narrowing of the cerebral vasculature, especially in the posterior circulation; whether this noninflammatory vasculopathy is a primary cause of the edema or occurs as a secondary phenomenon remains unclear. Other ancillary studies such as CSF analysis often yield nonspecific results. Many of the substances that have been implicated, such as cyclosporine, can cause this syndrome even at low doses or after years of treatment. Therefore, normal serum levels of these medications do not exclude them as inciting agents.

FIGURE 301-7

Axial fluid-attenuated inversion recovery (FLAIR) magnetic resonance imaging (MRI) of the brain in a patient taking cyclosporine after liver transplantation, who presented with seizures, headache, and cortical blindness. Increased signal is seen bilaterally in the occipital lobes predominantly involving the white matter, consistent with a hyperperfusion state secondary to calcineurin-inhibitor exposure.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Treatment involves judicious lowering of the blood pressure with IV agents such as labetalol or nicardipine, removal of the offending medication, and treatment of an underlying medical condition such as eclampsia. If the blood pressure is very elevated, it is reasonable to lower the MAP by ~20% initially, as further lowering of the pressure may cause secondary ischemia and possibly infarction as pressure drops below the lower range of the patient's autoregulatory capability. Seizures must be identified and controlled, often necessitating continuous EEG monitoring. Anticonvulsants are effective when seizure activity is identified, but in the special case of eclampsia, there is evidence to support the use of magnesium sulfate for seizure control.

POST-SOLID ORGAN TRANSPLANT BRAIN INJURY

Immunosuppressive medications are administered in high doses to patients after solid organ transplant, and many of these compounds have well-described neurologic complications. In patients with headache, seizures, or focal neurologic deficits taking calcineurin inhibitors, the diagnosis of hyperperfusion syndrome should be considered, as discussed above. This neurotoxicity occurs mainly with cyclosporine and tacrolimus and can present even in the setting of normal serum drug levels. Treatment primarily involves lowering the drug dosage or discontinuing the drug. Sirolimus has very few recorded cases of neurotoxicity and may be a reasonable alternative for some patients. Other examples of immunosuppressive medications and their neurologic complications include OKT3-associated akinetic mutism and the leukoencephalopathy seen with methotrexate, especially when it is administered intrathecally or with concurrent radiotherapy. In any solid organ transplant patient with neurologic complaints, a careful examination of the medication list is required to search for these possible drug effects.

Cerebrovascular complications of solid organ transplant are often first recognized in the immediate postoperative period. Border zone territory infarctions can occur, especially in the setting of systemic hypotension during cardiac transplant surgery. Embolic infarctions classically complicate cardiac transplantation, but all solid organ transplant procedures place patients at risk for systemic emboli. When cerebral embolization accompanies renal or liver transplantation surgery, a careful search for right-to-left shunting should include evaluation of the heart with agitated saline echocardiography (i.e., "bubble study"), as well as looking for intrapulmonary shunting. Renal and some cardiac transplant patients often have advanced atherosclerosis, providing a risk for stroke. Imaging with CT or MRI should be done when cerebrovascular complications are suspected to confirm the diagnosis and to exclude intracerebral hemorrhage, which most often occurs in the setting of coagulopathy secondary to liver failure or after cardiac bypass procedures.

Given that patients with solid organ transplants are chronically immunosuppressed, infections are a common concern **(Chap. 138)**. In any transplant patient with new CNS signs or symptoms such as seizure, confusion, or focal deficit, the diagnosis of a CNS infection should be considered and evaluated through imaging (usually MRI) and possibly lumbar puncture. The most common pathogens responsible for CNS infections in these patients vary based on time since transplant. In the first month posttransplant, common pathogens include the usual bacterial organisms associated with surgical procedures and indwelling catheters. Starting in the second month posttransplant, opportunistic infections of the CNS become more common, including *Nocardia* and *Toxoplasma* species as well as fungal infections such as aspergillosis. Viral infections that can affect the brain of the immunosuppressed patient, such as herpes simplex virus, cytomegalovirus, human herpesvirus type 6 (HHV-6), and varicella, also become more common after the first month posttransplant. Beyond 6 months, immunosuppressed posttransplant patients still remain at risk for these opportunistic bacterial, fungal, and viral infections but can also suffer late CNS infectious complications such as progressive multifocal leukoencephalopathy (PML) associated with JC virus **(Chap. 132)**, and Epstein-Barr virus–driven clonal expansions of B cells resulting in posttransplant lymphoproliferative disorder or CNS lymphoma **(Chap. 86)**.

CRITICAL CARE DISORDERS OF THE PERIPHERAL NERVOUS SYSTEM (PNS)

Critical illness with disorders of the PNS arises in two contexts: (1) primary neurologic diseases that require critical care interventions such as intubation and mechanical ventilation, and (2) secondary PNS manifestations of systemic critical illness, often involving multisystem organ failure. The former include acute polyneuropathies such as Guillain-Barré syndrome **(Chap. 439)**, neuromuscular junction disorders including myasthenia gravis **(Chap. 440)** and botulism **(Chap. 148)**, and primary muscle disorders such as polymyositis **(Chap. 358)**. The latter result either from the systemic disease itself or as a consequence of interventions and as a group are often referred to as ICU acquired weakness (ICUAW).

General principles of respiratory evaluation in patients with PNS involvement, regardless of cause, include assessment of pulmonary mechanics, such as maximal inspiratory force (MIF) and vital capacity (VC), and evaluation of strength of bulbar muscles. Regardless of the cause of weakness, endotracheal intubation should be considered when the MIF falls to below –25 cmH_2O or the VC is <1 L. Also, patients with severe palatal weakness may require endotracheal intubation in order to prevent acute upper airway obstruction or recurrent aspiration. Arterial blood gases and oxygen saturation from pulse oximetry are used to follow patients with potential respiratory compromise from PNS dysfunction. However, intubation and mechanical ventilation should be undertaken based on clinical assessment rather than waiting until oxygen saturation drops or CO_2 retention develops from hypoventilation. Noninvasive mechanical ventilation may be considered initially in lieu of endotracheal intubation in myasthenia gravis but is generally insufficient in patients with severe bulbar weakness or ventilatory failure with hypercarbia. **Principles of mechanical ventilation are discussed in Chap. 295.**

NEUROPATHY

Although encephalopathy may be the most obvious neurologic dysfunction in critically ill patients, dysfunction of the PNS is also quite common. It is typically present in patients with prolonged critical illnesses lasting several weeks and involving sepsis; clinical suspicion is aroused when there is failure to wean from mechanical ventilation despite improvement of the underlying sepsis and critical illness. *Critical illness polyneuropathy* refers to the most common PNS complication related to critical illness; it is seen in the setting of prolonged critical illness, sepsis, and multisystem organ failure. Neurologic findings include diffuse weakness, decreased reflexes, and distal sensory loss. Electrophysiologic studies demonstrate a diffuse, symmetric, distal axonal sensorimotor neuropathy, and pathologic studies have confirmed axonal degeneration. The precise mechanism of critical illness polyneuropathy remains unclear, but circulating factors such as cytokines, which are associated with sepsis and SIRS, are thought to play a role. It has been reported that up to 70% of patients with the sepsis syndrome have some degree of neuropathy, although far fewer have a clinical syndrome profound enough to cause severe respiratory muscle weakness requiring prolonged mechanical ventilation or resulting in failure to wean. Aggressive glycemic control with insulin infusions appears to decrease the risk of critical illness polyneuropathy. Treatment is otherwise supportive, with specific intervention directed at treating the underlying illness. Although spontaneous recovery is usually seen, the time course may extend over weeks to months and necessitate long-term ventilatory support and care even after the underlying critical illness has resolved.

DISORDERS OF NEUROMUSCULAR TRANSMISSION

A defect in neuromuscular transmission may be a source of weakness in critically ill patients. Botulism **(Chap. 148)** may be acquired by ingesting botulinum toxin from improperly stored food or may arise from an anaerobic abscess from *Clostridium botulinum* (wound botulism). Infants can present with generalized weakness from gut-derived *Clostridium* infection, especially if they are fed honey. Diplopia and dysphagia are early signs of food-borne botulism. Treatment is mostly supportive, although use of antitoxin early in the course may limit the duration of the neuromuscular blockade. General ICU care is similar to patients with Guillain-Barré syndrome or myasthenia gravis with focused care to avoid ulcer formation at pressure points, deep venous thromboprophylaxis, and infection prevention. Public health officers should be rapidly informed when the diagnosis is made to prevent further exposure to others from the tainted food or source of wound botulism (such as injection drug use).

Undiagnosed myasthenia gravis **(Chap. 440)** may be a consideration in weak ICU patients; however, persistent weakness secondary to impaired neuromuscular junction transmission is almost always due to administration of drugs. A number of medications impair neuromuscular transmission; these include antibiotics, especially aminoglycosides, and beta-blocking agents. In the ICU, the nondepolarizing neuromuscular blocking agents (nd-NMBAs), also known as muscle relaxants, are most commonly responsible. Included in this group of drugs are such agents as pancuronium, vecuronium, rocuronium, and cisatracurium. They are often used to facilitate mechanical ventilation or other critical care procedures, but with prolonged use persistent neuromuscular blockade may result in weakness even after discontinuation of these agents hours or days earlier. Risk factors for this prolonged action of neuromuscular blocking agents include female sex, metabolic acidosis, and renal failure.

Prolonged neuromuscular blockade does not appear to produce permanent damage to the PNS. Once the offending medications are discontinued, full strength is restored, although this may take days. In general, the lowest dose of neuromuscular blocking agent should be used to achieve the desired result and, when these agents are used in the ICU, a peripheral nerve stimulator should be used to monitor neuromuscular junction function.

MYOPATHY

Critically ill patients, especially those with sepsis, frequently develop muscle weakness and wasting, often in the face of seemingly adequate nutritional support. *Critical illness myopathy* is an overall term that describes several different discrete muscle disorders that may occur in critically ill patients. The assumption has been that a catabolic myopathy may develop as a result of multiple factors, including elevated cortisol and catecholamine release

and other circulating factors induced by the SIRS. In this syndrome, known as *cachectic myopathy*, serum creatine kinase levels and electromyography (EMG) are normal. Muscle biopsy shows type II fiber atrophy. Panfascicular muscle fiber necrosis may also occur in the setting of profound sepsis. This less common *acute necrotizing intensive care myopathy* is characterized clinically by weakness progressing to a profound level over just a few days. There may be associated elevations in serum creatine kinase and urine myoglobin. Both EMG and muscle biopsy may be normal initially but eventually show abnormal spontaneous activity and panfascicular necrosis with an accompanying inflammatory reaction. Acute rhabdomyolysis can occur from alcohol ingestion or from compartment syndromes.

A *thick-filament myopathy* may occur in the setting of glucocorticoid and nd-NMBA use. The most frequent scenario in which this is encountered is the asthmatic patient who requires high-dose glucocorticoids and nd-NMBA to facilitate mechanical ventilation. This muscle disorder is not due to prolonged action of nd-NMBAs at the neuromuscular junction but, rather, is an actual myopathy with muscle damage; it has occasionally been described with high-dose glucocorticoid use or sepsis alone. Clinically this syndrome is most often recognized when a patient fails to wean from mechanical ventilation despite resolution of the primary pulmonary process. Pathologically, there may be loss of thick (myosin) filaments. Thick-filament critical illness myopathy has a good prognosis. If patients survive their underlying critical illness, the myopathy invariably improves and most patients return to normal. However, because this syndrome is a result of true muscle damage, not just prolonged blockade at the neuromuscular junction, this process may take weeks or months, and tracheotomy with prolonged ventilatory support may be necessary. Some patients do have residual long-term weakness, with atrophy and fatigue limiting ambulation. At present, it is unclear how to prevent this myopathic complication, except by avoiding use of nd-NMBAs, a strategy not always possible. Monitoring with a peripheral nerve stimulator can help to avoid the overuse of these agents. However, this is more likely to prevent the complication of prolonged neuromuscular junction blockade than it is to prevent this myopathy.

FURTHER READING

Callaway CW et al: Part 4: Advanced life support: 2015 international consensus on cardiopulmonary resuscitation and emergency cardiovascular care science with treatment recommendations. Circulation 132:S84, 2015. [PubMed: 26472860]

Dhar R: Neurologic complications of transplantation. Handb Clin Neurol 141:545, 2017. [PubMed: 28190435]

Donnelly J et al: Regulation of the cerebral circulation: bedside assessment and clinical implications. Crit Care 20:129, 2016. [PubMed: 27145751]

Latronico N et al: Critical illness polyneuropathy and myopathy: A major cause of muscle weakness and paralysis. Lancet Neurol 10:931, 2011. [PubMed: 21939902]

Nielsen N et al: Targeted temperature management at 33 degrees C versus 36 degrees C after cardiac arrest. N Engl J Med 369:2197, 2013. [PubMed: 24237006]

Pandharipande PP et al: Long-term cognitive impairment after critical illness. N Engl J Med 369:1306, 2013. [PubMed: 24088092]

Posner JB et al: *Plum and Posner's Diagnosis of Stupor and Coma*, 4th ed. New York, Oxford University Press, 2007.

Quillinan N et al.: Neuropathophysiology of brain injury. Anesthesiol Clin 34:453, 2016.

Sandroni C et al: Prognostication in comatose survivors of cardiac arrest: An advisory statement from the European Resuscitation Council and the European Society of Intensive Care Medicine. Intensive Care Med 40:1816, 2014. [PubMed: 25398304]

Toledano M, Fugate JE: Posterior reversible encephalopathy in the intensive care unit. Handb Clin Neurol 141:467, 2017. [PubMed: 28190431]

Chapter 302: Subarachnoid Hemorrhage

J. Claude Hemphill, III; Wade S. Smith; Daryl R. Gress

INTRODUCTION

Subarachnoid hemorrhage (SAH) renders the brain critically ill from both primary and secondary brain insults. Excluding head trauma, the most common cause of SAH is rupture of a saccular aneurysm. Other causes include bleeding from a vascular malformation (arteriovenous malformation or dural arteriovenous fistula) and extension into the subarachnoid space from a primary intracerebral hemorrhage. Some idiopathic SAHs are localized to the perimesencephalic cisterns and are benign; they probably have a venous or capillary source, and angiography is unrevealing.

SACCULAR ("BERRY") ANEURYSM

Autopsy and angiography studies have found that about 2% of adults harbor intracranial aneurysms, for a prevalence of 4 million persons in the United States; the aneurysm will rupture, producing SAH, in 25,000–30,000 cases per year. For patients who arrive alive at hospital, the mortality rate over the next month is about 45%. Of those who survive, more than half are left with major neurologic deficits as a result of the initial hemorrhage, cerebral vasospasm with infarction, or hydrocephalus. If the patient survives but the aneurysm is not obliterated, the rate of rebleeding is about 20% in the first 2 weeks, 30% in the first month, and about 3% per year afterward. Given these alarming figures, the major therapeutic emphasis is on preventing the predictable early complications of the SAH.

Unruptured, asymptomatic aneurysms are much less dangerous than a recently ruptured aneurysm. The annual risk of rupture for aneurysms <10 mm in size is ~0.1%, and for aneurysms ≥10 mm in size is ~0.5–1%; the surgical morbidity rate far exceeds these percentages. Aneurysm location may also factor into risk, with basilar bifurcation aneurysms appearing to have a somewhat higher rupture risk. Because of the longer length of exposure to risk of rupture, younger patients with aneurysms >10 mm in size may benefit from prophylactic treatment. As with the treatment of asymptomatic carotid stenosis, this risk-benefit ratio strongly depends on the complication rate of treatment.

Giant aneurysms, those >2.5 cm in diameter, occur at the same sites (see below) as small aneurysms and account for 5% of cases. The three most common locations are the terminal internal carotid artery, middle cerebral artery (MCA) bifurcation, and top of the basilar artery. Their risk of rupture is ~6% in the first year

after identification and may remain high indefinitely. They often cause symptoms by compressing the adjacent brain or cranial nerves.

Mycotic aneurysms are usually located distal to the first bifurcation of major arteries of the circle of Willis. Most result from infected emboli due to bacterial endocarditis causing septic degeneration of arteries and subsequent dilation and rupture. Whether these lesions should be sought and repaired prior to rupture or left to heal spontaneously with antibiotic treatment remains controversial.

Pathophysiology

Saccular aneurysms occur at the bifurcations of the large- to medium-sized intracranial arteries; rupture is into the subarachnoid space in the basal cisterns and sometimes into the parenchyma of the adjacent brain. Approximately 85% of aneurysms occur in the anterior circulation, mostly on the circle of Willis. About 20% of patients have multiple aneurysms, many at mirror sites bilaterally. As an aneurysm develops, it typically forms a neck with a dome. The length of the neck and the size of the dome vary greatly and are important factors in planning neurosurgical obliteration or endovascular embolization. The arterial internal elastic lamina disappears at the base of the neck. The media thins, and connective tissue replaces smooth-muscle cells. At the site of rupture (most often the dome), the wall thins, and the tear that allows bleeding is often ≤0.5 mm long. Aneurysm size and site are important in predicting risk of rupture. Those >7 mm in diameter and those at the top of the basilar artery and at the origin of the posterior communicating artery are at greater risk of rupture.

Clinical Manifestations

Most unruptured intracranial aneurysms are completely asymptomatic. Symptoms are usually due to rupture and resultant SAH, although some unruptured aneurysms present with mass effect on cranial nerves or brain parenchyma. At the moment of aneurysmal rupture with major SAH, the intracranial pressure (ICP) suddenly rises. This may account for the sudden transient loss of consciousness that occurs in nearly half of patients. Sudden loss of consciousness may be preceded by a brief moment of excruciating headache, but most patients first complain of headache upon regaining consciousness. In 10% of cases, aneurysmal bleeding is severe enough to cause loss of consciousness for several days. In ~45% of cases, severe headache associated with exertion is the presenting complaint. The patient often calls the headache “the worst headache of my life”; however, the most important characteristic is sudden onset. Occasionally, these ruptures may present as headache of only moderate intensity or as a change in the patient’s usual headache pattern. The headache is usually generalized, often with neck stiffness, and vomiting is common.

Although sudden headache in the absence of focal neurologic symptoms is the hallmark of aneurysmal rupture, focal neurologic deficits may occur. Anterior communicating artery or MCA bifurcation aneurysms may rupture into the adjacent brain or subdural space and form a hematoma large enough to produce mass effect. The deficits that result can include hemiparesis, aphasia, and mental slowness (abulia).

Occasionally, prodromal symptoms suggest the location of a progressively enlarging unruptured aneurysm. A third cranial nerve palsy, particularly when associated with pupillary dilation, loss of ipsilateral (but retained contralateral) light reflex, and focal pain above or behind the eye, may occur with an expanding aneurysm at the junction of the posterior communicating artery and the internal carotid artery. A sixth nerve palsy may indicate an aneurysm in the cavernous sinus, and visual field defects can occur with an expanding supraclinoid carotid or anterior cerebral artery (ACA) aneurysm. Occipital and posterior cervical pain may signal a posterior inferior cerebellar artery or anterior inferior cerebellar artery aneurysm **(Chap. 419)**. Pain in or behind the eye and in the low temple can occur with an expanding MCA aneurysm. Thunderclap headache is a variant of migraine that simulates an SAH. Before concluding that a patient with sudden, severe headache has thunderclap migraine, a definitive workup for aneurysm or other intracranial pathology is required.

Aneurysms can undergo small ruptures and leaks of blood into the subarachnoid space, so-called *sentinel bleeds*. Sudden unexplained headache at any location should raise suspicion of SAH and be investigated, because a major hemorrhage may be imminent.

The initial clinical manifestations of SAH can be graded using the Hunt-Hess or World Federation of Neurosurgical Societies classification schemes **(Table 302-1)**. For ruptured aneurysms, prognosis for good outcomes falls as the grade increases. For example, it is unusual for a Hunt-Hess grade 1 patient to die if the aneurysm is treated, but the mortality rate for grade 4 and 5 patients may be as high as 60%.

TABLE 302-1

Grading Scales for Subarachnoid Hemorrhage

Grade	Hunt-Hess Scale	World Federation of Neurosurgical Societies (WFNS) Scale
1	Mild headache, normal mental status, no cranial nerve or motor findings	GCS[a] score 15, no motor deficits
2	Severe headache, normal mental status, may have cranial nerve deficit	GCS score 13–14, no motor deficits
3	Somnolent, confused, may have cranial nerve or mild motor deficit	GCS score 13–14, with motor deficits
4	Stupor, moderate to severe motor deficit, may have intermittent reflex posturing	GCS score 7–12, with or without motor deficits
5	Coma, reflex posturing or flaccid	GCS score 3–6, with or without motor deficits

[a]Glasgow Coma Scale; see Table 435-1.

Delayed Neurologic Deficits

There are four major causes of delayed neurologic deficits: rerupture, hydrocephalus, delayed cerebral ischemia (DCI), and hyponatremia.

1. *Rerupture*. The incidence of rerupture of an untreated aneurysm in the first month following SAH is ~30%, with the peak in the first 7 days. Rerupture is associated with a 50% mortality rate and poor outcome. Early treatment eliminates this risk.

2. *Hydrocephalus*. Acute hydrocephalus can cause stupor and coma and can be mitigated by placement of an external ventricular drain. More often, subacute hydrocephalus may develop over a few days or weeks and causes progressive drowsiness or slowed mentation with incontinence. Hydrocephalus is differentiated from cerebral vasospasm with a CT scan, CT angiogram, transcranial Doppler (TCD) ultrasound, or conventional x-ray angiography. Hydrocephalus may clear spontaneously or require temporary ventricular drainage. Chronic hydrocephalus may develop weeks to months after SAH and manifest as gait difficulty, incontinence, or impaired mentation. Subtle signs may be a lack of initiative in conversation or a failure to recover independence.

3. *Delayed cerebral ischemia*. Vasospasm is the narrowing of the arteries at the base of the brain following SAH. This may cause symptomatic ischemia and infarction in ~30% of patients and is the major cause of delayed morbidity and death. Signs of DCI appear 4–14 days after the hemorrhage, most often at 7 days. The severity and distribution of vasospasm determine whether infarction will occur.

    a. Vasospasm is believed to result from direct effects of clotted blood and its breakdown products on the arteries within the subarachnoid space. In general, the more blood that surrounds the arteries, the greater the chance of symptomatic vasospasm. Spasm of major arteries produces symptoms referable to the appropriate vascular territory **(Chap. 419)**. All of these focal symptoms may present abruptly, fluctuate, or develop over a few days. In most cases, focal spasm is preceded by a decline in mental status.

    b. Vasospasm can be detected reliably with conventional x-ray angiography, but this procedure is invasive and carries the risk of stroke and other complications. TCD ultrasound is based on the principle that the velocity of blood flow within an artery will rise as the lumen diameter is narrowed. By directing the probe along the MCA and proximal ACA, carotid terminus, and vertebral and basilar arteries on a daily or every-other-day basis, vasospasm can be reliably detected and treatments initiated to prevent cerebral ischemia (see below). CT angiography is another method that can detect vasospasm.

    c. Severe cerebral edema in patients with infarction from vasospasm may increase the ICP enough to reduce cerebral perfusion pressure. Treatment may include mannitol, hyperventilation, and for intractable cases hemicraniectomy; moderate hypothermia may have a role as well.

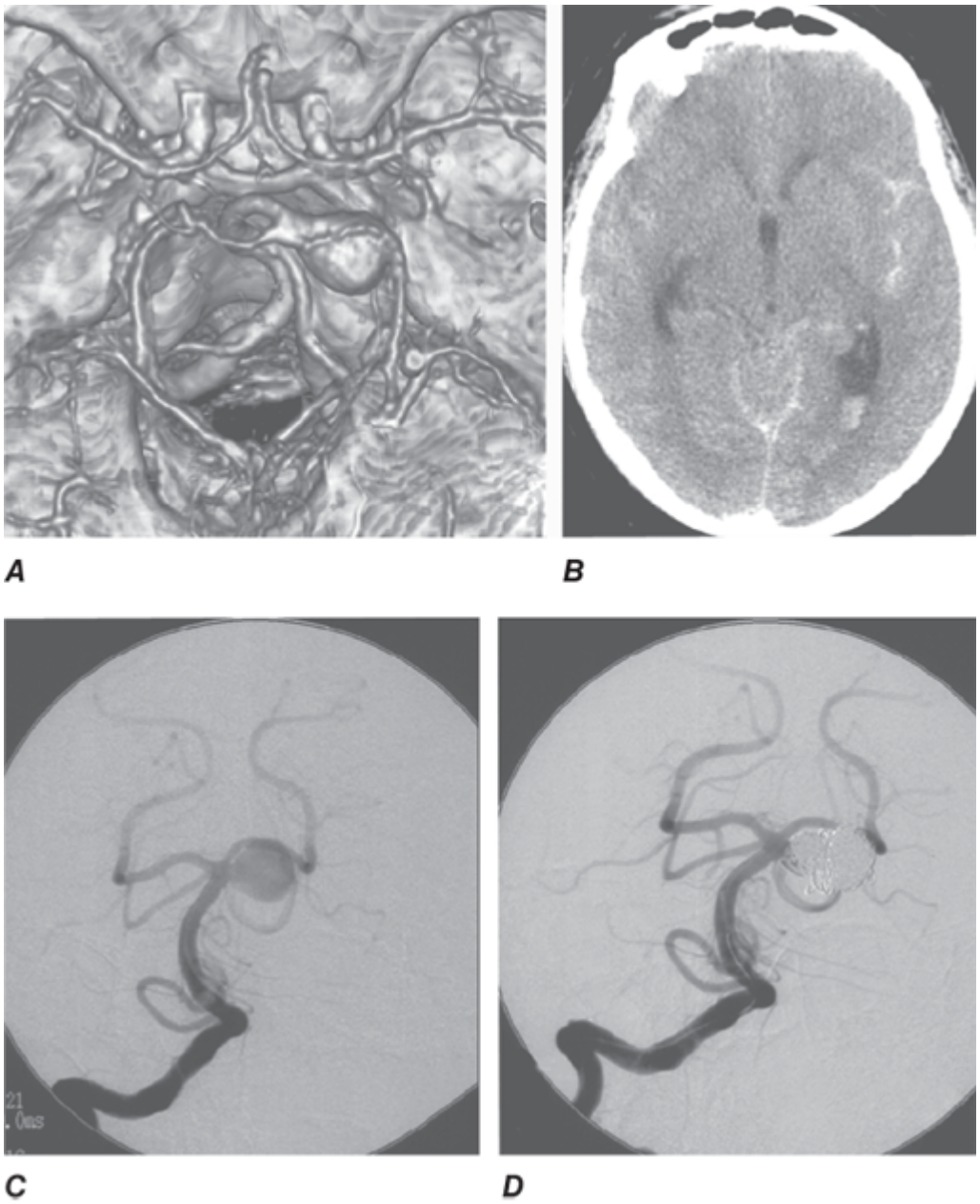
4. *Hyponatremia*. Hyponatremia may be profound and can develop quickly in the first 2 weeks following SAH. There is both natriuresis and volume depletion with SAH, so that patients become both hyponatremic and hypovolemic. Both atrial natriuretic peptide and brain natriuretic peptide have a role in producing this “cerebral salt-wasting syndrome.” Typically, it clears over the course of 1–2 weeks and, in the setting of SAH, should not be treated with free-water restriction as this may increase the risk of stroke (see below).

Laboratory Evaluation and Imaging

The hallmark of aneurysmal rupture is blood in the cerebrospinal fluid (CSF) **(Fig. 302-1)**. More than 95% of cases have enough blood to be visualized on a high-quality noncontrast CT scan obtained within 72 h. If the scan fails to establish the diagnosis of SAH and no mass lesion or obstructive hydrocephalus is found, a lumbar puncture should be performed to establish the presence of subarachnoid blood. Lysis of the red blood cells and subsequent conversion of hemoglobin to bilirubin stains the spinal fluid yellow within 6–12 h. This xanthochromic spinal fluid peaks in intensity at 48 h and lasts for 1–4 weeks, depending on the amount of subarachnoid blood.

FIGURE 302-1

Subarachnoid hemorrhage. ***A.*** Computed tomography (CT) angiography revealing an aneurysm of the left superior cerebellar artery. ***B.*** Noncontrast CT scan at the level of the third ventricle revealing subarachnoid blood (*bright*) in the left sylvian fissure and within the left lateral ventricle. ***C.*** Conventional anteroposterior x-ray angiogram of the right vertebral and basilar artery showing the large aneurysm. ***D.*** Conventional angiogram following coil embolization of the aneurysm, whereby the aneurysm body is filled with platinum coils delivered through a microcatheter navigated from the femoral artery into the aneurysm neck.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The extent and location of subarachnoid blood on a noncontrast CT scan help locate the underlying aneurysm, identify the cause of any neurologic deficit, and predict the occurrence of vasospasm. A high incidence of symptomatic vasospasm in the MCA and ACA has been found when early CT scans show subarachnoid clots >5 × 3 mm in the basal cisterns, or layers of blood >1 mm thick in the cerebral fissures. CT scans less reliably predict vasospasm in the vertebral, basilar, or posterior cerebral arteries.

Lumbar puncture prior to an imaging procedure is indicated only if a CT scan is not available at the time of the suspected SAH. Once the diagnosis of hemorrhage from a ruptured saccular aneurysm is suspected, four-vessel conventional x-ray angiography (both carotids and both vertebrals) is generally performed to localize and define the anatomic details of the aneurysm and to determine if other unruptured aneurysms exist (Fig. 302-1C). At some centers, the ruptured aneurysm can be treated using endovascular techniques at the time of the initial angiogram as a way to expedite treatment and minimize the number of invasive procedures. CT angiography is an alternative method for locating the aneurysm and may be sufficient to plan definitive therapy.

Close monitoring (daily or twice daily) of electrolytes is important because hyponatremia can occur precipitously during the first 2 weeks following SAH (see above).

The electrocardiogram (ECG) frequently shows ST-segment and T-wave changes similar to those associated with cardiac ischemia. A prolonged QRS complex, increased QT interval, and prominent “peaked” or deeply inverted symmetric T waves are usually secondary to the intracranial hemorrhage. There is evidence that structural myocardial lesions produced by circulating catecholamines and excessive discharge of sympathetic neurons may occur after SAH, causing these ECG changes and a reversible cardiomyopathy sufficient to cause shock or congestive heart failure. Echocardiography reveals a pattern of regional wall motion abnormalities that follow the distribution of sympathetic nerves rather than the major coronary arteries, with relative sparing of the ventricular wall apex. The sympathetic nerves themselves appear to be injured by direct toxicity from the excessive catecholamine release. An asymptomatic troponin elevation is common. Serious ventricular dysrhythmias occurring in-hospital are unusual.

TREATMENT

TREATMENT

Subarachnoid Hemorrhage

Early aneurysm repair prevents rerupture and allows the safe application of techniques to improve blood flow (e.g., induced hypertension) should vasospasm and DCI develop. An aneurysm can be “clipped” by a neurosurgeon or “coiled” by an endovascular surgeon. Surgical repair involves placing a metal clip across the aneurysm neck, thereby immediately eliminating the risk of rebleeding. This approach requires craniotomy and brain retraction, which is associated with neurologic morbidity. Endovascular techniques involve placing platinum coils, or other embolic material, within the aneurysm via a catheter that is passed from the femoral artery. The aneurysm is packed tightly to enhance thrombosis and over time is walled off from the circulation (Fig. 302-1D). There have been two prospective randomized trials of surgery versus endovascular treatment for ruptured aneurysms: the first was the International Subarachnoid Aneurysm Trial (ISAT), which was terminated early when 24% of patients treated with endovascular therapy were dead or dependent at 1 year compared to 31% treated with surgery, a significant 23% relative reduction. After 5 years, risk of death was lower in the coiling group, although the proportion of survivors who were independent was the same in both groups. Risk of rebleeding was low, but more common in the coiling group. These results favoring coiling at 1

year were confirmed in a second trial, although the differences in functional outcome were no longer significant at 3 years. Because some aneurysms have a morphology that is not amenable to endovascular treatment, surgery remains an important treatment option. Newer endovascular techniques using balloon-assisted coiling or placement of flow-diverting stents are increasing the types of aneurysms amenable to endovascular intervention. Centers that combine both endovascular and neurosurgical expertise likely offer the best outcomes for patients, and there are reliable data showing that specialized aneurysm treatment centers can improve mortality rates.

The medical management of SAH focuses on protecting the airway, managing blood pressure before and after aneurysm treatment, preventing rebleeding prior to treatment, managing vasospasm and DCI, treating hydrocephalus, treating hyponatremia, limiting secondary brain insults, and preventing pulmonary embolus (PE).

Intracranial hypertension following aneurysmal rupture occurs secondary to subarachnoid blood, parenchymal hematoma, acute hydrocephalus, or loss of vascular autoregulation. Patients who are stuporous should undergo emergent ventriculostomy to measure ICP and to treat high ICP in order to prevent cerebral ischemia. Medical therapies designed to combat raised ICP (e.g., osmotic therapy and sedation) can also be used as needed. High ICP refractory to treatment is a poor prognostic sign.

Prior to definitive treatment of the ruptured aneurysm, care is required to maintain adequate cerebral perfusion pressure while avoiding excessive elevation of arterial pressure. If the patient is alert, it is reasonable to lower the systolic blood pressure to below 160 mmHg using nicardipine, labetalol, or esmolol. If the patient has a depressed level of consciousness, ICP should be measured and the cerebral perfusion pressure targeted to 60–70 mmHg. If headache or neck pain is severe, mild sedation and analgesia are prescribed. Extreme sedation is avoided if possible because it can obscure the ability to clinically detect changes in neurologic status. Adequate hydration is necessary to avoid a decrease in blood volume predisposing to brain ischemia.

Seizures are uncommon at the onset of aneurysmal rupture. The quivering, jerking, and extensor posturing that often accompany loss of consciousness with SAH are probably related to the sharp rise in ICP rather than seizures. However, anticonvulsants are sometimes given as prophylactic therapy because a seizure could theoretically promote rebleeding.

Glucocorticoids may help reduce the head and neck ache caused by the irritative effect of the subarachnoid blood. There is no good evidence that they reduce cerebral edema, are neuroprotective, or reduce vascular injury, and their routine use therefore is not recommended.

Antifibrinolytic agents are not routinely prescribed but may be considered in patients in whom aneurysm treatment cannot proceed immediately. They are associated with a reduced incidence of aneurysmal rerupture but may also increase the risk of DCI and deep-vein thrombosis (DVT). Several recent studies suggest that a shorter duration of use (until the aneurysm is secured or for the first 3 days) may decrease rerupture and be safer than found in earlier studies of longer duration treatment.

DCI due to vasospasm remains the leading cause of morbidity and mortality following aneurysmal SAH. Treatment with the calcium channel antagonist nimodipine (60 mg PO every 4 h) improves outcome, perhaps by preventing ischemic injury rather than reducing the risk of vasospasm. Nimodipine can cause significant hypotension in some patients, which may worsen cerebral ischemia in patients with vasospasm. Symptomatic cerebral vasospasm can also be treated by increasing the cerebral perfusion pressure by raising mean arterial pressure through plasma volume expansion and the judicious use of IV vasopressor agents, usually phenylephrine or norepinephrine. Raised perfusion pressure has been associated with clinical improvement in many patients, but high arterial pressure may promote rebleeding in unprotected aneurysms. Treatment with induced hypertension and hypervolemia generally requires monitoring of arterial and central venous pressures; it is best to infuse pressors through a central venous line as well. Volume expansion helps prevent hypotension and augments cardiac output.

If DCI due to vasospasm persists despite optimal medical therapy, intraarterial vasodilators and percutaneous transluminal angioplasty are considered **(Fig. 302-2)**. Vasodilatation by direct angioplasty appears to be permanent, allowing hypertensive therapy to be tapered sooner. The pharmacologic vasodilators (verapamil and nicardipine) do not last more than about 24 h, and therefore multiple treatments may be required until the subarachnoid blood is reabsorbed. Although intraarterial papaverine is an effective vasodilator, there is evidence that papaverine may be neurotoxic, so its use should generally be avoided.

Acute hydrocephalus can cause stupor or coma. It may clear spontaneously or require temporary ventricular drainage. When chronic hydrocephalus develops, ventricular shunting is the treatment of choice.

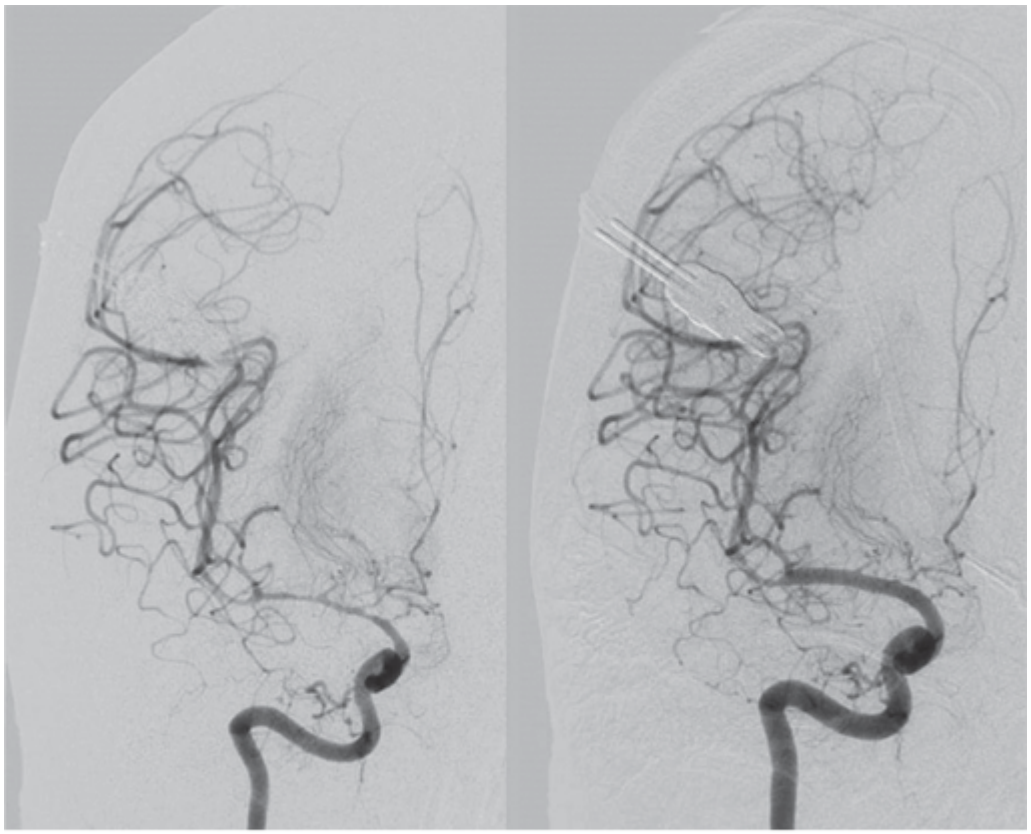
Free-water restriction is contraindicated in patients with SAH at risk for DCI because hypovolemia and hypotension may occur and precipitate cerebral ischemia. Many patients continue to experience a decline in serum sodium despite receiving parenteral fluids containing normal saline. Frequently, supplemental oral salt coupled with normal saline will mitigate hyponatremia, but often patients also require intravenous hypertonic saline. Care must be taken not to correct serum sodium too quickly in patients with marked hyponatremia of several days' duration, as the osmotic demyelination syndrome (Chap. 301) may occur.

All patients should have pneumatic compression stockings applied to prevent pulmonary embolism. Unfractionated heparin administered subcutaneously for DVT prophylaxis can be initiated within 1–2 days following endovascular treatment or craniotomy with surgical clipping and is a useful adjunct to pneumatic compression stockings. Treatment of PE depends on whether the aneurysm has been treated and whether or not the patient has had a craniotomy. Systemic anticoagulation with heparin is contraindicated in patients with ruptured and untreated aneurysms. It is a relative contraindication following craniotomy for several days, and it may delay thrombosis of a coiled aneurysm. If DVT or PE occurs within the first days following craniotomy, use of an inferior vena cava filter may be considered to prevent additional PEs, whereas systemic anticoagulation with heparin is preferred following successful endovascular treatment.

FIGURE 302-2

Vasospasm of the right middle cerebral artery. ***A.*** Catheter angiography demonstrates significant narrowing of the right middle cerebral artery (MCA). ***B.*** Because of symptomatic delayed cerebral ischemia, soft-balloon

angioplasty was used to dilate the proximal portion of the main MCA stem.



A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Diringer MN et al: Critical care management of patients following aneurysmal subarachnoid hemorrhage: Recommendations from the Neurocritical Care Society's Multidisciplinary Consensus Conference. Neurocrit Care 15:211, 2011. [PubMed: 21773873]

Etminan N, MacDonald RL: Management of aneurysmal subarachnoid hemorrhage. Handb Clin Neurol 140:195, 2017. [PubMed: 28187800]

Molyneux AJ et al: The durability of endovascular coiling versus neurosurgical clipping of ruptured cerebral aneurysms: 18 year follow-up of the UK cohort of the International Subarachnoid Aneurysm Trial (ISAT). Lancet 385:691, 2015. [PubMed: 25465111]

Chapter 303: Cellular and Molecular Biology of the Kidney

Alfred L. George, Jr.; Eric G. Neilson

INTRODUCTION

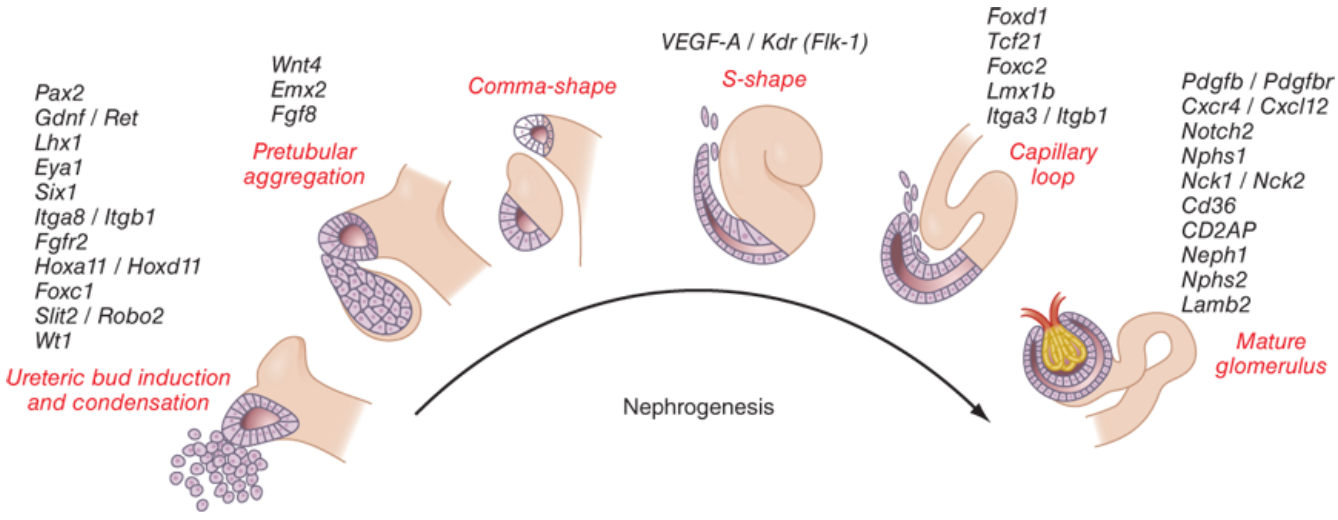
The kidney is one of the most highly differentiated organs in the body. At the conclusion of embryologic development, nearly 30 different cell types form a multitude of filtering capillaries and segmented nephrons enveloped by a dynamic interstitium. This cellular diversity modulates a variety of complex physiologic processes. Endocrine functions, the regulation of blood pressure and intraglomerular hemodynamics, solute and water transport, acid-base balance, and removal of drug metabolites are all accomplished by intricate mechanisms of renal response. This breadth of physiology hinges on the clever ingenuity of nephron architecture that evolved as complex organisms came out of water to live on land.

EMBRYOLOGIC DEVELOPMENT

Kidneys develop from intermediate mesoderm under the timed or sequential control of a growing number of genes, described in Fig. 303-1. The transcription of these genes is guided by morphogenic cues that invite two ureteric buds to each penetrate bilateral metanephric blastema, where they induce primary mesenchymal cells to form early nephrons. The two ureteric buds emerge from posterior nephric ducts and mature into separate collecting systems that eventually form a renal pelvis and ureter. Induced mesenchyme undergoes mesenchymal epithelial transitions to form comma-shaped bodies at the proximal end of each ureteric bud leading to the formation of S-shaped nephrons that cleft and enjoin with penetrating endothelial cells derived from sprouting angioblasts. Under the influence of vascular endothelial growth factor A (VEGF-A), these penetrating cells form capillaries with surrounding mesangial cells that differentiate into a glomerular filter for plasma water and solute. The ureteric buds branch and each branch produce a new set of nephrons. The number of branching events ultimately determines the total number of nephrons in each kidney. There are ~900,000 glomeruli in each kidney in normal birth weight adults and as few as 225,000 in low-birth-weight adults, with the latter producing numerous comorbid risks.

FIGURE 303-1

Genes controlling renal nephrogenesis. A growing number of genes have been identified at various stages of glomerulotubular development in the mammalian kidney. The genes listed have been tested in various genetically modified mice, and their location corresponds to the classical stages of kidney development postulated by Saxen in 1987.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Glomeruli evolve as complex capillary filters with fenestrated endothelia under the guiding influence of VEGF-A and angiopoietin-1 secreted by adjacently developing podocytes. Epithelial podocytes facing the urinary space envelop the exterior basement membrane supporting these emerging endothelial capillaries. Podocytes are partially polarized and periodically slough into the urinary space by epithelial-mesenchymal transition, and to a lesser extent apoptosis, only to be replenished by migrating parietal epithelia from Bowman capsule. Impaired replenishment results in heavy proteinuria. Podocytes attach to the basement membrane by special foot processes and share a slit-pore membrane with their neighbor. The slit-pore membrane forms a filter for plasma water and solute by the synthetic interaction of nephrin, annexin-4, CD2AP, FAT, ZO-1, P-cadherin, podocin, TRPC6, PLCE1, and Neph 1-3 proteins. Mutations in many of these proteins also result in heavy proteinuria. The glomerular capillaries are embedded in a mesangial matrix shrouded by parietal and proximal tubular epithelia forming Bowman capsule. Mesangial cells have an embryonic lineage consistent with arteriolar or juxtaglomerular cells and contain contractile actin-myosin fibers. These mesangial cells make contact with glomerular capillary loops, and their local matrix holds them in condensed arrangement.

Between nephrons lies the renal interstitium. This region forms a functional space surrounding glomeruli and their downstream tubules, which are home to resident and trafficking cells such as fibroblasts, dendritic cells, occasional lymphocytes, and lipid-laden macrophages. The cortical and medullary peritubular capillaries, which siphon off solute and water following tubular reclamation of glomerular filtrate, are also part of the interstitial fabric as well as a web of connective tissue that supports the kidney's emblematic architecture of folding tubules. The relational precision of these structures determines the unique physiology of the kidney.

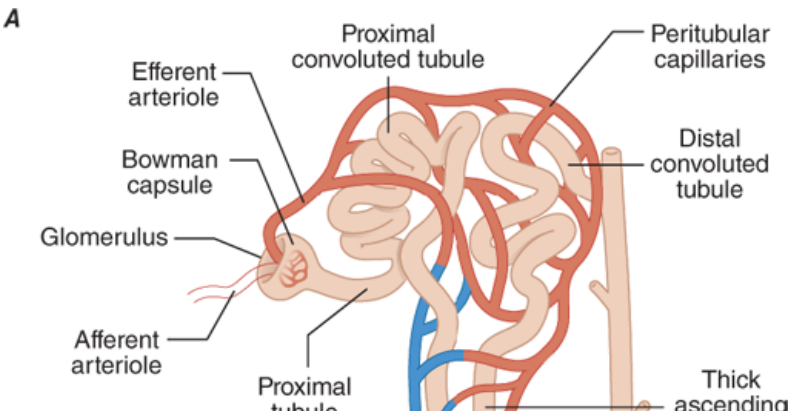
Each nephron is partitioned during embryologic development into a proximal tubule, descending and ascending limbs of the loop of Henle, distal tubule, and the collecting duct. These classic tubular segments build from subsegments lined by highly unique epithelia serving regional physiology. All nephrons have the same structural components, but there are two types whose structures depend on their location within the kidney. The majority of nephrons are cortical, with glomeruli located in the mid-to-outer cortex. Fewer nephrons are juxtamedullary, with glomeruli at the boundary of the cortex and outer medulla. Cortical nephrons have short loops of Henle, whereas juxtamedullary nephrons have long loops of Henle. There are critical differences in blood supply as well. The peritubular capillaries surrounding cortical nephrons are shared among adjacent nephrons. By contrast, juxtamedullary nephrons depend on individual capillaries called *vasa recta* that run alongside the long loops of Henle. Cortical nephrons perform most of the glomerular filtration because there are more of them and because their afferent arterioles are larger than their respective efferent arterioles. The juxtamedullary nephrons, with longer loops of Henle, create an osmotic gradient for concentrating urine. How developmental instructions specify the differentiation of all these unique epithelia among various tubular segments is still unknown.

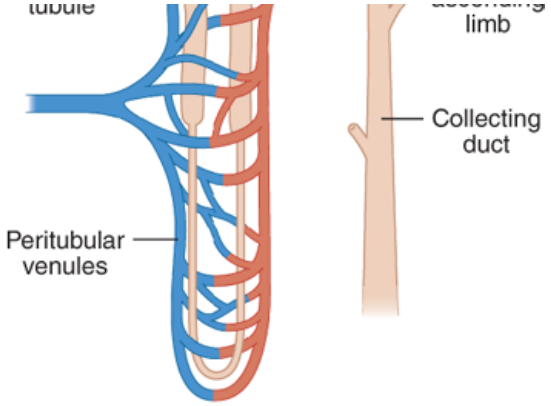
DETERMINANTS AND REGULATION OF GLOMERULAR FILTRATION

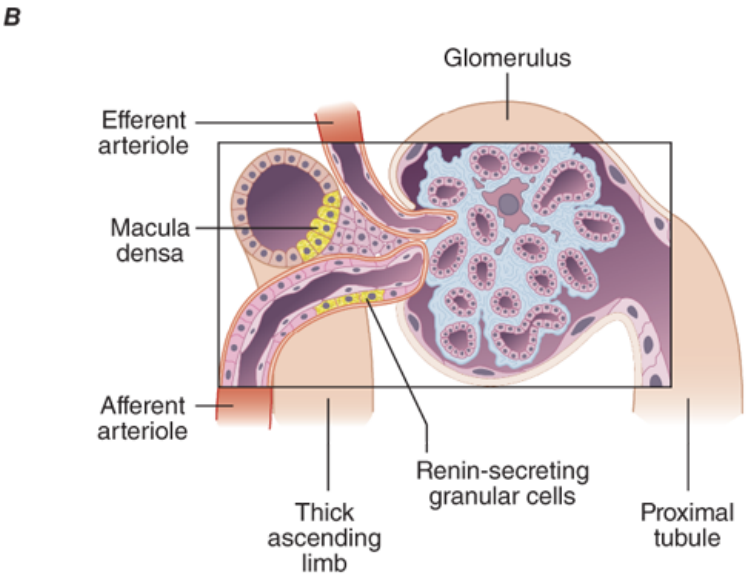
Renal blood flow normally drains ~20% of the cardiac output, or 1000 mL/min. Blood reaches each nephron through the afferent arteriole leading into a glomerular capillary where large amounts of fluid and solutes are filtered to form the tubular fluid. The distal ends of the glomerular capillaries coalesce to form an efferent arteriole leading to the first segment of a second capillary network (cortical peritubular capillaries or medullary vasa recta) surrounding the tubules (Fig. 303-2A). Thus, nephrons have two capillary beds arranged in a series separated by the efferent arteriole that regulates the hydrostatic pressure in both capillary beds. The distal capillaries empty into small venous branches that coalesce into larger veins to eventually form the renal vein.

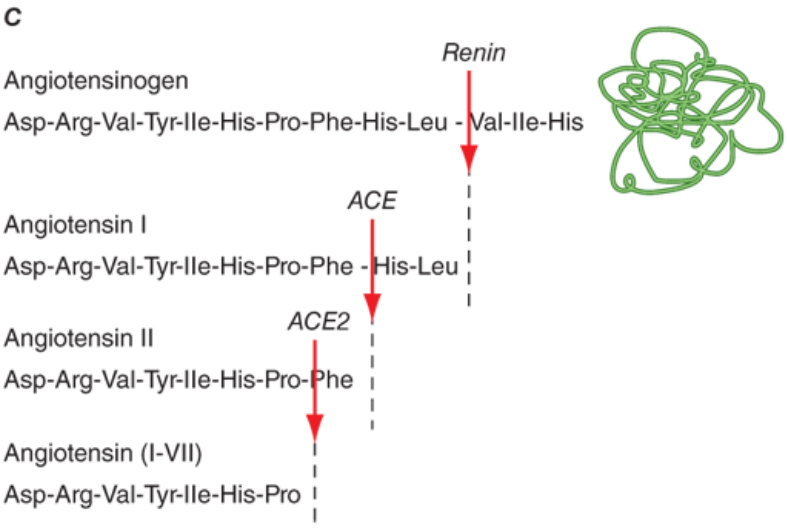
FIGURE 303-2

Renal microcirculation and the renin-angiotensin system. ***A.*** Diagram illustrating relationships of the nephron with glomerular and peritubular capillaries. ***B.*** Expanded view of the glomerulus with its juxtaglomerular apparatus including the macula densa and adjacent afferent arteriole. ***C.*** Proteolytic processing steps in the generation of angiotensins.









Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The hydrostatic pressure gradient across the glomerular capillary wall is the primary driving force for glomerular filtration. Oncotic pressure within the capillary lumen, determined by the concentration of unfiltered plasma proteins, partially offsets the hydrostatic pressure gradient and opposes filtration. As the oncotic pressure rises along the length of the glomerular capillary, the driving force for filtration falls to zero en route to the efferent arteriole. Approximately 20% of the renal plasma flow is filtered into Bowman space, and the ratio of glomerular filtration rate (GFR) to renal plasma flow determines the filtration fraction. Several factors, mostly hemodynamic, contribute to the regulation of filtration under physiologic conditions.

Although glomerular filtration is affected by renal artery pressure, this relationship is not linear across the range of physiologic blood pressures due to autoregulation of GFR. Autoregulation of glomerular filtration is the result of three major factors that modulate either afferent or efferent arteriolar tone: these include an autonomous vasoreactive (myogenic) reflex in the afferent arteriole, *tubuloglomerular feedback* (TGF), and angiotensin II-mediated vasoconstriction of the efferent arteriole. The myogenic reflex is a first line of defense against fluctuations in renal blood flow. Acute changes in renal perfusion pressure evoke reflex constriction or dilatation of the afferent arteriole in response to increased or decreased pressure, respectively. This phenomenon helps protect the glomerular capillary from sudden changes in systolic pressure.

TGF changes the rate of filtration and tubular flow by reflex vasoconstriction or dilatation of the afferent arteriole. TGF is mediated by specialized cells in the thick ascending limb of the loop of Henle called the *macula densa* that act as sensors of solute concentration and tubular fluid flow rate. With high tubular flow rates, a proxy for an inappropriately high filtration rate, there is increased solute delivery to the macula densa **(Fig. 303-2B)** that evokes vasoconstriction of the afferent arteriole causing GFR to return toward normal. One component of the soluble signal from the macula densa is adenosine triphosphate (ATP) released by the cells during increased NaCl reabsorption. ATP is metabolized in the extracellular space to generate adenosine, a potent vasoconstrictor of the afferent arteriole. During conditions associated with a fall in filtration rate, reduced solute delivery to the macula densa attenuates TGF, allowing afferent arteriolar dilatation and restoring GFR to normal levels. Angiotensin II and reactive oxygen species enhance, while nitric oxide (NO) blunts TGF.

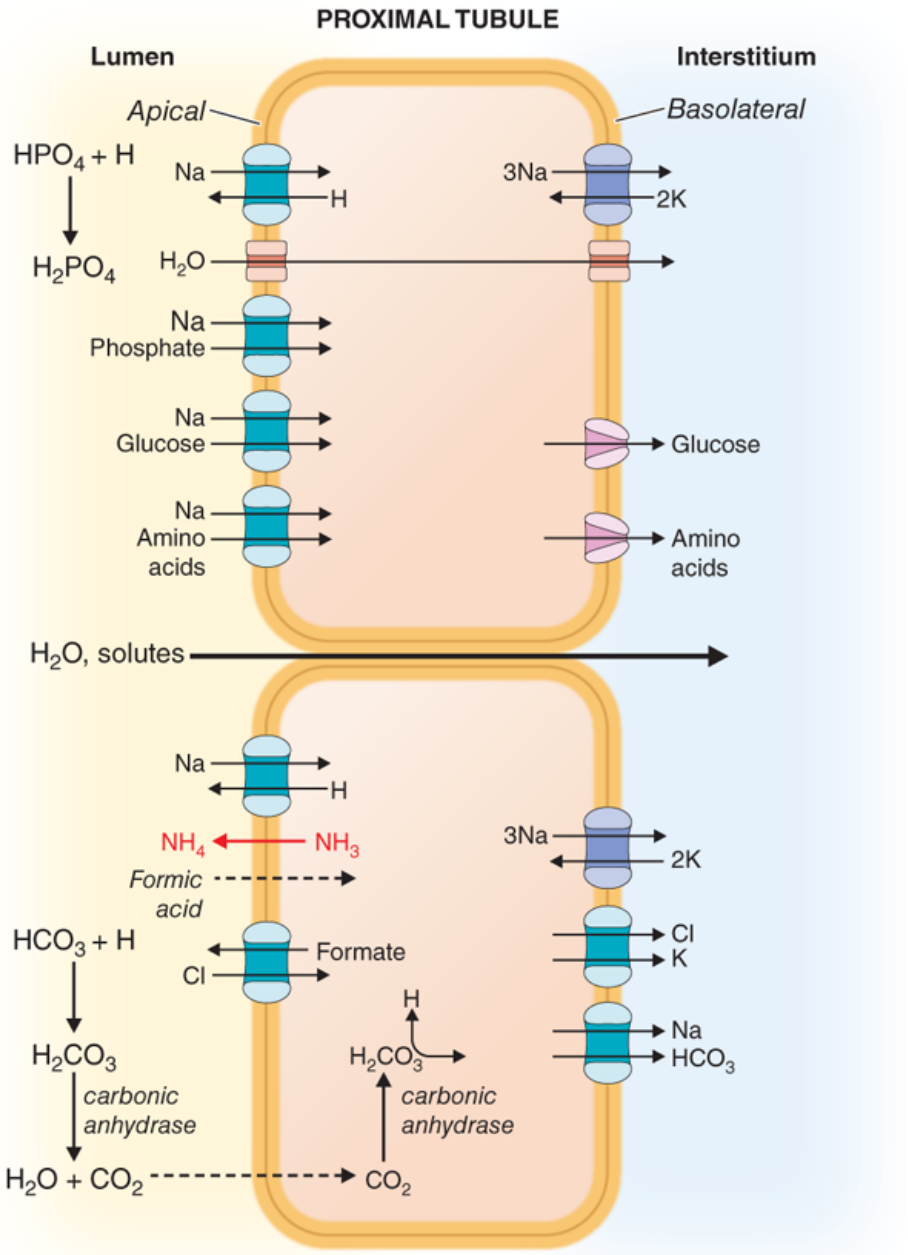
The third component underlying autoregulation of GFR involves angiotensin II. During states of reduced renal blood flow, renin is released from granular cells within the wall of the afferent arteriole near the macula densa in a region called the juxtaglomerular apparatus (Fig. 303-2B). Renin, a proteolytic enzyme, catalyzes the conversion of angiotensinogen to angiotensin I, which is subsequently converted to angiotensin II by angiotensin-converting enzyme (ACE) **(Fig. 303-2C)**. Angiotensin II evokes vasoconstriction of the efferent arteriole, and the resulting increased glomerular hydrostatic pressure elevates GFR to normal levels.

MECHANISMS OF RENAL TUBULAR TRANSPORT

The renal tubules are composed of highly differentiated epithelia that vary dramatically in morphology and function along the nephron **(Fig. 303-3)**. The cells lining the various tubular segments form monolayers connected to one another by a specialized region of the adjacent lateral membranes called the *tight junction*. Tight junctions form an occlusive barrier that separates the lumen of the tubule from the interstitial spaces surrounding the tubule and also apportions the cell membrane into discrete domains: the apical membrane facing the tubular lumen and the basolateral membrane facing the interstitium. This regionalization allows cells to allocate membrane proteins and lipids asymmetrically. Owing to this feature, renal epithelial cells are said to be *polarized*. The asymmetric assignment of membrane proteins, especially proteins mediating transport processes, provides the machinery for directional movement of fluid and solutes by the nephron.

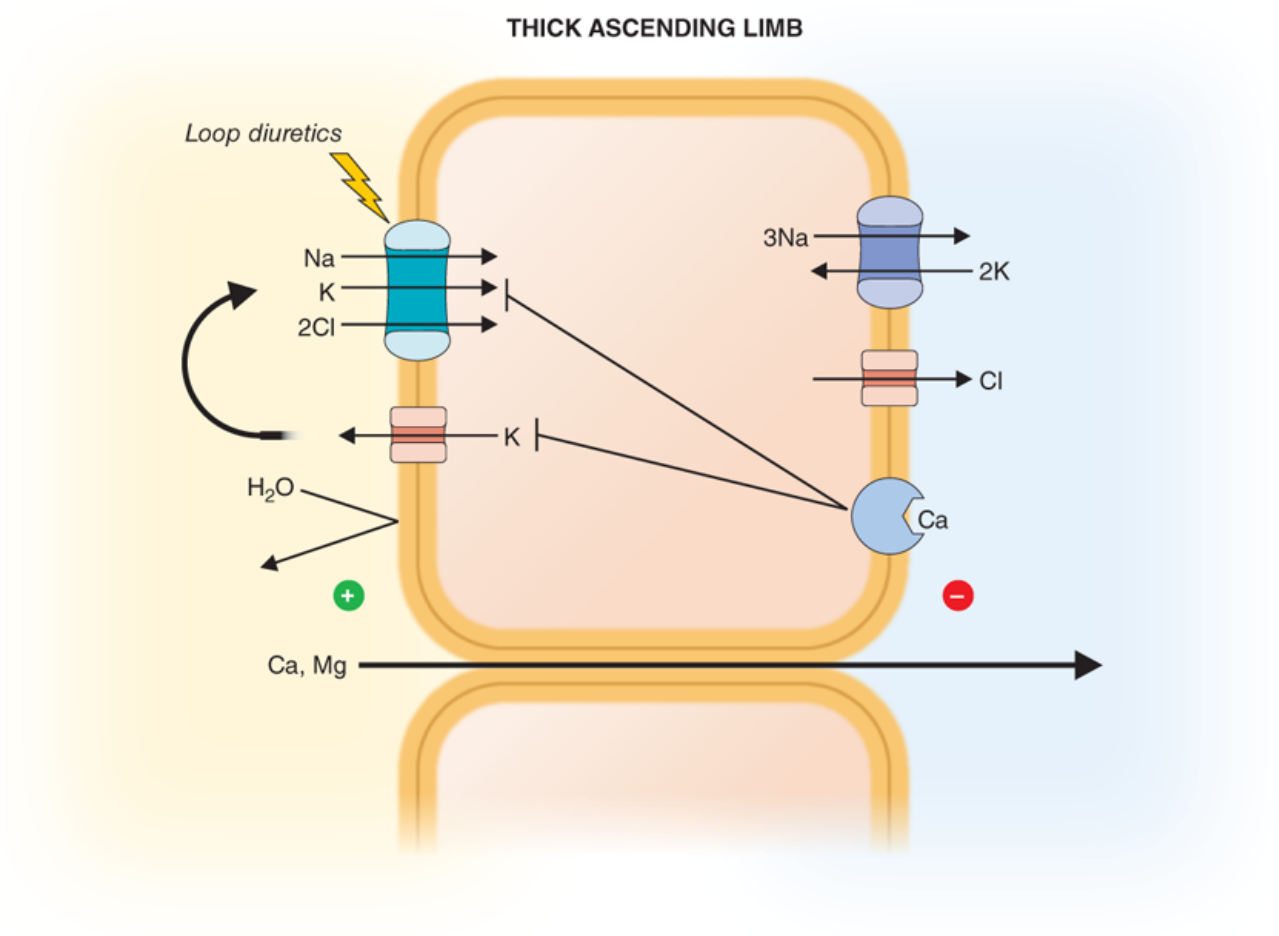
FIGURE 303-3

Transport activities of the major nephron segments. Representative cells from five major tubular segments are illustrated with the lumen side (apical membrane) facing left and interstitial side (basolateral membrane) facing right. ***A.*** Proximal tubular cells. ***B.*** Typical cell in the thick ascending limb of the loop of Henle. ***C.*** Distal convoluted tubular cell. ***D.*** Overview of entire nephron. ***E.*** Cortical collecting duct cells. ***F.*** Typical cell in the inner medullary collecting duct. The major membrane transporters, channels, and pumps are drawn with *arrows* indicating the direction of solute or water movement. For some events, the stoichiometry of transport is indicated by numerals preceding the solute. Targets for major diuretic agents are labeled. The actions of hormones are illustrated by *arrows with plus signs* for stimulatory effects and *lines with perpendicular ends* for inhibitory events. The *dashed line* indicates water impermeability of cell membranes in the thick ascending limb and distal convoluted tubule.



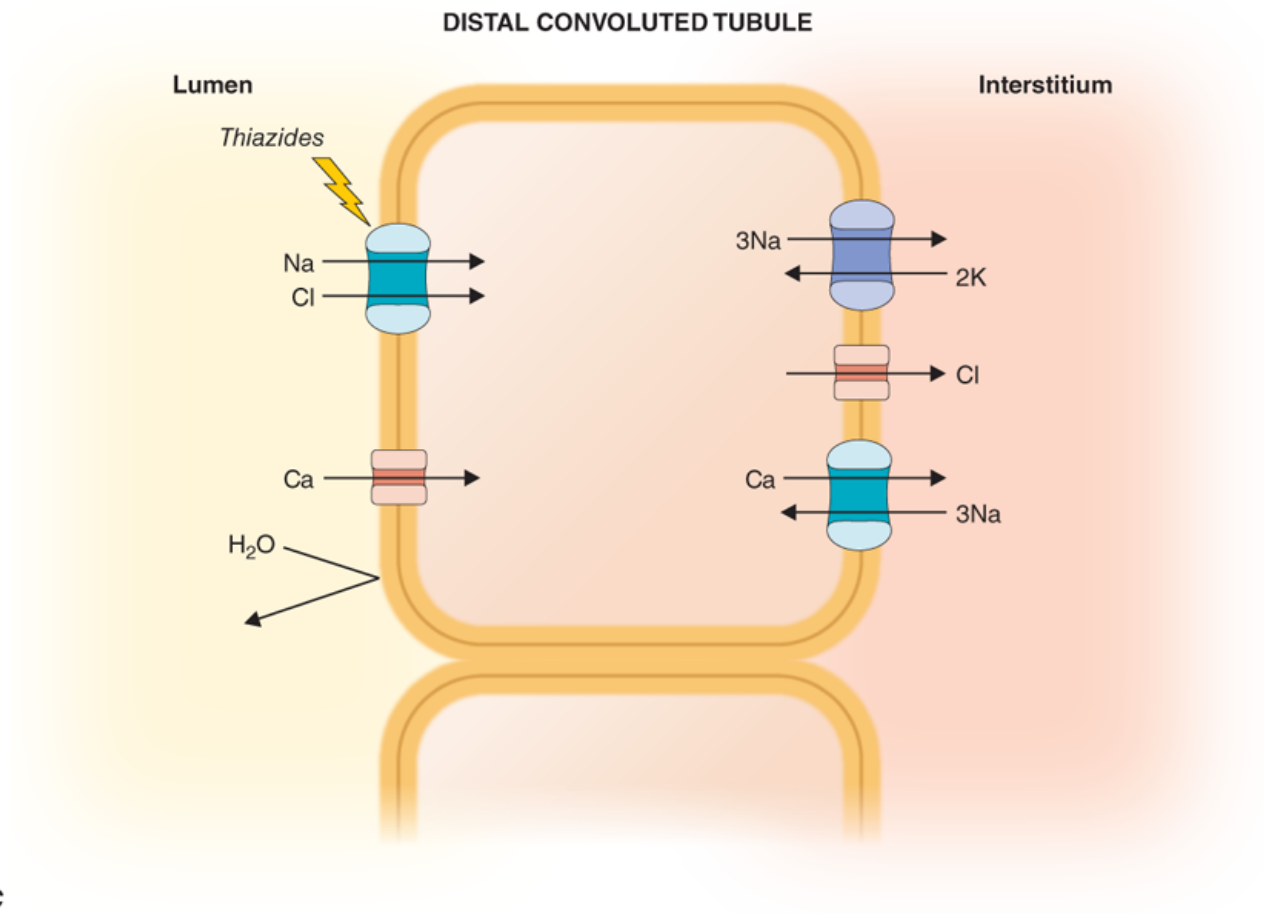
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



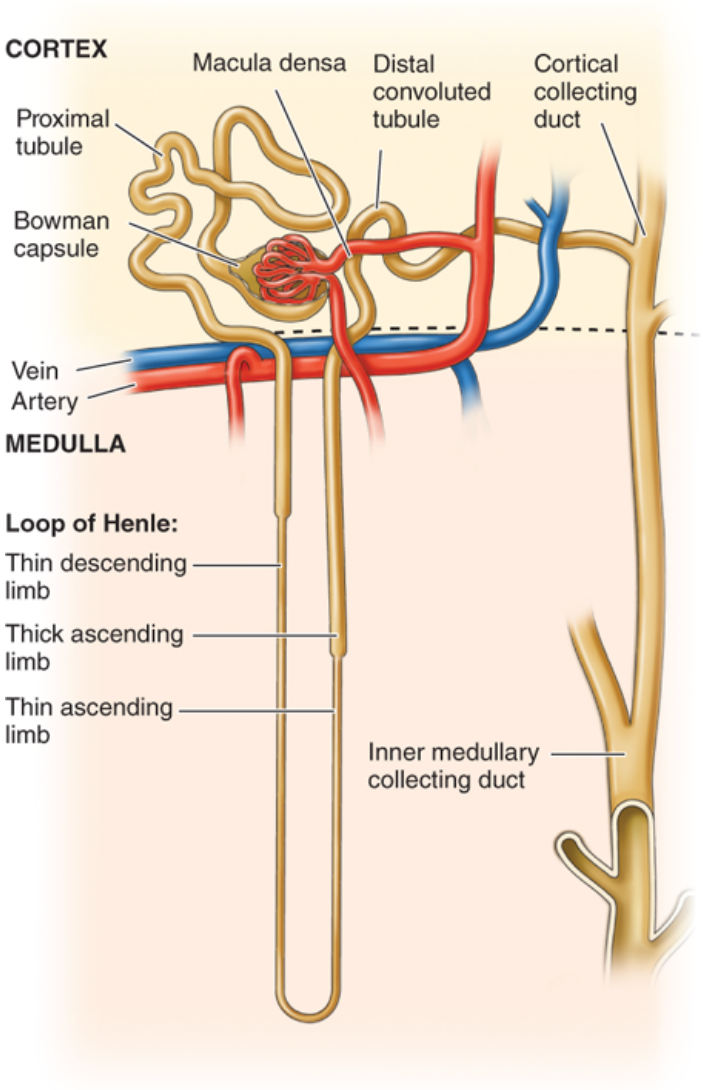
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



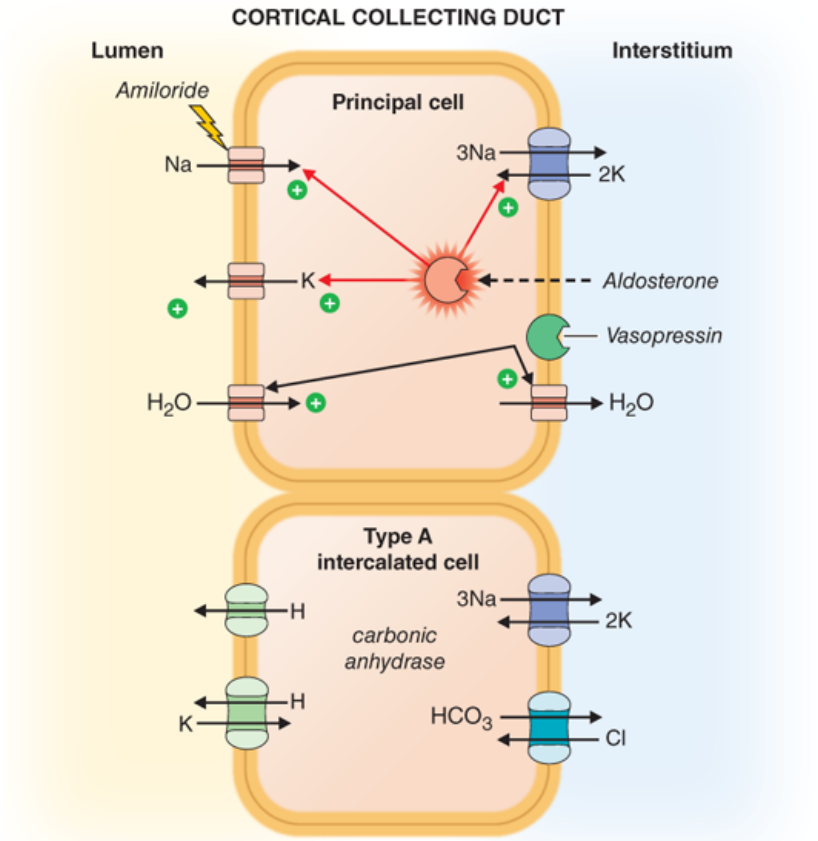
C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



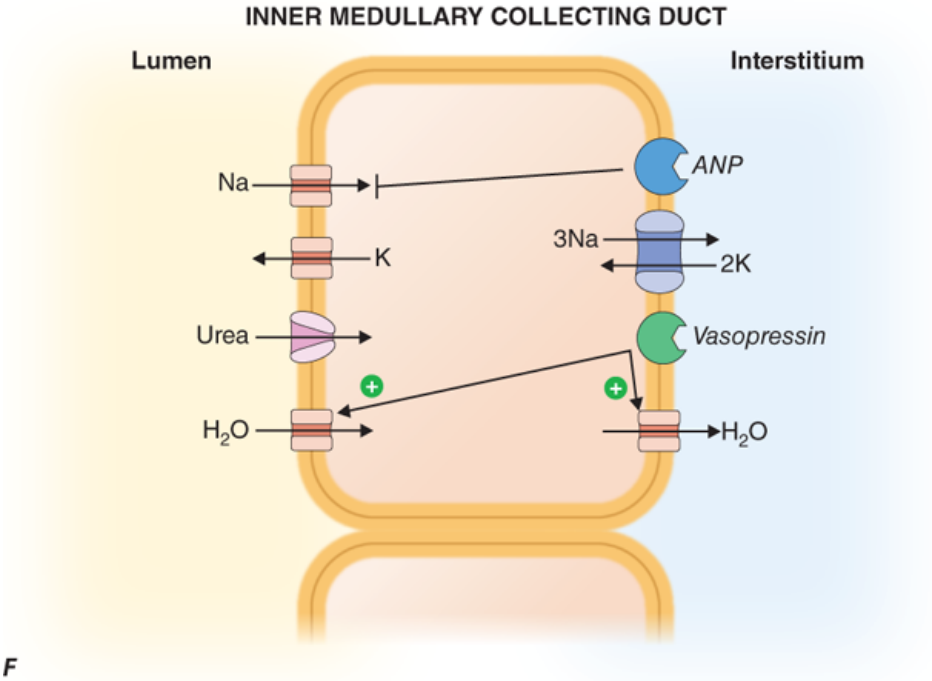
D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



E

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



F

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPITHELIAL SOLUTE TRANSPORT

There are two types of epithelial transport. Movement of fluid and solutes sequentially across the apical and basolateral cell membranes (or vice versa) mediated by transporters, channels, or pumps is called *cellular transport*. By contrast, movement of fluid and solutes through the narrow passageway between adjacent cells is called *paracellular transport*. Paracellular transport occurs through tight junctions, indicating that they are not completely "tight" or occlusive. Indeed, some epithelial cell layers allow rather robust paracellular transport to occur (*leaky epithelia*), whereas other epithelia have more restrictive tight junctions (*tight epithelia*). In addition, because the ability of ions to flow through the paracellular pathway determines the electrical resistance across the epithelial monolayer, leaky and tight epithelia are also referred to as low- or high-resistance epithelia, respectively. The proximal tubule contains leaky epithelia, whereas distal nephron segments, such as the collecting duct, contain tight epithelia. Leaky epithelia are most well suited for bulk fluid reabsorption, whereas tight epithelia allow for more refined control and regulation of transport.

MEMBRANE TRANSPORT

Cell membranes are composed of hydrophobic lipids that repel water and aqueous solutes. The movement of solutes and water across cell membranes is made possible by discrete classes of integral membrane proteins, including channels, pumps, and transporters. These different mechanisms mediate specific types of transport activities, including *active transport* (pumps), *passive transport* (channels), *facilitated diffusion* (transporters), and *secondary active transport* (cotransporters). Active transport requires metabolic energy generated by the hydrolysis of ATP. Active transport pumps are ion-translocating ATPases, including the ubiquitous Na^+/K^+-ATPase, the H^+-ATPases, and Ca^{2+}-ATPases. Active transport creates asymmetric ion concentrations across a cell membrane and can move ions against a chemical gradient. The potential energy stored in a concentration gradient of an ion such as Na^+ can be used to drive transport through other mechanisms (secondary active transport). Pumps are often *electrogenic*, meaning they can create an asymmetric distribution of electrostatic charges across the membrane and establish a voltage or membrane potential. The movement of solutes through a membrane protein by simple diffusion is called passive transport. This activity is mediated by channels created by selectively permeable membrane proteins, and it allows solute or water to move across a membrane driven by favorable *concentration gradients* or *electrochemical potential*. Facilitated diffusion is a specialized type of passive transport mediated by simple transporters called *carriers* or *uniporters*. For example, hexose transporters such as GLUT2 mediate glucose transport by tubular cells. These transporters are driven by the concentration gradient for glucose that is highest in extracellular fluids and lowest in the cytoplasm due to rapid metabolism. Many other transporters operate by translocating two or more ions/solutes in concert either in the same direction (*symporters* or *cotransporters*) or in opposite directions (*antiporters* or *exchangers*) across the cell membrane. The movement of two or more ions/solutes may produce no net change in the balance of electrostatic charges across the membrane (*electroneutral*), or a transport event may alter the balance of charges (*electrogenic*). Several inherited disorders of renal tubular solute and water transport occur as a consequence of mutations in genes encoding a variety of channels, transporter proteins, and their regulators (**Table 303-1**).

TABLE 303-1

Inherited Disorders Affecting Renal Tubular Ion and Solute Transport

DISEASE OR SYNDROME	GENE	OMIM[a]
Disorders Involving the Proximal Tubule		
Proximal renal tubular acidosis	Sodium bicarbonate cotransporter (*SLC4A4*, 4q21)	604278
Fanconi-Bickel syndrome	Glucose transporter, GLUT2 (*SLC2A2*, 3q26.2)	227810
Isolated renal glycosuria	Sodium glucose cotransporter (*SLC5A2*, 16p11.2)	233100
Cystinuria		
Type I	Cystine, dibasic and neutral amino acid transporter (*SLC3A1*, 2p16.3)	220100
Non-type I	Amino acid transporter, light subunit (*SLC7A9*, 19q13.1)	600918
Lysinuric protein intolerance	Amino acid transporter (*SLC7A7*, 4q11.2)	222700
Hartnup disorder	Neutral amino acid transporter (*SLC6A19*, 5p15.33)	34500
Hereditary hypophosphatemic rickets with hypercalcemia	Sodium phosphate cotransporter (*SLC34A3*, 9q34)	241530
Renal hypouricemia		
Type 1	Urate-anion exchanger (*SLC22A12*, 11q13)	220150
Type 2	Urate transporter, GLUT9 (*SLC2A9*, 4p16.1)	612076
Dent disease	Chloride channel, ClC-5 (*CLCN5*, Xp11.22)	300009
X-linked recessive nephrolithiasis with renal failure	Chloride channel, ClC-5 (*CLCN5*, Xp11.22)	310468
X-linked recessive hypophosphatemic rickets	Chloride channel, ClC-5 (*CLCN5*, Xp11.22)	307800
Disorders Involving the Loop of Henle		
Bartter's syndrome		
Type 1	Sodium, potassium chloride cotransporter (*SLC12A1*, 15q21.1)	241200
Type 2	Potassium channel, ROMK (*KCNJ1*, 11q24)	601678
Type 3	Chloride channel, ClC-Kb (*CLCNKB*, 1p36)	602023
with sensorineural deafness	Chloride channel accessory subunit, Barttin (*BSND*, 1p31)	602522
Autosomal dominant hypocalcemia with Bartter-like syndrome	Calcium-sensing receptor (*CASR*, 3q13.33)	601199
Familial hypocalciuric hypercalcemia	Calcium-sensing receptor (*CASR*, 3q13.33)	145980
Primary hypomagnesemia	Claudin-16 or paracellin-1 (*CLDN16* or *PCLN1*, 3q27)	248250

DISEASE OR SYNDROME	GENE	OMIM[a]
Isolated renal magnesium loss	Sodium potassium ATPase, γ_1-subunit (*ATP1G1*, 11q23)	154020
Disorders Involving the Distal Tubule and Collecting Duct		
Gitelman syndrome	Sodium chloride cotransporter (*SLC12A3*, 16q13)	263800
Primary hypomagnesemia with secondary hypocalcemia	Melastatin-related transient receptor potential cation channel 6 (*TRPM6*, 9q22)	602014
Pseudoaldosteronism (Liddle's syndrome)	Epithelial sodium channel β and γ subunits (*SCNN1B*, *SCNN1G*, 16p12.1)	177200
Recessive pseudohypoaldosteronism type 1	Epithelial sodium channel, α, β, and γ subunits (*SCNN1A*, 12p13; *SCNN1B*, *SCNN1G*, 16pp12.1)	264350
Pseudohypoaldosteronism type 2 (Gordon's hyperkalemia-hypertension syndrome)	Kinases WNK-1, WNK-4 (*WNK1*, 12p13; *WNK4*, 17q21.31)	145260
X-linked nephrogenic diabetes insipidus	Vasopressin V2 receptor (AVPR2, Xq28)	304800
Nephrogenic diabetes insipidus (autosomal)	Water channel, aquaporin-2 (*AQP2*, 12q13)	125800
Distal renal tubular acidosis		
autosomal dominant	Anion exchanger-1 (*SLC4A1*, 17q21.31)	179800
autosomal recessive	Anion exchanger-1 (*SLC4A1*, 17q21.31)	602722
with neural deafness	Proton ATPase, β1 subunit (*ATP6V1B1*, 2p13.3)	192132
with normal hearing	Proton ATPase, 116-kD subunit (*ATP6V0A4*, 7q34)	602722

[a]Online Mendelian Inheritance in Man database (*http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/Omim*).

SEGMENTAL NEPHRON FUNCTIONS

Each anatomic segment of the nephron has unique characteristics and specialized functions enabling selective transport of solutes and water (Fig. 303-3). Through sequential events of reabsorption and secretion along the nephron, tubular fluid is progressively conditioned into urine. Knowledge of the major tubular mechanisms responsible for solute and water transport is critical for understanding hormonal regulation of kidney function and the pharmacologic manipulation of renal excretion.

PROXIMAL TUBULE

The proximal tubule is responsible for reabsorbing ~60% of filtered NaCl and water, as well as ~90% of filtered bicarbonate and most critical nutrients such as glucose and amino acids. The proximal tubule uses both cellular and paracellular transport mechanisms. The apical membrane of proximal tubular cells has an expanded surface area available for reabsorptive work created by a dense array of microvilli called the *brush border*, and leaky tight junctions enable high-capacity fluid reabsorption.

Solute and water pass through these tight junctions to enter the lateral intercellular space where absorption by the peritubular capillaries occurs. Bulk fluid reabsorption by the proximal tubule is driven by high oncotic pressure and low hydrostatic pressure within the peritubular

capillaries. Cellular transport of most solutes by the proximal tubule is coupled to the Na^+ concentration gradient established by the activity of a basolateral Na^+/K^+-ATPase (**Fig. 303-3A**). This active transport mechanism maintains a steep Na^+ gradient by keeping intracellular Na^+ concentrations low. Solute reabsorption from the tubular lumen is coupled to the Na^+ gradient by Na^+-dependent transporters such as Na^+-glucose and Na^+-phosphate cotransporters present in apical membranes. In addition to the paracellular route, water reabsorption also occurs through the cellular pathway enabled by constitutively active water channels (aquaporin-1) present on both apical and basolateral membranes.

Proximal tubular cells reclaim nearly all filtered bicarbonate by a mechanism dependent on carbonic anhydrases. Filtered bicarbonate is first titrated by protons delivered to the lumen mainly by Na^+/H^+ exchange. The resulting carbonic acid (H_2CO_3) is metabolized by brush border carbonic anhydrase to water and carbon dioxide. Dissolved carbon dioxide then diffuses into the cell, where it is enzymatically hydrated by cytoplasmic carbonic anhydrase to re-form carbonic acid. Finally, intracellular carbonic acid dissociates into free protons and bicarbonate anions, and bicarbonate exits the cell through a basolateral Na^+/HCO_3^- cotransporter. This process is saturable, which can result in urinary bicarbonate excretion when plasma levels exceed the physiologically normal range (24–26 meq/L). Carbonic anhydrase inhibitors such as acetazolamide, a class of weak diuretic agents, block proximal tubule bicarbonate reabsorption and are useful for alkalinizing the urine.

The proximal tubule contributes to acid secretion by two mechanisms involving the titration of the urinary buffers ammonia (NH_3) and phosphate. Renal NH_3 is produced by glutamine metabolism in the proximal tubule. Subsequent diffusion of NH_3 out of the proximal tubular cell enables trapping of H^+, which is secreted by Na^+/H^+ exchange, in the lumen as ammonium ion (NH_4^+). Cellular K^+ levels inversely modulate proximal tubular ammoniagenesis, and in the setting of high serum K^+ from hypoaldosteronism, reduced ammoniagenesis promotes type IV renal tubular acidosis. Filtered hydrogen phosphate ion (HPO_4^{2-}) is also titrated in the proximal tubule by secreted H^+ to form $H_2PO_4^-$, and this reaction constitutes a major component of the urinary buffer referred to as titratable acid. Most filtered phosphate ion is reabsorbed by the proximal tubule through a sodium-coupled cotransport process that is regulated by parathyroid hormone (PTH).

Chloride is poorly reabsorbed throughout the first segment of the proximal tubule, and a rise in Cl^- concentration counterbalances the removal of bicarbonate anion from tubular fluid. In later proximal tubular segments, cellular Cl^- reabsorption is initiated by apical exchange of cellular formate for higher luminal concentrations of Cl^-. Once in the lumen, formate anions are titrated by H^+ (provided by Na^+/H^+ exchange) to generate neutral formic acid, which can diffuse passively across the apical membrane back into the cell where it dissociates a proton and is recycled. Basolateral Cl^- exit is mediated by a K^+/Cl^- cotransporter.

Reabsorption of glucose is nearly complete by the end of the proximal tubule. Cellular transport of glucose is mediated by apical Na^+-glucose cotransport coupled with basolateral, facilitated diffusion by a glucose transporter. This process is also saturable, leading to glycosuria when plasma levels exceed 180–200 mg/dL, as seen in untreated diabetes mellitus.

The proximal tubule possesses specific transporters capable of secreting a variety of organic acids (carboxylate anions) and bases (mostly primary amine cations). Organic anions transported by these systems include urate, dicarboxylic acid anions (succinate), ketoacid anions, and several protein-bound drugs not filtered at the glomerulus (penicillins, cephalosporins, and salicylates). Probenecid inhibits renal organic anion secretion and can be clinically useful for raising plasma concentrations of certain drugs like penicillin and oseltamivir. Organic cations secreted by the proximal tubule include various biogenic amine neurotransmitters (dopamine, acetylcholine, epinephrine, norepinephrine, and histamine) and creatinine. The ATP-dependent transporter P-glycoprotein is highly expressed in brush border membranes and secretes several medically important drugs, including cyclosporine, digoxin, tacrolimus, and various cancer chemotherapeutic agents. Certain drugs like cimetidine and trimethoprim compete with endogenous compounds for transport by the organic cation pathways. Although these drugs elevate serum creatinine levels, there is no actual change in GFR in this setting.

Calcium and phosphorus homeostasis depends upon normal functioning of the proximal tubule. Approximately 60–70% of filtered calcium and ~85% of filtered phosphorus (in the form of inorganic phosphate) are reabsorbed by the proximal tubule. Whereas calcium reabsorption is mostly by passive diffusion through the paracellular route, phosphate reabsorption is mediated by sodium-coupled cotransport. In addition to direct reabsorption, the proximal tubule contributes to systemic mineral balance by participating in specific endocrine pathways. Circulating 25-hydroxy vitamin D (calcidiol) is bioactivated by proximal tubular 1α-hydroxylase to produce 1,25-di-hydroxy vitamin D (calcitriol), the most active form of the hormone, that acts on the small intestine to promote calcium absorption. Phosphate balance is affected by circulating fibroblast growth hormone 23 (FGF23), a bone-derived hormone that interacts with its receptor (FGFR1) and co-receptor (Klotho) on proximal tubular cells to suppress sodium-phosphate cotransport and promote renal phosphate excretion. PTH stimulates proximal tubular 1α-

hydroxylation of vitamin D while it suppresses sodium-phosphate cotransport. Derangements in PTH and FGF23 account for abnormal calcium and phosphate balance in chronic kidney disease.

The proximal tubule, through distinct classes of Na^+-dependent and Na^+-independent transport systems, reabsorbs amino acids efficiently. These transporters are specific for different groups of amino acids. For example, cystine, lysine, arginine, and ornithine are transported by a system comprising two proteins encoded by the *SLC3A1* and *SLC7A9* genes. Mutations in either *SLC3A1* or *SLC7A9* impair reabsorption of these amino acids and cause the disease cystinuria. Peptide hormones, such as insulin and growth hormone, β_2-microglobulin, albumin, and other small proteins, are taken up by the proximal tubule through a process of absorptive endocytosis and are degraded in acidified endocytic lysosomes. Acidification of these vesicles depends on a vacuolar H^+-ATPase and Cl^- channel. Impaired acidification of endocytic vesicles because of mutations in a Cl^- channel gene (*CLCN5*) causes low-molecular-weight proteinuria in Dent disease.

LOOP OF HENLE

The loop of Henle consists of three major segments: descending thin limb, ascending thin limb, and ascending thick limb. These divisions are based on cellular morphology and anatomic location, but also correlate with specialization of function. Approximately 15–25% of filtered NaCl is reabsorbed in the loop of Henle, mainly by the thick ascending limb. The loop of Henle has an important role in urinary concentration by contributing to the generation of a hypertonic medullary interstitium in a process called *countercurrent multiplication*. The loop of Henle is the site of action for the most potent class of diuretic agents (loop diuretics) and also contributes to reabsorption of calcium and magnesium ions.

The descending thin limb is highly water permeable owing to dense expression of constitutively active aquaporin-1 water channels. By contrast, water permeability is negligible in the ascending limb. In the thick ascending limb, there is a high level of secondary active NaCl transport enabled by the $Na^+/K^+/2Cl^-$ cotransporter on the apical membrane in series with basolateral Cl^- channels and Na^+/K^+-ATPase (**Fig. 303-3B**). The $Na^+/K^+/2Cl^-$ cotransporter is the primary target for loop diuretics. Tubular fluid K^+ is the limiting substrate for this cotransporter (tubular concentration of K^+ is similar to plasma, about 4 meq/L), but transporter activity is maintained by K^+ recycling through an apical potassium channel. The cotransporter also enables reabsorption of NH_4^+ in lieu of K^+, and this leads to accumulation of both NH_4^+ and NH_3 in the medullary interstitium. An inherited disorder of the thick ascending limb, Bartter's syndrome, also results in a salt-wasting renal disease associated with hypokalemia and metabolic alkalosis. Loss-of-function mutations in one of five distinct genes encoding components of the $Na^+/K^+/2Cl^-$ cotransporter (*NKCC2*), apical K^+ channel (*KCNJ1*), basolateral Cl^- channel (*CLCNKB*, *BSND*), or calcium-sensing receptor (*CASR*) can cause Bartter's syndrome.

Potassium recycling also contributes to a positive electrostatic charge in the lumen relative to the interstitium that promotes divalent cation (Mg^{2+} and Ca^{2+}) reabsorption through a paracellular pathway. A Ca^{2+}-sensing, G-protein-coupled receptor (CaSR) on basolateral membranes regulates NaCl reabsorption in the thick ascending limb through dual signaling mechanisms using either cyclic AMP or eicosanoids. This receptor enables a steep relationship between plasma Ca^{2+} levels and renal Ca^{2+} excretion. Loss-of-function mutations in CaSR cause familial hypercalcemic hypocalciuria because of a blunted response of the thick ascending limb to extracellular Ca^{2+}. Mutations in *CLDN16* encoding paracellin-1, a transmembrane protein located within the tight junction complex, leads to familial hypomagnesemia with hypercalciuria and nephrocalcinosis, suggesting that the ion conductance of the paracellular pathway in the thick limb is regulated.

The loop of Henle contributes to urine-concentrating ability by establishing a *hypertonic medullary interstitium* that promotes water reabsorption by the downstream inner medullary collecting duct. *Countercurrent multiplication* produces a hypertonic medullary interstitium using two countercurrent systems: the loop of Henle (opposing descending and ascending limbs) and the vasa recta (medullary peritubular capillaries enveloping the loop). The countercurrent flow in these two systems helps maintain the hypertonic environment of the inner medulla, but NaCl reabsorption by the thick ascending limb is the primary initiating event. Reabsorption of NaCl without water dilutes the tubular fluid and adds new osmoles to medullary interstitial fluid. Because the descending thin limb is highly water permeable, osmotic equilibrium occurs between the descending limb tubular fluid and the interstitial space, leading to progressive solute trapping in the inner medulla. Maximum medullary interstitial osmolality also requires partial recycling of urea from the collecting duct.

DISTAL CONVOLUTED TUBULE

The distal convoluted tubule reabsorbs ~5% of the filtered NaCl. This segment is composed of a tight epithelium with little water permeability. The major NaCl-transporting pathway uses an apical membrane, electroneutral thiazide-sensitive Na^+/Cl^- cotransporter in tandem with

basolateral Na^+/K^+-ATPase and Cl^- channels (**Fig. 303-3C**). Apical Ca^{2+}-selective channels (TRPV5) and basolateral Na^+/Ca^{2+} exchange mediate calcium reabsorption in the distal convoluted tubule. Ca^{2+} reabsorption is inversely related to Na^+ reabsorption and is stimulated by PTH. Blocking apical Na^+/Cl^- cotransport will reduce intracellular Na^+, favoring increased basolateral Na^+/Ca^{2+} exchange and passive apical Ca^{2+} entry. Loss-of-function mutations of *SLC12A3* encoding the apical Na^+/Cl^- cotransporter cause Gitelman syndrome, a salt-wasting disorder associated with hypokalemic alkalosis and hypocalciuria. Mutations in genes encoding WNK kinases, WNK-1 and WNK-4, cause pseudohypoaldosteronism type II (Gordon syndrome) characterized by familial hypertension with hyperkalemia. WNK kinases influence the activity of several tubular ion transporters. Mutations in this disorder lead to overactivity of the apical Na^+/Cl^- cotransporter in the distal convoluted tubule as the primary stimulus for increased salt reabsorption, extracellular volume expansion, and hypertension. Hyperkalemia may be caused by diminished activity of apical K^+ channels in the collecting duct, a primary route for K^+ secretion. Mutations in *TRPM6* encoding Mg^{2+} permeable ion channels also cause familial hypomagnesemia with hypocalcemia. A molecular complex of TRPM6 and TRPM7 proteins is critical for Mg^{2+} reabsorption in the distal convoluted tubule.

COLLECTING DUCT

The collecting duct modulates the final composition of urine. The two major divisions, the cortical collecting duct and inner medullary collecting duct, contribute to reabsorbing ~4–5% of filtered Na^+ and are important for hormonal regulation of salt and water balance. Cells in both segments of the collecting duct express vasopressin-regulated water channels (aquaporin-2 on the apical membrane, aquaporin-3 and -4 on the basolateral membrane). The antidiuretic hormone vasopressin binds to the V2 receptor on the basolateral membrane and triggers an intracellular signaling cascade through G-protein-mediated activation of adenylyl cyclase, resulting in an increase in the cellular levels of cyclic AMP. This signaling cascade stimulates the insertion of water channels into the apical membrane of collecting duct cells to promote increased water permeability. This increase in permeability enables water reabsorption and production of concentrated urine. In the absence of vasopressin, collecting duct cells are water impermeable, and urine remains dilute.

The cortical collecting duct contains *high-resistance epithelia* with two cell types. Principal cells are the main water, Na^+-reabsorbing, and K^+-secreting cells, and the site of action of aldosterone, K^+-sparing diuretics, and mineralocorticoid receptor antagonists such as spironolactone and eplerenone. The other cells are type A and B intercalated cells. Type A intercalated cells mediate acid secretion and bicarbonate reabsorption also under the influence of aldosterone. Type B intercalated cells mediate bicarbonate secretion and acid reabsorption.

Virtually all transport is mediated through the cellular pathway for both principal cells and intercalated cells. In principal cells, passive apical Na^+ entry occurs through an amiloride-sensitive, epithelial Na^+ channel (ENaC) with basolateral exit mediated by the Na^+/K^+-ATPase (**Fig. 303-3D**). This Na^+ reabsorptive process is tightly regulated by aldosterone and is physiologically activated by a variety of proteolytic enzymes that cleave extracellular domains of ENaC; plasmin in the tubular fluid of nephrotic patients, for example, activates ENaC, leading to sodium retention. Aldosterone enters the cell across the basolateral membrane, binds to a cytoplasmic mineralocorticoid receptor, and then translocates into the nucleus, where it modulates gene transcription, resulting in increased Na^+ reabsorption and K^+ secretion. Activating mutations in ENaC increase Na^+ reclamation and produce hypokalemia, hypertension, and metabolic alkalosis (Liddle's syndrome). The potassium-sparing diuretics amiloride and triamterene block ENaC, causing reduced Na^+ reabsorption.

Principal cells secrete K^+ through an apical membrane potassium channel. Several forces govern the secretion of K^+. Most importantly, the high intracellular K^+ concentration generated by Na^+/K^+-ATPase creates a favorable concentration gradient for K^+ secretion into tubular fluid. With reabsorption of Na^+ without an accompanying anion, the tubular lumen becomes negative relative to the cell interior, creating a favorable electrical gradient for secretion of potassium. When Na^+ reabsorption is blocked, the electrical component of the driving force for K^+ secretion is blunted, and this explains lack of excess urinary K^+ loss during treatment with potassium-sparing diuretics or mineralocorticoid receptor antagonists. K^+ secretion is also promoted by aldosterone actions that increase regional Na^+ transport, which favor more lumen electronegativity, and by increasing the number and activity of potassium channels. Fast tubular fluid flow rates that occur during volume expansion or diuretics acting "upstream" of the cortical collecting duct also increase K^+ secretion, as does the presence of relatively nonreabsorbable anions (including bicarbonate and semisynthetic penicillins) that contribute to the lumen-negative potential. Off-target effects of certain antibiotics, such as trimethoprim and pentamidine, block ENaCs and predispose to hyperkalemia, especially when renal K^+ handling is impaired for other reasons. Principal cells, as described below, also participate in water reabsorption by increased water permeability in response to vasopressin.

Intercalated cells do not participate in Na^+ reabsorption but, instead, mediate acid-base secretion. These cells perform two types of transport: active H^+ transport mediated by H^+-ATPase (proton pump), and Cl^-/HCO_3^- exchange. Intercalated cells arrange the two transport mechanisms on opposite membranes to enable either acid or base secretion. Type A intercalated cells have an apical proton pump that mediates acid secretion and a basolateral Cl^-/HCO_3^- anion exchanger for bicarbonate reabsorption (Fig. 303-3E). Aldosterone increases the number of H^+-ATPase pumps, sometimes contributing to the development of metabolic alkalosis. Secreted H^+ is buffered by NH_3 that has diffused into the collecting duct lumen from the surrounding interstitium. By contrast, type B intercalated cells have the Cl^-/HCO_3^- exchanger on the apical membrane to mediate bicarbonate secretion while the proton pump resides on the basolateral membrane to enable acid reabsorption. Under conditions of acidemia, the kidney preferentially uses type A intercalated cells to secrete the excess H^+ and generate more HCO_3^-. The opposite is true in states of bicarbonate excess with alkalemia where the type B intercalated cells predominate. An extracellular protein called *hensin* mediates this adaptation.

Inner medullary collecting duct cells share many similarities with principal cells of the cortical collecting duct. They have apical Na^+ and K^+ channels that mediate Na^+ reabsorption and K^+ secretion, respectively (Fig. 303-3F). Sodium reabsorption by inner medullary collecting duct cells is also inhibited by the natriuretic peptides called *atrial natriuretic peptide* or *renal natriuretic peptide* (urodilatin); the same gene encodes both peptides but uses different posttranslational processing of a common preprohormone to generate different proteins. Atrial natriuretic peptides are secreted by atrial myocytes in response to volume expansion, whereas urodilatin is secreted by renal tubular epithelia. Natriuretic peptides interact with either apical (urodilatin) or basolateral (atrial natriuretic peptides) receptors on inner medullary collecting duct cells to stimulate guanylyl cyclase and increase levels of cytoplasmic cGMP. This effect in turn reduces the activity of the apical Na^+ channel in these cells and attenuates net Na^+ reabsorption, producing natriuresis.

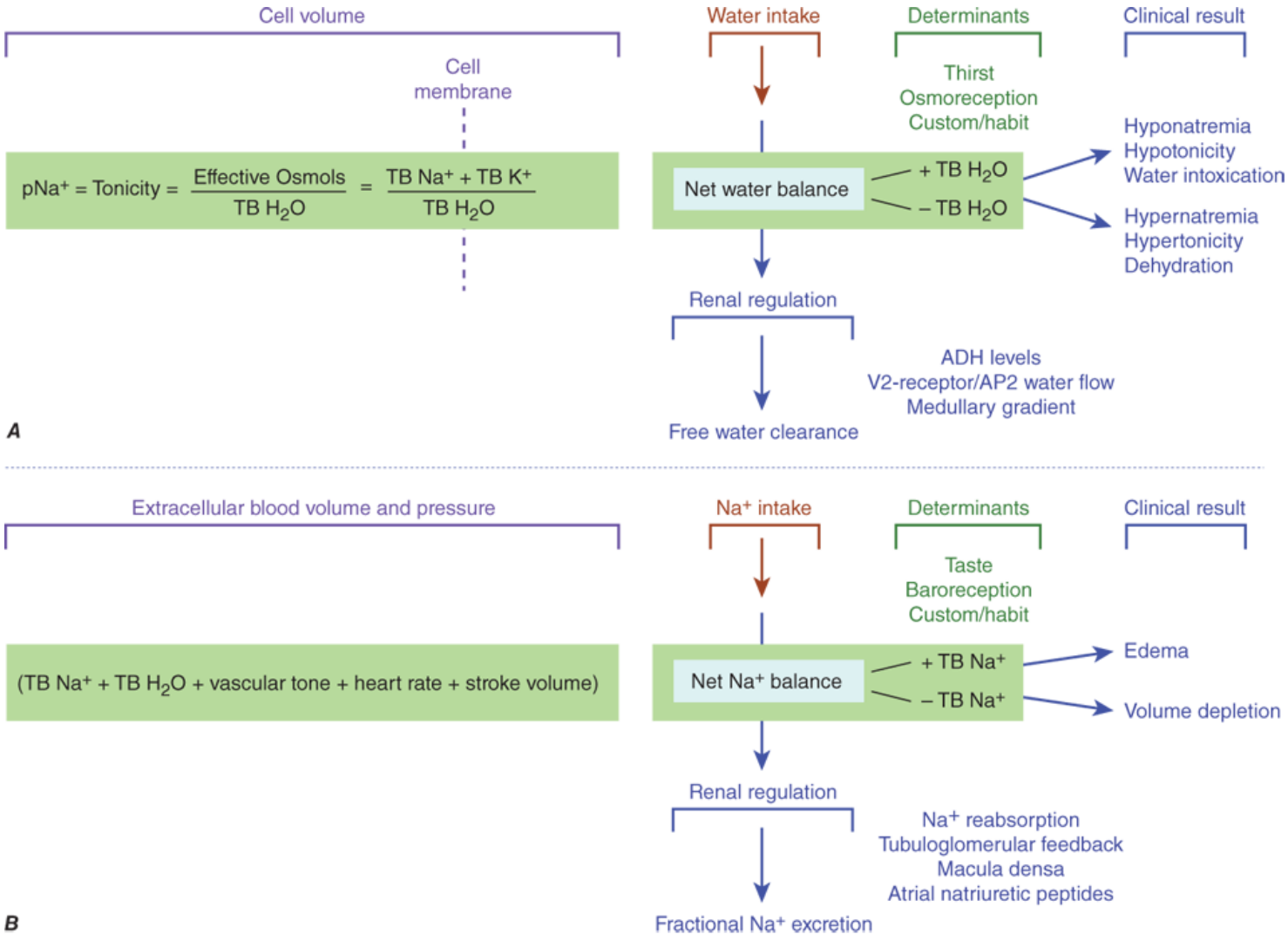
The inner medullary collecting duct transports urea out of the lumen, returning urea to the interstitium, where it contributes to the hypertonicity of the medullary interstitium. Urea is recycled by diffusing from the interstitium into the descending and ascending limbs of the loop of Henle.

HORMONAL REGULATION OF SODIUM AND WATER BALANCE

The balance of solute and water in the body is determined by the amounts ingested, distributed to various fluid compartments, and excreted by skin, bowel, and kidneys. *Tonicity*, the osmolar state determining the volume behavior of cells in a solution, is regulated by water balance **(Fig. 303-4A)**, and *extracellular blood volume* is regulated by Na^+ balance **(Fig. 303-4B)**. The kidney is a critical modulator of both physiologic processes.

FIGURE 303-4

Determinants of sodium and water balance. ***A.*** Plasma Na^+ concentration is a surrogate marker for plasma tonicity, the volume behavior of cells in a solution. Tonicity is determined by the number of effective osmoles in the body divided by the total body H_2O (TB H_2O), which translates simply into the total body Na (TB Na^+) and anions outside the cell separated from the total body K (TB K^+) inside the cell by the cell membrane. Net water balance is determined by the integrated functions of thirst, osmoreception, Na reabsorption, vasopressin release, and the strength of the medullary gradient in the kidney, keeping tonicity within a narrow range of osmolality around 280 mosmol/L. When water metabolism is disturbed and total body water increases, hyponatremia, hypotonicity, and water intoxication occur; when total body water decreases, hypernatremia, hypertonicity, and dehydration occur. ***B.*** Extracellular blood volume and pressure are an integrated function of total body Na^+ (TB Na^+), total body H_2O (TB H_2O), vascular tone, heart rate, and stroke volume that modulates volume and pressure in the vascular tree of the body. This extracellular blood volume is determined by net Na balance under the control of taste, baroreception, habit, Na^+ reabsorption, macula densa/tubuloglomerular feedback, and natriuretic peptides. When Na^+ metabolism is disturbed and total body Na^+ increases, edema occurs; when total body Na^+ is decreased, volume depletion occurs. ADH, antidiuretic hormone; AQP2, aquaporin-2.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

WATER BALANCE

Tonicity depends on the variable concentration of *effective osmoles* inside and outside the cell causing water to move in either direction across its membrane. Classic effective osmoles, like Na^+, K^+, and their anions, are solutes trapped on either side of a cell membrane, where they collectively partition and obligate water to move and find equilibrium in proportion to retained solute; Na^+/K^+-ATPase keeps most K^+ inside cells and most Na^+ outside. Normal tonicity (~280 mosmol/L) is rigorously defended by osmoregulatory mechanisms that control water balance to protect tissues from inadvertent *dehydration* (cell shrinkage) or *water intoxication* (cell swelling), both of which are deleterious to cell function (Fig. 303-4A).

The mechanisms that control osmoregulation are distinct from those governing extracellular volume, although there is some shared physiology in both processes. While cellular concentrations of K^+ have a determinant role in any level of tonicity, the routine surrogate marker for assessing clinical tonicity is the concentration of serum Na^+. Any reduction in total body water, which raises the Na^+ concentration, triggers a brisk sense of thirst and conservation of water by decreasing renal water excretion mediated by release of vasopressin from the posterior pituitary. Conversely, a decrease in plasma Na^+ concentration triggers an increase in renal water excretion by suppressing the secretion of vasopressin. Whereas all cells expressing mechanosensitive TRPV1, 2, or 4 channels, among potentially other sensors, respond to changes in tonicity by altering their volume and Ca^{2+} concentration, only $TRPV^+$ neuronal cells connected to the organum vasculosum of the lamina terminalis are *osmoreceptive*. Only these cells, because of their neural connectivity and adjacency to a minimal blood-brain barrier, modulate the downstream release of vasopressin by the posterior lobe of the pituitary gland. Secretion is stimulated primarily by changing tonicity and secondarily by other nonosmotic signals such as variable blood volume, stress, pain, nausea, and some drugs. The release of vasopressin by the posterior pituitary increases linearly as plasma tonicity rises above normal, although this varies, depending on the perception of

extracellular volume (one form of cross-talk between mechanisms that adjudicate blood volume and osmoregulation). Changing the intake or excretion of water provides a means for adjusting plasma tonicity; thus, osmoregulation governs water balance.

The kidneys play a vital role in maintaining water balance through the regulation of renal water excretion. The ability to concentrate urine to an osmolality exceeding that of plasma enables water conservation, whereas the ability to produce urine more dilute than plasma promotes excretion of excess water. For water to enter or exit a cell, the cell membrane must express aquaporins. In the kidney, aquaporin-1 is constitutively active in all water-permeable segments (e.g., proximal tubule, descending thin limb of the loop of Henle), whereas aquaporin-2, -3, and -4 in the collecting duct promote vasopressin-regulated water permeability. Net water reabsorption is ultimately driven by the osmotic gradient between dilute tubular fluid and a hypertonic medullary interstitium.

SODIUM BALANCE

The perception of *extracellular blood volume* is determined, in part, by the integration of arterial tone, cardiac stroke volume, heart rate, and the water and solute content of extracellular fluid. Na^+ and accompanying anions are the most abundant extracellular effective osmoles and together support a blood volume around which pressure is generated. Under normal conditions, this volume is regulated by sodium balance (Fig. 303-4B), and the balance between daily Na^+ intake and excretion is under the influence of *baroreceptors* in regional blood vessels and vascular hormone sensors modulated by atrial natriuretic peptides, the renin-angiotensin-aldosterone system, Ca^{2+} signaling, adenosine, vasopressin, and the neural adrenergic axis. If Na^+ intake exceeds Na^+ excretion (positive Na^+ balance), then an increase in blood volume will trigger a proportional increase in urinary Na^+ excretion. Conversely, when Na^+ intake is less than urinary excretion (negative Na^+ balance), blood volume will decrease and trigger enhanced renal Na^+ reabsorption, leading to decreased urinary Na^+ excretion.

The renin-angiotensin-aldosterone system is the best-understood hormonal system modulating renal Na^+ excretion. Renin is synthesized and secreted by granular cells in the wall of the afferent arteriole. Its secretion is controlled by several factors, including β_1-adrenergic stimulation to the afferent arteriole, input from the macula densa, and prostaglandins. Renin and ACE activity eventually produce angiotensin II that directly and indirectly promotes renal Na^+ and water reabsorption. Stimulation of proximal tubular Na^+/H^+ exchange by angiotensin II directly increases Na^+ reabsorption. Angiotensin II also promotes Na^+ reabsorption along the collecting duct by stimulating aldosterone secretion by the adrenal cortex. Constriction of the efferent glomerular arteriole by angiotensin II indirectly increases the filtration fraction and raises peritubular capillary oncotic pressure to promote tubular Na^+ reabsorption. Finally, angiotensin II inhibits renin secretion through a negative feedback loop. Alternative metabolism of angiotensin by ACE2 generates the vasodilatory peptide angiotensin 1-7 that acts through Mas receptors to counterbalance several actions of angiotensin II on blood pressure and renal function (Fig. 303-2C).

Aldosterone is synthesized and secreted by granulosa cells in the adrenal cortex. It binds to cytoplasmic mineralocorticoid receptors in the collecting duct principal cells that increase activity of ENaC, apical membrane K^+ channel, and basolateral Na^+/K^+-ATPase. These effects are mediated in part by aldosterone-stimulated transcription of the gene encoding serum/glucocorticoid-induced kinase 1 (SGK1). The activity of ENaC is increased by SGK1-mediated phosphorylation of Nedd4-2, a protein that promotes recycling of the Na^+ channel from the plasma membrane. Phosphorylated Nedd4-2 has impaired interactions with ENaC, leading to increased channel density at the plasma membrane and increased capacity for Na^+ reabsorption by the collecting duct.

Chronic exposure to aldosterone causes a decrease in urinary Na^+ excretion lasting only a few days, after which Na^+ excretion returns to previous levels. This phenomenon, called *aldosterone escape*, is explained by decreased proximal tubular Na^+ reabsorption following blood volume expansion. Excess Na^+ that is not reabsorbed by the proximal tubule overwhelms the reabsorptive capacity of more distal nephron segments. This escape may be facilitated by atrial natriuretic peptides that lose their effectiveness in the clinical settings of heart failure, nephrotic syndrome, and cirrhosis, leading to severe Na^+ retention and volume overload.

FURTHER READING

Bhave G, Neilson EG: Body fluid dynamics: Back to the future. J Am Soc Nephrol 22:2166, 2011. [PubMed: 22034644]

Carlstrom M et al: Renal autoregulation in health and disease. Physiol Rev 95:405, 2015. [PubMed: 25834230]

Combes AN et al: Cell-cell interactions driving kidney morphogenesis. Curr Top Dev Biol 112:467, 2015. [PubMed: 25733149]

Curthoys NP, Moe OW: Proximal tubule function and response to acidosis. Clin J Am Soc Nephrol 9:1627, 2014. [PubMed: 23908456]

Hadchouel J et al: Regulation of renal electrolyte transport by WNK and SPAK-OSR1 kinases. Annu Rev Physiol 78:367, 2016. [PubMed: 26863326]

Jung HJ, Kwon TH: Molecular mechanisms regulating aquaporin-2 in kidney collecting duct. Am J Physiol Renal Physiol 331:F1318, 2016.

Mount DB: Thick ascending limb of the loop of Henle. Clin J Am Soc Nephrol 9:1974, 2014. [PubMed: 25318757]

Nigam SK et al: Handling of drugs, metabolites, and uremic toxins by kidney proximal tubule drug transporters. Clin J Am Soc Nephrol 10:2039, 2015. [PubMed: 26490509]

Palmer LG, Schnermann J: Integrated control of Na transport along the nephron. Clin J Am Soc Nephrol 10:676, 2015. [PubMed: 25098598]

Pearce D et al: Collecting duct principal cell transport processes and their regulation. Clin J Am Soc Nephrol 10:135, 2015. [PubMed: 24875192]

Terker AS, Ellison DH: Renal mineralocorticoid receptor and electrolyte homeostasis. Am J Physiol Regul Integr Comp Physiol 309:R1068, 2015. [PubMed: 26136532]

Chapter 304: Acute Kidney Injury

Sushrut S. Waikar; Joseph V. Bonventre

INTRODUCTION

Acute kidney injury (AKI) is defined by the impairment of kidney filtration and excretory function over days to weeks, resulting in the retention of nitrogenous and other waste products normally cleared by the kidneys. AKI is not a single disease but, rather, a designation for a heterogeneous group of conditions that share common diagnostic features: specifically, an increase in serum creatinine (SCr) concentration often associated with a reduction in urine volume. It is important to recognize that AKI is a clinical diagnosis and not a structural one. A patient may have AKI with or without injury to the kidney parenchyma. AKI can range in severity from asymptomatic and transient changes in laboratory parameters of glomerular filtration rate (GFR), to overwhelming and rapidly fatal derangements in effective circulating volume regulation and electrolyte and acid-base composition of the plasma.

EPIDEMIOLOGY

AKI complicates 5–7% of acute care hospital admissions and up to 30% of admissions to the intensive care unit. The incidence of AKI has grown by more than fourfold in the United States since 1988 and is estimated to have a yearly incidence of 500 per 100,000 population, higher than the yearly incidence of stroke. AKI is associated with a markedly increased risk of death in hospitalized individuals, particularly in those admitted to the ICU where in-hospital mortality rates may exceed 50%. AKI increases the risk for the development or worsening of chronic kidney disease (CKD). Patients who survive and recover from an episode of severe AKI requiring dialysis are at increased risk for the later development of dialysis-requiring end-stage kidney disease. AKI may be community-acquired or hospital-acquired. Common causes of community-acquired AKI include volume depletion, heart failure, adverse effects of medications, obstruction of the urinary tract, or malignancy. The most common clinical settings for hospital-acquired AKI are sepsis, major surgical procedures, critical illness involving heart or liver failure, and nephrotoxic medication administration.

AKI IN THE DEVELOPING WORLD

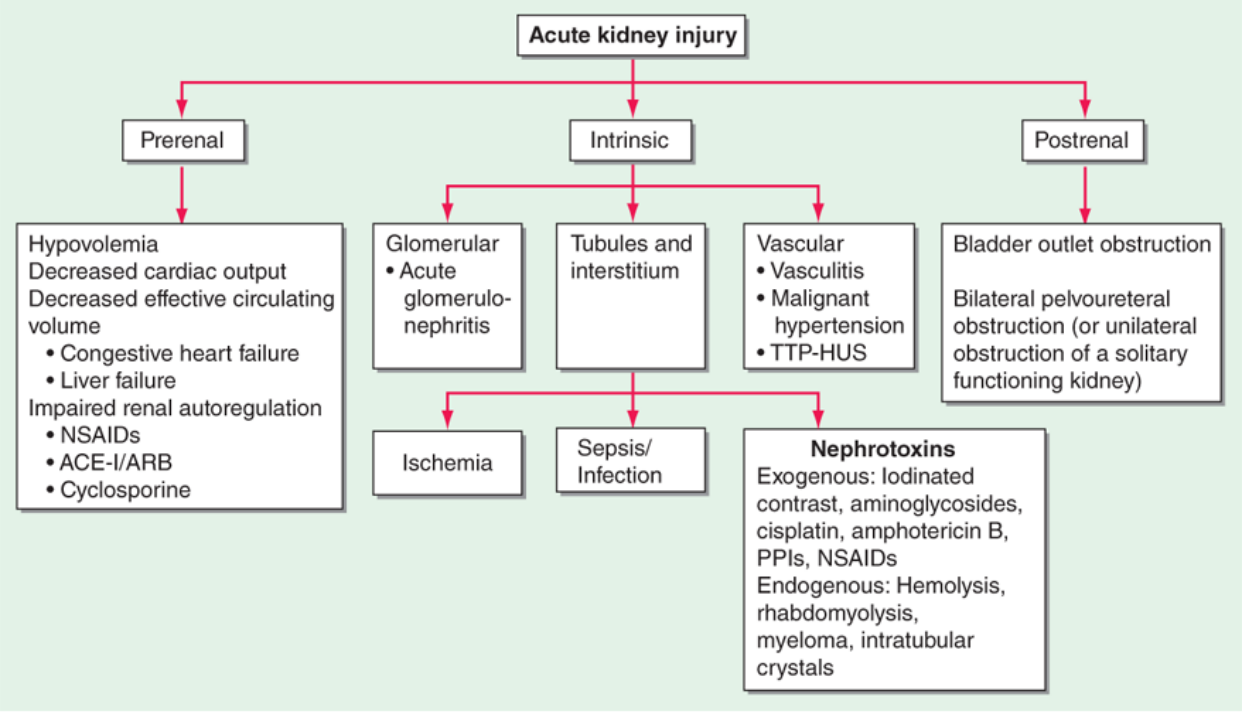
AKI is also a major medical complication in the developing world, where the epidemiology differs from that in developed countries due to differences in demographics, economics, environmental factors, and comorbid disease burden. While certain features of AKI are common in the developed and developing countries—particularly since urban centers of some developing countries increasingly resemble those in the developed world—many etiologies for AKI are region-specific such as envenomations from snakes, spiders, caterpillars, and bees; infectious causes such as malaria and leptospirosis; and crush injuries and resultant rhabdomyolysis from earthquakes.

ETIOLOGY AND PATHOPHYSIOLOGY

The causes of AKI have traditionally been divided into three broad categories: prerenal azotemia, intrinsic renal parenchymal disease, and postrenal obstruction (Fig. 304-1).

FIGURE 304-1

Classification of the major causes of acute kidney injury. ACE-I, angiotensin-converting enzyme inhibitor-I; ARB, angiotensin receptor blocker; NSAIDs, nonsteroidal anti-inflammatory drugs; PPI, proton pump inhibitors; TTP-HUS, thrombotic thrombocytopenic purpura–hemolytic-uremic syndrome.



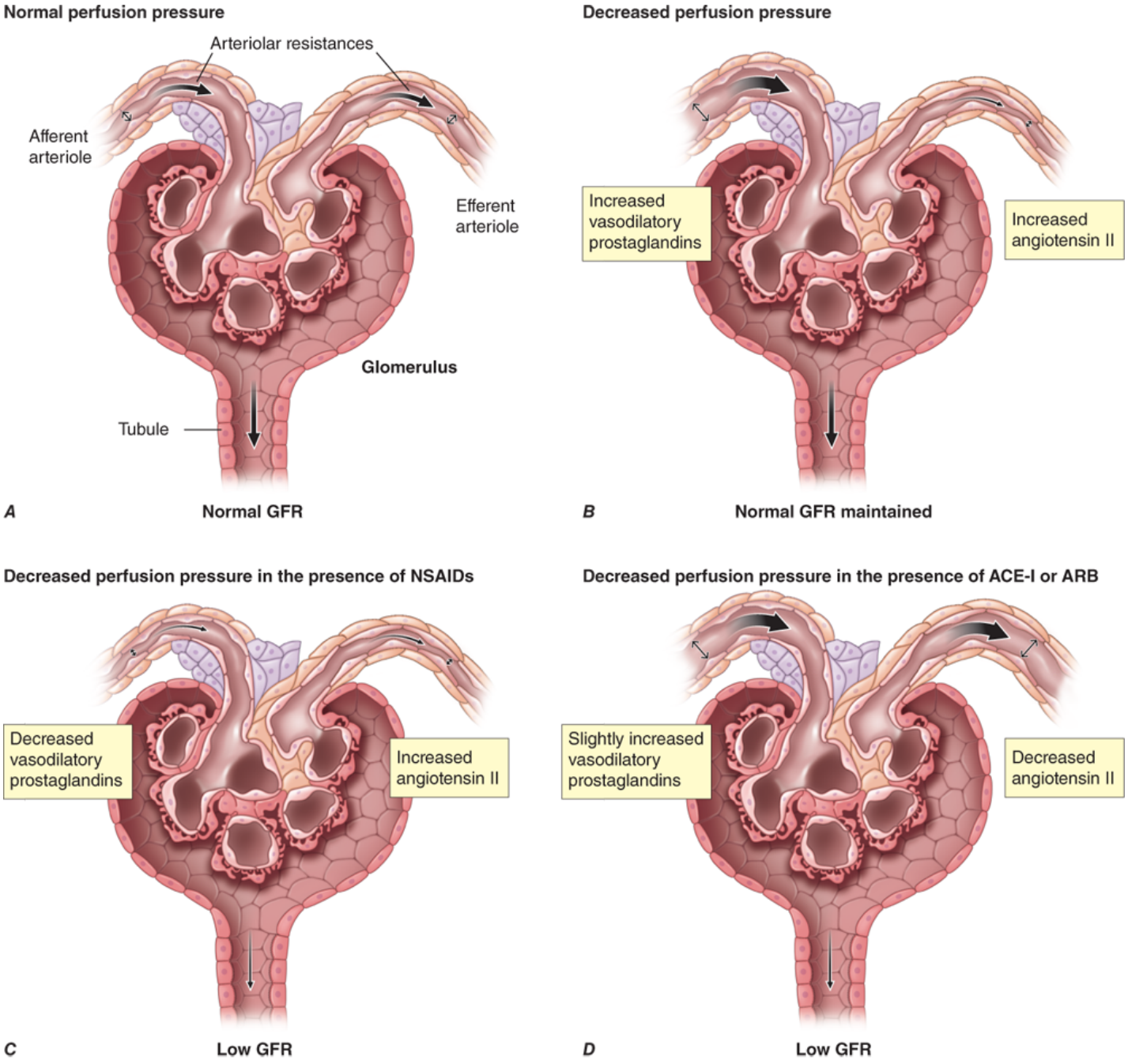
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PRERENAL AZOTEMIA

Prerenal azotemia (from “azo,” meaning nitrogen, and “-emia,” meaning in the blood) is the most common form of AKI. It is the designation for a rise in SCr or BUN concentration due to inadequate renal plasma flow and intraglomerular hydrostatic pressure to support normal glomerular filtration. The most common clinical conditions associated with prerenal azotemia are hypovolemia, decreased cardiac output, and medications that interfere with renal autoregulatory responses such as nonsteroidal anti-inflammatory drugs (NSAIDs) and inhibitors of angiotensin II (**Fig. 304-2**). Prerenal azotemia may coexist with other forms of intrinsic AKI associated with processes acting directly on the renal parenchyma. Prolonged periods of prerenal azotemia may lead to ischemic injury, often termed acute tubular necrosis (ATN). By definition, prerenal azotemia involves no parenchymal damage to the kidney and is rapidly reversible once parenchymal blood flow and intraglomerular hemodynamics are restored.

FIGURE 304-2

Intrarenal mechanisms for autoregulation of the glomerular filtration rate (GFR) under decreased perfusion pressure and reduction of the GFR by drugs. ***A.*** Normal conditions and a normal GFR. ***B.*** Reduced perfusion pressure within the autoregulatory range. Normal glomerular capillary pressure is maintained by afferent vasodilatation and efferent vasoconstriction. ***C.*** Reduced perfusion pressure with a nonsteroidal anti-inflammatory drug (NSAID). Loss of vasodilatory prostaglandins increases afferent resistance; this causes the glomerular capillary pressure to drop below normal values and the GFR to decrease. ***D.*** Reduced perfusion pressure with an angiotensin-converting enzyme inhibitor (ACE-I) or an angiotensin receptor blocker (ARB). Loss of angiotensin II action reduces efferent resistance; this causes the glomerular capillary pressure to drop below normal values and the GFR to decrease. *(From JG Abuelo: N Engl J Med 357:797-805, 2007; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Normal GFR is maintained in part by renal blood flow and the relative resistances of the afferent and efferent renal arterioles, which determine the glomerular plasma flow rate and the transcapillary hydraulic pressure gradient that drive glomerular ultrafiltration. Mild degrees of hypovolemia and reductions in cardiac output elicit compensatory renal physiologic changes. Because renal blood flow accounts for 20% of the cardiac output, renal vasoconstriction and salt and water reabsorption occur as homeostatic responses to decreased effective circulating volume or cardiac output in order to maintain blood pressure and increase intravascular volume to sustain perfusion to the cerebral and coronary vessels. Mediators of this response include angiotensin II, norepinephrine, and vasopressin (also termed antidiuretic hormone). Glomerular filtration can be maintained despite reduced renal blood flow by angiotensin II–mediated renal efferent vasoconstriction, which maintains glomerular capillary hydrostatic pressure closer to normal and thereby prevents marked reductions in GFR if renal blood flow reduction is not excessive.

In addition, a myogenic reflex within the afferent arteriole leads to dilation in the setting of low perfusion pressure, thereby maintaining glomerular perfusion. Intrarenal biosynthesis of vasodilator prostaglandins (prostacyclin, prostaglandin E_2), kallikrein and kinins, and possibly nitric oxide (NO) also increase in response to low renal perfusion pressure. Autoregulation is also accomplished by tubuloglomerular feedback, in which decreases in solute delivery to the macula densa (specialized cells within the distal tubule) elicit dilation of the juxtaposed afferent arteriole in order to maintain glomerular perfusion, a mechanism mediated, in part, by NO. There is a limit, however, to the ability of these counterregulatory mechanisms to maintain GFR in the face of systemic hypotension. Even in healthy adults, renal autoregulation usually fails once the systolic blood pressure falls below 80 mmHg.

A number of factors determine the robustness of the autoregulatory response and the risk of prerenal azotemia. Atherosclerosis, long-standing hypertension, and older age can lead to hyalinosis and myointimal hyperplasia, causing structural narrowing of the intrarenal arterioles and impaired capacity for renal afferent vasodilation. In CKD, renal afferent vasodilation may be operating at maximal capacity in order to maximize GFR in response to reduced functional renal mass. Drugs can affect the compensatory changes evoked to maintain GFR. Nonsteroidal anti-inflammatory agents (NSAIDs) inhibit renal prostaglandin production, limiting renal afferent vasodilation. Angiotensin-converting enzyme (ACE) inhibitors and angiotensin receptor blockers (ARBs) limit renal efferent vasoconstriction; this effect is particularly pronounced in patients with bilateral renal artery stenosis or unilateral renal artery stenosis (in the case of a solitary functioning kidney) because, as indicated above, efferent arteriolar vasoconstriction is needed to maintain GFR due to low renal perfusion. The combined use of NSAIDs with ACE inhibitors or ARBs poses a particularly high risk for developing prerenal azotemia.

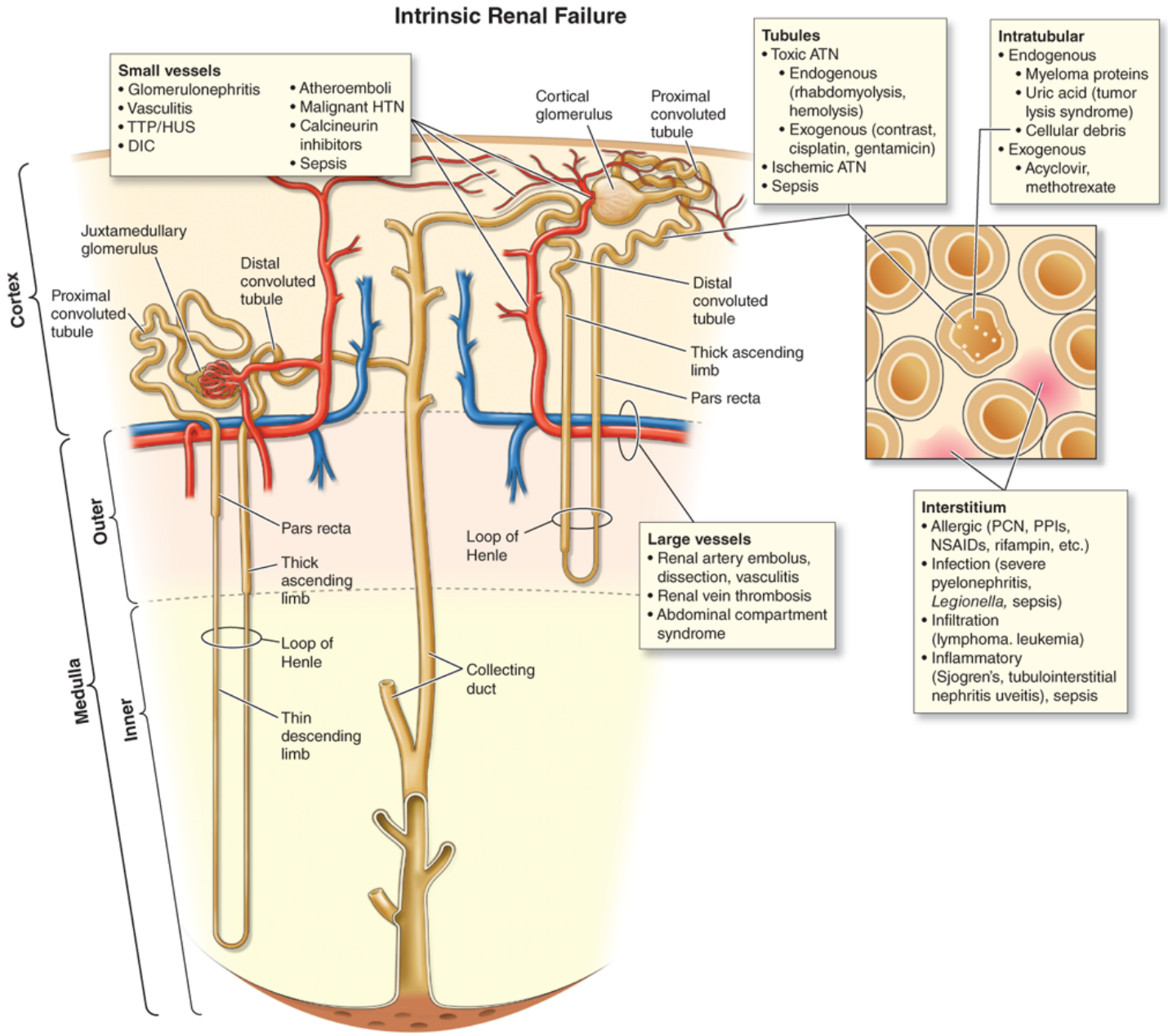
Many individuals with advanced liver disease exhibit a hemodynamic profile that resembles prerenal azotemia in the setting of total-body volume overload. Systemic vascular resistance is markedly reduced due to primary arterial vasodilation in the splanchnic circulation, resulting ultimately in activation of vasoconstrictor responses similar to those seen in hypovolemia. AKI is a common complication in this setting, and it can be triggered by volume depletion and spontaneous bacterial peritonitis. A particularly poor prognosis is seen in the case of type 1 hepatorenal syndrome, in which AKI, defined as >twofold increase in SCr to >2.5 mg/dL, within 2 weeks without an alternate cause (e.g., shock and nephrotoxic drugs), persists despite volume administration and withholding of diuretics. Type 2 hepatorenal syndrome is a less severe form characterized mainly by refractory ascites. The hepatorenal syndrome, defined as it is above, is difficult to distinguish from prerenal azotemia. An older way of characterizing hepatorenal was prerenal azotemia that would not improve, often leading to intrinsic renal AKI, unless a definitive procedure to improve hemodynamics, such as porto-systemic shunt placement or liver transplant, was performed. We still find this latter construct of use.

INTRINSIC AKI

The most common causes of intrinsic AKI are sepsis, ischemia, and nephrotoxins, both endogenous and exogenous (**Fig. 304-3**). In many cases, prerenal azotemia advances to tubular injury. Although classically termed “acute tubular necrosis,” human biopsy confirmation of tubular necrosis is, in general, often lacking in cases of sepsis and ischemia; indeed, processes such as inflammation, apoptosis, and altered regional perfusion may be important contributors pathophysiologically. ATN is also often diagnosed clinically without biopsy confirmation in settings such as sepsis with multiple alternate potential diagnoses, including drug-induced interstitial nephritis and immune complex glomerulonephritis. These and other causes of intrinsic AKI are considered to be less common and can be conceptualized anatomically according to the major site of renal parenchymal damage: glomeruli, tubulointerstitium, and vessels.

FIGURE 304-3

Major causes of intrinsic acute kidney injury. ATN, acute tubular necrosis; DIC, disseminated intravascular coagulation; HTN, hypertension; PCN, penicillin; PPI, proton pump inhibitors; TTP/HUS, thrombotic thrombocytopenic purpura/hemolytic-uremic syndrome; TINU, tubulointerstitial nephritis-uveitis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SEPSIS-ASSOCIATED AKI

In the United States, more than one million cases of sepsis occur each year. AKI complicates more than 50% of cases of severe sepsis and greatly increases the risk of death. Sepsis is also a very important cause of AKI in the developing world. Decreases in GFR with sepsis can occur even in the absence of overt hypotension, although most cases of severe AKI typically occur in the setting of hemodynamic collapse requiring vasopressor support. While there is clearly tubular injury associated with AKI in sepsis as manifest by the presence of tubular debris and casts in the urine, postmortem examinations of kidneys from individuals with severe sepsis suggest that other factors, perhaps related to inflammation, mitochondrial dysfunction, and interstitial edema, must also be considered in the pathophysiology of sepsis-induced AKI.

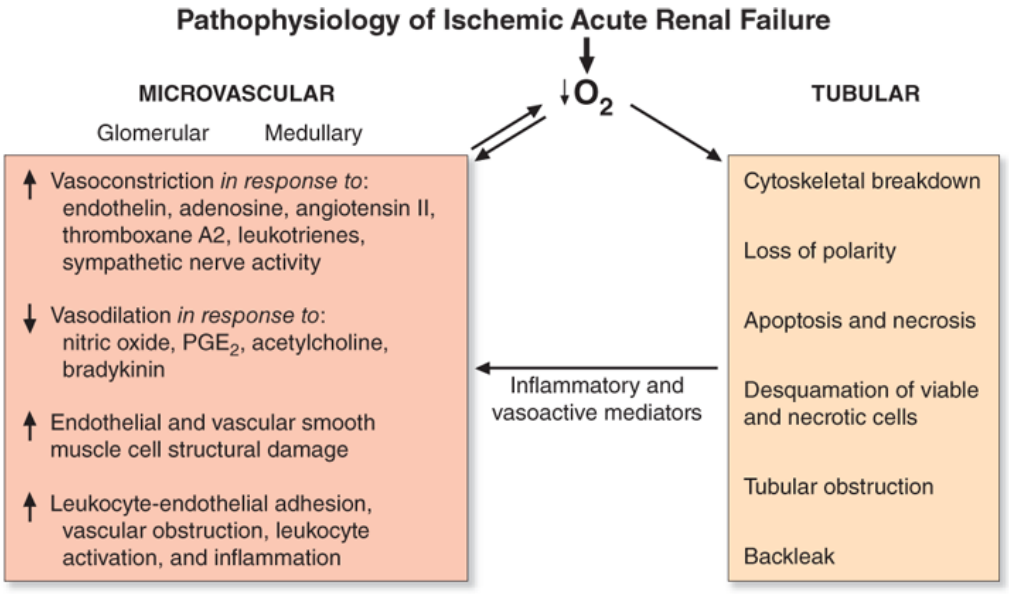
The hemodynamic effects of sepsis—arising from generalized arterial vasodilation, mediated in part by cytokines that upregulate the expression of inducible NO synthase in the vasculature—can lead to a reduction in GFR. The operative mechanisms may be excessive efferent arteriole vasodilation, particularly early in the course of sepsis, or renal vasoconstriction from activation of the sympathetic nervous system, the renin-angiotensin-aldosterone system, vasopressin, and endothelin. Sepsis may lead to endothelial damage, which results in increased microvascular leukocyte adhesion and migration, thrombosis, permeability, increased interstitial pressure, reduction in local flow to tubules, and activation of reactive oxygen species, all of which may injure renal tubular cells.

ISCHEMIA-ASSOCIATED AKI

Healthy kidneys receive 20% of the cardiac output and account for 10% of resting oxygen consumption, despite constituting only 0.5% of the human body mass. The kidneys are also the site of one of the most hypoxic regions in the body, the renal medulla. The outer medulla is particularly vulnerable to ischemic damage because of the architecture of the blood vessels that supply oxygen and nutrients to the tubules. In the outer medulla enhanced leukocyte-endothelial interactions in the small vessels lead to inflammation and reduced local blood flow to the metabolically very active S3 segment of the proximal tubule, which depends on oxidative metabolism for survival. Mitochondrial dysfunction due to ischemia and mitochondrial release of reactive oxygen species also play a role in renal tubular injury. Ischemia alone in a normal kidney is usually not sufficient to cause severe AKI, as evidenced by the relatively low risk of severe AKI even after total interruption of renal blood flow during suprarenal aortic clamping or cardiac arrest. Clinically, AKI more commonly develops when ischemia occurs in the context of limited renal reserve (e.g., CKD or older age) or coexisting insults such as sepsis, vasoactive or nephrotoxic drugs, rhabdomyolysis, or the systemic inflammatory states associated with burns and pancreatitis. Prerenal azotemia and ischemia-associated AKI represent a continuum of the manifestations of renal hypoperfusion. Persistent preglomerular vasoconstriction may be a common underlying cause of the reduction in GFR seen in AKI; implicated factors for vasoconstriction include activation of tubuloglomerular feedback from enhanced delivery of solute to the macula densa following proximal tubule injury, increased basal vascular tone and reactivity to vasoconstrictive agents, and decreased vasodilator responsiveness. Other contributors to low GFR include backleak of filtrate across damaged and denuded tubular epithelium and mechanical obstruction of tubules from necrotic debris **(Fig. 304-4)**.

FIGURE 304-4

Interacting microvascular and tubular events contributing to the pathophysiology of ischemic acute kidney injury. PGE_2, prostaglandin E_2. *(From JV Bonventre, JM Weinberg: J Am Soc Nephrol 14:2199, 2003.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Postoperative AKI

Ischemia-associated AKI is a serious complication in the postoperative period, especially after major operations involving significant blood loss and intraoperative hypotension. The procedures most commonly associated with AKI are cardiac surgery with cardiopulmonary bypass (particularly for combined valve and bypass procedures), vascular procedures with aortic cross clamping, and intraperitoneal procedures. Severe AKI requiring dialysis occurs in ~1% of cardiac and vascular surgery procedures. The risk of severe AKI has been less well studied for major intraperitoneal procedures but appears to be of comparable magnitude. Common risk factors for postoperative AKI include underlying CKD, older age, diabetes mellitus, congestive heart failure, and emergency procedures. The pathophysiology of AKI following cardiac surgery is multifactorial. Major AKI risk factors are common in the population undergoing cardiac surgery. The use of nephrotoxic agents, including iodinated contrast for cardiac imaging prior to surgery, may increase the risk of AKI. Cardiopulmonary bypass is a unique hemodynamic state characterized by nonpulsatile flow and exposure of the circulation to extracorporeal circuits. Longer duration of cardiopulmonary bypass is a risk factor for AKI. In addition to ischemic injury from sustained hypoperfusion, cardiopulmonary bypass may cause AKI through a number of mechanisms including extracorporeal circuit activation of leukocytes and inflammatory processes, hemolysis with resultant pigment nephropathy (see below), and aortic injury with resultant atheroemboli. AKI from atheroembolic disease, which can also occur following percutaneous catheterization of the aorta, or spontaneously, is due to cholesterol crystal embolization resulting in partial or total occlusion of multiple small arteries within the kidney. Over time, a foreign body reaction can result in intimal proliferation, giant cell formation, and further narrowing of the vascular lumen, accounting for the generally subacute (over a period of weeks rather than days) decline in renal function.

Burns and Acute Pancreatitis

Extensive fluid losses into the extravascular compartments of the body frequently accompany severe burns and acute pancreatitis. AKI is an ominous complication of burns, affecting 25% of individuals with >10% total body surface area involvement. In addition to severe hypovolemia resulting in decreased cardiac output and increased neurohormonal activation, burns and acute pancreatitis both lead to dysregulated inflammation and an increased risk of sepsis and acute lung injury, all of which may facilitate the development and progression of AKI. Individuals undergoing massive fluid resuscitation for trauma, burns, and acute pancreatitis can also develop the abdominal compartment syndrome, where markedly elevated intraabdominal pressures, usually >20 mmHg, lead to renal vein compression and reduced GFR.

Diseases of the Microvasculature Leading to Ischemia

Microvascular causes of AKI include the thrombotic microangiopathies (due to cocaine, certain chemotherapeutic agents, antiphospholipid antibody syndrome, radiation nephritis, malignant hypertensive nephrosclerosis, and thrombotic thrombocytopenic purpura/hemolytic-uremic syndrome [TTP-HUS]), scleroderma, and atheroembolic disease. Large-vessel diseases associated with AKI include renal artery dissection, thromboembolism, or thrombosis, and renal vein compression or thrombosis.

NEPHROTOXIN-ASSOCIATED AKI

The kidney has very high susceptibility to nephrotoxic agents due to extremely high blood perfusion and concentration of circulating substances along the nephron where water is reabsorbed and in the medullary interstitium; this results in high-concentration exposure of toxins to tubular, interstitial, and endothelial cells. Nephrotoxic injury occurs in response to a number of pharmacologic compounds with diverse structures, endogenous substances, and environmental exposures. All structures of the kidney are vulnerable to toxic injury, including the tubules, interstitium, vasculature, and collecting system. As with other forms of AKI, risk factors for nephrotoxicity include older age, CKD, and prerenal azotemia. Hypoalbuminemia may increase the risk of some forms of nephrotoxin-associated AKI due to increased free circulating drug concentrations.

Contrast Agents

Iodinated contrast agents used for cardiovascular and computed tomography (CT) imaging are a cause of AKI. The risk of AKI, or “contrast nephropathy,” is negligible in those with normal renal function but increases in the setting of CKD, particularly diabetic nephropathy. The most common clinical course of contrast nephropathy is characterized by a rise in SCr beginning 24–48 h following exposure, peaking within 3–5 days, and resolving within 1 week. More severe, dialysis-requiring AKI is uncommon except in the setting of significant preexisting CKD, often in association with congestive heart failure or other coexisting causes for ischemia-associated AKI. Patients with multiple myeloma and renal disease are particularly susceptible. Low fractional excretion of sodium (FeNa) and relatively benign urinary sediment without features of tubular necrosis (see below) are common findings. Contrast nephropathy is thought to occur from a combination of factors, including (1) hypoxia in the renal outer medulla due to perturbations in renal microcirculation and occlusion of small vessels; (2) cytotoxic damage to the tubules directly or via the generation of oxygen-free radicals, especially because the concentration of the agent within the tubule is markedly increased; and (3) transient tubule obstruction with precipitated contrast material. Other diagnostic agents implicated as a cause of AKI are high-dose gadolinium used for magnetic resonance imaging (MRI) and oral sodium phosphate solutions used as bowel purgatives.

Antibiotics

Several antimicrobial agents are commonly associated with AKI. *Vancomycin* may be associated with AKI, particularly when trough levels are high and when used in combination with other nephrotoxic antibiotics. *Aminoglycosides and amphotericin B* both cause tubular necrosis. Nonoliguric AKI (i.e., with a urine volume >400 mL/day) accompanies 10–30% of courses of aminoglycoside antibiotics, even when plasma levels are in the therapeutic range. Aminoglycosides are freely filtered across the glomerulus and then accumulate within the renal cortex, where concentrations can greatly exceed those of the plasma. AKI typically manifests after 5–7 days of therapy and can present even after the drug has been discontinued. Hypomagnesemia is a common finding.

Amphotericin B causes renal vasoconstriction from an increase in tubuloglomerular feedback as well as direct tubular toxicity mediated by reactive oxygen species. Nephrotoxicity from amphotericin B is dose and duration dependent. This drug binds to tubular membrane cholesterol and introduces pores. Clinical features of amphotericin B nephrotoxicity include polyuria, hypomagnesemia, hypocalcemia, and nongap metabolic acidosis.

Acyclovir can precipitate in tubules and cause AKI by tubular obstruction, particularly when given as an intravenous bolus at high doses (500 mg/m^2) or in the setting of hypovolemia. *Foscarnet, pentamidine, tenofovir*, and *cidofovir* are also frequently associated with AKI due to tubular toxicity. AKI secondary to acute interstitial nephritis can occur as a consequence of exposure to many antibiotics, including *penicillins, cephalosporins, quinolones, sulfonamides*, and *rifampin*.

Chemotherapeutic Agents

Cisplatin and *carboplatin* are accumulated by proximal tubular cells and cause necrosis and apoptosis. Intensive hydration regimens have reduced the incidence of cisplatin nephrotoxicity, but it remains a dose-limiting toxicity. *Ifosfamide* may cause hemorrhagic cystitis and tubular toxicity, manifested as type II renal tubular acidosis (Fanconi’s syndrome), polyuria, hypokalemia, and a modest decline in GFR. Antiangiogenesis agents, such as *bevacizumab,*

can cause proteinuria and hypertension via injury to the glomerular microvasculature (thrombotic microangiopathy). Other antineoplastic agents such as mitomycin C and gemcitabine may cause thrombotic microangiopathy with resultant AKI.

Toxic Ingestions

Ethylene glycol, present in automobile antifreeze, is metabolized to oxalic acid, glycolaldehyde, and glyoxylate, which may cause AKI through direct tubular injury and tubular obstruction. Diethylene glycol is an industrial agent that has caused outbreaks of severe AKI around the world due to adulteration of pharmaceutical preparations. The metabolite 2-hydroxyethoxyacetic acid (HEAA) is thought to be responsible for tubular injury. Melamine contamination of foodstuffs has led to nephrolithiasis and AKI, either through intratubular obstruction or possibly direct tubular toxicity. Aristolochic acid was found to be the cause of “Chinese herb nephropathy” and “Balkan nephropathy” due to contamination of medicinal herbs or farming. The list of environmental toxins is likely to grow and contribute to a better understanding of previously catalogued “idiopathic” chronic tubular interstitial disease, a common diagnosis in both the developed and developing world.

Endogenous Toxins

AKI may be caused by a number of endogenous compounds, including myoglobin, hemoglobin, uric acid, and myeloma light chains. Myoglobin can be released by injured muscle cells, and hemoglobin can be released during massive hemolysis leading to pigment nephropathy. Rhabdomyolysis may result from traumatic crush injuries, muscle ischemia during vascular or orthopedic surgery, compression during coma or immobilization, prolonged seizure activity, excessive exercise, heat stroke or malignant hyperthermia, infections, metabolic disorders (e.g., hypophosphatemia, severe hypothyroidism), and myopathies (drug-induced, metabolic, or inflammatory). Pathogenic factors for AKI due to endogenous toxins include intrarenal vasoconstriction, direct proximal tubular toxicity, and mechanical obstruction of the distal nephron lumen when myoglobin or hemoglobin precipitates with Tamm-Horsfall protein (uromodulin, the most common protein in urine and produced in the thick ascending limb of the loop of Henle), a process favored by acidic urine. Tumor lysis syndrome may follow initiation of cytotoxic therapy in patients with high-grade lymphomas and acute lymphoblastic leukemia; massive release of uric acid (with serum levels often exceeding 15 mg/dL) leads to precipitation of uric acid in the renal tubules and AKI **(Chap. 71)**. Other features of tumor lysis syndrome include hyperkalemia and hyperphosphatemia. The tumor lysis syndrome can also occasionally occur spontaneously or with treatment for solid tumors or multiple myeloma. Myeloma light chains can also cause AKI by direct tubular toxicity and by binding to Tamm-Horsfall protein to form obstructing intratubular casts. Hypercalcemia, which can also be seen in multiple myeloma, may cause AKI by intense renal vasoconstriction and volume depletion.

Other Causes of Acute Tubulointerstitial Disease Leading to AKI

While many of the ischemic and toxic causes of AKI previously described result in tubulointerstitial disease, many drugs are also associated with the development of an allergic response characterized by an inflammatory infiltrate and often peripheral and urinary eosinophilia. Proton pump inhibitors and NSAIDS are commonly used drugs that have been associated with acute tubulointerstitial nephritis. AKI may be also caused by severe infections and infiltrative malignant or nonmalignant (e.g., sarcoidosis) diseases.

Glomerulonephritis

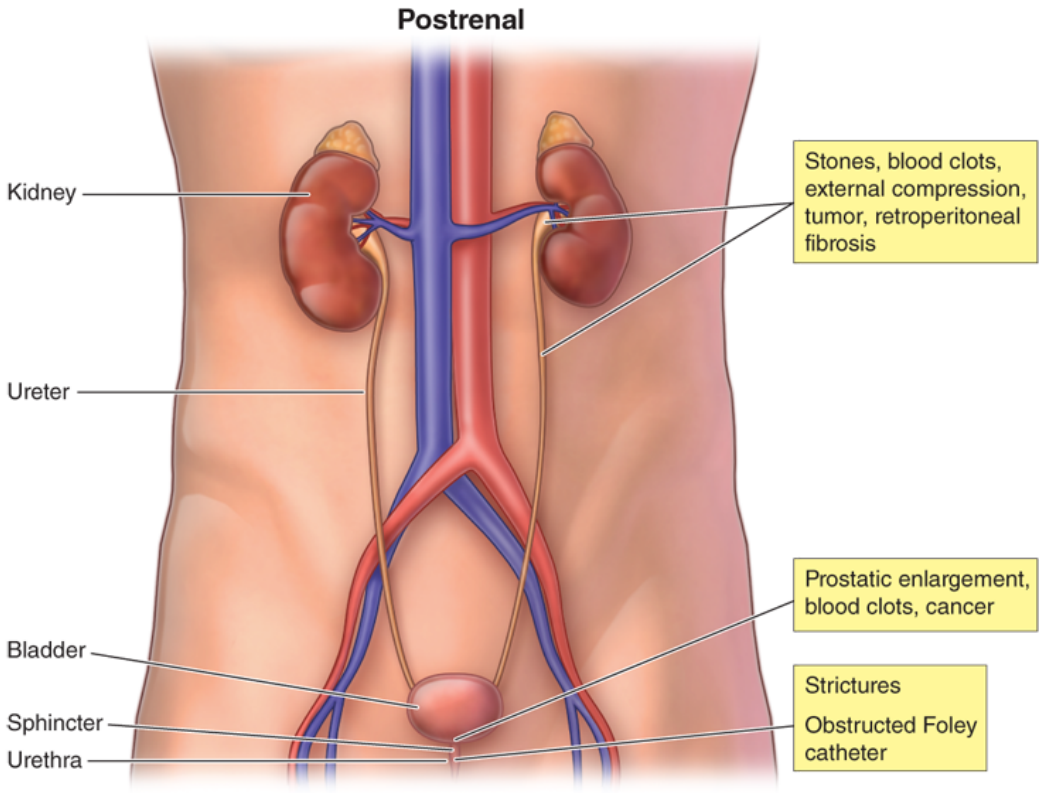
Diseases involving the glomerular podocytes, mesangial and endothelial cells can lead to AKI by compromising the filtration barrier and blood flow within the renal circulation. Although glomerulonephritis is a less common (~5%) cause of AKI, early recognition is particularly important because the diseases can respond to timely treatment with immunosuppressive agents or therapeutic plasma exchange, and the treatment may reverse the AKI.

POSTRENAL AKI

(See also Chap. 313) Postrenal AKI occurs when the normally unidirectional flow of urine is acutely blocked either partially or totally, leading to increased retrograde hydrostatic pressure and interference with glomerular filtration. Obstruction to urinary flow may be caused by functional or structural derangements anywhere from the renal pelvis to the tip of the urethra **(Fig. 304-5)**. Normal urinary flow rate does not rule out the presence of partial obstruction, because the GFR is normally two orders of magnitude higher than the urinary flow rate and hence a preservation of urine output may be misleading in hiding the postrenal partial obstruction. For AKI to occur in individuals with two healthy functional kidneys, obstruction must affect both kidneys in order to observe large increases in SCr. Unilateral obstruction may cause AKI in the setting of significant underlying CKD or, in rare cases, from reflex vasospasm of the contralateral kidney. Bladder neck obstruction is a common cause of postrenal AKI which impacts both kidneys. This can be due to prostate disease (benign prostatic hypertrophy or prostate cancer), neurogenic bladder, or therapy with anticholinergic drugs. Obstructed Foley catheters can cause postrenal AKI if not recognized and relieved. Other causes of lower tract obstruction are blood clots, calculi, and urethral strictures. Ureteric obstruction can occur from intraluminal obstruction (e.g., calculi, blood clots, sloughed renal papillae), infiltration of the ureteric wall (e.g., neoplasia), or external compression (e.g., retroperitoneal fibrosis, neoplasia, abscess, or inadvertent surgical damage). The pathophysiology of postrenal AKI involves hemodynamic alterations triggered by an abrupt increase in intratubular pressures. An initial period of hyperemia from afferent arteriolar dilation is followed by intrarenal vasoconstriction from the generation of angiotensin II, thromboxane A2, and vasopressin, and a reduction in NO production. Secondary reductions in glomerular function are due to underperfusion of glomeruli and, possibly, changes in the glomerular ultrafiltration coefficient.

FIGURE 304-5

Anatomic sites and causes of obstruction leading to postrenal acute kidney injury.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSTIC EVALUATION

By current definitions the presence of AKI is defined by an elevation in the SCr concentration or reduction in urine output (**Table 304-1**). AKI is currently defined by a rise from baseline of at least 0.3 mg/dL within 48 h or at least 50% higher than baseline within 1 week, or a reduction in urine output to <0.5 mL/kg per h for longer than 6 h. As indicated above, it is important to recognize that given this definition, some patients with AKI will not have tubular or glomerular damage (e.g., prerenal azotemia). The distinction between AKI and CKD is important for proper diagnosis and treatment. The distinction is straightforward when a recent baseline SCr concentration is available, but more difficult in the many instances in which the baseline is unknown. In such cases, clues suggestive of CKD can come from radiologic studies (e.g., small, shrunken kidneys with cortical thinning on renal ultrasound, or evidence of renal osteodystrophy) or laboratory tests such as normocytic anemia in the absence of blood loss or secondary hyperparathyroidism with hyperphosphatemia and hypocalcemia, consistent with CKD. No set of tests, however, can rule out AKI superimposed on CKD because AKI is a frequent complication in patients with CKD, further complicating the distinction. Serial blood tests showing a continued substantial rise of SCr represents clear evidence of AKI. Once the diagnosis of AKI is established, its cause needs to be determined since the elevation of SCr or reduction in urine output can be due to a large number of physiological and pathophysiological processes.

TABLE 304-1

Major Causes, Clinical Features, and Diagnostic Studies for Prerenal and Intrinsic Acute Kidney Injury

ETIOLOGY	CLINICAL FEATURES	LABORATORY FEATURES	COMMENTS
Prerenal azotemia	History of poor fluid intake or fluid loss (hemorrhage, diarrhea, vomiting, sequestration into extravascular space); NSAID/ACE-I/ARB; heart failure; evidence of volume depletion (tachycardia, absolute or postural hypotension, low jugular venous pressure, dry mucous membranes), decreased effective circulatory volume (cirrhosis, heart failure)	BUN/creatinine ratio above 20, FeNa <1%, hyaline casts in urine sediment, urine specific gravity >1.018, urine osmolality >500 mOsm/kg	Low FeNa, high specific gravity and osmolality may not be seen in the setting of CKD, diuretic use; BUN elevation out of proportion to creatinine may alternatively indicate upper GI bleed or increased catabolism. Response to restoration of hemodynamics is most diagnostic.
Sepsis-associated AKI	Sepsis, sepsis syndrome, or septic shock. Overt hypotension not always seen in mild to moderate AKI	Positive culture from normally sterile body fluid; urine sediment often contains granular casts, renal tubular epithelial cell casts	FeNa may be low (<1%), particularly early in the course, but is usually >1% with osmolality <500 mOsm/kg
Ischemia-associated AKI	Systemic hypotension, often superimposed upon sepsis and/or reasons for limited renal reserve such as older age, CKD	Urine sediment often contains granular casts, renal tubular epithelial cell casts. FeNa typically >1%.	
Nephrotoxin-Associated AKI: Endogenous			
Rhabdomyolysis	Traumatic crush injuries, seizures, immobilization	Elevated myoglobin, creatine kinase; urine heme positive with few red blood cells	FeNa may be low (<1%)
Hemolysis	Recent blood transfusion with transfusion reaction	Anemia, elevated LDH, low haptoglobin	FeNa may be low (<1%); evaluation for transfusion reaction
Tumor lysis	Recent chemotherapy	Hyperphosphatemia, hypocalcemia, hyperuricemia	
Multiple myeloma	Age >60 years, constitutional symptoms, bone pain	Monoclonal spike in urine or serum electrophoresis; low anion gap; anemia	Bone marrow or renal biopsy can be diagnostic
Nephrotoxin-Associated AKI: Exogenous			
Contrast nephropathy	Exposure to iodinated contrast	Characteristic course is rise in SCr within 1–2 d, peak within 3–5 d, recovery within 7 d	FeNa may be low (<1%)

ETIOLOGY	CLINICAL FEATURES	LABORATORY FEATURES	COMMENTS
Tubular injury	Aminoglycoside antibiotics, cisplatin, tenofovir, vancoycin, zoledronate, ethylene glycol, aristolochic acid, protein pump inhibitors, tacrolimus and melamine (to name a few)	Urine sediment often contains granular casts, renal tubular epithelial cell casts. FeNa typically >1%.	Can be oliguric or nonoliguric
Interstitial nephritis	Recent medication exposure (e.g., proton pump inhibitors, NSAIDs, antibiotics), can have fever, rash, arthralgias	Eosinophilia, sterile pyuria; often nonoliguric	Urine eosinophils have limited diagnostic accuracy; systemic signs of drug reaction often absent; kidney biopsy may be helpful
Other Causes of Intrinsic AKI			
Glomerulonephritis/vasculitis	Variable **(Chap. 308)** features include skin rash, arthralgias, sinusitis (AGBM disease), lung hemorrhage (AGBM, ANCA, lupus), recent skin infection or pharyngitis (poststreptococcal)	ANA, ANCA, AGBM antibody, hepatitis serologies, cryoglobulins, blood culture, decreased complement levels, ASO titer (abnormalities of these tests depending on etiology)	Kidney biopsy may be necessary
Interstitial nephritis	Nondrug-related causes include tubulointerstitial nephritis-uveitis (TINU) syndrome, *Legionella* infection	Eosinophilia, sterile pyuria; often nonoliguric	Urine eosinophils have limited diagnostic accuracy; kidney biopsy may be necessary
TTP/HUS	Neurologic abnormalities and/or AKI; recent diarrheal illness; use of calcineurin inhibitors; pregnancy or postpartum; spontaneous	Schistocytes on peripheral blood smear, elevated LDH, anemia, thrombocytopenia	"Typical HUS" refers to AKI with a diarrheal prodrome, often due to Shiga toxin released from *Escherichia coli* or other bacteria; "atypical HUS" is due to inherited or acquired complement dysregulation. "TTP-HUS" refers to sporadic cases in adults. Diagnosis may involve screening for ADAMTS13 activity, Shiga toxin–producing *E. coli*, genetic evaluation of complement regulatory proteins, and kidney biopsy.
Atheroembolic disease	Recent manipulation of the aorta or other large vessels; may occur spontaneously or after anticoagulation; retinal plaques, palpable purpura, livedo reticularis, GI bleed	Hypocomplementemia, eosinophiluria (variable), variable amounts of proteinuria	Skin or kidney biopsy can be diagnostic
Postrenal AKI	History of kidney stones, prostate disease, obstructed bladder catheter, retroperitoneal or pelvic neoplasm	No specific findings other than AKI; may have pyuria or hematuria	Imaging with computed tomography or ultrasound

Abbreviations: ACE-I, angiotensin-converting enzyme inhibitor-I; AGBM, antiglomerular basement membrane; AKI, acute kidney injury; ANA, antinuclear antibody; ANCA, antineutrophilic cytoplasmic antibody; ARB, angiotensin receptor blocker; ASO, antistreptolysin O; BUN, blood urea nitrogen; CKD, chronic kidney disease; FeNa, fractional excretion of sodium; GI, gastrointestinal; LDH, lactate dehydrogenase; NSAID, nonsteroidal anti-inflammatory drug; TTP/HUS, thrombotic thrombocytopenic purpura/hemolytic-uremic syndrome.

HISTORY AND PHYSICAL EXAMINATION

The clinical context, careful history taking, and physical examination often narrow the differential diagnosis for the cause of AKI. Prerenal azotemia should be suspected in the setting of vomiting, diarrhea, glycosuria causing polyuria, and several medications including diuretics, NSAIDs, ACE inhibitors, and ARBs. Physical signs of orthostatic hypotension, tachycardia, reduced jugular venous pressure, decreased skin turgor, and dry mucous membranes are often present in prerenal azotemia. Congestive heart failure, liver disease, and nephrotic syndrome can be associated with reductions in renal blood flow and/or alterations in intrarenal hemodynamics leading to reduced GFR. Extensive vascular disease raises the possibility of renal artery disease, especially if kidneys are known to be asymmetric in size. Atheroembolic disease can be associated with livedo reticularis and other signs of emboli to the legs. The presence of sepsis is an important clue to causation although, as described above, the detailed pathophysiology may be multifactorial.

A history of prostatic disease, nephrolithiasis, or pelvic or paraaortic malignancy would suggest the possibility of postrenal AKI. Whether or not symptoms are present early during obstruction of the urinary tract depends on the location of obstruction. Colicky flank pain radiating to the groin suggests acute ureteric obstruction. Nocturia and urinary frequency or hesitancy can be seen in prostatic disease. Abdominal fullness and suprapubic pain can accompany bladder enlargement. Definitive diagnosis of obstruction requires radiologic investigations.

A careful review of all medications is imperative in the evaluation of an individual with AKI. Not only are medications frequently a nephrotoxic cause of AKI, but doses of administered medications must be adjusted for reductions in kidney function. In this regard, it is important to recognize that reductions in true GFR are not reflected by equations which estimate GFR since those equations are dependent on SCr and the patient being in a steady state. With AKI, SCr will lag behind changes in filtration rate. Idiosyncratic reactions to a wide variety of medications can lead to allergic interstitial nephritis, which may be accompanied by fever, arthralgias, and a pruritic erythematous rash. The absence of systemic features of hypersensitivity, however, does not exclude the diagnosis of interstitial nephritis, and a kidney biopsy should be considered for definitive diagnosis.

AKI accompanied by palpable purpura, pulmonary hemorrhage, or sinusitis raises the possibility of systemic vasculitis with glomerulonephritis. A history of autoimmune disease, such as systemic lupus erythematosus, should lead to consideration of the possibility that the AKI is related to worsening of this underlying disease. Pregnancy should lead to the consideration of preeclampsia as a pathophysiological contributor to the AKI. A tense abdomen should prompt consideration of acute abdominal compartment syndrome, which requires measurement of bladder pressure. Signs of limb ischemia may be clues to the diagnosis of rhabdomyolysis.

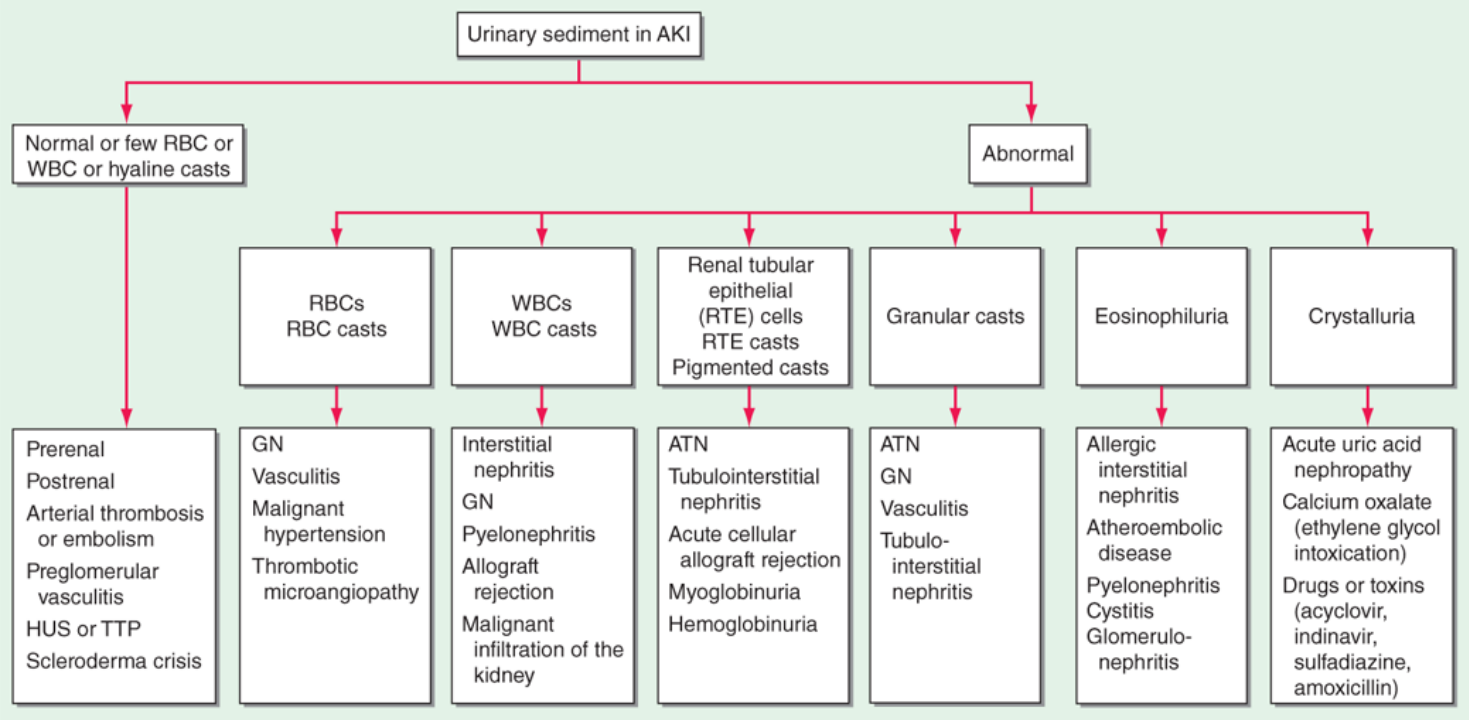
URINE FINDINGS

Complete anuria early in the course of AKI is uncommon except in the following situations: complete urinary tract obstruction, renal artery occlusion, overwhelming septic shock, severe ischemia (often with cortical necrosis), or severe proliferative glomerulonephritis or vasculitis. A reduction in urine output (oliguria, defined as <400 mL/24 h) usually denotes more severe AKI (i.e., lower GFR) than when urine output is preserved. Oliguria is associated with worse clinical outcomes in AKI. Preserved urine output can be seen in nephrogenic diabetes insipidus characteristic of long-standing urinary tract obstruction, tubulointerstitial disease, or nephrotoxicity from cisplatin or aminoglycosides, among other causes. Red or brown urine may be seen with or without gross hematuria; if the color persists in the supernatant after centrifugation, then pigment nephropathy from rhabdomyolysis or hemolysis should be suspected.

The urinalysis and urine sediment examination are invaluable tools, but they require clinical correlation because of generally limited sensitivity and specificity **(Fig. 304-6)** **(Chap. A3)**. In the absence of preexisting proteinuria from CKD, AKI from ischemia or nephrotoxins leads to mild proteinuria (<1 g/d). Greater proteinuria in AKI suggests damage to the glomerular ultrafiltration barrier or excretion of myeloma light chains; the latter are not detected with conventional urine dipsticks (which detect albumin) and require the sulfosalicylic acid test or immunoelectrophoresis. Atheroemboli can cause a variable degree of proteinuria. Extremely heavy proteinuria ("nephrotic range," >3.5 g/d) can occasionally be seen in glomerulonephritis, vasculitis, or toxins/medications that can affect the glomerulus as well as the tubulointerstitium (e.g., NSAIDs). AKI can also complicate cases of minimal change disease, a cause of the nephrotic syndrome **(Chap. 303)**. If the dipstick is positive for hemoglobin but few red blood cells are evident in the urine sediment, then rhabdomyolysis or hemolysis should be suspected.

FIGURE 304-6

Interpretation of urinary sediment findings in acute kidney injury (AKI). ATN, acute tubular necrosis; GN, glomerulonephritis; HUS, hemolytic-uremic syndrome; RBCs, red blood cells; RTE, renal tubular epithelial; TTP, thrombotic thrombocytopenic purpura; WBCs, white blood cells. *(Adapted from L Yang, JV Bonventre: Diagnosis and clinical evaluation of acute kidney injury. In Comprehensive Nephrology, 4th ed. J Floege et al [eds]. Philadelphia, Elsevier, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prerenal azotemia may present with hyaline casts or an unremarkable urine sediment examination. Postrenal AKI may also lead to an unremarkable sediment, but hematuria and pyuria may be seen depending on the cause of obstruction. AKI from ATN due to ischemic injury, sepsis, or certain nephrotoxins has characteristic urine sediment findings: pigmented "muddy brown" granular casts and tubular epithelial cell casts. These findings may be absent in more than 20% of cases, however. Glomerulonephritis may lead to dysmorphic red blood cells or red blood cell casts. Interstitial nephritis may lead to white blood cell casts. The urine sediment findings overlap somewhat in glomerulonephritis and interstitial nephritis, and a diagnosis is not always possible on the basis of the urine sediment alone. Urine eosinophils have a limited role in differential diagnosis; they can be seen in interstitial nephritis, pyelonephritis, cystitis, atheroembolic disease, or glomerulonephritis. Crystalluria may be important diagnostically. The finding of oxalate crystals in AKI should prompt an evaluation for ethylene glycol toxicity. Abundant uric acid crystals may be seen in the tumor lysis syndrome.

BLOOD LABORATORY FINDINGS

Certain forms of AKI are associated with characteristic patterns in the rise and fall of SCr. Prerenal azotemia typically leads to modest rises in SCr that return to baseline with improvement in hemodynamic status. Contrast nephropathy leads to a rise in SCr within 24–48 h, peak within 3–5 days, and resolution within 5–7 days. In comparison, atheroembolic disease usually manifests with more subacute rises in SCr, although severe AKI with rapid increases in SCr can occur in this setting. With many of the epithelial cell toxins such as aminoglycoside antibiotics and cisplatin, the rise in SCr is characteristically delayed for 3–5 days to 2 weeks after initial exposure.

A complete blood count may provide diagnostic clues. Anemia is common in AKI and is usually multifactorial in origin. It is not related to an effect of AKI solely on production of red blood cells because this effect in isolation takes longer to manifest. Peripheral eosinophilia can accompany interstitial nephritis, atheroembolic disease, polyarteritis nodosa, and Churg-Strauss vasculitis. Severe anemia in the absence of bleeding may reflect hemolysis, multiple myeloma, or thrombotic microangiopathy (e.g., hemolytic uremic syndrome [HUS] or TTP). Other laboratory findings of thrombotic microangiopathy include thrombocytopenia, schistocytes on peripheral blood smear, elevated lactate dehydrogenase level, and low haptoglobin content. Evaluation of patients suspected of having TTP or HUS includes measurement of levels of the von Willebrand factor cleaving protease (ADAMTS13) and testing for Shiga toxin–producing *Escherichia coli*. "Atypical HUS" constitutes the majority of adult cases of HUS; genetic testing is important because it is estimated that 60–70% of atypical HUS patients have mutations in genes encoding proteins that regulate the alternative complement pathway.

AKI often leads to hyperkalemia, hyperphosphatemia, and hypocalcemia. Marked hyperphosphatemia with accompanying hypocalcemia, however, suggests rhabdomyolysis or the tumor lysis syndrome. Serum creatine kinase and uric acid levels are often elevated in rhabdomyolysis, while tumor lysis syndrome shows normal or marginally elevated creatine kinase and markedly elevated serum uric acid. The anion gap may be increased with any cause of uremia due to retention of anions such as phosphate, hippurate, sulfate, and urate. The co-occurrence of an increased anion gap and an osmolal gap may suggest ethylene glycol poisoning, which may also cause oxalate crystalluria and oxalate deposition in kidney tissue. Low anion gap may provide a clue to the diagnosis of multiple myeloma due to the presence of unmeasured cationic proteins. Laboratory blood tests helpful for the diagnosis of glomerulonephritis and vasculitis include depressed complement levels and high titers of antinuclear antibodies (ANAs), antineutrophil cytoplasmic antibodies (ANCAs), antiglomerular basement membrane (Anti-GBM) antibodies, and cryoglobulins.

RENAL FAILURE INDICES

Several indices have been used to help differentiate prerenal azotemia from intrinsic AKI when the tubules are malfunctioning. The low tubular flow rate and increased renal medullary recycling of urea seen in prerenal azotemia may cause a disproportionate elevation of the BUN compared to creatinine. Other causes of disproportionate BUN elevation need to be kept in mind, however, including upper gastrointestinal bleeding, hyperalimentation, increased tissue catabolism, and glucocorticoid use.

The FeNa is the fraction of the filtered sodium load that is reabsorbed by the tubules, and is a measure of both the kidney's ability to reabsorb sodium as well as endogenously and exogenously administered factors that affect tubular reabsorption. As such, it depends on sodium intake, effective intravascular volume, GFR, diuretic intake, and intact tubular reabsorptive mechanisms. With prerenal azotemia, the FeNa may be <1%, suggesting avid tubular sodium reabsorption. In patients with CKD, a FeNa significantly >1% can be present despite a superimposed prerenal state. The FeNa may also be >1% despite hypovolemia due to treatment with diuretics. Low FeNa is often seen early in glomerulonephritis and other disorders and, hence, should not be taken as prima facie evidence of prerenal azotemia. Low FeNa is therefore suggestive, but not synonymous, with effective intravascular volume depletion, and should not be used as the sole guide for volume management. The response of urine output to crystalloid or colloid fluid administration may be both diagnostic and therapeutic in prerenal azotemia. In ischemic AKI, the FeNa is frequently >1% because of tubular injury and resultant inability to reabsorb sodium. Several causes of ischemia-associated and nephrotoxin-associated AKI can present with FeNa <1%, however, including sepsis (often early in the course), rhabdomyolysis, and contrast nephropathy.

The ability of the kidney to produce a concentrated urine is dependent upon many factors and reliant on good tubular function in multiple regions of the kidney. In the patient not taking diuretics and with good baseline kidney function, urine osmolality may be >500 mOsm/kg in prerenal azotemia, consistent with an intact medullary concentration gradient and elevated serum vasopressin levels causing water reabsorption resulting in concentrated urine. In elderly patients and those with CKD, however, baseline concentrating defects may exist, making urinary osmolality unreliable in many instances. Loss of concentrating ability is common in most forms of AKI that affect the tubules and interstitium, resulting in urine osmolality <350 mOsm/kg, but the finding is not specific.

RADIOLOGIC EVALUATION

Postrenal AKI should always be considered in the differential diagnosis of AKI because treatment is usually successful if instituted early. Simple bladder catheterization can rule out urethral obstruction. Imaging of the urinary tract with renal ultrasound or CT should be undertaken to investigate obstruction in individuals with AKI unless an alternate diagnosis is apparent. Findings of obstruction include dilation of the collecting system and hydroureteronephrosis. Obstruction can be present without radiologic abnormalities in the setting of volume depletion, retroperitoneal fibrosis, encasement with tumor, and also early in the course of obstruction. If a high-clinical index of suspicion for obstruction persists despite normal imaging, antegrade or retrograde pyelography should be performed. Imaging may also provide additional helpful information about kidney size and echogenicity to assist in the distinction between acute versus CKD. In CKD, kidneys are usually smaller unless the patient has diabetic nephropathy, HIV-associated nephropathy, or infiltrative diseases. Normal sized kidneys are expected in AKI. Enlarged kidneys in a patient with AKI suggests the possibility of acute interstitial nephritis or infiltrative diseases. Vascular imaging may be useful if venous or arterial obstruction is suspected, but the risks of contrast administration should be kept in mind. MRI with gadolinium-based contrast agents should be avoided if possible in severe AKI due to the possibility of inducing nephrogenic system fibrosis, a rare but serious complication seen most commonly in patients with end-stage renal disease.

KIDNEY BIOPSY

If the cause of AKI is not apparent based on the clinical context, physical examination, laboratory studies, and radiologic evaluation, kidney biopsy should be considered. The kidney biopsy can provide definitive diagnostic and prognostic information about acute kidney disease and CKD. The procedure is most often used in AKI when prerenal azotemia, postrenal AKI, and ischemic or nephrotoxic AKI have been deemed unlikely, and other possible diagnoses are being considered such as glomerulonephritis, vasculitis, interstitial nephritis, myeloma kidney, HUS and TTP, and allograft dysfunction. Kidney biopsy is associated with a risk of bleeding, which can be severe and organ- or life-threatening in patients with thrombocytopenia or coagulopathy.

NOVEL BIOMARKERS

BUN and creatinine are functional biomarkers of glomerular filtration rather than tissue injury biomarkers and, therefore, may be suboptimal for the diagnosis of actual parenchymal kidney damage. BUN and creatinine are also relatively slow to rise after kidney injury. Several novel biomarkers have been investigated and show promise for earlier and accurate diagnosis of AKI and for predicting AKI prognosis. In cases of oliguric AKI, the urinary flow rate in response to bolus intravenous furosemide 1.0–1.5 mg/kg can be used a prognostic test: urine output of less than 200 mL over 2 h after intravenous furosemide may identify patients at higher risk of progression to more severe AKI, and the need for renal replacement therapy. The severity or risk of progressive AKI may also be reflected in findings on urine microscopy. In one study involving review of fresh urine sediments by board-certified nephrologists, a greater number of renal tubular epithelial cells and/or granular casts in the urine sediment was associated with both the severity and worsening of AKI. Novel protein biomarkers of kidney injury have also been identified in animal models of AKI and further tested in humans. *Kidney injury molecule-1* (KIM-1) is a type 1 transmembrane protein that is abundantly expressed in proximal tubular cells injured by ischemia or nephrotoxins such as cisplatin. KIM-1 is not expressed in appreciable quantities in the absence of tubular injury or in extrarenal tissues. KIM-1's functional role may be to confer

phagocytic properties to tubular cells, enabling them to clear debris from the tubular lumen after kidney injury and also may reduce the inflammatory response to acute injury. KIM-1 can be detected shortly after ischemic or nephrotoxic injury in the urine and, therefore, may be an easily tested biomarker in the clinical setting. *Neutrophil gelatinase associated lipocalin* (NGAL, also known as lipocalin-2 or siderocalin) is another novel biomarker of AKI. NGAL was first discovered as a protein in granules of human neutrophils. NGAL can bind to iron siderophore complexes and may have tissue-protective effects in the proximal tubule. NGAL is highly upregulated after inflammation and kidney injury and can be detected in the plasma and urine within 2 h of cardiopulmonary bypass–associated AKI. In 2014, the U.S. Food and Drug Administration approved the marketing of a test based on the combination of the urinary concentrations of two cell-cycle arrest biomarkers, insulin-like growth factor binding protein 7 (IGFBP7) and tissue inhibitor of metalloproteinase-2 (TIMP-2) as predictive biomarkers for higher risk of the development of moderate to severe AKI in critically ill patients. A number of other biomarkers are under investigation for early and accurate identification of AKI and for risk stratification to identify individuals at increased risk. The optimal use of novel AKI biomarkers in clinical settings is an area of ongoing investigation.

COMPLICATIONS OF AKI

The kidney plays a central role in homeostatic control of volume status, blood pressure, plasma electrolyte composition, and acid-base balance, and for excretion of nitrogenous and other waste products. Complications associated with AKI are, therefore, protean, and depend on the severity of AKI and other associated conditions. Mild to moderate AKI may be entirely asymptomatic, particularly early in the course.

UREMIA

Buildup of nitrogenous waste products, manifested as an elevated BUN concentration, is a hallmark of AKI. BUN itself poses little direct toxicity at levels <100 mg/dL. At higher concentrations, mental status changes and bleeding complications can arise. Other toxins normally cleared by the kidney may be responsible for the symptom complex known as uremia. Few of the many possible uremic toxins have been definitively identified. The correlation of BUN and SCr concentrations with uremic symptoms is extremely variable, due in part to differences in urea and creatinine generation rates across individuals.

HYPERVOLEMIA AND HYPOVOLEMIA

Expansion of extracellular fluid volume is a major complication of oliguric and anuric AKI, due to impaired salt and water excretion. The result can be weight gain, dependent edema, increased jugular venous pressure, and pulmonary edema; the latter can be life threatening. Pulmonary edema can also occur from volume overload and hemorrhage in pulmonary renal syndromes. AKI may also induce or exacerbate acute lung injury characterized by increased vascular permeability and inflammatory cell infiltration in lung parenchyma. Recovery from AKI can sometimes be accompanied by polyuria, which, if untreated, can lead to significant volume depletion. The polyuric phase of recovery may be due to an osmotic diuresis from retained urea and other waste products as well as delayed recovery of tubular reabsorptive functions.

HYPONATREMIA

Abnormalities in plasma electrolyte composition can be mild or life threatening. The dysfunctional kidney has limited ability to regulate electrolyte balance. Administration of excessive hypotonic crystalloid or isotonic dextrose solutions can result in hypoosmolality and hyponatremia, which, if severe, can cause neurologic abnormalities, including seizures.

HYPERKALEMIA

An important complication of AKI is hyperkalemia. Marked hyperkalemia is particularly common in rhabdomyolysis, hemolysis, and tumor lysis syndrome due to release of intracellular potassium from damaged cells. Muscle weakness may be a symptom of hyperkalemia. Potassium affects the cellular membrane potential of cardiac and neuromuscular tissues. The more serious complication of hyperkalemia is due to effects on cardiac conduction, leading to potentially fatal arrhythmias.

ACIDOSIS

Metabolic acidosis, usually accompanied by an elevation in the anion gap, is common in AKI, and can further complicate acid-base and potassium balance in individuals with other causes of acidosis, including sepsis, diabetic ketoacidosis, or respiratory acidosis.

HYPERPHOSPHATEMIA AND HYPOCALCEMIA

AKI can lead to hyperphosphatemia, particularly in highly catabolic patients or those with AKI from rhabdomyolysis, hemolysis, and tumor lysis syndrome. Metastatic deposition of calcium phosphate can lead to hypocalcemia. AKI-associated hypocalcemia may also arise from derangements in the vitamin D–parathyroid hormone–fibroblast growth factor-23 axis. Hypocalcemia is often asymptomatic but can lead to perioral paresthesias, muscle cramps, seizures, carpopedal spasms, and prolongation of the QT interval on electrocardiography. Calcium levels should be corrected for the degree of hypoalbuminemia, if present, or ionized calcium levels should be followed. Mild, asymptomatic hypocalcemia does not require treatment.

BLEEDING

Hematologic complications of AKI include anemia and bleeding, both of which are exacerbated by coexisting disease processes such as sepsis, liver disease, and disseminated intravascular coagulation. Direct hematologic effects from AKI-related uremia include decreased erythropoiesis and platelet dysfunction.

INFECTIONS

Infections are a common precipitant of AKI and also a dreaded complication of AKI. Impaired host immunity has been described in end-stage renal disease and may be operative in severe AKI.

CARDIAC COMPLICATIONS

The major cardiac complications of AKI are arrhythmias, pericarditis, and pericardial effusion. In addition, volume overload and uremia may lead to cardiac injury and impaired cardiac function. In animal studies cellular apoptosis and capillary vascular congestion as well as mitochondrial dysfunction have been described in the heart after renal ischemia reperfusion.

MALNUTRITION

AKI is often a severely hypercatabolic state, and therefore, malnutrition is a major complication.

PREVENTION AND TREATMENT OF AKI

The management of individuals with and at risk for AKI varies according to the underlying cause **(Table 304-2)**. Common to all are several principles. Optimization of hemodynamics, correction of fluid and electrolyte imbalances, discontinuation of nephrotoxic medications, and dose adjustment of administered medications are all critical. Common causes of AKI such as sepsis and ischemic ATN do not yet have specific therapies once injury is established, but meticulous clinical attention is needed to support the patient until (if) AKI resolves. The kidney possesses remarkable capacity to repair itself after even severe, dialysis-requiring AKI, when baseline renal function was intact. However, many patients with AKI, particularly when superimposed on preexisting CKD, do not recover fully and may remain dialysis dependent. It has become increasingly apparent that AKI predisposes to accelerated progression of CKD, and CKD is an important risk factor for AKI.

TABLE 304-2

Management of Acute Kidney Injury

General Issues

1. Optimization of systemic and renal hemodynamics through volume resuscitation and judicious use of vasopressors
2. Elimination of nephrotoxic agents (e.g., ACE inhibitors, ARBs, NSAIDs, aminoglycosides) if possible
3. Initiation of renal replacement therapy when indicated

Specific Issues

1. Nephrotoxin-specific
   a. Rhabdomyolysis: aggressive intravenous fluids; consider forced alkaline diuresis
   b. Tumor lysis syndrome: aggressive intravenous fluids and allopurinol or rasburicase
2. Volume overload
   a. Salt and water restriction
   b. Diuretics
   c. Ultrafiltration
3. Hyponatremia
   a. Restriction of enteral free water intake, minimization of hypotonic intravenous solutions including those containing dextrose
   b. Hypertonic saline is rarely necessary in AKI. Vasopressin antagonists are generally not needed.
4. Hyperkalemia
   a. Restriction of dietary potassium intake
   b. Discontinuation of potassium-sparing diuretics, ACE inhibitors, ARBs, NSAIDs
   c. Loop diuretics to promote urinary potassium loss
   d. Potassium binding ion-exchange resin (sodium polystyrene sulfonate)
   e. Insulin (10 units regular) and glucose (50 mL of 50% dextrose) to promote entry of potassium intracellularly
   f. Inhaled beta-agonist therapy to promote entry of potassium intracellularly
   g. Calcium gluconate or calcium chloride (1 g) to stabilize the myocardium
5. Metabolic acidosis
   a. Sodium bicarbonate (if pH <7.2 to keep serum bicarbonate >15 mmol/L)
   b. Administration of other bases, e.g., THAM
   c. Renal replacement therapy
6. Hyperphosphatemia
   a. Restriction of dietary phosphate intake
   b. Phosphate binding agents (calcium acetate, sevelamer hydrochloride, aluminum hydroxide—taken with meals)
7. Hypocalcemia
   a. Calcium carbonate or calcium gluconate if symptomatic
8. Hypermagnesemia
   a. Discontinue Mg^{2+} containing antacids
9. Hyperuricemia
   a. Acute treatment is usually not required except in the setting of tumor lysis syndrome (see above)
10. Nutrition
    a. Sufficient protein and calorie intake (20–30 kcal/kg per day) to avoid negative nitrogen balance. Nutrition should be provided via the enteral route if possible.
11. Drug dosing
    a. Careful attention to dosages and frequency of administration of drugs, adjustment for degree of renal failure
    b. Note that serum creatinine concentration may overestimate renal function in the non–steady state characteristic of patients with AKI

Abbreviations: ACE, angiotensin-converting enzyme; ARBs, angiotensin receptor blocker; NSAIDs, nonsteroidal anti-inflammatory drug; THAM, tris (hydroxymethyl) aminomethane.

Prerenal Azotemia

Prevention and treatment of prerenal azotemia require optimization of renal perfusion. The composition of replacement fluids should be targeted to the type of fluid lost. Severe acute blood loss should be treated with packed red blood cells. Isotonic crystalloid and/or colloid should be used for less severe

acute hemorrhage or plasma loss in the case of burns and pancreatitis. Crystalloid solutions are less expensive and probably equally efficacious as colloid solutions. Hydroxyethyl starch solutions increase the risk of severe AKI and are contraindicated. Crystalloid has been reported to be preferable to albumin in the setting of traumatic brain injury. Isotonic crystalloid (e.g., 0.9% saline) or colloid should be used for volume resuscitation in severe hypovolemia, whereas hypotonic crystalloids (e.g., 0.45% saline) suffice for less severe hypovolemia and can also be used in the setting of hypernatremia. Excessive chloride administration from 0.9% saline may lead to hyperchloremic metabolic acidosis and may impair GFR. Bicarbonate-containing solutions (e.g., dextrose water with 150 mEq sodium bicarbonate) can be used if metabolic acidosis is a concern. Whether buffered crystalloid solutions containing bicarbonate or lactate offer advantages over normal saline for volume repletion in most critically ill patients is not yet established.

Optimization of cardiac function in AKI may require use of inotropic agents, preload- and afterload-reducing agents, antiarrhythmic drugs, and mechanical aids such as ventricular assist devices. Invasive hemodynamic monitoring to guide therapy may be necessary.

Cirrhosis and Hepatorenal Syndrome

Fluid management in individuals with cirrhosis, ascites, and AKI is challenging because of the frequent difficulty in ascertaining intravascular volume status. Administration of intravenous fluids as a volume challenge may be required diagnostically as well as therapeutically. Excessive volume administration may, however, result in worsening ascites and pulmonary compromise in the setting of hepatorenal syndrome or AKI due to superimposed spontaneous bacterial peritonitis. Peritonitis should be ruled out by culture of ascitic fluid. Albumin may prevent AKI in those treated with antibiotics for spontaneous bacterial peritonitis. The definitive treatment of the hepatorenal syndrome is orthotopic liver transplantation. Bridge therapies that have shown promise include terlipressin (a vasopressin analog), combination therapy with octreotide (a somatostatin analog) and midodrine (an α_1-adrenergic agonist), and norepinephrine, in combination with intravenous albumin (25–50 g, maximum 100 g/d).

Intrinsic AKI

Several agents have been tested and have failed to show benefit in the treatment of acute tubular injury. These include atrial natriuretic peptide, low-dose dopamine, endothelin antagonists, erythropoietin, loop diuretics, calcium channel blockers, α-adrenergic receptor blockers, prostaglandin analogs, antioxidants, antibodies against leukocyte adhesion molecules, and insulin-like growth factor, among many others. Most studies have enrolled patients with severe and well-established AKI, and treatment may have been initiated too late. Novel kidney injury biomarkers may provide an opportunity to test agents earlier in the course of AKI.

AKI due to acute glomerulonephritis or vasculitis may respond to immunosuppressive agents and/or plasmapheresis (Chap. 303). Allergic interstitial nephritis due to medications requires discontinuation of the offending agent. Glucocorticoids have been used, but not tested in randomized trials, in cases where AKI persists or worsens despite discontinuation of the suspected medication. AKI due to scleroderma (scleroderma renal crisis) should be treated with ACE inhibitors. Idiopathic TTP-HUS is a medical emergency and should be treated promptly with plasma exchange. Pharmacologic blockade of complement activation may be effective in atypical HUS.

Early and aggressive volume repletion is mandatory in patients with rhabdomyolysis, who may initially require 10 L of fluid per day. Alkaline fluids (e.g., 75 mmol/L sodium bicarbonate added to 0.45% saline) may be beneficial in preventing tubular injury and cast formation, but carry the risk of worsening hypocalcemia. Diuretics may be used if fluid repletion is adequate but unsuccessful in achieving urinary flow rates of 200–300 mL/h. There is no specific therapy for established AKI in rhabdomyolysis, other than dialysis in severe cases or general supportive care to maintain fluid and electrolyte balance and tissue perfusion. Careful attention must be focused on calcium and phosphate status because of precipitation in damaged tissue and release when the tissue heals.

Postrenal AKI

Prompt recognition and relief of urinary tract obstruction can forestall the development of permanent structural damage induced by urinary stasis. The site of obstruction defines the treatment approach. Transurethral or suprapubic bladder catheterization may be all that is needed initially for urethral strictures or functional bladder impairment. Ureteric obstruction may be treated by percutaneous nephrostomy tube placement or ureteral stent placement. Relief of obstruction is usually followed by an appropriate diuresis for several days. In rare cases, severe polyuria persists due to tubular dysfunction and may require continued administration of intravenous fluids and electrolytes for a period of time.

SUPPORTIVE MEASURES FOR AKI

Volume Management

Hypervolemia in oliguric or anuric AKI may be life threatening due to acute pulmonary edema, especially because many patients have coexisting pulmonary disease, and AKI likely increases pulmonary vascular permeability. Fluid and sodium should be restricted, and diuretics may be used to increase the urinary flow rate. There is no evidence that increasing urine output itself improves the natural history of AKI, but diuretics may help to avoid the need for dialysis in some cases. In severe cases of volume overload, furosemide may be given as a bolus (200 mg) followed by an intravenous drip (10–40 mg/h), with or without a thiazide diuretic. In decompensated heart failure, stepped diuretic therapy was found to be superior to ultrafiltration in preserving renal function. Diuretic therapy should be stopped if there is no response. Dopamine in low doses may transiently increase salt and water excretion by the kidney

in prerenal states, but clinical trials have failed to show any benefit in patients with intrinsic AKI. Because of the risk of arrhythmias and potential bowel ischemia, the risks of dopamine outweigh the benefits if used specifically for the treatment or prevention of AKI.

Electrolyte and Acid-Base Abnormalities

The treatment of dysnatremias and hyperkalemia is described in Chap. 49. Metabolic acidosis is generally not treated unless severe (pH <7.20 and serum bicarbonate <15 mmol/L). Acidosis can be treated with oral or intravenous sodium bicarbonate **(Chap. 51)**, but overcorrection should be avoided because of the possibility of metabolic alkalosis, hypocalcemia, hypokalemia, and volume overload. Hyperphosphatemia is common in AKI and can usually be treated by limiting intestinal absorption of phosphate using phosphate binders (calcium carbonate, calcium acetate, lanthanum, sevelamer, or aluminum hydroxide). Hypocalcemia does not usually require therapy unless symptoms are present. Ionized calcium should be monitored rather than total calcium when hypoalbuminemia is present.

Malnutrition

Protein energy wasting is common in AKI, particularly in the setting of multisystem organ failure. Inadequate nutrition may lead to starvation ketoacidosis and protein catabolism. Excessive nutrition may increase the generation of nitrogenous waste and lead to worsening azotemia. Total parenteral nutrition requires large volumes of fluid administration and may complicate efforts at volume control. According to the Kidney Disease Improving Global Outcomes (KDIGO) guidelines, patients with AKI should achieve a total energy intake of 20–30 kcal/kg per day. Protein intake should vary depending on the severity of AKI: 0.8–1.0 g/kg per day in noncatabolic AKI without the need for dialysis; 1.0–1.5 g/kg per day in patients on dialysis; and up to a maximum of 1.7 g/kg per day if hypercatabolic and receiving continuous renal replacement therapy. Trace elements and water-soluble vitamins should also be supplemented in AKI patients treated with dialysis and continuous renal replacement therapy.

Anemia

The anemia seen in AKI is usually multifactorial and is not improved by erythropoiesis-stimulating agents, due to their delayed onset of action and the presence of bone marrow resistance in critically ill patients. Uremic bleeding may respond to desmopressin or estrogens, but may require dialysis for treatment in the case of long-standing or severe uremia. Gastrointestinal prophylaxis with proton pump inhibitors or histamine (H_2) receptor blockers is required. It is important to recognize, however, that protein pump inhibitors have been associated with AKI from interstitial nephritis, a relationship that is increasingly being recognized. Venous thromboembolism prophylaxis is important and should be tailored to the clinical setting; low-molecular-weight heparins and factor Xa inhibitors have unpredictable pharmacokinetics in severe AKI and should be avoided.

Dialysis Indications and Modalities

Dialysis is indicated when medical management fails to control volume overload, hyperkalemia, or acidosis; in some toxic ingestions; and when there are severe complications of uremia (asterixis, pericardial rub or effusion, encephalopathy, uremic bleeding) **(See also Chap. 306)**. The timing of dialysis is still a matter of debate. Late initiation of dialysis carries the risk of avoidable volume, electrolyte, and metabolic complications of AKI. On the other hand, initiating dialysis too early may unnecessarily expose individuals to intravenous lines and invasive procedures, with the attendant risks of infection, bleeding, procedural complications, and hypotension. The initiation of dialysis should not await the development of a life-threatening complication of renal failure. Many nephrologists initiate dialysis for AKI empirically when the BUN exceeds a certain value (e.g., 100 mg/dL) in patients without clinical signs of recovery of kidney function. The available modes for renal replacement therapy in AKI require either access to the peritoneal cavity (for peritoneal dialysis) or the large blood vessels (for hemodialysis, hemofiltration, and other hybrid procedures). Small solutes are removed across a semipermeable membrane down their concentration gradient (“diffusive” clearance) and/or along with the movement of plasma water (“convective” clearance). The choice of modality is often dictated by the immediate availability of technology and the expertise of medical staff.

Hemodialysis can be used intermittently or continuously and can be done through convective clearance, diffusive clearance, or a combination of the two. Vascular access is through the femoral, internal jugular, or subclavian veins. Hemodialysis is an intermittent procedure that removes solutes through diffusive and convective clearance. Hemodialysis is typically performed 3–4 h per day, three to four times per week, and is the most common form of renal replacement therapy for AKI. One of the major complications of hemodialysis is hypotension, particularly in the critically ill, which can perpetuate AKI by causing ischemic injury to the recovering organ.

Continuous intravascular procedures were developed in the early 1980s to treat hemodynamically unstable patients without inducing the rapid shifts of volume, osmolarity, and electrolytes characteristic of intermittent hemodialysis. Continuous renal replacement therapy (CRRT) can be performed by convective clearance (continuous venovenous hemofiltration [CVVH]), in which large volumes of plasma water (and accompanying solutes) are forced across the semipermeable membrane by means of hydrostatic pressure; the plasma water is then replaced by a physiologic crystalloid solution. CRRT can also be performed by diffusive clearance (continuous venovenous hemodialysis [CVVHD]), a technology similar to hemodialysis except at lower blood flow and dialysate flow rates. A hybrid therapy combines both diffusive and convective clearance (continuous venovenous hemodiafiltration [CVVHDF]). To achieve some of the advantages of CRRT without the need for 24-h staffing of the procedure, some physicians favor slow low-efficiency dialysis (SLED) or extended daily dialysis (EDD). In this therapy, blood flow and dialysate flow are higher than in CVVHD, but the treatment time is reduced to ≤12 h.

The optimal dose of dialysis for AKI is not clear. Daily intermittent hemodialysis and high-dose CRRT do not confer a demonstrable survival or renal recovery advantage, but care should be taken to avoid undertreatment. Studies have failed to show that continuous therapies are superior to intermittent therapies. If available, CRRT is often preferred in patients with severe hemodynamic instability, cerebral edema, or significant volume overload.

Peritoneal dialysis can be performed through a temporary intraperitoneal catheter, although it is rarely used in the United States for AKI in adults. Peritoneal dialysis has enjoyed widespread use internationally, particularly when hemodialysis technology is not as readily available. Dialysate solution is instilled into and removed from the peritoneal cavity at regular intervals in order to achieve diffusive and convective clearance of solutes across the peritoneal membrane; ultrafiltration of water is achieved by the presence of an osmotic gradient across the peritoneal membrane achieved by high concentrations of dextrose in the dialysate solution. Because of its continuous nature, it is often better tolerated than intermittent procedures like hemodialysis in hypotensive patients. Peritoneal dialysis may not be sufficient for hypercatabolic patients due to inherent limitations in dialysis efficacy.

OUTCOME AND PROGNOSIS

The development of AKI is associated with a significantly increased risk of in-hospital and long-term mortality, longer length of stay, and increased costs. Prerenal azotemia, with the exception of the cardiorenal and hepatorenal syndromes, and postrenal azotemia carry a better prognosis than most cases of intrinsic AKI. The kidneys may recover even after severe, dialysis-requiring AKI. Survivors of an episode of AKI requiring temporary dialysis, however, are at extremely high risk for progressive CKD, and up to 10% may develop end-stage renal disease. Postdischarge care under the supervision of a nephrologist for aggressive secondary prevention of kidney disease is prudent. Patients with AKI are more likely to die prematurely after they leave the hospital even if their kidney function has recovered.

FURTHER READING

BONVENTRE JV, YANG L: Cellular pathophysiology of ischemic acute kidney injury. J Clin Invest 121:4210, 2011. [PubMed: 22045571]

GAUDRY S et al: Initiation strategies for renal-replacement therapy in the intensive care unit. N Engl J Med 375:122, 2016. [PubMed: 27181456]

KIDNEY DISEASE: IMPROVING GLOBAL OUTCOMES (KDIGO) Acute Kidney Injury Work Group. KDIGO Clinical Practice Guidelines for Acute Kidney Injury. Kidney Int. Supp 2:1, 2012.

LAMEIRE NH et al: Acute kidney injury: An increasing global concern. Lancet 382:170, 2013. [PubMed: 23727171]

ZENG X et al: Incidence, outcomes, and comparisons across definitions of AKI in hospitalized individuals. Clin J Am Soc Nephrol 9:12, 2014. [PubMed: 24178971]

Chapter 305: Chronic Kidney Disease

Joanne M. Bargman; Karl L. Skorecki

INTRODUCTION

Chronic kidney disease (CKD) encompasses a spectrum of pathophysiologic processes associated with abnormal kidney function and a progressive decline in glomerular filtration rate (GFR). The risk of CKD progression is closely linked to both the GFR and the amount of albuminuria. Figure 305-1 provides a staging of CKD stratified by the estimates of both of these parameters.

FIGURE 305-1

Kidney Disease Improving Global Outcome (KDIGO) classification of chronic kidney disease (CKD). Gradation of color from green to red corresponds to increasing risk and progression of CKD. GFR, glomerular filtration rate. *(Reproduced with permission from Kidney Int Suppl 3:5–14, 2013.)*

Prognosis of CKD by GFR and albuminuria categories: KDIGO 2012

GFR categories (ml/min/1.73 m²) description and range				Persistent albuminuria categories description and range		
				A1	A2	A3
				Normal to mildly increased	Moderately increased	Severely increased
				<30 mg/g <3 mg/mmol	30–300 mg/g 3–30 mg/mmol	>300 mg/g >30 mg/mmol
	G1	Normal or high	≥90			
	G2	Mildly decreased	60–89			
	G3a	Mildly to moderately decreased	45–59			
	G3b	Moderately to severely decreased	30–44			
	G4	Severely decreased	15–29			
	G5	Kidney failure	<15			

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The dispiriting term *end-stage renal disease* represents a stage of CKD where the accumulation of toxins, fluid, and electrolytes normally excreted by the kidneys leads to death unless the toxins are removed by renal replacement therapy, using dialysis or kidney transplantation. These interventions are discussed in Chaps. 306 and 307. *End-stage renal disease* will be supplanted in this chapter by the term *stage 5 CKD*.

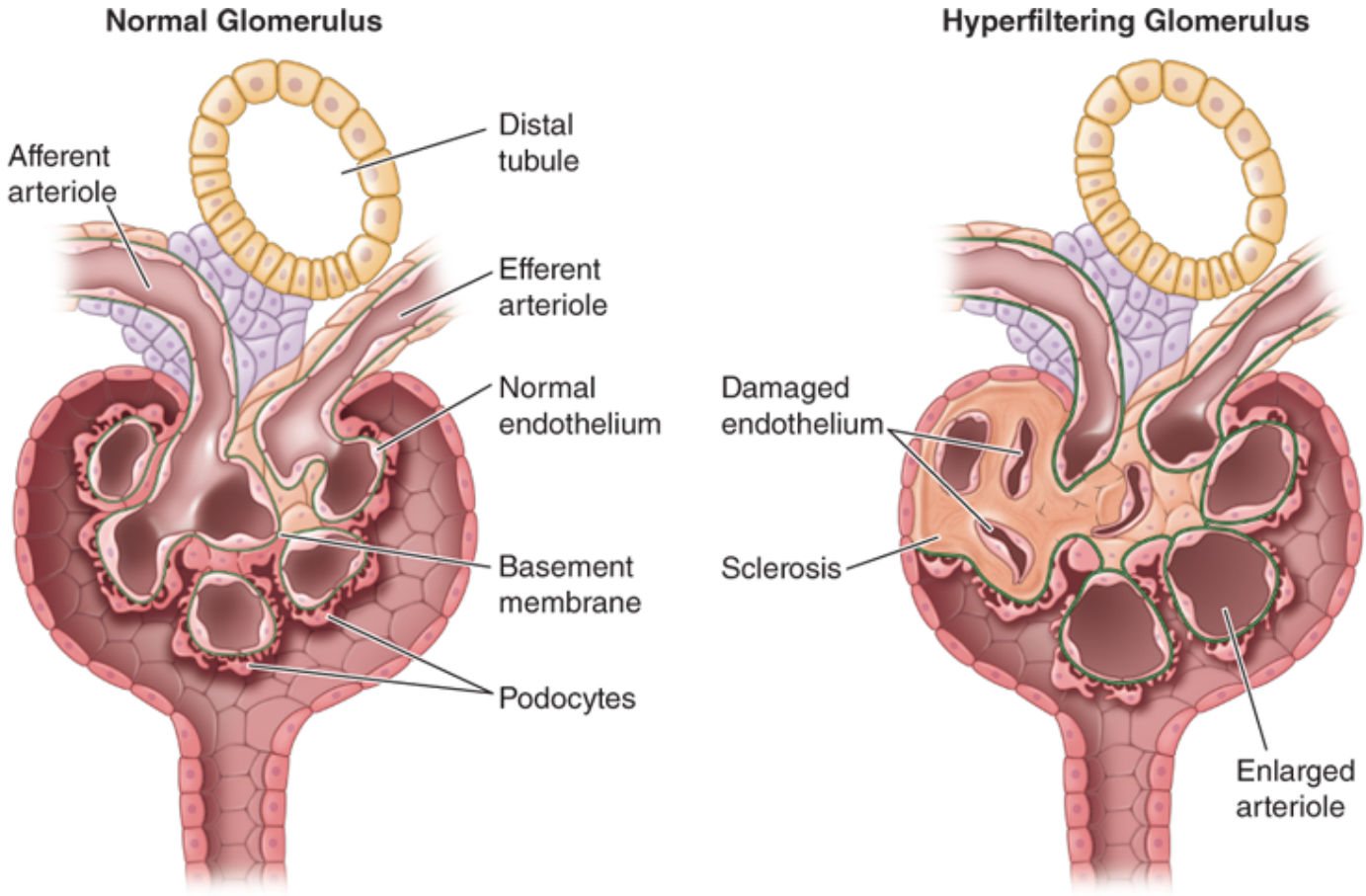
PATHOPHYSIOLOGY OF CKD

The pathophysiology of CKD involves two broad sets of mechanisms of damage: (1) initiating mechanisms specific to the underlying etiology (e.g., abnormalities in kidney development or integrity, immune complex deposition and inflammation in certain types of glomerulonephritis, or toxin exposure in certain diseases of the renal tubules and interstitium) and (2) hyperfiltration and hypertrophy of the remaining viable nephrons, that are a common consequence following long-term reduction of renal mass, irrespective of underlying etiology and lead to further decline in kidney function (Chap. 333e from the 19th edition of *Harrison's*). The responses to reduction in nephron number are mediated by vasoactive hormones, cytokines, and growth factors. Eventually, these short-term adaptations of hyperfiltration and hypertrophy to maintain GFR become maladaptive as the increased pressure and flow within the nephron predisposes to distortion of glomerular architecture, abnormal podocyte function, and

disruption of the filtration barrier leading to sclerosis and dropout of the remaining nephrons (Fig. 305-2). Increased intrarenal activity of the renin-angiotensin system (RAS) appears to contribute both to the initial compensatory hyperfiltration and to the subsequent maladaptive hypertrophy and sclerosis. This process explains why a reduction in renal mass from an isolated insult may lead to a progressive decline in renal function over many years (Fig. 305-3).

FIGURE 305-2

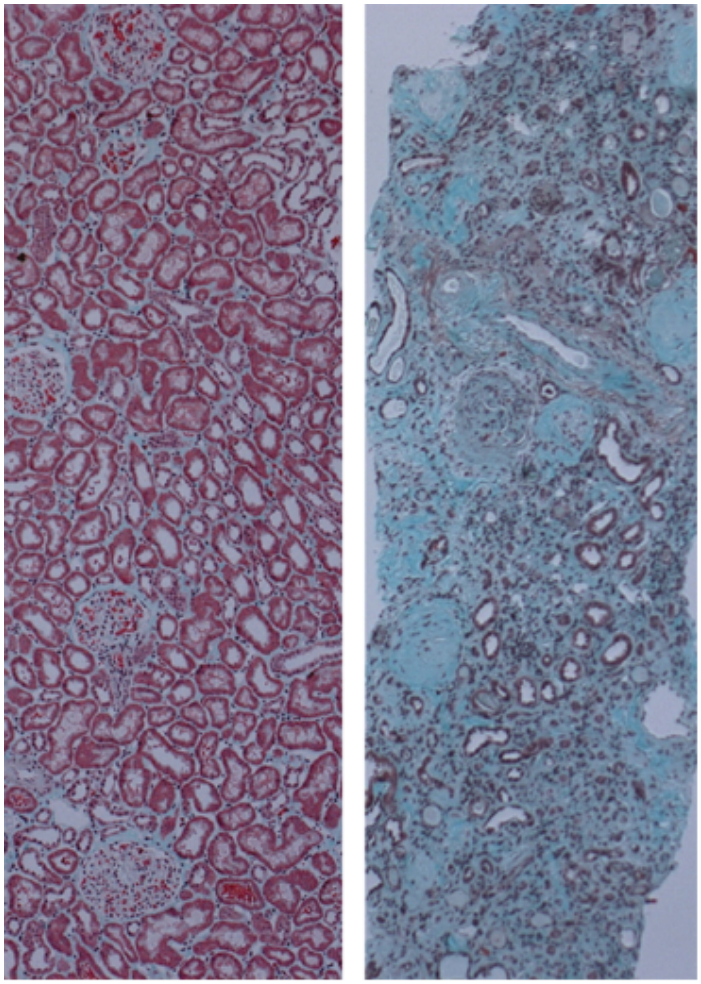
Left: Schema of the normal glomerular architecture. ***Right:*** Secondary glomerular changes associated with a reduction in nephron number, including enlargement of capillary lumens and focal adhesions, which are thought to occur consequent to compensatory hyperfiltration and hypertrophy in the remaining nephrons. *(Modified from JR Ingelfinger: N Engl J Med 348:99, 2003.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 305-3

Left: Low-power photomicrograph of a normal kidney showing normal glomeruli and healthy tubulointerstitium without fibrosis. ***Right:*** Low-power photomicrograph of chronic kidney disease with sclerosis of many glomeruli and severe tubulointerstitial fibrosis (Masson trichrome, 40× magnification). *(Slides courtesy of the late Dr. Andrew Herzenberg.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IDENTIFICATION OF RISK FACTORS AND STAGING OF CKD

It is important to identify factors that increase the risk for CKD, even in individuals with normal GFR. Risk factors include small for gestation birth weight, childhood obesity, hypertension, diabetes mellitus, autoimmune disease, advanced age, African ancestry, a family history of kidney disease, a previous episode of acute kidney injury, and the presence of proteinuria, abnormal urinary sediment, or structural abnormalities of the urinary tract. It has been increasingly recognized that one or more episodes of acute kidney injury are associated with an increased risk of developing CKD.

Many rare inherited forms of CKD follow a Mendelian inheritance pattern, often as part of a systemic syndrome, with the most common in this category being autosomal dominant polycystic kidney disease. In addition, recent research in the genetics of predisposition to common complex diseases (Chap. 456) has revealed DNA sequence variants at a number of genetic loci that are associated with common forms of CKD. A striking example is the finding of allelic versions of the *APOL1* gene, of West African population ancestry, which contributes to the several-fold higher frequency of certain common etiologies of nondiabetic CKD (e.g., focal segmental glomerulosclerosis) observed among African and Hispanic Americans, in major regions of continental Africa and the global African diaspora. The prevalence in West African populations seems to have arisen as an evolutionary adaptation

conferring protection from tropical pathogens. As in other common diseases with a heritable component, environmental triggers (such as a viral pathogen) transform genetic risk into disease.

To stage CKD, it is necessary to estimate the GFR rather than relying on serum creatinine concentration (Table 305-1). Many laboratories now report an estimated GFR, or eGFR, using one of these equations. These equations are valid only if the patient is in steady state, that is, the serum creatinine is neither rising nor falling over days.

TABLE 305-1

Recommended Equations for Estimation of Glomerular Filtration Rate (GFR) Using Serum Creatinine Concentration (S_{CR}), Age, Sex, Race, and Body Weight

1. Equation from the Modification of Diet in Renal Disease Study
Estimated GFR (mL/min per 1.73 m^2) = $1.86 \times (S_{Cr})^{-1.154} \times (age)^{-0.203}$
Multiply by 0.742 for women
Multiply by 1.21 for African ancestry
2. CKD-EPI Equation
$GFR = 141 \times \min(S_{Cr}/\kappa, 1)^{\alpha} \times \max(S_{Cr}/\kappa, 1)^{-1.209} \times 0.993^{Age}$
Multiply by 1.018 for women
Multiply by 1.159 for African ancestry
where S_{Cr} is serum creatinine in mg/dL, κ is 0.7 for females and 0.9 for males, α is –0.329 for females and –0.411 for males, min indicates the minimum of S_{Cr}/κ or 1, and max indicates the maximum of S_{Cr}/κ or 1.

Abbreviation: CKD-EPI, Chronic Kidney Disease Epidemiology Collaboration.

The normal annual mean decline in GFR with age from the peak GFR (~120 mL/min per 1.73 m^2) attained during the third decade of life is ~1 mL/min per year per 1.73 m^2, reaching a mean value of 70 mL/min per 1.73 m^2 at age 70, with considerable inter-individual variability. Although reduced GFR is expected with aging, the lower GFR signifies a true loss of kidney function with attendant consequences in terms of risk of CKD complications, and requirement for dose adjustment of medications. The mean GFR is lower in women than in men. For example, a woman in her eighties with a laboratory report of serum creatinine in the normal range may have a GFR of <50 mL/min per 1.73 m^2. Relatedly, even a mild elevation in serum creatinine concentration often signifies a substantial reduction in GFR in older individuals.

Measurement of albuminuria is also helpful for monitoring nephron injury and the response to therapy in many forms of CKD, especially chronic glomerular diseases. The cumbersome 24-h urine collection has been replaced by measurement of urinary albumin to creatinine ratio (UACR) in one and preferably several spot first-morning urine samples as a measure pointing to glomerular injury. Even in patients with negative conventional dipstick tests for elevated total protein excretion, UACR above 17 mg albumin/g creatinine in men and 25 mg albumin/g creatinine in women serves as a marker not only for early detection of primary kidney disease, but for systemic microvascular disease as well. The presence of albuminuria in general serves as a well-studied screening marker for the presence of systemic microvascular disease and endothelial dysfunction.

A Kidney Failure Risk (KFR) equation has been devised to predict the risk of progression to stage 5 dialysis-dependent kidney disease. The equation is available on many sites online (for example, *www.kidneyfailurerisk.com*) and uses age, sex, region (North American or non-North American), GFR and the urine albumin/creatinine. It has been validated in several cohorts around the world, although the risk for progression appears to be greater in North America, accounting for the regional adjustment in the equation.

Stages 1 and 2 CKD are usually asymptomatic, such that the recognition of CKD occurs more often as a result of laboratory testing in clinical settings other than suspicion of kidney disease. Moreover, in the absence of the risk factors noted above, population-wide screening is not recommended. With progression to CKD stages 3 and 4, clinical and laboratory complications become more prominent. Virtually all organ systems are affected, but the most evident complications include anemia and associated easy fatigability; decreased appetite with progressive malnutrition; abnormalities in calcium, phosphorus, and mineral-regulating hormones, such as $1,25(OH)_2D_3$ (calcitriol), parathyroid hormone (PTH), and fibroblast growth factor 23 (FGF-23); and abnormalities in sodium, potassium, water, and acid-base homeostasis. Many patients, especially the elderly, will have eGFR values compatible with stage 2 or 3 CKD. However, the majority of these patients will show no further deterioration of renal function. The primary care physician is advised to recheck kidney function, and if it is stable and not associated with proteinuria, the patient can usually be followed with interval repeat testing without referral to nephrologist. However, caution should be exercised in terms of potential exposure to nephrotoxins or interventions that risk acute kidney injury (AKI) and also with respect to medication dose adjustment. If repeat testing shows declining GFR, albuminuria, or uncontrolled hypertension, referral to a nephrologist is appropriate. If the patient progresses to stage 5 CKD, toxins accumulate such that patients usually experience a marked disturbance in their activities of daily living, well-being, nutritional status, and water and electrolyte homeostasis, eventuating in the *uremic syndrome*.

ETIOLOGY AND EPIDEMIOLOGY

It has been estimated from population data that at least 6% of the adult population in the United States has CKD at stages 1 and 2. An additional 4.5% of the U.S. population is estimated to have stages 3 and 4 CKD. **Table 305-2** lists the five most frequent categories of causes of CKD, cumulatively accounting for >90% of the CKD disease burden worldwide. The relative contribution of each category varies among different geographic regions. The most frequent cause of CKD in North America and Europe is diabetic nephropathy, most often secondary to type 2 diabetes mellitus. Patients with newly diagnosed CKD often have hypertension. When no overt evidence for a primary glomerular or tubulointerstitial kidney disease process is present, CKD is frequently attributed to hypertension. However, it is now appreciated that such individuals can be considered in two categories. The first includes patients with a subclinical primary glomerulopathy, such as focal segmental or global glomerulosclerosis **(Chap. 308)**. The second includes patients in whom progressive nephrosclerosis and hypertension is the renal correlate of a systemic vascular disease, often also involving large- and small-vessel cardiac and cerebral pathology. This latter combination is especially common in the elderly, in whom chronic renal ischemia as a cause of CKD may be underdiagnosed. The increasing incidence of CKD in the elderly has been ascribed, in part, to decreased mortality rate from the cardiac and cerebral complications of atherosclerotic vascular disease, enabling a greater segment of the population to progress to more advanced stages of CKD. Nevertheless, it should be appreciated that the majority of patients with early stages of CKD succumb to cardiovascular and cerebrovascular complications

before they progress to the more advanced stages of CKD. Indeed, even a minor decrement in GFR or the presence of albuminuria is now recognized as a major risk factor for cardiovascular disease.

TABLE 305-2

Leading Categories of Etiologies of CKD[a]

- Diabetic nephropathy
- Glomerulonephritis
- Hypertension-associated CKD (includes vascular and ischemic kidney disease and primary glomerular disease with associated hypertension)
- Autosomal dominant polycystic kidney disease
- Other cystic and tubulointerstitial nephropathy

[a]Relative contribution of each category varies with geographic region and race.

PATHOPHYSIOLOGY AND BIOCHEMISTRY OF UREMIA

Although serum urea and creatinine concentrations are used to measure the excretory capacity of the kidneys, accumulation of these two molecules themselves does not account for the many symptoms and signs that characterize the uremic syndrome in advanced renal failure. Large numbers of toxins that accumulate when GFR declines have been implicated in the uremic syndrome. These include water-soluble, hydrophobic, protein-bound, charged, and uncharged nitrogen-containing non-volatile products of metabolism. It is thus evident that the serum concentrations of urea and creatinine should be viewed as being readily measured, but very incomplete surrogate markers for retained toxins, and monitoring the levels of urea and creatinine in the patient with impaired kidney function represents a vast oversimplification of the uremic state.

The uremic syndrome involves more than renal excretory failure. A host of metabolic and endocrine functions normally performed by the kidneys is also impaired, and this results in anemia, malnutrition, and abnormal metabolism of carbohydrates, fats, and proteins. Furthermore, plasma levels of many hormones, including PTH, FGF-23, insulin, glucagon, steroid hormones including vitamin D and sex hormones, and prolactin change with CKD as a result of reduced excretion, decreased degradation, or abnormal regulation. Finally, CKD is associated with increased systemic inflammation. Elevated levels of C-reactive protein are detected along with other acute-phase reactants, whereas levels of so-called negative acute-phase reactants, such as albumin and fetuin, decline. Thus, the inflammation associated with CKD is important in the *malnutrition-inflammation-atherosclerosis/calcification syndrome*, which contributes in turn to the acceleration of vascular disease and comorbidity associated with advanced kidney disease.

In summary, the pathophysiology of the uremic syndrome can be divided into manifestations in three spheres of dysfunction: (1) those consequent to the accumulation of toxins that normally undergo renal excretion; (2) those consequent to the loss of other kidney functions, such as fluid and electrolyte homeostasis and hormone regulation; and (3) progressive systemic inflammation and its vascular and nutritional consequences.

CLINICAL AND LABORATORY MANIFESTATIONS OF CKD AND UREMIA

Uremia leads to disturbances in the function of virtually every organ system. Chronic dialysis can reduce the incidence and severity of many of these disturbances, so that the florid manifestations of uremia have largely disappeared in the modern health setting. However, even optimal dialysis therapy is not completely effective as renal replacement therapy, because some disturbances resulting from impaired kidney function fail to respond to dialysis.

FLUID, ELECTROLYTE, AND ACID-BASE DISORDERS

Sodium and Water Homeostasis

With normal renal function, tubular excretion of filtered sodium and water matches intake. Many forms of kidney disease (e.g., glomerulonephritis) disrupt this balance such that dietary intake of sodium exceeds its urinary excretion, leading to sodium retention and attendant extracellular fluid volume (ECFV) expansion. This expansion may contribute to hypertension, which itself can accelerate nephron injury. As long as water intake does not exceed the capacity for renal water clearance, the ECFV expansion will be isotonic and the patient will have a normal plasma sodium concentration. Hyponatremia is not commonly seen in CKD patients but, when present, often responds to water restriction. The patient with ECFV expansion (peripheral edema, sometimes hypertension poorly responsive to therapy) should be counseled regarding salt restriction. Thiazide diuretics have limited utility in stages 3–5 CKD, such that administration of loop diuretics, including furosemide, bumetanide, or torsemide, may also be needed. Resistance to loop diuretics in CKD often mandates use of higher doses than those used in patients with higher GFR. The combination of loop diuretics with metolazone may be helpful. Diuretic resistance with intractable edema and hypertension in advanced CKD may serve as an indication to initiate dialysis.

In addition to problems with salt and water excretion, some patients with CKD may instead have impaired renal conservation of sodium and water. When an extrarenal cause for fluid loss, such as gastrointestinal (GI) loss, is present, these patients may be prone to ECFV depletion because of the inability of the failing kidney to reclaim filtered sodium adequately. Furthermore, depletion of ECFV, whether due to GI losses or overzealous diuretic therapy, can further compromise kidney function through underperfusion, or a “prerenal” state, leading to acute-on-chronic kidney failure. In this setting, holding or adjusting the diuretic dose or even cautious volume repletion with normal saline may return the ECFV to normal and restore renal function to baseline.

Potassium Homeostasis

In CKD, the decline in GFR is not necessarily accompanied by a parallel decline in urinary potassium excretion, which is predominantly mediated by aldosterone-dependent secretion in the distal nephron. Another defense against potassium retention in these patients is augmented potassium excretion in the GI tract. Notwithstanding these two homeostatic responses, hyperkalemia may be precipitated in certain settings. These include increased dietary potassium intake, hemolysis, hemorrhage, transfusion of stored red blood cells, and metabolic acidosis. Importantly, a host of medications can inhibit renal potassium excretion and lead to hyperkalemia. The most important medications in this respect include the RAS inhibitors and spironolactone and other potassium-sparing diuretics such as amiloride, eplerenone, and triamterene. The benefits of the RAS inhibitors in ameliorating the

progression of CKD and its complications often favor their cautious and judicious use with very close monitoring of plasma potassium concentration.

Certain causes of CKD can be associated with earlier and more severe disruption of potassium-secretory mechanisms in the distal nephron, out of proportion to the decline in GFR. These include conditions associated with hyporeninemic hypoaldosteronism, such as diabetes, and renal diseases that preferentially affect the distal nephron, such as obstructive uropathy and sickle cell nephropathy.

Hypokalemia is not common in CKD and usually reflects markedly reduced dietary potassium intake, especially in association with excessive diuretic therapy or concurrent GI losses. The use of potassium supplements and potassium-sparing diuretics may be risky in patients with impaired renal function, and needs to be monitored closely.

Metabolic Acidosis

Metabolic acidosis is a common disturbance in advanced CKD. The majority of patients can still acidify the urine, but they produce less ammonia and, therefore cannot excrete the normal quantity of protons. Hyperkalemia, if present, further depresses ammonia production. The combination of hyperkalemia and hyperchloremic metabolic acidosis is often present, even at earlier stages of CKD (stages 1–3), in patients with diabetic nephropathy or in those with predominant tubulointerstitial disease or obstructive uropathy.

With worsening renal function, the total urinary net daily acid excretion is usually limited to 30–40 mmol, and the anions of retained organic acids can then lead to an anion-gap metabolic acidosis. Thus, the non-anion-gap metabolic acidosis seen in earlier stages of CKD may be complicated by the addition of an anion-gap metabolic acidosis as CKD progresses. In most patients, the metabolic acidosis is mild; the pH is rarely <7.32 and can usually be corrected with oral sodium bicarbonate supplementation. Animal and human studies have suggested that even modest degrees of metabolic acidosis may be associated with the development of protein catabolism. Alkali supplementation may, in addition, attenuate the catabolic state and possibly slow CKD progression and is recommended when the serum bicarbonate concentration falls below 20–23 mmol/L. The concomitant sodium load mandates careful attention to volume status and the need for diuretic agents.

TREATMENT

TREATMENT

Fluid, Electrolyte, and Acid-Base Disorders

Dietary salt restriction and the use of loop diuretics, occasionally in combination with metolazone, may be needed to maintain euvolemia. Water restriction is indicated only if there is a problem with hyponatremia.

Hyperkalemia often responds to dietary restriction of potassium, the use of kaliuretic diuretics, and avoidance of both potassium supplements (including occult sources, such as dietary salt substitutes) and dose reduction or avoidance of potassium-retaining medications (especially angiotensin-converting enzyme [ACE] inhibitors or angiotensin receptor blockers [ARBs]). Kaliuretic diuretics promote urinary potassium excretion, whereas potassium-binding resins, such as calcium resonium, sodium polystyrene or patiromer can promote potassium loss through the GI tract and may reduce the incidence of hyperkalemia. Intractable hyperkalemia is an indication

(although uncommon) to consider institution of dialysis in a CKD patient. The renal tubular acidosis and subsequent anion-gap metabolic acidosis in progressive CKD will respond to alkali supplementation, typically with sodium bicarbonate. Recent studies suggest that this replacement should be considered when the serum bicarbonate concentration falls below 20–23 mmol/L to avoid the protein catabolic state seen with even mild degrees of metabolic acidosis and to slow the progression of CKD.

DISORDERS OF CALCIUM AND PHOSPHATE METABOLISM

The principal complications of abnormalities of calcium and phosphate metabolism in CKD occur in the skeleton and the vascular bed, with occasional severe involvement of soft tissues. It is likely that disorders of bone turnover and disorders of vascular and soft tissue calcification are related to each other (Fig. 305-3).

Bone Manifestations of CKD

The major disorders of bone disease can be classified into those associated with high bone turnover with increased PTH levels (including *osteitis fibrosa cystica*, the classic lesion of secondary hyperparathyroidism), osteomalacia due to reduced action of the active forms of vitamin D, and low bone turnover with low or normal PTH levels (adynamic bone disease) or most often combinations of the foregoing.

The pathophysiology of secondary hyperparathyroidism and the consequent high-turnover bone disease is related to abnormal mineral metabolism through the following events: (1) declining GFR leads to reduced excretion of phosphate and, thus, phosphate retention; (2) the retained phosphate stimulates increased synthesis of both FGF-23 by osteocytes and PTH and stimulates growth of parathyroid gland mass; and (3) decreased levels of ionized calcium, resulting from suppression of calcitriol production by FGF-23 and by the failing kidney, as well as phosphate retention, also stimulate PTH production. Low calcitriol levels contribute to hyperparathyroidism, both by leading to hypocalcemia and also by a direct effect on PTH gene transcription. These changes start to occur when the GFR falls below 60 mL/min.

FGF-23 is part of a family of phosphatonins that promotes renal phosphate excretion. Recent studies have shown that levels of this hormone, secreted by osteocytes, increase early in the course of CKD, even before phosphate retention and hyperphosphatemia. FGF-23 may defend normal serum phosphorus in at least three ways: (1) increased renal phosphate excretion; (2) stimulation of PTH, which also increases renal phosphate excretion; and (3) suppression of the formation of $1,25(OH)_2D_3$, leading to diminished phosphorus absorption from the GI tract.

Interestingly, high levels of FGF-23 are also an independent risk factor for left ventricular hypertrophy and mortality in CKD, dialysis, and kidney transplant patients. Moreover, elevated levels of FGF-23 may indicate the need for therapeutic intervention (e.g., phosphate restriction), even when serum phosphate levels are within the normal range.

Hyperparathyroidism stimulates bone turnover and leads to *osteitis fibrosa cystica*. Bone histology shows abnormal osteoid, bone and bone marrow fibrosis, and in advanced stages, the formation of bone cysts, sometimes with hemorrhagic elements so that they appear brown in color, hence the term *brown tumor*. Clinical manifestations of severe hyperparathyroidism include bone pain and fragility, brown tumors, compression syndromes, and erythropoietin (EPO) resistance in part related to the bone marrow fibrosis. Furthermore, PTH itself is considered a

uremic toxin, and high levels are associated with muscle weakness, fibrosis of cardiac muscle, and nonspecific constitutional symptoms.

Adynamic bone disease is increasing in prevalence, especially among diabetics and the elderly. It is characterized by reduced bone volume and mineralization and may result from excessive suppression of PTH production, chronic inflammation, or both. Suppression of PTH can result from the use of vitamin D preparations or from excessive calcium exposure in the form of calcium-containing phosphate binders or high-calcium dialysis solutions. Complications of adynamic bone disease include an increased incidence of fracture and bone pain and an association with increased vascular and cardiac calcification. Occasionally the calcium will precipitate in the soft tissues into large concretions termed “tumoral calcinosis” (**Fig. 305-4**). Patients with adynamic bone disease often experience the most severe symptoms of musculoskeletal pain, owing to the inability to repair the microfractures that occur properly as a part of healthy skeletal homeostasis with regular physical activity. Osteomalacia is a distinct process, consequent to reduced production and action of $1,25(OH)_2D_3$, leading to non-mineralized osteoid.

FIGURE 305-4

Tumoral calcinosis. This patient was on hemodialysis for many years and was nonadherent to dietary phosphorus restriction or the use of phosphate binders. He was chronically severely hyperphosphatemic. He developed an enlarging painful mass on his arm that was extensively calcified.

Tumoral Calcinosis in a Dialysis Patient



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Calcium, Phosphorus, and the Cardiovascular System

Recent epidemiologic evidence has shown a strong association between hyperphosphatemia and increased cardiovascular mortality in patients with stage 5 and earlier stages of CKD. Hyperphosphatemia and hypercalcemia are associated with increased vascular calcification, but it is unclear whether the excessive mortality is mediated by this mechanism. Studies using computed tomography (CT) and electron-beam CT scanning show that CKD patients have calcification of the media in coronary arteries and even heart valves that appear to be orders of magnitude

greater than that in patients without renal disease. The magnitude of the calcification is proportional to age and hyperphosphatemia and is also associated with low PTH levels and low bone turnover. It is possible that in CKD patients ingested calcium cannot be incorporated into bones with low turnover and, therefore, is deposited at extraosseous sites, such as the vascular bed and soft tissues. It is interesting in this regard that there is also an association between osteoporosis and vascular calcification in the general population. Finally, hyperphosphatemia can induce a change in gene expression in vascular cells to an osteoblast-like profile, leading to vascular calcification and even ossification.

Other Complications of Abnormal Mineral Metabolism

Calciphylaxis is a devastating condition seen almost exclusively in patients with advanced CKD. It is heralded by livedo reticularis and advances to patches of ischemic necrosis, especially on the legs, thighs, abdomen, and breasts **(Fig. 305-5)**. Pathologically, there is evidence of vascular occlusion in association with extensive vascular and soft tissue calcification. It appears that this condition is increasing in incidence. Originally it was ascribed to severe abnormalities in calcium and phosphorus control in dialysis patients, usually associated with advanced hyperparathyroidism. However, more recently, calciphylaxis has been seen with increasing frequency in the absence of severe hyperparathyroidism. Other etiologies have been suggested, including the increased use of oral calcium as a phosphate binder. Warfarin is commonly used in hemodialysis patients in whom most direct oral anticoagulants (DOACs) are contraindicated, and one of the effects of warfarin therapy is to decrease the vitamin K–dependent regeneration of matrix GLA protein. This latter protein is important in preventing vascular calcification. Thus, warfarin treatment is considered a risk factor for calciphylaxis, and if a patient develops this syndrome, this medication should be discontinued and replaced with another anticoagulant.

FIGURE 305-5

Calciphylaxis. This peritoneal dialysis patient was on chronic warfarin therapy for atrial fibrillation. She noticed a small painful nodule on the abdomen that was followed by progressive skin necrosis and ulceration of the anterior abdominal wall. She was treated with hyperbaric oxygen, intravenous thiosulfate, and discontinuation of warfarin, with slow resolution of the ulceration.

Calciphylaxis



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Disorders of Calcium and Phosphate Metabolism

The optimal management of secondary hyperparathyroidism and *osteitis fibrosa* is prevention. Once the parathyroid gland mass is very large, it is difficult to control the disease. Careful attention should be paid to the plasma phosphate concentration in CKD patients, who should be counseled on a low-phosphate diet as well as the appropriate use of phosphate-binding agents. These are agents that are taken with meals and complex the dietary phosphate to limit its GI absorption. Examples of phosphate binders are calcium acetate and calcium carbonate. A major side effect of calcium-based phosphate binders is calcium accumulation and hypercalcemia, especially in patients with low-turnover bone disease. Sevelamer and lanthanum are non-calcium-containing polymers that also function as phosphate binders; they do not predispose CKD patients to hypercalcemia and may attenuate calcium deposition in the vascular bed.

Calcitriol exerts a direct suppressive effect on PTH secretion and also indirectly suppresses PTH secretion by raising the concentration of ionized calcium. However, calcitriol therapy may result in hypercalcemia and/or hyperphosphatemia through increased GI absorption of these minerals. Certain analogues of calcitriol are available (e.g., paricalcitol) that suppress PTH secretion with less attendant hypercalcemia.

Recognition of the role of the extracellular calcium-sensing receptor has led to the development of calcimimetic agents that enhance the sensitivity of the parathyroid cell to the suppressive effect of calcium. This class of drug,

which includes cinacalcet, produces a dose-dependent reduction in PTH and plasma calcium concentration in some patients.

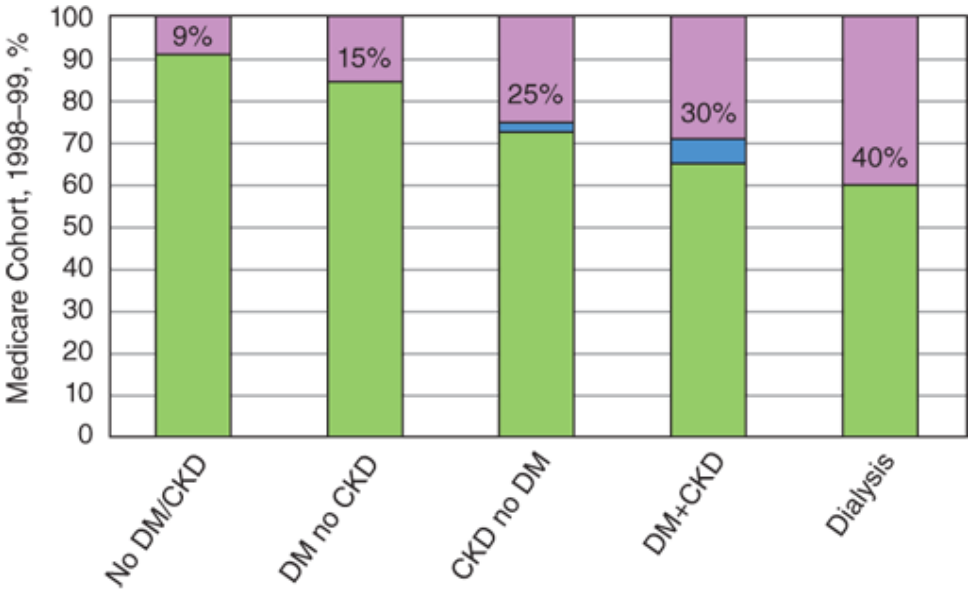
Current National Kidney Foundation Kidney Disease Outcomes Quality Initiative guidelines recommend a target PTH level between 150 and 300 pg/mL, recognizing that very low PTH levels are associated with adynamic bone disease and possible consequences of fracture and ectopic calcification.

CARDIOVASCULAR ABNORMALITIES

Cardiovascular disease is the leading cause of morbidity and mortality in patients at every stage of CKD. The incremental risk of cardiovascular disease in those with CKD compared to the age- and sex-matched general population ranges from 10- to 200-fold, depending on the stage of CKD. As a result, most patients with CKD succumb to cardiovascular disease **(Fig. 305-6)** before ever reaching stage 5 CKD. Between 30 and 45% of those patients who do reach stage 5 CKD have advanced cardiovascular complications. Thus, the focus of patient care in earlier CKD stages should be directed to prevention of cardiovascular complications.

FIGURE 305-6

U.S. Renal Data System showing increased likelihood of dying rather than starting dialysis or reaching stage 5 chronic kidney disease (CKD). 1, Death; 2, ESRD; 3, event-free. DM; diabetes mellitus. *(Data from RN Foley et al: J Am Soc Nephrol 16:489–495, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ischemic Vascular Disease

The increased prevalence of vascular disease in CKD patients derives from both traditional (“classic”) and nontraditional (CKD-related) risk factors. Traditional risk factors include hypertension, hypervolemia, dyslipidemia, sympathetic overactivity, and hyperhomocysteinemia. The CKD-related risk factors comprise anemia, hyperphosphatemia, hyperparathyroidism, increased FGF-23, sleep apnea, and generalized inflammation. The inflammatory state appears to accelerate vascular occlusive disease, and low levels of fetuin may permit more rapid vascular calcification, especially in the face of hyperphosphatemia. Other abnormalities seen in CKD may augment

myocardial ischemia, including left ventricular hypertrophy and microvascular disease. In addition, hemodialysis, with its attendant episodes of hypotension and hypovolemia, may further aggravate coronary ischemia and repeatedly stun the myocardium. Interestingly, however, the largest increment in cardiovascular mortality rate in dialysis patients is not necessarily directly associated with documented acute myocardial infarction but, instead, is the result of congestive heart failure and sudden death.

Cardiac troponin levels are frequently elevated in CKD without evidence of acute ischemia. The elevation complicates the diagnosis of acute myocardial infarction in this population. Serial measurements may be needed. Therefore, the trend in levels over the hours after presentation may be more informative than a single, elevated level. Interestingly, consistently elevated levels are an independent prognostic factor for adverse cardiovascular events in this population.

Heart Failure

Abnormal cardiac function secondary to myocardial ischemia, left ventricular hypertrophy, diastolic dysfunction, and frank cardiomyopathy, in combination with the salt and water retention often results in heart failure or even pulmonary edema. Heart failure can be a consequence of diastolic or systolic dysfunction, or both. A form of “low-pressure” pulmonary edema can also occur in advanced CKD, manifesting as shortness of breath and a “bat wing” distribution of alveolar edema fluid on the chest x-ray. This finding can occur even in the absence of ECFV overload and is associated with normal or mildly elevated pulmonary capillary wedge pressure. This process has been ascribed to increased permeability of alveolar capillary membranes as a manifestation of the uremic state, and it responds to dialysis. Other CKD-related risk factors, including anemia and sleep apnea, may contribute to the risk of heart failure.

Hypertension and Left Ventricular Hypertrophy

Hypertension is one of the most common complications of CKD. It usually develops early during the course of CKD and is associated with adverse outcomes, including the development of ventricular hypertrophy and a more rapid loss of renal function. Left ventricular hypertrophy and dilated cardiomyopathy are among the strongest risk factors for cardiovascular morbidity and mortality in patients with CKD and are thought to be related primarily, but not exclusively, to prolonged hypertension and ECFV overload. In addition, anemia and the placement of an arteriovenous fistula for hemodialysis can generate a high cardiac output state and consequent heart failure.

The absence of hypertension may signify poor left ventricular function. Indeed, in epidemiologic studies of dialysis patients, low blood pressure actually carries a worse prognosis than does high blood pressure. This mechanism, in part, accounts for the “reverse causation” seen in dialysis patients, wherein the presence of traditional risk factors, such as hypertension, hyperlipidemia, and obesity, appear to portend a better prognosis. Importantly, these observations derive from cross-sectional studies of late-stage CKD patients and should not be interpreted to discourage appropriate management of these risk factors in CKD patients, especially at early stages. In contrast to the general population, it is possible that in late-stage CKD, low blood pressure, reduced body mass index, and hypolipidemia indicate the presence of an advanced malnutrition-inflammation state, with poor prognosis.

The use of exogenous erythropoiesis-stimulating agents can increase blood pressure and the requirement for antihypertensive drugs. Chronic ECFV overload is also a contributor to hypertension, and improvement in blood pressure can often be seen with the use of dietary sodium restriction, diuretics, and fluid removal with dialysis.

Nevertheless, because of activation of the RAS and other disturbances in the balance of vasoconstrictors and vasodilators, some patients remain hypertensive despite careful attention to ECFV status.

TREATMENT

TREATMENT

Cardiovascular Abnormalities

MANAGEMENT OF HYPERTENSION

The overarching goal of hypertension therapy in CKD is to prevent the extrarenal complications of high blood pressure, such as cardiovascular disease and stroke. Although a clear-cut generalizable benefit in slowing progression of CKD remains as yet unproven, the benefit for cardiac and cerebrovascular health is compelling. In all patients with CKD, blood pressure should be controlled to levels recommended by national guideline panels. In CKD patients with diabetes or proteinuria >1 g per 24 h, blood pressure should be reduced to <130/80 mmHg, if achievable without prohibitive adverse effects. Salt restriction should be the first line of therapy. When volume management alone is not sufficient, the choice of antihypertensive agent is similar to that in the general population. ACE inhibitors and ARBs appear to slow the rate of decline of kidney function in a manner that extends beyond reduction of systemic arterial pressure and that involves correction of the intraglomerular hyperfiltration and hypertension. Occasionally, introduction of ACE inhibitors and ARBs can actually precipitate an episode of acute kidney injury, especially when used in combination in patients with ischemic renovascular disease. Slight reduction of GFR (<30% of baseline) may signify a salutary reduction in intra-glomerular hypertension and hyperfiltration, and, if stable over time, can be tolerated with continued monitoring. Progressive decline in GFR should prompt discontinuation of these agents. The use of ACE inhibitors and ARBs may also be complicated by the development of hyperkalemia. Often the concomitant use of a combination of kaliuretic diuretics (e.g., furosemide with metolazone), or a potassium-lowering GI tract binder, such as patrimer, can improve potassium excretion in addition to improving blood pressure control. Potassium-sparing diuretics should be used with caution or avoided altogether in most patients.

The recent movement to even lower blood pressure targets in the general population may not be applicable to patients with CKD, who often lack autoregulation to maintain GFR in the face of low perfusion pressure. If a patient experiences sudden decline in kidney function with intensification of antihypertensive therapy, consideration should be given to reducing therapy.

MANAGEMENT OF CARDIOVASCULAR DISEASE

There are many strategies available to treat the traditional and nontraditional risk factors in CKD patients. Although these have proved effective in the general population, there is little evidence for their benefit in patients with advanced CKD, especially those on dialysis. Certainly hypertension, and dyslipidemia promote atherosclerotic disease and are treatable complications of CKD. Renal disease complicated by nephrotic syndrome is associated with a very atherogenic lipid profile and hypercoagulability, which increases the risk of occlusive vascular disease. Because diabetes mellitus and hypertension are the two most frequent causes of advanced CKD, it is not surprising that cardiovascular disease is the most frequent cause of death in dialysis patients. The role of “inflammation” may be quantitatively more important in patients with kidney disease, and the treatment of more traditional risk factors may result in only modest success. However, modulation of traditional risk factors may be the only weapon in the

therapeutic armamentarium for these patients until the nature of inflammation in CKD and its treatment are better understood.

Pericardial Disease

Chest pain with respiratory accentuation, accompanied by a friction rub, is diagnostic of pericarditis. Classic electrocardiographic abnormalities include PR-interval depression and diffuse ST-segment elevation. Pericarditis can be accompanied by pericardial effusion that is seen on echocardiography and can rarely lead to tamponade. However, the pericardial effusion can be asymptomatic, and pericarditis can be seen without significant effusion.

Pericarditis is observed in advanced uremia, and with the advent of timely initiation of dialysis, is not as common as it once was. It is now more often observed in underdialyzed, non-adherent patients than in those starting dialysis.

TREATMENT

TREATMENT

Pericardial Disease

Uremic pericarditis is an absolute indication for the urgent initiation of dialysis or for intensification of the dialysis prescription in those already receiving dialysis. Because of the propensity to hemorrhage in pericardial fluid, hemodialysis should be performed without heparin. A pericardial drainage procedure should be considered in patients with recurrent pericardial effusion, especially with echocardiographic signs of impending tamponade. Non-uremic causes of pericarditis and effusion include viral, malignant, tuberculous, and autoimmune etiologies. It may also be seen after myocardial infarction and as a complication of treatment with the antihypertensive drug minoxidil.

HEMATOLOGIC ABNORMALITIES

Anemia

A normocytic, normochromic anemia is observed as early as stage 3 CKD and is almost universal by stage 4. The primary cause is insufficient production of EPO by the diseased kidneys. Additional factors are reviewed in **Table 305-3**.

TABLE 305-3

Causes of Anemia in CKD

Relative deficiency of erythropoietin
Diminished red blood cell survival
Bleeding diathesis
Iron deficiency due to poor dietary absorption and gastrointestinal blood loss
Hyperparathyroidism/bone marrow fibrosis
Chronic inflammation
Folate or vitamin B_{12} deficiency
Hemoglobinopathy
Comorbid conditions: hypo-/hyperthyroidism, pregnancy, HIV-associated disease, autoimmune disease, immunosuppressive drugs

The anemia of CKD is associated with a number of adverse pathophysiologic consequences, including decreased tissue oxygen delivery and utilization, increased cardiac output, ventricular dilation, and ventricular hypertrophy. Clinical manifestations include fatigue and diminished exercise tolerance, angina, heart failure, decreased cognition and mental acuity, and impaired host defense against infection. In addition, anemia may play a role in growth restriction in children with CKD. Although many studies in CKD patients have found that anemia and resistance to exogenous erythropoietic-stimulating agents (ESA) are associated with a poor prognosis, the relative contribution to a poor outcome of the low hematocrit itself, versus inflammation as a cause of the anemia and ESA resistance, remains unclear.

TREATMENT

TREATMENT

Anemia

The availability of recombinant human ESA has been one of the most significant advances in the care of renal patients since the introduction of dialysis and renal transplantation. Its routine use has obviated the need for regular blood transfusions in severely anemic CKD patients, thus dramatically reducing the incidence of transfusion-associated infections and iron overload. Frequent blood transfusions in dialysis patients also lead to the development of alloantibodies that can sensitize the patient to donor kidney antigens and make renal transplantation more problematic.

Adequate bone marrow iron stores should be available before treatment with ESA is initiated. Iron supplementation is usually essential to ensure an optimal response to ESA in patients with CKD because the demand for iron by the marrow frequently exceeds the amount of iron that is immediately available for erythropoiesis (measured by percent transferrin saturation), as well as the amount in iron stores (measured by serum ferritin). For the CKD patient not yet on dialysis or the patient treated with peritoneal dialysis, oral iron supplementation should be attempted. If there is GI intolerance or poor GI absorption, the patient may have to undergo IV iron infusion. For patients on hemodialysis, IV iron can be administered during dialysis, keeping in mind that iron therapy can

increase the susceptibility to bacterial infections, and that the adverse effects of free serum iron are still under investigation. In addition to iron, an adequate supply of other major substrates and cofactors for red cell production must be ensured, including vitamin B_{12} and folate. Anemia resistant to recommended doses of ESA in the face of adequate iron stores may be due to some combination of the following: acute or chronic inflammation, inadequate dialysis, severe hyperparathyroidism, chronic blood loss or hemolysis, chronic infection, or malignancy.

Randomized, controlled trials of ESA in CKD have failed to show an improvement in cardiovascular outcomes with this therapy. Indeed, there has been an indication that the use of ESA in CKD may be associated with an increased risk of stroke in those with type 2 diabetes, an increase in thromboembolic events, and perhaps a faster progression of renal decline. Therefore, any benefit in terms of improvement of anemic symptoms needs to be balanced against the potential cardiovascular risk. Although further studies are needed, it is quite clear that complete normalization of the hemoglobin concentration has not been demonstrated to be of incremental benefit to CKD patients. Current practice is to target a hemoglobin concentration of 100–115 g/L.

Abnormal Hemostasis

Patients with later stages of CKD may have a prolonged bleeding time, decreased activity of platelet factor III, abnormal platelet aggregation and adhesiveness, and impaired prothrombin consumption. Clinical manifestations include an increased tendency to bleeding and bruising, prolonged bleeding from surgical incisions, menorrhagia, and GI bleeding. Interestingly, CKD patients also have a greater susceptibility to thromboembolism, especially if they have renal disease that includes nephrotic-range proteinuria. The latter condition results in hypoalbuminemia and renal loss of anticoagulant factors, which can lead to a thrombophilic state.

TREATMENT

TREATMENT

Abnormal Hemostasis

Abnormal bleeding time and coagulopathy in patients with renal failure may be reversed temporarily with desmopressin (DDAVP), cryoprecipitate, IV conjugated estrogens, blood transfusions, and ESA therapy. Optimal dialysis will usually correct a prolonged bleeding time.

Given the coexistence of bleeding disorders and a propensity to thrombosis that is unique in the CKD patient, decisions about anticoagulation that have a favorable risk-benefit profile in the general population may not be applicable to the patient with advanced CKD. One example is warfarin anticoagulation for atrial fibrillation; the decision to anticoagulate should be made on an individual basis in the CKD patient because there appears to be a greater risk of bleeding complications.

Certain anticoagulants, such as fractionated low-molecular-weight heparin, may need to be avoided or dose-adjusted in these patients, with monitoring of factor Xa activity where available. It is often more prudent to use conventional unfractionated heparin, titrated to the measured partial thromboplastin time, in hospitalized patients requiring an alternative to warfarin anticoagulation. The new classes of oral anticoagulants are all, in part, renally eliminated and need to be avoided or dose adjusted in the face of decreased GFR (**Chap. 114**).

NEUROMUSCULAR ABNORMALITIES

Central nervous system (CNS), peripheral, and autonomic neuropathy as well as abnormalities in muscle structure and function are all well-recognized complications of CKD. Subtle clinical manifestations of uremic neuromuscular disease usually become evident at stage 3 CKD. Early manifestations of CNS complications include mild disturbances in memory and concentration and sleep disturbance. Neuromuscular irritability, including hiccups, cramps, and twitching, becomes evident at later stages. In advanced untreated kidney failure, asterixis, myoclonus, seizures, and coma can be seen.

Peripheral neuropathy usually becomes clinically evident after the patient reaches stage 4 CKD, although electrophysiologic and histologic evidence occurs earlier. Initially, sensory nerves are involved more than motor, lower extremities more than upper, and distal parts of the extremities more than proximal. The “restless leg syndrome” is characterized by ill-defined sensations of sometimes debilitating discomfort in the legs and feet relieved by frequent leg movement. Evidence of peripheral neuropathy without another cause (e.g., diabetes mellitus) is an indication for starting renal replacement therapy. Many of the complications described above will resolve with dialysis, although subtle nonspecific abnormalities may persist.

GASTROINTESTINAL AND NUTRITIONAL ABNORMALITIES

Uremic fetor, a urine-like odor on the breath, derives from the breakdown of urea to ammonia in saliva and is often associated with an unpleasant metallic taste (dysgeusia). Gastritis, peptic disease, and mucosal ulcerations at any level of the GI tract occur in uremic patients and can lead to abdominal pain, nausea, vomiting, and GI bleeding. These patients are also prone to constipation, which can be worsened by the administration of calcium and iron supplements. The retention of uremic toxins also leads to anorexia, nausea, and vomiting.

Protein restriction may be useful to decrease nausea and vomiting; however, it may put the patient at risk for malnutrition and should be carried out, if possible, in consultation with a registered dietitian specializing in the management of CKD patients. Weight loss and protein-energy malnutrition, a consequence of low protein and caloric intake, is common in advanced CKD and is often an indication for initiation of renal replacement therapy. Metabolic acidosis and the activation of inflammatory cytokines can promote protein catabolism. A number of indices are useful in nutritional assessment and include dietary history, including food diary and subjective global assessment; edema-free body weight; and measurement of urinary protein nitrogen appearance. Dual-energy x-ray absorptiometry is now widely used to estimate lean body mass versus fluid weight. Nutritional guidelines for patients with CKD are summarized in the “Treatment” section.

ENDOCRINE-METABOLIC DISTURBANCES

Glucose metabolism is impaired in CKD. However, fasting blood glucose is usually normal or only slightly elevated, and mild glucose intolerance does not require specific therapy. Because the kidney contributes to insulin removal from the circulation, plasma levels of insulin are slightly to moderately elevated in most uremic patients, both in the fasting and postprandial states. Because of this diminished renal degradation of insulin, patients on insulin therapy may need progressive reduction in dose as their renal function worsens. Many anti-hyperglycemic agents, including the gliptins, require dose reduction in renal failure, and some, such as metformin and sulfonylureas are contraindicated when the GFR is less than half of normal. A recent exception is the class of drugs that inhibit sodium-glucose transport in the proximal tubule, resulting in glucose lowering, accompanied by striking reductions in kidney function decline and in cardiovascular events. The stabilization of GFR in many patients with this

therapeutic intervention represents a major, important added beneficial effect of these drugs. Their long-term stabilizing effect on GFR and urine albumin excretion appears to result from correction of hyperfiltration early in type 2 diabetes mellitus *via* re-activation of the tubuloglomerular feedback loop. This represents a fortunate convergence of pathophysiology of glomerular hyperfiltration in diabetes with drug discovery.

In women with CKD, estrogen levels are low, and menstrual abnormalities, infertility, and inability to carry pregnancies to term are common. When the GFR has declined to ~40 mL/min, pregnancy is associated with a high rate of spontaneous abortion, with only ~20% of pregnancies leading to live births, and pregnancy may hasten the progression of the kidney disease itself. Women with CKD who are contemplating pregnancy should consult first with a nephrologist in conjunction with an obstetrician specializing in high-risk pregnancy. Men with CKD have reduced plasma testosterone levels, and sexual dysfunction and oligospermia may supervene. Sexual maturation may be delayed or impaired in adolescent children with CKD, even among those treated with dialysis. Many of these abnormalities improve or reverse with intensive dialysis or with successful renal transplantation.

DERMATOLOGIC ABNORMALITIES

Abnormalities of the skin are prevalent in progressive CKD. Pruritus is quite common and one of the most vexing manifestations of the uremic state. In advanced CKD, even on dialysis, patients may become more pigmented, and this is felt to reflect the deposition of retained pigmented metabolites, or *urochromes*. Although many of the cutaneous abnormalities improve with dialysis, pruritus is often tenacious. The first lines of management are to rule out unrelated skin disorders, such as scabies, and to treat hyperphosphatemia, which can cause itch. Local moisturizers, mild topical glucocorticoids, oral antihistamines, and ultraviolet radiation have been reported to be helpful.

A skin condition unique to CKD patients called *nephrogenic fibrosing dermopathy* consists of progressive subcutaneous induration, especially on the arms and legs. The condition is seen very rarely in patients with CKD who have been exposed to the magnetic resonance contrast agent gadolinium. Current recommendations are that patients with CKD stage 3 (GFR 30–59 mL/min) should minimize exposure to gadolinium, and those with CKD stages 4–5 (GFR <30 mL/min) should avoid the use of gadolinium agents unless it is medically necessary. However, no patient should be denied an imaging investigation that is critical to management, and under such circumstances, rapid removal of gadolinium by hemodialysis (even in patient's not yet receiving renal replacement therapy) shortly after the procedure may mitigate this sometimes devastating complication.

EVALUATION AND MANAGEMENT OF PATIENTS WITH CKD

INITIAL APPROACH

History and Physical Examination

Symptoms and overt signs of kidney disease are often subtle or absent until renal failure supervenes. Thus, the diagnosis of kidney disease often surprises patients and may be a cause of skepticism and denial. Particular aspects of the history that are germane to renal disease include a history of hypertension (which can cause CKD or more commonly be a consequence of CKD), diabetes mellitus, abnormal urinalyses, and problems with pregnancy such as preeclampsia or early pregnancy loss. A careful drug history should be elicited. Drugs to consider include

nonsteroidal anti-inflammatory agents, cyclooxygenase-2 (COX-2) inhibitors, antimicrobials, chemotherapeutic agents, antiretroviral agents, proton pump inhibitors, phosphate-containing bowel cathartics, and lithium. In evaluating the uremic syndrome, questions about appetite, weight loss, nausea, hiccups, peripheral edema, muscle cramps, pruritus, and restless legs are especially helpful. A family history of kidney disease, together with assessment of manifestations in other organ systems such as auditory, visual, and integumentary, may lead to the diagnosis of a heritable form of CKD (e.g., Alport or Fabry disease, cystinosis) or shared environmental exposure to nephrotoxic agents (e.g., heavy metals, aristolochic acid). It should be noted that clustering of CKD, sometimes of different etiologies, is often observed within families.

The physical examination should focus on blood pressure and target organ damage from hypertension. Thus, funduscopy and precordial examination should be carried out. Funduscopy is important in the diabetic patient, because it may show evidence of diabetic retinopathy, which is associated with nephropathy. Other physical examination manifestations of CKD include edema and sensory polyneuropathy. The finding of asterixis or a pericardial friction rub not attributable to other causes usually signifies the presence of the uremic syndrome.

Laboratory Investigation

Laboratory studies should focus on a search for clues to an underlying causative or aggravating disease process and on the degree of renal damage and its consequences. Serum and urine protein electrophoresis, looking for multiple myeloma, should be obtained in all patients >35 years with unexplained CKD, especially if there is associated anemia and elevated, or even inappropriately normal, serum calcium concentration in the face of renal insufficiency. In the presence of glomerulonephritis, autoimmune diseases such as lupus and underlying infectious etiologies such as hepatitis B and C and HIV should be tested. Serial measurements of renal function should be obtained to determine the pace of renal deterioration and ensure that the disease is truly chronic rather than acute or subacute and hence potentially reversible. Serum concentrations of calcium, phosphorus, vitamin D, and PTH should be measured to evaluate metabolic bone disease. Hemoglobin concentration, iron, vitamin B_{12}, and folate should also be evaluated. A 24-h urine collection may be helpful, because protein excretion >300 mg may be an indication for therapy with ACE inhibitors or ARBs.

Imaging Studies

The most useful imaging study is a renal ultrasound, which can verify the presence of two kidneys, determine if they are symmetric, provide an estimate of kidney size, and rule out renal masses and evidence of obstruction. Because it takes time for kidneys to shrink as a result of chronic disease, the finding of bilaterally small kidneys supports the diagnosis of CKD of long-standing duration. If the kidney size is normal, it is possible that the renal disease is acute or subacute. The exceptions are diabetic nephropathy (where kidney size is increased at the onset of diabetic nephropathy before CKD supervenes), amyloidosis, and HIV nephropathy, where kidney size may be normal in the face of CKD. Polycystic kidney disease that has reached some degree of renal failure will almost always present with enlarged kidneys with multiple cysts (Chap. 309). A discrepancy >1 cm in kidney length suggests either a unilateral developmental abnormality or disease process or renovascular disease with arterial insufficiency affecting one kidney more than the other. The diagnosis of renovascular disease can be undertaken with different techniques, including Doppler sonography, nuclear medicine studies, or CT or magnetic resonance imaging (MRI) studies. If there is a suspicion of reflux nephropathy (recurrent childhood urinary tract infection, asymmetric renal size with scars on the renal poles), a voiding cystogram may be indicated. However, in most cases, by the time the patient has

CKD, the reflux has resolved, and even if still present, repair does not improve renal function. Radiographic contrast imaging studies are not particularly helpful in the investigation of CKD. Intravenous or intraarterial dye should be avoided where possible in the CKD patient, especially with diabetic nephropathy, because of the risk of radiographic contrast dye–induced renal failure. When unavoidable, appropriate precautionary measures include avoidance of hypovolemia at the time of contrast exposure, minimization of the dye load, and choice of radiographic contrast preparations with the least nephrotoxic potential. Additional measures thought to attenuate contrast-induced worsening of renal function include judicious administration of sodium bicarbonate–containing solutions and *N* acetylcysteine.

Kidney Biopsy

In the patient with bilaterally small kidneys, renal biopsy is not advised because (1) it is technically difficult and has a greater likelihood of causing bleeding and other adverse consequences, (2) there is usually so much scarring that the underlying disease may not be apparent, and (3) the window of opportunity to render disease-specific therapy has passed. Other contraindications to renal biopsy include uncontrolled hypertension, active urinary tract infection, bleeding diathesis (including ongoing anticoagulation), and severe obesity. Ultrasound-guided percutaneous biopsy is the favored approach, but a surgical or laparoscopic approach can be considered, especially in the patient with a single kidney where direct visualization and control of bleeding are crucial. In the CKD patient in whom a kidney biopsy is indicated (e.g., suspicion of a concomitant or superimposed active process such as interstitial nephritis or in the face of accelerated loss of GFR), the bleeding time should be measured, and if increased, desmopressin should be administered immediately prior to the procedure.

A brief run of hemodialysis (without heparin) may also be considered prior to renal biopsy to normalize the bleeding time.

ESTABLISHING THE DIAGNOSIS AND ETIOLOGY OF CKD

The most important initial diagnostic step is to distinguish newly diagnosed CKD from acute or subacute renal failure, because the latter two conditions may respond to targeted therapy. Previous measurements of serum creatinine concentration are particularly helpful in this regard. Normal values from recent months or even years suggest that the current extent of renal dysfunction could be more acute, and hence reversible, than might otherwise be appreciated. In contrast, elevated serum creatinine concentration in the past suggests that the renal disease represents a chronic process. Even if there is evidence of chronicity, there is the possibility of a superimposed acute process (e.g., ECFV depletion, urinary infection or obstruction, or nephrotoxin exposure) supervening on the chronic condition. If the history suggests multiple systemic manifestations of recent onset (e.g., fever, polyarthritis, rash), it should be assumed that renal insufficiency is part of an acute systemic illness.

Although kidney biopsy can usually be performed in early CKD (stages 1–3), it is not always indicated. For example, in a patient with a history of type 1 diabetes mellitus for 15–20 years with retinopathy, nephrotic-range proteinuria, and absence of hematuria, the diagnosis of diabetic nephropathy is very likely and biopsy is usually not necessary. However, if there were some other finding not typical of diabetic nephropathy, such as hematuria or white blood cell casts, or absence of diabetic retinopathy, some other disease may be present and a biopsy may be indicated.

In the absence of a clinical diagnosis, kidney biopsy may be the only recourse to establish an etiology in early-stage CKD. However, as noted above, once the CKD is advanced and the kidneys are small and scarred, there is little utility

and significant risk in attempting to arrive at a specific diagnosis. Genetic testing is increasingly entering the repertoire of diagnostic tests, since the patterns of injury and kidney morphologic abnormalities often reflect overlapping causal mechanisms, whose origins can sometimes be attributed to a genetic predisposition or cause.

TREATMENT

TREATMENT

Chronic Kidney Disease

Treatments aimed at specific causes of CKD are discussed elsewhere. The optimal timing of both specific and nonspecific therapy is usually well before there has been a measurable decline in GFR and certainly before CKD is established. It is helpful to measure sequentially and plot the rate of decline of GFR in all patients. Any acceleration in the rate of decline should prompt a search for superimposed acute or subacute processes that may be reversible. These include ECFV depletion, uncontrolled hypertension, urinary tract infection, new obstructive uropathy, exposure to nephrotoxic agents (such as nonsteroidal anti-inflammatory drugs [NSAIDs] or radiographic dye), and reactivation or flare of the original disease, such as lupus or vasculitis.

SLOWING THE PROGRESSION OF CKD

There is variation in the rate of decline of GFR among patients with CKD. However, the following interventions should be considered in an effort to stabilize or slow the decline of renal function.

Reducing Intraglomerular Hypertension and Proteinuria

Increased intraglomerular filtration pressures and glomerular hypertrophy develop as a response to loss of nephron number. This response is maladaptive, as it promotes the ongoing decline of kidney function even if the inciting process has been treated or spontaneously resolved. Control of glomerular hypertension is important in slowing the progression of CKD. Moreover, elevated blood pressure increases proteinuria by increasing its flux across the glomerular capillaries. Conversely, the renoprotective effect of antihypertensive medications is gauged through the consequent reduction of proteinuria. Thus, the more effective a given treatment is in lowering protein excretion, the greater the subsequent impact on protection from decline in GFR. This observation is the basis for the treatment guideline establishing 130/80 mmHg as a target blood pressure in proteinuric CKD patients.

Several controlled studies have shown that ACE inhibitors and ARBs are effective in slowing the progression of renal failure in patients with advanced stages of both diabetic and nondiabetic CKD, in large part through effects on efferent vasodilatation and the subsequent decline in glomerular hypertension. In the absence of an anti-proteinuric response with either agent alone, combined treatment with both ACE inhibitors and ARBs has been considered. The combination is associated with a greater reduction in proteinuria compared to either agent alone. Insofar as reduction in proteinuria is a surrogate for improved renal outcome, the combination would appear to be advantageous. However, there is a greater incidence of acute kidney injury and adverse cardiac events from such combination therapy. On balance, therefore, ACE inhibitor plus ARB therapy should be avoided. A progressive increase in serum creatinine concentration with these agents may suggest the presence of renovascular disease within the large or small arteries. Development of side effects may mandate the use of second-line antihypertensive agents instead of ACE inhibitors or ARBs. Among the calcium channel blockers, diltiazem and verapamil may exhibit superior antiproteinuric and renoprotective effects compared to the dihydropyridines. At least two different categories of response can be considered: one in which progression is strongly associated with systemic and

intraglomerular hypertension and proteinuria (e.g., diabetic nephropathy, glomerular diseases) and in which ACE inhibitors and ARBs are likely to be the first choice; and another in which proteinuria is mild or absent initially (e.g., adult polycystic kidney disease and other tubulointerstitial diseases), where the contribution of intraglomerular hypertension is less prominent and other antihypertensive agents can be useful for control of systemic hypertension.

SLOWING THE PROGRESSION OF DIABETIC NEPHROPATHY

See Chap. 397

MANAGING OTHER COMPLICATIONS OF CKD

Medication Dose Adjustment

Although the loading dose of most drugs is not affected by CKD because renal elimination is not used in the calculation, the maintenance doses of many drugs will need to be adjusted. For those agents in which >70% excretion is by a nonrenal route, such as hepatic elimination, dose adjustment may not be needed. Some drugs that should be avoided include metformin, meperidine, and oral anti-hyperglycemics that are eliminated by the kidney. NSAIDs should be avoided because of the risk of further worsening of kidney function. Many antibiotics, antihypertensives, and antiarrhythmics may require a reduction in dosage or change in the dose interval. Several online Web-based databases for dose adjustment of medications according to stage of CKD or estimated GFR are available (e.g., *http://www.globalrph.com/index_renal.htm*). Nephrotoxic radiocontrast agents and gadolinium should be avoided or used according to strict guidelines when medically necessary as discussed above.

PREPARATION FOR RENAL REPLACEMENT THERAPY

(See also Chap. 307) Temporary relief of symptoms and signs of impending uremia, such as anorexia, nausea, vomiting, lassitude, and pruritus, may sometimes be achieved with dietary protein restriction. However, this carries a risk of malnutrition, and thus plans for more long-term management should be in place.

Maintenance dialysis and kidney transplantation have extended the lives of hundreds of thousands of patients with CKD worldwide. Clear indications for initiation of renal replacement therapy for patients with CKD include uremic pericarditis, encephalopathy, intractable muscle cramping, anorexia, and nausea not attributable to reversible causes such as peptic ulcer disease, evidence of malnutrition, and fluid and electrolyte abnormalities, principally hyperkalemia or ECFV overload, that are refractory to other measures.

Recommendations for the Optimal Time for Initiation of Renal Replacement Therapy

Because of the individual variability in the severity of uremic symptoms and renal function, it is ill-advised to assign an arbitrary urea nitrogen or creatinine level to the need to start dialysis. Moreover, patients may become accustomed to chronic uremia and deny symptoms, only to find that they feel better with dialysis and realize in retrospect how poorly they were feeling before its initiation.

Previous studies suggested that starting dialysis before the onset of severe symptoms and signs of uremia was associated with prolongation of survival. This led to the concept of “healthy” start and is congruent with the philosophy that it is better to keep patients feeling well rather than allowing them to become ill with uremia and then attempting to return them to better health with dialysis or transplantation. Although recent studies have not confirmed an association of early-start dialysis with improved patient survival, there may be merit in this approach for some patients. On a practical level, advanced preparation may help to avoid problems with the dialysis process

itself (e.g., a poorly functioning fistula for hemodialysis or malfunctioning peritoneal dialysis catheter) and, thus, preempt the morbidity associated with resorting to the insertion of temporary hemodialysis access with its attendant risks of sepsis, bleeding, thrombosis, and association with accelerated mortality.

Patient Education

Social, psychological, and physical preparation for the transition to renal replacement therapy and the choice of the optimal initial modality are best accomplished with a gradual approach involving a multidisciplinary team. Along with conservative measures discussed in the sections above, it is important to prepare patients with an intensive educational program, explaining the likelihood and timing of initiation of renal replacement therapy and the various forms of therapy available, and the option of nondialytic conservative care. The more knowledgeable that patients are about hemodialysis (both in-center and home-based), peritoneal dialysis, and kidney transplantation, the easier and more appropriate will be their decisions. Patients who are provided with education are more likely to choose home-based dialysis therapy. This approach is of societal benefit because home-based therapy is less expensive and is associated with improved quality of life. The educational programs should be commenced no later than stage 4 CKD so that the patient has sufficient time and cognitive function to learn the important concepts, make informed choices, and implement preparatory measures for renal replacement therapy.

Exploration of social support is also important. Early education of family members for selection and preparation of a home dialysis helper or a biologically or emotionally related potential living kidney donor should occur long before the onset of symptomatic renal failure.

Kidney transplantation **(Chap. 307)** offers the best potential for complete rehabilitation, because dialysis replaces only a small fraction of the kidneys' filtration function and none of the other renal functions, including endocrine and anti-inflammatory effects. Generally, kidney transplantation follows a period of dialysis treatment, although preemptive kidney transplantation (usually from a living donor) can be carried out if it is certain that the renal failure is irreversible.

IMPLICATIONS FOR GLOBAL HEALTH

 In contrast to the natural decline and successful eradication of many devastating infectious diseases, there is rapid growth in the prevalence of metabolic and vascular disease in developing countries. Diabetes mellitus is becoming increasingly prevalent in these countries, perhaps due in part to change in dietary habits, diminished physical activity, and weight gain. Therefore, it follows that there will be a proportionate increase in vascular and renal disease. Health care agencies must plan for improved screening of high-risk individuals for early detection, prevention, and treatment plans in these nations and must start considering options for improved availability of renal replacement therapies.

There is also increasing recognition of endemic nephropathies in developing countries that particularly target young males working in agriculture. The extent of morbidity and mortality associated with these nephropathies is only starting to be appreciated. It is unclear what the cause is, but a combination of population genetic risk with endemic nephrotoxins, exposure to pesticides, NSAID use, and chronic volume depletion have all been suggested to contribute.

FURTHER READING

COCA SG et al: Chronic kidney disease after acute kidney injury: A systematic review and meta-analysis. Kidney Int 81:442, 2012. [PubMed: 22113526]

CORESH J et al: Decline in estimated glomerular filtration rate and subsequent risk of end-stage renal disease and mortality. JAMA 311:2518, 2014. [PubMed: 24892770]

CORREA-ROTTER R et al: CKD of unknown origin in Central America: The case for a Mesoamerican nephropathy. Am J Kidney Dis 63:506, 2014. [PubMed: 24412050]

TANGRI N et al: Multinational assessment of accuracy of equations for predicting risk of kidney failure: A meta-analysis. JAMA 315:164, 2016. [PubMed: 26757465]

WANNER C et al: Empagliflozin and progression of kidney disease in type 2 diabetes. N Engl J Med 375:323, 2016. [PubMed: 27299675]

Chapter 306: Dialysis in the Treatment of Renal Failure

Kathleen D. Liu; Glenn M. Chertow

FIGURE 306-1

INTRODUCTION

Dialysis may be required for the treatment of either acute or chronic kidney disease (CKD). The use of continuous renal replacement therapies (CRRT) and prolonged intermittent renal replacement therapy (PIRRT)/slow low-efficiency dialysis (SLED) is specific to the management of acute renal failure and is discussed in **Chap. 304**. These modalities are performed continuously (CRRT) or over 6–12 h per session (PIRRT/SLED), in contrast to the 3–4 h of an intermittent hemodialysis session. Advantages and disadvantages of CRRT and PIRRT/SLED are discussed in **Chap. 304.**

Peritoneal dialysis is rarely used in developed countries for the treatment of acute renal failure because of the increased risk of infection and (as will be discussed in more detail below) less efficient clearance per unit of time. The focus of this chapter will be on the use of peritoneal and hemodialysis for end-stage renal disease (ESRD).

With the widespread availability of dialysis, the lives of hundreds of thousands of patients with ESRD have been prolonged. In the United States alone, there are now ~675,000 patients with treated ESRD (kidney failure requiring dialysis or transplantation), the vast majority of whom require dialysis. Since 2000, the prevalence of treated ESRD has increased 74%, which reflects both a small increase in the incidence rate and marginally enhanced survival of patients receiving dialysis. The incidence rate for treated ESRD in the United States is 370 cases per million population per year; ESRD is disproportionately higher in African Americans (875 per million population per year) as compared with white Americans (285 per million population per year). In the United States, the leading cause of ESRD is diabetes mellitus, currently accounting for almost 45% of newly diagnosed cases of ESRD. Approximately 30% of patients have ESRD that has been attributed to hypertension, although it is unclear whether in these cases hypertension is the cause or a consequence of vascular disease or other unknown causes of kidney failure. Other prevalent causes of ESRD include glomerulonephritis, polycystic kidney disease, and obstructive uropathy. A fraction of the excess incidence of ESRD in African Americans is likely related to transmission of high-risk alleles for the *APOL1* gene.

Globally, mortality rates for patients with ESRD are lowest in Europe and Japan but very high in the developing world because of the limited availability of dialysis. In the United States, the mortality rate of patients on dialysis has decreased slightly but remains extremely high, with a 5-year survival rate of ~40% for patients receiving dialysis. Deaths are due mainly to cardiovascular diseases and infections (~40 and 10% of deaths, respectively). Older age, male sex, nonblack race, diabetes mellitus, malnutrition, and underlying heart disease are important predictors of death.

TREATMENT OPTIONS FOR PATIENTS WITH ESRD

Commonly accepted criteria for initiating patients on maintenance dialysis include the presence of uremic symptoms, the presence of hyperkalemia unresponsive to conservative measures, persistent extracellular volume expansion despite diuretic therapy, acidosis refractory to medical therapy, a bleeding diathesis, and a creatinine clearance or estimated glomerular filtration rate (GFR) <10 mL/min per 1.73 m^2 (see **Chap. 305** for estimating equations). Timely referral to a nephrologist for advanced planning and creation of a dialysis access, education about ESRD treatment options, and management of the complications of advanced CKD, including hypertension, anemia, acidosis, and secondary hyperparathyroidism, is advisable. Recent data have suggested that a sizable fraction of ESRD cases result following episodes of acute renal failure, particularly among persons with underlying CKD. Furthermore, there is no benefit to initiating dialysis preemptively at a GFR of 10–14 mL/min per 1.73 m^2 compared to initiating dialysis for symptoms of uremia.

In ESRD, treatment options include hemodialysis (in center or at home); peritoneal dialysis, as either continuous ambulatory peritoneal dialysis (CAPD) or continuous cyclic peritoneal dialysis (CCPD); or transplantation **(Chap. 307)**. Although there are significant geographic variations and differences in practice patterns, in-center hemodialysis remains the most common therapeutic modality for ESRD (>90% of patients) in the United States. In contrast to hemodialysis, peritoneal dialysis is continuous, but much less efficient, in terms of solute clearance. While no large-scale clinical trials have been completed comparing outcomes among patients randomized to either hemodialysis or peritoneal dialysis, outcomes associated with both therapies are similar in most reports, and the decision of which modality to select is often based on personal preferences and quality-of-life considerations.

HEMODIALYSIS

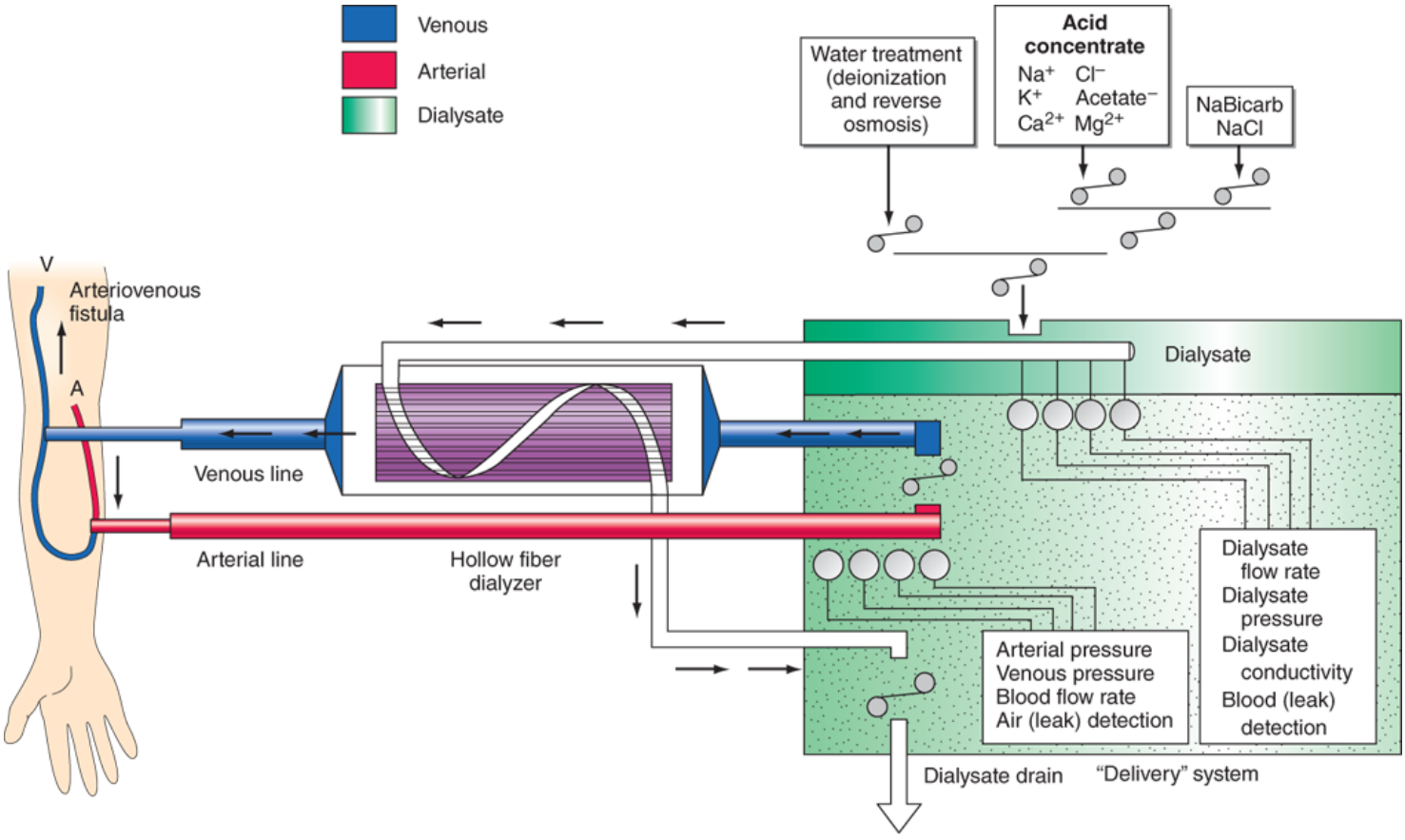
Hemodialysis relies on the principles of solute diffusion across a semipermeable membrane. Movement of metabolic waste products takes place down a concentration gradient from the circulation into the dialysate. The rate of diffusive transport increases in response to several factors, including the magnitude of the concentration gradient, the membrane surface area, and the mass transfer coefficient of the membrane. The latter is a function of the porosity and thickness of the membrane, the size of the solute molecule, and the conditions of flow on the two sides of the membrane. According to laws of diffusion, the larger the molecule, the slower its rate of transfer across the membrane. A small molecule, such as urea (60 Da), undergoes substantial clearance, whereas a larger molecule, such as creatinine (113 Da), is cleared less efficiently. In addition to diffusive clearance, movement of waste products from the circulation into the dialysate may occur as a result of ultrafiltration. Convective clearance occurs because of solvent drag, with solutes being swept along with water across the semipermeable dialysis membrane.

THE DIALYZER

There are three essential components to hemodialysis: the dialyzer, the composition and delivery of the dialysate, and the blood delivery system **(Fig. 306-1)**. The dialyzer is a plastic chamber with the ability to perfuse blood and dialysate compartments simultaneously at very high flow rates. The hollow-fiber dialyzer is the most common in use in the United States. These dialyzers are composed of bundles of capillary tubes through which blood circulates while dialysate travels on the outside of the fiber bundle. Virtually all dialyzers now manufactured in the United States are “biocompatible” synthetic membranes derived from polysulfone or related compounds (versus older cellulose “bioincompatible” membranes that activated the complement cascade). The frequency of reprocessing and reuse of hemodialyzers and blood lines varies across the world. In general as the cost of disposable supplies has decreased, their use has increased. In the United States, reprocessing of dialyzers is now extremely rare. Formaldehyde, peracetic acid–hydrogen peroxide, glutaraldehyde, and bleach have all been used as reprocessing agents.

FIGURE 306-1

Schema for hemodialysis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIALYSATE

The potassium concentration of dialysate may be varied from 0 to 4 mmol/L depending on the predialysis serum potassium concentration. The use of 0 or 1 mmol/L potassium dialysate is becoming less common owing to data suggesting that patients who undergo treatments with very low potassium dialysate have an increased risk of sudden death, perhaps due to arrhythmias in the setting of potassium shifts. The usual dialysate calcium concentration is 1.25 mmol/L (2.5 mEq/L), although modification may be required in selected settings (e.g., higher dialysate calcium concentrations

may be used in patients with hypocalcemia associated with secondary hyperparathyroidism or with “hungry bone syndrome” following parathyroidectomy). The usual dialysate sodium concentration is 136–140 mmol/L. In patients who frequently develop hypotension during their dialysis run, “sodium modeling” to counterbalance urea-related osmolar gradients may be employed. With sodium modeling, the dialysate sodium concentration is gradually lowered from the range of 145–155 mmol/L to isotonic concentrations (136–140 mmol/L) near the end of the dialysis treatment, typically declining either in steps or in a linear or exponential fashion. However, higher dialysate sodium concentrations and sodium modeling may predispose patients to positive sodium balance and increased thirst; thus, these strategies to ameliorate intradialytic hypotension may be undesirable in patients with hypertension or in patients with large interdialytic weight gains. Because patients are exposed to ~120 L of water during each dialysis treatment, water used for the dialysate is subjected to filtration, softening, deionization, and, ultimately, reverse osmosis to remove microbiologic contaminants and dissolved ions.

BLOOD DELIVERY SYSTEM

The blood delivery system is composed of the extracorporeal circuit and the dialysis access. The dialysis machine consists of a blood pump, dialysis solution delivery system, and various safety monitors. The blood pump moves blood from the access site, through the dialyzer, and back to the patient. The blood flow rate typically ranges from 250 to 450 mL/min, depending on the type and integrity of the vascular access. Negative hydrostatic pressure on the dialysate side can be manipulated to achieve desirable fluid removal or *ultrafiltration*. Dialysis membranes have different ultrafiltration coefficients (i.e., mL removed/min per mmHg) so that along with hydrostatic changes, fluid removal can be varied. The dialysis solution delivery system dilutes the concentrated dialysate with water and monitors the temperature, conductivity, and flow of dialysate.

DIALYSIS ACCESS

The fistula, graft, or catheter through which blood is obtained for hemodialysis is often referred to as a *hemodialysis (or vascular) access*. A native fistula created by the anastomosis of an artery to a vein (e.g., the Brescia-Cimino fistula, in which the cephalic vein is anastomosed end-to-side to the radial artery) results in arterialization of the vein. This facilitates its subsequent use in the placement of large needles (typically 15 gauge) to access the circulation. Fistulas have the highest long-term patency rate of all hemodialysis access options. For patients in whom fistulas fail to mature, or in patients whose vasculature does not allow creation of a successful fistula (i.e., poor arterial inflow or recipient veins of inadequate caliber), patients undergo placement of an arteriovenous graft (i.e., the interposition of prosthetic material, usually polytetrafluoroethylene, between an artery and a vein) or a tunneled hemodialysis catheter. In recent years, nephrologists, vascular surgeons, and health care policy makers in the United States have encouraged creation of arteriovenous fistulas in a larger fraction of patients (the “fistula first” initiative). Unfortunately, even when created, arteriovenous fistulas may not mature sufficiently to provide reliable access to the circulation, or they may thrombose early in their development.

The most important complication of arteriovenous grafts is thrombosis of the graft and graft failure, due principally to intimal hyperplasia at the anastomosis between the graft and recipient vein. When grafts (or fistulas) fail, catheter-guided angioplasty can be used to dilate stenoses; monitoring of venous pressures on dialysis and of access flow, although not routinely performed, may assist in the early recognition of impending vascular access failure. In addition to increased rates of access failure, grafts and (in particular) catheters are associated with much higher rates of infection than fistulas.

Intravenous large-bore catheters are often used in patients with acute renal failure and CKD. For persons on maintenance hemodialysis, tunneled catheters (either two separate catheters or a single catheter with two lumens) are often used when arteriovenous fistulas and grafts have failed or are not feasible due to anatomic considerations. These catheters are tunneled under the skin; the tunnel reduces bacterial translocation from the skin, resulting in a lower infection rate than with nontunneled temporary catheters. Most tunneled catheters are placed in the internal jugular veins; the external jugular, femoral, and subclavian veins may also be used.

Nephrologists, interventional radiologists, and vascular surgeons generally prefer to avoid placement of catheters into the subclavian veins; while flow rates are usually excellent, subclavian stenosis is a frequent complication and, if present, will likely prohibit permanent vascular access (i.e., a fistula or graft) in the ipsilateral extremity. Infection rates may be higher with femoral catheters. For patients with multiple vascular access complications and no other options for permanent vascular access, tunneled catheters may be the last “lifeline” for hemodialysis. Translumbar or transhepatic approaches into the inferior vena cava may be required if the superior vena cava or other central veins draining the upper extremities are stenosed or thrombosed.

GOALS OF DIALYSIS

The hemodialysis procedure consists of pumping heparinized blood through the dialyzer at a flow rate of 250–450 mL/min, while dialysate flows in an opposite *counter-current* direction at 500–800 mL/min. The efficiency of dialysis is determined by blood and dialysate flow through the dialyzer as well as dialyzer characteristics (i.e., its efficiency in removing solute). The *dose* of dialysis, which is currently defined as a derivation of the fractional urea clearance during a single treatment, is further governed by patient size, residual kidney function, dietary protein intake, the degree of anabolism or catabolism, and the presence of comorbid conditions.

Since the landmark studies of Sargent and Gotch relating the measurement of the dose of dialysis using urea concentrations with morbidity in the National Cooperative Dialysis Study, the *delivered* dose of dialysis has been measured and considered as a quality assurance and improvement tool. While the fractional removal of urea nitrogen and derivations thereof are considered to be the standard methods by which “adequacy of dialysis” is

measured, a large multicenter randomized clinical trial (the HEMO Study) failed to show a difference in mortality associated with a large difference in per-session urea clearance. Current targets include a urea reduction ratio (the fractional reduction in blood urea nitrogen per hemodialysis session) of >65–70% and a body water–indexed clearance × time product (Kt/V) >1.2 or 1.05, depending on whether urea concentrations are "equilibrated." For the majority of patients with ESRD, between 9 and 12 h of dialysis are required each week, usually divided into three equal sessions. Several studies have suggested that longer hemodialysis session lengths may be beneficial (independent of urea clearance), although these studies are confounded by a variety of patient characteristics, including body size and nutritional status. Hemodialysis "dose" should be individualized, and factors other than the urea nitrogen should be considered, including the adequacy of ultrafiltration or fluid removal and control of hyperkalemia, hyperphosphatemia, and metabolic acidosis. A randomized clinical trial comparing 6 versus 3 times per week hemodialysis (the "Frequent Hemodialysis Network Daily Trial") demonstrated improved control of hypertension and hyperphosphatemia, reduced left ventricular mass, and improved self-reported physical health with more frequent hemodialysis. Secondary analyses also demonstrated improvements in other metrics of health-related quality of life, including improved self-reported general health and a reduced "time to recovery" (time until usual activities can be resumed) among patients randomized to more frequent hemodialysis. A companion trial in which frequent nocturnal hemodialysis was compared to conventional hemodialysis at home showed no significant effect on left ventricular mass or self-reported physical health. Finally, an evaluation of the U.S. Renal Data System registry showed a significant increase in mortality and hospitalization for heart failure after the longer interdialytic interval that occurs over the dialysis "weekend."

COMPLICATIONS DURING HEMODIALYSIS

Hypotension is the most common acute complication of hemodialysis, particularly among patients with diabetes mellitus. Numerous factors appear to increase the risk of hypotension, including excessive ultrafiltration with inadequate compensatory vascular filling, impaired vasoactive or autonomic responses, osmolar shifts, overzealous use of antihypertensive agents, and reduced cardiac reserve. Patients with arteriovenous fistulas and grafts may develop high-output cardiac failure due to shunting of blood through the dialysis access; on rare occasions, this may necessitate ligation of the fistula or graft. The management of hypotension during dialysis consists of discontinuing ultrafiltration, the administration of 100–250 mL of isotonic saline, or administration of salt-poor albumin. Hypotension during dialysis can frequently be prevented by careful evaluation of the dry weight and by ultrafiltration modeling, such that more fluid is removed at the beginning rather than the end of the dialysis procedure. Excessively rapid fluid removal (>13 mL/kg per h) should be avoided, as rapid fluid removal has been associated with adverse outcomes, including cardiovascular deaths. Additional maneuvers to prevent intradialytic hypotension include the performance of sequential ultrafiltration followed by dialysis, cooling of the dialysate during dialysis treatment, and avoiding heavy meals during dialysis. Midodrine, an oral selective α1 adrenergic agent, has been advocated by some practitioners, although there is insufficient evidence of its safety and efficacy to support its routine use.

Muscle cramps during dialysis are also a common complication. The etiology of dialysis-associated cramps remains obscure. Changes in muscle perfusion because of excessively rapid volume removal or targeted removal below the patient's estimated dry weight often precipitate dialysis-associated cramps. Strategies that may be used to prevent cramps include reducing volume removal during dialysis, ultrafiltration profiling, and the use of sodium modeling (see above).

Anaphylactoid reactions to the dialyzer, particularly on its first use, have been reported most frequently with the bioincompatible cellulosic-containing membranes. Dialyzer reactions can be divided into two types, A and B. Type A reactions are attributed to an IgE-mediated intermediate hypersensitivity reaction to ethylene oxide used in the sterilization of new dialyzers. This reaction typically occurs soon after the initiation of a treatment (within the first few minutes) and can progress to full-blown anaphylaxis if the therapy is not promptly discontinued. Treatment with steroids or epinephrine may be needed if symptoms are severe. The type B reaction consists of a symptom complex of nonspecific chest and back pain, which appears to result from complement activation and cytokine release. These symptoms typically occur several minutes into the dialysis run and typically resolve over time with continued dialysis.

PERITONEAL DIALYSIS

In peritoneal dialysis, 1.5–3 L of a dextrose-containing solution is infused into the peritoneal cavity and allowed to dwell for a set period of time, usually 2–4 h. As with hemodialysis, metabolic byproducts are removed through a combination of convective clearance generated through ultrafiltration and diffusive clearance down a concentration gradient. The clearance of solutes and water during a peritoneal dialysis exchange depends on the balance between the movement of solute and water into the peritoneal cavity versus absorption from the peritoneal cavity. The rate of diffusion diminishes with time and eventually stops when equilibration between plasma and dialysate is reached. Absorption of solutes and water from the peritoneal cavity occurs across the peritoneal membrane into the peritoneal capillary circulation and via peritoneal lymphatics into the lymphatic circulation. The rate of peritoneal solute transport varies from patient to patient and may be altered by the presence of infection (peritonitis), drugs, and physical factors such as position and exercise.

FORMS OF PERITONEAL DIALYSIS

Peritoneal dialysis may be carried out as CAPD, CCPD, or a combination of both. In CAPD, dialysate is manually infused into the peritoneal cavity and exchanged three to five times during the day. A nighttime dwell is frequently instilled at bedtime and remains in the peritoneal cavity through the night. In CCPD, exchanges are performed in an automated fashion, usually at night; the patient is connected to an automated cycler that performs a series of

exchange cycles while the patient sleeps. The number of exchange cycles required to optimize peritoneal solute clearance varies by the peritoneal membrane characteristics; as with hemodialysis, solute clearance should be tracked to ensure dialysis "adequacy."

Peritoneal dialysis solutions are available in volumes typically ranging from 1.5 to 3 L. The major difference between the dialysate used for peritoneal rather than hemodialysis is that the hypertonicity of peritoneal dialysis solutions drives solute and fluid removal, whereas solute removal in hemodialysis depends on concentration gradients, and fluid removal requires transmembrane pressure. Typically, dextrose at varying concentrations contributes to the hypertonicity of peritoneal dialysate. Icodextrin is a nonabsorbable carbohydrate that can be used in place of dextrose. Studies have demonstrated more efficient ultrafiltration with icodextrin than with dextrose-containing solutions. Icodextrin is typically used as the "last fill" for patients on CCPD or for the longest dwell in patients on CAPD. The most common additives to peritoneal dialysis solutions are heparin to prevent obstruction of the dialysis catheter lumen with fibrin and antibiotics during an episode of acute peritonitis. Insulin may also be added in patients with diabetes mellitus.

ACCESS TO THE PERITONEAL CAVITY

Access to the peritoneal cavity is obtained through a peritoneal catheter. Catheters used for maintenance peritoneal dialysis are flexible, being made of silicone rubber with numerous side holes at the distal end. These catheters usually have two Dacron cuffs. The scarring that occurs around the cuffs anchors the catheter and seals it from bacteria tracking from the skin surface into the peritoneal cavity; it also prevents the external leakage of fluid from the peritoneal cavity. The cuffs are placed in the preperitoneal plane and ~2 cm from the skin surface.

The *peritoneal equilibrium test* is a formal evaluation of peritoneal membrane characteristics that measures the transfer rates of creatinine and glucose across the peritoneal membrane. Patients are classified as low, low–average, high–average, and high transporters. Patients with rapid equilibration (i.e., high transporters) tend to absorb more glucose and lose efficiency of ultrafiltration with long daytime dwells. High transporters also tend to lose larger quantities of albumin and other proteins across the peritoneal membrane. In general, patients with rapid transporting characteristics require more frequent, shorter dwell time exchanges, nearly always obligating use of a cycler. Slower (low and low–average) transporters tend to do well with fewer exchanges. The efficiency of solute clearance also depends on the volume of dialysate infused. Larger volumes allow for greater solute clearance, particularly with CAPD in patients with low and low–average transport characteristics.

As with hemodialysis, the optimal dose of peritoneal dialysis is unknown. Several observational studies have suggested that higher rates of urea and creatinine clearance (the latter generally measured in L/week) are associated with lower mortality rates and fewer uremic complications. However, a randomized clinical trial (Adequacy of Peritoneal Dialysis in Mexico [ADEMEX]) failed to show a significant reduction in mortality or complications with a relatively large increment in urea clearance. In general, patients on peritoneal dialysis do well when they retain residual kidney function. Rates of technique failure increase with years on dialysis and have been correlated with loss of residual function to a greater extent than loss of peritoneal membrane capacity. For some patients in whom CCPD does not provide sufficient solute clearance, a hybrid approach can be adopted where one or more daytime exchanges are added to the CCPD regimen. While this approach can enhance solute clearance and prolong a patient's capacity to remain on peritoneal dialysis, the burden of the hybrid approach can be overwhelming.

COMPLICATIONS DURING PERITONEAL DIALYSIS

The major complications of peritoneal dialysis are peritonitis, catheter-associated nonperitonitis infections, weight gain and other metabolic disturbances, and residual uremia (especially among patients with little or no residual kidney function).

Peritonitis typically develops when there has been a break in sterile technique during one or more of the exchange procedures. Peritonitis is usually defined by an elevated peritoneal fluid leukocyte count (100/mm^3, of which at least 50% are polymorphonuclear neutrophils); these cutoffs are lower than in spontaneous bacterial peritonitis because of the presence of dextrose in peritoneal dialysis solutions and rapid bacterial proliferation in this environment without antibiotic therapy. The clinical presentation typically consists of pain and cloudy dialysate, often with fever and other constitutional symptoms. The most common culprit organisms are gram-positive cocci, including *Staphylococcus,* reflecting the origin from the skin. Gram-negative rod infections are less common; fungal and mycobacterial infections can be seen in selected patients, particularly after antibacterial therapy. Most cases of peritonitis can be managed either with intraperitoneal or oral antibiotics, depending on the organism; many patients with peritonitis do not require hospitalization. In cases where peritonitis is due to hydrophilic gram-negative rods (e.g., *Pseudomonas* sp.) or yeast, antimicrobial therapy is usually not sufficient, and catheter removal is required to ensure complete eradication of infection. Nonperitonitis catheter-associated infections (often termed *tunnel infections*) vary widely in severity. Some cases can be managed with local antibiotic or silver nitrate administration, while others are severe enough to require parenteral antibiotic therapy and catheter removal.

Peritoneal dialysis is associated with a variety of metabolic complications. Albumin and other proteins can be lost across the peritoneal membrane in concert with the loss of metabolic wastes. Hypoproteinemia obligates a higher dietary protein intake in order to maintain nitrogen balance. Hyperglycemia and weight gain are also common complications of peritoneal dialysis. Several hundred calories in the form of dextrose are absorbed each day, depending on the concentration of dextrose employed. Patients receiving peritoneal dialysis, particularly those with diabetes mellitus, are prone to other complications of insulin resistance, including hypertriglyceridemia. On the positive side, the continuous nature of peritoneal dialysis

usually allows for a more liberal diet, due to continuous removal of potassium and phosphorus—two major dietary components whose accumulation can be hazardous in ESRD.

LONG-TERM OUTCOMES IN ESRD

Cardiovascular disease constitutes the major cause of death in patients with ESRD. Cardiovascular mortality and event rates are higher in patients receiving dialysis than in patients posttransplantation, although rates are extraordinarily high in both populations. The underlying cause of cardiovascular disease is unclear but may be related to shared risk factors (e.g., diabetes mellitus, hypertension, atherosclerotic and arteriosclerotic vascular disease), chronic inflammation, massive changes in extracellular volume (especially with high interdialytic weight gains), inadequate treatment of hypertension, dyslipidemia, anemia, dystrophic (vascular) calcification, and, perhaps, alterations in cardiovascular dynamics during the dialysis treatment. Few studies have targeted cardiovascular risk reduction in ESRD patients; none have demonstrated consistent benefit. Two clinical trials of statin agents in ESRD demonstrated significant reductions in low-density lipoprotein (LDL) cholesterol concentrations but no significant reductions in death or cardiovascular events (Die Deutsche Diabetes Dialyse Studie [4D] and AURORA studies). The Study of Heart and Renal Protection (SHARP) which included patients on dialysis and others with nondialysis-requiring CKD showed a 17% reduction in the rate of major cardiovascular events or cardiovascular death with simvastatin-ezatamide treatment. Most experts recommend conventional cardioprotective strategies (e.g., lipid-lowering agents, aspirin, inhibitors of the renin-angiotensin-aldosterone system, and β-adrenergic antagonists) in patients receiving dialysis based on the patients' cardiovascular risk profile, which appears to be increased by more than an order of magnitude relative to persons unaffected by kidney disease. Other complications of ESRD include a high incidence of infection, progressive debility and frailty, protein-energy malnutrition, and impaired cognitive function.

GLOBAL PERSPECTIVE

 The incidence of ESRD is increasing worldwide with longer life expectancies and improved care of infectious and cardiovascular diseases. The management of ESRD varies widely by country and within country by region, and it is influenced by economic and other major factors. In general, peritoneal dialysis is more commonly performed in poorer countries owing to its lower expense and the high cost of establishing in-center hemodialysis units.

FURTHER READING

COOPER ba et al: A randomized, controlled trial of early versus late initiation of dialysis. N Engl J Med 363:609, 2010. [PubMed: 20581422]

CORREA-ROTTER R et al: Peritoneal dialysis, in *Brenner and Rector's The Kidney,* 9th ed, MW Taal, GM Chertow, PA Marsden, K Skorecki, and ASL Yu (eds). Philadelphia, Elsevier, 2011.

FELLSTROM BC et al: Rosuvastatin and cardiovascular events in patients undergoing hemodialysis. N Engl J Med 360:1395, 2009. [PubMed: 19332456]

FLYTHE JE et al: Rapid fluid removal during dialysis is associated with cardiovascular morbidity and mortality. Kidney Int 79:250, 2011. [PubMed: 20927040]

FOLEY rn et al: Long interdialytic interval and mortality among patients receiving hemodialysis. N Engl J Med 365:1099, 2011. [PubMed: 21992122]

FREQUENT HEMODIALYSIS NETWORK TRIAL GROUP: In-center hemodialysis six times per week versus three times per week. N Engl J Med 363:2287, 2010. [PubMed: 21091062]

NATIONAL KIDNEY FOUNDATION: Kidney disease quality initiative clinical practice guidelines: Hemodialysis and peritoneal dialysis adequacy, 2006. Available online: *http://www.kidney.org/professionals/kdoqi/guidelines.cfm*.

ROCCO mv et al: The effects of frequent nocturnal home hemodialysis: The frequent hemodialysis network nocturnal trial. Kidney Int 80:1080, 2011. [PubMed: 21775973]

U.S. RENAL DATA SYSTEM: USRDS 2016 Annual Data Report: Atlas of End-Stage Renal Disease in the United States. Bethesda, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Disease, 2016.

Chapter 307: Transplantation in the Treatment of Renal Failure

Jamil Azzi; Edgar L. Milford; Mohamed H. Sayegh; Anil Chandraker

INTRODUCTION

Transplantation of the human kidney is the treatment of choice for advanced chronic renal failure. Worldwide, tens of thousands of these procedures have been performed with >180,000 patients bearing functioning kidney transplants in the United States today. When azathioprine and prednisone initially were used as immunosuppressive drugs in the 1960s, the results with properly matched familial donors were superior to those with organs from deceased donors: 75–90% compared with 50–60% graft survival rates at 1 year. During the 1970s and 1980s, the success rate at the 1-year mark for deceased-donor transplants rose progressively. Currently, deceased-donor grafts have a 92% 1-year survival and living-donor grafts have a 97% 1-year survival. Although there has been improvement in long-term survival, it has not been as impressive as the short-term survival, and currently the “average” ($t1/2$) life expectancy of a living-donor graft is around 14 years and that of a deceased-donor graft is close to 10 years.

Mortality rates after transplantation are highest in the first year and are age-related: 2% for ages 18–34 years, 3% for ages 35–49 years, and 6.8% for ages ≥50–60 years. These rates compare favorably with those in the chronic dialysis population even after risk adjustments for age, diabetes, and cardiovascular status. While the loss of kidney transplant due to acute rejection is currently rare, most allografts succumb at varying rates to a chronic process consisting of interstitial fibrosis, tubular atrophy, vasculopathy, and glomerulopathy, the pathogenesis of which is incompletely understood. Overall, transplantation returns most patients to an improved lifestyle and an improved life expectancy compared with patients on dialysis.

RECENT ACTIVITY AND RESULTS

In 2014 there were more than 12,328 deceased-donor kidney transplants and 5574 living-donor transplants in the United States, with the ratio of deceased to living donors remaining stable over the last few years. The backlog of patients with end-stage renal disease (ESRD) has been increasing every year, and it always lags behind the number of available donors. As the number of patients with end-stage kidney disease increases, the demand for kidney transplants continues to increase. As of 2015, there were 50,692 active adult candidates on the waiting list, and <18,000 patients were transplanted. This imbalance is set to worsen over the coming years with the predicted increased rates of obesity and diabetes worldwide. In an attempt to

increase utilization of marginal kidneys while insuring longevity-matching, a new allocation system was developed and recently implemented. The main rule is that patients expected to survive the longest receive the allografts expected to last the longest. For this purpose, the Kidney Donor Profile Index (KDPI) score from 0 to 100% has been introduced to quantify the potential risk of graft failure after kidney transplant based on 10 donor factors. The lower KDPI values are associated with higher expected post-transplant survival. Hence, kidneys with KDPI <20% are allocated to the 20% of the potential recipients with the highest expected postransplant survival. The kidneys with KDPI >85% (previously called expanded criteria donor or ECD kidneys) are usually used for older patients who are expected to fare less well on dialysis. Kidneys from donors after cardiac death (DCD) are also been used to overcome the increasing demand on the waiting list (Table 307-1).

TABLE 307-1

Definition of a Non-Heart-Beating Donor (Donation After Cardiac Death[a] [DCD])

I: Brought in dead
II: Unsuccessful resuscitation
III: Awaiting cardiac arrest
IV: Cardiac arrest after brainstem death
V: Cardiac arrest in a hospital patient

[a]Kidneys can be used for transplantation from categories II–V but are commonly only used from categories III and IV. The survival of these kidneys has not been shown to be inferior to that of deceased-donor kidneys.

Note: Kidneys can both have a KDPI >85 % and be DCD. High KDPI kidneys have been shown to have a poorer survival, and there is a separate shorter waiting list for those kidneys. They are generally utilized for patients for whom the benefits of being transplanted earlier outweigh the associated risks of using a lower quality kidney.

The overall results of transplantation are presented in Table 307-2 as the survival of grafts and of patients. At the 1-year mark, graft survival is higher for living-donor recipients, most likely because those grafts are not subject to as much ischemic injury. The more effective drugs now in use for immunosuppression have almost equalized the risk of graft rejection in all patients for the first year. At 5 and 10 years, however, there has been a steeper decline in survival of those with deceased-donor kidneys.

TABLE 307-2

Mean Rates of Graft and Patient Survival for Kidneys Transplanted in the United States from 1998 to 2008[a]

	1-Year Follow-Up		5-Year Follow-Up		10-Year Follow-Up	
	Grafts, %	Patients, %	Grafts, %	Patients, %	Grafts, %	Patients, %
Deceased donor	92	96	72	84	46	64
Living donor	96	99	81	91	59	77

[a]All patients transplanted are included, and the follow-up unadjusted survival data from the 1-, 5- , and 10-year periods are presented to show the attrition rates over time within the two types of organ donors.

Source: Data from Summary Tables, 2009 Annual Reports, Scientific Registry of Transplant Recipients.

RECIPIENT SELECTION

There are few absolute contraindications to renal transplantation. The transplant procedure is relatively noninvasive, as the organ is generally placed in the inguinal fossa without entering the peritoneal cavity. Recipients without perioperative complications often can be discharged from the hospital in excellent condition within 5 days of the operation.

Virtually all patients with ESRD who receive a transplant have a higher life expectancy than do risk-matched patients who remain on dialysis. Even though diabetic patients and older candidates have a higher mortality rate than other transplant recipients, their survival is improved with transplantation compared with those remaining on dialysis. This global benefit of transplantation as a treatment modality poses substantial ethical issues for policy makers, as the number of deceased donor kidneys available is far from sufficient to meet the current needs of the candidates. The current standard of care is that the candidate should have a life expectancy of >5 years to be put on a deceased organ wait list. Even for living donation, the candidate should have >5 years of life expectancy. This standard has been established because the benefits of kidney transplantation over dialysis are realized only after a perioperative period in which the mortality rate is higher in transplanted patients than in dialysis patients with comparable risk profiles.

All candidates must have a thorough risk-versus-benefit evaluation before being approved for transplantation. In particular, an aggressive approach to diagnosis of correctable coronary artery disease, presence of latent or indolent infection (HIV, hepatitis B or C, tuberculosis), and neoplasm should be a routine part of the candidate workup. Most transplant centers consider overt AIDS and active hepatitis absolute contraindications to transplantation because of the high risk of opportunistic infection. Some centers are now transplanting individuals with hepatitis and even HIV infection under strict protocols to

determine whether the risks and benefits favor transplantation over dialysis. Over the last few years, new direct acting hepatitis C antiviral medications have been introduced and have been shown to be very effective therapies both pre- and posttransplant. Those medications are reshaping our approach to patients with hepatitis C.

Among the few absolute “immunologic” contraindications to transplantation is the presence of antibodies against the donor kidney at the time of the anticipated transplant that can cause hyperacute rejection. Those harmful antibodies include natural antibodies against the ABO blood group antigens and antibodies against human leukocyte antigen (HLA) class I (A, B, C) or class II (DR, DQ, DP) antigens. These antibodies are routinely excluded by proper screening of the candidate’s ABO compatibility and direct cytotoxic cross-matching of candidate serum with lymphocytes of the donor.

TISSUE TYPING AND CLINICAL IMMUNOGENETICS

Matching for antigens of the HLA major histocompatibility gene complex (Chap. 343) is an important criterion for selection of donors for renal allografts. Each mammalian species has a single chromosomal region that encodes the strong, or major, transplantation antigens, and this region on the human sixth chromosome is called *HLA*. HLA antigens have been classically defined by serologic techniques, but methods to define specific nucleotide sequences in genomic DNA are increasingly being used. Other “minor” antigens may play crucial roles, in addition to the ABH(O) blood groups and endothelial antigens that are not shared with lymphocytes. The Rh system is not expressed on graft tissue. Evidence for designation of HLA as the genetic region that encodes major transplantation antigens comes from the success rate in living related donor renal and bone marrow transplantation, with superior results in HLA-identical sibling pairs. Nevertheless, 5% of HLA-identical renal allografts are rejected, often within the first weeks after transplantation. These failures represent states of prior sensitization to non-HLA antigens. Non-HLA minor antigens are relatively weak when initially encountered and are, therefore, suppressible by conventional immunosuppressive therapy. Once priming has occurred, however, secondary responses are much more refractory to treatment.

DONOR SELECTION

Donors can be deceased or volunteer living donors. When first-degree relatives are donors, graft survival rates at 1 year are 5–7% greater than those for deceased-donor grafts. The 5-year survival rates still favor a partially matched (3/6 HLA mismatched) family donor over a randomly selected cadaver donor. In addition, living donors provide the advantage of immediate availability. For both living and deceased donors, the 5-year outcomes are somewhat poorer if there is a complete (6/6) HLA mismatch.

The survival rate of living unrelated renal allografts is as high as that of perfectly HLA-matched cadaver renal transplants and comparable to that of kidneys from living relatives. This outcome is probably a consequence of both short cold ischemia time and the extra care taken to document that the condition and renal function of the donor are optimal before proceeding with a living unrelated donation. It is illegal in the United States to purchase organs for transplantation.

Living volunteer donors should be cleared of any medical conditions that may cause morbidity and mortality after kidney transplantation. Concern has been expressed about the potential risk to a volunteer kidney donor of premature renal failure after several years of increased blood flow and hyperfiltration per nephron in the remaining kidney. There are a few reports of the development of hypertension, proteinuria, and even lesions of focal segmental sclerosis in donors over long-term follow-up. It is also desirable to consider the risk of development of type 1 diabetes mellitus in a family member who is a potential donor to a diabetic renal failure patient. Anti-insulin and anti-islet cell antibodies should be measured, and glucose tolerance tests should be performed in such donors to exclude a prediabetic state. Selective renal arteriography should be performed on donors to rule out the presence of multiple or abnormal renal arteries, because the surgical procedure is difficult, and the ischemic time of the transplanted kidney is long when there are vascular abnormalities. Transplant surgeons commonly use a laparoscopic approach to isolate and remove the living donor's kidney. This operation has the advantage of less evident surgical scars, and, as there is less tissue trauma, laparoscopic donors have a substantially shorter hospital stay and less discomfort than those who undergo an open nephrectomy.

Deceased donors should be free of malignant neoplastic disease, hepatitis, and HIV owing to possible transmission to the recipient, although under certain circumstances hepatitis C- and HIV-positive organs may be used in previously infected recipients. Increased risk of graft failure exists when the donor is elderly or has acute renal failure or when the kidney has a prolonged period of ischemia.

In the United States, there is a coordinated national system of regulations, allocation support, and outcomes analysis for kidney transplantation called the Organ Procurement Transplant Network. It is now possible to remove deceased-donor kidneys and maintain them for up to 48 h on cold pulsatile perfusion or with simple flushing and cooling. Although generally an ischemic time of <24 h is preferred, this approach permits adequate time for typing, cross-matching, transportation, and selection problems to be solved.

PRESENSITIZATION

A positive cytotoxic cross-match of recipient serum with donor T lymphocytes indicates the presence of pre-formed donor specific anti-HLA class I antibodies and is usually predictive of an acute vasculitic event termed *hyperacute rejection*. This finding, along with ABO incompatibility, represents the only absolute immunologic contraindication for kidney transplantation. Recently, an increasing number of tissue typing laboratories have shifted to a flow cytometric-based crossmatch assay, which detects the presence of anti HLA antibodies that are not necessarily detected on a cytotoxic crossmatch assay and may not be an absolute contraindication to transplantation. The known sources of such sensitization are blood transfusion, a prior transplant, pregnancy, and vaccination/infection. Patients sustained by dialysis often show fluctuating antibody titers and specificity patterns. At the time of assignment of a cadaveric kidney, cross-matches are performed with at least a current serum. Previously analyzed antibody specificities and additional cross-matches are performed accordingly. Flow cytometry detects binding of anti-HLA antibodies of candidate serum by recipient's lymphocytes. This highly sensitive test can be useful for avoidance of accelerated, and often untreatable, early graft rejection in patients receiving second or third transplants.

For the purposes of crossmatching, donor T lymphocytes, which express class I but not class II antigens, are used as targets for detection of anti–class I (HLA-A and -B) antibodies that are expressed on all nucleated cells. Preformed anti–class II (HLA-DR and -DQ) antibodies against the donor also carry a higher risk of graft loss, particularly in recipients who have suffered early loss of a prior kidney transplant. B lymphocytes, which express both class I and class II antigens, are used as targets in these assays. Furthermore, donor-specific HLA antibodies that fix complements have been shown to strongly correlate with antibody mediated rejection and worse long-term outcome.

Some non-HLA antigens restricted in expression to endothelium and monocytes have been described, but their clinical relevance is not well established. A series of minor histocompatibility antigens do not elicit antibodies, and sensitization to these antigens is detectable only by cytotoxic T cells, an assay too cumbersome for routine use.

Desensitization before transplantation by reducing the level of anti-donor antibodies utilizing plasmapheresis and administration of pooled immunoglobulin, or both, has been useful in reducing the risk of hyperacute rejection following transplantation.

IMMUNOLOGY OF REJECTION

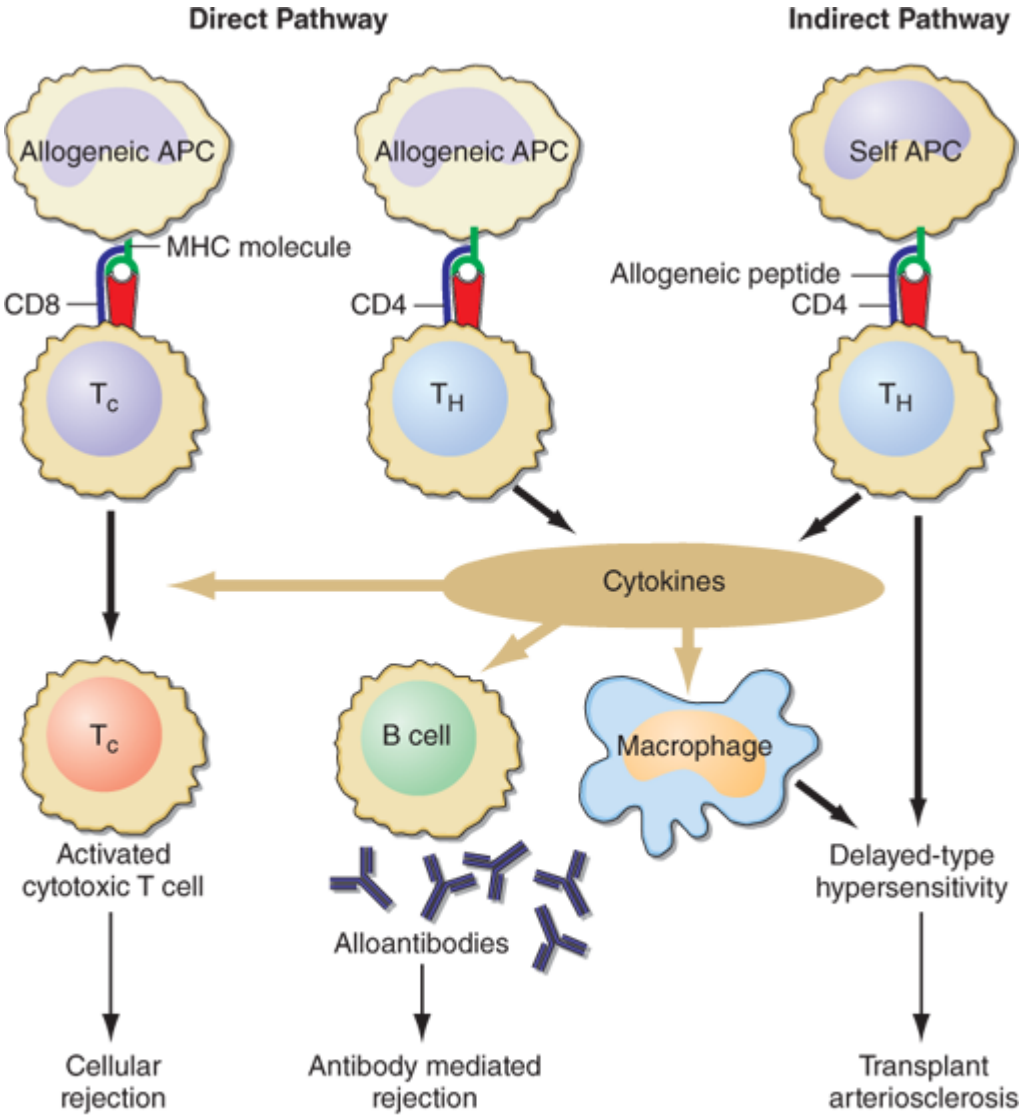
Both cellular and humoral (antibody-mediated) effector mechanisms can play roles in kidney transplant rejection.

Cellular rejection is mediated by lymphocytes that respond to HLA antigens expressed within the organ. The CD4+ lymphocyte responds to class II (HLA-DR) incompatibility by proliferating and releasing proinflammatory cytokines that augment the proliferative response of the immune system. CD8+ cytotoxic lymphocyte precursors respond primarily to class I (HLA-A, -B) antigens and mature into cytotoxic effector cells that cause organ damage through direct contact and lysis of donor target cells. Full T cell activation requires not only T cell receptor binding to the allo-antigens presented by self or donor HLA molecules (indirect and direct presentation respectively), but also engaging costimulatory molecules such as CD28 on T cells and CD80 and CD86 ligands on antigen presenting cells **(Fig. 307-1)**. Signaling through both of these pathways induces activation of the kinase activity of calcineurin which, in turn, activates transcription factors leading to upregulation of multiple genes, including interleukin-2 (IL-2) and interferon gamma. IL-2 signals through the target of rapamycin (TOR) to induce cell proliferation in an autocrine fashion. There is evidence that non-HLA antigens can also play a role in renal transplant rejection episodes. Recipients who receive a kidney from an HLA-identical sibling can have rejection episodes and require maintenance immunosuppression, whereas identical twin transplants require no immunosuppression. There are documented non-HLA antigens, such as an endothelial-specific antigen system with limited polymorphism and a tubular antigen, which can act as targets of humoral or cellular rejection responses, respectively.

FIGURE 307-1

Recognition pathways for major histocompatibility complex (MHC) antigens. Graft rejection is initiated by CD4 helper T lymphocytes (TH) having antigen receptors that bind to specific complexes of peptides and

MHC class II molecules on antigen-presenting cells (APC). In transplantation, in contrast to other immunologic responses, there are two sets of T cell clones involved in rejection. In the direct pathway the class II MHC of donor allogeneic APCs is recognized by CD4 TH cells that bind to the intact MHC molecule, and class I MHC allogeneic cells are recognized by CD8 T cells. The latter generally proliferate into cytotoxic cells (TC). In the indirect pathway, the incompatible MHC molecules are processed into peptides that are presented by the self-APCs of the recipient. The indirect, but not the direct, pathway is the normal physiologic process in T cell recognition of foreign antigens. Once TH cells are activated, they proliferate and, by secretion of cytokines and direct contact, exert strong helper effects on macrophages, TC, and B cells. *(From MH Sayegh, LH Turka: N Engl J Med, 338:1813, 1998. Copyright 1998, Massachusetts Medical Society. All rights reserved.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IMMUNOSUPPRESSIVE TREATMENT

Immunosuppressive therapy, as currently available, generally suppresses all immune responses, including those to bacteria, fungi, and even malignant tumors. In general, all clinically useful drugs are more selective to primary than to memory immune responses. Agents to suppress the immune response are classically divided into induction and maintenance agents, and will be discussed in the following paragraphs. Those currently in clinical use are listed in Table 307-3.

TABLE 307-3

Maintenance Immunosuppressive Drugs

Agent	Pharmacology	Mechanisms	Side Effects
Glucocorticoids	Increased bioavailability with hypoalbuminemia and liver disease; prednisone/prednisolone generally used	Binds cytosolic receptors and heat shock proteins. Blocks transcription of IL-1, -2, -3, -6, TNF-α, and IFN-γ	Hypertension, glucose intolerance, dyslipidemia, osteoporosis
Cyclosporine (CsA)	Lipid-soluble polypeptide, variable absorption, microemulsion more predictable	Trimolecular complex with cyclophilin and calcineurin → block in cytokine (e.g., IL-2) production; however, stimulates TGF-β production	Nephrotoxicity, hypertension, dyslipidemia, glucose intolerance, hirsutism/hyperplasia of gums
Tacrolimus	Macrolide, well absorbed	Trimolecular complex with FKBP-12 and calcineurin → block in cytokine (e.g., IL-2) production; may stimulate TGF-β production	Similar to CsA, but hirsutism/hyperplasia of gums unusual, and diabetes more likely
Azathioprine	Mercaptopurine analogue	Hepatic metabolites inhibit purine synthesis	Marrow suppression (WBC > RBC > platelets)
Mycophenolate mofetil/sodium	Metabolized to mycophenolic acid	Inhibits purine synthesis via inosine monophosphate dehydrogenase	Diarrhea/cramps; dose-related liver and marrow suppression is uncommon
Sirolimus/Everolimus	Macrolide, poor oral bioavailability	Complexes with FKBP-12 and then blocks p70 S6 kinase in the IL-2 receptor pathway for proliferation	Hyperlipidemia, thrombocytopenia

Agent	Pharmacology	Mechanisms	Side Effects
Belatacept	Fusion protein, intravenous injections	Binds CD80 and CD86, prevents CD28 binding and T cell activation	Posttransplant Lymphoproliferative Disease (PTLD)

Abbreviations: FKBP-12, FK506 binding protein 12; IFN, interferon; IL, interleukin; RBC, red blood cells; TGF, transforming growth factor; TNF, tumor necrosis factor; WBC, white blood cells.

INDUCTION THERAPY

Induction therapy is currently given to most kidney transplant recipients in the United States at the time of transplant to reduce the risk of early acute rejection and to minimize or eliminate the use of either steroids or calcineurin inhibitors and their associated toxicities. Induction therapy consists of antibodies that could be monoclonal or polyclonal, depletional or nondepletional.

Depleting Agents

Anti-thymocyte globulin (ATG): peripheral human lymphocytes, thymocytes, or lymphocytes from spleens or thoracic duct fistulas are injected into horses, rabbits, or goats to produce anti-lymphocyte serum, from which the globulin fraction is then separated. Those polyclonal antibodies induce lymphocyte depletion, and the immune system may take several months to recover.

Monoclonal antibodies against defined lymphocyte subsets offer a more precise and standardized form of therapy. Alemtuzumab is directed to the CD52 protein, widely distributed on immune cells such as B and T cells, natural killer cells, macrophages, and some granulocytes.

Nondepleting Agents

Another approach to more selective therapy is to target the 55-kDa alpha chain of the IL-2 receptor, which is expressed only on T cells that have been recently activated. This approach is used as prophylaxis for acute rejection in the immediate posttransplant period, and is effective at decreasing the early acute rejection rate with few adverse side effects.

The next step in the evolution of this therapeutic strategy, which has already been achieved in the short term in small numbers of immunologically well-matched patients, is the elimination of all maintenance immunosuppression therapy.

MAINTENANCE THERAPY

All kidney transplant recipients should receive maintenance immunosuppressive therapies except identical twins. The most frequently used combination is triple therapy with prednisone, a calcineurin inhibitor, and an antimetabolite; mTOR inhibitors can replace one of the last two agents. More recently, the FDA approved a new costimulatory blocking antibody, belatacept, as a new strategy to prevent long-term CNI toxicity.

Antimetabolites

Azathioprine, an analogue of mercaptopurine, was for two decades the keystone to immunosuppressive therapy in humans, but has given way to more effective agents. This agent can inhibit synthesis of DNA, RNA, or both. Azathioprine is administered in doses of 1.5–2 mg/kg per day. Reduction in the dose is required because of leukopenia and occasionally thrombocytopenia. Excessive amounts of azathioprine may also cause jaundice, anemia, and alopecia. If it is essential to administer allopurinol concurrently, the azathioprine dose must be reduced. As inhibition of xanthine oxidase delays degradation, this combination is best avoided.

Mycophenolate mofetil or mycophenolate sodium, both of which are metabolized to mycophenolic acid, is now used in place of azathioprine in most centers. It has a similar mode of action and a mild degree of gastrointestinal toxicity but produces less bone marrow suppression. Its advantage is its increased potency in preventing or reversing rejection.

Steroids

Glucocorticoids are important adjuncts to immunosuppressive therapy. Among all the agents employed, prednisone has effects that are easiest to assess, and in large doses it is usually effective for the reversal of rejection. In general, 200–300 mg prednisone is given immediately before or at the time of transplantation, and the dose is reduced to 30 mg within a week. The side effects of the glucocorticoids, particularly impairment of wound healing and predisposition to infection, make it desirable to taper the dose as rapidly as possible in the immediate postoperative period. Many centers now have protocols for early discontinuance or avoidance of steroids because of long-term adverse effects on bone, skin, and glucose metabolism. For treatment of acute rejection, methylprednisolone, 0.5–1 g IV, is administered immediately upon diagnosis of beginning rejection and continued once daily for 3 days. Such “pulse” doses are not effective in chronic rejection. Most patients whose renal function is stable after 6 months or a year do not require large doses of prednisone; maintenance doses of 5–10 mg/d are the rule. A major effect of steroids is preventing the release of interleukin (IL) 6 and IL-1 by monocytes-macrophages.

Calcineurin Inhibitors

Cyclosporine is a fungal peptide with potent immunosuppressive activity. It acts on the calcineurin pathway to block transcription of mRNA for IL-2 and other proinflammatory cytokines, thereby inhibiting T cell proliferation. Although it works alone, cyclosporine is more effective in conjunction with glucocorticoids and mycophenolate. Clinical results with tens of thousands of renal transplants have been impressive. Among its

toxic effects (nephrotoxicity, hepatotoxicity, hirsutism, tremor, gingival hyperplasia, and diabetes), only nephrotoxicity presents a serious management problem and is further discussed below.

Tacrolimus (previously called FK506) is a fungal macrolide that has the same mode of action as cyclosporine as well as a similar side-effect profile; it does not, however, produce hirsutism or gingival hyperplasia. De novo diabetes mellitus is more common with tacrolimus. The drug was first used in liver transplantation and may substitute for cyclosporine entirely or as an alternative in renal patients whose rejections are poorly controlled by cyclosporine. An extended release formulation of tacrolimus is now available and is given once daily.

TOR Inhibitors

Sirolimus (previously called rapamycin) is another fungal macrolide but has a different mode of action; i.e., it inhibits T cell growth factor signaling pathways, preventing the response to IL-2 and other cytokines. Sirolimus can be used in conjunction with cyclosporine or tacrolimus, or with mycophenolic acid, to avoid the use of calcineurin inhibitors.

Everolimus is another mTOR inhibitor with similar mechanism of action as *Sirolimus* but with better bioavailability.

Belatacept

Belatacept is a fusion protein that binds costimulatory ligands (CD80 and CD86) present on antigen presenting cells, interrupting their binding to CD28 on T cells. This inhibition leads to T cell anergy and apoptosis. Belatacept is FDA-approved for kidney transplant recipients and is given monthly as an intravenous infusion. The 7 years follow-up of the Belatacept Evaluation of Nephroprotection and Efficacy as First-line Immunosuppression Trial (BENEFIT) showed higher patient and graft survival for the belatacept treated group compared to cyclosporine.

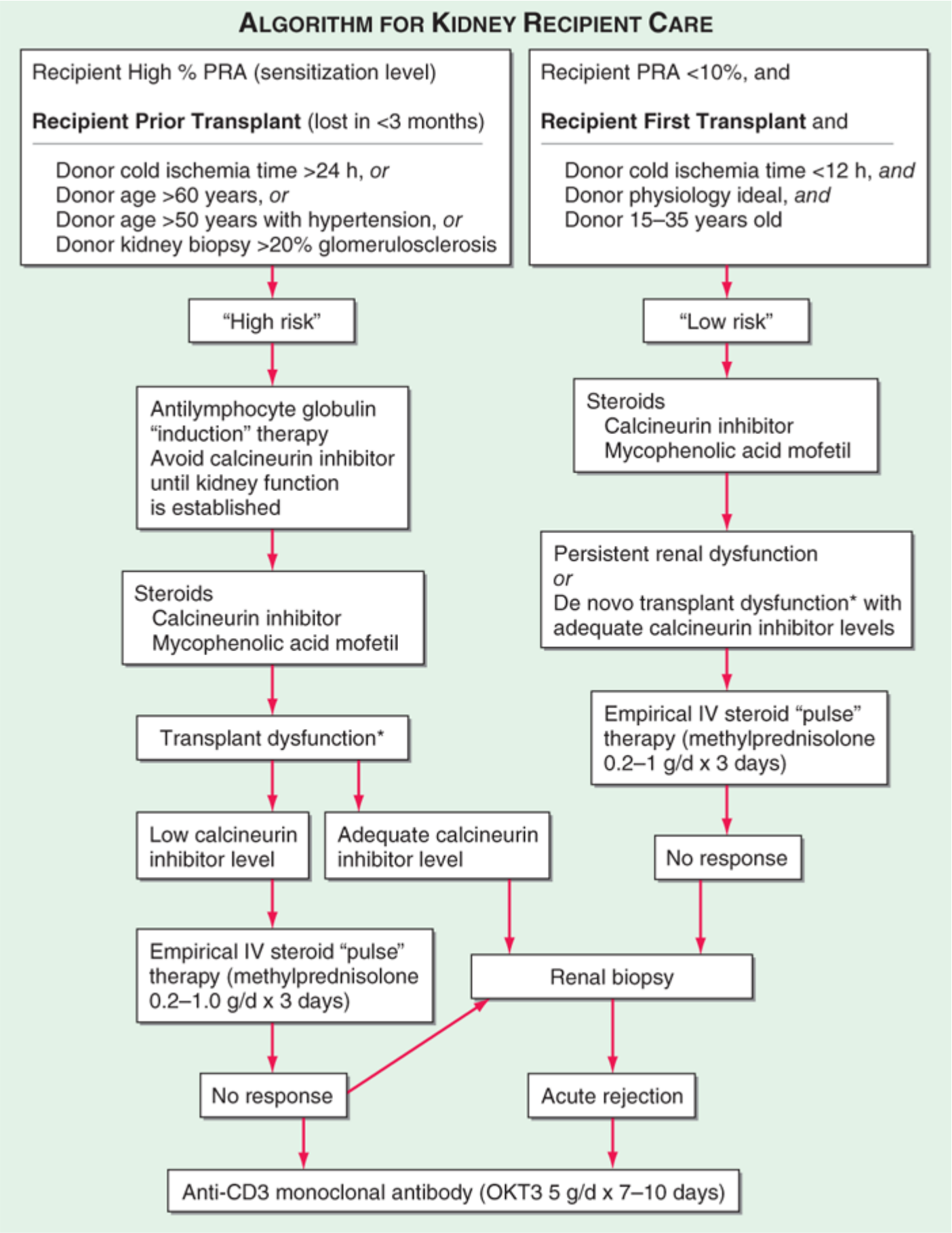
CLINICAL COURSE AND MANAGEMENT OF THE RECIPIENT

Adequate hemodialysis should be performed within 48 h of surgery as necessary, and care should be taken that the serum potassium level is not markedly elevated so that intraoperative cardiac arrhythmias can be averted. The diuresis that commonly occurs postoperatively must be carefully monitored. In some instances, it may be massive, reflecting the inability of ischemic tubules to regulate sodium and water excretion; with large diureses, large amounts of potassium may be lost. Most chronically uremic patients have some excess of extracellular fluid, and it is useful to maintain an expanded fluid volume in the immediate postoperative period. Acute tubular necrosis (ATN) due to ischemia may cause immediate oliguria or may follow an initial short period of graft function. Recovery usually occurs within 3 weeks, although periods as long as 6 weeks have been reported. Superimposition of rejection on ATN is common, and the differential diagnosis may be difficult without a graft biopsy. Cyclosporine therapy prolongs ATN, and some patients do not diurese until the dose is reduced drastically. Many centers avoid starting calcinenrins for the first several days, using ALG

or a monoclonal antibody along with mycophenolic acid and prednisone until renal function is established. Figure 307-2 illustrates an algorithm followed by many transplant centers for early posttransplant management of recipients at high or low risk of early renal dysfunction.

FIGURE 307-2

A typical algorithm for early posttransplant care of a kidney recipient. If any of the recipient or donor "high-risk" factors exist, more aggressive management is called for. Low-risk patients can be treated with a standard immunosuppressive regimen. Patients at higher risk of rejection or early ischemic and nephrotoxic transplant dysfunction are often induced with an antilymphocyte globulin to provide more potent early immunosuppression or to spare calcineurin nephrotoxicity. *When there is early transplant dysfunction, prerenal, obstructive, and vascular causes must be ruled out by ultrasonographic examination. The panel reactive antibody (PRA) is a quantitation of how much antibody is present in a candidate against a panel of cells representing the distribution of antigens in the donor pool. APC, antigen-presenting cell; MHC, major histocompatibility complex.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE REJECTION EPISODE

Early diagnosis of rejection allows prompt institution of therapy to preserve renal function and prevent irreversible damage. Clinical evidence of rejection is rarely characterized by fever, swelling, and tenderness over the allograft. Rejection may present only with a rise in serum creatinine, with or without a reduction in urine volume. The focus should be on ruling out other causes of functional deterioration.

Doppler ultrasonography may be useful in ascertaining changes in the renal vasculature and in renal blood flow. Thrombosis of the renal vein occurs rarely; it may be reversible if it is caused by technical factors and intervention is prompt. Diagnostic ultrasound is the procedure of choice to rule out urinary obstruction or to confirm the presence of perirenal collections of urine, blood, or lymph. A rise in the serum creatinine level is a late marker of rejection, but it may be the only sign. Novel biomarkers are needed for early noninvasive detection of allograft rejection.

Calcineurin inhibitors (cyclosporine and tacrolimus) have an afferent arteriolar constrictor effect on the kidney, and may produce permanent vascular and interstitial injury after sustained high-dose therapy. This action will lead to a deterioration in renal function difficult to distinguish from rejection without a renal biopsy. Interstitial fibrosis, isometric tubular vacuolization, and thickening of arteriolar walls are suggestive of this side effect, but not diagnostic. Hence, if no rejection is detected on the biopsy, serum creatinine may respond to a reduction in dose. However, if cellular rejection activity is present in the biopsy, appropriate therapy is indicated. The first rejection episode is usually treated with IV administration of methylprednisolone, 500–1000 mg daily for 3 days. Failure to respond is an indication for antibody therapy, usually with antithymocyte globulin.

Evidence of antibody-mediated injury is present when endothelial injury and deposition of complement component c4d is detected by fluorescence labeling. This is usually accompanied by detection of the antibody in the recipient blood. The prognosis is poor, and aggressive use of plasmapheresis, immunoglobulin infusions, anti-CD20 monoclonal antibody (rituximab) to target B lymphocytes, bortezomib to target antibody producing plasma cells, and eculizumab to inhibit complement is indicated.

MANAGEMENT PROBLEMS

The typical times after transplantation when the most common opportunistic infections occur are shown in **Table 307-4**. Prophylaxis for cytomegalovirus (CMV) and Pneumocystis *Jirovecci* pneumonia is given for 6–12 months after transplantation.

TABLE 307-4

The Most Common Opportunistic Infections in Renal Transplant Recipients

Peritransplant (<1 month)	Late (>6 months)
Wound infections	*Aspergillus*
Herpesvirus	*Nocardia*
Oral candidiasis	BK virus (polyoma)
Urinary tract infection	Herpes zoster
Early (1–6 months)	Hepatitis B
Pneumocystis carinii	Hepatitis C
Cytomegalovirus	
Legionella	
Listeria	
Hepatitis B	
Hepatitis C	

The signs and symptoms of infection may be masked or distorted. Fever without obvious cause is common, and only after days or weeks may it become apparent that it has a viral or fungal origin. Bacterial infections are most common during the first month after transplantation. The importance of blood cultures in such patients cannot be overemphasized because systemic infection without obvious foci is common. Particularly ominous are rapidly occurring pulmonary lesions, which may result in death within 5 days of onset. When these lesions become apparent, immunosuppressive agents should be discontinued, except for maintenance doses of prednisone.

Aggressive diagnostic procedures, including transbronchial and open-lung biopsy, are frequently indicated. In the case of *Pneumocystis carinii* **(Chap. 215)** infection, trimethoprim-sulfamethoxazole (TMP-SMX) is the treatment of choice; amphotericin B has been used effectively in systemic fungal infections. Prophylaxis against *P. jirovecci* with daily or alternate-day low-dose TMP-SMX is very effective. Involvement of the oropharynx with *Candida* **(Chap. 211)** may be treated with local nystatin. Tissue-invasive fungal infections require treatment with systemic agents such as fluconazole. Small doses (a total of 300 mg) of amphotericin

given over a period of 2 weeks may be effective in fungal infections refractory to fluconazole. Macrolide antibiotics, especially ketoconazole and erythromycin, and some calcium channel blockers (diltiazem, verapamil) compete with calcineurin inhibitors for P450 catabolism and cause elevated levels of these immunosuppressive drugs. Analeptics, such as phenytoin and carbamazepine, will increase catabolism to result in low levels. *Aspergillus* **(Chap. 212)**, *Nocardia* **(Chap. 169)**, and especially CMV **(Chap. 190)** infections also occur.

CMV is a common and dangerous DNA virus in transplant recipients. It does not generally appear until the end of the first posttransplant month. Active CMV infection is sometimes associated, or occasionally confused, with rejection episodes. Patients at highest risk for severe CMV disease are those without anti-CMV antibodies who receive a graft from a CMV antibody–positive donor (15% mortality). Valganciclovir is a cost-effective and bioavailable oral form of ganciclovir that has been proved effective in both prophylaxis and treatment of CMV disease. Early diagnosis in a febrile patient with clinical suspicion of CMV disease can be made by determining CMV viral load in the blood. A rise in IgM antibodies to CMV is also diagnostic. Culture of CMV from blood may be less sensitive. Tissue invasion of CMV is common in the gastrointestinal tract and lungs. CMV retinopathy occurs late in the course, if untreated. Treatment of active CMV disease with valganciclovir is always indicated. In many patients immune to CMV, viral activation can occur with major immunosuppressive regimens.

The polyoma group (BK, JC, SV40) is another class of DNA viruses that can become dormant in kidneys and can be activated by immunosuppression. When reactivation occurs with BK, if left untreated, there is a 50% chance of progressive fibrosis and loss of the graft within 1 year by the activated virus. Risk of infection is associated with the overall degree of immunosuppression rather than the individual immunosuppressive drugs used. Renal biopsy is necessary for the diagnosis. There have been variable results with leflunomide, cidofovir, and quinolone anitibiotics (which are effective against polyoma helicase), but it is most important to reduce the immunosuppressive load.

The complications of glucocorticoid therapy are well known and include gastrointestinal bleeding, impairment of wound healing, osteoporosis, diabetes mellitus, cataract formation, and hemorrhagic pancreatitis. The treatment of unexplained jaundice in transplant patients should include cessation or reduction of immunosuppressive drugs if hepatitis or drug toxicity is suspected. Therapy in such circumstances often does not result in rejection of a graft, at least for several weeks. Acyclovir is effective in therapy for herpes simplex virus infections.

CHRONIC LESIONS OF THE TRANSPLANTED KIDNEY

Although 1-year transplant survival is excellent, most recipients experience progressive decline in kidney function over time thereafter. Chronic renal transplant dysfunction can be caused by recurrent disease, hypertension, cyclosporine or tacrolimus nephrotoxicity, chronic immunologic rejection, secondary focal glomerulosclerosis, or a combination of these pathophysiologies. Chronic vascular changes with intimal proliferation and medial hypertrophy are commonly found. Control of systemic and intrarenal hypertension with ACE inhibitors is thought to have a beneficial influence on the rate of progression of chronic renal

transplant dysfunction. Renal biopsy can distinguish subacute cellular rejection from recurrent disease or secondary focal sclerosis.

MALIGNANCY

The incidence of tumors in patients on immunosuppressive therapy is 5–6%, or ~100 times greater than that in the general population in the same age range. The most common lesions are cancer of the skin and lips and carcinoma in situ of the cervix, as well as lymphomas such as non-Hodgkin's lymphoma. The risks are increased in proportion to the total immunosuppressive load administered and the time elapsed since transplantation. Surveillance for skin and cervical cancers is necessary.

OTHER COMPLICATIONS

Both chronic dialysis and renal transplant patients have a higher incidence of death from myocardial infarction and stroke than does the population at large, and this is particularly true in diabetic patients. Contributing factors are the use of glucocorticoids and sirolimus, as well as hypertension. Recipients of renal transplants have a high prevalence of coronary artery and peripheral vascular diseases. The percentage of deaths from these causes has been slowly rising as the numbers of transplanted diabetic patients and the average age of all recipients increase. More than 50% of renal recipient mortality is attributable to cardiovascular disease. In addition to strict control of blood pressure and blood lipid levels, close monitoring of patients for indications of further medical or surgical intervention is an important part of management.

Hypertension may be caused by (1) native kidney disease, (2) rejection activity in the transplant, (3) renal artery stenosis if an end-to-end anastomosis was constructed with an iliac artery branch, and (4) renal calcineurin inhibitor toxicity, which may improve with reduction in dose. Whereas ACE inhibitors may be useful, calcium channel blockers are more frequently used initially. Amelioration of hypertension to the range of 120–130/70–80 mmHg should be the goal in all patients.

Hypercalcemia after transplantation may indicate failure of hyperplastic parathyroid glands to regress. Aseptic necrosis of the head of the femur is probably due to preexisting hyperparathyroidism, with aggravation by glucocorticoid treatment. With improved management of calcium and phosphorus metabolism during chronic dialysis, the incidence of parathyroid-related complications has fallen dramatically. Persistent hyperparathyroid activity may require subtotal parathyroidectomy.

Although most transplant patients have robust production of erythropoietin and normalization of hemoglobin, *anemia* is commonly seen in the posttransplant period. Often the anemia is attributable to bone marrow–suppressant immunosuppressive medications such as azathioprine, mycophenolic acid, and sirolimus. Gastrointestinal bleeding is a common side effect of high-dose and long-term steroid administration. Many transplant patients have creatinine clearances of 30–50 mL/min and can be considered in the same way as other patients with chronic renal insufficiency for anemia management, including supplemental erythropoietin.

Chronic hepatitis, particularly when due to hepatitis B virus, can be a progressive, fatal disease over a decade or so. Patients who are persistently hepatitis B surface antigen–positive are at higher risk, according to some studies, but the presence of hepatitis C virus is also a concern when one embarks on a course of immunosuppression in a transplant recipient. However, the introduction of the new highly effective direct acting hepatitis C antiviral medications promises to reduce this risk significantly.

FURTHER READING

Birnbaum LM et al: Management of chronic allograft nephropathy: A systematic review. Clin J Am Soc Nephrol 4:860, 2009. [PubMed: 19339427]

Briganti EM et al: Risk of renal allograft loss from recurrent glomerulonephritis. N Engl J Med 347:103, 2002. [PubMed: 12110738]

Freeman RB Jr: Survival benefit: Quality versus quantity and trade-offs in developing new renal allocation systems. Am J Transplant 7:1043, 2007. [PubMed: 17359499]

Hirsch HH et al: Polyomavirus-associated nephropathy in renal transplantation: Interdisciplinary analyses and recommendations. Transplantation 79:1277, 2005. [PubMed: 15912088]

Hodson EM et al: Antiviral medications for preventing cytomegalovirus disease in solid organ transplant recipients. Cochrane Database Syst Rev CD003774, 2008.

Ibrahim HN et al: Long-term consequences of kidney donation. N Engl J Med 360:459, 2009. [PubMed: 19179315]

KDIGO clinical practice guideline for the care of kidney transplant recipients. Am J Transplant 9 Suppl 3:S1, 2009. [PubMed: 19845597]

Lentine KL et al: Incidence and predictors of myocardial infarction after kidney transplantation. J Am Soc Nephrol 16:496, 2005.

Loupy A et al: Complement-binding anti-HLA antibodies and kidney-allograft survival. N Engl J Med 369:1215, 2013. [PubMed: 24066742]

Matas AJ et al: OPTN/SRTR 2013 Annual Data Report: Kidney. Am J Transplant 15 Suppl 2:1, 2015.

Riella LV, Sheridan AM: Testing for high-risk APOL1 alleles in potential living kidney donors. Am J Kidney Dis 66:396, 2015. [PubMed: 26049628]

Sawinski D et al: Successful treatment of Hepatitis C in renal transplant recipients with direct-acting antiviral agents. Am J Transplant 16:1588, 2017.

Stegall MD: Clinical management of renal transplant patients with donor-specific alloantibody: The state of the art. Clin Transpl 307, 2010.

Stewart DE et al: Changes in deceased donor kidney transplantation one year after KAS implementation. Am J Transplant 16:1834, 2016. [PubMed: 26932731]

Vincenti F et al: Belatacept and long-term outcomes in kidney transplantation. N Engl J Med 374:333, 2016. [PubMed: 26816011]

Webster AC et al: Identifying high risk groups and quantifying absolute risk of cancer after kidney transplantation: a cohort study of 15,183 recipients. Am J Transplant 7:2140, 2007. [PubMed: 17640312]

Wolfe RA et al: Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. N Engl J Med 341:1725, 1999. [PubMed: 10580071]

Chapter 308: Glomerular Diseases

Julia B. Lewis; Eric G. Neilson

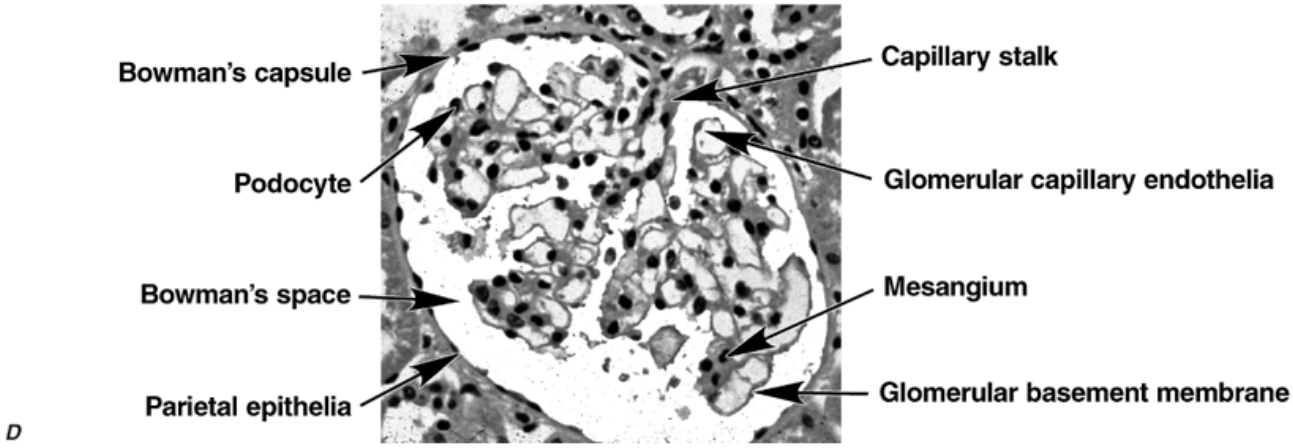
INTRODUCTION

Two human kidneys harbor nearly 1.8 million glomerular capillary tufts. Each glomerular tuft resides within Bowman's space. The capsule circumscribing this space is lined by parietal epithelial cells that transition into tubular epithelia forming the proximal nephron or migrate into the tuft to replenish podocytes. The glomerular capillary tuft derives from an afferent arteriole that forms a branching capillary bed embedded in mesangial matrix (Fig. 308-1). This capillary network funnels into an efferent arteriole, which passes filtered blood into cortical peritubular capillaries or medullary vasa recta that supply and exchange with a folded tubular architecture. Hence the glomerular capillary tuft, fed and drained by arterioles, represents an arteriolar portal system. Fenestrated endothelial cells resting on a glomerular basement membrane (GBM) line glomerular capillaries. Delicate foot processes extending from epithelial podocytes shroud the outer surface of these capillaries, and adjacent podocytes interconnect to each other by slit-pore membranes forming a selective filtration barrier.

FIGURE 308-1

Glomerular architecture. ***A.*** The glomerular capillaries form from a branching network of renal arteries, arterioles, leading to an afferent arteriole, glomerular capillary bed (tuft), and a draining efferent arteriole. *(From VH Gattone II et al: Hypertension 5:8, 1983.)* ***B.*** Scanning electron micrograph of podocytes that line the outer surface of the glomerular capillaries (*arrow* shows foot process). ***C.*** Scanning electron micrograph of the fenestrated endothelia lining the glomerular capillary. ***D.*** The various normal regions of the glomerulus on light microscopy. *(A–C: Courtesy of Dr. Vincent Gattone, Indiana University; with permission.)*





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The glomerular capillaries filter 120–180 L/d of plasma water containing various solutes for reclamation or discharge by downstream tubules. Most large proteins and all cells are excluded from filtration by a physicochemical barrier governed by pore size and negative electrostatic charge. The mechanics of filtration and reclamation are quite complicated for many solutes (Chap. 303). For example, in the case of serum albumin, the glomerulus is an imperfect barrier. Although albumin has a negative charge, which would tend to repel the negatively charged GBM, it only has a physical radius of 3.6 nm, while pores in the GBM and slit-pore membranes have a radius of 4 nm. Consequently, variable amounts of albumin inevitably cross the filtration barrier to be reclaimed by megalin and cubilin receptors along the proximal tubule. Remarkably, humans with normal nephrons excrete on average 8–10 mg of albumin in daily voided urine, ~20–60% of total excreted protein. This amount of albumin, and other proteins, can rise to gram quantities following glomerular injury.

The breadth of diseases affecting the glomerulus is expansive because the microenvironment supporting the glomerular capillaries can be injured in a variety of ways, producing many different lesions. Some order to this vast subject is brought by grouping all of these diseases into a smaller number of clinical syndromes.

PATHOGENESIS OF GLOMERULAR DISEASE

There are many forms of glomerular disease with pathogenesis variably linked to the presence of genetic mutations, infection, toxin exposure, autoimmunity, atherosclerosis, hypertension, emboli, thrombosis, or diabetes mellitus. Even after careful study, however, the cause often remains unknown, and the lesion is called *idiopathic*. Specific or unique features of pathogenesis are mentioned with the description of each of the glomerular diseases later in this chapter.


Some glomerular diseases result from genetic mutations producing familial disease or a founder effect: congenital nephrotic syndrome from mutations in *NPHS1* (nephrin) and *NPHS2* (podocin) affect the slit-pore membrane at birth, and *TRPC6* cation channel mutations produce *focal segmental glomerulosclerosis* (*FSGS*) in adulthood; polymorphisms in the gene encoding apolipoprotein L1, *APOL1*, are a major risk for nearly 70% of African Americans with nondiabetic end-stage renal disease (ESRD), particularly FSGS; mutations

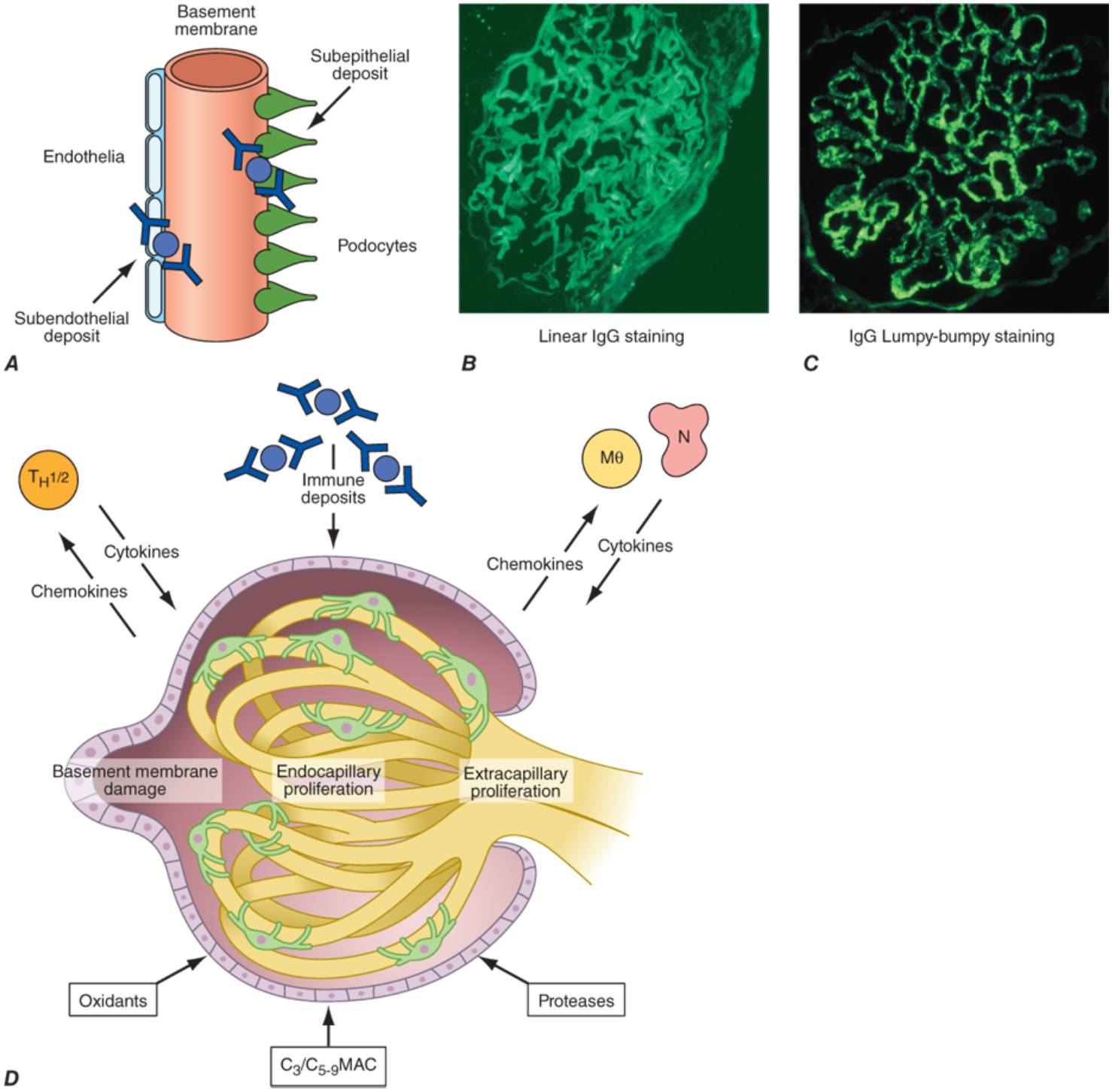
in complement factor H associate with *membranoproliferative glomerulonephritis* (*MPGN*), C_3 glomerulopathies, or *atypical hemolytic uremic syndrome* (*aHUS*), type II partial lipodystrophy from mutations in genes encoding lamin A/C, or PPARγ cause a metabolic syndrome associated with MPGN, or C_3 glomerulopathies, which is sometimes accompanied by dense deposits and C3 nephritic factor; Alport's syndrome, from mutations in the genes encoding for the α3, α4, or α5 chains of type IV collagen, produces *split-basement membranes* with *glomerulosclerosis*; and lysosomal storage diseases, such as α-galactosidase A deficiency causing Fabry's disease and *N* acetylneuraminic acid hydrolase deficiency causing nephrosialidosis, produce FSGS.

Systemic hypertension and atherosclerosis can produce pressure stress, ischemia, or lipid oxidants that lead to *chronic glomerulosclerosis. Malignant hypertension* can quickly complicate glomerulosclerosis with fibrinoid necrosis of arterioles and glomeruli, thrombotic microangiopathy, and acute renal failure. *Diabetic nephropathy* is an acquired sclerotic injury associated with thickening of the GBM secondary to the long-standing effects of hyperglycemia, advanced glycosylation end products, and reactive oxygen species.

Inflammation of the glomerular capillaries is called *glomerulonephritis*. Most glomerular or mesangial antigens involved in *immune-mediated glomerulonephritis* are unknown **(Fig. 308-2)**. Glomerular epithelial or mesangial cells may shed or express epitopes that mimic other immunogenic proteins made elsewhere in the body. Bacteria, fungi, and viruses can directly infect the kidney producing their own antigens. Autoimmune diseases like idiopathic *membranous glomerulonephritis* (*MGN*) or MPGN are confined to the kidney, whereas systemic inflammatory diseases like *lupus nephritis* or *granulomatosis with polyangiitis* spread to the kidney, causing secondary glomerular injury. *Antiglomerular basement membrane disease* producing Goodpasture's syndrome primarily injures both the lung and kidney because of the narrow distribution of the α3 NC1 domain of type IV collagen that is the target antigen.

FIGURE 308-2

The glomerulus is injured by a variety of mechanisms. ***A.*** Preformed immune deposits can precipitate from the circulation and collect along the glomerular basement membrane (GBM) in the subendothelial space or can form in situ along the subepithelial space. ***B.*** Immunofluorescent staining of glomeruli with labeled anti-IgG demonstrating linear staining from a patient with anti-GBM disease or immune deposits from a patient with membranous glomerulonephritis. ***C.*** The mechanisms of glomerular injury have a complicated pathogenesis. Immune deposits and complement deposition classically draw macrophages and neutrophils into the glomerulus. T lymphocytes may follow to participate in the injury pattern as well. ***D.*** Amplification mediators as locally derived oxidants and proteases expand this inflammation, and, depending on the location of the target antigen and the genetic polymorphisms of the host, basement membranes are damaged with either endocapillary or extracapillary proliferation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Local activation of Toll-like receptors on glomerular cells, deposition of immune complexes, or complement injury to glomerular structures induces mononuclear cell infiltration, which subsequently leads to an adaptive immune response attracted to the kidney by local release of chemokines. Neutrophils, macrophages, and T cells are drawn by chemokines into the glomerular tuft, where they react with antigens and epitopes on or near somatic cells or their structures, producing more cytokines and proteases that damage the mesangium, capillaries, and/or the GBM. While the adaptive immune response is similar to that of other tissues, early T cell activation plays an important role in the mechanism of glomerulonephritis. Antigens presented by class II major histocompatibility complex (MHC) molecules on macrophages and dendritic cells in conjunction with associative recognition molecules engage the CD4/8 T cell repertoire.

Mononuclear cells by themselves can injure the kidney, but autoimmune events that damage glomeruli classically produce a humoral immune response. *Poststreptococcal glomerulonephritis, lupus nephritis*, and idiopathic *membranous nephritis* typically are associated with immune deposits along the GBM, while anti-GBM antibodies produce the linear binding of anti-GBM disease. Preformed circulating

immune complexes can precipitate along the subendothelial side of the GBM, while other immune deposits form in situ on the subepithelial side. These latter deposits accumulate when circulating autoantibodies find their antigen trapped along the subepithelial edge of the GBM. Immune deposits in the glomerular mesangium may result from the deposition of preformed circulating complexes or in situ antigen-antibody interactions. Immune deposits stimulate the release of local proteases and activate the complement cascade, producing C_{5-9} attack complexes. In addition, local oxidants damage glomerular structures, producing proteinuria and effacement of the podocytes. Overlapping etiologies or pathophysiologic mechanisms can produce similar glomerular lesions, suggesting that downstream molecular and cellular responses often converge toward common patterns of injury.

PROGRESSION OF GLOMERULAR DISEASE

Persistent glomerulonephritis that worsens renal function is always accompanied by interstitial nephritis, renal fibrosis, and tubular atrophy **(see Fig. A3-27)**. What is not so obvious, however, is that renal failure in glomerulonephritis best correlates histologically with the appearance of tubulointerstitial nephritis rather than with the type of inciting glomerular injury.

Loss of renal function due to interstitial damage is explained hypothetically by several mechanisms. The simplest explanation is that urine flow is impeded by tubular obstruction as a result of interstitial inflammation and fibrosis. Thus, obstruction of the tubules with debris or by extrinsic compression functionally results in aglomerular nephrons. A second mechanism suggests that interstitial changes, including interstitial edema or fibrosis, alter tubular and vascular architecture and thereby compromise the normal tubular transport of solutes and water from tubular lumen to vascular space. This failure increases the solute and water content of the tubule fluid, resulting in isosthenuria and polyuria. Adaptive mechanisms related to tubuloglomerular feedback also fail, resulting in a reduction of renin output from the juxtaglomerular apparatus trapped by interstitial inflammation. Consequently, the local vasoconstrictive influence of angiotensin II on the glomerular arterioles decreases, and filtration drops owing to a generalized decrease in arteriolar tone. A third mechanism involves changes in vascular resistance due to damage of peritubular capillaries. The cross-sectional volume of these capillaries is decreased by interstitial inflammation, edema, or fibrosis. These structural alterations in vascular resistance affect renal function through two mechanisms. First, tubular cells are very metabolically active, and, as a result, decreased perfusion leads to tubular ischemic injury. Second, impairment of glomerular arteriolar outflow leads to increased intravascular hypertension in less-involved glomeruli; this selective intraglomerular hypertension aggravates and extends *mesangial sclerosis* and *glomerulosclerosis* to less-involved glomeruli. Regardless of the exact mechanism, early *acute tubulointerstitial nephritis* **(see Fig. A3-27)** suggests potentially recoverable renal function, whereas the development of *chronic interstitial fibrosis* prognosticates permanent loss **(see Fig. A3-30)**.

Persistent damage to glomerular capillaries spreads to the tubulointerstitium in association with proteinuria. There is a hypothesis that efferent arterioles leading from inflamed glomeruli carry forward inflammatory mediators, which induces downstream interstitial nephritis, resulting in fibrosis. Glomerular filtrate from injured glomerular capillaries adherent to Bowman's capsule may also be misdirected to the periglomerular interstitium. Most nephrologists believe, however, that proteinuric glomerular filtrate forming tubular fluid is the primary route to downstream tubulointerstitial injury, although none of these hypotheses are mutually exclusive.

The simplest explanation for the effect of proteinuria on the development of interstitial nephritis is that increasingly severe proteinuria, carrying activated cytokines and lipoproteins producing reactive oxygen species, triggers a downstream inflammatory cascade in and around epithelial cells lining the tubular nephron. These effects induce T lymphocyte and macrophage infiltrates in the interstitial spaces along with fibrosis and tubular atrophy.

Tubules disaggregate following direct damage to their basement membranes, leading to more interstitial fibroblasts and fibrosis at the site of injury; recent comprehensive evidence suggests that renal fibroblasts increase through several mechanisms: epithelial or endothelial-mesenchymal transitions (15%), bone marrow-derived fibrocytes (35%), and the proliferation of resident fibroblasts (50%). Transforming growth factor-β (TGF-β), fibroblast growth factor 2 (FGF-2), hypoxemia-inducible factor 1α (HIF-1α), and platelet-derived growth factor (PDGF) are particularly active in this transition. With persistent nephritis, fibroblasts multiply and lay down tenascin and a fibronectin scaffold for the polymerization of new interstitial collagen types I/III. These events form scar tissue through a process called fibrogenesis. In experimental studies, bone morphogenetic protein 7 and hepatocyte growth factor can reverse early fibrogenesis and preserve tubular architecture. When fibroblasts outdistance their survival factors, apoptosis occurs, and the permanent renal scar becomes acellular, leading to irreversible renal failure.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Glomerular Disease

HEMATURIA, PROTEINURIA, AND PYURIA

Patients with glomerular disease usually have some hematuria with varying degrees of proteinuria. Hematuria is typically asymptomatic. As few as 3–5 red blood cells in the spun sediment from first-voided morning urine is suspicious. The diagnosis of glomerular injury can be delayed because patients will not realize they have *microscopic hematuria*, and only rarely with the exception of IgA nephropathy and sickle cell disease is *gross hematuria* present. When working up microscopic hematuria, perhaps accompanied by minimal proteinuria (<500 mg/24 h), it is important to exclude anatomic lesions, such as malignancy of the urinary tract, particularly in older men. Microscopic hematuria may also appear with the onset of benign prostatic hypertrophy, interstitial nephritis, papillary necrosis, hypercalciuria, renal stones, cystic kidney diseases, or renal vascular injury. However, when red blood cell casts **(see Fig. A3-34)** or dysmorphic red blood cells are found in the sediment, glomerulonephritis is likely. A mean of 8–10 mg/24 h of albumin appears in the urine in the absence of kidney disease. In early nephropathy, such as in diabetic nephropathy, proteinuria increases to 30–300 mg/24 h and is called microalbuminuria and represents the presence of renal disease. Greater than 300 mg/24 h of albuminuria represents frank proteinuria and more advanced renal disease **(Table 308-1)**.

Sustained proteinuria >1–2 g/24 h is also commonly associated with glomerular disease. Patients often will not know they have proteinuria unless they become edematous or notice foaming urine on voiding. *Sustained proteinuria* has to be distinguished from lesser amounts of so-called *benign proteinuria* in the normal population. (Table 308-1). This latter class of proteinuria is nonsustained, generally <1 g/24 h, and is sometimes called *functional* or *transient proteinuria*. Fever, exercise, obesity, sleep apnea, emotional stress, and congestive heart failure can explain transient proteinuria. Proteinuria only seen with upright posture is called *orthostatic proteinuria* and has a benign prognosis. Isolated proteinuria sustained over multiple clinic visits is found in many glomerular lesions. Proteinuria in most adults with glomerular disease is *nonselective*, containing albumin and a mixture of other serum proteins, whereas in children with *minimal change disease* (MCD), the proteinuria is *selective* and composed largely of albumin.

Some patients with inflammatory glomerular disease, such as acute poststreptococcal glomerulonephritis or MPGN, have *pyuria* characterized by the presence of considerable numbers of leukocytes. This latter finding has to be distinguished from urine infected with bacteria.

CLINICAL SYNDROMES

Various forms of glomerular injury can also be parsed into several distinct syndromes on clinical grounds **(Table 308-2)**. These syndromes, however, are not always mutually exclusive. There is an *acute nephritic syndrome* producing 1–2 g/24 h of proteinuria, hematuria with red blood cell casts, pyuria, hypertension, fluid retention, and a rise in serum creatinine associated with a reduction in glomerular filtration. If glomerular inflammation develops slowly, the serum creatinine will rise gradually over many weeks, but if the serum creatinine rises quickly, particularly over a few days, acute nephritis is sometimes called *rapidly progressive glomerulonephritis* (RPGN); the histopathologic term *crescentic glomerulonephritis* is the pathologic equivalent of the clinical presentation of RPGN. When patients with RPGN present with lung hemorrhage from Goodpasture's syndrome, antineutrophil cytoplasmic antibodies (ANCA)-associated small-vessel vasculitis, lupus erythematosus, or cryoglobulinemia, they are often diagnosed as having a *pulmonary-renal syndrome. Nephrotic syndrome* describes the onset of heavy proteinuria (>3.0 g/24 h), hypertension, hypercholesterolemia, hypoalbuminemia, edema/anasarca, and microscopic hematuria; if only large amounts of proteinuria are present without clinical manifestations, the condition is sometimes called *nephrotic-range proteinuria*. The glomerular filtration rate (GFR) in these patients may initially be normal or, rarely, higher than normal, but with persistent hyperfiltration and continued nephron loss, it typically declines over months to years. Patients with a *basement membrane syndrome* either have genetically abnormal basement membranes (Alport's syndrome) or an autoimmune response to basement membrane collagen IV (Goodpasture's syndrome) associated with microscopic hematuria, mild to heavy proteinuria, and hypertension with variable elevations in serum creatinine. *Glomerular–vascular syndrome* describes patients with vascular injury producing hematuria and moderate proteinuria. Affected individuals can have vasculitis, thrombotic microangiopathy, antiphospholipid syndrome, or, more commonly, a systemic disease such as atherosclerosis, cholesterol emboli, hypertension, sickle cell anemia, and autoimmunity. *Infectious disease–associated syndrome* is most important if one has a global perspective. Save for subacute bacterial endocarditis (SBE) in the Western Hemisphere, malaria, and schistosomiasis may be the most common causes of glomerulonephritis throughout the world, closely followed by HIV and chronic hepatitis B and C. These infectious diseases produce a variety of inflammatory reactions in glomerular capillaries, ranging from nephrotic syndrome to acute nephritic injury, and urinalyses that demonstrate a combination of hematuria and proteinuria.

These six general categories of syndromes are usually determined at the bedside with the help of a history and physical examination, blood chemistries, renal ultrasound, and urinalysis. These initial studies help frame further diagnostic workup that typically involves testing of the serum for the presence of various proteins (HIV and hepatitis B and C antigens), antibodies (anti-GBM, antiphospholipid, antistreptolysin O [ASO], anti-DNAse, antihyaluronidase, ANCA, anti-DNA, cryoglobulins, anti-HIV, and anti-hepatitis B and C antibodies) or depletion of complement components (C_3 and C_4). The bedside history and physical examination can also help determine whether the glomerulonephritis is isolated to the kidney (*primary glomerulonephritis*) or is part of a systemic disease (*secondary glomerulonephritis*).

When confronted with an abnormal urinalysis and elevated serum creatinine, with or without edema or congestive heart failure, one must consider whether the glomerulonephritis is *acute* or *chronic*. This assessment is best made by careful history (last known urinalysis or serum creatinine during pregnancy or insurance physical, evidence of infection, or use of medication or recreational drugs); the size of the kidneys on renal ultrasound examination; and how the patient feels at presentation. Chronic glomerular disease often presents with decreased kidney size. Patients who quickly develop renal failure are fatigued and weak and often have uremic symptoms associated with nausea, vomiting, fluid retention, and somnolence. Primary glomerulonephritis presenting with renal failure that has progressed slowly, however, can be remarkably asymptomatic, as are patients with acute glomerulonephritis without much loss in renal function. Once this initial information is collected, selected patients who are clinically stable, have adequate blood clotting parameters, and are willing and able to receive treatment are encouraged to have a renal biopsy.

TABLE 308-1

Urine Assays for Albuminuria/Proteinuria

	24-h Albumin[a] (mg/24 h)	Albumin[a]/Creatinine Ratio (mg/g)	Dipstick Proteinuria	24-h Urine Protein[b] (mg/24 h)
Normal	8–10	<30	–	<150
Microalbuminuria	30–300	30–300	–/Trace/1+	–
Proteinuria	>300	>300	Trace–3+	>150

[a]Albumin detected by radioimmunoassay. [b]Albumin represents 20–60% of the total protein excreted in the urine.

TABLE 308-2

Patterns of Clinical Glomerulonephritis

Glomerular Syndromes	Proteinuria	Hematuria	Vascular Injury
Acute Nephritic Syndromes			
Poststreptococcal glomerulonephritis[a]	+/++	++/+++	–
Subacute bacterial endocarditis[a]	+/++	++	–
Lupus nephritis[a]	+/++	++/+++	+
Antiglomerular basement membrane disease[a]	++	++/+++	–
IgA nephropathy[a]	+/++	+++[c]	–
ANCA small-vessel vasculitis[a]			
Granulomatosis with polyangiitis (Wegener's)	+/++	++/+++	++++
Microscopic polyangiitis	+/++	++/+++	++++
Churg-Strauss syndrome	+/++	++/+++	++++
Henoch-Schönlein purpura[a]	+/++	++/+++	++++
Cryoglobulinemia[a]	+/++	++/+++	++++
Membranoproliferative glomerulonephritis[a]	++	++/+++	–
C_3 Glomerulopathies	++	++/+++	-
Mesangioproliferative glomerulonephritis	+	+/++	–
Pulmonary-Renal Syndromes			
Goodpasture's syndrome[a]	++	++/+++	–
ANCA small-vessel vasculitis[a]			
Granulomatosis with polyangiitis (Wegener's)	+/++	++/+++	++++
Microscopic polyangiitis	+/++	++/+++	++++
Churg-Strauss syndrome	+/++	++/+++	++++
Henoch-Schönlein purpura[a]	+/++	++/+++	++++
Cryoglobulinemia[a]	+/++	++/+++	++++

Glomerular Syndromes	Proteinuria	Hematuria	Vascular Injury
Nephrotic Syndromes			
Minimal change disease	++++	–	–
Focal segmental glomerulosclerosis	+++/++++	+	–
Membranous glomerulonephritis	++++	+	–
Diabetic nephropathy	++/++++	–/+	–
AL and AA amyloidosis	+++/++++	+	+/++
Light-chain deposition disease	+++	+	–
Fibrillary-immunotactoid disease	+++/++++	+	+
Fabry's disease	+	+	–
Basement Membrane Syndromes			
Anti-GBM disease[a]	++	++/+++	–
Alport's syndrome	++	++	–
Thin basement membrane disease	+	++	–
Nail-patella syndrome	++/+++	++	–
Glomerular Vascular Syndromes			
Atherosclerotic nephropathy	+	+	+++
Hypertensive nephropathy[b]	+/++	+/++	++
Cholesterol emboli	+/++	++	+++
Sickle cell disease	+/++	+++[c]	+++
Thrombotic microangiopathies	++	++	+++
Antiphospholipid syndrome	++	++	+++
ANCA small-vessel vasculitis[a]			
Granulomatosis with polyangiitis (Wegener's)	+/++	++/+++	++++
Microscopic polyangiitis	+/++	++/+++	++++
Churg-Strauss syndrome	+++	++/+++	++++
Henoch-Schönlein purpura[a]	+/++	++/+++	++++

Glomerular Syndromes	Proteinuria	Hematuria	Vascular Injury
Cryoglobulinemia[a]	+/++	++/+++	++++
AL and AA amyloidosis	+++/++++	+	+/++
Infectious Disease–Associated Syndromes			
Poststreptococcal glomerulonephritis[a]	+/++	++/+++	–
Subacute bacterial endocarditis[a]	+/++	++	–
HIV	+++	+/++	–
Hepatitis B and C	+++	+/++	–
Syphilis	+++	+	–
Leprosy	+++	+	–
Malaria	+++	+/++	–
Schistosomiasis	+++	+/++	–

[a]Can present as rapidly progressive glomerulonephritis (RPGN); sometimes called crescentic glomerulonephritis. [b]Can present as a malignant hypertensive crisis producing an aggressive fibrinoid necrosis in arterioles and small arteries with microangiopathic hemolytic anemia. [c]Can present with gross hematuria.

Abbreviations: AA, amyloid A; AL, amyloid L; ANCA, antineutrophil cytoplasmic antibodies; GBM, glomerular basement membrane.

RENAL PATHOLOGY

A renal biopsy in the setting of glomerulonephritis quickly identifies the type of glomerular injury and often suggests a course of treatment. The biopsy is processed for light microscopy using stains for *hematoxylin and eosin* (*H&E*) to assess cellularity and architecture, *periodic acid–Schiff* (*PAS*) to stain carbohydrate moieties in the membranes of the glomerular tuft and tubules, *Jones-methenamine silver* to enhance basement membrane structure, *Congo red* for amyloid deposits, and *Masson's trichrome* to identify collagen deposition and assess the degree of glomerulosclerosis and interstitial fibrosis. Biopsies are also processed for direct immunofluorescence using conjugated antibodies against IgG, IgM, and IgA to detect the presence of "lumpy-bumpy" immune deposits or "linear" IgG or IgA antibodies bound to GBM, antibodies against trapped complement proteins (C_3 and C_4), or specific antibodies against a relevant antigen. High-resolution electron microscopy can clarify the principal location of immune deposits and the status of the basement membrane.

Each region of a renal biopsy is assessed separately. By light microscopy, glomeruli (ideally 20) are reviewed individually for discrete lesions; <50% involvement is considered *focal*, and >50% is *diffuse*. Injury in each glomerular tuft can be *segmental,* involving a portion of the tuft, or *global,* involving most of the glomerulus. Glomeruli having *proliferative* characteristics show increased cellularity. When cells in the capillary tuft proliferate, it is called *endocapillary*, and when cellular proliferation extends into Bowman's space, it is called *extracapillary. Synechiae* are formed when epithelial podocytes attach to Bowman's capsule in the setting of glomerular injury; *crescents*, which in some cases may be the extension of synechiae, develop when fibrocellular/fibrin collections fill all or part of Bowman's space; and *sclerotic* glomeruli show acellular, amorphous accumulations of proteinaceous material throughout the tuft with loss of functional capillaries and normal mesangium. Since *age-related glomerulosclerosis* is common in adults, one can estimate the background percentage of sclerosis by dividing the patient's age in half and subtracting 10. Immunofluorescent and electron microscopy can detect the presence and location of *subepithelial, subendothelial, or mesangial* immune deposits, or *reduplication* or *splitting* of the basement

membrane. In the other regions of the biopsy, the vasculature surrounding glomeruli and tubules can show *angiopathy, vasculitis*, the presence of *fibrils*, or *thrombi*. The tubules can be assessed for adjacency to one another; separation can be the result of edema, tubular dropout, or collagen deposition resulting from interstitial fibrosis. Interstitial fibrosis is an ominous sign of irreversibility and progression to renal failure.

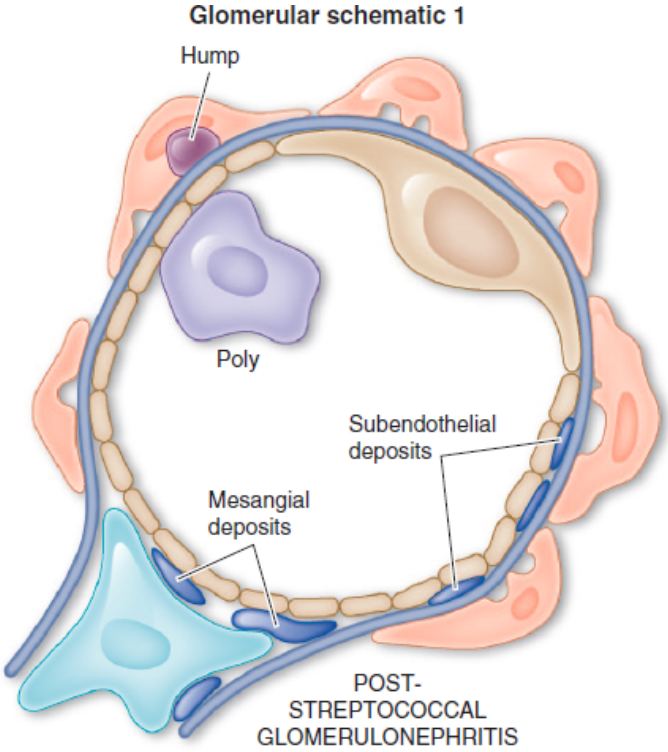
ACUTE NEPHRITIC SYNDROMES

Acute nephritic syndromes classically present with hypertension, hematuria, red blood cell casts, pyuria, and mild to moderate proteinuria. Extensive inflammatory damage to glomeruli causes a fall in GFR and eventually produces uremic symptoms with salt and water retention, leading to edema and hypertension.

POSTSTREPTOCOCCAL GLOMERULONEPHRITIS

Poststreptococcal glomerulonephritis is prototypical for *acute endocapillary proliferative glomerulonephritis*. The incidence of poststreptococcal glomerulonephritis has dramatically decreased in developed countries and in these locations is typically sporadic. Acute poststreptococcal glomerulonephritis in underdeveloped countries is epidemic and usually affects children between the ages of 2 and 14 years, but in developed countries is more typical in the elderly, especially in association with debilitating conditions. It is more common in males, and the familial or cohabitant incidence is as high as 40%. Skin and throat infections with particular M types of streptococci (nephritogenic strains) antedate glomerular disease; M types 47, 49, 55, 2, 60, and 57 are seen following impetigo and M types 1, 2, 4, 3, 25, 49, and 12 with pharyngitis. Poststreptococcal glomerulonephritis due to impetigo develops 2–6 weeks after skin infection and 1–3 weeks after streptococcal pharyngitis.

The renal biopsy in poststreptococcal glomerulonephritis demonstrates hypercellularity of mesangial and endothelial cells, glomerular infiltrates of polymorphonuclear leukocytes, granular subendothelial immune deposits of IgG, IgM, C_3, C_4, and C_{5-9}, and subepithelial deposits (which appear as “humps”) **(see Fig. A3-6)**. **(See Glomerular Schematic 1.)** Poststreptococcal glomerulonephritis is an immune-mediated disease involving putative streptococcal antigens, circulating immune complexes, and activation of complement in association with cell-mediated injury. Many candidate antigens have been proposed over the years; candidates from nephritogenic streptococci of interest at the moment are: a cationic cysteine proteinase known as streptococcal pyrogenic exotoxin B (SPEB) that is generated by proteolysis of a zymogen precursor (zSPEB), and NAPlr, the nephritis-associated plasmin receptor. These two antigens have biochemical affinity for plasmin, bind as complexes facilitated by this relationship, and activate the alternate complement pathway. The nephritogenic antigen, SPEB, has been demonstrated inside the subepithelial “humps” on biopsy.



The classic presentation is an acute nephritic picture with hematuria, pyuria, red blood cell casts, edema, hypertension, and oliguric renal failure, which may be severe enough to appear as RPGN. Systemic symptoms of headache, malaise, anorexia, and flank pain (due to swelling of the renal capsule) are reported in as many as 50% of cases. Five percent of children and 20% of adults have proteinuria in the nephrotic range. In the first week of symptoms, 90% of patients will have a depressed CH_{50} and decreased levels of C_3 with normal levels of C_4. Positive rheumatoid factor (30–40%), cryoglobulins and circulating immune complexes (60–70%), and ANCA against myeloperoxidase (10%) are also reported. Positive cultures for streptococcal infection are inconsistently present (10–70%), but increased titers of ASO (30%), anti-DNAse (70%), or antihyaluronidase antibodies (40%) can help confirm the diagnosis. Consequently, the diagnosis of poststreptococcal glomerulonephritis rarely requires a renal biopsy. A subclinical disease is reported in some series to be 4–5 times as common as clinical nephritis, and these latter cases are characterized by asymptomatic microscopic hematuria with low serum C_3 complement levels.

Treatment is supportive, with control of hypertension, edema, and dialysis as needed. Antibiotic treatment for streptococcal infection should be given to all patients and their cohabitants. There is no role for immunosuppressive therapy, even in the setting of crescents. Recurrent poststreptococcal glomerulonephritis is rare despite repeated streptococcal infections. Early death is rare in children but does occur in the elderly. Overall, the prognosis is good, with permanent renal failure being reported as very uncommon in the past (<1%) but with recent reports of an increased risk of chronic kidney disease in adulthood. Complete resolution of the hematuria and proteinuria in the majority of children occurs within 3–6 weeks of the onset of nephritis but 3–10% of children may have persistent microscopic hematuria, nonnephrotic proteinuria, or hypertension. The prognosis in elderly patients is worse with a high incidence of azotemia (up to 60%), nephrotic-range proteinuria, and ESRD.

SUBACUTE BACTERIAL ENDOCARDITIS

Endocarditis-associated glomerulonephritis is typically a complication of SBE, particularly in patients who remain untreated for a long time, have negative blood cultures, or have right-sided endocarditis. Common comorbidities are valvular heart disease, intravenous drug use, hepatitis C, and diabetes mellitus. Glomerulonephritis is unusual in acute bacterial endocarditis because it takes 10–14 days to develop immune complex–mediated injury, by which time the patient has been treated, often with emergent surgery. Grossly, the kidneys in SBE have subcapsular hemorrhages with a “flea-bitten” appearance, and microscopy on renal biopsy reveals focal proliferation around foci of necrosis associated with abundant mesangial, subendothelial, and subepithelial immune deposits of IgG, IgM, and C_3. Commonly patients present with a clinical picture of RPGN and have crescents on biopsy. Embolic infarcts or septic abscesses may also be present. The pathogenesis hinges on the renal deposition of circulating immune complexes in the kidney with complement activation. Patients present with gross or microscopic hematuria, pyuria, and mild proteinuria, acute kidney injury or, RPGN with rapid loss of renal function. A normocytic anemia, elevated erythrocyte sedimentation rate, hypocomplementemia, high titers of rheumatoid factor, type III cryoglobulins, circulating immune complexes, and ANCAs may be present. Levels of serum creatinine may be elevated at diagnosis, but with modern therapy there is little progression to chronic renal failure. Primary treatment is eradication of the infection with 4–6 weeks of antibiotics, and if accomplished expeditiously, the prognosis for renal recovery is good. ANCA-associated vasculitis sometimes accompanies or is confused with SBE and should be ruled out, as the treatment is different.

As variants of persistent bacterial infection in blood-associated glomerulonephritis, postinfectious glomerulonephritis can occur in patients with ventriculoatrial and ventriculoperitoneal shunts; pulmonary, intraabdominal, pelvic, or cutaneous infections; and infected vascular prostheses. In developed countries, a significant proportion of cases afflict adults, especially the immunocompromised, and the predominant organism is *Staphylococcus*. The clinical presentation of these conditions is variable and includes proteinuria, microscopic hematuria, acute renal failure, and hypertension. Serum complement levels are low, and there may be elevated levels of C-reactive proteins, rheumatoid factor, antinuclear antibodies, and cryoglobulins. Renal lesions include MPGN, diffuse proliferative and exudative glomerulonephritis (DPGN), or mesangioproliferative glomerulonephritis, sometimes leading to RPGN. Treatment focuses on eradicating the infection, with most patients treated as if they have endocarditis. The prognosis is guarded.

LUPUS NEPHRITIS

Lupus nephritis is a common and serious complication of systemic lupus erythematosus (SLE) and most severe in African-American female adolescents. Thirty to 50% of patients will have clinical manifestations of renal disease at the time of diagnosis, and 60% of adults and 80% of children develop renal abnormalities at some point in the course of their disease. Lupus nephritis results from the deposition of circulating immune complexes, which activate the complement cascade leading to complement-mediated damage, leukocyte infiltration, activation of procoagulant factors, and release of various cytokines. In situ immune complex formation following glomerular binding of

nuclear antigens, particularly necrotic nucleosomes, also plays a role in renal injury. The presence of antiphospholipid antibodies may also trigger a thrombotic microangiopathy in a minority of patients.

The clinical manifestations, course of disease, and treatment of lupus nephritis are closely linked to renal pathology. The most common clinical sign of renal disease is proteinuria, but hematuria, hypertension, varying degrees of renal failure, and active urine sediment with red blood cell casts can all be present. Although significant renal pathology can be found on biopsy even in the absence of major abnormalities in the urinalysis, most nephrologists do not biopsy patients until the urinalysis is convincingly abnormal. The extrarenal manifestations of lupus are important in establishing a firm diagnosis of systemic lupus because, while serologic abnormalities are common in lupus nephritis, they are not diagnostic. Anti-dsDNA antibodies that fix complement correlate best with the presence of renal disease. Hypocomplementemia is common in patients with acute lupus nephritis (70–90%) and declining complement levels may herald a flare. Although urinary biomarkers of lupus nephritis are being identified to assist in predicting renal flares, renal biopsy is the only reliable method of identifying the morphologic variants of lupus nephritis.

The World Health Organization (WHO) workshop in 1974 first outlined several distinct patterns of lupus-related glomerular injury; these were modified in 1982. In 2004 the International Society of Nephrology in conjunction with the Renal Pathology Society again updated the classification. This latest version of lesions seen on biopsy **(Table 308-3)** forms the basis for modern treatment recommendations. Class I nephritis describes normal glomerular histology by any technique or normal light microscopy with minimal mesangial deposits on immunofluorescent or electron microscopy. Class II designates mesangial immune complexes with *mesangial proliferation*. Both class I and II lesions are typically associated with minimal renal manifestation and normal renal function; nephrotic syndrome is rare. Patients with lesions limited to the renal mesangium have an excellent prognosis and generally do not need therapy for their lupus nephritis.

TABLE 308-3

Classification for Lupus Nephritis

Class I	Minimal mesangial	Normal histology with mesangial deposits
Class II	Mesangial proliferation	Mesangial hypercellularity with expansion of the mesangial matrix
Class III	Focal nephritis	Focal endocapillary ± extracapillary proliferation with focal subendothelial immune deposits and mild mesangial expansion
Class IV	Diffuse nephritis	Diffuse endocapillary ± extracapillary proliferation with diffuse subendothelial immune deposits and mesangial alterations
Class V	Membranous nephritis	Thickened basement membranes with diffuse subepithelial immune deposits; may occur with class III or IV lesions and is sometimes called mixed membranous and proliferative nephritis
Class VI	Sclerotic nephritis	Global sclerosis of nearly all glomerular capillaries

Note: Revised in 2004 by the International Society of Nephrology-Renal Pathology Society Study Group.

The subject of lupus nephritis is presented under acute nephritic syndromes because of the aggressive and important proliferative lesions seen in class III–V renal diseases. Class III describes *focal lesions with proliferation or scarring*, often involving only a segment of the glomerulus **(see Fig. A3-12)**. Class III lesions have the most varied course. Hypertension, an active urinary sediment, and proteinuria are common with nephrotic-range proteinuria in 25–33% of patients. Elevated serum creatinine is present in 25% of patients. Patients with mild proliferation involving a small percentage of glomeruli respond well to therapy with steroids alone, and fewer than 5% progress to renal failure over 5 years. Patients with more severe proliferation involving a greater percentage of glomeruli have a far worse prognosis and lower remission rates. Treatment of those patients is the same as that for class IV lesions. Many nephrologists believe that class III lesions are simply an early presentation of class IV disease. Others believe severe class III disease is a discrete vasculitic lesion requiring

aggressive therapy. Class IV describes *global, diffuse proliferative lesions* involving the vast majority of glomeruli. Patients with class IV lesions commonly have high anti-DNA antibody titers, low serum complement, hematuria, red blood cell casts, proteinuria, hypertension, and decreased renal function; 50% of patients have nephrotic-range proteinuria. Patients with crescents on biopsy often have a rapidly progressive decline in renal function **(see Fig. A3-12)**. Without treatment, this aggressive lesion has the worst renal prognosis. However, if a remission—defined as a return to near-normal renal function and proteinuria ≤330 mg/dL per day—is achieved with treatment, renal outcomes are excellent. Current evidence suggests that inducing a remission with administration of high-dose steroids and either cyclophosphamide or mycophenolate mofetil for 2–6 months, followed by maintenance therapy with lower doses of steroids and mycophenolate mofetil or azathioprine, best balances the likelihood of successful remission with the side effects of therapy. There is no consensus on use of high-dose intravenous methylprednisolone versus oral prednisone, monthly intravenous cyclophosphamide versus daily oral cyclophosphamide, or other immunosuppressants such as cyclosporine, tacrolimus, rituximab, or belimumab. Nephrologists tend to avoid prolonged use of cyclophosphamide in patients of childbearing age without first banking eggs or sperm.

The class V lesion describes subepithelial immune deposits producing a *membranous pattern*; a subcategory of class V lesions is associated with proliferative lesions and is sometimes called *mixed membranous and proliferative disease* **(see Fig. A3-11)**; this category of injury is treated like class IV glomerulonephritis. Sixty percent of patients present with nephrotic syndrome or lesser amounts of proteinuria. Patients with lupus nephritis class V, like patients with *idiopathic membranous nephropathy* (IMN), are predisposed to renal-vein thrombosis and other thrombotic complications. A minority of patients with class V will present with hypertension and renal dysfunction. There are conflicting data on the clinical course, prognosis, and appropriate therapy for patients with class V disease, which may reflect the heterogeneity of this group of patients. Patients with severe nephrotic syndrome, elevated serum creatinine, and a progressive course will probably benefit from therapy with steroids in combination with other immunosuppressive agents. Therapy with inhibitors of the renin-angiotensin system also may attenuate the proteinuria. Antiphospholipid antibodies present in lupus may result in glomerular microthromboses and complicate the course in up to 20% of lupus nephritis patients. The renal prognosis is worse despite anticoagulant therapy.

Patients with any of the above lesions also can transform to another lesion; hence patients often require reevaluation, including repeat renal biopsy. Lupus patients with class VI lesions have >90% *sclerotic glomeruli* and ESRD with interstitial fibrosis. As a group, ~20% of patients with lupus nephritis will reach end-stage disease, requiring dialysis or transplantation. Patients with lupus nephritis have a markedly increased mortality compared with the general population. Renal transplantation in renal failure from lupus, usually performed after ~6 months of inactive disease, results in allograft survival rates comparable to patients transplanted for other reasons.

ANTIGLOMERULAR BASEMENT MEMBRANE DISEASE

Patients who develop autoantibodies directed against glomerular basement antigens frequently develop a glomerulonephritis termed *antiglomerular basement membrane* (*anti-GBM*) *disease*. When they present with lung hemorrhage and glomerulonephritis, they have a pulmonary-renal syndrome called *Goodpasture's syndrome*. The target epitopes for this autoimmune disease lie in the quaternary structure of α3 NC1 domain of collagen IV. Indeed, anti-GBM disease may be considered an autoimmune "conformeropathy" that involves the perturbation of quaternary structure of the α 345NC1 hexamer. MHC-restricted T cells initiate the autoantibody response because humans are not tolerant to the epitopes created by this quaternary structure. The epitopes are normally sequestered in the collagen IV hexamer and can be exposed by infection, smoking, oxidants, or solvents. Goodpasture's syndrome appears in two age groups: in young men in their late twenties and in men and women in their sixties and seventies. Disease in the younger age group is usually explosive, with hemoptysis, a sudden fall in hemoglobin, fever, dyspnea, and hematuria. Hemoptysis is largely confined to smokers, and those who present with lung hemorrhage as a group do better than older populations who have prolonged, asymptomatic renal injury; presentation with oliguria is often associated with a particularly bad outcome. The performance of an urgent kidney biopsy is important in suspected cases of Goodpasture's syndrome to confirm the diagnosis and assess prognosis. Renal biopsies typically show *focal or segmental necrosis* that later, with aggressive destruction of the capillaries by cellular proliferation, leads to crescent formation in Bowman's space **(see Fig. A3-14)**. As these lesions progress, there is concomitant interstitial nephritis with fibrosis and tubular atrophy.

The presence of anti-GBM antibodies and complement is recognized on biopsy by linear immunofluorescent staining for IgG (rarely IgA). In testing serum for anti-GBM antibodies, it is particularly important that the α3 NC1 domain of collagen IV alone be used as the target. This is because nonnephritic antibodies against the α1 NC1 domain are seen in paraneoplastic syndromes and cannot be discerned from assays that use whole basement membrane fragments as the binding target. Between 10 and 15% of sera from patients with Goodpasture's syndrome also contain ANCA antibodies against myeloperoxidase. This subset of patients has a vasculitis-associated variant, which has a surprisingly good prognosis with treatment. Prognosis at presentation is worse if there are >50% crescents on renal biopsy with advanced

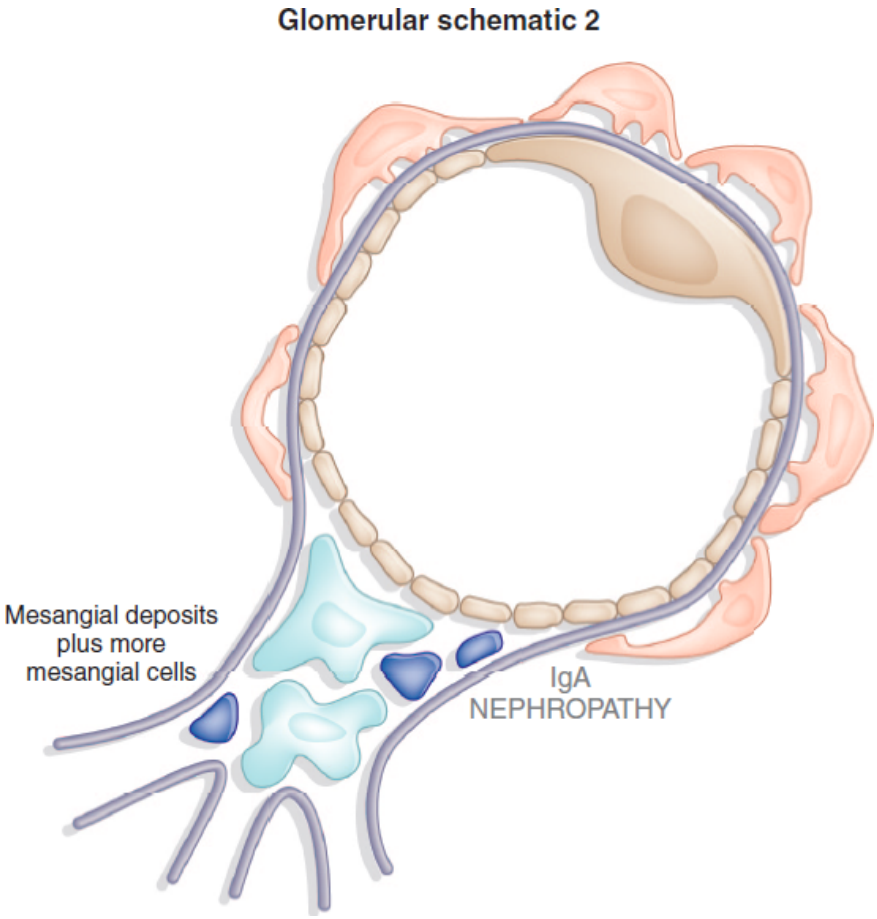
fibrosis, if serum creatinine is >5–6 mg/dL, if oliguria is present, or if there is a need for acute dialysis. Although frequently attempted, most of these latter patients will not respond to plasmapheresis and steroids. Patients with advanced renal failure who present with hemoptysis should still be treated for their lung hemorrhage, as it responds to plasmapheresis and can be lifesaving. Treated patients with less severe disease typically respond to 8–10 treatments of plasmapheresis accompanied by oral prednisone and cyclophosphamide in the first 2 weeks. Kidney transplantation is possible, but because there is risk of recurrence, experience suggests that patients should wait for 6 months and until serum antibodies are undetectable.

IgA NEPHROPATHY

Berger first described the glomerulonephritis now termed *IgA nephropathy*. It is classically characterized by episodic hematuria associated with the deposition of IgA in the mesangium. IgA nephropathy is one of the most common forms of glomerulonephritis worldwide. There is a male preponderance, a peak incidence in the second and third decades of life, and rare familial clustering. There are geographic differences in the prevalence of IgA nephropathy, with 30% prevalence along the Asian and Pacific Rim and 20% in southern Europe, compared to a much lower prevalence in northern Europe and North America. It was initially hypothesized that variation in detection, in part, accounted for regional differences. With clinical care in nephrology becoming more uniform, this variation in prevalence more likely reflects true differences among racial and ethnic groups.

IgA nephropathy is predominantly a sporadic disease but susceptibility to it has been shown uncommonly to have a genetic component depending on geography and the existence of “founder effects.” Familial forms of IgA nephropathy are more common in northern Italy and eastern Kentucky. No single causal gene has been identified. Clinical and laboratory evidence suggests close similarities between Henoch-Schönlein purpura and IgA nephropathy. Henoch-Schönlein purpura is distinguished clinically from IgA nephropathy by prominent systemic symptoms, a younger age (<20 years old), preceding infection, and abdominal complaints. Deposits of IgA are also found in the glomerular mesangium in a variety of systemic diseases, including chronic liver disease, Crohn’s disease, gastrointestinal adenocarcinoma, chronic bronchiectasis, idiopathic interstitial pneumonia, dermatitis herpetiformis, mycosis fungoides, leprosy, ankylosing spondylitis, relapsing polychondritis, and Sjögren’s syndrome. IgA deposition in these entities is not usually associated with clinically significant glomerular inflammation or renal dysfunction and thus is not called IgA nephropathy.

IgA nephropathy is an immune complex–mediated glomerulonephritis defined by the presence of diffuse mesangial IgA deposits often associated with mesangial hypercellularity. **(See Glomerular Schematic 2.)** IgM, IgG, C_3, or immunoglobulin light chains may be codistributed with IgA. IgA deposited in the mesangium is typically polymeric and of the IgA1 subclass, the pathogenic significance of which is not clear. Abnormalities have been described in IgA production by plasma cells; in IgA clearance, by the liver and in mesangial IgA clearance and receptors for IgA. Currently, however, abnormalities in the *O* glycosylation of the hinge region of primarily polymeric IgA1 seem to best account for the pathogenesis of sporadic IgA nephropathy. Synthesis of poorly galactosylated IgA1 results in exposure of N-acetyl-galactosomine in truncated IgA1 hinge regions which is recognized by IgG or IgA1 antibodies leading to formation of immune complexes in the circulation or in situ after glomerular deposition of galactose-deficient IgA1. The galactose-deficient IgA1 may evade liver catabolism and preferentially deposit in the mesangium. A second hit, such as a viral or other antigen exposure, may be necessary for disease manifestation. Despite the presence of elevated serum IgA levels in 20–50% of patients, and IgA deposition in skin biopsies in 15–55% of patients, a renal biopsy is necessary to confirm the diagnosis. Although the immunofluorescent pattern of IgA on renal biopsy defines IgA nephropathy in the proper clinical context, a variety of histologic lesions may be seen on light microscopy **(see Fig. A3-8)**, including DPGN; *segmental sclerosis*; and, rarely, *segmental necrosis with cellular crescent formation*, which typically presents as RPGN.



The two most common presentations of IgA nephropathy are recurrent episodes of macroscopic hematuria during or immediately following an upper respiratory infection often accompanied by proteinuria or persistent asymptomatic microscopic hematuria. Nephrotic syndrome is uncommon. Proteinuria can also first appear late in the course of the disease. Rarely patients present with acute renal failure and a rapidly progressive clinical picture. IgA nephropathy is a benign disease for the majority of patients, and 5–30% of patients may go into a complete remission, with others having hematuria but well preserved renal function. In the minority of patients who have progressive disease, progression is slow, with renal failure seen in only 25–30% of patients with IgA nephropathy over 20–25 years. This risk varies considerably among populations. Cumulatively, risk factors for the loss of renal function identified thus far account for <50% of the variation in observed outcome but include the presence of hypertension or proteinuria, the absence of episodes of macroscopic hematuria, male sex, older age of onset, and extensive glomerulosclerosis or interstitial fibrosis on renal biopsy. Several analyses in large populations of patients found persistent proteinuria for 6 months or longer to have the greatest predictive power for adverse renal outcomes.

There is no agreement on optimal treatment. Both large studies that include patients with multiple glomerular diseases and small studies of patients with IgA nephropathy support the use of angiotensin-converting enzyme (ACE) inhibitors in patients with proteinuria or declining renal function. In patients with persistent proteinuria after ACE inhibitor therapy, steroid treatment or other immunosuppressives have demonstrated conflicting results. Tonsillectomy and fish oil have also been suggested in small studies to benefit select patients. When presenting as RPGN, patients typically receive steroids, cytotoxic agents, and plasmapheresis.

ANCA SMALL-VESSEL VASCULITIS

A group of patients with small-vessel vasculitis (arterioles, capillaries, and venules; rarely small arteries) and glomerulonephritis have serum ANCA; the antibodies are of two types, anti-proteinase 3 (PR3) or anti-myeloperoxidase (MPO) (Chap. 356); Lamp-2 antibodies have also been reported experimentally as potentially pathogenic. ANCA are produced with the help of T cells and activate leukocytes and monocytes, which together damage the walls of small vessels. Endothelial injury also attracts more leukocytes and extends the inflammation. Granulomatosis with polyangiitis, microscopic polyangiitis, Churg-Strauss syndrome, and renal-limited vasculitis belong to this group because they are ANCA-positive and have a *pauci-immune glomerulonephritis* with few immune complexes in small vessels and glomerular capillaries. Patients with any of these diseases can have any combination of the above serum antibodies, but anti-PR3

antibodies are more common in granulomatosis with polyangiitis and anti-MPO antibodies are more common in microscopic polyangiitis or Churg-Strauss. Although each of these diseases has some unique clinical features, most features do not predict relapse or progression, and as a group, they are generally treated in the same way. Once diagnosed ANCA monitoring has limited value, but targeted determination of ANCA levels may be useful if a relapse is clinically suspected. Since mortality is high without treatment, virtually all patients receive urgent treatment. Induction therapy usually includes glucocorticoids and either cyclophosphamide or rituximab. Plasmapheresis is recommended in rapidly progressive renal failure or pulmonary hemorrhage. Monthly "pulse" IV cyclophosphamide to induce remission of ANCA-associated vasculitis is as effective as daily oral cyclophosphamide but may be associated with increased relapses. Steroids are tapered soon after acute inflammation subsides, and patients are maintained on cyclophosphamide or less toxic agents such as azathioprine, methotrexate, or rituximab for up to a year to minimize the risk of relapse.

Granulomatosis with Polyangiitis

Patients with this disease classically present with fever, purulent rhinorrhea, nasal ulcers, sinus pain, polyarthralgias/arthritis, cough, hemoptysis, shortness of breath, microscopic hematuria, and 0.5–1 g/24 h of proteinuria; occasionally there may be cutaneous purpura and mononeuritis multiplex. Presentation without renal involvement is termed *limited granulomatosis with polyangiitis,* although some of these patients will show signs of renal injury later. Chest x-ray often reveals nodules and persistent infiltrates, sometimes with cavities. Biopsy of involved tissue will show a small-vessel vasculitis and adjacent noncaseating granulomas. Renal biopsies during active disease demonstrate *segmental necrotizing glomerulonephritis* without immune deposits and have been classified as focal, mixed, crescentic or sclerotic **(see Fig. A3-13)**. The disease is more common in patients exposed to silica dust and those with α_1-antitrypsin deficiency, which is an inhibitor of PR3. Relapse after achieving remission is common and is more common in patients with granulomatosis with polyangiitis than the other ANCA-associated vasculitis, necessitating diligent follow-up care. Although associated with an unacceptable high mortality rate without treatment, the greatest threat to patients, especially elderly patients in the first year of therapy, is from adverse events, which are often secondary to treatment, rather than active vasculitis. Patients should also be monitored long term for malignancy after immunosuppressive therapy.

Microscopic Polyangiitis

Clinically, these patients look somewhat similar to those with granulomatosis with polyangiitis, except they rarely have significant lung disease or destructive sinusitis. The distinction is made on biopsy, where the vasculitis in microscopic polyangiitis is without granulomas. Some patients will also have injury limited to the capillaries and venules.

Churg-Strauss Syndrome

When small-vessel vasculitis is associated with peripheral eosinophilia, cutaneous purpura, mononeuritis, asthma, and allergic rhinitis, a diagnosis of Churg-Strauss syndrome is considered. Hypergammaglobulinemia, elevated levels of serum IgE, or the presence of rheumatoid factor sometimes accompanies the allergic state. Lung inflammation, including fleeting cough and pulmonary infiltrates, often precedes the systemic manifestations of disease by years; lung manifestations are rarely absent. A third of patients may have exudative pleural effusions associated with eosinophils. Small-vessel vasculitis and *focal segmental necrotizing glomerulonephritis* can be seen on renal biopsy, usually absent eosinophils or granulomas. The cause of Churg-Strauss syndrome is autoimmune, but the inciting factors are unknown.

C_3 Glomerulopathies

C_3 glomerulopathy is a recent disease classification that is defined by the glomerular accumulation of C_3 with little or no immunoglobulin and encompasses dense deposit disease (DDD), formerly MPGN type II (see below), and C_3 glomerulonephritis (C_3GN), **(Table 308-4)**. DDD is defined morphologically with dense deposits forming ribbons in the GBM. In the absence of this specific morphology the entity is categorized as C_3GN. Both are associated with the presence of a complement mutation believed to cause the renal pathology, including mutations in the complement factor H regulatory proteins (CFHR's) genes. DDD is primarily a disease of children and young adults while the other C_3 glomerulopathies are reported to present in an older age group (mean age 30). By definition kidneys with C_3 glomerulopathy show sole or dominant staining for C_3 but can have variable light microscopy with mesangial proliferative or membranoproliferative patterns seen most commonly. Morphologically, many cases are not distinguishable from recovering post-infections GN. Patients with DDD

present with proteinuria and/or hematuria with nephrotic range proteinuria in up to 2/3 of patients. Partial lipodystrophy and Drusen bodies in the retina may also be present. Prognosis is poor with 50% of patients progressing to ESRD. C_3GN patients are clinically less well defined but ~2/3 have hematuria and 1/3 proteinuria. In addition to renal biopsy serological and genetic evaluation may be indicated including measurement of C_3 levels which are typically low with normal C_4 levels, C_3 nephritic factor, Factor H, paraprotein detection and specific CFHR genetic mutations. The optimal therapies remain undefined but include inhibition of the renin-angiotensin system, anticoagulants, steroids and other immunosuppressants. Increasing evidence suggests a benefit of therapy with eculizumab, a monoclonal antibody directed at C_5 which is activated by $C_{3..}$

TABLE 308-4

Membranoproliferative Glomerulonephritis: Immunoglobulin-Mediated

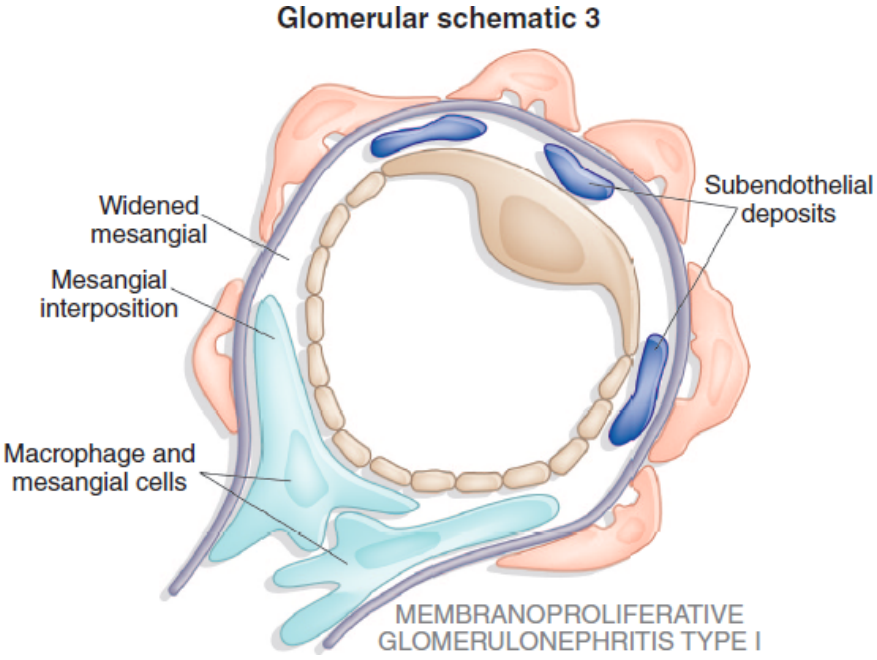
Type I Disease—Most Common
Idiopathic Subacute bacterial endocarditis Systemic lupus erythematosus Hepatitis C and cryoglobulinemia Mixed cryoglobulinemia Hepatitis C Cancer: lung, breast and ovary (germinal)
Type II Disease
Idiopathic Dense Deposit Disease (immunoglobulin-mediated)
Type III Disease
Idiopathic
C_3 Glomerulopathy: C_3 Dominant, Non-Immunoglobulin-mediated
Dense Deposit Disease (C_3 dominant)
Idiopathic Specific genetic mutations and/or autoantibodies to alternate complement pathway factors or regulatory factors of alternate complement pathway
C_3 Glomerulonephritis
Specific genetic mutations and/or autoantibodies to alternate complement pathway factors or regulatory factors of alternate complement pathway

MEMBRANOPROLIFERATIVE GLOMERULONEPHRITIS

MPGN is sometimes called *mesangiocapillary glomerulonephritis* or *lobar glomerulonephritis*. It is an immune-mediated glomerulonephritis characterized by thickening of the GBM with mesangioproliferative changes; 70% of patients have hypocomplementemia. MPGN is rare in African Americans, and idiopathic disease usually presents in childhood or young adulthood. MPGN has been subdivided pathologically into type I, type II, and type III disease. *Type I MPGN* is commonly associated with persistent hepatitis C infections, autoimmune diseases like lupus or cryoglobulinemia, or neoplastic diseases (Table 308-4). A minority of cases of MPGN type I

have C_3 but not immunoglobulin deposits on biopsy and are best considered as in the category of a C_3 glomerulopathy. *Types II and III MPGN* can be idiopathic, and immunoglobulin-mediated disease (driven by the classical complement pathway) but the majority of cases formerly defined as MPGN type II and III are non-immunoglobulin-mediated and driven by the alternate complement pathway.

Type I MPGN, the most proliferative of the three types, shows mesangial proliferation with lobular segmentation on renal biopsy and mesangial interposition between the capillary basement membrane and endothelial cells, producing a double contour sometimes called *tram-tracking* **(see Fig. A3-9)**. **(See Glomerular Schematic 3.)** Subendothelial deposits with low serum levels of C_3 are typical, although 50% of patients have normal levels of C_3 and occasional intramesangial deposits. Low serum C_3 and a dense thickening of the GBM containing ribbons of dense deposits and C_3 characterize type II MPGN*, dense deposit disease* **(see Fig. A3-10)**. Classically, the glomerular tuft has a lobular appearance; intramesangial deposits are rarely present and subendothelial deposits are generally absent. Proliferation in type III MPGN is less common than the other two types and is often focal; mesangial interposition is rare, and subepithelial deposits can occur along widened segments of the GBM that appear laminated and disrupted.



Classic type I MPGN is secondary to glomerular deposition of circulating immune complexes or their in situ formation. Patients with MPGN present with proteinuria, hematuria, and pyuria (30%); systemic symptoms of fatigue and malaise that are most common in children with type I disease; or an acute nephritic picture with RPGN and a speedy deterioration in renal function in up to 25% of patients. Low serum C_3 levels are common. Fifty percent of patients with MPGN develop ESRD 10 years after diagnosis, and 90% have renal insufficiency after 20 years. Nephrotic syndrome, hypertension, and renal insufficiency all predict poor outcome. In the presence of proteinuria, treatment with inhibitors of the renin-angiotensin system is prudent. Evidence for treatment with dipyridamole, Coumadin (warfarin), or cyclophosphamide is not strongly established. There is some evidence supporting the efficacy of treatment of *primary MPGN* with steroids, particularly in children, as well as reports of efficacy with plasma exchange and other immunosuppressive drugs. If defects in the complement pathway are found, treatment with eculizumab is of hypothetical but unproven benefit. In *secondary MPGN,* treating the associated infection, autoimmune disease, or neoplasms is of demonstrated benefit. In particular, pegylated interferon and ribavirin are useful in reducing viral load. Although all primary renal diseases can recur over time in transplanted renal allografts, patients with MPGN are well known to be at risk for not only a histologic recurrence but also a clinically significant recurrence with loss of graft function.

MESANGIOPROLIFERATIVE GLOMERULONEPHRITIS

Mesangioproliferative glomerulonephritis is characterized by expansion of the mesangium, sometimes associated with mesangial hypercellularity; thin, single contoured capillary walls; and mesangial immune deposits. Clinically, it can present with varying degrees of proteinuria and, commonly, hematuria. Mesangioproliferative disease may be seen in IgA nephropathy, *Plasmodium falciparum* malaria, resolving postinfectious glomerulonephritis, and class II nephritis from lupus, all of which can have a similar histologic appearance. With these secondary entities excluded, the diagnosis of *primary mesangioproliferative glomerulonephritis* is made in <15% of renal biopsies.

As an immune-mediated renal lesion with deposits of IgM, C1q, and C_3, the clinical course is variable. Patients with isolated hematuria may have a very benign course, and those with heavy proteinuria occasionally progress to renal failure. There is little agreement on treatment, but some clinical reports suggest benefit from use of inhibitors of the renin-angiotensin system, steroid therapy, and even cytotoxic agents.

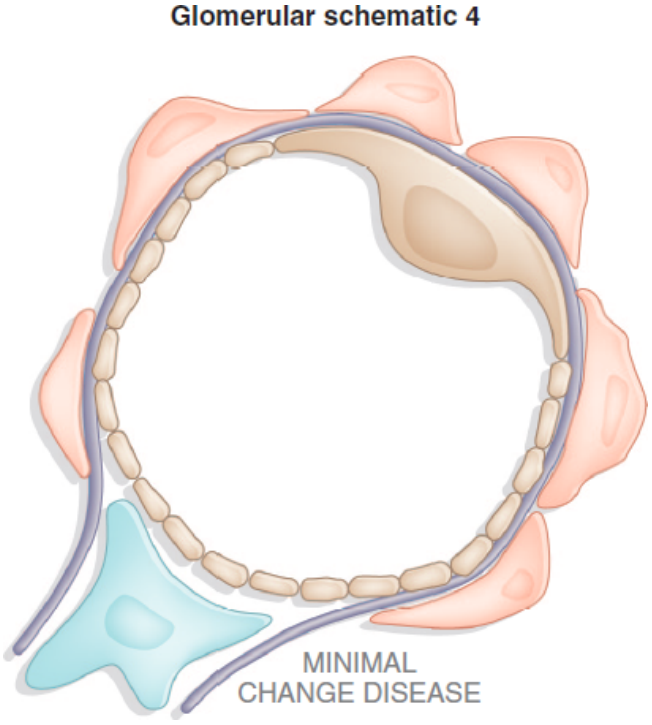
NEPHROTIC SYNDROME

Nephrotic syndrome classically presents with heavy proteinuria, minimal hematuria, hypoalbuminemia, hypercholesterolemia, edema, and hypertension. If left undiagnosed or untreated, some of these syndromes will progressively damage enough glomeruli to cause a fall in GFR, producing renal failure. Multiple studies have noted that the higher the 24-h urine protein excretion, the more rapid is the decline in GFR.

Therapies for various causes of nephrotic syndrome are noted under individual disease headings below. In general, all patients with hypercholesterolemia secondary to nephrotic syndrome should be treated with lipid-lowering agents because they are at increased risk for cardiovascular disease. Edema secondary to salt and water retention can be controlled with the judicious use of diuretics, avoiding intravascular volume depletion. Venous complications secondary to the hypercoagulable state associated with nephrotic syndrome can be treated with anticoagulants. The losses of various serum binding proteins, such as thyroid-binding globulin, lead to alterations in functional tests. Lastly, proteinuria itself is hypothesized to be nephrotoxic, and treatment of proteinuria with inhibitors of the renin-angiotensin system can lower urinary protein excretion.

MINIMAL CHANGE DISEASE

MCD, sometimes known as *nil lesion,* causes 70–90% of nephrotic syndrome in childhood but only 10–15% of nephrotic syndrome in adults. MCD usually presents as a primary renal disease but can be associated with several other conditions, including Hodgkin's disease, allergies, or use of nonsteroidal anti-inflammatory agents; significant interstitial nephritis often accompanies cases associated with nonsteroidal drug use. MCD on renal biopsy shows no obvious glomerular lesion by light microscopy and is negative for deposits by immunofluorescent microscopy, or occasionally shows small amounts of IgM in the mesangium **(see Fig. A3-1)**. **(See Glomerular Schematic 4.)** Electron microscopy, however, consistently demonstrates an effacement of the foot processes supporting the epithelial podocytes with weakening of slit-pore membranes. The pathophysiology of this lesion is uncertain. Most agree there is a circulating cytokine, perhaps related to a T cell response that alters capillary charge and podocyte integrity. The evidence for cytokine-related immune injury is circumstantial and is suggested by the presence of preceding allergies, altered cell-mediated immunity during viral infections, and the high frequency of remissions with steroids.



MCD presents clinically with the abrupt onset of edema and nephrotic syndrome accompanied by acellular urinary sediment. Average urine protein excretion reported in 24 h is 10 g with severe hypoalbuminemia. Less common clinical features include hypertension (30% in children, 50% in adults), microscopic hematuria (20% in children, 33% in adults), atopy or allergic symptoms (40% in children, 30% in adults), and decreased renal function (<5% in children, 30% in adults). The appearance of acute renal failure in adults is often seen more commonly in patients with low serum albumin and intrarenal edema (nephrosarca) that is responsive to intravenous albumin and diuretics. This presentation must be distinguished from acute renal failure secondary to hypovolemia. Acute tubular necrosis and interstitial inflammation are also reported. In children, the abnormal urine principally contains albumin with minimal amounts of higher-molecular-weight proteins, and is sometimes called *selective proteinuria*. Although up to 30% of children have a spontaneous remission, all children today are treated with steroids; only children who are nonresponders are biopsied in this setting. Primary responders are patients who have a complete remission (<0.2 mg/24 h of proteinuria) after a single course of prednisone; steroid-dependent patients relapse as their steroid dose is tapered. Frequent relapsers have two or more relapses in the 6 months following taper, and steroid-resistant patients fail to respond to steroid therapy. Adults are not considered steroid-resistant until after 4 months of therapy. Ninety to 95% of children will develop a complete remission after 8 weeks of steroid therapy, and 80–85% of adults will achieve complete remission, but only after a longer course of 20–24 weeks. Patients with steroid resistance may have FSGS on repeat biopsy. Some hypothesize that if the first renal biopsy does not have a sample of deeper corticomedullary glomeruli, then the correct diagnosis of FSGS may be missed.

Relapses occur in 70–75% of children after the first remission, and early relapse predicts multiple subsequent relapses, as do high levels of basal proteinuria. The frequency of relapses decreases after puberty. There is an increased risk of relapse following the rapid tapering of steroids in all groups. Relapses are less common in adults but are more resistant to subsequent therapy. Prednisone is first-line therapy, either given daily or on alternate days. Other immunosuppressive drugs, such as cyclophosphamide, chlorambucil, and mycophenolate mofetil, are saved for frequent relapsers, steroid-dependent patients, or steroid-resistant patients. Cyclosporine can induce remission, but relapse is also common when cyclosporine is withdrawn. The long-term prognosis in adults is less favorable when acute renal failure or steroid resistance occurs.

FOCAL SEGMENTAL GLOMERULOSCLEROSIS

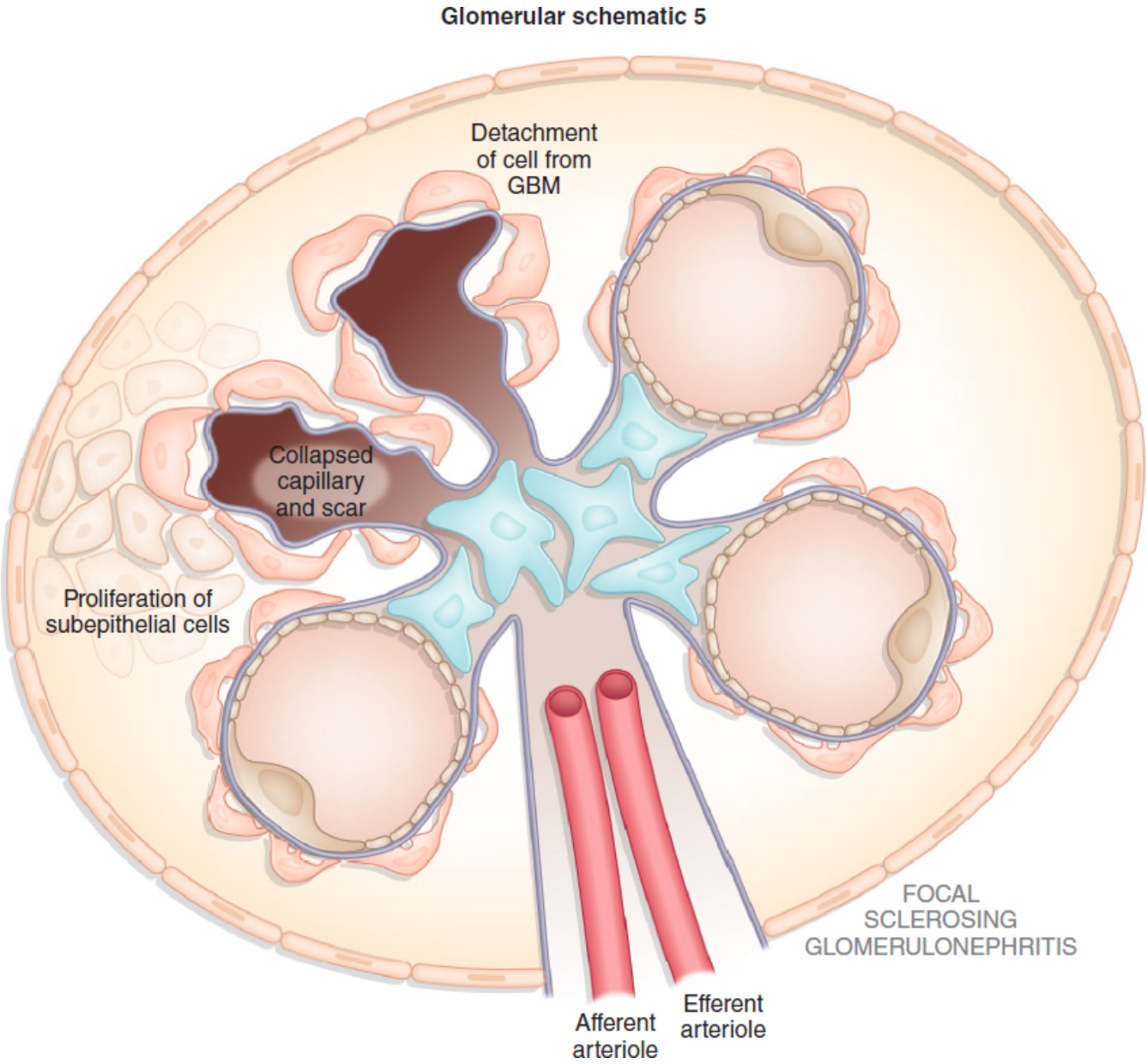
FSGS refers to a pattern of renal injury characterized by segmental glomerular scars that involve some but not all glomeruli; the clinical findings of FSGS largely manifest as proteinuria. When the secondary causes of FSGS are eliminated **(Table 308-5)**, the remaining patients are considered to have primary FSGS. The incidence of this disease is increasing, and it now represents up to one-third of cases of nephrotic syndrome in adults and one-half of cases of nephrotic syndrome in African Americans, in whom it is seen more commonly. The pathogenesis of FSGS is probably multifactorial. Possible mechanisms include a T cell–mediated circulating permeability factor, increased soluble urokinase receptor levels, TGF-β–mediated cellular proliferation and matrix synthesis, and podocyte abnormalities associated with genetic mutations. Risk polymorphisms at the *APOL1* locus encoding apolipoprotein L1 expressed in podocytes substantially explain the increased burden of FSGS among African Americans with or without HIV-associated disease.

TABLE 308-5

Focal Segmental Glomerulosclerosis

Primary focal segmental glomerulosclerosis
Secondary focal segmental glomerulosclerosis
Viruses: HIV/hepatitis B/parvovirus
Hypertensive nephropathy
Reflux nephropathy
Cholesterol emboli
Drugs: Heroin/analgesics/bisphosphonates/ecstasy
Oligomeganephronia
Renal dysgenesis
Alport's syndrome
Sickle cell disease
Lymphoma
Radiation nephritis
Familial podocytopathies
NPHS1 mutation/nephrin
NPHS2 mutation/podocin
TRPC6 mutation/cation channel
ACTN4 mutation/actinin
α-Galactosidase A deficiency/Fabry's disease
N acetylneuraminic acid hydrolase deficiency/nephrosialidosis

The pathologic changes of FSGS are most prominent in glomeruli located at the corticomedullary junction **(see Fig. A3-2)**, so if the renal biopsy specimen is from superficial tissue, the lesions can be missed, which sometimes leads to a misdiagnosis of MCD. In addition to focal and segmental scarring, other variants have been described, including cellular lesions with *endocapillary hypercellularity* and heavy proteinuria; *collapsing glomerulopathy* **(see Fig. A3-3)** with segmental or global glomerular collapse and a rapid decline in renal function; a hilar stalk lesion **(see Fig. A3-4)**; or the *glomerular tip lesion* **(see Fig. A3-5)**, which may have a better prognosis. **(See Glomerular Schematic 5.)**



FSGS can present with hematuria, hypertension, any level of proteinuria, or renal insufficiency. Nephrotic-range proteinuria, African-American race, and renal insufficiency are associated with a poor outcome, with 50% of patients reaching renal failure in 6–8 years. FSGS rarely remits spontaneously, but treatment-induced remission of proteinuria significantly improves prognosis. Treatment of patients with *primary FSGS* should include inhibitors of the renin-angiotensin system. Based on retrospective studies, patients with nephrotic-range proteinuria can be treated with steroids but respond far less often and after a longer course of therapy than patients with MCD. Proteinuria remits in only 20–45% of patients receiving a course of steroids over 6–9 months. Limited evidence suggests the use of cyclosporine in steroid-responsive patients helps ensure remissions. Relapse frequently occurs after cessation of cyclosporine therapy, and cyclosporine itself can lead to a deterioration of renal function due to its nephrotoxic effects. A role for other agents that suppress the immune system such as rituximab or mycophenolate mofetil has not been firmly established. Primary FSGS recurs in 25–40% of patients given allografts at ESRD, leading to graft loss in half of those cases. In recurrent post-transplant FSGS many patients will achieve a full or partial remission with plasmapheresis. The treatment of *secondary FSGS* typically involves treating the underlying cause and controlling proteinuria. There is no role for steroids or other immunosuppressive agents in secondary FSGS.

MEMBRANOUS GLOMERULONEPHRITIS

MGN, or *membranous nephropathy* as it is sometimes called, accounts for ~20% of cases of nephrotic syndrome in adults, with a peak incidence between the ages of 30 and 50 years and a male to female ratio of 2:1. IMN is rare in childhood and the most common cause of nephrotic syndrome in the elderly. In 20–30% of cases, MGN is secondary and is associated with a malignancy (solid tumors of the breast,

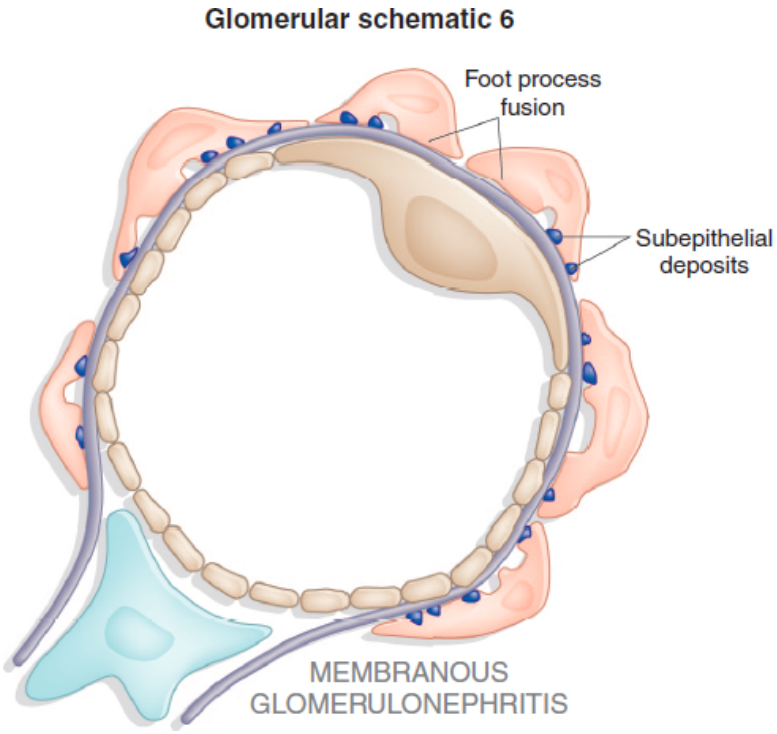
lung, colon), infection (hepatitis B, syphilis, malaria, schistosomiasis), rheumatologic disorders like lupus, rheumatoid arthritis, IgG4 diseases or drug exposure **(Table 308-6)**.

TABLE 308-6

Membranous Glomerulonephritis

Primary/idiopathic membranous glomerulonephritis
Secondary membranous glomerulonephritis
Infection: Hepatitis B and C, syphilis, malaria, schistosomiasis, leprosy, filariasis
Cancer: Breast, colon, lung, stomach, kidney, esophagus, neuroblastoma
Drugs: Gold, mercury, penicillamine, nonsteroidal anti-inflammatory agents, probenecid
Autoimmune diseases: Systemic lupus erythematosus, rheumatoid arthritis, primary biliary cirrhosis, dermatitis herpetiformis, bullous pemphigoid, myasthenia gravis, Sjögren's syndrome, Hashimoto's thyroiditis
Other systemic diseases: Fanconi's syndrome, sickle cell anemia, diabetes, Crohn's disease, sarcoidosis, Guillain-Barré syndrome, Weber-Christian disease, angiofollicular lymph node hyperplasia

Uniform thickening of the basement membrane along the peripheral capillary loops is seen by light microscopy on renal biopsy **(see Fig. A3-7)**; this thickening needs to be distinguished from that seen in diabetes and amyloidosis. **(See Glomerular Schematic 6.)** Immunofluorescence demonstrates diffuse granular deposits of IgG and C_3, and electron microscopy typically reveals electron-dense subepithelial deposits. While different stages (I–V) of progressive membranous lesions have been described, some published analyses indicate the degree of tubular atrophy or interstitial fibrosis is more predictive of progression than is the stage of glomerular disease. The presence of subendothelial deposits or the presence of tubuloreticular inclusions strongly points to a diagnosis of membranous lupus nephritis, which may precede the extrarenal manifestations of lupus. In 70% of cases of IMN, autoantibodies against the M-type phospholipase A_2 receptor circulate and bind to a conformational epitope present in the PLA2R on human podocytes, producing characteristic in situ deposits. 5–10% of IMN patients alternatively have autoantibodies to thrombospondin type-1 domain containing 7A. Both antigens co-localize within glomerular subepithelial deposits with IgG4 (PLA2R). Other renal diseases do not involve these autoantibodies. In most cases of secondary membranous nephropathy, these autoantibodies are absent with rare reports of autoantibodies to PLA2R in membranous glomerulopathy associated with hepatitis B and sarcoidosis. Circulating deposits and glomerular deposits of these autoantibodies have correlated with the likelihood of a spontaneous remission, severity of IMN, and the response to therapy. Eighty percent of patients with MGN present with nephrotic syndrome and nonselective proteinuria. Microscopic hematuria is seen but less commonly than in IgA nephropathy or FSGS. Spontaneous remissions occur in 20–33% of patients and often occur late in the course which make treatment decisions difficult. Low or absent levels of autoantibodies to PLA2R assist in predicting both spontaneous and treatment associated remissions. One-third of patients continue to have relapsing nephrotic syndrome but maintain normal renal function, and approximately another third of patients develop renal failure or die from the complications of nephrotic syndrome. Male gender, older age, hypertension, and the persistence of proteinuria are associated with worse prognosis. Although thrombotic complications are a feature of all nephrotic syndromes, MGN has the highest reported incidences of renal vein thrombosis, pulmonary embolism, and deep-vein thrombosis. Prophylactic anticoagulation is controversial but has been recommended for patients with severe or prolonged proteinuria in the absence of risk factors for bleeding.



In addition to the treatment of edema, dyslipidemia, and hypertension, inhibition of the renin-angiotensin system is recommended. Therapy with immunosuppressive drugs is also recommended for patients with primary MGN and persistent proteinuria (>3.0 g/24 h). The choice of immunosuppressive drugs for therapy is controversial, but current recommendations are to treat with steroids and cyclophosphamide, chlorambucil, mycophenolate mofetil, or cyclosporine or rituximab, an anti-CD20 antibody directed at B cells.

DIABETIC NEPHROPATHY

Diabetic nephropathy is the single most common cause of chronic renal failure in the United States, accounting for 45% of patients receiving renal replacement therapy, and is a rapidly growing problem worldwide. The dramatic increase in the number of patients with diabetic nephropathy reflects the epidemic increase in obesity, metabolic syndrome, and type 2 diabetes mellitus. Approximately 40% of patients with types 1 or 2 diabetes develop nephropathy, but due to the higher prevalence of type 2 diabetes (90%) compared to type 1 (10%), the majority of patients with diabetic nephropathy have type 2 disease. Renal lesions are more common in African-American, Native American, Polynesian, and Maori populations. Risk factors for the development of diabetic nephropathy include hyperglycemia, hypertension, dyslipidemia, smoking, a family history of diabetic nephropathy, and gene polymorphisms affecting the activity of the renin-angiotensin-aldosterone axis.

Within 1–2 years after the onset of clinical diabetes, morphologic changes appear in the kidney. Thickening of the GBM is a sensitive indicator for the presence of diabetes but correlates poorly with the presence or absence of clinically significant nephropathy. The composition of the GBM is altered notably with a loss of heparan sulfate moieties that form the negatively charged filtration barrier. This change results in increased filtration of serum proteins into the urine, predominately negatively charged albumin. The expansion of the mesangium due to the accumulation of extracellular matrix correlates with the clinical manifestations of diabetic nephropathy **(see stages in Fig. A3-20)**. This expansion in mesangial matrix is associated with the development of *mesangial sclerosis.* Some patients also develop eosinophilic, PAS^+ nodules called *nodular glomerulosclerosis* or *Kimmelstiel-Wilson nodules*. Immunofluorescence microscopy often reveals the nonspecific deposition of IgG (at times in a linear pattern) or complement staining without immune deposits on electron microscopy. Prominent vascular changes are frequently seen with hyaline and hypertensive arteriosclerosis. This is associated with varying degrees of chronic glomerulosclerosis and tubulointerstitial changes. Renal biopsies from patients with types 1 and 2 diabetes are largely indistinguishable.

These pathologic changes are the result of a number of postulated factors. Multiple lines of evidence support an important role for increases in glomerular capillary pressure (intraglomerular hypertension) in alterations in renal structure and function. Direct effects of hyperglycemia on the actin cytoskeleton of renal mesangial and vascular smooth-muscle cells as well as diabetes-associated changes in circulating factors such as atrial natriuretic factor, angiotensin II, and insulin-like growth factor (IGF) may account for this. Sustained

glomerular hypertension increases matrix production, alterations in the GBM with disruption in the filtration barrier (and hence proteinuria), and glomerulosclerosis. A number of factors have also been identified that alter matrix production, including the accumulation of advanced glycosylation end products, circulating factors including growth hormone, IGF-I, angiotensin II, connective tissue growth factor, TGF-β, and dyslipidemia.

The natural history of diabetic nephropathy in patients with types 1 and 2 diabetes is similar. However, since the onset of type 1 diabetes is readily identifiable and the onset of type 2 diabetes is not, a patient newly diagnosed with type 2 diabetes may present with *advanced diabetic nephropathy*. At the onset of diabetes, renal hypertrophy and glomerular hyperfiltration are present. The degree of glomerular hyperfiltration correlates with the subsequent risk of clinically significant nephropathy. In the ~40% of patients with diabetes who develop diabetic nephropathy, the earliest manifestation is an increase in albuminuria detected by sensitive radioimmunoassay. Albuminuria in the range of 30–300 mg/24 h is called *microalbuminuria* (Table 308-1). Microalbuminuria appears 5–10 years after the onset of diabetes. It is currently recommended to test patients with type 1 disease for microalbuminuria 5 years after diagnosis of diabetes and yearly thereafter and, because the time of onset of type 2 diabetes is often unknown, to test type 2 patients at the time of diagnosis of diabetes and yearly thereafter.

Patients with small increases in albuminuria increase their levels of urinary albumin excretion, typically reaching dipstick positive levels of proteinuria (>300 mg albuminuria) 5–10 years after the onset of early albuminuria. Microalbuminuria is a potent risk factor for cardiovascular events and death in patients with type 2 diabetes. Many patients with type 2 diabetes and microalbuminuria succumb to cardiovascular events before they progress to proteinuria or renal failure. Proteinuria in frank diabetic nephropathy can be variable, ranging from 500 mg to 25 g/24 h, and is often associated with nephrotic syndrome. More than 90% of patients with type 1 diabetes and nephropathy have diabetic retinopathy, so the absence of retinopathy in type 1 patients with proteinuria should prompt consideration of a diagnosis other than diabetic nephropathy; only 60% of patients with type 2 diabetes with nephropathy have diabetic retinopathy. There is a significant correlation between the presence of retinopathy and the presence of Kimmelstiel-Wilson nodules **(see Fig. A3-20)**. Also, characteristically, patients with advanced diabetic nephropathy have normal to enlarged kidneys, in contrast to many other glomerular diseases where kidney size is usually decreased. Using the above epidemiologic and clinical data, and in the absence of other clinical or serologic data suggesting another disease, diabetic nephropathy is usually diagnosed without a renal biopsy. After the onset of proteinuria, renal function inexorably declines, with 50% of patients reaching renal failure over another 5–10 years; thus, from the earliest stages of microalbuminuria, it usually takes 10–20 years to reach ESRD. However, up to 20–25% of patients with type 2 diabetes and chronic kidney disease have never had albuminuria documented. It is not known if this represents an altered natural history of diabetic nephropathy or another kidney disease that happens to occur in a patient with diabetes. Once renal failure appears, survival on dialysis is shorter for patients with diabetes compared to other dialysis patients. Survival is best for patients who receive a transplant from a living related donor.

Good evidence supports the benefits of blood sugar and blood pressure control as well as inhibition of the renin-angiotensin system in retarding the progression of diabetic nephropathy. In patients with type 1 diabetes, intensive control of blood sugar clearly prevents the development or progression of diabetic nephropathy. The evidence for benefit of intensive blood glucose control in patients with type 2 diabetes is less certain, with current studies reporting conflicting results.

Controlling systemic blood pressure decreases renal and cardiovascular adverse events in this high-risk population. The vast majority of patients with diabetic nephropathy require three or more antihypertensive drugs including ACE inhibitors or angiotensin receptor blockers (ARB) to achieve this goal. Drugs that inhibit the renin-angiotensin system, independent of their effects on systemic blood pressure, have been shown in numerous large clinical trials to slow the progression of diabetic nephropathy at early (microalbuminuria) and late (proteinuria with reduced glomerular filtration) stages. Since angiotensin II increases efferent arteriolar resistance and, hence, glomerular capillary pressure, one key mechanism for the efficacy of inhibitors of the renin angiotensin system is reducing glomerular hypertension. Evidence suggests increased risk for cardiovascular adverse events with little evidence of efficacy in some patients with a combination of two drugs (ACE inhibitors, ARBs, or renin inhibitors) that suppress several components of the renin-angiotensin system. Ongoing trials are examining the hypotheses that other agents may be of benefit including sodium glucose transport 2 inhibitors, endothelin antagonists, and aldosterone antagonists.

GLOMERULAR DEPOSITION DISEASES

Plasma cell dyscrasias producing excess light chain immunoglobulin sometimes lead to the formation of glomerular and tubular deposits that cause heavy proteinuria and renal failure; the same is true for the accumulation of serum amyloid A protein fragments seen in several inflammatory diseases. This broad group of proteinuric patients has *glomerular deposition disease*.

Light Chain Deposition Disease

The biochemical characteristics of nephrotoxic light chains produced in patients with light chain malignancies often confer a specific pattern of renal injury; that of either *cast nephropathy* **(see Fig. A3-17)**, which causes renal failure but not heavy proteinuria or amyloidosis, or light chain deposition disease **(see Fig. A3-16)**, which produces nephrotic syndrome with renal failure. These latter patients produce kappa light chains that do not have the biochemical features necessary to form amyloid fibrils. Instead, they self-aggregate and form granular deposits along the glomerular capillary and mesangium, tubular basement membrane, and Bowman's capsule. When predominant in glomeruli, nephrotic syndrome develops, and about 70% of patients progress to dialysis. Light-chain deposits are not fibrillar and do not stain with Congo red, but they are easily detected with anti–light chain antibody using immunofluorescence or as granular deposits on electron microscopy. A combination of the light chain rearrangement, self-aggregating properties at neutral pH, and abnormal metabolism probably contribute to the deposition. Treatment for light chain deposition disease is treatment of the primary disease and, if possible, autologous stem cell transplantation.

Renal Amyloidosis

Most *renal amyloidosis* is either the result of primary fibrillar deposits of immunoglobulin light chains known as amyloid L (AL), or secondary to fibrillar deposits of serum amyloid A (AA) protein fragments **(Chap. 108)**. Even though both occur for different reasons, their clinicopathophysiology is quite similar and will be discussed together. Amyloid infiltrates the liver, heart, peripheral nerves, carpal tunnel, upper pharynx, and kidney, producing restrictive cardiomyopathy, hepatomegaly, macroglossia, and heavy proteinuria sometimes associated with renal vein thrombosis. In systemic AL amyloidosis, also called *primary amyloidosis,* light chains produced in excess by clonal plasma cell dyscrasias are made into fragments by macrophages so they can self-aggregate at acid pH. A disproportionate number of these light chains (75%) are of the *lambda* class. About 10% of these patients have overt myeloma with lytic bone lesions and infiltration of the bone marrow with >30% plasma cells; nephrotic syndrome is common, and about 20% of patients progress to dialysis. AA amyloidosis is sometimes called *secondary amyloidosis* and also presents as nephrotic syndrome. It is due to deposition of β-pleated sheets of serum amyloid A protein, an acute phase reactant whose physiologic functions include cholesterol transport, immune cell attraction, and metalloproteases activation. Forty percent of patients with AA amyloid have rheumatoid arthritis, and another 10% have ankylosing spondylitis or psoriatic arthritis; the rest derive from other lesser causes. Less common in Western countries but more common in Mediterranean regions, particularly in Sephardic and Iraqi Jews, is familial Mediterranean fever (FMF). FMF is caused by a mutation in the gene encoding pyrin, whereas Muckle-Wells syndrome, a related disorder, results from a mutation in cryopyrin; both proteins are important in the apoptosis of leukocytes early in inflammation; such proteins with pyrin domains are part of a pathway called the *inflammasome*. Receptor mutations in tumor necrosis factor receptor 1 (TNFR1)-associated periodic syndrome also produce chronic inflammation and secondary amyloidosis. Fragments of serum amyloid A protein increase and self-aggregate by attaching to receptors for advanced glycation end products in the extracellular environment; nephrotic syndrome is common, and about 40–60% of patient's progress to dialysis. AA and AL amyloid fibrils are detectable with Congo red or in more detail with electron microscopy **(see Fig. A3-15)**. Serum-free light chain nephelometry assays are useful in the early diagnosis and follow-up of disease progression. Biopsy of involved liver or kidney is diagnostic 90% of the time when the pretest probability is high; abdominal fat pad aspirates are positive about 70% of the time, but apparently less so when looking for AA amyloid. Amyloid deposits are distributed along blood vessels and in the mesangial regions of the kidney. The treatment for primary amyloidosis, melphalan, and autologous hematopoietic stem cell transplantation can delay the course of disease in about 30% of patients. Secondary amyloidosis is also relentless unless the primary disease can be controlled. Some new drugs in development that disrupt the formation of fibrils may be available in the future.

Fibrillary-Immunotactoid Glomerulopathy

Fibrillary-immunotactoid glomerulopathy is a rare (<1.0% of renal biopsies), morphologically defined disease characterized by glomerular accumulation of nonbranching randomly arranged fibrils. Some classify amyloid and nonamyloid fibril-associated renal diseases all as fibrillary glomerulopathies with immunotactoid glomerulopathy reserved for nonamyloid fibrillary disease not associated with a systemic illness. Others define fibrillary glomerulonephritis as a nonamyloid fibrillary disease with fibrils 12–24 nm and immunotactoid glomerulonephritis with fibrils >30 nm. In either case, fibrillar/microtubular deposits of oligoclonal or oligotypic immunoglobulins and complement appear in the mesangium and along the glomerular capillary wall. Congo red stains are negative. The cause of this "nonamyloid" glomerulopathy is mostly idiopathic; reports of immunotactoid glomerulonephritis describe an occasional association with chronic lymphocytic leukemia or B cell lymphoma. Both disorders appear in adults in the fourth decade with moderate to heavy proteinuria, hematuria, and a wide variety of histologic lesions, including DPGN, MPGN, MGN, or mesangioproliferative glomerulonephritis.

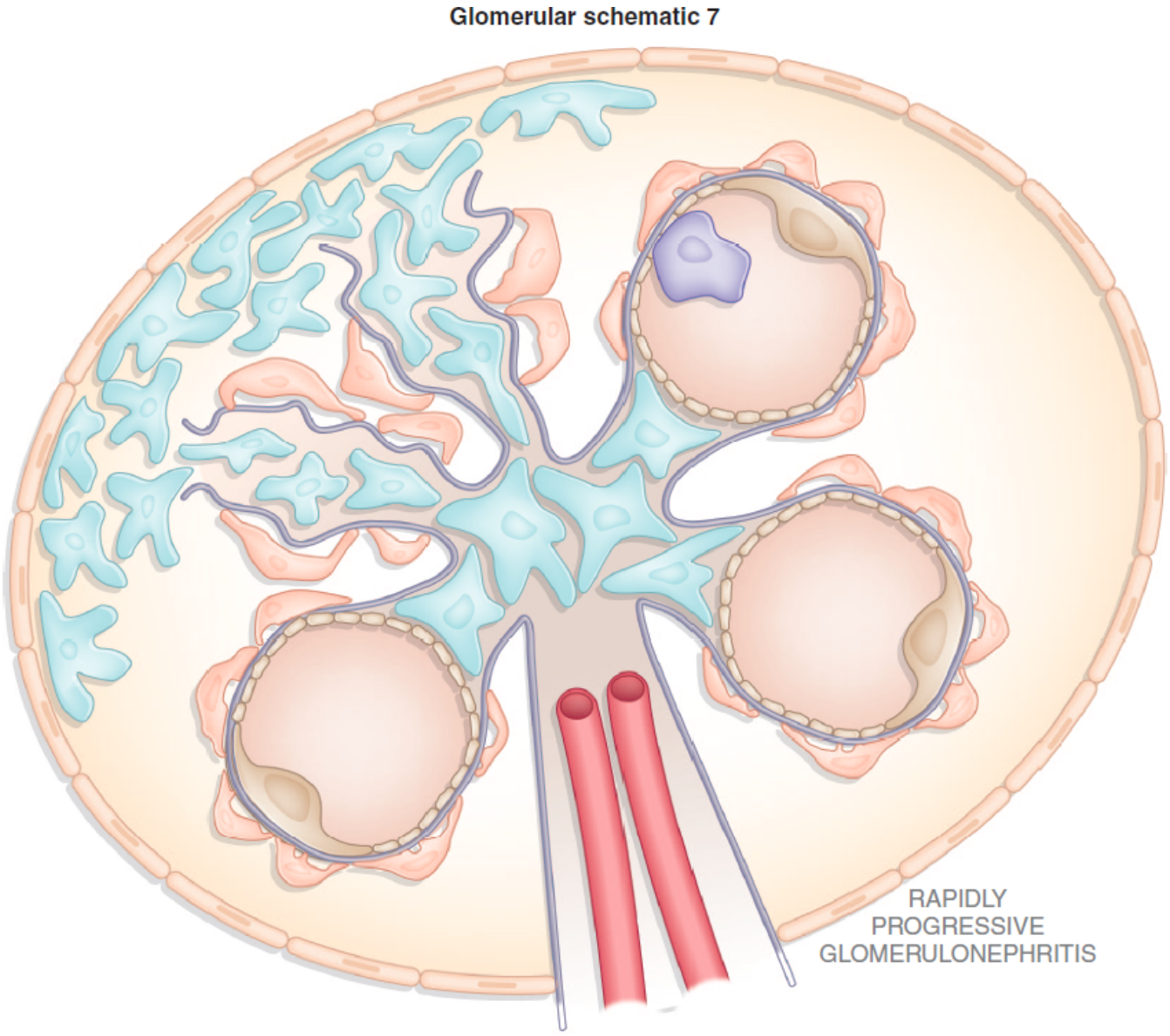
Nearly half of patients develop renal failure over a few years. There is no consensus on treatment of this uncommon disorder. The disease has been reported to recur following renal transplantation in a minority of cases.

FABRY'S DISEASE

Fabry's disease is an X-linked inborn error of globotriaosylceramide metabolism secondary to deficient lysosomal α-galactosidase A activity, resulting in excessive intracellular storage of globotriaosylceramide. Affected organs include the vascular endothelium, heart, brain, and kidneys. Classically, Fabry's disease presents in childhood in males with acroparesthesias, angiokeratoma, and hypohidrosis. Over time male patients develop cardiomyopathy, cerebrovascular disease, and renal injury, with an average age of death around 50 years of age. Hemizygotes with hypomorphic mutations sometimes present in the fourth to sixth decade with single-organ involvement. Rarely, dominant-negative α-galactosidase A mutations or female heterozygotes with unfavorable X inactivation present with mild single-organ involvement. Rare females develop severe manifestations including renal failure but do so later in life than males. Renal biopsy reveals enlarged glomerular visceral epithelial cells packed with small clear vacuoles containing globotriaosylceramide; vacuoles may also be found in parietal and tubular epithelia **(see Fig. A3-18)**. These vacuoles of electron-dense materials in parallel arrays (zebra bodies) are easily seen on electron microscopy. Ultimately, renal biopsies reveal FSGS. The nephropathy of Fabry's disease typically presents in the third decade as mild to moderate proteinuria, sometimes with microscopic hematuria or nephrotic syndrome. Urinalysis may reveal oval fat bodies and birefringent glycolipid globules under polarized light (Maltese cross). Renal biopsy is necessary for definitive diagnosis. Progression to renal failure occurs by the fourth or fifth decade. Treatment with inhibitors of the renin-angiotensin system is recommended. Treatment with recombinant α-galactosidase A clears microvascular endothelial deposits of globotriaosylceramide from the kidneys, heart, and skin. In patients with advanced organ involvement including chronic kidney disease, progression of disease occurs despite enzyme replacement therapy. Variable responses to enzyme therapy may be due to the occurrence of neutralizing antibodies or differences in uptake of the enzyme. Graft and patient survival following renal transplantation in patients with Fabry's are similar to other causes of ESRD.

PULMONARY-RENAL SYNDROMES

Several diseases can present with catastrophic hemoptysis and glomerulonephritis associated with varying degrees of renal failure. The usual causes include Goodpasture's syndrome, granulomatosis with polyangiitis, microscopic polyangiitis, Churg-Strauss vasculitis, and, rarely, Henoch-Schönlein purpura or cryoglobulinemia. Each of these diseases can also present without hemoptysis and are discussed in detail earlier in "Acute Nephritic Syndromes." **(See Glomerular Schematic 7.)** Pulmonary bleeding in this setting is life-threatening and often results in airway intubation, and acute renal failure requires dialysis. Diagnosis is difficult initially because biopsies and serologic testing take time. Treatment with plasmapheresis and methylprednisolone is often empirical and temporizing until results of testing are available.



BASEMENT MEMBRANE SYNDROMES

All kidney epithelia, including podocytes, rest on basement membranes assembled into a planar surface through the interweaving of collagen IV with laminins, nidogen, and sulfated proteoglycans. Structural abnormalities in GBM associated with hematuria are characteristic of several familial disorders related to the expression of collagen IV genes. The extended family of collagen IV contains six chains, which are expressed in different tissues at different stages of embryonic development. All epithelial basement membranes early in human development are composed of interconnected triple-helical protomers rich in α1.α1.α2(IV) collagen. Some specialized tissues undergo a developmental switch replacing α1.α1.α2(IV) protomers with an α3.α4.α5(IV) collagen network; this switch occurs in the kidney (glomerular and tubular basement membrane), lung, testis, cochlea, and eye, while an α5.α5.α6(IV) network appears in skin, smooth muscle, and esophagus and along Bowman's capsule in the kidney. This switch probably occurs because the α3.α4.α5(IV) network is more resistant to proteases and ensures the structural longevity of critical tissues. When basement membranes are the target of glomerular disease, they produce moderate proteinuria, some hematuria, and progressive renal failure.

ANTI-GBM DISEASE

Autoimmune disease where antibodies are directed against the α3 NC1 domain of collagen IV produces an *anti-GBM disease* often associated with RPGN and/or a pulmonary-renal syndrome called *Goodpasture's syndrome*. Discussion of this disease is covered earlier in "Acute Nephritic Syndromes."

ALPORT'S SYNDROME

Classically, patients with Alport's syndrome develop hematuria, thinning and splitting of the GBMs, mild proteinuria (<1–2 g/24 h), which appears late in the course, followed by chronic glomerulosclerosis leading to renal failure in association with sensorineural deafness. Some patients develop lenticonus of the anterior lens capsule, "dot and fleck" retinopathy, and rarely, mental retardation or leiomyomatosis. Approximately 85% of patients with Alport's syndrome have an X-linked inheritance of mutations in the α5(IV) collagen chain on chromosome Xq22–24. Female carriers have variable penetrance depending on the type of mutation or the degree of mosaicism created by X inactivation. Fifteen percent of patients have autosomal recessive disease of the α3(IV) or α4(IV) chains on chromosome 2q35–37. Rarely, some kindred have an autosomal dominant inheritance of dominant-negative mutations in α3(IV) or α4(IV) chains.

Pedigrees with the X-linked syndrome are quite variable in their rate and frequency of tissue damage leading to organ failure. Seventy percent of patients have the juvenile form with nonsense or missense mutations, reading frame shifts, or large deletions and generally develop renal failure and sensorineural deafness by age 30. Patients with splice variants, exon skipping, or missense mutations of α-helical glycines generally deteriorate after the age of 30 (adult form) with mild or late deafness. Early severe deafness, lenticonus, or proteinuria suggests a poorer prognosis. Usually females from X-linked pedigrees have only microhematuria, but up to 25% of carrier females have been reported to have more severe renal manifestations. Pedigrees with the autosomal recessive form of the disease have severe early disease in both females and males with asymptomatic parents.

Clinical evaluation should include a careful eye examination and hearing tests. However, the absence of extrarenal symptoms does not rule out the diagnosis. Since α5(IV) collagen is expressed in the skin, some X-linked Alport's patients can be diagnosed with a skin biopsy revealing the lack of the α5(IV) collagen chain on immunofluorescent analysis. Patients with mutations in α3(IV) or α4(IV) require a renal biopsy. Genetic testing can be used for the diagnosis of Alport's syndrome and the demonstration of the mode of inheritance. Early in their disease, Alport's patients typically have thin basement membranes on renal biopsy **(see Fig. A3-19)**, which thicken over time into multilamellations surrounding lucent areas that often contain granules of varying density—the so-called split basement membrane. In any Alport's kidney, there are areas of thinning mixed with splitting of the GBM. Tubules drop out, glomeruli scar, and the kidney eventually succumbs to interstitial fibrosis. All affected members of a family with X-linked Alport's syndrome should be identified and followed, including mothers of affected males. Primary treatment is control of systemic hypertension and use of ACE inhibitors to slow renal progression. Although patients who receive renal allografts usually develop anti-GBM antibodies directed toward the collagen epitopes absent in their native kidney, overt Goodpasture's syndrome is rare and graft survival is good.

THIN BASEMENT MEMBRANE DISEASE

Thin basement membrane disease (TBMD) characterized by persistent or recurrent hematuria is not typically associated with proteinuria, hypertension, or loss of renal function or extrarenal disease. Although not all cases are familial (perhaps a founder effect), it usually presents in childhood in multiple family members and is also called *benign familial hematuria*. Cases of TBMD have genetic defects in type IV collagen but in contrast to Alport behave as an autosomal dominant disorder that in ~40% of families segregates with the *COL(IV) α3/COL(IV) α4* loci. Mutations in these loci can result in a spectrum of disease ranging from TBMD to autosomal dominant or recessive Alport's. The GBM shows diffuse thinning compared to normal values for the patient's age in otherwise normal biopsies **(see Fig. A3-19)**. The vast majority of patients have a benign course.

NAIL-PATELLA SYNDROME

Patients with nail-patella syndrome develop iliac horns on the pelvis and dysplasia of the dorsal limbs involving the patella, elbows, and nails, variably associated with neural-sensory hearing impairment, glaucoma, and abnormalities of the GBM and podocytes, leading to hematuria, proteinuria, and FSGS. The syndrome is autosomal dominant, with haploinsufficiency for the *LIM* homeodomain transcription factor LMX1B; pedigrees are extremely variable in the penetrance for all features of the disease. LMX1B regulates the expression of genes encoding α3 and α4 chains of collagen IV, interstitial type III collagen, podocin, and CD2AP that help form the slit-pore membranes connecting podocytes. Mutations in the LIM domain region of *LMX1B* associate with glomerulopathy, and renal failure appears in as many as 30% of patients. Proteinuria or isolated hematuria is discovered throughout life, but usually by the third decade, and is inexplicably more common in females. On renal biopsy there is focal sclerosing glomerulonephritis with specific lucent damage to the lamina densa of the GBM, an increase in collagen III fibrils along glomerular capillaries and in the mesangium, and damage to the slit-pore membrane, producing heavy proteinuria not unlike that seen in congenital nephrotic syndrome. Patients with renal failure do well with transplantation.

GLOMERULAR-VASCULAR SYNDROMES

A variety of diseases result in classic vascular injury to the glomerular capillaries. Most of these processes also damage blood vessels elsewhere in the body. The group of diseases discussed here lead to vasculitis, renal endothelial injury, thrombosis, ischemia, and/or lipid-based occlusions.

ATHEROSCLEROTIC NEPHROPATHY

Aging in the developed world is commonly associated with the occlusion of coronary and systemic blood vessels. The reasons for this include obesity, insulin resistance, smoking, hypertension, and diets rich in lipids that deposit in the arterial and arteriolar circulation, producing local inflammation and fibrosis of small blood vessels. When the renal arterial circulation is involved, the glomerular microcirculation is damaged, leading to *chronic nephrosclerosis*. Patients with GFRs <60 mL/min have more cardiovascular events and hospitalizations than those with higher filtration rates. Several aggressive lipid disorders can accelerate this process, but most of the time atherosclerotic progression to chronic nephrosclerosis is associated with poorly controlled hypertension. Approximately 10% of glomeruli are normally sclerotic by age 40, rising to 20% by age 60 and 30% by age 80. Serum lipid profiles in humans are greatly affected by *apolipoprotein E* polymorphisms; the E4 allele is accompanied by increases in serum cholesterol and is more closely associated with atherogenic profiles in patients with renal failure. Mutations in E2 alleles, particularly in Japanese patients, produce a specific renal abnormality called *lipoprotein glomerulopathy* associated with glomerular lipoprotein thrombi and capillary dilation.

HYPERTENSIVE NEPHROSCLEROSIS

Systemic hypertension causes permanent damage to the kidneys in about 6% of patients with elevated blood pressure. As many as 27% of patients with end-stage kidney disease have hypertension as a primary cause. Although there is not a clear correlation between the extent or duration of hypertension and the risk of end-organ damage, *hypertensive nephrosclerosis* is fivefold more frequent in African Americans than whites. Risk alleles associated with *APOL1,* a functional gene for apolipoprotein L1 expressed in podocytes, substantially explains the increased burden of ESRD among African Americans. Associated risk factors for progression to end-stage kidney disease include increased age, male gender, race, smoking, hypercholesterolemia, duration of hypertension, low birth weight, and preexisting renal injury. Kidney biopsies in patients with hypertension, microhematuria, and moderate proteinuria demonstrate arteriolosclerosis, chronic nephrosclerosis, and interstitial fibrosis in the absence of immune deposits **(see Fig. A3-21)**. Today, based on a careful history, physical examination, urinalysis, and some serologic testing, the diagnosis of chronic nephrosclerosis is usually inferred without a biopsy. Recent studies suggest, in the absence of diabetes, adults with hypertension and cardiovascular risk factors benefit from achieving a systolic BP <120 mmHg compared to <140 mmHg. In the presence of kidney disease, most patients begin antihypertensive therapy with two drugs, classically a thiazide diuretic and an ACE inhibitor; most will require three drugs. There is strong evidence in African Americans with hypertensive nephrosclerosis that therapy initiated with an ACE inhibitor can slow the rate of decline in renal function independent of effects on systemic blood pressure. Malignant acceleration of hypertension complicates the course of chronic nephrosclerosis, particularly in the setting of scleroderma or cocaine use **(see Fig. A3-24)**. The hemodynamic stress of malignant hypertension leads to fibrinoid necrosis of small blood vessels, thrombotic microangiography, a nephritic urinalysis, and acute renal failure. In the setting of renal failure, chest pain, or papilledema, the condition is treated as a hypertensive emergency.

CHOLESTEROL EMBOLI

Aging patients with clinical complications from atherosclerosis sometimes shower cholesterol crystals into the circulation—either spontaneously or, more commonly, following an endovascular procedure with manipulation of the aorta—or with use of systemic anticoagulation. Spontaneous emboli may shower acutely or shower subacutely and somewhat more silently. Irregular emboli trapped in the microcirculation produce ischemic damage that induces an inflammatory reaction. Depending on the location of the atherosclerotic plaques releasing these cholesterol fragments, one may see cerebral transient ischemic attacks; livedo reticularis in the lower extremities; Hollenhorst plaques in the retina with visual field cuts; necrosis of the toes; and acute glomerular capillary injury leading to FSGS sometimes associated with hematuria, mild proteinuria, and loss of renal function, which typically progresses over a few years. Occasional patients have fever, eosinophilia, or eosinophiluria. A skin biopsy of an involved area may be diagnostic. Since tissue fixation dissolves the cholesterol, one typically sees only residual, biconvex clefts in involved vessels **(see Fig. A3-22)**. There is no therapy to reverse embolic occlusions, and steroids do not help. Controlling blood pressure and lipids and cessation of smoking are usually recommended for prevention.

SICKLE CELL DISEASE

Although individuals with SA-hemoglobin are usually asymptomatic, most will gradually develop hyposthenuria due to subclinical infarction of the renal medulla, thus predisposing them to volume depletion. There is an unexpectedly high prevalence of sickle trait among dialysis patients who are African American. Patients with homozygous SS-sickle cell disease and less commonly SC-sickle cell disease develop chronic vasoocclusive disease in many organs. Polymers of deoxygenated SS-hemoglobin distort the shape of red blood cells. These cells attach to endothelia and obstruct small blood vessels, producing frequent and painful sickle cell crises over time. Vessel occlusions in the kidney produce glomerular hypertension, FSGS, interstitial nephritis, and renal infarction associated with hyposthenuria, microscopic hematuria, and even gross hematuria; some patients also present with MPGN. Renal function can be overestimated due to the increased tubular secretion of creatinine seen in many patients with SS-sickle cell. By the second or third decade of life, persistent vasoocclusive disease in the kidney leads to varying degrees of renal failure, and some patients end up on dialysis. Their prognosis on dialysis is poor and anemia management with erythropoiesis-stimulating agents complicated. Treatment is directed to reducing the frequency of painful crises and administering ACE inhibitors in the hope of delaying a progressive decline in renal function. In sickle cell patients undergoing renal transplantation, renal graft survival is comparable to African Americans in the general transplant population.

THROMBOTIC MICROANGIOPATHIES

Thrombotic thrombocytopenic purpura (TTP) and *hemolytic-uremic syndrome* (HUS) represent a spectrum of thrombotic microangiopathies. TTP and HUS share the general features of idiopathic thrombocytopenic purpura, hemolytic anemia, fever, renal failure, and neurologic disturbances. When patients, particularly children, have more evidence of renal injury, their condition tends to be called HUS. In adults with neurologic disease, it is considered to be TTP. In adults there is often a mixture of both, which is why they are often referred to as having TTP/HUS. On examination of kidney tissue, there is evidence of *glomerular capillary endotheliosis* associated with platelet thrombi, damage to the capillary wall, and formation of fibrin material in and around glomeruli **(see Fig. A3-23)**. These tissue findings are similar to what is seen in preeclampsia/HELLP (*h*emolysis, *e*levated *l*iver enzymes, and *l*ow *p*latelet count syndrome), malignant hypertension, and the antiphospholipid syndrome. TTP/HUS is also seen in pregnancy; with the use of oral contraceptives or quinine; in renal transplant patients given OKT3 for rejection; in patients taking the calcineurin inhibitors, cyclosporine and tacrolimus, or in patients taking the antiplatelet agents, ticlopidine and clopidogrel; or following HIV infection.

Although there is no agreement on how much they share a final common pathophysiology, two general groups of patients are recognized: childhood HUS associated with enterohemorrhagic diarrhea and TTP/HUS in adults. Childhood HUS is caused by a toxin released by *Escherichia coli* 0157:H7 and occasionally by *Shigella dysenteriae*. This shiga toxin (verotoxin) directly injures endothelia, enterocytes, and renal cells, causing apoptosis, platelet clumping, and intravascular hemolysis by binding to the glycolipid receptors (Gb3). These receptors are more abundant along endothelia in children compared to adults. Shiga toxin also inhibits the endothelial production of ADAMTS13. In familial cases of adult TTP/HUS, there is a genetic deficiency of the ADAMTS13 metalloprotease that cleaves large multimers of von Willebrand's factor. Absent ADAMTS13, these large multimers cause platelet clumping and intravascular hemolysis. An antibody to ADAMTS13 is found in many sporadic cases of adult TTP/HUS, but not all; many patients also have antibodies to the thrombospondin receptor on selected endothelial cells in small vessels or increased levels of plasminogen-activator inhibitor 1 (PAI-1). Patients can be tested for ADAMTS13 activity and, if low, the presence of antibodies to ADAMTS13 distinguishes the deficiency from the immune-mediated disease. Some children with complement protein deficiencies express atypical HUS (aHUS), which can be treated with liver transplant. The treatment of adult TTP/HUS with ADAMTS13 antibodies is daily plasmapheresis, which can be lifesaving. Plasmapheresis with fresh frozen plasma is given until the platelet count rises, but in relapsing patients it normally is continued well after the platelet count improves, and in resistant patients twice-daily exchange may be helpful. Most patients respond within 2 weeks of daily plasmapheresis. Since TTP/HUS often has an autoimmune basis, there is an anecdotal role in relapsing patients for using splenectomy, steroids, immunosuppressive drugs, bortezomib, or rituximab, an anti-CD20 antibody. Patients without antibodies and a genetic deficiency of ADAMTS13 production can potentially be treated with fresh frozen plasma alone. Patients with childhood HUS from infectious diarrhea are not given antibiotics, because antibiotics are thought to accelerate the release of the toxin and the diarrhea is usually self-limited. No intervention appears superior to supportive therapy in children with postdiarrheal HUS.

ANTIPHOSPHOLIPID ANTIBODY SYNDROME

(See Chap. 350)

GLOBAL CONSIDERATIONS

INFECTIOUS DISEASE–ASSOCIATED SYNDROMES

 A number of infectious diseases will injure the glomerular capillaries as part of a systemic reaction producing an immune response or from direct infection of renal tissue. Evidence of this immune response is collected by glomeruli in the form of immune deposits that damage the kidney, producing moderate proteinuria and hematuria. A high prevalence of many of these infectious diseases in developing countries results in infection-associated renal disease being the most common cause of glomerulonephritis in many parts of the world.

Poststreptococcal Glomerulonephritis

This form of glomerulonephritis is one of the classic complications of streptococcal infection. The discussion of this disease can be found earlier, in the section “Acute Nephritic Syndromes.”

Subacute Bacterial Endocarditis

Renal injury from persistent bacteremia absent the continued presence of a foreign body, regardless of cause, is treated presumptively as if the patient has endocarditis. The discussion of this disease can be found earlier, in the section “Acute Nephritic Syndromes.”

Human Immunodeficiency Virus

Renal disease is an important complication of HIV disease. The risk of development of ESRD is much higher in HIV-infected African Americans than in HIV-infected whites. About 50% of HIV-infected patients with kidney disease have HIV-associated nephropathy (HIVAN) on biopsy. The lesion in HIVAN is FSGS, characteristically revealing a collapsing glomerulopathy **(see Fig. A3-3)** with visceral epithelial cell swelling, microcystic dilatation of renal tubules, and tubuloreticular inclusion. Renal epithelial cells express replicating HIV virus, but host immune responses also play a role in the pathogenesis. HIVAN develops almost exclusively in patients of black race origin who have the APOL1 variant. HIVICK, HIV immune complex kidney disease is a group of immune complex-mediated glomerular lesions seen in HIV patients, and on biopsy can look like a constellation of other glomerular lesions, including postinfectious glomerulonephritis, MGN, MPGN, DPGN, MCD, and IgA nephropathy. The HIVICK effect is a complication of active HIV viremia.

HIV patients with FSGS typically present with nephrotic-range proteinuria and hypoalbuminemia, but unlike patients with other etiologies for nephrotic syndrome, they do not commonly have hypertension, edema, or hyperlipidemia. Renal ultrasound also reveals large, echogenic kidneys despite the finding that renal function in some patients declines rapidly. Treatment with inhibitors of the renin-angiotensin system decreases the proteinuria. Effective antiretroviral therapy benefits both the patient and the kidney and improves survival of HIV-infected patients with HIVAN and in some cases HIVICK-associated chronic kidney disease or ESRD. In HIV-infected patients not yet on therapy, the presence of HIVAN is an indication to initiate therapy. Following the introduction of antiretroviral therapy, survival on dialysis for the HIV-infected patient has improved dramatically. Renal transplantations in HIV-infected patients without detectable viral loads or histories of opportunistic infections provide a better survival benefit over dialysis. Following transplantation, patient and graft survival are similar to the general transplant population despite frequent rejections.

Hepatitis B and C

Typically, infected patients present with microscopic hematuria, nonnephrotic or nephrotic-range proteinuria, and hypertension. There is a close association between hepatitis B infection and polyarteritis nodosa with vasculitis appearing generally in the first 6 months following infection. Renal manifestations include renal artery aneurysms, renal infarction, and ischemic scars. Alternatively, the hepatitis B carrier state can produce a MGN with predominant IgG1 deposition that is more common in children than adults, or MPGN that is more common in adults than in children. Renal histology is indistinguishable from idiopathic MGN, type I or type 3 MPGN. Viral antigens most commonly, HBeAG, are found in the renal deposits. Cryoglobulinemic glomerulonephritis has also been reported. There are no good treatment guidelines, but interferon α-2b and antiviral agents which consist of either nucleotide or nucleoside reverse transcription inhibitors have been used to some effect. Children have a good prognosis, with 60–65% achieving spontaneous remission within 4 years. In contrast, 30% of adults have renal insufficiency and 10% have renal failure 5 years after diagnosis.

Up to 30% of patients with chronic hepatitis C infection have some renal manifestations. Patients often present with type II mixed cryoglobulinemia, nephrotic syndrome, microscopic hematuria, abnormal liver function tests, depressed C3 levels, anti–hepatitis C virus (HCV) antibodies, and viral RNA in the blood. The renal lesions most commonly seen, in order of decreasing frequency, are cryoglobulinemic glomerulonephritis, MGN, and type I MPGN but PAN, IgA and FSGS have been reported. With the availability of direct-

acting antivirals, including ledipasvir/sofosbuvir which can achieve a viral remission in >95% of patients, the prevalence of glomerular disease in HCV patients should decline. These drugs are currently the treatment of choice for patients with HCV-related MPGN or PAN.

Other Viruses

Other viral infections are occasionally associated with glomerular lesions, but cause and effect are not well established. These viral infections and their respective glomerular lesions include: cytomegalovirus producing MPGN or FSGS; influenza and anti-GBM disease; measles-associated endocapillary proliferative glomerulonephritis, with measles antigen in the capillary loops and mesangium; parvovirus causing mild proliferative or mesangioproliferative glomerulonephritis or FSGS; mumps and mesangioproliferative glomerulonephritis; Epstein-Barr virus producing MPGN, diffuse proliferative nephritis, or IgA nephropathy; dengue hemorrhagic fever causing endocapillary proliferative glomerulonephritis; Hanta virus and mesangial proliferative glomerulonephritis and coxsackievirus producing *focal glomerulonephritis* or DPGN.

Syphilis

Secondary syphilis, with rash and constitutional symptoms, develops weeks to months after the chancre first appears and occasionally presents with the nephrotic syndrome from MGN caused by subepithelial immune deposits containing treponemal antigens. Other lesions have also rarely been described including interstitial syphilitic nephritis. The diagnosis is confirmed with nontreponemal and treponemal tests for *Treponema pallidum*. The renal lesion responds to treatment with penicillin or an alternative drug, if allergic. Additional testing for other sexually transmitted diseases is an important part of disease management.

Leprosy

Despite aggressive eradication programs, ~400,000 new cases of leprosy appear annually worldwide. The diagnosis is best made in patients with multiple skin lesions accompanied by sensory loss in affected areas, using skin smears showing paucibacillary or multibacillary infection (WHO criteria). Leprosy is caused by infection with *Mycobacterium leprae* and can be classified by Ridley-Jopling criteria into various types: tuberculoid, borderline tuberculoid, mid-borderline and borderline lepromatous, and lepromatous. Renal involvement in leprosy is related to the quantity of bacilli in the body, and the kidney is one of the target organs during splanchnic localization. In some series, all cases with borderline lepromatous and lepromatous types of leprosy have various forms of renal involvement including FSGS, mesangioproliferative glomerulonephritis, or renal amyloidosis; much less common are the renal lesions of DPGN and MPGN. Treatment of the infection with multi-drug therapy can reduce the incidence of renal disease or produce remission of the renal disease.

Malaria

There are 300–500 million incident cases of malaria each year worldwide, and the kidney is commonly involved. Glomerulonephritis is due to immune complexes containing malarial antigens that are implanted in the glomerulus. In malaria from *P. falciparum*, mild proteinuria is associated with subendothelial deposits, mesangial deposits, and mesangioproliferative glomerulonephritis that usually resolve with treatment. In quartan malaria from infection with *Plasmodium malariae*, children are more commonly affected and renal involvement is more severe. Transient proteinuria and microscopic hematuria can resolve with treatment of the infection. However, resistant nephrotic syndrome with progression to renal failure over 3–5 years does happen, as <50% of patients respond to steroid therapy. Affected patients with nephrotic syndrome have thickening of the glomerular capillary walls, with subendothelial deposits of IgG, IgM, and C_3 associated with a sparse membranoproliferative lesion. The rare mesangioproliferative glomerulonephritis reported with *Plasmodium vivax* or *Plasmodium ovale* typically has a benign course. Acute kidney injury can often complicate these glomerulopathies.

Schistosomiasis

Schistosomiasis affects >300 million people worldwide and primarily involves the urinary and gastrointestinal tracts. Glomerular involvement varies with the specific strain of schistosomiasis; *Schistosoma mansoni* is most commonly associated with clinical renal disease, and the glomerular lesions can be classified: Class I is a *mesangioproliferative glomerulonephritis*; class II is an *extracapillary proliferative glomerulonephritis*; class III is a *membranoproliferative glomerulonephritis*; class IV is a *focal segmental glomerulonephritis*; and class V is *amyloidosis*. Classes I–II often remit with treatment of the infection, but classes III and IV lesions are associated with IgA immune deposits and progress despite antiparasitic and/or immunosuppressive therapy.

Other Parasites

Renal involvement with toxoplasmosis infections is rare. When it occurs, patients present with nephrotic syndrome and have a histologic picture of MPGN. Fifty percent of patients with leishmaniasis will have mild to moderate proteinuria and microscopic hematuria, but renal insufficiency is rare. Acute DPGN, MGN, and mesangioproliferative glomerulonephritis have all been observed on biopsy. Filariasis and trichinosis are caused by nematodes and are sometimes associated with glomerular injury presenting with proteinuria, hematuria, and a variety of histologic lesions that typically resolve with eradication of the infection.

FURTHER READING

The GISEN Group (Gruppo Italiano di Studi Epidemiologici in Nefrologia): Randomised placebo-controlled trial of effect of ramipril on decline in glomerular filtration rate and risk of terminal renal failure in proteinuric, non-diabetic nephropathy. Lancet 349:1857, 1997. [PubMed: 9217756]

Kopp JB et al: APOL1 genetic variants in focal segmental glomerulosclerosis and HIV-associated nephropathy. J Am Soc Nephrol 22:2129, 2011. [PubMed: 21997394]

Kupin WL: Viral-associated GN: Hepatitis C and HIV. Clin J Am Soc Nephrol 12:1337, 2017.

Lewis EJ et al: Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. N Engl J Med 345:851, 2001. [PubMed: 11565517]

Pickering MC et al: C_3 glomerulopathy: Consensus report. Kidney Int 84:1079, 2013. [PubMed: 24172683]

Ronco P, Debiec H: Membranous nephropathy: A fairy tale for immunopathologists, nephrologists and patients. Mol Immunol 68:57, 2015. [PubMed: 26597209]

Chapter 309: Polycystic Kidney Disease and Other Inherited Disorders of Tubule Growth and Development

Jing Zhou; Martin R. Pollak

INTRODUCTION

The polycystic kidney diseases are a group of genetically heterogeneous disorders and a leading cause of kidney failure. The autosomal dominant form of polycystic kidney disease (ADPKD) is the most common life-threatening monogenic disease, affecting 12 million people worldwide. The autosomal recessive form of polycystic kidney disease (ARPKD) is rarer but affects the pediatric population. Kidney cysts are often seen in a wide range of syndromic diseases. Recent studies have shown that defects in the structure or function of the primary cilia may underline this group of genetic diseases collectively termed *ciliopathies* (Table 309-1).

TABLE 309-1

Inherited Diseases Commonly Associated with a Cystic Phenotype

Disease	Mode of Inheritance	Renal Abnormalities	Other Clinical Features	Genes
Autosomal dominant polycystic kidney disease	AD	Cortical and medullary cysts	Liver, pancreatic cysts, hypertension, subarachnoid hemorrhage	*PKD1, PKD2*
Autosomal recessive polycystic kidney disease	AR	Distal and collecting duct cysts	Oligohydramnios if severe, hypertension, ascending cholangitis, liver fibrosis	*PKHD1*
Medullary cystic kidney (Autosomal dominant tubulointerstital kidney disease)	AD	Small fibrotic kidneys; medullary cysts	In adults, gout	*UMOD* *MUC1* *REN*
Nephronophthisis	AR	Small fibrotic kidneys; medullary cysts	Growth retardation, anemia, (visual loss, liver fibrosis, cerebellar ataxia if associated with another syndrome)	*NPHP1-20, IQCB1, CEP290, GLIS2, RPGRIP1L, NEK8, SDCCAG8, TMEM67, TTC21B*
Senior-Loken syndrome	AR	Renal cysts	Juvenile nephronophthisis, Leber amaurosis	*NPHP1-6, SDCCAG8*
Leber congenital amaurosis	AR	Renal cysts	Visual impairment in first year of life; pigmentary retinopathy	*GUCY2D, RPE65, LCA3-14* (including *LCA10, CEP290*)
Meckel-Gruber syndrome	AR	Cortical and medullary cysts	CNS anomalies, polydactyly, congenital heart defects	*MKS1, TMEM216, TMEM67, CEP290, RPGRIP1L, CC2D2A, TCTN2, B9D1, B9D2, NPHP3*
Bardet-Biedl syndrome	AR	Renal cysts	Obesity, polydactyly, retinitis pigmentosa, anosmia, congenital heart defects, mental retardation	*BBS1, 2, ARL6, BBS4,5, MKKS, BBS7, TTC8, BBS9, 10, TRIM32, BBS12, MKS1, CEP290, C2ORF86; modifiers MKS1, MKS3, CCDC28B*
Oral-facial-digital syndrome type I	AR	Renal cysts	Oral cavity, face, and digit anomalies; CNS abnormalities; cystic kidney disease; X-linked with male lethality, primary ciliary dyskinesia	*OFD1*

Disease	Mode of Inheritance	Renal Abnormalities	Other Clinical Features	Genes
Cranioectodermal dysplasia (Sensenbrenner syndrome)	AR	Renal cysts	Skeletal dysplasia; thoracic deformities; polydactyly; renal cysts; retinitis pigmentosa	*IFT80*
Tuberous sclerosis	AD	Renal cysts	Angiomyolipomas; renal cell carcinoma Facial angiofibromas; CNS hamartomas	*TSC1, TSC2*
Von Hippel-Lindau disease	AD	Renal cysts	Renal cell carcinoma, retinal angiomas; CNS hemangioblastomas; pheochromocytomas	*VHL*

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; CNS, central nervous system.

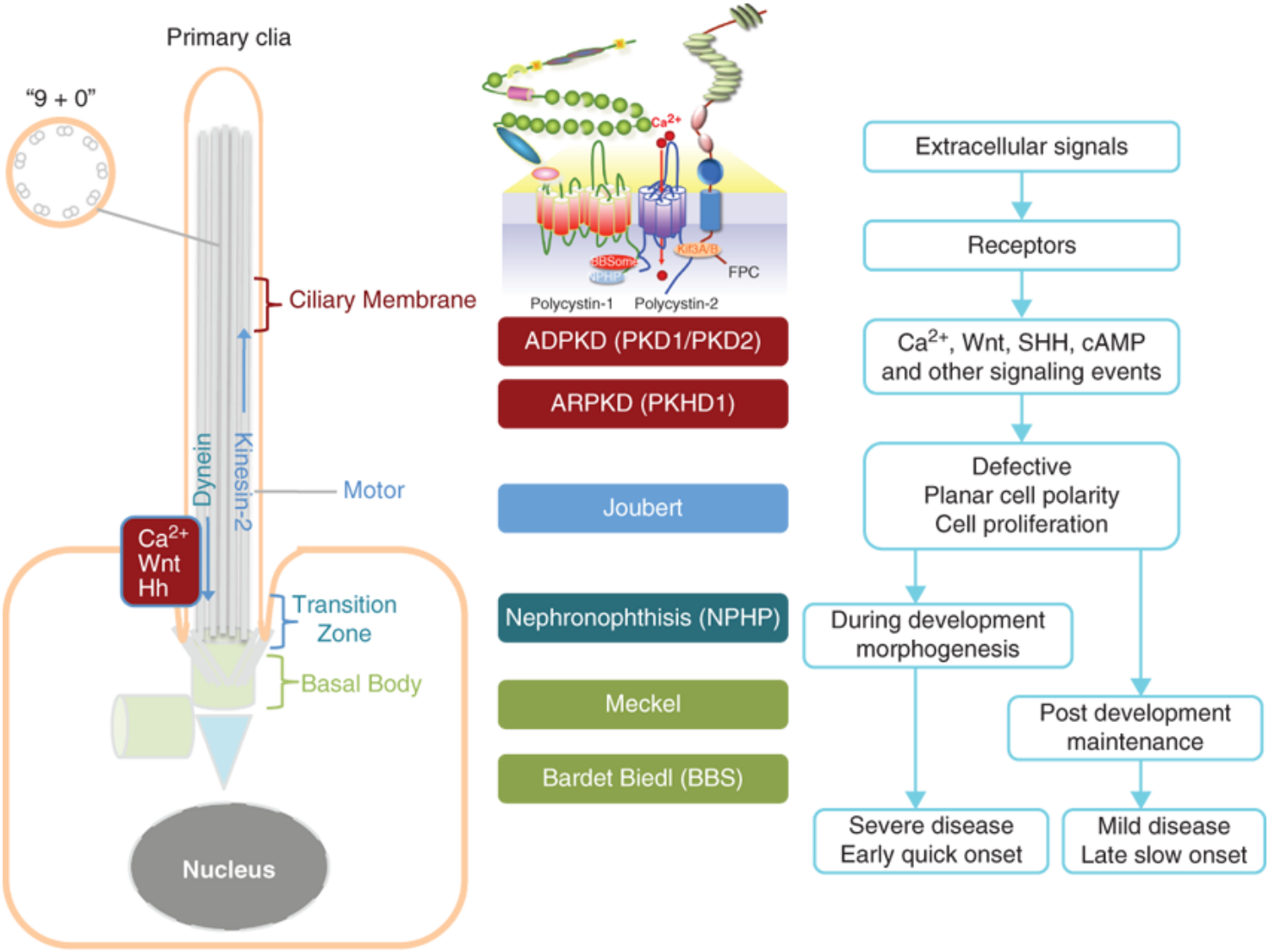
AUTOSOMAL DOMINANT POLYCYSTIC KIDNEY DISEASE

Etiology and Pathogenesis

ADPDK is characterized by progressive formation of epithelial lined cysts in the kidney **(Fig. 309-1)**. Although cysts only occur in 5% of the tubules in the kidney, the enormous growth of these cysts ultimately leads to the loss of normal surrounding tissues and loss of renal function. The cellular defects in ADPKD that have been known for a long time are increased cell proliferation and fluid secretion, decreased cell differentiation, and abnormal extracellular matrix. ADPKD is caused by mutations in *PKD1* and *PKD2* which, respectively, code for polycystin-1 (PC1) and polycystin-2 (PC2). PC1 is a large 11- transmembrane protein that functions like a G-protein coupled receptor. PC2 is a calcium-permeable six transmembrane protein that structurally belongs to the transient receptor potential (TRP) cation channel family. PC1 and PC2 are widely expressed in almost all tissues and organs. PC1 expression is high in development and low in the adult, whereas PC2 expression is relatively constant. PC1/2 are found on the primary cilium, a hair-like structure present on the apical membrane of a cell, in addition to the cell membranes and cell-cell junctions of tubular epithelial cells. Defects in the primary cilia are linked to a wide spectrum of human diseases, collectively termed ciliopathies. The most common phenotype shared by many ciliopathies is kidney cysts. PC1 and PC2 bind to each other via their respective C-terminal tails to form a receptor-channel complex and regulate each other's function. The PC1/2 protein complex serves as a mechanosensor or chemical sensor and regulates calcium and G-protein signaling. The PC1/2 protein complex may also directly regulate a number of cellular functions including the cell cycle, the actin cytoskeleton, planar cell polarity (PCP), and cell migration. This protein complex has also been implicated in regulating a number of signaling pathways, including Wnt, mammalian target of rapamycin (mTOR), STAT3, cMET, phosphoinositide 3-kinase (PI3K/Akt), G protein–coupled receptor (GPCR), and epidermal growth factor receptor (EGFR), as well as in the localization and activity of cystic fibrosis transmembrane conductance regulator (CFTR). One hypothesis is that loss of ciliary function of PC1 and PC2 leads to aberrant calcium signaling and a subsequent increase of adenylyl cyclase activity and decrease of phosphodiesterase activity, which, in turn, causes increased cellular cAMP. Increased cAMP promotes protein kinase A activity, among other effectors, and, in turn, leads to cyst growth by promoting proliferation and fluid secretion of cyst-lining cells through chloride and aquarporin channels in ADPKD kidneys.

FIGURE 309-1

Scheme of the primary cilium and cystic kidney disease proteins. ***Left:*** a scheme of the primary cilium. Primary cilia share a "9+0" organization of microtubule doublets. Proteins are transported into the cilium by motor protein kinesin 2 and transported out of the cilium by dynein. The cilium is connected to the basal body through the transition zone. ***Middle:*** topology of ADPKD and ARPKD proteins polycystin 1, polycystin 2, and FPC are shown. Localization of disease proteins in the cilium, the transition zone and the basal body are color coded. ***Right:*** potential disease mechanisms due to cilium mediated signaling events.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ADPKD is inherited as an autosomal dominant trait with complete penetrance, but variable expressivity. The disease affects all ethnic groups worldwide with an estimated prevalence of 1:1000 to 1:400. Only half of the patients with ADPKD are clinically diagnosed during their lifetime. ADPKD is genetically heterogeneous. The first disease gene (*PKD1*) was localized to the region of the alpha-globin gene on chromosome 16p13 in 1985, and a second disease gene (*PKD2*) locus was mapped to chromosome 4q21-q23 in 1993. Mutations of *PKD1* and *PKD2* are responsible for ~85% and ~15% of ADPKD cases, respectively. However, patients with *PKD2* mutations may be >15% because they tend to have milder clinical disease and, as a result, under-diagnosed. Embryonic lethality of *Pkd1* and *Pkd2* knockout mice suggest human homozygotes may be lethal, thus not clinically recognized.

PKD1 is comprised of 46 exons occupying ~52 kb of genomic DNA. It produces a ~14 kb transcript that encodes polycystin-1, a protein of ~4300 amino acids. A feature of the *PKD1* gene is that the 5' three-quarters of *PKD1* have been duplicated at six other

sites on chromosome 16p, and many of them produce mRNA transcripts, which provides a major challenge for genetic analysis of the duplicated region. *PKD2* is a single-copy gene with 15 exons producing a ~5.3 kb mRNA transcript that encodes polycystin-2, a protein of 968 amino acids. A third gene *GANAB*, encoding the glucosidase IIa subunit, was recently reported to cause ADPKD, but patients with mutations in this gene all appear to have polycystic liver disease, and their kidney disease is milder than that in classic ADPKD.

In ADPKD patients, every cell carries a germline mutant allele of either *PKD1* or *PKD2*. However, cysts develop in only a small fraction of the nephrons. Cysts are thought to originate from clonal growth of single cells that have received a somatic "second hit" mutation in the "normal" allele of the *PKD1* or *PKD2* gene. Accumulating evidence in mouse models now shows that partial loss of function of the second allele of *Pkd1* in a proliferative environment is sufficient for cystogenesis, suggesting that a critical amount of *PKD1* is needed in a cell. Somatic inactivation of the second allele of *Pkd1* in adult mice results in very slow onset of cyst development in the kidney, but a "third hit" such as an additional genetic or epigenetic event, the inactivation of a growth suppressor gene, the activation of a growth promoting gene(s), or an event like renal injury that activates the developmental program, may promote rapid cyst formation.

Clinical Manifestations

ADPKD is characterized by the progressive bilateral formation of renal cysts. Focal renal cysts are typically detected in affected subjects aged <30 years. Hundreds to thousands of cysts are usually present in the kidneys of most patients in the fifth decade (Fig. 309-2). Enlarged kidneys can each reach a fourfold increase in length, and weigh up to 20 times the normal weight. The clinical presentations of ADPKD are highly variable. While many patients are asymptomatic until the fourth to fifth decade of life and are diagnosed by incidental discoveries of hypertension or abdominal masses, back or flank pain is a frequent symptom in ~60% of patients with ADPKD. The pain may result from renal cyst infection, hemorrhage, or nephrolithiasis. Gross hematuria resulting from cyst rupture occurs in ~ 40% of patients during the course of their disease, and many of them will have recurrent episodes. Flank pain and hematuria may coexist if the cyst that ruptures is connected with the collecting system. Proteinuria is usually a minor feature of ADPKD. Infection is the second most common cause of death for patients with ADPKD. Up to half of patients with ADPKD will have one or more episodes of renal infection during their lifetime. An infected cyst and acute pyelonephritis are the most common renal infections often due to gram-negative bacteria, which are associated with fever and flank pain, with or without bacteremia. These complications and renal insufficiency often correlate with structural abnormality of the renal parenchyma. Kidney stones occur in ~20% of patients with ADPKD. Different from the general population, more than half of the stones in patients with ADPKD are composed of uric acid, with the remainder due to calcium oxalate. Distal acidification defects, abnormal ammonium transport, low urine pH, and hypocitraturia may be important in the pathogenesis of renal stones in ADPKD. Renal cell carcinoma is a rare complication of ADPKD with no apparent increased frequency compared to the general population. However, in ADPKD these tumors are more often bilateral at presentation, multicentric, and sarcomatoid in type. Radiological imaging is often not helpful in distinguishing cyst infection and cyst hemorrhage because of their complexity. CT scan and magnetic resonance imaging (MRI) are often useful in distinguishing a malignancy from a complex cyst. Cardiovascular complications are the major cause of mortality in patients with ADPKD. Hypertension is common, and typically occurs before any reduction in glomerular filtration rate (GFR). Hypertension is a risk factor for both cardiovascular and kidney disease progression in ADPKD. Notably, some normotensive patients with ADPKD may also have left ventricular hypertrophy. Hypertension in ADPKD may result from the increased activation of the renin-angiotensin-aldosterone system, increased sympathetic nerve activity, and impaired endothelial cilium function-dependent relaxation of small resistant blood vessels.

FIGURE 309-2

Photograph showing a kidney from a patient with autosomal dominant polycystic kidney disease. The kidney has been cut open to expose the parenchyma and internal aspects of cysts.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The progression of ADPKD has striking inter- and intrafamilial variability. The disease can present as early as *in utero*, but end-stage renal disease (ESRD) typically occurs in late middle age. Risk factors include early diagnosis of ADPKD, hypertension, gross hematuria, multiple pregnancies, and large kidney size. Liver cysts derived from the biliary epithelia are the most common extrarenal complication. Polycystic liver disease associated with ADPKD is different from autosomal dominant polycystic liver disease (ADPLD), which is caused by mutations in at least two distinct genes (*PRKCSH* and *SEC63*) and does not progress to renal failure. Massive polycystic liver disease occurs almost exclusively in women with ADPKD, particularly those with multiple pregnancies. Heterozygous loss-of-function variants in *PKHD1, ALG8, GANAB*, and *SEC61B* are now found in ADPLD. *ALG8, GANAB,* and *SEC61B*, all encode ER proteins that are involved in the same pathway as GIIβ and SEC63, and each appears to affect PC1 biogenesis.

Intracranial aneurysm (ICA) occurs four to five times more frequent in APDKD patients than that seen in the general population and cause high mortality. The disease gene products PC1 and PC2 may be directly responsible for defects in arterial smooth muscle cells and myofibroblasts. The focal nature and the natural history of ICA in ADPKD remain unclear. A family history of ICA is a risk factor of aneurysm rupture in ADPKD, whether hypertension and cigarette smoking are independent risk factors is not clear. About 20–50% of patients may experience “warning headaches” preceding the index episode of subarachnoid hemorrhage due to ruptured ICA. A CT scan is generally used as the first diagnostic test. A lumbar puncture may be used to

confirm the diagnosis. The role of radiological screening for ICA in asymptomatic patients with ADPKD remains unclear. ADPKD patients with a positive family history of ICAs may undergo pre-symptomatic screening of ICAs by MR angiography. Other vascular abnormalities in ADPKD patients include diffuse arterial dolichoectasias of the anterior and posterior cerebral circulation, which can predispose to arterial dissection and stroke. Mitral valve prolapse occurs in up to 30% of patients with ADPKD, and tricuspid valve prolapse is less common. Other valvular abnormalities occurring with increased frequency in ADPKD patients include insufficiency of the mitral, aortic, and tricuspid valves. Most patients are asymptomatic but some may progress and require valve replacement. The prevalence of colonic diverticulae and abdominal wall hernias are also increased in ADPKD patients.

Diagnosis

Diagnosis is typically made from a positive family history consistent with autosomal dominant inheritance and multiple kidney cysts bilaterally. Renal ultrasonography is often used for presymptomatic screening of at-risk subjects and for evaluation of potential living-related kidney donors from ADPKD families. The presence of *at least two renal cysts (unilateral or bilateral)* is sufficient for diagnosis among at-risk subjects between 15 and 29 years of age with a sensitivity value of 96% and specificity value of 100%. The presence of *at least two cysts in each kidney* and *at least four cysts in each kidney*, respectively, are required for the diagnosis among at-risk subjects aged 30–59 years and aged ≥60 years with a sensitivity value of 100% and specificity value of 100%. This is because there is an increased frequency of developing simple renal cysts with age. Conversely, in subjects aged between 30 and 59 years the absence of *at least two cysts in each kidney*, which is associated with a false negative rate of 0%, can be used for disease exclusion. These criteria have a lower sensitivity for patients with a *PKD2* mutation because a late onset of ADPKD2. CT scan and T2-MRI, with and without contrast enhancement, are more sensitive than ultrasonography and can detect cysts of smaller size. However, a CT scan exposes the patient to radiation and radiocontrast, which may cause serious allergic reactions and nephrotoxicity in patients with renal insufficiency. T2-MRI, with gadolinium as a contrast agent, has minimal renal toxicity and can detect cysts of only 2–3 mm in diameter. However, a large majority of cysts may still be below the detection level. Genetic testing by linkage analyses and mutational analyses are available for ambiguous cases. Because of the large size of *PKD1* gene and the presence of multiple highly homologous pseudogenes, mutational analysis of *PKD1* gene is difficult and costly. Application of new technologies such as paired-end next generation sequencing with multiplexing individually bar-coded long range PCR libraries may reduce the costs and improve the sensitivity for clinical genetic testing.

TREATMENT

Treatment

Autosomal Dominant Polycystic Kidney Disease

No specific treatment to prevent cyst growth or the decline of renal function has been approved by U.S. Food and Drug Administration. Blood pressure control to a target of 140/90 mmHg is recommended according to the guidelines from the eighth report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC VIII report) for reducing cardiovascular complications in ADPKD and renal disease progression. More rigorous blood pressure control does not equal greater clinical benefits. Maintaining a target systolic blood pressure to 110 mmHg in patients with moderate or advanced disease may increase the risk of renal disease progression by reducing renal blood flow. Lipid-soluble antibiotics against common gram-negative enteric organisms include trimethoprim-sulfamethoxazole, quinolones, and chloramphenicol, and are preferred for cyst infection because most renal cysts are not connected to glomerular filtration and antibiotics that are capable to penetrate the cyst walls are likely to be more effective. Treatment often requires 4–6 weeks. The treatment of kidney stones in ADPKD includes standard measures such as analgesics for pain relief, and hydration to ensure adequate urine flow. Management of chronic flank, back, or abdominal pain due to renal enlargement may include both pharmacologic (non-narcotic and narcotic analgesics) and non-pharmacological (transcutaneous electrical nerve stimulation, acupuncture, and biofeedback). Occasionally surgical decompression of cysts may be necessary. More than half of

ADPKD patients eventually require peritoneal dialysis, hemodialysis, or kidney transplantation. Peritoneal dialysis may not be suitable for some patients with massively enlarged polycystic kidneys due to the small intraabdominal space for efficient peritoneal exchange of fluid and solutes and increased chance of abdominal hernia and back pain. Patients with very large polycystic kidneys and recurrent renal cyst infection may require pretransplant nephrectomy or bilateral nephrectomy to accommodate the allograft and reduce the pain.

Specific treatment strategies to ADPKD have focused on slowing renal disease progression and lowering cardiovascular risk. For the latter, the main approach is to control blood pressure by inhibiting the renin-angiotensin-aldosterone system. The HALT PKD trial was set to evaluate the impact of intensive blockade of the renin-angiotensin-aldosterone system, and levels of blood pressure control on progressive renal disease. This trial found that rigorous blood-pressure control could slow cyst growth. Most approaches target the slowing of renal disease progression by inhibiting cell proliferation and fluid secretion. Several clinical trials have been conducted targeting cell proliferation: sirolimus and everolimus, inhibitors of the mammalian target of rapamycin (mTOR) pathway; OPC31260 and tolvaptan, which inhibits cyclic adenosine monophosphate (cAMP) pathways by antagonizing the activation of vasopressin V2 receptor (V2R) in collecting ducts and reduces cell proliferation by decreasing renal cAMP levels; and somatostatin analogues, which reduces cAMP levels by binding to several G-protein coupled receptors. The TAMPO and ALADIN trials showed that V2R antagonists and somatostatin analogues (octreotide-LAR groups) respectively slowed the decline of renal function. Some side effects, such as liver function impairment, polydipsia, and diarrhea, have been observed for tolvaptan and cholecystitis for octreotide-LAR. A recent report also showed that tolvaptan reduces renal pain. DIPAK, a small multi-center European study, showed that nerve block may be used to relieve pain in ADPKD patients suffering with refractory chronic pain. A combination of different growth inhibitors may enhance efficacy and reduce side effects.

Additional preclinical studies in animal models include the use of inhibitors to nonreceptor tyrosine kinase Src, B-raf, cycline-dependent kinase (CDK), transcription factors STAT3 and STAT6 (pyrimethamine and leflunomide), purinergic receptors, hepatocyte growth factor receptor, glucosylceramide, and agonists to peroxisome proliferator-activated receptor-gamma (PPARγ) receptors (thiazolidodinediones). Recently, several microRNAs have been identified that mediate disease progression, which may prove to be a new therapeutic target. Food restriction in mouse models of the disease was reported to reduce cyst area, kidney fibrosis, inflammation, and injury. Branched chain amino acids appear to enhance cyst development in a mouse model.

AUTOSOMAL RECESSIVE POLYCYSTIC KIDNEY DISEASE

Genetic Considerations



ARPKD is a significant hereditary renal disease in childhood, with an estimated prevalence of 1 in 20,000 live births. A carrier frequency of up to 1:70 has been reported. Mutations in a single gene, *PKHD1*, are responsible for all the clinical presentations of ARPKD. *PKHD1*, localized on human chromosome region 6p21.1-6p12.2, is one of the largest genes in the genome, occupies ~450 kb of DNA, and contains at least 86 exons. It produces multiple alternatively spliced transcripts. The largest transcript encodes fibrocystin/polyductin (FPC), which is a large receptor-like integral membrane protein of 4074 amino acids. FPC has a single transmembrane, a large N-terminal extracellular region, and a short intracellular cytoplasmic domain. FPC is localized on the primary cilia of epithelia cells of cortical and medullary collecting ducts and cholangiocytes of bile ducts, similar to polycystins and several other ciliopathy proteins. FPC is also expressed on the basal body and plasma membrane. The large extracellular domain of FPC is presumed to bind to an as yet unknown ligand(s), and is involved in cell-cell and cell-matrix interactions. FPC interacts with ADPKD protein PC2, and may also participate in regulation of the mechanosensory function of the primary cilia, calcium signaling, and PCP, suggesting a common mechanism underlying cystogenesis between ADPKD and ARPKD. FPC is also found on the centrosomes and mitotic spindle, and may regulate centrosome duplication and mitotic spindle assembly during cell division. A large number of various mutations have been

found throughout *PKHD1*, and are unique to individual families. Most patients are compound heterozygotes for *PKHD1* mutations. Patients with two truncation mutations appear to have an earlier onset of the disease.

Clinical Features

Classic ARPKD is generally diagnosed in utero or within the neonatal period, and characterized by greatly enlarged echogenic kidneys in diseased fetuses. Reduced fetal urine production may contribute to oligohydroaminos and pulmonary hypoplasia. About 30% of affected neonates die shortly after birth due to respiratory insufficiency. Close to 60% of mortality occurs within the first month of life. In the classic group, most patients are born with renal insufficiency and ESRD. However, infants often have a transient improvement in their GFR; death from renal insufficiency at this stage is rare. Some patients are diagnosed after the neonatal stage, which form the older group. Morbidity and mortality in this group often involve systemic hypertension, progressive renal insufficiency, and liver manifestations. The hallmarks of ARPKD liver disease are biliary dysgenesis due to a primary ductal plate malformation with associated periportal fibrosis, namely congenital hepatic fibrosis (CHF) and dilatation of intrahepatic bile ducts (Caroli disease). CHF and Caroli disease can then lead to portal hypertension exhibiting hepatosplenomegaly, variceal bleeding, and cholangitis. Some patients with the diagnosis of ARPKD at 1 year of age with nephromegaly exhibit slowly declining renal function over 20 years with only minimally enlarged kidneys at ESRD, and markedly atrophic kidneys following renal transplantation. The slow progression of renal disease is likely due to increasing fibrosis rather than the development of cysts. Systemic hypertension is common in all ARPKD patients, even those with normal renal function.

Diagnosis

Ultrasonography, CT, and MRI all can be used for diagnosis. Ultrasonography reveals large, echogenic kidneys with poor corticomedullary differentiation. The diagnosis can be made in utero after 24 weeks of gestation in severe cases. Macrocysts generally are not common at birth in ARPKD patients. The absence of renal cysts in either parent, particularly if they are >40 years of age on ultrasonography helps distinguish ARPKD from ADPKD in older patients. Clinical, laboratory, or radiographic evidence of hepatic fibrosis, hepatic pathology demonstrating characteristic ductal plate abnormalities, family history of affected siblings, or parental consanguinity suggestive of autosomal recessive inheritance is helpful. The lack of mutational hot spots and the large and complex genomic structure of *PKHD1* make molecular diagnosis difficult, however, presymptomatic screen of other at-risk members in a family with already identified ARPKD mutations is straightforward and inexpensive.

TREATMENT

Treatment

Autosomal Recessive Polycystic Kidney Disease

There is no specific therapy for ARPKD. Appropriate neonatal intensive care, blood pressure control, dialysis, and kidney transplantation increase survival into adulthood. Complications of hepatic fibrosis may necessitate liver transplantation. Patients with severe Caroli disease may need porto-systemic shunting. Upcoming therapies may target abnormal cell signaling mechanisms, as described above for ADPKD.

OTHER DISEASES CHARACTERIZED BY LARGE KIDNEY CYSTS

TUBEROUS SCLEROSIS

Tuberous sclerosis (TS) is a rare autosomal dominant syndrome caused by mutations in one of two genes, *TSC1*, encoding hamartin, or, *TSC2*, encoding tuberin. Published estimates of prevalence vary widely, but it certainly occurs in <1:5,000 births. Kidney cysts are a frequent feature of this condition, as are two other abnormalities of kidney growth, renal cell carcinoma and

renal angiomyolipomas. TS is a syndrome affecting multiple organ systems. Other features of TS include benign growths in the nervous system, eyes, heart, lung, liver, and the skin. Essentially all TS patients have such skin lesions, and a large proportion of patients have neurologic and cognitive manifestations. The *TSC2* gene is adjacent to *PKD1* in the human genome. Some patients have deletions in their genomic DNA that inactivate these two genes. Such individuals may have manifestations of both ADPKD and TS.

The most common kidney finding in TS is the presence of angiomyolipomas. These growths tend to be multiple and bilateral. While they are usually benign, they may bleed. Surgical removal is often recommended as a prophylactic measure in people with angiomyolipomas >4 cm in diameter. The cysts in TS are radiographically similar to those seen in ADPKD. In contrast to ADPKD, there is a clearly increased risk of renal cell carcinoma in TS patients. Regular periodic imaging is recommended in TS patients with kidney involvement to screen for the development of renal cell carcinoma. These cysts may rarely become large and hemorrhagic, occasionally requiring nephrectomy when nephron-sparing surgery is not possible.

Although not common, TS may lead to significant chronic kidney disease (CKD) and progress to end-stage kidney failure. Patients with TS and CKD typically have an unremarkable urine sediment and only minimal to mild amounts of proteinuria.

Mechanistically, the *TSC1* and *TSC2* gene products tuberin and hamartin interact physically. This protein complex is localized to the base of the cilia and inhibits intracellular signaling processes mediated by mTOR (mammalian target of rapamycin), leading to abnormal growth in a number of tissues. Investigation of mTOR inhibitors as therapy for TS is ongoing. There is increasing optimism that this class of drugs will become commonplace for prevention of the renal and non-renal manifestations of TS.

VON HIPPEL-LINDAU DISEASE

Von Hippel-Lindau disease (VHL) is an inherited cancer syndrome with renal manifestations. VHL is an autosomal dominant condition caused by mutations in the VHL tumor-suppressor gene. VHL is localized to the primary cilia and is necessary for the formation of primary cilia. Like many autosomal dominant cancer syndromes, VHL is recessive at the cellular level: a somatic mutation in the second VHL allele leads to loss of VHL in the cell and abnormal growth. Kidney manifestations of VHL include multiple bilateral kidney cysts, and renal cell carcinomas. Kidney cysts and carcinoma affects the majority of VHL patients. Non-renal features of VHL include pheochromocytomas, cerebellar hemangioblastomas, and retinal hemangiomas. While much rarer than ADPKD, it is important for this entity to be considered in the differential diagnosis of an individual with newly recognized kidney cysts.

In these patients, annual screening of the kidneys by imaging with CT or MRI scanning is recommended for early detection of renal cell carcinomas. Increasingly, nephron-sparing surgical approaches are being used for removal of cancerous lesions in order to preserve kidney function.

OTHER INHERITED DISEASES OF TUBULE GROWTH AND DEVELOPMENT

ADPKD is by far the most common adult onset single gene form of adult onset kidney disease. The large cysts that are sometimes seen in VHL and TS are similar in appearance to the cysts seen in ADPKD. A variety of other inherited disorders affecting primarily tubule and renal interstitial function can lead to CKD and eventual end-stage kidney disease in the absence of large tubule-derived cysts.

Inherited diseases affecting the tubulointerstitial compartment of the kidney can lead to secondary glomerular stress and glomerulosclerosis with some degree of concomitant proteinuria. Similarly, disorders of glomerular function will typically lead to secondary interstitial fibrosis and tubule atrophy. From a clinical perspective, therefore, distinguishing between a genetic disease of the renal tubules and a disease of the glomerulus may not be easy, particularly in the absence of a gross phenotype such as large kidney cysts.

AUTOSOMAL DOMINANT INTERSTITIAL KIDNEY DISEASE (MEDULLARY CYSTIC KIDNEY DISEASE)

The medullary cystic kidney diseases (MCKD) are autosomal dominant disorders. The term autosomal dominant tubulointerstitial kidney disease (ADTKD) is replacing MCKD as the preferred designation. Despite the old nosology, kidney cysts are not invariably present. Older literature often grouped MCKD together with the childhood-onset disorders known as the nephronophthises, but these are distinct clinical and genetic entities.

ADTKD-MUC1

Patients with medullary cystic kidney disease type I (MCKD I) have mutations in the mucin 1 gene *MUC1*. In contrast to MCKD II patients, individuals with MCKD I do not have elevated uric acid levels. The disease-causing *MUC1* mutations that have been reported all alter a highly repetitive region within the *MUC1* gene, leading to a large “neoprotein” fragment that may lead to toxic effects on the kidney tubule.

Clinically, patients with MCKD I exhibit slowly progressive CKD in adulthood, with only minimal amounts of increased urine protein and occasional renal cysts seen on ultrasound examination. Kidney histology shows tubulointerstitial fibrosis and tubular atrophy. The mechanisms by which *MUC1* mutations cause human kidney disease are not known. Disease does not recur in transplanted kidneys.

ADTKD-UMOD

ADTKD-IUMOD (also called MCKD II) is caused by mutations in the *UMOD* gene, which encodes the protein uromodulin, also known as Tamm-Horsfall protein. Uromodulin is also found on the centrosome, the mitotic spindle, and the primary cilia; it colocalizes with nephrocystin-1 and KIF3A on the cilia. *UMOD* mutations also cause the conditions that have been referred to as familial juvenile hyperuricemic nephropathy (HNFJ1) and glomerulocystic kidney disease (GCKD), although it is not clear that these different names represent clearly distinct disorders. The term *uromodulin-associated kidney disease* (or UAKD) has been suggested as a better name for MCKD II and the various other related *UMOD*-associated diseases. Despite the name, kidney cysts are not a common feature of MCKD II. MCKD II should be suspected clinically in patients with a family history of late onset kidney disease, benign urine sediments, absence of significant proteinuria, and hyperuricemia. Large genome-wide association studies have suggested that certain common non-coding sequence variants in *UMOD* are associated with a moderately increased risk of CKD in the general population. *UMOD*-associated disease is often associated with gout.

Other Forms of Familial Tubulointerstitial Kidney Disease

A small number of families have been identified with autosomal dominant tubulointerstitial kidney disease and hyperuricemia who lack *UMOD* mutations. Some of these families carry disease-segregating mutations in the renin gene *REN* (disease designation ADTKD-REN). ADKTKD-REN patients demonstrate hyporeninemia with mild hyperkalemia, and often have hyperuricemia and gout. There are other families who lack mutations in *UMOD, MUC1,* or *REN* mutation. Thus, mutations in other yet-to-be identified genes are able to produce similar interstitial kidney disease, both with and without hyperuricemia.

Kidney biopsies in patients with any of various forms of MCKD typically show interstitial fibrosis. These histologic features are not diagnostic of any particular genetic entity, and the specific diagnosis must be made by other means. Genetic tests for alterations in specific genes are increasingly available in the clinical setting.

Those patients with autosomal dominant interstitial kidney disease, *UMOD* or *REN* mutations, with hyperuricemia and gout should be treated similarly to others with these findings, with uric-acid lowering agents, such as allopurinol or febuxostat.

NEPHRONOPHTHISIS

A large and growing number of genetically distinct but related set of autosomal recessive disorders are referred to as nephronophthisis, or nephronophthisis-related ciliopathies. These entities should not be confused with the adult onset autosomal dominant MCKD discussed above, despite the often confusing nomenclature seen in older medical literature. Each of the individual forms of nephronophthisis is quite rare, but together this category constitutes the most common inherited childhood form of kidney failure requiring kidney replacement therapy.

Like ADPKD and ARPKD, the various genetically heterogeneous entities that fall under the category of nephronophthisis (NPHP) are disorders of ciliary function. Mutations in >90 genes have been identified that lead to NPHP under an autosomal recessive pattern of inheritance. Some of these gene defects cause limited kidney disease, while many cause ciliopathies characterized by multiple organ involvement. The various forms of NPHP share common features, including tubulointerstitial fibrosis, corticomedullary cysts, and progressive CKD, leading to renal failure. Proteinuria is absent or mild, and the urine sediment is not active.

NPHP is often divided into infantile, juvenile, and adolescent forms. The juvenile form is the most frequent, and usually caused by mutations in the *NPHP2* gene. The infantile form, usually caused by *NPHP2* mutations, is associated with end-stage kidney failure in early childhood. Patients with the adolescent form of NPHP typically develop end–stage kidney failure in early adulthood. Hypertension, if present, tends to be a late finding in the course of the NPHPs. The products of the NPHP genes are referred to as nephrocystins. *NPHP1* through *NPHP20* have been reported; some are referred to by other names, as well.

NPHP can present as an isolated finding, or be part of several multi-organ syndromes. Neurologic abnormalities are present in a significant number of patients. Bone and liver abnormities are seen in some NPHP patients. Senior-Loken syndrome is defined by the presence of NPHP with retinitis pigmentosa. Joubert syndrome is defined by multiple neurologic findings, including hypoplasia of the cerebellar vermis. Some forms of this genetically heterogeneous syndrome include NPHP as a component.

The multisystem disease Bardet-Biedl syndrome (BBS) is defined clinically by a spectrum of features, including truncal obesity, cognitive impairment, retinal dystrophy, polydactyly, developmental urogenital abnormalities, and kidney cysts. The kidney phenotype is NPHP-like, with small cysts deriving from the tubules, tubulointerstitial and often secondary glomerular disease, and urine concentrating defects. There are 19 BBS genes cloned. BBS follows autosomal recessive inheritance. Like ADPKD, ARPKD, and NPHP, BBS is a disease of abnormal ciliary function.

The multiple genes and gene products (nephrocystins) that are responsible for NPHP are expressed in cilia, basal bodies, and the centrosomes of kidney tubules cells. It has been hypothesized that all of the NPHP gene defects lead to a clinical phenotype by interfering with the regulation of PCP.

There are no specific clinical tests that define nephronophthisis. Genetic diagnosis is possible, cumbersome because of the large number of genes that can be responsible, but increasingly feasible due to new DNA sequencing technologies. There are no specific therapies for NPHP. Rather, therapy is aimed at treating signs of these diseases as well as those systemic abnormalities seen with all CKDs. Chronic dialysis or kidney transplantation are eventually required for NPHP-affected individuals.

KARYOMEGALIC TUBULOINTERSTITIAL NEPHRITIS

Karyomegalic tubulointerstitial nephritis is an exceptionally rare form of kidney disease with adult-onset progressive kidney failure. Kidney biopsy shows chronic tubulointerstitial nephritis, as well as interstitial fibrosis. This is a recessive disorder caused by inheritance of two mutant copies of the *FAN1* gene. *FAN1* encodes a component of a DNA repair machinery complex. Individuals with two mutant *FAN1* gene are genetically sensitized to the effect of DNA damage. Kidney histology shows karyomegaly in addition to the non-specific findings of interstitial fibrosis and tubular atrophy.

MEDULLARY SPONGE KIDNEY

Medullary sponge kidney (MSK) is often grouped together with inherited disorders of the kidney affecting tubule growth and development, although it is usually a sporadic finding rather than an inherited phenotype. MSK is caused by developmental malformation and cystic dilatation of the renal collecting ducts. The medullary cysts seen in this entity can be quite variable in size.

MSK is usually a benign entity. The diagnosis of MSK is often made incidentally. In the past, the diagnosis of MSK was often made by intravenous pyelography (IVP). CT urography, which has replaced IVPs for much routine kidney imaging, is not as sensitive in detecting MSK.

MSK is associated with an increased frequency of calcium phosphate and calcium oxalate kidney stones. Altered flow characteristics in the kidney tubules may lead to the development of formation of a nidus for stone formation. Kidney stones in this group are treated the same as are kidney stones in the general population. MSK patients also often exhibit reduced kidney concentrating ability and an increased frequency of urinary tract infections.

CAKUT

The structural abnormalities known as CAKUT (Congenital Abnormalities of the Kidney and Urinary Tract) are a group of etiologically and phenotypically heterogeneous disorders. Some form of CAKUT is estimated to occur in up to 1 in 500 live births. Specific abnormalities classified as part of the CAKUT spectrum include kidney hypoplasia, kidney agenesis, ureteropelvic junction obstruction, and vesicoureteral reflux.

CAKUT can be the cause of clinically significant problems in both adults and children. However, it is a major contributor to kidney failure in children, accounting for more than one-third of end-stage kidney disease in this group.

CAKUT is typically a sporadic finding, but can also cluster in families. Familial forms can be observed as parts of multisystem developmental syndromes. A growing list of specific genes have been identified which when mutated lead to syndromic forms of CAKUT. For example, the branchio-oto-renal syndrome, characterized by developmental abnormalities in the neck, ears, and kidney, can be caused by mutations in the *EYA1* and *SIX1* genes. Mutations in the *PAX2* transcription factor gene can cause the autosomal dominant renal coloboma syndrome, characterized by optic nerve malformations and hypoplastic kidneys. Recent work has demonstrated that a non-trivial fraction of children with CKD have an unsuspected genomic imbalance, often disrupting genes known to relevant to CAKUT and kidney development. It is not uncommon for such genetic lesions that affect both kidney and neurocognitive function.

In many instances, CAKUT is caused by environmental influences rather than genetic alterations. For example, renal tubular dysgenesis, defined by altered tubule development, can be caused by prenatal exposure of angiotensin converting enzyme inhibitors or angiotensin receptor blockers.

MITOCHONDRIAL DISEASE

Inherited disorders of the mitochondrial genome (discussed elsewhere in this text [**see also** Chap. 472]) commonly affect kidney function. Thirteen of the genes involved in encoding components of the mitochondrial respiratory chain are located on the mitochondrial genome that is inherited maternally. The remainder of these components is encoded by the nuclear genome. These defects of oxidative phosphorylation may affect multiple organs and tissues.

Neuromuscular disease is the best recognized part of this complex phenotype. Kidney disease is now recognized as a common component, as well. Tubulointerstitial disease may be seen on kidney biopsy, and progression to kidney failure may occur. Glomerular involvement, manifest as proteinuria and glomerulosclerosis, can also develop. Changes in proximal tubule

activity are the most common renal phenotype. Patients may have several defects in proximal tubule transport, including the Fanconi syndrome. Some patients may also have acidosis, hypophosphatemic rickets, hypercalciuria, glycosuria, and tubular proteinuria. Decreased urine concentrating ability is common.

GLOBAL CONSIDERATIONS


The disorders discussed above are all seen worldwide. In addition, a previously unrecognized epidemic of kidney disease is leading to very high rates of kidney failure in and near the western coast of Central America. This mesoamerican nephropathy is particularly common in Nicaragua and El Salvador. Mesoamerican nephropathy patients do not have significant proteinuria, suggesting that this is a disease of the kidney tubules and interstitium. The cause is unknown, but some have suggested that a combination of toxic environmental factors and heat stress underlie the development of this kidney disease, which has a striking male predominance. However, the fact that in many families, a large fraction of the men are affected with kidney disease has suggested that a strong genetic component is involved, as well.

FURTHER READING

Arts HH, Knoers NV: Current insights into renal ciliopathies: What can genetics teach us? Pediatr Nephrol 28:863, 2013. [PubMed: 22829176]

Bleyer AJ et al: Autosomal dominant tubulointerstitial kidney disease. Adv Chronic Kidney Dis 24:86, 2017. [PubMed: 28284384]

Grantham JJ et al: Detected renal cysts are tips of the iceberg in adults with ADPKD. Clin J Am Soc Nephrol 7:1087, 2012. [PubMed: 22580785]

Henske EP: Tuberous sclerosis and the kidney: From mesenchyme to epithelium, and beyond. Pediatr Nephrol 20:854, 2005. [PubMed: 15856327]

LaRiviere WB et al: Novel therapeutic approaches to autosomal dominant polycystic kidney disease. Transl Res 165:488, 2015. [PubMed: 25438190]

Ong AC, Harris PC: A polycystin-centric view of cyst formation and disease: The polycystins revisited. Kidney Int 88:699, 2015. [PubMed: 26200945]

Porath B et al: Mutations in GANAB, encoding the glucosidase IIalpha subunit, cause autosomal-dominant polycystic kidney and liver disease. Am J Hum Genet 98:1193, 2016. [PubMed: 27259053]

Reddy BV, Chapman AB: The spectrum of autosomal dominant polycystic kidney disease in children and adolescents. Pediatr Nephrol 32:31, 2017. [PubMed: 27034070]

Vivante A, Hildebrandt F: Exploring the genetic basis of early-onset chronic kidney disease. Nat Rev Nephrol 12:133, 2016. [PubMed: 26750453]

Zhou J: Polycystins and primary cilia, primers for cell cycle progression. Ann Rev Physiol 71:83, 2009.

Chapter 310: Tubulointerstitial Diseases of the Kidney

Laurence H. Beck Jr.; David J. Salant

INTRODUCTION

Inflammation or fibrosis of the renal interstitium and atrophy of the tubular compartment are common consequences of diseases that target the glomeruli or vasculature. Distinct from these secondary phenomena, however, are a group of disorders that primarily affect the tubules and interstitium, with relative sparing of the glomeruli and renal vessels. Such disorders are conveniently divided into acute and chronic tubulointerstitial nephritis (TIN) **(Table 310-1)**.

TABLE 310-1

Classification of the Causes of Tubulointerstitial Diseases of the Kidney

Acute Tubulointerstitial Disorders

Acute Interstitial Nephritis

Therapeutic agents

- Antibiotics (β-lactams, sulfonamides, quinolones, vancomycin, erythromycin, linezolid, minocycline, rifampin, ethambutol, acyclovir)
- Nonsteroidal anti-inflammatory drugs, COX-2 inhibitors
- Diuretics (rarely thiazides, loop diuretics, triamterene)
- Anticonvulsants (phenytoin, valproate, carbamazepine, phenobarbital)
- Miscellaneous (proton pump inhibitors, H_2 blockers, captopril, mesalazine, indinavir, allopurinol, lenalidomide)

Infection

- Bacteria (*Streptococcus, Staphylococcus, Legionella, Salmonella, Brucella, Yersinia, Corynebacterium diphtheriae*)
- Viruses (EBV, CMV, hantavirus, polyomavirus, HIV)
- Miscellaneous (*Leptospira, Rickettsia, Mycoplasma, Histoplasma*)

Autoimmune

- Tubulointerstitial nephritis with uveitis (TINU)
- Sjögren's syndrome
- Systemic lupus erythematosus
- Granulomatous interstitial nephritis
- IgG4-related systemic disease
- Idiopathic autoimmune interstitial nephritis

Acute obstructive disorders

- Light chain cast nephropathy ("myeloma kidney")
- Acute phosphate nephropathy
- Acute urate nephropathy

Chronic Tubulointerstitial Disorders

- Vesicoureteral reflux/reflux nephropathy
- Sickle cell disease
- Chronic exposure to toxins or therapeutic agents
- Analgesics, especially those containing phenacetin
- Lithium

- Heavy metals (lead, cadmium)
- Aristolochic acid (Chinese herbal and Balkan endemic nephropathies)
- Calcineurin inhibitors (cyclosporine, tacrolimus)

Metabolic Disturbances

- Hypercalcemia and/or nephrocalcinosis
- Hyperuricemia
- Prolonged hypokalemia
- Hyperoxaluria
- Cystinosis **(see Chap. 309)**

Cystic and Hereditary Disorders (see Chap. 309)

- Polycystic kidney disease
- Nephronophthisis
- Adult medullary cystic disease
- Medullary sponge kidney

Miscellaneous

- Aging
- Chronic glomerulonephritis
- Chronic urinary tract obstruction
- Ischemia and vascular disease
- Radiation nephritis (rare)

Abbreviations: CMV, cytomegalovirus; COX, cyclooxygenase; EBV, Epstein-Barr virus.

Acute TIN most often presents with acute renal failure **(Chap. 304)**. The acute nature of this group of disorders may be caused by aggressive inflammatory infiltrates that lead to tissue edema, tubular cell injury, and compromised tubular flow, or by frank obstruction of the tubules with casts, cellular debris, or crystals. There is sometimes flank pain due to distention of the renal capsule. Urinary sediment is often active with leukocytes and cellular casts, but depends on the exact nature of the disorder in question.

The clinical features of chronic TIN are more indolent and may manifest with disorders of tubular function, including polyuria from impaired concentrating ability (nephrogenic diabetes insipidus), defective proximal tubular reabsorption leading to features of Fanconi's syndrome (glycosuria, phosphaturia, aminoaciduria, hypokalemia, and type II renal tubular acidosis [RTA] from bicarbonaturia), or non-anion-gap metabolic

acidosis and hyperkalemia (type IV RTA) due to impaired ammoniagenesis, as well as progressive azotemia (rising creatinine and blood urea nitrogen [BUN]). There is often modest proteinuria (rarely >2 g/d) attributable to decreased tubular reabsorption of filtered proteins; however, nephrotic-range albuminuria may occur in some conditions due to the development of secondary focal segmental glomerulosclerosis (FSGS). Renal ultrasonography may reveal changes of "medical renal disease," such as increased echogenicity of the renal parenchyma with loss of corticomedullary differentiation, prominence of the renal pyramids, and cortical scarring in some conditions. The predominant pathology in chronic TIN is interstitial fibrosis with patchy mononuclear cell infiltration and widespread tubular atrophy, luminal dilation, and thickening of tubular basement membranes. Because of the nonspecific nature of the histopathology, biopsy specimens rarely provide a specific diagnosis. Thus, diagnosis relies on careful analysis of history, drug or toxin exposure, associated symptoms, and imaging studies.

ACUTE INTERSTITIAL NEPHRITIS

In 1897, Councilman reported on eight cases of acute interstitial nephritis (AIN) in the Medical and Surgical Reports of the Boston City Hospital; three as a postinfectious complication of scarlet fever and two from diphtheria. Later, he described the lesion as "an acute inflammation of the kidney characterized by cellular and fluid exudation in the interstitial tissue, accompanied by, but not dependant on, degeneration of the epithelium; the exudation is not purulent in character, and the lesions may be both diffuse and focal." Today AIN is far more often encountered as an allergic reaction to a drug (Table 310-1). Immune-mediated AIN may also occur as part of a known autoimmune syndrome, but in some cases there is no identifiable cause despite features suggestive of an immunologic etiology (Table 310-1).

ALLERGIC INTERSTITIAL NEPHRITIS

Although biopsy-proven AIN accounts for no more than ~15% of cases of unexplained acute renal failure, this is likely a substantial underestimate of the true incidence. This is because potentially offending medications are more often identified and empirically discontinued in a patient noted to have a rising serum creatinine, without the benefit of a renal biopsy to establish the diagnosis of AIN.

Clinical Features

The classic presentation of AIN, namely, fever, rash, peripheral eosinophilia, and oliguric renal failure occurring after 7–10 days of treatment with methicillin or another β-lactam antibiotic, is the exception rather than the rule. More often, patients are found incidentally to have a rising serum creatinine or present with symptoms attributable to acute renal failure **(Chap. 304)**. Atypical reactions can occur, most notably nonsteroidal anti-inflammatory drug (NSAID)-induced AIN, in which fever, rash, and eosinophilia are rare, but acute renal failure with heavy proteinuria is common. A particularly severe and rapid-onset AIN may occur upon reintroduction of rifampin after a drug-free period. More insidious reactions to the agents listed in Table 310-1 may lead to progressive tubulointerstitial damage. Examples include proton pump inhibitors and, rarely, sulfonamide and 5-aminosalicylate (mesalazine and sulfasalazine) derivatives and antiretrovirals. It is

not clear if the recent association of proton pump inhibitors with incident chronic kidney disease involves an intermediate step of prolonged, subclinical interstitial nephritis.

Diagnosis

Finding otherwise unexplained renal failure with or without oliguria and exposure to a potentially offending agent usually points to the diagnosis. Peripheral blood eosinophilia adds supporting evidence but is present in only a minority of patients. Urinalysis reveals pyuria with white blood cell casts and hematuria. Urinary eosinophils are neither sensitive nor specific for AIN; therefore, testing is not recommended. Renal biopsy is generally not required for diagnosis but reveals extensive interstitial and tubular infiltration of leukocytes, including eosinophils.

TREATMENT

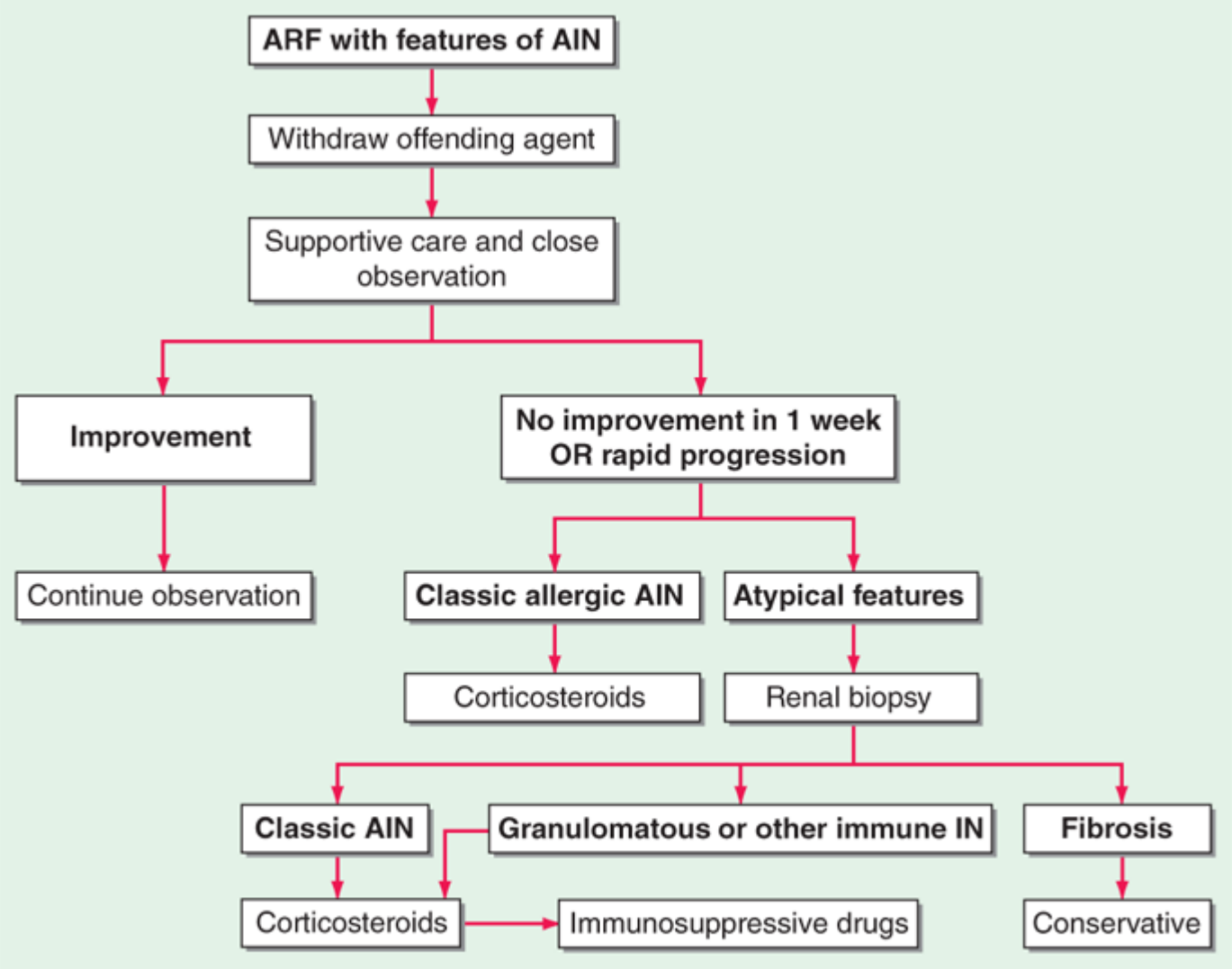
TREATMENT

Allergic Interstitial Nephritis

Discontinuation of the offending agent often leads to reversal of the renal injury. However, depending on the duration of exposure and degree of tubular atrophy and interstitial fibrosis that has occurred, the renal damage may not be completely reversible. Glucocorticoid therapy may accelerate renal recovery, but does not appear to impact long-term renal survival. It is best reserved for those cases with severe renal failure in which dialysis is imminent or if renal function continues to deteriorate despite stopping the offending drug (Fig. 310-1 and Table 310-2).

FIGURE 310-1

Algorithm for the treatment of allergic and other immune-mediated acute interstitial nephritis (AIN). ARF, acute renal failure; IN, interstitial nephritis. See text for immunosuppressive drugs used for refractory or relapsing AIN. *(Modified from S Reddy, DJ Salant: Ren Fail 20:829, 1998.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 310-2

Indications for Corticosteroids and Immunosuppressives in Interstitial Nephritis

Absolute Indications
• Sjögren's syndrome • Sarcoidosis • SLE interstitial nephritis • Adults with TINU • Idiopathic and other granulomatous interstitial nephritis
Relative Indications
• Drug-induced or idiopathic AIN with: Rapid progression of renal failure Diffuse infiltrates on biopsy Impending need for dialysis Delayed recovery • Children with TINU • Postinfectious AIN with delayed recovery (?)

Abbreviations: AIN, acute interstitial nephritis; SLE, systemic lupus erythematosus; TINU, tubulointerstitial nephritis with uveitis.

Source: Modified from S Reddy, DJ Salant: Ren Fail 20:829, 1998.

SJÖGREN'S SYNDROME

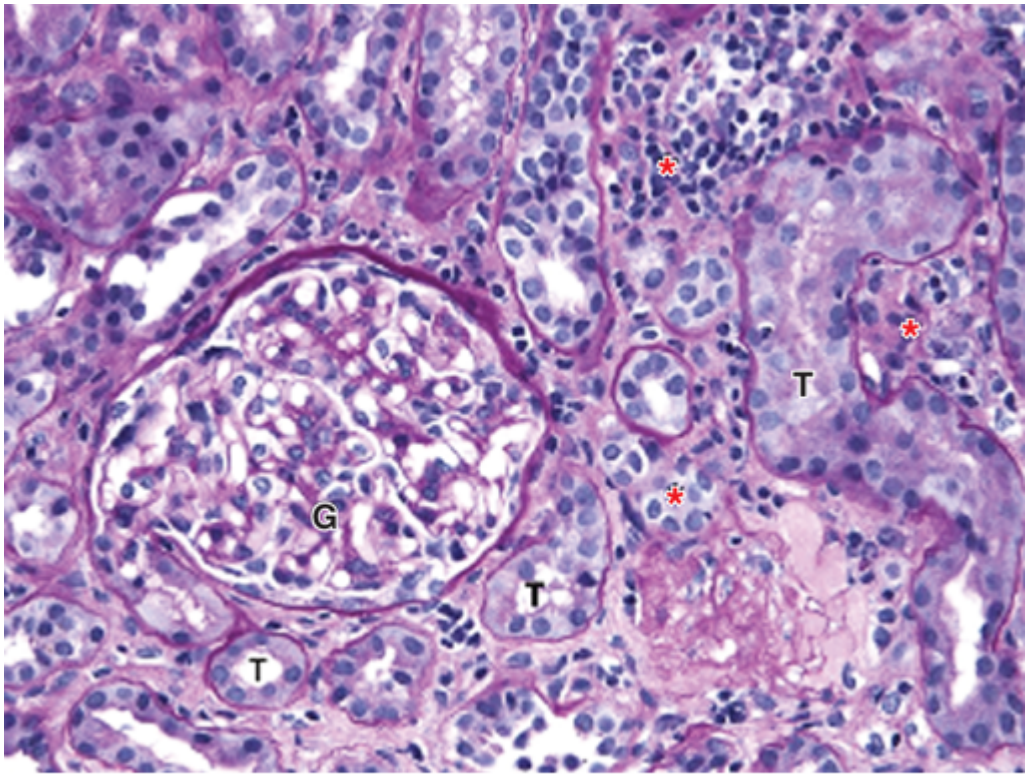
Sjögren's syndrome is a systemic autoimmune disorder that primarily targets the exocrine glands, especially the lacrimal and salivary glands, and thus results in symptoms, such as dry eyes and mouth, that constitute the "sicca syndrome" **(Chap. 354)**. TIN with a predominant lymphocytic infiltrate is the most common renal manifestation of Sjögren's syndrome and can be associated with distal RTA, nephrogenic diabetes insipidus, and moderate renal failure. Diagnosis is strongly supported by positive serologic testing for anti-Ro (SS-A) and anti-La (SS-B) antibodies. A large proportion of patients with Sjögren's syndrome also have polyclonal hypergammaglobulinemia. Treatment is initially with glucocorticoids, although patients may require maintenance therapy with azathioprine or mycophenolate mofetil to prevent relapse (Fig. 310-1 and Table 310-2).

TUBULOINTERSTITIAL NEPHRITIS WITH UVEITIS (TINU)

TINU is a systemic autoimmune disease of unknown etiology. It accounts for fewer than 5% of all cases of AIN, affects females three times more often than males, and has a median age of onset of 15 years. Its hallmark feature, in addition to a lymphocyte-predominant interstitial nephritis **(Fig. 310-2)**, is a painful anterior uveitis, often bilateral and accompanied by blurred vision and photophobia. Diagnosis is often confounded by the fact that the ocular symptoms precede or accompany the renal disease in only one-third of cases. Additional extrarenal features include fever, anorexia, weight loss, abdominal pain, and arthralgia. The presence of such symptoms as well as elevated creatinine, sterile pyuria, mild proteinuria, features of Fanconi's syndrome, and elevated erythrocyte sedimentation rate should raise suspicion for this disorder. Serologies suggestive of the more common autoimmune diseases are usually negative, and TINU is often a diagnosis of exclusion after other causes of uveitis and renal disease, such as Sjögren's syndrome, Behçet's disease, sarcoidosis, and systemic lupus erythematosus, have been considered. Clinical symptoms are typically self-limited in children, but are more apt to follow a relapsing course in adults. The renal and ocular manifestations generally respond well to oral glucocorticoids, although maintenance therapy with agents such as methotrexate, azathioprine, or mycophenolate may be necessary to prevent relapses (Fig. 310-1 and Table 310-2).

FIGURE 310-2

Acute interstitial nephritis (AIN) in a patient who presented with acute iritis, low-grade fever, erythrocyte sedimentation rate of 103, pyuria and cellular casts on urinalysis, and a newly elevated serum creatinine of 2.4 mg/dL. Both the iritis and AIN improved after intravenous methylprednisolone. This PAS-stained renal biopsy shows a mononuclear cell interstitial infiltrate (*asterisks*) and edema separating the tubules (T) and a normal glomerulus (G). Some of the tubules contain cellular debris and infiltrating inflammatory cells. The findings in this biopsy are indistinguishable from those that would be seen in a case of drug-induced AIN. PAS, Periodic acid–Schiff.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SYSTEMIC LUPUS ERYTHEMATOSUS

An interstitial mononuclear cell inflammatory reaction often accompanies the glomerular lesion in most cases of class III or IV lupus nephritis **(Chap. 308)**, and deposits of immune complexes can be identified in tubule basement membranes in ~50% of cases. Occasionally, however, the tubulointerstitial inflammation predominates and may manifest with azotemia and type IV RTA rather than features of glomerulonephritis.

GRANULOMATOUS INTERSTITIAL NEPHRITIS

Some patients may present with features of AIN but follow a protracted and relapsing course. Renal biopsy in such patients reveals a more chronic inflammatory infiltrate with granulomas and multinucleated giant cells. Most often, no associated disease or cause is found; however, some of these cases may have or subsequently develop the pulmonary, cutaneous, or other systemic manifestations of *sarcoidosis* such as hypercalcemia. Most patients experience some improvement in renal function if treated early with glucocorticoids before the development of significant interstitial fibrosis and tubular atrophy (Table 310-2). Other immunosuppressive agents may be required for those who relapse frequently upon steroid withdrawal. Other immunosuppressive agents may be required for those who relapse frequently upon steroid withdrawal (Fig. 310-1). Tuberculosis should be ruled out before starting treatment because this too is a rare cause of granulomatous interstitial nephritis.

IgG4-RELATED SYSTEMIC DISEASE

A form of AIN characterized by a dense inflammatory infiltrate containing IgG4-expressing plasma cells can occur as a part of a syndrome known as IgG4-related systemic disease. Autoimmune pancreatitis, sclerosing cholangitis, retroperitoneal fibrosis, and a chronic sclerosing sialadenitis (mimicking Sjögren's syndrome) may variably be present as well. Fibrotic lesions that form pseudotumors in the affected organs soon replace the initial inflammatory infiltrates and often lead to biopsy or excision for fear of true malignancy. Although the involvement of IgG4 in the pathogenesis is not understood, glucocorticoids have been successfully used as first-line treatment in this group of disorders, once they are correctly diagnosed.

IDIOPATHIC AIN

Some patients present with typical clinical and histologic features of AIN but have no evidence of drug exposure or clinical or serologic features of an autoimmune disease. The presence in some cases of autoantibodies to a tubular antigen, similar to that identified in rats with an induced form of interstitial nephritis, suggests that an autoimmune response may be involved. Like TINU and granulomatous interstitial nephritis, idiopathic AIN is responsive to glucocorticoid therapy but may follow a relapsing course requiring maintenance treatment with another immunosuppressive agent (Fig. 310-1 and Table 310-2). Recently, cases have been identified in which autoantibodies that may be important in disease pathogenesis were seen to target antigens expressed by collecting duct or proximal tubular brush border.

INFECTION-ASSOCIATED AIN

AIN may also occur as a local inflammatory reaction to microbial infection (Table 310-1) and should be distinguished from acute bacterial pyelonephritis **(Chap. 130)**. Acute bacterial pyelonephritis does not generally cause acute renal failure unless it affects both kidneys or causes septic shock. Presently, infection-associated AIN is most often seen in immunocompromised patients, particularly renal transplant recipients with reactivation of polyomavirus BK **(Chaps. 138 and 307)**.

CRYSTAL DEPOSITION DISORDERS AND OBSTRUCTIVE TUBULOPATHIES

Acute renal failure may occur when crystals of various types are deposited in tubular cells and interstitium or when they obstruct tubules. Oliguric acute renal failure, often accompanied by flank pain from tubular obstruction, may occur in patients treated with sulfadiazine for toxoplasmosis, indinavir and atazanavir for HIV, and intravenous acyclovir for severe herpesvirus infections. Urinalysis reveals "sheaf of wheat" sulfonamide crystals, individual or parallel clusters of needle-shaped indinavir crystals, or red-green birefringement needle-shaped crystals of acyclovir. This adverse effect is generally precipitated by hypovolemia and is reversible with saline volume repletion and drug withdrawal. Distinct from the obstructive disease, a frank AIN from indinavir crystal deposition has also been reported.

Acute tubular obstruction is also the cause of oliguric renal failure in patients with *acute urate nephropathy*. It typically results from severe hyperuricemia from tumor lysis syndrome in patients with lympho- or myeloproliferative disorders treated with cytotoxic agents, but also may occur spontaneously before the treatment has been initiated **(Chap. 71)**. Uric acid crystallization in the tubules and collecting system leads to

partial or complete obstruction of the collecting ducts, renal pelvis, or ureter. A dense precipitate of birefringent uric acid crystals is found in the urine, usually in association with microscopic or gross hematuria. Prophylactic allopurinol reduces the risk of uric acid nephropathy but is of no benefit once tumor lysis has occurred. Once oliguria has developed, attempts to increase tubular flow and solubility of uric acid with alkaline diuresis may be of some benefit; however, emergent treatment with hemodialysis or rasburicase, a recombinant urate oxidase, is usually required to rapidly lower uric acid levels and restore renal function.

Calcium oxalate crystal deposition in tubular cells and interstitium may lead to permanent renal dysfunction in patients who survive ethylene glycol intoxication, in patients with enteric hyperoxaluria from ileal resection or small-bowel bypass surgery, and in patients with hereditary hyperoxaluria **(Chap. 312)**. *Acute phosphate nephropathy* is an uncommon but serious complication of oral Phosphosoda used as a laxative or for bowel preparation for colonoscopy. It results from calcium phosphate crystal deposition in tubules and interstitium and occurs especially in subjects with underlying renal impairment and hypovolemia. Consequently, Phosphosoda should be avoided in patients with chronic kidney disease.

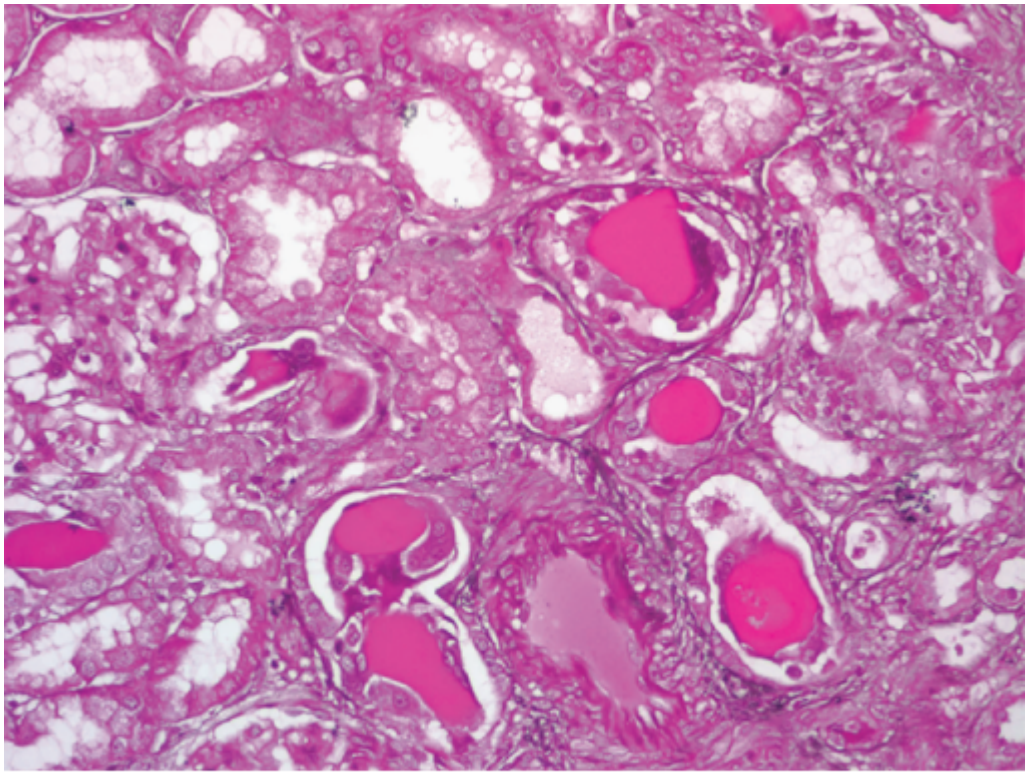
LIGHT CHAIN CAST NEPHROPATHY

Patients with multiple myeloma may develop acute renal failure in the setting of hypovolemia, infection, or hypercalcemia or after exposure to NSAIDs or radiographic contrast media. The diagnosis of light chain cast nephropathy (LCCN)—commonly known as *myeloma kidney*—should be considered in patients who fail to recover when the precipitating factor is corrected or in any elderly patient with otherwise unexplained acute renal failure.

In this disorder, filtered monoclonal immunoglobulin light chains (Bence-Jones proteins) form intratubular aggregates with secreted Tamm-Horsfall protein in the distal tubule. Casts, in addition to obstructing the tubular flow in affected nephrons, incite a giant cell or foreign body reaction and can lead to tubular rupture, resulting in interstitial fibrosis **(Fig. 310-3)**. Although LCCN generally occurs in patients with known multiple myeloma and a large plasma cell burden, the disorder should also be considered as a possible diagnosis in patients who have known monoclonal gammopathy even in the absence of frank myeloma. Filtered monoclonal light chains may also cause less pronounced renal manifestations in the absence of obstruction, due to direct toxicity to proximal tubular cells and intracellular crystal formation. This may result in isolated tubular disorders such as RTA or full Fanconi's syndrome.

FIGURE 310-3

Histologic appearance of myeloma cast nephropathy. A hematoxylin-eosin–stained kidney biopsy shows many atrophic tubules filled with eosinophilic casts (consisting of Bence-Jones protein), which are surrounded by giant cell reactions. *(Courtesy of Dr. Michael N. Koss, University of Southern California Keck School of Medicine; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Diagnosis

Clinical clues to the diagnosis include anemia, bone pain, hypercalcemia, and an abnormally narrow anion gap due to hypoalbuminemia and hypergammaglobulinemia. Urinary dipsticks detect albumin but not immunoglobulin light chains; however, laboratory detection of increased amounts of protein in a spot urine specimen and a negative dipstick result are highly suggestive that the urine contains Bence-Jones protein. Serum and urine should both be sent for protein electrophoresis and for immunofixation for the detection and identification of a potential monoclonal band. A sensitive method is available to detect urine and serum free light chains.

TREATMENT

TREATMENT

Light Chain Cast Nephropathy

The goals of treatment are to correct precipitating factors such as hypovolemia and hypercalcemia, discontinue potential nephrotoxic agents, and treat the underlying plasma cell dyscrasia (Chap. 107); plasmapheresis to remove light chains is of questionable value for LCCN.

LYMPHOMATOUS INFILTRATION OF THE KIDNEY

Interstitial infiltration by malignant B lymphocytes is a common autopsy finding in patients dying of chronic lymphocytic leukemia and non-Hodgkin's lymphoma; however, this is usually an incidental finding. Rarely, such infiltrates may cause massive enlargement of the kidneys and oliguric acute renal failure. Although high-dose glucocorticoids and subsequent chemotherapy often result in recovery of renal function, the prognosis in such cases is generally poor.

CHRONIC TUBULOINTERSTITIAL DISEASES

Improved occupational and public health measures, together with the banning of over-the-counter phenacetin-containing analgesics, has led to a dramatic decline in the incidence of chronic interstitial nephritis (CIN) from heavy metal—particularly lead and cadmium—exposure and analgesic nephropathy in North America. Today, CIN is most often the result of renal ischemia or secondary to a primary glomerular disease **(Chap. 308)**. Other important forms of CIN are the result of developmental anomalies or inherited diseases such as reflux nephropathy or sickle cell nephropathy and may not be recognized until adolescence or adulthood. Although it is impossible to reverse damage that has already occurred, further deterioration may be prevented or at least slowed in such cases by treating glomerular hypertension, a common denominator in the development of secondary FSGS and progressive loss of functioning nephrons. Therefore, awareness and early detection of patients at risk may prevent them from developing end-stage renal disease (ESRD).

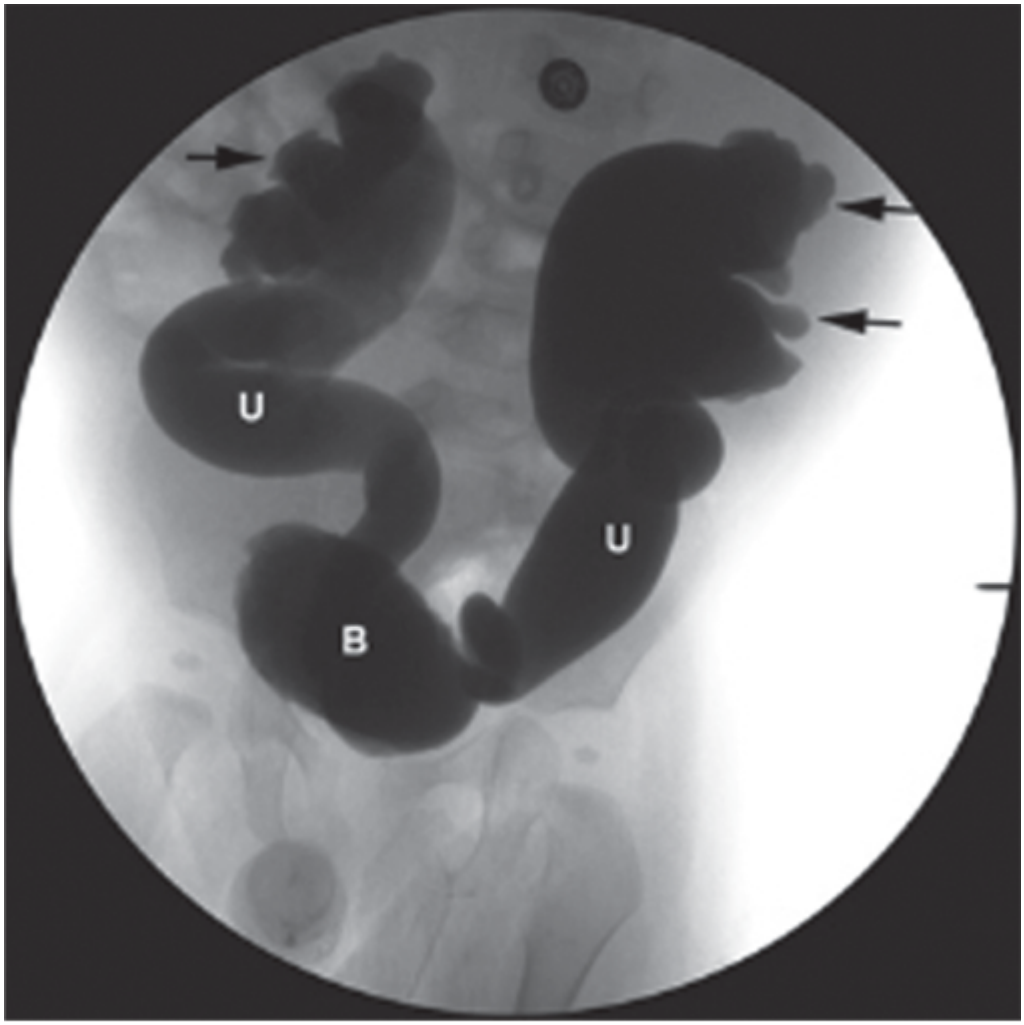
VESICOURETERAL REFLUX AND REFLUX NEPHROPATHY

Reflux nephropathy is the consequence of vesicoureteral reflux (VUR) or other urologic anomalies in early childhood. It was previously called *chronic pyelonephritis* because it was believed to result from recurrent urinary tract infections (UTIs) in childhood. VUR stems from abnormal retrograde urine flow from the bladder into one or both ureters and kidneys because of mislocated and incompetent ureterovesical valves **(Fig. 310-4)**. Although high-pressure sterile reflux may impair normal growth of the kidneys, when coupled with recurrent UTIs in early childhood, the result is patchy interstitial scarring and tubular atrophy. Loss of functioning nephrons leads to hypertrophy of the remnant glomeruli and eventual secondary FSGS. Reflux nephropathy often goes unnoticed until early adulthood when chronic kidney disease is detected during routine evaluation or during pregnancy. Affected adults are frequently asymptomatic, but may give a history of prolonged bed-wetting or recurrent UTIs during childhood, and exhibit variable renal insufficiency, hypertension, mild to moderate proteinuria, and unremarkable urine sediment. When both kidneys are affected, the disease often progresses inexorably over several years to ESRD, despite the absence of ongoing urinary infections or reflux. A single affected kidney may go undetected, except for the presence of hypertension. Renal ultrasound in adults characteristically shows asymmetric small kidneys with irregular outlines, thinned cortices, and regions of compensatory hypertrophy (Fig. 310-4).

FIGURE 310-4

Radiographs of vesicoureteral reflux (VUR) and reflux nephropathy. *A.* Voiding cystourethrogram in a 7-month-old baby with bilateral high-grade VUR evidenced by clubbed calyces (*arrows*) and dilated tortuous ureters (U) entering the bladder (B). ***B.*** Abdominal computed tomography scan (coronal plane

reconstruction) in a child showing severe scarring of the lower portion of the right kidney (*arrow*). ***C.*** Sonogram of the right kidney showing loss of parenchyma at the lower pole due to scarring (*arrow*) and hypertrophy of the mid-region (*arrowhead*). *(Courtesy of Dr. George Gross, University of Maryland Medical Center; with permission.)*



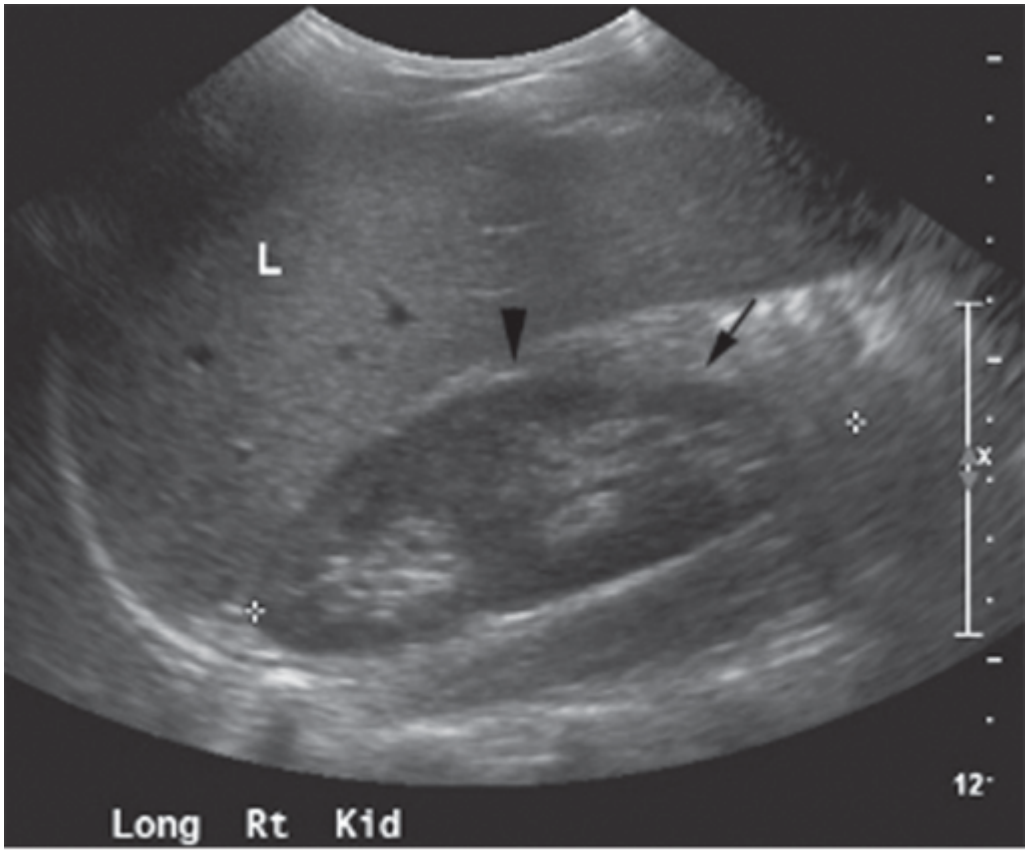
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Vesicoureteral Reflux and Reflux Nephropathy

Maintenance of sterile urine in childhood has been shown to limit scarring of the kidneys. Surgical reimplantation of the ureters into the bladder to restore competency is indicated in young children with persistent high-grade reflux, but is ineffective and is not indicated in adolescents or adults after scarring has occurred. Aggressive control of blood pressure with an angiotensin-converting enzyme inhibitor (ACEI) or angiotensin receptor blocker (ARB) and other agents is effective in reducing proteinuria and may significantly forestall further deterioration of renal function.

SICKLE CELL NEPHROPATHY

The pathogenesis and clinical manifestations of sickle cell nephropathy are described in **Chap. 311**. Evidence of tubular injury may be evident in childhood and early adolescence in the form of polyuria due to decreased concentrating ability or type IV RTA years before there is significant nephron loss and proteinuria from secondary FSGS. Early recognition of these subtle renal abnormalities or development of microalbuminuria in a child with sickle cell disease may warrant consultation with a nephrologist and/or therapy with low-dose

ACEIs. Papillary necrosis may result from ischemia due to sickling of red cells in the relatively hypoxemic and hypertonic medullary vasculature and present with gross hematuria and ureteric obstruction by sloughed ischemic papillae **(Table 310-3)**.

TABLE 310-3

Major Causes of Papilary Necrosis

Analgesic nephropathy
Sickle cell nephropathy
Diabetes with urinary tract infection
Prolonged NSAID use (rare)

Abbreviation: NSAID, nonsteroidal anti-inflammatory drug.

TUBULOINTERSTITIAL ABNORMALITIES ASSOCIATED WITH GLOMERULONEPHRITIS

Primary glomerulopathies are often associated with damage to tubules and interstitium. This may occasionally be due to the same pathologic process affecting the glomerulus and tubulointerstitium, as is the case with immune-complex deposition in lupus nephritis. More often, however, chronic tubulointerstitial changes occur as a secondary consequence of prolonged glomerular dysfunction. Potential mechanisms by which glomerular disease might cause tubulointerstitial injury include proteinuria-mediated damage to the epithelial cells, activation of tubular cells by cytokines and complement, or reduced peritubular blood flow leading to downstream tubulointerstitial ischemia, especially in the case of glomeruli that are globally obsolescent due to severe glomerulonephritis. It is often difficult to discern the initial cause of injury by renal biopsy in a patient who presents with advanced renal disease in this setting.

ANALGESIC NEPHROPATHY

Analgesic nephropathy results from the long-term use of compound analgesic preparations containing phenacetin (banned in the United States since 1983), aspirin, and caffeine. In its classic form, analgesic nephropathy is characterized by renal insufficiency, papillary necrosis (Table 310-3) attributable to the presumed concentration of the drug to toxic levels in the inner medulla, and a radiographic constellation of small, scarred kidneys with papillary calcifications best appreciated by computed tomography **(Fig. 310-5)**. Patients may also have polyuria due to impaired concentrating ability and non-anion-gap metabolic acidosis from tubular damage. Shedding of a sloughed necrotic papilla can cause gross hematuria and ureteric colic due to ureteral obstruction. Individuals with ESRD as a result of analgesic nephropathy are at increased risk of a urothelial malignancy compared to patients with other causes of renal failure. Recent cohort studies in

individuals with normal baseline renal function suggest that the moderate chronic use of current analgesic preparations available in the United States, including acetaminophen and NSAIDs, does not seem to cause the constellation of findings known as analgesic nephropathy, although volume-depleted individuals and those with chronic kidney disease are at higher risk of NSAID-related renal toxicity. Nonetheless, it is recommended that heavy users of acetaminophen and NSAIDs be screened for evidence of renal disease.

FIGURE 310-5

Radiologic appearance of analgesic nephropathy. A noncontrast computed tomography scan shows an atrophic left kidney with papillary calcifications in a garland pattern. *(Reprinted by permission from Macmillan Publishers, Ltd., MM Elseviers et al: Kidney International 48:1316, 1995.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ARISTOLOCHIC ACID NEPHROPATHY

Two seemingly unrelated forms of CIN, Chinese herbal nephropathy and Balkan endemic nephropathy, have recently been linked by the underlying etiologic agent aristolochic acid and are now collectively termed aristolochic acid nephropathy (AAN). In Chinese herbal nephropathy, first described in the early 1990s in young women taking traditional Chinese herbal preparations as part of a weight-loss regimen, one of the offending agents has been identified as aristolochic acid, a known carcinogen from the plant *Aristolochia*. Multiple *Aristolochia* species have been used in traditional herbal remedies for centuries and continue to be available despite official bans on their use in many countries. Molecular evidence has also implicated aristolochic acid in Balkan endemic nephropathy, a chronic TIN found primarily in towns along the tributaries of the Danube River and first described in the 1950s. Although the exact route of exposure is not known with certainty, contamination of local grain preparations with the seeds of *Aristolochia* species seems most likely. Aristolochic acid, after prolonged exposure, produces renal interstitial fibrosis with a relative

paucity of cellular infiltrates. The urine sediment is bland, with rare leukocytes and only mild proteinuria. Anemia may be disproportionately severe relative to the level of renal dysfunction. Definitive diagnosis of AAN requires two of the following three features: characteristic histology on kidney biopsy; confirmation of aristolochic acid ingestion; and detection of aristolactam-DNA adducts in kidney or urinary tract tissue. These latter lesions represent a molecular signature of aristolochic acid–derived DNA damage and often consist of characteristic A:T-to-T:A transversions. Due to this mutagenic activity, AAN is associated with a very high incidence of upper urinary tract urothelial cancers, with risk related to cumulative dose. Surveillance with computed tomography, ureteroscopy, and urine cytology is warranted, and consideration should be given to bilateral nephroureterectomy once a patient has reached ESRD.

KARYOMEGALIC INTERSTITIAL NEPHRITIS

Karyomegalic interstitial nephritis is an unusual form of slowly progressive chronic kidney disease with mild proteinuria, interstitial fibrosis, tubular atrophy, and oddly enlarged nuclei of proximal tubular epithelial cells. It has been linked to mutations in *FAN1*, a nuclease involved in DNA repair, which may render carriers of the mutation susceptible to environmental DNA-damaging agents.

LITHIUM-ASSOCIATED NEPHROPATHY

The use of lithium salts for the treatment of manic-depressive illness may have several renal sequelae, the most common of which is nephrogenic diabetes insipidus manifesting as polyuria and polydipsia. Lithium accumulates in principal cells of the collecting duct by entering through the epithelial sodium channel (ENaC), where it inhibits glycogen synthase kinase 3β and downregulates vasopressin-regulated aquaporin water channels. Less frequently, chronic TIN develops after prolonged (>10–20 years) lithium use and is most likely to occur in patients who have experienced repeated episodes of toxic lithium levels. Findings on renal biopsy include interstitial fibrosis and tubular atrophy that are out of proportion to the degree of glomerulosclerosis or vascular disease, a sparse lymphocytic infiltrate, and small cysts or dilation of the distal tubule and collecting duct that are highly characteristic of this disorder. The degree of interstitial fibrosis correlates with both duration and cumulative dose of lithium. Individuals with lithium-associated nephropathy are typically asymptomatic, with minimal proteinuria, few urinary leukocytes, and normal blood pressure. Some patients develop more severe proteinuria due to secondary FSGS, which may contribute to further loss of renal function.

TREATMENT

TREATMENT

Lithium-Associated Nephropathy

Renal function should be followed regularly in patients taking lithium, and caution should be exercised in patients with underlying renal disease. The use of amiloride to inhibit lithium entry via ENaC has been effective to prevent and treat lithium-induced nephrogenic diabetes insipidus, but it is not clear if it will prevent lithium-induced CIN. Once lithium-associated nephropathy is detected, the discontinuation of

lithium in attempt to forestall further renal deterioration can be problematic, as lithium is an effective mood stabilizer that is often incompletely substituted by other agents. Furthermore, despite discontinuation of lithium, chronic renal disease in such patients is often irreversible and can slowly progress to ESRD. The most prudent approach is to monitor lithium levels frequently and adjust dosing to avoid toxic levels (preferably <1 meq/L). This is especially important because lithium is cleared less effectively as renal function declines. In patients who develop significant proteinuria, ACEI or ARB treatment should be initiated.

CALCINEURIN-INHIBITOR NEPHROTOXICITY

The calcineurin inhibitor (CNI) immunosuppressive agents cyclosporine and tacrolimus can cause both acute and chronic renal injury. Acute forms can result from vascular causes such as vasoconstriction or the development of thrombotic microangiopathy, or can be due to a toxic tubulopathy. Chronic CNI-induced renal injury is typically seen in solid organ (including heart-lung and liver) transplant recipients and manifests with a slow but irreversible reduction of glomerular filtration rate, with mild proteinuria and arterial hypertension. Hyperkalemia is a relatively common complication and is caused, in part, by tubular resistance to aldosterone. The histologic changes in renal tissue include patchy interstitial fibrosis and tubular atrophy, often in a “striped” pattern. In addition, the intrarenal vasculature often demonstrates hyalinosis, and focal glomerulosclerosis can be present as well. Similar changes may occur in patients receiving CNIs for autoimmune diseases, although the doses are generally lower than those used for organ transplantation. Dose reduction or CNI avoidance appears to mitigate the chronic tubulointerstitial changes, but may increase the risk of rejection and graft loss.

HEAVY METAL (LEAD) NEPHROPATHY

Heavy metals, such as lead or cadmium, can lead to a chronic tubulointerstitial process after prolonged exposure. The disease entity is no longer commonly diagnosed, because such heavy metal exposure has been greatly reduced due to the known health risks from lead and the consequent removal of lead from most commercial products and fuels. Nonetheless, occupational exposure is possible in workers involved in the manufacture or destruction of batteries, removal of lead paint, or manufacture of alloys and electrical equipment (cadmium) in countries where industrial regulation is less stringent. In addition, ingestion of moonshine whiskey distilled in lead-tainted containers has been one of the more frequent sources of lead exposure.

Early signs of chronic lead intoxication are attributable to proximal tubule dysfunction, particularly hyperuricemia as a result of diminished urate secretion. The triad of “saturnine gout,” hypertension, and renal insufficiency should prompt a practitioner to ask specifically about lead exposure. Unfortunately, evaluating lead burden is not as straightforward as ordering a blood test; the preferred methods involve measuring urinary lead after infusion of a chelating agent or by radiographic fluoroscopy of bone. Several recent studies have shown an association between chronic low-level lead exposure and decreased renal function, although either of these two factors may have been the primary event. In those patients who have

CIN of unclear origin and an elevated total body lead burden, repeated treatments of lead chelation therapy have been shown to slow the decline in renal function.

METABOLIC DISORDERS

Disorders leading to excessively high or low levels of certain electrolytes and products of metabolism can also lead to chronic kidney disease if untreated.

CHRONIC URIC ACID NEPHROPATHY

The constellation of pathologic findings that represent *gouty nephropathy* are very uncommon nowadays and are more of historical interest than clinical importance, as gout is typically well managed with allopurinol and other agents. However, there is emerging evidence that hyperuricemia is an independent risk factor for the development of chronic kidney disease, perhaps through endothelial damage. The complex interactions of hyperuricemia, hypertension, and renal failure are still incompletely understood.

Presently, gouty nephropathy is most likely to be encountered in patients with severe tophaceous gout and prolonged hyperuricemia from a hereditary disorder of purine metabolism **(Chap. 410)**. This should be distinguished from juvenile hyperuricemic nephropathy, a form of medullary cystic kidney disease caused by mutations in uromodulin (UMOD) **(Chap. 309)** and now grouped into the larger category of autosomal dominant tubulointerstitial kidney disease. Histologically, the distinctive feature of gouty nephropathy is the presence of crystalline deposits of uric acid and monosodium urate salts in the kidney parenchyma. These deposits not only cause intrarenal obstruction but also incite an inflammatory response, leading to lymphocytic infiltration, foreign-body giant cell reaction, and eventual fibrosis, especially in the medullary and papillary regions of the kidney. Since patients with gout frequently suffer from hypertension and hyperlipidemia, degenerative changes of the renal arterioles may constitute a striking feature of the histologic abnormality, out of proportion to the other morphologic defects. Clinically, gouty nephropathy is an insidious cause of chronic kidney disease. Early in its course, glomerular filtration rate may be near normal, often despite morphologic changes in medullary and cortical interstitium, proteinuria, and diminished urinary concentrating ability. Treatment with allopurinol and urine alkalinization is generally effective in preventing uric acid nephrolithiasis and the consequences of recurrent kidney stones; however, gouty nephropathy may be intractable to such measures. Furthermore, the use of allopurinol in asymptomatic hyperuricemia has not been consistently shown to improve renal function.

HYPERCALCEMIC NEPHROPATHY

(See also Chap. 403) Chronic hypercalcemia, as occurs in primary hyperparathyroidism, sarcoidosis, multiple myeloma, vitamin D intoxication, or metastatic bone disease, can cause tubulointerstitial disease and progressive renal failure. The earliest lesion is a focal degenerative change in renal epithelia, primarily in collecting ducts, distal tubules, and loops of Henle. Tubular cell necrosis leads to nephron obstruction and stasis of intrarenal urine, favoring local precipitation of calcium salts and infection. Dilation and atrophy of tubules eventually occur, as do interstitial fibrosis, mononuclear leukocyte infiltration, and interstitial

calcium deposition (nephrocalcinosis). Calcium deposition may also occur in glomeruli and the walls of renal arterioles.

Clinically, the most striking defect is an inability to maximally concentrate the urine, due to reduced collecting duct responsiveness to arginine vasopressin and defective transport of sodium and chloride in the loop of Henle. Reductions in both glomerular filtration rate and renal blood flow can occur, both in acute and in prolonged hypercalcemia. Eventually, uncontrolled hypercalcemia leads to severe tubulointerstitial damage and overt renal failure. Abdominal x-rays may demonstrate nephrocalcinosis as well as nephrolithiasis, the latter due to the hypercalciuria that often accompanies hypercalcemia.

Treatment consists of reducing the serum calcium concentration toward normal and correcting the primary abnormality of calcium metabolism **(Chap. 403).** Renal dysfunction of acute hypercalcemia may be completely reversible. Gradual progressive renal insufficiency related to chronic hypercalcemia, however, may not improve even with correction of the calcium disorder.

HYPOKALEMIC NEPHROPATHY

Patients with prolonged and severe hypokalemia from chronic laxative or diuretic abuse, surreptitious vomiting, or primary aldosteronism may develop a reversible tubular lesion characterized by vacuolar degeneration of proximal and distal tubular cells. Eventually, tubular atrophy and cystic dilation accompanied by interstitial fibrosis may ensue, leading to irreversible chronic kidney disease. Timely correction of the hypokalemia will prevent further progression, but persistent hypokalemia can cause ESRD.

GLOBAL PERSPECTIVE



The causes of acute and CIN vary widely across the globe. Analgesic nephropathy continues to be seen in countries where phenacetin-containing compound analgesic preparations are readily available. Adulterants in unregulated herbal and traditional medicaments pose a threat of toxic interstitial nephritis, as exemplified by aristolochic acid contamination of herbal slimming preparations. Contamination of food sources with toxins, such as the recent outbreak of nephrolithiasis and acute renal failure from melamine contamination of infant milk formula, poses a continuing risk. Large-scale exposure to aristolochic acid remains prevalent in many Asian countries where traditional herbal medicine use is common. Although industrial exposure to lead and cadmium has largely disappeared as a cause of CIN in developed nations, it remains a risk for nephrotoxicity in countries where such exposure is less well controlled. New endemic forms of chronic kidney disease continue to be described. Most notable is the nephropathy found among Pacific coastal plantation workers in Central America that is estimated to have claimed 20,000 lives thus far due to the development of end-stage kidney disease. This entity has been named Mesoamerican nephropathy, although similar pathophysiologic mechanisms may also be at play in other regional forms of chronic kidney disease in Sri Lanka and southern India. Although a variety of etiologic factors have been proposed, the most likely cause appears to be related to repetitive episodes of heat exposure, dehydration or volume depletion, and consequent metabolic changes leading to uricosuria and elevated levels of vasopressin. Global warming

and regional temperature variability have been proposed as contributors to these newly described forms of kidney disease.

FURTHER READING

Bleyer AJ, Kmoch S: Autosomal dominant tubulointerstitial kidney disease: Of names and genes. Kidney Int 86:459, 2014. [PubMed: 25168494]

Correa-Rotter R et al: CKD of unknown origin in Central America: The case for a Mesoamerican nephropathy. Am J Kidney Dis 63:506, 2014. [PubMed: 24412050]

Muriithi AK et al: Clinical characteristics, causes and outcomes of acute interstitial nephritis in the elderly. Kidney Int 87:458, 2015. [PubMed: 25185078]

Praga M et al: Changes in the aetiology, clinical presentation and management of acute interstitial nephritis, an increasingly common cause of acute kidney injury. Nephrol Dial Transplant 30:1472, 2015. [PubMed: 25324356]

Raissian Y et al: Diagnosis of IgG4-related tubulointerstitial nephritis. J Am Soc Nephrol 22:1343, 2011. [PubMed: 21719792]

Chapter 311: Vascular Injury to the Kidney

Ronald S. Go; Nelson Leung

INTRODUCTION

The renal circulation is complex and is characterized by a highly perfused arteriolar network, reaching cortical glomerular structures adjacent to lower-flow vasa recta that descend into medullary segments. Disorders of the larger vessels, including renal artery stenosis and atheroembolic disease, are discussed elsewhere (Chap. 322). This chapter examines primary disorders of the renal microvessels, many of which are associated with thrombosis and hemolysis.

THROMBOTIC MICROANGIOPATHY

Thrombotic microangiopathy (TMA) is a pathologic lesion characterized by endothelial cell injury in the terminal arterioles and capillaries. Platelet and hyaline thrombi causing partial or complete occlusion are integral to the histopathology of TMA. TMA is usually accompanied by microangiopathic hemolytic anemia (MAHA) with its typical features of thrombocytopenia and schistocytes, but not always. In the kidney, TMA is characterized by swollen endocapillary cells (endotheliosis), fibrin thrombi, platelet plugs, arterial intimal fibrosis, and a membranoproliferative pattern in the glomerulus. Fibrin thrombi may extend into the arteriolar vascular pole, producing glomerular collapse and at times cortical necrosis. In kidneys that recover from acute TMA, secondary focal segmental glomerulosclerosis may develop. Diseases associated with this lesion include thrombotic thrombocytopenic purpura (TTP), hemolytic-uremic syndrome (HUS), malignant hypertension, scleroderma renal crisis, antiphospholipid syndrome, preeclampsia/HELLP (hemolysis, elevated liver enzymes, low platelet count) syndrome, HIV infection, and radiation nephropathy. TMA can also be seen in myeloproliferative neoplasm (MPN)-related glomerulopathy and POEMS (polyneuropathy, endocrinopathy, organomegaly, monoclonal gammopathy and skin changes) syndrome which are not associated with MAHA.

HEMOLYTIC-UREMIC SYNDROME/THROMBOTIC THROMBOCYTOPENIC PURPURA

HUS and TTP are the prototypes for MAHA. Historically, HUS and TTP were distinguished mainly by their clinical and epidemiologic differences. TTP develops more commonly in adults and was thought to have more neurologic complications while HUS occurs more frequently in children, particularly when associated with hemorrhagic diarrhea. However, atypical HUS (aHUS) can have its first appearance in adulthood, and

better testing has revealed that neurologic involvement is as common in HUS as in TTP. Currently, HUS and TTP can be differentiated etiologically and treated according to their specific pathophysiologic features.

Hemolytic-Uremic Syndrome

HUS is loosely defined by the presence of MAHA and renal impairment. At least four variants are recognized. The most common is Shiga toxin–producing *Escherichia coli* (STEC) HUS, which is also known as D^+ (diarrhea-associated) HUS or enterohemorrhagic *E. coli* (EHEC) HUS. Most cases involve children <5 years of age, but adults also are susceptible, as evidenced by a 2011 outbreak in northern Europe. Diarrhea, often bloody, precedes MAHA within 1 week in >80% of cases. Abdominal pain, cramping, and vomiting are frequent, whereas fever is typically absent. Neurologic symptoms, including dysphasia, hyperreflexia, blurred vision, memory deficits, encephalopathy, perseveration, and agraphia, often develop, especially in adults. Seizures and cerebral infarction can occur in severe cases. STEC HUS is caused by the Shiga toxins (Stx1 and Stx2), which are also referred to as *verotoxins.* These toxins are produced by certain strains of *E. coli* and *Shigella dysenteriae*. In the United States and Europe, the most common STEC strain is O157:H7, but HUS has been reported with other strains (O157/H^-, O111:H^-, O26:H11/H^-, O145:H28, and O104:H4). After entry into the circulation, Shiga toxin binds to the glycolipid surface receptor globotriaosylceramide (Gb3), which is richly expressed on cells of the renal microvasculature. Upon binding, the toxin enters the cells, inducing inflammatory cytokines (interleukin 8 [IL-8], monocyte chemotactic protein 1 [MCP-1], and stromal cell–derived factor 1 [SDF-1]) and chemokine receptors (CXCR4 and CXCR7); this action results in platelet aggregation and the microangiopathic process. *Streptococcus pneumoniae* can also cause HUS. Certain strains produce a neuraminidase that cleaves the N-acetylneuraminic acid moieties normally covering the Thomsen-Friedenreich antigen on platelets and endothelial cells. Exposure of this cryptic antigen to preformed IgM results in severe MAHA.

Atypical HUS or complement medicated HUS is the result of complement dysregulation. The complement dysregulation can be congenital or acquired. The affected patients often have a low C3 and a normal C4 levels characteristic of alternative pathway activation. Factor H deficiency, the most common defect, has been linked to families with aHUS. Factor H competes with factor B to prevent the formation of C3bBb and acts as a cofactor for factor I, which proteolytically degrades C3b. More than 70 mutations of the factor H gene have been identified. Most are missense mutations that produce abnormalities in the C-terminus region, affecting its binding to C3b but not its concentration. Other mutations result in low levels or the complete absence of the protein. Deficiencies in other complement-regulatory proteins, such as factor I, factor B, membrane cofactor protein (CD46), C3, complement factor H–related protein 1 (CFHR1), CFHR3, CFHR5, and thrombomodulin, have also been reported. Finally, an autoimmune variant of aHUS has been discovered. DEAP (deficiency of CFHR plasma proteins and CFH autoantibody positive) HUS occurs when an autoantibody to factor H is formed. DEAP HUS is often associated with a deletion of an 84-kb fragment of the chromosome that encodes for CFHR1 and CFHR3. The autoantibody blocks the binding of factor H to C3b and surface-bound C3 convertase. Renal injury is often severe resulting in end stage renal disease. The severity of the renal injury and recurrence after kidney transplant depend on the complement regulatory protein.

Traditionally, TTP is characterized by the pentad: MAHA, thrombocytopenia, neurologic symptoms, fever, and renal failure. The pathophysiology of TTP involves the accumulation of ultra-large multimers of von Willebrand factor as a result of the absence or markedly decreased activity of the plasma protease ADAMTS13, a disintegrin and metalloproteinase with a thrombospondin type 1 motif, member 13. TTP is now defined as MAHA associated with ADAMTS13 activity of (<5–10%). These ultra-large multimers form clots and shear erythrocytes, resulting in MAHA; however, the absence of ADAMTS13 alone may not by itself produce TTP. Often, an additional inflammatory trigger (such as infection, surgery, pancreatitis, or pregnancy) is required to initiate clinical TTP. This may be mediated by human neutrophil peptides that inhibit cleavage of von Willebrand factor by ADAMTS13.

Data from the Oklahoma TTP/HUS Registry suggest an incidence rate of 2.9 cases/10^6 patients in the United States. The median age of onset is 40 years. The incidence is more than nine times higher among blacks than among non-blacks. Like that of systemic lupus erythematosus, the incidence of TTP is nearly three times higher among women than among men. If untreated, TTP has a mortality rate exceeding 90%. Even with modern therapy, 20% of patients die within the first month from complications of microvascular thrombosis.

The classic form of TTP is idiopathic TTP, which is the result of a severe deficiency in ADAMTS13. In the past, TTP had traditionally been associated with infection, malignancy, and intense inflammation (e.g., pancreatitis), but ADAMTS13 activity is typically not decreased in these conditions and therefore should be not considered TTP. In idiopathic TTP, the formation of an autoantibody to ADAMTS13 (IgG or IgM) either increases its clearance or inhibits its activity. Upshaw-Schülman syndrome is a hereditary condition characterized by congenital deficiency of ADAMTS13. TTP in these patients can start within the first weeks of life but in some instances may not present until adulthood, especially during pregnancy. Both environmental and genetic factors are thought to influence the development of TTP. Plasma transfusion is an effective strategy for prevention and treatment.

Drug-induced TMA is a recognized complication of treatment with some chemotherapeutic agents, immunosuppressive agents, and quinine. Two different mechanisms are now recognized. Toxic or endothelial damage (pathologically similar to that of HUS) is the main cause of the TMA that develops in association with chemotherapeutic agents (e.g., mitomycin C, gemcitabine) and immunosuppressive agents (cyclosporine, interferon, sirolimus, and tacrolimus). This process is usually dose-dependent. Alternatively, TMA may develop as a result of drug-induced autoantibodies. This form is less likely to be dose-dependent and can, in fact, occur after a single dose in patients with previous exposure. ADAMTS13 deficiency is found in fewer than half of patients with clopidogrel-associated TTP. Quinine appears to induce autoantibodies to granulocytes, lymphocytes, endothelial cells, and platelet glycoprotein Ib/IX or IIb/IIIa complexes, but not to ADAMTS13. Quinine-associated TTP is more common among women. TMA has also been reported with drugs that inhibit vascular endothelial growth factor, such as bevacizumab; the mechanism is not completely understood.

TREATMENT

TREATMENT

HUS/TTP

Treatment should be based on pathophysiology. Autoantibody-mediated TTP and DEAP HUS respond to the combination of plasma exchange and prednisone. In addition to removing the autoantibodies, plasma exchange with fresh-frozen plasma replaces ADAMTS13. Twice-daily plasma exchanges with administration of rituximab may be effective in refractory cases. Plasma infusion is usually sufficient to replace the ADAMTS13 in Upshaw-Schülman syndrome. Plasma exchange should be considered if larger volumes are necessary. Drug-induced TMA secondary to endothelial damage typically does not respond to plasma exchange and is treated primarily by discontinuing use of the agent and providing supportive care. Similarly, STEC HUS should be treated with supportive measures as plasma exchange has not been found to be effective. Antimotility agents and antibiotics increase the incidence of HUS among children, but azithromycin was recently found to decrease the duration of bacterial shedding by adults. Plasma infusion/exchange is effective in certain types of aHUS as it replaces complement-regulatory proteins. Eculizumab is a monoclonal antibody to C5 that is approved for use in aHUS which has been shown to abort MAHA and improve renal function. Antibiotics and washed red cells should be given in neuraminidase-associated HUS, and plasmapheresis may be helpful. However, plasma and whole-blood transfusion should be avoided since these products contain IgM, which may exacerbate MAHA. Finally, combined factor H and ADAMTS13 deficiency have been reported. The affected patients are generally less responsive to plasma infusion, a result illustrating the complexity of the management of these cases.

HEMATOPOIETIC STEM CELL TRANSPLANTATION–ASSOCIATED THROMBOTIC MICROANGIOPATHY (HSCT-TMA)

HSCT-TMA develops after HSCT, with an incidence of ~8%. Etiologic factors include conditioning regimens, immunosuppression, infections, and graft-versus-host disease. Other risk factors include female sex and human leukocyte antigen (HLA)–mismatched donor grafts. HSCT-TMA usually occurs within the first 100 days of HSCT. **Table 311-1** lists definitions of HSCT-TMA currently used for clinical trials. Diagnosis may be difficult since thrombocytopenia, anemia, and renal insufficiency are common after HSCT. HSCT-TMA carries a high mortality rate (75% within 3 months). The majority of patients have >10% ADAMTS13 activity, and plasma exchange is beneficial in <25% of patients. Discontinuation of calcineurin inhibitors and treatment of infections or sinusoidal obstruction syndrome (if present) are recommended. Treatment with rituximab and defibrotide may also be helpful, but clinical trial data are lacking.

TABLE 311-1

Criteria for Establishing Microangiopathic Kidney Injury Associated with Hematopoietic Stem Cell Transplantation

International Working Group	Blood and Marrow Transplant Clinical Trials Network Toxicity Committee
4% schistocytes in the blood	RBC fragmentation and at least 2 schistocytes per high-power field
De novo, prolonged, or progressive thrombocytopenia	Concurrent increase in LDH concentration above baseline
A sudden and persistent increase in LDH concentration	Negative direct and indirect Coombs test
Decrease in hemoglobin level or increased RBC transfusion requirement	Concurrent renal and/or neurologic dysfunction without other explanations
Decrease in haptoglobin concentration	

Note: These features underscore the need to identify pathways of hemolysis and thrombocytopenia that accompany deterioration of kidney function.

Abbreviations: LDH, lactate dehydrogenase; RBC, red blood cell.

HIV-RELATED TMA

HIV-related TMA is a complication encountered mainly before widespread use of highly active antiretroviral therapy. It is seen in patients with advanced AIDS and low CD4+ T cell counts although it can be the first manifestation of HIV infection. The presence of MAHA, thrombocytopenia, and renal failure are suggestive, but renal biopsy is required for diagnosis since other renal diseases are also associated with HIV infection. Thrombocytopenia may prohibit renal biopsy in some patients. The mechanism of injury is unclear, although HIV can induce apoptosis in endothelial cells. ADAMTS13 activity is not reduced in these patients. Cytomegalovirus co-infection may also be a risk factor. Effective antiviral therapy is key, while plasma exchange should be limited to patients who have evidence of TTP.

RADIATION NEPHROPATHY

Either local or total body irradiation can produce microangiopathic injury. The kidney is one of the most radiosensitive organs, and injury can result with as little as 4–5 Gy. Such injury is characterized by renal

insufficiency, proteinuria, and hypertension usually developing ≥6 months after radiation exposure. Renal biopsy reveals classic TMA with damage to glomerular, tubular, and vascular cells, but systemic evidence of MAHA is uncommon. Because of its high incidence after allogeneic HSCT, radiation nephropathy is often referred to as *bone marrow transplant nephropathy*. No specific therapy is available, although observational evidence supports renin-angiotensin system blockade.

SCLERODERMA (PROGRESSIVE SYSTEMIC SCLEROSIS)

Kidney involvement is common (up to 52%) in patients with widespread scleroderma, with 20% of cases resulting directly from scleroderma renal crisis. Other renal manifestations in scleroderma include transient (prerenal) or medication-related forms of acute kidney injury (e.g., associated with *D*-penicillamine, nonsteroidal anti-inflammatory drugs, or cyclosporine). Scleroderma renal crisis occurs in 12% of patients with diffuse systemic sclerosis but in only 2% of those with limited systemic sclerosis. Scleroderma renal crisis is the most severe manifestation of renal involvement, and is characterized by accelerated hypertension, a rapid decline in renal function, nephrotic range proteinuria, and hematuria. Retinopathy and encephalopathy may accompany the hypertension. Salt and water retention with microvascular injury can lead to pulmonary edema. Cardiac manifestations, including myocarditis, pericarditis, and arrhythmias, denote an especially poor prognosis. Although MAHA is present in more than half of patients, coagulopathy is rare.

The renal lesion in scleroderma renal crisis is characterized by arcuate artery intimal and medial proliferation with luminal narrowing. This lesion is described as “onion-skinning” and can be accompanied by glomerular collapse due to reduced blood flow. Histologically, scleroderma renal crisis is indistinguishable from malignant hypertension, with which it can coexist. Fibrinoid necrosis and thrombosis are common. Before the availability of angiotensin-converting enzyme (ACE) inhibitors, the mortality rate for scleroderma renal crisis was >90% at 1 month. Introduction of renin-angiotensin system blockade has lowered the mortality rate to 30% at 3 years. Nearly two-thirds of patients with scleroderma renal crisis may require dialysis support, with recovery of renal function in 50% (median time, 1 year). Glomerulonephritis and vasculitis associated with antineutrophil cytoplasmic antibodies and systemic lupus erythematosus have been described in patients with scleroderma. An association has been found with a speckled pattern of antinuclear antibodies and with antibodies to RNA polymerases I and III. Anti-U3-RNP may identify young patients at risk for scleroderma renal crisis. Anticentromere antibody, in contrast, is a negative predictor of this disorder. Because of the overlap between scleroderma renal crisis and other autoimmune disorders, a renal biopsy is recommended for patients with atypical renal involvement, especially if hypertension is absent.

Treatment with ACE inhibition is the first-line therapy unless contraindicated. The goal of therapy is to reduce systolic and diastolic blood pressure by 20 mmHg and 10 mmHg, respectively, every 24 h until blood pressure is normal. Additional antihypertensive therapy may be given once the dose of drug for ACE inhibition is maximized. Both ACE inhibitors and angiotensin II receptor antagonists are effective, although data suggest that treatment with ACE inhibitors is superior. ACE inhibition alone does not prevent scleroderma renal crisis,

but it does reduce the impact of hypertension. Intravenous iloprost has been used in Europe for blood pressure management and improvement of renal perfusion. Kidney transplantation is not recommended for 2 years after the start of dialysis since delayed recovery may occur.

ANTIPHOSPHOLIPID SYNDROME

Antiphospholipid syndrome (Chap. 350) can be either primary or secondary to systemic lupus erythematosus. It is characterized by a predisposition to systemic thrombosis (arterial and venous) and fetal morbidity mediated by antiphospholipid antibodies—mainly anticardiolipin antibodies (IgG, IgM, or IgA), lupus anticoagulant, or anti-β-2 glycoprotein I antibodies (antiβ2GPI). Patients with both anticardiolipin antibodies and antiβ2GPI appear to have the highest risk of thrombosis. The vascular compartment within the kidney is the main site of renal involvement. Arteriosclerosis is commonly present in the arcuate and intralobular arteries. In the intralobular arteries, fibrous intimal hyperplasia characterized by intimal thickening secondary to intense myofibroblastic intimal cellular proliferation with extracellular matrix deposition is frequently seen along with onion-skinning. Arterial and arteriolar fibrous and fibrocellular occlusions are present in more than two-thirds of biopsy samples. Cortical necrosis and focal cortical atrophy may result from vascular occlusion. TMA is commonly present in renal biopsies, although signs of MAHA and platelet consumption are usually absent. TMA is especially common in the catastrophic variant of antiphospholipid syndrome. In patients with secondary antiphospholipid syndrome, other glomerulopathies may be present, including membranous nephropathy, minimal change disease, focal segmental glomerulosclerosis, and pauci-immune crescentic glomerulonephritis.

Large vessels can be involved in antiphospholipid syndrome and may form the proximal nidus near the ostium for thrombosis of the renal artery. Renal vein thrombosis can occur and should be suspected in patients with lupus anticoagulant who develop nephrotic-range proteinuria. Progression to end-stage renal disease can occur, and a thrombosis may form in the vascular access and the renal allografts. Hypertension is common. Treatment entails lifelong anticoagulation. Glucocorticoids may be beneficial in accelerated hypertension. Immunosuppression and plasma exchange may be helpful for catastrophic episodes of antiphospholipid syndrome but by themselves do not reduce recurrent thrombosis.

HELLP SYNDROME

HELLP (hemolysis, elevated liver enzymes, low platelets) syndrome is a dangerous complication of pregnancy associated with microvascular injury. Occurring in 0.2–0.9% of all pregnancies and in 10–20% of women with severe preeclampsia, this syndrome carries a mortality rate of 7.4–34%. Most commonly developing in the third trimester, 10% of cases occur before week 27 and 30% post-partum. Although a strong association exists between HELLP syndrome and preeclampsia, nearly 20% of cases are not preceded by recognized preeclampsia. Risk factors include abnormal placentation, family history, and elevated levels of fetal mRNA for FLT1 (vascular endothelial growth factor receptor 1) and endoglin. Patients with HELLP syndrome have higher levels of inflammatory markers (C-reactive protein, IL-1Ra, and IL-6) and soluble HLA-DR than do those with preeclampsia alone.

Renal failure occurs in half of patients with HELLP syndrome, although the etiology is not well understood. Limited data suggest that renal failure is the result of both preeclampsia and acute tubular necrosis. Renal histologic findings are those of TMA with endothelial cell swelling and occlusion of the capillary lumens, but luminal thrombi are typically absent. However, thrombi become more common in severe eclampsia and HELLP syndrome. Although renal failure is common, the organ that defines this syndrome is the liver. Subcapsular hepatic hematomas sometimes produce spontaneous rupture of the liver and can be life-threatening. Neurologic complications such as cerebral infarction, cerebral and brainstem hemorrhage, and cerebral edema are other potentially life-threatening complications. Nonfatal complications include placental abruption, permanent vision loss due to Purtscher-like (hemorrhagic and vaso-occlusive vasculopathy) retinopathy, pulmonary edema, bleeding, and fetal demise.

Many features are shared by HELLP syndrome and MAHA. Diagnosis of HELLP syndrome is complicated by the fact that aHUS and TTP also can be triggered by pregnancy and complement mutations are common (30–40%) among patients with HELLP syndrome. Patients with antiphospholipid syndrome also have an elevated risk of HELLP syndrome. A history of MAHA before pregnancy is of diagnostic value. Serum levels of ADAMTS13 activity are reduced (by 30–60%) in HELLP syndrome but not to the levels seen in TTP (<5%). Determination of the ratio of lactate dehydrogenase to aspartate aminotransferase may be helpful. This ratio is 13:1 in patients with HELLP syndrome and preeclampsia as opposed to 29:1 in patients without preeclampsia. Other markers, such as antithrombin III (decreased in HELLP syndrome but not in TTP) and D-dimer (elevated in HELLP syndrome but not in TTP), may also be useful. HELLP syndrome usually resolves spontaneously after delivery, although a small percentage of HELLP cases occur post-partum. Glucocorticoids may decrease inflammatory markers, although two randomized controlled trials failed to show much benefit. Plasma exchange should be considered if hemolysis is refractory to glucocorticoids and/or delivery, especially if TTP has not been ruled out.

Myeloproliferative Neoplasm-Related Glomerulopathy

While MAHA is often present in TMA, this is not true for all lesions. Two conditions are now recognized to present with renal TMA but no evidence of systemic MAHA. The first is MPN-related glomerulopathy. MPN represents a group of clonal disorders that includes chronic myelogenous leukemia (CML), polycythemia vera (PV), essential thrombocythemia (ET), primary myelofibrosis (PMF), systemic mastocytosis, chronic eosinophilic leukemia not otherwise specified, chronic neutrophilic leukemia, and unclassifiable MPN. These patients present with renal impairment and nephrotic range proteinuria. MPN-related glomerulopathy usually occurs late in the course of the hematologic condition as median time from diagnosis of MPN to glomerulopathy was 7.2 years. Renal biopsy shows mesangial expansion, hypercellularity, mesangial and segmental sclerosis, luminal hyalinosis, loss of overlying podocytes, and adhesions to Bowman's capsule and duplication of glomerular basement membranes. Foot process effacement ranges from 30 to 95%. Arteriosclerosis is common and ranges from mild to severe. Arteriolar hyalinosis can also be seen. Extremedullary hematopoiesis can sometimes be seen especially in patients with myelofibrosis. MPGN-related glomerulopathy may develop while patients are on treatment with hydroxyurea and JAK2 inhibitors. No standard treatment is available. RAS blockade and corticosteroids have been tried with mixed results.

POEMS Syndrome

POEMS syndrome is a systemic disease characterized by polyneuropathy, organomegaly, endocrinopathy, monoclonal gammopathy, and skin changes. Peripheral neuropathy with severe motorsensory deficit is the hallmark of the disease. Another characteristic is that >95% of monoclonal light chain is of the lambda isotype. IgA also makes up about 50% of the monoclonal protein. Organomegaly can involve any organ and often presents as lymphadenopathy. In the kidney, the hypertrophy frequently is unilateral. One study suggests the difference in kidney size is due to unilateral contraction; however, a volumemetric study showed that enlargement is responsible for the difference in kidney size in some patients. Glomerulomegaly is not uncommon. Lobular appearance, endothelial cell swelling, hypercellularity, mesangiolysis, microaneurysm, and glomerular enlargement are reminiscent of membranoproliferative glomerulonephritis. Most patients present with mild to moderate renal impairment and low grade proteinuria. Progression to end stage renal disease is rare.

SICKLE CELL NEPHROPATHY

Renal complications in sickle cell disease result from occlusion of the vasa recta in the renal medulla. The low partial pressure of oxygen and high osmolarity predispose to hemoglobin S polymerization and erythrocyte sickling. Sequelae include hyposthenuria, hematuria, and papillary necrosis (which can also occur in sickle trait). The kidney responds by increases in blood flow and glomerular filtration rate mediated by prostaglandins. This dependence on prostaglandins may explain the greater reduction of glomerular filtration rate by nonsteroidal anti-inflammatory drugs in these patients than in others. The glomeruli are typically enlarged. Intracapillary fragmentation and phagocytosis of sickled erythrocytes are thought to be responsible for the membranoproliferative glomerulonephritis–like lesion, and focal segmental glomerulosclerosis is seen in more advanced cases. Proteinuria is present in 20–30%, and nephrotic-range proteinuria is associated with progression to renal failure. ACE inhibitors reduce proteinuria, although data are lacking on prevention of renal failure. Patients with sickle cell disease are also more prone to acute renal failure. The cause is thought to reflect microvascular occlusion associated with nontraumatic rhabdomyolysis, high fever, infection, and generalized sickling. Chronic kidney disease is present in 12–20% of patients. Despite the frequency of renal disease, hypertension is uncommon in patients with sickle cell disease.

RENAL VEIN THROMBOSIS

Renal vein thrombosis either can present with flank pain, tenderness, hematuria, rapid decline in renal function, and proteinuria or can be silent. Occasionally, renal vein thrombosis is identified during a workup for pulmonary embolism. The left renal vein is more commonly involved, and two-thirds of cases are bilateral. Etiologies can be divided into three broad categories: endothelial damage, venous stasis, and hypercoagulability. Homocystinuria, endovascular intervention, and surgery can produce vascular endothelial damage. Dehydration, which is more common among male patients, is a common cause of stasis in the pediatric population. Stasis also can result from compression and kinking of the renal veins from

retroperitoneal processes such as retroperitoneal fibrosis and abdominal neoplasms. Thrombosis can occur throughout the renal circulation, including the renal veins, with antiphospholipid syndrome. Renal vein thrombosis can also be secondary to nephrotic syndrome, particularly membranous nephropathy. Other hypercoagulable states less commonly associated with renal vein thrombosis include proteins C and S, antithrombin deficiency, factor V Leiden, disseminated malignancy, and oral contraceptives. Severe nephrotic syndrome may also predispose patients to renal vein thrombosis.

Diagnostic screening can be performed with Doppler ultrasonography, which is more sensitive than ultrasonography alone. Computed tomography angiography is ~100% sensitive. Magnetic resonance angiography is another option but is more expensive. Treatment for renal vein thrombosis consists of anticoagulation and therapy for the underlying cause. Endovascular thrombolysis may be considered in severe cases. Occasionally, nephrectomy may be undertaken for life-threatening complications. Vena caval filters are often used to prevent migration of thrombi.

FURTHER READING

Al-Nouri ZL et al: Drug-induced thrombotic microangiopathy: A systematic review of published reports. Blood 125:616, 2015. [PubMed: 25414441]

Bose N et al: Scleroderma renal crisis. Sem Arthritis and Rheumatism 44:687, 2015.

Bresin E et al: Combined complement gene mutations in atypical hemolytic uremic syndrome influence clinical phenotype. J Am Soc Nephrol 24:475, 2013. [PubMed: 23431077]

George JN, Nester CM: Syndromes of thrombotic microangiopathy. N Engl J Med 371:1847, 2014. [PubMed: 25372103]

Go RS et al: Thrombotic Microangiopathy Care Pathway: A Consensus Statement for the Mayo Clinic Complement Alternative Pathway-Thrombotic Microangiopathy (CAP-TMA) Disease-Oriented Group. Mayo Clin Proc 91:1189, 2016. [PubMed: 27497856]

Chapter 312: Nephrolithiasis

Gary C. Curhan

INTRODUCTION

Nephrolithiasis, or kidney stone disease, is a common, painful, and costly condition. Each year, billions of dollars are spent on nephrolithiasis-related activity, with the majority of expenditures on surgical treatment of existing stones. While a stone may form due to crystallization of lithogenic factors in the upper urinary tract, it can subsequently move into the ureter and cause renal colic. Although nephrolithiasis is rarely fatal, patients who have had renal colic report that it is the worst pain they have ever experienced. The evidence on which to base clinical recommendations is not as strong as desired; nonetheless, most experts agree that the recurrence of most, if not all, types of stones can be prevented with careful evaluation and targeted recommendations. Preventive treatment may be lifelong; therefore, an in-depth understanding of this condition must inform the implementation of tailored interventions that are most appropriate for and acceptable to the patient.

There are several types of kidney stones. It is clinically important to identify the stone type, which informs prognosis and selection of the optimal preventive regimen. Calcium oxalate stones are most common (~75%); next, in order, are calcium phosphate (~15%), uric acid (~8%), struvite (~1%), and cystine (<1%) stones. Many stones are a mixture of crystal types (e.g., calcium oxalate and calcium phosphate) and also contain protein in the stone matrix. Rarely, stones are composed of medications, such as acyclovir, atazanavir, and triamterene. Stones that form as a result of an upper tract infection, if not appropriately treated, can have devastating consequences and lead to end-stage renal disease. Consideration should be given to teaching practitioners strategies to prevent recurrence of all stone types and the related morbidity.

EPIDEMIOLOGY


Nephrolithiasis is a global disease. Data suggest an increasing prevalence, likely due to Westernization of lifestyle habits (e.g., dietary changes, increasing body mass index). National Health and Nutrition Examination Survey data for 2007–2010 indicate that up to 19% of men and 9% of women will develop at least one stone during their lifetime. The prevalence is ~50% lower among black individuals than among whites. The incidence of nephrolithiasis (i.e., the rate at which previously unaffected individuals develop their first stone) also varies by age, sex, and race. Among white men, the peak annual incidence is ~3.5 cases/1000 at age 40 and declines to ~2 cases/1000 by age 70. Among white women in their thirties, the annual incidence is ~2.5 cases/1000; the figure decreases to ~1.5/1000 at age 50 and beyond. In addition to

the medical costs associated with nephrolithiasis, this condition also has a substantial economic impact, as those affected are often of working age. Once an individual has had a stone, the prevention of a recurrence is essential. Published recurrence rates vary by the definitions and diagnostic methods used. Some reports have relied on symptomatic events, while others have been based on imaging. Most experts agree that radiographic evidence of a second stone should be considered to represent a recurrence, even if the stone has not yet caused symptoms.

ASSOCIATED MEDICAL CONDITIONS

Nephrolithiasis is a systemic disorder. Several conditions predispose to stone formation, including gastrointestinal malabsorption (e.g., Crohn's disease, gastric bypass surgery), primary hyperparathyroidism, obesity, type 2 diabetes mellitus, and distal renal tubular acidosis. A number of other medical conditions are more likely to be present in individuals with a history of nephrolithiasis, including hypertension, gout, cardiovascular disease, cholelithiasis, reduced bone mineral density, and chronic kidney disease.

Although nephrolithiasis does not directly cause upper urinary tract infections (UTIs), a UTI in the setting of an obstructing stone is a urologic emergency ("pus under pressure") and requires urgent intervention to reestablish drainage.

PATHOGENESIS

In the consideration of the processes involved in crystal formation, it is helpful to view urine as a complex solution. A clinically useful concept is *supersaturation* (the point at which the concentration product exceeds the solubility product). However, even though the urine in most individuals is supersaturated with respect to one or more types of crystals, the presence of inhibitors of crystallization prevents the majority of the population from continuously forming stones. The most clinically important inhibitor of calcium-containing stones is urine citrate. While the calculated supersaturation value does not perfectly predict stone formation, it is a useful guide as it integrates the multiple factors that are measured in a 24-h urine collection.

Recent studies have changed the paradigm for the site of initiation of stone formation. Renal biopsies of stone formers have revealed calcium phosphate in the renal interstitium. It is hypothesized that this calcium phosphate deposits at the thin limb of the loop of Henle, and then extends down to the papilla and erodes through the papillary epithelium, where it provides a site for deposition of calcium oxalate and calcium phosphate crystals. The majority of calcium oxalate stones grow on calcium phosphate at the tip of the renal papilla (*Randall's plaque*). Tubular plugs of calcium phosphate may be the initiating event in calcium phosphate stone development. Thus, the process of stone formation may begin years before a clinically detectable stone is identified. The processes involved in interstitial deposition are under active investigation.

RISK FACTORS

Risk factors for nephrolithiasis can be categorized as dietary, nondietary, or urinary. These risk factors vary by stone type and by clinical characteristics.

Dietary Risk Factors

Patients who develop stones often change their diet; therefore, studies that retrospectively assess diet may be hampered by recall bias. Some studies have examined the relation between diet and changes in the lithogenic composition of the urine, often using calculated supersaturation. However, the composition of the urine does not perfectly predict risk, and not all components that modify risk are included in the calculation of supersaturation. Thus, dietary associations are best investigated by prospective studies that examine actual stone formation as the outcome. Dietary factors that are associated with an increased risk of nephrolithiasis include animal protein, oxalate, sodium, sucrose, and fructose. Dietary factors associated with a lower risk include calcium, potassium, and phytate.

CALCIUM

The role of dietary calcium deserves special attention. Although in the distant past dietary calcium had been suspected of increasing the risk of stone disease, several prospective observational studies and a randomized controlled trial have demonstrated that higher dietary calcium intake is related to a *lower* risk of stone formation. The reduction in risk associated with higher calcium intake may be due to a reduction in intestinal absorption of dietary oxalate that results in lower urine oxalate. Low calcium intake is contraindicated as it increases the risk of stone formation and may contribute to lower bone density in stone formers.

Despite similar bioavailability, supplemental calcium may increase the risk of stone formation. The discrepancy between the risks from dietary calcium and calcium supplements may be due to the timing of supplemental calcium intake or to higher total calcium consumption leading to higher urinary calcium excretion.

OXALATE

Urinary oxalate is derived from both endogenous production and absorption of dietary oxalate. Owing to its low and often variable bioavailability, much of the oxalate in food may not be readily absorbed. However, absorption may be higher in stone formers. Although observational studies demonstrate that dietary oxalate is only a weak risk factor for stone formation, urinary oxalate is a strong risk factor for calcium oxalate stone formation, and efforts to avoid high oxalate intake should thus be beneficial.

OTHER NUTRIENTS

Several other nutrients have been studied and implicated in stone formation. Higher intake of animal protein may lead to increased excretion of calcium and uric acid as well as to decreased urinary excretion of citrate, all of which increase the risk of stone formation. Higher sodium and sucrose intake increases calcium

excretion independent of calcium intake. Higher potassium intake decreases calcium excretion, and many potassium-rich foods increase urinary citrate excretion due to their alkali content. Other dietary factors that have been inconsistently associated with lower stone risk include magnesium and phytate.

Vitamin C supplements are associated with an increased risk of calcium oxalate stone formation in men, possibly because of raised levels of oxalate in urine. Thus, male calcium oxalate stone formers should be advised to avoid vitamin C supplements. Although high doses of supplemental vitamin B_6 may be beneficial in selected patients with type 1 primary hyperoxaluria, the risk is not reduced in other patients.

FLUIDS AND BEVERAGES

The risk of stone formation increases as urine volume decreases. When the urine output is <1 L/d, the risk of stone formation more than doubles. Fluid intake is the main determinant of urine volume, and the importance of fluid intake in preventing stone formation has been demonstrated in observational studies and in a randomized controlled trial. Observational studies have found that coffee, tea, beer, wine, and orange juice are associated with a reduced risk of stone formation. Sugar-sweetened beverage consumption may increase risk.

Nondietary Risk Factors

Age, race, body size, and environment are important risk factors for nephrolithiasis. The incidence of stone disease is highest in middle-aged white men, but stones can form in infants as well as in the elderly. There is geographic variability, with the highest prevalence in the southeastern United States. Weight gain increases the risk of stone formation, and the increasing prevalence of nephrolithiasis in the United States may be due in part to the increasing prevalence of obesity. Environmental and occupational influences that may lead to lower urine volume, such as working in a hot environment or lack of ready access to water or a bathroom, are important considerations.

Urinary Risk Factors

URINE VOLUME

As mentioned above, lower urine volume results in higher concentrations of lithogenic factors and is a common and readily modifiable risk factor. A randomized trial has demonstrated the effectiveness of higher fluid intake in increasing urine volume and reducing the risk of stone recurrence.

URINE CALCIUM

Higher urine calcium excretion increases the likelihood of formation of calcium oxalate and calcium phosphate stones. While the term *hypercalciuria* is often used, there is no widely accepted cutoff that distinguishes between normal and abnormal urine calcium excretion. In fact, the relation between urine calcium and stone risk appears to be continuous; thus, the use of an arbitrary threshold should be avoided.

Levels of urine calcium excretion are higher in individuals with a history of nephrolithiasis; however, the mechanisms remain poorly understood. Greater gastrointestinal calcium absorption is one important contributor, and greater bone turnover (with a resultant reduction in bone mineral density) may be another. Primary renal calcium loss, with lower serum calcium concentrations and elevated serum levels of parathyroid hormone (PTH) (and a normal 25-hydroxy vitamin D level), is rare.

URINE OXALATE

Higher urine oxalate excretion increases the likelihood of calcium oxalate stone formation. As for urine calcium, no definition for “abnormal” urine oxalate excretion is widely accepted. Given that the relation between urine oxalate and stone risk is continuous, simple dichotomization of urine oxalate excretion is not helpful in assessing risk. The two sources of urine oxalate are endogenous generation and dietary intake. Dietary oxalate is the major contributor and also the source that can be modified. Notably, higher dietary calcium intake reduces gastrointestinal oxalate absorption and thereby reduces urine oxalate.

URINE CITRATE

Urine citrate is a natural inhibitor of calcium-containing stones; thus, lower urine citrate excretion increases the risk of stone formation. Citrate reabsorption is influenced by the intracellular pH of proximal tubular cells. Metabolic acidosis, including that due to higher animal flesh intake, will lead to a reduction in citrate excretion by increasing reabsorption of filtered citrate. However, a notable proportion of patients have lower urine citrate for reasons that remain unclear.

URINE URIC ACID

Higher urine levels of uric acid—a risk factor for uric acid stone formation—are found in individuals with excess purine consumption and rare genetic conditions that lead to overproduction of uric acid. This characteristic does not appear to be associated with the risk of calcium oxalate stone formation.

URINE pH

Urine pH influences the solubility of some crystal types. Uric acid stones form only when the urine pH is consistently ≤5.5 or lower, whereas calcium phosphate stones are more likely to form when the urine pH is ≥6.5 or higher. Cystine is more soluble at higher urine pH. Calcium oxalate stones are not influenced by urine pH.

Genetic Risk Factors



The risk of nephrolithiasis is more than twofold greater in individuals with a family history of stone disease. This association is likely due to a combination of genetic predisposition and similar environmental

exposures. While a number of rare monogenic disorders cause nephrolithiasis, the genetic contributors to common forms of stone disease remain to be determined.

The two most common and well-characterized rare monogenic disorders that lead to stone formation are primary hyperoxaluria and cystinuria. *Primary hyperoxaluria* is an autosomal recessive disorder that causes excessive endogenous oxalate generation by the liver, with consequent calcium oxalate stone formation and crystal deposition in organs. Intraparenchymal calcium oxalate deposition in the kidney can eventually lead to renal failure. *Cystinuria* is an autosomal recessive disorder that causes abnormal reabsorption of filtered basic amino acids. The excessive urinary excretion of cystine, which is poorly soluble, leads to cystine stone formation. Cystine stones are visible on plain radiographs and often manifest as staghorn calculi or multiple bilateral stones. Repeat episodes of obstruction and instrumentation can cause a reduction in the glomerular filtration rate (GFR).

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Nephrolithiasis

Evidence-based guidelines for the evaluation and treatment of nephrolithiasis have been recently published. Although there is limited evidence for several aspects, there are standard approaches to patients with acute and chronic presentations that can reasonably guide the clinical evaluation.

It typically requires weeks to months (and often much longer) for a kidney stone to grow to a clinically detectable size. Although the passage of a stone is a dramatic event, stone formation and growth are characteristically clinically silent. A stone can remain asymptomatic in the kidney for years or even decades before signs (e.g., hematuria) or symptoms (e.g., pain) become apparent. Thus, it is important to remember that the onset of symptoms, typically attributable to a stone moving into the ureter, does not provide insight into when the stone actually formed. The factors that induce stone movement are unknown.

Clinical Presentation and Differential Diagnosis

There are two common presentations for individuals with an acute stone event: renal colic and painless gross hematuria. *Renal colic* is a misnomer because pain typically does not subside completely; rather, it varies in intensity. When a stone moves into the ureter, the discomfort often begins with a sudden onset of unilateral flank pain.

The intensity of the pain can increase rapidly, and there are no alleviating factors. This pain, which is accompanied often by nausea and occasionally by vomiting, may radiate, depending on the location of the stone. If the stone lodges in the upper part of the ureter, pain may radiate anteriorly; if the stone is in the lower part of the ureter, pain can radiate to the ipsilateral testicle in men or the ipsilateral labium in women. Occasionally, a patient has gross hematuria without pain.

Other diagnoses may be confused with acute renal colic. If the stone is lodged at the right ureteral pelvic junction, symptoms may mimic those of acute cholecystitis. If the stone blocks the ureter as it crosses over the right pelvic brim, symptoms may mimic acute appendicitis, whereas blockage at the left pelvic brim may be confused with acute diverticulitis. If the stone lodges in the ureter at the ureterovesical junction, the patient may experience urinary urgency and frequency. In female patients, the latter symptoms may lead to an incorrect diagnosis of bacterial cystitis; the urine will contain red and white blood cells, but the urine culture will be negative. An obstructing stone with proximal infection may present as acute pyelonephritis. A UTI in the setting of ureteral obstruction is a medical emergency that requires immediate restoration of drainage by placement of either a ureteral stent or a percutaneous nephrostomy tube. Other conditions to consider in the differential diagnosis include muscular or skeletal pain, *herpes zoster*, duodenal ulcer, abdominal aortic aneurysm, gynecologic conditions, ureteral stricture, and ureteral obstruction by materials other than a stone, such as a blood clot or sloughed papilla. Extraluminal processes can lead to ureteral compression and obstruction; however, because of the gradual onset, these conditions do not typically present with renal colic.

Diagnosis and Intervention

Serum chemistry findings are typically normal, but the white blood cell count may be elevated. Examination of the urine sediment will usually reveal red and white blood cells and occasionally crystals **(Fig. 312-1)**. The absence of hematuria does not exclude a stone, particularly when urine flow is completely obstructed by a stone.

The diagnosis is often made on the basis of the history, physical examination, and urinalysis. Thus, it may not be necessary to wait for radiographic confirmation before treating the symptoms. The diagnosis is confirmed by an appropriate imaging study—preferably helical computed tomography (CT), which is highly sensitive, allows visualization of uric acid stones (traditionally considered “radiolucent”), and does not require radiocontrast **(Fig. 312-2)**. Helical CT detects stones as small as 1 mm that may be missed by other imaging modalities.

Typically, helical CT reveals a ureteral stone or evidence of recent passage (e.g., perinephric stranding or hydronephrosis), whereas a plain abdominal radiograph (kidney/ureter/bladder, or KUB) can miss a stone in the ureter or kidney, even if it is radiopaque, and does not provide information on obstruction. Abdominal ultrasound offers the advantage of avoiding radiation and provides information on hydronephrosis, but it is not as sensitive as CT and images only the kidney and possibly the proximal segment of the ureter; thus most ureteral stones are not detectable by ultrasound.

Many patients who experience their first episode of colic seek emergent medical care. Randomized trials have demonstrated that parenterally administered nonsteroidal anti-inflammatory drugs (such as ketorolac) are just as effective as opioids in relieving symptoms and have fewer side effects. Excessive fluid administration has not been shown to be beneficial; therefore, the goal should be to maintain euvolemia. If the pain can be adequately controlled and the patient is able to take fluids orally, hospitalization can be avoided. Use of an alpha blocker may increase the rate of spontaneous stone passage.

Urologic intervention should be postponed unless there is evidence of UTI, a low probability of spontaneous stone passage (e.g., a stone measuring ≥6 mm or an anatomic abnormality), or intractable pain. A ureteral stent may be placed cystoscopically, but this procedure typically requires general anesthesia, and the stent can be quite uncomfortable, may cause gross hematuria, and may increase the risk of UTI.

If an intervention is indicated, the selection of the most appropriate intervention is determined by the size, location, and composition of the stone; the urinary tract anatomy; and the experience of the urologist. Extracorporeal shockwave lithotripsy (ESWL), the least invasive option, uses shock waves generated outside the body to fragment the stone, but is being used less frequently. An endourologic approach, now more frequently used than ESWL, can remove a stone by basket extraction or laser fragmentation. For large upper-tract stones, percutaneous nephrostolithotomy has the highest likelihood of rendering the patient stone-free. Advances in urologic approaches and instruments have nearly eliminated the need for open surgical procedures such as ureterolithotomy or pyelolithotomy.

Evaluation for Stone Prevention

More than half of first-time stone formers will have a recurrence within 10 years. A careful evaluation is indicated to identify predisposing factors, which can then be modified to reduce the risk of new stone formation. It is appropriate to proceed with an evaluation even after the first stone if the patient is interested because recurrences are common and are usually preventable with inexpensive lifestyle modifications or other treatments.

History

A detailed history, obtained from the patient and from a thorough review of medical records, should include the number and frequency of episodes (distinguishing stone passage from stone formation) and previous imaging studies, interventions, evaluations, and treatments. Inquiries about the patient's medical history should cover UTIs, bariatric surgery, gout, hypertension, and diabetes mellitus. A family history of stone disease may reveal a genetic predisposition. A complete list of current prescription and over-the-counter medications as well as vitamin and mineral supplements is essential. The review of systems should focus on identifying possible etiologic factors related to low urine volume (e.g., high insensible losses) and gastrointestinal malabsorption as well as on ascertaining how frequently the patient voids during the day and overnight.

A large body of compelling evidence has demonstrated the important role of diet in stone disease. Thus, the dietary history should encompass information on usual dietary habits (meals and snacks), calcium intake, consumption of high-oxalate foods (spinach, rhubarb, potatoes), and fluid intake (including amount of specific beverages typically consumed). Amount and frequency of use of vitamin and mineral supplements should be carefully assessed.

Physical Examination

The physical examination should assess weight, blood pressure, costovertebral angle tenderness, and lower-extremity edema as well as signs of other systemic conditions such as primary hyperparathyroidism and gout.

Laboratory Evaluation

If not recently measured, the following serum levels should be determined: electrolytes (to uncover hypokalemia or renal tubular acidosis), creatinine, calcium, and uric acid. The PTH level should be measured if indicated by high-normal or elevated serum and urine calcium concentrations. Often, 25-hydroxy vitamin D is measured in concert with PTH to investigate the possible role of secondarily elevated PTH levels in the setting of vitamin D insufficiency.

The urinalysis, including examination of the sediment, can provide useful information. In individuals with asymptomatic residual renal stones, red and white blood cells are frequently present in urine. If there is concern about the possibility of an infection, a urine culture should be performed. The sediment may also reveal crystals (Fig. 312-1), which may help identify the stone type and also provide prognostic information, as crystalluria is a strong risk factor for new stone formation.

The results from 24-h urine collections serve as the cornerstone on which therapeutic recommendations are based. Recommendations on lifestyle modification should be deferred until urine collection is complete. As a baseline assessment, patients should collect at least two 24-h urine samples while consuming their usual diet and usual volume of fluid. The following factors should be measured: total volume, calcium, oxalate, citrate, uric acid, sodium, potassium, phosphorus, pH, and creatinine. When available, the calculated supersaturation is also informative. There is substantial day-to-day variability in the 24-h excretion of many relevant factors; therefore, obtaining values from two collections is important before committing a patient to long-term lifestyle changes or medication. The interpretation of the 24-h urine results should take into account that the collections are usually performed on a weekend day when the patient is staying at home; an individual's habits may differ dramatically (beneficially or detrimentally) at work or outside the home. Specialized testing, such as calcium loading or restriction, is not recommended as it does not influence clinical recommendations.

Stone composition analysis is essential if a stone or fragment is available; patients should be encouraged to retrieve passed stones. The stone type cannot be determined with certainty from 24-h urine results, but pure uric acid stones can be identified by low Hounsfield units on CT.

Imaging

The "gold standard" diagnostic test is helical CT without contrast. If not already performed during an acute episode, a low-dose CT should be considered to definitively establish the baseline stone burden. A suboptimal imaging study may not detect a residual stone that, if subsequently passed, would be mistaken for a new stone. In this instance, the preventive medical regimen might be unnecessarily changed as the result of a preexisting stone.

Recommendations for follow-up imaging should be tailored to the individual patient. While CT provides the best information, the radiation dose is higher than from other modalities; therefore, CT should be performed only if the results will lead to a change in clinical recommendations. Although they are less sensitive, renal ultrasound or a KUB examinaion is typically used to minimize radiation exposure, with recognition of the limitations.

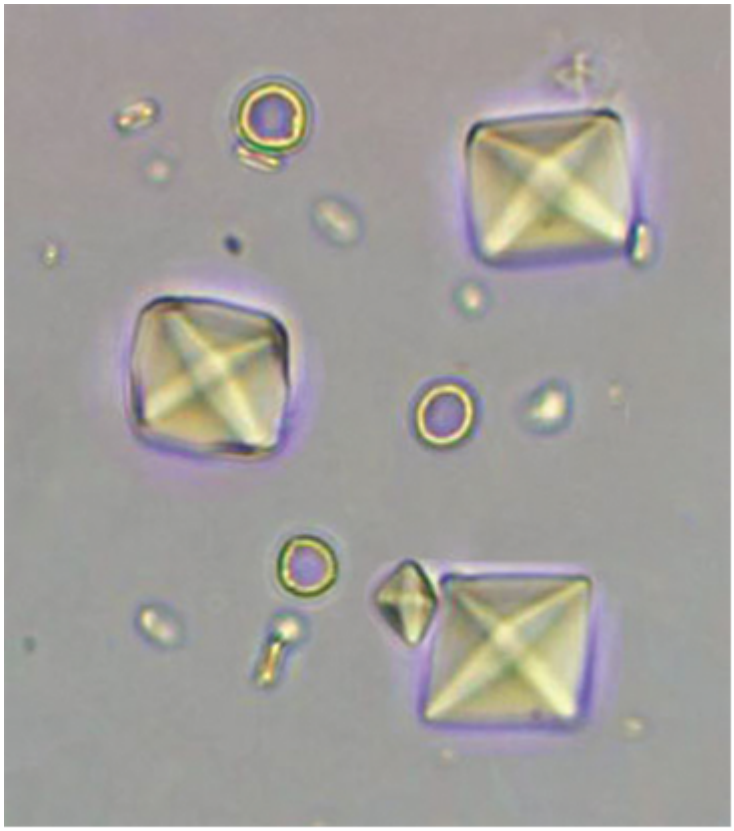
Prevention of New Stone Formation

Recommendations for preventing stone formation depend on the stone type and the results of metabolic evaluation. After remediable secondary causes of stone formation (e.g., primary hyperparathyroidism) are excluded, the focus should turn to modification of the urine composition to reduce the risk of new stone formation. The urinary constituents are continuous variables, and the associated risk is continuous; thus, there are no definitive thresholds. Dichotomization into “normal” and “abnormal” can be misleading and should be avoided.

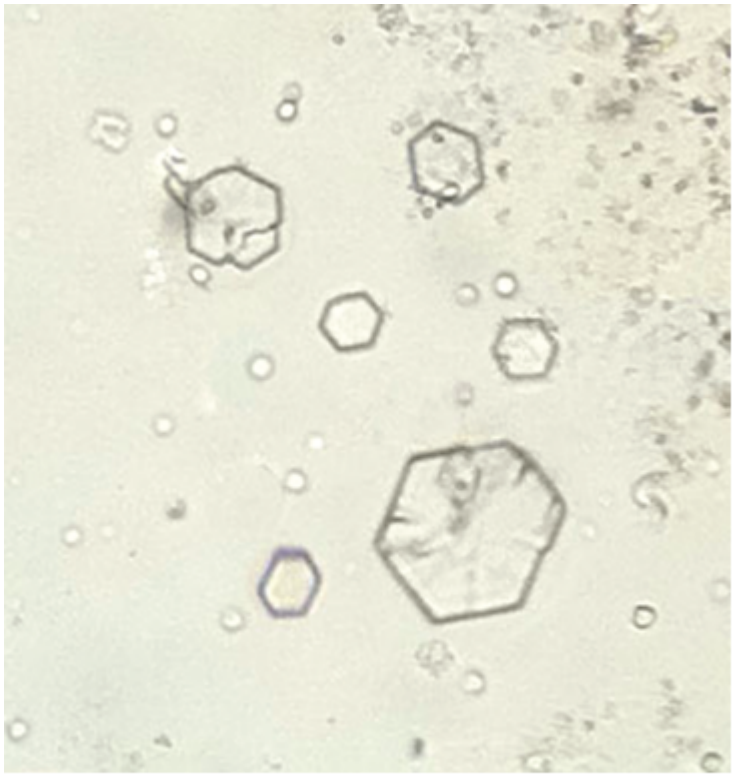
For all stone types, consistently diluted urine reduces the likelihood of crystal formation. The urine volume should be at least 2 L/d. Because of differences in insensible fluid losses and fluid intake from food sources, the required total fluid intake will vary from person to person. Rather than specify how much to drink, it is more helpful to educate patients about how much *more* they need to drink in light of their 24-h urine volume. For example, if the daily urine volume is 1.5 L, then the patient should be advised to drink at least 0.5 L more per day in order to increase the urine volume to the goal of 2 L/day.

FIGURE 312-1

Urine sediment from a patient with calcium oxalate stones (*left*) and a patient with cystine stones (*right*). Calcium oxalate dihydrate crystals are bipyramidally shaped, and cystine crystals are hexagonal. *(Left panel image courtesy of Dr. John Lieske, Mayo Clinic.)*



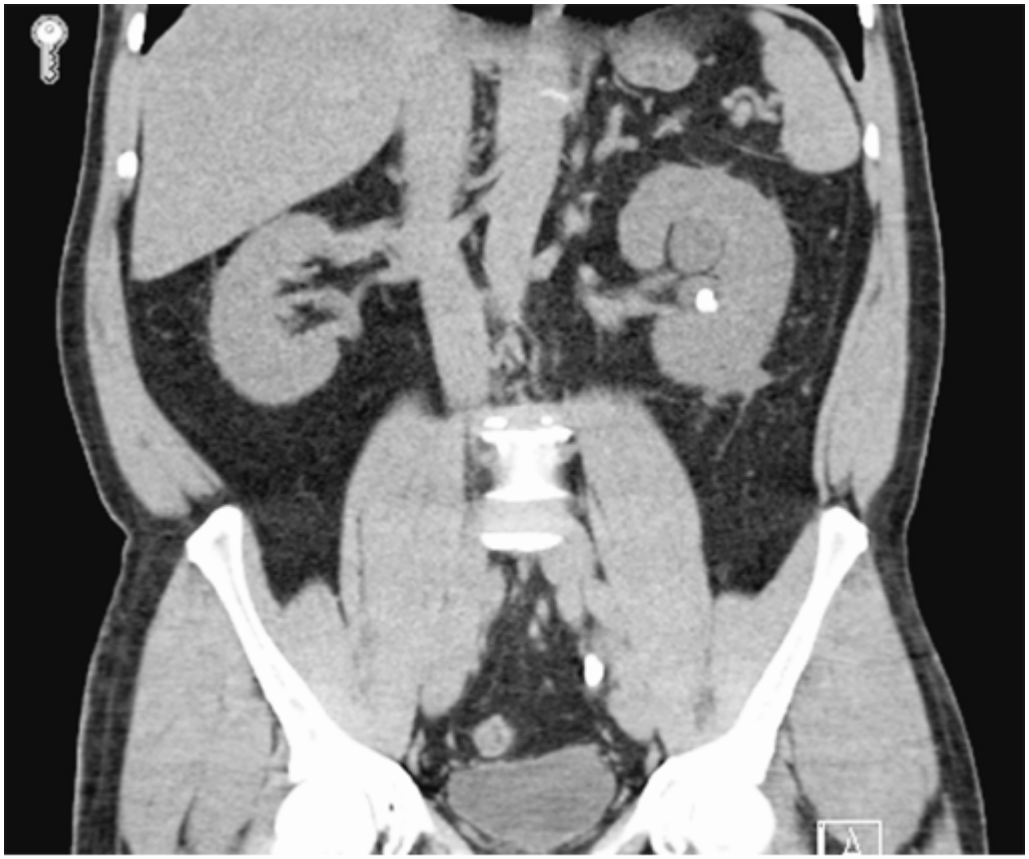
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 312-2

Coronal noncontrast CT image from a patient who presented with left-sided renal colic. An obstructing calculus, present in the distal left ureter at the level of S1, measures 10 mm in maximal dimension. There is severe left hydroureteronephrosis and associated left perinephric fat stranding. In addition, there is a nonobstructing 6-mm left renal calculus in the interpolar region. *(Image courtesy of Dr. Stuart Silverman, Brigham and Women's Hospital.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

RECOMMENDATIONS FOR SPECIFIC STONE TYPES

Calcium Oxalate

Risk factors for calcium oxalate stones include higher urine calcium, higher urine oxalate, and lower urine citrate. This stone type is insensitive to pH in the physiologic range.

Individuals with higher urine calcium excretion tend to absorb a higher percentage of ingested calcium. Nevertheless, dietary calcium restriction is not beneficial and, in fact, is likely to be harmful (see "Dietary Risk Factors," above). In a randomized trial in men with high urine calcium and recurrent calcium oxalate stones, a diet containing 1200 mg of calcium and a low intake of sodium and animal protein significantly reduced subsequent stone formation from that with a low-calcium diet (400 mg/d). Excessive calcium intake (>1200 mg/d) should be avoided.

A thiazide diuretic, in doses higher than those used to treat hypertension, can substantially lower urine calcium excretion. Several randomized controlled trials have demonstrated that thiazide diuretics, most commonly chlorthalidone, can reduce calcium oxalate stone recurrence by ~50%. When a thiazide is prescribed, dietary sodium restriction is essential to obtain the desired reduction in urinary calcium excretion and minimize urinary potassium losses. While bisphosphonates may reduce urine calcium excretion in some individuals, there are no data on whether this class of medication can reduce stone

formation; therefore, bisphosphonates cannot be recommended solely for stone prevention at present but they can be used to treat those individuals with low bone density.

A reduction in urine oxalate will in turn reduce the supersaturation of calcium oxalate. In patients with the common form of nephrolithiasis, avoiding high-dose vitamin C supplements is the only known strategy that reduces endogenous oxalate production.

Oxalate is a metabolic end product; therefore, any dietary oxalate that is absorbed will be excreted in the urine. Reducing absorption of exogenous oxalate involves two approaches. First, the avoidance of foods that contain high amounts of oxalate, such as spinach, rhubarb, almonds, and potatoes, is prudent. However, extreme oxalate restriction has not been demonstrated to reduce stone recurrence and could be harmful to overall health, given other health benefits of many foods that are erroneously considered to be high in oxalate. Controversy exists regarding the most clinically relevant measure of the oxalate content of foods (e.g., bioavailability). Notably, the absorption of oxalate is reduced by higher calcium intake; therefore, individuals with higher-than-desired urinary oxalate should be counseled to consume adequate calcium. Oxalate absorption can be influenced by the intestinal microbiota, depending on the presence of oxalate-degrading bacteria. Currently, however, there are no available therapies to alter the microbiota that beneficially affect urinary oxalate excretion over the long term.

Citrate is a natural inhibitor of calcium oxalate and calcium phosphate stones. Higher-level consumption of foods rich in alkali (i.e., fruits and vegetables) can increase urine citrate. For patients with lower urine citrate in whom dietary modification does not adequately increase urine citrate, the addition of supplemental alkali (typically potassium citrate or bicarbonate) will lead to an increase in urinary citrate excretion. Sodium salts, such as sodium bicarbonate, while successful in raising urine citrate, are typically avoided due to the adverse effects of sodium on urine calcium excretion. Urine pH in the physiologic range does not influence calcium oxalate stone formation.

Past reports suggested that higher levels of urine uric acid may increase the risk of calcium oxalate stones, but more recent studies do not support this association. However, allopurinol reduced stone recurrence in one randomized controlled trial in patients with calcium oxalate stones and high urine uric acid levels. The lack of association between urine uric acid level and calcium oxalate stones suggests that a different mechanism underlies the observed beneficial effect of allopurinol.

Additional dietary modifications may be beneficial in reducing stone recurrence. Restriction of nondairy animal protein (e.g., meat, chicken, seafood) is a reasonable approach and may result in higher excretion of citrate and lower excretion of calcium. In addition, reducing sodium intake to <2.5 g/d may decrease urinary excretion of calcium. Sucrose and fructose intake should be minimized.

For adherence to a dietary pattern that is more manageable for patients than manipulating individual nutrients, the DASH (Dietary Approaches to Stop Hypertension) diet provides an appropriate and readily available option. Randomized trials have conclusively shown the DASH diet to reduce blood pressure. At

present, only data from observational studies are available, but these demonstrate a strong and consistent inverse association between the DASH diet and risk of stone formation.

Calcium Phosphate

Calcium phosphate stones share risk factors with calcium oxalate stones, including higher concentrations of urine calcium and lower concentrations of urine citrate, but additional factors deserve attention. Higher urine phosphate levels and higher urine pH (typically ≥6.5) are associated with an increased likelihood of calcium phosphate stone formation. Calcium phosphate stones are more common in patients with distal renal tubular acidosis and primary hyperparathyroidism.

There are no randomized trials on which to base preventive recommendations for calcium phosphate stone formers, so the interventions are focused on modification of the recognized risk factors. Thiazide diuretics (with sodium restriction) may be used to reduce urine calcium, as described above for calcium oxalate stones. In patients with low urine citrate levels, alkali supplements (e.g., potassium citrate or bicarbonate) may be used to increase these concentrations. However, the urine pH of these patients should be monitored initially because supplemental alkali can raise urine pH, thereby potentially increasing the risk of stone formation. Because these patients tend to have a urinary acidification defect, reducing the urine pH is not an option. Reduction of dietary phosphate may be beneficial by reducing urine phosphate excretion.

Uric Acid

The two main risk factors for uric acid stones are persistently low urine pH and higher uric acid excretion. Urine pH is the predominant influence on uric acid solubility; therefore, the mainstay of prevention of uric acid stone formation entails increasing urine pH. Alkalinizing the urine can be readily achieved by increasing the intake of foods rich in alkali (e.g., fruits and vegetables) and reducing the intake of foods that produce acid (e.g., animal flesh). If necessary, supplementation with bicarbonate or citrate salts (preferably potassium-based) can be used to reach the recommended pH goal of 6.5 throughout the day and night.

Urine uric acid excretion is determined by uric acid generation. Uric acid is the end product of purine metabolism; thus, reduced consumption of purine-containing foods can lower urine uric acid excretion. It is noteworthy that the serum uric acid level is dependent on the fractional excretion of uric acid and therefore does not provide information on urine uric acid excretion. For example, an individual with high uric acid generation and concurrent high fractional excretion of uric acid will have high urine uric acid excretion with a normal (or even low) serum uric acid level. If alkalinization of the urine alone is not successful and if dietary modifications do not reduce urine uric acid sufficiently, then the use of a xanthine oxidase inhibitor, such as allopurinol or febuxostat, can reduce urine uric acid excretion by 40–50%.

Cystine

Cystine excretion is not easily modified. Long-term dietary cystine restriction is not feasible and is unlikely to be successful; thus the focus for cystine stone prevention is on increasing cystine solubility. This goal may be

achieved by treatment with medication that covalently binds to cystine (tiopronin or penicillamine) and a medication that raises urine pH. Tiopronin is the preferred choice due to its better adverse event profile. The preferred alkalinizing agent to achieve a urine pH of 7.5 is potassium citrate or bicarbonate as sodium salts may increase cystine excretion. As with all stone types, and especially in patients with cystinuria, maintaining a high urine volume is an essential component of the preventive regimen.

Struvite

Struvite stones, also known as *infection stones* or *triple-phosphate stones*, form only when the upper urinary tract is infected with urease-producing bacteria such as *Proteus mirabilis, Klebsiella pneumoniae*, or *Providencia* species. Urease produced by these bacteria hydrolyzes urea and may elevate the urine pH to a supraphysiologic level (>8.0). Struvite stones may grow quickly and fill the renal pelvis (*staghorn calculi*).

Struvite stones require complete removal by a urologist. New stone formation can be avoided by the prevention of UTIs. In patients with recurrent upper UTIs (e.g., some individuals with surgically altered urinary drainage or spinal cord injury), the urease inhibitor acetohydroxamic acid can be considered; however, this agent should be used with caution because of potential side effects.

LONG-TERM FOLLOW-UP

In general, the preventive regimens described above do not cure the underlying pathophysiologic process. Thus these recommendations typically need to be followed for the patient's lifetime, and it is essential to tailor recommendations in a way that is acceptable to the patient. Because the memory of the acute stone event fades and patients often return to old habits (e.g., insufficient fluid intake), long-term follow-up, including repeat 24-h urine collections, is important to ensure that the preventive regimen has been implemented and has resulted in the desired reduction in the risk of new stone formation.

Follow-up imaging should be planned thoughtfully. Many patients with recurrent episodes of renal colic that lead to emergency room visits often undergo repeat CT studies. While CT does provide the best information, the radiation dose is substantially higher than that with plain abdominal radiography (KUB). Small stones may be missed by KUB, and ultrasound has a limited ability to determine size and number of stones. Minimizing radiation exposure should be a goal of the long-term follow-up plan and must be balanced against the gain in diagnostic information.

FURTHER READING

Pearle MS et al: Medical management of kidney stones: AUA guideline. J Urol 192:316, 2014. [PubMed: 24857648]

Prochaska ML et al: Insights into nephrolithiasis from the Nurses' Health Studies. Am J Public Health 106:1638, 2016. [PubMed: 27459448]

Chapter 313: Urinary Tract Obstruction

Julian L. Seifter

FIGURE 313-1

INTRODUCTION

Obstruction to the flow of urine, with attendant stasis and elevation in urinary tract pressure, impairs renal and urinary conduit functions and is a common cause of acute and chronic kidney disease (obstructive nephropathy). Early recognition and prompt treatment of urinary tract obstruction (UTO) can prevent or reverse devastating effects on kidney structure and function, and decrease susceptibility to hypertension, infection, and stone formation. Chronic obstruction may lead to permanent loss of renal mass (renal atrophy) and excretory capability. Since obstructive disease may be secondary to serious underlying inflammatory, vascular, or malignant disease, familiarity with clinical findings, appropriate diagnostic testing, and therapeutic approach is of great importance to the clinician.

ETIOLOGY

Obstruction to urine flow can result from *intrinsic* or *extrinsic mechanical blockade* as well as from *functional defects* not associated with fixed occlusion of the urinary drainage system. Mechanical obstruction can occur at any level of the urinary tract, from within the renal tubules, or the renal calyces to the external urethral meatus (obstructive uropathy). Normal points of narrowing, such as the ureteropelvic and ureterovesical junctions, bladder neck, and urethral meatus, are common sites of obstruction. When lower UTO is above the level of the bladder, unilateral dilatation of the ureter (*hydroureter*) and renal pyelocalyceal system (*hydronephrosis*) occurs; lesions at or below the level of the bladder cause bilateral involvement.

Common forms of obstruction are listed in Table 313-1. Childhood causes include *congenital malformations*, such as narrowing of the ureteropelvic junction (UPJ) and abnormal insertion of the ureter into the bladder, the most common cause. Vesicoureteral reflux in the absence of urinary tract infection or bladder neck obstruction often resolves with age. Reinsertion of the ureter into the bladder is indicated if reflux is severe and unlikely to improve spontaneously, if renal function deteriorates, or if urinary tract infections recur despite chronic antimicrobial therapy. Vesicoureteral reflux may cause prenatal hydronephrosis and, if severe, can lead to recurrent urinary infections, hypertension and renal scarring in childhood. Posterior urethral valves are the most common cause of bilateral hydronephrosis in boys. In adults, UTO is due mainly to *acquired defects*. Pelvic tumors, calculi, and urethral stricture predominate. Ligation of, or injury to, the ureter during pelvic or colonic surgery can lead to hydronephrosis which, if unilateral, may remain undetected. Obstructive uropathy may also result from extrinsic neoplastic (carcinoma of cervix or colon) or inflammatory disorders. Lymphomas and pelvic or colonic neoplasms with retroperitoneal involvement are causes of ureteral obstruction. As many as 50% of men aged >40 years may have lower urinary tract symptoms associated with benign prostatic hypertrophy, but these symptoms may occur without bladder outlet obstruction.

TABLE 313-1

Common Mechanical Causes of Urinary Tract Obstruction

Ureter	Bladder Outlet	Urethra
Congenital		
Ureteropelvic junction narrowing or obstruction Ureterovesical junction narrowing or obstruction and reflux Ureterocele Retrocaval ureter	Bladder neck obstruction Ureterocele	Posterior urethral valves Anterior urethral valves Stricture Meatal stenosis Phimosis
Acquired Intrinsic Defects		
Calculi Inflammation Infection Trauma Sloughed papillae Tumor Blood clots	Benign prostatic hyperplasia Cancer of prostate Cancer of bladder Calculi Diabetic neuropathy Spinal cord disease Anticholinergic drugs and α-adrenergic antagonists	Stricture Tumor Calculi Trauma Phimosis
Acquired Extrinsic Defects		
Pregnant uterus Retroperitoneal fibrosis Aortic aneurysm Uterine leiomyomata Carcinoma of uterus, prostate, bladder, colon, rectum Lymphoma Pelvic inflammatory disease, endometriosis Accidental surgical ligation	Carcinoma of cervix, colon Trauma	Trauma

Functional impairment of urine flow occurs when voiding is altered by abnormal pontine or sacral centers of micturition control. It may be asymptomatic or associated with lower urinary tract symptoms such as frequency, urgency, and postmicturition incontinence, nocturia, straining to void, slow stream, hesitancy, or a feeling of incomplete emptying. A history should be sought for trauma, back injury, surgery, diabetes, neurologic or psychiatric conditions, and medications. Causes include neurogenic bladder, often with adynamic ureter, and vesicoureteral reflux. Reflux in children may result in severe unilateral or bilateral hydroureter and hydronephrosis. Overflow urinary incontinence

combined with fecal incontinence may require an urgent evaluation for cauda equina syndrome. Urinary retention may be the consequence of α-adrenergic and anticholinergic agents, as well as opiates. Hydronephrosis in pregnancy is due to relaxational effects of progesterone on smooth muscle of the renal pelvis, as well as ureteral compression by the enlarged uterus, more often on the right side.

Diagnostic tools to identify anatomic obstruction include urinary flow measurements and a postvoid residual. Bladder volume may be readily assessed by bedside ultrasound. Cystourethroscopy and urodynamic studies may be reserved for the symptomatic patient to assess the filling phase (cystometry), pressure-volume relationship of the bladder, bladder compliance, and capacity. Pressure-flow analysis evaluates bladder contractility and bladder outlet resistance during voiding. Bladder obstruction is characterized by high pressures in women, whereas in men, a diagnosis of bladder outlet obstruction is based on flow rate and voiding pressures. A voiding cystourethrogram may be useful in evaluating incomplete emptying and bladder neck and urethral pathology.

CLINICAL FEATURES AND PATHOPHYSIOLOGY

The pathophysiology and clinical features of UTO are summarized in **Table 313-2**. Flank *pain*, the symptom that most commonly leads to medical attention, is due to distention of the collecting system or renal capsule. Pain severity is influenced more by the rate at which distention develops than by the degree of distention. Acute supravesical obstruction, as from a stone lodged in a ureter **(Chap. 312)**, is associated with excruciating, sometimes intermittent, pain, known as *renal colic*. This pain often radiates to the lower abdomen, testes, or labia. By contrast, more insidious causes of obstruction, such as chronic narrowing of the UPJ, may produce little or no pain and yet result in total destruction of the affected kidney. Flank pain that occurs only with micturition is pathognomonic of vesicoureteral reflux.

TABLE 313-2

Pathophysiology of Bilateral Ureteral Obstruction

Hemodynamic Effects	Tubule Effects	Clinical Features
Acute		
↑ Renal blood flow ↓ GFR ↓ Medullary blood flow ↑ Vasodilator prostaglandins, nitric oxide	↑ Ureteral and tubule pressures ↑ Reabsorption of Na^+, urea, water	Pain (capsule distention) Azotemia Oliguria or anuria
Chronic		
↓ Renal blood flow ↓↓ GFR ↑ Vasoconstrictor prostaglandins ↑ Renin-angiotensin production	↓ Medullary osmolarity ↓ Concentrating ability Structural damage; parenchymal atrophy ↓ Transport functions for Na^+, K^+, H^+	Azotemia Hypertension AVP-insensitive polyuria Natriuresis Hyperkalemic, hyperchloremic acidosis
Release of Obstruction		
Slow ↑ in GFR (variable)	↓ Tubule pressure ↑ Solute load per nephron (urea, NaCl) Natriuretic factors present	Postobstructive diuresis Potential for volume depletion and electrolyte imbalance due to losses of Na^+, K^+, PO_4^{2-}, Mg^{2+}, and water

Abbreviations: AVP, arginine vasopressin; GFR, glomerular filtration rate.

Obstruction of urine flow results in an increase in hydrostatic pressures proximal to the site of obstruction. It is this buildup of pressure that leads to the accompanying pain, the distention of the collecting system in the kidney, and elevated intratubular pressures that initiate tubular dysfunction. In the first days of obstruction, the dilatation of the poorly compliant collecting system may be minimal. As the increased hydrostatic pressure is expressed in the urinary space of the glomeruli, further filtration decreases or stops completely.

Azotemia develops when overall excretory function is impaired, often in the setting of bladder outlet obstruction, bilateral renal pelvic or ureteric obstruction, or unilateral disease in a patient with a solitary functioning kidney. Complete bilateral obstruction should be suspected when acute renal failure is accompanied by anuria. Any patient with renal failure otherwise unexplained, or with a history of nephrolithiasis, hematuria, diabetes mellitus, prostatic enlargement, pelvic surgery, trauma, or tumor should be evaluated for UTO.

In the acute setting, partial, bilateral obstruction may mimic prerenal azotemia with a high blood urea nitrogen-to-creatinine ratio, concentrated urine and sodium retention. Renal vascular resistance may be increased. However, with more prolonged obstruction, symptoms of *polyuria* and *nocturia* commonly accompany partial UTO and result from loss of medullary hypertonicity with diminished renal concentrating ability. Failure to produce urine free of salt (natriuresis) is due to downregulation of salt reabsorption in the proximal tubule and of transport proteins including the Na^+, K^+ adenosine triphosphatase (ATPase), NaK_2Cl cotransporter (NaK_2Cl) in the thick ascending limb, and the epithelial Na^+ channel (ENaC) in collecting duct cells. In addition to direct effects on renal transport mechanisms, increased prostaglandin E_2 (PGE_2) (due to induction of cyclooxygenase-2 [COX-2]), angiotensin II (with its downregulation of Na^+ transporters), and atrial or B-type natriuretic peptides (ANP or BNP) (due to volume expansion in the azotemic patient) contribute to decreased salt reabsorption along the nephron.

Dysregulation of aquaporin-2 water channels in the collecting duct contributes to the polyuria. The defect usually does not improve with administration of vasopressin and is therefore a form of acquired nephrogenic diabetes insipidus.

Wide fluctuations in urine output in a patient with azotemia should always raise the possibility of intermittent or partial UTO. If fluid intake is inadequate, severe dehydration and hypernatremia may develop. However, as with other causes of poor renal function, excesses of salt and water intake may result in edema and hyponatremia.

Partial bilateral UTO often results in *acquired distal renal tubular acidosis, hyperkalemia*, and *renal salt wasting*. The H^+-ATPase, situated on the apical membrane of the intercalated cells of the collecting duct, is critical for distal H^+ secretion. The trafficking of intracellular H^+ pumps from the cytoplasm to the cell membrane is disrupted in UTO. The decreased function of the ENaC, in the apical membrane of neighboring collecting duct principal cells, contributes to decreased Na^+ reabsorption (salt-wasting), and, therefore, decreased K^+ secretion via K^+ channels. Ammonium (NH_4^+) excretion important to the elimination of H^+ is impaired. These defects in tubule function are often accompanied by renal tubulointerstitial damage. Azotemia with hyperkalemia and metabolic acidosis should prompt consideration of UTO.

The renal interstitium becomes edematous and infiltrated with mononuclear inflammatory cells early in UTO. Later, interstitial fibrosis and atrophy of the papillae and medulla occur and precede these processes in the cortex. The increase in angiotensin II noted in UTO contributes to the inflammatory response and fibroblast accumulation through mechanisms involving profibrotic cytokines. With time, this process leads to chronic kidney damage.

UTO must always be considered in patients with urinary tract infections or urolithiasis. Urinary stasis encourages the growth of organisms. Urea-splitting bacteria are associated with magnesium ammonium phosphate (struvite) calculi that may take on a staghorn appearance. *Hypertension* is frequent in acute and subacute unilateral obstruction and is usually a consequence of increased release of renin by the involved kidney. Chronic kidney disease from bilateral UTO, often associated with extracellular volume expansion, may result in significant hypertension. *Erythrocytosis*, an infrequent complication of obstructive uropathy, is secondary to increased erythropoietin production.

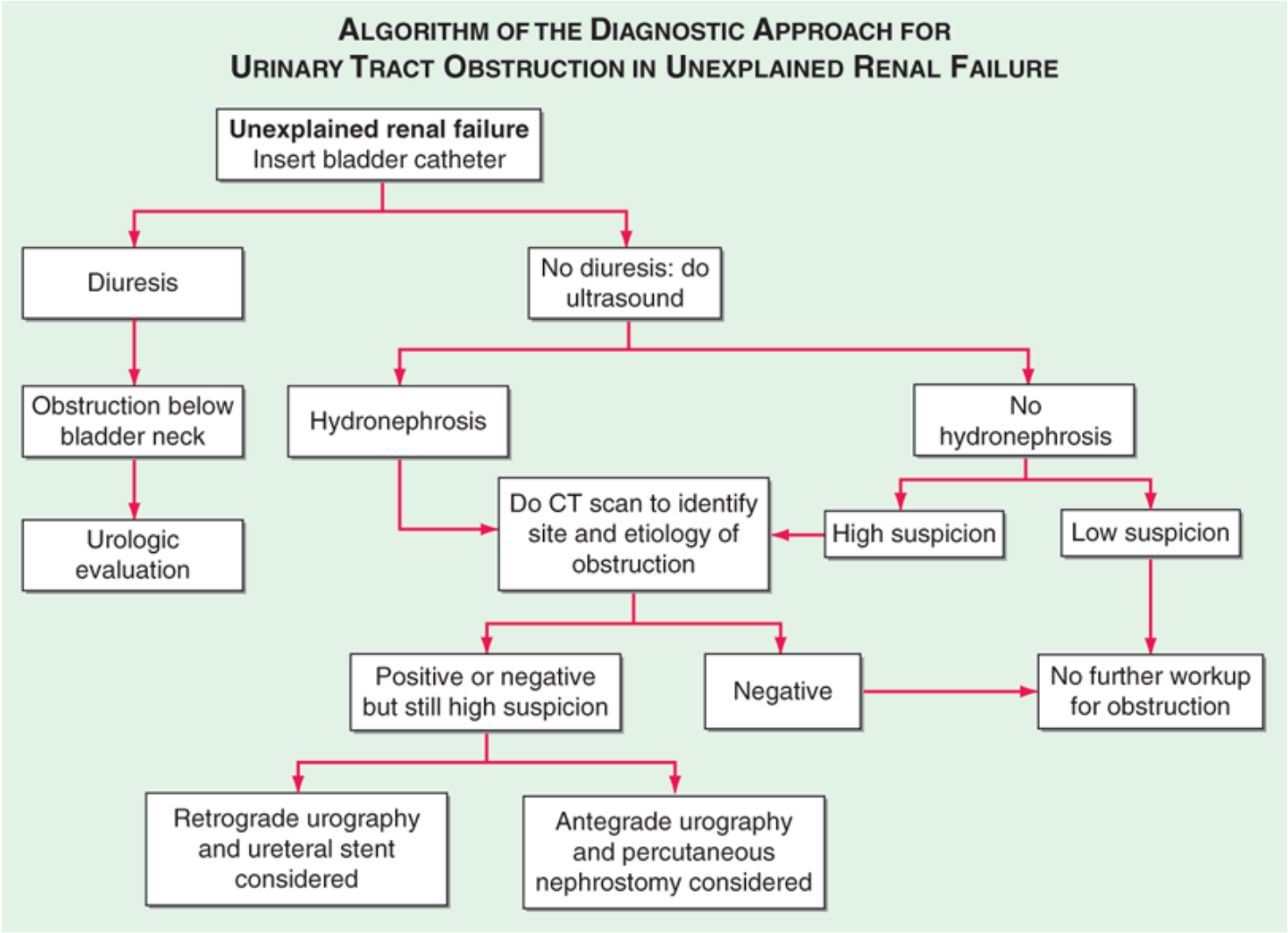
DIAGNOSIS

A history of difficulty in voiding, pain, infection, or change in urinary volume is common. Evidence for distention of the kidney or urinary bladder can often be obtained by palpation and percussion of the abdomen. A careful rectal and genital examination may reveal enlargement or nodularity of the prostate, abnormal rectal sphincter tone, or a rectal or pelvic mass.

Urinalysis may reveal hematuria, pyuria, and bacteriuria. The urine sediment is often normal, even when obstruction leads to marked azotemia and extensive structural damage. An abdominal scout film, although insensitive, may detect nephrocalcinosis or a radiopaque stone. As indicated in Fig. 313-1, if UTO is suspected, a bladder catheter should be inserted. Abdominal ultrasonography should be performed to evaluate renal and bladder size, as well as pyelocalyceal contour. Ultrasonography is ~90% specific and sensitive for detection of hydronephrosis. False-positive results are associated with diuresis, renal cysts, or the presence of an extrarenal pelvis, a normal congenital variant. Congenital UPJ obstruction may be mistaken for renal cystic disease. Hydronephrosis may be absent on ultrasound when obstruction is less than 48 h in duration or associated with volume contraction, staghorn calculi, retroperitoneal fibrosis, or infiltrative renal disease. Duplex Doppler ultrasonography may detect an increased resistive index in urinary obstruction.

FIGURE 313-1

Diagnostic approach for urinary tract obstruction in unexplained renal failure. CT, computed tomography.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Recent advances in technology have led to alternatives and have largely replaced the once standard intravenous urogram in the further evaluation of UTO. The high-resolution multidetector row computed tomography (CT) scan in particular has advantages of visualizing the retroperitoneum, as well as identifying both intrinsic and extrinsic sites of obstruction. Noncontrast CT scans improve visualization of the urinary tract in the patient with renal impairment and are safer for patients at risk for contrast nephropathy. Magnetic resonance urography is a promising technique but, at this time, not superior to the CT scan and carries the risk of certain gadolinium agents in patients with renal insufficiency. CT

scanning may define the site of obstruction, identify and characterize kidney stones, and demonstrate dilatation of the calyces, renal pelvis, and ureter above the obstruction. The ureter may be tortuous in chronic obstruction. Radionuclide scans are able to give differential renal function but give less anatomic detail than CT scans. Furosemide is sometimes given to increase detection with imaging, and to distinguish functional from anatomic obstruction. The increase in urinary flow may bring out the pain of an acute obstructive process.

To facilitate visualization of a suspected lesion in a ureter or renal pelvis, *retrograde* or *antegrade urography* should be attempted. These procedures do not carry risk of contrast-induced acute kidney injury in patients with renal insufficiency. The retrograde approach involves catheterization of the involved ureter under cystoscopic control, whereas the antegrade technique necessitates percutaneous placement of a catheter into the renal pelvis. Although the antegrade approach may provide immediate decompression of a unilateral obstructing lesion, many urologists initially attempt the retrograde approach unless the catheterization is unsuccessful.

Voiding cystourethrography is of value in the diagnosis of vesicoureteral reflux and bladder neck and urethral obstructions. Postvoiding films reveal residual urine. Endoscopic visualization by the urologist often permits precise identification of lesions involving the urethra, prostate, bladder, and ureteral orifices.

TREATMENT

TREATMENT

Urinary Tract Obstruction

UTO complicated by infection requires immediate relief of obstruction to prevent development of generalized sepsis and progressive renal damage. Sepsis necessitates prompt urologic intervention. Drainage may be achieved by nephrostomy, ureterostomy, or ureteral, urethral, or suprapubic catheterization. Prolonged antibiotic treatment may be necessary. Chronic or recurrent infections in a poorly functioning obstructed kidney may necessitate nephrectomy. When infection is not present, surgery is often delayed until acid-base, fluid, and electrolyte status is restored. Nevertheless, the site of obstruction should be ascertained as soon as feasible. Elective relief of obstruction is usually recommended in patients with urinary retention, recurrent urinary tract infections, persistent pain, or progressive loss of renal function. Benign prostatic hypertrophy may be treated medically with α-adrenergic blockers and 5α-reductase inhibitors. Renal colic may be treated with anti-inflammatory medication as edema often contributes to an obstructing ureteral stone, and α-adrenergic blockers may also be of benefit. Use of opiates in patients with decreased renal function may be dangerous and should be used with caution. Functional obstruction secondary to neurogenic bladder may be decreased with the combination of frequent voiding and cholinergic drugs.

PROGNOSIS

With relief of obstruction, the prognosis regarding return of renal function depends largely on whether irreversible renal damage has occurred. When obstruction is not relieved, the course will depend mainly on whether the obstruction is complete or incomplete and bilateral or unilateral, as well as whether or not urinary tract infection is also present. Complete obstruction with infection can lead to total destruction of the kidney within days. Partial return of glomerular filtration rate may follow relief of complete obstruction of 1 and 2 weeks' duration, but after 8 weeks of obstruction, recovery is unlikely. In the absence of definitive evidence of irreversibility, every effort should be made to decompress the obstruction in the hope of restoring renal function at least partially. A renal radionuclide scan, performed after a prolonged period of decompression, may be used to predict the reversibility of renal dysfunction.

POSTOBSTRUCTIVE DIURESIS

Relief of bilateral, but not unilateral, complete obstruction commonly results in polyuria, which may be massive. The urine is usually hypotonic and may contain large amounts of sodium chloride, potassium, phosphate, and magnesium. The natriuresis is due in part to the correction of extracellular volume expansion, the increase in natriuretic factors accumulated during the period of renal failure, and depressed salt and water reabsorption when urine flow is reestablished. The retained urea is excreted with improved GFR, resulting in an osmotic diuresis which increases the urine volume of electrolyte-free water. The urinary concentrations of sodium and potassium that when added are less than the serum sodium is evidence of electrolyte free-water excretion. In the majority of patients, this diuresis results in the *appropriate* excretion of the excesses of retained salt and water. When extracellular volume and composition return to normal, the diuresis usually abates spontaneously. Occasionally, iatrogenic expansion of extracellular volume is responsible for, or sustains, the diuresis observed in the postobstructive period. Replacement with intravenous fluids in amounts less than urinary losses usually prevents this complication. More aggressive fluid management is required in the setting of hypovolemia, hypotension, or disturbances in serum electrolyte concentrations.

The loss of electrolyte-free water with urea may result in hypernatremia. Measured urinary output and serum and urine sodium, potassium and osmolal concentrations should guide the use of appropriate intravenous replacement. Often replacement with 0.45% saline is required. Relief of obstruction may be followed by urinary salt and water losses severe enough to provoke profound dehydration and vascular collapse. In these patients, decreased tubule reabsorptive capacity is probably responsible for the marked diuresis. Appropriate therapy in such patients includes intravenous administration of salt-containing solutions to replace sodium and volume deficits.

FURTHER READING

Frokiaer J, Zeidel ML: Urinary tract obstruction, in *Brenner and Rector's The Kidney*, 10th ed, BM Brenner (ed). Philadelphia, Saunders, 2015, pp 1257–1282.

Meldrum KK: Pathophysiology of urinary tract obstruction, in *Campbell-Walsh Urology*, 11th ed, AJ Wein et al (eds). Philadelphia, Elsevier; 2016, pp 1089–1103.

Smith-Bindman R et al: Ultrasonography versus computed tomography for suspected nephrolithiasis. N Engl J Med 371:1100, 2014. [PubMed: 25229916]

Stoller ML: Urinary obstruction & stasis, in *Smith and Tanagho's General Urology*, 18th ed, JW McAninch, TF Lue (eds). New York, McGraw-Hill, 2013, pp. 170–182.

Tanagho EA, Nguyen HT: Vesicoureteral reflux, in Smith and Tanagho's *General Urology*, 18th ed, WJ McAninch, TF Lue (eds). New York, McGraw-Hill, 2013, pp. 182–197.

Chapter 314: Approach to the Patient with Gastrointestinal Disease

William L. Hasler; Chung Owyang

ANATOMIC CONSIDERATIONS

The gastrointestinal (GI) tract extends from the mouth to the anus and is composed of several organs with distinct functions. Specialized sphincters that assist in gut compartmentalization separate the organs. The gut wall is organized into distinct layers that contribute to regional activities. The mucosa is a barrier to luminal contents or a site for fluid and nutrient transfer. Gut smooth muscle in association with the enteric nervous system mediates propulsion from one region to the next. Many GI organs possess a serosal layer that provides a supportive foundation and permits external input.

Interactions with other systems serve the needs of the gut and the body. Pancreaticobiliary conduits deliver bile and enzymes into the duodenum. The vascular supply is modulated by GI activity. Lymphatic channels assist in gut immune activities. Intrinsic nerves provide the controls for propulsion and fluid regulation. Extrinsic neural input provides volitional or involuntary control that is specific for each gut region.

FUNCTIONS OF THE GI TRACT

The GI tract serves two main functions—assimilating nutrients and eliminating waste. In the mouth, food is processed, mixed with salivary amylase, and delivered to the gut lumen. The esophagus propels the bolus into the stomach; the lower esophageal sphincter prevents oral reflux of gastric contents. The squamous esophageal mucosa protects against significant diffusion or absorption. Aboral esophageal contractions coordinate with relaxation of the upper and lower esophageal sphincters on swallowing.

The stomach triturates and mixes the food bolus with pepsin and acid. Gastric acid also sterilizes the upper gut. The proximal stomach serves a storage function by relaxing to accommodate the meal. Phasic contractions in the distal stomach propel food residue against the pylorus, where it is ground and thrust proximally for further mixing before it is emptied into the duodenum. The stomach secretes intrinsic factor for vitamin B_{12} absorption.

Most nutrient absorption occurs in the small intestine. The intestinal mucosal villus architecture provides maximal surface area for absorption and is endowed with specialized enzymes and transporters. Triturated food from the stomach mixes with pancreatic juice and bile in the duodenum. Pancreatic juice contains

enzymes for carbohydrate, protein, and fat digestion as well as bicarbonate to optimize the pH for enzyme activation. Bile secreted by the liver and stored in the gallbladder is essential for lipid digestion. The proximal intestine is optimized for rapid absorption of most nutrients and minerals, whereas the ileum is better suited for absorbing vitamin B_{12} and bile acids. Bile contains by-products of erythrocyte degradation, toxins, medications, and cholesterol for fecal evacuation. Small intestinal motor function delivers indigestible residue into the colon for processing. The ileocecal junction is a sphincteric structure that prevents coloileal reflux, maintaining small-intestinal sterility.

The colon prepares waste for evacuation. The colonic mucosa dehydrates the stool, decreasing daily volumes of 1000–1500 mL in the ileum to 100–200 mL expelled from the rectum. The colon possesses a dense bacterial colonization that ferments undigested carbohydrates and short-chain fatty acids. Additional roles for the gut microbiome include modulation of immune and physiologic activity. Transit in the esophagus takes seconds and times in the stomach and small intestine range from minutes to a few hours, but colon propagation requires more than 1 day in most individuals. Colon contractions exhibit a to-and-fro character that promotes fecal desiccation. The proximal colon mixes and absorbs fluid, while the distal colon exhibits peristaltic contractions and mass movements to expel the stool. The colon terminates in the anus, which possesses volitional and involuntary controls to permit fecal retention until it can be released in a convenient setting.

EXTRINSIC MODULATION OF GUT FUNCTION

GI function is modified by influences outside the gut. Unlike other organs, the gut is in continuity with the outside environment. Protective mechanisms are vigilant against damage from foods, medications, toxins, and infectious organisms. Mucosal immune mechanisms include epithelial and lamina propria lymphocyte and plasma cell populations supported by lymph node chains to prevent noxious agents from entering the circulation. Antimicrobial peptides secreted by intestinal Paneth cells also defend against luminal pathogens. All drugs and toxins absorbed into the bloodstream are filtered and detoxified in the liver via the portal venous circulation. Although intrinsic nerves control most basic gut activities, extrinsic neural input modulates many functions. Many GI reflexes involve extrinsic vagus or splanchnic nerve pathways. The brain-gut axis alters function in regions not under volitional regulation. As an example, stress has potent effects on gut motor, secretory, and sensory functions.

OVERVIEW OF GI DISEASES

GI diseases develop as a result of abnormalities within or outside of the gut and range in severity from those that produce mild symptoms and no long-term morbidity to those with intractable symptoms or adverse outcomes. Diseases may be localized to one organ or exhibit diffuse involvement at many sites.

CLASSIFICATION OF GI DISEASES

GI diseases are manifestations of alterations in nutrient assimilation or waste evacuation or in the activities supporting these main functions.

Impaired Digestion and Absorption

Diseases of the stomach, intestine, biliary tree, and pancreas can disrupt digestion and absorption. The most common intestinal maldigestion syndrome, lactase deficiency, produces gas and diarrhea after ingestion of dairy products and has no adverse outcomes. Other intestinal enzyme deficiencies produce similar symptoms after ingestion of other simple sugars. Conversely, celiac disease, bacterial overgrowth, infectious enteritis, Crohn's ileitis, and radiation damage, which affect digestion and/or absorption more diffusely, produce anemia, dehydration, electrolyte disorders, or malnutrition. Gastric hypersecretory conditions such as Zollinger-Ellison syndrome damage the intestinal mucosa, impair pancreatic enzyme activation, and accelerate transit due to excess gastric acid. Biliary obstruction from stricture or neoplasm impairs fat digestion. Impaired pancreatic enzyme release in chronic pancreatitis or pancreatic cancer decreases intraluminal digestion and can lead to malnutrition.

Altered Secretion

Selected GI diseases result from dysregulation of gut secretion. Gastric acid hypersecretion occurs in Zollinger-Ellison syndrome, G cell hyperplasia, retained antrum syndrome, and some individuals with duodenal ulcers. Conversely, patients with atrophic gastritis or pernicious anemia release little or no gastric acid. Inflammatory and infectious small-intestinal and colonic diseases produce fluid loss through impaired absorption or enhanced secretion. Common intestinal and colonic hypersecretory conditions cause diarrhea and include acute bacterial or viral infection, chronic *Giardia* or cryptosporidia infections, small-intestinal bacterial overgrowth, bile salt diarrhea, microscopic colitis, diabetic diarrhea, and abuse of certain laxatives. Less common causes include large colonic villus adenomas and endocrine neoplasias with tumor overproduction of secretagogue transmitters like vasoactive intestinal polypeptide.

Altered Gut Transit

Impaired gut transit may be secondary to mechanical obstruction. Esophageal occlusion most often results from stricture (due to acid exposure or eosinophilic esophagitis) or neoplasm. Gastric outlet obstruction develops from peptic ulcer disease or gastric cancer. Small-intestinal obstruction most commonly results from adhesions but may also occur with Crohn's disease, radiation- or drug-induced strictures, and less likely malignancy. The most common cause of colonic obstruction is colon cancer, although inflammatory strictures develop in patients with inflammatory bowel disease (IBD), after certain infections such as diverticulitis, or with some drugs.

Retardation of propulsion also develops from disordered motor function. Achalasia is characterized by impaired esophageal body peristalsis and incomplete lower esophageal sphincter relaxation. Gastroparesis is the symptomatic delay in gastric emptying of meals due to impaired gastric motility. Intestinal pseudoobstruction causes marked delays in small-bowel transit due to enteric nerve or intestinal smooth-muscle injury. Slow-transit constipation is produced by diffusely impaired colonic propulsion. Constipation also is produced by outlet abnormalities such as rectal prolapse, intussusception, or dyssynergia—a failure of anal or puborectalis relaxation upon attempted defecation.

Disorders of rapid propulsion are less common than those with delayed transit. Rapid gastric emptying occurs in postvagotomy dumping syndrome, with gastric hypersecretion, and in some cases of functional dyspepsia and cyclic vomiting syndrome. Exaggerated intestinal or colonic motor patterns may be responsible for diarrhea in irritable bowel syndrome (IBS). Accelerated transit with hyperdefecation is noted in hyperthyroidism.

Immune Dysregulation

Many inflammatory GI conditions are consequences of altered gut immune function. The mucosal inflammation of celiac disease results from dietary ingestion of gluten-containing grains. Some patients with food allergy also exhibit altered immune populations. Eosinophilic esophagitis and eosinophilic gastroenteritis are inflammatory disorders with prominent mucosal eosinophils. Ulcerative colitis and Crohn's disease are disorders of uncertain etiology that produce mucosal injury primarily in the lower gut. The microscopic colitides, lymphocytic and collagenous colitis, exhibit colonic subepithelial infiltrates without visible mucosal damage. Bacterial, viral, and protozoal organisms may produce ileitis or colitis in selected patient populations. Furthermore, alterations in the gut microbiome (termed dysbiosis) are postulated to trigger flares of IBD, celiac disease, and IBS.

Impaired Gut Blood Flow

Different GI regions are at variable risk for ischemic damage from impaired blood flow. Rare cases of gastroparesis result from blockage of the celiac and superior mesenteric arteries. More commonly encountered are intestinal and colonic ischemia that are consequences of arterial embolus, arterial thrombosis, venous thrombosis, or hypoperfusion from dehydration, sepsis, hemorrhage, or reduced cardiac output. These may produce mucosal injury, hemorrhage, or even perforation. Chronic ischemia may result in intestinal stricture. Some cases of radiation enterocolitis exhibit reduced mucosal blood flow.

Neoplastic Degeneration

All GI regions are susceptible to malignant degeneration to varying degrees. In the United States, colorectal cancer is most common and usually presents after age 50 years. Worldwide, gastric cancer is prevalent especially in certain Asian regions. Esophageal cancer develops with chronic acid reflux or after an extensive alcohol or tobacco use history. Small-intestinal neoplasms are rare and occur with underlying inflammatory disease. Anal cancers arise after prior anal infection or inflammation. Pancreatic and biliary cancers elicit severe pain, weight loss, and jaundice and have poor prognoses. Hepatocellular carcinoma usually arises in the setting of chronic viral hepatitis or cirrhosis secondary to other causes. Most GI cancers exhibit carcinomatous histology; however, lymphomas and other cell types also are observed.

Disorders without Obvious Organic Abnormalities

The most common GI disorders show no abnormalities on biochemical or structural testing and include IBS, functional dyspepsia, and functional heartburn. These disorders exhibit altered gut motor function; however, the pathogenic relevance of these abnormalities is uncertain. Exaggerated visceral sensory responses to

noxious stimulation may cause discomfort in these disorders. Symptoms in other patients result from altered processing of visceral pain sensations in the central nervous system. Functional bowel patients with severe symptoms may exhibit significant emotional disturbances on psychometric testing. Subtle immunologic defects may contribute to functional symptoms as well.

Genetic Influences

Although many GI diseases result from environmental factors, others exhibit hereditary components. Family members of IBD patients show a genetic predisposition to disease development themselves. Colonic, esophageal, and pancreatic malignancies arise in certain inherited disorders. Rare genetic dysmotility syndromes are described. Familial clustering is observed in the functional bowel disorders, although this may be secondary learned familial illness behavior rather than a true hereditary factor.

SYMPTOMS OF GI DISEASE

Symptoms of GI disease include abdominal pain, heartburn, nausea and vomiting, altered bowel habits, GI bleeding, jaundice, and other manifestations (Table 314-1).

TABLE 314-1

Common Causes of Common Gastrointestinal (GI) Symptoms

Abdominal Pain	Nausea and Vomiting	Diarrhea	GI Bleeding	Obstructive Jaundice
Appendicitis	Medications	Infection	Ulcer disease	Bile duct stones
Gallstone disease	GI obstruction	Poorly absorbed sugars	Esophagitis	Cholangiocarcinoma
Pancreatitis	Motor disorders	Inflammatory bowel disease	Varices	Cholangitis
Diverticulitis	Functional bowel disorder	Microscopic colitis	Vascular lesions	Sclerosing cholangitis
Ulcer disease	Enteric infection	Functional bowel disorder	Neoplasm	Ampullary stenosis
Esophagitis	Pregnancy	Celiac disease	Diverticula	Ampullary carcinoma
GI obstruction	Endocrine disease	Pancreatic insufficiency	Hemorrhoids	Pancreatitis
Inflammatory bowel disease	Motion sickness	Hyperthyroidism	Fissures	Pancreatic tumor
Functional bowel disorder	Central nervous system disease	Ischemia	Inflammatory bowel disease	
Vascular disease		Endocrine tumor	Infectious colitis	
Gynecologic causes				
Renal stone				

Abdominal Pain

Abdominal pain results from GI disease and extra-intestinal conditions involving the genitourinary tract, abdominal wall, thorax, or spine. Visceral pain generally is midline in location and vague in character, whereas parietal pain is localized and precisely described. Painful inflammatory diseases include peptic ulcer, appendicitis, diverticulitis, IBD, pancreatitis, cholecystitis, and infectious enterocolitis. Noninflammatory visceral sources include biliary colic, mesenteric ischemia, and neoplasia. The most common causes of abdominal pain are IBS and functional dyspepsia.

Heartburn

Heartburn, a burning substernal sensation, is reported intermittently by 40% of the population. Classically, heartburn results from excess gastroesophageal acid reflux, but some cases exhibit normal esophageal acid exposure and are caused by reflux of nonacidic material or heightened sensitivity of esophageal nerves.

Nausea and Vomiting

Nausea and vomiting are caused by GI diseases, medications, toxins, infection, endocrine disorders, labyrinthine conditions, and central nervous system disease. Mechanical obstructions of the upper gut are commonly excluded as causes of chronic nausea and vomiting, but disorders of propulsion including gastroparesis and intestinal pseudoobstruction elicit similar symptoms. Nausea and vomiting also are commonly reported by patients with IBS and functional disorders of the upper gut (including chronic nausea vomiting syndrome and cyclic vomiting syndrome).

Altered Bowel Habits

Altered bowel habits are common complaints of patients with GI disease. Constipation may be reported as infrequent defecation, straining with defecation, passage of hard stools, or a sense of incomplete fecal evacuation and is caused by obstruction, colonic motor disorders, medications, and endocrine diseases like hypothyroidism and hyperparathyroidism. Diarrhea may be reported as frequent defecation, passage of loose or watery stools, fecal urgency, or a similar sense of incomplete evacuation. The differential diagnosis of diarrhea includes infections, inflammatory causes, malabsorption, and medications. IBS produces constipation, diarrhea, or an alternating bowel pattern. Fecal mucus is common in IBS, whereas pus and blood characterize IBD. Steatorrhea develops with malabsorption.

GI Bleeding

Hemorrhage may develop from any gut organ. Upper GI bleeding presents with melena or hematemesis, whereas lower GI bleeding produces passage of bright red or maroon stools. However, briskly bleeding upper sites can elicit voluminous red rectal bleeding, whereas slowly bleeding ascending colon sites may produce melena. Chronic occult GI bleeding may present with iron deficiency anemia. Causes of upper GI bleeding include ulcer disease, gastroduodenitis, esophagitis, portal hypertensive etiologies, malignancy, tears across the gastroesophageal junction, and vascular lesions. Lower GI sources of hemorrhage include hemorrhoids, anal fissures, diverticula, ischemic colitis, neoplasm, IBD, infectious colitis, drug-induced colitis, arteriovenous malformations, and other vascular lesions.

Jaundice

Jaundice results from prehepatic, intrahepatic, or post-hepatic disease. Posthepatic causes of jaundice include biliary diseases, like choledocholithiasis, acute cholangitis, primary sclerosing cholangitis, other strictures, and neoplasm, and pancreatic disorders, like acute and chronic pancreatitis, stricture, and malignancy.

Other Symptoms

Other symptoms are manifestations of GI disease. Dysphagia, odynophagia, and unexplained chest pain suggest esophageal disease. A globus sensation is reported with esophagopharyngeal conditions, but also occurs with functional GI disorders. Weight loss, anorexia, and fatigue present with neoplastic, inflammatory, motility, pancreatic, and psychiatric conditions. IBD is associated with hepatobiliary dysfunction, skin and eye lesions, and arthritis. Celiac disease may present with dermatitis herpetiformis. Jaundice can produce pruritus. Conversely, systemic diseases have GI consequences. Systemic lupus may cause gut ischemia, presenting with pain or bleeding. Severe burns may lead to gastric ulcer formation.

EVALUATION OF THE PATIENT WITH GI DISEASE

Evaluation of the patient with suspected GI disease begins with a careful history and examination. Subsequent investigation with tools to test gut structure or function and luminal constituents are indicated in selected cases. In patients with normal findings on diagnostic testing, validated symptom profiles are used to confidently diagnose a functional bowel disorder.

HISTORY

The history in suspected GI disease has several components. Symptom timing, patterns, and duration suggest specific etiologies. Short duration symptoms commonly result from acute infection or inflammation, toxin exposure, or ischemia. Long-standing symptoms point to chronic inflammation, neoplasia, or functional bowel disorders. Symptoms from mechanical obstruction, ischemia, IBD, and functional bowel disorders are worsened by meals, while ulcer symptoms may be relieved by eating or antacids. Ulcer pain occurs intermittently over weeks to months, whereas biliary colic has a sudden onset and lasts up to several hours. Acute pancreatitis pain is severe and persists for days to weeks. Meals elicit diarrhea while defecation relieves discomfort in some cases of IBD and IBS. Functional bowel disorders are exacerbated by stress. Sudden awakening from sound sleep by pain suggests organic rather than functional disease. Diarrhea from malabsorption usually improves with fasting, whereas secretory diarrhea persists without oral intake.

Symptom relation to other factors narrows the list of diagnostic possibilities. Obstructive symptoms with prior abdominal surgery raise concern for adhesions. Loose stools after gastrectomy or cholecystectomy suggest dumping syndrome or postcholecystectomy diarrhea. Symptom onset after travel prompts consideration of infection. Medications produce pain, altered bowel habits, or GI bleeding. Celiac disease is prevalent in people

of northern European descent, whereas IBD is more common in Jewish populations. A sexual history may raise concern for infection or immunodeficiency.

For nearly 40 years, working groups have devised symptom criteria to improve diagnosis of functional bowel disorders and to minimize the numbers of unnecessary diagnostic tests performed. The best accepted symptom-based criteria are the Rome criteria. However, when tested against findings of structural investigations in IBS and functional dyspepsia, the Rome criteria exhibit sensitivities and specificities of only 55–75% indicating the need for careful test selection in patients at high risk of organic disease.

PHYSICAL EXAMINATION

The physical examination complements information from the history. Abnormal vital signs provide diagnostic clues and determine the need for acute intervention. Fever suggests inflammation or neoplasm. Orthostasis is produced by significant blood loss, dehydration, sepsis, or autonomic neuropathy. Skin, eye, or joint findings may point to specific diagnoses. Neck examination with swallowing assessment evaluates dysphagia. Lung and cardiac examinations evaluate for cardiopulmonary disease as causes of abdominal pain or nausea. Pelvic examination tests for a gynecologic source of abdominal pain. Rectal examination may detect blood, indicating mucosal injury or neoplasm or a palpable inflammatory mass in appendicitis. Metabolic conditions and gut motor disorders have associated peripheral neuropathy.

Abdominal inspection may reveal distention from obstruction, tumor, or ascites or vascular abnormalities with liver disease. Ecchymoses develop with severe pancreatitis. Auscultation detects bruits or friction rubs from vascular disease or hepatic tumors. Loss of bowel sounds signifies ileus, whereas high-pitched, hyperactive sounds characterize intestinal obstruction. Percussion assesses liver size and detects shifting dullness from ascites. Palpation assesses for hepatosplenomegaly and neoplastic or inflammatory masses. Intestinal ischemia elicits severe pain but little tenderness. Patients with visceral pain may exhibit generalized discomfort, whereas those with parietal pain or peritonitis have localized pain with involuntary guarding, rigidity, or rebound. Patients with musculoskeletal abdominal wall pain may note tenderness exacerbated by Valsalva or leg lift maneuvers.

TOOLS FOR PATIENT EVALUATION

Laboratory, radiographic, and functional tests assist in diagnosis of suspected GI disease. The GI tract also is amenable to internal evaluation using endoscopy and to examination of luminal contents. Histopathologic examinations of GI tissues complement these tests.

Laboratory

Laboratory tests facilitate diagnosis of GI disease. Iron-deficiency anemia suggests mucosal blood loss, whereas vitamin B_{12} deficiency results from intestinal, gastric, or pancreatic disease. Either can result from inadequate oral intake. Leukocytosis and increased sedimentation rates and C-reactive proteins are found in inflammation, whereas leukopenia is seen in viremic illness. Severe vomiting or diarrhea elicits electrolyte

disturbances, acid-base abnormalities, and elevated blood urea nitrogen. Pancreaticobiliary or liver disease produce elevated pancreatic or liver chemistries. Thyroid chemistries, cortisol, and calcium levels evaluate for endocrinologic causes of symptoms. Pregnancy testing is considered for women with unexplained nausea. Serologic tests screen for celiac disease, IBD, connective tissue diseases, and paraneoplastic dysmotility syndromes. Hormone levels are obtained for suspected endocrine neoplasia. Intraabdominal malignancies produce tumor markers including the carcinoembryonic antigen CA 19-9 and α-fetoprotein. Blood testing also monitors medication therapy, as with thiopurine metabolite levels in IBD. Pharmacogenetic methods are being adopted to determine optimal patient populations for GI medication use. In areas including IBD, research into novel biomarkers is being conducted to predict longitudinal course and treatment response. Other body fluids are sampled under certain circumstances. Ascitic fluid is analyzed for infection, malignancy, or findings of portal hypertension. Urine samples screen for carcinoid, porphyria, and heavy metal intoxication.

Luminal Contents

Luminal contents can provide diagnostic clues. Stool samples are cultured for bacterial pathogens, examined for leukocytes and parasites, or tested for *Giardia* antigen. Duodenal aspirates can be examined for parasites or cultured for bacterial overgrowth. Fecal fat is quantified in possible malabsorption. Elevations in fecal calprotectin or lactoferrin are found in inflammatory conditions like IBD. Stool electrolytes can be measured in diarrheal conditions. Laxative screens are performed for suspected laxative abuse. Fecal immunochemical and DNA tests are assuming emerging roles in colon cancer screening in low risk populations. Gastric acid is quantified to exclude gastrinoma. Esophageal pH testing is done for refractory symptoms of acid reflux, whereas impedance techniques quantify nonacidic reflux. Pancreatic juice is analyzed for enzyme or bicarbonate content to exclude exocrine insufficiency.

Endoscopy

The gut is accessible with endoscopy, which can diagnose causes of bleeding, pain, nausea and vomiting, weight loss, altered bowel function, and fever. **Table 314-2** lists common indications for endoscopic procedures. Upper endoscopy evaluates the esophagus, stomach, and duodenum, whereas colonoscopy assesses the colon and distal ileum. Upper endoscopy is advocated as the initial test performed for suspected ulcer disease, esophagitis, neoplasm, malabsorption, and Barrett's metaplasia because of its abilities to visualize and biopsy any abnormality. Colonoscopy is the preferred procedure for colon cancer screening and surveillance and to biopsy colitis secondary to IBD, infection, ischemia, and radiation. Sigmoidoscopy examines the colon to the splenic flexure and excludes distal inflammation or obstruction in young patients not at significant risk for colon cancer. For elusive GI bleeding from arteriovenous malformations or superficial ulcers, small-intestinal examination is performed with push enteroscopy, capsule endoscopy, or double-balloon enteroscopy. Capsule endoscopy also visualizes small-intestinal Crohn's disease in individuals with negative radiography. Endoscopic retrograde cholangiopancreaticography (ERCP) provides diagnoses of pancreatic and biliary disease. Endoscopic ultrasound (EUS) diagnoses and stages GI malignancy, excludes choledocholithiasis, evaluates pancreatitis, and assesses anal continuity.

TABLE 314-2

Common Indications for Endoscopy

Upper Endoscopy	Colonoscopy	Endoscopic Retrograde Cholangiopancreatography	Endoscopic Ultrasound	Capsule Endoscopy	Double-Balloon Endoscopy
Dyspepsia despite treatment Dyspepsia with signs of organic disease Refractory vomiting Dysphagia Upper GI bleeding Anemia Weight loss Malabsorption Biopsy radiologic abnormality Polypectomy Place gastrostomy Barrett's surveillance Palliate neoplasm Sample duodenal tissue/fluid Remove foreign body Endoscopic mucosal resection or	Cancer screening Lower GI bleeding Anemia Diarrhea Polypectomy Obstruction Biopsy radiologic abnormality Cancer surveillance: family history prior polyp/cancer, colitis Palliate neoplasm Remove foreign body Place stent across stenosis	Jaundice Postbiliary surgery complaints Cholangitis Gallstone pancreatitis Pancreatic/biliary/ampullary tumor Unexplained pancreatitis Pancreatitis with unrelenting pain Fistulas Biopsy radiologic abnormality Pancreaticobiliary drainage Sample bile Sphincter of Oddi manometry	Staging of malignancy Characterize and biopsy submucosal mass Bile duct stones Chronic pancreatitis Drain pseudocyst Anal continuity	Obscure gastrointestinal (GI) bleeding Suspected Crohn's disease of the small intestine	Ablation of small-intestinal bleeding sources Biopsy of suspicious small-intestinal masses/ulcers

Upper Endoscopy	Colonoscopy	Endoscopic Retrograde Cholangiopancreatography	Endoscopic Ultrasound	Capsule Endoscopy	Double-Balloon Endoscopy
ablation of dysplastic Barrett's mucosa Place stent across stenosis					

A newer area involves development of novel imaging protocols which permit optical biopsies to define mucosal histology and detect dysplasia in selected settings. Methods employed include narrow band imaging and chromoendoscopy in colitis and confocal laser endomicroscopy and optical coherence tomography in Barrett's esophagus and gastric cancer surveillance.

Radiography/Nuclear Medicine

Radiographic tests evaluate gut diseases and extraluminal structures. Contrast radiography with barium provides mucosal definition and can assess gut transit and pelvic floor dysfunction. Barium swallow is the initial procedure to exclude subtle rings, strictures, or achalasia as causes of dysphagia, whereas small-bowel contrast radiology detects intestinal tumors and Crohn's ileitis. Contrast enemas are performed when colonoscopy is unsuccessful or contraindicated. Ultrasound and computed tomography (CT) evaluate regions not accessible by endoscopy or contrast studies, including the liver, pancreas, gallbladder, kidneys, and retroperitoneum and are useful for diagnosing mass lesions, fluid collections, organ enlargement, and, in the case of ultrasound, gallstones. CT and magnetic resonance (MR) colonography have been considered as alternatives to colonoscopy for colon cancer screening. MR methods image the pancreaticobiliary ducts to exclude neoplasm, stones, and sclerosing cholangitis, and the liver to characterize benign and malignant tumors. Specialized CT or MR enterography quantifies IBD intensity. Angiography excludes mesenteric ischemia and determines spread of malignancy. Angiographic techniques also access the biliary tree in obstructive jaundice. CT and MR techniques screen for mesenteric occlusion, thereby limiting exposure to angiographic dyes. Positron emission tomography can distinguish malignant from benign disease in several organ systems.

Scintigraphy evaluates structural abnormalities and quantifies luminal transit. Radionuclide scans localize bleeding sites in patients with brisk hemorrhage to direct therapy with endoscopy, angiography, or surgery. Radiolabeled leukocyte scans search for intraabdominal abscesses not visualized on CT. Biliary scintigraphy complements ultrasound in assessing for cholecystitis. Scintigraphy to quantify esophageal and gastric emptying is well established, whereas techniques to measure small-intestinal or colonic transit are less widely used.

Histopathology

Endoscopic mucosal biopsies evaluate for inflammatory, infectious, and neoplastic disease. Deep rectal biopsies facilitate diagnosis of Hirschsprung's disease or amyloid. Liver biopsy is performed for abnormal liver chemistries, in unexplained jaundice, following liver transplant to exclude rejection, and to characterize inflammation in chronic viral hepatitis prior to initiating antiviral therapy. Biopsies obtained during CT or ultrasound evaluate for intraabdominal conditions not accessible by endoscopy.

Functional Testing

Tests of gut function provide important data when structural testing is nondiagnostic. Functional testing of motor activity is provided by newer high resolution manometric techniques. Esophageal manometry is useful for suspected achalasia, whereas small-intestinal manometry tests for pseudoobstruction and colon manometry evaluates for colonic inertia. A wireless motility capsule measures transit and contractile activity in the stomach, small intestine, and colon in a single test. Anorectal manometry with balloon expulsion testing is used for unexplained incontinence or constipation from outlet dysfunction. Biliary manometry tests for sphincter of Oddi dysfunction with unexplained biliary pain. A novel endoluminal functional lumen imaging probe is available to measure heightened distensibility in the lower esophageal sphincter in achalasia and pylorus in gastroparesis. Measurement of breath hydrogen while fasting and after oral mono- or oligosaccharide challenge can screen for carbohydrate intolerance and small-intestinal bacterial overgrowth. Urea breath testing assesses for persistent *Helicobacter pylori* infection, while a recently approved gastric emptying breath test is an alternative to scintigraphy for gastroparesis diagnosis.

TREATMENT

TREATMENT

Gastrointestinal Disease

Management options for GI diseases depend on the cause of symptoms. Available treatments include modifications in dietary intake, medications, treatment of gut dysbiosis, interventional endoscopy or radiology techniques, surgery, and therapies directed to external influences. Given the hereditary predisposition of many GI diseases, genetic testing may be indicated in some patients.

NUTRITIONAL MANIPULATION

Dietary modifications for GI disease include treatments that only reduce symptoms, therapies correct pathologic defects, or replace normal food intake with enteral or parenteral formulations. Changes that improve symptoms but do not reverse organic abnormalities include lactose restriction for lactase deficiency, liquid meals in gastroparesis, carbohydrate restrictions with dumping syndrome, and low-FODMAP (fermentable oligo-di-monosaccharides and polyols) diets in IBS. The gluten-free diet for celiac disease exemplifies a primary therapy to reduce mucosal inflammation. Likewise, elimination diets may improve histology in some cases of eosinophilic esophagitis. Medium-chain triglycerides replace normal fats in short-gut syndrome or severe ileal disease. Perfusing liquid meals through a gastrostomy is performed in those who

cannot swallow safely. Enteral jejunostomy feedings are considered for gastric dysmotility syndromes that preclude feeding into the stomach. Intravenous hyperalimentation is used for generalized gut malfunction which does not permit enteral nutrition.

PHARMACOTHERAPY

Several medications can treat GI diseases. Considerable resources are expended on over-the-counter remedies. Many prescription drug classes are offered as short-term or continuous therapy of GI illness. Alternative treatments have gained popularity in conditions for which traditional therapies provide incomplete relief.

Over-the-Counter Agents

Over-the-counter agents are reserved for mild GI symptoms. Antacids and histamine H_2 antagonists decrease symptoms in gastroesophageal reflux disease (GERD) and dyspepsia. Potent acid inhibitors are available over the counter for treatment of more persistent GERD. Fiber supplements, stool softeners, enemas, and laxatives are used for constipation. Laxatives are categorized as stimulants, osmotic agents (including isotonic preparations containing polyethylene glycol), and poorly absorbed sugars. Nonprescription antidiarrheal agents include bismuth subsalicylate, kaolin-pectin combinations, and loperamide. Supplemental enzymes include lactase pills for lactose intolerance and bacterial α-galactosidase to treat excess gas. Capsules containing peppermint oil are available over the counter for treating discomfort in IBS and dyspepsia, while antiflatulents and adsorbents reduce gaseous symptoms. In general, using a nonprescription preparation for more than a short time for chronic persistent symptoms should be supervised by a health care provider.

Prescription Drugs

Prescription drugs are approved for a broad range of GI diseases. Higher dose prescription proton pump inhibitors are advocated for GERD when over-the-counter preparations are inadequate. Cytoprotective agents are available for upper gut ulcers but are less frequently prescribed. Prokinetic drugs stimulate GI propulsion in gastroparesis and pseudoobstruction. Prosecretory drugs are prescribed for constipation refractory to other agents, while peripheral opiate antagonists are offered for opiate-induced constipation. Prescription antidiarrheals include opiate drugs, anticholinergic antispasmodics, tricyclics, bile acid binders, and serotonin antagonists. Antispasmodics and antidepressants also are useful for functional GI disorders, whereas narcotics are used for pain control in organic conditions such as disseminated malignancy and chronic pancreatitis. Antiemetics reduce nausea and vomiting. Potent pancreatic enzymes decrease malabsorption and pain from pancreatic disease. Antisecretory drugs such as the somatostatin analogue octreotide treat hypersecretory states. Antibiotics treat *Helicobacter pylori*-induced ulcers, infectious diarrhea, diverticulitis, intestinal bacterial overgrowth, and Crohn's disease. Anti-inflammatory and immunomodulatory drugs are used in IBD, microscopic colitis, refractory celiac disease, and gut vasculitis. Over the past decade, several newer biologic agents including those with anti-tumor necrosis factor activity have had dramatic impact in Crohn's disease and ulcerative colitis. Chemotherapy with or without radiotherapy is offered for GI malignancies. Most GI carcinomas respond poorly to such therapy, whereas lymphomas may be cured with such intervention.

Complementary and Alternative Medicine Treatments

Alternative treatments are marketed to treat selected GI symptoms. Ginger, acupressure, and acustimulation have been advocated for nausea, whereas pyridoxine has been investigated for nausea of first-trimester pregnancy. Herbal preparations like STW 5 (Iberogast, a mixture of nine herbs) are useful in cases of functional dyspepsia and IBS. Low-potency pancreatic enzyme preparations are sold as general digestive aids but have little evidence to support their efficacy.

THERAPIES TARGETING GUT DYSBIOSIS

Some cases of diarrhea predominant-IBS respond to nonabsorbable antibiotics. Oral antibiotics also are the mainstay of managing intestinal bacterial overgrowth. Probiotics containing active bacterial cultures and prebiotics that selectively nourish non-noxious commensal bacteria are used as adjuncts in some cases of infectious diarrhea and IBS. Transplantation of donor feces into the colon by colonoscopy or enema has become accepted and effective treatment for recurrent, refractory *Clostridium difficile* colitis.

INTERVENTIONAL ENDOSCOPY AND RADIOLOGY

Gut luminal intubations are performed in some situations. Nasogastric tube suction decompresses the upper gut in ileus or mechanical obstruction. Nasogastric lavage of saline or water in the patient with upper GI hemorrhage determines the rate of bleeding and helps evacuate blood before therapeutic endoscopy. Enteral feedings can be delivered through nasogastric or nasoenteric tubes. Enemas relieve fecal impaction or assist in gas evacuation in acute colonic pseudoobstruction. A rectal tube can be placed to vent the distal colon in colonic pseudoobstruction and other colonic distention disorders.

In addition to its diagnostic role, endoscopy has therapeutic capabilities in many settings. Cautery techniques and injection of vasoconstrictor substances can stop hemorrhage from ulcers and vascular malformations. Endoscopic encirclement of varices and hemorrhoids with constricting bands stops hemorrhage from these sites, whereas endoscopically placed clips can occlude arterial bleeding sites. Cyanoacrylate and hemostatic powder sprays have been evaluated for abilities to stop brisk GI bleeding. Endoscopy can remove polyps or debulk lumen-narrowing malignancies. Colonoscopy is used to withdraw luminal gas in some cases of acute colonic pseudoobstruction. Endoscopic mucosal resection, submucosal dissection, and radiofrequency techniques can ablate some cases of Barrett's esophagus with dysplasia or superficial cancer and early gastric malignancies. Obstructions of the gut lumen and pancreaticobiliary tree are relieved by endoscopic dilation or placing plastic or expandable metal stents. Endoscopic sphincterotomy of the ampulla of Vater relieves symptoms of choledocholithiasis. Cholangioscopy can help with stone lithotripsy in the common bile duct, ablation of small ductal tumors, and placement of gallbladder stents to facilitate drainage in non-operative candidates. Endoscopic methods have been developed for pancreatic cyst gastrostomy, pancreatic necrosectomy, and placement of fiducial markers to direct pancreatic and rectal radiotherapy. Endoscopy is commonly used to insert gastric feeding tubes. Peroral endoscopic myotomy is now being performed on the lower esophageal sphincter in achalasia and on the pylorus in gastroparesis by selected endoscopists. Endoscopic treatments for acid reflux including radiofrequency therapy, transoral fundoplication, endoscopic stapling, and antireflux mucosectomy have been devised. Similarly, endoscopic bariatric methodologies including intragastric balloons, aspiration therapy, gastroplasty, and duodenal bypass are in use or in development.

Radiologic measures also are useful in GI disease. Angiographic embolization or vasoconstriction decreases bleeding from gut sites not amenable to endoscopic intervention. Dilatation or stenting with fluoroscopic guidance relieves luminal strictures. Contrast enemas can reduce volvulus and evacuate air in acute colonic pseudoobstruction. CT and ultrasound help drain abdominal fluid collections, in many cases obviating the need for surgery. Percutaneous transhepatic cholangiography relieves biliary obstruction when ERCP is contraindicated. Transjugular intrahepatic portosystemic shunts are commonly performed by interventional radiologists for variceal hemorrhage not amenable to endoscopic therapy. Lithotripsy can fragment gallstones in patients who are not candidates for surgery. In some instances, radiologic approaches offer advantages over endoscopy for gastroenterostomy placement. Finally, central venous catheters for parenteral nutrition may be placed using radiographic techniques.

SURGERY

Surgery is performed to cure disease, control symptoms without cure, maintain nutrition, or palliate unresectable neoplasm. Medication-unresponsive ulcerative colitis, diverticulitis, cholecystitis, appendicitis, and intraabdominal abscess are curable with surgery, whereas symptom control without cure is only possible with Crohn's disease. Surgery is mandated for ulcer complications such as bleeding, obstruction, or perforation and intestinal obstructions that persist after conservative care. Fundoplication of the gastroesophageal junction is performed for severe ulcerative esophagitis and drug-refractory symptomatic acid reflux. Achalasia responds to operations to reduce lower esophageal sphincter tone. Operations for motor disorders have been introduced including implanted electrical stimulators for gastroparesis and electrical devices and artificial sphincters for fecal incontinence. Surgery may be needed to place a jejunostomy for long-term enteral feedings. The threshold for performing surgery depends on the clinical setting. In all cases, the benefits of operation must be weighed against the potential for postoperative complications.

THERAPY DIRECTED TO EXTERNAL INFLUENCES

In some conditions, GI symptoms respond to treatments directed outside the gut. Psychological therapies including psychotherapy, behavior modification, and hypnosis, have shown efficacy in functional bowel disorders. Patients with significant psychological dysfunction and those with little response to treatments targeting the gut are likely to benefit from this form of therapy. Biofeedback methods administered by physical therapies are accepted for treating refractory fecal incontinence or constipation secondary to dyssynergia.

FURTHER READING

Beg S et al: The use of optical imaging techniques in the gastrointestinal tract. Frontline Gastroenterol 7:207, 2016. [PubMed: 27429735]

Goodman RP, Chung DC: Clinical genetic testing in gastroenterology. Clin Transl Gastroenterol 7:e167, 2016. [PubMed: 27124700]

Panes J et al: Advances in use of endoscopy, radiology, and biomarkers to monitor inflammatory bowel diseases. Gastroenterology 152:362, 2017. [PubMed: 27751880]

Peery AF et al: Burden of gastrointestinal, liver, and pancreatic diseases in the United States. Gastroenterology 149:1731, 2015. [PubMed: 26327134]

Schreuders EH et al: Advances in fecal tests for colorectal cancer screening. Curr Treat Options Gastroenterol 14:152, 2016. [PubMed: 26825703]

Weilert F, Binmoeller KF: New endoscopic technologies and procedural advances for endoscopic hemostasis. Clin Gastroenterol Hepatol 14:1234, 2016. [PubMed: 27215365]

Chapter 315: Gastrointestinal Endoscopy

Louis Michel Wong Kee Song; Mark Topazian

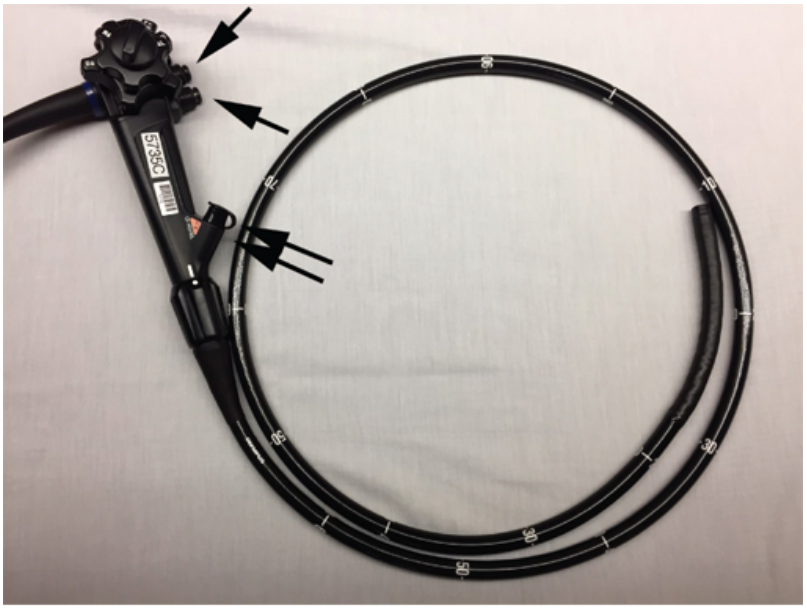
INTRODUCTION

Gastrointestinal endoscopy has been attempted for over 200 years, but the introduction of semirigid gastroscopes in the middle of the twentieth century marked the dawn of the modern endoscopic era. Since then rapid advances in endoscopic technology have led to dramatic changes in the diagnosis and treatment of many digestive diseases. Innovative endoscopic devices and new endoscopic treatment modalities continue to expand the use of endoscopy in patient care.

Flexible endoscopes provide an electronic video image generated by a charge-coupled device in the tip of the endoscope. Operator controls permit deflection of the endoscope tip; fiberoptic bundles or light-emitting diodes provide light at the tip of the endoscope; and working channels allow washing, suctioning, and the passage of instruments **(Fig. 315-1)**. Progressive changes in the diameter and stiffness of endoscopes have improved the ease and patient tolerance of endoscopy.

FIGURE 315-1

Gastrointestinal endoscope. Shown here is a conventional colonoscope with control knobs for tip deflection, push buttons for suction and air insufflation (*single arrows*), and a working channel for passage of accessories *(double arrows)*.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ENDOSCOPIC PROCEDURES

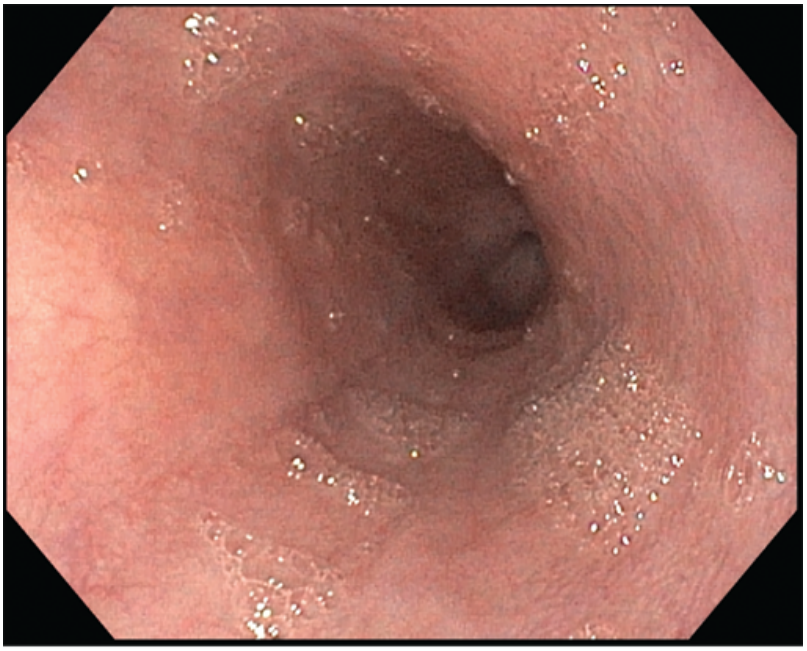
UPPER ENDOSCOPY

Upper endoscopy, also referred to as esophagogastroduodenoscopy (EGD), is performed by passing a flexible endoscope through the mouth into the esophagus, stomach, and duodenum. The procedure is the best method for examining the upper gastrointestinal mucosa **(Fig. 315-2)**. While the upper gastrointestinal radiographic series has similar accuracy for diagnosis of duodenal ulcer **(Fig. 315-3)**, EGD is superior for detection of gastric ulcers **(Fig. 315-4)** and flat mucosal lesions, such as Barrett's esophagus **(Fig. 315-5)**, and it permits directed biopsy and endoscopic therapy. Intravenous conscious sedation is given to most patients in the United States to ease the anxiety and

discomfort of the procedure, although in many countries EGD is routinely performed with topical pharyngeal anesthesia only. Patient tolerance of unsedated EGD is improved by the use of an ultrathin, 5-mm diameter endoscope that can be passed transorally or transnasally.

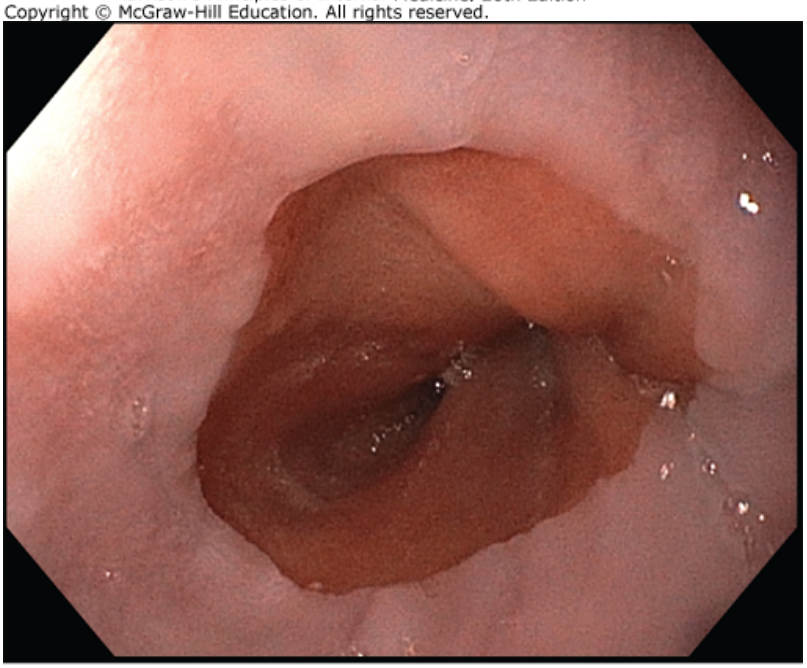
FIGURE 315-2

Normal upper endoscopic examination. ***A.*** Esophagus. ***B.*** Gastroesophageal junction. ***C.*** Gastric fundus. ***D.*** Gastric body. ***E.*** Gastric antrum. ***F.*** Pylorus. ***G.*** Duodenal bulb. ***H.*** Second portion of the duodenum.



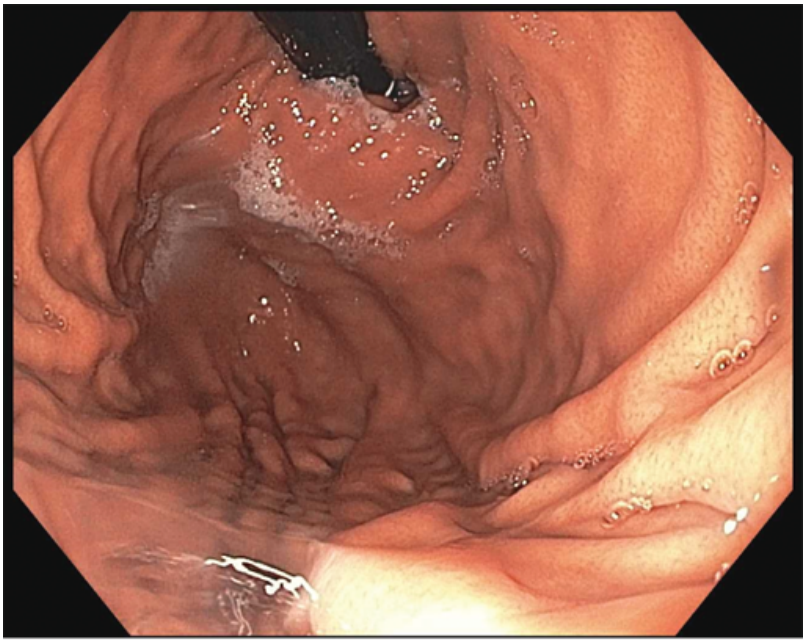
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



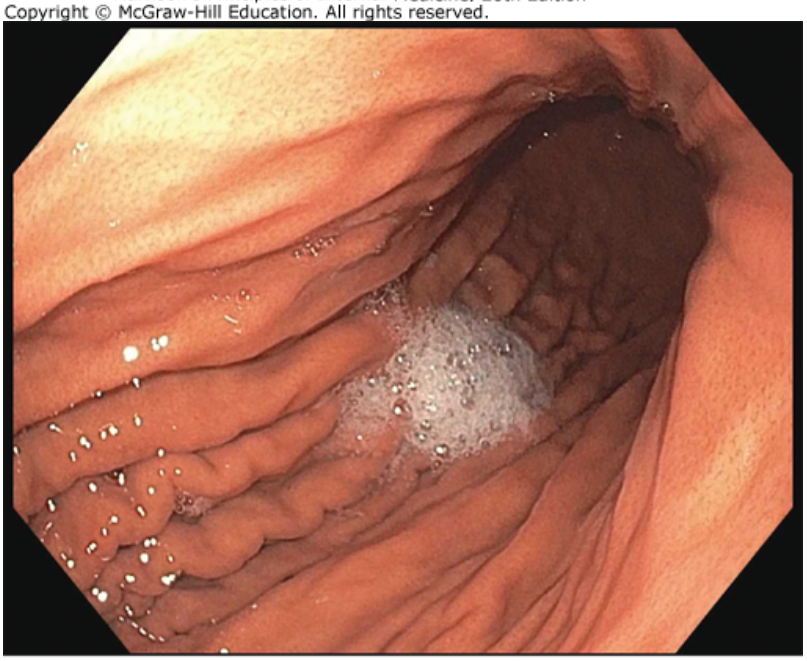
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



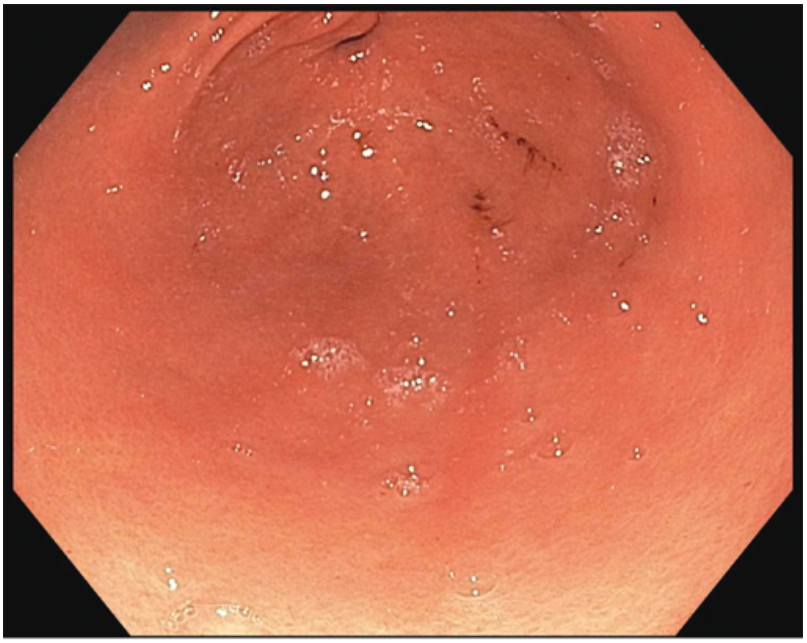
C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



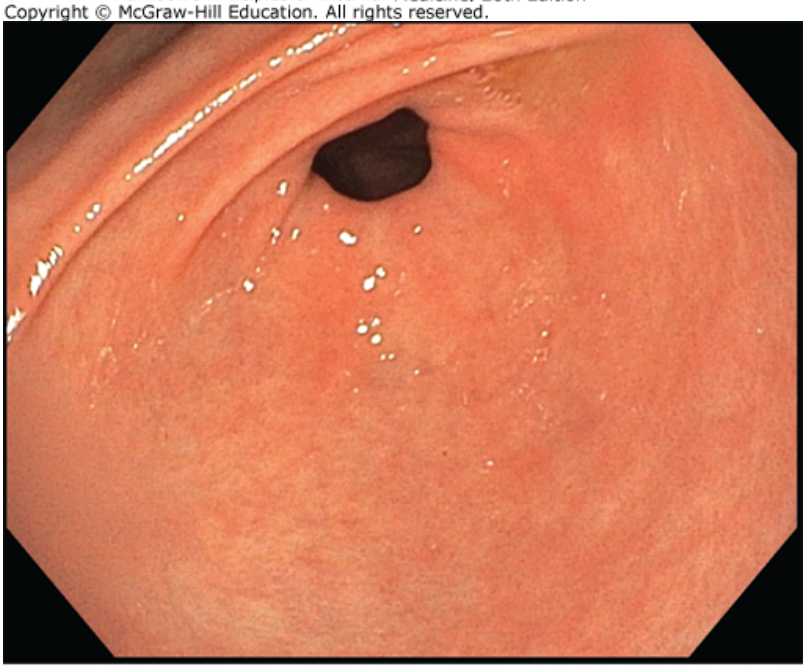
D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



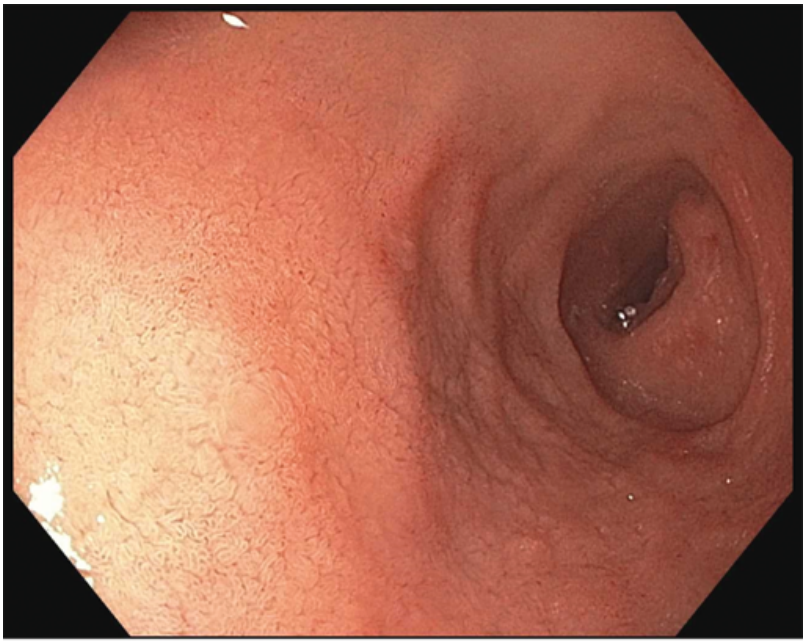
E

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



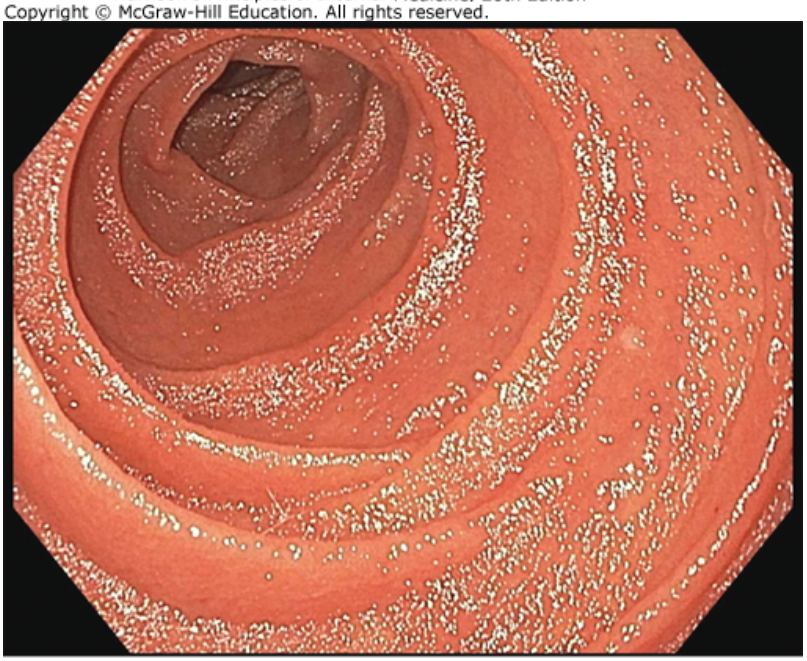
F

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



G

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

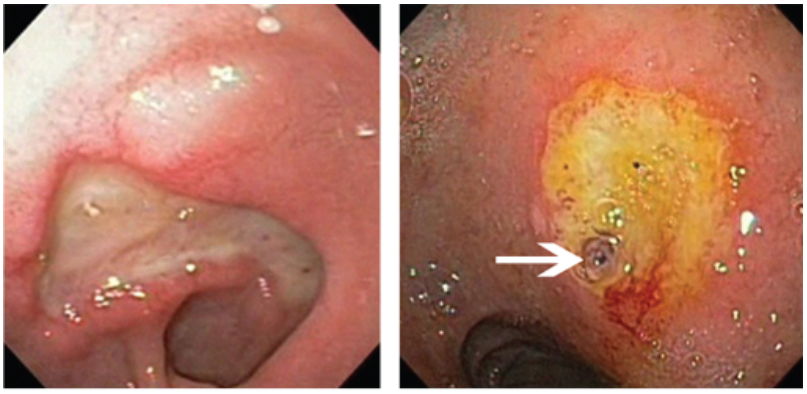


H

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-3

Duodenal ulcers. ***A.*** Ulcer with a clean base. ***B.*** Ulcer with a visible vessel *(arrow)* in a patient with recent hemorrhage.

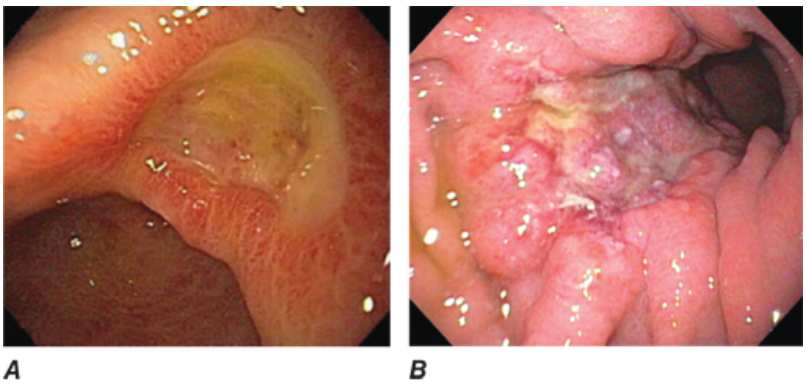


A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-4

Gastric ulcers. ***A.*** Benign gastric ulcer. ***B.*** Malignant gastric ulcer involving greater curvature of stomach.

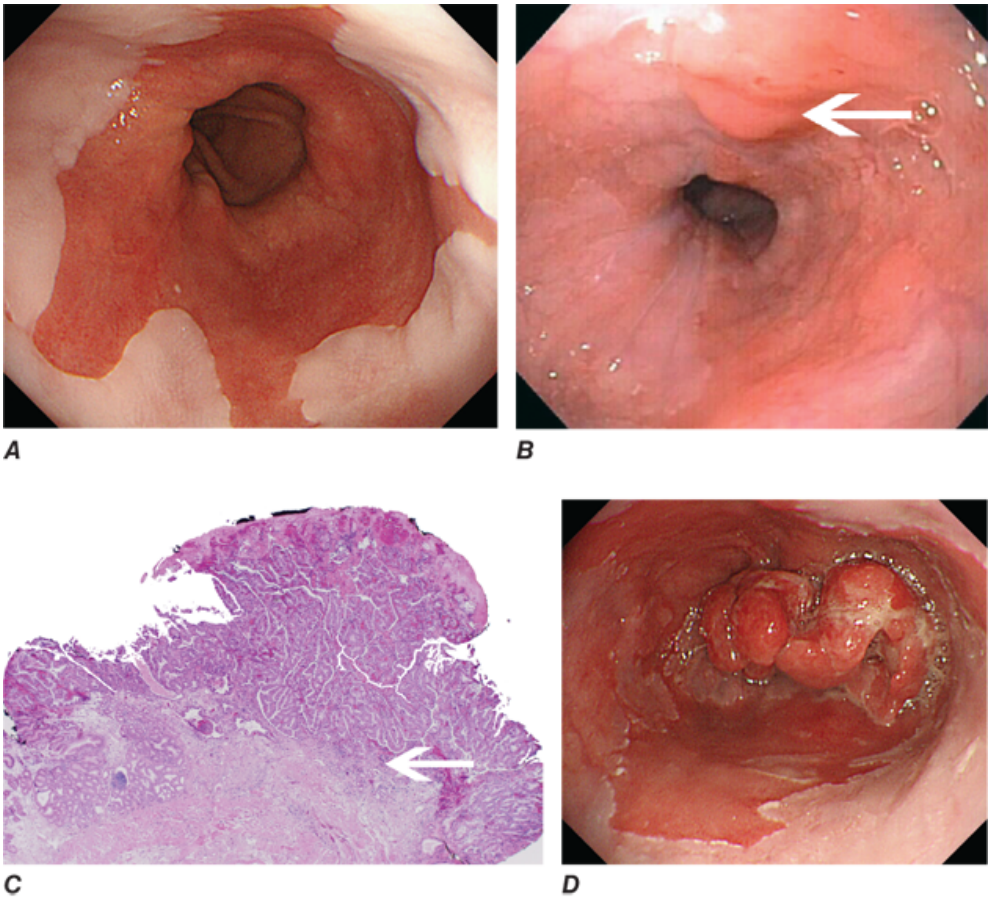


A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-5

Barrett's esophagus. ***A.*** Pink tongues of Barrett's mucosa extending proximally from the gastroesophageal junction. ***B.*** Barrett's esophagus with a suspicious nodule *(arrow)* identified during endoscopic surveillance. ***C.*** Histologic finding of intramucosal adenocarcinoma in the endoscopically resected nodule. Tumor extends into the esophageal submucosa *(arrow).* ***D.*** Barrett's esophagus with locally advanced adenocarcinoma.



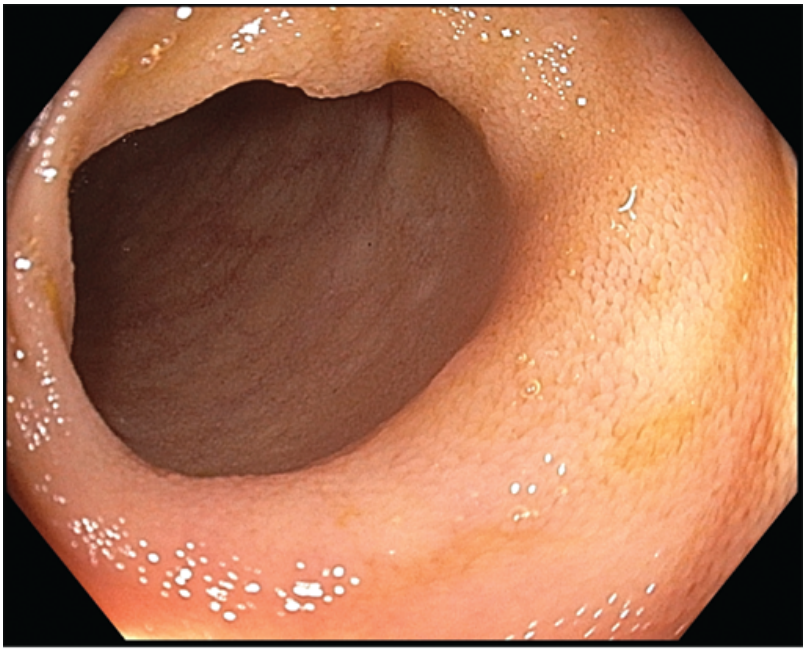
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

COLONOSCOPY

Colonoscopy is performed by passing a flexible colonoscope through the anal canal into the rectum and colon. The cecum is reached in >95% of cases and the terminal ileum **(Fig. 315-6)** can often be examined. Colonoscopy is the gold standard for imaging the colonic mucosa **(Fig. 315-7)**. Colonoscopy has greater sensitivity than barium enema for colitis **(Fig. 315-8)**, polyps **(Fig. 315-9)**, and cancer **(Fig. 315-10)**. CT colonography rivals the accuracy of colonoscopy for detection of some polyps and cancer, although it is not as sensitive for the detection of flat lesions, such as serrated polyps **(Fig. 315-11)**. Conscious sedation is usually given before colonoscopy in the United States, although a willing patient and a skilled examiner can complete the procedure without sedation in many cases.

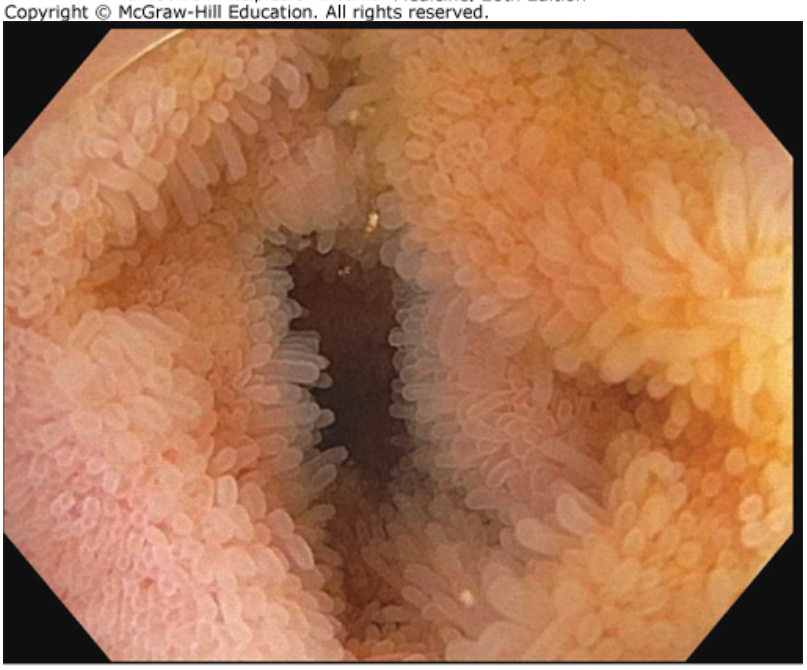
FIGURE 315-6

Colonoscopic view of terminal ileum. ***A.*** Normal-appearing terminal ileum (TI). ***B.*** View of normal villi of TI enhanced by examination under water immersion.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

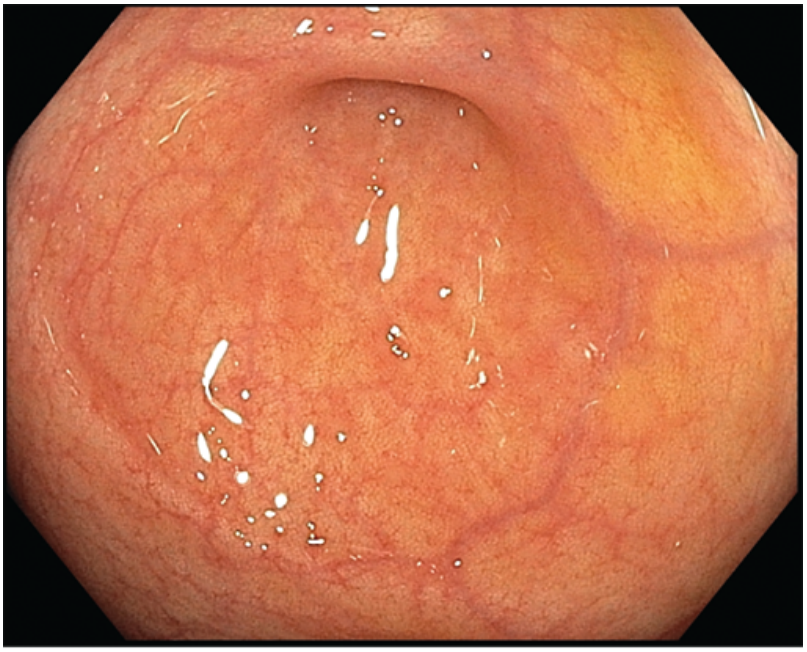


B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

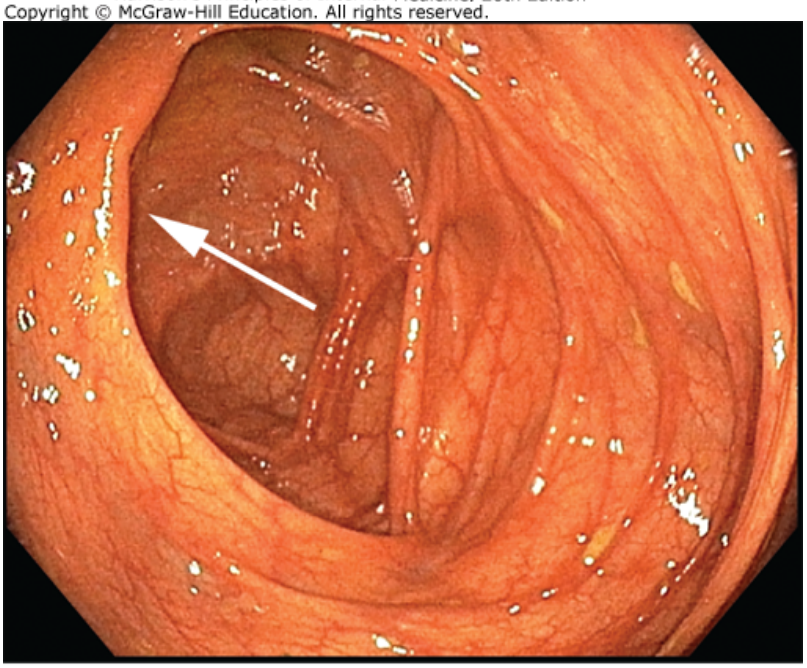
FIGURE 315-7

Normal colonoscopic examination. ***A.*** Cecum with view of appendiceal orifice. ***B.*** Ileocecal valve. ***C.*** Normal-appearing colon. ***D.*** Rectum (retroflexed view).



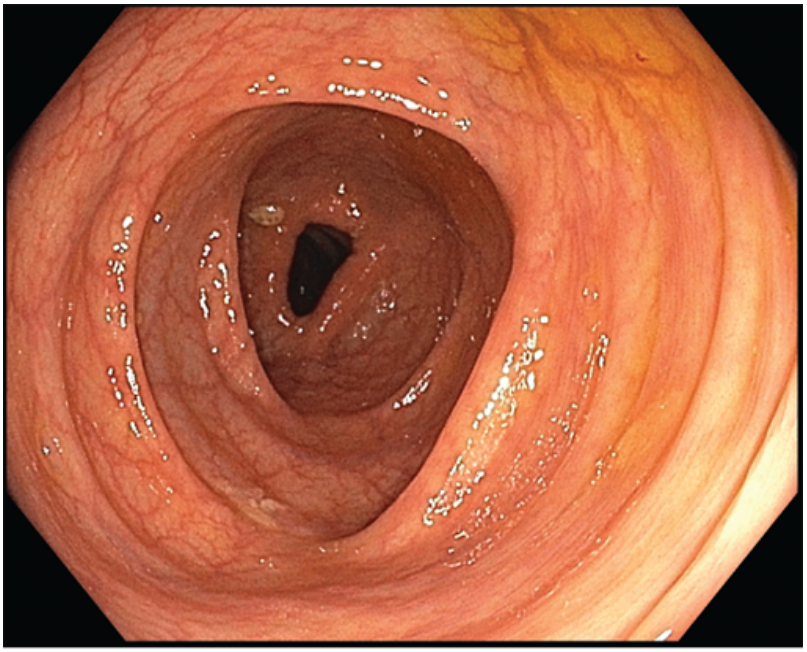
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



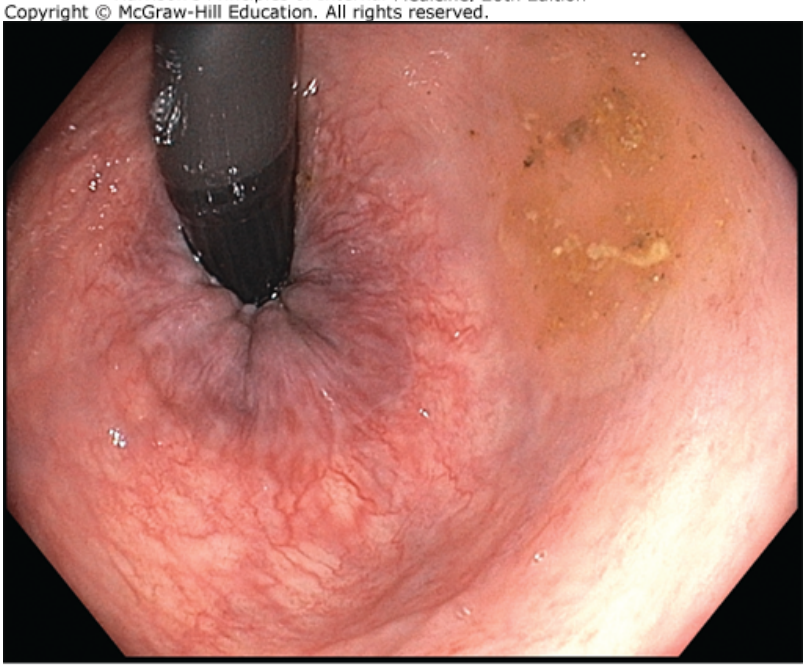
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

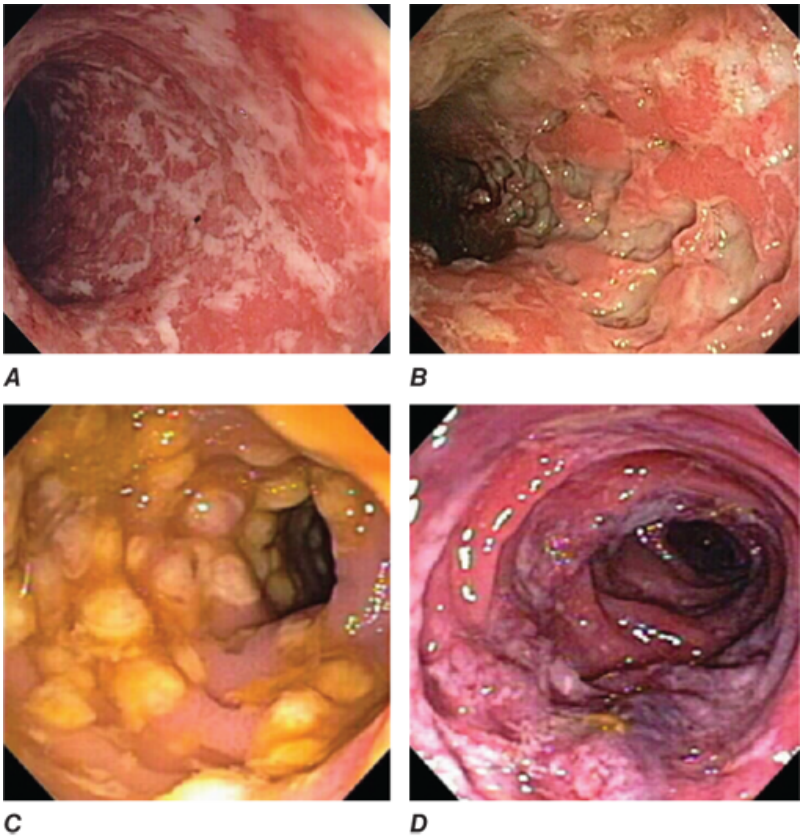


D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-8

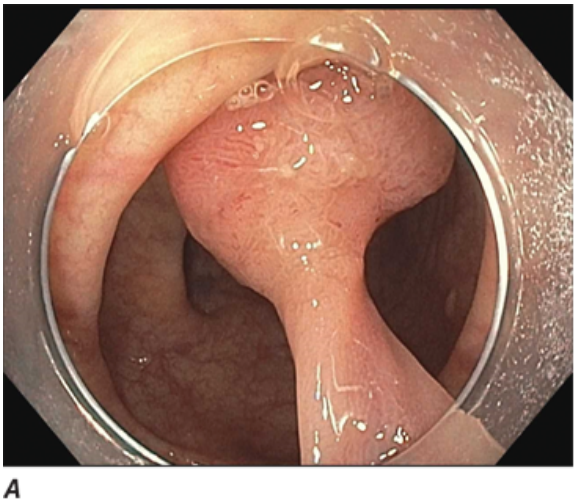
Causes of colitis. ***A.*** Chronic ulcerative colitis with diffuse ulcerations and exudates. ***B.*** Severe Crohn's colitis with deep ulcers. ***C.*** Pseudomembranous colitis with yellow, adherent pseudomembranes. ***D.*** Ischemic colitis with patchy mucosal edema, subepithelial hemorrhage, and cyanosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-9

Colonic polyps. ***A.*** Pedunculated polyp on a stalk. ***B.*** Sessile polyp.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

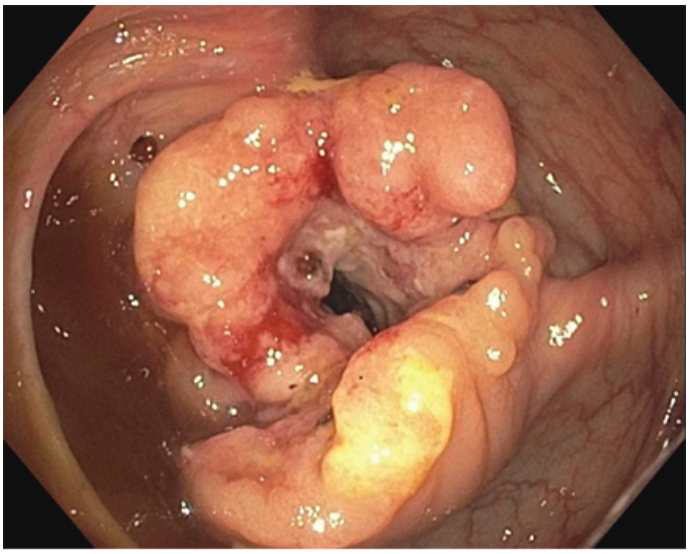


B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-10

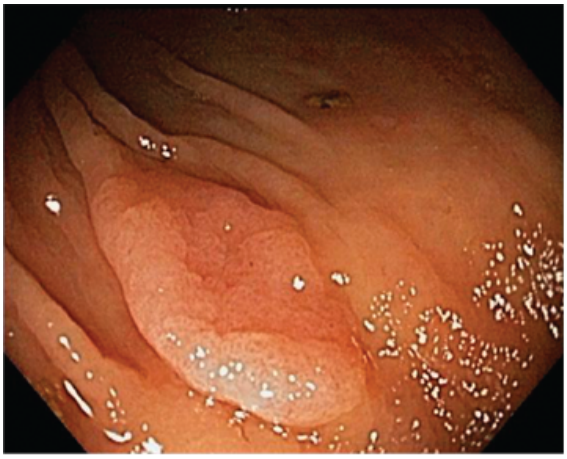
Ulcerated colon adenocarcinoma narrowing the colonic lumen.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-11

Flat serrated polyp in the cecum. ***A.*** Appearance of the lesion under conventional white-light imaging. ***B.*** Mucosal patterns and boundary of the lesion enhanced with narrow band imaging. ***C.*** Submucosal lifting of the lesion with dye (methylene blue) injection prior to resection.



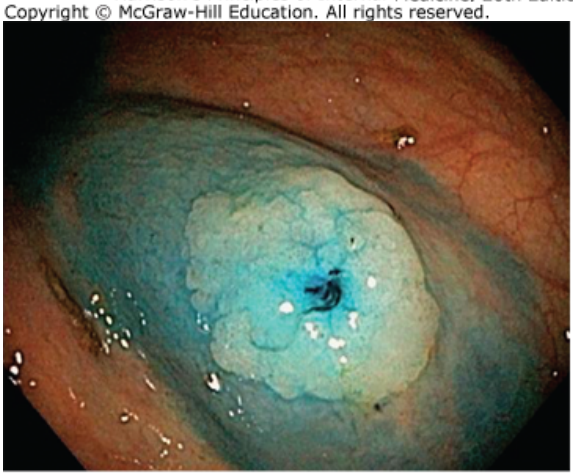
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FLEXIBLE SIGMOIDOSCOPY

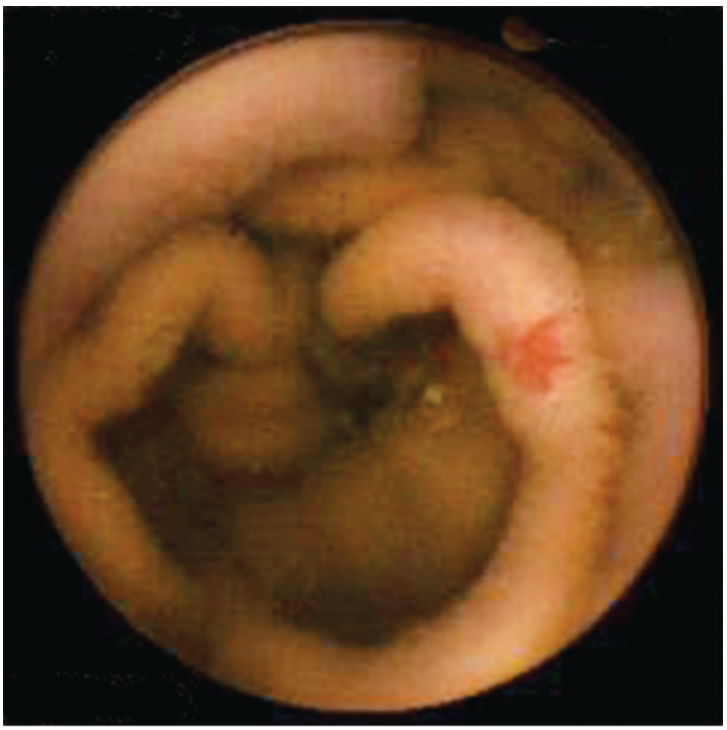
Flexible sigmoidoscopy is similar to colonoscopy, but it visualizes only the rectum and a variable portion of the left colon, typically to 60 cm from the anal verge. This procedure causes abdominal cramping, but it is brief and is usually performed without sedation. Flexible sigmoidoscopy is primarily used for evaluation of diarrhea and rectal outlet bleeding.

SMALL BOWEL ENDOSCOPY

Three endoscopic techniques are currently used to evaluate the small intestine, most often in patients presenting with presumed small bowel bleeding. For *capsule endoscopy*, the patient swallows a disposable capsule that contains a complementary metal oxide silicon (CMOS) chip camera. Color still images **(Fig. 315-12)** are transmitted wirelessly to an external receiver at several frames per second until the capsule's battery is exhausted or it is passed into the toilet. Capsule endoscopy enables visualization of the small bowel mucosa beyond the reach of a conventional endoscope, and at present it is solely a diagnostic procedure. Patients with a history of prior intestinal surgery or Crohn's disease are at risk for capsule retention at the site of a clinically unsuspected small bowel stricture, and ingestion of a "patency capsule" composed of radiologically opaque biodegradable material may be indicated prior to capsule endoscopy in such patients.

FIGURE 315-12

Capsule endoscopy image of jejunal vascular ectasia.



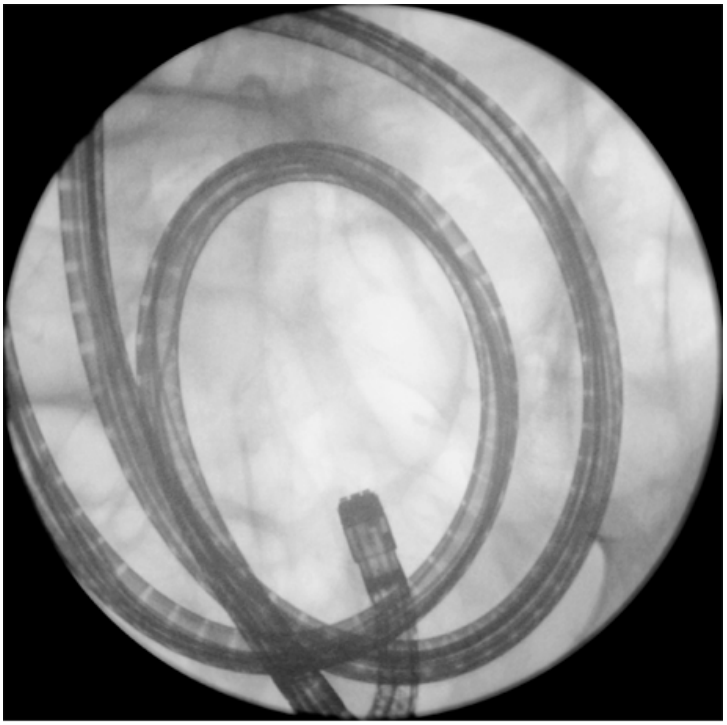
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Push enteroscopy is performed with a long endoscope similar in design to an upper endoscope. The enteroscope is pushed down the small bowel, sometimes with the help of a stiffening overtube that extends from the mouth to the small intestine. The proximal to mid-jejunum is usually reached, and the instrument channel of the endoscope allows for biopsy or endoscopic therapy.

Deeper insertion into the small bowel can be accomplished by *device-assisted enteroscopy,* which may utilize inflatable balloons at the tip of the enteroscope and/or an overtube (*single-* or *double-balloon enteroscopy*) or a rotating, screw-like overtube (*spiral enteroscopy*) to pleat the small intestine onto the endoscope **(Fig. 315-13, Video V5-1)**. Using device-assisted enteroscopy the entire small intestine can be visualized in some patients when both the oral and anal routes of insertion are used. Biopsies and endoscopic therapy can be performed throughout the visualized small bowel **(Fig. 315-14)**.

FIGURE 315-13

Radiograph of a double-balloon enteroscope in the small intestine.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-14

Nonsteroidal anti-inflammatory drug (NSAID)-induced proximal ileal stricture diagnosed by double-balloon endoscopy. ***A.*** Ileal stricture causing obstructive symptoms. ***B.*** Balloon dilation of the ileal stricture. ***C.*** Appearance of stricture after dilation.



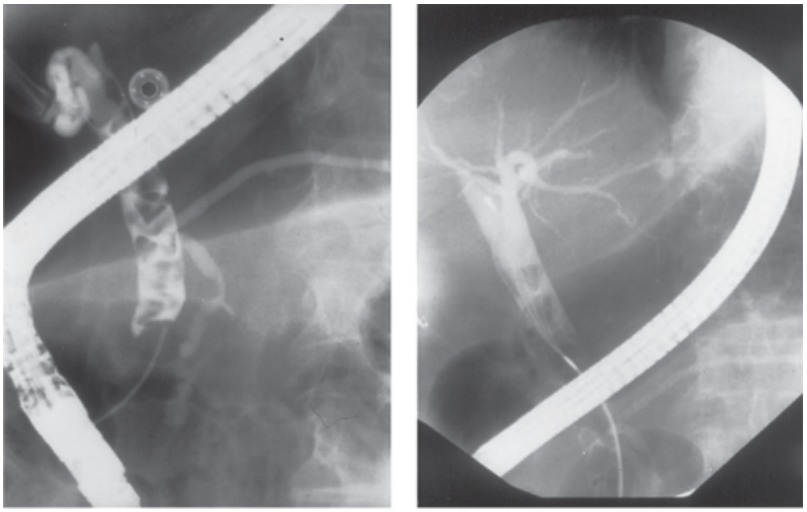
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ENDOSCOPIC RETROGRADE CHOLANGIOPANCREATOGRAPHY (ERCP)

During ERCP a side-viewing endoscope is passed through the mouth to the duodenum, the ampulla of Vater is identified and cannulated with a thin plastic catheter, and radiographic contrast material is injected into the bile duct and pancreatic duct under fluoroscopic guidance (Fig. 315-15). When indicated, the major papilla can be opened using the technique of endoscopic sphincterotomy (Fig. 315-16). Stones can be retrieved from the ducts, biopsies can be performed, and strictures can be dilated and/or stented (Fig. 315-17), and ductal leaks can be treated (Fig. 315-18). ERCP is usually performed for therapy but is also important diagnostically, and facilitates tissue sampling of biliary or pancreatic ductal strictures.

FIGURE 315-15

Endoscopic retrograde cholangiopancreatography (ERCP) for bile duct stones with cholangitis. ***A.*** Faceted bile duct stones are demonstrated in the common bile duct. ***B.*** After endoscopic sphincterotomy, the stones are extracted with a Dormia basket. A small abscess communicates with the left hepatic duct.

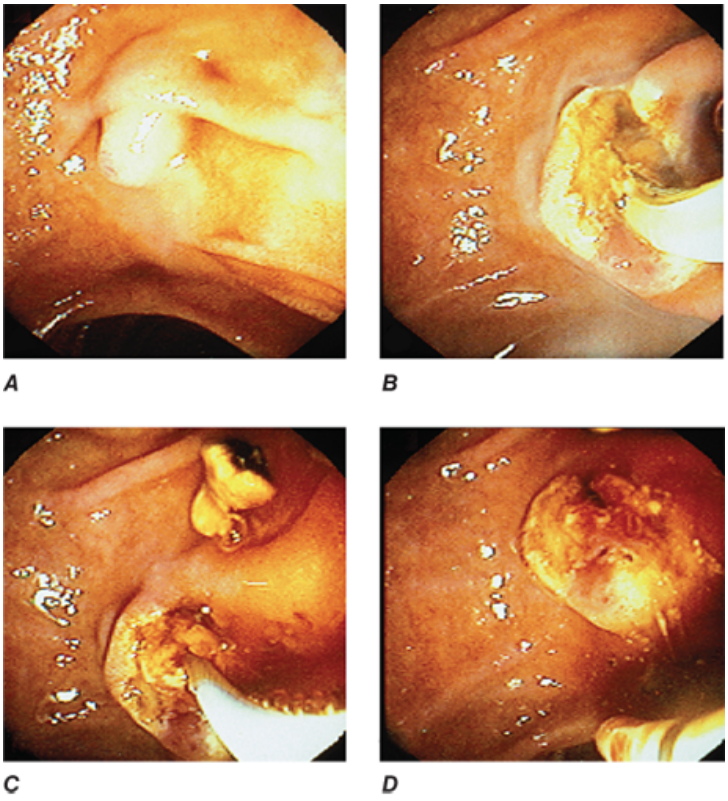


A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-16

Endoscopic sphincterotomy. ***A.*** A normal-appearing ampulla of Vater. ***B.*** Sphincterotomy is performed with electrosurgery. ***C.*** Bile duct stones are extracted with a balloon catheter. ***D.*** Final appearance of the sphincterotomy.

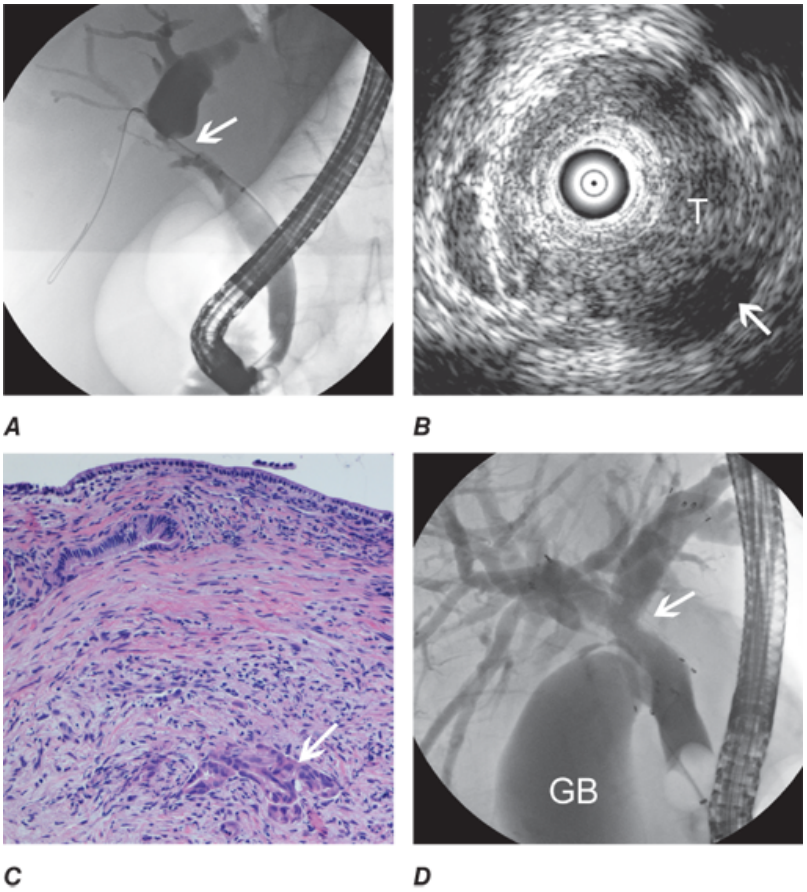


A B C D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-17

Endoscopic diagnosis, staging, and palliation of hilar cholangiocarcinoma. ***A.*** Endoscopic retrograde cholangiopancreatography (ERCP) in a patient with obstructive jaundice demonstrates a malignant-appearing stricture of the biliary confluence extending into the left and right intrahepatic ducts. ***B.*** Intraductal ultrasound of the biliary stricture demonstrates marked bile duct wall thickening due to tumor (T) with partial encasement of the hepatic artery *(arrow)*. ***C.*** Intraductal biopsy obtained during ERCP demonstrates malignant cells infiltrating the submucosa of the bile duct wall *(arrow)*. ***D.*** Endoscopic placement of bilateral self-expanding metal stents *(arrow)* relieves the biliary obstruction. GB, gallbladder. *(Image C courtesy of Dr. Thomas Smyrk; with permission.)*

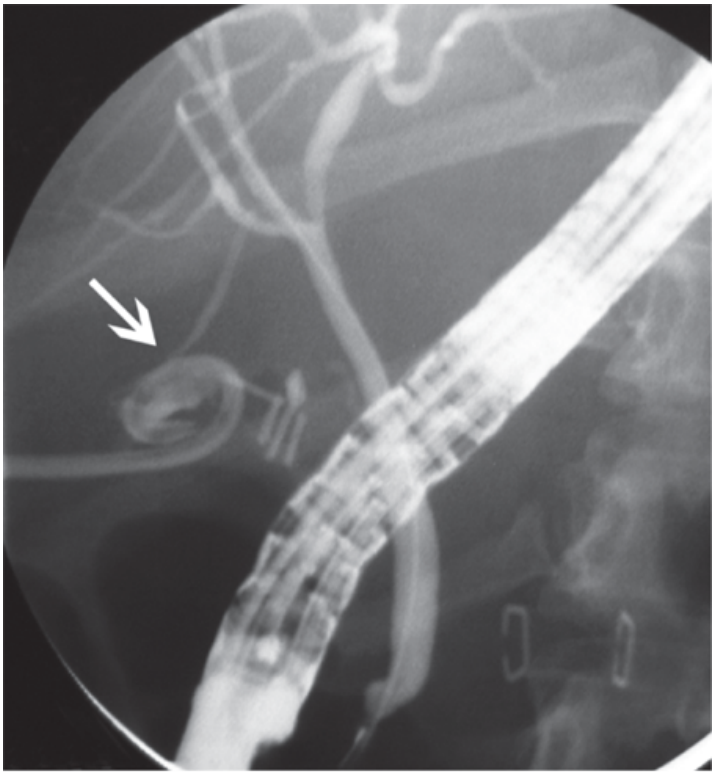


A B C D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-18

Bile leak *(arrow)* from a duct of Luschka after laparoscopic cholecystectomy. Contrast leaks from a small right intrahepatic duct into the gallbladder fossa, then flows into the pigtail of a percutaneous drainage catheter.



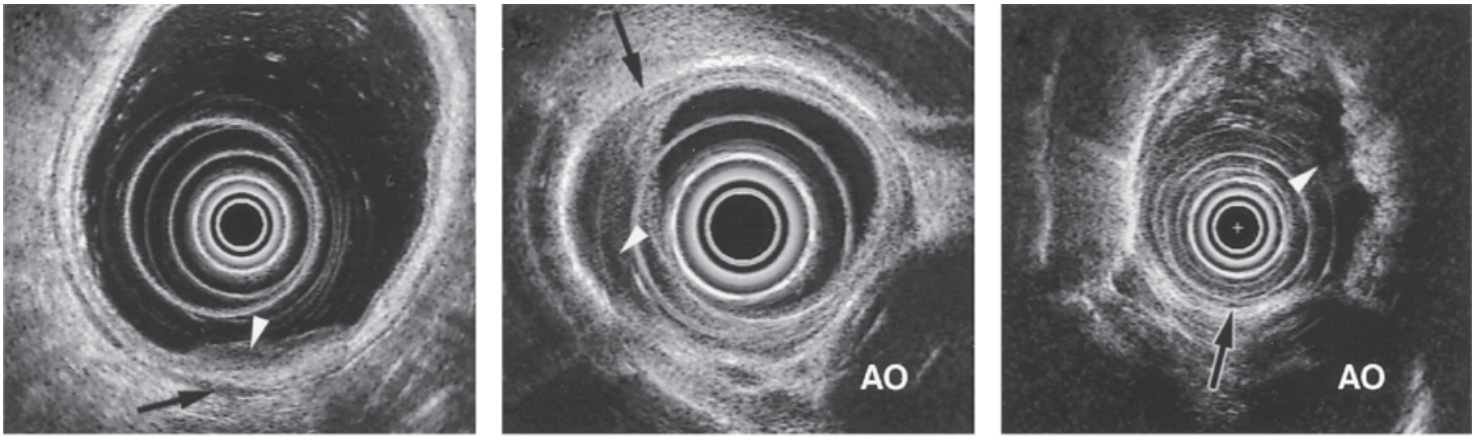
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ENDOSCOPIC ULTRASOUND (EUS)

EUS utilizes ultrasound transducers incorporated into the tip of a flexible endoscope. Ultrasound images are obtained of the gut wall and adjacent organs, vessels, lymph nodes, and other structures. High-resolution images are obtained by bringing a high-frequency ultrasound transducer close to the area of interest via endoscopy. EUS provides the most accurate preoperative local staging of esophageal, pancreatic, and rectal malignancies (Fig. 315-19), but it does not detect most distant metastases. EUS is also useful for diagnosis of bile duct stones, gallbladder disease, subepithelial gastrointestinal lesions, and chronic pancreatitis. Fine-needle aspirates and core biopsies of organs, masses and lymph nodes in the posterior mediastinum, abdomen, pancreas, retroperitoneum, and pelvis can be obtained under EUS guidance (Fig. 315-20). EUS-guided therapeutic procedures are increasingly performed, including drainage of abscesses, pseudocysts, and pancreatic necrosis into the gut lumen (Video V5-2), celiac plexus neurolysis for treatment of pancreatic pain, ethanol ablation of pancreatic neuroendocrine tumors, treatment of gastrointestinal hemorrhage, and drainage of obstructed biliary and pancreatic ducts.

FIGURE 315-19

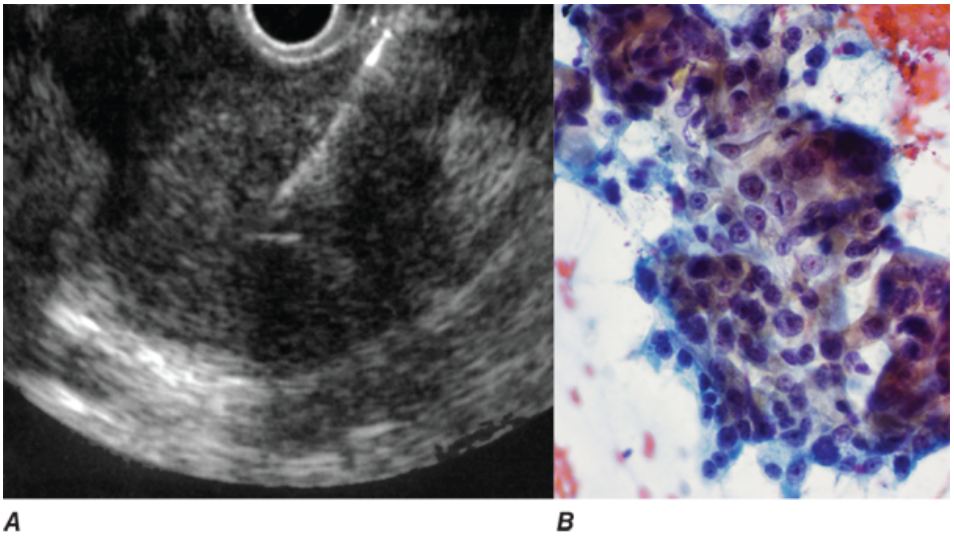
Local staging of gastrointestinal cancers with endoscopic ultrasound. In each example the white arrowhead marks the primary tumor and the black arrow indicates the muscularis propria (mp) of the intestinal wall. ***A.*** T1 gastric cancer. The tumor does not invade the mp. ***B.*** T2 esophageal cancer. The tumor invades the mp. ***C.*** T3 esophageal cancer. The tumor extends through the mp into the surrounding tissue, and focally abuts the aorta. AO, aorta.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-20

Endoscopic ultrasound (EUS)-guided fine-needle aspiration (FNA). ***A.*** Ultrasound image of a 22-gauge needle passed through the duodenal wall and positioned in a hypoechoic pancreatic head mass. ***B.*** Micrograph of aspirated malignant cells. *(Image B courtesy of Dr. Michael R. Henry; with permission.)*



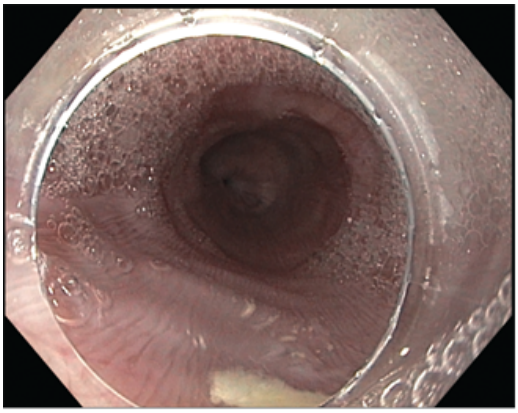
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

NATURAL ORIFICE TRANSLUMINAL ENDOSCOPIC SURGERY (NOTES)

NOTES is an evolving collection of endoscopic methods that entail passage of an endoscope or its accessories into or through the wall of the gastrointestinal tract to perform diagnostic or therapeutic interventions. Some NOTES procedures, such as percutaneous endoscopic gastrostomy (PEG) or endoscopic necrosectomy of pancreatic necrosis, are well-established clinical procedures (Video V5-2); others such as peroral endoscopic myotomy (POEM) for achalasia (Fig. 315-21), peroral endoscopic tumorectomy (POET) (Fig. 315-22), and endoscopic full-thickness resection (EFTR) of gastrointestinal mural lesions (Fig. 315-23, Video V5-3), are emerging as minimally invasive therapeutic options. NOTES is an area of continuing innovation and endoscopic research.

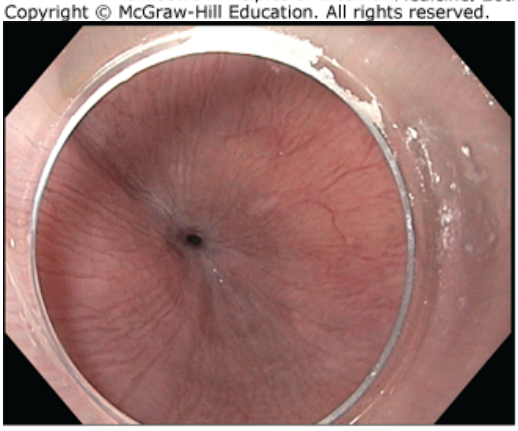
FIGURE 315-21

Peroral endoscopic myotomy (POEM) for achalasia. ***A.*** Dilated aperistaltic esophagus with retained secretions. ***B.*** Hypertonic lower esophageal sphincter (LES) region. ***C.*** Mucosal incision (mucosotomy) 10 cm proximal to the LES. ***D.*** Submucosal dissection using an electrosurgical knife following endoscope entry through the mucosotomy site into the submucosal space. ***E.*** Completion of submucosal tunnel to the cardia. ***F.*** Initiation of myotomy of the muscularis propria distal to the mucosotomy site. ***G.*** Completion of myotomy to the cardia. ***H.*** Closure of mucosotomy site with clips. ***I.*** Patulous gastroesophageal junction following myotomy.



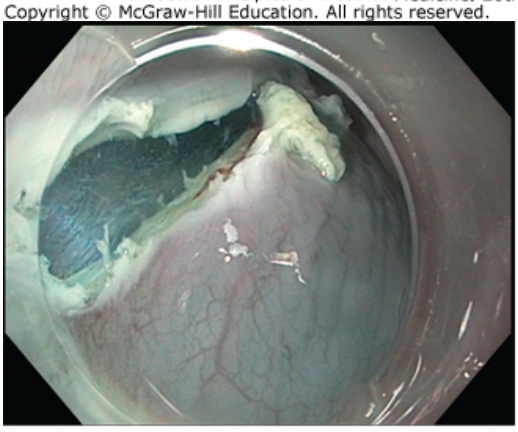
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



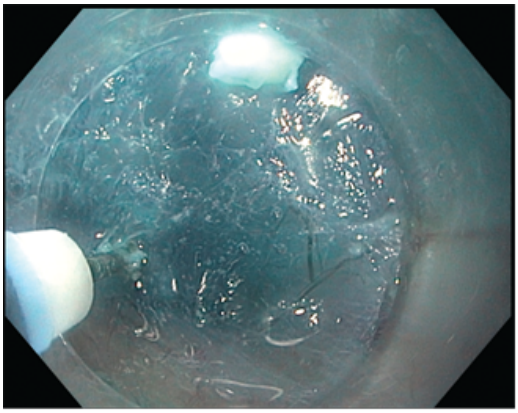
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



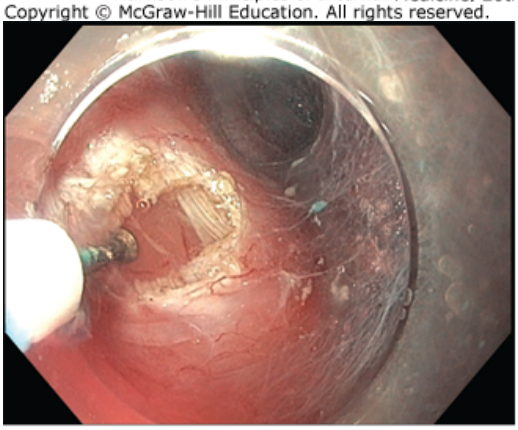
D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



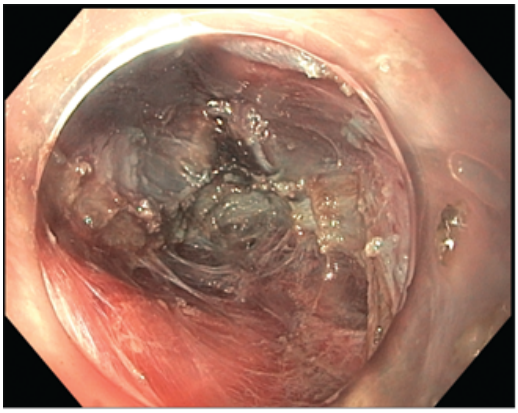
E

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



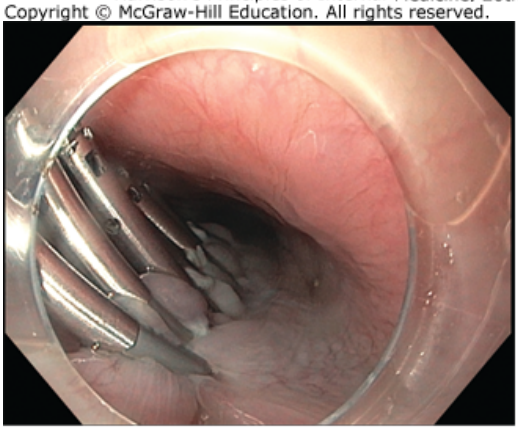
F

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



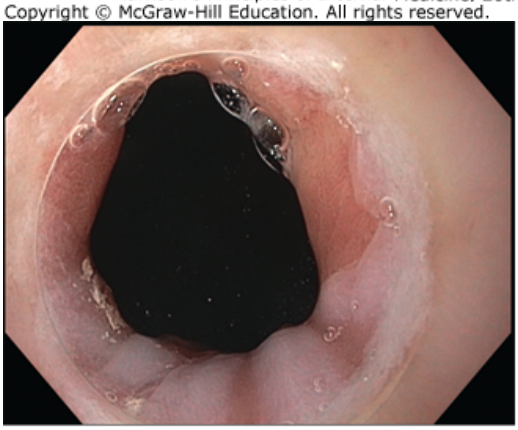
G

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



H

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

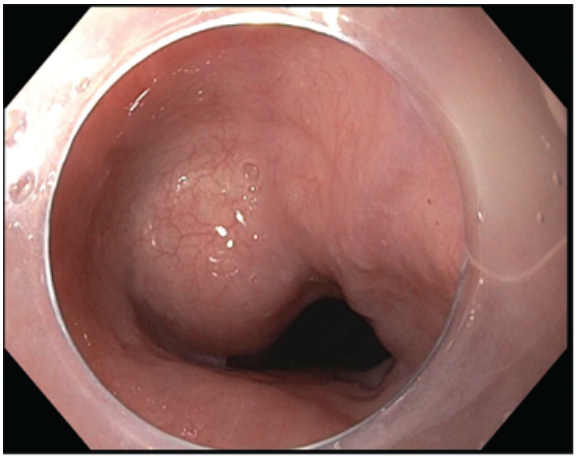


I

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-22

Peroral endoscopic tumorectomy (POET). ***A.*** Midesophageal subepithelial lesion. ***B.*** Mucosal incision (mucosotomy) 5 cm proximal to the lesion. ***C.*** Submucosal dissection and tunneling to the site of the lesion. ***D.*** Dissection of the lesion from its attachment to the muscularis propria. ***E.*** Postresection defect through the muscularis propria. ***F.*** Mucosotomy site. ***G.*** Closure of mucosotomy site with clips. ***H.*** Resected specimen (leiomyoma).



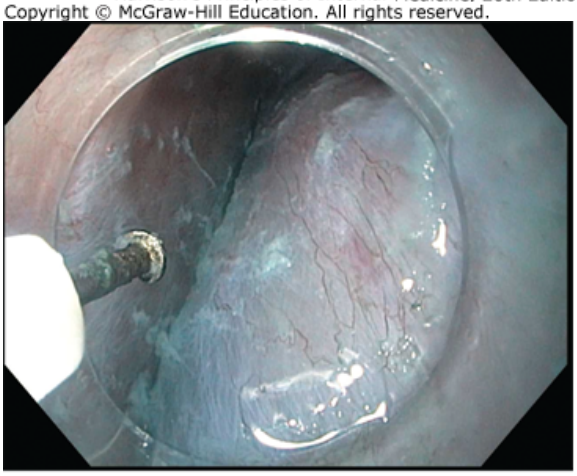
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



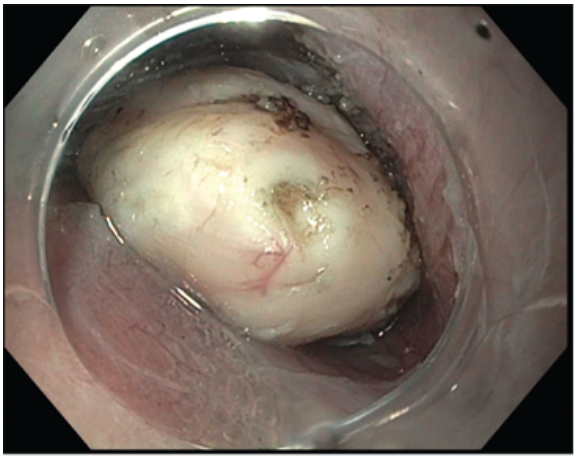
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



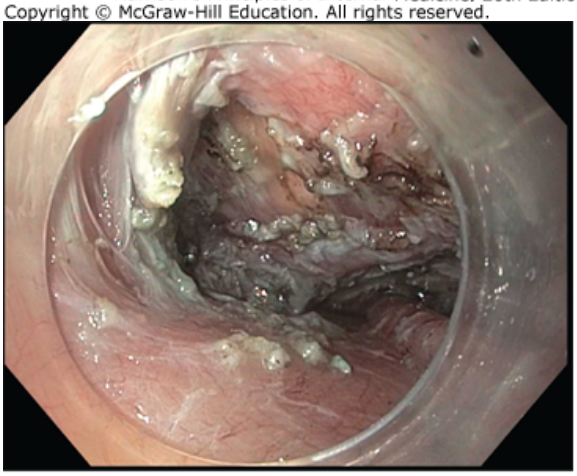
C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



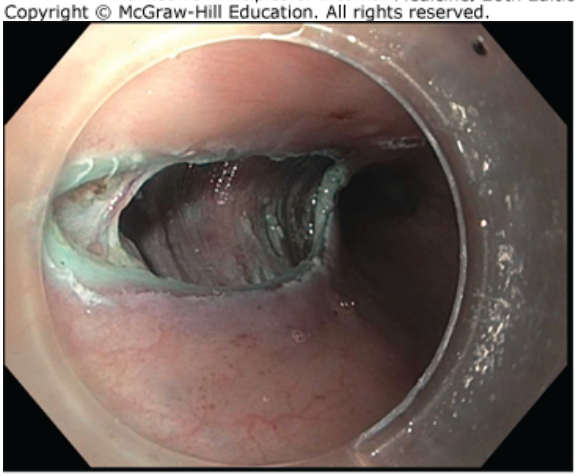
D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



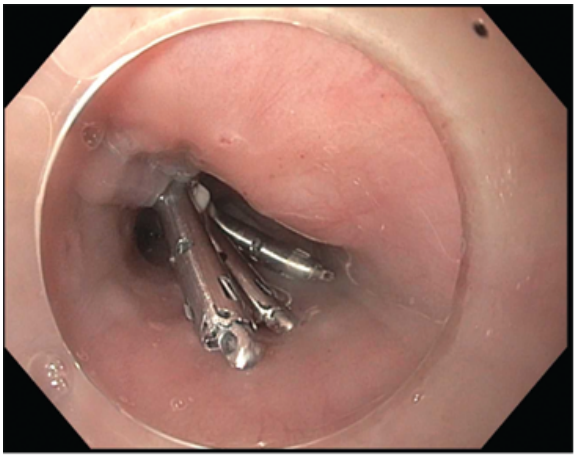
E

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



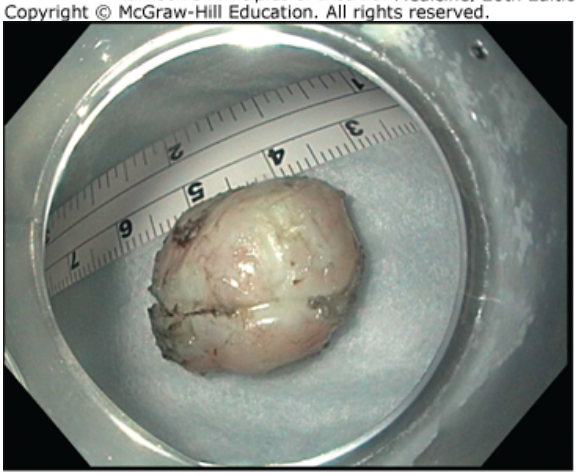
F

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



G

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

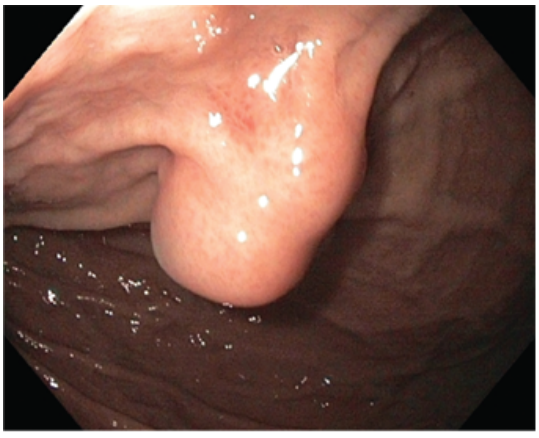


H

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-23

Endoscopic full-thickness resection (EFTR) of a gastrointestinal stromal tumor. ***A.*** Subepithelial lesion in the proximal stomach. ***B.*** Hypoechoic lesion arising from the fourth layer (muscularis propria) at endoscopic ultrasound. ***C.*** Full-thickness resection defect. ***D.*** Closure of defect using an over-the-scope clip.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



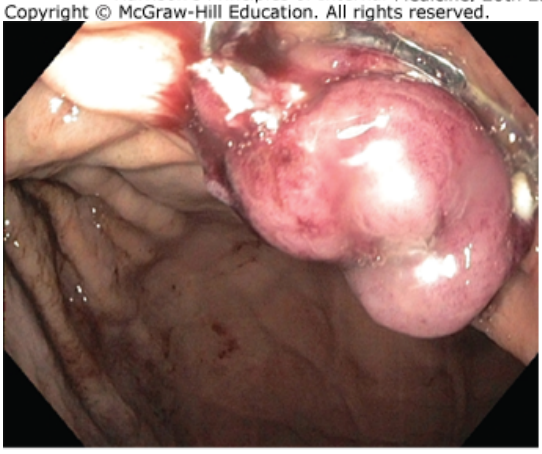
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

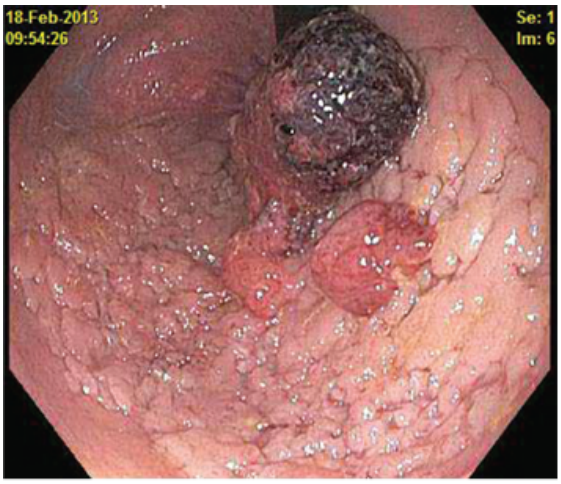
ENDOSCOPIC RESECTION AND CLOSURE TECHNIQUES

Endoscopic mucosal resection (EMR) (Video V5-4) and endoscopic submucosal dissection (ESD) (Fig. 315-24, Video V5-5) are the two commonly used techniques for the resection of benign and early-stage malignant gastrointestinal neoplasms. In addition to providing larger specimens for more accurate histopathological assessment and diagnosis, these techniques can be potentially curative for certain dysplastic lesions and focal intramucosal carcinomas involving the esophagus, stomach, and colon. Several devices are available for closure

of EMR and ESD defects as well as gastrointestinal fistulas and perforations. Endoscopic clips deployed through the working channel of an endoscope have been used for many years to treat bleeding lesions, and the development of larger over-the-scope clips has facilitated endoscopic closure of gastrointestinal fistulas and perforations not previously amenable to endoscopic therapy (Video V5-6). Endoscopic suturing can be used to close perforations and large defects (Fig. 315-25), anastomotic leaks, and fistulas. Other potential indications for endoscopic suturing include stent fixation to prevent migration (Fig. 315-26, Video V5-7) and endoscopic bariatric procedures. These technologies are likely to have an expanding role in patient care.

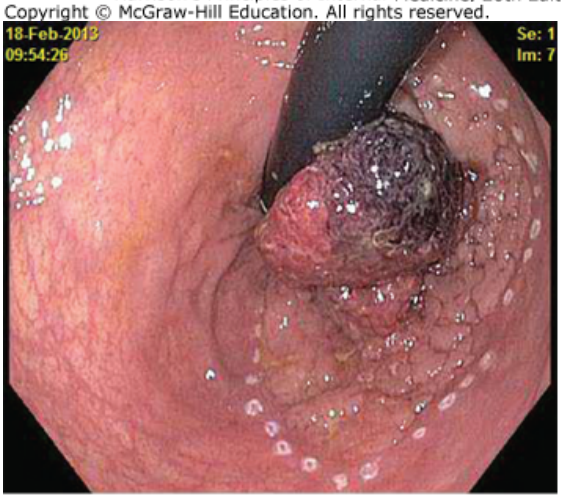
FIGURE 315-24

Endoscopic submucosal dissection (ESD). ***A.*** Large flat distal rectal adenoma with central lobulation. ***B.*** Marking the periphery of the lesion with coagulation dots. ***C.*** Rectal defect following ESD. ***D.*** Specimen resected en bloc.



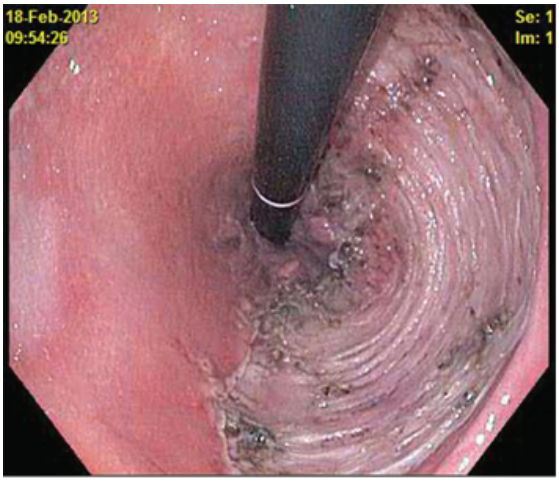
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

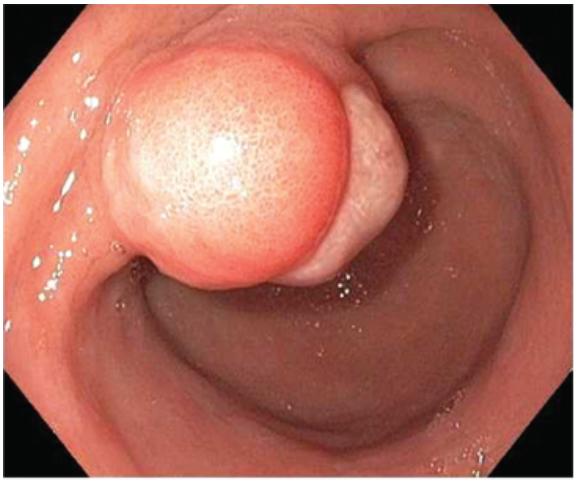


D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

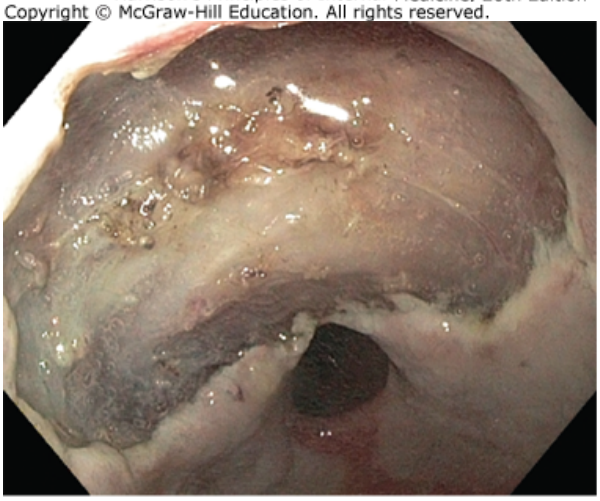
FIGURE 315-25

Closure of large defect using an endoscopic suturing device. ***A.*** Ulcerated inflammatory fibroid polyp in the antrum. ***B.*** Large defect following endoscopic submucosal dissection of the lesion. ***C.*** Closure of the defect using endoscopic sutures *(arrows)*. ***D.*** Resected specimen.



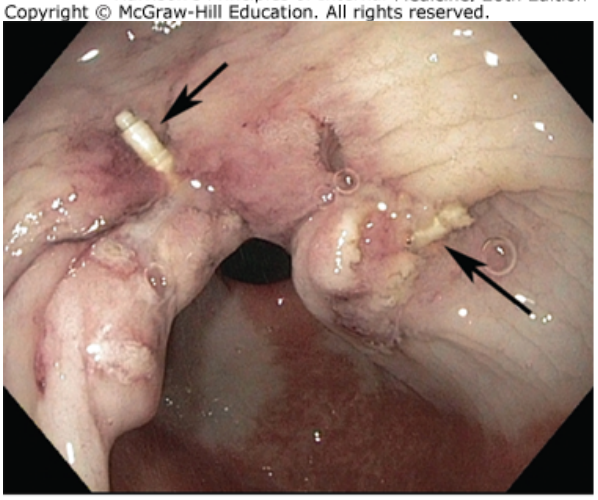
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

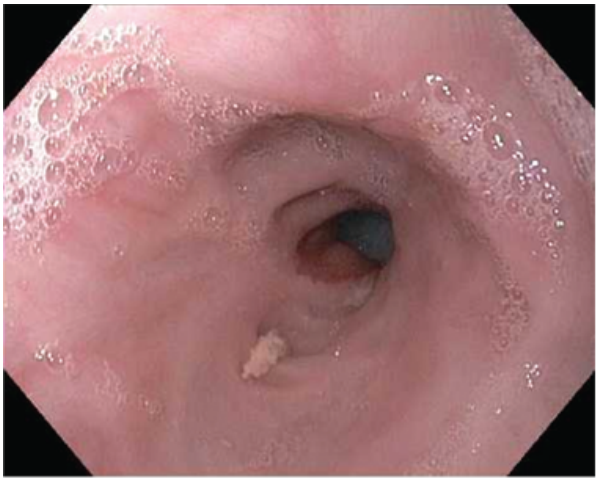


D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-26

Prevention of stent migration using endoscopic sutures. ***A.*** Esophagogastric anastomotic stricture refractory to balloon dilation. ***B.*** Temporary placement of a covered esophageal stent. ***C.*** Endoscopic suturing device to anchor the stent to the esophageal wall. ***D.*** Stent fixation with endoscopic sutures *(arrows)*.



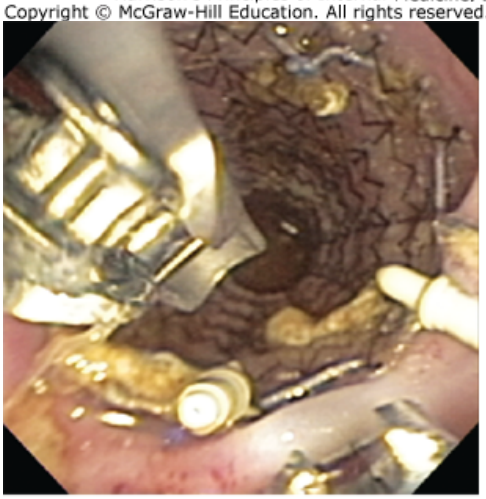
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



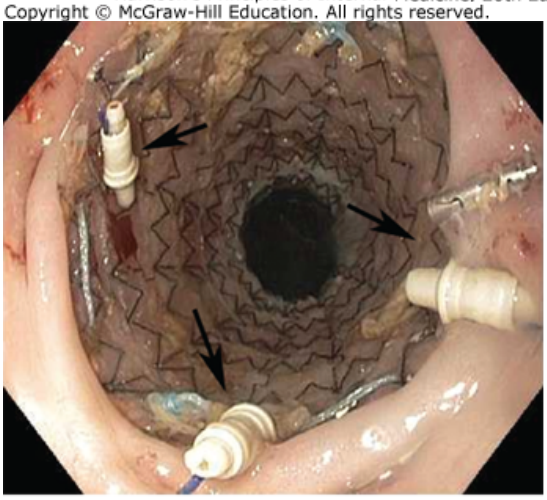
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

RISKS OF ENDOSCOPY

Medications used during conscious sedation may cause respiratory depression or allergic reactions. All endoscopic procedures carry some risk of bleeding and gastrointestinal perforation. The risk is small with diagnostic upper endoscopy, flexible sigmoidoscopy, and colonoscopy (<1:1000 procedures), but it ranges from 0.5 to 5% when therapeutic procedures such as polypectomy, EMR, ESD, control of hemorrhage, or stricture dilation are performed. The risk of adverse events for diagnostic EUS (without needle aspiration) is similar to that for diagnostic upper endoscopy.

Infectious complications are uncommon with most endoscopic procedures. Some procedures carry a higher incidence of postprocedure bacteremia, and prophylactic antibiotics may be indicated (Table 315-1). Management of antithrombotic agents prior to endoscopic procedures should take into account the procedural risk of hemorrhage, the agent, and the patient condition, as summarized in Table 315-2.

TABLE 315-1

Antibiotic Prophylaxis for Endoscopic Procedures

Patient Condition	Procedure Contemplated	Goal of Prophylaxis	Periprocedural Antibiotic Prophylaxis
All cardiac conditions	Any endoscopic procedure	Prevention of infective endocarditis	Not indicated
Bile duct obstruction in the absence of cholangitis	ERCP with complete drainage	Prevention of cholangitis	Not recommended
Bile duct obstruction in absence of cholangitis	ERCP with anticipated incomplete drainage (e.g., sclerosing cholangitis, hilar strictures)	Prevention of cholangitis	Recommended; continue antibiotics after the procedure
Sterile pancreatic fluid collection (e.g., pseudocyst, necrosis), which communicates with pancreatic duct	ERCP	Prevention of cyst infection	Recommended; continue antibiotics after the procedure
Sterile pancreatic fluid collection	Transmural drainage	Prevention of cyst infection	Recommended
Solid lesion along upper GI tract	EUS-FNA	Prevention of local infection	Not recommended[a]
Solid lesion along lower GI tract	EUS-FNA	Prevention of local infection	Not recommended[a]
Cystic lesions along GI tract (including mediastinum and pancreas)	EUS-FNA	Prevention of cyst infection	Recommended
All patients	Percutaneous endoscopic feeding tube placement	Prevention of peristomal infection	Recommended[b]
Cirrhosis with acute GI bleeding	Required for all such patients, regardless of endoscopic procedures	Prevention of infectious complications and reduction of mortality	Recommended, upon admission[c]
Continuous peritoneal dialysis	Lower GI tract endoscopy	Prevention of bacterial peritonitis	Recommended
Synthetic vascular graft and other nonvalvular cardiovascular devices	Any endoscopic procedure	Prevention of graft and device infection	Not recommended[d]
Prosthetic joints	Any endoscopic procedure	Prevention of septic arthritis	Not recommended[e]

[a]Low rates of bacteremia and local infection.

[b]Cefazolin or an antibiotic with equivalent coverage of oral and skin flora.

[c]Risk for bacterial infection associated with cirrhosis and GI bleeding is well established; ceftriaxone or a quinolone antibiotic recommended.

[d]No reported cases of infection associated with endoscopy.

[e]Very low risk of infection.

Abbreviations: ERCP, endoscopic retrograde cholangiopancreatography; EUS-FNA, endoscopic ultrasound–fine-needle aspiration; GI, gastrointestinal.

Source: Adapted from MA Kashab et al: Gastrointest Endosc 81:81, 2015; with permission from Elsevier.

TABLE 315-2

Management of Antithrombotic Drugs Prior to Endoscopic Procedures

Drug	Bleeding Risk of Procedure	Management	Interval between Last dose and Procedure	Comments
Warfarin	Low[a]	Continue	N/A	Ensure that INR is not supratherapeutic
	High[b]	Discontinue	3–7 days (usually 5), INR should be ≤1.5 for procedure	Consider bridging therapy with heparin;[c] usually safe to resume warfarin on the same or next day
Dabigatran, rivaroxaban, apixaban, edoxaban	Low[a]	Continue or hold morning dose on day of procedure	N/A	
Dabigatran	High[b]	Discontinue	2–3 days if GFR is ≥50 mL/min, 3–4 days if GFR is 30–49 mL/min	Bridging therapy not recommended; resume drug when bleeding risk is low
Rivaroxaban, apixaban, edoxaban	High[a]	Discontinue	2 days if GFR is ≥60 mL/min, 3 days if GFR is 30–59 mL/min, 4 days if GFR is <30 mL/min	Bridging therapy not recommended; resume drug when bleeding risk is low
Heparin	Low[a]	Continue	N/A	
	High[b]	Discontinue	4–6 h for unfractionated heparin	Skip one dose if using low-molecular-weight heparin
Aspirin	Any	Continue	N/A	Low-dose aspirin does not substantially increase the risk of endoscopic procedures
Aspirin with dipyridamole	Low[a]	Continue	N/A	
	High[b]	Discontinue	2–7 days	Consider continuing aspirin monotherapy
P2Y12 receptor antagonists (clopidogrel, prasugrel, ticlopidine, ticagrelor, cangrelor	Low[a]	Continue	N/A	
	High[b]	Coronary stent in place: discuss with cardiologist	5 days (clopidogrel or ticagrelor), 7 days (prasugrel), 10–4 days (ticlopidine)	Risk of stent thrombosis for 12 months after insertion of drug-eluting coronary stent or 1 month after insertion of bare metal coronary stent

Drug	Bleeding Risk of Procedure	Management	Interval between Last dose and Procedure	Comments
		No coronary stent: Discontinue, consider substituting to aspirin		

[a]Low-risk endoscopic procedures include EGD or colonoscopy with or without biopsy, EUS without FNA, ERCP with stent exchange.

[b]High-risk endoscopic procedures include EGD or colonoscopy with dilation, polypectomy, or thermal ablation; PEG; EUS with FNA; ERCP with sphincterotomy or pseudocyst drainage.

[c]Bridging therapy with low-molecular-weight heparin should be considered for patients discontinuing warfarin who are at high risk for thromboembolism, including those with (1) atrial fibrillation with a CHA2DS2-VASc score ≥2, mechanical valves, or history of stroke; (2) mechanical mitral valve; (3) mechanical aortic valve with other thromboembolic risk factors or older-generation mechanical aortic valve; (4) venous thromboembolism within the past 3 months.

Abbreviations: GFR, glomerular filtration rate; INR, international normalized ratio; N/A, not applicable.

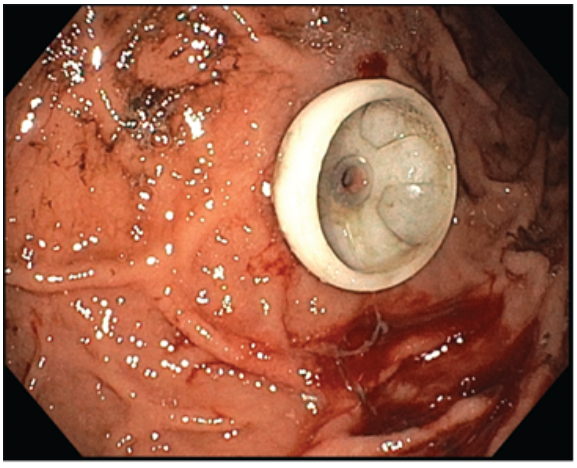
Source: Data from RD Acosta et al: Gastrointest Endosc 83:3, 2016; and AM Veitch et al: Gut 65:374, 2016.

ERCP carries additional risks. Pancreatitis occurs in ~5% of patients undergoing the procedure and in up to 30% of patients with sphincter of Oddi dysfunction. Young anicteric patients with normal ducts are at increased risk. Post-ERCP pancreatitis is usually mild and self-limited, but it may result in prolonged hospitalization, surgery, diabetes, or death when severe. Significant bleeding occurs after endoscopic sphincterotomy in ~1% of cases. Ascending cholangitis, pseudocyst infection, duodenal perforation, and abscess formation may occur as a result of ERCP.

Percutaneous gastrostomy tube placement during EGD is associated with a 10–15% incidence of adverse events, most often wound infections. Fasciitis, pneumonia, bleeding (Fig. 315-27), buried bumper syndrome, and colonic injury may result from gastrostomy tube placement.

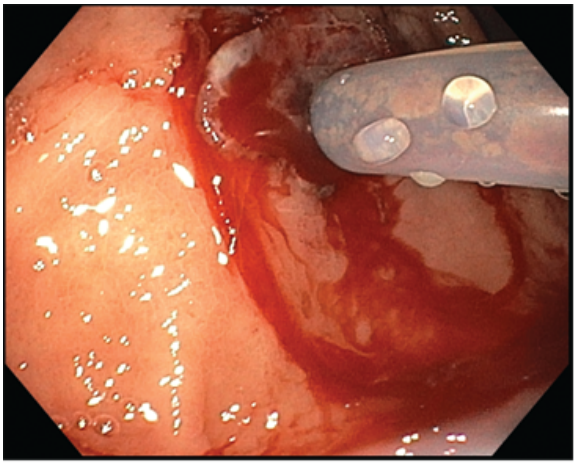
FIGURE 315-27

Bleeding from percutaneous endoscopic gastrostomy (PEG) tube placement. ***A.*** Patient with melena from a recently placed PEG tube. ***B.*** Loosening of the internal disk bumper of the PEG tube revealed active bleeding from within the PEG tract.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

URGENT ENDOSCOPY

ACUTE GASTROINTESTINAL HEMORRHAGE

Endoscopy is the primary diagnostic and therapeutic technique for patients with acute gastrointestinal hemorrhage. Although gastrointestinal bleeding stops spontaneously in most cases, some patients will have persistent or recurrent hemorrhage that may be life-threatening. Clinical predictors of rebleeding help identify patients most likely to benefit from urgent endoscopy and endoscopic, angiographic, or surgical hemostasis.

Initial Evaluation

The initial evaluation of the bleeding patient focuses on the severity of hemorrhage as reflected by the presence of supine hypotension or tachycardia, postural vital sign changes, and the frequency of hematemesis or melena. Decreases in hematocrit and hemoglobin lag behind the clinical course and are not reliable gauges of the magnitude of acute bleeding. Nasogastric tube aspiration and lavage can also be used to judge the severity of bleeding, but these are no longer routinely performed for this purpose. The bedside initial evaluation, completed well before the bleeding source is confidently identified, guides immediate supportive care of the patient, triage to the ward or intensive care unit, and timing of endoscopy. The severity of the initial hemorrhage is the most important indication for urgent endoscopy, since a large initial bleed increases the likelihood of ongoing or recurrent bleeding. Patients with resting hypotension or orthostatic change in vital signs, repeated hematemesis, bloody nasogastric aspirate that does not clear with large volume lavage, or those requiring blood transfusions should be considered for urgent endoscopy. In addition, patients with cirrhosis, coagulopathy, respiratory or renal failure, and those over 70 years of age are more likely to have significant rebleeding and to benefit from prompt evaluation and treatment.

Bedside evaluation also suggests an upper or lower gastrointestinal source of bleeding in most patients. Over 90% of patients with melena are bleeding proximal to the ligament of Treitz, and ~85% of patients with hematochezia are bleeding from the colon. Melena can result from bleeding in the small bowel or right colon, especially in older patients with slow colonic transit. Conversely, some patients with massive hematochezia may be bleeding from an upper gastrointestinal source, such as a gastric Dieulafoy lesion or duodenal ulcer, with rapid intestinal transit. Early upper endoscopy should be considered in such patients.

Endoscopy should be performed after the patient has been resuscitated with intravenous fluids and transfusions, as necessary. Marked coagulopathy or thrombocytopenia is usually treated before endoscopy, since correction of these abnormalities may lead to resolution of bleeding, and techniques for endoscopic hemostasis are limited in such patients. Metabolic derangements should also be addressed. Tracheal intubation for airway protection should be considered before upper endoscopy in patients with repeated recent hematemesis, encephalopathy and suspected variceal hemorrhage. A single dose of erythromycin (3–4 mg/kg or 250 mg) administered intravenously 30–90 min prior to upper endoscopy increases gastric emptying and may clear blood and clots from the stomach to improve endoscopic visualization.

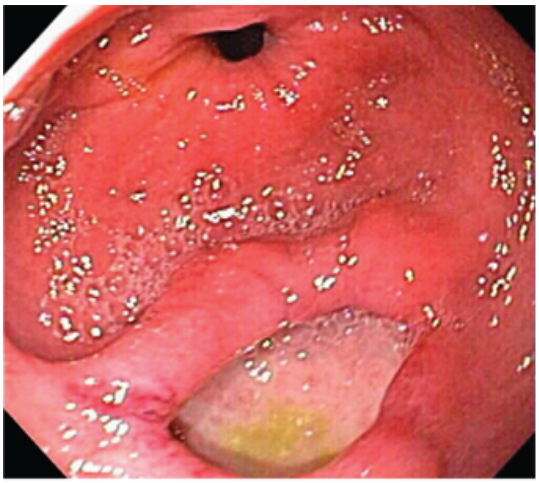
Most patients with hematochezia who are otherwise stable can undergo semielective colonoscopy. Controlled trials have not shown a benefit to urgent colonoscopy in patients hospitalized with hematochezia, although patients with massive or recurrent large-volume episodes of hematochezia should undergo urgent colonoscopy after a rapid colonic purge with a polyethylene glycol solution. Colonoscopy has a higher diagnostic yield than radionuclide bleeding scans or angiography in lower gastrointestinal bleeding, and endoscopic therapy can be applied in some cases. Urgent colonoscopy can be hindered by poor visualization due to persistent vigorous bleeding with recurrent hemodynamic instability, and other techniques (such as angiography or even emergent subtotal colectomy) must be employed. In such patients, massive bleeding originating from an upper gastrointestinal source should also be considered and excluded promptly by upper endoscopy. The anal and rectal mucosa should also be visualized endoscopically early in the course of massive rectal bleeding, as bleeding lesions in or close to the anal canal may be identified that are amenable to endoscopic or surgical transanal hemostatic techniques.

Peptic Ulcer

The endoscopic appearance of peptic ulcers provides useful prognostic information and guides the need for endoscopic therapy in patients with acute hemorrhage **(Fig. 315-28)**. A clean-based ulcer is associated with a low risk (3–5%) of rebleeding; patients with melena and a clean-based ulcer are often discharged home from the emergency room or endoscopy suite if they are young, reliable, and otherwise healthy. Flat pigmented spots and adherent clots covering the ulcer base have a 10% and 20% risk of rebleeding, respectively. Endoscopic therapy may be considered for an ulcer with an adherent clot. When a fibrin plug is seen protruding from a vessel wall in the base of an ulcer (so-called sentinel clot or visible vessel), the risk of rebleeding from the ulcer is 40%. This finding generally leads to endoscopic therapy to decrease the rebleeding rate. When active spurting from an ulcer is seen, there is a 90% risk of ongoing bleeding without therapy.

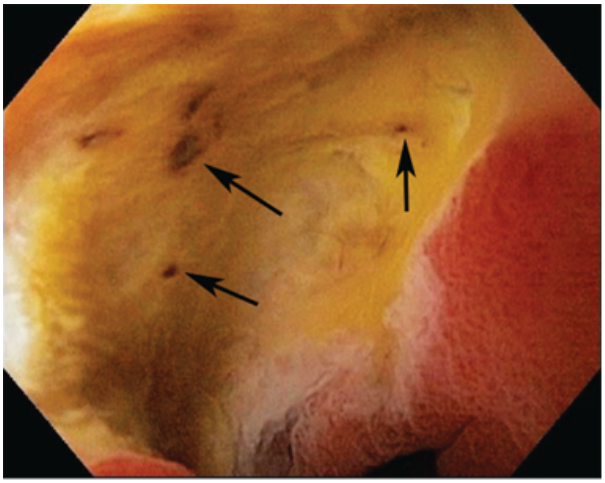
FIGURE 315-28

Stigmata of hemorrhage in peptic ulcers. ***A.*** Gastric antral ulcer with a clean base. ***B.*** Duodenal ulcer with flat pigmented spots *(arrows).* ***C.*** Duodenal ulcer with a dense adherent clot. ***D.*** Gastric ulcer with a pigmented protuberance/visible vessel. ***E.*** Duodenal ulcer with active spurting *(arrow).*



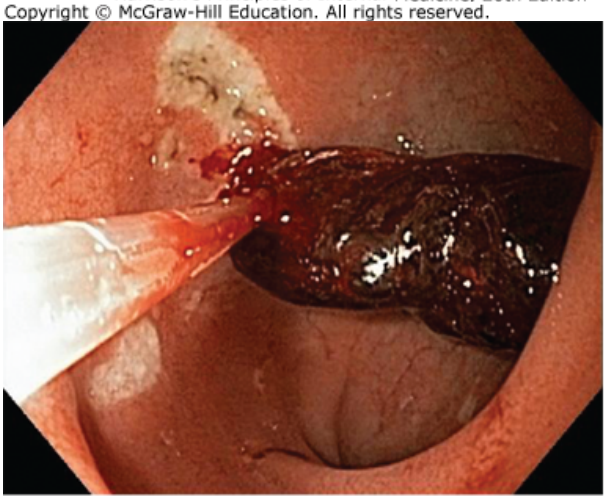
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



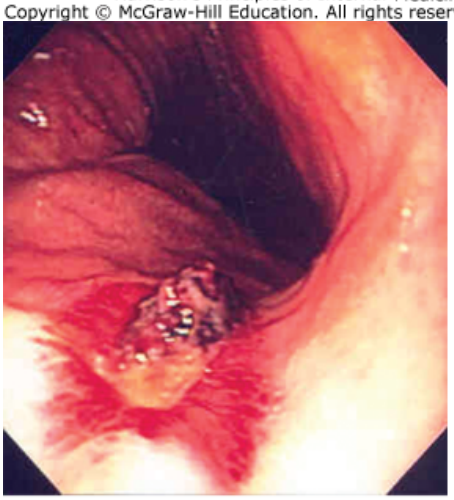
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



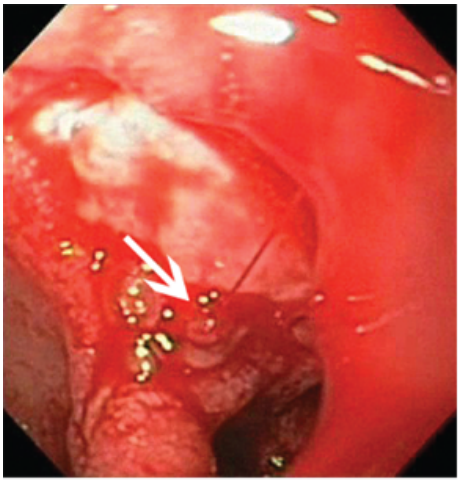
C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



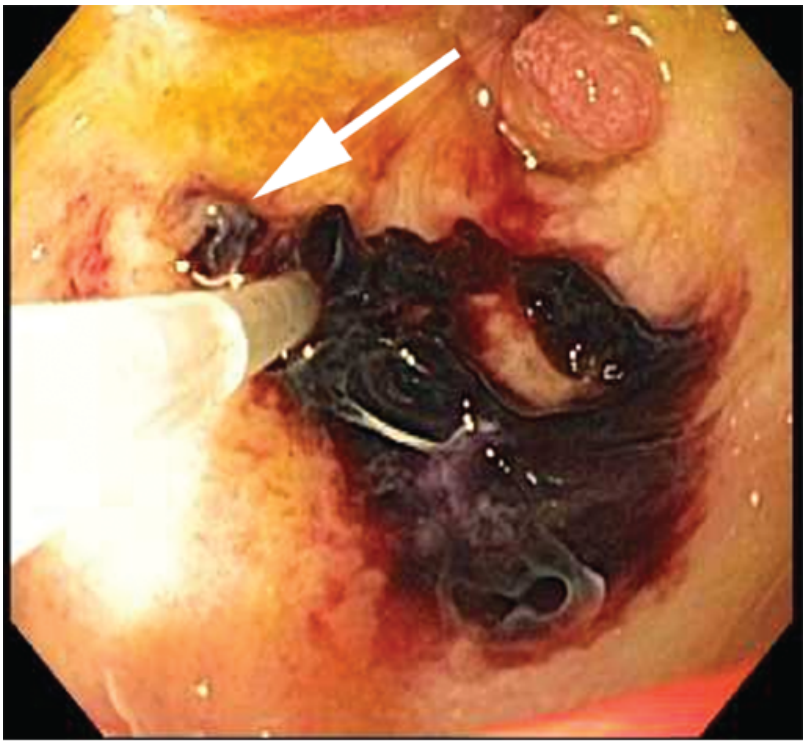
E

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Endoscopic therapy of ulcers with high-risk stigmata typically lowers the rebleeding rate to 5–10%. Several hemostatic techniques are available, including injection of epinephrine or a sclerosant into and around the vessel **(Fig. 315-29)**, "coaptive coagulation" of the vessel in the base of the ulcer using a thermal probe that is pressed against the site of bleeding **(Fig. 315-30)**, placement of through-the-scope hemoclips **(Fig. 315-31)** or an over-the-scope clip **(Fig. 315-32)**, or a combination of these modalities **(Video V5-8)**. In conjunction with endoscopic therapy, the administration of a proton pump inhibitor decreases the risk of rebleeding and improves patient outcome.

FIGURE 315-29

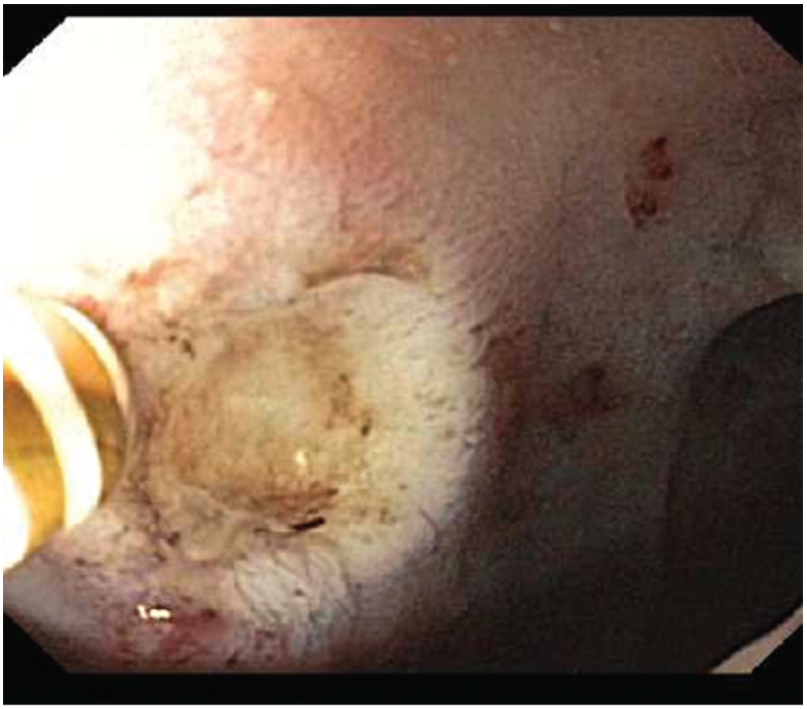
Epinephrine injection into a duodenal ulcer with visible vessel *(arrow)* and adherent clot.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-30

Treated duodenal ulcer by contact coagulation with a thermal probe.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

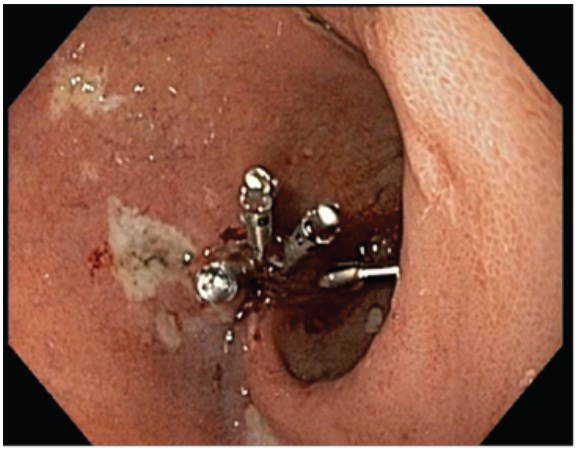
FIGURE 315-31

Ulcer hemostasis using through-the-scope clips. ***A.*** Superficial duodenal ulcer with visible vessel *(arrow).* ***B.*** Hemostasis secured following placement of multiple clips.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

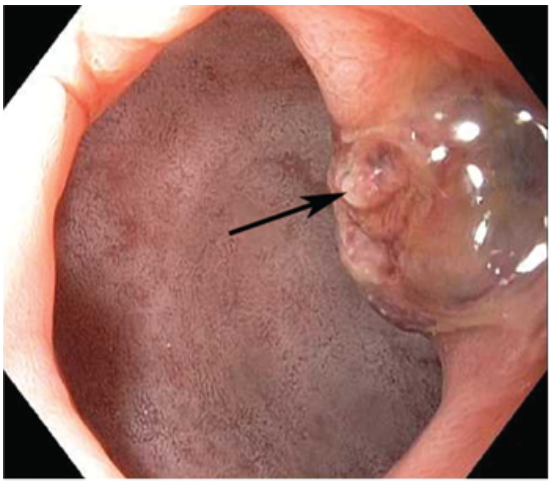


B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

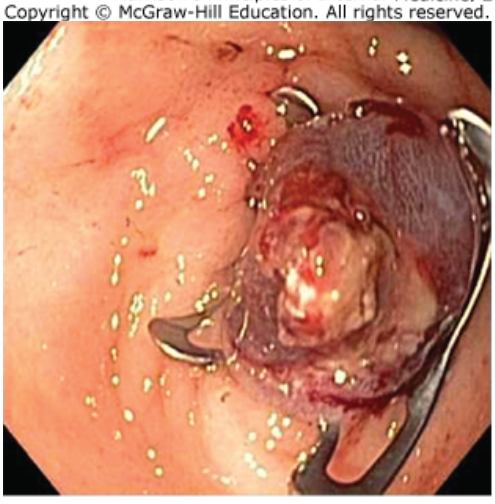
FIGURE 315-32

Endoscopic hemostasis of ulcer bleeding. ***A.*** Pyloric channel ulcer with visible vessel *(arrow).* ***B.*** Ulcer hemostasis with placement of an over-the-scope clip.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

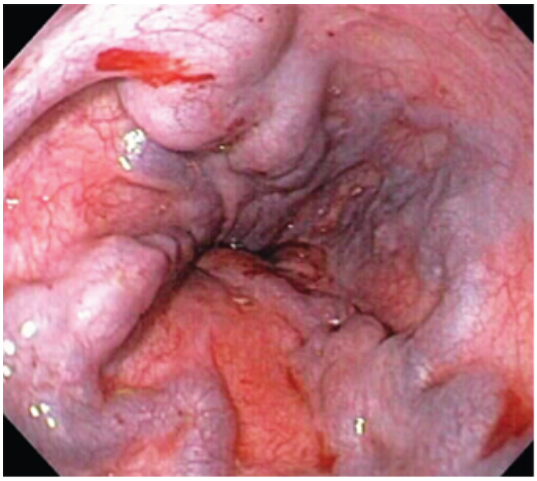
Varices

Two complementary strategies guide therapy of bleeding varices: local treatment of the bleeding varices and treatment of the underlying portal hypertension. Local therapies, including endoscopic variceal band ligation, endoscopic variceal sclerotherapy (EVS), stent placement and balloon tamponade with a Sengstaken-Blakemore tube, effectively control acute hemorrhage in most patients, although therapies that decrease portal pressure (pharmacologic treatment, surgical shunts, or radiologically placed intrahepatic portosystemic shunts) also play an important role.

Endoscopic variceal ligation (EVL) is indicated for the prevention of a first bleed (primary prophylaxis) from large esophageal varices **(Fig. 315-33)**, particularly in patients in whom nonselective beta blockers are contraindicated or not tolerated. EVL is also the preferred endoscopic therapy for control of active esophageal variceal bleeding and for subsequent eradication of esophageal varices (secondary prophylaxis). During EVL a varix is suctioned into a cap fitted on the end of the endoscope, and a rubber band is released from the cap, ligating the varix **(Fig. 315-34, Video V5-9)**. EVL controls acute hemorrhage in up to 90% of patients. Complications of EVL, such as postligation ulcer bleeding and esophageal stenosis, are uncommon. EVS involves the injection of a sclerosing, thrombogenic solution into or next to esophageal varices. EVS also controls acute hemorrhage in most patients, but it is generally used as salvage therapy when band ligation fails because of its higher complication rate. Bleeding from large gastric fundic varices **(Fig. 315-35)** is best treated with endoscopic cyanoacrylate ("glue") injection **(Video V5-10)**, since EVL or EVS of these varices is associated with a high rebleeding rate. Complications of cyanoacrylate injection include infection and glue embolization to other organs, such as the lungs, brain, and spleen.

FIGURE 315-33

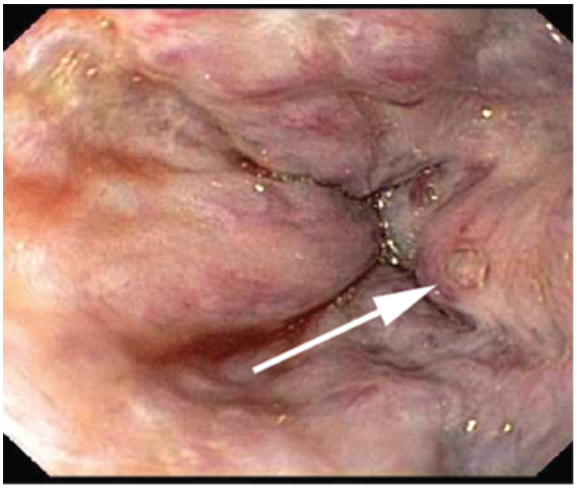
Esophageal varices.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

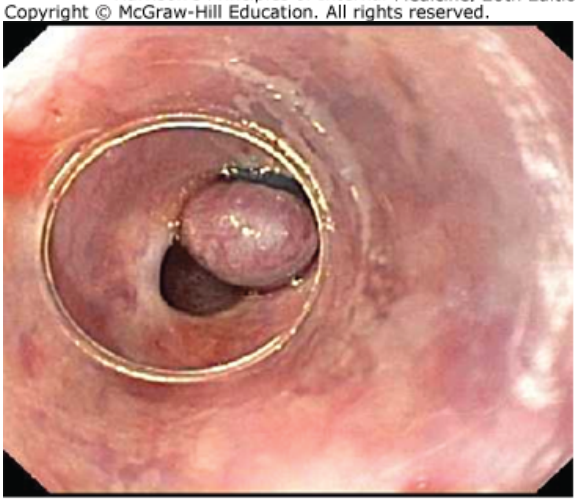
FIGURE 315-34

Endoscopic band ligation of esophageal varices. ***A.*** Large esophageal varices with stigmata of recent bleeding characterized by a fibrin plug *(arrow).* ***B.*** Band ligation of varices.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

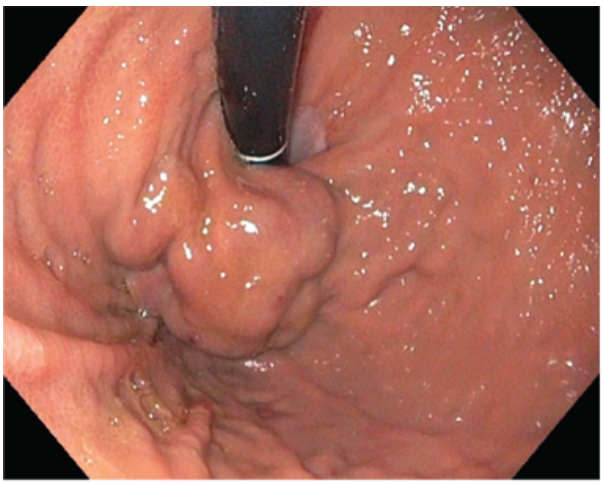


B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

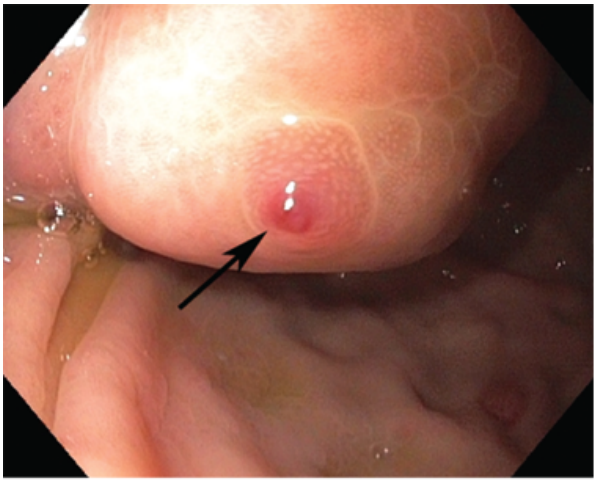
FIGURE 315-35

Gastric varices. ***A.*** Large gastric fundal varices. ***B.*** Stigmata of recent bleeding from the same gastric varices *(arrow)*.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

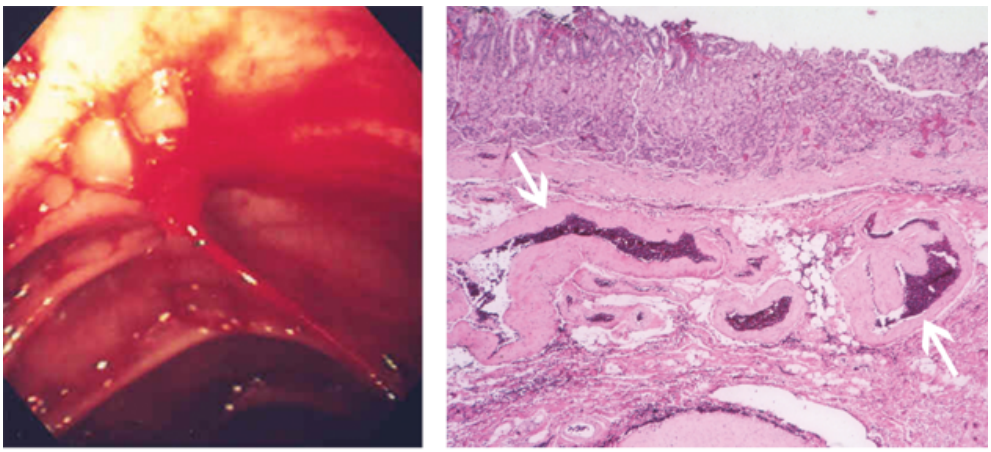
After treatment of the acute hemorrhage, an elective course of endoscopic therapy can be undertaken with the goal of eradicating esophageal varices and preventing rebleeding months to years later. However, this chronic therapy is less successful, preventing long-term rebleeding in ~50% of patients. Pharmacologic therapies that decrease portal pressure have similar efficacy, and the two modalities are usually combined.

Dieulafoy's Lesion

This lesion, also called persistent caliber artery, is a large-caliber arteriole that runs immediately beneath the gastrointestinal mucosa and bleeds through a focal mucosal erosion (Fig. 315-36). Dieulafoy's lesion is seen most commonly on the lesser curvature of the proximal stomach, causes impressive arterial hemorrhage, and may be difficult to diagnose when not actively bleeding; it is often recognized only after repeated endoscopy for recurrent bleeding. Endoscopic therapy, such as thermal coagulation, band ligation, or endoscopic suturing, is typically effective for control of bleeding and sealing of the underlying vessel once the lesion has been identified (Video V5-11). Rescue therapies, such as angiographic embolization or surgical oversewing, are considered in situations where endoscopic therapy has failed.

FIGURE 315-36

Dieulafoy's lesion. ***A.*** Actively spurting jejunal Dieulafoy's lesion. There is no underlying mucosal lesion. ***B.*** Histology of a gastric Dieulafoy's lesion. A persistent caliber artery *(arrows)* is present in the gastric submucosa, immediately beneath the mucosa.



A *B*

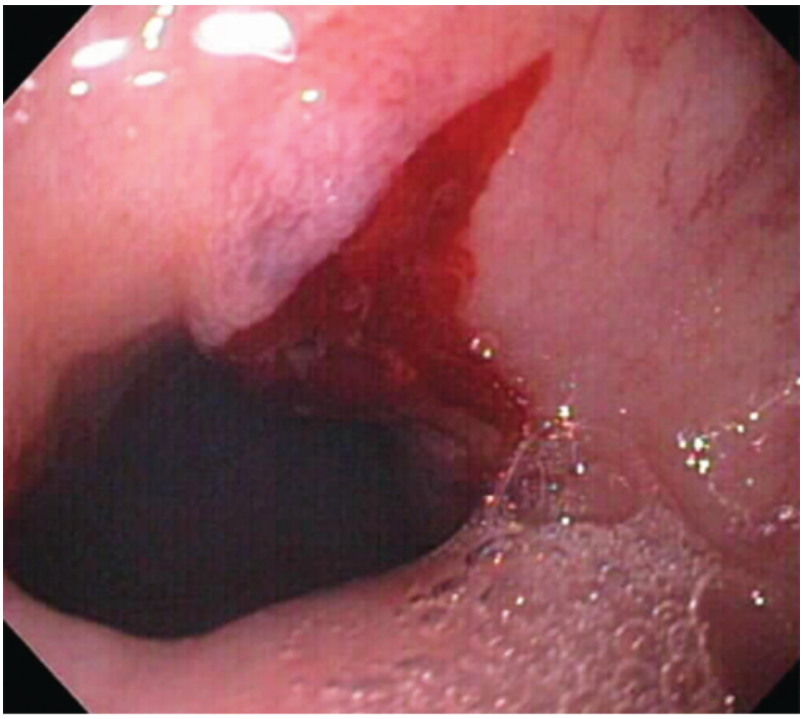
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Mallory-Weiss Tear

A Mallory-Weiss tear is a linear mucosal rent near or across the gastroesophageal junction that is often associated with retching or vomiting **(Fig. 315-37)**. When the tear disrupts a submucosal arteriole, brisk hemorrhage may result. Endoscopy is the best method for diagnosis, and an actively bleeding tear can be treated endoscopically with epinephrine injection, coaptive coagulation, band ligation, or hemoclips **(Video V5-12)**. Unlike peptic ulcer, a Mallory-Weiss tear with a nonbleeding sentinel clot in its base rarely rebleeds and thus does not necessitate endoscopic therapy.

FIGURE 315-37

Mallory-Weiss tear at the gastroesophageal junction.



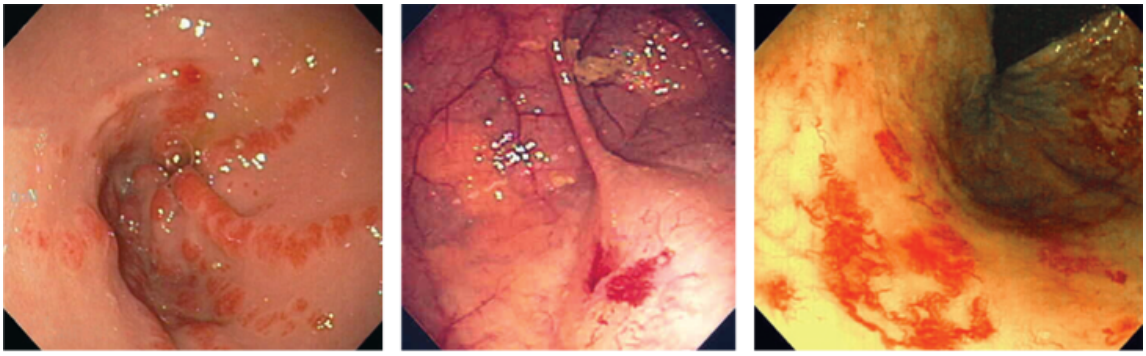
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Vascular Ectasias

Vascular ectasias are flat mucosal vascular anomalies that are best diagnosed by endoscopy. They usually cause slow intestinal blood loss and occur either in a sporadic fashion or in a well-defined pattern of distribution (e.g., gastric antral vascular ectasia [GAVE] or "watermelon stomach") **(Fig. 315-38)**. Cecal vascular ectasias, GAVE, and radiation-induced rectal ectasias are often responsive to local endoscopic ablative therapy, such as argon plasma coagulation **(Video V5-13)**. Patients with diffuse small bowel vascular ectasias (associated with chronic renal failure and with hereditary hemorrhagic telangiectasia) may continue to bleed despite endoscopic treatment of easily accessible lesions by conventional endoscopy. These patients may benefit from device-assisted enteroscopy with endoscopic therapy or pharmacologic treatment with octreotide or estrogen/progesterone.

FIGURE 315-38

Gastrointestinal vascular ectasias. ***A.*** Gastric antral vascular ectasia ("watermelon stomach") characterized by stripes of prominent flat or raised vascular ectasias. ***B.*** Cecal vascular ectasias. ***C.*** Radiation-induced vascular ectasias of the rectum in a patient previously treated for prostate cancer.

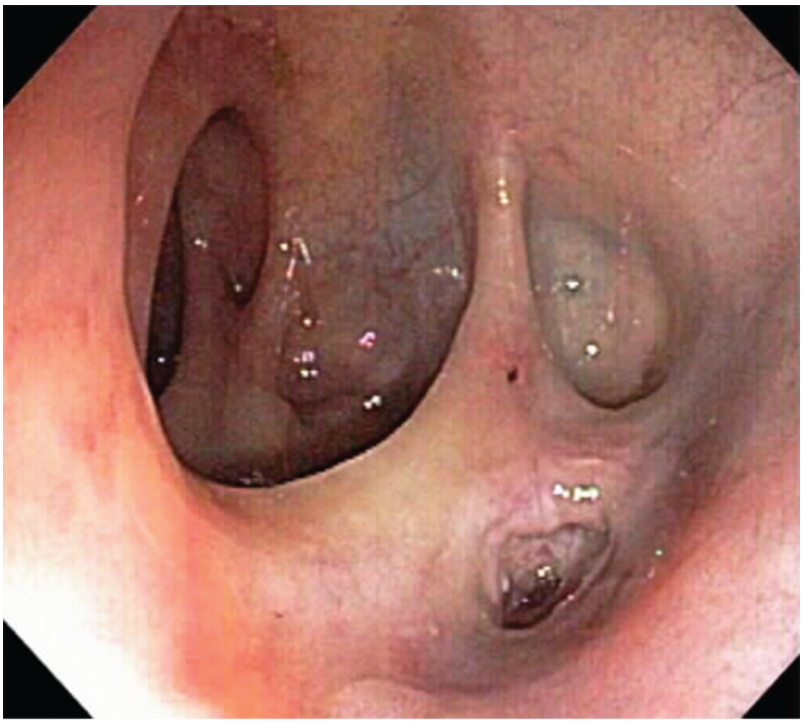


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Colonic Diverticula

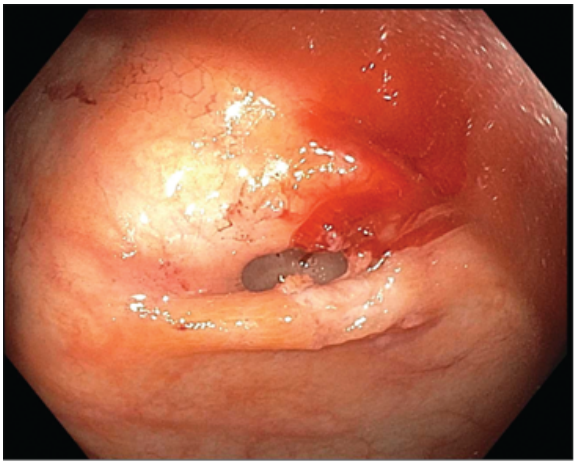
Diverticula form where nutrient arteries penetrate the muscular wall of the colon en route to the colonic mucosa (Fig. 315-39). The artery found in the base of a diverticulum may bleed, causing painless and impressive hematochezia. Colonoscopy is indicated in patients with hematochezia and suspected diverticular hemorrhage, since other causes of bleeding (such as vascular ectasias, colitis, and colon cancer) must be excluded. In addition an actively bleeding diverticulum may be seen and treated during colonoscopy (Fig. 315-40, Video V5-14).

FIGURE 315-39
Colonic diverticula.



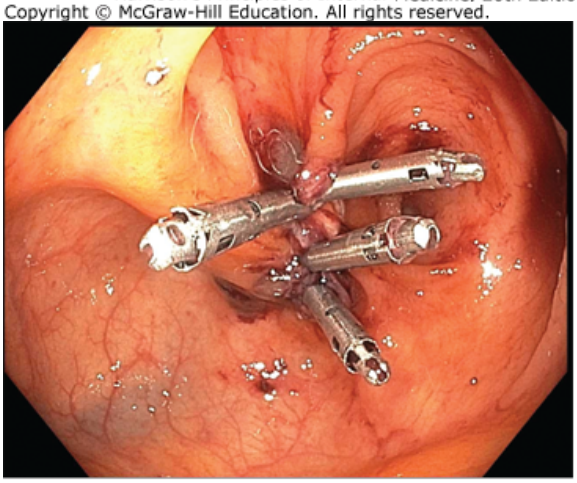
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-40
Diverticular hemorrhage. ***A.*** Actively bleeding sigmoid diverticulum. ***B.*** Hemostasis achieved using endoscopic clips.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GASTROINTESTINAL OBSTRUCTION AND PSEUDOOBSTRUCTION

Endoscopy is useful for evaluation and treatment of some forms of gastrointestinal obstruction. An important exception is small-bowel obstruction due to surgical adhesions, which is generally not diagnosed or treated endoscopically. Esophageal, gastroduodenal, and colonic obstruction or pseudoobstruction can all be diagnosed and often managed endoscopically.

Acute Esophageal Obstruction

Esophageal obstruction by impacted food (Fig. 315-41) or an ingested foreign body (Fig. 315-42) is a potentially life-threatening event and represents an endoscopic emergency. Left untreated, the patient may develop esophageal ulceration, ischemia, and perforation. Patients with persistent esophageal obstruction often have hypersalivation and are usually unable to swallow water. Sips of a carbonated beverage, sublingual nifedipine or nitrates, or intravenous glucagon may resolve an esophageal food impaction, but in most patients an underlying web, ring, or stricture is present and endoscopic removal of the obstructing food bolus is necessary. Endoscopy is generally the best initial test in such patients since endoscopic removal of the obstructing material is usually possible, and the presence of an underlying esophageal pathology can often be determined. Radiographs of the chest and neck should be considered before endoscopy in patients with fever, obstruction for ≥24 h, or ingestion of a sharp object, such as a fishbone. Radiographic contrast studies interfere with subsequent endoscopy and are not advisable in most patients with a clinical picture of esophageal obstruction.

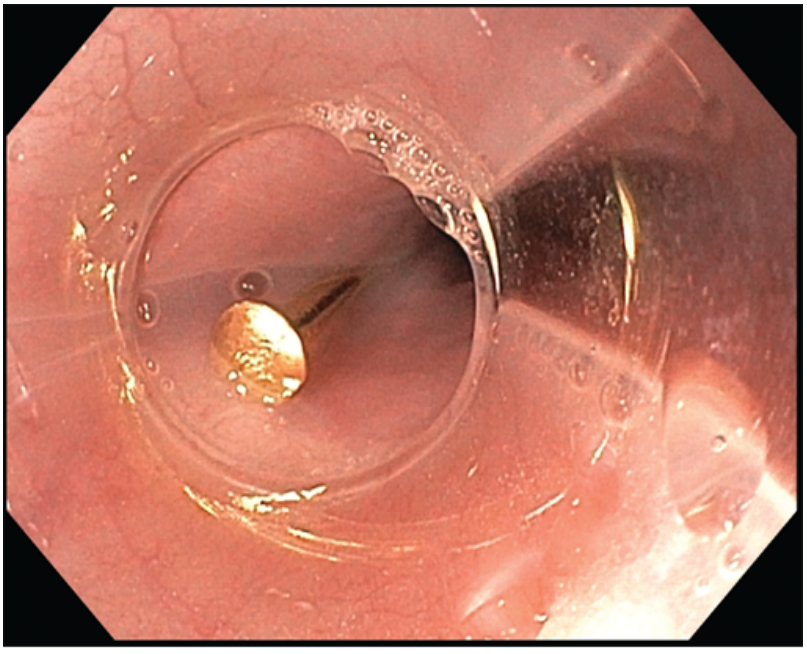
FIGURE 315-41
Esophageal food (meat) impaction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-42

Impacted nail in the esophagus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

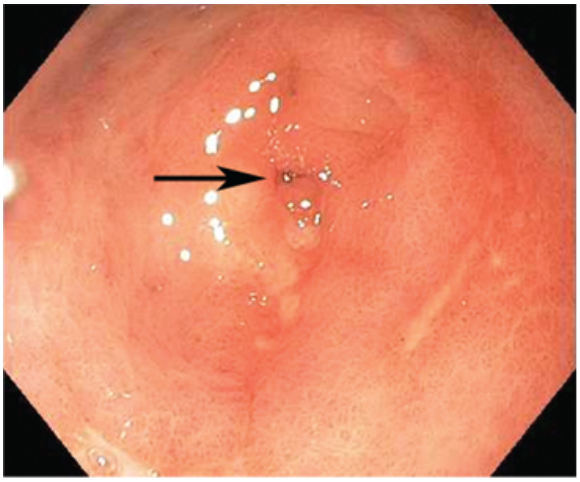
Gastric Outlet Obstruction

Obstruction of the gastric outlet is commonly caused by gastric, duodenal, or pancreatic malignancy, or chronic peptic ulceration with stenosis of the pylorus **(Fig. 315-43)**. Patients vomit partially digested food many hours after eating. Gastric decompression with a nasogastric tube and subsequent lavage for removal of retained material is the first step in treatment. The diagnosis can then be confirmed with a saline load test, if desired. Endoscopy is useful for diagnosis and treatment. Patients with benign pyloric stenosis may be treated with endoscopic balloon dilation of the pylorus, and a course of endoscopic dilation results in long-term relief of symptoms in ~50% of patients. Removable, fully covered lumen-apposing metal stents (LAMS) may also be used to treat benign pyloric stenosis **(Video V5-15)**. Malignant

gastric outlet obstruction can be relieved with endoscopically placed expandable stents in patients with inoperable malignancy (**Video V5-16**).

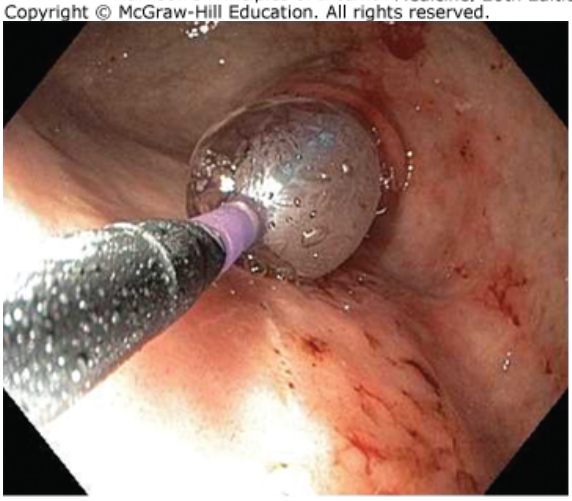
FIGURE 315-43

Gastric outlet obstruction due to pyloric stenosis. ***A.*** Nonsteroidal anti-inflammatory agent-induced ulcer disease with severe stenosis of the pylorus *(arrow).* ***B.*** Balloon dilation of the stenosis. ***C.*** Appearance of pyloric ring post dilation.



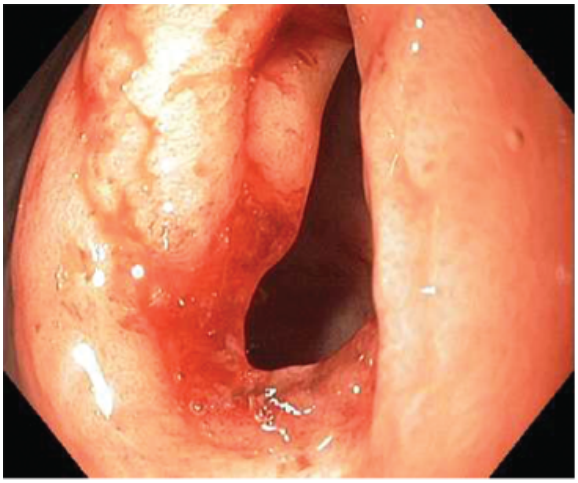
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Colonic Obstruction and Pseudoobstruction

These conditions both present with abdominal distention and discomfort, tympany, and a dilated colon on plain abdominal radiography. The radiographic appearance may be characteristic of a particular condition, such as sigmoid volvulus **(Fig. 315-44)**. Both obstruction and pseudoobstruction may lead to colonic perforation if left untreated. Acute colonic pseudoobstruction is a form of colonic ileus that is usually attributable to electrolyte disorders, narcotic and anticholinergic medications, immobility (as after surgery), or retroperitoneal hemorrhage or mass. Multiple causative factors are often present. Colonoscopy, water-soluble contrast enema, or CT may be used to assess for an obstructing lesion and differentiate obstruction from pseudoobstruction. One of these diagnostic studies should be strongly considered if the patient does not have clear risk factors for pseudoobstruction, if radiographs do not show air in the rectum, or if the patient fails to improve when underlying causes of pseudoobstruction have been addressed. The risk of cecal perforation in pseudoobstruction rises when the cecal diameter exceeds 12 cm, and decompression of the colon may be achieved using intravenous neostigmine or via colonoscopic decompression **(Fig. 315-45)**. Most patients should receive a trial of conservative therapy (with correction of electrolyte disorders, removal of offending medications, and increased mobilization) before undergoing an invasive decompressive procedure for colonic pseudoobstruction.

FIGURE 315-44

Sigmoid volvulus with the characteristic radiologic appearance of a “bent inner tube.”



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-45

Acute colonic pseudoobstruction. ***A.*** Acute colonic dilation occurring in a patient soon after knee surgery. ***B.*** Colonoscopic placement of decompression tube with marked improvement in colonic dilation.



A



B

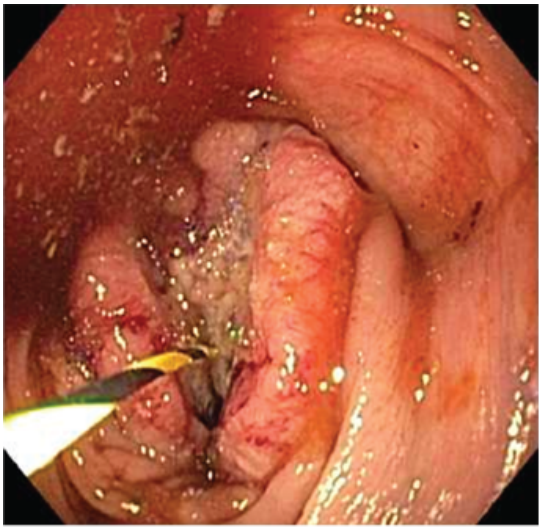
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Colonic obstruction is an indication for urgent intervention. In the past, emergent diverting colostomy was usually performed with a subsequent second operation after bowel preparation to treat the underlying cause of obstruction. Colonoscopic placement of an

expandable stent is an alternative treatment option that can relieve malignant colonic obstruction without emergency surgery and permit bowel preparation for an elective one-stage operation (Fig. 315-46, Video V5-17).

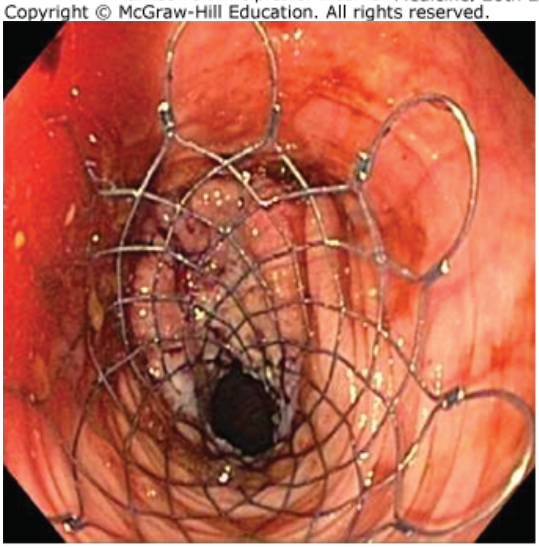
FIGURE 315-46

Obstructing colonic carcinoma. ***A.*** Colonic adenocarcinoma causing marked luminal narrowing of the distal transverse colon. ***B.*** Endoscopic placement of a self-expandable metal stent. ***C.*** Radiograph of expanded stent across the obstructing tumor with a residual waist *(arrow)*.



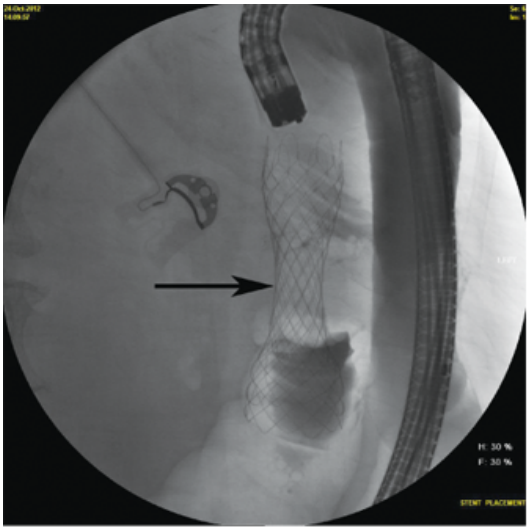
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ACUTE BILIARY OBSTRUCTION

The steady, severe pain that occurs when a gallstone acutely obstructs the common bile duct often brings patients to a hospital. The diagnosis of a ductal stone is suspected when the patient is jaundiced or when serum liver tests or pancreatic enzyme levels are elevated; it is confirmed by EUS, magnetic resonance cholangiography (MRCP), or direct cholangiography (performed endoscopically, percutaneously, or during surgery). ERCP is the primary means of treating common bile duct stones (Figs. 315-15 and 315-16), although they can also be removed by laparoscopic bile duct exploration at the time of cholecystectomy. Radiologic percutaneous biliary drainage may be required in some cases.

Bile Duct Imaging

While transabdominal ultrasound diagnoses only a minority of bile duct stones, MRCP and EUS are >90% accurate and have an important role in diagnosis. Examples of these modalities are shown in **Fig. 315-47**.

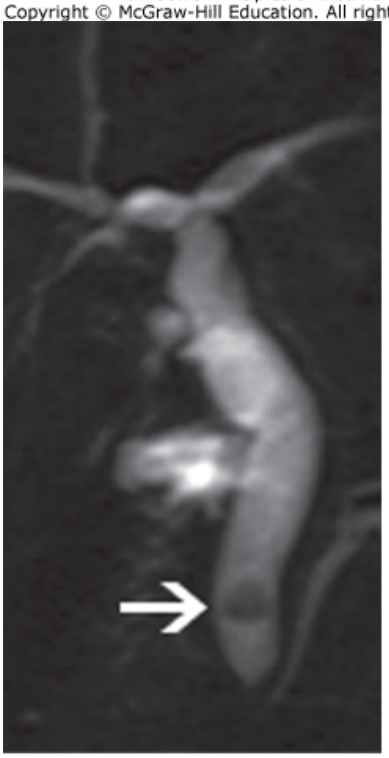
FIGURE 315-47

Methods of bile duct imaging. Arrows mark bile duct stones. ***A.*** Endoscopic ultrasound (EUS). ***B.*** Magnetic resonance cholangiopancreatography (MRCP). ***C.*** Helical computed tomography (CT).



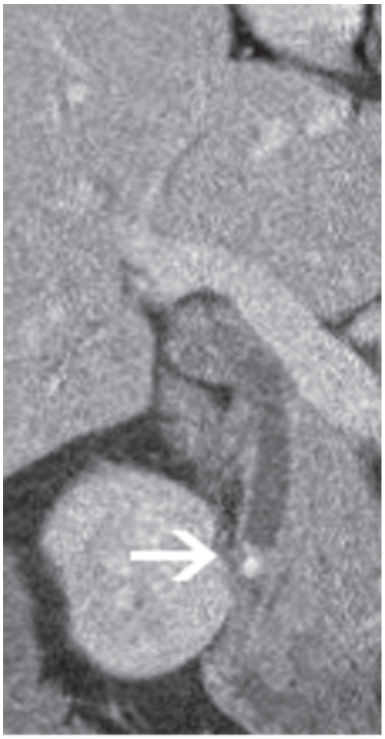
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

If the suspicion for a bile duct stone is high and urgent treatment is required (as in a patient with obstructive jaundice and biliary sepsis), ERCP is the procedure of choice since it remains the gold standard for diagnosis and allows for immediate treatment (Video V5-18). If a persistent bile duct stone is relatively unlikely (as in a patient with gallstone pancreatitis), ERCP may be supplanted by less invasive imaging techniques, such as EUS, MRCP, or intraoperative cholangiography performed during cholecystectomy, sparing some patients the risk and discomfort of ERCP.

Ascending Cholangitis

Charcot's triad of jaundice, abdominal pain, and fever is present in ~70% of patients with ascending cholangitis and biliary sepsis. These patients are managed initially with fluid resuscitation and intravenous antibiotics. Abdominal ultrasound is often performed to assess for gallbladder stones and bile duct dilation. However, the bile duct may not be dilated early in the course of acute biliary obstruction. Medical management usually improves the patient's clinical status, providing a window of ~24 h during which biliary drainage should be established, typically by ERCP. Undue delay can result in recrudescence of overt sepsis and increased morbidity and mortality rates. In addition to Charcot's triad, the additional presence of shock and confusion (Reynolds's pentad) is associated with a high mortality rate and should prompt urgent intervention to restore biliary drainage.

Gallstone Pancreatitis

Gallstones may cause acute pancreatitis as they pass through the ampulla of Vater. The occurrence of gallstone pancreatitis usually implies passage of a stone into the duodenum, and only ~20% of patients harbor a persistent stone in the ampulla or the common bile duct. Retained stones are more common in patients with jaundice, rising serum liver tests following hospitalization, severe pancreatitis, or superimposed ascending cholangitis.

Urgent ERCP decreases the morbidity rate of gallstone pancreatitis in a subset of patients with retained bile duct stones. It is unclear whether the benefit of ERCP is mainly attributable to treatment and prevention of ascending cholangitis or to relief of pancreatic ductal obstruction. ERCP is warranted early in the course of gallstone pancreatitis if ascending cholangitis is suspected, especially in a jaundiced patient. Urgent ERCP may also benefit patients predicted to have severe pancreatitis using a clinical index of severity, such as the Glasgow or Ranson score. Since the benefit of ERCP is limited to patients with a retained bile duct stone, a strategy of initial MRCP or EUS for diagnosis decreases the utilization of ERCP in gallstone pancreatitis and improves clinical outcomes by limiting the occurrence of ERCP-related adverse events.

ELECTIVE ENDOSCOPY

DYSPEPSIA

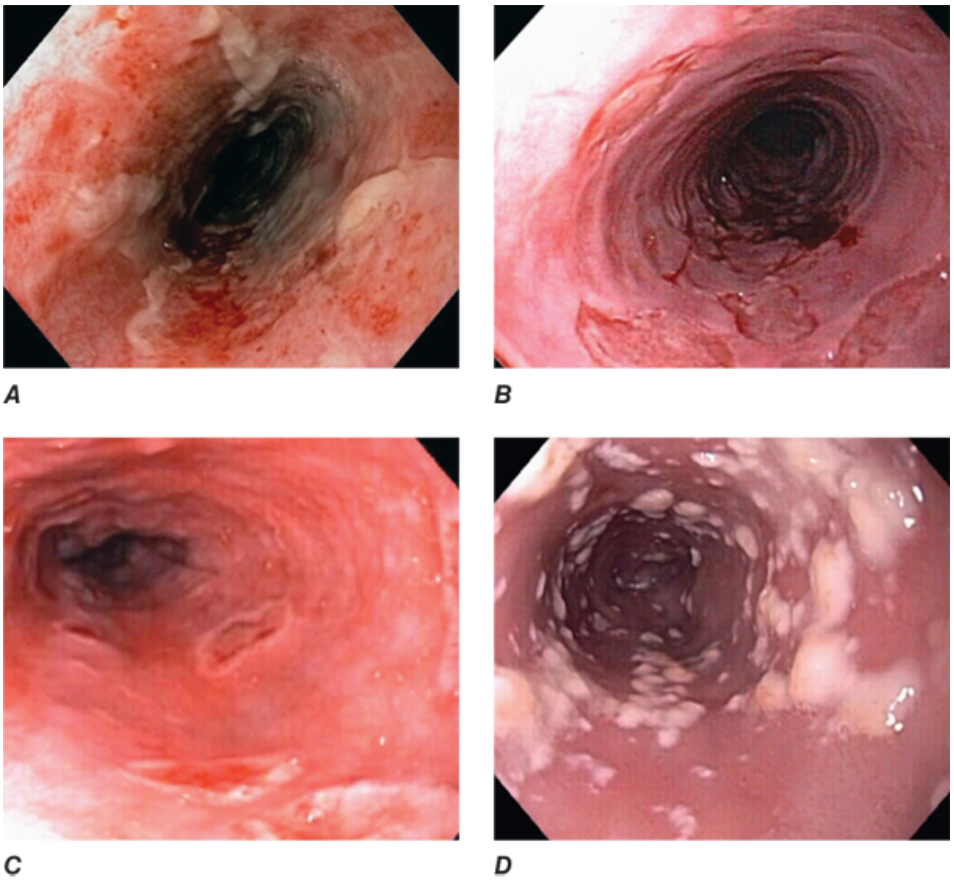
Dyspepsia is a chronic or recurrent burning discomfort or pain in the upper abdomen that may be caused by diverse processes, such as gastroesophageal reflux, peptic ulcer disease, and “nonulcer dyspepsia,” a heterogeneous category that includes disorders of motility, sensation and somatization. Gastric and esophageal malignancies are less common causes of dyspepsia. Careful history-taking allows accurate differential diagnosis of dyspepsia in only about half of patients. In the remainder, endoscopy can be a useful diagnostic tool, especially in patients whose symptoms are not resolved by *Helicobacter pylori* treatment or an empirical trial of acid-reducing therapy. Endoscopy should be performed at the outset in patients with dyspepsia and alarm features, such as weight loss or iron-deficiency anemia.

GASTROESOPHAGEAL REFLUX DISEASE (GERD)

When classic symptoms of gastroesophageal reflux are present, such as water brash and substernal heartburn, presumptive diagnosis and empirical treatment are often sufficient. Endoscopy is a sensitive test for diagnosis of esophagitis (Fig. 315-48), but it will miss nonerosive reflux disease (NERD) since some patients have symptomatic reflux without esophagitis. The most sensitive test for diagnosis of GERD is 24-h ambulatory pH monitoring. Endoscopy is indicated in patients with reflux symptoms refractory to antisecretory therapy; in those with alarm symptoms, such as dysphagia, weight loss, or gastrointestinal bleeding; and in those with recurrent dyspepsia after treatment that is not clearly due to reflux on clinical grounds alone. Endoscopy should be considered in patients with long-standing (≥10 years) GERD, as they have a sixfold increased risk of harboring Barrett’s esophagus compared to patients with <1 year of reflux symptoms.

FIGURE 315-48

Causes of esophagitis. ***A.*** Severe reflux esophagitis with mucosal ulceration and friability. ***B.*** Cytomegalovirus esophagitis. ***C.*** Herpes simplex virus esophagitis with target-type shallow ulcerations. ***D.*** Candida esophagitis with white plaques adherent to the esophageal mucosa.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Barrett’s Esophagus

Barrett's esophagus is specialized columnar metaplasia that replaces the normal squamous mucosa of the distal esophagus in some persons with GERD. Barrett's epithelium is a major risk factor for adenocarcinoma of the esophagus and is readily detected endoscopically, due to proximal displacement of the squamocolumnar junction (Fig. 315-5). A screening EGD for Barrett's esophagus should be considered in patients with a chronic (≥10 year) history of GERD symptoms. Endoscopic biopsy is the gold standard for confirmation of Barrett's esophagus and for dysplasia or cancer arising in Barrett's mucosa.

Periodic EGD with biopsies is recommended for surveillance of patients with Barrett's esophagus. Endoscopic resection (EMR or ESD) and/or ablation are performed when high-grade dysplasia or intramucosal cancer are found in the Barrett's mucosa. Although guidelines recommend observation and surveillance of low-grade dysplasia in Barrett's mucosa, recent evidence suggests that endoscopic treatment may be appropriate in select patients. Radiofrequency ablation (RFA) is the commonest ablative modality used for endoscopic treatment of Barrett's esophagus, and other modalities, such as cryotherapy, are also available.

PEPTIC ULCER

Peptic ulcer classically causes epigastric gnawing or burning, often occurring nocturnally and promptly relieved by food or antacids. Although endoscopy is the most sensitive diagnostic test for peptic ulcer, it is not a cost-effective strategy in young patients with ulcer-like dyspeptic symptoms unless endoscopy is available at low cost. Patients with suspected peptic ulcer should be evaluated for *H. pylori* infection. Serology (past or present infection), urea breath testing (current infection), and stool tests are noninvasive and less costly than endoscopy with biopsy. Patients aged >50 and those with alarm symptoms or persistent symptoms despite treatment should undergo endoscopy to exclude malignancy.

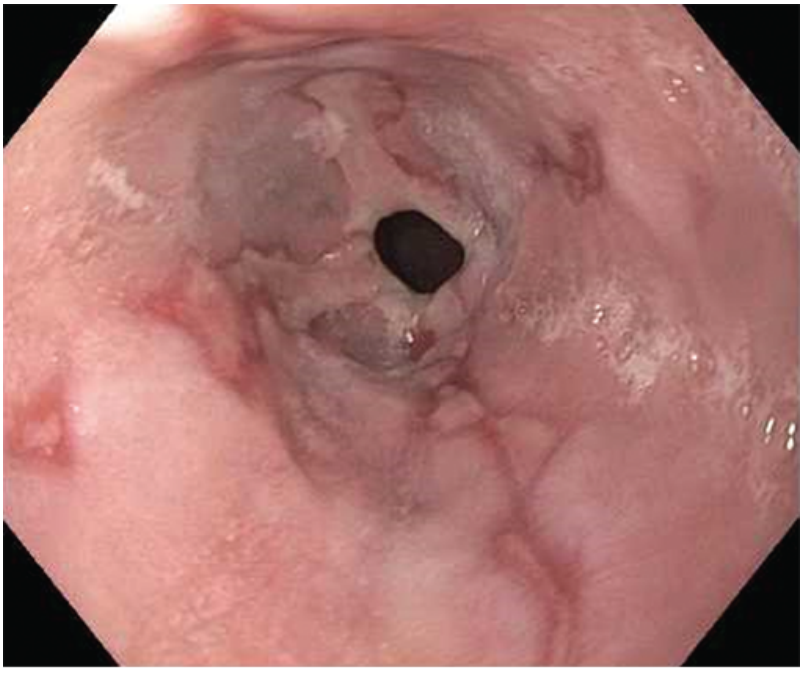
NONULCER DYSPEPSIA

Nonulcer dyspepsia may be associated with bloating and, unlike peptic ulcer, tends not to remit and recur. Most patients describe persistent symptoms despite acid-reducing, prokinetic, or anti-*Helicobacter* therapy, and are referred for endoscopy to exclude a refractory ulcer and assess for other causes. Although endoscopy is useful for excluding other diagnoses, its impact on the treatment of patients with nonulcer dyspepsia is limited.

DYSPHAGIA

About 50% of patients presenting with difficulty swallowing have a mechanical obstruction; the remainder has a motility disorder, such as achalasia or diffuse esophageal spasm. Careful history-taking often points to a presumptive diagnosis and leads to the appropriate use of diagnostic tests. Esophageal strictures **(Fig. 315-49)** typically cause progressive dysphagia, first for solids, then for liquids; motility disorders often cause intermittent dysphagia for both solids and liquids. Some underlying disorders have characteristic historic features: Schatzki's ring **(Fig. 315-50)** causes episodic dysphagia for solids, typically at the beginning of a meal; oropharyngeal motor disorders typically present with difficulty initiating deglutition (*transfer dysphagia*) and nasal reflux or coughing with swallowing; and achalasia may cause nocturnal regurgitation of undigested food.

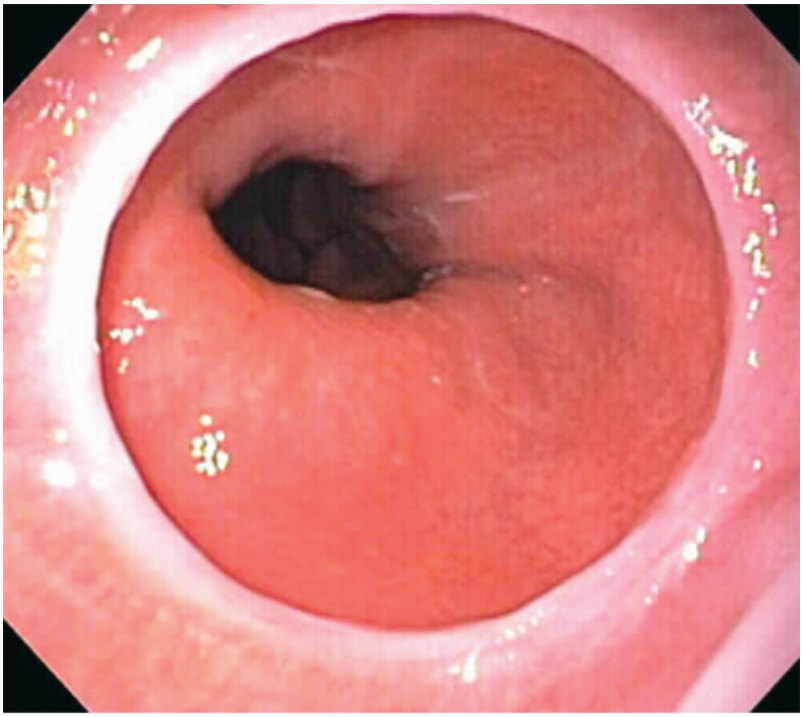
FIGURE 315-49
Peptic esophageal stricture associated with esophagitis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-50

Schatzki's ring at the gastroesophageal junction.

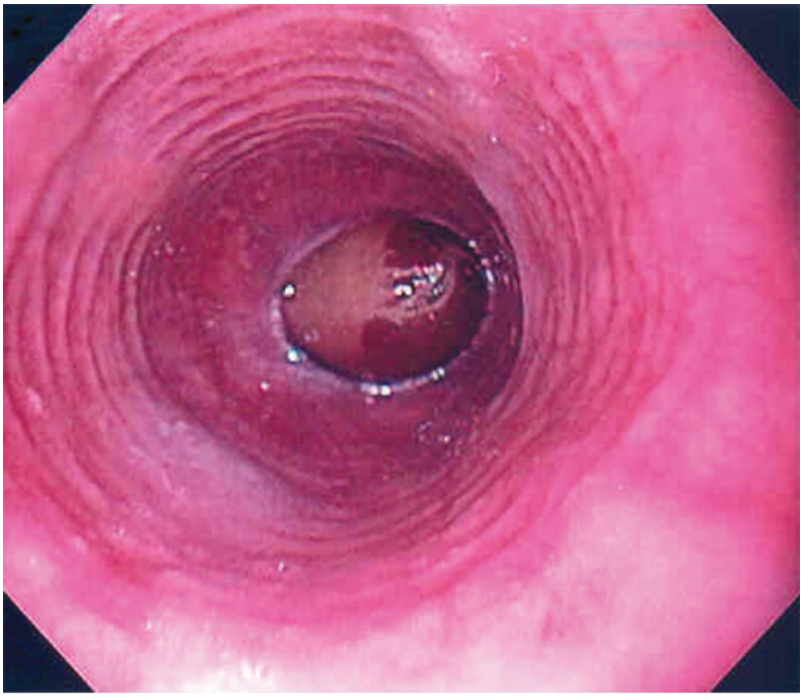


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

When mechanical obstruction is suspected, endoscopy is a useful initial diagnostic test, since it permits immediate biopsy and/or dilation of strictures, masses, or rings. The presence of linear furrows and multiple corrugated rings throughout a narrowed esophagus (*feline esophagus*) should raise suspicion for eosinophilic esophagitis, an increasingly recognized cause for recurrent dysphagia and food impaction (Fig. 315-51). Blind or forceful passage of an endoscope may lead to perforation in a patient with stenosis of the cervical esophagus or a Zenker's diverticulum, but gentle passage of an endoscope under direct visual guidance is reasonably safe. Endoscopy can miss a subtle stricture or ring in some patients.

FIGURE 315-51

Eosinophilic esophagitis with multiple circular rings of the esophagus creating a corrugated appearance, and an impacted grape at the narrowed esophagogastric junction. The diagnosis requires biopsy with histologic finding of >15–20 eosinophils/high-power field.



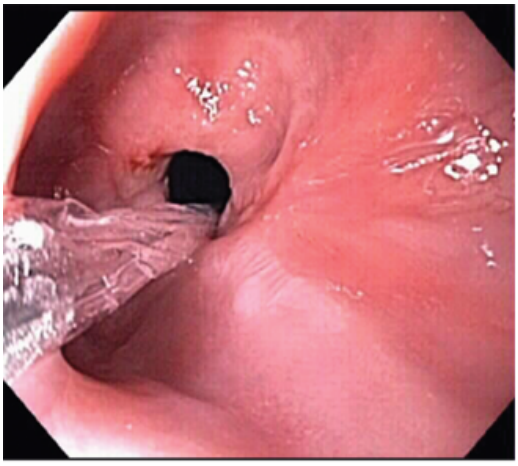
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

When transfer dysphagia is evident or an esophageal motility disorder is suspected, esophageal radiography and/or a video-swallow study are the best initial diagnostic tests. The oropharyngeal swallowing mechanism, esophageal peristalsis, and the lower esophageal sphincter can all be assessed. In some disorders, subsequent esophageal manometry may also be important for diagnosis.

Various causes of dysphagia are amenable to endoscopic therapy. Benign strictures, rings, and webs can be dilated using a through-the-scope balloon **(Fig. 315-52)** or a polyvinyl dilator passed over a guide wire. In some instances, thin fibrotic strictures may respond to needle-knife electroincision **(Fig. 315-53)** when they prove refractory to dilation. Esophageal covered stents can be used to palliate dysphagia from malignant obstruction **(Fig. 315-54)**, and flexible endoscopic myotomy is an option for Zenker's diverticulum **(Video V5-19)**. Recent advances in submucosal endoscopy have enabled the development of procedures, such as POEM **(Video V5-20)** and POET **(Video V5-21)** for the management of achalasia and select subepithelial esophageal tumors, respectively.

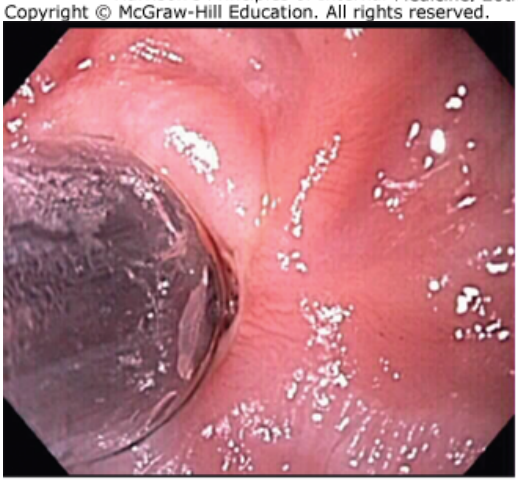
FIGURE 315-52

Endoscopic management of peptic stricture. ***A.*** Peptic stricture. ***B.*** Through-the-scope balloon dilation of stricture. ***C.*** Improvement in luminal diameter postdilation.



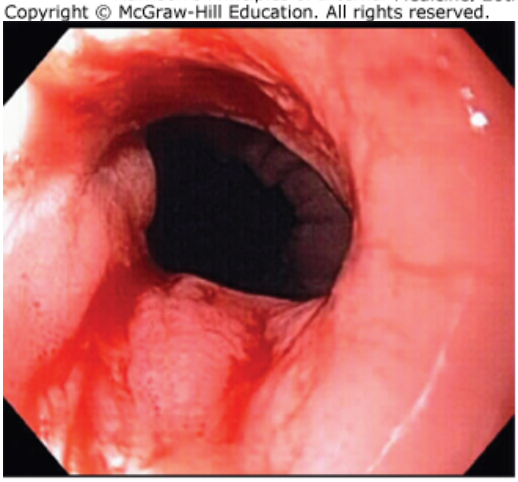
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

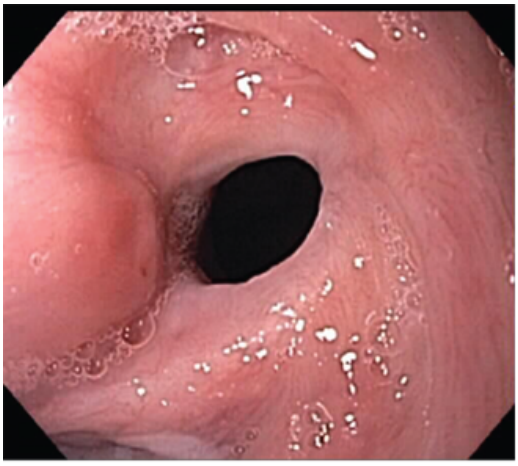


C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

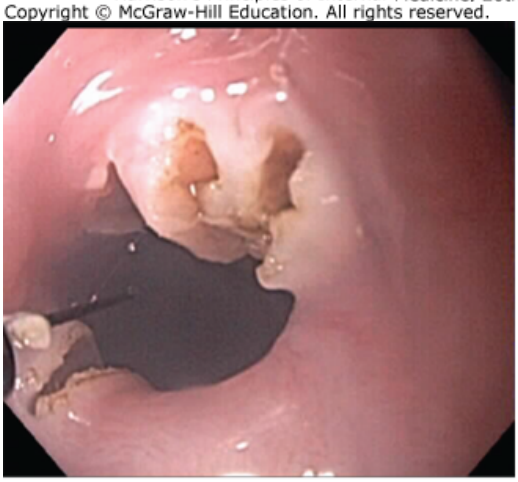
FIGURE 315-53

Endoscopic management of an esophagogastric anastomotic stricture. ***A.*** Recurrent anastomotic stricture despite periodic balloon dilation. ***B.*** Needle-knife electroincision of stricture. ***C.*** Improvement in luminal opening post therapy.



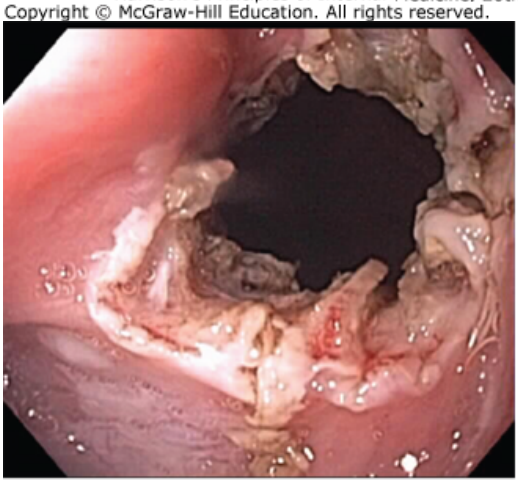
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

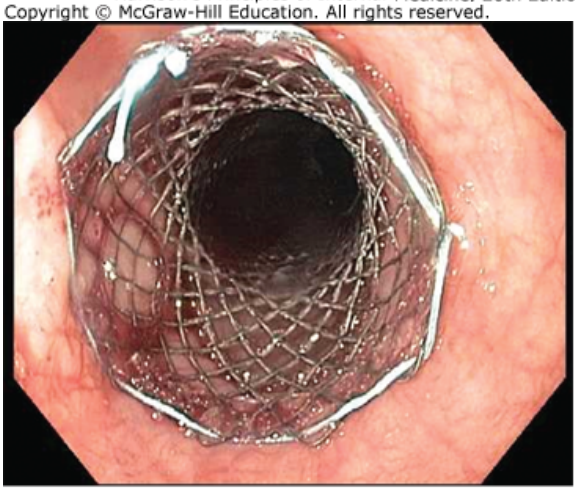
FIGURE 315-54

Palliation of malignant dysphagia. ***A.*** Obstructing distal esophageal cancer. ***B.*** Palliative stent placement.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ENDOSCOPIC TREATMENT OF OBESITY

The majority of Americans are overweight or obese, and obesity-associated diabetes has become a major public health problem. Bariatric surgery is the most effective weight-loss intervention, and it has been shown to decrease long-term mortality in obese persons, but many patients do not undergo surgery. Endoscopic treatments for obesity have been developed and include insertion of an intragastric balloon or duodenojejunal bypass liner, placement of a percutaneous gastric tube for aspiration of gastric contents after meals, or endoscopic sleeve gastroplasty, which utilizes endoscopic suturing to narrow the lumen of the gastric body (**Video V5-22**). Prospective trials show that these treatments induce total body weight loss of 7–20% and varying degrees of glycemic control. Additional endoscopic modalities are undergoing initial clinical trials. The long-term efficacy of endoscopic bariatric treatment is currently unknown.

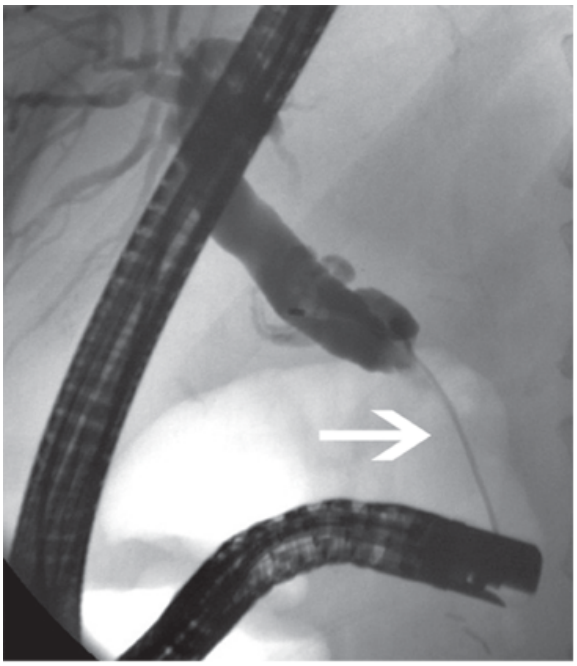
TREATMENT OF MALIGNANCIES

Endoscopy plays an important role in the treatment of gastrointestinal malignancies. Early-stage malignancies limited to the superficial layers of the gastrointestinal mucosa may be resected using the techniques of EMR (Video V5-4) or ESD (Video V5-5). Photodynamic therapy (PDT) and RFA are effective modalities for ablative treatment of high-grade dysplasia and intramucosal cancer in Barrett's esophagus (**Video V5-23**). Gastrointestinal stromal tumors can be removed en bloc by EFTR (Video V5-3). In general, endoscopic techniques offer the advantage of a minimally invasive approach to treatment but rely on other imaging techniques (such as CT, MRI, positron emission tomography [PET], and EUS) to exclude distant metastases or locally advanced disease better treated by surgery or other modalities. The decision to treat an early-stage gastrointestinal malignancy endoscopically is often made in collaboration with a surgeon and/or oncologist.

Endoscopic palliation of gastrointestinal malignancies relieves symptoms and in many cases prolongs survival. Malignant obstruction can be relieved by endoscopic stent placement (**Figs. 315-17**, 315-54, and **315-55**; Videos V5-16, V5-17), and malignant gastrointestinal bleeding can often be palliated endoscopically as well. EUS-guided celiac plexus neurolysis may relieve pancreatic cancer pain.

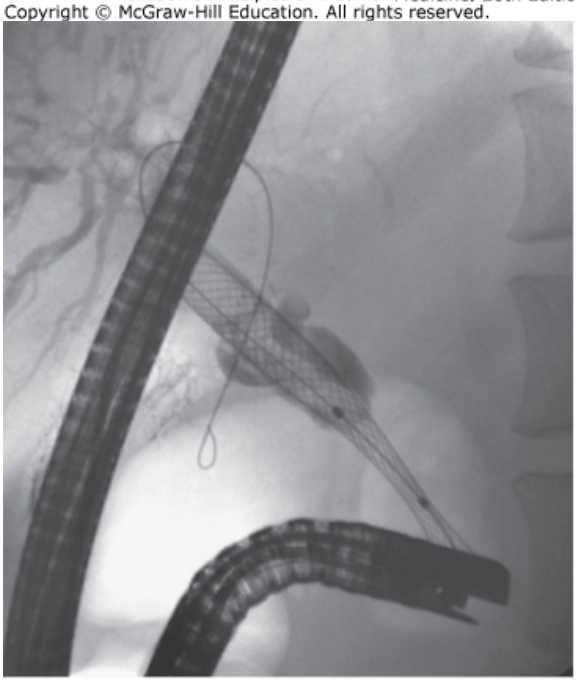
FIGURE 315-55

Biliary and duodenal self-expanding metal stents (SEMS) for obstruction caused by pancreatic cancer. ***A.*** Endoscopic retrograde cholangiopancreatography (ERCP) demonstrates a distal bile duct stricture *(arrow)*. ***B.*** A biliary SEMS is placed. ***C.*** Contrast injection demonstrates a duodenal stricture *(arrow)*. ***D.*** Biliary and duodenal SEMS in place.



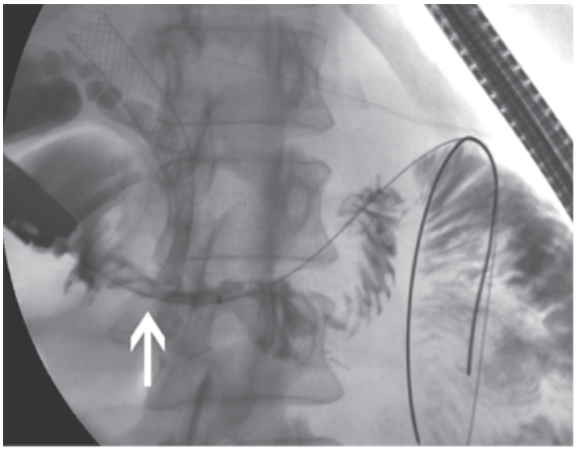
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



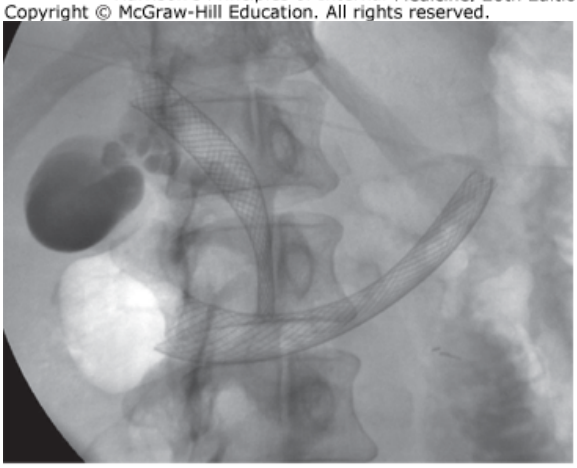
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



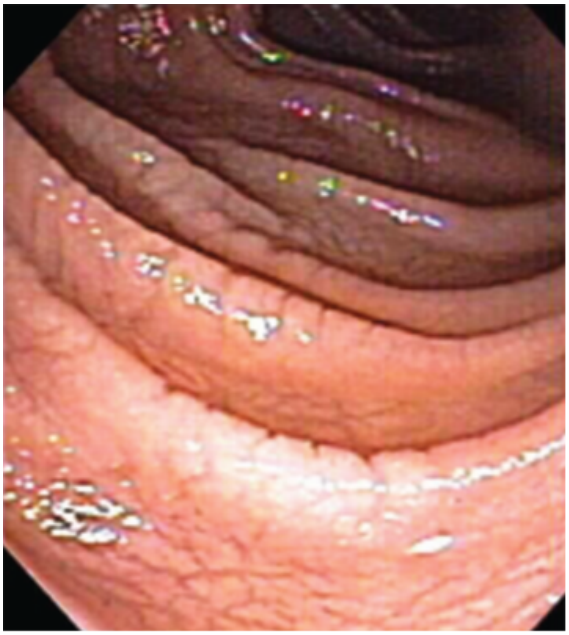
D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ANEMIA AND OCCULT BLOOD IN THE STOOL

Iron-deficiency anemia may be attributed to poor iron absorption (as in celiac sprue) or, more commonly, chronic blood loss. Intestinal bleeding should be strongly suspected in men and postmenopausal women with iron-deficiency anemia, and colonoscopy is indicated in such patients, even in the absence of detectable occult blood in the stool. Approximately 30% will have large colonic polyps, 10% will have colorectal cancer, and a few patients will have colonic vascular lesions. When a convincing source of blood loss is not found in the colon, upper gastrointestinal endoscopy should be considered; if no lesion is found, duodenal biopsies should be obtained to exclude sprue **(Fig. 315-56)**. Small bowel evaluation with capsule endoscopy **(Fig. 315-57)**, CT or MR enterography, or device-assisted enteroscopy may be appropriate if both EGD and colonoscopy are unrevealing.

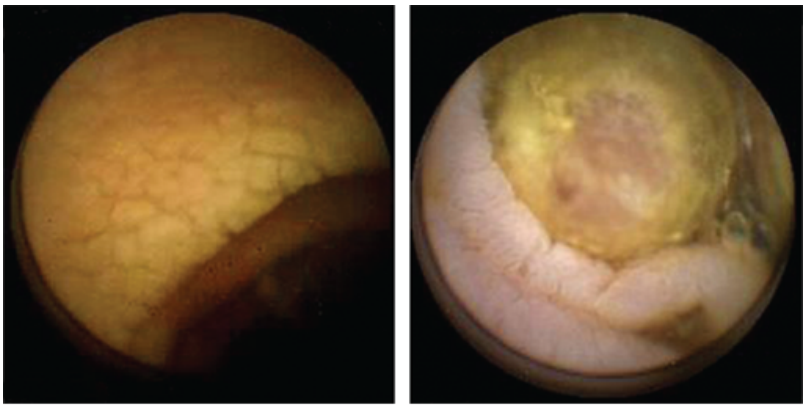
FIGURE 315-56
Scalloped duodenal folds in a patient with celiac sprue.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 315-57

Capsule endoscopy images of a mildly scalloped jejunal fold (*left*) and an ileal tumor (*right*) in a patient with celiac sprue. *(Images courtesy of Dr. Elizabeth Rajan; with permission.)*



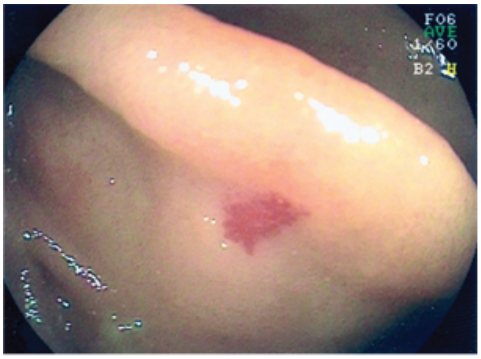
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Tests for occult blood in the stool detect hemoglobin or the heme moiety and are most sensitive for colonic blood loss, although they will also detect larger amounts of upper gastrointestinal bleeding. Patients with occult blood in normal-appearing stool should undergo colonoscopy to diagnose or exclude colorectal neoplasia, especially if they are over 50 years of age or have a family history of colonic neoplasia. Whether upper endoscopy is also indicated depends on the patient's symptoms.

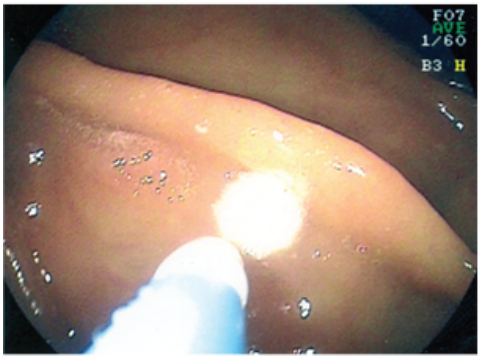
The small intestine may be the source of chronic intestinal bleeding, especially if colonoscopy and upper endoscopy are not diagnostic. The utility of small bowel evaluation varies with the clinical setting and is most important in patients in whom bleeding causes chronic or recurrent anemia. In contrast to the low diagnostic yield of small bowel radiography, positive findings on capsule endoscopy are seen in 50–70% of patients with suspected small intestinal bleeding. The most common finding is mucosal vascular ectasia. CT or MR enterography accurately detects small bowel masses and inflammation, and are also useful for initial small bowel evaluation. Deep enteroscopy may follow capsule endoscopy for biopsy of lesions or to provide specific therapy, such as argon plasma coagulation of vascular ectasias **(Fig. 315-58)**.

FIGURE 315-58

A. Mid-jejunal vascular ectasia identified by double-balloon endoscopy. ***B.*** Ablation of vascular ectasia with argon plasma coagulation.



A



B

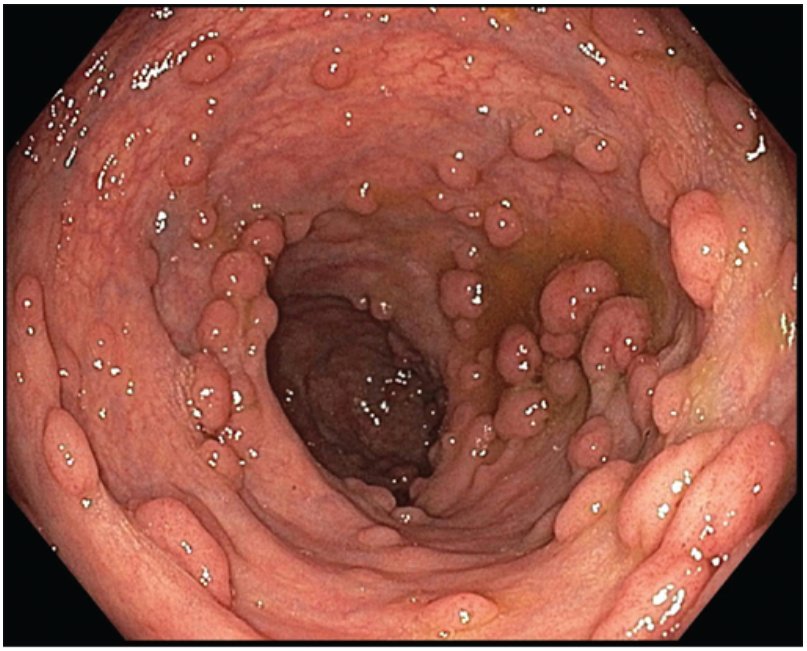
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

COLORECTAL CANCER SCREENING

The majority of colon cancers develop from preexisting colonic adenomas, and colorectal cancer can be largely prevented by the detection and removal of adenomatous polyps (Video V5-24). The choice of screening strategy for an asymptomatic person depends on personal and family history. Individuals with inflammatory bowel disease, a history of colorectal polyps or cancer, family members with adenomatous polyps or cancer, or certain familial cancer syndromes (Fig. 315-59) are at increased risk for colorectal cancer. An individual without these factors is generally considered at average risk.

FIGURE 315-59

Numerous colon polyps in a patient with familial adenomatous polyposis syndrome.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Screening strategies are summarized in Table 315-3. While stool tests for occult blood have been shown to decrease mortality rate from colorectal cancer, they do not detect some cancers and many polyps, and direct visualization of the colon is a more effective screening strategy. Either sigmoidoscopy or colonoscopy may be used for cancer screening in asymptomatic average-risk individuals. The use of sigmoidoscopy was based on the historical finding that the majority of colorectal cancers occurred in the rectum and left colon, and that patients with right-sided colon cancers had left-sided polyps. Over the past several decades, however, the distribution of colon cancers has changed in the United States, with proportionally fewer rectal and left-sided cancers than in the past. Large American studies of colonoscopy for screening of average-risk individuals show that cancers are roughly equally distributed between left and right colon and half of patients with right-sided lesions have no polyps in the left colon. Visualization of the entire colon thus appears to be the optimal strategy for colorectal cancer screening and prevention.

TABLE 315-3

Colorectal Cancer Screening Strategies

	Choices/Recommendations	Comments
Average-Risk Patients		
Asymptomatic individuals ≥50 years of age (≥45 years of age for African Americans)	Colonoscopy every 10 years[a]	Preferred cancer prevention strategy
	Annual FIT or FOBT, multiple take-home specimen cards	Cancer detection strategy, does not detect most polyps; colonoscopy if results are positive
	CT colonography every 5 years	Colonoscopy if results are positive
	Flexible sigmoidoscopy every 5 years	Does not detect proximal colon polyps and cancers; colonoscopy if results are positive
	Double-contrast barium enema every 5 years	Less sensitive than colonoscopy or CT colonography, misses some cancers and polyps; colonoscopy if results are positive
	Stool DNA test every 3 years	Does not detect many polyps; colonoscopy if results are positive
Personal History of Polyps or CRC		
1 or 2 small (<1 cm) adenomas with low-grade dysplasia	Repeat colonoscopy in 5–10 years	Assuming complete polyp resection. Interval may vary based on prior history, family history
3–10 adenomas, or any high-risk adenoma[b]	Repeat colonoscopy in 3 years; subsequent colonoscopy based on findings	Assuming complete polyp resection
>10 adenomas	Repeat colonoscopy in <3 years based on clinical judgment	Consider evaluation for FAP or HNPCC; see recommendations below
Piecemeal removal of a sessile polyp	Exam in 2–6 months to verify complete removal	
Small (<1 cm) hyperplastic polyps of sigmoid and rectum	Repeat colonoscopy in 10 years	Those with hyperplastic polyposis syndrome merit more frequent follow-up
Sessile serrated adenoma/polyp <10 mm, without dysplasia	Repeat colonoscopy in 5 years	
Sessile serrated adenoma/polyp ≥10 mm or with dysplasia, or ≥2 serrated polyps	Repeat colonoscopy in 3 years	Serrated polyposis syndrome merits more frequent follow-up
Incompletely removed serrated polyp ≥1 cm	Exam in 2–6 months to verify complete removal	

	Choices/Recommendations	Comments
Colon cancer	Evaluate entire colon around the time of resection, then repeat colonoscopy in 1 year	Subsequent colonoscopy in 3 years if the 1-year examination is normal
Inflammatory Bowel Disease		
Long-standing (>8 years) ulcerative pancolitis or Crohn's colitis, or left-sided ulcerative colitis of >15 years' duration	Colonoscopy with biopsies every 1–2 years	Consider chromoendoscopy or other advanced imaging techniques for detection of flat dysplasia during colonoscopy
Family History of Polyps or CRC		
First-degree relatives with only small tubular adenomas	Same as average risk	
Single first-degree relative with CRC or advanced adenoma at age ≥60 years	Colonoscopy every 10 years starting at age 40	
Single first-degree relative with CRC or advanced adenoma at age <60 years, OR two first-degree relatives with CRC or advanced adenomas at any age	Colonoscopy every 5 years beginning at age 40 years or 10 years younger than age at diagnosis of the youngest affected relative, whichever is earlier	
FAP	Sigmoidoscopy or colonoscopy annually, beginning at age 10–12 years	Consider genetic counseling and testing
HNPCC	Colonoscopy every 2 years beginning at age 20–25 years (or 10 years younger than the youngest affected first-degree relative) until age 40, then annually thereafter	Consider histologic evaluation for microsatellite instability in tumor specimens of patients who meet modified Bethesda criteria; consider genetic counseling and testing

[a]Assumes good colonic preparation and complete examination to cecum.

[b]High-risk adenoma: any adenoma ≥ 1 cm in size, or containing high-grade dysplasia or villous features

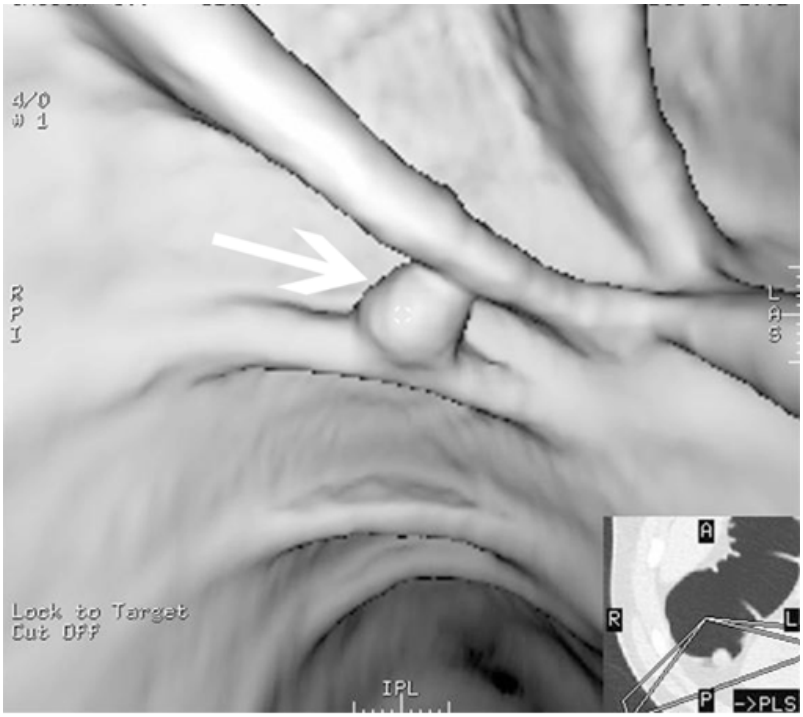
Abbreviations: CRC, colorectal cancer; FAP, familial adenomatous polyposis; FIT, fecal immunochemical test; FOBT, fecal occult blood test; HNPCC, hereditary nonpolyposis colorectal cancer.

Source: Adapted from United States Preventative Services Task Force Guidelines released in 2016 (*https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/colorectal-cancer-screening2?ds=1&s=colorectal)* accessed on Jan 2, 2017, and American Cancer Society Guidelines updated in 2016 (*http://www.cancer.org/cancer/colonandrectumcancer/moreinformation/colonandrectumcancerearlydetection/colorectal-cancer-early-detection-acs-recommendations*) accessed Jan 2, 2017.

Virtual colonoscopy (VC) is a radiologic technique that images the colon with CT following rectal insufflation of the colonic lumen. Computer rendering of CT images generates an electronic display of a virtual "flight" along the colonic lumen, simulating colonoscopy **(Fig. 315-60)**. Findings detected during VC often require subsequent conventional colonoscopy for confirmation and treatment.

FIGURE 315-60

Virtual colonoscopy image of a colon polyp (*arrow*). *(Image courtesy of Dr. Jeff Fidler; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

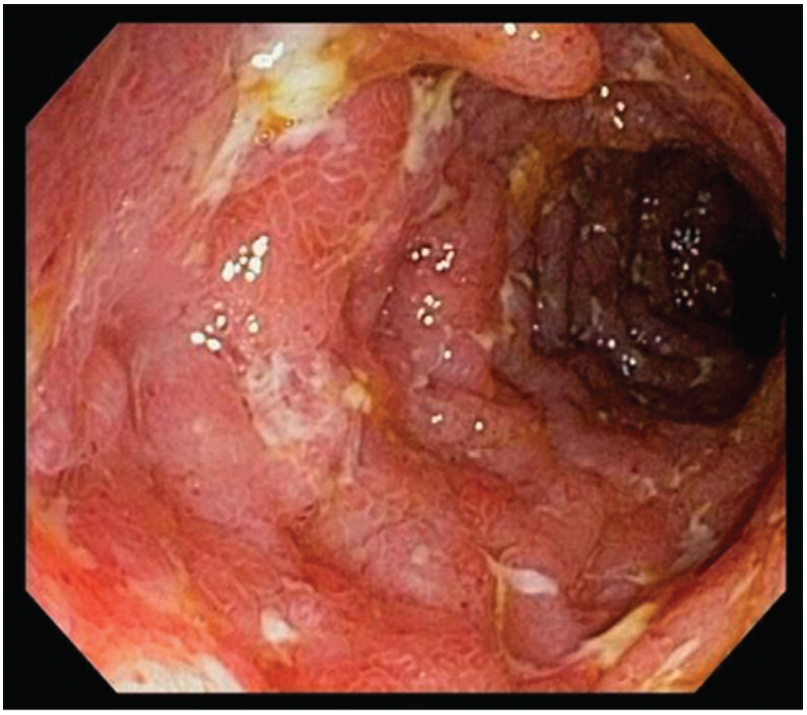
DIARRHEA

Most cases of diarrhea are acute, self-limited, and due to infections or medication. Chronic diarrhea (lasting >6 weeks) is more often due to a primary inflammatory, malabsorptive, or motility disorder, is less likely to resolve spontaneously, and generally requires diagnostic evaluation. Patients with chronic diarrhea or severe, unexplained acute diarrhea often undergo endoscopy if stool tests for pathogens are unrevealing. The choice of endoscopic testing depends on the clinical setting.

Patients with colonic symptoms and findings such as bloody diarrhea, tenesmus, fever, or leukocytes in stool generally undergo sigmoidoscopy or colonoscopy to assess for colitis (Fig. 315-8). Sigmoidoscopy is an appropriate initial test in most patients. Conversely, patients with symptoms and findings suggesting small bowel disease, such as large-volume watery stools, substantial weight loss, and malabsorption of iron, calcium, or fat, may undergo upper endoscopy with duodenal aspirates for assessment of bacterial overgrowth and biopsies for assessment of mucosal diseases, such as celiac sprue.

Many patients with chronic diarrhea do not fit either of these patterns. In the setting of a long-standing history of alternating constipation and diarrhea dating to early adulthood, without findings such as blood in the stool or anemia, a diagnosis of irritable bowel syndrome may be made without direct visualization of the bowel. Steatorrhea and upper abdominal pain may prompt evaluation of the pancreas rather than the gut. Patients whose chronic diarrhea is not easily categorized often undergo initial colonoscopy to examine the entire colon and terminal ileum for inflammatory or neoplastic disease **(Fig. 315-61)**.

FIGURE 315-61
Crohn's ileitis.

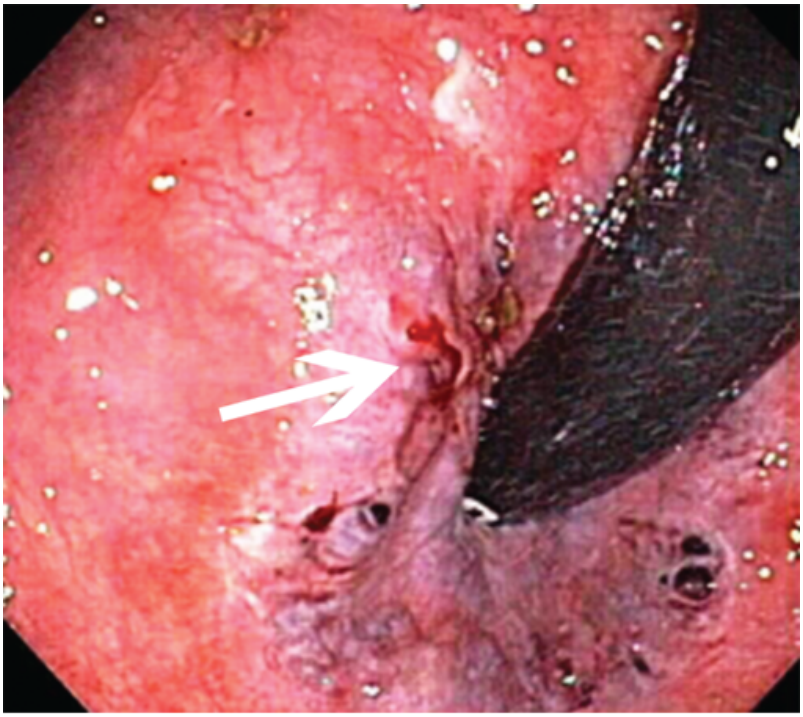


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MINOR HEMATOCHEZIA

Bright red blood passed with or on formed brown stool usually has a rectal, anal, or distal sigmoid source **(Fig. 315-62)**. Patients with even trivial amounts of hematochezia should be investigated with flexible sigmoidoscopy and anoscopy to exclude polyps or cancers in the distal colon. Patients reporting red blood on the toilet tissue only, without blood in the toilet or on the stool, are generally bleeding from a lesion in the anal canal; careful external inspection, digital examination, and proctoscopy with anoscopy may be sufficient for diagnosis in such cases.

FIGURE 315-62
Internal hemorrhoids with bleeding *(arrow)* as seen on a retroflexed view of the rectum.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PANCREATITIS

About 20% of patients with pancreatitis have no identified cause after routine clinical investigation (including a review of medication and alcohol use, measurement of serum triglyceride and calcium levels, abdominal ultrasonography, and CT or MR). Endoscopic assessment leads to a specific diagnosis in the majority of such patients, often altering clinical management. Endoscopic investigation is particularly appropriate if the patient has had more than one episode of pancreatitis.

Microlithiasis, or the presence of microscopic crystals in bile, is a leading cause of previously unexplained acute pancreatitis and is sometimes seen during abdominal ultrasonography as layering sludge or flecks of floating, echogenic material in the gallbladder. EUS may identify previously undetected microlithiasis.

Previously undetected chronic pancreatitis, pancreatic malignancy, or pancreas divisum may be diagnosed by either ERCP or EUS. Autoimmune pancreatitis is often suspected on the basis of CT, MR, or serologic findings, but it may first become apparent during EUS and may require EUS-guided pancreatic biopsy for histologic diagnosis.

Severe pancreatitis often results in pancreatic fluid collections. Symptomatic pseudocysts and areas of walled-off pancreatic necrosis can be drained into the stomach or duodenum endoscopically, using transpapillary and transmural endoscopic techniques. Pancreatic necrosis can be treated by direct endoscopic necrosectomy (Video V5-2) via an endoscopically created transmural drainage site.

CANCER STAGING

Local staging of esophageal, gastric, pancreatic, bile duct, and rectal cancers can be obtained with EUS (Fig. 315-19). EUS with fine-needle aspiration (Fig. 315-20) currently provides the most accurate preoperative assessment of local tumor and nodal staging, but it does not detect many distant metastases. Details of the local tumor stage can guide treatment decisions including resectability and need for neoadjuvant therapy. EUS with transesophageal needle biopsy may also be used to assess the presence of non-small-cell lung cancer in mediastinal nodes.

OPEN-ACCESS ENDOSCOPY

Direct scheduling of endoscopic procedures by primary care physicians without preceding gastroenterology consultation, or *open-access endoscopy*, is common. When the indications for endoscopy are clear-cut and appropriate, the procedural risks are low, and the patient

understands what to expect, open-access endoscopy streamlines patient care and decreases costs.

Patients referred for open-access endoscopy should have a recent history, physical examination, and medication review. A copy of such an evaluation should be available when the patient comes to the endoscopy suite. Patients with unstable cardiovascular or respiratory conditions should not be referred directly for open-access endoscopy. Patients with particular conditions and undergoing certain procedures should be prescribed prophylactic antibiotics prior to endoscopy (Table 315-1). In addition, patients taking anticoagulants and/or antiplatelet drugs may require adjustment of these agents before endoscopy based on the procedural risk for bleeding and their underlying risk for a thromboembolic event (Table 315-2).

Common indications for open-access EGD include dyspepsia resistant to a trial of appropriate therapy, dysphagia, gastrointestinal bleeding, and persistent anorexia or early satiety. Open-access colonoscopy is often requested in men or postmenopausal women with iron-deficiency anemia, in patients age >50 with occult blood in the stool, in patients with a previous history of colorectal adenomatous polyps or cancer, and for colorectal cancer screening. Flexible sigmoidoscopy is commonly performed as an open-access procedure.

When patients are referred for open-access colonoscopy, the primary care provider may need to choose a colonic preparation. Commonly used oral preparations include polyethylene glycol lavage solution, with or without citric acid. A “split-dose” regimen improves the quality of colonic preparation. Sodium phosphate purgatives may cause fluid and electrolyte abnormalities and renal toxicity, especially in patients with renal failure or congestive heart failure and those >70 years of age.

FURTHER READING

ASGE Standards of Practice Committee et al: Appropriate use of GI endoscopy. Gastrointest Endosc 75:1127, 2012. [PubMed: 22624807]

ASGE Standards of Practice Committee et al: Antibiotic prophylaxis for GI endoscopy. Gastrointest Endosc 81:81, 2015. [PubMed: 25442089]

ASGE Standards of Practice Committee et al: Open-access endoscopy. Gastrointest Endosc 81:1326, 2015. [PubMed: 25865387]

Garcia-Tsao G et al: Portal hypertensive bleeding in cirrhosis: Risk stratification, diagnosis, and management: 2016 practice guidance by the American Association for the study of liver diseases. Hepatology 65:310, 2017. [PubMed: 27786365]

Gralnek IM et al: Diagnosis and management of nonvariceal upper gastrointestinal hemorrhage: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 47:a1, 2015. [PubMed: 26417980]

Lin JS et al: Screening for colorectal cancer: Updated evidence report and systematic review. JAMA 315:2576, 2016. [PubMed: 27305422]

Shaheen NJ et al: ACG Clinical Guideline: Diagnosis and management of Barrett’s esophagus. Am J Gastroenterol 111:30, 2016. [PubMed: 26526079]

Strate LL et al: ACG Clinical Guideline: Management of patients with acute lower gastrointestinal bleeding. Am J Gastroenterol 111:459, 2016. [PubMed: 26925883]

Chapter 316: Diseases of the Esophagus

Peter J. Kahrilas; Ikuo Hirano

ESOPHAGEAL STRUCTURE AND FUNCTION

The esophagus is a hollow, muscular tube coursing through the posterior mediastinum joining the hypopharynx to the stomach with a sphincter at each end. It functions to transport food and fluid between these ends, otherwise remaining empty. The physiology of swallowing, esophageal motility, and oral and pharyngeal dysphagia are described in **Chap. 40**. Esophageal diseases can be manifested by impaired function or pain. Key functional impairments are swallowing disorders and excessive gastroesophageal reflux. Pain, sometimes indistinguishable from cardiac chest pain, can result from inflammation, infection, dysmotility, or neoplasm.

SYMPTOMS OF ESOPHAGEAL DISEASE

The clinical history remains central to the evaluation of esophageal symptoms. A thoughtfully obtained history will often expedite management. Important details include weight gain or loss, gastrointestinal bleeding, dietary habits including the timing of meals, smoking, and alcohol consumption. The major esophageal symptoms are heartburn, regurgitation, chest pain, dysphagia, odynophagia, and globus sensation.

Heartburn (pyrosis), the most common esophageal symptom, is characterized by a discomfort or burning sensation behind the sternum that arises from the epigastrium and may radiate toward the neck. Heartburn is an intermittent symptom, most commonly experienced after eating, during exercise, and while lying recumbent. The discomfort is relieved with drinking water or antacid but can occur frequently interfering with normal activities including sleep. The association between heartburn and gastroesophageal reflux disease (GERD) is so strong that empirical therapy for GERD has become accepted management. However, the term “heartburn” is often misused and/or referred to with other terms such as “indigestion” or “repeating,” making it important to clarify the intended meaning.

Regurgitation is the effortless return of food or fluid into the pharynx without nausea or retching. Patients report a sour or burning fluid in the throat or mouth that may also contain undigested food particles. Bending, belching, or maneuvers that increase intraabdominal pressure can provoke regurgitation. A clinician needs to discriminate among regurgitation, vomiting, and rumination. *Vomiting* is preceded by nausea and accompanied by retching. *Rumination* is a behavior in which recently swallowed food is regurgitated and then reswallowed repetitively for up to an hour. Although there is some linkage between rumination and cognitive deficiency, the behavior is also exhibited by unimpaired individuals.

Chest pain is a common esophageal symptom with characteristics similar to cardiac pain, sometimes making this distinction difficult. Esophageal pain is usually experienced as a pressure type sensation in the mid chest, radiating to the mid back, arms, or jaws. The similarity to cardiac pain is likely because the two organs share a nerve plexus and the nerve endings in the esophageal wall have poor discriminative ability among stimuli. Esophageal distention or even chemostimulation (e.g., with acid) will often be perceived as chest pain. Gastroesophageal reflux is the most common cause of esophageal chest pain.

Esophageal *dysphagia* **(Chap. 40)** is often described as a feeling of food “sticking” or even lodging in the chest. Important distinctions are between uniquely solid food dysphagia as opposed to liquid and solid, episodic versus constant dysphagia, and progressive versus static dysphagia. If the dysphagia is for liquids as well as solid food, it suggests a motility disorder such as

achalasia. Conversely, uniquely solid food dysphagia is suggestive of a stricture, ring, or tumor. Of note, a patient's localization of food hang-up in the esophagus is notoriously imprecise. Approximately 30% of distal esophageal obstructions are perceived as cervical dysphagia. In such instances, the absence of concomitant symptoms generally associated with oropharyngeal dysphagia such as aspiration, nasopharyngeal regurgitation, cough, drooling, or obvious neuromuscular compromise should suggest an esophageal etiology.

Odynophagia is pain either caused by or exacerbated by swallowing. Although typically considered distinct from dysphagia, odynophagia may manifest concurrently with dysphagia. Odynophagia is more common with pill or infectious esophagitis than with reflux esophagitis and should prompt a search for these entities. When odynophagia does occur in GERD, it is likely related to an esophageal ulcer or extensive erosions.

Globus sensation, also known as globus pharyngeus, is the perception of a lump or fullness in the throat that is felt irrespective of swallowing. Although such patients are frequently referred for an evaluation of dysphagia, globus sensation is often relieved by the act of swallowing. As implied by its alternative name, "globus hystericus," globus sensation often occurs in the setting of anxiety or obsessive-compulsive disorders. Clinical experience teaches that it is often attributable to GERD.

Water brash is excessive salivation resulting from a vagal reflex triggered by acidification of the esophageal mucosa. This is not a common symptom. Afflicted individuals will describe the unpleasant sensation of the mouth rapidly filling with salty thin fluid, often in the setting of concomitant heartburn.

DIAGNOSTIC STUDIES

ENDOSCOPY

Endoscopy, also known as esophagogastroduodenoscopy (EGD), is the most useful test for the evaluation of the proximal gastrointestinal tract. Modern instruments produce high-quality, color images of the esophageal, gastric, and duodenal lumen. Endoscopes also have an instrumentation channel through which biopsy forceps, injection catheters for local delivery of therapeutic agents, balloon dilators, or hemostatic devices can be used. The key advantages of endoscopy over barium radiography are: (1) increased sensitivity for the detection of mucosal lesions, (2) vastly increased sensitivity for the detection of abnormalities mainly identifiable by color such as Barrett's metaplasia or vascular lesions, (3) the ability to obtain biopsy specimens for histologic examination of suspected abnormalities, and (4) the ability to dilate strictures during the examination. The main disadvantages of endoscopy are low sensitivity for detection of diffuse, non-focal esophageal strictures, cost, and the utilization of sedatives or anesthetics.

RADIOGRAPHY

Contrast radiography of the esophagus, stomach, and duodenum can demonstrate reflux of the contrast media, hiatal hernia, mucosal granularity, erosions, ulcerations, and strictures. The sensitivity of radiography compared with endoscopy for detecting reflux esophagitis reportedly ranges from 22 to 95%, with higher grades of esophagitis (i.e., ulceration or stricture) exhibiting greater detection rates. Conversely, the sensitivity of barium radiography for detecting esophageal strictures is greater than that of endoscopy, especially when the study is done in conjunction with a 13-mm barium tablet. Barium studies also provide an assessment of esophageal function and morphology that may be undetected on endoscopy. Tracheoesophageal fistula, altered postsurgical anatomy, and extrinsic esophageal compression are conditions where radiographic imaging complements endoscopic assessment. Hypopharyngeal pathology and disorders of the cricopharyngeus muscle are better appreciated on radiographic examination than with endoscopy, particularly with rapid sequence or video fluoroscopic recording. The major shortcoming of barium radiography is that it rarely obviates the need for endoscopy. Either a positive or a negative study is usually followed by an endoscopic evaluation either to obtain biopsies, provide therapy, or clarify findings in the case of a positive examination or to add a level of certainty in the case of a negative one.

ENDOSCOPIC ULTRASOUND

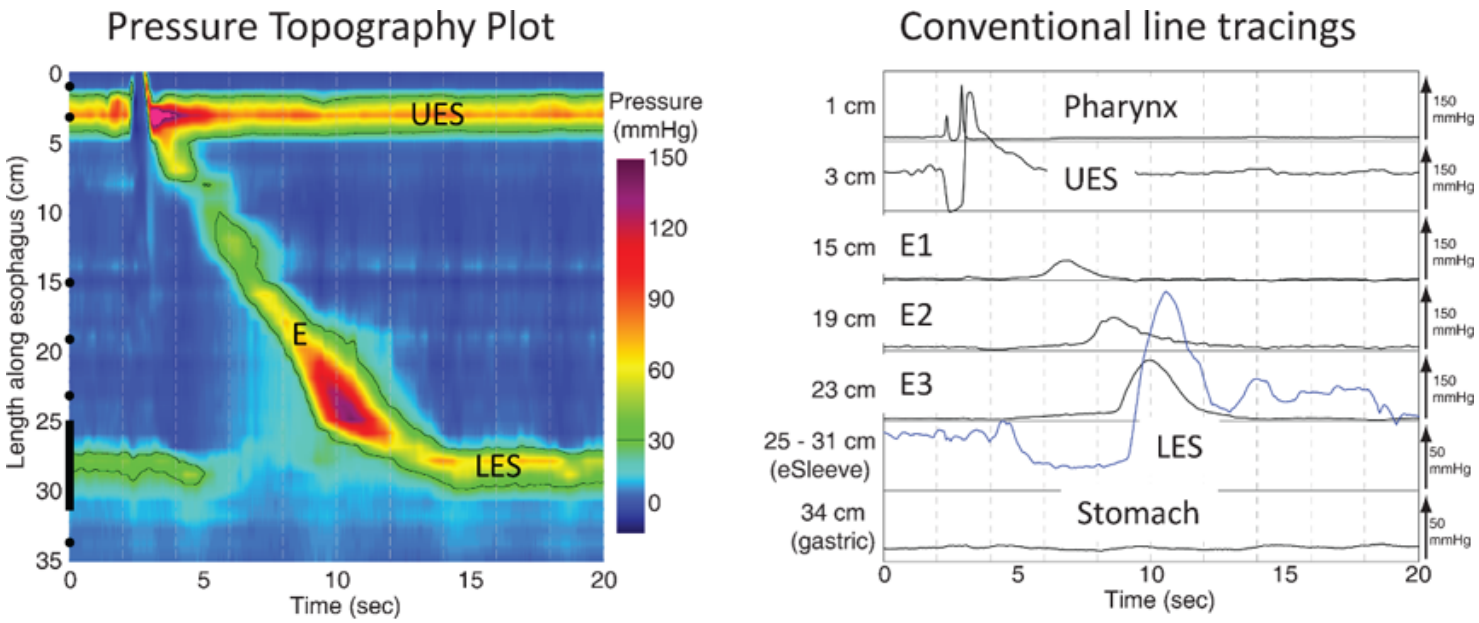
Endoscopic ultrasound (EUS) instruments combine an endoscope with an ultrasound transducer to create a transmural image of the tissue surrounding the endoscope tip. The key advantage of EUS over alternative radiologic imaging techniques is much greater resolution attributable to the proximity of the ultrasound transducer to the area being examined. Available devices can provide either radial imaging (360-degree, cross-sectional) or a curved linear image that can guide fine-needle aspiration of imaged structures such as lymph nodes or tumors. Major esophageal applications of EUS are to stage esophageal cancer, to evaluate dysplasia in Barrett's esophagus, and to assess submucosal lesions.

ESOPHAGEAL MANOMETRY

Esophageal manometry, or motility testing, entails positioning a pressure-sensing catheter within the esophagus and then observing the contractility following test swallows. The upper and lower esophageal sphincters (LESs) appear as zones of high pressure that relax on swallowing, while the intersphincteric esophagus exhibits peristaltic contractions. Manometry is used to diagnose motility disorders (achalasia, diffuse esophageal spasm [DES]) and to assess peristaltic integrity prior to the surgery for reflux disease. Technologic advances have enhanced esophageal manometry as high-resolution esophageal pressure topography **(Fig. 316-1)**. Manometry can also be combined with intraluminal impedance monitoring. Impedance recordings use a series of paired electrodes added to the manometry catheter. Esophageal luminal contents in contact with the electrodes decrease (liquid) or increase (air) the impedance signal, allowing detection of anterograde or retrograde esophageal bolus transit.

FIGURE 316-1

High-resolution esophageal pressure topography (*right*) and conventional manometry (*left*) of a normal swallow. E, esophageal body; LES, lower esophageal sphincter; UES, upper esophageal sphincter.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

REFLUX TESTING

GERD is often diagnosed in the absence of endoscopic esophagitis, which would otherwise define the disease. This occurs in the settings of partially treated disease, an abnormally sensitive esophageal mucosa, or without obvious explanation. In such instances, reflux testing can demonstrate excessive esophageal exposure to refluxed gastric juice, the physiologic abnormality of GERD. This can be done by ambulatory 24- to 96-h esophageal pH recording using either a wireless pH-sensitive transmitter that

is affixed to the esophageal mucosa or a transnasally positioned wire electrode with the tip stationed in the distal esophagus. Either way, the outcome is expressed as the percentage of the day that the pH was <4 (indicative of recent acid reflux), with values exceeding 5% indicative of GERD. Reflux testing is useful in the evaluation of patients presenting with atypical symptoms or an inexplicably poor response to therapy. Intraluminal impedance monitoring can be added to pH monitoring to detect reflux events irrespective of whether or not they are acidic, potentially increasing the sensitivity of the study.

STRUCTURAL DISORDERS

HIATAL HERNIA

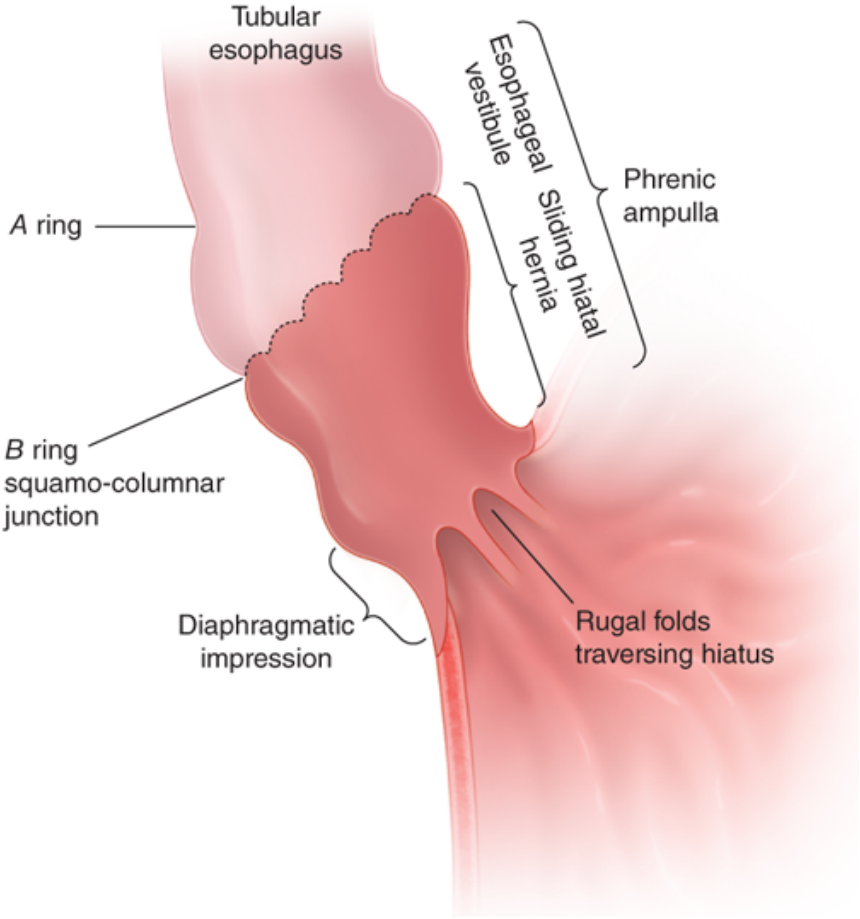
Hiatus hernia is a herniation of viscera, most commonly the stomach, into the mediastinum through the esophageal hiatus of the diaphragm. Four types of hiatus hernia are distinguished with type I, or sliding hiatal hernia, comprising at least 95% of the overall total. A sliding hiatal hernia is one in which the gastroesophageal junction and gastric cardia translocate cephalad as a result of weakening of the phrenoesophageal ligament attaching the gastroesophageal junction to the diaphragm at the hiatus and dilatation of the diaphragmatic hiatus. The incidence of sliding hernia increases with age. True to its name, sliding hernias enlarge with increased intraabdominal pressure, swallowing, and respiration. Conceptually, sliding hernias are the result of wear and tear: increased intraabdominal pressure from abdominal obesity, pregnancy, etc., along with hereditary factors predisposing to the condition. The main significance of sliding hernias is the propensity of affected individuals to have GERD.

Type II, III, and IV hiatal hernias are all subtypes of paraesophageal hernia in which the herniation into the mediastinum includes a visceral structure other than the gastric cardia. With type II and III paraesophageal hernias, the gastric fundus also herniates with the distinction being that in type II, the gastroesophageal junction remains fixed at the hiatus, whereas type III is a combined sliding and paraesophageal hernia. With type IV hiatal hernias, viscera other than the stomach herniate into the mediastinum, most commonly the colon. With type II and III paraesophageal hernias, the stomach inverts as it herniates and large paraesophageal hernias can lead to an “upside down stomach,” gastric volvulus, and even strangulation of the stomach. Because of this risk, surgical repair is often advocated for large paraesophageal hernias particularly when they are symptomatic.

RINGS AND WEBS

A lower esophageal mucosal ring, also called a *B ring*, is a thin membranous narrowing at the squamocolumnar mucosal junction (Fig. 316-2). Its origin is unknown, but B rings are demonstrable in about 10–15% of the general population and are usually asymptomatic. When the lumen diameter is <13 mm, distal rings are usually associated with episodic solid food dysphagia and are called *Schatzki rings*. Patients typically present older than 40 years, consistent with an acquired rather than congenital origin. Schatzki ring is one of the most common causes of intermittent food impaction, also known as “steakhouse syndrome” because meat is a typical instigator. Symptomatic rings are readily treated by dilation.

FIGURE 316-2
Radiographic anatomy of the gastroesophageal junction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

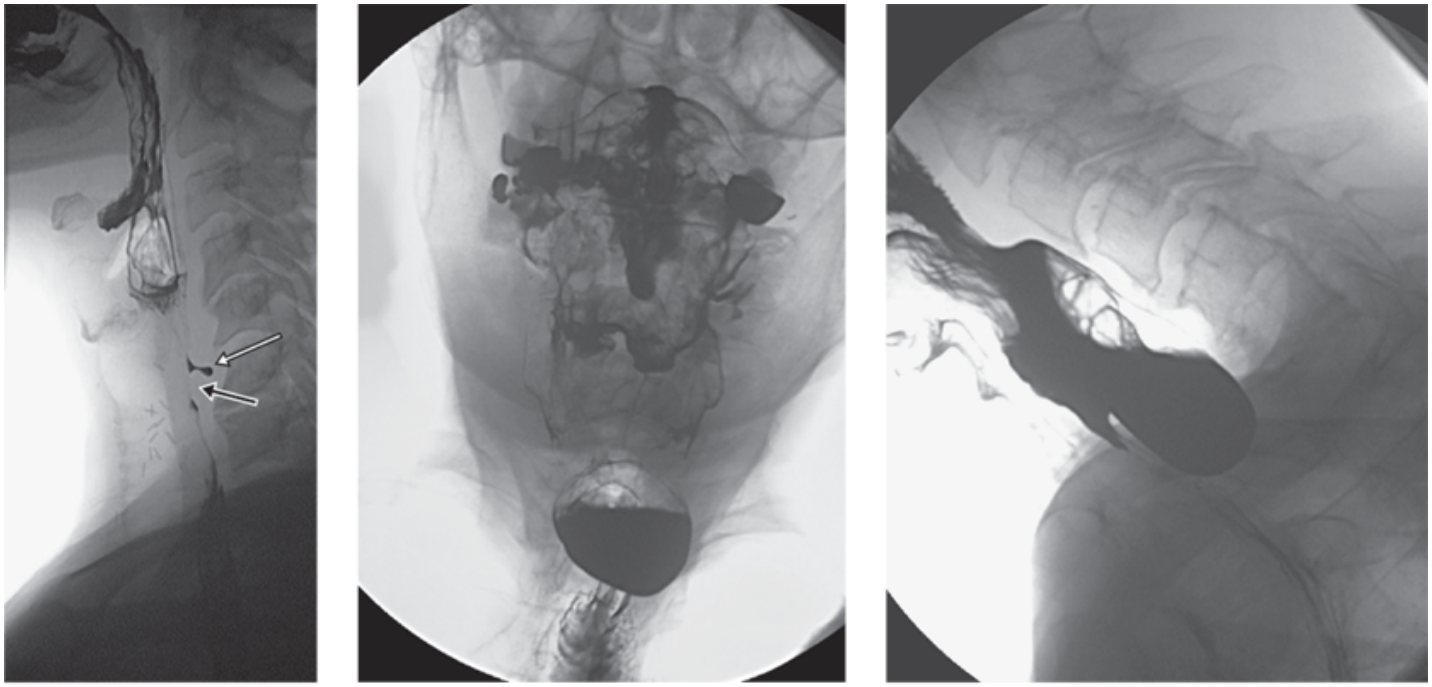
Web-like constrictions higher in the esophagus can be of congenital or inflammatory origin. Asymptomatic cervical esophageal webs are demonstrated in about 10% of people and typically originate along the anterior aspect of the esophagus. When circumferential, they can cause intermittent dysphagia to solids similar to Schatzki rings and are similarly treated with dilation. The combination of symptomatic proximal esophageal webs and iron-deficiency anemia in middle-aged women constitutes Plummer-Vinson syndrome.

DIVERTICULA

Esophageal diverticula are categorized by location with the most common being epiphrenic, hypopharyngeal (Zenker's), and midesophageal. Epiphrenic and Zenker's diverticula are false diverticula involving herniation of the mucosa and submucosa through the muscular layer of the esophagus. These lesions result from increased intraluminal pressure associated with distal obstruction. In the case of Zenker's, the obstruction is a stenotic cricopharyngeus muscle (upper esophageal sphincter), and the hypopharyngeal herniation most commonly occurs in an area of natural weakness proximal to the cricopharyngeus known as *Killian's triangle* (Fig. 316-3). Small Zenker's diverticula are usually asymptomatic, but when they enlarge sufficiently to retain food and saliva they can be associated with dysphagia, halitosis, and aspiration. Treatment is by surgical diverticulectomy and cricopharyngeal myotomy or a marsupialization procedure in which an endoscopic stapling device is used to divide the cricopharyngeus.

FIGURE 316-3

Examples of small (*A*) and large (*B, C*) Zenker's diverticula arising from Killian's triangle in the distal hypopharynx. Smaller diverticula are evident only during the swallow, whereas larger ones retain food and fluid.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Epiphrenic diverticula are often associated with achalasia, esophageal hypercontractile disorders, or a distal esophageal stricture. Midesophageal diverticula may be caused by traction from adjacent inflammation (classically tuberculosis) in which case they are true diverticula involving all layers of the esophageal wall, or by pulsion associated with esophageal motor disorders. Midesophageal and epiphrenic diverticula are usually asymptomatic until they enlarge sufficiently to retain food and cause dysphagia and regurgitation. Symptoms attributable to the diverticula tend to correlate more with the underlying esophageal disorder than the size of the diverticula. Large diverticula can be removed surgically, usually in conjunction with a myotomy if the underlying motility disorder is identified. Diffuse intramural esophageal pseudodiverticulosis is a rare entity that results from dilatation of the excretory ducts of submucosal esophageal glands **(Fig. 316-4)**. Esophageal candidiasis and proximal esophageal strictures are commonly found in association with this disorder.

FIGURE 316-4

Intramural esophageal pseudodiverticulosis associated with chronic obstruction. Invaginations of contrast into the esophageal wall outline deep esophageal glands.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TUMORS

Esophageal cancer occurs in about 4.5:100,000 people in the United States with the associated mortality being only slightly less at 4.4:100,000. It is about 10 times less common than colorectal cancer but kills about one-quarter as many patients. These statistics emphasize both the rarity and lethality of esophageal cancer. One notable trend is the shift of dominant esophageal cancer type from squamous cell to adenocarcinoma, strongly linked to reflux disease and Barrett's metaplasia. Other distinctions between cell types are the predilection for adenocarcinoma to affect the distal esophagus in white males and squamous cell to affect the more proximal esophagus in black males with the added risk factors of smoking, alcohol consumption, caustic injury, and human papilloma virus infection **(Chap. 76)**.

The typical presentation of esophageal cancer is of progressive solid food dysphagia and weight loss. Associated symptoms may include odynophagia, iron deficiency, and, with midesophageal tumors, hoarseness from left recurrent laryngeal nerve injury. Generally, these are indications of locally invasive or even metastatic disease manifest by tracheoesophageal fistulas and vocal cord paralysis. Even when detected as a small lesion, esophageal cancer has poor survival because of the abundant esophageal lymphatics leading to regional lymph node metastases.

Benign esophageal tumors are uncommon and usually discovered incidentally. In decreasing frequency of occurrence, cell types include leiomyoma, fibrovascular polyps, squamous papilloma, granular cell tumors, lipomas, neurofibromas, and inflammatory fibroid polyps. These generally become symptomatic only when they are associated with dysphagia and merit removal only under the same circumstances.

CONGENITAL ANOMALIES

The most common congenital esophageal anomaly is esophageal atresia, occurring in about 1 in 5000 live births. Atresia can occur in several permutations, the common denominator being developmental failure of fusion between the proximal and distal esophagus associated with a tracheoesophageal fistula, most commonly with the distal segment excluded. Alternatively, there can be an H-type configuration in which esophageal fusion has occurred, but with a tracheoesophageal fistula. Esophageal atresia is usually recognized and corrected surgically within the first few days of life. Later life complications include dysphagia from anastomotic strictures or absent peristalsis and reflux, which can be severe. Less common developmental anomalies include congenital esophageal stenosis, webs, and duplications.

Dysphagia can also result from congenital abnormalities that cause extrinsic compression of the esophagus. In dysphagia lusoria, the esophagus is compressed by an aberrant right subclavian artery arising from the descending aorta and passing behind the esophagus. Alternatively, vascular rings may surround and constrict the esophagus.

Heterotopic gastric mucosa, also known as an esophageal inlet patch, is a focus of gastric type epithelium in the proximal cervical esophagus; the estimated prevalence is 4–5%. The inlet patch is thought to result from incomplete replacement of embryonic columnar epithelium with squamous epithelium. The majority of inlet patches are asymptomatic, but acid production can occur as most contain fundic type gastric epithelium with parietal cells.

ESOPHAGEAL MOTILITY DISORDERS

Esophageal motility disorders are diseases attributable to esophageal neuromuscular dysfunction commonly associated with dysphagia, chest pain, or heartburn. The major entities are achalasia, DES, and GERD. Motility disorders can also be secondary to broader disease processes as is the case with pseudoachalasia, Chagas' disease, and scleroderma. Not included in this discussion are diseases affecting the pharynx and proximal esophagus, the impairment of which is almost always part of a more global neuromuscular disease process.

ACHALASIA

Achalasia is a rare disease caused by loss of ganglion cells within the esophageal myenteric plexus with a population incidence estimated to be 1–3 per 100,000 and usually presenting between age 25 and 60. With long-standing disease, aganglionosis is noted. The disease involves both excitatory (cholinergic) and inhibitory (nitric oxide) ganglionic neurons. Functionally, inhibitory neurons mediate deglutitive LES relaxation and the sequential propagation of peristalsis. Their absence leads to impaired deglutitive LES relaxation and absent peristalsis. Increasing evidence suggests that the ultimate cause of ganglion cell degeneration in achalasia is an autoimmune process attributable to a latent infection with human herpes simplex virus 1 combined with genetic susceptibility.

Long-standing achalasia is characterized by progressive dilatation and sigmoid deformity of the esophagus with hypertrophy of the LES. Clinical manifestations may include dysphagia, regurgitation, chest pain, and weight loss. Most patients report solid and liquid food dysphagia. Regurgitation occurs when food, fluid, and secretions are retained in the dilated esophagus. Patients with advanced achalasia are at risk for bronchitis, pneumonia, or lung abscess from chronic regurgitation and aspiration. Chest pain is frequent early in the course of achalasia, thought to result from esophageal spasm. Patients describe a squeezing, pressure-like retrosternal pain, sometimes radiating to the neck, arms, jaw, and back. Paradoxically, some patients complain of heartburn

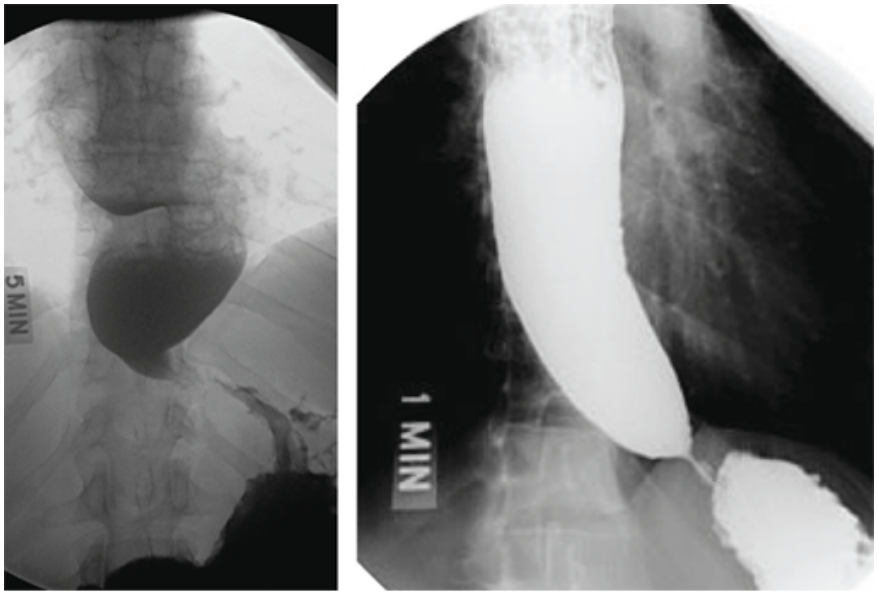
that may be a chest pain equivalent. Treatment of achalasia is less effective in relieving chest pain than it is in relieving dysphagia or regurgitation.

The differential diagnosis of achalasia includes DES, Chagas' disease, and pseudoachalasia. Chagas' disease is endemic in areas of central Brazil, Venezuela, and northern Argentina and spread by the bite of the reduviid (kissing) bug that transmits the protozoan, *Trypanosoma cruzi.* The chronic phase of the disease develops years after infection and results from destruction of autonomic ganglion cells throughout the body, including the heart, gut, urinary tract, and respiratory tract. Tumor infiltration, most commonly seen with carcinoma in the gastric fundus or distal esophagus, can mimic idiopathic achalasia. The resultant "pseudoachalasia" accounts for up to 5% of suspected cases and is more likely with advanced age, abrupt onset of symptoms (<1 year), and weight loss. Hence, endoscopy is a necessary part of the evaluation of achalasia. When the clinical suspicion for pseudoachalasia is high and endoscopy nondiagnostic, computed tomography (CT) scanning or EUS may be of value. Rarely, pseudoachalasia can result from a paraneoplastic syndrome with circulating antineuronal antibodies.

Achalasia is diagnosed by barium swallow x-ray and/or esophageal manometry; endoscopy has a relatively minor role other than to exclude pseudoachalasia. The barium swallow x-ray appearance is of a dilated esophagus with poor emptying, an air-fluid level, and tapering at the LES giving it a beak-like appearance **(Fig. 316-5)**. Occasionally, an epiphrenic diverticulum is observed. In long-standing achalasia, the esophagus may assume a sigmoid configuration. The diagnostic criteria for achalasia with esophageal manometry are impaired LES relaxation and absent peristalsis. High-resolution manometry has somewhat advanced this diagnosis; three subtypes of achalasia are differentiated based on the pattern of pressurization in the nonperistaltic esophagus **(Fig. 316-6)**. Because manometry identifies early disease before esophageal dilatation and food retention, it is the most sensitive diagnostic test.

FIGURE 316-5

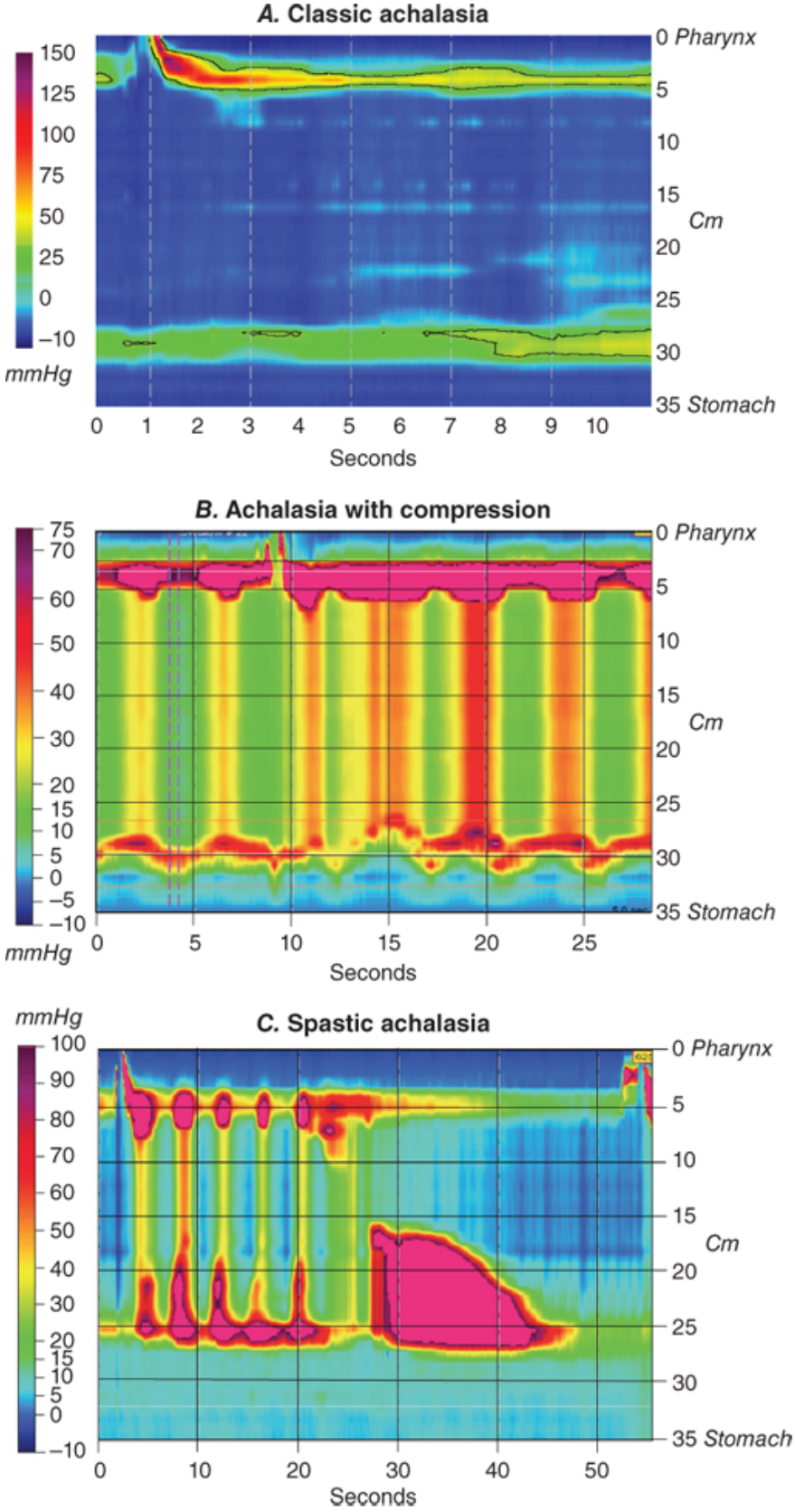
Achalasia with esophageal dilatation, tapering at the gastro-esophageal junction, and an air-fluid level within the esophagus. The example on the *left* shows sigmoid deformity with very advanced disease.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 316-6

Three subtypes of achalasia: classic (*A*), with esophageal compression (*B*), and spastic achalasia (*C*) imaged with pressure topography. All are characterized by impaired lower esophageal sphincter (LES) relaxation and absent peristalsis. However, classic achalasia has minimal pressurization of the esophageal body, whereas substantial fluid pressurization is observed in achalasia with esophageal compression, and spastic esophageal contractions are observed with spastic achalasia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

There is no known way of preventing or reversing achalasia. Therapy is directed at reducing LES pressure so that gravity and esophageal pressurization promote esophageal emptying. Peristalsis does not recover. However, in many instances, remnants of peristalsis masked by esophageal pressurization and dilatation prior to therapy are demonstrable following effective treatment. LES pressure can be reduced by pharmacologic therapy, pneumatic balloon dilation, or surgical myotomy. No large, controlled

trials of the therapeutic alternatives exist, and the optimal approach is debated. Pharmacologic therapies are relatively ineffective but are often used as temporizing therapies. Nitrates or calcium channel blockers are administered before eating, advising caution because of their effects on blood pressure. Botulinum toxin, injected into the LES under endoscopic guidance, inhibits acetylcholine release from nerve endings and improves dysphagia in about 66% of cases for at least 6 months. Sildenafil and alternative phosphodiesterase inhibitors effectively decrease LES pressure, but practicalities limit their clinical use in achalasia.

The only durable therapies for achalasia are pneumatic dilation and Heller myotomy. Pneumatic dilation, with a reported efficacy ranging from 32 to 98%, is an endoscopic technique using a noncompliant, cylindrical balloon dilator positioned across the LES and inflated to a diameter of 3–4 cm. The major complication is perforation with a reported incidence of 0.5–5%. The most common surgical procedure for achalasia is laparoscopic Heller myotomy, usually performed in conjunction with an antireflux procedure (partial fundoplication); good to excellent results are reported in 62–100% of cases. A European-randomized controlled trial demonstrated an equivalent response rate of ~90% for both pneumatic dilation and laparoscopic Heller myotomy at 5-year follow-up. Occasionally, patients with advanced disease fail to respond to pneumatic dilation or Heller myotomy. In such refractory cases, esophageal resection with gastric pull-up or interposition of a segment of transverse colon may be the only option other than gastrostomy feeding.

An endoscopic approach to LES myotomy has been introduced, referred to as per oral esophageal myotomy (POEM). This technique involves the creation of a submucosal tunnel within the esophageal wall through which the circular muscle of the LES and distal esophagus are transected with electrocautery. Short-term studies of efficacy have been favorable. Potential advantages over the conventional laparoscopic approach include avoidance of surgical disruption of the diaphragmatic hiatus and more rapid recovery.

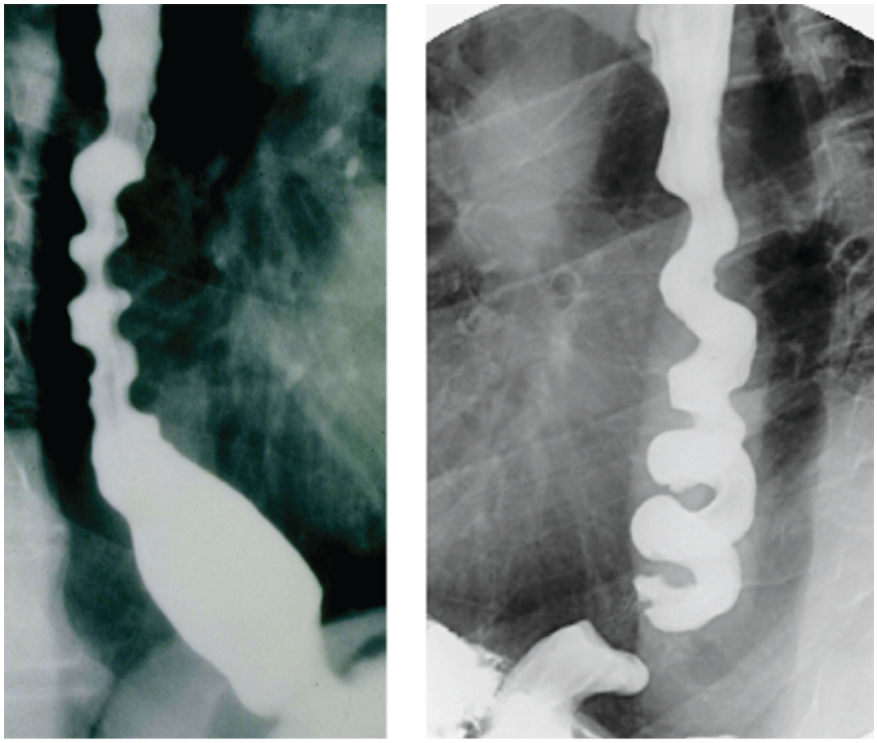
In untreated or inadequately treated achalasia, esophageal dilatation predisposes to stasis esophagitis. Prolonged stasis esophagitis is the likely explanation for the association between achalasia and esophageal squamous cell cancer. Tumors develop after years of achalasia, usually in the setting of extreme esophageal dilatation with the overall squamous cell cancer risk increased seventeenfold compared to controls.

DIFFUSE ESOPHAGEAL SPASM

DES is manifested by episodes of dysphagia and chest pain attributable to abnormal esophageal contractions with normal deglutitive LES relaxation. Beyond that, there is little consensus. The pathophysiology and natural history of DES are ill defined. Radiographically, DES has been characterized by tertiary contractions or a “corkscrew esophagus” **(Fig. 316-7)**, but in many instances, these abnormalities are actually indicative of achalasia. Manometrically, a variety of defining features have been proposed including uncoordinated (“spastic”) activity in the distal esophagus, spontaneous and repetitive contractions, or high-amplitude and prolonged contractions. The current consensus, derived from high-resolution manometry studies, is to define spasm by the occurrence of contractions in the distal esophagus with short latency relative to the time of the pharyngeal contraction, a dysfunction indicative of impairment of inhibitory myenteric plexus neurons. When defined in this restrictive fashion **(Fig. 316-8)**, DES is actually much less common than achalasia.

FIGURE 316-7

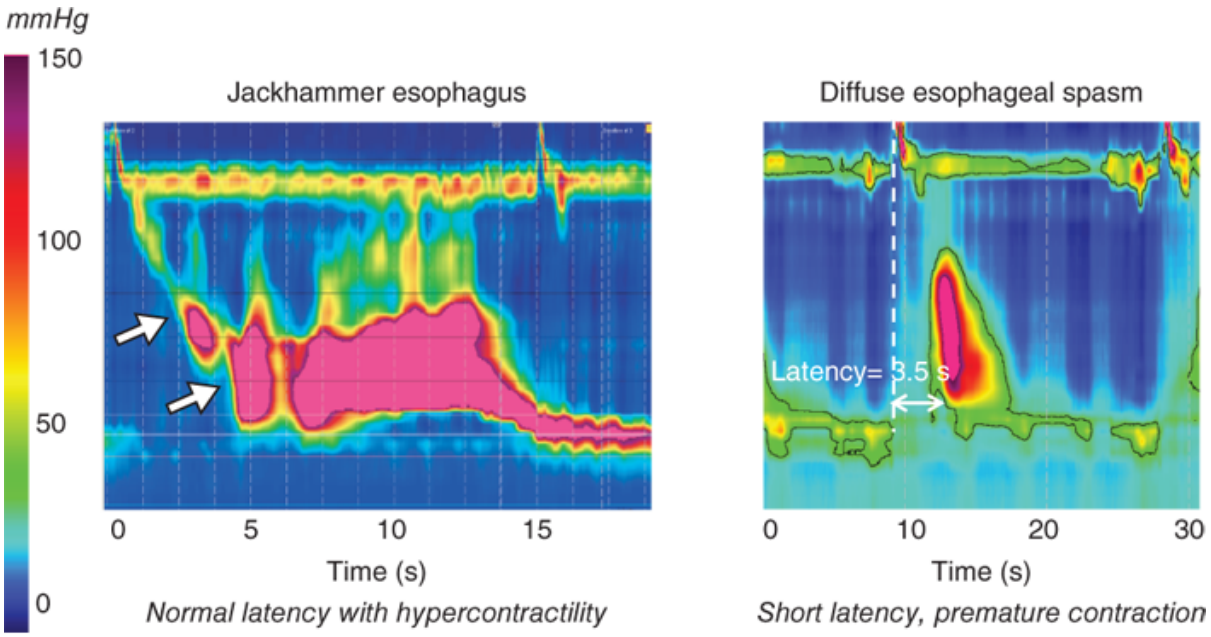
Diffuse esophageal spasm. The characteristic “corkscrew” esophagus results from spastic contraction of the circular muscle in the esophageal wall; more precisely, this is actually a helical array of muscle. These findings are also seen with spastic achalasia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 316-8

Esophageal pressure topography of the two major variants of esophageal spasm: jackhammer esophagus (*left*) and diffuse esophageal spasm (*right*). Jackhammer esophagus is defined by the extraordinarily vigorous and repetitive contractions with normal peristaltic onset and normal latency of the contraction. Diffuse esophageal spasm is similar but primarily defined by a short latency (premature) contraction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Esophageal chest pain closely mimics angina pectoris. Features suggesting esophageal pain include pain that is nonexertional, prolonged, interrupts sleep, meal-related, relieved with antacids, and accompanied by heartburn, dysphagia, or regurgitation. However, all of these features exhibit overlap with cardiac pain, which still must be the primary consideration. Furthermore, even within the spectrum of esophageal diseases, both chest pain and dysphagia are also characteristic of peptic or infectious

esophagitis. Only after these more common entities have been excluded by evaluation and/or treatment should a diagnosis of DES be pursued.

Although DES is diagnosed by manometry, endoscopy is useful to identify alternative structural and inflammatory lesions that may cause chest pain. Radiographically, a “corkscrew esophagus,” “rosary bead esophagus,” pseudodiverticula, or curling can be indicative of DES, but these are also found with spastic achalasia. Given these vagaries of defining DES, and the resultant heterogeneity of patients identified for inclusion in therapeutic trials, it is not surprising that trial results have been disappointing. Only small, uncontrolled trials exist, reporting response to nitrates, calcium channel blockers, hydralazine, botulinum toxin, and anxiolytics. The only controlled trial showing efficacy was with an anxiolytic. Surgical therapy (long myotomy or even esophagectomy) should be considered only with severe weight loss or unbearable pain. These indications are extremely rare.

NONSPECIFIC MANOMETRIC FINDINGS

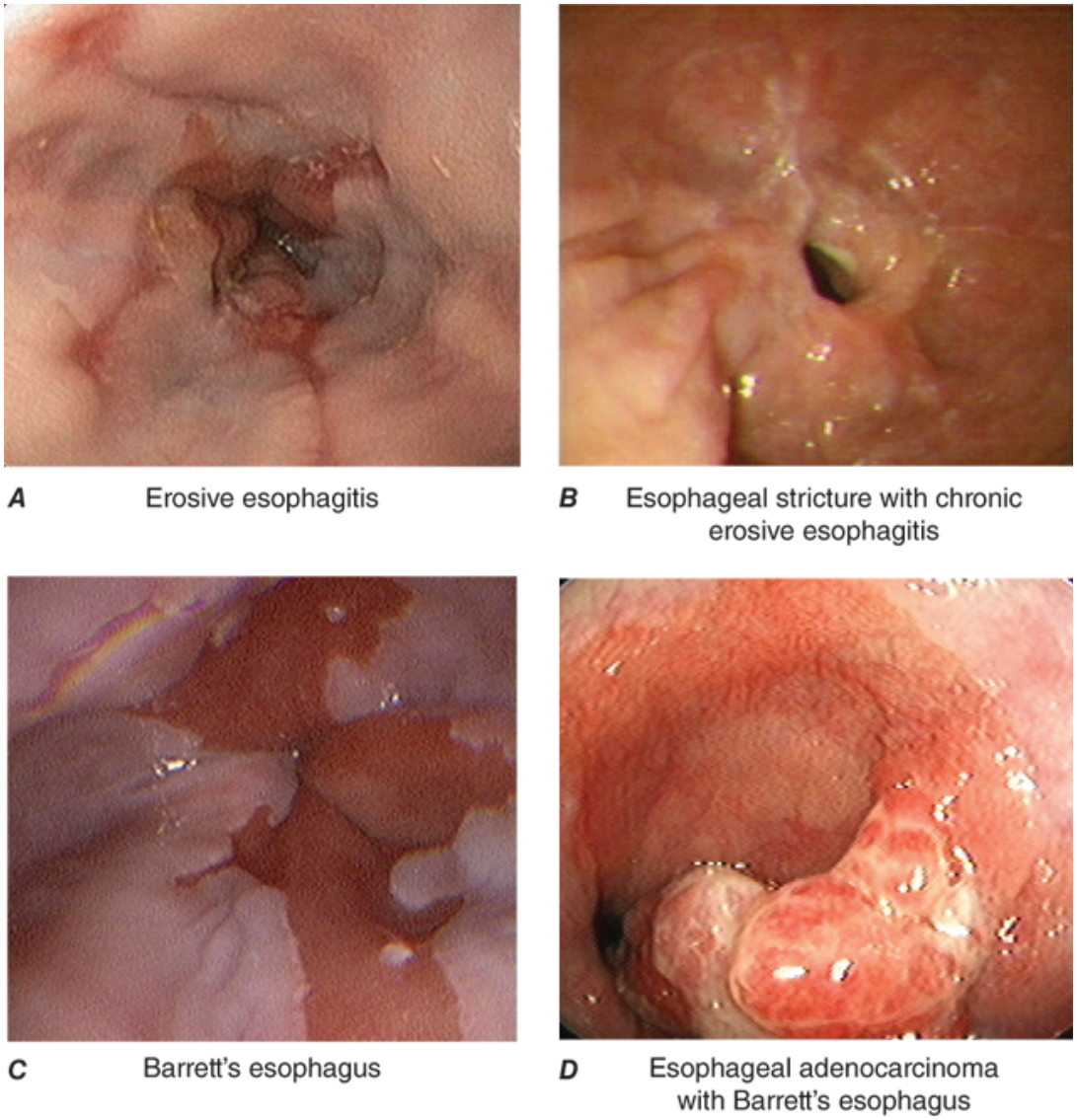
Manometric studies done to evaluate chest pain and/or dysphagia often report minor abnormalities (e.g., hypertensive or hypotensive peristalsis, hypertensive LES) that are insufficient to diagnose either achalasia or DES. These findings are of unclear significance. Reflux and psychiatric diagnoses, particularly anxiety and depression, are common among such individuals. A lower visceral pain threshold and symptoms of irritable bowel syndrome are noted in more than half of such patients. Consequently, therapy for these individuals should either target the most common esophageal disorder, GERD, or more global conditions such as depression or somatization neurosis that are found to be coexistent.

GASTROESOPHAGEAL REFLUX DISEASE

The current conception of GERD is to encompass a family of conditions with the commonality that they are caused by gastroesophageal reflux resulting in either troublesome symptoms or an array of potential esophageal and extraesophageal manifestations. It is estimated that 10–15% of adults in the United States are affected by GERD, although such estimates are based only on population studies of self-reported chronic heartburn. With respect to the esophagus, the spectrum of injury includes esophagitis, stricture, Barrett’s esophagus, and adenocarcinoma **(Fig. 316-9)**. Of particular concern is the rising incidence of esophageal adenocarcinoma, an epidemiologic trend that parallels the increasing incidence of GERD. There were about 8000 incident cases of esophageal adenocarcinoma in the United States in 2013 (half of all esophageal cancers); it is estimated that this disease burden has increased two- to sixfold in the last 20 years.

FIGURE 316-9

Endoscopic appearance of (*A*) peptic esophagitis, (*B*) a peptic stricture, (*C*) Barrett’s metaplasia, and (*D*) adenocarcinoma developing within an area of Barrett’s esophagus.



A Erosive esophagitis

B Esophageal stricture with chronic erosive esophagitis

C Barrett's esophagus

D Esophageal adenocarcinoma with Barrett's esophagus

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOPHYSIOLOGY

The best-defined subset of GERD patients, albeit a minority overall, have esophagitis. Esophagitis occurs when refluxed gastric acid and pepsin cause necrosis of the esophageal mucosa causing erosions and ulcers. Note that some degree of gastroesophageal reflux is normal, physiologically intertwined with the mechanism of belching (transient LES relaxation), but esophagitis results from excessive reflux, often accompanied by impaired clearance of the refluxed gastric juice. Restricting reflux to that which is physiologically intended depends on the anatomic and physiologic integrity of the esophagogastric junction, a complex sphincter comprised of both the LES and the surrounding crural diaphragm. Three dominant mechanisms of esophagogastric junction incompetence are recognized: (1) transient LES relaxations (a vagovagal reflex in which LES relaxation is elicited by gastric distention), (2) LES hypotension, or (3) anatomic distortion of the esophagogastric junction inclusive of hiatus hernia. Of note, the third factor, esophagogastric junction anatomic disruption, is both significant unto itself and also because it interacts with the first two mechanisms. Transient LES relaxations account for about 90% of reflux in normal subjects or GERD patients without hiatus hernia, but patients with hiatus hernia have a more heterogeneous mechanistic profile. Factors tending to exacerbate reflux regardless of mechanism are abdominal obesity, pregnancy, gastric hypersecretory states, delayed gastric emptying, disruption of esophageal peristalsis, and gluttony.

After acid reflux, peristalsis returns the refluxed fluid to the stomach and acid clearance is completed by titration of the residual acid by bicarbonate contained in swallowed saliva. Consequently, two causes of prolonged acid clearance are impaired peristalsis and reduced salivation. Impaired peristaltic emptying can be attributable to disrupted peristalsis or superimposed reflux associated with a hiatal hernia. With superimposed reflux, fluid retained within a sliding hiatal hernia refluxes back into the esophagus during swallow-related LES relaxation, a phenomenon that does not normally occur.

Inherent in the pathophysiologic model of GERD is that gastric juice is harmful to the esophageal epithelium. However, gastric acid hypersecretion is usually not a dominant factor in the development of esophagitis. An obvious exception is with Zollinger-Ellison syndrome, which is associated with severe esophagitis in about 50% of patients. Another caveat is with chronic *Helicobacter pylori* gastritis, which may have a protective effect by inducing atrophic gastritis with concomitant hypoacidity. Pepsin, bile, and pancreatic enzymes within gastric secretions can also injure the esophageal epithelium, but their noxious properties are either lessened without an acidic environment or dependent on acidity for activation. Bile warrants attention because it persists in refluxate despite acid-suppressing medications. Bile can transverse the cell membrane, imparting severe cellular injury in a weakly acidic environment, and has also been invoked as a cofactor in the pathogenesis of Barrett's metaplasia and adenocarcinoma. Hence, the causticity of gastric refluxate extends beyond hydrochloric acid.

SYMPTOMS

Heartburn and regurgitation are the typical symptoms of GERD. Somewhat less common are dysphagia and chest pain. In each case, multiple potential mechanisms for symptom genesis operate that extend beyond the basic concepts of mucosal erosion and activation of afferent sensory nerves. Specifically, hypersensitivity and functional pain are increasingly recognized as cofactors. Nonetheless, the dominant clinical strategy is empirical treatment with acid inhibitors, reserving further evaluation for those who fail to respond. Important exceptions to this are patients with chest pain or persistent dysphagia, each of which may be indicative of more morbid conditions. With chest pain, cardiac disease must be carefully considered. In the case of persistent dysphagia, chronic reflux can lead to the development of a peptic stricture or adenocarcinoma, each of which benefits from early detection and/or specific therapy.

Extraesophageal syndromes with an established association to GERD include chronic cough, laryngitis, asthma, and dental erosions. A multitude of other conditions including pharyngitis, chronic bronchitis, pulmonary fibrosis, chronic sinusitis, cardiac arrhythmias, sleep apnea, and recurrent aspiration pneumonia have proposed associations with GERD. However, in both cases, it is important to emphasize the word *association* as opposed to *causation*. In many instances, the disorders likely coexist because of shared pathogenetic mechanisms rather than strict causality. Potential mechanisms for extraesophageal GERD manifestations are either regurgitation with direct contact between the refluxate and supraesophageal structures or via a vagovagal reflex wherein reflux activation of esophageal afferent nerves triggers efferent vagal reflexes such as bronchospasm, cough, or arrhythmias.

DIFFERENTIAL DIAGNOSIS

Although generally quite characteristic, symptoms from GERD need to be distinguished from symptoms related to infectious, pill, or eosinophilic esophagitis (EoE), peptic ulcer disease, dyspepsia, biliary colic, coronary artery disease, and esophageal motility disorders. It is especially important that coronary artery disease be given early consideration because of its potentially lethal implications. The remaining elements of the differential diagnosis can be addressed by endoscopy, upper gastrointestinal series, or esophageal manometry as appropriate. Erosive esophagitis at the esophagogastric junction is the endoscopic hallmark of GERD, but identified in only about one third of patients with GERD. The distinction among etiologies of esophagitis is readily made by endoscopic appearance but mucosal biopsies may be helpful to evaluate for infectious or eosinophilic inflammation. In terms of endoscopic appearance, the ulcerations seen in peptic esophagitis are usually solitary and distal, whereas infectious ulcerations are punctate and diffuse. EoE characteristically exhibits multiple esophageal rings, linear furrows, white punctate exudate, and strictures. Esophageal ulcerations from pill esophagitis are usually singular and deep at points of luminal narrowing, especially near the carina, with sparing of the distal esophagus.

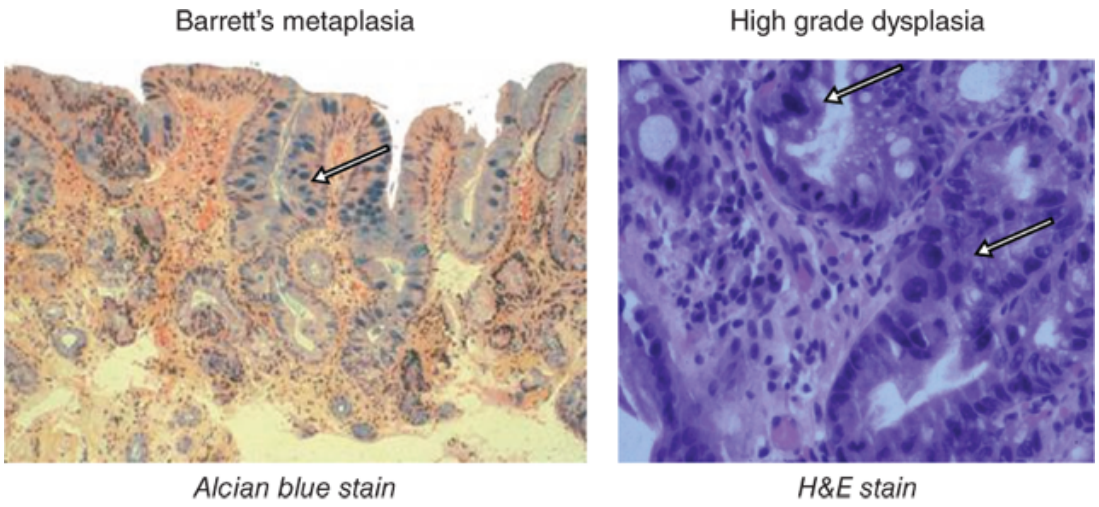
COMPLICATIONS

The complications of GERD are related to chronic esophagitis (bleeding and stricture) and the relationship between GERD and esophageal adenocarcinoma. However, both erosive esophagitis and peptic strictures have become increasingly rare in the era of potent antisecretory medications. Conversely, the most severe histologic consequence of GERD is Barrett's metaplasia with the associated risk of esophageal adenocarcinoma, and the incidence of these lesions has increased, not decreased, in the era of potent acid suppression. Barrett's metaplasia, endoscopically recognized by tongues of salmon-colored mucosa extending proximally from the gastroesophageal junction (Fig. 316-9) or histopathologically by the finding of specialized columnar metaplasia, is associated with a significantly increased risk for development of esophageal adenocarcinoma.

Barrett's metaplasia can progress to adenocarcinoma through the intermediate stages of low- and high-grade dysplasia **(Fig. 316-10)**. Owing to this risk, areas of Barrett's and especially any included areas of mucosal irregularity should be carefully inspected and extensively biopsied. The rate of cancer development is estimated at 0.1–0.3% per year, but vagaries in definitional criteria and of the extent of Barrett's metaplasia requisite to establish the diagnosis have contributed to variability and inconsistency in this risk assessment. The group at greatest risk is obese white males in their sixth decade of life. However, despite common practice, the utility of endoscopic screening and surveillance programs intended to control the adenocarcinoma risk has not been established. Also of note, no high-level evidence confirms that aggressive antisecretory therapy or antireflux surgery causes regression of Barrett's esophagus or prevents adenocarcinoma.

FIGURE 316-10

Histopathology of Barrett's metaplasia and Barrett's with high-grade dysplasia. H&E, hematoxylin and eosin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although the management of Barrett's esophagus remains controversial, the finding of dysplasia in Barrett's, particularly high-grade dysplasia, mandates further intervention. In addition to the high rate of progression to adenocarcinoma, there is also a high prevalence of unrecognized coexisting cancer with high-grade dysplasia. Treatment recommendations for Barrett's esophagus with high-grade dysplasia have evolved over the past several years. Historically, esophagectomy was the gold standard treatment for high-grade dysplasia. However, esophagectomy has a mortality ranging from 3 to 10%, along with substantial morbidity and recent prospective studies have demonstrated the efficacy of mucosal ablation therapy with substantially less morbidity and essentially no mortality. Consequently, current societal guidelines endorse endoscopic mucosal ablation therapies for the management of high-grade dysplasia.

TREATMENT

TREATMENT

Lifestyle modifications are routinely advocated as GERD therapy. Broadly speaking, these fall into three categories: (1) avoidance of foods that reduce LES pressure, making them “refluxogenic” (these commonly include fatty foods, alcohol, spearmint, peppermint, and possibly coffee and tea); (2) avoidance of acidic foods that are inherently irritating (citrus fruits, tomato-based foods); and (3) adoption of behaviors to minimize reflux and/or heartburn. In general, minimal evidence supports the efficacy of these measures. However, clinical experience dictates that subsets of patients are benefitted by specific recommendations, based on their individual history and symptom profile. A patient with sleep disturbance from nighttime heartburn is more likely to benefit from elevation of the head of the bed and avoidance of eating before retiring. The most broadly applicable recommendation is for weight reduction. Even though the benefit with respect to reflux cannot be assured, the strong epidemiologic relationship between body mass index and GERD and the secondary health gains of weight reduction is beyond dispute.

The dominant pharmacologic approach to GERD management is with inhibitors of gastric acid secretion, and abundant data support the effectiveness of this approach. Pharmacologically reducing the acidity of gastric juice does not prevent reflux, but it ameliorates reflux symptoms and allows esophagitis to heal. The hierarchy of effectiveness among pharmaceuticals parallels their antisecretory potency. Proton pump inhibitors (PPIs) are more efficacious than histamine$_2$ receptor antagonists (H_2RAs), and both are superior to placebo. No major differences exist among PPIs, and only modest gain is achieved by increased dosage.

Paradoxically, the perceived frequency and severity of heartburn correlate poorly with the presence or severity of esophagitis. When GERD treatments are assessed in terms of resolving heartburn, both efficacy and differences among pharmaceuticals are less clear-cut than with the objective of healing esophagitis. Although the same overall hierarchy of effectiveness exists, observed efficacy rates are lower and vary widely, likely reflecting patient heterogeneity.

Reflux symptoms tend to be chronic, irrespective of esophagitis. Thus, a common management strategy is indefinite treatment with PPIs or H_2RAs as necessary for symptom control. The side effects of PPI therapy are generally minimal. Rare cases of interstitial nephritis and severe, reversible hypomagnesemia have been reported. Vitamin B_{12} and iron absorption may be compromised and susceptibility to enteric infections, particularly *Clostridium difficile* colitis, increased with treatment. Observational data have also noted an association between PPI exposure and renal disease, dementia, and cardiovascular disease, but the hazard ratios reported in these studies were small and potential for unrecognized residual confounding bias was substantial. Population studies have also suggested a slight increased risk of bone fracture with chronic PPI use suggesting an impairment of calcium absorption, but prospective studies have failed to corroborate this. Nonetheless, as with any medication, PPI dosage should be minimized to that necessary for the clinical indication.

Laparoscopic Nissen fundoplication, wherein the proximal stomach is wrapped around the distal esophagus to create an antireflux barrier, is a surgical alternative to the management of chronic GERD. Just as with PPI therapy, evidence on the utility of fundoplication is strongest for treating esophagitis, and controlled trials suggest similar efficacy to PPI therapy. However, the benefits of fundoplication must be weighed against potential deleterious effects, including surgical morbidity and mortality, postoperative dysphagia, failure or breakdown requiring reoperation, an inability to belch, and increased bloating, flatulence, and bowel symptoms after surgery.

EOSINOPHILIC ESOPHAGITIS

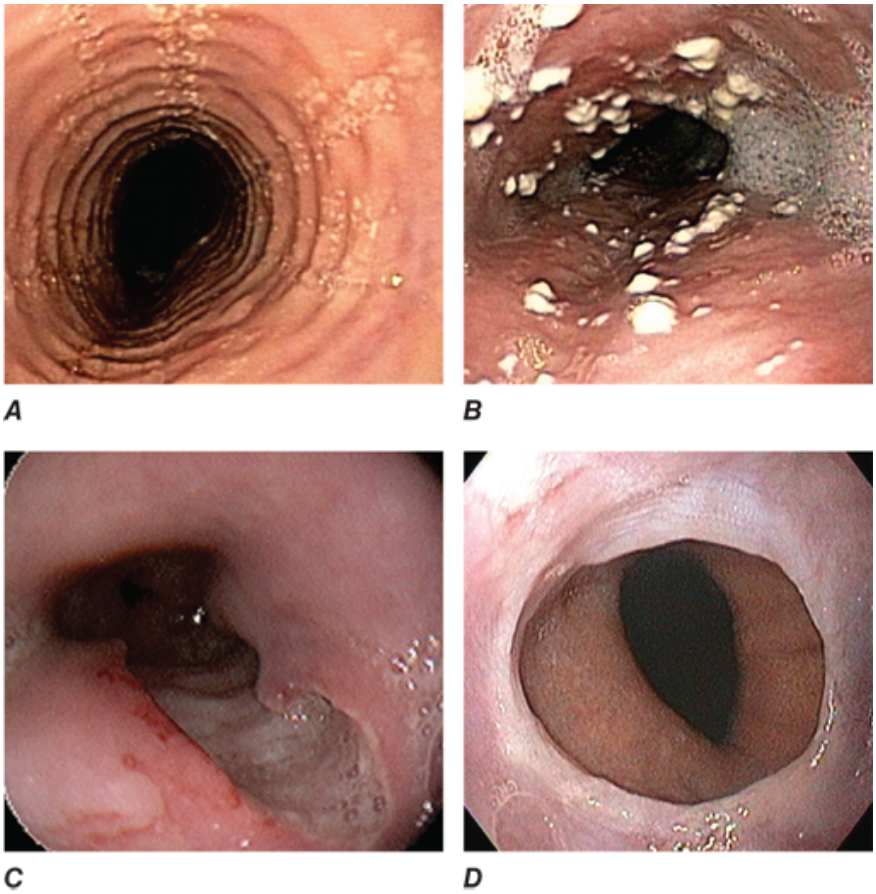
EoE is increasingly recognized in adults and children around the world. Current prevalence estimates in the United States identified 4–6 cases per 10,000 with a predilection for white males between 30 and 40 years of age. The increasing prevalence of EoE is attributable to a combination of an increasing incidence and a growing recognition of the condition. There is also an incompletely understood, but important, overlap between EoE and GERD that may confound the diagnosis of the disease.

EoE is diagnosed based on the combination of typical esophageal symptoms and esophageal mucosal biopsies demonstrating squamous epithelial eosinophil-predominant inflammation. Alternative etiologies of esophageal eosinophilia include GERD, drug hypersensitivity, connective tissue disorders, hypereosinophilic syndrome, Crohn's disease, and infection. Current evidence indicates that EoE is an immunologic disorder induced by antigen sensitization in susceptible individuals. Dietary factors play an important role in both the pathogenesis and treatment of EoE. Aeroallergens may also contribute, but the evidence is weaker. The natural history of EoE is unclear, but an increased risk of esophageal stricture development paralleling the duration of untreated disease has been noted.

EoE should be strongly considered in children and adults with dysphagia and esophageal food impactions. In preadolescent children, symptom presentations of EoE include chest or abdominal pain, nausea, vomiting, and food aversion. Other symptoms in adults may include atypical chest pain and heartburn, particularly heartburn that is refractory to PPI therapy. An atopic history of food allergy, asthma, eczema, or allergic rhinitis is present in the majority of patients. Peripheral blood eosinophilia is demonstrable in up to 50% of patients, but the specificity of this finding is problematic in the setting of concomitant atopy. The characteristic endoscopically identified esophageal findings include loss of vascular markings (edema), multiple esophageal rings, longitudinally oriented furrows, and punctate exudate **(Fig. 316-11)**. Histologic confirmation is made with the demonstration of esophageal mucosal eosinophilia (peak density ≥15 eosinophils per high-power field) **(Fig. 316-12)**. Complications of EoE include esophageal stricture, narrow-caliber esophagus, food impaction, and esophageal perforation.

FIGURE 316-11

Endoscopic features of (***A***) eosinophilic esophagitis (EoE), (***B***) *Candida* esophagitis, (***C***) giant ulcer associated with HIV, and (***D***) a Schatzki ring.

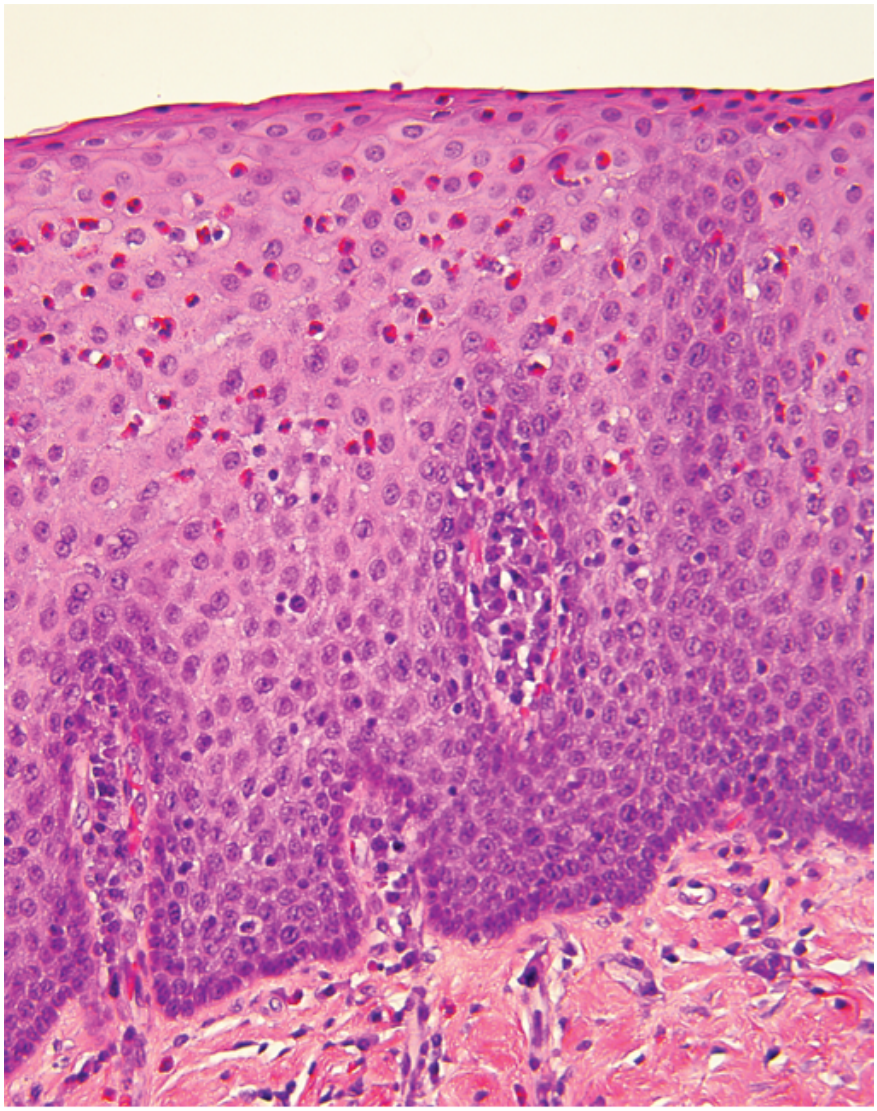


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 316-12

Histopathology of eosinophilic esophagitis (EoE) showing infiltration of the esophageal squamous epithelium with eosinophils. Additional features of basal cell hyperplasia and lamina propria fibrosis are present. Eosinophilic inflammation can also be seen

with gastroesophageal reflux disease.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The goals of EoE management are symptom control and the prevention of complications. Once esophageal eosinophilia is demonstrated, patients typically undergo a trial of PPI therapy as a means of excluding a contribution of GERD to the symptoms and esophageal mucosal inflammation. PPI-responsive esophageal eosinophilia, characterized by elimination of mucosal eosinophilia, occurs in 30–50% of cases of suspected EoE. Patients with persistent symptoms and eosinophilic inflammation following PPI therapy are subsequently considered for treatments such as elimination diets or swallowed topical glucocorticoids. Elemental formula diets are a highly effective therapy but are limited by palatability. Notably, allergy testing by means of either serum IgE or skin prick testing has demonstrated poor sensitivity and specificity in the identification of foods that incite the esophageal inflammatory response. Allergy testing combining skin prick and atopy patch testing has been effective in children with EoE, but additional validation is needed. Empiric elimination of common food allergies (milk, wheat, egg, soy, nuts, and seafood) followed by systematic reintroduction has been an effective diet therapy in both children and adults with EoE. The intent of the elimination diet approach is the identification of a single or a small number of food triggers. Swallowed, topical glucocorticoids (e.g., fluticasone propionate or budesonide) are highly effective, but recurrence of disease is common following the cessation of short-term therapy. Systemic glucocorticoids are reserved for severely afflicted patients refractory to less morbid treatments. Esophageal dilation is very effective at relieving dysphagia in patients with fibrostenosis. Dilation should be approached conservatively because of the risk of deep, esophageal mural laceration or perforation in the stiff-walled esophagus that is characteristic of the disease.

INFECTIOUS ESOPHAGITIS

With the increased use of immunosuppression for organ transplantation as well as chronic inflammatory diseases and chemotherapy along with the AIDS epidemic, infections with *Candida* species, herpesvirus, and cytomegalovirus (CMV) have become relatively common. Although rare, infectious esophagitis also occurs among the non-immunocompromised, with herpes simplex and *Candida albicans* being the most common pathogens. Among AIDS patients, infectious esophagitis becomes more common as the CD4 count declines; cases are rare with a CD4 count >200 and common when <100. HIV itself may also be associated with a self-limited syndrome of acute esophageal ulceration with oral ulcers and a maculopapular skin rash at the time of seroconversion. Additionally, some patients with advanced disease have deep, persistent esophageal ulcers treated with oral glucocorticoids or thalidomide. However, with the widespread use of highly effective anti-viral therapies, a reduction in these HIV complications has been noted.

Regardless of the infectious agent, odynophagia is a characteristic symptom of infectious esophagitis; dysphagia, chest pain, and hemorrhage are also common. Odynophagia is uncommon with reflux esophagitis, so its presence should always raise suspicion of an alternative etiology.

CANDIDA ESOPHAGITIS

Candida is normally found in the throat, but can become pathogenic and produce esophagitis in a compromised host; *C. albicans* is most common. *Candida* esophagitis also occurs with esophageal stasis secondary to esophageal motor disorders and diverticula. Patients complain of odynophagia and dysphagia. If oral thrush is present, empirical therapy is appropriate, but co-infection is common, and persistent symptoms should lead to prompt endoscopy with biopsy, which is the most useful diagnostic evaluation. *Candida* esophagitis has a characteristic appearance of white plaques with friability. Rarely, *Candida* esophagitis is complicated by bleeding, perforation, stricture, or systemic invasion. Oral fluconazole (200–400 mg on the first day, followed by 100–200 mg daily) for 14–21 days is the preferred treatment. Patients refractory to fluconazole may respond to voriconazole, or posaconazole. Alternatively, poorly responsive patients or those who cannot swallow medications can be treated with an intravenous echinocandin.

HERPETIC ESOPHAGITIS

Herpes simplex virus type 1 or 2 may cause esophagitis. Vesicles on the nose and lips may coexist and are suggestive of a herpetic etiology. Varicella-zoster virus can also cause esophagitis in children with chickenpox or adults with zoster. The characteristic endoscopic findings are vesicles and small, punched-out ulcerations. Because herpes simplex infections are limited to squamous epithelium, biopsies from the ulcer margins are most likely to reveal the characteristic ground-glass nuclei, eosinophilic Cowdry's type A inclusion bodies, and giant cells. Culture or polymerase chain reaction (PCR) assays are helpful to identify acyclovir-resistant strains. Acyclovir (200 mg orally five times a day for 7–10 days) can be used for immunocompetent hosts, although the disease is typically self-limited after a 1- to 2-week period in such patients. Immunocompromised patients are treated with acyclovir (400 mg orally five times a day for 14–21 days), famciclovir (500 mg orally three times a day), or valacyclovir (1 g orally three times a day). In patients with severe odynophagia, intravenous acyclovir, 5 mg/kg every 8 h for 7–14 days, reduces this morbidity.

CYTOMEGALOVIRUS

CMV esophagitis occurs primarily in immunocompromised patients, particularly organ transplant recipients. CMV is usually activated from a latent stage. Endoscopically, CMV lesions appear as serpiginous ulcers in an otherwise normal mucosa, particularly in the distal esophagus. Biopsies from the ulcer bases have the greatest diagnostic yield for finding the pathognomonic large nuclear or cytoplasmic inclusion bodies. Immunohistology with monoclonal antibodies to CMV and in situ hybridization tests are useful for early diagnosis. Data on therapy for CMV esophagitis are limited. Treatment studies of CMV gastrointestinal disease have demonstrated effectiveness of both ganciclovir (5 mg/kg every 12 h intravenously) and

valganciclovir (900 mg orally every 12 h). Therapy is continued until healing, which may take 3–6 weeks. Maintenance therapy may be needed for patients with relapsing disease.

MECHANICAL TRAUMA AND IATROGENIC INJURY

ESOPHAGEAL PERFORATION

Most cases of esophageal perforation are from instrumentation of the esophagus or trauma. Alternatively, forceful vomiting or retching can lead to spontaneous rupture at the gastroesophageal junction (Boerhaave's syndrome). More rarely, corrosive esophagitis or neoplasms lead to perforation. Instrument perforation from endoscopy or nasogastric tube placement typically occurs in the hypopharynx or at the gastroesophageal junction. Perforation may also occur at the site of a stricture in the setting of endoscopic food disimpaction or esophageal dilation. Esophageal perforation causes pleuritic retrosternal pain that can be associated with pneumomediastinum and subcutaneous emphysema. Mediastinitis is a major complication of esophageal perforation, and prompt recognition is key to optimizing outcome. CT of the chest is most sensitive in detecting mediastinal air. Esophageal perforation is confirmed by a contrast swallow, usually Gastrografin followed by thin barium. Treatment includes nasogastric suction and parenteral broad-spectrum antibiotics with prompt surgical drainage and repair in noncontained leaks. Conservative therapy with NPO status and antibiotics without surgery may be appropriate in cases of contained perforation that are detected early. Endoscopic clipping or stent placement may be indicated in nonoperated iatrogenic perforations or nonoperable cases such as perforated tumors.

MALLORY-WEISS TEAR

Vomiting, retching, or vigorous coughing can cause a nontransmural tear at the gastroesophageal junction that is a common cause of upper gastrointestinal bleeding. Most patients present with hematemesis. Antecedent vomiting is the norm, but not always evident. Bleeding usually abates spontaneously, but protracted bleeding may respond to local epinephrine or cauterization therapy, endoscopic clipping, or angiographic embolization. Surgery is rarely needed.

RADIATION ESOPHAGITIS

Radiation esophagitis can complicate treatment for thoracic cancers, especially breast and lung, with the risk proportional to radiation dosage. Radiosensitizing drugs such as doxorubicin, bleomycin, cyclophosphamide, and cisplatin also increase the risk. Dysphagia and odynophagia may last weeks to months after therapy. The esophageal mucosa becomes erythematous, edematous, and friable. Submucosal fibrosis and degenerative tissue changes and stricturing may occur years after the radiation exposure. Radiation exposure in excess of 5000 cGy has been associated with increased risk of esophageal stricture. Treatment for acute radiation esophagitis is supportive. Chronic strictures are managed with esophageal dilation.

CORROSIVE ESOPHAGITIS

Caustic esophageal injury from ingestion of alkali or, less commonly, acid can be accidental or from attempted suicide. Absence of oral injury does not exclude possible esophageal involvement. Thus, early endoscopic evaluation is recommended to assess and grade the injury to the esophageal mucosa. Severe corrosive injury may lead to esophageal perforation, bleeding, stricture, and death. Glucocorticoids have not been shown to improve the clinical outcome of acute corrosive esophagitis and are not recommended. Healing of more severe grades of caustic injury is commonly associated with severe stricture formation and often requires repeated dilation.

PILL ESOPHAGITIS

Pill-induced esophagitis occurs when a swallowed pill fails to traverse the entire esophagus and lodges within the lumen. Generally, this is attributed to poor "pill taking habits": inadequate liquid with the pill or lying down immediately after taking a

pill. The most common location for the pill to lodge is in the mid-esophagus near the crossing of the aorta or carina. Extrinsic compression from these structures halts the movement of the pill or capsule. Since initially reported in 1970, more than 1000 cases of pill esophagitis have been reported, suggesting that this is not an unusual occurrence. A wide variety of medications are implicated with the most common being doxycycline, tetracycline, quinidine, phenytoin, potassium chloride, ferrous sulfate, nonsteroidal anti-inflammatory drugs (NSAIDs), and bisphosphonates.

Typical symptoms of pill esophagitis are the sudden onset of chest pain and odynophagia. Characteristically, the pain will develop over a period of hours or will awaken the individual from sleep. A classic history in the setting of ingestion of recognized pill offenders obviates the need for diagnostic testing in most patients. When endoscopy is performed, localized ulceration or inflammation is evident. Histologically, acute inflammation is typical. Chest CT imaging will sometimes reveal esophageal thickening consistent with transmural inflammation. Although the condition usually resolves within days to weeks, symptoms may persist for months and stricture can develop in severe cases. No specific therapy is known to hasten the healing process, but antisecretory medications are frequently prescribed to remove concomitant reflux as an aggravating factor. When healing results in stricture formation, dilation is indicated.

FOREIGN BODIES AND FOOD IMPACTION

Food or foreign bodies may lodge in the esophagus causing complete obstruction, which in turn can cause an inability to handle secretions (foaming at the mouth) and severe chest pain. Food impaction may occur due to peptic stricture, carcinoma, Schatzki ring, EoE, or simply inattentive eating. If it does not spontaneously resolve, impacted food can be removed endoscopically. Use of meat tenderizer enzymes to facilitate passage of a meat bolus is discouraged because of potential esophageal injury. Glucagon (1 mg IV) is sometimes tried before endoscopic dislodgement. After emergent treatment, patients should be evaluated for potential causes of the impaction with treatment rendered as indicated.

ESOPHAGEAL MANIFESTATIONS OF SYSTEMIC DISEASE

SCLERODERMA AND COLLAGEN VASCULAR DISEASES

Scleroderma esophagus (hypotensive LES and absent esophageal peristalsis) was initially described as a manifestation of scleroderma or other collagen vascular diseases and thought to be specific for these disorders. However, this nomenclature subsequently proved unfortunate and has been discarded because an estimated half of qualifying patients do not have an identifiable systemic disease, and reflux disease is often the only identifiable association. When scleroderma esophagus occurs as a manifestation of a collagen vascular disease, the histopathologic findings are of infiltration and destruction of the esophageal muscularis propria with collagen deposition and fibrosis. The pathogenesis of absent peristalsis and LES hypotension in the absence of a collagen vascular disease is unknown. Regardless of the underlying cause, the manometric abnormalities predispose patients to severe GERD due to inadequate LES barrier function combined with poor esophageal clearance of refluxed acid. Dysphagia may also be manifest but is generally mild and alleviated by eating in an upright position and using liquids to facilitate solid emptying.

DERMATOLOGIC DISEASES

A host of dermatologic disorders (pemphigus vulgaris, bullous pemphigoid, cicatricial pemphigoid, Behçet's syndrome, and epidermolysis bullosa) can affect the oropharynx and esophagus, particularly the proximal esophagus with blisters, bullae, webs, and strictures. Glucocorticoid treatment is usually effective. Erosive lichen planus, Stevens-Johnson syndrome, and graft-versus-host disease can also involve the esophagus. Esophageal dilation may be necessary to treat strictures.

FURTHER READING

Dellon ES et al: ACG Clinical Guideline: Evidence based approach to the diagnosis and management of esophageal eosinophilia and eosinophilic esophagitis. Am J Gastroenterol 108:679, 2013. [PubMed: 23567357]

Furuta GT, Katzka DA: Eosinophilic esophagitis. N Engl J Med 373:1640, 2015. [PubMed: 26488694]

Kahrilas PJ: Clinical practice. Gastroesophageal reflux disease. N Engl J Med 359:1700, 2008. [PubMed: 18923172]

Kahrilas PJ, Boeckxstaens G: The spectrum of achalasia: Lessons from studies of pathophysiology and high-resolution manometry. Gastroenterology 145:954, 2013. [PubMed: 23973923]

Kahrilas PJ et al: American Gastroenterological Association Institute technical review on the management of gastroesophageal reflux disease. Gastroenterology 135:1392, 2008. [PubMed: 18801365]

Pandolfino JE, Gawron AJ: Achalasia: A systematic review. JAMA 313:1841, 2015. [PubMed: 25965233]

Shaheen NJ et al: Diagnosis and Management of Barrett's Esophagus. Am J Gastroenterol 111:30, 2016. [PubMed: 26526079]

Spechler SJ, Souza RF: Barrett's esophagus. N Engl J Med 371:836, 2014. [PubMed: 25162890]

Chapter 317: Peptic Ulcer Disease and Related Disorders

John Del Valle

PEPTIC ULCER DISEASE

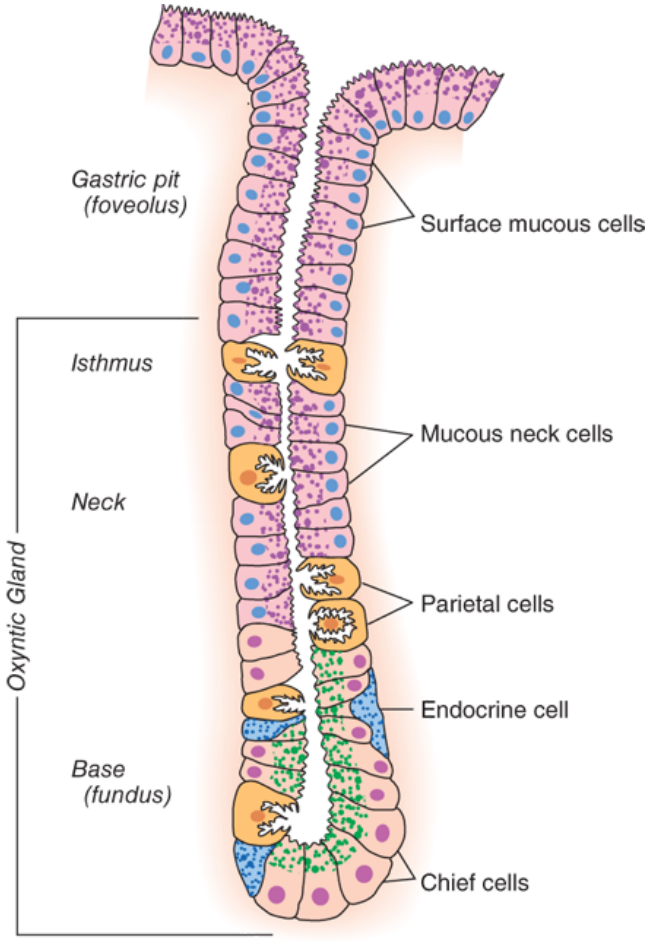
A *peptic ulcer* is defined as disruption of the mucosal integrity of the stomach and/or duodenum leading to a local defect or excavation due to active inflammation. Although burning epigastric pain exacerbated by fasting and improved with meals is a symptom complex associated with peptic ulcer disease (PUD), it is now clear that >90% patients with this symptom complex (dyspepsia) do not have ulcers and that the majority of patients with peptic ulcers may be asymptomatic. Ulcers occur within the stomach and/or duodenum and are often chronic in nature. Acid peptic disorders are very common in the United States, with 4 million individuals (new cases and recurrences) affected per year. Lifetime prevalence of PUD in the United States is ~12% in men and 10% in women. PUD significantly affects quality of life by impairing overall patient well-being and contributing substantially to work absenteeism. Moreover, an estimated 15,000 deaths per year occur as a consequence of complicated PUD. The financial impact of these common disorders has been substantial, with an estimated burden on direct and indirect health care costs of ~$6 billion per year in the United States, with $3 billion spent on hospitalizations, $2 billion on physician office visits, and $1 billion in decreased productivity and days lost from work.

GASTRIC PHYSIOLOGY

Gastric Anatomy

The gastric epithelial lining consists of rugae that contain microscopic gastric pits, each branching into four or five gastric glands made up of highly specialized epithelial cells. The makeup of gastric glands varies with their anatomic location. Glands within the gastric cardia comprise <5% of the gastric gland area and contain mucous and endocrine cells. The 75% of gastric glands are found within the oxyntic mucosa and contain mucous neck, parietal, chief, endocrine, enterochromaffin, and enterochromaffin-like (ECL) cells **(Fig. 317-1)**. Pyloric glands contain mucous and endocrine cells (including gastrin cells) and are found in the antrum.

FIGURE 317-1
Diagrammatic representation of the oxyntic gastric gland. *(Adapted from S Ito, RJ Winchester: J Cell Biol 16:541, 1963. © 1963 Ito and Winchester.)*

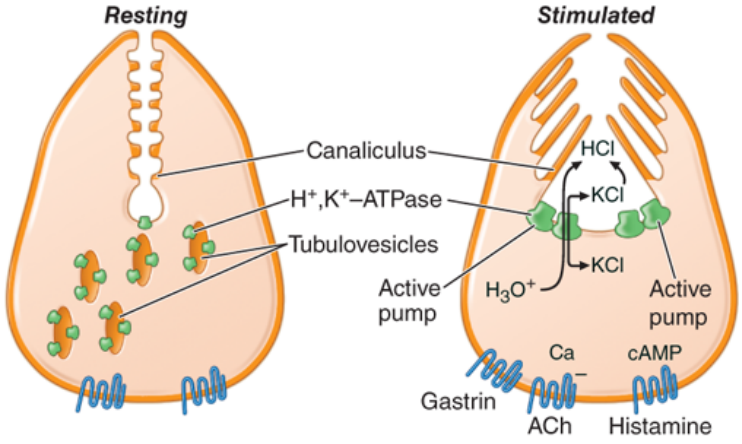


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The parietal cell, also known as the oxyntic cell, is usually found in the neck, or isthmus, or in the oxyntic gland. The resting, or unstimulated, parietal cell has prominent cytoplasmic tubulovesicles and intracellular canaliculi containing short microvilli along its apical surface (Fig. 317-2). H^+,K^+-adenosine triphosphatase (ATPase) is expressed in the tubulovesicle membrane; upon cell stimulation, this membrane, along with apical membranes, transforms into a dense network of apical intracellular canaliculi containing long microvilli. Acid secretion, a process requiring high energy, occurs at the apical canalicular surface. Numerous mitochondria (30–40% of total cell volume) generate the energy required for secretion.

FIGURE 317-2

Gastric parietal cell undergoing transformation after secretagogue-mediated stimulation. cAMP, cyclic adenosine monophosphate. *(Adapted from SJ Hersey, G Sachs: Physiol Rev 75:155, 1995.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

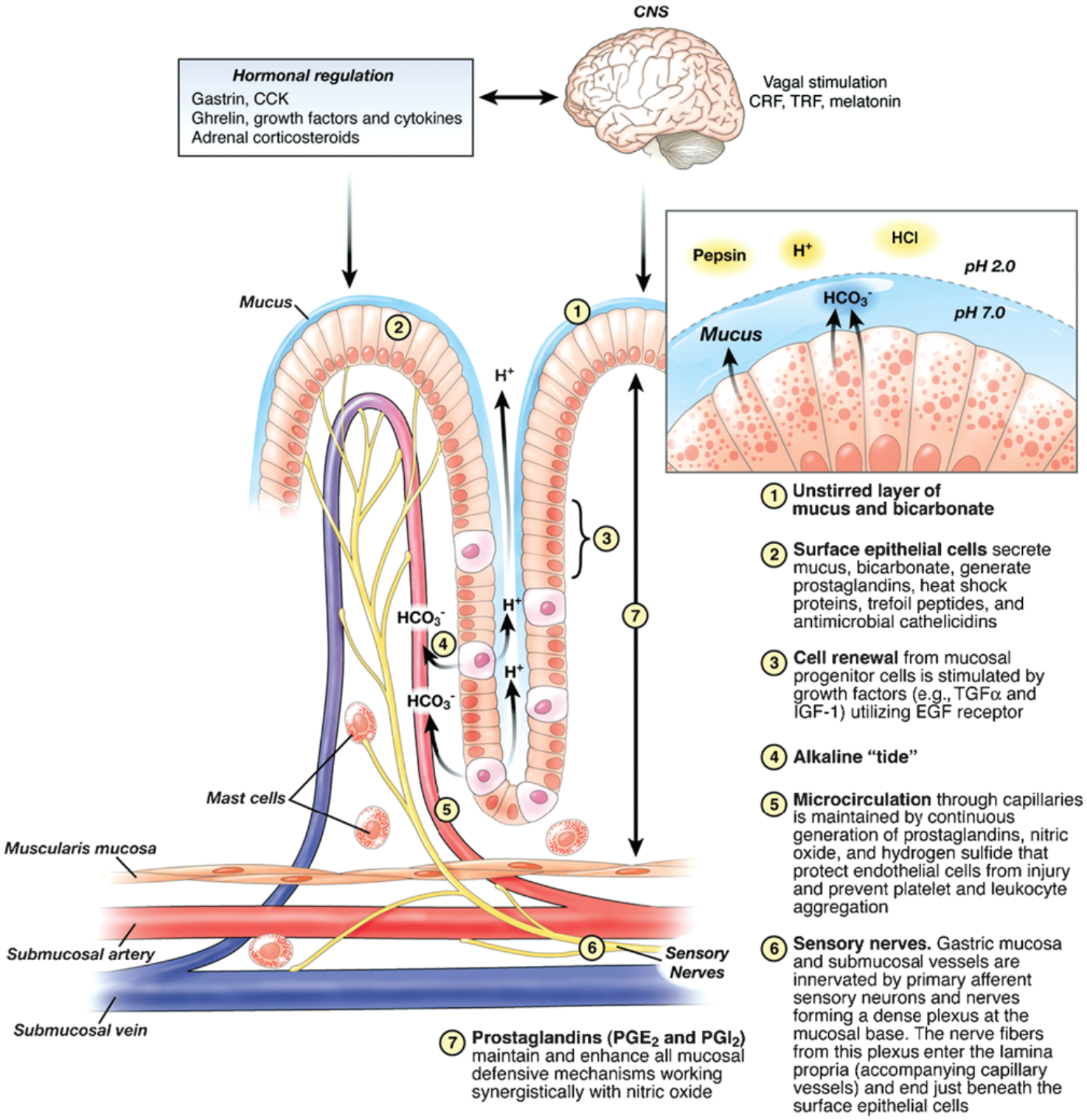
Gastroduodenal Mucosal Defense

The gastric epithelium is under constant assault by a series of endogenous noxious factors, including hydrochloric acid (HCl), pepsinogen/pepsin, and bile salts. In addition, a steady flow of exogenous substances such as medications, alcohol, and bacteria encounter the gastric mucosa. A highly intricate biologic system is in place to provide defense from mucosal injury and to repair any injury that may occur.

The mucosal defense system can be envisioned as a three-level barrier, composed of preepithelial, epithelial, and subepithelial elements **(Fig. 317-3)**. The first line of defense is a mucus-bicarbonate-phospholipid layer, which serves as a physicochemical barrier to multiple molecules, including hydrogen ions. Mucus is secreted in a regulated fashion by gastroduodenal surface epithelial cells. It consists primarily of water (95%) and a mixture of phospholipids and glycoproteins (mucin). The mucous gel functions as a nonstirred water layer impeding diffusion of ions and molecules such as pepsin. Bicarbonate, secreted in a regulated manner by surface epithelial cells of the gastroduodenal mucosa into the mucous gel, forms a pH gradient ranging from 1 to 2 at the gastric luminal surface and reaching 6–7 along the epithelial cell surface.

FIGURE 317-3

Components involved in providing gastroduodenal mucosal defense and repair. CCK, cholecystokinin; CRF, corticotropin-releasing factor; EGF, epidermal growth factor; HCl, hydrochloride; IGF, insulin-like growth factor; TGFα, transforming growth factor α; TRF, thyrotropin releasing factor. *(Modified and updated from A Tarnawski: Cellular and molecular mechanisms of mucosal defense and repair, in T Yoshikawa, T Arakawa [eds]: Bioregulation and Its Disorders in the Gastrointestinal Tract. Tokyo, Japan: Blackwell Science, 1998:3–17.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Surface epithelial cells provide the next line of defense through several factors, including mucus production, epithelial cell ionic transporters that maintain intracellular pH and bicarbonate production, and intracellular tight junctions. Surface epithelial cells generate heat shock proteins that prevent protein denaturation and protect cells from certain factors such as increased temperature, cytotoxic agents, or oxidative stress. Epithelial cells also generate trefoil factor family peptides and cathelicidins, which also play a role in surface cell protection and regeneration. If the preepithelial barrier were breached, gastric epithelial cells bordering a site of injury can migrate to restore a damaged region (*restitution*). This process occurs independent of cell division and requires uninterrupted blood flow and an alkaline pH in the surrounding environment. Several growth factors, including epidermal growth factor (EGF), transforming growth factor (TGF) α, and basic fibroblast growth factor (FGF), modulate the process of restitution. Larger defects that are not effectively repaired by restitution require cell proliferation. Epithelial cell regeneration is regulated by prostaglandins and growth factors such as EGF and TGF-α. In tandem with epithelial cell renewal, formation of new vessels (*angiogenesis*) within the injured microvascular bed occurs. Both FGF and vascular endothelial growth factor (VEGF) are important in regulating angiogenesis in the gastric

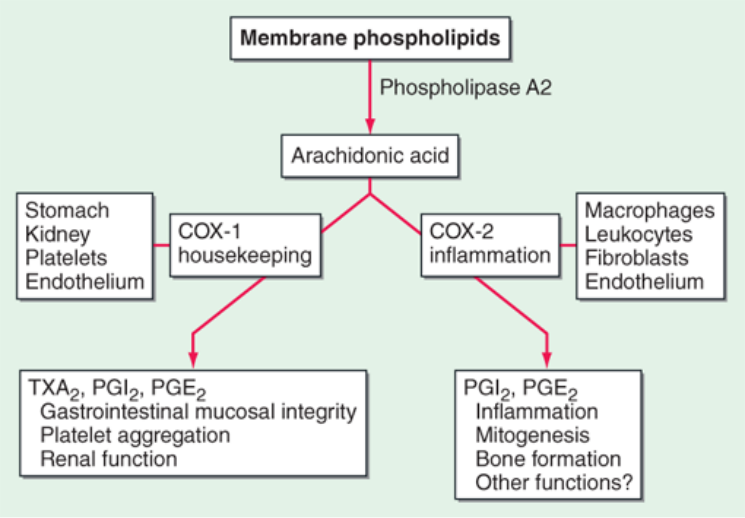
mucosa. In addition, the gastric peptide gastrin (see below) has been recently found to stimulate cell proliferation, migration, invasion and angiogenesis. It may also have the ability to regulate gastric stem cells.

An elaborate microvascular system within the gastric submucosal layer is the key component of the subepithelial defense/repair system, providing HCO_3^-, which neutralizes the acid generated by the parietal cell. Moreover, this microcirculatory bed provides an adequate supply of micronutrients and oxygen while removing toxic metabolic by-products. Several locally produced factors including nitric oxide (NO) (see below), hydrogen sulfide and prostacyclin contribute to the vascular protective pathway through vasodilation of the microcirculation.

Prostaglandins play a central role in gastric epithelial defense/repair **(Fig. 317-4)**. The gastric mucosa contains abundant levels of prostaglandins that regulate the release of mucosal bicarbonate and mucus, inhibit parietal cell secretion, and are important in maintaining mucosal blood flow and epithelial cell restitution. Prostaglandins are derived from esterified arachidonic acid, which is formed from phospholipids (cell membrane) by the action of phospholipase A_2. A key enzyme that controls the rate-limiting step in prostaglandin synthesis is cyclooxygenase (COX), which is present in two isoforms (COX-1, COX-2), each having distinct characteristics regarding structure, tissue distribution, and expression. COX-1 is expressed in a host of tissues, including the stomach, platelets, kidneys, and endothelial cells. This isoform is expressed in a constitutive manner and plays an important role in maintaining the integrity of renal function, platelet aggregation, and gastrointestinal (GI) mucosal integrity. In contrast, the expression of COX-2 is inducible by inflammatory stimuli, and it is expressed in macrophages, leukocytes, fibroblasts, and synovial cells. The beneficial effects of nonsteroidal anti-inflammatory drugs (NSAIDs) on tissue inflammation are due to inhibition of COX-2; the toxicity of these drugs (e.g., GI mucosal ulceration and renal dysfunction) is related to inhibition of the COX-1 isoform. The highly COX-2–selective NSAIDs have the potential to provide the beneficial effect of decreasing tissue inflammation while minimizing toxicity in the GI tract. Selective COX-2 inhibitors have had adverse effects on the cardiovascular (CV) system, leading to increased risk of myocardial infarction. Therefore, the U.S. Food and Drug Administration (FDA) has removed two of these agents (valdecoxib and rofecoxib) from the market (see below).

FIGURE 317-4

Schematic representation of the steps involved in synthesis of prostaglandin E_2 (PGE_2) and prostacyclin (PGI_2). Characteristics and distribution of the cyclooxygenase (COX) enzymes 1 and 2 are also shown. TXA_2, thromboxane A_2.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

NO is important in the maintenance of gastric mucosal integrity. The key enzyme NO synthase is constitutively expressed in the mucosa and contributes to cytoprotection by stimulating gastric mucus, increasing mucosal blood flow, and maintaining epithelial cell barrier function. The central nervous system (CNS) and hormonal factors also play a role in regulating mucosal defense through multiple pathways (Fig. 317-3).

Since the discovery of *Helicobacter pylori* and its impact on gastric pathology, it has become clear that the stomach has an elaborate and complex inherent immunological system in place. Although a detailed description of the gastric immune system is beyond the scope of this chapter several features are worth highlighting. The gastric immune response to certain pathogens such as *H. pylori* (see below) involves extensive interplay between innate (dendritic cell, epithelial cells, neutrophils and macrophages) and adaptive (B and T cells) components. Helper T cells (Th and Th Reg cells) have been extensively studied and appear to play an important role in a broad array of gastric physiology extending from gastric secretion to epithelial cell turnover via production of a number of cytokines.

The discovery of *H. pylori* has also led to the understanding that the stomach, once thought to be devoid of microorganisms due to its highly adverse environment (acid and pepsin) can serve as host for bacterial communities consisting of hundreds of phylotypes, otherwise known as its microbiota.

The conceptual framework of the microbiome has been receiving extensive attention in light of its importance in human health and disease. The overall relevance of the gastric microbiome and its impact on gastric pathology remains to be established but it is likely that alteration of microorganism homeostasis will play a role in aspects of certain disorders like PUD, gastritis and gastric cancer.

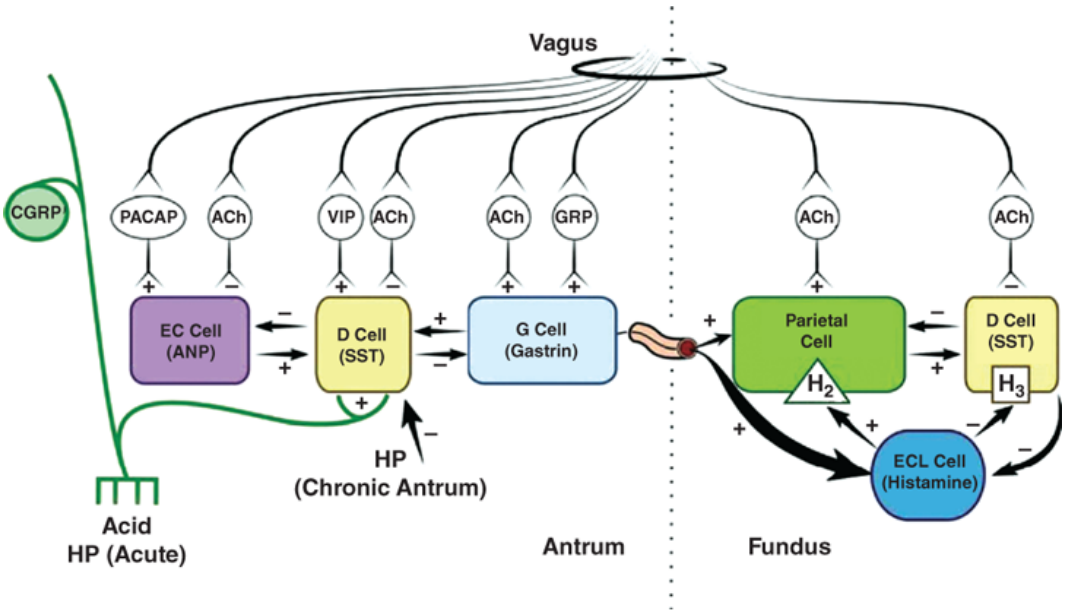
Physiology of Gastric Secretion

HCl and pepsinogen are the two principal gastric secretory products capable of inducing mucosal injury. Gastric acid and pepsinogen play a physiologic role in protein digestion; absorption of iron, calcium, magnesium, and vitamin B_{12}; and killing ingested bacteria. Acid secretion should be viewed as occurring under basal and stimulated conditions. Basal acid production occurs in a circadian pattern, with highest levels occurring during the night and lowest levels during the morning hours. Cholinergic input via the vagus nerve and histaminergic input from local gastric sources are the principal contributors to basal acid secretion. Stimulated gastric acid secretion occurs primarily in three phases based on the site where the signal originates (cephalic, gastric, and intestinal). Sight, smell, and taste of food are the components of the cephalic phase, which stimulates gastric secretion via the vagus nerve. The gastric phase is activated once food enters the stomach. This component of secretion is driven by nutrients (amino acids and amines) that directly (via peptone and amino acid receptors) and indirectly (via stimulation of intramural gastrin releasing peptide neurons) stimulate the G cell to release gastrin, which in turn activates the parietal cell via direct and indirect mechanisms. Distention of the stomach wall also leads to gastrin release and acid production. The last phase of gastric acid secretion is initiated as food enters the intestine and is mediated by luminal distention and nutrient assimilation. A series of pathways that inhibit gastric acid production are also set into motion during these phases. The GI hormone somatostatin is released from endocrine cells found in the gastric mucosa (D cells) in response to HCl. Somatostatin can inhibit acid production by both direct (parietal cell) and indirect mechanisms (decreased histamine release from ECL cells, ghrelin release from Gr cells and gastrin release from G cells). Additional neural (central and peripheral) and humoral (amylin, atrial natriuretic peptide [ANP], cholecystokinin, ghrelin, interleukin 11 [IL-11], obestatin, secretin, and serotonin) factors play a role in counterbalancing acid secretion. Under physiologic circumstances, these phases occur simultaneously. Ghrelin, the appetite-regulating hormone expressed in Gr cells in the stomach, and its related peptide motilin (released from the duodenum) may increase gastric acid secretion through stimulation of histamine release from ECL cells, but this remains to be confirmed.

The acid-secreting parietal cell is located in the oxyntic gland, adjacent to other cellular elements (ECL cell, D cell) important in the gastric secretory process **(Fig. 317-5)**. This unique cell also secretes intrinsic factor (IF) and IL-11. The parietal cell expresses receptors for several stimulants of acid secretion, including histamine (H_2), gastrin (cholecystokinin 2/gastrin receptor), and acetylcholine (muscarinic, M_3). Binding of histamine to the H_2 receptor leads to activation of adenylate cyclase and the phosphoinositol pathways in turn resulting in an increase in cyclic adenosine monophosphate (AMP) and intracellular calcium respectively. Activation of the gastrin and muscarinic receptors results in activation of the protein kinase C/phosphoinositide signaling pathway. Each of these signaling pathways in turn regulates a series of downstream kinase cascades that control the acid-secreting pump, H^+,K^+-ATPase. The discovery that different ligands and their corresponding receptors lead to activation of different signaling pathways explains the potentiation of acid secretion that occurs when histamine and gastrin or acetylcholine are combined. More importantly, this observation explains why blocking one receptor type (H_2) decreases acid secretion stimulated by agents that activate a different pathway (gastrin, acetylcholine). Parietal cells also express receptors for ligands that inhibit acid production (glucagon-like peptide-1, prostaglandins, somatostatin, and EGF). Histamine also stimulates gastric acid secretion indirectly by activating the histamine H_3 receptor on D-cells, which inhibits somatostatin release.

FIGURE 317-5

Regulation of gastric acid secretion at the cellular level. ACh, acetylcholine; ANP, atrial natriuretic peptide; CGRP, calcitonin gene-related peptide; EC, enterochromaffin; ECL, enterochromaffin-like; GRP, gastrin-releasing peptide; PACAP, pituitary adenylate-cyclase activating peptide; SST, somatostatin; VIP, vasoactive intestinal peptide.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The enzyme H^+,K^+-ATPase is responsible for generating the large concentration of H^+. It is a membrane-bound protein that consists of two subunits, α and β. The active catalytic site is found within the α subunit; the function of the β subunit is unclear. This enzyme uses the chemical energy of adenosine triphosphate (ATP) to transfer H^+ ions from parietal cell cytoplasm to the secretory canaliculi in exchange for K^+. The H^+,K^+-ATPase is located within the secretory canaliculus and in nonsecretory cytoplasmic tubulovesicles. The tubulovesicles are impermeable to K^+, which leads to an inactive pump in this location. The distribution of pumps between the nonsecretory vesicles and the secretory canaliculus varies according to parietal cell activity (Fig. 317-2). Proton pumps are recycled back to the inactive state in cytoplasmic vesicles once parietal cell activation ceases. Ezrin (an actin binding protein), actin, myosin, soluble N-ethylmaleimide-sensitive factor attachment protein receptors (SNAREs), small G proteins of the Rab family and secretory carrier membrane proteins (SCAMPS) are postulated to participate in parietal cell membrane translocation. In addition, acid secretion requires a number of apical and basolateral parietal cell membrane chloride and potassium channels.

The chief cell, found primarily in the gastric fundus, synthesizes and secretes pepsinogen, the inactive precursor of the proteolytic enzyme pepsin. The acid environment within the stomach leads to cleavage of the inactive precursor to pepsin and provides the low pH (<2) required for pepsin activity. Pepsin activity is significantly diminished at a pH of 4 and irreversibly inactivated and denatured at a pH of ≥7. Many of the secretagogues that stimulate acid secretion also stimulate pepsinogen release. The precise role of pepsin in the pathogenesis of PUD remains to be established.

PATHOPHYSIOLOGIC BASIS OF PUD

PUD encompasses both gastric and duodenal ulcers (DUs). *Ulcers* are defined as breaks in the mucosal surface >5 mm in size, with depth to the submucosa. DUs and gastric ulcers (GUs) share many common features in terms of pathogenesis, diagnosis, and treatment, but several factors distinguish them from one another. *H. pylori* and NSAIDs are the most common risk factors for PUD, with estimated odds ratios in the United States of 3.7 and 3.3, respectively. Additional risk factors (odds ratio) include chronic obstructive lung disease (2.34), chronic renal insufficiency (2.29), current tobacco use (1.99), former tobacco use (1.55), older age (1.67), three or more doctor visits in a year (1.49), coronary heart disease (1.46), former alcohol use (1.29), African-American race (1.20), obesity (1.18), and diabetes (1.13). The mechanisms by which some of these risk factors lead to ulcer disease are highlighted below.

Epidemiology

DUODENAL ULCERS

DUs are estimated to occur in 6–15% of the Western population. The incidence of DUs declined steadily from 1960 to 1980 and has remained stable since then. The death rates, need for surgery, and physician visits have decreased by >50% over the past 30 years. The reason for the reduction in the frequency of DUs is likely related to the decreasing frequency of *H. pylori.* Before the discovery of *H. pylori*, the natural history of DUs was typified by frequent recurrences after initial therapy. Eradication of *H. pylori* has reduced these recurrence rates by >80 %.

GASTRIC ULCERS

GUs tend to occur later in life than duodenal lesions, with a peak incidence reported in the sixth decade. More than one-half of GUs occur in males and are less common than DUs, perhaps due to the higher likelihood of GUs being silent and presenting only after a complication develops. Autopsy studies suggest a similar incidence of DUs and GUs.

Pathology

DUODENAL ULCERS

DUs occur most often in the first portion of the duodenum (>95%), with ~90% located within 3 cm of the pylorus. They are usually ≤1 cm in diameter but can occasionally reach 3–6 cm (giant ulcer). Ulcers are sharply demarcated, with depth at times reaching the muscularis propria. The base of the ulcer often consists of a zone of eosinophilic necrosis with surrounding fibrosis. Malignant DUs are extremely rare.

GASTRIC ULCERS

In contrast to DUs, GUs can represent a malignancy and should be biopsied upon discovery. Benign GUs are most often found distal to the junction between the antrum and the acid secretory mucosa. Benign GUs are quite rare in the gastric fundus and are histologically similar to DUs. Benign GUs associated with *H. pylori* are also associated with antral gastritis. In contrast, NSAID-related GUs are not accompanied by chronic active gastritis but may instead have evidence of a chemical gastropathy, typified by foveolar hyperplasia, edema of the lamina propria, and epithelial regeneration in the absence of *H. pylori.* Extension of smooth-muscle fibers into the upper portions of the mucosa, where they are not typically found, may also occur.

Pathophysiology

DUODENAL ULCERS

H. pylori and NSAID-induced injuries account for the majority of DUs. Many acid secretory abnormalities have been described in DU patients. Of these, average basal and nocturnal gastric acid secretion appears to be increased in DU patients as compared to controls; however, the level of overlap between DU patients and control subjects is substantial. The reason for this altered secretory process is unclear, but *H. pylori* infection may contribute. Bicarbonate secretion is significantly decreased in the duodenal bulb of patients with an active DU as compared to control subjects. *H. pylori* infection may also play a role in this process (see below).

GASTRIC ULCERS

As in DUs, the majority of GUs can be attributed to either *H. pylori* or NSAID-induced mucosal damage. Prepyloric GUs or those in the body associated with a DU or a duodenal scar are similar in pathogenesis to DUs. Gastric acid output (basal and stimulated) tends to be normal or decreased in GU patients. When GUs develop in the presence of minimal acid levels, impairment of mucosal defense factors may be present. GUs have been classified based on their location: Type I occur in the gastric body and tend to be associated with low gastric acid production; type II occur in the antrum and gastric acid can vary from low to normal; type III occur within 3 cm of the pylorus and are commonly accompanied by DUs and normal or high gastric acid production; and type IV are found in the cardia and are associated with low gastric acid production.

H. PYLORI AND ACID PEPTIC DISORDERS

Gastric infection with the bacterium *H. pylori* accounts for the majority of PUD **(Chap. 158)**. This organism also plays a role in the development of gastric mucosa-associated lymphoid tissue (MALT) lymphoma and gastric adenocarcinoma. Although the entire genome of *H. pylori* has been sequenced, it is still not clear how this organism, which resides in the stomach, causes ulceration in the duodenum. *H. pylori* eradication efforts may lead to a decrease in gastric cancer in high-risk populations particularly in individuals who have not developed chronic atrophic gastritis and gastric metaplasia.

The Bacterium

The bacterium, initially named *Campylobacter pyloridis*, is a gram-negative microaerophilic rod found most commonly in the deeper portions of the mucous gel coating the gastric mucosa or between the mucous layer and the gastric epithelium. It may attach to gastric epithelium but under normal circumstances does not appear to invade cells. It is strategically designed to live within the aggressive environment of the stomach. It is S-shaped (~0.5–3 μm in size) and contains multiple sheathed flagella. Initially, *H. pylori* resides in the antrum but, over time, migrates toward the more proximal segments of the stomach. The organism is capable of transforming into a coccoid form, which represents a dormant state that may facilitate survival in adverse conditions. The genome of *H. pylori* (1.65 million base pairs) encodes ~1500 proteins. Among this multitude of proteins there are factors that are essential determinants of *H. pylori*–mediated pathogenesis and colonization such as the outer membrane protein (Hop proteins), urease, and the vacuolating cytotoxin (Vac A). Moreover, the majority of *H. pylori* strains contain a genomic fragment that encodes the cag pathogenicity island (cag-PAI). Several of the genes that make up cag-PAI encode components of a type IV secretion island that translocates Cag A into host cells. Once in the cell, Cag A activates a series of cellular events important in cell growth and cytokine production. *H. pylori* also has extensive genetic diversity that in turn enhances its ability to promote disease. The first step in infection by *H. pylori* is dependent on the bacteria's motility and its ability to produce urease. Urease produces ammonia from urea, an essential step in alkalinizing the surrounding pH. Additional bacterial factors include catalase, lipase, adhesins, platelet-activating factor, and pic B (induces cytokines). Multiple strains of *H. pylori* exist and are characterized by their ability to express several of these factors (Cag A, Vac A, etc.). It is possible that the different diseases related to *H. pylori* infection can be attributed to different strains of the organism with distinct pathogenic features.

Epidemiology

The prevalence of *H. pylori* varies throughout the world and depends largely on the overall standard of living in the region. In developing parts of the world, 80% of the population may be infected by the age of 20, whereas the prevalence is 20–50% in industrialized countries. In contrast, in the United States this organism is rare in childhood. The overall prevalence of *H. pylori* in the United States is ~30%, with individuals born before 1950 having a higher rate of infection than those born later. About 10% of Americans <30 years of age are colonized with the bacteria. The rate of infection with *H.*

pylori in industrialized countries has decreased substantially in recent decades. The steady increase in the prevalence of *H. pylori* noted with increasing age is due primarily to a cohort effect, reflecting higher transmission during a period in which the earlier cohorts were children. It has been calculated through mathematical models that improved sanitation during the latter half of the nineteenth century dramatically decreased transmission of *H. pylori.* Moreover, with the present rate of intervention, the organism will be ultimately eliminated from the United States. Two factors that predispose to higher colonization rates include poor socioeconomic status and less education. These factors, not race, are responsible for the rate of *H. pylori* infection in blacks and Hispanic Americans being double the rate seen in whites of comparable age. Other risk factors for *H. pylori* infection are (1) birth or residence in a developing country, (2) domestic crowding, (3) unsanitary living conditions, (4) unclean food or water, and (5) exposure to gastric contents of an infected individual.

Transmission of *H. pylori* occurs from person to person, following an oral-oral or fecal-oral route. The risk of *H. pylori* infection is declining in developing countries. The rate of infection in the United States has fallen by >50% when compared to 30 years ago.

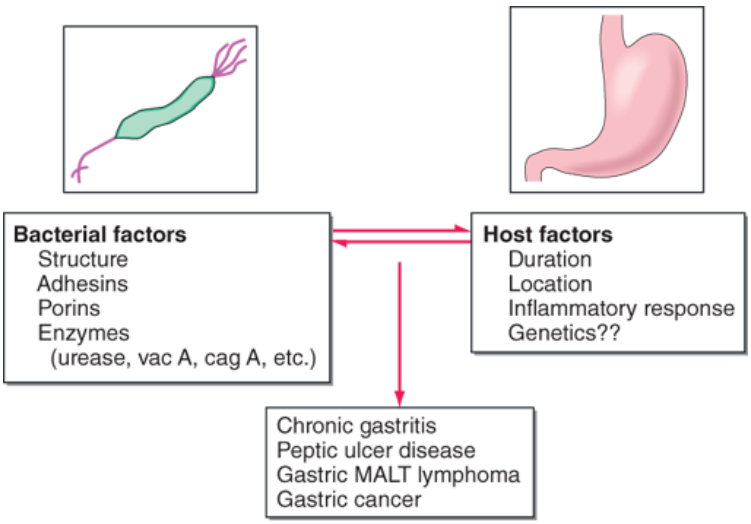
Pathophysiology

H. pylori infection is virtually always associated with a chronic active gastritis, but only 10–15% of infected individuals develop frank peptic ulceration. The basis for this difference is unknown, but is likely due to a combination of host and bacterial factors some of which are outlined below. Initial studies suggested that >90% of all DUs were associated with *H. pylori*, but *H. pylori* is present in only 30–60% of individuals with GUs and 50–70% of patients with DUs. The pathophysiology of ulcers not associated with *H. pylori* or NSAID ingestion (or the rare Zollinger-Ellison syndrome [ZES]) is becoming more relevant as the incidence of *H. pylori* is dropping, particularly in the Western world (see below).

The particular end result of *H. pylori* infection (gastritis, PUD, gastric MALT lymphoma, gastric cancer) is determined by a complex interplay between bacterial and host factors **(Fig. 317-6)**.

FIGURE 317-6

Outline of the bacterial and host factors important in determining *H. pylori*–induced gastrointestinal disease. MALT, mucosal-associated lymphoid tissue.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Bacterial factors: H. pylori is able to facilitate gastric residence, induce mucosal injury, and avoid host defense. Different strains of *H. pylori* produce different virulence factors including γ-glutamyl transpeptidase (GGT), cytotoxin-associated gene A (cagA) product, and virulence components vacuolating toxin (vacA), in addition to pathogen-associated molecular patterns (PAMPs) such as flagella and lipopolysaccharide (LPS). A specific region of the bacterial genome, the pathogenicity island (cag-PAI), encodes the virulence factors Cag A and pic B. Vac A also contributes to pathogenicity, although it is not encoded within the pathogenicity island. These virulence factors, in conjunction with additional bacterial constituents, can cause mucosal damage, in part through their ability to target the host immune cells. For example, Vac A targets human CD4 T cells, inhibiting their proliferation and in addition can disrupt normal function of B cells, CD8 T cells, macrophages, and mast cells. Multiple studies have demonstrated that *H. pylori* strains that are cag-PAI positive are associated with a higher risk of PUD, premalignant gastric lesions, and gastric cancer than are strains that lack the cag-PAI. In addition, *H. pylori* may directly inhibit parietal cell H^+,K^+-ATPase activity through a Cag A–dependent mechanism, leading in part to the low acid production observed after acute infection with the organism. Urease, which allows the bacteria to reside in the acidic stomach, generates NH_3, which can damage epithelial cells. The bacteria produce surface factors that are chemotactic for neutrophils and monocytes, which in turn contribute to epithelial cell injury (see below). *H. pylori* makes proteases and phospholipases that break down the glycoprotein lipid complex of the mucous gel, thus reducing the efficacy of this first line of mucosal defense. *H. pylori* expresses adhesins (OMPs like

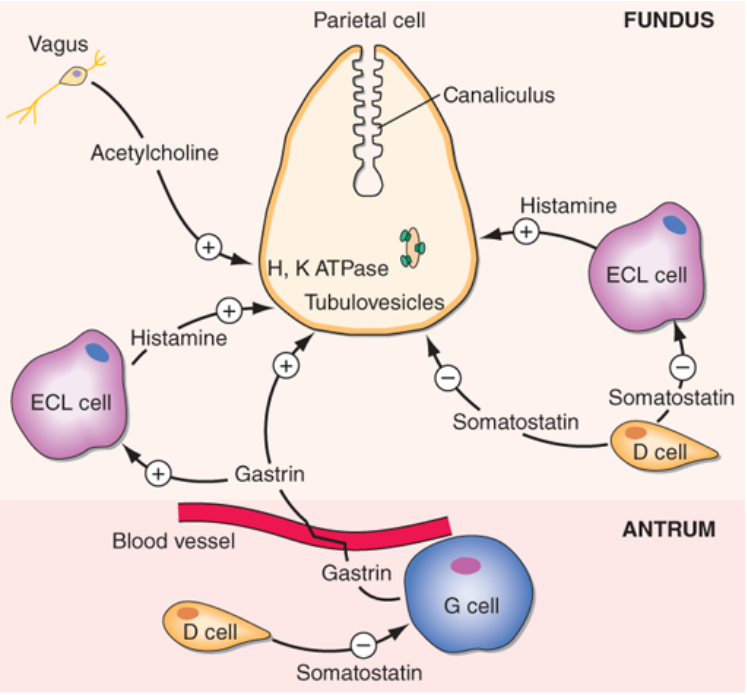
BabA), which facilitate attachment of the bacteria to gastric epithelial cells. Although LPS of gram-negative bacteria often plays an important role in the infection, *H. pylori* LPS has low immunologic activity compared to that of other organisms. It may promote a smoldering chronic inflammation.

Host factors: Studies in twins suggest that there may be genetic predisposition to acquire *H. pylori.* The inflammatory response to *H. pylori* includes recruitment of neutrophils, lymphocytes (T and B), macrophages, and plasma cells. The pathogen leads to local injury by binding to class II major histocompatibility complex (MHC) molecules expressed on gastric epithelial cells, leading to cell death (*apoptosis*). Moreover, bacterial strains that encode cag-PAI can introduce Cag A into the host cells, leading to further cell injury and activation of cellular pathways involved in cytokine production and repression of tumor-suppressor genes. Elevated concentrations of multiple cytokines are found in the gastric epithelium of *H. pylori*–infected individuals, including interleukin (IL) 1α/β, IL-2, IL-6, IL-8, tumor necrosis factor (TNF) α, and interferon (IFN) γ. *H. pylori* infection also leads to both a mucosal and a systemic humoral response, which does not lead to eradication of the bacteria but further compounds epithelial cell injury. Additional mechanisms by which *H. pylori* may cause epithelial cell injury include (1) activated neutrophil-mediated production of reactive oxygen or nitrogen species and enhanced epithelial cell turnover and (2) apoptosis related to interaction with T cells (T helper 1, or T_H1, cells) and IFN-γ. Finally, the human stomach is colonized by a host of commensal organisms that may affect the likelihood of *H. pylori*–infection and subsequent mucosal injury. Moreover, colonization of the stomach with *H. pylori* likely alters the composition of the gastric microbiota. The impact of the latter on gastric pathophysiology remains unknown. *H. pylori* also appears to regulate NO formation via different mechanisms that in turn may contribute to the organism's cytotoxic effects. Specifically, *H. pylori* derived factors, such as urease, or the bacterium itself, stimulate NO synthase (NOS2) expression in macrophages and in gastric epithelial cells leading to NO release and subsequent cytotoxic effect on surrounding cells. *H. pylori* also leads to the formation of 8-nitroguanine (8-NO2-Gua), which in conjunction with oncoprotein CagA, may contribute to the development of gastric cancer.

The reason for *H. pylori*–mediated duodenal ulceration remains unclear. Studies suggest that *H. pylori* associated with duodenal ulceration may be more virulent. In addition, certain specific bacterial factors such as the DU-promoting gene A (*dupA*), may be associated with the development of DUs. Another potential contributing factor is that gastric metaplasia in the duodenum of DU patients, which may be due to high acid exposure (see below), permits *H. pylori* to bind to it and produce local injury secondary to the host response. Another hypothesis is that *H. pylori* antral infection could lead to increased acid production, increased duodenal acid, and mucosal injury. Basal and stimulated (meal, gastrin-releasing peptide [GRP]) gastrin release are increased in *H. pylori*–infected individuals, and somatostatin-secreting D cells may be decreased. *H. pylori* infection might induce increased acid secretion through both direct and indirect actions of *H. pylori* and proinflammatory cytokines (IL-8, TNF, and IL-1) on G, D, and parietal cells **(Fig. 317-7)**. GUs, in contrast, are associated with *H. pylori*–induced pangastritis and normal or low gastric acid secretion. *H. pylori* infection has also been associated with decreased duodenal mucosal bicarbonate production. Data supporting and contradicting each of these interesting theories have been demonstrated. Thus, the mechanism by which *H. pylori* infection of the stomach leads to duodenal ulceration remains to be established.

FIGURE 317-7

Summary of potential mechanisms by which *H. pylori* may lead to gastric secretory abnormalities. D, somatostatin cell; ECL, enterochromaffin-like cell; G, G cell. *(Adapted from J Calam et al: Gastroenterology 113:543, 1997.)*

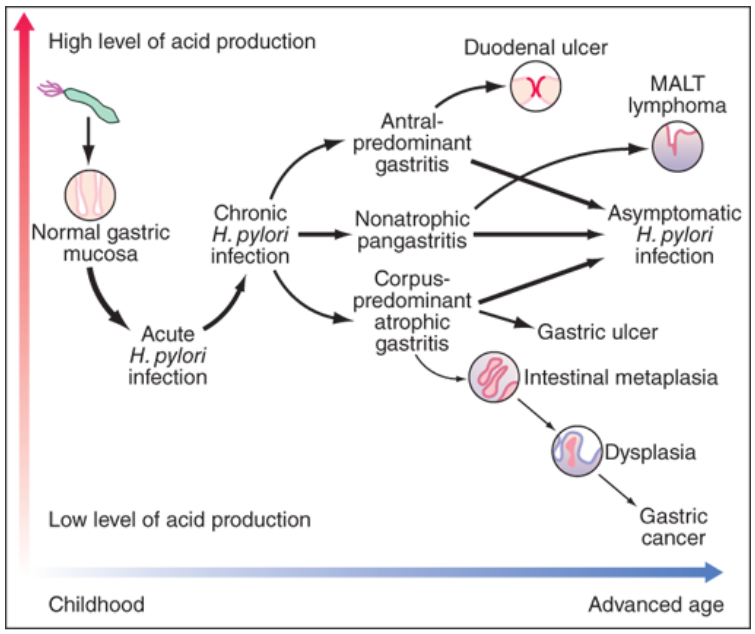


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In summary, the final effect of *H. pylori* on the GI tract is variable and determined by microbial and host factors. The type and distribution of gastritis correlate with the ultimate gastric and duodenal pathology observed. Specifically, the presence of antral-predominant gastritis is associated with DU formation; gastritis involving primarily the corpus predisposes to the development of GUs, gastric atrophy, and ultimately gastric carcinoma **(Fig. 317-8)**.

FIGURE 317-8

Natural history of *H. pylori* infection. MALT, mucosal-associated lymphoid tissue. *(Used with permission from S Suerbaum, P Michetti: N Engl J Med 347:1175, 2002.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

NSAID-INDUCED DISEASE

Epidemiology

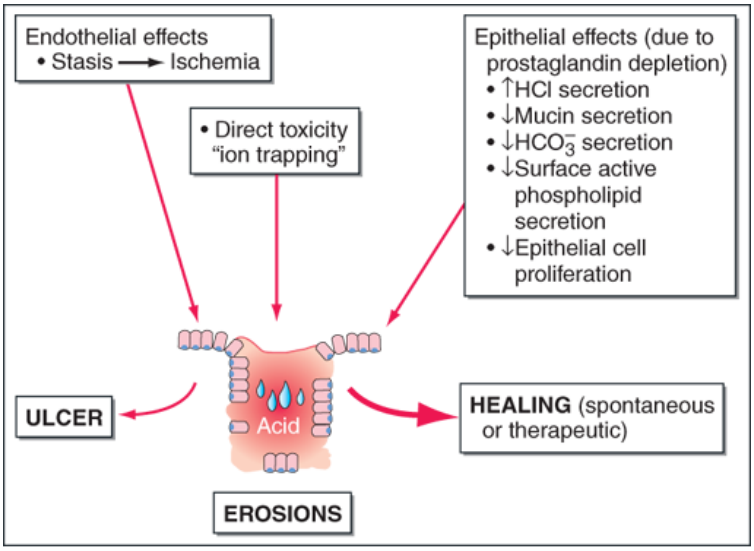
NSAIDs represent a group of the most commonly used medications in the world and the United States. It is estimated that 7 billion dollars per year are spent on NSAIDs world wide with more than 30 billion over-the-counter tablets and over 100 million prescriptions sold yearly in the United States alone. In fact, after the introduction of COX-2 inhibitors in the year 2000, the number of prescriptions written for NSAIDs was >111 million at a cost of $4.8 billion. Side effects and complications due to NSAIDs are considered the most common drug-related toxicities in the United States. The spectrum of NSAID-induced morbidity ranges from nausea and dyspepsia (prevalence reported as high as 50–60%) to a serious GI complication such as endoscopy-documented peptic ulceration (15–30% of individuals taking NSAIDs regularly) complicated by bleeding or perforation in as many as 1.5% of users per year. It is estimated that NSAID-induced GI bleeding accounts for 60,000–120,000 hospital admissions per year, and deaths related to NSAID-induced toxicity may be as high as 16,000 per year in the United States. Approximately 4–5% of patients develop symptomatic ulcers within 1 year. Unfortunately, dyspeptic symptoms do not correlate with NSAID-induced pathology. Over 80% of patients with serious NSAID-related complications did not have preceding dyspepsia. In view of the lack of warning signs, it is important to identify patients who are at increased risk for morbidity and mortality related to NSAID usage. Even 75 mg/d of aspirin may lead to serious GI ulceration; thus, no dose of NSAID is completely safe. In fact, the incidence of mucosal injury (ulcers and erosions) in patients taking low-dose aspirin (75–325 mg) has been estimated to range from as low as 8 to as high as 60%. It appears that *H. pylori* infection increases the risk of PUD-associated GI bleeding in chronic users of low-dose aspirin. Established risk factors include advanced age, history of ulcer, concomitant use of glucocorticoids, high-dose NSAIDs, multiple NSAIDs, concomitant use of anticoagulants, clopidogrel, and serious or multisystem disease. Possible risk factors include concomitant infection with *H. pylori*, cigarette smoking, and alcohol consumption.

Pathophysiology

Prostaglandins play a critical role in maintaining gastroduodenal mucosal integrity and repair. It therefore follows that interruption of prostaglandin synthesis can impair mucosal defense and repair, thus facilitating mucosal injury via a systemic mechanism. Animal studies have demonstrated that neutrophil adherence to the gastric microcirculation plays an essential role in the initiation of NSAID-induced mucosal injury. A summary of the pathogenetic pathways by which systemically administered NSAIDs may lead to mucosal injury is shown in **Fig. 317-9**. Single nucleotide polymorphisms (SNPs) have been found in several genes, including those encoding certain subtypes of cytochrome P450 (see below), interleukin-1β (*IL-1β*), angiotensinogen (*AGT*), and an organic ion transporting polypeptide (*SLCO1B1*), but these findings need confirmation in larger scale studies.

FIGURE 317-9

Mechanisms by which nonsteroidal anti-inflammatory drugs may induce mucosal injury. *(Adapted from J Scheiman et al: J Clin Outcomes Management 3:23, 1996. Copyright 2003 Turner White Communications, Inc., www.turner-white.com. Used with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Injury to the mucosa also occurs as a result of the topical encounter with NSAIDs leading to increased epithelial surface permeability. Aspirin and many NSAIDs are weak acids that remain in a nonionized lipophilic form when found within the acid environment of the stomach. Under these conditions, NSAIDs migrate across lipid membranes of epithelial cells, leading to cell injury once trapped intracellularly in an ionized form. Topical NSAIDs can also alter the surface mucous layer, permitting back diffusion of H^+ and pepsin, leading to further epithelial cell damage. Moreover, enteric-coated or buffered preparations are also associated with risk of peptic ulceration. NSAIDs can also lead to mucosal injury via production of additional pro-inflammatory mediators like TNF and leukotrienes through simultaneous activation of the lipoxygenase pathway.

The interplay between *H. pylori* and NSAIDs in the pathogenesis of PUD is complex. Meta-analysis supports the conclusion that each of these aggressive factors is independent and synergistic risk factors for PUD and its complications such as GI bleeding. For example, eradication of *H. pylori* reduces the likelihood of GI complications in high-risk individuals to levels observed in individuals with average risk of NSAID-induced complications.

PATHOGENETIC FACTORS UNRELATED TO *H. PYLORI* AND NSAIDS IN ACID PEPTIC DISEASE

Cigarette smoking has been implicated in the pathogenesis of PUD. Not only have smokers been found to have ulcers more frequently than do nonsmokers, but smoking appears to decrease healing rates, impair response to therapy, and increase ulcer-related complications such as perforation. The mechanism responsible for increased ulcer diathesis in smokers is unknown. Theories have included altered gastric emptying, decreased proximal duodenal bicarbonate production, increased risk for *H. pylori* infection, and cigarette-induced generation of noxious mucosal free radicals. Genetic predisposition may play a role in ulcer development. First-degree relatives of DU patients are three times as likely to develop an ulcer; however, the potential role of *H. pylori* infection in contacts is a major consideration. Increased frequencies of blood group O and of the nonsecretor status have also been implicated as genetic risk factors for peptic diathesis. However, *H. pylori* preferentially binds to group O antigens. Additional genetic factors have been postulated to predispose certain individuals to developing PUD and/or upper GI bleeding. Specifically, genes encoding the NSAID-metabolizing enzymes cytochrome P450 2C9 and 2C8 (CYP2C9 and CYP2C8) are potential susceptibility genes for NSAID-induced PUD, but unfortunately, the studies have not been consistent in demonstrating this association. In a United Kingdom study, the *CYP2C19*17* gain-of-function polymorphism was associated with PUD in a Caucasian cohort, irrespective of ulcer etiology. These findings need to be confirmed in broader studies. Psychological stress has been thought to contribute to PUD, but studies examining the role of psychological factors in its pathogenesis have generated conflicting results. Although PUD is associated with certain personality traits (neuroticism), these same traits are also present in individuals with nonulcer dyspepsia (NUD) and other functional and organic disorders.

Diet has also been thought to play a role in peptic diseases. Certain foods and beverages can cause dyspepsia, but no convincing studies indicate an association between ulcer formation and a specific diet. Specific chronic disorders have been shown to have a strong association with PUD: (1) advanced age, (2) chronic pulmonary disease, (3) chronic renal failure, (4) cirrhosis, (5) nephrolithiasis, (6) α_1-antitrypsin deficiency, and (7) systemic mastocytosis. Disorders with a possible association are (1) hyperparathyroidism, (2) coronary artery disease, (3) polycythemia vera, (4) chronic pancreatitis, (5) former alcohol use, (6) obesity, (7) African-American race, and (8) three or more doctor visits in a year.

Multiple factors play a role in the pathogenesis of PUD. The two predominant causes are *H. pylori* infection and NSAID ingestion. PUD not related to *H. pylori* or NSAIDs is increasing. Other less common causes of PUD are shown in **Table 317-1**. These etiologic agents should be considered as the

incidence of *H. pylori* is decreasing. Independent of the inciting or injurious agent, peptic ulcers develop as a result of an imbalance between mucosal protection/repair and aggressive factors. Gastric acid plays an important role in mucosal injury.

TABLE 317-1

Causes of Ulcers Not Caused by *Helicobacter pylori* and NSAIDs

Pathogenesis of Non-Hp and Non-NSAID Ulcer Disease
Infection
Cytomegalovirus
Herpes simplex virus
Helicobacter heilmannii
Drug/Toxin
Bisphosphonates
Chemotherapy
Clopidogrel
Crack cocaine
Glucocorticoids (when combined with NSAIDs)
Mycophenolate mofetil
Potassium chloride
Miscellaneous
Basophilia in myeloproliferative disease
Duodenal obstruction (e.g., annular pancreas)
Infiltrating disease
Ischemia
Radiation therapy
Eosinophilic infiltration
Sarcoidosis
Crohn's disease
Idiopathic hypersecretory state

Abbreviations: Hp, *H. pylori*; NSAIDs, nonsteroidal anti-inflammatory drugs.

CLINICAL FEATURES

History

Abdominal pain is common to many GI disorders, including DU and GU, but has a poor predictive value for the presence of either DU or GU. Up to 87% of patients with NSAID-induced mucosal disease can present with a complication (bleeding, perforation, and obstruction) without antecedent symptoms. Despite this poor correlation, a careful history and physical examination are essential components of the approach to a patient suspected of having peptic ulcers.

Epigastric pain described as a burning or gnawing discomfort can be present in both DU and GU. The discomfort is also described as an ill-defined, aching sensation or as hunger pain. The typical pain pattern in DU occurs 90 min to 3 h after a meal and is frequently relieved by antacids or food. Pain that awakes the patient from sleep (between midnight and 3 A.M.) is the most discriminating symptom, with two-thirds of DU patients describing this complaint. Unfortunately, this symptom is also present in one-third of patients with NUD (see below). Elderly patients are less likely to have abdominal pain as a manifestation of PUD and may instead present with a complication such as ulcer bleeding or perforation. The pain pattern in GU patients may be different from that in DU patients, where discomfort may actually be precipitated by food. Nausea and weight loss occur more commonly in GU patients. Endoscopy detects ulcers in <30% of patients who have dyspepsia.

The mechanism for development of abdominal pain in ulcer patients is unknown. Several possible explanations include acid-induced activation of chemical receptors in the duodenum, enhanced duodenal sensitivity to bile acids and pepsin, or altered gastroduodenal motility.

Variation in the intensity or distribution of the abdominal pain, as well as the onset of associated symptoms such as nausea and/or vomiting, may be indicative of an ulcer complication. Dyspepsia that becomes constant, is no longer relieved by food or antacids, or radiates to the back may indicate a penetrating ulcer (pancreas). Sudden onset of severe, generalized abdominal pain may indicate perforation. Pain worsening with meals, nausea, and vomiting of undigested food suggest gastric outlet obstruction. Tarry stools or coffee-ground emesis indicate bleeding.

Physical Examination

Epigastric tenderness is the most frequent finding in patients with GU or DU. Pain may be found to the right of the midline in 20% of patients. Unfortunately, the predictive value of this finding is low. Physical examination is critically important for discovering evidence of ulcer complication. Tachycardia and orthostasis suggest dehydration secondary to vomiting or active GI blood loss. A severely tender, board-like abdomen suggests a perforation. Presence of a succussion splash indicates retained fluid in the stomach, suggesting gastric outlet obstruction.

PUD-Related Complications

GASTROINTESTINAL BLEEDING

GI bleeding is the most common complication observed in PUD. Bleeding is estimated to occur in 19.4–57 per 100,000 individuals in a general population or in ~15% of patients. Bleeding and complications of ulcer disease occur more often in individuals >60 years of age. The 30-day mortality rate is as high as 2.5–10%. The higher incidence in the elderly is likely due to the increased use of NSAIDs in this group. In addition, up to 80% of the mortality in PUD-related bleeding is due to nonbleeding causes such as multiorgan failure (24%), pulmonary complications (24%), and malignancy (34%).

Greater than 50% of patients with ulcer-related hemorrhage bleed without any preceding warning signs or symptoms.

PERFORATION

The second most common ulcer-related complication is perforation, being reported in as many as 6–7% of PUD patients with an estimated 30-day mortality of >20%. As in the case of bleeding, the incidence of perforation in the elderly appears to be increasing secondary to increased use of NSAIDs. Perforation of DUs has become less common in light of the increased rates of *H. pylori* eradication with NSAID induced GUs leading to perforation occurring more commonly. *Penetration* is a form of perforation in which the ulcer bed tunnels into an adjacent organ. DUs tend to penetrate posteriorly into the pancreas, leading to pancreatitis, whereas GUs tend to penetrate into the left hepatic lobe. Gastrocolic fistulas associated with GUs have also been described.

GASTRIC OUTLET OBSTRUCTION

Gastric outlet obstruction is the least common ulcer-related complication, occurring in 1–2% of patients. A patient may have relative obstruction secondary to ulcer-related inflammation and edema in the peripyloric region. This process often resolves with ulcer healing. A fixed, mechanical obstruction secondary to scar formation in the peripyloric areas is also possible. The latter requires endoscopic (balloon dilation) or surgical intervention. Signs and symptoms relative to mechanical obstruction may develop insidiously. New onset of early satiety, nausea, vomiting, increase of postprandial abdominal pain, and weight loss should make gastric outlet obstruction a possible diagnosis.

Differential Diagnosis

The list of GI and non-GI disorders that can mimic ulceration of the stomach or duodenum is quite extensive. The most commonly encountered diagnosis among patients seen for upper abdominal discomfort is functional dyspepsia (FD) or *essential dyspepsia* which refers to a group of heterogeneous disorders typified by upper abdominal pain without the presence of an ulcer. The symptoms can range from postprandial fullness and early satiety to epigastric burning pain. The dichotomy of this symptom complex has led to the identification of two subcategories of FD including

postprandial distress syndrome (PDS) and epigastric pain syndrome (EPS). Dyspepsia has been reported to occur in up to 30% of the U.S. population. Up to 80% of patients seeking medical care for dyspepsia have a negative diagnostic evaluation. The etiology of FD is not established, but recent studies suggest that post-infectious states, certain foods and *H. pylori* infection may contribute to the pathogenesis of this common disorder.

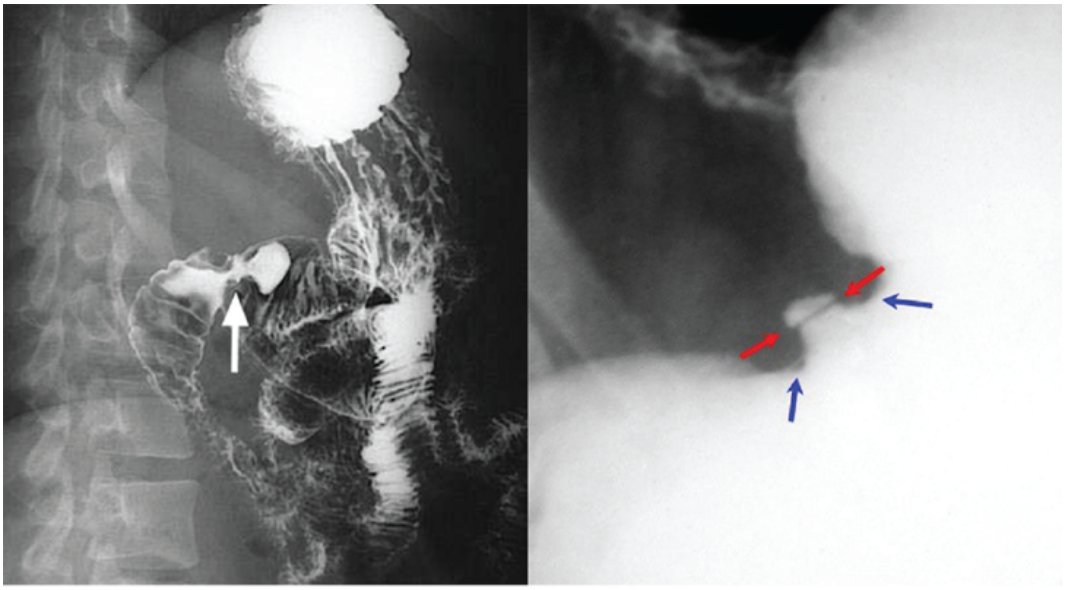
Several additional disease processes that may present with "ulcer-like" symptoms include proximal GI tumors, gastroesophageal reflux, vascular disease, pancreaticobiliary disease (biliary colic, chronic pancreatitis), and gastroduodenal Crohn's disease.

Diagnostic Evaluation

In view of the poor predictive value of abdominal pain for the presence of a gastroduodenal ulcer and the multiple disease processes that can mimic this disease, the clinician is often confronted with having to establish the presence of an ulcer. Documentation of an ulcer requires either a radiographic (barium study) or an endoscopic procedure. However, a large percentage of patients with symptoms suggestive of an ulcer have NUD; testing for *H. pylori* and antibiotic therapy (see below) is appropriate for individuals who are otherwise healthy and <45 years of age, before embarking on a diagnostic evaluation **(Chap. 41)**.

Barium studies of the proximal GI tract are rarely used as a first test for documenting an ulcer. The sensitivity of older single-contrast barium meals for detecting a DU is as high as 80%, with a double-contrast study providing detection rates as high as 90%. Sensitivity for detection is decreased in small ulcers (<0.5 cm), with presence of previous scarring, or in postoperative patients. A DU appears as a well-demarcated crater, most often seen in the bulb **(Fig. 317-10A)**. A GU may represent benign or malignant disease. Typically, a benign GU also appears as a discrete crater with radiating mucosal folds originating from the ulcer margin **(Fig. 317-10B)**. Ulcers >3 cm in size or those associated with a mass are more often malignant. Unfortunately, up to 8% of GUs that appear to be benign by radiographic appearance are malignant by endoscopy or surgery. Radiographic studies that show a GU must be followed by endoscopy and biopsy.

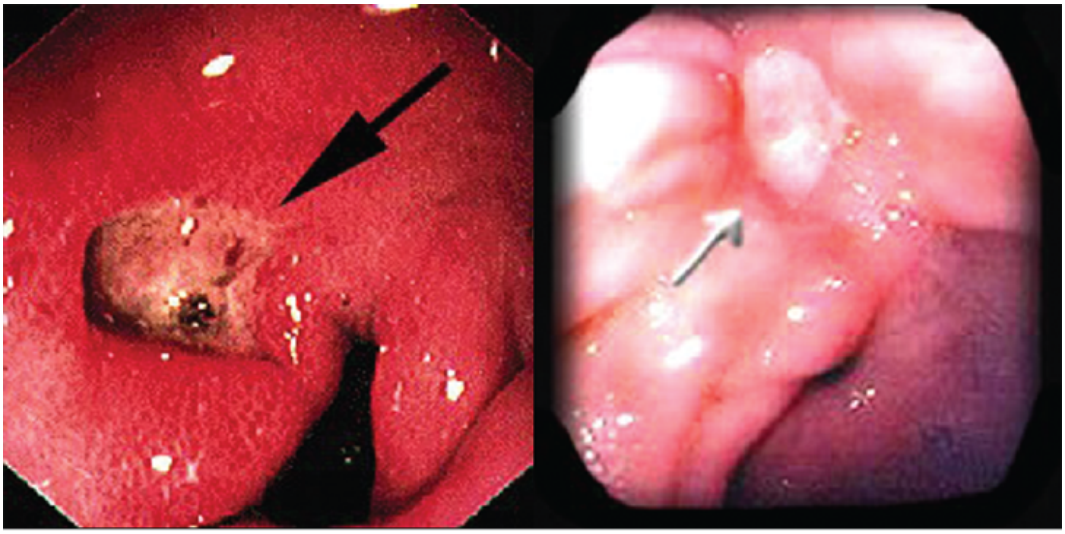
FIGURE 317-10
Barium study demonstrating (***A***) a benign duodenal ulcer and (***B***) a benign gastric ulcer.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Endoscopy provides the most sensitive and specific approach for examining the upper GI tract **(Fig. 317-11)**. In addition to permitting direct visualization of the mucosa, endoscopy facilitates photographic documentation of a mucosal defect and tissue biopsy to rule out malignancy (GU) or *H. pylori.* Endoscopic examination is particularly helpful in identifying lesions too small to detect by radiographic examination, for evaluation of atypical radiographic abnormalities, or to determine if an ulcer is a source of blood loss.

FIGURE 317-11
Endoscopy demonstrating (***A***) a benign duodenal ulcer and (***B***) a benign gastric ulcer.



A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although the methods for diagnosing *H. pylori* are outlined in **Chap. 158**, a brief summary will be included here **(Table 317-2)**. Several biopsy urease tests have been developed (PyloriTek, CLOtest, Hpfast, Pronto Dry) that have a sensitivity and specificity of >90–95%. Several noninvasive methods for detecting this organism have been developed. Three types of studies routinely used include serologic testing, the ^{13}C- or ^{14}C-urea breath test, and the fecal *H. pylori* (Hp) antigen test (monoclonal antibody test). A urinary Hp antigen test appears promising.

TABLE 317-2

Tests for Detection of *Helicobacter pylori*

Test	Sensitivity/Specificity, %	Comments
Invasive (Endoscopy/Biopsy Required)		
Rapid urease	80–95/95–100	Simple, false negative with recent use of PPIs, antibiotics, or bismuth compounds
Histology	80–90/>95	Requires pathology processing and staining; provides histologic information
Culture	—/—	Time-consuming, expensive, dependent on experience; allows determination of antibiotic susceptibility
Noninvasive		
Serology	>80/>90	Inexpensive, convenient; not useful for early follow-up
Urea breath test	>90/>90	Simple, rapid; useful for early follow-up; false negatives with recent therapy (see rapid urease test); exposure to low-dose radiation with ^{14}C test
Stool antigen	>90/>90	Inexpensive, convenient

Abbreviation: PPIs, proton pump inhibitors.

Occasionally, specialized testing such as serum gastrin and gastric acid analysis or sham feeding may be needed in individuals with complicated or refractory PUD (see "Zollinger-Ellison Syndrome [ZES]," below). Screening for aspirin or NSAIDs (blood or urine) may also be necessary in refractory *H. pylori*–negative PUD patients.

TREATMENT

TREATMENT

Peptic Ulcer Disease

Before the discovery of *H. pylori*, the therapy of PUD was centered on the old dictum by Schwartz of "no acid, no ulcer." Although acid secretion is still important in the pathogenesis of PUD, eradication of *H. pylori* and therapy/prevention of NSAID-induced disease is the mainstay of treatment. A summary of commonly used drugs for treatment of acid peptic disorders is shown in Table 317-3.

ACID-NEUTRALIZING/INHIBITORY DRUGS

Antacids

Before we understood the important role of histamine in stimulating parietal cell activity, neutralization of secreted acid with antacids constituted the main form of therapy for peptic ulcers. They are now rarely, if ever, used as the primary therapeutic agent but instead are often used by patients for symptomatic relief of dyspepsia. The most commonly used agents are mixtures of aluminum hydroxide and magnesium hydroxide. Aluminum hydroxide can produce constipation and phosphate depletion; magnesium hydroxide may cause loose stools. Many of the commonly used antacids (e.g., Maalox, Mylanta) have a combination of both aluminum and magnesium hydroxide in order to avoid these side effects. The magnesium-containing preparation should not be used in chronic renal failure patients because of possible hypermagnesemia, and aluminum may cause chronic neurotoxicity in these patients.

Calcium carbonate and sodium bicarbonate are potent antacids with varying levels of potential problems. The long-term use of calcium carbonate (converts to calcium chloride in the stomach) can lead to milk-alkali syndrome (hypercalcemia, hyperphosphatemia with possible renal calcinosis and progression to renal insufficiency). Sodium bicarbonate may induce systemic alkalosis.

H_2 Receptor Antagonists

Four of these agents are presently available (cimetidine, ranitidine, famotidine, and nizatidine), and their structures share homology with histamine. Although each has different potency, all will significantly inhibit basal and stimulated acid secretion to comparable levels when used at therapeutic doses. Moreover, similar ulcer-healing rates are achieved with each drug when used at the correct dosage. Presently, this class of drug is often used for treatment of active ulcers (4–6 weeks) in combination with antibiotics directed at eradicating *H. pylori* (see below).

Cimetidine was the first H_2 receptor antagonist used for the treatment of acid peptic disorders.

Cimetidine may have weak antiandrogenic side effects resulting in reversible gynecomastia and impotence, primarily in patients receiving high doses for prolonged periods of time (months to years). In view of cimetidine's ability to inhibit cytochrome P450, careful monitoring of drugs such as warfarin, phenytoin, and theophylline is indicated with long-term usage. Other rare reversible adverse effects reported with cimetidine include confusion and elevated levels of serum aminotransferases, creatinine, and serum prolactin. Ranitidine, famotidine, and nizatidine are more potent H_2 receptor antagonists than cimetidine. Each can be used once a day at bedtime for ulcer prevention, which was commonly done before the discovery of *H. pylori* and the development of proton pump inhibitors (PPIs). Patients may develop tolerance to H_2 blockers, a rare event with PPIs (see below). Comparable nighttime dosing regimens are cimetidine 800 mg, ranitidine 300 mg, famotidine 40 mg, and nizatidine 300 mg.

Additional rare, reversible systemic toxicities reported with H_2 receptor antagonists include pancytopenia, neutropenia, anemia, and thrombocytopenia, with a prevalence rate varying from 0.01 to 0.2%. Cimetidine and ranitidine (to a lesser extent) can bind to hepatic cytochrome P450; famotidine and nizatidine do not.

Proton Pump (H^+,K^+-ATPase) Inhibitors

Omeprazole, esomeprazole, lansoprazole, rabeprazole, and pantoprazole are substituted benzimidazole derivatives that covalently bind and irreversibly inhibit H^+,K^+-ATPase. Esomeprazole is the S-enantiomer of omeprazole, which is a racemic mixture of both S- and R-optical isomers. The R-isomer of lansoprazole, dexlansoprazole, is the most recent PPI approved for clinical use. Its reported advantage is a dual delayed-release system, aimed at improving treatment of gastroesophageal reflux disease (GERD). These are the most potent acid inhibitory agents available. Omeprazole and lansoprazole are the PPIs that have been used for the longest time. Both are acid-labile and are administered as enteric-coated granules in a sustained-release capsule that dissolves within the small intestine at a pH of 6. Lansoprazole is available in an orally disintegrating tablet that can be taken with or without water, an advantage for individuals who have significant dysphagia. Absorption kinetics are similar to the capsule. In addition, a lansoprazole-naproxen combination preparation that has been made available is targeted at decreasing NSAID-related GI injury (see below). Omeprazole is available as nonenteric-coated granules mixed with sodium bicarbonate in a powder form that can be administered orally or via gastric tube. The sodium bicarbonate has two purposes: to protect the omeprazole from acid degradation and to promote rapid gastric alkalinization and subsequent proton pump activation, which facilitates rapid action of the PPI. Pantoprazole and rabeprazole are available as enteric-coated tablets. Pantoprazole is also available as a parenteral formulation for intravenous use. These agents are lipophilic compounds; upon entering the parietal cell, they are protonated and trapped within the acid environment of the tubulovesicular and canalicular system. These agents potently inhibit all phases of gastric acid secretion. Onset of action is rapid, with a maximum acid inhibitory effect between 2 and 6 h after administration and duration of

inhibition lasting up to 72–96 h. With repeated daily dosing, progressive acid inhibitory effects are observed, with basal and secretagogue-stimulated acid production being inhibited by >95% after 1 week of therapy. The half-life of PPIs is ~18 h; thus, it can take between 2 and 5 days for gastric acid secretion to return to normal levels once these drugs have been discontinued. Because the pumps need to be activated for these agents to be effective, their efficacy is maximized if they are administered before a meal (except for the immediate-release formulation of omeprazole) (e.g., in the morning before breakfast). Mild to moderate hypergastrinemia has been observed in patients taking these drugs. Carcinoid tumors developed in some animals given the drugs preclinically; however, extensive experience has failed to demonstrate gastric carcinoid tumor development in humans. Serum gastrin levels return to normal levels within 1–2 weeks after drug cessation. Rebound gastric acid hypersecretion has been described in *H. pylori*–negative individuals after discontinuation of PPIs. It occurs even after relatively short-term usage (2 months) and may last for up to 2 months after the PPI has been discontinued. The mechanism involves gastrin-induced hyperplasia and hypertrophy of histamine-secreting ECL cells. The clinical relevance of this observation is that individuals may have worsening symptoms of GERD or dyspepsia upon stopping the PPI. Gradual tapering of the PPI and switching to an H_2 receptor antagonist may prevent this from occurring. *H. pylori*–induced inflammation and concomitant decrease in acid production may explain why this does not occur in *H. pylori*–positive patients. IF production is also inhibited, but vitamin B_{12}-deficiency anemia is uncommon, probably because of the large stores of the vitamin. As with any agent that leads to significant hypochlorhydria, PPIs may interfere with absorption of drugs such as ketoconazole, ampicillin, iron, and digoxin. Hepatic cytochrome P450 can be inhibited by the earlier PPIs (omeprazole, lansoprazole). Rabeprazole, pantoprazole, and esomeprazole do not appear to interact significantly with drugs metabolized by the cytochrome P450 system. The overall clinical significance of this observation is not definitely established. Caution should be taken when using theophylline, warfarin, diazepam, atazanavir, and phenytoin concomitantly with PPIs.

The list of potential side effects with long-term PPI use has steadily grown over the years. These agents are commonly used since several formulations have become available as over the counter medications. Moreover, up to 70% of current prescriptions for long-term PPIs may be unwarranted. Interpretation of the multiple studies should take into consideration that the vast majority were retrospective observational studies in which confounding factors could not be accounted for entirely.

Long-term acid suppression, especially with PPIs, has been associated with a higher incidence of community-acquired pneumonia as well as community and hospital acquired *Clostridium difficile*–associated disease. A meta-analysis showed a 74% increased risk of *Clostridium difficile* infection and a 2.5-fold higher risk of reinfection as compared to non-users. In light of these concerns the FDA published a safety alert regarding the association between *Clostridium difficile* infection and PPI use. Although the risk of spontaneous bacterial peritonitis in cirrhotics was thought to be increased, the data here are less supportive. The impact of PPI-induced changes in the host microbiome is postulated to play a role in the increased risk of infection, but this theory needs to be confirmed. These observations require confirmation but should alert the practitioner to take caution when recommending these agents for long-term use, especially in elderly patients at risk for developing pneumonia or *Clostridium difficile* infection.

A population-based study revealed that long-term use of PPIs was associated with the development of hip fractures in older women. The absolute risk of fracture remained low despite an observed increase associated with the dose and duration of acid suppression. The mechanism for this observation is not clear, and this finding must be confirmed before making broad recommendations regarding the discontinuation of these agents in patients who benefit from them. Long-term use of PPIs has also been implicated in the development of iron, vitamin B12, and magnesium deficiency. A meta-analysis of nine observational studies found a 40% increase in hypomagnesemia in PPI users as compared to non-users. One approach to consider in patients needing to take PPI's long term, is to check a complete blood count looking for evidence of anemia due to iron or B12 deficiency, vitamin B12 level and a magnesium level after 1–2 years of PPI use but these recommendations are not evidence-based nor recommended by expert opinion. PPIs may exert a negative effect on the antiplatelet effect of clopidogrel. Although the evidence is mixed and inconclusive, a small increase in mortality and readmission rate for coronary events was seen in patients receiving a PPI while on clopidogrel in earlier studies. Subsequently, three meta-analyses reported an inverse correlation between clopidogrel and PPI use; therefore, the influence of this drug interaction on mortality is not clearly established. The mechanism involves the competition of the PPI and clopidogrel with the same cytochrome P450 (CYP2C19). Whether this is a class effect of PPIs is unclear; there appears to be at least a theoretical advantage of pantoprazole over the other PPIs, but this has not been confirmed. This drug interaction is particularly relevant in light of the common use of aspirin and clopidogrel for prevention of coronary events and the efficacy of PPIs in preventing GI bleeding in these patients. The FDA has made several recommendations while awaiting further evidence to clarify the impact of PPI therapy on clopidogrel use. Health care providers should continue to prescribe clopidogrel to patients who require it and should reevaluate the need for starting or continuing treatment with a PPI. From a practical standpoint, additional recommendations to consider include the following: Patients taking clopidogrel with aspirin, especially with other GI risk factors for bleeding, should receive GI protective therapy. Although high-dose H_2 blockers have been considered an option, these do not appear to be as effective as PPIs. If PPIs are to be given, some have recommended that there be a 12-h separation between administration of the PPI and clopidogrel to minimize competition of the two agents with the involved cytochrome P450. One option is to give the PPI 30 min before breakfast and the clopidogrel at bedtime. Insufficient data are available to firmly recommend one PPI over another. Additional concerning side effects with long-term PPI use include increased cardiac risks independent of clopidogrel use, dementia, acute and chronic kidney injury. Again, the data are often retrospective and confounding variables were not consistently eliminated thus making it difficult to develop definitive association between PPIs and the toxicities outlined. A summary of the side effects with the corresponding relative risks is shown in **Table 317-4**. Ultimately, heightened awareness of inappropriate long-term use of PPIs is paramount. Patients aged ≥65 years of age have a higher risk for some of the long-term side effects of PPIs highlighted above, in part due to the higher prevalence of concomitant chronic diseases. It is

therefore essential to carefully select individuals, especially among the elderly, who need long-term PPI therapy and discontinue it in those individuals who do not need it.

Development of novel acid inhibitory agents continues in an attempt to primarily address the need for better agents to treat GERD. For example, modified H_2 blockers with greater potency and duration as well as novel PPIs with longer half-life and potency are under study. For example, tenatoprazole is a PPI containing an imidazopyridine ring instead of a benzimidazole ring, which promotes irreversible proton pump inhibition. This agent has a longer half-life than the other PPIs and may be beneficial for inhibiting nocturnal acid secretion, which has significant relevance in GERD. Additional PPIs with longer half-life and combined with other agents are being studied but the details are beyond the scope of this chapter. A second new class of agents is the potassium-competitive acid pump antagonists (P-CABs). These compounds inhibit gastric acid secretion via potassium competitive binding of the H^+,K^+-ATPase. Revaprazan and venoprazan are the first two agents approved for use in Korea and Japan, respectively.

CYTOPROTECTIVE AGENTS

Sucralfate

Sucralfate is a complex sucrose salt in which the hydroxyl groups have been substituted by aluminum hydroxide and sulfate. This compound is insoluble in water and becomes a viscous paste within the stomach and duodenum, binding primarily to sites of active ulceration. Sucralfate may act by several mechanisms: serving as a physicochemical barrier, promoting a trophic action by binding growth factors such as EGF, enhancing prostaglandin synthesis, stimulating mucus and bicarbonate secretion, and enhancing mucosal defense and repair. Toxicity from this drug is rare, with constipation being most common (2–3%). It should be avoided in patients with chronic renal insufficiency to prevent aluminum-induced neurotoxicity. Hypophosphatemia and gastric bezoar formation have also been reported rarely. Standard dosing of sucralfate is 1 g qid.

Bismuth-Containing Preparations

Sir William Osler considered bismuth-containing compounds the drug of choice for treating PUD. The resurgence in the use of these agents is due to their effect against *H. pylori.* Colloidal bismuth subcitrate (CBS) and bismuth subsalicylate (BSS, Pepto-Bismol) are the most widely used preparations. The mechanism by which these agents induce ulcer healing is unclear. Adverse effects with short-term use include black stools, constipation, and darkening of the tongue. Long-term use with high doses, especially with the avidly absorbed CBS, may lead to neurotoxicity. These compounds are commonly used as one of the agents in an anti-*H. pylori* regimen (see below).

Prostaglandin Analogues

In view of their central role in maintaining mucosal integrity and repair, stable prostaglandin analogues were developed for the treatment of PUD. The mechanism by which this rapidly absorbed drug provides its therapeutic effect is through enhancement of mucosal defense and repair. The most common toxicity noted with this drug is diarrhea (10–30% incidence). Other major toxicities include uterine bleeding and contractions; misoprostol is contraindicated in women who may be pregnant, and women of childbearing age must be made clearly aware of this potential drug toxicity. The standard therapeutic dose is 200 μg qid.

Miscellaneous Drugs

A number of drugs including anticholinergic agents and tricyclic antidepressants were used for treating acid peptic disorders, but in light of their toxicity and the development of potent antisecretory agents, these are rarely, if ever, used today.

THERAPY OF *H. PYLORI*

The physician's goal in treating PUD is to provide relief of symptoms (pain or dyspepsia), promote ulcer healing, and ultimately prevent ulcer recurrence and complications. The greatest influence of understanding the role of *H. pylori* in peptic disease has been the ability to prevent recurrence. Documented eradication of *H. pylori* in patients with PUD is associated with a dramatic decrease in ulcer recurrence to <10–20% as compared to 59% in GU patients and 67% in DU patients when the organism is not eliminated. Eradication of the organism may lead to diminished recurrent ulcer bleeding. The effect of its eradication on ulcer perforation is unclear.

Extensive effort has been made in determining who of the many individuals with *H. pylori* infection should be treated. The common conclusion arrived at by multiple consensus conferences around the world is that *H. pylori* should be eradicated in patients with documented PUD. This holds true independent of time of presentation (first episode or not), severity of symptoms, presence of confounding factors such as ingestion of NSAIDs, or whether the ulcer is in remission. Some have advocated treating patients with a history of documented PUD who are found to be *H. pylori*–positive by stool antigen or breath testing. Between 60 and 90% of patients with gastric MALT lymphoma experience complete remission of the tumor in response to *H. pylori* eradication. The Maastricht IV/Florence Consensus Report recommends a test-and-treat approach for patients with uninvestigated dyspepsia if the local incidence of *H. pylori* is >20%. The American College of Gastroenterology (ACG) clinical guidelines (Developed for North America) recommend that individuals aged <60 with uninvestigated dyspepsia should be tested and treated for *H. pylori*. In addition, recommendations from this consensus report and the ACG clinical guidelines include testing and offering eradication of *H. pylori* in patients who will be using NSAIDs (including low-dose aspirin) on a long-term basis, especially if there is a prior history of PUD. These individuals will require continued PPI treatment as well as eradication treatment, because eradication of the organism alone does not eliminate the risk of gastroduodenal ulcers in patients already receiving long-term NSAIDs. Treating patients with NUD to prevent gastric cancer or patients with GERD requiring long-term acid suppression remains

controversial. Guidelines from the ACG suggest eradication of *H. pylori* in patients who have undergone resection of early gastric cancer. The Maastricht IV/Florence Consensus Report also evaluated *H. pylori* treatment in gastric cancer prevention and recommends that eradication should be considered in the following situations: first-degree relatives of family members with gastric cancer; patients with previous gastric neoplasm treated by endoscopic or subtotal resection; individuals with a risk of gastritis (severe pangastritis or body-predominant gastritis) or severe atrophy; patients with gastric acid inhibition for >1 year; individuals with strong environmental risk factors for gastric cancer (heavy smoking; high exposure to dust, coal, quartz, or cement; and/or work in quarries); and *H. pylori*–positive patients with a fear of gastric cancer. Finally the ACG clinical guidelines recommend testing and offering *H. pylori* eradication to patients with unexplained iron deficiency anemia and idiopathic thrombocytopenic purpura.

Multiple drugs have been evaluated in the therapy of *H. pylori.* No single agent is effective in eradicating the organism. Combination therapy for 14 days provides the greatest efficacy, although regimens based on sequential administration of antibiotics also appear promising (see below). A shorter administration course (7–10 days), although attractive, has not proved as successful as the 14-day regimens. The agents used with the greatest frequency include amoxicillin, metronidazole, tetracycline, clarithromycin, and bismuth compounds.

Suggested treatment regimens for *H. pylori* are outlined in **Table 317-5** Choice of a particular regimen will be influenced by several factors, including efficacy, patient tolerance, existing antibiotic resistance, prior antibiotic use and cost of the drugs. The aim for initial eradication rates should be 85–90%. Dual therapy (PPI plus amoxicillin, PPI plus clarithromycin, ranitidine bismuth citrate [Tritec] plus clarithromycin) is not recommended in view of studies demonstrating eradication rates of <80–85%. The combination of bismuth, metronidazole, and tetracycline was the first triple regimen found effective against *H. pylori.* The combination of two antibiotics plus either a PPI, H_2 blocker, or bismuth compound has comparable success rates. Addition of acid suppression assists in providing early symptom relief and enhances bacterial eradication.

Triple therapy, although effective, has several drawbacks, including the potential for poor patient compliance and drug-induced side effects. Compliance is being addressed by simplifying the regimens so that patients can take the medications twice a day. Simpler (dual therapy) and shorter regimens (7 and 10 days) are not as effective as triple therapy for 14 days. Two anti-*H. pylori* regimens are available in prepackaged formulation: Prevpac (lansoprazole, clarithromycin, and amoxicillin) and Helidac (BSS, tetracycline, and metronidazole). The contents of the Prevpac are to be taken twice per day for 14 days, whereas Helidac constituents are taken four times per day with an antisecretory agent (PPI or H_2 blocker), also for at least 14 days. Clarithromycin-based triple therapy should be avoided in settings where *H. pylori* resistance to this agent exceeds 15%.

Side effects have been reported in up to 20–30% of patients on triple therapy. Bismuth may cause black stools, constipation, or darkening of the tongue. The most feared complication with amoxicillin is pseudomembranous colitis, but this occurs in <1–2% of patients. Amoxicillin can also lead to antibiotic-associated diarrhea, nausea, vomiting, skin rash, and allergic reaction. Concomitant use of probiotics may ameliorate some of the antibiotic side effects (see below). Tetracycline has been reported to cause rashes and, very rarely, hepatotoxicity and anaphylaxis.

One important concern with treating patients who may not need therapy is the potential for development of antibiotic-resistant strains. The incidence and type of antibiotic-resistant *H. pylori* strains vary worldwide. Strains resistant to metronidazole, clarithromycin, amoxicillin, and tetracycline have been described, with the latter two being uncommon. Antibiotic-resistant strains are the most common cause for treatment failure in compliant patients. Unfortunately, in vitro resistance does not predict outcome in patients. Culture and sensitivity testing of *H. pylori* is not performed routinely. Although resistance to metronidazole has been found in as many as 30% of isolates in North America and 80% in developing countries, triple therapy is effective in eradicating the organism in >50% of patients infected with a resistant strain. Clarithromycin resistance is seen in 13–16% of individuals in the United States, with resistance to amoxicillin being <1% and resistance to both metronidazole and clarithromycin in the 5% range. Resistance to tetracycline and rifabutin (see below) is reported to be <2% in the United States. In light of the paucity of *H. pylori* antibiotic real time resistance data, asking the patient about prior antibiotic exposure should be included in the decision-making and used as a surrogate for potential antibiotic resistance especially when it comes to prior macrolide use. Clarithromycin use should be excluded in patients with prior macrolide usage. An approach to antibiotic selection for *H. pylori* therapy has been recommended in the ACG clinical guidelines **(Fig. 317-12)**.

Failure of *H. pylori* eradication with triple therapy in a compliant patient is usually due to infection with a resistant organism. A series of salvage therapies for *H. pylori* are shown in **Table 317-6**. Quadruple therapy (Table 317-4), where clarithromycin is substituted for metronidazole (or vice versa), should be the next step. The combination of PPI, amoxicillin, and rifabutin for 10 days has also been used successfully (86% cure rate) in patients infected with resistant strains. Additional regimens considered for second-line therapy include levofloxacin-based triple therapy (levofloxacin, amoxicillin, PPI) for 10 days and furazolidone-based triple therapy (furazolidone, amoxicillin, PPI) for 14 days. Unfortunately, there is no universally accepted treatment regimen recommended for patients in whom two courses of antibiotics have failed. If eradication is still not achieved in a compliant patient, then culture and sensitivity of the organism should be considered. One challenge with this approach is that culture and sensitivity testing is cumbersome and not widely available, thus *H. pylori* resistance data within specific communities are often not available. Non-culture-based approaches using molecular markers to determine potential resistance through stool testing are being developed but are not widely available. Additional factors that may lower eradication rates include the patient's country of origin (higher in Northeast Asia than other parts of Asia or Europe) and cigarette smoking. In addition, meta-analysis suggests that even the most effective regimens (quadruple therapy including PPI, bismuth, tetracycline, and metronidazole and triple therapy including PPI, clarithromycin, and amoxicillin) may have suboptimal eradication rates (<80%), thus, demonstrating the need for the development of more efficacious treatments.

In view of the observation that 15–25% of patients treated with first-line therapy may still remain infected with the organism, new approaches to treatment have been explored. One promising approach is sequential therapy. Regimens examined consist of 5 days of amoxicillin and a PPI, followed by an additional 5 days of PPI plus tinidazole and clarithromycin or levofloxacin. One promising regimen that has the benefit of being shorter in duration, easier to take, and less expensive is 5 days of concomitant therapy (PPI twice daily, amoxicillin 1 g twice daily, levofloxacin 500 mg twice daily, and tinidazole 500 mg twice daily). Initial studies have demonstrated eradication rates of >90% with good patient tolerance. Confirmation of these findings and applicability of this approach in the United States are needed, although some experts are recommending abandoning clarithromycin-based triple therapy in the United States for the concomitant therapy or the alternative sequential therapies highlighted above.

Innovative non–antibiotic-mediated approaches have been explored in an effort to improve eradication rates of *H. pylori*. Pretreatment of patients with *N*-acetylcysteine as a mucolytic agent to destroy the *H. pylori* biofilm and therefore impair antibiotic resistance has been examined, but more studies are needed to confirm the applicability of this approach. In vitro studies suggest that certain probiotics like *Lactobacillus* or its metabolites can inhibit *H. pylori*. Administration of probiotics has been attempted in several clinical studies in an effort to maximize antibiotic-mediated eradication with varying results. Overall, it appears that the use of certain probiotics, such as *Lactobacillus* spp., *Saccharomyces* spp., *Bifidobacterium* spp., and *Bacillus clausii*, did not alter eradication rates but importantly decreased antibiotic-associated side effects including nausea, dysgeusia, diarrhea, and abdominal discomfort/pain, resulting in enhanced tolerability of *H. pylori* therapies. Additional studies are needed to confirm the potential benefits of probiotics in this setting.

Reinfection after successful eradication of *H. pylori* is rare in the United States (<1% per year). If recurrent infection occurs within the first 6 months after completing therapy, the most likely explanation is recrudescence as opposed to reinfection.

THERAPY OF NSAID-RELATED GASTRIC OR DUODENAL INJURY

Medical intervention for NSAID-related mucosal injury includes treatment of an active ulcer and primary prevention of future injury. Recommendations for the treatment and primary prevention of NSAID-related mucosal injury are listed in **Table 317-7**. Ideally, the injurious agent should be stopped as the first step in the therapy of an active NSAID-induced ulcer. If that is possible, then treatment with one of the acid inhibitory agents (H_2 blockers, PPIs) is indicated. Cessation of NSAIDs is not always possible because of the patient's severe underlying disease. Only PPIs can heal GUs or DUs, independent of whether NSAIDs are discontinued.

The widespread use of NSAIDs has created some concern due to the increasing likelihood of GI and CV side effects associated with these agents. The approach to primary prevention has included avoiding the agent, using the lowest possible dose of the agent for the shortest period of time possible, using NSAIDs that are theoretically less injurious, using newer topical NSAID preparations, and/or using concomitant medical therapy to prevent NSAID-induced injury. Several nonselective NSAIDs that are associated with a lower likelihood of GI and CV toxicity include naproxen and ibuprofen, although the beneficial effect may be eliminated if higher dosages of the agents are used. Primary prevention of NSAID-induced ulceration can be accomplished by misoprostol (200 μg qid) or a PPI. High-dose H_2 blockers (famotidine, 40 mg bid) have also shown some promise in preventing endoscopically documented ulcers, although PPIs are superior. The highly selective COX-2 inhibitors, celecoxib and rofecoxib, are 100 times more selective inhibitors of COX-2 than standard NSAIDs, leading to gastric or duodenal mucosal injury that is comparable to placebo; their utilization led to an increase in CV events and withdrawal from the market. Additional caution was engendered when the CLASS study demonstrated that the advantage of celecoxib in preventing GI complications was offset when low-dose aspirin was used simultaneously. Therefore, gastric protection therapy is required in individuals taking COX-2 inhibitors and aspirin prophylaxis. Finally, much of the work demonstrating the benefit of COX-2 inhibitors and PPIs on GI injury has been performed in individuals of average risk; it is unclear if the same level of benefit will be achieved in high-risk patients. For example, concomitant use of warfarin and a COX-2 inhibitor was associated with rates of GI bleeding similar to those observed in patients taking nonselective NSAIDs. A combination of factors, including withdrawal of the majority of COX-2 inhibitors from the market, the observation that low-dose aspirin appears to diminish the beneficial effect of COX-2 selective inhibitors, and the growing use of aspirin for prophylaxis of CV events, have significantly altered the approach to gastric protective therapy during the use of NSAIDs. A set of guidelines for the approach to the use of NSAIDs was published by the ACG and is shown in **Table 317-8**. Individuals who are not at risk for CV events do not use aspirin and are without risk for GI complications can receive nonselective NSAIDs without gastric protection. In those without CV risk factors but with a high potential risk (prior GI bleeding or multiple GI risk factors) for NSAID-induced GI toxicity, cautious use of a selective COX-2 inhibitor and co-therapy with misoprostol or high-dose PPI are recommended. Individuals at moderate GI risk without cardiac risk factors can be treated with a COX-2 inhibitor alone or with a nonselective NSAID with misoprostol or a PPI. Individuals with CV risk factors, who require low-dose aspirin and have low potential for NSAID-induced toxicity, should be considered for a non-NSAID agent or use of a traditional NSAID such as naproxen (lower CV side effects) in combination with gastric protection, if warranted. Finally, individuals with CV and GI risks who require aspirin must be considered for non-NSAID therapy, but if that is not an option, then gastric protection with any type of NSAID must be considered. Any patient, regardless of risk status, who is being considered for long-term traditional NSAID therapy, should also be considered for *H. pylori* testing and treatment if positive. Assuring the use of GI protective agents with NSAIDs is difficult, even in high-risk patients. This is in part due to under prescribing of the appropriate protective agent; other times the difficulty is related to patient compliance. The latter may be due to patients forgetting to take multiple pills or preferring not to take the extra pill, especially if they have no GI symptoms. Several NSAID gastroprotective-containing combination pills are now commercially available, including double-dose famotidine with ibuprofen, diclofenac with misoprostol, and naproxen with esomeprazole. Although initial studies suggested improved compliance

and a cost advantage when taking these combination drugs, their clinical benefit over the use of separate pills has not been established. One additional concern with NSAID-induced GI complications is the relatively low rate of primary care provider compliance with established guidelines outlining preventative measures. An intervention including professional education, informatics to facilitate review, and financial incentives for practices to review patients' charts to assess appropriateness showed reduced rate of high-risk prescribing of antiplatelet medications and NSAIDs with a tendency towards improved clinical outcomes. Efforts continue toward developing safer NSAIDs, including topical NSAIDs, NSAID formulations that are rapidly absorbed (diclofenac potassium powder mixed with a buffering agent, Prosorb and SoluMatrix technology), NO–releasing NSAIDs, hydrogen sulfide–releasing NSAIDs, dual COX/5-LOX inhibitors, NSAID prodrugs, or agents that can effectively sequester unbound NSAIDs without interfering with their efficacy.

APPROACH AND THERAPY: SUMMARY

Controversy continues regarding the best approach to the patient who presents with dyspepsia **(Chap. 41)**. The discovery of *H. pylori* and its role in pathogenesis of ulcers has added a new variable to the equation. Previously, if a patient <50 years of age presented with dyspepsia and without alarming signs or symptoms suggestive of an ulcer complication or malignancy, an empirical therapeutic trial with acid suppression was commonly recommended. Although this approach is practiced by some today, an approach presently gaining approval for the treatment of patients with dyspepsia is outlined in **Fig. 317-13**. The referral to a gastroenterologist is for the potential need of endoscopy and subsequent evaluation and treatment if the endoscopy is negative.

Once an ulcer (GU or DU) is documented, the main issue at stake is whether *H. pylori* or an NSAID is involved. With *H. pylori* present, independent of the NSAID status, triple therapy is recommended for 14 days, followed by continued acid-suppressing drugs (H_2 receptor antagonist or PPIs) for a total of 4–6 weeks. *H. pylori* eradication should be documented 4 weeks after completing antibiotics. The test of choice for documenting eradication is the laboratory-based validated monoclonal stool antigen test or a urea breath test (UBT). The patient must be off antisecretory agents when being tested for eradication of *H. pylori* with UBT or stool antigen. Serologic testing is not useful for the purpose of documenting eradication because antibody titers fall slowly and often do not become undetectable. Some recommend that patients with complicated ulcer disease, or who are frail, should be treated with long-term acid suppression, thus making documentation of *H. pylori* eradication a moot point. In view of this discrepancy in practice, it would be best to discuss with the patient the different options available.

Several issues differentiate the approach to a GU versus a DU. GUs, especially of the body and fundus, have the potential of being malignant. Multiple biopsies of a GU should be taken initially; even if these are negative for neoplasm, repeat endoscopy to document healing at 8–12 weeks should be performed, with biopsy if the ulcer is still present. About 70% of GUs eventually found to be malignant undergo significant (usually incomplete) healing. Repeat endoscopy is warranted in patients with DU if symptoms persist despite medical therapy or a complication is suspected.

The majority (>90%) of GUs and DUs heal with the conventional therapy outlined above. A GU that fails to heal after 12 weeks and a DU that does not heal after 8 weeks of therapy should be considered refractory. Once poor compliance and persistent *H. pylori* infection have been excluded, NSAID use, either inadvertent or surreptitious, must be excluded. In addition, cigarette smoking must be eliminated. For a GU, malignancy must be meticulously excluded. Next, consideration should be given to a gastric acid hypersecretory state such as ZES (see "Zollinger-Ellison Syndrome," below) or the idiopathic form, which can be excluded with gastric acid analysis. Although a subset of patients have gastric acid hypersecretion of unclear etiology as a contributing factor to refractory ulcers, ZES should be excluded with a fasting gastrin or secretin stimulation test (see below). More than 90% of refractory ulcers (either DUs or GUs) heal after 8 weeks of treatment with higher doses of PPI (omeprazole, 40 mg/d; lansoprazole 30–60 mg/d). This higher dose is also effective in maintaining remission. Surgical intervention may be a consideration at this point; however, other rare causes of refractory ulcers must be excluded before recommending surgery. Rare etiologies of refractory ulcers that may be diagnosed by gastric or duodenal biopsies include ischemia, Crohn's disease, amyloidosis, sarcoidosis, lymphoma, eosinophilic gastroenteritis, smoking crack cocaine or infection (cytomegalovirus [CMV], tuberculosis, or syphilis).

SURGICAL THERAPY

Surgical intervention in PUD can be viewed as being either elective, for treatment of medically refractory disease, or as urgent/emergent, for the treatment of an ulcer-related complication. The development of pharmacologic and endoscopic approaches for the treatment of peptic disease and its complications has led to a substantial decrease in the number of operations needed for this disorder with a drop of >90% for elective ulcer surgery over the last four decades. Refractory ulcers are an exceedingly rare occurrence. Surgery is more often required for treatment of an ulcer-related complication.

Hemorrhage is the most common ulcer-related complication, occurring in ~15–25% of patients. Bleeding may occur in any age group but is most often seen in older patients (sixth decade or beyond). The majority of patients stop bleeding spontaneously, but endoscopic therapy **(Chap. 315)** is necessary in some. Parenterally and orally administered PPIs also decrease ulcer rebleeding in patients who have undergone endoscopic therapy. Patients unresponsive or refractory to endoscopic intervention will require angiographic intervention or surgery (~5% of transfusion-requiring patients).

Free peritoneal perforation occurs in ~2–3% of DU patients with NSAID-induced GU perforations occurring more commonly. Sudden onset of severe abdominal pain with peritoneal signs and evidence of pneumoperitoneum on abdominal imaging is the classic presentation of a perforated viscous but this presentation occurs in only two-thirds of patients. The latter is especially true in elderly patients (>70 years old), obese individuals and in immunocompromised patients. It is important to keep in mind that as in the case of bleeding, up to 10% of these patients will not have antecedent ulcer symptoms. Delay in diagnosis clearly leads to higher mortality thus early suspicion and intervention with nasogastric suction, intravenous PPI, antibiotics and surgical consultation is essential. Concomitant bleeding may occur in up to 10% of patients with perforation, with mortality being increased substantially. Peptic ulcer can also penetrate into adjacent organs, especially with a posterior DU, which can penetrate into the pancreas, colon, liver, or biliary tree.

Pyloric channel ulcers or DUs can lead to gastric outlet obstruction in ~2–3% of patients. This can result from chronic scarring or from impaired motility due to inflammation and/or edema with pylorospasm. Patients may present with early satiety, nausea, vomiting of undigested food, and weight loss. Conservative management with nasogastric suction, intravenous hydration/nutrition, and antisecretory agents is indicated for 7–10 days with the hope that a functional obstruction will reverse. If a mechanical obstruction persists, endoscopic intervention with balloon dilation may be effective. Surgery should be considered if all else fails.

Specific Operations for Duodenal Ulcers

Surgical treatment was originally designed to decrease gastric acid secretion. Operations most commonly performed include (1) vagotomy and drainage (by pyloroplasty, gastroduodenostomy, or gastrojejunostomy), (2) highly selective vagotomy (which does not require a drainage procedure), and (3) vagotomy with antrectomy. The specific procedure performed is dictated by the underlying circumstances: elective versus emergency, the degree and extent of duodenal ulceration, the etiology of the ulcer (*H. pylori*, NSAIDs, malignancy), and the expertise of the surgeon. Moreover, the trend has been toward a dramatic decrease in the need for surgery for treatment of refractory PUD, and when needed, minimally invasive and anatomy-preserving operations are preferred.

Vagotomy is a component of each of these procedures and is aimed at decreasing acid secretion through ablating cholinergic input to the stomach. Unfortunately, both truncal and selective vagotomy (preserves the celiac and hepatic branches) result in gastric atony despite successful reduction of both basal acid output (BAO; decreased by 85%) and maximal acid output (MAO; decreased by 50%). Drainage through pyloroplasty or gastroduodenostomy is required in an effort to compensate for the vagotomy-induced gastric motility disorder. This procedure has an intermediate complication rate and a 10% ulcer recurrence rate. To minimize gastric dysmotility, highly selective vagotomy (also known as parietal cell, super-selective, or proximal vagotomy) was developed. Only the vagal fibers innervating the portion of the stomach that contains parietal cells is transected, thus leaving fibers important for regulating gastric motility intact. Although this procedure leads to an immediate decrease in both BAO and stimulated acid output, acid secretion recovers over time. By the end of the first postoperative year, basal and stimulated acid output are ~30 and 50%, respectively, of preoperative levels. Ulcer recurrence rates are higher with highly selective vagotomy (≥10%), although the overall complication rates are the lowest of the three procedures.

The procedure that provides the lowest rates of ulcer recurrence (1%) but has the highest complication rate is vagotomy (truncal or selective) in combination with antrectomy. Antrectomy is aimed at eliminating an additional stimulant of gastric acid secretion, gastrin. Two principal types of reanastomoses are used after antrectomy: gastroduodenostomy (Billroth I) or gastrojejunostomy (Billroth II) **(Fig. 317-14)**. Although Billroth I is often preferred over II, severe duodenal inflammation or scarring may preclude its performance. Prospective, randomized studies confirm that partial gastrectomy followed by Roux-en-Y reconstruction leads to a significantly better clinical, endoscopic, and histologic outcome than Billroth II reconstruction.

Of these procedures, highly selective vagotomy may be the one of choice in the elective setting, except in situations where ulcer recurrence rates are high (prepyloric ulcers and those refractory to medical therapy). Selection of vagotomy and antrectomy may be more appropriate in these circumstances.

These procedures have been traditionally performed by standard laparotomy. The advent of laparoscopic surgery has led several surgical teams to successfully perform highly selective vagotomy, truncal vagotomy/pyloroplasty, and truncal vagotomy/antrectomy through this approach. An increase in the number of laparoscopic procedures for treatment of PUD has occurred. Laparoscopic repair of perforated peptic ulcers is safe, feasible for the experienced surgeon and is associated with decreased postoperative pain, although it does take longer than an open approach. Moreover, no difference between the two approaches is noted in postoperative complications or length of hospital stay.

Specific Operations for GUs

The location and the presence of a concomitant DU dictate the operative procedure performed for a GU. Antrectomy (including the ulcer) with a Billroth I anastomosis is the treatment of choice for an antral ulcer. Vagotomy is performed only if a DU is present. Although ulcer excision with vagotomy and drainage procedure has been proposed, the higher incidence of ulcer recurrence makes this a less desirable approach. Ulcers located near the esophagogastric junction may require a more radical approach, a subtotal gastrectomy with a Roux-en-Y esophagogastrojejunostomy (Csendes' procedure). A less aggressive approach, including antrectomy, intraoperative ulcer biopsy, and vagotomy (Kelling-Madlener procedure), may be indicated in fragile patients with a high GU. Ulcer recurrence approaches 30% with this procedure.

Surgery-Related Complications

Complications seen after surgery for PUD are related primarily to the extent of the anatomic modification performed. Minimal alteration (highly selective vagotomy) is associated with higher rates of ulcer recurrence and less GI disturbance. More aggressive surgical procedures have a lower rate of ulcer recurrence but a greater incidence of GI dysfunction. Overall, morbidity and mortality related to these procedures are quite low. Morbidity associated with vagotomy and antrectomy or pyloroplasty is ≤5%, with mortality ~1%. Highly selective vagotomy has lower morbidity and mortality rates of 1 and 0.3%, respectively.

In addition to the potential early consequences of any intraabdominal procedure (bleeding, infection, thromboembolism), gastroparesis, duodenal stump leak, and efferent loop obstruction can be observed.

Recurrent Ulceration

The risk of ulcer recurrence is directly related to the procedure performed. Ulcers that recur after partial gastric resection tend to develop at the anastomosis (stomal or marginal ulcer). Epigastric abdominal pain is the most frequent presenting complaint (>90%). Severity and duration of pain tend to be more progressive than observed with DUs before surgery.

Ulcers may recur for several reasons, including incomplete vagotomy, inadequate drainage, retained antrum, and, less likely, persistent or recurrent *H. pylori* infection. ZES should have been excluded preoperatively. Surreptitious use of NSAIDs is an important reason for recurrent ulcers after surgery, especially if the initial procedure was done for an NSAID-induced ulcer. Once *H. pylori* and NSAIDs have been excluded as etiologic factors, the question of incomplete vagotomy or retained gastric antrum should be explored. For the latter, fasting plasma gastrin levels should be determined. If elevated, retained antrum or ZES (see below) should be considered. Incomplete vagotomy can be ruled out by gastric acid analysis coupled with sham feeding. In this test, gastric acid output is measured while the patient sees, smells, and chews a meal (without swallowing). The cephalic phase of gastric secretion, which is mediated by the vagus, is being assessed with this study. An increase in gastric acid output in response to sham feeding is evidence that the vagus nerve is intact. A rise in serum pancreatic polypeptide >50% within 30 min of sham feeding is also suggestive of an intact vagus nerve.

Medical therapy with H_2 blockers will heal postoperative ulceration in 70–90% of patients. The efficacy of PPIs has not been fully assessed in this group, but one may anticipate greater rates of ulcer healing compared to those obtained with H_2 blockers. Repeat operation (complete vagotomy, partial gastrectomy) may be required in a small subgroup of patients who have not responded to aggressive medical management.

Afferent Loop Syndromes

Although rarely seen today as a result of the decrease in the performance of Billroth II anastomosis, two types of afferent loop syndrome can occur in patients who have undergone this type of partial gastric resection. The more common of the two is bacterial overgrowth in the afferent limb secondary to stasis. Patients may experience postprandial abdominal pain, bloating, and diarrhea with concomitant malabsorption of fats and vitamin B_{12}. Cases refractory to antibiotics may require surgical revision of the loop. The less common afferent loop syndrome can present with severe abdominal pain and bloating that occur 20–60 min after meals. Pain is often followed by nausea and vomiting of bile-containing material. The pain and bloating may improve after emesis. The cause of this clinical picture is theorized to be incomplete drainage of bile and pancreatic secretions from an afferent loop that is partially obstructed. Cases refractory to dietary measures may need surgical revision or conversion of the Billroth II anastomosis to a Roux-en-Y gastrojejunostomy.

Dumping Syndrome

Dumping syndrome consists of a series of vasomotor and GI signs and symptoms and occurs in patients who have undergone vagotomy and drainage (especially Billroth procedures). Two phases of dumping, early and late, can occur. Early dumping takes place 15–30 min after meals and consists of crampy abdominal discomfort, nausea, diarrhea, belching, tachycardia, palpitations, diaphoresis, light-headedness, and, rarely, syncope. These signs and symptoms arise from the rapid emptying of hyperosmolar gastric contents into the small intestine, resulting in a fluid shift into the gut lumen with plasma volume contraction and acute intestinal distention. Release of vasoactive GI hormones (vasoactive intestinal polypeptide, neurotensin, motilin) is also theorized to play a role in early dumping.

The late phase of dumping typically occurs 90 min to 3 h after meals. Vasomotor symptoms (light-headedness, diaphoresis, palpitations, tachycardia, and syncope) predominate during this phase. This component of dumping is thought to be secondary to hypoglycemia from excessive insulin release.

Dumping syndrome is most noticeable after meals rich in simple carbohydrates (especially sucrose) and high osmolarity. Ingestion of large amounts of fluids may also contribute. Up to 50% of postvagotomy and drainage patients will experience dumping syndrome to some degree early on. Signs and symptoms often improve with time, but a severe protracted picture can occur in up to 1% of patients.

Dietary modification is the cornerstone of therapy for patients with dumping syndrome. Small, multiple (six) meals devoid of simple carbohydrates coupled with elimination of liquids during meals is important. Antidiarrheals and anticholinergic agents are complementary to diet. Guar and pectin, which increase the viscosity of intraluminal contents, may be beneficial in more symptomatic individuals. Acarbose, an α-glucosidase inhibitor that

delays digestion of ingested carbohydrates, has also been shown to be beneficial in the treatment of the late phases of dumping. The somatostatin analogue octreotide has been successful in diet-refractory cases. This drug is administered subcutaneously (50 µg tid), titrated according to clinical response. A long-acting depot formulation of octreotide can be administered once every 28 days and provides symptom relief comparable to the short-acting agent. In addition, patient weight gain and quality of life appear to be superior with the long-acting form.

Postvagotomy Diarrhea

Up to 10% of patients may seek medical attention for the treatment of postvagotomy diarrhea. This complication is most commonly observed after truncal vagotomy, which is rarely performed today. Patients may complain of intermittent diarrhea that occurs typically 1–2 h after meals. Occasionally the symptoms may be severe and relentless. This is due to a motility disorder from interruption of the vagal fibers supplying the luminal gut. Other contributing factors may include decreased absorption of nutrients (see below), increased excretion of bile acids, and release of luminal factors that promote secretion. Diphenoxylate or loperamide is often useful in symptom control. The bile salt–binding agent cholestyramine may be helpful in severe cases. Surgical reversal of a 10-cm segment of jejunum may yield a substantial improvement in bowel frequency in a subset of patients.

Bile Reflux Gastropathy

A subset of post–partial gastrectomy patients who present with abdominal pain, early satiety, nausea, and vomiting will have mucosal erythema of the gastric remnant as the only finding. Histologic examination of the gastric mucosa reveals minimal inflammation but the presence of epithelial cell injury. This clinical picture is categorized as bile or alkaline reflux gastropathy/gastritis. Although reflux of bile is implicated as the reason for this disorder, the mechanism is unknown. Prokinetic agents, cholestyramine, and sucralfate have been somewhat effective treatments. Severe refractory symptoms may require using either nuclear scanning with ^{99m}Tc-HIDA to document reflux or an alkaline challenge test, where 0.1 *N* NaOH is infused into the stomach in an effort to reproduce the patient's symptoms. Surgical diversion of pancreaticobiliary secretions away from the gastric remnant with a Roux-en-Y gastrojejunostomy consisting of a long (50–60 cm) Roux limb has been used in severe cases. Bilious vomiting improves, but early satiety and bloating may persist in up to 50% of patients.

Maldigestion And Malabsorption

Weight loss can be observed in up to 60% of patients after partial gastric resection. Patients can experience a 10% loss of body weight, which stabilizes 3 months postoperatively. A significant component of this weight reduction is due to decreased oral intake. However, mild steatorrhea can also develop. Reasons for maldigestion/malabsorption include decreased gastric acid production, rapid gastric emptying, decreased food dispersion in the stomach, reduced luminal bile concentration, reduced pancreatic secretory response to feeding, and rapid intestinal transit.

Decreased serum vitamin B_{12} levels can be observed after partial gastrectomy. This is usually not due to deficiency of intrinsic factor (IF), since a minimal amount of parietal cells (source of IF) are removed during antrectomy. Reduced vitamin B_{12} may be due to competition for the vitamin by bacterial overgrowth or inability to split the vitamin from its protein-bound source due to hypochlorhydria.

Iron-deficiency anemia may be a consequence of impaired absorption of dietary iron in patients with a Billroth II gastrojejunostomy. Absorption of iron salts is normal in these individuals; thus, a favorable response to oral iron supplementation can be anticipated. Folate deficiency with concomitant anemia can also develop in these patients. This deficiency may be secondary to decreased absorption or diminished oral intake.

Malabsorption of vitamin D and calcium resulting in osteoporosis and osteomalacia is common after partial gastrectomy and gastrojejunostomy (Billroth II). Osteomalacia can occur as a late complication in up to 25% of post–partial gastrectomy patients. Bone fractures occur twice as commonly in men after gastric surgery as in a control population. It may take years before x-ray findings demonstrate diminished bone density. Elevated alkaline phosphatase, reduced serum calcium, bone pain, and pathologic fractures may be seen in patients with osteomalacia. The high incidence of these abnormalities in this subgroup of patients justifies treating them with vitamin D and calcium supplementation indefinitely. Therapy is especially important in females. Copper deficiency has also been reported in patients undergoing surgeries that bypass the duodenum, where copper is primarily absorbed. Patients may present with a rare syndrome that includes ataxia, myelopathy, and peripheral neuropathy.

Gastric Adenocarcinoma

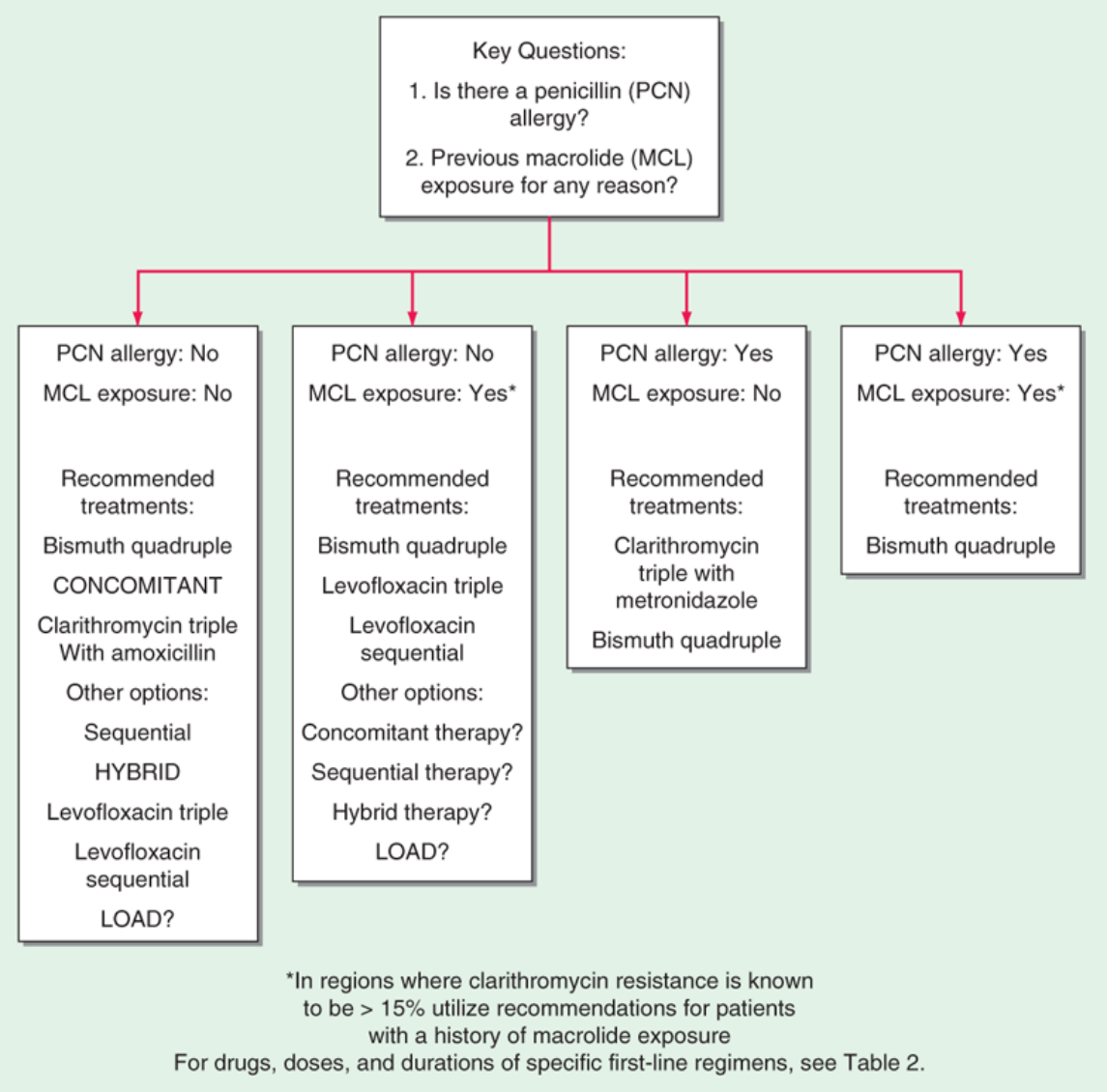
The incidence of adenocarcinoma in the gastric stump is increased 15 years after resection. Some have reported a four- to fivefold increase in gastric cancer 20–25 years after resection. The pathogenesis is unclear but may involve alkaline reflux, bacterial proliferation, or hypochlorhydria. The role of endoscopic screening is not clear, and most guidelines do not support its use.

Additional Complications

Reflux esophagitis and a higher incidence of gallstones and cholecystitis have been reported to patients undergoing subtotal gastrectomy. The latter is thought to be due to decreased gallbladder contractility associated with vagotomy and bypass of the duodenum, leading to decreased postprandial release of cholecystokinin.

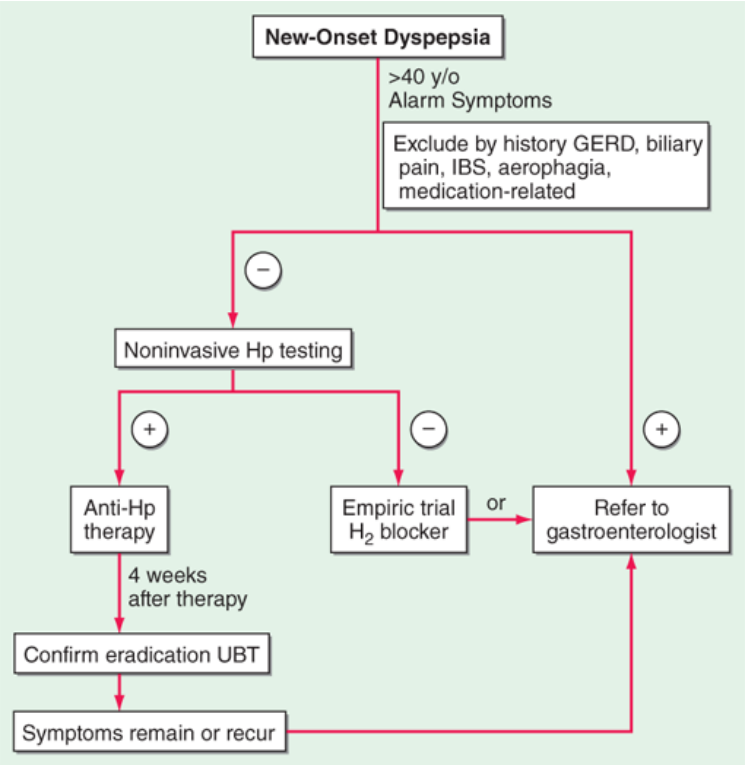
FIGURE 317-12

Approach to selecting antibiotics for patients with *H. pylori* infection. *(Adapted from WD Chey et al: ACG Clinical Guideline: Treatment of Helicobacter pylori Infection. Am J Gastroenterol 112:212, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.
FIGURE 317-13

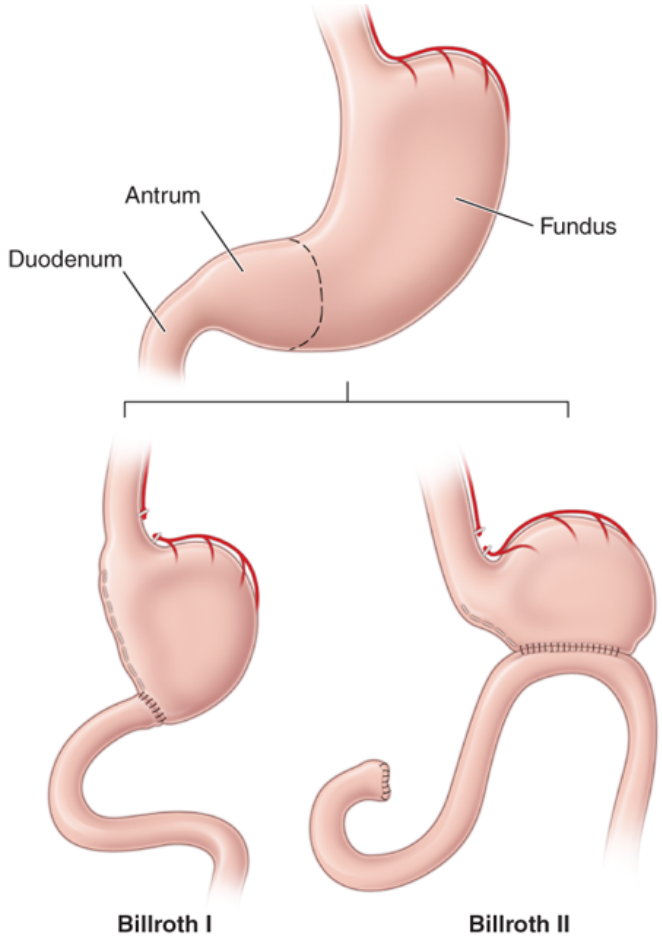
Overview of new-onset dyspepsia. GERD, gastroesophageal reflux disease; Hp, *Helicobacter pylori*; IBS, irritable bowel syndrome; UBT, urea breath test. *(Adapted from BS Anand, DY Graham: Endoscopy 31:215, 1999.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 317-14

Schematic representation of Billroth I and II procedures.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 317-3

Drugs Used in the Treatment of Peptic Ulcer Disease

Drug Type/Mechanism	Examples	Dose
Acid-Suppressing Drugs		
Antacids	Mylanta, Maalox, Tums, Gaviscon	100–140 meq/L 1 and 3 h after meals and hs
H_2 receptor antagonists	Cimetidine	400 mg bid
	Ranitidine	300 mg hs
	Famotidine	40 mg hs
	Nizatidine	300 mg hs
Proton pump inhibitors	Omeprazole	20 mg/d
	Lansoprazole	30 mg/d
	Rabeprazole	20 mg/d
	Pantoprazole	40 mg/d
	Esomeprazole	20 mg/d
	Dexlansoprazole	30 mg/d
Mucosal Protective Agents		
Sucralfate	Sucralfate	1 g qid
Prostaglandin analogue	Misoprostol	200 μg qid
Bismuth-containing compounds	Bismuth subsalicylate (BSS)	See anti-*H. pylori* regimens (Table 317-4)

Abbreviation: hs, at bedtime (*hora somni*).

TABLE 317-4

Evidence Supporting the Potential Adverse Effects of Proton Pump Inhibitor Drugs

Adverse Effect	Adjusted OR (95% CI)
Chronic kidney disease	1.50 (1.11–1.90)
Acute kidney disease	2.52 (2.27–2.79)
Acute interstitial nephritis	3.00 (1.47–6.14)
Hypomagnesemia	1.43 (1.08–1.88)
Clostridium difficile	1.74 (1.47–2.85)
Community-acquired pneumonia	1.34 (1.14–1.57)
Community-acquired pneumonia	1.05 (0.89–1.25)
Bone fracture	1.33 (1.15–1.54)

Abbreviation: OR, odds ratio.

Source: Adapted from AJ Schoenfeld, D Grady: Adverse effects associated with proton pump inhibitors. JAMA Intern Med 176:172, 2016.

TABLE 317-5

Recommended First-Line Therapies for *H. pylori* Infection

Regimen	Drugs (doses)	Dosing Frequency	Duration (days)	FDA Approval
Clarithromycin triple	PPI (standard or double dose)	BID	14	Yes[a]
	Clarithromycin (500mg)			
	Amoxicillin (1 g) or Metronidazole (500 mg TID)			
Bismuth quadruple	PPI (standard dose)	BID	10–14	No[b]
	Bismuth subcitrate (120–300 mg) or Subsalicylate (300 mg)	QID		
	Tetracycline (500 mg)	QID		
	Metronidazole (250–500 mg)	QID (250) TID to QID (500)		
Concomitant	PPI (standard dose)	BID	10–14	No
	Clarithromycin (500 mg)			
	Amoxicillin (1 g)			
	Nitroimidazole (500 mg)[c]			
Sequential	PPI (standard dose)	BID	5–7	No
	PPI, Clarithromycin (500 mg) + Nitroimidazole (500 mg)[c]	BID	5–7	
Hybrid	PPI (standard dose) + Amox (1 g)	BID	7	No
	PPI, Amox, Clarithromycin (500mg), Nitroimidazole (500 mg)[c]	BID	7	
Levofloxacin triple	PPI (standard or double dose) + Amox (1 g)	BID	5–7	No
	Levofloxacin (500 mg)	QD		
	Amox (1 g)	BID		
Levofloxacin sequential	PPI (standard or double dose) + Amox (1 g)	BID	5–7	No
	PPI, Amox, Levofloxacin (500 mg QD), Nitroimidazole (500 mg)[c]	BID	5–7	
LOAD	Levofloxacin (250 mg)	QD	7–10	No
	PPI (double dose)	QD		
	Nitazoxanide (500 mg)	BID		
	Doxycycline (100 mg)	QD		

[a]Several PPI, clarithromycin, and amoxicillin combinations have achieved FDA approval. PPI, clarithromycin and metronidazole is not an FDA-approved treatment regimen.

[b]PPI, bismuth, tetracycline, and metronidazole combined with a PPI for 10 days is an FDA-approved treatment regimen.

[C] Metronidazole or tinidazole.

Abbreviations: BID, twice daily; FDA, Food and Drug Administration; PPI, proton pump inhibitor; TID, three times daily; QD, once daily; QID, four times daily.

Source: Adapted from WD Chey et al: ACG Clinical Guideline: Treatment of *Helicobacter pylori* Infection. Am J Gastroenterol 112:212, 2017.

TABLE 317-6

Salvage Therapies for *H. pylori* Infection

Regimen	Drugs (doses)	Dosing Frequency	Duration (days)	FDA Approval
Bismuth quadruple	PPI (standard dose)	BID	14	No[a]
	Bismuth subcitrate (120–300 mg) or subsalicylate (300 mg)	QID		
	Tetracycline (500 mg)	QID		
	Metronidazole (500 mg)	TID or QID		
Levofloxacin triple	PPI (standard dose)	BID	14	No
	Levofloxacin (500 mg)	QD		
	Amox (1 g)	BID		
Concomitant	PPI (standard dose)	BID	10–14	No
	Clarithromycin (500 mg)	BID		
	Amoxicillin (1 g)	BID		
	Nitroimidazole (500 mg)	BID or TID		
Rifabutin triple	PPI (standard dose)	BID	10	No
	Rifabutin (300 mg)	QD		
	Amox (1 g)	BID		
High-dose dual	PPI (standard to double dose)	TID or QID	14	No
	Amox (1 g TID or 750 mg QID)	TID or QID		

[a]PPI, bismuth, tetracycline, and metronidazole prescribed separately is not an FDA-approved treatment regimen. However, Pylera, a combination product containing bismuth subcitrate, tetracycline, and metronidazole combined with a PPI for 10 days is an FDA-approved treatment regimen.

Abbreviations: BID, twice daily; FDA, Food and Drug Administration; PPI, proton pump inhibitor; TID, three times daily; QD, once daily; QID, four times daily.

Source: Adapted from WD Chey et al: ACG Clinical Guideline: Treatment of *Helicobacter pylori* Infection. Am J Gastroenterol 112:212–238, 2017.

TABLE 317-7

Recommendations for Treatment of NSAID-Related Mucosal Injury

Clinical Setting	Recommendation
Active ulcer	
NSAID discontinued	H_2 receptor antagonist or PPI
NSAID continued	PPI
Prophylactic therapy	Misoprostol
	PPI
	Selective COX-2 inhibitor
H. pylori infection	Eradication if active ulcer present or there is a past history of peptic ulcer disease

Abbreviations: COX-2, isoenzyme of cyclooxygenase; NSAID, nonsteroidal anti-inflammatory drug; PPI, proton pump inhibitor.

TABLE 317-8

Guide to NSAID Therapy

	No/Low NSAID GI Risk	NSAID GI Risk
No CV risk (no aspirin)	Traditional NSAID	Coxib *or* Traditional NSAID + PPI *or* misoprostol Consider non-NSAID therapy
CV risk (consider aspirin)	Traditional NSAID + PPI *or* misoprostol if GI risk warrants gastroprotection Consider non-NSAID therapy	A gastroprotective agent must be added if a traditional NSAID is prescribed Consider non-NSAID therapy

Abbreviations: CV, cardiovascular; GI, gastrointestinal; NSAID, nonsteroidal anti-inflammatory drug; PPI, proton pump inhibitor.

Source: Adapted from AM Fendrick: Am J Manag Care 10:740, 2004. Reproduced with permission of INTELLISPHERE, LLC via Copyright Clearance Center.

RELATED CONDITIONS

ZOLLINGER–ELLISON SYNDROME

Severe peptic ulcer diathesis secondary to gastric acid hypersecretion due to unregulated gastrin release from a non-β cell often well-differentiated neuroendocrine tumor (gastrinoma) defines the components of ZES. Initially, ZES was typified by aggressive and refractory ulceration in which total gastrectomy provided the only chance for enhancing survival. Today it can be cured by surgical resection in up to 40% of patients.

Epidemiology

The true incidence of ZES is unknown but estimates suggest that it varies from 0.1 to 1% of individuals presenting with PUD with 0.1–3 individuals per year having this rare diagnosis. Females are slightly more commonly affected than males, and the majority of patients are diagnosed between ages 30 and 50. Gastrinomas are classified into sporadic tumors (80%) and those associated with multiple endocrine neoplasia (MEN) type 1 (see below). The widespread availability and use of PPIs has led to a decreased patient referral for gastrinoma evaluation, delay in diagnosis, and an increase in false-positive diagnoses of ZES. In fact, diagnosis may be delayed for ≥6 years after symptoms consistent with ZES are displayed.

Pathophysiology

Hypergastrinemia originating from an autonomous neoplasm is the driving force responsible for the clinical manifestations in ZES. Gastrin stimulates acid secretion through gastrin receptors on parietal cells and by inducing histamine release from ECL cells. Gastrin also has a trophic action on gastric epithelial cells. Long-standing hypergastrinemia leads to markedly increased gastric acid secretion through both parietal cell stimulation and increased parietal cell mass. The increased gastric acid output leads to peptic ulcer diathesis, erosive esophagitis, and diarrhea.

Tumor Distribution

Although early studies suggested that the vast majority of gastrinomas occurred within the pancreas, a significant number of these lesions are extrapancreatic. Over 80% of these tumors are found within the hypothetical gastrinoma triangle (confluence of the cystic and common bile ducts superiorly, junction of the second and third portions of the duodenum inferiorly, and junction of the neck and body of the pancreas medially). Duodenal tumors constitute the most common nonpancreatic lesion; between 50 and 75% of gastrinomas are found here. Duodenal tumors are smaller, slower growing, and less likely to metastasize than pancreatic lesions. Less common extrapancreatic sites include stomach, bones, ovaries, heart, liver, and lymph nodes. More than 60% of tumors are considered malignant, with up to 30–50% of patients having multiple lesions or metastatic disease at presentation. Histologically, gastrin-producing cells appear well-differentiated, expressing markers typically found in endocrine neoplasms (chromogranin, neuron-specific enolase).

Clinical Manifestations

Gastric acid hypersecretion is responsible for the signs and symptoms observed in patients with ZES. Peptic ulcer is the most common clinical manifestation, occurring in >90% of gastrinoma patients. Initial presentation and ulcer location (duodenal bulb) may be indistinguishable from common PUD. Clinical situations that should create suspicion of gastrinoma are ulcers in unusual locations (second part of the duodenum and beyond), ulcers refractory to standard medical therapy, ulcer recurrence after acid-reducing surgery, ulcers presenting with frank complications (bleeding, obstruction, and perforation), or ulcers in the absence of *H. pylori* or NSAID ingestion. Symptoms of esophageal origin are present in up to two-thirds of patients with ZES, with a spectrum ranging from mild esophagitis to frank ulceration with stricture and Barrett's mucosa.

Diarrhea, the next most common clinical manifestation, is found in up to 50% of patients. Although diarrhea often occurs concomitantly with acid peptic disease, it may also occur independent of an ulcer. Etiology of the diarrhea is multifactorial, resulting from marked volume overload to the small bowel, pancreatic enzyme inactivation by acid, and damage of the intestinal epithelial surface by acid. The epithelial damage can lead to a mild degree of maldigestion and malabsorption of nutrients. The diarrhea may also have a secretory component due to the direct stimulatory effect of gastrin on enterocytes or the co-secretion of additional hormones from the tumor such as vasoactive intestinal peptide.

Gastrinomas can develop in the presence of MEN 1 syndrome **(Chaps. 80 and 381)** in ~25% of patients. This autosomal dominant disorder involves primarily three organ sites: the parathyroid glands (80–90%), pancreas (40–80%), and pituitary gland (30–60%). The syndrome is caused by inactivating mutations of the *MEN1* tumor suppressor gene found on the long arm of chromosome 11q13. The gene encodes for Menin, which has an important role in DNA replication and transcriptional regulation. A genetic diagnosis is obtained by sequencing of the *MEN1* gene, which can reveal mutations in 70–90% of typical MEN 1 cases. A family may have an unknown mutation, making a genetic diagnosis impossible, and therefore certain individuals will require a clinical diagnosis, which is determined by whether a patient has tumors in two of the three endocrine organs (parathyroid, pancreas/duodenum, or pituitary) or has a family history of MEN 1 and one of the endocrine organ tumors. In view of the stimulatory effect of calcium on gastric secretion, the hyperparathyroidism and hypercalcemia seen in MEN 1 patients may have a direct effect on ulcer disease. Resolution of hypercalcemia by parathyroidectomy reduces gastrin and gastric acid output in gastrinoma patients. An additional distinguishing feature in ZES patients with MEN 1 is the higher incidence of gastric carcinoid tumor development (as compared to patients with sporadic gastrinomas). ZES presents and is diagnosed earlier in MEN 1 patients, and they have a more indolent course as compared to patients with sporadic gastrinoma. Gastrinomas tend to be smaller, multiple, and located in the duodenal wall more often than is seen in patients with sporadic ZES. Establishing the diagnosis of MEN 1 is critical in order to provide genetic counseling to the patient and his or her family and also to determine the recommended surgical approach. Therefore, gastrinoma patients should be screened for MEN I performing a detailed family history and obtaining several serum markers including calcium, parathyroid, prolactin and pancreatic polypeptide levels.

Diagnosis

Biochemical measurements of gastrin and acid secretion in patients suspected of ZES play an important role is establishing this rare diagnosis. Often, patients suspected of having ZES will be treated with a PPI in an effort to ameliorate symptoms and decrease the likelihood of possible acid-related complications. The presence of the PPI, which will lower acid secretion and potentially elevate fasting gastrin levels in normal individuals, will make the diagnostic approach in these individuals somewhat difficult. Significant morbidity related to peptic diathesis has been described when stopping PPIs in gastrinoma patients; therefore, a systematic approach in stopping these agents is warranted (see below). The first step in the evaluation of a patient suspected of having ZES is to obtain a fasting gastrin level. A list of clinical scenarios that should arouse suspicion regarding this diagnosis is shown in **Table 317-9**. Fasting gastrin levels obtained using a dependable assay are usually <150 pg/mL. A normal fasting gastrin, on two separate occasions, especially if the patient is on a PPI, virtually excludes this diagnosis. Virtually all gastrinoma patients will have a gastrin level >150–200

pg/mL. Measurement of fasting gastrin should be repeated to confirm the clinical suspicion. Some of the commercial biochemical assays used for measuring serum gastrin may be inaccurate. Variable specificity of the antibodies used have led to both false-positive and false-negative fasting gastrin levels, placing in jeopardy the ability to make an accurate diagnosis of ZES.

TABLE 317-9

When to Obtain a Fasting Serum Gastrin Level

Multiple ulcers
Ulcers in unusual locations; associated with severe esophagitis; resistant to therapy with frequent recurrences; in the absence of nonsteroidal anti-inflammatory drug ingestion or *H. pylori* infection
Ulcer patients awaiting surgery
Extensive family history for peptic ulcer disease
Postoperative ulcer recurrence
Basal hyperchlorhydria
Unexplained diarrhea or steatorrhea
Hypercalcemia
Family history of pancreatic islet, pituitary, or parathyroid tumor
Prominent gastric or duodenal folds

Multiple processes can lead to an elevated fasting gastrin level, the most frequent of which are gastric hypochlorhydria and achlorhydria, with or without pernicious anemia. Gastric acid induces feedback inhibition of gastrin release. A decrease in acid production will subsequently lead to failure of the feedback inhibitory pathway, resulting in net hypergastrinemia. Gastrin levels will thus be high in patients using antisecretory agents for the treatment of acid peptic disorders and dyspepsia. *H. pylori* infection can also cause hypergastrinemia. Additional causes of elevated gastrin include retained gastric antrum; G cell hyperplasia; gastric outlet obstruction; renal insufficiency; massive small-bowel obstruction; and conditions such as rheumatoid arthritis, vitiligo, diabetes mellitus, and pheochromocytoma. Although a fasting gastrin >10 times normal is highly suggestive of ZES, two-thirds of patients will have fasting gastrin levels that overlap with levels found in the more common disorders outlined above, especially if a PPI is being taken by the patient. The effect of the PPI on gastrin levels and acid secretion will linger several days after stopping the PPI; therefore, it should be stopped for a minimum of 7 days before testing. During this period, the patient should be placed on a histamine H_2 antagonist, such as famotidine, twice to three times per day. Although this type of agent has a short-term effect on gastrin and acid secretion, it needs to be stopped 24 h before repeating fasting gastrin levels or performing some of the tests highlighted below. The patient may take antacids for the final day, stopping them ~12 h before testing is performed. Heightened awareness of complications related to gastric acid hypersecretion during the period of PPI cessation is critical.

The next step at times needed for establishing a biochemical diagnosis of gastrinoma is to assess acid secretion. Nothing further needs to be done if decreased acid output in the absence of a PPI is observed. A pH can be measured on gastric fluid obtained either during endoscopy or through nasogastric aspiration; a pH <3 is suggestive of a gastrinoma, but a pH >3 is not helpful in excluding the diagnosis. In those situations where the pH is >3, formal gastric acid analysis should be performed if available. Normal BAO in nongastric surgery patients is typically <5 meq/h. A BAO >15 meq/h in the presence of hypergastrinemia is considered pathognomonic of ZES, but up to 12% of patients with common PUD may have elevated BAO to a lesser degree that can overlap with levels seen in ZES patients. In an effort to improve the sensitivity and specificity of gastric secretory studies, a BAO/MAO ratio was established using pentagastrin infusion as a way to maximally stimulate acid production, with a BAO/MAO ratio >0.6 being highly suggestive of ZES. Pentagastrin is no longer available in the United States, making measurement of MAO virtually impossible. An endoscopic method for measuring gastric acid output has been developed but requires further validation.

Gastrin provocative tests have been developed in an effort to differentiate between the causes of hypergastrinemia and are especially helpful in patients with indeterminate acid secretory studies. The tests are the secretin stimulation test and the calcium infusion study. The most sensitive and specific gastrin provocative test for the diagnosis of gastrinoma is the secretin study. An increase in gastrin of ≥120 pg within 15 min of secretin injection has a sensitivity and specificity of >90% for ZES. PPI-induced hypochlorhydria or achlorhydria may lead to a false-positive secretin test; thus, this agent must be stopped for 1 week before testing.

The calcium infusion study is less sensitive and specific than the secretin test, which, coupled with it being a more cumbersome study with greater potential for adverse effects, relegates it to rare utilization in the cases where the patient's clinical characteristics are highly suggestive of ZES but the secretin stimulation is inconclusive.

Tumor Localization

Once the biochemical diagnosis of gastrinoma has been confirmed, the tumor must be located. Multiple imaging studies have been used in an effort to enhance tumor localization **(Table 317-10)**. The broad range of sensitivity is due to the variable success rates achieved by the different investigative groups. Endoscopic ultrasound (EUS) permits imaging of the pancreas with a high degree of resolution (<5 mm). This modality is particularly helpful in excluding small neoplasms within the pancreas and in assessing the presence of surrounding lymph nodes and vascular involvement, but it is not very sensitive for finding duodenal lesions. Several types of endocrine tumors express cell-surface receptors for somatostatin, in particular the sub-type 2 (SSTR2). This permits the localization, staging, and prediction of therapeutic response to somatostatin analogues (see below) bygastrinomas. The original functional scinitigraphic tool developed measuring the uptake of the stable somatostatin analogue ^{111}In-pentreotide (OctreoScan) has demonstrated sensitivity and specificity rates of >80%. More recently, PET-CT with ^{68}Ga-DOTATATE has been developed and is superior than Octreoscan for assessing tumor presence in patients with well-differentiated neuroendocrine tumors such as gastrinomas, with sensitivity and specificity of >90%, making it the functional imaging study of choice when available.

TABLE 317-10

Sensitivity of Imaging Studies in Zollinger-Ellison Syndrome

Study	Sensitivity, %	
	Primary Gastrinoma	Metastatic Gastrinoma
Ultrasound	21–28	14
CT scan	55–70	>85
Selective angiography	35–68	33–86
Portal venous sampling	70–90	N/A
SASI	55–78	41
MRI	55–70	>85
OctreoScan	67–86	80–100
EUS	80–100	N/A

Abbreviations: CT, computed tomography; EUS, endoscopic ultrasonography; MRI, magnetic resonance imaging; N/A, not applicable; OctreoScan, imaging with ^{111}In-pentreotide; SASI, selective arterial secretin injection.

Up to 50% of patients have metastatic disease at diagnosis. Success in controlling gastric acid hypersecretion has shifted the emphasis of therapy toward providing a surgical cure. Detecting the primary tumor and excluding metastatic disease are critical in view of this paradigm shift. Once a biochemical diagnosis has been confirmed, the patient should first undergo an abdominal computed tomography (CT) scan, magnetic resonance imaging (MRI), or OctreoScan/PET-CT with ^{68}Ga-DOTATATE (depending on availability) to exclude metastatic disease. Once metastatic disease has been excluded, an experienced endocrine surgeon may opt for exploratory laparotomy with intraoperative ultrasound or transillumination. In other centers, careful examination of the peripancreatic area with EUS, accompanied by endoscopic exploration of the duodenum for primary tumors, will be performed before surgery. Selective arterial secretin injection may be a useful adjuvant for localizing tumors in a subset of patients. The extent of the diagnostic and surgical approach must be carefully balanced with the patient's overall physiologic condition and the natural history of a slow-growing gastrinoma.

TREATMENT

TREATMENT

Zollinger-Ellison Syndrome

Treatment of functional endocrine tumors is directed at ameliorating the signs and symptoms related to hormone overproduction, curative resection of the neoplasm, and attempts to control tumor growth in metastatic disease.

PPIs are the treatment of choice and have decreased the need for total gastrectomy. Initial PPI doses tend to be higher than those used for treatment of GERD or PUD. The initial dose of omeprazole, lansoprazole, rabeprazole, or esomeprazole should be in the range of 60 mg in divided doses in a 24-h period. Dosing can be adjusted to achieve a BAO <10 meq/h (at the drug trough) in surgery-naive patients and to <5 meq/h in individuals who have previously undergone an acid-reducing operation. Although the somatostatin analogue has inhibitory effects on gastrin release from receptor-bearing tumors and inhibits gastric acid secretion to some extent, PPIs have the advantage of reducing parietal cell activity to a greater degree. Despite this, octreotide or lanreotide may be considered as adjunctive therapy to the PPI in patients with tumors that express somatostatin receptors and have peptic symptoms that are difficult to control with high-dose PPI.

The ultimate goal of surgery would be to provide a definitive cure. Improved understanding of tumor distribution has led to immediate cure rates as high as 33% with 10-year disease-free intervals as high as 95% in sporadic gastrinoma patients undergoing surgery. A positive outcome is highly dependent on the experience of the surgical team treating these rare tumors. Surgical therapy of gastrinoma patients with MEN 1 remains controversial because of the difficulty in rendering these patients disease-free with surgery. In contrast to the encouraging postoperative results observed in patients with sporadic disease, only 6% of MEN 1 patients are disease-free 5 years after an operation. Moreover, in contrast to patients with sporadic ZES, the clinical course of MEN 1 patients is benign and rarely leads to disease-related mortality, recommending that early surgery be deferred. Some groups suggest surgery only if a clearly identifiable, nonmetastatic lesion is documented by structural studies. Others advocate a more aggressive approach, where all patients free of hepatic metastasis are explored and all detected tumors in the duodenum are resected; this is followed by enucleation of lesions in the pancreatic head, with a distal pancreatectomy to follow. The outcome of the two approaches has not been clearly defined. Laparoscopic surgical interventions may provide attractive approaches in the future but currently seem to be of some limited benefit in patients with gastrinoma because a significant percentage of the tumors may be extrapancreatic and difficult to localize with a laparoscopic approach. Finally, patients selected for surgery should be individuals whose health status would lead them to tolerate a more aggressive operation and obtain the long-term benefits from such aggressive surgery, which are often witnessed after 10 years.

Therapy of metastatic endocrine tumors in general remains suboptimal; gastrinomas are no exception. In light of the observation that in many instances tumor growth is indolent and that many individuals with metastatic disease remain relatively stable for significant periods of time, many advocate not instituting systemic tumor-targeted therapy until evidence of tumor progression or refractory symptoms not controlled with PPIs are noted. Medical approaches, including biological therapy (IFN-α, long-acting somatostatin analogues, peptide receptor radionuclides), systemic chemotherapy (streptozotocin, 5-fluorouracil, and doxorubicin), and hepatic artery embolization, may lead to significant toxicity without a substantial improvement in overall survival. Use of temozolomide with capecitabine has demonstrated radiographic regression and progression-free survival in patients with well-differentiated NETs in the range of 70% and 18 months respectively. Systemic therapy with radiolabeled somatostatin analogues (Peptide Receptor Radiotherapy, PRRT) has been used in the therapy of metastatic neuroendocrine tumors and appears to be very promising in terms of radiographic, symptom, and progression-free survival, but additional studies are warranted. Several promising therapies are being explored, including radiofrequency ablation or cryoablation of liver lesions and use of agents that block the vascular endothelial growth receptor pathway (sunitinib) or the mammalian target of rapamycin **(Chap. 80)**.

Surgical approaches, including debulking surgery and liver transplantation for hepatic metastasis, have also produced limited benefit.

The overall 5- and 10-year survival rates for gastrinoma patients are 62–75% and 47–53%, respectively. Individuals with the entire tumor resected or those with a negative laparotomy have 5- and 10-year survival rates >90%. Patients with incompletely resected tumors have 5- and 10-year survival rates of 43 and 25%, respectively. Patients with hepatic metastasis have <20% survival at 5 years. Favorable prognostic indicators include primary duodenal wall tumors, isolated lymph node tumor, the presence of MEN 1, and undetectable tumor upon surgical exploration. Poor outcome is seen in patients with shorter disease duration; higher gastrin levels (>10,000 pg/mL); large pancreatic primary tumors (>3 cm); metastatic disease to lymph nodes, liver, and bone; and Cushing's syndrome. Rapid growth of hepatic metastases is also predictive of poor outcome.

STRESS-RELATED MUCOSAL INJURY

Patients suffering from shock, sepsis, massive burns, severe trauma, or head injury can develop acute erosive gastric mucosal changes or frank ulceration with bleeding. Classified as stress-induced gastritis or ulcers, injury is most commonly observed in the acid-producing (fundus and body) portions of the stomach. The most common presentation is GI bleeding, which is usually minimal but can occasionally be life-threatening. Respiratory failure requiring mechanical ventilation and underlying coagulopathy are risk factors for bleeding, which tends to occur 48–72 h after the acute injury or insult.

Histologically, stress injury does not contain inflammation or *H. pylori*; thus, "gastritis" is a misnomer. Although elevated gastric acid secretion may be noted in patients with stress ulceration after head trauma (Cushing's ulcer) and severe burns (Curling's ulcer), mucosal ischemia, breakdown of the

normal protective barriers of the stomach, systemic release of cytokines, poor GI motility, and oxidative stress also play an important role in the pathogenesis. Acid must contribute to injury in view of the significant drop in bleeding noted when acid inhibitors are used as prophylaxis for stress gastritis.

Improvement in the general management of intensive care unit patients has led to a significant decrease in the incidence of GI bleeding due to stress ulceration. The estimated decrease in bleeding is from 20–30% to <5%. This improvement has led to some debate regarding the need for prophylactic therapy. The high mortality associated with stress-induced clinically important GI bleeding (>40%) and the limited benefit of medical (endoscopic, angiographic) and surgical therapy in a patient with hemodynamically compromising bleeding associated with stress ulcer/gastritis support the use of preventive measures in high-risk patients (mechanically ventilated, coagulopathy, multiorgan failure, or severe burns). Metaanalysis comparing H_2 blockers with PPIs for the prevention of stress-associated clinically important and overt GI bleeding demonstrates superiority of the latter without increasing the risk of nosocomial infections, increasing mortality, or prolonging intensive care unit length of stay. Therefore, PPIs are the treatment of choice for stress prophylaxis. Oral PPI is the best option if the patient can tolerate enteral administration. Pantoprazole is available as an intravenous formulation for individuals in whom enteral administration is not possible. If bleeding occurs despite these measures, endoscopy, intraarterial vasopressin, and embolization are options. If all else fails, then surgery should be considered. Although vagotomy and antrectomy may be used, the better approach would be a total gastrectomy, which has an exceedingly high mortality rate in this setting.

GASTRITIS

The term *gastritis* should be reserved for histologically documented inflammation of the gastric mucosa. Gastritis is not the mucosal erythema seen during endoscopy and is not interchangeable with “dyspepsia.” The etiologic factors leading to gastritis are broad and heterogeneous. Gastritis has been classified based on time course (acute vs chronic), histologic features, and anatomic distribution or proposed pathogenic mechanism **(Table 317-11)**.

TABLE 317-11

Classification of Gastritis

I. Acute gastritis
A. Acute *H. pylori* infection
B. Other acute infectious gastritides
1. Bacterial (other than *H. pylori)*
2. *H. heilmannii*
3. Phlegmonous
4. Mycobacterial
5. Syphilitic
6. Viral
7. Parasitic
8. Fungal
II. Chronic atrophic gastritis
A. Type A: Autoimmune, body-predominant
B. Type B: *H. pylori*–related, antral-predominant
C. Indeterminate
III. Uncommon forms of gastritis
A. Lymphocytic
B. Eosinophilic
C. Crohn's disease
D. Sarcoidosis
E. Isolated granulomatous gastritis
F. Russell body gastritis

The correlation between the histologic findings of gastritis, the clinical picture of abdominal pain or dyspepsia, and endoscopic findings noted on gross inspection of the gastric mucosa is poor. Therefore, there is no typical clinical manifestation of gastritis.

Acute Gastritis

The most common causes of acute gastritis are infectious. Acute infection with *H. pylori* induces gastritis. However, *H. pylori* acute gastritis has not been extensively studied. It is reported as presenting with sudden onset of epigastric pain, nausea, and vomiting, and limited mucosal histologic studies demonstrate a marked infiltrate of neutrophils with edema and hyperemia. If not treated, this picture will evolve into one of chronic gastritis. Hypochlorhydria lasting for up to 1 year may follow acute *H. pylori* infection.

Bacterial infection of the stomach or phlegmonous gastritis is a rare, potentially life-threatening disorder characterized by marked and diffuse acute inflammatory infiltrates of the entire gastric wall, at times accompanied by necrosis. Elderly individuals, alcoholics, and AIDS patients may be affected. Potential iatrogenic causes include polypectomy and mucosal injection with India ink. Organisms associated with this entity include streptococci, staphylococci, *Escherichia coli*, *Proteus*, and *Haemophilus* species. Failure of supportive measures and antibiotics may result in gastrectomy.

Other types of infectious gastritis may occur in immunocompromised individuals such as AIDS patients. Examples include herpetic (herpes simplex) or CMV gastritis. The histologic finding of intranuclear inclusions would be observed in the latter.

Chronic Gastritis

Chronic gastritis is identified histologically by an inflammatory cell infiltrate consisting primarily of lymphocytes and plasma cells, with very scant neutrophil involvement. Distribution of the inflammation may be patchy, initially involving superficial and glandular portions of the gastric mucosa. This picture may progress to more severe glandular destruction, with atrophy and metaplasia. Chronic gastritis has been classified according to histologic characteristics. These include superficial atrophic changes and gastric atrophy. The association of atrophic gastritis with the development of gastric cancer has led to the development of endoscopic and serologic markers of severity. Some of these include gross inspection and classification of mucosal abnormalities during standard endoscopy, magnification endoscopy, endoscopy with narrow band imaging and/or autofluorescence imaging, and measurement of several serum biomarkers including pepsinogen I and II levels, gastrin-17, and anti-*H. pylori* serologies. The clinical utility of these tools is currently being explored.

The early phase of chronic gastritis is *superficial gastritis.* The inflammatory changes are limited to the lamina propria of the surface mucosa, with edema and cellular infiltrates separating intact gastric glands. The next stage is *atrophic gastritis.* The inflammatory infiltrate extends deeper into the mucosa, with progressive distortion and destruction of the glands. The final stage of chronic gastritis is *gastric atrophy.* Glandular structures are lost, and there is a paucity of inflammatory infiltrates. Endoscopically, the mucosa may be substantially thin, permitting clear visualization of the underlying blood vessels.

Gastric glands may undergo morphologic transformation in chronic gastritis. Intestinal metaplasia denotes the conversion of gastric glands to a small intestinal phenotype with small-bowel mucosal glands containing goblet cells. The metaplastic changes may vary in distribution from patchy to fairly extensive gastric involvement. Intestinal metaplasia is an important predisposing factor for gastric cancer **(Chap. 76)**.

Chronic gastritis is also classified according to the predominant site of involvement. Type A refers to the body-predominant form (autoimmune), and type B is the antral-predominant form (*H. pylori*–related). This classification is artificial in view of the difficulty in distinguishing between these two entities. The term *AB gastritis* has been used to refer to a mixed antral/body picture.

TYPE A GASTRITIS

The less common of the two forms involves primarily the fundus and body, with antral sparing. Traditionally, this form of gastritis has been associated with pernicious anemia **(Chap. 95)** in the presence of circulating antibodies against parietal cells and IF; thus, it is also called *autoimmune gastritis. H. pylori* infection can lead to a similar distribution of gastritis. The characteristics of an autoimmune picture are not always present.

Antibodies to parietal cells have been detected in >90% of patients with pernicious anemia and in up to 50% of patients with type A gastritis. The parietal cell antibody is directed against H^+,K^+-ATPase. T cells are also implicated in the injury pattern of this form of gastritis. A subset of patients infected with *H. pylori* develop antibodies against H^+,K^+-ATPase, potentially leading to the atrophic gastritis pattern seen in some patients infected with this organism. The mechanism is thought to involve molecular mimicry between *H. pylori* LPS and H^+,K^+-ATPase.

Parietal cell antibodies and atrophic gastritis are observed in family members of patients with pernicious anemia. These antibodies are observed in up to 20% of individuals aged >60 and in ~20% of patients with vitiligo and Addison's disease. About one-half of patients with pernicious anemia have antibodies to thyroid antigens, and about 30% of patients with thyroid disease have circulating antiparietal cell antibodies. Anti-IF antibodies are more specific than parietal cell antibodies for type A gastritis, being present in ~40% of patients with pernicious anemia. Another parameter consistent with this form of gastritis being autoimmune in origin is the higher incidence of specific familial histocompatibility haplotypes such as HLA-B8 and HLA-DR3.

The parietal cell–containing gastric gland is preferentially targeted in this form of gastritis, and achlorhydria results. Parietal cells are the source of IF, the lack of which will lead to vitamin B_{12} deficiency and its sequelae (megaloblastic anemia, neurologic dysfunction).

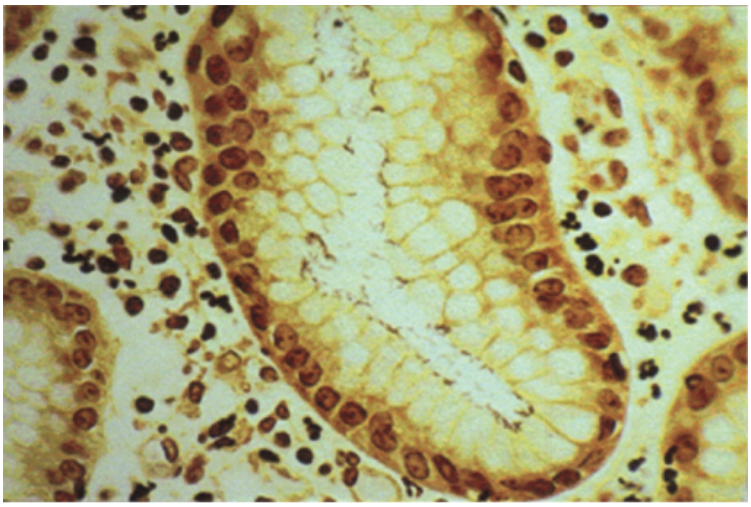
Gastric acid plays an important role in feedback inhibition of gastrin release from G cells. Achlorhydria, coupled with relative sparing of the antral mucosa (site of G cells), leads to hypergastrinemia. Gastrin levels can be markedly elevated (>500 pg/mL) in patients with pernicious anemia. ECL cell hyperplasia with frank development of gastric carcinoid tumors may result from gastrin trophic effects. Hypergastrinemia and achlorhydria may also be seen in nonpernicious anemia–associated type A gastritis.

TYPE B GASTRITIS

Type B, or antral-predominant, gastritis is the more common form of chronic gastritis. *H. pylori* infection is the cause of this entity. Although described as “antral-predominant,” this is likely a misnomer in view of studies documenting the progression of the inflammatory process toward the body and fundus of infected individuals. The conversion to a pangastritis is time-dependent and estimated to require 15–20 years. This form of gastritis increases with age, being present in up to 100% of persons aged >70. Histology improves after *H. pylori* eradication. The number of *H. pylori* organisms decreases dramatically with progression to gastric atrophy, and the degree of inflammation correlates with the level of these organisms. Early on, with antral-predominant findings, the quantity of *H. pylori* is highest and a dense chronic inflammatory infiltrate of the lamina propria is noted, accompanied by epithelial cell infiltration with polymorphonuclear leukocytes **(Fig. 317-15)**.

FIGURE 317-15

Chronic gastritis and *H. pylori* organisms. Steiner silver stain of superficial gastric mucosa showing abundant darkly stained microorganisms layered over the apical portion of the surface epithelium. Note that there is no tissue invasion.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Multifocal atrophic gastritis, gastric atrophy with subsequent metaplasia, has been observed in chronic *H. pylori*–induced gastritis. This may ultimately lead to development of gastric adenocarcinoma **(Fig. 317-8; Chap. 76)**. *H. pylori* infection is now considered an independent risk factor for gastric cancer. Worldwide epidemiologic studies have documented a higher incidence of *H. pylori* infection in patients with adenocarcinoma of the stomach as compared to control subjects. Seropositivity for *H. pylori* is associated with a three- to sixfold increased risk of gastric cancer. This risk may be as high as ninefold after adjusting for the inaccuracy of serologic testing in the elderly. The mechanism by which *H. pylori* infection leads to cancer is unknown, but it appears to be related to the chronic inflammation induced by the organism. Eradication of *H. pylori* as a general preventative measure for gastric cancer is being evaluated but is not yet recommended.

Infection with *H. pylori* is also associated with development of a low-grade B cell lymphoma, gastric MALT lymphoma **(Chap. 104)**. The chronic T cell stimulation caused by the infection leads to production of cytokines that promote the B cell tumor. The tumor should be initially staged with a CT scan of the abdomen and EUS. Tumor growth remains dependent on the presence of *H. pylori*, and its eradication is often associated with complete regression of the tumor. The tumor may take more than a year to regress after treating the infection. Such patients should be followed by EUS every 2–3 months. If the tumor is stable or decreasing in size, no other therapy is necessary. If the tumor grows, it may have become a high-grade B cell lymphoma. When the tumor becomes a high-grade aggressive lymphoma histologically, it loses responsiveness to *H. pylori* eradication.

TREATMENT

TREATMENT

Chronic Gastritis

Treatment in chronic gastritis is aimed at the sequelae and not the underlying inflammation. Patients with pernicious anemia will require parenteral vitamin B_{12} supplementation on a long-term basis. Eradication of *H. pylori* is often recommended even if PUD or a low-grade MALT lymphoma is not present. Expert opinion suggests that patients with atrophic gastritis complicated by intestinal metaplasia without dysplasia should undergo surveillance endoscopy every 3 years.

Miscellaneous Forms of Gastritis

Lymphocytic gastritis is characterized histologically by intense infiltration of the surface epithelium with lymphocytes. The infiltrative process is primarily in the body of the stomach and consists of mature T cells and plasmacytes. The etiology of this form of chronic gastritis is unknown. It has been described in patients with celiac sprue, but whether there is a common factor associating these two entities is unknown. No specific symptoms

suggest lymphocytic gastritis. A subgroup of patients have thickened folds noted on endoscopy. These folds are often capped by small nodules that contain a central depression or erosion; this form of the disease is called *varioliform gastritis. H. pylori* probably plays no significant role in lymphocytic gastritis. Therapy with glucocorticoids or sodium cromoglycate has obtained unclear results.

Marked eosinophilic infiltration involving any layer of the stomach (mucosa, muscularis propria, and serosa) is characteristic of *eosinophilic gastritis.* Affected individuals will often have circulating eosinophilia with clinical manifestation of systemic allergy. Involvement may range from isolated gastric disease to diffuse eosinophilic gastroenteritis. Antral involvement predominates, with prominent edematous folds being observed on endoscopy. These prominent antral folds can lead to outlet obstruction. Patients can present with epigastric discomfort, nausea, and vomiting. Treatment with glucocorticoids has been successful.

Several systemic disorders may be associated with *granulomatous gastritis.* Gastric involvement has been observed in Crohn's disease. Involvement may range from granulomatous infiltrates noted only on gastric biopsies to frank ulceration and stricture formation. Gastric Crohn's disease usually occurs in the presence of small-intestinal disease. Several rare infectious processes can lead to granulomatous gastritis, including histoplasmosis, candidiasis, syphilis, and tuberculosis. Other unusual causes of this form of gastritis include sarcoidosis, idiopathic granulomatous gastritis, and eosinophilic granulomas involving the stomach. Establishing the specific etiologic agent in this form of gastritis can be difficult, at times requiring repeat endoscopy with biopsy and cytology. Occasionally, a surgically obtained full-thickness biopsy of the stomach may be required to exclude malignancy.

Russell body gastritis (RBG) is a mucosal lesion of unknown etiology that has a pseudotumoral endoscopic appearance. Histologically, it is defined by the presence of numerous plasma cells containing Russell bodies (RBs) that express kappa and lambda light chains. Only 10 cases have been reported, and 7 of these have been associated with *H. pylori* infection. The lesion can be confused with a neoplastic process, but it is benign in nature, and the natural history of the lesion is not known. There have been cases of resolution of the lesion when *H. pylori* was eradicated.

MÉNÉTRIER'S DISEASE

Ménétrier's disease (MD) is a very rare gastropathy characterized by large, tortuous mucosal folds. MD has an average age of onset of 40–60 years with a male predominance. The differential diagnosis of large gastric folds includes ZES, malignancy (lymphoma, infiltrating carcinoma), infectious etiologies (CMV, histoplasmosis, syphilis, tuberculosis), gastritis polyposa profunda, and infiltrative disorders such as sarcoidosis. MD is most commonly confused with large or multiple gastric polyps (prolonged PPI use) or familial polyposis syndromes. The mucosal folds in MD are often most prominent in the body and fundus, sparing the antrum. Histologically, massive foveolar hyperplasia (hyperplasia of surface and glandular mucous cells) and a marked reduction in oxyntic glands and parietal cells and chief cells are noted. This hyperplasia produces the prominent folds observed. The pits of the gastric glands elongate and may become extremely dilated and tortuous. Although the lamina propria may contain a mild chronic inflammatory infiltrate including eosinophils and plasma cells, MD is not considered a form of gastritis. The etiology of this unusual clinical picture in children is often CMV, but the etiology in adults is unknown. Overexpression of the growth factor TGF-α has been demonstrated in patients with MD. The overexpression of TGF-α in turn results in overstimulation of the epidermal growth factor receptor (EGFR) pathway and increased proliferation of mucus cells, resulting in the observed foveolar hyperplasia.

The clinical presentation in adults is usually insidious and progressive. Epigastric pain, nausea, vomiting, anorexia, peripheral edema, and weight loss are signs and symptoms in patients with MD. Occult GI bleeding may occur, but overt bleeding is unusual and, when present, is due to superficial mucosal erosions. In fact, bleeding is more often seen in one of the common mimics of MD, gastric polyposis. Twenty to 100% of patients (depending on time of presentation) develop a protein-losing gastropathy due to hypersecretion of gastric mucus accompanied by hypoalbuminemia and edema. Gastric acid secretion is usually reduced or absent because of the decreased parietal cells. Large gastric folds are readily detectable by either radiographic (barium meal) or endoscopic methods. Endoscopy with deep mucosal biopsy, preferably full thickness with a snare technique, is required to establish the diagnosis and exclude other entities that may present similarly. A nondiagnostic biopsy may lead to a surgically obtained full-thickness biopsy to exclude malignancy. Although MD is considered premalignant by some, the risk of neoplastic progression is not defined. Complete blood count, serum gastrin, serum albumin, CMV and *H. pylori* serology, and pH testing of gastric aspirate during endoscopy should be included as part of the initial evaluation of patients with large gastric folds.

TREATMENT

TREATMENT

Ménétrier's Disease

Medical therapy with anticholinergic agents, prostaglandins, PPIs, prednisone, somatostatin analogues (octreotide) and H_2 receptor antagonists yields varying results. Ulcers should be treated with a standard approach. The discovery that MD is associated with overstimulation of the EGFR pathway has led to the successful use of the EGF inhibitory antibody, cetuximab, in these patients. Specifically, four of seven patients who completed a 1-month

trial with this agent demonstrated near complete histologic remission and improvement in symptoms. Cetuximab is now considered the first-line treatment for MD, leaving total gastrectomy for severe disease with persistent and substantial protein loss despite therapy with this agent.

FURTHER READING

Basuroy R et al: Neuroendocrine tumors. Gastroenterol Clin N Am 45:487, 2016.

Chan JA, Kulke MH: Medical management of pancreatic neuroendocrine tumors: Current and future therapy. Surg Oncol Clin N Am 25:423, 2016. [PubMed: 27013373]

Chey WD et al: ACG Clinical Guideline: Treatment of *Helicobacter pylori* Infection. Am J Gastroenterol 112:212, 2017. [PubMed: 28071659]

Clancy TE: Surgical management of pancreatic neuroendocrine tumors. Hematol Oncol Clin N Am 30:103, 2016.

Davis JL, Ripley RT: Postgastrectomy syndromes and nutritional considerations following gastric surgery. Surg Clin N Am 97:277, 2017. [PubMed: 28325187]

Dreischulte T: Safer prescribing—A trial of education, informatics, and financial incentives. N Engl J Med 374:1053, 2016. [PubMed: 26981935]

Epelboym I, Mazeh H: Zollinger-Ellison Syndrome: Classical considerations and current controversies. Oncologist 19:44, 2014. [PubMed: 24319020]

Graham DY: *Helicobacter pylori* update: Gastric cancer, reliable therapy, and possible benefits. Gastroenterology 148:719, 2015. [PubMed: 25655557]

Hunt RH et al: The stomach in health and disease. Gut 64:1650, 2015. [PubMed: 26342014]

Kim HU: Diagnostic and treatment approaches for refractory peptic ulcers. Clin Endosc 48:285, 2015. [PubMed: 26240800]

Laine L: Upper gastrointestinal bleeding due to a peptic ulcer. N Engl J Med 374:2367, 2016. [PubMed: 27305194]

McCarberg BH, Cryer B: Evolving therapeutic strategies to improve nonsteroidal anti-inflammatory drug safety. Am J Ther 22:167, 2015. [PubMed: 25943389]

Melcarne L et al: Management of NSAID-associated peptic ulcer disease. Expert Rev Gastroenterol Hepatol 6:723, 2016.

Patel KA, Howden CW: Update on the diagnosis and management of *Helicobacter pylori* infection in adults. J Clin Gastroenterol 49:461, 2015. [PubMed: 25811119]

Pelletier JP et al: Efficacy and safety of oral and analgesics in the management of osteoarthritis: Evidence from real-life setting trials and surveys. Semin Arthritis Rheum 45:S22, 2016. [PubMed: 26806184]

Scarpignato C et al: Effective and safe proton pump inhibitor therapy in acid-related disease—A position paper addressing benefits and potential harms of acid suppression. BMC Med 14:179, 2016. [PubMed: 27825371]

Schoenfeld AJ, Grady D: Adverse effects associate with proton pump inhibitors. JAMA Int Med 176:172, 2016.

Schubert ML: Gastric acid secretion. Curr Opin Gastroenterol 32:452, 2016. [PubMed: 27607343]

Singh S et al: Diagnosis and management of gastrointestinal neuroendocrine tumors: An evidence-based Canadian consensus. Cancer Treat Rev 47:32, 2016. [PubMed: 27236421]

Sugano K et al: Kyoto global consensus report on *Helicobacter pylori* gastritis. Gut 64:1353, 2015. [PubMed: 26187502]

Talley NJ: Functional dyspepsia: New insights into pathogenesis and therapy. Korean J Intern Med 31:444, 2016. [PubMed: 27048251]

Vinik AI, Chaya C: Clinical presentation and diagnosis of neuroendocrine tumors. Hematol Oncol Clin N Am 30:21, 2016.

Chapter 318: Disorders of Absorption

Henry J. Binder

INTRODUCTION

Disorders of absorption constitute a broad spectrum of conditions with multiple etiologies and varied clinical manifestations. Almost all of these clinical problems are associated with *diminished* intestinal absorption of one or more dietary nutrients and are often referred to as the *malabsorption syndrome.* This term is not ideal as it represents a pathophysiologic state, does *not* provide an etiologic explanation for the underlying problem, and should not be considered an adequate final diagnosis. The only clinical conditions in which absorption is *increased* are hemochromatosis and Wilson's disease, in which absorption of iron and copper, respectively, is elevated.

Most malabsorption syndromes are associated with *steatorrhea*, an increase in stool fat excretion to >7% of dietary fat intake. Some malabsorption disorders are not associated with steatorrhea: primary lactase deficiency, a congenital absence of the small-intestinal brush border disaccharidase enzyme lactase, is associated with lactose "malabsorption," and pernicious anemia is associated with a marked decrease in intestinal absorption of cobalamin (vitamin B_{12}) due to an absence of gastric parietal-cell intrinsic factor, which is required for cobalamin absorption.

Disorders of absorption must be included in the differential diagnosis of diarrhea **(Chap. 42)**. First, diarrhea is frequently associated with and/or is a consequence of the diminished absorption of one or more dietary nutrients. The diarrhea may be secondary either to the intestinal process that is responsible for the steatorrhea or to steatorrhea per se. Thus, celiac disease (see below) is associated with both extensive morphologic changes in the small-intestinal mucosa and reduced absorption of several dietary nutrients; in contrast, the diarrhea of steatorrhea is the result of the effect of nonabsorbed dietary fatty acids on intestinal (usually colonic) ion transport. For example, oleic and ricinoleic acids (a bacterially hydroxylated fatty acid that is also the active ingredient in castor oil, a widely used laxative) induce active colonic Cl ion secretion, most likely secondary to increasing intracellular Ca. In addition, diarrhea per se may result in mild steatorrhea (<11 g of fat excretion while on a 100-g fat diet). Second, most patients will indicate that they have diarrhea, not that they have fat malabsorption. Third, many intestinal disorders that have diarrhea as a prominent symptom (e.g., ulcerative colitis, traveler's diarrhea secondary to an enterotoxin produced by *Escherichia coli*) do not necessarily have diminished absorption of any dietary nutrient.

Diarrhea as a *symptom* (i.e., when the term is used by patients to describe their bowel movement pattern) may reflect a decrease in stool consistency, an increase in stool volume, an increase in number of bowel movements, or any combination of these three changes. In contrast, diarrhea as a *sign* is a quantitative increase in stool water or weight of >200–225 mL or g per 24 h when a Western-type diet is consumed. Individuals consuming a diet with higher-fiber content may normally have a stool weight of up to 400 g/24 h. Thus, the clinician must clarify what an individual patient means by diarrhea. Some 10% of patients referred to gastroenterologists for further evaluation of unexplained diarrhea do not have an increase in stool water when this variable is determined quantitatively. Such patients may have small, frequent, somewhat loose bowel movements with stool urgency that is indicative of proctitis, but do not have an increase in stool weight or volume. In addition, an occasional patient will describe their fecal incontinence as diarrhea due to social embarrassment.

It is also critical to establish whether a patient's diarrhea is secondary to diminished absorption of one or more dietary nutrients rather than being due to small- and/or large-intestinal fluid and electrolyte secretion. The former has often been termed *osmotic diarrhea*, while the latter has been referred to as *secretory diarrhea.* Unfortunately, both secretory and osmotic elements can be present simultaneously in the same disorder; thus, this distinction is not always precise. Nonetheless, two studies—determination of stool electrolytes and observation of the effect of a fast on stool output—can help make this distinction.

The demonstration of the effect of prolonged (>24 h) fasting on stool output can suggest that a *dietary nutrient* is responsible for the individual's diarrhea. Secretory diarrhea associated with enterotoxin-induced traveler's diarrhea would not be affected by prolonged fasting, as enterotoxin-induced stimulation of intestinal fluid and electrolyte secretion is not altered by eating. In contrast, diarrhea secondary to lactose malabsorption in primary lactase deficiency would undoubtedly cease during a prolonged fast. Thus, a substantial decrease in stool output by a fasting patient during quantitative stool collection lasting at least 24 h is presumptive evidence that the diarrhea is related to malabsorption of one or more dietary nutrients. The persistence of stool output during fasting indicates that the diarrhea is likely secretory and that its cause is *not* a dietary nutrient. Either a luminal (e.g., *E. coli* enterotoxin) or a circulating (e.g., vasoactive intestinal peptide) secretagogue could be responsible for unaltered persistence of a patient's diarrhea during a prolonged fast. The observed effects of fasting can be compared and correlated with stool electrolyte and osmolality determinations.

Measurement of stool electrolytes and osmolality requires comparison of Na^+ and K^+ concentrations in liquid stool with the osmolality of the stool in order to determine the presence or absence of a so-called stool osmotic gap. The following formula is used:

$2 \times (\text{stool } [Na^+] + \text{stool } [K^+]) \leq \text{stool osmolality}$

The cation concentrations are doubled to estimate stool anion concentrations. The presence of a significant osmotic gap suggests the presence in stool water of a substance (or substances) other than Na/K/anions, which is presumably responsible for the patient's diarrhea. Originally, stool osmolality was measured, but it is almost invariably greater than the required 290–300 mosmol/kg H_2O, reflecting bacterial degradation of

nonabsorbed carbohydrate either immediately before defecation or in the stool jar while specimen awaits chemical analysis, even when the stool is refrigerated. As a result, the stool osmolality should be assumed to be 300 mosmol/kg H_2O. A low stool osmolality (<290 mosmol/kg H_2O) reflects the addition of either dilute urine or water, indicating either collection of urine and stool together or so-called factitious diarrhea, a form of Münchausen's syndrome. When the calculated difference in the formula above is >50, an osmotic gap exists; its presence suggests that the diarrhea is due to a nonabsorbed dietary nutrient—for example a fatty acid and/or a carbohydrate. When this difference is <25, it is presumed that a dietary nutrient is not responsible for the diarrhea. Since elements of both osmotic diarrhea (i.e., due to malabsorption of a dietary nutrient) and secretory diarrhea may be present, this distinction at times is less clear-cut at the bedside than when used as a teaching example. Ideally, the presence of an osmotic gap will be associated with a marked decrease in stool output during a prolonged fast, while an osmotic gap will likely be absent in an individual whose stool output is not reduced substantially during a period of fasting.

NUTRIENT DIGESTION AND ABSORPTION

The lengths of the small intestine and the colon are ~300 and ~80 cm, respectively. However, the effective functional surface area is ~600-fold greater than that of a hollow tube as a result of folds, villi (in the small intestine), and microvilli. The functional surface area of the small intestine is somewhat greater than that of a doubles tennis court. In addition to nutrient digestion and absorption, the intestinal epithelia have several other functions:

1. *Barrier and immune defense.* The intestine is exposed to a large number of potential antigens and enteric and invasive microorganisms, and it is extremely effective at preventing the entry of almost all of these agents. The intestinal mucosa also synthesizes and secretes secretory IgA.

2. *Fluid and electrolyte absorption and secretion.* The intestine absorbs ~7–8 L of fluid daily, a volume comprising dietary fluid intake (1–2 L/d) and salivary, gastric, pancreatic, biliary, and intestinal fluid (6–7 L/d). Several stimuli, especially bacteria and bacterial enterotoxins, induce fluid and electrolyte secretion that may lead to diarrhea **(Chap. 128)**.

3. *Synthesis and secretion of several proteins.* The intestinal mucosa is a major site for the production of proteins, including apolipoproteins.

4. *Production of several bioactive amines and peptides.* The intestine is one of the largest endocrine organs in the body and produces several amines (e.g., 5-hydroxytryptophan) and peptides that serve as paracrine and hormonal mediators of intestinal function.

The small and large intestines are distinct anatomically (villi are present in the small intestine but are absent in the colon) and functionally (nutrient digestion and absorption take place in the small intestine but not in the colon). No precise anatomic characteristics separate duodenum, jejunum, and ileum, although certain nutrients are absorbed exclusively in specific areas of the small intestine. However, villous cells in the small intestine (surface epithelial cells in the colon) and crypt cells have distinct anatomic and functional

characteristics. Intestinal epithelial cells are continuously renewed; new proliferating epithelial cells at the base of the crypt migrate over 48–72 h to the tip of the villus (or surface of the colon), where they exist as well-developed epithelial cells with digestive and absorptive function. This high rate of cell turnover explains the relatively rapid resolution of diarrhea and other digestive-tract side effects during chemotherapy as new cells not exposed to these toxic agents are produced. Equally important is the paradigm of separation of villous/surface cell and crypt cell functions. Digestive hydrolytic enzymes are present primarily in the brush border of villous epithelial cells. Absorptive and secretory functions are also separate: villous/surface cells are primarily, but not exclusively, the site for absorptive function, while secretory function is located in crypts of both the small and large intestines.

Nutrients, minerals, and vitamins are absorbed by one or more active-transport mechanisms. These mechanisms are energy dependent and are mediated by membrane transport proteins. These processes will result in the *net* movement of a substance against or in the absence of an electrochemical concentration gradient. Intestinal absorption of amino acids and monosaccharides (e.g., glucose) is also a specialized form of active transport—*secondary active transport.* The movement of actively transported nutrients against a concentration gradient is Na^+ dependent and is due to a Na^+ gradient across the apical membrane. The Na^+ gradient is maintained by Na^+,K^+-adenosine triphosphatase (ATPase), the so-called Na^+ pump located on the basolateral membrane, which extrudes Na^+ and maintains low intracellular [Na] as well as the Na^+ gradient across the apical membrane. As a result, active glucose absorption and glucose-stimulated Na^+ absorption require both the apical membrane transport protein SGLT1 and the basolateral Na^+,K^+-ATPase. In addition to requiring Na^+ for its absorption, glucose stimulates Na^+ and fluid absorption; this effect is the physiologic basis of oral rehydration therapy for the treatment of diarrhea (Chap. 42). **The mechanisms of intestinal fluid and electrolyte absorption and secretion are discussed in Chap. 42.**

Although the intestinal epithelial cells are crucial mediators of absorption and of ion and water flow, the several cell types in the lamina propria (e.g., mast cells, macrophages, myofibroblasts) and the enteric nervous system interact with the epithelium to regulate mucosal cell function. Intestinal function results from the integrated responses and interactions of intestinal epithelial cells and intestinal muscle.

ENTEROHEPATIC CIRCULATION OF BILE ACIDS

Bile acids are not present in the diet but are synthesized in the liver by a series of enzymatic steps that also represent cholesterol catabolism. Indeed, interruption of the enterohepatic circulation of bile acids can reduce serum cholesterol levels by 10% before a new steady state is established. Bile acids are either primary or secondary. Primary bile acids are synthesized in the liver from cholesterol, and secondary bile acids are synthesized from primary bile acids in the intestine by colonic bacterial enzymes. The two primary bile acids in humans are cholic acid and chenodeoxycholic acid; the two most abundant secondary bile acids are deoxycholic acid and lithocholic acid. The liver synthesizes ~500 mg of bile acids daily; the bile acids are conjugated to either taurine or glycine (to form tauroconjugated and glycoconjugated bile acids, respectively) and are secreted into the duodenum in bile. The primary functions of bile acids are to (1)

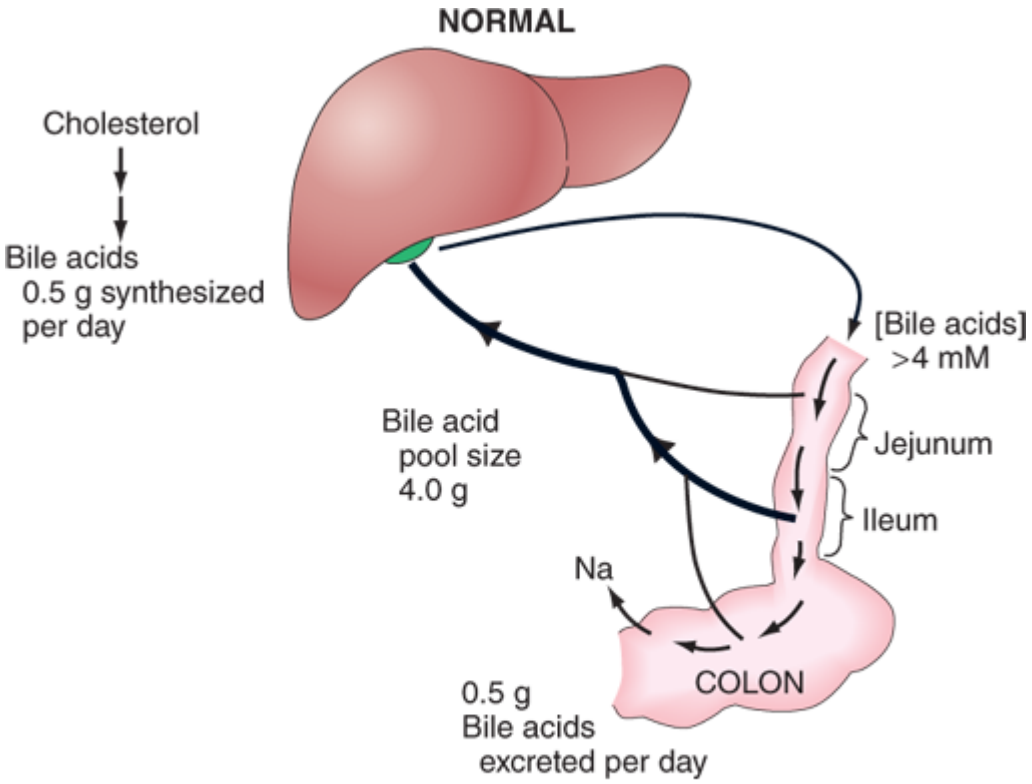
promote bile flow, (2) solubilize cholesterol and phospholipid in the gallbladder by mixed micelle formation, and (3) enhance dietary lipid digestion and absorption by forming mixed micelles in the proximal small intestine.

Bile acids are primarily absorbed by an active, Na^+-dependent process that takes place exclusively in the ileum; to a lesser extent, they are absorbed by non-carrier-mediated transport processes in the jejunum, ileum, and colon. Conjugated bile acids that enter the colon are deconjugated by colonic bacterial enzymes. The unconjugated bile acids are rapidly absorbed by nonionic diffusion. Colonic bacterial enzymes also dehydroxylate bile acids to secondary bile acids.

Bile acids absorbed from the intestine return to the liver via the portal vein and are then re-secreted **(Fig. 318-1)**. Bile-acid synthesis is largely autoregulated by 7α-hydroxylase, the initial enzyme in cholesterol degradation. A decrease in the quantity of bile acids returning to the liver from the intestine is associated with an increase in bile-acid synthesis/cholesterol catabolism (mediated by fibroblast growth factor [FGF19]), which helps keep the bile-acid pool size relatively constant. However, the capacity to increase bile-acid synthesis is limited to ~2- to 2.5-fold (see below). The bile-acid pool size is ~4 g. The pool is circulated via the enterohepatic circulation about twice during each meal, or six to eight times during a 24-h period. A relatively small quantity of bile acids is not absorbed and is excreted in stool daily; this fecal loss is matched by hepatic bile-acid synthesis.

FIGURE 318-1

Schematic representation of the enterohepatic circulation of bile acids. Bile-acid synthesis is cholesterol catabolism and occurs in the liver. Bile acids are secreted in bile and are stored in the gallbladder between meals and at night. Food in the duodenum induces the release of cholecystokinin, a potent stimulus for gallbladder contraction resulting in bile-acid entry into the duodenum. Bile acids are primarily absorbed via an Na-dependent transport process that is located only in the ileum. A relatively small quantity of bile acids (~500 mg) is not absorbed in a 24-h period and is lost in stool. Fecal bile-acid losses are matched by bile-acid synthesis. The bile-acid pool (the total amount of bile acids in the body) is ~4 g and is circulated twice during each meal or six to eight times in a 24-h period.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Defects in any of the steps in enterohepatic circulation of bile acids can result in a decrease in the duodenal concentration of conjugated bile acids and consequently in the development of steatorrhea. Thus, steatorrhea can be caused by abnormalities in bile-acid synthesis and excretion, their physical state in the intestinal lumen, and reabsorption (Table 318-1).

TABLE 318-1

Defects in Enterohepatic Circulation of Bile Acids

Process	Pathophysiologic Defect	Disease Example
Synthesis	Decreased hepatic function	Cirrhosis
Biliary secretion	Altered canalicular function	Primary biliary cirrhosis
Maintenance of conjugated bile acids	Bacterial overgrowth	Jejunal diverticulosis
Reabsorption	Abnormal ileal function	Crohn's disease

Synthesis

Decreased bile-acid synthesis and steatorrhea have been demonstrated in chronic liver disease, but steatorrhea often is not a major component of illness in these patients.

Secretion

Although bile-acid secretion may be reduced or absent in biliary obstruction, steatorrhea is rarely a significant medical problem in these patients. In contrast, primary biliary cirrhosis represents a defect in canalicular excretion of organic anions, including bile acids, and not infrequently is associated with steatorrhea and its consequences (e.g., chronic bone disease). Thus, the osteopenia/osteomalacia and other chronic bone abnormalities often present in patients with primary biliary cirrhosis and other cholestatic syndromes are secondary to steatorrhea that then leads to calcium and vitamin D malabsorption as well as to the effects of cholestasis (e.g., bile acids and inflammatory cytokines).

Maintenance of Conjugated Bile Acids

In bacterial overgrowth syndromes associated with diarrhea, steatorrhea, and macrocytic anemia, a colonic type of bacterial flora is increased in the small intestine. Steatorrhea is primarily a result of the decrease in conjugated bile acids secondary to their deconjugation by colonic-type bacteria. Two complementary explanations account for the resulting impairment of micelle formation: (1) Unconjugated bile acids are rapidly absorbed in the jejunum by nonionic diffusion, and the result is a reduced concentration of duodenal bile acids; and (2) the critical micellar concentration (CMC) of unconjugated bile acids is higher than that of conjugated bile acids; therefore, unconjugated bile acids are less effective than conjugated bile acids in micelle formation.

Reabsorption

Ileal dysfunction caused by either Crohn's disease or surgical resection results in a decrease in bile-acid reabsorption in the ileum and an *increase* in the delivery of bile acids to the large intestine. The resulting clinical consequences—diarrhea with or without steatorrhea—are determined by the *degree* of ileal dysfunction and the *response* of the enterohepatic circulation to bile-acid losses (Table 318-2). Patients with limited ileal disease or resection often have diarrhea but not steatorrhea. The diarrhea, a result of stimulation of active Cl secretion by bile acids in the colon, has been called *bile-acid diarrhea* or *choleretic enteropathy* and responds promptly to cholestyramine, an anion-binding resin. Steatorrhea does not develop because hepatic synthesis of bile acids increases to compensate for the rate of fecal bile-acid losses, resulting in maintenance of both the bile-acid pool size and the intraduodenal concentrations of bile acids. In contrast, patients with greater degrees of ileal disease and/or resection often have diarrhea and steatorrhea that do not respond to cholestyramine. In this situation, ileal disease is also associated with increased volumes of bile acids entering the colon; however, hepatic synthesis can no longer increase sufficiently to maintain the bile-acid pool size. As a consequence, the intraduodenal concentration of bile acids is reduced to less than the CMC, and the result is impaired micelle formation and steatorrhea. This second situation is often called *fatty acid diarrhea.* Cholestyramine may not be effective (and may even exacerbate the diarrhea by further depleting the intraduodenal bile-acid concentration); however, a low-fat diet to reduce fatty acid entry into

the colon can be effective. Two clinical features—the length of the ileal section removed and the degree of steatorrhea—can predict whether an individual patient will respond to cholestyramine. Unfortunately, these predictors are imperfect, and a therapeutic trial of cholestyramine is often necessary to establish whether an individual patient will benefit from cholestyramine. Table 318-2 contrasts the characteristics of bile-acid diarrhea (small ileal dysfunction) and fatty acid diarrhea (large ileal dysfunction).

TABLE 318-2

Comparison of Bile Acid and Fatty Acid Diarrhea

	Bile-Acid Diarrhea	Fatty Acid Diarrhea
Extent of ileal disease	Limited	Extensive
Ileal bile-acid absorption	Reduced	Reduced
Fecal bile-acid excretion	Increased	Increased
Fecal bile-acid loss compensated by hepatic synthesis	Yes	No
Bile-acid pool size	Normal	Reduced
Intraduodenal (bile acid)	Normal	Reduced
Steatorrhea	None or mild	>20 g
Response to cholestyramine	Yes	No
Response to low-fat diet	No	Yes

Bile-acid diarrhea can also occur in the absence of ileal inflammation and/or resection and is characterized by an abnormal ^{75}SeHCAT retention study and reduced ileal release of FGF19, a negative regulator of bile-acid synthesis, with a consequent increase in bile-acid synthesis and secretion that exceeds ileal bile-acid absorption. The diarrhea in these patients also responds to cholestyramine.

LIPIDS

Steatorrhea is caused by one or more defects in the digestion and absorption of dietary fat. The average intake of dietary fat in the United States is ~120–150 g/d, and fat absorption is linear to dietary fat intake. The total load of fat presented to the small intestine is considerably greater, as substantial amounts of lipid are secreted in bile each day (see "Enterohepatic Circulation of Bile Acids," above). Three types of fatty acids compose fats: long-chain fatty acids (LCFAs), medium-chain fatty acids (MCFAs), and short-chain fatty acids

(SCFAs) (Table 318-3). Dietary fat is exclusively composed of long-chain triglycerides (LCTs)—that is, glycerol that is bound via ester linkages to three LCFAs. While the majority of dietary LCFAs have carbon chain lengths of 16 or 18, all fatty acids of carbon chain length >12 are metabolized in the same manner; saturated and unsaturated fatty acids are handled identically.

TABLE 318-3

Comparison of Different Types of Fatty Acids

	Long-Chain	Medium-Chain	Short-Chain
Carbon chain length	>12	8–12	<8
Present in diet	In large amounts	In small amounts	No
Origin	In diet as triglycerides	Only in small amounts in diet as triglycerides	Bacterial degradation in colon of nonabsorbed carbohydrate to fatty acids
Primary site of absorption	Small intestine	Small intestine	Colon
Requires pancreatic lipolysis	Yes	No	No
Requires micelle formation	Yes	No	No
Present in stool	Minimal	No	Substantial

Assimilation of dietary lipid requires three integrated processes: (1) an intraluminal, or digestive, phase; (2) a mucosal, or absorptive, phase; and (3) a delivery, or postabsorptive, phase. An abnormality at any site involved in these processes can cause steatorrhea (Table 318-4). Therefore, it is essential that any patient with steatorrhea be evaluated to identify the specific physiologic defect in overall lipid digestion/absorption, as therapy will be determined by the specific etiology.

TABLE 318-4

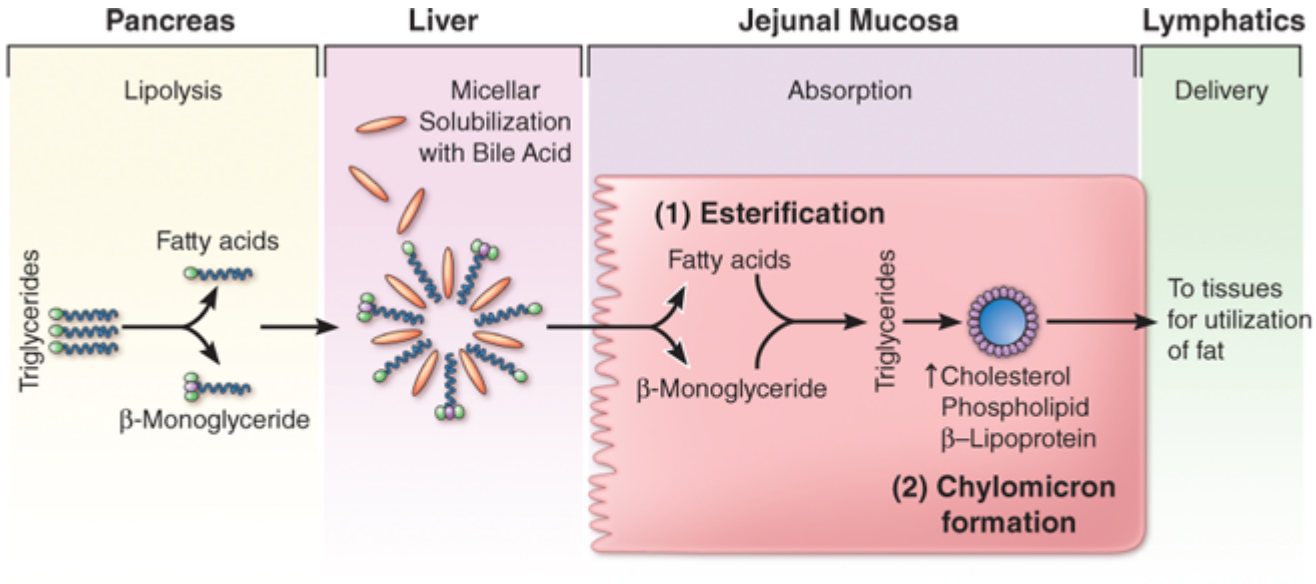
Defects in Lipid Digestion and Absorption in Steatorrhea

Phase, Process	Pathophysiologic Defect	Disease Example
Digestive		
Lipolysis formation	Decreased lipase secretion	Chronic pancreatitis
Micelle formation	Decreased intraduodenal bile acids	See Table 318-1
Absorptive		
Mucosal uptake and re-esterification	Mucosal dysfunction	Celiac disease
Postabsorptive		
Chylomicron formation	Absent β-lipoproteins	Abetalipoproteinemia
Delivery from intestine	Abnormal lymphatics	Intestinal lymphangiectasia

The digestive phase has two components: *lipolysis* and *micelle formation.* Although dietary lipid is in the form of LCTs, the intestinal mucosa does not absorb triglycerides; they must first be hydrolyzed **(Fig. 318-2)**. The initial step in lipid digestion is the formation of emulsions of finely dispersed lipid, which is accomplished by mastication and gastric contractions. Lipolysis, the hydrolysis of triglycerides to free fatty acids, monoglycerides, and glycerol by lipase, is initiated in the stomach by lingual and gastric lipases that have a pH optimum of 4.5–6.0. About 20–30% of total lipolysis occurs in the stomach. Lipolysis is completed in the duodenum and jejunum by pancreatic lipase, which is inactivated by a pH <7.0. Pancreatic lipolysis is greatly enhanced by the presence of a second pancreatic enzyme, colipase, which facilitates the movement of lipase to the triglyceride.

FIGURE 318-2

Schematic representation of lipid digestion and absorption. Dietary lipid is in the form of long-chain triglycerides. The overall process can be divided into (1) a digestive phase that includes both lipolysis and micelle formation requiring pancreatic lipase and conjugated bile acids, respectively, in the duodenum; (2) an absorptive phase for mucosal uptake and re-esterification; and (3) a postabsorptive phase that includes chylomicron formation and exit from the intestinal epithelial cell via lymphatics. *(Courtesy of John M. Dietschy, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Impaired lipolysis can lead to steatorrhea and can occur in the presence of pancreatic insufficiency due to chronic pancreatitis in adults or cystic fibrosis in children and adolescents. Normal lipolysis can be maintained by ~5% of maximal pancreatic lipase secretion; thus, steatorrhea is a late manifestation of these disorders. A reduction in intraduodenal pH can also result in altered lipolysis, as pancreatic lipase is inactivated at pH <7. Thus, ~15% of patients who have gastrinoma (Chap. 317), with substantial increases in gastric acid secretion from ectopic production of gastrin (usually from an islet cell adenoma), have diarrhea, and some have steatorrhea believed to be secondary to acid inactivation of pancreatic lipase. Similarly, patients who have chronic pancreatitis (with reduced lipase secretion) often have a decrease in pancreatic bicarbonate secretion, which will also result in a lowering of intraduodenal pH and inactivation of endogenous pancreatic lipase or of therapeutically administered lipase.

Overlying the microvillus membrane of the small intestine is the so-called unstirred water layer, a relatively stagnant aqueous phase that must be traversed by the products of lipolysis that are primarily water insoluble. Water-soluble mixed micelles provide a mechanism by which the water-insoluble products of lipolysis can reach the luminal plasma membrane of villous epithelial cells—the site for lipid absorption. Mixed micelles are molecular aggregates composed of fatty acids, monoglycerides, phospholipids, cholesterol, and conjugated bile acids. These mixed micelles are formed when the concentration of conjugated bile acids is greater than its CMC, which differs among the several bile acids present in the small-intestinal lumen. Conjugated bile acids, synthesized in the liver and excreted into the duodenum in bile, are regulated by the enterohepatic circulation (see above). Steatorrhea can result from impaired movement of fatty acids across the unstirred aqueous fluid layer in two situations: (1) an increase in the relative thickness of the unstirred water layer that occurs in bacterial overgrowth syndromes (see below) secondary to functional stasis (e.g., scleroderma); and (2) a decrease in the *duodenal* concentration of conjugated bile acids below the CMC, resulting in impaired micelle formation. Thus, steatorrhea can be caused by one or more defects in the enterohepatic circulation of bile acids.

Uptake and re-esterification constitute the *absorptive phase* of lipid digestion/absorption. Although passive diffusion has been thought to be responsible, a carrier-mediated process may mediate fatty acid and monoglyceride uptake. Regardless of the uptake process, fatty acids and monoglycerides are re-esterified by a series of enzymatic steps in the endoplasmic reticulum to form triglycerides, in which lipid exits from the intestinal epithelial cell. Impaired lipid absorption as a result of mucosal inflammation (e.g., celiac disease) and/or intestinal resection can also lead to steatorrhea.

The re-esterified triglycerides require the formation of *chylomicrons* to permit their exit from the small-intestinal epithelial cell and their delivery to the liver via the lymphatics. Chylomicrons are composed of β-lipoprotein and contain triglycerides, cholesterol, cholesterol esters, and phospholipids and enter the lymphatics, not the portal vein. Defects in the *postabsorptive phase* of lipid digestion/absorption can also result in steatorrhea, but these disorders are uncommon. Abetalipoproteinemia, or acanthocytosis, is a rare disorder of impaired synthesis of β-lipoprotein associated with abnormal erythrocytes (acanthocytes), neurologic problems, and steatorrhea **(Chap. 400)**. Lipolysis, micelle formation, and lipid uptake are all normal in patients with abetalipoproteinemia, but the re-esterified triglyceride cannot exit the epithelial cell because of the failure to produce chylomicrons. Small-intestinal biopsy samples obtained from these rare patients in the postprandial state reveal lipid-laden small-intestinal epithelial cells that become perfectly normal in appearance after a 72- to 96-h fast. Similarly, abnormalities of intestinal lymphatics (e.g., intestinal lymphangiectasia) may also be associated with steatorrhea as well as protein loss (see below). Steatorrhea can result from defects at any of the several steps in lipid digestion/absorption.

The mechanism of lipid digestion/absorption outlined above is limited to *dietary* lipid, which is almost exclusively in the form of LCTs (Table 318-3). Medium-chain triglycerides (MCTs), composed of fatty acids with carbon chain lengths of 8–12, are present in large amounts in coconut oil and are used as a nutritional supplement. MCTs can be digested and absorbed by a pathway different from that involved in LCT digestion and absorption; at one time, MCTs held promise as an important treatment for steatorrhea of almost all etiologies. Unfortunately, they have been less therapeutically effective than expected because, for reasons that are not completely understood, their use often is not associated with an increase in body weight.

In contrast to LCTs, MCTs do not require pancreatic lipolysis as they can be absorbed intact by the intestinal epithelial cell. Further, micelle formation is not necessary for the absorption of MCTs (or MCFAs, if hydrolyzed by pancreatic lipase). MCTs are absorbed more efficiently than LCTs for the following reasons: (1) The rate of absorption is greater for MCTs than for LCFAs; (2) after absorption, MCFAs are not re-esterified; (3) after absorption, MCTs are hydrolyzed to MCFAs; (4) MCTs do not require chylomicron formation to exit intestinal epithelial cells; and (5) the route of MCT exit is via the portal vein and not via lymphatics. Thus, the absorption of MCTs is greater than that of LCTs in pancreatic insufficiency, conditions with reduced intraduodenal bile-acid concentrations, small-intestinal mucosal disease, abetalipoproteinemia, and intestinal lymphangiectasia.

SCFAs are not dietary lipids but are synthesized by colonic bacterial enzymes from nonabsorbed carbohydrate and are the anions present at the highest concentration in stool (80–130 m*M*). The SCFAs in stool are primarily acetate, propionate, and butyrate, whose carbon chain lengths are 2, 3, and 4,

respectively. Butyrate is the primary nutrient for colonic epithelial cells, and its deficiency can be associated with one or more colitides. SCFAs conserve calories and carbohydrate: carbohydrates that are not completely absorbed in the small intestine will not be absorbed in the large intestine because of the absence of both disaccharidases and SGLT1, the transport protein that mediates monosaccharide absorption. In contrast, SCFAs are rapidly absorbed and stimulate colonic NaCl and fluid absorption. Most antibiotic-associated diarrhea not caused by *Clostridium difficile* is due to antibiotic suppression of the colonic microbiota, with a resulting decrease in SCFA production. As *C. difficile* accounts for only ~15–20% of all antibiotic-associated diarrhea, a relative decrease in colonic production of SCFA is likely the cause of most antibiotic-associated diarrhea.

The clinical manifestations of steatorrhea are a consequence both of the underlying disorder responsible for its development and of steatorrhea per se. Depending on the degree of steatorrhea and the level of dietary intake, significant fat malabsorption may lead to weight loss. Steatorrhea per se can be responsible for diarrhea; if the primary cause of the steatorrhea has not been identified, a low-fat diet can often ameliorate the diarrhea by decreasing fecal fat excretion. Steatorrhea is commonly associated with fat-soluble vitamin deficiency, which requires replacement with water-soluble preparations of these vitamins.

Disorders of absorption may also be associated with malabsorption of other dietary nutrients—most often carbohydrates—with or without a decrease in dietary lipid digestion and absorption. Therefore, knowledge of the mechanisms of digestion and absorption of carbohydrates, proteins, and other minerals and vitamins is useful in the evaluation of patients with altered intestinal nutrient absorption.

CARBOHYDRATES

Carbohydrates in the diet are present in the form of starch, disaccharides (sucrose and lactose), and glucose. Carbohydrates are absorbed only in the small intestine and only in the form of monosaccharides. Therefore, before their absorption, starch and disaccharides must first be digested by pancreatic amylase and intestinal brush border disaccharidases to monosaccharides. Monosaccharide absorption occurs by a Na-dependent process mediated by the brush border transport protein SGLT1.

Lactose malabsorption is the only clinically important disorder of carbohydrate absorption. Lactose, the disaccharide present in milk, requires digestion by brush border lactase to its two constituent monosaccharides, glucose and galactose. Lactase is present in almost all species in the postnatal period but then disappears throughout the animal kingdom, except in humans. Lactase activity persists in many individuals throughout life. Two different types of lactase deficiency exist—primary and secondary. In *primary lactase deficiency*, a genetically determined decrease or absence of lactase is noted, while all other aspects of both intestinal absorption and brush border enzymes are normal. In a number of nonwhite groups, primary lactase deficiency is common in adulthood. In fact, Northern European and North American whites are the only groups to maintain small-intestinal lactase activity throughout adult life. **Table 318-5** presents the incidence of primary lactase deficiency in several ethnic groups. Lactase persistence in adults is an abnormality due to a defect in the regulation of its maturation. In contrast, *secondary lactase deficiency* occurs in association with small-intestinal mucosal disease, with abnormalities in both structure and

function of other brush border enzymes and transport processes. Secondary lactase deficiency is often seen in celiac disease.

TABLE 318-5

Primary Lactase Deficiency in Adult Ethnic Groups

Ethnic Group	Prevalence of Lactase Deficiency, %
Northern European	5–15
Mediterranean	60–85
African black	85–100
African American	45–80
American white	10–25
Native American	50–95
Mexican American	40–75
Asian	90–100

Source: From FJ Simoons: Am J Dig Dis 23:963, 1978.

As lactose digestion is rate-limiting compared to glucose/galactose absorption, lactase deficiency is associated with significant lactose malabsorption. Some individuals with lactose malabsorption develop symptoms such as diarrhea, abdominal pain, cramps, and/or flatus. Most individuals with primary lactase deficiency do not have symptoms. Since lactose intolerance may be associated with symptoms suggestive of irritable bowel syndrome, persistence of such symptoms in an individual who exhibits lactose intolerance while on a strict lactose-free diet suggests that the person's symptoms were related to irritable bowel syndrome.

The development of symptoms of lactose intolerance is related to several factors:

1. *Amount of lactose in the diet*.
2. *Rate of gastric emptying.* Symptoms are more likely when gastric emptying is rapid than when it is slower. Therefore, skim milk is more likely to be associated with symptoms of lactose intolerance than whole milk, as the rate of gastric emptying after skim milk intake is more rapid. Similarly, diarrhea following

subtotal gastrectomy is often a result of lactose intolerance, as gastric emptying is accelerated in patients with a gastrojejunostomy.

3. *Small-intestinal transit time.* Although the small and large intestines both contribute to the development of symptoms, many symptoms of lactase deficiency are related to the interaction of colonic bacteria and nonabsorbed lactose. More rapid small-intestinal transit makes symptoms more likely.

4. *Colonic compensation by production of SCFAs from nonabsorbed lactose.* Reduced levels of colonic microflora, which can follow antibiotic use, are associated with increased symptoms after lactose ingestion, especially in a lactase-deficient individual.

Glucose-galactose or monosaccharide malabsorption may also be associated with diarrhea and is due to a congenital absence of SGLT1. Diarrhea develops when individuals with this disorder ingest carbohydrates that contain actively transported monosaccharides (e.g., glucose, galactose) but not when they ingest monosaccharides that are not actively transported (e.g., fructose). Fructose is absorbed by the brush border transport protein GLUT 5, a facilitated diffusion process that is not Na-dependent and is distinct from SGLT1. In contrast, some individuals develop diarrhea as a result of the consumption of large quantities of sorbitol, a sugar used in diabetic candy; sorbitol is only minimally absorbed because of the absence of an intestinal absorptive transport mechanism for this sugar.

PROTEINS

Protein is present in food almost exclusively as polypeptides and requires extensive hydrolysis to di- and tripeptides and amino acids before absorption. Proteolysis occurs in both the stomach and the small intestine; it is mediated by pepsin, which is secreted as pepsinogen by gastric chief cells, and by trypsinogen and other peptidases from pancreatic acinar cells. The proenzymes pepsinogen and trypsinogen must be activated to pepsin (by pepsin at a pH <5) and trypsin (by the intestinal brush border enzyme enterokinase and subsequently by trypsin), respectively. Proteins are absorbed by separate transport systems for di- and tripeptides and for different types of amino acids—for example neutral and dibasic. Alterations in either protein or amino acid digestion and absorption are rarely observed clinically, even in the presence of extensive small-intestinal mucosal inflammation. However, three rare genetic disorders involve protein digestion/absorption: (1) *Enterokinase deficiency* is due to an absence of the brush border enzyme that converts the proenzyme trypsinogen to trypsin and is associated with diarrhea, growth retardation, and hypoproteinemia; (2) *Hartnup's syndrome*, a defect in neutral amino acid transport, is characterized by a pellagra-like rash and neuropsychiatric symptoms; and (3) *cystinuria*, a defect in dibasic amino acid transport, is associated with renal calculi and chronic pancreatitis.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Malabsorption

The clues provided by the history, symptoms, and initial preliminary observations will serve to limit extensive, ill-focused, and expensive laboratory and imaging studies. For example, a clinician evaluating a patient who has symptoms suggestive of malabsorption and has recently undergone extensive small-intestinal resection for mesenteric ischemia should direct the initial assessment almost exclusively to defining whether a short-bowel syndrome might explain the entire clinical picture. Similarly, the development of a pattern of bowel movements suggestive of steatorrhea in a patient with long-standing alcohol abuse and chronic pancreatitis should prompt an assessment of pancreatic exocrine function.

The classic picture of malabsorption is rarely seen today in most parts of the United States. As a consequence, diseases with malabsorption must be suspected in individuals who have less severe symptoms and signs and subtle evidence of the altered absorption of only a *single* nutrient rather than obvious evidence of the malabsorption of multiple nutrients.

Although diarrhea can be caused by changes in fluid and electrolyte movement in either the small or the large intestine, dietary nutrients are absorbed almost exclusively in the small intestine. Therefore, the demonstration of diminished absorption of a dietary nutrient provides unequivocal evidence for small-intestinal disease, although colonic dysfunction may also be present (e.g., Crohn's disease may involve both the small and large intestines). Dietary nutrient absorption may be segmental or diffuse along the small intestine and is site specific. Thus, for example, calcium, iron, and folic acid are exclusively absorbed by active-transport processes in the proximal small intestine, especially the duodenum; in contrast, the active-transport mechanisms for both cobalamin and bile acids are operative only in the ileum. Therefore, in an individual who years previously has had an intestinal resection, the details of which are not presently available, a presentation with evidence of calcium, folic acid, and/or iron malabsorption but without cobalamin deficiency makes it likely that the duodenum and proximal jejunum, but not the ileum, were resected.

Some nutrients—for example glucose, amino acids, and lipids—are absorbed throughout the small intestine, although their rate of absorption is greater in the proximal than in the distal segments. However, after segmental resection of the small intestine, the remaining segments undergo both morphologic and functional "adaptation" to enhance absorption. Such adaptation is secondary to the presence of luminal nutrients and hormonal stimuli and may not be complete in humans for several months after resection. Adaptation is critical for the survival of individuals who have undergone massive resection of the small intestine and/or colon.

Establishing the presence of steatorrhea and identifying its specific cause are often quite difficult. The "gold standard" remains a timed, quantitative stool-fat determination. From a practical standpoint, stool collections are invariably difficult and often incomplete, as nobody wants to handle stool. A qualitative test—Sudan III staining—has long been available to document an increase in stool fat. This test is rapid and inexpensive but, as a qualitative test, does not establish the degree of fat malabsorption and is best used as a preliminary screening study. Many of the blood, breath, and isotopic tests that have been developed (1) do not directly measure fat absorption; (2) exhibit excellent sensitivity when steatorrhea is obvious and severe but poor sensitivity when steatorrhea is mild (e.g., assays for stool chymotrypsin and elastase, which can

potentially distinguish pancreatic from nonpancreatic etiologies of steatorrhea); or (3) have not survived the transition from the research laboratory to commercial application.

Nevertheless, routine laboratory studies (i.e., complete blood count, prothrombin time, serum protein determination, alkaline phosphatase) may suggest dietary nutrient depletion, especially deficiencies of iron, folate, cobalamin, and vitamins D and K. Additional studies include measurement of serum carotene, cholesterol, albumin, iron, folate, and cobalamin levels. The serum carotene level can also be reduced if the patient's dietary intake of leafy vegetables is poor.

If steatorrhea and/or altered absorption of other nutrients are suspected, history, clinical observations, and laboratory testing can help detect deficiency of a nutrient, especially of a fat-soluble vitamin (A, D, E, or K). Thus, evidence of metabolic bone disease with elevated alkaline phosphatase concentrations and/or reduced serum calcium levels suggests vitamin D malabsorption. A deficiency of vitamin K is suggested by an elevated prothrombin time in an individual without liver disease who is not taking anticoagulants. Macrocytic anemia leads to an evaluation for possible cobalamin or folic acid malabsorption. Iron-deficiency anemia in the absence of occult bleeding from the gastrointestinal tract in either a male patient or a nonmenstruating female patient requires an evaluation for iron malabsorption and the exclusion of celiac disease, as iron is absorbed exclusively in the proximal small intestine.

At times, however, a timed (72-h) quantitative stool collection, preferably while the patient is on a defined diet, must be undertaken in order to determine stool fat content and establish the diagnosis of steatorrhea. The presence of steatorrhea then requires further assessment to identify the pathophysiologic process(es) responsible for the defect in dietary lipid digestion/absorption (Table 318-4). Other studies include the D-xylose test, duodenal mucosal biopsy, small-intestinal radiologic examination, and tests of pancreatic exocrine function.

URINARY D-XYLOSE TEST

The urinary D-xylose test for carbohydrate absorption provides an assessment of proximal small-intestinal mucosal function. D-Xylose, a pentose, is absorbed almost exclusively in the proximal small intestine. The D-xylose test is usually performed by administering 25 g of D-xylose and collecting urine for 5 h. An abnormal test (excretion of <4.5 g) primarily reflects duodenal/jejunal mucosal disease. The D-xylose test can also be abnormal in patients with blind loop syndrome (as a consequence primarily of an abnormal intestinal mucosa) and, as a false-positive study, in patients with large collections of fluid in a third space (i.e., ascites, pleural fluid). The ease of obtaining a mucosal biopsy of the small intestine by endoscopy and the false-negative rate of the D-xylose test have led to its diminished use. When small-intestinal mucosal disease is suspected, a small-intestinal mucosal biopsy should be performed.

RADIOLOGIC EXAMINATION

Radiologic examination of the small intestine using barium contrast (small-bowel series or study) can provide important information in the evaluation of the patient with presumed or suspected malabsorption. This study is most often performed in conjunction with an examination of the esophagus, stomach, and

duodenal bulb. Because insufficient barium is given to the patient to permit an adequate examination of the small-intestinal mucosa, especially in the ileum, many gastrointestinal radiologists alter the procedure by performing either a small-bowel series in which a large amount of barium is given by mouth, without concurrent examination of the esophagus and stomach, or an enteroclysis study in which a large amount of barium is introduced into the duodenum via a fluoroscopically placed tube. In addition, many of the diagnostic features initially described by radiologists to denote the presence of small-intestinal disease (e.g., flocculation, segmentation) are rarely seen with current barium suspensions. Nonetheless, in skilled hands, barium contrast examination of the small intestine can yield important information. For example, with extensive mucosal disease, intestinal dilation can be seen as a dilution of barium from increased intestinal fluid secretion **(Fig. 318-3)**. A normal barium contrast study does *not* exclude the possibility of small-intestinal disease. However, a small-bowel series remains useful in the search for anatomic abnormalities, such as strictures and fistulas (as in Crohn's disease) or blind loop syndrome (e.g., multiple jejunal diverticula) and to define the extent of a previous surgical resection. Other imaging studies that assess the integrity of small-intestinal morphology are CT enterography and magnetic resonance enterography. Capsule endoscopy and double-balloon enteroscopy are other useful aids in the diagnostic assessment of small-intestinal pathology and most often are used to identify a small-intestinal bleeding site.

BIOPSY OF SMALL-INTESTINAL MUCOSA

A small-intestinal mucosal biopsy is essential in the evaluation of a patient with documented steatorrhea or chronic diarrhea (i.e., that lasting >3 weeks) **(Chap. 42)**. The ready availability of endoscopic equipment to examine the stomach and duodenum has led to its almost uniform use as the preferred method of obtaining histologic material from the proximal small-intestinal mucosa. The primary indications for a small-intestinal biopsy are evaluation of a patient (1) either with documented or suspected steatorrhea or with chronic diarrhea, and (2) with diffuse or focal abnormalities of the small intestine defined on a small-intestinal series. Lesions seen on small-bowel biopsy can be classified into three categories **(Table 318-6)**:

1. *Diffuse, specific lesions.* Relatively few diseases associated with altered nutrient absorption have specific histopathologic abnormalities on small-intestinal mucosal biopsy, and these diseases are uncommon. *Whipple's disease* is characterized by the presence of periodic acid–Schiff (PAS)–positive macrophages in the lamina propria; the bacilli that are also present may require electron microscopic examination for identification **(Fig. 318-4)**. *Abetalipoproteinemia* is characterized by a normal mucosal appearance except for the presence of mucosal absorptive cells that contain lipid postprandially and disappear after a prolonged period of either fat-free intake or fasting. *Immune globulin deficiency* is associated with a variety of histopathologic findings on small-intestinal mucosal biopsy. The characteristic feature is the absence of or substantial reduction in the number of plasma cells in the lamina propria; the mucosal architecture may be either perfectly normal or flat (i.e., villous atrophy). As patients with immune globulin deficiency are often infected with *Giardia lamblia, Giardia* trophozoites may also be seen in the biopsy.

2. *Patchy, specific lesions.* Several diseases feature an abnormal small-intestinal mucosa with a patchy distribution. As a result, biopsy samples obtained randomly or in the absence of endoscopically

visualized abnormalities may not reveal diagnostic features. Intestinal *lymphoma* can at times be diagnosed on mucosal biopsy by the identification of malignant lymphoma cells in the lamina propria and submucosa (Chap. 104). Dilated lymphatics in the submucosa and sometimes in the lamina propria indicate *lymphangiectasia* associated with hypoproteinemia secondary to protein loss into the intestine. *Eosinophilic gastroenteritis* comprises a heterogeneous group of disorders with a spectrum of presentations and symptoms, with an eosinophilic infiltrate of the lamina propria, and with or without peripheral eosinophilia. The patchy nature of the infiltrate and its presence in the submucosa often lead to an absence of histopathologic findings on mucosal biopsy. As the involvement of the duodenum in *Crohn's disease* is also submucosal and not necessarily continuous, mucosal biopsies are not the most direct approach to the diagnosis of duodenal Crohn's disease (Chap. 319). Amyloid deposition can be identified by Congo Red staining in some patients with *amyloidosis* involving the duodenum (Chap. 108).

3. *Diffuse, nonspecific lesions. Celiac disease* presents with a characteristic mucosal appearance on duodenal/proximal jejunal mucosal biopsy that is *not* diagnostic of the disease. The diagnosis of celiac disease is established by clinical, histologic, and immunologic responses to a gluten-free diet. *Tropical sprue* (see below) is associated with histologic findings similar to those of celiac disease after a tropical or subtropical exposure but does not respond to gluten restriction; most often symptoms improve with antibiotics and folate administration.

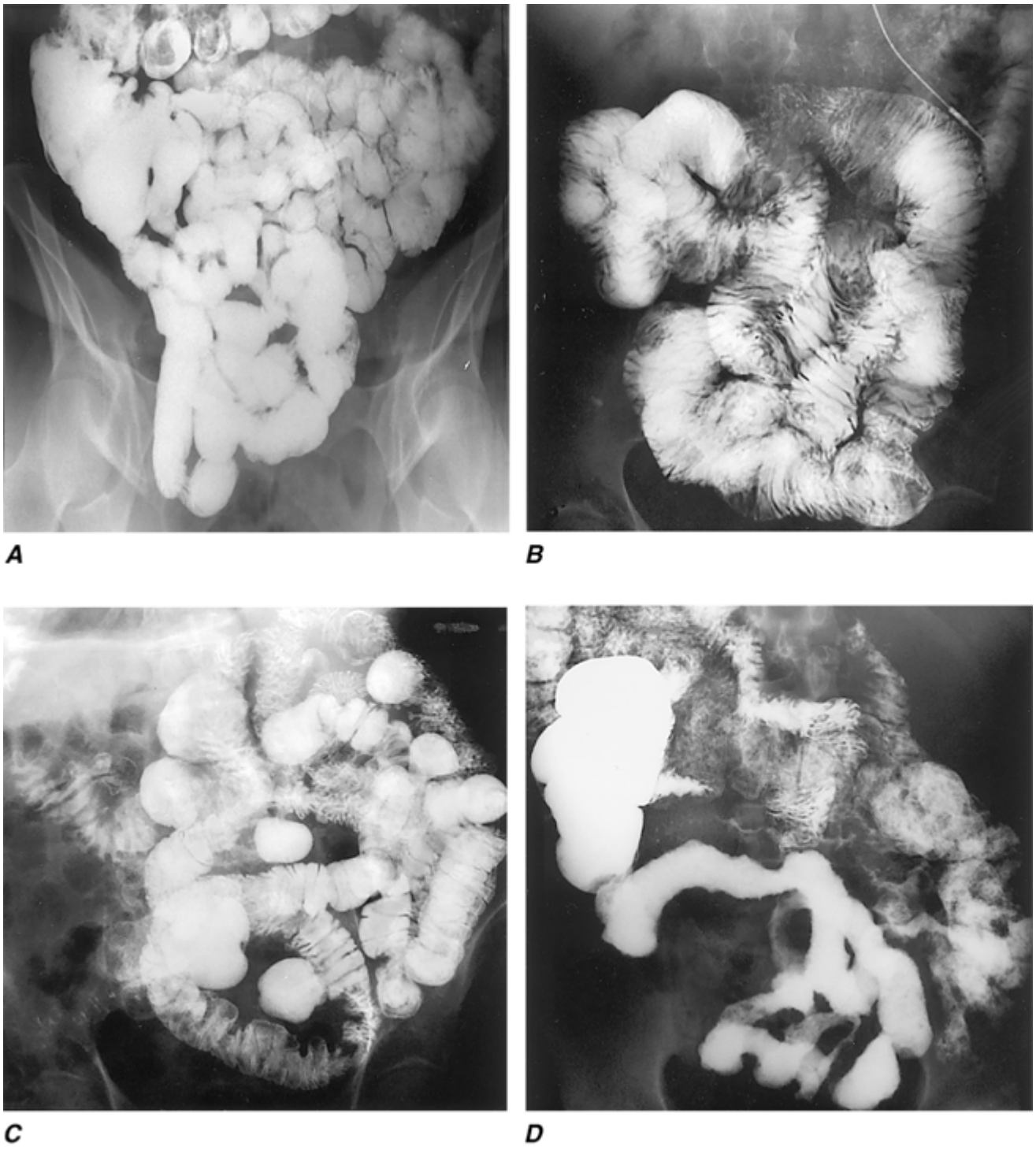
Several microorganisms can be identified in small-intestinal biopsy samples, establishing a correct diagnosis. At times, the biopsy is performed specifically to diagnose infection (e.g., Whipple's disease or giardiasis). In most other instances, the infection is detected incidentally during the workup for diarrhea or other abdominal symptoms. Many of these infections occur in immunocompromised patients with diarrhea; the etiologic agents include *Cryptosporidium*, *Isospora belli*, microsporidia, *Cyclospora*, *Toxoplasma*, cytomegalovirus, adenovirus, *Mycobacterium avium-intracellulare*, and *G. lamblia.* In immunocompromised patients, when *Candida, Aspergillus, Cryptococcus*, or *Histoplasma* organisms are seen on duodenal biopsy, their presence generally reflects systemic infection. Apart from Whipple's disease and infections in the immunocompromised host, small-bowel biopsy is seldom used as the primary mode of diagnosis of infection. Even giardiasis is more easily diagnosed by stool antigen studies and/or duodenal aspiration than by duodenal biopsy.

Patients with steatorrhea require assessment of *pancreatic exocrine function*, which is often abnormal in chronic pancreatitis. The secretin test that collects pancreatic secretions by duodenal intubation following intravenous administration of secretin is the only test that directly measures pancreatic exocrine function but is available only at a few specialized centers. Endoscopic approaches (endoscopic retrograde cholangiopancreatography, endoscopic ultrasound) provide an excellent assessment of pancreatic duct anatomy but do *not* assess exocrine function (Chap. 340).

Table 318-7 summarizes the results of the D-xylose test, the Schilling test, and small-intestinal mucosal biopsy in patients with steatorrhea of various etiologies.

FIGURE 318-3

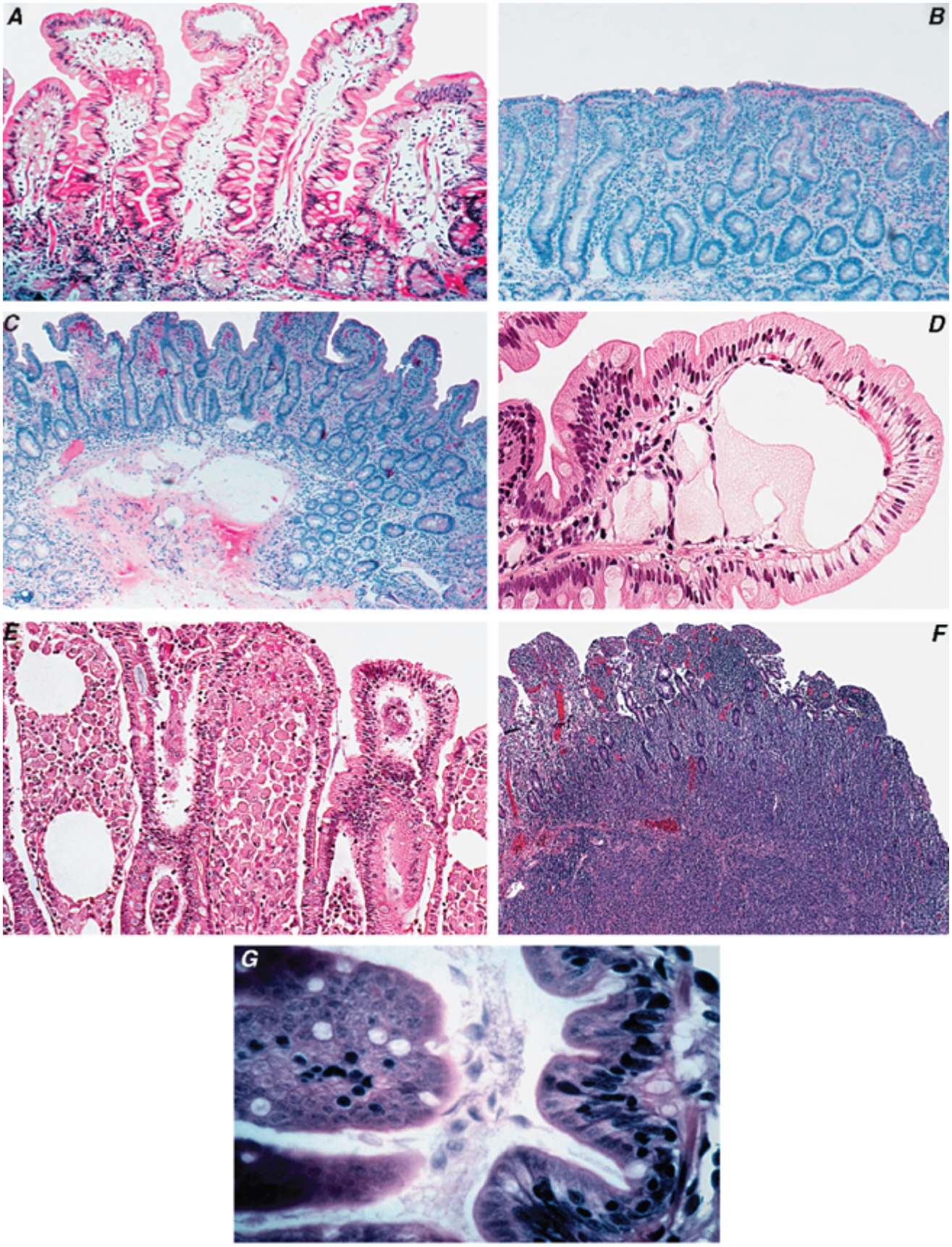
Barium contrast small-intestinal radiologic examinations. ***A.*** Normal individual. ***B.*** Celiac disease. ***C.*** Jejunal diverticulosis. ***D.*** Crohn's disease. *(Courtesy of Morton Burrell, MD, Yale University; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 318-4

Small-intestinal mucosal biopsies. ***A.*** Normal individual. ***B.*** Untreated celiac disease. ***C.*** Treated celiac disease. ***D.*** Intestinal lymphangiectasia. ***E.*** Whipple's disease. ***F.*** Lymphoma. ***G.*** Giardiasis. *(Courtesy of Marie Robert, MD, Yale University; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 318-6

Diseases That Can Be Diagnosed by Small-Intestinal Mucosal Biopsies

Lesions	Pathologic Findings
Diffuse, Specific	
Whipple's disease	Lamina propria includes macrophages containing material positive on periodic acid–Schiff staining
Agammaglobulinemia	No plasma cells; either normal or absent villi ("flat mucosa")
Abetalipoproteinemia	Normal villi; epithelial cells vacuolated with fat postprandially
Patchy, Specific	
Intestinal lymphoma	Malignant cells in lamina propria and submucosa
Intestinal lymphangiectasia	Dilated lymphatics; clubbed villi
Eosinophilic gastroenteritis	Eosinophil infiltration of lamina propria and mucosa
Amyloidosis	Amyloid deposits
Crohn's disease	Noncaseating granulomas
Infection by one or more microorganisms (see text)	Specific organisms
Mastocytosis	Mast cell infiltration of lamina propria
Diffuse, Nonspecific	
Celiac disease	Short or absent villi; mononuclear infiltrate; epithelial cell damage; hypertrophy of crypts
Tropical sprue	Similar to celiac disease
Bacterial overgrowth	Patchy damage to villi; lymphocyte infiltration
Folate deficiency	Short villi; decreased mitosis in crypts; megalocytosis

Lesions	Pathologic Findings
Vitamin B_{12} deficiency	Similar to folate deficiency
Radiation enteritis	Similar to folate deficiency
Zollinger-Ellison syndrome	Mucosal ulceration and erosion from acid
Protein-calorie malnutrition	Villous atrophy; secondary bacterial overgrowth
Drug-induced enteritis	Variable histology

TABLE 318-7

Results of Diagnostic Studies in Steatorrhea of Various Etiologies

	D-Xylose Test	Schilling Test	Duodenal Mucosal Biopsy
Chronic pancreatitis	Normal	50% abnormal; if abnormal, normal with pancreatic enzyme treatment	Normal
Bacterial overgrowth syndromes	Normal or only modestly abnormal	Often abnormal; if abnormal, normal after antibiotic treatment	Usually normal
Ileal disease	Normal	Abnormal	Normal
Celiac disease	Decreased	Normal	Abnormal: probably “flat”
Intestinal lymphangiectasia	Normal	Normal	Abnormal: “dilated lymphatics”

SPECIFIC DISEASE ENTITIES

CELIAC DISEASE

Celiac disease is a common cause of malabsorption of one or more nutrients. Although celiac disease was originally considered largely a disease of white individuals, especially persons of European descent, recent observations have established that it is a common disease with protean manifestations, a worldwide distribution, and an estimated incidence in the United States that is as high as 1 in 113 people. Its incidence has increased over the past 50 years. Celiac disease has had several other names, including nontropical sprue, celiac sprue, adult celiac disease, and gluten-sensitive enteropathy. The etiology of celiac disease is not known, but environmental, immunologic, and genetic factors are important. Celiac disease is considered an "iceberg" disease. A small number of individuals have classic symptoms and manifestations related to nutrient malabsorption along with a varied natural history; the onset of symptoms can occur at all points from the first year of life through the eighth decade. A much larger number of individuals have "atypical celiac disease," with manifestations that are not obviously related to intestinal malabsorption (e.g., anemia, osteopenia, infertility, and neurologic symptoms). Finally, an even larger number of persons have "silent celiac disease"; they are essentially asymptomatic despite abnormal small-intestinal histopathology and serologies (see below).

The hallmark of celiac disease is an abnormal small-intestinal biopsy (Fig. 318-4) and the response of the condition (including symptoms and histologic changes on small-intestinal biopsy) to the elimination of gluten from the diet. The histologic changes have a proximal-to-distal intestinal distribution of severity, which probably reflects the exposure of the intestinal mucosa to varied amounts of dietary gluten. The symptoms do not necessarily correlate with histologic changes, especially as many newly diagnosed patients with celiac disease may be asymptomatic or only minimally symptomatic (often with no gastrointestinal symptoms).

The symptoms of celiac disease may appear with the introduction of cereals into an infant's diet, although spontaneous remissions often occur during the second decade of life that may be either permanent or followed by the reappearance of symptoms over several years. Alternatively, the symptoms of celiac disease may first become evident at almost any age throughout adulthood. In many patients, frequent spontaneous remissions and exacerbations occur. The symptoms range from significant malabsorption of multiple nutrients, with diarrhea, steatorrhea, weight loss, and the consequences of nutrient depletion (i.e., anemia and metabolic bone disease), to the total absence of gastrointestinal symptoms despite evidence of the depletion of a single nutrient (e.g., iron or folate deficiency, osteomalacia, edema from protein loss). Asymptomatic relatives of patients with celiac disease have been identified as having this disease either by small-intestinal biopsy or by serologic studies (e.g., antiendomysial antibodies, tissue transglutaminase [tTG], deamidated gliadin peptide). The availability of these "celiac serologies" has led to a substantial increase in the frequency of diagnosis of celiac disease, and the diagnosis is now being made primarily in patients without "classic" symptoms but with atypical and subclinical presentations.

Etiology

The etiology of celiac disease is not known, but environmental, immunologic, and genetic factors all appear to contribute to the disease. One *environmental* factor is the clear association of the disease with gliadin, a

component of gluten that is present in wheat, barley, and rye. In addition to the role of gluten restriction in treatment, the instillation of gluten into the normal-appearing rectum and the distal ileum of patients with celiac disease results in morphologic changes within hours.

An *immunologic* component in the pathogenesis of celiac disease is critical and involves both adaptive and innate immune responses. Serum antibodies—IgA antigliadin, antiendomysial, and anti-tTG antibodies—are present, but it is not known whether such antibodies are primary or secondary to the tissue damage. The presence of antiendomysial antibody is 90–95% sensitive and 90–95% specific; the antigen recognized by antiendomysial antibody is tTG, which deaminates gliadin, which is presented to HLA-DQ2 or HLA-DQ8 (see below). Antibody studies are frequently used to identify patients with celiac disease; patients with these antibodies should undergo duodenal biopsy. This autoantibody has not been linked to a pathogenetic mechanism (or mechanisms) responsible for celiac disease. Nonetheless, this antibody is useful in establishing the true prevalence of celiac disease in the general population. A 4-week course of treatment with prednisolone induces a remission in a patient with celiac disease who continues to eat gluten and converts the “flat” abnormal duodenal biopsy to a more normal-appearing one. In addition, gliadin peptides interact with gliadin-specific T cells that mediate tissue injury and induce the release of one or more cytokines (e.g., interferon γ) that cause tissue injury.


Genetic factor(s) are also involved in celiac disease. The incidence of symptomatic celiac disease varies widely in different population groups (high among whites, low among blacks and Asians) and is 10% among first-degree relatives of celiac disease patients. However, serologic studies provide clear evidence that celiac disease is present worldwide. Furthermore, all patients with celiac disease express the HLA-DQ2 or HLA-DQ8 allele, although only a minority of people expressing DQ2/DQ8 have celiac disease. Absence of DQ2/DQ8 excludes the diagnosis of celiac disease.

Diagnosis

A small-intestinal biopsy is required to establish a diagnosis of celiac disease (Fig. 318-4). A biopsy should be performed when patients have symptoms and laboratory findings suggestive of nutrient malabsorption and/or deficiency as well as a positive tTG antibody test. Since the presentation of celiac disease is often subtle, without overt evidence of malabsorption or nutrient deficiency, a relatively low threshold for biopsy performance is important. It is more prudent to perform a biopsy than another test of intestinal absorption that can never completely exclude or establish this diagnosis.

The diagnosis of celiac disease requires the detection of characteristic histologic changes on small-intestinal biopsy together with a prompt clinical and histologic response after the institution of a gluten-free diet. If IgA antiendomysial or tTG antibodies have been detected in serologic studies, they too should disappear after a gluten-free diet is started. With the increase in the number of patients diagnosed with celiac disease (mostly by serologic studies), the spectrum of histologic changes seen on duodenal biopsy has increased and includes findings that are not as severe as the classic changes shown in Fig. 318-4. The classic changes seen on duodenal/jejunal biopsy are restricted to the mucosa and include (1) an increase in the number of

intraepithelial lymphocytes; (2) absence or a reduced height of villi, which causes a flat appearance with increased crypt cell proliferation resulting in crypt hyperplasia and loss of villous structure, with consequent villous, but not mucosal, atrophy; (3) a cuboidal appearance and nuclei that are no longer oriented basally in surface epithelial cells; and (4) increased numbers of lymphocytes and plasma cells in the lamina propria (Fig. 318-4B). Although these features are characteristic of celiac disease, they are not diagnostic because a similar appearance can develop eosinophilic enteritis, and milk-protein intolerance in children and occasionally in lymphoma, bacterial overgrowth, Crohn's disease, and gastrinoma with acid hypersecretion. However, a characteristic histologic appearance that reverts toward normal after the initiation of a gluten-free diet establishes the diagnosis of celiac disease (Fig. 318-4C). Readministration of gluten, with or without an additional small-intestinal biopsy, is not necessary.

A number of patients exhibit *nonceliac gluten sensitivity*; that is, they have gastrointestinal symptoms that respond to gluten restriction but do not have celiac disease. The basis for such gluten sensitivity is not known.

Failure to Respond to Gluten Restriction

The most common cause of persistent symptoms in a patient who fulfills all the criteria for the diagnosis of celiac disease is *continued intake of gluten*. Gluten is ubiquitous, and a significant effort must be made to exclude all gluten from the diet. Use of rice flour in place of wheat flour is very helpful, and several support groups provide important aid to patients with celiac disease and to their families. More than 90% of patients who have the characteristic findings of celiac disease respond to complete dietary gluten restriction. The remainder constitute a heterogeneous group (whose condition is often called *refractory celiac disease* or *refractory sprue*) that includes some patients who (1) respond to restriction of other dietary protein (e.g., soy); (2) respond to glucocorticoid treatment; (3) are "temporary" (i.e., whose clinical and morphologic findings disappear after several months or years); or (4) fail to respond to all measures and have a fatal outcome, with or without documented complications of celiac disease, such as the development of intestinal T cell lymphoma or autoimmune enteropathy.

Therapeutic approaches that do not include a gluten-free diet are being developed and include the use of peptidases to inactivate toxic gliadin peptides and of small molecules to block toxic peptide uptake across intestinal tight junctions.

Mechanism of Diarrhea

The diarrhea in celiac disease has several pathogenetic mechanisms. Diarrhea may be secondary to (1) steatorrhea, which is primarily a result of changes in jejunal mucosal function; (2) secondary lactase deficiency, a consequence of changes in jejunal brush border enzymatic function; (3) bile-acid malabsorption resulting in bile-acid–induced fluid secretion in the colon (in cases with more extensive disease involving the ileum); and (4) endogenous fluid secretion resulting from crypt hyperplasia. Celiac disease patients with more severe involvement may improve temporarily with *dietary lactose and fat restriction* while awaiting the full effects of total gluten restriction, which constitutes primary therapy.

Associated Diseases

Celiac disease is associated with dermatitis herpetiformis (DH), but this association has not been explained. Patients with DH have characteristic papulovesicular lesions that respond to dapsone. Almost all patients with DH have histologic changes in the small intestine consistent with celiac disease, although usually much milder and less diffuse in distribution. Most patients with DH have mild or no gastrointestinal symptoms. In contrast, relatively few patients with celiac disease have DH.

Celiac disease is also associated with diabetes mellitus type 1, IgA deficiency, Down's syndrome, and Turner's syndrome. The clinical importance of the association with diabetes is that, although severe watery diarrhea without evidence of malabsorption is most often diagnosed as "diabetic diarrhea" **(Chap. 396)**, assay of antiendomysial antibodies and/or a small-intestinal biopsy must be considered to exclude celiac disease.

Complications

The most important complication of celiac disease is the development of cancer. The incidences of both gastrointestinal and nongastrointestinal neoplasms as well as intestinal lymphoma are elevated among patients with celiac disease. For unexplained reasons, the frequency of lymphoma in patients with celiac disease is higher in Ireland and the United Kingdom than in the United States. The possibility of lymphoma must be considered whenever a patient with celiac disease who has previously done well on a gluten-free diet is no longer responsive to gluten restriction or a patient who presents with clinical and histologic features consistent with celiac disease does not respond to a gluten-free diet. Other complications of celiac disease include the development of intestinal ulceration independent of lymphoma and so-called refractory sprue (see above) and collagenous sprue. In *collagenous sprue*, a layer of collagen-like material is present beneath the basement membrane; patients with collagenous sprue generally do not respond to a gluten-free diet and often have a poor prognosis.

TROPICAL SPRUE



Tropical sprue is a poorly understood syndrome that affects both expatriates and natives in certain but not all tropical areas and is manifested by chronic diarrhea, steatorrhea, weight loss, and nutritional deficiencies, including those of both folate and cobalamin. This disease affects 5–10% of the population in some tropical areas.

Chronic diarrhea in a tropical environment is most often caused by infectious agents, including *G. lamblia*, *Yersinia enterocolitica*, *C. difficile*, *Cryptosporidium parvum*, and *Cyclospora cayetanensis*. Tropical sprue should not be entertained as a possible diagnosis until the presence of cysts and trophozoites has been excluded in three stool samples. **Chronic infections of the gastrointestinal tract and diarrhea in patients with or without AIDS are discussed in Chaps. 128, 129, and 197.**

The small-intestinal mucosa of individuals living in tropical areas is not identical to that of individuals who reside in temperate climates. In residents of tropical areas, biopsies reveal a mild alteration of villous

architecture with a modest increase in mononuclear cells in the lamina propria, which on occasion can be as severe as that seen in celiac disease. These changes are observed both in native residents and in expatriates living in tropical regions and are usually associated with mild decreases in absorptive function, but they revert to “normal” when an individual moves or returns to a temperate area. Some have suggested that the changes seen in tropical enteropathy and in tropical sprue represent different ends of the spectrum of a single entity, but convincing evidence to support this concept is lacking.

In the past few years the term “environmental enteropathy” has been introduced as the diagnosis of many patients (especially infants and children) who had previously been diagnosed as tropical sprue. Such patients have physical and/or mental shunting. However, exact delineation of this newly designated entity is lacking.

Etiology

Because tropical sprue responds to antibiotics, the consensus is that it may be caused by one or more infectious agents. Nonetheless, the etiology and pathogenesis of tropical sprue are uncertain. First, its occurrence is not evenly distributed in all tropical areas; rather, it is found in specific locations, including southern India, the Philippines, and several Caribbean islands (e.g., Puerto Rico, Haiti), but is rarely observed in Africa, Jamaica, or Southeast Asia. Second, an occasional individual does not develop symptoms of tropical sprue until long after having left an endemic area. For this reason, celiac disease (often referred to as celiac sprue) was originally called *nontropical sprue* to distinguish it from tropical sprue. Third, multiple microorganisms have been identified in jejunal aspirates, with relatively little consistency among studies. *Klebsiella pneumoniae*, *Enterobacter cloacae*, and *E. coli* have been implicated in some studies of tropical sprue, while other studies have favored a role for a toxin produced by one or more of these bacteria. Fourth, the incidence of tropical sprue appears to have decreased substantially during the past two or three decades, perhaps in relation to improved sanitation in many tropical countries during this time. Some have speculated that the reduced occurrence is attributable to the wider use of antibiotics in acute diarrhea, especially in travelers to tropical areas from temperate countries. Fifth, the role of folic acid deficiency in the pathogenesis of tropical sprue requires clarification. Folic acid is absorbed exclusively in the duodenum and proximal jejunum, and most patients with tropical sprue have evidence of folate malabsorption and depletion. Although folate deficiency can cause changes in small-intestinal mucosa that are corrected by folate replacement, several earlier studies reporting that tropical sprue could be cured by folic acid did not provide an explanation for the “insult” that was initially responsible for folate malabsorption.

The clinical pattern of tropical sprue varies in different areas of the world (e.g., India vs Puerto Rico). Not infrequently, individuals in southern India initially report the occurrence of acute enteritis before the development of steatorrhea and malabsorption. In contrast, in Puerto Rico, a more insidious onset of symptoms and a more dramatic response to antibiotics are seen than in some other locations. Tropical sprue in different areas of the world may not be the same disease, and similar clinical entities may have different etiologies.

Diagnosis

The diagnosis of tropical sprue is best based on an abnormal small-intestinal mucosal biopsy in an individual with chronic diarrhea and evidence of malabsorption who is either residing or has recently lived in a tropical country. The small-intestinal biopsy in tropical sprue does not reveal pathognomonic features but resembles, and can often be indistinguishable from, that seen in celiac disease (Fig. 318-4). The biopsy sample in tropical sprue has less villous architectural alteration and more mononuclear cell infiltrate in the lamina propria. In contrast to those of celiac disease, the histologic features of tropical sprue manifest with a similar degree of severity throughout the small intestine, and a gluten-free diet does not result in either clinical or histologic improvement in tropical sprue.

TREATMENT

TREATMENT

Tropical Sprue

Broad-spectrum antibiotics and folic acid are most often curative, especially if the patient leaves the tropical area and does not return. Tetracycline should be used for up to 6 months and may be associated with improvement within 1–2 weeks. Folic acid alone induces hematologic remission as well as improvement in appetite, weight gain, and some morphologic changes in small-intestinal biopsy. Because of marked folate deficiency, folic acid is most often given together with antibiotics.

SHORT-BOWEL SYNDROME

Short-bowel syndrome is a descriptive term for the myriad clinical problems that follow resection of various lengths of the small intestine or, on rare occasions, are congenital (e.g., microvillous inclusion disease). The factors that determine both the type and degree of symptoms include (1) the specific segment (jejunum vs ileum) resected, (2) the length of the resected segment, (3) the integrity of the ileocecal valve, (4) whether any large intestine has also been removed, (5) residual disease in the remaining small and/or large intestine (e.g., Crohn's disease, mesenteric artery disease), and (6) the degree of adaptation in the remaining intestine. Short-bowel syndrome can occur in persons of any age, from neonates to the elderly.

Three different situations in adults mandate intestinal resection: (1) mesenteric vascular disease, including atherosclerosis, thrombotic phenomena, and vasculitides; (2) primary mucosal and submucosal disease (e.g., Crohn's disease); and (3) operations without preexisting small-intestinal disease (e.g., after trauma).

After resection of the small intestine, the residual intestine undergoes adaptation of both structure and function that may last for up to 6–12 months. Continued intake of dietary nutrients and calories is required to stimulate adaptation via direct contact with the intestinal mucosa, the release of one or more intestinal hormones, and pancreatic and biliary secretions. Thus, enteral nutrition with calorie administration must be maintained, especially in the early postoperative period, even if an extensive intestinal resection requiring parenteral nutrition (PN) has been performed. The subsequent ability of such patients to absorb nutrients will not be known for several months, until adaptation is complete.

Multiple factors besides the absence of intestinal mucosa (required for lipid, fluid, and electrolyte absorption) contribute to diarrhea and steatorrhea in these patients. Removal of the ileum, and especially the ileocecal valve, is often associated with more severe diarrhea than jejunal resection. Without part or all of the ileum, diarrhea can be caused by an increase in bile acids entering the colon; these acids stimulate colonic fluid and electrolyte secretion. Absence of the ileocecal valve is also associated with a decrease in intestinal transit time and bacterial overgrowth from the colon. The presence of the colon (or a major portion) is associated with substantially less diarrhea and a lower likelihood of *intestinal failure* (an inability to maintain nutrition without parenteral support) as a result of fermentation of nonabsorbed carbohydrates to SCFAs. The latter are absorbed in the colon and stimulate Na and water absorption, improving overall fluid balance. Lactose intolerance as a result of the removal of lactase-containing mucosa as well as gastric hypersecretion may also contribute to the diarrhea.

In addition to diarrhea and/or steatorrhea, a range of nonintestinal symptoms is observed in some patients. The frequency of renal calcium oxalate calculi increases significantly in patients with a small-intestinal resection and an intact colon; this greater frequency is due to an increase in oxalate absorption by the large intestine, with subsequent *enteric hyperoxaluria*. Two possible mechanisms for the increase in oxalate absorption in the colon have been suggested: (1) increased bile acids and fatty acids that augment colonic mucosal permeability, resulting in enhanced oxalate absorption; and (2) increased fatty acids that bind calcium, resulting in an enhanced amount of soluble oxalate that is then absorbed. Since oxalate is high in relatively few foods (e.g., spinach, rhubarb, tea), dietary restrictions alone do not constitute adequate treatment. Cholestyramine (an anion-binding resin) and calcium have proved useful in reducing hyperoxaluria. Similarly, an increase in cholesterol gallstones is related to a decrease in the bile-acid pool size, which results in the generation of cholesterol supersaturation in gallbladder bile. Gastric hypersecretion of acid occurs in many patients after large resections of the small intestine. The etiology is unclear but may be related to either reduced hormonal inhibition of acid secretion or increased gastrin levels due to reduced small-intestinal catabolism of circulating gastrin. The resulting gastric acid secretion may be an important factor contributing to diarrhea and steatorrhea. A reduced pH in the duodenum can inactivate pancreatic lipase and/or precipitate duodenal bile acids, thereby increasing steatorrhea, and an increase in gastric secretion can create a volume overload relative to the reduced small-intestinal absorptive capacity. Inhibition of gastric acid secretion with proton pump inhibitors can help reduce diarrhea and steatorrhea.

TREATMENT

TREATMENT

Short-Bowel Syndrome

Treatment of short-bowel syndrome depends on the severity of symptoms and on whether the individual is able to maintain caloric and electrolyte balance with oral intake alone. Initial treatment includes judicious use of opiates (including codeine) to reduce stool output and establish an effective diet. If the colon is in situ, the initial diet should be low in fat and high in carbohydrate in order to minimize diarrhea from fatty acid stimulation of colonic fluid secretion. MCTs (see Table 318-3), a low-lactose diet, and various soluble fiber–

containing diets should also be tried. In the absence of an ileocecal valve, possible bacterial overgrowth must be considered and treated. If gastric acid hypersecretion is contributing to diarrhea and steatorrhea, a proton pump inhibitor may be helpful. Usually none of these therapeutic approaches provides an instant solution, but each can contribute to the reduction of disabling diarrhea.

The patient's vitamin and mineral status must also be monitored; replacement therapy should be initiated if indicated. Fat-soluble vitamins, folate, cobalamin, calcium, iron, magnesium, and zinc are the most critical factors to monitor on a regular basis. If these approaches are not successful, home PN is an established therapy that can be maintained for many years. Small-intestinal transplantation is becoming established as a possible approach for individuals with extensive intestinal resection who cannot be maintained without PN—that is, those with intestinal failure. A recombinant analogue of glucagon-like peptide 2 (GLP-2; teduglutide) is approved for use in patients with PN-dependent short-bowel syndrome on the basis of its ability to increase intestinal growth, improve absorption, and reduce requirement for PN.

BACTERIAL OVERGROWTH SYNDROMES

Bacterial overgrowth syndromes comprise a group of disorders with diarrhea, steatorrhea, and macrocytic anemia whose common feature is the proliferation of colonic-type bacteria within the small intestine. This bacterial proliferation is due to stasis caused by impaired peristalsis (*functional stasis*), changes in intestinal anatomy (*anatomic stasis*), or direct communication between the small and large intestine. These conditions have also been referred to as *stagnant bowel syndrome* or *blind loop syndrome.*

Pathogenesis

The manifestations of bacterial overgrowth syndromes are a direct consequence of the presence of increased amounts of a colonic-type bacterial flora, such as *E. coli* or *Bacteroides*, in the small intestine. *Macrocytic anemia* is due to cobalamin—not folate—deficiency. Most bacteria require cobalamin for growth, and increased concentrations of bacteria use up the relatively small amounts of dietary cobalamin. *Steatorrhea* is due to impaired micelle formation as a consequence of a reduced intraduodenal concentration of conjugated bile acids and the presence of unconjugated bile acids. Certain bacteria, including *Bacteroides*, deconjugate conjugated bile acids to unconjugated bile acids. Unconjugated bile acids are absorbed more rapidly than conjugated bile acids; as a result, the intraduodenal concentration of bile acids is reduced. In addition, the CMC of unconjugated bile acids is higher than that of conjugated bile acids, and the result is a decrease in micelle formation. *Diarrhea* is due, at least in part, to steatorrhea, when it is present. However, some patients manifest diarrhea *without* steatorrhea, and it is assumed that the colonic-type bacteria in these patients are producing one or more bacterial enterotoxins that are responsible for fluid secretion and diarrhea.

Etiology

The etiology of these different disorders is bacterial proliferation in the small-intestinal lumen secondary to anatomic or functional stasis or to a communication between the relatively sterile small intestine and the colon, with its high levels of aerobic and anaerobic bacteria. Several examples of *anatomic* stasis have been

identified: (1) one or more diverticula (both duodenal and jejunal) (Fig. 318-3C); (2) fistulas and strictures related to Crohn's disease (Fig. 318-3D); (3) a proximal duodenal afferent loop following subtotal gastrectomy and gastrojejunostomy; (4) a bypass of the intestine (e.g., a jejunoileal bypass for obesity); and (5) dilation at the site of a previous intestinal anastomosis. These anatomic derangements are often associated with the presence of a segment (or segments) of intestine out of continuity of propagated peristalsis, with consequent stasis and bacterial proliferation. Bacterial overgrowth syndromes can also occur in the *absence* of an anatomic blind loop when *functional* stasis is present. Impaired peristalsis and bacterial overgrowth in the absence of a blind loop occur in scleroderma, where motility abnormalities exist in both the esophagus and the small intestine **(Chap. 353)**. Functional stasis and bacterial overgrowth can also develop in association with diabetes mellitus and in the small intestine when a direct connection exists between the small and large intestines, including an ileocolonic resection, or occasionally after an enterocolic anastomosis that permits entry of bacteria into the small intestine as a result of bypassing the ileocecal valve.

Diagnosis

The diagnosis may be suspected from the combination of a low serum cobalamin level and an elevated serum folate level, as enteric bacteria frequently produce folate compounds that are absorbed in the duodenum. Ideally, the bacterial overgrowth syndromes are diagnosed by the demonstration of increased levels of aerobic and/or anaerobic colonic-type bacteria in a jejunal aspirate obtained by intubation. However, this specialized test is rarely available. Breath hydrogen testing with administration of lactulose (a nondigestible disaccharide) has also been used to detect bacterial overgrowth. Often, the diagnosis is suspected clinically and confirmed by the response to treatment.

TREATMENT

TREATMENT

Bacterial Overgrowth Syndromes

Primary treatment should be directed, if at all possible, to the surgical correction of an anatomic blind loop. In the absence of functional stasis, it is important to define the anatomic relationships responsible for stasis and bacterial overgrowth. For example, bacterial overgrowth secondary to strictures, one or more diverticula, or a proximal afferent loop can potentially be cured by surgical correction of the anatomic state. In contrast, the functional stasis of scleroderma or certain anatomic stasis states (e.g., multiple jejunal diverticula) cannot be corrected surgically, and these conditions should be treated with broad-spectrum antibiotics. Tetracycline used to be the initial drug of choice; because of increasing resistance, however, other antibiotics, such as metronidazole, amoxicillin/clavulanic acid, rifaximin and cephalosporins, have been employed. The antibiotic should be given for ~3 weeks or until symptoms remit. Although the natural history of these conditions is chronic, antibiotics should not be given continuously. Symptoms usually remit within 2–3 weeks of initial antibiotic therapy. Treatment need not be repeated until symptoms recur. For frequent recurrences, several treatment strategies exist, but the use of antibiotics for 1 week per month, whether or not symptoms are present, is often most effective.

Unfortunately, therapy for bacterial overgrowth syndromes is largely empirical, with an absence of clinical trials on which to base rational decisions regarding antibiotic choice, treatment duration, and/or the best approach to therapy for recurrences. Bacterial overgrowth may also occur as a component of another chronic disease, such as Crohn's disease, radiation enteritis, or short-bowel syndrome. Treatment of the bacterial overgrowth in these settings will not cure the underlying problem but may be very important in ameliorating a subset of clinical problems that are related to bacterial overgrowth.

WHIPPLE'S DISEASE

Whipple's disease is a chronic multisystemic disease associated with diarrhea, steatorrhea, weight loss, arthralgia, and central nervous system (CNS) and cardiac problems; it is caused by the bacterium *Tropheryma whipplei.* Until the identification of *T. whipplei* by polymerase chain reaction, the hallmark of Whipple's disease had been the presence of PAS-positive macrophages in the small intestine (Fig. 318-4E) and other organs with evidence of disease.

Etiology

T. whipplei, a small (50–500 nm) gram-positive bacillus in the group Actinobacteria, has low virulence but high infectivity. Symptoms of Whipple's disease are relatively minimal compared to the bacterial burden in multiple tissues.

Clinical Presentation

The onset of Whipple's disease is insidious and is characterized by diarrhea, steatorrhea, abdominal pain, weight loss, migratory large-joint arthropathy, and fever as well as ophthalmologic and CNS symptoms. Dementia is a relatively late symptom and an extremely poor prognostic sign, especially in patients who experience relapse after the induction of a remission with antibiotics. For unexplained reasons, the disease occurs primarily in middle-aged white men. The steatorrhea in these patients is generally believed to be secondary to both small-intestinal mucosal injury and lymphatic obstruction due to the increased number of PAS-positive macrophages in the lamina propria of the small intestine.

Diagnosis

The diagnosis of Whipple's disease is suggested by a multisystemic disease in a patient with diarrhea and steatorrhea. Tissue biopsy of the small intestine and/or other organs that may be involved (e.g., liver, lymph nodes, heart, eyes, CNS, or synovial membranes), given the patient's symptoms, is the primary approach. The presence of PAS-positive macrophages containing the characteristic small bacilli is suggestive of this diagnosis. However, *T. whipplei*–containing macrophages can be confused with PAS-positive macrophages containing *M. avium* complex, which may be a cause of diarrhea in AIDS. The presence of the *T. whipplei* bacillus outside of macrophages is a more important indicator of active disease than is their presence within the macrophages. *T. whipplei* has now been successfully grown in culture.

TREATMENT

TREATMENT

Whipple's Disease

The treatment for Whipple's disease is prolonged use of antibiotics. The current regimen of choice is ceftriaxone or meropenem for 2 weeks followed by oral TMP-SMX (160/800 mg) twice a day for 1 year. PAS-positive macrophages can persist after successful treatment, and the presence of bacilli outside of macrophages is indicative of persistent infection or an early sign of recurrence. Recurrence of disease activity, especially with dementia, is an extremely poor prognostic sign and requires an antibiotic that crosses the blood-brain barrier. If trimethoprim-sulfamethoxazole is not tolerated, chloramphenicol is an appropriate second choice.

PROTEIN-LOSING ENTEROPATHY

Protein-losing enteropathy is not a specific disease but rather a group of gastrointestinal and nongastrointestinal disorders with hypoproteinemia and edema in the absence of either proteinuria or defects in protein synthesis (e.g., chronic liver disease). These diseases are characterized by excess protein loss into the gastrointestinal tract. Normally, ~10% of total protein catabolism occurs via the gastrointestinal tract. Evidence of increased protein loss into the gastrointestinal tract is found in >65 different diseases, which can be classified into three groups: (1) mucosal ulceration, such that the protein loss primarily represents exudation across damaged mucosa (e.g., ulcerative colitis, gastrointestinal carcinomas, and peptic ulcer); (2) nonulcerated mucosa, but with evidence of mucosal damage so that the protein loss represents loss across epithelia with altered permeability (e.g., celiac disease and Ménétrier's disease in the small intestine and stomach, respectively); and (3) lymphatic dysfunction, representing either primary lymphatic disease or lymphatic disease secondary to partial lymphatic obstruction that may occur as a result of enlarged lymph nodes or cardiac disease.

Diagnosis

The diagnosis of protein-losing enteropathy is suggested by peripheral edema and low serum albumin and globulin levels in the absence of renal and hepatic disease. An individual with protein-losing enteropathy only rarely has selective loss of *only* albumin or *only* globulins. Therefore, marked reduction of serum albumin with normal serum globulins should not initiate an evaluation for protein-losing enteropathy but should suggest renal and/or hepatic disease. Likewise, reduced serum globulins with normal serum albumin levels are more likely a result of reduced globulin synthesis rather than enhanced globulin loss into the intestine. An increase in protein loss into the gastrointestinal tract has been documented by the administration of one of several radiolabeled proteins and its quantitation in stool during a 24- or 48-h period. Unfortunately, none of these radiolabeled proteins is available for routine clinical use. α_1-Antitrypsin, a protein that accounts for ~4% of total serum proteins and is resistant to proteolysis, can be used to detect enhanced rates of serum protein loss into the intestinal tract but cannot be used to assess gastric protein loss

because of its degradation in an acid milieu. α_1-Antitrypsin clearance is measured by determining stool volume as well as both stool and plasma α_1-antitrypsin concentrations. In addition to the loss of protein via abnormal and distended lymphatics, peripheral lymphocytes may be lost via lymphatics, with consequent relative lymphopenia. Thus, lymphopenia in a patient with hypoproteinemia indicates increased loss of protein into the gastrointestinal tract.

Patients with increased protein loss into the gastrointestinal tract from lymphatic obstruction often have steatorrhea and diarrhea. The steatorrhea is a result of altered lymphatic flow as lipid-containing chylomicrons exit from intestinal epithelial cells via intestinal lymphatics (Table 318-4; Fig. 318-4). In the absence of mechanical or anatomic lymphatic obstruction, intrinsic intestinal lymphatic dysfunction—with or without lymphatic dysfunction in the peripheral extremities—has been designated *intestinal lymphangiectasia.* Similarly, ~50% of individuals with intrinsic peripheral lymphatic disease (Milroy's disease) also have intestinal lymphangiectasia and hypoproteinemia. Other than steatorrhea and enhanced protein loss into the gastrointestinal tract, all other aspects of intestinal absorptive function are normal in intestinal lymphangiectasia.

Other Causes

Patients who appear to have idiopathic protein-losing enteropathy without evidence of gastrointestinal disease should be examined for cardiac disease—especially right-sided valvular disease and chronic pericarditis **(Chaps. 263 and 265)**. On occasion, hypoproteinemia can be the only presenting manifestation in these two types of heart disease. Ménétrier's disease (also called *hypertrophic gastropathy*) is an uncommon entity that involves the body and fundus of the stomach and is characterized by large gastric folds, reduced gastric acid secretion, and, at times, enhanced protein loss into the stomach.

TREATMENT

TREATMENT

Protein-Losing Enteropathy

As excess protein loss into the gastrointestinal tract is most often secondary to a specific disease, treatment should be directed primarily to the underlying disease process and not to the hypoproteinemia. For example, if significant hypoproteinemia with resulting peripheral edema is secondary to celiac disease or ulcerative colitis, a gluten-free diet and mesalamine, respectively, would be the initial therapy. When enhanced protein loss is secondary to lymphatic obstruction, it is critical to establish the nature of this obstruction. Identification of mesenteric nodes or lymphoma may be possible by imaging studies. Similarly, it is important to exclude cardiac disease as a cause of protein-losing enteropathy, either by echosonography or, on occasion, by a right-heart catheterization.

The increased protein loss that occurs in intestinal lymphangiectasia is a result of distended lymphatics associated with lipid malabsorption. The hypoproteinemia is treated with a low-fat diet and the

administration of MCTs (Table 318-3), which do not exit from the intestinal epithelial cells via lymphatics but are delivered to the body via the portal vein.

SUMMARY

The many conditions that can produce malabsorption are classified by their pathophysiology in **Table 318-8**. The pathophysiology of the various clinical manifestations of malabsorption is summarized in **Table 318-9**.

TABLE 318-8

Classification of Malabsorption Syndromes

Inadequate digestion	
Postgastrectomy[a]	
Deficiency or inactivation of pancreatic lipase	
Exocrine pancreatic insufficiency	
Chronic pancreatitis	
Pancreatic carcinoma	
Cystic fibrosis	
Pancreatic insufficiency—congenital or acquired	
Gastrinoma—acid inactivation of lipase	
Drugs—orlistat	
Reduced intraduodenal bile-acid concentration/impaired micelle formation	
Liver disease	
Parenchymal liver disease	
Cholestatic liver disease	
Bacterial overgrowth in small intestine:	
Anatomic stasis	Functional stasis
Afferent loop	Diabetes[a]
Stasis/blind	Scleroderma[a]
Loop/strictures/fistulae	Intestinal pseudo-obstruction
Interrupted enterohepatic circulation of bile salts	

Ileal resection	
Crohn’s disease	
Drugs (binding or precipitating bile salts)—neomycin, cholestyramine, calcium carbonate	
Impaired mucosal absorption/mucosal loss or defect	
Intestinal resection or bypass[a]	
Inflammation, infiltration, or infection:	
Crohn’s disease[a]	Celiac disease
Amyloidosis	Collagenous sprue
Scleroderma[a]	Whipple’s disease[a]
Lymphoma[a]	Radiation enteritis[a]
Eosinophilic enteritis	Folate and vitamin B_{12} deficiency
Mastocytosis	Infections—giardiasis
Tropical sprue	Graft versus host disease
Genetic disorders	
Disaccharidase deficiency	
Agammaglobulinemia	
Abetalipoproteinemia	
Hartnup’s disease	
Cystinuria	
Impaired nutrient delivery to and/or from intestine:	
Lymphatic obstruction	Circulatory disorders

Lymphoma[a]	Congestive heart failure
Lymphangiectasia	Constrictive pericarditis
	Mesenteric artery atherosclerosis
	Vasculitis
Endocrine and metabolic disorders	
Diabetes[a]	
Hypoparathyroidism	
Adrenal insufficiency	
Hyperthyroidism	
Carcinoid syndrome	

[a]Malabsorption caused by more than one mechanism.

TABLE 318-9

Pathophysiology of Clinical Manifestations of Malabsorption Disorders

Symptom or Sign	Mechanism
Weight loss/malnutrition	Anorexia, malabsorption of nutrients
Diarrhea	Impaired absorption or secretion of water and electrolytes; colonic fluid secretion secondary to unabsorbed dihydroxy bile acids and fatty acids
Flatus	Bacterial fermentation of unabsorbed carbohydrate
Glossitis, cheilosis, stomatitis	Deficiency of iron, vitamin B_{12}, folate, and vitamin A
Abdominal pain	Bowel distention or inflammation, pancreatitis
Bone pain	Calcium, vitamin D malabsorption, protein deficiency, osteoporosis
Tetany, paresthesia	Calcium and magnesium malabsorption
Weakness	Anemia, electrolyte depletion (particularly K^+)
Azotemia, hypotension	Fluid and electrolyte depletion
Amenorrhea, decreased libido	Protein depletion, decreased calories, secondary hypopituitarism
Anemia	Impaired absorption of iron, folate, vitamin B_{12}
Bleeding	Vitamin K malabsorption, hypoprothrombinemia
Night blindness/xerophthalmia	Vitamin A malabsorption
Peripheral neuropathy	Vitamin B_{12} and thiamine deficiency
Dermatitis	Deficiency of vitamin A, zinc, and essential fatty acid

FURTHER READING

Binder HJ: Role of colonic short chain fatty acid transport in diarrhea. Annu Rev Physiol 72:297, 2010. [PubMed: 20148677]

Carroll RE et al: Management and complications of short bowel syndrome: An updated review. Curr Gastroenterol Rep 18:40, 2016. [PubMed: 27324885]

Desnues B, Al Moussawi K, Fenollar F: New insights into Whipple's disease and *Tropheryma whipplei* infections. Microbes Infect 12:1102, 2010. [PubMed: 20708091]

Ghoshal UC et al: Tropical sprue in 2014: The new face of an old disease. Curr Gastroenterol Rep 16:391, 2014. [PubMed: 24781741]

Hill ID et al: NASPGHAN Clinical Report on the diagnosis and treatment of gluten-related disorders. J Pediatr Gastroenerol Nutr 63:156, 2016.

Kelly CP et al: Advances in diagnosis and management of celiac disease. Gastroenterology 148:1175, 2015. [PubMed: 25662623]

Korpe PS, Petri WA: Environmental enteropathy: critical implications of a poorly studied condition. Trends Mol Med 18:328, 2012. [PubMed: 22633998]

Rezaeie A, Pimentel M, Rao SS: How to test and treat small intestinal bacterial overgrowth: An evidence-based approach. Curr Gastroenterol Rep 18:8, 2016. [PubMed: 26780631]

Chapter 319: Inflammatory Bowel Disease

Sonia Friedman; Richard S. Blumberg

INTRODUCTION

Inflammatory bowel disease (IBD) is an immune-mediated chronic intestinal condition. Ulcerative colitis (UC) and Crohn's disease (CD) are the two major types of IBD.

GLOBAL CONSIDERATIONS: EPIDEMIOLOGY

The highest incidence rates of CD and UC have been reported in northern Europe, the United Kingdom, and North America. Countries in the Pacific, including New Zealand and Australia, which share many possible environmental risk factors and similar genetic background as northwest Europe and North America, have high incidence rates of IBD.

The highest reported incidence rates are in Canada (19.2 per 1000,000 for UC and 20.2 for CD), Northern Europe (24.3 per 100,000 for UC in Iceland and 10.6 per 100,000 for CD in the United Kingdom), and Australia (17.4 per 100,000 for UC and 29.3 per 100,000 for CD). Prevalence is highest in Europe (505 per 100,000 for UC in Norway and 322 per 100,000 for CD in Italy) and Canada (248 per 100,000 for UC and 319 per 100,000 for CD) (Table 319-1). Based on these estimates ~0.6% of the Canadian population has IBD.

TABLE 319-1

Epidemiology of IBD

	Ulcerative Colitis	Crohn's Disease
Incidence (North America) per person-years	0–19.2 per 100,000	0–20.2 per 100,000
Age of onset	Second to fourth decades and seventh to ninth decades	Second to fourth decades and seventh to ninth decades
Ethnicity	Jewish > non-Jewish white > African American > Hispanic > Asian	
Female/male ratio	0.51–1.58	0.34–1.65
Smoking	May prevent disease (odds ratio 0.58)	May cause disease (odds ratio 1.76)
Oral contraceptives	No increased risk	Hazard ratio 2.82
Appendectomy	Protective (risk reduction 13–26%)	Not protective
Monozygotic twins	6–18% concordance	38–58% concordance
Dizygotic twins	0–2% concordance	4% concordance
Antibiotic use in the first year of life	2.9× the risk of developing childhood IBD	

Abbreviation: IBD, inflammatory bowel disease.

In countries that are becoming more Westernized, including China, South Korea, India, Lebanon, Iran, Thailand, and countries in the French West Indies and North Africa, IBD appears to be emerging, emphasizing the importance of environmental factors in disease pathogenesis. In Japan, the prevalence of CD has risen rapidly from 2.9 cases per 100,000 in 1986 to 13.5 per 100,000 in 1998, whereas in South Korea, the prevalence of UC has quadrupled from 7.6 per 100,000 in 1997 to 30.9 per 100,000 in 2005. In Hong Kong, the prevalence of UC almost tripled from 2.3 in 1997 to 6.3 per 100,000 over a 9-year period.

In Singapore, the prevalence of CD increased from 1.3 in 1990 to 7.2 per 100,000 in 2004. In China the number of cases of UC has increased by fourfold between 1981–1990 and 1991–2000.

Increasing immigration to Western societies also has an impact on the incidence and prevalence of IBD. The prevalence of UC among southern Asians who immigrated to the United Kingdom (UK) was higher in comparison to the European UK population (17 cases per 100,000 persons vs 7 per 100,000). Spanish patients who emigrated within Europe, but not those who immigrated to Latin America, developed IBD more frequently than controls. Individuals who have immigrated to Westernized countries and then returned to their country of birth also continue to demonstrate an increased risk of developing IBD.

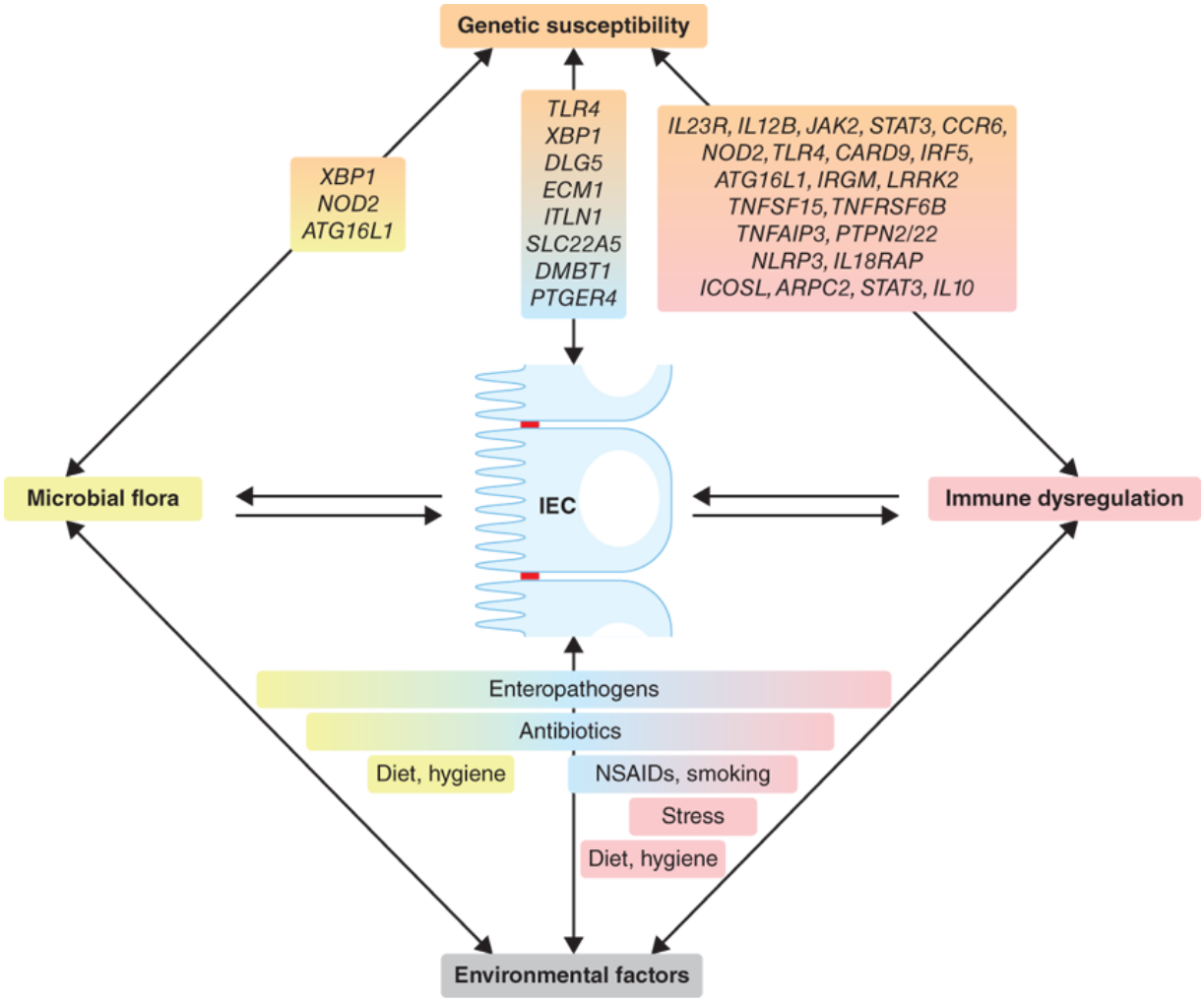
Peak incidence of UC and CD is in the second to fourth decades, with 78% of CD studies and 51% of UC studies reporting the highest incidence among those age 20–29 years old. A second modest rise in incidence occurs between the seventh and ninth decades of life. The female-to-male ratio ranges from 0.51 to 1.58 for UC studies and 0.34 to 1.65 for CD studies, suggesting that the diagnosis of IBD is not gender-specific. Pediatric IBD (patients <18 years of age) comprises ~20–25% of all IBD patients and about 5% of all IBD patients are <10 years of age. Very early onset IBD (VEOIBD) has been defined as IBD that occurs in children <6 years of age and infantile IBD in children <2 years of age. VEOIBD and infantile IBD mainly affect the colon, are resistant to standard medications, and patients often have a strong family history of IBD, with at least one first-degree related affected. Twenty-five percent of patients have an underlying immunodeficiency. In some cases, infantile IBD or VEOIBD can be caused by a number of rare, single genetic mutations.

The greatest incidence of IBD is among white and Jewish people, but the incidence of IBD in Hispanic and Asian people is increasing, as noted above. Urban areas have a higher prevalence of IBD than rural areas, and high socioeconomic classes have a higher prevalence than lower socioeconomic classes.

Epidemiologic studies have identified a number of potential environmental factors that are associated with disease risk **(Fig. 319-1)**. In Caucasian populations, smoking is an important risk factor in IBD with opposite effects on UC (odds ratio [OR] 0.58) and CD (OR 1.76), whereas in other ethnic groups with different genetic susceptibility, smoking may play a lesser role. There is a protective effect of previous appendectomy with confirmed appendicitis (risk reduction of 13–26%), particularly at a young age, on the development of UC across different geographical regions and populations. There is a modest association with the development of CD but this may be due to diagnostic bias. Oral contraceptive use is associated with the risk of CD with a reported hazard ratio as high as 2.82 among current users and 1.39 among past users. The association between oral contraceptive use and UC is limited to women with a history of smoking. There is an association between antibiotic use and the development of childhood IBD with children who received one or more dispensations of antibiotics during the first year of life having a 2.9-fold increase in the risk of developing IBD during childhood. Breastfeeding may also protect against the development of IBD. Infectious gastroenteritis with pathogens (e.g., *Salmonella*, *Shigella*, *Campylobacter* spp., *Clostridium difficile*) increases IBD risk by two- to threefold. Diets high in animal protein, sugars, sweets, oils, fish and shellfish, and dietary fat, especially ω-6 fatty acids, and low in ω-3 fatty acids have been implicated in increasing the risk of IBD. A protective effect of vitamin D on the risk of CD has been reported.

FIGURE 319-1

Pathogenesis of inflammatory bowel disease (IBD). In IBD, the tridirectional relationship between the commensal flora (microbiota), intestinal epithelial cells (IECs), and mucosal immune system is dysregulated, leading to chronic inflammation. Each of these three factors is affected by genetic and environmental factors that determine risk for the disease. NSAIDs, nonsteroidal anti-inflammatory drugs. *(Adapted from A Kaser et al: Annu Rev Immunol 28:573, 2010.)*

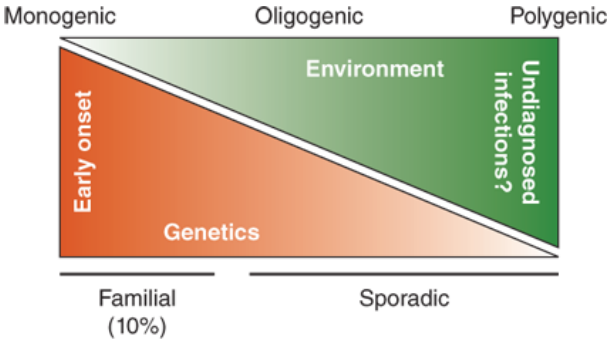


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IBD is a familial disease in 5–10% of patients (Fig. 319-2). Some of these patients may exhibit early-onset disease during the first decade of life and, in CD, a concordance of anatomic site and clinical type within families. In the remainder of patients, IBD is observed in the absence of a family history (i.e., sporadic disease). If a patient has IBD, the lifetime risk that a first-degree relative will be affected is ~10%. If two parents have IBD, each child has a 36% chance of being affected. In twin studies, 38–58% of monozygotic twins are concordant for CD and 6–18% are concordant for UC, whereas 4% of dizygotic twins are concordant for CD and 0–2% are concordant for UC in Swedish and Danish cohorts. The risks of developing IBD are higher in first-degree relatives of Jewish versus non-Jewish patients: 7.8% versus 5.2% for CD and 4.5% versus 1.6% for UC.

FIGURE 319-2

A model for the syndromic nature of inflammatory bowel disease. Genetic and environmental factors variably influence the development and phenotypic manifestations of IBD. At the one extreme, IBD is exemplified as a simple Mendelian disorder as observed in "early-onset IBD" due to single gene defects such as *IL10*, *IL10RA*, and *IL10RB*; and at the other extreme, it may be exemplified by as yet to be described emerging infectious diseases. *(Adapted from A Kaser et al: Dig Dis 28:395, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GLOBAL CONSIDERATIONS: IBD PHENOTYPES

There are racial differences in IBD location and behavior that may reflect underlying genetic variations and have important implications for diagnosis and management of disease. African Americans and Hispanics tend to have an ileocolonic CD distribution. Data from East Asia have observed that ileocolonic CD is the most common CD phenotype (50.5–71%) and perianal disease is more common in East Asian patients (30.3–58.8%) than Caucasians (25.1–29.6%). Pancolonic disease is more common than left-sided colitis or proctitis among African Americans, Hispanics, and Asian patients with UC. Older Asian patients with UC (age >60) tend to have a more aggressive disease course. Among African Americans, joint involvement is the predominant extra intestinal manifestation (EIM) reported and ranges from 15.7 to 29.6%. Ocular involvement is also common in African Americans and ranges from 7.1 to 13%. Dermatologic manifestations are the most common EIM reported in Hispanics (10–13%). There are few data on all aspects of disease in Hispanics, on the incidence and prevalence of IBD in African Americans, and in Asians with IBD outside Asia. These ethnic variations implicate the importance of different genetic and/or environmental factors in the pathogenesis of this disorder.

ETIOLOGY AND PATHOGENESIS

Under physiologic conditions, homeostasis normally exists between the commensal microbiota, epithelial cells that line the interior of the intestines (intestinal epithelial cells [IECs]) and immune cells within the tissues (Fig. 319-1). A consensus hypothesis is that each of these three major host compartments that function together as an integrated “supraorganism” (microbiota, IECs, and immune cells) are affected by specific environmental (e.g., smoking, antibiotics, enteropathogens) and genetic factors that, in a susceptible host, cumulatively and interactively disrupt homeostasis during the course of one’s life, which in so doing culminates in a chronic state of dysregulated inflammation; that is IBD. Although chronic activation of the mucosal immune system may represent an appropriate response to an infectious agent, a search for such an agent has thus far been unrewarding in IBD. As such, IBD is currently considered an inappropriate immune response to the endogenous (autochthonous) commensal microbiota within the intestines, with or without some component of autoimmunity. Importantly, the normal, uninflamed intestines contain a large number of immune cells that are in a unique state of activation, in which the gut is restrained from full immunologic responses to the commensal microbiota and dietary antigens by very powerful regulatory pathways that function within the immune system (e.g., T-regulatory cells that express the FoxP3 transcription factor and suppress inflammation). During the course of infections or other environmental stimuli in the normal host, full activation of the lymphoid tissues in the intestines occurs but is rapidly superseded by dampening of the immune response and tissue repair. In IBD such processes may not be regulated normally.

GENETIC CONSIDERATIONS

The genetic underpinning of IBD is known from its occurrence in the context of several genetic syndromes and the development of severe, refractory IBD in early life in the setting of single gene defects that affect the immune system **(Table 319-2)**. These include mutations in genes encoding, for example, interleukin-10 (IL-10), the IL-10 receptor (IL-10R), cytotoxic T-lymphocyte associated protein-4 (*CTLA4*), neutrophil cytosolic factor 2 protein (*NCF2*), X-linked inhibitor of apoptosis protein (*XIAP*), lipopolysaccharide responsive and beige-like anchor protein (*LRBA*), or tetratricopeptide repeat domain 7A protein (*TTC7*), among many other genes that are involved in host-commensal interactions. In addition, IBD has a familial origin in at least 10% of afflicted individuals (Fig. 319-2). In the majority of patients, IBD is considered to be a polygenic disorder that gives rise to multiple clinical subgroups within UC and CD. A variety of genetic approaches including candidate gene studies, linkage analysis, and genome-wide association studies (GWAS) that focus on the identification of disease-associated, single-nucleotide polymorphisms (SNPs) within the human genome and, more recently, whole-genome sequencing have elucidated many of the genetic factors that affect risk for these diseases. GWAS have, to date, identified ~200 genetic loci with two-thirds of these loci observed to be associated with both disease phenotypes with the remainder being specific for either CD or UC **(Table 319-3)**. These genetic similarities account for the overlapping immunopathogenesis and consequently epidemiologic observations of both diseases in the same families and similarities in response to therapies. Because the specific causal variants for each identified gene or locus are mostly unknown as most risk loci are contained with regulatory regions of the associated genes, it is not clear whether the similarities in the genetic risk factors associated with CD and UC that are observed are shared at structural or functional levels. The risk conferred by each identified gene or locus is unequal and generally small, such that only ~20% of the genetic variance is considered to be explained by the current genetic information. Further, many of the genetic risk factors identified are also observed to be associated with risk for other immune-mediated diseases, suggesting that related immunogenetic pathways are involved in the pathogenesis of multiple different disorders accounting for the common responsiveness to similar types of biologic therapies (e.g., anti–tumor necrosis factor [TNF] therapies) and possibly the simultaneous occurrence of these disorders. The diseases and the genetic risk factors that are shared with IBD include rheumatoid arthritis (*TNFAIP3*), psoriasis (*IL23R*, *IL12B*), ankylosing spondylitis (*IL23R*), type 1 diabetes mellitus (*IL10*, *PTPN2*), asthma (*ORMDL3*), and systemic lupus erythematosus (*TNFAIP3*, *IL10*) among others.

TABLE 319-2

Primary Genetic Disorders Associated with IBD

Name	Genetic Association	Phenotype
Turner's syndrome	Loss of part or all of X chromosome	Associated with UC and colonic CD
Hermansky-Pudlak	Autosomal recessive chromosome 10q23	Granulomatous colitis, oculocutaneous albinism, platelet dysfunction, pulmonary fibrosis
Wiskott-Aldrich syndrome (WAS)	X-linked recessive disorder, loss of WAS protein function	Colitis, immunodeficiency, severely dysfunctional platelets, and thrombocytopenia
Glycogen storage disease	Deficiency of the glucose-6-phosphate transport protein type B1	Granulomatous colitis, presents in infancy with hypoglycemia, growth failure, hepatomegaly, and neutropenia
Immune dysregulation polyendocrinopathy, enteropathy X-linked (IPEX)	Loss of FoxP3 transcription factor and T regulatory cell function	UC-like autoimmune enteropathy, with endocrinopathy (neonatal type 1 diabetes or thyroiditis), dermatitis
Early-onset IBD	Deficient IL-10 and IL-10 receptor function	Severe, refractory IBD in early life

Abbreviations: CD, Crohn's disease; IBD, inflammatory bowel disease; IL, interleukin; UC, ulcerative colitis.

TABLE 319-3

Some Genetic Loci Associated with Crohn's Disease and/or Ulcerative Colitis

Chromosome	Putative Gene	Gene Name	Protein Function	CD	UC
Unfolded Protein Response, Autophagy and Metabolism					
2q37	*ATG16L1*	ATG16 autophagy related 16-like 1	Autophagy	+	
5q31	*SLC22A5*	Solute carrier family 22, member 5	β carnitine transporter	+	
5q33	*IRGM*	Immunity-related GTPase family, M	Autophagy	+	
7p21	*AGR2*	Anterior gradient 2	Unfolded protein response	+	+
12q12	*LRRK2*	Leucine-rich repeat kinase 2	Autophagy	+	
13q14	*C13orf1*	FAMIN/ LACC1	Immunometabolic regulator	+	
17q21	*ORMDL3*	Orosomucoid related member 1-like 3	Unfolded protein response and lipid synthesis	+	+
22q12	*XBP1*	X-box binding protein 1	Unfolded protein response	+	+
Innate Immunity					
1q23	*ITLN1*	Intelectin 1	Bacterial binding	+	
16q12	*NOD2*	Nucleotide-binding oligomerization domain containing 2	Bacterial sensing and autophagy activation	+	
Adaptive Immunity					
1p31	*IL23R*	Interleukin 23 receptor	Th17 cell stimulation	+	+
1q32	*IL10*	Interleukin 10	Treg-associated cytokine		+
5q33	*IL12B*	Interleukin 12B	IL-12 p40 chain of IL-12/IL-23	+	+
18p11	*PTPN2*	Protein tyrosine phosphatase, nonreceptor type 2	T cell regulation	+	
Inflammation					
3p21	*MST1*	Macrophage stimulating 1	Macrophage activation	+	+
5p13	*PTGER4*	Prostaglandin E receptor 4	PGE_2 receptor	+	+
6q23	*TNFAIP3*	Tumor necrosis factor, alpha-induced protein 3 (A20)	Toll-like receptor regulation	+	
6q27	*CCR6*	Chemokine (C-C motif) receptor 6	Dendritic cell migration	+	
9p24	*JAK2*	Janus kinase 2	IL-6R and IL-23R signaling	+	+
17q21	*STAT3*	Signal transducer and activator of transcription 3	IL-6R, IL-23R, and IL-10R signaling	+	+

Abbreviations: CD, Crohn's disease; GTPase, guanosine triphosphatase; IL, interleukin; PGE_2, prostaglandin E_2; UC, ulcerative colitis.

Source: Adapted from A Kaser et al: Ann Rev Immunol 28:573, 2010; B Khor et al: Nature 474:307, 2011; and L Jostins et al: Nature 491:119, 2012.

The genetic factors defined to date that are recognized to mediate risk for IBD have highlighted the importance of several common mechanisms of disease (Table 319-3). These include the following: those genes that are associated with fundamental cell biologic processes such as endoplasmic reticulum (ER) and metabolic stress (e.g., *XBP1*, *ORMDL3*, *OCTN*), which serve to regulate the secretory activity of cells involved in responses to the commensal microbiota such as Paneth and goblet cells and the manner in which intestinal cells respond to the metabolic products of bacteria; those associated with innate immunity and autophagy (e.g., *NOD2, ATG16L1*, *IRGM, JAK2, STAT3, C13orf31*) that function in innate immune cells (both parenchymal and hematopoietic) to respond to and effectively clear bacteria, mycobacteria, and viruses; those that are associated with the regulation of adaptive immunity (e.g., *IL23R, IL12B*, *IL10*, *PTPN2*), which regulate the balance between inflammatory and anti-inflammatory (regulatory) cytokines; and, finally, those that are involved in the development and resolution of inflammation (e.g., *MST1*, *CCR6, TNFAIP3, PTGER4*) and ultimately leukocyte recruitment and inflammatory mediator production. Some of these loci are associated with specific subtypes of disease such as the association between *NOD2* polymorphisms and fibrostenosing CD or *ATG16L1* and fistulizing disease, especially within the ileum. However, the clinical utility of these genetic risk factors for the diagnosis or determination of prognosis and therapeutic responses remains to be defined.

COMMENSAL MICROBIOTA AND IBD

The endogenous commensal microbiota within the intestines plays a central role in the pathogenesis of IBD. Humans are born sterile and acquire their commensal microbiota initially from the mother during egress through the birth canal and subsequently from environmental sources. A stable configuration of up to 1000 species of bacteria that achieves a biomass of ~10^{12} colony-forming units per gram of feces is achieved by 3 years of age, which likely persists into adult life, with each individual human possessing a unique combination of species. In addition, the intestines contain other microbial life forms including archae, viruses, and protists. The microbiota is thus considered as a critical and sustaining component of the human organism. The establishment and maintenance of the intestinal microbiota composition and function is under the control of host (e.g., immune and epithelial responses), environmental (e.g., diet and antibiotics), and likely genetic (e.g., *NOD2*) factors (Fig. 319-1). In turn, the microbiota, through its structural components and metabolic activity, has major influences on the epithelial and immune function of the host, which, through epigenetic effects, may have durable consequences. During early life when the commensal microbiota is being established, these microbial effects on the host may be particularly important in determining later life risk for IBD. Specific components of the microbiota can promote or protect from disease. The commensal microbiota in patients with both UC and CD is demonstrably different from nonafflicted individuals, a state of dysbiosis, suggesting the presence of microorganisms that drive disease (e.g., Proteobacteria such as enteroinvasive and adherent *Escherichia coli*) and to which the immune response is directed and/or the loss of microorganisms that hinder inflammation (e.g., Firmicutes such as *Faecalibacterium prausnitzii*). Many of the changes in the commensal microbiota occur as a consequence of the inflammation. In addition, agents that alter the intestinal microbiota such as metronidazole, ciprofloxacin, and elemental diets, may improve CD. CD may also respond to fecal diversion, demonstrating the ability of luminal contents to exacerbate disease.

DEFECTIVE IMMUNE REGULATION IN IBD

The mucosal immune system does not normally elicit an inflammatory immune response to luminal contents due to oral (mucosal) tolerance. Administration of soluble antigens orally, rather than subcutaneously or intramuscularly, leads to antigen-specific control of the response and the host's ability to tolerate the antigen. Multiple mechanisms are involved in the induction of oral tolerance and include deletion or anergy of antigen-reactive T cells or induction of CD4+ T cells that suppress gut inflammation (e.g., T-regulatory cells expressing the FoxP3 transcription factor) that secrete anti-inflammatory cytokines such as IL-10, IL-35, and transforming growth factor β (TGF-β). Oral tolerance may be responsible for the lack of immune responsiveness to dietary antigens and the commensal microbiota in the intestinal lumen. In IBD this suppression of inflammation is altered, leading to uncontrolled inflammation. The mechanisms of this regulated immune suppression are incompletely known.

Gene knockout ($^{-/-}$) or transgenic (Tg) mouse models of IBD, which include those that are directed at genes demonstrated to be associated with risk for the human disease, have revealed that deleting specific cytokines (e.g., IL-2, IL-10, TGF-β) or their receptors, deleting molecules associated with T-cell antigen recognition (e.g., T-cell antigen receptors), or interfering with IEC barrier function and the regulation of responses to commensal bacteria (e.g., XBP1, mucus glycoproteins, or nuclear factor-κB [NF-κB]) leads to spontaneous colitis or enteritis. In the majority of circumstances, intestinal inflammation in these animal models requires the presence of the commensal microbiota. However, in some cases, activation of certain elements of the intestinal immune system may be exacerbated by the absence of bacteria resulting in severe colitis emphasizing the presence of protective properties that are also contained within the commensal microbiota. Thus, a variety of specific alterations in either the microbiota or host can lead to uncontrolled immune activation and inflammation directed at the intestines in mice. How these relate to human IBD remains to be defined, but they are consistent with inappropriate responses of the genetically susceptible host to the commensal microbiota.

THE INFLAMMATORY CASCADE IN IBD

In both UC and CD, inflammation thus likely emerges from the genetic predisposition of the host in the context of yet-to-be defined environmental factors. Once initiated in IBD by abnormal innate immune sensing of bacteria by parenchymal cells (e.g., IECs) and hematopoietic cells (e.g., dendritic cells), the immune inflammatory response is perpetuated by T-cell activation when coupled together with inadequate regulatory pathways. A sequential cascade of inflammatory mediators extends the response making each step a potential target for therapy. Inflammatory cytokines from innate immune cells such as IL-1, IL-6, and TNF have diverse effects on tissues. They promote fibrogenesis, collagen production, activation of tissue metalloproteinases, and the production of other inflammatory mediators; they also activate the coagulation cascade in local blood vessels (e.g., increased production of von

Willebrand's factor). These cytokines are normally produced in response to infection but are usually turned off or inhibited at the appropriate time to limit tissue damage. In IBD their activity is not regulated, resulting in an imbalance between the pro-inflammatory and anti-inflammatory mediators. Some cytokines activate other inflammatory cells (macrophages and B cells), and others act indirectly to recruit other lymphocytes, inflammatory leukocytes, and mononuclear cells from the bloodstream into the gut through interactions between homing receptors on leukocytes (e.g., α4β7 integrin) and addressins on vascular endothelium (e.g., MadCAM1). CD4+ T-helper (T_H) cells that promote inflammation are of three major types, all of which may be associated with colitis in animal models and perhaps humans: T_H1 cells (secrete interferon [IFN] γ), T_H2 cells (secrete IL-4, IL-5, IL-13), and T_H17 cells (secrete IL-17, IL-21). T_H17 cells may however also provide protective functions. Innate immune-like cells that lack T-cell receptors are also present in intestines, polarize to the same functional fates and may similarly participate in IBD. T_H1 cells induce transmural granulomatous inflammation that resembles CD; T_H2 cells, and related natural killer T cells that secrete IL-4, IL-5, and IL-13, induce superficial mucosal inflammation resembling UC in animal models; and T_H17 cells may be responsible for neutrophilic recruitment. However, neutralization of the cytokines produced by these cells, such as IFN-γ or IL-17, has yet to show efficacy in therapeutic trials. Each of these T-cell subsets cross-regulate each other. The T_H1 cytokine pathway is initiated by IL-12, a key cytokine in the pathogenesis of experimental models of mucosal inflammation. IL-4 and IL-23, together with IL-6 and TGF-β, induce T_H2 and T_H17 cells, respectively, and IL-23 inhibits the suppressive function of regulatory T cells. Activated macrophages secrete TNF and IL-6.

These characteristics of the immune response in IBD explain the beneficial therapeutic effects of antibodies to block pro-inflammatory cytokines or the signaling by their receptors (e.g., anti-TNF, anti-IL-12, anti-IL-23, anti-IL-6, or Janus kinase [JAK] inhibitors) or molecules associated with leukocyte recruitment (e.g., anti-α4β7). They also highlight the potential usefulness of cytokines that inhibit inflammation and promote regulatory T cells or promote intestinal barrier (e.g., IL-10) in the treatment of IBD. Therapies such as the 5-aminosalicylic acid (5-ASA) compounds and glucocorticoids are also potent inhibitors of these inflammatory mediators through inhibition of transcription factors such as NF-κB that regulate their expression.

PATHOLOGY

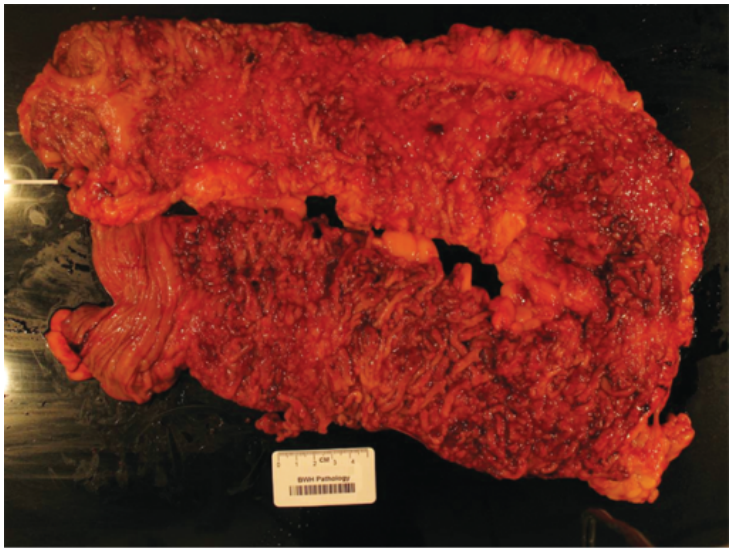
ULCERATIVE COLITIS: MACROSCOPIC FEATURES

UC is a mucosal disease that usually involves the rectum and extends proximally to involve all or part of the colon. About 40–50% of patients have disease limited to the rectum and rectosigmoid, 30–40% have disease extending beyond the sigmoid but not involving the whole colon, and 20% have a total colitis. Proximal spread occurs in continuity without areas of uninvolved mucosa. When the whole colon is involved, the inflammation extends 2–3 cm into the terminal ileum in 10–20% of patients. The endoscopic changes of *backwash ileitis* are superficial and mild and are of little clinical significance. Although variations in macroscopic activity may suggest skip areas, biopsies from normal-appearing mucosa are usually abnormal. Thus, it is important to obtain multiple biopsies from apparently uninvolved mucosa, whether proximal or distal, during endoscopy. One caveat is that effective medical therapy can change the appearance of the mucosa such that either skip areas or the entire colon can be microscopically normal.

With mild inflammation, the mucosa is erythematous and has a fine granular surface that resembles sandpaper. In more severe disease, the mucosa is hemorrhagic, edematous, and ulcerated **(Fig. 319-3)**. In long-standing disease, inflammatory polyps (pseudopolyps) may be present as a result of epithelial regeneration. The mucosa may appear normal in remission, but in patients with many years of disease it appears atrophic and featureless, and the entire colon becomes narrowed and shortened. Patients with fulminant disease can develop a toxic colitis or megacolon where the bowel wall thins and the mucosa is severely ulcerated; this may lead to perforation.

FIGURE 319-3

Ulcerative colitis. Diffuse (nonsegmental) mucosal disease, with broad areas of ulceration. The bowel wall is not thickened, and there is no cobblestoning. *(Courtesy of Dr. R. Odze, Division of Gastrointestinal Pathology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*



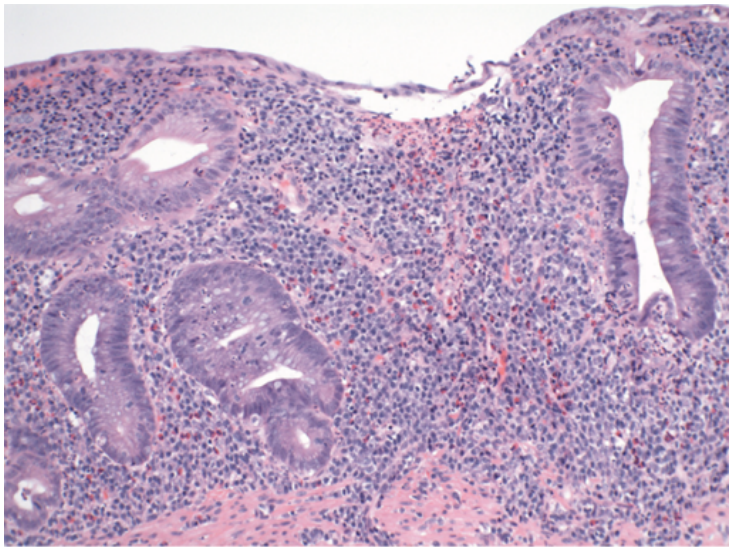
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ULCERATIVE COLITIS: MICROSCOPIC FEATURES

Histologic findings correlate well with the endoscopic appearance and clinical course of UC. The process is limited to the mucosa and superficial submucosa, with deeper layers unaffected except in fulminant disease. In UC, two major histologic features suggest chronicity and help distinguish it from infectious or acute self-limited colitis. First, the crypt architecture of the colon is distorted; crypts may be bifid and reduced in number, often with a gap between the crypt bases and the muscularis mucosae. Second, some patients have basal plasma cells and multiple basal lymphoid aggregates. Mucosal vascular congestion, with edema and focal hemorrhage, and an inflammatory cell infiltrate of neutrophils, lymphocytes, plasma cells, and macrophages may be present. The neutrophils invade the epithelium, usually in the crypts, giving rise to cryptitis and, ultimately, to crypt abscesses **(Fig. 319-4)**. Ileal changes in patients with backwash ileitis include villous atrophy and crypt regeneration with increased inflammation, increased neutrophil and mononuclear inflammation in the lamina propria, and patchy cryptitis and crypt abscesses.

FIGURE 319-4

Medium-power view of colonic mucosa in ulcerative colitis showing diffuse mixed inflammation, basal lymphoplasmacytosis, crypt atrophy and irregularity, and superficial erosion. These features are typical of chronic active ulcerative colitis. *(Courtesy of Dr. R. Odze, Division of Gastrointestinal Pathology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CROHN'S DISEASE: MACROSCOPIC FEATURES

CD can affect any part of the gastrointestinal (GI) tract from the mouth to the anus. Some 30–40% of patients have small bowel disease alone, 40–55% have disease involving both the small and large intestines, and 15–25% have colitis alone. In the 75% of patients with small intestinal disease, the terminal

ileum is involved in 90%. Unlike UC, which almost always involves the rectum, the rectum is often spared in CD. CD is segmental with skip areas in the midst of diseased intestine (Fig. 319-5). Perirectal fistulas, fissures, abscesses, and anal stenosis are present in one-third of patients with CD, particularly those with colonic involvement. Rarely, CD may also involve the liver and the pancreas.

FIGURE 319-5

Crohn's disease of the colon showing thickening of the wall, with stenosis, linear serpiginous ulcers and cobblestoning of the mucosa. *(Courtesy of Dr. R Odze, Division of Gastrointestinal Pathology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Unlike UC, CD is a transmural process. Endoscopically, aphthous or small superficial ulcerations characterize mild disease; in more active disease, stellate ulcerations fuse longitudinally and transversely to demarcate islands of mucosa that frequently are histologically normal. This "cobblestone" appearance is characteristic of CD, both endoscopically and by barium radiography. As in UC, pseudopolyps can form in CD.

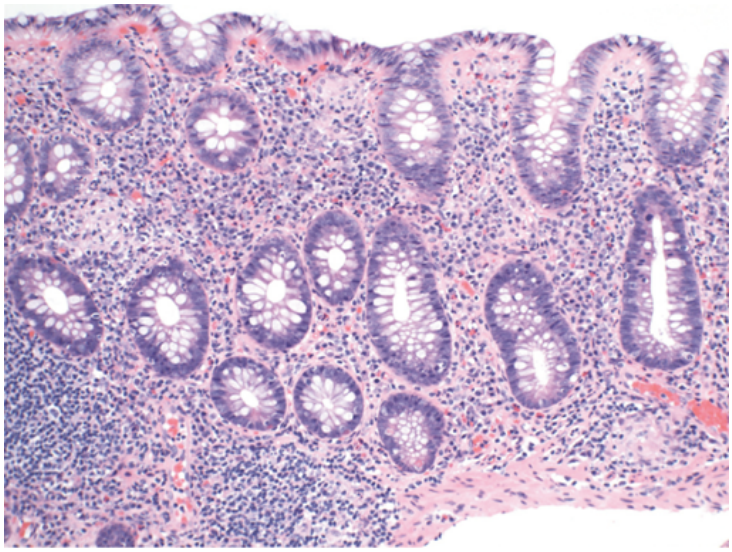
Active CD is characterized by focal inflammation and formation of fistula tracts, which resolve by fibrosis and stricturing of the bowel. The bowel wall thickens and becomes narrowed and fibrotic, leading to chronic, recurrent bowel obstructions. Projections of thickened mesentery encase the bowel ("creeping fat"), and serosal and mesenteric inflammation promotes adhesions and fistula formation.

CROHN'S DISEASE: MICROSCOPIC FEATURES

The earliest lesions are aphthoid ulcerations and focal crypt abscesses with loose aggregations of macrophages, which form noncaseating granulomas in all layers of the bowel wall (Fig. 319-6). Granulomas can be seen in lymph nodes, mesentery, peritoneum, liver, and pancreas. Granulomas are a characteristic feature of CD. They are less commonly found on mucosal biopsies than on surgical resection specimens. Other histologic features of CD include submucosal or subserosal lymphoid aggregates, particularly away from areas of ulceration, gross and microscopic skip areas, and transmural inflammation that is accompanied by fissures that penetrate deeply into the bowel wall and sometimes form fistulous tracts or local abscesses.

FIGURE 319-6

Medium-power view of Crohn's colitis showing mixed acute and chronic inflammation, crypt atrophy, and multiple small epithelioid granulomas in the mucosa. *(Courtesy of Dr. R Odze, Division of Gastrointestinal Pathology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CLINICAL PRESENTATION

ULCERATIVE COLITIS

Signs and Symptoms

The major symptoms of UC are diarrhea, rectal bleeding, tenesmus, passage of mucus, and crampy abdominal pain. The severity of symptoms correlates with the extent of disease. Although UC can present acutely, symptoms usually have been present for weeks to months. Occasionally, diarrhea and bleeding are so intermittent and mild that the patient does not seek medical attention.

Patients with proctitis usually pass fresh blood or blood-stained mucus, either mixed with stool or streaked onto the surface of a normal or hard stool. They also have tenesmus, or urgency with a feeling of incomplete evacuation, but rarely have abdominal pain. With proctitis or proctosigmoiditis, proximal transit slows, which may account for the constipation commonly seen in patients with distal disease.

When the disease extends beyond the rectum, blood is usually mixed with stool or grossly bloody diarrhea may be noted. Colonic motility is altered by inflammation with rapid transit through the inflamed intestine. When the disease is severe, patients pass a liquid stool containing blood, pus, and fecal matter. Diarrhea is often nocturnal and/or postprandial. Although severe pain is not a prominent symptom, some patients with active disease may experience lower abdominal discomfort or mild central abdominal cramping. Severe cramping and abdominal pain can occur with severe attacks of the disease. Other symptoms in moderate to severe disease include anorexia, nausea, vomiting, fever, and weight loss.

Physical signs of proctitis include a tender anal canal and blood on rectal examination. With more extensive disease, patients have tenderness to palpation directly over the colon. Patients with a toxic colitis have severe pain and bleeding, and those with megacolon have hepatic tympany. Both may have signs of peritonitis if a perforation has occurred. The classification of disease activity is shown in Table 319-4.

TABLE 319-4

Ulcerative Colitis: Disease Presentation

	Mild	Moderate	Severe
Bowel movements	<4 per day	4–6 per day	>6 per day
Blood in stool	Small	Moderate	Severe
Fever	None	<37.5°C mean (<99.5°F)	>37.5°C mean (>99.5°F)
Tachycardia	None	<90 mean pulse	>90 mean pulse
Anemia	Mild	>75% of a normal hemoglobin	≤75% of a normal hemoglobin
Sedimentation rate	<30 mm		>30 mm
Endoscopic appearance	Erythema, decreased vascular pattern, fine granularity	Marked erythema, coarse granularity, absent vascular markings, contact bleeding, no ulcerations	Spontaneous bleeding, ulcerations

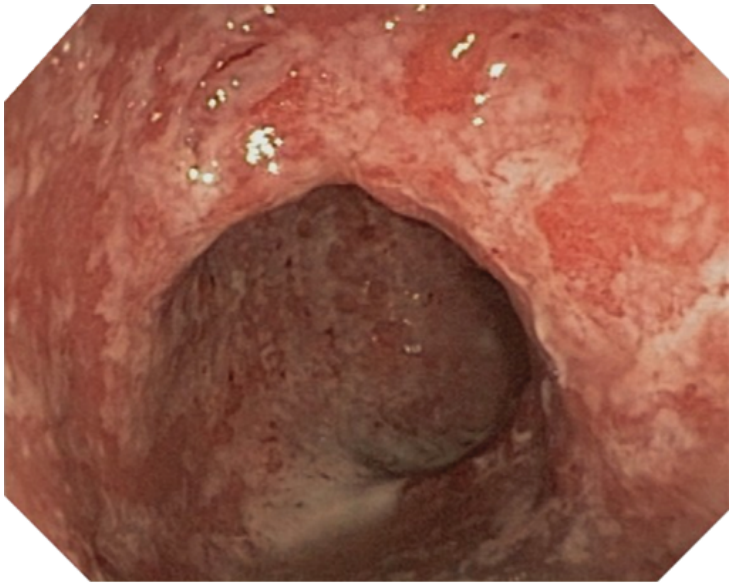
Laboratory, Endoscopic, and Radiographic Features

Active disease can be associated with a rise in acute-phase reactants (C-reactive protein [CRP]), platelet count, and erythrocyte sedimentation rate (ESR), and a decrease in hemoglobin. Fecal lactoferrin, a glycoprotein present in activated neutrophils, is a highly sensitive and specific marker for detecting intestinal inflammation. Fecal calprotectin is present in neutrophils and monocytes and levels correlate well with histologic inflammation, predict relapses, and detect pouchitis. Both fecal lactoferrin and calprotectin are becoming an integral part of IBD management and are used frequently to rule out active inflammation versus symptoms of irritable bowel or bacterial overgrowth. In severely ill patients, the serum albumin level will fall rather quickly. Leukocytosis may be present but is not a specific indicator of disease activity. Proctitis or proctosigmoiditis rarely causes a rise in CRP. Diagnosis relies on the patient's history; clinical symptoms; negative stool examination for bacteria, *C. difficile* toxin, and ova and parasites; sigmoidoscopic appearance **(see Fig. 315-4A)**; and histology of rectal or colonic biopsy specimens.

Sigmoidoscopy is used to assess disease activity and is usually performed before treatment. If the patient is not having an acute flare, colonoscopy is used to assess disease extent and activity **(Fig. 319-7)**. Endoscopically mild disease is characterized by erythema, decreased vascular pattern, and mild friability. Moderate disease is characterized by marked erythema, absent vascular pattern, friability and erosions, and severe disease by spontaneous bleeding and ulcerations. Histologic features change more slowly than clinical features but can also be used to grade disease activity.

FIGURE 319-7

Colonoscopy with acute ulcerative colitis: severe colon inflammation with erythema, friability, and exudates. *(Courtesy of Dr. M. Hamilton, Gastroenterology Division, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Complications

Only 15% of patients with UC present initially with catastrophic illness. Massive hemorrhage occurs with severe attacks of disease in 1% of patients, and treatment for the disease usually stops the bleeding. However, if a patient requires 6–8 units of blood within 24–48 h, colectomy is indicated. *Toxic megacolon* is defined as a transverse or right colon with a diameter of >6 cm, with loss of haustration in patients with severe attacks of UC. It occurs in about 5% of attacks and can be triggered by electrolyte abnormalities and narcotics. About 50% of acute dilations will resolve with medical therapy alone, but urgent colectomy is required for those that do not improve. Perforation is the most dangerous of the local complications, and the physical signs of peritonitis may not be obvious, especially if the patient is receiving glucocorticoids. Although perforation is rare, the mortality rate for perforation complicating a toxic megacolon is about 15%. In addition, patients can develop a toxic colitis and such severe ulcerations that the bowel may perforate without first dilating.

Strictures occur in 5–10% of patients and are always a concern in UC because of the possibility of underlying neoplasia. Although benign strictures can form from the inflammation and fibrosis of UC, strictures that are impassable with the colonoscope should be presumed malignant until proven otherwise. A stricture that prevents passage of the colonoscope is an indication for surgery. UC patients occasionally develop anal fissures, perianal abscesses, or hemorrhoids, but the occurrence of extensive perianal lesions should suggest CD.

CROHN'S DISEASE

Signs and Symptoms

Although CD usually presents as acute or chronic bowel inflammation, the inflammatory process evolves toward one of two patterns of disease: a fibrostenotic obstructing pattern or a penetrating fistulous pattern, each with different treatments and prognoses. The site of disease influences the clinical manifestations.

ILEOCOLITIS

Because the most common site of inflammation is the terminal ileum, the usual presentation of ileocolitis is a chronic history of recurrent episodes of right lower quadrant pain and diarrhea. Sometimes the initial presentation mimics acute appendicitis with pronounced right lower quadrant pain, a palpable mass, fever, and leukocytosis. Pain is usually colicky; it precedes and is relieved by defecation. A low-grade fever is usually noted. High-spiking fever suggests intraabdominal abscess formation. Weight loss is common—typically 10–20% of body weight—and develops as a consequence of diarrhea, anorexia, and fear of eating.

An inflammatory mass may be palpated in the right lower quadrant of the abdomen. The mass is composed of inflamed bowel, induration of the mesentery, and enlarged abdominal lymph nodes. Extension of the mass can cause obstruction of the right ureter or bladder inflammation, manifested by dysuria and fever. The ''string sign'' on barium studies results from a severely narrowed loop of bowel, which makes the lumen resemble a frayed cotton string. It is caused by incomplete filling of the lumen as the result of edema, irritability, and spasms associated with inflammation and ulcerations. The sign may be seen in both nonstenotic and stenotic phases of the disease.

Bowel obstruction may take several forms. In the early stages of disease, bowel wall edema and spasm produce intermittent obstructive manifestations and increasing symptoms of postprandial pain. Over several years, persistent inflammation gradually progresses to fibrostenotic narrowing and stricture.

Diarrhea will decrease and be replaced by chronic bowel obstruction. Acute episodes of obstruction occur as well, precipitated by bowel inflammation and spasm or sometimes by impaction of undigested food or medication. These episodes usually resolve with intravenous fluids and gastric decompression.

Severe inflammation of the ileocecal region may lead to localized wall thinning, with microperforation and fistula formation to the adjacent bowel, the skin, or the urinary bladder, or to an abscess cavity in the mesentery. Enterovesical fistulas typically present as dysuria or recurrent bladder infections or, less commonly, as pneumaturia or fecaluria. Enterocutaneous fistulas follow tissue planes of least resistance, usually draining through abdominal surgical scars. Enterovaginal fistulas are rare and present as dyspareunia or as a feculent or foul-smelling, often painful vaginal discharge. They are unlikely to develop without a prior hysterectomy.

JEJUNOILEITIS

Extensive inflammatory disease is associated with a loss of digestive and absorptive surface, resulting in malabsorption and steatorrhea. Nutritional deficiencies can also result from poor intake and enteric losses of protein and other nutrients. Intestinal malabsorption can cause anemia, hypoalbuminemia, hypocalcemia, hypomagnesemia, coagulopathy, and hyperoxaluria with nephrolithiasis in patients with an intact colon. Many patients need to take oral and often intravenous iron. Vertebral fractures are caused by a combination of vitamin D deficiency, hypocalcemia, and prolonged glucocorticoid use. Pellagra from niacin deficiency can occur in extensive small-bowel disease, and malabsorption of vitamin B_{12} can lead to megaloblastic anemia and neurologic symptoms. Other important nutrients to measure and replete if low are folate and vitamins A, E, and K. Levels of minerals such as zinc, selenium, copper, and magnesium are often low in patients with extensive small-bowel inflammation or resections, and these should be repleted as well. Most patients should take a daily multivitamin, calcium, and vitamin D supplements.

Diarrhea is characteristic of active disease; its causes include (1) bacterial overgrowth in obstructive stasis or fistulization, (2) bile-acid malabsorption due to a diseased or resected terminal ileum, and (3) intestinal inflammation with decreased water absorption and increased secretion of electrolytes.

COLITIS AND PERIANAL DISEASE

Patients with colitis present with low-grade fevers, malaise, diarrhea, crampy abdominal pain, and sometimes hematochezia. Gross bleeding is not as common as in UC and appears in about one-half of patients with exclusively colonic disease. Only 1–2% exhibit massive bleeding. Pain is caused by passage of fecal material through narrowed and inflamed segments of the large bowel. Decreased rectal compliance is another cause for diarrhea in Crohn's colitis patients. Toxic megacolon is rare but may be seen with severe inflammation and short duration disease.

Stricturing can occur in the colon in 4–16% of patients and produce symptoms of bowel obstruction. If the endoscopist is unable to traverse a stricture in Crohn's colitis, surgical resection should be considered, especially if the patient has symptoms of chronic obstruction. Colonic disease may fistulize into the stomach or duodenum, causing feculent vomiting, or to the proximal or mid-small bowel, causing malabsorption by "short circuiting" and bacterial overgrowth. Ten percent of women with Crohn's colitis will develop a rectovaginal fistula.

Perianal disease affects about one-third of patients with Crohn's colitis and is manifested by incontinence, large hemorrhoidal tags, anal strictures, anorectal fistulae, and perirectal abscesses. Not all patients with perianal fistula will have endoscopic evidence of colonic inflammation.

GASTRODUODENAL DISEASE

Symptoms and signs of upper GI tract disease include nausea, vomiting, and epigastric pain. Patients usually have an *Helicobacter pylori*–negative gastritis. The second portion of the duodenum is more commonly involved than the bulb. Fistulas involving the stomach or duodenum arise from the small or large bowel and do not necessarily signify the presence of upper GI tract involvement. Patients with advanced gastroduodenal CD may develop a chronic gastric outlet obstruction. About 30% of children diagnosed with CD have esophagogastroduodenal involvement.

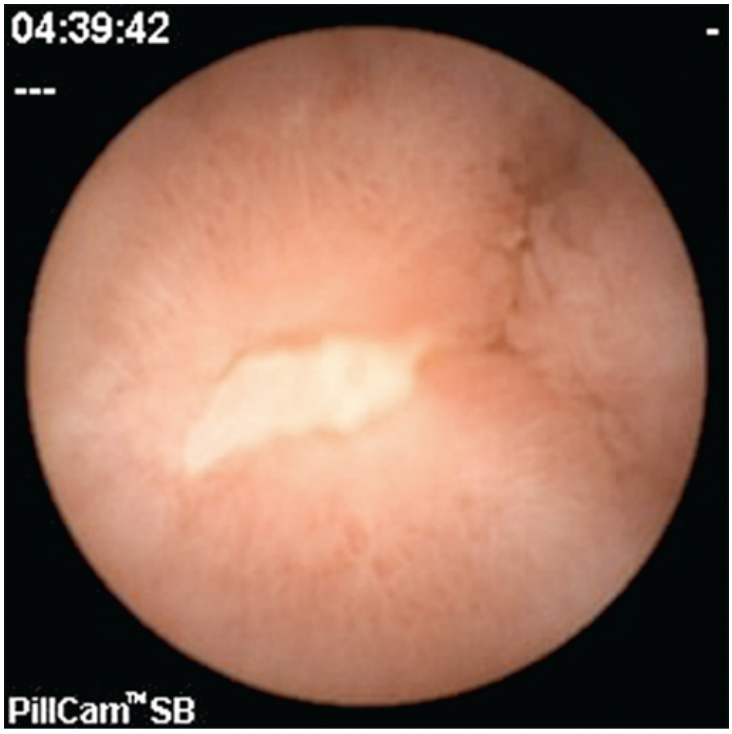
Laboratory, Endoscopic, and Radiographic Features

Laboratory abnormalities include elevated ESR and CRP. In more severe disease, findings include hypoalbuminemia, anemia, and leukocytosis. Fecal calprotectin and lactoferrin levels have been used to distinguish IBD from irritable bowel syndrome (IBS), assess whether CD is active, and to detect postoperative recurrence of CD.

Endoscopic features of CD include rectal sparing, aphthous ulcerations, fistulas, and skip lesions. Colonoscopy allows examination and biopsy of mass lesions or strictures and biopsy of the terminal ileum. Upper endoscopy is useful in diagnosing gastroduodenal involvement in patients with upper tract symptoms. Ileal or colonic strictures may be dilated with balloons introduced through the colonoscope. Strictures ≤4 cm in length and those at anastomotic sites respond better to endoscopic dilation. The perforation rate is as high as 10%. Most endoscopists dilate only fibrotic strictures and not those associated with active inflammation. Wireless capsule endoscopy (WCE) allows direct visualization of the entire small-bowel mucosa **(Fig. 319-8)**. The diagnostic yield of detecting lesions suggestive of active CD is higher with WCE than CT or magnetic resonance (MR) enterography or small-bowel series. WCE cannot be used in the setting of a small-bowel stricture. Capsule retention occurs in <1% of patients with suspected CD, but retention rates of 4–6% are seen in patients with established CD. It is helpful to give the patient with CD a patency capsule, which is made of barium and starts to dissolve 30 h after ingestion. An abdominal x-ray can be taken at around 30 h after ingestion to see if the capsule is still present in the small bowel, which would indicate a stricture.

FIGURE 319-8

Wireless capsule endoscopy image in a patient with Crohn's disease of the ileum shows ulcerations and narrowing of the intestinal lumen. *(Courtesy of Dr. S. Reddy, Gastroenterology Division, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In CD, early radiographic findings in the small bowel include thickened folds and aphthous ulcerations. “Cobblestoning” from longitudinal and transverse ulcerations most frequently involves the small bowel. In more advanced disease, strictures, fistulas, inflammatory masses, and abscesses may be detected. The earliest macroscopic findings of colonic CD are aphthous ulcers. These small ulcers are often multiple and separated by normal intervening mucosa. As the disease progresses, aphthous ulcers become enlarged, deeper, and occasionally connected to one another, forming longitudinal stellate, serpiginous, and linear ulcers **(see Fig. 315-4B)**.

The transmural inflammation of CD leads to decreased luminal diameter and limited distensibility. As ulcers progress deeper, they can lead to fistula formation. The segmental nature of CD results in wide gaps of normal or dilated bowel between involved segments.

Although CT enterography (CTE), MR enterography (MRE), and small-bowel follow-through (SBFT) have been shown to be equally accurate in the identification of active small-bowel inflammation, CTE and MRE have been shown to be superior to SBFT in the detection of extraluminal complications, including fistulas, sinus tracts, and abscesses. MRI is thought to offer superior soft tissue contrast and has the added advantage of avoiding radiation exposure changes **(Figs. 319-9 and 319-10)**. The lack of ionizing radiation is particularly appealing in younger patients and when monitoring response to therapy where serial images will be obtained. Ultrasound is becoming increasingly more popular, especially in Europe, for measuring CD extent and activity. Pelvic MRI is superior to CT for demonstrating pelvic lesions such as ischiorectal abscesses and perianal fistulae **(Fig. 319-11)**.

FIGURE 319-9

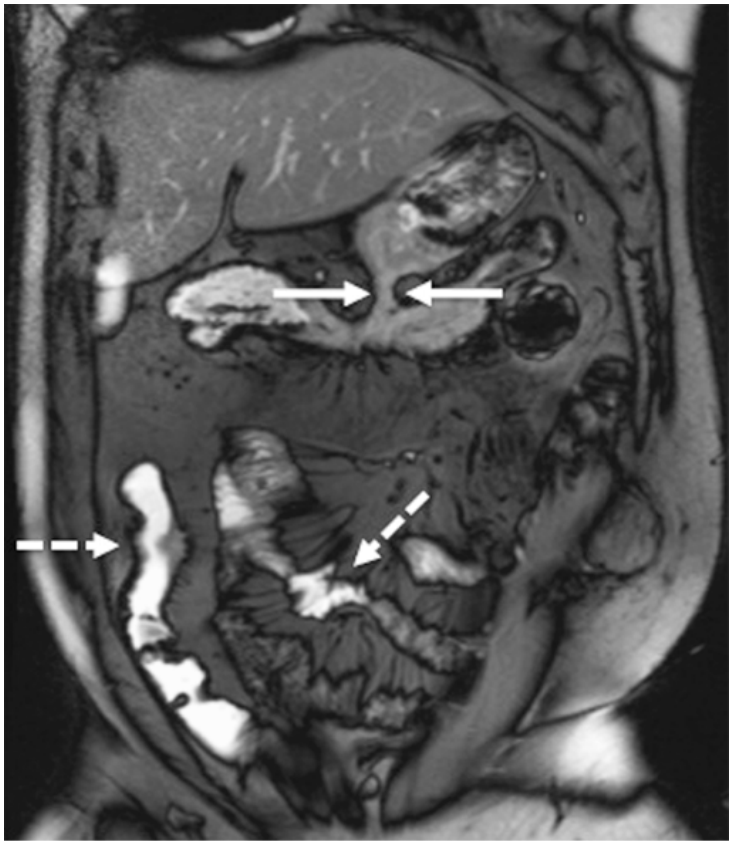
A coronal magnetic resonance image was obtained using a half Fourier single-shot T2-weighted acquisition with fat saturation in a 27-year-old pregnant (23 weeks’ gestation) woman. The patient had Crohn’s disease and was maintained on 6-mercaptopurine and prednisone. She presented with abdominal pain, distension, vomiting, and small-bowel obstruction. The image reveals a 7- to 10-cm long stricture at the terminal ileum (*white arrows*) causing obstruction and significant dilatation of the proximal small bowel (*white asterisk*). A fetus is seen in the uterus (*dashed white arrows*). *(Courtesy of Drs. J. F. B. Chick and P. B. Shyn, Abdominal Imaging and Intervention, Department of Radiology, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 319-10

A coronal balanced, steady-state, free precession, T2-weighted image with fat saturation was obtained in a 32-year-old man with Crohn's disease and prior episodes of bowel obstruction, fistulas, and abscesses. He was being treated with 6-mercaptopurine and presented with abdominal distention and diarrhea. The image demonstrates a new gastrocolic fistula (*solid white arrows*). Multifocal involvement of the small bowel and terminal ileum is also present (*dashed white arrows*). *(Courtesy of Drs. J. F. B. Chick and P. B. Shyn, Abdominal Imaging and Intervention, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 319-11

Axial T2-weighted fat-saturated image obtained in a 39-year-old male with Crohn's disease shows a defect in the internal sphincter at the 6:00 position of the mid anal canal (*open white arrow*) communicating with a 1.1-cm intersphincteric collection (*black arrow*). Wide defect in the external sphincter at the 7:00 position *(solid white arrow)* leads to a moderate sized perianal abscess in the ischioanal fossa (*asterisk*). *(Courtesy of Drs. J.S. Quon and P.B. Shyn, Abdominal Imaging and Intervention, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Complications

Because CD is a transmural process, serosal adhesions develop that provide direct pathways for fistula formation and reduce the incidence of free perforation. Perforation occurs in 1–2% of patients, usually in the ileum but occasionally in the jejunum or as a complication of toxic megacolon. The peritonitis of free perforation, especially colonic, may be fatal. Intraabdominal and pelvic abscesses occur in 10–30% of patients with CD at some time in the course of their illness. CT-guided percutaneous drainage of the abscess is standard therapy. Despite adequate drainage, most patients need resection of the offending bowel segment. Percutaneous drainage has an especially high failure rate in abdominal wall abscesses. Systemic glucocorticoid therapy increases the risk of intraabdominal and pelvic abscesses in CD patients who have never had an operation. Other complications include intestinal obstruction in 40%, massive hemorrhage, malabsorption, and severe perianal disease.

Serologic Markers

Patients with CD show a wide variation in the way they present and progress over time. Some patients present with mild disease activity and do well with generally safe and mild medications, but many others exhibit more severe disease and can develop serious complications that will require surgery. Current and developing biologic therapies can help halt progression of disease and give patients with moderate to severe CD a better quality of life. There are potential risks of biologic therapies such as infection and malignancy, and it would be optimal to determine by genetic or serologic markers at the time of diagnosis which patients will require more aggressive medical therapy. This same argument holds true for UC patients as well.

For success in diagnosing IBD and in differentiating between CD and UC, the efficacy of these serologic tests depends on the prevalence of IBD in a specific population. Increased titers of anti-Saccharomyces cerevisiae antibodies (ASCAs) have been associated with CD, whereas increased levels of perinuclear antineutrophil cytoplasmic antibodies (pANCA) are more commonly seen in patients with UC. However, when evaluated in a meta-analysis of 60 studies, the sensitivity and specificity of a ASCA+/pANCA- pattern for identification of CD was 55% and 93% respectively. In addition to ASCA, multiple other antibodies to bacterial proteins (Omp-C and I2), flagellin (CBir1) and bacterial carbohydrates have been studied and associated with CD, including laminaribioside (ALCA), chitobioside (ACCA) and mannobioside (SMCA). These serologic markers tend to have low sensitivity and specificity though due to elevation in levels cause by other autoimmune diseases, infections and inflammation outside the GI tract.

Clinical factors described at diagnosis are more helpful than serologies at predicting the natural history of CD. The initial requirements for glucocorticoid use, an age at diagnosis below 40 years and the presence of perianal disease at diagnosis, have been shown to be independently associated with subsequent disabling CD after 5 years. Except in special circumstances (such as before consideration of an ileoanal pouch anastomosis [IPAA] in a patient with indeterminate colitis), serologic markers have only minimal clinical utility.

DIFFERENTIAL DIAGNOSIS OF UC AND CD

UC and CD have similar features to many other diseases. In the absence of a key diagnostic test, a combination of features is used (Table 319-5). Once a diagnosis of IBD is made, distinguishing between UC and CD is impossible initially in up to 15% of cases. These are termed *indeterminate colitis.* Fortunately, in most cases, the true nature of the underlying colitis becomes evident later in the course of the patient's disease. Approximately 5% (range 1–20%) of colon resection specimens are difficult to classify as either UC or CD because they exhibit overlapping histologic features.

TABLE 319-5

Different Clinical, Endoscopic, and Radiographic Features

	Ulcerative Colitis	Crohn's Disease
Clinical		
Gross blood in stool	Yes	Occasionally
Mucus	Yes	Occasionally
Systemic symptoms	Occasionally	Frequently
Pain	Occasionally	Frequently
Abdominal mass	Rarely	Yes
Significant perineal disease	No	Frequently
Fistulas	No	Yes
Small intestinal obstruction	No	Frequently
Colonic obstruction	Rarely	Frequently
Response to antibiotics	No	Yes
Recurrence after surgery	No	Yes
Endoscopic		
Rectal sparing	Rarely	Frequently
Continuous disease	Yes	Occasionally
"Cobblestoning"	No	Yes
Granuloma on biopsy	No	Occasionally
Radiographic		
Small bowel significantly abnormal	No	Yes
Abnormal terminal ileum	No	Yes
Segmental colitis	No	Yes
Asymmetric colitis	No	Yes
Stricture	Occasionally	Frequently

INFECTIOUS DISEASES

Infections of the small intestines and colon can mimic CD or UC. They may be bacterial, fungal, viral, or protozoal in origin **(Table 319-6)**. *Campylobacter* colitis can mimic the endoscopic appearance of severe UC and can cause a relapse of established UC. *Salmonella* can cause watery or bloody diarrhea, nausea, and vomiting. Shigellosis causes watery diarrhea, abdominal pain, and fever followed by rectal tenesmus and by the passage of blood and mucus per rectum. All three are usually self-limited, but 1% of patients infected with *Salmonella* become asymptomatic carriers. *Yersinia enterocolitica* infection occurs mainly in the terminal ileum and causes mucosal ulceration, neutrophil invasion, and thickening of the ileal wall. Other bacterial infections that may mimic IBD include *C. difficile*, which presents with watery diarrhea, tenesmus, nausea, and vomiting; and *E. coli*, three categories of which can cause colitis. These are enterohemorrhagic, enteroinvasive, and enteroadherent *E. coli*, all of which can cause bloody diarrhea and abdominal tenderness. Diagnosis of bacterial colitis is made by sending stool specimens for bacterial culture and *C. difficile* toxin analysis. Gonorrhea, *Chlamydia*, and syphilis can also cause proctitis.

TABLE 319-6

Diseases That Mimic IBD

Infectious Etiologies		
Bacterial	**Mycobacterial**	**Viral**
Salmonella	Tuberculosis	Cytomegalovirus
Shigella	*Mycobacterium avium*	Herpes simplex
Toxigenic		HIV
Escherichia coli	**Parasitic**	**Fungal**
Campylobacter	Amebiasis	Histoplasmosis
Yersinia	*Isospora*	*Candida*
Clostridium difficile	*Trichuris trichiura*	*Aspergillus*
Gonorrhea	Hookworm	
Chlamydia trachomatis	*Strongyloides*	
Noninfectious Etiologies		
Inflammatory	**Neoplastic**	**Drugs and Chemicals**
Appendicitis	Lymphoma	NSAIDs
Diverticulitis	Metastatic	Phosphosoda
Diversion colitis	Carcinoma	Cathartic colon
Collagenous/lymphocytic colitis	Carcinoma of the ileum	Gold
		Oral contraceptives
Ischemic colitis	Carcinoid	Cocaine
Radiation colitis/enteritis	Familial polyposis	Ipilimumab
		Mycophenolate mofetil
Solitary rectal ulcer syndrome		
Eosinophilic gastroenteritis		
Neutropenic colitis		
Behçet's syndrome		
Graft-versus-host disease		

Abbreviation: IBD, inflammatory bowel disease; NSAIDs, nonsteroidal anti-inflammatory drugs.

GI involvement with mycobacterial infection occurs primarily in the immunosuppressed patient but may occur in patients with normal immunity. Distal ileal and cecal involvement predominates, and patients present with symptoms of small-bowel obstruction and a tender abdominal mass. The diagnosis is

made most directly by colonoscopy with biopsy and culture. *Mycobacterium avium-intracellulare* complex infection occurs in advanced stages of HIV infection and in other immunocompromised states; it usually manifests as a systemic infection with diarrhea, abdominal pain, weight loss, fever, and malabsorption. Diagnosis is established by acid-fast smear and culture of mucosal biopsies.

Although most of the patients with viral colitis are immunosuppressed, cytomegalovirus (CMV) and herpes simplex proctitis may occur in immunocompetent individuals. CMV occurs most commonly in the esophagus, colon, and rectum but may also involve the small intestine. Symptoms include abdominal pain, bloody diarrhea, fever, and weight loss. With severe disease, necrosis and perforation can occur. Diagnosis is made by identification of characteristic intranuclear inclusions in mucosal cells on biopsy. Herpes simplex infection of the GI tract is limited to the oropharynx, anorectum, and perianal areas. Symptoms include anorectal pain, tenesmus, constipation, inguinal adenopathy, difficulty with urinary voiding, and sacral paresthesias. Diagnosis is made by rectal biopsy with identification of characteristic cellular inclusions and viral culture. HIV itself can cause diarrhea, nausea, vomiting, and anorexia. Small intestinal biopsies show partial villous atrophy; small bowel bacterial overgrowth and fat malabsorption may also be noted.

Protozoan parasites include *Isospora belli*, which can cause a self-limited infection in healthy hosts but causes a chronic profuse, watery diarrhea, and weight loss in AIDS patients. *Entamoeba histolytica* or related species infect about 10% of the world's population; symptoms include abdominal pain, tenesmus, frequent loose stools containing blood and mucus, and abdominal tenderness. Colonoscopy reveals focal punctate ulcers with normal intervening mucosa; diagnosis is made by biopsy or serum amebic antibodies. Fulminant amebic colitis is rare but has a mortality rate of >50%.

Other parasitic infections that may mimic IBD include hookworm (*Necator americanus*), whipworm (*Trichuris trichiura*), and *Strongyloides stercoralis.* In severely immunocompromised patients, *Candida* or *Aspergillus* can be identified in the submucosa. Disseminated histoplasmosis can involve the ileocecal area.

NONINFECTIOUS DISEASES

Diverticulitis can be confused with CD clinically and radiographically. Both diseases cause fever, abdominal pain, tender abdominal mass, leukocytosis, elevated ESR, partial obstruction, and fistulas. Perianal disease or ileitis on small-bowel series favors the diagnosis of CD. Significant endoscopic mucosal abnormalities are more likely in CD than in diverticulitis. Endoscopic or clinical recurrence following segmental resection favors CD. Diverticular-associated colitis is similar to CD, but mucosal abnormalities are limited to the sigmoid and descending colon.

Ischemic colitis is commonly confused with IBD. The ischemic process can be chronic and diffuse, as in UC, or segmental, as in CD. Colonic inflammation due to ischemia may resolve quickly or may persist and result in transmural scarring and stricture formation. Ischemic bowel disease should be considered in the elderly following abdominal aortic aneurysm repair or when a patient has a hypercoagulable state or a severe cardiac or peripheral vascular disorder. Patients usually present with sudden onset of left lower quadrant pain, urgency to defecate, and the passage of bright red blood per rectum. Endoscopic examination often demonstrates a normal-appearing rectum and a sharp transition to an area of inflammation in the descending colon and splenic flexure.

The effects of radiotherapy on the GI tract can be difficult to distinguish from IBD. Acute symptoms can occur within 1–2 weeks of starting radiotherapy. When the rectum and sigmoid are irradiated, patients develop bloody, mucoid diarrhea and tenesmus, as in distal UC. With small-bowel involvement, diarrhea is common. Late symptoms include malabsorption and weight loss. Stricturing with obstruction and bacterial overgrowth may occur. Fistulas can penetrate the bladder, vagina, or abdominal wall. Flexible sigmoidoscopy reveals mucosal granularity, friability, numerous telangiectasias, and occasionally discrete ulcerations. Biopsy can be diagnostic.

Solitary rectal ulcer syndrome is uncommon and can be confused with IBD. It occurs in persons of all ages and may be caused by impaired evacuation and failure of relaxation of the puborectalis muscle. Single or multiple ulcerations may arise from anal sphincter overactivity, higher intrarectal pressures during defecation, and digital removal of stool. Patients complain of constipation with straining and pass blood and mucus per rectum. Other symptoms include abdominal pain, diarrhea, tenesmus, and perineal pain. Ulceration, that may be as large as 5 cm in diameter, is usually observed anteriorly or anterior-laterally 3–15 cm from the anal verge. Biopsies can be diagnostic.

Several types of colitis are associated with nonsteroidal anti-inflammatory drugs (NSAIDs), including de novo colitis, reactivation of IBD, and proctitis caused by use of suppositories. Most patients with NSAID-related colitis present with diarrhea and abdominal pain, and complications include stricture, bleeding, obstruction, perforation, and fistulization. Withdrawal of these agents is crucial, and in cases of reactivated IBD, standard therapies are indicated.

There are complications of two common drugs used in a hospital setting that mimic IBD. The first is ipilimumab, a drug that targets cytotoxic T lymphocyte antigen 4 (CTLA-4) and reverses T cell inhibition and is used to treat metastatic melanoma. Ipilimumab can cause an autoimmune colitis that is commonly associated with diarrhea: patients with diarrhea of grade 3 or greater and those who have colitis on colonoscopy often require glucocorticoid or infliximab therapy. The second is mycophenolate mofetil (MMF), an immunosuppressive agent that is anti-proliferative and commonly used to prevent post-transplant rejection. The colitis associated with MMF is common and can occur in more than one-third of patients taking the drug. Treatment is dose reduction or cessation of the drug. There have been case reports of entanercept (TNF receptor-Fc fusion protein) associated with de novo CD and UC.

THE ATYPICAL COLITIDES

Two atypical colitides—collagenous colitis and lymphocytic colitis—have completely normal endoscopic appearances. Collagenous colitis has two main histologic components: increased subepithelial collagen deposition and colitis with increased intraepithelial lymphocytes. The female to male ratio is 9:1, and most patients present in the sixth or seventh decades of life. The main symptom is chronic watery diarrhea. Treatments range from sulfasalazine or mesalamine and diphenoxylate/atropine (Lomotil) to bismuth to budesonide to prednisone or azathioprine/6-mercaptopurine for refractory disease. Risk factors include smoking, use of NSAIDs, proton pump inhibitors, or beta blockers; and a history of autoimmune disease.

Lymphocytic colitis has features similar to collagenous colitis, including age at onset and clinical presentation, but it has almost equal incidence in men and women and no subepithelial collagen deposition on pathologic section. However, intraepithelial lymphocytes are increased. Use of sertraline (but not beta blockers) is an additional risk factor. The frequency of celiac disease is increased in lymphocytic colitis and ranges from 9 to 27%. Celiac disease should be excluded in all patients with lymphocytic colitis, particularly if diarrhea does not respond to conventional therapy. Treatment is similar to that of collagenous colitis with the exception of a gluten-free diet for those who have celiac disease.

Diversion colitis is an inflammatory process that arises in segments of the large intestine that are excluded from the fecal stream. It usually occurs in patients with ileostomy or colostomy when a mucus fistula or a Hartmann's pouch has been created. Clinically, patients have mucus or bloody discharge from the rectum. Erythema, granularity, friability, and, in more severe cases, ulceration can be seen on endoscopy. Histopathology shows areas of active inflammation with foci of cryptitis and crypt abscesses. Crypt architecture is normal, which differentiates it from UC. It may be impossible to distinguish from CD. Short-chain fatty acid enemas may help in diversion colitis, but the definitive therapy is surgical reanastomosis.

EXTRAINTESTINAL MANIFESTATIONS

Up to one-third of IBD patients have at least one extraintestinal disease manifestation.

DERMATOLOGIC

Erythema nodosum (EN) occurs in up to 15% of CD patients and 10% of UC patients. Attacks usually correlate with bowel activity; skin lesions develop after the onset of bowel symptoms, and patients frequently have concomitant active peripheral arthritis. The lesions of EN are hot, red, tender nodules measuring 1–5 cm in diameter and are found on the anterior surface of the lower legs, ankles, calves, thighs, and arms. Therapy is directed toward the underlying bowel disease.

Pyoderma gangrenosum (PG) is seen in 1–12% of UC patients and less commonly in Crohn's colitis. Although it usually presents after the diagnosis of IBD, PG may occur years before the onset of bowel symptoms, run a course independent of the bowel disease, respond poorly to colectomy, and even develop years after proctocolectomy. It is usually associated with severe disease. Lesions are commonly found on the dorsal surface of the feet and legs but may occur on the arms, chest, stoma, and even the face. PG usually begins as a pustule and then spreads concentrically to rapidly undermine healthy skin. Lesions then ulcerate, with violaceous edges surrounded by a margin of erythema. Centrally, they contain necrotic tissue with blood and exudates. Lesions may be single or multiple and grow as large as 30 cm. They are sometimes very difficult to treat and often require IV antibiotics, IV glucocorticoids, dapsone, azathioprine, thalidomide, IV cyclosporine (CSA), infliximab or adalimumab.

Other dermatologic manifestations include pyoderma vegetans, which occurs in intertriginous areas; pyostomatitis vegetans, which involves the mucous membranes; Sweet syndrome, a neutrophilic dermatosis; and metastatic CD, a rare disorder defined by cutaneous granuloma formation. Psoriasis affects 5–10% of patients with IBD and is unrelated to bowel activity consistent with the potential shared immunogenetic basis of these diseases. Perianal skin tags are found in 75–80% of patients with CD, especially those with colon involvement. Oral mucosal lesions, seen often in CD and rarely in UC, include aphthous stomatitis and "cobblestone" lesions of the buccal mucosa.

RHEUMATOLOGIC

Peripheral arthritis develops in 15–20% of IBD patients, is more common in CD, and worsens with exacerbations of bowel activity. It is asymmetric, polyarticular, and migratory and most often affects large joints of the upper and lower extremities. Treatment is directed at reducing bowel inflammation. In severe UC, colectomy frequently cures the arthritis.

Ankylosing spondylitis (AS) occurs in about 10% of IBD patients and is more common in CD than UC. About two-thirds of IBD patients with AS express the HLA-B27 antigen. The AS activity is not related to bowel activity and does not remit with glucocorticoids or colectomy. It most often affects the spine and pelvis, producing symptoms of diffuse low-back pain, buttock pain, and morning stiffness. The course is continuous and progressive, leading to permanent skeletal damage and deformity. Anti-TNF therapy reduces spinal inflammation and improves functional status and quality of life.

Sacroiliitis is symmetric, occurs equally in UC and CD, is often asymptomatic, does not correlate with bowel activity, and does not always progress to AS. Other rheumatic manifestations include hypertrophic osteoarthropathy, pelvic/femoral osteomyelitis, and relapsing polychondritis.

OCULAR

The incidence of ocular complications in IBD patients is 1–10%. The most common are conjunctivitis, anterior uveitis/iritis, and episcleritis. Uveitis is associated with both UC and Crohn's colitis, may be found during periods of remission, and may develop in patients following bowel resection. Symptoms include ocular pain, photophobia, blurred vision, and headache. Prompt intervention, sometimes with systemic glucocorticoids, is required to prevent scarring and visual impairment. Episcleritis is a benign disorder that presents with symptoms of mild ocular burning. It occurs in 3–4% of IBD patients, more commonly in Crohn's colitis, and is treated with topical glucocorticoids.

HEPATOBILIARY

Hepatic steatosis is detectable in about one-half of the abnormal liver biopsies from patients with CD and UC; patients usually present with hepatomegaly. Fatty liver usually results from a combination of chronic debilitating illness, malnutrition, and glucocorticoid therapy. Cholelithiasis occurs in 10–35% of CD patients with ileitis or ileal resection. Gallstone formation is caused by malabsorption of bile acids, resulting in depletion of the bile salt pool and the secretion of lithogenic bile.

Primary sclerosing cholangitis (PSC) is a disorder characterized by both intrahepatic and extrahepatic bile duct inflammation and fibrosis, frequently leading to biliary cirrhosis and hepatic failure; ~5% of patients with UC have PSC, but 50–75% of patients with PSC have IBD. PSC occurs less often in patients with CD. Although it can be recognized after the diagnosis of IBD, PSC can be detected earlier or even years after proctocolectomy. Consistent with this, the immunogenetic basis for PSC appears to be overlapping but distinct from UC based on GWAS, although both IBD and PSC are commonly pANCA positive. Most patients have no symptoms at the time of diagnosis; when symptoms are present, they consist of fatigue, jaundice, abdominal pain, fever, anorexia, and malaise. The traditional gold standard diagnostic test is endoscopic retrograde cholangiopancreatography (ERCP), but magnetic resonance cholangiopancreatography (MRCP) is sensitive, specific and safer. MRCP is reasonable as an initial diagnostic test in children and adults and can visualize irregularities, multifocal strictures, and dilatations of all levels of the biliary tree. In patients with PSC, both ERCP and MRCP demonstrate multiple bile duct strictures alternating with relatively normal segments.

Gallbladder polyps in patients with PSC have a high incidence of malignancy and cholecystectomy is recommended, even if a mass lesion is less than 1 cm in diameter. Gallbladder surveillance with ultrasound should be performed annually. Endoscopic stenting may be palliative for cholestasis secondary to bile duct obstruction. Patients with symptomatic disease develop cirrhosis and liver failure over 5–10 years and eventually require liver transplantation. PSC patients have a 10–15% lifetime risk of developing cholangiocarcinoma and then cannot be transplanted. Patients with IBD and PSC are at increased risk of colon cancer and should be surveyed yearly by colonoscopy and biopsy.

In addition, cholangiography is normal in a small percentage of patients who have a variant of PSC known as *small duct primary sclerosing cholangitis*. This variant (sometimes referred to as "pericholangitis") is probably a form of PSC involving small-caliber bile ducts. It has similar biochemical and histologic features to classic PSC. It appears to have a significantly better prognosis than classic PSC, although it may evolve into classic PSC. Granulomatous hepatitis and hepatic amyloidosis are much rarer extraintestinal manifestations of IBD.

UROLOGIC

The most frequent genitourinary complications are calculi, ureteral obstruction, and ileal bladder fistulas. The highest frequency of nephrolithiasis (10–20%) occurs in patients with CD following small bowel resection. Calcium oxalate stones develop secondary to hyperoxaluria, which results from increased absorption of dietary oxalate. Normally, dietary calcium combines with luminal oxalate to form insoluble calcium oxalate, which is eliminated in the stool. In patients with ileal dysfunction, however, nonabsorbed fatty acids bind calcium and leave oxalate unbound. The unbound oxalate is then delivered to the colon, where it is readily absorbed, especially in the presence of inflammation.

METABOLIC BONE DISORDERS

Low bone mass occurs in 14–42% of IBD patients. The risk is increased by glucocorticoids, CSA, methotrexate (MTX), and total parenteral nutrition (TPN). Malabsorption and inflammation mediated by IL-1, IL-6, TNF, and other inflammatory mediators also contribute to low bone density. An increased incidence of hip, spine, wrist, and rib fractures has been noted: 36% in CD and 45% in UC. The absolute risk of an osteoporotic fracture is about 1% per person per year. Fracture rates, particularly in the spine and hip, are highest among the elderly (age >60). One study noted an OR of 1.72 for vertebral fracture and an OR of 1.59 for hip fracture. The disease severity predicted the risk of a fracture. Only 13% of IBD patients who had a fracture were on any kind of antifracture treatment. Up to 20% of bone mass can be lost per year with chronic glucocorticoid use. The effect is dosage-dependent. Budesonide may also suppress the pituitary-adrenal axis and thus carries a risk of causing osteoporosis.

Osteonecrosis is characterized by death of osteocytes and adipocytes and eventual bone collapse. The pain is aggravated by motion and swelling of the joints. It affects the hips more often than knees and shoulders, and in one series, 4.3% of patients developed osteonecrosis within 6 months of starting glucocorticoids. Diagnosis is made by bone scan or MRI, and treatment consists of pain control, cord decompression, osteotomy, and joint replacement.

THROMBOEMBOLIC DISORDERS

Patients with IBD have an increased risk of both venous and arterial thrombosis even if the disease is not active. Factors responsible for the hypercoagulable state have included abnormalities of the platelet-endothelial interaction, hyperhomocysteinemia, alterations in the coagulation cascade,

impaired fibrinolysis, involvement of tissue factor-bearing microvesicles, disruption of the normal coagulation system by autoantibodies, and a genetic predisposition. A spectrum of vasculitides involving small, medium, and large vessels has also been observed.

OTHER DISORDERS

More common cardiopulmonary manifestations include endocarditis, myocarditis, pleuropericarditis, and interstitial lung disease. A secondary or reactive amyloidosis can occur in patients with long-standing IBD, especially in patients with CD. Amyloid material is deposited systemically and can cause diarrhea, constipation, and renal failure. The renal disease can be successfully treated with colchicine. Pancreatitis is a rare extraintestinal manifestation of IBD and results from duodenal fistulas; ampullary CD; gallstones; PSC; drugs such as 6-mercaptopurine, azathioprine, or, very rarely, 5-ASA agents; autoimmune pancreatitis; and primary CD of the pancreas.

TREATMENT

TREATMENT

Inflammatory Bowel Disease

5-ASA AGENTS

These agents are effective at inducing and maintaining remission in UC. They may have a limited role in inducing remission in CD but no clear role in maintenance of CD. Newer sulfa-free aminosalicylate preparations deliver increased amounts of the pharmacologically active ingredient of sulfasalazine (5-ASA, mesalamine) to the site of active bowel disease while limiting systemic toxicity. Peroxisome proliferator activated receptor γ (PPAR-γ) may mediate 5-ASA therapeutic action by decreasing nuclear localization of NF-κB. Sulfa-free aminosalicylate formulations include alternative azo-bonded carriers, 5-ASA dimers, and delayed-release and controlled-release preparations. Each has the same efficacy as sulfasalazine when equimolar concentrations are used.

Sulfasalazine's molecular structure provides a convenient delivery system to the colon by allowing the intact molecule to pass through the small intestine after only partial absorption and to be broken down in the colon by bacterial azo reductases that cleave the azo bond linking the sulfa and 5-ASA moieties. Sulfasalazine is effective treatment for mild to moderate UC and is occasionally used in Crohn's colitis, but its high rate of side effects limits its use. Although sulfasalazine is more effective at higher doses, at 6 or 8 g/d up to 30% of patients experience allergic reactions or intolerable side effects such as headache, anorexia, nausea, and vomiting that are attributable to the sulfapyridine moiety. Hypersensitivity reactions, independent of sulfapyridine levels, include rash, fever, hepatitis, agranulocytosis, hypersensitivity pneumonitis, pancreatitis, worsening of colitis, and reversible sperm abnormalities. Sulfasalazine can also impair folate absorption, and patients should be given folic acid supplements.

Balsalazide contains an azo bond binding mesalamine to the carrier molecule 4-aminobenzoyl-β-alanine; it is effective in the colon.

Delzicol and Asacol HD (high dose) are enteric-coated forms of mesalamine with the 5-ASA being released at pH >7. They disintegrate with complete breakup of the tablet occurring in many different parts of the gut ranging from the small intestine to the splenic flexure; they have increased gastric residence when taken with a meal. Asacol has been discontinued and replaced with Delzicol, which lacks dibutyl phthalate (DBP), an inactive ingredient in Asacol's enteric coating. DBP has been associated with adverse effects on the male reproductive system in animals at very high doses.

Lialda is a once-a-day formulation of mesalamine (Multi-Matrix System [MMX]) designed to release mesalamine in the colon. The MMX technology incorporates mesalamine into a lipophilic matrix within a hydrophilic matrix encapsulated in a polymer resistant to degradation at a low pH (<7) to delay release throughout the colon. The safety profile appears to be comparable to other 5-ASA formulations.

Apriso is a formulation containing encapsulated mesalamine granules that delivers mesalamine to the terminal ileum and colon via a proprietary extended-release mechanism (Intellicor). The outer coating of this agent (Eudragit L) dissolves at a pH >6. In addition, there is a polymer matrix core that aids in sustained release throughout the colon. Because Lialda and Apriso are given once daily, an anticipated benefit is improved compliance compared with two to four daily doses required for other mesalamine preparations.

Pentasa is another mesalamine formulation that uses an ethylcellulose coating to allow water absorption into small beads containing the mesalamine. Water dissolves the 5-ASA, which then diffuses out of the bead into the lumen. Disintegration of the capsule occurs in the stomach. The microspheres then disperse throughout the entire GI tract from the small intestine through the distal colon in both fasted and fed conditions.

Salofalk® Granu-Stix, an unencapsulated version of mesalamine, has been in use in Europe for induction and maintenance of remission for several years.

Appropriate doses of the 5-ASA compounds are shown in **Table 319-7**. Some 50–75% of patients with mild to moderate UC improve when treated with 5-ASA doses equivalent to 2 g/d of mesalamine; the dose response continues up to at least 4.8 g/d.

More common side effects of the 5-ASA medications include headaches, nausea, hair loss, and abdominal pain. Rare side effects of the 5-ASA medications include renal impairment, hematuria, pancreatitis, and paradoxical worsening of colitis. Renal function tests and urinalysis should be checked yearly.

Topical *Rowasa* enemas are composed of mesalamine and are effective in mild-to-moderate distal UC. Combination therapy with mesalamine in both oral and enema form is more effective than either treatment alone for both distal and extensive UC.

Canasa suppositories composed of mesalamine are effective in treating proctitis.

GLUCOCORTICOIDS

The majority of patients with moderate to severe UC benefit from oral or parenteral glucocorticoids. Prednisone is usually started at doses of 40–60 mg/d for active UC that is unresponsive to 5-ASA therapy. Parenteral glucocorticoids may be administered as hydrocortisone, 300 mg/d, or methylprednisolone, 40–60 mg/d. A new glucocorticoid for UC, budesonide (Uceris), is released entirely in the colon and has minimal to no glucocorticoid side effects. The dose is 9 mg/d for 8 weeks, and no taper is required. Topically applied glucocorticoids are also beneficial for distal colitis and may serve as an adjunct in those who have rectal involvement plus more proximal disease. Hydrocortisone enemas or foam may control active disease, although they have no proven role as maintenance therapy. These glucocorticoids are significantly absorbed from the rectum and can lead to adrenal suppression with prolonged administration. Topical 5-ASA therapy is more effective than topical steroid therapy in the treatment of distal UC.

Glucocorticoids are also effective for treatment of moderate to severe CD and induce a 60–70% remission rate compared to a 30% placebo response. The systemic effects of standard glucocorticoid formulations have led to the development of more potent formulations that are less well-absorbed and have increased first-pass metabolism. Controlled ileal-release budesonide has been nearly equal to prednisone for ileocolonic CD with fewer glucocorticoid side effects. Budesonide is used for 2–3 months at a dose of 9 mg/d, and then tapered. Glucocorticoids play no role in maintenance therapy in either UC or CD. Once clinical remission has been induced, they should be tapered according to the clinical activity, normally at a rate of no more than 5 mg/week. They can usually be tapered to 20 mg/d within 4–5 weeks but often take several months to be discontinued altogether. The side effects are numerous, including fluid retention, abdominal striae, fat redistribution, hyperglycemia, subcapsular cataracts, osteonecrosis, osteoporosis, myopathy, emotional disturbances, and withdrawal symptoms. Most of these side effects, aside from osteonecrosis, are related to the dose and duration of therapy.

ANTIBIOTICS

Antibiotics have no role in the treatment of active or quiescent UC. However, pouchitis, which occurs in about 30–50% of UC patients after colectomy and IPAA, usually responds to treatment with metronidazole and/or ciprofloxacin.

Metronidazole is effective in active inflammatory, fistulizing, and perianal CD and may prevent recurrence after ileal resection. The most effective dose is 15–20 mg/kg per day in three divided doses; it is usually continued for several months. Common side effects include nausea, metallic taste, and disulfiram-like reaction. Peripheral neuropathy can occur with prolonged administration (several months) and on rare occasions is permanent despite discontinuation. *Ciprofloxacin* (500 mg bid) is also beneficial for inflammatory, perianal, and fistulizing CD but has been associated with tendinitis and tendon rupture. Both ciprofloxacin and metronidazole antibiotics can be used only for short period of time due to side effects.

AZATHIOPRINE AND 6-MERCAPTOPURINE

Azathioprine and 6-mercaptopurine (6-MP) are purine analogues used concomitantly with biologic therapy or, less often, as the sole immunosuppressants. Azathioprine is rapidly absorbed and converted to 6-MP, which is then metabolized to the active end product, thioinosinic acid, an inhibitor of purine ribonucleotide synthesis and cell proliferation. Efficacy can be seen as early as 3–4 weeks but can take up to 4–6 months. Adherence can be monitored by measuring the levels of 6-thioguanine and 6-methyl-mercaptopurine, end products of 6-MP metabolism. The doses used range from 2–3 mg/kg per day for azathioprine and 1–1.5 mg/kg per day for 6-MP.

Although azathioprine and 6-MP are usually well tolerated, pancreatitis occurs in 3–4% of patients, typically presents within the first few weeks of therapy, and is completely reversible when the drug is stopped. Other side effects include nausea, fever, rash, and hepatitis. Bone marrow suppression (particularly leukopenia) is dose-related and often delayed, necessitating regular monitoring of the complete blood cell count (CBC). Additionally, 1 in 300 individuals lacks thiopurine methyltransferase, the enzyme responsible for drug metabolism to inactive end-products (6-methylmercaptopurine); an additional 11% of the population are heterozygotes with intermediate enzyme activity. Both are at increased risk of toxicity because of increased accumulation of active 6-thioguanine metabolites. Although 6-thioguanine and 6-methylmercaptopurine levels can be followed to determine correct drug dosing and reduce toxicity, weight-based dosing is an acceptable alternative. CBCs and liver function tests should be monitored frequently regardless of dosing strategy.

METHOTREXATE

MTX inhibits dihydrofolate reductase, resulting in impaired DNA synthesis. Additional anti-inflammatory properties may be related to decrease in the production of IL-1. It is used most often concomitantly with biologic therapy to decrease antibody formation and improve disease response. Intramuscular (IM) or subcutaneous (SC) doses range from 15 to 25 mg/week. Potential toxicities include leukopenia and hepatic fibrosis, necessitating periodic evaluation of CBCs and liver enzymes. The role of liver biopsy in patients on long-term MTX is uncertain but is probably limited to those with increased liver enzymes. Hypersensitivity pneumonitis is a rare but serious complication of therapy.

CYCLOSPORINE

CSA is a lipophilic peptide with inhibitory effects on both the cellular and humoral immune systems. CSA blocks the production of IL-2 by T helper lymphocytes. CSA binds to cyclophilin, and this complex inhibits calcineurin, a cytoplasmic phosphatase enzyme involved in the activation of T cells. CSA also indirectly inhibits B cell function by blocking helper T cells. CSA has a more rapid onset of action than 6-MP and azathioprine.

CSA is most effective when given at 2–4 mg/kg per day IV in severe UC that is refractory to IV glucocorticoids, with 82% of patients responding. CSA can be an alternative to colectomy. The long-term success of oral CSA is not as dramatic, but if patients are started on 6-MP or azathioprine at the time of hospital discharge, remission can be maintained. For the 2 mg/kg dose, levels as measured by monoclonal radioimmunoassay or by the high-performance liquid chromatography assay should be maintained between 150 and 350 ng/mL.

CSA may cause significant toxicity; renal function should be monitored frequently. Hypertension, gingival hyperplasia, hypertrichosis, paresthesias, tremors, headaches, and electrolyte abnormalities are common side effects. Creatinine elevation calls for dose reduction or discontinuation. Seizures may also complicate therapy, especially if the patient is hypomagnesemic or if serum cholesterol levels are <3.1 mmol/L (<120 mg/dL). Opportunistic infections, most notably *Pneumocystis carinii* pneumonia, may occur with combination immunosuppressive treatment; prophylaxis should be given. Major adverse events occurred in 15% of patients in one large study, including nephrotoxicity not responding to dose adjustment, serious infections, seizures, anaphylaxis, and death of two patients. This high incidence suggests that vigorous monitoring by experienced clinicians at tertiary care centers may be required. To compare IV CSA versus infliximab, a large trial was conducted in Europe by the GETAID group. The results indicated identical 7-day response rates between CSA 2 mg/kg (with doses adjusted for levels of 150–250 ng/mL) and infliximab 5 mg/kg, with both groups achieving response rates of 85%. Serious infections occurred in 5 of 55 CSA patients and 4 of 56 infliximab patients. Response rates were similar in the two groups at day 98 among patients treated with oral CSA versus infliximab at the usual induction dose and maintenance dose regimen (40 and 46%, respectively). In light of data showing equal efficacy of CSA and infliximab in severe UC, more physicians are relying on infliximab rather than CSA in these patients.

TACROLIMUS

Tacrolimus is a macrolide antibiotic with immunomodulatory properties similar to CSA. It is 100 times as potent as CSA and is not dependent on bile or mucosal integrity for absorption. These pharmacologic properties enable tacrolimus to have good oral absorption despite proximal small bowel Crohn's involvement. It has shown efficacy in children with refractory IBD and in adults with extensive involvement of the small bowel. It is also effective in adults with glucocorticoid-dependent or refractory UC and CD as well as refractory fistulizing CD.

BIOLOGIC THERAPIES

Biologic therapy is now commonly given as an initial therapy for patients with moderate to severe CD and UC. Patients who respond to biologic therapies enjoy an improvement in clinical symptoms; a better quality of life; less disability, fatigue, and depression; and fewer surgeries and hospitalizations.

Anti-TNF Therapies

The first biologic therapy approved for moderate to severely active CD and also UC was *infliximab*, a chimeric IgG1 antibody against TNF-α,. Of active CD patients refractory to glucocorticoids, 6-MP, or 5-ASA, 65% will respond to IV infliximab (5 mg/kg); one-third will enter complete remission. The ACCENT I (A Crohn's Disease Clinical Trial Evaluating Infliximab in a New Long-Term Treatment Regimen) study showed that of the patients who experience an initial response, 40% will maintain remission for at least 1 year with repeated infusions of infliximab every 8 weeks.

Infliximab is also effective in CD patients with refractory perianal and enterocutaneous fistulas, with the ACCENT II trial showing a 68% response rate (50% reduction in fistula drainage) and a 50% complete remission rate. Reinfusion, typically every 8 weeks, is necessary to continue therapeutic benefits in many patients.

The SONIC (Study of Biologic and Immunomodulator-Naive Patients with Crohn's Disease) trial compared infliximab plus azathioprine, infliximab alone, and azathioprine alone in immunomodulator- and biologic-naive patients with moderate to severe CD. At 1 year, the infliximab plus azathioprine group had a glucocorticoid-free remission rate of 46% compared with 35% for infliximab alone and 24% for azathioprine alone. There was also complete mucosal healing at week 26 with the combined approach relative to either infliximab or azathioprine alone (44 vs 30 vs 17%). The adverse events were equal between groups.

Two large trials of infliximab in moderate to severe UC also showed efficacy with a response rate of 37–49%, with about one-fifth of patients maintaining remission after 54 weeks. Dosing for UC and CD are identical, with induction dosing at 0, 2, and 6 weeks and every 8 weeks thereafter. There is a similar study to SONIC in patients with moderate to severe UC. After 16 weeks of therapy, UC patients taking azathioprine plus infliximab had a glucocorticoid-free remission rate of 40% compared to 24% and 22% of those on azathioprine and infliximab alone, respectively. This is even further evidence for "top-down" or more aggressive therapy for both moderate to severe CD and UC.

Adalimumab is a recombinant human monoclonal IgG1 antibody containing only human peptide sequences and is injected subcutaneously. Adalimumab binds TNF and neutralizes its function by blocking the interaction between TNF and its cell-surface receptor. Therefore, it seems to have a similar mechanism of action to infliximab but with less immunogenicity. Adalimumab is approved for treatment of moderate to severe CD and UC. CHARM (Crohn's Trial of the Fully Human Adalimumab for Remission Maintenance) is an adalimumab maintenance study in patients who responded to adalimumab induction therapy. About 50% of the patients in this trial were previously treated with infliximab. Remission rates ranged from 42 to 48% of infliximab-naïve patients at 1 year compared with remission rates of 31–34% in patients who had previously received infliximab. Another trial showed a remission rate of 21% at 4 weeks in patients who had initially responded to and then failed infliximab. UC results are similar with a sustained remission rate at one year of 22% (12.4% placebo) among anti–TNF-naïve patients and a sustained remission rate at 1 year of 10.2% (3% placebo) among patients who had previously received anti-TNF agents. In clinical practice, the remission rate in both CD and UC patients taking adalimumab increases with a dose increase to 40 mg weekly instead of every other week.

Certolizumab pegol is a pegylated form of an anti-TNF Fab portion of an antibody administered SC once monthly. SC certolizumab pegol was effective for induction of clinical response in patients with active inflammatory CD. In the PRECISE II (Pegylated Antibody Fragment Evaluation in Crohn's Disease) trial of maintenance therapy with certolizumab in patients who responded to certolizumab induction, the results were similar to the CHARM trial. At week 26, the subgroup of patients who were infliximab naïve had a response of 69% as compared to 44% in patients who had previously received infliximab.

Golimumab is another fully human IgG1 antibody against TNF-α and is currently approved for the treatment of moderately to severely active UC. Like adalimumab and certolizumab, golimumab is injected SC.

Side Effects of Anti-TNF Therapies

Development of Antibodies

The development of antibodies to infliximab is associated with an increased risk of infusion reactions and a decreased response to treatment. Current practice does not include giving on-demand or episodic infusions in contrast to periodic (every 8 week) infusions because patients are most likely to develop antibodies. Anti-infliximab antibodies are generally present when the quality of response or the response duration to infliximab infusion decreases. There are commercial assays for both infliximab and adalimumab antibodies and trough levels to determine optimal dosing. If a patient has high anti-infliximab antibodies and a low trough level of infliximab, it is best to switch to another anti-TNF therapy. Most acute infusion reactions and serum sickness can be managed with glucocorticoids and antihistamines. Some reactions can be serious and would necessitate a change in therapy, especially if a patient has anti-infliximab antibodies. It is now common practice to add an immunomodulator such as azathioprine, 6-mercaptopurine or MTX to anti-TNF therapy in order to prevent antibody formation.

Non-Hodgkin's Lymphoma (NHL)

The baseline risk of NHL in CD patients is 2:10,000, which is slightly higher than in the general population. Azathioprine and/or 6-MP therapy increases the risk to about 4:10,000. The highest risk for thiopurine-associated NHL is in patients over 65 years old actively using thiopurines (yearly incidence rate per 1000 patient years of 5.41), with a moderate risk in those between the ages of 50 and 65 (incidence rate of 2.58 compared to an incidence rate of 0.37 in patients <50 years old). It is difficult to assess whether anti-TNF medications are associated with lymphoma because most patients are also receiving thiopurines. After adjustment for co-treatments, no excess risk of lymphoma was found in a recent adequately powered Danish study of a cohort of IBD patients exposed to anti-TNF medications.

Hepatosplenic T-Cell Lymphoma (HSTCL)

HSTCL is a nearly universally fatal lymphoma in patients with or without CD. In patients with CD, events reported to the Food and Drug Administration Adverse Event Reporting System (FDA AERS) and search of PubMed and Embase published case reports demonstrate a total of 37 unique cases. Eighty-six percent of the patients were male, with a median age of 26 years. Patients had CD for a mean of 10 years before the diagnosis of HSTCL. Thirty-six cases had used either 6-MP or azathioprine, and 28 cases had used infliximab. Of these 28 cases, 27 had also used 6-MP or azathioprine. The other case had a history of both infliximab and adalimumab exposure.

Skin Lesions

New-onset psoriasiform skin lesions develop in nearly 5% of IBD patients treated with anti-TNF therapy. Most often, these can be treated topically, and occasionally, anti-TNF therapy must be decreased, switched, or stopped. Patients with IBD may have a slight unexplained intrinsic higher risk of developing melanoma. The risk of melanoma is increased almost twofold with anti-TNF and not thiopurine use. The risk of nonmelanoma skin cancer is increased with thiopurines and biologics, especially with 1 year of follow-up or greater. Patients on these medications should have a skin check at least once a year.

Infections

All of the anti-TNF drugs are associated with an increased risk of infections, particularly reactivation of latent tuberculosis and opportunistic fungal infections including disseminated histoplasmosis and coccidioidomycosis. It is recommended that patients have a purified protein derivative (PPD) or a QuantiFERON-TB gold test as well as a chest x-ray before initiation of anti-TNF therapy. Patients >65 have a higher rate of infections and death on infliximab or adalimumab than those <65 years of age.

Other

Acute liver injury due to reactivation of hepatitis B virus and to autoimmune effects and cholestasis has been reported. Rarely, infliximab and the other anti-TNF drugs have been associated with optic neuritis, seizures, new onset or exacerbation of clinical symptoms, and radiographic evidence of central nervous system demyelinating disorders, including multiple sclerosis. They may exacerbate symptoms in patients with New York Heart Association functional class III/IV heart failure.

ANTI-INTEGRINS

Integrins are expressed on the cell surface of leukocytes and serve as mediators of leukocyte adhesion to vascular endothelium. α4-Integrin along with its β1 or β7 subunit interact with endothelial ligands termed adhesion molecules. Interaction between α4β7 and mucosal addressin cellular adhesion molecule (MAdCAM-1) is important in lymphocyte trafficking to gut mucosa.

Natalizumab is a recombinant humanized IgG4 antibody against α4-integrin that has been shown to be effective in induction and maintenance of patients with CD. It has been approved since February 2008 for the treatment of patients with CD refractory or intolerant to anti-TNF therapy. The rates of response and remission at 3 months are about 60 and 40%, respectively, with a sustained remission rate of about 40% at 36 weeks.

Natalizumab is no longer widely used for CD due to the risk of progressive multifocal leukoencephalopathy (PML). The most important risk factor for development of PML is exposure to the John Cunningham (JC) polyomavirus, seen in 50–55% of the adult population. The other two risk factors for development of PML are longer duration of treatment, especially beyond 2 years, and prior treatment with an immunosuppressant medication. Patients with all three risk factors have an estimated risk of 11:1000.

The FDA approved a commercial enzyme-linked immunosorbent assay (ELISA) kit to assay anti-JC viral antibodies (Stratify JCV Antibody ELISA; Focus Diagnostics, Cypress, CA) in early 2012. The test is 99% accurate in stratifying risk of PML. It is recommended that all patients be tested prior to initiating natalizumab therapy. JC virus serologies are then measured every 6 months because 1–2% of patients will seroconvert yearly. Natalizumab is administered IV, 300 mg every 4 weeks. Labeling requirements mandate that it should not be used in combination with any immunosuppressant medications.

Vedolizumab, another leukocyte trafficking inhibitor, is a monoclonal antibody directed against α4β7 integrin specifically and has the ability to convey gut-selective immunosuppression. Vedolizumab is indicated for CD and UC patients who have had an inadequate response or lost response to, or were intolerant of a TNF blocker or immunomodulator; or had an inadequate response or were intolerant to, or demonstrated dependence on, glucocorticoids. It is also an option for patients who are JC antibody positive since unlike natalizumab it inhibits adhesion of a discrete gut-homing subset of T lymphocytes to MAdCAM-1, but not to vascular adhesion molecule-1. Vedolizumab decreases GI inflammation without inhibiting systemic immune responses or affecting T-cell trafficking to the central nervous system. Vedolizumab is given intravenously every 8 weeks after 3 induction doses at 0, 2, and 6 weeks. In the GEMINI I trial (A Phase 3, Randomized, Placebo-Controlled, Blinded, Multicenter Study of the Induction and Maintenance of Clinical Response and Remission by MLN002 in Patients With Moderate to Severe Ulcerative Colitis), 42% of the UC patients treated every 8 weeks and 45% of those treated every 4 weeks were in clinical remission at week 52 compared with 16% placebo. In the GEMINI II trial, the clinical remission rates for CD patients treated with vedolizumab were 36 to 39%, compared with 22% placebo at 52 weeks.

Ustekinumab, a fully human IgG1 monoclonal antibody, blocks the biologic activity of IL-12 and IL-23 through their common p40 subunit by inhibiting the interaction of these cytokines with their receptors on T cells, natural killer cells, and antigen presenting cells. It has recently been FDA-approved for use in Crohn's patients who have failed or were intolerant to immunomodulator or corticosteroid therapy but who never failed treatment with anti-TNF therapy or who failed or were intolerant to treatment with one or more anti-TNF medications. The result for the highest 6 mg/kg IV induction dose and subsequent 90 mg every 8 week dose in one major clinical trial was 41.7% remission rate versus 27.4% placebo at 22 weeks in Crohn's patients failing anti-TNF therapy.

THERAPIES IN CLINICAL DEVELOPMENT

Tofacitinib is an oral inhibitor of Janus kinases 1, 3, and, to a lesser extent, 2. It is expected to block signaling involving common gamma chain–containing cytokines including IL-2, IL-4, IL-7, IL-9, IL-15, and IL-21. These cytokines are integral to lymphocyte activation, function, and proliferation. It is effective in moderate to severe UC in clinical trials.

Biosimilars The FDA defines a biosimilar drug as a "biological product that is highly similar to the reference product not withstanding minor differences in clinically inactive components." There are no clinically meaningful differences between the biological product and the reference product in terms of the safety, purity, and potency of the product. The infliximab biosimilar CT-P13 is approved and available for use in almost 70 countries and many other biosimilars to infliximab and adalimumab are currently being manufactured. Biosimilars may be approved without randomized-controlled trials. The FDA examines quality considerations such as the expression system, manufacturing process, assessment of physiochemical properties, functional activities, receptor binding and immunochemical properties, measurement of impurities, stability under multiple stress conditions and effect of product formulation and shipping, Since biosimilars will likely cost about a third of the reference drug, they will likely be widely used in the near future in the United States.

Ozanimod is an oral agonist of the sphingosine-1-phosphate receptor subtypes 1 and 5 that causes peripheral lymphocyte sequestration, potentially decreasing the number of activated lymphocytes circulating to the GI tract. In a phase 2 trial of ozanimod in 197 patients with moderate to severe UC, at 32 weeks, 21% of patients who received 1 mg of ozanimod versus 6% achieved clinical remission. Phase 3 trials are now in progress.

NUTRITIONAL THERAPIES

Dietary antigens may stimulate the mucosal immune response. Patients with active CD respond to bowel rest, along with TPN. Bowel rest and TPN are as effective as glucocorticoids at inducing remission of active CD but are not effective as maintenance therapy. Enteral nutrition in the form of elemental or peptide-based preparations is also as effective as glucocorticoids or TPN, but these diets are not palatable. Enteral diets may provide the small intestine with nutrients vital to cell growth and do not have the complications of TPN. In contrast to CD, dietary intervention does not reduce inflammation in UC. Standard medical management of UC and CD is shown in **Fig. 319-12**.

SURGICAL THERAPY

Ulcerative Colitis

Nearly one-half of patients with extensive chronic UC undergo surgery within the first 10 years of their illness. The indications for surgery are listed in **Table 319-8**. Morbidity is about 20% for elective, 30% for urgent, and 40% for emergency proctocolectomy. The risks are primarily hemorrhage, contamination and sepsis, and neural injury. The operation of choice is an IPAA.

Because UC is a mucosal disease, the rectal mucosa can be dissected and removed down to the dentate line of the anus or about 2 cm proximal to this landmark. The ileum is fashioned into a pouch that serves as a neorectum. This ileal pouch is then sutured circumferentially to the anus in an end-to-end fashion. If performed carefully, this operation preserves the anal sphincter and maintains continence. The overall operative morbidity is 10%, with the major complication being bowel obstruction. Pouch failure necessitating conversion to permanent ileostomy occurs in 5–10% of patients. Some inflamed rectal mucosa is usually left behind, and thus endoscopic surveillance is necessary. Primary dysplasia of the ileal mucosa of the pouch has occurred rarely.

Patients with IPAA usually have about 6–10 bowel movements a day. On validated quality-of-life indices, they report better performance in sports and sexual activities than ileostomy patients. The most frequent complication of IPAA is pouchitis in about 30–50% of patients with UC. This syndrome consists of increased stool frequency, watery stools, cramping, urgency, nocturnal leakage of stool, arthralgias, malaise, and fever. Pouch biopsies may distinguish true pouchitis from underlying CD. Although pouchitis usually responds to antibiotics, 3–5% of patients remain refractory and may require glucocorticoids, immunomodulators, biologics or even pouch removal. A highly concentrated probiotic preparation with four strains of *Lactobacillus*, three strains of *Bifidobacterium*, and one strain of *Streptococcus salivarius* may prevent the recurrence of pouchitis when taken daily.

Crohn's Disease

Most patients with CD require at least one operation in their lifetime. The need for surgery is related to duration of disease and the site of involvement. Patients with small-bowel disease have an 80% chance of requiring surgery. Those with colitis alone have a 50% chance. Surgery is an option only when medical treatment has failed or complications dictate its necessity. The indications for surgery are shown in Table 319-8.

Small Intestinal Disease

Because CD is chronic and recurrent, with no clear surgical cure, as little intestine as possible is resected. Current surgical alternatives for treatment of obstructing CD include resection of the diseased segment and strictureplasty. Surgical resection of the diseased segment is the most frequently performed operation, and in most cases, primary anastomosis can be done to restore continuity. If much of the small bowel has already been resected and the strictures are short, with intervening areas of normal mucosa, strictureplasties should be done to avoid a functionally insufficient length of bowel. The strictured area of intestine is incised longitudinally and the incision sutured transversely, thus widening the narrowed area. Complications of strictureplasty include prolonged ileus, hemorrhage, fistula, abscess, leak, and restricture.

There is evidence that mesalamine, nitroimidazole antibiotics, 6-MP/azathioprine, infliximab, and adalimumab are all superior to placebo for the prevention of postoperative recurrence of CD. Mesalamine is the least effective, and the side effects of the nitroimidazole antibiotics limit their use. Risk factors for early recurrence of disease include cigarette smoking, penetrating disease (internal fistulas, abscesses, or other evidence of penetration through the wall of the bowel), early recurrence since a previous surgery, multiple surgeries, and a young age at the time of the first surgery. Aggressive postoperative treatment with 6-MP/azathioprine, infliximab, or adalimumab should be considered for this group of patients. It is also recommended to evaluate for endoscopic recurrence of CD via a colonoscopy, if possible, 6 months after surgery.

Colorectal Disease

A greater percentage of patients with Crohn's colitis require surgery for intractability, fulminant disease, and anorectal disease. Several alternatives are available, ranging from the use of a temporary loop ileostomy to resection of segments of diseased colon or even the entire colon and rectum. For patients with segmental involvement, segmental colon resection with primary anastomosis can be performed. In 20–25% of patients with extensive colitis, the rectum is spared sufficiently to consider rectal preservation. Most surgeons believe that an IPAA is contraindicated in CD due to the high incidence of pouch failure. A diverting colostomy may help heal severe perianal disease or rectovaginal fistulas, but disease almost always recurs with reanastomosis. These patients often require a total proctocolectomy and ileostomy.

TABLE 319-7

Oral 5-ASA Preparations

Preparation	Formulation	Delivery	Dosing Per Day
Azo-Bond			
Sulfasalazine (500 mg) (Azulfidine)	Sulfapyridine-5-ASA	Colon	3–6 g (acute) 2–4 g (maintenance)
Balsalazide (750 mg) (Colazal)	Aminobenzoyl-alanine–5-ASA	Colon	6.75–9 g
Delayed-Release			
Mesalamine (400, 800 mg) (Delzicol, Asacol HD)	Eudragit S (pH 7)	Distal ileum-colon	2.4–4.8 g (acute) 1.6–4.8 g (maintenance)
Mesalamine (1.2 g) (Lialda)	MMX mesalamine (SPD476)	Ileum-colon	2.4–4.8 g
Controlled-Release			
Mesalamine (250, 500, 1000 mg) (Pentasa)	Ethylcellulose microgranules	Stomach-colon	2–4 g (acute) 1.5–4 g (maintenance)
Delayed- and Extended-Release			
Mesalamine (0.375 g) (Apriso)	Intellicor extended-release mechanism	Ileum-colon	1.5 g (maintenance)

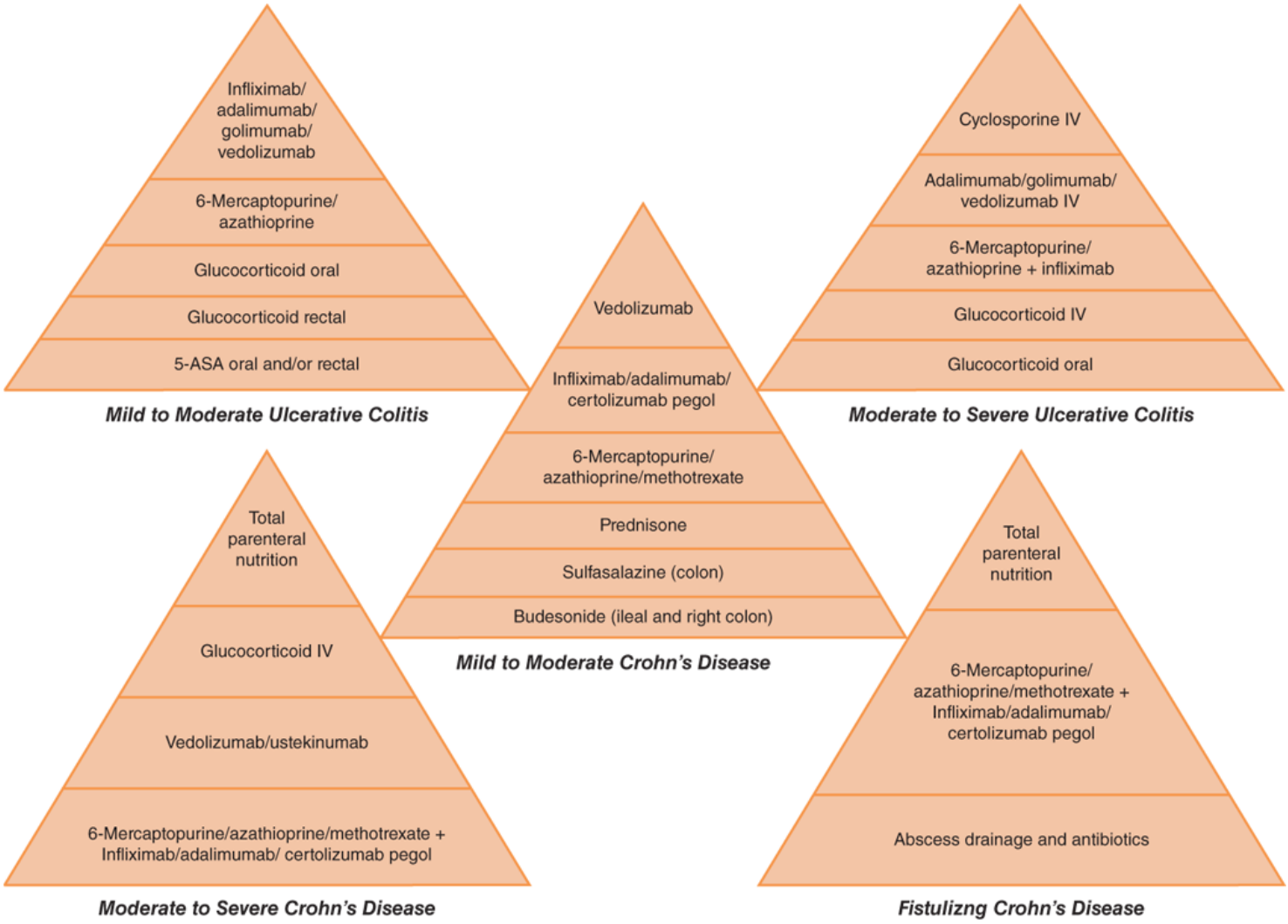
TABLE 319-8

Indications for Surgery

Ulcerative Colitis	Crohn's Disease
Intractable disease	Small Intestine
Fulminant disease	Stricture and obstruction
Toxic megacolon	unresponsive to medical therapy
Colonic perforation	Massive hemorrhage
Massive colonic hemorrhage	Refractory fistula
Extracolonic disease	Abscess
Colonic obstruction	Colon and rectum
Colon cancer prophylaxis	Intractable disease
Colon dysplasia or cancer	Fulminant disease
	Perianal disease unresponsive to medical therapy
	Refractory fistula
	Colonic obstruction
	Cancer prophylaxis
	Colon dysplasia or cancer

FIGURE 319-12

Medical management of inflammatory bowel disease. 5-ASA, 5-aminosalicylic acid; CD, Crohn's disease; UC, ulcerative colitis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IBD AND PREGNANCY

Patients with quiescent UC and CD have normal fertility rates; the fallopian tubes can be scarred by the inflammatory process of CD, especially on the right side because of the proximity of the terminal ileum. In addition, perirectal, perineal, and rectovaginal abscesses and fistulae can result in dyspareunia. Infertility in men can be caused by sulfasalazine but reverses when treatment is stopped. In women who have an IPAA, most studies show that the fertility rate is reduced to about 50–80% of normal. This is due to scarring or occlusion of the fallopian tubes secondary to pelvic inflammation.

In mild or quiescent UC and CD, fetal outcome is nearly normal. The courses of CD and UC during pregnancy mostly correlate with disease activity at the time of conception. Patients should be in remission for 6 months before conceiving. Most CD patients can deliver vaginally, but cesarean delivery may be the preferred route of delivery for patients with anorectal and perirectal abscesses and fistulas to reduce the likelihood of fistulas developing or extending into the episiotomy scar. Unless they desire multiple children, UC patients with an IPAA should consider a cesarean delivery due to an increased risk of future fecal incontinence.

Sulfasalazine, Lialda, Apriso, Delzicol, balsalazide and now Asacol HD since the DBP has been removed from the capsule are safe for use in pregnancy and nursing with the caveat that additional folate supplementation must be given with sulfasalazine. Topical 5-ASA agents are safe during pregnancy and nursing. Glucocorticoids are generally safe for use during pregnancy and are indicated for patients with moderate to severe disease activity. The amount of glucocorticoids received by the nursing infant is minimal. The safest antibiotics to use for CD in pregnancy for short periods of time (weeks, not months) are ampicillin and cephalosporins. Metronidazole can be used in the second or third trimester. Ciprofloxacin causes cartilage lesions in immature animals and should be avoided because of the absence of data on its effects on growth and development in humans.

6-MP and azathioprine pose minimal or no risk during pregnancy, but experience is limited. If the patient cannot be weaned from the drug or has an exacerbation that requires 6-MP/azathioprine during pregnancy, she should continue the drug with informed consent. Breast milk has been shown to contain negligible levels of 6-MP/azathioprine when measured in a limited number of patients.

Little data exist on CSA in pregnancy. In a small number of patients with severe IBD treated with IV CSA during pregnancy, 80% of pregnancies were successfully completed without development of renal toxicity or congenital malformations. However, because of the lack of data, CSA should probably be avoided unless the patient would otherwise require surgery.

MTX is contraindicated in pregnancy and nursing. In a large prospective and multiple retrospective studies, no increased risk of stillbirths, miscarriages, or spontaneous abortions was seen with infliximab, adalimumab, or certolizumab. Infliximab and adalimumab are IgG1 antibodies and are actively transported across the placenta in the late second and third trimester. Infants can have serum levels of infliximab and adalimumab up to 12 months of age, and live vaccines should be avoided during this time. Certolizumab crosses the placenta by passive diffusion, and infant serum and cord blood levels are minimal. The anti-TNF drugs are relatively safe in nursing. Miniscule levels of infliximab, adalimumab, and certolizumab have been reported in breast milk. These levels are of no clinical significance. It is recommended that drugs should not be switched during pregnancy unless necessitated by the medical condition of the IBD. Vedolizumab and natalizumab appear safe during pregnancy although the data are limited. There are very little data available on ustekinemab use during pregnancy.

Surgery in UC should be performed only for emergency indications, including severe hemorrhage, perforation, and megacolon refractory to medical therapy. Total colectomy and ileostomy carry a 50% risk of postoperative spontaneous abortion. Fetal mortality is also high in CD requiring surgery. Patients with IPAAs have increased nighttime stool frequency during pregnancy that resolves postpartum. Transient small-bowel obstruction or ileus has been noted in up to 8% of patients with ileostomies.

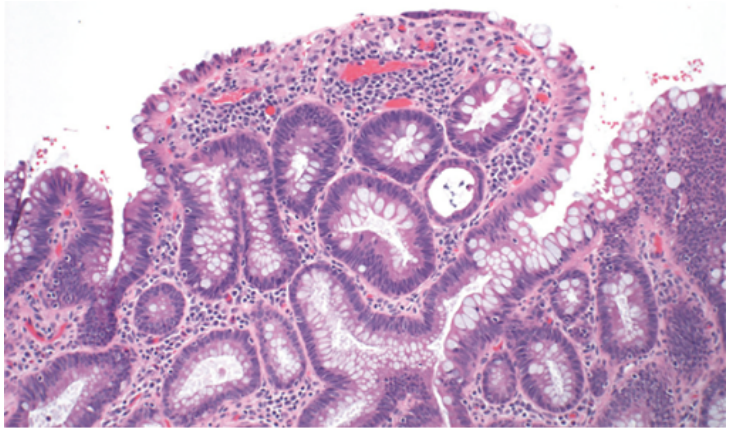
CANCER IN IBD

ULCERATIVE COLITIS

Patients with long-standing UC are at increased risk for developing colonic epithelial dysplasia and carcinoma **(Fig. 319-13)**.

FIGURE 319-13

Medium-power view of low-grade dysplasia in a patient with chronic ulcerative colitis. Low-grade dysplastic crypts are interspersed among regenerating crypts. *(Courtesy of Dr. R. Odze, Division of Gastrointestinal Pathology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The risk of neoplasia in chronic UC increases with duration and extent of disease. In contrast to the relatively high risk in one large meta-analysis (2% after 10 years, 8% after 20 years, and 18% after 30 years of disease), a decrease in the risk of colorectal cancer has been noted over time potentially due to better control of inflammation, better colonoscopic surveillance, more frequent colectomies and use of 5-ASA chemoprophylaxis. The rates of colon cancer are still about 1.5 to 2 × higher than in the general population, and colonoscopic surveillance is the standard of care.

Annual or biennial colonoscopy with multiple biopsies is recommended for patients with >8–10 years of extensive colitis (greater than one-third of the colon involved) or 12–15 years of proctosigmoiditis (less than one-third but more than just the rectum) and has been widely used to screen and survey for subsequent dysplasia and carcinoma. Risk factors for cancer in UC include long-duration disease, extensive disease, family history of colon cancer, PSC, a colon stricture, and the presence of post inflammatory pseudopolyps on colonoscopy.

CROHN'S DISEASE

Risk factors for developing cancer in Crohn's colitis are long-duration and extensive disease, bypassed colon segments, colon strictures, PSC, and family history of colon cancer. The cancer risks in CD and UC are probably equivalent for similar extent and duration of disease. In the CESAME study, a prospective observational cohort of IBD patients in France, the standardized incidence ratios of colorectal cancer were 2.2 for all IBD patients (95% confidence interval [CI], 1.5–3.0; $p < 0.001$) and 7.0 for patients with long-standing extensive colitis (both Crohn's and UC) (95% CI, 4.4–10.5; $p < 0.001$).

Thus, the same endoscopic surveillance strategy used for UC is recommended for patients with chronic Crohn's colitis. A pediatric colonoscope can be used to pass narrow strictures in CD patients, but surgery should be considered in symptomatic patients with impassable strictures.

MANAGEMENT OF DYSPLASIA AND CANCER

Dysplasia can be flat or polypoid. If flat high-grade dysplasia is encountered on colonoscopic surveillance, the usual treatment is colectomy for UC and either colectomy or segmental resection for CD. If flat low-grade dysplasia is found (Fig. 319-13), most investigators recommend immediate colectomy. Adenomas may occur coincidently in UC and CD patients with chronic colitis and can be removed endoscopically provided that biopsies of the surrounding mucosa are free of dysplasia. High-definition and high-magnification colonoscopes and dye sprays have increased the rate of dysplasia detection.

IBD patients are also at greater risk for other malignancies. Patients with CD may have an increased risk of NHL, leukemia, and myelodysplastic syndromes. Severe, chronic, complicated perianal disease in CD patients may be associated with an increased risk of cancer in the lower rectum and anal canal (squamous cell cancers). Although the absolute risk of small-bowel adenocarcinoma in CD is low (2.2% at 25 years in one study), patients with long-standing, extensive, small-bowel disease should consider screening.

FURTHER READING

Beaugerie L, Itzkowitz SH: Cancers complicating inflammatory bowel disease. N Engl J Med 372:15, 2014.

Bonovas S: Biologic therapies and risk of infection and malignancy in patients with inflammatory bowel disease: a systemic review and network meta-analysis. Clin Gastroenetrol Hepatol 14:1385, 2016.

Columbel JF et al: The safety of vedolizumab for ulcerative colitis and Crohn's disease. Gut 66:839, 2017.

Ha C, Kornbluth A: A critical review of biosimilars in IBD: The confluence of biologic drug development, regulatory requirements, clinical outcomes, and big business. Inflamm Bowel Dis 22:2513, 2016. [PubMed: 27564646]

Julsgaard M: Concentrations of adalimumab and infliximab in mothers and newborns and effects on infection. Gastroenterology 151:110, 2016. [PubMed: 27063728]

Ng SC: Geographical variability and environmental risk factors in inflammatory bowel disease. Gut 62:630, 2013. [PubMed: 23335431]

Panes J, Jairath V, Levesque BG: Advances in use of endoscopy, radiology and biomarkers to monitor inflammatory bowel disease. Gastroenterology 152:362, 2017. [PubMed: 27751880]

Regueiro M: Infliximab reduces endoscopic, but not clinical, recurrence of Crohn's disease after ileocolonic resection. Gastroenterology 150:1568, 2016. [PubMed: 26946343]

Sandborn WJ: Ustekinumab induction and maintenance therapy in refractory Crohn's disease. N Engl J Med 367:12, 2012.

Sandborn WJ: Tofactitinib as induction and maintenance therapy for ulcerative colitis. N Engl J Med 376:1723, 2017. [PubMed: 28467869]

Chapter 320: Irritable Bowel Syndrome

Chung Owyang

INTRODUCTION

Irritable bowel syndrome (IBS) is a functional bowel disorder characterized by abdominal pain or discomfort and altered bowel habits in the absence of detectable structural abnormalities. No clear diagnostic markers exist for IBS; thus the diagnosis of the disorder is based on clinical presentation. In 2016, the Rome III criteria for the diagnosis of IBS were updated to Rome IV **(Table 320-1)**. Throughout the world, about 10–20% of adults and adolescents have symptoms consistent with IBS, and most studies show a female predominance. IBS symptoms tend to come and go over time and often overlap with other functional disorders such as fibromyalgia, headache, backache, and genitourinary symptoms. Severity of symptoms varies and can significantly impair quality of life, resulting in high health care costs. Advances in basic, mechanistic, and clinical investigations have improved our understanding of this disorder and its physiologic and psychosocial determinants. Altered gastrointestinal (GI) motility, visceral hyperalgesia, disturbance of brain–gut interaction, abnormal central processing, autonomic and hormonal events, genetic and environmental factors, and psychosocial disturbances are variably involved, depending on the individual. This progress may result in improved methods of treatment.

TABLE 320-1

Rome IV Diagnostic Criteria for Irritable Bowel Syndrome[a]

Recurrent abdominal pain, on average, at least 1 day per week in the last 3 months, associated with ≥2 of the following criteria:

1. Related to defecation
2. Associated with a change in frequency of stool
3. Associated with a change in form (appearance) of stool

[a]Criteria fulfilled for the last 3 months with symptom onset at least 6 months prior to diagnosis.

CLINICAL FEATURES

IBS is a disorder that affects all ages, although most patients have their first symptoms before age 45. Older individuals have a lower reporting frequency. Women are diagnosed with IBS two to three times as often as men and make up 80% of the population with severe IBS. As indicated in Table 320-1, pain is a key symptom for the diagnosis of IBS. This symptom should be associated with defecation and/or have their onset associated with a change in frequency or form of stool. In comparison to Rome III, the Rome IV criteria is more stringent, requiring abdominal pain to occur at a minimum of once a week and eliminates "discomfort" as one of the criteria. Painless diarrhea or constipation does not fulfill the diagnostic criteria to be classified as IBS. Supportive symptoms that are not part of the diagnostic criteria include defecation straining, urgency or a feeling of incomplete bowel movement, passing mucus, and bloating.

Abdominal Pain

According to the current IBS diagnostic criteria, abdominal pain is a prerequisite clinical feature of IBS. Abdominal pain in IBS is highly variable in intensity and location. It is frequently episodic and crampy, but it may be superimposed on a background of constant ache. Pain may be mild enough to be ignored or it may interfere with daily activities. Despite this, malnutrition due to inadequate caloric intake is exceedingly rare with IBS. Sleep deprivation is also unusual because abdominal pain is almost uniformly present only during waking hours. However, patients with severe IBS frequently wake repeatedly during the night; thus, nocturnal pain is a poor discriminating factor between organic and functional bowel disease. Pain is often exacerbated by eating or emotional stress and improved by passage of flatus or stools. In addition, female patients with IBS commonly report worsening symptoms during the premenstrual and menstrual phases.

Altered Bowel Habits

Alteration in bowel habits is the most consistent clinical feature in IBS. The most common pattern is constipation alternating with diarrhea, usually with one of these symptoms predominating. At first, constipation may be episodic, but eventually it becomes continuous and increasingly intractable to treatment with laxatives. Stools are usually hard with narrowed caliber, possibly reflecting excessive dehydration caused by prolonged colonic retention and spasm. Most patients also experience a sense of incomplete evacuation, thus leading to repeated attempts at defecation in a short time span. Patients whose predominant symptom is constipation may have weeks or months of constipation interrupted with brief periods of diarrhea. In other patients, diarrhea may be the predominant symptom. Diarrhea resulting from IBS usually consists of small volumes of loose stools. Most patients have stool volumes of <200 mL. Nocturnal diarrhea does not occur in IBS. Diarrhea may be aggravated by emotional stress or eating. Stool may be accompanied by passage of large amounts of mucus. Bleeding is not a feature of IBS unless hemorrhoids are present, and malabsorption or weight loss does not occur.

Bowel pattern subtypes are highly unstable. In a patient population with ~33% prevalence rates of IBS-diarrhea predominant (IBS-D), IBS-constipation predominant (IBS-C), and IBS-mixed (IBS-M) forms, 75% of patients change subtypes and 29% switch between IBS-C and IBS-D over 1 year. The heterogeneity and

variable natural history of bowel habits in IBS increase the difficulty of conducting pathophysiology studies and clinical trials.

Gas and Flatulence

Patients with IBS frequently complain of abdominal distention and increased belching or flatulence, all of which they attribute to increased gas. Although some patients with these symptoms actually may have a larger amount of gas, quantitative measurements reveal that most patients who complain of increased gas generate no more than a normal amount of intestinal gas. Most IBS patients have impaired transit and tolerance of intestinal gas loads. In addition, patients with IBS tend to reflux gas from the distal to the more proximal intestine, which may explain the belching. Some patients with bloating may also experience visible distention with increase in abdominal girth. Both symptoms are more common among female patients and in those with higher overall Somatic Symptom Checklist scores.

Upper GI Symptoms

Between 25 and 50% of patients with IBS complain of dyspepsia, heartburn, nausea, and vomiting. This suggests that other areas of the gut apart from the colon may be involved. Prolonged ambulant recordings of small-bowel motility in patients with IBS show a high incidence of abnormalities in the small bowel during the diurnal (waking) period; nocturnal motor patterns are not different from those of healthy controls. The overlap between dyspepsia and IBS is great. The prevalence of IBS is higher among patients with dyspepsia (31.7%) than among those who reported no symptoms of dyspepsia (7.9%). Conversely, among patients with IBS, 55.6% reported symptoms of dyspepsia. In addition, the functional abdominal symptoms can change over time. Those with predominant dyspepsia or IBS can flux between the two. Although the prevalence of functional GI disorders is stable over time, the turnover in symptom status is high. Many episodes of symptom disappearance are due to subjects changing symptoms rather than total symptom resolution. Thus it is conceivable that functional dyspepsia and IBS are two manifestations of a single, more extensive digestive system disorder. Furthermore, IBS symptoms are prevalent in noncardiac chest pain patients, suggesting overlap with other functional gut disorders.

PATHOPHYSIOLOGY

The pathogenesis of IBS is poorly understood, although roles of abnormal gut motor and sensory activity, central neural dysfunction, psychological disturbances, mucosal inflammation, stress, and luminal factors such as bile acid malabsorption and gut dysbiosis have been proposed **(Fig. 320-1)**.

GI Motor Abnormalities

Studies of colonic myoelectrical and motor activity under unstimulated conditions have not shown consistent abnormalities in IBS. In contrast, colonic motor abnormalities are more prominent under stimulated conditions in IBS. IBS patients may exhibit increased rectosigmoid motor activity for up to 3 h after eating. Similarly, inflation of rectal balloons both in IBS-D and IBS-C patients leads to marked and

prolonged distention-evoked contractile activity. Recordings from the transverse, descending, and sigmoid colon showed that the motility index and peak amplitude of high-amplitude propagating contractions (HAPCs) in diarrhea-prone IBS patients were greatly increased compared to those in healthy subjects and were associated with rapid colonic transit and accompanied by abdominal pain.

Visceral Hypersensitivity

As with studies of motor activity, IBS patients frequently exhibit exaggerated sensory responses to visceral stimulation. The frequency of perceptions of food intolerance is at least twofold more common than in the general population. Postprandial pain has been temporally related to entry of the food bolus into the cecum in 74% of patients. On the other hand, prolonged fasting in IBS patients is often associated with significant improvement in symptoms. Rectal balloon inflation produces nonpainful and painful sensations at lower volumes in IBS patients than in healthy controls without altering rectal tension, suggestive of visceral afferent dysfunction in IBS. Similar studies show gastric and esophageal hypersensitivity in patients with nonulcer dyspepsia and noncardiac chest pain, raising the possibility that these conditions have a similar pathophysiologic basis. Lipids lower the thresholds for the first sensation of gas, discomfort, and pain in IBS patients. Hence, postprandial symptoms in IBS patients may be explained in part by a nutrient-dependent exaggerated sensory component of the gastrocolonic response. In contrast to enhanced gut sensitivity, IBS patients do not exhibit heightened sensitivity elsewhere in the body. Thus, the afferent pathway disturbances in IBS appear to be selective for visceral innervation with sparing of somatic pathways. The mechanisms responsible for visceral hypersensitivity are still under investigation. It has been proposed that these exaggerated responses may be due to (1) increased end-organ sensitivity with recruitment of “silent” nociceptors; (2) spinal hyperexcitability with activation of nitric oxide and possibly other neurotransmitters; (3) endogenous (cortical and brainstem) modulation of caudad nociceptive transmission; and (4) over time, the possible development of long-term hyperalgesia due to development of neuroplasticity, resulting in permanent or semipermanent changes in neural responses to chronic or recurrent visceral stimulation.

Central Neural Dysregulation

The role of central nervous system (CNS) factors in the pathogenesis of IBS is strongly suggested by the clinical association of emotional disorders and stress with symptom exacerbation and the therapeutic response to therapies that act on cerebral cortical sites. Functional brain imaging studies such as magnetic resonance imaging (MRI) have shown that in response to distal colonic stimulation, the mid-cingulate cortex —a brain region concerned with attention processes and response selection—shows greater activation in IBS patients. Modulation of this region is associated with changes in the subjective unpleasantness of pain. In addition, IBS patients also show preferential activation of the prefrontal lobe, which contains a vigilance network within the brain that increases alertness. These may represent a form of cerebral dysfunction leading to the increased perception of visceral pain.

Abnormal Psychological Features

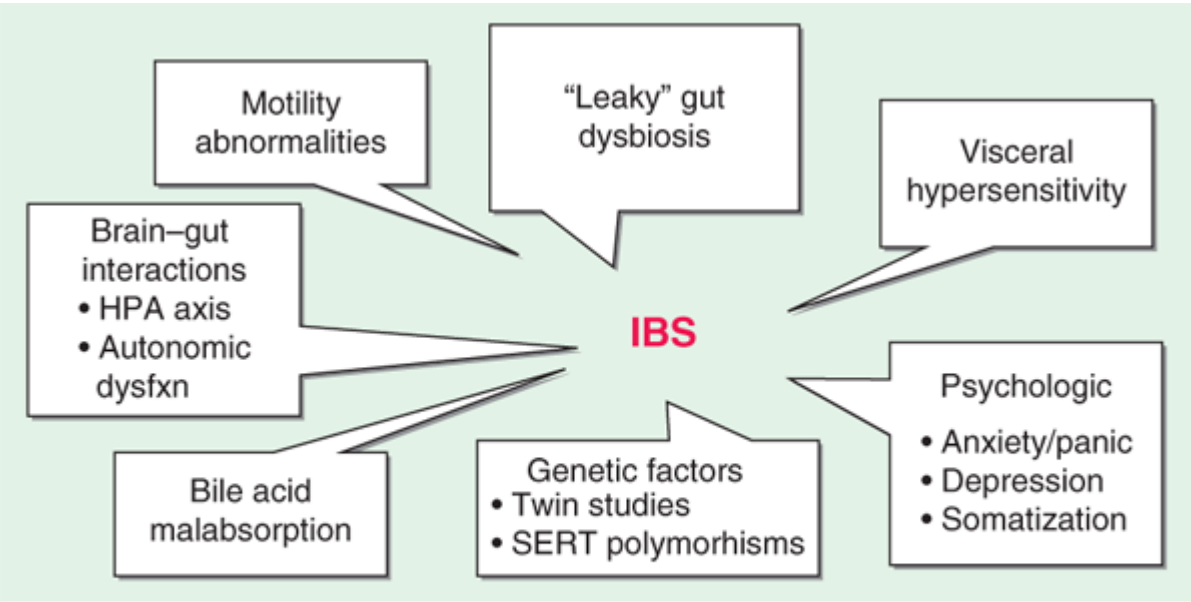
Abnormal psychiatric features are recorded in up to 80% of IBS patients, especially in referral centers; however, no single psychiatric diagnosis predominates. Most of these patients demonstrated exaggerated symptoms in response to visceral distention, and this abnormality persists even after exclusion of psychological factors.

Psychological factors influence pain thresholds in IBS patients, as stress alters sensory thresholds. An association between prior sexual or physical abuse and development of IBS has been reported. Clinical studies suggest that IBS has a strong developmental component which involves interactions of genetic and epigenetic factors early in life. These may modulate brain networks related to emotional arousal and/or central autonomic control, salience and somatosensory integration. Abuse is associated with greater pain reporting, psychological distress, and poor health outcome. Brain functional MRI studies show greater activation of the posterior and middle dorsal cingulate cortex, which is implicated in affect processing in IBS patients with a past history of sexual abuse.

Thus, patients with IBS frequently demonstrate increased motor reactivity of the colon and small bowel to a variety of stimuli and altered visceral sensation associated with lowered sensation thresholds. These may result from CNS–enteric nervous system dysregulation.

FIGURE 320-1

Pathophysiology of IBS. The cause of IBS is likely to be multifactorial. Patients often show evidence of visceral hypersensitivity and motility abnormalities. Many IBS patients have increased anxiety and/or depression and their symptoms are often exacerbated by mental or physical stress suggesting abnormal brain–gut interaction. Genetic studies suggest a few IBS patients may have genetic abnormalities affecting the serotonin transport system in the enteric nerves. Up to 30% of IBS patients may have bile acid malabsorption. Gut dysbiosis and impaired mucosa permeability also have been reported in many IBS patients. This may lead to subclinical mucosa inflammation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Postinfectious IBS

IBS may be induced by GI infection. In an investigation of 544 patients with confirmed bacterial gastroenteritis, one-quarter developed IBS subsequently. Conversely, about a third of IBS patients experienced an acute “gastroenteritis-like” illness at the onset of their chronic IBS symptomatology. This group of “postinfective” IBS occurs more commonly in females and affects younger rather than older patients. Risk factors for developing postinfectious IBS include, in order of importance, prolonged duration of initial illness, toxicity of infecting bacterial strain, smoking, mucosal markers of inflammation, female sex, depression, hypochondriasis, and adverse life events in the preceding 3 months. Age older than 60 years might protect against postinfectious IBS, whereas treatment with antibiotics has been associated with increased risk. The microbes involved in the initial infection are *Campylobacter, Salmonella*, and *Shigella*. Those patients with *Campylobacter* infection who are toxin-positive are more likely to develop postinfective IBS. Increased rectal mucosal enteroendocrine cells, T lymphocytes, and increased gut permeability are acute changes following *Campylobacter* enteritis that could persist for more than a year and may contribute to postinfective IBS.

Immune Activation and Mucosal Inflammation

Some patients with IBS display persistent signs of low-grade mucosal inflammation with activated lymphocytes, mast cells, and enhanced expression of proinflammatory cytokines. Other studies also indicate that peripheral blood mononuclear cells (PBMCs) from IBS patients show abnormal release of proinflammatory cytokines such as IL6, IL1β, and TNF. These abnormalities may contribute to abnormal epithelial secretion and visceral hypersensitivity. There is increasing evidence that some members of the superfamily of transient receptor potential (TRP) cation channels such as TRPV1 (vanilloid) channels are central to the initiation and persistence of visceral hypersensitivity. Mucosal inflammation can lead to increased expression of TRPV1 in the enteric nervous system. Enhanced expression of TRPV1 channels in the sensory neurons of the gut has been observed in IBS, and such expression appears to correlate with visceral hypersensitivity and abdominal pain. Interestingly, clinical studies have also shown increased intestinal permeability in patients with IBS-D. Psychological stress and anxiety can increase the release of proinflammatory cytokine, and this in turn may alter intestinal permeability. A clinical study shows 39% of IBS-D patients had increased intestinal permeability as measured by the lactulose/mannitol ratio. These IBS patients also demonstrated higher Functional Bowel Disorder Severity Index (FBDSI) score and increased hypersensitivity to visceral nociceptive pain stimuli. This provides a functional link between psychological stress, immune activation, and symptom generation in patients with IBS.

Altered Gut Flora

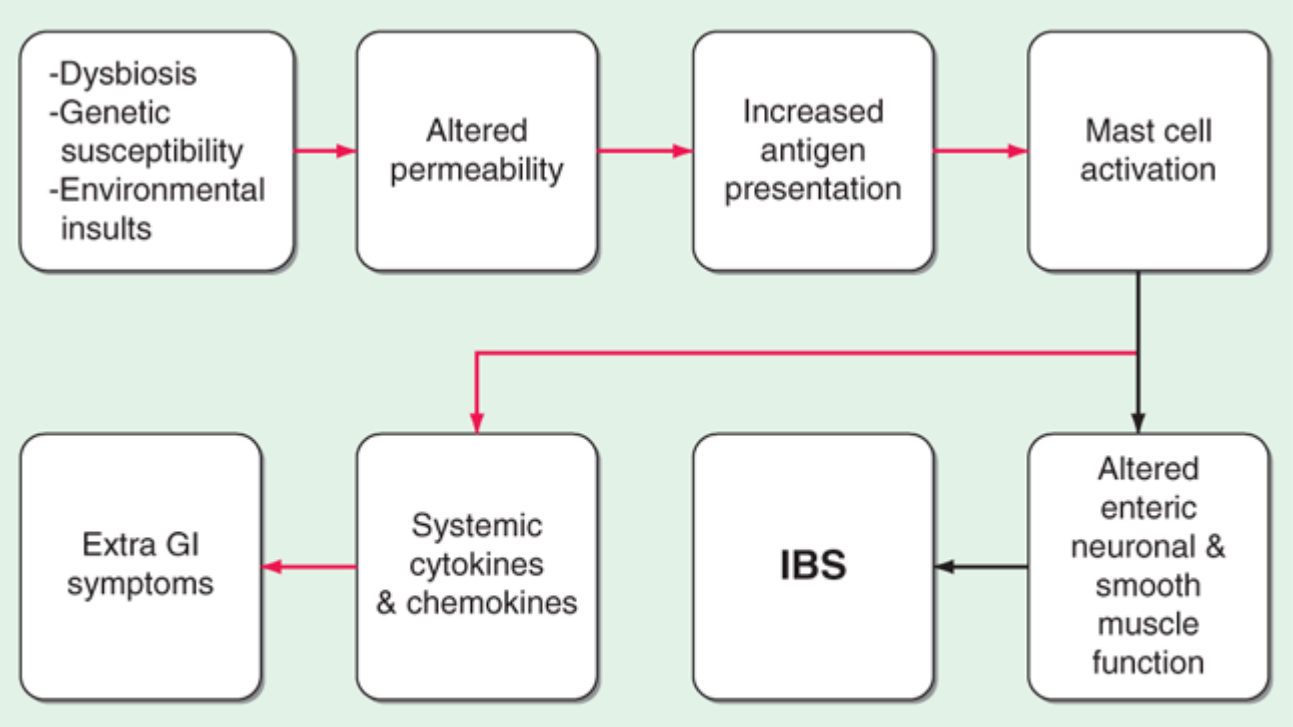
A high prevalence of small intestinal bacterial overgrowth in IBS patients has been noted based on positive lactulose hydrogen breath test. This finding, however, has been challenged by a number of other studies that found no increased incidence of bacterial overgrowth based on jejunal aspirate culture. Abnormal H_2 breath

test can occur because of small-bowel rapid transit and may lead to erroneous interpretation. Hence, the role of testing for small intestinal bacterial overgrowth in IBS patients remains unclear.

Studies using culture-independent approaches such as 16S rRNA gene-based analysis found significant differences between the molecular profile of the fecal microbiota of IBS patients and that of healthy subjects. Twenty-two studies, comprising 827 subjects, reported significant changes in the microbial communities of healthy individuals versus patients with different subtypes of IBS. Despite a lack of consensus on the exact microbial differences between IBS patients and controls, in general IBS patients had decreased proportions of the genera *Bifidobacterium* and *Lactobacillus* and increased ratios of Firmicutes:Bacteroidetes. It has been speculated that these changes may be related to stress and diet. A temporary reduction in lactobacilli has been reported in animal models of early-life stress. On the other hand, Firmicutes is the dominant phylum in adults consuming a diet high in animal fat and protein. It is conceivable that gut dysbiosis acting in concert with genetic susceptibility and environmental insults may alter mucosal permeability and increase antigen presentation to the immune cells in the lamina propria. This may result in mast cell activation and altered enteric neuronal and smooth muscle function causing IBS symptoms. In addition, release of cytokines and chemokines from mucosal inflammation may generate extra GI symptoms such as chronic fatigue, muscle pain, and anxiety **(Fig. 320-2).**

FIGURE 320-2

Gut dysbiosis and IBS. Gut dysbiosis acting in concert with genetic and environmental factors may alter intestinal permeability, increase antigen presentation resulting in mast cell activation. Products of mast cell degranulation may alter neuronal and smooth muscle function causing IBS symptoms. The cytokines and chemokines generated from mucosal inflammation may cause symptoms such as fibromyalgia, chronic fatigue, and mood changes. (*Adapted from NJ Talley, AA Fodor: Gastroenterology 141: 1555, 2011.*)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Abnormal Serotonin Pathways

The serotonin (5-HT)-containing enterochromaffin cells in the colon are increased in a subset of IBS-D patients compared to healthy individuals or patients with ulcerative colitis. Furthermore, postprandial plasma 5-HT levels were significantly higher in this group of patients compared to healthy controls. Tryptophan hydroxylase 1 (TPH1) is the rate-limiting enzyme in enterochromaffin cell 5-HT biosynthesis, functional TPH1 polymorphism has been shown to be associated with IBS habit subtypes. Because serotonin plays an important role in the regulation of GI motility and visceral perception, the increased release of serotonin may contribute to the postprandial symptoms of these patients and provides a rationale for the use of serotonin antagonists in the treatment of this disorder.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Irritable Bowel Syndrome

Because IBS is a disorder for which no pathognomonic abnormalities have been identified, its diagnosis relies on recognition of positive clinical features and elimination of other organic diseases. Symptom-based criteria have been developed for the purpose of differentiating patients with IBS from those with organic diseases. These include the Manning, Rome I, Rome II, Rome III, and Rome IV criteria. Rome IV criteria for the diagnosis of IBS were published in 2016 (Table 320-1) and defined IBS on the basis of abdominal pain and altered bowel habits that occur with sufficient frequency in affected patients. A careful history and physical

examination are frequently helpful in establishing the diagnosis. Clinical features suggestive of IBS include the following: recurrence of lower abdominal pain with altered bowel habits over a period of time without progressive deterioration, onset of symptoms during periods of stress or emotional upset, absence of other systemic symptoms such as fever and weight loss, and small-volume stool without any evidence of blood.

On the other hand, the appearance of the disorder for the first time in old age, progressive course from time of onset, persistent diarrhea after a 48-h fast, and presence of nocturnal diarrhea or steatorrheal stools argue against the diagnosis of IBS.

Because the major symptoms of IBS—abdominal pain, abdominal bloating, and alteration in bowel habits—are common complaints of many GI organic disorders, the list of differential diagnoses is a long one. The quality, location, and timing of pain may be helpful to suggest specific disorders. Pain due to IBS that occurs in the epigastric or periumbilical area must be differentiated from biliary tract disease, peptic ulcer disorders, intestinal ischemia, and carcinoma of the stomach and pancreas. If pain occurs mainly in the lower abdomen, the possibility of diverticular disease of the colon, inflammatory bowel disease (including ulcerative colitis and Crohn's disease), and carcinoma of the colon must be considered. Postprandial pain accompanied by bloating, nausea, and vomiting suggests gastroparesis or partial intestinal obstruction. Intestinal infestation with *Giardia lamblia* or other parasites may cause similar symptoms. When diarrhea is the major complaint, the possibility of lactase deficiency, laxative abuse, malabsorption, celiac sprue, hyperthyroidism, inflammatory bowel disease, and infectious diarrhea must be ruled out. On the other hand, constipation may be a side effect of many different drugs, such as anticholinergic, antihypertensive, and antidepressant medications. Endocrinopathies such as hypothyroidism and hypoparathyroidism must also be considered in the differential diagnosis of constipation, particularly if other systemic signs or symptoms of these endocrinopathies are present. In addition, acute intermittent porphyria and lead poisoning may present in a fashion similar to IBS, with painful constipation as the major complaint. These possibilities are suspected on the basis of their clinical presentations and are confirmed by appropriate serum and urine tests.

Few tests are required for patients who have typical IBS symptoms and no alarm features. Unnecessary investigations may be costly and even harmful. The American Gastroenterological Association has delineated factors to be considered when determining the aggressiveness of the diagnostic evaluation. These include the duration of symptoms, the change in symptoms over time, the age and sex of the patient, the referral status of the patient, prior diagnostic studies, a family history of colorectal malignancy, and the degree of psychosocial dysfunction. Thus, a younger individual with mild symptoms requires a minimal diagnostic evaluation, while an older person or an individual with rapidly progressive symptoms should undergo a more thorough exclusion of organic disease. Most patients should have a complete blood count and sigmoidoscopic examination; in addition, stool specimens should be examined for ova and parasites in those who have diarrhea. In patients with persistent diarrhea not responding to simple antidiarrheal agents, a sigmoid colon biopsy should be performed to rule out microscopic colitis. In those age >40 years, an air-contrast barium enema or colonoscopy should also be performed. If the main symptoms are diarrhea and increased gas, the possibility of lactase deficiency should be ruled out with a hydrogen breath test or with

evaluation after a 3-week lactose-free diet. Excessive gas with bloating also raises the possibility of small bowel bacteria overgrowth and should be ruled out with a glucose hydrogen breath test. Some patients with IBS-D may have undiagnosed celiac sprue. Because the symptoms of celiac sprue respond to a gluten-free diet, testing for celiac sprue in IBS may prevent years of morbidity and attendant expense. Decision-analysis studies show that serology testing for celiac sprue in patients with IBS-D has an acceptable cost when the prevalence of celiac sprue is >1% and is the dominant strategy when the prevalence is >8%. In patients with concurrent symptoms of dyspepsia, upper GI radiographs or esophagogastroduodenoscopy may be advisable. In patients with postprandial right upper quadrant pain, an ultrasonogram of the gallbladder should be obtained. Laboratory features that argue against IBS include evidence of anemia, elevated sedimentation rate, presence of leukocytes or blood in stool, and stool volume >200–300 mL/d. These findings would necessitate other diagnostic considerations.

TREATMENT

TREATMENT

Irritable Bowel Syndrome

Patient Counseling and Dietary Alterations

Reassurance and careful explanation of the functional nature of the disorder and of how to avoid obvious food precipitants are important first steps in patient counseling and dietary change. Occasionally, a meticulous dietary history may reveal substances (such as coffee, disaccharides, legumes, and cabbage) that aggravate symptoms. Excessive fructose and artificial sweeteners, such as sorbitol or mannitol, may cause diarrhea, bloating, cramping, or flatulence. As a therapeutic trial, patients should be encouraged to eliminate any foodstuffs that appear to produce symptoms. However patients should avoid nutritionally depleted diets. A diet low in fermentable oligosaccharides, disaccharides, monosaccharides, and polyols (FODMAPs) **(Table 320-2)** has been shown to be helpful in IBS patients (see Low FODMAP Diet).

Stool-Bulking Agents

High-fiber diets and bulking agents, such as bran or hydrophilic colloid, are frequently used in treating IBS. The water-holding action of fibers may contribute to increased stool bulk because of the ability of fiber to increase fecal output of bacteria. Fiber also speeds up colonic transit in most persons. In diarrhea-prone patients, whole-colonic transit is faster than average; however, dietary fiber can delay transit. Furthermore, because of their hydrophilic properties, stool-bulking agents bind water and thus prevent both excessive hydration and dehydration of stool. The latter observation may explain the clinical experience that a high-fiber diet relieves diarrhea in some IBS patients. Fiber supplementation with psyllium has been shown to reduce perception of rectal distention, indicating that fiber may have a positive effect on visceral afferent function.

The beneficial effects of dietary fiber on colonic physiology suggest that dietary fiber should be an effective treatment for IBS patients, but controlled trials of dietary fiber have produced variable results. This is not surprising since IBS is a heterogeneous disorder, with some patients being constipated and other having

predominant diarrhea. Most investigations report increases in stool weight, decreases in colonic transit times, and improvement in constipation. Others have noted benefits in patients with alternating diarrhea and constipation, pain, and bloating. However, most studies observe no responses in patients with diarrhea- or pain-predominant IBS. It is possible that different fiber preparations may have dissimilar effects on selected symptoms in IBS. A cross-over comparison of different fiber preparations found that psyllium produced greater improvements in stool pattern and abdominal pain than bran. Furthermore, psyllium preparations tend to produce less bloating and distention. Despite the equivocal data regarding efficacy, most gastroenterologists consider stool-bulking agents worth trying in patients with IBS-C. Fiber should be started at a nominal dose and slowly titrated up as tolerated over the course of several weeks to a targeted dose of 20–30 g of total dietary and supplementary fiber per day. Even when used judiciously, fiber can exacerbate bloating, flatulence, constipation, and diarrhea.

Antispasmodics

Clinicians have observed that anticholinergic drugs may provide temporary relief for symptoms such as painful cramps related to intestinal spasm. Although controlled clinical trials have produced mixed results, evidence generally supports beneficial effects of anticholinergic drugs for pain. A meta-analysis of 26 double-blind clinical trials of antispasmodic agents in IBS reported better global improvement (62%) and abdominal pain reductions (64%) compared to placebo (35 and 45%, respectively), suggesting efficacy in some patients. The drugs are most effective when prescribed in anticipation of predictable pain. Physiologic studies demonstrate that anticholinergic drugs inhibit the gastrocolic reflex; hence, postprandial pain is best managed by giving antispasmodics 30 min before meals so that effective blood levels are achieved shortly before the anticipated onset of pain. Most anticholinergics contain natural belladonna alkaloids, which may cause xerostomia, urinary hesitancy and retention, blurred vision, and drowsiness. They should be used in the elderly with caution. Some physicians prefer to use synthetic anticholinergics such as dicyclomine that have less effect on mucous membrane secretions and produce fewer undesirable side effects.

Antidiarrheal Agents

Peripherally acting opiate-based agents are the initial therapy of choice for IBS-D. Physiologic studies demonstrate increases in segmenting colonic contractions, delays in fecal transit, increases in anal pressures, and reductions in rectal perception with these drugs. When diarrhea is severe, especially in the painless diarrhea variant of IBS, small doses of loperamide, 2–4 mg every 4–6 h up to a maximum of 12 g/d, can be prescribed. These agents are less addictive than paregoric, codeine, or tincture of opium. In general, the intestines do not become tolerant of the antidiarrheal effect of opiates, and increasing doses are not required to maintain antidiarrheal potency. These agents are most useful if taken before anticipated stressful events that are known to cause diarrhea. However, not infrequently, a high dose of loperamide may cause cramping because of increases in segmenting colonic contractions. Another antidiarrheal agent that may be used in IBS patients is the bile acid binder cholestyramine resin as up to 30% of IBS-D patients may have bile acid malabsorption.

Antidepressant Drugs

In addition to their mood-elevating effects, antidepressant medications have several physiologic effects that suggest they may be beneficial in IBS. In IBS-D patients, the tricyclic antidepressant imipramine slows jejunal migrating motor complex transit propagation and delays orocecal and whole-gut transit, indicative of a motor inhibitory effect. Some studies also suggest that tricyclic agents may alter visceral afferent neural function.

A number of studies indicate that tricyclic antidepressants may be effective in some IBS patients. In a 2-month study of desipramine, abdominal pain improved in 86% of patients compared to 59% given placebo. Another study of desipramine in 28 IBS patients showed improvement in stool frequency, diarrhea, pain, and depression. When stratified according to the predominant symptoms, improvements were observed in IBS-D patients, with no improvement being noted in IBS-C patients. The beneficial effects of the tricyclic compounds in the treatment of IBS appear to be independent of their effects on depression. The therapeutic benefits for the bowel symptoms occur faster and at a lower dosage. The efficacy of antidepressant agents in other chemical classes in the management of IBS is less well evaluated. In contrast to tricyclic agents, the selective serotonin reuptake inhibitor (SSRI) paroxetine accelerates orocecal transit, raising the possibility that this drug class may be useful in IBS-C patients. The SSRI citalopram blunts perception of rectal distention and reduces the magnitude of the gastrocolonic response in healthy volunteers. A small placebo-controlled study of citalopram in IBS patients reported reductions in pain. However, these findings could not be confirmed in another randomized controlled trial that showed that citalopram at 20 mg/d for 4 weeks was not superior to placebo in treating nondepressed IBS patients. Hence, the efficacy of SSRIs in the treatment of IBS needs further confirmation.

Antiflatulence Therapy

The management of excessive gas is seldom satisfactory, except when there is obvious aerophagia or disaccharidase deficiency. Patients should be advised to eat slowly and not chew gum or drink carbonated beverages. Bloating may decrease if an associated gut syndrome such as IBS or constipation is improved. If bloating is accompanied by diarrhea and worsens after ingesting dairy products, fresh fruits, vegetables, or juices, further investigation or a dietary exclusion trial may be worthwhile. Avoiding flatogenic foods, exercising, losing excess weight, and taking activated charcoal are safe but unproven remedies. A low FODMAP diet has been shown to be quite effective to reduce gas and bloating (see Low FODMAP Diet). Data regarding the use of surfactants such as simethicone are conflicting. Antibiotics may help in a subgroup of IBS patients with predominant symptoms of bloating. Beano, an over-the-counter oral β-glycosidase solution, may reduce rectal passage of gas without decreasing bloating and pain. Pancreatic enzymes reduce bloating, gas, and fullness during and after high-calorie, high-fat meal ingestion.

Modulation of Gut Flora

Because altered colonic flora (gut dysbiosis) may contribute to the pathogenesis of IBS, this has led to great interest in using antibiotics, prebiotics, and probiotics to treat IBS **(Fig. 320-3)**.

Antibiotics

Antibiotic treatment benefits a subset of IBS patients. In a double-blind, randomized, placebo-controlled study, neomycin dosed at 500 mg twice daily for 10 days was more effective than placebo at improving symptom scores among IBS patients. The nonabsorbed oral antibiotic rifaximin is the most thoroughly studies antibiotic of the treatment of IBS. In a double-blind, placebo-controlled study, patients receiving rifaximin at a dose of 550 mg two times daily for 2 weeks experienced substantial improvement of global IBS symptoms over placebo. Rifaximin is the only antibiotic with demonstrated sustained benefit beyond therapy cessation in IBS patients. The drug has a favorable safety and tolerability profile compared with systemic antibiotics. A systemic review and meta-analysis of five studies of IBS patients found that rifaximin is more effective than placebo for global symptoms and bloating (odds ratio 1.57) with a number needed to treat (NNT) of 10.2. The modest therapeutic gain was similar to that yielded by other current available therapies for IBS. However, currently there are still insufficient data to recommend routine use of this antibiotic in the treatment of IBS.

Prebiotics

These are nondigestible food ingredients that stimulate growth and/or activity of bacteria in the GI tract. There have been four randomized trials to examine the effects of prebiotics. Three of the four studies reported that prebiotics worsened or did not improve IBS symptoms. This is not surprising given the adverse effects of high carbohydrate diet on IBS symptoms.

Probiotics

These are defined as live microorganisms that when administered in adequate amounts confer a health benefit on the host. A meta-analysis of 10 probiotic studies in IBS patients found significant relief of pain and bloating with the use of *Bifidobacterium breve, B longum,* and *Lactobacillus acidophilus* species compared to placebo. However, there was no change in stool frequency or consistency. Large-scale studies of well-phenotyped IBS patients are needed to establish the efficacy of these probiotics.

Low Fodmap Diet

A diet rich in FODMAP (fermentable oligosaccharides, disaccharides, monosaccharides, and polyols) often triggers symptoms in IBS patients. FODMAPs are poorly absorbed by the small intestine and fermented by bacteria in the colon to produce gas and osmotically active carbohydrates **(Fig. 320-4)**. At the same time, on entering the colon, FODMAPs may serve as nutrient for the colonic bacteria and promote the growth of gram negative commensal bacteria which may induce epithelial damage and subclinical mucosa inflammation. Fructose and fructans induce IBS symptoms in a dose-dependent manner. In contrast, a low FODMAP diet reduces IBS symptoms. A randomized controlled trial showed a 4-week low FODMAP diet improved symptoms in 68% of IBS patients compared with 23% on a habitual diet. Low FOMAP diets appeared to be superior to national guidelines for IBS management. These observations were confirmed by a double blind controlled study of 30 IBS patients and 18 healthy controls, in which a low FODMAP diet significantly reduced bloating, pain, passage of gas, and diarrhea. A double blind randomized control trial involving 92 IBS patients showed >50% of patients on the low FODMAP diet had major improvement of their abdominal pain compared with 20% of the control group. These observations demonstrate the impressive efficacy of low

FODMAP diet for many IBS patients, and if confirmed may justify the recommendation of a low FODMAP diet as first-line treatment for IBS patients.

Serotonin Receptor Agonist and Antagonists

Serotonin receptor antagonists have been evaluated as therapies for IBS-D. Serotonin acting on 5-HT_3 receptors enhances the sensitivity of afferent neurons projecting from the gut. In humans, a 5-HT_3 receptor antagonist such as alosetron reduces perception of painful visceral stimulation in IBS. It also induces rectal relaxation, increases rectal compliance, and delays colonic transit. Meta-analysis of 14 randomized controlled trials of alosetron or cilansetron showed that these antagonists are more effective than placebo in achieving global improvement in IBS symptoms and relief of abdominal pain and discomfort. These agents are more likely to cause constipation in IBS patients with diarrhea alternating with constipation. Also, 0.2% of patients using 5-HT_3 antagonists developed ischemic colitis versus none in the control group. In postrelease surveillance, 84 cases of ischemic colitis were observed, including 44 cases that required surgery and 4 deaths. As a consequence, the medication was voluntarily withdrawn by the manufacturer in 2000. Alosetron has been reintroduced under a new risk-management program where patients have to sign a patient-physician agreement. This has significantly limited its usage.

Novel 5-HT_4 receptor agonists such as tegaserod exhibit prokinetic activity by stimulating peristalsis. In IBS patients with constipation, tegaserod accelerated intestinal and ascending colon transit. Clinical trials involving >4000 IBS-C patients reported reductions in discomfort and improvements in constipation and bloating, compared to placebo. Diarrhea is the major side effect. However, tegaserod has been withdrawn from the market; a meta-analysis revealed an increase in serious cardiovascular events.

Chloride Channel Activators

Lubiprostone is a bicyclic fatty acid that stimulates chloride channels in the apical membrane of intestinal epithelial cells. Chloride secretion induces passive movement of sodium and water into the bowel lumen and improves bowel function. Oral lubiprostone was effective in the treatment of patients with constipation-predominant IBS in large phase II and phase III randomized, double-blinded, placebo-controlled multicenter trials. Responses were significantly greater in patients receiving lubiprostone 8 µg twice daily for 3 months than in those receiving placebo. In general, the drug was quite well tolerated. The major side effects are nausea and diarrhea. Lubiprostone is a new class of compounds for treatment of chronic constipation with or without IBS.

Guanylate Cyclase-C Agonist

Linaclotide is a minimally absorbed 14-amino-acid peptide guanylate cyclase-C (GC-C) agonist that binds to and activates GC-C on the luminal surface of intestinal epithelium. Activation of GC-C results in generation of cyclic guanosine monophosphate (cGMP), which triggers secretion of fluid, sodium, and bicarbonate. In animal models, linaclotide accelerates GI transit and reduces visceral nociception. The analgesic action of linaclotide appears to be mediated by cGMP acting on afferent pain fibers innervating the GI tract. A phase III, double-blind, controlled trial showed that linaclotide, 290 µg given once daily, significantly improved

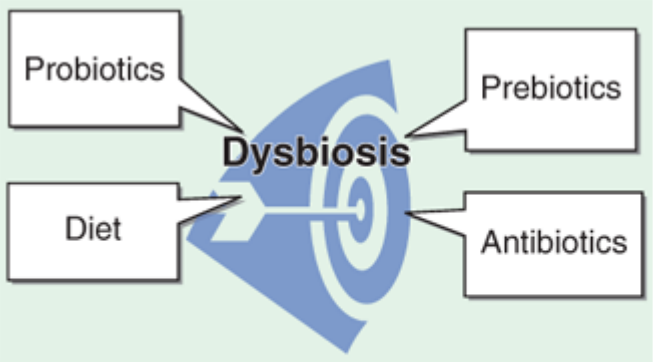
abdominal pain, bloating, and spontaneous bowel movement. The only significant side effect was diarrhea, which occurred in 4.5% of the patients. The drug has been approved for treatment of constipation in IBS-C patients.

Summary

The treatment strategy of IBS depends on the severity of the disorder **(Table 320-3)**. Most IBS patients have mild symptoms. They are usually cared for in primary care practices, have little or no psychosocial difficulties, and do not seek health care often. Treatment usually involves education, reassurance, and dietary/lifestyle changes. A smaller portion have moderate symptoms that are usually intermittent and correlate with altered gut physiology, e.g., worsened with eating or stress and relieved by defecation. For IBS-D patients, treatments include gut-acting pharmacologic agents such as antispasmodics, antidiarrheals, bile acid binders, and the newer gut serotonin modulators **(Table 320-4)**. In IBS-C patients, increased fiber intake and the use of osmotic agents such as polyethylene glycol may achieve satisfactory results. For patients with more severe constipation, a chloride channel opener (lubiprostone) or GC-C agonist (linaclotide) may be considered. For IBS patients with predominant gas and bloating, a low-FODMAP diet may provide significant relief. Some patients may benefit from probiotics and rifaximin treatment. A small proportion of IBS patients have severe and refractory symptoms, are usually seen in referral centers, and frequently have constant pain and psychosocial difficulties. This group of patients is best managed with antidepressants and other psychological treatments (Table 320-4). Clinical trials demonstrating success of low FODMAP diet in improving IBS symptoms and quality of life provide strong evidence supporting the use of this dietary approach in the treatment of IBS. These observations, if confirmed, may lead to the use of low FODMAP diet as the first line of treatment of IBS patients with moderate to severe symptoms.

FIGURE 320-3

Gut dysbiosis: a potential treatment target. Prebiotics, probiotics, and low FODMAP diet may be used to modulate gut flora and treat IBS.

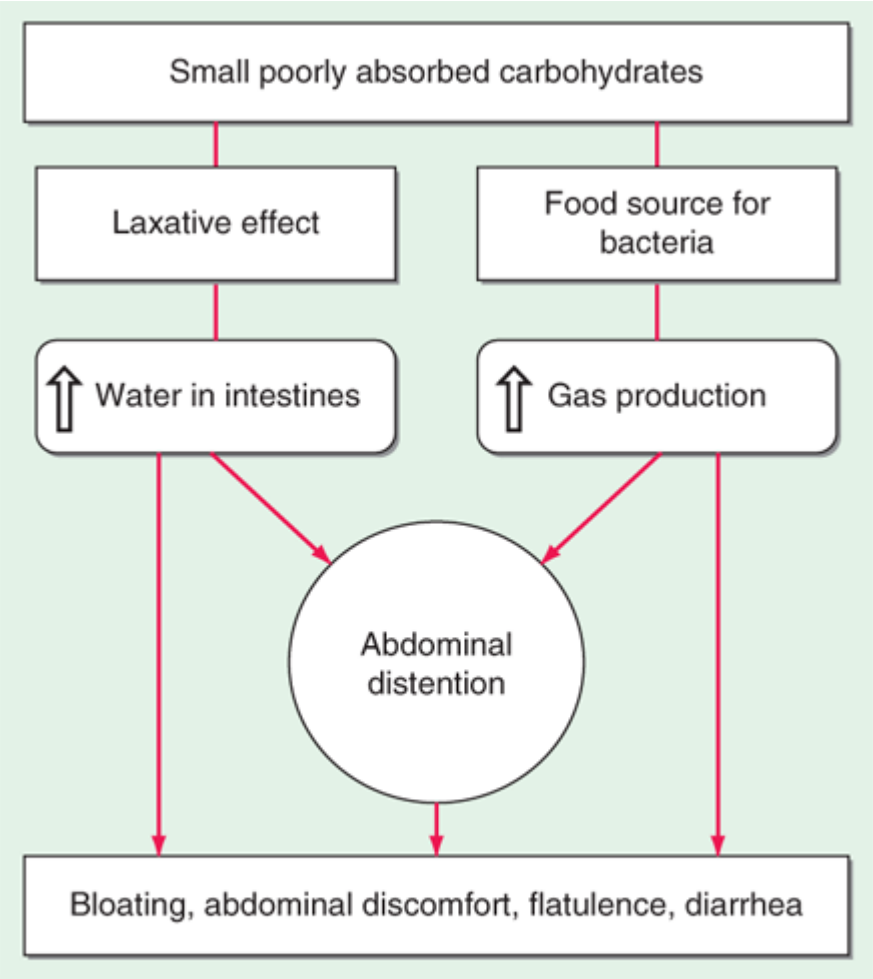


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 320-4

Pathogenesis of FODMAP-related symptoms. FODMAPs are poorly absorbed by the small intestine and fermented by gut bacteria to produce gas and osmotically active carbohydrates. These events act in concert to cause bloating, flatulence, and diarrhea. FODMAP may also serve as nutrients for colonic bacteria which

may induce mucosa inflammation. *(Figure created using data from http://www.nutritiontoyou.com/wp-content/uploads/2014/06/IBS-symptoms.png.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 320-2

Some Common Food Sources of FODMAPs

Food Type	Free Fructose	Lactose	Fructans	Galacto-oligosaccharides	Polyols
Fruits	Apple, cherry, mango, pear, watermelon		Peach, persimmon, watermelon		Apple, apricot, pear, avocado, blackberries, cherry, nectarine, plum, prune
Vegetables	Asparagus, artichokes, sugar snap peas		Artichokes, beetroot, Brussels sprout, chicory, fennel, garlic, leek, onion, peas		Cauliflower, mushroom, snow peas
Grains and cereals			Wheat, rye, barley		
Nuts and seeds			Pistachios		
Milk and milk products		Milk, yogurt, ice cream, custard, soft cheeses			
Legumes			Legumes, lentils, chickpeas	Legumes, chickpeas, lentils	
Other	Honey, high-fructose corn syrup		Chicory drinks		

Food Type	Free Fructose	Lactose	Fructans	Galacto-oligosaccharides	Polyols
Food additives			Inulin, FOS		Sorbitol, mannitol, maltitol, xylitol, isomalt

Abbreviations: FODMAPs, fermentable oligosaccharides, disaccharides, monosaccharides, and polyols; FOS, fructo-oligosaccharides.

Source: Adapted from PR Gibson et al: Am J Gastroenterol 107:657, 2012.

TABLE 320-3

Spectrum of Severity in IBS

	Mild	Moderate	Severe
Clinical Features			
Prevalence	70%	25%	5%
Correlations with gut physiology	+++	++	+
Symptoms constant	0	+	+++
Psychosocial difficulties	0	+	+++
Health care issues	+	++	+++
Practice type	Primary	Specialty	Referral

TABLE 320-4

Possible Drugs for a Dominant Symptom in IBS

Symptom	Drug	Dose
Diarrhea	Loperamide	2–4 mg when necessary/maximum 12 g/d
	Cholestyramine resin	4 g with meals
	Alosetron[a]	0.5–1 mg bid (for severe IBS, women)
Constipation	Psyllium husk	3–4 g bid with meals, then adjust
	Methylcellulose	2 g bid with meals, then adjust
	Calcium polycarbophil	1 g qd to qid
	Lactulose syrup	10–20 g bid
	70% sorbitol	15 mL bid
	Polyethylene glycol 3350	17 g in 250 mL water qd
	Lubiprostone (Amitiza)	24 mg bid
	Magnesium hydroxide	30–60 mL qd
	Linaclotide	290 µg qd
Abdominal pain	Smooth-muscle relaxant	qd to qid ac
	Tricyclic antidepressants	Start 25–50 mg hs, then adjust
	Selective serotonin reuptake inhibitors	Begin small dose, increase as needed
Gas and bloating	Low FODMAP diet	
	Probiotics	qd
	Rifaximin	550 mg bid

[a]Available only in the United States.

Abbreviation: FODMAP, fermentable oligosaccharides, disaccharides, monosaccharides, and polyols.

Source: Adapted from GF Longstreth et al: Gastroenterology 130:1480, 2006.

FURTHER READING

Barbara G et al: The immune system in irritable bowel syndrome. J Neurogastroenterol Motil 17:349, 2011. [PubMed: 22148103]

Drossman DA: Functional gastrointestinal disorders: History, pathophysiology, clinical features, and Rome IV. Gastroenterology 150:1262, 2016.

Eswaran SL et al: A randomized controlled trial comparing the low FODMAP diet vs. modified NICE guidelines in US adults with IBS-D. Am J Gastroenterol 111:1824, 2016. [PubMed: 27725652]

Mayer EA et al: Brain-gut microbiome interactions and functional bowel disorders. Gastroenterology 146:1500; 2014. [PubMed: 24583088]

Mayer EA et al: Towards a systems view of IBS. Nar Rev Gastroenterol Hepatol 12:592, 2015.

Natividad JM, Verdu EF: Modulation of intestinal barrier by intestinal microbiota: Pathological and therapeutic implications. Pharmacological Research 69:42, 2013. [PubMed: 23089410]

Simren M et al: Intestinal microbiota in functional bowel disorders: A Rome foundation report. Gut 62:159, 2013. [PubMed: 22730468]

Talley NJ, Fodor AA: Bugs, stool, and the irritable bowel syndrome: Too much is as bad as too little? Gastroenterology 141:1555, 2011. [PubMed: 21945058]

Chapter 321: Diverticular Disease and Common Anorectal Disorders

Rizwan Ahmed; Susan L. Gearhart

DIVERTICULAR DISEASE

Incidence and Epidemiology

In the United States, diverticulosis affects 60% of the population aged >60 and up to 30% of individuals with diverticular disease will experience recurrent symptoms. Diverticular disease has become the fifth most costly gastrointestinal disorder in the United States and is the leading indication for elective colon resection. The incidence of diverticular disease is on the rise. Fortunately, only 20% of patients with diverticulosis develop diverticular disease and 4% require hospitalization. Previously overlooked, the majority of patients with diverticular disease report a lower health-related quality of life and more depression as compared to matched controls, thus adding to health care costs. Formerly, diverticular disease was confined to developed countries; however, with the adoption of westernized diets in underdeveloped countries, diverticulosis is on the rise across the globe. Immigrants to the United States develop diverticular disease at the same rate as U.S. natives. Although the prevalence among females and males is similar, males tend to present at a younger age. The mean age at presentation is now shifting to affect younger populations.

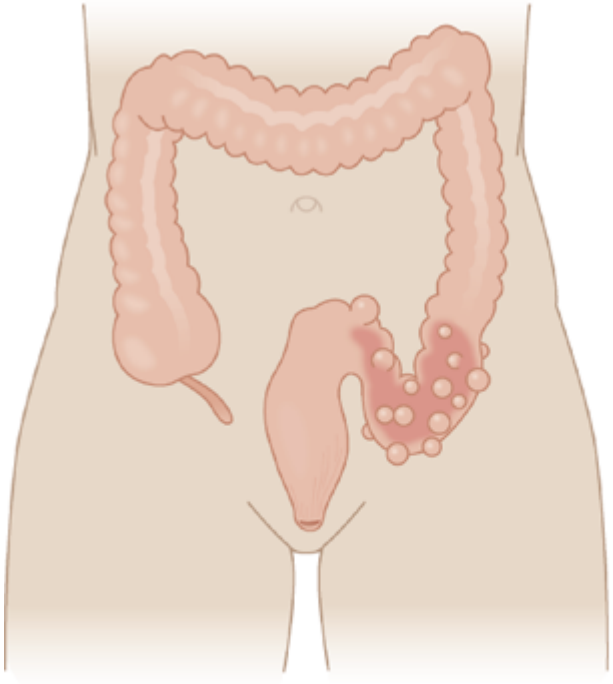
Anatomy and Pathophysiology

Two types of diverticula occur in the intestine: true and false (or pseudo diverticula). A true diverticulum is a saclike herniation of the entire bowel wall, whereas a pseudo diverticulum involves only a protrusion of the mucosa and submucosa through the muscularis propria of the colon **(Fig. 321-1)**. The type of diverticulum most commonly affecting the colon is the pseudo diverticulum. Diverticula commonly affect the left and sigmoid colon; the rectum is always spared. However, in Asian populations, 70% of diverticula are seen in the right colon and cecum as well. Yamanda et al. found right-side colonic diverticulosis in 22% of Japanese patients undergoing colonoscopy. *Diverticulitis* is inflammation of a diverticulum. Previous understanding of the pathogenesis of diverticulosis attributed a low-fiber diet as the sole culprit, and onset of diverticulitis would occur acutely when these diverticula become obstructed. However, evidence now suggests that the pathogenesis is more complex and multifactorial. The diverticula occur at the point where the nutrient artery, or *vasa recti*, penetrates through the muscularis propria, resulting in a break in the integrity of the colonic wall. This anatomic restriction may be a result of the relative high-pressure zone within the muscular

sigmoid colon. Thus, higher-amplitude contractions combined with constipated, high-fat-content stool within the sigmoid lumen in an area of weakness in the colonic wall results in the creation of these diverticula. Consequently, the vasa recti is either compressed or eroded, leading to either perforation or bleeding. Chronic low-grade inflammation is thought to play a key role in neuronal degeneration leading to dysmotility and high intraluminal pressure. As a consequence, pockets or outpouchings develop in the colonic wall where it is weakest. Furthermore, better understanding of the gut microbiota suggests that dysbiosis is an important aspect of disease.

FIGURE 321-1

Gross and microscopic view of sigmoid diverticular disease. *Arrows* mark an inflamed diverticulum with the diverticular wall made up only of mucosa.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Presentation, Evaluation, and Management of Diverticular Bleeding

Hemorrhage from a colonic diverticulum is the most common cause of hematochezia in patients >60 years, yet only 20% of patients with diverticulosis will have gastrointestinal bleeding. Patients at increased risk for bleeding tend to be hypertensive, have atherosclerosis, and regularly use aspirin and nonsteroidal anti-inflammatory agents. Most bleeds are self-limited and stop spontaneously with bowel rest. The lifetime risk of rebleeding is 25%.

Initial localization of diverticular bleeding may include colonoscopy, multiplanar computed tomography (CT) angiogram, or nuclear medicine tagged red cell scan. If the patient is stable, ongoing bleeding is best managed by angiography. If mesenteric angiography can localize the bleeding site, the vessel can be occluded successfully with a coil in 80% of cases. The patient can then be followed closely with repetitive colonoscopy, if necessary, looking for evidence of colonic ischemia. Alternatively, a segmental resection of the colon can be undertaken to eliminate the risk of further bleeding. This may be advantageous in patients on chronic anticoagulation. However, with highly selective coil embolization, the rate of colonic ischemia is <10% and the risk of acute rebleeding is <25%. Long-term results (40 months) indicate that >50% of patients with acute diverticular bleeds treated with highly selective angiography have had definitive treatment. As

another alternative, a selective infusion of vasopressin can be given to stop the hemorrhage, although this has been associated with significant complications, including myocardial infarction and intestinal ischemia. Furthermore, bleeding recurs in 50% of patients once the infusion is stopped.

If the patient is unstable or has had a 6-unit bleed within 24 h, current recommendations are that surgery should be performed. If the bleeding has been localized, a segmental resection can be performed. If the site of bleeding has not been definitively identified, a subtotal colectomy may be required. In patients without severe comorbidities, surgical resection can be performed with a primary anastomosis. A higher anastomotic leak rate has been reported in patients who received >10 units of blood.

Presentation, Evaluation, and Staging of Diverticulitis

Acute uncomplicated diverticulitis (also known as Symptomatic Uncomplicated Diverticular Disease, SUDD) characteristically presents with fever, anorexia, left lower quadrant abdominal pain, and obstipation **(Table 321-1)**. In <25% of cases, patients may present with generalized peritonitis indicating the presence of a diverticular perforation. If a pericolonic abscess has formed, the patient may have abdominal distention and signs of localized peritonitis. Laboratory investigations will demonstrate a leukocytosis. Rarely, a patient may present with an air-fluid level in the left lower quadrant on plain abdominal film. This is a giant diverticulum of the sigmoid colon and is managed with resection to avoid impending perforation.

TABLE 321-1

Presentation of Diverticular Disease

Uncomplicated Diverticular Disease—75%
Abdominal pain
Fever
Leukocytosis
Anorexia/obstipation
Complicated Diverticular Disease—25%
Abscess 16%
Perforation 10%
Stricture 5%
Fistula 2%

The diagnosis of diverticulitis is best made on CT with the following findings: sigmoid diverticula, thickened colonic wall >4 mm, and inflammation within the periodic fat ± the collection of contrast material or fluid. In up to 20% of patients, an abdominal abscess may be present. Symptoms of irritable bowel syndrome **(Chap. 320)** may mimic those of diverticulitis. Therefore, suspected diverticulitis that does not meet CT criteria or is not associated with a leukocytosis or fever is not diverticular disease. Other conditions that can mimic diverticular disease include an ovarian cyst, endometriosis, acute appendicitis, and pelvic inflammatory disease.

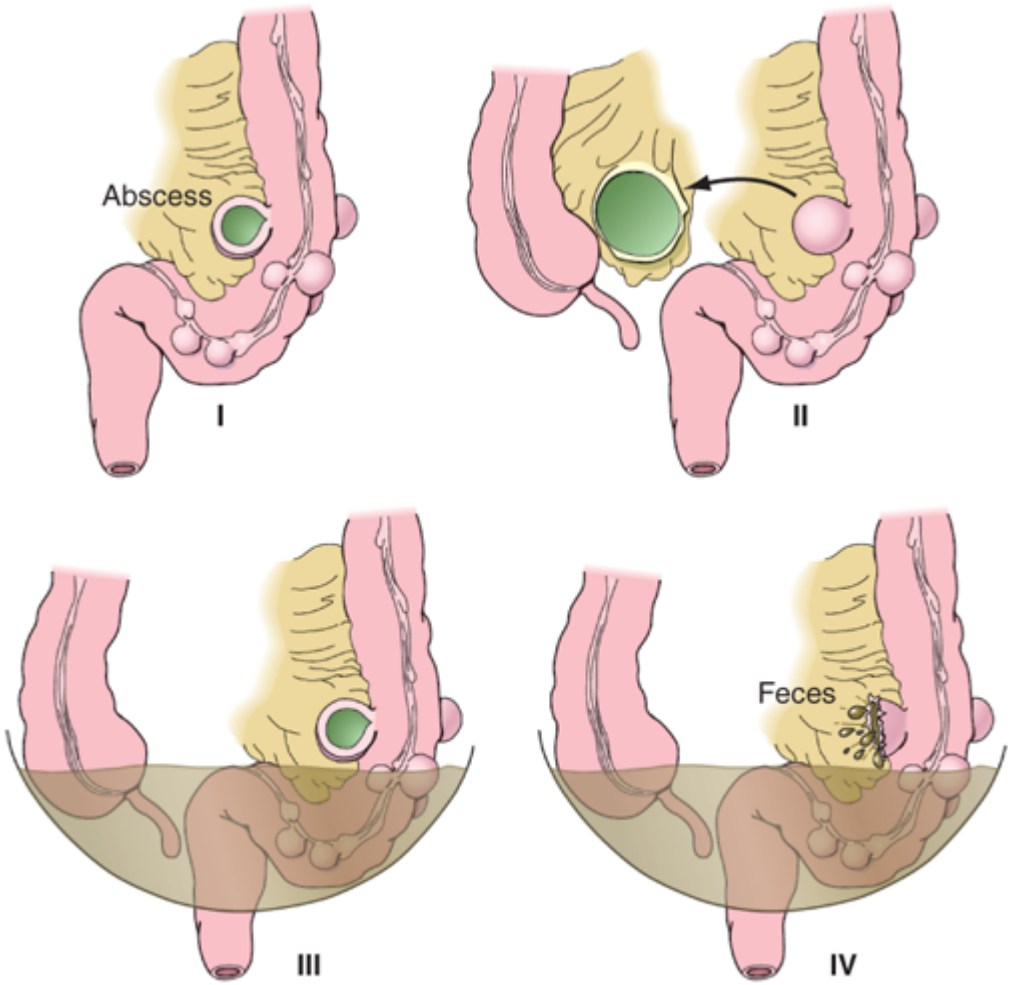
Although the benefit of colonoscopy in the evaluation of patients with diverticular disease has been called into question, its use is still considered important in the exclusion of colorectal cancer. The parallel epidemiology of colorectal cancer and diverticular disease provides enough concern for an endoscopic evaluation before operative management. Therefore, a colonoscopy should be performed ~6 weeks after an attack of diverticular disease.

Complicated diverticular disease is defined as diverticular disease associated with an abscess or perforation and less commonly with a fistula (Table 321-1). Perforated diverticular disease is staged using the Hinchey classification system **(Fig. 321-2)**. This staging system was developed to predict outcomes following the surgical management of complicated diverticular disease. In recent years, the Hinchey staging system has

been modified to include the development of a phlegmon or early abscess (Hinchey stage Ia). A pericolic abscess is then considered Hinchey stage Ib. In complicated diverticular disease with fistula formation, common locations include cutaneous, vaginal, or vesicle fistulas. These conditions present with either passage of stool through the skin or vagina or the presence of air in the urinary stream (pneumaturia). Colovaginal fistulas are more common in women who have undergone a hysterectomy.

FIGURE 321-2

Hinchey classification of diverticulitis. Stage I: Perforated diverticulitis with a confined paracolic abscess. Stage II: Perforated diverticulitis that has closed spontaneously with distant abscess formation. Stage III: Noncommunicating perforated diverticulitis with fecal peritonitis (the diverticular neck is closed off, and, therefore, contrast will not freely expel on radiographic images). Stage IV: Perforation and free communication with the peritoneum, resulting in fecal peritonitis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Diverticular Disease

MEDICAL MANAGEMENT

Asymptomatic diverticular disease discovered on imaging studies or at the time of colonoscopy is best managed by lifestyle changes. Although the data regarding dietary risks and symptomatic diverticular disease are limited **(see Table 321-2)**, patients may benefit from a fiber-enriched diet that includes 30 g of fiber each day. Supplementary fiber products such as Metamucil, Fibercon, or Citrucel are useful. The use of fiber increases colonic transit time, and, therefore, preventing increased intraluminal pressure leading to the development of diverticulosis. The incidence of complicated diverticular disease appears to also be increased in patients who smoke. Therefore, patients should be encouraged to refrain from smoking. The historical recommendation to avoid eating nuts is based on no more than anecdotal data.

SUDD with confirmation of inflammation and infection within the colon should be treated initially with bowel rest. The routine use of antibiotics in uncomplicated diverticular disease did not demonstrate any benefit in time to symptom resolution, complications, or risk of recurrence. However, the data are limited and antibiotics remain in the treatment paradigm. Hospitalization is recommended if the patient is unable to take oral therapy, affected by several comorbidities, fails to improve with outpatient therapy, and if the patient is affected by complicated diverticulitis. Nearly 75% of patients hospitalized for acute diverticulitis will respond to nonoperative treatment with a suitable antimicrobial regimen. The current recommended antimicrobial coverage is a third-generation cephalosporin or ciprofloxacin and metronidazole targeting aerobic gram-negative rods and anaerobic bacteria. Unfortunately, these agents do not cover enterococci, and the addition of ampicillin to this regimen for nonresponders is recommended. Alternatively, single-agent therapy with a third-generation penicillin such as IV piperacillin or oral penicillin/clavulanic acid may be effective. The usual course of antibiotics is 7–10 days, although this length of time is being investigated. Patients should remain on a limited diet until their pain resolves.

Once the acute attack has resolved, the mainstay medical management of diverticular disease to prevent symptoms has evolved. Established risk factors for symptomatic recurrence include younger age, the formation of a diverticular abscess, and more frequent attacks (>2 per year). Newer directions are targeted at colonic inflammation and dysbiosis. Diverticular disease is now considered a functional bowel disorder associated with low-grade inflammation. However, the use of anti-inflammatory medications (mesalazine) in randomized clinical trials has not demonstrated any effect on recurrence rates over placebo alone. Some authors have suggested that the use of anti-inflammatory medications is most helpful in patients with diverticular disease who also have segmental colitis (Segmental Colitis-Associated Diverticular Disease [SCADD]). Treatment strategies targeting dysbiosis in diverticular disease have also been evaluated using polymerase chain reaction (PCR) on stool specimens. Stool samples from consumers of a high-fiber diet have different bacterial content than stool samples from consumers of a low-fiber, high-fat diet. Probiotics are increasingly used by gastroenterologists for multiple bowel disorders and may prevent recurrence of diverticulitis. Specifically, probiotics containing *Lactobacillus acidophilus* and *Bifidobacterium* strains may be beneficial, however, a recent systematic review was unable to show any benefit to the use of probiotics. Rifaximin (a poorly absorbed broad-spectrum antibiotic), when compared to fiber alone for the treatment of SUDD, is associated with 30% less frequent recurrent symptoms from uncomplicated diverticular disease.

SURGICAL MANAGEMENT

Preoperative risk factors influencing postoperative mortality rates include higher American Society of Anesthesiologists (ASA) physical status class **(Table 321-3)** and preexisting organ failure. In patients who are low risk (ASA P1 and P2), surgical therapy can be offered to those who do not rapidly improve on medical therapy. For uncomplicated diverticular disease, medical therapy can be continued beyond two attacks without an increased risk of perforation requiring a colostomy. However, patients on immunosuppressive therapy, in chronic renal failure, or with a collagen-vascular disease have a fivefold greater risk of perforation during recurrent attacks. Surgical therapy is indicated in all low-surgical-risk patients with complicated diverticular disease.

The goals of surgical management of diverticular disease include controlling sepsis, eliminating complications such as fistula or obstruction, removing the diseased colonic segment, and restoring intestinal continuity. These goals must be obtained while minimizing morbidity rate, length of hospitalization, and cost in addition to maximizing survival and quality of life. **Table 321-4** lists the operations most commonly indicated based on the Hinchey classification and the predicted postoperative outcomes. The current options for uncomplicated diverticular disease include an open or a laparoscopic resection of the diseased area with reanastomosis to the rectosigmoid. Preservation of portions of the sigmoid colon may lead to early recurrence of the disease. The benefits of laparoscopic resection over open surgical techniques include early discharge (by at least 1 day), less narcotic use, less postoperative complications, and an earlier return to work.

The options for the surgical management of complicated diverticular disease **(Fig. 321-3)** include the following open or laparoscopic procedures: (1) proximal diversion of the fecal stream with an ileostomy or colostomy and sutured omental patch with drainage, (2) resection with colostomy and mucous fistula or closure of distal bowel with formation of a Hartmann's pouch (Hartmann's procedure), (3) resection with anastomosis (coloproctostomy), or (4) resection with anastomosis and diversion (coloproctostomy with loop ileostomy or colostomy). (5) Laparoscopic technique of washout and drainage without diversion has been described for Hinchey III patients; however, a threefold increased risk of recurrent peritonitis requiring reoperation with washout alone has been reported.

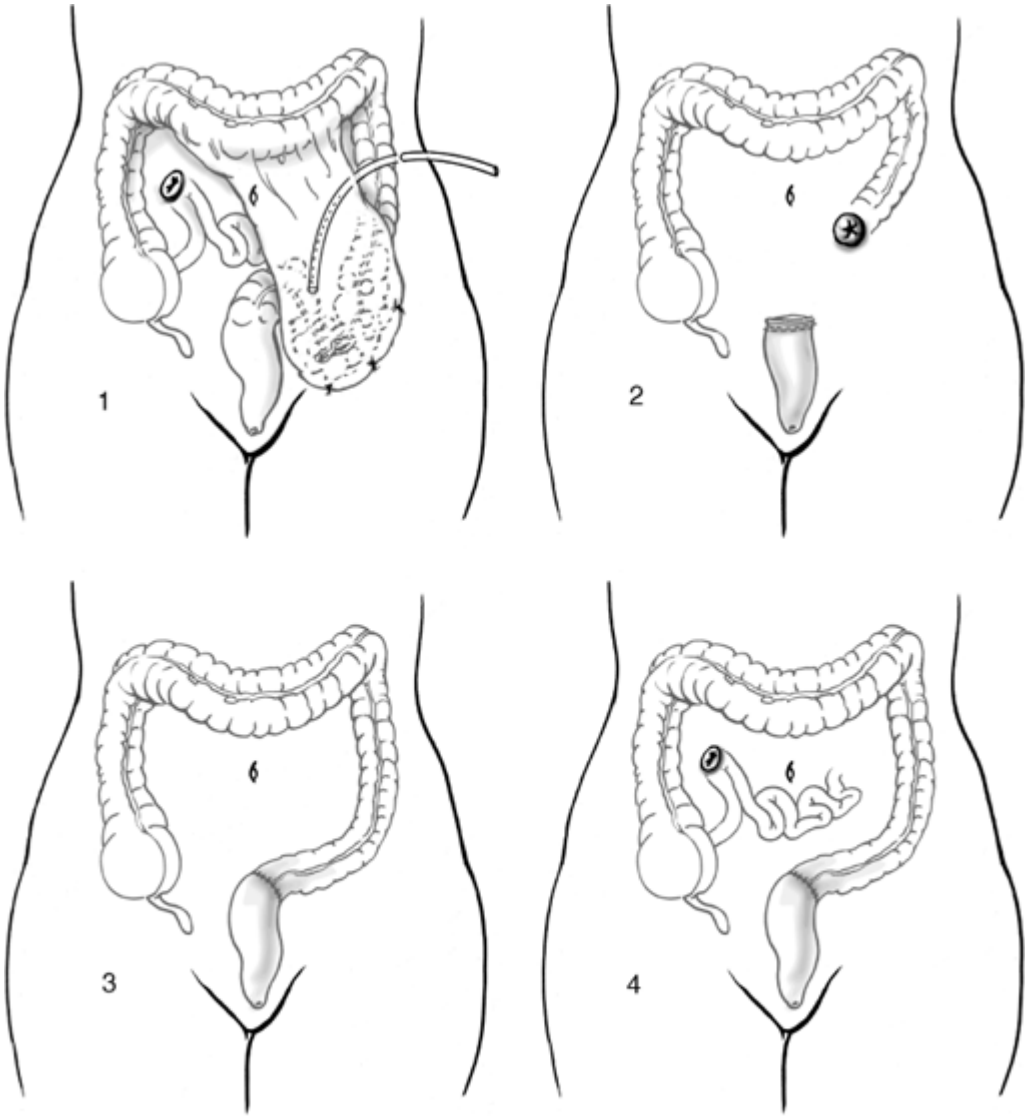
Patients with Hinchey stage Ia are managed with antibiotic therapy only followed by resection with anastomosis at 6 weeks. Patients with Hinchey stages Ib and II disease are managed with percutaneous drainage followed by resection with anastomosis about 6 weeks later. Current guidelines put forth by the American Society of Colon and Rectal Surgeons suggest, in addition to antibiotic therapy, CT-guided percutaneous drainage of diverticular abscesses that are >3 cm and have a well-defined wall. Abscesses that are <5 cm may resolve with antibiotic therapy alone. Contraindications to percutaneous drainage are no percutaneous access route, pneumoperitoneum, and fecal peritonitis. Drainage of a diverticular abscess is associated with a 20–25% failure rate. Urgent operative intervention is undertaken if percutaneous drainage fails and patients develop generalized peritonitis, and most will need to be managed with a Hartmann's procedure (resection of the sigmoid colon with end colostomy and rectal stump). In selected cases, nonoperative therapy may be considered. In one nonrandomized study, nonoperative management of

isolated paracolic abscesses (Hinchey stage I) was associated with only a 20% recurrence rate at 2 years. More than 80% of patients with distant abscesses (Hinchey stage II) required surgical resection for recurrent symptoms.

The management of Hinchey stage III disease is under debate. In this population of patients, no fecal peritonitis is present and it is presumed that the perforation has sealed. Historically, Hinchey stage III has been managed with a Hartmann's procedure or with primary anastomosis and proximal diversion. Several studies have examined short- and long-term outcomes for laparoscopic peritoneal lavage to remove the peritoneal contamination and place drainage catheters should a communication to the bowel still exist. However, this procedure has been associated with an increased risk of requiring reoperation for ongoing peritonitis. Overall, ostomy rates are lower with the use of laparoscopic peritoneal lavage. No anastomosis of any type should be attempted in Hinchey stage IV disease, or the presence of fecal peritonitis. A limited approach to these patients is associated with a decreased mortality rate.

FIGURE 321-3

Methods of surgical management of complicated diverticular disease. ***1.*** Drainage, omental pedicle graft, and proximal diversion. ***2.*** Hartmann's procedure. ***3.*** Sigmoid resection with coloproctostomy. ***4.*** Sigmoid resection with coloproctostomy and proximal diversion.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 321-2

The Use of Fiber in the Management of Diverticular Disease

Journal, Study Year	Patients (n)	Intervention	Study Length	Findings
Lancet, 1977	18	Wheat or bran crisp bread	3 months	Significant reduction of symptoms score
BMJ, 1981	58	Bran, ispaghula, placebo	16 weeks	No difference
J Gastroenterol, 1977	30	Methylcellulose	3 months	Significant reduction in symptoms
BMJ, 2011	47,033	Vegetarian vs nonvegetarian	11.6 years	Vegetarians had a 31% lower risk of DD
Gastroenterology, 2012	2104	Fiber consumption	12 years	Fiber associated with great risk of DD
Jama, 2008	47,288	Nut, corn, popcorn consumption	18 years	Higher nut, corn, and popcorn had lower risk of recurrence
Ann R Coll Surg Engl, 1985	56	Fiber consumption	66 months	Higher fiber associated with 19% reduction in symptom recurrence

Modified from A Turis, A Papa, S Danese: Review Article: The pathophysiology and medical management of diverticulosis and diverticular disease of the colon. Aliment Parmacol Ther 42:664, 2015.

TABLE 321-3

American Society of Anesthesiologists Physical Status Classification System

P1	A normal healthy patient
P2	A patient with mild systemic disease
P3	A patient with severe systemic disease
P4	A patient with severe systemic disease that is a constant threat to life
P5	A moribund patient who is not expected to survive without the operation
P6	A declared brain-dead patient whose organs are being removed for donor purposes

TABLE 321-4

Outcome Following Surgical Therapy for Complicated Diverticular Disease Based Upon Modified Hinchey Staging

Hinchey Stage	Operative Procedure	Anastomotic Leak Rate, %	Overall Morbidity Rate, %
Ia (pericolic phlegmon)	Laparoscopic or open colon resection	43	15
Ib (pericolic abscess)	Percutaneous drainage followed by laparoscopic or open colon resection	3	15
II	Percutaneous drainage followed by laparoscopic or open colon resection +/− proximal diversion with an ostomy	3	15
III	Laparoscopic washout and drainage *or* Laparoscopic or open resection with proximal diversion (ostomy) *or* Hartmann's procedure	3	30% risk of peritonitis requiring reoperation if no resection is performed. Overall morbidity 50% Overall mortality 15%
IV	Hartmann's procedure *or* Washout with proximal diversion	—	Overall morbidity 50% Overall mortality 15%

Recurrent Symptoms

Recurrent abdominal symptoms following surgical resection for diverticular disease occur in 10% of patients. Recurrent diverticular disease develops in patients following inadequate surgical resection. A retained segment of diseased rectosigmoid colon is associated with twice the incidence of recurrence. The presence of irritable bowel syndrome may also cause recurrence of initial symptoms. Patients undergoing surgical resection for presumed diverticulitis and symptoms of chronic abdominal cramping and irregular loose bowel movements consistent with irritable bowel syndrome have poorer functional outcomes.

COMMON DISEASES OF THE ANORECTUM

RECTAL PROLAPSE (PROCIDENTIA)

Incidence and Epidemiology

Rectal prolapse is six times more common in women than in men. The incidence of rectal prolapse peaks in women >60 years. Women with rectal prolapse have a higher incidence of associated pelvic floor disorders including urinary incontinence, rectocele, cystocele, and enterocele. About 20% of children with rectal prolapse will have cystic fibrosis. All children presenting with prolapse should undergo a sweat chloride test. Less common associations include Ehlers-Danlos syndrome, solitary rectal ulcer syndrome, congenital hypothyroidism, Hirschsprung's disease, dementia, mental retardation, and schizophrenia.

Anatomy and Pathophysiology

Rectal prolapse (procidentia) is a circumferential, full-thickness protrusion of the rectal wall through the anal orifice. It is often associated with a redundant sigmoid colon, pelvic laxity, and a deep rectovaginal septum (pouch of Douglas). Initially, rectal prolapse was felt to be the result of early internal rectal intussusception, which occurs in the upper to mid rectum. This was considered to be the first step in an inevitable progression to full-thickness external prolapse. However, only 1 of 38 patients with internal prolapse followed for >5 years developed full-thickness prolapse. Others have suggested that full-thickness prolapse is the result of damage to the nerve supply to the pelvic floor muscles or pudendal nerves from repeated stretching with straining to defecate. Damage to the pudendal nerves would weaken the pelvic floor muscles, including the external anal sphincter muscles. Bilateral pudendal nerve injury is more significantly associated with prolapse and incontinence than unilateral injury.

Presentation and Evaluation

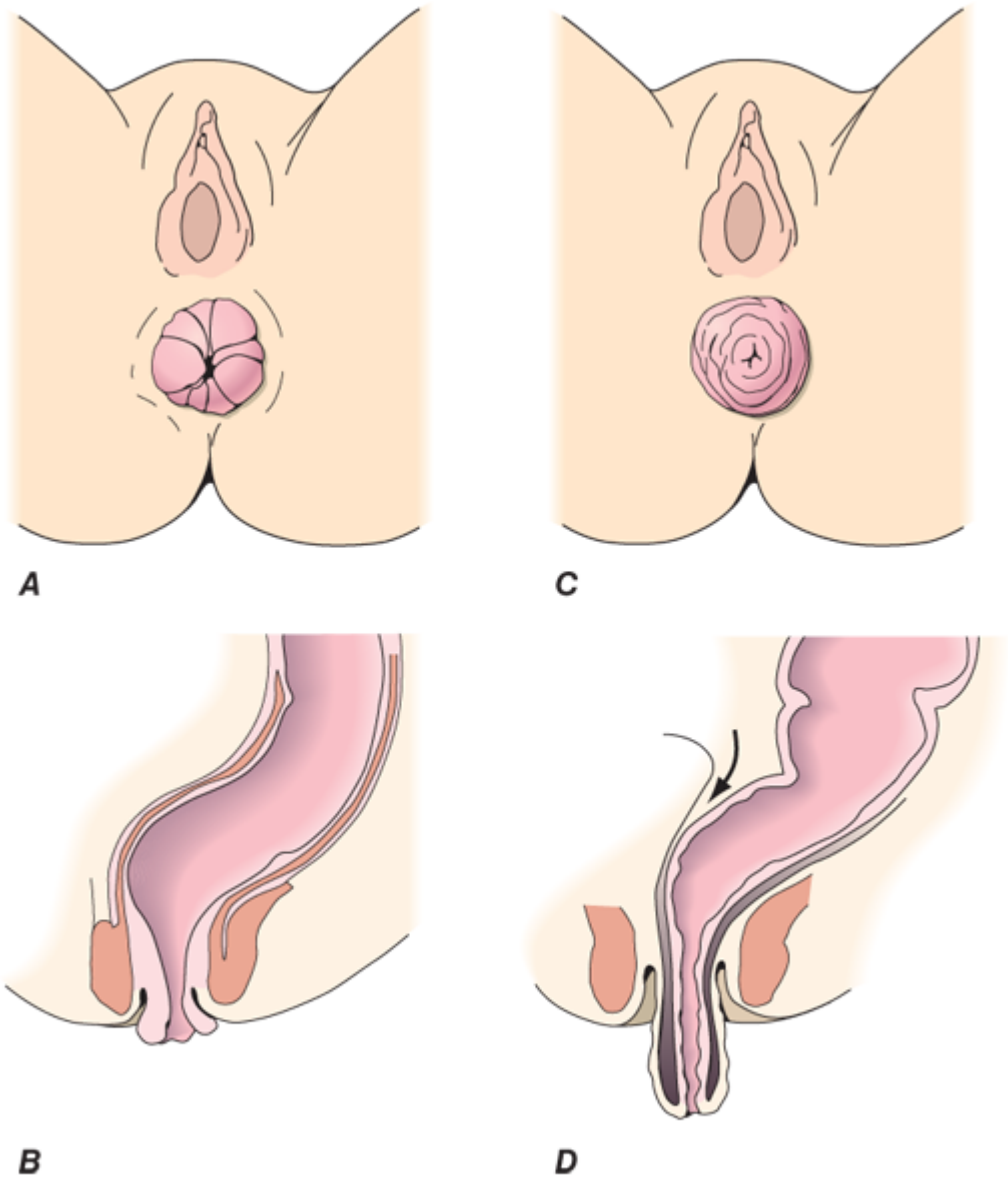
In external prolapse, the majority of patient complaints include anal mass, bleeding per rectum, and poor perianal hygiene. Prolapse of the rectum usually occurs following defecation and will spontaneously reduce or require the patient to manually reduce the prolapse. Constipation occurs in ~30–67% of patients with rectal prolapse. Differing degrees of fecal incontinence occur in 50–70% of patients. Patients with internal rectal prolapse will present with symptoms of both constipation and incontinence. Other associated findings include outlet obstruction (anismus) in 30%, colonic inertia in 10%, and solitary rectal ulcer syndrome in 12%.

Office evaluation is best performed after the patient has been given an enema, which enables the prolapse to protrude. An important distinction should be made between full-thickness rectal prolapse and isolated mucosal prolapse associated with hemorrhoidal disease **(Fig. 321-4)**. Mucosal prolapse is known for radial grooves rather than circumferential folds around the anus and is due to increased laxity of the connective tissue between the submucosa and underlying muscle of the anal canal. The evaluation of prolapse should also include cystoproctography and colonoscopy. These examinations evaluate for associated pelvic floor

disorders and rule out a malignancy or a polyp as the lead point for prolapse. If rectal prolapse is associated with chronic constipation, the patient should undergo a defecating proctogram and a sitzmark study. This will evaluate for the presence of anismus or colonic inertia. Anismus is the result of attempting to defecate against a closed pelvic floor and is also known as *nonrelaxing puborectalis*. This can be seen when straightening of the rectum fails to occur on fluoroscopy while the patient is attempting to defecate. In colonic inertia, a sitzmark study will demonstrate retention of >20% of markers on abdominal x-ray 5 days after swallowing. For patients with fecal incontinence, endoanal ultrasound and manometric evaluation, including pudendal nerve testing of their anal sphincter muscles, may be performed before surgery for prolapse (see "Fecal Incontinence," below).

FIGURE 321-4

Degrees of rectal prolapse. Mucosal prolapse only (***A, B,*** sagittal view). Full-thickness prolapse associated with redundant rectosigmoid and deep pouch of Douglas (***C, D,*** sagittal view).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

Rectal Prolapse

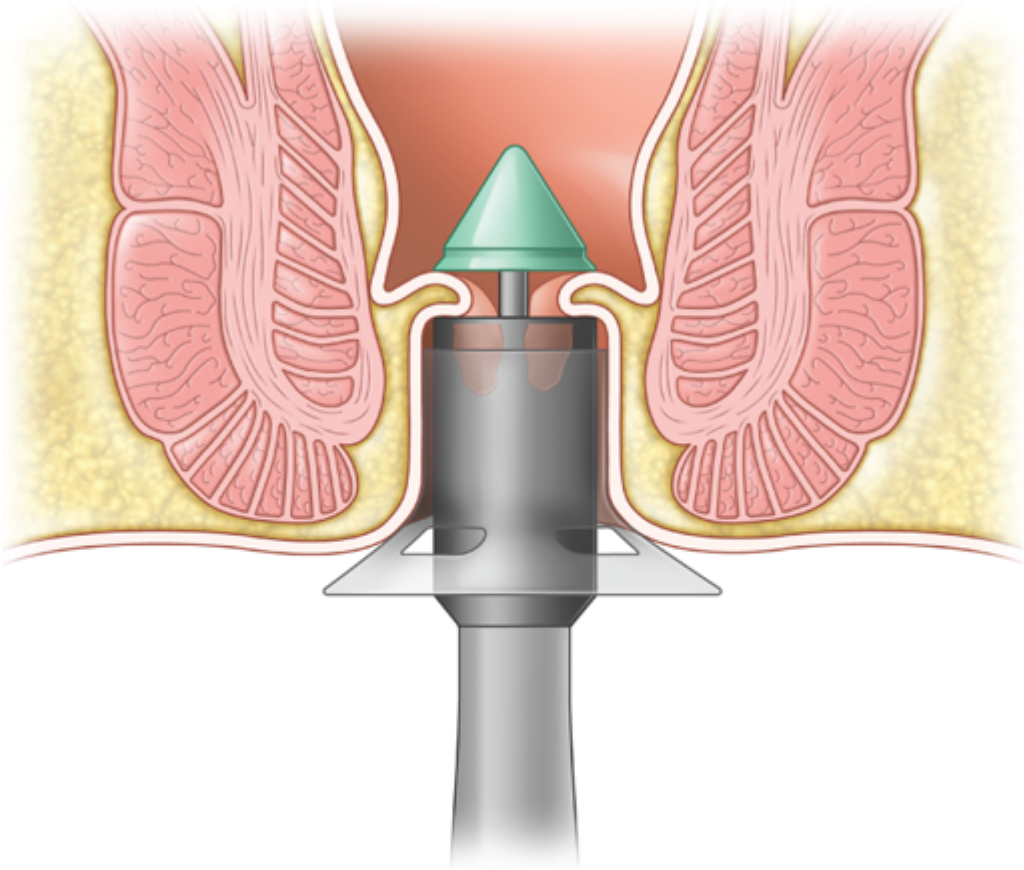
The medical approach to the management of rectal prolapse is limited and includes stool-bulking agents or fiber supplementation to ease the process of evacuation. Surgical correction of rectal prolapse is the mainstay of therapy. Two approaches are commonly considered, transabdominal and transperineal. Transabdominal approaches have been associated with lower recurrence rates, but some patients with significant comorbidities are better served by a transperineal approach.

Common transperineal approaches include a transanal proctectomy (Altmeier procedure), mucosal proctectomy (Delorme procedure), or placement of a Tirsch wire encircling the anus. The goal of the transperineal approach is to remove the redundant rectosigmoid colon. Common transabdominal approaches include presacral suture or mesh rectopexy (Ripstein) with (Frykman-Goldberg) or without resection of the redundant sigmoid. Colon resection, in general, is reserved for patients with constipation and outlet obstruction. Ventral rectopexy is an effective method of abdominal repair of full-thickness prolapse that does not require sigmoid resection (see description below). This repair may have improved functional results over other abdominal repairs. Transabdominal procedures can be performed effectively with laparoscopic and, more recently, robotic techniques without increased incidence of recurrence. The goal of the transabdominal approach is to restore normal anatomy by removing redundant bowel and reattaching the supportive tissue of the rectum to the presacral fascia. The final alternative is abdominal proctectomy with end-sigmoid colostomy. If total colonic inertia is present, as defined by a history of constipation and a positive sitzmark study, a subtotal colectomy with an ileosigmoid or rectal anastomosis may be required at the time of rectopexy.

Previously, the presence of internal rectal prolapse identified on imaging studies has been considered a nonsurgical disorder, and biofeedback was recommended. However, only one-third of patients will have successful resolution of symptoms from biofeedback. Two surgical procedures more effective than biofeedback are the Stapled Transanal Rectal Resection (STARR) and the Laparoscopic Ventral Rectopexy (LVR). The STARR procedure **(Fig. 321-5)** is performed through the anus in patients with internal prolapse. A circular stapling device is inserted through the anus; the internal prolapse is identified and ligated with the stapling device. LVR **(Fig. 321-6)** is performed through an abdominal approach. An opening in the peritoneum is created on the left side of the rectosigmoid junction, and this opening continues down anterior on the rectum into the pouch of Douglas. No rectal mobilization is performed, thus avoiding any autonomic nerve injury. Mesh is secured to the anterior and lateral portion of the rectum, the vaginal fornix, and the sacral promontory, allowing for closure of the rectovaginal septum and correction of the internal prolapse. In both procedures, recurrence at 1 year was low (<10%) and symptoms improved in more than three-fourths of patients.

FIGURE 321-5

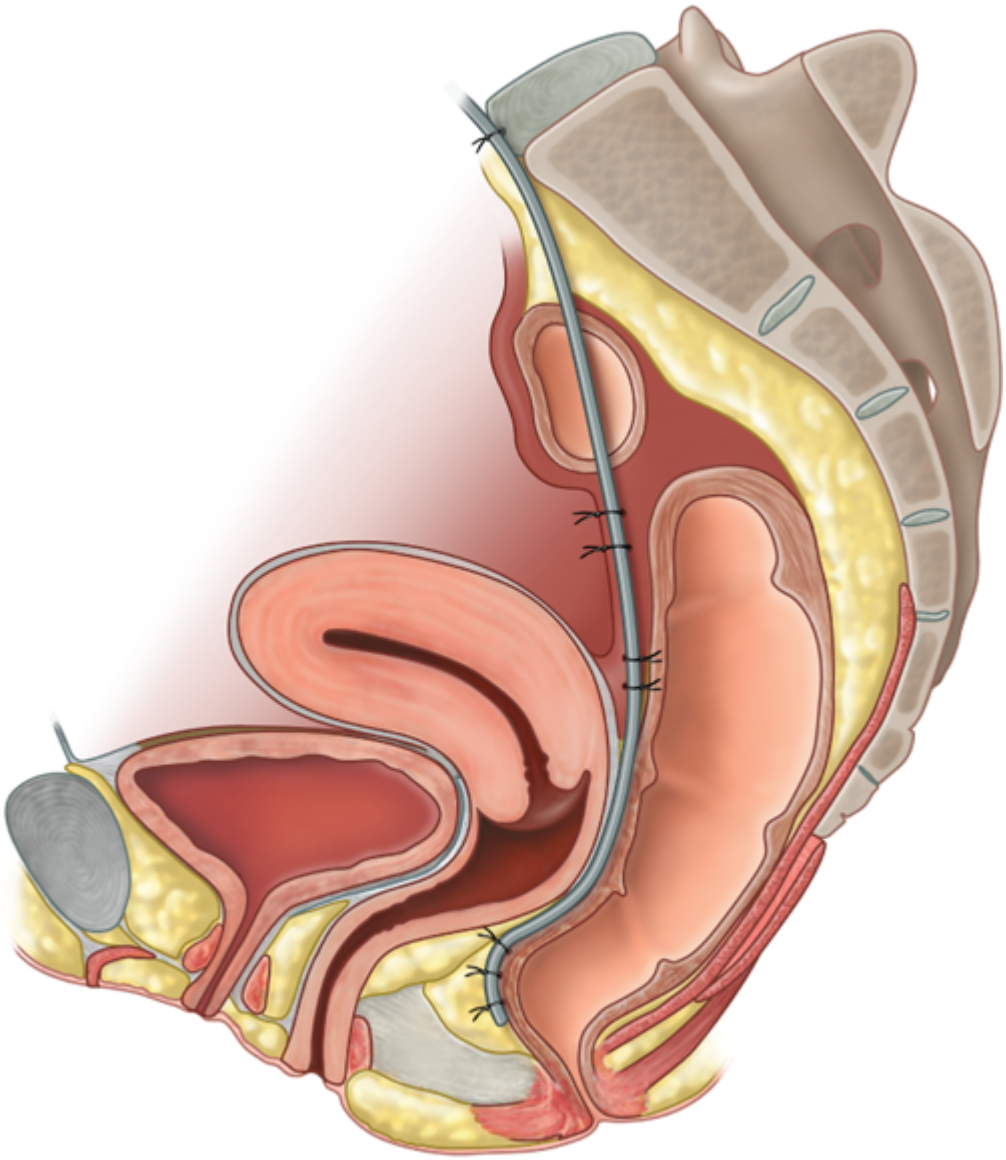
Stapled transanal rectal resection. Schematic of placement of the circular stapling device.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 321-6

Laparoscopic ventral rectopexy (LVR). To reduce the internal prolapse and close any rectovaginal septal defect, the pouch of Douglas is opened and mesh is secured to the anterolateral rectum, vaginal fornix, and sacrum. *(From A D'Hoore et al: Br J Surg 91:1500, 2004.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FECAL INCONTINENCE

Incidence and Epidemiology

Fecal incontinence is the involuntary passage of fecal material for at least 1 month in an individual with a developmental age of at least 4 years. The prevalence of fecal incontinence in the United States is 0.5–11%. The majority of patients are women and aged >65. A higher incidence of incontinence is seen among parous women. One-half of patients with fecal incontinence also suffer from urinary incontinence. The majority of incontinence is a result of obstetric injury to the pelvic floor, either while carrying a fetus or during the delivery. An anatomic sphincter defect may occur in up to 32% of women following childbirth regardless of visible damage to the perineum. Risk factors at the time of delivery include prolonged labor, the use of forceps, and the need for an episiotomy. Symptoms of incontinence can present after two or more decades

following obstetric injury. Medical conditions known to contribute to the development of fecal incontinence are listed in **Table 321-5**.

TABLE 321-5

Medical Conditions That Contribute to Symptoms of Fecal Incontinence

Neurologic Disorders
• Dementia
• Brain tumor
• Stroke
• Multiple sclerosis
• Tabes dorsalis
• Cauda equina lesions
Skeletal Muscle Disorders
• Myasthenia gravis
• Myopathies, muscular dystrophy
Miscellaneous
• Hypothyroidism
• Irritable bowel syndrome
• Diabetes
• Severe diarrhea
• Scleroderma

Anatomy and Pathophysiology

The anal sphincter complex is made up of the internal and external anal sphincter. The internal sphincter is smooth muscle and a continuation of the circular fibers of the rectal wall. It is innervated by the intestinal myenteric plexus and is therefore not under voluntary control. The external anal sphincter is formed in continuation with the levator ani muscles and is under voluntary control. The pudendal nerve supplies motor innervation to the external anal sphincter. Obstetric injury may result in tearing of the muscle fibers anteriorly at the time of the delivery. This results in an obvious anterior defect on endoanal ultrasound. Injury may also be the result of stretching of the pudendal nerves during pregnancy or delivery of the fetus through the birth canal.

Presentation and Evaluation

Patients may suffer with varying degrees of fecal incontinence. Minor incontinence includes incontinence to flatus and occasional seepage of liquid stool. Major incontinence is frequent inability to control solid waste. As a result of fecal incontinence, patients suffer from poor perianal hygiene. Beyond the immediate problems associated with fecal incontinence, these patients are often withdrawn and suffer from depression. For this reason, quality-of-life measures are an important component in the evaluation of patients with fecal incontinence.

The evaluation of fecal incontinence should include a thorough history and physical examination including digital rectal examination (DRE). Weak sphincter tone on DRE and loss of the “anal wink” reflex (S1-level control) may indicate a neurogenic dysfunction. Perianal scars may represent surgical injury. Other studies helpful in the diagnosis of fecal incontinence include anal manometry, pudendal nerve terminal motor latency (PNTML), and endoanal ultrasound. Centers that care for patients with fecal incontinence will have an anorectal physiology laboratory that uses standardized methods of evaluating anorectal physiology. Anorectal manometry (ARM) measures resting and squeeze pressures within the anal canal using an intraluminal water-perfused catheter. Current methods of ARM include use of a three-dimensional, high-resolution system with a 12-catheter perfusion system, which allows physiologic delineation of anatomic abnormalities. Pudendal nerve studies evaluate the function of the nerves innervating the anal canal using a finger electrode placed in the anal canal. Stretch injuries to these nerves will result in a delayed response of the sphincter muscle to a stimulus, indicating a prolonged latency. Finally, endoanal ultrasound will evaluate the extent of the injury to the sphincter muscles before surgical repair. Unfortunately, all of these investigations are user-dependent, and very few studies demonstrate that these studies predict outcome following an intervention. Magnetic resonance imaging (MRI) has been used, but its routine use for imaging in fecal incontinence is not well established.

Rarely does a pelvic floor disorder exist alone. The majority of patients with fecal incontinence will have some degree of urinary incontinence. Similarly, fecal incontinence is a part of the spectrum of pelvic organ prolapse. For this reason, patients may present with symptoms of obstructed defecation as well as fecal incontinence. Careful evaluation including dynamic MRI or cinedefecography should be performed to search for other associated defects. Surgical repair of incontinence without attention to other associated defects may decrease the success of the repair.

TREATMENT

Fecal Incontinence

Medical management of fecal incontinence includes strategies to bulk up the stool, which help in increasing fecal sensation. These include fiber supplementation, loperamide, diphenoxylate, and bile acid binders. These agents harden the stool and delay frequency of bowel movements and are helpful in patients with minimal to mild symptoms. Furthermore, patients can be offered a form of physical therapy called biofeedback. This therapy helps strengthen the external sphincter muscle while training the patient to relax with defecation to avoid unnecessary straining and further injury to the sphincter muscles. Biofeedback has had variable success and is dependent on the motivation of the patient. At a minimum, biofeedback is risk-free. Most patients will have some improvement. For this reason, it should be incorporated into the initial recommendation to all patients with fecal incontinence.

Historically, the “gold standard” for the treatment of fecal incontinence with an isolated sphincter defect has been the overlapping sphincteroplasty. The external anal sphincter muscle and scar tissue as well as any identifiable internal sphincter muscle are dissected free from the surrounding adipose and connective tissue and then an overlapping repair is performed in an attempt to rebuild the muscular ring and restore its function. However, long-term results following overlapping sphincteroplasty have been poor with a 50% failure rate over 5 years.

Alternative therapies such as Sacral Nerve Stimulation (SNS), collagen-enhancing injectables, and magnetic “Fenix” ring are other options. SNS is an adaptation of a procedure developed for the management of urinary incontinence. SNS is ideally suited for patients with intact but weak anal sphincters. A temporary nerve stimulator is placed on the third sacral nerve. If there is at least a 50% improvement in symptoms, a permanent nerve stimulator is placed under the skin. Long-term results for sacral stimulation have been promising, with nearly 80% of patients having a reduction in incontinence episodes by at least 50%. This reduction has been sustainable in studies out to 5 years. Collagen-enhancing injectables have been around for several years. More than 50% of incontinent patients treated with nonanimal stabilized hyaluronic acid (NASHA/DX) achieved a 50% reduction in incontinence episodes, and these results were sustainable up to 2 years. Currently, this injectable is not universally available. The Fenix is a magnetic ring that is implanted around the anal sphincter muscles. Its long-term outcomes are still being studied and it is currently only available for compassionate use.

Finally, the use of stem cells to increase the bulk of the sphincter muscles is currently being tested. Stem cells can be harvested from the patient’s own muscle, grown, and then implanted into their sphincter complex. Concern for cost and the need for an additional procedure dampen enthusiasm. Trial results are awaited.

HEMORRHOIDAL DISEASE

Incidence and Epidemiology

Symptomatic hemorrhoids affect >1 million individuals in the Western world per year. The prevalence of hemorrhoidal disease is not selective for age or sex. However, age is known to be a risk factor. The prevalence of hemorrhoidal disease is less in underdeveloped countries. The typical low-fiber, high-fat Western diet is associated with constipation and straining and the development of symptomatic hemorrhoids.

Anatomy and Pathophysiology

Hemorrhoidal cushions are a normal part of the anal canal. The vascular structures contained within this tissue aid in continence by preventing damage to the sphincter muscle. Three main hemorrhoidal complexes traverse the anal canal—the left lateral, the right anterior, and the right posterior. Engorgement and straining lead to prolapse of this tissue into the anal canal. Over time, the anatomic support system of the hemorrhoidal complex weakens, exposing this tissue to the outside of the anal canal where it is susceptible to injury. Hemorrhoids are commonly classified as external or internal. External hemorrhoids originate below the dentate line and are covered with squamous epithelium and are associated with an internal component. External hemorrhoids are painful when thrombosed. Internal hemorrhoids originate above the dentate line and are covered with mucosa and transitional zone epithelium and represent the majority of hemorrhoids. The standard classification of hemorrhoidal disease is based on the progression of the disease from their normal internal location to the prolapsing external position **(Table 321-6)**.

TABLE 321-6

The Staging and Treatment of Hemorrhoids

Stage	Description of Classification	Treatment
I	Enlargement with bleeding	Fiber supplementation Short course of cortisone suppository Sclerotherapy Infared coagulation
II	Protrusion with spontaneous reduction	Fiber supplementation Short course of cortisone suppository Sclerotherapy Infared coagulation
III	Protrusion requiring manual reduction	Fiber supplementation Short course of cortisone suppository Rubber band ligation Operative hemorrhoidectomy
IV	Irreducible protrusion	Fiber supplementation Cortisone suppository Operative hemorrhoidectomy

Presentation and Evaluation

Patients commonly present to a physician for two reasons: bleeding and protrusion. Pain is less common than with fissures and, if present, is described as a dull ache from engorgement of the hemorrhoidal tissue. Severe pain may indicate a thrombosed hemorrhoid. Hemorrhoidal bleeding is described as painless bright red blood seen either in the toilet or upon wiping. Occasional patients can present with significant bleeding, which may be a cause of anemia; however, the presence of a colonic neoplasm must be ruled out in anemic patients. Patients who present with a protruding mass complain about inability to maintain perianal hygiene and are often concerned about the presence of a malignancy.

The diagnosis of hemorrhoidal disease is made on physical examination. Inspection of the perianal region for evidence of thrombosis or excoriation is performed, followed by a careful digital examination. Anoscopy is performed paying particular attention to the known position of hemorrhoidal disease. The patient is asked to strain. If this is difficult for the patient, the maneuver can be performed while sitting on a toilet. The physician is notified when the tissue prolapses. It is important to differentiate the circumferential

appearance of a full-thickness rectal prolapse from the radial nature of prolapsing hemorrhoids (see "Rectal Prolapse," above). The stage and location of the hemorrhoidal complexes are defined.

TREATMENT

TREATMENT

Hemorrhoidal Disease

The treatment for bleeding hemorrhoids is based on the stage of the disease (Table 321-6). In all patients with bleeding, the possibility of other causes must be considered. In young patients without a family history of colorectal cancer, the hemorrhoidal disease may be treated first and a colonoscopic examination performed if the bleeding continues. Older patients who have not had colorectal cancer screening should undergo colonoscopy or flexible sigmoidoscopy.

With rare exceptions, the acutely thrombosed hemorrhoid can be excised within the first 72 h by performing an elliptical excision. Sitz baths, fiber, and stool softeners are prescribed. Additional therapy for bleeding hemorrhoids includes the office procedures of rubber band ligation, infrared coagulation, and sclerotherapy. Sensation begins at the dentate line; therefore, all procedures can be performed without discomfort either endoscopically or in the office. Bands are placed around the engorged tissue, causing ischemia and fibrosis. This aids in fixing the tissue proximally in the anal canal. Patients may complain of a dull ache for 24 h following band application. During sclerotherapy, 1–2 mL of a sclerosant (usually sodium tetradecyl sulfate) is injected using a 25-gauge needle into the submucosa of the hemorrhoidal complex. Care must be taken not to inject the anal canal circumferentially, or stenosis may occur.

For surgical management of hemorrhoidal disease, excisional hemorrhoidectomy, transhemorrhoidal dearterialization (THD), or stapled hemorrhoidectomy ("the procedure for prolapse or hemorrhoids" [PPH]) is the procedure of choice. All surgical methods of management are equally effective in the treatment of symptomatic third- and fourth-degree hemorrhoids. However, because the sutured hemorrhoidectomy involves the removal of redundant tissue down to the anal verge, unpleasant anal skin tags are removed as well. The stapled hemorrhoidectomy is associated with less discomfort; however, this procedure does not remove anal skin tags and an increased number of complications are associated with use of the stapling device. THD uses ultrasound guidance to ligate the blood supply to the anal tissue, hence reducing hemorrhoidal engorgement. No procedures on hemorrhoids should be done in patients who are immunocompromised or who have active proctitis. Furthermore, emergent hemorrhoidectomy for bleeding hemorrhoids is associated with a higher complication rate.

Acute complications associated with the treatment of hemorrhoids include pain, infection, recurrent bleeding, and urinary retention. Care should be taken to place bands properly and to avoid overhydration in patients undergoing operative hemorrhoidectomy. Late complications include fecal incontinence as a result of injury to the sphincter during the dissection. Anal stenosis may develop from overzealous excision, with loss of mucosal skin bridges for reepithelialization. Finally, an *ectropion* (prolapse of rectal mucosa from the

anal canal) may develop. Patients with an ectropion complain of a “wet” anus as a result of inability to prevent soiling once the rectal mucosa is exposed below the dentate line.

ANORECTAL ABSCESS

Incidence and Epidemiology

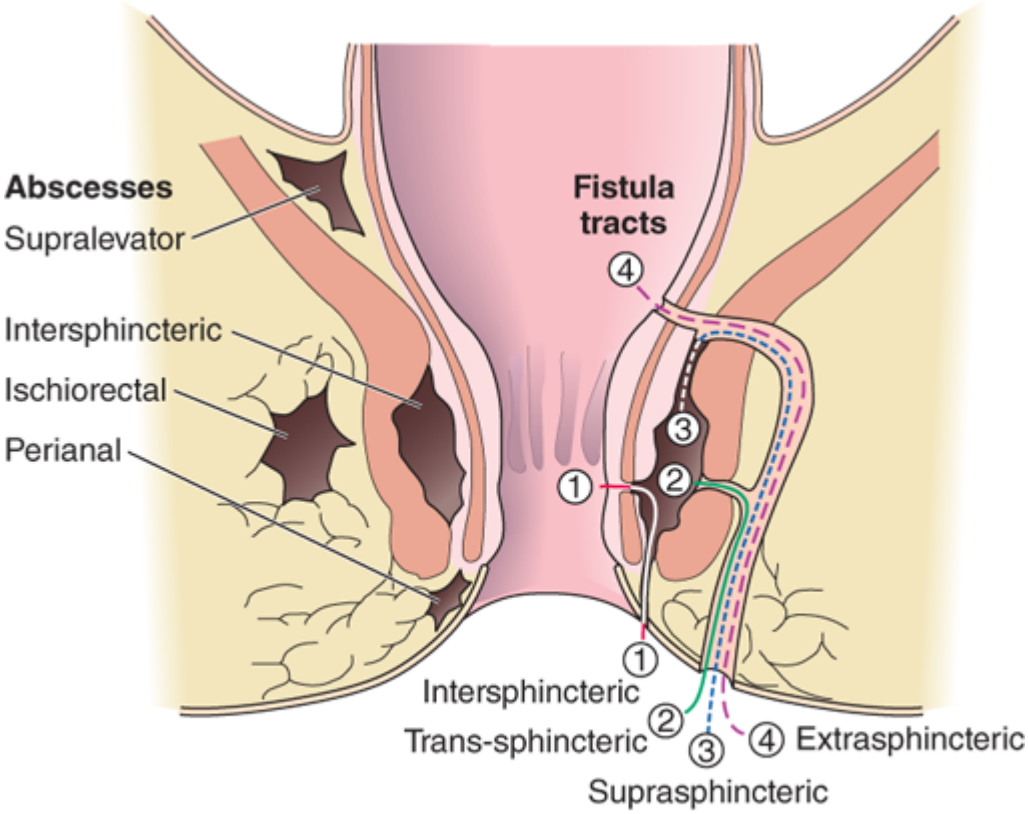
The development of a perianal abscess is more common in men than women by a ratio of 3:1. The peak incidence is in the third to fifth decade of life. Perianal pain associated with the presence of an abscess accounts for 15% of office visits to a colorectal surgeon. The disease is more prevalent in immunocompromised patients such as those with diabetes, hematologic disorders, or inflammatory bowel disease (IBD) and persons who are HIV positive. These disorders should be considered in patients with recurrent perianal infections.

Anatomy and Pathophysiology

An anorectal abscess is an abnormal fluid-containing cavity in the anorectal region. Anorectal abscess results from an infection involving the glands surrounding the anal canal. Normally, these glands release mucus into the anal canal, which aids in defecation. When stool accidentally enters the anal glands, the glands become infected and an abscess develops. Anorectal abscesses are perianal in 40–50% of patients, ischiorectal in 20–25%, intersphincteric in 2–5%, and supralevator in 2.5% **(Fig. 321-7)**.

FIGURE 321-7

Common locations of anorectal abscess (***left***) and fistula in ano (***right***).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Presentation and Evaluation

Perianal pain and fever are the hallmarks of an abscess. Patients may have difficulty voiding and have blood in the stool. A prostatic abscess may present with similar complaints, including dysuria. Patients with a prostatic abscess will often have a history of recurrent sexually transmitted diseases. On physical examination, a large fluctuant area is usually readily visible. Routine laboratory evaluation shows an elevated white blood cell count. Diagnostic procedures are rarely necessary unless evaluating a recurrent abscess. A CT scan or MRI has an accuracy of 80% in determining incomplete drainage. If there is a concern about the presence of IBD, a rigid or flexible sigmoidoscopic examination may be done at the time of drainage to evaluate for inflammation within the rectosigmoid region. A more complete evaluation for Crohn's disease would include a full colonoscopy and small-bowel series.

TREATMENT

TREATMENT

Anorectal Abscess

As with all abscesses, the "gold standard" is drainage. Office drainage of an uncomplicated anorectal abscess may suffice. A small incision close to the anal verge is made, and a Mallenkot drain is advanced into the abscess cavity. For patients who have a complicated abscess or who are diabetic or immunocompromised, drainage should be performed in an operating room under anesthesia. These patients are at greater risk for

developing necrotizing fasciitis. The role of antibiotics in the management of anorectal abscesses is limited. Antibiotics are only warranted in patients who are immunocompromised or have prosthetic heart valves, artificial joints, diabetes, or IBD.

FISTULA IN ANO

Incidence and Epidemiology

The incidence and prevalence of fistulating perianal disease parallel the incidence of anorectal abscess, estimating to be 1 in 10,000 individuals. Some 30–40% of abscesses will give rise to fistula in ano. Although the majority of the fistulas are cryptoglandular in origin, 10% are associated with IBD, tuberculosis, malignancy, and radiation.

Anatomy and Pathophysiology

A fistula in ano is defined as a communication of an abscess cavity with an identifiable internal opening within the anal canal. This identifiable opening is most commonly located at the dentate line where the anal glands enter the anal canal. Patients experiencing continuous drainage following the treatment of a perianal abscess likely have a fistula in ano. These fistulas are classified by their relationship to the anal sphincter muscles, with 70% being intersphincteric, 23% transsphincteric, 5% suprasphincteric, and 2% extrasphincteric (Fig. 321-7).

Presentation and Evaluation

A patient with a fistula in ano will complain of constant drainage from the perianal region associated with a firm mass. The drainage may increase with defecation. Perianal hygiene is difficult to maintain. Examination under anesthesia is the best way to evaluate a fistula. At the time of the examination, anoscopy is performed to look for an internal opening. Diluted hydrogen peroxide will aid in identifying such an opening. In lieu of anesthesia, MRI with an endoanal coil will also identify tracts in 80% of the cases. After drainage of an abscess with insertion of a Mallenkot catheter, a fistulagram through the catheter can be obtained in search of an occult fistula tract. Goodsall's rule states that a posterior external fistula will enter the anal canal in the posterior midline, whereas an anterior fistula will enter at the nearest crypt. A fistula exiting >3 cm from the anal verge may have a complicated upward extension and may not obey Goodsall's rule.

TREATMENT

TREATMENT

Fistula in Ano

A newly diagnosed draining fistula is best managed with placement of a seton, a vessel loop or silk tie placed through the fistula tract, which maintains the tract open and quiets down the surrounding inflammation that occurs from repeated blockage of the tract. Once the inflammation is less, the exact relationship of the fistula

tract to the anal sphincters can be ascertained. A simple fistulotomy can be performed for intersphincteric and low (less than one-third of the muscle) transsphincteric fistulas without compromising continence. For a higher transsphincteric fistula, an anorectal advancement flap in combination with a drainage catheter or fibrin glue may be used. Very long (>2 cm) and narrow tracts respond better to fibrin glue than shorter tracts. Simple ligation of the internal fistula tract (LIFT procedure) has also been used in the management of simple fistula with good success.

Patients should be maintained on stool-bulking agents, nonnarcotic pain medication, and sitz baths following surgery for a fistula. Early complications from these procedures include urinary retention and bleeding. Later complications are rare (<10%) and include temporary and permanent incontinence. Recurrence is 0–18% following fistulotomy and 20–30% following anorectal advancement flap and the LIFT procedure. The use of stem cell implants at the time of repair for recalcitrant fistulizing disease of the anus is being studied.

ANAL FISSURE

Incidence and Epidemiology

Anal fissures occur at all ages but are more common in the third through the fifth decades. A fissure is the most common cause of rectal bleeding in infancy. The prevalence is equal in males and females. It is associated with constipation, diarrhea, infectious etiologies, perianal trauma, and Crohn’s disease.

Anatomy and Pathophysiology

Trauma to the anal canal occurs following defecation. This injury occurs in the anterior or, more commonly, the posterior anal canal. Irritation caused by the trauma to the anal canal results in an increased resting pressure of the internal sphincter. The blood supply to the sphincter and anal mucosa enters laterally. Therefore, increased anal sphincter tone results in a relative ischemia in the region of the fissure and leads to poor healing of the anal injury. A fissure that is not in the posterior or anterior position should raise suspicion for other causes, including tuberculosis, syphilis, Crohn’s disease, and malignancy.

Presentation and Evaluation

A fissure can be easily diagnosed on history alone. The classic complaint is pain, which is strongly associated with defecation and is relentless. The bright red bleeding that can be associated with a fissure is less extensive than that associated with hemorrhoids. On examination, most fissures are located in either the posterior or anterior position. A lateral fissure is worrisome because it may have a less benign nature, and systemic disorders should be ruled out. A chronic fissure is indicated by the presence of a hypertrophied anal papilla at the proximal end of the fissure and a sentinel pile or skin tag at the distal end. Often the circular fibers of the hypertrophied internal sphincter are visible within the base of the fissure. If anal manometry is performed, elevation in anal resting pressure and a sawtooth deformity with paradoxical contractions of the sphincter muscles are pathognomonic.

TREATMENT

TREATMENT

Anal Fissure

The management of the acute fissure is conservative. Stool softeners for those with constipation, increased dietary fiber, topical anesthetics, glucocorticoids, and sitz baths are prescribed and will heal 60–90% of fissures. Chronic fissures are those present for >6 weeks. These can be treated with modalities aimed at decreasing the anal canal resting pressure including nifedipine ointment applied three times a day and botulinum toxin type A, up to 20 units, injected into the internal sphincter on each side of the fissure. Surgical management includes anal dilatation and lateral internal sphincterotomy. Usually, one-third of the internal sphincter muscle is divided; it is easily identified because it is hypertrophied. Recurrence rates from medical therapy are higher, but this is offset by a risk of incontinence following sphincterotomy. Lateral internal sphincterotomy may lead to incontinence more commonly in women.

ACKNOWLEDGMENT

We would like to thank Cory Sandore for providing some illustrations for this chapter. Gregory Bulkley, MD, contributed to this chapter in an earlier edition and some of that material has been retained here.

FURTHER READING

Bharucha AE, Rao SSC, Shin AS: Surgical interventions and the use of device-aided therapy for the treatment of fecal incontinence and defecatory disorders. Clin Gastroenterol Hepatol 15:1844, 2017.

Cotter TG et al: Approach to the patient with hematochezia. Mayo Clin Proc 92:797, 2017. [PubMed: 28473039]

Guttenplan M: The evaluation and office management of hemorrhoids for the gastroenterologist. Curr Gastroenterol Rep 19:30, 2017. [PubMed: 28567655]

Prichard D, Bharucha AE: Management of pelvic floor disorders: Biofeedback and more. Curr Treat Options Gastroenterol 12:456, 2014. [PubMed: 25267107]

Salfity HV et al: Minimally invasive incision and drainage technique in the treatment of simple subcutaneous abscess in adults. Am Surg 83:699, 2017. [PubMed: 28738938]

Sugrue J et al: Sphincter-sparing anal fistula repair: Are we getting better? Dis Colon Rectum 60:1071, 2017. [PubMed: 28891851]

Tursi A: Dietary pattern and colonic diverticulosis. Curr Opin Clin Nutr Metab Care 20:409, 2017. [PubMed: 28768296]

Chapter 322: Mesenteric Vascular Insufficiency

Satinderjit Locham; Mahmoud Malas

FIGURE 322-1

INTESTINAL ISCHEMIA

INCIDENCE AND EPIDEMIOLOGY

Intestinal ischemia occurs when splanchnic perfusion fails to meet the metabolic demands of the intestines, resulting in ischemic tissue injury. Mesenteric ischemia affects 2–3 people per 100,000, with an increasing incidence in the aging population. Delay in diagnosis and management results in a high mortality, and prompt interventions may be lifesaving. Intestinal ischemia is further classified based on etiology, which dictates management: (1) arterioocclusive mesenteric ischemia, (2) nonocclusive mesenteric ischemia, and (3) mesenteric venous thrombosis.

Risk factors for arterioocclusive mesenteric ischemia are generally acute in onset that include atrial fibrillation, recent myocardial infarction, valvular heart disease, and recent cardiac or vascular catheterization, all of which result in embolic clots reaching the mesenteric circulation. Nonocclusive mesenteric ischemia, also known as “intestinal angina,” is generally more insidious and often seen in the aging population affected by atherosclerotic disease. Patients with chronic atherosclerotic disease could also suffer an acute insult from emboli leading to complete occlusion. Nonocclusive mesenteric ischemia is also seen in patients receiving high-dose vasopressor infusions, patients presenting with cardiogenic or septic shock, and cocaine overdose. It is the most prevalent gastrointestinal disease complicating cardiovascular surgery. The incidence of ischemic colitis following elective aortic repair is 5–9%, and the incidence triples in patients following emergent repair. Mesenteric venous thrombosis is less common and is associated with the presence of a hypercoagulable state including protein C or S deficiency, antithrombin III deficiency, polycythemia vera, and carcinoma.

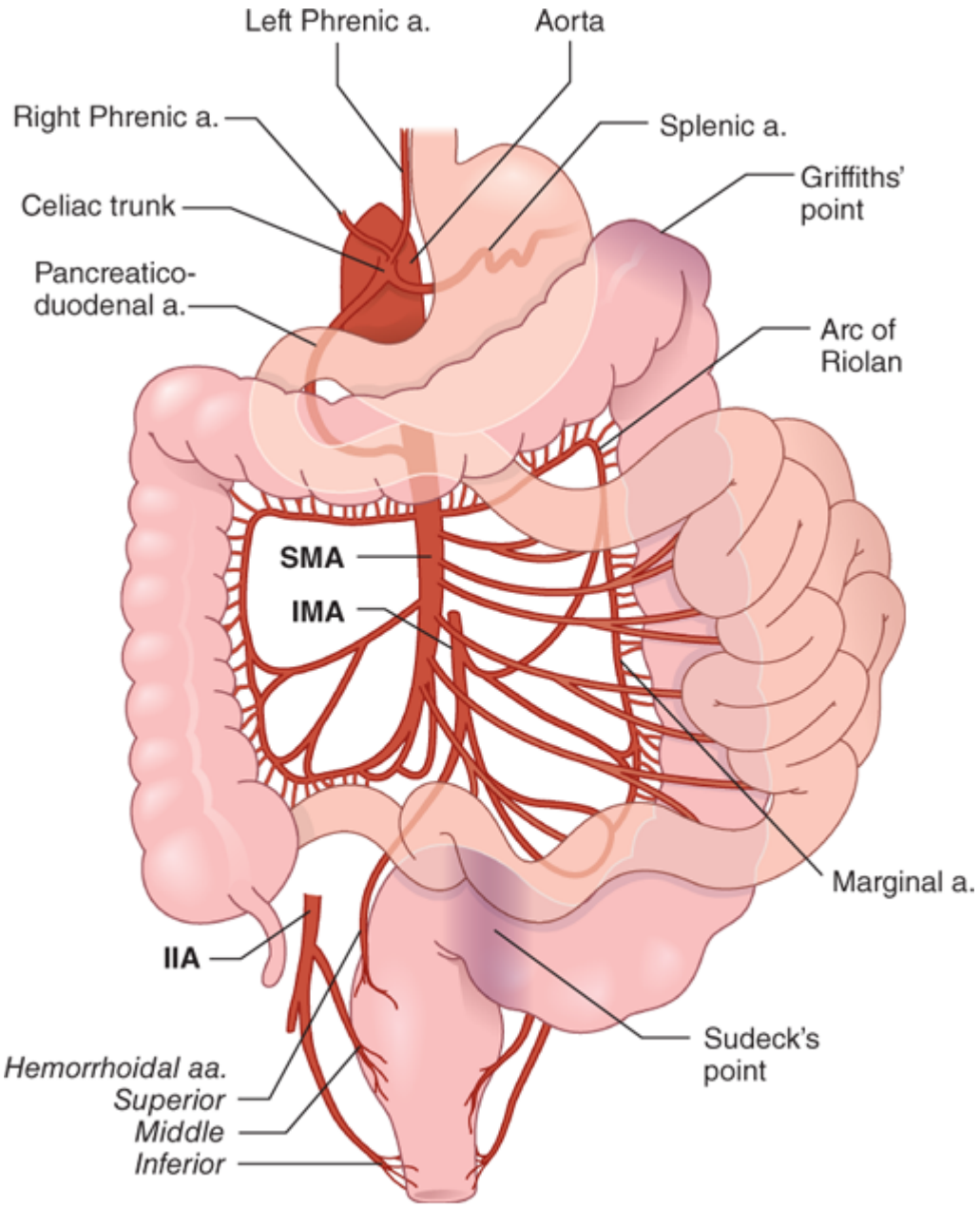
ANATOMY AND PATHOPHYSIOLOGY

The blood supply to the intestines is depicted in **Fig. 322-1**. To prevent ischemic injury, extensive collateralization occurs between major mesenteric trunks and branches of the mesenteric arcades. Collateral vessels within the small bowel are numerous and meet within the duodenum and the bed of the pancreas. Collateral vessels within the colon meet at the splenic flexure and descending/sigmoid colon. These areas, which are inherently at risk for decreased blood flow, are known as *Griffiths’ point* and *Sudeck’s point*,

respectively, and are the most common locations for colonic ischemia (Fig. 322-1, shaded areas). The splanchnic circulation can receive up to 30% of the cardiac output. Protective responses to prevent intestinal ischemia include abundant collateralization, autoregulation of blood flow, and the ability to increase oxygen extraction from the blood.

FIGURE 322-1

Blood supply to the intestines includes the celiac artery, superior mesenteric artery (SMA), inferior mesenteric artery (IMA), and branches of the internal iliac artery (IIA). Griffiths' and Sudeck's points, indicated by shaded areas, are watershed areas within the colonic blood supply and common locations for ischemia.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Occlusive ischemia is a result of disruption of blood flow by an embolus or progressive thrombosis in a major artery supplying the intestine. In >75% of cases, emboli originate from the heart and preferentially lodge in

the superior mesenteric artery (SMA) just distal to the origin of the middle colic artery. Progressive thrombosis of at least two of the major vessels supplying the intestine is required for the development of chronic intestinal angina. Nonocclusive ischemia is disproportionate mesenteric vasoconstriction (arteriolar vasospasm) in response to a severe physiologic stress such as shock. If left untreated, early mucosal stress ulceration will progress to full-thickness injury. Even in the early stages of ischemia, there is translocation of bacteria across the intestinal mucosa, resulting in bacteremia that can lead to sepsis.

PRESENTATION, EVALUATION, AND MANAGEMENT

Intestinal ischemia remains one of the most challenging diagnoses. The mortality rate is >50%. The most significant indicator of survival is the timeliness of diagnosis and treatment. An overview of diagnosis and management of each form of intestinal ischemia is given in **Table 322-1**.

TABLE 322-1

Overview of the Management of Acute Intestinal Ischemia

Condition	Key to Early Diagnosis	Treatment of Underlying Cause	Treatment of Specific Lesion	Treatment of Systemic Consequence
Arterioocclusive mesenteric ischemia 1. Arterial embolus	Computed tomography (CT) angiography Early laparotomy	Anticoagulation Cardioversion Proximal thrombectomy	Laparotomy Embolectomy Vascular bypass Assess viability and resect dead bowel	Ensure hydration Give antibiotics Reverse acidosis Optimize oxygen delivery Avoid vasoconstrictors
2. Arterial thrombosis	Duplex ultrasound Angiography	Anticoagulation Hydration	Endovascular approach: thrombolysis, angioplasty and stenting Endarterectomy/thrombectomy or vascular bypass Assess viability and resect dead bowel	Give antibiotics Reverse acidosis Optimize oxygen delivery Support cardiac output Avoid vasoconstrictors
Mesenteric venous thrombosis Venous thrombosis	Spiral CT Angiography with venous phase	Anticoagulation Massive hydration	Anticoagulation ± laparotomy/thrombectomy/catheter-directed thrombolysis Assess viability and resect dead bowel	Give antibiotics Reverse acidosis Optimize oxygen delivery Support cardiac output Avoid vasoconstrictors

Condition	Key to Early Diagnosis	Treatment of Underlying Cause	Treatment of Specific Lesion	Treatment of Systemic Consequence
Nonocclusive mesenteric ischemia	Vasospasm: Angiography Hypoperfusion: Spiral CT or colonoscopy	Ensure hydration Support cardiac output Avoid vasoconstrictors	Vasospasm Intraarterial vasodilators Hypoperfusion Delayed laparotomy Assess viability and resect dead bowel	Ensure hydration Give antibiotics Reverse acidosis Optimize oxygen delivery Support cardiac output Avoid vasoconstrictors

Source: Modified from GB Bulkley, in JL Cameron (ed): *Current Surgical Therapy*, 2nd ed. Toronto, BC Decker, 1986.

Acute mesenteric ischemia resulting from arterial embolus or thrombosis presents with severe acute, nonremitting abdominal pain strikingly out of proportion to the physical findings. Associated symptoms may include nausea and vomiting, transient diarrhea, anorexia, and bloody stools. With the exception of minimal abdominal distention and hypoactive bowel sounds, early abdominal examination is unimpressive. Later findings will demonstrate peritonitis and cardiovascular collapse. In the evaluation of acute intestinal ischemia, routine laboratory tests should be obtained, including complete blood count, serum chemistry, coagulation profile, arterial blood gas, amylase, lipase, lactic acid, blood type and cross match, and cardiac enzymes. Regardless of the need for urgent surgery, emergent admission to a monitored bed or intensive care unit is recommended for resuscitation and further evaluation. If the diagnosis of intestinal ischemia is being considered, consultation with a surgical service is necessary. Often the decision to operate is made on a high index of suspicion from the history and physical examination despite normal laboratory findings.

Other diagnostic modalities that may be useful in diagnosis but should not delay surgical therapy include electrocardiogram (ECG), echocardiogram, abdominal radiographs, computed tomography (CT), and mesenteric angiography. More recently, mesentery duplex scanning and visible light spectroscopy during colonoscopy have been demonstrated to be beneficial. The ECG may demonstrate an arrhythmia, indicating the possible source of the emboli. A plain abdominal film may show evidence of free intraperitoneal air, indicating a perforated viscus and the need for emergent exploration. Earlier features of intestinal ischemia seen on abdominal radiographs include bowel-wall edema, known as “thumbprinting.” If the ischemia progresses, air can be seen within the bowel wall (pneumatosis intestinalis) and within the portal venous system. Other features include calcifications of the aorta and its tributaries, indicating atherosclerotic disease. With the administration of oral and IV contrast, dynamic CT angiography with three-dimensional

reconstruction is a highly sensitive test for intestinal ischemia. In acute embolic disease, mesenteric angiography is best performed intraoperatively. A mesenteric duplex scan demonstrating a high peak velocity of flow in the SMA is associated with an ~80% positive predictive value of mesenteric ischemia. More significantly, a negative duplex scan virtually precludes the diagnosis of mesenteric ischemia. Duplex imaging serves as a screening test; however, further investigations with angiography are usually needed. One of the biggest limitations of duplex scanning is patients' body habitus. The duplex imaging yields poor results in obese patients. Nevertheless, "food fear" in patients with chronic disease often leads to a decreased appetite and lower abdominal fat, thus, yielding high duplex imaging results. The endoscopic techniques such as visible light spectroscopy can also be used in the diagnosis of chronic ischemia. When suspecting mesenteric ischemia involving the colon, performing an endoscopy to evaluate up to the splenic flexure is high yield. This is often an excellent diagnostic tool in patients with chronic renal insufficiency who cannot tolerate IV contrast.

The "gold standard" for the diagnosis of acute arterial occlusive disease is angiography, and management is laparotomy. Surgical exploration should not be delayed if suspicion of acute occlusive mesenteric ischemia is high or evidence of clinical deterioration or frank peritonitis is present. The goal of operative exploration is to resect compromised bowel and restore blood supply. The entire length of the small and large bowel beginning at the ligament of Treitz should be evaluated. The pattern of intestinal ischemia may indicate the level of arterial occlusion. In the case of SMA occlusion where the embolus usually lies just proximal to the origin of the middle colic artery, the proximal jejunum is often spared while the remainder of the small bowel up to the transverse colon will be ischemic. The surgical management of acute mesenteric ischemia of the small bowel is embolectomy via arteriotomy; a small incision is made in the artery through which the clot is retrieved. Another way to manage acute thrombosis is thrombolysis therapy and angioplasty with stent placement. However, this approach is more commonly applied to treat chronic mesenteric ischemia. If this is unsuccessful, a bypass from the aorta or iliac artery to the SMA is performed.

Nonocclusive or vasospastic mesenteric ischemia presents with generalized abdominal pain, anorexia, bloody stools, and abdominal distention. Often these patients are obtunded, and physical findings may not assist in the diagnosis. The presence of a leukocytosis, metabolic acidosis, elevated amylase or creatinine phosphokinase levels, and/or lactic acidosis is useful in support of the diagnosis of advanced intestinal ischemia; however, these markers may not be indicative of either reversible ischemia or frank necrosis. Investigational markers for intestinal ischemia include D-dimer, glutathione *S*-transferase, platelet-activating factor (PAF), and mucosal pH monitoring. Regardless of the need for urgent surgery, emergent admission to a monitored bed or intensive care unit is recommended for resuscitation and further evaluation. Early manifestations of intestinal ischemia include fluid sequestration within the bowel wall leading to a loss of interstitial volume. Aggressive fluid resuscitation may be necessary. To optimize oxygen delivery, nasal O_2 and blood transfusions may be given. Broad-spectrum antibiotics should be given to provide sufficient coverage for enteric pathogens, including gram-negative and anaerobic organisms. Frequent monitoring of the patient's vital signs, urine output, blood gases, and lactate levels is paramount, as is frequent abdominal examination. All vasoconstricting agents should be avoided; fluid resuscitation is the intervention of choice to maintain hemodynamics.

If ischemic colitis is a concern, colonoscopy should be performed to assess the integrity of the colon mucosa. Visualization of the rectosigmoid region may demonstrate decreased mucosal integrity, associated more commonly with nonocclusive mesenteric ischemia, or, on occasion, occlusive disease as a result of acute loss of inferior mesenteric arterial flow following aortic surgery. Ischemia of the colonic mucosa is graded as *mild* with minimal mucosal erythema or as *moderate* with pale mucosal ulcerations and evidence of extension to the muscular layer of the bowel wall. *Severe* ischemic colitis presents with severe ulcerations resulting in black or green discoloration of the mucosa, consistent with full-thickness bowel-wall necrosis. The degree of reversibility can be predicted from the mucosal findings: mild erythema is nearly 100% reversible, moderate is ~50% reversible, and frank necrosis is simply dead bowel. Follow-up colonoscopy can be performed to rule out progression of ischemic colitis.

Laparotomy for nonocclusive mesenteric ischemia is warranted in patients with signs of peritonitis or worsening endoscopic findings and if the patient's condition does not improve with aggressive resuscitation. Ischemic colitis is optimally treated with resection of the ischemic bowel and formation of a proximal stoma. Primary anastomosis should not be performed in patients with acute intestinal ischemia.

Patients with mesenteric venous thrombosis may present with a gradual or sudden onset of pain. Symptoms include vague abdominal pain, nausea, and vomiting. Physical examination findings include abdominal distention with mild to moderate tenderness and signs of dehydration. The diagnosis of mesenteric thrombosis is frequently made on abdominal spiral CT with oral and IV contrast. Findings on CT angiography with venous phase include bowel-wall thickening and ascites. IV contrast will demonstrate a delayed arterial phase and clot within the superior mesenteric vein. The goal of management is to optimize hemodynamics and correct electrolyte abnormalities with massive fluid resuscitation. Intravenous antibiotics as well as anticoagulation should be initiated. If laparotomy is performed and mesenteric venous thrombosis is suspected, heparin anticoagulation is immediately initiated, and compromised bowel is resected. Of all acute intestinal disorders, mesenteric venous insufficiency is associated with the best prognosis.

Chronic intestinal ischemia presents with intestinal angina or postprandial abdominal pain associated with increased need of blood flow to the intestine following meals. Patients report abdominal cramping and pain following ingestion of a meal. Weight loss and chronic diarrhea may also be noted. Abdominal pain without weight loss is not chronic mesenteric angina. Physical examination will often reveal a malnourished patient with an abdominal bruit as well as other manifestations of atherosclerosis. Duplex ultrasound evaluation of the mesenteric vessels has gained in popularity. It is important to perform the test while the patient is fasting because the presence of increased bowel gas prevents adequate visualization of flow disturbances within the vessels or the lack of a vasodilation response to feeding during the test. This tool is frequently used as a screening test for patients with symptoms suggestive of chronic mesenteric ischemia. The gold standard for confirmation of mesenteric arterial occlusion is mesenteric angiography. Evaluation with mesenteric angiography allows for identification and possible intervention for the treatment of atherosclerosis within the vessel lumen and will also evaluate the patency of remaining mesenteric vessels. The use of mesenteric angiography may be limited in the presence of renal failure or contrast allergy. Magnetic resonance angiography is an alternative if the administration of contrast dye is contraindicated.

The management of chronic intestinal ischemia includes medical management of atherosclerotic disease by exercise, cessation of smoking, and antiplatelet and lipid-lowering medications. A full cardiac evaluation should be performed before intervention on chronic mesenteric ischemia. Newer endovascular procedures may avoid an operative intervention in selected patient populations. Angioplasty with endovascular stenting in the treatment of chronic mesenteric ischemia is associated with an 80% long-term success rate. In patients requiring surgical exploration, the approach used is determined by findings of the mesenteric angiogram. The entire length of the small and large bowel should be evaluated, beginning at the ligament of Treitz. Restoration of blood flow at the time of laparotomy is accomplished with mesenteric vessel endarterectomy or bypass.

Determination of intestinal viability intraoperatively in patients with suspected intestinal ischemia can be challenging. After revascularization, peristalsis and return of a pink color of the bowel wall should be observed. Palpation of major arterial mesenteric vessels can be performed, as well as applying a Doppler flowmeter to the antimesenteric border of the bowel wall, but neither is a definitive indicator of viability. In equivocal cases, 1 g of IV sodium fluorescein is administered, and the pattern of bowel reperfusion is observed under ultraviolet illumination with a standard (3600 A) Wood's lamp. An area of nonfluorescence >5 mm in diameter suggests nonviability. If doubt persists, reexploration performed 24–48 h following surgery will allow demarcation of nonviable bowel. Primary intestinal anastomosis in patients with ischemic bowel is always worrisome; thus, delayed bowel reconstruction and reanastomosis should be deferred to the time of second-look laparotomy.

ACKNOWLEDGMENTS

We thank Cory Sandore for providing the illustration for this chapter. Susan Gearhart contributed to this chapter in the 18th edition and Rizwan Ahmed contributed to the 19th edition.

FURTHER READING

Deng QW et al: Risk factors for postoperative acute mesenteric ischemia among adult patients undergoing cardiac surgery: A systematic review and meta-analysis. J Crit Care 42:294, 2017. [PubMed: 28820981]

Salsano G et al: What is the best revascularization strategy for acute occlusive arterial mesenteric ischemia: Systematic review and meta-analysis. Cardiovasc Intervent Radiol 41:27, 2018.

Sise MJ: Acute mesenteric ischemia. Surg Clin North Am 94:165–81, 2014. [PubMed: 24267504]

Chapter 323: Acute Intestinal Obstruction

Danny O. Jacobs

EPIDEMIOLOGY

Morbidity and mortality from acute intestinal obstruction have been decreasing over the past several decades. Nevertheless, the diagnosis can still be challenging, and the type of complications that patients suffer has not changed significantly. The extent of mechanical obstruction is typically described as partial, high-grade, or complete—generally correlating with the risk of complications and the urgency with which the underlying disease process must be addressed. Obstruction is also commonly described as being either “simple” or, alternatively, “strangulated” if vascular insufficiency and intestinal ischemia are evident.

Acute intestinal obstruction occurs either *mechanically* from blockage or from intestinal dysmotility when there is no blockage. In the latter instance, the abnormality is described as being *functional*. Mechanical bowel obstruction may be caused by extrinsic processes, intrinsic abnormalities of the bowel wall, or intraluminal abnormalities (Table 323-1). Within each of these broad categories are many diseases that can impede intestinal propulsion. Intrinsic diseases that can cause intestinal obstruction are usually congenital, inflammatory, neoplastic, or traumatic in origin, although intussusception and radiation injury can also be etiologic.

TABLE 323-1

Most Common Causes of Acute Intestinal Obstruction

Extrinsic Disease
Adhesions (especially due to previous abdominal surgery), internal or external hernias, neoplasms (including carcinomatosis and extraintestinal malignancies, mostly commonly ovarian), endometriosis or intraperitoneal abscesses, and idiopathic sclerosis
Intrinsic Disease
Congenital (e.g., malrotation, atresia, stenosis, intestinal duplication, cyst formation, and congenital bands—the latter rarely in adults)
Inflammation (e.g., inflammatory bowel disease, especially Crohn's disease, but also diverticulitis, radiation, tuberculosis, lymphogranuloma venereum, and schistosomiasis)
Neoplasia (note: primary small-bowel cancer is rare; obstructive colon cancer may mimic small-bowel obstruction if the ileocecal valve is incompetent)
Traumatic (e.g., hematoma formation, anastomotic strictures)
Other, including intussusception (where the lead point is typically a polyp or tumor in adults), volvulus, obstruction of duodenum by superior mesenteric artery, radiation or ischemic injury, and aganglionosis, which is Hirschsprung's disease
Intraluminal Abnormalities
Bezoars, feces, foreign bodies including inspissated barium, gallstones (entering the lumen via a cholecystoenteric fistula), enteroliths

Acute intestinal obstruction accounts for ~1–3% of all hospitalizations and a quarter of all urgent or emergent general surgery admissions. Approximately 80% of cases involve the small bowel, and about one-third of these patients show evidence of significant ischemia. The mortality rate for patients with strangulation who are operated on within 24–30 h of the onset of symptoms is ~8% but triples shortly thereafter.

Extrinsic diseases most commonly cause mechanical obstruction of the small intestine. In the United States and Europe, almost all cases are caused by postoperative adhesions, carcinomatosis, or herniation of the anterior abdominal wall. Carcinomatosis most often originates from the ovary, pancreas, stomach, or colon, although rarely, metastasis from distant organs like the breast and skin can occur. Adhesions are responsible

for the majority of cases of early postoperative obstruction that require intervention. It is important to note many patients who are successfully treated for adhesive small-bowel instruction will recur. Approximately 20% of patients who were treated conservatively and between 5 and 30% of patients who were managed operatively will require readmission within 10 years.

Open operations of the lower abdomen, including appendectomy and colorectal and gynecologic procedures, are especially likely to create adhesions that can cause bowel obstruction (Table 323-2). The risk of internal herniation is increased by abdominal procedures such as laparoscopic or open Roux-en-Y gastric bypass. Although laparoscopic procedures may generate fewer postoperative adhesions compared with open surgery, the risk of obstructive adhesion formation is not eliminated.

TABLE 323-2

Acute Small-Intestinal and Colonic Obstruction Incidences

Cause	Incidence
Postoperative adhesions	>50%
Neoplasms	~20%
Hernias (especially ventral or internal types, where the risk of strangulation is increased)	~10%
Inflammatory bowel disease, other inflammation (obstruction may resolve if acute inflammation and edema subside)	~5%
Intussusception, volvulus, other miscellaneous diseases	<15%

Volvulus, which occurs when bowel twists on its mesenteric axis, can cause partial or complete obstruction and vascular insufficiency. The sigmoid colon is most commonly affected, accounting for approximately two-thirds of all cases of volvulus and 4% of all cases of large-bowel obstruction. The cecum and terminal ileum can also volvulize, or the cecum alone may be involved as a cecal bascule. Risk factors include institutionalization, the presence of neuropsychiatric conditions requiring psychotropic medication, chronic constipation, and aging; patients typically present in their seventies or eighties.

Colonic volvulus is more common in Eastern Europe, Russia, and Africa than it is in the United States. It is rare for adhesions or hernias to obstruct the colon. Cancer of the descending colon and rectum is responsible for approximately two-thirds of all cases, followed by diverticulitis and volvulus.

Functional obstruction, also known as *ileus* and *pseudo-obstruction*, is present when dysmotility prevents intestinal contents from being propelled distally and no mechanical blockage exists. Ileus that occurs after intraabdominal surgery is the most commonly identified form of functional bowel obstruction, but there are

many other causes (Table 323-3). Although postoperative ileus is most often transient, it is often the most common reason why hospital discharge is delayed. Pseudo-obstruction of the colon, also known as Ogilvie's syndrome, is a relatively rare disease. Some patients with Ogilvie's syndrome have colonic dysmotility due to abnormalities of their autonomic nervous system that may be inherited.

TABLE 323-3

Most Common Causes of Ileus (Functional or Pseudo-Obstruction of the Intestine)

Intraabdominal procedures, lumbar spinal injuries, or surgical procedures on the lumbar spine and pelvis
Metabolic or electrolyte abnormalities, especially hypokalemia and hypomagnesemia, but also hyponatremia, uremia, and severe hyperglycemia
Drugs such as opiates, antihistamines, and some psychotropic (e.g., haloperidol, tricyclic antidepressants) and anticholinergic agents
Intestinal ischemia
Intraabdominal or retroperitoneal inflammation or hemorrhage
Lower lobe pneumonias
Intraoperative radiation (likely due to muscle damage)
Systemic sepsis
Hyperparathyroidism
Pseudo-obstruction (Ogilvie's syndrome)
Ileus secondary to hereditary or acquired visceral myopathies and neuropathies that disrupt myocellular neural coordination
Some collagen vascular diseases such as lupus erythematosus or scleroderma

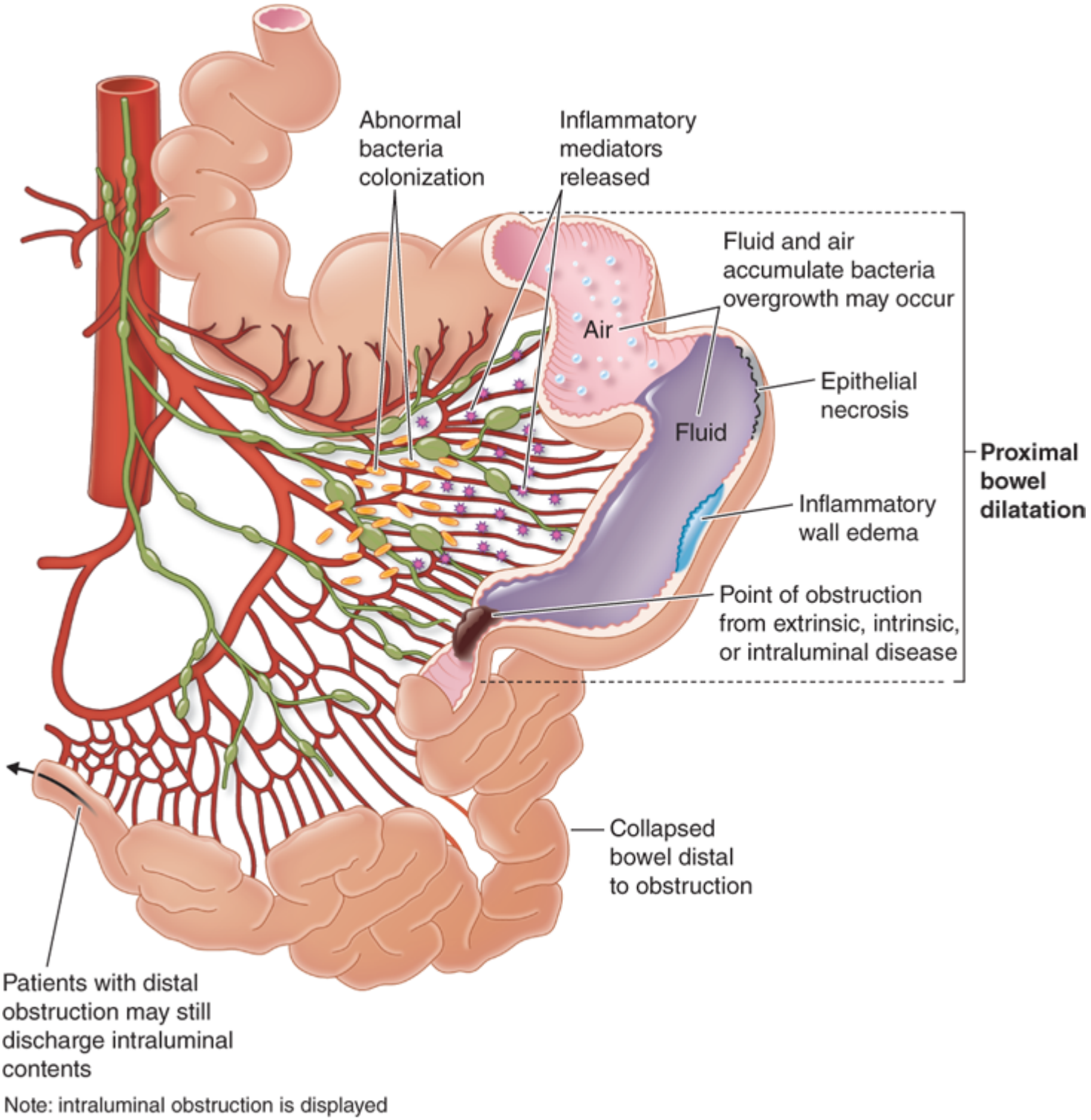
PATHOPHYSIOLOGY

The manifestations of acute intestinal obstruction depend on the nature of the underlying disease process, its location, and changes in blood flow (Fig. 323-1). Increased intestinal contractility, which occurs proximally *and* distal to the obstruction, is a characteristic response. Subsequently, intestinal peristalsis slows as the

intestine or stomach proximal to the point of obstruction dilates and fills with gastrointestinal secretions and swallowed air. Although swallowed air is the primary contributor to intestinal distension, intraluminal air may also accumulate from fermentation, local carbon dioxide production, and altered gaseous diffusion.

FIGURE 323-1

Pathophysiologic changes of small-bowel obstruction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Intraluminal dilation also increases intraluminal pressure. When luminal pressure exceeds venous pressure, venous and lymphatic drainage is impeded. Edema ensues, and the bowel wall proximal to the site of blockage may become hypoxemic. Epithelial necrosis can be identified within 12 h of obstruction. Ultimately, arterial blood supply may become so compromised that full-thickness ischemia, necrosis, and perforation result. Stasis increases the bacteria counts within the jejunum and ileum. Bacteria like *Escherichia coli*, *Streptococcus faecalis*, and *Klebsiella*, and other pathogens may be recovered from intestinal cultures, mesenteric lymph nodes, the blood stream, and other sites.

Other manifestations depend on the degree of hypovolemia, the patient's metabolic response, and the presence or absence of associated intestinal ischemia. Inflammatory edema eventually increases the production of reactive oxygen species and activates neutrophils and macrophages, which accumulate within the bowel wall. Their accumulation, along with changes in innate immunity, disrupts secretory and neuromotor processes. Dehydration is caused by loss of the normal intestinal absorptive capacity as well as fluid accumulation in the gastric or intestinal wall and intraperitoneally.

Anorexia and emesis tend to exacerbate intravascular volume depletion. In the worst case scenario that is most commonly identified after high-grade distal obstruction, emesis leads to losses of gastric potassium, hydrogen, and chloride, while dehydration stimulates proximal renal tubule bicarbonate reabsorption. Intraperitoneal fluid accumulation, especially in patients with severe distal bowel obstruction, may increase intraabdominal pressure enough to elevate the diaphragm, inhibit respiration and to impede systemic venous return and promote vascular instability. Severe hemodynamic compromise may elicit a systemic inflammatory response and generalized microvascular leakage.

Closed-loop obstruction results when the proximal and distal openings of a given bowel segment are both occluded, for example, due to volvulus or a hernia. It is the most common precursor for strangulation, but not every closed loop strangulates. The risk of vascular insufficiency, systemic inflammation, hemodynamic compromise, and irreversible intestinal ischemia is much greater in patients with closed-loop obstruction. Pathologic changes may occur more rapidly, and emergency intervention is indicated. Irreversible bowel ischemia may progress to transmural necrosis even if obstruction is relieved. It is also important to remember that patients with high-grade distal colonic obstruction who have competent ileocecal valves may present with closed-loop obstruction. In the latter instance, the cecum may progressively dilate such that ischemic necrosis results in perforation especially when the cecal diameter exceeds 10–12 cm, as informed by Laplace's law. Patients with distal colonic obstruction whose ileocecal valves are incompetent tend to present later in the course of disease and mimic patients with distal small-bowel obstruction.

HISTORY AND PHYSICAL FINDINGS

Even though the presenting signs and symptoms can be misleading, many patients with acute obstruction can be accurately diagnosed after a thorough history and physical examination is performed. However, small-bowel obstruction with strangulation can be especially difficult to diagnosis promptly. Early recognition allows earlier treatment that decreases the risk of progression or other excess morbidity.

The cardinal signs are colicky abdominal pain, abdominal distention, emesis, and obstipation. More intraluminal fluid accumulates in patients with distal obstruction, which typically leads to greater distention, more discomfort, and delayed emesis. This emesis is feculent when there is bacterial overgrowth. Patients with more proximal obstruction commonly present with less abdominal distention but more pronounced vomiting. Elements of the history that might be helpful include any prior history of surgery, including herniorrhaphy, as well as any history of cancer or inflammatory bowel disease.

Most patients, even those with simple obstruction, appear to be critically ill. Many may be oliguric, hypotensive, and tachycardic because of severe intravascular volume depletion. Fever is worrisome for strangulation or systemic inflammation. Bowel sounds and bowel functional activity are notoriously difficult to interpret. Classically, many patients with early small-bowel obstruction will have high-pitched, “musical” tinkling bowel sounds and peristaltic “rushes” known as borborygmi. Later in the course of disease, the bowel sounds may be absent or hypoactive as peristaltic activity decreases. This is in contrast to the common findings in patients with ileus or pseudo-obstruction where bowel sounds are typically absent or hypoactive from the beginning. Lastly, patients with partial blockage may continue to pass flatus and stool, and those with complete blockage may evacuate bowel contents present downstream beyond their obstruction.

All surgical incisions should be examined and the presence of a tender abdominal or groin mass strongly suggests that an incarcerated hernia may be the cause of obstruction. The presence of tenderness should increase the concern about the presence of complications such as ischemia, necrosis, or peritonitis. Severe pain with localization or signs of peritoneal irritation is suspicious for strangulated or closed-loop obstruction. It is important to remember that the discomfort may be out of proportion to physical findings mimicking the complaints of patients with acute mesenteric ischemia. Patients with colonic volvulus present with the classic manifestations of closed-loop obstruction: severe abdominal pain, vomiting, and obstipation. Asymmetrical abdominal distension and a tympanic mass may be evident.

Patients with ileus or pseudo-obstruction may have signs and symptoms similar to those of bowel obstruction. Although abdominal distention is present, colicky abdominal pain is typically absent, and patients may not have nausea or emesis. Ongoing, regular discharge of stool or flatus can sometimes help distinguish patients with ileus from those with complete mechanical bowel obstruction.

LABORATORY AND IMAGING STUDIES

Laboratory testing should include a complete blood count and serum electrolyte and creatinine measurements. Serial assessments are often useful. Mild hemoconcentration and slight elevation of the white blood cell count commonly occur after simple bowel obstruction. Emesis and dehydration may cause hypokalemia, hypochloremia, elevated blood urea nitrogen–to–creatinine ratios, and metabolic alkalosis. Patients may be hyponatremic on admission because many have attempted to rehydrate themselves with hypotonic fluids. The presence of guaiac-positive stools and iron-deficiency anemia are strongly suggestive of malignancy.

Higher white blood cell counts with the presence of immature forms or the presence of metabolic acidosis are worrisome for severe volume depletion or ischemic necrosis and sepsis. Presently, there are no laboratory tests that are especially useful for identifying the presence of simple or strangulated obstruction, although increases in serum D-lactate, creatine kinase BB isoenzymes, or intestinal fatty acid binding protein levels may be suggestive of the latter.

Recommendations for diagnostic imaging continue to evolve. In all cases, the key is not to operative intervention unnecessarily when the patient's signs or symptoms strongly suggest that high-grade or complete obstruction or bowel compromise is present. Abdominal radiography, which must include upright or cross-table lateral views, can be completed quickly and may indicate the need for emergency surgical intervention in patients who are not in the immediate postoperative period. A "staircasing" pattern of dilated air and fluid-filled small-bowel loops >2.5 cm in diameter with little or no air seen in the colon are classical findings in patients with small-bowel obstruction, although findings may be equivocal in some patients with documented disease. Little bowel gas appears in patients with proximal bowel obstruction or in patients whose intestinal lumens are filled with fluid. Upright plain films of the abdomen of patients with large-bowel obstruction typically show colon dilatation. Small-bowel air-fluid levels will not be obvious if the ileocecal valve is competent. Although it can be difficult to distinguish from ileus, small-bowel obstruction is more likely when air-fluid levels are seen without significant colonic distension. Free air suggests that perforation has occurred in patients who have not recently undergone surgical procedures. A gas-filled, "coffee bean"–shaped dilated shadow may be seen in patients with volvulus.

More sophisticated imaging, which may be unnecessarily time consuming and expensive, can nevertheless be beneficial when the diagnosis is unclear. Computed tomography (CT) is the most commonly used imaging modality. Its sensitivity for detecting bowel obstruction is ~95% (78–100%) in patients with high-grade obstruction, with a specificity of 96% and an accuracy of ≥95%. Its accuracy in diagnosing closed-loop obstruction is much lower (60%). CT may also provide useful information regarding location or to identify particular circumstances where surgical intervention is needed urgently. Patients who have evidence of contrast appearing within the cecum within 4–24 h of oral administration of water-soluble contrast can be expected to improve with high sensitivity and specificity (~95% each). For example, contrast studies may demonstrate a "bird's beak," a "c-loop," or "whorl" deformity on CT imaging at the site where twisting obstructs the lumen when a colonic volvulus is present. Although abdominal radiography is usually the initial examination, unlike CT imaging, it may not accurately distinguish obstruction from other causes of colonic dysmotility. Examples of some CT images are reproduced in **Fig. 323-2**.

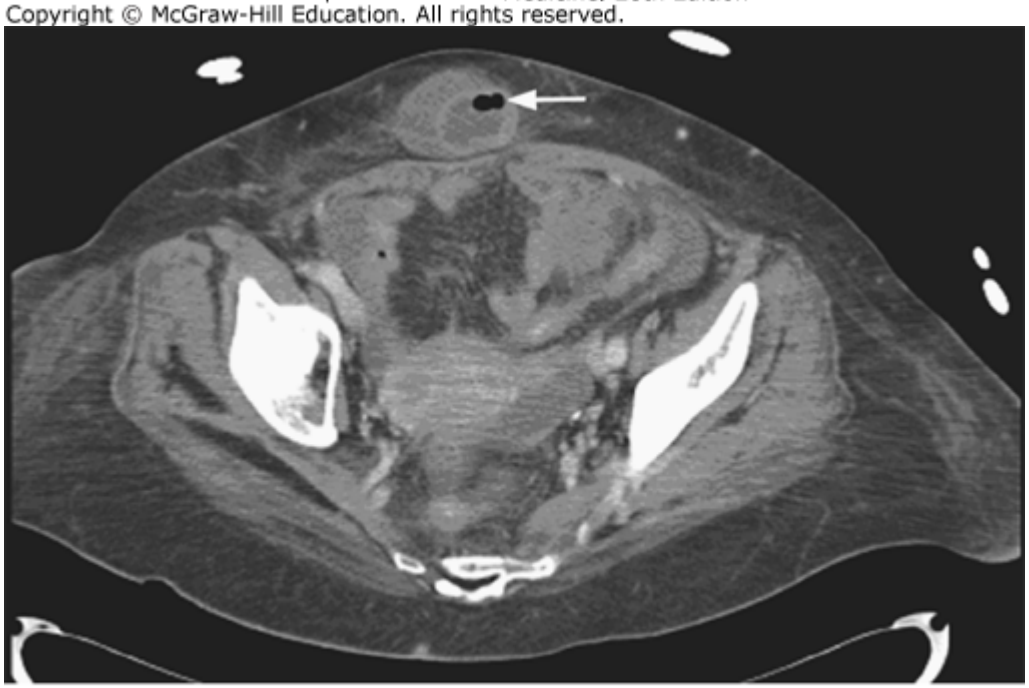
FIGURE 323-2

Computed tomography with oral and intravenous contrast demonstrating (***A***) evidence of small-bowel dilatation with air-fluid levels consistent with a small-bowel obstruction; (***B***) a partial small-bowel obstruction from an incarcerated ventral hernia (*arrow*); and (***C***) decompressed bowel seen distal to the hernia (*arrow*). *(From W Silen: Acute intestinal obstruction, in DL Longo et al [eds]: Harrison's Principles of Internal Medicine, 18th ed. New York, McGraw-Hill, 2012.)*



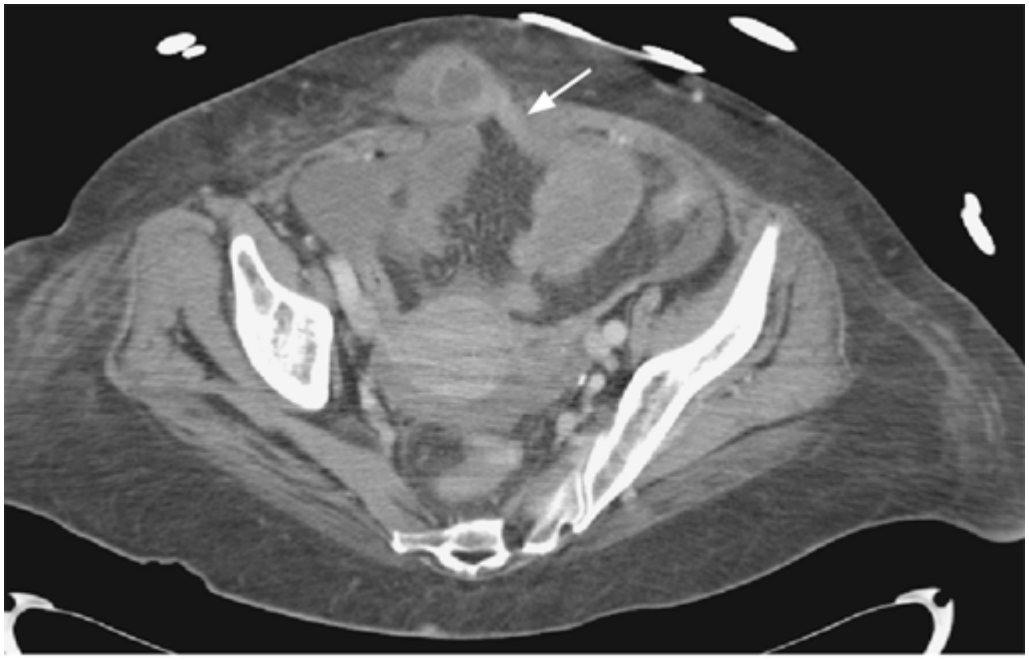
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ultrasonographic evaluations are especially difficult to interpret but may be sensitive and appropriate studies to evaluate patients who are pregnant or for whom x-ray exposure is otherwise contraindicated or inappropriate.

CT imaging with enteral and IV contrast can also identify ischemia. Altered bowel wall enhancement is the most specific early finding, but its sensitivity is low. Mesenteric venous gas, pneumoperitoneum, and pneumatosis intestinalis are late findings indicating the presence of bowel necrosis. CT scanning after a water-soluble contrast enema may help distinguish ileus or pseudo-obstruction from distal large-bowel obstruction in patients who present with evidence of small-bowel and colonic distention. CT enteroclysis, though rarely performed, can accurately identify neoplasia as a cause of bowel obstruction. Contrast enemas or colonoscopies are almost always needed to identify causes of acute colonic obstruction.

Barium studies are generally contraindicated in patients with firm evidence of complete or high-grade bowel obstruction, especially when they present acutely. Barium should never be given orally to a patient with possible obstruction until that diagnosis has been excluded. In every other case, such investigations should only be performed in exceptional circumstances and with great caution because patients with significant obstruction may develop barium concretions as an additional source of blockage and some who would have otherwise recovered will require operative intervention. Barium opacification also renders cross-sectional imaging studies or angiography uninterpretable.

TREATMENT

TREATMENT

Acute Intestinal Obstruction

An improved understanding of the pathophysiology of bowel obstruction and the importance of fluid resuscitation, electrolyte repletion, intestinal decompression, and the selected use of antibiotics have likely contributed to a reduction in mortality from acute bowel obstruction. Every patient should be stabilized as quickly as possible. Nasogastric tube suction decompresses the stomach, minimizes further distention from swallowed air, improves patient comfort, and reduces the risk of aspiration. Urine output should be assessed using a Foley catheter. In some cases, for example, in patients with cardiac disease, central venous pressures should be monitored. The use of antibiotics is controversial, although prophylactic administration may be warranted if operation is anticipated. Complete bowel obstruction is an indication for intervention. Stenting may be possible and warranted for some patients with high-grade obstruction due to unresectable stage IV malignancy. Stenting may also allow elective mechanical bowel preparation before surgery is undertaken. Because treatment options are so variable, it is helpful to make as precise a diagnosis as possible preoperatively.

ILEUS

Patients with ileus are treated supportively with intravenous fluids and nasogastric decompression while any underlying pathology is treated. Pharmacologic therapy is not yet proven to be efficacious or cost-effective. However, peripherally active μ-opioid receptor antagonists (e.g., alvimopan and methylnaltrexone) may accelerate gastrointestinal recovery in some patients who have undergone abdominal surgery.

COLONIC PSEUDO-OBSTRUCTION (OGILVIE'S DISEASE)

Neostigmine is an acetylcholinesterase inhibitor that increases cholinergic (parasympathetic) activity, which can stimulate colonic motility. Some studies have shown it to be moderately effective in alleviating acute colonic pseudo-obstruction. It is the most common therapeutic approach and can be used once it is certain that there is no mechanical obstruction. Cardiac monitoring is required, and atropine should be immediately available. Intravenous administration induces defecation and flatus within 10 min in the majority of patients who will respond. Sympathetic blockade by epidural anesthesia can successfully ameliorate pseudo-obstruction in some patients.

VOLVULUS

Patients with sigmoid volvulus can often be decompressed using a flexible tube inserted through a rigid proctoscope or using a flexible sigmoidoscope. Successful decompression results in sudden release of gas and fluid with evidence of decreased abdominal distension and allows definitive correction to be scheduled electively. Cecal volvulus most often requires laparotomy or laparoscopic correction.

INTRAOPERATIVE STRATEGIES

Approximately 60–80% of selected patients with mechanical bowel obstruction can be successfully treated conservatively. Indeed, most cases of radiation-induced obstruction should also be managed nonoperatively if possible. In most circumstances, early consultation with a surgeon is prudent when there is concern about strangulation obstruction or other abnormality that needs to be addressed urgently. Deterioration signifies a need for intervention. At this time, the decision as to whether the patient can continue to be treated nonoperatively can only be based on clinical judgment, although, as described earlier, imaging studies can

sometimes be helpful. The frequency of major complications after operation ranges from 12 to 47%, with greater risk being attributed to resection therapies and the patient's overall health. Risk is increased for patients with American Society of Anesthesiologists (ASA) class III or higher.

At operation, dilation proximal to the site of blockage with distal collapse is a defining feature of bowel obstruction. Intraoperative strategies depend on the underlying problem and range from lysis of adhesions to resection with or without diverting ostomy to primary resection with anastomosis. Resection is warranted when there is concern about the bowel's viability after the obstructive process is relieved. Laparoscopic approaches can be useful for patients with early obstruction when extensive adhesions are not expected to be present. Some patients with high-grade obstruction secondary to malignant disease that is not amendable to resection will benefit from bypass procedures.

ADULT INTUSSUSCEPTION AND GALLSTONE ILEUS

Primary resection is prudent. Careful manual reduction of any involved bowel may limit the amount of intestine that needs to be removed. A proximal ostomy may be required if unprepped colon is involved. The most common site of intestinal obstruction in patients with gallstone "ileus" is the ileum (60% of patients). The gallstone enters the intestinal tract most often via a cholecystoduodenal fistula. It can usually be removed by operative enterolithotomy. Addressing the gallbladder disease during urgent or emergent surgery is not recommended.

POSTOPERATIVE BOWEL OBSTRUCTION

Early postoperative mechanical bowel obstruction is that which occurs within the first 6 weeks of operation. Most are partial and can be expected to resolve spontaneously. It tends to respond and behave differently from classic mechanical bowel obstruction and may be very difficult to distinguish from postoperative ileus. A higher index of suspicion for a definitive site of obstruction is warranted for patients who undergo laparoscopic surgical procedures. Patients who first had ileus and then subsequently develop obstructive symptoms after an initial return of normal bowel function are more likely to have true postoperative small-bowel obstruction. The longer it takes for a patient's obstructive symptoms to resolve after hospitalization, the more likely the patient is to require surgical intervention.

ACKNOWLEDGMENT

The wisdom and expertise of Dr. William Silen are gratefully acknowledged.

FURTHER READING

Catena F et al: Adhesive small bowel adhesions obstruction: Evolutions in diagnosis, management and prevention. World J Gastrointest Surg 27:222, 2016.

Ferrada P et al: Surgery or stenting for colonic obstruction: A practice management guideline from the Eastern Association for the Surgery of Trauma. J Trauma Acute Care Surg 80:659, 2016. [PubMed: 26808035]

Jaffe T, Thompson WM: Large-bowel obstruction in the adults: Classic radiographic and CT findings, etiology and mimics. Radiology 275:651, 2015. [PubMed: 25997131]

Paulson EK, Thompson WM: Review of small-bowel obstruction: The diagnosis and when to worry. Radiology 275:332, 2015. [PubMed: 25906301]

Perry H et al: Relative accuracy of emergency CT in adults with non-traumatic abdominal pain. Brit Inst Rad 89:20150416, 2016.

Taylor MR, Lalani N: Adult small bowel obstruction. Acad Emerg Med 20:528, 2013. [PubMed: 23758299]

Chapter 324: Acute Appendicitis and Peritonitis

Danny O. Jacobs

ACUTE APPENDICITIS

INCIDENCE AND EPIDEMIOLOGY

Appendicitis occurs more frequently in Westernized societies but its incidence is decreasing for uncertain reasons. Nevertheless, acute appendicitis remains the most common emergency general surgical disease affecting the abdomen, with a rate of ~100 per 100,000 person-years in Europe and the Americas or about 11 cases per 10,000 people annually. Approximately 9% of men and 7% of women will experience an episode during their lifetime. Appendicitis occurs most commonly in 10- to 19-year-olds; however, the average age at diagnosis appears to be gradually increasing, as is the frequency of the disease in African Americans, Asians, and Native Americans. Overall, 70% of patients are <30 years old and most are men.

One of the more common complications and most important causes of excess morbidity and mortality is perforation, whether it is contained and localized or unconstrained within the peritoneal cavity. In contrast to the trend observed for appendicitis and appendectomy, the incidence of perforated appendicitis (~20 cases per 100,000 person-years) is increasing. The explanation for this trend is unknown. Approximately 20% of all patients will present with evidence of perforation, but the percentage risk is much higher in patients under 5 or over 65 years of age.

PATHOGENESIS OF APPENDICITIS AND APPENDICEAL PERFORATION

Appendicitis was first described in 1886 by Reginald Fitz. Its etiology is still not completely understood. Fecaliths, incompletely digested food residue, lymphoid hyperplasia, intraluminal scarring, tumors, bacteria, viruses, and inflammatory bowel disease have all been associated with inflammation of the appendix and appendicitis.

Although not proven, obstruction of the appendiceal lumen is believed to be an important step in the development of appendicitis—at least in some cases. Here, obstruction leads to bacterial overgrowth and luminal distension, with an increase in intraluminal pressure that can inhibit the flow of lymph and blood. Then, vascular thrombosis and ischemic necrosis with perforation of the distal appendix may occur. Therefore, perforation that occurs near the base of the appendix should raise concerns about another disease process. Most patients who will perforate do so before they are evaluated by surgeons.

Appendiceal fecaliths (or appendicoliths) are found in ~50% of patients with gangrenous appendicitis who perforate but are rarely identified in those who have simple disease. As mentioned earlier, the incidence of perforated, but not simple, appendicitis is increasing. The rate of perforated and nonperforated appendicitis is correlated in men but not in women. Together these observations suggest that the underlying pathophysiologic processes are different and that simple appendicitis does not always progress to perforation. It appears that some cases of simple acute appendicitis may resolve spontaneously or with antibiotic therapy with limited risk of recurrent disease. The use of antibiotics to treat uncomplicated appendicitis is currently being studied intensively. Preliminary data indicate that as many as 70% of patients who present with uncomplicated appendicitis based on computed tomography (CT) and who are treated with antibiotics alone will be free of recurrent disease for at least a year. These findings highlight the importance of clinical decision-making and risk assessment when deciding and discussing treatment options with patients who presumably have simple disease, for example, deciding who is an appropriate candidate for non-operative management and who is not. The latter is especially pertinent given the difficulty in assessing which patients might progress to perforation and which will not.

Increasingly it appears that there are two broad categories of patients with appendicitis—those with complicated disease like gangrene or perforation and those without. When perforation occurs, the resultant leak may be contained by the omentum or other surrounding tissues to form an abscess. Free perforation normally causes severe peritonitis. These patients may also develop infective suppurative thrombosis of the portal vein and its tributaries along with intrahepatic abscesses. The prognosis of the very unfortunate patients who develop this rare but dreaded complication is very poor.

CLINICAL MANIFESTATIONS

Improved diagnosis, supportive care, and surgical interventions are likely responsible for the remarkable decrease in the risk of mortality from simple appendicitis to currently <1%. Nevertheless, it is still important to identify patients who might have appendicitis as early as possible. Patients who have persistent symptoms that haven’t improved over 48 h may be more likely to perforate or develop other complications.

Appendicitis should be included in the differential diagnosis of abdominal pain for every patient in any age group unless it is certain that the organ has been previously removed (Table 324-1).

TABLE 324-1

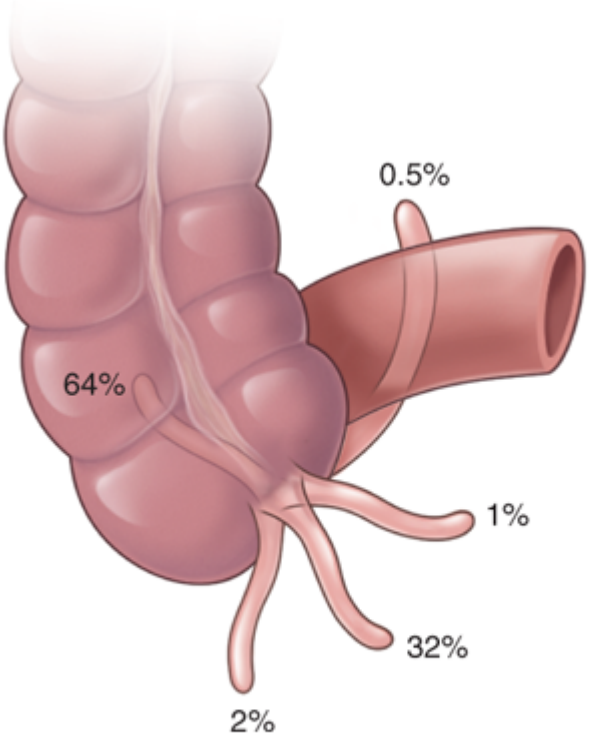
Some Conditions That Mimic Appendicitis

Crohn's disease	Meckel's diverticulitis
Cholecystitis or other gallbladder disease	Mittelschmerz
Diverticulitis	Mesenteric adenitis
Ectopic pregnancy	Omental torsion
Endometriosis	Pancreatitis
Gastroenteritis or colitis	Lower lobe pneumonia
Gastric or duodenal ulceration	Pelvic inflammatory disease
Hepatitis	Ruptured ovarian cyst or other cystic disease of the ovaries
Kidney disease, including nephrolithiasis	Small-bowel obstruction
Liver abscess	Urinary tract infection

The appendix's anatomical location, which varies, may directly influence how the patient presents. Where the appendix can be "found" ranges from local differences in how the appendiceal body and tip lie relative to its attachment to the cecum (**Figs. 324-1 and 324-2**), to where the appendix is actually situated in the peritoneal cavity—for example, from its typical location in the right lower quadrant, to the pelvis, right flank, right upper quadrant (as may be observed during pregnancy), or even the left side of the abdomen for patients with malrotation or who have severely redundant colons.

FIGURE 324-1

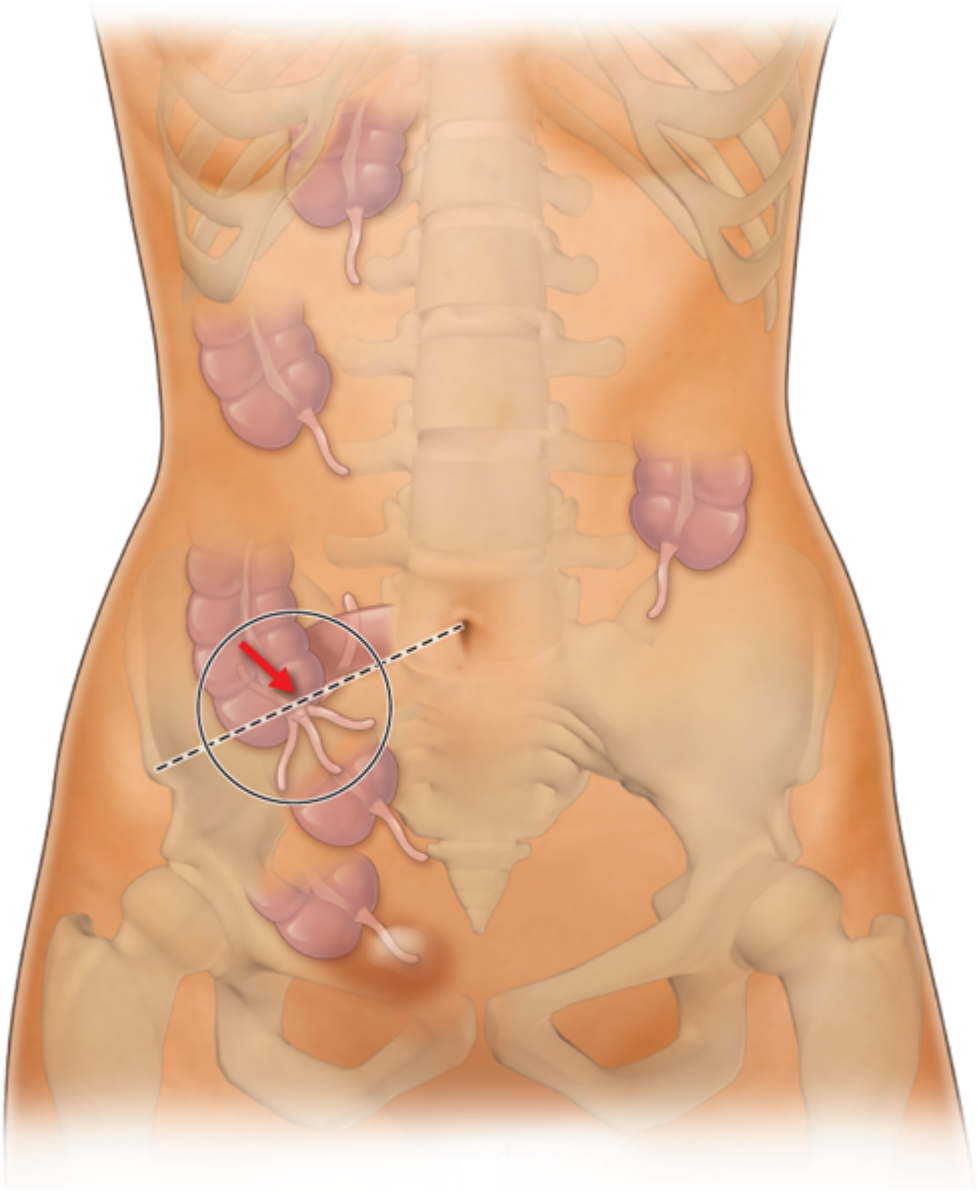
Regional anatomical variations of the appendix.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 324-2

Locations of the appendix and cecum.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Because the differential diagnosis of appendicitis is so extensive, deciding if a patient has appendicitis can be difficult (Table 324-2). Many patients may not present with the classically described history or physical findings and some may not have any abdominal discomfort early in the disease process. Soliciting an appropriate history requires detecting and evaluating symptoms that might suggest alternative diagnoses.

TABLE 324-2

Relative Frequency of Common Presenting Symptoms

Symptoms	Frequency
Abdominal pain	>95%
Anorexia	>70%
Constipation	4–16%
Diarrhea	4–16%
Fever	10–20%
Migration of pain to right lower quadrant	50–60%
Nausea	>65%
Vomiting	50–75%

What is the classic history? Nonspecific complaints occur first. Patients may notice changes in bowel habits or malaise and vague, perhaps intermittent, crampy, abdominal pain in the epigastric or periumbilical region. The pain subsequently migrates to the right lower quadrant over 12–24 h, where it is sharper and can be definitively localized as transmural inflammation when the appendix irritates the parietal peritoneum. Parietal peritoneal irritation may be associated with local muscle rigidity and stiffness. Patients with appendicitis will most often observe that their nausea, if present, followed the development of abdominal pain, which can help distinguish them from patients with gastroenteritis, for example, where nausea occurs first. Emesis, if present, also occurs after the onset of pain and is typically mild and scant. Thus, timing of the onset of symptoms and the characteristics of the patient's pain and any associated findings must be rigorously assessed. Anorexia is so common that the diagnosis of appendicitis should be questioned in its absence.

Arriving at the correct diagnosis is even more challenging when the appendix is not located in the right lower quadrant, in women of childbearing age, and in the very young or elderly. Because the differential diagnosis of appendicitis is so broad, often the key question to answer expeditiously is whether the patient has appendicitis or some other condition that requires immediate operative intervention. A major concern is that the likelihood of a delay in diagnosis is greater if the appendix is unusually positioned. All patients should undergo a rectal examination. An inflamed appendix located behind the cecum or below the pelvic brim may prompt very little tenderness of the anterior abdominal wall.

Patients with pelvic appendicitis are more likely to present with dysuria, urinary frequency, diarrhea, or tenesmus. They may only experience pain in the suprapubic region on palpation or on rectal or pelvic examination. A pelvic examination in women is mandatory to rule out conditions affecting urogynecologic organs that can cause abdominal pain and mimic appendicitis such as pelvic inflammatory disease, ectopic pregnancy, and ovarian torsion. Interest in the ability of various clinical scoring systems to predict appendicitis or the need for imaging studies continues. However, none of the currently available decision tools yet appear to be able to circumvene or obviate the need for expert clinical opinion. The relative frequencies of some presenting signs are displayed in Table 324-3.

TABLE 324-3

Relative Frequency of Some Presenting Signs

Signs	Frequency (%)
Abdominal tenderness	>95%
Right lower quadrant tenderness	>90%
Rebound tenderness	30–70%
Rectal tenderness	30–40%
Cervical motion tenderness	30%
Rigidity	~10%
Psoas sign	3–5%
Obturator sign	5–10%
Rovsing's sign	5%
Palpable mass	<5%

Patients with simple appendicitis normally only appear mildly ill with a pulse and temperature that are usually only slightly above normal. The provider should be concerned about other disease processes beside appendicitis or the presence of complications such as perforation, phlegmon, or abscess formation if the temperature is >38.3°C (~101°F) and if there are rigors.

Patients with appendicitis will be found to lie quite still to avoid peritoneal irritation caused by movement, and some will report discomfort caused by a bumpy car ride on the way to the hospital or clinic, coughing,

sneezing, or other actions that replicate a Valsalva maneuver. The entire abdomen should be examined systematically starting in an area where the patient does not report discomfort if possible. Classically, maximal tenderness is identified in the right lower quadrant at or near McBurney's point, which is located approximately one-third of the way along a line originating at the anterior iliac spine and running to the umbilicus. Gentle pressure in the left lower quadrant may elicit pain in the right lower quadrant if the appendix is located there. This is Rovsing's sign **(Table 324-4)**. Evidence of parietal peritoneal irritation is often best elicited by gentle abdominal percussion, jiggling the patient's gurney or bed, or mildly bumping the feet.

TABLE 324-4

Classic Signs of Appendicitis in Patients with Abdominal Pain

Maneuver	Findings
Rovsing's sign	Palpating in the left lower quadrant causes pain in the right lower quadrant
Obturator sign	Internal rotation of the hip causes pain, suggesting the possibility of an inflamed appendix located in the pelvis
Iliopsoas sign	Extending the right hip causes pain along posterolateral back and hip, suggesting retrocecal appendicitis

Atypical presentation and pain patterns are common, especially in the very old or the very young. Diagnosing appendicitis in children can be especially challenging because they tend to respond so dramatically to stimulation and obtaining an accurate history may be difficult. In addition, it is important to remember that the smaller omentum found in children may be less likely to wall off an appendiceal perforation. Observing the child in a quiet surrounding may be helpful.

Signs and symptoms of appendicitis can be subtle in the elderly who may not react as vigorously to appendicitis as younger people. Pain, if noticed, may be minimal and have originated in the right lower quadrant or, otherwise, where the appendix is located. It may never have been noticed to be intermittent, or there may only be significant discomfort with deep palpation. Nausea, anorexia, and emesis may be the predominant complaints. The rare patient may even present with signs and symptoms of distal bowel obstruction secondary to appendiceal inflammation and phlegmon or abscess formation.

LABORATORY TESTING

Laboratory testing does not identify patients with appendicitis. The white blood cell count is only mildly to moderately elevated in ~70% of patients with simple appendicitis (with a leukocytosis of 10,000–18,000 cells/

μL). A “left shift” toward immature polymorphonuclear leukocytes is present in >95% of cases. A sickle cell preparation may be prudent to obtain in those of African, Spanish, Mediterranean, or Indian ancestry. Serum amylase and lipase levels should be measured.

Urinalysis is indicated to help exclude genitourinary conditions that may mimic acute appendicitis, but a few red or white blood cells may be present as a nonspecific finding. However, an inflamed appendix that abuts the ureter or bladder may cause sterile pyuria or hematuria. Every woman of childbearing age should have a pregnancy test. Cervical cultures are indicated if pelvic inflammatory disease is suspected. Anemia and guaiac-positive stools should raise concern about the presence of other diseases or complications such as cancer.

IMAGING

Plain films of the abdomen are rarely helpful and so are not routinely obtained unless the clinician is worried about other conditions such as intestinal obstruction, perforated viscus, or ureterolithiasis. Less than 5% of patients will present with an opaque fecalith in the right lower quadrant. The presence of a fecalith is not diagnostic of appendicitis, although its presence in an appropriate location where the patient complains of pain is suggestive and is associated with a greater likelihood of complications.

The effectiveness of ultrasonography as a tool to diagnosis appendicitis is highly operator dependent. Even in very skilled hands, the appendix may not be visualized. Its overall sensitivity is 0.86, with a specificity of 0.81. Ultrasonography, especially intravaginal techniques, appears to be most useful for identifying pelvic pathology in women. Ultrasonographic findings suggesting the presence of appendicitis include wall thickening, an increased appendiceal diameter, and the presence of free fluid. Current practice in many institutions is to first perform ultrasonography and progress to other imaging studies only if the findings are equivocal.

The sensitivity and specificity of CT are at least 0.94 and 0.95, respectively. Thus, CT imaging, given its high negative predictive value, may be helpful if the diagnosis is in doubt, although studies performed early in the course of disease may not have any typical radiographic findings. In patients where the diagnosis is uncertain, delaying operation at the time of presentation to obtain CT does not appear to increase the risk of perforation. CT scanning is a superior method for assessing the severity of acute appendicitis in the absence of peritoneal findings indicative of perforation, abscess, or suspicion of an associated malignancy.

Suggestive findings on CT examination include dilatation >6 mm with wall thickening, a lumen that does not fill with enteric contrast, and fatty tissue stranding or air surrounding the appendix, which suggests inflammation **(Figs. 324-3 and 324-4)**. The presence of luminal air or contrast is not consistent with a diagnosis of appendicitis. Furthermore, nonvisualization of the appendix is a nonspecific finding that should not be used to rule out the presence of appendiceal or periappendiceal inflammation.

FIGURE 324-3

Computed tomography with oral and intravenous contrast of acute appendicitis. There is thickening of the wall of the appendix and periappendiceal stranding (*arrow*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 324-4

Appendiceal fecalith (*arrow*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SPECIAL PATIENT POPULATIONS

Appendicitis in the most common extrauterine general surgical emergency observed during pregnancy. Early symptoms of appendicitis such as nausea and anorexia may be overlooked. Diagnosing appendicitis in pregnant patients may be especially difficult because as the uterus enlarges the appendix may be pushed higher along the right flank even to the right upper quadrant or because the gravid uterus may obscure typical physical findings. Ultrasonography may facilitate early diagnosis. A high index of suspicion is required because of the effects of unrecognized and untreated appendicitis on the fetus. For example, the fetal mortality rate is four times greater (from 5 to 20%) in patients with perforation.

Immunocompromised patients may present with only mild tenderness and may have many other disease processes in their differential diagnosis, including atypical infections from mycobacteria, *Cytomegalovirus*, or other fungi. Enterocolitis is a concern and may be present in patients who present with abdominal pain, fever, and neutropenia due to chemotherapy. CT imaging may be very helpful, although it is important not to be overly cautious and delay operative intervention for those patients who are believed to have appendicitis.

TREATMENT

TREATMENT

Acute Appendicitis

In the absence of contraindications, most patients who have strongly suggestive medical histories and physical examinations with supportive laboratory findings are candidates for appendectomy. In many instances, imaging studies are not required but are often obtained before surgical consultation is requested. Certainly, imaging and further study is appropriate in patients whose evaluations are suggestive but not convincing.

CT may accurately indicate the presence of appendicitis or other intraabdominal processes that warrant intervention. Whenever the diagnosis is uncertain, it is prudent to observe the patient and repeat the abdominal examination over 6–8 h. Any evidence of progression is an indication for operation. Narcotics can be given to patients with severe discomfort.

All patients should be fully prepared for surgery and have any fluid and electrolyte abnormalities corrected. Either laparoscopic or open appendectomy is a satisfactory choice for patients with uncomplicated appendicitis though most procedures are performed in a minimally invasive fashion. Management of those who present with a mass representing a phlegmon or abscess can be more difficult. Such patients are best served by treatment with broad-spectrum antibiotics, drainage if there is an abscess >3 cm in diameter, and parenteral fluids and bowel rest if they appear to respond to conservative management. The appendix can then be more safely removed 6–12 weeks later when inflammation has diminished.

Laparoscopic appendectomy now accounts for the majority of all appendectomies performed in Western cultures and is associated with less postoperative pain, shorter lengths of stay, faster return to normal activity and likely fewer superficial wound complications—although the risk of intraabdominal abscess formation may be higher.

A laparoscopic approach may also be useful when the exact diagnosis is uncertain. A laparoscopic approach may also facilitate exposure in those who are very obese. Absent complications, most patients can be discharged within 24–40 h of operation. The most common postoperative complications are fever and leukocytosis. Continuation of these findings beyond 5 days should raise concern for the presence of an intraabdominal abscess. The mortality rate for uncomplicated, nonperforated appendicitis is 0.1–0.5%, which approximates the overall risk of general anesthesia. The mortality rate for perforated appendicitis or other complicated disease is much higher, ranging from 3% overall to a high as 15% in the elderly.

ACUTE PERITONITIS

Acute peritonitis, or inflammation of the visceral and parietal peritoneum, is most often but not always infectious in origin, resulting from perforation of a hollow viscus. This is called *secondary peritonitis*, as opposed to *primary* or *spontaneous peritonitis*, when a specific intraabdominal source cannot be identified. In either instance, the inflammation can be localized or diffuse.

ETIOLOGY

Infective organisms may contaminate the peritoneal cavity after spillage from a hollow viscus, because of a penetrating wound of the abdominal wall, or because of the introduction of a foreign object like a peritoneal dialysis catheter or port that becomes infected. Secondary peritonitis most commonly results from perforation of the appendix, colonic diverticuli, or the stomach and duodenum. It may also occur as a complication of bowel infarction or incarceration, cancer, inflammatory bowel disease, and intestinal obstruction or volvulus. Conditions that may cause secondary bacterial peritonitis and their mechanisms are listed in Table 324-5. Over 90% of the cases of primary or spontaneous bacterial peritonitis occur in patients with ascites or hypoproteinemia (<1 g/L).

TABLE 324-5

Conditions Leading to Secondary Bacterial Peritonitis

Bowel perforation	**Perforation or leakage of other organs**
Appendicitis trauma (blunt or penetrating)	Biliary leakage (e.g., after liver biopsy)
Anastomotic leakage	Cholecystitis
Adhesion	Intraperitoneal bleeding
Diverticulitis	Pancreatitis
Iatrogenic (including endoscopic perforation)	Salpingitis
Ingested foreign body	Traumatic or other rupture of urinary bladder
Inflammation	**Loss of peritoneal integrity**
Intussusception	Intraperitoneal chemotherapy
Neoplasms	Iatrogenic (e.g., postoperative foreign body)
Obstruction	Perinephric abscess
Peptic ulcer disease	Peritoneal dialysis or other indwelling devices
Strangulated hernia	Trauma
Vascular (including ischemia or embolus)	

Aseptic peritonitis is most commonly caused by the abnormal presence of physiologic fluids like gastric juice, bile, pancreatic enzymes, blood, or urine. It can also be caused by the effects of normally sterile foreign bodies like surgical sponges or instruments. More rarely, it occurs as a complication of systemic diseases like lupus erythematosus, porphyria, and familial Mediterranean fever. The chemical irritation caused by stomach acid and activated pancreatic enzymes is extreme and secondary bacterial infection may occur.

CLINICAL FEATURES

The cardinal signs and symptoms of peritonitis are acute, typically severe, abdominal pain with tenderness and fever. How the patient's complaints of pain are manifested depends on their overall physical health and whether the inflammation is diffuse or localized. Elderly and immunosuppressed patients may not respond as aggressively to the irritation. Diffuse, generalized peritonitis is most often recognized as diffuse abdominal tenderness with local guarding, rigidity, and other evidence of parietal peritoneal irritation. Physical findings

may only be identified in a specific region of the abdomen if the intraperitoneal inflammatory process is limited or otherwise contained as may occur in patients with uncomplicated appendicitis or diverticulitis. Bowel sounds are usually absent to hypoactive.

Most patients present with tachycardia and signs of volume depletion with hypotension. Laboratory testing typically reveals a significant leukocytosis, and patients may be severely acidotic. Radiographic studies may show dilatation of the bowel and associated bowel wall edema. Free air, or other evidence of leakage, requires attention and could represent a surgical emergency. In stable patients in whom ascites is present, diagnostic paracentesis is indicated, where the fluid is tested for protein and lactate dehydrogenase and the cell count is measured.

THERAPY AND PROGNOSIS

Whereas mortality rates can be <10% for reasonably healthy patients with relatively uncomplicated, localized peritonitis, mortality rates >40% have been reported for the elderly or immunocompromised. Successful treatment depends on correcting any electrolyte abnormalities, restoration of fluid volume and stabilization of the cardiovascular system, appropriate antibiotic therapy, and surgical correction of any underlying abnormalities.

ACKNOWLEDGMENT

The wisdom and expertise of Dr. William Silen is gratefully acknowledged in this updated chapter on acute appendicitis and peritonitis.

FURTHER READING

Andersson RE: Short-term complications and long-term morbidity of laparoscopic and open appendicectomy in a national cohort. Br J Surg 101:1135, 2014. [PubMed: 24979720]

Buckius MT et al: Changing epidemiology of acute appendicitis in the United States: Study period 1993–2008. J Surg Res 175:185, 2012. [PubMed: 22099604]

Drake FT et al: Time to appendectomy and risk of perforation in acute appendicitis. JAMA Surg 149:837, 2014. [PubMed: 24990687]

Flum DR: Acute appendicitis—appendectomy of the “antibiotics first” strategy. N Engl J Med 372:1937, 2015. [PubMed: 25970051]

Ohle R et al: The Alvarado score for predicting acute appendicitis: A systematic review. BMC Med 9:139, 2011. [PubMed: 22204638]

Salminen P et al: Antibiotic therapy versus appendectomy for treatment of uncomplicated acute appendicitis: The APPAC Randomized Clinical Trial. JAMA 313:2340, 2015. [PubMed: 26080338]

Vons C et al: Amoxicillin plus clavulanic acid versus appendicectomy for treatment of acute uncomplicated appendicitis: An open-label, non-inferiority, randomised controlled trial. Lancet 377:1573, 2011. [PubMed: 21550483]

Chapter 325: Nutrient Requirements and Dietary Assessment

Johanna Dwyer

INTRODUCTION

Nutrients are substances that are not synthesized in sufficient amounts in the body and therefore must be supplied by the diet. Nutrient requirements for groups of healthy persons have been determined experimentally. The absence of essential nutrients leads to growth impairment, organ dysfunction, and failure to maintain nitrogen balance or adequate status of protein and other nutrients. For good health, we require energy-providing nutrients (protein, fat, and carbohydrate), vitamins, minerals, and water. Requirements for organic nutrients include 9 essential amino acids, several fatty acids, glucose, 4 fat-soluble vitamins, 10 water-soluble vitamins, dietary fiber, and choline. Several inorganic substances, including 4 minerals, 7 trace minerals, 3 electrolytes, and the ultratrace elements, must also be supplied by diet.

The amounts of essential nutrients required by individuals differ by their age and physiologic state. Conditionally essential nutrients are not required in the diet but must be supplied to certain individuals who do not synthesize them in adequate amounts, such as those with genetic defects, those with pathologies such as infection, disease or trauma with nutritional implications, and developmentally immature infants. For example, inositol, taurine, arginine, and glutamine may be needed by premature infants. Many other organic and inorganic compounds that are present in foods, such as pesticides, lead, phytochemicals, zoochemicals, and microbial products may also have health effects.

ESSENTIAL NUTRIENT REQUIREMENTS

Energy

For weight to remain stable, energy intake must match energy output. The major components of energy output are resting energy expenditure (REE) and physical activity; minor components include the energy cost of metabolizing food (thermic effect of food, or specific dynamic action) and shivering thermogenesis (e.g., cold-induced thermogenesis). The average energy intake is ~2600 kcal/d for American men and ~1800 kcal/d for American women, although these estimates vary with body size and activity level. Formulas for roughly estimating REE are useful in assessing the energy needs of an individual whose weight is stable. Thus, for males, REE = 900 + 10*m*, and for females, REE = 700 + 7*m*, where *m* is mass in kilograms. The calculated REE is then adjusted for physical activity level by multiplying by 1.2 for sedentary, 1.4 for moderately active, or 1.8 for very active individuals. The final figure, the estimated energy requirement (EER), provides an approximation of total caloric needs in a state of energy balance for a person of a certain age, sex, weight, height, and physical activity level. **For further discussion of energy balance in health and disease, see Chap. 327.**

Protein

Dietary protein consists of both essential and nonessential amino acids that are required for protein synthesis. The nine essential amino acids are histidine, isoleucine, leucine, lysine, methionine/cystine, phenylalanine/tyrosine, threonine, tryptophan, and valine. Certain amino acids, such as alanine, can also be used for energy and gluconeogenesis. When energy intake is inadequate, protein intake must be increased, because ingested amino acids are diverted into pathways of glucose synthesis and oxidation. In extreme energy deprivation, protein-calorie malnutrition may ensue **(Chap. 327)**.

For adults, the recommended dietary allowance (RDA) for protein is ~0.8 g/kg desirable body mass per day, assuming that energy needs are met and that the protein is of relatively high biologic value. Current recommendations for a healthy diet call for at least 10–14% of calories from protein. Most American diets provide at least those amounts. Biological value tends to be highest for animal proteins, followed by proteins from legumes (beans), cereals (rice, wheat, corn), and roots. Combinations of plant proteins that complement one another in their essential amino acid profiles or combinations of animal and plant proteins can increase biological value and lower total protein intakes necessary to meet requirements. In healthy people with adequate diets, the timing of protein intake over the course of the day has little effect.

Protein needs increase during growth, pregnancy, lactation, and rehabilitation after injury or malnutrition. Tolerance to dietary protein is decreased in renal insufficiency (with consequent uremia) and in liver failure. Usual protein intakes can precipitate encephalopathy in patients with cirrhosis of the liver.

Fat and Carbohydrate

Fats are a concentrated source of energy and constitute, on average, 34% of calories in U.S. diets. However, for optimal health, fat intake should total no more than 30% of calories. Saturated fat and trans fat should be limited to <10% of calories and polyunsaturated fats to <10% of calories, with monounsaturated fats accounting for the remainder of fat intake. At least 45–55% of total calories should be derived from carbohydrates. The brain requires ~100 g of glucose per day for fuel; other tissues use about 50 g/d. Some tissues (e.g., brain and red blood cells) rely on glucose supplied either exogenously or from muscle proteolysis. Over time, during

hypocaloric states, adaptations in carbohydrate needs are possible. Like fat (9 kcal/g), carbohydrate (4 kcal/g), and protein (4 kcal/g), alcohol (ethanol) provides energy (7 kcal/g). However, it is not a nutrient.

Water

For adults, 1–1.5 mL of water per kilocalorie of energy expenditure is sufficient under usual conditions to allow for normal variations in physical activity, sweating, and solute load of the diet. Water losses include 50–100 mL/d in the feces; 500–1000 mL/d by evaporation or exhalation; and, depending on the renal solute load, ≥1000 mL/d in the urine. If external losses increase, intakes must increase accordingly to avoid underhydration. Fever increases water losses by ~200 mL/d per °C; diarrheal losses vary but may be as great as 5 L/d in severe diarrhea. Heavy sweating, vigorous exercise, and vomiting also increase water losses. When renal function is normal and solute intakes are adequate, the kidneys can adjust to increased water intake by excreting up to 18 L of excess water per day **(Chap. 374)**. However, obligatory urine outputs can compromise hydration status when there is inadequate water intake or when losses increase in disease or kidney damage.

Infants have high requirements for water because of their large surface area to volume ratios, their inability to communicate their thirst, and the limited capacity of the immature kidney to handle high renal solute loads. Increased water needs during pregnancy are ~30 mL/d. During lactation, milk production increases daily water requirements so that ~1000 mL of additional water is needed, or 1 mL for each milliliter of milk produced. Special attention must also be paid to the water needs of the elderly, who have reduced total body water and blunted thirst sensation and are more likely to be taking medications such as diuretics.

Other Nutrients

See Chap. 326 for detailed descriptions of vitamins and minerals.

DIETARY REFERENCE INTAKES AND RDAS

Fortunately, human life and well-being can be maintained within a fairly wide range with most nutrient intakes. However, the capacity for adaptation is not infinite—too much, as well as too little, intake of a nutrient can have adverse effects or alter the health benefits conferred by another nutrient. Therefore, benchmark recommendations regarding nutrient intakes have been developed to guide clinical practice. These quantitative estimates of nutrient intakes are collectively referred to as the *dietary reference intakes* (DRIs). The DRIs have supplanted the RDAs—the single reference values used in the United States until the early 1990s. DRIs include an *estimated average requirement* (EAR) for nutrients as well as other reference values used for dietary planning: the RDA, the *adequate intake* (AI), and the tolerable *upper level* (UL). The DRIs also include acceptable macronutrient distribution ranges (AMDRs) for protein, fat, and carbohydrate. The current DRIs for vitamins and elements are provided in **Tables 325-1** and **325-2**, respectively. **Table 325-3** provides DRIs for water and macronutrients. **EERs are discussed in Chap. 327 on energy balance in health and disease.**

TABLE 325-1

Dietary Reference Intakes (DRIs): Recommended Dietary Allowances and Adequate Intakes for Vitamins

Life-Stage Group	Vitamin A (µg/d)[a]	Vitamin C (mg/d)	Vitamin D (µg/d)[b,c]	Vitamin E (mg/d)[d]	Vitamin K (µg/d)	Thiamin (mg/d)	Riboflavin (mg/d)	Niacin (mg/d)[e]	Vitamin B_6 (mg/d)	Folate (µg/d)[f]	Vitamin B_{12} (µg/d)	Pantothenic Acid (mg/d)	Biotin (µg/d)	Choline (mg/d)[g]
Infants														
Birth to 6 mo	400*	40*	10	4*	2.0*	0.2*	0.3*	2*	0.1*	65*	0.4*	1.7*	5*	125*
6–12 mo	500*	50*	10	5*	2.5*	0.3*	0.4*	4*	0.3*	80*	0.5*	1.8*	6*	150*
Children														
1–3 y	**300**	**15**	**15**	**6**	30*	**0.5**	**0.5**	**6**	**0.5**	**150**	**0.9**	2*	8*	200*
4–8 y	**400**	**25**	**15**	**7**	55*	**0.6**	**0.6**	**8**	**0.6**	**200**	**1.2**	3*	12*	250*
Males														
9–13 y	**600**	**45**	**15**	**11**	60*	**0.9**	**0.9**	**12**	**1.0**	**300**	**1.8**	4*	20*	375*
14–18 y	**900**	**75**	**15**	**15**	75*	**1.2**	**1.3**	**16**	**1.3**	**400**	**2.4**	5*	25*	550*
19–30 y	**900**	**90**	**15**	**15**	120*	**1.2**	**1.3**	**16**	**1.3**	**400**	**2.4**	5*	30*	550*
31–50 y	**900**	**90**	**15**	**15**	120*	**1.2**	**1.3**	**16**	**1.3**	**400**	**2.4**	5*	30*	550*
51–70 y	**900**	**90**	**15**	**15**	120*	**1.2**	**1.3**	**16**	**1.7**	**400**	**2.4**[h]	5*	30*	550*
>70 y	**900**	**90**	**20**	**15**	120*	**1.2**	**1.3**	**16**	**1.7**	**400**	**2.4**[h]	5*	30*	550*
Females														
9–13 y	**600**	**45**	**15**	**11**	60*	**0.9**	**0.9**	**12**	**1.0**	**300**	**1.8**	4*	20*	375*
14–18 y	**700**	**65**	**15**	**15**	75*	**1.0**	**1.0**	**14**	**1.2**	**400**[i]	**2.4**	5*	25*	400*
19–30 y	**700**	**75**	**15**	**15**	90*	**1.1**	**1.1**	**14**	**1.3**	**400**[i]	**2.4**	5*	30*	425*
31–50 y	**700**	**75**	**15**	**15**	90*	**1.1**	**1.1**	**14**	**1.3**	**400**[i]	**2.4**	5*	30*	425*
51–70 y	**700**	**75**	**15**	**15**	90*	**1.1**	**1.1**	**14**	**1.5**	**400**	**2.4**[h]	5*	30*	425*
>70 y	**700**	**75**	**20**	**15**	90*	**1.1**	**1.1**	**14**	**1.5**	**400**	**2.4**[h]	5*	30*	425*
Pregnant Women														

Life-Stage Group	Vitamin A (µg/d)[a]	Vitamin C (mg/d)	Vitamin D (µg/d)[b,c]	Vitamin E (mg/d)[d]	Vitamin K (µg/d)	Thiamin (mg/d)	Riboflavin (mg/d)	Niacin (mg/d)[e]	Vitamin B_6 (mg/d)	Folate (µg/d)[f]	Vitamin B_{12} (µg/d)	Pantothenic Acid (mg/d)	Biotin (µg/d)	Choline (mg/d)[g]
14–18 y	**750**	**80**	**15**	**15**	75*	**1.4**	**1.4**	**18**	**1.9**	**600**[i]	**2.6**	6*	30*	450*
19–30 y	**770**	**85**	**15**	**15**	90*	**1.4**	**1.4**	**18**	**1.9**	**600**[i]	**2.6**	6*	30*	450*
31–50 y	**770**	**85**	**15**	**15**	90*	**1.4**	**1.4**	**18**	**1.9**	**600**[i]	**2.6**	6*	30*	450*
Lactating Women														
14–18 y	**1200**	**115**	**15**	**19**	75*	**1.4**	**1.6**	**17**	**2.0**	**500**	**2.8**	7*	35*	550*
19–30 y	**1300**	**120**	**15**	**19**	90*	**1.4**	**1.6**	**17**	**2.0**	**500**	**2.8**	7*	35*	550*
31–50 y	**1300**	**120**	**15**	**19**	90*	**1.4**	**1.6**	**17**	**2.0**	**500**	**2.8**	7*	35*	550*

Note: This table (taken from the DRI reports; see *www.nap.edu*) presents recommended dietary allowances (RDAs) in **bold type** and adequate intakes (AIs) in ordinary type followed by an asterisk (*). An RDA is the average daily dietary intake level sufficient to meet the nutrient requirements of nearly all healthy individuals (97–98%) in a group. The RDA is calculated from an estimated average requirement (EAR). If sufficient scientific evidence is not available to establish an EAR and thus to calculate an RDA, an AI is usually developed. For healthy breast-fed infants, an AI is the mean intake. The AI for other life-stage and sex-specific groups is believed to cover the needs of all healthy individuals in those groups, but lack of data or uncertainty in the data makes it impossible to specify with confidence the percentage of individuals covered by this intake.

[a]As retinol activity equivalents (RAEs). 1 RAE = 1 µg retinol, 12 µg β-carotene, 24 µg α-carotene, or 24 µg β-cryptoxanthin. The RAE for dietary provitamin A carotenoids is twofold greater than the retinol equivalent (RE), whereas the RAE for preformed vitamin A is the same as the RE. [b]As cholecalciferol. 1 µg cholecalciferol = 40 IU vitamin D. [c]Under the assumption of minimal sunlight. [d]As α-tocopherol. α-Tocopherol includes *RRR*-α-tocopherol, the only form of α-tocopherol that occurs naturally in foods, and the *2R*-stereoisomeric forms of α-tocopherol (*RRR-, RSR-, RRS-*, and *RSS*-α-tocopherol) that occur in fortified foods and supplements. It does not include the *2S*-stereoisomeric forms of α-tocopherol (*SRR-, SSR-, SRS-*, and *SSS*-α-tocopherol) also found in fortified foods and supplements. [e]As niacin equivalents (NEs). 1 mg of niacin = 60 mg of tryptophan; 0–6 months = preformed niacin (not NE). [f]As dietary folate equivalents (DFEs). 1 DFE = 1 µg food folate = 0.6 µg of folic acid from fortified food or as a supplement consumed with food = 0.5 µg of a supplement taken on an empty stomach. [g]Although AIs have been set for choline, there are few data to assess whether a dietary supply of choline is needed at all stages of the life cycle, and it may be that the choline requirement can be met by endogenous synthesis at some of these stages. [h]Because 10–30% of older people may malabsorb food-bound B12, it is advisable for those >50 years of age to meet their RDA mainly by consuming foods fortified with B_{12} or a supplement containing B_{12}. [i]In view of evidence linking inadequate folate intake with neural tube defects in the fetus, it is recommended that all women capable of becoming pregnant consume 400 µg of folate from supplements or fortified foods in addition to intake of food folate from a varied diet. [j] It is assumed that women will continue consuming 400 µg from supplements or fortified food until their pregnancy is confirmed and they enter prenatal care, which ordinarily occurs after the end of the periconceptional period—the critical time for formation of the neural tube.

Source: Food and Nutrition Board, Institute of Medicine, National Academies (*http://www.iom.edu/Activities/Nutrition/SummaryDRIs/DRI-Tables.aspx*).

TABLE 325-2

Dietary Reference Intakes (DRIs): Recommended Dietary Allowances and Adequate Intakes for Elements

Life-Stage Group	Calcium (mg/d)	Chromium (µg/d)	Copper (µg/d)	Fluoride (mg/d)	Iodine (µg/d)	Iron (mg/d)	Magnesium (mg/d)	Manganese (mg/d)	Molybdenum (µg/d)	Phosphorus (mg/d)	Selenium (µg/d)	Zinc (mg/d)	Potassium (g/d)
Infants													
Birth to 6 mo	200*	0.2*	200*	0.01*	110*	0.27*	30*	0.003*	2*	100*	15*	2*	0.4*
6–12 mo	260*	5.5*	220*	0.5*	130*	**11**	75*	0.6*	3*	275*	20*	**3**	0.7*
Children													
1–3 y	**700**	11*	**340**	0.7*	**90**	**7**	**80**	1.2*	**17**	**460**	**20**	**3**	3.0*
4–8 y	**1000**	15*	**440**	1*	**90**	**10**	**130**	1.5*	**22**	**500**	**30**	**5**	3.8*
Males													
9–13 y	**1300**	25*	**700**	2*	**120**	**8**	**240**	1.9*	**34**	**1250**	**40**	**8**	4.5*
14–18 y	**1300**	35*	**890**	3*	**150**	**11**	**410**	2.2*	**43**	**1250**	**55**	**11**	4.7*
19–30 y	**1000**	35*	**900**	4*	**150**	**8**	**400**	2.3*	**45**	**700**	**55**	**11**	4.7*
31–50 y	**1000**	35*	**900**	4*	**150**	**8**	**420**	2.3*	**45**	**700**	**55**	**11**	4.7*
51–70 y	**1000**	30*	**900**	4*	**150**	**8**	**420**	2.3*	**45**	**700**	**55**	**11**	4.7*
>70 y	**1200**	30*	**900**	4*	**150**	**8**	**420**	2.3*	**45**	**700**	**55**	**11**	4.7*
Females													
9–13 y	**1300**	21*	**700**	2*	**120**	**8**	**240**	1.6*	**34**	**1250**	**40**	**8**	4.5*
14–18 y	**1300**	24*	**890**	3*	**150**	**15**	**360**	1.6*	**43**	**1250**	**55**	**9**	4.7*
19–30 y	**1000**	25*	**900**	3*	**150**	**18**	**310**	1.8*	**45**	**700**	**55**	**8**	4.7*
31–50 y	**1000**	25*	**900**	3*	**150**	**18**	320	1.8*	**45**	**700**	**55**	**8**	4.7*
51–70 y	**1200**	20*	**900**	3*	**150**	**8**	**320**	1.8*	**45**	**700**	**55**	**8**	4.7*
>70 y	**1200**	20*	**900**	3*	**150**	**8**	**320**	1.8*	**45**	**700**	**55**	**8**	4.7*
Pregnant Women													

Life-Stage Group	Calcium (mg/d)	Chromium (µg/d)	Copper (µg/d)	Fluoride (mg/d)	Iodine (µg/d)	Iron (mg/d)	Magnesium (mg/d)	Manganese (mg/d)	Molybdenum (µg/d)	Phosphorus (mg/d)	Selenium (µg/d)	Zinc (mg/d)	Potassium (g/d)
14–18 y	**1300**	29*	**1000**	3*	**220**	**27**	**400**	2.0*	**50**	**1250**	**60**	**12**	4.7*
19–30 y	**1000**	30*	**1000**	3*	**220**	**27**	**350**	2.0*	**50**	**700**	**60**	**11**	4.7*
31–50 y	**1000**	30*	**1000**	3*	**220**	**27**	**360**	2.0*	**50**	**700**	**60**	**11**	4.7*
Lactating Women													
14–18 y	**1300**	44*	**1300**	3*	**290**	**10**	**360**	2.6*	**50**	**1250**	**70**	**13**	5.1*
19–30 y	**1000**	45*	**1300**	3*	**290**	**9**	**310**	2.6*	**50**	**700**	**70**	**12**	5.1*
31–50 y	**1000**	45*	**1300**	3*	**290**	**9**	**320**	2.6*	**50**	**700**	**70**	**12**	5.1*

Note: This table (taken from the DRI reports; see *www.nap.edu*) presents recommended dietary allowances (RDAs) in **bold type** and adequate intakes (AIs) in ordinary type followed by an asterisk (*). An RDA is the average daily dietary intake level sufficient to meet the nutrient requirements of nearly all healthy individuals (97–98%) in a group. The RDA is calculated from an estimated average requirement (EAR). If sufficient scientific evidence is not available to establish an EAR and thus to calculate an RDA, an AI is usually developed. For healthy breast-fed infants, an AI is the mean intake. The AI for other life-stage and sex-specific groups is believed to cover the needs of all healthy individuals in those groups, but lack of data or uncertainty in the data makes it impossible to specify with confidence the percentage of individuals covered by this intake.

Sources: Food and Nutrition Board, Institute of Medicine, National Academies (*http://www.iom.edu/Activities/Nutrition/SummaryDRIs/DRI-Tables.aspx*), based on: *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride* (1997); *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline* (1998); *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids* (2000); and *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc* (2001); *Dietary Reference Intakes for Water, Potassium, Sodium, Chloride, and Sulfate* (2005); and *Dietary Reference Intakes for Calcium and Vitamin D* (2011). These reports can be accessed via *www.nap.edu*.

TABLE 325-3

Dietary Reference Intakes (DRIs): Recommended Dietary Allowances and Adequate Intakes for Total Water and Macronutrients

Life-Stage Group	Total Water[a] (L/d)	Carbohydrate (g/d)	Total Fiber (g/d)	Fat (g/d)	Linoleic Acid (g/d)	α-Linolenic Acid (g/d)	Protein[b] (g/d)
Infants							
Birth to 6 mo	0.7*	60*	ND[c]	31*	4.4*	0.5*	9.1*
6–12 mo	0.8*	95*	ND	30*	4.6*	0.5*	**11.0**
Children							
1–3 y	1.3*	**130**	19*	ND	7*	0.7*	**13**
4–8 y	1.7*	**130**	25*	ND	10*	0.9*	**19**
Males							
9–13 y	2.4*	**130**	31*	ND	12*	1.2*	**34**
14–18 y	3.3*	**130**	38*	ND	16*	1.6*	**52**
19–30 y	3.7*	**130**	38*	ND	17*	1.6*	**56**
31–50 y	3.7*	**130**	38*	ND	17*	1.6*	**56**
51–70 y	3.7*	**130**	30*	ND	14*	1.6*	**56**
>70 y	3.7*	**130**	30*	ND	14*	1.6*	**56**
Females							
9–13 y	2.1*	**130**	26*	ND	10*	1.0*	**34**
14–18 y	2.3*	**130**	26*	ND	11*	1.1*	**46**
19–30 y	2.7*	**130**	25*	ND	12*	1.1*	**46**
31–50 y	2.7*	**130**	25*	ND	12*	1.1*	**46**
51–70 y	2.7*	**130**	21*	ND	11*	1.1*	**46**
>70 y	2.7*	**130**	21*	ND	11*	1.1*	**46**
Pregnant Women							
14–18 y	3.0*	**175**	28*	ND	13*	1.4*	**71**
19–30 y	3.0*	**175**	28*	ND	13*	1.4*	**71**
31–50 y	3.0*	**175**	28*	ND	13*	1.4*	**71**
Lactating Women							
14–18	3.8*	**210**	29*	ND	13*	1.3*	**71**
19–30 y	3.8*	**210**	29*	ND	13*	1.3*	**71**

Life-Stage Group	Total Water[a] (L/d)	Carbohydrate (g/d)	Total Fiber (g/d)	Fat (g/d)	Linoleic Acid (g/d)	α-Linolenic Acid (g/d)	Protein[b] (g/d)
31–50 y	3.8*	**210**	29*	ND	13*	1.3*	**71**

Note: This table (taken from the DRI reports; see *www.nap.edu*) presents recommended dietary allowances (RDAs) in **bold type** and adequate intakes (AIs) in ordinary type followed by an asterisk (*). An RDA is the average daily dietary intake level sufficient to meet the nutrient requirements of nearly all healthy individuals (97–98%) in a group. The RDA is calculated from an estimated average requirement (EAR). If sufficient scientific evidence is not available to establish an EAR and thus to calculate an RDA, an AI is usually developed. For healthy breast-fed infants, an AI is the mean intake. The AI for other life-stage and sex-specific groups is believed to cover the needs of all healthy individuals in those groups, but lack of data or uncertainty in the data make it impossible to specify with confidence the percentage of individuals covered by this intake.

[a]Total water includes all water contained in food, beverages, and drinking water. [b]Based on grams of protein per kilogram of body weight for the reference body weight (e.g., for adults: 0.8 g/kg body weight for the reference body weight). [c]Not determined.

Source: Food and Nutrition Board, Institute of Medicine, National Academies (*http://www.iom.edu/Activities/Nutrition/SummaryDRIs/DRI-Tables.aspx*), based on: *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids* (2002/2005) and *Dietary Reference Intakes for Water, Potassium, Sodium, Chloride, and Sulfate* (2005). These reports can be accessed via *www.nap.edu*.

Estimated Average Requirement

When florid manifestations of the classic dietary-deficiency diseases such as rickets (deficiency of vitamin D and calcium), scurvy (deficiency of vitamin C), xerophthalmia (deficiency of vitamin A), and protein-calorie malnutrition were common, nutrient adequacy was inferred from the absence of their clinical deficiency signs. Later, biochemical and other changes were used that became evident long before the deficiency was clinically apparent. Consequently, criteria of adequacy are now based on biological markers when they are available. Priority is given to sensitive biochemical, physiologic, or behavioral tests that reflect early changes in regulatory processes; maintenance of body stores of nutrients; or, if available, the amount of a nutrient that minimizes the risk of chronic degenerative disease. Current efforts focus on this last variable, but relevant markers often are not available, and long time lags between intake and disease outcomes further complicate the picture.

The types of evidence and criteria used to establish nutrient requirements vary by nutrient, age, and physiologic group. The EAR is the amount of a nutrient estimated to be adequate for half of the healthy individuals of a specific age and sex. The EAR is not an effective estimate of nutrient adequacy in individuals because it is a median requirement for a group; 50% of individuals in a group fall below the requirement and 50% fall above it. Thus, a person with a usual intake at the EAR has a 50% risk of inadequate intake. For these reasons the other standards described below are more useful for clinical purposes.

Recommended Dietary Allowances

The RDA, the nutrient intake goal for planning diets of individuals, is the average daily dietary intake level that meets the nutrient requirements of nearly all healthy persons of a specific sex, age, life stage, or physiologic condition (e.g., pregnancy or lactation). It is defined statistically as two standard deviations above the EAR to ensure that the needs of any given individual are met. The online tool at *http://fnic.nal.usda.gov/interactiveDRI/* allows health professionals to calculate individualized daily nutrient recommendations for dietary planning based on the DRIs. The RDAs are used to formulate food guides such as the U.S. Department of Agriculture (USDA) MyPlate Food Guide for individuals (*www.supertracker.usda.gov/default.aspx*), to create food-exchange lists for therapeutic diet planning, and as a standard for describing the nutritional content of foods and nutrient-containing dietary supplements on labels.

The risk of dietary inadequacy increases as one's intake falls below the RDA. However, the RDA is an overly generous criterion for evaluating nutrient adequacy. For example, by definition, the RDA exceeds the actual requirements of all but ~2–3% of the population. Therefore, many people whose intake fall below the RDA are still getting enough of the nutrient. On food labels, the nutrient content in a food is stated by weight or as a percent of the daily value (DV), a variant of the RDA used on the nutrition facts panel that, for an adult, represents the highest RDA for an adult consuming 2000 kcal.

Adequate Intake

It is not possible to set an RDA for some nutrients that lack an established EAR. In this circumstance, the AI is based on observed or experimentally determined approximations of nutrient intakes in healthy people. In the DRIs, AIs rather than RDAs are proposed for nutrients consumed by infants (up to age 1 year) as well as for chromium, fluoride, manganese, sodium, potassium, pantothenic acid, biotin, choline, and water consumed by persons of all ages. Vitamin D and calcium recommendations were recently revised, and more precise estimates are now available.

Tolerable Upper Levels of Nutrient Intake

Healthy individuals gain no established benefit from consuming nutrient levels above the RDA or AI. In fact, excessive nutrient intake can disturb body functions and cause acute, progressive, or permanent disabilities. The tolerable UL is the highest level of chronic nutrient intake (usually daily) that is unlikely to pose a risk of adverse health effects for most of the population. Data on the adverse effects of large amounts of many nutrients are unavailable or too limited to establish a UL. Therefore, the lack of a UL does *not* mean that the risk of adverse effects from high intake is nonexistent. Nutrient levels in commonly eaten foods rarely exceed the UL. However, very highly fortified foods and dietary supplements provide more concentrated amounts of nutrients per serving and thus pose a potential risk of toxicity. Nutrient dietary supplements are labeled with Supplement Facts that express the amount of nutrient in absolute units or as the percentage of the DV

provided per recommended serving size. Total nutrient consumption, including that in foods, supplements, and over-the-counter medications (e.g., antacids), should not exceed RDA levels.

Acceptable Macronutrient Distribution Ranges

The AMDRs are not experimentally determined; rather they are rough ranges for energy-providing macronutrient intakes (protein, carbohydrate, and fat) that the National Academy of Medicine's (formerly Institute of Medicine, IOM) Food and Nutrition Board considers to be healthful. These ranges are 10–35% of calories for protein, 20–35% of calories for fat, and 45–65% of calories for carbohydrate. Alcohol, which also provides energy, is not a nutrient; therefore, no recommendations are not provided.

FACTORS ALTERING NUTRIENT NEEDS

The DRIs are affected by age, sex, growth rate, pregnancy, lactation, physical activity level, concomitant diseases, drugs, and dietary composition. If requirements for nutrient sufficiency are close to intake levels indicating excess of a nutrient, dietary planning is difficult.

Physiologic Factors

Growth, strenuous physical activity, pregnancy, and lactation all increase needs for energy and several essential nutrients. Energy needs rise during pregnancy due to fetal growth demands and increased energy required for milk production during lactation. Energy needs decrease with loss of lean body mass, the major determinant of REE. The energy needs of older persons, especially those aged >70 years, tend to be lower than those of younger persons because lean tissue, physical activity, and health often decline with age.

Dietary Composition

Dietary composition affects the biological availability and use of nutrients. For example, iron absorption may be impaired by large amounts of calcium or lead; likewise, non-heme iron uptake may be impaired by a lack of ascorbic acid and amino acids in the meal. Bodily protein may be decreased when essential amino acids are not present in sufficient amounts—a rare scenario in U.S. diets. Animal foods, such as milk, eggs, and meat, have high biologic values, with most of the needed amino acids present in adequate amounts. Plant proteins in corn (maize), soy, rice, and wheat have lower biological values and must be combined with other plant or animal proteins or fortified with the amino acids that are deficient to achieve optimal use by the body.

Route of Intake

The RDAs apply only to oral intakes. When nutrients are administered parenterally, similar values can sometimes be used for amino acids, glucose (carbohydrate), fats, sodium, chloride, potassium, and most vitamins because their intestinal absorption rate is nearly 100%. However, the oral bioavailability of most mineral elements may be only half that obtained by parenteral administration. For some nutrients that are not readily stored in the body or that cannot be stored in large amounts, timing of administration may also be important. For example, amino acids cannot be used for protein synthesis if they are not supplied together; instead, they will be used for energy production, although in healthy individuals eating adequate diets, the distribution of protein intake over the course of the day has little effect on health.

Disease

Dietary deficiency diseases include protein-calorie malnutrition, iron-deficiency anemia, goiter (due to iodine deficiency), rickets and osteomalacia (vitamin D deficiency), and xeropthalmia (vitamin A deficiency), megaloblastic anemia (vitamin B_{12} or folic acid deficiency), scurvy (vitamin C/ascorbic acid deficiency), beriberi (thiamin deficiency), and pellagra (niacin and tryptophan deficiency) **(Chaps. 326 and 327)**. Each deficiency disease is characterized by imbalances at the cellular level between the supply of nutrients or energy and the body's nutritional needs for growth, maintenance, and other functions. Imbalances and excesses in nutrient intakes are recognized as risk factors for certain chronic degenerative diseases, such as saturated fat and cholesterol in coronary artery disease; sodium in hypertension; obesity in hormone-dependent cancers (endometrial and breast); and ethanol in alcoholism. Diet is only one of many risk factors because the etiology and pathogenesis of these disorders are multifactorial. Osteoporosis, for example, is associated with calcium deficiency, sometimes secondary to vitamin D deficiency, as well as with environment related risk factors (e.g., smoking, sedentary lifestyle), physiology (e.g., estrogen deficiency), genetic determinants (e.g., defects in collagen metabolism), and drug use (chronic steroid and aromatase inhibitors) **(Chap. 404)**.

DIETARY ASSESSMENT

Nutrition assessment in clinical situations is an iterative process that involves: (1) screening for malnutrition, (2) assessing the diet and other data to establish either the absence or the presence of malnutrition and its possible causes, (3) planning and implementing the most appropriate nutritional therapy, and (4) reassessing intakes to make sure that they have been consumed. Some disease states affect the bioavailability, requirements, use, or excretion of specific nutrients. In these circumstances, specific measurements of various nutrients or their biomarkers may be required to ensure adequate replacement **(Chap. 326)**.

Most health care facilities have nutrition-screening processes in place for identifying possible malnutrition after hospital admission. Nutritional screening is required by the Joint Commission, which accredits and certifies health care organizations in the United States. However, no universally recognized or validated standards exist. The factors that are usually assessed include abnormal weight for height or body mass index (e.g., BMI <19 or >25); reported weight change (involuntary loss or gain of >5 kg in the past 6 months) **(Chap. 43)**; diagnoses with known nutritional implications (e.g., metabolic disease, any disease affecting the gastrointestinal tract, alcoholism); present therapeutic dietary prescription; chronic poor appetite; presence of chewing and swallowing problems or major food intolerances; need for

assistance with preparing or shopping for food, eating, or other aspects of self-care; and social isolation. The nutritional status of hospitalized patients should be reassessed periodically—at least once every week.

A more complete dietary assessment is indicated for patients who exhibit a high risk of or frank malnutrition on nutritional screening. The type of assessment varies with the clinical setting, the severity of the patient's illness, and the stability of the patient's condition.

Acute-Care Settings

In acute-care settings, anorexia, various other diseases, test procedures, and medications can compromise dietary intake. Under such circumstances, the goal is to identify and avoid inadequate intake and to assure appropriate alimentation. Dietary assessment focuses on what patients are currently eating, whether or not they are able and willing to eat, and whether or not they experience any problems with eating. Dietary intake assessment is based on information from observed intakes; medical records; history; clinical examination; and anthropometric, biochemical, and functional status evaluations. The objective is to gather enough information to establish the likelihood of malnutrition due to poor dietary intake or other causes in order to assess whether nutritional therapy is indicated **(Chap. 328)**.

Simple observations may suffice to suggest inadequate oral intake. These include dietitians' and nurses' notes; observation of a patient's frequent refusal to eat or the amount of food eaten on trays; the frequent performance of tests and procedures that are likely to cause meals to be skipped; adherence to nutritionally inadequate diet orders (e.g., clear liquids or full liquids) for more than a few days; the occurrence of fever, gastrointestinal distress, vomiting, diarrhea, or a comatose state; and the presence of diseases or use of treatments that involve any part of the alimentary tract. Acutely ill patients with diet-related diseases such as diabetes need assessment because an inappropriate diet may exacerbate these conditions and adversely affect other therapies. Abnormal biochemical values (serum albumin levels <35 g/L [<3.5 mg/dL]; serum cholesterol levels <3.9 mmol/L [<150 mg/dL]) are nonspecific but may indicate a need for further nutritional assessment.

Most therapeutic diets offered in hospitals are calculated to meet individual nutrient requirements and the RDA *if they are eaten*. Exceptions include clear liquids, some full-liquid diets, and test diets (such as those adhered to in preparation for gastrointestinal procedures), which are inadequate for several nutrients and should not be used, if possible, for more than 24 h. However, because as much as half of the food served to hospitalized patients is not eaten, it cannot be assumed that the intakes of hospitalized patients are adequate. Dietary assessment should compare how much and what kinds of food the patient has consumed with the diet that has been provided. Major deviations in intakes of energy, protein, fluids, or other nutrients of special concern for the patient's illness should be noted and corrected, especially for long-staying patients.

Nutritional monitoring is especially important for patients who are very ill and who have extended lengths of hospital stay. Patients who are fed by enteral and parenteral routes also require special nutritional assessment and monitoring by physicians and/or dietitians with certification in nutritional support **(Chap. 328)**.

Ambulatory Settings

The aim of dietary assessment in the outpatient setting is to determine whether or not the patient's usual diet is a health risk in itself or if it contributes to existing chronic disease-related problems. Dietary assessment also provides the basis for planning a diet that fulfills therapeutic goals while ensuring patient adherence. The outpatient's dietary assessment should review the adequacy of present and usual food intakes, including vitamin and mineral supplements, oral nutritional supplements, medical foods, other dietary supplements, medications, and alcohol, because all of these may affect the patient's nutritional status. The assessment should focus on the dietary constituents that are most likely to be involved or compromised by a specific diagnosis as well as on any comorbidities that are present. More than one day's intake should be reviewed to provide a better representation of the usual diet, upon which personalized dietary recommendations can be based.

There are many ways to assess the adequacy of a patient's habitual diet. These include use of a food guide, a food-exchange list, a diet history, or a food-frequency questionnaire. A commonly used food guide for healthy persons is the USDA's Choose My Plate, which is useful as a rough guide for avoiding inadequate intakes of essential nutrients as well as likely excesses in the amounts of fat (especially saturated and trans fats), sodium, sugar, and alcohol consumed **(Table 325-4)**. The Choose My Plate graphic emphasizes a balance between calories and nutritional needs, encouraging increased intake of fruits and vegetables, whole grains, and low-fat milk in conjunction with reduced intake of sodium and high-calorie sugary drinks. The Web version of the guide provides a calculator that tailors the number of servings suggested for healthy patients of different weights, sexes, ages, and life-cycle stages to help them to meet their needs while avoiding excess (*http://www.supertracker.usda.gov/default.aspx* and *www.ChooseMyPlate.gov*). Patients who follow ethnic or unusual dietary patterns may need extra instruction on how foods should be categorized and on the appropriate portion sizes that constitute a serving. The process of reviewing the guide with patients helps them transition to healthier dietary patterns and identifies food groups eaten in excess of recommendations or in insufficient quantities. For persons on therapeutic diets, assessment against food-exchange lists may be useful. These include, for example, American Diabetes Association food-exchange lists for diabetes and the Academy of Nutrition and Dietetics food-exchange lists for renal disease.

TABLE 325-4

Choose My Plate: A Guide to Individualized Dietary Planning

	Examples of Standard Portion Sizes at Indicated Energy Level		
Dietary Factor, Unit of Measure (Advice)	**Lower: 1600 kcal**	**Moderate: 2200 kcal**	**Higher: 2800 kcal**
Fruits, cups (Focus on fruits.)	1.5	2	2.5
Vegetables, cups (Vary vegetables.)	2	3	3.5
Grains, oz eq (Make at least half of grains whole.)[a]	5	7	10
Protein foods, oz eq (Go lean with protein.)[b]	5	6	7
Dairy, cups or oz[c] (Choose calcium-rich foods.)	3	3	3
"Empty" calories, kcal[d]	120	260	400
Sodium, mg	<2300 at all energy levels		
Physical activity, min	At least 150 min vigorous physical activity per week at all energy levels		

Note: Oils (formerly listed with portions of 5, 6, and 8 teaspoons for the lower, moderate, and higher energy levels, respectively) are no longer singled out in Choose My Plate, but rather are included in the empty calories/added sugar category with SOFAS (calories from solid fats and added sugars). The limit is the remaining number of calories in each food pattern above after intake of the recommended amounts of the nutrient-dense foods.

[a]For example, 1 serving equals 1 slice bread, 1 cup ready-to-eat cereal, or 0.5 cup cooked rice, pasta, or cooked cereal. [b]For example, 1 serving equals 1 oz lean meat, poultry, or fish; 1 egg; 1 tablespoon peanut butter; 0.25 cup cooked dry beans; or 0.5 oz nuts or seeds. [c]For example, 1 serving equals 1 cup milk or yogurt, 1.5 oz natural cheese, or 2 oz processed cheese. [d]Formerly called "discretionary calorie allowance." Portions are calculated as the number of calories remaining after all of the above allotments are accounted for.

Abbreviation: oz eq, ounce equivalent.

Source: Data from U.S. Department of Agriculture (*http://www.Choosemyplate.gov*).

NUTRITIONAL STATUS ASSESSMENT

Full nutritional status assessment is reserved for seriously ill patients and those at very high nutritional risk when the cause of malnutrition is still uncertain after the initial clinical evaluation and dietary assessment. It involves multiple dimensions, including documentation of dietary intake, anthropometric measurements, biochemical measurements of blood and urine, clinical examination, health history elicitation, and functional status evaluation. Therapeutic dietary prescriptions and menu plans for most diseases are available from most hospitals and from the Academy of Nutrition and Dietetics. **For further discussion of nutritional assessment, see Chap. 327.**

GLOBAL CONSIDERATIONS

The DRIs (e.g., the EAR, the UL, and energy needs) are estimates of physiologic requirements based on experimental evidence. Assuming that appropriate adjustments are made for age, sex, body size, and physical activity level, these estimates should be applicable to individuals in most parts of the world. However, other values are not transportable. The AIs are based on customary and adequate intakes in U.S. and Canadian populations, which appear to be compatible with good health, rather than on a large body of direct experimental evidence. Similarly, the AMDRs represent expert opinion regarding the approximate intakes of energy-providing nutrients that are healthful in these North American populations. Thus these measures should be used with caution in other settings. Nutrient-based standards like the DRIs have also been developed by the World Health Organization/Food and Agricultural Organization of the United Nations and are available on the Web (*http://www.who.int/nutrition/topics/nutrecomm/en/index.html*). The European Food Safety Authority (EFSA) Panel on Dietetic Products, Nutrition and Allergies periodically publishes its recommendations in the EFSA's on-line *journal*. Other countries have promulgated similar recommendations. The different standards have many similarities in their basic concepts, definitions, and nutrient recommendation levels, but there are some differences from the DRIs as a result of the functional criteria chosen, environmental differences, the timeliness of the evidence reviewed, and expert judgment.

FURTHER READING

Brannon PM et al: Scanning for new evidene to prioritize updates to the Dietary Reference Intakes: Case studies for thiamin and phosphorus. Am J Clin Nur 104:1366, 2016.

Forster H et al: Personalised nutrition: The role of new dietary assessment methods. Proc Nutr Soc 75:96, 2016. [PubMed: 26032731]

Gibson RS: *Principles of Nutritional Assessment*, 2nd ed. Oxford, Oxford University Press, 2005.

Lupton JR et al: Nutrient reference value: Non-communicable disease endpoints—A conference report. Eur J Nutr 55 SUPPL 1:S1–10, 2016. [PubMed: 26983608]

Ocke MC: Evaluation of methodologies for assessing the overall diet: Dietary quality scales and dietary pattern analysis. Proc Nutr Soc 72:191, 2013. [PubMed: 23360896]

Otten JJ et al: *Dietary Reference Intakes: The Essential Guide to Nutrient Requirements*. Washington, DC, National Academies Press, 2006.

Potischman N, Freudenheim JL: Biomarkers of nutritional exposure and nutritional status: An overview. J Nutr 133:873S, 2003. [PubMed: 12612172]

Report of the Subcommittee on Interpretation and Uses of Dietary Reference Intakes and Upper Reference Levels of Nutrients, and the Steering Committtee on the Scientific Evaluation of Dietary Reference Intakes, Food and Nutrition Board: Dietary Reference Intakes: Applications in Dietary Assessment. Washington, DC, National Academies Press, 2008.

Willett WC: *Nutritonal Epidemiology*, 3rd ed. Oxford, Oxford University Press, 2012.

Yetley EA et al: Options for basing Dietary Reference Intakes (DRIs) on chronic disease endpoints report from a joint US-/Canadian-sponsored working group. Am J Clin Nutr 105:249S, 2017. [PubMed: 27927637]

Chapter 326: Vitamin and Trace Mineral Deficiency and Excess

Paolo M. Suter; Robert M. Russell

FIGURE 326-1

INTRODUCTION

Vitamins are required constituents of the human diet since they are synthesized inadequately or not at all in the human body. Only small amounts of these substances are needed to carry out essential biochemical reactions (e.g., by acting as coenzymes or prosthetic groups). Overt vitamin or trace mineral deficiencies are rare in Western countries because of a plentiful, varied, and inexpensive food supply; food fortification; and use of supplements. However, multiple nutrient deficiencies may appear together in persons who are chronically ill or alcoholic. After gastric bypass surgery, patients are at high risk for multiple nutrient deficiencies. Moreover, subclinical vitamin and trace mineral deficiencies, as diagnosed by laboratory testing, are quite common in the normal population, especially in the geriatric age group. Conversely, because of the widespread use of nutrient supplements, nutrient toxicities are gaining pathophysiologic and clinical importance.

Victims of famine, emergency-affected and displaced populations, and refugees are at increased risk for protein-energy malnutrition and classic micronutrient deficiencies (vitamin A, iron, iodine) as well as for overt deficiencies in thiamine (beriberi), riboflavin, vitamin C (scurvy), and niacin (pellagra).

Body stores of vitamins and minerals vary tremendously. For example, stores of vitamins B_{12} and A are large, and an adult may not become deficient until ≥1 year after beginning to eat a deficient diet. However, folate and thiamine may become depleted within weeks among those eating a deficient diet. Therapeutic modalities can deplete essential nutrients from the body; for example, hemodialysis or diuretics remove water-soluble vitamins, which must be replaced by supplementation.

Vitamins and trace minerals play several roles in diseases: (1) Deficiencies of vitamins and minerals may be caused by disease states such as malabsorption; (2) either deficiency or excess of vitamins and minerals can cause disease in and of itself (e.g., vitamin A intoxication and liver disease); and (3) vitamins and minerals in high doses may be used as drugs (e.g., niacin for hypercholesterolemia). Since they are covered elsewhere, the hematologic-related vitamins and minerals **(Chaps. 93 and 95)** either are not considered or are

considered only briefly in this chapter, as are the bone-related vitamins and minerals (vitamin D, calcium, phosphorus, magnesium; **Chap. 402**).

VITAMINS

See also Table 326-1 and Fig. 326-1.

TABLE 326-1

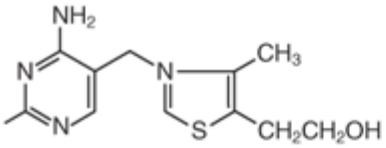
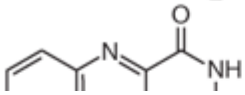
Principal Clinical Findings of Vitamin Malnutrition

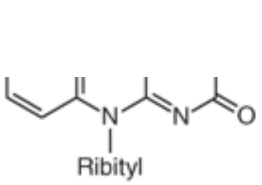
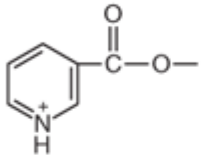
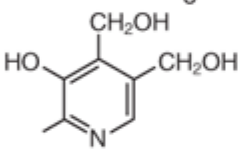
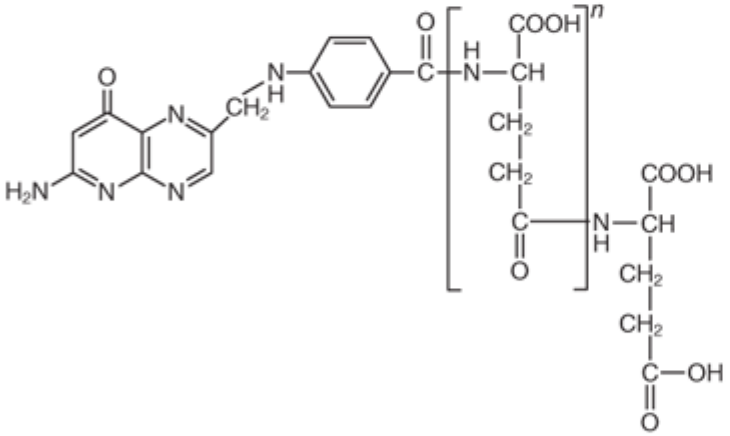
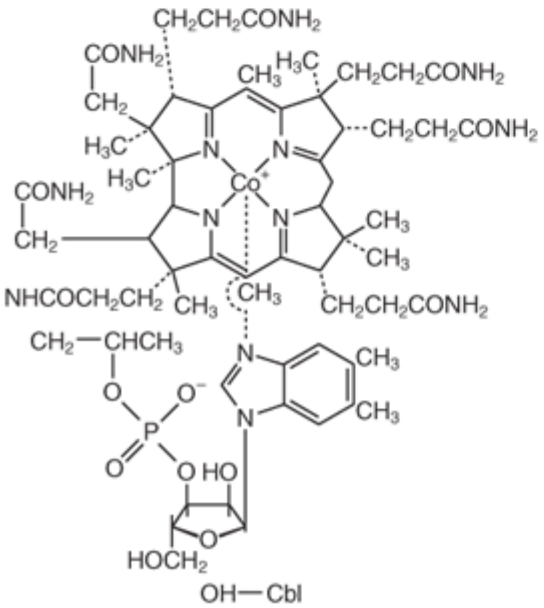
Nutrient	Clinical Finding	Dietary Level per Day Associated with Overt Deficiency in Adults	Contributing Factors to Deficiency
Thiamine	Beriberi: neuropathy, muscle weakness and wasting, cardiomegaly, edema, ophthalmoplegia, confabulation	<0.3 mg/1000 kcal	Alcoholism, chronic diuretic use, hyperemesis, thiaminases in food
Riboflavin	Magenta tongue, angular stomatitis, seborrhea, cheilosis, ocular symptoms, corneal vascularization	<0.4 mg	Alcoholism, individuals with poor diets and low intake of milk products
Niacin	Pellagra: pigmented rash of sun-exposed areas, bright red tongue, diarrhea, apathy, memory loss, disorientation	<9.0 niacin equivalents	Alcoholism, vitamin B_6 deficiency, riboflavin deficiency, tryptophan deficiency
Vitamin B_6	Seborrhea, glossitis, convulsions, neuropathy, depression, confusion, microcytic anemia	<0.2 mg	Alcoholism, isoniazid
Folate	Megaloblastic anemia, atrophic glossitis, depression, ↑ homocysteine	<100 µg/d	Alcoholism, sulfasalazine, pyrimethamine, triamterene
Vitamin B_{12}	Megaloblastic anemia, loss of vibratory and position sense, abnormal gait, dementia, impotence, loss of bladder and bowel control, ↑ homocysteine, ↑ methylmalonic acid	<1.0 µg/d	Gastric atrophy (pernicious anemia), terminal ileal disease, strict vegetarianism, acid-reducing drugs (e.g., H_2 blockers), metformin
Vitamin C	Scurvy: petechiae, ecchymosis, coiled hairs, inflamed and bleeding gums, joint effusion, poor wound healing, fatigue	<10 mg/d	Smoking, alcoholism

Nutrient	Clinical Finding	Dietary Level per Day Associated with Overt Deficiency in Adults	Contributing Factors to Deficiency
Vitamin A	Xerophthalmia, night blindness, Bitot's spots, follicular hyperkeratosis, impaired embryonic development, immune dysfunction	<300 µg/d	Fat malabsorption, infection, measles, alcoholism, protein-energy malnutrition
Vitamin D	Rickets: skeletal deformation, rachitic rosary, bowed legs; osteomalacia	<2.0 µg/d	Aging, lack of sunlight exposure, fat malabsorption, deeply pigmented skin
Vitamin E	Peripheral neuropathy, spinocerebellar ataxia, skeletal muscle atrophy, retinopathy	Not described unless underlying contributing factor is present	Occurs only with fat malabsorption or genetic abnormalities of vitamin E metabolism/transport
Vitamin K	Elevated prothrombin time, bleeding	<10 µg/d	Fat malabsorption, liver disease, antibiotic use

FIGURE 326-1

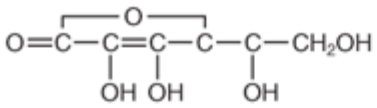
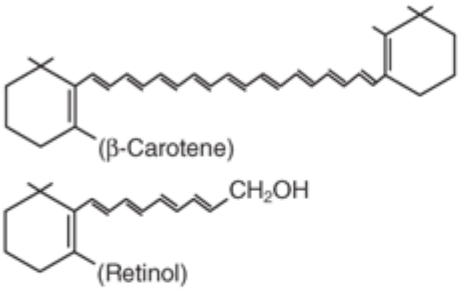
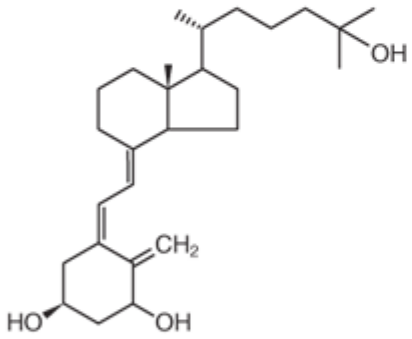
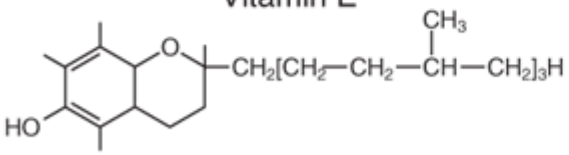
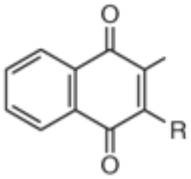
Structures and principal functions of vitamins associated with human disorders.

Vitamin	Active derivative or cofactor form	Principal function
Thiamine (B_1) 	Thiamine pyrophosphate	Coenzyme for cleavage of carbon-carbon bonds; amino acid and carbohydrate metabolism
Riboflavin (B_2) 	Flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD)	Cofactor for oxidation, reduction reactions

	(FAD)	reactions, and covalently attached prosthetic groups for some enzymes
Niacin 	Nicotinamide adenine dinucleotide phosphate (NADP) and nicotinamide adenine dinucleotide (NAD)	Coenzymes for oxidation and reduction reactions
Vitamin B_6 	Pyridoxal phosphate	Cofactor for enzymes of amino acid metabolism
Folate 	Polyglutamate forms of (5, 6, 7, 8) tetrahydrofolate with carbon unit attachments	Coenzyme for one carbon transfer in nucleic acid and amino acid metabolism
Vitamin B_{12} 	Methylcobalamine Adenosylcobalamin	Coenzyme for methionine synthase and L-methylmalonyl-CoA mutase

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition

J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Vitamin	Active derivative or cofactor form	Principal function
Vitamin C 	Ascorbic acid and dehydroascorbic acid	Participation as a redox ion in many biologic oxidation and hydrogen transfer reactions
Vitamin A 	Retinol, retinaldehyde, and retinoic acid	Formation of rhodopsin (vision) and glycoproteins (epithelial cell function); also regulates gene transcription
Vitamin D 	1,25-Dihydroxyvitamin D	Maintenance of blood calcium and phosphorus levels; antiproliferative hormone
Vitamin E 	Tocopherols and tocotrienols	Antioxidants
Vitamin K 	Vitamin K hydroquinone	Cofactor for posttranslation carboxylation of many proteins including essential clotting factors

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THIAMINE (VITAMIN B_1)

Thiamine was the first B vitamin to be identified and therefore is referred to as vitamin B_1. Thiamine functions in the decarboxylation of α-ketoacids (e.g., pyruvate α-ketoglutarate) and branched-chain amino acids and thus is essential for energy generation. In addition, thiamine pyrophosphate acts as a coenzyme for a transketolase reaction that mediates the conversion of hexose and pentose phosphates. It has been postulated that thiamine plays a role in peripheral nerve conduction, although the exact chemical reactions underlying this function are not known.

Food Sources

The median intake of thiamine in the United States from food alone is 2 mg/d. Primary food sources for thiamine include yeast, organ meat, pork, legumes, beef, whole grains, and nuts. Milled rice and grains contain little thiamine. Thiamine deficiency is therefore more common in cultures that rely heavily on a rice-based diet. Tea, coffee (regular and decaffeinated), raw fish, and shellfish contain thiaminases, which can destroy the vitamin. Thus, drinking large amounts of tea or coffee could theoretically lower thiamine body stores.

Deficiency

Most dietary deficiency of thiamine worldwide is the result of poor dietary intake. In Western countries, the primary causes of thiamine deficiency are alcoholism and chronic illnesses such as cancer. Alcohol interferes directly with the absorption of thiamine and with the synthesis of thiamine pyrophosphate, and it increases urinary excretion. Thiamine should always be replenished when a patient with alcoholism is being refed, as carbohydrate repletion without adequate thiamine can precipitate acute thiamine deficiency with lactic acidosis. Other at-risk populations are women with prolonged hyperemesis gravidarum and anorexia, patients with overall poor nutritional status who are receiving parenteral glucose, patients who have had bariatric/metabolic surgery (*bariatric Wernicke*), and patients receiving chronic diuretic therapy (e.g., in hypertension or heart failure) due to increased urinary thiamine losses. Maternal thiamine deficiency can lead to infantile beriberi in breast-fed children. Thiamine deficiency could be an underlying factor in motor vehicle accidents and could be overlooked in the setting of head injury.

Thiamine deficiency in its early stage induces anorexia and nonspecific symptoms (e.g., irritability, decrease in short-term memory). Prolonged thiamine deficiency causes *beriberi*, which is classically categorized as wet or dry although there is considerable overlap between the two categories. In either form of *beriberi*, patients may complain of pain and paresthesia. *Wet beriberi* presents primarily with cardiovascular symptoms that are due to impaired myocardial energy metabolism and dysautonomia; it can occur after 3 months of a thiamine-deficient diet. Patients present with an enlarged heart, tachycardia, high-output congestive heart failure, peripheral edema, and peripheral neuritis. Patients with *dry beriberi* present with a symmetric peripheral neuropathy of the motor and sensory systems, with diminished reflexes. The neuropathy affects the legs most markedly, and patients have difficulty rising from a squatting position.

Alcoholic patients with chronic thiamine deficiency also may have central nervous system (CNS) manifestations known as *Wernicke's encephalopathy*, which consists of horizontal nystagmus,

ophthalmoplegia (due to weakness of one or more extraocular muscles), cerebellar ataxia, and mental impairment **(Chap. 445)**. When there is an additional loss of memory and a confabulatory psychosis, the syndrome is known as *Wernicke-Korsakoff syndrome*. Despite the typical clinical picture and history, Wernicke-Korsakoff syndrome is underdiagnosed.

The laboratory diagnosis of thiamine deficiency usually is made by a functional enzymatic assay of transketolase activity measured before and after the addition of thiamine pyrophosphate. A >25% stimulation in response to the addition of thiamine pyrophosphate (i.e., an activity coefficient of 1.25) is interpreted as abnormal. Thiamine or the phosphorylated esters of thiamine in serum or blood also can be measured by high-performance liquid chromatography to detect deficiency.

TREATMENT

TREATMENT

Thiamine Deficiency

In acute thiamine deficiency with either cardiovascular or neurologic signs, 200 mg of thiamine three times daily should be given intravenously until there is no further improvement in acute symptoms; oral thiamine (10 mg/d) should subsequently be given until recovery is complete. Cardiovascular and ophthalmoplegic improvement occurs within 24 h. Other manifestations gradually clear, although psychosis in Wernicke-Korsakoff syndrome may be permanent or may persist for several months. Other nutrient deficiencies should be corrected concomitantly.

Toxicity

Although anaphylaxis has been reported after high intravenous doses of thiamine, no adverse effects have been recorded from either food or supplements at high doses. Thiamine supplements may be bought over the counter in doses of up to 50 mg/d.

RIBOFLAVIN (VITAMIN B_2)

Riboflavin is important for the metabolism of fat, carbohydrate, and protein, acting as a respiratory coenzyme and an electron donor. Enzymes that contain flavin adenine dinucleotide (FAD) or flavin mononucleotide (FMN) as prosthetic groups are known as *flavoenzymes* (e.g., succinic acid dehydrogenase, monoamine oxidase, glutathione reductase). FAD is a cofactor for methyltetrahydrofolate reductase and therefore modulates homocysteine metabolism. The vitamin also plays a role in drug and steroid metabolism, including detoxification reactions.

Although much is known about the chemical and enzymatic reactions of riboflavin, the clinical manifestations of riboflavin deficiency are nonspecific and are similar to those of other deficiencies of B vitamins. Riboflavin deficiency is manifested principally by lesions of the mucocutaneous surfaces of the

mouth and skin. In addition, corneal vascularization, anemia, and personality changes have been described with riboflavin deficiency.

Deficiency and Excess

Riboflavin deficiency almost always is due to dietary deficiency. Milk, other dairy products, and enriched breads and cereals are the most important dietary sources of riboflavin in the United States, although lean meat, fish, eggs, broccoli, and legumes are also good sources. Riboflavin is extremely sensitive to light, and milk should be stored in containers that protect against photodegradation. Laboratory diagnosis of riboflavin deficiency can be made by determination of red blood cell or urinary riboflavin concentrations or by measurement of erythrocyte glutathione reductase activity, with and without added FAD. Because the capacity of the gastrointestinal tract to absorb riboflavin is limited (~20 mg after one oral dose), riboflavin toxicity has not been described.

NIACIN (VITAMIN B_3)

The term *niacin* refers to nicotinic acid and nicotinamide and their biologically active derivatives. Nicotinic acid and nicotinamide serve as precursors of two coenzymes, nicotinamide adenine dinucleotide (NAD) and NAD phosphate (NADP), which are important in numerous oxidation and reduction reactions in the body. In addition, NAD and NADP are active in adenine diphosphate–ribose transfer reactions involved in DNA repair and calcium mobilization.

Metabolism and Requirements

Nicotinic acid and nicotinamide are absorbed well from the stomach and small intestine. The bioavailability of niacin from beans, milk, meat, and eggs is high; bioavailability from cereal grains is lower. Since flour is enriched with “free” niacin (i.e., the non-coenzyme form), bioavailability is excellent. Median intakes of niacin in the United States considerably exceed the recommended dietary allowance (RDA).

The amino acid tryptophan can be converted to niacin with an efficiency of 60:1 by weight. Thus, the RDA for niacin is expressed in niacin equivalents. A lower-level conversion of tryptophan to niacin occurs in vitamin B_6 and/or riboflavin deficiencies and in the presence of isoniazid. The urinary excretion products of niacin include 2-pyridone and 2-methyl nicotinamide, measurements of which are used in the diagnosis of niacin deficiency.

Deficiency

Niacin deficiency causes *pellagra*, which is found mostly among people eating corn-based diets in parts of China, Africa, and India. Pellagra in North America is found mainly among alcoholics; among patients with congenital defects of intestinal and kidney absorption of tryptophan (Hartnup disease; **Chap. 413**); and among patients with carcinoid syndrome **(Chap. 80)**, in which there is increased conversion of tryptophan to serotonin. The antituberculosis drug isoniazid is a structural analog of niacin and can precipitate pellagra. In

the setting of famine or population displacement, pellagra results from the absolute lack of niacin but also from the deficiency of micronutrients required for the conversion of tryptophan to niacin (e.g., iron, riboflavin, and pyridoxine). The early symptoms of pellagra include loss of appetite, generalized weakness and irritability, abdominal pain, and vomiting. Bright red glossitis then ensues and is followed by a characteristic skin rash that is pigmented and scaling, particularly in skin areas exposed to sunlight. This rash is known as *Casal's necklace* because it forms a ring around the neck; it is seen in advanced cases. Vaginitis and esophagitis also may occur. Diarrhea (due in part to proctitis and in part to malabsorption), depression, seizures, and dementia are also part of the pellagra syndrome. The primary manifestations of this syndrome are sometimes referred to as "the four *D*s": *d*ermatitis, *d*iarrhea, and *d*ementia leading to *d*eath.

TREATMENT

TREATMENT

Pellagra

Treatment of pellagra consists of oral supplementation with 100–200 mg of nicotinamide or nicotinic acid three times daily for 5 days. High doses of nicotinic acid (2 g/d in a time-release form) are used for the treatment of elevated cholesterol and triglyceride levels and/or low high-density lipoprotein cholesterol levels without, however, a proven benefit on cardiovascular endpoints (Chap. 400).

Toxicity

Prostaglandin-mediated flushing due to binding of the vitamin to a G protein–coupled receptor has been observed at daily nicotinic acid doses as low as 30 mg taken as a supplement or as therapy for dyslipidemia. There is no evidence of toxicity from niacin that is derived from food sources. Flushing always starts in the face and may be accompanied by skin dryness, itching, paresthesia, and headache. Flushing is subject to tachyphylaxis and often improves with time; premedication with aspirin may alleviate these symptoms. Nausea, vomiting, and abdominal pain also occur at similar doses of niacin. Hepatic toxicity is the most serious toxic reaction caused by sustained-release niacin and may present as jaundice with elevated aspartate aminotransferase (AST) and alanine aminotransferase (ALT) levels. A few cases of fulminant hepatitis requiring liver transplantation have been reported at doses of 3–9 g/d. Other toxic reactions include glucose intolerance, hyperuricemia, macular edema, and macular cysts. The combination of nicotinic acid preparations for dyslipidemia with 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase inhibitors may increase the risk of rhabdomyolysis. The upper limit for daily niacin intake has been set at 35 mg. However, this upper limit does not pertain to the therapeutic use of niacin.

PYRIDOXINE (VITAMIN B_6)

Vitamin B_6 refers to a family of compounds that includes pyridoxine, pyridoxal, pyridoxamine, and their 5′-phosphate derivatives. 5′-Pyridoxal phosphate (PLP) is a cofactor for >100 enzymes involved in amino acid metabolism. Vitamin B_6 also is involved in heme and neurotransmitter synthesis and in the metabolism of

glycogen, lipids, steroids, sphingoid bases, and several vitamins, including the conversion of tryptophan to niacin.

Dietary Sources

Plants contain vitamin B_6 in the form of pyridoxine, whereas animal tissues contain PLP and pyridoxamine phosphate. The vitamin B_6 contained in plants is less bioavailable than that in animal tissues. Rich food sources of vitamin B_6 include legumes, nuts, wheat bran, and meat, although it is present in all food groups.

Deficiency

Symptoms of vitamin B_6 deficiency include epithelial changes, as seen frequently with other B vitamin deficiencies. In addition, severe vitamin B_6 deficiency can lead to peripheral neuropathy, abnormal electroencephalograms, and personality changes that include depression and confusion. In infants, diarrhea, seizures, and anemia have been reported. Microcytic hypochromic anemia is due to diminished hemoglobin synthesis, since the first enzyme involved in heme biosynthesis (aminolevulinate synthase) requires PLP as a cofactor **(Chap. 93)**. In some case reports, platelet dysfunction has been reported. Since vitamin B_6 is necessary for the conversion of homocysteine to cystathionine, it is possible that chronic low-grade vitamin B_6 deficiency may result in hyperhomocysteinemia and increased risk of cardiovascular disease **(Chap. 413)**. Independent of homocysteine, low levels of circulating vitamin B_6 have been associated with inflammation and elevated levels of C-reactive protein.

Certain medications, such as isoniazid, L-dopa, penicillamine, and cycloserine, interact with PLP due to a reaction with carbonyl groups. Pyridoxine should be given concurrently with isoniazid to avoid neuropathy. The increased ratio of AST to ALT seen in alcoholic liver disease reflects the relative vitamin B_6 dependence of ALT. Vitamin B_6 dependency syndromes that require pharmacologic doses of vitamin B_6 are rare; they include cystathionine β-synthase deficiency, pyridoxine-responsive (primarily sideroblastic) anemias, and gyrate atrophy with chorioretinal degeneration due to decreased activity of the mitochondrial enzyme ornithine aminotransferase. In these situations, 100–200 mg/d of oral vitamin B_6 is required for treatment.

Severe nausea and vomiting in pregnancy might respond to pyridoxine combined with doxylamine. High doses of vitamin B_6 have been used to treat carpal tunnel syndrome, premenstrual syndrome, schizophrenia, autism, and diabetic neuropathy but have not been found to be effective.

The laboratory diagnosis of vitamin B_6 deficiency is generally based on low plasma PLP values (<20 nmol/L). Vitamin B_6 deficiency is treated with 50 mg/d; higher doses of 100–200 mg/d are given if the deficiency is related to medication use. Vitamin B_6 should not be given with L-dopa, since the vitamin interferes with the action of this drug.

Toxicity

The safe upper limit for vitamin B_6 has been set at 100 mg/d, although no adverse effects have been associated with high intakes of vitamin B_6 from food sources only. When toxicity occurs, it causes severe sensory neuropathy, leaving patients unable to walk. Some cases of photosensitivity and dermatitis have been reported.

FOLATE (VITAMIN B_{12})

See Chap. 95.

VITAMIN C

Both ascorbic acid (only the L-isomer) and its oxidized product dehydroascorbic acid are biologically active. Actions of vitamin C include antioxidant activity, promotion of nonheme iron absorption, carnitine biosynthesis, conversion of dopamine to norepinephrine, and synthesis of many peptide hormones. Vitamin C is also important for connective tissue metabolism and cross-linking (proline hydroxylation), and it is a component of many drug-metabolizing enzyme systems, particularly the mixed-function oxidase systems.

Absorption and Dietary Sources

Vitamin C is almost completely absorbed if <100 mg is administered in a single dose; however, only ≤50% is absorbed at doses >1 g. Enhanced degradation and fecal and urinary excretion of vitamin C occur at higher intake levels.

Good dietary sources of vitamin C include citrus fruits, green vegetables (especially broccoli), tomatoes, and potatoes. Consumption of five servings of fruits and vegetables a day provides vitamin C in excess of the RDA of 90 mg/d for men and 75 mg/d for women. In addition, ~40% of the U.S. population consumes vitamin C as a dietary supplement in which “natural forms” of the vitamin are no more bioavailable than synthetic forms. Smoking, hemodialysis, pregnancy, lactation, and stress (e.g., infection, trauma) appear to increase vitamin C requirements.

Deficiency

Vitamin C deficiency causes scurvy. In the United States, this condition is seen primarily among the poor and the elderly, in alcoholics who consume <10 mg/d of vitamin C, and in individuals consuming macrobiotic diets. Vitamin C deficiency also can occur in young adults who eat severely unbalanced diets. In addition to generalized fatigue, symptoms of scurvy primarily reflect impaired formation of mature connective tissue and include bleeding into the skin (petechiae, ecchymoses, perifollicular hemorrhages); inflamed and bleeding gums; and manifestations of bleeding into joints, the peritoneal cavity, the pericardium, and the adrenal glands. In children, vitamin C deficiency may cause impaired bone growth. Laboratory diagnosis of vitamin C deficiency is based on low plasma or leukocyte levels.

Administration of vitamin C (200 mg/d) improves the symptoms of scurvy within several days. High-dose vitamin C supplementation (e.g., 0.2 up to several grams per day) may slightly decrease the symptoms and duration of upper respiratory tract infections. Vitamin C supplementation has also been reported to be useful in Chédiak-Higashi syndrome (Chap. 60) and osteogenesis imperfecta (Chap. 406). Diets high in vitamin C have been claimed to lower the incidence of certain cancers, particularly esophageal and gastric cancers. If proved, this effect may be because vitamin C can prevent the conversion of nitrites and secondary amines to carcinogenic nitrosamines. Evidence for a potential role of pro-oxidative effects of parenteral ascorbic acid in the treatment of advanced cancers is emerging in laboratory studies.

Toxicity

Taking >2 g of vitamin C in a single dose may result in abdominal pain, diarrhea, and nausea. Since vitamin C may be metabolized to oxalate, it is feared that chronic high-dose vitamin C supplementation could result in an increased prevalence of kidney stones. However, except in patients with preexisting renal disease, this association has not been borne out in several trials. Nevertheless, it is reasonable to advise patients with a history of kidney stones and renal insufficiency not to take large doses of vitamin C. There is also an unproven but possible risk that chronic high doses of vitamin C could promote iron overload and iron toxicity. High doses of vitamin C can induce hemolysis in patients with glucose-6-phosphate dehydrogenase deficiency, and doses >1 g/d can cause false-negative guaiac reactions and interfere with tests for urinary glucose. High doses may interfere with the activity of certain drugs (e.g., bortezomib in myeloma patients).

BIOTIN

Biotin is a water-soluble vitamin that plays a role in gene expression, gluconeogenesis, and fatty acid synthesis and serves as a CO_2 carrier on the surface of both cytosolic and mitochondrial carboxylase enzymes. The vitamin also functions in the catabolism of specific amino acids (e.g., leucine) and in gene regulation by histone biotinylation. Excellent food sources of biotin include organ meat such as liver or kidney, soy and other beans, yeast, and egg yolks; however, egg white contains the protein avidin, which strongly binds the vitamin and reduces its bioavailability.

Biotin deficiency due to low dietary intake is rare; rather, deficiency is due to inborn errors of metabolism. Biotin deficiency has been induced by experimental feeding of egg white diets and by biotin-free parenteral nutrition in patients with short bowels. In adults, biotin deficiency results in mental changes (depression, hallucinations), paresthesia, anorexia, and nausea. A scaling, seborrheic, and erythematous rash may occur around the eyes, nose, and mouth as well as on the extremities. In infants, biotin deficiency presents as hypotonia, lethargy, and apathy. In addition, infants may develop alopecia and a characteristic rash that includes the ears. The laboratory diagnosis of biotin deficiency can be established on the basis of a decreased concentration of urinary biotin (or its major metabolites), increased urinary excretion of 3-hydroxyisovaleric acid after a leucine challenge, or decreased activity of biotin-dependent enzymes in lymphocytes (e.g., propionyl-CoA carboxylase). Treatment requires pharmacologic doses of biotin, that is, up to 10 mg/d. No toxicity is known.

PANTOTHENIC ACID (VITAMIN B_5)

Pantothenic acid is a component of coenzyme A and phosphopantetheine, which are involved in fatty acid metabolism and the synthesis of cholesterol, steroid hormones, and all compounds formed from isoprenoid units. In addition, pantothenic acid is involved in the acetylation of proteins. The vitamin is excreted in the urine, and the laboratory diagnosis of deficiency is based on low urinary vitamin levels.

The vitamin is ubiquitous in the food supply. Liver, yeast, egg yolks, whole grains, and vegetables are particularly good sources. Human pantothenic acid deficiency has been demonstrated only by experimental feeding of diets low in pantothenic acid or by administration of a specific pantothenic acid antagonist. The symptoms of pantothenic acid deficiency are nonspecific and include gastrointestinal disturbance, depression, muscle cramps, paresthesia, ataxia, and hypoglycemia. Pantothenic acid deficiency is believed to have caused the “burning feet syndrome” seen in prisoners of war during World War II. No toxicity of this vitamin has been reported.

CHOLINE

Choline is a precursor for acetylcholine, phospholipids, and betaine. Choline is necessary for the structural integrity of cell membranes, cholinergic neurotransmission, lipid and cholesterol metabolism, methyl-group metabolism, and transmembrane signaling. Recently, a recommended adequate intake was set at 550 mg/d for men and 425 mg/d for women, although certain genetic polymorphisms can increase an individual’s requirement. Choline is thought to be a “conditionally essential” nutrient in that its de novo synthesis occurs in the liver and results in lesser-than-used amounts only under certain stress conditions (e.g., alcoholic liver disease). The dietary requirement for choline depends on the status of other nutrients involved in methyl-group metabolism (folate, vitamin B_{12}, vitamin B_6, and methionine) and thus varies widely. Choline is widely distributed in food (e.g., egg yolks, wheat germ, organ meat, milk) in the form of lecithin (phosphatidylcholine). Choline deficiency has occurred in patients receiving parenteral nutrition devoid of choline. Deficiency results in fatty liver, elevated aminotransferase levels, and skeletal muscle damage with high creatine phosphokinase values. The diagnosis of choline deficiency is currently based on low plasma levels, although nonspecific conditions (e.g., heavy exercise) may also suppress plasma levels.

Toxicity from choline results in hypotension, cholinergic sweating, diarrhea, salivation, and a fishy body odor. The upper limit for choline intake has been set at 3.5 g/d. Because of its ability to lower cholesterol and homocysteine levels, choline treatment has been suggested for patients with dementia and patients at high risk of cardiovascular disease. However, the benefits of such treatment have not been firmly documented. Choline- and betaine-restricted diets are of therapeutic value in trimethylaminuria (“fish odor syndrome”).

FLAVONOIDS

Flavonoids constitute a large family of polyphenols that contribute to the aroma, taste, and color of fruits and vegetables. Major groups of dietary flavonoids include anthocyanidins in berries; catechins in green tea and chocolate; flavonols (e.g., quercetin) in broccoli, kale, leeks, onions, and the skins of grapes and apples;

and isoflavones (e.g., genistein) in legumes. Isoflavones have a low bioavailability and are partially metabolized by the intestinal flora. The dietary intake of flavonoids is estimated at 10–100 mg/d; this figure is almost certainly an underestimate attributable to a lack of information on their concentrations in many foods. Several flavonoids have antioxidant activity and affect cell signaling. From observational epidemiologic studies and limited clinical (human and animal) studies, flavonoids have been postulated to play a role in the prevention of several chronic diseases, including neurodegenerative disease, diabetes, and osteoporosis. The ultimate importance and usefulness of these compounds against human disease have not been consistently demonstrated.

VITAMIN A

Vitamin A, in the strictest sense, refers to retinol. However, the oxidized metabolites retinaldehyde and retinoic acid are also biologically active compounds. The term *retinoids* includes all molecules (including synthetic molecules) that are chemically related to retinol. Retinaldehyde (11-*cis*) is the form of vitamin A that is required for normal vision, whereas retinoic acid is necessary for normal morphogenesis, growth, and cell differentiation. Retinoic acid does not function directly in vision and, in contrast to retinol, is not involved in reproduction. Vitamin A also plays a role in iron utilization, humoral immunity, T cell–mediated immunity, natural killer cell activity, and phagocytosis.

Vitamin A is found in the human food supply in two forms: preformed as esters and provitamin A in carotenoids. There are >600 carotenoids in nature, ~50 of which can be metabolized to vitamin A. β-Carotene is the most prevalent carotenoid with provitamin A activity in the food supply. In humans, significant fractions of carotenoids are absorbed intact and are stored in liver and fat. It is estimated that ≥12 μg (range, 4–27 μg) of dietary all-*trans* β-carotene is equivalent to 1 μg of retinol activity, whereas the figure is ≥24 μg for other dietary provitamin A carotenoids (e.g., cryptoxanthin, α-carotene). The vitamin A equivalency for a β-carotene supplement in an oily solution is 2:1.

Metabolism

The liver contains ~90% of the vitamin A reserves and secretes vitamin A in the form of retinol, which is bound in the circulation to retinol-binding protein. Once binding has occurred, the retinol-binding protein complex interacts with a second protein, transthyretin. This trimolecular complex functions to prevent vitamin A from being filtered by the kidney glomerulus, thus protecting the body against the toxicity of retinol and allowing retinol to be taken up by specific cell-surface receptors that recognize retinol-binding protein. A certain amount of vitamin A enters peripheral cells even if it is not bound to retinol-binding protein. After retinol is internalized by the cell, it becomes bound to a series of cellular retinol-binding proteins, which function as sequestering and transporting agents as well as co-ligands for enzymatic reactions. Certain cells also contain retinoic acid–binding proteins, which have sequestering functions but also shuttle retinoic acid to the nucleus and enable its metabolism.

Retinoic acid is a ligand for certain nuclear receptors that act as transcription factors. Two families of receptors (retinoic acid receptors [RARs] and retinoid X receptors [RXRs]) are active in retinoid-mediated gene

transcription. Retinoid receptors regulate transcription by binding as dimeric complexes to specific DNA sites—the retinoic acid response elements—in target genes **(Chap. 370)**. The receptors can either stimulate or repress gene expression in response to their ligands. RARs bind all-*trans* retinoic acid and 9-*cis*-retinoic acid, whereas RXRs bind only 9-*cis*-retinoic acid.

The retinoid receptors play an important role in controlling cell proliferation and differentiation. RXRs dimerize with other nuclear receptors to function as coregulators of genes responsive to retinoids, but also to thyroid hormone and calcitriol. RXR agonists induce insulin sensitivity experimentally, perhaps because RXRs are cofactors for the peroxisome proliferator-activated receptors, which mediate also fatty acid and carbohydrate metabolism and are targets for different drugs including thiazolidinedione drugs (e.g., rosiglitazone and pioglitazone) **(Chap. 397)**.

Dietary Sources

The retinol activity equivalent (RAE) is used to express the vitamin A value of food: 1 RAE is defined as 1 µg of retinol (0.003491 mmol), 12 µg of β-carotene, and 24 µg of other provitamin A carotenoids. In older literature, vitamin A often was expressed in international units (IUs), with 1 µg of retinol equal to 3.33 IU of retinol and 20 IU of β-carotene. Although these IUs are no longer in scientific use, they can still be found in reports of the food industry and in public health interventions in low-income countries.

Liver, fish, and eggs are excellent food sources for preformed vitamin A; vegetable sources of provitamin A carotenoids include dark green and deeply colored fruits and vegetables. Moderate cooking of vegetables enhances carotenoid release for uptake in the gut. Carotenoid absorption is also aided by some fat in a meal. Exclusive breast-feeding can cover the vitamin A needs of infants if the mother has an adequate vitamin A status and a large enough volume of milk. If the nursing mother has inadequate vitamin A intake, concomitant diseases, or her infant was a preterm delivery, breast milk probably will not supply enough vitamin A to prevent deficiency. In developing countries, chronic dietary deficiency is the main cause of vitamin A deficiency and is exacerbated by infection. In early childhood, low vitamin A status results from inadequate intakes of animal food sources and edible oils, both of which are expensive, coupled with seasonal unavailability of vegetables and fruits and lack of marketed fortified food products. Factors that interfere with vitamin A metabolism may also affect status or function. For example, concurrent zinc deficiency can interfere with the mobilization of vitamin A from liver stores. Alcohol interferes with the conversion of retinol to retinaldehyde in the eye by competing for alcohol (retinol) dehydrogenase. Drugs that interfere with the absorption of vitamin A include mineral oil, neomycin, and cholestyramine.

Deficiency


Vitamin A deficiency is endemic in areas where diets are chronically poor, especially in southern Asia, sub-Saharan Africa, some parts of Latin America, and the western Pacific, including parts of China. Vitamin A status is usually assessed by measuring serum retinol (normal range, 1.05–3.50 µmol/L [30–100 µg/dL]) or blood-spot retinol or by tests of dark adaptation. Stable isotopic or invasive liver biopsy methods are available to estimate total body stores of vitamin A. As judged by deficient serum retinol (<0.70 µmol/L [20

µg/dL]), vitamin A deficiency worldwide is present in 190 million preschool-age children, among whom >5 million have an ocular manifestation of deficiency termed *xerophthalmia*. This condition includes milder stages of night blindness and conjunctival *xerosis* (dryness) with *Bitot's spots* (white patches of keratinized epithelium appearing on the sclera) that may affect 1–5% of children in deficient populations as well as rare, potentially blinding corneal ulceration and necrosis. *Keratomalacia* (softening of the cornea) leads to corneal scarring that blinds an estimated quarter of a million children each year and is associated with fatality rates of 4–25%. However, vitamin A deficiency severe enough to cause any clinical stage poses an increased risk of death from diarrhea, dysentery, measles, malaria, or respiratory disease. This is because vitamin A deficiency can compromise barrier, innate, and acquired immune defenses to infection. In areas where deficiency is widely prevalent, vitamin A supplementation can markedly reduce the risk of childhood mortality (by 23–34%, on average). About 10% of pregnant women in undernourished settings also develop night blindness (assessed by history) during the latter half of pregnancy; this level of moderate to severe vitamin A deficiency is associated with an increased risk of maternal infection and death. Maternal vitamin A deficiency may also exacerbate already low vitamin A nutrition and associated risks for the newborn. In South Asia, where maternal deficiency is prominent, giving infants a single oral dose (50,000 IU) of vitamin A shortly after birth has reduced infant mortality by ≥10%, whereas in African settings less affected by maternal vitamin A deficiency, no effect has been noted, revealing differences in risk of deficiency and benefit of supplementation across regions.

TREATMENT

TREATMENT

Vitamin A Deficiency

Vitamin A is commercially available for treatment and prevention in esterified forms (e.g., acetate, palmitate), which are more stable than other forms. Any stage of xerophthalmia should be treated with 60 mg (or RAE) or 200,000 IU of vitamin A in oily solution, usually contained in a soft-gel capsule. The same dose is repeated 1 and 14 days later. Doses should be reduced by half for patients 6–11 months of age. Mothers with night blindness or Bitot's spots should be given vitamin A orally 3 mg daily for at least 3 months. These regimens are efficacious, and they are far less expensive and more widely available than injectable water-miscible vitamin A. A common approach to prevention is to provide vitamin A supplementation every 4–6 months to young children 6 months to 5 years of age (both HIV-positive and HIV-negative) in high-risk areas. For prevention, infants 6–11 months of age should receive 30 mg vitamin A; children 12–59 months of age, 60 mg. For reasons that are not clear, while early neonatal vitamin A may reduce infant mortality, vitamin A given between 1 and 5 months of age has not proven effective in improving survival in high-risk settings.

Uncomplicated vitamin A deficiency is rare in industrialized countries. One high-risk group—extremely low-birth-weight (<1000-g) infants—is likely to be vitamin A–deficient and should receive a supplement of 1500 µg (or RAE) three times a week for 4 weeks. Severe measles in any society can lead to secondary vitamin A deficiency. Children hospitalized with measles should receive two 60-mg doses of vitamin A on two consecutive days. Vitamin A deficiency most often occurs in patients with malabsorptive diseases (e.g., celiac

sprue, short-bowel syndrome) who have abnormal dark adaptation or symptoms of night blindness without other ocular changes. Typically, such patients are diagnosed in advanced care settings where they are treated for 1 month with 15 mg/d of a water-miscible preparation of vitamin A. This treatment is followed by a lower maintenance dose, with the exact amount determined by monitoring serum retinol. Finding application elsewhere in medicine, retinoic acid is useful in the treatment of promyelocytic leukemia **(Chap. 100)** and also is used in the treatment of cystic acne because it inhibits keratinization, decreases sebum secretion, and possibly alters the inflammatory reaction **(Chap. 53)**.

No specific signs or symptoms result from carotenoid deficiency. It was postulated that β-carotene would be an effective chemopreventive agent for cancer because numerous epidemiologic studies had shown that diets high in β-carotene were associated with lower incidences of cancers of the respiratory and digestive systems. However, intervention studies in smokers found that treatment with high doses of β-carotene actually resulted in more lung cancers than did treatment with placebo. Non–provitamin A carotenoids such as lutein and zeaxanthin have been suggested to confer protection against macular degeneration, and one large-scale intervention study did not show a beneficial effect except in those with a low lutein status. The use of the non–provitamin A carotenoid lycopene to protect against prostate cancer has been proposed. Again, however, the effectiveness of these agents has not been proved by intervention studies, and the mechanisms underlying these purported biologic actions are unknown.

Selective plant-breeding techniques that lead to a higher provitamin A carotenoid content in staple foods may decrease vitamin A malnutrition in low-income countries. Moreover, a recently developed genetically modified food (Golden Rice) has an improved β-carotene–to–vitamin A conversion ratio of ~3:1.

Toxicity

The acute toxicity of vitamin A was first noted in Arctic explorers who ate polar bear liver and has also been seen after administration of 150 mg to adults or 100 mg to children. Acute toxicity is manifested by increased intracranial pressure, vertigo, diplopia, bulging fontanels (in children), seizures, and exfoliative dermatitis; it may result in death. Among children being treated for vitamin A deficiency according to the protocols outlined above, transient bulging of fontanels occurs in 2% of infants, and transient nausea, vomiting, and headache occur in 5% of preschoolers. Chronic vitamin A intoxication is largely a concern in industrialized countries and has been seen in otherwise healthy adults who ingest 15 mg/d and children who ingest 6 mg/d over a period of several months. Manifestations include dry skin, cheilosis, glossitis, vomiting, alopecia, bone demineralization and pain, hypercalcemia, lymph node enlargement, hyperlipidemia, amenorrhea, and features of pseudotumor cerebri with increased intracranial pressure and papilledema. Liver fibrosis with portal hypertension may also result from chronic vitamin A intoxication. Provision of vitamin A in excess to pregnant women has resulted in spontaneous abortion and in congenital malformations, including craniofacial abnormalities and valvular heart disease. In pregnancy, the daily dose of vitamin A should not exceed 3 mg. Commercially available retinoid derivatives are also toxic, including 13-*cis*-retinoic acid, which has been associated with birth defects. Thus contraception should be continued for at least 1 year and possibly longer in women who have taken 13-*cis*-retinoic acid.

In malnourished children, vitamin A supplements (30–60 mg), in amounts calculated as a function of age and given in several rounds over 2 years, are considered to amplify nonspecific effects of vaccines. However, for unclear reasons, in one African setting there has been a negative effect on mortality rates in incompletely vaccinated girls.

High doses of carotenoids do not result in toxic symptoms but should be avoided in smokers due to an increased risk of lung cancer. Very high doses of β-carotene (~200 mg/d) have been used to treat or prevent the skin rashes of erythropoietic protoporphyria. Carotenemia, which is characterized by a yellowing of the skin (increases of the palms and soles) but not the sclerae, may follow ingestion of >30 mg of β-carotene daily. Hypothyroid patients are particularly susceptible to the development of carotenemia due to impaired breakdown of carotene to vitamin A. Reduction of carotenes in the diet results in the disappearance of skin yellowing and carotenemia over a period of 30–60 days.

VITAMIN D

The metabolism of the fat-soluble vitamin D is described in detail in **Chap. 402**. The biologic effects of this vitamin are mediated by vitamin D receptors, which are found in most tissues; binding with these receptors potentially expands vitamin D actions to many different cell systems and organs (e.g., immune cells, brain, breast, colon, and prostate) in addition to the classic endocrine effects on calcium and phosphate metabolism and bone health. Vitamin D is thought to be important for maintaining normal function of many nonskeletal tissues such as muscle (including heart muscle), for immune function, and for inflammation as well as for cell proliferation and differentiation. Studies have shown that vitamin D may be useful as adjunctive treatment for tuberculosis, psoriasis, and multiple sclerosis or for the prevention of certain cancers. Vitamin D insufficiency may increase the risk of type 1 diabetes mellitus, cardiovascular disease (insulin resistance, hypertension, or low-grade inflammation), or brain dysfunction (e.g., depression). However, the exact physiologic roles of vitamin D in these nonskeletal diseases and the importance of these roles have not been clarified.

The skin is a major source of vitamin D, which is synthesized upon skin exposure to ultraviolet B radiation (UV-B; wavelength, 290–320 nm). Except for fish, food (unless fortified) contains only limited amounts of vitamin D. Vitamin D_2 (ergocalciferol) is obtained from plant sources and is the chemical form found in some supplements.

Deficiency

Vitamin D status has been assessed by measuring serum levels of 25-dihydroxyvitamin D (25[OH] vitamin D); however, there is no consensus on a uniform assay or on optimal serum levels. The optimal level might, in fact, differ according to the targeted disease entity. Epidemiologic and experimental data indicate that a 25(OH) vitamin D level of >20 ng/mL (≥50 nmol/L; to convert ng/mL to nmol/L, multiply by 2.496) is sufficient for good bone health. Some experts, however, advocate higher serum levels (e.g., >30 ng/mL) for other desirable endpoints of vitamin D action. There is insufficient evidence to recommend combined vitamin D

and calcium supplementation as a primary preventive strategy (as opposed to secondary prevention) for reduction of the incidence of fractures in healthy men and premenopausal women.

Risk factors for vitamin D deficiency are old age, lack of sun exposure, dark skin (especially among residents of northern latitudes), fat malabsorption, and obesity. *Rickets* represents the classic disease of vitamin D deficiency. Signs of deficiency are muscle soreness, weakness, and bone pain. Some of these effects are independent of calcium intake.

The U.S. National Academy of Sciences recently advised that the majority of adult North Americans should receive 600 IU/d of vitamin D (RDA = 15 µg/d or 600 IU/d; **Chap. 325**). However, for people aged >70 years, the RDA is set at 20 µg/d (800 IU/d). The consumption of fortified or enriched foods as well as suberythemal sun exposure should be encouraged for people at risk for vitamin D deficiency. If adequate intake is impossible, vitamin D supplements should be taken, especially during the winter months. Vitamin D deficiency can be treated by the oral administration of 50,000 IU/week for 6–8 weeks followed by a maintenance dose of 800 IU/d (20 µg/d) from food and supplements once normal plasma levels have been attained. There is uncertainty regarding the optimal therapeutic dosage (high vs low) for elderly at risk of falls. The physiologic effects of vitamin D_2 and vitamin D_3 are similar when these vitamins are ingested over long periods.

Toxicity

The upper limit of intake has been set at 4000 IU/d. Contrary to earlier beliefs, acute vitamin D intoxication is rare and usually is caused by the uncontrolled and excessive ingestion of supplements or by faulty food fortification practices. High plasma levels of $1,25(OH)_2$ vitamin D and calcium are central features of toxicity and mandate discontinuation of vitamin D and calcium supplements; in addition, treatment of hypercalcemia may be required.

VITAMIN E

Vitamin E is the collective designation for all stereoisomers of tocopherols and tocotrienols, although only the RR tocopherols meet human requirements. Vitamin E acts as a chain-breaking antioxidant and is an efficient peroxyl radical scavenger that protects low-density lipoproteins and polyunsaturated fats in membranes from oxidation. A network of other antioxidants (e.g., vitamin C, glutathione) and enzymes maintains vitamin E in a reduced state. Vitamin E also inhibits prostaglandin synthesis and the activities of protein kinase C and phospholipase A_2.

Absorption and Metabolism

After absorption, vitamin E is taken up from chylomicrons by the liver, and a hepatic α-tocopherol transport protein mediates intracellular vitamin E transport and incorporation into very low density lipoprotein. The transport protein has a particular affinity for the RRR isomeric form of α-tocopherol; thus, this natural isomer has the most biologic activity.

Requirement

Vitamin E is widely distributed in the food supply, with particularly high levels in sunflower oil, safflower oil, and wheat germ oil; γ-tocotrienols are notably present in soybean and corn oils. Vitamin E is also found in meats, nuts, and cereal grains, and small amounts are present in fruits and vegetables. Vitamin E pills containing doses of 50–1000 mg are ingested by ~10% of the U.S. population. The RDA for vitamin E is 15 mg/d (34.9 μmol or 22.5 IU) for all adults. Diets high in polyunsaturated fats may necessitate a slightly higher intake of vitamin E.

Dietary deficiency of vitamin E does not exist. Vitamin E deficiency is seen only in severe and prolonged malabsorptive diseases, such as celiac disease, or after small-intestinal resection or bariatric surgery. Children with cystic fibrosis or prolonged cholestasis may develop vitamin E deficiency characterized by areflexia and hemolytic anemia. Children with abetalipoproteinemia cannot absorb or transport vitamin E and become deficient quite rapidly. A familial form of isolated vitamin E deficiency also exists; it is due to a defect in the α-tocopherol transport protein. Vitamin E deficiency causes axonal degeneration of the large myelinated axons and results in posterior column and spinocerebellar symptoms. Peripheral neuropathy is initially characterized by areflexia, with progression to an ataxic gait, and by decreased vibration and position sensations. Ophthalmoplegia, skeletal myopathy, and pigmented retinopathy may also be features of vitamin E deficiency. A deficiency of either vitamin E or selenium in the host has been shown to increase certain viral mutations and, therefore, virulence. The laboratory diagnosis of vitamin E deficiency is based on low blood levels of α-tocopherol (<5 μg/mL, or <0.8 mg of α-tocopherol per gram of total lipids).

TREATMENT

TREATMENT

Vitamin E Deficiency

Symptomatic vitamin E deficiency should be treated with 800–1200 mg of α-tocopherol per day. Patients with abetalipoproteinemia may need as much as 5000–7000 mg/d. Children with symptomatic vitamin E deficiency should be treated orally with water-miscible esters (400 mg/d); alternatively, 2 mg/kg/d may be administered intramuscularly. Vitamin E in high doses may protect against oxygen-induced retrolental fibroplasia and bronchopulmonary dysplasia as well as intraventricular hemorrhage of prematurity. Vitamin E has been suggested to increase sexual performance, treat intermittent claudication, and slow the aging process, but convincing evidence for these properties is lacking. When given in combination with other antioxidants, vitamin E may help prevent macular degeneration. Vitamin E may have favorable therapeutic effects in noncirrhotic nondiabetic patients with NASH (nonalcoholic steatohepatitis). High doses (60–800 mg/d) of vitamin E have been shown in controlled trials to improve parameters of immune function and reduce colds in nursing home residents, but intervention studies using vitamin E to prevent cardiovascular disease or cancer have not shown efficacy, and, at doses >400 mg/d, vitamin E may even increase all-cause mortality rates and prostate cancer risk.

Toxicity

All forms of vitamin E are absorbed and could contribute to toxicity; however, the toxicity risk seems to be rather low as long as liver function is normal. High doses of vitamin E (>800 mg/d) may reduce platelet aggregation and interfere with vitamin K metabolism and are therefore contraindicated in patients taking warfarin and antiplatelet agents (such as aspirin or clopidogrel). Nausea, flatulence, and diarrhea have been reported at doses >1 g/d.

VITAMIN K

There are two natural forms of vitamin K: vitamin K_1, also known as *phylloquinone*, from vegetable sources, and vitamin K_2, or *menaquinones*, which are synthesized by bacterial flora and found in hepatic tissue. Phylloquinone can be converted to menaquinone in some organs.

Vitamin K is required for the posttranslational carboxylation of glutamic acid, which is necessary for calcium binding to γ-carboxylated proteins such as prothrombin (factor II); factors VII, IX, and X; protein C; protein S; and proteins found in bone (osteocalcin) and vascular smooth muscle (e.g., matrix Gla protein). However, the importance of vitamin K for bone mineralization and prevention of vascular calcification is not known. Warfarin-type drugs inhibit γ-carboxylation by preventing the conversion of vitamin K to its active hydroquinone form.

Dietary Sources

Vitamin K is found in green leafy vegetables such as kale and spinach, and appreciable amounts are also present in margarine and liver. Vitamin K is present in vegetable oils; olive, canola, and soybean oils are particularly rich sources. The average daily intake by Americans is estimated to be ~100 μg/d.

Deficiency

The symptoms of vitamin K deficiency are due to hemorrhage; newborns are particularly susceptible because of low fat stores, low breast milk levels of vitamin K, relative sterility of the infantile intestinal tract, liver immaturity, and poor placental transport. Intracranial bleeding as well as gastrointestinal and skin bleeding can occur in vitamin K–deficient infants 1–7 days after birth. Thus, vitamin K (0.5–1 mg IM) is given prophylactically at delivery.

Vitamin K deficiency in adults may be seen in patients with chronic small-intestinal disease (e.g., celiac disease, Crohn's disease), in those with obstructed biliary tracts, or after small-bowel resection. Broad-spectrum antibiotic treatment can precipitate vitamin K deficiency by reducing numbers of gut bacteria, which synthesize menaquinones, and by inhibiting the metabolism of vitamin K. In patients with warfarin therapy, the antiobesity drug orlistat can lead to changes in international normalized ratio due to vitamin K malabsorption. Vitamin K deficiency usually is diagnosed on the basis of an elevated prothrombin time or reduced clotting factors, although vitamin K may also be measured directly by high-pressure liquid chromatography. Vitamin K deficiency is treated with a parenteral dose of 10 mg. For patients with chronic malabsorption, 1–2 mg/d should be given orally or 1–2 mg per week can be taken parenterally. Patients with

liver disease may have an elevated prothrombin time because of liver cell destruction as well as vitamin K deficiency. If an elevated prothrombin time does not improve during vitamin K therapy, it can be deduced that this abnormality is not the result of vitamin K deficiency.

Toxicity

Toxicity from dietary phylloquinones and menaquinones has not been described. High doses of vitamin K can impair the actions of oral vitamin K antagonist anticoagulants.

MINERALS

See also Table 326-2.

TABLE 326-2

Deficiencies and Toxicities of Metals

Element	Deficiency	Toxicity	Tolerable Upper (Dietary) Intake Level
Boron	No biologic function determined	Developmental defects, male sterility, testicular atrophy	20 mg/d (extrapolated from animal data)
Calcium	Reduced bone mass, osteoporosis	Renal insufficiency (milk-alkali syndrome), nephrolithiasis, impaired iron absorption, thiazide diuretics	2500 mg/d (milk-alkali)
Copper	Anemia, growth retardation, defective keratinization and pigmentation of hair, hypothermia, degenerative changes in aortic elastin, osteopenia, mental deterioration	Nausea, vomiting, diarrhea, hepatic failure, tremor, mental deterioration, hemolytic anemia, renal dysfunction	10 mg/d (liver toxicity)
Chromium	Impaired glucose tolerance	*Occupational:* Renal failure, dermatitis, pulmonary cancer	Not determined
Fluoride	↑ Dental caries	Dental and skeletal fluorosis, osteosclerosis	10 mg/d (fluorosis)
Iodine	Thyroid enlargement, ↓ T_4, cretinism	Thyroid dysfunction, acne-like eruptions	1100 µg/d (thyroid dysfunction)

Element	Deficiency	Toxicity	Tolerable Upper (Dietary) Intake Level
Iron	Muscle abnormalities, koilonychia, pica, anemia, ↓ work performance, impaired cognitive development, premature labor, ↑ perinatal maternal death	Gastrointestinal effects (nausea, vomiting, diarrhea, constipation), iron overload with organ damage, acute and chronic systemic toxicity, increased susceptibility to malaria, increased risk association with certain chronic diseases (e.g., diabetes)	45 mg/d of elemental iron (gastrointestinal side effects)
Manganese	Impaired growth and skeletal development, reproduction, lipid and carbohydrate metabolism; upper body rash	*General:* Neurotoxicity, Parkinson-like symptoms *Occupational:* Encephalitis-like syndrome, Parkinson-like syndrome, psychosis, pneumoconiosis	11 mg/d (neurotoxicity)
Molybdenum	Severe neurologic abnormalities	Reproductive and fetal abnormalities	2 mg/d (extrapolated from animal data)
Selenium	Cardiomyopathy, heart failure, striated muscle degeneration	*General:* Alopecia, nausea, vomiting, abnormal nails, emotional lability, peripheral neuropathy, lassitude, garlic odor to breath, dermatitis *Occupational:* Lung and nasal carcinomas, liver necrosis, pulmonary inflammation	400 µg/d (hair, nail changes)
Phosphorus	Rickets (osteomalacia), proximal muscle weakness, rhabdomyolysis, paresthesia, ataxia, seizure, confusion, heart failure, hemolysis, acidosis	Hyperphosphatemia	4000 mg/d

Element	Deficiency	Toxicity	Tolerable Upper (Dietary) Intake Level
Zinc	Growth retardation, ↓ taste and smell, alopecia, dermatitis, diarrhea, immune dysfunction, failure to thrive, gonadal atrophy, congenital malformations	*General:* Reduced copper absorption, gastritis, sweating, fever, nausea, vomiting *Occupational:* Respiratory distress, pulmonary fibrosis	40 mg/d (impaired copper metabolism)

CALCIUM

See Chap. 402.

ZINC

Zinc is an integral component of many metalloenzymes in the body; it is involved in the synthesis and stabilization of proteins, DNA, and RNA, and plays a structural role in ribosomes and membranes. Zinc is necessary for the binding of steroid hormone receptors and several other transcription factors to DNA. Zinc is absolutely required for normal spermatogenesis, fetal growth, and embryonic development.

Absorption

The absorption of zinc from the diet is inhibited by dietary phytate, fiber, oxalate, iron, and copper as well as by certain drugs, including penicillamine, sodium valproate, and ethambutol. Meat, shellfish, nuts, and legumes are good sources of bioavailable zinc, whereas zinc in grains and legumes is less available for absorption.

Deficiency

Mild zinc deficiency has been described in many diseases, including diabetes mellitus, HIV/AIDS, cirrhosis, alcoholism, inflammatory bowel disease, malabsorption syndromes, and sickle cell disease. In these diseases, mild chronic zinc deficiency can cause stunted growth in children, decreased taste sensation (*hypogeusia*), and impaired immune function. Severe chronic zinc deficiency has been described as a cause of hypogonadism and dwarfism in several Middle Eastern countries. In these children, hypopigmented hair is also part of the syndrome. Acrodermatitis enteropathica is a rare autosomal recessive disorder characterized by abnormalities in zinc absorption. Clinical manifestations include diarrhea, alopecia, muscle wasting, depression, irritability, and a rash involving the extremities, face, and perineum. The rash is characterized by

vesicular and pustular crusting with scaling and erythema. Occasional patients with Wilson's disease have developed zinc deficiency as a consequence of penicillamine therapy **(Chap. 408)**.

Zinc deficiency is prevalent in many developing countries and usually coexists with other micronutrient deficiencies (especially iron deficiency). Zinc (20 mg/d until recovery) may be an effective adjunctive therapeutic strategy for diarrheal disease and pneumonia in children ≥6 months of age.

The diagnosis of zinc deficiency is usually based on a serum zinc level <12 µmol/L (<70 µg/dL). Pregnancy and birth control pills may cause a slight depression in serum zinc levels, and hypoalbuminemia from any cause can result in hypozincemia. In acute stress situations (illness but also post-exercise recovery), zinc may be redistributed from serum into tissues. Zinc deficiency may be treated with 60 mg of elemental zinc taken by mouth twice a day. Zinc gluconate lozenges (13 mg of elemental zinc every 2 h while awake) have been reported to reduce the duration and symptoms of the common cold in adults, but study results are conflicting.

Toxicity

Acute zinc toxicity after oral ingestion causes nausea, vomiting, and fever. Zinc fumes from welding may also be toxic and cause fever, respiratory distress, excessive salivation, sweating, and headache. Chronic large doses of zinc may depress immune function and cause hypochromic anemia as a result of copper deficiency. Intranasal zinc preparations should be avoided because they may lead to irreversible damage of the nasal mucosa and anosmia.

COPPER

Copper is an integral part of numerous enzyme systems, including amine oxidases, ferroxidase (ceruloplasmin), cytochrome c oxidase, superoxide dismutase, and dopamine hydroxylase. Copper is also a component of ferroprotein, a transport protein involved in the basolateral transfer of iron during absorption from the enterocyte. As such, copper plays a role in iron metabolism, melanin synthesis, energy production, neurotransmitter synthesis, and CNS function; the synthesis and cross-linking of elastin and collagen; and the scavenging of superoxide radicals. Dietary sources of copper include shellfish, liver, nuts, legumes, bran, and organ meats.

Deficiency

Dietary copper deficiency is relatively rare, although it has been described in premature infants who are fed milk diets and in infants with malabsorption (Table 326-2). Copper-deficiency anemia (refractory to therapeutic iron) has been reported in patients with malabsorptive diseases and nephrotic syndrome and in patients treated for Wilson's disease with chronic high doses of oral zinc, which can interfere with copper absorption. *Menkes kinky hair syndrome* is an X-linked metabolic disturbance of copper metabolism characterized by mental retardation, hypocupremia, and decreased circulating ceruloplasmin **(Chap. 406)**. This syndrome is caused by mutations in the copper-transporting *ATP7A* gene. Children with this disease

often die within 5 years because of dissecting aneurysms or cardiac rupture. Aceruloplasminemia is a rare autosomal recessive disease characterized by tissue iron overload, mental deterioration, microcytic anemia, and low serum iron and copper concentrations.

The diagnosis of copper deficiency is usually based on low serum levels of copper (<65 µg/dL) and low ceruloplasmin levels (<20 mg/dL). Serum levels of copper may be elevated in pregnancy or stress conditions since ceruloplasmin is an acute-phase reactant and 90% of circulating copper is bound to ceruloplasmin.

Toxicity

Copper toxicity is usually accidental (Table 326-2). In severe cases, kidney failure, liver failure, and coma may ensue. In Wilson's disease, mutations in the copper-transporting *ATP7B* gene lead to accumulation of copper in the liver and brain, with low blood levels due to decreased ceruloplasmin **(Chap. 408)**.

SELENIUM


Selenium, in the form of selenocysteine, is a component of the enzyme glutathione peroxidase, which serves to protect proteins, cell membranes, lipids, and nucleic acids from oxidant molecules. As such, selenium is being actively studied as a chemopreventive agent against certain cancers, such as prostate cancer. However, it remains unclear whether selenium is effective as a chemopreventive agent or whether it increases cancer risk (e.g., prostate cancer). Selenocysteine is also found in the deiodinase enzymes, which mediate the deiodination of thyroxine to triiodothyronine **(Chap. 375)**. Rich dietary sources of selenium include seafood, muscle meat, and cereals, although the selenium content of cereal is determined by the soil concentration. Countries with low soil concentrations include parts of Scandinavia, China, and New Zealand. *Keshan disease* is an endemic cardiomyopathy found in children and young women residing in regions of China where dietary intake of selenium is low (<20 µg/d). Concomitant deficiencies of iodine and selenium may worsen the clinical manifestations of cretinism. Chronic ingestion of large amounts of selenium leads to selenosis, characterized by hair and nail brittleness and loss, garlic breath odor, skin rash, myopathy, irritability, and other abnormalities of the nervous system.

CHROMIUM

Chromium potentiates the action of insulin in patients with impaired glucose tolerance, presumably by increasing insulin receptor–mediated signaling, although its usefulness in treating type 2 diabetes is uncertain. In addition, improvement in blood lipid profiles has been reported in some patients. The usefulness of chromium supplements in muscle building has not been substantiated. Rich food sources of chromium include yeast, meat, and grain products. Chromium in the trivalent state is found in supplements and is largely nontoxic; however, chromium-6 is a product of stainless steel welding and is a known pulmonary carcinogen as well as a cause of liver, kidney, and CNS damage.

MAGNESIUM

See Chap. 402.

FLUORIDE, MANGANESE, AND ULTRATRACE ELEMENTS

An essential function for fluoride in humans has not been described, although it is useful for the maintenance of structure in teeth and bones. Adult fluorosis results in mottled and pitted defects in tooth enamel as well as brittle bone (skeletal fluorosis).

Manganese and molybdenum deficiencies have been reported in patients with rare genetic abnormalities and in a few patients receiving prolonged total parenteral nutrition. Several manganese-specific enzymes have been identified (e.g., manganese superoxide dismutase). Deficiencies of manganese have been reported to result in bone demineralization, poor growth, ataxia, disturbances in carbohydrate and lipid metabolism, and convulsions.

Ultratrace elements are defined as those needed in amounts <1 mg/d. Essentiality has not been established for most ultratrace elements, although selenium, chromium, and iodine are clearly essential **(Chap. 375)**. Molybdenum is necessary for the activity of sulfite and xanthine oxidase, and molybdenum deficiency may result in skeletal and brain lesions.

FURTHER READING

Boyce SG et al: Can composite nutritional supplement based on the current guidelines prevent vitamin and mineral deficiency after weight loss surgery? Obes Surg 26:966, 2016. [PubMed: 26319661]

Colleen R et al: Revisiting vitamin C and cancer. Science 350:1317, 2015. [PubMed: 26659042]

Day E et al: Thiamine for prevention and treatment of Wernicke-Korsakoff syndrome in people who abuse alcohol. Cochrane Database Syst Rev 7:CD004033, 2013.

Imdad A et al: Vitamin A supplementation for preventing morbidity and mortality in children from six months to five years of age. Cochrane Database Syst Rev 3:CD008524, 2017. [PubMed: 28282701]

Stevens GA et al: Trends and mortality effects of vitamin A deficiency in children in 138 low-income and middle-income countries between 1991 and 2013: A pooled analysis of population-based surveys. Lancet Glob Health 3:e528, 2015. [PubMed: 26275329]

Vinceti M, Rothman KJ: More results but no clear conclusion on selenium and cancer. Am J Clin Nutr 104:245, 2016. [PubMed: 27413135]

WHO: Guideline: Vitamin A supplementation in pregnant women. Geneva, World Health Organization, 2011.

Chapter 327: Malnutrition and Nutritional Assessment

Gordon L. Jensen

INTRODUCTION

Malnutrition occurs in 30–50% of hospitalized patients depending upon the setting and criteria that are used. Poor wound healing, compromised immune status, impaired organ function, increased length of hospital stay, and increased mortality are among the notable adverse outcomes associated with malnutrition. It is now widely appreciated that acute or chronic inflammation contribute to the pathophysiology of disease-related or injury-related malnutrition. The presence of inflammation can also render historic nutrition assessment indicators, like albumin and prealbumin, unreliable and inflammation diminishes favorable responses to nutrition therapies. In order to guide appropriate care, it is necessary to properly assess and diagnose malnutrition. Nutrition assessment is a comprehensive evaluation to diagnose a malnutrition syndrome and to guide intervention and expected outcomes. Patients are often targeted for assessment after being identified at nutritional risk based upon screening procedures conducted by nursing or nutrition personnel within 24 h of hospital admission. Screening tends to focus explicitly on a few risk variables like weight loss, compromised dietary intake, and high risk medical/surgical diagnoses. Preferably, health professionals complement this screening with a systematic approach to comprehensive nutrition assessment that incorporates an appreciation for the contributions of inflammation that serve as the basis for new approaches to the diagnosis and management of malnutrition syndromes.

MALNUTRITION SYNDROMES

 Famine and starvation have long been leading causes of malnutrition and remain so in developing countries. However, with improvements in agriculture, education, public health, healthcare, and living standards, malnutrition in the settings of disease, surgery, and injury has become a prevalent concern throughout the world. Malnutrition now encompasses the full continuum of undernutrition and over-nutrition (obesity). For the objectives of this chapter, we will focus upon the former. Historic definitions for malnutrition syndromes are problematic in their use of diagnostic criteria that suffer poor sensitivity, sensitivity, and inter-observer reliability. Definitions overlap and confusion and misdiagnosis are frequent. In addition, some approaches do not recognize undernutrition in obese persons. While the historic syndromes of marasmus, kwashiorkor, and protein-calorie malnutrition remain in use, this chapter will instead highlight new insights to the diagnosis of malnutrition syndromes.

The Subjective Global Assessment, a comprehensive nutrition assessment that included a metabolic stress of disease component, was described and validated in the 1980s. In 2010, an International Consensus Guideline Committee incorporated a new appreciation for the role of inflammatory response into their proposed nomenclature for nutrition diagnosis in adults in the clinical practice setting. *Starvation-associated malnutrition*, when there is chronic starvation without inflammation (anorexia nervosa or major depression with lack of interest in eating), *chronic disease-associated malnutrition*, when inflammation is chronic and of mild to moderate degree (e.g., organ failure, pancreatic cancer, or sarcopenic obesity), and *acute disease or injury-associated malnutrition*, when inflammation is acute and of severe degree (e.g., major infection, burns, trauma, or closed head injury). In 2012, the Academy of Nutrition and Dietetics and the American Society for Parenteral and Enteral Nutrition (ASPEN) extended this approach with corresponding nomenclature that included malnutrition in the setting of social and environmental circumstances, malnutrition in the setting of chronic illness, and malnutrition in the setting of acute illness or injury. Clinical characteristics were proposed to support a diagnosis that encompasses the presence of illness or injury, poor food intake, weight loss, and physical findings of fat loss, muscle loss, edema, or reduced grip strength. In 2016, the European Society for Parenteral and Enteral Nutrition (ESPEN) formally adopted an inflammation-based construct similar to these earlier approaches. Also in 2016, the Global Leadership Initiative on Malnutrition, a collaborative effort of ASPEN, ESPEN, the Latin American Federation of Parenteral and Enteral Nutrition, the Parenteral and Enteral Society of Asia, and other nutrition societies embarked on an effort to identify evidence-based criteria that will be disseminated throughout the world for use as dictated by regional preference.

Recent studies suggest that these newer approaches to diagnosis of malnutrition have similar utility in predicting adverse outcomes. This is not surprising since they share a number of common criteria including a metabolic stress of disease component that is a proxy indicator of inflammation. Irrespective of the approach that is selected, assessment of patients can be facilitated using the indicators of malnutrition and inflammation described below.

NUTRITION ASSESSMENT

There is unfortunately no single clinical or laboratory indicator of comprehensive nutritional status. Assessment therefore requires systematic integration of data from a variety of sources. Micronutrient deficiencies of clinical relevance may be detected in association with any of the malnutrition syndromes, but a detailed discussion of their assessment is beyond the scope of this chapter **(see Chap. 326)**. Physical findings characteristic of micronutrient deficiencies are however summarized in **Table 327-1**.

TABLE 327-1

History and Physical Examination Elements

Element	Notes
Historical Data	
Body weight	Ask about usual weight, peak weight, and deliberate weight loss. A 4.5 kg (10-lb) weight loss over 6 months is noteworthy and a weight loss of >10% of usual body weight is prognostic of clinical outcomes. Use medical records, family, and caregivers as information resources.
Medical and surgical conditions; chronic disease	Look for medical or surgical conditions or chronic disease that can place one at nutritional risk secondary to increased requirements, or compromised intake or assimilation like: critical illness, severe burns, major abdominal surgery, multi-trauma, closed head injury, previous gastrointestinal surgery, severe gastrointestinal hemorrhage, enterocutaneous fistula, gastrointestinal obstruction, mesenteric ischemia, severe acute pancreatitis, chronic pancreatitis, inflammatory bowel disease, celiac disease, bacterial overgrowth, solid or hematologic malignancy, bone marrow transplant, acquired immune deficiency syndrome, and organ failure/transplant—kidney, liver, heart, lung, or gut. A number of conditions or diseases are characterized by severe acute inflammatory response including critical illness, major infection/sepsis, adult respiratory distress syndrome, systemic inflammatory response syndrome, severe burns, major abdominal surgery, multi-trauma, and closed head injury. Many conditions or diseases are more typically associated with mild to moderate chronic inflammatory response. Examples include cardiovascular disease, congestive heart failure, cystic fibrosis, inflammatory bowel disease, celiac disease, chronic pancreatitis, rheumatoid arthritis, solid tumors, hematologic malignancies, sarcopenic obesity, diabetes mellitus, metabolic syndrome, cerebrovascular accident, neuromuscular disease, dementia, organ failure/transplant—kidney, liver, heart, lung, or gut, periodontal disease, pressure wounds, and chronic obstructive pulmonary disease. Note that acute exacerbations, infections, or other complications may superimpose acute inflammatory response on such conditions or diseases. Examples of starvation-associated conditions that generally have little or no inflammatory component include anorexia nervosa or compromised intake in the setting of major depression.

Element	Notes
Constitutional signs/symptoms	Fever or hypothermia can indicate active inflammatory response. Tachycardia is also common. Anorexia is another manifestation of inflammatory response and is also often a side effect of treatments and medications.
Eating difficulties/gastrointestinal complaints	Poor dentition or problems swallowing can compromise oral intake. Vomiting, nausea, abdominal pain, abdominal distension, diarrhea, constipation, and gastrointestinal bleeding can be signs of gastrointestinal pathology that may place one at nutritional risk.
Eating disorders	Look for distorted body image, compulsive exercise, amenorrhea, vomiting, tooth loss, dental caries, and use of laxatives, diuretics or Ipecac.
Medication use	Many medications can adversely affect nutrient intake or assimilation. Review potential drug–drug and drug–nutrient interactions. A pharmacist consultant can be helpful.
Dietary practices and supplement use	Look for dietary practices including therapeutic, weight reduction, vegetarian, macrobiotic, and fad diets. Also record use of dietary supplements, including vitamins, minerals, and herbals. Ask about dietary intake. Recall, record, and food frequency tools are available. It is estimated that 50% or more of adults take dietary supplements.
Influences on nutritional status	Ask about factors such as living environment, functional status (activities of daily living and instrumental activities of daily living), dependency, caregiver status, resources, dentition, alcohol or substance abuse, mental health (depression or dementia), and lifestyle.
Physical Examination Data	
Body mass index (BMI)	BMI = weight in kg/(height in meters)^2 BMI <18.5 kg/m^2 proposed screen for malnutrition per National Institutes of Health guidelines. BMI ≤15 kg/m^2 or less is associated with increased mortality. Comparison with ideal body weight for stature can also be determined from reference tables. Note hydration status and edema at the time body weight is determined.

Element	Notes
Weight loss	Look for loss of muscle mass and subcutaneous fat. Temporal and neck muscle wasting may be readily observed. Anthropometrics including skin-folds and circumferences can be useful but require training to achieve reliability.
Weakness/loss of strength	Decreased hand-grip and leg extensor strength have been related to loss of muscle mass in malnourished states. Lower extremity weakness may be observed in thiamine deficiency.
Peripheral edema	Peripheral edema may confound weight measurements and is often observed with reduced visceral proteins as well as inflammatory states. Edema may also be observed with thiamine deficiency.
Hair examination	Hair findings are indicative of certain nutrient deficiencies. Loss: protein, B12, folate Brittle: biotin Color change: zinc Dry: vitamins A and E Easy pluckability: protein, biotin, zinc Coiled, corkscrew: vitamins A and C Alopecia is common in severely malnourished persons. Ask about excessive hair loss on pillow or when combing hair.
Skin examination	Skin findings are indicative of certain nutrient deficiencies. Desquammation: riboflavin Petechiae: vitamins A and C Perifollicular hemorrhage: vitamin C Ecchymosis: vitamins C and K Xerosis, bran-like desquamation: essential fatty acid Pigmentation, cracking, crusting: niacin Acneiform lesions, follicular keratosis, xerosis: vitamin A Acro-orificial dermatitis, erythematous, vesiculbullous, and pustular: zinc Characteristic nutritional dermatitis and skin findings may be observed with a number of nutrient deficiencies. Wounds and pressure sores should also be noted as indicators of compromised nutritional status.

Element	Notes
Eye examination	Ocular findings are indicative of certain nutrient deficiencies. Bitot's spots: vitamin A Xerosis: vitamin A Angular palpebritis: riboflavin Also ask about difficulties with night vision/night blindness; indicates vitamin A deficiency.
Perioral examination	Perioral findings are indicative of certain nutrient deficiencies. Angular stomatitis and cheilosis: B complex, iron, and protein Glossitis: niacin, folate, and vitamin B12 Magenta tongue: riboflavin Bleeding gums, gingivitis, tooth loss: vitamin C Angular stomatitis, cheilosis, and glossitis are associated with vitamin and mineral deficiencies. Note poor dentition, caries, and tooth loss. Difficulty swallowing and impairment of gag should also be recognized.
Extremity examination	Extremity findings indicate certain nutrient deficiencies Arthralgia: vitamin C Calf pain: thiamine Extremities may also exhibit loss of muscle mass and/or peripheral edema. Neurological findings in the extremities may also result from deficiencies described below.
Mental status/nervous system examination	Mental and nervous system findings indicate certain nutrient deficiencies. Ophthalmoplegia and foot drop: thiamine Paresthesia: thiamine, vitamin B12, and biotin Depressed vibratory and position senses: vitamin B12 Anxiety, depression, and hallucinations: niacin Memory disturbance: vitamin B12 Hyporeflexia, loss of lower extremity deep tendon reflexes: thiamine and vitamin B12 Conduct formal cognitive and depression assessments as appropriate. Dementia and depression are common causes of malnutrition among older persons. Wernicke-Korsakoff syndrome may be observed with severe thiamine deficiency.

Element	Notes
Functional assessment	Observe and test physical performance as indicated: gait, chair stands, stair steps, and balance. These provide complex measures of integrated neurological status, coordination, and strength.

Source: Adapted with permission from G Jensen: *Nutritional Syndromes. Smart Medicine/PIER.* Philadelphia, American College of Physicians, 2013.

Medical/Surgical History and Clinical Diagnosis

Knowledge of a patient's medical/surgical history and associated clinical diagnoses is especially helpful in discerning the likelihood of malnutrition and inflammation. Non-volitional weight loss is a well validated nutrition assessment indicator and is often also associated with underlying disease or inflammatory condition. The degree and duration of weight loss determine its clinical significance. A 10% loss of body weight over 6-months is of clinical relevance, while a 30% loss of body weight over the same duration is severe and life-threatening. Since weight loss history is often unavailable or unreliable, one should query the patient as well as the medical records, family, and caregivers as appropriate to secure a valid weight trajectory.

A number of conditions or diseases are characterized by severe acute inflammatory response whereas others are more typically associated with a chronic inflammatory response that is mild to moderate in severity and may be relapsing and remitting (Table 327-1). It is also common for acute inflammatory events to be superimposed on those with chronic conditions; for example, a patient with chronic renal disease is admitted to the hospital with sepsis. The inflammatory milieu, especially when severe, may modify nutrient requirements by elevating resting energy expenditure and promoting muscle catabolism and nitrogen losses. Inflammation also promotes anorexia, decreasing food intake and further compromising nutritional status. A deteriorating course may result because the presence of inflammation may reduce the benefit of nutritional interventions and the associated malnutrition may in turn diminish the effectiveness of medical therapies. It is also imperative to recognize medical/surgical conditions or diseases that place the patient at increased risk to become malnourished because they have increased nutritional requirements, or compromised intake or assimilation (Table 327-1).

Nutrition assessment should also include a review of medications with attention to undesirable side effects including anorexia, xerostomia, nausea, diarrhea, and constipation. Potential drug/nutrient interactions should also be identified.

Clinical Signs and Physical Examination

Nonspecific clinical indicators of inflammation include fever, hypothermia, and tachycardia. The nutrition-focused physical examination should identify edema as well as signs of weight gain/loss and specific nutrient deficiencies. Thorough examination should be particularly directed to those parts of the body where high cell turnover occurs (e.g., hair, skin, mouth, tongue) as they are most likely to exhibit observable signs of nutritional deficiencies (Table 327-1). Physical findings of weight loss associated with decreased muscle and subcutaneous fat mass should not be overlooked, but when appreciable edema is present, these changes may not be readily appreciated.

Anthropometric Data

Body weight measurements are recommended with each clinic visit or hospitalization so that a reliable weight change trajectory may be monitored. Patients should be weighed in a consistent manner without over-garments or shoes. In order to secure valid measurements, calibration of scales and appropriate staff training are essential. Chair or bed scales may be used for those who cannot stand. For those who are able, height should be measured in a standing position without shoes using a stadiometer. If an adult cannot safely stand, height can be estimated by doubling the arm span measurement (from the patient's sternal notch to the end of the longest finger). Stature of frail older persons can also be estimated from measurement of knee height using a caliper device.

Body weight is often standardized for height to obtain an ideal weight for comparison, but available reference tables require subjective assessment of frame size and offer limited reference data for many relevant population groups, including older persons. A simple measure of body size and an indirect measure of body fatness is provided by body mass index (BMI), defined as weight (kg) / height (m^2). The National Institutes of Health BMI categories for adults are: BMI <18.5 = underweight, BMI 18.5–24.9 = desirable, BMI 25.0–29.9 = overweight, and BMI ≥30 = obese. A higher desirable BMI range for persons 65 years of age and older has been proposed by the Centers for Medicare and Medicaid Services in its quality indicators system: BMI ≥23 and <30.

While classical anthropometric measurements including skin-folds and circumferences can be helpful, their utility in routine patient care has been limited because practitioner training is required to achieve suitable reliability. Body composition assessment methodologies include bioelectrical impedance analysis (BIA), dual-energy x-ray absorptiometry (DEXA), computed tomography (CT), and magnetic resonance imaging (MRI). The imaging modalities have become the state of the art for precise measurements of muscle mass. It is possible to take advantage of CT or MRI studies that are being done for other clinical purposes to evaluate musculature.

Laboratory Indicators

Laboratory findings **(Table 327-2)** are but one part of the comprehensive nutrition assessment and must be used in combination with other domains of assessment to appropriately diagnose a malnutrition syndrome. Although serum albumin or prealbumin are often measured in patients with suspected malnutrition, their utility is limited due to their poor sensitivity and specificity as indicators of nutritional status. Patients with

low albumin or prealbumin may or may not prove to be malnourished when evaluated by comprehensive nutrition assessment because these proteins are readily reduced by the systemic response to injury, disease or inflammation. C-reactive protein is a positive acute phase reactant that may be measured to help discern whether active inflammation is manifest. If C-reactive protein is increased, and albumin or prealbumin decreased, then inflammation is likely to be a contributing factor. Since it is recognized that C-reactive protein suffers limitations as a point in time measure, trends in levels over the clinical course may be helpful. Research suggests that interleukin 6, and perhaps other cytokines, may also offer promise as indicators of inflammatory status. Nonspecific laboratory indicators that are often associated with inflammatory response include leukocytosis and hyperglycemia. Additional tests that may be obtained to help confirm the presence of inflammatory response include 24-h urine urea nitrogen and indirect calorimetry. In the setting of severe acute systemic inflammatory response, negative nitrogen balance and elevated resting energy expenditure are anticipated.

TABLE 327-2

Body Composition, Laboratories, and Other Studies

Test	Notes
Body Composition Studies	
Anthropometrics	Skin folds and circumferences require training for reliability. Typical coefficient of variation is ≥10%.
Bioelectrical impedance	Based upon differential resistance of body tissues. Equipment easily portable. Good measure of body water. Requires population specific validation of regression equations.
Water displacement	Impractical for most clinical settings. Weighed in water tank. Historic reference measure.
Whole body counting and isotope dilution techniques	Research methodologies. Naturally occurring ^{40}K isotope to measure body cell mass by whole body counting. Total body water measurement by dilution volume of tritium, deuterium, or ^{18}O-labeled water.
Air plethysmography	Research methodology. Subject sits inside moderately sized BodPod chamber. Validated against water displacement and impedance.
Dual energy x-ray absorptiometry (DEXA)	Often used for bone density but can be used for soft tissue measurements with appropriate software. Can compare truncal and appendicular components. Modest x-ray exposure.
Imaging with computed tomography (CT) or magnetic resonance imaging (MRI)	State of the art research methods for visualizing body tissue compartments. Can quantify visceral fat. Costly and CT entails X-ray exposure.
Laboratories and Other Studies	

Test	Notes
Albumin	Lacks sensitivity and specificity for malnutrition. Potent risk indicator for morbidity and mortality. Proxy measure for underlying injury, disease or inflammation. Half-life is 14–20 days. Also consider liver disease, nephrotic syndrome, and protein-wasting enteropathy.
Prealbumin	Sensitive to short-term changes in inflammation and protein nutrition with half-life of 2–3 days. Otherwise suffers the same limitations of albumin with limited sensitivity and specificity for malnutrition. Levels may be decreased in liver failure and increased in renal failure.
Transferrin	Acute phase reactant also altered by perturbation in iron status. Half-life is 8–10 days. Lacks sensitivity and specificity for malnutrition.
Retinol-binding protein	Responds to very short-term changes in nutritional status but utility is also limited by response to stress and inflammation. Half-life is 12 h. Also affected by vitamin A deficiency and renal disease.
C-reactive protein	C-reactive protein is a positive acute phase reactant. It is generally elevated if an active inflammatory process is manifest.
Cholesterol	Low cholesterol (<160 mg/dL) is often observed in malnourished persons with serious underlying disease. It is unrelated to dietary intake in many clinical settings. Increased complications and mortality are observed. It appears that low cholesterol is again a nonspecific feature of poor health status that reflects cytokine-mediated inflammatory condition. Vegans and patients with hyperthyroidism may also exhibit low cholesterol.
Carotene	Nonspecific indicator of malabsorption and poor nutritional intake.
Cytokines	Research is exploring prognostic use of cytokine measurements as indicators of inflammatory status.
Electrolytes, blood urea nitrogen, creatinine, and glucose	Monitor for abnormalities consistent with under- or over-hydration status and purging (contraction alkalosis). BUN may also be low in the setting of markedly reduced body cell mass. Blood urea nitrogen and creatinine are elevated in renal failure. Hyperglycemia may be nonspecific indicator of inflammatory response.

Test	Notes
Complete blood count with differential	Screen for nutritional anemias (iron, B12, and folate), lymphopenia (malnutrition) and thrombocytopenia (vitamin C and folate). Leukocytosis may be observed with inflammatory response.
Total lymphocyte count	Relative lymphopenia (total lymphocyte count <1200/mm^3) is a nonspecific marker for malnutrition.
Helper/suppressor T cell ratio	Ratio may be reduced in severely undernourished patients. Not specific for nutritional status.
Nitrogen balance	24-h urine can be analyzed for urine urea nitrogen (UUN) to determine nitrogen balance and give indication of degree of catabolism and adequacy of protein replacement. Requires accurate urine collection and normal renal function. Nitrogen balance = (protein/6.25) − (UUN+4). Generally negative in the setting of acute severe inflammatory response.
Urine 3-methylhistidine	Indicator of muscle catabolism and protein sufficiency. Released upon breakdown of myofibrillar protein and excreted without reutilization. Urine measurement requires a meat-free diet for 3 days prior to collection.
Creatinine height index (CHI)	CHI = (24-h urinary creatinine excretion/ideal urinary creatinine for gender and height) × 100. Indicator of muscle depletion. Requires accurate urine collection and normal renal function.
Prothrombin time/international normalized ratio (INR)	Nonspecific indicator of vitamin K status. Prolonged in liver failure.
Specific micronutrients	When suspected a variety of specific micronutrient levels may be measured: thiamine, riboflavin, niacin, folate, pyridoxine, vitamins A, C, D, E, B12, zinc, iron, selenium, carnitine, and homocysteine—indicator of B12, folate, and pyridoxine status.
Skin testing—recall antigens	Delayed hypersensitivity testing. While malnourished patients are often anergic, this is not specific for nutritional status.
Electrocardiogram	Severely malnourished patients with reduced body cell mass may exhibit low voltage and prolonged QT interval. These findings are not specific for malnutrition.

Test	Notes
Video fluoroscopy	Helpful to evaluate suspected swallowing disorders.
Endoscopic and x-ray studies of gastrointestinal tract	Useful to evaluate impaired function, motility, and obstruction.
Fat absorption	72-h fecal fat can be used to quantitate degree of malabsorption.
Schilling test	Identify the cause for impaired vitamin B12 absorption.
Indirect calorimetry	Metabolic cart can be used to determine resting energy expenditure (REE) for accurate estimation of energy needs. Elevated REE is a sign of systemic inflammatory response.

Source: Adapted with permission from G Jensen: *Nutritional Syndromes. Smart Medicine/PIER*. Philadelphia, American College of Physicians, 2013.

Dietary Assessment

Dietary assessment can be used to detect inadequate or imbalanced food or nutrient intakes. While dietary assessment in patient care settings can be quite challenging, 24-h recall and modified diet history approaches are sometimes used. A modified diet history is targeted to query types and frequencies of intake of specific foods of interest. It is often necessary to access diverse resources for diet history information including the patient, medical records, family, and caregivers. Consultation of a registered dietitian nutritionist is highly recommended. Dietary practices and supplements should be carefully reviewed for potential inadequacies and toxicities. Since patients will often present to healthcare practitioners with acute medical events superimposed upon chronic health conditions, it is common for patients to have had decreased food intakes and malnutrition for extended periods prior to assessment. It is therefore imperative that compromised dietary intake should not be overlooked so that appropriate intervention may be undertaken.

Ongoing assessment is indicated when parenteral or enteral feedings are initiated, because it is necessary to discern what amount of formula is actually being administered to and received by the patient. Enteral feedings, in particular, often interrupted or held for procedures, tolerance issues, and feeding tube displacements. It is therefore not unusual for such patients to be appreciably underfed for extended periods. When a patient is beginning to transition to oral feedings, it is imperative to monitor quantities of food and/or supplements that are actually consumed as well as patient tolerance to feeding. Meals are often

delayed or missed for tests or procedures. If possible, the patient should be queried about intake since tray inspection is notoriously unreliable as an indicator of consumption.

Functional Outcomes

Advanced malnutrition is accompanied by declines in muscle mass and function that can be detected by strength and physical performance measures. Hand-grip strength measured with a simple handgrip dynamometer is the most practical routine clinical assessment. Physical performance tests like timed gait, chair stands, and stair steps are used in the comprehensive assessment of integrated functions in frail older persons.

The decline in overall functional status observed in advanced malnutrition is associated with nutrient deficiencies and impairment of organ system functions. Poor wound healing and immune compromise are examples of such impairments. Improved wound healing parameters and restored responsiveness to recall antigens by delayed hypersensitivity testing may be measured to demonstrate improvements with nutritional repletion, though it must be appreciated that these are multivariable outcomes for which improved nutritional status is but one variable.

FURTHER READING

Cederholm T et al: ESPEN guidelines on definitions and terminology of clinical nutrition. Clin Nutr 36:49, 2017.

Cederholm T, Jensen GL: To create a consensus on malnutrition diagnostic criteria. JPEN J Parenter Enteral Nutr 41:311, 2017.

Detsky AS et al: What is Subjective Global Assessment of nutritional status? J Parenter Enteral Nutr 11:8, 1987.

Guerra RS et al: Usefulness of six diagnostic and screening measures for undernutrition in predicting length of hospital stay: A comparative analysis. J Acad Nutr Diet 115:927, 2015. [PubMed: 25634094]

Jensen GL: Inflammation as the key interface of the medical and nutrition universes: A provocative examination of the future of clinical nutrition and medicine. J Parenter Enteral Nutr 30:453, 2006.

Jensen GL: Malnutrition and inflammation—“Burning down the house”: Inflammation as an adaptive physiologic response versus self-destruction? J Parenter Enteral Nutr 39:56, 2015.

Jensen GL: Global leadership conversation: Addressing malnutrition. J Parenter Enteral Nutr 40:455, 2016.

Jensen GL, Hsiao PY, Wheeler D: Adult nutrition assessment tutorial. J Parenter Enteral Nutr 36:267, 2012.

Jensen GL et al: Adult starvation and disease-related malnutrition: A proposal for etiology-based diagnosis in the clinical practice setting from the International Consensus Guideline Committee. J Parenter Enteral Nutr 34:156, 2010.

White JV et al: Consensus statement: Academy of Nutrition and Dietetics and American Society for Parenteral and Enteral Nutrition. Characteristics recommended for the identification and documentation of adult malnutrition (under-nutrition). J Parenter Enteral Nutr 36:275, 2012.

Chapter 328: Enteral and Parenteral Nutrition

L. John Hoffer; Bruce R. Bistrian; David F. Driscoll

INTRODUCTION

There are three kinds of specialized nutritional support (SNS): (1) optimized voluntary nutritional support, which is used when a patient's barriers to adequate nutrition can be overcome by special attention to the details of how their food is constituted, prepared, served, and its consumption monitored, (2) forced enteral nutrition (EN), in which a liquid nutrient formula is delivered through a tube placed in the stomach or small intestine, and (3) parenteral nutrition (PN), in which a nutritionally complete mixture of crystalline amino acids, dextrose, triglyceride emulsions, minerals, electrolytes, and micronutrients is infused directly into the bloodstream.

When does a hospitalized patient need SNS? When SNS is indicated, how should it be provided? This chapter reviews the physiological principles that underlie the correct use of SNS, and provides practical information about the diagnosis and management of nutritional disorders in adult hospitalized patients.

The management of in-hospital nutritional disorders follows 3 steps: (1) screening and diagnosis; (2) determination of the severity and urgency of treating a diagnosed nutritional disorder in its overall clinical context; and (3) selection of the modality of SNS, its composition, and the details of providing it. To follow these steps effectively, physicians require a general understanding of nutritional physiology, nutrient requirements, the pathophysiology and diagnosis of the nutritional disorders, and familiarity with the indications, advantages, risks, and administration of the different kinds of SNS. Because most physicians are incompletely trained in clinical nutrition, they must collaborate with clinical dietitians and specialized pharmacists when ordering EN and PN.

NUTRITIONAL PHYSIOLOGY

(See Chaps. 325, 326, 327)

Energy

Total energy expenditure (TEE) is comprised of resting energy expenditure (REE, ~24 kcal/kg normal adult body weight/day), activity energy expenditure (~12 kcal/kg in healthy sedentary individuals) and the thermic effect of food (10% of TEE). The TEE of a healthy adult is ~36 kcal/kg. REE can be measured by indirect calorimetry or estimated using a variety of predictive equations that input weight, height, age, sex, and

sometimes disease-related factors. Fever and some forms of critical illness increase REE. Prolonged semi-starvation normally triggers an adaptive reduction in REE, voluntary physical activity, and the thermic effect of food. Broadly speaking, a patient's TEE identifies the amount of dietary energy they have to consume and metabolize to maintain their existing store of body fat (and protein). The amount of energy they actually *require* may be less than TEE (as in obesity therapy) or greater than TEE (when rehabilitating nutritionally depleted patients).

Protein and Amino Acids

Protein is an essential nutrient because whole body protein turnover—a continuous process of protein synthesis and breakdown to its constituent amino acids—is associated with obligatory amino acid catabolism to carbon dioxide, water, carbohydrates, ammonium, urea, and sulfuric acid. Amino acid catabolism can be adaptively reduced when protein intake decreases, but not below a certain lowest rate known as the protein minimum. After adjustment for the inefficiency of exogenous protein retention, this rate defines the minimum dietary protein intake necessary to maintain whole body protein homeostasis and zero nitrogen (N) balance. Proteins of nutritional interest, including the ones that form the metabolically active tissues of the body known as the body cell mass (BCM), are 16% nitrogen (N) by weight. Skeletal muscle makes up ~80% of the BCM and 40–50% of the body weight of normal-weight young adults. The excretion of 1 g N from the body—mostly as urinary urea N—implies the loss of 6.25 g formed protein and (since the BCM is ~80% tissue water) ~31 g of BCM, almost of all it as muscle tissue.

The lowest daily protein intake compatible with zero N balance—approximately 0.65 g/kg body weight—is the average minimum protein requirement of a healthy adult. To account for inter-individual variability, two standard deviations are added to calculate the "safe" or "recommended" daily minimum protein intake of 0.80 g/kg normal body weight. Protein intakes greater than the minimum requirement are usually inconsequential, as surfeit amino acids are normally easily catabolized. The average protein consumption in wealthy societies is approximately twice the average minimum requirement.

Many diseases (or their treatment) increase the dietary protein requirement, by (1) increasing amino acid loss from the body (as in malabsorption and protein loss via wound exudates, fistulas or inflammatory diarrhea), removing amino acids from the circulation (renal replacement therapy), or (2) increasing muscle protein catabolism (as with high-dose glucocorticoid therapy). Protein-catabolic critically ill patients often excrete 15 g N/day or more in their urine in the absence of dietary protein provision. This rate of N loss implies a loss of 15 × 6.25 = 94 g muscle protein—equivalent to one pound of muscle lost from the body every day. Sufficiently generous protein provision can minimize this kind of muscle atrophy, and there is widespread agreement that protein-catabolic patients require much more dietary protein requirement than healthy people. The exact magnitude of the increase has not yet been determined, but the most frequent recommendation for patients with protein-catabolic diseases is 1.5 g protein/kg normal body weight/day, close to the habitual protein intake of healthy people in wealthy societies.

Protein-Energy Interaction

Energy deficiency—whether deliberate, as in weight reduction therapy, or inadvertent, as frequently occurs in hospitalized patients—increases amino acid loss from muscle and increases the dietary protein requirement. The mechanism responsible for energy deficiency's protein-wasting effect differs from the one that mediates inflammation-induced muscle atrophy, and the interaction between these different processes is imperfectly understood. It does appear, however, that systemic inflammation diminishes, but doesn't prevent the protein-sparing effect of generous protein provision as long as the protein is combined with some exogenous energy (e.g., 50% of TEE). Energy provision more generous than ~50–70% of TEE exerts little further protein-sparing effect in this situation, and the additional amounts of glucose and fluid volume required to provide it often have adverse effects.

Micronutrients

Minimum amounts of nine water-soluble vitamins (the B vitamins and vitamin C) and four fat-soluble vitamins (A, D, E, and K), eight minerals (calcium, phosphate, potassium, sodium, chloride, magnesium, zinc, and iron), essential fatty acids, and several essential trace elements (notably including selenium, copper, and iodine) are required throughout life to avoid deficiency diseases and death. Patients who have been hospitalized for more than a few days commonly consume inadequate amounts of food and the micronutrients it provides.

Overt deficiencies of potassium, sodium, magnesium, phosphate, and iron occur so often in hospitalized patients that it is standard practice to monitor for and correct them. Many drugs used in acute-care medicine induce renal potassium, magnesium, or zinc wasting that necessitate appropriate increases in their provision. Gastrointestinal losses from nasogastric drainage tubes or intestinal losses from fistulas or diarrhea incur losses of potassium, sodium, calcium, magnesium, and zinc which add to their normal daily requirement.

Less studied, but nonetheless common, are subclinical or unrecognized deficiencies of calcium, zinc, vitamin D, vitamin C, and possibly other micronutrients. Physicians often assume that consumption of the regular hospital diet will protect patients from these deficiencies. This assumption is not warranted when the patient's nutritional status was borderline or deficient when they were admitted to hospital and remains inadequate throughout their hospital stay. These patients are at risk of a variety of micronutrient deficiency diseases in addition to the symptoms and disability created by continuing in-hospital starvation.

PROTEIN-ENERGY MALNUTRITION AND ITS VARIANTS

The decision to embark on SNS should be justified by a well-formulated nutritional diagnosis and explicitly defined therapeutic goal. This chapter focuses on the diagnosis, treatment, and prevention of in-hospital *protein-energy malnutrition* (PEM). PEM is the disease caused by prolonged inadequate energy and protein consumption—starvation—with consequent depletion of the BCM and body fat. The pathologic features that differentiate PEM—which is a disease—from the physiological process of starvation that leads to it, emerge when the BCM has become depleted seriously enough to impair specific physiological functions. There are

many synonyms for simple, starvation-induced PEM: starvation disease, hunger disease, inanition and, most recently, starvation-related malnutrition (SM).

The body normally adapts to starvation by reducing energy expenditure and curtailing protein catabolism, partly by hormone- and nervous system-regulated alterations in cellular metabolism, and partly by reducing its muscle mass. These adaptations enable prolonged survival during sub-lethal starvation, but survival comes at a cost that includes lethargy, a tendency to hypothermia, muscle atrophy (including of the cardiac and respiratory muscles), skin thinning, and functional disability. The cardinal diagnostic features of PEM—generalized muscle atrophy and subcutaneous adipose tissue depletion—are easy to detect by simple physical examination. Too often, however, in-hospital PEM remains undiagnosed, partly because of health care worker unawareness and inattention, and partly because PEM overlaps and is often confused with other common conditions that also cause muscle atrophy.

Terminology

The terms used to describe nutritional disorders are often ambiguous and confusing. “Malnutrition” is a blanket term that indiscriminately refers to the sum total of the nutritional environments that give rise to starvation, the physiological process of starvation, and the PEM that may result from it. Until this terminology is better standardized, we suggest that health care workers explicitly distinguish among (1) situations that create a risk of inadequate nutrient intake, (2) situations in which inadequate intake of specific nutrients actually occurs and creates a discernable risk of developing a specific nutritional disease, and (3) the specific diseases themselves, as enumerated below.

Starvation-Related Malnutrition (Uncomplicated Protein-Energy Malnutrition)

PEM is diagnosed when the physical examination reveals generalized muscle atrophy and diminished (but not necessarily depleted) subcutaneous adipose tissue due to prolonged inadequate food consumption or malabsorption. PEM is always associated with weight loss, but weight loss alone may not reveal its full severity, because extracellular fluid (ECF) volume and hence body weight increase, sometimes seriously enough to cause edema. SM is “uncomplicated” PEM due solely to prolonged starvation. Adaptation is an important feature of SM that increases the likelihood of survival by reducing energy expenditure and slowing body protein turnover, thus reducing or halting the loss of body protein and fat. The risk of complications, including death, increases as the depletion of BCM worsens. A 50% depletion of BCM puts otherwise uncompromised young adults at the cusp of survival; older patients with co-morbidities are at even greater risk. People with SM feel unwell, lack strength, are frail, and are at risk of hypothermia. The severity of SM is revealed by physical examination and an evaluation of the patient’s strength and physical function.

The main cause of SM worldwide is involuntary food deprivation. The causes of SM in hospitalized patients are many. They include inadvertent or physician-ordered food deprivation, psychologic depression or distress, poorly controlled pain or nausea, badly presented unappealing food, communication barriers, anorexia nervosa, physical or sensory disability (including dysgeusia), thrush, dysphagia and other

mechanical difficulties ingesting food, partial obstruction of the esophagus, stomach or intestinal tract, intestinal angina, and very commonly, combinations of these causes.

Chronic Disease-Related Malnutrition and Cachexia

These two terms refer to starvation-induced PEM that is complicated by chronic systemic inflammation. Chronic disease-related malnutrition (CDM) is highly prevalent among patients with chronic infections, inflammatory autoimmune diseases, chronic severe hepatic, renal, cardiac and pulmonary disease, and certain inflammatory cancers. CDM both causes and is worsened by anorexia: a strong disinclination to eat much food even when there is no physical barrier to eating. CDM is characterized by a moderately increased rate of muscle protein catabolism, muscle atrophy and weakness, fatigue and reduced voluntary activity, and a subverted adaption to starvation, all of which contribute to a vicious cycle of worsening CDM. Fortunately, the nutrient deficit on the input side (anorexia-driven inadequate food consumption) is often a stronger driver of the patient's CDM than increased nutrient loss on the output side (increased amino acid catabolism and REE), thus making CDM amenable to nutritional intervention while effective treatment of the primary disease is underway. The anorexia of CDM is less inhibiting to people who have a hearty pre-morbid appetite and obesity-prone phenotype than it is to previously thin, habitual under-eaters.

Acute Disease-Related Malnutrition

The term acute disease-related malnutrition (ADM) refers to a specific metabolic-nutritional environment that creates a very high risk of severe PEM, rather than an existing disease entity. A synonym for ADM is "catabolic critical illness." The intense systemic inflammation that accompanies severe tissue injury and sepsis causes extremely rapid and severe muscle atrophy (and increases REE to a variable extent) in a setting in which voluntary food intake is impossible. People with ADM are usually treated in intensive care units. They may or may not have PEM at the onset of ADM, but it will surely develop within days to a few weeks unless their medical or surgical disease is rapidly and effectively treated and SNS is appropriately provided.

NUTRITIONAL DIAGNOSIS

The cardinal anatomic features of PEM—generalized muscle atrophy and diminished body fat—are revealed by a discerning physical examination, but what ought to be an easy diagnosis is commonly missed. This section explains the details and pitfalls of diagnosing PEM and judging its severity.

Muscle Mass

Once the examiner's attention has been drawn to it, generalized muscle atrophy is easy to identify and its severity determinable almost at a glance. The problem with diagnosing PEM in the hospital setting, apart from simple inattention, is that generalized muscle atrophy has many causes: old age-related muscle atrophy (sarcopenia), disuse muscle atrophy due to reduced mobility and bed rest, high-dose glucocorticoid therapy, certain endocrine diseases, and primary muscle or neuromuscular diseases. Indeed, muscle atrophy is so common among hospitalized patients that health care workers often mistakenly regard it is a usual, even

defining, characteristic of the patient's primary disease. In reality, muscle atrophy is very commonly at least partly due to SM or CDM. As such, it represents a potentially remediable *complication* of the patient's primary disease rather than a necessary *feature* of it. Whenever a health care worker detects generalized muscle atrophy, they should review the patient's overall situation and identify which of its potential causes are pertinent, and in particular, which of them are treatable and reversible. The combination of old age, disuse muscle atrophy, and starvation is very common. Old age is irreversible, but adequate protein and energy provision combined with physical rehabilitation can be lifesaving.

Muscle atrophy, no matter what its cause, is especially dangerous in ADM, because patients in this situation are closer to the cliff-edge of lethal BCM depletion. As well, their reduced muscle protein mass is unable to release amino acids into the circulation at a rate sufficient to meet the need for protein synthesis at sites of injury and healing, and within the central protein pool to regulate the immuno-inflammatory process.

Subcutaneous Adipose Tissue

Severe adipose tissue depletion is sufficient to diagnose PEM, but it is not a *necessary* criterion. The modern obesity epidemic has created a population of obese patients with chronic inflammation and starvation whose muscle atrophy outpaces their fat loss. A targeted physical examination easily differentiates these patients' muscle groups from their subcutaneous fat, and reveals their moderate and sometimes severe PEM.

ECF Volume

The ECF volume normally represents ~20% of body weight. Chronic starvation increases the ECF volume, occasionally enough to cause dependent edema ("starvation edema"). Hospitalized patients with CDM commonly have other edema-causing conditions, including hypoalbuminemia with its associated reduction in plasma oncotic pressure. Unless recognized and accounted for, increased ECF volume can mask the true extent of muscle and adipose tissue depletion in patients with PEM.

Body Mass Index

Body mass index (BMI) is defined as body weight (kg) divided by the square of height (m^2). Normal BMI ranges from 20 to 25 kg/m^2. BMI >25 usually indicates increased body fat; BMI <20 indicates decreased muscle mass and body fat. Survival during prolonged, severe starvation depends both on fat and protein stores. A BMI of 11–13 is usually incompatible with life. Some guidelines and clinical trial enrollment criteria define "malnutrition"—in this context a synonym for PEM—as a BMI <16 or 17. This oversimplification can lead to serious error. A BMI of 17 certainly is consistent with PEM, because the body architecture of such a BMI can only be created by jettisoning a large amount of BCM and adipose tissue. But a BMI >17 does not rule out PEM. Many patients with PEM have normal or above-normal BMIs due to residual obesity or an expanded ECF volume.

Visual BMI

After some practice, health care workers can accurately predict the BMI of non-obese, non-edematous patients by examining their muscular architecture. Visual BMI is useful for quantifying and communicating the severity of a patient's PEM. Once acquired, this skill can be used to estimate the severity of PEM in obese or edematous patients—in whom measured BMI is unreliable—by focusing attention on their muscular architecture while simultaneously discounting their subcutaneous fat and edema. Visual BMI may also be used to estimate a patient's "normalized dry body weight." The normalized dry body weight of a 1.75 m tall patient with visual BMI 17 = $1.75^2 \times 17 = 52$ kg. Since protein and energy targets are based on normal body weight, this calculation is useful in situations in which actual body weight is unreliable or difficult to measure.

Laboratory and Technical Assessment

Laboratory measurements have three main purposes in the evaluation and management of PEM.

MUSCLE MASS

Bedside ultrasound is a potentially valuable technique for quantifying muscle mass at specific body sites, but it need not, nor should it, replace the immediate and comprehensive evaluation provided by the eyes, hands and mind of the discerning bedside examiner.

SYSTEMIC INFLAMMATION

The presence or absence of systemic inflammation distinguishes SM from CDM/cachexia. The most useful laboratory indicators of systemic inflammation are a reduced serum albumin concentration and an increased serum C-reactive protein concentration. Systemic inflammation increases the permeability of capillary walls to large molecules; the resulting osmotic shift increases the ECF volume. Intravascular albumin pool redistributes into this large volume, decreasing the serum albumin concentration. Increased albumin catabolism likely also contributes. Muscle atrophy and dietary protein deficiency perpetuate inflammation-induced hypoalbuminemia, because muscle protein and the diet provide the amino acids required for hepatic albumin synthesis.

Hypoalbuminemia is often claimed to indicate or diagnose "malnutrition." This is incorrect. Hypoalbuminemia indicates the presence of systemic inflammation, which, by inducing anorexia and increasing muscle catabolism, creates a high risk that PEM could develop; but PEM may or may not exist at the present moment. Hypoalbuminemia will not improve as long as systemic inflammation persists, even with prolonged optimal nutritional therapy. After systemic inflammation has subsided, several weeks of optimal nutrition may be required for serum albumin concentrations to renormalize.

PROTEIN-CATABOLIC INTENSITY

The defining feature of protein-catabolic disease (as occurs most intensely in ADM, but also in CDM) is increased muscle amino acid catabolism. Conditions that increase body protein loss can be identified by

measuring the rate of body N loss. Most N leaves the body in the urine (almost all of it in urea, ammonium, and creatinine), the feces, skin, and by other minor routes. Total N is not usually measured in hospital laboratories, but urinary urea concentrations are routinely available. Urea normally accounts for ~85% of urinary N. Formulas are available that estimate that total N loss solely from 24-h urinary urea excretion. A recent, validated formula estimates daily total N loss (g) = g N in urinary urea/0.85 + 2.

Net muscle protein catabolism follows approximately first-order ("decay") kinetics, such that the rate of N loss from muscle is proportional to the existing total amount of N available to be lost. Muscle atrophic, protein-catabolic patients lose less body N/day in absolute terms than an equivalently catabolic patients with normal muscle mass, but they are at nevertheless at greater risk of succumbing to their critical illness. The interpretation of a patient's rate of N loss should be tempered by a consideration of their existing muscle mass.

Instrumental Nutritional Assessment

Many nutritional assessment instruments claim to identify "malnutrition" by enumerating and summing a list of risk factors, laboratory results, and physical findings. These tools are often hindered by ambiguity about the intended meaning of "malnutrition" and failure to distinguish between screening and diagnosis. Diagnosis is the process of identifying a known pathological entity in a particular patient—SM or CDM, for example—by considering the patient's medical history, pertinent findings on physical examination, and laboratory or imaging reports. Diagnosis also involves an estimation of the probability that the diagnosis is correct and a judgment about its severity. By contrast, screening is the application of a test that identifies people at sufficiently high risk of a certain disease to warrant carrying out definitive procedures to establish the diagnosis or rule it out, or which identifies people at sufficient risk of developing the disease to warrant specific preventive interventions. Screening tools and risk predictors are useful, but it's a mistake to confuse them with clinical diagnosis.

Subjective Global Assessment

The best-validated and most useful formal bedside instrument for diagnosing SM and CDM (and judging their severity) is subjective global assessment. With this method the examiner reflects on the totality of (1) the patient's history (for evidence of inadequate food intake, weight loss, and the presence of factors, such as gastrointestinal disease, and systemic inflammation, that strongly predict diminished ability consume enough food), (2) the patient's current body composition (muscle mass, subcutaneous fat, and ECF volume), and (3) their functional status (strength and mobility), then takes a moment to form an intuitive judgment as to whether the patient has (A) no SM or CDM, (B) is in the gray zone of possible or mild SM or CDM, or (C) definitely has SM or CDM. A judgment is also reached as to how urgently nutritional intervention is required.

SPECIALIZED NUTRITION SUPPORT

Optimized Voluntary Nutritional Support

When feasible, this is the approach of choice because it engages and empowers the patient, encourages mobilization and reconditioning, is consistent with the objectives of patient-centered medicine, and is risk free. Its disadvantage is that it is time-consuming and labor-intensive, and demands interest in and attention to the specific needs of individual patients.

Enteral Nutrition

This is nutrition provided through a feeding tube placed through the nose into the stomach or beyond it into the duodenum, via a mini-surgical procedure in which a feeding tube is inserted through the abdominal wall into the stomach or beyond it into the jejunum using an endoscope, or by an open surgical approach to access the stomach or small intestine. EN is the treatment of choice when optimized voluntary nutritional support is impossible or has failed. It is relatively simple, safe, inexpensive, and maintains the digestive, absorptive, and immunologic barrier functions of the gastrointestinal tract. EN allows the delivery of accurately known amounts of nutrients. Pliable, small-bore feeding tubes make placement relatively easy and acceptable to most patients. Constant-rate infusion pumps increase the reliability of nutrient delivery. Patients are candidates for EN when optimized voluntary nutrition is not feasible or has failed, and their GI tract is adequately functional and can be accessed.

EN Products

The commonest forms of EN used are commercially manufactured formulas with defined compositions.

STANDARD POLYMERIC FORMULAS

These are the most widely used sources of EN. They are available in a wide variety of formats that generally meet the nutritional requirements of a normal, healthy person. Carbohydrates provide most of the energy. The proteins (from casein, whey, or soy) are intact and require normal pancreatic enzyme function for digestion and absorption. These products are isotonic or nearly so, and provide from 1000 to 2000 kcal and 50–70 g protein/L.

POLYMERIC FORMULAS WITH FIBER

The addition of dietary fiber to formulas sometimes improves bowel function and feeding tolerance. Fermentable (soluble) fibers such as pectin and guar are metabolized by colonic bacteria, yielding short-chain fatty acids that fuel colonocytes. Nonfermentable (insoluble) fibers increase fecal bulk, improve peristalsis, and may improve diarrhea.

ELEMENTAL AND SEMI-ELEMENTAL FORMULAS

The macronutrients in these formulas are partially or completely hydrolyzed. They are primarily designed for patients with known maldigestion and malabsorption, but they are sometimes used empirically for patients

who have had prolonged bowel rest or are critically ill without strong evidence of their superiority, or when a patient is intolerant of a standard polymeric formula.

IMMUNE-ENHANCING FORMULAS

In addition to providing macronutrients and conventional amounts of micronutrients, these products (IEFs) contain large amounts of certain nutrients designed to favorably modulate the immune response: arginine and n-3 fatty acids especially, but also various combinations of glutamine, nucleotides, and antioxidants. It is difficult to evaluate the IEFs because there are many different formulations on the market, and considerable heterogeneity in the patient populations studied. Good evidence of benefit has emerged from clinical trials of perioperative IEFs in patients undergoing elective gastrointestinal surgery and patients with traumatic brain injury. EIFs have not yet demonstrated benefit in other kinds of critical illness.

PROTEIN-ENRICHED FORMULAS

Most EN formulas provide calories and protein in a ratio appropriate for a healthy person, whereas protein-enriched formulas provide ~90 g protein and 1000 kcal/L. Originally marketed to meet increased protein requirement associated with hypocaloric nutrition in obese patients, these products are increasingly used to provide protein-catabolic patients an appropriately generous amount of protein without calorie-overfeeding.

OTHER FORMULAS

A wide variety of disease-specific EN products are available for patients with diabetes, hepatic, and renal or pulmonary disease. The details and evidence related to these products go beyond the scope of this chapter.

Parenteral Nutrition

PN delivers a complete nutritional regimen directly into the bloodstream in the form of crystalline amino acids, dextrose, triglyceride emulsions, minerals (calcium, phosphate, magnesium, and zinc), electrolytes, and micronutrients. Because of its high osmolarity (>1200 mOsm/L) and often large volume, PN is infused into a central vein in adults. Ready-to-use PN admixtures typically containing 4–7% hydrous amino acids and 20–25% dextrose (with or without electrolytes) are available in 2-chamber (amino acids and dextrose) or 3-chamber (amino acids, dextrose, and lipid) bags that are intermixed and vitamins, trace minerals and additional electrolytes added just prior to infusion. Although convenient and cost-effective, these products have fixed nutrient composition and thus are dosed according to the volume required to meet calorie requirements. In some situations—especially ADM—a costlier approach is justified that uses a computer-controlled sterile compounder to generate combinations of amino acid and dextrose that meet the precise protein and energy requirements of individual patients.

For example, 1 L of a standard ready-to-use admixture of 5% amino acids and 25% dextrose provides 50 g of amino acids (equivalent to 41.5 g protein substrate) and 1000 kcal; the use of this solution to meet the 1.5–2.0-g/kg protein requirement of an acutely ill 70-kg patient requires the infusion of 2.5–3.4 L of fluid and a

potentially excessive amount of energy (2500–3300 kcal). When the patient has an adequate fat store, clinical evidence increasingly supports the safety and efficacy of high-protein, moderately hypocaloric nutrition in ADM. A sterile compounder can accurately generate an appropriate recipe for such a patient. For example, 1 L of an admixture of 600 mL of 15% amino acids, 300 mL of 70% dextrose, and 100 mL of electrolyte/micronutrient mix contains 90 g amino acids (equivalent to 75 g of protein substrate) and 1020 kcal. (1 g mixed hydrated amino acids provides 3.3 kcal; 1 g dextrose provides 3.4 kcal; 1 g lipid emulsion provides ~10 kcal.) When a compounder is used, this patient's increased protein requirement can be met with less volume and avoid excessive calorie provision.

Amino Acids

Protein synthesis requires 21 alpha-amino acids, 9 of which are essential and 11 are non-essential because they are readily synthesized from an essential amino acid (methionine or phenylalanine) or from widely available carbohydrate precursors and the amine groups provided by the body's large and rapidly interconverting pool of non-essential amino acids. One amino acid—arginine—is conditionally essential. PN amino acid admixtures vary somewhat but all of them provide appropriate amounts of the essential amino acids and arginine while compensating for their lack of glutamine (and sometimes glutamate or aspartate) by including large amounts of glycine or other non-essential amino acids. The specific contribution of each non-essential amino acid to a nutritional admixture is less important than the total amount of non-essential N it provides. The hydrated status of the free amino acids in PN reduces their calorie density from 4.0 to 3.3 kcal/g, and reduces the amount of protein substrate they provide by 17%. For example, 100 g of free mixed amino acids provide 83 g protein substrate and 340 kcal.

Carbohydrate and Lipids

The carbohydrate in PN is dextrose monohydrate (3.4 kcal/g). Lipid emulsions provide calories and the essential n-6 and n-3 fatty acids. Traditional lipid emulsions are based solely on soy bean oil, but they are giving way to mixed emulsions that include medium-chain triglycerides, n-9 monounsaturated fatty acids, and n-3 fatty acids. Emulsions of pure soybean oil, a mixture of 80% olive oil and 20% soybean oil, and a mixture of 30% soybean oil, 30% medium chain triglycerides, 25% olive oil and 15% fish oil are available. The more complex lipid emulsions are more highly enriched in n-3 fatty acids and fewer n-6 polyunsaturated fatty acids than soybean lipid, which is more prone to lipid peroxidation and could promote the formation of the pro-inflammatory n-6 derivatives. As a general rule, lipid infusion rates should not exceed 8 g/h (for a 70 kg patient) or 175 g (1925 kcal)/day.

Minerals, Micronutrients, and Trace Elements

The default concentrations of electrolytes, minerals, and micronutrients in PN solutions are designed to meet normal requirements, and adjusted to meet the frequently abnormal and often-changing requirements of individual patients. Because they are unstable, multivitamin mixtures are injected into PN bags just prior to their delivery to the floor. Parenteral water-soluble vitamin requirements are greater than standard oral requirements because hospitalized patients often have vitamin deficiencies or increased requirements, and

because intravenous administration of vitamins increases urinary losses. Vitamin C spontaneously degrades in PN solutions even when light-protected. The amount of vitamin D in currently used intravenous vitamin products is inadequate. Iron salts are incompatible with most PN solutions.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Indications, Selection, and Provision of Specialized Nutritional Support

SNS is complicated, and the quantity and quality of the formal clinical evidence supporting its different uses is limited. Yet malnutrition-caused diseases are highly prevalent; they worsen clinical outcomes; and they are preventable and treatable. Physicians are charged with preventing, diagnosing, and treating these diseases, and they carry out their duty properly when guided by a sound understanding of nutritional principles, astute observation, rigorous clinical reasoning, and collaboration with dietitians and specialized clinical pharmacists.

Most hospitalized patients should not require SNS. Many of them will improve with appropriate management of their primary disease. Others have a terminal disease whose downward course will not be halted by SNS. Patients who can't eat enough hospital food and who have, or are at high risk of SM or CDM, are candidates for optimized voluntary nutrition support. When this approach is inappropriate or has failed, the pros and cons of invasive SNS are considered. The decision to provide or withhold EN or PN is based on a synthesis of four factors: (1) the determination that nutrient ingestion will likely continue to be inadequate for many days; (2) the patient has important muscle atrophy (of any cause) or fat depletion; (3) the patient's nutrient requirements are increased (as from inflammatory diarrhea, enterocutaneous fistulas or exudates, or a pronounced inflammatory protein-catabolic state); and (4) the reasoned judgment that SNS has a reasonable prospect of improving the patient's clinical outcome or quality of life. When the patient already has PEM (and treating it is likely to improve their well-being and clinical outcome), the decision tips more steeply in favor of intervention. It is important to formally diagnose and document PEM and its variants. Formal diagnosis focuses attention on the urgency of the situation and guides the selection, composition, and urgency of nutritional therapy.

EN THERAPY

EN is indicated when patients cannot eat enough food and unlikely to do so for a long time, their gastrointestinal tract is functional and accessible, and optimized voluntary nutrition is impossible or inadequate to meet the patient's nutritional needs. EN is most commonly used in settings of impaired consciousness, severe dysphagia, severe upper gastrointestinal tract dysfunction or obstruction, the requirement for mechanical ventilation, and critical illness in general. As well, situations frequently arise in which a patient's voluntary food intake is seriously curtailed by combinations of anorexia, nausea, vomiting, pain, delirium, depression, distress, chewing difficulties, undiagnosed thrush, unappealing food, and physical and sensory disability (including dysgeusia). In these complicated, difficult, and evolving situations, the diagnosis of SM or CDM should tip the decision towards EN.

Intestinal ischemia, mechanical obstruction, peritonitis, and gastrointestinal hemorrhage are contraindications to EN. Severe coagulopathy, esophageal varices, absent gag reflex, hypotension, adynamic ileus, pancreatitis, diarrhea, and nausea and vomiting are not absolute contraindications, but they increase the risk of complications and make it less likely that EN will succeed in achieving its nutritional goal.

Initiation, Progression, and Monitoring

Feeding tube insertion and EN delivery are medical procedures that require voluntary informed consent. Nasogastric tube feeding may proceed when gastrointestinal function is adequate with regard to gastric contractility (e.g., nasogastric tube output <1200 mL/d), intestinal contractility (absence of a known or suspected intraabdominal pathologic process, non-distended abdomen and detectable bowel sounds, although the *absence* of bowel sounds is not, in itself, a contraindication), and adequate colonic function (passage of stools and flatus). After consent has been obtained and the appropriate feeding tube (usually a nasogastric tube for short-term feeding) has been placed and its position verified, the head of the patient's bed is raised to at least 30° and kept raised to reduce the risk of regurgitation. A standard polymeric formula is infused, usually at a starting rate of 50 mL/h and advanced by 25 mL/h every 4–8 h until the goal rate is met. Elemental formulas commence at a slower rate and progress more slowly. Intragastric feeding allows a higher formula osmolarity. Intragastric bolus feeding is an option (200–400 mL feeding solution infused over 15–60 min at regular intervals with verification of residual gastric contents every 4 h). Bolus feeding is not possible with jejunal feeding, which requires the patient to be monitored for abdominal pain and abdominal distention and bowel sounds every 4 h.

Complications and Their Management

The most common complications of EN are aspiration of regurgitated or vomited formula, diarrhea, fluid volume and electrolyte derangements, hyperglycemia, nausea, abdominal pain, constipation, and failure to achieve the nutritional goal.

Aspiration

Debilitated patients with delayed gastric emptying, impaired gag reflex, and ineffective cough are at high risk of aspiration pneumonia. Aspiration is particularly common in mechanically ventilated patients. Ventilator-associated pneumonia is mostly caused by aspiration of microbial pathogens in the mouth and throat past the cuffs of endotracheal or tracheostomy tubes, but tracheal suctioning induces coughing and gastric regurgitation. Measures to prevent ventilator-associated pneumonia include elevation of the head of the bed, mouth hygiene and gastrointestinal decontamination, nurse-directed algorithms for formula advancement and, sometimes, post–pyloric feeding. EN does not have to be held for gastric residual volumes less than 300 to 400 mL in the absence of other signs of gastrointestinal intolerance (nausea, vomiting, severe abdominal pain, abdominal distention). Continuous EN is often tolerated better than bolus feeding, and it is the only option with jejunal feeding.

Diarrhea

Diarrhea is common when bowel function is compromised by disease or drugs (most often, broad-spectrum antibiotics). Once infectious and inflammatory causes have been ruled out, EN-associated diarrhea may be

controlled by the use of a fiber-containing formula or the addition of an antidiarrheal agent to it. H2 blockers or proton pump inhibitors may help reduce the net volume of fluid presented to the colon. Diarrhea does not usually impair macronutrient absorption, since amino acids, lipids, and glucose are mostly absorbed in the proximal-to-middle small intestine. Since luminal nutrients have trophic effects on the intestinal mucosa, it is often appropriate to persist with tube feeding despite moderate, tolerable diarrhea, even it necessitates supplemental parenteral fluid support. Except for patients with markedly impaired small-intestinal absorptive function, there are no well-established indications for elemental formulas, but they may be used empirically when diarrhea persists despite the use of fiber-enriched formulas and antidiarrheal agents.

Gastrointestinal Intolerance

Abnormally high gastric residual volumes, abdominal distention, pain, and nausea are distressing for patients, increase the nursing workload, and delay the progression of EN. These problems can be avoided or minimized by ensuring normal fluid and electrolyte balance, preventing severe hyperglycemia, and, when a patient experiences nausea, vomiting or abdominal distention, by the judicious use of antiemetic and prokinetic drugs (and sometimes proton pump inhibitors) on a regular—rather than as needed—basis. Patients with gastroparesis require post-pyloric feeding.

Fluid Volume, Electrolyte, and Blood Glucose Abnormalities

The essential purpose of EN is to provide macronutrients at an appropriate rate. EN formulas provide standard amounts of fluid, electrolytes, minerals, and micronutrients. They are not designed to manage abnormal fluid volume, electrolyte, and mineral requirements, which vary considerably among different patients and can change rapidly. Altered fluid volume requirements can, to a certain extent, be accommodated by selecting EN formulas with appropriate osmolarities, but medically active patients have widely varying and rapidly changing requirements for fluid, electrolyte, and glucose control. Blood glucose status must be monitored regularly and measures—including intravenous fluid, electrolyte, and insulin therapy—taken to maintain homeostasis.

Failure to Reach the Nutritional Goal

EN is frequently delayed or interrupted in medically active patients. The reasons are many: diagnostic tests and procedures (including dialysis), physical or occupational therapy, a clogged or pulled out tube, and intolerance to EN. When the flow rate is low, formula at the tip of the feeding tube may be precipitated by gastric acid and cause an occlusion. Inadequate tube flushing, dense formulas, and the introduction of inadequately homogenized solid medications also cause tube clogging. The end result is prolonged delay in the progression of EN and failure to meet the patient's nutrient requirements. Brief periods of inadequate calorie provision are usually inconsequential in patients who have adequate fat reserves, but deficient provision of protein is a serious but, unfortunately, under-investigated problem. The protein-to-calorie ratio of most EN formulas is appropriate for healthy people and hence too low for patients whose protein requirement is increased. High-protein formulas are appropriate when EN is progressing too slowly or the patient has an increased protein requirement, existing PEM, or is obese and will directly benefit from hypocaloric nutrition.

EN in the Intensive Care Unit

Most critically ill patients cannot eat anything—they depend entirely on SNS. EN has two purposes in the intensive care unit. The first purpose is to meet the patient's macronutrient requirements—especially their often dramatically increased protein requirement. The second purpose is to infuse nutrients into the intestines at a rate that sustains normal intestinal barrier and immunological functions in the face of a systemic inflammatory response that threatens intestinal integrity and immune function. Current guidelines recommend that EN commence as soon as possible after a critically ill patient has been fluid resuscitated and stabilized. The initial rate is 10–20 kcal/h. This rate of EN delivery provides ~25 g protein and ~500 kcal/d to patients who may require >100 g protein and >1800 kcal/d. Once EN is underway, the rate of delivery is increased as tolerated toward the patient's nutritional goal. Too often, however, EN falls far short of the protein provision target, even after a week or longer in the intensive care unit. Newer, high-protein EN products may reduce the severity of this protein shortfall.

PN THERAPY

PN is costlier, more resource-intensive, potentially riskier, and requires more expertise than EN. It is used when invasive SNS is indicated and EN is impossible, inappropriate, or unable to meet the patient's nutritional needs. The risks of PN are those of inserting and maintaining a central venous catheter (traumatic injury from the insertion, serious infection, and venous thrombosis), allergy to some of its components, glucose, electrolyte, magnesium, phosphate, and acid-base balance abnormalities, the adverse effects of the large intravenous fluid volumes. Prolonged PN—especially when it delivers excess calories—can lead to hepatic dysfunction.

INITIATION, PROGRESSION, MONITORING, AND DISCONTINUATION

PN may commence after the patient has been hemodynamically resuscitated, glucose, electrolyte, and acid-base homeostasis are established, and the patient is able to tolerate the fluid volumes involved. In adults, the high osmolarity of PN solutions, fluid demands, and need for strict sterility require their infusion through a dedicated port in a central venous catheter. Peripherally inserted central catheters (PICCs) are increasingly favored, although they limit monitoring for catheter infection by wire exchange. Jugular or femoral vein catheters are discouraged because it is very difficult to maintain a dry, sterile dressing over the insertion site. The initial dose of dextrose should not exceed 200 g/day to avoid hyperglycemia (or, in susceptible patients, the refeeding syndrome). On the other hand, the full requirement dose of amino acids can be administered from the first day onwards (this option is unavailable when premixed PN solutions are used). The glucose dose increases on a daily basis until it approaches the patient's energy requirement. Lipid emulsions are added after the first week.

Capillary blood glucose is monitored several times daily and subcutaneous regular insulin is added to the PN admixture as required to maintain average serum glucose concentrations <140 mg/dL and >80 mg/dL. The dose of regular insulin required to do this on a given day can be added to the following day's PN solution. The insulin dose increases roughly proportionately to the increasing glucose dose. Certain benchmarks are useful. Basal endogenous insulin secretion is ~30 units/day in normal people. When insulin is required for non-diabetic, non-catabolic patients, 10 units of regular insulin will roughly cover 100 g infused dextrose. Patients with non-insulin dependent diabetes require ~20 units/100 g dextrose. Non-catabolic patients with

insulin-dependent diabetes usually require approximately twice the at-home insulin dose, because parenteral glucose is a more potent insulin secretagogue than oral carbohydrate and because some insulin adheres to the infusion bag. As a general rule (for a patient with dry body weight 70 kg) the glucose infusion rate should not exceed ~500 g (1700 kcal)/day in non-critically ill patients, and should not exceed ~350 g (1200 kcal)/day in critically ill patients. Even lower glucose infusion rates (e.g., 200 g/day) are advisable if they prevent or minimize hyperglycemia in insulin-resistant patients.

In general, the benefits of constraining glucose, lipid, and fluid volume provision in ADM justify hypocaloric nutrition for the first 2 weeks of SNS as long as the calorie deficit is counterbalanced by generous amino acid provision. Lipids are commonly introduced after the first week of PN and can make up calorie shortfalls. Serum triglyceride concentrations are measured before commencing lipid infusions in order to detect preexisting hypertriglyceridemia (usually defined as >400 mg/dL), which is a relative contraindication. They may be infused daily or 2–3 times weekly. As appreciation or the adverse effects of excessive glucose administration has increased, interest has grown in the use of lipid emulsions as an energy source. Lipid infusion rates should not usually exceed ~50 g (500 kcal)/day in critical illness.

Biochemical Monitoring

Serum biochemistry (urea, creatinine, electrolytes, glucose, magnesium, phosphate, calcium, and albumin) are measured prior to starting PN and followed daily for the first few days, then twice weekly or as required. Serum triglycerides and liver function tests (and often ferritin) are measured at baseline and again after PN is underway to confirm that the lipid infusions are well tolerated. N balance (calculated from 24-h urinary urea N excretion) is useful at the outset for evaluating the severity of protein catabolism in patients with CDM or ADM to identify patients who require more generous amino acid provision, and during PN to determine whether the patient's N balance is improving with therapy. Serum ferritin should be measured every 2 months, although the duration of most in-hospital PN is shorter than this.

Discontinuation

PN is tapered and discontinued as soon as the patient can be adequately nourished by the enteral route. The dose of PN is reduced as food intake increases. As a general rule, once a patient is tolerating one-half to two-thirds of their food requirement by the enteral route and there is no mechanical or other barrier to further improvements in intake, PN can be discontinued. The transition to oral nutrition is slow for many patients with CDM. Optimized voluntary nutrition is much preferred to invasive EN for these patients because it is safe, effective, fosters well-being, and prepares them for discharge. It is tempting to stop PN as a way to stimulate more food consumption; the temptation should be resisted. PN does not create anorexia, nor does discontinuing it stimulate appetite. Too-early discontinuation of PN may delay a patient's progression to full voluntary food consumption by inducing anxiety and recreating starvation conditions. PN is most successfully weaned by encouraging physical activity, optimizing voluntary nutrition (including food from home), emotional support, and having patience. On the other hand, patients who are at the cusp of adequate oral nutrition commonly benefit from hospital discharge, where the security and pleasure of being at home and eating home-made food are potent stimuli.

Drawbacks, Side Effects, and Complications

Complications of Central Catheters

Patients receiving PN are at greater risk of bloodstream infections than other patients with central venous catheters. Proper aseptic insertion technique, meticulous dressing care, and one port dedicated solely to PN reduce this risk. Catheter-induced upper arm venous thrombosis is an uncommon but important complication.

Hyperglycemia

The most frequent metabolic complication of PN is hyperglycemia in patients with insulin resistance due to non-insulin dependent diabetes mellitus, high-dose glucocorticoid therapy, or severe systemic inflammation; the problem is exacerbated by excessively high rates of glucose provision. Glucose concentrations are most easily kept at <140 mg/dL with the least risk of hypoglycemia by infusing hypocaloric amounts of glucose and, if necessary, meeting the patient's energy requirement with intravenous lipid. In ADM, the benefits of using the lowest insulin dose—minimal hyperinsulinemia and a reduced risk of hypoglycemia—almost always outweigh the doubtful goal of rapidly matching calorie provision to the patient's energy expenditure rate.

Hypoglycemia

Reactive hypoglycemia is uncommon but may occur when high-dextrose, non-insulin containing PN is abruptly discontinued. It is prevented by slowing the PN infusion rate to 50 mL/h for 1 or 2 h prior to discontinuing it (or replacing it with 10% dextrose), or, when the oral route is available, providing a snack. More often, hypoglycemia occurs when the intensity of the patient's metabolic stress (or their glucocorticoid dose) decreases without an appropriate downward adjustment of the insulin dose. This problem is avoided by frequent capillary glucose determinations and careful attention to medication doses and the patient's general condition.

Artefactual Hyperglycemia and Hyperkalemia

Blood samples must be carefully and appropriately drawn from dual-port PICC. Intermixing of the sample with even a tiny volume of PN solution will falsely indicate hyperglycemia and hyperkalemia, and may trigger a treatment error. The problem is identified when the patient's apparent serum glucose (and potassium) concentrations abruptly increase without reason and the apparently very high glucose concentration is out of keeping with concurrent capillary glucose readings.

Volume Overload

Hypertonic intravenous dextrose stimulates a much more intense insulin response than oral glucose. A potent antinatriuretic and antidiuretic hormone, insulin potentiates sodium and water retention. In this setting net fluid retention is likely when total fluid provision exceeds 2 L/day in patients not experiencing large gastrointestinal losses. The problem of volume overload can be reduced by preparing PN solutions with a compounder, infusing glucose at a rate that minimizes the need for exogenous insulin therapy, and avoiding overfeeding in general. Sodium intake plays an important role in fluid retention. Net fluid retention can be minimized by limiting sodium delivery to 20–30 mmol/day (with an adjustment for gastric or intestinal

sodium losses) since, when renal function is normal and antidiuretic hormone secretion is not increased, urinary sodium concentrations are usually <10 mmol/L.

Hypertriglyceridemia

This complication occurs when the rate of lipid infusion exceeds plasma triglyceride clearance capacity. Renal failure, sepsis, excessive glucose (which stimulates lipogenesis), diabetes mellitus, high-dose glucocorticoid therapy, and multiple-organ failure reduce triglyceride clearance. An impaired immune response, increased risk of acute pancreatitis, and altered pulmonary hemodynamics are potential but not well documented complications of PN-induced severe hypertriglyceridemia.

Hepatic Dsyfunction

Mild elevations of serum liver enzyme concentrations can occur within 2–4 weeks of initiating PN, but in most cases they return to normal even when PN is continued. Clinically important hepatic dysfunction, although common in children, is uncommon in adults as long as energy overfeeding and resultant fatty liver are avoided. Intrahepatic cholestasis occasionally occurs after many weeks of continuous PN and is most often multifactorial in origin. Cyclic PN—in which PN is infused for only 12 h/day—may prevent or reduce the severity of this complication.

PN IN THE INTENSIVE CARE UNIT

Current guidelines recommend starting EN as soon as a critically ill patient has been resuscitated, stabilized, and enteral access established. EN is then advanced over the following days to toward the patient's nutritional goal. If the goal has not been achieved after 7–10 days, amino acid-rich PN is recommended, especially when the patient remains protein catabolic. PN that is generous in amino acids (85–140 g protein substrate/day for a 70 kg patient) and hypocaloric (1200–1400 kcal/day) limits the risk and severity of hyperglycemia and volume overload and may improve clinical outcomes in this setting. Soy-based lipid emulsions should be avoided during the first week of PN in the intensive care unit; alternative lipid emulsions may prove to be safe and beneficial.

SPECIAL CLINICAL CONDITIONS

Old Age

In addition to their other physiological frailties, elderly people commonly suffer from age-related muscle atrophy that is compounded by disuse muscle atrophy. These factors place them at high risk of PEM and make them plausible candidates for early SNS.

Inactivity

Physical activity and adequate nutrition are closely interdependent. Reduced physical activity reduces appetite, and physical rehabilitation and its associated emotional benefits restore optimism and appetite. Full nutrient provision will maintain or normalize many physiological functions in bedridden patients, but they will not increase muscle mass.

Renal Failure

Protein provision should not be reduced in the presence of renal injury unless renal replacement therapy is unavailable. Protein and vitamin C provision should increase when renal replacement therapy is used, for it removes large amounts of amino acids and vitamin C from the circulation.

Liver Failure

Patients with severe hepatic disease are plausible candidates for SNS because they are relatively intolerant of starvation and commonly already have CDM when they are admitted to hospital. SNS should be generous in calories and especially protein, which should be provided despite an increased risk of hepatic encephalopathy. The risk of encephalopathy can be mitigated by meticulous attention to fluid balance and electrolyte status and by spreading protein provision over the day to accommodate the liver's reduced capacity to clear amino acid-derived ammonia.

Perioperative SNS

Patients awaiting elective major surgery benefit from 7 to 10 days of preoperative SNS if SM (or especially CDM) is present. When feasible and properly implemented, optimized voluntary nutrition is greatly preferred, but when a patient has been admitted to hospital their condition is by definition semi-urgent, and EN or PN will meet the patient's nutritional goal more quickly. Preoperative SNS improves immunity and reduces postoperative complications, but it will not increase serum albumin concentrations and it should not be provided for >7–10 days with that goal in mind. More prolonged preoperative EN or PN may confer slight additional nutritional benefits, but they are counterbalanced by its risks and the consequences of prolonged hospitalization and delayed surgery. Surgery should not be delayed for starving patients whose muscle mass is normal or only mildly depleted and who are not experiencing systemic inflammation, since they tolerate even major uncomplicated surgery well without preoperative SNS. The urgency of surgery often precludes otherwise indicated preoperative SNS. Early postoperative PN is usually indicated for these patients, for they are at increased risk of postoperative complications and are highly unlikely to consume an adequate amount of food voluntarily over the next many days. Patients with only mild muscle atrophy, no systemic inflammation, and no postoperative complications do not require postoperative PN unless (1) adequate feeding by mouth has not been achieved by day 5–7 after surgery or (2) there are indications that voluntary feeding will be further delayed. Warning indicators include inadequately controlled nausea or pain, impaired gastric, small intestinal or colonic function, serum electrolyte imbalance, altered mental status, inability to mobilize from the bed, or a suspected surgical-site infection or anastomotic leak. Perioperative immune-enhancing EN appears to reduce morbidity in patients undergoing major elective gastrointestinal surgery.

Iron and PN

Iron deficiency is common in acutely ill hospitalized patients. Risk factors for in-hospital iron deficiency include inadequate nutritional intake, gastrointestinal disease, and frequent blood withdrawals. In-hospital iron deficiency is often missed because inflammation-associated anemia is much commoner and increases serum concentrations of ferritin, a positive acute-phase protein. The parenteral iron requirement is normally only ~1 mg/day, but since iron is a highly reactive catalyst of oxidative reactions it is not included in PN mixtures. Serum ferritin concentrations should usually be measured when PN commences and re-measured at ~8 week intervals. A falling mean red cell volume (even within the low-normal range) or an intermediate

serum ferritin concentration in the presence of systemic inflammation strongly suggest iron deficiency. Intravenous iron should be ordered according to standard guidelines. A termination order should also be written to avoid the risk of inadvertent iron overdosing. Iron replacement should be avoided during the acute phase of critical illness because a substantial rise in the serum iron concentration could increase susceptibility to some bacterial infections, but it is indicated in SM and CDM both to improve iron-deficiency anemia and provide the non-hematologic benefits of adequate iron nutrition.

Zinc

It is insufficiently appreciated that 1 L of secretory diarrhea contains ~12 mg of zinc. Patients with intestinal fistulas or high volume chronic diarrhea require this amount of zinc in addition to their daily requirement of 15 mg to avoid zinc deficiency. Zinc may be provided parenterally or enterally (because of its low bioavailability, 12 mg parenteral zinc is equivalent to 30 mg oral elemental zinc).

Cancer

SNS plays a crucial role in cancer therapy. Many cancers (especially those that involve the gastrointestinal tract or induce systemic inflammation) and most cancer therapies create the conditions for starvation and commonly lead to SM or CDM. The prevention or treatment of these diseases may improve patients' quality of life and tolerance to anti-cancer therapy. As a general rule, EN and PN are not prescribed to patients who are not undergoing active anti-cancer therapy because the side effects and complications of invasive SNS are not counterbalanced by an improved disease trajectory. In some cases, a disease may be inexorably progressing but so slowly that the patient will die of starvation long before they would from the cancer. EN or PN is usually appropriate in these cases.

Advanced Dementia

Optimized voluntary nutrition is the key approach in this situation, and it can be used to deal with problems such as disability and dysphagia in patients who show discernable pleasure from eating. There is no evidence that EN or PN improve quality or length of life in patients who have advanced dementia and show little or no interest in eating food, and their side effects and complications are unpleasant and sometimes dangerous.

REFEEDING SYNDROME

The refeeding syndrome can occur in starving patients during the first week of nutritional repletion if carbohydrates are introduced too rapidly. Carbohydrate provision stimulates insulin secretion, which, owing to its antinatriuretic effect, expands the ECF volume when excessive sodium is provided. Refeeding edema can be minimized by severely limiting sodium provision and increasing carbohydrate provision slowly. Carbohydrate refeeding may stimulate enough intracellular glucose-6-phosphate and glycogen synthesis to seriously lower serum phosphate concentrations. Refeeding increases the adaptively down-regulated REE of patients with SM and stimulates N retention, new cell synthesis, and cellular rehydration. Phosphate, potassium, magnesium, and zinc are common and dangerous during refeeding. Their status must be monitored and appropriate supplements provided. Acute thiamine deficiency is a devastating but preventable complication of refeeding, even with simple dextrose infusions. Left heart failure may occur in predisposed patients. The precipitants of left heart failure are an abrupt increase of intravascular volume due

to the administration of fluids and of glucose, which stimulates insulin secretion with associated renal sodium retention; increased cardiac demand because of increased REE in a patient with an atrophic left ventricle and low stroke volume; and myocardial deficiencies of potassium, phosphorus, or magnesium. Cardiac arrhythmias may occur. Refeeding may precipitate acute wet beriberi in malnourished, poverty-stricken populations.

GLOBAL CONSIDERATIONS



The macro-nutritional diseases are the same worldwide, but their prevalence varies in accordance with regional variations in the prevalence of the primary medical and surgical diseases that induce them and the two most important causes of starvation: poverty and social iniquity. Longstanding borderline SM is so prevalent in some parts of the world that the World Health Organization has coined the term *chronic energy deficiency* (CED) to describe it. CED is classified into three grades of severity: Grade I: BMI 17.0–18.4, Grade II: BMI 16.0–16.9, and Grade III: BMI <16. In otherwise healthy adults, CED of Grades II and III are associated with an increasing probability of days of illness, reduced work capacity, poorer reproductive function, and poorer lactation performance. Voluntary physical activity is decreased in Grade III CED, which is equivalent to the diagnosis of SM.

FURTHER READING

Derenski K et al: Parenteral nutrition basics for the clinician caring for the Adult Patient. Nutr Clin Pract 31:578, 2016. [PubMed: 27440772]

Esuro AA, Hummell AC: Enteral formulas in nutrition support practice: Is there a better choice for your patient? Nutr Clin Pract 31:709, 2016. [PubMed: 27694642]

Lembeck ME et al: The role of intravenous fluids and enteral or parenteral nutrition in patients with life-limiting illness. Med Clin North Am 100:1131, 2016. [PubMed: 27542432]

McClave SA et al: Guidelines for the provision and assessment of nutrition support therapy in the adult critically ill patient: Society of Critical Care Medicine (SCCM) and American Society for Parenteral and Enteral Nutrition (ASPEN). J Parenter Enteral Nutr 40:159, 2016.

White JV et al: Consensus statement: Academy of Nutrition and Dietetics and American Society for Parenteral and Enteral Nutrition: Characteristics recommended for the identification and documentation of adult malnutrition (undernutrition). J Parenter Enteral Nutr 36:275, 2012.

Chapter 329: Approach to the Patient with Liver Disease

Marc G. Ghany; Jay H. Hoofnagle

FIGURE 329-1

INTRODUCTION

A diagnosis of liver disease usually can be made accurately by careful elicitation of the patient's history, physical examination, and application of a few laboratory tests. In some circumstances, radiologic examinations are helpful or, indeed, diagnostic. Liver biopsy is considered the criterion standard in evaluation of liver disease, but is now needed less for diagnosis than for grading (activity) and staging (fibrosis) of disease. Non-invasive means of assessing fibrosis stage have become increasingly helpful and may allow for avoidance of biopsy in a proportion of patients. This chapter provides an introduction to diagnosis and management of liver disease, briefly reviewing the structure and function of the liver; the major clinical manifestations of liver disease; and the use of clinical history, physical examination, laboratory tests, imaging studies, and liver biopsy.

LIVER STRUCTURE AND FUNCTION

The liver is the largest organ of the body, weighing 1–1.5 kg and representing 1.5–2.5% of the lean body mass. The size and shape of the liver vary and generally match the general body shape—long and lean or squat and square. This organ is located in the right upper quadrant of the abdomen under the right lower rib cage against the diaphragm and projects for a variable extent into the left upper quadrant. It is held in place by ligamentous attachments to the diaphragm, peritoneum, great vessels, and upper gastrointestinal organs. The liver receives a dual blood supply; ~20% of the blood flow is oxygen-rich blood from the hepatic artery, and 80% is nutrient-rich blood from the portal vein arising from the stomach, intestines, pancreas, and spleen.

The majority of cells in the liver are hepatocytes, which constitute two-thirds of the organ's mass. The remaining cell types are Kupffer cells (members of the reticuloendothelial system), stellate (Ito or fat-storing) cells, endothelial and blood vessel cells, bile ductular cells, and cells of supporting structures. Viewed by light microscopy, the liver appears to be organized in lobules, with portal areas at the periphery and central veins in the center of each lobule. However, from a functional point of view, the liver is organized into acini, with both hepatic arterial and portal venous blood entering the acinus from the portal areas (zone 1) and

then flowing through the sinusoids to the terminal hepatic veins (zone 3); the intervening hepatocytes constitute zone 2. The advantage of viewing the acinus as the physiologic unit of the liver is that this perspective helps to explain the morphologic patterns and zonality of many vascular and biliary diseases not explained by the lobular arrangement.

Portal areas of the liver consist of small veins, arteries, bile ducts, and lymphatics organized in a loose stroma of supporting matrix and small amounts of collagen. Blood flowing into the portal areas is distributed through the sinusoids, passing from zone 1 to zone 3 of the acinus and draining into the terminal hepatic veins (“central veins”). Secreted bile flows in the opposite direction—that is, in a counter-current pattern from zone 3 to zone 1. The sinusoids are lined by unique endothelial cells that have prominent fenestrae of variable sizes, allowing the free flow of plasma but not of cellular elements. The plasma is thus in direct contact with hepatocytes in the subendothelial space of Disse.

Hepatocytes have distinct polarity. The basolateral side of the hepatocyte lines the space of Disse and is richly lined with microvilli; it exhibits endocytotic and pinocytotic activity, with passive and active uptake of nutrients, proteins, and other molecules. The apical pole of the hepatocyte forms the canalicular membranes through which bile components are secreted. The canaliculi of hepatocytes form a fine network, which fuses into the bile ductular elements near the portal areas. Kupffer cells usually lie within the sinusoidal vascular space and represent the largest group of fixed macrophages in the body. The stellate cells are located in the space of Disse but are not usually prominent unless activated, when they produce collagen and matrix. Red blood cells stay in the sinusoidal space as blood flows through the lobules, but white blood cells can migrate through or around endothelial cells into the space of Disse and from there to portal areas, where they can return to the circulation through lymphatics.

Hepatocytes perform numerous and vital roles in maintaining homeostasis and health. These functions include the synthesis of most essential serum proteins (albumin, carrier proteins, coagulation factors, many hormonal and growth factors), the production of bile and its carriers (bile acids, cholesterol, lecithin, phospholipids), the regulation of nutrients (glucose, glycogen, lipids, cholesterol, amino acids), and the metabolism and conjugation of lipophilic compounds (bilirubin, anions, cations, drugs) for excretion in the bile or urine. Measurement of these activities to assess liver function is complicated by the multiplicity and variability of these functions. The most commonly used liver “function” tests are measurements of serum bilirubin, serum albumin, and prothrombin time. The serum bilirubin level is a measure of hepatic conjugation and excretion; the serum albumin level and prothrombin time are measures of protein synthesis. Abnormalities of bilirubin, albumin, and prothrombin time are typical of hepatic dysfunction. Frank liver failure is incompatible with life, and the functions of the liver are too complex and diverse to be subserved by a mechanical pump; a dialysis membrane; or a concoction of infused hormones, proteins, and growth factors.

LIVER DISEASES

While there are many causes of liver disease (Table 329-1), these disorders generally present clinically in a few distinct patterns and are usually classified as hepatocellular, cholestatic (obstructive), or mixed. In *hepatocellular diseases* (such as viral hepatitis and alcoholic liver disease), features of liver injury, inflammation, and necrosis predominate. In *cholestatic diseases*, such as gallstone or malignant obstruction, primary biliary cholangitis (previously referred to as primary biliary cirrhosis), and some drug-induced liver diseases, features of inhibition of bile flow predominate. In a mixed pattern, features of both hepatocellular and cholestatic injury are present (such as in cholestatic forms of viral hepatitis and many drug-induced liver diseases). The pattern of onset and prominence of symptoms can rapidly suggest a diagnosis, particularly if major risk factors are considered, such as the age and sex of the patient and a history of exposure or risk behaviors.

TABLE 329-1

Liver Diseases

Inherited hyperbilirubinemia
- Gilbert syndrome
- Crigler-Najjar syndrome, types I and II
- Dubin-Johnson syndrome
- Rotor syndrome

Viral hepatitis
- Hepatitis A
- Hepatitis B
- Hepatitis C
- Hepatitis D
- Hepatitis E
- Others (Epstein-Barr virus [mononucleosis] herpesvirus, cytomegalovirus, adenovirus hepatitis)
- Cryptogenic hepatitis

Immune and autoimmune liver diseases
- Primary biliary cholangitis
- Autoimmune hepatitis
- Sclerosing cholangitis
- Overlap syndromes
- Graft-versus-host disease
- Allograft rejection

Genetic liver diseases
- α_1 antitrypsin deficiency
- Hemochromatosis
- Wilson disease
- Benign recurrent intrahepatic cholestasis
- Progressive familial intrahepatic cholestasis, types I–III
- Others (galactosemia, tyrosinemia, cystic fibrosis, Niemann-Pick-disease, Gaucher's disease)

Alcoholic liver disease
- Acute fatty liver
- Acute alcoholic hepatitis
- Laënnec cirrhosis

Nonalcoholic fatty liver
- Steatosis
- Steatohepatitis

Acute fatty liver of pregnancy

Liver involvement in systemic diseases
- Sarcoidosis
- Amyloidosis
- Glycogen storage diseases
- Celiac disease
- Tuberculosis
- *Mycobacterium avium-intracellulare* infection

Cholestatic syndromes
- Benign postoperative cholestasis
- Jaundice of sepsis
- Total parenteral-nutrition–induced jaundice
- Cholestasis of pregnancy
- Cholangitis and cholecystitis
- Extrahepatic biliary obstruction (stone, stricture, cancer)
- Biliary atresia
- Caroli disease
- Cryptosporidiosis

Drug-induced liver disease
- Hepatocellular patterns (isoniazid, acetaminophen)
- Cholestatic patterns (methyltestosterone)
- Mixed patterns (sulfonamides, phenytoin)
- Micro- and macrovesicular steatosis (methotrexate, fialuridine)

Vascular injury
- Sinusoidal obstruction syndrome
- Budd-Chiari syndrome
- Ischemic hepatitis
- Passive congestion
- Portal vein thrombosis
- Nodular regenerative hyperplasia

Mass lesions

	Hepatocellular carcinoma Cholangiocarcinoma Adenoma Focal nodular hyperplasia Metastatic tumors Abscess Cysts Hemangioma

Typical presenting symptoms of liver disease include jaundice, fatigue, itching, right-upper-quadrant pain, nausea, poor appetite, abdominal distention, and intestinal bleeding. At present, however, many patients are diagnosed with liver disease who have no symptoms and who have been found to have abnormalities in biochemical liver tests as a part of a routine physical examination or screening for blood donation or for insurance or employment. The wide availability of batteries of liver tests makes it relatively simple to demonstrate the presence of liver injury as well as to rule it out in someone in whom liver disease is suspected.

Evaluation of patients with liver disease should be directed at (1) establishing the etiologic diagnosis, (2) estimating disease severity (*grading*), and (3) establishing the disease stage (*staging*). *Diagnosis* should focus on the category of disease (hepatocellular, cholestatic, or mixed injury) as well as on the specific etiologic diagnosis. *Grading* refers to assessment of the severity or activity of disease—active or inactive as well as mild, moderate, or severe. *Staging* refers to estimation of the point in the course of the natural history of the disease, whether early or late; or precirrhotic, cirrhotic, or end-stage. This chapter introduces general, salient concepts in the evaluation of patients with liver disease that help lead to the diagnoses discussed in subsequent chapters.

CLINICAL HISTORY

The clinical history should focus on the symptoms of liver disease—their nature, patterns of onset, and progression—and on potential risk factors for liver disease. The manifestations of liver disease include constitutional symptoms such as fatigue, weakness, nausea, poor appetite, and malaise and the more liver-specific symptoms of jaundice, dark urine, light stools, itching, abdominal pain, and bloating. Symptoms can also suggest the presence of cirrhosis, end-stage liver disease, or complications of cirrhosis such as portal hypertension. Generally, the constellation of symptoms and their patterns of onset rather than a specific symptom points to an etiology.

Fatigue is the most common and most characteristic symptom of liver disease. It is variously described as lethargy, weakness, listlessness, malaise, increased need for sleep, lack of stamina, and poor energy. The fatigue of liver disease typically arises after activity or exercise and is rarely present or severe after adequate rest; that is, it is “afternoon” rather than “morning” fatigue. Fatigue in liver disease is often intermittent and variable in severity from hour to hour and day to day. In some patients, it may not be clear whether fatigue is

due to the liver disease or due to other problems such as stress, anxiety, sleep disturbance, or a concurrent illness.

Nausea occurs with more severe liver disease and may accompany fatigue or be provoked by smelling food odors or eating fatty foods. Vomiting can occur but is rarely persistent or prominent. Poor appetite with weight loss occurs frequently in acute liver disease, but is rare in chronic disease except when cirrhosis is present and advanced. Diarrhea is uncommon in liver disease except with severe jaundice, in which a lack of bile acids reaching the intestine can lead to steatorrhea.

Right-upper-quadrant discomfort or ache (“liver pain”) occurs in many liver diseases and is usually marked by tenderness over the liver area. The pain arises from stretching or irritation of Glisson’s capsule, which surrounds the liver and is rich in nerve endings. Severe pain is most typical of gallbladder disease, liver abscess, and severe sinusoidal obstruction syndrome (previously known as veno-occlusive disease) but is also an occasional accompaniment of acute hepatitis.

Itching occurs with acute liver disease, appearing early in obstructive jaundice (from biliary obstruction or drug-induced cholestasis) and somewhat later in hepatocellular disease (acute hepatitis). Itching also occurs in chronic liver diseases—typically the cholestatic forms such as primary biliary cholangitis and sclerosing cholangitis, in which it is often the presenting symptom, preceding the onset of jaundice. However, itching can occur in any liver disease, particularly once cirrhosis develops.

Jaundice is the hallmark symptom of liver disease and perhaps the most reliable marker of severity. Patients usually report darkening of the urine before they notice scleral icterus. Jaundice is rarely detectable with a bilirubin level <43 µmol/L (2.5 mg/dL). With severe cholestasis, there will also be lightening of the color of the stools and steatorrhea. Jaundice without dark urine usually indicates indirect (unconjugated) hyperbilirubinemia and is typical of hemolytic anemia and the genetic disorders of bilirubin conjugation, the common and benign form being Gilbert syndrome and the rare and severe form being Crigler-Najjar syndrome. Gilbert syndrome affects up to 5% of the general population; the jaundice in this condition is more noticeable after fasting and with stress.

Major risk factors for liver disease that should be sought in the clinical history include details of alcohol use, medication use (including herbal compounds, birth control pills, and over-the-counter medications), personal habits, sexual activity, travel, exposure to jaundiced or other high-risk persons, injection drug use, recent surgery, remote or recent transfusion of blood or blood products, occupation, accidental exposure to blood or needlestick, and familial history of liver disease.

For assessing the risk of viral hepatitis, a careful history of sexual activity is of particular importance and should include the number of lifetime sexual partners and, for men, a history of having sex with men. Sexual exposure is a common mode of spread of hepatitis B but is uncommon for hepatitis C. A family history of hepatitis, liver disease, and liver cancer is also important. Maternal-infant transmission occurs with both hepatitis B and C. Vertical spread of hepatitis B can now be prevented by passive and active immunization of the infant at birth, although addition of antiviral therapy during the third trimester of pregnancy is now

recommended for mothers with levels of HBV DNA >200,000 IU/mL. Vertical spread of hepatitis C is uncommon, but there are no reliable means of prevention. Transmission is more common among HIV-co-infected mothers and is also linked to prolonged and difficult labor and delivery, early rupture of membranes, internal fetal monitoring, and a high viral load. A history of injection drug use, even in the remote past, is of great importance in assessing the risk for hepatitis B and C. Injection drug use is now the single most common risk factor for hepatitis C. Transfusion with blood or blood products is no longer an important risk factor for acute viral hepatitis. However, blood transfusions received before the introduction of sensitive enzyme immunoassays for antibody to hepatitis C virus in 1992 is an important risk factor for chronic hepatitis C. Blood transfusion before 1986, when screening for antibody to hepatitis B core antigen was introduced, is also a risk factor for hepatitis B. Travel to a developing area of the world, exposure to persons with jaundice, and exposure to young children in day-care centers are risk factors for hepatitis A. Tattooing and body piercing (for hepatitis B and C) and eating shellfish (for hepatitis A) are frequently mentioned but are actually types of exposure that quite rarely lead to the acquisition of hepatitis.

 Hepatitis E is one of the more common causes of jaundice in Asia and Africa but is uncommon in developed nations. In endemic areas transmission is usually through exposure to fecally contaminated water. Recently, non-travel-related (*autochthonous*) cases of hepatitis E have been described in developed countries, including the United States. These cases appear to be due to strains of hepatitis E virus that are endemic in swine and some wild animals (genotypes 3 and 4). While occasional cases are associated with eating raw or undercooked pork or game (deer and wild boars), most cases of hepatitis E occur without known exposure, predominantly in elderly men without typical risk factors for viral hepatitis. Hepatitis E infection can become chronic in immunosuppressed individuals (such as transplant recipients, patients receiving chemotherapy, or patients with HIV infection), in whom it presents with abnormal serum enzymes in the absence of markers of hepatitis B or C.

A history of alcohol intake is important in assessing the cause of liver disease and also in planning management and recommendations. In the United States, for example, at least 70% of adults drink alcohol to some degree, but significant alcohol intake is less common; in population-based surveys, only 5% of individuals have more than two drinks per day, the average drink representing 11–15 g of alcohol. Alcohol consumption associated with an increased rate of alcoholic liver disease is probably more than two drinks (22–30 g) per day in women and three drinks (33–45 g) in men. Most patients with alcoholic cirrhosis have a much higher daily intake and have drunk excessively for ≥10 years before onset of liver disease. In assessing alcohol intake, the history should also focus on whether alcohol abuse or dependence is present. Alcoholism is usually defined by the behavioral patterns and consequences of alcohol intake, not by the amount. *Abuse* is defined by a repetitive pattern of drinking alcohol that has adverse effects on social, family, occupational, or health status. *Dependence* is defined by alcohol-seeking behavior, despite its adverse effects. Many alcoholics demonstrate both dependence and abuse, and dependence is considered the more serious and advanced form of alcoholism. A clinically helpful approach to diagnosis of alcohol dependence and abuse is the use of the CAGE questionnaire **(Table 329-2)**, which is recommended for all medical history-taking.

TABLE 329-2

CAGE Questions[a]

Acronym	Question
C	Have you ever felt you ought to *c*ut down on your drinking?
A	Have people *a*nnoyed you by criticizing your drinking?
G	Have you ever felt *g*uilty or bad about your drinking?
E	Have you ever had a drink first thing in the morning to steady your nerves or get rid of a hangover (*e*ye-opener)?

[a]One "yes" response should raise suspicion of an alcohol use problem, and more than one is a strong indication of abuse or dependence.

Family history can be helpful in assessing liver disease. Familial causes of liver disease include Wilson disease; hemochromatosis and α_1 antitrypsin deficiency; and the more uncommon inherited pediatric liver diseases—that is, familial intrahepatic cholestasis, benign recurrent intrahepatic cholestasis, and Alagille syndrome. Onset of severe liver disease in childhood or adolescence in conjunction with a family history of liver disease or neuropsychiatric disturbance should lead to investigation for Wilson's disease. A family history of cirrhosis, diabetes, or endocrine failure and the appearance of liver disease in adulthood suggests hemochromatosis and should prompt investigation of iron status. Abnormal iron studies in adult patients warrant genotyping of the *HFE* gene for the C282Y and H63D mutations typical of genetic hemochromatosis. In children and adolescents with iron overload, other non-*HFE* causes of hemochromatosis should be sought. A family history of emphysema should lead to investigation of α_1 antitrypsin levels and, if levels are low, for protease inhibitor (Pi) genotype.

PHYSICAL EXAMINATION

The physical examination rarely uncovers evidence of liver dysfunction in a patient without symptoms or laboratory findings, nor are most signs of liver disease specific to one diagnosis. Thus, the physical examination complements rather than replaces the need for other diagnostic approaches. In many patients, the physical examination is normal unless the disease is acute or severe and advanced. Nevertheless, the physical examination is important in that it can yield the first evidence of hepatic failure, portal hypertension, and liver decompensation. In addition, the physical examination can reveal signs—related either to risk factors or to associated diseases or findings—that point to a specific diagnosis.

Typical physical findings in liver disease are icterus, hepatomegaly, hepatic tenderness, splenomegaly, spider angiomata, palmar erythema, and skin excoriations. Signs of advanced disease include muscle wasting, ascites, edema, dilated abdominal veins, hepatic fetor, asterixis, mental confusion, stupor, and coma. In male patients with cirrhosis, particularly that related to alcohol use, signs of hyperestrogenemia such as gynecomastia, testicular atrophy, and loss of male-pattern hair distribution may be found.

Icterus is best appreciated when the sclera is inspected under natural light. In fair-skinned individuals, a yellow tinge to the skin may be obvious. In dark-skinned individuals, examination of the mucous membranes below the tongue can demonstrate jaundice. Jaundice is rarely detectable if the serum bilirubin level is <43 µmol/L (2.5 mg/dL) but may remain detectable below this level during recovery from jaundice (because of protein and tissue binding of conjugated bilirubin).

Spider angiomata and palmar erythema occur in both acute and chronic liver disease; these manifestations may be especially prominent in persons with cirrhosis but can develop in normal individuals and are frequently found during pregnancy. Spider angiomata are superficial, tortuous arterioles, and—unlike simple telangiectases—typically fill from the center outward. Spider angiomata occur only on the arms, face, and upper torso; they can be pulsatile and may be difficult to detect in dark-skinned individuals.

Hepatomegaly is not a highly reliable sign of liver disease because of variability in the liver's size and shape and the physical impediments to assessment of liver size by percussion and palpation. Marked hepatomegaly is typical of cirrhosis, sinusoidal obstruction syndrome, infiltrative disorders such as amyloidosis, metastatic, or primary cancers of the liver, and alcoholic hepatitis. Careful assessment of the liver edge may also reveal unusual firmness, irregularity of the surface, or frank nodules. Perhaps the most reliable physical finding in the liver examination is hepatic tenderness. Discomfort when the liver is touched or pressed upon should be carefully sought with percussive comparison of the right and left upper quadrants.

Splenomegaly, which occurs in many medical conditions, can be a subtle but significant physical finding in liver disease. The availability of ultrasound (US) methods for assessment of the spleen allows confirmation of the physical finding.

Signs of advanced liver disease include muscle wasting and weight loss as well as hepatomegaly, bruising, ascites, and edema. Ascites is best appreciated by attempts to detect shifting dullness by careful percussion. US examination will confirm the finding of ascites in equivocal cases. Peripheral edema can occur with or without ascites. In patients with advanced liver disease, other factors frequently contribute to edema formation, including hypoalbuminemia, venous insufficiency, heart failure, and medications.

Hepatic failure is defined as the occurrence of signs or symptoms of hepatic encephalopathy in a person with severe acute or chronic liver disease. The first signs of hepatic encephalopathy can be subtle and nonspecific—change in sleep patterns, change in personality, irritability, and mental dullness. Thereafter, confusion, disorientation, stupor, and eventually coma supervene. In acute liver failure, excitability and mania may be present. Physical findings include asterixis and flapping tremors of the body and tongue. *Fetor hepaticus* refers to the slightly sweet, ammoniacal odor that can develop in patients with liver failure, particularly if

there is portal-venous shunting of blood around the liver. Other causes of coma and disorientation should be excluded, mainly electrolyte imbalances, sedative use, and renal or respiratory failure. The appearance of hepatic encephalopathy during acute hepatitis is the major criterion for diagnosis of fulminant hepatitis and indicates a poor prognosis. In chronic liver disease, encephalopathy is usually triggered by a medical complication such as gastrointestinal bleeding, over-diuresis, uremia, dehydration, electrolyte imbalance, infection, constipation, or use of narcotic analgesics.

A helpful measure of hepatic encephalopathy is a careful mental-status examination and use of the trail-making test, which consists of a series of 25 numbered circles that the patient is asked to connect as rapidly as possible using a pencil. The normal range for the connect-the-dot test is 15–30 sec; it is considerably longer in patients with early hepatic encephalopathy. Other tests include drawing of abstract objects or comparison of a signature to previous examples. More sophisticated testing—for example, with electroencephalography and visual evoked potentials—can detect mild forms of encephalopathy but are rarely clinically useful.

Other signs of advanced liver disease include umbilical hernia from ascites, hydrothorax, prominent veins over the abdomen, and *caput medusa*, a condition that consists of collateral veins radiating from the umbilicus and results from recanulation of the umbilical vein. Widened pulse pressure and signs of a hyperdynamic circulation can occur in patients with cirrhosis as a result of fluid and sodium retention, increased cardiac output, and reduced peripheral resistance. Patients with long-standing cirrhosis and portal hypertension are prone to develop the hepatopulmonary syndrome, which is defined by the triad of liver disease, hypoxemia, and pulmonary arteriovenous shunting. The hepatopulmonary syndrome is characterized by platypnea and orthodeoxia: shortness of breath and oxygen desaturation that occur paradoxically upon the assumption of an upright position. Measurement of oxygen saturation by pulse oximetry is a reliable screening test for hepatopulmonary syndrome.

Several skin disorders and changes are common in liver disease. Hyperpigmentation is typical of advanced chronic cholestatic diseases such as primary biliary cholangitis and sclerosing cholangitis. In these same conditions, xanthelasma and tendon xanthomata occur as a result of retention and high serum levels of lipids and cholesterol. Slate-gray pigmentation of the skin is also seen with hemochromatosis if iron levels are high for a prolonged period. Mucocutaneous vasculitis with palpable purpura, especially on the lower extremities, is typical of cryoglobulinemia of chronic hepatitis C but can also occur in chronic hepatitis B.

Some physical signs point to specific liver diseases. Kayser-Fleischer rings occur in Wilson disease and consist of a golden-brown copper pigment deposited in Descemet’s membrane at the periphery of the cornea; they are best seen by slit-lamp examination. Dupuytren contracture and parotid enlargement are suggestive of chronic alcoholism and alcoholic liver disease. In metastatic liver disease or primary hepatocellular carcinoma, signs of cachexia and wasting as well as firm hepatomegaly and a hepatic bruit may be prominent.

DIAGNOSIS OF LIVER DISEASE

The major causes of liver disease and key diagnostic features are outlined in **Table 329-3**, and an algorithm for evaluation of the patient with suspected liver disease is shown in **Fig. 329-1**. Specifics of diagnosis are discussed in later chapters. The most common causes of acute liver disease are viral hepatitis (particularly hepatitis A, B, and C), drug-induced liver injury, cholangitis, and alcoholic liver disease. Liver biopsy usually is not needed in the diagnosis and management of acute liver disease, exceptions being situations where the diagnosis remains unclear despite thorough clinical and laboratory investigation. Liver biopsy can be helpful in diagnosing drug-induced liver disease and acute alcoholic hepatitis.

TABLE 329-3

Important Diagnostic Tests in Common Liver Diseases

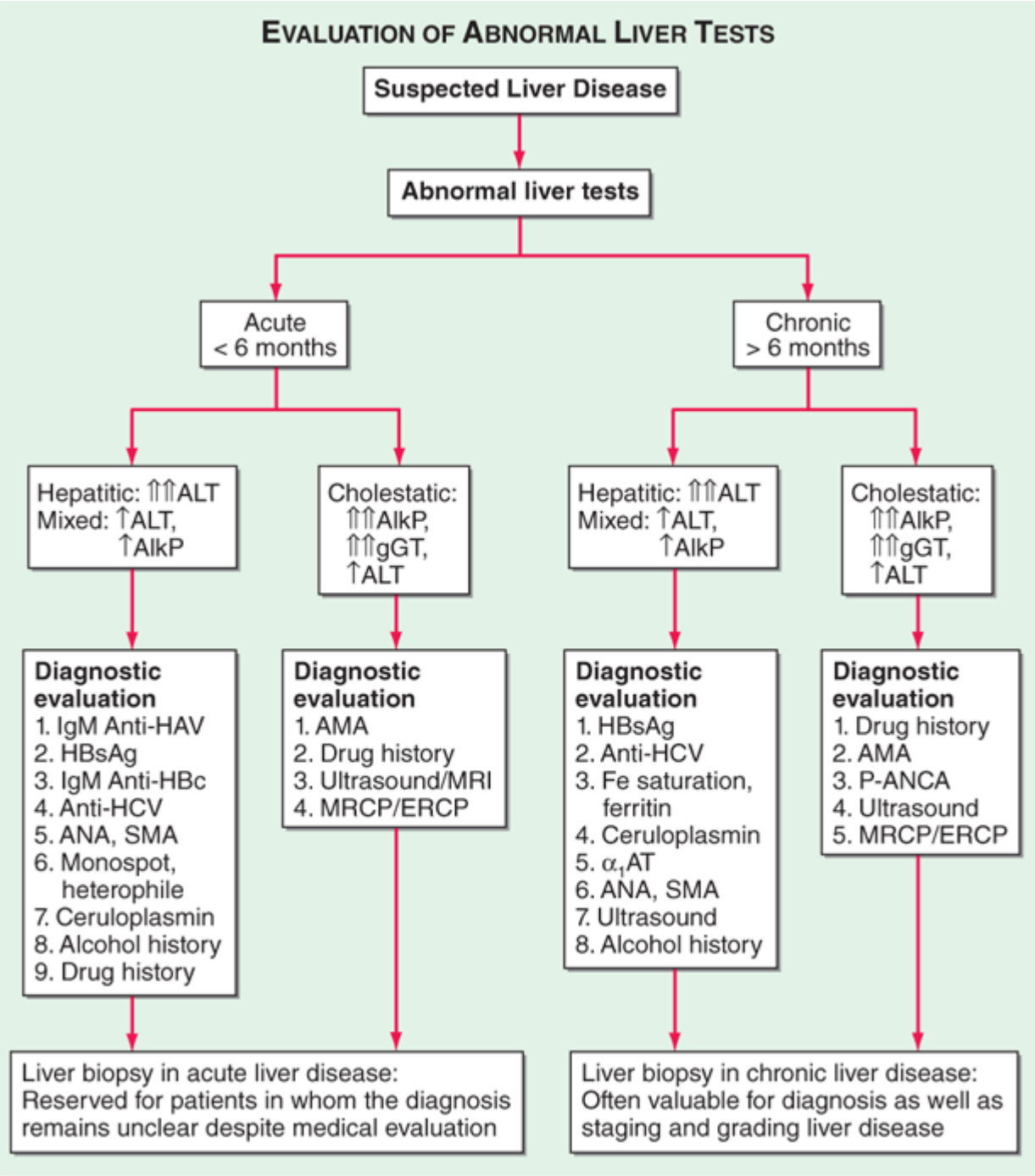
Disease	Diagnostic Test
Hepatitis A	Anti-HAV IgM
Hepatitis B	
Acute	HBsAg and anti-HBc IgM
Chronic	HBsAg and HBeAg and/or HBV DNA
Hepatitis C	Anti-HCV and HCV RNA
Hepatitis D (delta)	HBsAg and anti-HDV
Hepatitis E	Anti-HEV IgM and HEV RNA
Autoimmune hepatitis	ANA or SMA, elevated IgG levels, and compatible histology
Primary biliary cholangitis	Mitochondrial antibody, elevated IgM levels, and compatible histology
Primary sclerosing cholangitis	P-ANCA, cholangiography
Drug-induced liver disease	History of drug ingestion
Alcoholic liver disease	History of excessive alcohol intake and compatible histology
Nonalcoholic steatohepatitis	Ultrasound or CT evidence of fatty liver and compatible histology
α_1 antitrypsin disease	Reduced α_1 antitrypsin levels, phenotype PiZZ or PiSZ
Wilson disease	Decreased serum ceruloplasmin and increased urinary copper; increased hepatic copper level

Disease	Diagnostic Test
Hemochromatosis	Elevated iron saturation and serum ferritin; genetic testing for *HFE* gene mutations
Hepatocellular cancer	Elevated α-fetoprotein level (to >500 ng/mL); ultrasound or CT image of mass

Abbreviations: ANA, antinuclear antibody; anti-HBc, antibody to hepatitis B core (antigen); HAV, HBV, HCV, HDV, HEV: hepatitis A, B, C, D, E virus; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; P-ANCA, peripheral antineutrophil cytoplasmic antibody; SMA, smooth-muscle antibody.

FIGURE 329-1

Algorithm for evaluation of abnormal liver tests. For patients with suspected liver disease, an appropriate approach to evaluation is initial routine liver testing—for example, measurement of serum bilirubin, albumin, alanine aminotransferase (ALT), AST, and AlkP. These results (sometimes complemented by testing of γ-glutamyl transpeptidase; gGT) will establish whether the pattern of abnormalities is hepatic, cholestatic, or mixed. In addition, the duration of symptoms or abnormalities will indicate whether the disease is acute or chronic. If the disease is acute and if history, laboratory tests, and imaging studies do not reveal a diagnosis, liver biopsy is appropriate to help establish the diagnosis. If the disease is chronic, liver biopsy can be helpful not only for diagnosis but also for grading of the activity and staging the progression of disease. This approach is generally applicable to patients without immune deficiency. In patients with HIV infection or recipients of bone marrow or solid organ transplants, the diagnostic evaluation should also include evaluation for opportunistic infections (e.g., with adenovirus, cytomegalovirus, *Coccidioides*, hepatitis E virus) as well as for vascular and immunologic conditions (venoocclusive disease, graft-versus-host disease). α_1 AT, α_1 antitrypsin; AMA; antimitochondrial antibody; ANA, antinuclear antibody; anti-HBc, antibody to hepatitis B core (antigen); ERCP, endoscopic retrograde cholangiopancreatography; HAV, hepatitis A virus; HBsAg, hepatitis B surface antigen; HCV, hepatitis C virus; MRCP, magnetic resonance cholangiopancreatography; P-ANCA, peripheral antineutrophil cytoplasmic antibody; SMA, smooth-muscle antibody.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The most common causes of chronic liver disease, in general order of frequency, are chronic hepatitis C, alcoholic liver disease, nonalcoholic steatohepatitis, chronic hepatitis B, autoimmune hepatitis, sclerosing cholangitis, primary biliary cholangitis, hemochromatosis, and Wilson disease. Hepatitis E virus is a rare cause of chronic hepatitis, with cases occurring mostly in persons who are immunosuppressed or immunodeficient. Strict diagnostic criteria have not been developed for most liver diseases, but liver biopsy plays an important role in the diagnosis of autoimmune hepatitis, primary biliary cholangitis, nonalcoholic and alcoholic steatohepatitis, and Wilson disease (with a quantitative hepatic copper level in the last instance).

Laboratory Testing

Diagnosis of liver disease is greatly aided by the availability of reliable and sensitive tests of liver injury and function. A typical battery of blood tests used for initial assessment of liver disease includes measurement of levels of serum alanine (ALT) and aspartate aminotransferases (AST), alkaline phosphatase (AlkP), direct and total serum bilirubin and albumin, and prothrombin time. The pattern of abnormalities generally points to hepatocellular versus cholestatic liver disease and helps determine whether the disease is acute or chronic and whether cirrhosis and hepatic failure are present. On the basis of these results, further testing over time may be necessary. Other laboratory tests may be helpful, such as γ-glutamyl transpeptidase (γGT) to define whether AlkP elevations are due to liver disease; hepatitis serology to define the type of viral hepatitis; and autoimmune markers to diagnose primary biliary cholangitis (antimitochondrial antibody), sclerosing cholangitis (peripheral antineutrophil cytoplasmic antibody), and autoimmune hepatitis (antinuclear, smooth-muscle, and liver-kidney microsomal antibody). A simple delineation of laboratory abnormalities and common liver diseases is given in Table 329-3.

The use and interpretation of liver function tests are summarized in Chap. 330.

Diagnostic Imaging

Great advances have been made in hepatobiliary imaging, although no method is adequately accurate in demonstrating underlying cirrhosis in its early stages. Of the many modalities available for imaging the liver, US, computerized tomography (CT), and magnetic resonance imaging (MRI) are the most commonly employed and are complementary to one another. In general, US and CT are highly sensitive for detecting biliary duct dilation and are the first-line options for investigating cases of suspected obstructive jaundice. All three modalities can detect a fatty liver, which appears bright on imaging studies. Modifications of CT and MRI can be used to quantify liver fat, and this information may ultimately be valuable in monitoring therapy in patients with fatty liver disease. Magnetic resonance cholangiopancreatography (MRCP) and endoscopic retrograde cholangiopancreatography (ERCP) are the procedures of choice for visualization of the biliary tree. MRCP offers several advantages over ERCP: there is no need for contrast media or ionizing radiation, images can be acquired faster, the procedure is less operator dependent, and it carries no risk of pancreatitis. MRCP is superior to US and CT for detecting choledocholithiasis but is less specific. MRCP is useful in the diagnosis of bile duct obstruction and congenital biliary abnormalities, but ERCP is more valuable in evaluating ampullary lesions and primary sclerosing cholangitis. ERCP permits biopsy, direct visualization of the ampulla and common bile duct, and intraductal ultrasonography and brushings for cytological evaluation of malignancy. It also provides several therapeutic options in patients with obstructive jaundice, such as sphincterotomy, stone extraction, and placement of nasobiliary catheters and biliary stents. Doppler US and MRI are used to assess hepatic vasculature and hemodynamics and to monitor surgically or radiologically placed vascular shunts, including transjugular intrahepatic portosystemic shunts. Multidetector or spiral CT and MRI with contrast-enhancement are the procedures of choice for the identification and evaluation of hepatic masses, the staging of liver tumors, and preoperative assessment. With regard to mass lesions, the sensitivity of hepatic imaging continues to increase; unfortunately,

specificity remains a problem, and often two and sometimes three studies are needed before a diagnosis can be reached. An emerging imaging modality for the investigation of hepatic lesions is contrast-enhanced US. This procedure permits enhancement of liver lesions in a similar fashion as contrast-enhanced, cross-sectional CT, or MR imaging. Major advantages are real-time assessment of liver perfusion throughout the vascular phases without risk of nephrotoxicity and radiation exposure. Other advantages are its widespread availability and lower cost. Limitations include body habitus of the patient and skill of the operator. US is the recommended modality for hepatocellular carcinoma screening. Contrast-enhanced US, CT, and MRI are appropriate for further investigation of lesions detected on screening US. The American College of Radiologists has developed a Liver Imaging Reporting and Data System (LI-RADS) to standardize the reporting and data collection of CT, MR, and contrast-enhanced US imaging for hepatocellular carcinoma (HCC). This system allows for more consistent reporting and reduces imaging interpretation variability and errors.

Recently, US transient elastography has been approved for the measurement of hepatic stiffness—providing an indirect assessment of fibrosis and cirrhosis; this technique can eliminate the need for liver biopsy if the only indication is the assessment of disease stage. MR elastography is more sensitive than US elastography, but is also more expensive and requires advanced scheduling and special equipment. Studies are ongoing to determine whether hepatic elastography is an appropriate means of monitoring fibrosis and disease progression. Finally, interventional radiologic techniques allow the biopsy of solitary lesions, the radiofrequency ablation and chemoembolization of cancerous lesions, the insertion of drains into hepatic abscesses, the measurement of portal pressure, and the creation of vascular shunts in patients with portal hypertension. Which modality to use depends on factors such as availability, cost, and experience of the radiologist with each technique.

Liver Biopsy

Liver biopsy remains the gold standard in the evaluation of patients with liver disease, particularly chronic liver disease. Liver biopsy is necessary for diagnosis in selected instances but is more often useful for assessment of the severity (grade) and stage of liver damage, prediction of prognosis, and monitoring of the response to treatment. The size of the liver biopsy sample is an important determinant of reliability; a length of 1.5–2 cm is necessary for accurate assessment of fibrosis. Because liver biopsy is an invasive procedure and not without complications, it should be used only when it will contribute materially to decisions about management and therapy. In the future, noninvasive means of assessing disease activity (batteries of blood tests) and fibrosis (elastography and fibrosis markers) may replace liver biopsy for the staging and grading of disease.

GRADING AND STAGING OF LIVER DISEASE

Grading refers to an assessment of the severity or activity of liver disease, whether acute or chronic; active or inactive; and mild, moderate, or severe. Liver biopsy is the most accurate means of assessing severity, particularly in chronic liver disease. Serum aminotransferase levels serve as convenient and noninvasive markers for disease activity but do not always reliably reflect disease severity. Thus, normal serum

aminotransferase levels in patients with hepatitis B surface antigen in serum may indicate the inactive carrier state or may reflect mild chronic hepatitis B or hepatitis B with fluctuating disease activity. Serum testing for hepatitis B e antigen and hepatitis B virus DNA can help sort out these different patterns, but these markers can also fluctuate and change over time. Similarly, in chronic hepatitis C, serum aminotransferase levels can be normal despite moderate disease activity. Finally, in both alcoholic and nonalcoholic steatohepatitis, aminotransferase levels are quite unreliable in reflecting severity. In these conditions, liver biopsy is helpful in guiding management and identifying appropriate therapy, particularly if treatment is difficult, prolonged, and expensive, as is often the case in chronic viral hepatitis. Of the several well-verified numerical scales for grading activity in chronic liver disease, the most commonly used are the METAVIR, histology activity index and the Ishak fibrosis scale.

Liver biopsy is also the most accurate means of assessing stage of disease as early or advanced, precirrhotic, and cirrhotic. Staging of disease pertains largely to chronic liver diseases in which progression to cirrhosis and end-stage disease can occur but may require years or decades. Clinical features, biochemical tests, and hepatic imaging studies are helpful in assessing stage but generally become abnormal only in the middle to late stages of cirrhosis. Noninvasive tests that suggest advanced fibrosis include mild elevations of bilirubin, prolongation of prothrombin time, slight decreases in serum albumin, and mild thrombocytopenia (which is often the first indication of worsening fibrosis). Combinations of blood test results that include clinical features, routine laboratory tests, and special laboratory tests such as serum proteins or small molecules that are affected by or involved with fibrogenesis have been used to create models for predicting advanced liver disease, but these models are not reliable enough to use on a regular basis or for repeated measures and only separate advanced from early disease (Table 329-4). Recently, elastography and noninvasive breath tests using ^{13}C-labeled compounds have been proposed as a means of detecting early stages of fibrosis and liver dysfunction, but their reliability and reproducibility remain to be proven. A major limitation of noninvasive markers is that they can be affected by disease activity. Even elastography is limited in this regard, in that it measures liver stiffness, not fibrosis per se, and can be affected by inflammation, edema, hepatocyte necrosis, and intrasinusoidal cellularity (inflammatory, malignant, or sickled cells). Thus, at present, mild to moderate stages of hepatic fibrosis are detectable only by liver biopsy. In the assessment of stage, the degree of fibrosis is usually used as the quantitative measure. The amount of fibrosis is generally staged on a scale of 0 to 4+ (METAVIR scale) or 0 to 6+ (Ishak scale). The importance of staging relates primarily to prognosis, recommendation of therapy and to optimal management of complications. Patients with cirrhosis are candidates for screening and surveillance for esophageal varices and hepatocellular carcinoma. Patients without advanced fibrosis need not undergo screening.

TABLE 329-4

Selected Noninvasive Methods of Assessing Hepatic Fibrosis and Cirrhosis

Method	Parameters	Advanced Fibrosis	Cirrhosis
APRI	AST, platelet count	>1	>1.5 (1–2)
ELF	Age, hyaluronic acid, MMP-3, TIMP-1	>7.7	>9.3
FIB-4	Age, AST, ALT, platelet count	>1.45	>3.25
*Fibro Test	Haptoglobin, α2-macroglobulin, apolipoprotein A1, γGT, total bilirubin	>0.45	>0.63
TE	Measures speed of a shear wave generated by vibration through liver tissue	>7.3 kPa	>15 kPa (9–26.5 kPa)
ARFI	Measures speed of shear wave generated by acoustic radiation force through liver tissue	>1.3 m/s	>1.87 m/s

*Patented models.

Note: The cutpoints presented in the table were mostly derived from patients with chronic hepatitis C. The cutpoints for the non-invasive models and tests presented in the table varies among different liver disease and among patients with the same disease among different populations.

Abbreviations: ALT, alanine aminotransferase; APRI, AST-to-Platelet Ratio; ARFI, acoustic radiation force imaging; AST, aspartate aminotransferase; ELF, Enhanced Liver Fibrosis Panel; γGT, gamma glutamyl-transpeptidase; MMP-3, metalloproteinase-3; TIMP-1, tissue inhibitor of metalloproteinase-1; TE, Transient Elastography.

Once cirrhosis develops, other scoring systems are employed to assess compensated versus decompensated disease and prognosis. The initial staging system used for this purpose was the modified Child-Pugh classification, with a scoring system of 5–15: scores of 5 and 6 represent Child-Pugh class A (consistent with "compensated cirrhosis"), scores of 7–9 represent class B, and scores of 10–15 represent class C **(Table 329-5)**. This scoring system was initially devised to stratify patients with cirrhosis into risk groups before portal decompressive surgery. The Child-Pugh score is a reasonably reliable predictor of survival in many liver diseases and predicts the likelihood of major complications of cirrhosis, such as bleeding from varices and spontaneous bacterial peritonitis. This classification scheme was used to assess prognosis in cirrhosis and to provide standard criteria for listing a patient as a candidate for liver transplantation (Child-Pugh class B). More recently, the Child-Pugh system has been replaced by the Model for End-Stage Liver Disease (MELD)

system for the latter purpose. The MELD score is a prospectively derived system designed to predict the prognosis of patients with liver disease and portal hypertension. Initially, this score was calculated from three noninvasive variables: the prothrombin time expressed as the international normalized ratio (INR), the serum bilirubin level, and the serum creatinine concentration. The ability of the MELD score to predict outcome after liver transplantation is regularly monitored and was modified to increase its accuracy and improve allocation of donated livers. These modifications include serum sodium concentration as a factor in the model and a reweighting of the MELD components. A separate scoring system pediatric end-stage liver disease (PELD) is used for children (<12 years of age). Transient elastography has also been used to stage cirrhosis and has been shown to be useful in predicting complications such as variceal hemorrhage, ascites development and liver-related death.

TABLE 329-5

Child-Pugh Classification of Cirrhosis

Factor	Points Toward Total Score			
	Units	**1**	**2**	**3**
Serum bilirubin	μmol/L	<34	34–51	>51
	mg/dL	<2.0	2.0–3.0	>3.0
Serum albumin	g/L	>35	30–35	<30
	g/dL	>3.5	3.0–3.5	<3.0
Prothrombin time	seconds prolonged	<4	4–6	>6
	INR[a]	<1.7	1.7–2.3	>2.3
Ascites		None	Easily controlled	Poorly controlled
Hepatic encephalopathy		None	Minimal	Advanced

[a]International normalized ratio.

Note: The Child-Pugh score is calculated by adding the scores for the five factors and can range from 5 to 15. The resulting Child-Pugh class can be A (a score of 5–6), B (7–9), or C (≥10). Decompensation indicates cirrhosis, with a Child-Pugh score of ≥7 (class B). This level has been the accepted criterion for listing a patient for liver transplantation.

The MELD system provides a more objective means of assessing disease severity and has less center-to-center variation than the Child-Pugh score as well as a wider range of values. The MELD and PELD systems are currently used to establish priority listing for liver transplantation in the United States. Convenient MELD and PELD calculators are available via the internet: *(http://optn.transplant.hrsa.gov/resources/MeldPeldCalculator.asp?index=98)*.

NONSPECIFIC ISSUES IN THE MANAGEMENT OF PATIENTS WITH LIVER DISEASE

Specifics on the management of different forms of acute or chronic liver disease are supplied in subsequent chapters, but certain issues are applicable to any patient with liver disease. These issues include advice regarding alcohol use, medication use, vaccination, and surveillance for complications of liver disease. Alcohol should be used sparingly, if at all, by patients with liver disease. Abstinence from alcohol should be encouraged for all patients with alcohol-related liver disease, patients with cirrhosis, and patients receiving interferon-based therapy for hepatitis B and during antiviral therapy of hepatitis C. With regard to vaccinations, all patients with liver disease should receive hepatitis A vaccine, and those with risk factors should receive hepatitis B vaccine as well. Influenza and pneumococcal vaccination should also be encouraged, with adherence to the recommendations of the Centers for Disease Control and Prevention. Patients with liver disease should exercise caution in using any medications other than those that are most necessary. Drug-induced hepatotoxicity can mimic many forms of liver disease and can cause exacerbations of chronic hepatitis and cirrhosis; drugs should be suspected in any situation in which the cause of exacerbation is unknown. Finally, consideration should be given to surveillance for complications of chronic liver disease such as variceal hemorrhage and hepatocellular carcinoma. Cirrhosis warrants upper endoscopy to assess the presence of varices, and the patient should receive chronic therapy with beta blockers or should be offered endoscopic obliteration if large varices are found. Moreover, cirrhosis warrants screening and long-term surveillance for development of hepatocellular carcinoma. While the optimal regimen for such surveillance has not been established, an appropriate approach is US of the liver at 6- to 12-month intervals.

FURTHER READING

Patel K, Bedossa P, Castera L: Diagnosis of liver fibrosis: Present and future. Semin Liver Dis 35:166, 2015. [PubMed: 25974902]

Chapter 330: Evaluation of Liver Function

Daniel S. Pratt

FIGURE 330-1

INTRODUCTION

There are a number of tests that can be used to evaluate liver function. These tests include biochemical tests, radiologic tests, and pathologic tests.

Serum biochemical tests, also commonly referred to as “liver function tests,” can be used to (1) detect the presence of liver disease, (2) distinguish among different types of liver disorders, (3) gauge the extent of known liver damage, and (4) follow the response to treatment. However, serum biochemical tests have shortcomings. They lack sensitivity and specificity; they can be normal in patients with serious liver disease and abnormal in patients with diseases that do not affect the liver. Liver tests rarely suggest a specific diagnosis; rather, they suggest a general category of liver disease, such as hepatocellular or cholestatic, which then further directs the evaluation. The liver carries out thousands of biochemical functions, most of which cannot be easily measured by blood tests. Laboratory tests measure only a limited number of these functions. In fact, many tests, such as the aminotransferases and alkaline phosphatase, do not measure liver function at all. Rather, they detect liver cell damage or interference with bile flow. Thus, no one biochemical test enables the clinician to accurately assess the liver’s total functional capacity.

To increase the sensitivity and the specificity of biochemical tests in the detection of liver disease, it is best to use them as a battery. Tests usually employed in clinical practice include the bilirubin, aminotransferases, alkaline phosphatase, albumin, and prothrombin time tests. When more than one of these tests provide abnormal findings or the findings are persistently abnormal on serial determinations, the probability of liver disease is high. When all test results are normal, the probability of missing occult liver disease is low.

Serum Bilirubin

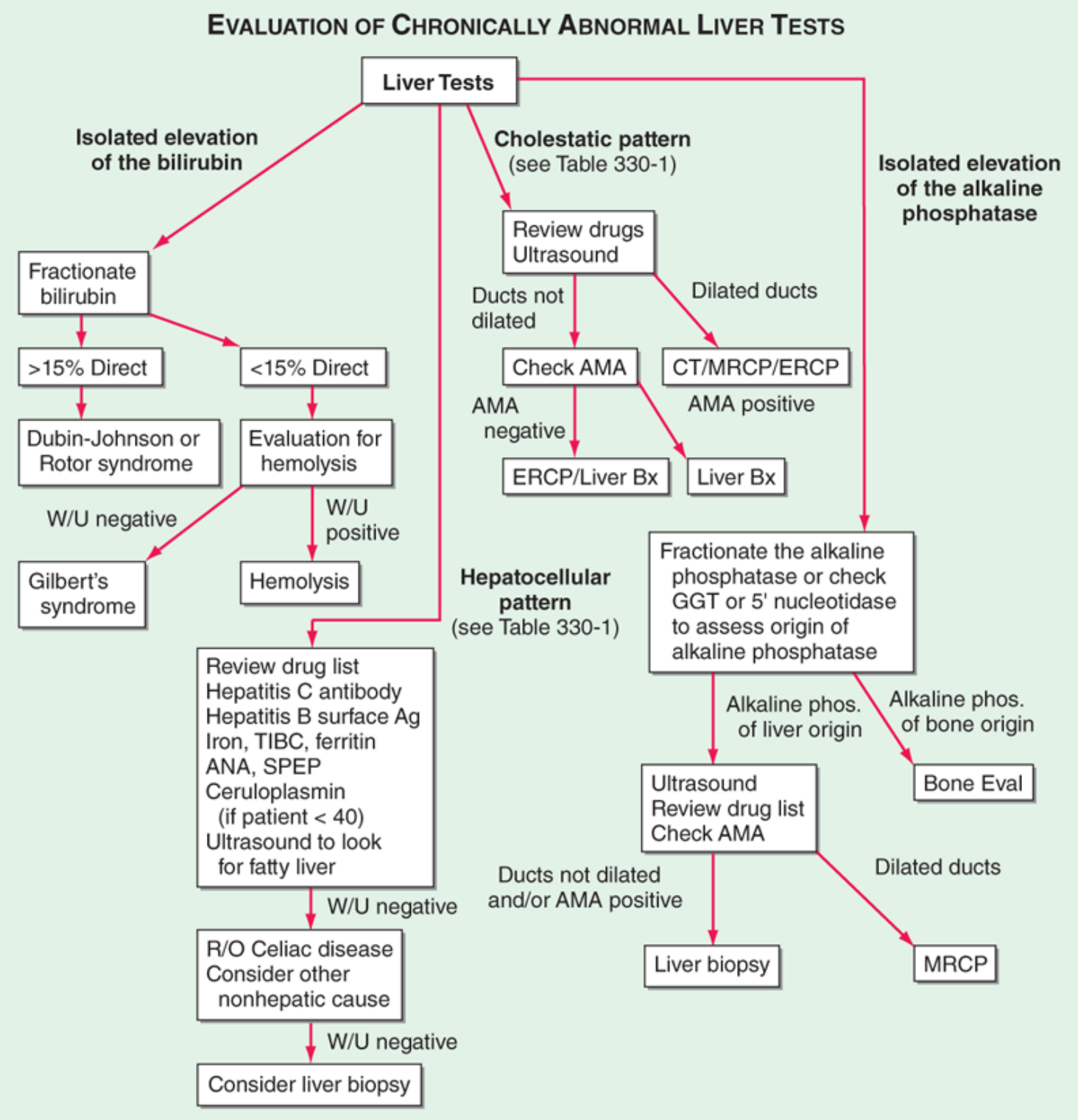
Bilirubin, a breakdown product of the porphyrin ring of heme-containing proteins, is found in the blood in two fractions—conjugated and unconjugated **(See also Chap. 45)**. The unconjugated fraction, also termed the *indirect fraction*, is insoluble in water and is bound to albumin in the blood. The conjugated (direct) bilirubin fraction is water-soluble and can therefore be excreted by the kidney. Normal values of total serum bilirubin are reported between 1 and 1.5 mg/dL with 95% of a normal population falling between 0.2 and 0.9 mg/dL. If the direct-acting fraction is <15% of the total, the bilirubin can be considered to all be indirect. The most frequently reported upper limit of normal for conjugated bilirubin is 0.3 mg/dL.

Elevation of the unconjugated fraction of bilirubin is rarely due to liver disease. An isolated elevation of unconjugated bilirubin is seen primarily in hemolytic disorders and in a number of genetic conditions such as Crigler-Najjar and Gilbert’s syndromes **(Chap. 45)**. Isolated unconjugated hyperbilirubinemia (bilirubin elevated but <15% direct) should

prompt a workup for hemolysis (Fig. 330-1). In the absence of hemolysis, an isolated, unconjugated hyperbilirubinemia in an otherwise healthy patient can be attributed to Gilbert's syndrome, and no further evaluation is required.

FIGURE 330-1

Algorithm for the evaluation of chronically abnormal liver tests. AMA, antimitochondrial antibody; ANA, antinuclear antibody; Bx, biopsy; CT, computed tomography; ERCP, endoscopic retrograde cholangiopancreatography; GGT, γ glutamyl transpeptidase; MRCP, magnetic resonance cholangiopancreatography; R/O, rule out; SPEP, serum protein electrophoresis; TIBC, total iron-binding capacity; W/U, workup.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In contrast, conjugated hyperbilirubinemia almost always implies liver or biliary tract disease. The rate-limiting step in bilirubin metabolism is not conjugation of bilirubin, but rather the transport of conjugated bilirubin into the bile canaliculi. Thus, elevation of the conjugated fraction may be seen in any type of liver disease including fulminant liver failure. In most liver diseases, both conjugated and unconjugated fractions of the bilirubin tend to be elevated. Except in the presence of a purely unconjugated hyperbilirubinemia, fractionation of the bilirubin is rarely helpful in determining the cause of jaundice.

Although the degree of elevation of the serum bilirubin has not been critically assessed as a prognostic marker, it is important in a number of conditions. In viral hepatitis, the higher the serum bilirubin, the greater is the hepatocellular damage. Total serum bilirubin correlates with poor outcomes in alcoholic hepatitis. It is also a critical component of the Model for End-Stage Liver Disease (MELD) score, a tool used to estimate survival of patients with end-stage liver disease and assess operative risk of patients with cirrhosis. An elevated total serum bilirubin in patients with drug-induced liver disease indicates more severe injury.

Unconjugated bilirubin always binds to albumin in the serum and is not filtered by the kidney. Therefore, any bilirubin found in the urine is conjugated bilirubin; the presence of bilirubinuria implies the presence of liver disease. A urine dipstick test can theoretically give the same information as fractionation of the serum bilirubin. This test is almost 100% accurate. Phenothiazines may give a false-positive reading with the Ictotest tablet. In patients recovering from jaundice, the urine bilirubin clears prior to the serum bilirubin.

Serum Enzymes

The liver contains thousands of enzymes, some of which are also present in the serum in very low concentrations. These enzymes have no known function in the serum and behave like other serum proteins. They are distributed in the plasma and in interstitial fluid and have characteristic half-lives, which are usually measured in days. Very little is known about the catabolism of serum enzymes, although they are probably cleared by cells in the reticuloendothelial system. The elevation of a given enzyme activity in the serum is thought to primarily reflect its increased rate of entrance into serum from damaged liver cells.

Serum enzyme tests can be grouped into two categories: (1) enzymes whose elevation in serum reflects damage to hepatocytes and (2) enzymes whose elevation in serum reflects cholestasis.

ENZYMES THAT REFLECT DAMAGE TO HEPATOCYTES

The aminotransferases (transaminases) are sensitive indicators of liver cell injury and are most helpful in recognizing acute hepatocellular diseases such as hepatitis. They include aspartate aminotransferase (AST) and alanine aminotransferase (ALT). AST is found in the liver, cardiac muscle, skeletal muscle, kidneys, brain, pancreas, lungs, leukocytes, and erythrocytes in decreasing order of concentration. ALT is found primarily in the liver and is therefore a more specific indicator of liver injury. The aminotransferases are normally present in the serum in low concentrations. These enzymes are released into the blood in greater amounts when there is damage to the liver cell membrane resulting in increased permeability. Liver cell necrosis is not required for the release of the aminotransferases, and there is a poor correlation between the degree of liver cell damage and the level of the aminotransferases. Thus, the absolute elevation of the aminotransferases is of no prognostic significance in acute hepatocellular disorders.

The normal range for aminotransferases varies widely among laboratories, but generally ranges from 10 to 40 IU/L. The interlaboratory variation in normal range is due to technical reasons; no reference standards exist to establish upper limits of normal for ALT and AST. Some have recommended revisions of normal limits of the aminotransferases to adjust for sex and body mass index, but others have noted the potential costs and unclear benefits of implementing this change.

Any type of liver cell injury can cause modest elevations in the serum aminotransferases. Levels of up to 300 IU/L are nonspecific and may be found in any type of liver disorder. Minimal ALT elevations in asymptomatic blood donors rarely indicate severe liver disease; studies have shown that fatty liver disease is the most likely explanation. Striking elevations—that is, aminotransferases >1000 IU/L—occur almost exclusively in disorders associated with extensive

hepatocellular injury such as (1) viral hepatitis, (2) ischemic liver injury (prolonged hypotension or acute heart failure), or (3) toxin- or drug-induced liver injury.

The pattern of the aminotransferase elevation can be helpful diagnostically. In most acute hepatocellular disorders, the ALT is higher than or equal to the AST. Whereas the AST:ALT ratio is typically <1 in patients with chronic viral hepatitis and nonalcoholic fatty liver disease, a number of groups have noted that as cirrhosis develops, this ratio rises to >1. An AST:ALT ratio >2:1 is suggestive, whereas a ratio >3:1 is highly suggestive, of alcoholic liver disease. The AST in alcoholic liver disease is rarely >300 IU/L, and the ALT is often normal. A low level of ALT in the serum is due to an alcohol-induced deficiency of pyridoxal phosphate.

The aminotransferases are usually not greatly elevated in obstructive jaundice. One notable exception occurs during the acute phase of biliary obstruction caused by the passage of a gallstone into the common bile duct. In this setting, the aminotransferases can briefly be in the 1000–2000 IU/L range. However, aminotransferase levels decrease quickly, and the biochemical tests rapidly evolve into those typical of cholestasis.

ENZYMES THAT REFLECT CHOLESTASIS

The activities of three enzymes—alkaline phosphatase, 5′-nucleotidase, and γ-glutamyl transpeptidase (GGT)—are usually elevated in cholestasis. Alkaline phosphatase and 5′-nucleotidase are found in or near the bile canalicular membrane of hepatocytes, whereas GGT is located in the endoplasmic reticulum and in bile duct epithelial cells. Reflecting its more diffuse localization in the liver, GGT elevation in serum is less specific for cholestasis than are elevations of alkaline phosphatase or 5′-nucleotidase. Some have advocated the use of GGT to identify patients with occult alcohol use. Its lack of specificity makes its use in this setting questionable.

The normal serum alkaline phosphatase consists of many distinct isoenzymes found in the liver, bone, placenta, and, less commonly, in the small intestine. Patients over age 60 can have a mildly elevated alkaline phosphatase (1–1.5 times normal), whereas individuals with blood types O and B can have an elevation of the serum alkaline phosphatase after eating a fatty meal due to the influx of intestinal alkaline phosphatase into the blood. It is also elevated in children and adolescents undergoing rapid bone growth because of bone alkaline phosphatase, and late in normal pregnancies due to the influx of placental alkaline phosphatase.

Elevation of liver-derived alkaline phosphatase is not totally specific for cholestasis, and a less than threefold elevation can be seen in almost any type of liver disease. Alkaline phosphatase elevations greater than four times normal occur primarily in patients with cholestatic liver disorders, infiltrative liver diseases such as cancer and amyloidosis, and bone conditions characterized by rapid bone turnover (e.g., Paget's disease). In bone diseases, the elevation is due to increased amounts of the bone isoenzymes. In liver diseases, the elevation is almost always due to increased amounts of the liver isoenzyme.

If an elevated serum alkaline phosphatase is the only abnormal finding in an apparently healthy person, or if the degree of elevation is higher than expected in the clinical setting, identification of the source of elevated isoenzymes is helpful (Fig. 330-1). This problem can be approached in two ways. First, and most precise, is the fractionation of the alkaline phosphatase by electrophoresis. The second, best substantiated, and most available approach involves the measurement of serum 5′-nucleotidase or GGT. These enzymes are rarely elevated in conditions other than liver disease.

In the absence of jaundice or elevated aminotransferases, an elevated alkaline phosphatase of liver origin often, but not always, suggests early cholestasis and, less often, hepatic infiltration by tumor or granulomata. Other conditions

that cause isolated elevations of the alkaline phosphatase include Hodgkin's disease, diabetes, hyperthyroidism, congestive heart failure, amyloidosis, and inflammatory bowel disease.

The level of serum alkaline phosphatase elevation is not helpful in distinguishing between intrahepatic and extrahepatic cholestasis. There is essentially no difference among the values found in obstructive jaundice due to cancer, common duct stone, sclerosing cholangitis, or bile duct stricture. Values are similarly increased in patients with intrahepatic cholestasis due to drug-induced hepatitis, primary biliary cirrhosis, rejection of transplanted livers, and, rarely, alcohol-induced steatohepatitis. Values are also greatly elevated in hepatobiliary disorders seen in patients with AIDS (e.g., AIDS cholangiopathy due to cytomegalovirus or cryptosporidial infection and tuberculosis with hepatic involvement).

TESTS THAT MEASURE BIOSYNTHETIC FUNCTION OF THE LIVER

Serum Albumin

Serum albumin is synthesized exclusively by hepatocytes. Serum albumin has a long half-life: 18–20 days, with ~4% degraded per day. Because of this slow turnover, the serum albumin is not a good indicator of acute or mild hepatic dysfunction; only minimal changes in the serum albumin are seen in acute liver conditions such as viral hepatitis, drug-related hepatotoxicity, and obstructive jaundice. In hepatitis, albumin levels <3 g/dL should raise the possibility of chronic liver disease. Hypoalbuminemia is more common in chronic liver disorders such as cirrhosis and usually reflects severe liver damage and decreased albumin synthesis. One exception is the patient with ascites in whom synthesis may be normal or even increased, but levels are low because of the increased volume of distribution. However, hypoalbuminemia is not specific for liver disease and may occur in protein malnutrition of any cause, as well as protein-losing enteropathies, nephrotic syndrome, and chronic infections that are associated with prolonged increases in levels of serum interleukin 1 and/or tumor necrosis factor, cytokines that inhibit albumin synthesis. Serum albumin should not be measured for screening in patients in whom there is no suspicion of liver disease. A general medical clinic study of consecutive patients in whom no indications were present for albumin measurement showed that although 12% of patients had abnormal test results, the finding was of clinical importance in only 0.4%.

Serum Globulins

Serum globulins are a group of proteins made up of γ globulins (immunoglobulins) produced by B lymphocytes and α and β globulins produced primarily in hepatocytes. γ globulins are increased in chronic liver disease, such as chronic hepatitis and cirrhosis. In cirrhosis, the increased serum γ globulin concentration is due to the increased synthesis of antibodies, some of which are directed against intestinal bacteria. This occurs because the cirrhotic liver fails to clear bacterial antigens that normally reach the liver through the hepatic circulation.

Increases in the concentration of specific isotypes of γ globulins are often helpful in the recognition of certain chronic liver diseases. Diffuse polyclonal increases in IgG levels are common in autoimmune hepatitis; increases >100% should alert the clinician to this possibility. Increases in the IgM levels are common in primary biliary cirrhosis, whereas increases in the IgA levels occur in alcoholic liver disease.

COAGULATION FACTORS

With the exception of factor VIII, which is produced by vascular endothelial cells, the blood clotting factors are made exclusively in hepatocytes. Their serum half-lives are much shorter than albumin, ranging from 6 h for factor VII to 5

days for fibrinogen. Because of their rapid turnover, measurement of the clotting factors is the single best acute measure of hepatic synthetic function and helpful in both diagnosis and assessing the prognosis of acute parenchymal liver disease. Useful for this purpose is the *serum prothrombin time*, which collectively measures factors II, V, VII, and X. Biosynthesis of factors II, VII, IX, and X depends on vitamin K. The international normalized ratio (INR) is used to express the degree of anticoagulation on warfarin therapy. The INR standardizes prothrombin time measurement according to the characteristics of the thromboplastin reagent used in a particular lab, which is expressed as an International Sensitivity Index (ISI); the ISI is then used in calculating the INR.

The prothrombin time may be elevated in hepatitis and cirrhosis as well as in disorders that lead to vitamin K deficiency such as obstructive jaundice or fat malabsorption of any kind. Marked prolongation of the prothrombin time, >5 s above control and not corrected by parenteral vitamin K administration, is a poor prognostic sign in acute viral hepatitis and other acute and chronic liver diseases. The INR, along with the total serum bilirubin and creatinine, are components of the MELD score, which is used as a measure of hepatic decompensation and to allocate organs for liver transplantation.

OTHER DIAGNOSTIC TESTS

Although tests may direct the physician to a category of liver disease, additional biochemical testing, radiologic testing, and procedures are often necessary to make the proper diagnosis, as shown in Fig. 330-1. The most commonly used ancillary tests are reviewed here, as are the noninvasive tests available for assessing hepatic fibrosis.

Ammonia

Ammonia is produced in the body during normal protein metabolism and by intestinal bacteria, primarily those in the colon. The liver plays a role in the detoxification of ammonia by converting it to urea, which is excreted by the kidneys. Striated muscle also plays a role in detoxification of ammonia, where it is combined with glutamic acid to form glutamine. Patients with advanced liver disease typically have significant muscle wasting, which likely contributes to hyperammonemia. Some physicians use the blood ammonia for detecting encephalopathy or for monitoring hepatic synthetic function, although its use for either of these indications has problems. There is very poor correlation between either the presence or the severity of acute encephalopathy and elevation of blood ammonia; it can be occasionally useful for identifying occult liver disease in patients with mental status changes. There is also a poor correlation of the blood serum ammonia and hepatic function. The ammonia can be elevated in patients with severe portal hypertension and portal blood shunting around the liver even in the presence of normal or near-normal hepatic function. Elevated arterial ammonia levels have been shown to correlate with outcome in fulminant hepatic failure.

Liver Biopsy

Percutaneous biopsy of the liver is a safe procedure that can be easily performed at the bedside with local anesthesia and ultrasound guidance. Liver biopsy is of proven value in the following situations: (1) hepatocellular disease of uncertain cause, (2) prolonged hepatitis with the possibility of autoimmune hepatitis, (3) unexplained hepatomegaly, (4) unexplained splenomegaly, (5) hepatic lesions uncharacterized by radiologic imaging, (6) fever of unknown origin, and (7) staging of malignant lymphoma. Liver biopsy is most accurate in disorders causing diffuse changes throughout the liver and is subject to sampling error in focal disorders. Liver biopsy should not be the initial procedure in the diagnosis of cholestasis. The biliary tree should first be assessed for signs of obstruction. Contraindications to

performing a percutaneous liver biopsy include significant ascites and prolonged INR. Under these circumstances, the biopsy can be performed via the transjugular approach.

Noninvasive Tests to Detect Hepatic Fibrosis

Although liver biopsy is the standard for the assessment of hepatic fibrosis, noninvasive measures of hepatic fibrosis have been developed and show promise. These measures include multiparameter tests aimed at detecting and staging the degree of hepatic fibrosis and imaging techniques. FibroTest (marketed as FibroSure in the United States) is the best evaluated of the multiparameter blood tests. The test incorporates haptoglobin, bilirubin, GGT, apolipoprotein A-I, and α2-macroglobulin and has been found to have high positive and negative predictive values for diagnosing advanced fibrosis in patients with chronic hepatitis C, chronic hepatitis B, and alcoholic liver disease and patients taking methotrexate for psoriasis. Transient elastography (TE), marketed as FibroScan, and magnetic resonance elastography (MRE) both have gained U.S. Food and Drug Administration approval for use in the management of patients with liver disease. TE uses ultrasound waves to measure hepatic stiffness noninvasively. TE has been shown to be accurate for identifying advanced fibrosis in patients with chronic hepatitis C, primary biliary cirrhosis, hemochromatosis, nonalcoholic fatty liver disease, and recurrent chronic hepatitis after liver transplantation. MRE has been found to be superior to TE for staging liver fibrosis in patients with a variety of chronic liver diseases, but requires access to a magnetic resonance imaging scanner.

Ultrasonography

Ultrasonography is the first diagnostic test to use in patients whose liver tests suggest cholestasis, to look for the presence of a dilated intrahepatic or extrahepatic biliary tree or to identify gallstones. In addition, it shows space-occupying lesions within the liver, enables the clinician to distinguish between cystic and solid masses, and helps direct percutaneous biopsies. Ultrasound with Doppler imaging can detect the patency of the portal vein, hepatic artery, and hepatic veins and determine the direction of blood flow. This is the first test ordered in patients suspected of having Budd-Chiari syndrome.

USE OF LIVER TESTS

As previously noted, the best way to increase the sensitivity and specificity of laboratory tests in the detection of liver disease is to employ a battery of tests that includes the aminotransferases, alkaline phosphatase, bilirubin, albumin, and prothrombin time along with the judicious use of the other tests described in this chapter. **Table 330-1** shows how patterns of liver tests can lead the clinician to a category of disease that will direct further evaluation. However, it is important to remember that no single set of liver tests will necessarily provide a diagnosis. It is often necessary to repeat these tests on several occasions over days to weeks for a diagnostic pattern to emerge. Figure 330-1 is an algorithm for the evaluation of chronically abnormal liver tests.

TABLE 330-1

Liver Test Patterns in Hepatobiliary Disorders

Type of Disorder	Bilirubin	Aminotransferases	Alkaline Phosphatase	Albumin	Prothrombin Time
Hemolysis/Gilbert's syndrome	Normal to 86 µmol/L (5 mg/dL) 85% due to indirect fractions No bilirubinuria	Normal	Normal	Normal	Normal
Acute hepatocellular necrosis (viral and drug hepatitis, hepatotoxins, acute heart failure)	Both fractions may be elevated Peak usually follows aminotransferases Bilirubinuria	Elevated, often >500 IU, ALT > AST	Normal to <3× normal elevation	Normal	Usually normal. If >5× above control and not corrected by parenteral vitamin K, suggests poor prognosis
Chronic hepatocellular disorders	Both fractions may be elevated Bilirubinuria	Elevated, but usually <300 IU	Normal to <3× normal elevation	Often decreased	Often prolonged Fails to correct with parenteral vitamin K
Alcoholic hepatitis, cirrhosis	Both fractions may be elevated Bilirubinuria	AST:ALT >2 suggests alcoholic hepatitis or cirrhosis	Normal to <3× normal elevation	Often decreased	Often prolonged Fails to correct with parenteral vitamin K
Intra- and extrahepatic cholestasis	Both fractions may be elevated	Normal to moderate elevation	Elevated, often >4× normal elevation	Normal, unless chronic	Normal If prolonged, will correct with parenteral vitamin K

Type of Disorder	Bilirubin	Aminotransferases	Alkaline Phosphatase	Albumin	Prothrombin Time
(Obstructive jaundice) Infiltrative diseases (tumor, granulomata); partial bile duct obstruction	Bilirubinuria Usually normal	Rarely >500 IU Normal to slight elevation	Elevated, often >4× normal elevation Fractionate, or confirm liver origin with 5′-nucleotidase or γ glutamyl transpeptidase	Normal	Normal

GLOBAL CONSIDERATIONS

The tests and principles presented in this chapter are applicable worldwide. The causes of liver test abnormalities vary according to region. In developing nations, infectious diseases are more commonly the etiology of abnormal serum liver tests than in developed nations.

ACKNOWLEDGMENT

This chapter represents a revised version of a chapter in previous editions of Harrison's in which Marshall M. Kaplan was a co-author.

FURTHER READING

Kamath PS, Kim WR: The Model for End-Stage Liver Disease (MELD). Hepatology 45:797, 2007. [PubMed: 17326206]

Kaplan M: Alkaline phosphatase. Gastroenterology 62:452, 1972. [PubMed: 4551808]

Martínez SM et al: Noninvasive assessment of liver fibrosis. Hepatology 53:325, 2011. [PubMed: 21254180]

Prati D et al: Updated definitions of healthy ranges for serum alanine aminotransferase levels. Ann Intern Med 137:1, 2002. [PubMed: 12093239]

Chapter 331: The Hyperbilirubinemias

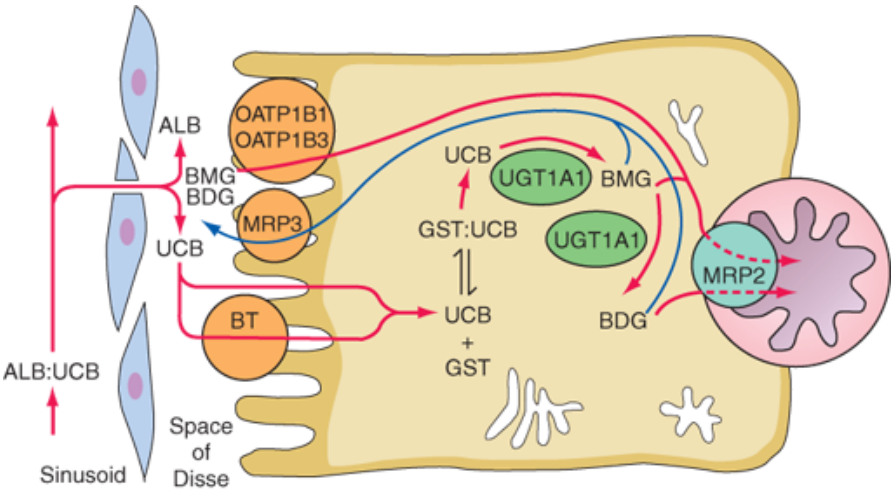
Allan W. Wolkoff

BILIRUBIN METABOLISM

The details of bilirubin metabolism are presented in Chap. 45. However, the hyperbilirubinemias are best understood in terms of perturbations of specific aspects of bilirubin metabolism and transport, and these will be briefly reviewed here as depicted in Fig. 331-1.

FIGURE 331-1

Hepatocellular bilirubin transport. Albumin-bound bilirubin in sinusoidal blood passes through endothelial cell fenestrae to reach the hepatocyte surface, entering the cell by both facilitated and simple diffusional processes. Within the cell, it is bound to glutathione-S-transferases and conjugated by bilirubin-UDP-glucuronosyltransferase (UGT1A1) to mono- and diglucuronides, which are actively transported across the canalicular membrane into the bile. In addition to this direct excretion of bilirubin glucuronides, a portion are transported into the portal circulation by MRP3 and subjected to reuptake into the hepatocyte by OATP1B1 and OATP1B3. ALB, albumin; BDG, bilirubin diglucuronide; BMG, bilirubin monoglucuronide; BT, proposed bilirubin transporter; GST, glutathione-S-transferase; MRP2 and MRP3, multidrug resistance–associated proteins 2 and 3; OATP1B1 and OATP1B3, organic anion transport proteins 1B1 and 1B3; UCB, unconjugated bilirubin; UGT1A1, bilirubin-UDP-glucuronosyltransferase.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Bilirubin is the end product of heme degradation. Some 70–90% of bilirubin is derived from degradation of the hemoglobin of senescent red blood cells. Bilirubin produced in the periphery is transported to the liver within the plasma, where, due to its insolubility in aqueous solutions, it is tightly bound to albumin. Under normal circumstances, bilirubin is removed from the circulation rapidly and efficiently by hepatocytes. Transfer of bilirubin from blood to bile involves four distinct but interrelated steps (Fig. 331-1).

1. *Hepatocellular uptake:* Uptake of bilirubin by the hepatocyte has carrier-mediated kinetics. Although a number of candidate bilirubin transporters have been proposed, the actual transporter remains elusive.

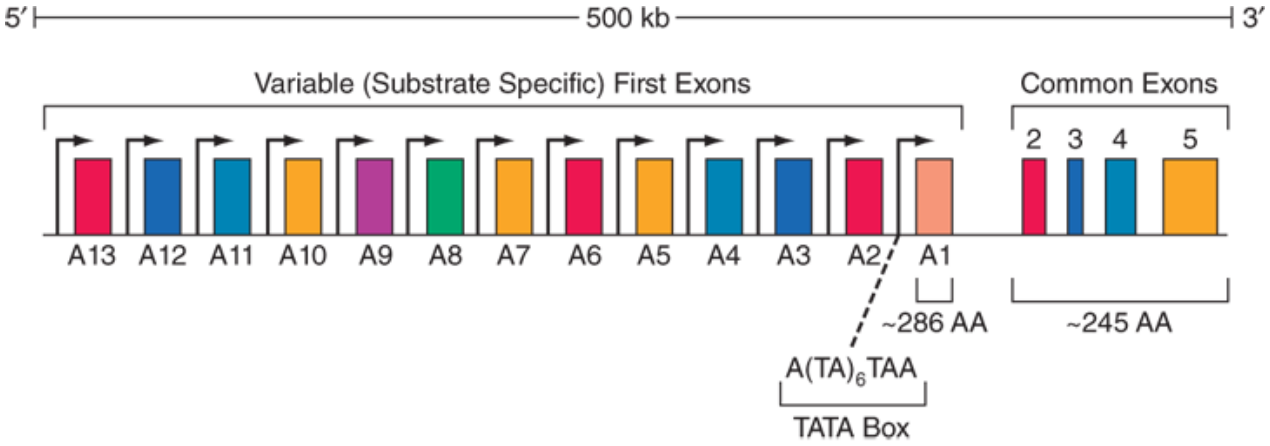
2. *Intracellular binding:* Within the hepatocyte, bilirubin is kept in solution by binding as a nonsubstrate ligand to several of the glutathione-S-transferases, formerly called ligandins.

3. *Conjugation:* Bilirubin is conjugated with one or two glucuronic acid moieties by a specific UDP-glucuronosyltransferase to form bilirubin mono- and diglucuronide, respectively. Conjugation disrupts the internal hydrogen bonding that limits aqueous solubility of bilirubin, and the resulting glucuronide conjugates are highly soluble in water. Conjugation is obligatory for excretion of bilirubin across the bile canalicular membrane into bile. The UDP-glucuronosyltransferases have been classified into gene families based on the degree of homology among the mRNAs for the various isoforms. Those that conjugate bilirubin and certain other substrates have been designated the *UGT1* family. These are expressed from a single gene complex by alternative promoter usage. This gene complex contains multiple substrate-specific first exons, designated A1, A2, etc. **(Fig. 331-2)**, each with its own promoter and each encoding the amino-terminal half of a specific isoform. In addition, there are four common exons (exons 2–5) that encode the shared carboxyl-terminal half of all of the *UGT1* isoforms. The various first exons encode the specific aglycone substrate binding sites for each isoform, while the shared exons encode the binding site for the sugar donor, UDP-glucuronic acid, and the transmembrane domain. Exon A1 and the four common exons, collectively designated as the UGT1A1 gene (Fig. 331-2), encode the physiologically critical enzyme bilirubin-UDP-glucuronosyltransferase (UGT1A1). A functional corollary of the organization of the *UGT1* gene is that a mutation in one of the first exons will affect only a single enzyme isoform. By contrast, a mutation in exons 2–5 will alter all isoforms encoded by the *UGT1* gene complex.

4. *Biliary excretion:* It has been thought until recently that bilirubin mono- and diglucuronides are excreted directly across the canalicular plasma membrane into the bile canaliculus by an ATP-dependent transport process mediated by a canalicular membrane protein called *multidrug resistance–associated protein 2* (MRP2). Mutations of MRP2 result in the Dubin-Johnson syndrome (see below). However, studies in patients with Rotor syndrome (see below) indicate that after formation, a portion of the glucuronides is transported into the portal circulation by a sinusoidal membrane protein called *multidrug resistance–associated protein 3* (MRP3) and is subjected to reuptake into the hepatocyte by the sinusoidal membrane uptake transporters *organic anion transport protein 1B1* (OATP1B1) and OATP1B3.

FIGURE 331-2

Structural organization of the human *UGT1* gene complex. This large complex on chromosome 2 contains at least 13 substrate-specific first exons (A1, A2, etc.). Since four of these are pseudogenes, nine UGT1 isoforms with differing substrate specificities are expressed. Each exon 1 has its own promoter and encodes the amino-terminal substrate-specific ~286 amino acids of the various *UGT1*-encoded isoforms, and common exons 2–5 that encode the 245 carboxyl-terminal amino acids common to all of the isoforms. mRNAs for specific isoforms are assembled by splicing a particular first exon such as the bilirubin-specific exon A1 to exons 2 to 5. The resulting message encodes a complete enzyme, in this particular case bilirubin-UDP-glucuronosyltransferase (UGT1A1). Mutations in a first exon affect only a single isoform. Those in exons 2–5 affect all enzymes encoded by the UGT1 complex.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EXTRAHEPATIC ASPECTS OF BILIRUBIN DISPOSITION

Bilirubin in the Gut

Following secretion into bile, conjugated bilirubin reaches the duodenum and passes down the gastrointestinal tract without reabsorption by the intestinal mucosa. An appreciable fraction is converted by bacterial metabolism in the gut to the water-soluble colorless compound urobilinogen. Urobilinogen undergoes enterohepatic cycling. Urobilinogen not taken up by the liver reaches the systemic circulation, from which some is cleared by the kidneys. Unconjugated bilirubin ordinarily does not reach the gut except in neonates or, by ill-defined alternative pathways, in the presence of severe unconjugated hyperbilirubinemia (e.g., Crigler-Najjar syndrome, type I [CN-I]). Unconjugated bilirubin that reaches the gut is partly reabsorbed, amplifying any underlying hyperbilirubinemia.

Renal Excretion of Bilirubin Conjugates

Unconjugated bilirubin is not excreted in urine, as it is too tightly bound to albumin for effective glomerular filtration and there is no tubular mechanism for its renal secretion. In contrast, the bilirubin conjugates are readily filtered at the glomerulus and can appear in urine in disorders characterized by increased bilirubin conjugates in the circulation. It should be kept in mind that the kidney can serve as an “overflow valve” for conjugated bilirubin. Consequently, the level of jaundice in individuals with conjugated hyperbilirubinemia can be amplified in the presence of renal failure.

DISORDERS OF BILIRUBIN METABOLISM LEADING TO UNCONJUGATED HYPERBILIRUBINEMIA

INCREASED BILIRUBIN PRODUCTION

Hemolysis

Increased destruction of erythrocytes leads to increased bilirubin turnover and unconjugated hyperbilirubinemia; the hyperbilirubinemia is usually modest in the presence of normal liver function. In particular, the bone marrow is only capable of a sustained eightfold increase in erythrocyte production in response to a hemolytic stress. Therefore, hemolysis alone cannot result in a sustained hyperbilirubinemia of more than ~68 µmol/L (4 mg/dL). Higher values imply concomitant hepatic dysfunction. When hemolysis is the only abnormality in an otherwise healthy individual, the result is a purely unconjugated hyperbilirubinemia, with the direct-reacting fraction as measured in a typical clinical laboratory being ≤15% of the total

serum bilirubin. In the presence of systemic disease, which may include a degree of hepatic dysfunction, hemolysis may produce a component of conjugated hyperbilirubinemia in addition to an elevated unconjugated bilirubin concentration. Prolonged hemolysis may lead to the precipitation of bilirubin salts within the gallbladder or biliary tree, resulting in the formation of gallstones in which bilirubin, rather than cholesterol, is the major component. Such pigment stones may lead to acute or chronic cholecystitis, biliary obstruction, or any other biliary tract consequence of calculous disease.

Ineffective Erythropoiesis

During erythroid maturation, small amounts of hemoglobin may be lost at the time of nuclear extrusion, and a fraction of developing erythroid cells is destroyed within the marrow. These processes normally account for a small proportion of bilirubin that is produced. In various disorders, including thalassemia major, megaloblastic anemias due to folate or vitamin B_{12} deficiency, congenital erythropoietic porphyria, lead poisoning, and various congenital and acquired dyserythropoietic anemias, the fraction of total bilirubin production derived from ineffective erythropoiesis is increased, reaching as much as 70% of the total. This may be sufficient to produce modest degrees of unconjugated hyperbilirubinemia.

Miscellaneous

Degradation of the hemoglobin of extravascular collections of erythrocytes, such as those seen in massive tissue infarctions or large hematomas, may lead transiently to unconjugated hyperbilirubinemia.

DECREASED HEPATIC BILIRUBIN CLEARANCE

Decreased Hepatic Uptake

Decreased hepatic bilirubin uptake is believed to contribute to the unconjugated hyperbilirubinemia of Gilbert's syndrome (GS), although the molecular basis for this finding remains unclear (see below). Several drugs, including flavaspidic acid, novobiocin, and rifampin, as well as various cholecystographic contrast agents, have been reported to inhibit bilirubin uptake. The resulting unconjugated hyperbilirubinemia resolves with cessation of the medication.

Impaired Conjugation

PHYSIOLOGIC NEONATAL JAUNDICE

Bilirubin produced by the fetus is cleared by the placenta and eliminated by the maternal liver. Immediately after birth, the neonatal liver must assume responsibility for bilirubin clearance and excretion. However, many hepatic physiologic processes are incompletely developed at birth. Levels of UGT1A1 are low, and alternative excretory pathways allow passage of unconjugated bilirubin into the gut. Since the intestinal flora that convert bilirubin to urobilinogen are also undeveloped, an enterohepatic circulation of unconjugated bilirubin ensues. As a consequence, most neonates develop mild unconjugated hyperbilirubinemia between days 2 and 5 after birth. Peak levels are typically <85–170 µmol/L (5–10 mg/dL) and decline to normal adult concentrations within 2 weeks, as mechanisms required for bilirubin disposition mature. Prematurity, often associated with more profound immaturity of hepatic function and hemolysis, can result in higher levels of unconjugated hyperbilirubinemia. A rapidly rising unconjugated bilirubin concentration, or absolute levels >340 µmol/L (20 mg/dL), puts the infant at risk for bilirubin encephalopathy, or kernicterus. Under these circumstances, bilirubin crosses an immature blood-brain barrier and precipitates in the basal ganglia and other areas of the brain. The consequences range from appreciable neurologic deficits to death. Treatment options include phototherapy, which converts bilirubin into water-soluble photoisomers that are excreted directly into bile, and exchange transfusion. The canalicular mechanisms responsible for bilirubin excretion are also immature at birth, and their maturation may lag behind that of UGT1A1; this can lead to transient conjugated neonatal hyperbilirubinemia, especially in infants with hemolysis.

ACQUIRED CONJUGATION DEFECTS

A modest reduction in bilirubin conjugating capacity may be observed in advanced hepatitis or cirrhosis. However, in this setting, conjugation is better preserved than other aspects of bilirubin disposition, such as canalicular excretion. Various drugs, including pregnanediol, novobiocin, chloramphenicol, gentamicin, and atazanavir may produce unconjugated hyperbilirubinemia by inhibiting UGT1A1 activity. Bilirubin conjugation may be inhibited by certain fatty acids that are present in breast milk, but not serum of mothers whose infants have excessive neonatal hyperbilirubinemia (*breast milk jaundice*). Alternatively, there may be increased enterohepatic circulation of bilirubin in these infants. The pathogenesis of breast milk jaundice appears to differ from that of transient familial neonatal hyperbilirubinemia (Lucey-Driscoll syndrome), in which there may be a UGT1A1 inhibitor in maternal serum.

HEREDITARY DEFECTS IN BILIRUBIN CONJUGATION

Three familial disorders characterized by differing degrees of unconjugated hyperbilirubinemia have long been recognized. The defining clinical features of each are described below **(Table 331-1)**. While these disorders have been recognized for decades to reflect differing degrees of deficiency in the ability to conjugate bilirubin, recent advances in the molecular biology of the *UGT1* gene complex have elucidated their interrelationships and clarified previously puzzling features.

TABLE 331-1

Principal Differential Characteristics of Gilbert's and Crigler-Najjar Syndromes

Feature	Crigler-Najjar Syndrome		Gilbert's Syndrome
	Type I	Type II	
Total serum bilirubin, μmol/L (mg/dL)	310–755 (usually >345) (18–45 [usually >20])	100–430 (usually ≤345) (6–25 [usually ≤20])	Typically ≤70 μmol/L (≤4 mg/dL) in absence of fasting or hemolysis
Routine liver tests Response to phenobarbital Kernicterus Hepatic histology	Normal None Usual Normal	Normal Decreases bilirubin by >25% Rare Normal	Normal Decreases bilirubin to normal No Usually normal; increased lipofuscin pigment in some
Bile characteristics Color Bilirubin fractions	Pale or colorless >90% unconjugated	Pigmented Largest fraction (mean: 57%) monoconjugates	Normal dark color Mainly diconjugates but monoconjugates increased (mean: 23%)
Bilirubin UDP-glucuronosyltransferase activity Inheritance (all autosomal)	Typically absent; traces in some patients Recessive	Markedly reduced: 0–10% of normal Predominantly recessive	Reduced: typically 10–33% of normal Promoter mutation: recessive Missense mutations: 7 of 8 dominant; 1 reportedly recessive

Crigler-Najjar Syndrome, Type I

CN-I is characterized by striking unconjugated hyperbilirubinemia of about 340–765 µmol/L (20–45 mg/dL) that appears in the neonatal period and persists for life. Other conventional hepatic biochemical tests such as serum aminotransferases and alkaline phosphatase are normal, and there is no evidence of hemolysis. Hepatic histology is also essentially normal except for the occasional presence of bile plugs within canaliculi. Bilirubin glucuronides are virtually absent from the bile, and there is no detectable constitutive expression of UGT1A1 activity in hepatic tissue. Neither UGT1A1 activity nor the serum bilirubin concentration responds to administration of phenobarbital or other enzyme inducers. Unconjugated bilirubin accumulates in plasma, from which it is eliminated very slowly by alternative pathways that include direct passage into the bile and small intestine, possibly via bilirubin photoisomers. This accounts for the small amount of urobilinogen found in feces. No bilirubin is found in the urine. First described in 1952, the disorder is rare (estimated prevalence, 0.6–1.0 per million). Many patients are from geographically or socially isolated communities in which consanguinity is common, and pedigree analyses show an autosomal recessive pattern of inheritance. The majority of patients (type IA) exhibit defects in the glucuronide conjugation of a spectrum of substrates in addition to bilirubin, including various drugs and other xenobiotics. These individuals have mutations in one of the common exons (2–5) of the *UGT1* gene (Fig. 331-2). In a smaller subset (type IB), the defect is limited largely to bilirubin conjugation, and the causative mutation is in the bilirubin-specific exon A1. Estrogen glucuronidation is mediated by UGT1A1 and is defective in all CN-I patients. More than 30 different genetic lesions of *UGT1A1* responsible for CN-I have been identified, including deletions, insertions, alterations in intron splice donor and acceptor sites, exon skipping, and point mutations that introduce premature stop codons or alter critical amino acids. Their common feature is that they all encode proteins with absent or, at most, traces of bilirubin-UDP-glucuronosyltransferase enzymatic activity.

Prior to the use of phototherapy, most patients with CN-I died of bilirubin encephalopathy (*kernicterus*) in infancy or early childhood. A few lived as long as early adult life without overt neurologic damage, although more subtle testing usually indicated mild but progressive brain damage. In the absence of liver transplantation, death eventually supervened from late-onset bilirubin encephalopathy, which often followed a nonspecific febrile illness. Although isolated hepatocyte transplantation has been used in a small number of cases of CN-I, early liver transplantation **(Chap. 338)** remains the best hope to prevent brain injury and death at present. It is anticipated that gene replacement therapy may be an option in the future.

Crigler-Najjar Syndrome, Type II (CN-II)

This condition was recognized as a distinct entity in 1962 and is characterized by marked unconjugated hyperbilirubinemia in the absence of abnormalities of other conventional hepatic biochemical tests, hepatic histology, or hemolysis. It differs from CN-I in several specific ways (Table 331-1): (1) Although there is considerable overlap, average bilirubin concentrations are lower in CN-II; (2) accordingly, CN-II is only infrequently associated with kernicterus; (3) bile is deeply colored, and bilirubin glucuronides are present, with a striking, characteristic increase in the proportion of monoglucuronides; (4) UGT1A1 in liver is usually present at reduced levels (typically ≤10% of normal); and (5) while typically detected in infancy, hyperbilirubinemia was not recognized in some cases until later in life and, in one instance, at age 34. As with CN-I, most CN-II cases exhibit abnormalities in the conjugation of other compounds, such as salicylamide and menthol, but in some instances, the defect appears limited to bilirubin. Reduction of serum bilirubin concentrations by >25% in response to enzyme inducers such as phenobarbital distinguishes CN-II from CN-I, although this response may not be elicited in early infancy and often is not accompanied by measurable UGT1A1 induction. Bilirubin concentrations during phenobarbital administration do not return to normal but are typically in the range of 51–86 µmol/L (3–5 mg/dL). Although the incidence of kernicterus in CN-II is low, instances have occurred, not only in infants but also in adolescents and adults, often in the setting of an intercurrent illness, fasting, or another factor that temporarily raises the serum bilirubin concentration above baseline and reduces serum albumin levels. For this reason, phenobarbital therapy is widely recommended, a single bedtime dose often sufficing to maintain clinically safe serum bilirubin concentrations.

Over 100 different mutations in the *UGT1* gene have been identified as causing CN-I or CN-II. It was found that missense mutations are more common in CN-II patients, as would be expected in this less severe phenotype. Their common feature is

that they encode for a bilirubin-UDP-glucuronosyltransferase with markedly reduced, but detectable, enzymatic activity. The spectrum of residual enzyme activity explains the spectrum of phenotypic severity of the resulting hyperbilirubinemia. Molecular analysis has established that a large majority of CN-II patients are either homozygotes or compound heterozygotes for CN-II mutations and that individuals carrying one mutated and one entirely normal allele have normal bilirubin concentrations.

Gilbert's Syndrome

This syndrome is characterized by mild unconjugated hyperbilirubinemia, normal values for standard hepatic biochemical tests, and normal hepatic histology other than a modest increase of lipofuscin pigment in some patients. Serum bilirubin concentrations are most often <51 µmol/L (<3 mg/dL), although both higher and lower values are frequent. The clinical spectrum of hyperbilirubinemia fades into that of CN-II at serum bilirubin concentrations of 86–136 µmol/L (5–8 mg/dL). At the other end of the scale, the distinction between mild cases of GS and a normal state is often blurred. Bilirubin concentrations may fluctuate substantially in any given individual, and at least 25% of patients will exhibit temporarily normal values during prolonged follow-up. More elevated values are associated with stress, fatigue, alcohol use, reduced caloric intake, and intercurrent illness, while increased caloric intake or administration of enzyme-inducing agents produces lower bilirubin levels. GS is most often diagnosed at or shortly after puberty or in adult life during routine examinations that include multichannel biochemical analyses. UGT1A1 activity is typically reduced to 10–35% of normal, and bile pigments exhibit a characteristic increase in bilirubin monoglucuronides. Studies of radiobilirubin kinetics indicate that hepatic bilirubin clearance is reduced to an average of one-third of normal. Administration of phenobarbital normalizes both the serum bilirubin concentration and hepatic bilirubin clearance; however, failure of UGT1A1 activity to improve in many such instances suggests the possible coexistence of an additional defect. Compartmental analysis of bilirubin kinetic data suggests that GS patients may have a defect in bilirubin uptake as well as in conjugation, although this has not been shown directly. Defect(s) in the hepatic uptake of other organic anions that at least partially share an uptake mechanism with bilirubin, such as sulfobromophthalein and indocyanine green (ICG), are observed in a minority of patients. The metabolism and transport of bile acids that do not utilize the bilirubin uptake mechanism are normal. The magnitude of changes in the serum bilirubin concentration induced by provocation tests such as 48 hours of fasting or the IV administration of nicotinic acid have been reported to be of help in separating GS patients from normal individuals. Other studies dispute this assertion. Moreover, on theoretical grounds, the results of such studies should provide no more information than simple measurements of the baseline serum bilirubin concentration. Family studies indicate that GS and hereditary hemolytic anemias such as hereditary spherocytosis, glucose-6-phosphate dehydrogenase deficiency, and β-thalassemia trait sort independently. Reports of hemolysis in up to 50% of GS patients are believed to reflect better case finding, since patients with both GS and hemolysis have higher bilirubin concentrations, and are more likely to be jaundiced, than patients with either defect alone.

GS is common, with many series placing its prevalence as high as 8%. Males predominate over females by reported ratios ranging from 1.5:1 to >7:1. However, these ratios may have a large artifactual component since normal males have higher mean bilirubin levels than normal females, but the diagnosis of GS is often based on comparison to normal ranges established in men. The high prevalence of GS in the general population may explain the reported frequency of mild unconjugated hyperbilirubinemia in liver transplant recipients. The disposition of most xenobiotics metabolized by glucuronidation appears to be normal in GS, as is oxidative drug metabolism in the majority of reported studies. The principal exception is the metabolism of the antitumor agent irinotecan (CPT-11), whose active metabolite (SN-38) is glucuronidated specifically by bilirubin-UDP-glucuronosyltransferase. Administration of CPT-11 to patients with GS has resulted in several toxicities, including intractable diarrhea and myelosuppression. Some reports also suggest abnormal disposition of menthol, estradiol benzoate, acetaminophen, tolbutamide, and rifamycin SV. Although some of these studies have been disputed, and there have been no reports of clinical complications from use of these agents in GS, prudence should be exercised in prescribing them, or any agents metabolized primarily by glucuronidation in this condition. It should also be noted that the

HIV protease inhibitors indinavir and atazanavir (Chap. 197) can inhibit UGT1A1, resulting in hyperbilirubinemia that is most pronounced in patients with preexisting GS.

Most older pedigree studies of GS were consistent with autosomal dominant inheritance with variable expressivity. However, studies of the *UGT1* gene in GS have indicated a variety of molecular genetic bases for the phenotypic picture and several different patterns of inheritance. Studies in Europe and the United States found that nearly all patients had normal coding regions for UGT1A1, but were homozygous for the insertion of an extra TA (i.e., $A[TA]_7TAA$ rather than $A[TA]_6TAA$) in the promoter region of the first exon. This appeared to be necessary, but not sufficient, for clinically expressed GS, since 15% of normal controls were also homozygous for this variant. While normal by standard criteria, these individuals had somewhat higher bilirubin concentrations than the rest of the controls studied. Heterozygotes for this abnormality had bilirubin concentrations identical to those homozygous for the normal $A[TA]_6TAA$ allele. The prevalence of the $A[TA]_7TAA$ allele in a general Western population is 30%, in which case 9% would be homozygotes. This is slightly higher than the prevalence of GS based on purely phenotypic parameters. It was suggested that additional variables, such as mild hemolysis or a defect in bilirubin uptake, might be among the factors enhancing phenotypic expression of the defect.

Phenotypic expression of GS due solely to the $A[TA]_7TAA$ promoter abnormality is inherited as an autosomal recessive trait. A number of CN-II kindreds have been identified in whom there is also an allele containing a normal coding region but the $A[TA]_7TAA$ promoter abnormality. CN-II heterozygotes, who have the $A[TA]_6TAA$ promoter, are phenotypically normal, whereas those with the $A[TA]_7TAA$ promoter express the phenotypic picture of GS. GS in such kindreds may also result from homozygosity for the $A[TA]_7TAA$ promoter abnormality. Seven different missense mutations in the *UGT1* gene that reportedly cause GS with dominant inheritance have been found in Japanese individuals. Another Japanese patient with mild unconjugated hyperbilirubinemia was homozygous for a missense mutation in exon 5. GS in her family appeared to be recessive.

DISORDERS OF BILIRUBIN METABOLISM LEADING TO MIXED OR PREDOMINANTLY CONJUGATED HYPERBILIRUBINEMIA

In hyperbilirubinemia due to acquired liver disease (e.g., acute hepatitis, common bile duct stone), there are usually elevations in the serum concentrations of both conjugated and unconjugated bilirubin. Although biliary tract obstruction or hepatocellular cholestatic injury may present on occasion with a predominantly conjugated hyperbilirubinemia, it is generally not possible to differentiate intrahepatic from extrahepatic causes of jaundice based on the serum levels or relative proportions of unconjugated and conjugated bilirubin. The major reason for determining the amounts of conjugated and unconjugated bilirubin in the serum is for the initial differentiation of hepatic parenchymal and obstructive disorders (mixed conjugated and unconjugated hyperbilirubinemia) from the inheritable and hemolytic disorders discussed above that are associated with unconjugated hyperbilirubinemia.

FAMILIAL DEFECTS IN HEPATIC EXCRETORY FUNCTION

Dubin-Johnson Syndrome (DJS)

This benign, relatively rare disorder is characterized by low-grade, predominantly conjugated hyperbilirubinemia (Table 331-2). Total bilirubin concentrations are typically between 34 and 85 μmol/L (2 and 5 mg/dL) but on occasion can be in the normal range or as high as 340–430 μmol/L (20–25 mg/dL) and can fluctuate widely in any given patient. The degree of hyperbilirubinemia may be increased by intercurrent illness, oral contraceptive use, and pregnancy. Because the hyperbilirubinemia is due to a predominant rise in conjugated bilirubin, bilirubinuria is characteristically present. Aside from

elevated serum bilirubin levels, other routine laboratory tests are normal. Physical examination is usually normal except for jaundice, although an occasional patient may have hepatosplenomegaly.

TABLE 331-2

Principal Differential Characteristics of Inheritable Disorders of Bile Canalicular Function

	DJS	Rotor	PFIC1	BRIC1	PFIC2	BRIC2	PFIC3
Gene Protein Cholestasis	*ABCCA* MRP2 No	*SLCO1B1/SLCO1B3* OATP1B1/1B3 No	*ATP8B1* FIC1 Yes	*ATP8B1* FIC1 Episodic	*ABCB11* BSEP Yes	*ABCB11* BSEP Episodic	*ABCB4* MDR3 Yes
Serum GGT Serum bile acids	Normal Normal	Normal Normal	Normal ↑↑	Normal ↑↑ during episodes	Normal ↑↑	Normal ↑↑ during episodes	↑↑ ↑↑
Clinical features	Mild conjugated hyperbilirubinemia; otherwise normal liver function; dark pigment in liver; characteristic pattern of urinary coproporphyrins	Mild conjugated hyperbilirubinemia; otherwise normal liver function; liver without abnormal pigmentation	Severe cholestasis beginning in childhood	Recurrent episodes of cholestasis beginning at any age	Severe cholestasis beginning in childhood	Recurrent episodes of cholestasis beginning at any age	Severe cholestasis beginning in childhood; decreased phospholipids in bile

Abbreviations: BRIC, benign recurrent intrahepatic cholestasis; BSEP, bile salt excretory protein; DJS, Dubin-Johnson syndrome; GGT, γ-glutamyltransferase; MRP2, multidrug resistance–associated protein 2; OATP1A/1B, organic anion transport proteins 1B1 and 1B3; PFIC, progressive familial intrahepatic cholestasis; ↑↑, increased.

Patients with DJS are usually asymptomatic, although some may have vague constitutional symptoms. These latter patients have usually undergone extensive and often unnecessary diagnostic examinations for unexplained jaundice and have high levels of anxiety. In women, the condition may be subclinical until the patient becomes pregnant or receives oral contraceptives, at which time chemical hyperbilirubinemia becomes frank jaundice. Even in these situations, other routine liver function tests, including serum alkaline phosphatase and transaminase activities, are normal.

A cardinal feature of DJS is the accumulation in the lysosomes of centrilobular hepatocytes of dark, coarsely granular pigment. As a result, the liver may be grossly black in appearance. This pigment is thought to be derived from epinephrine metabolites that are not excreted normally. The pigment may disappear during bouts of viral hepatitis, only to reaccumulate slowly after recovery.

Biliary excretion of a number of anionic compounds is compromised in DJS. These include various cholecystographic agents, as well as sulfobromophthalein (Bromsulphalein, BSP), a synthetic dye formerly used in a test of liver function. In this test, the rate of disappearance of BSP from plasma was determined following bolus IV administration. BSP is conjugated with glutathione in the hepatocyte; the resulting conjugate is normally excreted rapidly into the bile canaliculus. Patients with DJS exhibit characteristic rises in plasma concentrations at 90 min after injection, due to reflux of conjugated BSP into the circulation from the hepatocyte. Dyes such as ICG that are taken up by hepatocytes but are not further metabolized prior to biliary excretion do not show this reflux phenomenon. Continuous BSP infusion studies suggest a reduction in the time to

maximum plasma concentration (t_{max}) for biliary excretion. Bile acid disposition, including hepatocellular uptake and biliary excretion, is normal in DJS. These patients have normal serum and biliary bile acid concentrations and do not have pruritus.

By analogy with findings in several mutant rat strains, the selective defect—in biliary excretion of bilirubin conjugates and certain other classes of organic compounds, but not of bile acids—that characterizes DJS in humans was found to reflect defective expression of MRP2, an ATP-dependent canalicular membrane transporter. Several different mutations in the *MRP2* gene produce the Dubin-Johnson phenotype, which has an autosomal recessive pattern of inheritance. Although MRP2 is undoubtedly important in the biliary excretion of conjugated bilirubin, the fact that this pigment is still excreted in the absence of MRP2 suggests that other, as yet uncharacterized, transport proteins may serve in a secondary role in this process.

Patients with DJS also have a diagnostic abnormality in urinary coproporphyrin excretion. There are two naturally occurring coproporphyrin isomers, I and III. Normally, ~75% of the coproporphyrin in urine is isomer III. In urine from DJS patients, total coproporphyrin content is normal, but >80% is isomer I. Heterozygotes for the syndrome show an intermediate pattern. The molecular basis for this phenomenon remains unclear.

Rotor Syndrome

This benign, autosomal recessive disorder is clinically similar to DJS (Table 331-2), although it is seen even less frequently. A major phenotypic difference is that the liver in patients with Rotor syndrome has no increased pigmentation and appears totally normal. The only abnormality in routine laboratory tests is an elevation of total serum bilirubin, due to a predominant rise in conjugated bilirubin. This is accompanied by bilirubinuria. Several additional features differentiate Rotor syndrome from DJS. In Rotor syndrome, the gallbladder is usually visualized on oral cholecystography, in contrast to the nonvisualization that is typical of DJS. The pattern of urinary coproporphyrin excretion also differs. The pattern in Rotor syndrome resembles that of many acquired disorders of hepatobiliary function, in which coproporphyrin I, the major coproporphyrin isomer in bile, refluxes from the hepatocyte back into the circulation and is excreted in urine. Thus, total urinary coproporphyrin excretion is substantially increased in Rotor syndrome, in contrast to the normal levels seen in DJS. Although the fraction of coproporphyrin I in urine is elevated, it is usually <70% of the total, compared with ≥80% in DJS. The disorders also can be distinguished by their patterns of BSP excretion. Although clearance of BSP from plasma is delayed in Rotor syndrome, there is no reflux of conjugated BSP back into the circulation as seen in DJS. Kinetic analysis of plasma BSP infusion studies suggests the presence of a defect in intrahepatocellular storage of this compound. This has never been demonstrated directly. Recent studies indicate that the molecular basis of Rotor syndrome results from simultaneous deficiency of the hepatocyte plasma membrane transporters OATP1B1 and OATP1B3. This results in reduced reuptake by these transporters of conjugated bilirubin that has been pumped out of the hepatocyte into the portal circulation by MRP3 (Fig. 331-1).

Benign Recurrent Intrahepatic Cholestasis (BRIC)

This rare disorder is characterized by recurrent attacks of pruritus and jaundice. The typical episode begins with mild malaise and elevations in serum aminotransferase levels, followed rapidly by rises in alkaline phosphatase and conjugated bilirubin and onset of jaundice and itching. The first one or two episodes may be misdiagnosed as acute viral hepatitis. The cholestatic episodes, which may begin in childhood or adulthood, can vary in duration from several weeks to months, followed by a complete clinical and biochemical resolution. Intervals between attacks may vary from several months to years. Between episodes, physical examination is normal, as are serum levels of bile acids, bilirubin, transaminases, and alkaline phosphatase. The disorder is familial and has an autosomal recessive pattern of inheritance. BRIC is considered a benign disorder in that it does not lead to cirrhosis or end-stage liver disease. However, the episodes of jaundice and pruritus can be prolonged and debilitating, and some patients have undergone liver transplantation to relieve the intractable and disabling symptoms. Treatment during the cholestatic episodes is symptomatic; there is no specific treatment to prevent or shorten the occurrence of episodes.

A gene termed *FIC1* was recently identified and found to be mutated in patients with BRIC. Curiously, this gene is expressed strongly in the small intestine but only weakly in the liver. The protein encoded by *FIC1* shows little similarity to those that have been shown to play a role in bile canalicular excretion of various compounds. Rather, it appears to be a member of a P-type ATPase family that transports aminophospholipids from the outer to the inner leaflet of a variety of cell membranes. Its relationship to the pathobiology of this disorder remains unclear. A second phenotypically identical form of BRIC, termed BRIC type 2, has been described resulting from mutations in the bile salt excretory protein (BSEP), the protein that is defective in progressive familial intrahepatic cholestasis (PFIC) type 2 (Table 331-2). How some mutations in this protein result in the episodic BRIC phenotype is unknown.

Progressive Familial Intrahepatic Cholestasis (FIC)

This name is applied to three phenotypically related syndromes (Table 331-2). PFIC type 1 (Byler disease) presents in early infancy as cholestasis that may be initially episodic. However, in contrast to BRIC, Byler disease progresses to malnutrition, growth retardation, and end-stage liver disease during childhood. This disorder is also a consequence of a *FIC1* mutation. The functional relationship of the FIC1 protein to the pathogenesis of cholestasis in these disorders is unknown. Two other types of PFIC (types 2 and 3) have been described. PFIC type 2 is associated with a mutation in the protein originally named *sister of p-glycoprotein*, now known as *bile salt excretory protein*, which is the major bile canalicular exporter of bile acids. As noted above, some mutations of this protein are associated with BRIC type 2, rather than the progressive FIC type 2 phenotype. Progressive FIC type 3 has been associated with a mutation of MDR3, a protein that is essential for normal hepatocellular excretion of phospholipids across the bile canaliculus. Although all three types of PFIC have similar clinical phenotypes, only type 3 is associated with high serum levels of γ-glutamyltransferase (GGT) activity. In contrast, activity of this enzyme is normal or only mildly elevated in symptomatic BRIC and progressive FIC types 1 and 2. Interestingly, mutations in FIC1 or BSEP are not found in approximately one-third of patients with clinical PFIC and normal GGT. Recent studies have shown that patients with mutations in *NR1H4*, the gene encoding the farnesoid X receptor (FXR), a nuclear hormone receptor activated by bile acids, have a syndrome identical to PFIC2 with absent expression of BSEP. Mutations in tight junction protein 2 (TJP2) have also been associated with severe cholestasis with normal GGT levels, likely due to disruption of tight junctions at the bile canaliculus.

FURTHER READING

Canu G et al: Gilbert and Crigler Najjar syndromes: An update of the UDP-glucuronosyltransferase 1A1 (UGT1A1) gene mutation database. Blood Cells Mol Dis 50:273, 2013. [PubMed: 23403257]

Erlinger S, Arias IM, Dhumeaux D: Inherited disorders of bilirubin transport and conjugation: New insights into molecular mechanisms and consequences. Gastroenterology 146:1625, 2014. [PubMed: 24704527]

Gomez-Ospina N et al: Mutations in the nuclear bile acid receptor FXR cause progressive familial intrahepatic cholestasis. Nat Commun 7:10713, 2016. [PubMed: 26888176]

Hansen TW: Biology of bilirubin photoisomers. Clin Perinatol 43:277, 2016. [PubMed: 27235207]

Lamola AA: A pharmacologic view of phototherapy. Clin Perinatol 43:259, 2016. [PubMed: 27235206]

Memon N et al: Inherited disorders of bilirubin clearance. Pediatr Res 79:378, 2016. [PubMed: 26595536]

Rodrigues C et al: Impact of UGT1A1 gene variants on total bilirubin levels in Gilbert syndrome patients and in healthy subjects. Blood Cells Mol Dis 48:166, 2012. [PubMed: 22325916]

Sambrotta M et al: Mutations in TJP2 cause progressive cholestatic liver disease. Nat Genet 46:326, 2014. [PubMed: 24614073]

Soroka CJ, Boyer JL: Biosynthesis and trafficking of the bile salt export pump, BSEP: Therapeutic implications of BSEP mutations. Mol Aspects Med 37:3, 2014. [PubMed: 23685087]

van de Steeg E et al: Complete OATP1B1 and OATP1B3 deficiency causes human Rotor syndrome by interrupting conjugated bilirubin reuptake into the liver. J Clin Invest 122:519, 2012. [PubMed: 22232210]

Wolkoff AW: Organic anion uptake by hepatocytes. Compr Physiol 4:1715, 2014. [PubMed: 25428858]

Chapter 332: Acute Viral Hepatitis

Jules L. Dienstag

INTRODUCTION

Acute viral hepatitis is a systemic infection affecting the liver predominantly. Almost all cases of acute viral hepatitis are caused by one of five viral agents: hepatitis A virus (HAV), hepatitis B virus (HBV), hepatitis C virus (HCV), the HBV-associated delta agent or hepatitis D virus (HDV), and hepatitis E virus (HEV). All these human hepatitis viruses are RNA viruses, except for hepatitis B, which is a DNA virus but replicates like a retrovirus. Although these agents can be distinguished by their molecular and antigenic properties, all types of viral hepatitis produce clinically similar illnesses. These range from asymptomatic and inapparent to fulminant and fatal acute infections common to all types, on the one hand, and from subclinical persistent infections to rapidly progressive chronic liver disease with cirrhosis and even hepatocellular carcinoma, common to the bloodborne types (HBV, HCV, and HDV), on the other.

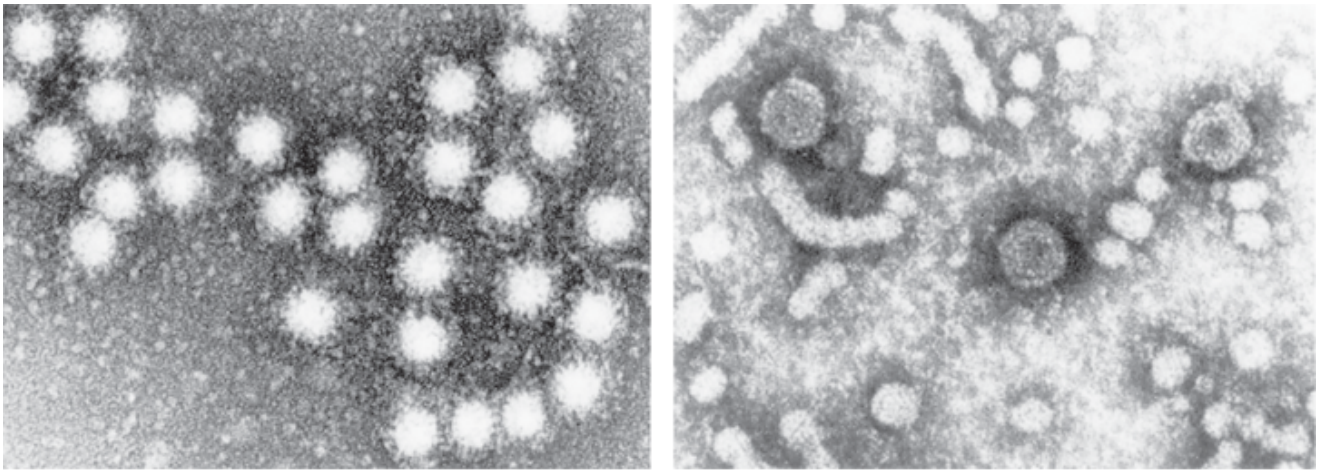
VIROLOGY AND ETIOLOGY

Hepatitis A

HAV is a nonenveloped 27-nm, heat-, acid-, and ether-resistant RNA virus in the *Hepatovirus* genus of the picornavirus family (**Fig. 332-1**). Its virion contains four capsid polypeptides, designated VP1–VP4, which are cleaved posttranslationally from the polyprotein product of a 7500-nucleotide genome. Inactivation of viral activity can be achieved by boiling for 1 min, by contact with formaldehyde and chlorine, or by ultraviolet irradiation. Despite nucleotide sequence variation of up to 20% among isolates of HAV, and despite the recognition of four genotypes affecting humans, all strains of this virus are immunologically indistinguishable and belong to one serotype. Human HAV can infect and cause hepatitis in chimpanzees, tamarins (marmosets), and several monkey species. Recently, a hepatotropic *Hepatovirus* related to, and likely to have shared common evolutionary ancestry with, human HAV has been identified in several species of harbor seals, albeit without histologic evidence for liver injury or inflammation. Hepatitis A has an incubation period of ~4 weeks. Its replication is limited to the liver, but the virus is present in the liver, bile, stools, and blood during the late incubation period and acute preicteric/presymptomatic phase of illness. Despite slightly longer persistence of virus in the liver, fecal shedding, viremia, and infectivity diminish rapidly once jaundice becomes apparent. HAV can be cultivated reproducibly in vitro.

FIGURE 332-1

Electron micrographs of hepatitis A virus particles and serum from a patient with hepatitis B. ***Left:*** 27-nm hepatitis A virus particles purified from stool of a patient with acute hepatitis A and aggregated by antibody to hepatitis A virus. ***Right:*** Concentrated serum from a patient with hepatitis B, demonstrating the 42-nm virions, tubular forms, and spherical 22-nm particles of hepatitis B surface antigen. 132,000×. (Hepatitis D resembles 42-nm virions of hepatitis B but is smaller, 35–37 nm; hepatitis E resembles hepatitis A virus but is slightly larger, 32–34 nm; hepatitis C has been visualized as a 55-nm particle.)

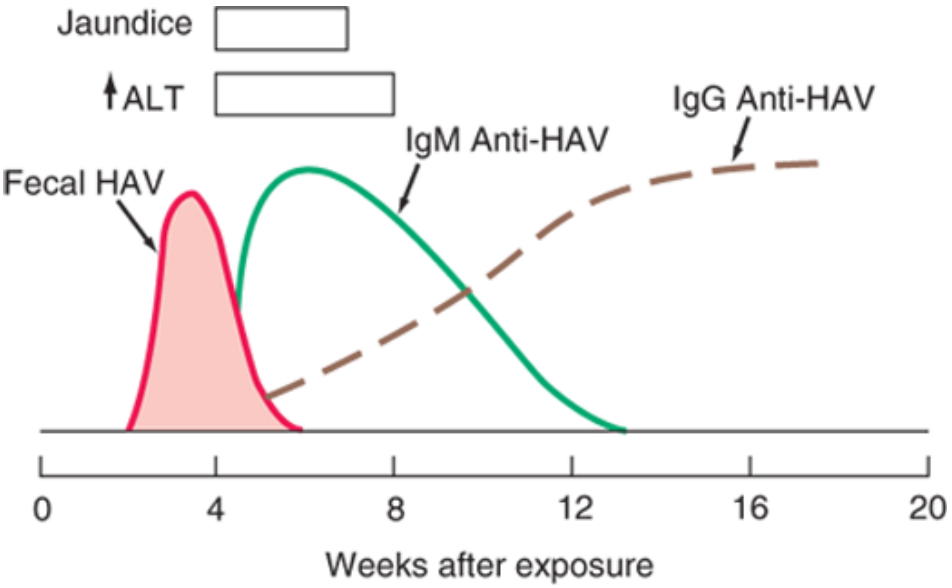


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Antibodies to HAV (anti-HAV) can be detected during acute illness when serum aminotransferase activity is elevated and fecal HAV shedding is still occurring. This early antibody response is predominantly of the IgM class and persists for several (~3) months, rarely for 6–12 months. During convalescence, however, anti-HAV of the IgG class becomes the predominant antibody (Fig. 332-2). Therefore, the diagnosis of hepatitis A is made during acute illness by demonstrating anti-HAV of the IgM class. After acute illness, anti-HAV of the IgG class remains detectable indefinitely, and patients with serum anti-HAV are immune to reinfection. Neutralizing antibody activity parallels the appearance of anti-HAV, and the IgG anti-HAV present in immune globulin accounts for the protection it affords against HAV infection.

FIGURE 332-2

Scheme of typical clinical and laboratory features of hepatitis A virus (HAV). ALT, alanine aminotransferase.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

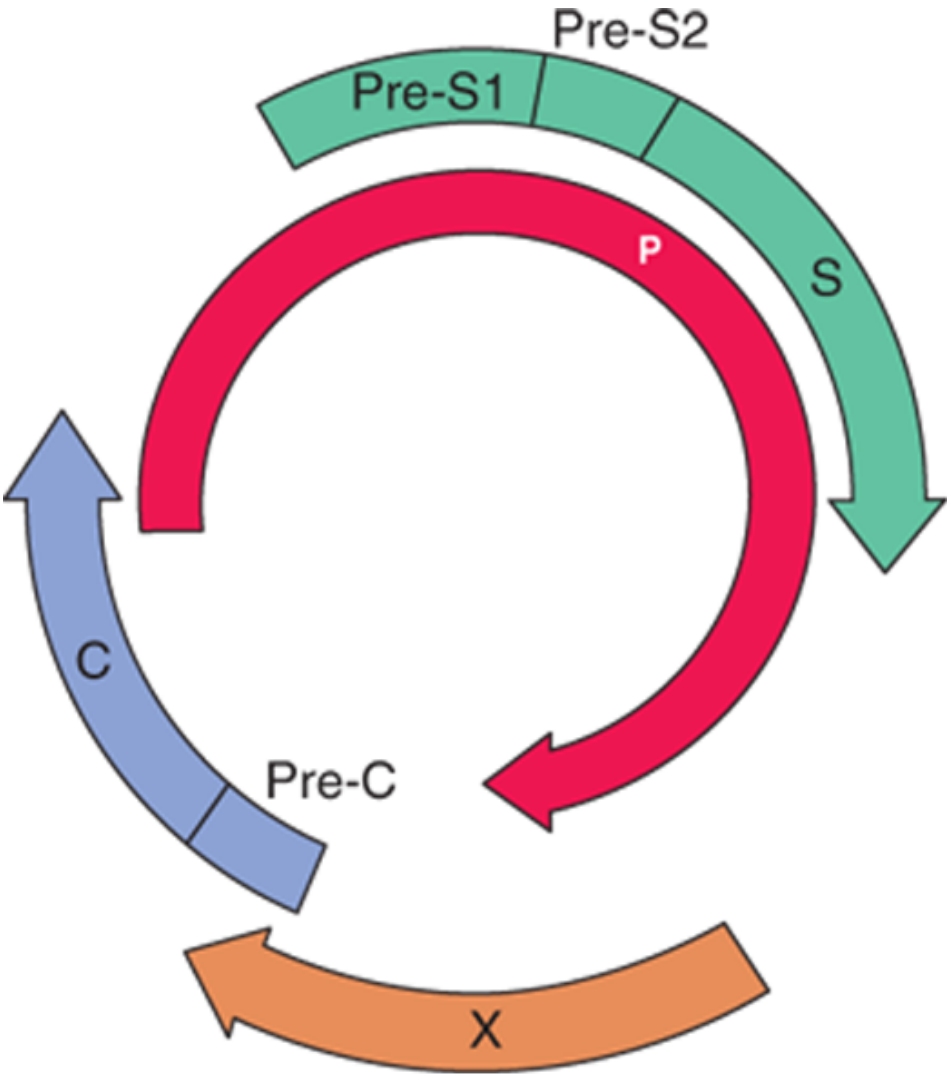
Hepatitis B

HBV is a DNA virus with a remarkably compact genomic structure; despite its small, circular, 3200-bp size, HBV DNA codes for four sets of viral products with a complex, multiparticle structure. HBV achieves its genomic economy by relying on an efficient strategy of encoding proteins from four overlapping genes: S, C, P, and X (Fig. 332-3), as detailed

below. Once thought to be unique among viruses, HBV is now recognized as one of a family of animal viruses, hepadnaviruses (hepatotropic DNA viruses), and is classified as hepadnavirus type 1. Similar viruses infect certain species of woodchucks, ground and tree squirrels, and Pekin ducks, to mention the most carefully characterized; genetic evidence of ancient HBV-like virus forbears has been found in fossils of ancient birds, and a HBV-like virus has been identified in contemporary fish. Like HBV, all have the same distinctive three morphologic forms, have counterparts to the envelope and nucleocapsid virus antigens of HBV, replicate in the liver but exist in extrahepatic sites, contain their own endogenous DNA polymerase, have partially double-strand and partially single-strand genomes, are associated with acute and chronic hepatitis and hepatocellular carcinoma, and rely on a replicative strategy unique among DNA viruses but typical of retroviruses. Entry of HBV into hepatocytes is mediated by binding to the sodium taurocholate cotransporting polypeptide receptor. Instead of DNA replication directly from a DNA template, hepadnaviruses rely on reverse transcription (effected by the DNA polymerase) of minus-strand DNA from a “pregenomic” RNA intermediate. Then, plus-strand DNA is transcribed from the minus-strand DNA template by the DNA-dependent DNA polymerase and converted in the hepatocyte nucleus to a covalently closed circular DNA, which serves as a template for messenger RNA and pregenomic RNA. Viral proteins are translated by the messenger RNA, and the proteins and genome are packaged into virions and secreted from the hepatocyte. Although HBV is difficult to cultivate in vitro in the conventional sense from clinical material, several cell lines have been transfected with HBV DNA. Such transfected cells support in vitro replication of the intact virus and its component proteins.

FIGURE 332-3

Compact genomic structure of hepatitis B virus (HBV). This structure, with overlapping genes, permits HBV to code for multiple proteins. The S gene codes for the “major” envelope protein, HBsAg. Pre-S1 and pre-S2, upstream of S, combine with S to code for two larger proteins, “middle” protein, the product of pre-S2 + S, and “large” protein, the product of pre-S1 + pre-S2 + S. The largest gene, P, codes for DNA polymerase. The C gene codes for two nucleocapsid proteins, HBeAg, a soluble, secreted protein (initiation from the pre-C region of the gene), and HBcAg, the intracellular core protein (initiation after pre-C). The X gene codes for HBxAg, which can transactivate the transcription of cellular and viral genes; its clinical relevance is not known, but it may contribute to carcinogenesis by binding to p53.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

VIRAL PROTEINS AND PARTICLES

Of the three particulate forms of HBV (Table 332-1), the most numerous are the 22-nm particles, which appear as spherical or long filamentous forms; these are antigenically indistinguishable from the outer surface or envelope protein of HBV and are thought to represent excess viral envelope protein. Outnumbered in serum by a factor of 100 or 1000 to 1 compared with the spheres and tubules are large, 42-nm, double-shelled spherical particles, which represent the intact hepatitis B virion (Fig. 332-1). The envelope protein expressed on the outer surface of the virion and on the smaller spherical and tubular structures is referred to as *hepatitis B surface antigen* (HBsAg). The concentration of HBsAg and virus particles in the blood may reach 500 µg/mL and 10 trillion particles per milliliter, respectively. The envelope protein, HBsAg, is the product of the S gene of HBV.

TABLE 332-1

Nomenclature and Features of Hepatitis Viruses

HEPATITIS TYPE	VIRUS PARTICLE, NM	MORPHOLOGY	GENOME[a]	CLASSIFICATION	ANTIGEN(s)	ANTIBODIES	REMARKS
HAV	27	Icosahedral nonenveloped	7.5-kb RNA, linear, ss, +	Hepatovirus	HAV	Anti-HAV	Early fecal shedding Diagnosis: IgM anti-HAV Previous infection: IgG anti-HAV
HBV	42 27 22	Double-shelled virion (surface and core) spherical Nucleocapsid core Spherical and filamentous; represents excess virus coat material	3.2-kb DNA, circular, ss/ds	Hepadnavirus	HBsAg HBcAg HBeAg HBcAg HBeAg HBsAg	Anti-HBs Anti-HBc Anti-HBe Anti-HBc Anti-HBe Anti-HBs	Bloodborne virus; carrier state Acute diagnosis: HBsAg, IgM anti-HBc Chronic diagnosis: IgG anti-HBc, HBsAg Markers of replication: HBeAg, HBV DNA Liver, lymphocytes, other organs Nucleocapsid contains DNA and DNA polymerase; present in hepatocyte nucleus; HBcAg does not circulate; HBeAg (soluble, nonparticulate) and HBV DNA circulate—correlate with infectivity and complete virions

HEPATITIS TYPE	VIRUS PARTICLE, NM	MORPHOLOGY	GENOME[a]	CLASSIFICATION	ANTIGEN(s)	ANTIBODIES	REMARKS
							HBsAg detectable in >95% of patients with acute hepatitis B; found in serum, body fluids, hepatocyte cytoplasm; anti-HBs appears following infection—protective antibody
HCV	Approx. 50–80	Enveloped	9.4-kb RNA, linear, ss, +	Hepacivirus	HCV core antigen	Anti-HCV	Bloodborne agent, formerly labeled non-A, non-B hepatitis Acute diagnosis: anti-HCV, HCV RNA Chronic diagnosis: anti-HCV, HCV RNA; cytoplasmic location in hepatocytes

HEPATITIS TYPE	VIRUS PARTICLE, NM	MORPHOLOGY	GENOME[a]	CLASSIFICATION	ANTIGEN(s)	ANTIBODIES	REMARKS
HDV	35–37	Enveloped hybrid particle with HBsAg coat and HDV core	1.7-kb RNA, circular, ss, –	Resembles viroids and plant satellite viruses (genus Deltavirus)	HBsAg HDAg	Anti-HBs Anti-HDV	Defective RNA virus, requires helper function of HBV (hepadnaviruses); HDV antigen (HDAg) present in hepatocyte nucleus Diagnosis: anti-HDV, HDV RNA; HBV/HDV co-infection—IgM anti-HBc and anti-HDV; HDV superinfection—IgG anti-HBc and anti-HDV
HEV	32–34	Nonenveloped icosahedral	7.6-kb RNA, linear, ss, +	Hepevirus	HEV antigen	Anti-HEV	Agent of enterically transmitted hepatitis; rare in the United States; occurs in Asia, Mediterranean countries, Central America Diagnosis: IgM/IgG anti-HEV (assays not routinely available); virus in stool, bile, hepatocyte cytoplasm

[a]ss, single-strand; ss/ds, partially single-strand, partially double-strand; –, minus-strand; +, plus-strand.

Note: See text for abbreviations.

Envelope HBsAg subdeterminants include a common group-reactive antigen, *a*, shared by all HBsAg isolates and one of several subtype-specific antigens—*d* or *y*, *w* or *r*—as well as other specificities. Hepatitis B isolates fall into one of at least 8 subtypes and 10 genotypes (A–J). Geographic distribution of genotypes and subtypes varies; genotypes A (corresponding to subtype *adw*) and D (*ayw*) predominate in the United States and Europe, whereas genotypes B (*adw*) and C (*adr*) predominate in Asia. Clinical course and outcome are independent of subtype, but genotype B appears to be associated with less rapidly progressive liver disease and cirrhosis and a lower likelihood, or delayed appearance, of hepatocellular carcinoma than genotype C or D. Patients with genotype A are more likely to clear circulating viremia and achieve hepatitis B e antigen (HBeAg) and HBsAg seroconversion, both spontaneously and in response to antiviral therapy. In addition, “precore” mutations are favored by certain genotypes (see below).

Upstream of the S gene are the pre-S genes (Fig. 332-3), which code for pre-S gene products, including receptors on the HBV surface for polymerized human serum albumin and for hepatocyte membrane proteins. The pre-S region actually consists of both pre-S1 and pre-S2. Depending on where translation is initiated, three potential HBsAg gene products are synthesized. The protein product of the S gene is HBsAg (*major protein*), the product of the S region plus the adjacent pre-S2 region is the *middle protein*, and the product of the pre-S1 plus pre-S2 plus S regions is the *large protein*. Compared with the smaller spherical and tubular particles of HBV, complete 42-nm virions are enriched in the large protein. Both pre-S proteins and their respective antibodies can be detected during HBV infection, and the period of pre-S antigenemia appears to coincide with other markers of virus replication, as detailed below; however, pre-S proteins have little clinical relevance and are not included in routine serologic testing repertoires.

The intact 42-nm virion contains a 27-nm nucleocapsid core particle. Nucleocapsid proteins are coded for by the C gene. The antigen expressed on the surface of the nucleocapsid core is *hepatitis B core antigen* (HBcAg), and its corresponding antibody is anti-HBc. A third HBV antigen is *hepatitis B e antigen* (HBeAg), a soluble, nonparticulate, nucleocapsid protein that is immunologically distinct from intact HBcAg but is a product of the same C gene. The C gene has two initiation codons: a precore and a core region (Fig. 332-3). If translation is initiated at the precore region, the protein product is HBeAg, which has a signal peptide that binds it to the smooth endoplasmic reticulum, the secretory apparatus of the cell, leading to its secretion into the circulation. If translation begins at the core region, HBcAg is the protein product; it has no signal peptide and is not secreted, but it assembles into nucleocapsid particles, which bind to and incorporate RNA, and which, ultimately, contain HBV DNA. Also packaged within the nucleocapsid core is a DNA polymerase, which directs replication and repair of HBV DNA. When packaging within viral proteins is complete, synthesis of the incomplete plus strand stops; this accounts for the single-strand gap and for differences in the size of the gap. HBcAg particles remain in the hepatocyte, where they are readily detectable by immunohistochemical staining and are exported after encapsidation by an envelope of HBsAg. Therefore, naked core particles do not circulate in the serum. The secreted nucleocapsid protein, HBeAg, provides a convenient, readily detectable, qualitative marker of HBV replication and relative infectivity.

HBsAg-positive serum containing HBeAg is more likely to be highly infectious and to be associated with the presence of hepatitis B virions (and detectable HBV DNA, see below) than HBeAg-negative or anti-HBe-positive serum. For example, HBsAg-positive mothers who are HBeAg-positive almost invariably (>90%) transmit hepatitis B infection to their offspring, whereas HBsAg-positive mothers with anti-HBe rarely (10–15%) infect their offspring.

Early during the course of acute hepatitis B, HBeAg appears transiently; its disappearance may be a harbinger of clinical improvement and resolution of infection. Persistence of HBeAg in serum beyond the first 3 months of acute infection may be predictive of the development of chronic infection, and the presence of HBeAg during chronic hepatitis B tends to be associated with ongoing viral replication, infectivity, and inflammatory liver injury (except during the early decades after perinatally acquired HBV infection; see below).

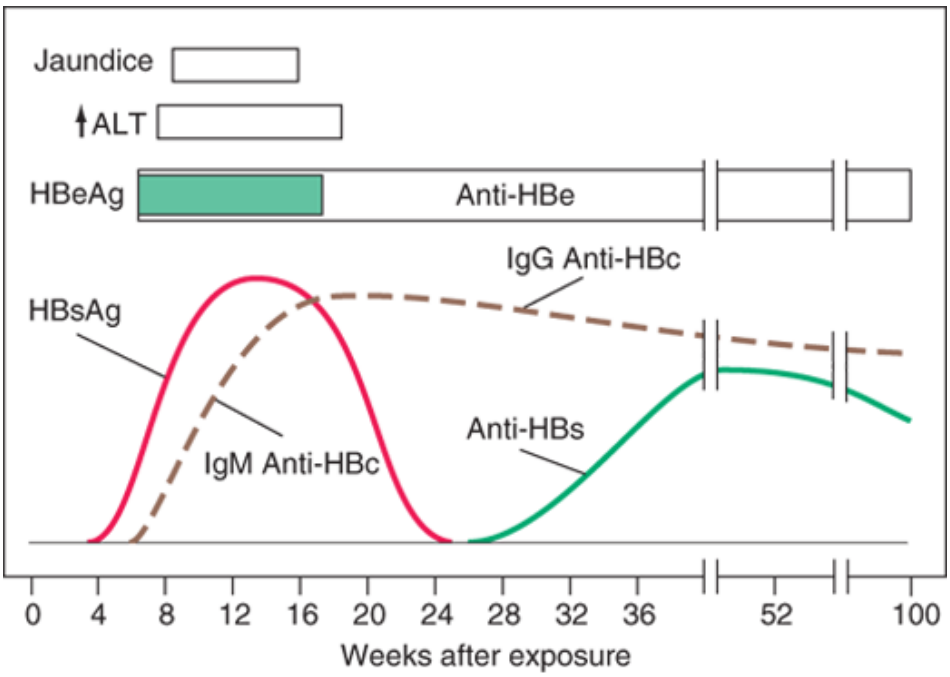
The third and largest of the HBV genes, the P gene (Fig. 332-3), codes for HBV DNA polymerase; as noted above, this enzyme has both DNA-dependent DNA polymerase and RNA-dependent reverse transcriptase activities. The fourth gene, X, codes for a small, nonparticulate protein, *hepatitis B x antigen* (HBxAg), that is capable of transactivating the transcription of both viral and cellular genes (Fig. 332-3). In the cytoplasm, HBxAg effects calcium release (possibly from mitochondria), which activates signal-transduction pathways that lead to stimulation of HBV reverse transcription and HBV DNA replication. Such transactivation may enhance the replication of HBV, leading to the clinical association observed between the expression of HBxAg and antibodies to it in patients with severe chronic hepatitis and hepatocellular carcinoma. The transactivating activity can enhance the transcription and replication of other viruses besides HBV, such as HIV. Cellular processes transactivated by X include the human interferon-γ gene and class I major histocompatibility genes; potentially, these effects could contribute to enhanced susceptibility of HBV-infected hepatocytes to cytolytic T cells. The expression of X can also induce programmed cell death (apoptosis). The clinical relevance of HBxAg is limited, however, and testing for it is not part of routine clinical practice.

SEROLOGIC AND VIROLOGIC MARKERS

After a person is infected with HBV, the first virologic marker detectable in serum within 1–12 weeks, usually between 8 and 12 weeks, is HBsAg **(Fig. 332-4)**. Circulating HBsAg precedes elevations of serum aminotransferase activity and clinical symptoms by 2–6 weeks and remains detectable during the entire icteric or symptomatic phase of acute hepatitis B and beyond. In typical cases, HBsAg becomes undetectable 1–2 months after the onset of jaundice and rarely persists beyond 6 months. After HBsAg disappears, antibody to HBsAg (anti-HBs) becomes detectable in serum and remains detectable indefinitely thereafter. Because HBcAg is intracellular and, when in the serum, sequestered within an HBsAg coat, naked core particles do not circulate in serum, and therefore HBcAg is not detectable routinely in the serum of patients with HBV infection. By contrast, anti-HBc is readily demonstrable in serum, beginning within the first 1–2 weeks after the appearance of HBsAg and preceding detectable levels of anti-HBs by weeks to months. Because variability exists in the time of appearance of anti-HBs after HBV infection, occasionally a gap of several weeks or longer may separate the disappearance of HBsAg and the appearance of anti-HBs. During this “gap” or “window” period, anti-HBc may represent the only serologic evidence of current or recent HBV infection, and blood containing anti-HBc in the absence of HBsAg and anti-HBs has been implicated in transfusion-associated hepatitis B. In part because the sensitivity of immunoassays for HBsAg and anti-HBs has increased, however, this window period is rarely encountered. In some persons, years after HBV infection, anti-HBc may persist in the circulation longer than anti-HBs. Therefore, isolated anti-HBc does not necessarily indicate active virus replication; most instances of isolated anti-HBc represent hepatitis B infection in the remote past. Rarely, however, isolated anti-HBc represents low-level hepatitis B viremia, with HBsAg below the detection threshold, and, occasionally, isolated anti-HBc represents a cross-reacting or false-positive immunologic specificity. Recent and remote HBV infections can be distinguished by determination of the immunoglobulin class of anti-HBc. Anti-HBc of the IgM class (IgM anti-HBc) predominates during the first 6 months after acute infection, whereas IgG anti-HBc is the predominant class of anti-HBc beyond 6 months. Therefore, patients with current or recent acute hepatitis B, including those in the anti-HBc window, have IgM anti-HBc in their serum. In patients who have recovered from hepatitis B in the remote past as well as those with chronic HBV infection, anti-HBc is predominantly of the IgG class. Infrequently, in ≤1–5% of patients with acute HBV infection, levels of HBsAg are too low to be detected; in such cases, the presence of IgM anti-HBc establishes the diagnosis of acute hepatitis B. When isolated anti-HBc occurs in the rare patient with chronic hepatitis B whose HBsAg level is below the sensitivity threshold of contemporary immunoassays (a low-level carrier), anti-HBc is of the IgG class. Generally, in persons who have recovered from hepatitis B, anti-HBs and anti-HBc persist indefinitely.

FIGURE 332-4

Scheme of typical clinical and laboratory features of acute hepatitis B. ALT, alanine aminotransferase.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

The temporal association between the appearance of anti-HBs and resolution of HBV infection as well as the observation that persons with anti-HBs in serum are protected against reinfection with HBV suggests that *anti-HBs is the protective antibody*. Therefore, strategies for prevention of HBV infection are based on providing susceptible persons with circulating anti-HBs (see below). Occasionally, in ~10% of patients with chronic hepatitis B, low-level, low-affinity anti-HBs can be detected. This antibody is directed against a subtype determinant different from that represented by the patient's HBsAg; its presence is thought to reflect the stimulation of a related clone of antibody-forming cells, but it has no clinical relevance and does not signal imminent clearance of hepatitis B. These patients with HBsAg and such nonneutralizing anti-HBs should be categorized as having chronic HBV infection.

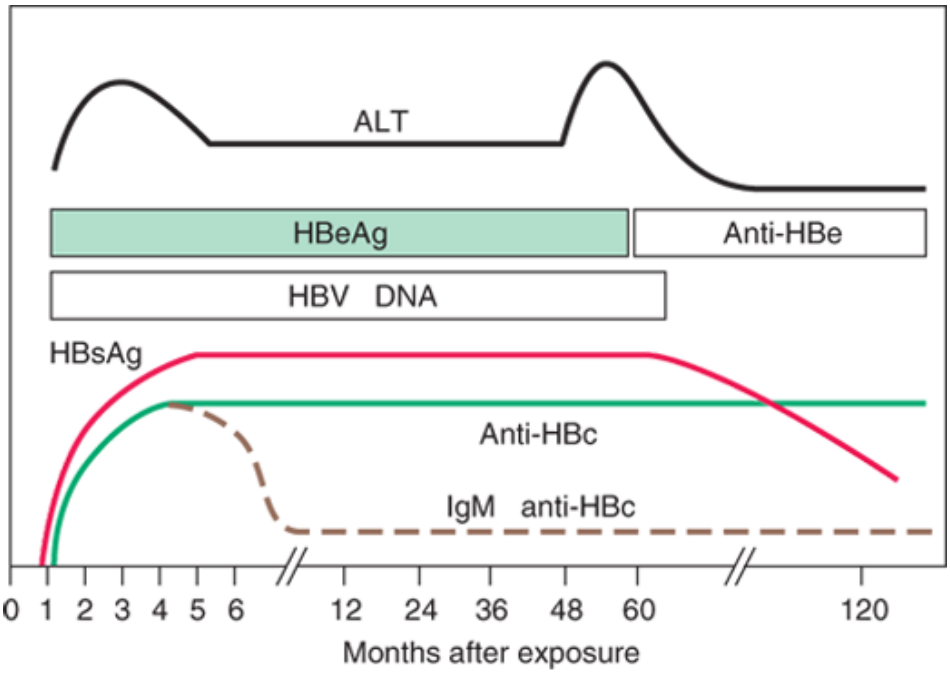
The other readily detectable serologic marker of HBV infection, HBeAg, appears concurrently with or shortly after HBsAg. Its appearance coincides temporally with high levels of virus replication and reflects the presence of circulating intact virions and detectable HBV DNA (with the notable exception of patients with precore mutations who cannot synthesize HBeAg—see "Molecular Variants"). Pre-S1 and pre-S2 proteins are also expressed during periods of peak replication, but assays for these gene products are not routinely available. In self-limited HBV infections, HBeAg becomes undetectable shortly after peak elevations in aminotransferase activity, before the disappearance of HBsAg, and anti-HBe then becomes detectable, coinciding with a period of relatively lower infectivity (Fig. 332-4). Because markers of HBV replication appear transiently during acute infection, testing for such markers is of little clinical utility in typical cases of acute HBV infection. In contrast, markers of HBV replication provide valuable information in patients with protracted infections.

Departing from the pattern typical of acute HBV infections, in chronic HBV infection, HBsAg remains detectable beyond 6 months, anti-HBc is primarily of the IgG class, and anti-HBs is either undetectable or detectable at low levels (see "Laboratory Features") **(Fig. 332-5)**. During early chronic HBV infection, HBV DNA can be detected both in serum and in hepatocyte nuclei, where it is present in free or episomal form. This relatively highly *replicative stage* of HBV infection is the time of maximal infectivity and liver injury; HBeAg is a qualitative marker and HBV DNA a quantitative marker of this replicative phase, during which all three forms of HBV circulate, including intact virions. Over time, the relatively replicative phase of chronic HBV infection gives way to a relatively *nonreplicative phase*. This occurs at a rate of ~10% per year and is accompanied by seroconversion from HBeAg to anti-HBe. In many cases, this seroconversion coincides

with a transient, usually mild, acute hepatitis-like elevation in aminotransferase activity, believed to reflect cell-mediated immune clearance of virus-infected hepatocytes. In the nonreplicative phase of chronic infection, when HBV DNA is demonstrable in hepatocyte nuclei, it tends to be integrated into the host genome. In this phase, only spherical and tubular forms of HBV, *not intact virions*, circulate, and liver injury tends to subside. Most such patients would be characterized as *inactive HBV carriers*. In reality, the designations *replicative* and *nonreplicative* are only relative; even in the so-called nonreplicative phase, HBV replication can be detected at levels of approximately ≤10^3 virions/mL with highly sensitive amplification probes such as the polymerase chain reaction (PCR); below this replication threshold, liver injury and infectivity of HBV are limited to negligible. Still, the distinctions are pathophysiologically and clinically meaningful. Occasionally, nonreplicative HBV infection converts back to replicative infection. Such spontaneous reactivations are accompanied by reexpression of HBeAg and HBV DNA, and sometimes of IgM anti-HBc, as well as by exacerbations of liver injury. Because high-titer IgM anti-HBc can reappear during acute exacerbations of chronic hepatitis B, relying on IgM anti-HBc versus IgG anti-HBc to distinguish between acute and chronic hepatitis B infection, respectively, may not always be reliable; in such cases, patient history and additional follow-up monitoring over time are invaluable in helping to distinguish de novo acute hepatitis B infection from acute exacerbation of chronic hepatitis B infection.

FIGURE 332-5

Scheme of typical laboratory features of wild-type chronic hepatitis B. HBeAg and hepatitis B virus (HBV) DNA can be detected in serum during the relatively *replicative phase* of chronic infection, which is associated with infectivity and liver injury. Seroconversion from the replicative phase to the relatively *nonreplicative phase* occurs at a rate of ~10% per year and is heralded by an acute hepatitis–like elevation of alanine aminotransferase (ALT) activity; during the nonreplicative phase, infectivity and liver injury are limited. In HBeAg-negative chronic hepatitis B associated with mutations in the precore region of the HBV genome, replicative chronic hepatitis B occurs in the absence of HBeAg.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

MOLECULAR VARIANTS

Variation occurs throughout the HBV genome, and clinical isolates of HBV that do not express typical viral proteins have been attributed to mutations in individual or even multiple gene locations. For example, variants have been described that lack nucleocapsid proteins (commonly), envelope proteins (very rarely), or both. Two categories of naturally

occurring HBV variants have attracted the most attention. One of these was identified initially in Mediterranean countries among patients with severe chronic HBV infection and detectable HBV DNA but with anti-HBe instead of HBeAg. These patients were found to be infected with an HBV mutant that contained an alteration in the precore region rendering the virus incapable of encoding HBeAg. Although several potential mutation sites exist in the pre-C region, the region of the C gene necessary for the expression of HBeAg (see "Virology and Etiology"), the most commonly encountered in such patients is a single base substitution, from G to A in the second to last codon of the pre-C gene at nucleotide 1896. This substitution results in the replacement of the TGG tryptophan codon by a stop codon (TAG), which prevents the translation of HBeAg. Another mutation, in the core-promoter region, prevents transcription of the coding region for HBeAg and yields an HBeAg-negative phenotype. Patients with such mutations in the precore region and who are unable to secrete HBeAg may have severe liver disease that progresses more rapidly to cirrhosis, or alternatively, they are identified clinically later in the course of the natural history of chronic hepatitis B, when the disease is more advanced. Both "wild-type" HBV and precore-mutant HBV can coexist in the same patient, or mutant HBV may arise late during wild-type HBV infection. In addition, clusters of fulminant hepatitis B in Israel and Japan were attributed to common-source infection with a precore mutant. Fulminant hepatitis B in North America and western Europe, however, occurs in patients infected with wild-type HBV, in the absence of precore mutants, and both precore mutants and other mutations throughout the HBV genome occur commonly, even in patients with typical, self-limited, milder forms of HBV infection. HBeAg-negative chronic hepatitis with mutations in the precore region is now the most frequently encountered form of hepatitis B in Mediterranean countries and in Europe. In the United States, where HBV genotype A (less prone to G1896A mutation) is prevalent, precore-mutant HBV is much less common; however, as a result of immigration from Asia and Europe, the proportion of HBeAg-negative hepatitis B–infected individuals has increased in the United States, and they now represent ~30–40% of patients with chronic hepatitis B. Characteristic of such HBeAg-negative chronic hepatitis B are lower levels of HBV DNA (usually $\leq 10^5$ IU/mL) and one of several patterns of aminotransferase activity—persistent elevations, periodic fluctuations above the normal range, and periodic fluctuations between the normal and elevated range.

The second important category of HBV mutants consists of *escape mutants*, in which a single amino acid substitution, from glycine to arginine, occurs at position 145 of the immunodominant *a* determinant common to all HBsAg subtypes. This HBsAg alteration leads to a critical conformational change that results in a loss of neutralizing activity by anti-HBs. This specific HBV/*a* mutant has been observed in two situations, active and passive immunization, in which humoral immunologic pressure may favor evolutionary change ("escape") in the virus—in a small number of hepatitis B vaccine recipients who acquired HBV infection despite the prior appearance of neutralizing anti-HBs and in HBV-infected liver transplant recipients treated with a high-potency human monoclonal anti-HBs preparation. Although such mutants have not been recognized frequently, their existence raises a concern that may complicate vaccination strategies and serologic diagnosis.

Different types of mutations emerge during antiviral therapy of chronic hepatitis B with nucleoside analogues; such "YMDD" and similar mutations in the polymerase motif of HBV are described in Chap. 334.

EXTRAHEPATIC SITES

Hepatitis B antigens and HBV DNA have been identified in extrahepatic sites, including the lymph nodes, bone marrow, circulating lymphocytes, spleen, and pancreas. Although the virus does not appear to be associated with tissue injury in any of these extrahepatic sites, its presence in these "remote" reservoirs has been invoked (but is not necessary) to explain the recurrence of HBV infection after orthotopic liver transplantation. The clinical relevance of such extrahepatic HBV is limited.

Hepatitis D

The delta hepatitis agent, or HDV, the only member of the genus *Deltavirus*, is a defective RNA virus that co-infects with and requires the helper function of HBV (or other hepadnaviruses) for its replication and expression. Slightly smaller than HBV, HDV is a formalin-sensitive, 35- to 37-nm virus with a hybrid structure. Its nucleocapsid expresses HDV antigen (HDAg), which bears no antigenic homology with any of the HBV antigens, and contains the virus genome. The HDV core is "encapsidated" by an outer envelope of HBsAg, indistinguishable from that of HBV except in its relative compositions of major, middle, and large HBsAg component proteins. The genome is a small, 1700-nucleotide, circular, single-strand RNA of negative polarity that is nonhomologous with HBV DNA (except for a small area of the polymerase gene) but that has features and the rolling circle model of replication common to genomes of plant satellite viruses or viroids. HDV RNA contains many areas of internal complementarity; therefore, it can fold on itself by internal base pairing to form an unusual, very stable, rodlike structure that contains a very stable, self-cleaving and self-ligating ribozyme. HDV RNA requires host RNA polymerase II for its replication in the hepatocyte nucleus via RNA-directed RNA synthesis by transcription of genomic RNA to a complementary antigenomic (plus strand) RNA; the antigenomic RNA, in turn, serves as a template for subsequent genomic RNA synthesis effected by host RNA polymerase I. HDV RNA has only one open reading frame, and HDAg, a product of the antigenomic strand, is the only known HDV protein; HDAg exists in two forms: a small, 195-amino-acid species, which plays a role in facilitating HDV RNA replication, and a large, 214-amino-acid species, which appears to suppress replication but is required for assembly of the antigen into virions. HDV antigens have been shown to bind directly to RNA polymerase II, resulting in stimulation of transcription. Although complete hepatitis D virions and liver injury require the cooperative helper function of HBV, intracellular replication of HDV RNA can occur without HBV. Genomic heterogeneity among HDV isolates has been described; however, pathophysiologic and clinical consequences of this genetic diversity have not been recognized. The clinical spectrum of hepatitis D is common to all eight genotypes identified, the predominant of which is genotype 1.

HDV can either infect a person simultaneously with HBV (*co-infection*) or superinfect a person already infected with HBV (*superinfection*); when HDV infection is transmitted from a donor with one HBsAg subtype to an HBsAg-positive recipient with a different subtype, HDV assumes the HBsAg subtype of the recipient, rather than the donor. Because HDV relies absolutely on HBV, the duration of HDV infection is determined by the duration of (and cannot outlast) HBV infection. HDV replication tends to suppress HBV replication; therefore, patients with hepatitis D tend to have lower levels of HBV replication. HDV antigen is expressed primarily in hepatocyte nuclei and is occasionally detectable in serum. During acute HDV infection, anti-HDV of the IgM class predominates, and 30–40 days may elapse after symptoms appear before anti-HDV can be detected. In self-limited infection, anti-HDV is low-titer and transient, rarely remaining detectable beyond the clearance of HBsAg and HDV antigen. In chronic HDV infection, anti-HDV circulates in high titer, and both IgM and IgG anti-HDV can be detected. HDV antigen in the liver and HDV RNA in serum and liver can be detected during HDV replication.

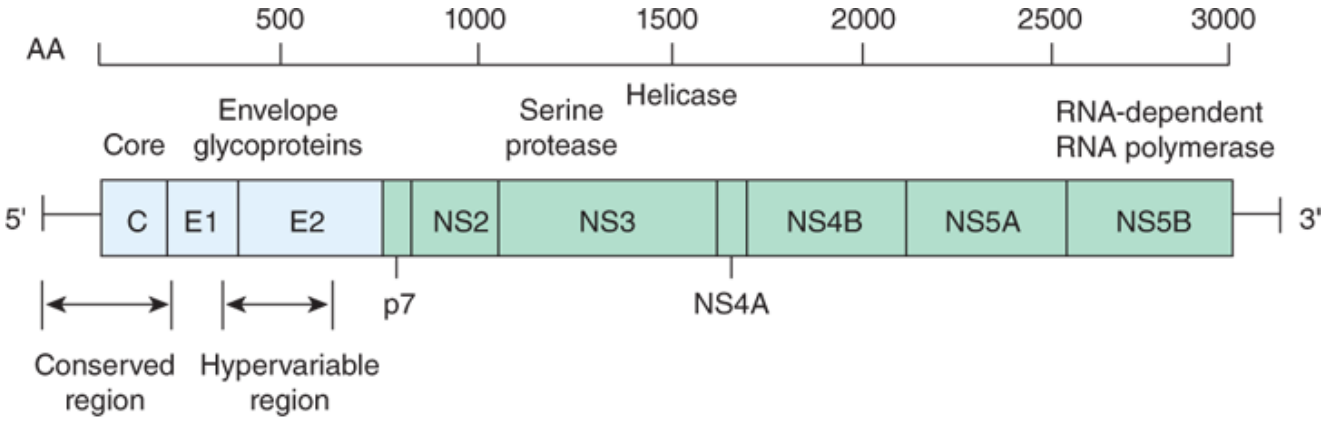
Hepatitis C

Hepatitis C virus, which, before its identification was labeled "non-A, non-B hepatitis," is a linear, single-strand, positive-sense, 9600-nucleotide RNA virus, the genome of which is similar in organization to that of flaviviruses and pestiviruses; HCV is the only member of the genus *Hepacivirus* in the family Flaviviridae. The HCV genome contains a single, large open reading frame (ORF) (gene) that codes for a virus polyprotein of ~3000 amino acids, which is cleaved after translation to yield 10 viral proteins. The 5′ end of the genome consists of an untranslated region (containing an internal ribosomal entry site [IRES]) adjacent to the genes for three structural proteins, the nucleocapsid core protein, C, and two envelope glycoproteins, E1 and E2. The 5′ untranslated region and core gene are highly conserved among genotypes, but the envelope proteins are coded for by the hypervariable region, which varies from isolate to isolate and may allow the virus to evade host immunologic containment directed at accessible virus-envelope proteins. The 3′ end of the genome

also includes an untranslated region and contains the genes for seven nonstructural (NS) proteins: p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B. p7 is a membrane ion channel protein necessary for efficient assembly and release of HCV. The NS2 cysteine protease cleaves NS3 from NS2, and the NS3-4A serine protease cleaves all the downstream proteins from the polyprotein. Important NS proteins involved in virus replication include the NS3 helicase; NS3-4A serine protease; the multifunctional membrane-associated phosphoprotein NS5A, an essential component of the viral replication membranous web (along with NS4B); and the NS5B RNA-dependent RNA polymerase **(Fig. 332-6)**. Because HCV does not replicate via a DNA intermediate, it does not integrate into the host genome. Because HCV tends to circulate in relatively low titer, 10^3–10^7 virions/mL, visualization of the 50- to 80-nm virus particles remains difficult. Still, the replication rate of HCV is very high, 10^{12} virions per day; its half-life is 2.7 h. The chimpanzee is a helpful but cumbersome animal model. Although a robust, reproducible, small animal model is lacking, HCV replication has been documented in an immunodeficient mouse model containing explants of human liver and in transgenic mouse and rat models. Although in vitro replication is difficult, replicons in hepatocellular carcinoma–derived cell lines support replication of genetically manipulated, truncated, or full-length HCV RNA (but not intact virions); infectious pseudotyped retroviral HCV particles have been shown to yield functioning envelope proteins. In 2005, complete replication of HCV and intact 55-nm virions were described in cell culture systems. HCV entry into the hepatocyte occurs via the nonliver-specific CD81 receptor and the liver-specific tight junction protein claudin-1. A growing list of additional host receptors to which HCV binds on cell entry includes occludin, low-density lipoprotein receptors, glycosaminoglycans, scavenger receptor B1, and epidermal growth factor receptor, among others. Relying on the same assembly and secretion pathway as low-density and very-low-density lipoproteins, HCV is a lipoviroparticle and masquerades as a lipoprotein, which may limit its visibility to the adaptive immune system and explain its ability to evade immune containment and clearance. After viral entry and uncoating, translation is initiated by the IRES on the endoplasmic reticulum membrane, and the HCV polyprotein is cleaved during translation and posttranslationally by host cellular proteases as well as HCV NS2-3 and NS3-4A proteases. Host cofactors involved in HCV replication include cyclophilin A, which binds to NS5A and yields conformational changes required for viral replication, and liver-specific host microRNA miR-122.

FIGURE 332-6

Organization of the hepatitis C virus genome and its associated, 3000-amino-acid (AA) proteins. The three structural genes at the 5′ end are the core region, C, which codes for the nucleocapsid, and the envelope regions, E1 and E2, which code for envelope glycoproteins. The 5′ untranslated region and the C region are highly conserved among isolates, whereas the envelope domain E2 contains the hypervariable region. At the 3′ end are seven nonstructural (NS) regions—p7, a membrane protein adjacent to the structural proteins that appears to function as an ion channel; NS2, which codes for a cysteine protease; NS3, which codes for a serine protease and an RNA helicase; NS4 and NS4B; NS5A, a multifunctional membrane-associated phosphoprotein, an essential component of the viral replication membranous web; and NS5B, which codes for an RNA-dependent RNA polymerase. After translation of the entire polyprotein, individual proteins are cleaved by both host and viral proteases.



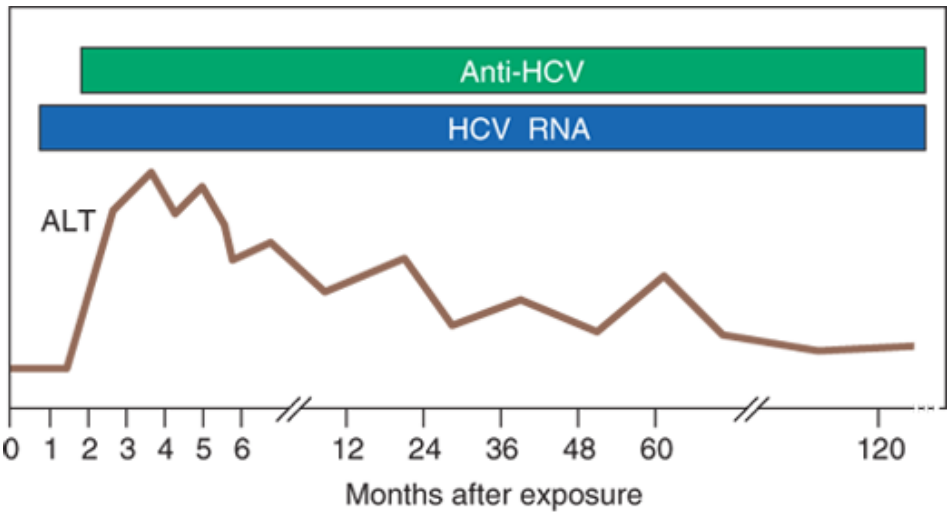
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

At least six distinct major genotypes (and a minor genotype 7), as well as >50 subtypes within genotypes, of HCV have been identified by nucleotide sequencing. Genotypes differ from one another in sequence homology by ≥30%, and subtypes differ by ~20%. Because divergence of HCV isolates within a genotype or subtype and within the same host may vary insufficiently to define a distinct genotype, these intragenotypic differences are referred to as *quasispecies* and differ in sequence homology by only a few percent. The genotypic and quasispecies diversity of HCV, resulting from its high mutation rate, interferes with effective humoral immunity. Neutralizing antibodies to HCV have been demonstrated, but they tend to be short-lived, and HCV infection does not induce lasting immunity against reinfection with different virus isolates or even the same virus isolate. Thus, neither *heterologous* nor *homologous* immunity appears to develop commonly after acute HCV infection. Some HCV genotypes are distributed worldwide, whereas others are more geographically confined (see "Epidemiology and Global Features"). In addition, differences exist among genotypes in responsiveness to antiviral therapy but not in pathogenicity or clinical progression (except for genotype 3, in which hepatic steatosis and clinical progression are more likely).

Currently available, third-generation immunoassays, which incorporate proteins from the core, NS3, and NS5 regions, detect anti-HCV antibodies during acute infection. The most sensitive indicator of HCV infection is the presence of HCV RNA, which requires molecular amplification by PCR or transcription-mediated amplification (TMA) **(Fig. 332-7)**. To allow standardization of the quantification of HCV RNA among laboratories and commercial assays, HCV RNA is reported as international units (IUs) per milliliter; quantitative assays with a broad dynamic range are available that allow detection of HCV RNA with a sensitivity as low as 5 IU/mL. HCV RNA can be detected within a few days of exposure to HCV—well before the appearance of anti-HCV—and tends to persist for the duration of HCV infection. Application of sensitive molecular probes for HCV RNA has revealed the presence of replicative HCV in peripheral blood lymphocytes of infected persons; however, as is the case for HBV in lymphocytes, the clinical relevance of HCV lymphocyte infection is not known.

FIGURE 332-7

Scheme of typical laboratory features during acute hepatitis C progressing to chronicity. Hepatitis C virus (HCV) RNA is the first detectable event, preceding alanine aminotransferase (ALT) elevation and the appearance of anti-HCV.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Hepatitis E

Previously labeled *epidemic* or *enterically transmitted non-A, non-B hepatitis*, HEV is an enterically transmitted virus that causes clinically apparent hepatitis primarily in India, Asia, Africa, and Central America; in those geographic areas, HEV is the most common cause of acute hepatitis; one-third of the global population appears to have been infected. This agent, with epidemiologic features resembling those of hepatitis A, is a 27- to 34-nm, nonenveloped, heat-stable, HAV-like virus with a 7200-nucleotide, single-strand, positive-sense RNA genome. HEV has three overlapping ORFs (genes), the largest of which, *ORF1*, encodes nonstructural proteins involved in virus replication (viral replicase). A middle-sized gene, *ORF2*, encodes the nucleocapsid protein, the major structural protein, and the smallest, *ORF3*, encodes a small structural protein involved in virus particle secretion. All HEV isolates appear to belong to a single serotype, despite genomic heterogeneity of up to 25% and the existence of five genotypes, only four of which have been detected in humans; genotypes 1 and 2 (common in developing countries) appear to be more virulent, whereas genotypes 3 (the most common in the United States and Europe) and 4 (seen in China) are more attenuated and account for subclinical infections. Contributing to the perpetuation of this virus are animal reservoirs, most notably in swine but also in camels, deer, rats, and rabbits, among others. No genomic or antigenic homology, however, exists between HEV and HAV or other picornaviruses; and HEV, although resembling caliciviruses, is sufficiently distinct from any known agent to merit its own classification as a unique genus, *Hepevirus*, within the family Hepeviridae. The virus has been detected in stool, bile, and liver, and is excreted in the stool during the late incubation period. Both IgM anti-HEV during early acute infection and IgG anti-HEV predominating after the first 3 months can be detected. The presence of HEV RNA in serum and stool accompanies acute infection; viremia resolves as clinical-biochemical recovery ensues, while HEV RNA in stool may outlast viremia by several weeks. Currently, availability and reliability of serologic/virologic testing for HEV infection is limited—and not FDA-approved or licensed—but can be done in specialized laboratories (e.g., the Centers for Disease Control and Prevention).

PATHOGENESIS

Under ordinary circumstances, none of the hepatitis viruses is known to be directly cytopathic to hepatocytes. Evidence suggests that the clinical manifestations and outcomes after acute liver injury associated with viral hepatitis are determined by the immunologic responses of the host. Among the viral hepatitides, the immunopathogenesis of hepatitis B and C has been studied most extensively.

For HBV, the existence of inactive hepatitis B carriers with normal liver histology and function suggests that the virus is not directly cytopathic. The fact that patients with defects in cellular immune competence are more likely to remain chronically infected rather than to clear HBV supports the role of cellular immune responses in the pathogenesis of hepatitis B–related liver injury. The model that has the most experimental support involves cytolytic T cells sensitized specifically to recognize host and hepatitis B viral antigens on the liver cell surface. Nucleocapsid proteins (HBcAg and possibly HBeAg), present on the cell membrane in minute quantities, are the viral target antigens that, with host antigens, invite cytolytic T cells to destroy HBV-infected hepatocytes. Differences in the robustness and broad polyclonality of CD8+ cytolytic T cell responsiveness; in the level of HBV-specific helper CD4+ T cells; in attenuation, depletion, and exhaustion of virus-specific T cells; in viral T cell epitope escape mutations that allow the virus to evade T cell containment; and in the elaboration of antiviral cytokines by T cells have been invoked to explain differences in outcomes between those who recover after acute hepatitis and those who progress to chronic hepatitis, or between those with mild and those with severe (fulminant) acute HBV infection.

Although a robust cytolytic T cell response occurs and eliminates virus-infected liver cells during acute hepatitis B, >90% of HBV DNA has been found in experimentally infected chimpanzees to disappear from the liver and blood before maximal T cell infiltration of the liver and before most of the biochemical and histologic evidence of liver injury. This observation suggests that components of the innate immune system and inflammatory cytokines, independent of cytopathic antiviral mechanisms, participate in the early immune response to HBV infection; this effect has been shown to represent elimination of HBV replicative intermediates from the cytoplasm and covalently closed circular viral DNA from the nucleus of infected hepatocytes. In turn, the innate immune response to HBV infection is mediated largely by natural killer (NK) cell cytotoxicity, activated by immunosuppressive cytokines (e.g., interleukin [IL] 10 and transforming growth factor [TGF] β), reduced signals from inhibitory receptor expression (e.g., major histocompatibility complex), or increased signals from activating receptor expression on infected hepatocytes. In addition, NK cells reduce helper CD4+ cells, which results in reduced CD8+ cells and exhaustion of the virus-specific T cell response to HBV infection. Ultimately, HBV-HLA–specific cytolytic T cell responses of the adaptive immune system are felt to be responsible for recovery from HBV infection.

Debate continues over the relative importance of viral and host factors in the pathogenesis of HBV-associated liver injury and its outcome. As noted above, precore genetic mutants of HBV have been associated with the more severe outcomes of HBV infection (severe chronic and fulminant hepatitis), suggesting that, under certain circumstances, relative pathogenicity is a property of the virus, not the host. The facts that concomitant HDV and HBV infections are associated with more severe liver injury than HBV infection alone and that cells transfected in vitro with the gene for HDV antigen express HDV antigen and then become necrotic in the absence of any immunologic influences are also consistent with a viral effect on pathogenicity. Similarly, in patients who undergo liver transplantation for end-stage chronic hepatitis B, occasionally, rapidly progressive liver injury appears in the new liver. This clinical pattern is associated with an unusual histologic pattern in the new liver, *fibrosing cholestatic hepatitis*, which, ultrastructurally, appears to represent a choking of the cell with overwhelming quantities of HBsAg. This observation suggests that, under the influence of the potent immunosuppressive agents required to prevent allograft rejection, HBV may have a direct cytopathic effect on liver cells, independent of the immune system.

Although the precise mechanism of liver injury in HBV infection remains elusive, studies of nucleocapsid proteins have shed light on the profound immunologic tolerance to HBV of babies born to mothers with highly replicative (HBeAg-positive), chronic HBV infection. In HBeAg-expressing transgenic mice, in utero exposure to HBeAg, which is sufficiently small to traverse the placenta, induces T cell tolerance to both nucleocapsid proteins. This, in turn, may explain why,

when infection occurs so early in life, immunologic clearance does not occur, and protracted, lifelong infection ensues. An alternative explanation proposed to explain why robust liver injury does not accompany neonatal HBV infection but predisposes to chronic infection is defective priming of HBV-specific T cells during in utero exposure to HBV.

An important distinction should be drawn between HBV infection acquired at birth, common in endemic areas, such as East Asia, and infection acquired in adulthood, common in the West. Infection in the neonatal period is associated with the acquisition of what appears to be a high level of immunologic tolerance to HBV and absence of an acute hepatitis illness, but the almost invariable establishment of chronic, often lifelong infection. Neonatally acquired HBV infection can culminate decades later in cirrhosis and hepatocellular carcinoma (see "Complications and Sequelae"). In contrast, when HBV infection is acquired during adolescence or early adulthood, the host immune response to HBV-infected hepatocytes tends to be robust, an acute hepatitis-like illness is the rule, and failure to recover is the exception. After adulthood-acquired infection, chronicity is uncommon, and the risk of hepatocellular carcinoma is very low. Based on these observations, some authorities categorize HBV infection into an "immunotolerant" phase, an "immunoreactive" phase, and an "inactive" phase. This somewhat simplistic formulation does not apply at all to the typical adult in the West with self-limited acute hepatitis B, in whom no period of immunologic tolerance occurs. Even among those with neonatally acquired HBV infection, in whom immunologic tolerance appears to be established definitively, immunologic responses to HBV infection have been demonstrated, and intermittent bursts of hepatic necroinflammatory activity punctuate the early decades of life during which liver injury appears to be quiescent (labeled by some as the "immunotolerant" phase; however, it more accurately is a period of dissociation between high-level HBV replication and a paucity of inflammatory liver injury). In addition, even when clinically apparent liver injury and progressive fibrosis emerge during later decades (the so-called immunoreactive, or immunointolerant, phase), the level of immunologic tolerance to HBV remains substantial. More accurately, in patients with neonatally acquired HBV infection, a dynamic equilibrium exists between tolerance and intolerance, the outcome of which determines the clinical expression of chronic infection. Persons infected as neonates tend to have a relatively higher level of immunologic tolerance (high replication, low necroinflammatory activity) during the early decades of life and a relatively lower level (but only rarely a loss) of tolerance (and necroinflammatory activity reflecting the level of virus replication) in the later decades of life.

Hepatitis C

Cell-mediated immune responses and elaboration by T cells of antiviral cytokines contribute to the multicellular innate and adaptive immune responses involved in the containment of infection and pathogenesis of liver injury associated with hepatitis C. The fact that HCV is so efficient in evading these immune mechanisms is a testament to its highly evolved ability to disrupt host immune responses at multiple levels. After exposure to HCV, the host cell identifies viral product motifs (pattern recognition receptors) that distinguish the virus from "self," resulting in the elaboration of interferons and other cytokines that result in activation of innate and adaptive immune responses. Intrahepatic HLA class I–restricted cytolytic T cells directed at nucleocapsid, envelope, and nonstructural viral protein antigens have been demonstrated in patients with chronic hepatitis C; however, such virus-specific cytolytic T cell responses do not correlate adequately with the degree of liver injury or with recovery. Yet, a consensus has emerged supporting a role in the pathogenesis of HCV-associated liver injury of virus-activated CD4+ helper T cells that stimulate, via the cytokines they elaborate, HCV-specific CD8+ cytotoxic T cells. These responses appear to be more robust (higher in number, more diverse in viral antigen specificity, more functionally effective, and more long lasting) in those who recover from HCV infection than in those who have chronic infection. Contributing to chronic infection are a CD4+ proliferative defect that results in rapid contraction of CD4+ responses, mutations in CD8+ T cell–targeted viral epitopes that allow HCV to escape immune-mediated clearance, and upregulation of inhibitory receptors on functionally impaired, exhausted T cells. Although attention has focused on adaptive immunity, HCV proteins have been shown to interfere with innate immunity

by resulting in blocking of type 1 interferon responses and inhibition of interferon signaling and effector molecules in the interferon signaling cascade. Several HLA alleles have been linked with self-limited hepatitis C, the most convincing of which is the CC haplotype of the *IL28B* gene, which codes for interferon λ3, a component of innate immune antiviral defense. The *IL28B* association is even stronger when combined with HLA class II *DQB1*03:01*. The link between non-CC *IL28B* polymorphisms and failure to clear HCV infection has been explained by a chromosome 19q13.13 frameshift variant upstream of *IL28B*, the ΔG polymorphism of which creates an ORF in a novel interferon gene (*IFN-l4*) associated with impaired HCV clearance. Also shown to contribute to limiting HCV infection are NK cells of the innate immune system that function when HLA class I molecules required for successful adaptive immunity are underexpressed. Both peripheral cytotoxicity and intrahepatic NK cell cytotoxicity are dysfunctional in persistent HCV infection. Adding to the complexity of the immune response, HCV core, NS4B, and NS5B have been shown to suppress the immunoregulatory nuclear factor (NF)-κB pathway, resulting in reduced antiapoptotic proteins and a resultant increased vulnerability to tumor necrosis factor (TNF) α–mediated cell death. Patients with hepatitis C and unfavorable (non-CC, associated with reduced HCV clearance) *IL28B* alleles have been shown to have depressed NK cell/innate immune function. Of note, the emergence of substantial viral quasispecies diversity and HCV sequence variation allow the virus to evade attempts by the host to contain HCV infection by both humoral and cellular immunity.

Finally, cross-reactivity between viral antigens (HCV NS3 and NS5A) and host autoantigens (cytochrome P450 2D6) has been invoked to explain the association between hepatitis C and a subset of patients with autoimmune hepatitis and antibodies to liver-kidney microsomal (LKM) antigen (anti-LKM) **(Chap. 334)**.

EXTRAHEPATIC MANIFESTATIONS

Immune complex–mediated tissue damage appears to play a pathogenetic role in the extrahepatic manifestations of acute hepatitis B. The occasional prodromal serum sickness–like syndrome observed in acute hepatitis B appears to be related to the deposition in tissue blood vessel walls of HBsAg-anti-HBs circulating immune complexes, leading to activation of the complement system and depressed serum complement levels.

In patients with chronic hepatitis B, other types of immune-complex disease may be seen. Glomerulonephritis with the nephrotic syndrome is observed occasionally; HBsAg, immunoglobulin, and C3 deposition has been found in the glomerular basement membrane. Whereas generalized vasculitis (polyarteritis nodosa) develops in considerably <1% of patients with chronic HBV infection, 20–30% of patients with polyarteritis nodosa have HBsAg in serum **(Chap. 356)**. In these patients, the affected small- and medium-size arterioles contain HBsAg, immunoglobulins, and complement components. Another extrahepatic manifestation of viral hepatitis, essential mixed cryoglobulinemia (EMC), was reported initially to be associated with hepatitis B. The disorder is characterized clinically by arthritis, cutaneous vasculitis (palpable purpura), and, occasionally, glomerulonephritis and serologically by the presence of circulating cryoprecipitable immune complexes of more than one immunoglobulin class **(Chaps. 308 and 356)**. Many patients with this syndrome have chronic liver disease, but the association with HBV infection is limited; instead, a substantial proportion has chronic HCV infection, with circulating immune complexes containing HCV RNA. Immune-complex glomerulonephritis is another recognized extrahepatic manifestation of chronic hepatitis C. Immune-complex disorders have been linked, albeit rarely, with both hepatitis A and E. In hepatitis E, rare neurologic complications have been postulated to result from both immunologic mechanisms and/or direct CNS infection with the virus.

PATHOLOGY

The typical morphologic lesions of all types of viral hepatitis are similar and consist of panlobular infiltration with mononuclear cells, hepatic cell necrosis, hyperplasia of Kupffer cells, and variable degrees of cholestasis. Hepatic cell

regeneration is present, as evidenced by numerous mitotic figures, multinucleated cells, and “rosette” or “pseudoacinar” formation. The mononuclear infiltration consists primarily of small lymphocytes, although plasma cells and eosinophils occasionally are present. Liver cell damage consists of hepatic cell degeneration and necrosis, cell dropout, ballooning of cells, and acidophilic degeneration of hepatocytes (forming so-called Councilman or apoptotic bodies). Large hepatocytes with a ground-glass appearance of the cytoplasm may be seen in chronic but not in acute HBV infection; these cells contain HBsAg and can be identified histochemically with orcein or aldehyde fuchsin. In uncomplicated viral hepatitis, the reticulin framework is preserved.

In hepatitis C, the histologic lesion is often remarkable for a relative paucity of inflammation, a marked increase in activation of sinusoidal lining cells, lymphoid aggregates, the presence of fat (more frequent in genotype 3 and linked to increased fibrosis), and, occasionally, bile duct lesions in which biliary epithelial cells appear to be piled up without interruption of the basement membrane. Occasionally, microvesicular steatosis occurs in hepatitis D. In hepatitis E, a common histologic feature is marked cholestasis. A cholestatic variant of slowly resolving acute hepatitis A also has been described.

A more severe histologic lesion, *bridging hepatic necrosis*, also termed *subacute* or *confluent necrosis* or *interface hepatitis*, is observed occasionally in acute hepatitis. “Bridging” between lobules results from large areas of hepatic cell dropout, with collapse of the reticulin framework. Characteristically, the bridge consists of condensed reticulum, inflammatory debris, and degenerating liver cells that span adjacent portal areas, portal to central veins, or central vein to central vein. This lesion had been thought to have prognostic significance; in many of the originally described patients with this lesion, a subacute course terminated in death within several weeks to months, or severe chronic hepatitis and cirrhosis developed; however, the association between bridging necrosis and a poor prognosis in patients with acute hepatitis has not been upheld. Therefore, although demonstration of this lesion in patients with chronic hepatitis has prognostic significance (**Chap. 334**), its demonstration during acute hepatitis is less meaningful, and liver biopsies to identify this lesion are no longer undertaken routinely in patients with acute hepatitis. In *massive hepatic necrosis* (fulminant hepatitis, “acute yellow atrophy”), the striking feature at postmortem examination is the finding of a small, shrunken, soft liver. Histologic examination reveals massive necrosis and dropout of liver cells of most lobules with extensive collapse and condensation of the reticulin framework. When histologic documentation is required in the management of fulminant or very severe hepatitis, a biopsy can be done by the angiographically guided transjugular route, which permits the performance of this invasive procedure in the presence of severe coagulopathy.

Immunohistochemical and electron-microscopic studies have localized HBsAg to the cytoplasm and plasma membrane of infected liver cells. In contrast, HBcAg predominates in the nucleus, but, occasionally, scant amounts are also seen in the cytoplasm and on the cell membrane. HDV antigen is localized to the hepatocyte nucleus, whereas HAV, HCV, and HEV antigens are localized to the cytoplasm.

EPIDEMIOLOGY AND GLOBAL FEATURES


Before the availability of serologic tests for hepatitis viruses, all viral hepatitis cases were labeled either as “infectious” or “serum” hepatitis. Modes of transmission overlap, however, and *a clear distinction among the different types of viral hepatitis cannot be made solely on the basis of clinical or epidemiologic features* (**Table 332-2**). The most accurate means to distinguish the various types of viral hepatitis involves specific serologic testing.

TABLE 332-2

Clinical and Epidemiologic Features of Viral Hepatitis

FEATURE	HAV	HBV	HCV	HDV	HEV
Incubation (days)	15–45, mean 30	30–180, mean 60–90	15–160, mean 50	30–180, mean 60–90	14–60, mean 40
Onset	Acute	Insidious or acute	Insidious or acute	Insidious or acute	Acute
Age preference	Children, young adults	Young adults (sexual and percutaneous), babies, toddlers	Any age, but more common in adults	Any age (similar to HBV)	Epidemic cases: young adults (20–40 years); sporadic cases: older adults (>60)
Transmission Fecal-oral Percutaneous Perinatal Sexual	+++ Unusual – ±	– +++ +++ ++	– +++ ±[a] ±[a]	– +++ + ++	+++ – – –
Clinical Severity Fulminant Progression to chronicity Carrier Cancer Prognosis	Mild 0.1% None None Excellent	Occasionally severe 0.1–1% Occasional (1–10%) (90% of neonates) 0.1–30%[c] + (neonatal infection) Worse with age, debility	Moderate 0.1% Common (85%) 1.5–3.2% + Moderate	Occasionally severe 5–20%[b] Common[d] Variable[g] ± Acute, good Chronic, poor	Mild 1–2%[e] None[f] None None Good
Prophylaxis	Ig, inactivated vaccine	HBIG, recombinant vaccine	None	HBV vaccine (none for HBV carriers)	Vaccine

FEATURE	HAV	HBV	HCV	HDV	HEV
Therapy	None	Interferon Lamivudine Adefovir Pegylated interferon[h] Entecavir[h] Telbivudine Tenofovir[h]	Pegylated interferon ribavirin telaprevir,[i] boceprevir,[i] simeprevir, sofosbuvir, ledipasvir, paritaprevir/ritonavir ombitasvir, dasabuvir daclatasvir, velpatasvir, grazoprevir, elbasvir	Pegylated interferon ±	None[j]

[a]Primarily with HIV co-infection and high-level viremia in index case; more likely in persons with multiple sex partners or sexually transmitted diseases; risk ~5%. [b]Up to 5% in acute HBV/HDV co-infection; up to 20% in HDV superinfection of chronic HBV infection.[c] Varies considerably throughout the world and in subpopulations within countries; see text. [d]In acute HBV/HDV co-infection, the frequency of chronicity is the same as that for HBV; in HDV superinfection, chronicity is invariable. [e]10–20% in pregnant women. [f]Except as observed in immunosuppressed liver allograft recipients or other immunosuppressed hosts. [g]Common in Mediterranean countries; rare in North America and western Europe. [h]First-line agents. [i]No longer recommended. [j]Anecdotal reports and retrospective studies suggest that pegylated interferon and/or ribavirin are effective in treating chronic hepatitis E, observed in immunocompromised persons; ribavirin monotherapy has been used successfully in acute, severe hepatitis E.

Abbreviation: HBIG, hepatitis B immunoglobulin. See text for other abbreviations.

Hepatitis A

This agent is transmitted almost exclusively by the fecal-oral route. Person-to-person spread of HAV is enhanced by poor personal hygiene and overcrowding; large outbreaks as well as sporadic cases have been traced to contaminated food, water, milk, frozen raspberries and strawberries, green onions imported from Mexico, and shellfish (e.g., scallops imported from the Philippines used to make sushi, the culprit identified in a 2016 Hawaiian outbreak). Intrafamily and intrainstitutional spreads are also common. Early epidemiologic observations supported a predilection for hepatitis A to occur in late fall and early winter. In temperate zones, epidemic waves have been recorded every 5–20 years as new segments of nonimmune population appeared; however, in developed countries, the incidence of hepatitis A has been declining, presumably as a function of improved sanitation, and these cyclic patterns are no longer observed. No HAV carrier state has been identified after acute hepatitis A; perpetuation of the virus in nature depends presumably on nonepidemic, inapparent subclinical infection, ingestion of contaminated food or water in, or imported from, endemic areas, and/or contamination linked to environmental reservoirs.

In the general population, anti-HAV, a marker for previous HAV infection, increases in prevalence as a function of increasing age and of decreasing socioeconomic status. In the 1970s, serologic evidence of prior hepatitis A infection occurred in ~40% of urban populations in the United States, most of whose members never recalled having had a symptomatic case of hepatitis. In subsequent decades, however, the prevalence of anti-HAV has been declining in the United States. In developing countries, exposure, infection, and subsequent immunity are almost universal in childhood. As the frequency of subclinical childhood infections declines in developed countries, a susceptible cohort of adults

emerges. Hepatitis A tends to be more symptomatic in adults; therefore, paradoxically, as the frequency of HAV infection declines, the likelihood of clinically apparent, even severe, HAV illnesses increases in the susceptible adult population. Travel to endemic areas is a common source of infection for adults from nonendemic areas. More recently recognized epidemiologic foci of HAV infection include child care centers, neonatal intensive care units, promiscuous men who have sex with men, injection drug users, and unvaccinated close contacts of newly arrived international adopted children, most of whom emanate from countries with intermediate-to-high hepatitis A endemicity. Although hepatitis A is rarely bloodborne, several outbreaks have been recognized in recipients of clotting-factor concentrates. In the United States, the introduction of hepatitis A vaccination programs among children from high-incidence states has resulted in a >70% reduction in the annual incidence of new HAV infections and has shifted the burden of new infections from children to adults. In the 2007–2012 U.S. Public Health Service National Health and Nutrition Examination Survey (NHANES), the prevalence of anti-HAV in the U.S. population aged ≥20 years had declined to 24.2% from the 29.5% measured in NHANES 1999–2006. While universal childhood vaccination accounted for a high prevalence of vaccine-induced immunity in children aged 2–19 years, the lowest age-specific prevalence of anti-HAV (16.1–17.6%) occurred in adults in the fourth and fifth decades, respectively (aged 30–49 years). This is a subgroup of the population who remain susceptible to acute hepatitis A acquired during travel to endemic areas and from contaminated foods, especially those imported from endemic countries.

Hepatitis B

Percutaneous inoculation has long been recognized as a major route of hepatitis B transmission, but the outmoded designation “serum hepatitis” is an inaccurate label for the epidemiologic spectrum of HBV infection. As detailed below, most of the hepatitis transmitted by blood transfusion is not caused by HBV; moreover, in approximately two-thirds of patients with acute type B hepatitis, no history of an identifiable percutaneous exposure can be elicited. We now recognize that many cases of hepatitis B result from less obvious modes of nonpercutaneous or covert percutaneous transmission. HBsAg has been identified in almost every body fluid from infected persons, and at least some of these body fluids—most notably semen and saliva—are infectious, albeit less so than serum, when administered percutaneously or nonpercutaneously to experimental animals. Among the nonpercutaneous modes of HBV transmission, oral ingestion has been documented as a potential but inefficient route of exposure. By contrast, the two nonpercutaneous routes considered to have the greatest impact are intimate (especially sexual) contact and perinatal transmission.

In sub-Saharan Africa, intimate contact among toddlers is considered instrumental in contributing to the maintenance of the high frequency of hepatitis B in the population. Perinatal transmission occurs primarily in infants born to mothers with chronic hepatitis B or (rarely) mothers with acute hepatitis B during the third trimester of pregnancy or during the early postpartum period. Perinatal transmission is uncommon in North America and western Europe but occurs with great frequency and is the most important mode of HBV perpetuation in East Asia and developing countries. Although the precise mode of perinatal transmission is unknown, and although ~10% of infections may be acquired in utero, epidemiologic evidence suggests that most infections occur approximately at the time of delivery and are not related to breast-feeding (which is not contraindicated in women with hepatitis B). The likelihood of perinatal transmission of HBV correlates with the presence of HBeAg and high-level viral replication; 90% of HBeAg-positive mothers but only 10–15% of anti-HBe-positive mothers transmit HBV infection to their offspring. In most cases, acute infection in the neonate is clinically asymptomatic, but the child is very likely to remain chronically infected.

The >350–400 million persons with chronic HBV infection in the world constitute the main reservoir of hepatitis B in human beings. Whereas serum HBsAg is infrequent (0.1–0.5%) in normal populations in the United States and western Europe, a prevalence of up to 5–20% has been found in East Asia and in some tropical countries; in persons with Down’s

syndrome, lepromatous leprosy, leukemia, Hodgkin's disease, or polyarteritis nodosa; in patients with chronic renal disease on hemodialysis; and in injection drug users.

Other groups with high rates of HBV infection include spouses of acutely infected persons; sexually promiscuous persons (especially promiscuous men who have sex with men); health care workers exposed to blood; persons who require repeated transfusions especially with pooled blood-product concentrates (e.g., hemophiliacs); residents and staff of custodial institutions for the developmentally handicapped; prisoners; and, to a lesser extent, family members of chronically infected patients. In volunteer blood donors, the prevalence of anti-HBs, a reflection of previous HBV infection, ranges from 5% to 10%, but the prevalence is higher in lower socioeconomic strata, older age groups, and persons—including those mentioned above—exposed to blood products. Because of highly sensitive virologic screening of donor blood, the risk of acquiring HBV infection from a blood transfusion is 1 in 230,000.

Prevalence of infection, modes of transmission, and human behavior conspire to mold geographically different epidemiologic patterns of HBV infection. In East Asia and Africa, hepatitis B, a disease of the newborn and young children, is perpetuated by a cycle of maternal-neonatal spread. In North America and western Europe, hepatitis B is primarily a disease of adolescence and early adulthood, the time of life when intimate sexual contact and recreational and occupational percutaneous exposures tend to occur. To some degree, however, this dichotomy between high-prevalence and low-prevalence geographic regions has been minimized by immigration from high-prevalence to low-prevalence areas. For example, in the United States, NHANES data from 2007 to 2012 revealed an overall prevalence of current HBV infection (detectable HBsAg) of 0.3%; however, the prevalence in Asian persons, 93% of whom were foreign-born, was 10-fold higher, 3.1%, representing 50% of the U.S. national disease burden. The introduction of hepatitis B vaccine in the early 1980s and adoption of universal childhood vaccination policies in many countries resulted in a dramatic, ~90% decline in the incidence of new HBV infections in those countries as well as in the dire consequences of chronic infection, including hepatocellular carcinoma. In the United States, as demonstrated in NHANES 2007–2012, following the 1991 implementation of universal childhood vaccination, HBsAg seropositivity had declined in children aged 6–19 years to as low as 0.03%, an ~85% reduction. Populations and groups for whom HBV infection screening is recommended are listed in **Table 332-3**.

TABLE 332-3

High-Risk Populations for Whom Hbv Infection Screening Is Recommended

Persons born in countries/regions with a high (≥8%) and intermediate (≥2%) prevalence of HBV infection including immigrants and adopted children and including persons born in the United States who were not vaccinated as infants and whose parents emigrated from areas of high HBV endemicity
Household and sexual contacts of persons with hepatitis B
Babies born to HBsAg-positive mothers
Persons who have used injection drugs
Persons with multiple sexual contacts or a history of sexually transmitted disease
Men who have sex with men
Inmates of correctional facilities
Persons with elevated alanine or aspartate aminotransferase levels
Blood/plasma/organ/tissue/semen donors
Persons with HCV or HIV infection
Hemodialysis patients
Pregnant women
Persons who are the source of blood or body fluids that would be an indication for postexposure prophylaxis (e.g., needlestick, mucosal exposure, sexual assault)
Persons who require immunosuppressive or cytotoxic therapy (including anti–tumor necrosis factor α therapy for rheumatologic or inflammatory bowel disorders)

Hepatitis D

Infection with HDV has a worldwide distribution, but two epidemiologic patterns exist. In Mediterranean countries (northern Africa, southern Europe, the Middle East), HDV infection is endemic among those with hepatitis B, and the disease is transmitted predominantly by nonpercutaneous means, especially close personal contact. In nonendemic areas, such as the United States (where hepatitis D is rare among persons with chronic hepatitis B) and northern Europe, HDV infection is confined to persons exposed frequently to blood and blood products, primarily injection drug users, (especially so in HIV-infected injection drug users) and hemophiliacs. In the United States, the prevalence of HDV infection in the national population is 0.02% (NHANES 1999–2012); however, among HBsAg-positive persons, the prevalence of HDV infection is highest in injection drug users (11%) and hemophiliacs (19%). HDV infection can be introduced into a population through drug users or by migration of persons from endemic to nonendemic areas. Thus, patterns of population migration and human behavior facilitating percutaneous contact play important roles in the introduction and amplification of HDV infection. Occasionally, the migrating epidemiology of hepatitis D is expressed in explosive outbreaks of severe hepatitis, such as those that have occurred in remote South American villages (e.g., “Lábrea fever” in the Amazon basin) as well as in urban centers in the United States. Ultimately, such outbreaks of hepatitis D—either of co-infections with acute hepatitis B or of superinfections in those already infected with HBV—may blur the distinctions between endemic and nonendemic areas. On a global scale, HDV infection declined at the end of the 1990s. Even in Italy, an HDV-endemic area, public health measures introduced to control HBV infection (e.g., mass hepatitis B vaccination) resulted during the 1990s in a 1.5%/year reduction in the prevalence of HDV infection. Still, the frequency of HDV infection during the first decade of the twenty-first century has not fallen below levels reached during

the 1990s; the reservoir has been sustained by survivors infected during 1970–1980 and recent immigrants from still-endemic (e.g., Eastern Europe and Central Asia) to less-endemic countries.

Hepatitis C

Routine screening of blood donors for HBsAg and the elimination of commercial blood sources in the early 1970s reduced the frequency of, but did not eliminate, transfusion-associated hepatitis. During the 1970s, the likelihood of acquiring hepatitis after transfusion of voluntarily donated, HBsAg-screened blood was ~10% per patient (up to 0.9% per unit transfused); 90–95% of these cases were classified, based on serologic exclusion of hepatitis A and B, as "non-A, non-B" hepatitis. For patients requiring transfusion of pooled products, such as clotting factor concentrates, the risk was even higher, up to 20–30%.

During the 1980s, voluntary self-exclusion of blood donors with risk factors for AIDS and then the introduction of donor screening for anti-HIV reduced further the likelihood of transfusion-associated hepatitis to <5%. During the late 1980s and early 1990s, the introduction first of "surrogate" screening tests for non-A, non-B hepatitis (alanine aminotransferase [ALT] and anti-HBc, both shown to identify blood donors with a higher likelihood of transmitting non-A, non-B hepatitis to recipients) and, subsequently, after the discovery of HCV, first-generation immunoassays for anti-HCV reduced the frequency of transfusion-associated hepatitis even further. A prospective analysis of transfusion-associated hepatitis conducted between 1986 and 1990 showed that the frequency of transfusion-associated hepatitis at one urban university hospital fell from a baseline of 3.8% per patient (0.45% per unit transfused) to 1.5% per patient (0.19% per unit) after the introduction of surrogate testing and to 0.6% per patient (0.03% per unit) after the introduction of first-generation anti-HCV assays. The introduction of second-generation anti-HCV assays reduced the frequency of transfusion-associated hepatitis C to almost imperceptible levels—1 in 100,000—and these gains were reinforced by the application of third-generation anti-HCV assays and of automated PCR testing of donated blood for HCV RNA, which has resulted in a reduction in the risk of transfusion-associated HCV infection to 1 in 2.3 million transfusions.

In addition to being transmitted by transfusion, hepatitis C can be transmitted by other percutaneous routes, such as injection drug use. In addition, this virus can be transmitted by occupational exposure to blood, and the likelihood of infection is increased in hemodialysis units. Although the frequency of transfusion-associated hepatitis C fell as a result of blood-donor screening, the overall frequency of hepatitis C remained the same until the early 1990s, when the overall frequency of reported cases fell by 80%, in parallel with a reduction in the number of new cases in injection drug users. After the exclusion of anti-HCV-positive plasma units from the donor pool, rare, sporadic instances have occurred of hepatitis C among recipients of immunoglobulin preparations for intravenous (but not intramuscular) use.

Serologic evidence for HCV infection occurs in 90% of patients with a history of transfusion-associated hepatitis (almost all occurring before 1992, when second-generation HCV screening tests were introduced); hemophiliacs and others treated with clotting factors; injection drug users; 60–70% of patients with sporadic "non-A, non-B" hepatitis who lack identifiable risk factors; 0.5% of volunteer blood donors; and, in the NHANES survey conducted in the United States between 1999 and 2002, 1.6% of the general population in the United States, which translates into 4.1 million persons (3.2 million with viremia), the majority of whom are unaware of their infections. Moreover, such population surveys do not include higher-risk groups such as incarcerated persons, homeless persons, and active injection drug users, indicating that the actual prevalence is even higher. Comparable frequencies of HCV infection occur in most countries around the world, with 170 million persons infected worldwide, but extraordinarily high prevalences of HCV infection occur in certain countries such as Egypt, where >20% of the population (as high as 50% in persons born prior to 1960) in some cities is infected. The high frequency in Egypt is attributable to contaminated equipment used for medical procedures and unsafe injection practices in the 1950s to 1980s (during a campaign to eradicate schistosomiasis with

intravenous tartar emetic). In the United States, African Americans and Mexican Americans have higher frequencies of HCV infection than whites. Data from NHANES showed that between 1988 and 1994, 30- to 40-year-old men had the highest prevalence of HCV infection; however, in a survey conducted between 1999 and 2002, the peak age decile had shifted to those age 40–49 years; an increase in hepatitis C–related mortality has paralleled this secular trend, increasing since 1995 predominantly in the 45- to 65-year age group. Thus, despite an 80% reduction in new HCV infections during the 1990s, the prevalence of HCV infection in the population was sustained by an aging cohort that had acquired their infections three to four decades earlier, during the 1960s and 1970s, as a result predominantly of self-inoculation with recreational drugs. Retrospective phylogenetic mapping of >45,000 HCV genotype 1a isolates revealed that the hepatitis C epidemic emerged in the United States between 1940 and 1965, peaking in 1950 and aligning temporally with the post-World-War-II expansion of medical procedures (including re-use of glass syringes). Thus, HCV was amplified iatrogenically not only in Egypt but also in the United States; in the United States, the seeds sewn by medical procedures in the 1950s were reaped in the 1960s and 1970s among transfusion recipients and injection drug users, even those whose drug use was confined to brief adolescent experimentation.

In NHANES 2003–2010, the prevalence of HCV infection (HCV RNA reactivity) in the United States had actually fallen to 1% (2.7 million persons) from 1.3% (3.2 million) the decade before (NHANES 1999–2002), attributable to deaths among the HCV-infected population. As death resulting from HIV infection fell after 1999, age-adjusted mortality associated with HCV infection surpassed that of HIV infection in 2007; >70% of HCV-associated deaths occurred in the “baby boomer” cohort born between 1945 and 1965. By 2012, HCV mortality had surpassed deaths from HIV, tuberculosis, hepatitis B, and 57 other notifiable infectious diseases (i.e., *all* infectious diseases) reported to the Centers for Disease Control and Prevention. In NHANES 1999–2002, compared to the 1.6% prevalence of HCV infection in the population at large, the prevalence in the 1945–1965 birth cohort was 3.2%, representing three-quarters of all infected persons. Therefore, in 2012, the Centers for Disease Control and Prevention recommended that all persons born between 1945 and 1965 be screened for hepatitis C, without ascertainment of risk, a recommendation shown to be cost-effective and predicted to identify 800,000 infected persons. Because of the availability of highly effective antiviral therapy, such screening would have the potential to avert 200,000 cases of cirrhosis and 47,000 cases of hepatocellular carcinoma and to prevent 120,000 hepatitis-related deaths; with the availability of the new generation of direct-acting antivirals (efficacy >95%, see Chap 334), screening baby boomers and treating those with hepatitis C have been predicted to reduce the HCV-associated disease burden by 50–70% through 2050.

Hepatitis C accounts for 40% of chronic liver disease, is the most frequent indication for liver transplantation, and is estimated to account for 8000–10,000 deaths per year in the United States. The distribution of HCV genotypes varies in different parts of the world. Worldwide, genotype 1 is the most common. In the United States, genotype 1 accounts for 70% of HCV infections, whereas genotypes 2 and 3 account for the remaining 30%; among African Americans, the frequency of genotype 1 is even higher (i.e., 90%). Genotype 4 predominates in Egypt; genotype 5 is localized to South Africa, genotype 6 to Hong Kong, and genotype 7 to Central Africa. Most asymptomatic blood donors found to have anti-HCV and ~20–30% of persons with reported cases of acute hepatitis C do not fall into a recognized risk group; however, many such blood donors do recall risk-associated behaviors when questioned carefully.

As a bloodborne infection, HCV potentially can be transmitted sexually and perinatally; however, both of these modes of transmission are inefficient for hepatitis C. Although 10–15% of patients with acute hepatitis C report having potential sexual sources of infection, most studies have failed to identify sexual transmission of this agent. The chances of sexual and perinatal transmission have been estimated to be ~5% but shown in a prospective study to be only 1% between monogamous sexual partners, well below comparable rates for HIV and HBV infections. Moreover, sexual transmission appears to be confined to such subgroups as persons with multiple sexual partners and sexually transmitted diseases; for example, isolated clusters of sexually transmitted HCV infection have been reported in HIV-infected men who have

sex with men. Breast-feeding does not increase the risk of HCV infection between an infected mother and her infant. Infection of health workers is not dramatically higher than among the general population; however, health workers are more likely to acquire HCV infection through accidental needle punctures, the efficiency of which is ~3%. Infection of household contacts is rare as well.

Besides persons born between 1945 and 1965, other groups with an increased frequency of HCV infection are listed in Table 332-4. In immunosuppressed individuals, levels of anti-HCV may be undetectable, and a diagnosis may require testing for HCV RNA. Although new acute cases of hepatitis C are rare outside of the injection-drug using community, newly diagnosed cases are common among otherwise healthy persons who experimented briefly with injection drugs, as noted above, three or four decades earlier. Such instances usually remain unrecognized for years, until unearthed by laboratory screening for routine medical examinations, insurance applications, and attempted blood donation. Although, overall, the annual incidence of new HCV infections has continued to fall, the rate of new infections has been increasing since 2002, amplified by the recent epidemic of opioid use, in a new cohort of young injection drug users, age 15–24 years (accounting for more than two-thirds of all acute cases), who, unlike older cohorts, had not learned to take precautions to prevent bloodborne infections.

TABLE 332-4

High-Risk Populations for Whom Hcv-Infection Screening is Recommended

Persons born between 1945 and 1965
Persons who have ever used injection drugs
Persons with HIV infection
Hemophiliacs treated with clotting factor concentrates prior to 1987
Persons who have ever undergone long-term hemodialysis
Persons with unexplained elevations of aminotransferase levels
Transfusion or transplantation recipients prior to July 1992
Recipients of blood or organs from a donor found to be positive for hepatitis C
Children born to women with hepatitis C
Health care, public safety, and emergency medical personnel following needle injury or mucosal exposure to HCV-contaminated blood
Sexual partners of persons with hepatitis C infection

Hepatitis E

This type of hepatitis, identified in India, Asia, Africa, the Middle East, and Central America, resembles hepatitis A in its primarily enteric mode of spread. The commonly recognized cases occur after contamination of water supplies such as after monsoon flooding, but sporadic, isolated cases occur. An epidemiologic feature that distinguishes HEV from other enteric agents is the rarity of secondary person-to-person spread from infected persons to their close contacts. Large waterborne outbreaks in endemic areas are linked to genotypes 1 and 2, arise in populations that are immune to HAV, favor young adults, and account for antibody prevalences of 30–80%. In nonendemic areas of the world, such as the United States, clinically apparent acute hepatitis E is extremely rare; however, during the 1988–1994 NHANES survey conducted by the U.S. Public Health Service, the prevalence of anti-HEV was 21%, reflecting subclinical infections, infection with genotypes 3 and 4, predominantly in older males (>60 years). In nonendemic areas, HEV accounts hardly

at all for cases of sporadic (labeled “autochthonous” or indigenous) hepatitis; however, cases imported from endemic areas have been found in the United States. Evidence supports a zoonotic reservoir for HEV primarily in swine, which may account for the mostly subclinical infections in nonendemic areas. A previously unrecognized high distribution of HEV infection, linked to pork-product ingestion, has been discovered in western Europe (e.g., in Germany, an estimated annual incidence of 300,000 cases and a 17% prevalence of anti-HEV among adults; in France, a 22% prevalence of anti-HEV in healthy blood donors).

CLINICAL AND LABORATORY FEATURES

Symptoms and Signs

Acute viral hepatitis occurs after an incubation period that varies according to the responsible agent. Generally, incubation periods for hepatitis A range from 15 to 45 days (mean, 4 weeks), for hepatitis B and D from 30 to 180 days (mean, 8–12 weeks), for hepatitis C from 15 to 160 days (mean, 7 weeks), and for hepatitis E from 14 to 60 days (mean, 5–6 weeks). The *prodromal symptoms* of acute viral hepatitis are systemic and quite variable. Constitutional symptoms of anorexia, nausea and vomiting, fatigue, malaise, arthralgias, myalgias, headache, photophobia, pharyngitis, cough, and coryza may precede the onset of jaundice by 1–2 weeks. The nausea, vomiting, and anorexia are frequently associated with alterations in olfaction and taste. A low-grade fever between 38° and 39°C (100°–102°F) is more often present in hepatitis A and E than in hepatitis B or C, except when hepatitis B is heralded by a serum sickness–like syndrome; rarely, a fever of 39.5°–40°C (103°–104°F) may accompany the constitutional symptoms. Dark urine and clay-colored stools may be noticed by the patient from 1–5 days before the onset of clinical jaundice.

With the onset of *clinical jaundice*, the constitutional prodromal symptoms usually diminish, but in some patients, mild weight loss (2.5–5 kg) is common and may continue during the entire icteric phase. The liver becomes enlarged and tender and may be associated with right upper quadrant pain and discomfort. Infrequently, patients present with a cholestatic picture, suggesting extrahepatic biliary obstruction. Splenomegaly and cervical adenopathy are present in 10–20% of patients with acute hepatitis. Rarely, a few spider angiomas appear during the icteric phase and disappear during convalescence. During the *recovery phase*, constitutional symptoms disappear, but usually some liver enlargement and abnormalities in liver biochemical tests are still evident. The duration of the posticteric phase is variable, ranging from 2 to 12 weeks, and is usually more prolonged in acute hepatitis B and C. Complete clinical and biochemical recovery is to be expected 1–2 months after all cases of hepatitis A and E and 3–4 months after the onset of jaundice in three-quarters of uncomplicated, self-limited cases of hepatitis B and C (among healthy adults, acute hepatitis B is self-limited in 95–99%, whereas hepatitis C is self-limited in only ~15–20%). In the remainder, biochemical recovery may be delayed. A substantial proportion of patients with viral hepatitis never become icteric.

Infection with HDV can occur in the presence of acute or chronic HBV infection; the duration of HBV infection determines the duration of HDV infection. When acute HDV and HBV infections occur simultaneously, clinical and biochemical features may be indistinguishable from those of HBV infection alone, although occasionally they are more severe. As opposed to patients with *acute* HBV infection, patients with *chronic* HBV infection can support HDV replication indefinitely, as when acute HDV infection occurs in the presence of a nonresolving acute HBV infection or, more commonly, when acute hepatitis D is superimposed on underlying chronic hepatitis B. In such cases, the HDV superinfection appears as a clinical exacerbation or an episode resembling acute viral hepatitis in someone already chronically infected with HBV. Superinfection with HDV in a patient with chronic hepatitis B often leads to clinical deterioration (see below).

In addition to superinfections with other hepatitis agents, acute hepatitis-like clinical events in persons with chronic hepatitis B may accompany spontaneous HBeAg to anti-HBe seroconversion or spontaneous reactivation (i.e., reversion from relatively nonreplicative to replicative infection). Such reactivations can occur as well in therapeutically immunosuppressed patients with chronic HBV infection when cytotoxic/immunosuppressive drugs are withdrawn; in these cases, restoration of immune competence is thought to allow resumption of previously checked cell-mediated immune cytolysis of HBV-infected hepatocytes. Occasionally, acute clinical exacerbations of chronic hepatitis B may represent the emergence of a precore mutant (see "Virology and Etiology"), and the subsequent course in such patients may be characterized by periodic exacerbations. Cytotoxic chemotherapy can lead to reactivation of chronic hepatitis C as well, and anti-TNF-α therapy can lead to reactivation of both hepatitis B and C.

Laboratory Features

The serum aminotransferases aspartate aminotransferase (AST) and ALT (previously designated SGOT and SGPT) increase to a variable degree during the prodromal phase of acute viral hepatitis and precede the rise in bilirubin level (Figs. 332-2 and 332-4). The level of these enzymes, however, does not correlate well with the degree of liver cell damage. Peak levels vary from ~400 to ~4000 IU or more; these levels are usually reached at the time the patient is clinically icteric and diminish progressively during the recovery phase of acute hepatitis. The diagnosis of anicteric hepatitis is based on clinical features and on aminotransferase elevations.

Jaundice is usually visible in the sclera or skin when the serum bilirubin value is >43 µmol/L (2.5 mg/dL). When jaundice appears, the serum bilirubin typically rises to levels ranging from 85 to 340 µmol/L (5–20 mg/dL). The serum bilirubin may continue to rise despite falling serum aminotransferase levels. In most instances, the total bilirubin is equally divided between the conjugated and unconjugated fractions. Bilirubin levels >340 µmol/L (20 mg/dL) extending and persisting late into the course of viral hepatitis are more likely to be associated with severe disease. In certain patients with underlying hemolytic anemia, however, such as glucose-6-phosphate dehydrogenase deficiency and sickle cell anemia, a high serum bilirubin level is common, resulting from superimposed hemolysis. In such patients, bilirubin levels >513 µmol/L (30 mg/dL) have been observed and are not necessarily associated with a poor prognosis.

Neutropenia and lymphopenia are transient and are followed by a relative lymphocytosis. Atypical lymphocytes (varying between 2 and 20%) are common during the acute phase. Measurement of the prothrombin time (PT) is important in patients with acute viral hepatitis, because a prolonged value may reflect a severe hepatic synthetic defect, signify extensive hepatocellular necrosis, and indicate a worse prognosis. Occasionally, a prolonged PT may occur with only mild increases in the serum bilirubin and aminotransferase levels. Prolonged nausea and vomiting, inadequate carbohydrate intake, and poor hepatic glycogen reserves may contribute to hypoglycemia noted occasionally in patients with severe viral hepatitis. Serum alkaline phosphatase may be normal or only mildly elevated, whereas a fall in serum albumin is uncommon in uncomplicated acute viral hepatitis. In some patients, mild and transient steatorrhea has been noted, as well as slight microscopic hematuria and minimal proteinuria.

A diffuse but mild elevation of the γ globulin fraction is common during acute viral hepatitis. Serum IgG and IgM levels are elevated in about one-third of patients during the acute phase of viral hepatitis, but the serum IgM level is elevated more characteristically during acute hepatitis A. During the acute phase of viral hepatitis, antibodies to smooth muscle and other cell constituents may be present, and low titers of rheumatoid factor, nuclear antibody, and heterophile antibody can also be found occasionally. In hepatitis C and D, antibodies to LKM may occur; however, the species of LKM antibodies in the two types of hepatitis are different from each other as well as from the LKM antibody species characteristic of autoimmune hepatitis type 2 **(Chap. 334)**. The autoantibodies in viral hepatitis are nonspecific and can

also be associated with other viral and systemic diseases. In contrast, virus-specific antibodies, which appear during and after hepatitis virus infection, are serologic markers of diagnostic importance.

As described above, serologic tests are available routinely with which to establish a diagnosis of hepatitis A, B, D, and C. Tests for fecal or serum HAV are not routinely available. Therefore, a diagnosis of hepatitis A is based on detection of IgM anti-HAV during acute illness (Fig. 332-2). Rheumatoid factor can give rise to false-positive results in this test.

A diagnosis of HBV infection can usually be made by detection of HBsAg in serum. Infrequently, levels of HBsAg are too low to be detected during acute HBV infection, even with contemporary, highly sensitive immunoassays. In such cases, the diagnosis can be established by the presence of IgM anti-HBc.

The titer of HBsAg bears little relation to the severity of clinical disease. Indeed, an inverse correlation exists between the serum concentration of HBsAg and the degree of liver cell damage. For example, titers are highest in immunosuppressed patients, lower in patients with chronic liver disease (but higher in mild chronic than in severe chronic hepatitis), and very low in patients with acute fulminant hepatitis. These observations suggest that in hepatitis B the degree of liver cell damage and the clinical course are related to variations in the patient's immune response to HBV rather than to the amount of circulating HBsAg. In immunocompetent persons, however, a correlation exists between markers of HBV *replication* and liver injury (see below).

Another important serologic marker in patients with hepatitis B is HBeAg. Its principal clinical usefulness is as an indicator of relative infectivity. Because HBeAg is invariably present during early acute hepatitis B, HBeAg testing is indicated primarily in chronic infection.

In patients with hepatitis B surface antigenemia of unknown duration (e.g., blood donors found to be HBsAg-positive) testing for IgM anti-HBc may be useful to distinguish between acute or recent infection (IgM anti-HBc-positive) and chronic HBV infection (IgM anti-HBc-negative, IgG anti-HBc-positive). A false-positive test for IgM anti-HBc may be encountered in patients with high-titer rheumatoid factor. Also, IgM anti-HBc may be reexpressed during acute reactivation of chronic hepatitis B.

Anti-HBs is rarely detectable in the presence of HBsAg in patients with *acute* hepatitis B, but 10–20% of persons with *chronic* HBV infection may harbor low-level anti-HBs. This antibody is directed not against the common group determinant, *a*, but against the heterotypic subtype determinant (e.g., HBsAg of subtype *ad* with anti-HBs of subtype *y*). In most cases, this serologic pattern cannot be attributed to infection with two different HBV subtypes but, instead, is thought (based on the clonal selection theory of antibody diversity) to reflect the stimulation of a related clone of antibody-forming cells and is not a harbinger of imminent HBsAg clearance. When such antibody is detected, its presence is of no recognized clinical significance (see "Virology and Etiology").

After immunization with hepatitis B vaccine, which consists of HBsAg alone, anti-HBs is the only serologic marker to appear. The commonly encountered serologic patterns of hepatitis B and their interpretations are summarized in **Table 332-5**. Tests for the detection of HBV DNA in liver and serum are now available. Like HBeAg, serum HBV DNA is an indicator of HBV replication, but tests for HBV DNA are more sensitive and quantitative. First-generation hybridization assays for HBV DNA had a sensitivity of 10^5–10^6 virions/mL, a relative threshold below which infectivity and liver injury are limited and HBeAg is usually undetectable. Currently, testing for HBV DNA has shifted from insensitive hybridization assays to amplification assays (e.g., the PCR-based assay, which can detect as few as 10 or 100 virions/mL); among the commercially available PCR assays, the most useful are those with the highest sensitivity (5–10 IU/mL) and the largest dynamic range (10^0–10^9 IU/mL). With increased sensitivity, amplification assays remain reactive well below the current

10^3–10^4 IU/mL threshold for infectivity and liver injury. These markers are useful in following the course of HBV replication in patients with chronic hepatitis B receiving antiviral chemotherapy (**Chap. 334**). Except for the early decades of life after perinatally acquired HBV infection (see above), in immunocompetent adults with chronic hepatitis B, a general correlation exists between the level of HBV replication, as reflected by the level of serum HBV DNA, and the degree of liver injury. High-serum HBV DNA levels, increased expression of viral antigens, and necroinflammatory activity in the liver go hand in hand unless immunosuppression interferes with cytolytic T cell responses to virus-infected cells; reduction of HBV replication with antiviral drugs tends to be accompanied by an improvement in liver histology. Among patients with chronic hepatitis B, high levels of HBV DNA increase the risk of cirrhosis, hepatic decompensation, and hepatocellular carcinoma (see "Complications and Sequelae").

TABLE 332-5

Commonly Encountered Serologic Patterns of Hepatitis B Infection

HBsAg	ANTI-HBs	ANTI-HBc	HBeAg	ANTI-HBe	INTERPRETATION
+	−	IgM	+	−	Acute hepatitis B, high infectivity[a]
+	−	IgG	+	−	Chronic hepatitis B, high infectivity
+	−	IgG	−	+	1. Late acute or chronic hepatitis B, low infectivity 2. HBeAg-negative ("precore-mutant") hepatitis B (chronic or, rarely, acute)
+	+	+	+/−	+/−	1. HBsAg of one subtype and heterotypic anti-HBs (common) 2. Process of seroconversion from HBsAg to anti-HBs (rare)
−	−	IgM	+/−	+/−	1. Acute hepatitis B[a] 2. Anti-HBc "window"
−	−	IgG	−	+/−	1. Low-level hepatitis B carrier 2. Hepatitis B in remote past
−	+	IgG	−	+/−	Recovery from hepatitis B
−	+	−	−	−	1. Immunization with HBsAg (after vaccination) 2. Hepatitis B in the remote past (?) 3. False-positive

[a]IgM anti-HBc may reappear during acute reactivation of chronic hepatitis B.

Note: See text for abbreviations.

In patients with hepatitis C, an episodic pattern of aminotransferase elevation is common. A specific serologic diagnosis of hepatitis C can be made by demonstrating the presence in serum of anti-HCV. When contemporary immunoassays are

used, anti-HCV can be detected in acute hepatitis C during the initial phase of elevated aminotransferase activity and remains detectable after recovery (rare) and during chronic infection (common). Nonspecificity can confound immunoassays for anti-HCV, especially in persons with a low prior probability of infection, such as volunteer blood donors, or in persons with circulating rheumatoid factor, which can bind nonspecifically to assay reagents; testing for HCV RNA can be used in such settings to distinguish between true-positive and false-positive anti-HCV determinations. Assays for HCV RNA are the most sensitive tests for HCV infection and represent the “gold standard” in establishing a diagnosis of hepatitis C. HCV RNA can be detected even before acute elevation of aminotransferase activity and before the appearance of anti-HCV in patients with acute hepatitis C. In addition, HCV RNA remains detectable indefinitely, continuously in most but intermittently in some, in patients with chronic hepatitis C (detectable as well in some persons with normal liver tests, i.e., inactive carriers). In the very small minority of patients with hepatitis C who lack anti-HCV, a diagnosis can be supported by detection of HCV RNA. If all these tests are negative and the patient has a well-characterized case of hepatitis after percutaneous exposure to blood or blood products, a diagnosis of hepatitis caused by an unidentified agent can be entertained.

Amplification techniques are required to detect HCV RNA. Currently, such target amplification (i.e., synthesis of multiple copies of the viral genome) is achieved by PCR, in which the viral RNA is reverse transcribed to complementary DNA and then amplified by repeated cycles of DNA synthesis. Quantitative PCR assays provide a measurement of relative “viral load”; current PCR assays have a sensitivity of 10 (lower limit of detection)-25 (lower limit of quantitation) IU/mL and a wide dynamic range (10–10^7 IU/mL). Determination of HCV RNA level is not a reliable marker of disease severity or prognosis but is helpful in predicting relative responsiveness to antiviral therapy. The same is true for determinations of HCV genotype (**Chap. 334**). Of course, HCV RNA monitoring during and after antiviral therapy is the *sine qua non* for determining on-treatment and durable responsiveness.

A proportion of patients with hepatitis C have isolated anti-HBc in their blood, a reflection of a common risk in certain populations of exposure to multiple bloodborne hepatitis agents. The anti-HBc in such cases is almost invariably of the IgG class and usually represents HBV infection in the remote past (HBV DNA undetectable); it rarely represents current HBV infection with low-level virus carriage. Detectable anti-HCV in the absence of HCV RNA signifies spontaneous or therapeutically induced recovery from (“cured”) hepatitis C.

The presence of HDV infection can be identified by demonstrating intrahepatic HDV antigen or, more practically, an anti-HDV seroconversion (a rise in titer of anti-HDV or de novo appearance of anti-HDV). Circulating HDV antigen, also diagnostic of acute infection, is detectable only briefly, if at all. Because anti-HDV is often undetectable once HBsAg disappears, retrospective serodiagnosis of acute self-limited, simultaneous HBV and HDV infection is difficult. Early diagnosis of acute infection may be hampered by a delay of up to 30–40 days in the appearance of anti-HDV.

When a patient presents with acute hepatitis and has HBsAg and anti-HDV in serum, determination of the class of anti-HBc is helpful in establishing the relationship between infection with HBV and HDV. Although IgM anti-HBc does not distinguish *absolutely* between acute and chronic HBV infection, its presence is a reliable indicator of recent infection and its absence a reliable indicator of infection in the remote past. In simultaneous acute HBV and HDV infections, IgM anti-HBc will be detectable, whereas in acute HDV infection superimposed on chronic HBV infection, anti-HBc will be of the IgG class. Assays for HDV RNA, available in specialized laboratories and yet to be standardized, can be used to confirm HDV infection and to monitor treatment during chronic infection.

The serologic/virologic course of events during acute hepatitis E is entirely analogous to that of acute hepatitis A, with brief fecal shedding of virus and viremia and an early IgM anti-HEV response that predominates during approximately the first 3 months but is eclipsed thereafter by long-lasting IgG anti-HEV. Diagnostic tests of varying reliability for

hepatitis E are commercially available but used routinely primarily outside the United States; in the United States, diagnostic serologic/virologic assays can be performed at the Centers for Disease Control and Prevention or other specialized reference laboratories.

Liver biopsy is rarely necessary or indicated in acute viral hepatitis, except when the diagnosis is questionable or when clinical evidence suggests a diagnosis of chronic hepatitis.

A diagnostic algorithm can be applied in the evaluation of cases of acute viral hepatitis. A patient with acute hepatitis should undergo four serologic tests: HBsAg, IgM anti-HAV, IgM anti-HBc, and anti-HCV (**Table 332-6**). The presence of HBsAg, with or without IgM anti-HBc, represents HBV infection. If IgM anti-HBc is present, the HBV infection is considered acute; if IgM anti-HBc is absent, the HBV infection is considered chronic. A diagnosis of acute hepatitis B can be made in the absence of HBsAg when IgM anti-HBc is detectable. A diagnosis of acute hepatitis A is based on the presence of IgM anti-HAV. If IgM anti-HAV coexists with HBsAg, a diagnosis of simultaneous HAV and HBV infections can be made; if IgM anti-HBc (with or without HBsAg) is detectable, the patient has simultaneous acute hepatitis A and B, and if IgM anti-HBc is undetectable, the patient has acute hepatitis A superimposed on chronic HBV infection. The presence of anti-HCV supports a diagnosis of acute hepatitis C. Occasionally, testing for HCV RNA or repeat anti-HCV testing later during the illness is necessary to establish the diagnosis. Absence of all serologic markers is consistent with a diagnosis of "non-A, non-B, non-C" hepatitis (no other proven human hepatitis viruses have been identified), if the epidemiologic setting is appropriate.

TABLE 332-6

Simplified Diagnostic Approach in Patients Presenting with Acute Hepatitis

SEROLOGIC TESTS OF PATIENT'S SERUM				
HBsAg	**IgM ANTI-HAV**	**IgM ANTI-HBc**	**ANTI-HCV**	**DIAGNOSTIC INTERPRETATION**
+	−	+	−	Acute hepatitis B
+	−	−	−	Chronic hepatitis B
+	+	−	−	Acute hepatitis A superimposed on chronic hepatitis B
+	+	+	−	Acute hepatitis A and B
−	+	−	−	Acute hepatitis A
−	+	+	−	Acute hepatitis A and B (HBsAg below detection threshold)
−	−	+	−	Acute hepatitis B (HBsAg below detection threshold)
−	−	−	+	Acute hepatitis C

Note: See text for abbreviations.

In patients with chronic hepatitis, initial testing should consist of HBsAg and anti-HCV. Anti-HCV supports and HCV RNA testing establishes the diagnosis of chronic hepatitis C. If a serologic diagnosis of chronic hepatitis B is made, testing for HBeAg and anti-HBe is indicated to evaluate relative infectivity. Testing for HBV DNA in such patients provides a more quantitative and sensitive measure of the level of virus replication, and therefore is very helpful during antiviral therapy (Chap. 334). In patients with chronic hepatitis B and normal aminotransferase activity in the absence of HBeAg, serial testing over time is often required to distinguish between inactive carriage and HBeAg-negative chronic hepatitis B with fluctuating virologic and necroinflammatory activity. In persons with hepatitis B, testing for anti-HDV is useful in those with severe and fulminant disease, with severe chronic disease, with chronic hepatitis B and acute hepatitis-like exacerbations, with frequent percutaneous exposures, and from areas where HDV infection is endemic.

PROGNOSIS

Virtually all previously healthy patients with hepatitis A recover completely with no clinical sequelae. Similarly, in acute hepatitis B, 95–99% of previously healthy adults have a favorable course and recover completely. Certain clinical and laboratory features, however, suggest a more complicated and protracted course. Patients of advanced age and with serious underlying medical disorders may have a prolonged course and are more likely to experience severe hepatitis. Initial presenting features such as ascites, peripheral edema, and symptoms of hepatic encephalopathy suggest a poorer prognosis. In addition, a prolonged PT, low serum albumin level, hypoglycemia, and very high serum bilirubin values suggest severe hepatocellular disease. Patients with these clinical and laboratory features deserve prompt hospital admission. The case fatality rate in hepatitis A and B is very low (~0.1%) but is increased by advanced age and underlying debilitating disorders. Among patients ill enough to be hospitalized for acute hepatitis B, the fatality rate is 1%. Hepatitis C is less severe during the acute phase than hepatitis B and is more likely to be anicteric; fatalities are rare, but the precise case fatality rate is not known. In outbreaks of waterborne hepatitis E in India and Asia, the case fatality rate is 1–2% and up to 10–20% in pregnant women. Contributing to fulminant hepatitis E in endemic countries (but only very rarely or not at all in nonendemic countries) are instances of acute hepatitis E superimposed on underlying chronic liver disease (“acute-on-chronic” liver disease). Patients with simultaneous acute hepatitis B and D do not necessarily experience a higher mortality rate than do patients with acute hepatitis B alone; however, in several outbreaks of acute simultaneous HBV and HDV infection among injection drug users, the case fatality rate was ~5%. When HDV superinfection occurs in a person with chronic hepatitis B, the likelihood of fulminant hepatitis and death is increased substantially. Although the case fatality rate for hepatitis D is not known definitively, in outbreaks of severe HDV superinfection in isolated populations with a high hepatitis B carrier rate, a mortality rate >20% has been recorded.

COMPLICATIONS AND SEQUELAE

A small proportion of patients with hepatitis A experience *relapsing hepatitis* weeks to months after apparent recovery from acute hepatitis. Relapses are characterized by recurrence of symptoms, aminotransferase elevations, occasional jaundice, and fecal excretion of HAV. Another unusual variant of acute hepatitis A is *cholestatic hepatitis*, characterized by protracted cholestatic jaundice and pruritus. Rarely, liver test abnormalities persist for many months, even up to 1 year. Even when these complications occur, hepatitis A remains self-limited and does not progress to chronic liver disease. During the prodromal phase of acute hepatitis B, a serum sickness–like syndrome characterized by arthralgia or arthritis, rash, angioedema, and, rarely, hematuria and proteinuria may develop in 5–10% of patients. This syndrome occurs before the onset of clinical jaundice, and these patients are often diagnosed erroneously as having rheumatologic diseases. The diagnosis can be established by measuring serum aminotransferase levels, which are almost invariably elevated, and serum HBsAg. As noted above, EMC is an immune-complex disease that can complicate chronic hepatitis C and is part of a spectrum of B cell lymphoproliferative disorders, which, in rare instances, can evolve to B cell lymphoma

(Chap. 104). Attention has been drawn as well to associations between hepatitis C and such cutaneous disorders as porphyria cutanea tarda and lichen planus. A mechanism for these associations is unknown. Finally, related to the reliance of HCV on lipoprotein secretion and assembly pathways and on interactions of HCV with glucose metabolism, HCV infection may be complicated by hepatic steatosis, hypercholesterolemia, insulin resistance (and other manifestations of the metabolic syndrome), and type 2 diabetes mellitus; both hepatic steatosis and insulin resistance appear to accelerate hepatic fibrosis and blunt responsiveness to interferon-based antiviral therapy (Chap. 334).

The most feared complication of viral hepatitis is *fulminant hepatitis* (massive hepatic necrosis); fortunately, this is a rare event. Fulminant hepatitis is seen primarily in hepatitis B, D, and E, but rare fulminant cases of hepatitis A occur primarily in older adults and in persons with underlying chronic liver disease, including, according to some reports, chronic hepatitis B and C. Hepatitis B accounts for >50% of fulminant cases of viral hepatitis, a sizable proportion of which are associated with HDV infection and another proportion with underlying chronic hepatitis C. Fulminant hepatitis is hardly ever seen in hepatitis C, but hepatitis E, as noted above, can be complicated by fatal fulminant hepatitis in 1–2% of all cases and in up to 20% of cases in pregnant women. Patients usually present with signs and symptoms of encephalopathy that may evolve to deep coma. The liver is usually small and the PT excessively prolonged. The combination of rapidly shrinking liver size, rapidly rising bilirubin level, and marked prolongation of the PT, even as aminotransferase levels fall, together with clinical signs of confusion, disorientation, somnolence, ascites, and edema, indicates that the patient has hepatic failure with encephalopathy. Cerebral edema is common; brainstem compression, gastrointestinal bleeding, sepsis, respiratory failure, cardiovascular collapse, and renal failure are terminal events. The mortality rate is exceedingly high (>80% in patients with deep coma), but patients who survive may have a complete biochemical and histologic recovery. If a donor liver can be located in time, liver transplantation may be lifesaving in patients with fulminant hepatitis (Chap. 338).

Documenting the disappearance of HBsAg after apparent clinical recovery from acute hepatitis B is particularly important. Before laboratory methods were available to distinguish between acute hepatitis and acute hepatitis-like exacerbations (*spontaneous reactivations*) of chronic hepatitis B, observations suggested that ~10% of previously healthy patients remained HBsAg-positive for >6 months after the onset of clinically apparent acute hepatitis B. One-half of these persons cleared the antigen from their circulations during the next several years, but the other 5% remained chronically HBsAg-positive. More recent observations suggest that the true rate of chronic infection after clinically apparent acute hepatitis B is as low as 1% in normal, immunocompetent, young adults. Earlier, higher estimates may have been confounded by inadvertent inclusion of acute exacerbations in chronically infected patients; these patients, chronically HBsAg-positive before exacerbation, were unlikely to seroconvert to HBsAg-negative thereafter. Whether the rate of chronicity is 10% or 1%, such patients have IgG anti-HBc in serum; anti-HBs is either undetected or detected at low titer against the opposite subtype specificity of the antigen (see “Laboratory Features”). These patients may (1) be inactive carriers; (2) have low-grade, mild chronic hepatitis; or (3) have moderate to severe chronic hepatitis with or without cirrhosis. The likelihood of remaining chronically infected after acute HBV infection is especially high among neonates, persons with Down’s syndrome, chronically hemodialyzed patients, and immunosuppressed patients, including persons with HIV infection.

Chronic hepatitis is an important late complication of acute hepatitis B occurring in a small proportion of patients with acute disease but more common in those who present with chronic infection without having experienced an acute illness, as occurs typically after neonatal infection or after infection in an immunosuppressed host (Chap. 334). The following clinical and laboratory features suggest progression of acute hepatitis to chronic hepatitis: (1) lack of complete resolution of clinical symptoms of anorexia, weight loss, fatigue, and the persistence of hepatomegaly; (2) the presence of bridging/interface or multilobular hepatic necrosis on liver biopsy during protracted, severe acute viral hepatitis; (3)

failure of the serum aminotransferase, bilirubin, and globulin levels to return to normal within 6–12 months after the acute illness; and (4) the persistence of HBeAg for >3 months or HBsAg for >6 months after acute hepatitis.

Although acute hepatitis D infection does not increase the likelihood of chronicity of simultaneous acute hepatitis B, hepatitis D has the potential for contributing to the severity of chronic hepatitis B. Hepatitis D superinfection can transform inactive or mild chronic hepatitis B into severe, progressive chronic hepatitis and cirrhosis; it also can accelerate the course of chronic hepatitis B. Some HDV superinfections in patients with chronic hepatitis B lead to fulminant hepatitis. As defined in longitudinal studies over three decades, the annual rate of cirrhosis in patients with chronic hepatitis D is 4%. Although HDV and HBV infections are associated with severe liver disease, mild hepatitis and even inactive carriage have been identified in some patients, and the disease may become indolent beyond the early years of infection.

After acute HCV infection, the likelihood of remaining chronically *infected* approaches 85–90%. Although many patients with chronic hepatitis C have no symptoms, cirrhosis may develop in as many as 20% within 10–20 years of acute illness; in some series of cases reported by referral centers, cirrhosis has been reported in as many as 50% of patients with chronic hepatitis C. Among cirrhotic patients with chronic hepatitis C, the annual risk of hepatic decompensation is ~4%. Although chronic hepatitis C accounts for at least 40% of cases of chronic liver disease and of patients undergoing liver transplantation for end-stage liver disease in the United States and Europe, in the majority of patients with chronic hepatitis C, morbidity and mortality are limited during the initial 20 years after the onset of infection. Progression of chronic hepatitis C may be influenced by advanced age of acquisition, long duration of infection, immunosuppression, coexisting excessive alcohol use, concomitant hepatic steatosis, other hepatitis virus infection, or HIV co-infection. In fact, instances of severe and rapidly progressive chronic hepatitis B and C are being recognized with increasing frequency in patients with HIV infection **(Chap. 197)**. In contrast, neither HAV nor HEV causes chronic liver disease in immunocompetent hosts; however, cases of chronic hepatitis E (including cirrhosis and end-stage liver disease) have been observed in immunosuppressed organ-transplant recipients, persons receiving cytotoxic chemotherapy, and persons with HIV infection. Among patients with chronic hepatitis (e.g., caused by hepatitis B or C, alcohol, etc.) in endemic countries, hepatitis E has been reported as the cause of acute-on-chronic liver failure; however, in most experiences among patients from nonendemic countries, HEV has not been found to contribute to hepatic decompensation in patients with chronic hepatitis.

Persons with chronic hepatitis B, particularly those infected in infancy or early childhood and especially those with HBeAg and/or high-level HBV DNA, have an enhanced risk of hepatocellular carcinoma. The risks of cirrhosis and hepatocellular carcinoma increase with the level of HBV replication. The annual rate of hepatocellular carcinoma in patients with chronic hepatitis D and cirrhosis is ~3%. The risk of hepatocellular carcinoma is increased as well in patients with chronic hepatitis C, almost exclusively in patients with cirrhosis, and almost always after at least several decades, usually after three decades of disease **(Chap. 78)**. Among such cirrhotic patients with chronic hepatitis C, the annual risk of hepatocellular carcinoma is ~1–4%.

Rare complications of viral hepatitis include pancreatitis, myocarditis, atypical pneumonia, aplastic anemia, transverse myelitis, and peripheral neuropathy. In children, hepatitis B may present rarely with anicteric hepatitis, a nonpruritic papular rash of the face, buttocks, and limbs, and lymphadenopathy (papular acrodermatitis of childhood or Gianotti-Crosti syndrome).

Rarely, autoimmune hepatitis **(Chap. 334)** can be triggered by a bout of otherwise self-limited acute hepatitis, as reported after acute hepatitis A, B, and C.

DIFFERENTIAL DIAGNOSIS

Viral diseases such as infectious mononucleosis; those due to cytomegalovirus, herpes simplex, and coxsackieviruses; and toxoplasmosis may share certain clinical features with viral hepatitis and cause elevations in serum aminotransferase and, less commonly, in serum bilirubin levels. Tests such as the differential heterophile and serologic tests for these agents may be helpful in the differential diagnosis if HBsAg, anti-HBc, IgM anti-HAV, and anti-HCV determinations are negative. Aminotransferase elevations can accompany almost any systemic viral infection; other rare causes of liver injury confused with viral hepatitis are infections with *Leptospira*, *Candida*, *Brucella*, *Mycobacteria*, and *Pneumocystis*. A complete drug history is particularly important because many drugs and certain anesthetic agents can produce a picture of either acute hepatitis or cholestasis (Chap. 333). Equally important is a past history of unexplained "repeated episodes" of acute hepatitis. This history should alert the physician to the possibility that the underlying disorder is chronic hepatitis, for example autoimmune hepatitis (Chap. 334). Alcoholic hepatitis must also be considered, but usually the serum aminotransferase levels are not as markedly elevated, and other stigmata of alcoholism may be present. The finding on liver biopsy of fatty infiltration, a neutrophilic inflammatory reaction, and "alcoholic hyaline" would be consistent with alcohol-induced rather than viral liver injury. Because acute hepatitis may present with right upper quadrant abdominal pain, nausea and vomiting, fever, and icterus, it is often confused with acute cholecystitis, common duct stone, or ascending cholangitis. Patients with acute viral hepatitis may tolerate surgery poorly; therefore, it is important to exclude this diagnosis, and in confusing cases, a percutaneous liver biopsy may be necessary before laparotomy. Viral hepatitis in the elderly is often misdiagnosed as obstructive jaundice resulting from a common duct stone or carcinoma of the pancreas. Because acute hepatitis in the elderly may be quite severe and the operative mortality high, a thorough evaluation including biochemical tests, radiographic studies of the biliary tree, and even liver biopsy may be necessary to exclude primary parenchymal liver disease. Another clinical constellation that may mimic acute hepatitis is right ventricular failure with passive hepatic congestion or hypoperfusion syndromes, such as those associated with shock, severe hypotension, and severe left ventricular failure. Also included in this general category is any disorder that interferes with venous return to the heart, such as right atrial myxoma, constrictive pericarditis, hepatic vein occlusion (Budd-Chiari syndrome), or venoocclusive disease. Clinical features are usually sufficient to distinguish among these vascular disorders and viral hepatitis. Acute fatty liver of pregnancy, cholestasis of pregnancy, eclampsia, and the HELLP (*h*emolysis, *e*levated *l*iver tests, and *l*ow *p*latelets) syndrome can be confused with viral hepatitis during pregnancy. Very rarely, malignancies metastatic to the liver can mimic acute or even fulminant viral hepatitis. Occasionally, genetic or metabolic liver disorders (e.g., Wilson's disease, α_1 antitrypsin deficiency) and nonalcoholic fatty liver disease are confused with acute viral hepatitis.

TREATMENT

TREATMENT: Acute Viral Hepatitis

Most persons with acute hepatitis (especially hepatitis A, B, and E) recover spontaneously and do not require specific antiviral therapy. In hepatitis B, among previously healthy adults who present with clinically apparent acute hepatitis, recovery occurs in ~99%; therefore, antiviral therapy is not likely to improve the rate of recovery and is not required. In rare instances of severe acute hepatitis B, treatment with a nucleoside analogue at oral doses used to treat chronic hepatitis B (Chap. 334) has been attempted successfully. Although clinical trials have not been done to establish the efficacy or duration of this approach, most authorities would recommend institution of antiviral therapy with a nucleoside analogue (entecavir or tenofovir, the most potent and least resistance-prone agents) for severe, but not mild–moderate, acute hepatitis B. Treatment should continue until 3 months after HBsAg seroconversion or 6 months after HBeAg seroconversion.

In typical cases of acute hepatitis C, recovery is rare (~15–20% in most experiences), progression to chronic hepatitis is the rule, and small clinical trials during the era of interferon-based regimens suggested that antiviral therapy with courses (usually 24 weeks) of standard or pegylated interferon α monotherapy reduced the rate of chronicity considerably by inducing sustained responses in 30–70% of patients (according to a meta-analyses of published studies) and in up to 98% in a small German multicenter study (treatment initiated an average of 3 months after infection). In the current interferon-free therapy era, as of 2016, six different all-oral, brief-duration (most lasting 12 weeks), very well-tolerated, highly effective (sustained virologic response rates exceeding 90–95%) combination regimens (of polymerase inhibitors, protease inhibitors, and/or NS5A inhibitors) are available to treat patients with chronic hepatitis C (**see Chap. 334**); the same regimens are available and recommended to treat patients with acute hepatitis C. Although the duration of therapy for acute hepatitis C has not been determined definitively, in a study of 20 patients, acute hepatitis C resolved after treatment lasting only 6 weeks. In 2016, the European Association for the Study of the Liver (EASL) recommended 8 weeks of treatment for acute hepatitis C with a genotype-appropriate (**see Chap. 334**) direct-acting antiviral regimen consisting of sofosbuvir plus one of the three approved NS5A inhibitors without ribavirin (12 weeks for patients with acute hepatitis C and either a baseline HCV RNA level >1 million IU/mL or HIV co-infection).

Because spontaneous recovery can occur and because most cases of acute hepatitis C are not clinically severe or rapidly progressive, delaying antiviral therapy of acute hepatitis C for at least 12–16 weeks and even up to 6 months (after which recovery is unlikely) is a recommended approach. Patients with jaundice, those with HCV genotype 1, women, earlier age of infection, lower level of HCV RNA, HBV co-infection, and absence of current injection-drug use are more likely to recover from acute hepatitis C, as are persons who have genetic markers associated with spontaneous recovery (*IL28B* CC haplotype). Because of the marked reduction over the past three decades in the frequency of acute hepatitis C, opportunities to identify and treat patients with acute hepatitis C are rare, except in two population subsets: (1) In health workers who sustain hepatitis C–contaminated needle sticks (occupational accidents), monitoring for ALT elevations and the presence of HCV RNA identify acute hepatitis C in ~3%, and this group should be treated; (2) in injection-drug users, the risk of acute hepatitis C has been on the rise, and the epidemic of opioid use has contributed to an amplification of HCV infection among drug users. Such patients are candidates for antiviral therapy, and efforts to combine antiviral therapy with drug-rehabilitation therapy have been very successful.

Notwithstanding these specific therapeutic considerations, in most cases of typical acute viral hepatitis, specific treatment generally is not necessary. Although hospitalization may be required for clinically severe illness, most patients do not require hospital care. Forced and prolonged bed rest is not essential for full recovery, but many patients will feel better with restricted physical activity. A high-calorie diet is desirable, and because many patients may experience nausea late in the day, the major caloric intake is best tolerated in the morning. Intravenous feeding is necessary in the acute stage if the patient has persistent vomiting and cannot maintain oral intake. Drugs capable of producing adverse reactions such as cholestasis and drugs metabolized by the liver should be avoided. If severe pruritus is present, the use of the bile salt-sequestering resin cholestyramine is helpful. Glucocorticoid therapy has no value in acute viral hepatitis, even in severe cases, and may be deleterious, even increasing the risk of chronicity (e.g., of acute hepatitis B).

Physical isolation of patients with hepatitis to a single room and bathroom is rarely necessary except in the case of fecal incontinence for hepatitis A and E or uncontrolled, voluminous bleeding for hepatitis B (with or without concomitant hepatitis D) and C. Because most patients hospitalized with hepatitis A excrete little, if any, HAV, the likelihood of HAV transmission from these patients during their hospitalization is low. Therefore, burdensome *enteric precautions are no longer recommended*. Although gloves should be worn when the bed pans or fecal material of patients with hepatitis A are handled, these precautions do not represent a departure from sensible procedure and contemporary universal precautions for all hospitalized patients. For patients with hepatitis B and C, emphasis should be placed on blood precautions (i.e., avoiding direct, ungloved hand contact with blood and other body fluids). Enteric precautions are

unnecessary. The importance of simple hygienic precautions such as hand washing cannot be overemphasized. Universal precautions that have been adopted for all patients apply to patients with viral hepatitis. Hospitalized patients may be discharged following substantial symptomatic improvement, a significant downward trend in the serum aminotransferase and bilirubin values, and a return to normal of the PT. Mild aminotransferase elevations should not be considered contraindications to the gradual resumption of normal activity.

In *fulminant hepatitis*, the goal of therapy is to support the patient by maintenance of fluid balance, support of circulation and respiration, control of bleeding, correction of hypoglycemia, and treatment of other complications of the comatose state in anticipation of liver regeneration and repair. Protein intake should be restricted, and oral lactulose administered. Glucocorticoid therapy has been shown in controlled trials to be ineffective. Likewise, exchange transfusion, plasmapheresis, human cross-circulation, porcine liver cross-perfusion, hemoperfusion, and extracorporeal liver-assist devices have not been proven to enhance survival. Meticulous intensive care that includes prophylactic antibiotic coverage is the one factor that appears to improve survival. Orthotopic liver transplantation is resorted to with increasing frequency, with excellent results, in patients with fulminant hepatitis (**Chap. 338**). Fulminant hepatitis C is very rare; however, in fulminant hepatitis B, oral antiviral therapy has been used successfully, as reported anecdotally. In clinically severe hepatitis E (with jaundice and coagulopathy), successful therapy with ribavirin (600 mg twice daily, 15 mg/kg) has been reported anecdotally. Unfortunately, when fulminant hepatitis E occurs in pregnant women (as it does in up to 20% of pregnant women with acute hepatitis E), ribavirin, which is teratogenic, is contraindicated.

PROPHYLAXIS

Because application of therapy for acute viral hepatitis is limited and because antiviral therapy for chronic viral hepatitis is cumbersome, costly, and not effective in all patients (**Chap. 334**), emphasis is placed on prevention through immunization. The prophylactic approach differs for each of the types of viral hepatitis. In the past, immunoprophylaxis relied exclusively on passive immunization with antibody-containing globulin preparations purified by cold ethanol fractionation from the plasma of hundreds of normal donors. Currently, for hepatitis A, B, and E, active immunization with vaccines is the preferable approach to prevention.

Hepatitis A

Both passive immunization with IG and active immunization with killed vaccines are available. All preparations of IG contain anti-HAV concentrations sufficient to be protective. When administered before exposure or during the early incubation period, IG is effective in preventing clinically apparent hepatitis A. For postexposure prophylaxis of intimate contacts (household, sexual, institutional) of persons with hepatitis A, the administration of 0.02 mL/kg is recommended as early after exposure as possible; it may be effective even when administered as late as 2 weeks after exposure. Prophylaxis is not necessary for those who have already received hepatitis A vaccine, for casual contacts (office, factory, school, or hospital), for most elderly persons, who are very likely to be immune, or for those known to have anti-HAV in their serum. In day care centers, recognition of hepatitis A in children or staff should provide a stimulus for immunoprophylaxis in the center and in the children's family members. By the time most common-source outbreaks of hepatitis A are recognized, it is usually too late in the incubation period for IG to be effective; however, prophylaxis may limit the frequency of secondary cases. For travelers to tropical countries, developing countries, and other areas outside standard tourist routes, IG prophylaxis had been recommended before a vaccine became available. When such travel lasted <3 months, 0.02 mL/kg was given; for longer travel or residence in these areas, a dose of 0.06 mL/kg every 4–6 months was recommended. Administration of plasma-derived globulin is safe; all contemporary lots of IG are subjected to viral inactivation steps and must be free of HCV RNA as determined by PCR testing. Administration of IM lots of IG has not been associated with transmission of HBV, HCV, or HIV.

Formalin-inactivated vaccines made from strains of HAV attenuated in tissue culture have been shown to be safe, immunogenic, and effective in preventing hepatitis A. Hepatitis A vaccines are approved for use in persons who are at least 1 year old and appear to provide adequate protection beginning 4 weeks after a primary inoculation. If it can be given within 4 weeks of an expected exposure, such as by travel to an endemic area, hepatitis A vaccine is the preferred approach to *preexposure* immunoprophylaxis. If travel is more imminent, IG (0.02 mL/kg) should be administered at a different injection site, along with the first dose of vaccine. Because vaccination provides long-lasting protection (protective levels of anti-HAV should last 20 years after vaccination), persons whose risk will be sustained (e.g., frequent travelers or those remaining in endemic areas for prolonged periods) should be vaccinated, and vaccine should supplant the need for repeated IG injections. Shortly after its introduction, hepatitis A vaccine was recommended for children living in communities with a high incidence of HAV infection; in 1999, this recommendation was extended to include all children living in states, counties, and communities with high rates of HAV infection. As of 2006, the Advisory Committee on Immunization Practices of the U.S. Public Health Service recommended *routine hepatitis A vaccination of all children*. Other groups considered being at increased risk for HAV infection and who are candidates for hepatitis A vaccination include military personnel, populations with cyclic outbreaks of hepatitis A (e.g., Alaskan natives), employees of day care centers, primate handlers, laboratory workers exposed to hepatitis A or fecal specimens, and patients with chronic liver disease. Because of an increased risk of fulminant hepatitis A—observed in some experiences but not confirmed in others—among patients with chronic hepatitis C, patients with chronic hepatitis C are candidates for hepatitis A vaccination, as are persons with chronic hepatitis B. Other populations whose recognized risk of hepatitis A is increased should be vaccinated, including men who have sex with men, injection drug users, persons with clotting disorders who require frequent administration of clotting-factor concentrates, persons traveling from the United States to countries with high or intermediate hepatitis A endemicity, postexposure prophylaxis for contacts of persons with hepatitis A, and household members and other close contacts of adopted children arriving from countries with high and moderate hepatitis A endemicity. Recommendations for dose and frequency differ for the two approved vaccine preparations (Table 332-7); all injections are IM. Hepatitis A vaccine has been reported to be effective in preventing secondary household and day care center–associated cases of acute hepatitis A. Because the vaccine provides long-lasting protection and is simpler to use, in 2006, the Immunization Practices Advisory Committee of the U.S. Public Health Service favored hepatitis A vaccine to IG for postexposure prophylaxis of healthy persons age 2–40 years; for younger or older persons, for immunosuppressed patients, and for patients with chronic liver disease, IG should continue to be used. In the United States, reported mortality resulting from hepatitis A declined in parallel with hepatitis A vaccine–associated reductions in the annual incidence of new infections.

TABLE 332-7

Hepatitis A Vaccination Schedules

AGE, YEARS	NO. OF DOSES	DOSE	SCHEDULE, MONTHS
HAVRIX (GlaxoSmithKline)[a]			
1–18 ≥19	2 2	720 ELU[b] (0.5 mL) 1440 ELU (1 mL)	0, 6–12 0, 6–12
VAQTA (Merck)			
1–18 ≥19	2 2	25 units (0.5 mL) 50 units (1 mL)	0, 6–18 0, 6–18

[a]A combination of this hepatitis A vaccine and hepatitis B vaccine, TWINRIX, is licensed for simultaneous protection against both of these viruses among adults (age ≥18 years). Each 1-mL dose contains 720 ELU of hepatitis A vaccine and 20 μg of hepatitis B vaccine. These doses are recommended at months 0, 1, and 6. [b]Enzyme-linked immunoassay units.

Hepatitis B

Until 1982, prevention of hepatitis B was based on *passive* immunoprophylaxis either with standard immunoglobulin, containing modest levels of anti-HBs, or hepatitis B immunoglobulin (HBIG), containing high-titer anti-HBs. The efficacy of standard IG has never been established and remains questionable; even the efficacy of HBIG, demonstrated in several clinical trials, has been challenged, and its contribution appears to be in reducing the frequency of clinical *illness*, not in preventing *infection*. The first vaccine for *active* immunization, introduced in 1982, was prepared from purified, noninfectious 22-nm spherical HBsAg particles derived from the plasma of healthy HBsAg carriers. In 1987, the plasma-derived vaccine was supplanted by a genetically engineered vaccine derived from recombinant yeast. The latter vaccine consists of HBsAg particles that are nonglycosylated but are otherwise indistinguishable from natural HBsAg; two recombinant vaccines are licensed for use in the United States. Current recommendations can be divided into those for preexposure and postexposure prophylaxis.

For *preexposure* prophylaxis against hepatitis B in settings of frequent exposure (health workers exposed to blood; first-responder public safety workers; hemodialysis patients and staff; residents and staff of custodial institutions for the developmentally handicapped; injection drug users; inmates of long-term correctional facilities; persons with multiple sexual partners or who have had a sexually transmitted disease; men who have sex with men; persons such as hemophiliacs who require long-term, high-volume therapy with blood derivatives; household and sexual contacts of persons with chronic HBV infection; persons living in or traveling extensively in endemic areas; unvaccinated children aged <18; unvaccinated children who are Alaskan natives, Pacific Islanders, or residents in households of first-generation immigrants from endemic countries; persons born in countries with a prevalence of HBV infection ≥2%; patients with chronic liver disease; persons < age 60 with diabetes mellitus [those ≥60 at the discretion of their physicians]; persons with end-stage renal disease; and persons with HIV infection), three IM (deltoid, not gluteal) injections of hepatitis B vaccine are recommended at 0, 1, and 6 months (other, optional schedules are summarized in **Table 332-8**). Pregnancy is

not a contraindication to vaccination. In areas of low HBV endemicity such as the United States, despite the availability of safe and effective hepatitis B vaccines, a strategy of vaccinating persons in high-risk groups was not effective. The incidence of new hepatitis B cases continued to increase in the United States after the introduction of vaccines; <10% of all targeted persons in high-risk groups were actually vaccinated, and ~30% of persons with sporadic acute hepatitis B did not fall into any high-risk-group category. Therefore, to have an impact on the frequency of HBV infection in an area of low endemicity such as the United States, universal hepatitis B vaccination in childhood has been recommended. For unvaccinated children born after the implementation of universal infant vaccination, vaccination during early adolescence, at age 11–12 years, was recommended, and this recommendation has been extended to include all unvaccinated children age 0–19 years. In HBV-hyperendemic areas (e.g., Asia), universal vaccination of children has resulted in a marked (~70–90%) 30-year decline in complications of hepatitis B, including liver-related mortality and hepatocellular carcinoma.

TABLE 332-8

Preexposure Hepatitis B Vaccination Schedules

TARGET GROUP	NO. OF DOSES	DOSE	SCHEDULE, MONTHS
RECOMBIVAX-HB (Merck)[a]			
Infants, children (<1–10 years)	3	5 µg (0.5 mL)	0, 1–2, 4–6
Adolescents (11–19 years)	3 or 4	5 µg (0.5 mL)	0–2, 1–4, 4–6 *or* 0, 12, 24 *or* 0, 1, 2, 12
Adults (≥20 years)	*Or*	10 µg (1 mL)	0, 4–6 (age 11–15)
Hemodialysis patients[b]	2	10 µg (1 mL)	0–2, 1–4, 4–6
<20 years	3	5 µg (0.5 mL)	0, 1, 6
≥20 years	3	40 µg (4 mL)	0, 1, 6
	3		
ENGERIX-B (GlaxoSmithKline)[c]			
Infants, children (<1–10 years)	3 or 4	10 µg (0.5 mL)	0, 1–2, 4–6 *or* 0, 1, 2, 12
Adolescents (10–19 years)	3 or 4	10 µg (0.5 mL)	0, 1–2, 4–6 *or* 0, 12, 24 *or* 0, 1, 2, 12
Adults (≥20 years)	3 or 4	20 µg (1 mL)	0–2, 1–4, 4–6 or 0, 1, 2, 12
Hemodialysis patients[b]	4	10 µg (0.5 mL)	0, 1, 2, 6
<20 years	4	40 µg (2 mL)	0, 1, 2, 6
≥20 years			

[a]This manufacturer produces a licensed combination of hepatitis B vaccine and vaccines against *Haemophilus influenzae* type b and *Neisseria meningitides*, Comvax, for use in infants and young children. Please consult product insert for dose and schedule. [b]This group also includes other immunocompromised persons. [c]This manufacturer produces two licensed combination hepatitis B vaccines: (1) Twinrix, recombinant hepatitis B vaccine plus inactivated hepatitis A vaccine, is licensed for simultaneous protection against both of these viruses among adults (age ≥18 years). Each 1-mL dose contains 720 ELU of hepatitis A vaccine and 20 µg of hepatitis B vaccine. These doses are recommended at months 0, 1, and 6. (2) Pediarix, recombinant hepatitis B vaccine plus diphtheria and tetanus toxoid, pertussis, and inactivated poliovirus, is licensed for use in infants and young children. Please consult product insert for doses and schedules.

The two available recombinant hepatitis B vaccines are comparable, one containing 10 µg of HBsAg (Recombivax-HB) and the other containing 20 µg of HBsAg (Engerix-B), and recommended doses for each injection vary for the two preparations (Table 332-8). Combinations of hepatitis B vaccine with other childhood vaccines are available as well (Table 332-8).

For unvaccinated persons sustaining an exposure to HBV, *postexposure* prophylaxis with a combination of HBIG (for rapid achievement of high-titer circulating anti-HBs) and hepatitis B vaccine (for achievement of long-lasting immunity as well as its apparent efficacy in attenuating clinical illness after exposure) is recommended. For *perinatal* exposure of infants born to HBsAg-positive mothers, a single dose of HBIG, 0.5 mL, should be administered IM in the thigh *immediately after birth*, followed by a complete course of three injections of recombinant hepatitis B vaccine (see doses

above) to be started within the first 12 h of life. For those experiencing a direct percutaneous inoculation or transmucosal exposure to HBsAg-positive blood or body fluids (e.g., accidental *needle stick*, other mucosal penetration, or ingestion), a single IM dose of HBIG, 0.06 mL/kg, administered as soon after exposure as possible, is followed by a complete course of hepatitis B vaccine to begin within the first week. For pregnant mothers with high-level HBV DNA (>2 × 10^5 IU/mL), adding antiviral nucleoside analogues (e.g., pregnancy class B tenofovir, see **Chap 334**) during the third trimester of pregnancy reduces perinatal transmission even further. For persons exposed by *sexual* contact to a patient with acute hepatitis B, a single IM dose of HBIG, 0.06 mL/kg, should be given within 14 days of exposure, to be followed by a complete course of hepatitis B vaccine. When both HBIG and hepatitis B vaccine are recommended, they may be given at the same time but at separate sites. Testing adults for anti-HBs after a course of vaccine is advisable to document the acquisition of immunity, but, because hepatitis B vaccine immunogenicity is nearly universal in infants, postvaccination anti-HBs testing of children is not recommended.

The precise duration of protection afforded by hepatitis B vaccine is unknown; however, ~80–90% of immunocompetent adult vaccinees retain protective levels of anti-HBs for at least 5 years, and 60–80% for 10 years, and protective antibody has been documented to last for at least two decades after vaccination in infancy. Thereafter and even after anti-HBs becomes undetectable, protection persists against clinical hepatitis B, hepatitis B surface antigenemia, and chronic HBV infection. Currently, *booster* immunizations are not recommended routinely, except in immunosuppressed persons who have lost detectable anti-HBs or immunocompetent persons who sustain percutaneous HBsAg-positive inoculations after losing detectable antibody. Specifically, for hemodialysis patients, annual anti-HBs testing is recommended after vaccination; booster doses are recommended when anti-HBs levels fall to <10 mIU/mL. As noted above, for persons at risk of both hepatitis A and B, a combined vaccine is available containing 720 enzyme-linked immunoassay units (ELUs) of inactivated HAV and 20 μg of recombinant HBsAg (at 0, 1, and 6 months).

Hepatitis D

Infection with hepatitis D can be prevented by vaccinating susceptible persons with hepatitis B vaccine. No product is available for immunoprophylaxis to prevent HDV superinfection in persons with chronic HBV infection; for them, avoidance of percutaneous exposures and limitation of intimate contact with persons who have HDV infection are recommended.

Hepatitis C

IG is ineffective in preventing hepatitis C and is no longer recommended for postexposure prophylaxis in cases of perinatal, needle stick, or sexual exposure. Although prototype vaccines that induce antibodies to HCV envelope proteins have been developed, currently, hepatitis C vaccination is not feasible practically. Genotype and quasispecies viral heterogeneity, as well as rapid evasion of neutralizing antibodies by this rapidly mutating virus, conspire to render HCV a difficult target for immunoprophylaxis with a vaccine. Prevention of transfusion-associated hepatitis C has been accomplished by the following successively introduced measures: exclusion of commercial blood donors and reliance on a volunteer blood supply; screening donor blood with surrogate markers such as ALT (no longer recommended) and anti-HBc, markers that identify segments of the blood donor population with an increased risk of bloodborne infections; exclusion of blood donors in high-risk groups for AIDS and the introduction of anti-HIV screening tests; and progressively sensitive serologic and virologic screening tests for HCV infection.

In the absence of active or passive immunization, prevention of hepatitis C includes behavior changes and precautions to limit exposures to infected persons. Recommendations designed to identify patients with clinically inapparent hepatitis as candidates for medical management have as a secondary benefit the identification of persons whose

contacts could be at risk of becoming infected. A so-called look-back program has been recommended to identify persons who were transfused before 1992 with blood from a donor found subsequently to have hepatitis C. In addition, anti-HCV testing is recommended for persons born between 1945 and 1965, anyone who received a blood transfusion or a transplanted organ before the introduction of second-generation screening tests in 1992, those who ever used injection drugs (or took other illicit drugs by noninjection routes), chronically hemodialyzed patients, persons with clotting disorders who received clotting factors made before 1987 from pooled blood products, persons with elevated aminotransferase levels, health workers exposed to HCV-positive blood or contaminated needles, recipients of blood or organs from a donor found to be positive for hepatitis C, persons with HIV infection, health care and public safety personnel following a needle stick or other nonpercutaneous exposure to HCV-infected material, sexual partners of persons with hepatitis C, and children born to HCV-positive mothers (Table 332-4).

For stable, monogamous sexual partners, sexual transmission of hepatitis C is unlikely, and sexual barrier precautions are not recommended. For persons with multiple sexual partners or with sexually transmitted diseases, the risk of sexual transmission of hepatitis C is increased, and barrier precautions (latex condoms) are recommended. A person with hepatitis C should avoid sharing such items as razors, toothbrushes, and nail clippers with sexual partners and family members. No special precautions are recommended for babies born to mothers with hepatitis C, and breast-feeding does not have to be restricted.

Hepatitis E

Whether IG prevents hepatitis E remains undetermined. Safe and effective recombinant genotype 1 vaccines, which protect against other genotypes as well, have been developed and are available in endemic areas but not in the United States. Protection provided by the Chinese hepatitis E vaccine is long-lasting, documented in a clinical trial up to 4.5 years.

FURTHER READING

Buckley GJ, Strom BL (eds): *Eliminating the Public Health Problem of Hepatitis B and C in the United States: Phase One Report.* Washington DC, National Academies Press, 2016. Available from *http://www.nap.edu/23407*. Accessed September 22, 2016.

Centers for Disease Control and Prevention: Recommendations for the identification of chronic hepatitis C virus infection among persons born during 1945–1965. MMWR 61(RR-4):1, 2012.

Debing Y et al: Update on hepatitis E virology: Implications for clinical practice. J Hepatol 65:200, 2016. [PubMed: 26966047]

Denniston MM et al: Chronic hepatitis C virus infection in the United States, National Health and Nutrition Examination Survey 2003–2010. Ann Intern Med 160:293, 2014. [PubMed: 24737271]

Douam F et al: The mechanism of HCV entry into host cells. Prog Mol Biol Transl Sci 129:63, 2015. [PubMed: 25595801]

Joy JB et al: The spread of hepatitis C virus genotype 1a in North America: A retrospective phylogenetic study. Lancet Infect Dis 16:698, 2016. [PubMed: 27039040]

Lin H-H et al: Changing hepatitis D virus epidemiology in a hepatitis B virus endemic area with a national vaccination program. Hepatology 61:1870, 2016.

Pan CQ et al: Tenofovir to prevent hepatitis B transmission in mothers with high viral load. N Engl J Med 374:2324, 2016. [PubMed: 27305192]

Roberts H et al: Prevalence of chronic hepatitis B virus (HBV) infection in U.S. households: National Health and Nutrition Examination Survey (NHANES), 1988–2012. Hepatology 63:388, 2016. [PubMed: 26251317]

Trépo C et al: Hepatitis B virus infection. Lancet 384:2053, 2014. [PubMed: 24954675]

ONLINE BIBLIOGRAPHY

Bartosch B: Hepatitis C virus and its complex interplay with hepatic glucose and lipid metabolism. J Hepatol 50:845, 2009. [PubMed: 19329215]

Centers for Disease Control and Prevention: A comprehensive immunization strategy to eliminate transmission of hepatitis B virus infection in the United States: Recommendations of the Advisory Committee on Immunization Practices (ACIP) part 1: Immunization of infants, children, and adolescents. MMWR 54 (RR-16):1, 2005; part II: Immunization of adults. MMWR 55(RR-16):1, 2006.

Centers for Disease Control and Prevention: Prevention of hepatitis A through active or passive immunization: Recommendations of the Advisory Committee on Immunization Practices (ACIP). MMWR 55 (RR-7):1, 2006.

Centers for Disease Control and Prevention: Recommendations for identification and public health management of persons with chronic hepatitis B virus infection. MMWR 57(RR-8):1, 2008.

Centers for Disease Control and Prevention: Updated recommendations from the Advisory Committee on Immunization Practices (ACIP) for use of hepatitis A vaccine in close contacts of newly arriving international adoptees. MMWR 58:1006, 2009. [PubMed: 19763077]

Chang M-H et al: Long-term effects of hepatitis B immunization of infants in preventing liver cancer. Gastroenterology 151:472, 2016. [PubMed: 27269245]

Dienstag JL: Hepatitis B virus infection. N Engl J Med 359:1486, 2008. [PubMed: 18832247]

Holmberg SD et al: Hepatitis C in the United States. N Engl J Med 368:1859, 2013. [PubMed: 23675657]

Hughes SA et al: Hepatitis delta virus. Lancet 378:73, 2011. [PubMed: 21511329]

Lindenbach BD, Rice CM: Unravelling hepatitis C virus replication from genome to function. Nature 436:933, 2005. [PubMed: 16107832]

Lindenbach BD et al: Complete replication of hepatitis C virus in cell culture. Science 309:623, 2005. [PubMed: 15947137]

Ly KN et al: Rising mortality associated with hepatitis C virus in the United States 2003–2013. Clin Infect Dis 62:1287, 2016. [PubMed: 26936668]

Miller RH et al: Compact organization of the hepatitis B virus genome. Hepatology 9:322, 1989. [PubMed: 2643549]

Moradpour D et al: Replication of hepatitis C virus. Nature Rev Microbiol 5:453, 2007.

Murphy TV et al: Progress toward eliminating hepatitis A disease in the United States. MMWR 65:29, 2016.

Rehermann B: Pathogenesis of chronic viral hepatitis: differential role of T cells and NK cells. Nature Med 19:859, 2013. [PubMed: 23836236]

Rizzetto M, Alavian SM: Hepatitis delta: the rediscovery. Clin Liver Dis 17:475, 2013. [PubMed: 23905817]

Seeger C et al: Biochemical and genetic evidence for the hepatitis B virus replication strategy. Science 232:477, 1986. [PubMed: 3961490]

Zhang J et al: Long-term efficacy of a hepatitis E vaccine. N Engl J Med 372:914, 2015. [PubMed: 25738667]

Chapter 333: Toxic and Drug-Induced Hepatitis

William M. Lee; Jules L. Dienstag

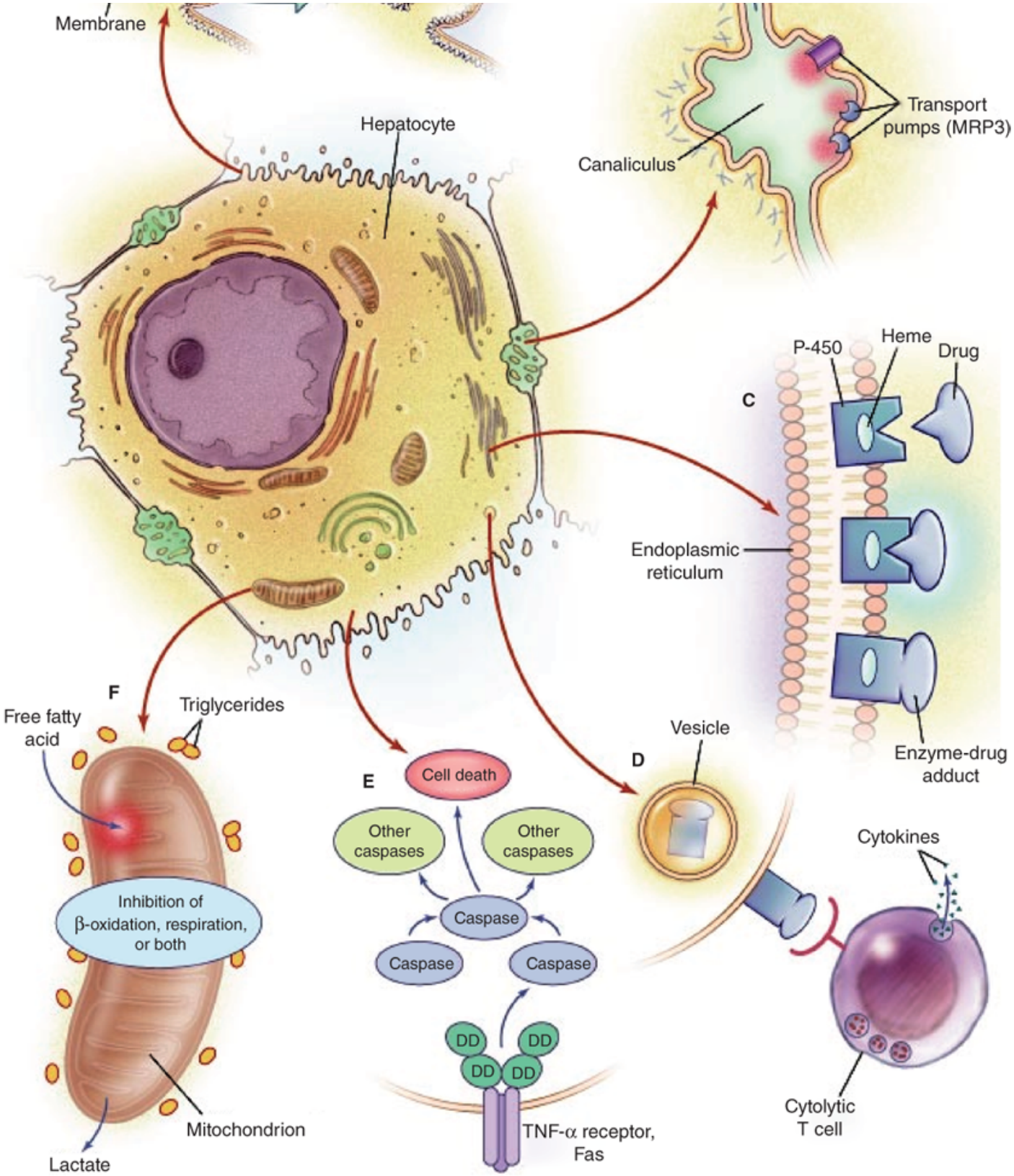
INTRODUCTION

Liver injury is a possible consequence of ingestion of any xenobiotic, including industrial toxins, pharmacologic agents, and complementary and alternative medications (CAMs). Among patients with acute liver failure, drug-induced liver injury (DILI) is the most common cause, and evidence for hepatotoxicity detected during clinical trials for drug development is the most common reason for failure of compounds to reach approval status. DILI requires careful history taking to identify unrecognized exposure to chemicals used in work or at home, drugs taken by prescription or bought over the counter, and herbal or dietary supplement medicines. Hepatotoxic drugs can injure the hepatocyte directly, for example, via a free-radical or metabolic intermediate that causes peroxidation of membrane lipids and that results in liver cell injury. Alternatively, a drug or its metabolite may activate components of the innate or adaptive immune system, stimulate apoptotic pathways, or initiate damage to bile excretory pathways **(Fig. 333-1)**. Interference with bile canalicular pumps can allow endogenous bile acids, which can injure the liver, to accumulate. Such secondary injury, in turn, may lead to necrosis of hepatocytes; injure bile ducts, producing cholestasis; or block pathways of lipid movement, inhibit protein synthesis, or impair mitochondrial oxidation of fatty acids, resulting in lactic acidosis and intracellular triglyceride accumulation (expressed histologically as microvesicular steatosis). In other instances, drug metabolites sensitize hepatocytes to toxic cytokines. The differences observed between susceptible and nonsusceptible drug recipients may be attributable to human leukocyte antigens (HLA) haplotypes that determine binding of drug-related haptens on the cell surface as well as to polymorphisms in elaboration of competing, protective cytokines, as has been suggested for acetaminophen hepatotoxicity (see below). Immune mechanisms may include cytotoxic lymphocytes or antibody-mediated cellular cytotoxicity. In addition, a role has been shown for activation of nuclear transporters, such as the constitutive androstane receptor (CAR) or, more recently, the pregnane X receptor (PXR), in the induction of drug hepatotoxicity.

FIGURE 333-1

Potential mechanisms of drug-induced liver injury. The normal hepatocyte may be affected adversely by drugs through (***A***) disruption of intracellular calcium homeostasis that leads to the disassembly of actin fibrils at the surface of the hepatocyte, resulting in blebbing of the cell membrane, rupture, and cell lysis; (***B***) disruption of actin filaments next to the canaliculus (the specialized portion of the cell responsible for bile excretion), leading to loss of villous processes and interruption of transport pumps such as multidrug resistance–associated protein 3 (MRP3), which, in turn, prevents the excretion of bilirubin and other organic compounds; (***C***) covalent binding of the heme-containing cytochrome P450 enzyme to the drug, thus creating nonfunctioning adducts; (***D***) migration of these enzyme-drug adducts to the cell surface in vesicles to serve as target immunogens for cytolytic attack by T cells, stimulating an immune response involving cytolytic T cells and cytokines; (***E***) activation of apoptotic pathways by tumor necrosis factor α (TNF-α) receptor or Fas (DD denotes death domain), triggering the cascade of intercellular caspases, resulting in programmed cell death; or (***F***) inhibition of mitochondrial function by a dual effect on both β-oxidation and the respiratory-chain enzymes, leading to failure of free fatty acid metabolism, a lack of aerobic respiration, and accumulation of lactate and reactive oxygen species (which may disrupt mitochondrial DNA). Toxic metabolites excreted in bile may damage bile-duct epithelium (not shown). CTLs, cytolytic T lymphocytes. *(Reproduced from WM Lee: Drug-induced hepatotoxicity. N Engl J Med 349:474, 2003, with permission.)*





A. Rupture of cell membrane.
B. Injury of bile canaliculus (disruption of transport pumps).
C. P-450-drug covalent binding (drug adducts).
D. Drug adducts targeted by CTLs/cytokines.
E. Activation of apoptotic pathway by TNFα/Fas.
F. Inhibition of mitochondrial function.

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DRUG METABOLISM

Most drugs, which are water-insoluble, undergo a series of metabolic steps, culminating in a water-soluble form appropriate for renal or biliary excretion. This process begins with oxidation or methylation mediated initially by the microsomal mixed-function oxygenases, cytochrome P450 (phase I reaction), followed by glucuronidation or sulfation (phase II reaction) or inactivation by glutathione. Most drug hepatotoxicity is mediated by a phase I toxic metabolite, but glutathione depletion, precluding inactivation of harmful compounds by glutathione S-transferase, can contribute as well.

LIVER INJURY CAUSED BY DRUGS

In general, two major types of chemical hepatotoxicity have been recognized: (1) direct toxic and (2) idiosyncratic. As shown in Table 333-1, direct toxic hepatitis occurs with predictable regularity in individuals exposed to the offending agent and is dose-dependent. The latent period between exposure and liver injury is usually short (often several hours), although clinical manifestations may be delayed for 24–48 h. Agents producing toxic hepatitis are generally systemic poisons or are converted in the liver to toxic metabolites. The direct hepatotoxins result in morphologic abnormalities that are reasonably characteristic and reproducible for each toxin. For example, carbon tetrachloride and trichloroethylene characteristically produce a centrilobular zonal necrosis, whereas yellow phosphorus poisoning typically results in periportal injury. The hepatotoxic octapeptides of *Amanita phalloides* usually produce massive hepatic necrosis; the lethal dose of the toxin is ~10 mg, the amount found in a single deathcap mushroom. Liver injury, which is often only one facet of the toxicity produced by the direct hepatotoxins, may go unrecognized until jaundice appears.

TABLE 333-1

Some Features of Toxic and Drug-Induced Hepatic Injury

Features	Direct Toxic Effect[a]		Idiosyncratic[a]			Other[a]
	Carbon Tetrachloride	Acetaminophen	Amoxicillin-Clavulanate	Isoniazid	Ciprofloxacin	Estrogens/Androgenic Steroids
Predictable and dose-related toxicity	+	+	0	0	0	+
Latent period	Short	Short	Delayed onset	Variable	May be short	Variable
Arthralgia, fever, rash, eosinophilia	0	0	0	0	0	0
Liver morphology	Necrosis, fatty infiltration	Centrilobular necrosis	Mixed hepatocellular/cholestatic	Hepatocellular injury resembling viral hepatitis	Hepatocellular injury resembling viral hepatitis	Cholestasis *without* portal inflammation

[a]The drugs listed are typical examples.

In idiosyncratic drug reactions, the occurrence of hepatitis is usually infrequent (1 in 10^3–10^5 patients) and unpredictable; the response is not as clearly dose-dependent as is injury associated with direct hepatotoxins, and liver injury may occur at any time

during or shortly after exposure to the drug. That said, recent data suggest that most agents causing idiosyncratic toxicity are given at a daily dose exceeding 100 mg, suggesting a role for dose—drugs with low potency must be given in higher doses that engender greater chances for "off-target" effects. Adding to the difficulty of predicting or identifying idiosyncratic drug hepatotoxicity is the occurrence of mild, transient, nonprogressive serum aminotransferase elevations that resolve with continued drug use. Such "adaptation," the mechanism of which is unknown, is well recognized for drugs such as isoniazid (INH), valproate, phenytoin, and HMG-CoA reductase inhibitors (statins). Extrahepatic manifestations of hypersensitivity, such as rash, arthralgias, fever, leukocytosis, and eosinophilia, occur in about one-quarter of patients with idiosyncratic hepatotoxic drug reactions but are characteristic for certain drugs and not others. Both primary immunologic injury and direct hepatotoxicity related to idiosyncratic differences in generation of toxic metabolites have been invoked to explain idiosyncratic drug reactions. The most current data appear to implicate the adaptive immune system responding to the formation of immune stimulatory compounds resulting from phase I metabolic activation of the offending drug. Differences in host susceptibility may result from varying kinetics of toxic metabolite generation and genetic polymorphisms in downstream drug-metabolizing pathways or cytokine activation; in addition, certain HLA haplotypes have been associated with hepatotoxicity of certain drugs such as amoxicillin-clavulanate and flucloxacillin. Occasionally, however, the clinical features of an allergic reaction (prominent tissue eosinophilia, autoantibodies, etc.) are difficult to ignore and suggest activation of IgE pathways. A few instances of drug hepatotoxicity are observed to be associated with autoantibodies, including a class of antibodies to liver-kidney microsomes, anti-LKM2, directed against a cytochrome P450 enzyme.

Idiosyncratic reactions lead to a morphologic pattern that is more variable than those produced by direct toxins; a single agent is often capable of causing a variety of lesions, although certain patterns tend to predominate. Depending on the agent involved, idiosyncratic hepatitis may result in a clinical and morphologic picture indistinguishable from that of viral hepatitis (e.g., INH or ciprofloxacin). So-called hepatocellular injury is the most common form, featuring spotty necrosis in the liver lobule with a predominantly lymphocytic infiltrate resembling that observed in acute hepatitis A, B, or C. Drug-induced cholestasis ranges from mild to increasingly severe: (1) bland cholestasis with limited hepatocellular injury (e.g., estrogens, 17,α-substituted androgens); (2) inflammatory cholestasis (e.g., amoxicillin-clavulanic acid [the most frequently implicated antibiotic among cases of DILI], oxacillin, erythromycin estolate); (3) sclerosing cholangitis (e.g., after intrahepatic infusion of the chemotherapeutic agent floxuridine for hepatic metastases from a primary colonic carcinoma); and (4) disappearance of bile ducts, "ductopenic" cholestasis, similar to that observed in chronic rejection **(Chap. 338)** following liver transplantation (e.g., carbamazepine, levofloxacin). Cholestasis may result from binding of drugs to canalicular membrane transporters, accumulation of toxic bile acids resulting from canalicular pump failure, or genetic defects in canalicular transporter proteins. Clinically, the distinction between a hepatocellular and a cholestatic reaction is indicated by the R value, the ratio of alanine aminotransferase (ALT) to alkaline phosphatase values, both expressed as multiples of the upper limit of normal. An R value of >5.0 is associated with hepatocellular injury, R <2.0 with cholestatic injury, and R between 2.0 and 5.0 with mixed hepatocellular-cholestatic injury.

Morphologic alterations may also include bridging hepatic necrosis (e.g., methyldopa) or, infrequently, hepatic granulomas (e.g., sulfonamides). Some drugs result in macrovesicular or microvesicular steatosis or steatohepatitis, which, in some cases, has been linked to mitochondrial dysfunction and lipid peroxidation. Severe hepatotoxicity associated with steatohepatitis, most likely a result of mitochondrial toxicity, is being recognized with increasing frequency among patients receiving antiretroviral therapy with reverse transcriptase inhibitors for HIV infection (e.g., zidovudine, didanosine), although many of these drugs have been withdrawn because of such hepatotoxicity **(Chap. 197)**. Generally, such mitochondrial hepatotoxicity of these antiretroviral agents is reversible, but dramatic, nonreversible hepatotoxicity associated with mitochondrial injury (inhibition of DNA polymerase γ) was the cause of acute liver failure encountered during early clinical trials of now-abandoned fialuridine, a fluorinated pyrimidine analogue with potent antiviral activity against hepatitis B virus. Another potential target for idiosyncratic drug hepatotoxicity is sinusoidal lining cells; when these are injured, such as by high-dose chemotherapeutic agents (e.g., cyclophosphamide, melphalan, busulfan) administered prior to bone marrow transplantation, venoocclusive disease can result. Nodular regenerative hyperplasia, a subtle form of portal hypertension, may also result from vascular injury to portal venous endothelium following systemic chemotherapy, such as with oxaliplatin, as part of adjuvant treatment for colon cancer.

Not all adverse hepatic drug reactions can be classified as either toxic or idiosyncratic. For example, oral contraceptives, which combine estrogenic and progestational compounds, may result in impairment of liver tests and, occasionally, jaundice; however, they do not produce necrosis or fatty change, manifestations of hypersensitivity are generally absent, and susceptibility to the development of oral contraceptive–induced cholestasis appears to be genetically determined. Such estrogen-induced cholestasis is more common in women with cholestasis of pregnancy, a disorder linked to genetic defects in multidrug resistance–associated canalicular transporter proteins. A rare complication of oral contraceptive therapy is hepatic sinusoidal dilatation localized to the periportal zone of the liver lobule.

Any idiosyncratic reaction that occurs in <1:10,000 recipients will go unrecognized in most clinical trials, which involve only several thousand recipients. The U.S. Food and Drug Administration (FDA) and pharmaceutical companies have learned to look for even subtle indications of serious toxicity and monitor regularly the number of trial subjects in whom any aminotransferase elevations develop, as a possible surrogate for more serious toxicity. Even more valid as a predictor of severe hepatotoxicity is the occurrence of jaundice in patients enrolled in a clinical drug trial, so-called “Hy’s Law,” named after Hyman Zimmerman, one of the pioneers of the field of drug hepatotoxicity. He recognized that, if jaundice occurred during a phase III trial, more serious liver injury was likely, with a 10:1 ratio between cases of jaundice and liver failure—10 patients with jaundice to 1 patient with acute liver failure. Thus, the finding of such Hy’s Law cases during drug development often portends failure of approval, particularly if any of the subjects sustains a bad outcome. Troglitazone, a peroxisome proliferator-activated receptor γ agonist, was the first in its class of thiazolidinedione insulin-sensitizing agents. Although in retrospect, Hy’s Law cases of jaundice had occurred during phase III trials, no instances of liver failure were recognized until well after the drug was introduced, emphasizing the importance of postmarketing surveillance in identifying toxic drugs and in leading to their withdrawal from use. Fortunately, such hepatotoxicity is not characteristic of the second-generation thiazolidinedione insulin-sensitizing agents rosiglitazone and pioglitazone; in clinical trials, the frequency of aminotransferase elevations in patients treated with these medications did not differ from that in placebo recipients, and isolated reports of liver injury among recipients are extremely rare.

Proving that an episode of liver injury is caused by a drug is difficult in many cases. DILI is nearly always a presumptive diagnosis, and many other disorders produce a similar clinicopathologic picture. Thus, causality may be difficult to establish and requires several separate supportive assessment variables to lead to a high level of certainty, including temporal association (time of onset, time to resolution), clinical-biochemical features, type of injury (hepatocellular versus cholestatic), extrahepatic features, likelihood that a given agent is to blame based on its past record, and exclusion of other potential causes. Scoring systems such as the Roussel-Uclaf Causality Assessment Method (RUCAM) yield residual uncertainty and have not been adopted widely. Currently, the U.S. Drug-Induced Liver Injury Network (DILIN) relies on a structured expert opinion process requiring detailed data on each case and a comprehensive review by three experts who arrive at a consensus on a five-degree scale of likelihood (definite, highly likely, probable, possible, unlikely); however, this approach is not practical for routine clinical application.

Generally, drug hepatotoxicity is not more frequent in persons with underlying chronic liver disease, although the severity of the outcome may be amplified. Reported exceptions include hepatotoxicity of aspirin, methotrexate, INH (only in certain experiences), antiretroviral therapy for HIV infection, and certain drugs such as conditioning regimens for bone marrow transplantation in the presence of hepatitis C.

TREATMENT

TREATMENT

Toxic and Drug-Induced Hepatic Disease

Treatment is largely supportive, except in acetaminophen hepatotoxicity (for which *N*-acetylcysteine is effective, see below). In patients with fulminant hepatitis resulting from drug hepatotoxicity, liver transplantation may be lifesaving (Chap. 338). Withdrawal of the suspected agent is indicated at the first sign of an adverse reaction. A number of studies have suggested that lethal outcomes follow continued use of an agent in the face of symptoms and signs of liver injury. In the case of the direct toxins, liver involvement should not divert attention from renal or other organ involvement, which may also threaten survival. A number

of agents are used occasionally but are of questionable value: glucocorticoids for drug hepatotoxicity with allergic features, silibinin for hepatotoxic mushroom poisoning, and ursodeoxycholic acid for cholestatic drug hepatotoxicity have not been shown to be effective and cannot be recommended. A double-blind, randomized controlled trial of the use of *N*-acetylcysteine for non-acetaminophen acute liver failure, including cases of DILI demonstrated benefit particularly for patients with early-stage hepatic encephalopathy; however, the drug is not FDA-approved for this indication.

In Table 333-2, several classes of chemical agents are listed together with examples of the pattern of liver injury they produce. Certain drugs appear to be responsible for the development of chronic as well as acute hepatic injury. For example, nitrofurantoin, minocycline, hydralazine, and methyldopa have been associated with moderate to severe chronic hepatitis with autoimmune features. Methotrexate, tamoxifen, and amiodarone have been implicated in the development of cirrhosis. Portal hypertension in the absence of cirrhosis may result from alterations in hepatic architecture produced by vitamin A or arsenic intoxication, industrial exposure to polyvinyl chloride, or administration of thorium dioxide. The latter three agents have also been associated with angiosarcoma of the liver. Oral contraceptives have been implicated in the development of hepatic adenoma and, rarely, hepatocellular carcinoma and hepatic vein occlusion (Budd-Chiari syndrome). Another unusual lesion, peliosis hepatis (blood cysts of the liver), has been observed in some patients treated with anabolic or contraceptive steroids. The existence of these hepatic disorders expands the spectrum of liver injury induced by chemical agents and emphasizes the need for a thorough drug history in all patients with liver dysfunction. The comprehensive, authoritative LiverTox website, which contains up-to-date information on drug-induced liver injury, is available as a valuable reference through the National Institutes of Health and the National Library of Medicine (*www.livertox.nih.gov*).

TABLE 333-2

Principal Alterations of Hepatic Morphology Produced by Some Commonly Used Drugs and Chemicals[a]

Principal Morphologic Change	Class of Agent	Example
Cholestasis	Anabolic steroid	Methyl testosterone, many other body-building supplements
	Antibiotic	Erythromycin estolate, nitrofurantoin, rifampin, amoxicillin-clavulanic acid, oxacillin
	Anticonvulsant	Carbamazepine
	Antidepressant	Duloxetine, mirtazapine, tricyclic antidepressants
	Anti-inflammatory	Sulindac
	Antiplatelet	Clopidogrel
	Antihypertensive	Irbesartan, fosinopril
	Antithyroid	Methimazole
	Calcium channel blocker	Nifedipine, verapamil
	Immunosuppressive	Cyclosporine
	Lipid-lowering	Ezetimibe
	Oncotherapeutic	Anabolic steroids, busulfan, tamoxifen, irinotecan, cytarabine, temozolomide
	Oral contraceptive	Norethynodrel with mestranol
	Oral hypoglycemic	Chlorpropamide
	Tranquilizer	Chlorpromazine[b]
Fatty liver	Antiarrhythmic	Amiodarone
	Antibiotic	Tetracycline (high-dose, IV)
	Anticonvulsant	Valproic acid
	Antiviral	Dideoxynucleosides (e.g., zidovudine), protease inhibitors (e.g., indinavir, ritonavir)
	Oncotherapeutic	Asparaginase, methotrexate, tamoxifen

Principal Morphologic Change	Class of Agent	Example
Hepatitis	Anesthetic Antiandrogen Antibiotic Anticonvulsant Antidepressant Antifungal Antihypertensive Anti-inflammatory Antipsychotic Antiviral Calcium channel blocker Cholinesterase inhibitor Diuretic Laxative Norepinephrine-reuptake inhibitor Oral hypoglycemic	Halothane, fluothane Flutamide Isoniazid,[c] rifampicin, nitrofurantoin, telithromycin, minocycline,[d] pyrazinamide, trovafloxacin[e] Phenytoin, carbamazepine, valproic acid, phenobarbital Iproniazid, amitriptyline, trazodone, venlafaxine, fluoxetine, paroxetine, duloxetine, sertraline, nefazodone[e] Ketoconazole, fluconazole, itraconazole Methyldopa,[c] captopril, enalapril, lisinopril, losartan Ibuprofen, indomethacin, diclofenac, sulindac, bromfenac Risperidone Zidovudine, didanosine, stavudine, nevirapine, ritonavir, indinavir, tipranavir, zalcitabine Nifedipine, verapamil, diltiazem Tacrine Chlorothiazide Oxyphenisatin[c,e] Atomoxetine Troglitazone,[e] acarbose
Mixed hepatitis/cholestatic	Antibiotic Antibacterial Antifungal Antihistamine Immunosuppressive Lipid-lowering	Amoxicillin-clavulanic acid, trimethoprim-sulfamethoxazole Clindamycin Terbinafine Cyproheptadine Azathioprine Nicotinic acid, lovastatin, ezetimibe
Toxic (necrosis)	Analgesic Hydrocarbon Metal Mushroom Solvent	Acetaminophen Carbon tetrachloride Yellow phosphorus *Amanita phalloides* Dimethylformamide
Granulomas	Antiarrhythmic Antibiotic Anticonvulsant Anti-inflammatory Xanthine oxidase inhibitor	Quinidine, diltiazem Sulfonamides Carbamazepine Phenylbutazone Allopurinol
Vascular injury	Chemotherapeutic	Oxaliplatin, melphalan

[a]Several agents cause more than one type of liver lesion and appear under more than one category. [b]Rarely associated with primary biliary cirrhosis–like lesion. [c]Occasionally associated with chronic hepatitis or bridging hepatic necrosis or cirrhosis. [d]Associated with an autoimmune hepatitis–like syndrome. [e]Withdrawn from use because of severe hepatotoxicity.

The following are patterns of adverse hepatic reactions for some prototypic agents.

ACETAMINOPHEN HEPATOTOXICITY (DIRECT TOXIN)

Acetaminophen represents the most prevalent cause of acute liver failure in the Western world; up to 72% of patients with acetaminophen hepatotoxicity in Scandinavia—somewhat lower frequencies in the United Kingdom and the United States—progress to encephalopathy and coagulopathy. Acetaminophen causes dose-related centrilobular hepatic necrosis after single-time-point ingestions, as intentional self-harm, or over extended periods, as unintentional overdoses, when multiple drug preparations or inappropriate drug amounts are used daily for several days, for example, for relief of pain or fever. In these instances, 8 g/d, twice the daily recommended maximum dose, over several days can readily lead to liver failure. Use of opioid-acetaminophen combinations appears to be particularly harmful, because habituation to the opioid may occur with a gradual increase in opioid-acetaminophen combination dosing over days or weeks. A single dose of 10–15 g, occasionally less, may produce clinical evidence of liver injury. Fatal fulminant disease is usually (although not invariably) associated with ingestion of ≥25 g. Blood levels of acetaminophen correlate with severity of hepatic injury (levels >300 μg/mL 4 h after ingestion are predictive of the development of severe damage; levels <150 μg/mL suggest that hepatic injury is highly unlikely). Nausea, vomiting, diarrhea, abdominal pain, and shock are early manifestations occurring 4–12 h after ingestion. Then 24–48 h later, when these features are abating, hepatic injury becomes apparent. Maximal abnormalities and hepatic failure are evident 3–5 days after ingestion, and aminotransferase levels exceeding 10,000 IU/L are not uncommon (i.e., levels far exceeding those in patients with viral hepatitis). Renal failure and myocardial injury may be present. Whether or not a clear history of overdose can be elicited, clinical suspicion of acetaminophen hepatotoxicity should be raised by the presence of the extremely high aminotransferase levels in association with low bilirubin levels that are characteristic of this hyperacute injury. This biochemical signature should trigger further questioning of the subject if possible; however, denial or altered mentation may confound diagnostic efforts. In this setting, a presumptive diagnosis is reasonable, and the proven antidote, *N*-acetylcysteine—both safe and presumed to be effective even when injury has already begun to evolve—should be instituted.

Acetaminophen is metabolized predominantly by a phase II reaction to innocuous sulfate and glucuronide metabolites; however, a small proportion of acetaminophen is metabolized by a phase I reaction to a hepatotoxic metabolite formed from the parent compound by cytochrome P450 CYP2E1. This metabolite, *N*-acetyl-p-benzoquinone-imine (NAPQI), is detoxified by binding to "hepatoprotective" glutathione to become harmless, water-soluble mercapturic acid, which undergoes renal excretion. When excessive amounts of NAPQI are formed, or when glutathione levels are low, glutathione levels are depleted and overwhelmed, permitting covalent binding to nucleophilic hepatocyte macromolecules forming acetaminophen-protein "adducts." These adducts, which can be measured in serum by high-performance liquid chromatography, hold promise as diagnostic markers of acetaminophen hepatotoxicity, and a point-of-care assay for acetaminophen-Cys adducts is under development. The binding of acetaminophen to hepatocyte macromolecules is believed to lead to hepatocyte necrosis; the precise sequence and mechanism are unknown. Hepatic injury may be potentiated by prior administration of alcohol, phenobarbital, INH, or other drugs; by conditions that stimulate the mixed-function oxidase system; or by conditions such as starvation (including inability to maintain oral intake during severe febrile illnesses) that reduce hepatic glutathione levels. Alcohol induces cytochrome P450 CYP2E1; consequently, increased levels of the toxic metabolite NAPQI may be produced in chronic alcoholics after acetaminophen ingestion, but the role of alcohol in potentiating acute acetaminophen injury is still debated. Alcohol also suppresses hepatic glutathione production. Therefore, in chronic alcoholics, the toxic dose of acetaminophen may be as low as 2 g, and alcoholic patients should be warned specifically about the dangers of even standard doses of this commonly used drug. In a 2006 study, aminotransferase elevations were identified in 31–44% of normal subjects treated for 14 days with the maximal recommended dose of acetaminophen, 4 g daily (administered alone or as part of an acetaminophen-opioid combination); because these changes were transient and never associated with bilirubin elevation, the clinical relevance of these findings remains to be determined. Although underlying hepatitis C virus (HCV) infection was found to be associated with an increased risk of acute liver

injury in patients hospitalized for acetaminophen overdose, generally, in patients with nonalcoholic liver disease, acetaminophen taken in recommended doses is well tolerated. Acetaminophen use in cirrhotic patients has not been associated with hepatic decompensation. On the other hand, because of the link between acetaminophen use and liver injury, and because of the limited safety margin between safe and toxic doses, the FDA has recommended that the daily dose of acetaminophen be reduced from 4 g to 3 g (even lower for persons with chronic alcohol use), that all acetaminophen-containing products be labeled prominently as containing acetaminophen, and that the potential for liver injury be prominent in the packaging of acetaminophen and acetaminophen-containing products. Within opioid combination products, the limit for the acetaminophen component has been lowered to 325 mg per tablet.

TREATMENT

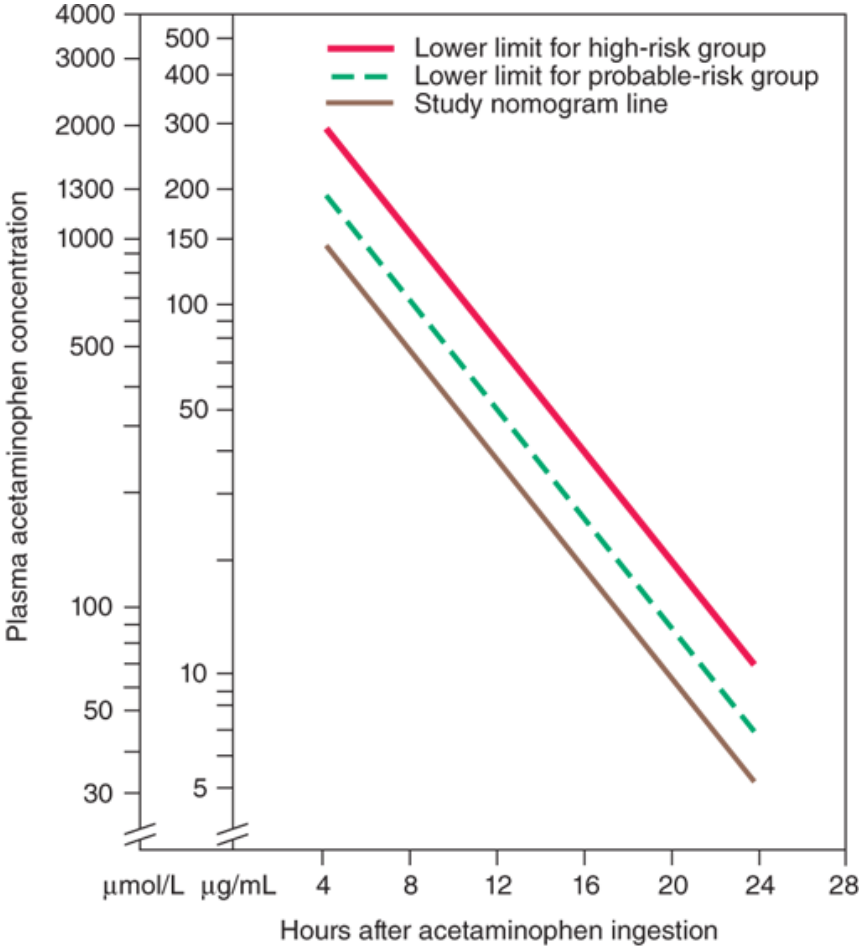
TREATMENT

Acetaminophen Overdosage

Treatment includes gastric lavage, supportive measures, and oral administration of activated charcoal or cholestyramine to prevent absorption of residual drug. Neither charcoal nor cholestyramine appears to be effective if given >30 min after acetaminophen ingestion; if they are used, the stomach lavage should be done before other agents are administered orally. The chances of possible, probable, and high-risk hepatotoxicity can be derived from a nomogram plot **(Fig. 333-2)**, readily available in emergency departments, as a function of measuring acetaminophen plasma levels 8 h after ingestion. In patients with high acetaminophen blood levels (>200 µg/mL measured at 4 h or >100 µg/mL at 8 h after ingestion), the administration of *N*-acetylcysteine reduces the severity of hepatic necrosis. This agent provides sulfhydryl donor groups to replete glutathione, which is required to render harmless toxic metabolites that would otherwise bind covalently via sulfhydryl linkages to cell proteins, resulting in the formation of drug metabolite-protein adducts. Therapy should be begun within 8 h of ingestion but may be at least partially effective when given as late as 24–36 h after overdose. Later administration of sulfhydryl compounds is of uncertain value. Routine use of *N*-acetylcysteine has substantially reduced the occurrence of fatal acetaminophen hepatotoxicity. *N*-acetylcysteine may be given orally but is more commonly used as an IV solution, with a loading dose of 140 mg/kg over 1 h, followed by 70 mg/kg every 4 h for 15–20 doses. Whenever a patient with potential acetaminophen hepatotoxicity is encountered, a local poison control center should be contacted. Treatment can be stopped when plasma acetaminophen levels indicate that the risk of liver damage is low. If signs of hepatic failure (e.g., progressive jaundice, coagulopathy, confusion) occur despite *N*-acetylcysteine therapy for acetaminophen hepatotoxicity, liver transplantation may be the only option. Early arterial blood lactate levels among such patients with acute liver failure may distinguish patients highly likely to require liver transplantation (lactate levels >3.5 mmol/L) from those likely to survive without liver replacement. Acute renal injury occurs in nearly 75% of patients with severe acetaminophen injury but is virtually always self-limited.

FIGURE 333-2

Nomogram to define risk of acetaminophen hepatotoxicity according to initial plasma acetaminophen concentration. *(After BH Rumack, H Matthew: Pediatrics 55:871, 1975.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Survivors of acute acetaminophen overdose rarely, if ever, have ongoing liver injury or sequelae.

ISONIAZID HEPATOTOXICITY (TOXIC AND IDIOSYNCRATIC REACTION)

INH remains central to most antituberculous prophylactic and therapeutic regimens, despite its long-standing recognition as a hepatotoxin. In 10% of patients treated with INH, elevated serum aminotransferase levels develop during the first few weeks of therapy; however, these elevations in most cases are self-limited, mild (values for ALT <200 IU/L), and resolve despite continued drug use. This adaptive response allows continuation of the agent if symptoms and progressive enzyme elevations do not follow the initial elevations. Acute hepatocellular drug-induced liver injury secondary to INH is evident with a variable latency period up to 6 months and is more frequent in alcoholics and patients taking certain other medications, such as barbiturates, rifampin, and pyrazinamide. If the clinical threshold of encephalopathy is reached, severe hepatic injury is likely to be fatal or to require liver transplantation. Liver biopsy reveals morphologic changes similar to those of viral hepatitis or bridging hepatic necrosis. Substantial liver injury appears to be age-related, increasing substantially after age 35; the highest frequency is in patients over age 50, and the lowest is in patients under the age of 20. Even for patients >50 years of age monitored carefully during therapy, hepatotoxicity occurs in only ~2%, well below the risk estimate derived from earlier experiences. Fever, rash, eosinophilia, and other manifestations of drug allergy are distinctly unusual. Antibodies to INH have been detected in INH recipients, but a link to causality of liver injury remains unclear. A clinical picture resembling chronic hepatitis has been observed in a few patients. Many public health programs that require INH prophylaxis for a positive tuberculin skin test or blood test (Quantiferon or T-Spot) include monthly monitoring of aminotransferase levels, although this practice has been called into question. Even more effective in limiting serious outcomes may be encouraging patients to be alert for symptoms such as nausea, fatigue, or jaundice, because most fatalities occur in the setting of continued INH use despite clinically apparent illness.

SODIUM VALPROATE HEPATOTOXICITY (TOXIC AND IDIOSYNCRATIC REACTION)

Sodium valproate, an anticonvulsant useful in the treatment of petit mal and other seizure disorders, has been associated with the development of severe hepatic toxicity and, rarely, fatalities, predominantly in children but also in adults. Among children listed as candidates for liver transplantation, valproate is the most common antiepileptic drug implicated. Asymptomatic elevations of serum aminotransferase levels have been recognized in as many as 45% of treated patients. These “adaptive” changes, however, appear to have no clinical importance, because major hepatotoxicity is not seen in the majority of patients despite continuation of drug therapy. In the rare patients in whom jaundice, encephalopathy, and evidence of hepatic failure are found, examination of liver tissue reveals microvesicular fat and bridging hepatic necrosis, predominantly in the centrilobular zone. Bile duct injury may also be apparent. Most likely, sodium valproate is not directly hepatotoxic, but its metabolite, 4-pentenoic acid, may be responsible for hepatic injury. Valproate hepatotoxicity is more common in persons with mitochondrial enzyme deficiencies and may be ameliorated by IV administration of carnitine, which valproate therapy can deplete. Valproate toxicity has been linked to HLA haplotypes (*DR4* and *B*1502*) and to mutations in mitochondrial DNA polymerase gamma 1.

NITROFURANTOIN HEPATOTOXICITY (IDIOSYNCRATIC REACTION)

This commonly used antibiotic for urinary tract infections may cause an acute hepatitis leading to fatal outcome or, more frequently, chronic hepatitis of varying severity but indistinguishable from autoimmune hepatitis. These two scenarios may reflect the frequent use and reuse of the drug for treatment of recurrent cystitis in women. Although most toxic agents manifest injury within 6 months of first ingestion, nitrofurantoin may have a longer latency period, in part perhaps because of its intermittent, recurrent use. Autoantibodies to nuclear components, smooth muscle, and mitochondria are seen and may subside after resolution of infection; however, glucocorticoid or other immunosuppressive medication may be necessary to resolve the autoimmune injury, and cirrhosis may be seen in cases that are not recognized quickly. Interstitial pulmonary fibrosis presenting as chronic cough and dyspnea may be present and resolve slowly with medication withdrawal. Histologic findings are identical to those of autoimmune hepatitis. A similar disease pattern can be observed with minocycline that is used repeatedly for the treatment of acne in teenagers as well as with hydralazine and alpha methyldopa.

AMOXICILLIN-CLAVULANATE HEPATOTOXICITY (IDIOSYNCRATIC MIXED REACTION)

Currently, the most common agent implicated as causing drug-induced liver injury in the United States and in Europe is amoxicillin-clavulanate (most frequent brand name: Augmentin). This medication causes a very specific syndrome of mixed or primarily cholestatic injury. Because hepatotoxicity may follow amoxicillin-clavulanate therapy after a relatively long latency period, the liver injury may begin to manifest at the time of drug withdrawal or after the drug has been withdrawn. The high prevalence of hepatotoxicity reflects in part the very frequent use of this drug for respiratory tract infections, including community-acquired pneumonia. The mechanism of hepatotoxicity is unclear, but the liver injury is thought to be caused by amoxicillin toxicity that is potentiated in some way by clavulanate, which itself appears not to be toxic. Symptoms include nausea, anorexia, fatigue, and jaundice—which may be prolonged—with pruritus. Rash is quite uncommon. On occasion, amoxicillin-clavulanate, like other cholestatic hepatotoxic drugs, causes permanent injury to small bile ducts, leading to the so-called “vanishing bile duct syndrome.” In vanishing bile duct syndrome, initially, liver injury is minimal except for severe cholestasis; however, over time, histologic evidence of bile duct abnormalities is replaced by a paucity and eventual absence of discernible ducts on subsequent biopsies.

PHENYTOIN HEPATOTOXICITY (IDIOSYNCRATIC REACTION)

Phenytoin, formerly diphenylhydantoin, a mainstay in the treatment of seizure disorders, has been associated in rare instances with the development of severe hepatitis-like liver injury leading to fulminant hepatic failure. In many patients, the hepatitis is associated with striking fever, lymphadenopathy, rash (Stevens-Johnson syndrome or exfoliative dermatitis), leukocytosis, and eosinophilia, suggesting an immunologically mediated hypersensitivity mechanism. Despite these observations, evidence

suggests that metabolic idiosyncrasy may be responsible for hepatic injury. In the liver, phenytoin is converted by cytochrome P450 to metabolites that include the highly reactive electrophilic arene oxides. These metabolites are normally metabolized further by epoxide hydrolases. A defect (genetic or acquired) in epoxide hydrolase activity could permit covalent binding of arene oxides to hepatic macromolecules, thereby leading to hepatic injury. Hepatic injury is usually manifest within the first 2 months after beginning phenytoin therapy. With the exception of an abundance of eosinophils in the liver, the clinical, biochemical, and histologic picture resembles that of viral hepatitis. In rare instances, bile duct injury may be the salient feature of phenytoin hepatotoxicity, with striking features of intrahepatic cholestasis. Asymptomatic elevations of aminotransferase and alkaline phosphatase levels have been observed in a sizable proportion of patients receiving long-term phenytoin therapy. These liver changes are believed by some authorities to represent the potent hepatic enzyme-inducing properties of phenytoin and are accompanied histologically by swelling of hepatocytes in the absence of necroinflammatory activity or evidence of chronic liver disease.

AMIODARONE HEPATOTOXICITY (TOXIC AND IDIOSYNCRATIC REACTION)

Therapy with this potent antiarrhythmic drug is accompanied in 15–50% of patients by modest elevations of serum aminotransferase levels that may remain stable or diminish despite continuation of the drug. Such abnormalities may appear days to many months after beginning therapy. A proportion of those with elevated aminotransferase levels have detectable hepatomegaly, and clinically important liver disease develops in <5% of patients. Features that represent a direct effect of the drug on the liver and that are common to the majority of long-term recipients are ultrastructural phospholipidosis, unaccompanied by clinical liver disease, and interference with hepatic mixed-function oxidase metabolism of other drugs. The cationic amphiphilic drug and its major metabolite desethylamiodarone accumulate in hepatocyte lysosomes and mitochondria and in bile duct epithelium. The relatively common elevations in aminotransferase levels are also considered a predictable, dose-dependent, direct hepatotoxic effect. On the other hand, in the rare patient with clinically apparent, symptomatic liver disease, liver injury resembling that seen in alcoholic liver disease is observed. The so-called pseudoalcoholic liver injury can range from steatosis, to alcoholic hepatitis–like neutrophilic infiltration and Mallory's hyaline, to cirrhosis. Electron-microscopic demonstration of phospholipid-laden lysosomal lamellar bodies can help to distinguish amiodarone hepatotoxicity from typical alcoholic hepatitis. This category of liver injury appears to be a metabolic idiosyncrasy that allows hepatotoxic metabolites to be generated. Rarely, an acute idiosyncratic hepatocellular injury resembling viral hepatitis or cholestatic hepatitis occurs. Hepatic granulomas have occasionally been observed. Because amiodarone has a long half-life, liver injury may persist for months after the drug is stopped.

ERYTHROMYCIN HEPATOTOXICITY (CHOLESTATIC IDIOSYNCRATIC REACTION)

The most important adverse effect associated with erythromycin, more common in children than adults, is the infrequent occurrence of a cholestatic reaction. Although most of these reactions have been associated with erythromycin estolate, other erythromycins may also be responsible. The reaction usually begins during the first 2 or 3 weeks of therapy and includes nausea, vomiting, fever, right upper quadrant abdominal pain, jaundice, leukocytosis, and moderately elevated aminotransferase and alkaline phosphatase levels. The clinical picture can resemble acute cholecystitis or bacterial cholangitis. Liver biopsy reveals variable cholestasis; portal inflammation comprising lymphocytes, polymorphonuclear leukocytes, and eosinophils; and scattered foci of hepatocyte necrosis. Symptoms and laboratory findings usually subside within a few days of drug withdrawal, and evidence of chronic liver disease has not been found on follow-up. The precise mechanism remains ill-defined.

ORAL CONTRACEPTIVE HEPATOTOXICITY (CHOLESTATIC REACTION)

The administration of oral contraceptive combinations of estrogenic and progestational steroids leads to intrahepatic cholestasis with pruritus and jaundice in a small proportion of patients weeks to months after taking these agents. Especially susceptible seem to be patients with recurrent idiopathic jaundice of pregnancy, severe pruritus of pregnancy, or a family history of these disorders. With the exception of liver biochemical tests, laboratory studies are normal, and extrahepatic manifestations of hypersensitivity are absent. Liver biopsy reveals cholestasis with bile plugs in dilated canaliculi and striking bilirubin staining of liver cells. In

contrast to chlorpromazine-induced cholestasis, portal inflammation is absent. The lesion is reversible on withdrawal of the agent. The two steroid components appear to act synergistically on hepatic function, although the estrogen may be primarily responsible. Oral contraceptives are contraindicated in patients with a history of recurrent jaundice of pregnancy. Primarily benign, but rarely malignant, neoplasms of the liver, hepatic vein occlusion, and peripheral sinusoidal dilatation have also been associated with oral contraceptive therapy. Focal nodular hyperplasia of the liver is not more frequent among users of oral contraceptives.

ANABOLIC STEROIDS (CHOLESTATIC REACTION)

The most common form of liver injury caused by CAMs is the profound cholestasis associated with anabolic steroids used by body builders. Unregulated agents sold in gyms and health food stores as diet supplements, which are taken by athletes to improve their performance, may contain anabolic steroids. In a young male, jaundice that is accompanied by a cholestatic, rather than a hepatitic, laboratory profile almost invariably will turn out to be caused by the use of one of a variety of androgen congeners. Such agents have the potential to injure bile transport pumps and to cause intense cholestasis; the time to onset is variable, and resolution, which is the rule, may require many weeks to months. Initially, anorexia, nausea, and malaise may occur, followed by pruritus in some but not all patients. Serum aminotransferase levels are usually <100 IU/L and serum alkaline phosphatase levels are generally moderately elevated with bilirubin levels frequently exceeding 342 µmol/L (20 mg/dL). Examination of liver tissue reveals cholestasis without substantial inflammation or necrosis. Anabolic steroids have also been used by prescription to treat bone marrow failure. In this setting, hepatic centrizonal sinusoidal dilatation and peliosis hepatis have been reported in rare patients, as have hepatic adenomas and hepatocellular carcinoma.

TRIMETHOPRIM-SULFAMETHOXAZOLE HEPATOTOXICITY (IDIOSYNCRATIC REACTION)

This antibiotic combination is used routinely for urinary tract infections in immunocompetent persons and for prophylaxis against and therapy of *Pneumocystis carinii* pneumonia in immunosuppressed persons (transplant recipients, patients with AIDS). With its increasing use, its occasional hepatotoxicity is being recognized with growing frequency. Its likelihood is unpredictable, but when it occurs, trimethoprim-sulfamethoxazole hepatotoxicity follows a relatively uniform latency period of several weeks and is often accompanied by eosinophilia, rash, and other features of a hypersensitivity reaction. Biochemically and histologically, acute hepatocellular necrosis predominates, but cholestatic features are quite frequent. Occasionally, cholestasis without necrosis occurs, and, very rarely, a severe cholangiolytic pattern of liver injury is observed. In most cases, liver injury is self-limited, but rare fatalities have been recorded. The hepatotoxicity is attributable to the sulfamethoxazole component of the drug and is similar in features to that seen with other sulfonamides; tissue eosinophilia and granulomas may be seen. The risk of trimethoprim-sulfamethoxazole hepatotoxicity is increased in persons with HIV infection.

HMG-COA REDUCTASE INHIBITORS (STATINS) (IDIOSYNCRATIC MIXED HEPATOCELLULAR AND CHOLESTATIC REACTION)

Between 1 and 2% of patients taking lovastatin, simvastatin, pravastatin, fluvastatin, or one of the newer statin drugs for the treatment of hypercholesterolemia experience asymptomatic, reversible elevations (>threefold) of aminotransferase activity. Acute hepatitis-like histologic changes, centrilobular necrosis, and centrilobular cholestasis have been described in a very small number of cases. In a larger proportion, minor aminotransferase elevations appear during the first several weeks of therapy. Careful laboratory monitoring can distinguish between patients with minor, transitory changes, who may continue therapy and those with more profound and sustained abnormalities, who should discontinue therapy. Because clinically meaningful aminotransferase elevations are so rare after statin use and do not differ in meta-analyses from the frequency of such laboratory abnormalities in placebo recipients, a panel of liver experts recommended to the National Lipid Association's Safety Task Force that liver test monitoring was not necessary in patients treated with statins and that statin therapy need not be discontinued in patients found to have asymptomatic isolated aminotransferase elevations during therapy. Statin hepatotoxicity is not increased in patients with chronic hepatitis C, hepatic steatosis, or other underlying liver diseases, and statins can be used safely in these patients.

TOTAL PARENTERAL NUTRITION (STEATOSIS, CHOLESTASIS)

Total parenteral nutrition (TPN) is often complicated by cholestatic hepatitis attributable to steatosis, cholestasis, or gallstones (or gallbladder sludge). Steatosis or steatohepatitis may result from the excess carbohydrate calories in these nutritional supplements and is the predominant form of TPN-associated liver disorder in adults. The frequency of this complication has been reduced substantially by the introduction of balanced TPN formulas that rely on lipid as an alternative caloric source. Cholestasis and cholelithiasis, caused by the absence of stimulation of bile flow and secretion resulting from the lack of oral intake, is the predominant form of TPN-associated liver disease in infants, especially in premature neonates. Often, cholestasis in such neonates is multifactorial, contributed to by other factors such as sepsis, hypoxemia, and hypotension; occasionally, TPN-induced cholestasis in neonates culminates in chronic liver disease and liver failure. When TPN-associated liver test abnormalities occur in adults, balancing the TPN formula with more lipid is the intervention of first recourse. In infants with TPN-associated cholestasis, the addition of oral feeding may ameliorate the problem. Therapeutic interventions suggested, but not shown, to be of proven benefit, include cholecystokinin, ursodeoxycholic acid, *S* adenosyl methionine, and taurine.

ALTERNATIVE AND COMPLEMENTARY MEDICINES (IDIOSYNCRATIC HEPATITIS, STEATOSIS)

Herbal medications that are of scientifically unproven efficacy and that lack prospective safety oversight by regulatory agencies account currently for more than 20% of drug-induced liver injury in the United States. Besides anabolic steroids, the most common category of dietary or herbal products is weight loss agents. Included among the herbal remedies associated with toxic hepatitis are Jin Bu Huan, xiao-chai-hu-tang, germander, chaparral, senna, mistletoe, skullcap, gentian, comfrey (containing pyrrolizidine alkaloids), ma huang, bee pollen, valerian root, pennyroyal oil, kava, celandine, Impila (*Callilepis laureola*), LipoKinetix, Hydroxycut, herbal nutritional supplements, and herbal teas containing *Camellia sinensis* (green tea extract). Well characterized are the acute hepatitis-like histologic lesions following Jin Bu Huan use: focal hepatocellular necrosis, mixed mononuclear portal tract infiltration, coagulative necrosis, apoptotic hepatocyte degeneration, tissue eosinophilia, and microvesicular steatosis. Megadoses of vitamin A can injure the liver, as can pyrrolizidine alkaloids, which often contaminate Chinese herbal preparations and can cause a venoocclusive injury leading to sinusoidal hepatic vein obstruction. Because some alternative medicines induce toxicity via active metabolites, alcohol and drugs that stimulate cytochrome P450 enzymes may enhance the toxicity of some of these products. Conversely, some alternative medicines also stimulate cytochrome P450 and may result in or amplify the toxicity of recognized drug hepatotoxins. Given the widespread use of such poorly defined herbal preparations, hepatotoxicity is likely to be encountered with increasing frequency; therefore, a drug history in patients with acute and chronic liver disease should include use of “alternative medicines” and other nonprescription preparations sold in so-called health food stores.

HIGHLY ACTIVE ANTIRETROVIRAL THERAPY (HAART) FOR HIV INFECTION (MITOCHONDRIAL TOXIC, IDIOSYNCRATIC, STEATOSIS; HEPATOCELLULAR, CHOLESTATIC, AND MIXED)

The recognition of drug hepatotoxicity in persons with HIV infection is complicated in this population by the many alternative causes of liver injury (chronic viral hepatitis, fatty infiltration, infiltrative disorders, mycobacterial infection, etc.), but drug hepatotoxicity associated with HAART is an emerging and common type of liver injury in HIV-infected persons **(Chap. 197)**. Although no one antiviral agent is recognized as a potent hepatotoxin, combination regimens including reverse transcriptase and protease inhibitors cause hepatotoxicity in ~10% of treated patients. Implicated most frequently are combinations including nucleoside analogue reverse transcriptase inhibitors zidovudine, didanosine, and, to a lesser extent, stavudine; protease inhibitors ritonavir and indinavir (and amprenavir when used together with ritonavir), as well as tipranavir; and nonnucleoside reverse transcriptase inhibitors nevirapine and, to a lesser extent, efavirenz. These drugs cause predominantly hepatocellular injury but cholestatic injury as well, and prolonged (>6 months) use of reverse transcriptase inhibitors has been associated with mitochondrial injury, steatosis, and lactic acidosis. Indirect hyperbilirubinemia, resulting from direct inhibition of bilirubin-conjugating activity by UDP-glucuronosyltransferase, usually without elevation of aminotransferase or alkaline phosphatase activities, occurs in ~10% of patients treated with the protease inhibitor indinavir. Distinguishing the impact of HAART hepatotoxicity in patients with HIV and hepatitis virus co-infection is made challenging by the following: (1) both chronic hepatitis B and hepatitis C can affect the natural history of HIV infection and the response to HAART, and (2) HAART can have an impact on chronic viral hepatitis. For example, immunologic reconstitution with HAART can result in immunologically mediated liver-cell injury in patients with chronic hepatitis B co-infection if treatment with an antiviral agent for hepatitis B (e.g., the nucleoside

analogue lamivudine) is withdrawn or if nucleoside analogue resistance emerges. Infection with HIV, especially with low CD4+ T cell counts, has been reported to increase the rate of hepatic fibrosis associated with chronic hepatitis C, and HAART therapy can increase levels of serum aminotransferases and HCV RNA in patients with hepatitis C co-infection. Didanosine or stavudine should not be used with ribavirin in patients with HIV/HCV co-infection because of an increased risk of severe mitochondrial toxicity and lactic acidosis.

ACKNOWLEDGMENT

Kurt J. Isselbacher, MD, contributed to this chapter in previous editions of Harrison's.

FURTHER READING

Björnsson ES, Hoofnagle JL: Categorization of drugs implicated in causing liver injury: Critical assessment based upon published case reports. Hepatology 63:590, 2016. [PubMed: 26517184]

Chalasani N et al: Features and outcomes of 899 patients with drug-induced liver injury: The DILIN prospective study. Gastroenterology 148:1340, 2015. [PubMed: 25754159]

Chen M et al: A model to predict severity of drug-induced liver injury in humans. Hepatology 64:931, 2016. [PubMed: 27302180]

Fontana RJ et al: Idiosyncratic drug-induced liver injury is associated with substantial morbidity and mortality within 6 months from onset. Gastroenterology 147:96, 2014. [PubMed: 24681128]

Goldberg DS et al: Population-representative incidence of drug-induced acute liver failure based on an analysis of an integrated health care system. Gastroenterology 148:1353, 2015. [PubMed: 25733099]

Hayashi PH et al: Under-reporting and poor adherence to monitoring guidelines for severe cases of isoniazid hepatotoxicity. Clin Gastroenterol Hepatol 13:1676, 2015. [PubMed: 25724701]

Hoofnagle JH et al: LiverTox: A website on drug-induced liver injury. Hepatology 57:873, 2013. Available from *www.livertox.nih.gov*. Accessed September 26, 2016. [PubMed: 23456678]

Kaplowitz N, Deleve LD (eds): *Drug-Induced Liver Disease*, 3rd ed. London, Elsevier/Academic Press, 2013.

Lee WM et al: Intravenous N-acetylcysteine improves transplant-free survival in early stage non-acetaminophen acute liver failure. Gastroenterology 137:856, 2009. [PubMed: 19524577]

Medina-Caliz I et al: Definition and risk factors for chronicity following acute idiosyncratic drug-induced liver injury. J Hepatol 65:532, 2016. [PubMed: 27184533]

Chapter 334: Chronic Hepatitis

Jules L. Dienstag

INTRODUCTION

Chronic hepatitis represents a series of liver disorders of varying causes and severity in which hepatic inflammation and necrosis continue for at least 6 months. Milder forms are nonprogressive or only slowly progressive, while more severe forms may be associated with scarring and architectural reorganization, which, when advanced, lead ultimately to cirrhosis. Several categories of chronic hepatitis have been recognized. These include chronic viral hepatitis, drug-induced chronic hepatitis **(Chap. 333)**, and autoimmune chronic hepatitis. In many cases, clinical and laboratory features are insufficient to allow assignment into one of these three categories; these "idiopathic" cases are also believed to represent autoimmune chronic hepatitis. Finally, clinical and laboratory features of chronic hepatitis are observed occasionally in patients with such hereditary/metabolic disorders as Wilson's disease (copper overload), α_1 antitrypsin deficiency **(Chaps. 337 and 408)**, and nonalcoholic fatty liver disease **(Chap. 336)** and even occasionally in patients with alcoholic liver injury **(Chap. 335)**. Although all types of chronic hepatitis share certain clinical, laboratory, and histopathologic features, chronic viral and chronic autoimmune hepatitis are sufficiently distinct to merit separate discussions. **For discussion of acute hepatitis, see Chap. 332.**

CLASSIFICATION OF CHRONIC HEPATITIS

Common to all forms of chronic hepatitis are histopathologic distinctions based on localization and extent of liver injury. These vary from the milder forms, previously labeled *chronic persistent hepatitis* and *chronic lobular hepatitis*, to the more severe form, formerly called *chronic active hepatitis*. When first defined, these designations were believed to have prognostic implications, which were not corroborated by subsequent observations. Categorization of chronic hepatitis based primarily on histopathologic features has been replaced by a more informative classification based on a combination of clinical, serologic, and histologic variables. Classification of chronic hepatitis is based on (1) its *cause*; (2) its histologic activity, or *grade*; and (3) its degree of progression based on level of fibrosis, or *stage*. Thus, neither clinical features alone nor histologic features—requiring liver biopsy or noninvasive markers of fibrosis—alone are sufficient to characterize and distinguish among the several categories of chronic hepatitis.

CLASSIFICATION BY CAUSE

Clinical and serologic features allow the establishment of a diagnosis of *chronic viral hepatitis*, caused by hepatitis B, hepatitis B plus D, or hepatitis C; *autoimmune hepatitis*, including several subcategories, I and II, based on serologic distinctions; *drug-associated chronic hepatitis*; and a category of unknown cause, or *cryptogenic chronic hepatitis* (**Table 334-1**). These are addressed in more detail below.

TABLE 334-1

Clinical and Laboratory Features of Chronic Hepatitis

TYPE OF HEPATITIS	DIAGNOSTIC TEST(s)	AUTOANTIBODIES	THERAPY
Chronic hepatitis B	HBsAg, IgG anti-HBc, HBeAg, HBV DNA	Uncommon	IFN-α, PEG IFN-α Oral agents: First-line: entecavir, tenofovir Second-line: lamivudine, adefovir, telbivudine
Chronic hepatitis C	Anti-HCV, HCV RNA	Anti-LKM1[a]	PEG IFN-α plus ribavirin[b] Direct-acting oral agents: sofosbuvir, ledipasvir, velpatasvir ritonavir-boosted paritaprevir, ombitasvir, dasabavir elbasvir, grazoprevir daclatasvir, simeprevir
Chronic hepatitis D	Anti-HDV, HDV RNA, HBsAg, IgG anti-HBc	Anti-LKM3	IFN-α, PEG IFN-α[c]
Autoimmune hepatitis	ANA[d] (homogeneous), anti-LKM1 (±) Hyperglobulinemia	ANA, anti-LKM1 anti-SLA[e]	Prednisone, azathioprine
Drug-associated	—	Uncommon	Withdraw drug
Cryptogenic	All negative	None	Prednisone (?), azathioprine (?)

[a]Antibodies to liver-kidney microsomes type 1 (autoimmune hepatitis type II and some cases of hepatitis C). [b]Supplanted in almost all cases by combinations of the direct-acting antiviral agents listed (see *www.hcvguidelines.org*). [c]Early clinical trials suggested benefit of IFN-α therapy; PEG IFN-α is as effective, if not more so, and has supplanted standard IFN-α. [d]Antinuclear antibody (autoimmune hepatitis type I). [e]Antibodies to soluble liver antigen (autoimmune hepatitis type III).

Abbreviations: HBc, hepatitis B core; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; HBV, hepatitis B virus; HCV, hepatitis C virus; HDV, hepatitis D virus; IFN-α, interferon α; IgG, immunoglobulin G; LKM, liver-kidney microsome; PEG IFN-α, pegylated interferon α; SLA, soluble liver antigen.

CLASSIFICATION BY GRADE

Grade, a histologic assessment of necroinflammatory activity, is based on examination of the liver biopsy. An assessment of important histologic features includes the degree of *periportal necrosis* and the disruption of the limiting plate of periportal hepatocytes by inflammatory cells (so-called *piecemeal necrosis* or *interface hepatitis*); the degree of confluent necrosis that links or forms bridges between vascular structures—between portal tract and portal tract or even more important bridges between portal tract and central vein—referred to as *bridging necrosis*; the degree of hepatocyte degeneration and focal necrosis within the lobule; and the degree of *portal inflammation*. Several scoring systems that take these histologic features into account have been devised, and the most popular are the histologic activity index (HAI), used commonly in the United States, and the METAVIR score, used in Europe (**Table 334-2**). Based on the presence and degree of these features of histologic activity, chronic hepatitis can be graded as mild, moderate, or severe.

TABLE 334-2

Histologic Grading and Staging of Chronic Hepatitis

	HISTOLOGIC ACTIVITY INDEX (HAI)[a]		METAVIR[b]	
HISTOLOGIC FEATURE	**SEVERITY**	**SCORE**	**SEVERITY**	**SCORE**
Necroinflammatory Activity (grade)				
Periportal necrosis, including piecemeal necrosis and/or bridging necrosis (BN)	None Mild Mild/moderate Moderate Severe	0 1 2 3 4	None Mild Moderate Severe Bridging necrosis	0 1 2 3 Yes No
Intralobular necrosis Confluent	—None —Focal —Zone 3 some —Zone 3 most —Zone 3 + BN few —Zone 3 + BN multiple —Panacinar/multiacinar	0 1 2 3 4 5 6	None or mild Moderate Severe	0 1 2
Focal	—None —≤1 focus/10× field —2–4 foci/10× field —5–10 foci/10× field —>10 foci/10× field	0 1 2 3 4		
Portal Inflammation	None Mild Moderate Moderate/marked Marked Total	0 1 2 3 4 0–18	A0–A3[c]	
Fibrosis (stage)				

	HISTOLOGIC ACTIVITY INDEX (HAI)[a]		METAVIR[b]	
HISTOLOGIC FEATURE	**SEVERITY**	**SCORE**	**SEVERITY**	**SCORE**
Necroinflammatory Activity (grade)				
None	Total	0	F0	
		1	F1	
		2	F1	
		3	F2	
		4	F3	
		5	F4	
		6	F4	
Portal fibrosis—some		6	4	
Portal fibrosis—most				
Bridging fibrosis—few				
Bridging fibrosis—many				
Incomplete cirrhosis				
Cirrhosis				

[a]Ishak K, Baptista A, Bianchi L, et al: Histologic grading and staging of chronic hepatitis. J Hepatol 22:696, 1995. [b]Bedossa P, Poynard T, French METAVIR Cooperative Study Group: An algorithm for grading activity in chronic hepatitis C. Hepatology 24:289, 1996. [c]Necroinflammatory grade: A0 = none; A1 = mild; A2 = moderate; A3 = severe.

CLASSIFICATION BY STAGE

The stage of chronic hepatitis, which reflects the level of progression of the disease, is based on the degree of hepatic fibrosis. When fibrosis is so extensive that fibrous septa surround parenchymal nodules and alter the normal architecture of the liver lobule, the histologic lesion is defined as *cirrhosis*. Staging is based on the degree of fibrosis as categorized on a numerical scale 0–6 (HAI) or 0–4 (METAVIR) (Table 334-2). Several noninvasive approaches have been introduced to provide approximations of hepatic histologic stage, including serum biomarkers of fibrosis and imaging determinations of liver elasticity.

CHRONIC VIRAL HEPATITIS

Both the enterically transmitted forms of viral hepatitis, hepatitis A and E, are self-limited and do not cause chronic hepatitis (rare reports notwithstanding in which acute hepatitis A serves as a trigger for the onset of autoimmune hepatitis in genetically susceptible patients or in which hepatitis E **(Chap. 332)** can cause chronic liver disease in immunosuppressed hosts, for example, after liver transplantation). In contrast, the entire

clinicopathologic spectrum of chronic hepatitis occurs in patients with chronic viral hepatitis B and C as well as in patients with chronic hepatitis D superimposed on chronic hepatitis B.

CHRONIC HEPATITIS B

The likelihood of chronicity after acute hepatitis B varies as a function of age. Infection at birth is associated with clinically silent acute infection but a 90% chance of chronic infection, whereas infection in young adulthood in immunocompetent persons is typically associated with clinically apparent acute hepatitis but a risk of chronicity of only ~1%. Most cases of chronic hepatitis B among adults, however, occur in patients who never had a recognized episode of clinically apparent acute viral hepatitis. The degree of liver injury (grade) in patients with chronic hepatitis B is variable, ranging from none in inactive carriers to mild to moderate to severe. Among adults with chronic hepatitis B, histologic features are of prognostic importance. In one long-term study of patients with chronic hepatitis B, investigators found a 5-year survival rate of 97% for patients with mild chronic hepatitis, 86% for patients with moderate to severe chronic hepatitis, and only 55% for patients with chronic hepatitis and postnecrotic cirrhosis. The 15-year survival in these cohorts was 77%, 66%, and 40%, respectively. On the other hand, more recent observations do not allow us to be so sanguine about the prognosis in patients with mild chronic hepatitis; among such patients followed for 1–13 years, progression to more severe chronic hepatitis and cirrhosis has been observed in more than a quarter of cases.

More important to consider than histology alone in patients with chronic hepatitis B is the degree of hepatitis B virus (HBV) replication. As reviewed in **Chap. 332,** chronic HBV infection can occur in the presence or absence of serum hepatitis B e antigen (HBeAg), and generally, for both HBeAg-reactive and HBeAg-negative chronic hepatitis B, the level of HBV DNA correlates with the level of liver injury and risk of progression. In *HBeAg-reactive chronic hepatitis B*, two phases have been recognized based on the relative level of HBV replication. The relatively *replicative phase* is characterized by the presence in the serum of HBeAg and HBV DNA levels well in excess of 10^3–10^4 IU/mL, sometimes exceeding 10^9 IU/mL; by the presence in the liver of detectable intrahepatocyte nucleocapsid antigens (primarily hepatitis B core antigen [HBcAg]); by high infectivity; and by accompanying liver injury. In contrast, the relatively *nonreplicative phase* is characterized by the absence of the conventional serum marker of HBV replication (HBeAg), the appearance of anti-HBe, levels of HBV DNA below a threshold of ~10^3 IU/mL, the absence of intrahepatocytic HBcAg, limited infectivity, and minimal liver injury. Patients in the relatively replicative phase tend to have more severe chronic hepatitis, whereas those in the relatively nonreplicative phase tend to have minimal or mild chronic hepatitis or to be inactive hepatitis B carriers. The likelihood in a patient with HBeAg-reactive chronic hepatitis B of converting spontaneously from relatively replicative to nonreplicative infection is ~10% per year. Distinctions in HBV replication and in histologic category, however, do not always coincide. In patients with HBeAg-reactive chronic HBV infection, especially when acquired at birth or in early childhood, as recognized commonly in Asian countries, a dichotomy is common between very high levels of HBV replication during the early decades of life (when the level of apparent host immunologic tolerance of HBV is relatively high) and negligible levels of liver injury. Yet despite the relatively immediate, apparently benign nature of liver disease for many decades in this population, in the middle decades, activation of liver injury emerges as what appears to be the relative tolerance of the host to HBV declines, and these patients with childhood-acquired HBV infection are ultimately

at increased risk later in life of cirrhosis, hepatocellular carcinoma (HCC) **(Chap. 78)**, and liver-related death. **A discussion of the pathogenesis of liver injury in patients with chronic hepatitis B appears in Chap. 332.**

HBeAg-negative chronic hepatitis B (i.e., chronic HBV infection with active virus replication, readily detectable HBV DNA but without HBeAg [anti-HBe-reactive]), is more common than HBeAg-reactive chronic hepatitis B in Mediterranean and European countries and in Asia (and, correspondingly, in HBV genotypes other than A). Compared to patients with HBeAg-reactive chronic hepatitis B, patients with HBeAg-negative chronic hepatitis B have HBV DNA levels several orders of magnitude lower (no more than 10^5–10^6 IU/mL) than those observed in the HBeAg-reactive subset. Most such cases represent precore or core-promoter mutations acquired late in the natural history of the disease (mostly early-life onset; age range 40–55 years, older than that for HBeAg-reactive chronic hepatitis B); these mutations prevent translation of HBeAg from the precore component of the HBV genome (precore mutants) or are characterized by downregulated transcription of precore mRNA (core-promoter mutants; **Chap. 332**). Although their levels of HBV DNA tend to be lower than among patients with HBeAg-reactive chronic hepatitis B, patients with HBeAg-negative chronic hepatitis B can have progressive liver injury (complicated by cirrhosis and HCC) and experience episodic reactivation of liver disease reflected in fluctuating levels of aminotransferase activity ("flares"). The biochemical and histologic activity of HBeAg-negative disease tends to correlate closely with levels of HBV replication, unlike the case mentioned above of Asian patients with HBeAg-reactive chronic hepatitis B during the early decades of their HBV infection. Worth reiterating, the level of HBV replication is the most important risk factor for the ultimate development of cirrhosis and HCC in both HBeAg-reactive (beyond the early decades of "relatively nonreplicative" infection) and HBeAg-negative patients. Although levels of HBV DNA are lower and more readily suppressed by therapy to undetectable levels in HBeAg-negative (compared to HBeAg-reactive) chronic hepatitis B, achieving sustained responses that permit discontinuation of antiviral therapy is less likely in HBeAg-negative patients (see below). Inactive carriers are patients with circulating hepatitis B surface antigen (HBsAg), normal serum aminotransferase levels, undetectable HBeAg, and levels of HBV DNA that are either undetectable or present at a threshold of $\leq 10^3$ IU/mL. This serologic profile occurs not only in inactive carriers but also in patients with HBeAg-negative chronic hepatitis B during periods of relative inactivity; distinguishing between the two requires sequential biochemical and virologic monitoring over many months.

The spectrum of *clinical features* of chronic hepatitis B is broad, ranging from asymptomatic infection to debilitating disease or even end-stage, fatal hepatic failure. As noted above, the onset of the disease tends to be insidious in most patients, with the exception of the very few in whom chronic disease follows failure of resolution of clinically apparent acute hepatitis B. **The clinical and laboratory features associated with progression from acute to chronic hepatitis B are discussed in Chap. 332.**

Fatigue is a common symptom, and persistent or intermittent *jaundice* is a common feature in severe or advanced cases. Intermittent deepening of jaundice and recurrence of malaise and anorexia, as well as worsening fatigue, are reminiscent of acute hepatitis; such exacerbations may occur spontaneously, often coinciding with evidence of virologic reactivation; may lead to progressive liver injury; and, when superimposed on well-established cirrhosis, may cause hepatic decompensation. Complications of cirrhosis occur in end-stage chronic hepatitis and include ascites, edema, bleeding gastroesophageal varices, hepatic

encephalopathy, coagulopathy, and hypersplenism. Occasionally, these complications bring the patient to initial clinical attention. Extrahepatic complications of chronic hepatitis B, similar to those seen during the prodromal phase of acute hepatitis B, are associated with tissue deposition of circulating hepatitis B antigen–antibody immune complexes. These include arthralgias and arthritis, which are common, and the more rare purpuric cutaneous lesions (leukocytoclastic vasculitis), immune-complex glomerulonephritis, and generalized vasculitis (polyarteritis nodosa) **(Chaps. 332 and 356)**.

Laboratory features of chronic hepatitis B do not distinguish adequately between histologically mild and severe hepatitis. Aminotransferase elevations tend to be modest for chronic hepatitis B but may fluctuate in the range of 100−1000 units. As is true for acute viral hepatitis B, alanine aminotransferase (ALT) tends to be more elevated than aspartate aminotransferase (AST); however, once cirrhosis is established, AST tends to exceed ALT. Levels of alkaline phosphatase activity tend to be normal or only marginally elevated. In severe cases, moderate elevations in serum bilirubin (51.3−171 µmol/L [3−10 mg/dL]) occur. Hypoalbuminemia and prolongation of the prothrombin time occur in severe or end-stage cases. Hyperglobulinemia and detectable circulating autoantibodies are distinctly absent in chronic hepatitis B (in contrast to autoimmune hepatitis). **Viral markers of chronic HBV infection are discussed in Chap. 332**.

TREATMENT

TREATMENT: Chronic Hepatitis B

Although progression to cirrhosis is more likely in severe than in mild or moderate chronic hepatitis B, all forms of chronic hepatitis B can be progressive, and progression occurs primarily in patients with active HBV replication. Moreover, in populations of patients with chronic hepatitis B who are at risk for HCC **(Chap. 78)**, the risk is highest for those with continued, high-level HBV replication and lower for persons in whom initially high-level HBV DNA falls spontaneously over time. Therefore, management of chronic hepatitis B is directed at suppressing the level of virus replication. Although clinical trials tend to focus on clinical endpoints achieved over 1−2 years (e.g., suppression of HBV DNA to undetectable levels, loss of HBeAg/HBsAg, improvement in histology, normalization of ALT), these short-term gains translate into reductions in the risk of clinical progression, hepatic decompensation, HCC, liver transplantation, and death; regression of cirrhosis and of esophageal varices have been documented to follow long-term pharmacologic suppression of HBV replication. In addition, restoration of impaired HBV-specific T-cell function has been shown following successful suppression of HBV replication with antiviral therapy. To date, seven drugs have been approved for treatment of chronic hepatitis B: injectable interferon (IFN) α and pegylated interferon (long-acting IFN bound to polyethylene glycol, PEG [PEG IFN]) and the oral agents lamivudine, adefovir dipivoxil, entecavir, telbivudine, and tenofovir disoproxil fumarate (TDF).

Antiviral therapy for hepatitis B has evolved rapidly since the mid-1990s, as has the sensitivity of tests for HBV DNA. When IFN and lamivudine were evaluated in clinical trials, HBV DNA was measured by insensitive hybridization assays with detection thresholds of 10^5–10^6 virions/mL; when adefovir, entecavir, telbivudine, tenofovir, and PEG IFN were studied in clinical trials, HBV DNA was measured by sensitive amplification assays (polymerase chain reaction [PCR]) with detection thresholds of 10^1–10^3 viral copies/mL or IU/mL. Recognition

of these distinctions is helpful when comparing results of clinical trials that established the efficacy of these therapies (reviewed below in chronological order of publication of these efficacy trials).

INTERFERON

IFN-α was the first approved therapy (1992) for chronic hepatitis B. Although it is no longer used to treat hepatitis B, standard IFN is important historically, having provided important lessons about antiviral therapy in general. For immunocompetent adults with HBeAg-reactive chronic hepatitis B (who tend to have high-level HBV DNA [$>10^5$–10^6 virions/mL] and histologic evidence of chronic hepatitis on liver biopsy), a 16-week course of IFN given subcutaneously at a daily dose of 5 million units, or three times a week at a dose of 10 million units, resulted in a loss of HBeAg and hybridization-detectable HBV DNA (i.e., a reduction to levels below 10^5–10^6 virions/mL) in ~30% of patients, with a concomitant improvement in liver histology. Seroconversion from HBeAg to anti-HBe occurred in ~20%, and, in early trials, ~8% lost HBsAg. Successful IFN therapy and seroconversion were often accompanied by an acute hepatitis-like elevation in aminotransferase activity, postulated to result from enhanced cytolytic T cell clearance of HBV-infected hepatocytes. Relapse after successful therapy was rare (1 or 2%). The likelihood of responding to IFN was higher in patients with lower levels of HBV DNA and substantial elevations of ALT. Although children can respond as well as adults, IFN therapy was not effective in very young children infected at birth. Similarly, IFN therapy was not effective in immunosuppressed persons, Asian patients with neonatal acquisition of infection and minimal-to-mild ALT elevations, or patients with decompensated chronic hepatitis B (in whom such therapy was actually detrimental, sometimes precipitating decompensation, often associated with severe adverse effects). Among patients with HBeAg loss during therapy, long-term follow-up demonstrated that 80% experienced eventual loss of HBsAg (i.e., all serologic markers of infection, and normalization of ALT over a 9-year posttreatment period). In addition, improved long-term and complication-free survival as well as a reduction in the frequency of HCC were documented among IFN responders, supporting the conclusion that successful antiviral therapy improves the natural history of chronic hepatitis B.

Initial trials of brief-duration IFN therapy in patients with *HBeAg-negative chronic hepatitis B* were disappointing, suppressing HBV replication transiently during therapy but almost never resulting in sustained antiviral responses. In subsequent IFN trials among patients with HBeAg-negative chronic hepatitis B, however, more protracted courses, lasting up to 1.5 years, were reported to result in sustained remissions documented to last for several years, with suppressed HBV DNA and aminotransferase activity, in ~20%.

Complications of IFN therapy include systemic “flu-like” symptoms; marrow suppression; emotional lability (irritability, depression, anxiety); autoimmune reactions (especially autoimmune thyroiditis); and miscellaneous side effects such as alopecia, rashes, diarrhea, and numbness and tingling of the extremities. With the possible exception of autoimmune thyroiditis, all these side effects are reversible upon dose lowering or cessation of therapy.

Although no longer competitive with the newer generation of antivirals, IFN did represent the first successful antiviral approach and set a standard against which to measure subsequent drugs in the achievement of durable virologic, serologic, biochemical, and histologic responses; consolidation of virologic and biochemical benefit in the ensuing years after therapy; and improvement in the natural history of chronic hepatitis B.

Standard IFN has been supplanted by long-acting PEG IFN (see below), and IFN nonresponders are now treated with one of the newer oral nucleoside analogues.

LAMIVUDINE

The first of the nucleoside analogues to be approved (in 1998) for hepatitis B, the dideoxynucleoside lamivudine inhibits reverse transcriptase activity of both HIV and HBV and is an effective agent for patients with chronic hepatitis B. Although generally superseded by newer, more potent, less resistance-prone agents, lamivudine is still used in regions of the world where newer agents are not yet available or affordable. In clinical trials among patients with HBeAg-reactive chronic hepatitis B, lamivudine therapy at daily doses of 100 mg for 48–52 weeks suppressed HBV DNA by a median of ~5.5 $\log_{10}$ copies/mL and to undetectable levels, as measured by PCR amplification assays, in ~40% of patients. Therapy was associated with HBeAg loss in 32–33%, HBeAg seroconversion (i.e., conversion from HBeAg-reactive to anti-HBe-reactive) in 16–21%, normalization of ALT in 40–75%, improvement in histology in 50–60%, retardation in hepatic fibrosis in 20–30%, and prevention of progression to cirrhosis. HBeAg responses occur even in patients resistant to IFN (e.g., those with high-level HBV DNA) or who failed in the past to respond to it. As is true for IFN therapy of chronic hepatitis B, patients with near-normal ALT activity tend not to experience HBeAg responses (despite suppression of HBV DNA), whereas those with ALT levels exceeding 5 × the upper limit of normal can expect 1-year HBeAg seroconversion rates of 50–60%. Generally, HBeAg seroconversions are confined to patients who achieve suppression of HBV DNA to $<10^4$ copies/mL (equivalent to ~10^3 IU/mL). Lamivudine-associated HBeAg responses are accompanied by a delayed posttreatment HBsAg seroconversion rate comparable to that seen after IFN-induced HBeAg responses. Among Western patients who undergo HBeAg responses during a year-long course of therapy and in whom the response is sustained for 4–6 months after cessation of therapy, the response is durable thereafter in the vast majority (>80%); therefore, the achievement of an HBeAg response represents a viable stopping point in therapy. Reduced durability has been reported in Asian patients; therefore, to support the durability of HBeAg responses, patients should receive a period of consolidation therapy of ≥6 months in Western patients and ≥1 year in Asian patients after HBeAg seroconversion (see treatment guidelines below). Close posttreatment monitoring is necessary to identify HBV reactivation promptly and to resume therapy. If HBeAg is unaffected by lamivudine therapy, the current approach is to continue therapy until an HBeAg response occurs, but long-term therapy may be required to suppress HBV replication and, in turn, limit liver injury; HBeAg seroconversions can increase to a level of 50% after 5 years of therapy. Histologic improvement continues to accrue with therapy beyond the first year; after a cumulative course of 3 years of lamivudine therapy, necroinflammatory activity is reduced in the majority of patients, and even cirrhosis has been shown to regress to precirrhotic stages in as many as three-quarters of patients.

Losses of HBsAg have been few during the first year of lamivudine therapy, and this observation had been cited as an advantage of IFN-based over lamivudine therapy; however, in head-to-head comparisons between standard IFN and lamivudine monotherapy, HBsAg losses were rare in both groups. Trials in which lamivudine and IFN were administered in combination failed to show a benefit of combination therapy over lamivudine monotherapy for either treatment-naïve patients or prior IFN nonresponders.

In patients with *HBeAg-negative chronic hepatitis B* (i.e., in those with precore and core-promoter HBV mutations), 1 year of lamivudine therapy results in HBV DNA suppression and normalization of ALT in three-

quarters of patients and in histologic improvement in approximately two-thirds. Therapy has been shown to suppress HBV DNA by ~4.5 $\log_{10}$ copies/mL (baseline HBV DNA levels are lower than in patients with HBeAg-reactive hepatitis B) and to undetectable levels in ~70%, as measured by sensitive PCR amplification assays. Lacking HBeAg at the outset, patients with HBeAg-negative chronic hepatitis B cannot achieve an HBeAg response—a stopping point in HBeAg-reactive patients; almost invariably, when therapy is discontinued, reactivation is the rule. Therefore, these patients require long-term therapy; with successive years, the proportion with suppressed HBV DNA and normal ALT increases.

Clinical and laboratory side effects of lamivudine are negligible and indistinguishable from those observed in placebo recipients. Still, lamivudine doses should be reduced in patients with reduced creatinine clearance. During lamivudine therapy, transient ALT elevations, resembling those seen during IFN therapy and during spontaneous HBeAg-to-anti-HBe seroconversions, occur in one-fourth of patients. These ALT elevations may result from restored cytolytic T cell activation permitted by suppression of HBV replication. Similar ALT elevations, however, occurred at an identical frequency in placebo recipients; however, ALT elevations associated with HBeAg seroconversion in clinical trials were confined to lamivudine-treated patients. When therapy is stopped after a year of therapy, two- to threefold ALT elevations occur in 20–30% of lamivudine-treated patients, representing renewed liver-cell injury as HBV replication returns. Although these posttreatment flares are almost always transient and mild, rare severe exacerbations, especially in cirrhotic patients, have been observed, mandating close and careful clinical and virologic monitoring after discontinuation of treatment. Many authorities caution against discontinuing therapy in patients with cirrhosis, in whom posttreatment flares could precipitate decompensation.

Long-term monotherapy with lamivudine is associated with methionine-to-valine (M204V) or methionine-to-isoleucine (M204I) mutations, primarily at amino acid 204 in the tyrosine-methionine-aspartate-aspartate (YMDD) motif of the C domain of HBV DNA polymerase, analogous to mutations that occur in HIV-infected patients treated with this drug. During a year of therapy, YMDD mutations occur in 15–30% of patients; the frequency increases with each year of therapy, reaching 70% at year 5. Ultimately, patients with YMDD mutants experience degradation of clinical, biochemical, and histologic responses; therefore, if treatment is begun with lamivudine monotherapy, the emergence of lamivudine resistance, reflected clinically by a breakthrough from suppressed levels of HBV DNA and ALT, is managed by adding another antiviral to which YMDD variants are sensitive (e.g., adefovir, tenofovir; see below).

Currently, although lamivudine is very safe and still used widely in other parts of the world, in the United States and Europe, lamivudine has been eclipsed by more potent antivirals that have superior resistance profiles (see below); it is no longer recommended as first-line therapy. Still, as the first successful oral antiviral agent for use in hepatitis B, lamivudine provided proof of principle that polymerase inhibitors can achieve virologic, serologic, biochemical, and histologic benefits. In addition, lamivudine has been shown to be effective in the treatment of patients with decompensated hepatitis B (for whom IFN is contraindicated), in some of whom decompensation can be reversed. Moreover, among patients with cirrhosis or advanced fibrosis, lamivudine has been shown to be effective in reducing the risk of progression to hepatic decompensation and, marginally, the risk of HCC. In the half decade following the introduction in the United States of lamivudine therapy for hepatitis B, referral of patients with HBV-associated end-stage liver disease for liver transplantation was

reduced by ~30%, supporting further the beneficial impact of oral antiviral therapy on the natural history of chronic hepatitis B.

Because lamivudine monotherapy can result universally in the rapid emergence of YMDD variants in persons with HIV infection, patients with chronic hepatitis B should be tested for anti-HIV prior to therapy; if HIV infection is identified, lamivudine monotherapy at the HBV daily dose of 100 mg is contraindicated. These patients should be treated for both HIV and HBV with an HIV drug regimen that includes or is supplemented by at least two drugs active against HBV; antiretroviral therapy (ART) often contains two drugs with antiviral activity against HBV (e.g., tenofovir and emtricitabine), but if lamivudine is part of the regimen, the daily dose should be 300 mg **(Chap. 197)**. The safety of lamivudine during pregnancy has not been established; however, the drug is not teratogenic in rodents and has been used safely in pregnant women with HIV infection and with HBV infection. Administration of lamivudine during the last months of pregnancy to mothers with high-level hepatitis B viremia ($\geq 10^8$ IU/mL) can reduce the likelihood of perinatal transmission of hepatitis B.

ADEFOVIR DIPIVOXIL

At an oral daily dose of 10 mg, the acyclic nucleotide analogue adefovir dipivoxil, the prodrug of adefovir (approved for hepatitis B in 2002), reduces HBV DNA by ~3.5–4 $\log_{10}$ copies/mL and is equally effective in treatment-naïve patients and prior IFN nonresponders. In HBeAg-reactive chronic hepatitis B, a 48-week course of adefovir dipivoxil was shown to achieve histologic improvement (and reduce the progression of fibrosis) and normalization of ALT in just over one-half of patients, HBeAg seroconversion in 12%, HBeAg loss in 23%, and suppression to an undetectable level of HBV DNA in 13–21%, as measured by PCR. Similar to IFN and lamivudine, adefovir dipivoxil is more likely to achieve an HBeAg response in patients with high baseline ALT; among adefovir-treated patients with ALT level >5 × the upper limit of normal, HBeAg seroconversions occurred in 25%. The durability of adefovir-induced HBeAg responses is high (91% in one study); therefore, HBeAg response can be relied upon as a stopping point for adefovir therapy, after a period of consolidation therapy, as outlined above. Although data on the impact of additional therapy beyond 1 year are limited, biochemical, serologic, and virologic outcomes improve progressively as therapy is continued.

In patients with *HBeAg-negative chronic hepatitis B*, a 48-week course of 10 mg/d of adefovir dipivoxil resulted in histologic improvement in two-thirds, normalization of ALT in three-fourths, and suppression of HBV DNA to PCR-undetectable levels in one-half to two-thirds. As was true for lamivudine, because HBeAg responses—a potential stopping point—cannot be achieved in this group, reactivation is the rule when adefovir therapy is discontinued, and indefinite, long-term therapy is required. Treatment beyond the first year consolidates the gain of the first year; after 5 years of therapy, improvement in hepatic inflammation and regression of fibrosis were observed in three-fourths of patients, ALT was normal in 70%, and HBV DNA was undetectable in almost 70%. In one study, stopping adefovir after 5 years was followed by sustained suppression of HBV DNA and ALT, but most HBeAg-negative patients are treated indefinitely unless HBsAg loss, albeit very rare, is achieved.

Adefovir contains a flexible acyclic linker instead of the L-nucleoside ring of lamivudine, avoiding steric hindrance by mutated amino acids. In addition, the molecular structure of phosphorylated adefovir is very similar to that of its natural substrate; therefore, mutations to adefovir would also affect binding of the natural substrate, dATP. Hypothetically, these are among the reasons that resistance to adefovir dipivoxil is much less

likely than resistance to lamivudine; no resistance was encountered in 1 year of clinical trial therapy. In subsequent years, however, adefovir resistance begins to emerge (asparagine to threonine at amino acid 236 [N236T] and alanine to valine or threonine at amino acid 181 [A181V/T], primarily), occurring in 2.5% after 2 years, but in 29% after 5 years of therapy (reported in HBeAg-negative patients). Among patients co-infected with HBV and HIV and who have normal CD4+ T cell counts, adefovir dipivoxil is effective in suppressing HBV dramatically (by 5 $\log_{10}$ in one study). Moreover, adefovir dipivoxil is effective in lamivudine-resistant, YMDD-mutant HBV and can be used when such lamivudine-induced variants emerge. When lamivudine resistance occurs, adding adefovir (i.e., maintaining lamivudine to preempt the emergence of adefovir resistance) is superior to switching to adefovir. Almost invariably, patients with adefovir-induced HBV mutations respond to lamivudine (or newer agents, such as entecavir, see below). When, in the past, adefovir had been evaluated as therapy for HIV infection, doses of 60–120 mg were required to suppress HIV, and, at these doses, the drug was nephrotoxic. Even at 30 mg/d, creatinine elevations of 44 µmol/L (0.5 mg/dL) occurred in 10% of patients; however, at the HBV-effective dose of 10 mg, such creatinine elevations are rarely encountered. If any nephrotoxicity does occur, it rarely appears before 6–8 months of therapy. Although renal tubular injury is a rare potential side effect, and although creatinine monitoring is recommended during treatment, the therapeutic index of adefovir dipivoxil is high, and the nephrotoxicity observed in clinical trials at higher doses was reversible. For patients with underlying renal disease, frequency of administration of adefovir dipivoxil should be reduced to every 48 h for creatinine clearances of 30–49 mL/min; to every 72 h for creatinine clearances of 10–29 mL/min; and to once a week, following dialysis, for patients undergoing hemodialysis. Adefovir dipivoxil is very well tolerated, and ALT elevations during and after withdrawal of therapy are similar to those observed and described above in clinical trials of lamivudine. An advantage of adefovir is its relatively favorable resistance profile; however, it is not as potent as the other approved oral agents, it does not suppress HBV DNA as rapidly or as uniformly as the others, it is the least likely of all agents to result in HBeAg seroconversion, and 20–50% of patients fail to suppress HBV DNA by 2 $\log_{10}$ (“primary nonresponders”). For these reasons, adefovir, which has been supplanted in both treatment-naïve and lamivudine-resistant patients by the more potent, less resistance-prone nucleotide analogue tenofovir (see below), is no longer recommended as first-line therapy.

PEGYLATED IFN

After long-acting PEG IFN was shown to be effective in the treatment of hepatitis C (see below), this more convenient drug was evaluated in the treatment of chronic hepatitis B. Once-a-week PEG IFN is more effective than the more frequently administered, standard IFN, and several large-scale trials of PEG IFN versus oral nucleoside analogues were conducted among patients with HBeAg-reactive and HBeAg-negative chronic hepatitis B.

In HBeAg-reactive chronic hepatitis B, two large-scale studies were done. In one study, PEG IFN-α 2b (100 µg weekly for 32 weeks, then 50 µg weekly for another 20 weeks for a total of 52 weeks) was evaluated against a comparison arm of combination PEG IFN with oral lamivudine in 307 subjects. The other study involved PEG IFN-α 2a (180 µg weekly for 48 weeks) in 814 primarily Asian patients, three-fourths of whom had ALT ≥2 × the upper limit of normal, with comparison arms of lamivudine monotherapy and combination PEG IFN plus lamivudine. At the end of therapy (48–52 weeks) in the PEG IFN monotherapy arms, HBeAg loss occurred in

~30%, HBeAg seroconversion in 22−27%, undetectable HBV DNA (<400 copies/mL by PCR) in 10−25%, normal ALT in 34−39%, and a mean reduction in HBV DNA of 2 $\log_{10}$ copies/mL (PEG IFN-α 2b) to 4.5 $\log_{10}$ copies/mL (PEG IFN-α 2a). Six months after completing PEG IFN monotherapy in these trials, HBeAg losses were present in ~35%, HBeAg seroconversion in ~30%, undetectable HBV DNA in 7−14%, normal ALT in 32−41%, and a mean reduction in HBV DNA of 2−2.4 $\log_{10}$ copies/mL. Although the combination of PEG IFN and lamivudine was superior at the end of therapy in one or more serologic, virologic, or biochemical outcomes, neither the combination arm (in both studies) nor the lamivudine monotherapy arm (in the PEG IFN-α 2a trial) demonstrated any benefit compared to the PEG IFN monotherapy arms 6 months after therapy. Moreover, HBsAg seroconversion occurred in 3−7% of PEG IFN recipients (with or without lamivudine); some of these seroconversions were identified by the end of therapy, but many were identified during the posttreatment follow-up period. The likelihood of HBeAg loss in PEG IFN-treated HBeAg-reactive patients is associated with HBV genotype A > B > C > D (shown for PEG IFN-α2b but not for α-2a). PEG IFN-α 2a was approved in the US for hepatitis B in 2005; PEG IFN-α 2b, not approved for hepatitis B in the US, is used in other countries.

Based on these results, some authorities concluded that PEG IFN monotherapy should be the first-line therapy of choice in HBeAg-reactive chronic hepatitis B; however, this conclusion has been challenged. Although a finite, 1-year course of PEG IFN results in a higher rate of sustained response (6 months after treatment) than is achieved with oral nucleoside/nucleotide analogue therapy, the comparison is confounded by the fact that oral agents are not discontinued at the end of 1 year. Instead, taken orally and free of side effects, therapy with oral agents is extended indefinitely or until after the occurrence of an HBeAg response. The rate of HBeAg responses after 2 years of oral-agent nucleoside analogue therapy is at least as high as, if not higher than, that achieved with PEG IFN after 1 year; favoring oral agents is the absence of injections, difficult-to-tolerate side effects, and laboratory monitoring as well as lower direct and indirect medical care costs and inconvenience. The association of HBsAg responses with PEG IFN therapy occurs in such a small proportion of patients that subjecting everyone to PEG IFN for the marginal gain of HBsAg responses during or immediately after therapy in such a very small minority is questionable. Moreover, HBsAg responses occur in a comparable proportion of patients treated with early-generation nucleoside/nucleotide analogues in the years *after* therapy, and, with the newer, more potent nucleoside analogues, the frequency of HBsAg loss during the first year of therapy equals that of PEG IFN and is exceeded during year 2 and beyond (see below). Of course, resistance is not an issue during PEG IFN therapy, but the risk of resistance is much lower with new agents (≤1% up to 3−8 years in previously treatment-naïve, entecavir-treated and 0% of tenofovir-treated patients; see below). Finally, the level of HBV DNA inhibition that can be achieved with the newer agents, and even with lamivudine, exceeds that which can be achieved with PEG IFN, in some cases by several orders of magnitude.

In HBeAg-negative chronic hepatitis B, a trial of PEG IFN-α 2a (180 µg weekly for 48 weeks versus comparison arms of lamivudine monotherapy and of combination therapy) in 564 patients showed that PEG IFN monotherapy resulted at the end of therapy in suppression of HBV DNA by a mean of 4.1 $\log_{10}$ copies/mL, undetectable HBV DNA (<400 copies/mL by PCR) in 63%, normal ALT in 38%, and loss of HBsAg in 4%. Although lamivudine monotherapy and combination lamivudine−PEG IFN therapy were both superior to PEG IFN at the end of therapy, no advantage of lamivudine monotherapy or combination therapy was apparent over PEG IFN monotherapy 6 months after therapy—suppression of HBV DNA by a mean of 2.3 $\log_{10}$ copies/mL, undetectable

HBV DNA in 19%, and normal ALT in 59%. In subjects involved in this trial followed for up to 5 years, among the two-thirds followed who had been treated initially with PEG IFN, 17% maintained HBV DNA suppression to <400 copies/mL, but ALT remained normal in only 22%; HBsAg loss increased gradually to 12%. Among the half followed who had been treated initially with lamivudine monotherapy, HBV DNA remained <400 copies/mL in 7% and ALT normal in 16%; by year 5, 3.5% had lost HBsAg. As was the case for standard IFN therapy in HBeAg-negative patients, only a small proportion maintained responsiveness after completion of PEG IFN therapy, raising questions about the relative value of a finite period of PEG IFN, versus a longer course with a potent, low-resistance oral nucleoside analogue in these patients. Moreover, the value of PEG IFN for HBeAg-negative chronic hepatitis B has not been confirmed. In the only other controlled clinical trial of PEG IFN for HBeAg-negative chronic hepatitis B, the hepatitis C regimen of PEG IFN plus ribavirin was compared to PEG IFN monotherapy. In this trial, HBV DNA suppression (<400 copies/mL) occurred in only 7.5% of the two groups combined, and no study subject lost HBsAg.

In patients treated with PEG IFN, HBeAg and HBsAg responses have been associated with *IL28B* genotype CC, the favorable genotype identified in trials of PEG IFN for chronic hepatitis C. Also, reductions in quantitative HBsAg levels have been shown to correlate with and to be predictive of responsiveness to PEG IFN in chronic hepatitis B. If HBsAg levels fail to fall within the first 12–24 weeks or to reach <20,000 IU/mL by week 24, PEG IFN therapy is unlikely to be effective and should be discontinued. (Similar observations of HBsAg levels in oral-agent-treated patients are of interest, but of limited clinical relevance, given the very high likelihood of virologic responses during such therapy.)

ENTECAVIR

Entecavir, an oral cyclopentyl guanosine analogue polymerase inhibitor (approved 2005), appears to be the most potent of the HBV antivirals and is just as well tolerated as lamivudine. In a 709-subject clinical trial among HBeAg-reactive patients, oral entecavir, 0.5 mg daily, was compared to lamivudine, 100 mg daily. At 48 weeks, entecavir was superior to lamivudine in suppression of HBV DNA (mean 6.9 vs 5.5 $\log_{10}$ copies/mL), percentage with undetectable HBV DNA (<300 copies/mL by PCR; 67% vs 36%), histologic improvement (≥2-point improvement in necroinflammatory HAI score; 72% vs 62%), and normal ALT (68% vs 60%). The two treatments were indistinguishable in percentage with HBeAg loss (22% vs 20%) and seroconversion (21% vs 18%). Among patients treated with entecavir for 96 weeks, HBV DNA was undetectable cumulatively in 80% (vs 39% for lamivudine), and HBeAg seroconversions had occurred in 31% (vs 26% for lamivudine). After 3–6 years of entecavir, HBeAg seroconversions have been observed in 39–44% and HBsAg loss in 5–6%. Similarly, in a 638-subject clinical trial among HBeAg-negative patients, at week 48, oral entecavir, 0.5 mg daily, was superior to lamivudine, 100 mg daily, in suppression of HBV DNA (mean 5.0 vs 4.5 $\log_{10}$ copies/mL) and in percentage with undetectable HBV DNA (90% vs 72%), histologic improvement (70% vs 61%), and normal ALT (78% vs 71%). No resistance mutations were encountered in previously treatment-naïve, entecavir-treated patients during 96 weeks of therapy, and in a cohort of subjects treated for up to 6 years, resistance emerged in only 1.2%. Entecavir-induced HBeAg seroconversions are as durable as those achieved with other antivirals. Its high barrier to resistance coupled with its high potency renders entecavir a first-line drug for patients with chronic hepatitis B.

Entecavir is also effective against lamivudine-resistant HBV infection. In a trial of 286 lamivudine-resistant patients, entecavir, at a higher daily dose of 1 mg, was superior to lamivudine, as measured at week 48, in achieving suppression of HBV DNA (mean 5.1 vs 0.48 $\log_{10}$ copies/mL), undetectable HBV DNA (72% vs 19%), normal ALT (61% versus 15%), HBeAg loss (10% vs 3%), and HBeAg seroconversion (8% vs 3%). In this population of lamivudine-experienced patients, however, entecavir resistance emerged in 7% at 48 weeks. Although entecavir resistance requires both a YMDD mutation and a second mutation at one of several other sites (e.g., T184A, S202G/I, or M250V), resistance to entecavir in lamivudine-resistant chronic hepatitis B has been recorded to increase progressively to 43% at 4 years and 57% at 6 years; therefore, entecavir is not as attractive a choice (and is not recommended, despite its approval for this indication) as adefovir or tenofovir for patients with lamivudine-resistant hepatitis B.

In clinical trials, entecavir had an excellent safety profile. In addition, on-treatment and posttreatment ALT flares are relatively uncommon and relatively mild in entecavir-treated patients. Doses should be reduced for patients with reduced creatinine clearance. Entecavir does have low-level antiviral activity against HIV and cannot be used as monotherapy to treat HBV infection in HIV/HBV co-infected persons.

TELBIVUDINE

Telbivudine, a cytosine analogue (approved 2006), is similar in efficacy to entecavir but slightly less potent in suppressing HBV DNA (a slightly less profound median 6.4 $\log_{10}$ reduction in HBeAg-reactive disease and a similar 5.2 $\log_{10}$ reduction in HBeAg-negative disease). In its registration trial, telbivudine at an oral daily dose of 600 mg suppressed HBV DNA to <300 copies/mL in 60% of HBeAg-positive and 88% of HBeAg-negative patients, reduced ALT to normal in 77% of HBeAg-positive and 74% of HBeAg-negative patients, and improved histology in 65% of HBeAg-positive and 67% of HBeAg-negative patients. Although resistance to telbivudine (M204I, not M204V, mutations) was less frequent than resistance to lamivudine at the end of 1 year, resistance mutations after 2 years of treatment occurred in up to 22%. Generally well tolerated, telbivudine has been associated with a low frequency of asymptomatic creatine kinase elevations and with a very low frequency of peripheral neuropathy; frequency of administration should be reduced for patients with impaired creatinine clearance. Its excellent potency notwithstanding, the inferior resistance and safety profile of telbivudine has limited its appeal; telbivudine is neither recommended as first-line therapy nor widely used.

TENOFOVIR

TDF, an acyclic nucleotide analogue and potent antiretroviral agent used to treat HIV infection (approved for hepatitis B in 2008), is similar to adefovir but more potent in suppressing HBV DNA and inducing HBeAg responses; it is highly active against both wild-type and lamivudine-resistant HBV and active in patients whose response to adefovir is slow and/or limited. At an oral once-daily dose of 300 mg for 48 weeks, tenofovir suppressed HBV DNA by 6.2 $\log_{10}$ (to undetectable levels [<400 copies/mL] in 76%) in HBeAg-positive patients and by 4.6 $\log_{10}$ (to undetectable levels in 93%) in HBeAg-negative patients; reduced ALT to normal in 68% of HBeAg-positive and 76% of HBeAg-negative patients; and improved histology in 74% of HBeAg-positive and 72% of HBeAg-negative patients. In HBeAg-positive patients, HBeAg seroconversions occurred in 21% by the end of year 1, 27% by year 2, 34% by year 3, and 40% by year 5 of tenofovir treatment; HBsAg loss occurred in 3% by the end of year 1 and 6% at year 2, and 8% by year 5. After 5 years of tenofovir therapy, 87% of patients

experienced histologic improvement, including reduction in fibrosis score (51%) and regression of cirrhosis (71%). The 5-year safety (negligible renal toxicity, in 1%, and mild reduction in bone density, in ~0.5%) and resistance profiles (none recorded through 8 years) of tenofovir are very favorable as well; therefore, tenofovir has supplanted adefovir both as first-line therapy for chronic hepatitis B and as add-on therapy for lamivudine-resistant chronic hepatitis B. Studies of tenofovir and entecavir reviewed in 2015 showed no difference in long-term risks of renal and bone toxicity; however, among patients treated with tenofovir, instances of acute renal failure and of low blood phosphate levels have been reported. Thus, in patients receiving tenofovir, monitoring bone density is not recommended, but periodic (at least annual) monitoring for renal injury is (serum creatinine and phosphate, urine glucose and protein). Frequency of tenofovir administration should be reduced for patients with impaired creatinine clearance.

A comparison of the six antiviral therapies in current use appears in Table 334-3; their relative potencies in suppressing HBV DNA are shown in Fig. 334-1.

COMBINATION THERAPY

Although the combination of lamivudine and PEG IFN suppresses HBV DNA more profoundly during therapy than does monotherapy with either drug alone (and is much less likely to be associated with lamivudine resistance), this combination used for a year is no better than a year of PEG IFN in achieving sustained responses. To date, combinations of oral nucleoside/nucleotide agents have not achieved an enhancement in virologic, serologic, or biochemical efficacy over that achieved by the more potent of the combined drugs given individually. In a 2-year trial of combination entecavir and tenofovir versus entecavir monotherapy, for a small subgroup of patients with very high HBV DNA levels ($\geq 10^8$ IU/mL), a reduction in HBV DNA to <50 IU/mL was higher in the combination group (79% vs 62%); however, no differences in HBeAg responses or any other endpoint were observed between the combination-therapy and monotherapy groups, even in the high-HBV DNA subgroup. On the other hand, combining agents that are not cross-resistant (e.g., lamivudine or entecavir with adefovir or tenofovir) has the potential to reduce the risk or perhaps even to preempt entirely the emergence of drug resistance. In the future, the treatment paradigm may shift from the current approach of sequential monotherapy to preemptive combination therapy, perhaps not for all patients but for subsets (e.g., patients with very high levels of HBV DNA, immunosuppressed patients); however, designing and executing clinical trials that demonstrate superior efficacy and resistance profile of combination therapy over monotherapy with entecavir or tenofovir will remain challenging. Whereas, initially, in clinical studies of adefovir as rescue therapy for lamividune resistance, *adding* adefovir to lamivudine (combination therapy) was considered a better strategy than replacing lamivudine with adefovir monotherapy, according to the 2016 treatment recommendations of the American Association for the Study of Liver Diseases (AASLD), data to support adding or switching agents are insufficient. Therefore, while sound virologic principles would favor adding as opposed to switching, according to current recommendations involving the more potent first-line agents, entecavir for tenofovir resistance and tenofovir for entecavir resistance, either strategy is acceptable. For patients who already have acquired multidrug resistance (to both nucleoside analogues [lamivudine, entecavir, telbivudine] and nucleotide analogues [adefovir, tenofovir]), treatment with a combination of entecavir and tenofovir has been shown to be highly effective in suppressive HBV DNA and overcoming drug resistance.

NOVEL ANTIVIRALS AND STRATEGIES

In addition to the seven approved antiviral drugs for hepatitis B, emtricitabine, a fluorinated cytosine analogue very similar to lamivudine in structure, efficacy, and resistance profile, offers no advantage over lamivudine. A combination of emtricitabine and tenofovir is approved for the treatment of HIV infection and is an appealing combination therapy for hepatitis B, especially for lamivudine-resistant disease; however, neither emtricitabine nor the combination is approved for hepatitis B. Several initially promising antiviral agents have been abandoned because of toxicity (e.g., clevudine, which was linked to myopathy during its clinical development). As noted above, the current formulation of tenofovir, TDF, has been associated with renal toxicity and loss of bone density, especially in patients with HIV infection, less so in patients with HBV infection. A new formulation, tenofovir alafenamide (TAF), is a prodrug of tenofovir that is metabolized to the active agent in its target organ (the liver for HBV infection); such targeting permits higher dose delivery to the liver with markedly reduced systemic exposure. Studies in patients with chronic hepatitis B treated with 25 mg of TAF or 300 mg of TDF demonstrate comparable virologic efficacy as well as less reduction in bone mineral density and estimated glomerular filtration rate for TAF. Based on its better renal and bone safety profile than TDF, TAF has been approved for HBV infection and provides an alternative to TDF in patients with TDF-associated elevations in serum creatinine and/or reductions in serum phosphorus. Direct-acting antivirals (DAAs) have been very successful in the management of chronic hepatitis B; however, most patients require long-duration, usually indefinite, therapy. Ideally, an approach to achieving “cure” (eradication of HBV infection) with finite-duration therapy would be welcome. Currently, innovative approaches being investigated include viral entry inhibitors, nucleocapsid assembly inhibitors, HBV secretion (HBsAg release) inhibitors, immunomodulators (e.g., toll receptor agonists, T-cell vaccines, programmed cell death [PD-1] blockade, reconstitution of innate immune responses [e.g., retinoic acid-inducible gene-1, RIG-1, HBV mRNA recognition, and activation of innate immune signaling) and of adaptive immune responses, covalently closed circular (ccc) DNA silencing/inhibition/cleavage, RNA interference, and HBx inhibitors. While data supporting several of these unconventional approaches have begun to appear, none has been shown to “cure” hepatitis B, and none is likely to be competitive, unless it can be shown to go beyond current antivirals in achieving recovery (HBsAg seroconversion) from HBV infection. Finally, initial emphasis in the development of antiviral therapy for hepatitis B was placed on monotherapy; whether combination regimens will yield additive or synergistic efficacy remains to be determined.

TREATMENT RECOMMENDATIONS

Several learned societies and groups of expert physicians have issued treatment recommendations for patients with chronic hepatitis B; the most authoritative and updated (and free of financial support by pharmaceutical companies) are those of the AASLD and of the European Association for the Study of the Liver (EASL). Although the recommendations differ slightly, a consensus has emerged on most of the important points **(Table 334-4)**. No treatment is recommended or available for inactive “nonreplicative” hepatitis B carriers (undetectable HBeAg with normal ALT and HBV DNA $\leq 10^3$ IU/mL documented serially over time). In patients with detectable HBeAg and HBV DNA levels $>2 \times 10^4$ IU/mL, treatment is recommended by the AASLD for those with ALT levels above 2 × the upper limit of normal. (The EASL recommends treatment in HBeAg-positive patients for HBV DNA levels $>2 \times 10^3$ IU/mL and ALT above the upper limit of normal.) For HBeAg-positive patients with ALT ≤2 × the upper limit of normal, in whom sustained responses are not likely and who would require multiyear therapy,

antiviral therapy is not recommended currently. This pattern is common during the early decades of life among Asian patients infected at birth; even in this group, therapy would be considered for those >40 years of age, ALT persistently at the high end of the twofold range, and/or with a family history of HCC, especially if the liver biopsy shows moderate to severe necroinflammatory activity or fibrosis. In this group, when, eventually, ALT becomes elevated later in life, antiviral therapy should be instituted. For patients with HBeAg-negative chronic hepatitis B, ALT >2 × the upper limit of normal (above the upper limit of normal according to EASL), and HBV DNA >2 × 10^3 IU/mL, antiviral therapy is recommended. If HBV DNA is >2 × 10^3 IU/mL and ALT is 1 to >2 × the upper limit of normal, liver biopsy should be considered to help in arriving at a decision to treat if substantial liver injury is present (treatment in this subset would be recommended according to EASL guidelines, because ALT is elevated). Per current AASLD recommendations, antiviral treatment with oral agents can be stopped after HBeAg seroconversion in noncirrhotics, and the suggested period of consolidation therapy is 12 months with close monitoring for recurrent viremia (monthly × 6, then every 3 months for the rest of a year) after cessation of therapy. For patients with HBeAg-negative chronic hepatitis, the current recommendation with oral agents is for indefinite therapy; although sufficient data are lacking, stopping therapy in this group can be considered after HBsAg loss.

For patients with compensated cirrhosis, because antiviral therapy has been shown to retard clinical progression, treatment is recommended regardless of HBeAg status and ALT as long as HBV DNA is detectable at >2 × 10^3 IU/mL (detectable at any level according to the EASL); monitoring without therapy is recommended for those with HBV DNA <2 × 10^3 IU/mL, unless ALT is elevated. For patients with decompensated cirrhosis, treatment is recommended regardless of serologic and biochemical status, as long as HBV DNA is detectable. Patients with decompensated cirrhosis should be evaluated as candidates for liver transplantation.

Among the seven available drugs for hepatitis B, PEG IFN has supplanted standard IFN, entecavir has supplanted lamivudine, and tenofovir has supplanted adefovir. PEG IFN, entecavir, or tenofovir is recommended as first-line therapy (Table 334-3). PEG IFN requires finite-duration therapy, achieves the highest rate of HBeAg responses after a year of therapy, and does not support viral mutations, but it requires subcutaneous injections and is associated with inconvenience, more intensive clinical and laboratory monitoring, and intolerability. Oral nucleoside analogues require long-term therapy in most patients, and when used alone, lamivudine and telbivudine foster the emergence of viral mutations, adefovir somewhat less so, and entecavir (except in lamivudine-experienced patients) and tenofovir rarely at all. Oral agents do not require injections or cumbersome laboratory monitoring, are very well tolerated, lead to improved histology in 50–90% of patients, suppress HBV DNA more profoundly than PEG IFN, and are effective even in patients who fail to respond to IFN-based therapy. Although oral agents are less likely to result in HBeAg responses during the first year of therapy, as compared to PEG IFN, treatment with oral agents tends to be extended beyond the first year and, by the end of the second year, yields HBeAg responses (and even HBsAg responses) comparable in frequency to those achieved after 1 year of PEG IFN (and without the associated side effects) (**Table 334-5**). In a 2016 systematic review of 1716 patients involved in 25 clinical trials, responses after oral-agent therapy were found to be durable. Among patients with HBeAg-reactive chronic hepatitis B, the pooled rates of durable HBeAg seroconversions maintained after cessation of nucleoside/nucleotide analogue therapy (including all the oral agents) were 92% and 88% at posttreatment months 12 and 24, respectively, unaffected by the

duration of post-HBeAg-response consolidation therapy (>6 months in all studies evaluated); the pooled rate of durable biochemical remission after therapy in this population was 76%. Even for HBeAg-negative chronic hepatitis B, for which most authorities recommend indefinite therapy, pooled rates of virologic remissions maintained after cessation of oral-agent therapy were 44%, 31%, and 30% at posttreatment months 12, 24, and 36, and the pooled rate of durable biochemical remission in this population was 57%.

Although adefovir and tenofovir are safe, renal monitoring (e.g., serum creatinine and phosphate, urine glucose and protein) is recommended. Substantial experience with lamivudine during pregnancy (see above) has identified no teratogenicity; although widely used during pregnancy, lamivudine remains classified as pregnancy category C. Although IFNs do not appear to cause congenital anomalies, these have antiproliferative properties and should be avoided during pregnancy. Adefovir during pregnancy has not been associated with birth defects; however, the risk of spontaneous abortion may be increased, and adefovir is categorized as pregnancy category C. Data on the safety of entecavir during pregnancy have not been published (pregnancy category C). Sufficient data in animals and limited data in humans suggest that telbivudine and tenofovir (both pregnancy category B) can be used safely during pregnancy; however, telbivudine is not an acceptable first-line drug. In general, then, except for lamivudine and tenofovir, and until additional data become available, the other antivirals for hepatitis B should be avoided or used with extreme caution during pregnancy.

For children aged 2 to <18 with HBeAg-reactive hepatitis B (most children will be HBeAg-reactive; no studies have been done in children with HBeAg-negative chronic hepatitis B), treatment is recommended if HBV DNA is detectable and ALT levels are elevated, but not if ALT levels are normal. Each of the available drugs, except telbivudine, is approved for different childhood age groups (standard IFN α-2b age ≥1 year; PEG IFN α-2a age ≥5 years [approved for hepatitis C, not B, but can be used in hepatitis B]; lamivudine and entecavir age ≥2 years; adefovir and tenofovir age ≥12 years). Package inserts should be consulted for childhood doses.

As noted above, some physicians prefer to begin with PEG IFN, while other physicians and patients prefer oral agents as first-line therapy. For patients with decompensated cirrhosis, the emergence of resistance can result in further deterioration and loss of antiviral effectiveness. Therefore, in this patient subset, the threshold for relying on therapy with a very favorable resistance profile (e.g., entecavir or tenofovir) or on combination therapy is low. PEG IFN should not be used in patients with compensated or decompensated cirrhosis.

For patients with end-stage chronic hepatitis B who undergo liver transplantation, reinfection of the new liver is almost universal in the absence of antiviral therapy. The majority of patients become high-level viremic carriers with minimal liver injury. Before the availability of antiviral therapy, an unpredictable proportion experienced severe hepatitis B–related liver injury, sometimes a fulminant-like hepatitis and sometimes a rapid recapitulation of the original severe chronic hepatitis B (**Chap. 332**). Currently, however, prevention of recurrent hepatitis B after liver transplantation has been achieved definitively by *combining* hepatitis B immune globulin with one of the low-resistance oral nucleoside (entecavir) or nucleotide analogues (tenofovir) (**Chap. 338**); preliminary data suggest that the newer, more potent, and less resistance-prone oral agents may be used instead of hepatitis B immune globulin for posttransplantation therapy. In patients documented at the time of liver transplantation to have undetectable HBV DNA in serum and cccDNA in the liver (i.e., with low risk for recurrence of HBV infection), a preliminary clinical trial suggested that, after patients received 5 years of

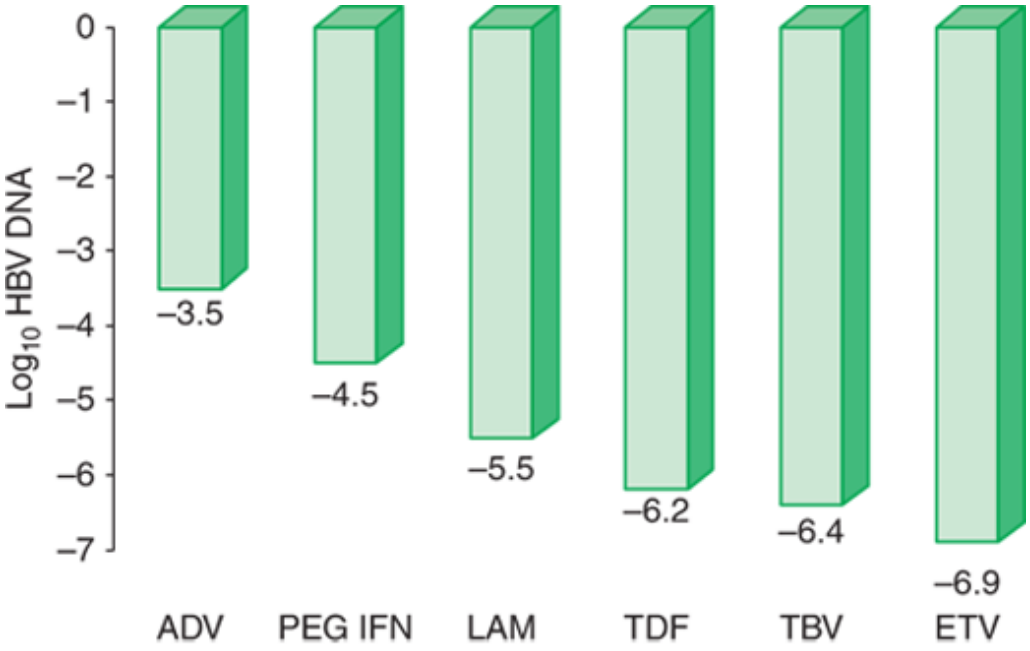
combined therapy, both hepatitis B immune globulin and oral-agent therapy can be withdrawn sequentially (over two 6-month periods) with a success rate, as monitored over a median of 6 years postwithdrawal, of 90% and an anti-HBs seroconversion rate of 60% (some with transient reappearance of HBV DNA and/or HBsAg).

Patients with HBV-HIV co-infection can have progressive HBV-associated liver disease and, occasionally, a severe exacerbation of hepatitis B resulting from immunologic reconstitution following ART. Lamivudine should never be used as monotherapy in patients with HBV-HIV infection because HIV resistance emerges rapidly to both viruses. Adefovir has been used successfully to treat chronic hepatitis B in HBV-HIV co-infected patients but is no longer considered a first-line agent for HBV. Entecavir has low-level activity against HIV and can result in selection of HIV resistance; therefore, it should be avoided in HBV-HIV co-infection. Tenofovir and the combination of tenofovir and emtricitabine in one pill are approved therapies for HIV and represent excellent choices for treating HBV infection in HBV-HIV co-infected patients. Generally, even for HBV-HIV co-infected patients who do not yet meet treatment criteria for HIV infection, treating for both HBV and HIV is recommended.

Patients with chronic hepatitis B who undergo cytotoxic chemotherapy for treatment of malignancies as well as patients treated with immunosuppressive, anticytokine, or antitumor necrosis factor therapies (the risk varies, from highest [e.g., B-cell-depleting agents, anthracycline derivatives, moderate/high-dose corticosteroids for ≥4 weeks] to moderate [e.g., tumor necrosis factor alpha inhibitors, cytokine or integrin inhibitors, tyrosine kinase inhibitors, low-dose corticosteroids for ≥4 weeks], to lowest [e.g., immunosuppressive agents like methotrexate and azathioprine, intraarticular corticosteroids, any dose of corticosteroids for ≤1 week]) experience enhanced HBV replication and viral expression on hepatocyte membranes during chemotherapy coupled with suppression of cellular immunity. When chemotherapy is withdrawn, such patients are at risk for reactivation of hepatitis B, often severe and occasionally fatal. Such rebound reactivation represents restoration of cytolytic T cell function against a target organ enriched in HBV expression. Preemptive treatment with the first of the oral HBV antivirals, lamivudine, prior to the initiation of chemotherapy was shown to reduce the risk of such reactivation substantially; treating *after* reactivation has occurred is less effective. The newer, more potent oral antiviral agents, entecavir and tenofovir, which are even more effective in preventing hepatitis B reactivation and with a lower risk of antiviral drug resistance, are preferred. The optimal duration of antiviral therapy after completion of chemotherapy is not known, but a suggested approach is 6 months (12 months for B-cell-depleting agents) for inactive hepatitis B carriers and longer-duration therapy in patients with baseline HBV DNA levels >2 × 10^3 IU/mL, until standard clinical endpoints are met (Table 334-4). Such chemotherapy-associated reactivation of hepatitis B is common (4–68%, median 25%, in a meta-analysis) in persons with ongoing HBV infection (HBsAg-reactive); however, such reactivation can occur albeit less commonly in persons who have cleared HBsAg, but express anti-HBc (moderate risk, <10%) and rarely (<5%) even in persons with serologic evidence of recovery from HBV infection (anti-HBs-reactive, anti-HBc-reactive). Therefore, most authorities (e.g., Centers for Disease Control and Prevention; AASLD; American Gastroenterological Association; EASL) recommend HBsAg and anti-HBc (± anti-HBs) screening of all patients undergoing such chemotherapy and preemptive antiviral prophylaxis for HBsAg-reactive persons and close on-therapy monitoring of anti-HBc-reactive/anti-HBs-reactive persons with treatment if and when reactivation occurs.

FIGURE 334-1

Relative potency of antiviral drugs for hepatitis B, as reflected by median log_{10} HBV DNA reduction in HBeAg-positive chronic hepatitis B. These data are from individual reports of large, randomized controlled registration trials that were the basis for approval of the drugs. In most instances, these data do not represent direct comparisons among the drugs, because study populations were different, baseline patient variables were not always uniform, and the sensitivity and dynamic range of the HBV DNA assays used in the trials varied. ADV, adefovir dipivoxil; ETV, entecavir; LAM, lamivudine; PEG IFN, pegylated interferon α2a; TBV, telbivudine; TDF, tenofovir disoproxil fumarate.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 334-3

Comparison of Pegylated Interferon (PEG IFN), Lamivudine, Adefovir, Entecavir, Telbivudine, and Tenofovir Therapy for Chronic Hepatitis B[a]

FEATURE	PEG IFN[b]	LAMIVUDINE	ADEFOVIR	ENTECAVIR	TELBIVUDINE	TENOFOVIR
Route of administration	Subcutaneous injection	Oral	Oral	Oral	Oral	Oral
Duration of therapy[c]	48–52 weeks	≥52 weeks	≥48 weeks	≥48 weeks	≥52 weeks	≥48 weeks
Tolerability	Poorly tolerated	Well tolerated	Well tolerated; creatinine monitoring recommended	Well tolerated	Well tolerated	Well tolerated; creatinine monitoring recommended
HBeAg seroconversion 1 yr Rx >1 yr Rx	18–20% NA	16–21% up to 50% @ 5 yrs	12% 43% @ 3 yrs[d]	21% 31% @ 2 yrs 44% @ 6 yrs	22% 30% @ 2 yrs	21% 40% @ 5 yrs
Log_{10} HBV DNA reduction (mean copies/mL) HBeAg-reactive HBeAg-negative	4.5 4.1	5.5 4.4–4.7	median 3.5–5 median 3.5–3.9	6.9 5.0	6.4 5.2	6.2 4.6

FEATURE	PEG IFN[b]	LAMIVUDINE	ADEFOVIR	ENTECAVIR	TELBIVUDINE	TENOFOVIR
HBV DNA PCR negative (<300–400 copies/mL; <1000 copies/mL for adefovir) at end of yr 1 HBeAg-reactive HBeAg-negative	10–25% 63%	36–44% 60–73%	13–21% 48–77%	67% (91% @ 4 yrs) 90%	60% 88%	76% 93%
ALT normalization at end of yr 1 HBeAg-reactive HBeAg-negative	39% 34–38%	41–75% 62–79%	48–61% 48–77%	68% 78%	77% 74%	68% 76%
HBsAg loss yr 1 >yr 1	3–4% 12% 5 yr after 1 yr of Rx	≤1% No data	0% 5% at yr 5	2% 6% at yr 6	<1% No data	3% 8% at yr 5
Histologic improvement (≥2 point reduction in HAI) at yr 1 HBeAg-reactive HBeAg-negative	38% 6 months after 48% 6 months after	49–62% 61–66%	53–68% 64%	72% 70%	65% 67%	74% 72%
Viral resistance	None	15–30% @ 1 yr 70% @ 5 yrs	None @ 1 yr 29% @ 5 yrs	≤1% @ 1 yr[e] 1.2% @ 6 yrs[e]	Up to 5% @ yr 1 Up to 22% @ yr 2	0% @ yr 1 0% through yr 8

FEATURE	PEG IFN[b]	LAMIVUDINE	ADEFOVIR	ENTECAVIR	TELBIVUDINE	TENOFOVIR
Pregnancy category	C	C[f]	C	C	B	B
Cost (US$) for 1 yr	~$18,000	~$2,500	~$6,500	~$8,700[g]	~$6,000	~$6,000

[a]Generally, these comparisons are based on data on each drug tested individually versus placebo in registration clinical trials; because, with rare exception, these comparisons are not based on head-to-head testing of these drugs, relative advantages and disadvantages should be interpreted cautiously. [b]Although standard interferon α administered daily or three times a week is approved as therapy for chronic hepatitis B, it has been supplanted by PEG IFN, which is administered once a week and is more effective. Standard interferon has no advantages over PEG IFN. [c]Duration of therapy in clinical efficacy trials; use in clinical practice may vary. [d]Because of a computer-generated randomization error that resulted in misallocation of drug versus placebo during the second year of clinical trial treatment, the frequency of HBeAg seroconversion beyond the first year is an estimate (Kaplan-Meier analysis) based on the small subset in whom adefovir was administered correctly. [e]7% during a year of therapy (43% at year 4) in lamivudine-resistant patients. [f]Despite its Class C designation, lamivudine has an extensive pregnancy safety record in women with HIV/AIDS. [g]Approximately $17,400 for lamivudine-refractory patients.

Abbreviations: ALT, alanine aminotransferase; HAI, histologic activity index; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; HBV, hepatitis B virus; NA, not applicable; PEG IFN, pegylated interferon; PCR, polymerase chain reaction; Rx, therapy; yr, year.

TABLE 334-4

Recommendations for Treatment of Chronic Hepatitis B[a]

HBeAg STATUS	CLINICAL	HBV DNA (IU/mL)	ALT	RECOMMENDATION
HBeAg-reactive	[b] Chronic hepatitis Cirrhosis compensated Cirrhosis decompensated	$>2 \times 10^4$ $>2 \times 10^4$[d] $>2 \times 10^3$ $<2 \times 10^3$ Detectable Undetectable	≤2 × ULN[c,d] >2 × ULN[d] < or > ULN >ULN < or > ULN < or > ULN	No treatment; monitor. In patients >40, with family history of hepatocellular carcinoma, and/or ALT persistently at the high end of the twofold range, liver biopsy may help in decision to treat Treat[e] Treat[e] with oral agents, not PEG IFN Consider treatment[f] Treat[e] with oral agents[g], not PEG IFN; refer for liver transplantation Observe; refer for liver transplantation
HBeAg-negative	[b] Chronic hepatitis Chronic hepatitis Cirrhosis compensated Cirrhosis decompensated	$\leq 2 \times 10^3$ $>10^3$ $>10^4$ $>2 \times 10^3$ $<2 \times 10^3$ Detectable Undetectable	≤ULN 1 to >2 × ULN[d] >2 × ULN[d] < or > ULN >ULN < or > ULN < or > ULN	Inactive carrier; treatment not necessary Consider liver biopsy; treat[h] if biopsy shows moderate to severe inflammation or fibrosis Treat[h,i] Treat[e] with oral agents, not PEG IFN Consider treatment[f] Treat[h] with oral agents[g], not PEG IFN; refer for liver transplantation Observe; refer for liver transplantation

[a]Based on practice guidelines of the American Association for the Study of Liver Diseases (AASLD). Except as indicated in footnotes, these guidelines are similar to those issued by the European Association for the Study of the Liver (EASL). [b]Liver disease tends to be mild or inactive clinically; most such patients do not undergo liver biopsy. [c]This pattern is common during early decades of life in Asian patients infected at birth. [d]According to the EASL guidelines, treat if HBV DNA is $>2 \times 10^3$ IU/mL and ALT >ULN. [e]One of the potent oral drugs with a high barrier to resistance (entecavir or tenofovir) or PEG IFN can be used as first-line therapy (see text). These oral agents, but not PEG IFN, should be used for

interferon-refractory/intolerant and immunocompromised patients. PEG IFN is administered weekly by subcutaneous injection for a year; the oral agents are administered daily for at least a year and continued indefinitely or until at least 6 months after HBeAg seroconversion. [f]According to EASL guidelines, patients with compensated cirrhosis and detectable HBV DNA at any level, even with normal ALT, are candidates for therapy. Most authorities would treat indefinitely, even in HBeAg-positive disease after HBeAg seroconversion. [g]Because the emergence of resistance can lead to loss of antiviral benefit and further deterioration in decompensated cirrhosis, a low-resistance regimen is recommended—entecavir or tenofovir monotherapy or combination therapy with the more resistance-prone lamivudine (or telbivudine) plus adefovir. Therapy should be instituted urgently. [h]Because HBeAg seroconversion is not an option, the goal of therapy is to suppress HBV DNA and maintain a normal ALT. PEG IFN is administered by subcutaneous injection weekly for a year; caution is warranted in relying on a 6-month posttreatment interval to define a sustained response, because the majority of such responses are lost thereafter. Oral agents, entecavir or tenofovir, are administered daily, usually indefinitely or until, as very rarely occurs, virologic and biochemical responses are accompanied by HBsAg seroconversion. [i]For older patients and those with advanced fibrosis, consider lowering the HBV DNA threshold to $>2 \times 10^3$ IU/mL.

Abbreviations: AASLD, American Association for the Study of Liver Diseases; ALT, alanine aminotransferase; EASL, European Association for the Study of the Liver; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; HBV, hepatitis B virus; PEG IFN, pegylated interferon; ULN, upper limit of normal.

TABLE 334-5

Pegylated Interferon Versus Oral Nucleoside Analogues for the Treatment of Chronic Hepatitis B

	PEG IFN	NUCLEOSIDE ANALOGUES
Administration	Weekly injection	Daily, orally
Tolerability	Poorly tolerated, intensive monitoring	Well tolerated, limited monitoring
Duration of therapy	Finite 48 weeks	≥1 year, indefinite in most patients
Maximum mean HBV DNA suppression	4.5 $\log_{10}$	6.9 $\log_{10}$
Effective in high-level HBV DNA (≥10^9 IU/mL)	No	Yes
HBeAg seroconversion During 1 year of therapy During >1 year of therapy	~30% Not applicable	~20% 30% (year 2) to up to 50% (year 5)
HBeAg-negative posttreatment HBV DNA suppression	17% @ 5 years	7% @ 4 years (lamivudine)
HBsAg loss During 1 year of therapy During >1 year of therapy After 1 year of therapy–HBeAg-negative	3–4% Not applicable 12% @ 5 years	0–3% 3–8% @ 5 years of therapy 3.5% @ 5 years
Antiviral resistance	None	Lamivudine: ~30% year 1, ~70% year 5 Adefovir: 0% year 1, ~30% year 5 Telbivudine: up to 4% year 1, 22% year 2 Entecavir: ≤1.2% through year 6 Tenofovir: 0% through year 8
Use in cirrhosis, transplantation, immunosuppressed	No	Yes

	PEG IFN	NUCLEOSIDE ANALOGUES
Cost, 1 year of therapy	++++	+ to ++

Abbreviations: HBV, hepatitis B virus; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; PEG IFN, pegylated interferon.

CHRONIC HEPATITIS D (DELTA HEPATITIS)

Chronic hepatitis D virus (HDV) may follow acute co-infection with HBV but at a rate no higher than the rate of chronicity of acute hepatitis B. That is, although HDV co-infection can increase the severity of acute hepatitis B, HDV does not increase the likelihood of progression to chronic hepatitis B. When, however, HDV superinfection occurs in a person who is already chronically infected with HBV, long-term HDV infection is the rule, and a worsening of the liver disease is the expected consequence. Except for severity, chronic hepatitis B plus D has similar clinical and laboratory features to those seen in chronic hepatitis B alone. Relatively severe and progressive chronic hepatitis, with or without cirrhosis, is the rule, and mild chronic hepatitis is the exception. Occasionally, however, mild hepatitis or even, rarely, inactive carriage occurs in patients with chronic hepatitis B plus D, and the disease may become indolent after several years of infection. A distinguishing serologic feature of chronic hepatitis D is the presence in the circulation of antibodies to liver-kidney microsomes (anti-LKM); however, the anti-LKM seen in hepatitis D, anti-LKM3, are directed against uridine diphosphate glucuronosyltransferase and are distinct from anti-LKM1 seen in patients with autoimmune hepatitis and in a subset of patients with chronic hepatitis C (see below). **The clinical and laboratory features of chronic HDV infection are summarized in Chap. 332.**

TREATMENT

TREATMENT: Chronic Hepatitis D

Management is not well defined, and the host cellular RNA polymerase upon which HDV replication depends cannot be targeted by conventional antiviral agents. Glucocorticoids are ineffective and are not used. Preliminary experimental trials of IFN-α suggested that conventional doses and durations of therapy lower levels of HDV RNA and aminotransferase activity only transiently during treatment but have no impact on the natural history of the disease. In contrast, high-dose IFN-α (9 million units three times a week) for 12 months was reported to be associated with a sustained loss of HDV replication and clinical improvement in up to 50% of patients. Moreover, in anecdotal reports, the beneficial impact of treatment has been observed to persist for 15 years and to be associated with a reduction in grade of hepatic necrosis and inflammation, reversion of advanced fibrosis (improved stage), and clearance of HDV RNA in some patients. A suggested approach to therapy has been high-dose, long-term IFN for at least a year and, in responders, extension of therapy until HDV RNA and HBsAg clearance; however, extension of therapy to a second year provided no advantage, and sustained responses after completion of therapy have been rare. PEG IFN has also been shown to be more

effective in the treatment of chronic hepatitis D (e.g., after 48 weeks of therapy, associated with undetectable HDV RNA, durable for at least 24 posttreatment weeks, in a quarter to a half of patients) and is a more convenient replacement for standard IFN; however, loss of virologic responses (reappearance of HDV RNA) was observed during long-term (median 4.5-year) monitoring in over half of initial, 24-week-posttreatment responders. Even extending PEG IFN therapy for 5 years and driving treatment doses up to 270 µg weekly (of PEG IFN- α2a), as reported in a small trial among 13 patients, while achieving serologic, virologic, histologic, biochemical, and clinical improvement, yielded sustained virologic responses (SVRs) in only 3 patients (58–246 weeks of posttreatment observation). None of the nucleoside analogue antiviral agents for hepatitis B is effective in hepatitis D, and adding oral nucleoside agents to PEG IFN is no more effective than PEG IFN monotherapy. While recommended, PEG IFN therapy is far from satisfactory. Preliminary trials have been performed with an oral prenylation inhibitor, lonafarnib, and with an inhibitor of HBV/HDV viral entry into hepatocytes, myrcludex B. Prenylation, the posttranslational covalent addition of the prenyl lipid farnesyl to large HDV antigen, is required for this HDV protein to interact and form secreted viral particles with HBsAg. In 14 patients treated twice daily for 28 days with 100 or 200 mg of lonafarnib, HDV RNA fell by 0.73 $\log_{10}$ IU/mL and 1.54 $\log_{10}$ IU/mL, respectively, before rebounding after completion of therapy. Hepatitis B virus entry into hepatocytes requires the binding of the myristolated N-terminal pre-S1 peptide of large HBsAg to sodium taurocholate co-transporting peptide, the functional receptor for HBV into hepatocytes. The application of myrcludex B, a synthetic homologous myristolated lipopeptide that competes for binding with HBsAg, was reported in a study of 24 patients (with a baseline mean of 4.1–4.2 $\log_{10}$ copies/mL of HDV RNA) randomized to 24 weeks of treatment with myrcludex B (2 mg daily subcutaneously) as monotherapy or combined with PEG IFN compared to PEG IFN alone. A reduction in HDV RNA occurred in all three groups, by 1.67 $\log_{10}$ copies/mL (in two of eight patients RNA became undetectable), 2.59 $\log_{10}$ copies/mL (in five of eight patients RNA became undetectable), and 2.17 $\log_{10}$ copies/mL (in two of eight patients RNA became undetectable), respectively. No change occurred, however, in the level of HBsAg, which would have been expected. In these two exploratory brief-duration trials, sustained responses were not achieved, and toxicities were encountered (e.g., intermittent vomiting and weight loss [lonafarnib] and transient amylase and lipase elevations [myrcludex B]); however, from these proof-of-principle trials, potentially, more definitive and larger-scale studies will follow.

In patients with end-stage liver disease secondary to chronic hepatitis D, liver transplantation has been effective. If hepatitis D recurs in the new liver without the expression of hepatitis B (an unusual serologic profile in immunocompetent persons but common in transplant patients), liver injury is limited. In fact, the outcome of transplantation for chronic hepatitis D is superior to that for chronic hepatitis B; in such patients, combination hepatitis B immune globulin and nucleoside analogue therapy for hepatitis B is indicated (Chap. 338).

CHRONIC HEPATITIS C

Regardless of the epidemiologic mode of acquisition of hepatitis C virus (HCV) infection, chronic hepatitis follows acute hepatitis C in 50−70% of cases; chronic infection is common even in those with a return to normal in aminotransferase levels after acute hepatitis C, adding up to an 85% likelihood of chronic HCV infection after

acute hepatitis C. Few clues had emerged to explain host differences associated with chronic infection until recently, when variation in a single nucleotide polymorphism (SNP) on chromosome 19, *IL28B* (which codes for IFN-λ3), was identified that distinguished between responders and nonresponders to IFN-based antiviral therapy (see below). The same variants correlated with spontaneous resolution after acute infection: 53% in genotype C/C, 30% in genotype C/T, but only 23% in genotype T/T. The association with HCV clearance after acute infection is even stronger when *IL28B* haplotype is combined with haplotype G/G of a SNP near human leukocyte antigen (HLA) Class II *DBQ1*03:01*.

In patients with chronic hepatitis C followed for 20 years, progression to cirrhosis occurs in about 20–25%. Such is the case even for patients with relatively clinically mild chronic hepatitis, including those without symptoms, with only modest elevations of aminotransferase activity, and with mild chronic hepatitis on liver biopsy. Even in cohorts of well compensated patients with chronic hepatitis C referred for clinical research trials (no complications of chronic liver disease and with normal hepatic synthetic function), the prevalence of cirrhosis may be as high as 50%. Most cases of hepatitis C are identified initially in asymptomatic patients who have no history of acute hepatitis C (e.g., those discovered while attempting to donate blood, while undergoing lab testing as part of an application for life insurance, or as a result of routine laboratory tests). The source of HCV infection in many of these cases is not defined, although a long-forgotten percutaneous exposure (e.g., injection drug use) in the remote past can be elicited in a substantial proportion and probably accounts for most infections; most of these infections were acquired in the 1960s and 1970s, coming to clinical attention decades later.

Approximately one-third of patients with chronic hepatitis C have normal or near-normal aminotransferase activity; although one-third to one-half of these patients have chronic hepatitis on liver biopsy, the grade of liver injury and stage of fibrosis tend to be mild in the vast majority. In some cases, more severe liver injury has been reported—even, rarely, cirrhosis, most likely the result of previous histologic activity. Among patients with persistent normal aminotransferase activity sustained over ≥5–10 years, histologic progression has been shown to be rare; however, approximately one-fourth of patients with normal aminotransferase activity experience subsequent aminotransferase elevations, and histologic injury can be progressive once abnormal biochemical activity resumes. Therefore, continued clinical monitoring and antiviral therapy are indicated, even for patients with normal aminotransferase activity.

Despite this substantial rate of progression of chronic hepatitis C, and despite the fact that liver failure can result from end-stage chronic hepatitis C, the long-term prognosis over 1–2 decades for chronic hepatitis C in a majority of patients is relatively benign. Mortality >10–20 years among patients with transfusion-associated chronic hepatitis C has been shown not to differ from mortality in a matched population of transfused patients in whom hepatitis C did not develop. Although death in the hepatitis group is more likely to result from liver failure, and although hepatic decompensation may occur in ~15% of such patients over the course of a decade, the majority (almost 60%) of patients remain asymptomatic and well compensated, with no clinical sequelae of chronic liver disease. Overall, chronic hepatitis C tends to be very slowly and insidiously progressive, if at all, in the vast majority of patients, whereas in approximately one-fourth of cases, chronic hepatitis C will progress eventually to end-stage cirrhosis. In fact, because HCV infection is so prevalent, and because a proportion of patients progress inexorably to end-stage liver disease, hepatitis C is the most frequent indication for liver

transplantation (Chap. 338). In the United States, hepatitis C accounts for up to 40% of all chronic liver disease; as of 2007, mortality caused by hepatitis C surpassed that associated with HIV/AIDS, and as of 2012, reported deaths caused by hepatitis C surpassed those associated with all other notifiable infectious diseases (HIV, tuberculosis, hepatitis B, and 57 other infectious diseases). Moreover, because the prevalence of HCV infection is so much higher in the "baby boomer" cohort born between 1945 and 1965, three-quarters of the mortality associated with hepatitis C occurs in this age cohort. Referral bias may account for the more severe outcomes described in cohorts of patients reported from tertiary care centers (20-year progression of ≥20%) versus the more benign outcomes in cohorts of patients monitored from initial blood-product-associated acute hepatitis or identified in community settings (20-year progression of only 4–7%). Still unexplained, however, are the wide ranges in reported progression to cirrhosis, from 2% over 17 years in a population of Irish women with hepatitis C infection acquired from contaminated anti-D immune globulin to 30% over ≤11 years in recipients of contaminated intravenous immune globulin.

Progression of liver disease in patients with chronic hepatitis C has been reported to be more likely in patients with older age, longer duration of infection, advanced histologic stage and grade, more complex HCV quasispecies diversity, increased hepatic iron, concomitant other liver disorders (alcoholic liver disease, chronic hepatitis B, hemochromatosis, α_1 antitrypsin deficiency, and steatohepatitis), HIV infection, and obesity. Among these variables, however, duration of infection appears to be one of the most important, and some of the others probably reflect disease duration to some extent (e.g., quasispecies diversity, hepatic iron accumulation). No other epidemiologic or clinical features of chronic hepatitis C (e.g., severity of acute hepatitis, level of aminotransferase activity, level of HCV RNA, presence or absence of jaundice during acute hepatitis) are predictive of eventual outcome. Despite the relatively benign nature of chronic hepatitis C over time in many patients, cirrhosis following chronic hepatitis C has been associated with the late development, after several decades, of HCC (Chap. 78); the annual rate of HCC in cirrhotic patients with hepatitis C is 1–4%, occurring primarily in patients who have had HCV infection for 30 years or more.

Perhaps the best prognostic indicator in chronic hepatitis C is liver histology; the rate of hepatic fibrosis may be slow, moderate, or rapid. Patients with mild necrosis and inflammation as well as those with limited fibrosis have an excellent prognosis and limited progression to cirrhosis. In contrast, among patients with moderate to severe necroinflammatory activity or fibrosis, including septal or bridging fibrosis, progression to cirrhosis is highly likely over the course of 10–20 years. The pace of fibrosis progression may be accelerated by such factors as concomitant HIV infection, other causes of liver disease, excessive alcohol use, and hepatic steatosis. Among patients with compensated cirrhosis associated with hepatitis C, the 10-year survival rate is close to 80%; mortality occurs at a rate of 2–6% per year; decompensation at a rate of 4–5% per year; and, as noted above, HCC at a rate of 1–4% per year. Estimates of the natural history of chronic hepatitis C have been made, based on data available on the prevalence of HCV infection in the US population and on the rate of disease progression. Weighted primarily by the concentration of chronic hepatitis C in the baby boomer generation, the peak prevalence was estimated to have occurred in 2015. The calculated frequency of cirrhosis in US patients with hepatitis C was 5% in 1990, 25% in 2010, and is projected to be 37% in 2020. Estimated peak mortality has been predicted to occur in 2032. **A discussion of the pathogenesis of liver injury in patients with chronic hepatitis C appears in Chap. 332.**

Clinical features of chronic hepatitis C are similar to those described above for chronic hepatitis B. Generally, fatigue is the most common symptom; jaundice is rare. Immune complex–mediated extrahepatic complications of chronic hepatitis C are less common than in chronic hepatitis B (despite the fact that assays for immune complexes are often positive in patients with chronic hepatitis C), with the exception of essential mixed cryoglobulinemia **(Chap. 332)**, which is linked to cutaneous vasculitis and membranoproliferative glomerulonephritis as well as lymphoproliferative disorders such as B-cell lymphoma and unexplained monoclonal gammopathy. In addition, chronic hepatitis C has been associated with extrahepatic complications unrelated to immune-complex injury. These include Sjögren's syndrome, lichen planus, porphyria cutanea tarda, type 2 diabetes mellitus, and the metabolic syndrome (including insulin resistance and steatohepatitis).

Laboratory features of chronic hepatitis C are similar to those in patients with chronic hepatitis B, but aminotransferase levels tend to fluctuate more (the characteristic episodic pattern of aminotransferase activity) and to be lower, especially in patients with long-standing disease. An interesting and occasionally confusing finding in patients with chronic hepatitis C is the presence of autoantibodies. Rarely, patients with autoimmune hepatitis (see below) and hyperglobulinemia have false-positive immunoassays for anti-HCV. On the other hand, some patients with serologically confirmable chronic hepatitis C have circulating anti-LKM. These antibodies are anti-LKM1, as seen in patients with autoimmune hepatitis type 2 (see below), and are directed against a 33-amino-acid sequence of cytochrome P450 IID6. The occurrence of anti-LKM1 in some patients with chronic hepatitis C may result from the partial sequence homology between the epitope recognized by anti-LKM1 and two segments of the HCV polyprotein. In addition, the presence of this autoantibody in some patients with chronic hepatitis C suggests that autoimmunity may be playing a role in the pathogenesis of chronic hepatitis C.

Histopathologic features of chronic hepatitis C, especially those that distinguish hepatitis C from hepatitis B, are described in Chap. 332.

TREATMENT

TREATMENT: Chronic Hepatitis C

Therapy for chronic hepatitis C has evolved substantially in the 25 years since IFN-α was introduced for this indication in 1991. The therapeutic armamentarium grew to include PEG IFN with ribavirin and, then, in 2011, the introduction of the first protease inhibitors, telaprevir and boceprevir, used in combination with PEG IFN and ribavirin in patients with HCV genotype 1. The field of antiviral therapy for hepatitis C was transformed beginning in 2013, with the approval of the first nucleoside analogue, sofosbuvir. As of 2016, no fewer than six, all-oral, highly effective (>95%), low-resistance, well tolerated, short-duration (usually 12 weeks) combination regimens of DAA drugs are available. The remarkable historical evolution of antiviral therapy for hepatitis C is instructive.

THE INTERFEON ERA (1991–2011)

IFN-based therapy has been supplanted by DAA agents introduced in the second decade of the twenty-first century; however, many important lessons about antiviral therapy for chronic hepatitis C were learned from

the experience with IFN-based treatment, and many of the limitations of—and disparities in responsiveness to—IFN-based therapy have been overcome by current-generation DAA treatments. When first approved, IFN-α was administered via subcutaneous injection three times a week for 6 months but achieved an SVR **(Fig. 334-2)** (defined then as a reduction of HCV RNA to undetectable levels by PCR when measured ≥24 weeks after completion of therapy) <10%. Doubling the duration of therapy—but not increasing the dose or changing IFN preparations—increased the SVR rate to ~20%, and addition to the regimen of daily ribavirin, an oral guanosine nucleoside, increased the SVR rate to 40%. When used alone, ribavirin is ineffective and does not reduce HCV RNA levels appreciably, but ribavirin enhances the efficacy of IFN by reducing the likelihood of virologic relapse after the achievement of an end-treatment response (Fig. 334-2) (response measured during, and maintained to the end of, treatment). Proposed mechanisms to explain the role of ribavirin include subtle direct reduction of HCV replication, inhibition of host inosine monophosphate dehydrogenase activity (and associated depletion of guanosine pools), immune modulation, induction of virologic mutational catastrophe, and enhancement of IFN-stimulated gene expression. Ribavirin, despite its poorly understood mechanism of action, retains a modest role in supporting DAA agents as well (see below). IFN therapy results in activation of the JAK-STAT signal transduction pathway, which culminates in the intracellular elaboration of genes and their protein products that have antiviral properties. Hepatitis C proteins inhibit JAK-STAT signaling at several steps along the pathway, and exogenous IFN restores expression of IFN-stimulated genes and their antiviral effects.

Treatment with the combination of PEG IFN and ribavirin increased responsiveness (frequency of SVR) to as high as 55% overall—to >40% in genotypes 1 and 4, and to >80% in genotypes 2 and 3. Even in the absence of biochemical and virologic responses, histologic improvement occurred in approximately three-fourths of all treated patients. In chronic hepatitis C, ALT levels fall precipitously during therapy, and up to 90% of virologic responses are achieved within the first 12 weeks of therapy; responses thereafter are rare. Most relapses occur within the first 12 weeks after treatment; therefore, an SVR at week 12 posttreatment (SVR_{12}) is roughly equivalent to a 24-week SVR, and SVR_{12} has become the new standard. SVRs are very durable; normal ALT, improved histology, and absence of HCV RNA in serum and liver have been documented a decade after successful therapy, and “relapses” 2 years after sustained responses are almost unheard of. Thus, an SVR to antiviral therapy of chronic hepatitis C is tantamount to a cure, which is followed by marked improvements in liver-disease outcomes (see below).

Patient variables that correlate with sustained virologic responsiveness to IFN-based therapy include favorable genotype (genotypes 2 and 3 as opposed to genotypes 1 and 4; genotype 1b as opposed to genotype 1a); low baseline HCV RNA level (<800,000 IU/mL), low HCV quasispecies diversity, and histologically mild hepatitis and minimal fibrosis, especially absence of cirrhosis; immunocompetence, low liver iron levels, age <40; female gender; and absence of obesity, insulin resistance, type 2 diabetes mellitus, and hepatic steatosis. High levels of HCV RNA, more histologically advanced liver disease, and high HCV quasispecies diversity all go hand in hand with advanced duration of infection and reduced IFN responsiveness. Also associated with poor responses to IFN-based therapy are African-American ethnicity (contributed to, but not explained entirely by, a higher proportion with genotype 1, slower early treatment viral kinetics, impaired HCV-specific immunity, and host genetic differences in *IL28B* alleles, described below), Latino ethnicity, and poor treatment adherence (<80% of IFN and ribavirin doses and <80% of prescribed duration of therapy). Ironically, patients whose

disease was least likely to progress were the ones *most* likely to respond to IFN and vice versa. For patients treated with combination IFN-ribavirin, therapy for those with genotype 1 usually required a full 48 weeks with SVRs in the range of 40–45%, whereas in those with genotypes 2 and 3, a 24-week course of therapy sufficed with SVRs in the range of 80% (although refined tailoring of treatment duration could be indicated based on rapidity of response or associated cofactors, see below).

Genetic changes in the virus may explain differences in treatment responsiveness in some patients (e.g., among patients with genotype 1b, responsiveness to IFN is enhanced in those with amino-acid-substitution mutations in the nonstructural protein 5A gene). As described above in the discussion of spontaneous recovery from acute hepatitis C, IFN gene variants discovered in genome-wide association studies were shown to have a substantial impact on responsiveness of patients with genotype 1 to antiviral therapy. In studies of patients treated with PEG IFN and ribavirin, variants of the *IL28B* SNP that code for IFN-λ3 (a type III IFN, the receptors for which are more discretely distributed than IFN-α receptors and more concentrated in hepatocytes) correlate significantly with responsiveness. Patients homozygous for the C allele at this locus have the highest frequency of achieving an SVR (~80%), those homozygous for the T allele at this locus are least likely to achieve an SVR (~25%), and those heterozygous at this locus (C/T) have an intermediate level of responsiveness (SVRs in ~35%).

Side effects of IFN therapy are described in the section on treatment of chronic hepatitis B. The most pronounced side effect of ribavirin therapy is hemolysis—an expected reduction in hemoglobin of up to 2−3 g or in hematocrit of 5−10% but also a small, unpredictable proportion with profound, brisk hemolysis, resulting in symptomatic anemia; therefore, close monitoring of blood counts is crucial, and ribavirin should be avoided in patients with anemia or hemoglobinopathies; in patients with coronary artery disease or cerebrovascular disease, in whom anemia can precipitate an ischemic event; in patients with renal insufficiency (the drug is excreted renally); and in pregnancy (the drug is teratogenic, mandating scrupulous use of efficient contraception during, and for several months after, therapy in women of child-bearing age [because of their antiproliferative properties, IFNs also are contraindicated during pregnancy]). When symptomatic anemia occurs, ribavirin dose reductions or addition of erythropoietin to boost red blood cell levels may be required; erythropoietin was shown to improve patients' quality of life but not the likelihood of achieving an SVR. If ribavirin was stopped during therapy, SVR rates fell, but responsiveness could be maintained as long as ribavirin was not stopped and the total ribavirin dose exceeded 60% of the planned dose.

Ribavirin can also cause nasal and chest congestion, pruritus, and precipitation of gout. Combination IFN-ribavirin therapy is more difficult to tolerate than IFN monotherapy and more likely to lead to dose reductions and discontinuation of therapy.

Studies of viral kinetics have shown that despite a virion half-life in serum of only 2−3 h, the level of HCV is maintained by a high replication rate of 10^{12} hepatitis C virions per day. IFN-α blocks virion production or release with an efficacy that increases with increasing drug doses; moreover, the calculated death rate for infected cells during IFN therapy is inversely related to the level of HCV RNA. Patients with the most rapid death rate of infected hepatocytes are more likely to achieve undetectable HCV RNA at 3 months; in practice, failure to achieve an early virologic response (EVR), a ≥2-$\log_{10}$ reduction in HCV RNA by week 12, predicts failure to

experience a subsequent SVR. Similarly, patients in whom HCV RNA becomes undetectable within 4 weeks (i.e., who achieve a rapid virologic response [RVR]) have a very high likelihood of achieving an SVR (Fig. 334-2). Surprisingly, however, high-dose induction with IFN-based therapy did not yield higher SVR rates.

For the treatment of chronic hepatitis C, standard IFNs were supplanted beginning in 2001 by PEG IFNs. These have elimination times up to sevenfold longer than standard IFNs (i.e., a substantially longer half-life) and achieve prolonged concentrations, permitting administration once (rather than three times) a week. Instead of the frequent drug peaks (linked to side effects) and troughs (when drug is absent) associated with frequent administration of short-acting IFNs, administration of PEG IFNs results in drug concentrations that are more stable and sustained over time. Once-a-week PEG IFN monotherapy is twice as effective as monotherapy with its standard IFN counterpart, approaches the efficacy of combination standard IFN plus ribavirin, and is as well tolerated as standard IFNs, without more difficult-to-manage thrombocytopenia and leukopenia than standard IFNs. For most of the decade prior to 2011, when protease inhibitors were introduced for HCV genotype 1 (see below), the standard of care was a combination of PEG IFN plus ribavirin for all HCV genotypes.

Two PEG IFNs are available: PEG IFN-α2b, a 12-kD, linear PEG molecule bound to IFN-α2b, and PEG IFN-α2a, a larger, 40-kD, branched PEG molecule bound to IFN-α2a; because of its larger size and smaller volume of extravascular distribution, PEG IFN-α2a can be given at a uniform dose independent of weight, whereas the dose of the smaller PEG IFN-α2b, which has a much wider volume distribution, must be weight-based. The standard dose of PEG IFN α2a was 180 μg and of PEG IFN-α2b 1.5 μg/kg. The ribavirin dose adopted for both PEG IFNs was, for genotype 1, 1000 mg (for patients <75 kg) to 1200 mg (for patients ≥75 kg) and, for genotypes 2 and 3, 800 mg; a broader ribavirin dose/weight range was approved subsequently for PEG IFN-α2b in patients with genotype 1: <65 kg, 800 mg; 65–85 kg, 1000 mg; >85–105 kg, 1200 mg; and >105 kg, 1400 mg. For both drugs, recommended treatment durations were 48 weeks for genotype 1 and 24 weeks for genotypes 2 and 3 (somewhat more refractory, justifying a full 48 weeks especially for advanced hepatic fibrosis or cirrhosis and/or high-level HCV RNA). Between the two PEG IFNs, PEG IFN-α2a appeared to be slightly better tolerated and slightly more effective than PEG IFN-α2b in registration trials (SVRs for genotype 1: 41–51% vs 40–42%, respectively) as well as in subsequent head-to-head trials and a systematic review of randomized trials (SVR in genotypes 1–4: 48–55% vs 32–40%, respectively).

Until the 2011 introduction of protease inhibitors, unless ribavirin was contraindicated (see above), combination PEG IFN plus ribavirin was the recommended course of therapy. Even after the introduction of protease inhibitors for genotypes 1 and 4, however, PEG IFN–ribavirin remained the standard of care for patients with genotypes 2 and 3 until late 2013. For patients treated with combination PEG IFN–ribavirin, measurement of quantitative HCV RNA levels at 12 weeks was helpful in guiding therapy; if a 2-$\log_{10}$ drop in HCV RNA had not been achieved by this time, chances for an SVR were negligible, and additional therapy was futile. If the 12-week HCV RNA had fallen by 2 $\log_{10}$ (EVR), the chances for an SVR at the end of therapy were approximately two-thirds; if the 12-week HCV RNA was undetectable (“complete” EVR), the chances for an SVR exceeded 80% (Fig. 334-2).

The frequency of an SVR to PEG IFN–ribavirin therapy could be increased by tailoring therapy according to baseline variables and on-treatment virologic responsiveness. In patients with baseline variables weighing

against a response (e.g., HCV RNA >800,000 IU/mL, weight >85 kg), by raising the dose of PEG IFN (e.g., to as high as 270 µg of PEG IFN-α2a) and/or the dose of ribavirin to as high as 1600 mg daily (if tolerated or supplemented by erythropoietin); or by extending therapy from 48 to 72 weeks for patients with genotype 1 and a slow virologic response (i.e., failure of HCV RNA to fall rapidly to undetectable levels within 4 weeks [absence of a RVR]), SVR rates could be improved somewhat. In contradistinction, in patients with genotype 1 (and 4) who had a 4-week RVR (which occurred in ≤20%), especially in the subset with low baseline HCV RNA, abbreviating the duration of therapy to 24 weeks, resulted in SVR rates of ~90%. Responsiveness to IFN-ribavirin-based therapy was diminished in immunocompromised patients and in patients with HIV-HCV co-infection and contraindicated in patients with decompensated liver disease or end-stage renal disease. The cumbersome nature of IFN-ribavirin-based therapy (injections, complicated laboratory monitoring, side effects and poor tolerability, modest efficacy, variables and patient subsets associated with poor responsiveness, tailored therapy, futility rules, etc.) was supplanted eventually (in 2016) by DAAs for all genotypes (see below). Most of the variables associated with poor responsiveness to IFN-based therapy became irrelevant, and difficult-to-treat patient subpopulations began to experience responses to DAAs that were indistinguishable from responses in standard patients (see below).

Persons with chronic HCV infection have been shown to suffer increased liver-related mortality. On the other hand, successful antiviral therapy of chronic hepatitis C resulting in an SVR has been shown to improve survival (and to reduce the need for liver transplantation); to lower the risk of liver failure, liver-related death, and all-cause mortality; to slow the progression of chronic hepatitis C; and to reverse fibrosis and even cirrhosis. Whereas the 10-year and 20-year survival in the absence of an SVR is reduced in cirrhotic patients with chronic hepatitis C, survival at these intervals after an SVR has been found to be indistinguishable from that of the general population. Although successful treatment reduces mortality and liver failure (3-4-fold 10-year reduction) in cirrhotic patients (and in those with advanced fibrosis) and reduces the need for liver transplantation and the likelihood of HCC (14-fold 10-year reduction), the risk of liver-related death and HCC persists, albeit at a much reduced level, necessitating continued clinical monitoring and cancer surveillance after SVR in cirrhotics. On the other hand, in the absence of an SVR, IFN-based therapy does not reduce the risk of HCC. Similarly, for nonresponders to PEG IFN–ribavirin therapy, three trials of long-term maintenance therapy with PEG IFN showed no benefit in reducing the risk of histologic progression or clinical decompensation, including the development of HCC. Fortunately, PEG IFN-ribavirin nonresponders can now be retreated with DAAs and experience SVR rates comparable to those in treatment-naïve persons (see below).

FIRST-GENERATION PROTEASE INHIBITORS (2011–2013)

The HCV RNA genome encodes a single polyprotein, which is cleaved during and after translation by host and viral-encoded proteases. One protease involved in the cleavage of the viral polyprotein is an NS3/4A viral protein that has serine protease activity. Telaprevir and boceprevir are serine protease inhibitors that target NS3/4A. In 2011, telaprevir and boceprevir used in combination with PEG IFN and ribavirin were approved by the U.S. Food and Drug Administration (FDA) as the first oral DAA agents for the treatment of hepatitis C genotype 1 (not other genotypes) in adults with stable liver disease, both in patients who had not been treated before or who had failed previous treatment. Although now replaced by more effective, all-oral regimens, these first-in-class agents represented a breakthrough in the treatment of chronic hepatitis C and established milestones against which subsequent therapies could be measured.

Because resistance developed rapidly during monotherapy with telaprevir and boceprevir, these drugs had to be used in combination with PEG IFN and ribavirin. Ribavirin in particular appeared to reduce relapse rates significantly in protease inhibitor-based regimens, such that those who could not take or were intolerant to ribavirin were unlikely to benefit from the addition of these agents. Telaprevir and boceprevir regimens consisted of periods of triple therapy (protease inhibitor plus PEG IFN plus ribavirin) and periods of dual therapy (PEG IFN plus ribavirin). Telaprevir regimens began with 12 weeks of triple therapy followed by dual therapy of a duration based on HCV RNA status at weeks 4 and 12 ("response-guided therapy") and prior treatment status. Boceprevir-based regimens consisted of a 4-week lead-in period of dual (PEG IFN–ribavirin) therapy followed by triple therapy and, in some instances, a further extension of dual therapy, with duration of response-guided therapy based on HCV RNA status at weeks 4, 8, and 24 and prior treatment status.

For patients with HCV genotype 1, protease inhibitors improved the frequency of RVRs and SVRs significantly as compared to PEG IFN plus ribavirin alone. In treatment-naïve patients, telaprevir-based SVRs were achieved in up to 79% of patients who received 12 weeks of triple therapy followed by 12–36 weeks of dual therapy, and among those with EVRs (undetectable HCV RNA at weeks 4 and 12) and response-guided therapy stopped at week 24 (12 weeks of triple therapy, then 12 weeks of dual therapy), SVRs occurred in 83–92%. In studies with boceprevir in treatment-naïve patients, SVRs occurred in 59–66% of patients, and among those with undetectable HCV RNA at 8 weeks, the SVR rate increased to 86–88%. Adding to the complexity of treatment with these protease inhibitors were absolute stopping rules for futility, that is, absence of HCV RNA reductions at critical treatment milestones, which were shown to be invariably predictive of nonresponse (telaprevir: HCV RNA >1000 IU/mL at weeks 4 or 12, or detectable at week 24; boceprevir: HCV RNA ≥100 IU/mL at week 12, or detectable at week 24).

In patients previously treated unsuccessfully with PEG IFN plus ribavirin, telaprevir-based treatment achieved SVRs in 83–88% of prior relapsers, 54–59% of partial responders (HCV RNA reduced by ≥2 $\log_{10}$ IU/mL but not to undetectable levels), and 29–33% of null responders (HCV RNA reduced by <2 $\log_{10}$ IU/mL). With boceprevir, a similar degradation in SVR rate occurred as a function of prior responsiveness—in 75% of prior relapsers, in 40–52% of previous partial responders; in ~30–40% of null responders. In a substantial proportion of protease inhibitor nonresponders, resistance-associated substitutions (RASs, previously referred to as resistance-associated variants, RAVs) could be identified, but these variants were not archived, and wild-type HCV reemerged in almost all cases within 1.5 to 2 years. SVRs to these protease inhibitors were highest in prior relapsers and treatment-naïve patients (white > black ethnicity), lower in prior partial responders, lower still in prior null responders, and lowest in cirrhotic prior null responders, for whom no benefit accrued over PEG IFN/ribavirin treatment. Responses to protease inhibitor triple-drug regimens were higher in patients with *IL28B* C than non-C genotypes, HCV genotype 1b than genotype 1a, less advanced than more advanced fibrosis stage, whites than blacks, lower body mass index (BMI) than elevated BMI, and, for boceprevir, achievement of a >1 $\log_{10}$ HCV RNA reduction during 4 weeks of PEG IFN–ribavirin lead-in therapy. Age and HCV RNA level were less influential and insulin resistance was noninfluential on response to these antiviral agents.

Both of these protease inhibitors had substantial toxicities. Telaprevir was associated with a severe, generalized (trunk and extremities), often confluent, maculopapular, pruritic rash in ~6% of treated patients

(that required careful dermatologic monitoring in all patients and systemic corticosteroid therapy in the most severely affected). Other common side effects included pruritus, rectal burning, nausea, diarrhea, fatigue, dysgeusia (altered or unpleasant taste), and anemia, which required close monitoring, could be relatively refractory, occasionally requiring transfusion and even hospitalization (especially in cirrhotic prior nonresponders). Anemia occurred in half of boceprevir-treated patients, neutropenia in up to 30% and thrombocytopenia in 3–4%. Other side effects of boceprevir include fatigue, nausea, headache, dysgeusia, dry mouth, vomiting, and diarrhea.

Both drugs came with an inconveniently high pill burden and had to be administered every 8 hours with food (TVR with a 20-g fat meal). Use of protease inhibitors was further complicated by numerous drug-drug interactions. As telaprevir and boceprevir are both eliminated by and inhibit CYP3A4, these agents could not be administered with other medications that induce CYP3A4 or are dependent on CYP3A4 for elimination. Care had to be taken to examine for any potential interactions between these protease inhibitors and other medications the patient was taking, and a convenient website became available to check for such drug-drug interactions (*www.hep-druginteractions.org*).

Despite the improvement in SVRs with protease-inhibitor-based regimens for genotype 1 compared to PEG IFN-ribavirin (e.g., in treatment-naïve patients 66–79% vs 38–44%), triple-drug protease-inhibitor therapy was hampered by amplified intolerability, the complexity of response-guided regimens and futility stopping rules, the inconvenience of thrice-daily dosing with meals and a high pill burden, the need for PEG IFN injections and ribavirin with all their intolerability, and multiple drug-drug interactions. Moreover, side effects appeared to be more severe and burdensome once these drugs entered practice, especially in cirrhotic nonresponders, in whom studies reported from Europe showed serious adverse events in up to 45% and deaths in up to 3%. All these issues, as well as rapidly accelerating progress on next-generation and all-oral DAA therapy (see below), conspired to temper enthusiasm for these new antivirals; after a brief stint as recommended therapy (2011–2013), these drugs became obsolete and are no longer recommended.

CONTEMPORARY DIRECT-ACTING ANTIVIRAL COMBINATION THERAPY (2013–)

Since late 2013, the number of new antiviral agents for hepatitis C has expanded substantially, and, currently, PEG IFN-based treatments have been supplanted by six therapeutic regimens: all oral, IFN-free, highly efficacious (>95% SVR), well tolerated, with high barriers to resistance, simple dosing and low pill burdens, treatment durations as brief as 8 to 12 weeks, and, in many cases, pangenotypic efficacy (**Table 334-6**). These drugs are distributed among three classes of DAAs: NS3/4 protease inhibitors (which cleave the single HCV polyprotein into constituent structural and nonstructural proteins), NS5B nucleoside and nonnucleoside polymerase inhibitors (which interfere with the RNA-dependent RNA polymerase [a replicase] involved in synthesis of viral RNA), and NS5A inhibitors (which interfere with a membrane-associated phosphoprotein essential to the HVC RNA replication complex).

The first of the new DAA agents (approved in November 2013) was simeprevir, a second-generation protease inhibitor for genotype 1, followed shortly thereafter (December 2013) by sofosbuvir, a pangenotypic nucleoside polymerase inhibitor. For genotype 1, both of these agents had to be combined with PEG IFN and ribavirin; for genotypes 2 and 3, sofosbuvir was administered with ribavirin, without PEG IFN; however, these treatment

regimens have been supplanted by combinations of all-oral, IFN-free, DAAs, and ribavirin is rarely needed, retained only for very limited indications.

Simeprevir: When simeprevir was used with PEG IFN, its efficacy (genotype 1b > 1a) was similar to that of first-generation protease inhibitors, but required only once-a-day dosing without the complexity of response-guided therapy. Similar to first-generation protease inhibitors, simeprevir was hampered by many drug-drug interactions and side effects (including photosensitivity, rash, and mild hyperbilirubinemia); moreover, patients, with HCV NS3 polymorphism Q80K had markedly reduced drug efficacy, necessitating pretreatment genetic testing and disqualifying a substantial proportion (approximately a third) of potential treatment candidates. Little about simeprevir supported its adoption in combination with PEG IFN and ribavirin. On the other hand, the combination of simeprevir (150 mg) along with sofosbuvir (400 mg) for 12 weeks was found to be effective in treatment-naïve (97% SVR_{12}) or treatment-experienced (95% SVR_{12}) patients without cirrhosis and in treatment-naïve (88% SVR_{12}) or treatment-refractory (79% SVR_{12}) patients with cirrhosis (it remains one of the recommended regimens for genotype 1).

Sofosbuvir: Sofosbuvir, the first nonprotease inhibitor DAA to be approved, has an excellent profile—high potency, high barrier to resistance, pangenotypic activity, very well tolerated with limited adverse effects (most commonly mild fatigue, insomnia, headache, and nausea), once-daily oral administration, and relative freedom from major drug-drug interactions. Sofosbuvir has efficacy in all genotypes (1 to 6); in treatment-naïve subjects and prior nonresponders to PEG IFN-based and protease-inhibitor-based therapy; with PEG IFN-RBV or in IFN-free regimens; in combination with RBV or with NS5A inhibitors; and for treatment periods as brief as 8 to 12 weeks to as long as 24 weeks. Currently, sofosbuvir is used in combination with either the protease inhibitor simeprevir (as described above) or, more commonly, with one of three NS5A inhibitors. Thus, sofosbuvir is a component of four of the six recommended DAA regimens for genotype 1, two of the four regimens for genotype 4, and both of the regimens for genotypes 2, 3, 5, and 6 (Table 334-6).

Sofosbuvir/ledipasvir: The DAA combination that has had a dominant role in the treatment of hepatitis C is sofosbuvir (400 mg) plus the NS5A inhibitor ledipasvir (90 mg) in a once-a-day, fixed-dose, single pill, approved in October 2014 for genotype 1 and in November 2015 for genotypes 4, 5, and 6. Phase-III trials were conducted in treatment-naïve noncirrhotic patients, in treatment-naïve cirrhotic and noncirrhotic patients, and in treatment-experienced cirrhotic and noncirrhotic patients treated for 8, 12, or 24 weeks, both with and without ribavirin. In treatment-naïve noncirrhotics, an SVR_{12} was achieved in 97–99% of subjects, and no benefit was observed by extending therapy from 12 to 24 weeks or by adding ribavirin. Moreover, for treatment-naïve, noncirrhotic patients with baseline HCV RNA $<6 \times 10^6$ IU/mL, a treatment duration of 8 weeks was as effective as one of 12 weeks (94–95% SVR_{12}), which may be a consideration for a proportion of patients. In cirrhotic patients, SVR_{12} was achieved in 97–100% of treatment-naïve subjects (no advantage of extending therapy from 12 to 24 weeks or of adding ribavirin); however, for cirrhotic prior nonresponders to IFN-based therapy, 12 weeks of therapy was inferior (86% SVR_{12}) to 24 weeks of therapy (100% SVR_{12}). This combination, which is equally effective in patients with HIV-HCV co-infection and in African-American patients, has been shown to be highly effective in patients with decompensated cirrhosis and in patients with hepatitis C after liver

transplantation and after kidney transplantation. On the other hand, the safety and efficacy of sofosbuvir/ledipasvir in patients with advanced renal failure have not been established, and all sofosbuvir-containing regimens can be associated with severe bradycardia in patients taking the antiarrhythmic agent amiodarone, especially along with beta blockers; sofosbuvir-containing combinations are contraindicated with amiodarone. Drug-drug interactions are few, but P-gp inducers, like St. John's wort and rifampin, and proton-pump gastric acid inhibitors, like omeprazole, may reduce sofosbuvir/ledipasvir concentrations. Generally, responsiveness to sofosbuvir/ledipasvir is not reduced in patients with baseline RASs to these agents, with the exception of treatment-experienced patients who have baseline NS5A RASs (for whom EASL recommends adding ribavirin or, if ribavirin in contraindicated, extending treatment to 24 weeks).

Paritaprevir/ritonavir, ombitasvir, and dasabuvir: The combination of ritonavir (100 mg)-boosted paritaprevir (150 mg), a protease inhibitor; ombitasvir (25 mg), an NS5A inhibitor; dasabuvir (250 mg), a nonnucleoside polymerase inhibitor; ± weight-based ribavirin (total of five drugs) was approved in December 2014 for genotypes 1 and 4. Paritaprevir/ritonavir and ombitasvir, formulated in a single tablet, are taken once daily, and both dasabuvir (a separate pill) and weight-based ribavirin (when included in the regimen) are taken twice daily. In clinical trials, this combination achieved SVR_{12} rates of 87–100% in treatment-naïve and treatment-experienced patients with genotype 1; without ribavirin, this combination in genotype 1a is ~7% less responsive than genotype 1b. Therefore, in treatment-naïve patients with genotype 1a, this combination is administered *with* ribavirin for 12 weeks in the absence of cirrhosis (95–97% SVR_{12}) or for 24 weeks in the presence of compensated cirrhosis (94% SVR_{12}), while in patients with genotype 1b, the combination does not require ribavirin, and the duration of therapy is 12 weeks for both noncirrhotics and cirrhotics (99–100% SVR_{12}). In prior nonresponders without cirrhosis, the combination is administered for 12 weeks, *with* ribavirin in genotype 1a (96% SVR_{12}), *without* ribavirin in genotype 1b (100% SVR_{12}). In prior nonresponders with cirrhosis, the combination is administered for 24 weeks *with* ribavirin in genotype 1a (SVR_{12} 100% in prior relapsers and partial responders, 95% in prior null responders [in whom treatment without ribavirin was associated with an 80% SVR_{12}]), but only for 12 weeks and *without* ribavirin in genotype 1b (100% SVR_{12}). For genotype 4, the regimen is given for 12 weeks with ribavirin, but without dasabuvir in treatment-naïve and treatment-experienced patients (100% SVR_{12}), including those with compensated cirrhosis. In July 2016, the FDA approved a long-acting formulation of dasabuvir, allowing once-a-day instead of twice-a-day treatment; for genotype 1a, twice-daily ribavirin dosing remains.

This combination is well tolerated with generally mild side effects, for example, fatigue, asthenia, insomnia, headache, and pruritus. Hyperbilirubinemia (primarily unconjugated) and elevations in alanine aminotransferase activity may occur but resolve during or shortly after treatment. Because of occasional hyperbilirubinemia and potential hepatotoxicity (FDA warning letter issued October 2015 regarding hepatic failure/decompensation reported in treated cirrhotic patients), this combination is not recommended in patients with decompensated cirrhosis, and treated cirrhotic patients should be monitored closely for decompensation; however, the safety and efficacy of this combination have been demonstrated for patients with advanced renal insufficiency. Similar to other regimens containing protease inhibitors, drug-drug interactions are common with other drugs that induce CYP3A4 or are dependent on CYP3A4 for elimination.

Checking for potential drug-drug interactions is important prior to initiating therapy with this drug combination (www.hep-druginteractions.org). Responsiveness to this multidrug regimen is not reduced in patients with baseline RASs to these agents.

Compared to sofosbuvir/ledipasvir, this regimen has the disadvantage of requiring twice-a-day ribavirin therapy for genotype 1a and of being contraindicated in decompensated cirrhosis; however, it has the advantage of offering a 12-week, ribavirin-free regimen for prior null responders with cirrhosis and providing an option for patients with renal failure.

Sofosbuvir and Daclatasvir: Daclatasvir, an NS5A inhibitor, along with the polymerase inhibitor sofosbuvir, was approved by the FDA in July 2015 for genotype 3 and in February 2016 for genotype 1 (AASLD-Infectious Diseases Society of America [IDSA] guidelines [see below] include its recommendation as well for genotype 2; in August 2014, this combination was approved in Europe for genotypes 1, 2, 3, and 4, and EASL recommends it for all these genotypes as well as for genotypes 5 and 6). At the time of its approval for genotype 3, daclatasvir filled a need inadequately met by other available combination DAAs. Although data on genotype 3 are the most robust, clinical trials of this combination in genotypes 1 and 2 support its efficacy and recommendations for first-line (genotype 1) and alternative (genotype 2) treatment, in some cases with ribavirin (Table 334-6). Daclatasvir, a 60-mg tablet, and sofosbuvir, a separate 400 mg tablet are taken once-a-day for 12 to 24 weeks.

In clinical trials among treatment-naïve or treatment-experienced patients, SVR_{12} rates for 12 weeks of daclatasvir plus sofosbuvir were 98% with genotype 1 (comparable results in genotypes 1a and 1b), 92% for genotype 2, and 89% for genotype 3. For noncirrhotic patients, the addition of ribavirin or the extension of therapy to 24 weeks did not improve efficacy. In patients with compensated cirrhosis, limited prospective data and data from observational cohorts suggested that extending therapy to 24 weeks, with or without ribavirin, improved efficacy. In cirrhotics, SVR_{12} was achieved in 93% with Child Class-Pugh A and B but in only 56% with Class-C decompensated cirrhosis. For patients with genotype 3 and cirrhosis, the combination was effective in treatment-naïve patients (94% SVR_{12}), but less so in prior nonresponders (69% SVR_{12}). Outcomes in patients with HIV-HCV co-infection were comparable.

Like other sofosbuvir-NS5A inhibitor combinations, daclatasvir plus sofosbuvir is well tolerated (mild fatigue, headache, nausea, diarrhea in 5–14%), but can cause severe bradycardia when administered with amiodarone (contraindicated), especially along with beta blockers. Because daclatasvir is a substrate for CYP3A, CYP3A inducers can reduce daclatasvir levels, and CYP3A inhibitors reduce daclatasvir levels. Similarly, daclatasvir, an inhibitor of P-gp, OATP1B1 and 1B3, and BCP, can increase the levels of drugs that are substrates of these transporters. As noted above for other DAAs, checking for potential drug-drug interactions is advisable prior to initiating therapy (www.hep-druginteractions.org). Responsiveness to daclatasvir-containing drug-combination therapy is reduced in cirrhotic patients with genotype 1a and in both cirrhotic and noncirrhotic patients with genotype 3 who have baseline daclatasvir-associated NS5A RASs.

Although daclatasvir-sofosbuvir is approved for genotypes 1 and 3 and recommended as an alternative for genotype 2, better documented efficacy and simplicity of other regimens have limited the popularity of this drug combination.

Elbasvir/Grazoprevir: Elbasvir (50 mg), an NS5A inhibitor, combined in a single, fixed-dose pill with grazoprevir (100 mg), an NS3/4 protease inhibitor, was approved in January 2016 as a once-a-day (with or without food) treatment for genotypes 1 and 4. In clinical trials, a 12-week course was effective in treatment-naïve and treatment-experienced patients without cirrhosis or with compensated cirrhosis. In treatment-naïve patients, this combination yielded an SVR_{12} in 92% of patients with genotype 1a, 99% with genotype 1b, and 100% with genotype 4 (very small numbers, however); 10 patients with genotype 6 were included, but only 80% achieved SVR_{12}. Cirrhotic and noncirrhotic patients had comparable rates of SVR_{12}, 97% and 94%, respectively. For this drug combination, however, ~11% of patients with genotype 1a harbor NS5A polymorphisms, that is, RASs, at baseline. If present, these NS5A RASs reduce efficacy of elbasvir/grazoprevir (unlike baseline RASs to the most of the other combination DAA regimens described above and below) from 99% to 58% in treatment-naïve patients. Therefore, all patients with genotype 1a require baseline RAS testing; if these RASs are present, treatment extension to 16 weeks and the addition of weight-based ribavirin bring the SVR_{12} up to expected levels of close to 100%. In treatment-experienced patients, both extending treatment to 16 weeks and adding ribavirin were studied; however, generally, in the absence of baseline NS5A RASs, SVR_{12} rates were not increased over those without ribavirin for 12 weeks (94–97%). For genotype 1a, among prior nonresponders to PEG IFN/ribavirin, 12 weeks of elbasvir/grazoprevir suffices without ribavirin except for patients with baseline NS5A RASs, who require 16 weeks of therapy and ribavirin. Among nonresponders to prior protease-inhibitor therapy, even in the absence of baseline NS5A RASs, ribavirin should be added to a 12-week regimen; in the presence of baseline NS5A RASs, treatment should be extended to 16 weeks and ribavirin added. For genotype 1b, NS5A RASs are not an issue, and the only subgroup requiring modification of a 12-week course of therapy are prior nonreponders to protease-inhibitor regimens, for whom ribavirin is added. For genotype 4, the recommended regimen for all prior nonresponders (whether to PEG IFN/ribavirin or protease inhibitor regimens) is 16 weeks of elbasvir/grazoprevir plus ribavirin (Table 334-6).

This combination is just as effective in patients with HIV-HCV co-infection and in patients with advanced renal failure (including those requiring hemodialysis); however, it is contraindicated in decompensated cirrhosis. Like other protease inhibitor regimens, elbasvir/grazoprevir can be associated with aminotransferase elevations and potential hepatotoxicity; because these drugs are excreted by the liver, in decompensated liver disease, plasma drug concentrations may become elevated substantially. Therefore, all treated patients should have alanine aminotransferase screening periodically during therapy, and the drug should be stopped for elevations exceeding 10-fold or for elevations of conjugated bilirubin, alkaline phosphatase, or prothrombin time.

Elbasvir/grazoprevir is well tolerated, with only low levels of mild adverse effects (fatigue, headache, nausea in 5–11%) seen just as frequently in placebo recipients. Both elbasvir and grazoprevir are substrates for CYP3A and are subject to multiple potential drug-drug interactions. Therefore, this combination should not be used with potent CYP3A inducers; conversely, CYP3A and OATP1B1 inhibitors can lead to untoward elevations of plasma elbasvir/grazoprevir concentrations. Checking for potential drug-drug interactions is advisable prior to initiating therapy (www.hep-druginteractions.org).

Compared to other available regimens for genotypes 1 and 4, elbasvir/grazoprevir has the disadvantage/inconvenience of requiring baseline NS5A RAS testing but the advantages of a comparable regimen for cirrhotics and noncirrhotics, for treatment-naïve and treatment-experienced patients, and for patients with normal renal function and with renal failure.

Sofosbuvir/velpatasvir: The combination in a single, fixed-dose pill of velpatasvir (100 mg), a highly potent, pangenotypic NS5A inhibitor, along with the polymerase inhibitor sofosbuvir (400 mg) was approved in June 2016 for genotypes 1–6, in treatment-naïve and treatment-experienced noncirrhotics and cirrhotics. Ribavirin is not required, including in patients with genotypes 2 and 3, except in patients with decompensated cirrhosis.

In a series of clinical trials, this combination for 12 weeks in the absence of ribavirin was shown to yield 99% SVR_{12} (range 97–100%) in genotypes 1, 2, 4, 5, and 6 and 95% in genotype 3. Baseline NS5A RASs had no impact on responsiveness.

Prior to the availability of this drug combination, patients with genotype 3, especially those with cirrhosis and prior null response to other therapies, proved to be the most refractory subset of patients. In treatment-naïve patients with genotype 3, 12 weeks of sofosbuvir/velpatasvir (95% SVR_{12}) was superior to 24 weeks of sofosbuvir plus ribavirin (80% SVR_{12}). In patients with genotype 3, the combination of sofosbuvir/velpatasvir for 12 weeks was comparable in noncirrhotics (97% SVR_{12}) and cirrhotics (91% SVR_{12}) and in treatment-naïve (97% SVR_{12}) and treatment-experienced (90% SVR_{12}) patients, superior in all these categories to 24 weeks of sofosbuvir plus ribavirin (87%, 66%, 86%, and 63%, respectively). In cirrhotic null responders, most available IFN-free regimens for genotype 3 (including daclatasvir plus sofosbuvir, approved specifically for this genotype) achieved SVR_{12} rates in the range of ~60–75%, while the combination of PEG IFN, ribavirin, and sofosbuvir could boost SVR_{12} to the mid-80% range. For treatment-experienced patients with genotype 3, sofosbuvir/velpatasvir in noncirrhotics and cirrhotics had similarly high efficacy (91% and 89% SVR_{12}, respectively); this was the highest recorded SVR_{12} for genotype-3 cirrhotic null responders treated with IFN-free DAA regimens. Finally, in patients with genotypes 1–4 and 6 and with decompensated, Class-B cirrhosis (55% treatment-experienced), sofosbuvir/velpatasvir plus ribavirin for 12 weeks yielded an SVR_{12} in 94%; this result was better than sofosbuvir/velpatasvir without ribavirin for 12 weeks (83% SVR_{12}) or 24 weeks (86% SVR_{12}).

Like other all-oral DAAs, sofosbuvir/velpatasvir was very well tolerated; in noncirrhotic and compensated cirrhotic patients, mild headache and fatigue was seen in >10%—this occurred in a comparable proportion of placebo recipients; in decompensated cirrhosis, mild fatigue, headache, nausea, insomnia, diarrhea, and anemia (ribavirin was part of the regimen) was seen in >10%. Like other sofosbuvir-containing regimens, sofosbuvir/velpatasvir should not be administered along with amiodarone (potential serious bradycardia); in addition, P-gp inducers and moderate-to-potent CYP3A inducers can reduce plasma levels of sofosbuvir and/or velpatasvir. Checking for drug-drug interactions prior to therapy is advisable (www.hep-druginteractions.org). Baseline RASs do not influence responsiveness to this combination.

FUTURE DIRECT-ACTING ANTIVIRAL COMBINATION THERAPY (2017–)

Most treatment needs have been met by contemporary DAA regimens described above; however, several additional, highly potent, pangenotypic drug combinations are in development. For example, an investigative protease inhibitor (voxilaprevir) added to the polymerase inhibitor/NS5A inhibitor combination of sofosbuvir/velpatasvir yields a very well tolerated *triple-drug* combination with 97% SVR_{12} across all HCV genotypes and patient subgroups. These include noncirrhotic/cirrhotic, treatment-naïve/treatment-experienced groups, including those who had prior NS5A treatment and results were independent of the number of prior DAA drug classes received; no effects of baseline NS5A RASs were noted. Several experimental combinations may allow even briefer durations of therapy. In a small, exploratory trial, a *6-week* combination of sofosbuvir plus an experimental pangenotypic, very high potency, very low resistance NS5A inhibitor (odalasvir) achieved SVR_{12} in 100% of 12 patients with genotype 1. Similarly, in a *6-week* triple combination of odalasvir with the protease inhibitor simeprevir and an experimental polymerase inhibitor ("AL-335"), SVR_{12} was observed in 100% of 20 treatment-naïve noncirrhotic patients with genotype 1. In phase-II clinical trials, *8 weeks* of an experimental combination of two high-potency, pangenotypic DAAs, a protease inhibitor ("ABT-493") plus an NS5A inhibitor ("ABT-530"), yielded 100% SVR_{12} in treatment-naïve noncirrhotic patients with genotypes 1, 2, and 3. In cirrhotics with genotype 3 and in patients with genotypes 4, 5, and 6, 12 weeks of therapy with this DAA combination yielded 100% SVR_{12}. In patients with prior DAA treatment failure, 12 weeks of this double-combination sufficed to achieve a ≥95% SVR_{12}; neither baseline NS5A nor protease inhibitor RASs influenced SVR_{12} rates. No safety issues have been encountered, and the potential for drug-drug interactions is limited. These promising combinations are undergoing phase-II and phase-III trials.

Less advanced is the development of inhibitors of host proteins, such as oral, nonimmunosuppressive inhibitors of cyclophilin A (which interacts with NS5A during HCV replication) and subcutaneous antisense antagonists of host liver-expressed micro-RNA-122 (which promotes HCV replication). Given the accelerated progress of all-oral, short-treatment-duration, high-efficacy, DAAs, these alternative approaches may not be practical or competitive; moreover, development of both approaches has been retarded by emerging toxicities such as pancreatitis associated with cyclophilin inhibitors and jaundice associated with micro-RNA-122.

Although data on the impact of DAAs on the natural history of chronic hepatitis C are still limited, preliminary findings are that successful therapy is associated with a gradual reduction in fibrosis progression and a regression of advanced fibrosis (cirrhosis), improvement in survival among patients with decompensated cirrhosis, and a decline in the number of patients with hepatitis C being referred for liver transplantation. Based on the known prevalence, natural history, and rate of progression of chronic hepatitis C and on the efficacy of DAA therapies and their impact on the complications of hepatitis C, modeling estimates have suggested that the availability and application of these therapies have the potential to reduce the hepatitis C-associated disease burden including liver-related death, HCC, decompensated cirrhosis, and liver transplantation by 50–70% between 2015 and 2050.

TREATMENT RECOMMENDATIONS

Because the pace of new drug development and approval has been so rapid, the AASLD and the IDSA have been providing a consensus of updated treatment recommendations for patients with hepatitis C; these recommendations, which continue to be revised regularly based on new data, are available online at www.hcvguidelines.org and should be consulted before initiating therapy (Table 334-6). The EASL issues similar (but not identical) treatment recommendations annually for hepatitis C (www.easl.eu), most recently in September 2016. Divergences between AASLD-IDSA and EASL recommendations are noted in Table 334-6.

Prior to therapy, HCV genotype should be determined, because the genotype dictates which treatment regimens are indicated (Table 334-6). Monitoring of serum HCV RNA levels pretreatment, during treatment, and posttreatment is crucial in assessing response to therapy; moreover, the baseline level may contribute to determining the duration of therapy (e.g., in noncirrhotic patients with genotype 1 and HCV RNA $<6 \times 10^6$ IU/mL, 8 [instead of the usual 12] weeks of sofosbuvir/ledipasvir may be a consideration). The goal of treatment is to eradicate HCV RNA during therapy and to document that the virus remains undetectable for at least 12 weeks after completion of therapy (SVR_{12}). Several reports have appeared describing hepatitis B reactivation, often severe, during and after DAA therapy in patients coinfected with HCV and HBV who were not being treated for their HBV infections. Therefore, screening for HBV infection is recommended prior to initiating DAA therapy for hepatitis C (which should have been done to determine HBV-immunity status as a prelude to recommended hepatitis B vaccination in patients with chronic hepatitis C), and therapy for HBV infection (for those meeting HBV treatment criteria, see above) should be initiated prior to or simultaneously with HCV therapy.

INDICATIONS FOR ANTIVIRAL THERAPY

Patients with chronic hepatitis C who have detectable HCV RNA in serum, whether or not aminotransferase levels are increased, and chronic hepatitis of any grade and stage are candidates for antiviral therapy with DAA agents. The only exception would be patients with short life expectancies, for whom treating hepatitis C would have no influence on longevity. Certainly, for patients with advanced liver disease, early treatment merits a high priority. Although patients with persistently normal aminotransferase activity tend to progress histologically very slowly or not at all, they respond to antiviral therapy just as well as do patients with elevated aminotransferase levels; therefore, such patients are potential candidates for antiviral therapy. As noted above, antiviral therapy has been shown to improve survival and complication-free survival and to slow progression of and to reverse fibrosis.

HCV genotype determines the regimen to be selected (Table 334-6). Similarly, the absence or presence of cirrhosis/advanced fibrosis determines the treatment options from which to select, including the antiviral agents to be used, the duration of therapy, and the need for ribavirin (Table 334-6). A pretreatment liver biopsy to assess histologic grade and stage provides substantial information about progression of hepatitis C in the past, has prognostic value for future progression, and can identify such histologic factors as steatosis and stage of fibrosis, which can influence responsiveness to therapy. As therapy has improved for patients with a broad range of histologic severity, and as noninvasive measures of the stage of fibrosis (e.g., assessment of liver elasticity by imaging) have gained in accuracy and popularity, noninvasive approaches have supplanted histology in most cases. If cirrhosis/advanced fibrosis is present prior to therapy, the risk of HCC, although reduced substantially by successful therapy, is not eliminated, and twice yearly posttreatment imaging for HCC

surveillance (and endoscopic surveillance for esophageal varices at intervals of 1–3 years) is indicated even after an SVR. In patients with low-level fibrosis at baseline, achievement of an SVR allows the cessation of such surveillance.

Patients who have relapsed after, or failed to respond to, a course of IFN-based or DAA agent-based therapy are candidates for retreatment with a DAA therapy regimen (Table 334-6). For patients who have failed to respond to a DAA combination, options include increasing the duration of therapy with the failed regimen, adding ribavirin, or changing the drug class (e.g., after failed protease and polymerase inhibitors, switching to an NS5A-containing combination). In the presence of cirrhosis or a need for urgent retreatment, patients who have failed protease inhibitor plus polymerase inhibitor combination therapy or who have failed an NS5A combination are candidates for RAS testing and tailored therapy based on such resistance testing. If reliable RAS testing is not available, adding ribavirin or extending the duration of therapy are options. For prior nonresponders to IFN-based therapy, NS5A inhibitor-containing regimens are highly effective; however, reduced responsiveness can be encountered, especially in cirrhotic patients. For this relatively refractory group, ideally, the most potent/effective NS5A regimen should be selected to give such patients the best chance of responding and to avoid treatment-emergent NS5A RASs. Additional details for treatment of such patient subgroups can be found at *www.hcvguidelines.org*.

Persons with acute hepatitis C are also candidates for antiviral therapy (**Chap. 332**) with the same DAA agents approved for chronic hepatitis C; delaying the initiation of therapy for an observation period of 12–16 weeks (and even up to 6 months) has been recommended to allow for spontaneous recovery, especially in light of the fact that most cases of acute hepatitis C are not clinically severe or rapidly progressive. The duration of therapy for acute hepatitis C has not been determined definitively; however, in a small study of 20 patients, 6 weeks of sofosbuvir/ledipasvir sufficed for a 100% SVR_{12}. According to 2016 EASL recommendations, patients with acute hepatitis C should be treated for 8 weeks with a genotype-appropriate DAA regimen consisting of sofosbuvir plus one of the three approved NS5A inhibitors without ribavirin (extended to 12 weeks for patients with acute hepatitis C and HIV co-infection or for patients with acute hepatitis C and a baseline HCV RNA level >1 million IU/mL). In patients with biochemically and histologically mild chronic hepatitis C, the rate of progression is slow; however, such patients respond just as well to antiviral therapy as those with elevated aminotransferase levels and more histologically severe hepatitis. Because of the high cost of DAA treatments, initially a higher priority was assigned to patients with advanced fibrosis/cirrhosis; however, this controversial approach was relied upon by some medical insurers and pharmacy benefit management organizations to withhold therapy from patients with low-level fibrosis. Unfortunately, delaying therapy until fibrosis becomes advanced misses the opportunity to prevent all the dire consequences of chronic hepatitis C (liver failure, death/transplantation, HCC), which can be reduced, but not eliminated completely once advanced fibrosis is established. Therefore, therapy for patients with mild disease is justified as well as cost-effective.

Patients with compensated cirrhosis can respond to therapy, and their likelihood of a sustained response with DAAs is comparable to that in noncirrhotics. Patients with decompensated cirrhosis, who were not candidates for IFN-based antiviral therapy, respond well to DAA therapy regimens consisting of combinations of polymerase inhibitors and NS5A inhibitors (e.g., sofosbuvir/ledipasvir, sofosbuvir/velpatasvir); however, protease-inhibitor-containing combinations have been associated with potential hepatotoxicity and hepatic

decompensation and are contraindicated in this patient subset. Patients with decompensated cirrhosis should be referred to a liver transplantation center. DAAs are highly effective not only for patients with end-stage liver disease awaiting liver transplantation but also for patients with recurrent hepatitis C after liver transplantation. Ideally, patients should be treated prior to liver transplantation; however, a concern is that eradication of HCV infection will disqualify such patients from accepting donor livers from persons with HCV infection, thus contracting the potential donor pool and limiting accessibility to donor organs and timely transplantation. In addition, responsiveness to DAA therapy appears to be reduced in patients with decompensated cirrhosis and with high model for end-stage liver disease (MELD) scores; in this subgroup, responsiveness after liver transplantation would be substantially better. Therefore, advocacy has been expressed (recommended by EASL) for postponing DAA therapy in patients with high-MELD HCV-associated end-stage liver disease until after liver transplantation; the decision whether to treat pretransplantation or posttransplantation should be individualized thoughtfully for each patient, based on such factors as MELD score, time anticipated prior to availability of a donor organ, relative clinical stability, and co-morbidities **(Chap. 338)**. The cutaneous and renal vasculitis of HCV-associated essential mixed cryoglobulinemia **(Chap. 332)** may respond to antiviral therapy, but sustained responses were rare after discontinuation of therapy in the IFN era, and prolonged, potentially indefinite, therapy was recommended. Now that more effective DAAs are available, a 12-week course of sofosbuvir-based combination therapy has been shown to yield an SVR_{12} rate exceeding 80% in cryoglobulinemic vasculitis. Anecdotal reports suggest that IFN-based antiviral therapy may be effective in porphyria cutanea tarda or lichen planus associated with hepatitis C; whether the more appealing DAAs are effective in these groups remains to be documented.

In patients with HCV/HIV co-infection, hepatitis C is more progressive and severe than in HCV-monoinfected patients. Although patients with HCV/HIV co-infection responded less well to IFN-based antiviral therapy for hepatitis C, they respond as well as patients with HCV infection alone to DAA combination regimens. In HCV/HIV-infected patients, ribavirin can potentiate the toxicity of didanosine (e.g., lactic acidosis) and the lipoatrophy of stavudine, and zidovudine can exacerbate ribavirin-associated hemolytic anemia; therefore, these drug combinations should be avoided.

Patients with a history of injection drug use and alcoholism can be treated successfully for chronic hepatitis C, preferably in conjunction with drug and alcohol treatment programs. Moreover, because injection-drug users, as a source of transmission to others, account disproportionately for perpetuating the spread of HCV infection in the population, the impact of treating active injection-drug users is amplified by reducing such transmission. The approved oral combinations of DAAs are effective in patients with mild-modest renal failure and require no dose adjustments; however, in patients with severe renal impairment (creatinine clearances <30 mL/minute), data are limited on the use of sofosbuvir-containing combinations. For such patients, including those undergoing hemodialysis, recommended combinations are 12 weeks of elbasvir/grazoprevir for genotypes 1a, 1b, and 4 or 12 weeks of paritaprevir/ritonavir, ombitasvir, and dasabuvir for genotype 1b. In genotype 1a, the addition of 200 mg/day of ribavirin to paritaprevir/ritonavir, ombitasvir, and dasabuvir, if the hemoglobin level exceeds 10 g/dL, is an alternative regimen but requires vigilance for the onset of ribavirin-induced hemolytic anemia. For patients with severe renal impairment and HCV genotypes 2, 3, 5, or 6, PEG IFN with low-dose

ribavirin (200 mg daily, if the hemoglobin exceeds 10 g/dL) is recommended. After renal transplantation, levels of SVR_{12} in patients treated with the approved oral DAA combinations have approached 100%.

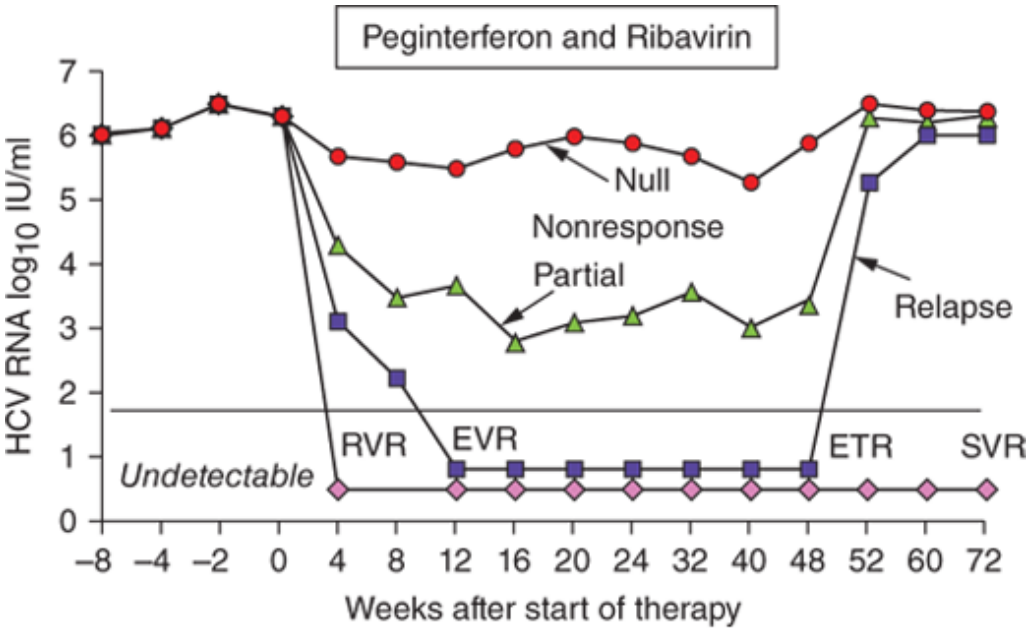
No clinical studies of the use of DAAs during pregnancy are available. Ribavirin is contraindicated during pregnancy; therefore, any regimen including ribavirin should not be used. Sofosbuvir; sofosbuvir + ledipasvir; and paritaprevir/ritonavir, ombitasvir, and dasabuvir are classified as pregnancy category B; the other DAAs do not have a pregnancy classification. Therefore, these therapies are not indicated routinely in pregnancy and should be used, with caution, only if the benefit of treatment is compelling and justified compared to the potential for fetal risk.

Choosing among available treatment options: The large number of recommended all-oral DAA combinations can be daunting to treating clinicians. In some instances, the combination approved is determined by insurance payers; however, cost considerations aside, how is the clinician to choose among the options? The most popular of the regimens has been fixed-dose, single-pill sofosbuvir/ledipasvir, which is effective for all genotypes except 2 and 3, which requires no baseline RAS testing, and which can be used in noncirrhotic patients with genotype 1 and low-level viremia for as brief a period as 8 weeks. For genotypes 2 and 3, fixed-dose, single-pill sofosbuvir/velpatasvir appears to be the combination of choice; because this combination is so effective across all genotypes, in the future, for simplicity, clinicians may resort to a "one-size-fits-all" regimen such as this one in all patients (except for those with advanced renal failure). In addition, this regimen is the only one that can be used in almost all situations (independent of genotype, treatment experience, and cirrhosis) without ribavirin, and the duration of which is almost always 12 weeks; exceptions: a) ribavirin recommended for decompensated cirrhosis, b) EASL recommends adding ribavirin in treatment-experienced patients with genotype 3 or, if ribavirin is contraindicated, extending treatment to 24 weeks (Table 334-6, footnote c). As noted above, protease-inhibitor-containing DAA regimens (elbasvir/grazoprevir; paritaprevir/ritonavir, ombitasvir, and dasabuvir; simeprevir and sofosbuvir) are contraindicated in decompensated cirrhosis. For advanced renal failure, safety and efficacy have been documented for elbasvir/grazoprevir and paritaprevir/ritonavir, ombitasvir, and dasabuvir, but not for sofosbuvir-NS5A combinations.

FIGURE 334-2

Classification of virologic responses based on outcomes during and after a 48-week course of pegylated interferon (PEG IFN) plus ribavirin antiviral therapy in patients with hepatitis C, genotype 1 or 4 (for genotype 2 or 3, the course would be 24 weeks). Nonresponders can be classified as null responders (hepatitis C virus [HCV] RNA reduction of <2 $\log_{10}$ IU/mL) or partial responders (HCV RNA reduction ≥2 $\log_{10}$ IU/mL but not suppressed to undetectable) by week 24 of therapy. In responders, HCV RNA can become undetectable, as shown with sensitive amplification assays, within 4 weeks (RVR, rapid virologic response); can be reduced by ≥2 $\log_{10}$ IU/mL within 12 weeks (early virologic response, EVR; if HCV RNA is undetectable at 12 weeks, the designation is "complete" EVR); or at the end of therapy, 48 weeks (ETR, end-treatment response). In responders, if HCV RNA remains undetectable for 24 weeks after ETR, week 72, the patient has a sustained virologic response (SVR), but if HCV RNA becomes detectable again, the patient is considered to have relapsed. The posttreatment week-24 SVR (SVR_{24}) has been supplanted by an SVR at week 12 (SVR_{12}), which has been

shown to be equivalent to an SVR_{24}. In patients treated with DAA therapy, RVR and EVR milestones are largely irrelevant, being met by almost all patients. *(Reproduced with permission, courtesy of Marc G. Ghany, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health and the American Association for the Study of Liver Diseases. Hepatology 49:1335, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 334-6

Indications and Recommendations for Antiviral Therapy of Chronic Hepatitis C[a]

Standard Indications for Therapy

All patients with chronic HCV infection (detectable HCV RNA, with or without elevated ALT) except for those with short life expectancies owing to comorbid conditions.

Any stage of fibrosis; highest priority for advanced fibrosis [METAVIR stage 3]/cirrhosis [METAVIR stage 4] (pretreatment biopsy is no longer embraced and has been supplanted by noninvasive measures of fibrosis, e.g., imaging to determine liver elasticity)

Responsiveness in groups previously refractory to interferon-based therapy (HIV-HCV co-infection, renal insufficiency, African American and Latino ethnicity, *IL28B* non-C haplotype, obesity, insulin resistance, hepatic decompensation, etc.) is not diminished to contemporary direct-acting oral combination regimens.

Retreatment Recommended

Relapsers, partial responders, or nonresponders after a previous course of interferon-based therapy or prior direct-acting antiviral therapy (see genotype-specific recommendations below).

Antiviral Therapy Not Recommended

Pregnancy: No clinical studies of direct-acting antivirals during pregnancy are available. Ribavirin is contraindicated during pregnancy; therefore, any regimen including ribavirin should not be used. Sofosbuvir; sofosbuvir + ledipasvir; and paritaprevir/ritonavir + ombitasvir + dasabuvir are classified as pregnancy category B, but the other direct-acting antivirals do not have a pregnancy classification. Therefore, these therapies are not indicated routinely in pregnancy and should be used, with caution, only if the benefit of treatment outweighs the potential for fetal risk.

Therapeutic Regimens (based on AASLD-IDSA recommendations, *www.hcvguidelines.org*)[b]

The European Association for the Study of the Liver (EASL) issued recommendations in 2016; divergences from AASLD-IDSA recommendations are summarized as a footnote below.[c]

TREATMENT-NAÏVE OR RELAPSED AFTER PRIOR PEG IFN/RIBAVIRIN THERAPY

Genotype 1a

ledipasvir + sofosbuvir 12 weeks (consider 8 weeks for noncirrhotic patients with HCV RNA <6 × 10^6 IU/mL)

paritaprevir/ritonavir + ombitasvir + dasabuvir + RBV 12 weeks (no cirrhosis) or 24 weeks (cirrhosis)

sofosbuvir + simeprevir 12 weeks (no cirrhosis) or ± RBV 24 weeks (cirrhosis)

daclatasvir + sofosbuvir 12 weeks (no cirrhosis) or ± RBV 24 weeks (cirrhosis)

FAILED PRIOR PEG IFN/RIBAVIRIN THERAPY, NO CIRRHOSIS

Genotype 1a

ledipasvir + sofosbuvir 12 weeks

paritaprevir/ritonavir + ombitasvir + dasabuvir + RBV 12 weeks

sofosbuvir + simeprevir 12 weeks

daclatasvir + sofosbuvir 12 weeks

grazoprevir + elbasvir 12 weeks (without ELB NS5A RASs) or + RBV x 16 weeks (ELB NS5A RASs)

sofosbuvir + velpatasvir 12 weeks

Genotype 1b

ledipasvir + sofosbuvir 12 weeks

paritaprevir/ritonavir + ombitasvir + dasabuvir 12 weeks

sofosbuvir + simeprevir 12 weeks

daclatasvir + sofosbuvir 12 weeks

grazoprevir + elbasvir 12 weeks

sofosbuvir + velpatasvir 12 weeks

grazoprevir + elbasvir 12 weeks (no cirrhosis or cirrhosis sans ELB NS5A RASs) or + RBV × 16 weeks (ELB NS5A RASs)

sofosbuvir + velpatasvir 12 weeks

Genotype 1b

ledipasvir + sofosbuvir 12 weeks (consider 8 weeks for noncirrhotic patients with HCV RNA <6 × 10^6 IU/mL)

paritaprevir/ritonavir + ombitasvir + dasabuvir 12 weeks

sofosbuvir + simeprevir 12 weeks (no cirrhosis) or ± RBV 24 weeks (cirrhosis)

daclatasvir + sofosbuvir 12 weeks (no cirrhosis) or ± RBV 24 weeks (cirrhosis)

grazoprevir + elbasvir 12 weeks

sofosbuvir + velpatasvir 12 weeks

Genotype 2

sofosbuvir + velpatasvir 12 weeks

daclatasvir + sofosbuvir (no cirrhosis) 12 weeks or 16–24 weeks (cirrhosis)

Genotype 3

sofosbuvir + velpatasvir 12 weeks

daclatasvir + sofosbuvir 12 weeks (no cirrhosis) or ± RBV 24 weeks (cirrhosis)

Genotype 4

sofosbuvir + velpatasvir 12 weeks

ledipasvir + sofosbuvir 12 weeks

paritaprevir/r + ombitasvir + RBV 12 weeks (no dasabuvir)

grazoprevir + elbasvir 12 weeks

Genotypes 5, 6

sofosbuvir + velpatasvir 12 weeks

ledipasvir + sofosbuvir 12 weeks

paritaprevir/ritonavir + ombitasvir + RBV 12 weeks (no dasabuvir)

grazoprevir + elbasvir 12 weeks (prior relapse) or + RBV 16 weeks (prior nonresponse)

ledipasvir + sofosbuvir 24 weeks

Genotypes 5, 6

sofosbuvir + velpatasvir 12 weeks

ledipasvir + sofosbuvir 12 weeks

Genotype 2

sofosbuvir + velpatasvir 12 weeks

daclatasvir + sofosbuvir 12 weeks

Genotype 3

sofosbuvir + velpatasvir 12 weeks

daclatasvir + sofosbuvir 12 weeks

Genotype 4

sofosbuvir + velpatasvir 12 weeks

ledipasvir + sofosbuvir 12 weeks

paritaprevir/r + ombitasvir + RBV 12 weeks (no dasabuvir)

grazoprevir + elbasvir 12 weeks (prior relapse) or + RBV 16 weeks (prior nonresponse)

Genotypes 5, 6

sofosbuvir + velpatasvir 12 weeks

ledipasvir + sofosbuvir 12 weeks

FAILED PRIOR PEG IFN/RIBAVIRIN THERAPY, COMPENSATED CIRRHOSIS

Genotype 1a

ledipasvir + sofosbuvir + RBV 12 weeks

ledipasvir + sofosbuvir 24 weeks

sofosbuvir + velpatasvir 12 weeks

grazoprevir + elbasvir 12 weeks (without ELB NS5A RASs) or + RBV × 16 weeks (ELB NS5A RASs)

paritaprevir/ritonavir + ombitasvir + dasabuvir + RBV 24 weeks

sofosbuvir + simeprevir ± RBV 24 weeks (no Q80K variant)

daclatasvir + sofosbuvir ± RBV 24 weeks

Genotype 1b

ledipasvir + sofosbuvir + RBV 12 weeks

ledipasvir + sofosbuvir 24 weeks

sofosbuvir + velpatasvir 12 weeks

grazoprevir + elbasvir 12 weeks

paritaprevir/ritonavir + ombitasvir + dasabuvir 12 weeks

sofosbuvir + simeprevir ± RBV 24 weeks

daclatasvir + sofosbuvir ± RBV 24 weeks

Genotype 2

sofosbuvir + velpatasvir 12 weeks

sofosbuvir + daclatasvir 16 or 24

	weeks *Genotype 3* **sofosbuvir + velpatasvir 12 weeks** daclatasvir + sofosbuvir + RBV 24 weeks *Genotype 4* **sofosbuvir + velpatasvir 12 weeks** ledipasvir + sofosbuvir + RBV 12 weeks **FEATURES ASSOCIATED WITH REDUCED RESPONSIVENESS TO DIRECT-ACTING ANTIVIRAL COMBINATION THERAPY** Genotype and subtype (genotype 1a less responsive than genotype 1b for several drugs) Treatment experience Advanced fibrosis (bridging fibrosis, cirrhosis) Reduced adherence

[a]Rapidly evolving new recommendations continue to be issued; for up-to-date treatment recommendations, please see www.hcvguidelines.org. [b]Class-I recommendations in **bold** font, all others are Class-II recommendations [c]The following EASL recommendations differ from those of AASLD-IDSA (Please note that, although mentioned in EASL recommendations, testing for baseline RASs is not recommended routinely, but, if reliable resistance testing available, results can be used to guide therapy.):

Genotype 1

For genotype 1, simeprevir + sofosbuvir is not recommended.

For genotype 1a, treatment-experienced patients (IFN-based regimen failures) treated with sofosbuvir + ledipasvir should have weight-based ribavirin added. If reliable testing for RASs is available, ribavirin is needed only if baseline RASs are present, and, in such patients, if ribavirin is contraindicated, sofosbuvir + ledipasvir should be extended to 24 weeks.

For genotype 1b, in treatment-naïve, noncirrhotic patients receiving paritaprevir/ritonavir + ombitasvir + dasabuvir a treatment duration of 8 weeks can be considered.

For genotype 1a, in patients treatment with grazoprevir + elbasvir, EASL recommends testing for ELB RASs even in noncirrhotics. If resistance testing is not done, the level of baseline HCV RNA should determine whether ribavirin is added and the duration of therapy. If HCV RNA >800,000 IU/mL, add ribavirin and treat for 16 weeks; if HCV RNA ≤800,000 IU/mL, ribavirin is not added, and treatment for 12 weeks suffices. If baseline testing for RASs is available, patients with HCV RNA >800,000 IU/mL and detectable RASs should be treated with ribavirin for 16 weeks. Treatment without ribavirin and for 12 weeks suffices if HCV RNA ≤800,000 IU/mL even with detectable RASs or even if HCV RNA >800,000 IU/mL with undetectable RASs.

For genotype 1a, in treatment-experienced patients (IFN-based regimen failures) treated with daclatasvir + sofosbuvir, follow the same recommendations described above for ledipasvir + sofosbuvir regarding the addition of ribavirin.

Genotype 2

EASL recommendations are the same as those of AASLD-IDSA.

Genotype 3

For treatment-experienced patients (IFN-based regimen failures) treated with sofosbuvir + velpatasvir or sofosbuvir + daclatasvir, if testing for baseline RASs is not available, add weight-based ribavirin. If resistance testing is available, ribavirin is needed only if baseline RASs are present, and, in such patients, if ribavirin is contraindicated, treatment should be extended to 24 weeks.

Genotype 4

Treatment-experienced patients (IFN-based regimen failures) treated with sofosbuvir + ledipasvir should have weight-based ribavirin added, and, in such patients, if ribavirin is contraindicated, treatment should be extended to 24 weeks.

In treatment-experienced patients (IFN-based regimen failures) treated with grazoprevir + elbasvir, if HCV RNA >800,000 IU/mL, weight-based ribavirin should be added, and treatment should be extended to 16 weeks.

EASL recommends two additional treatment options for genotype 4 (noncirrhotic or cirrhotic) that are not included in AASLD-IDSA guidelines: sofosbuvir + daclatasvir and sofosbuvir + simeprevir. For both these options, treatment-naïve patients should be treated for 12 weeks without ribavirin; treatment-experienced (IFN-based regimen failures) patients should be treated with ribavirin for 12 weeks or, if ribavirin is contraindicated, without ribavirin for 24 weeks.

Genotypes 5 and 6

Treatment-experienced patients (IFN-based regimen failures) treated with sofosbuvir + ledipasvir should have weight-based ribavirin added, and, in such patients, if ribavirin is contraindicated, treatment should be extended to 24 weeks.

EASL recommends an additional treatment option for genotype 5 and 6 (noncirrhotic or cirrhotic) that is not included in AASLD-IDSA guidelines: sofosbuvir + daclatasvir. Treatment-naïve patients should be treated for 12 weeks without ribavirin; treatment-experienced (IFN-based regimen failures) patients should be treated with ribavirin for 12 weeks or, if ribavirin is contraindicated, without ribavirin for 24 weeks.

Drug doses: sofosbuvir 400 mg; ledipasvir 90 mg; paritaprevir 150 mg; ritonavir 100 mg; ombitasvir 25 mg; dasabuvir 250 mg; ribavirin, weight-based: 1,000 mg (<75 Kg)-1,200 mg (≥75 kg); simeprevir 150 mg; daclatasvir 60 mg; elbasvir 50 mg; grazoprevir 100 mg; velpatasvir 100 mg.

Abbreviations: AASLD, American Association for the Study of Liver Diseases; ALT, alanine aminotransferase; ELB NS5A RASs, elbasvir NS5A resistance-associated substitutions; HCV, hepatitis C virus; IFN, interferon; IDSA, Infectious Diseases Society of America; PEG IFN, pegylated interferon; IU, international units (1 IU/mL is equivalent to ~2.5 copies/mL); RASs, resistance-associated substitutions; RBV, ribavirin.

AUTOIMMUNE HEPATITIS

DEFINITION

Autoimmune hepatitis is a chronic disorder characterized by continuing hepatocellular necrosis and inflammation, usually with fibrosis, which can progress to cirrhosis and liver failure. When fulfilling criteria of severity, this type of chronic hepatitis, when untreated, may have a 6-month mortality of as high as 40%. Based on contemporary estimates of the natural history of autoimmune hepatitis, the 10-year survival is 80–98% for treated and 67% for untreated patients. The prominence of extrahepatic features of autoimmunity and seroimmunologic abnormalities in this disorder supports an autoimmune process in its pathogenesis; this concept is reflected in the prior labels *lupoid* and *plasma cell hepatitis*. Autoantibodies and other typical features of autoimmunity, however, do not occur in all cases; among the broader categories of “idiopathic” or cryptogenic chronic hepatitis, many, perhaps the majority, are probably autoimmune in origin. Cases in which hepatotropic viruses, metabolic/genetic derangements (including nonalcoholic fatty liver disease), and hepatotoxic drugs have been excluded represent a spectrum of heterogeneous liver disorders of unknown cause, a proportion of which are most likely autoimmune hepatitis.

IMMUNOPATHOGENESIS

The weight of evidence suggests that the progressive liver injury in patients with autoimmune hepatitis is the result of a cell-mediated immunologic attack directed against liver cells in the setting of a loss of, or failed, immunologic tolerance for self liver antigens. In all likelihood, predisposition to autoimmunity is inherited, whereas the liver specificity of this injury is triggered by environmental (e.g., chemical, drug [e.g., minocycline], or viral) factors. For example, patients have been described in whom apparently self-limited cases of acute hepatitis A, B, or C led to autoimmune hepatitis, presumably because of genetic susceptibility or

predisposition. Evidence to support an autoimmune pathogenesis in this type of hepatitis includes the following: (1) in the liver, the histopathologic lesions are composed predominantly of cytotoxic T cells and plasma cells; (2) circulating autoantibodies (nuclear, smooth muscle, thyroid, etc.; see below), rheumatoid factor, and hyperglobulinemia are common; (3) other autoimmune disorders—such as autoimmune thyroiditis, rheumatoid arthritis, autoimmune hemolytic anemia, ulcerative colitis, membranoproliferative glomerulonephritis, juvenile diabetes mellitus, vitiligo, celiac disease, and Sjögren's syndrome—occur with increased frequency in patients and in their relatives who have autoimmune hepatitis; (4) histocompatibility haplotypes associated with autoimmune diseases, such as HLA-B1, B8, DR3, and DR4 as well as extended haplotype *DRB1*0301* and *DRB1*0401* alleles, are common in patients with autoimmune hepatitis; and (5) this type of chronic hepatitis is responsive to glucocorticoid/immunosuppressive therapy, effective in a variety of autoimmune disorders.

Cellular immune mechanisms appear to be important in the pathogenesis of autoimmune hepatitis. In vitro studies have suggested that in patients with this disorder, $CD4^+$ T lymphocytes are capable of becoming sensitized to hepatocyte membrane proteins and of destroying liver cells. Molecular mimicry by cross-reacting antigens that contain epitopes similar to liver antigens is postulated to activate these T cells, which infiltrate, and result in injury to, the liver. Abnormalities of immunoregulatory control over cytotoxic lymphocytes (impaired regulatory CD4+CD25+ T cell influences) may play a role as well. Studies of genetic predisposition to autoimmune hepatitis demonstrate that certain haplotypes are associated with the disorder, as enumerated above, as are polymorphisms in cytotoxic T lymphocyte antigens (*CTLA-4*) and tumor necrosis factor α (*TNFA*2*). The precise triggering factors, genetic influences, and cytotoxic and immunoregulatory mechanisms involved in this type of liver injury remain incompletely defined.

Intriguing clues into the pathogenesis of autoimmune hepatitis come from the observation that circulating autoantibodies are prevalent in patients with this disorder. Among the autoantibodies described in these patients are antibodies to nuclei (so-called antinuclear antibodies [ANAs], primarily in a homogeneous pattern) and smooth muscle (so-called anti-smooth-muscle antibodies, directed at actin, vimentin, and skeletin), antibodies to F-actin, anti-LKM (see below), antibodies to "soluble liver antigen" (directed against a uracil-guanine-adenine transfer RNA suppressor protein), antibodies to α-actinin, and antibodies to the liver-specific asialoglycoprotein receptor (or "hepatic lectin") and other hepatocyte membrane proteins. Although some of these provide helpful diagnostic markers, their involvement in the pathogenesis of autoimmune hepatitis has not been established.

Humoral immune mechanisms have been shown to play a role in the extrahepatic manifestations of autoimmune and idiopathic hepatitis. Arthralgias, arthritis, cutaneous vasculitis, and glomerulonephritis occurring in patients with autoimmune hepatitis appear to be mediated by the deposition of circulating immune complexes in affected tissue vessels, followed by complement activation, inflammation, and tissue injury. While specific viral antigen-antibody complexes can be identified in acute and chronic viral hepatitis, the nature of the immune complexes in autoimmune hepatitis has not been defined.

CLINICAL FEATURES

Many of the *clinical features* of autoimmune hepatitis are similar to those described for chronic viral hepatitis. The onset of disease may be insidious or abrupt; the disease may present initially like, and be confused with, acute viral hepatitis; a history of recurrent bouts of what had been labeled *acute hepatitis* is not uncommon. In approximately a quarter of patients, the diagnosis is made in the absence of symptoms, based on abnormal liver laboratory tests. A subset of patients with autoimmune hepatitis has distinct features. Such patients are predominantly young to middle-aged women with marked hyperglobulinemia and high-titer circulating ANAs. This is the group with positive lupus erythematosus (LE) preparations (initially labeled "*lupoid*" hepatitis) in whom other autoimmune features are common. Fatigue, malaise, anorexia, amenorrhea, acne, arthralgias, and jaundice are common. Occasionally, arthritis, maculopapular eruptions (including cutaneous vasculitis), erythema nodosum, colitis, pleurisy, pericarditis, anemia, azotemia, and sicca syndrome (keratoconjunctivitis, xerostomia) occur. In some patients, complications of cirrhosis, such as ascites and edema (associated with portal hypertension and hypoalbuminemia), encephalopathy, hypersplenism, coagulopathy, or variceal bleeding may bring the patient to initial medical attention.

The course of autoimmune hepatitis may be variable. In patients with mild disease or limited histologic lesions (e.g., piecemeal necrosis without bridging), progression to cirrhosis is limited, but, even in this subset, clinical monitoring is important to identify progression; up to half left untreated can progress to cirrhosis over the course of 15 years. In North America, cirrhosis at presentation is more common in African Americans than in whites. In those with severe symptomatic autoimmune hepatitis (aminotransferase levels >10 times normal, marked hyperglobulinemia, "aggressive" histologic lesions—bridging necrosis or multilobular collapse, cirrhosis), the 6-month mortality without therapy may be as high as 40%. Such severe disease accounts for only 20% of cases; the natural history of milder disease is variable, often accentuated by spontaneous remissions and exacerbations. Especially poor prognostic signs include the presence histologically of multilobular collapse at the time of initial presentation and failure of serum bilirubin to improve after 2 weeks of therapy. Death may result from hepatic failure, hepatic coma, other complications of cirrhosis (e.g., variceal hemorrhage), and intercurrent infection. In patients with established cirrhosis, HCC may be a late complication **(Chap. 78)** but occurs less frequently than in cirrhosis associated with viral hepatitis.

Laboratory features of autoimmune hepatitis are similar to those seen in chronic viral hepatitis. Liver biochemical tests are invariably abnormal but may not correlate with the clinical severity or histopathologic features in individual cases. Many patients with autoimmune hepatitis have normal serum bilirubin, alkaline phosphatase, and globulin levels with only minimal aminotransferase elevations. Serum AST and ALT levels are increased and fluctuate in the range of 100−1000 units. In severe cases, the serum bilirubin level is moderately elevated (51−171 µmol/L [3−10 mg/dL]). Hypoalbuminemia occurs in patients with very active or advanced disease. Serum alkaline phosphatase levels may be moderately elevated or near normal. In a small proportion of patients, marked elevations of alkaline phosphatase activity occur; in such patients, clinical and laboratory features overlap with those of primary biliary cirrhosis **(Chap. 337)**. The prothrombin time is often prolonged, particularly late in the disease or during active phases.

Polyclonal hypergammaglobulinemia (>2.5 g/dL) is common in autoimmune hepatitis, as is the presence of rheumatoid factor. As noted above, circulating autoantibodies are also prevalent, most characteristically ANAs in a homogeneous staining pattern. Smooth-muscle antibodies are less specific, seen just as frequently in

chronic viral hepatitis. Because of the high levels of globulins achieved in the circulation of some patients with autoimmune hepatitis, occasionally the globulins may bind nonspecifically in solid-phase binding immunoassays for viral antibodies. This has been recognized most commonly in tests for antibodies to hepatitis C virus, as noted above. In fact, studies of autoantibodies in autoimmune hepatitis have led to the recognition of new categories of autoimmune hepatitis. *Type I autoimmune hepatitis* is the classic syndrome prevalent in North America and northern Europe occurring in young women, associated with marked hyperglobulinemia, lupoid features, circulating ANAs, and HLA-DR3 or HLA-DR4 (especially *B8-DRB1*03*). Also associated with type I autoimmune hepatitis are autoantibodies against actin and atypical perinuclear antineutrophilic cytoplasmic antibodies (pANCA). Included in the spectrum of type-I autoimmune hepatitis is a subset of patients who lack ANA and anti-LKM1, but who have circulating antibodies to soluble liver antigen. Most of these patients are women and have clinical features similar to, perhaps more severe than, those of other patients with type I autoimmune hepatitis.

Type II autoimmune hepatitis, often seen in children, more common in Mediterranean populations, and linked to HLA- DRB1 and HLA-DQB1 haplotypes, is associated not with ANA but with anti-LKM. Actually, anti-LKM represent a heterogeneous group of antibodies. In type II autoimmune hepatitis, the antibody is anti-LKM1, directed against cytochrome P450 2D6. This is the same anti-LKM seen in some patients with chronic hepatitis C. Anti-LKM2 is seen in drug-induced hepatitis, and anti-LKM3 (directed against uridine diphosphate glucuronyltransferases) is seen in patients with chronic hepatitis D. Another autoantibody observed in type II autoimmune hepatitis is directed against liver cytosol formiminotransferase cyclodeaminase (anti-liver cytosol 1).

Liver biopsy abnormalities are similar to those described for chronic viral hepatitis. Expanding portal tracts and extending beyond the plate of periportal hepatocytes into the parenchyma (designated *interface hepatitis* or *piecemeal necrosis*) is a mononuclear cell infiltrate that, in autoimmune hepatitis, may include the presence of plasma cells. Necroinflammatory activity characterizes the lobular parenchyma, and evidence of hepatocellular regeneration is reflected by “rosette” formation, the occurrence of thickened liver cell plates, and regenerative “pseudolobules.” Septal fibrosis, bridging fibrosis, and cirrhosis are frequent. In patients with early autoimmune hepatitis presenting as an acute-hepatitis-like illness, lobular and centrilobular (as opposed to the more common periportal) necrosis has been reported. Bile duct injury and granulomas are uncommon; however, a subgroup of patients with autoimmune hepatitis has histologic, biochemical, and serologic features overlapping those of primary biliary cirrhosis **(Chap. 337)**.

DIAGNOSTIC CRITERIA

An international group has suggested a set of criteria for establishing a diagnosis of autoimmune hepatitis. Exclusion of liver disease caused by genetic disorders, viral hepatitis, drug hepatotoxicity, and alcohol are linked with such inclusive diagnostic criteria as hyperglobulinemia, autoantibodies, and characteristic histologic features. This international group has also suggested a comprehensive diagnostic scoring system that, rarely required for typical cases, may be helpful when typical features are not present. Factors that weigh in favor of the diagnosis include female gender; predominant aminotransferase elevation; presence and level of globulin elevation; presence of nuclear, smooth muscle, LKM1, and other autoantibodies; concurrent other

autoimmune diseases; characteristic histologic features (interface hepatitis, plasma cells, rosettes); HLA-DR3 or DR4 markers; and response to treatment (see below). A more simplified, more specific scoring system relies on four variables: autoantibodies, serum IgG level, typical or compatible histologic features, and absence of viral hepatitis markers. Weighing against the diagnosis are predominant alkaline phosphatase elevation, mitochondrial antibodies, markers of viral hepatitis, history of hepatotoxic drugs or excessive alcohol, histologic evidence of bile duct injury, or such atypical histologic features as fatty infiltration, iron overload, and viral inclusions.

DIFFERENTIAL DIAGNOSIS

Early during the course of chronic hepatitis, autoimmune hepatitis may resemble typical *acute viral hepatitis* (**Chap. 332**). Without histologic assessment, severe chronic hepatitis cannot be readily distinguished based on clinical or biochemical criteria from mild chronic hepatitis. In adolescence, *Wilson's disease* (**Chaps. 337 and 408**) may present with features of chronic hepatitis long before neurologic manifestations become apparent and before the formation of Kayser-Fleischer rings (copper deposition in Descemet's membrane in the periphery of the cornea). In this age group, serum ceruloplasmin and serum and urinary copper determinations plus measurement of liver copper levels establish the correct diagnosis. *Postnecrotic* or *cryptogenic cirrhosis* and *primary biliary cirrhosis* (**Chap. 337**) share clinical features with autoimmune hepatitis, and both alcoholic hepatitis (**Chap. 335**) and nonalcoholic steatohepatitis (**Chap. 336**) may present with many features common to autoimmune hepatitis; historic, biochemical, serologic, and histologic assessments are usually sufficient to allow these entities to be distinguished from autoimmune hepatitis. Of course, the distinction between autoimmune and chronic viral hepatitis is not always straightforward, especially when viral antibodies occur in patients with autoimmune disease or when autoantibodies occur in patients with viral disease. Furthermore, the presence of extrahepatic features such as arthritis, cutaneous vasculitis, or pleuritis—not to mention the presence of circulating autoantibodies—may cause confusion with *rheumatologic disorders* such as rheumatoid arthritis and systemic LE. The existence of clinical and biochemical features of progressive necroinflammatory liver disease distinguishes chronic hepatitis from these other disorders, which are not associated with severe liver disease. Rarely, hepatic venous outflow obstruction (Budd-Chiari syndrome) may present with features suggestive of autoimmune hepatitis, but painful hepatomegaly, ascites, and vascular imaging provide distinguishing diagnostic clues. Other diagnostic considerations would include celiac disease and ischemic liver disease, which would be readily distinguishable by clinical and laboratory features from autoimmune hepatitis.

Finally, occasionally, features of autoimmune hepatitis overlap with features of autoimmune biliary disorders such as primary biliary cirrhosis, primary sclerosing cholangitis (**Chaps. 337 and 339**), or, even more rarely, mitochondrial antibody-negative autoimmune cholangitis. Such overlap syndromes are difficult to categorize, and often response to therapy may be the distinguishing factor that establishes the diagnosis.

TREATMENT

TREATMENT: Autoimmune Hepatitis

The mainstay of management in autoimmune hepatitis is glucocorticoid therapy. Several controlled clinical trials have documented that such therapy leads to symptomatic, clinical, biochemical, and histologic improvement as well as increased survival. A therapeutic response can be expected in up to 80% of patients. Unfortunately, therapy has not been shown in clinical trials to prevent ultimate progression to cirrhosis; however, instances of reversal of fibrosis and cirrhosis have been reported in patients responding to treatment, and rapid treatment responses within 1 year do translate into a reduction in progression to cirrhosis. Although some advocate the use of prednisolone (the hepatic metabolite of prednisone), prednisone is just as effective and is favored by most authorities. Therapy may be initiated at 20 mg/d, but a popular regimen in the United States relies on an initiation dose of 60 mg/d. This high dose is tapered successively over the course of a month down to a maintenance level of 20 mg/d. An alternative, but equally effective, more appealing approach is to begin with half the prednisone dose (30 mg/d) along with azathioprine (50 mg/d). With azathioprine maintained at 50 mg/d, the prednisone dose is tapered over the course of a month down to a maintenance level of 10 mg/d. The advantage of the combination approach is a reduction, over the span of an 18-month course of therapy, in serious, life-threatening complications of steroid therapy (e.g., cushingoid features, hypertension, diabetes, osteoporosis) from 66% down to under 20%. Genetic analysis for thiopurine S-methyltransferase allelic variants does not correlate with azathioprine-associated cytopenias or efficacy and is not assessed routinely in patients with autoimmune hepatitis. In combination regimens, 6-mercaptopurine may be substituted for its prodrug azathioprine, but this is rarely required. Azathioprine alone, however, is not effective in achieving remission, nor is alternate-day glucocorticoid therapy. Limited experience with budesonide in noncirrhotic patients suggests that this steroid side effect–sparing drug may be effective; however, the few randomized controlled trials of budesonide have not consistently shown efficacy. Although therapy has been shown to be effective for severe autoimmune hepatitis (AST ≥10 × the upper limit of normal or ≥5 × the upper limit of normal in conjunction with serum globulin greater than or equal to twice normal; bridging necrosis or multilobular necrosis on liver biopsy; presence of symptoms), therapy is not indicated for mild forms of chronic hepatitis, and the efficacy of therapy in mild or asymptomatic autoimmune hepatitis has not been established.

Improvement of fatigue, anorexia, malaise, and jaundice tends to occur within days to several weeks; biochemical improvement occurs over the course of several weeks to months, with a fall in serum bilirubin and globulin levels and an increase in serum albumin. Serum aminotransferase levels usually drop promptly, but improvements in AST and ALT alone do not appear to be reliable markers of recovery in individual patients; histologic improvement, characterized by a decrease in mononuclear infiltration and in hepatocellular necrosis, may be delayed for 6–24 months. Still, if interpreted cautiously, aminotransferase levels are valuable indicators of relative disease activity, and, although recommended, many authorities do *not* advocate for serial liver biopsies to assess therapeutic success or to guide decisions to alter or stop therapy. Rapidity of response is more common in older patients (≥69 years) and those with HLA *DBR1*04*; although rapid responders may progress less slowly to cirrhosis and liver transplantation, they are no less likely than slower responders to relapse after therapy. Therapy should continue for at least 12–18 months. After tapering and cessation of therapy, the likelihood of relapse is at least 50%, even if posttreatment histology has improved to show mild chronic hepatitis, and the majority of patients require therapy at maintenance doses indefinitely. Continuing azathioprine alone (2 mg/kg body weight daily) after cessation of prednisone therapy has been shown to

reduce the frequency of relapse. Long-term maintenance with low-dose prednisone (≤10 mg daily) has also been shown to keep autoimmune hepatitis in check without the theoretical risk of azathioprine marrow suppression and, in young women of child-bearing age, teratogenicity; however, maintenance azathioprine is more effective in preserving remission.

In medically refractory cases, an attempt should be made to intensify treatment with high-dose glucocorticoid monotherapy (60 mg daily) or combination glucocorticoid (30 mg daily) plus high-dose azathioprine (150 mg daily) therapy. After a month, doses of prednisone can be reduced by 10 mg a month, and doses of azathioprine can be reduced by 50 mg a month toward ultimate, conventional maintenance doses. Patients refractory to this regimen may be treated with cyclosporine, tacrolimus, or mycophenolate mofetil. Similarly, in exploratory studies, infusions of monoclonal antibodies directed at tumor necrosis factor (infliximab) and against the B-lymphocyte antigen CD20 (rituximab) have been reported to be of clinical benefit (improved aminotransferase levels, immunoglobulin G levels, histologic inflammatory activity) as rescue therapy for refractory autoimmune hepatitis. To date, however, only limited, often anecdotal, data in small numbers of patients support these alternative approaches. If medical therapy fails, or when chronic hepatitis progresses to cirrhosis and is associated with life-threatening complications of liver decompensation, liver transplantation is the only recourse **(Chap. 338)**; in patients with severe autoimmune hepatitis, failure of the bilirubin to improve after 2 weeks of therapy should prompt early consideration of the patient for liver transplantation. Recurrence of autoimmune hepatitis in the new liver occurs rarely in most experiences but in as many as 35–40% of cases in others; nonetheless, 5-year patient and graft survival exceed 80%.

Like all patients with chronic liver disease, patients with autoimmune hepatitis should be vaccinated against hepatitis A and B, ideally before immunosuppressive therapy is begun, if practical. Patients with autoimmune hepatitis and cirrhosis should be screened for HCC with ultrasound at 6-month intervals and for gastroesophageal varices with upper gastrointestinal endoscopy at intervals of 1–3 years, based on severity of liver disease.

FURTHER READING

AASLD/IDSA HCV GUIDANCE PANEL: Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology 62:932, 2015. Updated regularly and available at *http://www.hcvguidelines.org*. Accessed September 22, 2016. [PubMed: 26111063]

Carbone M, Neuberger JM: Autoimmune liver disease, autoimmunity and liver transplantation. J Hepatol 60:210, 2014. [PubMed: 24084655]

European Association for the Study of the Liver: EASL clinical practice guidelines: Management of chronic hepatitis B virus infection. J Hepatol 57:167, 2012. [PubMed: 22436845]

European Association for the Study of the Liver: EASL clinical practice guidelines: autoimmune hepatitis. J Hepatol 63:971, 2015.

European Association for the Study of the Liver: EASL Recommendations on treatment of hepatitis C 2016. J Hepatol 2016 in press. Available at http://dx.doi.org.ezproxy1.library.usyd.edu.au/10.1016/j. jhep.2016.09.001. Accessed September 22, 2016.

Lok ASG et al: Antiviral therapy for chronic hepatitis B viral infection in adults: A systematic review and meta-analysis. Hepatology 63:284, 2016. [PubMed: 26566246]

Manns MP et al: AASLD Practice Guidelines: Diagnosis and management of autoimmune hepatitis. Hepatology 51;2193, 2010.

Pawlotsky J-M et al: From non-A, non-B hepatitis to hepatitis C virus cure. J Hepatol 62:S87, 2015. [PubMed: 25920094]

Terrault N et al: AASLD guidelines for treatment of chronic hepatitis B. Hepatology 63:261, 2016. [PubMed: 26566064]

Wedemeyer H et al: Peginterferon plus adefovir versus either drug alone for hepatitis delta. N Engl J Med 364:322, 2011. [PubMed: 21268724]

ADDITIONAL ONLINE BIBLIOGRAPHY

Charlton M et al: Ledipasvir and sofosbuvir plus ribavirin for treatment of HCV infection in patients with advanced liver disease. Gastroenterology 149:649, 2015. [PubMed: 25985734]

Chhatwal J et al: Hepatitis C disease burden in the United States in the era of oral direct-acting antivirals. Hepatology 64:1442, 2016. [PubMed: 27015107]

Curry MP et al: Sofosbuvir and velpatasvir for HCV in patients with decompensated cirrhosis. N Engl J Med 373:2618, 2015. [PubMed: 26569658]

Czaja AJ: Rapidity of treatment response and outcome in type 1 autoimmune hepatitis. J Hepatol 51:161, 2009. [PubMed: 19446908]

Dienstag JL: Hepatitis B virus infection. N Engl J Med 359:1486, 2008. [PubMed: 18832247]

Feld JJ et al: Sofosbuvir and velpatasvir for HCV genotype 1, 2, 4, 5, and 6 infection. N Engl J Med 373:2599, 2015. [PubMed: 26571066]

Foster GR et al: Sofosbuvir and velpatasvir for HCV genotype 2 and 3 infection. N Engl J Med 373:2608, 2015. [PubMed: 26575258]

Foster GR et al: Impact of direct acting antiviral therapy in patients with chronic hepatitis C and decompensated cirrhosis. J Hepatol 64:1224, 2016. [PubMed: 26829205]

Heller T et al: Long-term therapy of chronic delta hepatitis with peginterferon alfa. Aliment Pharmacol Ther 40:93, 2014. [PubMed: 24815494]

Liberal R et al: Update on autoimmune hepatitis. J Clin Trans Hepatol 3:42, 2015.

Paul S et al: Hepatitis B virus reactivation and prophylaxis during solid tumor chemotherapy: A systematic review and meta-analysis. Ann Intern Med 164:30, 2016. [PubMed: 26595058]

Pawlotsky JM: Hepatitis C virus resistance to direct-acting antiviral drugs in interferon-free regimens. Gastroenterology 151:70, 2016. [PubMed: 27080301]

Perrillo RP et al: American Gastroenterological Association Institute technical review on prevention and treatment of hepatitis B virus reactivation during immunosuppressive drug therapy. Gastroenterology 148:221, 2015. [PubMed: 25447852]

Reddy KR et al: American Gastroenterological Association Institute guideline on the prevention and treatment of hepatitis B virus reactivation during immunosuppressive drug therapy. Gastroenterology 148:215, 2015. [PubMed: 25447850]

Rizzetto M: Hepatitis D: Thirty years after. J Hepatol 50:1043, 2009. [PubMed: 19285743]

Rosen HR: Chronic hepatitis C infection. N Engl J Med 364:2429, 2011. [PubMed: 21696309]

Sarrazin C: The importance of resistance to direct antiviral drugs in HCV infection in clinical practice. J Hepatol 64:486, 2016. [PubMed: 26409317]

Sarrazin C et al: Prevalence of resistance-associated substitutions in HCV NS5A, NS5B, or NS3 and outcomes of treatment with ledipasvir and sofosbuvir. Gastroenterology 151:501, 2016. [PubMed: 27296509]

Summerskill WHJ et al: Prednisone for chronic active liver disease: Dose titration, standard dose, and combination with azathioprine compared. Gut 16:876, 1975. [PubMed: 1104411]

Van der meer AJ et al: Association between sustained virological response and all-cause mortality among patients with chronic hepatitis C and advanced hepatic fibrosis. JAMA 308:2584, 2012. [PubMed: 23268517]

Webster DP et al: Hepatitis C. Lancet 385:1124, 2015. [PubMed: 25687730]

Chapter 335: Alcoholic Liver Disease

Mark E. Mailliard; Michael F. Sorrell

INTRODUCTION

Chronic and excessive alcohol ingestion is a major cause of liver disease and is responsible for nearly 50% of the mortality from all cirrhosis. The pathology of alcoholic liver disease consists of three major lesions, with the progressive injury rarely existing in a pure form: (1) fatty liver, (2) alcoholic hepatitis, and (3) cirrhosis. Fatty liver is present in >90% of daily as well as binge drinkers. A much smaller percentage of heavy drinkers will progress to alcoholic hepatitis, thought to be a precursor to cirrhosis. The prognosis of severe alcoholic liver disease is dismal; the mortality of patients with alcoholic hepatitis concurrent with cirrhosis is nearly 60% at 4 years. Although alcohol is considered a direct hepatotoxin, only between 10 and 20% of alcoholics will develop alcoholic hepatitis. The explanation for this apparent paradox is unclear but involves the complex interaction of facilitating factors such as drinking patterns, diet, obesity, and gender. There are no diagnostic tools that can predict individual susceptibility to alcoholic liver disease.

GLOBAL CONSIDERATIONS

Alcohol is the world's third largest risk factor for disease burden. The harmful use of alcohol results in about 3.5 million deaths worldwide each year. Most of the mortality attributed to alcohol is secondary to cirrhosis. Mortality from cirrhosis is directly related to alcohol consumption, with the Eastern European countries the most significantly burdened. Cirrhosis and its complications are closely correlated with volume of alcohol consumed per capita population and are regardless of gender.

ETIOLOGY AND PATHOGENESIS

Quantity and duration of alcohol intake are the most important risk factors involved in the development of alcoholic liver disease **(Table 335-1)**. The roles of beverage type(s), i.e., wine, beer, or spirits, and pattern of drinking (daily versus binge drinking) are less clear. Progress beyond the fatty liver stage seems to require additional risk factors that remain incompletely defined. Although there are genetic predispositions for alcoholism **(Chap. 445)**, gender is a strong determinant for alcoholic liver disease. Women are more susceptible to alcoholic liver injury when compared to men. They develop advanced liver disease with substantially less alcohol intake. In general, the time it takes to develop liver disease is directly related to the amount of alcohol consumed. It is useful in estimating alcohol consumption to understand that one beer,

four ounces of wine, or one ounce of 80% spirits all contain ~12 g of alcohol. The threshold for developing alcoholic liver disease is higher in men (>14 drinks per week), while women are at increased risk for liver injury by consuming >7 drinks per week. Gender-dependent differences result from poorly understood effects of estrogen, proportion of body fat, and the gastric metabolism of alcohol. Obesity, a high-fat diet, and the protective effect of coffee have been postulated to play a part in the development of the pathogenic process.

TABLE 335-1

Risk Factors for Alcoholic Liver Disease

Risk Factor	Comment
Quantity	In men, 40–80 g/d of ethanol produces fatty liver; 160 g/d for 10–20 years causes hepatitis or cirrhosis. Only 15% of alcoholics develop alcoholic liver disease.
Gender	Women exhibit increased susceptibility to alcoholic liver disease at amounts >20 g/d; two drinks per day is probably safe.
Hepatitis C	HCV infection concurrent with alcoholic liver disease is associated with younger age for severity, more advanced histology, and decreased survival.
Genetics	Patatin-like phospholipase domain-containing protein 3 (PNPLA3) has been associated with alcoholic cirrhosis.
Fatty liver	Alcohol injury does not require malnutrition, but obesity and nonalcoholic fatty liver are risk factors. Patients should receive vigorous attention to nutritional support.

Chronic infection with hepatitis C virus (HCV) **(Chap. 334)** is an important comorbidity in the progression of alcoholic liver disease to cirrhosis in chronic drinkers. Even light to moderate alcohol intake of 15–30 g/d increases the risk of cirrhosis and hepatocellular cancer in HCV-infected individuals. Patients with both alcoholic liver injury and HCV infection develop decompensated liver disease at a younger age and have poorer overall survival. Increased liver iron stores and, rarely, porphyria cutanea tarda can occur as a consequence of the overlapping injurious processes secondary to alcohol and HCV infection.

The pathogenesis of alcoholic liver injury is unclear. The present conceptual foundation is that alcohol acts as a direct hepatotoxin and that malnutrition does not have a major role. Ingestion of alcohol initiates an inflammatory cascade by its metabolism, resulting in a variety of metabolic responses. Steatosis from lipogenesis, fatty acid synthesis, and depression of fatty acid oxidation appears secondary to effects on sterol regulatory transcription factor and peroxisome proliferator-activated receptor α (PPAR-α). Intestinal-derived

endotoxin initiates a pathogenic process through toll-like receptor 4 and tumor necrosis factor α (TNF-α) that facilitates hepatocyte apoptosis and necrosis. The cell injury and endotoxin release initiated by ethanol and its metabolites also activate innate and adaptive immunity pathways releasing proinflammatory cytokines (e.g., TNF-α), chemokines, and proliferation of T and B cells. The effect of chronic ethanol ingestion on intestinal permeability influences liposaccharide hepatic influx as well as microbiome dysbiosis, further contributing to the pathogenic process. The production of toxic protein-aldehyde adducts, generation of reducing equivalents, and oxidative stress also play a role. Hepatocyte injury and impaired regeneration following chronic alcohol ingestion are ultimately associated with stellate cell activation and collagen production; key events in fibrogenesis. The resulting fibrosis from continuing alcohol use determines the architectural derangement of the liver and associated pathophysiology.

PATHOLOGY

The liver has a limited repertoire in response to injury. Fatty liver is the initial and most common histologic response to hepatotoxic stimuli, including excessive alcohol ingestion. The accumulation of fat within the perivenular hepatocytes coincides with the location of alcohol dehydrogenase, the major enzyme responsible for alcohol metabolism. Continuing alcohol ingestion results in fat accumulation throughout the entire hepatic lobule. Despite extensive fatty change and distortion of the hepatocytes with macrovesicular fat, the cessation of drinking results in normalization of hepatic architecture and fat content. Alcoholic fatty liver has traditionally been regarded as entirely benign, but similar to the spectrum of nonalcoholic fatty liver disease, the appearance of steatohepatitis and certain pathologic features such as giant mitochondria, perivenular fibrosis, and macrovesicular fat may be associated with progressive liver injury.

The transition between fatty liver and the development of alcoholic hepatitis is blurred. The hallmark of alcoholic hepatitis is hepatocyte injury characterized by ballooning degeneration, spotty necrosis, polymorphonuclear infiltrate, and fibrosis in the perivenular and perisinusoidal space of Disse. Mallory-Denk bodies are often present in florid cases but are neither specific nor necessary to establish the diagnosis. Alcoholic hepatitis is thought to be a precursor to the development of cirrhosis. However, like fatty liver, it is potentially reversible with cessation of drinking. Cirrhosis is present in up to 50% of patients with biopsy-proven alcoholic hepatitis, and its regression is uncertain, even with abstention.

CLINICAL FEATURES

The clinical manifestations of alcoholic fatty liver are subtle and characteristically detected as a consequence of the patient's visit for a seemingly unrelated matter. Previously unsuspected hepatomegaly is often the only clinical finding. Occasionally, patients with fatty liver will present with right upper quadrant discomfort, nausea, and, rarely, jaundice. Differentiation of alcoholic fatty liver from nonalcoholic fatty liver is difficult unless an accurate drinking history is ascertained. In every instance where liver disease is present, a thoughtful and sensitive drinking history should be obtained. Standard, validated questions accurately detect alcohol-related problems **(Chap. 445)**. Alcoholic hepatitis is associated with a wide gamut of clinical features. Fever, spider nevi, jaundice, and abdominal pain simulating an acute abdomen represent the

extreme end of the spectrum, while many patients will be entirely asymptomatic. Portal hypertension, ascites, or variceal bleeding can occur in the absence of cirrhosis. Recognition of the clinical features of alcoholic hepatitis is central to the initiation of an effective and appropriate diagnostic and therapeutic strategy. It is important to recognize that patients with alcoholic cirrhosis often exhibit clinical features identical to other causes of cirrhosis.

LABORATORY FEATURES

Patients with alcoholic liver disease are often identified through routine screening tests. The typical laboratory abnormalities seen in fatty liver are nonspecific and include modest elevations of aspartate aminotransferase (AST), alanine aminotransferase (ALT), and γ-glutamyl transpeptidase (GGTP), often accompanied by hypertriglyceridemia and hyperbilirubinemia. In alcoholic hepatitis and in contrast to other causes of fatty liver, AST and ALT are usually elevated two- to sevenfold. They are rarely >400 IU, and the AST/ALT ratio is >1 **(Table 335-2)**. Hyperbilirubinemia is accompanied by modest increases in the alkaline phosphatase level. Derangement in hepatocyte synthetic function indicates more serious disease. Hypoalbuminemia and coagulopathy are common in advanced liver injury. Ultrasonography is useful in detecting fatty infiltration of the liver and determining liver size. The demonstration by ultrasound of portal vein flow reversal, ascites, and intraabdominal venous collaterals indicates serious liver injury with less potential for complete reversal.

TABLE 335-2

Laboratory Diagnosis of Alcoholic Fatty Liver and Alcoholic Hepatitis

Test	Comment
AST	Increased two- to sevenfold, <400 IU/L, greater than ALT
ALT	Increased two- to sevenfold, <400 IU/L
AST/ALT	Usually >1
GGTP	Not specific to alcohol, easily inducible, elevated in all forms of fatty liver
Bilirubin	May be markedly increased in alcoholic hepatitis despite modest elevation in alkaline phosphatase

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGTP, γ-glutamyl transpeptidase.

PROGNOSIS

Critically ill patients with alcoholic hepatitis have short-term (30-day) mortality rates >50%. Severe alcoholic hepatitis is heralded by coagulopathy (prothrombin time increased >5 s), anemia, serum albumin concentrations <25 g/L (2.5 mg/dL), serum bilirubin levels >137 µmol/L (8 mg/dL), renal failure, and ascites. A discriminant function calculated as 4.6 X (the prolongation of the prothrombin time above control [seconds]) + serum bilirubin (mg/dL) can identify patients with a poor prognosis (discriminant function >32). A Model for End-Stage Liver Disease (MELD) score **(Chap. 338)** ≥21 also is associated with significant mortality in alcoholic hepatitis. The presence of ascites, variceal hemorrhage, deep encephalopathy, or hepatorenal syndrome predicts a dismal prognosis. The pathologic stage of the injury can be helpful in predicting prognosis. Liver biopsy should be performed whenever possible to establish the diagnosis and to guide the therapeutic decisions.

TREATMENT

TREATMENT

Alcoholic Liver Disease

Complete abstinence from alcohol is the cornerstone in the treatment of alcoholic liver disease. Improved survival and the potential for reversal of histologic injury regardless of the initial clinical presentation are associated with total avoidance of alcohol ingestion. Referral of patients to experienced alcohol counselors and/or alcohol treatment programs should be routine in the management of patients with alcoholic liver disease. Attention should be directed to the nutritional and psychosocial states during the evaluation and treatment periods. Because of data suggesting that the pathogenic mechanisms in alcoholic hepatitis involve cytokine release and the perpetuation of injury by immunologic processes, glucocorticoids have been extensively evaluated in the treatment of alcoholic hepatitis. Patients with severe alcoholic hepatitis, defined as a discriminant function >32 or MELD >20, should be given prednisone, 40 mg/d, or prednisolone, 32 mg/d, for 4 weeks, followed by a steroid taper **(Fig. 335-1)**. Exclusion criteria include active gastrointestinal bleeding, renal failure, or pancreatitis. Patients with infection can be concurrently treated with antibiotics and steroids. Women with encephalopathy from severe alcoholic hepatitis may be particularly good candidates for glucocorticoids. A Lille score >0.45, at *http://www.lillemodel.com*, uses pretreatment variables plus the change in total bilirubin at day 7 of glucocorticoids to identify those patients unresponsive to therapy.

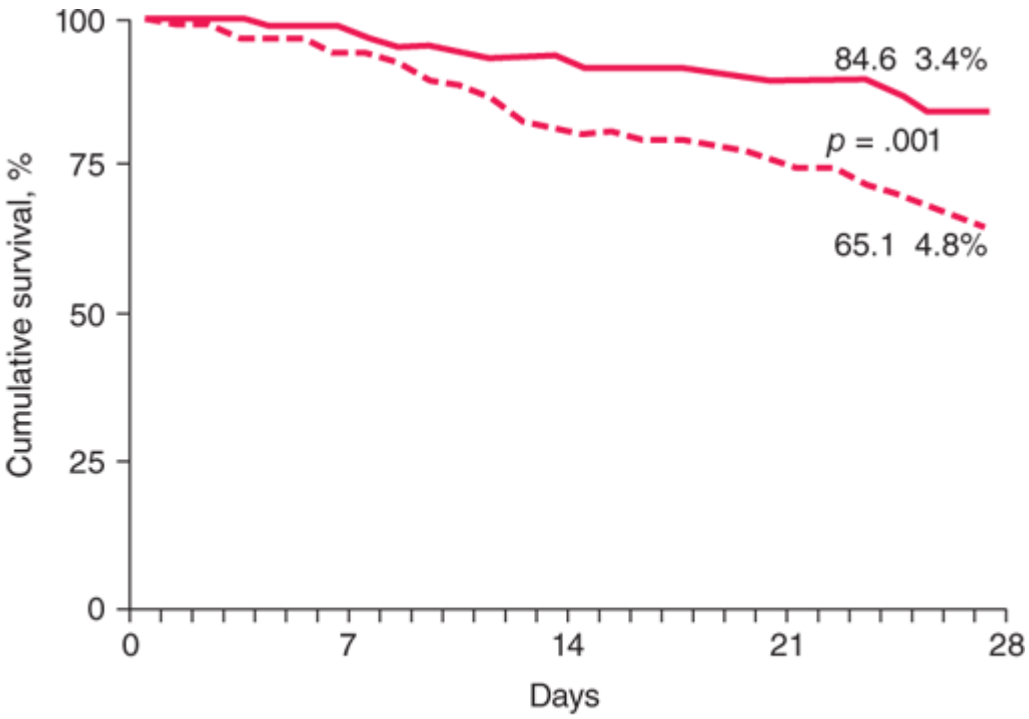
The role of TNF-α expression and receptor activity in alcoholic liver injury has led to an examination of pentoxifylline, the nonspecific TNF inhibitor, either by itself, or with glucocorticoids for severe alcoholic hepatitis. In one study, pentoxifylline demonstrated an improved survival in the therapy of severe alcoholic hepatitis, primarily due to a decrease in hepatorenal syndrome. Subsequent clinical trials failed to find an increased benefit from pentoxifylline, either by itself or in combination with prednisolone **(Fig. 335-2)**.

Liver transplantation is an accepted indication for treatment in select patients with complications of cirrhosis secondary to alcohol abuse. Outcomes are equal or superior to other indications for transplantation. In general, transplant candidacy should be reevaluated after a defined period of sobriety. Patients presenting

with alcoholic hepatitis have been largely excluded from transplant candidacy because of the perceived risk of increased surgical mortality and high rates of recidivism following transplantation. A European multidisciplinary group has reported excellent long-term transplant outcomes in highly selected patients with florid alcoholic hepatitis. General application of transplantation in such patients must await confirmatory outcomes.

FIGURE 335-1

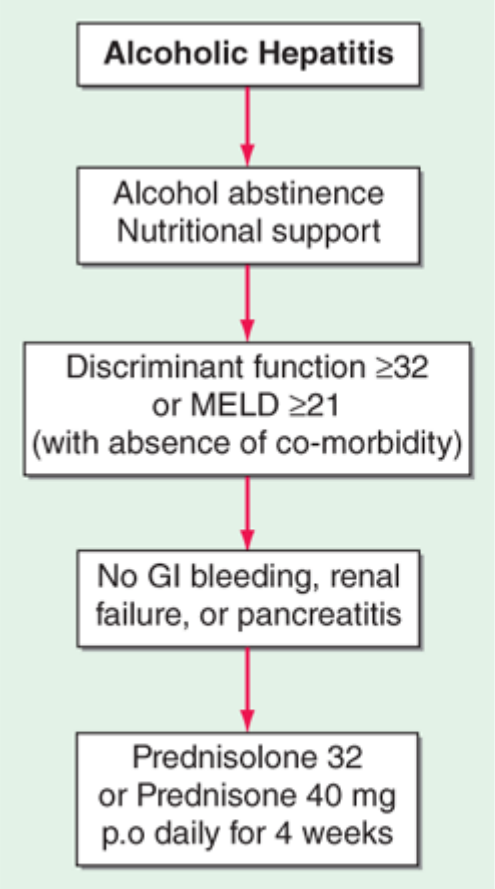
Effect of glucocorticoid therapy of severe alcoholic hepatitis on short-term survival: the result of a meta-analysis of individual data from three studies. Prednisolone, solid line; placebo, dotted line. *(Adapted from P Mathurin et al: J Hepatol 36:480, 2002, with permission from Elsevier Science.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 335-2

Treatment algorithm for alcoholic hepatitis. As identified by a calculated discriminant function >32 or MELD >20 (see text), patients with severe alcoholic hepatitis, without the presence of gastrointestinal bleeding, renal failure or pancreatitis would be candidates for glucocorticoids.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Mathurin P et al: Corticosteroids improves short-term survival in patients with severe alcoholic hepatitis (AH): Individual data analysis of the last three randomized placebo controlled double blind trials of corticosteroids in severe AH. J Hepatol 36:480, 2002. [PubMed: 11943418]

Sanyal AJ et al: Alcoholic and nonalcoholic fatty liver disease. Gastroenterology 150:8 (suppl), 2016.

Thurz MR et al: Prednisolone or pentoxifylline for alcoholic hepatitis. N Engl J Med 372:1619, 2015. [PubMed: 25901427]

Chapter 336: Nonalcoholic Fatty Liver Diseases and Nonalcoholic Steatohepatitis

Manal F. Abdelmalek; Anna Mae Diehl

INCIDENCE, PREVALENCE, AND NATURAL HISTORY

Nonalcoholic fatty liver disease (NAFLD) is the most common chronic liver disease in many parts of the world, including the United States. Population-based abdominal imaging studies have demonstrated fatty liver in at least 25% of American adults. Because the vast majority of these subjects deny hazardous levels of alcohol consumption (defined as greater than one drink per day in women or two drinks per day in men), they are considered to have NAFLD. NAFLD is strongly associated with overweight/obesity and insulin resistance. However, it can also occur in lean individuals and is particularly common in those with a paucity of adipose depots (i.e., lipodystrophy). Ethnic/racial factors also appear to influence liver fat accumulation; the documented prevalence of NAFLD is lowest in African Americans (~25%), highest in Americans of Hispanic ancestry (~50%), and intermediate in American whites (~33%).

NAFLD encompasses a spectrum of liver pathology with different clinical prognoses. The simple accumulation of triglyceride within hepatocytes (hepatic steatosis) is on the most clinically benign extreme of the spectrum. On the opposite, most clinically ominous extreme, are cirrhosis **(Chap. 337)** and primary liver cancer **(Chap. 78)**. The risk of developing cirrhosis is extremely low in individuals with chronic hepatic steatosis, but increases as steatosis becomes complicated by histologically conspicuous hepatocyte death and inflammation (i.e., nonalcoholic steatohepatitis [NASH]). NASH itself is also a heterogeneous condition; sometimes it improves to steatosis or normal histology, sometimes it remains relatively stable for years, but sometimes it results in progressive accumulation of fibrous scar that eventuates in cirrhosis. Once NAFLD-related cirrhosis develops, the annual incidence of primary liver cancer can be as high as 3%.

Abdominal imaging is not able to determine which individuals with NAFLD have associated liver cell death and inflammation (i.e., NASH), and specific blood tests to diagnose NASH are not yet available. However, population-based studies that have used elevated serum ALT as a marker of liver injury indicate that about 6–8% of American adults have serum ALT elevations that cannot be explained by excessive alcohol consumption, other known causes of fatty liver disease **(Table 336-1)**, viral hepatitis, or drug-induced or congenital liver diseases. Because the prevalence of such “cryptogenic” ALT elevations increases with body mass index, it is presumed that they are due to NASH. Hence, at any given point in time, NASH is present in about 25% of individuals who have NAFLD (i.e., about 6% of the general U.S. adult population has NASH).

Smaller cross-sectional studies in which liver biopsies have been performed on NASH patients at tertiary referral centers consistently demonstrate advanced fibrosis or cirrhosis in about 25% of those cohorts. By extrapolation, therefore, cirrhosis develops in about 6% of individuals with NAFLD (i.e., in about 1.5–2% of the general U.S. population). The risk for advanced liver fibrosis is highest in individuals with NASH who are aged >45–50 years and overweight/obese or afflicted with type 2 diabetes.

TABLE 336-1

Alternative Causes of Hepatic Steatosis

- Alcoholic liver disease
- Hepatitis C (particularly genotype 3)
- Inborn errors of metabolism
  - Abetalipoproteinemia
  - Cholesterol ester storage disease
  - Galactosemia
  - Glycogen storage disease
  - Hereditary fructose intolerance
  - Homocystinuria
  - Systemic carnitine deficiency
  - Tyrosinemia
  - Weber-Christian syndrome
  - Wilson's disease
  - Wolman's disease
- Medications **(see Table 336-2)**
- Miscellaneous
  - Industrial exposure to petrochemical
  - Inflammatory bowel disease
  - Lipodystrophy
  - Bacterial overgrowth
  - Starvation
  - Parenteral nutrition
- Surgical procedures
  - Bilopancreatic diversion
  - Extensive small-bowel resection
  - Gastric bypass
  - Jejunoileal bypass
- Reye's syndrome
- Acute fatty liver of pregnancy
- HELLP syndrome (*h*emolytic anemia, *e*levated *l*iver enzymes, *l*ow *p*latelet count)

Heritable factors clearly impact susceptibility to hepatic steatosis, NASH, liver fibrosis, and liver cancer. Indeed, recent twin studies suggest that inheritance accounts for about half the risk for developing cirrhosis. Certain variants in PNPLA3 (a gene that encodes an enzyme involved in intracellular trafficking of lipids) consistently correlates with susceptibility to hepatic steatosis, cirrhosis, and liver cancer. Polymorphisms in

other genes involved in lipid homeostasis (e.g., TM6SF2 and MBOAT7) are also emerging as potential genetic risk factors for NAFLD. Epigenetic factors (i.e., heritable traits that do not result from direct changes in DNA) may also influence NAFLD pathogenesis and/or progression based on evidence that intra-uterine exposures influence susceptibility to obesity and the metabolic syndrome in adolescence.

Experts have predicted that NAFLD will be the leading indication for liver transplantation in the United States by 2020. Similar to cirrhosis caused by other liver diseases, cirrhosis caused by NAFLD increases the risk for primary liver cancer. Both hepatocellular carcinoma and intrahepatic cholangiocarcinoma (ICC) have also been reported to occur in NAFLD patients without cirrhosis, suggesting that NAFLD per se may be a premalignant condition. NAFLD, NASH, and NAFLD-related cirrhosis are not limited to adults. All have been well documented in children. As in adults, obesity and insulin resistance are the main risk factors for pediatric NAFLD. Thus, the rising incidence and prevalence of childhood obesity suggests that NAFLD is likely to become an even greater contributor to society's burden of liver disease in the future.

PATHOGENESIS

The mechanisms underlying the pathogenesis and progression of NAFLD are not entirely clear. The best-understood mechanisms pertain to hepatic steatosis. This is proven to result when hepatocyte mechanisms for triglyceride synthesis (e.g., lipid uptake and de novo lipogenesis) overwhelm mechanisms for triglyceride disposal (e.g., degradative metabolism and lipoprotein export), leading to accumulation of fat (i.e., triglyceride) within hepatocytes. Obesity stimulates hepatocyte triglyceride accumulation by altering the intestinal microbiota to enhance both energy harvest from dietary sources and intestinal permeability. Reduced intestinal barrier function increases hepatic exposure to gut-derived products, which stimulate liver cells to generate inflammatory mediators that inhibit insulin actions. Obese adipose depots also produce excessive soluble factors (adipokines) that inhibit tissue insulin sensitivity. Insulin resistance promotes hyperglycemia. This drives the pancreas to produce more insulin to maintain glucose homeostasis. However, hyperinsulinemia also promotes lipid uptake, fat synthesis, and fat storage. The net result is hepatic triglyceride accumulation (i.e., steatosis).

Triglyceride per se is not hepatotoxic. However, its precursors (e.g., fatty acids and diacylglycerols) and metabolic by-products (e.g., reactive oxygen species) may damage hepatocytes, leading to hepatocyte lipotoxicity. Lipotoxicity also triggers the generation of other factors (e.g., inflammatory cytokines, hormonal mediators) that deregulate systems that normally maintain hepatocyte viability. The net result is increased hepatocyte death. Dying hepatocytes, in turn, release various factors that trigger wound healing responses that aim to replace (regenerate) lost hepatocytes. Such repair involves transient expansion of other cell types, such as myofibroblasts and progenitor cells, that make and degrade matrix, remodel the vasculature, and generate replacement hepatocytes, as well as the recruitment of immune cells that release factors that modulate liver injury and repair. NASH is the morphologic manifestation of lipotoxicity and resultant wound healing responses. Because the severity and duration of lipotoxic liver injury dictate the intensity and duration of repair, the histologic features and outcomes of NASH are variable. Cirrhosis and liver cancer are potential outcomes of chronic NASH. Cirrhosis results from futile repair, i.e., progressive accumulation of

wound healing cells, fibrous matrix, and abnormal vasculature (scarring), rather than efficient reconstruction/regeneration of healthy hepatic parenchyma. Primary liver cancers develop when malignantly transformed liver cells escape mechanisms that normally control regenerative growth. The mechanisms responsible for futile repair (cirrhosis) and liver carcinogenesis are not well understood. Because normal liver regeneration is a very complex process, there are multiple opportunities for deregulation and, thus, pathogenic heterogeneity. To date, this heterogeneity has confounded development of both diagnostic tests and treatments for defective/deregulated liver repair (i.e., cirrhosis and cancer). Hence, current strategies focus on circumventing misrepair by preventing and/or reducing lipotoxic liver injury.

DIAGNOSIS

Diagnosing NAFLD requires demonstration of increased liver fat in the absence of hazardous levels of alcohol consumption. Thresholds for potentially dangerous alcohol ingestion have been set at more than one drink per day in women and two drinks per day in men based on epidemiologic evidence that the prevalence of serum aminotransferase elevations increases when alcohol consumption habitually exceeds these levels. In those studies, one drink was defined as having 10 g of ethanol and, thus, is equivalent to one can of beer, 4 ounces of wine, or 1.5 ounces (one shot) of distilled spirits. Other causes of liver fat accumulation (particularly exposure to certain drugs; **Table 336-2**) and liver injury (e.g., viral hepatitis, autoimmune liver disease, iron or copper overload, α_1 antitrypsin deficiency) must also be excluded. Thus, establishing the diagnosis of NAFLD does not require invasive testing: it can be accomplished by history and physical examination, liver imaging (ultrasound is an acceptable first-line test; computed tomography [CT] or magnetic resonance imaging [MRI] enhances sensitivity for liver fat detection but adds expense), and blood tests to exclude other liver diseases. It is important to emphasize that the liver may not be enlarged, and serum aminotransferases and liver function tests (e.g., bilirubin, albumin, prothrombin time) may be completely normal, in individuals with NAFLD. Because there is yet no one specific blood test for NAFLD, confidence in the diagnosis of NAFLD is increased by identification of NAFLD risk factors. The latter include increased body mass index, insulin resistance/type 2 diabetes mellitus, and other parameters indicative of the metabolic syndrome (e.g., systemic hypertension, dyslipidemia, hyperuricemia/gout, cardiovascular disease; **Chap. 401**) in the patient or family members.

TABLE 336-2

Medications Associated with Hepatic Steatosis

- Cytotoxic and cytostatic drugs
  - L-Asparaginase
  - Azacitidine
  - Azaserine
  - Bleomycin
  - Methotrexate
  - Puromycin
  - Tetracycline
  - Doxycycline
- Metals
  - Antimony
  - Barium salts
  - Chromates
  - Phosphorus
  - Rare earths of low atomic number
  - Thallium compounds
  - Uranium compounds
- Other drugs and toxins
  - Amiodarone
  - 4,4′-Diethylaminoethoxyhexesterol
  - Ethionine
  - Ethyl bromide
  - Estrogens
  - Glucocorticoids
  - Highly active antiretroviral therapy
  - Hydralazine
  - Hypoglycin
  - Orotate
  - Perhexiline maleate
  - Safrole
  - Tamoxifen

Establishing the severity of NAFLD-related liver injury and related scarring (i.e., staging NAFLD) is more difficult than simply diagnosing NAFLD. Staging is critically important, however, because it is necessary to define prognosis and thereby determine treatment recommendations. The goal of staging is to distinguish

patients with NASH from those with simple steatosis and to identify which of the NASH patients have advanced fibrosis. The 10-year probability of developing liver-related morbidity or mortality in steatosis is negligible, and hence, this subgroup of NAFLD patients tends to be managed conservatively (see below). In contrast, more intensive follow-up and therapy are justified in NASH patients, and the subgroup with advanced fibrosis merits the most intensive scrutiny and intervention because their 10-year risk of liver-related morbidity and mortality is clearly increased.

Staging approaches can be separated into noninvasive testing (i.e., blood testing, physical examination, and imaging) and invasive approaches (i.e., liver biopsy). Blood test evidence of hepatic dysfunction (e.g., hyperbilirubinemia, hypoalbuminemia, prothrombin time prolongation) or portal hypertension (e.g., thrombocytopenia) and stigmata of portal hypertension on physical examination (e.g., spider angiomata, palmar erythema, splenomegaly, ascites, clubbing, encephalopathy) suggest a diagnosis of advanced NAFLD. Currently, however, liver biopsy is the gold standard for establishing the severity of liver injury and fibrosis because it is both more sensitive and specific than these other tests for establishing NAFLD severity. Although invasive, liver biopsy is seldom complicated by serious adverse sequelae such as significant bleeding, pain, or inadvertent puncture of other organs and thus is relatively safe. However, biopsy suffers from potential sampling error unless tissue cores of 2 cm or longer are acquired. Also, examination of tissue at a single point in time is not reliable for determining whether the pathologic processes are progressing or regressing. The risk of serial liver biopsies within short time intervals is generally deemed as unacceptable outside of research studies. These limitations of liver biopsy have stimulated efforts to develop noninvasive approaches to stage NAFLD. As is true for many other types of chronic liver disease, in NAFLD the levels of serum aminotransferases (aspartate aminotransferase [AST] and alanine aminotransferase [ALT]) do not reliably reflect the severity of liver cell injury, extent of liver cell death, or related liver inflammation and fibrosis. Thus, they are imperfect for determining which individuals with NAFLD have NASH. This has prompted efforts to identify superior markers of NASH, and particularly liver fibrosis, because fibrosis stage predicts eventual liver outcomes and mortality in NASH. Algorithms that combine various laboratory tests (e.g., ELF score, BARD Score, NAFLD Fibrosis Score, APRI score) are somewhat helpful in separating NASH patients with advanced versus mild liver fibrosis. Combining these tests with new imaging approaches that permit noninvasive quantification of liver fat (e.g., MRI using proton density fat fraction [MRI-PDFF]) and liver stiffness, a surrogate marker of liver fibrosis (e.g., magnetic resonance elastography, MRE, and transient elastography, FibroScan) improves their predictive power (Chap. 330). Increasingly, these new tools are being used serially to monitor fibrosis progression and regression in NAFLD patients. As a result, liver biopsy staging is becoming restricted to patients who cannot be stratified reliably using these noninvasive assessments.

CLINICAL FEATURES OF NAFLD

Most subjects with NAFLD are asymptomatic. The diagnosis is often made when abnormal liver aminotransferases or features of fatty liver are noted during an evaluation performed for other reasons. NAFLD may also be diagnosed during the workup of vague right upper quadrant abdominal pain, hepatomegaly, or an abnormal-appearing liver at time of abdominal surgery. Obesity is present in 50–90% of

subjects. Most patients with NAFLD also have other features of the metabolic syndrome **(Chap. 401)**. Some have subtle stigmata of chronic liver disease, such as spider angiomata, palmer erythema, or splenomegaly. In a small minority of patients with advanced NAFLD, complications of end-stage liver disease (e.g., jaundice, features of portal hypertension such as ascites or variceal hemorrhage) may be the initial findings.

The association of NAFLD with obesity, diabetes, hypertriglyceridemia, hypertension, and cardiovascular disease is well known. Other associations include chronic fatigue, mood alterations, obstructive sleep apnea, thyroid dysfunction, and chronic pain syndrome. NAFLD is an independent risk factor for metabolic syndrome **(Chap. 401)**. Longitudinal studies suggest that patients with NASH are at two- to threefold increased risk for the development of metabolic syndrome. Similarly, studies have shown that patients with NASH have a higher risk for the development of hypertension and diabetes mellitus. The presence of NAFLD is also independently associated with endothelial dysfunction, increased carotid intimal thickness, and the number of plaques in carotid and coronary arteries. Such data indicate that NAFLD has many deleterious effects on health in general.

TREATMENT OF NAFLD

Treatment of NAFLD can be divided into three components: (1) specific therapy of NAFLD-related liver disease; (2) treatment of NAFLD-associated comorbidities; and (3) treatment of the complications of advanced NAFLD. The subsequent discussion focuses on specific therapies for NAFLD, with some mention of their impact on major NAFLD comorbidities (insulin resistance/diabetes, obesity, and dyslipidemia). Treatment of the complications of advanced NAFLD involves management of the complications of cirrhosis and portal hypertension, including primary liver cancers. Approaches to accomplish these objectives are similar to those used in other chronic liver diseases and are covered elsewhere in the textbook **(Chaps. 337 and 78)**.

At present, there are no Food and Drug Administration (FDA)-approved therapies for the treatment of NAFLD. Thus, the current approach to NAFLD management focuses on treatment to improve the risk factors for NASH (i.e., obesity, insulin resistance, metabolic syndrome, dyslipidemia). Based on our understanding of the natural history of NAFLD, only patients with NASH or those with features of hepatic fibrosis on liver biopsy are considered currently for targeted pharmacologic therapies. This approach may change as our understanding of disease pathophysiology improves and potential targets of therapy evolve.

Diet and Exercise

Lifestyle changes and dietary modification are the foundation for NAFLD treatment. Many studies indicate that lifestyle modification can improve serum aminotransferases and hepatic steatosis, with loss of at least 3–5% of body weight improving steatosis, but greater weight loss (up to 10%) necessary to improve steatohepatitis. The benefits of different dietary macronutrient contents (e.g., low-carbohydrate vs low-fat diets, saturated vs unsaturated fat diets) and different intensities of calorie restriction appear to be comparable. In adults with NAFLD, exercise regimens that improve fitness may be sufficient to reduce hepatic

steatosis, but their impact on other aspects of liver histology remains unknown. Unfortunately, most NAFLD patients are unable to achieve sustained weight loss. Although pharmacologic therapies such as orlistat, topiramate, and phentermine to facilitate weight loss are available, their role in the treatment of NAFLD remains experimental.

Pharmacologic Therapies

Several drug therapies have been tried in both research and clinical settings. No agent has yet been approved by the FDA for the treatment of NAFLD. Hence, this remains an area of active research. Because NAFLD is strongly associated with the metabolic syndrome and type 2 diabetes **(Chaps. 396 and 397)**, the efficacy of various insulin-sensitizing agents has been examined. *Metformin*, an agent that mainly improves hepatic insulin sensitivity, has been evaluated in several small, open-label studies in adults and a recent larger, prospectively randomized trial in children (dubbed the TONIC study). Although several of the adult NASH studies suggested improvements in aminotransferases and/or liver histology, metformin did not improve liver histology in the TONIC study of children with NASH. Thus, it is not currently recommended as a treatment for NASH. Uncontrolled open-label studies have also investigated *thiazolidinediones* (*pioglitazone and rosiglitazone*) in adults with NASH. This class of drugs is known to improve systemic insulin resistance. Both pioglitazone and rosiglitazone reduced aminotransferases and improved some of the histologic features of NASH in small, uncontrolled studies. A large, National Institutes of Health–sponsored, randomized placebo-controlled clinical trial, the PIVENS Study (Pioglitazone vs Vitamin E vs Placebo for the Treatment of 247 Nondiabetic Adults with NASH), demonstrated that resolution of histologic NASH occurred more often in subjects treated with pioglitazone (30 mg/d) than with placebo for 18 months (47 vs 21%, $p = .001$). However, many subjects in the pioglitazone group gained weight, and liver fibrosis did not improve. Also, it should be noted that the long-term safety and efficacy of thiazolidinediones in patients with NASH has not been established. Five-year follow-up of subjects treated with rosiglitazone demonstrated no reduction in liver fibrosis, and rosiglitazone has been associated with increased long-term risk for cardiovascular mortality. Hence, it is not recommended as a treatment for NAFLD. Pioglitazone may be safer because in a recent large meta-analysis it was associated with reduced overall morality, myocardial infarction, and stroke. However, caution must be exercised when considering its use in patients with impaired myocardial function.

Antioxidants have also been evaluated for the treatment of NAFLD because oxidant stress is thought to contribute to the pathogenesis of NASH. *Vitamin E*, an inexpensive yet potent antioxidant, has been examined in several small pediatric and adult studies with varying results. In all of those studies, vitamin E was well tolerated, and most showed modest improvements in aminotransferase levels, radiographic features of hepatic steatosis, and/or histologic features of NASH. Vitamin E (800 IU/d) was also compared to placebo in the PIVENS and TONIC studies. In PIVENS, vitamin E was the only agent that achieved the predetermined primary endpoint (i.e., improvement in steatohepatitis, lobular inflammation, and steatosis score, without an increase in the fibrosis score). This endpoint was met in 43% of patients in the vitamin E group ($p = .001$ vs placebo), 34% in the pioglitazone group ($p = .04$ vs placebo), and 19% in the placebo group. Vitamin E also improved NASH histology in pediatric patients with NASH (TONIC trial). However, a recent population-based study suggested that chronic vitamin E therapy may increase the risk for cardiovascular

mortality. Thus, vitamin E should only be considered as a first-line pharmacotherapy for nondiabetic NASH patients. Also, given its potentially negative effects on cardiovascular health, caution should be exercised until the risk-to-benefit ratio and long-term therapeutic efficacy of vitamin E are better defined. Ursodeoxycholic acid (a bile acid that improves certain cholestatic liver diseases) and *betaine* (metabolite of choline that raises SAM levels and decreases cellular oxidative damage) offer no histologic benefit over placebo in patients with NASH. Experimental evidence to support the use of *omega-3 fatty acids* in NAFLD exists; however, a recent large, multicenter, placebo-controlled study failed to demonstrate a histologic benefit. Other pharmacotherapies are also being evaluated in NAFLD (e.g., *probiotics, farnesoid X receptor agonists, intestinal bile acid transport inhibitors, fibroblast growth factor agonists, anticytokine agents, glucagon-like peptide agonists, dipeptidyl IV antagonists, dual PPAR-alpha/PPAR-delta agonists, modulators of liver fibrosis*); however, sufficient data do not yet exist to justify their use as NASH treatments in standard clinical practice.

Statins are an important class of agents to treat dyslipidemia and decrease cardiovascular risk. There is no evidence to suggest that statins cause liver failure in patients with any chronic liver disease, including NAFLD. The incidence of liver enzyme elevations in NAFLD patients taking statins is also no different than that of healthy controls or patients with other chronic liver diseases. Moreover, several studies have suggested that statins may improve aminotransferases and histology in patients with NASH. Yet, there is continued reluctance to use statins in patients with NAFLD. The lack of evidence that statins harm the liver in NAFLD patients, combined with the increase risk for cardiovascular morbidity and mortality in NAFLD patients, warrants the use of statins to treat dyslipidemia in patients with NAFLD/NASH.

Bariatric Surgery

Although interest in bariatric surgery as a treatment for NAFLD exists, a recently published Cochrane review concluded that lack of randomized clinical trials or adequate clinical studies prevents definitive assessment of benefits and harms of bariatric surgery as a treatment for NASH. Most studies of bariatric surgery have shown that bariatric surgery is generally safe in individuals with well-compensated chronic liver disease and improves hepatic steatosis and necroinflammation (i.e., features of NAFLD/NASH); however, effects on hepatic fibrosis have been variable. Concern lingers because some of the largest prospective studies suggest that hepatic fibrosis might progress after bariatric surgery. Thus, the Cochrane review deemed it premature to recommend bariatric surgery as a primary treatment for NASH. This opinion was challenged by a recently study which demonstrated that fibrosis stage had improved by 5 years after surgery in about half the patients in one large bariatric surgery cohort. However, most of those individuals had relatively mild fibrosis initially and thus, it is unclear if similar outcomes would occur in individuals with more advanced liver disease. Indeed there is general agreement that patients with NAFLD-related cirrhosis and portal hypertension should be excluded as candidates for bariatric surgery. However, given growing evidence for the benefits of bariatric surgery on metabolic syndrome complications in individuals with refractory obesity, it is not contraindicated in otherwise eligible patients with NAFLD or NASH.

Liver Transplantation

Patients with NAFLD in whom end-stage liver disease develops should be evaluated for liver transplantation (Chap. 338). The outcomes of liver transplantation in well-selected patients with NAFLD are generally good, but comorbid medical conditions associated with NAFLD, such as diabetes mellitus, obesity, and cardiovascular disease, often limit transplant candidacy. NAFLD may recur after liver transplantation. The risk factors for recurrent or de novo NAFLD after liver transplantation are multifactorial and include hypertriglyceridemia, obesity, diabetes mellitus, and immunosuppressive therapies, particularly glucocorticoids.

GLOBAL HEALTH CONSIDERATIONS



The epidemic of obesity is now a global and accelerating phenomenon. Worldwide, there are >1 billion overweight adults, of whom at least 300 million are obese. The worldwide prevalence of obesity has more than doubled since 1980. In the wake of the obesity epidemic follow numerous comorbidities, including NAFLD. NAFLD is the most common liver disease identified in Western countries and the fastest rising form of chronic liver disease worldwide. The economic burden directly attributable to NAFLD is already enormous (estimated as $103 billion/year in the United States and nearly 35 billion Euros/year for four European Union countries) and predicted to increase tenfold by the year 2025. Present understanding of NAFLD natural history is based mainly on studies in whites who became overweight/obese and developed the metabolic syndrome in adulthood. The impact of the global childhood obesity epidemic on NAFLD pathogenesis/progression is unknown. Emerging evidence demonstrates that advanced NAFLD, including cirrhosis and primary liver cancer, can occur in children, prompting concerns that childhood-onset NAFLD might follow a more aggressive course than typical adult-acquired NAFLD. Some of the most populated parts of the world are in the midst of industrial revolutions, and certain environmental pollutants seem to exacerbate NAFLD. Some studies also suggest that the risk for NASH and NAFLD-related cirrhosis may be higher in certain ethnic groups such as Asians, certain Hispanics, and Native Americans and lower in others such as African Americans, compared with whites. Although all of these variables confound efforts to predict the net impact of this obesity-related liver disease on global health, it seems likely that NAFLD will remain a major cause of chronic liver disease worldwide for the foreseeable future.

FURTHER READING

Angulo P et al: Fibrosis in nonalcoholic fatty liver disease: Mechanisms and clinical implications. Semin Liver Dis 35:132, 2015. [PubMed: 25974899]

Chalasani N et al: The diagnosis and management of non-alcoholic fatty liver disease: Practice guideline by the American Gastroenterological Association, American Association for the Study of Liver Diseases, and American College of Gastroenterology. Gastroenterology 142:1592, 2012. [PubMed: 22656328]

European Association for the Study of the Liver (EASL); European Association for the Study of Diabetes (EASD); European Association for the Study of Obesity (EASO): EASL-EASD-EASO Clinical Practice Guidelines

for the management of non-alcoholic fatty liver disease. J Hepatol 64:1388, 2016. [PubMed: 27062661]

Sanyal AJ et al: Pioglitazone, vitamin E, or placebo for nonalcoholic steatohepatitis. N Engl J Med 362:1675, 2010. [PubMed: 20427778]

Chapter 337: Cirrhosis and Its Complications

Bruce R. Bacon

INTRODUCTION

Cirrhosis is a condition that is defined histopathologically and has a variety of clinical manifestations and complications, some of which can be life-threatening. In the past, it has been thought that cirrhosis was never reversible; however, it has become apparent that when the underlying insult that has caused the cirrhosis has been removed, there can be reversal of fibrosis. This is most apparent with the successful treatment of chronic hepatitis C; however, reversal of fibrosis is also seen in patients with hemochromatosis who have been successfully treated and in patients with alcoholic liver disease who have discontinued alcohol use.

Regardless of the cause of cirrhosis, the pathologic features consist of the development of fibrosis to the point that there is architectural distortion with the formation of regenerative nodules. This results in a decrease in hepatocellular mass, and thus function, and an alteration of blood flow. The induction of fibrosis occurs with activation of hepatic stellate cells, resulting in the formation of increased amounts of collagen and other components of the extracellular matrix.

Clinical features of cirrhosis are the result of pathologic changes and mirror the severity of the liver disease. Most hepatic pathologists provide an assessment of grading and staging when evaluating liver biopsy samples. These grading and staging schemes vary between disease states and have been developed for most conditions, including chronic viral hepatitis, nonalcoholic fatty liver disease, and primary biliary cholangitis. Advanced fibrosis usually includes bridging fibrosis with nodularity designated as stage 3 and cirrhosis designated as stage 4. Patients who have cirrhosis have varying degrees of compensated liver function, and clinicians need to differentiate between those who have stable, compensated cirrhosis and those who have decompensated cirrhosis. Patients who have developed complications of their liver disease and have become decompensated should be considered for liver transplantation. Many of the complications of cirrhosis will require specific therapy. *Portal hypertension* is a significant complicating feature of decompensated cirrhosis and is responsible for the development of ascites and bleeding from esophagogastric varices, two complications that signify decompensated cirrhosis. Loss of hepatocellular function results in jaundice, coagulation disorders, and hypoalbuminemia and contributes to the causes of portosystemic encephalopathy. The complications of cirrhosis are basically the same regardless of the etiology. Nonetheless, it is useful to classify patients by the cause of their liver disease **(Table 337-1)**; patients can be divided into broad groups with alcoholic cirrhosis, cirrhosis due to chronic viral hepatitis, biliary

cirrhosis, and other, less common causes such as cardiac cirrhosis, cryptogenic cirrhosis, and other miscellaneous causes.

TABLE 337-1

Causes of Cirrhosis

Alcoholism	Cardiac cirrhosis
Chronic viral hepatitis	Inherited metabolic liver disease
Hepatitis B	Hemochromatosis
Hepatitis C	Wilson's disease
Autoimmune hepatitis	α_1 Antitrypsin deficiency
Nonalcoholic steatohepatitis	Cystic fibrosis
Biliary cirrhosis	Cryptogenic cirrhosis
Primary biliary cholangitis	
Primary sclerosing cholangitis	
Autoimmune cholangiopathy	

ALCOHOLIC CIRRHOSIS

Excessive chronic alcohol use can cause several different types of chronic liver disease, including alcoholic fatty liver, alcoholic hepatitis, and alcoholic cirrhosis. Furthermore, use of excessive alcohol can contribute to liver damage in patients with other liver diseases, such as hepatitis C, hemochromatosis, and fatty liver disease related to obesity. Chronic alcohol use can produce fibrosis in the absence of accompanying inflammation and/or necrosis. Fibrosis can be centrilobular, pericellular, or periportal. When fibrosis reaches a certain degree, there is disruption of the normal liver architecture and replacement of liver cells by regenerative nodules. In alcoholic cirrhosis, the nodules are usually <3 mm in diameter; this form of cirrhosis is referred to as *micronodular*. With cessation of alcohol use, larger nodules may form, resulting in a mixed micronodular and macronodular cirrhosis.

Pathogenesis

Alcohol is the most commonly used drug in the United States, and more than two-thirds of adults drink alcohol each year. Thirty percent have had a binge within the past month, and over 7% of adults regularly consume more than two drinks per day. Unfortunately, more than 14 million adults in the United States meet the diagnostic criteria for alcohol abuse or dependence. In the United States, chronic liver disease is the tenth most common cause of death in adults, and alcoholic cirrhosis accounts for ~40% of deaths due to cirrhosis.

Ethanol is mainly absorbed by the small intestine and, to a lesser degree, through the stomach. Gastric alcohol dehydrogenase (ADH) initiates alcohol metabolism. Three enzyme systems account for metabolism of alcohol in the liver. These include cytosolic ADH, the microsomal ethanol oxidizing system (MEOS), and peroxisomal catalase. The majority of ethanol oxidation occurs via ADH to form acetaldehyde, which is a highly reactive molecule that may have multiple effects. Ultimately, acetaldehyde is metabolized to acetate by aldehyde dehydrogenase (ALDH). Intake of ethanol increases intracellular accumulation of triglycerides by increasing fatty acid uptake and by reducing fatty acid oxidation and lipoprotein secretion. Protein synthesis, glycosylation, and secretion are impaired. Oxidative damage to hepatocyte membranes occurs due to the formation of reactive oxygen species; acetaldehyde is a highly reactive molecule that combines with proteins to form protein-acetaldehyde adducts. These adducts may interfere with specific enzyme activities, including microtubular formation and hepatic protein trafficking. With acetaldehyde-mediated hepatocyte damage, certain reactive oxygen species can result in Kupffer cell activation. As a result, profibrogenic cytokines are produced that initiate and perpetuate stellate cell activation, with the resultant production of excess collagen and extracellular matrix. Connective tissue appears in both periportal and pericentral zones and eventually connects portal triads with central veins forming regenerative nodules. Hepatocyte loss occurs, and with increased collagen production and deposition, together with continuing hepatocyte destruction, the liver contracts and shrinks in size. This process generally takes from years to decades to occur and requires repeated insults.

Clinical Features

The diagnosis of alcoholic liver disease requires an accurate history regarding both amount and duration of alcohol consumption. Patients with alcoholic liver disease can present with nonspecific symptoms such as vague right upper quadrant abdominal pain, fever, nausea and vomiting, diarrhea, anorexia, and malaise. Alternatively, they may present with more specific complications of chronic liver disease, including ascites, edema, or upper gastrointestinal (GI) hemorrhage. Many cases present incidentally at the time of autopsy or elective surgery. Other clinical manifestations include the development of jaundice or encephalopathy. The abrupt onset of any of these complications may be the first event prompting the patient to seek medical attention. Other patients may be identified in the course of an evaluation of routine laboratory studies that are found to be abnormal. On physical examination, the liver and spleen may be enlarged, with the liver edge being firm and nodular. Other frequent findings include scleral icterus, palmar erythema **(Fig. 337-1)**, spider angiomas **(Fig. 337-2)**, parotid gland enlargement, digital clubbing, muscle wasting, or the development of edema and ascites. Men may have decreased body hair and gynecomastia as well as testicular atrophy, which may be a consequence of hormonal abnormalities or a direct toxic effect of alcohol on the testes. In

women with advanced alcoholic cirrhosis, menstrual irregularities usually occur, and some women may be amenorrheic. These changes are often reversible following cessation of alcohol ingestion.

FIGURE 337-1

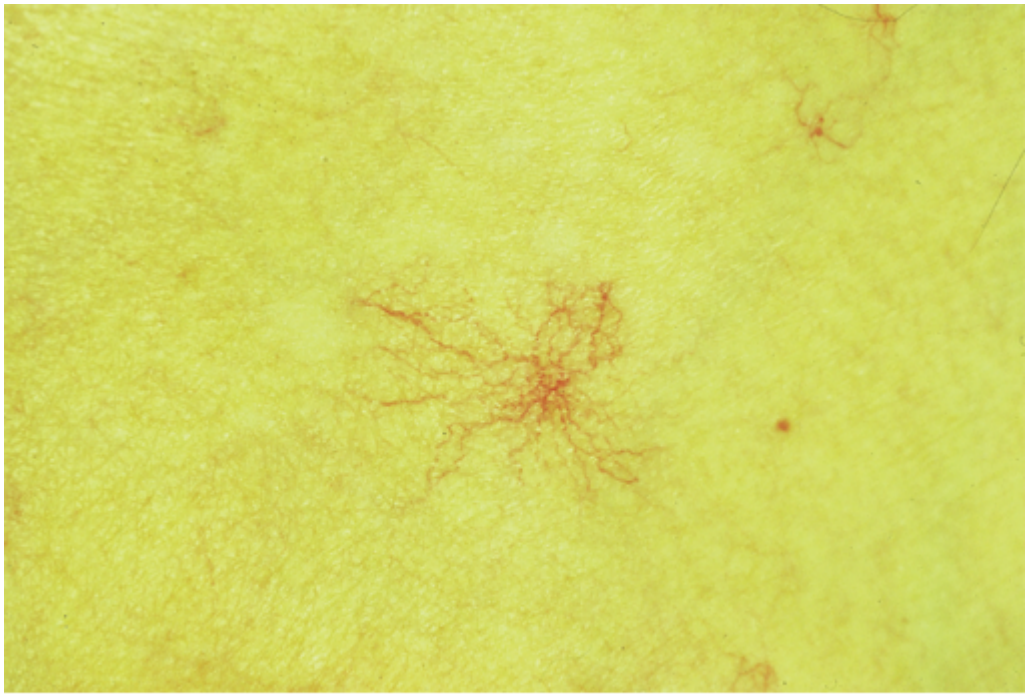
Palmar erythema. This figure shows palmar erythema in a patient with alcoholic cirrhosis. The erythema is peripheral over the palm with central pallor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 337-2

Spider angioma. This figure shows a spider angioma in a patient with hepatitis C cirrhosis. With release of central compression, the arteriole fills from the center and spreads out peripherally.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Laboratory tests may be completely normal in patients with early compensated alcoholic cirrhosis. Alternatively, in advanced liver disease, many abnormalities usually are present. Patients may be anemic either from chronic GI blood loss, nutritional deficiencies, or hypersplenism related to portal hypertension, or as a direct suppressive effect of alcohol on the bone marrow. A unique form of hemolytic anemia (with spur cells and acanthocytes) called *Zieve's syndrome* can occur in patients with severe alcoholic hepatitis. Platelet counts are often reduced early in the disease, reflective of portal hypertension with hypersplenism. Serum total bilirubin can be normal or elevated with advanced disease. Direct bilirubin is frequently mildly elevated in patients with a normal total bilirubin, but the abnormality typically progresses as the disease worsens. Prothrombin times are often prolonged and usually do not respond to administration of parenteral vitamin K. Serum sodium levels are usually normal unless patients have ascites and then can be depressed, largely due to ingestion of excess free water. Serum alanine and aspartate aminotransferases (ALT, AST) are typically elevated, particularly in patients who continue to drink, with AST levels being higher than ALT levels, usually by a 2:1 ratio.

Diagnosis

Patients who have any of the above-mentioned clinical features, physical examination findings, or laboratory studies should be considered to have alcoholic liver disease. The diagnosis, however, requires accurate knowledge that the patient is continuing to use and abuse alcohol. Furthermore, other forms of chronic liver disease (e.g., chronic viral hepatitis or metabolic or autoimmune liver diseases) must be considered or ruled out, or if present, an estimate of relative causality along with the alcohol use should be determined. Liver biopsy can be helpful to confirm a diagnosis, but generally when patients present with alcoholic hepatitis and are still drinking, liver biopsy is withheld until abstinence has been maintained for at least 6 months to determine residual, nonreversible disease.

In patients who have had complications of cirrhosis and who continue to drink, there is a <50% 5-year survival. In contrast, in patients who are able to remain abstinent, the prognosis is significantly improved. In patients with advanced liver disease, the prognosis remains poor; however, in individuals who are able to remain abstinent, liver transplantation is a viable option.

TREATMENT

TREATMENT

Alcoholic Cirrhosis

Abstinence is the cornerstone of therapy for patients with alcoholic liver disease. In addition, patients require good nutrition and long-term medical supervision to manage underlying complications that may develop. Complications such as the development of ascites and edema, variceal hemorrhage, or portosystemic encephalopathy all require specific management and treatment. Glucocorticoids are occasionally used in patients with severe alcoholic hepatitis in the absence of infection. Survival has been shown to improve in certain studies. Treatment is restricted to patients with a discriminant function (DF) value of >32. The DF is calculated as the serum total bilirubin plus the difference in the patient's prothrombin time compared to control (in seconds) multiplied by 4.6. In patients for whom this value is >32, there is improved survival at 28 days with the use of glucocorticoids.

Other therapies that have been used include oral pentoxifylline, which decreases the production of tumor necrosis factor α (TNF-α) and other proinflammatory cytokines. In contrast to glucocorticoids, with which complications can occur, pentoxifylline is relatively easy to administer and has few, if any, side effects. A variety of nutritional therapies have been tried with either parenteral or enteral feedings; however, it is unclear whether any of these modalities have significantly improved survival.

Recent studies have used parenterally administered inhibitors of TNF-α such as infliximab or etanercept. Early results have shown no adverse events; however, there was no clear-cut improvement in survival. Anabolic steroids, propylthiouracil, antioxidants, colchicine, and penicillamine have all been used but do not show clear-cut benefits and are not recommended.

As mentioned above, the cornerstone to treatment is cessation of alcohol use. Recent experience with medications that reduce craving for alcohol, such as acamprosate calcium, has been favorable. Patients may take other necessary medications even in the presence of cirrhosis. Acetaminophen use is often discouraged in patients with liver disease; however, if no more than 2 g of acetaminophen per day are consumed, there generally are no problems.

CIRRHOSIS DUE TO CHRONIC VIRAL HEPATITIS B OR C


Of patients exposed to the hepatitis C virus (HCV), ~80% develop chronic hepatitis C, and of those, about 20–30% will develop cirrhosis over 20–30 years. Many of these patients have had concomitant alcohol use, and the true incidence of cirrhosis due to hepatitis C alone is unknown. Nonetheless, this represents a

significant number of patients. It is expected that an even higher percentage will go on to develop cirrhosis over longer periods of time. In the United States, ~5 to 6 million people have been exposed to HCV, with about 4–5 million who are chronically viremic. Worldwide, about 170 million individuals have hepatitis C, with some areas of the world (e.g., Egypt) having up to 15% of the population infected. HCV is a noncytopathic virus, and liver damage is probably immune-mediated. Progression of liver disease due to chronic hepatitis C is characterized by portal-based fibrosis with bridging fibrosis and nodularity developing, ultimately culminating in the development of cirrhosis. In cirrhosis due to chronic hepatitis C, the liver is small and shrunken with characteristic features of a mixed micro- and macronodular cirrhosis seen on liver biopsy. In addition to the increased fibrosis that is seen in cirrhosis due to hepatitis C, an inflammatory infiltrate is found in portal areas with interface hepatitis and occasionally some lobular hepatocellular injury and inflammation. In patients with HCV genotype 3, steatosis is often present.

 Similar findings are seen in patients with cirrhosis due to chronic hepatitis B. Of adult patients exposed to hepatitis B, about 5% develop chronic hepatitis B, and about 20% of those patients will go on to develop cirrhosis. Special stains for hepatitis B core (HBc) and hepatitis B surface (HBs) antigen will be positive, and ground-glass hepatocytes signifying HBs antigen (HBsAg) may be present. In the United States, there are about 2 million carriers of hepatitis B, whereas in other parts of the world where hepatitis B virus (HBV) is endemic (i.e., Asia, Southeast Asia, sub-Saharan Africa), up to 15% of the population may be infected, having acquired the infection vertically at the time of birth. Thus, over 300–400 million individuals are thought to have hepatitis B worldwide. Approximately 25% of these individuals may ultimately develop cirrhosis.

Clinical Features and Diagnosis

Patients with cirrhosis due to either chronic hepatitis C or B can present with the usual symptoms and signs of chronic liver disease. Fatigue, malaise, vague right upper quadrant pain, and laboratory abnormalities are frequent presenting features. Diagnosis requires a thorough laboratory evaluation, including quantitative HCV RNA testing and analysis for HCV genotype, or hepatitis B serologies to include HBsAg, anti-HBs, HBeAg (hepatitis B e antigen), anti-HBe, and quantitative HBV DNA levels.

TREATMENT

TREATMENT

Cirrhosis due to Chronic Viral Hepatitis B or C

Management of complications of cirrhosis revolves around specific therapy for treatment of whatever complications occur (e.g., esophageal variceal hemorrhage, development of ascites and edema, or encephalopathy). In patients with chronic hepatitis B, numerous studies have shown beneficial effects of antiviral therapy, which is effective at viral suppression, as evidenced by reducing aminotransferase levels and HBV DNA levels, and improving histology by reducing inflammation and fibrosis. Several clinical trials and case series have demonstrated that patients with decompensated liver disease can become compensated with the use of antiviral therapy directed against hepatitis B. Currently available therapy

includes lamivudine, adefovir, telbivudine, entecavir, and tenofovir. Interferon α can also be used for treating hepatitis B, but it should not be used in cirrhotics. The majority of patients being treated for hepatitis B are receiving either entecavir or tenofovir **(see Chap. 334)**.

Treatment of patients with cirrhosis due to hepatitis C used to be more difficult because the side effects of pegylated interferon and ribavirin therapy were difficult to manage. Over the last several years, interferon-based regimens have been replaced by direct-acting antiviral protocols that are highly successful (>95% cure rate), well tolerated, usually of short duration (8–12 weeks), but costly. These medications have truly revolutionized the treatment of hepatitis C **(see Chap. 334)**.

CIRRHOSIS FROM AUTOIMMUNE HEPATITIS AND NONALCOHOLIC FATTY LIVER DISEASE

Other causes of posthepatitic cirrhosis include autoimmune hepatitis and cirrhosis due to nonalcoholic steatohepatitis. Many patients with autoimmune hepatitis (AIH) present with cirrhosis that is already established. Typically, these patients will not benefit from immunosuppressive therapy with glucocorticoids or azathioprine because the AIH is “burned out.” In this situation, liver biopsy does not show a significant inflammatory infiltrate. Diagnosis in this setting requires positive autoimmune markers such as antinuclear antibody (ANA) or anti-smooth-muscle antibody (ASMA). When patients with AIH present with cirrhosis and active inflammation accompanied by elevated liver enzymes, there can be considerable benefit from the use of immunosuppressive therapy.

Patients with nonalcoholic steatohepatitis are increasingly being found to have progressed to cirrhosis. With the epidemic of obesity that continues in Western countries, more and more patients are identified with nonalcoholic fatty liver disease **(Chap. 336)**. Of these, a significant subset has nonalcoholic steatohepatitis and can progress to increased fibrosis and cirrhosis. Over the past several years, it has been increasingly recognized that many patients who were thought to have cryptogenic cirrhosis in fact have nonalcoholic steatohepatitis. As their cirrhosis progresses, they become catabolic and then lose the telltale signs of steatosis seen on biopsy. Management of complications of cirrhosis due to either AIH or nonalcoholic steatohepatitis is similar to that for other forms of cirrhosis.

BILIARY CIRRHOSIS

Biliary cirrhosis has pathologic features that are different from either alcoholic cirrhosis or posthepatitic cirrhosis, yet the manifestations of end-stage liver disease are the same. Cholestatic liver disease may result from necroinflammatory lesions, congenital or metabolic processes, or external bile duct compression. Thus, two broad categories reflect the anatomic sites of abnormal bile retention: *intrahepatic* and *extrahepatic*. The distinction is important for obvious therapeutic reasons. Extrahepatic obstruction may benefit from surgical or endoscopic biliary tract decompression, whereas intrahepatic cholestatic processes will not improve with such interventions and require a different approach.

The major causes of chronic cholestatic syndromes are primary biliary cholangitis (PBC), autoimmune cholangitis (AIC), primary sclerosing cholangitis (PSC), and idiopathic adulthood ductopenia. These syndromes are usually clinically distinguished from each other by antibody testing, cholangiographic findings, and clinical presentation. However, they all share the histopathologic features of chronic cholestasis, such as cholate stasis; copper deposition; xanthomatous transformation of hepatocytes; and irregular, so-called biliary fibrosis. In addition, there may be chronic portal inflammation, interface activity, and chronic lobular inflammation. Ductopenia is a result of this progressive disease as patients develop cirrhosis.

PRIMARY BILIARY CHOLANGITIS

PBC is seen in about 100–200 individuals per million, with a strong female preponderance and a median age of around 50 years at the time of diagnosis. The cause of PBC is unknown; it is characterized by portal inflammation and necrosis of cholangiocytes in small- and medium-sized bile ducts. Cholestatic features prevail, and biliary cirrhosis is characterized by an elevated bilirubin level and progressive liver failure. Liver transplantation is the treatment of choice for patients with decompensated cirrhosis due to PBC. A variety of therapies have been proposed, but ursodeoxycholic acid (UDCA) has been the only approved treatment that has some degree of efficacy by slowing the rate of progression of the disease. In 2016, obeticholic acid was approved for use in PBC patients with an inadequate response to UDCA. Several other agents are in varying stages of development.

Antimitochondrial antibodies (AMA) are present in about 90% of patients with PBC. These autoantibodies recognize intermitochondrial membrane proteins that are enzymes of the pyruvate dehydrogenase complex (PDC), the branched-chain 2-oxoacid dehydrogenase complex, and the 2-oxogluterate dehydrogenase complex. Most relate to pyruvate dehydrogenase. These autoantibodies are not pathogenic but rather are useful markers for making a diagnosis of PBC.

Pathology

Histopathologic analyses of liver biopsies of patients with PBC have resulted in identifying four distinct stages of the disease as it progresses. The earliest lesion is termed *chronic nonsuppurative destructive cholangitis* and is a necrotizing inflammatory process of the portal tracts. Medium and small bile ducts are infiltrated with lymphocytes and undergo duct destruction. Mild fibrosis and sometimes bile stasis can occur. With progression, the inflammatory infiltrate becomes less prominent, but the number of bile ducts is reduced and there is proliferation of smaller bile ductules. Increased fibrosis ensues with the expansion of periportal fibrosis to bridging fibrosis. Finally, cirrhosis, which may be micronodular or macronodular, develops.

Clinical Features

Currently, most patients with PBC are diagnosed well before the end-stage manifestations of the disease are present, and, as such, most patients are actually asymptomatic. When symptoms are present, they most

prominently include a significant degree of fatigue out of proportion to what would be expected for either the severity of the liver disease or the age of the patient. Pruritus is seen in ~50% of patients at the time of diagnosis, and it can be debilitating. It might be intermittent and usually is most bothersome in the evening. In some patients, pruritus can develop toward the end of pregnancy, and there are examples of patients having been diagnosed with cholestasis of pregnancy rather than PBC. Pruritus that presents prior to the development of jaundice indicates severe disease and a poor prognosis.

Physical examination can show jaundice and other complications of chronic liver disease including hepatomegaly, splenomegaly, ascites, and edema. Other features that are unique to PBC include hyperpigmentation, xanthelasma, and xanthomata, which are related to the altered cholesterol metabolism seen in this disease. Hyperpigmentation is evident on the trunk and the arms and is seen in areas of exfoliation and lichenification associated with progressive scratching related to the pruritus. Bone pain resulting from osteopenia or osteoporosis is occasionally seen at the time of diagnosis.

Laboratory Findings

Laboratory findings in PBC show cholestatic liver enzyme abnormalities with an elevation in γ-glutamyl transpeptidase and alkaline phosphatase (ALP) along with mild elevations in aminotransferases (ALT and AST). Immunoglobulins, particularly IgM, are typically increased. Hyperbilirubinemia usually is seen once cirrhosis has developed. Thrombocytopenia, leukopenia, and anemia may be seen in patients with portal hypertension and hypersplenism. Liver biopsy shows characteristic features as described above and should be evident to any experienced hepatopathologist. Up to 10% of patients with characteristic PBC will have features of AIH as well and are defined as having “overlap” syndrome. These patients are usually treated as PBC patients and may progress to cirrhosis with the same frequency as typical PBC patients. Some patients require immunosuppressive medications as well.

Diagnosis

PBC should be considered in patients with chronic cholestatic liver enzyme abnormalities. It is most often seen in middle-aged women. AMA testing may be negative, and it should be remembered that as many as 10% of patients with PBC may be AMA-negative. Liver biopsy is most important in this setting of AMA-negative PBC. In patients who are AMA-negative with cholestatic liver enzymes, PSC should be ruled out by way of cholangiography.

TREATMENT

TREATMENT

Primary Biliary Cholangitis

Treatment of the typical manifestations of cirrhosis is no different for PBC than for other forms of cirrhosis. UDCA has been shown to improve both biochemical and histologic features of the disease. Improvement is greatest when therapy is initiated early; the likelihood of significant improvement with UDCA is low in

patients with PBC who present with manifestations of cirrhosis. UDCA is given in doses of 13–15 mg/kg per day; the medication is usually well-tolerated, although some patients have worsening pruritus with initiation of therapy. A small proportion of patients may have diarrhea or headache as a side effect of the drug. UDCA has been shown to slow the rate of progression of PBC, but it does not reverse or cure the disease. About 30 to 40% of patients with PBC do not have a satisfactory response to UCDA; about half of these patients will have significant improvement with obeticholic acid. Patients with PBC require long-term follow-up by a physician experienced with the disease. Certain patients may need to be considered for liver transplantation should their liver disease decompensate.

The main symptoms of PBC are fatigue and pruritus, and symptom management is important. Several therapies have been tried for treatment of fatigue, but none of them has been successful; frequent naps should be encouraged. Pruritus is treated with antihistamines, narcotic receptor antagonists (naltrexone), and rifampin. Cholestyramine, a bile salt–sequestering agent, has been helpful in some patients but is somewhat tedious and difficult to take. Plasmapheresis has been used rarely in patients with severe intractable pruritus. There is an increased incidence of osteopenia and osteoporosis in patients with cholestatic liver disease, and bone density testing should be performed. Treatment with a bisphosphonate should be instituted when bone disease is identified.

PRIMARY SCLEROSING CHOLANGITIS

As in PBC, the cause of PSC remains unknown. PSC is a chronic cholestatic syndrome that is characterized by diffuse inflammation and fibrosis involving the entire biliary tree, resulting in chronic cholestasis. This pathologic process ultimately results in obliteration of both the intra- and extrahepatic biliary tree, leading to biliary cirrhosis, portal hypertension, and liver failure. The cause of PSC remains unknown despite extensive investigation into various mechanisms related to bacterial and viral infections, toxins, genetic predisposition, and immunologic mechanisms, all of which have been postulated to contribute to the pathogenesis and progression of this syndrome.

Pathologic changes that can occur in PSC show bile duct proliferation as well as ductopenia and fibrous cholangitis (pericholangitis). Often, liver biopsy changes in PSC are not pathognomonic, and establishing the diagnosis of PSC must involve imaging of the biliary tree. Periductal fibrosis is occasionally seen on biopsy specimens and can be quite helpful in making the diagnosis. As the disease progresses, biliary cirrhosis is the final, end-stage manifestation of PSC.

Clinical Features

The usual clinical features of PSC are those found in cholestatic liver disease, with fatigue, pruritus, steatorrhea, deficiencies of fat-soluble vitamins, and the associated consequences. As in PBC, the fatigue is profound and nonspecific. Pruritus can often be debilitating and is related to the cholestasis. The severity of pruritus does not correlate with the severity of the disease. Metabolic bone disease, as seen in PBC, can occur with PSC and should be treated (see above).

Laboratory Findings

Patients with PSC typically are identified in the course of an evaluation of abnormal liver enzymes. Most patients have at least a twofold increase in ALP and may have elevated aminotransferases as well. Albumin levels may be decreased, and prothrombin times are prolonged in a substantial proportion of patients at the time of diagnosis. Some degree of correction of a prolonged prothrombin time may occur with parenteral vitamin K. A small subset of patients have aminotransferase elevations greater than five times the upper limit of normal and may have features of AIH on biopsy. These individuals are thought to have an overlap syndrome between PSC and AIH. Autoantibodies are frequently positive in patients with the overlap syndrome, but are typically negative in patients who only have PSC. One autoantibody, the perinuclear antineutrophil cytoplasmic antibody (p-ANCA), is positive in about 65% of patients with PSC. Over 50% of patients with PSC also have ulcerative colitis (UC); accordingly, once a diagnosis of PSC is established, colonoscopy should be performed to look for evidence of UC.

Diagnosis

The definitive diagnosis of PSC requires cholangiographic imaging. Over the last several years, magnetic resonance imaging (MRI) with magnetic resonance cholangiopancreatography (MRCP) has been used as the imaging technique of choice for initial evaluation. Once patients are screened in this manner, some investigators feel that endoscopic retrograde cholangiopancreatography (ERCP) should also be performed to be certain whether or not a dominant stricture is present. Typical cholangiographic findings in PSC are multifocal stricturing and beading involving both the intrahepatic and extrahepatic biliary tree. However, although involvement may be of the intrahepatic bile ducts alone or of the extrahepatic bile ducts alone, more commonly, both are involved. These strictures are typically short and with intervening segments of normal or slightly dilated bile ducts that are distributed diffusely, producing the classic beaded appearance. The gallbladder and cystic duct can be involved in up to 15% of cases. Patients with high-grade, diffuse stricturing of the intrahepatic bile ducts have an overall poor prognosis. Gradually, biliary cirrhosis develops, and patients will progress to decompensated liver disease with all the manifestations of ascites, esophageal variceal hemorrhage, and encephalopathy.

TREATMENT

TREATMENT

Primary Sclerosing Cholangitis

There is no specific proven treatment for PSC. A recently completed study of high-dose (20 mg/kg per day) UDCA was found to be harmful. Some clinicians use UDCA at “PBC dosages” of 13–15 mg/kg per day with anecdotal improvement. Endoscopic dilatation of dominant strictures can be helpful, but the ultimate treatment is liver transplantation. A dreaded complication of PSC is the development of cholangiocarcinoma, which is a relative contraindication to liver transplantation. Symptoms of pruritus are common, and the approach is as mentioned previously for this problem in patients with PBC (see above).

CARDIAC CIRRHOSIS

Definition

Patients with long-standing right-sided congestive heart failure may develop chronic liver injury and cardiac cirrhosis. This is an increasingly uncommon, if not rare, cause of chronic liver disease given the advances made in the care of patients with heart failure.

Etiology and Pathology

In the case of long-term right-sided heart failure, there is an elevated venous pressure transmitted via the inferior vena cava and hepatic veins to the sinusoids of the liver, which become dilated and engorged with blood. The liver becomes enlarged and swollen, and with long-term passive congestion and relative ischemia due to poor circulation, centrilobular hepatocytes can become necrotic, leading to pericentral fibrosis. This fibrotic pattern can extend to the periphery of the lobule outward until a unique pattern of fibrosis causing cirrhosis can occur.

Clinical Features

Patients typically have signs of congestive heart failure and will manifest an enlarged firm liver on physical examination. ALP levels are characteristically elevated, and aminotransferases may be normal or slightly increased with AST usually higher than ALT. It is unlikely that patients will develop variceal hemorrhage or encephalopathy.

Diagnosis

The diagnosis is usually made in someone with clear-cut cardiac disease who has an elevated ALP and an enlarged liver. Liver biopsy shows a pattern of fibrosis that can be recognized by an experienced hepatopathologist. Differentiation from Budd-Chiari syndrome (BCS) can be made by seeing extravasation of red blood cells in BCS, but not in cardiac hepatopathy. Venoocclusive disease can also affect hepatic outflow and has characteristic features on liver biopsy. Venoocclusive disease can be seen under the circumstances of conditioning for bone marrow transplant with radiation and chemotherapy; it can also be seen with the ingestion of certain herbal teas as well as pyrrolizidine alkaloids. This is typically seen in Caribbean countries and rarely in the United States. Treatment is based on management of the underlying cardiac disease.

OTHER TYPES OF CIRRHOSIS

There are several other less common causes of chronic liver disease that can progress to cirrhosis. These include inherited metabolic liver diseases such as hemochromatosis, Wilson's disease, α_1 antitrypsin (α_1AT) deficiency, and cystic fibrosis. For all of these disorders, the manifestations of cirrhosis are similar, with some minor variations, to those seen in other patients with other causes of cirrhosis.

Hemochromatosis is an inherited disorder of iron metabolism that results in a progressive increase in hepatic iron deposition, which, over time, can lead to a portal-based fibrosis progressing to cirrhosis, liver failure, and hepatocellular cancer. While the frequency of hemochromatosis is relatively common, with genetic susceptibility occurring in 1 in 250 individuals, the frequency of end-stage manifestations due to the disease is relatively low, and fewer than 5% of those patients who are genotypically susceptible will go on to develop severe liver disease from hemochromatosis. Diagnosis is made with serum iron studies showing an elevated transferrin saturation and an elevated ferritin level, along with abnormalities identified by *HFE* mutation analysis. Treatment is straightforward, with regular therapeutic phlebotomy.

Wilson's disease is an inherited disorder of copper homeostasis with failure to excrete excess amounts of copper, leading to an accumulation in the liver. This disorder is relatively uncommon, affecting 1 in 30,000 individuals. Wilson's disease typically affects adolescents and young adults. Prompt diagnosis before end-stage manifestations become irreversible can lead to significant clinical improvement. Diagnosis requires determination of ceruloplasmin levels, which are low; 24-h urine copper levels, which are elevated; typical physical examination findings, including Kayser-Fleischer corneal rings; and characteristic liver biopsy findings. Treatment consists of copper-chelating medications.

α_1*AT deficiency* results from an inherited disorder that causes abnormal folding of the α_1AT protein, resulting in failure of secretion of that protein from the liver. It is unknown how the retained protein leads to liver disease. Patients with α_1AT deficiency at greatest risk for developing chronic liver disease have the ZZ phenotype, but only about 10–20% of such individuals will develop chronic liver disease. Diagnosis is made by determining α_1AT levels and phenotype. Characteristic periodic acid–Schiff (PAS)-positive, diastase-resistant globules are seen on liver biopsy. The only effective treatment is liver transplantation, which is curative.

Cystic fibrosis is an uncommon inherited disorder affecting whites of northern European descent. A biliary-type cirrhosis can occur, and some patients derive benefit from the chronic use of UDCA.

MAJOR COMPLICATIONS OF CIRRHOSIS

The clinical course of patients with advanced cirrhosis is often complicated by a number of important sequelae that can occur regardless of the underlying cause of the liver disease. These include portal hypertension and its consequences of gastroesophageal variceal hemorrhage, splenomegaly, ascites, hepatic encephalopathy, spontaneous bacterial peritonitis (SBP), hepatorenal syndrome (HRS), and hepatocellular carcinoma (Table 337-2).

TABLE 337-2

Complications of Cirrhosis

Portal hypertension	Coagulopathy
Gastroesophageal varices	Factor deficiency
Portal hypertensive gastropathy	Fibrinolysis
Splenomegaly, hypersplenism	Thrombocytopenia
Ascites	Bone disease
Spontaneous bacterial peritonitis	Osteopenia
Hepatorenal syndrome	Osteoporosis
Type 1	Osteomalacia
Type 2	Hematologic abnormalities
Hepatic encephalopathy	Anemia
Hepatopulmonary syndrome	Hemolysis
Portopulmonary hypertension	Thrombocytopenia
Malnutrition	Neutropenia

PORTAL HYPERTENSION

Portal hypertension is defined as the elevation of the hepatic venous pressure gradient (HVPG) to >5 mmHg. Portal hypertension is caused by a combination of two simultaneously occurring hemodynamic processes: (1) increased intrahepatic resistance to the passage of blood flow through the liver due to cirrhosis and regenerative nodules, and (2) increased splanchnic blood flow secondary to vasodilation within the splanchnic vascular bed. Portal hypertension is directly responsible for the two major complications of cirrhosis: variceal hemorrhage and ascites. *Variceal hemorrhage* is an immediate life-threatening problem with a 20–30% mortality rate associated with each episode of bleeding. The portal venous system normally drains blood from the stomach, intestines, spleen, pancreas, and gallbladder, and the portal vein is formed by the confluence of the superior mesenteric and splenic veins. Deoxygenated blood from the small bowel

drains into the superior mesenteric vein along with blood from the head of the pancreas, the ascending colon, and part of the transverse colon. Conversely, the splenic vein drains the spleen and the pancreas and is joined by the inferior mesenteric vein, which brings blood from the transverse and descending colon as well as from the superior two-thirds of the rectum. Thus, the portal vein normally receives blood from almost the entire GI tract.

The causes of portal hypertension are usually subcategorized as prehepatic, intrahepatic, and posthepatic (Table 337-3). Prehepatic causes of portal hypertension are those affecting the portal venous system before it enters the liver; they include portal vein thrombosis and splenic vein thrombosis. Posthepatic causes encompass those affecting the hepatic veins and venous drainage to the heart; they include BCS, venoocclusive disease, and chronic right-sided cardiac congestion. Intrahepatic causes account for over 95% of cases of portal hypertension and are represented by the major forms of cirrhosis. Intrahepatic causes of portal hypertension can be further subdivided into presinusoidal, sinusoidal, and postsinusoidal causes. Postsinusoidal causes include venoocclusive disease, whereas presinusoidal causes include congenital hepatic fibrosis and schistosomiasis. Sinusoidal causes are related to cirrhosis from various causes.

TABLE 337-3

Classification of Portal Hypertension

Prehepatic
Portal vein thrombosis
Splenic vein thrombosis
Massive splenomegaly (Banti's syndrome)
Hepatic
Presinusoidal
Schistosomiasis
Congenital hepatic fibrosis
Sinusoidal
Cirrhosis—many causes
Alcoholic hepatitis
Postsinusoidal
Hepatic sinusoidal obstruction (venoocclusive syndrome)
Posthepatic
Budd-Chiari syndrome
Inferior vena caval webs
Cardiac causes
Restrictive cardiomyopathy
Constrictive pericarditis
Severe congestive heart failure

Cirrhosis is the most common cause of portal hypertension in the United States, and clinically significant portal hypertension is present in >60% of patients with cirrhosis. Portal vein obstruction may be idiopathic or can occur in association with cirrhosis or with infection, pancreatitis, or abdominal trauma.

Coagulation disorders that can lead to the development of portal vein thrombosis include polycythemia vera; essential thrombocytosis; deficiencies in protein C, protein S, antithrombin 3, and factor V Leiden; and abnormalities in the gene-regulating prothrombin production. Some patients may have a subclinical myeloproliferative disorder.

Clinical Features

The three primary complications of portal hypertension are gastroesophageal varices with hemorrhage, ascites, and hypersplenism. Thus, patients may present with upper GI bleeding, which, on endoscopy, is found to be due to esophageal or gastric varices; with the development of ascites along with peripheral edema; or with an enlarged spleen with associated reduction in platelets and white blood cells on routine laboratory testing.

ESOPHAGEAL VARICES

Over the last decade, it has become common practice to screen known cirrhotics with endoscopy to look for esophageal varices. Such screening studies have shown that approximately one-third of patients with histologically confirmed cirrhosis have varices. Approximately 5–15% of cirrhotics per year develop varices, and it is estimated that the majority of patients with cirrhosis will develop varices over their lifetimes. Furthermore, it is anticipated that roughly one-third of patients with varices will develop bleeding. Several factors predict the risk of bleeding, including the severity of cirrhosis (Child's class, MELD score); the height of wedged-hepatic vein pressure; the size of the varix; the location of the varix; and certain endoscopic stigmata, including red wale signs, hematocystic spots, diffuse erythema, bluish color, cherry red spots, or white-nipple spots. Patients with tense ascites are also at increased risk for bleeding from varices.

Diagnosis

In patients with cirrhosis who are being followed chronically, the development of portal hypertension is usually revealed by the presence of thrombocytopenia; the appearance of an enlarged spleen; or the development of ascites, encephalopathy, and/or esophageal varices with or without bleeding. In previously undiagnosed patients, any of these features should prompt further evaluation to determine the presence of portal hypertension and liver disease. Varices should be identified by endoscopy. Abdominal imaging, either by computed tomography (CT) or MRI, can be helpful in demonstrating a nodular liver and in finding changes of portal hypertension with intraabdominal collateral circulation. If necessary, interventional radiologic procedures can be performed to determine wedged and free hepatic vein pressures that will allow for the calculation of a wedged-to-free gradient, which is equivalent to the portal pressure. The average normal wedged-to-free gradient is 5 mmHg, and patients with a gradient >12 mmHg are at risk for variceal hemorrhage.

TREATMENT

TREATMENT

Variceal Hemorrhage

Treatment for variceal hemorrhage as a complication of portal hypertension is divided into two main categories: (1) primary prophylaxis and (2) prevention of rebleeding once there has been an initial variceal hemorrhage. Primary prophylaxis requires routine screening by endoscopy of all patients with cirrhosis. Once varices that are at increased risk for bleeding are identified, primary prophylaxis can be achieved either through nonselective beta blockade or by variceal band ligation. Numerous placebo-controlled clinical trials of either propranolol or nadolol have been reported in the literature. The most rigorous studies were those that only included patients with significantly enlarged varices or with hepatic vein pressure gradients >12 mmHg. Patients treated with beta blockers have a lower risk of variceal hemorrhage than those treated with placebo over 1 and 2 years of follow-up. There is also a decrease in mortality related to variceal hemorrhage. Unfortunately, overall survival was improved in only one study. Further studies have demonstrated that the degree of reduction of portal pressure is a significant feature to determine success of therapy. Therefore, it has been suggested that repeat measurements of hepatic vein pressure gradients may be used to guide pharmacologic therapy; however, this may be cost-prohibitive. Several studies have evaluated variceal band ligation and variceal sclerotherapy as methods for providing primary prophylaxis.

Endoscopic variceal ligation (EVL) has achieved a level of success and comfort with most gastroenterologists who see patients with these complications of portal hypertension. Thus, in patients with cirrhosis who are screened for portal hypertension and are found to have large varices, it is recommended that they receive either beta blockade or primary prophylaxis with EVL.

The approach to patients once they have had a variceal bleed is first to treat the acute bleed, which can be life-threatening, and then to prevent further bleeding. Prevention of further bleeding is usually accomplished with repeated variceal band ligation until varices are obliterated. Treatment of acute bleeding requires both fluid and blood-product replacement as well as prevention of subsequent bleeding with EVL.

The medical management of acute variceal hemorrhage includes the use of vasoconstricting agents, usually somatostatin or octreotide. Vasopressin was used in the past but is no longer commonly used. Balloon tamponade (Sengstaken-Blakemore tube or Minnesota tube) can be used in patients who cannot get endoscopic therapy immediately or who need stabilization prior to endoscopic therapy. Control of bleeding can be achieved in the vast majority of cases; however, bleeding recurs in the majority of patients if definitive endoscopic therapy has not been instituted. Octreotide, a direct splanchnic vasoconstrictor, is given at dosages of 50–100 µg/h by continuous infusion. Endoscopic intervention is used as first-line treatment to control bleeding acutely. Some endoscopists will use variceal injection therapy (sclerotherapy) as initial therapy, particularly when bleeding is vigorous. Variceal band ligation is used to control acute bleeding in over 90% of cases and should be repeated until obliteration of all varices is accomplished. When esophageal varices extend into the proximal stomach, band ligation is less successful. In these situations, when bleeding continues from gastric varices, consideration for a transjugular intrahepatic portosystemic shunt (TIPS)

should be made. This technique creates a portosystemic shunt by a percutaneous approach using an expandable metal stent, which is advanced under angiographic guidance to the hepatic veins and then through the substance of the liver to create a direct portocaval shunt. This offers an alternative to surgery for acute decompression of portal hypertension. Encephalopathy can occur in as many as 20% of patients after TIPS and is particularly problematic in elderly patients and in patients with preexisting encephalopathy. TIPS should be reserved for individuals who fail endoscopic or medical management or who are poor surgical risks. TIPS can sometimes be used as a bridge to transplantation. Surgical esophageal transsection is a procedure that is rarely used and generally is associated with a poor outcome.

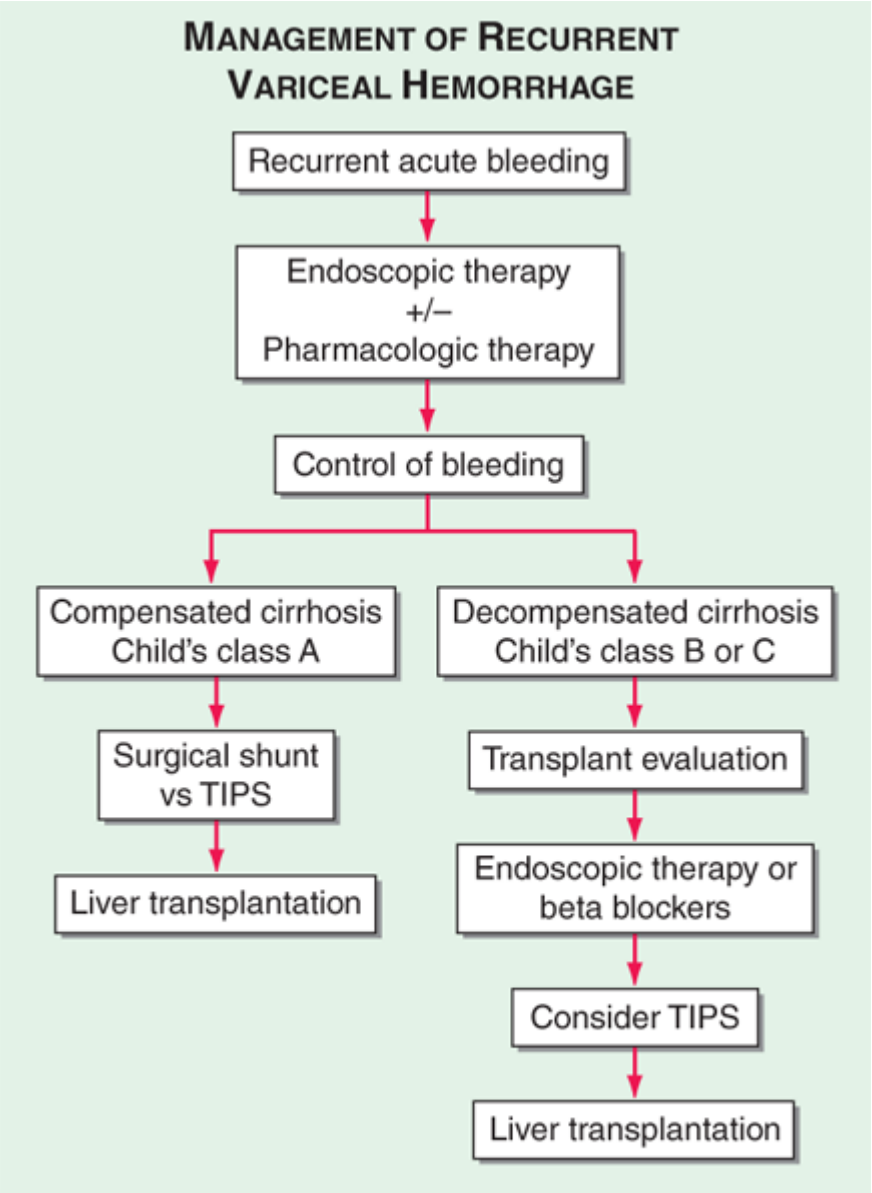
MANAGEMENT OF RECURRENT VARICEAL HEMORRHAGE

Once patients have had an acute bleed and have been managed successfully, attention should be paid to preventing recurrent bleeding. This usually requires repeated variceal band ligation until varices are obliterated. Beta blockade may be of adjunctive benefit in patients who are having recurrent variceal band ligation; however, once varices have been obliterated, the need for beta blockade is lessened. Despite successful variceal obliteration, many patients will still have portal hypertensive gastropathy from which bleeding can occur. Nonselective beta blockade may be helpful to prevent further bleeding from portal hypertensive gastropathy once varices have been obliterated **(Fig. 337-3)**.

Portosystemic shunt surgery is less commonly performed with the advent of TIPS; nonetheless, this procedure should be considered for patients with good hepatic synthetic function who could benefit by having portal decompressive surgery.

FIGURE 337-3

Management of recurrent variceal hemorrhage. This algorithm describes an approach to management of patients who have recurrent bleeding from esophageal varices. Initial therapy is generally with endoscopic therapy often supplemented by pharmacologic therapy. With control of bleeding, a decision needs to be made as to whether patients should go on to a surgical shunt or TIPS (if they are Child's class A) and be considered for transplant, or if they should have TIPS and be considered for transplant (if they are Child's class B or C). TIPS, transjugular intrahepatic portosystemic shunt.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SPLENOMEGALY AND HYPERSPLENISM

Congestive splenomegaly is common in patients with portal hypertension. Clinical features include the presence of an enlarged spleen on physical examination and the development of thrombocytopenia and leukopenia in patients who have cirrhosis. Some patients will have fairly significant left-sided and left upper quadrant abdominal pain related to an enlarged and engorged spleen. Splenomegaly itself usually requires no specific treatment, although splenectomy can be successfully performed under very special circumstances.

Hypersplenism with the development of thrombocytopenia is a common feature of patients with cirrhosis and is usually the first indication of portal hypertension.

ASCITES

Definition

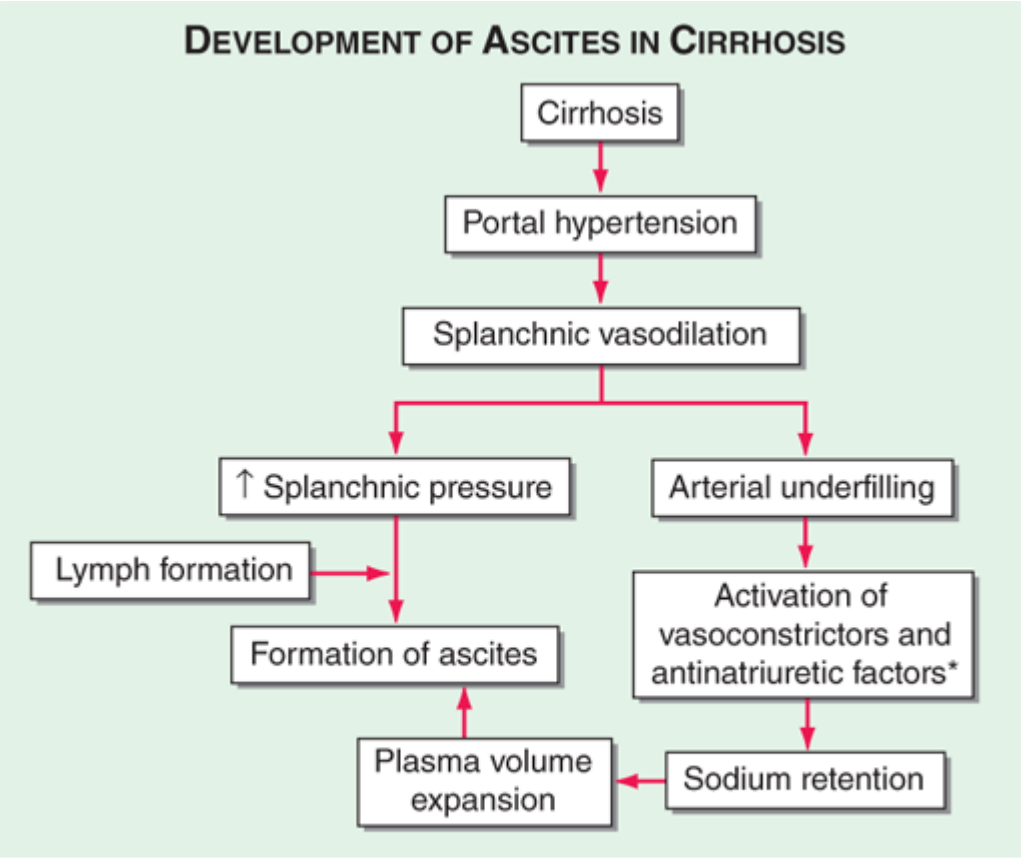
Ascites is the accumulation of fluid within the peritoneal cavity. Overwhelmingly, the most common cause of ascites is portal hypertension related to cirrhosis; however, clinicians should remember that malignant or infectious causes of ascites can be present as well, and careful differentiation of these other causes are obviously important for patient care.

Pathogenesis

The presence of portal hypertension contributes to the development of ascites in patients who have cirrhosis (Fig. 337-4). There is an increase in intrahepatic resistance, causing increased portal pressure, but there is also vasodilation of the splanchnic arterial system, which, in turn, results in an increase in portal venous inflow. Both of these abnormalities result in increased production of splanchnic lymph. Vasodilating factors such as nitric oxide are responsible for the vasodilatory effect. These hemodynamic changes result in sodium retention by causing activation of the renin-angiotensin-aldosterone system with the development of hyperaldosteronism. The renal effects of increased aldosterone leading to sodium retention also contribute to the development of ascites. Sodium retention causes fluid accumulation and expansion of the extracellular fluid volume, which results in the formation of peripheral edema and ascites. Sodium retention is the consequence of a homeostatic response caused by underfilling of the arterial circulation secondary to arterial vasodilation in the splanchnic vascular bed. Because the retained fluid is constantly leaking out of the intravascular compartment into the peritoneal cavity, the sensation of vascular filling is not achieved, and the process continues. Hypoalbuminemia and reduced plasma oncotic pressure also contribute to the loss of fluid from the vascular compartment into the peritoneal cavity. Hypoalbuminemia is due to decreased synthetic function in a cirrhotic liver.

FIGURE 337-4

Development of ascites in cirrhosis. This flow diagram illustrates the importance of portal hypertension with splanchnic vasodilation in the development of ascites. *Antinatriuretic factors include the renin-angiotensin-aldosterone system and the sympathetic nervous system.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical Features

Patients typically note an increase in abdominal girth that is often accompanied by the development of peripheral edema. The development of ascites is often insidious, and it is surprising that some patients wait so long and become so distended before seeking medical attention. Patients usually have at least 1–2 L of fluid in the abdomen before they are aware that there is an increase. If ascitic fluid is massive, respiratory function can be compromised, and patients will complain of shortness of breath. Hepatic hydrothorax may also occur in this setting, contributing to respiratory symptoms. Patients with massive ascites are often malnourished and have muscle wasting and excessive fatigue and weakness.

Diagnosis

Diagnosis of ascites is by physical examination and is often aided by abdominal imaging. Patients will have bulging flanks, may have a fluid wave, or may have the presence of shifting dullness. This is determined by taking patients from a supine position to lying on either their left or right side and noting the movement of the dullness to percussion. Subtle amounts of ascites can be detected by ultrasound or CT scanning. Hepatic hydrothorax is more common on the right side and implicates a rent in the diaphragm with free flow of ascitic fluid into the thoracic cavity.

When patients present with ascites for the first time, it is recommended that a diagnostic paracentesis be performed to characterize the fluid. This should include the determination of total protein and albumin content, blood cell counts with differential, and cultures. In the appropriate setting, amylase may be measured and cytology performed. In patients with cirrhosis, the protein concentration of the ascitic fluid is quite low, with the majority of patients having an ascitic fluid protein concentration <1 g/dL. The development of the serum ascites-to-albumin gradient (SAAG) has replaced the description of exudative or transudative fluid. When the gradient between the serum albumin level and the ascitic fluid albumin level is >1.1 g/dL, the cause of the ascites is most likely due to portal hypertension; this is usually in the setting of cirrhosis. When the gradient is <1.1 g/dL, infectious or malignant causes of ascites should be considered. When levels of ascitic fluid proteins are very low, patients are at increased risk for developing SBP. A high level of red blood cells in the ascitic fluid signifies a traumatic tap or perhaps a hepatocellular cancer or a ruptured omental varix. When the absolute level of polymorphonuclear leukocytes is >250/μL, the question of ascitic fluid infection should be strongly considered. Ascitic fluid cultures should be obtained using bedside inoculation of culture media.

TREATMENT

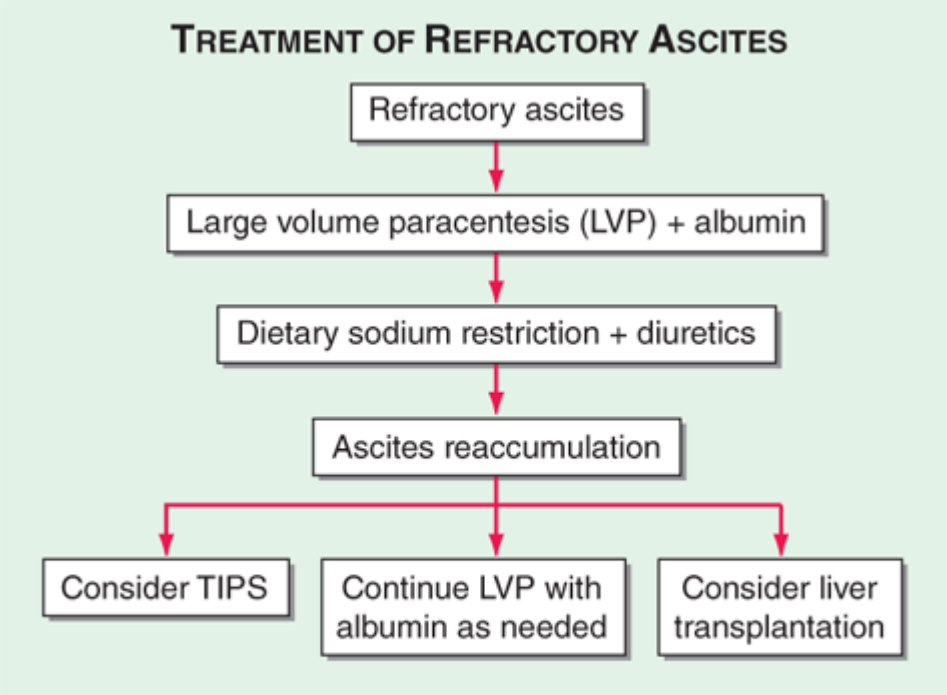
TREATMENT

Ascites

Patients with small amounts of ascites can usually be managed with dietary sodium restriction alone. Most average diets in the United States contain 6–8 g of sodium per day, and if patients eat at restaurants or fast-food outlets, the amount of sodium in their diet can exceed this amount. Thus, it is often extremely difficult to get patients to change their dietary habits to ingest <2 g of sodium per day, which is the recommended amount. Patients are frequently surprised to realize how much sodium is in the standard U.S. diet; thus, it is important to make educational pamphlets available to the patient. Often, a simple recommendation is to eat fresh or frozen foods, avoiding canned or processed foods, which are usually preserved with sodium. When a moderate amount of ascites is present, diuretic therapy is usually necessary. Traditionally, spironolactone at 100–200 mg/d as a single dose is started, and furosemide may be added at 40–80 mg/d, particularly in patients who have peripheral edema. In patients who have never received diuretics before, the failure of the above-mentioned dosages suggests that they are not being compliant with a low-sodium diet. If compliance is confirmed and ascitic fluid is not being mobilized, spironolactone can be increased to 400–600 mg/d and furosemide increased to 120–160 mg/d. If ascites is still present with these dosages of diuretics in patients who are compliant with a low-sodium diet, then they are defined as having *refractory ascites*, and alternative treatment modalities including repeated large-volume paracentesis or a TIPS procedure should be considered **(Fig. 337-5)**. Recent studies have shown that TIPS, while managing the ascites, does not improve survival in these patients. Unfortunately, TIPS is often associated with an increased frequency of hepatic encephalopathy and must be considered carefully on a case-by-case basis. The prognosis for patients with cirrhosis with ascites is poor, and some studies have shown that <50% of patients survive 2 years after the onset of ascites. Thus, there should be consideration for liver transplantation in patients with the onset of ascites.

FIGURE 337-5

Treatment of refractory ascites. In patients who develop azotemia in the course of receiving diuretics in the management of their ascites, some will require repeated large-volume paracentesis (LVP), some may be considered for transjugular intrahepatic portosystemic shunt (TIPS), and some would be good candidates for liver transplantation. These decisions are all individualized.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SPONTANEOUS BACTERIAL PERITONITIS

SBP is a common and severe complication of ascites characterized by spontaneous infection of the ascitic fluid without an intraabdominal source. In patients with cirrhosis and ascites severe enough for hospitalization, SBP can occur in up to 30% of individuals and can have a 25% in-hospital mortality rate. Bacterial translocation is the presumed mechanism for development of SBP, with gut flora traversing the intestine into mesenteric lymph nodes, leading to bacteremia and seeding of the ascitic fluid. The most common organisms are *Escherichia coli* and other gut bacteria; however, gram-positive bacteria, including *Streptococcus viridans*, *Staphylococcus aureus*, and *Enterococcus* sp., can also be found. If more than two organisms are identified, secondary bacterial peritonitis due to a perforated viscus should be considered. The diagnosis of SBP is made when the fluid sample has an absolute neutrophil count >250/μL. Bedside cultures should be obtained when ascitic fluid is tapped. Patients with ascites may present with fever, altered mental status, elevated white blood cell count, and abdominal pain or discomfort, or they may present without any of these features. Therefore, it is necessary to have a high degree of clinical suspicion, and peritoneal taps are important for making the diagnosis. Treatment is commonly with a third-generation cephalosporin. In patients with variceal hemorrhage, the frequency of SBP is significantly increased, and prophylaxis against SBP is recommended when a patient presents with upper GI bleeding. Furthermore, in

patients who have had an episode(s) of SBP and recovered, once-weekly administration of antibiotics is used as prophylaxis for recurrent SBP.

HEPATORENAL SYNDROME

HRS is a form of functional renal failure without renal pathology that occurs in about 10% of patients with advanced cirrhosis or acute liver failure. There are marked disturbances in the arterial renal circulation in patients with HRS; these include an increase in vascular resistance accompanied by a reduction in systemic vascular resistance. The reason for renal vasoconstriction is most likely multifactorial and is poorly understood. The diagnosis is made usually in the presence of a large amount of ascites in patients who have a stepwise progressive increase in creatinine. Type 1 HRS is characterized by a progressive impairment in renal function and a significant reduction in creatinine clearance within 1–2 weeks of presentation. Type 2 HRS is characterized by a reduction in glomerular filtration rate with an elevation of serum creatinine level, but it is fairly stable and is associated with a better outcome than that of type 1 HRS.

HRS is often seen in patients with refractory ascites and requires exclusion of other causes of acute renal failure. Treatment has, unfortunately, been difficult, and in the past, dopamine or prostaglandin analogues were used as renal vasodilating medications. Carefully performed studies have failed to show clear-cut benefit from these therapeutic approaches. Currently, patients are treated with midodrine, an α-agonist, along with octreotide and intravenous albumin. The best therapy for HRS is liver transplantation; recovery of renal function is typical in this setting. In patients with either type 1 or type 2 HRS, the prognosis is poor unless transplant can be achieved within a short period of time.

HEPATIC ENCEPHALOPATHY

Portosystemic encephalopathy is a serious complication of chronic liver disease and is broadly defined as an alteration in mental status and cognitive function occurring in the presence of liver failure. In acute liver injury with fulminant hepatic failure, the development of encephalopathy is a requirement for a diagnosis of fulminant failure. Encephalopathy is much more commonly seen in patients with chronic liver disease. Gut-derived neurotoxins that are not removed by the liver because of vascular shunting and decreased hepatic mass get to the brain and cause the symptoms that we know of as hepatic encephalopathy. Ammonia levels are typically elevated in patients with hepatic encephalopathy, but the correlation between severity of liver disease and height of ammonia levels is often poor, and most hepatologists do not rely on ammonia levels to make a diagnosis. Other compounds and metabolites that may contribute to the development of encephalopathy include certain false neurotransmitters and mercaptans.

Clinical Features

In acute liver failure, changes in mental status can occur within weeks to months. Brain edema can be seen in these patients, with severe encephalopathy associated with swelling of the gray matter. Cerebral herniation is a feared complication of brain edema in acute liver failure, and treatment is meant to decrease edema with mannitol and judicious use of intravenous fluids.

In patients with cirrhosis, encephalopathy is often found as a result of certain precipitating events such as hypokalemia, infection, an increased dietary protein load, or electrolyte disturbances. Patients may be confused or exhibit a change in personality. They may actually be quite violent and difficult to manage; alternatively, patients may be very sleepy and difficult to rouse. Because precipitating events are so commonly found, they should be sought carefully. If patients have ascites, this should be tapped to rule out infection. Evidence of GI bleeding should be sought, and patients should be appropriately hydrated. Electrolytes should be measured and abnormalities corrected. In patients presenting with encephalopathy, asterixis is often present. Asterixis can be elicited by having patients extend their arms and bend their wrists back. In this maneuver, patients who are encephalopathic have a “liver flap”—that is, a sudden forward movement of the wrist. This requires patients to be able to cooperate with the examiner and obviously cannot be elicited in patients who are severely encephalopathic or in hepatic coma.

The diagnosis of hepatic encephalopathy is clinical and requires an experienced clinician to recognize and put together all of the various features. Often when patients have encephalopathy for the first time, they (and/or their caregivers) are unaware of what is transpiring, but once they have been through the experience for the first time, they can identify when this is developing in subsequent situations and can often self-medicate to impair the development or worsening of encephalopathy.

TREATMENT

TREATMENT

Hepatic Encephalopathy

Treatment is multifactorial and includes management of the above-mentioned precipitating factors. Sometimes hydration and correction of electrolyte imbalance are all that is necessary. In the past, restriction of dietary protein was considered for patients with encephalopathy; however, the negative impact of that maneuver on overall nutrition is thought to outweigh the benefit when treating encephalopathy, and it is thus discouraged. There may be some benefit to replacing animal-based protein with vegetable-based protein in some patients with encephalopathy that is difficult to manage. The mainstay of treatment for encephalopathy, in addition to correcting precipitating factors, is to use lactulose, a nonabsorbable disaccharide, which results in colonic acidification. Catharsis ensues, contributing to the elimination of nitrogenous products in the gut that are responsible for the development of encephalopathy. The goal of lactulose therapy is to promote 2–3 soft stools per day. Patients are asked to titrate their amount of ingested lactulose to achieve the desired effect. Poorly absorbed antibiotics are often used as adjunctive therapies for patients who have a difficult time with lactulose. The alternating administration of neomycin and metronidazole has been used in the past to reduce the individual side effects of each: neomycin for renal insufficiency and ototoxicity and metronidazole for peripheral neuropathy. More recently, rifaximin at 550 mg twice daily has been very effective in treating encephalopathy without the known side effects of neomycin or metronidazole. Zinc supplementation is sometimes helpful in patients with encephalopathy and is relatively harmless. The development of encephalopathy in patients with chronic liver disease is a poor prognostic sign, but this complication can be managed in the vast majority of patients.

MALNUTRITION IN CIRRHOSIS

Because the liver is principally involved in the regulation of protein and energy metabolism in the body, it is not surprising that patients with advanced liver disease are commonly malnourished. Once patients become cirrhotic, they are more catabolic, and muscle protein is metabolized. There are multiple factors that contribute to the malnutrition of cirrhosis, including poor dietary intake, alterations in gut nutrient absorption, and alterations in protein metabolism. Dietary supplementation for patients with cirrhosis is helpful in preventing patients from becoming catabolic.

ABNORMALITIES IN COAGULATION

Coagulopathy is almost universal in patients with cirrhosis. There is decreased synthesis of clotting factors and impaired clearance of anticoagulants. In addition, patients may have thrombocytopenia from hypersplenism due to portal hypertension. Vitamin K–dependent clotting factors are factors II, VII, IX, and X. Vitamin K requires biliary excretion for its subsequent absorption; thus, in patients with chronic cholestatic syndromes, vitamin K absorption is frequently diminished. Intravenous or intramuscular vitamin K can quickly correct this abnormality. More commonly, the synthesis of vitamin K–dependent clotting factors is diminished because of a decrease in hepatic mass, and, under these circumstances, administration of parenteral vitamin K does not improve the clotting factors or the prothrombin time. Platelet function is often abnormal in patients with chronic liver disease, in addition to decreases in platelet levels due to hypersplenism.

BONE DISEASE IN CIRRHOSIS

Osteoporosis is common in patients with chronic cholestatic liver disease because of malabsorption of vitamin D and decreased calcium ingestion. The rate of bone resorption exceeds that of new bone formation in patients with cirrhosis, resulting in bone loss. Dual x-ray absorptiometry (DEXA) is a useful method for determining osteoporosis or osteopenia in patients with chronic liver disease. When a DEXA scan shows decreased bone mass, treatment should be administered with bisphosphonates that are effective at inhibiting resorption of bone and efficacious in the treatment of osteoporosis.

HEMATOLOGIC ABNORMALITIES IN CIRRHOSIS

Numerous hematologic manifestations of cirrhosis are present, including anemia from a variety of causes including hypersplenism, hemolysis, iron deficiency, and perhaps folate deficiency from malnutrition. Macrocytosis is a common abnormality in red blood cell morphology seen in patients with chronic liver disease, and neutropenia may be seen as a result of hypersplenism.

FURTHER READING

AASLD/IDSA HCV Guidance Panel: Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology 62:932, 2015. [PubMed: 26111063]

Garcia-Tsao G et al: Prevention and management of gastroesophageal varices and variceal hemorrhage in cirrhosis. Am J Gastroenterol 102:2086, 2007. [PubMed: 17727436]

Runyon BA: Introduction to the revised American Association for the Study of Liver Diseases practice guideline management of adult patients with ascites due to cirrhosis: Update 2012. Hepatology 57:1651, 2013. [PubMed: 23463403]

Terrault NA et al: AASLD guidelines for treatment of chronic hepatitis B. Hepatology 63:261, 2016. [PubMed: 26566064]

Vilstrup H et al: Hepatic encephalopathy in chronic liver disease: 2014 Practice Guideline by the American Association for the Study of Liver Diseases and the European Association for the Study of the Liver. Hepatology 60:715, 2014. [PubMed: 25042402]

Chapter 338: Liver Transplantation

Raymond T. Chung; Jules L. Dienstag

FIGURE 338-1

INTRODUCTION

Liver transplantation—the replacement of the native, diseased liver by a normal organ (allograft)—has matured from an experimental procedure reserved for desperately ill patients to an accepted, lifesaving operation applied more optimally in the natural history of end-stage liver disease. The preferred and technically most advanced approach is *orthotopic transplantation*, in which the native organ is removed and the donor organ is inserted in the same anatomic location. Pioneered in the 1960s by Thomas Starzl at the University of Colorado and, later, at the University of Pittsburgh and by Roy Calne in Cambridge, England, liver transplantation is now performed routinely worldwide. Success measured as 1-year survival has improved from ~30% in the 1970s to >90% today. These improved prospects for prolonged survival resulted from refinements in operative technique, improvements in organ procurement and preservation, advances in immunosuppressive therapy, and, perhaps most influentially, more enlightened patient selection and timing. Despite the perioperative morbidity and mortality, the technical and management challenges of the procedure, and its costs, liver transplantation has become the approach of choice for selected patients whose chronic or acute liver disease is progressive, life-threatening, and unresponsive to medical therapy. Based on the current level of success, the number of liver transplants has continued to grow each year; in 2017, 8082 patients received liver allografts in the United States. Still, the demand for new livers continues to outpace availability; as of 2018, 13,925 patients in the United States were on a waiting list for a donor liver. In response to this drastic shortage of donor organs, many transplantation centers supplement cadaver-organ liver transplantation with living-donor transplantation.

INDICATIONS

Potential candidates for liver transplantation are children and adults who, in the absence of contraindications (see below), suffer from severe, irreversible liver disease for which alternative medical or surgical treatments have been exhausted or are unavailable. *Timing of the operation is of critical importance*. Indeed, improved timing and better patient selection are felt to have contributed more to the increased success of liver transplantation in the 1980s and beyond than all the impressive technical and immunologic advances combined. Although the disease should be advanced, and although opportunities for spontaneous or medically induced stabilization or recovery should be allowed, the procedure should be done sufficiently

early to give the surgical procedure a fair chance for success. Ideally, transplantation should be considered in patients with end-stage liver disease who are experiencing or have experienced a life-threatening complication of hepatic decompensation or whose quality of life has deteriorated to unacceptable levels. Although patients with well-compensated cirrhosis can survive for many years, many patients with quasi-stable chronic liver disease have much more advanced disease than may be apparent. As discussed below, the better the status of the patient prior to transplantation, the higher will be its anticipated success rate. The decision about *when* to transplant is complex and requires the combined judgment of an experienced team of hepatologists, transplant surgeons, anesthesiologists, and specialists in support services, not to mention the well-informed consent of the patient and the patient's family.

TRANSPLANTATION IN CHILDREN

Indications for transplantation in children are listed in **Table 338-1**. The most common is *biliary atresia*. *Inherited or genetic disorders of metabolism* associated with liver failure constitute another major indication for transplantation in children and adolescents. In Crigler-Najjar disease type I and in certain hereditary disorders of the urea cycle and of amino acid or lactate-pyruvate metabolism, transplantation may be the only way to prevent impending deterioration of central nervous system function, despite the fact that the native liver is structurally normal. Combined heart and liver transplantation has yielded dramatic improvement in cardiac function and in cholesterol levels in children with homozygous familial hypercholesterolemia; combined liver and kidney transplantation has been successful in patients with primary hyperoxaluria type I. In hemophiliacs with transfusion-associated hepatitis and liver failure, liver transplantation has been associated with recovery of normal factor VIII synthesis.

TABLE 338-1

Indications for Liver Transplantation

Children	Adults
Biliary atresia	Primary biliary cirrhosis
Neonatal hepatitis	Secondary biliary cirrhosis
Congenital hepatic fibrosis	Primary sclerosing cholangitis
Alagille's syndrome[a]	Autoimmune hepatitis
Byler's disease[b]	Caroli's disease[c]
α_1-Antitrypsin deficiency	Cryptogenic cirrhosis
Inherited disorders of metabolism	Chronic hepatitis with cirrhosis
Wilson's disease	Hepatic vein thrombosis
Tyrosinemia	Fulminant hepatitis
Glycogen storage diseases	Alcoholic cirrhosis
Lysosomal storage diseases	Chronic viral hepatitis
Protoporphyria	Primary hepatocellular malignancies
Crigler-Najjar disease type I	Hepatic adenomas
Familial hypercholesterolemia	Nonalcoholic steatohepatitis
Primary hyperoxaluria type I	Familial amyloid polyneuropathy
Hemophilia	

[a]Arteriohepatic dysplasia, with paucity of bile ducts, and congenital malformations, including pulmonary stenosis. [b]Intrahepatic cholestasis, progressive liver failure, and mental and growth retardation. [c]Multiple cystic dilatations of the intrahepatic biliary tree.

TRANSPLANTATION IN ADULTS

Liver transplantation is indicated for end-stage *cirrhosis* of all causes (Table 338-1). In *sclerosing cholangitis* and *Caroli's disease* (multiple cystic dilatations of the intrahepatic biliary tree), recurrent infections and sepsis associated with inflammatory and fibrotic obstruction of the biliary tree may be an indication for transplantation. Because prior biliary surgery complicates and is a relative contraindication for liver transplantation, surgical diversion of the biliary tree has been all but abandoned for patients with sclerosing cholangitis. In patients who undergo transplantation for *hepatic vein thrombosis* (*Budd-Chiari syndrome*), postoperative anticoagulation is essential; underlying myeloproliferative disorders may have to be treated but are not a contraindication to liver transplantation. If a donor organ can be located quickly, before life-threatening complications—including cerebral edema—set in, patients with acute liver failure are candidates for liver transplantation. Routine candidates for liver transplantation are patients with *alcoholic cirrhosis*, *chronic viral hepatitis*, and *primary hepatocellular malignancies*. Although all three of these categories are considered to be high risk, liver transplantation can be offered to carefully selected patients. Currently, chronic hepatitis C and alcoholic liver disease are the most common indications for liver transplantation, accounting for over 40% of all adult candidates who undergo the procedure. Patients with alcoholic cirrhosis can be considered as candidates for transplantation if they meet strict criteria for abstinence and reform; however, these criteria still do not prevent recidivism in up to a quarter of cases. In highly selected cases in a limited number of centers, transplantation for severe *acute* alcoholic hepatitis has been performed with success; however, because patients with acute alcoholic hepatitis are still actively using alcohol, and because continued alcohol abuse remains a concern, acute alcoholic hepatitis is not a routine indication for liver transplantation. Patients with chronic hepatitis C have early allograft and patient survival comparable to those of other subsets of patients after transplantation; however, reinfection in the donor organ is universal, recurrent hepatitis C is insidiously progressive, allograft cirrhosis develops in 20–30% at 5 years, and cirrhosis and late organ failure occur at a higher frequency beyond 5 years. With the introduction of highly effective direct acting antiviral agents targeting hepatitis C virus (HCV), allograft outcomes have already improved substantially. In patients with chronic hepatitis B, in the absence of measures to prevent recurrent hepatitis B, survival after transplantation is reduced by ~10–20%; however, prophylactic use of hepatitis B immune globulin (HBIg) during and after transplantation increases the success of transplantation to a level comparable to that seen in patients with nonviral causes of liver decompensation. Specific oral antiviral drugs (e.g., entecavir, tenofovir disoproxil fumarate) **(Chap. 334)** can be used both for prophylaxis against and for treatment of recurrent hepatitis B, facilitating further the management of patients undergoing liver transplantation for end-stage hepatitis B; most transplantation centers rely on antiviral drugs with or without HBIg to manage patients with hepatitis B. Issues of disease recurrence are discussed in more detail below. Patients with nonmetastatic primary hepatobiliary tumors—primary hepatocellular carcinoma (HCC), cholangiocarcinoma, hepatoblastoma, angiosarcoma, epithelioid hemangioendothelioma, and multiple or massive hepatic adenomata—have undergone liver transplantation; however, for some hepatobiliary malignancies, overall survival is significantly lower than that for other categories of liver disease. Most transplantation centers have reported 5-year recurrence-free survival rates in patients with unresectable HCC for single tumors <5 cm in diameter or for three or fewer lesions all <3 cm comparable to those seen in patients undergoing transplantation for nonmalignant indications. Consequently, liver transplantation is

currently restricted to patients whose hepatic malignancies meet these criteria. Expanded criteria for patients with HCC continue to be evaluated. Because the likelihood of recurrent cholangiocarcinoma is very high, only highly selected patients with limited disease are being evaluated for transplantation after intensive chemotherapy and radiation.

CONTRAINDICATIONS

Absolute contraindications for transplantation include life-threatening systemic diseases, uncontrolled extrahepatic bacterial or fungal infections, preexisting advanced cardiovascular or pulmonary disease, multiple uncorrectable life-threatening congenital anomalies, metastatic malignancy, and active drug or alcohol abuse **(Table 338–2)**. Because carefully selected patients in their sixties and even seventies have undergone transplantation successfully, advanced age per se is no longer considered an absolute contraindication; however, in older patients a more thorough preoperative evaluation should be undertaken to exclude ischemic cardiac disease and other comorbid conditions. Advanced age (>70 years), however, should be considered a *relative contraindication*—that is, a factor to be taken into account with other relative contraindications. Other relative contraindications include portal vein thrombosis, preexisting renal disease not associated with liver disease (which may prompt consideration of combined liver and kidney transplantation), intrahepatic or biliary sepsis, severe hypoxemia (Po_2 <50 mmHg) resulting from right-to-left intrapulmonary shunts, portopulmonary hypertension with high mean pulmonary artery pressures (>35 mmHg), previous extensive hepatobiliary surgery, any uncontrolled serious psychiatric disorder, and lack of sufficient social supports. Any one of these relative contraindications is insufficient in and of itself to preclude transplantation. For example, the problem of portal vein thrombosis can be overcome by constructing a graft from the donor liver portal vein to the recipient's superior mesenteric vein. Now that combination antiretroviral therapy has dramatically improved the survival of persons with HIV infection **(Chap. 197)**, and because end-stage liver disease caused by chronic hepatitis C and B has emerged as a serious source of morbidity and mortality in the HIV-infected population, liver transplantation has now been performed successfully in selected HIV-positive persons who have excellent control of HIV infection. Selected patients with CD4± T cell counts >100/μL and with pharmacologic suppression of HIV viremia have undergone transplantation for end-stage liver disease. HIV-infected persons who have received liver allografts for end-stage liver disease resulting from chronic hepatitis B have experienced survival rates compared to those of HIV-negative persons undergoing transplantation for the same indication. In contrast, recurrent HCV in the allograft has until recently limited long-term success in persons with HCV-related end-stage liver disease. Again, it is expected that the availability of direct-acting antiviral (DAA) agents targeting HCV (see below and **Chap 334**) will significantly improve allograft outcomes.

TABLE 338-2

Contraindications to Liver Transplantation

Absolute	Relative
Uncontrolled extrahepatobiliary infection	Age >70
Active, untreated sepsis	Prior extensive hepatobiliary surgery
Uncorrectable, life-limiting congenital anomalies	Portal vein thrombosis
Active substance or alcohol abuse	Renal failure not attributable to liver disease
Advanced cardiopulmonary disease	Previous extrahepatic malignancy (not including nonmelanoma skin cancer)
Extrahepatobiliary malignancy (not including nonmelanoma malignancy skin cancer)	Severe obesity
Metastatic malignancy to the liver	Severe malnutrition/wasting
Cholangiocarcinoma	Medical noncompliance
AIDS	HIV seropositivity with failure to control HIV viremia or CD4 <100/μL
Life-threatening systemic diseases	Intrahepatic sepsis
	Severe hypoxemia secondary to right-to-left intrapulmonary shunts (Po_2 <50 mmHg)
	Severe pulmonary hypertension (mean pulmonary artery pressure >35 mmHg)
	Uncontrolled psychiatric disorder

TECHNICAL CONSIDERATIONS

CADAVER DONOR SELECTION

Cadaver donor livers for transplantation are procured primarily from victims of head trauma. Organs from brain-dead donors up to age 60 are acceptable if the following criteria are met: hemodynamic stability, adequate oxygenation, absence of bacterial or fungal infection, absence of abdominal trauma, absence of hepatic dysfunction, and serologic exclusion of hepatitis B (HBV) and C viruses and HIV. Occasionally, organs from donors with hepatitis B and C are used (e.g., for recipients with prior hepatitis B and C, respectively). Organs from donors with antibodies to hepatitis B core antigen (anti-HBc) can also be used when the need is especially urgent, and recipients of these organs are treated prophylactically with antiviral drugs. Cardiovascular and respiratory functions are maintained artificially until the liver can be removed. Transplantation of organs procured from deceased donors who have succumbed to cardiac death can be performed successfully under selected circumstances, when ischemic time is minimized and liver histology preserved. Compatibility in ABO blood group and organ size between donor and recipient are important considerations in donor selection; however, ABO-incompatible, split liver, or reduced-donor-organ transplants can be performed in emergencies or marked donor scarcity. Tissue typing for human leukocyte antigen (HLA) matching is not required, and preformed cytotoxic HLA antibodies do not preclude liver transplantation. Following perfusion with cold electrolyte solution, the donor liver is removed and packed in ice. The use of University of Wisconsin (UW) solution, rich in lactobionate and raffinose, has permitted the extension of cold ischemic time up to 20 h; however, 12 h may be a more reasonable limit. Improved techniques for harvesting multiple organs from the same donor have increased the availability of donor livers, but the availability of donor livers is far outstripped by the demand. Currently in the United States, all donor livers are distributed through a nationwide organ-sharing network (United Network for Organ Sharing [UNOS]) designed to allocate available organs based on regional considerations and recipient acuity. Recipients who have the highest disease severity generally have the highest priority, but allocation strategies that balance highest urgency against best outcomes continue to evolve to distribute cadaver organs most effectively. Allocation based on the Child-Turcotte-Pugh (CTP) score, which uses five clinical variables (encephalopathy stage, ascites, bilirubin, albumin, and prothrombin time) and waiting time, has been replaced by allocation based on urgency alone, calculated by the Model for End-Stage Liver Disease (MELD) score. The MELD score is based on a mathematical model that includes bilirubin, creatinine, and prothrombin time expressed as international normalized ratio (INR) **(Table 338-3)**. Neither waiting time (except as a tie breaker between two potential recipients with the same MELD scores) nor posttransplantation outcome is taken into account, but use of the MELD score has been shown to reduce waiting list mortality, to reduce waiting time prior to transplantation, to be the best predictor of pretransplantation mortality, to satisfy the prevailing view that medical need should be the decisive determinant, and to eliminate both the subjectivity inherent in the CTP scoring system (presence and degree of ascites and hepatic encephalopathy) and the differences in waiting times among different regions of the country. Recent data indicate that liver recipients with MELD scores <15 experienced higher posttransplantation mortality rates than similarly classified patients who remained on the wait list. This observation led to the modification of UNOS policy to allocate donor organs to candidates with MELD scores exceeding 15 within the local or regional procurement organization before offering the organ to local patients whose scores are <15. In 2016, the MELD score was modified to incorporate serum sodium, another important predictor of survival in liver transplantation candidates (the MELD-Na score).

TABLE 338-3

United Network for Organ Sharing (UNOS) Liver Transplantation Waiting List Criteria

Status 1	Fulminant hepatic failure (including primary graft nonfunction and hepatic artery thrombosis within 7 days after transplantation as well as acute decompensated Wilson's disease)[a]
The Model for End-Stage Liver Disease (MELD)-Na score, on a continuous scale,[b] determines allocation of the remainder of donor organs. This model is based on the following calculation:	
MELD = 3.78 × $\log_e$ bilirubin (mg/100 mL) ± 11.2 × $\log_e$ international normalized ratio (INR) ± 9.57 × $\log_e$ creatinine (mg/100 mL) + 6.43.[c,d,e] MELD-Na = MELD + 1.59 * (135 – Na [mEq/L])	
Online calculators to determine MELD scores are available, such as the following: *https://optn.transplant.hrsa.gov/resources/allocation-calculators/meld-calculator/*	

[a]For children <18 years of age, status 1 includes acute or chronic liver failure plus hospitalization in an intensive care unit or inborn errors of metabolism. Status 1 is retained for those persons with fulminant hepatic failure and supersedes the MELD score. [b]The MELD scale is continuous, with 34 levels ranging between 6 and 40 (scores above 40 are categorized as 40). Donor organs usually do not become available unless the MELD score exceeds 20. [c]Patients with stage T2 hepatocellular carcinoma receive 22 disease-specific points. [d]Creatinine is included because renal function is a validated predictor of survival in patients with liver disease. For adults undergoing dialysis twice a week, the creatinine in the equation is set to 4 mg/100 mL. [e]For children <18 years of age, the Pediatric End-Stage Liver Disease (PELD) scale is used. This scale is based on albumin, bilirubin, INR, growth failure, and age. Status 1 is retained.

The highest priority (status 1) continues to be reserved for patients with fulminant hepatic failure or primary graft nonfunction. Because candidates for liver transplantation who have HCC may not be sufficiently decompensated to compete for donor organs based on urgency criteria alone, and because protracted waiting for cadaver donor organs often results in tumor growth beyond acceptable limits for transplantation, such patients are assigned disease-specific MELD points (Table 338–3). Other disease-specific MELD exceptions include portopulmonary hypertension, hepatopulmonary syndrome, familial amyloid polyneuropathy, primary hyperoxaluria (necessitating liver-kidney transplantation), cystic fibrosis liver disease, and highly selected cases of hilar cholangiocarcinoma.

LIVING DONOR TRANSPLANTATION

Occasionally, especially for liver transplantation in children, one cadaver organ can be split between two recipients (one adult and one child). A more viable alternative, transplantation of the right lobe of the liver from a healthy adult donor into an adult recipient, has gained increased popularity. Living donor

transplantation of the left lobe (left lateral segment), introduced in the early 1990s to alleviate the extreme shortage of donor organs for small children, accounts currently for approximately one-third of all liver transplantation procedures in children. Driven by the shortage of cadaver organs, living donor transplantation involving the more sizable right lobe is being considered with increasing frequency in adults; however, living donor liver transplantation cannot be expected to solve the donor organ shortage; 367 such procedures were done in 2017, representing only about 5% of all liver transplant operations done in the United States.

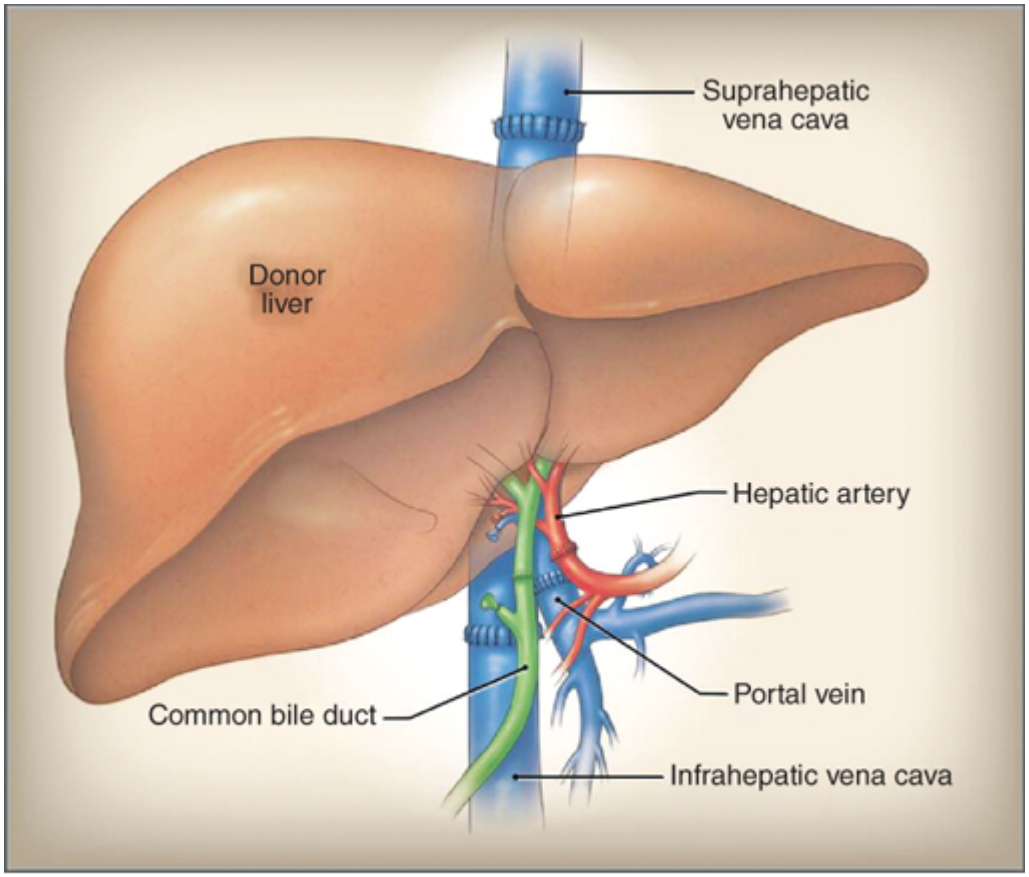
Living donor transplantation can reduce waiting time and cold-ischemia time; it is done under elective, rather than emergency, circumstances; and it may be lifesaving in recipients who cannot afford to wait for a cadaver donor. The downside, of course, is the risk to the healthy donor (a mean of 10 weeks of medical disability; biliary complications in ~5%; postoperative complications such as wound infection, small-bowel obstruction, and incisional hernias in 9–19%; and even, in 0.2–0.4%, death) as well as the increased frequency of biliary (15–32%) and vascular (10%) complications in the recipient. Potential donors must participate voluntarily without coercion, and transplantation teams should go to great lengths to exclude subtle coercive or inappropriate psychological factors as well as outline carefully to both donor and recipient the potential benefits and risks of the procedure. Donors for the procedure should be 18–60 years old; have a compatible blood type with the recipient; have no chronic medical problems or history of major abdominal surgery; should be related genetically or emotionally to the recipient; and pass an exhaustive series of clinical, biochemical, and serologic evaluations to unearth disqualifying medical disorders. The recipient should meet the same UNOS criteria for liver transplantation as recipients of a cadaver donor allograft. Comprehensive outcome data on adult-to-adult living donor liver transplantation are being collected (*www.nih-a2all.org*).

SURGICAL TECHNIQUE

Removal of the recipient's native liver is technically difficult, particularly in the presence of portal hypertension with its associated collateral circulation and extensive varices and especially in the presence of scarring from previous abdominal operations. The combination of portal hypertension and coagulopathy (elevated prothrombin time and thrombocytopenia) may translate into large blood product transfusion requirements. After the portal vein and infrahepatic and suprahepatic inferior vena cavae are dissected, the hepatic artery and common bile duct are dissected. Then the native liver is removed and the donor organ inserted. During the anhepatic phase, coagulopathy, hypoglycemia, hypocalcemia, and hypothermia are encountered and must be managed by the anesthesiology team. Caval, portal vein, hepatic artery, and bile duct anastomoses are performed in succession, the last by end-to-end suturing of the donor and recipient common bile ducts **(Fig. 338-1)** or by choledochojejunostomy to a Roux-en-Y loop if the recipient common bile duct cannot be used for reconstruction (e.g., in sclerosing cholangitis). A typical transplant operation lasts 8 h, with a range of 6–18 h. Because of excessive bleeding, large volumes of blood, blood products, and volume expanders may be required during surgery; however, blood requirements have fallen sharply with improvements in surgical technique, blood-salvage interventions, and experience.

FIGURE 338-1

The anastomoses in orthotopic liver transplantation. The anastomoses are performed in the following sequence: (1) suprahepatic and infrahepatic vena cava, (2) portal vein, (3) hepatic artery, and (4) common bile duct-to-duct anastomosis. *(Adapted from JL Dienstag, AB Cosimi: N Engl J Med 367:1483, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

As noted above, emerging alternatives to orthotopic liver transplantation include split-liver grafts, in which one donor organ is divided and inserted into two recipients; and living donor procedures, in which part of the left (for children), the left (for children or small adults), or the right (for adults) lobe of the liver is harvested from a living donor for transplantation into the recipient. In the adult procedure, once the right lobe is removed from the donor, the donor right hepatic vein is anastomosed to the recipient right hepatic vein remnant, followed by donor-to-recipient anastomoses of the portal vein and then the hepatic artery. Finally, the biliary anastomosis is performed, duct-to-duct if practical or via Roux-en-Y anastomosis. Heterotopic liver transplantation, in which the donor liver is inserted without removal of the native liver, has met with very limited success and acceptance, except in a very small number of centers. In attempts to support desperately ill patients until a suitable donor organ can be identified, several transplantation centers are studying extracorporeal perfusion with bioartificial liver cartridges constructed from hepatocytes bound to hollow fiber systems and used as temporary hepatic-assist devices, but their efficacy remains to be established. Areas of research with the potential to overcome the shortage of donor organs include hepatocyte transplantation and xenotransplantation with genetically modified organs of nonhuman origin (e.g., swine).

POSTOPERATIVE COURSE AND MANAGEMENT

IMMUNOSUPPRESSIVE THERAPY

The introduction in 1980 of cyclosporine as an immunosuppressive agent contributed substantially to the improvement in survival after liver transplantation. Cyclosporine, a calcineurin inhibitor, blocks early activation of T cells and is specific for T cell functions that result from the interaction of the T cell with its receptor and that involve the calcium-dependent signal transduction pathway. As a result, the activity of cyclosporine leads to inhibition of lymphokine gene activation, blocking interleukins 2, 3, and 4, tumor necrosis factor a, and other lymphokines. Cyclosporine also inhibits B cell functions. This process occurs without affecting rapidly dividing cells in the bone marrow, which may account for the reduced frequency of posttransplantation systemic infections. The most common and important side effect of cyclosporine therapy is nephrotoxicity. Cyclosporine causes dose-dependent renal tubular injury and direct renal artery vasospasm. Following renal function is therefore important in monitoring cyclosporine therapy, perhaps even a more reliable indicator than blood levels of the drug. Nephrotoxicity is reversible and can be managed by dose reduction. Other adverse effects of cyclosporine therapy include hypertension, hyperkalemia, tremor, hirsutism, glucose intolerance, and gingival hyperplasia.

Tacrolimus, a macrolide lactone antibiotic isolated from a Japanese soil fungus, *Streptomyces tsukubaensis*, has the same mechanism of action as cyclosporine but is 10–100 times more potent. Initially applied as "rescue" therapy for patients in whom rejection occurred despite the use of cyclosporine, tacrolimus was shown to be associated with a reduced frequency of acute, refractory, and chronic rejection. Although patient and graft survival are the same with these two drugs, the advantage of tacrolimus in minimizing episodes of rejection, reducing the need for additional glucocorticoid doses, and reducing the likelihood of bacterial and cytomegalovirus (CMV) infection has simplified the management of patients undergoing liver transplantation. In addition, the oral absorption of tacrolimus is more predictable than that of cyclosporine, especially during the early postoperative period when T-tube drainage interferes with the enterohepatic circulation of cyclosporine. As a result, in most transplantation centers, tacrolimus has now supplanted cyclosporine for primary immunosuppression, and many centers rely on oral rather than IV administration from the outset. For transplantation centers that prefer cyclosporine, a better-absorbed microemulsion preparation is available.

Although more potent than cyclosporine, tacrolimus is also more toxic and more likely to be discontinued for adverse events. The toxicity of tacrolimus is similar to that of cyclosporine; nephrotoxicity and neurotoxicity are the most commonly encountered adverse effects, and neurotoxicity (tremor, seizures, hallucinations, psychoses, coma) is more likely and more severe in tacrolimus-treated patients. Both drugs can cause diabetes mellitus, but tacrolimus does not cause hirsutism or gingival hyperplasia. Because of overlapping toxicity between cyclosporine and tacrolimus, especially nephrotoxicity, and because tacrolimus reduces cyclosporine clearance, these two drugs should not be used together. Because 99% of tacrolimus is metabolized by the liver, hepatic dysfunction reduces its clearance; in primary graft nonfunction (when, for technical reasons or because of ischemic damage prior to its insertion, the allograft is defective and does not function normally from the outset), tacrolimus doses have to be reduced substantially, especially in children. Both cyclosporine and tacrolimus are metabolized by the cytochrome P450 IIIA system, and, therefore, drugs

that induce cytochrome P450 (e.g., phenytoin, phenobarbital, carbamazepine, rifampin) reduce available levels of cyclosporine and tacrolimus; and drugs that inhibit cytochrome P450 (e.g., erythromycin, fluconazole, ketoconazole, clotrimazole, itraconazole, verapamil, diltiazem, danazol, metoclopramide, the HIV protease inhibitor ritonavir, and the HCV protease inhibitor paritaprevir) increase cyclosporine and tacrolimus blood levels. Indeed, itraconazole is used occasionally to help boost tacrolimus levels. Like azathioprine, cyclosporine and tacrolimus appear to be associated with a risk of lymphoproliferative malignancies (see below), which may occur earlier after cyclosporine or tacrolimus than after azathioprine therapy. Because of these side effects, combinations of cyclosporine or tacrolimus with prednisone and an antimetabolite (azathioprine or mycophenolic acid, see below)—all at reduced doses—are preferable regimens for immunosuppressive therapy.

Mycophenolic acid, a nonnucleoside purine metabolism inhibitor derived as a fermentation product from several *Penicillium* species, is another immunosuppressive drug being used for patients undergoing liver transplantation. Mycophenolate has been shown to be better than azathioprine, when used with other standard immunosuppressive drugs, in preventing rejection after renal transplantation and has been adopted widely as well for use in liver transplantation. The most common adverse effects of mycophenolate are bone marrow suppression and gastrointestinal complaints.

In patients with pretransplantation renal dysfunction or renal deterioration that occurs intraoperatively or immediately postoperatively, tacrolimus or cyclosporine therapy may not be practical; under these circumstances, induction or maintenance of immunosuppression with antithymocyte globulin (ATG, thymoglobulin) or monoclonal antibodies to T cells, OKT3, may be appropriate. Therapy with these agents has been especially effective in reversing acute rejection in the posttransplantation period and is the standard treatment for acute rejection that fails to respond to methylprednisolone boluses. Available data support the use of thymoglobulin induction to delay calcineurin inhibitor use and its attendant nephrotoxicity. IV infusions of thymoglobulin may be complicated by fever and chills, which can be ameliorated by premedication with antipyretics and a low dose of glucocorticoids. Infusions of OKT3 may be complicated by fever, chills, and diarrhea, or by pulmonary edema, which can be fatal. Because OKT3 is such a potent immunosuppressive agent, its use is also more likely to be complicated by opportunistic infection or lymphoproliferative disorders; therefore, because of the availability of alternative immunosuppressive drugs, OKT3 is now used sparingly.

Sirolimus, an inhibitor of the mammalian target of rapamycin (mTOR), blocks later events in T cell activation, is approved for use in kidney transplantation, but is not formally approved for use in liver transplant recipients because of the reported association with an increased frequency of hepatic artery thrombosis in the first month posttransplantation. In patients with calcineurin inhibitor–related nephrotoxicity, conversion to sirolimus has been demonstrated to be effective in preventing rejection with accompanying improvements in renal function. Because of its profound antiproliferative effects, sirolimus has also been suggested to be a useful immunosuppressive agent in patients with a prior or current history of malignancy, such as HCC. Side effects include hyperlipidemia, peripheral edema, oral ulcers, and interstitial pneumonitis. Everolimus is a hydroxyethyl derivative of sirolimus that, when used in conjunction with low-dose

tacrolimus, also provides successful protection against acute rejection, with decreased renal impairment compared to that associated with standard tacrolimus dosing. Everolimus and sirolimus share a similar adverse events profile. The most important principle of immunosuppression is that the ideal approach strikes a balance between immunosuppression and immunologic competence. In general, given sufficient immunosuppression, acute liver allograft rejection is nearly always reversible. On one hand, incompletely treated acute rejection predisposes to the development of chronic rejection, which can threaten graft survival. On the other hand, if the cumulative dose of immunosuppressive therapy is too large, the patient may succumb to opportunistic infection. In hepatitis C, pulse glucocorticoids or OKT3 use accelerate recurrent allograft hepatitis, although the routine use of DAA therapy to clear the allograft of HCV should remove or greatly diminish this concern. Further complicating matters, acute rejection can be difficult to distinguish histologically from recurrent hepatitis C. Therefore, immunosuppressive drugs must be used judiciously, with strict attention to the infectious consequences of such therapy and careful confirmation of the diagnosis of acute rejection. In this vein, efforts have been made to minimize the use of glucocorticoids, a mainstay of immunosuppressive regimens, and steroid-free immunosuppression can be achieved in some instances. Patients who undergo liver transplantation for autoimmune diseases such as primary biliary cirrhosis, autoimmune hepatitis, and primary sclerosing cholangitis are less likely to achieve freedom from glucocorticoids.

POSTOPERATIVE COMPLICATIONS

Complications of liver transplantation can be divided into nonhepatic and hepatic categories (**Tables 338-4 and 338-5**). In addition, both immediate postoperative and late complications are encountered. As a rule, patients who undergo liver transplantation have been chronically ill for protracted periods and may be malnourished and wasted. The impact of such chronic illness and the multisystem failure that accompanies liver failure continue to require attention in the postoperative period. Because of the massive fluid losses and fluid shifts that occur during the operation, patients may remain fluid-overloaded during the immediate postoperative period, straining cardiovascular reserve; this effect can be amplified in the face of transient renal dysfunction and pulmonary capillary vascular permeability. Continuous monitoring of cardiovascular and pulmonary function, measures to maintain the integrity of the intravascular compartment and to treat extravascular volume overload, and scrupulous attention to potential sources and sites of infection are of paramount importance. Cardiovascular instability may also result from the electrolyte imbalance that may accompany reperfusion of the donor liver as well as from restoration of systemic vascular resistance following implantation. Pulmonary function may be compromised further by paralysis of the right hemidiaphragm associated with phrenic nerve injury. The hyperdynamic state with increased cardiac output that is characteristic of patients with liver failure reverses rapidly after successful liver transplantation.

TABLE 338-4

Nonhepatic Complications of Liver Transplantation

Category	Complication
Cardiovascular instability	Arrhythmias
	Congestive heart failure
	Cardiomyopathy
Pulmonary compromise	Pneumonia
	Pulmonary capillary vascular permeability
	Fluid overload
Renal dysfunction	Prerenal azotemia
	Hypoperfusion injury (acute tubular necrosis)
	Drug nephrotoxicity
	↓ Renal blood flow secondary to ↑ intraabdominal pressure
Hematologic	Anemia secondary to gastrointestinal and/or intraabdominal bleeding
	Hemolytic anemia, aplastic anemia
	Thrombocytopenia
Infection	Bacterial: early, common postoperative infections
	Fungal/parasitic: late, opportunistic infections
	Viral: late, opportunistic infections, recurrent hepatitis
Neuropsychiatric	Seizures
	Metabolic encephalopathy
	Depression

Category	Complication
	Difficult psychosocial adjustment
Diseases of donor	Infectious
	Malignant
Malignancy	B cell lymphoma (posttransplantation lymphoproliferative disorders)
	De novo neoplasms (particularly squamous cell skin carcinoma)

TABLE 338-5

Hepatic Complications of Liver Transplantation

Hepatic Dysfunction Common after Major Surgery	
Prehepatic	Pigment load
	Hemolysis
	Blood collections (hematomas, abdominal collections)
Intrahepatic	
Early	Hepatotoxic drugs and anesthesia
	Hypoperfusion (hypotension, shock, sepsis)
	Benign postoperative cholestasis
Late	Transfusion-associated hepatitis
	Exacerbation of primary hepatic disease
Posthepatic	Biliary obstruction
	↓ Renal clearance of conjugated bilirubin (renal dysfunction)
Hepatic Dysfunction Unique to Liver Transplantation	
Primary graft nonfunction	
Vascular compromise	Portal vein obstruction
	Hepatic artery thrombosis
	Anastomotic leak with intraabdominal bleeding
Bile duct disorder	Stenosis, obstruction, leak
Rejection	
Recurrent primary hepatic disease	

Other immediate management issues include renal dysfunction. Prerenal azotemia, acute kidney injury associated with hypoperfusion (acute tubular necrosis), and renal toxicity caused by antibiotics, tacrolimus, or cyclosporine are encountered frequently in the postoperative period, sometimes necessitating dialysis. Hemolytic-uremic syndrome can be associated with cyclosporine, tacrolimus, or OKT3. Occasionally, postoperative intraperitoneal bleeding may be sufficient to increase intraabdominal pressure, which, in turn, may reduce renal blood flow; this effect is rapidly reversible when abdominal distention is relieved by exploratory laparotomy to identify and ligate the bleeding site and to remove intraperitoneal clot.

Anemia may also result from acute upper gastrointestinal bleeding or from transient hemolytic anemia, which may be autoimmune, especially when blood group O livers are transplanted into blood group A or B recipients. This autoimmune hemolytic anemia is mediated by donor intrahepatic lymphocytes that recognize red blood cell A or B antigens on recipient erythrocytes. Transient in nature, this process resolves once the donor liver is repopulated by recipient bone marrow–derived lymphocytes; the hemolysis can be treated by transfusing blood group O red blood cells and/or by administering higher doses of glucocorticoids. Transient thrombocytopenia is also commonly encountered. Aplastic anemia, a late occurrence, is rare but has been reported in almost 30% of patients who underwent liver transplantation for acute, severe hepatitis of unknown cause.

Bacterial, fungal, or viral infections are common and may be life-threatening postoperatively. Early after transplant surgery, common postoperative infections predominate—pneumonia, wound infections, infected intraabdominal collections, urinary tract infections, and IV line infections—rather than opportunistic infections; these infections may involve the biliary tree and liver as well. Beyond the first postoperative month, the toll of immunosuppression becomes evident, and opportunistic infections—CMV, herpes viruses, fungal infections (*Aspergillus*, *Candida*, cryptococcal disease), mycobacterial infections, parasitic infections (*Pneumocystis*, *Toxoplasma*), bacterial infections (*Nocardia*, *Legionella*, *Listeria*)—predominate. Rarely, early infections represent those transmitted with the donor liver, either infections present in the donor or infections acquired during procurement processing. De novo viral hepatitis infections acquired from the donor organ or, almost unheard of now, from transfused blood products occur after typical incubation periods for these agents (well beyond the first month). Obviously, infections in an immunosuppressed host demand early recognition and prompt management; prophylactic antibiotic therapy is administered routinely in the immediate postoperative period. Use of sulfamethoxazole with trimethoprim reduces the incidence of postoperative *Pneumocystis carinii* pneumonia. Antiviral prophylaxis for CMV with ganciclovir should be administered in patients at high risk (e.g., when a CMV-seropositive donor organ is implanted into a CMV-seronegative recipient).

Neuropsychiatric complications include seizures (commonly associated with cyclosporine and tacrolimus toxicity), metabolic encephalopathy, depression, and difficult psychosocial adjustment. Rarely, diseases are transmitted by the allograft from the donor to the recipient. In addition to viral and bacterial infections, malignancies of donor origin have occurred. Posttransplantation lymphoproliferative disorders, especially B cell lymphoma, are a recognized complication associated with immunosuppressive drugs such as

azathioprine, tacrolimus, and cyclosporine (see above). Epstein-Barr virus has been shown to play a contributory role in some of these tumors, which may regress when immunosuppressive therapy is reduced. De novo neoplasms appear at increased frequency after liver transplantation, particularly squamous cell carcinomas of the skin. Routine screening should be performed.

Long-term complications after liver transplantation attributable primarily to immunosuppressive medications include diabetes mellitus and osteoporosis (associated with glucocorticoids and calcineurin inhibitors) as well as hypertension, hyperlipidemia, and chronic renal insufficiency (associated with cyclosporine and tacrolimus). Monitoring and treating these disorders are routine components of posttransplantation care; in some cases, they respond to changes in immunosuppressive regimen, while in others, specific treatment of the disorder is introduced. Data from a large U.S. database showed that the prevalence of renal failure was 18% at year 5 and 25% at year 10 after liver transplantation. Similarly, the high frequency of diabetes, hypertension, hyperlipidemia, obesity, and the metabolic syndrome renders patients susceptible to cardiovascular disease after liver transplantation; although hepatic complications account for most of the mortality after liver transplantation, renal failure and cardiovascular disease are the other leading causes of late mortality after liver transplantation.

HEPATIC COMPLICATIONS

Hepatic dysfunction after liver transplantation is similar to the hepatic complications encountered after major abdominal and cardiothoracic surgery; however, in addition, hepatic complications include primary graft failure, vascular compromise, failure or stricture of the biliary anastomoses, and rejection. As in nontransplantation surgery, postoperative jaundice may result from prehepatic, intrahepatic, and posthepatic sources. *Prehepatic* sources represent the massive hemoglobin pigment load from transfusions, hemolysis, hematomas, ecchymoses, and other collections of blood. *Early intrahepatic* liver injury includes effects of hepatotoxic drugs and anesthesia; hypoperfusion injury associated with hypotension, sepsis, and shock; and benign postoperative cholestasis. *Late intrahepatic* sources of liver injury include exacerbation of primary disease. *Posthepatic* sources of hepatic dysfunction include biliary obstruction and reduced renal clearance of conjugated bilirubin. Hepatic complications unique to liver transplantation include primary graft failure associated with ischemic injury to the organ during harvesting; vascular compromise associated with thrombosis or stenosis of the portal vein or hepatic artery anastomoses; vascular anastomotic leak; stenosis, obstruction, or leakage of the anastomosed common bile duct; recurrence of primary hepatic disorder (see below); and rejection.

TRANSPLANT REJECTION

Despite the use of immunosuppressive drugs, rejection of the transplanted liver still occurs in a proportion of patients, beginning 1–2 weeks after surgery. Clinical signs suggesting rejection are fever, right upper quadrant pain, and reduced bile pigment and volume. Leukocytosis may occur, but the most reliable indicators are increases in serum bilirubin and aminotransferase levels. Because these tests lack specificity, distinguishing among rejection, biliary obstruction, primary graft nonfunction, vascular compromise, viral

hepatitis, CMV infection, drug hepatotoxicity, and recurrent primary disease may be difficult. Radiographic visualization of the biliary tree and/or percutaneous liver biopsy often help to establish the correct diagnosis. Morphologic features of acute rejection include a mixed portal cellular infiltrate, bile duct injury, and/or endothelial inflammation ("endothelialitis"); some of these findings are reminiscent of graft-versus-host disease, primary biliary cirrhosis, or recurrent allograft hepatitis C. As soon as transplant rejection is suspected, treatment consists of IV methylprednisolone in repeated boluses; if this fails to abort rejection, many centers use thymoglobulin or OKT3. Caution should be exercised when managing acute rejection with pulse glucocorticoids or OKT3 in patients with HCV infection, because of the high risk of triggering recurrent allograft hepatitis C. The availability of DAAs for HCV should greatly alleviate this concern.

Chronic rejection is a relatively rare outcome that can follow repeated bouts of acute rejection or that occurs unrelated to preceding rejection episodes. Morphologically, chronic rejection is characterized by progressive cholestasis, focal parenchymal necrosis, mononuclear infiltration, vascular lesions (intimal fibrosis, subintimal foam cells, fibrinoid necrosis), and fibrosis. This process may be reflected as ductopenia—the vanishing bile duct syndrome, which is more common in patients undergoing liver transplantation for autoimmune liver disease. Reversibility of chronic rejection is limited; in patients with therapy-resistant chronic rejection, retransplantation has yielded encouraging results.

OUTCOME

SURVIVAL

The survival rate for patients undergoing liver transplantation has improved steadily since 1983. One-year survival rates have increased from ~70% in the early 1980s to 85–90% from 2003 to the present time. Currently, the 5-year survival rate exceeds 60%. An important observation is the relationship between clinical status before transplantation and outcome. For patients who undergo liver transplantation when their level of compensation is high (e.g., still working or only partially disabled), a 1-year survival rate of >85% is common. For those whose level of decompensation mandates continuous in-hospital care prior to transplantation, the 1-year survival rate is ~70%, whereas for those who are so decompensated that they require life support in an intensive care unit, the 1-year survival rate is ~50%. Since the adoption by UNOS in 2002 of the MELD system for organ allocation, posttransplantation survival has been found to be affected adversely for candidates with MELD scores >25, considered high disease severity. Thus, irrespective of allocation scheme, high disease severity before transplantation corresponds to diminished posttransplantation survival. Another important distinction in survival has been drawn between high- and low-risk patient categories. For patients who do not fit any "high-risk" designations, 1-year and 5-year survival rates of 85 and 80%, respectively, have been recorded. In contrast, among patients in high-risk categories—cancer, fulminant hepatitis, age >65, concurrent renal failure, respirator dependence, portal vein thrombosis, and history of a portacaval shunt or multiple right upper quadrant operations—survival statistics fall into the range of 60% at 1 year and 35% at 5 years. Survival after retransplantation for primary graft nonfunction is ~50%. Causes of failure of liver transplantation vary with time. Failures within the first 3 months result primarily from technical complications, postoperative infections, and hemorrhage. Transplant

failures after the first 3 months are more likely to result from infection, rejection, or recurrent disease (such as malignancy or viral hepatitis).

RECURRENCE OF PRIMARY DISEASE

Features of autoimmune hepatitis, primary sclerosing cholangitis, and primary biliary cirrhosis overlap with those of rejection or posttransplantation bile duct injury. Whether autoimmune hepatitis and sclerosing cholangitis recur after liver transplantation is controversial; data supporting recurrent autoimmune hepatitis (in up to one-third of patients in some series) are more convincing than those supporting recurrent sclerosing cholangitis. Similarly, reports of recurrent primary biliary cirrhosis after liver transplantation have appeared; however, the histologic features of primary biliary cirrhosis and chronic rejection are virtually indistinguishable and occur as frequently in patients with primary biliary cirrhosis as in patients undergoing transplantation for other reasons. The presence of a florid inflammatory bile duct lesion is highly suggestive of the recurrence of primary biliary cirrhosis, but even this lesion can be observed in acute rejection. Hereditary disorders such as Wilson's disease and α_1-antitrypsin deficiency have not recurred after liver transplantation; however, recurrence of disordered iron metabolism has been observed in some patients with hemochromatosis. Hepatic vein thrombosis (Budd-Chiari syndrome) may recur; this can be minimized by treating underlying myeloproliferative disorders and by anticoagulation. Because cholangiocarcinoma recurs almost invariably, few centers now offer transplantation to such patients; however, a few highly selected patients with operatively confirmed stage I or II cholangiocarcinoma who undergo liver transplantation combined with neoadjuvant chemoradiation may experience excellent outcomes. In patients with intrahepatic HCC who meet criteria for transplantation, 1- and 5-year survivals are similar to those observed in patients undergoing liver transplantation for nonmalignant disease. Finally, metabolic disorders such as nonalcoholic steatohepatitis recur frequently, especially if the underlying metabolic predisposition is not altered. The metabolic syndrome occurs commonly after liver transplantation as a result of recurrent nonalcoholic fatty liver, immunosuppressive medications, and/or, in patients with hepatitis C related to the impact of HCV infection on insulin resistance, diabetes and fatty liver.

Hepatitis A can recur after transplantation for fulminant hepatitis A, but such acute reinfection has no serious clinical sequelae. In fulminant hepatitis B, recurrence is not the rule; however, in the absence of any prophylactic measures, hepatitis B usually recurs after transplantation for end-stage chronic hepatitis B. Before the introduction of prophylactic antiviral therapy, immunosuppressive therapy sufficient to prevent allograft rejection led inevitably to marked increases in hepatitis B viremia, regardless of pretransplantation levels. Overall graft and patient survival were poor, and some patients experienced a rapid recapitulation of severe injury—severe chronic hepatitis or even fulminant hepatitis—after transplantation. Also recognized in the era before availability of antiviral regimens was *fibrosing cholestatic hepatitis*, rapidly progressive liver injury associated with marked hyperbilirubinemia, substantial prolongation of the prothrombin time (both out of proportion to relatively modest elevations of aminotransferase activity), and rapidly progressive liver failure. This lesion has been suggested to represent a "choking off" of the hepatocyte by an overwhelming density of HBV proteins. Complications such as sepsis and pancreatitis were also observed more frequently in patients undergoing liver transplantation for hepatitis B prior to the introduction of antiviral therapy. The

introduction of long-term prophylaxis with HBIg revolutionized liver transplantation for chronic hepatitis B. Preoperative hepatitis B vaccination, preoperative or postoperative interferon (IFN) therapy, or short-term (≤2 months) HBIg prophylaxis has not been shown to be effective, but a retrospective analysis of data from several hundred European patients followed for 3 years after transplantation has shown that long-term (≥6 months) prophylaxis with HBIg is associated with a lowering of the risk of HBV reinfection from ~75 to 35% and a reduction in mortality from ~50 to 20%.

As a result of long-term HBIg use following liver transplantation for chronic hepatitis B, similar improvements in outcome have been observed in the United States, with 1-year survival rates between 75 and 90%. Currently, with HBIg prophylaxis, the outcome of liver transplantation for chronic hepatitis B is indistinguishable from that for chronic liver disease unassociated with chronic hepatitis B; essentially, medical concerns regarding liver transplantation for chronic hepatitis B have been eliminated. Passive immunoprophylaxis with HBIg is begun during the anhepatic stage of surgery, repeated daily for the first 6 postoperative days, and then continued with infusions that are given either at regular intervals of 4–6 weeks or, alternatively, when anti-hepatitis B surface (HBs) levels fall below a threshold of 100 mIU/mL. The current approach in most centers is to continue HBIg indefinitely, which can add ~$20,000 per year to the cost of care; some centers are evaluating regimens that shift to less frequent administration or to IM administration in the late posttransplantation period or, in low-risk patients, maintenance with antiviral therapy (see below) alone. Still, “breakthrough” HBV infection occasionally occurs.

Further improving the outcome of liver transplantation for chronic hepatitis B is the current availability of such antiviral drugs as entecavir and tenofovir disoproxil fumarate **(Chap. 334)**. When these drugs are administered to patients with decompensated liver disease, a proportion improves sufficiently to postpone imminent liver transplantation. In addition, antiviral therapy can be used to prevent recurrence of HBV infection when administered *prior to* transplantation; to treat hepatitis B that recurs *after* transplantation, including in patients who break through HBIg prophylaxis; and to reverse the course of otherwise fatal fibrosing cholestatic hepatitis. Clinical trials have shown that entecavir or tenofovir antiviral therapy reduces the level of HBV replication substantially, sometimes even resulting in clearance of hepatitis B surface antigen (HBsAg); reduces alanine aminotransferase (ALT) levels; and improves histologic features of necrosis and inflammation. Currently, most liver transplantation centers combine HBIg plus one of the low-resistance oral nucleoside (entecavir) or nucleotide analogues (tenofovir). In low-risk patients with no detectable hepatitis B viremia at the time of transplantation, a number of clinical trials have suggested that antiviral prophylaxis can suffice, without HBIg or with a finite duration of HBIg, to prevent recurrent HBV infection of the allograft. In patients documented at the time of liver transplantation to have undetectable HBV DNA in serum and cccDNA in the liver (i.e., with low risk for recurrence of HBV infection), a preliminary clinical trial suggested that, after receipt of 5 years of combined therapy, both HBIg and oral-agent therapy can be withdrawn sequentially (over two 6-month periods) with a success rate, as monitored over a median of 6 years postwithdrawal, of 90% and an anti-HBs seroconversion rate of 60% (despite transient reappearance of HBV DNA and/or HBsAg in some of these patients).

Antiviral prophylactic approaches applied to patients undergoing liver transplantation for chronic hepatitis B are being used as well for patients without hepatitis B who receive organs from donors with antibody to hepatitis B core antigen (anti-HBc) but do not have HBsAg. Patients who undergo liver transplantation for chronic hepatitis B plus D are less likely to experience recurrent liver injury than patients undergoing liver transplantation for hepatitis B alone; still, such co-infected patients would also be offered standard posttransplantation prophylactic therapy for hepatitis B.

Until recently, the most common indication for liver transplantation was end-stage liver disease resulting from chronic hepatitis C. For patients undergoing liver transplantation for hepatitis C, because of an aggressive natural history of recurrent allograft hepatitis C, graft and patient survival were diminished substantially compared to other indications for transplantation.

The recent approval of several new DAA agents and of IFN-free DAA regimens against HCV has already had a major impact on the management and outcome of both pretransplantation and posttransplantation HCV infection. Such therapeutic approaches (1) permit the clearance of viremia in a substantial proportion of decompensated cirrhotics, thereby preventing recurrent allograft infection and even improving the clinical status of most of these patients, delaying or obviating the need for liver replacement; and (2) achieve sustained virologic responses in a much higher proportion of persons with allograft HCV infection, because of improvements in antiviral treatment efficacy and tolerability. Ideally, patients should be treated prior to liver transplantation. A concern, however, is that eradication of HCV infection will reduce the MELD score and lower the priority for a donor organ in some patients who still require transplantation because of continued hepatic decompensation and profound reduction in quality of life. In addition, elimination of HCV infection prior to transplantation would disqualify such patients from accepting donor livers from persons with HCV infection, contracting the potential donor pool and limiting accessibility to donor organs and timely transplantation. Therefore, consideration should be given to postponing DAA therapy in patients with high-MELD HCV-associated end-stage liver disease until after liver transplantation; however, a distinct threshold at which to treat pretransplantation or posttransplantation has not yet been established. Regardless, the approach to treatment should be individualized thoughtfully for each patient, based on such factors as MELD score, time anticipated prior to availability of a donor organ, relative clinical stability, and co-morbidities.

Recent DAA combinations that have been used successfully against allograft HCV include ledipasvir + sofosbuvir + ribavirin; velpatasvir + sofosbuvir + ribavirin, and grazoprevir/pibrentasvir. (For updated guidelines, see *www.hcvguidelines.org*). In patients with recurrent HCV infection after liver transplantation, each of these regimens has yielded response rates approaching those seen in compensated non-transplant patient populations.

A small number of allograft recipients have historically succumbed to early HCV-associated liver injury, and a syndrome reminiscent of fibrosing cholestatic hepatitis (see above) has been observed rarely. Currently, however, the routine use of DAA regimens early after transplantation, before the onset of these variant presentations, should have a profound impact on the frequency of severe recurrent allograft hepatitis C.

Patients who undergo liver transplantation for end-stage alcoholic cirrhosis are at risk of resorting to drinking again after transplantation, a potential source of recurrent alcoholic liver injury. Currently, alcoholic liver disease is one of the more common indications for liver transplantation, accounting for 20–25% of all liver transplantation procedures, and most transplantation centers screen candidates carefully for predictors of continued abstinence. Recidivism is more likely in patients whose sobriety prior to transplantation was <6 months. For abstinent patients with alcoholic cirrhosis, liver transplantation can be undertaken successfully, with outcomes comparable to those for other categories of patients with chronic liver disease, when coordinated by a team approach that includes substance abuse counseling.

POSTTRANSPLANTATION QUALITY OF LIFE

Full rehabilitation is achieved in the majority of patients who survive the early postoperative months and escape chronic rejection or unmanageable infection. Psychosocial maladjustment interferes with medical compliance in a small number of patients, but most manage to adhere to immunosuppressive regimens, which must be continued indefinitely. In one study, 85% of patients who survived their transplant operations returned to gainful activities. In fact, some women have conceived and carried pregnancies to term after transplantation without demonstrable injury to their infants.

FURTHER READING

AASLD/IDSA HCV Guidance Panel: Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology 62:932, 2015. (Updated regularly, available at *http://www.hcvguidelines.org*). Accessed Mar 28, 2018. [PubMed: 26111063]

European Association for the Study of the Liver: EASL clinical practice guidelines: Liver transplantation. J Hepatol 64:433, 2016. [PubMed: 26597456]

Fung J et al: Outcomes including liver histology after liver transplantation for chronic hepatitis B using oral antiviral therapy alone. Liver Transpl 21:1504, 2015. [PubMed: 26194746]

Gane EJ et al: Combination of lamivudine and adefovir without hepatitis B immune globulin is safe and effective prophylaxis against hepatitis B virus recurrence in hepatitis B surface antigen-positive liver transplant candidates. Liver Transpl 19:268, 2013. [PubMed: 23447403]

Kwo PY et al: An interferon-free antiviral regimen for HCV after liver transplantation. N Engl J Med 371:2375, 2014. [PubMed: 25386767]

Lenci I et al: Complete hepatitis B virus prophylaxis withdrawal in hepatitis B surface antigen-positive liver transplant recipients after longterm minimal immunosuppression. Liver Transpl 22:1205, 2016. [PubMed: 27272189]

Lucey MR et al: Long-term management of the successful adult liver transplant: 2012 practice guideline by the American Association for the Study of Liver Diseases and the American Society of Transplantation. Liver Transpl 19:3, 2013. [PubMed: 23281277]

Manns M et al: Ledipasvir and sofosbuvir plus ribavirin in patients with genotype 1 or 4 hepatitis C virus infection and advanced liver disease: A multicentre, open-label, randomised, phase 2 trial. Lancet Infectious Diseases 16:685, 2016. [PubMed: 26907736]

Martin P et al: Evaluation for liver transplantation in adults: 2013 practice guideline by the American Association for the Study of Liver Diseases and the American Society of Transplantation. Hepatology 59:1145, 2014.

Poordad F et al: Daclatasvir with sofosbuvir and ribavirin for hepatitis C virus infection with advanced cirrhosis or post-liver transplantation recurrence. Hepatology 63:1493, 2016. [PubMed: 26754432]

Chapter 339: Diseases of the Gallbladder and Bile Ducts

Norton J. Greenberger; Gustav Paumgartner

PHYSIOLOGY OF BILE PRODUCTION AND FLOW

BILE SECRETION AND COMPOSITION

Bile formed in the hepatic lobules is secreted into a complex network of canaliculi, small bile ductules, and larger bile ducts that run with lymphatics and branches of the portal vein and hepatic artery in portal tracts situated between hepatic lobules. These interlobular bile ducts coalesce to form larger septal bile ducts that join to form the right and left hepatic ducts, which in turn, unite to form the common hepatic duct. The common hepatic duct is joined by the cystic duct of the gallbladder to form the common bile duct (CBD), which enters the duodenum (often after joining the main pancreatic duct) through the ampulla of Vater.

Hepatic bile is an isotonic fluid with an electrolyte composition resembling blood plasma. The electrolyte composition of gallbladder bile differs from that of hepatic bile because most of the inorganic anions, chloride, and bicarbonate have been removed by reabsorption across the gallbladder epithelium. As a result of water reabsorption, total solute concentration of bile increases from 3–4 g/dL in hepatic bile to 10–15 g/dL in gallbladder bile.

Major solute components of bile by moles percent include bile acids (80%), lecithin and traces of other phospholipids (16%), and unesterified cholesterol (4.0%). In the lithogenic state, the cholesterol value can be as high as 8–10%. Other constituents include conjugated bilirubin; proteins (all immunoglobulins, albumin, metabolites of hormones, and other proteins metabolized in the liver); electrolytes; mucus; and, often, drugs and their metabolites.

The total daily basal secretion of hepatic bile is ~500–600 mL. Many substances taken up or synthesized by the hepatocyte are secreted into the bile canaliculi. The canalicular membrane forms microvilli and is associated with microfilaments of actin, microtubules, and other contractile elements. Prior to their secretion into the bile, many substances are taken up into the hepatocyte, while others, such as phospholipids, a portion of primary bile acids, and some cholesterol, are synthesized de novo in the hepatocyte. Three mechanisms are important in regulating bile flow: (1) active transport of bile acids from hepatocytes into the bile canaliculi, (2) active transport of other organic anions, and (3) cholangiocellular secretion. The last is a secretin-mediated and cyclic AMP–dependent mechanism that results in the secretion of a sodium- and bicarbonate-rich fluid into the bile ducts.

Active vectorial secretion of biliary constituents from the portal blood into the bile canaliculi is driven by a set of polarized transport systems at the basolateral (sinusoidal) and the canalicular apical plasma membrane domains of the hepatocyte. Two sinusoidal bile salt uptake systems have been cloned in humans, the Na^+/taurocholate cotransporter (NTCP, SLC10A1) and the organic anion–transporting proteins (OATPs), which also transport a large variety of non-bile salt organic anions. Several ATP-dependent canalicular transport systems, “export pumps,” (ATP-binding cassette transport proteins, also known as ABC transporters) have been identified, the most important of which are: the bile salt export pump (BSEP, ABCB11); the anionic conjugate export pump (MRP2, ABCC2), which mediates the canalicular excretion of various amphiphilic conjugates formed by phase II conjugation (e.g., bilirubin mono- and diglucuronides and drugs); the multidrug export pump (MDR1, ABCB1) for hydrophobic cationic compounds; and the phospholipid export pump (MDR3, ABCB4). Two hemitransporters ABCG5/G8, functioning as a couple, constitute the canalicular cholesterol and phytosterol transporter. F1C1 (ATP8B1) is an aminophospholipid transferase (“flippase”) essential for maintaining the lipid asymmetry of the canalicular membrane. The canalicular membrane also contains ATP-independent transport systems such as the Cl/HCO_3 anion exchanger isoform 2 (AE2, SLC4A2) for canalicular bicarbonate secretion. For most of these transporters, genetic defects have been identified that are associated with various forms of cholestasis or defects of biliary excretion. F1C1 is defective in progressive familial intrahepatic cholestasis type 1 (PFIC1) and benign recurrent intrahepatic cholestasis type 1 (BRIC1) and results in ablation of all other ATP-dependent transporter functions. BSEP is defective in PFIC2 and BRIC2. Mutations of MRP2 (ABCC2) cause the Dubin-Johnson syndrome, an inherited form of conjugated hyperbilirubinemia (Chap. 331). A defective MDR3 (ABCB4) results in PFIC3. ABCG5/G8, the canalicular half transporters for cholesterol and other neutral sterols, are defective in sitosterolemia. The cystic fibrosis transmembrane regulator (CFTR, ABCC7) located on bile duct epithelial cells but not on canalicular membranes is defective in cystic fibrosis, which is associated with impaired cholangiocellular pH regulation during ductular bile formation and chronic cholestatic liver disease, occasionally resulting in biliary cirrhosis.

THE BILE ACIDS

The primary bile acids, cholic acid and chenodeoxycholic acid (CDCA), are synthesized from cholesterol in the liver, conjugated with glycine or taurine, and secreted into the bile. Secondary bile acids, including deoxycholate and lithocholate, are formed in the colon as bacterial metabolites of the primary bile acids. However, lithocholic acid is much less efficiently absorbed from the colon than deoxycholic acid. Another secondary bile acid, found in low concentration, is ursodeoxycholic acid (UDCA), a stereoisomer of CDCA. In healthy subjects, the ratio of glycine to taurine conjugates in bile is ~3:1.

Bile acids are detergent-like molecules that in aqueous solutions and above a critical concentration of about 2 m*M* form molecular aggregates called *micelles.* Cholesterol alone is sparingly soluble in aqueous environments, and its solubility in bile depends on both the total lipid concentration and the relative molar percentages of bile acids and lecithin. Normal ratios of these constituents favor the formation of solubilizing *mixed micelles*, while abnormal ratios promote the precipitation of cholesterol crystals in bile via an intermediate liquid crystal phase.

In addition to facilitating the biliary excretion of cholesterol, bile acids facilitate the normal intestinal absorption of dietary fats, mainly cholesterol and fat-soluble vitamins, via a micellar transport mechanism **(Chap. 318)**. Bile acids also serve as a major physiologic driving force for hepatic bile flow and aid in water and electrolyte transport in the small bowel and colon.

ENTEROHEPATIC CIRCULATION

Bile acids are efficiently conserved under normal conditions. Unconjugated, and to a lesser degree also conjugated, bile acids are absorbed by *passive diffusion* along the entire gut. Quantitatively much more important for bile salt recirculation, however, is the *active transport* mechanism for conjugated bile acids in the distal ileum **(Chap. 318)**. The reabsorbed bile acids enter the portal bloodstream and are taken up rapidly by hepatocytes, reconjugated, and resecreted into bile (enterohepatic circulation).

The normal bile acid pool size is ~2–4 g. During digestion of a meal, the bile acid pool undergoes at least one or more enterohepatic cycles, depending on the size and composition of the meal. Normally, the bile acid pool circulates ~5–10 times daily. Intestinal reabsorption of the pool is about 95% efficient; therefore, fecal loss of bile acids is in the range of 0.2–0.4 g/d. In the steady state, this fecal loss is compensated by an equal daily synthesis of bile acids by the liver, and, thus, the size of the bile acid pool is maintained. Bile acids in the intestine release fibroblast growth factor 19 (FGF19) into the circulation, which is transported to the liver where it suppresses synthesis of bile acids from cholesterol by inhibiting the rate-limiting enzyme cytochrome P450 7A1 (CYP7A1) and also promotes gallbladder relaxation. While the loss of bile salts in stool is usually matched by increased hepatic synthesis, the maximum rate of synthesis is ~5 g/d, which may be insufficient to replete the bile acid pool size when there is pronounced impairment of intestinal bile salt reabsorption.

The expression of ABC transporters in the enterohepatic circulation and of the rate-limiting enzymes of bile acid and cholesterol synthesis are regulated in a coordinated fashion by nuclear receptors, which are ligand-activated transcription factors. The hepatic BSEP (ABCB11) is upregulated by the farnesoid X receptor (FXR), a bile acid sensor that also represses bile acid synthesis. The expression of the cholesterol transporter, ABCG5/G8, is upregulated by the liver X receptor (LXR), which is an oxysterol sensor.

GALLBLADDER AND SPHINCTERIC FUNCTIONS

In the fasting state, the sphincter of Oddi (SOD) offers a high-pressure zone of resistance to bile flow from the CBD into the duodenum. Its tonic contraction serves to (1) prevent reflux of duodenal contents into the pancreatic and bile ducts and (2) promote filling of the gallbladder. The major factor controlling the evacuation of the gallbladder is the peptide hormone cholecystokinin (CCK), which is released from the duodenal mucosa in response to the ingestion of fats and amino acids. CCK produces (1) powerful contraction of the gallbladder, (2) decreased resistance of the SOD, and (3) enhanced flow of biliary contents into the duodenum.

Hepatic bile is “concentrated” within the gallbladder by energy-dependent transmucosal absorption of water and electrolytes. Almost the entire bile acid pool may be sequestered in the gallbladder following an overnight fast for delivery into the duodenum with the first meal of the day. The normal capacity of the gallbladder is ~30 mL of bile.

DISEASES OF THE GALLBLADDER

CONGENITAL ANOMALIES

Anomalies of the biliary tract are not uncommon and include abnormalities in number, size, and shape (e.g., agenesis of the gallbladder, duplications, rudimentary or oversized “giant” gallbladders, and diverticula). *Phrygian cap* is a clinically innocuous entity in which a partial or complete septum (or fold) separates the fundus from the body. Anomalies of position or suspension are not uncommon and include left-sided gallbladder, intrahepatic gallbladder, retrodisplacement of the gallbladder, and “floating” gallbladder. The latter condition predisposes to acute torsion, volvulus, or herniation of the gallbladder.

GALLSTONES

Epidemiology and Pathogenesis

Gallstones are quite prevalent in most Western countries. Gallstone formation increases after age 50. In the United States, the third National Health and Nutrition Examination Survey (NHANES III) has revealed an overall prevalence of gallstones of 7.9% in men and 16.6% in women. The prevalence was high in Mexican Americans (8.9% in men, 26.7% in women), intermediate for non-Hispanic whites (8.6% in men, 16.6% in women), and low for African Americans (5.3% in men, 13.9% in women).

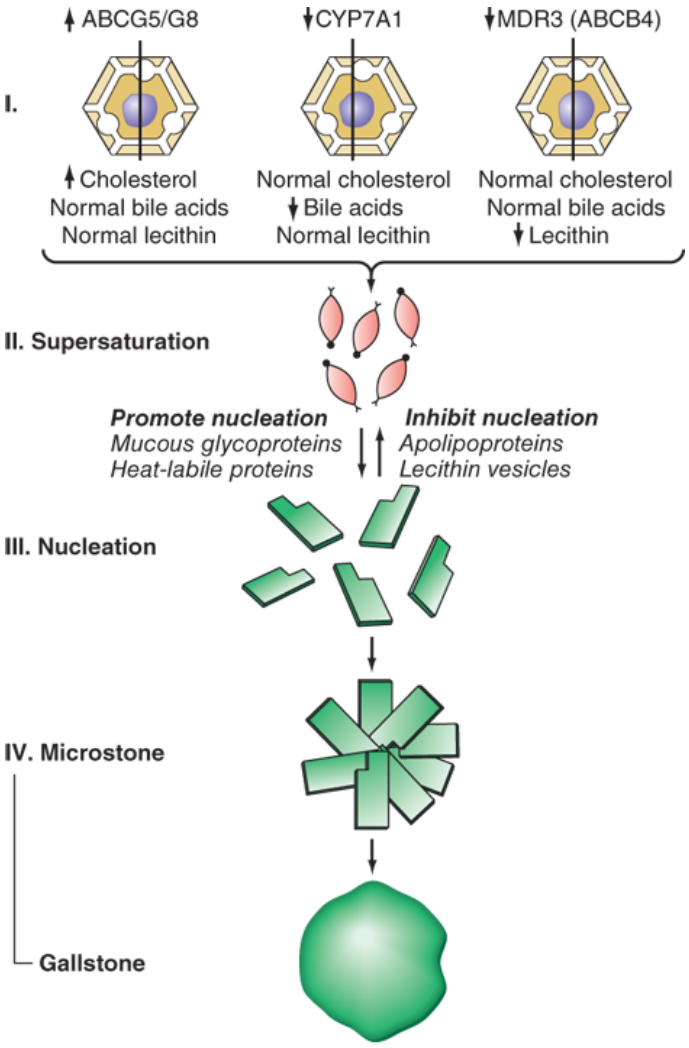
Gallstones are formed because of abnormal bile composition. They are divided into two major types: cholesterol stones and pigment stones. Cholesterol stones account for >90% of all gallstones in Western industrialized countries. Cholesterol gallstones usually contain >50% cholesterol monohydrate plus an admixture of calcium salts, bile pigments, proteins, and fatty acids. Pigment stones are composed primarily of calcium bilirubinate; they contain <20% cholesterol and are classified into “black” and “brown” types, the latter forming secondary to chronic biliary infection.

CHOLESTEROL STONES AND BILIARY SLUDGE

Cholesterol is essentially water-insoluble and requires aqueous dispersion into either micelles or vesicles, both of which require the presence of a second lipid to solubilize the cholesterol. Cholesterol and phospholipids are secreted into bile as unilamellar bilayered vesicles, which are converted into mixed micelles consisting of bile acids, phospholipids, and cholesterol by the action of bile acids. If there is an excess of cholesterol in relation to phospholipids and bile acids, unstable, cholesterol-rich vesicles remain, which aggregate into large multilamellar vesicles from which cholesterol crystals precipitate (Fig. 339-1).

FIGURE 339-1

Scheme showing pathogenesis of cholesterol gallstone formation. Conditions or factors that increase the ratio of cholesterol to bile acids and phospholipids (lecithin) favor gallstone formation. ABCB4, ATP-binding cassette transporter; ABCG5/8, ATP-binding cassette (ABC) transporter G5/G8; CYP7A1, cytochrome P450 7A1; MDR3, multidrug resistance protein 3, also called phospholipid export pump.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

There are several important mechanisms in the formation of lithogenic (stone-forming) bile. The most important is increased biliary secretion of cholesterol. This may occur in association with obesity, the metabolic syndrome, high-caloric and cholesterol-rich diets, or drugs (e.g., clofibrate) and

may result from increased activity of hydroxymethylglutaryl-coenzyme A (HMG-CoA) reductase, the rate-limiting enzyme of hepatic cholesterol synthesis, and increased hepatic uptake of cholesterol from blood. In patients with gallstones, dietary cholesterol *increases* biliary cholesterol secretion. This does not occur in non-gallstone patients on high-cholesterol diets. In addition to environmental factors such as high-caloric and cholesterol-rich diets, genetic factors play an important role in gallstone disease. A large study of symptomatic gallstones in Swedish twins provided strong evidence for a role of genetic factors in gallstone pathogenesis. Genetic factors accounted for 25%, shared environmental factors for 13%, and individual environmental factors for 62% of the phenotypic variation among monozygotic twins. A single nucleotide polymorphism of the gene encoding the hepatic cholesterol transporter ABCG5/G8 has been found in 21% of patients with gallstones, but only in 9% of the general population. It is thought to cause a gain of function of the cholesterol transporter and to contribute to cholesterol hypersecretion. A high prevalence of gallstones is found among first-degree relatives of gallstone carriers and in certain ethnic populations such as American Indians, Chilean Indians, and Chilean Hispanics. A common genetic trait has been identified for some of these populations by mitochondrial DNA analysis. In some patients, impaired hepatic conversion of cholesterol to bile acids may also occur, resulting in an increase of the lithogenic cholesterol/bile acid ratio. Although most cholesterol stones have a polygenic basis, there are rare monogenic (Mendelian) causes. Recently, a mutation in the *CYP7A1* gene has been described that results in a deficiency of the enzyme cholesterol 7-hydroxylase, which catalyzes the initial step in cholesterol catabolism and bile acid synthesis. The homozygous state is associated with hypercholesterolemia and gallstones. Because the phenotype is expressed in the heterozygote state, mutations in the *CYP7A1* gene may contribute to the susceptibility to cholesterol gallstone disease in the population. Mutations in the *MDR3* (ABCB4) gene, which encodes the phospholipid export pump in the canalicular membrane of the hepatocyte, may cause defective phospholipid secretion into bile, resulting in cholesterol supersaturation of bile and formation of cholesterol gallstones in the gallbladder and in the bile ducts. Thus, an excess of biliary cholesterol in relation to bile acids and phospholipids is primarily due to hypersecretion of cholesterol, but hyposecretion of bile acids or phospholipids may contribute. An additional disturbance of bile acid metabolism that is likely to contribute to supersaturation of bile with cholesterol is enhanced conversion of cholic acid to deoxycholic acid, with replacement of the cholic acid pool by an expanded deoxycholic acid pool. It may result from enhanced dehydroxylation of cholic acid and increased absorption of newly formed deoxycholic acid. An increased deoxycholate secretion is associated with hypersecretion of cholesterol into bile.

While supersaturation of bile with cholesterol is an important prerequisite for gallstone formation, it is generally not sufficient by itself to produce cholesterol precipitation in vivo. Most individuals with supersaturated bile do not develop stones because the time required for cholesterol crystals to nucleate and grow is longer than the time bile remains in the gallbladder.

An important mechanism is *nucleation* of cholesterol monohydrate crystals, which is greatly accelerated in human lithogenic bile. Accelerated nucleation of cholesterol monohydrate in bile may be due to either an *excess of pronucleating factors* or a *deficiency of antinucleating factors*. Mucin and certain nonmucin glycoproteins, principally immunoglobulins, appear to be pronucleating factors, while apolipoproteins A-I and A-II and other glycoproteins appear to be antinucleating factors. Pigment particles may possibly play a role as nucleating factors. In a genome-wide analysis of serum bilirubin levels, the uridine diphosphate-glucuronyltransferase 1A1 (*UGT1A1*) Gilbert's syndrome gene variant was associated with the presence of gallstone disease. Because most gallstones associated with the *UGT1A1* variant were cholesterol stones, this finding points to the role of pigment particles in the pathogenesis of gallbladder stones. Cholesterol monohydrate crystal nucleation and crystal growth probably occur within the mucin gel layer. Vesicle fusion leads to liquid crystals, which, in turn, nucleate into solid cholesterol monohydrate crystals. Continued growth of the crystals occurs by direct nucleation of cholesterol molecules from supersaturated unilamellar or multilamellar biliary vesicles.

A third important mechanism in cholesterol gallstone formation is *gallbladder hypomotility*. If the gallbladder emptied all supersaturated or crystal-containing bile completely, stones would not be able to grow. A high percentage of patients with gallstones exhibits abnormalities of gallbladder emptying. Ultrasonographic studies show that gallstone patients display an increased gallbladder volume during fasting and also after a test meal (residual volume) and that fractional emptying after gallbladder stimulation is decreased. The incidence of gallstones is increased in conditions associated with infrequent or impaired gallbladder emptying such as fasting, parenteral nutrition, or pregnancy and in patients using drugs that inhibit gallbladder motility.

Biliary sludge is a thick, mucous material that, upon microscopic examination, reveals lecithin-cholesterol liquid crystals, cholesterol monohydrate crystals, calcium bilirubinate, and mucin gels. Biliary sludge typically forms a crescent-like layer in the most dependent portion of the gallbladder and is recognized by characteristic echoes on ultrasonography (see below). The presence of biliary sludge implies two abnormalities: (1) the normal balance between gallbladder mucin secretion and elimination has become deranged, and (2) nucleation of biliary solutes has occurred. That biliary sludge may be a precursor form of gallstone disease is evident from several observations. In one study, 96 patients with gallbladder sludge were followed prospectively by serial ultrasound studies. In 18%, biliary sludge disappeared and did not recur for at least 2 years. In 60%, biliary sludge disappeared and reappeared; in 14%, gallstones (8% asymptomatic, 6% symptomatic) developed; and in 6%, severe biliary pain with or without acute pancreatitis occurred. In 12 patients, cholecystectomies were performed, 6 for gallstone-associated biliary pain and 3 in symptomatic patients with sludge but without gallstones who had prior attacks of pancreatitis; the latter did not recur after cholecystectomy. It should be emphasized that biliary sludge can develop with disorders that cause gallbladder hypomotility; that is, surgery, burns, total parenteral nutrition, pregnancy, and oral contraceptives—all of which are associated with gallstone formation. However, the presence of biliary sludge implies supersaturation of bile with either cholesterol or calcium bilirubinate.

Two other conditions are associated with cholesterol-stone or biliary-sludge formation: pregnancy and rapid weight reduction through a very-low-calorie diet. There appear to be two key changes during pregnancy that contribute to a “cholelithogenic state”: (1) a marked increase in cholesterol saturation of bile during the third trimester and (2) sluggish gallbladder contraction in response to a standard meal, resulting in impaired gallbladder emptying. That these changes are related to pregnancy per se is supported by several studies that show reversal of these abnormalities quite rapidly after delivery. During pregnancy, gallbladder sludge develops in 20–30% of women and gallstones in 5–12%. Although biliary sludge is a common finding during pregnancy, it is usually asymptomatic and often resolves spontaneously after delivery. Gallstones, which are less common than sludge and frequently associated with biliary colic, may also disappear after delivery because of spontaneous dissolution related to bile becoming unsaturated with cholesterol postpartum.

Approximately 10–20% of persons with rapid weight reduction achieved through very-low-calorie dieting develop gallstones. In a study involving 600 patients who completed a 3-month, 520-kcal/d diet, UDCA in a dosage of 600 mg/d proved highly effective in preventing gallstone formation; gallstones developed in only 3% of UDCA recipients, compared to 28% of placebo-treated patients. In obese patients treated by gastric banding, 500 mg/d of UDCA reduced the risk of gallstone formation from 30 to 8% within a follow-up of 6 months.

To summarize, cholesterol gallstone disease occurs because of several defects, which include (1) bile supersaturation with cholesterol, (2) nucleation of cholesterol monohydrate with subsequent crystal retention and stone growth, and (3) abnormal gallbladder motor function with delayed emptying and stasis. Other important factors known to predispose to cholesterol-stone formation are summarized in **Table 339-1**.

TABLE 339-1

Predisposing Factors for Cholesterol and Pigment Gallstone Formation

Cholesterol Stones
1. Demographic/genetic factors: Prevalence highest in North American Indians, Chilean Indians, and Chilean Hispanics, greater in Northern Europe and North America than in Asia, lowest in Japan; familial disposition; hereditary aspects 2. Obesity, metabolic syndrome: Normal bile acid pool and secretion but increased biliary secretion of cholesterol 3. Weight loss: Mobilization of tissue cholesterol leads to increased biliary cholesterol secretion while enterohepatic circulation of bile acids is decreased 4. Female sex hormones a. Estrogens stimulate hepatic lipoprotein receptors, increase uptake of dietary cholesterol, and increase biliary cholesterol secretion b. Natural estrogens, other estrogens, and oral contraceptives lead to decreased bile salt secretion and decreased conversion of cholesterol to cholesteryl esters 5. Pregnancy: Impaired gallbladder emptying caused by progesterone combined with the influence of estrogens, which increase biliary cholesterol secretion 6. Increasing age: Increased biliary secretion of cholesterol, decreased size of bile acid pool, decreased secretion of bile salts 7. Gallbladder hypomotility leading to stasis and formation of sludge a. Prolonged parenteral nutrition b. Fasting c. Pregnancy d. Drugs such as octreotide 8. Clofibrate therapy: Increased biliary secretion of cholesterol 9. Decreased bile acid secretion a. Primary biliary cirrhosis b. Genetic defect of the *CYP7A1* gene 10. Decreased phospholipid secretion: Genetic defect of the *MDR3* gene 11. Miscellaneous a. High-calorie, high-fat diet b. Spinal cord injury
Pigment Stones
1. Demographic/genetic factors: Asia, rural setting 2. Chronic hemolysis 3. Alcoholic liver cirrhosis 4. Pernicious anemia 5. Cystic fibrosis 6. Chronic biliary tract infection, parasite infections 7. Increasing age 8. Ileal disease, ileal resection or bypass

PIGMENT STONES

Black pigment stones are composed of either pure calcium bilirubinate or polymer-like complexes with calcium and mucin glycoproteins. They are more common in patients who have chronic hemolytic states (with increased conjugated bilirubin in bile), liver cirrhosis, Gilbert's syndrome, or cystic fibrosis. Gallbladder stones in patients with ileal diseases, ileal resection, or ileal bypass generally are also black pigment stones. Enterohepatic recycling of bilirubin in ileal disease states contributes to their pathogenesis. Brown pigment stones are composed of calcium salts of unconjugated bilirubin with varying amounts of cholesterol and protein. They are caused by the presence of increased amounts of unconjugated, insoluble bilirubin in bile that precipitates to form stones. Deconjugation of an excess of soluble bilirubin mono- and diglucuronides may be mediated by endogenous β-glucuronidase but may also occur by spontaneous hydrolysis. Sometimes, the enzyme is also produced when bile is chronically infected by bacteria, and such stones are brown. Pigment stone formation is frequent in Asia and is often associated with infections in the gallbladder and biliary tree (Table 339-1).

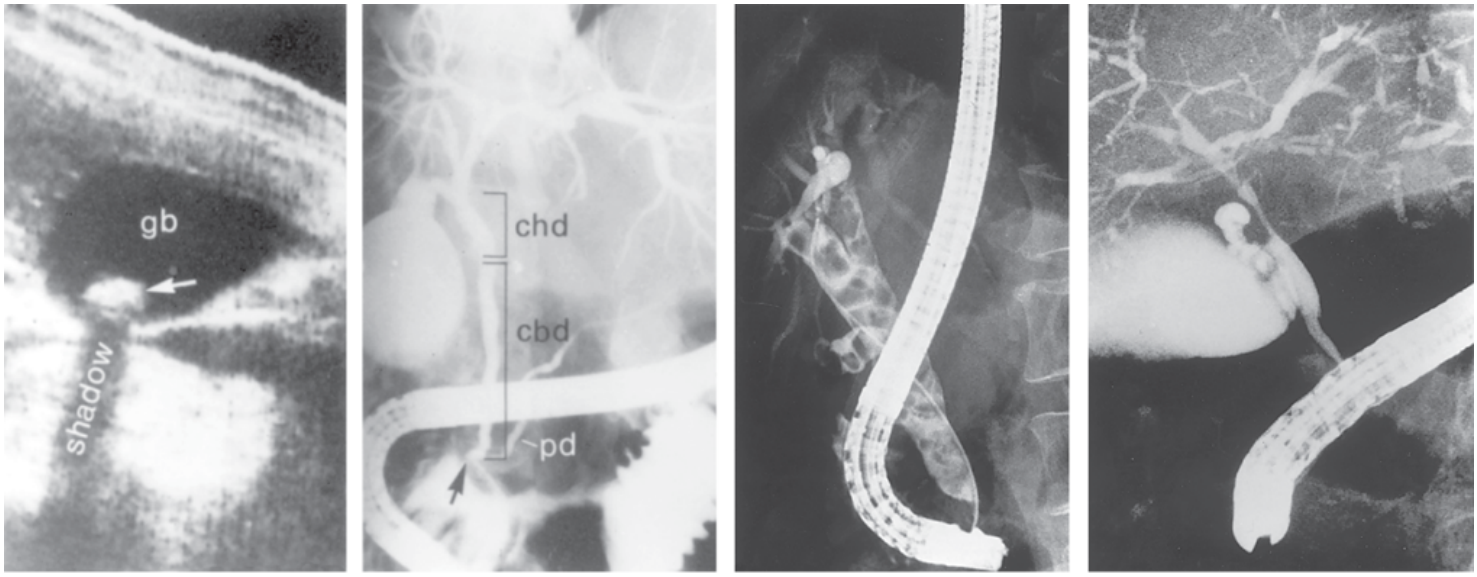
Diagnosis

Procedures of potential use in the diagnosis of cholelithiasis and other diseases of the gallbladder are detailed in **Table 339-2**. Ultrasonography of the gallbladder is very accurate in the identification of cholelithiasis and has replaced oral cholecystography (OCG) **(Fig. 339-2A)**. Stones as small as 1.5 mm

in diameter may be confidently identified provided that firm criteria are used (e.g., acoustic “shadowing” of opacities that are within the gallbladder lumen and that change with the patient’s position [by gravity]). In major medical centers, the false-negative and false-positive rates for ultrasound in gallstone patients are ~2–4%. Biliary sludge is material of low echogenic activity that typically forms a layer in the most dependent position of the gallbladder. This layer shifts with postural changes but fails to produce acoustic shadowing; these two characteristics distinguish sludges from gallstones. Ultrasound can also be used to assess the emptying function of the gallbladder.

FIGURE 339-2

Examples of ultrasound and radiologic studies of the biliary tract. ***A.*** An ultrasound study showing a distended gallbladder (GB) containing a single large stone (*arrow*), which casts an acoustic shadow. ***B.*** Endoscopic retrograde cholangiopancreatogram (ERCP) showing normal biliary tract anatomy. In addition to the endoscope and large vertical gallbladder filled with contrast dye, the common hepatic duct (CHD), common bile duct (CBD), and pancreatic duct (PD) are shown. The *arrow* points to the ampulla of Vater. ***C.*** Endoscopic retrograde cholangiogram (ERC) showing choledocholithiasis. The biliary tract is dilated and contains multiple radiolucent calculi. ***D.*** ERCP showing sclerosing cholangitis. The CBD shows areas that are strictured and narrowed.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 339-2

Diagnostic Evaluation of the Gallbladder

Diagnostic Advantages	Diagnostic Limitations	Comment
Gallbladder Ultrasound		
Rapid Accurate identification of gallstones (>95%) Simultaneous scanning of GB, liver, bile ducts, pancreas "Real-time" scanning allows assessment of GB volume, contractility Not limited by jaundice, pregnancy May detect very small stones	Bowel gas Massive obesity Ascites	Procedure of choice for detection of stones
Plain Abdominal X-Ray		
Low cost Readily available	Relatively low yield Contraindicated in pregnancy	Pathognomonic findings in: calcified gallstones Limey bile, porcelain GB Emphysematous cholecystitis Gallstone ileus
Radioisotope Scans (HIDA, DIDA, etc.)		
Accurate identification of cystic duct obstruction Simultaneous assessment of bile ducts	Contraindicated in pregnancy Serum bilirubin >103–205 µmol/L (6–12 mg/dL) Cholecystogram of low resolution	Indicated for confirmation of suspected acute cholecystitis; less sensitive and less specific in chronic cholecystitis; useful indiagnosis of acalculous cholecystopathy, especially if given with CCK to assess gallbladder emptying

Abbreviations: CCK, cholecystokinin; GB, gallbladder.

The plain abdominal film may detect gallstones containing sufficient calcium to be radiopaque (10–15% of cholesterol and ~50% of pigment stones). Plain radiography may also be of use in the diagnosis of emphysematous cholecystitis, porcelain gallbladder, limey bile, and gallstone ileus.

OCG has historically been a useful procedure for the diagnosis of gallstones but has been replaced by ultrasound and is regarded as obsolete. It may be used to assess the patency of the cystic duct and gallbladder emptying function. Further, OCG can also delineate the size and number of gallstones and determine whether they are calcified.

Radiopharmaceuticals such as ^{99m}Tc-labeled *N*-substituted iminodiacetic acids (HIDA, DIDA, DISIDA, etc.) are rapidly extracted from the blood and are excreted into the biliary tree in high concentration even in the presence of mild to moderate serum bilirubin elevations. Failure to image the gallbladder in the presence of biliary ductal visualization may indicate cystic duct obstruction, acute or chronic cholecystitis, or surgical absence of the organ. Such scans have some application in the diagnosis of acute cholecystitis.

Symptoms of Gallstone Disease

Gallstones usually produce symptoms by causing inflammation or obstruction following their migration into the cystic duct or CBD. The most specific and characteristic symptom of gallstone disease is biliary colic that is a constant and often long-lasting pain (see below). Obstruction of the cystic duct or CBD by a stone produces increased intraluminal pressure and distention of the viscus that cannot be relieved by repetitive biliary contractions. The

resultant visceral pain is characteristically a severe, steady ache or fullness in the epigastrium or right upper quadrant (RUQ) of the abdomen with frequent radiation to the interscapular area, right scapula, or shoulder.

Biliary colic begins quite suddenly and may persist with severe intensity for 30 min to 5 h, subsiding gradually or rapidly. It is steady rather than intermittent, as would be suggested by the word *colic*, which must be regarded as a misnomer, although it is in widespread use. An episode of biliary pain persisting beyond 5 h should raise the suspicion of acute cholecystitis (see below). Nausea and vomiting frequently accompany episodes of biliary pain. An elevated level of serum bilirubin and/or alkaline phosphatase suggests a common duct stone. Fever or chills (rigors) with biliary pain usually imply a complication, that is, cholecystitis, pancreatitis, or cholangitis. Complaints of short-lasting, vague epigastric fullness, dyspepsia, eructation, or flatulence, especially following a fatty meal, should not be confused with biliary pain. Such symptoms are frequently elicited from patients with or without gallstone disease but are not specific for biliary calculi. Biliary colic may be precipitated by eating a fatty meal, by consumption of a large meal following a period of prolonged fasting, or by eating a normal meal; it is frequently nocturnal, occurring within a few hours of retiring.

Natural History

Gallstone disease discovered in an asymptomatic patient or in a patient whose symptoms are not referable to cholelithiasis is a common clinical problem. Sixty to 80% of persons with asymptomatic gallstones remain asymptomatic over follow-up periods of up to 25 years. The probability of developing symptoms within 5 years after diagnosis is 2–4% per year and decreases in the years thereafter to 1–2%. The yearly incidence of complications is about 0.1–0.3%. Patients remaining asymptomatic for 15 years were found to be unlikely to develop symptoms during further follow-up, and most patients who did develop complications from their gallstones experienced *prior* warning symptoms. Similar conclusions apply to diabetic patients with silent gallstones. Decision analysis has suggested that (1) the cumulative risk of death due to gallstone disease while on expectant management is small, and (2) prophylactic cholecystectomy is not warranted.

Complications requiring cholecystectomy are much more common in gallstone patients who have developed symptoms of biliary pain. Patients found to have gallstones at a young age are more likely to develop symptoms from cholelithiasis than are patients >60 years at the time of initial diagnosis. Patients with diabetes mellitus and gallstones may be somewhat more susceptible to septic complications, but the magnitude of risk of septic biliary complications in diabetic patients is incompletely defined.

TREATMENT

TREATMENT

Gallstones

SURGICAL THERAPY

In asymptomatic gallstone patients, the risk of developing symptoms or complications requiring surgery is quite small (see above). Thus, a recommendation for cholecystectomy in a patient with gallstones should probably be based on assessment of three factors: (1) the presence of symptoms that are frequent enough or severe enough to interfere with the patient's general routine; (2) the presence of a prior complication of gallstone disease, that is, history of acute cholecystitis, pancreatitis, gallstone fistula, etc.; or (3) the presence of an underlying condition predisposing the patient to increased risk of gallstone complications (e.g., calcified or porcelain gallbladder and/or a previous attack of acute cholecystitis regardless of current symptomatic status). Patients with very large gallstones (>3 cm in diameter) and patients harboring gallstones in a congenitally anomalous gallbladder might also be considered for prophylactic cholecystectomy. Although young age is a worrisome factor in asymptomatic gallstone patients, few authorities would now recommend routine cholecystectomy in all young patients with silent stones. Laparoscopic cholecystectomy is a minimal-access approach for the removal of the gallbladder together with its stones. Its advantages include a markedly shortened hospital stay, minimal disability, and decreased cost, and it is the procedure of choice for most patients referred for elective cholecystectomy.

From several studies involving >4000 patients undergoing laparoscopic cholecystectomy, the following key points emerge: (1) complications develop in ~4% of patients, (2) conversion to laparotomy occurs in 5%, (3) the death rate is remarkably low (i.e., <0.1%), and (4) the rate of bile duct injuries is low (i.e., 0.2–0.6%) and comparable with open cholecystectomy. These data indicate why laparoscopic cholecystectomy has become the "gold standard" for treating symptomatic cholelithiasis.

MEDICAL THERAPY—GALLSTONE DISSOLUTION

In carefully selected patients with a functioning gallbladder and with radiolucent stones <10 mm in diameter, complete dissolution can be achieved in ~50% of patients within 6 months to 2 years. For good results within a reasonable time period, this therapy should be limited to radiolucent stones <5 mm in diameter. The dose of UDCA should be 10–15 mg/kg per day. Stones >10 mm in size rarely dissolve. Pigment stones are not responsive to UDCA therapy. Probably ≤10% of patients with *symptomatic* cholelithiasis are candidates for such treatment. However, in addition to the vexing problem of recurrent stones (30–50% over 3–5 years of follow-up), there is also the factor of taking an expensive drug for up to 2 years. The advantages and success of laparoscopic cholecystectomy have largely reduced the role of gallstone dissolution to patients who wish to avoid or are not candidates for elective cholecystectomy. However, patients with cholesterol gallstone disease who develop recurrent choledocholithiasis after cholecystectomy should be on long-term treatment with UDCA.

ACUTE AND CHRONIC CHOLECYSTITIS

Acute Cholecystitis

Acute inflammation of the gallbladder wall usually follows obstruction of the cystic duct by a stone. Inflammatory response can be evoked by three factors: (1) *mechanical inflammation* produced by increased intraluminal pressure and distention with resulting ischemia of the gallbladder mucosa and wall, (2) *chemical inflammation* caused by the release of lysolecithin (due to the action of phospholipase on lecithin in bile) and other local tissue factors, and (3) *bacterial inflammation*, which may play a role in 50–85% of patients with acute cholecystitis. The organisms most frequently isolated by culture of gallbladder bile in these patients include *Escherichia coli*, *Klebsiella* spp., *Streptococcus* spp., and *Clostridium* spp.

Acute cholecystitis often begins as an attack of biliary pain that progressively worsens. Approximately 60–70% of patients report having experienced prior attacks that resolved spontaneously. As the episode progresses, however, the pain of acute cholecystitis becomes more generalized in the right upper abdomen. As with biliary colic, the pain of cholecystitis may radiate to the interscapular area, right scapula, or shoulder. Peritoneal signs of inflammation such as increased pain with jarring or on deep respiration may be apparent. The patient is anorectic and often nauseated. Vomiting is relatively common and may produce symptoms and signs of vascular and extracellular volume depletion. Jaundice is unusual early in the course of acute cholecystitis but may occur when edematous inflammatory changes involve the bile ducts and surrounding lymph nodes.

A low-grade fever is characteristically present, but shaking chills or rigors are not uncommon. The RUQ of the abdomen is almost invariably tender to palpation. An enlarged, tense gallbladder is palpable in 25–50% of patients. Deep inspiration or cough during subcostal palpation of the RUQ usually produces increased pain and inspiratory arrest (Murphy's sign). Localized rebound tenderness in the RUQ is common, as are abdominal distention and hypoactive bowel sounds from paralytic ileus, but generalized peritoneal signs and abdominal rigidity are usually lacking, in the absence of perforation.

The diagnosis of acute cholecystitis is usually made on the basis of a characteristic history and physical examination. The triad of sudden onset of RUQ tenderness, fever, and leukocytosis is highly suggestive. Typically, leukocytosis in the range of 10,000–15,000 cells per microliter with a left shift on differential count is found. The serum bilirubin is mildly elevated (<85.5 µmol/L [5 mg/dL]) in fewer than half of patients, whereas about one-fourth have modest elevations in serum aminotransferases (usually less than a fivefold elevation). Ultrasound will demonstrate calculi in 90–95% of cases and is useful for detection of signs of gallbladder inflammation including thickening of the wall, pericholecystic fluid, and dilatation of the bile duct. The radionuclide (e.g., HIDA) biliary scan may be confirmatory if bile duct imaging is seen without visualization of the gallbladder.

Approximately 75% of patients treated medically have remission of acute symptoms within 2–7 days following hospitalization. In 25%, however, a complication of acute cholecystitis will occur despite conservative treatment (see below). In this setting, prompt surgical intervention is required. Of the 75% of patients with acute cholecystitis who undergo remission of symptoms, ~25% will experience a recurrence of cholecystitis within 1 year, and 60% will have at least one recurrent bout within 6 years. In view of the natural history of the disease, acute cholecystitis is best treated by early surgery whenever possible. Mirizzi's syndrome is a rare complication in which a gallstone becomes impacted in the cystic duct or neck of the gallbladder causing compression of the CBD, resulting in CBD obstruction and jaundice. Ultrasound shows gallstone(s) lying outside the hepatic duct. Endoscopic retrograde cholangiopancreatography (ERCP) **(Fig. 339-2B)**, percutaneous transhepatic cholangiography (PTC), or magnetic resonance cholangiopancreatography (MRCP) will usually demonstrate the characteristic extrinsic compression of the CBD. Surgery consists of removing the cystic duct, diseased gallbladder, and the impacted stone. The preoperative diagnosis of Mirizzi's syndrome is important to avoid CBD injury.

ACALCULOUS CHOLECYSTITIS

In 5–10% of patients with acute cholecystitis, calculi obstructing the cystic duct are not found at surgery. In >50% of such cases, an underlying explanation for acalculous inflammation is not found. An increased risk for the development of acalculous cholecystitis is especially associated with prolonged fasting, serious trauma or burns, with the postpartum period following prolonged labor, and with orthopedic and other nonbiliary major surgical operations in the postoperative period. It may possibly complicate periods of prolonged parenteral hyperalimentation. For some of these cases, biliary sludge in the cystic duct may be responsible. Other precipitating factors include vasculitis, obstructing adenocarcinoma of the gallbladder, diabetes mellitus, torsion of the gallbladder, "unusual" bacterial infections of the gallbladder (e.g., *Leptospira*, *Streptococcus*, *Salmonella*, or *Vibrio cholerae*), and parasitic infestation of the gallbladder. Acalculous cholecystitis may also be seen with a variety of other systemic disease processes (e.g., sarcoidosis, cardiovascular disease, tuberculosis, syphilis, actinomycosis).

Although the clinical manifestations of acalculous cholecystitis are indistinguishable from those of calculous cholecystitis, the setting of acute gallbladder inflammation complicating severe underlying illness is characteristic of acalculous disease. Ultrasound or computed tomography (CT) examinations demonstrating a large, tense, static gallbladder without stones and with evidence of poor emptying over a prolonged period may be diagnostically useful in some cases. The complication rate for acalculous cholecystitis exceeds that for calculous cholecystitis. Successful management of acute acalculous cholecystitis appears to depend primarily on early diagnosis and surgical intervention, with meticulous attention to postoperative care.

ACALCULOUS CHOLECYSTOPATHY

Disordered motility of the gallbladder can produce recurrent biliary pain in patients without gallstones. Infusion of an octapeptide of CCK can be used to measure the gallbladder ejection fraction during cholescintigraphy. The surgical findings have included abnormalities such as chronic cholecystitis, gallbladder muscle hypertrophy, and/or a markedly narrowed cystic duct. Some of these patients may well have had antecedent gallbladder disease.

The following criteria can be used to identify patients with acalculous cholecystopathy: (1) recurrent episodes of typical RUQ pain characteristic of biliary tract pain, (2) abnormal CCK cholescintigraphy demonstrating a gallbladder ejection fraction of <40%, and (3) infusion of CCK reproducing the patient's pain. An additional clue would be the identification of a large gallbladder on ultrasound examination. Importantly, it should be noted that SOD dysfunction can also give rise to recurrent RUQ pain and CCK-scintigraphic abnormalities.

EMPHYSEMATOUS CHOLECYSTITIS

So-called emphysematous cholecystitis is thought to begin with acute cholecystitis (calculous or acalculous) followed by ischemia or gangrene of the gallbladder wall and infection by gas-producing organisms. Bacteria most frequently cultured in this setting include anaerobes, such as *Clostridium welchii* or *C. perfringens*, and aerobes, such as *E. coli.* This condition occurs most frequently in elderly men and in patients with diabetes mellitus. The clinical manifestations are essentially indistinguishable from those of nongaseous cholecystitis. The diagnosis is usually made on plain abdominal film by finding gas within the gallbladder lumen, dissecting within the gallbladder wall to form a gaseous ring, or in the pericholecystic tissues. The morbidity and mortality rates with emphysematous cholecystitis are considerable. Prompt surgical intervention coupled with appropriate antibiotics is mandatory.

Chronic Cholecystitis

Chronic inflammation of the gallbladder wall is almost always associated with the presence of gallstones and is thought to result from repeated bouts of subacute or acute cholecystitis or from persistent mechanical irritation of the gallbladder wall by gallstones. The presence of bacteria in the bile occurs in >25% of patients with chronic cholecystitis. The presence of infected bile in a patient with *chronic* cholecystitis undergoing elective cholecystectomy probably adds little to the operative risk. Chronic cholecystitis may be asymptomatic for years, which may progress to symptomatic gallbladder disease or to acute cholecystitis, or may present with complications (see below).

Complications of Cholecystitis

EMPYEMA AND HYDROPS

Empyema of the gallbladder usually results from progression of acute cholecystitis with persistent cystic duct obstruction to superinfection of the stagnant bile with a pus-forming bacterial organism. The clinical picture resembles that of cholangitis with high fever; severe RUQ pain; marked leukocytosis; and often, prostration. Empyema of the gallbladder carries a high risk of gram-negative sepsis and/or perforation. Emergency surgical intervention with proper antibiotic coverage is required as soon as the diagnosis is suspected.

Hydrops or mucocele of the gallbladder may also result from prolonged obstruction of the cystic duct, usually by a large solitary calculus. In this instance, the obstructed gallbladder lumen is progressively distended, over a period of time, by mucus (mucocele) or by a clear transudate (hydrops) produced by mucosal epithelial cells. A visible, easily palpable, nontender mass sometimes extending from the RUQ into the right iliac fossa may be found on physical examination. The patient with hydrops of the gallbladder frequently remains asymptomatic, although chronic RUQ pain may also occur. Cholecystectomy is indicated, because empyema, perforation, or gangrene may complicate the condition.

GANGRENE AND PERFORATION

Gangrene of the gallbladder results from ischemia of the wall and patchy or complete tissue necrosis. Underlying conditions often include marked distention of the gallbladder, vasculitis, diabetes mellitus, empyema, or torsion resulting in arterial occlusion. Gangrene usually predisposes to perforation of the gallbladder, but perforation may also occur in chronic cholecystitis without premonitory warning symptoms. *Localized perforations* are usually contained by the omentum or by adhesions produced by recurrent inflammation of the gallbladder. Bacterial superinfection of the walled-off gallbladder contents results in abscess formation. Most patients are best treated with cholecystectomy, but some seriously ill patients may be managed with cholecystostomy and drainage of the abscess. *Free perforation* is less common but is associated with a mortality rate of ~30%. Such patients may experience a sudden transient relief of RUQ pain as the distended gallbladder decompresses; this is followed by signs of generalized peritonitis.

FISTULA FORMATION AND GALLSTONE ILEUS

Fistula formation into an adjacent organ adherent to the gallbladder wall may result from inflammation and adhesion formation. Fistulas into the duodenum are most common, followed in frequency by those involving the hepatic flexure of the colon, stomach or jejunum, abdominal wall, and renal pelvis. Clinically "silent" biliary-enteric fistulas occurring as a complication of acute cholecystitis have been found in up to 5% of patients undergoing cholecystectomy. Asymptomatic cholecystoenteric fistulas may sometimes be diagnosed by finding gas in the biliary tree on plain abdominal films. Barium contrast studies or endoscopy of the upper gastrointestinal tract or colon may demonstrate the fistula. Treatment in the symptomatic patient usually consists of cholecystectomy, CBD exploration, and closure of the fistulous tract.

Gallstone ileus refers to mechanical intestinal obstruction resulting from the passage of a large gallstone into the bowel lumen. The stone customarily enters the duodenum through a cholecystoenteric fistula at that level. The site of obstruction by the impacted gallstone is usually at the ileocecal valve, provided that the more proximal small bowel is of normal caliber. The majority of patients do not give a history of either prior biliary tract symptoms or complaints suggestive of acute cholecystitis or fistula formation. Large stones, >2.5 cm in diameter, are thought to predispose to fistula formation by gradual erosion through the gallbladder fundus. Diagnostic confirmation may occasionally be found on the plain abdominal film (e.g., small-intestinal obstruction with gas in the biliary tree [pneumobilia] and a calcified, ectopic gallstone) or following an upper gastrointestinal series

(cholecystoduodenal fistula with small-bowel obstruction at the ileocecal valve). Laparotomy with stone extraction (or propulsion into the colon) remains the procedure of choice to relieve obstruction. Evacuation of large stones within the gallbladder should also be performed. In general, the gallbladder and its attachment to the intestines should be left alone.

LIMEY (MILK OF CALCIUM) BILE AND PORCELAIN GALLBLADDER

Calcium salts in the lumen of the gallbladder in sufficient concentration may produce calcium precipitation and diffuse, hazy opacification of bile or a layering effect on plain abdominal roentgenography. This so-called limey bile, or milk of calcium bile, is usually clinically innocuous, but cholecystectomy is recommended, especially when it occurs in a hydropic gallbladder. In the entity called *porcelain gallbladder*, calcium salt deposition within the wall of a chronically inflamed gallbladder may be detected on the plain abdominal film. Cholecystectomy is advised in all patients with porcelain gallbladder because in a high percentage of cases this finding appears to be associated with the development of carcinoma of the gallbladder.

TREATMENT

TREATMENT

Acute Cholecystitis

MEDICAL THERAPY

Although surgical intervention remains the mainstay of therapy for acute cholecystitis and its complications, a period of in-hospital stabilization may be required before cholecystectomy. Oral intake is eliminated, nasogastric suction may be indicated, and extracellular volume depletion and electrolyte abnormalities are repaired. Meperidine or nonsteroidal anti-inflammatory drugs (NSAIDs) such as ketoralac or opioids, that is, morphine and hydromorphone, are usually employed for analgesia. Intravenous antibiotic therapy is usually indicated in patients with severe acute cholecystitis, even though bacterial superinfection of bile may not have occurred in the early stages of the inflammatory process. Antibiotic therapy is guided by the most common gram-negative organisms and anaerobes likely to be present, which are *E. coli*, *Klebsiella* spp., and *Streptococcus* spp. Effective antibiotics include piperacillin plus tazobactam, ceftriaxone plus metronidazole, levofloxacin plus metronidazole. Anaerobic coverage by a drug such as metronidazole should be added if gangrenous or emphysematous cholecystitis is suspected. Imipenem and meropenem should be reserved for the most severe, life-threatening infections when other regimens have failed **(Chap. 156)**. Postoperative complications of wound infection, abscess formation, and sepsis are reduced in antibiotic-treated patients.

SURGICAL THERAPY

The optimal timing of surgical intervention in patients with acute cholecystitis depends on stabilization of the patient. The clear trend is toward earlier surgery, and this is due in part to requirements for shorter hospital stays. Urgent (emergency) cholecystectomy or cholecystostomy is probably appropriate in most patients in whom a complication of acute cholecystitis such as empyema, emphysematous cholecystitis, or perforation is suspected or confirmed. Patients with uncomplicated acute cholecystitis should undergo early elective laparoscopic cholecystectomy, ideally within 48–72 h after diagnosis. The complication rate is not increased in patients undergoing early as opposed to delayed (>6 weeks after diagnosis) cholecystectomy. Delayed surgical intervention is probably best reserved for (1) patients in whom the overall medical condition imposes an unacceptable risk for early surgery and (2) patients in whom the diagnosis of acute cholecystitis is in doubt. Thus, early cholecystectomy (within 72 h) is the treatment of choice for most patients with acute cholecystitis. Mortality figures for emergency cholecystectomy in most centers range from 1 to 3%, whereas the mortality risk for early elective cholecystectomy is ~0.5% in patients under age 60. Of course, the operative risks increase with age-related diseases of other organ systems and with the presence of long- or short-term complications of gallbladder disease. Seriously ill or debilitated patients with cholecystitis may be managed with cholecystostomy and tube drainage of the gallbladder. Elective cholecystectomy may then be done at a later date.

Postcholecystectomy Complications

Early complications following cholecystectomy include atelectasis and other pulmonary disorders, abscess formation (often subphrenic), external or internal hemorrhage, biliary-enteric fistula, and bile leaks. Jaundice may indicate absorption of bile from an intraabdominal collection following a biliary leak or mechanical obstruction of the CBD by retained calculi, intraductal blood clots, or extrinsic compression.

Overall, cholecystectomy is a very successful operation that provides total or near-total relief of preoperative symptoms in 75–90% of patients. The most common cause of persistent postcholecystectomy symptoms is an overlooked symptomatic nonbiliary disorder (e.g., reflux esophagitis, peptic ulceration, pancreatitis, or—most often—irritable bowel syndrome). In a small percentage of patients, however, a disorder of the extrahepatic bile ducts may result in persistent symptomatology. These so-called postcholecystectomy syndromes may be due to (1) biliary strictures, (2) retained biliary calculi, (3) cystic duct stump syndrome, (4) stenosis or dyskinesia of the SOD, or (5) bile salt–induced diarrhea or gastritis.

CYSTIC DUCT STUMP SYNDROME

In the absence of cholangiographically demonstrable retained stones, symptoms resembling biliary pain or cholecystitis in the postcholecystectomy patient have frequently been attributed to disease in a long (>1 cm) cystic duct remnant (cystic duct stump syndrome). Careful analysis, however, reveals that postcholecystectomy complaints are attributable to other causes in almost all patients in whom the symptom complex was originally

thought to result from the existence of a long cystic duct stump. Accordingly, considerable care should be taken to investigate the possible role of other factors in the production of postcholecystectomy symptoms before attributing them to cystic duct stump syndrome.

SOD STENOSIS AND SOD DYSKINESIA, AND BILIARY DYSKINESIA

Symptoms of biliary colic accompanied by signs of recurrent, intermittent biliary obstruction may be produced by acalculous cholecystopathy, SOD stenosis, or SOD dyskinesia. SOD stenosis is thought to result from acute or chronic inflammation of the papilla of Vater or from glandular hyperplasia of the papillary segment. Five criteria have been used to define SOD stenosis: (1) upper abdominal pain, usually RUQ or epigastric; (2) abnormal liver tests; (3) dilatation of the CBD upon MRCP or ERCP examination; (4) delayed (>45 min) drainage of contrast material from the duct; and (5) increased basal pressure of the SOD. After exclusion of acalculous cholecystopathy, treatment consists of endoscopic or surgical sphincteroplasty to ensure wide patency of the distal portions of both the bile and pancreatic ducts. The greater the number of the preceding criteria present, the greater is the likelihood that a patient does have a degree of SOD sufficient to justify correction. The factors usually considered as indications for sphincterotomy include (1) prolonged duration of symptoms, (2) lack of response to symptomatic treatment, (3) presence of severe disability, and (4) the patient's choice of sphincterotomy over surgery (given a clear understanding on his or her part of the risks involved in both procedures).

Biliary SOD disorders are characterized by three criteria: (1) biliary pain, (2) absence of bile duct stones or other abnormalities, and (3) elevated liver enzymes or a dilated CBD, but not both. In this setting, either hepatobiliary scintigraphy or SOD manometry can support the diagnosis. Importantly, the presence of both elevated liver enzymes and a dilated CBD should raise the question of obstruction. Proposed mechanisms to account for SOD dysfunction include spasm of the sphincter, denervation sensitivity resulting in hypertonicity, and abnormalities in the sequencing or frequency rates of the sphincteric-contraction waves. When thorough evaluation has failed to demonstrate another cause for the pain and when cholangiographic and manometric criteria suggest a diagnosis of SOD dyskinesia, medical treatment with nitrites or anticholinergics to attempt pharmacologic relaxation of SOD has been proposed but not evaluated in detailed studies. Endoscopic biliary sphincterotomy (EBS) or surgical sphincterotomy may be indicated in patients who fail to respond to a 2–3 month trial of medical therapy, especially if SOD pressures are elevated. Approximately 45% of such patients have long-term pain relief after EBS. Endoscopic biliary sphincterotomy has become the procedure of choice for removing bile duct stones and for other biliary and pancreatic problems.

BILE SALT–INDUCED DIARRHEA AND GASTRITIS

Postcholecystectomy patients may develop symptoms of dyspepsia, which have been attributed to duodenogastric reflux of bile. However, firm data linking these symptoms to bile gastritis after surgical removal of the gallbladder are lacking. Cholecystectomy induces persistent changes in gut transit, and these changes effect a noticeable modification of bowel habits. Cholecystectomy shortens gut transit time by accelerating passage of the fecal bolus through the colon with marked acceleration in the right colon, thus causing an increase in colonic bile acid output and a shift in bile acid composition toward the more diarrheagenic secondary bile acids, that is, deoxycholic acid. Diarrhea that is severe enough, that is, three or more watery movements per day, can be classified as postcholecystectomy diarrhea, and this occurs in 5–10% of patients undergoing elective cholecystectomy. Treatment with bile acid–sequestering agents such as cholestyramine or colestipol is often effective in ameliorating troublesome diarrhea.

THE HYPERPLASTIC CHOLECYSTOSES

The term *hyperplastic cholecystoses* is used to denote a group of disorders of the gallbladder characterized by excessive proliferation of normal tissue components.

Adenomyomatosis is characterized by a benign proliferation of gallbladder surface epithelium with glandlike formations, extramural sinuses, transverse strictures, and/or fundal nodule ("adenoma" or "adenomyoma") formation.

Cholesterolosis is characterized by abnormal deposition of lipid, especially cholesteryl esters, within macrophages in the lamina propria of the gallbladder wall. In its diffuse form ("strawberry gallbladder"), the gallbladder mucosa is brick red and speckled with bright yellow flecks of lipid. The localized form shows solitary or multiple "cholesterol polyps" studding the gallbladder wall. Cholesterol stones of the gallbladder are found in nearly half the cases. Cholecystectomy is indicated in both adenomyomatosis and cholesterolosis when symptomatic or when cholelithiasis is present.

The prevalence of gallbladder polyps in the adult population is 1–4% with a marked male predominance. Types of gallbladder polyps include cholesterol polyps, adenomyomas, inflammatory polyps, and adenomas (rare). No significant changes have been found over a 5-year period in asymptomatic patients with gallbladder polyps <6 mm and few changes in polyps 7–9 mm. Cholecystectomy is recommended in symptomatic patients as well as in asymptomatic patients >50 years whose polyps are >10 mm or associated with gallstones or polyp growth on serial ultrasonography.

DISEASES OF THE BILE DUCTS

CONGENITAL ANOMALIES

Biliary Atresia and Hypoplasia

Atretic and hypoplastic lesions of the extrahepatic and large intrahepatic bile ducts are the most common biliary anomalies of clinical relevance encountered in infancy. The clinical picture is one of severe obstructive jaundice during the first month of life, with pale stools. When biliary atresia is

suspected on the basis of clinical, laboratory, and imaging findings, the diagnosis is confirmed by surgical exploration and operative cholangiography. Approximately 10% of cases of biliary atresia are treatable with Roux-en-Y choledochojejunostomy, with the Kasai procedure (hepatic portoenterostomy) being attempted in the remainder in an effort to restore some bile flow. Most patients, even those having successful biliary-enteric anastomoses, eventually develop chronic cholangitis, extensive hepatic fibrosis, and portal hypertension.

Choledochal Cysts

Cystic dilatation may involve the free portion of the CBD, that is, choledochal cyst, or may present as diverticulum formation in the intraduodenal segment. In the latter situation, chronic reflux of pancreatic juice into the biliary tree can produce inflammation and stenosis of the extrahepatic bile ducts leading to cholangitis or biliary obstruction. Because the process may be gradual, ~50% of patients present with onset of symptoms after age 10. The diagnosis may be made by ultrasound, abdominal CT, MRC, or cholangiography. Only one-third of patients show the classic triad of abdominal pain, jaundice, and an abdominal mass. Ultrasonographic detection of a cyst separate from the gallbladder should suggest the diagnosis of choledochal cyst, which can be confirmed by demonstrating the entrance of extrahepatic bile ducts into the cyst. Surgical treatment involves excision of the "cyst" and biliary-enteric anastomosis. Patients with choledochal cysts are at increased risk for the subsequent development of cholangiocarcinoma.

Congenital Biliary Ectasia

Dilatation of intrahepatic bile ducts may involve either the major intrahepatic radicles (Caroli's disease), the inter- and intralobular ducts (congenital hepatic fibrosis), or both. In Caroli's disease, clinical manifestations include recurrent cholangitis, abscess formation in and around the affected ducts, and, often, brown pigment gallstone formation within portions of ectatic intrahepatic biliary radicles. Ultrasound, MRC, and CT are of great diagnostic value in demonstrating cystic dilatation of the intrahepatic bile ducts. Treatment with ongoing antibiotic therapy is usually undertaken in an effort to limit the frequency and severity of recurrent bouts of cholangitis. Progression to secondary biliary cirrhosis with portal hypertension, extrahepatic biliary obstruction, cholangiocarcinoma, or recurrent episodes of sepsis with hepatic abscess formation is common.

CHOLEDOCHOLITHIASIS

Pathophysiology and Clinical Manifestations

Passage of gallstones into the CBD occurs in ~10–15% of patients with cholelithiasis. The incidence of common duct stones increases with increasing age of the patient, so that up to 25% of elderly patients may have calculi in the common duct at the time of cholecystectomy. Undetected duct stones are left behind in ~1–5% of cholecystectomy patients. The overwhelming majority of bile duct stones are cholesterol stones formed in the gallbladder, which then migrate into the extrahepatic biliary tree through the cystic duct. Primary calculi arising de novo in the ducts are usually brown pigment stones developing in patients with (1) hepatobiliary parasitism or chronic, recurrent cholangitis; (2) congenital anomalies of the bile ducts (especially Caroli's disease); (3) dilated, sclerosed, or strictured ducts; or (4) an *MDR3* (ABCB4) gene defect leading to impaired biliary phospholipids secretion (low phospholipid–associated cholesterol cholelithiasis). Common duct stones may remain asymptomatic for years, may pass spontaneously into the duodenum, or (most often) may present with biliary colic or a complication.

Complications

CHOLANGITIS

Cholangitis may be acute or chronic, and symptoms result from inflammation, which usually is caused by at least partial obstruction to the flow of bile. Bacteria are present on bile culture in ~75% of patients with acute cholangitis early in the symptomatic course. The characteristic presentation of acute cholangitis involves biliary pain, jaundice, and spiking fevers with chills (Charcot's triad). Blood cultures are frequently positive, and leukocytosis is typical. *Nonsuppurative acute cholangitis* is most common and may respond relatively rapidly to supportive measures and to treatment with antibiotics. In *suppurative acute cholangitis*, however, the presence of pus under pressure in a completely obstructed ductal system leads to symptoms of severe toxicity—mental confusion, bacteremia, and septic shock. Response to antibiotics alone in this setting is relatively poor, multiple hepatic abscesses are often present, and the mortality rate approaches 100% unless prompt endoscopic or surgical relief of the obstruction and drainage of infected bile are carried out. Endoscopic management of bacterial cholangitis is as effective as surgical intervention. ERCP with endoscopic sphincterotomy is safe and the preferred initial procedure for both establishing a definitive diagnosis and providing effective therapy.

OBSTRUCTIVE JAUNDICE

Gradual obstruction of the CBD over a period of weeks or months usually leads to initial manifestations of jaundice or pruritus without associated symptoms of biliary colic or cholangitis. Painless jaundice may occur in patients with choledocholithiasis, but is much more characteristic of biliary obstruction secondary to malignancy of the head of the pancreas, bile ducts, or ampulla of Vater.

In patients whose obstruction is secondary to choledocholithiasis, associated chronic calculous cholecystitis is very common, and the gallbladder in this setting may be unable to distend. The absence of a palpable gallbladder in most patients with biliary obstruction from duct stones is the basis for Courvoisier's law, that is, that the presence of a palpably enlarged gallbladder suggests that the biliary obstruction is secondary to an underlying malignancy rather than to calculous disease. Biliary obstruction causes progressive dilatation of the intrahepatic bile ducts as intrabiliary pressures

rise. Hepatic bile flow is suppressed, and reabsorption and regurgitation of conjugated bilirubin into the bloodstream lead to jaundice accompanied by dark urine (bilirubinuria) and light-colored (acholic) stools.

CBD stones should be suspected in any patient with cholecystitis whose serum bilirubin level is >85.5 µmol/L (5 mg/dL). The maximum bilirubin level is seldom >256.5 µmol/L (15.0 mg/dL) in patients with choledocholithiasis unless concomitant hepatic or renal disease or another factor leading to marked hyperbilirubinemia exists. Serum bilirubin levels ≥342.0 µmol/L (20 mg/dL) should suggest the possibility of neoplastic obstruction. The serum alkaline phosphatase level is almost always elevated in biliary obstruction. A rise in alkaline phosphatase often precedes clinical jaundice and may be the only abnormality in routine liver function tests. There may be a two- to tenfold elevation of serum aminotransferases, especially in association with acute obstruction. Following relief of the obstructing process, serum aminotransferase elevations usually return rapidly to normal, while the serum bilirubin level may take 1–2 weeks to return to normal. The alkaline phosphatase level usually falls slowly, lagging behind the decrease in serum bilirubin.

PANCREATITIS

The most common associated entity discovered in patients with nonalcoholic acute pancreatitis is biliary tract disease. Biochemical evidence of pancreatic inflammation complicates acute cholecystitis in 15% of cases and choledocholithiasis in >30%, and the common factor appears to be the passage of gallstones through the common duct. Coexisting pancreatitis should be suspected in patients with symptoms of cholecystitis who develop (1) back pain or pain to the left of the abdominal midline, (2) prolonged vomiting with paralytic ileus, or (3) a pleural effusion, especially on the left side. Surgical treatment of gallstone disease is usually associated with resolution of the pancreatitis.

SECONDARY BILIARY CIRRHOSIS

Secondary biliary cirrhosis may complicate prolonged or intermittent duct obstruction with or without recurrent cholangitis. Although this complication may be seen in patients with choledocholithiasis, it is more common in cases of prolonged obstruction from stricture or neoplasm. Once established, secondary biliary cirrhosis may be progressive even after correction of the obstructing process, and increasingly severe hepatic cirrhosis may lead to portal hypertension or to hepatic failure and death. Prolonged biliary obstruction may also be associated with clinically relevant deficiencies of the fat-soluble vitamins A, D, E, and K.

Diagnosis and Treatment

The diagnosis of choledocholithiasis is usually made by cholangiography **(Table 339-3)**, either preoperatively by endoscopic retrograde cholangiogram (ERC) **(Fig. 339-2C)** or MRCP or intraoperatively at the time of cholecystectomy. As many as 15% of patients undergoing cholecystectomy will prove to have CBD stones. When CBD stones are suspected prior to laparoscopic cholecystectomy, preoperative ERCP with endoscopic papillotomy and stone extraction is the preferred approach. It not only provides stone clearance but also defines the anatomy of the biliary tree in relationship to the cystic duct. CBD stones should be suspected in gallstone patients who have any of the following risk factors: (1) a history of jaundice or pancreatitis, (2) abnormal tests of liver function, and (3) ultrasonographic or MRCP evidence of a dilated CBD or stones in the duct. Alternatively, if intraoperative cholangiography reveals retained stones, postoperative ERCP can be carried out. The need for preoperative ERCP is expected to decrease further as laparoscopic techniques for bile duct exploration improve.

TABLE 339-3

Diagnostic Evaluation of the Bile Ducts

Diagnostic Advantages	Diagnostic Limitations	Contraindications	Complications	Comment
Hepatobiliary Ultrasound				
Rapid Simultaneous scanning of GB, liver, bile ducts, pancreas Accurate identification of dilated bile ducts Not limited by jaundice, pregnancy Guidance for fine-needle biopsy	Bowel gas Massive obesity Ascites Barium Partial bile duct obstruction Poor visualization of distal CBD	None	None	Initial procedure of choice in investigating possible biliary tract obstruction
Computed Tomography				
Simultaneous scanning of GB, liver, bile ducts, pancreas Accurate identification of dilated bile ducts, masses Not limited by jaundice, gas, obesity, ascites High-resolution image Guidance for fine-needle biopsy	Extreme cachexia Movement artifact Ileus Partial bile duct obstruction	Pregnancy	Reaction to iodinated contrast, if used	Indicated for evaluation of hepatic or pancreatic masses Procedure of choice in investigating possible biliary obstruction if diagnostic limitations prevent HBUS
Magnetic Resonance Cholangiopancreatography				
Useful modality for visualizing pancreatic and biliary ducts Has excellent sensitivity for bile duct dilatation, biliary stricture, and intraductal abnormalities Can identify pancreatic duct dilatation or stricture, pancreatic duct stenosis, and pancreas divisum	Cannot offer therapeutic intervention High cost	Claustrophobia Certain metals (iron)	None	
Endoscopic Retrograde Cholangiopancreatography				
Simultaneous pancreatography Best visualization of distal biliary tract Bile or pancreatic cytology Endoscopic sphincterotomy and stone removal Biliary manometry	Gastroduodenal obstruction Roux-en-Ybiliary-entericanastomosis	Pregnancy Acute pancreatitis Severe cardiopulmonary disease	Pancreatitis Cholangitis, sepsis Infected pancreatic pseudocyst Perforation (rare) Hypoxemia, aspiration	Cholangiogram of choice in: Absence of dilated ducts Pancreatic, ampullary or gastroduodenal disease Prior biliary surgery Endoscopic sphincterotomy treatment possibility
Percutaneous Transhepatic Cholangiogram				

Diagnostic Advantages	Diagnostic Limitations	Contraindications	Complications	Comment
Extremely successful when bile ducts dilated Best visualization of proximal biliary tract Bile cytology/culture Percutaneous transhepatic drainage	Nondilated or sclerosed ducts	Pregnancy Uncorrectable coagulopathy Massive ascites Hepatic abscess	Bleeding Hemobilia Bile peritonitis Bacteremia, sepsis	Indicated when ERCP is contraindicated or failed
Endoscopic Ultrasound				
Most sensitive method to detect ampullary stones				Excellent for detecting choledocholithiasis

Abbreviations: CBD, common bile duct; ERCP, endoscopic retrograde cholangiopancreatography; GB, gallbladder; HBUS, hepatobiliary ultrasound.

The widespread use of laparoscopic cholecystectomy and ERCP has decreased the incidence of complicated biliary tract disease and the need for choledocholithotomy and T-tube drainage of the bile ducts. EBS followed by spontaneous passage or stone extraction is the treatment of choice in the management of patients with common duct stones, especially in elderly or poor-risk patients.

TRAUMA, STRICTURES, AND HEMOBILIA

Most benign strictures of the extrahepatic bile ducts result from surgical trauma and occur in about 1 in 500 cholecystectomies. Strictures may present with bile leak or abscess formation in the immediate postoperative period or with biliary obstruction or cholangitis as long as 2 years or more following the inciting trauma. The diagnosis is established by percutaneous or endoscopic cholangiography. Endoscopic brushing of biliary strictures may be helpful in establishing the nature of the lesion and is more accurate than bile cytology alone. When positive exfoliative cytology is obtained, the diagnosis of a neoplastic stricture is established. This procedure is especially important in patients with primary sclerosing cholangitis (PSC) who are predisposed to the development of cholangiocarcinomas. Successful operative correction of non-PSC bile duct strictures by a skillful surgeon with duct-to-bowel anastomosis is usually possible, although mortality rates from surgical complications, recurrent cholangitis, or secondary biliary cirrhosis are high.

Hemobilia may follow traumatic or operative injury to the liver or bile ducts, intraductal rupture of a hepatic abscess or aneurysm of the hepatic artery, biliary or hepatic tumor hemorrhage, or mechanical complications of choledocholithiasis or hepatobiliary parasitism. Diagnostic procedures such as liver biopsy, PTC, and transhepatic biliary drainage catheter placement may also be complicated by hemobilia. Patients often present with a classic triad of biliary pain, obstructive jaundice, and melena or occult blood in the stools. The diagnosis is sometimes made by cholangiographic evidence of blood clot in the biliary tree, but selective angiographic verification may be required. Although minor episodes of hemobilia may resolve without operative intervention, surgical ligation of the bleeding vessel is frequently required.

EXTRINSIC COMPRESSION OF THE BILE DUCTS

Partial or complete biliary obstruction may be produced by extrinsic compression of the ducts. The most common cause of this form of obstructive jaundice is carcinoma of the head of the pancreas. Biliary obstruction may also occur as a complication of either acute or chronic pancreatitis or involvement of lymph nodes in the porta hepatis by lymphoma or metastatic carcinoma. The latter should be distinguished from cholestasis resulting from massive replacement of the liver by tumor.

HEPATOBILIARY PARASITISM

Infestation of the biliary tract by adult helminths or their ova may produce a chronic, recurrent pyogenic cholangitis with or without multiple hepatic abscesses, ductal stones, or biliary obstruction. This condition is relatively rare but does occur in inhabitants of southern China and elsewhere in Southeast Asia. The organisms most commonly involved are trematodes or flukes, including *Clonorchis sinensis*, *Opisthorchis viverrini* or *Opisthorchis felineus*, and *Fasciola hepatica*. The biliary tract also may be involved by intraductal migration of adult *Ascaris lumbricoides* from the duodenum or by intrabiliary rupture of hydatid cysts of the liver produced by *Echinococcus* spp. The diagnosis is made by cholangiography and the presence of characteristic ova on stool examination. When obstruction is present, the treatment of choice is laparotomy under antibiotic coverage, with common duct exploration and a biliary drainage procedure.

SCLEROSING CHOLANGITIS

Primary or idiopathic sclerosing cholangitis (PSC) is characterized by a progressive, inflammatory, sclerosing, and obliterative process affecting the extrahepatic and/or the intrahepatic bile ducts. The disorder occurs up to 90% in association with inflammatory bowel disease, especially ulcerative colitis. It may also be associated with autoimmune pancreatitis; multifocal fibrosclerosis syndromes such as retroperitoneal, mediastinal, and/or periureteral fibrosis; Riedel's struma; or pseudotumor of the orbit.

Immunoglobulin G4 (IgG4)–associated cholangitis is a recently described biliary disease of unknown etiology that presents with biochemical and cholangiographic features indistinguishable from PSC, is often associated with autoimmune pancreatitis and other fibrosing conditions, and is characterized by elevated serum IgG4 and infiltration of IgG4-positive plasma cells in bile ducts and liver tissue. All newly diagnosed PSC patients should have a serum IgG4 level checked. In contrast to PSC, IgG4-associated cholangitis is not associated with inflammatory bowel disease and should be suspected if associated with increased serum IgG4 and unexplained pancreatic disease. Glucocorticoids are regarded as the initial treatment of choice. Relapse is common after steroid withdrawal, especially with proximal strictures. Long-term treatment with glucocorticoids and/or azathioprine may be needed after relapse or for inadequate response **(Chap. 341)**.

Patients with PSC often present with signs and symptoms of chronic or intermittent biliary obstruction: RUQ abdominal pain, pruritus, jaundice, or acute cholangitis. Late in the course, complete biliary obstruction, secondary biliary cirrhosis, hepatic failure, or portal hypertension with bleeding varices may occur. The diagnosis is usually established by finding multifocal, diffusely distributed strictures with intervening segments of normal or dilated ducts, producing a beaded appearance on cholangiography **(Fig. 339-2D)**. The cholangiographic techniques of choice in suspected cases are MRCP and ERCP. When a diagnosis of sclerosing cholangitis has been established, a search for associated diseases, especially for chronic inflammatory bowel disease, should be carried out.

A recent study describes the natural history and outcome for 305 patients of Swedish descent with PSC; 134 (44%) of the patients were asymptomatic at the time of diagnosis and, not surprisingly, had a significantly higher survival rate. The independent predictors of a bad prognosis were advanced age, serum bilirubin concentration, and liver histologic changes. Cholangiocarcinoma was found in 24 patients (8%). Inflammatory bowel disease was closely associated with PSC and had a prevalence of 81% in this study population. The PSC Autoimmune Hepatitis (AIH) Overlap syndrome is characterized by clinical, biochemical, and histological features of AIH and cholangiographic features of PSC.

Small duct PSC is defined by the presence of chronic cholestasis and hepatic histology consistent with PSC but with normal findings on cholangiography. Small duct PSC is found in ~5% of patients with PSC and may represent an earlier stage of PSC associated with a significantly better long-term prognosis. However, such patients may progress to classic PSC and/or end-stage liver disease with consequent necessity of liver transplantation.

In patients with AIDS, cholangiopancreatography may demonstrate a broad range of biliary tract changes as well as pancreatic duct obstruction and occasionally pancreatitis **(Chap. 197)**. Further, biliary tract lesions in AIDS include infection and cholangiopancreatographic changes similar to those of PSC. Changes noted include: (1) diffuse involvement of intrahepatic bile ducts alone, (2) involvement of both intra- and extrahepatic bile ducts, (3) ampullary stenosis, (4) stricture of the intrapancreatic portion of the CBD, and (5) pancreatic duct involvement. Associated infectious organisms include *Cryptosporidium*, *Mycobacterium avium-intracellulare*, cytomegalovirus, *Microsporidia*, and *Isospora*. ERCP sphincterotomy can provide significant pain reduction in patients with AIDS-associated papillary stenosis.

TREATMENT

TREATMENT

Sclerosing Cholangitis

Therapy with cholestyramine may help control symptoms of pruritus, and antibiotics are useful when cholangitis complicates the clinical picture. Vitamin D and calcium supplementation may help prevent the loss of bone mass frequently seen in patients with chronic cholestasis. Glucocorticoids, methotrexate, and cyclosporine have not been shown to be efficacious in PSC. UDCA in high dosage (28–30 mg/kg) was not effective in prolonging survival. In cases where high-grade biliary obstruction (dominant strictures) has occurred, balloon dilatation or stenting may be appropriate. Only rarely is surgical intervention indicated. Efforts at biliary-enteric anastomosis or stent placement may, however, be complicated by recurrent cholangitis and further progression of the stenosing process. The prognosis is unfavorable, with a median survival of 9–12 years following the diagnosis, regardless of therapy. Four variables (age, serum bilirubin level, histologic stage, and splenomegaly) predict survival in patients with PSC and serve as the basis for a risk score. PSC is one of the most common indications for liver transplantation.

FURTHER READING

Baron TH et al: Interventional approaches to gallbladder disease. N Engl J Med 373:357, 2015. [PubMed: 26200981]

Lindor K et al: American College of Gastroenterology (ACG) Guidelines: Primary Sclerosing Cholangitis. Hepatology 51:660, 2010. [PubMed: 20101749]

Ryl JK et al: Clinical features of acute acalculous cholecystitis. J Clin Gastroenterol 36:166, 2003. [PubMed: 12544202]

Strasberg S: Clinical practice. Acute calculous cholecystitis. N Engl J Med 358:2804, 2008. [PubMed: 18579815]

Chapter 340: Approach to the Patient with Pancreatic Disease

Darwin L. Conwell; Norton J. Greenberger; Peter A. Banks

FIGURE 340-1

GENERAL CONSIDERATIONS

As emphasized in **Chap. 341**, the etiologies as well as clinical manifestations of pancreatitis are quite varied. Although it is well-appreciated that pancreatitis is frequently secondary to biliary tract disease and alcohol abuse, it can also be caused by drugs, genetic mutations, trauma, and viral infections and is associated with metabolic and connective tissue disorders. In ~30% of patients with acute pancreatitis and 25–40% of patients with chronic pancreatitis, the etiology initially can be obscure.

The incidence of acute pancreatitis is about 5–35 per 100,000 new cases per year worldwide, with a mortality rate of about 3%. The incidence of chronic pancreatitis is about 4–8 per 100,000 persons with a prevalence of 26–42 cases per 100,000. The number of patients admitted to the hospital who suffer with both acute and chronic pancreatitis in the United States is largely increasing and is now estimated to be 274,119 for acute pancreatitis and 19,724 for chronic pancreatitis. Acute pancreatitis is the most common gastrointestinal diagnosis requiring hospitalization in the United States. Acute and chronic pancreatic disease costs an estimated $3 billion annually in health care expenditures. These numbers may underestimate the true incidence and prevalence, because non–alcohol-induced pancreatitis has been largely ignored. At autopsy, the prevalence of chronic pancreatitis ranges from 0.04 to 5%.

The diagnosis of acute pancreatitis is generally clearly defined based on a combination of laboratory, imaging, and clinical symptoms. The diagnosis of chronic pancreatitis, especially in mild disease, is hampered by the relative inaccessibility of the pancreas to direct examination and the nonspecificity of the abdominal pain associated with chronic pancreatitis. Many patients with chronic pancreatitis do not have elevated blood amylase or lipase levels. Some patients with chronic pancreatitis develop signs and symptoms of pancreatic exocrine insufficiency (PEI), and thus, objective evidence for pancreatic disease can be demonstrated. However, there is a very large reservoir of pancreatic exocrine function. More than 90% of the pancreas must be damaged before maldigestion of fat and protein is manifested. Noninvasive, indirect tests of pancreatic exocrine function (fecal elastase) are much more likely to give abnormal results in patients with obvious advanced pancreatic disease (i.e., pancreatic calcification, steatorrhea, or diabetes mellitus) than in patients with occult disease. Invasive, direct tests of pancreatic secretory function (secretin tests) are the most sensitive and specific tests to detect early chronic pancreatic disease when imaging is equivocal or normal.

TESTS USEFUL IN THE DIAGNOSIS OF PANCREATIC DISEASE

Several tests have proved of value in the evaluation of pancreatic disease. Examples of specific tests and their usefulness in the diagnosis of acute and chronic pancreatitis are summarized in **Table 340-1** and **Fig. 340-1**. At some institutions, pancreatic function tests are available and performed if the diagnosis of chronic pancreatic disease remains a possibility after noninvasive tests (ultrasound, computed tomography [CT], magnetic resonance cholangiopancreatography [MRCP]) or

invasive tests (endoscopic retrograde cholangiopancreatography [ERCP], endoscopic ultrasonography [EUS]) have given normal or inconclusive results. In this regard, tests using *direct* stimulation of the pancreas with secretin are the most sensitive.

TABLE 340-1

Tests Useful in the Diagnosis of Acute and Chronic Pancreatitis and Pancreatic Tumors

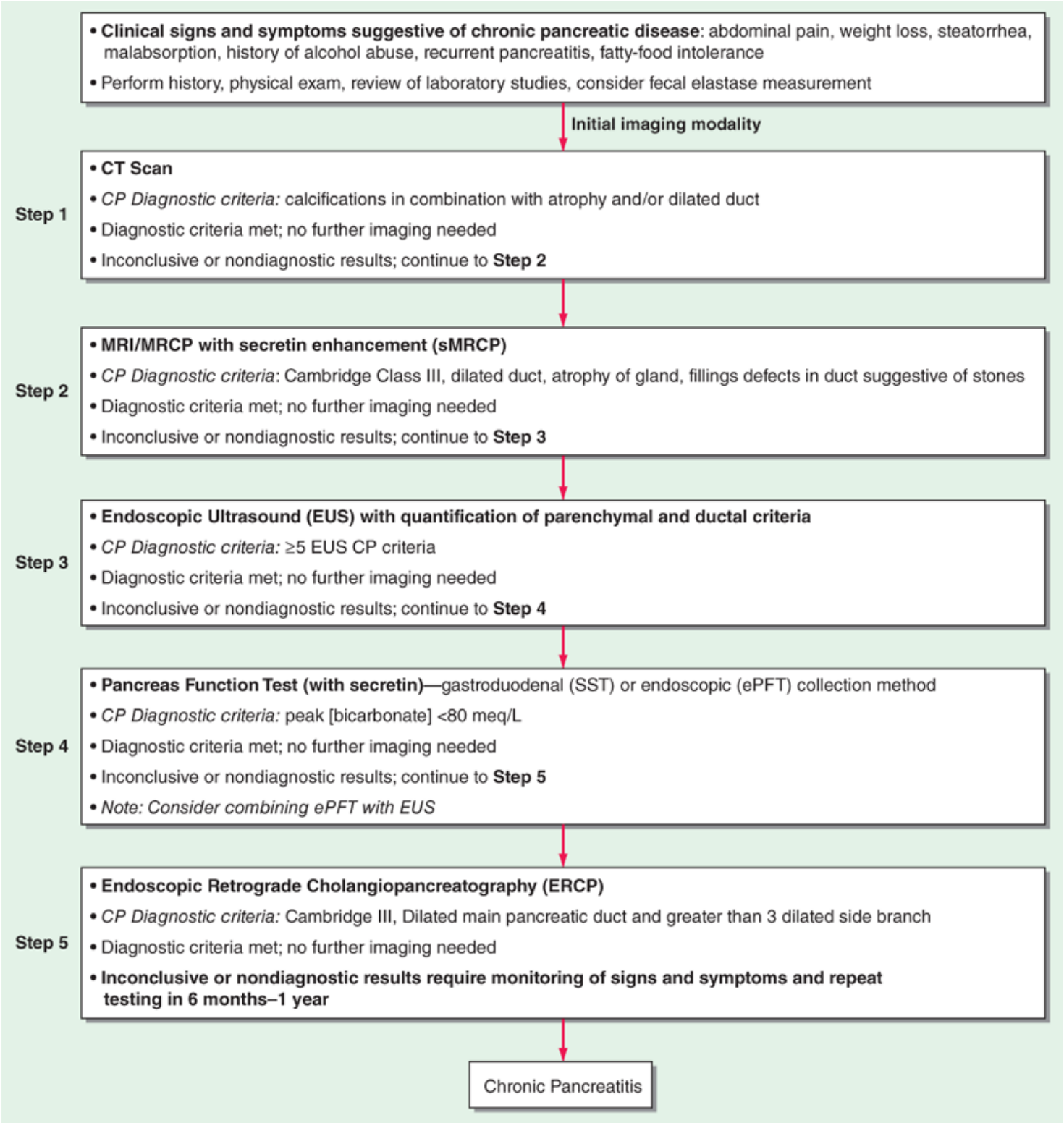
Test	Principle	Comment
Pancreatic Enzymes in Body Fluids		
Serum lipase	Pancreatic inflammation leads to increased serum enzyme levels	Enzyme measurement of choice for diagnosis of acute pancreatitis
Amylase		
1. Serum	Pancreatic inflammation leads to increased serum enzyme levels	Simple; reliable if test results are three times the upper limit of normal (3× ULN)
2. Urine	Renal clearance of amylase is increased in acute pancreatitis	Infrequently used
3. Ascitic fluid	Disruption of gland or main pancreatic duct leads to increased amylase concentration	Can help establish source of ascites; false positives occur with intestinal obstruction and perforated ulcer; can also measure lipase
4. Pleural fluid	Exudative pleural effusion with pancreatitis	False positives occur with carcinoma of the lung and esophageal perforation
Studies Pertaining to Pancreatic Structure		
Radiologic and radionuclide tests		
1. Plain film of the abdomen	Can be abnormal in acute and chronic pancreatitis	Infrequently used
2. Upper gastrointestinal x-rays		Infrequently used
3. Ultrasonography (US)	Can provide information on edema, inflammation, calcification, pseudocysts, and mass lesions	Simple, noninvasive; sequential studies quite feasible; useful in diagnosis of gallstones; pancreas visualization limited by interference from overlying bowel gas

Test	Principle	Comment
4. Computed tomography (CT) scan	Permits detailed visualization of pancreas and surrounding structures, pancreatic fluid collection, pseudocyst; assessment of necrosis or interstitial disease	Useful in the diagnosis of pancreatic calcification, dilated pancreatic ducts, and pancreatic tumors; may not be able to distinguish between inflammatory and neoplastic mass lesions
5. Magnetic resonance cholangiopancreatography (MRCP)	Three-dimensional imaging has been used to produce very good images of the pancreatic-biliary ductal system by a noninvasive technique	Has replaced ERCP as a diagnostic test; noninvasive
6. Endoscopic ultrasonography (EUS)	High-frequency transducer used with EUS can produce very high-resolution images and depict changes in the pancreatic duct and parenchyma with great detail	Can be used to assess gallstones, chronic pancreatitis, and pancreatic carcinoma
7. Endoscopic retrograde cholangiopancreatography (ERCP)	Cannulation of pancreatic and common bile duct permits visualization of pancreatic-biliary ductal system	Primarily a therapeutic procedure; invasive
Pancreatic biopsy with US or CT guidance	Percutaneous aspiration biopsy of mass-forming lesions of the pancreas	High diagnostic yield; laparotomy avoided; can be done with EUS for the evaluation of chronic pancreatitis, autoimmune pancreatitis, and pancreatic carcinoma
Tests of Exocrine Pancreatic Function		
Direct stimulation of the pancreas with analysis of duodenal contents		
1. Secretin test	Secretin leads to increased output of pancreatic juice and HCO_3^-; pancreatic secretory response is related to the functional mass of pancreatic tissue	Sensitive enough to detect occult disease; involves duodenal intubation and fluoroscopic placement of gastroduodenal tube; poorly defined normal enzyme response; overlap in chronic pancreatitis; large secretory reserve capacity of the pancreas; currently done at only a few medical centers
2. Endoscopic secretin test	Replaces need for tube placement duodenum	Sensitive enough to detect occult disease; high negative predictive value; avoids intubation and fluoroscopy; requires sedation

Test	Principle	Comment
Measurement of intraluminal digestion products		
1. Quantitative stool fat determination	Lack of lipolytic enzymes brings about impaired fat digestion	Reliable reference standard for defining severity of malabsorption; does not distinguish between maldigestion and malabsorption
Measurement of pancreatic enzymes in feces		
1. Fecal elastase	Pancreatic secretion of proteolytic enzymes; not degraded in intestine	Diagnostic accuracy best if value is <100 µg/g performed on a solid stool

FIGURE 340-1

A stepwise diagnostic approach to the patient with suspected chronic pancreatitis (CP). Endoscopic ultrasonography (EUS) and magnetic resonance cholangiopancreatography (sMRCP/MRCP) are appropriate diagnostic alternatives to endoscopic retrograde cholangiopancreatography (ERCP). CT, computed tomography.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pancreatic Enzymes in Body Fluids

The serum amylase and lipase levels are widely used as screening tests for acute pancreatitis in the patient with acute abdominal pain or back pain. Values greater than three times the upper limit of normal (3× ULN) in combination with epigastric pain strongly suggest the diagnosis of acute pancreatitis. In acute pancreatitis, the serum amylase and lipase are usually elevated within 24 h of onset and remain so for 3–7 days. Levels usually return to normal within 7 days unless there is

pancreatic ductal disruption, ductal obstruction, or pseudocyst formation. Approximately 85% of patients with acute pancreatitis have a threefold or greater elevated serum lipase and amylase levels. The values may be normal if (1) there is a delay (2–5 days) before blood samples are obtained, (2) the underlying disorder is chronic pancreatitis rather than acute pancreatitis, or (3) hypertriglyceridemia is present. Patients with hypertriglyceridemia and pancreatitis have been found to have spuriously low levels of amylase and perhaps lipase activity. In the absence of objective evidence of pancreatitis by abdominal ultrasound, CT scan, MRCP, or EUS, mild to moderate elevations of amylase and/or lipase are not helpful in making a diagnosis of chronic pancreatitis.

The serum amylase can be elevated in other conditions **(Table 340-2)**, in part because the enzyme is found in many organs. In addition to the pancreas and salivary glands, small quantities of amylase are found in the tissues of the fallopian tubes, lung, thyroid, and tonsils and can be produced by various tumors (carcinomas of the lung, esophagus, breast, and ovary). Isoamylase determinations do not accurately distinguish elevated blood amylase levels from pancreatic or nonpancreatic sources. In patients with unexplained hyperamylasemia, measurement of macroamylase can avoid numerous tests in patients with this rare disorder.

TABLE 340-2

Causes of Hyperamylasemia and Hyperamylasuria

Pancreatic Disease
I. Pancreatitis A. Acute B. Chronic: ductal obstruction C. Complications of pancreatitis 1. Pancreatic pseudocyst 2. Ascites caused by pancreatic duct disruption 3. Pancreatic necrosis II. Pancreatic trauma III. Pancreatic carcinoma
Non-Pancreatic Disorders
I. Renal insufficiency II. Salivary gland lesions A. Mumps B. Calculus C. Irradiation sialadenitis D. Maxillofacial surgery III. “Tumor” hyperamylasemia A. Carcinoma of the lung, esophagus, breast, or ovary IV. Macroamylasemia V. Burns VI. Diabetic ketoacidosis VII. Pregnancy VIII. Renal transplantation IX. Cerebral trauma X. Drugs: opiates
Other Abdominal Disorders
I. Biliary tract disease: cholecystitis, choledocholithiasis II. Intraabdominal disease A. Perforated or penetrating peptic ulcer B. Intestinal obstruction or inflammation C. Ruptured ectopic pregnancy D. Peritonitis E. Aortic aneurysm F. Postoperative hyperamylasemia

Elevation of ascitic fluid amylase occurs in acute pancreatitis as well as in (1) ascites due to disruption of the main pancreatic duct or a leaking pseudocyst and (2) other abdominal disorders that simulate pancreatitis (e.g., intestinal obstruction, intestinal infarction, or perforated peptic ulcer). Elevation of pleural fluid amylase can occur in acute pancreatitis, chronic pancreatitis, carcinoma of the lung, and esophageal perforation. Lipase is the single best enzyme to measure for the diagnosis of acute pancreatitis. No single blood test is reliable for the diagnosis of acute pancreatitis in patients with renal failure. Pancreatic enzyme elevations are usually less than three times the upper limit of normal. Determining whether a patient with renal failure and abdominal pain has pancreatitis remains a difficult clinical problem. One study found that serum amylase levels were elevated in patients with renal dysfunction only when creatinine clearance was <0.8 mL/s (<50 mL/min). In such patients, the serum amylase level was invariably <500 IU/L in the absence of objective evidence of acute pancreatitis. In that study, serum lipase and trypsin levels paralleled serum amylase values. With these limitations in mind, the recommended screening test for acute pancreatitis in renal disease is serum lipase.

Studies Pertaining to Pancreatic Structure

RADIOLOGIC TESTS

Plain films of the abdomen, which once provided useful information in patients with acute and chronic pancreatitis, have been superseded by other more detailed imaging procedures (ultrasound, EUS, CT, MRCP).

Ultrasonography can provide important information in patients with acute pancreatitis, chronic pancreatitis, pseudocysts, and pancreatic carcinoma. Sonographic changes can indicate the presence of edema, inflammation, and calcification (not obvious on plain films of the abdomen), as well as pseudocysts, mass lesions, and gallstones. In acute pancreatitis, the pancreas is characteristically enlarged. In pancreatic pseudocyst, the usual appearance is primarily that of a smooth, round fluid collection. Pancreatic carcinoma distorts the usual landmarks, and mass lesions >3.0 cm are usually detected as localized, solid lesions. US is often the initial investigation for most patients with suspected pancreatic disease. However, obesity and excess intestinal bowel gas can interfere with pancreatic imaging by US studies.

Computed tomography is the best imaging study for initial evaluation of a suspected pancreatic disorder and for the assessment of complications of acute and chronic pancreatitis. It is especially useful in the detection of pancreatic and peripancreatic acute fluid collections, fluid-containing lesions such as pseudocysts, walled-off necrosis, calcium deposits **(see Chap. 341, Figs. 341-1, 341-2, and 341-4)**, and pancreatic neoplasms. Acute pancreatitis is characterized by (1) enlargement of the pancreatic outline, (2) distortion of the pancreatic contour, and/or (3) a pancreatic fluid that has a different attenuation coefficient than normal pancreas. Oral, water-soluble contrast agents are used to opacify the stomach and duodenum during CT scans; this strategy permits more precise delineation of various organs as well as mass lesions. Dynamic CT (using rapid IV administration of contrast) is useful in estimating the extent of pancreatic necrosis and in predicting morbidity and mortality. CT provides clear images much more rapidly and essentially negates artifact caused by patient movement. If acute pancreatitis is confirmed with serology and physical examination findings, CT scan in the first 3 days is not recommended to avoid overuse and minimize costs. The major benefit of CT in acute pancreatitis is the diagnosis of pancreatic necrosis in patients not responding to conservative management within 72 h.

Endoscopic ultrasonography produces high-resolution images of the pancreatic parenchyma and pancreatic duct with a transducer fixed to an endoscope that can be directed onto the surface of the pancreas through the stomach or duodenum. EUS and MRCP have replaced ERCP for diagnostic purposes. EUS allows one to obtain information about the pancreatic duct as well as the parenchyma and has few procedure-related complications associated with it, in contrast to the 5–10% of post-ERCP pancreatitis observed. EUS is also helpful in detecting common bile duct stones in acute pancreatitis. Pancreatic masses can also be biopsied via EUS in cases with suspected pancreas cancer, and one can deliver nerve-blocking agents through EUS fine-needle injection in patients suffering from pancreatic pain from chronic pancreatitis or cancer. EUS has been studied as a diagnostic modality for chronic pancreatitis. Criteria for abnormalities on EUS in severe chronic pancreatic

disease have been developed. There is general agreement that the presence of five or more of the nine criteria listed in Table 340-3 is highly predictive of chronic pancreatitis. Recent studies comparing EUS and ERCP to the secretin test in patients with unexplained abdominal pain suspected of having chronic pancreatitis show similar diagnostic accuracy in detecting early changes of chronic pancreatitis. The exact role of EUS versus CT, ERCP, or function testing in the early diagnosis of chronic pancreatitis has yet to be clearly defined.

TABLE 340-3

Endoscopic Ultrasonographic Criteria for Chronic Pancreatitis (Total Criteria = 9)

Ductal	Parenchymal
Stones	Echogenic strands
Hyperechoic main duct margins	Echogenic foci
Main duct irregularity	Lobular contour
Main duct dilatation	Cyst
Visible side branches	

MRI and MRCP are now being used to view the bile ducts, pancreatic duct, and the pancreas parenchyma in both acute pancreatitis and chronic pancreatitis. For diagnostic imaging in chronic pancreatitis, non-breath-holding and three-dimensional turbo spin-echo techniques are being used to produce superb MRCP images. The main pancreatic duct and common bile duct can be seen well, but there is still a question as to whether changes can be detected consistently in the secondary ducts. The secondary ducts are not visualized in a normal pancreas. Secretin-enhanced MRCP is emerging as a method to better evaluate ductal changes. In anteroposterior imaging, T2 imaging of fluid collections can differentiate necrotic debris from fluid in suspected walled-off necrosis, and T1 imaging can diagnose hemorrhage in suspected pseudoaneurysm rupture.

As mentioned, EUS and MRCP have largely replaced ERCP in the diagnostic evaluation of pancreatic disease. As these techniques become more refined, especially with the administration of secretin, they may well be the diagnostic tests of choice to evaluate the pancreatic duct. ERCP is primarily of therapeutic value after CT, EUS, or MRCP has detected abnormalities requiring invasive endoscopic treatment. ERCP can also be helpful at clarification of equivocal findings discovered with other imaging techniques (see Chap. 341, Fig. 341-1). Pancreatic carcinoma is characterized by stenosis or obstruction of either the pancreatic duct or the common bile duct; both ductal systems are often abnormal (double-duct sign). In chronic pancreatitis, ERCP abnormalities in the main pancreatic duct and side branches have been outlined by the Cambridge classification. The presence of ductal stenosis and irregularity can make it difficult to distinguish chronic pancreatitis from carcinoma. It is important to be aware that ERCP changes interpreted as indicating chronic pancreatitis actually may be due to the effects of aging on the pancreatic duct or sequelae of a recent attack of acute pancreatitis. Although aging may cause impressive ductal alterations, it does not affect the results of pancreatic function tests (i.e., the secretin test). Elevated serum amylase levels after ERCP have been reported in the majority of patients, and clinical pancreatitis in 5–10% of patients. Recent data suggest that pancreatic duct stenting and rectal indomethacin can decrease the incidence of ERCP-induced pancreatitis. ERCP should not be done for diagnostic purposes and should especially be avoided in high-risk patients.

Percutaneous aspiration biopsy or a trucut biopsy of a pancreatic mass often distinguishes a pancreatic inflammatory mass from a pancreatic neoplasm.

TESTS OF EXOCRINE PANCREATIC FUNCTION

Pancreatic function tests (Table 340-1) can be divided into the following:

1. *Direct stimulation of the pancreas* by IV infusion of secretin followed by collection and measurement of duodenal contents

   The secretin test, used to detect diffuse pancreatic disease, is based on the physiologic principle that the pancreatic secretory response is directly related to the functional mass of pancreatic tissue. In the standard assay, secretin is given IV in a dose of 0.2 μg/kg of synthetic human secretin as a bolus. Normal values for the standard secretin test are (1) volume output >2 mL/kg per hour, (2) bicarbonate (HCO_3^-) concentration >80 mmol/L, and (3) HCO_3^- output >10 mmol/L in 1 h.

   The most reproducible measurement, giving the highest level of discrimination between normal subjects and patients with chronic PEI, appears to be the maximal bicarbonate concentration. A cutoff point below 80 mmol/L is considered abnormal and suggestive of abnormal secretory function that is most commonly observed in early chronic pancreatitis.

   There may be a dissociation between the results of the secretin test and other tests of absorptive function. For example, patients with chronic pancreatitis often have abnormally low outputs of HCO_3^- after secretin but have normal fecal fat excretion. Thus the secretin test measures the secretory capacity of ductular epithelium, whereas fecal fat excretion indirectly reflects intraluminal lipolytic activity. Steatorrhea does not occur until intraluminal levels of lipase are markedly reduced, underscoring the fact that only small amounts of enzymes are necessary for intraluminal digestive activities. It must be emphasized that an abnormal secretin test result suggests that pancreatic ductal secretory function is abnormal. This is commonly observed in chronic pancreatitis.

2. *Measurement of fecal pancreatic enzymes* such as elastase

   Measurement of *intraluminal digestion products* (i.e., undigested muscle fibers, stool fat, and fecal nitrogen) is discussed in **Chap. 318**. The amount of human elastase in stool reflects the pancreatic output of this proteolytic enzyme. Decreased elastase-1 activity (FE-1) in stool is an excellent test to detect severe PEI in patients with chronic pancreatitis and cystic fibrosis. FE-1 levels >200 μg/g are normal, levels of 100–200 μg/g are considered mild, and levels <100 μg/g are severe for PEI. Although the test is simple and noninvasive, it can give false-positive results and has a low sensitivity. Fecal levels <50 μg/g are definitive for PEI provided that the stool specimen is solid.

Tests useful in the diagnosis of exocrine pancreatic insufficiency and the differential diagnosis of malabsorption are also discussed in Chaps. 318 and 341.

FURTHER READING

Conwell DL, Wu BU: Chronic pancreatitis: Making the diagnosis. Clin Gastroenterol Hepatol 10:1088, 2012. [PubMed: 22642958]

Conwell DL et al: American Pancreatic Association practice guidelines in chronic pancreatitis: Evidence-based report on diagnostic guidelines. Pancreas 43:1143, 2014. [PubMed: 25333398]

Hart PA et al: Endoscopic pancreas fluid collection: Methods and relevance for clinical care and translational science. Am J Gastroenterol 111:1258, 2016. [PubMed: 27481304]

Yadav D, Lowenfels AB: The epidemiology of pancreatitis and pancreatic cancer. Gastroenterology 144:1252, 2013. [PubMed: 23622135]

Chapter 341: Acute and Chronic Pancreatitis

Darwin L. Conwell; Peter A. Banks; Norton J. Greenberger

BIOCHEMISTRY AND PHYSIOLOGY OF PANCREATIC EXOCRINE SECRETION

GENERAL CONSIDERATIONS

The pancreas secretes 1500–3000 mL of isosmotic alkaline (pH >8) fluid per day containing about 20 enzymes. The pancreatic secretions provide the enzymes and bicarbonate needed to affect the major digestive activity of the gastrointestinal tract and provide an optimal pH for the function of these enzymes.

REGULATION OF PANCREATIC SECRETION

The exocrine pancreas is influenced by intimately interacting hormonal and neural systems. *Gastric acid* is the stimulus for the release of secretin from the duodenal mucosa (S cells), which stimulates the secretion of water and electrolytes from pancreatic ductal cells. Release of cholecystokinin (CCK) from the duodenal and proximal jejunal mucosa (Ito cells) is largely triggered by long-chain fatty acids, essential amino acids (tryptophan, phenylalanine, valine, methionine), and gastric acid itself. CCK evokes an enzyme-rich secretion from acinar cells in the pancreas. The *parasympathetic nervous system* (via the vagus nerve) exerts significant control over pancreatic secretion. Secretion evoked by secretin and CCK depends on permissive roles of vagal afferent and efferent pathways. This is particularly true for enzyme secretion, whereas water and bicarbonate secretions are heavily dependent on the hormonal effects of secretin and to a lesser extent CCK. Also, vagal stimulation affects the release of vasoactive intestinal peptide (VIP), a secretin agonist. Pancreatic exocrine secretion is also influenced by inhibitory neuropeptides such as somatostatin, pancreatic polypeptide, peptide YY, neuropeptide Y, enkephalin, pancreastatin, calcitonin gene–related peptides, glucagon, and galanin. Although pancreatic polypeptide and peptide YY may act primarily on nerves outside the pancreas, somatostatin acts at multiple sites. Nitric oxide (NO) is also an important neurotransmitter.

WATER AND ELECTROLYTE SECRETION

Bicarbonate is the ion of primary physiologic importance within pancreatic secretion. The ductal cells secrete bicarbonate predominantly derived from plasma (93%) more than from intracellular metabolism (7%). Bicarbonate enters the duct lumen through the sodium bicarbonate cotransporter with depolarization caused by chloride efflux through the cystic fibrosis transmembrane conductance regulator (CFTR). Secretin and VIP bind at the basolateral surface and cause an increase in secondary messenger intracellular cyclic AMP, and act on the apical surface of the ductal cells opening the CFTR in promoting secretion. CCK, acting as a neuromodulator, markedly potentiates the stimulatory effects of secretin. Acetylcholine also plays an important role in ductal cell secretion. Intraluminal bicarbonate secreted from the ductal cells helps neutralize gastric acid and creates the appropriate pH for the activity of pancreatic enzymes and bile salts on ingested food.

ENZYME SECRETION

The acinar cell is highly compartmentalized and is concerned with the secretion of pancreatic enzymes. Proteins synthesized by the rough endoplasmic reticulum are processed in the Golgi and then targeted to the appropriate site, whether that be zymogen granules, lysosomes, or other cell compartments. The zymogen granules migrate to the apical region of the acinar cell awaiting the appropriate neural or hormonal stimulatory response. The pancreas secretes amylolytic, lipolytic, and proteolytic enzymes into the duct lumen. *Amylolytic enzymes*, such as amylase, hydrolyze starch to oligosaccharides and to the disaccharide maltose. The *lipolytic enzymes* include lipase, phospholipase A_2, and cholesterol esterase. Bile salts inhibit lipase in isolation, but colipase, another constituent of pancreatic secretion, binds to lipase and prevents this inhibition. Bile salts activate phospholipase A and cholesterol esterase. *Proteolytic enzymes* include endopeptidases (trypsin, chymotrypsin), which act on internal peptide bonds of proteins and polypeptides; exopeptidases (carboxypeptidases, aminopeptidases), which act on the free carboxyl- and amino-terminal ends of peptides, respectively; and elastase. The proteolytic enzymes are secreted as inactive zymogen precursors. Ribonucleases (deoxyribonucleases, ribonuclease) are also secreted. *Enterokinase*, an enzyme found in the duodenal mucosa, cleaves the lysine-isoleucine bond of trypsinogen to form trypsin. Trypsin then activates the other proteolytic zymogens and phospholipase A_2 in a cascade phenomenon. All pancreatic enzymes have pH optima in the alkaline range. The nervous system initiates pancreatic enzyme secretion. The neurologic stimulation is cholinergic, involving extrinsic innervation by the vagus nerve and subsequent innervation by intrapancreatic cholinergic nerves. The stimulatory neurotransmitters are acetylcholine and gastrin-releasing peptides. These neurotransmitters activate calcium-dependent secondary messenger systems, resulting in the release of zymogens into the pancreas

duct. VIP is present in intrapancreatic nerves and potentiates the effect of acetylcholine. In contrast to other species, there are no CCK receptors on acinar cells in humans. CCK in physiologic concentrations stimulates pancreatic secretion by stimulating afferent vagal and intrapancreatic nerves.

AUTOPROTECTION OF THE PANCREAS

Autodigestion of the pancreas is prevented by (1) the packaging of pancreatic proteases in precursor (proenzyme) form, (2) intracellular calcium homeostasis (low intracellular calcium in the cytosol of the acinar cell promotes the destruction of spontaneously activated trypsin), (3) acid-base balance, and (4) the synthesis of protective protease inhibitors (pancreatic secretory trypsin inhibitor [PSTI] or SPINK1), which can bind and inactivate about 20% of intracellular trypsin activity. Chymotrypsin C can also lyse and inactivate trypsin. These protease inhibitors are found in the acinar cell, the pancreatic secretions, and the α_1- and α_2-globulin fractions of plasma. Loss of any of these four protective mechanisms leads to premature enzyme activation, autodigestion, and acute pancreatitis.

ENTEROPANCREATIC AXIS AND FEEDBACK INHIBITION

Pancreatic enzyme secretion is controlled, at least in part, by a negative feedback mechanism induced by the presence of active serine proteases in the duodenum. To illustrate, perfusion of the duodenal lumen with phenylalanine (stimulates early digestion) causes a prompt increase in plasma CCK levels as well as increased secretion of chymotrypsin and other pancreatic enzymes. However, simultaneous perfusion with trypsin (stimulates late digestion) blunts both responses. Conversely, perfusion of the duodenal lumen with protease inhibitors actually leads to enzyme hypersecretion. The available evidence supports the concept that the duodenum contains a peptide called *CCK-releasing factor* (CCK-RF) that is involved in stimulating CCK release. It appears that serine proteases inhibit pancreatic secretion by inactivating a CCK-releasing peptide in the lumen of the small intestine. Thus, the integrative result of both bicarbonate and enzyme secretion depends on a feedback process for both bicarbonate and pancreatic enzymes. Acidification of the duodenum releases secretin, which stimulates vagal and other neural pathways to activate pancreatic duct cells, which secrete bicarbonate. This bicarbonate then neutralizes the duodenal acid, and the feedback loop is completed. Dietary proteins bind proteases, thereby leading to an increase in free CCK-RF. CCK is then released into the blood in physiologic concentrations, acting primarily through the neural pathways (vagal-vagal). This leads to acetylcholine-mediated pancreatic enzyme secretion. Proteases continue to be secreted from the pancreas until the protein within the duodenum is digested. At this point, pancreatic protease secretion is reduced to basic levels, thus completing this step in the feedback process.

ACUTE PANCREATITIS

GENERAL CONSIDERATIONS

Recent U.S. estimates from the Nationwide Inpatient Sample report that acute pancreatitis is the most common inpatient principal gastrointestinal diagnosis. The incidence of acute pancreatitis also varies in different countries and depends on cause (e.g., alcohol, gallstones, metabolic factors, drugs [Table 341-1]). The annual incidence ranges from 13 to 45 cases per 100,000 persons. Acute pancreatitis results in >250,000 hospitalizations per year. The median length of hospital stay is 4 days, with a median hospital cost of $6096 and a mortality of 1%. The estimated cost annually approaches $2.6 billion. Hospitalization rates increase with age, which are 88% higher among blacks, and are higher among males than females. The age-adjusted rate of hospital discharges with an acute pancreatitis diagnosis increased 62% between 1988 and 2004. From 2000 to 2009, the rate increased 30%. Thus, acute pancreatitis is increasing and is a significant burden on health care costs and resource utilization.

TABLE 341-1

Causes of Acute Pancreatitis

Common Causes
Gallstones (including microlithiasis)
Alcohol (acute and chronic alcoholism)
Hypertriglyceridemia
Endoscopic retrograde cholangiopancreatography (ERCP), especially after biliary manometry
Drugs (azathioprine, 6-mercaptopurine, sulfonamides, estrogens, tetracycline, valproic acid, anti-HIV medications, 5-aminosalicylic acid [5-ASA])
Trauma (especially blunt abdominal trauma)
Postoperative (abdominal and nonabdominal operations)
Uncommon Causes
Vascular causes and vasculitis (ischemic-hypoperfusion states after cardiac surgery)
Connective tissue disorders and thrombotic thrombocytopenic purpura (TTP)
Cancer of the pancreas
Hypercalcemia
Periampullary diverticulum
Pancreas divisum
Hereditary pancreatitis
Cystic fibrosis
Renal failure
Infections (mumps, coxsackievirus, cytomegalovirus, echovirus, parasites)
Autoimmune (e.g., type 1 and type 2)
Causes to Consider in Patients with Recurrent Bouts of Acute Pancreatitis without an Obvious Etiology
Occult disease of the biliary tree or pancreatic ducts, especially microlithiasis, biliary sludge
Drugs
Alcohol abuse
Metabolic: Hypertriglyceridemia, hypercalcemia
Anatomic: Pancreas divisum
Pancreatic cancer
Intraductal papillary mucinous neoplasm (IPMN)

Hereditary pancreatitis
Cystic fibrosis
Autoimmune
Idiopathic

ETIOLOGY AND PATHOGENESIS

There are many causes of acute pancreatitis (Table 341-1), but the mechanisms by which these conditions trigger pancreatic inflammation have not been fully elucidated. Gallstones and alcohol account for 80–90% of the acute pancreatitis cases in the United States. Gallstones continue to be the leading cause of acute pancreatitis in most series (30–60%). The risk of acute pancreatitis in patients with at least one gallstone <5 mm in diameter is fourfold greater than that in patients with larger stones. Alcohol is the second most common cause, responsible for 15–30% of cases in the United States. The incidence of pancreatitis in alcoholics is surprisingly low (5/100,000), indicating that in addition to the amount of alcohol ingested, other factors affect a person's susceptibility to pancreatic injury such as cigarette smoking. Acute pancreatitis occurs in 5–10% of patients following endoscopic retrograde cholangiopancreatography (ERCP). Use of a prophylactic pancreatic duct stent and rectal nonsteroidal anti-inflammatory drugs (NSAIDs, indomethicin) has been shown to reduce pancreatitis after ERCP. Risk factors for post-ERCP pancreatitis include minor papilla sphincterotomy, sphincter of Oddi dysfunction, prior history of post-ERCP pancreatitis, age <60 years, >2 contrast injections into the pancreatic duct, and endoscopic trainee involvement.

Hypertriglyceridemia is the cause of acute pancreatitis in 1.3–3.8% of cases; serum triglyceride levels are usually >1000 mg/dL. Most patients with hypertriglyceridemia, when subsequently examined, show evidence of an underlying derangement in lipid metabolism, probably unrelated to pancreatitis. Such patients are prone to recurrent episodes of pancreatitis. Any factor (e.g., drugs or alcohol) that causes an abrupt increase in serum triglycerides can precipitate a bout of acute pancreatitis. Patients with a deficiency of apolipoprotein CII have an increased incidence of pancreatitis; apolipoprotein CII activates lipoprotein lipase, which is important in clearing chylomicrons from the bloodstream. Patients with diabetes mellitus who have developed ketoacidosis and patients who are on certain medications such as oral contraceptives may also develop high triglyceride levels. Approximately 0.1–2% of cases of acute pancreatitis are drug related. Drugs cause pancreatitis either by a hypersensitivity reaction or by the generation of a toxic metabolite, although in some cases, it is not clear which of these mechanisms is operative (Table 341-1).

Pathologically, acute pancreatitis varies from *interstitial pancreatitis* (pancreas blood supply maintained), which is generally self-limited to *necrotizing pancreatitis* (pancreas blood supply interrupted), in which the extent of necrosis may correlate with the severity of the attack and its systemic complications. Autodigestion is a currently accepted pathogenic theory; according to this theory, pancreatitis results when proteolytic enzymes (e.g., trypsinogen, chymotrypsinogen, proelastase, and lipolytic enzymes such as phospholipase A_2) are activated in the pancreas acinar cell rather than in the intestinal lumen. A number of factors (e.g., endotoxins, exotoxins, viral infections, ischemia, oxidative stress, lysosomal calcium, and direct trauma) are believed to facilitate premature activation of trypsin. Activated proteolytic enzymes, especially trypsin, not only digest pancreatic and peripancreatic tissues but also can activate other enzymes, such as elastase and phospholipase A_2. Spontaneous activation of trypsin also can occur.

ACTIVATION OF PANCREATIC ENZYMES IN THE PATHOGENESIS OF ACUTE PANCREATITIS

Several recent studies have suggested that pancreatitis is a disease that evolves in three phases. The *initial phase* is characterized by intrapancreatic digestive enzyme activation and acinar cell injury. Trypsin activation appears to be mediated by lysosomal hydrolases such as cathepsin B that become colocalized with digestive enzymes in intracellular organelles; it is currently believed that acinar cell injury is the consequence of trypsin activation. The *second phase* of pancreatitis involves the activation, chemoattraction, and sequestration of leukocytes and macrophages in the pancreas, resulting in an enhanced intrapancreatic inflammatory reaction. Neutrophil depletion induced by prior administration of an antineutrophil serum has been shown to reduce the severity of experimentally induced pancreatitis. There is also evidence to support the concept that neutrophils can activate trypsinogen. Thus, intrapancreatic acinar cell activation of trypsinogen could be a two-step process (i.e., an early neutrophil-independent and a later neutrophil-dependent phase). The *third phase* of pancreatitis is due to the effects of activated proteolytic enzymes and cytokines, released by the inflamed pancreas, on distant organs. Activated proteolytic enzymes, especially trypsin, not only digest pancreatic and peripancreatic tissues but also activate other enzymes such as elastase and phospholipase A_2. The active enzymes and cytokines then digest cellular membranes and cause proteolysis, edema, interstitial hemorrhage, vascular damage, coagulation necrosis, fat necrosis, and parenchymal cell necrosis. Cellular injury and death result in the liberation of bradykinin peptides, vasoactive substances, and histamine that can produce vasodilation, increased vascular permeability, and edema with profound effects on many organs. The systemic inflammatory response syndrome (SIRS) and acute respiratory distress syndrome (ARDS), as well as multiorgan failure, may occur as a result of this cascade of local and distant effects.

A number of genetic factors can increase the susceptibility and/or modify the severity of pancreatic injury in acute pancreatitis, recurrent pancreatitis, and chronic pancreatitis. All of the major genetic susceptibility factors center on the control of trypsin activity within the pancreatic acinar cell, in part because they were identified as candidate genes linked to intrapancreatic trypsin control. Five genetic variants have been identified as being associated with susceptibility to pancreatitis. The genes that have been identified include (1) cationic trypsinogen gene (*PRSS1*), (2) pancreatic secretory trypsin inhibitor (*SPINK1*), (3) the cystic fibrosis transmembrane conductance regulator gene (*CFTR*), (4) the chymotrypsin C gene (*CTRC*), and (5) the calcium-sensing receptor (*CASR*). Investigations of other genetic variants are currently under way, and new genes will be added to this list in the future. Multiple medical, ethical, and psychological issues arise when these genes are discovered, and referral to genetic counselors is recommended.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Abdominal Pain

Abdominal pain is the major symptom of acute pancreatitis. Pain may vary from a mild discomfort to severe, constant, and incapacitating distress. Characteristically, the pain, which is steady and boring in character, is located in the epigastrium and periumbilical region, and may radiate to the back, chest, flanks, and lower abdomen. Nausea, vomiting, and abdominal distention due to gastric and intestinal hypomotility and chemical peritonitis are also frequent complaints.

Physical examination frequently reveals a distressed and anxious patient. Low-grade fever, tachycardia, and hypotension are fairly common. Shock is not unusual and may result from (1) hypovolemia secondary to exudation of blood and plasma proteins into the retroperitoneal space; (2) increased formation and release of kinin peptides, which cause vasodilation and increased vascular permeability; and (3) systemic effects of proteolytic and lipolytic enzymes released into the circulation. Jaundice occurs infrequently; when present, it usually is due to edema of the head of the pancreas with compression of the intrapancreatic portion of the common bile duct or passage of a biliary stone or sludge. Erythematous skin nodules due to subcutaneous fat necrosis may rarely occur. In 10–20% of patients, there are pulmonary findings, including basilar rales, atelectasis, and pleural effusion, the latter most frequently left sided. Abdominal tenderness and muscle rigidity are present to a variable degree, but compared with the intense pain, these signs may be less impressive. Bowel sounds are usually diminished or absent. An enlarged pancreas from acute fluid collection, walled off necrosis, or a pseudocyst may be palpable in the upper abdomen later in the course of the disease (i.e., 4–6 weeks). A faint blue discoloration around the umbilicus (Cullen's sign) may occur as the result of hemoperitoneum, and a blue-red-purple or green-brown discoloration of the flanks (Turner's sign) reflects tissue catabolism of hemoglobin from severe necrotizing pancreatitis with hemorrhage.

LABORATORY DATA

Serum amylase and lipase values threefold or more above normal virtually clinch the diagnosis if gut perforation, ischemia, and infarction are excluded. Serum lipase is the preferred test. However, it should be noted that there is no correlation between the severity of pancreatitis and the degree of serum lipase and amylase elevations. After 3–7 days, even with continuing evidence of pancreatitis, total serum amylase values tend to return toward normal. However, pancreatic lipase levels may remain elevated for 7–14 days. It should be recognized that amylase elevations in serum and urine occur in many conditions other than pancreatitis **(see Chap. 340, Table 340-2)**. Importantly, patients with *acidemia* (arterial pH ≤7.32) may have spurious elevations in serum amylase. This finding explains why patients with diabetic ketoacidosis may have marked elevations in serum amylase without any other evidence of acute pancreatitis. Serum lipase activity increases in parallel with amylase activity and is more specific than amylase. A serum lipase measurement can be instrumental in differentiating a pancreatic or nonpancreatic cause for hyperamylasemia. *Leukocytosis* (15,000–20,000 leukocytes/μL) occurs frequently. Patients with more severe disease may show hemoconcentration with hematocrit values >44% and/or prerenal azotemia with a blood urea nitrogen (BUN) level >22 mg/dL resulting from loss of plasma into the retroperitoneal space and peritoneal cavity.

Hemoconcentration may be the harbinger of more severe disease (i.e., pancreatic necrosis), whereas azotemia is a significant risk factor for mortality. *Hyperglycemia* is common and is due to multiple factors, including decreased insulin release, increased glucagon release, and an increased output of adrenal glucocorticoids and catecholamines. *Hypocalcemia* occurs in ~25% of patients, and its pathogenesis is incompletely understood. Although earlier studies suggested that the response of the parathyroid gland to a decrease in serum calcium is impaired, subsequent observations have failed to confirm this phenomenon. Intraperitoneal saponification of calcium by fatty acids in areas of fat necrosis occurs occasionally, with large amounts (up to 6.0 g) dissolved or suspended in ascitic fluid. Such "soap formation" may also be significant in patients with pancreatitis, mild hypocalcemia, and little or no obvious ascites. *Hyperbilirubinemia* (serum bilirubin >4.0 mg/dL) occurs in ~10% of patients. However, jaundice is transient, and serum bilirubin levels return to normal in 4–7 days. Serum alkaline phosphatase and aspartate aminotransferase levels are also transiently elevated, and they parallel serum bilirubin values and may point to gallbladder-related disease or inflammation in the pancreatic head. *Hypertriglyceridemia* occurs in 5–10% of patients, and serum amylase levels in these individuals are often spuriously normal **(Chap. 340)**. Approximately 5–10% of patients have *hypoxemia* (arterial PO_2 ≤60 mm Hg), which may herald the onset of ARDS. Finally, the electrocardiogram is occasionally abnormal in acute pancreatitis with ST-segment and T-wave abnormalities simulating myocardial ischemia.

An abdominal ultrasound is recommended in the emergency ward as the initial diagnostic imaging modality and is most useful to evaluate for gallstone disease and the pancreatic head.

The Revised Atlanta criteria have clearly outlined the morphologic features of acute pancreatitis on computed tomography (CT) scan as follows: (1) interstitial pancreatitis, (2) necrotizing pancreatitis, (3) acute pancreatic fluid collection, (4) pancreatic pseudocyst, (5) acute necrotic collection (ANC), and (6) walled-off necrosis (WON) **(Table 341-2 and Fig. 341-1)**. **Radiologic studies useful in the diagnosis of acute pancreatitis are discussed in Chap. 340 and listed in Table 340-1**.

TABLE 341-2

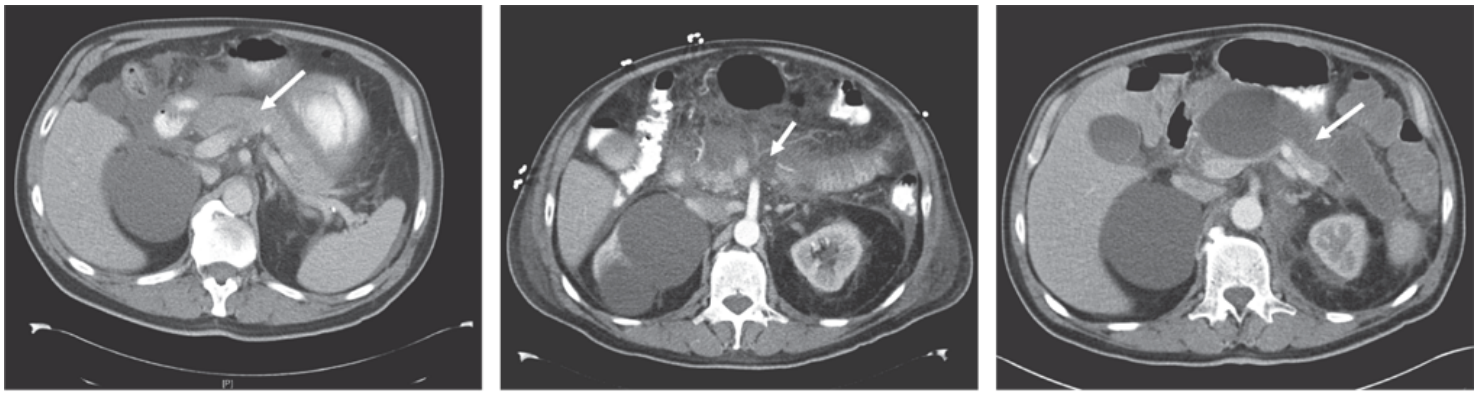
Revised Atlanta Definitions of Morphologic Features of Acute Pancreatitis

Morphologic Feature	Definition	Computed Tomography Criteria
Interstitial pancreatitis	Acute inflammation of the pancreatic parenchyma and peripancreatic tissues, but without recognizable tissue necrosis	Pancreatic parenchyma enhancement by IV contrast agent No findings of peripancreatic necrosis
Necrotizing pancreatitis	Inflammation associated with pancreatic parenchymal necrosis and/or peripancreatic necrosis	Lack of pancreatic parenchymal enhancement by IV contrast agent and/or presence of findings of peripancreatic necrosis (see below—ANC and WON)
Acute pancreatic fluid collection	Peripancreatic fluid associated with interstitial edematous pancreatitis with no associated peripancreatic necrosis. This term applies only to areas of peripancreatic fluid seen within the first 4 weeks after onset of interstitial edematous pancreatitis and without the features of a pseudocyst.	Occurs in the setting of interstitial edematous pancreatitis Homogeneous collection with fluid density Confined by normal peripancreatic fascial planes No definable wall encapsulating the collection Adjacent to pancreas (no intrapancreatic extension)
Pancreatic pseudocyst	An encapsulated collection of fluid with a well-defined inflammatory wall usually outside the pancreas with minimal or no necrosis. This entity usually occurs >4 weeks after onset of interstitial edematous pancreatitis.	Well circumscribed, usually round or oval Homogeneous fluid density No nonliquid component Well-defined wall; that is, completely encapsulated Maturation usually requires >4 weeks after onset of acute pancreatitis; occurs after interstitial edematous pancreatitis
Acute necrotic collection (ANC)	A collection containing variable amounts of both fluid and necrosis associated with necrotizing pancreatitis; the necrosis can involve the pancreatic parenchyma and/or the peripancreatic tissues.	Occurs only in the setting of acute necrotizing pancreatitis Heterogeneous and nonliquid density of varying degrees in different locations (some appear homogeneous early in their course) No definable wall encapsulating the collection Location—intrapancreatic and/or extrapancreatic
Walled-off necrosis (WON)	A mature, encapsulated collection of pancreatic and/or peripancreatic necrosis that has developed a well-defined inflammatory wall. WON usually occurs >4 weeks after onset of necrotizing pancreatitis.	Heterogeneous with liquid and nonliquid density with varying degrees of loculations (some may appear homogeneous) Well-defined wall; that is, completely encapsulated Location—intrapancreatic and/or extrapancreatic Maturation usually requires 4 weeks after onset of acute necrotizing pancreatitis

Source: Data from P Banks et al: Gut 62:102, 2013.

FIGURE 341-1

Acute pancreatitis: computed tomography (CT) evolution. ***A.*** Contrast-enhanced CT scan of the abdomen performed on admission for a patient with clinical and biochemical parameters suggestive of acute pancreatitis. Note the abnormal enhancement of the pancreatic parenchyma (*arrow*) suggestive of interstitial pancreatitis. ***B.*** Contrast-enhanced CT scan of the abdomen performed on the same patient 6 days later for persistent fever and systemic inflammatory response syndrome. The pancreas now demonstrates significant areas of nonenhancement consistent with development of necrosis, particularly in the body and neck region (*arrow*). Note that an early CT scan obtained within the first 48 h of hospitalization may underestimate or miss necrosis. ***C.*** Contrast-enhanced CT scan of the abdomen performed on the same patient 2 months after the initial episode of acute pancreatitis. CT now demonstrates evidence of a fluid collection consistent with walled-off pancreatic necrosis (*arrow*). *(Courtesy of Dr. KJ Mortele, Brigham and Women's Hospital; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS

Any severe acute pain in the abdomen or back should suggest the possibility of acute pancreatitis. The diagnosis is established by two of the following three criteria: (1) typical abdominal pain in the epigastrium that may radiate to the back, (2) threefold or greater elevation in serum lipase and/or amylase, and (3) confirmatory findings of acute pancreatitis on cross-sectional abdominal imaging. Patients also have associated nausea, emesis, fever, tachycardia, and abnormal findings on abdominal examination. Laboratory studies may reveal leukocytosis, hypocalcemia, and hyperglycemia. Although not required for diagnosis, markers of severity may include hemoconcentration (hematocrit >44%), admission azotemia (BUN >22 mg/dL), SIRS, and signs of organ failure (Table 341-3).

TABLE 341-3

Severe Acute Pancreatitis

Risk Factors for Severity
• Age >60 years
• Obesity, BMI >30
• Comorbid disease (Charlson Comorbidity Index)
Markers of Severity at Admission or within 24 h
• SIRS—defined by presence of 2 or more criteria: • Core temperature <36° or >38°C • Heart rate >90 beats/min • Respirations >20/min or Pco_2 <32 mmHg • White blood cell count >12,000/μL, <4000/μL, or 10% bands • APACHE II • Hemoconcentration (hematocrit >44%) • Admission BUN (>22 mg/dL) • BISAP Score   • (B) BUN >25 mg/dL   • (I) Impaired mental status   • (S) SIRS: ≥2 of 4 present   • (A) Age >60 years   • (P) Pleural effusion • Organ failure (Modified Marshall Score) • Cardiovascular: systolic BP <90 mm Hg, heart rate >130 beats/min • Pulmonary: Pao_2 <60 mm Hg • Renal: serum creatinine >2.0 mg %
Markers of Severity during Hospitalization
• Persistent organ failure
• Pancreatic necrosis

Abbreviations: APACHE II, Acute Physiology and Chronic Health Evaluation II; BMI, body mass index; BISAP, Bedside Index of Severity in Acute Pancreatitis; BP, blood pressure; BUN, blood urea nitrogen; SIRS, systemic inflammatory response syndrome.

The *differential diagnosis* should include the following disorders: (1) perforated viscus, especially peptic ulcer; (2) acute cholecystitis and biliary colic; (3) acute intestinal obstruction; (4) mesenteric vascular occlusion; (5) renal colic; (6) inferior myocardial infarction; (7) dissecting aortic aneurysm; (8) connective tissue disorders with vasculitis; (9) pneumonia; and (10) diabetic ketoacidosis. It may be difficult to differentiate acute cholecystitis from acute pancreatitis, because an elevated serum amylase may be found in both disorders. Pain of biliary tract origin is more right sided or epigastric than periumbilical or left upper quadrant and can be more severe; ileus is usually absent. Ultrasound is helpful in establishing the diagnosis of cholelithiasis and cholecystitis. Intestinal obstruction due to mechanical factors can be differentiated from pancreatitis by the history of crescendo-decrescendo pain, findings on abdominal examination, and CT of the abdomen showing changes characteristic of mechanical obstruction. Acute mesenteric vascular occlusion is usually suspected in elderly debilitated patients with brisk leukocytosis, abdominal distention, and bloody diarrhea, confirmed by CT or magnetic resonance angiography. Vasculitides secondary to systemic lupus erythematosus and polyarteritis nodosa may be confused with pancreatitis, especially because pancreatitis may develop as a complication of these diseases. Diabetic ketoacidosis is often accompanied by abdominal pain and elevated total serum amylase levels, thus closely mimicking acute pancreatitis. However, the serum lipase level is not elevated in diabetic ketoacidosis.

CLINICAL COURSE, DEFINITIONS, AND CLASSIFICATIONS

The Revised Atlanta Criteria (1) defines phases of acute pancreatitis, (2) outlines severity of acute pancreatitis, and (3) clarifies imaging definitions as outlined below.

Phases of Acute Pancreatitis

Two phases of acute pancreatitis have been defined, early (<2 weeks) and late (>2 weeks), which primarily describes the hospital course of the disease. In the *early phase* of acute pancreatitis, which lasts 1–2 weeks, severity is defined by clinical parameters rather than morphologic findings. Most patients exhibit SIRS, and if this persists, patients are predisposed to organ failure. Three organ systems should be assessed to define organ failure: respiratory, cardiovascular, and renal. Organ failure is defined as a score of 2 or more for one of these three organ systems using the modified Marshall scoring system. Persistent organ failure (>48 h) is the most important clinical finding in regard to severity of the acute pancreatitis episode. Organ failure that affects more than one organ is considered multisystem organ failure. CT imaging is usually not needed or recommended during the first 48 h of admission in acute pancreatitis.

The *late phase* is characterized by a protracted course of illness and may require imaging to evaluate for local complications. The important clinical parameter of severity, as in the early phase, is persistent organ failure. These patients may require supportive measures such as renal dialysis, ventilator support, or need for supplemental nutrition via the nasojejunal or parenteral route. The radiographic feature of greatest importance to recognize in this phase is the development of necrotizing pancreatitis on CT imaging. Necrosis generally prolongs hospitalization and, if infected, may require operative, endoscopic, or percutaneous intervention.

Severity of Acute Pancreatitis

Three severity classifications have also been defined: mild, moderately severe, and severe. *Mild acute pancreatitis* is without local complications or organ failure. Most patients with interstitial acute pancreatitis have mild pancreatitis. In mild acute pancreatitis, the disease is self-limited and subsides spontaneously, usually within 3–7 days after treatment is instituted. Oral intake can be resumed if the patient is hungry, has normal bowel function, and is without nausea and vomiting. Typically, a clear or full liquid diet has been recommended for the initial meal; however, a low-fat solid diet is a reasonable choice following recovery from mild acute pancreatitis.

Moderately severe acute pancreatitis is characterized by transient organ failure (resolves in <48 h) or local or systemic complications in the absence of persistent organ failure. These patients may or may not have necrosis, but may develop a local complication such as a fluid collection that requires a prolonged hospitalization >1 week.

Severe acute pancreatitis is characterized by persistent organ failure (>48 h). Organ failure can be single or multiple. A CT scan or magnetic resonance imaging (MRI) should be obtained to assess for necrosis and/or complications. If a local complication is encountered, management is dictated by clinical symptoms, evidence of infection, maturity of fluid collection, and clinical stability of the patient. Prophylactic antibiotics are not recommended.

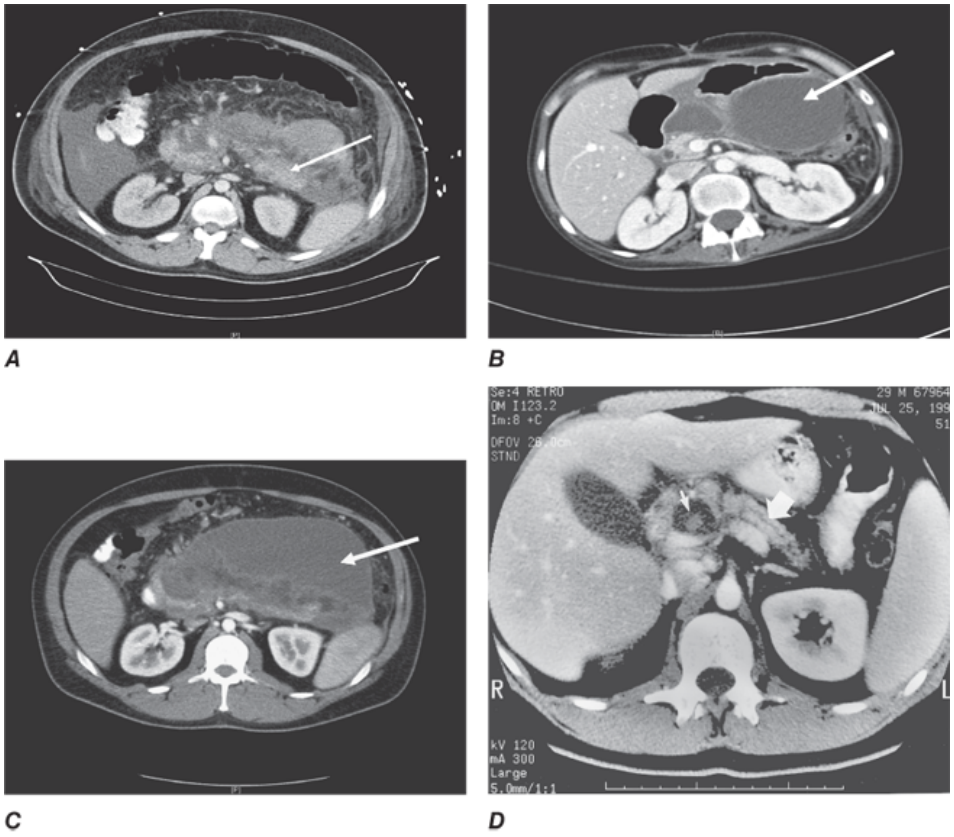
Imaging in Acute Pancreatitis

Two types of pancreatitis are recognized on imaging as *interstitial* or *necrotizing* based on pancreatic perfusion. CT imaging is best evaluated 3–5 days into hospitalization when patients are not responding to supportive care to look for local complications such as necrosis. Recent studies report the overutilization of CT imaging in acute pancreatitis and its inability to be better than clinical judgment in the early days of acute pancreatitis management. The Revised Atlanta criteria also outline the terminology for local complications and fluid collections along with a CT imaging template to guide reporting of findings. Local morphologic features are summarized in Table 341-1. *Interstitial pancreatitis* occurs in 90–95% of admissions for acute pancreatitis and is characterized by diffuse gland enlargement, homogenous contrast enhancement, and mild inflammatory changes or peripancreatic stranding. Symptoms generally resolve with a week of hospitalization. *Necrotizing pancreatitis* occurs in 5–10% of acute pancreatitis admissions and does not evolve until several days of hospitalization. It is characterized by lack of pancreatic parenchymal enhancement by intravenous contrast agent and/or presence of findings of peripancreatic necrosis. According to the Revised Atlanta criteria, the natural history of pancreatic and peripancreatic necrosis is variable because it may remain solid or liquefy, remain sterile or become infected, and persist or disappear over time. CT identification of local complications, particularly necrosis, is critical in patients who are not responding to therapy because patients with infected and sterile necrosis are at greatest risk of mortality **(Figs. 341-1B, 341-2, and 341-3)**. The median prevalence of organ failure is 54% in necrotizing pancreatitis. The prevalence of organ failure is perhaps slightly higher in infected versus sterile necrosis. With single-organ system failure, the mortality is 3–10% but increases to 47% with multisystem organ failure.

FIGURE 341-2

A. Acute necrotizing pancreatitis: computed tomography (CT) scan. Contrast-enhanced CT scan showing acute pancreatitis with necrosis. *Arrow* shows partially enhancing body/tail of pancreas surrounded by fluid with decreased enhancement in the neck/body of the pancreas. ***B.*** Acute fluid collection: CT scan. Contrast-enhanced CT scan showing fluid collection in the retroperitoneum (*arrow*) compressing the air-filled stomach arising from the pancreas in a patient with asparaginase-induced acute necrotizing pancreatitis. ***C.*** Walled-off pancreatic necrosis: CT scan. CT scan showing marked walled-off necrosis of the pancreas and peripancreatic area (*arrow*) in a patient with necrotizing pancreatitis. Addendum: In past years, both of these CT findings (Figs. 341-2B and 341-2C) would have been misinterpreted as pseudocysts. ***D.*** Spiral CT showing a pseudocyst (*small arrow*) with a

pseudoaneurysm (*light area in pseudocyst*). Note the demonstration of the main pancreatic duct (*big arrow*), even though this duct is minimally dilated by endoscopic retrograde cholangiopancreatography. *(A, B, C, courtesy of Dr. KJ Mortele, Brigham and Women's Hospital; D, courtesy of Dr. PR Ros, Brigham and Women's Hospital; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 341-3

A. Pancreaticopleural fistula: pancreatic duct leak on endoscopic retrograde cholangiopancreatography. Pancreatic duct leak (*arrow*) demonstrated at the time of retrograde pancreatogram in a patient with acute exacerbation of alcohol-induced acute or chronic pancreatitis. ***B.*** Pancreaticopleural fistula: computed tomography (CT) scan. Contrast-enhanced CT scan (coronal view) with *arrows* showing fistula tract from pancreatic duct disruption in the pancreatic pleural fistula. ***C.*** Pancreaticopleural fistula: chest x-ray. Large pleural effusion in the left hemithorax from a disrupted pancreatic duct. Analysis of pleural fluid revealed elevated amylase concentration. *(Courtesy of Dr. KJ Mortele, Brigham and Women's Hospital; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ACUTE PANCREATITIS MANAGEMENT

We will briefly outline the management of patients with acute pancreatitis from the time of diagnosis in the emergency ward to ongoing hospital admission and, finally, to time of discharge, highlighting salient features based on severity and complications. It is important to note that 85–90% of

cases of acute pancreatitis are self-limited and subside spontaneously, usually within 3–7 days after initiation of treatment, and do not exhibit organ failure or local complications.

The management of acute pancreatitis begins in the emergency ward. After a diagnosis has been confirmed, aggressive fluid resuscitation is initiated, intravenous analgesics are administered, severity is assessed, and a search for etiologies that may impact acute care is begun. Patients who do not respond to aggressive fluid resuscitation in the emergency ward should be considered for admission to a step-down or intensive care unit for aggressive fluid resuscitation, hemodynamic monitoring, and management of necrosis or organ failure.

Fluid Resuscitation and Monitoring Response to Therapy

The most important treatment intervention for acute pancreatitis is safe, aggressive intravenous fluid resuscitation. The patient is made NPO to rest the pancreas and is given intravenous narcotic analgesics to control abdominal pain and supplemental oxygen (2 L) via nasal cannula.

Intravenous fluids of lactated Ringer's or normal saline are initially bolused at 15–20 mL/kg (1050–1400 mL), followed by 2–3 mL/kg per hour (200–250 mL/h), to maintain urine output >0.5 mL/kg per hour. Serial bedside evaluations are required every 6–8 h to assess vital signs, oxygen saturation, and change in physical examination to optimize fluid resuscitation. Lactated Ringer's solution has been shown to decrease systemic inflammation (lower CRP levels from admission) and may be a better crystalloid than normal saline. A *targeted resuscitation strategy* with measurement of hematocrit and BUN every 8–12 h is recommended to ensure adequacy of fluid resuscitation and monitor response to therapy, noting less aggressive resuscitation strategy may be needed in milder forms of pancreatitis. A rising BUN during hospitalization is not only associated with inadequate hydration but also higher in-hospital mortality.

A decrease in hematocrit and BUN during the first 12–24 h is strong evidence that sufficient fluids are being administered. Serial measurements and bedside assessment for fluid overload are continued, and fluid rates are maintained at the current rate. Adjustments in fluid resuscitation may be required in patients with cardiac, pulmonary, or renal disease. A rise in hematocrit or BUN during serial measurement should be treated with a repeat volume challenge with a 2-L crystalloid bolus followed by increasing the fluid rate by 1.5 mg/kg per hour. If the BUN or hematocrit fails to respond (i.e., remains elevated or does not decrease) to this bolus challenge and increase in fluid rate, consideration of transfer to an intensive care unit is strongly recommended for hemodynamic monitoring.

Assessment of Severity and Hospital Triage

Severity of acute pancreatitis should be determined in the emergency ward to assist in patient triage to a regular hospital ward or step-down unit or direct admission to an intensive care unit. The Bedside Index of Severity in Acute Pancreatitis (BISAP) incorporates five clinical and laboratory parameters obtained within the first 24 h of hospitalization (Table 341-3)—BUN >25 mg/dL, impaired mental status (Glasgow coma score <15), SIRS, age >60 years, and pleural effusion on radiography—that can be useful in assessing severity. Presence of three or more of these factors was associated with substantially increased risk for in-hospital mortality among patients with acute pancreatitis. In addition, an elevated hematocrit >44% and admission BUN >22 mg/dL are also associated with more severe acute pancreatitis. Incorporating these indices with the overall patient response to initial fluid resuscitation in the emergency ward can be useful at triaging patients to the appropriate hospital acute care setting.

In general, patients with lower BISAP scores, hematocrits, and admission BUNs tend to respond to initial management and are triaged to a regular hospital ward for ongoing care. If SIRS is not present at 24 h, the patient is unlikely to develop organ failure or necrosis. Therefore, patients with persistent SIRS at 24 h or underlying comorbid illnesses (e.g., chronic obstructive pulmonary disease, congestive heart failure) should be considered for a step-down unit setting if available. Patients with higher BISAP scores and elevations in hematocrit and admission BUN that do not respond to initial fluid resuscitation and exhibit evidence of respiratory failure, hypotension, or organ failure should be considered for direct admission to an intensive care unit.

Special Considerations Based on Etiology

A careful history, review of medications, selected laboratory studies (liver profile, serum triglycerides, serum calcium), and an abdominal ultrasound are recommended in the emergency ward to assess for etiologies that may impact acute management. An abdominal ultrasound is the initial imaging modality of choice and will evaluate the gallbladder and common duct and assess the pancreatic head.

GALLSTONE PANCREATITIS

Patients with evidence of ascending cholangitis (rising white blood cell count, increasing liver enzymes) should undergo ERCP within 24–48 h of admission. Patients with gallstone pancreatitis are at increased risk of recurrence, and consideration should be given to performing a cholecystectomy during the same admission or within 4–6 weeks of discharge. An alternative for patients who are not surgical candidates would be to perform an endoscopic biliary sphincterotomy before discharge.

HYPERTRIGLYCERIDEMIA

Serum triglycerides >1000 mg/dL are associated with acute pancreatitis. Initial therapy may include insulin, heparin, or plasmapheresis. Outpatient therapies include control of diabetes if present, administration of lipid-lowering agents, weight loss, and avoidance of drugs that elevate lipid levels.

Other potential etiologies that may impact acute hospital care include *hypercalcemia*, *autoimmune pancreatitis* (AIP), *post-ERCP pancreatitis*, and *drug-induced pancreatitis*. Treatment of hyperparathyroidism or malignancy is effective at reducing serum calcium. AIP is responsive to glucocorticoid administration. Pancreatic duct stenting and rectal indomethacin administration are effective at decreasing pancreatitis after ERCP. Drugs that cause pancreatitis should be discontinued. Multiple drugs have been implicated, but only about 30 have been challenged (Class 1A) and found to be causative.

Nutritional Therapy

A low-fat solid diet can be administered to subjects with mild acute pancreatitis after the abdominal pain has resolved. Enteral nutrition should be considered 2–3 days after admission in subjects with more severe pancreatitis instead of total parenteral nutrition (TPN). Enteral feeding maintains gut barrier integrity, limits bacterial translocation, is less expensive, and has fewer complications than TPN. The choice of gastric versus nasojejunal enteral feeding is currently under investigation.

Management of Local Complications

Patients exhibiting signs of clinical deterioration despite aggressive fluid resuscitation and hemodynamic monitoring should be assessed for local complications, which may include necrosis, pseudocyst formation, pancreas duct disruption, peripancreatic vascular complications, and extrapancreatic infections **(Table 341-4)**. A multidisciplinary team approach is recommended including gastroenterology, surgery, interventional radiology, and intensive care specialists, and consideration should also be made for transfer to a pancreas center.

TABLE 341-4

Complications of Acute Pancreatitis

Local
Necrosis
Sterile
Infected
Walled-off necrosis
Pancreatic fluid collections
Pancreatic pseudocyst
Disruption of main pancreatic duct or secondary branches
Pancreatic ascites
Involvement of contiguous organs by necrotizing pancreatitis
Thrombosis of blood vessels (splenic vein, portal vein)
Pancreatic enteric fistula
Bowel infarction
Obstructive jaundice
Systemic
Pulmonary
Pleural effusion
Atelectasis
Mediastinal fluid
Pneumonitis
Acute respiratory distress syndrome
Hypoxemia (unrecognized)
Cardiovascular
Hypotension
Hypovolemia
Nonspecific ST-T changes in electrocardiogram simulating myocardial infarction
Pericardial effusion
Hematologic
Disseminated intravascular coagulation
Gastrointestinal hemorrhage
Peptic ulcer disease
Erosive gastritis
Hemorrhagic pancreatic necrosis with erosion into major blood vessels
Portal vein thrombosis, splenic vein thrombosis, variceal hemorrhage
Renal
Oliguria (<300 mL/d)
Azotemia
Renal artery and/or renal vein thrombosis
Acute tubular necrosis
Metabolic
Hyperglycemia
Hypertriglyceridemia
Hypocalcemia
Encephalopathy
Sudden blindness (Purtscher’s retinopathy)
Central nervous system
Psychosis
Fat emboli
Fat necrosis
Subcutaneous tissues (erythematous nodules)

Bone

Miscellaneous (mediastinum, pleura, nervous system)

NECROSIS

The management of necrosis requires a multidisciplinary team approach. The benefits of percutaneous aspiration of necrosis with Gram stain and culture should be considered or discussed if there are ongoing signs of possible pancreatic infection such as sustained leukocytosis, fever, or organ failure. There is currently no role for *prophylactic antibiotics* in necrotizing pancreatitis. It is reasonable to start broad-spectrum antibiotics in a patient who appears septic while awaiting the results of Gram stain and cultures. If cultures are negative, the antibiotics should be discontinued to minimize the risk of developing opportunistic or fungal superinfection. Repeated fine-needle aspiration and Gram stain with culture of pancreatic necrosis may be done every 5–7 days in the presence of persistent fever. Repeated CT or MRI imaging should also be considered with any change in clinical course to monitor for complications (e.g., thromboses, hemorrhage, abdominal compartment syndrome).

In general, *sterile necrosis* is most often managed conservatively unless complications arise. Once a diagnosis of *infected necrosis* is established and an organism identified, targeted antibiotics should be instituted. Pancreatic debridement (necrosectomy) should be considered for definitive management of *infected necrosis*, but clinical decisions are generally influenced by response to antibiotic treatment and overall clinical condition. Symptomatic local complications as outlined in the Revised Atlanta criteria may require definitive therapy.

A step-up approach (percutaneous or endoscopic transgastric drainage followed, if necessary, by open necrosectomy) has been successfully reported by some pancreatic centers. One-third of the patients successfully treated with the step-up approach did not require major abdominal surgery. A randomized trial reported advantages to an initial endoscopic approach compared to an initial surgical necrosectomy approach in select patients requiring intervention for symptomatic WON. Taken together, a more conservative approach to the management of infected pancreatic necrosis has evolved under the close supervision of a multidisciplinary team. If conservative therapy can be safely implemented for 4–6 weeks, to allow the pancreatic collections to resolve or “wall-off,” surgical or endoscopic intervention is generally much safer and better tolerated by the patient.

PSEUDOCYST

The incidence of pseudocyst is low, and most acute collections resolve over time. Less than 10% of patients have persistent fluid collections after 6 weeks that would meet the definition of a pseudocyst. Only symptomatic collections should be drained with surgery or endoscopy or by percutaneous route.

PANCREATIC DUCT DISRUPTION

Pancreatic duct disruption may present with symptoms of increasing abdominal pain or shortness of breath in the setting of an enlarging fluid collection. Diagnosis can be confirmed on magnetic resonance cholangiopancreatography (MRCP) or ERCP. Placement of a bridging pancreatic stent for at least 6 weeks is >90% effective at resolving the leak. Nonbridging stents are less effective (25–50%).

PERIVASCULAR COMPLICATIONS

Perivascular complications may include *splenic vein thrombosis* with gastric varices and pseudoaneurysms. *Gastric varices* bleed <5% of the time. Life-threatening bleeding from a ruptured *pseudoaneurysm* can be diagnosed and treated with mesenteric angiography and embolization.

EXTRAPANCREATIC INFECTIONS

Hospital-acquired infections occur in up to 20% of patients with acute pancreatitis. Patients should be continually monitored for the development pneumonia, urinary tract infection, and line infection. Continued culturing of urine, monitoring of chest x-rays, and routine changing of intravenous lines are important during hospitalization.

Follow-Up Care

Hospitalizations for moderately severe and severe acute pancreatitis can be prolonged and last weeks to months and often involve a period of intensive care unit admission and outpatient rehabilitation or subacute nursing care. Follow-up evaluation should assess for development of diabetes, exocrine insufficiency, recurrent cholangitis, or development of infected fluid collections. As mentioned previously, cholecystectomy should be performed during hospitalization or within 4–6 weeks of discharge if possible for patients with uncomplicated gallstone pancreatitis.

RECURRENT PANCREATITIS

Approximately 25% of patients who have had an attack of acute pancreatitis have a recurrence. The two most common etiologic factors are alcohol and cholelithiasis. In patients with recurrent pancreatitis without an obvious cause, the differential diagnosis should encompass occult biliary tract disease including microlithiasis, hypertriglyceridemia, drugs, pancreatic cancer, pancreas divisum, and cystic fibrosis (Table 341-1). In one series of 31 patients diagnosed initially as having idiopathic or recurrent acute pancreatitis, 23 were found to have occult gallstone disease. Thus, approximately two-thirds of patients with recurrent acute pancreatitis without an obvious cause actually have occult gallstone disease due to microlithiasis. Genetic defects as in hereditary pancreatitis and cystic fibrosis mutations can result in recurrent pancreatitis. Other diseases of the biliary tree and pancreatic ducts that can cause acute pancreatitis include choledochocele; ampullary tumors; pancreas divisum; and pancreatic duct stones, stricture, and tumor. Approximately 2–4% of patients with pancreatic carcinoma present with acute pancreatitis.

PANCREATITIS IN PATIENTS WITH AIDS

The incidence of acute pancreatitis is theoretically increased in patients with AIDS for two reasons: (1) the high incidence of infections involving the pancreas such as infections with cytomegalovirus, *Cryptosporidium*, and the *Mycobacterium avium* complex; and (2) the frequent use by patients with AIDS of medications such as didanosine, pentamidine, trimethoprim-sulfamethoxazole, and protease inhibitors. It should be noted that the incidence has been markedly reduced due to advances in therapy **(Chap. 197)**.

CHRONIC PANCREATITIS AND PANCREATIC EXOCRINE INSUFFICIENCY

PATHOPHYSIOLOGY

Chronic pancreatitis is a disease process characterized by irreversible damage to the pancreas as distinct from the reversible changes noted in acute pancreatitis (Table 341-4). The events that initiate and then perpetuate the inflammatory process in the pancreas are becoming more clearly understood. Irrespective of the mechanism of injury, it is becoming apparent that stellate cell activation that results in cytokine expression and production of extracellular matrix proteins cause acute and chronic inflammation and collagen deposition in the pancreas. Thus, the condition is defined by the presence of histologic abnormalities, including chronic inflammation, fibrosis, and progressive destruction of both exocrine and eventually endocrine tissue (atrophy). A number of etiologies have been associated with chronic pancreatitis resulting in the cardinal manifestations of the disease such as abdominal pain, steatorrhea, weight loss, and diabetes mellitus **(Table 341-5)**.

TABLE 341-5

Chronic Pancreatitis and Pancreatic Exocrine Insufficiency: Tigar-O Classification System

Toxic-metabolic
Alcoholic
Tobacco smoking
Hypercalcemia
Hyperlipidemia
Chronic renal failure
Medications—phenacetin abuse
Toxins—organotin compounds (e.g., dibutylin dichloride, DBTC)
Idiopathic
Early onset
Late onset
Tropical
Genetic
Cationic trypsinogen (*PRSS1*)
Cystic fibrosis transmembrane conductance regulator gene (*CFTR*)
Calcium-sensing receptor (*CASR*)
Chymotrypsin C gene (*CTRC*)
Pancreatic secretory trypsin inhibitor gene (*SPINK1*)
Autoimmune
Type 1 autoimmune chronic pancreatitis
IgG4 systemic
Type 2 autoimmune chronic pancreatitis
Recurrent and severe acute pancreatitis
Postnecrotic (severe acute pancreatitis)
Recurrent acute pancreatitis
Vascular diseases/ischemia
Radiation induced
Obstructive

Pancreas divisum
Duct obstruction (e.g., tumor)
Preampullary duodenal wall cysts
Posttraumatic pancreatic duct scars

Abbreviations: DBTC, dibutylin dichloride; TIGAR-O, toxic-metabolic, idiopathic, genetic, autoimmune, recurrent and severe acute pancreatitis, obstructive.

Although alcohol has been believed to be the primary cause of chronic pancreatitis, other factors contribute to the disease because not all heavy consumers of alcohol develop pancreatic disease. There is also a strong association between smoking and chronic pancreatitis. Cigarette smoke leads to an increased susceptibility to pancreatic autodigestion and predisposes to dysregulation of duct cell CFTR function. Smoking is an independent, dose-dependent risk factor for chronic pancreatitis and recurrent acute pancreatitis. Both continued alcohol and smoking exposure are associated with pancreatic fibrosis, calcifications, and progression of disease.

Recent characterization of pancreatic stellate cells (PSCs) has added insight into the underlying cellular responses behind development of chronic pancreatitis. Specifically, PSCs are believed to play a role in maintaining normal pancreatic architecture that can shift toward fibrogenesis in the case of chronic pancreatitis. The sentinel acute pancreatitis event (SAPE) hypothesis uniformly describes the events in the pathogenesis of chronic pancreatitis. It is believed that alcohol or additional stimuli lead to matrix metalloproteinase–mediated destruction of normal collagen in pancreatic parenchyma, which later allows for pancreatic remodeling. Proinflammatory cytokines, tumor necrosis factor α (TNF-α), interleukin 1 (IL-1), and interleukin 6 (IL-6), as well as oxidant complexes, are able to induce PSC activity with subsequent new collagen synthesis. In addition to being stimulated by cytokines, oxidants, or growth factors, PSCs also possess transforming growth factor β (TGF-β)–mediated self-activating autocrine pathways that may explain disease progression in chronic pancreatitis even after removal of noxious stimuli.

ETIOLOGIC CONSIDERATIONS

Among adults in the United States, alcoholism is the most common cause of clinically apparent chronic pancreatitis, whereas cystic fibrosis is the most frequent cause in children. As many as 25% of adults in the United States with chronic pancreatitis have the *idiopathic* form. Recent investigations have indicated that up to 15% of patients with idiopathic pancreatitis may have pancreatitis due to genetic defects (Table 341-5).

Whitcomb and associates studied several large families with hereditary chronic pancreatitis and were able to identify a genetic defect that affects the gene encoding for trypsinogen. Several additional defects of this gene have also been described. The defect prevents the destruction of prematurely activated trypsin and allows it to be resistant to the intracellular protective effect of trypsin inhibitor. It is hypothesized that this continual activation of digestive enzymes within the gland leads to acute injury and, finally, chronic pancreatitis. Since the initial discovery of the *PRSS1* mutation defect, other genetic diseases have been detected (Table 341-5).

Several other groups of investigators have documented mutations of *CFTR*. This gene functions as a cyclic AMP–regulated chloride channel. In patients with cystic fibrosis, the high concentration of macromolecules can block the pancreatic ducts. It must be appreciated, however, that there is a great deal of heterogeneity in relationship to the *CFTR* gene defect. More than 1000 putative mutations of the *CFTR* gene have been identified. Attempts to elucidate the relationship between the genotype and pancreatic manifestations have been hampered by the number of mutations. The ability to detect *CFTR* mutations has led to the recognition that the clinical spectrum of the disease is broader than previously thought. Two studies have clarified the association between mutations of the *CFTR* gene and another monosymptomatic form of cystic fibrosis (i.e., chronic pancreatitis). It is estimated that in patients with idiopathic pancreatitis, the frequency of a single *CFTR* mutation is 11 times the expected frequency and the frequency of two mutant alleles is 80 times the expected frequency. In these studies, the patients were adults when the diagnosis of pancreatitis was made; none had any clinical evidence of pulmonary disease, and sweat test results were not diagnostic of cystic fibrosis. The prevalence of such mutations is unclear, and further studies are certainly needed. In addition, the therapeutic and prognostic implication of these findings with respect to managing pancreatitis remains to be determined. Long-term follow-up of affected patients is needed. *CFTR* mutations are common in the general population. It is unclear whether the *CFTR* mutation alone can lead to pancreatitis as an autosomal recessive disease. A study evaluated 39 patients with idiopathic chronic pancreatitis to assess the risk associated with these mutations. Patients with two *CFTR* mutations (compound heterozygotes) demonstrated *CFTR* function at a level between that seen in typical cystic fibrosis and cystic fibrosis carriers and had a fortyfold increased risk of pancreatitis. The presence of an *N34S SPINK1* mutation increased the risk twentyfold. A combination of two *CFTR* mutations and an *N34S SPINK1* mutation increased the risk of pancreatitis 900-fold. Knowledge of the genetic defects and downstream alterations in protein expression has led to the development of novel genetic therapy in cystic fibrosis children that potentiates the CFTR channel resulting in improvement in lung function, quality of life, and weight gain. Table 341-5 lists recognized causes of chronic pancreatitis and pancreatic exocrine insufficiency.

AUTOIMMUNE PANCREATITIS

AIP is an uncommon disorder of presumed autoimmune causation with characteristic laboratory, histologic, and morphologic findings **(Table 341-6)**. The nomenclature has recently been simplified to include AIP and idiopathic duct centric pancreatitis (IDCP). In AIP, the pancreas is involved as part of an IgG4 systemic disease **(Chap. 361)** and meets HISORt criteria as defined below. The characteristic pancreatic histopathologic findings include lymphoplasmacytic infiltrate, storiform fibrosis, and abundant IgG4 cells. IDCP is histologically confirmed IDCP with granulocytic infiltration of the duct wall (termed GEL), but without IgG4 positive cells and systemic involvement. It is a disorder limited to the pancreas only.

TABLE 341-6

Clinical Features of Autoimmune Pancreatitis (AIP)

• Mild symptoms, usually abdominal pain, but without frequent attacks of acute pancreatitis
• Diffuse swelling and enlargement of the pancreas
• Two-thirds of patients present with either obstructive jaundice or a “mass” in the head of the pancreas mimicking carcinoma
• Diffuse irregular narrowing of the pancreatic duct (MRCP or ERCP)
• Increased levels of serum gamma globulins, especially IgG4
• Presence of other autoantibodies (ANA), rheumatoid factor (RF)
• Can occur with other autoimmune diseases: Sjögren’s syndrome, primary sclerosing cholangitis, ulcerative colitis, rheumatoid arthritis
• Extrapancreatic bile duct changes such as stricture of the common bile duct and intrahepatic ducts
• Pancreatic calcifications (rare)
• Pancreatic biopsies reveal extensive fibrosis and lymphoplasmacytic infiltration
• Glucocorticoids are effective in alleviating symptoms, decreasing size of the pancreas, and reversing histopathologic changes

Abbreviations: ERCP, endoscopic retrograde cholangiopancreatography; MRCP, magnetic resonance cholangiopancreatography.

Although AIP was initially described as a primary pancreatic disorder, it is now recognized that it is associated with other disorders of presumed autoimmune etiology, and this has been termed IgG4-related disease **(Chap. 361)**. The clinical features include IgG4-associated cholangitis, rheumatoid arthritis, Sjögren’s syndrome, ulcerative colitis, mediastinal fibrosis and adenopathy, autoimmune thyroiditis, tubulointerstitial nephritis, retroperitoneal fibrosis, chronic periaortitis, chronic sclerosing sialadenitis, and Mikulicz’s disease. Mild symptoms, usually abdominal pain, and recurrent acute pancreatitis are unusual. Furthermore, AIP is not a common cause of idiopathic recurrent pancreatitis.

Weight loss from pancreatic atrophy and new onset of diabetes may also occur in patients that smoke and consume alcohol. An obstructive pattern on liver tests is common (i.e., disproportionately elevated serum alkaline phosphatase and minimally elevated serum aminotransferases). Elevated serum levels of IgG4 provide a marker for the disease, particularly in Western populations. Serum IgG4 is elevated in 2/3 of those with AIP. CT scans reveal abnormalities in the majority of patients and include diffuse enlargement, focal enlargement, and a distinct enlargement at the head of the pancreas. ERCP or MRCP reveals strictures in the bile duct in more than one-third of patients with AIP; these may include common bile duct strictures, intrahepatic bile duct strictures, or proximal bile duct strictures, with accompanying narrowing of the pancreatic portion of the bile duct. This has been termed autoimmune IgG4 cholangitis. Characteristic histologic findings include extensive lymphoplasmacytic infiltrates with dense fibrosis around pancreatic ducts, as well as a lymphoplasmacytic infiltration, resulting in an obliterative phlebitis.

The Mayo Clinic HISORt criteria indicate that AIP can be diagnosed by the presence of at least one or more of the following: (1) *h*istology; (2) *i*maging; (3) *s*erology (elevated serum IgG4 levels); (4) *o*ther organ involvement; and (5) *r*esponse to glucocorticoid *t*herapy, with improvement in pancreatic and extrapancreatic manifestations.

Glucocorticoids have shown efficacy in alleviating symptoms, decreasing the size of the pancreas, and reversing histopathologic features in patients with AIP. Patients may respond dramatically to glucocorticoid therapy within a 2- to 4-week period. Prednisone is usually administered at an initial dose of 40 mg/d for 4 weeks followed by a taper of the daily dosage by 5 mg/wk based on monitoring of clinical parameters. Relief of symptoms, serial changes in abdominal imaging and improvements in liver tests are parameters to follow. A poor response to glucocorticoids over a 2- to 4-week period should raise suspicion of pancreatic cancer or other forms of chronic pancreatitis. A recent multicenter international report reviewed 1064 patients with AIP. Clinical remission was achieved in 99% of AIP and 92% of IDCP patients with steroids. However, disease relapse occurred in 31% of AIP and 9% of

IDCP patients. For treatment of disease relapse in AIP, glucocorticoids were successful in 201 of 295 (68%) patients, and azathioprine was successful in 52 of 58 patients (85%). A small number of patients responded favorably to 6-mercaptapurine, rituximab, cyclosporine, and cyclophosphamide. AIP and IDCP are highly responsive to initial glucocorticoid treatment. Relapse is common in AIP patients, especially those with biliary tract strictures. Most relapses occur after glucocorticoids are discontinued. Patients with refractory symptoms and strictures generally require immunomodulator therapy as noted above. Rituximab, a monoclonal antibody directed against B cells has been shown to be very effective at inducing and maintaining remission. Appearance of interval cancers following a diagnosis of AIP is uncommon.

Clinical Features of Chronic Pancreatitis

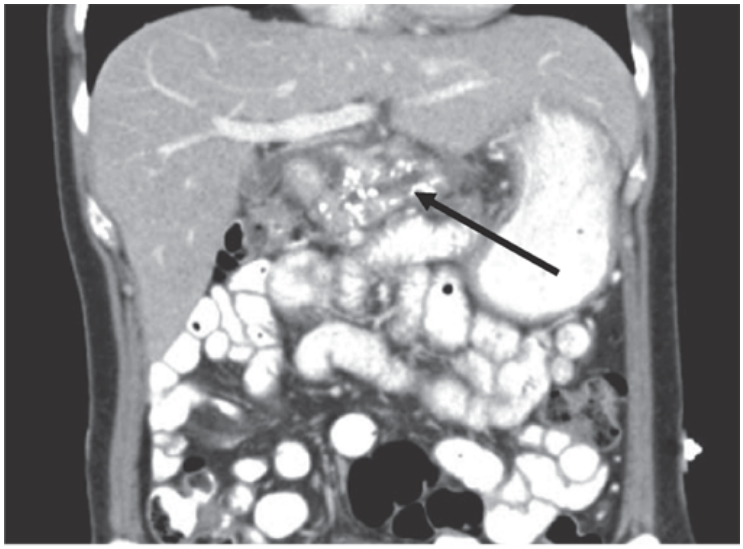
Patients with chronic pancreatitis seek medical attention predominantly because of two symptoms: abdominal pain or maldigestion and weight loss. The abdominal pain may be quite variable in location, severity, and frequency. The pain can be constant or intermittent with frequent pain-free intervals. Eating may exacerbate the pain, leading to a fear of eating with consequent weight loss. The spectrum of abdominal pain ranges from mild to quite severe, with narcotic dependence as a frequent consequence. Maldigestion is manifested as chronic diarrhea, steatorrhea, weight loss, and fatigue. Patients with chronic abdominal pain may or may not progress to maldigestion, and ~20% of patients will present with symptoms of maldigestion without a history of abdominal pain. Patients with chronic pancreatitis have significant morbidity and mortality and use appreciable amounts of societal resources. Despite steatorrhea, clinically apparent deficiencies of fat-soluble vitamins are surprisingly uncommon. Physical findings in these patients are usually unimpressive, so that there is a disparity between the severity of abdominal pain and the physical signs that usually consist of some mild tenderness.

The diagnosis of early or mild chronic pancreatitis can be challenging because there is no biomarker for the disease. In contrast to acute pancreatitis, the serum amylase and lipase levels are usually not strikingly elevated in chronic pancreatitis. Elevation of serum bilirubin and alkaline phosphatase may indicate cholestasis secondary to common bile duct stricture caused by chronic inflammation. Many patients have impaired glucose tolerance with elevated fasting blood glucose levels. The fecal elastase-1 and small-bowel biopsy are useful in the evaluation of patients with suspected pancreatic steatorrhea. The fecal elastase level will be abnormal and small-bowel histology will be normal in such patients. A decrease of fecal elastase level to <100 μg per gram of stool strongly suggests severe pancreatic exocrine insufficiency.

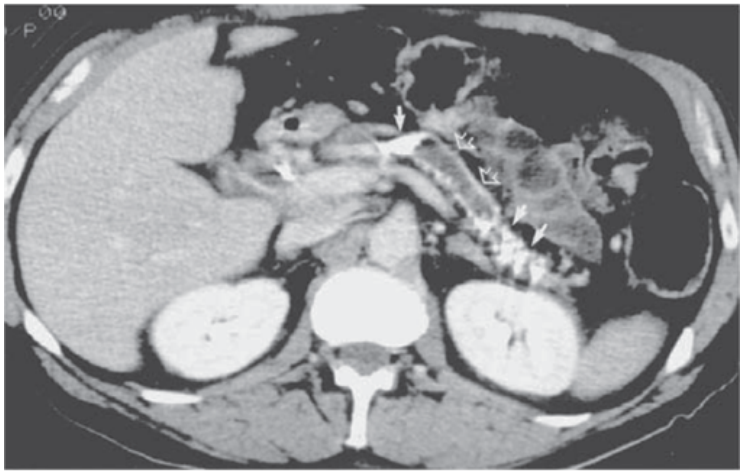
The radiographic evaluation of a patient with suspected chronic pancreatitis usually proceeds from a noninvasive to more invasive approach. Abdominal CT imaging **(Fig. 341-4A, B)** is the initial modality of choice, followed by MRI **(Fig. 341-4C)**, endoscopic ultrasound, and pancreas function testing. In addition to excluding a pseudocyst and pancreatic cancer, CT may show calcification, dilated ducts, or an atrophic pancreas. Although abdominal CT scanning and MRCP greatly aid in the diagnosis of pancreatic disease, the diagnostic test with the best sensitivity and specificity is the hormone stimulation test using secretin. The secretin test becomes abnormal when ≥60% of the pancreatic exocrine function has been lost. This usually correlates well with the onset of chronic abdominal pain. The role of endoscopic ultrasonography (EUS) in diagnosing early chronic pancreatitis is still being defined. A total of nine endosonographic features have been described in chronic pancreatitis. The presence of five or more features is considered diagnostic of chronic pancreatitis. EUS is not a sensitive enough test for detecting early chronic pancreatitis alone **(Chap. 340)** and may show positive features in patients who have dyspepsia or even normal aging individuals. Recent data suggest that EUS can be combined with endoscopic pancreatic function testing (EUS-ePFT) during a single endoscopy to screen for chronic pancreatitis in patients with chronic abdominal pain. Diffuse calcifications noted on plain film of the abdomen usually indicate significant damage to the pancreas and are pathognomic for chronic pancreatitis (Fig. 341-4A). Although alcohol is by far the most common cause of pancreatic calcification, such calcification may also be noted in hereditary pancreatitis, posttraumatic pancreatitis, hypercalcemic pancreatitis, idiopathic chronic pancreatitis, and tropical pancreatitis.

FIGURE 341-4

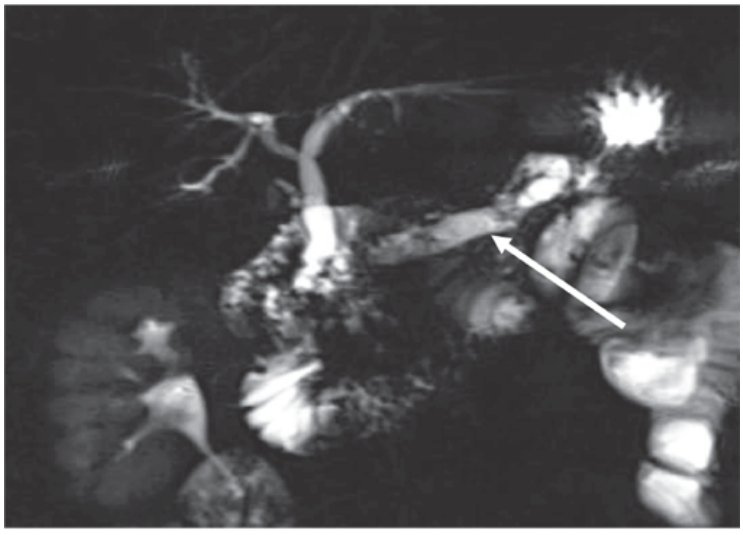
A. Chronic pancreatitis and pancreatic calculi: computed tomography (CT) scan. In this contrast-enhanced CT scan of the abdomen, there is evidence of an atrophic pancreas with multiple calcifications and stones in the parenchyma and dilated pancreatic duct (*arrow*). ***B.*** In this contrast-enhanced CT scan of the abdomen, there is evidence of an atrophic pancreas with multiple calcifications (*arrows*). Note the markedly dilated pancreatic duct seen in this section through the body and tail (*open arrows*). ***C.*** Chronic pancreatitis on magnetic resonance cholangiopancreatography (MRCP): dilated duct with filling defects. Gadolinium-enhanced magnetic resonance imaging/MRCP reveals a dilated pancreatic duct (*arrow*) in chronic pancreatitis with multiple filling defects suggestive of pancreatic duct calculi. *(A, C, courtesy of Dr. KJ Mortele, Brigham and Women's Hospital; with permission.)*



A



B



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Complications of Chronic Pancreatitis

The complications of chronic pancreatitis are protean and are listed in Table 341-7. Although most patients have impaired glucose tolerance, diabetic ketoacidosis and diabetic coma are uncommon. Likewise, end-organ damage (retinopathy, neuropathy, nephropathy) is also uncommon. A nondiabetic

retinopathy may be due to either vitamin A and/or zinc deficiency. Gastrointestinal bleeding may occur from peptic ulceration, gastritis, a pseudocyst eroding into the duodenum, arterial bleeding into the pancreatic duct (hemosuccus pancreaticus), or ruptured varices secondary to splenic vein thrombosis due to chronic inflammation of the tail of the pancreas. Jaundice, cholestasis, and biliary cirrhosis may occur from the chronic inflammatory reaction around the intrapancreatic portion of the common bile duct. Twenty years after the diagnosis of calcific chronic pancreatitis, the cumulative risk of pancreatic carcinoma is 4%. Patients with hereditary pancreatitis are at a tenfold higher risk for pancreatic cancer.

TABLE 341-7

Complications of Chronic Pancreatitis

Chronic abdominal pain	Jaundice
Narcotic addiction	Biliary stricture and/or biliary cirrhosis
Diabetes mellitus/impaired glucose tolerance	Pseudocyst
Gastroparesis	Metabolic bone disease
Malabsorption/maldigestion	Pancreatic cancer

TREATMENT

TREATMENT

Chronic Pancreatitis

STEATORRHEA

The treatment of steatorrhea with pancreatic enzymes is straightforward even though complete correction of steatorrhea is unusual. Enzyme therapy usually brings diarrhea under control and restores absorption of fat to an acceptable level and affects weight gain. Thus, pancreatic enzyme replacement has been the cornerstone of therapy. In treating steatorrhea, it is important to use a potent pancreatic formulation that will deliver sufficient lipase into the duodenum to correct maldigestion and decrease steatorrhea. In an attempt to standardize the enzyme activity, potency, and bioavailability, the U.S. Food and Drug Administration (FDA) required that all pancreas enzyme drugs in the United States obtain a New Drug Application (NDA) by April 2008. Table 341-8 lists frequently used formulations, but availability will be based on compliance with the FDA mandate. Recent data suggest that dosages up to 80,000–100,000 units of lipase taken during the meal may be necessary to normalize nutritional parameters in malnourished chronic pancreatitis patients, and some may require acid suppression with proton pump inhibitors.

ABDOMINAL PAIN

The management of pain in patients with chronic pancreatitis is problematic. Recent meta-analyses have shown no consistent benefit of enzyme therapy at reducing pain in chronic pancreatitis. In some patients with idiopathic chronic pancreatitis, conventional non-enteric-coated enzyme preparations containing high concentrations of serine proteases may relieve mild abdominal pain or discomfort. The pain relief experienced by these patients actually may be due to improvements in the dyspepsia from maldigestion.

Gastroparesis is also quite common in patients with chronic pancreatitis. It is important to recognize and treat with prokinetic drugs because treatment with enzymes may fail simply because gastric dysmotility is interfering with the delivery of enzymes into the upper intestine. A recent prospective study reported that pregabalin can improve pain in chronic pancreatitis and lower pain medication requirement.

Endoscopic treatment of chronic pancreatitis pain may involve sphincterotomy, stenting, stone extraction, and drainage of a pancreatic pseudocyst. Therapy directed to the pancreatic duct would seem to be most appropriate in the setting of a dominant stricture, especially if a ductal stone has led to obstruction. The use of endoscopic stenting for patients with chronic pain, but without a dominant stricture, has not been subjected to any controlled trials. It is now appreciated that significant complications can occur from stenting (i.e., bleeding, cholangitis, stent migration, pancreatitis, and stent clogging). In patients with large-duct disease usually from alcohol-induced chronic pancreatitis, ductal decompression with *surgical therapy* has been the therapy of choice. Among such patients, 80% seem to obtain immediate relief; however, at the end of 3 years, one-half of the patients have recurrence of pain. Two randomized prospective trials comparing endoscopic to surgical therapy for chronic pancreatitis demonstrated that surgical therapy was superior to endoscopy at decreasing pain and improving quality of life in selected patients with dilated ducts and abdominal pain. This would suggest that chronic pancreatitis patients with dilated ducts and pain should be considered for surgical intervention. The role of preoperative stenting prior to surgery as a predictor of response has yet to be proven.

A Whipple procedure, total pancreatectomy, and autologous islet cell transplantation have been used in selected patients with chronic pancreatitis and abdominal pain refractory to conventional therapy. The patients who have benefited the most from total pancreatectomy have chronic pancreatitis without prior pancreatic surgery or evidence of islet cell insufficiency. The role of this procedure remains to be fully defined but may be an option in lieu

of ductal decompression surgery or pancreatic resection in patients with intractable, painful small-duct disease, hereditary pancreatitis and particularly as the standard surgical procedures tend to decrease islet cell yield. Celiac plexus block has not resulted in long-lasting pain relief.

TABLE 341-8

FDA-Approved Pancreatic Enzyme (Pancrelipase) Preparations

Product	Enzyme Content/Unit Dose, U.S. Pharmacopeia Units		
	Lipase[a]	Amylase[a]	Protease[a]
Immediate-Release Capsule			
Non-enteric-coated			
Viokace 10,440	10,440	391,550	39,150
Viokace 20,880	20,880	78,300	78,300
Delayed-Release Capsules			
Enteric-coated mini-microspheres			
Creon 3000	3000	15,000	9500
Creon 6000	6000	30,000	19,000
Creon 12,000	12,000	60,000	38,000
Creon 24,000	24,000	120,000	76,000
Enteric-Coated Mini-Tablets			
Ultresa 13,800	13,800	27,600	27,600
Ultresa 20,700	20,700	41,400	41,400
Ultresa 23,000	23,000	46,000	46,000
Enteric-Coated Beads			
Zenpep 3000	3000	16,000	10,000
Zenpep 5000	5000	27,000	17,000
Zenpep 10,000	10,000	55,000	34,000
Zenpep 15,000	15,000	82,000	51,000
Zenpep 20,000	20,000	109,000	68,000
Zenpep 25,000	25,000	136,000	85,000
Enteric-Coated Micro-Tablets			
Pancreaze 4200	4200	17,500	10,000
Pancreaze 10,500	10,500	43,750	25,000
Pancreaze 16,800	16,800	70,000	40,000
Pancreaze 21,000	21,000	61,000	37,000

Product	Enzyme Content/Unit Dose, U.S. Pharmacopeia Units		
	Lipase[a]	Amylase[a]	Protease[a]
Bicarbonate-Buffered Enteric-Coated Microspheres			
Pertzye 8000	8000	30,250	28,750
Pertzye 16,000	16,000	60,500	57,500

[a]U.S. Pharmacopeia (USP) units per tablet or capsule.

Note: The FDA has mandated all enzyme manufacturers to submit New Drug Applications (NDAs) for all pancreatic extract drug products after reviewing data that showed substantial variations among currently marketed products. Numerous manufacturers have investigations under way to seek FDA approval for the treatment of exocrine pancreatic insufficiency due to cystic fibrosis or other conditions under the new guidelines for this class of drugs (*www.fda.gov*).

HEREDITARY PANCREATITIS

Hereditary pancreatitis is a rare disease that is similar to chronic pancreatitis except for an early age of onset and evidence of hereditary factors. A genomewide search using genetic linkage analysis identified the hereditary pancreatitis gene on chromosome 7. Mutations in ion codons 29 (exon 2) and 122 (exon 3) of the cationic trypsinogen gene cause autosomal dominant forms of hereditary pancreatitis. The codon 122 mutations lead to a substitution of the corresponding arginine with another amino acid, usually histidine. This substitution, when it occurs, eliminates a fail-safe trypsin self-destruction site necessary to eliminate trypsin that is prematurely activated within the acinar cell. These patients have recurring attacks of severe abdominal pain that may last from a few days to a few weeks. The serum amylase and lipase levels may be elevated during acute attacks but are usually normal. Patients frequently develop pancreatic calcification, diabetes mellitus, and steatorrhea; in addition, they have an increased incidence of pancreatic carcinoma, with the cumulative incidence being as high as 40% by age 70 years. A recent natural history study of hereditary pancreatitis in >200 patients from France reported that abdominal pain started in childhood at age 10 years, steatorrhea developed at age 29 years, diabetes at age 38 years, and pancreatic carcinoma at age 55 years. Such patients often require surgical ductal decompression for pain relief. Abdominal complaints in relatives of patients with hereditary pancreatitis should raise the question of pancreatic disease.

PSTI, or SPINK1, is a 56-amino-acid peptide that specifically inhibits trypsin by physically blocking its active site. SPINK1 acts as the first line of defense against prematurely activated trypsinogen in the acinar cell. Recently, it has been shown that the frequency of SPINK1 mutations in patients with idiopathic chronic pancreatitis is markedly increased, suggesting that these mutations may be associated with pancreatitis.

PANCREATIC ENDOCRINE TUMORS

Pancreatic endocrine tumors are discussed in Chap. 80.

OTHER CONDITIONS

ANNULAR PANCREAS

When the ventral pancreatic anlage fails to migrate correctly to make contact with the dorsal anlage, the result may be a ring of pancreatic tissue encircling the duodenum. Such an annular pancreas may cause intestinal obstruction in the neonate or the adult. Symptoms of postprandial fullness, epigastric pain, nausea, and vomiting may be present for years before the diagnosis is entertained. The radiographic findings are symmetric dilation of the proximal duodenum with bulging of the recesses on either side of the annular band, effacement but not destruction of the duodenal mucosa, accentuation of the findings in the right anterior oblique position, and lack of change on repeated examinations. The differential diagnosis should include duodenal webs, tumors of the pancreas or duodenum, postbulbar peptic ulcer, regional enteritis, and adhesions. Patients with annular pancreas have an increased incidence of pancreatitis and peptic ulcer. Because of these and other potential complications, the treatment is surgical even if the condition has been present for years. Retrocolic duodenojejeunostomy is the procedure of choice, although some surgeons advocate Billroth II gastrectomy, gastroenterostomy, and vagotomy.

PANCREAS DIVISUM

Pancreas divisum is present in 7–10% of the population and occurs when the embryologic ventral and dorsal pancreatic anlagen fail to fuse, so that pancreatic drainage is accomplished mainly through the accessory papilla. Pancreas divisum is the most common congenital anatomic variant of the

human pancreas. Current evidence indicates that this anomaly does not predispose to the development of pancreatitis in the great majority of patients who harbor it. However, the combination of pancreas divisum and a small accessory orifice could result in dorsal duct obstruction. The challenge is to identify this subset of patients with dorsal duct pathology. Cannulation of the dorsal duct by ERCP is not as easily done as is cannulation of the ventral duct. Patients with pancreatitis and pancreas divisum demonstrated by MRCP or ERCP should be treated with conservative measures. In many of these patients, pancreatitis is idiopathic and unrelated to the pancreas divisum. Endoscopic or surgical intervention is indicated only if pancreatitis recurs and no other cause can be found. If marked dilation of the dorsal duct can be demonstrated, surgical ductal decompression should be performed. It should be stressed that the ERCP/MRCP appearance of pancreas divisum (i.e., a small-caliber ventral duct with an arborizing pattern) may be mistaken as representing an obstructed main pancreatic duct secondary to a mass lesion.

MACROAMYLASEMIA

In macroamylasemia, amylase circulates in the blood in a polymer form too large to be easily excreted by the kidney. Patients with this condition demonstrate an elevated serum amylase value and a low urinary amylase value. The presence of macroamylase can be documented by chromatography of the serum. The prevalence of macroamylasemia is 1.5% of the nonalcoholic general adult hospital population. Usually macroamylasemia is an incidental finding and is not related to disease of the pancreas or other organs. Macrolipasemia has now been documented in a few patients with cirrhosis or non-Hodgkin's lymphoma. In these patients, the pancreas appeared normal on ultrasound and CT examination. Lipase was shown to be complexed with immunoglobulin A. Thus, the possibility of *both* macroamylasemia and macrolipasemia should be considered in patients with elevated blood levels of these enzymes.

ACKNOWLEDGMENT

This chapter represents a revised version of chapters by Drs. Norton J. Greenberger, Phillip P. Toskes, and Bechien Wu that were in previous editions of Harrison's.

FURTHER READING

Duggan SN et al: High prevalence of osteoporosis in patients with chronic pancreatitis: A systematic review and meta-analysis. Clin Gastroenterol Hepatol 12:219, 2014. [PubMed: 23856359]

Forsmark CE: Management of chronic pancreatitis. Gastroenterology 144:1282, 2013. [PubMed: 23622138]

Hart PA, Zen Y, Chari ST: Recent advances in autoimmune pancreatitis. Gastroenterology 149:39, 2015. [PubMed: 25770706]

Tenner S et al: American College of Gastroenterology guideline: Management of acute pancreatitis. Am J Gastroenterol 108:1400, 2013. [PubMed: 23896955]

Wu B, Banks PA: Clinical Management of patients with acute pancreaitis. Gastroenterology 144:1272, 2013. [PubMed: 23622137]

Yadav D, Lowenfels AB: The epidemiology of pancreatitis and pancreatic cancer. Gastroenterology 144:1252, 2013. [PubMed: 23622135]

Chapter 342: Introduction to the Immune System

Barton F. Haynes; Kelly A. Soderberg; Anthony S. Fauci

DEFINITIONS

Adaptive immune system—recently evolved system of immune responses mediated by T and B lymphocytes. Immune responses by these cells are based on specific antigen recognition by clonotypic receptors that are products of genes that rearrange during development and throughout the life of the organism. Additional cells of the adaptive immune system include various types of antigen-presenting cells (APCs).

Antibody—B cell–produced molecules encoded by genes that re-arrange during B cell development consisting of immunoglobulin heavy and light chains that together form the central component of the B cell receptor (BCR) for antigen. Antibody can exist as B cell–surface antigen-recognition molecules or as secreted molecules in plasma and other body fluids.

Antigens—foreign or self-molecules that are recognized by the adaptive and innate immune systems resulting in immune cell triggering, T cell activation, and/or B cell antibody production.

Antimicrobial peptides—small peptides <100 amino acids in length that are produced by cells of the innate immune system and have anti-infectious agent activity.

Apoptosis—the process of *programmed cell death* whereby signaling through various "death receptors" on the surface of cells (e.g., tumor necrosis factor [TNF] receptors, CD95) leads to a signaling cascade that involves activation of the caspase family of molecules and leads to DNA cleavage and cell death. Apoptosis, which does not lead to induction of inordinate inflammation, is to be contrasted with *cell necrosis*, which does lead to induction of inflammatory responses.

Autoimmune diseases—diseases such as systemic lupus erythematosus and rheumatoid arthritis in which cells of the adaptive immune system such as autoreactive T and B cells become overreactive and produce self-reactive T cell and antibody responses.

Autoinflammatory diseases—hereditary disorders such as hereditary periodic fevers (HPFs) characterized by recurrent episodes of severe inflammation and fever due to mutations in controls of the innate inflammatory response, i.e., the inflammasome (see below and Table 342-6). Patients with HPFs also have rashes and serosal and joint inflammation, and some can have neurologic symptoms. Autoinflammatory diseases are different from autoimmune diseases in that evidence for activation of adaptive immune cells such as autoreactive B cells is not present.

Autophagy—lysosomal degradation pathway mechanism of cells to dispose of intracellular debris and damaged organelles. Autophagy by cells of the innate immune system is used to control intracellular infectious agents such as mycobacteria, in part by initiation of phagosome maturation and enhancing major histocompatibility complex (MHC) class II antigen presentation to CD4 T cells.

B cell receptor for antigen—complex of surface molecules that rearrange during postnatal B cell development, made up of surface immunoglobulin (Ig) and associated Ig αβ chain molecules that recognize nominal antigen via Ig heavy- and light-chain variable regions, and signal the B cell to terminally differentiate to make antigen-specific antibody.

B lymphocytes—bone marrow-derived or bursal-equivalent lymphocytes that express surface immunoglobulin (the BCR for antigen) and secrete specific antibody after interaction with antigen.

B regulatory cells—a population of suppressive B cells that aid in the inhibition of inflammation through the release of cytokines such as IL-10.

CD classification of human lymphocyte differentiation antigens—the development of monoclonal antibody technology led to the discovery of a large number of new leukocyte surface molecules. In 1982, the First International Workshop on Leukocyte Differentiation Antigens was held to establish a nomenclature for cell-surface molecules of human leukocytes. From this and subsequent leukocyte differentiation workshops has come the *cluster of differentiation* (CD) classification of leukocyte antigens.

CD4 T cell—T lymphocyte subset that participates in adaptive immunity and helps B cells make antibody.

CD8 T cell—cytotoxic T lymphocyte subset that destroys tumor cells and cells infected with intracellular pathogens.

Chemokines—soluble molecules that direct and determine immune cell movement and circulation pathways.

Complement—cascading series of plasma enzymes and effector proteins whose function is to lyse pathogens and/or target them to be phagocytized by neutrophils and monocyte/macrophage lineage cells of the reticuloendothelial system.

Co-stimulatory molecules—molecules of APCs (such as B7-1 and B7-2 or CD40) that lead to T cell activation when bound by ligands on activated T cells (such as CD28 or CD40 ligand).

Crystallopathies—nanoparticle or microparticle-sized deposits of crystals, misfolded proteins or airborne particulate matter that can stimulate the inflammasome and initiate inflammation and tissue damage.

Cytokines—soluble proteins that interact with specific cellular receptors that are involved in the regulation of the growth and activation of immune cells and mediate normal and pathologic inflammatory and immune responses.

Dendritic cells—myeloid and/or lymphoid lineage APCs of the adaptive immune system. Immature dendritic cells (DCs), or DC precursors, are key components of the innate immune system by responding to infections with production of high levels of cytokines. DCs are key initiators both of innate immune responses via cytokine production and of adaptive immune responses via presentation of antigen to T lymphocytes.

Ig Fc receptors—receptors found on the surface of certain cells including B cells, natural killer (NK) cells, macrophages, neutrophils, and mast cells. Fc receptors bind to antibodies that have attached to invading pathogen-infected cells. They stimulate cytotoxic cells to destroy microbe-infected cells through a mechanism known as antibody-dependent cell-mediated cytotoxicity (ADCC). Examples of important Fc receptors include CD16 (FcγRIIIa), CD23 (FcεR), CD32 (FcγRII), CD64 (FcγRI), and CD89 (FcαR).

Inflammasome—large cytoplasmic complexes of intracellular proteins that link the sensing of microbial products and cellular stress to the proteolytic activation of interleukin (IL)-1β and IL-18 inflammatory cytokines. Activation of molecules in the inflammasome is a key step in the response of the innate immune system for intracellular recognition of microbial and other danger signals in both health and pathologic states.

Innate immune system—ancient immune recognition system of host cells bearing germline-encoded pattern recognition receptors (PRRs) that recognize pathogens and trigger a variety of mechanisms of pathogen elimination. Cells of the innate immune system include NK cell lymphocytes, monocytes/macrophages, DCs, neutrophils, basophils, eosinophils, tissue mast cells, and epithelial cells.

Large granular lymphocytes—lymphocytes of the innate immune system with azurophilic cytotoxic granules that have NK cell activity capable of killing foreign and host cells with few or no self-MHC class I molecules.

Natural killer cells—large granular lymphocytes (LGLs) that kill target cells expressing few or no human leukocyte antigen (HLA) class I molecules, such as malignantly transformed cells and virally infected cells. NK cells express receptors that inhibit killer cell function when self-MHC class I is present.

NK T cells—innate-like lymphocytes that use an invariant T cell receptor (TCR)-α chain combined with a limited set of TCR-β chains and coexpress receptors commonly found on NK cells. NK T cells recognize lipid antigens of bacterial, viral, fungal, and protozoal infectious agents.

Pathogen-associated molecular patterns (PAMPs)—invariant molecular structures expressed by large groups of microorganisms that are recognized by host cellular PRRs in the mediation of innate immunity.

Pattern recognition receptors—germline-encoded receptors expressed by cells of the innate immune system that recognize PAMPs.

Polyreactive natural antibodies—preexisting low-affinity antibodies produced by B cells that cross-react with multiple antigens and are available at the time of infection to bind to and coat the invading pathogen and harness innate responses to slow the infection until an adaptive high-affinity protective antibody response can be made.

T lymphocytes—thymus-derived lymphocytes that mediate adaptive cellular immune responses including T helper, T regulatory, and cytotoxic T lymphocyte effector cell functions.

T cell exhaustion—state of T cells when the persistence of antigen disrupts memory T cell function, resulting in defects in memory T cell responses. Most frequently occurs in malignancies and in chronic viral infections such as HIV-1 and hepatitis C.

T cell receptor (TCR) for antigen—complex of surface molecules that rearrange during postnatal T cell development made up of clonotypic TCR-α and β chains that are associated with the CD3 complex composed of invariant γ, δ, ε, ζ, and η chains. TCR-α and β chains recognize peptide fragments of protein antigen physically bound in APC MHC class I or II molecules, leading to signaling via the CD3 complex to mediate effector functions.

T follicular helper T cells (Tfh)—CD4 T cells regulated by bcl-6 in B cell follicle germinal centers that produce IL-4 and IL-21 and drive B cell differentiation and affinity maturation in peripheral lymphoid tissues such as lymph node and spleen.

T_h1 T cells—CD4 helper T cell subset regulated by transcription factor T-bet and produces interferon (IFN)-gamma, IL-2 and TNF-beta and participates in cell mediated immunity.

T_h2 T cells—CD4 helper T cell subset regulated by transcription factors STAT6 and GATA3 that produces IL-4, IL-5, IL-6, IL-9, IL-10, and IL-13 and regulates antibody and eosinophil responses.

T regulatory cell (Treg)—CD4 and CD8 T cells regulated by the transcription factor Foxp3 that play roles in modulating the immune system to prevent deleterious immune activation. Expression of Foxp3 is a defining Treg marker.

T_H9 T cells—CD4 T cells regulated by the transcription factor PU.1 that secrete IL-9 and enhance inflammation in atopic disease and inflammatory bowel disease as well as mediate antitumor immunity.

T_H17 T cells—CD4 T cells regulated by the transcription factor RORγt that secrete IL-17, IL-22, and IL-26 and play roles in autoimmune inflammatory disorders as well as defend against bacterial and fungal pathogens.

Tolerance—B and T cell nonresponsiveness to antigens that results from encounter with foreign or self-antigens by B and T lymphocytes in the absence of expression of APC co-stimulatory molecules. Tolerance to antigens may be induced and maintained by multiple mechanisms either centrally (B cell deletion in the thymus for T cells or bone marrow for B cells) or peripherally (by cell deletion or anergy at sites throughout the peripheral immune system).

INTRODUCTION

The human immune system has evolved over millions of years from both invertebrate and vertebrate organisms to develop sophisticated defense mechanisms that protect the host from microbes and their virulence factors. The normal immune system has three key properties: A highly diverse repertoire of antigen receptors that enables recognition of a nearly infinite range of pathogens; immune memory, to mount rapid recall immune responses; and immunologic tolerance, to avoid immune damage to normal self-tissues. From invertebrates, humans have inherited the *innate immune system*, an ancient defense system that uses germline-encoded proteins to recognize pathogens. Cells of the innate immune system, such as macrophages, DCs, and NK lymphocytes, recognize PAMPs that are highly conserved among many microbes and use a diverse set of PRR molecules. Important components of the recognition of microbes by the innate immune system include recognition by germline-encoded host molecules, recognition of key microbe virulence factors but not recognition of self-molecules, and nonrecognition of benign foreign molecules or microbes. Upon contact with pathogens, macrophages and NK cells may kill pathogens directly or, in concert with DCs, may activate a series of events that both slow the infection and recruit the more recently evolved arm of the human immune system, the *adaptive immune system*.

Adaptive immunity is found only in vertebrates and is based on the generation of antigen receptors on T and B lymphocytes by gene rearrangements, such that individual T or B cells express unique antigen receptors on their surface capable of specifically recognizing diverse antigens of the myriad infectious agents in the environment. Coupled with finely tuned specific recognition mechanisms that maintain tolerance (nonreactivity) to self-antigens, T and B lymphocytes bring both *specificity* and *immune memory* to vertebrate host defenses.

This chapter describes the cellular components, key molecules **(Table 342-1)**, and mechanisms that make up the innate and adaptive immune systems and describes how adaptive immunity is recruited to the defense of the host by innate immune responses. An appreciation of the cellular and molecular bases of innate and adaptive immune responses is critical to understanding the pathogenesis of inflammatory, autoimmune, infectious, and immunodeficiency diseases.

TABLE 342-1

Human Leukocyte Surface Antigens—The CD Classification of Leukocyte Differentiation Antigens

Surface Antigen (Other Names)	Family	Molecular Mass, kDa	Distribution	Ligand(s)	Function
CD1a (T6, HTA-1)	Ig	49	CD, cortical thymocytes, Langerhans type of DCs	TCRγδ T cells	CD1 molecules present lipid antigens of intracellular bacteria such as *Mycobacterium leprae* and *M. tuberculosis* to TCRγδT cells
CD1b	Ig	45	CD, cortical thymocytes, Langerhans type of DCs	TCRγδ T cells	
CD1c	Ig	43	DC, cortical thymocytes, subset of B cells, Langerhans type of DCs	TCRγδ T cells	
CD1d	Ig	37	Cortical thymocytes, intestinal epithelium, Langerhans type of DCs	TCRγδ T cells	
CD2 (T12, LFA-2)	Ig	50	T, NK	CD58, CD48, CD59, CD15	Alternative T cell activation, T cell anergy, T cell cytokine production, T- or NK-mediated cytolysis, T cell apoptosis, cell adhesion
CD3 (T3, Leu-4)	Ig	γ:25–28, δ:21–28, ε:20–25, η:21–22, ζ:16	T	Associates with the TCR	T cell activation and function; ζ is the signal transduction component of the CD3 complex
CD4 (T4, Leu-3)	Ig	55	T, myeloid	MHC-II, HIV, gp120, IL-16, SABP	T cell selection, T cell activation, signal transduction with p56*lck*, primary receptor for HIV
CD7 (3A1, Leu-9)	Ig	40	T, NK	K-12 (CD7L)	T and NK cell signal transduction and regulation of IFN-γ, TNF-α production
CD8 (T8, Leu-2)	Ig	34	T	MHC-I	T cell selection, T cell activation, signal transduction with p56*lck*
CD14 (LPS-receptor)	LRG	53–55	M, G (weak), not by myeloid progenitors	Endotoxin (lipopolysaccharide), lipoteichoic acid, PI	TLR4 mediates with LPS and other PAMP activation of innate immunity
CD16a (FcγRIIIa)	Ig	50–80	NK, macrophages, neutrophils	Fc portion of IgG	Mediates phagocytosis and ADCC
CD19 B4	Ig	95	B (except plasma cells), FDC	Not known	Associates with CD21 and CD81 to form a complex involved in signal transduction in B cell development, activation, and differentiation
CD20 (B1)	Unassigned	33–37	B (except plasma cells)	Not known	Cell signaling, may be important for B cell activation and proliferation
CD21 (B2, CR2, EBV-R, C3dR)	RCA	145	Mature B, FDC, subset of thymocytes	C3d, C3dg, iC3b, CD23, EBV	Associates with CD19 and CD81 to form a complex involved in signal transduction in B cell development, activation, and differentiation; Epstein-Barr virus receptor

Surface Antigen (Other Names)	Family	Molecular Mass, kDa	Distribution	Ligand(s)	Function
CD22 (BL-CAM)	Ig	130–140	Mature B	CDw75	Cell adhesion, signaling through association with p72*sky*, p53/56*lyn*, PI3 kinase, SHP1, fLCγ
CD23 (FcεRII, B6, Leu-20, BLAST-2)	C-type lectin	45	B, M, FDC	IgE, CD21, CD11b, CD11c	Regulates IgE synthesis, cytokine release by monocytes
CD28	Ig	44	T, plasma cells	CD80, CD86	Co-stimulatory for T cell activation; involved in the decision between T cell activation and anergy
CD32a (FcγRIIa)	Ig	40	NK, macrophages, neutrophils	Fc portion of IgG	Mediates phagocytosis and ADCC
CD40	TNFR	48–50	B, DC, EC, thymic epithelium, MP, cancers	CD154	B cell activation, proliferation, and differentiation; formation of GCs; isotype switching; rescue from apoptosis
CD45 (LCA, T200, B220)	PTP	180, 200, 210, 220	All leukocytes	Galectin-1, CD2, CD3, CD4	T and B activation, thymocyte development, signal transduction, apoptosis
CD45RA	PTP	210, 220	Subset T, medullary thymocytes, “naive” T	Galectin-1, CD2, CD3, CD4	Isoforms of CD45 containing exon 4 (A), restricted to a subset of T cells
CD45RB	PTP	200, 210, 220	All leukocytes	Galectin-1, CD2, CD3, CD4	Isoforms of CD45 containing exon 5 (B)
CD45RC	PTP	210, 220	Subset T, medullary thymocytes, “naive” T	Galectin-1, CD2, CD3, CD4	Isoforms of CD45 containing exon 6 (C), restricted to a subset of T cells
CD45RO	PTP	180	Subset T, cortical thymocytes, “memory” T	Galectin-1, CD2, CD3, CD4	Isoforms of CD45 containing no differentially spliced exons, restricted to a subset of T cells
CD64 (FcγRI)	Ig	45–55	Macrophages and monocytes	Fc portion of IgG	Mediates phagocytosis and ADCC
CD80 (B7-1, BB1)	Ig	60	Activated B and T, MP, DC	CD28, CD152	Co-regulator of T cell activation; signaling through CD28 stimulates and through CD152 inhibits T cell activation
CD86 (B7-2, B70)	Ig	80	Subset B, DC, EC, activated T, thymic epithelium	CD28, CD152	Co-regulator of T cell activation; signaling through CD28 stimulates and through CD152 inhibits T cell activation
CD89 (FCαR)	Ig	55–100	Neutrophils, eosinophils, monocytes, and MP	Fc portion of IgG	Mediates phagocytosis and ADCC of IgA-coated pathogens
CD95 (APO-1, Fas)	TNFR	43	Activated T and B	Fas ligand	Mediates apoptosis
CD152 (CTLA-4)	Ig	30–33	Activated T	CD80, CD86	Inhibits T cell proliferation

Surface Antigen (Other Names)	Family	Molecular Mass, kDa	Distribution	Ligand(s)	Function
CD154 (CD40L)	TNF	33	Activated CD4+ T, subset CD8+ T, NK, M, basophil	CD40	Co-stimulatory for T cell activation, B cell proliferation and differentiation
CD279 (PD-1)	Ig	50–55	B, T, TfH	PD-L1, PD-L2	Inhibits T cell proliferation

Abbreviations: ADCC, antibody-dependent cell-mediated cytotoxicity; CTLA, cytotoxic T lymphocyte–associated protein; DC, dendritic cells; EBV, Epstein-Barr virus; EC, endothelial cells; ECM, extracellular matrix; Fcγ RIII, low-affinity IgG receptor isoform A; FDC, follicular dendritic cells; G, granulocytes; GC, germinal center; GPI, glycosyl phosphatidylinositol; HTA, human thymocyte antigen; Ig, immunoglobulin; IgG, immunoglobulin G; LCA, leukocyte common antigen; LPS, lipopolysaccharide; MHC-I, major histocompatibility complex class I; MP, macrophages; Mr, relative molecular mass; NK, natural killer cells; P, platelets; PBT, peripheral blood T cells; PD-1, programmed cell death-1; PI, phosphatidylinositol; PI3K, phosphatidylinositol 3-kinase; PLC, phospholipase C; PTP, protein tyrosine phosphatase; TCR, T cell receptor; TfH, T follicular helper cells; TNF, tumor necrosis factor; TNFR, tumor necrosis factor receptor. For an expanded list of cluster of differentiation (CD) human antigens, see Harrison's Online at *http://www.accessmedicine.com.ezproxy1.library.usyd.edu.au*; and for a full list of CD human antigens from the most recent Human Workshop on Leukocyte Differentiation Antigens (VII), see *http://mpr.nci.nih.gov.ezproxy1.library.usyd.edu.au/prow/*.

Source: Compiled from T Kishimoto et al (eds): *Leukocyte Typing VI.* New York: Garland Publishing, 1997; R Brines et al: Immunology Today 18S:1, 1997; and S Shaw (ed): Protein reviews on the Web. *http://mpr.nci.nih.gov.ezproxy1.library.usyd.edu.au/prow/*.

THE INNATE IMMUNE SYSTEM

All multicellular organisms, including humans, have developed the use of a limited number of surface and intracellular germline-encoded molecules that recognize pathogens. Because of the myriad of human pathogens, host molecules of the human innate immune system sense "danger signals" and either recognize PAMPs, the common molecular structures shared by many pathogens, or recognize host cell molecules produced in response to infection such as heat shock proteins and fragments of the extracellular matrix. PAMPs must be conserved structures vital to pathogen virulence and survival, such as bacterial endotoxin, so that pathogens cannot mutate molecules of PAMPs to evade human innate immune responses. PRRs are host proteins of the innate immune system that recognize PAMPs as host danger signal molecules **(Tables 342-2 and 342-3)**. Thus, recognition of pathogen molecules by hematopoietic and nonhematopoietic cell types leads to activation/production of the complement cascade, cytokines, and antimicrobial peptides as effector molecules. In addition, pathogen PAMPs as host danger signal molecules activate DCs to mature and to express molecules on the DC surface that optimize antigen presentation to respond to foreign antigens.

TABLE 342-2

Major Components of the Innate Immune System

Pattern recognition receptors (PRRs)	Toll-like receptors (TLRs), C-type lectin receptors (CLRs), retinoic acid–inducible gene (RIG)-1-like receptors (RLRs), and NOD-like receptors (NLRs)
Antimicrobial peptides	α-Defensins, β-defensins, cathelin, protegrin, granulysin, histatin, secretory leukoprotease inhibitor, and probiotics
Cells	Macrophages, dendritic cells, NK cells, NK-T cells, neutrophils, eosinophils, mast cells, basophils, and epithelial cells
Complement components	Classic and alternative complement pathway, and proteins that bind complement components
Cytokines	Autocrine, paracrine, endocrine cytokines that mediate host defense and inflammation, as well as recruit, direct, and regulate adaptive immune responses

Abbreviation: NK, natural killer.

TABLE 342-3

Pattern Recognition Receptors (PRRs) and Their Ligands

PRRs	Localization	Ligand	Origin of the Ligand
TLR			
TLR1	Plasma membrane	Triacyl lipoprotein	Bacteria
TLR2	Plasma membrane	Lipoprotein	Bacteria, viruses, parasite, self
TLR3	Endolysosome	dsRNA	Virus
TLR4	Plasma membrane	LPS	Bacteria, viruses, self
TLR5	Plasma membrane	Flagellin	Bacteria
TLR6	Plasma membrane	Diacyl lipoprotein	Bacteria, viruses
TLR7 (human TLR8)	Endolysosome	ssRNA	Virus, bacteria, self
TLR9	Endolysosome	CpG-DNA	Virus, bacteria, protozoa, self
TLR10	Endolysosome	Unknown	Unknown
TLR11	Plasma membrane	Profilin-like molecule	Protozoa
RLR			
RIG-I	Cytoplasm	Short dsRNA, triphosphate dsRNA	RNA viruses, DNA virus
MDA5	Cytoplasm	Long dsRNA	RNA viruses (Picornaviridae)
LGP2	Cytoplasm	Unknown	RNA viruses
NLR			
NOD1	Cytoplasm	iE-DAP	Bacteria
NOD2	Cytoplasm	MDP	Bacteria
CLR			
Dectin-1	Plasma membrane	β^2-Glucan	Fungi
Dectin-2	Plasma membrane	β^2-Glucan	Fungi
MINCLE	Plasma membrane	SAP130	Self, fungi

Abbreviations: CLR, C-type lectin receptors; dsRNA, double-strand RNA; iE-DAP, D-glutamyl-meso-diaminopimelic acid moiety; LGP2, Laboratory of Genetics and Physiology 2 protein encoded by the gene *DHX58*; MDA5, melanoma differentiation-associated protein 5; MDP, MurNAc-L-Ala-D-isoGln, also known as muramyl dipeptide; MINCLE, macrophage-inducible C-type lectin; NLR, NOD-like receptor; NOD, NOTCH protein domain; RIG, retinoic acid–inducible gene; RLR, RIG-like receptors; TLR, Toll-like receptor.

Source: Adapted from O Takeuchi, S Akira: Cell 140:805, 2010, with permission.

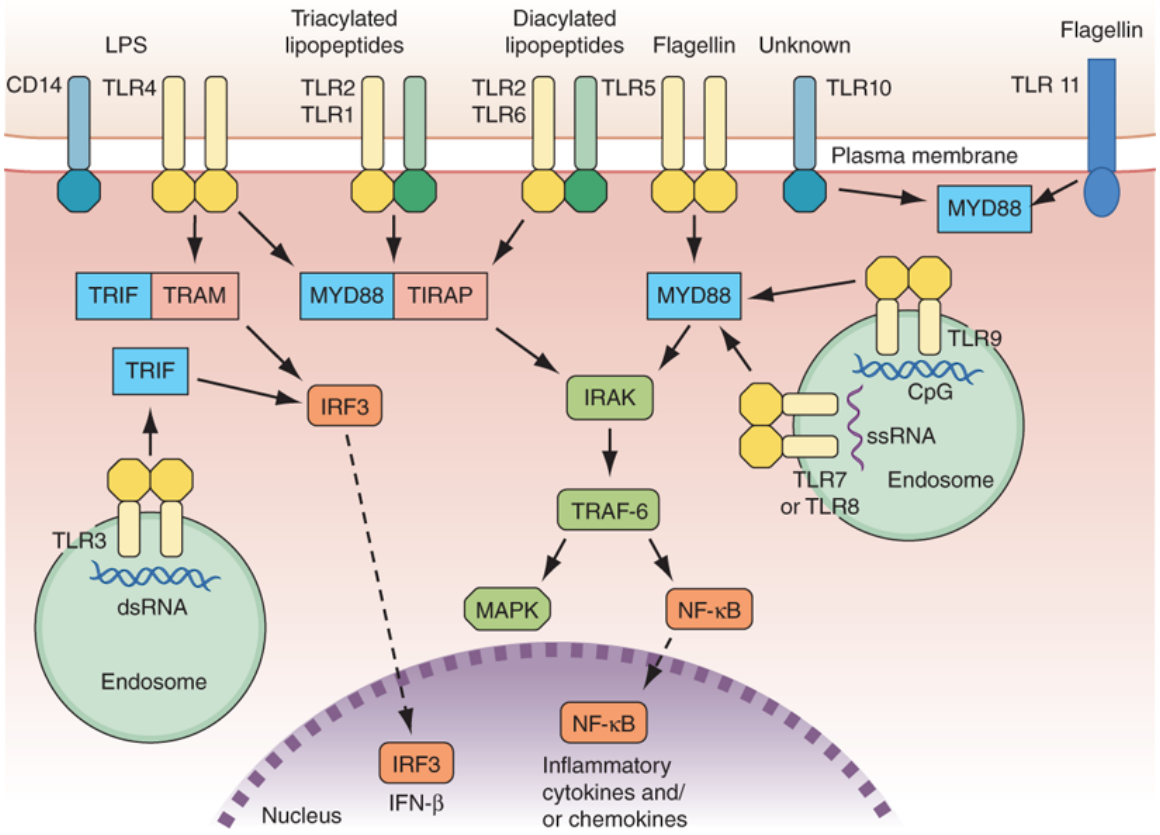
PATTERN RECOGNITION

Major PRR families of proteins include transmembrane proteins, such as the Toll-like receptors (TLRs) and C-type lectin receptors (CLRs), and cytoplasmic proteins, such as the retinoic acid–inducible gene (RIG)-1-like receptors (RLRs) and NOD-like receptors (NLRs) (Table 342-3). A major group of PRR collagenous glycoproteins with C-type lectin domains are termed *collectins* and include the serum protein mannose-binding lectin (MBL). MBL and other collectins, as well as two other protein families—the pentraxins (such as C-reactive protein and serum amyloid P) and macrophage scavenger receptors—all have the property of opsonizing (coating) bacteria for phagocytosis by macrophages and can also activate the complement cascade to lyse bacteria. Integrins are cell-surface adhesion molecules that affect attachment between cells and the extracellular matrix and mediate signal transduction that reflects the chemical composition of the cell environment. For example, integrins signal after cells bind bacterial lipopolysaccharide (LPS) and activate phagocytic cells to ingest pathogens.

There are multiple connections between the innate and adaptive immune systems; these include (1) a plasma protein, LPS-binding protein, that binds and transfers LPS to the macrophage LPS receptor, CD14; (2) a human family of proteins called *Toll-like receptor proteins* (TLRs), some of which are associated with CD14, bind LPS, and signal epithelial cells, DCs, and macrophages to produce cytokines and upregulate cell-surface molecules that signal the initiation of adaptive immune responses (**Fig. 342-1**, Tables 342-3 and 342-4), and (3) families of intracellular microbial sensors called NLRs and RLRs. Proteins in the Toll family can be expressed on macrophages, DCs, and B cells as well as on a variety of nonhematopoietic cell types, including respiratory epithelial cells. Eleven TLRs have been identified in humans, and 13 TLRs have been identified in mice **(Tables 342-4 and 342-5)**. Upon ligation, TLRs activate a series of intracellular events that lead to the killing of bacteria- and viral-infected cells as well as to the recruitment and ultimate activation of antigen-specific T and B lymphocytes (Fig. 342-1). Importantly, signaling by massive amounts of LPS through TLR4 leads to the release of large amounts of cytokines that mediate LPS-induced shock. Mutations in TLR4 proteins in mice protect from LPS shock, and TLR mutations in humans protect from LPS-induced inflammatory diseases such as LPS-induced asthma (Fig. 342-1).

FIGURE 342-1

Overview of major TLR signaling pathways. All TLRs signal through MYD88, with the exception of TLR3. TLR4 and the TLR2 subfamily (TLR1, TLR2, TLR6) also engage TIRAP (Toll-interleukin 1 receptor domain-containing adapter protein). TLR3 signals through TRIF (Toll-interleukin 1 receptor domain-containing adapter-inducing interferon-β). TRIF is also used in conjunction with TRAM (TRIF-related adaptor molecule) in the TLR4-MYD88-independent pathway. *Dashed arrows* indicate translocation into the nucleus. dsRNA, double-strand RNA; IFN, interferon; IRF3, interferon regulatory factor 3; LPS, lipopolysaccharide; MAPK, mitogen-activated protein kinases; NF-κB, nuclear factor-κB; ssRNA, single-strand RNA; TLR, Toll-like receptor. *(Adapted from D van Duin et al: Trends Immunol 27:49, 2006, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 342-4

The Role of Pattern Recognition Receptors (PRRs) in Modulation of Adaptive Immune Responses

PRR Family	PRRs	Ligand	DC or Macrophage Cytokine Response	Adaptive Immune Response
TLRs	TLR2 (heterodimer with TLR1 or 6)	Lipopeptides	Low IL-12p70	T_H1
		Pam-3-cys (TLR 2/1)	High IL-10	T_H2
		MALP (TLR 2/6)	IL-6	T regulatory
	TLR3	dsRNA	IL-12p70	T_H1
			IFN-α	
			IL-6	
	TLR4	*Escherichia coli* LPS	High IL-12p70	T_H1
			Intermediate IL-10	
			IL-6	
	TLR5	Flagellin	High IL-12p70	T_H1
			Low IL-12p70	T_H2
	TLR7/8	ssRNA	High IL-12p70	T_H1
		Imidazoquinolines	IFN-α	
			IL-6	
	TLR9	CpG DNA	High IL-12p70	T_H1
			Low IL-10	
			IL-6	
			IFN-α	
	TLR10	?	?	?
	TLR11	Profilin-like molecule uropathogenic bacteria	?	?
C-type lectins	DC-SIGN	Env of HIV; core protein of HCV; components of *Mycobacterium tuberculosis*; *Helicobacter pylori*, Lewis Ag	*H. pylori*, Lewis Ag	T_H2
			Suppresses IL-12p70	T regulatory
			Suppresses TLR signaling in DCs	
NOD	NOD2	Muramyl dipeptide of peptidoglycan of bacteria	Induces IL-10 in DCs	Weak T cell response (tolerogenic?)
Mannose receptor	Mannose receptor	Mannosylated lipoarabinomannans from bacillus Calmette-Guérin and *M. tuberculosis*	Suppresses IL-12 and TLR signaling in DCs	Weak T cell response? (tolerogenic?)

Abbreviations: CpG, sequences in DNA recognized by TLR-9; DC, dendritic cell; DC-SIGN, DC-specific C-type lectin; dsRNA, double-strand RNA; HCV, hepatitis C virus; HIV, human immunodeficiency virus; LPS, lipopolysaccharide; MALP, macrophage-activating lipopeptide; NOD, NOTCH protein domain; ssRNA, single-strand RNA; T_H1, helper T cell; T_H2, helper T cell; TLR, Toll-like receptor.

Source: B Pulendran: J Immunol 174:2457, 2005. Copyright 2005 The American Association of Immunologists, Inc., with permission.

TABLE 342-5

Cells of the Innate Immune System and Their Major Roles in Triggering Adaptive Immunity

Cell Type	Major Role in Innate Immunity	Major Role in Adaptive Immunity
Macrophages	Phagocytose and kill bacteria; produce antimicrobial peptides; bind LPS; produce inflammatory cytokines	Produce IL-1 and TNF-α to upregulate lymphocyte adhesion molecules and chemokines to attract antigen-specific lymphocyte. Produce IL-12 to recruit T_H1 T helper cell responses; upregulate co-stimulatory and MHC molecules to facilitate T and B lymphocyte recognition and activation. Macrophages and dendritic cells, after LPS signaling, upregulate co-stimulatory molecules B7-1 (CD80) and B7-2 (CD86) that are required for activation of pathogen-specific T cells. There are also Toll-like proteins on B cells and dendritic cells that, after LPS ligation, induce CD80 and CD86 on these cells for T cell antigen presentation.
Plasmacytoid dendritic cells (DCs) of lymphoid lineage	Produce large amounts of interferon-α (IFN-α), which has antitumor and antiviral activity, and are found in T cell zones of lymphoid organs; they circulate in blood	IFN-α is a potent activator of macrophage and mature DCs to phagocytose invading pathogens and present pathogen antigens to T and B cells
Myeloid dendritic cells are of two types: interstitial and Langerhans-derived	Interstitial DCs are strong producers of IL-12 and IL-10 and are located in T cell zones of lymphoid organs, circulate in blood, and are present in the interstices of the lung, heart, and kidney; Langerhans DCs are strong producers of IL-12; are located in T cell zones of lymph nodes, skin epithelia, and the thymic medulla; and circulate in blood	Interstitial DCs are potent activators of macrophage and mature DCs to phagocytose invading pathogens and present pathogen antigens to T and B cells
Natural killer (NK) cells	Kill foreign and host cells that have low levels of MHC+ self-peptides. Express NK receptors that inhibit NK function in the presence of high expression of self-MHC.	Produce TNF-α and IFN-γ, which recruit T_H1 helper T cell responses
NK-T cells	Lymphocytes with both T cell and NK surface markers that recognize lipid antigens of intracellular bacteria such as *Mycobacterium tuberculosis* by CD1 molecules and kill host cells infected with intracellular bacteria.	Produce IL-4 to recruit T_H2 helper T cell responses, IgG1 and IgE production
Neutrophils	Phagocytose and kill bacteria, produce antimicrobial peptides	Produce nitric oxide synthase and nitric oxide, which inhibit apoptosis in lymphocytes and can prolong adaptive immune responses
Eosinophils	Kill invading parasites	Produce IL-5, which recruits Ig-specific antibody responses
Mast cells and basophils	Release TNF-α, IL-6, and IFN-γ in response to a variety of bacterial PAMPs	Produce IL-4, which recruits T_H2 helper T cell responses, and recruit IgG1- and IgE-specific antibody responses
Epithelial cells	Produce antimicrobial peptides; tissue-specific epithelia produce mediator of local innate immunity; e.g., lung epithelial cells produce surfactant proteins (proteins within the collectin family) that bind and promote clearance of lung-invading microbes	Produces TGF-β, which triggers IgA-specific antibody responses

Abbreviations: IL-4, IL-5, IL-6, IL-10, and IL-12, interleukin 4, 5, 6, 10, and 12, respectively; MHC, major histocompatibility complex: LPS, lipopolysaccharide; PAMP, pathogen-associated molecular patterns; T_H, helper T cell; TNF-α, tumor necrosis factor-alpha.

Source: Adapted from R Medzhitov, CA Janeway: Curr Opinion Immunol 9:4, 1997. Copyright 1997, with permission from Elsevier.

Two other families of cytoplasmic PRRs are the NLRs and the RLRs. These families, unlike the TLRs, are composed primarily of soluble intracellular proteins that scan host cell cytoplasm for intracellular pathogens (Tables 342-2 and 342-3).

The intracellular microbial sensors, NLRs, after triggering, form large cytoplasmic complexes termed *inflammasomes*, which are aggregates of molecules including NOD-like receptor pyrin (NLRP) proteins (Table 342-3). Inflammasomes activate inflammatory caspases and IL-1β in the presence of nonbacterial danger signals (cell stress) and bacterial PAMPs. Mutations in inflammasome proteins can lead to chronic inflammation in a group of periodic febrile diseases called *autoinflammatory syndromes* **(Table 342-6)**. Inflammasomes are activated upon sensing of PAMPs. Crystallopathies are diseases caused by tissue crystal deposition such as monosodium urate that can activate the inflammasome and, in the case of urate deposition, can lead to gout with arthritis or renal disease.

TABLE 342-6

Diseases Associated with Inflammasome Activity

Disease	Clinical Features	Gene Mutated	Etiologic Agent	Inflammasome Involvement	Anakinra Response[a]
Familial cold autoinflammatory syndrome (FCAS)	Fever, arthralgia, cold-induced urticaria	*NALP3*		Overactive	Yes
Muckle-Wells syndrome (MWS)	Fever, arthralgia, urticaria, sensorineural deafness, amyloidosis	*NLRP3*		Overactive	Yes
Chronic infantile neurologic cutaneous and articular syndrome (CINCA, NOMID)	Fever, severe arthralgia, urticaria, neurologic problems, severe amyloidosis	*NALP3*		Overactive	Yes
Familial Mediterranean fever (FMF)	Fever, peritonitis, pleuritis, amyloidosis	Pyrin		Overactive	Partial
Pyogenic arthritis, pyoderma gangrenosum, and acne syndrome (PAPA)	Pyogenic sterile arthritis	*PSTPIP1*		Overactive	Yes
Hyperimmunoglobulin D syndrome (HIDS)	Arthralgia, abdominal pain, lymphadenopathy	Mevalonate kinase		To be demonstrated	Yes
Tumor necrosis factor receptor-1–associated syndrome (TRAPS)	Fever, abdominal pain, skin lesions	*TNF-R1*		To be demonstrated	Yes
Systemic-onset juvenile idiopathic arthritis (SOJIA)	Chronic joint inflammation		Unknown	To be demonstrated	Yes
Adult-onset Still's disease (AOSD)	Arthralgia, fever		Unknown	To be demonstrated	Yes
Behçet's disease	Arthralgia, uveitis, ulcers		Unknown	To be demonstrated	Yes
Schnitzler's syndrome	Urticaria, fever, arthralgia		Unknown	To be demonstrated	Yes
Gout	Metabolic arthritis, pain		Uric acid (MSU)	Activated	Yes
Pseudogout	Arthritis		CPPD	Activated	Yes
Contact dermatitis	Urticaria		Irritants	Activated	Unknown
Fever syndrome	Fever	*NALP12*		Unknown	Unknown
Hydatidiform mole	Hydatid mole	*NALP7*		Unknown	Unknown
Vitiligo	Skin depigmentation, autoimmunity	*NLRP1*		Overactive	Unknown
Crohn's disease		*NLRP3*		Underactive	Unknown
Multiple Sclerosis		*NLPR3*		Activated	Unknown
Psoriasitic arthritis		*NLRP3*		Activated	Yes

[a]Anakinra is a recombinant interleukin-1 (IL-1) receptor antagonist that functions to block the biologic activity of naturally occurring IL-1.

Abbreviation: CPPD, calcium pyrophosphate dehydrate.

Source: Adapted from F Martinon et al: Ann Rev Immunol 27:229, 2009. Copyright 2009. Reproduced with permission from Annual Reviews Inc.

EFFECTOR CELLS OF INNATE IMMUNITY

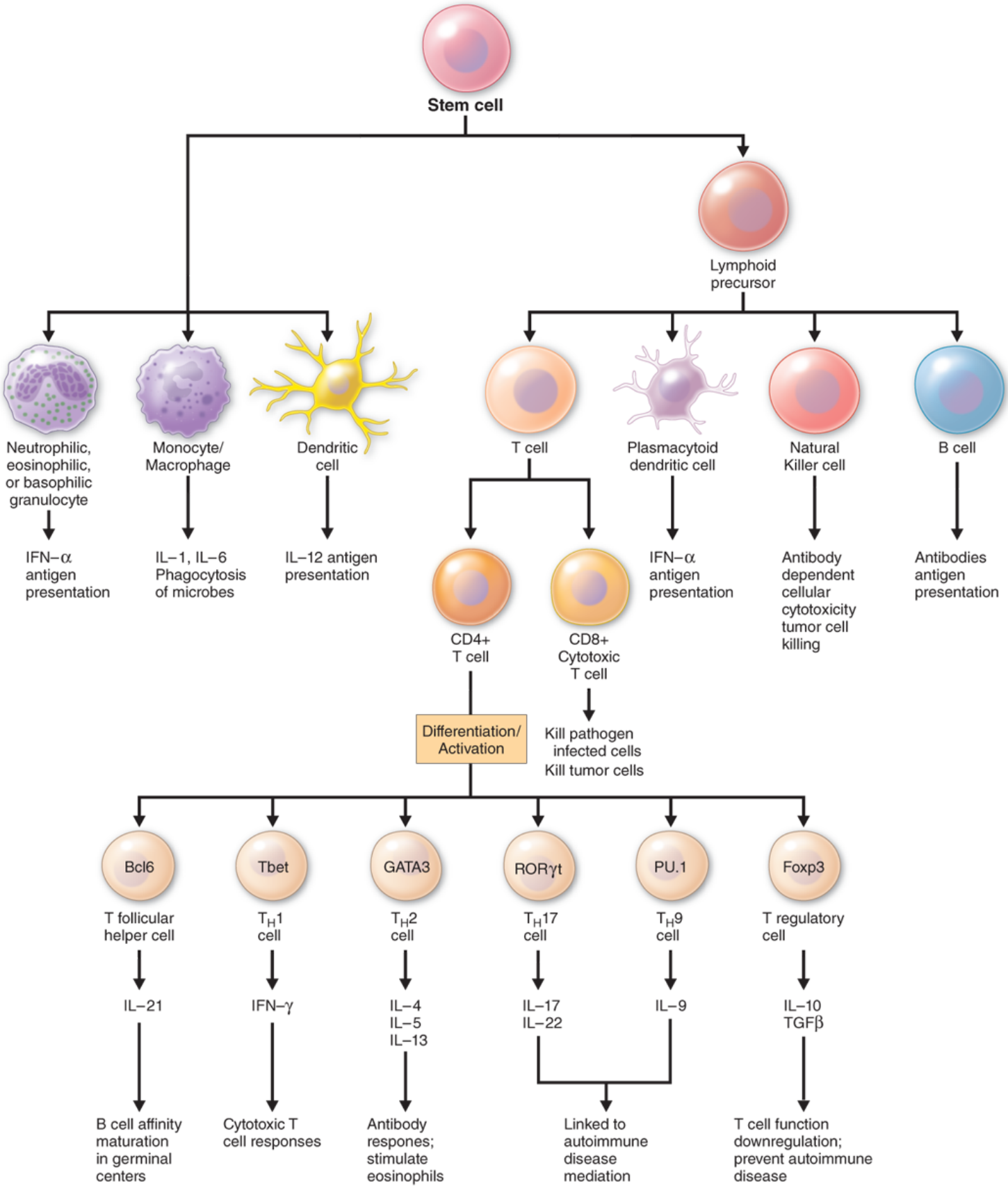
Cells of the innate immune system and their roles in the first line of host defense are listed in Table 342-5. Equally important as their roles in the mediation of innate immune responses are the roles that each cell type plays in recruiting T and B lymphocytes of the adaptive immune system to engage in specific pathogen responses.

Monocytes-Macrophages

Monocytes arise from precursor cells within bone marrow **(Fig. 342-2)** and circulate with a half-life ranging from 1 to 3 days. Monocytes leave the peripheral circulation via capillaries and migration into a vast extravascular cellular pool. Tissue macrophages arise from monocytes that have migrated out of the circulation and by in situ proliferation of macrophage precursors in tissue. Common locations where tissue macrophages (and certain of their specialized forms) are found are lymph node, spleen, bone marrow, perivascular connective tissue, serous cavities such as the peritoneum, pleura, skin connective tissue, lung (alveolar macrophages), liver (Kupffer cells), bone (osteoclasts), central nervous system (microglia cells), and synovium (type A lining cells).

FIGURE 342-2

Model of immune effector cell development. Hematopoietic stem cells differentiate into T cells, antigen-presenting dendritic cells, natural killer cells, macrophages, granulocytes, or B cells. Foreign antigen is processed by dendritic cells, macrophages and B cells, and peptide fragments of foreign antigen are presented to CD4+ and/or CD8+ T cells. CD8+ T cell activation leads to induction of cytotoxic T lymphocyte (CTL) or killer T cell generation, as well as induction of cytokine-producing CD8+ cytotoxic T cells. Granulocytes (neutrophils, eosinophils, or basophils) are effector cells of the innate immune system and mediate anti-infectious agent activity by cytokine production, infectious agent killing or both. T_H1 CD4+ T cells play an important role in defense against intracellular microbes and help in the generation of CD8+ cytotoxic T cells. T_H2 CD4+ T cells producing (IFN) γ or IL-4, IL-5, IL-13 regulate Ig class switching and determine the type of antibody produced. T_H17 cells secrete IL-17 and IL-22, and Th9 cells secrete IL-9. Both are linked to mediation of autoimmune disease. CD4+ T regulatory cells produce IL-10 and TGFβ and downregulate T and B cell responses once the microbe has been eliminated. Each of the types of CD4+ T cells are regulated by different transcription factors and the key transcription factors are shown in the circles above each CD4+ T cell type.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In general, monocytes-macrophages are on the first line of defense associated with innate immunity and ingest and destroy microorganisms through the release of toxic products such as hydrogen peroxide (H_2O_2) and nitric oxide (NO). Inflammatory mediators produced by macrophages attract additional effector cells such as neutrophils to the site of infection. Macrophage mediators include prostaglandins; leukotrienes; platelet activating factor; cytokines such as IL-1, TNF-α, IL-6, and IL-12; and chemokines **(Tables 342-7,342-8,342-9).**

TABLE 342-7

Cytokines and Cytokine Receptors

Cytokine	Receptor	Cell Source	Cell Target	Biologic Activity
IL-1α, β	Type I IL-1r, Type II IL-1r	Monocytes/macrophages, B cells, fibroblasts, most epithelial cells including thymic epithelium, endothelial cells	All cells	Upregulates adhesion molecule expression, neutrophil and macrophage emigration, mimics shock, fever, upregulates hepatic acute-phase protein production, facilitates hematopoiesis
IL-2	IL-2r α, β, common γ	T cells	T cells, B cells, NK cells, monocytes-macrophages	Promotes T cell activation and proliferation, B cell growth, NK cell proliferation and activation, enhanced monocyte/macrophage cytolytic activity
IL-3	IL-3r, common β	T cells, NK cells, mast cells	Monocytes-macrophages, mast cells, eosinophils, bone marrow progenitors	Stimulates hematopoietic progenitors
IL-4	IL-4r α, common γ	T cells, mast cells, basophils	T cells, B cells, NK cells, monocytes-macrophages, neutrophils, eosinophils, endothelial cells, fibroblasts	Stimulates T_H2 helper T cell differentiation and proliferation; stimulates B cell Ig class switch to IgG1 and IgE anti-inflammatory action on T cells, monocytes; produced by T follicular helper cells in B cell germinal centers that stimulate B cell maturation.
IL-5	IL-5r α, common γ	T cells, mast cells, eosinophils	Eosinophils, basophils, murine B cells	Regulates eosinophil migration and activation
IL-6	IL-6r, gp130	Monocytes-macrophages, B cells, fibroblasts, most epithelium including thymic epithelium, endothelial cells	T cells, B cells, epithelial cells, hepatocytes, monocytes-macrophages	Induces acute-phase protein production, T and B cell differentiation and growth, myeloma cell growth, and osteoclast growth and activation
IL-7	IL-7r α, common γ	Bone marrow, thymic epithelial cells	T cells, B cells, bone marrow cells	Differentiates B, T, and NK cell precursors, activates T and NK cells
IL-8	CXCR1, CXCR2	Monocytes-macrophages, T cells, neutrophils, fibroblasts, endothelial cells, epithelial cells	Neutrophils, T cells, monocytes-macrophages, endothelial cells, basophils	Induces neutrophil, monocyte, and T cell migration, induces neutrophil adherence to endothelial cells and histamine release from basophils, and stimulates angiogenesis; suppresses proliferation of hepatic precursors
IL-9	IL-9r α, common γ	T cells	Bone marrow progenitors, B cells, T cells, mast cells	Induces mast cell proliferation and function, synergizes with IL-4 in IgG and IgE production and T cell growth, activation, and differentiation
IL-10	IL-10r	Monocytes-macrophages, T cells, B cells, keratinocytes, mast cells	Monocytes-macrophages, T cells, B cells, NK cells, mast cells	Inhibits macrophage proinflammatory cytokine production, downregulates cytokine class II antigen and B7-1 and B7-2 expression, inhibits differentiation of T_H1 helper T cells, inhibits NK cell function, stimulates mast cell proliferation and function, B cell activation, and differentiation
IL-11	IL-11r α, gp130	Bone marrow stromal cells	Megakaryocytes, B cells, hepatocytes	Induces megakaryocyte colony formation and maturation, enhances antibody responses, stimulates acute-phase protein production

Cytokine	Receptor	Cell Source	Cell Target	Biologic Activity
IL-12 (35-kDa and 40-kDa subunits)	IL-12r	Activated macrophages, dendritic cells, neutrophils	T cells, NK cells	Induces T_H1 T helper cell formation and lymphokine-activated killer cell formation; increases CD8+ CTL cytolytic activity; ↓IL-17, ↑IFN-γ
IL-13	IL-13r/IL-4r α	T cells (T_H2)	Monocytes-macrophages, B cells, endothelial cells, keratinocytes	Upregulates VCAM-1 and C-C chemokine expression on endothelial cells and B cell activation and differentiation, and inhibits macrophage proinflammatory cytokine production
IL-14	Unknown	T cells	Normal and malignant B cells	Induces B cell proliferation, inhibits antibody secretion, and expands selected B cell subgroups
IL-15	IL-15r α, common γ, IL2r β	Monocytes-macrophages, epithelial cells, fibroblasts	T cells, NK cells	Promotes T cell activation and proliferation, angiogenesis, and NK cells
IL-16	CD4	Mast cells, eosinophils, CD8+ T cells, respiratory epithelium	CD4+ T cells, monocytes-macrophages, eosinophils	Promotes chemoattraction of CD4+ T cells, monocytes, and eosinophils; inhibits HIV replication; inhibits T cell activation through CD3/T cell receptor
IL-17	IL-17r	CD4+ T cells	Fibroblasts, endothelium, epithelium, macrophages	Enhances cytokine/chemokine secretion; promotes delayed-type reactions
IL-18	IL-18r (IL-1R-related protein)	Keratinocytes, macrophages	T cells, B cells, NK cells	Upregulates IFN-γ production, enhances NK cell cytotoxicity
IL-21	IL-δγ chain/IL-21R	CD4 T cells	NK cells	Downregulates NK cell–activating molecules, NKG2D/DAP10; produced by T follicular helper cells in B cell germinal centers that stimulate B cell maturation.
IL-22	IL-22 R1/IL-10R2	DC, T cells	Epithelial cells	Innate responses against bacterial pathogens; promotes hepatocyte survival
IL-23	IL-12Rb1/IL23R	Macrophages, other cell types	T cells	Opposite effects of IL-12 (↑IL-17, ↑IFN-γ)
IL-24	IL-20 R1/IL-20R2 IL-22R1/IL-20 R2	Macrophages, T_H2 cells	Nonhematopoietic cells such as fibroblasts	Promotes wound healing
IL-25 (also called IL-17E)	IL-17RB	CD4 T cells, mast cells	Fibroblasts, endothelium, epithelium, macrophages	Proinflammatory; induces cytokine production
IL-26	IL-20 R1/IL-10R2	T_H1, T_H17 T cells, synovial cells	Epithelial cells	Proinflammatory; induces cytokine production
IFN-α	Type I interferon receptor	All cells	All cells	Promotes antiviral activity; stimulates T cell, macrophage, and NK cell activity; direct antitumor effects; upregulates MHC class I antigen expression; used therapeutically in viral and autoimmune conditions
IFN-β	Type I interferon receptor	All cells	All cells	Antiviral activity; stimulates T cell, macrophage, and NK cell activity; direct antitumor effects; upregulates MHC class I antigen expression; used therapeutically in viral and autoimmune conditions

Cytokine	Receptor	Cell Source	Cell Target	Biologic Activity
IFN-γ	Type II interferon receptor	T cells, NK cells	All cells	Regulates macrophage and NK cell activations; stimulates immunoglobulin secretion by B cells; induction of class II histocompatibility antigens; T_H1 T cell differentiation
TNF-α	TNFrI, TNFrII	Monocytes-macrophages, mast cells, basophils, eosinophils, NK cells, B cells, T cells, keratinocytes, fibroblasts, thymic epithelial cells	All cells except erythrocytes	Fever, anorexia, shock, capillary leak syndrome, enhanced leukocyte cytotoxicity, enhanced NK cell function, acute phase protein synthesis, proinflammatory cytokine induction
TNF-β	TNFrI, TNFrII	T cells, B cells	All cells except erythrocytes	Cell cytotoxicity, lymph node and spleen development
LT-β	LTβR	T cells	All cells except erythrocytes	Cell cytotoxicity, normal lymph node development
G-CSF	G-CSFr; gp130	Monocytes-macrophages, fibroblasts, endothelial cells, thymic epithelial cells, stromal cells	Myeloid cells, endothelial cells	Regulates myelopoiesis; enhances survival and function of neutrophils; clinical use in reversing neutropenia after cytotoxic chemotherapy
GM-CSF	GM-CSFr, common β	T cells, monocytes-macrophages, fibroblasts, endothelial cells, thymic epithelial cells	Monocytes-macrophages, neutrophils, eosinophils, fibroblasts, endothelial cells	Regulates myelopoiesis; enhances macrophage bactericidal and tumoricidal activity; mediator of dendritic cell maturation and function; upregulates NK cell function; clinical use in reversing neutropenia after cytotoxic chemotherapy
M-CSF	M-CSFr (*c-fms* protooncogene)	Fibroblasts, endothelial cells, monocytes-macrophages, T cells, B cells, epithelial cells including thymic epithelium	Monocytes-macrophages	Regulates monocyte-macrophage production and function
LIF	LIFr-α; gp130	Activated T cells, bone marrow stromal cells, thymic epithelium	Megakaryocytes, monocytes, hepatocytes, possibly lymphocyte subpopulations	Induces hepatic acute-phase protein production; stimulates macrophage differentiation; promotes growth of myeloma cells and hematopoietic progenitors; stimulates thrombopoiesis
OSM	OSMr; LIFr; gp130	Activated monocytes-macrophages and T cells, bone marrow stromal cells, some breast carcinoma cell lines, myeloma cells	Neurons, hepatocytes, monocytes-macrophages, adipocytes, alveolar epithelial cells, embryonic stem cells, melanocytes, endothelial cells, fibroblasts, myeloma cells	Induces hepatic acute-phase protein production; stimulates macrophage differentiation; promotes growth of myeloma cells and hematopoietic progenitors; stimulates thrombopoiesis; stimulates growth of Kaposi's sarcoma cells
SCF	SCFr (*c-kit* protooncogene)	Bone marrow stromal cells and fibroblasts	Embryonic stem cells, myeloid and lymphoid precursors, mast cells	Stimulates hematopoietic progenitor cell growth, mast cell growth; promotes embryonic stem cell migration
TGF-β (3 isoforms)	Type I, II, III TGF-β receptor	Most cell types	Most cell types	Downregulates T cell, macrophage, and granulocyte responses; stimulates synthesis of matrix proteins; stimulates angiogenesis

Cytokine	Receptor	Cell Source	Cell Target	Biologic Activity
Lymphotactin/SCM-1	XCR1	NK cells, mast cells, double-negative thymocytes, activated CD8+ T cells	T cells, NK cells	Chemoattractant for lymphocytes; only known chemokine of C class
MCP-1	CCR2	Fibroblasts, smooth-muscle cells, activated PBMCs	Monocytes-macrophages, NK cells, memory T cells, basophils	Chemoattractant for monocytes, activated memory T cells, and NK cells; induces granule release from CD8+ T cells and NK cells; potent histamine-releasing factor for basophils; suppresses proliferation of hematopoietic precursors; regulates monocyte protease production
MCP-2	CCR1, CCR2	Fibroblasts, activated PBMCs	Monocytes-macrophages, T cells, eosinophils, basophils, NK cells	Chemoattractant for monocytes, memory and naïve T cells, eosinophils, ?NK cells; activates basophils and eosinophils; regulates monocyte protease production
MCP-3	CCR1, CCR2	Fibroblasts, activated PBMCs	Monocytes-macrophages, T cells, eosinophils, basophils, NK cells, dendritic cells	Chemoattractant for monocytes, memory and naïve T cells, dendritic cells, eosinophils, ?NK cells; activates basophils and eosinophils; regulates monocyte protease production
MCP-4	CCR2, CCR3	Lung, colon, small intestinal epithelial cells, activated endothelial cells	Monocytes-macrophages, T cells, eosinophils, basophils	Chemoattractant for monocytes, T cells, eosinophils, and basophils
Eotaxin	CCR3	Pulmonary epithelial cells, heart	Eosinophils, basophils	Potent chemoattractant for eosinophils and basophils; induces allergic airways disease; acts in concert with IL-5 to activate eosinophils; antibodies to eotaxin inhibit airway inflammation
TARC	CCR4	Thymus, dendritic cells, activated T cells	T cells, NK cells	Chemoattractant for T and NK cells
MDC	CCR4	Monocytes-macrophages, dendritic cells, thymus	Activated T cells	Chemoattractant for activated T cells; inhibits infection with T cell tropic HIV
MIP-1α	CCR1, CCR5	Monocytes-macrophages, T cells	Monocytes-macrophages, T cells, dendritic cells, NK cells, eosinophils, basophils	Chemoattractant for monocytes, T cells, dendritic cells, and NK cells, and weak chemoattractant for eosinophils and basophils; activates NK cell function; suppresses proliferation of hematopoietic precursors; necessary for myocarditis associated with coxsackievirus infection; inhibits infection with monocytotropic HIV
MIP-1β	CCR5	Monocytes-macrophages, T cells	Monocytes-macrophages, T cells, NK cells, dendritic cells	Chemoattractant for monocytes, T cells, and NK cells; activates NK cell function; inhibits infection with monocytotropic HIV
RANTES	CCR1, CCR2, CCR5	Monocytes-macrophages, T cells, fibroblasts, eosinophils	Monocytes-macrophages, T cells, NK cells, dendritic cells, eosinophils, basophils	Chemoattractant for monocytes-macrophages, CD4+, CD45Ro+ T cells, CD8+ T cells, NK cells, eosinophils, and basophils; induces histamine release from basophils; inhibits infections with monocytotropic HIV
LARC/MIP-3α/Exodus-1	CCR6	Dendritic cells, fetal liver cells, activated T cells	T cells, B cells	Chemoattractant for lymphocytes

Cytokine	Receptor	Cell Source	Cell Target	Biologic Activity
ELC/MIP-3β	CCR7	Thymus, lymph node, appendix	Activated T cells and B cells	Chemoattractant for B and T cells; receptor upregulated on EBV-infected B cells and HSV-infected T cells
I-309/TCA-3	CCR8	Activated T cells	Monocytes-macrophages, T cells	Chemoattractant for monocytes; prevents glucocorticoid-induced apoptosis in some T cell lines
SLC/TCA-4/Exodus-2	CCR7	Thymic epithelial cells, lymph node, appendix, and spleen	T cells	Chemoattractant for T lymphocytes; inhibits hematopoiesis
DC-CK1/PARC	Unknown	Dendritic cells in secondary lymphoid tissues	Naïve T cells	May have a role in induction of immune responses
TECK	CCR9	Dendritic cells, thymus, liver, small intestine	T cells, monocytes-macrophages, dendritic cells	Thymic dendritic cell-derived cytokine, possibly involved in T cell development
GRO-α/MGSA	CXCR2	Activated granulocytes, monocyte-macrophages, and epithelial cells	Neutrophils, epithelial cells, ?endothelial cells	Neutrophil chemoattractant and activator; mitogenic for some melanoma cell lines; suppresses proliferation of hematopoietic precursors; angiogenic activity
GRO-β/MIP-2α	CXCR2	Activated granulocytes and monocyte-macrophages	Neutrophils and ? endothelial cells	Neutrophil chemoattractant and activator; angiogenic activity
NAP-2	CXCR2	Platelets	Neutrophils, basophils	Derived from platelet basic protein; neutrophil chemoattractant and activator
IP-10	CXCR3	Monocytes-macrophages, T cells, fibroblasts, endothelial cells, epithelial cells	Activated T cells, tumor-infiltrating lymphocytes, ? endothelial cells, ?NK cells	IFN-γ-inducible protein that is a chemoattractant for T cells; suppresses proliferation of hematopoietic precursors
MIG	CXCR3	Monocytes-macrophages, T cells, fibroblasts	Activated T cells, tumor-infiltrating lymphocytes	IFN-γ-inducible protein that is a chemoattractant for T cells; suppresses proliferation of hematopoietic precursors
SDF-1	CXCR4	Fibroblasts	T cells, dendritic cells, ? basophils, ?endothelial cells	Low-potency, high-efficacy T cell chemoattractant; required for B-lymphocyte development; prevents infection of CD4+, CXCR4+ cells by T cell tropic HIV
Fractalkine	CX3CR1	Activated endothelial cells	NK cells, T cells, monocytes-macrophages	Cell-surface chemokine/mucin hybrid molecule that functions as a chemoattractant, leukocyte activator, and cell adhesion molecule
PF-4	Unknown	Platelets, megakaryocytes	Fibroblasts, endothelial cells	Chemoattractant for fibroblasts; suppresses proliferation of hematopoietic precursors; inhibits endothelial cell proliferation and angiogenesis

Abbreviations: B7-1, CD80; B7-2, CD86; CCR, CC-type chemokine receptor; CXCR, CXC-type chemokine receptor; DC-CK, dendritic cell chemokine; EBV, Epstein-Barr virus; ELC, EB11 ligand chemokine (MIP-1b); G-CSF, granulocyte colony-stimulating factor; GM-CSF, granulocyte-macrophage colony-stimulating factor; GRP, growth-related peptide; HSV, herpes simplex virus; IFN, interferon; Ig, immunoglobulin; IL, interleukin; IP-10, IFN-γ-inducible protein-10; LARC, liver- and activation-regulated chemokine; LIF, leukemia inhibitory factor; MCP, monocyte chemotactic protein; M-CSF, macrophage colony-stimulating factor; MDC, macrophage-derived chemokine; MGSA, melanoma growth-stimulating activity; MHC, major histocompatibility complex; MIG, monokine induced by IFN-γ; MIP, macrophage inflammatory protein; NAP, neutrophil-activating protein; NK, natural killer; OSM, oncostatin M; PARC, pulmonary- and activation-regulated chemokine; PBMC, peripheral blood mononuclear cells; PF, platelet factor; RANTES, regulated on activation, normally T cell–expressed and –secreted; SCF, stem cell factor; SDF, stromal cell–derived factor; SLC, secondary lymphoid tissue

chemokine; TARC, thymus- and activation-regulated chemokine; TCA, T cell activation protein; TECK, thymus-expressed chemokine; TGF, transforming growth factor; T_H1 and T_H2, helper T cell subsets; TNF, tumor necrosis factor; VCAM, vascular cell adhesion molecule.

Source: Data from JS Sundy et al: Appendix B, in *Inflammation, Basic Principles and Clinical Correlates*, 3rd ed, J Gallin and R Snyderman (eds). Philadelphia, Lippincott Williams and Wilkins, 1999.

TABLE 342-8

CC, CXC_1, CX_3, C_1, and XC Families of Chemokines and Chemokine Receptors

Chemokine Receptor	Chemokine Ligands	Cell Types	Disease Connection
CCR1	CCL3 (MIP-1α), CCL5 (RANTES), CCL7 (MCP-3), CCL14 (HCC1)	T cells, monocytes, eosinophils, basophils	Rheumatoid arthritis, multiple sclerosis
CCR2	CCL2 (MCP-1), CCL8 (MCP-2), CCL7 (MCP-3), CCL13 (MCP-4), CCL16 (HCC4)	Monocytes, dendritic cells (immature), memory T cells	Atherosclerosis, rheumatoid arthritis, multiple sclerosis, resistance to intracellular pathogens, type 2 diabetes mellitus
CCR3	CCL11 (eotaxin), CCL13 (eotaxin-2), CCL7 (MCP-3), CCL5 (RANTES), CCL8 (MCP-2), CCL13 (MCP-4)	Eosinophils, basophils, mast cells, T_H2, platelets	Allergic asthma and rhinitis
CCR4	CCL17 (TARC), CCL22 (MDC)	T cells (T_H2), dendritic cells (mature), basophils, macrophages, platelets	Parasitic infection, graft rejection, T cell homing to skin
CCR5	CCL3 (MIP-1α), CCL4 (MIP-1α), CCL5 (RANTES), CCL11 (eotaxin), CCL14 (HCC1), CCL16 (HCC4)	T cells, monocytes	HIV-1 co-receptor (T cell–tropic strains), transplant rejection
CCR6	CCL20 (MIP-3α, LARC)	T cells (T regulatory and memory), B cells, dendritic cells	Mucosal humoral immunity, allergic asthma, intestinal T cell homing
CCR7	CCL19 (ELC), CCL21 (SLC)	T cells, dendritic cells (mature)	Transport of T cells and dendritic cells to lymph nodes, antigen presentation, and cellular immunity
CCR8	CCL1 (1309)	T cells (T_H2), monocytes, dendritic cells	Dendritic cell migration to lymph node, type 2 cellular immunity, granuloma formation
CCR9	CCL25 (TECK)	T cells, IgA+ plasma cells	Homing of T cells and IgA+ plasma cells to the intestine, inflammatory bowel disease
CCR10	CCL27 (CTACK), CCL28 (MEC)	T cells	T cell homing to intestine and skin
CXCR1	CXCL8 (interleukin-8), CXCL6 (GCP2)	Neutrophils, monocytes	Inflammatory lung disease, COPD
CXCR2	CXCL8, CXCL1 (GROα), CXCL2 (GROα), CXCL3 (GROα), CXCL5 (ENA-78), CXCL6	Neutrophils, monocytes, microvascular endothelial cells	Inflammatory lung disease, COPD, angiogenic for tumor growth
CXCR3-A	CXCL9 (MIG), CXCL10 (IP-10), CXCL11 (I-TAC)	Type 1 helper cells, mast cells, mesangial cells	Inflammatory skin disease, multiple sclerosis, transplant rejection
CXCR3-B	CXCL4 (PF4), CXCL9 (MIG), CXCL10 (IP-10), CXCL11 (I-TAC)	Microvascular endothelial cells, neoplastic cells	Angiostatic for tumor growth
CXCR4	CXCL12 (SDF-1)	Widely expressed	HIV-1 co-receptor (T cell–tropic), tumor metastases, hematopoiesis
CXCR5	CXCL13 (BCA-1)	B cells, follicular helper T cells	Formation of B cell follicles
CXCR6	CXCL16 (SR-PSOX)	CD8+ T cells, natural killer cells, and memory CD4+ T cells	Inflammatory liver disease, atherosclerosis (CXCL16)

Chemokine Receptor	Chemokine Ligands	Cell Types	Disease Connection
CX_3CR1	CX3CL1 (fractalkine)	Macrophages, endothelial cells, smooth-muscle cells	Atherosclerosis
XCR1	XCL1 (lymphotactin), XCL2	T cells, natural killer cells	Rheumatoid arthritis, IgA nephropathy, tumor response

Abbreviations: BCA-1, B-cell chemoattractant 1; COPD, chronic obstructive pulmonary disease; CTACK, cutaneous T cell–attracting chemokine; ELC, Epstein-Barr I1-ligand chemokine; ENA, epithelial cell–derived neutrophil-activating peptide; GCP, granulocyte chemotactic protein; GRO, growth-regulated oncogene; HCC, hemofiltrate chemokine; IP-10, interferon inducible 10; I-TAC, interferon-inducible T cell alpha chemoattractant; LARC, liver- and activation-regulated chemokine; MCP, monocyte chemoattractant protein; MDC, macrophage-derived chemokine; MEC, mammary-enriched chemokine; MIG, monokine induced by interferon-γ; MIP, macrophage inflammatory protein; PF, platelet factor; SDF, stromal cell–derived factor; SLC, secondary lymphoid-tissue chemokine; SR-PSOX, scavenger receptor for phosphatidylserine-containing oxidized lipids; TARC, thymus- and activation-regulated chemokine; TECK, thymus-expressed chemokine; T_H2, type 2 helper T cells.

Source: From IF Charo, RM Ranshohoff: N Engl J Med 354:610, 2006, with permission. Copyright Massachusetts Medical Society. All rights reserved.

TABLE 342-9

Cytokine Families Grouped by Structural Similarity

Hematopoietins	IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-9, IL-11, IL-12, IL-15, IL-16, IL-17, IL-21, IL-23, EPO, LIF, GM-CSF, G-CSF, OSM, CNTF, GH, and TPO
	TNF-α, LT-α, LT-β, CD40L, CD30L, CD27L, 4-1BBL, OX40, OPG, and FasL
IL-1	IL-1α, IL-1β, IL-1ra, IL-18, bFGF, aFGF, and ECGF
PDGF	PDGF A, PDGF B, and M-CSF
TGF-β	TGF-β and BMPs (1, 2, 4, etc.)
C-X-C chemokines	IL-8, Gro-α/β/γ, NAP-2, ENA78, GCP-2, PF4, CTAP-3, MIG, and IP-10
C-C chemokines	MCP-1, MCP-2, MCP-3, MIP-1α, MIP-1β, RANTES

Abbreviations: aFGF, acidic fibroblast growth factor; 4-1BBL, 401 BB ligand; bFGF, basic fibroblast growth factor; BMP, bone marrow morphogenetic proteins; C-C, cysteine-cysteine; CD, cluster of differentiation; CNTF, ciliary neurotrophic factor; CTAP, connective tissue–activating peptide; C-X-C, cysteine-x-cysteine; ECGF, endothelial cell growth factor; EPO, erythropoietin; FasL, Fas ligand; GCP-2, granulocyte chemotactic protein 2; G-CSF, granulocyte colony-stimulating factor; GH, growth hormone; GM-CSF, granulocyte-macrophage colony-stimulating factor; Gro, growth-related gene products; IFN, interferon; IL, interleukin; IP, interferon-γ inducible protein; LIF, leukemia inhibitory factor; LT, lymphotoxin; MCP, monocyte chemoattractant; M-CSF, macrophage colony-stimulating factor; MIG, monokine induced by interferon-γ; MIP, macrophage inflammatory protein; NAP-2, neutrophil activating protein 2; OPG, osteoprotegerin; OSM, oncostatin M; PDGF, platelet-derived growth factor; PF, platelet factor; R, receptor; RANTES, regulated on activation, normal T cell–expressed and –secreted; TGF, transforming growth factor; TNF, tumor necrosis factor; TPO, thyroperoxidase.

Although monocytes-macrophages were originally thought to be the major APCs of the immune system, it is now clear that cell types called *dendritic cells* are the most potent and effective APCs in the body (see below). Monocytes-macrophages mediate innate immune effector functions such as destruction of antibody-coated bacteria, tumor cells, or even normal hematopoietic cells in certain types of autoimmune cytopenias. Monocytes-macrophages ingest bacteria or are infected by viruses, and in doing so, they frequently undergo programmed cell death or *apoptosis*. Macrophages that are infected by intracellular infectious agents are recognized by DCs as infected and apoptotic cells and are phagocytosed by DCs. In this manner, DCs "cross-present" infectious agent antigens of macrophages to T cells. Activated macrophages can also mediate antigen-nonspecific lytic activity and eliminate cell types such as tumor cells in the absence of antibody. This activity is largely mediated by cytokines (i.e., TNF-α and IL-1). Monocytes-macrophages express lineage-specific molecules (e.g., the cell-surface LPS receptor, CD14) as well as surface receptors for a number of molecules, including the Fc region of IgG, activated complement components, and various cytokines (Table 342-7).

Dendritic Cells

Human DCs contain several subsets, including myeloid DCs and plasmacytoid DCs. Myeloid DCs can differentiate into either macrophages-monocytes or tissue-specific DCs. In contrast to myeloid DCs, plasmacytoid DCs are inefficient APCs but are potent producers of type I IFN (e.g., IFN-α) in response to viral infections. The maturation of DCs is regulated through cell-to-cell contact and soluble factors, and DCs attract immune effectors through secretion of chemokines. When DCs come in contact with bacterial products, viral proteins, or host proteins released as danger signals from distressed host cells (Fig.

342-2), infectious agent molecules bind to various TLRs and activate DCs to release cytokines and chemokines that drive cells of the innate immune system to become activated to respond to invading organisms, and recruit T and B cells of the adaptive immune system to respond. Plasmacytoid DCs produce antiviral IFN-α that activates NK cell killing of pathogen-infected cells; IFN-α also activates T cells to mature into antipathogen cytotoxic (killer) T cells. Following contact with pathogens, both plasmacytoid and myeloid DCs produce chemokines that attract helper and cytotoxic T cells, B cells, polymorphonuclear cells, and naïve and memory T cells as well as regulatory T cells to ultimately dampen the immune response once the pathogen is controlled. TLR engagement on DCs upregulates MHC class II, B7-1 (CD80), and B7-2 (CD86), which enhance DC-specific antigen presentation and induce cytokine production (Table 342-7). Thus, DCs are important bridges between early (innate) and later (adaptive) immunity. DCs also modulate and determine the types of immune responses induced by pathogens via the TLRs expressed on DCs (TLR7–9 on plasmacytoid DCs, TLR4 on monocytoid DCs) and via the TLR adapter proteins that are induced to associate with TLRs (Fig. 342-1, Table 342-4). In addition, other PRRs, such as C-type lectins, NLRs, and mannose receptors, upon ligation by pathogen products, activate cells of the adaptive immune system and, like TLR stimulation, by a variety of factors, determine the type and quality of the adaptive immune response that is triggered (Table 342-4).

Large Granular Lymphocytes/Natural Killer Cells

LGLs or NK cells account for ~5–15% of peripheral blood lymphocytes. NK cells are nonadherent, nonphagocytic cells with large azurophilic cytoplasmic granules. NK cells express surface receptors for the Fc portion of IgG (FcR) (CD16) and for NCAM-I (CD56), and many NK cells express T lineage markers, particularly CD8, and proliferate in response to IL-2. NK cells arise in both bone marrow and thymic microenvironments.

Functionally, NK cells share features with both monocytes-macrophages and neutrophils in that they mediate both antibody-dependent cellular cytotoxicity (ADCC) and NK cell activity. ADCC is the binding of an opsonized (antibody-coated) target cell to an Fc receptor-bearing effector cell via the Fc region of antibody, resulting in lysis of the target by the effector cell. NK cell cytotoxicity is the nonimmune (i.e., effector cell never having had previous contact with the target), MHC-unrestricted, non-antibody-mediated killing of target cells, which are usually malignant cell types, transplanted foreign cells, or virus-infected cells. Thus, NK cell cytotoxicity may play an important role in immune surveillance and destruction of malignant and virus-infected host cells. NK cell hyporesponsiveness is also observed in patients with *Chédiak-Higashi syndrome*, an autosomal recessive disease associated with fusion of cytoplasmic granules and defective degranulation of neutrophil lysosomes.

NK cells have a variety of surface receptors that have inhibitory or activating functions and belong to two structural families. These families include the immunoglobulin superfamily and the lectin-like type II transmembrane proteins. NK immunoglobulin superfamily receptors include the killer cell immunoglobulin-like activating or inhibitory receptors (KIRs), many of which have been shown to have HLA class I ligands. The KIRs are made up proteins with either two (KIR2D) or three (KIR3D) extracellular immunoglobulin domains (D). Moreover, their nomenclature designates their function as either inhibitory KIRs with a long (L) cytoplasmic tail and immunoreceptor tyrosine-based inhibitory motif (ITIM) (KIRDL) or activating KIRs with a short (S) cytoplasmic tail (KIRDS). NK cell inactivation by KIRs is a central mechanism to prevent damage to normal host cells. Genetic studies have demonstrated the association of KIRs with viral infection outcome and autoimmune disease **(Table 342-10)**.

TABLE 342-10

Association of KIRS with Disease

Disease	KIR Association	Observation
Psoriatic arthritis	KIR2DS1/KIR2DS2; HLA-Cw group homozygosity	Susceptibility
Spondylarthritides	Increased KIR3DL2 expression Interaction HLA-B27 homodimers with KIR3DL1/KIR3DL2; independent of peptide	May contribute to disease pathology May contribute to disease pathogenesis
Ankylosing spondylitis	KIR3DL1/3DS1; HLA-B27 genotypes	Susceptibility
Rheumatoid vasculitis	KIR2DS2; HLA-Cw*03 Increased KIR2L2/2DS2 in patients with extraarticular manifestations	Susceptibility Clinical manifestations may have different genetic backgrounds with respect to KIR genotype
Rheumatoid arthritis	Decreased KIR2DS1/3DS1 in patients without bone erosions KIR2DS4; HLA-Cw4	Susceptibility Susceptibility
Scleroderma	KIR2DS2+/KIR2DL2–	Susceptibility
Behçet's disease	Altered KIR3DL1 expression	Associated with severe eye disease
Psoriasis vulgaris	2DS1; HLA-Cw*06 2DS1; 2DL5; haplotype B	Susceptibility Susceptibility
IDDM	KIR2DS2; HLA-C1	Susceptibility
Type 1 diabetes	KIR2DS2; HLA-C1 and no HLA-C2, no HLA-Bw4	Increased disease progression
Preeclampsia	KIR2DL1 with fewer KIR2DS (mother); HLA-C2 (fetus)	Increased disease progression
AIDS	KIR3DS1; HLA-Bw4Ile80 KIR3DS1 homozygous; no HLA-Bw4Ile80	Decreased disease progression Increased disease progression
HCV infection	KIR2DL3 homozygous; HLA-C1 homozygous	Decreased disease progression
Cervical neoplasia (HPV induced)	KIR3DS1; HLA-C1 homozygous and no HLA-Bw4	Increased disease progression
Malignant melanoma	KIR2DL2 and/or KIR2DL3; HLA-C1	Increased disease progression

Abbreviations: HCV, hepatitis C virus; HLA, human leukocyte antigen; HPV, human papillomavirus; IDDM, insulin-dependent diabetes mellitus; KIR, killer cell immunoglobulin-like receptor.

Source: Adapted from R Diaz-Pena et al: Adv Exp Med Biol 649:286, 2009.

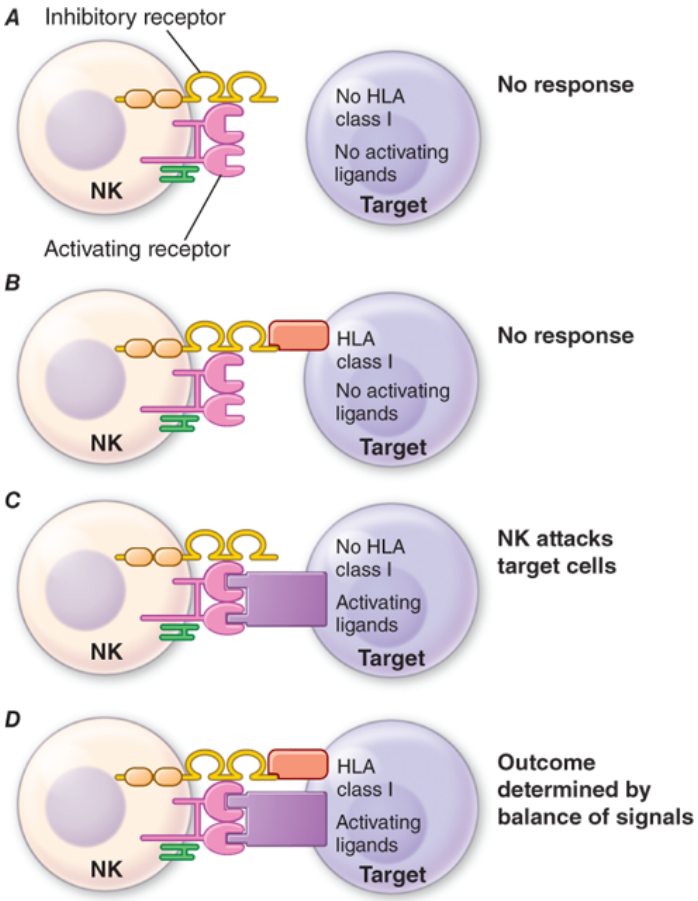
In addition to the KIRs, a second set of immunoglobulin superfamily receptors includes the natural cytotoxicity receptors (NCRs), which include NKp46, NKp30, and NKp44. These receptors help to mediate NK cell activation against target cells. The ligands to which NCRs bind on target cells have been recently recognized to be comprised of molecules of pathogens such as influenza, cytomegalovirus, and malaria as well as host molecules expressed on tumor cells.

NK cell signaling is, therefore, a highly coordinated series of inhibiting and activating signals that prevent NK cells from responding to uninfected, nonmalignant self-cells; however, they are activated to attack malignant and virally infected cells **(Fig. 342-3)**. Recent evidence suggests that NK cells, although not possessing rearranging immune recognition genes, may be able to mediate recall for NK cell responses to viruses and for immune responses such as contact hypersensitivity.

FIGURE 342-3

Encounters between NK cells: Potential targets and possible outcomes. The amount of activating and inhibitory receptors on the NK cells and the amount of ligands on the target cell, as well as the qualitative differences in the signals transduced, determine the extent of the NK response. ***A.*** When target cells have

no HLA class I or activating ligands, NK cells cannot kill target cells. ***B.*** When target cells bear self-HLA, NK cells cannot kill targets. ***C.*** When target cells are pathogen-infected and have downregulated HLA and express activating ligands, NK cells kill target cells. ***D.*** When NK cells encounter targets with both self-HLA and activating receptors, then the level of target killing is determined by the balance of inhibitory and activating signals to the NK cell. HLA, human leukocyte antigen; NK, natural killer. *(Adapted from L Lanier: Annu Rev Immunol 23:225, 2005; reproduced with permission from Annual Reviews Inc. Copyright 2011 by Annual Reviews Inc.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Some NK cells express CD3 and invariant TCR-α chains and are termed *NK T cells*. TCRs of NK T cells recognize lipid molecules of intracellular bacteria when presented in the context of CD1d molecules on APCs. Upon activation, NK T cells secrete effector cytokines such as IL-4 and IFN-γ. This mode of recognition of intracellular bacteria such as *Listeria monocytogenes* and *Mycobacterium tuberculosis* by NK T cells leads to induction of activation of DCs and is thought to be an important innate defense mechanism against these organisms.

The receptors for the Fc portion of IgG (FcγRs) are present on NK cells, B cells, macrophages, neutrophils, and mast cells and mediate interactions of IgG with antibody-coated target cells, such as virally infected cells. Antibody-NK interaction via antibody Fc and NK cell FcR links the adaptive and innate immune systems and regulates the mediation of IgG antibody effector functions such as ADCC. There are both activation and inhibitory FcγRs. Activation FcRs, such as FcγRI (CD64), FcγRIIa (CD32a), and FcγRIIIa (CD16a), are characterized by the presence of an immunoreceptor tyrosine-based activating motif (ITAM) sequence, whereas inhibitory FcRs, such as FcγRIIb (CD32b), contain an ITIM sequence. There is evidence that dysregulation in IgG-FcγR interactions plays roles in arthritis, multiple sclerosis, and systemic lupus erythematosus.

Neutrophils, Eosinophils, and Basophils

Granulocytes are present in nearly all forms of inflammation and are amplifiers and effectors of innate immune responses (Fig. 342-2). Unchecked accumulation and activation of granulocytes can lead to host tissue damage, as seen in neutrophil- and eosinophil-mediated *systemic necrotizing vasculitis*. Granulocytes are derived from stem cells in bone marrow. Each type of granulocyte (neutrophil, eosinophil, or basophil) is derived from a different subclass of progenitor cell that is stimulated to proliferate by colony-stimulating factors (Table 342-7). During terminal maturation of granulocytes, class-specific nuclear morphology and cytoplasmic granules appear that allow for histologic identification of granulocyte type.

Neutrophils express Fc receptor IIIa for IgG (CD16a) as well as receptors for activated complement components (C3b or CD35). Upon interaction of neutrophils with antibody-coated (opsonized) bacteria or immune complexes, azurophilic granules (containing myeloperoxidase, lysozyme, elastase, and other enzymes) and specific granules (containing lactoferrin, lysozyme, collagenase, and other enzymes) are released, and microbicidal superoxide radicals

(O_2^-) are generated at the neutrophil surface. The generation of superoxide leads to inflammation by direct injury to tissue and by alteration of macromolecules such as collagen and DNA.

Eosinophils are potent cytotoxic effector cells for various parasitic organisms. In *Nippostrongylus brasiliensis* helminth infection, eosinophils are important cytotoxic effector cells for removal of these parasites. Key to regulation of eosinophil cytotoxicity to *N. brasiliensis* worms are antigen-specific T helper cells that produce IL-4, thus providing an example of regulation of innate immune responses by adaptive immunity antigen-specific T cells. Intracytoplasmic contents of eosinophils, such as major basic protein, eosinophil cationic protein, and eosinophil-derived neurotoxin, are capable of directly damaging tissues and may be responsible in part for the organ system dysfunction in the *hypereosinophilic syndromes* **(Chap. 60)**. Because the eosinophil granule contains anti-inflammatory types of enzymes (histaminase, arylsulfatase, phospholipase D), eosinophils may homeostatically downregulate or terminate ongoing inflammatory responses.

Basophils and tissue mast cells are potent reservoirs of cytokines such as IL-4 and can respond to bacteria and viruses with antipathogen cytokine production through multiple TLRs expressed on their surface. Mast cells and basophils can also mediate immunity through the binding of antipathogen antibodies. This is a particularly important host defense mechanism against parasitic diseases. Basophils express high-affinity surface receptors for IgE (FcεRII) (CD23) and, upon cross-linking of basophil-bound IgE by antigen, can release histamine, eosinophil chemotactic factor of anaphylaxis, and neutral protease—all mediators of allergic immediate (anaphylaxis) hypersensitivity responses **(Table 342-11)**. In addition, basophils express surface receptors for activated complement components (C3a, C5a), through which mediator release can be directly affected. Thus, basophils, like most cells of the immune system, can be activated in the service of host defense against pathogens, or they can be activated for mediation release and cause pathogenic responses in allergic and inflammatory diseases. **For further discussion of tissue mast cells, see Chap. 347**.

TABLE 342-11

Examples of Mediators Released from Immune Cells and Basophils

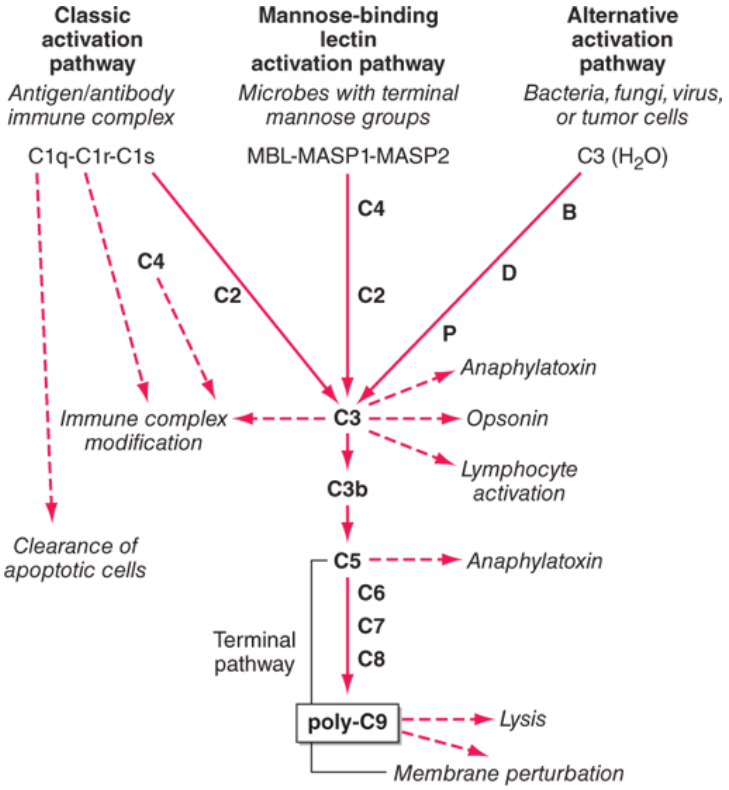
Mediator	Actions
Histamine	Smooth-muscle contraction, increased vascular permeability
Slow reacting substance of anaphylaxis (SRSA) (leukotriene C_4, D_4, E_4)	Smooth-muscle contraction
Eosinophil chemotactic factor of anaphylaxis (ECF-A)	Chemotactic attraction of eosinophils
Platelet-activating factor	Activates platelets to secrete serotonin and other mediators: smooth-muscle contraction; induces vascular permeability
Neutrophil chemotactic factor (NCF)	Chemotactic attraction of neutrophils
Leukotactic activity (leukotriene B_4)	Chemotactic attraction of neutrophils
Heparin	Anticoagulant
Basophil kallikrein of anaphylaxis (BK-A)	Cleaves kininogen to form bradykinin

The Complement System

The complement system, an important soluble component of the innate immune system, is a series of plasma enzymes, regulatory proteins, and proteins that are activated in a cascading fashion, resulting in cell lysis. There are four pathways of the complement system: the classic activation pathway activated by antigen/antibody immune complexes, the MBL (a serum collectin; Table 342-3) activation pathway activated by microbes with terminal mannose groups, the alternative activation pathway activated by microbes or tumor cells, and the terminal pathway that is common to the first three pathways and leads to the membrane attack complex that lyses cells **(Fig. 342-4)**. The series of enzymes of the complement system are serine proteases.

FIGURE 342-4

The four pathways and the effector mechanisms of the complement system. *Dashed arrows* indicate the functions of pathway components. *(After BJ Morley, MJ Walport: The Complement Facts Books. London, Academic Press, 2000, Chap. 2; with permission. Copyright Academic Press, London, 2000.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Activation of the classic complement pathway via immune complex binding to C1q links the innate and adaptive immune systems via specific antibody in the immune complex. The alternative complement activation pathway is antibody-independent and is activated by binding of C3 directly to pathogens and "altered self" such as tumor cells. In the renal glomerular inflammatory disease *IgA nephropathy,* IgA activates the alternative complement pathway and causes glomerular damage and decreased renal function. Activation of the classic complement pathway via C1, C4, and C2 and activation of the alternative pathway via factor D, C3, and factor B both lead to cleavage and activation of C3. C3 activation fragments, when bound to target surfaces such as bacteria and other foreign antigens, are critical for opsonization (coating by antibody and complement) in preparation for phagocytosis. The MBL pathway substitutes MBL-associated serine proteases (MASPs) 1 and 2 for C1q, C1r, and C1s to activate C4. The MBL activation pathway is activated by mannose on the surface of bacteria and viruses.

The three pathways of complement activation all converge on the final common terminal pathway. C3 cleavage by each pathway results in activation of C5, C6, C7, C8, and C9, resulting in the membrane attack complex that physically inserts into the membranes of target cells or bacteria and lyses them.

Thus, complement activation is a critical component of innate immunity for responding to microbial infection. The functional consequences of complement activation by the three initiating pathways and the terminal pathway are shown in Fig. 342-4. In general, the cleavage products of complement components facilitate microbe or damaged cell clearance (C1q, C4, C3), promote activation and enhancement of inflammation (anaphylatoxins, C3a, C5a), and promote microbe or opsonized cell lysis (membrane attack complex).

CYTOKINES

Cytokines are soluble proteins produced by a wide variety of cell types (Tables 342-7,342-8,342-9). They are critical for both normal innate and adaptive immune responses, and their expression may be perturbed in most immune, inflammatory, and infectious disease states.

Cytokines are involved in the regulation of the growth, development, and activation of immune system cells and in the mediation of the inflammatory response. In general, cytokines are characterized by considerable redundancy; different cytokines have similar functions. In addition, many cytokines are pleiotropic in that they are capable of acting on many different cell types. This pleiotropism results from the expression on multiple cell types of receptors for the same cytokine (see below), leading to the formation of "cytokine networks." The action of cytokines may be (1) autocrine when the target cell is the same cell that secretes the cytokine, (2) paracrine when the target cell is nearby, and (3) endocrine when the cytokine is secreted into the circulation and acts distal to the source.

Cytokines have been named based on presumed targets or based on presumed functions. Those cytokines that are thought to primarily target leukocytes have been named IL-1, 2, 3, etc. Many cytokines that were originally described as having a certain function have retained those names (e.g., granulocyte colony-stimulating factor [G-CSF]). Cytokines belong in general to three major structural families: the hematopoietin family; the TNF, IL-1, platelet-derived growth factor (PDGF), and transforming growth factor (TGF) β families; and the CXC and C-C chemokine families (Table 342-9). Chemokines are cytokines

that regulate cell movement and trafficking; they act through G protein-coupled receptors and have a distinctive three-dimensional structure. IL-8 is the only chemokine that early on was named an IL (Table 342-7).

In general, cytokines exert their effects by influencing gene activation that results in cellular activation, growth, differentiation, functional cell-surface molecule expression, and cellular effector function. In this regard, cytokines can have dramatic effects on the regulation of immune responses and the pathogenesis of a variety of diseases. Indeed, T cells have been categorized on the basis of the pattern of cytokines that they secrete, which results in either humoral immune response (T_H2) or cell-mediated immune response (T_H1). A third type of T helper cell is the T_H17 cell that contributes to host defense against extracellular bacteria and fungi, particularly at mucosal sites (Fig. 342-2).

Cytokine receptors can be grouped into five general families based on similarities in their extracellular amino acid sequences and conserved structural domains. The *immunoglobulin (Ig) superfamily* represents a large number of cell-surface and secreted proteins. The IL-1 receptors (type 1, type 2) are examples of cytokine receptors with extracellular Ig domains.

The hallmark of the *hematopoietic growth factor (type 1) receptor* family is that the extracellular regions of each receptor contain two conserved motifs. One motif, located at the N terminus, is rich in cysteine residues. The other motif is located at the C terminus proximal to the transmembrane region and comprises five amino acid residues, tryptophan-serine-X-tryptophan-serine (WSXWS). This family can be grouped on the basis of the number of receptor subunits they have and on the utilization of shared subunits. A number of cytokine receptors, i.e., IL-6, IL-11, IL-12, and leukemia inhibitory factor, are paired with gp130. There is also a common 150-kDa subunit shared by IL-3, IL-5, and granulocyte-macrophage colony-stimulating factor (GM-CSF) receptors. The gamma chain (γ_c) of the IL-2 receptor is common to the IL-2, IL-4, IL-7, IL-9, and IL-15 receptors. Thus, the specific cytokine receptor is responsible for ligand-specific binding, whereas the subunits such as gp130, the 150-kDa subunit, and γ_c are important in signal transduction. The γ_c gene is on the X chromosome, and mutations in the γ_c protein result in the *X-linked form of severe combined immune deficiency syndrome (X-SCID)* **(Chap. 344)**.

The members of the *interferon (type II) receptor* family include the receptors for IFN-γ and β, which share a similar 210-amino-acid binding domain with conserved cysteine pairs at both the amino and carboxy termini. The members of the *TNF (type III) receptor family* share a common binding domain composed of repeated cysteine-rich regions. Members of this family include the p55 and p75 receptors for TNF (TNF-R1 and TNF-R2, respectively); CD40 antigen, which is an important B cell-surface marker involved in immunoglobulin isotype switching; fas/Apo-1, whose triggering induces apoptosis; CD27 and CD30, which are found on activated T cells and B cells; and nerve growth factor receptor.

The common motif for the *seven transmembrane helix family* was originally found in receptors linked to GTP-binding proteins. This family includes receptors for chemokines (Table 342-8), β-adrenergic receptors, and retinal rhodopsin. It is important to note that two members of the chemokine receptor family, CXC chemokine receptor type 4 (CXCR4) and β chemokine receptor type 5 (CCR5), have been found to serve as the two major co-receptors for binding and entry of HIV into CD4-expressing host cells **(Chap. 197)**.

Significant advances have been made in defining the signaling pathways through which cytokines exert their intracellular effects. The Janus family of protein tyrosine kinases (JAK) is a critical element involved in signaling via the hematopoietin receptors. Four JAK kinases, JAK1, JAK2, JAK3, and Tyk2, preferentially bind different cytokine receptor subunits. Cytokine binding to its receptor brings the cytokine receptor subunits into apposition and allows a pair of JAKs to transphosphorylate and activate one another. The JAKs then phosphorylate the receptor on the tyrosine residues and allow signaling molecules to bind to the receptor, whereby the signaling molecules become phosphorylated. Signaling molecules bind the receptor because they have domains (SH2, or src homology 2 domains) that can bind phosphorylated tyrosine residues. There are a number of these important signaling molecules that bind the receptor, such as the adapter molecule SHC, which can couple the receptor to the activation of the mitogen-activated protein kinase pathway. In addition, an important class of substrate of the JAKs is the signal transducers and activators of transcription (STAT) family of transcription factors. STATs have SH2 domains that enable them to bind to phosphorylated receptors, where they are then phosphorylated by the JAKs. It appears that different STATs have specificity for different receptor subunits. The STATs then dissociate from the receptor and translocate to the nucleus, bind to DNA motifs that they recognize, and regulate gene expression. The STATs preferentially bind DNA motifs that are slightly different from one another and thereby control transcription of specific genes. The importance of this pathway is particularly relevant to lymphoid development. Mutations of JAK3 itself also result in a disorder identical to X-SCID; however, because JAK3 is found on chromosome 19 and not on the X chromosome, JAK3 deficiency occurs in boys and girls **(Chap. 344)**.

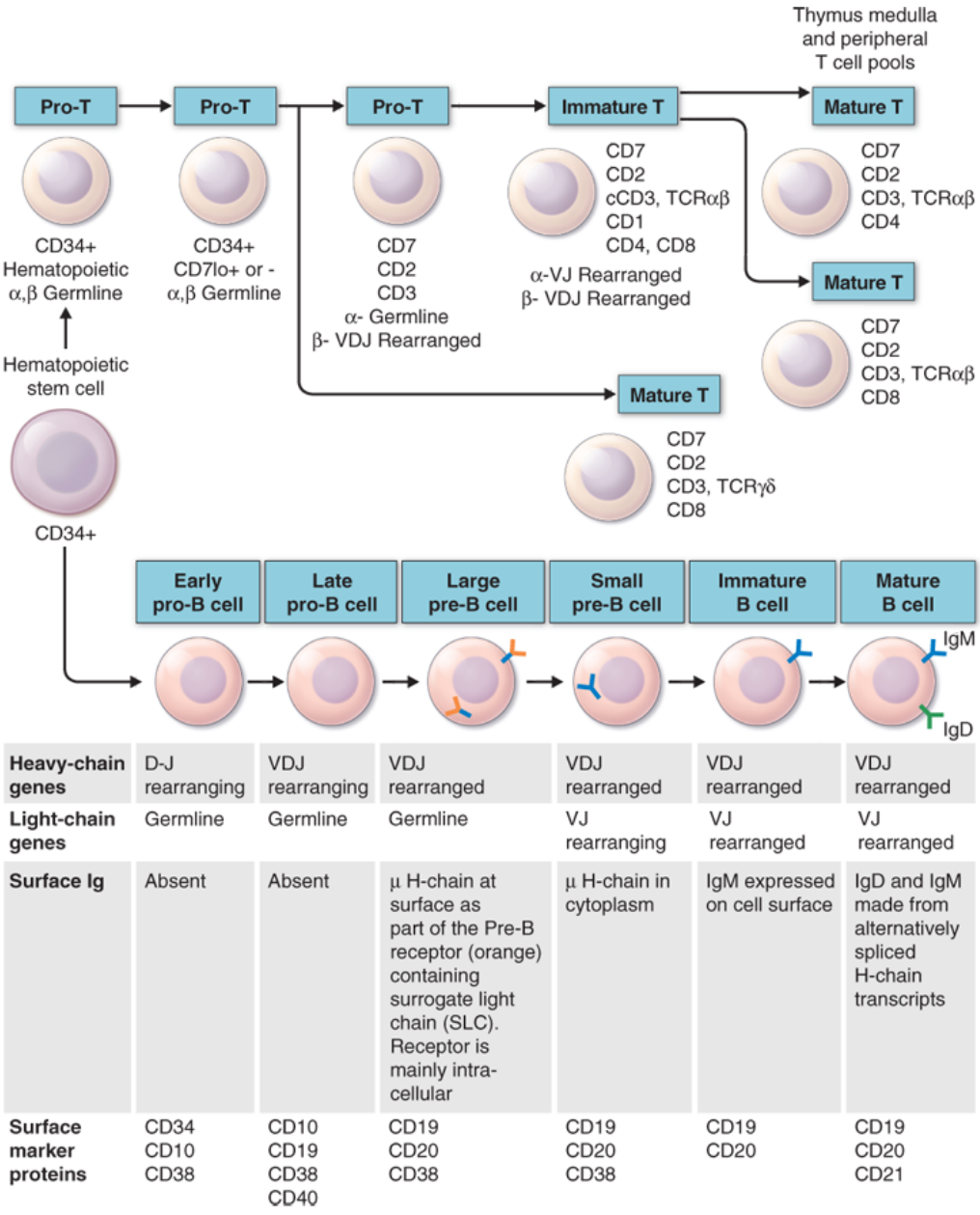
THE ADAPTIVE IMMUNE SYSTEM

Adaptive immunity is characterized by antigen-specific responses to a foreign antigen or pathogen. A key feature of adaptive immunity is that following the initial contact with antigen (*immunologic priming*), subsequent antigen exposure leads to more rapid and vigorous immune responses (*immunologic memory*). The adaptive immune system consists of dual limbs of cellular and humoral immunity. The principal effectors of cellular immunity are T lymphocytes, whereas the principal effectors of humoral immunity are B lymphocytes. Both B and T lymphocytes derive from a common stem cell **(Fig. 342-5)**.

FIGURE 342-5

Development stages of T and B cells. Elements of the developing T and B cell receptor for antigen are shown schematically. The classification into the various stages of B cell development is primarily defined by rearrangement of the immunoglobulin (Ig) heavy (H) and light (L) chain genes and by the

absence or presence of specific surface markers. The classification of stages of T cell development is primarily defined by cell-surface marker protein expression (sCD3, surface CD3 expression; cCD3, cytoplasmic CD3 expression; TCR, T cell receptor). For B cell development, the pre-B cell receptor is shown as a blue-orange B cell receptor. *(Adapted from CA Janeway et al [eds]: Immunobiology, 9th ed. New York, Garland, 2016; with permission.)*



Heavy-chain genes	D-J rearranging	VDJ rearranging	VDJ rearranged	VDJ rearranged	VDJ rearranged	VDJ rearranged
Light-chain genes	Germline	Germline	Germline	VJ rearranging	VJ rearranged	VJ rearranged
Surface Ig	Absent	Absent	μ H-chain at surface as part of the Pre-B receptor (orange) containing surrogate light chain (SLC). Receptor is mainly intra-cellular	μ H-chain in cytoplasm	IgM expressed on cell surface	IgD and IgM made from alternatively spliced H-chain transcripts
Surface marker proteins	CD34 CD10 CD38	CD10 CD19 CD38 CD40	CD19 CD20 CD38	CD19 CD20 CD38	CD19 CD20	CD19 CD20 CD21

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The proportion and distribution of immunocompetent cells in various tissues reflect cell traffic, homing patterns, and functional capabilities. Bone marrow is the major site of maturation of B cells, monocytes-macrophages, DCs, and granulocytes and contains pluripotent stem cells that, under the influence of various colony-stimulating factors, are capable of giving rise to all hematopoietic cell types. T cell precursors also arise from hematopoietic stem cells and home to the thymus for maturation. Mature T lymphocytes, B lymphocytes, monocytes, and DCs enter the circulation and home to peripheral lymphoid organs (lymph nodes, spleen) and mucosal surface-associated lymphoid tissue (gut, genitourinary, and respiratory tracts) as well as the skin and mucous membranes and await activation by foreign antigen.

T Cells

The pool of effector T cells is established in the thymus early in life and is maintained throughout life both by new T cell production in the thymus and by antigen-driven expansion of virgin peripheral T cells into “memory” T cells that reside in peripheral lymphoid organs. The thymus exports ~2% of the total number of thymocytes per day throughout life, with the total number of daily thymic emigrants decreasing by ~3% per year during the first four decades of life.

Mature T lymphocytes constitute 70–80% of normal peripheral blood lymphocytes (only 2% of the total-body lymphocytes are contained in peripheral blood), 90% of thoracic duct lymphocytes, 30–40% of lymph node cells, and 20–30% of spleen lymphoid cells. In lymph nodes, T cells occupy deep paracortical areas around B cell germinal centers, and in the spleen, they are located in periarteriolar areas of white pulp **(Chap. 62)**. T cells are the primary effectors of cell-mediated immunity, with subsets of T cells maturing into CD8+ cytotoxic T cells capable of lysis of virus-infected or foreign cells (short-lived effector T cells) and CD4+ T cells capable of T cell help for CD8+ T cell and B cell development. Two populations of long-lived memory T cells are triggered by infections: effector memory and central memory T cells. Effector memory T cells reside in nonlymphoid organs and respond rapidly to repeated pathogenic infections with cytokine production and cytotoxic functions to kill virus-infected cells. Central memory T cells home to lymphoid organs where they replenish long- and short-lived and effector memory T cells as needed.

In general, CD4+ T cells are the primary regulatory cells of T and B lymphocyte and monocyte function by the production of cytokines and by direct cell contact (Fig. 342-2). In addition, T cells regulate erythroid cell maturation in bone marrow and, through cell contact (CD40 ligand), have an important role in activation of B cells and induction of Ig isotype switching. Considerable evidence now exists that colonization of the gut by commensal bacteria (the gut microbiome) is responsible for expansion of the peripheral CD4+ T cell compartment in normal children and adults.

Human T cells express cell-surface proteins that mark stages of intrathymic T cell maturation or identify specific functional subpopulations of mature T cells. Many of these molecules mediate or participate in important T cell functions (Table 342-1, Fig. 342-5).

The earliest identifiable T cell precursors in bone marrow are CD34+ pro-T cells (i.e., cells in which TCR genes are neither rearranged nor expressed). In the thymus, CD34+ T cell precursors begin cytoplasmic (c) synthesis of components of the CD3 complex of TCR-associated molecules (Fig. 342-5). Within T cell precursors, TCR for antigen gene rearrangement yields two T cell lineages, expressing either TCR-αβ chains or TCR-γδ chains. T cells expressing the TCR-αβ chains constitute the majority of peripheral T cells in blood, lymph node, and spleen and terminally differentiate into either CD4+ or CD8+ cells. Cells expressing TCR-γδ chains circulate as a minor population in blood; their functions, although not fully understood, have been postulated to be those of immune surveillance at epithelial surfaces and cellular defenses against mycobacterial organisms and other intracellular bacteria through recognition of bacterial lipids.

In the thymus, the recognition of self-peptides on thymic epithelial cells, thymic macrophages, and DCs plays an important role in shaping T cell repertoire. As immature cortical thymocytes begin to express surface TCR for antigen, thymocytes with TCRs capable of interacting with self-peptides in the context of self-MHC antigens with low affinity are activated and survive (positive selection). Thymocytes with TCRs that are incapable of binding to self-MHC antigens or bind with high affinity, die of attrition (no selection) or by apoptosis (negative selection). Thymocytes that are positively selected undergo maturation into CD4 or CD8 single positive T cells, and then migrate to the thymus medulla where they interact with self-peptide-self-MHC molecules, where they can again undergo selection. The purpose of negative and positive thymocyte selection is to eliminate potential pathogenic autoreactive T cells, and at the same time, select a repertoire of mature T cells capable of recognizing foreign antigens.

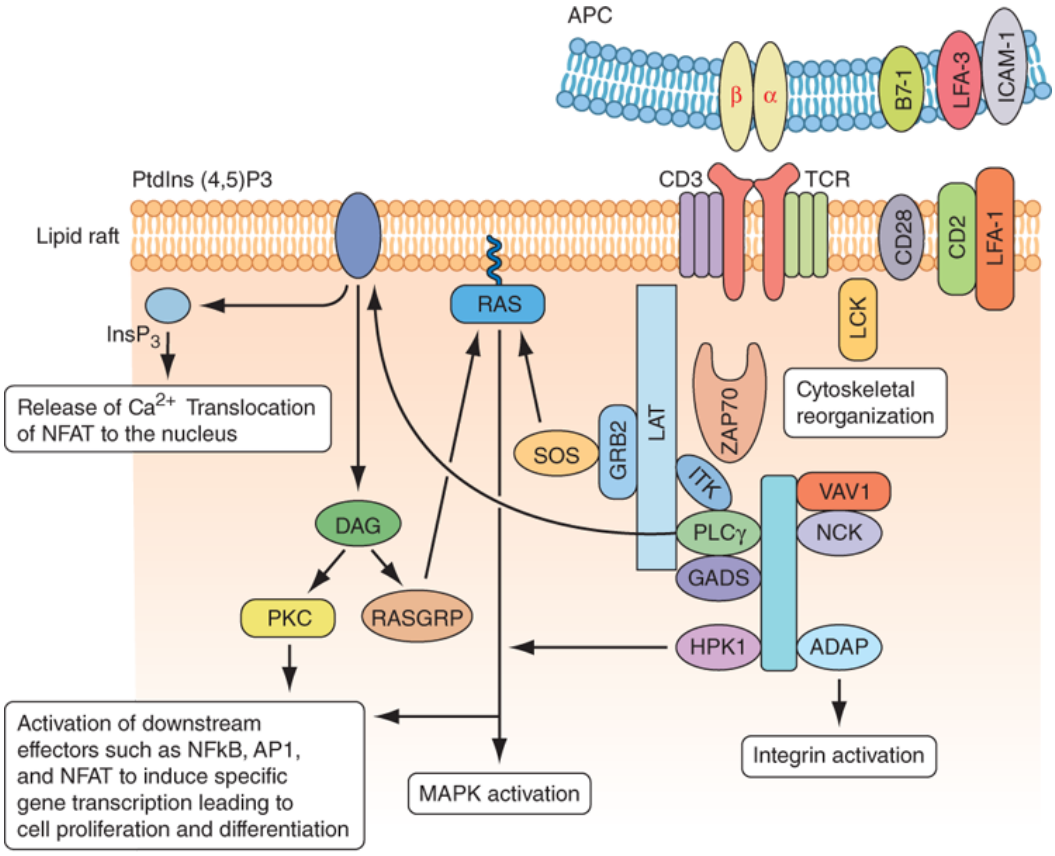
Mature thymocytes that are positively selected are functional MHC class II -restricted CD4+ T cells (Figure 342-2), or they are CD8+ T cells destined to become CD8+ MHC class I -restricted cytotoxic T cells. *MHC class I* - or *class II -restriction* means that T cells recognize antigen peptide fragments only when they are presented in the antigen-recognisite of a class I or class II MHC molecule, respectively **(Chap. 343)**. After thymocyte maturation and selection, CD4 and CD8 thymocytes leave the thymus and migrate to the peripheral immune system. The thymus can continue to be a contributor to the peripheral immune system well into adult life, both normally and when the peripheral T cell pool is damaged, such as occurs in AIDS and cancer chemotherapy.

MOLECULAR BASIS OF T CELL RECOGNITION OF ANTIGEN

The TCR for antigen is a complex of molecules consisting of an antigen-binding heterodimer of either αβ or γδ chains noncovalently linked with five CD3 subunits (γ, δ, ε, ζ, and η) **(Fig. 342-6)**. The CD3 ζ chains are either disulfide-linked homodimers (CD3-ζ_2) or disulfide-linked heterodimers composed of one ζ chain and one η chain. TCR-αβ or TCR-γδ molecules must be associated with CD3 molecules to be inserted into the T cell-surface membrane, TCRα being paired with TCR-β and TCR-γ being paired with TCR-δ. Molecules of the CD3 complex mediate transduction of T cell activation signals via TCRs, whereas TCR-α and β or γ and δ molecules combine to form the TCR antigen-binding site.

FIGURE 342-6

Signaling through the T cell receptor. Activation signals are mediated via immunoreceptor tyrosine-based activation (ITAM) sequences in LAT and CD3 chains (*blue bars*) that bind to enzymes and transduce activation signals to the nucleus via the indicated intracellular activation pathways. Ligation of the T cell receptor (TCR) by MHC complexed with antigen results in sequential activation of LCK and γ-chain-associated protein kinase of 70 kDa (ZAP70). ZAP70 phosphorylates several downstream targets, including LAT (linker for activation of T cells) and SLP76 (SCR homology 2 [SH2] domain-containing leukocyte protein of 76 kDa). SLP76 is recruited to membrane-bound LAT through its constitutive interaction with GADS (GRB2-related adaptor protein). Together, SLP76 and LAT nucleate a multimolecular signaling complex, which induces a host of downstream responses, including calcium flux, mitogen-activated protein kinase (MAPK) activation, integrin activation, and cytoskeletal reorganization. APC, antigen-presenting cell. *(Adapted from GA Koretzky et al: Nat Rev Immunol 6:67, 2006; with permission from Macmillan Publishers Ltd. Copyright 2006.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The α, β, γ, and δ TCR for antigen molecules have amino acid sequence homology and structural similarities to immunoglobulin heavy and light chains and are members of the *immunoglobulin gene superfamily* of molecules. The genes encoding TCR molecules are encoded as clusters of gene segments that rearrange during the course of T cell maturation. This creates an efficient and compact mechanism for housing the diversity requirements of antigen receptor molecules. The TCR-α chain is on chromosome 14 and consists of a series of V (variable), J (joining), and C (constant) regions. The TCR-β chain is on chromosome 7 and consists of multiple V, D (diversity), J, and C TCR-β loci. The TCR-γ chain is on chromosome 7, and the TCR-δ chain is in the middle of the TCR-α locus on chromosome 14. Thus, molecules of the TCR for antigen have constant (framework) and variable regions, and the gene segments encoding the α, β, γ, and δ chains of these molecules are recombined and selected in the thymus, culminating in synthesis of the completed molecule. In both T and B cell precursors (see below), DNA rearrangements of antigen receptor genes involve the same enzymes, recombinase activating gene RAG1 and RAG2, both DNA-dependent protein kinases.

TCR diversity is created by the different V, D, and J segments that are possible for each receptor chain by the many permutations of V, D, and J segment combinations, by "N-region diversification" due to the addition of nucleotides at the junction of rearranged gene segments, and by the pairing of individual chains to form a TCR dimer. As T cells mature in the thymus, the repertoire of antigen-reactive T cells is modified by selection processes that eliminate many autoreactive T cells, enhance the proliferation of cells that function appropriately with self-MHC molecules and antigen, and allow T cells with nonproductive TCR rearrangements to die.

TCR-αβ cells do not recognize native protein or carbohydrate antigens. Instead, T cells recognize only short (~9–13 amino acids) peptide fragments derived from protein antigens taken up or produced in APCs. Foreign antigens may be taken up by endocytosis into acidified intracellular vesicles or by phagocytosis and degraded into small peptides that associate with MHC class II molecules (exogenous antigen-presentation pathway). Other foreign antigens arise endogenously in the cytosol (such as from replicating viruses) and are broken down into small peptides that associate with MHC class I molecules (endogenous antigen-presenting pathway). Thus, APCs proteolytically degrade foreign proteins and display peptide fragments embedded in the MHC class I or II antigen-recognition site on the MHC molecule surface, where foreign peptide fragments are available to bind to TCR-αβ or TCR-γδ chains of reactive T cells. CD4 molecules act as adhesives and, by direct binding to MHC class II (DR, DQ, or DP) molecules, stabilize the interaction of TCR with peptide antigen (Fig. 342-6). Similarly, CD8 molecules also act as adhesives to stabilize the TCR-antigen interaction by direct CD8 molecule binding to MHC class I (A, B, or C) molecules.

Antigens that arise in the cytosol and are processed via the endogenous antigen-presentation pathway are cleaved into small peptides by a complex of proteases called the *proteasome*. From the proteasome, antigen peptide fragments are transported from the cytosol into the lumen of the endoplasmic reticulum by a heterodimeric complex termed *transporters associated with antigen processing*, or TAP proteins. There, MHC class I molecules in the endoplasmic reticulum membrane physically associate with processed cytosolic peptides. Following peptide association with class I molecules, peptide-class I complexes are exported to the Golgi apparatus, and then to the cell surface, for recognition by CD8+ T cells.

Antigens taken up from the extracellular space via endocytosis into intracellular acidified vesicles are degraded by vesicle proteases into peptide fragments. Intracellular vesicles containing MHC class II molecules fuse with peptide-containing vesicles, thus allowing peptide fragments to physically bind to MHC class II molecules. Peptide-MHC class II complexes are then transported to the cell surface for recognition by CD4+ T cells **(Chap. 343)**.

Whereas it is generally agreed that the TCR-αβ receptor recognizes peptide antigens in the context of MHC class I or class II molecules, lipids in the cell wall of intracellular bacteria such as *M. tuberculosis* can also be presented to a wide variety of T cells, including subsets of TCR-γδ T cells, and a subset of CD8+ TCR-αβ T cells. Importantly, bacterial lipid antigens are not presented in the context of MHC class I or II molecules, but rather are presented in the context of MHC-related CD1 molecules. Some γδ T cells that recognize lipid antigens via CD1 molecules have very restricted TCR usage, do not need antigen priming to respond to bacterial lipids, and may actually be a form of innate rather than acquired immunity to intracellular bacteria.

Just as foreign antigens are degraded and their peptide fragments presented in the context of MHC class I or class II molecules on APCs, endogenous self-proteins also are degraded, and self-peptide fragments are presented to T cells in the context of MHC class I or class II molecules on APCs. In peripheral lymphoid organs, there are T cells that are capable of recognizing self-protein fragments but normally are *anergic* or *tolerant*, i.e., nonresponsive to self-antigenic stimulation, due to lack of self-antigen upregulating APC *co-stimulatory molecules* such as B7-1 (CD80) and B7-2 (CD86) (see below).

Once engagement of mature T cell TCR by foreign peptide occurs in the context of self-MHC class I or class II molecules, binding of non-antigen-specific adhesion ligand pairs such as CD54-CD11/CD18 and CD58-CD2 stabilizes MHC peptide-TCR binding, and the expression of these adhesion molecules is upregulated (Fig. 342-6). Once antigen ligation of the TCR occurs, the T cell membrane is partitioned into *lipid membrane microdomains*, or *lipid rafts*, that coalesce the key signaling molecules TCR/CD3 complex, CD28, CD2, LAT (linker for activation of T cells), intracellular activated (dephosphorylated) src family protein tyrosine kinases (PTKs), and the key CD3ζ-associated protein-70 (ZAP-70) PTK (Fig. 342-6). Importantly, during T cell activation, the CD45 molecule, with protein tyrosine phosphatase activity, is partitioned away from the TCR complex to allow activating phosphorylation events to occur. The coalescence of signaling molecules of activated T lymphocytes in *microdomains* has suggested that T cell-APC interactions can be considered *immunologic synapses*, analogous in function to neuronal synapses.

After TCR-MHC binding is stabilized, activation signals are transmitted through the cell to the nucleus and lead to the expression of gene products important in mediating the wide diversity of T cell functions such as the secretion of IL-2. The TCR does not have intrinsic signaling activity but is linked to a variety of signaling pathways via ITAMs expressed on the various CD3 chains that bind to proteins that mediate signal transduction. Each of the pathways results in the activation of particular transcription factors that control the expression of cytokine and cytokine receptor genes. Thus, antigen-MHC binding to the TCR induces the activation of the src family of PTKs, Fyn and Lck (Lck is associated with CD4 or CD8 co-stimulatory molecules); phosphorylation of CD3ζ chain; activation of the related tyrosine kinases ZAP-70 and Syk; and downstream activation of the calcium-dependent calcineurin pathway, the ras pathway, and the protein kinase C pathway. Each of these pathways leads to activation of specific families of transcription factors (including *NF-AT*, *fos* and *jun*, and *rel/NF-κB*) that form heteromultimers capable of inducing expression of IL-2, IL-2 receptor, IL-4, TNF-α, and other T cell mediators.

In addition to the signals delivered to the T cell from the TCR complex and CD4 and CD8, molecules on the T cell, such as CD28 and inducible co-stimulator (ICOS), and molecules on DCs, such as B7-1 (CD80) and B7-2 (CD86), also deliver important co-stimulatory signals that upregulate T cell cytokine production and are essential for T cell activation. If signaling through CD28 or ICOS does not occur, or if CD28 is blocked, the T cell becomes anergic rather than activated (see "Immune Tolerance and Autoimmunity" below). CTLA-4 (CD152) is similar to CD28 in its ability to bind CD80 and CD86. Unlike CD28, CTLA-4 transmits an inhibitory signal to T cells, acting as an off switch.

T CELL EXHAUSTION IN VIRAL INFECTIONS AND CANCER

In chronic viral infections such as HIV-1, hepatitis C virus, and hepatitis B virus and in chronic malignancies, the persistence of antigen disrupts memory T cell function, resulting in defects in memory T cell responses. This has been defined as *T cell exhaustion* and is associated with T cell programmed cell death protein 1 (PD-1) (CD279) expression. Exhausted T cells have compromised proliferation and lose the ability to produce effector molecules, like IL-2, TNF-α, and IFN-γ. PD-1 downregulates T cell responses and is associated with T cell exhaustion and disease progression. For this reason, inhibition of T cell PD-1 activity to enhance effector T cell function is being explored as a target for immunotherapy in both viral infections and certain malignancies.

T CELL SUPERANTIGENS

Conventional antigens bind to MHC class I or II molecules in the groove of the αβ heterodimer and bind to T cells via the V regions of the TCR-α and β chains. In contrast, superantigens bind directly to the lateral portion of the TCR-β chain and MHC class II β chain and stimulate T cells based solely on the Vβ gene segment used independent of the D, J, and Vα sequences present. *Superantigens* are protein molecules capable of activating up to 20% of the peripheral T cell pool, whereas conventional antigens activate <1 in 10,000 T cells. T cell superantigens include staphylococcal enterotoxins and other bacterial products. Superantigen stimulation of human peripheral T cells occurs in the clinical setting of *staphylococcal toxic shock syndrome*, leading to massive overproduction of T cell cytokines that leads to hypotension and shock **(Chap. 142)**.

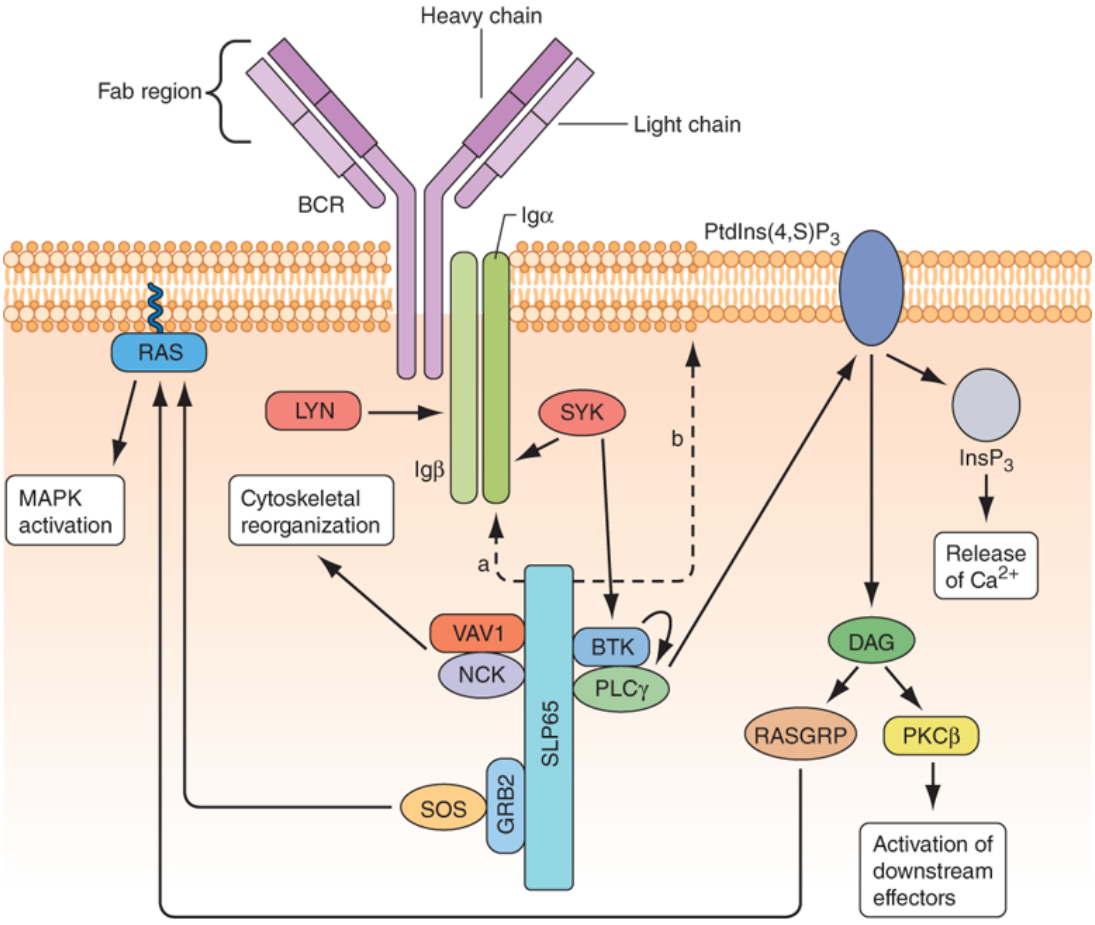
B CELLS

Mature B cells constitute 10–15% of human peripheral blood lymphocytes, 20–30% of lymph node cells, 50% of splenic lymphocytes, and ~10% of bone marrow lymphocytes. B cells express on their surface intramembrane immunoglobulin (Ig) molecules that function as BCRs for antigen in a complex of Ig-associated α and β signaling molecules with properties similar to those described in T cells **(Fig. 342-7)**. Unlike T cells, which recognize only processed

peptide fragments of conventional antigens embedded in the notches of MHC class I and class II antigens of APCs, B cells are capable of recognizing and proliferating to whole unprocessed native antigens via antigen binding to B cell–surface Ig (sIg) receptors. B cells also express surface receptors for the Fc region of IgG molecules (CD32) as well as receptors for activated complement components (C3d or CD21, C3b or CD35). The primary function of B cells is to produce antibodies. B cells also serve as APCs and are highly efficient at antigen processing. Their antigen-presenting function is enhanced by a variety of cytokines. Mature B cells are derived from bone marrow precursor cells that arise continuously throughout life (Fig. 342-5).

FIGURE 342-7

B cell receptor (BCR) activation results in the sequential activation of protein tyrosine kinases, which results in the formation of a signaling complex and activation of downstream pathways as shown. Whereas SLP76 is recruited to the membrane through GADS and LAT, the mechanism of SLP65 recruitment is unclear. Studies have indicated two mechanisms: (a) direct binding by the SH2 domain of SLP65 to immunoglobulin (Ig) of the BCR complex or (b) membrane recruitment through a leucine zipper in the amino terminus of SLP65 and an unknown binding partner. ADAP, adhesion- and degranulation-promoting adaptor protein; AP1, activator protein 1; BTK, Bruton's tyrosine kinase; DAG, diacylglycerol; GRB2, growth factor receptor-bound protein 2; HPK1, hematopoietic progenitor kinase 1; $InsP_3$, inositol-1,4,5-trisphosphate; ITK, interleukin-2-inducible T cell kinase; NCK, noncatalytic region of tyrosine kinase; NF-B, nuclear factor B; PKC, protein kinase C; PLC, phospholipase C; $PtdIns(4,5)P_2$, phosphatidylinositol-4,5-bisphosphate; RASGRP, RAS guanyl-releasing protein; SOS, son of sevenless homologue; SYK, spleen tyrosine kinase. *(Adapted from GA Koretzky et al: Nat Rev Immunol 6:67, 2006; with permission from Macmillan Publishers Ltd. Copyright 2006.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

B lymphocyte development can be separated into antigen-independent and antigen-dependent phases. Antigen-independent B cell development occurs in primary lymphoid organs and includes all stages of B cell maturation up to the sIg+ mature B cell. Antigen-dependent B cell maturation is driven by the interaction of antigen with the mature B cell sIg, leading to memory B cell induction, Ig class switching, and plasma cell formation. Antigen-dependent stages of B cell maturation occur in secondary lymphoid organs, including lymph node, spleen, and gut Peyer's patches. In contrast to the T cell repertoire that is generated intrathymically before contact with foreign antigen, the repertoire of B cells expressing diverse antigen-reactive sites is modified by further alteration of Ig genes after stimulation by antigen—a process called *somatic hypermutation*—that occurs in lymph node germinal centers.

During B cell development, diversity of the antigen-binding variable region of Ig is generated by an ordered set of Ig gene rearrangements that are similar to the rearrangements undergone by TCR α, β, γ, and δ genes. For the heavy chain, there is first a rearrangement of D segments to J segments, followed by a second rearrangement between a V gene segment and the newly formed D-J sequence; the C segment is aligned to the V-D-J complex to yield a functional Ig heavy chain gene (V-D-J-C). During later stages, a functional κ or γ light chain gene is generated by rearrangement of a V segment to a J segment, ultimately yielding an intact Ig molecule composed of heavy and light chains.

The process of Ig gene rearrangement is regulated and results in a single antibody specificity produced by each B cell, with each Ig molecule comprising one type of heavy chain and one type of light chain. Although each B cell contains two copies of Ig light and heavy chain genes, only one gene of each type is productively rearranged and expressed in each B cell, a process termed *allelic exclusion*.

There are ~300 V_κ genes and 5 J_κ genes, resulting in the pairing of V_κ and J_κ genes to create >1500 different kappa light chain combinations. There are ~70 V_λ genes and 4 J_λ genes for >280 different lambda light chain combinations. The number of distinct light chains that can be generated is increased by somatic mutations within the V and J genes, thus creating large numbers of possible specificities from a limited amount of germline genetic information. As noted above, in heavy chain Ig gene rearrangement, the VH domain is created by the joining of three types of germline genes called V_H, D_H, and J_H, thus allowing for even greater diversity in the variable region of heavy chains than of light chains.

The most immature B cell precursors (early pro-B cells) lack cytoplasmic Ig (cIg) and sIg (Fig. 342-5). The large pre-B cell is marked by the acquisition of the surface pre-BCR composed of μ heavy (H) chains and a pre-B light chain, termed V pre-B. V pre-B is a surrogate light chain receptor encoded by the nonrearranged V pre-B and the γ5 light chain locus (the pre-BCR). Pro- and pre-B cells are driven to proliferate and mature by signals from bone marrow stroma—in particular, IL-7. Light chain rearrangement occurs in the small pre-B cell stage such that the full BCR is expressed at the immature B cell stage. Immature B cells have rearranged Ig light chain genes and express sIgM. As immature B cells develop into mature B cells, sIgD is expressed as well as sIgM. At this point, B lineage development in bone marrow is complete, and B cells exit into the peripheral circulation and migrate to secondary lymphoid organs to encounter specific antigens.

Random rearrangements of Ig genes occasionally generate self-reactive antibodies, and mechanisms must be in place to correct these mistakes. One such mechanism is BCR editing, whereby autoreactive BCRs are mutated to not react with self-antigens. If receptor editing is unsuccessful in eliminating autoreactive B cells, then autoreactive B cells undergo negative selection in the bone marrow through induction of apoptosis after BCR engagement of self-antigen.

After leaving the bone marrow, B cells populate peripheral B cell sites, such as lymph node and spleen, and await contact with foreign antigens that react with each B cell's clonotypic receptor. Antigen-driven B cell activation occurs through the BCR, and a process known as *somatic hypermutation* takes place whereby point mutations in rearranged H- and L-genes give rise to mutant sIg molecules, some of which bind antigen better than the original sIg molecules. Somatic hypermutation, therefore, is a process whereby memory B cells in peripheral lymph organs have the best binding, or the highest-affinity antibodies. This overall process of generating the best antibodies is called *affinity maturation of antibody*.

Lymphocytes that synthesize IgG, IgA, and IgE are derived from sIgM+, sIgD+ mature B cells. Ig class switching occurs in lymph node and other peripheral lymphoid tissue germinal centers. CD40 on B cells and CD40 ligand on T cells constitute a critical co-stimulatory receptor-ligand pair of immune-stimulatory molecules. Pairs of CD40+ B cells and CD40 ligand+ T cells bind and drive B cell Ig class switching via T cell-produced cytokines such as IL-4 and TGF-β. IL-1, -2, -4, -5, and -6 synergize to drive mature B cells to proliferate and differentiate into Ig-secreting cells.

Humoral Mediators of Adaptive Immunity: Immunoglobulins

Immunoglobulins are the products of differentiated B cells and mediate the humoral arm of the immune response. The primary functions of antibodies are to bind specifically to antigen and bring about the inactivation or removal of the offending toxin, microbe, parasite, or other foreign substance from the body. The structural basis of Ig molecule function and Ig gene organization has provided insight into the role of antibodies in normal protective immunity, pathologic immune-mediated damage by immune complexes, and autoantibody formation against host determinants.

All immunoglobulins have the basic structure of two heavy and two light chains (Fig. 342-7). Immunoglobulin isotype (i.e., G, M, A, D, E) is determined by the type of Ig heavy chain present. IgG and IgA isotypes can be divided further into subclasses (G1, G2, G3, G4, and A1, A2) based on specific antigenic determinants on Ig heavy chains. The characteristics of human immunoglobulins are outlined in **Table 342-12**. The four chains are covalently linked by disulfide bonds. Each chain is made up of a V region and C regions (also called *domains*), themselves made up of units of ~110 amino acids. Light chains have one variable (V_L) and one constant (C_L) unit; heavy chains have one variable unit (V_H) and three or four constant (C_H) units, depending on isotype. As the name suggests, the constant, or C, regions of Ig molecules are made up of homologous sequences and share the same primary structure as all other Ig chains of the same isotype and subclass. Constant regions are involved in biologic functions of Ig molecules. The C_H2 domain of IgG and the C_H4 units of IgM are involved with the binding of the C1q portion of C1 during complement activation. The C_H region at the carboxy-terminal end of the IgG molecule, the Fc region, binds to surface Fc receptors (CD16, CD32, CD64) of macrophages, DCs, NK cells, B cells, neutrophils, and eosinophils. The Fc of IgA binds to FcαR (CD89), and the Fc of IgE binds to FcεR (CD23).

TABLE 342-12

Physical, Chemical, and Biologic Properties of Human Immunoglobulins

Property	IgG	IgA	IgM	IgD	IgE
Usual molecular form	Monomer	Monomer, dimer	Pentamer, hexamer	Monomer	Monomer
Other chains	None	J chain, SC	J chain	None	None
Subclasses	G1, G2, G3, G4	A1, A2	None	None	None
Heavy chain allotypes	Gm (=30)	No A1, A2m (2)	None	None	None
Molecular mass, kDa	150	160, 400	950, 1150	175	190
Serum level in average adult, mg/mL	9.5–12.5	1.5–2.6	0.7–1.7	0.04	0.0003
Percentage of total serum Ig	75–85	7–15	5–10	0.3	0.019
Serum half-life, days	23	6	5	3	2.5
Synthesis rate, mg/kg per day	33	65	7	0.4	0.016
Antibody valence	2	2, 4	10, 12	2	2
Classical complement activation	+(G1, 2?, 3)	–	++	–	–
Alternate complement activation	+(G4)	+	–	+	–
Binding cells via Fc	Macrophages, neutrophils, large granular lymphocytes	Lymphocytes	Lymphocytes	None	Mast cells, basophils, B cells
Biologic properties	Placental transfer, secondary Ab for most antipathogen responses	Secretory immunoglobulin	Primary Ab responses	Marker for mature B cells	Allergy, antiparasite responses

Source: After L Carayannopoulos, JD Capra, in WE Paul (ed): *Fundamental Immunology*, 3rd ed. New York, Raven, 1993; with permission.

Variable regions (V_L and V_H) constitute the antibody-binding (Fab) region of the molecule. Within the V_L and V_H regions are hypervariable regions (extreme sequence variability) that constitute the antigen-binding site unique to each Ig molecule. The idiotype is defined as the specific region of the Fab portion of the Ig molecule to which antigen binds. Antibodies against the idiotype portion of an antibody molecule are called *anti-idiotype antibodies*. The formation of such antibodies in vivo during a normal B cell antibody response may generate a negative (or "off") signal to B cells to terminate antibody production.

IgG constitutes ~75–85% of total serum immunoglobulin. The four IgG subclasses are numbered in order of their level in serum, IgG1 being found in greatest amounts and IgG4 the least. IgG subclasses have clinical relevance in their varying ability to bind macrophage and neutrophil Fc receptors and to activate complement (Table 342-12). Moreover, selective deficiencies of certain IgG subclasses give rise to clinical syndromes in which the patient is inordinately susceptible to bacterial infections. IgG antibodies are frequently the predominant antibody made after rechallenge of the host with antigen (secondary antibody response).

IgM antibodies normally circulate as a 950-kDa pentamer with 160-kDa bivalent monomers joined by a molecule called the *J chain*, a 15-kDa nonimmunoglobulin molecule that also effects polymerization of IgA molecules. IgM is the first immunoglobulin to appear in the immune response (primary antibody response) and is the initial type of antibody made by neonates. Membrane IgM in the monomeric form also functions as a major antigen receptor on the surface of mature B cells (Table 342-12). IgM is an important component of immune complexes in autoimmune diseases. For example, IgM antibodies

against IgG molecules (rheumatoid factors) are present in high titers in *rheumatoid arthritis*, other collagen diseases, and some infectious diseases (*subacute bacterial endocarditis*).

IgA constitutes only 7–15% of total serum immunoglobulin but is the predominant class of immunoglobulin in secretions. IgA in secretions (tears, saliva, nasal secretions, gastrointestinal tract fluid, and human milk) is in the form of secretory IgA (sIgA), a polymer consisting of two IgA monomers, a joining molecule, again termed the J chain, and a glycoprotein called the *secretory protein*. Of the two IgA subclasses, IgA1 is primarily found in serum, whereas IgA2 is more prevalent in secretions. IgA fixes complement via the alternative complement pathway and has potent antiviral activity in humans by prevention of virus binding to respiratory and gastrointestinal epithelial cells.

IgD is found in minute quantities in serum and, together with IgM, is a major receptor for antigen on the naïve B cell surface. IgE, which is present in serum in very low concentrations, is the major class of immunoglobulin involved in arming mast cells and basophils by binding to these cells via the Fc region. Antigen cross-linking of IgE molecules on basophil and mast cell surfaces results in release of mediators of the immediate hypersensitivity (allergic) response (Table 342-12).

CELLULAR INTERACTIONS IN REGULATION OF NORMAL IMMUNE RESPONSES

The net result of activation of the humoral (B cell) and cellular (T cell) arms of the adaptive immune system by foreign antigen is the elimination of antigen directly by specific effector T cells or in concert with specific antibody. Figure 342-2 is a simplified schematic diagram of the T and B cell responses indicating some of these cellular interactions.

The expression of adaptive immune cell function is the result of a complex series of immunoregulatory events that occur in phases. Both T and B lymphocytes mediate immune functions, and each of these cell types, when given appropriate signals, passes through stages, from activation and induction through proliferation, differentiation, and ultimately effector functions. The effector function expressed may be at the end point of a response, such as secretion of antibody by a differentiated plasma cell, or it might serve a regulatory function that modulates other functions, such as is seen with CD4+ and CD8+ T lymphocytes that modulate both differentiation of B cells and activation of CD8+ cytotoxic T cells.

CD4 helper T cells can be subdivided on the basis of cytokines produced (Fig. 342-2). Activated T_H1-type helper T cells secrete IL-2, IFN-γ, IL-3, TNF-a, GM-CSF, and TNF-β, whereas activated T_H2-type helper T cells secrete IL-3, -4, -5, -6, -10, and -13. T_H1 CD4+ T cells, through elaboration of IFN-γ, have a central role in mediating intracellular killing by a variety of pathogens. T_H1 CD4+ T cells also provide T cell help for generation of cytotoxic T cells and some types of opsonizing antibody, and they generally respond to antigens that lead to delayed hypersensitivity types of immune responses for many intracellular viruses and bacteria (such as HIV or *M. tuberculosis*). In contrast, T_H2 cells have a primary role in regulatory humoral immunity and isotype switching. T_H2 cells, through production of IL-4 and IL-10, have a regulatory role in limiting proinflammatory responses mediated by T_H1 cells (Fig. 342-2). In addition, T_H2 CD4+ T cells provide help to B cells for specific Ig production and respond to antigens that require high antibody levels for foreign antigen elimination (extracellular encapsulated bacteria such as *Streptococcus pneumoniae* and certain parasite infections). Additional subsets of the CD4 T_H cells have been described, one of which is termed T_H17 that secrete cytokines IL-17, -22, and -26. T_H17 cells have been shown to play a role in autoimmune inflammatory disorders in addition to defense against extracellular bacteria and fungi, particularly at mucosal surfaces. T_H9 cells are defined by their secretion of IL-9 and have been shown to play a role in atopic disease, inflammatory bowel disease, and in anti-tumor immunity. Moreover, the Tfh subset of helper T cells is crucial for providing the necessary signals to B cells in germinal centers to undergo affinity maturation. In summary, the type of T cell response generated in an immune response is determined by the microbe PAMPs presented to the DCs, the TLRs on the DCs that become activated, the types of DCs that are activated, and the cytokines that are produced (Table 342-4). Commonly, myeloid DCs produce IL-12 and activate T_H1 T cell responses that result in IFN-γ and cytotoxic T cell induction, and plasmacytoid DCs produce IFN-α and lead to T_H2 responses that result in IL-4 production and enhanced antibody responses.

As shown in Fig. 342-2, upon activation by DCs, T cell subsets that produce IL-2, IL-3, IFN-γ, and/or IL-4, -5, -6, -10, and -13 are generated and exert positive and negative influences on effector T and B cells. For B cells, trophic effects are mediated by a variety of cytokines, particularly T cell–derived IL-3, -4, -5, and -6, that act at sequential stages of B cell maturation, resulting in B cell proliferation, differentiation, and ultimately antibody secretion. For cytotoxic T cells, trophic factors include inducer T cell secretion of IL-2, IFN-γ, and IL-12.

Important types of immunomodulatory T cells that control immune responses are *CD4+ and CD8+ T regulatory cells*. These cells express the α chain of the IL-2 receptor (CD25), produce IL-10, and suppress both T and B cell responses. T regulatory cells are induced by immature DCs and play key roles in maintaining tolerance to self-antigens. Loss of T regulatory cells is the cause of organ-specific autoimmune disease in mice such as autoimmune thyroiditis, adrenalitis, and oophoritis (see "Immune Tolerance and Autoimmunity" below). T regulatory cells also play key roles in controlling the magnitude and duration of immune responses to microbes. Normally, after the initial immune response to a microbe has eliminated the invader, T regulatory cells are activated to suppress the antimicrobe response and prevent host injury. Some microbes have adapted to induce T regulatory cell activation at the site of infection to promote parasite infection and survival. In *Leishmania* infection, the parasite induces T regulatory cell accumulation at skin infection sites that dampens anti-*Leishmania* T cell responses and prevents parasite elimination. Although B cells recognize native antigen via B cell–surface Ig receptors, B cells require T cell help to produce high-affinity antibody of multiple isotypes that are the most effective in eliminating foreign antigen. T cell–B cell

interactions that lead to high-affinity antibody production require (1) processing of native antigen by B cells and expression of peptide fragments on the B cell surface for presentation to T_H cells, (2) the ligation of B cells by both the TCR complex and the CD40 ligand, (3) induction of the process termed *antibody isotype switching* in antigen-specific B cell clones, and (4) induction of the process of affinity maturation of antibody in the germinal centers of B cell follicles of lymph node and spleen.

Naïve B cells express cell-surface IgD and IgM, and initial contact of naïve B cells with antigen is via binding of native antigen to B cell–surface IgM. T cell cytokines, released following T_H2 cell contact with B cells or by a "bystander" effect, induce changes in Ig gene conformation that promote recombination of Ig genes. These events then result in the switching of expression of heavy chain exons in a triggered B cell, leading to the secretion of IgG, IgA, or, in some cases, IgE antibody with the same V region antigen specificity as the original IgM antibody, for response to a wide variety of extracellular bacteria, protozoa, and helminths. CD40 ligand expression by activated T cells is critical for induction of B cell antibody isotype switching and for B cell responsiveness to cytokines. Patients with mutations in T cell CD40 ligand have B cells that are unable to undergo isotype switching, resulting in lack of memory B cell generation and the immunodeficiency syndrome of *X-linked hyper-IgM syndrome* **(Chap. 344)**.

IMMUNE TOLERANCE AND AUTOIMMUNITY

Immune tolerance is defined as the absence of activation of pathogenic autoreactivity to self-antigens. *Autoimmune diseases* are syndromes caused by the activation of T or B cells or both, with no evidence of other causes such as infections or malignancies **(Chap. 348)**. Immune tolerance and autoimmunity are present normally in health; when abnormal, they represent extremes from the normal state. For example, low levels of autoreactivity of T and B cells with self-antigens in the periphery are critical to T and B cell survival. Similarly, low levels of autoreactivity and thymocyte recognition of self-antigens in the thymus are the mechanisms whereby normal T cells are positively selected to survive and leave the thymus to respond to foreign microbes in the periphery and T cells highly reactive to self-antigens are negatively selected and die to prevent overly self-reactive T cells from migrating to the periphery (central tolerance). However, not all self-antigens are expressed in the thymus to delete highly self-reactive T cells, and there are mechanisms for induction of tolerance in peripheral T cells as well. Unlike the presentation of microbial antigens by mature DCs, the presentation of self-antigens by immature DCs neither activates nor matures the DCs to express high levels of co-stimulatory molecules such as B7-1 (CD80) or B7-2 (CD86). When peripheral T cells are stimulated by DCs expressing self-antigens in the context of HLA molecules, sufficient stimulation of T cells occurs to keep them alive, but otherwise they remain anergic, or nonresponsive, until T cells contact a DC with high levels of co-stimulatory molecules expressing microbial antigens and become activated to respond to the microbe. If B cells have high self-reactive BCRs, they normally undergo either deletion in the bone marrow or receptor editing to express a less autoreactive receptor. Although many autoimmune diseases are characterized by abnormal or pathogenic autoantibody production **(Table 342-13)**, most autoimmune diseases are caused by a combination of excess T and B cell reactivity.

TABLE 342-13

Recombinant or Purified Autoantigens Recognized by Autoantibodies Associated with Human Autoimmune Disorders

Autoantigen	Autoimmune Diseases
Cell- or Organ-Specific Autoimmunity	
Acetylcholine receptor	Myasthenia gravis
Actin	Chronic active hepatitis, primary biliary cirrhosis
Adenine nucleotide translator (ANT)	Dilated cardiomyopathy, myocarditis
β-Adrenoreceptor	Dilated cardiomyopathy
Aromatic L-amino acid decarboxylase	Autoimmune polyendocrine syndrome type 1 (APS-1)
Asialoglycoprotein receptor	Autoimmune hepatitis
Bactericidal/permeability-increasing protein (Bpi)	Cystic fibrosis vasculitides
Calcium-sensing receptor	Acquired hypoparathyroidism
Cholesterol side-chain cleavage enzyme (CYP11a)	Autoimmune polyglandular syndrome-1
Collagen type IV-α3-chain	Goodpasture's syndrome
Cytochrome P450 2D6 (CYP2D6)	Autoimmune hepatitis
Desmin	Crohn's disease, coronary artery disease
Desmoglein 1	Pemphigus foliaceus
Desmoglein 3	Pemphigus vulgaris
F-actin	Autoimmune hepatitis
GM gangliosides	Guillain-Barré syndrome
Glutamate decarboxylase (GAD65)	Type 1 diabetes, stiff-person syndrome
Glutamate receptor (GLUR)	Rasmussen encephalitis
H/K ATPase	Autoimmune gastritis
17-α-Hydroxylase (CYP17)	Autoimmune polyglandular syndrome-1
21-Hydroxylase (CYP21)	Addison's disease
IA-2 (ICA512)	Type 1 diabetes
Insulin	Type 1 diabetes, insulin hypoglycemic syndrome (Hirata's disease)
Insulin receptor	Type B insulin resistance, acanthosis, systemic lupus erythematosus (SLE)
Intrinsic factor type 1	Pernicious anemia
Leukocyte function-associated antigen (LFA-1)	Treatment-resistant Lyme arthritis
Myelin-associated glycoprotein (MAG)	Polyneuropathy

Autoantigen	Autoimmune Diseases
Myelin-basic protein	Multiple sclerosis, demyelinating diseases
Myelin oligodendrocyte glycoprotein (MOG)	Multiple sclerosis
Myosin	Rheumatic fever
p-80-Collin	Atopic dermatitis
Pyruvate dehydrogenase complex-E2 (PDC-E2)	Primary biliary cirrhosis
Sodium iodide symporter (NIS)	Graves' disease, autoimmune hypothyroidism
SOX-10	Vitiligo
Thyroid and eye muscle shared protein	Thyroid-associated ophthalmopathy
Thyroglobulin	Autoimmune thyroiditis
Thyroid peroxidase	Autoimmune Hashimoto's thyroiditis
Thyrotropin receptor	Graves' disease
Tissue transglutaminase	Celiac disease
Transcription coactivator p75	Atopic dermatitis
Tryptophan hydroxylase	Autoimmune polyglandular syndrome-1
Tyrosinase	Vitiligo, metastatic melanoma
Tyrosine hydroxylase	Autoimmune polyglandular syndrome-1
Systemic Autoimmunity	
ACTH	ACTH deficiency
Aminoacyl-tRNA histidyl synthetase	Myositis, dermatomyositis
Aminoacyl-tRNA synthetase (several)	Polymyositis, dermatomyositis
Cardiolipin	Systemic lupus erythematosus, antiphospholipid syndrome
Carbonic anhydrase II	Systemic lupus erythematosus, Sjögren's syndrome, systemic sclerosis
Collagen (multiple types)	Rheumatoid arthritis, systemic lupus erythematosus, progressive systemic sclerosis
Centromere-associated proteins	Systemic sclerosis
DNA-dependent nucleoside-stimulated ATPase	Dermatomyositis
Fibrillarin	Scleroderma
Fibronectin	Systemic lupus erythematosus, rheumatoid arthritis, morphea
Glucose-6-phosphate isomerase	Rheumatoid arthritis
β2-Glycoprotein I (B2-GPI)	Primary antiphospholipid syndrome

Autoantigen	Autoimmune Diseases
Golgin (95, 97, 160, 180) Heat shock protein	Sjögren's syndrome, systemic lupus erythematosus, rheumatoid arthritis Various immune-related disorders
Hemidesmosomal protein 180	Bullous pemphigoid, herpes gestationis, cicatricial pemphigoid
Histone H2A-H2B-DNA	Systemic lupus erythematosus
IgE receptor	Chronic idiopathic urticaria
Keratin	Rheumatoid arthritis
Ku-DNA-protein kinase	Systemic lupus erythematosus
Ku-nucleoprotein La phosphoprotein (La 55-B)	Connective tissue syndrome Sjögren's syndrome
Myeloperoxidase	Necrotizing and crescentic glomerulonephritis (NCGN), systemic vasculitis
Proteinase 3 (PR3)	Granulomatosis with polyangiitis (Wegener's), Churg-Strauss syndrome
RNA polymerase I–III (RNP)	Systemic sclerosis, systemic lupus erythematosus
Signal recognition protein (SRP54)	Polymyositis
Topoisomerase-1 (Scl-70)	Scleroderma, Raynaud's syndrome
Tublin	Chronic liver disease, visceral leishmaniasis
Vimentin	Systemic autoimmune disease
Plasma Protein and Cytokine Autoimmunity	
C1 inhibitor	Autoimmune C1 deficiency
C1q	Systemic lupus erythematosus, membrane proliferative glomerulonephritis (MPGN)
Cytokines (IL-1α, IL-1β, IL-6, IL-10, LIF)	Rheumatoid arthritis, systemic sclerosis, normal subjects
Factor II, factor V, factor VII, factor VIII, factor IX, factor X, factor XI, thrombin vWF	Prolonged coagulation time
Glycoprotein IIb/IIIg and Ib/IX	Autoimmune thrombocytopenia purpura
IgA	Immunodeficiency associated with systemic lupus erythematosus, pernicious anemia, thyroiditis, Sjögren's syndrome, and chronic active hepatitis
Oxidized LDL (OxLDL)	Atherosclerosis
Cancer and Paraneoplastic Autoimmunity	
Amphiphysin	Neuropathy, small-cell lung cancer
Cyclin B1	Hepatocellular carcinoma
DNA topoisomerase II	Liver cancer
Desmoplakin	Paraneoplastic pemphigus
Gephyrin	Paraneoplastic stiff-person syndrome

Autoantigen	Autoimmune Diseases
Hu proteins	Paraneoplastic encephalomyelitis
Neuronal nicotinic acetylcholine receptor	Subacute autonomic neuropathy, cancer
p53	Cancer, systemic lupus erythematosus
p62 (IGF-II mRNA-binding protein)	Hepatocellular carcinoma (China)
Recoverin	Cancer-associated retinopathy
Ri protein	Paraneoplastic opsoclonus myoclonus ataxia
βIV spectrin	Lower motor neuron syndrome
Synaptotagmin	Lambert-Eaton myasthenic syndrome
Voltage-gated calcium channels	Lambert-Eaton myasthenic syndrome
Yo protein	Paraneoplastic cerebellar degeneration

Source: From A Lernmark et al: J Clin Invest 108:1091, 2001; with permission.

Multiple factors contribute to the genesis of autoimmune disease syndromes, including genetic susceptibility (HLAB27 with ankylosing spondylitis) (Table 342-13), environmental immune stimulants such as drugs (e.g., procainamide and phenytoin [Dilantin] with drug-induced systemic lupus erythematosus), infectious agent triggers (such as Epstein-Barr virus and autoantibody production against red blood cells and platelets), and loss of T regulatory cells (leading to thyroiditis, adrenalitis, and oophoritis).

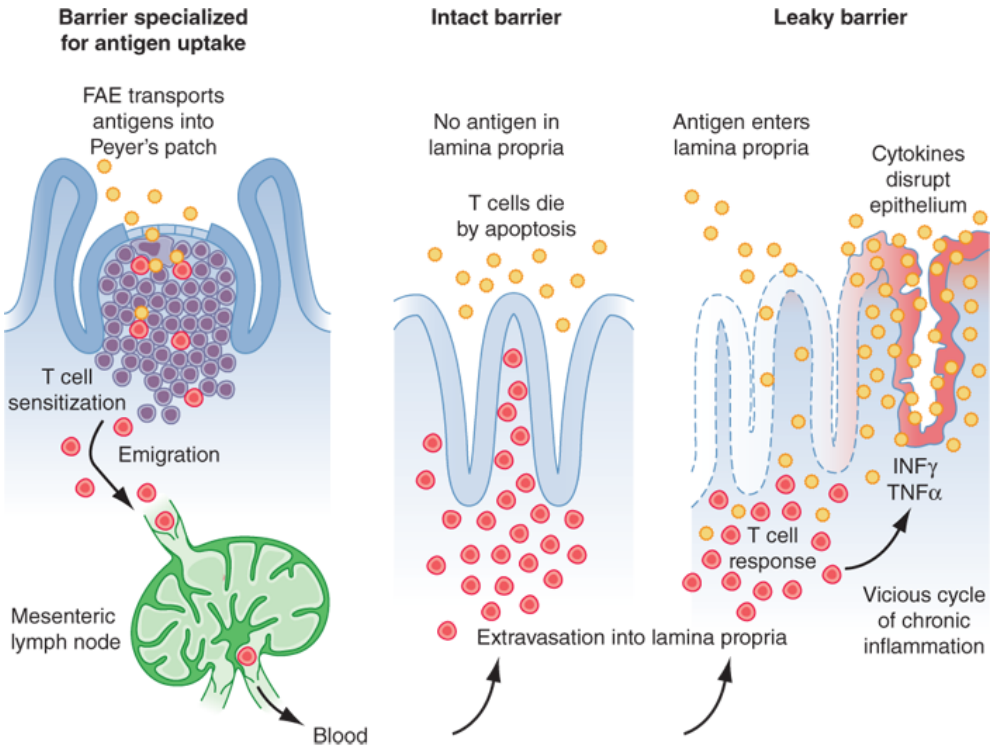
Immunity at Mucosal Surfaces

Mucosa covering the respiratory, digestive, and urogenital tracts; the eye conjunctiva; the inner ear; and the ducts of all exocrine glands contain cells of the innate and adaptive mucosal immune system that protect these surfaces against pathogens. In the healthy adult, mucosa-associated lymphoid tissue (MALT) contains 80% of all immune cells within the body and constitutes the largest mammalian lymphoid organ system.

MALT has three main functions: (1) to protect the mucous membranes from invasive pathogens; (2) to prevent uptake of foreign antigens from food, commensal organisms, and airborne pathogens and particulate matter; and (3) to prevent pathologic immune responses from foreign antigens if they do cross the mucosal barriers of the body **(Fig. 342-8)**.

FIGURE 342-8

Increased epithelial permeability may be important in the development of chronic gut T cell–mediated inflammation. CD4 T cells activated by gut antigens in Peyer's patches migrate to the lamina propria (LP). In healthy individuals, these cells die by apoptosis. Increased epithelial permeability may allow sufficient antigen to enter the LP to trigger T cell activation, breaking tolerance mediated by immunosuppressive cytokines and perhaps T regulatory cells. Proinflammatory cytokines then further increase epithelial permeability, setting up a vicious cycle of chronic inflammation. *(From TT MacDonald et al: Science 307:1920, 2005; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MALT is a compartmentalized system of immune cells that functions independently from systemic immune organs. Whereas the systemic immune organs are essentially sterile under normal conditions and respond vigorously to pathogens, MALT immune cells are continuously bathed in foreign proteins and commensal bacteria, and they must select those pathogenic antigens that must be eliminated. MALT contains anatomically defined foci of immune cells in the intestine, tonsil, appendix, and peribronchial areas that are inductive sites for mucosal immune responses. From these sites, immune T and B cells migrate to effector sites in mucosal parenchyma and exocrine glands where mucosal immune cells eliminate pathogen-infected cells. In addition to mucosal immune responses, all mucosal sites have strong mechanical and chemical barriers and cleansing functions to repel pathogens.

Key components of MALT include specialized epithelial cells called “membrane” or “M” cells that take up antigens and deliver them to DCs or other APCs. Effector cells in MALT include B cells producing antipathogen neutralizing antibodies of secretory IgA as well as IgG isotype, T cells producing similar cytokines as in systemic immune system response, and T helper and cytotoxic T cells that respond to pathogen-infected cells.

Secretory IgA is produced in amounts of >50 mg/kg of body weight per 24 h and functions to inhibit bacterial adhesion, inhibit macromolecule absorption in the gut, neutralize viruses, and enhance antigen elimination in tissue through binding to IgA and receptor-mediated transport of immune complexes through epithelial cells.

Recent studies have demonstrated the importance of commensal gut and other mucosal bacteria to the health of the human immune system. Normal commensal flora induces anti-inflammatory events in the gut and protects epithelial cells from pathogens through TLRs and other PRR signaling. When the gut is depleted of normal commensal flora, the immune system becomes abnormal, with loss of T_H1 T cell function. Restoration of the normal gut flora can reestablish the balance in T helper cell ratios characteristic of the normal immune system. Diet also has an impact on the gut microbiome. Altered microbiome composition has been etiologically related to obesity, insulin resistance and diabetes. When the gut barrier is intact, either antigens do not transverse the gut epithelium or, when pathogens are present, a self-limited, protective MALT immune response eliminates the pathogen (Fig. 342-8). However, when the gut barrier breaks down, immune responses to commensal flora antigens can cause inflammatory bowel diseases such as *Crohn's disease* and, perhaps, *ulcerative colitis* (Fig. 342-8) **(Chap. 319)**. Uncontrolled MALT immune responses to food antigens, such as gluten, can cause *celiac disease* **(Chap. 319)**.

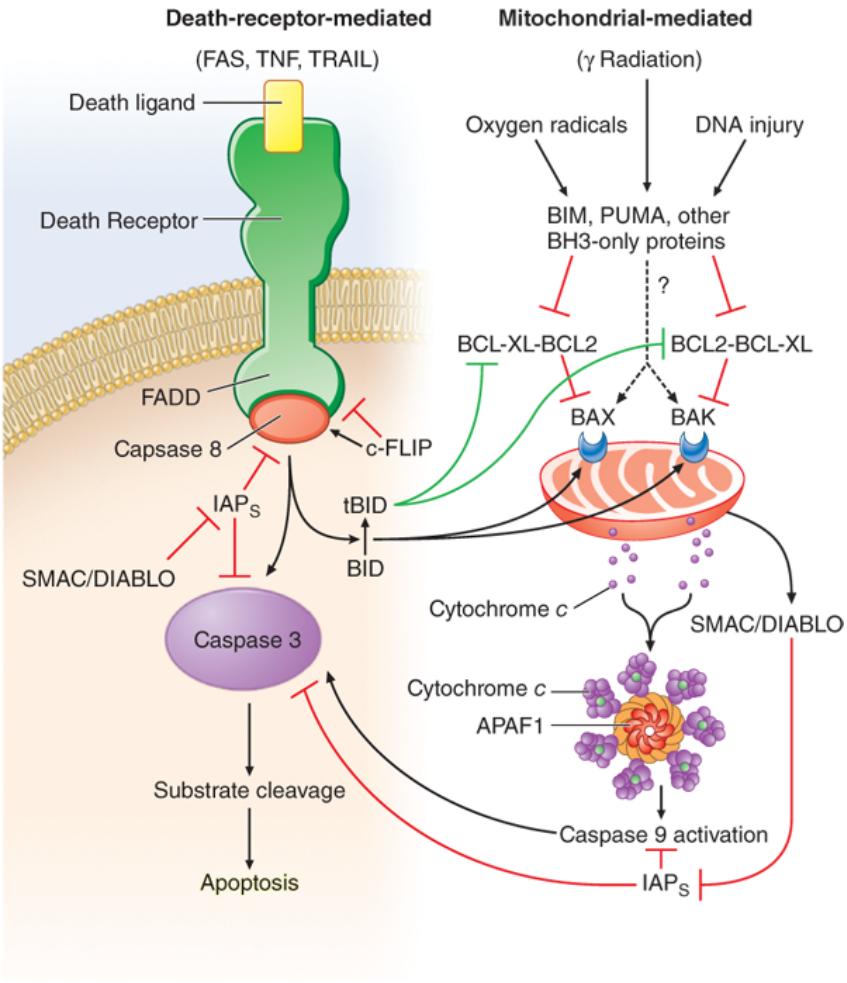
THE CELLULAR AND MOLECULAR CONTROL OF PROGRAMMED CELL DEATH

The process of apoptosis (programmed cell death) plays a crucial role in regulating normal immune responses to antigen. In general, a wide variety of stimuli trigger one of several apoptotic pathways to eliminate microbe-infected cells, eliminate cells with damaged DNA, or eliminate activated immune cells that are no longer needed **(Fig. 342-9)**. The largest known family of “death receptors” is the TNF receptor (TNF-R) family (TNF-R1, TNF-R2, Fas [CD95], death receptor 3 [DR3], death receptor 4 [DR4; TNF-related apoptosis-including ligand receptor 1, or TRAIL-R1], and death receptor 5 [DR5, TRAIL-R2]); their ligands are all in the TNF-α family. Binding of ligands to these death receptors leads to a signaling cascade that involves activation of the *caspase* family of molecules that leads to DNA cleavage and cell death. Two other pathways of programmed cell death involve nuclear *p53* in the elimination of cells with abnormal DNA and *mitochondrial cytochrome c* to induce cell death in damaged cells (Fig. 342-9). A number of human diseases have now been described

that result from, or are associated with, mutated apoptosis genes (Table 342-14). These include mutations in the Fas and Fas ligand genes in autoimmune and lymphoproliferation syndromes, and multiple associations of mutations in genes in the apoptotic pathway with malignant syndromes.

FIGURE 342-9

Pathways of cellular apoptosis. There are two major pathways of apoptosis: the death-receptor pathway, which is mediated by activation of death receptors, and the BCL2-regulated mitochondrial pathway, which is mediated by noxious stimuli that ultimately lead to mitochondrial injury. Ligation of death receptors recruits the adaptor protein FAS-associated death domain (FADD). FADD in turn recruits caspase 8, which ultimately activates caspase 3, the key "executioner" caspase. Cellular FLICE-inhibitory protein (c-FLIP) can either inhibit or potentiate binding of FADD and caspase 8, depending on its concentration. In the intrinsic pathway, proapoptotic BH3 proteins are activated by noxious stimuli, which interact with and inhibit antiapoptotic BCL2 or BCL-XL. Thus, BAX and BAK are free to induce mitochondrial permeabilization with release of cytochrome c, which ultimately results in the activation of caspase 9 through the apoptosome. Caspase 9 then activates caspase 3. SMAC/DIABLO is also released after mitochondrial permeabilization and acts to block the action of inhibitors of apoptosis protein (IAPs), which inhibit caspase activation. There is potential cross-talk between the two pathways, which is mediated by the truncated form of BID (tBID) that is produced by caspase 8–mediated BID cleavage; tBID acts to inhibit the BCL2-BCL-XL pathway and to activate BAX and BAK. There is debate (indicated by the *question mark*) as to whether proapoptotic BH3 molecules (e.g., BIM and PUMA) act directly on BAX and BAK to induce mitochondrial permeability or whether they act only on BCL2-BCL-XL. APAF1, apoptotic protease-activating factor 1; BH3, BCL homologue; TNF, tumor necrosis factor; TRAIL, TNF-related apoptosis-inducing ligand. *(From RS Hotchkiss et al: N Engl J Med 361:1570, 2009; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 342-14

Immune System Molecule Defects in Animals or Humans That Cause Autoimmune or Malignant Syndromes

Protein	Defect	Disease or Syndrome	Observation in Animal Models or Humans
Cytokines and Signaling Proteins			
Tumor necrosis factor (TNF) α	Overexpression	Inflammatory bowel disease (IBD), arthritis, vasculitis	Mice
TNF-α	Underexpression	Systemic lupus erythematosus (SLE)	Mice
Interleukin (IL)-1-receptor antagonist	Underexpression	Arthritis	Mice
IL-2	Overexpression	IBD	Mice
IL-7	Overexpression	IBD	Mice
IL-10	Overexpression	IBD	Mice
IL-2 receptor	Overexpression	IBD	Mice
IL-10 receptor	Overexpression	IBD	Mice
IL-3	Overexpression	Demyelinating syndrome	Mice
Interferon-δ	Overexpression in skin	SLE	Mice
STAT-3	Underexpression	IBD	Mice
STAT-4	Overexpression	IBD	Mice
Transforming growth factor (TGF) β	Underexpression	Systemic wasting syndrome and IBD	Mice
TGF-β receptor in T cells	Underexpression	SLE	Mice
Programmed death (CD279, PD-1)	Underexpression	SLE-like syndrome	Mice
Cytotoxic T lymphocyte, antigen-4 (CTLA-4)	Underexpression	Systemic lymphoproliferative disease	Mice
IL-10	Underexpression	IBD (mouse), type 1 diabetes, thyroid disease, primary (human)	Mice and humans
Major Histocompatibility Locus Molecules[a]			
HLA-B27	Allele expression or overexpression	Inflammatory bowel disease	Rats and humans
Complement deficiency of C1, 2, 3 or 4	Underexpression		Humans
LIGHT (TNF superfamily 14)	Overexpression	Systemic lymphoproliferative (mouse) and autoimmunity	Mice
HLA class II DQB10301, DQB10302	Allele expression	Juvenile-onset diabetes	Humans

Protein	Defect	Disease or Syndrome	Observation in Animal Models or Humans
HLA class II DQB10401, DQB10402	Allele expression	Rheumatoid arthritis	Humans
HLA class I B27	Allele expression	Ankylosing spondylitis, IBD	Rats and humans
Apoptosis Proteins			
TNF receptor 1 (TNF-R1)	Underexpression	Familial periodic fever syndrome	Humans
Fas (CD95; Apo-1)	Underexpression	Autoimmune lymphoproliferative syndrome type 1 (ALPS 1); malignant lymphoma; bladder cancer	Humans
Fas ligand	Underexpression	SLE (only one case identified)	Humans
Perforin	Underexpression	Familial hemophagocytic lymphohistiocytosis (FHL)	Humans
Caspase 10	Underexpression	Autoimmune lymphoproliferative syndrome type II (ALPS II)	Humans
bcl-10	Underexpression	Non-Hodgkin's lymphoma	Humans
P53	Underexpression	Various malignant neoplasms	Humans
Bax	Underexpression	Colon cancer; hematopoietic malignancies	Humans
bcl-2	Underexpression	Non-Hodgkin's lymphoma	Humans
c-IAP2	Underexpression	Low-grade MALT lymphoma	Humans
NAIP1	Underexpression	Spinal muscular atrophy	Humans

[a]Many autoimmune diseases are associated with a myriad of major histocompatibility complex gene allele (HLA) types. They are presented here as examples.

Abbreviation: MALT, mucosa-associated lymphoid tissue.

Source: Adapted from L Mullauer: Mutat Res 488:211, 2001; and A Davidson, B Diamond: N Engl J Med 345:340, 2001.

MECHANISMS OF IMMUNE-MEDIATED DAMAGE TO MICROBES OR HOST TISSUES

Several responses by the host innate and adaptive immune systems to foreign microbes culminate in rapid and efficient elimination of microbes. In these scenarios, the classic weapons of the adaptive immune system (T cells, B cells) interface with cells (macrophages, DCs, NK cells, neutrophils, eosinophils, basophils) and soluble products (microbial peptides, pentraxins, complement and coagulation systems) of the innate immune system **(Chaps. 60 and 345)**.

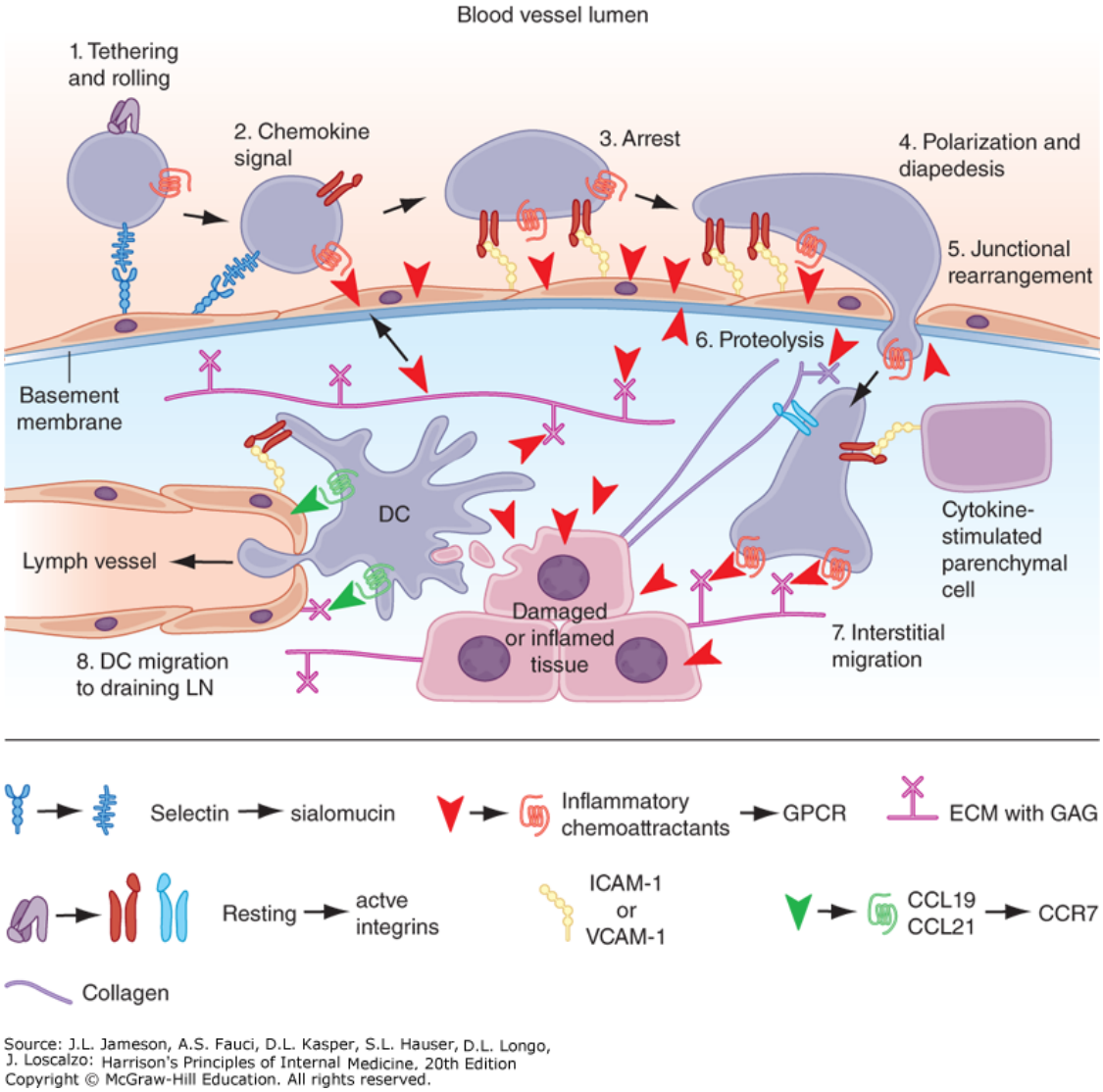
There are five general phases of host defenses: (1) migration of leukocytes to sites of antigen localization; (2) antigen-nonspecific recognition of pathogens by macrophages and other cells and systems of the innate immune system; (3) specific recognition of foreign antigens mediated by T and B lymphocytes; (4) amplification of the inflammatory response with recruitment of specific and nonspecific effector cells by complement components, cytokines, kinins, arachidonic acid metabolites, and mast cell–basophil products; and (5) macrophage, neutrophil, and lymphocyte participation in destruction of antigen with ultimate removal of antigen particles by phagocytosis (by macrophages or neutrophils) or by direct cytotoxic mechanisms (involving macrophages, neutrophils, DCs, and lymphocytes). Under normal circumstances, orderly progression of host defenses through these phases results in a well-controlled immune and inflammatory response that protects the host from the offending antigen. However, dysfunction of any of the host defense systems can damage host tissue and produce clinical disease. Furthermore, for certain pathogens or antigens, the normal immune response itself might contribute substantially to the tissue damage. For example, the immune and inflammatory response in the brain to certain pathogens such as *M. tuberculosis* may be responsible for much of the morbidity rate of this disease in that organ system **(Chap. 173)**. In addition, the morbidity rate associated with certain pneumonias such as that caused by *Pneumocystis jiroveci* may be associated more with inflammatory infiltrates than with the tissue-destructive effects of the microorganism itself **(Chap. 215)**.

Molecular Basis of Lymphocyte–Endothelial Cell Interactions

The control of lymphocyte circulatory patterns between the bloodstream and peripheral lymphoid organs operates at the level of lymphocyte–endothelial cell interactions to control the specificity of lymphocyte subset entry into organs. Similarly, lymphocyte–endothelial cell interactions regulate the entry of lymphocytes into inflamed tissue. Adhesion molecule expression on lymphocytes and endothelial cells regulates the retention and subsequent egress of lymphocytes within tissue sites of antigenic stimulation, delaying cell exit from tissue and preventing reentry into the circulating lymphocyte pool **(Fig. 342-10)**. All types of lymphocyte migration begin with lymphocyte attachment to specialized regions of vessels, termed *high endothelial venules* (HEVs). An important concept is that adhesion molecules do not generally bind their ligand until a conformational change (ligand activation) occurs in the adhesion molecule that allows ligand binding. Induction of a conformation-dependent determinant on an adhesion molecule can be accomplished by cytokines or via ligation of other adhesion molecules on the cell.

FIGURE 342-10

Key migration steps of immune cells at sites of inflammation. Inflammation due to tissue damage or infection induces the release of cytokines (not shown) and inflammatory chemoattractants (*red arrowheads*) from distressed stromal cells and "professional" sentinels, such as mast cells and macrophages (not shown). The inflammatory signals induce upregulation of endothelial selectins and immunoglobulin "superfamily" members, particularly ICAM-1 and/or VCAM-1. Chemoattractants, particularly chemokines, are produced by or translocated across venular endothelial cells (*red arrow*) and are displayed in the lumen to rolling leukocytes. Those leukocytes that express the appropriate set of trafficking molecules undergo a multistep adhesion cascade (steps 1–3) and then polarize and move by diapedesis across the venular wall (steps 4 and 5). Diapedesis involves transient disassembly of endothelial junctions and penetration through the underlying basement membrane (step 6). Once in the extravascular (interstitial) space, the migrating cell uses different integrins to gain "footholds" on collagen fibers and other ECM molecules, such as laminin and fibronectin, and on inflammation-induced ICAM-1 on the surface of parenchymal cells (step 7). The migrating cell receives guidance cues from distinct sets of chemoattractants, particularly chemokines, which may be immobilized on glycosaminoglycans (GAG) that "decorate" many ECM molecules and stromal cells. Inflammatory signals also induce tissue dendritic cells (DCs) to undergo maturation. Once DCs process material from damaged tissues and invading pathogens, they upregulate CCR7, which allows them to enter draining lymph vessels that express the CCR7 ligand CCL21 (and CCL19). In lymph nodes (LNs), these antigen-loaded mature DCs activate naïve T cells and expand pools of effector lymphocytes, which enter the blood and migrate back to the site of inflammation. T cells in tissue also use this CCR7-dependent route to migrate from peripheral sites to draining lymph nodes through afferent lymphatics. *(Adapted from AD Luster et al: Nat Immunol 6:1182, 2005; with permission from Macmillan Publishers Ltd. Copyright 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The first stage of lymphocyte–endothelial cell interactions, *attachment and rolling*, occurs when lymphocytes leave the stream of flowing blood cells in a postcapillary venule and roll along venule endothelial cells (Fig. 342-10). Lymphocyte rolling is mediated by the L-selectin molecule (LECAM-1, LAM-1, CD62L) and slows cell transit time through venules, allowing time for activation of adherent cells.

The second stage of lymphocyte–endothelial cell interactions, *firm adhesion with activation-dependent stable arrest*, requires stimulation of lymphocytes by chemoattractants or by endothelial cell–derived cytokines. Cytokines thought to participate in adherent cell activation include members of the IL-8 family, platelet-activation factor, leukotriene B_4, and C5a. In addition, HEVs express chemokines, SLC (CCL21) and ELC (CCL19), which participate in this process. Following activation by chemoattractants, lymphocytes shed L-selectin from the cell surface and upregulate cell CD11b/18 (MAC-1) or CD11a/18 (LFA-1) molecules, resulting in firm attachment of lymphocytes to HEVs.

Lymphocyte homing to peripheral lymph nodes involves adhesion of L-selectin to glycoprotein HEV ligands collectively referred to as *peripheral node addressin* (*PNAd*), whereas homing of lymphocytes to intestine Peyer's patches primarily involves adhesion of the a4β7 integrin to mucosal addressin cell adhesion molecule-1 (MAdCAM-1) on the Peyer's patch HEVs. However, for migration to mucosal Peyer's patch lymphoid aggregates, naïve lymphocytes primarily use L-selectin, whereas memory lymphocytes use a4β7 integrin. a4β1 integrin (CD49d/CD29, VLA-4)–VCAM-1 interactions are important in the initial interaction of memory lymphocytes with HEVs of multiple organs in sites of inflammation **(Table 342-15)**.

TABLE 342-15

Trafficking Molecules Involved In Inflammatory Disease Processes

		Proposed Leukocyte Receptors for Endothelial Traffic Signals		
Disease	**Key Effector Cell**	**L-Selectin, Ligand**	**GPCR**	**Integrin[a]**
Acute Inflammation				
Myocardial infarction	Neutrophil	PSGL-1	CXCR1, CXCR2, PAFR, BLT1	LFA-1, Mac-1
Stroke	Neutrophil	L-Selectin, PSGL-1	CXCR1, CXCR2, PAFR, BLT1	LFA-1, Mac-1
Ischemia-reperfusion	Neutrophil	PSGL-1	CXCR1, CXCR2, PAFR, BLT1	LFA-1, Mac-1
T_H1 Inflammation				
Atherosclerosis	Monocyte	PSGL-1	CCR1, CCR2, BLT1, CXCR2, CX3CR1	VLA-4
	T_H1	PSGL-1	CXCR3, CCR5	VLA-4
Multiple sclerosis	T_H1	PSGL-1 (?)	CXCR3, CXCR6	VLA-4, LFA-1
	Monocyte	PSGL-1 (?)	CCR2, CCR1	VLA-4, LFA-1
Rheumatoid arthritis	Monocyte	PSGL-1	CCR1, CCR2	VLA-1, VLA-2, VLA-4, LFA-1
	T_H1	PSGL-1	CXCR3, CXCR6	VLA-1, VLA-2, VLA-4, LFA-1
	Neutrophil	L-Selectin, PSGL-1	CXCR2, BLT1	LFA-1[b]
Psoriasis	Skin-homing T_H1	CLA	CCR4, CCR10, CXCR3	VLA-4[c], LFA-1
Crohn's disease	Gut-homing T_H1	PSGL-1	CCR9, CXCR3	α4, β7, LFA-1
Type 1 diabetes	T_H1	PSGL-1 (?)	CCR4, CCR5	VLA-4, LFA-1
	CD8	L-Selectin (?), PSGL-1 (?)	CXCR3	VLA-4, LFA-1
Allograft rejection	CD8	PSGL-1	CXCR3, CX3CR1, BLT1	VLA-4, LFA-1
	B cell	L-Selectin, PSGL-1	CXCR5, CXCR4	VLA-4, LFA-1
Hepatitis	CD8	PSGL-1	CXCR3, CCR5, CXCR6	VLA-4
Lupus	T_H1	None	CXCR6	VLA-4[d]
	Plasmacytoid DC	L-Selectin, CLA	CCR7, CXCR3, ChemR23	LFA-1, Mac-1
	B cell	CLA (?)	CXCR5, CXCR4	LFA-1
T_H2 Inflammation				
Asthma	T_H2	PSGL-1	CCR4, CCR8, BLT1	LFA-1
	Eosinophil	PSGL-1	CCR3, PAFR, BLT1	VLA-4, LFA-1
	Mast cells	PSGL-1	CCR2, CCR3, BLT1	VLA-4, LFA-1

		Proposed Leukocyte Receptors for Endothelial Traffic Signals		
Disease	**Key Effector Cell**	**L-Selectin, Ligand**	**GPCR**	**Integrin[a]**
Atopic dermatitis	Skin-homing T_H2	CLA	CCR4, CCR10	VLA-4, LFA-1

[a]Various β_1 integrins have been linked in different ways in basal lamina and interstitial migration of distinct cell types and inflammatory settings. [b]In some settings, Mac-1 has been linked to transmigration. [c]CD44 can act in concert with VLA-4 in particular models of leukocyte arrest. [d]T_H2 cells require VAP-1 to traffic to inflamed liver.

Source: From AD Luster et al: Nat Immunol 6:1182, 2005; with permission from Macmillan Publishers Ltd. Copyright 2005.

The third stage of leukocyte emigration in HEVs is *sticking and arrest*. Sticking of the lymphocyte to endothelial cells and arrest at the site of sticking are mediated predominantly by ligation of a1β2 integrin LFA-1 to the integrin ligand ICAM-1 on HEVs. Whereas the first three stages of lymphocyte attachment to HEVs take only a few seconds, the fourth stage of lymphocyte emigration, *transendothelial migration*, takes ~10 min. Although the molecular mechanisms that control lymphocyte transendothelial migration are not fully characterized, the HEV CD44 molecule and molecules of the HEV glycocalyx (extracellular matrix) are thought to play important regulatory roles in this process (Fig. 342-10). Finally, expression of matrix metalloproteases capable of digesting the subendothelial basement membrane, rich in nonfibrillar collagen, appears to be required for the penetration of lymphoid cells into the extravascular sites.

Abnormal induction of HEV formation and use of the molecules discussed above have been implicated in the induction and maintenance of inflammation in a number of chronic inflammatory diseases. In animal models of type 1 diabetes mellitus, MAdCAM-1 and GlyCAM-1 have been shown to be highly expressed on HEVs in inflamed pancreatic islets, and treatment of these animals with inhibitors of L-selectin and a4 integrin function blocked the development of type 1 diabetes mellitus **(Chap. 396)**. A similar role for abnormal induction of the adhesion molecules of lymphocyte emigration has been suggested in *rheumatoid arthritis* **(Chap. 351)**, *Hashimoto's thyroiditis* **(Chap. 375)**, *Graves' disease* **(Chap. 375)**, *multiple sclerosis* **(Chap. 436)**, *Crohn's disease* **(Chap. 319)**, and *ulcerative colitis* **(Chap. 319)**.

Immune-Complex Formation

Clearance of antigen by immune-complex formation between antigen, complement, and antibody is a highly effective mechanism of host defense. However, depending on the level of immune complexes formed and their physicochemical properties, immune complexes may or may not result in host and foreign cell damage. After antigen exposure, certain types of soluble antigen-antibody complexes freely circulate and, if not cleared by the reticuloendothelial system, can be deposited in blood vessel walls and in other tissues such as renal glomeruli and cause *vasculitis* or *glomerulonephritis* syndromes **(Chaps. 308 and 356)**. Deficiencies of early complement components are associated with inefficient clearance of immune complexes and immune complex mediated tissue damage in autoimmune syndromes, whereas deficiencies of the later complement components are associated with susceptibility to recurrent *Neisseria* infections **(Table 342-16)**.

TABLE 342-16

Complement Deficiencies and Associated Diseases

Component	Associated Diseases
Classic Pathway	
Clq, Clr, Cls, C4	Immune-complex syndromes,[a] pyogenic infections
C2	Immune-complex syndromes,[a] few with pyogenic infections
C1 inhibitor	Rare immune-complex disease, few with pyogenic infections
C3 and Alternative Pathway C3	
C3	Immune-complex syndromes,[a] pyogenic infections
D	Pyogenic infections
Properdin	*Neisseria* infections
I	Pyogenic infections
H	Hemolytic-uremic syndrome
Membrane Attack Complex	
C5, C6, C7, C8	Recurrent *Neisseria* infections, immune-complex disease
C9	Rare *Neisseria* infections

[a]Immune-complex syndromes include systemic lupus erythematosus (SLE) and SLE-like syndromes, glomerulonephritis, and vasculitis syndromes.

Source: After JA Schifferli, DK Peters: Lancet 322:957, 1983. Copyright 1983, with permission from Elsevier.

Immediate-Type Hypersensitivity

Helper T cells that drive antiallergen IgE responses are usually T_H2-type inducer T cells that secrete IL-4, IL-5, IL-6, and IL-10. Mast cells and basophils have high-affinity receptors for the Fc portion of IgE (FcRI), and cell-bound antiallergen IgE effectively "arms" basophils and mast cells. Mediator release is triggered by antigen (allergen) interaction with Fc receptor-bound IgE, and the mediators released are responsible for the pathophysiologic changes of *allergic diseases* (Table 342-11). Mediators released from mast cells and basophils can be divided into three broad functional types: (1) those that increase vascular permeability and contract smooth muscle (histamine, platelet-activating factor, SRS-A, BK-A), (2) those that are chemotactic for or activate other inflammatory cells (ECF-A, NCF, leukotriene B_4), and (3) those that modulate the release of other mediators (BK-A, platelet-activating factor) **(Chap. 345)**.

Cytotoxic Reactions of Antibody

In this type of immunologic injury, complement-fixing (C1-binding) antibodies against normal or foreign cells or tissues (IgM, IgG1, IgG2, IgG3) bind complement via the classic pathway and initiate a sequence of events similar to that initiated by immune-complex deposition, resulting in cell lysis or tissue injury. Examples of antibody-mediated cytotoxic reactions include red cell lysis in *transfusion reactions*, *Goodpasture's syndrome* with anti–glomerular basement membrane antibody formation, and *pemphigus vulgaris* with antiepidermal antibodies inducing blistering skin disease.

Delayed-Type Hypersensitivity Reactions

Inflammatory reactions initiated by mononuclear leukocytes and not by antibody alone have been termed *delayed-type hypersensitivity reactions*. The term *delayed* has been used to contrast a secondary cellular response that appears 48–72 h after antigen exposure with an *immediate* hypersensitivity response generally seen within 12 h of antigen challenge and initiated by basophil mediator release or preformed antibody. For example, in an individual previously infected with *M. tuberculosis* organisms, intradermal placement of tuberculin purified protein derivative as a skin test challenge results in an indurated area of skin at 48–72 h, indicating previous exposure to tuberculosis.

The cellular events that result in classic delayed-type hypersensitivity responses are centered on T cells (predominantly, although not exclusively, IFN-γ, IL-2, and TNF-α-secreting T_H1-type helper T cells) and macrophages. Recently, NK cells have been suggested to play a major role in the form of delayed hypersensitivity that occurs following skin contact with immunogens. First, local immune and inflammatory responses at the site of foreign antigen upregulate endothelial cell adhesion molecule expression, promoting the accumulation of lymphocytes at the tissue site. In the general scheme outlined in Fig. 342-2, antigen is processed by DCs and presented to small numbers of CD4+ T cells expressing a TCR specific for the antigen. IL-12 produced by APCs induces T cells to produce IFN-γ (T_H1 response). Macrophages frequently undergo epithelioid cell transformation and fuse to form multinucleated giant cells in response to IFN-γ. This type of mononuclear cell infiltrate is termed *granulomatous inflammation*. Examples of diseases in which delayed-type hypersensitivity plays a major role are fungal infections (*histoplasmosis*; **Chap. 207**), mycobacterial infections (*tuberculosis*, *leprosy*; **Chaps. 173 and 174**), chlamydial infections (*lymphogranuloma venereum*; **Chap. 184**), helminth infections (*schistosomiasis*; **Chap. 229**), reactions to toxins (*berylliosis*; **Chap. 283**), and hypersensitivity reactions to organic dusts (*hypersensitivity pneumonitis*; **Chap. 282**). In addition, delayed-type hypersensitivity responses play important roles in tissue damage in autoimmune diseases such as *rheumatoid arthritis*, *temporal arteritis*, and granulomatosis with polyangiitis (Wegener's) **(Chaps. 351 and 356)**.

Autophagy

Autophagy is a process that involves a lysosomal degradation pathway mechanism of cells to dispose of intracellular debris and damaged organelles. Autophagy by cells of the innate immune system is used to control intracellular infectious agents such as *Mycobacterium tuberculosis*, in part by initiation of phagosome maturation and enhancing MHC class II antigen presentation to CD4 T cells.

CLINICAL EVALUATION OF IMMUNE FUNCTION

Clinical assessment of immunity requires investigation of the four major components of the immune system that participate in host defense and in the pathogenesis of autoimmune diseases: (1) humoral immunity (B cells); (2) cell-mediated immunity (T cells, monocytes); (3) phagocytic cells of the reticuloendothelial system (macrophages), as well as polymorphonuclear leukocytes; and (4) complement. Clinical problems that require an evaluation of immunity include chronic infections, recurrent infections, unusual infecting agents, and certain autoimmune syndromes. The type of clinical syndrome under evaluation can provide information regarding possible immune defects **(Chap. 344)**. Defects in cellular immunity generally result in viral, mycobacterial, and fungal infections. An extreme example of deficiency in cellular immunity is AIDS **(Chap. 197)**. Antibody deficiencies result in recurrent bacterial infections, frequently with organisms such as *S. pneumoniae* and *Haemophilus influenzae* **(Chap. 344)**. Disorders of phagocyte function are frequently manifested by recurrent skin infections, often due to *Staphylococcus aureus* **(Chap. 60)**. Finally, deficiencies of early and late complement components are associated with autoimmune phenomena and recurrent *Neisseria* infections (Table 342-16). **For further discussion of useful initial screening tests of immune function, see Chap. 344**.

IMMUNOTHERAPY

Many therapies for autoimmune and inflammatory diseases involve the use of nonspecific immune-modulating or immunosuppressive agents such as glucocorticoids or cytotoxic drugs. The goal of development of new treatments for immune-mediated diseases is to design ways to specifically interrupt pathologic immune responses, leaving nonpathologic immune responses intact. Novel ways to interrupt pathologic immune responses that are under investigation include the use of anti-inflammatory cytokines or specific cytokine inhibitors as anti-inflammatory agents, the use of monoclonal antibodies against T or B lymphocytes as therapeutic agents, the use of intravenous Ig for certain infections and immune complex–mediated diseases, the use of specific cytokines to reconstitute components of the immune system, and bone marrow transplantation to replace the pathogenic immune system with a more normal immune system **(Chaps. 60, 344, and 197)**. In particular, the use of a monoclonal antibody to B cells (rituximab, anti-CD20 MAb) is approved in the United States for the treatment of non-Hodgkin's lymphoma **(Chap. 104)** and, in combination with methotrexate, for treatment of adult patients with severe rheumatoid arthritis resistant to TNF-α inhibitors **(Chap. 351)**. CTLA-4 inhibitors such as Ipilimumab and Tremelimumab and anti-PD-1 antibodies such as Nivolumab have been shown to reverse CD8 T cell exhaustion in melanoma and other solid tumors and induce immune cell control of tumor growth. CTLA4 and PD-1 inhibitors are currently being studied in HCV or HIV-1 infection to reverse anti-viral CD8 T cell dysfunction and promote the reduction of virus infected cells. A new technique that engineers autologous T cells to express antibody receptors that target leukemic cells, termed T cells with chimeric antigen receptors (T CARs), is currently showing promising results in clinical trials for the treatment of certain types of leukemias and lymphomas.

Cell-based therapies have been studied for many years, including ex vivo activation of NK cells for reinfusion into patients with malignancies, and DC therapy of ex vivo priming of DCs for enhanced presentation of cancer antigens, with reinfusion of primed DCs into the patient. One such strategy for DC therapy has been approved by the FDA for treatment of advanced prostate cancer.

Cytokines and Cytokine Inhibitors

Several TNF inhibitors are used as biological therapies in the treatment of rheumatoid arthritis; these include monoclonal antibodies, TNF-R Fc fusion proteins, and Fab fragments. Use of anti-TNF-α antibody therapies such as adalimumab, infliximab, and golimumab has resulted in clinical improvement in patients with these diseases and has opened the way for targeting TNF-α to treat other severe forms of autoimmune and/or inflammatory disease. Blockage of TNF-α has been effective in *rheumatoid arthritis*, *psoriasis*, *Crohn's disease*, and *ankylosing spondylitis*. Other cytokine inhibitors are recombinant soluble

TNF-α receptor (R) fused to human Ig and anakinra (soluble *IL-1 receptor antagonist*, or IL-1ra). The treatment of autoinflammatory syndromes (Table 342-6) with recombinant IL-1 receptor antagonist can prevent symptoms in these syndromes, because the overproduction of IL-1β is a hallmark of these diseases.

TNF-αR-Fc fusion protein (etanercept) and IL-1ra act to inhibit the activity of pathogenic cytokines in rheumatoid arthritis, i.e., TNF-α and IL-1, respectively. Similarly, anti-IL-6, IFN-β, and IL-11 act to inhibit pathogenic proinflammatory cytokines. Anti-IL-6 (tocilizumab) inhibits IL-6 activity, whereas IFN-β and IL-11 decrease IL-1 and TNF-α production.

Of particular note has been the successful use of IFN-γ in the treatment of the phagocytic cell defect in *chronic granulomatous disease* **(Chap. 60)**.

Th17 CD4 T cells have been implicated in the pathogenesis of psoriasis, ulcerative colitis, and other autoimmune diseases. Monoclonal antibodies have now been developed that target cytokines (IL-12, IL-23) that induce Th17 T cell differentiation, and are licensed by the FDA for treatment of psoriasis. Monoclonal antibodies that directly target IL-17 have also recently been licensed for psoriasis and psoriatic arthritis treatment.

Monoclonal Antibodies to T and B Cells

The OKT3 MAb against human T cells has been used for several years as a T cell–specific immunosuppressive agent that can substitute for horse antithymocyte globulin (ATG) in the treatment of solid organ transplant rejection. OKT3 produces fewer allergic reactions than ATG but does induce human anti-mouse Ig antibody—thus limiting its use. Anti-CD4 MAb therapy has been used in trials to treat patients with rheumatoid arthritis. While inducing profound immunosuppression, anti-CD4 MAb treatment also induces susceptibility to severe infections. Treatment of patients with a MAb against the T cell molecule CD40 ligand (CD154) is under investigation to induce tolerance to organ transplants, with promising results reported in animal studies. Monoclonal antibodies to the CD25 (IL-2a) receptor (basiliximab) are being used for treatment of graft-versus-host disease in bone marrow transplantation, and anti-CD20 MAb (rituximab) is used to treat hematologic neoplasms, autoimmune diseases, kidney transplant rejection, and rheumatoid arthritis. The anti-IgE monoclonal antibody (omalizumab) is used for blocking antigen-specific IgE that causes *hay fever* and *allergic rhinitis* **(Chap. 345)**; however, side effects of anti-IgE include increased risk of anaphylaxis. Studies have shown that T_H17 cells, in addition to T_H1, are mediators of inflammation in Crohn's disease, and anti–IL-12/IL-23p40 antibody therapy has been studied as a treatment.

It is important to realize the potential risks of these immunosuppressive monoclonal antibodies. Natalizumab is a humanized IgG antibody against an a4 integrin that inhibits leukocyte migration into tissues and has been approved for treatment of multiple sclerosis in the United States. Both it and anti-CD20 (rituximab) have been associated with the onset of progressive multifocal leukoencephalopathy (PML)—a serious and usually fatal CNS infection caused by JC polyomavirus. Efalizumab, a humanized IgG monoclonal antibody previously approved for treatment of plaque psoriasis, has now been taken off the market due to reactivation of JC virus leading to fatal PML. Thus, use of any currently approved immunosuppressant immunotherapies should be undertaken with caution and with careful monitoring of patients according to FDA guidelines.

Intravenous Immunoglobulin (IVIg)

IVIg has been used successfully to block reticuloendothelial cell function and immune complex clearance in various immune cytopenias such as immune thrombocytopenia **(Chap. 111)**. In addition, IVIg is useful for prevention of tissue damage in certain inflammatory syndromes such as Kawasaki disease **(Chap. 356)** and as Ig replacement therapy for certain types of immunoglobulin deficiencies **(Chap. 344)**. In addition, controlled clinical trials support the use of IVIg in selected patients with graft-versus-host disease, multiple sclerosis, myasthenia gravis, Guillain-Barré syndrome, and chronic demyelinating polyneuropathy.

Stem Cell Transplantation

Hematopoietic stem cell transplantation (SCT) is now being comprehensively studied to treat several autoimmune diseases including systemic lupus erythematosus, multiple sclerosis, and scleroderma. The goal of immune reconstitution in autoimmune disease syndromes is to replace a dysfunctional immune system with a normally reactive immune cell repertoire. Preliminary results in patients with scleroderma and lupus have showed encouraging results. Controlled clinical trials in these three diseases are now being launched in the United States and Europe to compare the toxicity and efficacy of conventional immunosuppression therapy with that of myeloablative autologous SCT. Recently, SCT was used in the setting of HIV-1 infection. HIV-1 infection of CD4+ T cells requires the presence of surface CD4 receptor and the chemokine receptor 5 (CCR5) co-receptor. Studies have demonstrated that patients who are homozygous for a 32-bp deletion in the CCR5 allele do not express CD4+ T cell CCR5 and thus are resistant to HIV-1 infection with HIV-1 strains that use this co-receptor. Stem cells from a homozygous CCR5 delta32 donor were transplanted to an HIV-infected patient following standard conditioning for such transplants, and the patient has maintained long-term control of the virus without antiretrovirals. Thus, a number of recent insights into immune system function have spawned a new field of interventional immunotherapy and have enhanced the prospect for development of more specific and nontoxic therapies for immune and inflammatory diseases.

FURTHER READING

Crotty S: T follicular helper cell differentiation, function, and roles in disease. Immunity 41:529, 2014. [PubMed: 25367570]

Dupage M, Bluestone JA: Harnessing the plasticity of CD4+T cells to treat immune-mediated disease. Nat Rev Immunol 3:149, 2016.

Kim BS et al: Targeting IL-17 in autoimmunity and inflammation. Arch Pharm Res 39:1537, 2016. [PubMed: 27576555]

Levine B et al: Autophagy in immunity and inflammation. Nature 469:323, 2011. [PubMed: 21248839]

Mulay SR, Anders HJ: Crystallopathies. N Engl J Med 374:25, 2016.

Pedersen HK et al: Human gut microbes impact host serum metabolome and insulin sensitivity. Nature 535:376, 2016. [PubMed: 27409811]

Pfeiffer JK, Virgin HW: Transkingdom control of viral infection and immunity in the mammalian intestine. Science 351:239, 2016.

Ratner D et al: Bacterial secretion systems and regulation of inflammasome activation. J Leuk Bio 101(1):165, 2017.

Yang F et al: The diverse biological functions of neutrophils, beyond the defense against infections. Inflammation 40:311, 2017. [PubMed: 27817110]

Chapter 343: The Major Histocompatibility Complex

Gerald T. Nepom

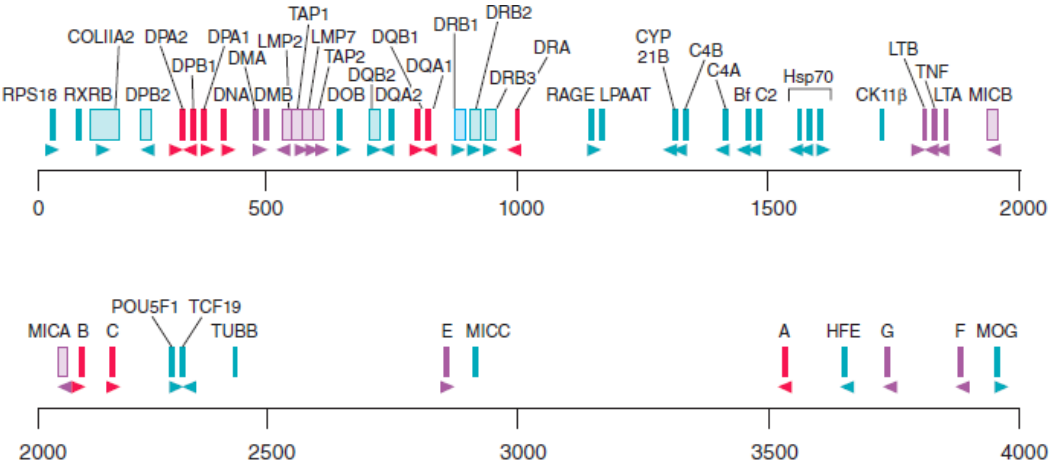
THE HLA COMPLEX AND ITS PRODUCTS

The human major histocompatibility complex (MHC), commonly called the human leukocyte antigen (HLA) complex, is a 4-megabase (Mb) region on chromosome 6 (6p21.3) that is densely packed with expressed genes. The best known of these genes are the *HLA class I* and *class II genes*, whose products are critical for immunologic specificity and transplantation histocompatibility, and they play a major role in susceptibility to a number of autoimmune diseases and some forms of drug hypersensitivity. Many other genes in the HLA region are also essential to the innate and antigen-specific functioning of the immune system. The HLA region shows extensive conservation with the MHC of other mammals in terms of genomic organization, gene sequence, and protein structure and function.

The *HLA class I genes* are located in a 2-Mb stretch of DNA at the telomeric end of the HLA region (**Fig. 343-1**). The classic (MHC class Ia) HLA-A, B, and C loci, the products of which are integral participants in the immune response to intracellular infections, tumors, and allografts, are expressed in all nucleated cells and are highly polymorphic in the population. *Polymorphism* refers to a high degree of allelic variation within a genetic locus that leads to extensive variation between different individuals expressing different alleles. More than 3400 alleles at HLA-A, 4300 alleles at HLA-B, and 3100 at HLA-C have been identified in different human populations, making this the most highly polymorphic segment known within the human genome. Each of the alleles at these loci encodes a *heavy chain* (also called an *α chain*) that associates noncovalently with the nonpolymorphic light chain *β_2-microglobulin*, encoded on chromosome 15.

FIGURE 343-1

Physical map of the HLA region, showing the class I and class II loci, other immunologically important loci, and a sampling of other genes mapped to this region. Gene orientation is indicated by *arrowheads*. Scale is in kilobase (kb). The approximate genetic distance from DP to A is 3.2 cM. This includes 0.8 cM between A and B (including 0.2 cM between C and B), 0.4–0.8 cM between B and DR-DQ, and 1.6–2.0 cM between DR-DQ and DP.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

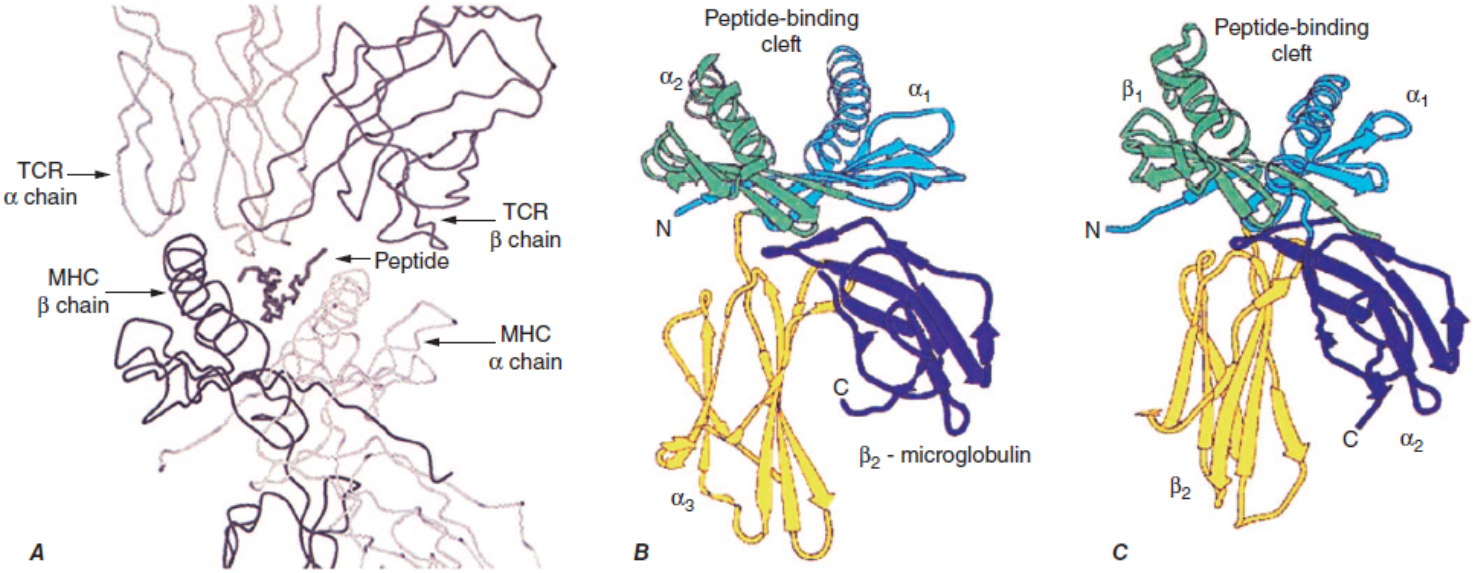
The designation of HLA genes and their products is based on a World Health Organization (WHO) nomenclature, in which alleles are given a single designation that indicates locus, allotype, and sequence-based subtype. For example, *HLA-A*02:01* indicates subtype 1 of a group of alleles that encode HLA-A2 molecules. Subtypes that differ from each other at the nucleotide but not the amino acid sequence level are designated by an extra numeral (e.g., *HLA-B*07:02:01* and *HLA-B*07:02:02* are two variants of *HLA-B*07:02*, both encoding the same HLA-B7 molecule). The nomenclature of class II genes, discussed below, is made more complicated by the fact that both chains of a class II molecule are encoded by closely linked HLA-encoded loci, each of which may be polymorphic, and by the presence of differing numbers of isotypic DRB loci in different individuals. It has become clear that accurate HLA genotyping requires DNA sequence analysis, and the identification of alleles at the

DNA sequence level has contributed greatly to the understanding of the role of HLA molecules as peptide-binding ligands, to the analysis of associations of HLA alleles with certain diseases, to the study of the population genetics of HLA, and to a clearer understanding of the contribution of HLA differences to allograft rejection and graft-versus-host disease. Current databases of HLA class I and class II sequences can be accessed by the Internet (e.g., from the IMGT/HLA Database, *http://www.ebi.ac.uk/imgt/hla*), and frequent updates of HLA gene lists are published in several journals. It is also possible to predict HLA genotypes by virtue of their linkage with single nucleotide polymorphisms (SNPs) prevalent in the genome. Imputation of HLA alleles using this technique is not as precise as targeted sequencing; however, the technology is much simpler and cheaper.

The biologic significance of this MHC genetic diversity, resulting in extreme variation in the human population, is evident from the perspective of the structure of MHC molecules. As shown in **Fig. 343-2**, the MHC class I and class II genes encode MHC molecules that bind small peptides, and together this complex (pMHC; peptide-MHC) forms the ligand for recognition by T lymphocytes, through the antigen-specific T cell receptor (TCR). There is a direct link between the genetic variation and this structural interaction: The allelic changes in genetic sequence result in diversification of the peptide-binding capabilities of each MHC molecule and in differences for specific TCR binding. Thus, different pMHC complexes bind different antigens and are targets for recognition by different T cells.

FIGURE 343-2

A. The trimolecular complex of TCR (*top*), MHC molecule (*bottom*), and a bound peptide form the structural determinants of specific antigen recognition. Other panels (***B*** and ***C***) show the domain structure of MHC class I (***B***) and class II (***C***) molecules. The α_1 and α_2 domains of class I and the α_1 and β_1 domains of class II form a β-sheet platform that forms the floor of the peptide-binding groove, and α helices that form the sides of the groove. The α_3 (***B***) and β_2 domains (***C***) project from the cell surface and form the contact sites for CD8 and CD4, respectively. *(Adapted from EL Reinherz et al: Science 286:1913, 1999; and C Janeway et al: Immunobiology Bookshelf, 2nd ed. Garland Publishing, New York, 1997; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The class I MHC and class II MHC structures, shown in Fig. 343-2B,C, are structurally closely related; however, there are a few key differences. While both bind peptides and present them to T cells, the binding pockets have different shapes, which influence the types of immune responses that result (discussed below). In addition, there are structural contact sites for T cell molecules known as CD8 and CD4, expressed on the class I or class II membrane-proximal domains, respectively. This ensures that when peptide antigens are presented by class I molecules, the responding T cells are predominantly of the CD8 class, and similarly, that T cells responding to class II pMHC complexes are predominantly CD4.

The nonclassic, or class Ib, MHC molecules, HLA-E, F, and G, are much less polymorphic than MHC Ia and appear to have distinct functions. The HLA-E molecule has a peptide repertoire displaying signal peptides cleaved from classic MHC class I molecules and is the major self-recognition target for the natural killer (NK) cell–inhibitory receptors NKG2A or NKG2C paired with CD94 (see below and **Chap. 342**). This appears to be a function of immune surveillance, because loss of MHC class I signal peptides serves as a surrogate marker for injured or infected cells, leading to release of the inhibitory signal and subsequent activation of NK cells. HLA-E can also bind and present peptides to CD8 T cells, albeit with a limited scope, as eight allelic HLA-E molecules are known. HLA-G was originally described in stem cells and in extravillous trophoblasts, where it

is implicated in regulation of maternal-fetal tolerance in pregnancy. It is now recognized as a widely expressed regulatory molecule that is expressed in multiple alternatively spliced forms, and provides inhibitory signals in both cell-bound and soluble forms; induction of expression is associated with downregulatory immunomodulation at sites of inflammation or malignancy. Eighteen allelic HLA-G molecules have been identified, interacting with receptors on NK, T cell, and dendritic cells. HLA-F occurs in four allelic forms, and is expressed on proliferating lymphoid and monocyte cells; its function is largely unknown, although it has been shown to form complexes that interact with specific NK receptors, sometimes together with other class I molecules in the absence of bound peptides. In general, the emerging view of non-classical class Ib molecules is a complex regulatory network for engaging immunomodulatory responses in the absence of traditional forms of antigen recognition attributed to classical class Ia molecules.

Additional class I–like genes have been identified, some HLA-linked and some encoded on other chromosomes, that show only distant homology to the class Ia and Ib molecules but share the three-dimensional class I structure. Those on chromosome 6p21 include MIC-A and MIC-B, which are encoded centromeric to HLA-B, and HLA-HFE, located 3 to 4 cM (centi-Morgan) telomeric of HLA-F. MIC-A and MIC-B do not bind peptide but are expressed on gut and other epithelium in a stress-inducible manner and serve as activation signals for certain γδ T cells, NK cells, CD8 T cells, and activated macrophages, acting through the activating NKG2D receptors. Over 100 MIC-A and 40 MIC-B alleles are known, and additional diversification comes from variable alanine repeat sequences in the transmembrane domain. Due to this structural diversity, MIC-A can be recognized as a foreign tissue target during organ transplantation, contributing to graft failure. HLA-HFE encodes the gene defective in hereditary hemochromatosis **(Chap. 407)**. Among the non-HLA, class I–like genes, CD1 refers to a family of molecules that present glycolipids or other nonpeptide ligands to certain T cells, including T cells with NK activity; FcRn binds IgG within lysosomes and protects it from catabolism **(Chap. 342)**; and Zn-α_2-glycoprotein 1 binds a nonpeptide ligand and promotes catabolism of triglycerides in adipose tissue. Like the HLA-A, B, C, E, and G heavy chains, each of which forms a heterodimer with β2-microglobulin (Fig. 343-2), the class I–like molecules, HLA-HFE, FcRn, and CD1 also bind to β_2-microglobulin, but MIC-A, MIC-B, and Zn-α_2-glycoprotein 1 do not.

The HLA class II region is also illustrated in Fig. 343-1. Multiple class II genes are arrayed within the centromeric 1 Mb of the HLA region, forming distinct haplotypes. A *haplotype* refers to an array of alleles at polymorphic loci along a chromosomal segment. Multiple class II genes are present on a single haplotype, clustered into three major subregions: HLA-DR, DQ, and DP. Each of these subregions contains at least one functional alpha (A) locus and one functional beta (B) locus. Together these encode proteins that form the α and β polypeptide chains of a mature class II HLA molecule. Thus, the DRA and DRB genes encode an HLA-DR molecule; DQA and DQB genes encode HLA-DQ molecules; and DPA and DPB genes encode HLA-DP molecules. There are several DRB genes (*DRB1*, *DRB2*, *DRB3*, etc.), so that two expressed DR molecules are encoded on most haplotypes by combining the α-chain product of the DRA gene with separate β chains. Nearly 2000 alleles have been identified at the HLA-DRB1 locus, with most of the variation occurring within limited segments encoding residues that interact with antigens. Detailed analysis of sequences and population distribution of these alleles strongly suggest that this diversity is actively selected by environmental pressures associated with pathogen diversity. In the DQ region, both DQA1 and DQB1 are polymorphic, with over 70 DQA1 alleles and 900 DQB1 alleles. The current nomenclature is largely analogous to that discussed above for class I, using the convention "locus * allele."

In addition to allelic polymorphism, products of different DQA alleles can, with some limitations, pair with products of different DQB alleles through both *cis* and *trans* pairing to create combinatorial complexity and expand the number of expressed class II molecules. Because of the enormous allelic diversity in the general population, most individuals are heterozygous at all of the class I and class II loci. Thus, most individuals express six classic class I molecules (two each of HLA-A, -B, and -C) and many class II molecules—two DP, two to four DR, and multiple DQ (both *cis* and *trans* dimers).

OTHER GENES IN THE MHC

In addition to the class I and class II genes themselves, there are numerous genes interspersed among the HLA loci that have interesting and important immunologic functions. Our current concept of the function of MHC genes now encompasses many of these additional genes, some of which are also highly polymorphic. Indeed, direct comparison of the complete DNA sequences for eight of the entire 4-Mb MHC regions from different haplotypes shows >44,000 nucleotide variations, encoding an extremely high potential for biologic diversity, and at least 97 genes located in this region are known to have coding region sequence variation. Specific examples include the TAP and LMP genes, as discussed in more detail below, which encode molecules that participate in intermediate steps in the HLA class I biosynthetic pathway. Another set of HLA genes, DMA and DMB, performs an analogous function for the class II pathway. These genes encode an intracellular molecule that facilitates the proper complexing of HLA class II molecules with antigen (see below). The *HLA class III region* is a name given to a cluster of genes between the class I and class II complexes, which includes genes for the two closely related cytokines tumor necrosis factor (TNF)-α and lymphotoxin (TNF-β); the complement components C2, C4, and Bf; heat shock protein (HSP) 70; and the enzyme 21-hydroxylase.

The class I genes HLA-A, B, and C are expressed in all nucleated cells, although generally to a higher degree on leukocytes than on nonleukocytes. In contrast, the class II genes show a more restricted distribution: HLA-DR and HLA-DP genes are constitutively expressed on most cells of the myeloid cell lineage, whereas all three class II gene families (HLA-DR, -DQ, and -DP) are inducible by certain stimuli provided by inflammatory cytokines such as interferon γ. Within the lymphoid lineage, expression of these class II genes is constitutive on B cells and inducible on human T cells. Most endothelial and epithelial cells in the body, including the vascular endothelium and the intestinal epithelium, are also inducible for class II gene expression, and some cells show specialized expression, such as HLA-DQA2 and HLA-DQB2 on Langerhans cells. While somatic tissues normally express only class I and not class II genes, during times of local inflammation, they are recruited by cytokine stimuli to express class II genes as well, thereby becoming active participants in ongoing immune responses. Class II expression is controlled largely at the transcriptional level through a conserved set of promoter elements that interact with a protein known as *CIITA*. Cytokine-mediated induction of CIITA is a principal method by which tissue-specific expression of HLA gene expression is controlled. Other HLA genes involved in the immune response, such as TAP and LMP, are also susceptible to upregulation by signals such as interferon γ.

LINKAGE DISEQUILIBRIUM

In addition to extensive polymorphism at the class I and class II loci, another characteristic feature of the HLA complex is *linkage disequilibrium*. This is formally defined as a deviation from Hardy-Weinberg equilibrium for alleles at linked loci. This is reflected in the very low recombination rates between certain loci within the HLA complex. For example, recombination between DR and DQ loci is almost never observed in family studies, and characteristic haplotypes with particular arrays of DR and DQ alleles are found in every population. Similarly, the complement components C2, C4, and Bf are almost invariably inherited together, and the alleles at these loci are found in characteristic haplotypes. In contrast, there is a recombinational hotspot between DQ and DP, which are separated by 1–2 cM of genetic distance, despite their close physical proximity. Certain extended haplotypes encompassing the interval from DQ into the class I region are commonly found, the most notable being the haplotype DR3-B8-A1, which is found, in whole or in part, in 10–30% of northern European whites. As discussed below under HLA and immunologic disease, one consequence of the phenomenon of linkage disequilibrium has been the resulting difficulty in assigning HLA-disease associations to a single allele at a single locus.

MHC STRUCTURE AND FUNCTION

Class I and class II molecules display a distinctive structural architecture, which contains specialized functional domains responsible for the unique genetic and immunologic properties of the HLA complex. The principal known function of both class I and class II HLA molecules is to bind antigenic peptides in order to present antigen to an appropriate T cell. The ability of a particular peptide to satisfactorily bind to an individual HLA molecule is a direct function of the molecular fit between the amino acid residues on the peptide with respect to the amino acid residues of the HLA molecule. The bound peptide forms a tertiary structure called the *MHC-peptide complex*, which communicates with T lymphocytes through binding to the TCR molecule. The first site of TCR-MHC-peptide interaction in the life of a T cell occurs in the thymus, where self-peptides are presented to developing thymocytes by MHC molecules expressed on thymic epithelium and hematopoietically derived antigen-presenting cells, which are primarily responsible for positive and negative selection, respectively **(Chap. 342)**. Thus, the population of MHC–T cell complexes expressed in the thymus shapes the TCR repertoire. Mature T cells encounter MHC molecules in the periphery both in the maintenance of tolerance **(Chap. 348)** and in the initiation of immune responses. The TCR-MHC-peptide interaction is the central event in the initiation of most antigen-specific immune responses, since it is the structural determinant of the specificity. For potentially immunogenic peptides, the ability of a given peptide to be generated and bound by an HLA molecule is a primary feature of whether or not an immune response to that peptide can be generated, and the repertoire of peptides that a particular individual's HLA molecules can bind exerts a major influence over the specificity of that individual's immune response.

When a TCR molecule binds to an HLA-peptide complex, it forms intermolecular contacts with both the antigenic peptide and with the HLA molecule itself. The outcome of this recognition event depends on the density and duration of the binding interaction, accounting for a dual specificity requirement for activation of the T cell. That is, the TCR must be specific both for the antigenic peptide and for the HLA molecule. The polymorphic nature of the presenting molecules, and the influence that this exerts on the peptide repertoire of each molecule, results in the phenomenon of *MHC restriction* of the T cell specificity for a given peptide. The binding of CD8 or CD4 molecules to the class I or class II molecule, respectively, also contributes to the interaction between T cell and the HLA-peptide complex, by providing for the selective activation of the appropriate T cell.

CLASS I STRUCTURE

(Fig. 343-2B) As noted above, MHC class I molecules provide a cell-surface display of peptides derived from intracellular proteins, and they also provide the signal for self-recognition by NK cells. Surface-expressed class I molecules consist of an MHC-encoded 44-kD glycoprotein heavy

chain, a non-MHC-encoded 12-kD light chain β_2-microglobulin, and an antigenic peptide, typically 8–11 amino acids in length and derived from intracellularly produced protein. The heavy chain displays a prominent peptide-binding groove. In HLA-A and B molecules, the groove is ~3 nm in length by 1.2 nm in maximum width (30 Å × 12 Å), whereas it is apparently somewhat wider in HLA-C. Antigenic peptides are noncovalently bound in an extended conformation within the peptide-binding groove, with both N- and C-terminal ends anchored in pockets within the groove (A and F pockets, respectively) and, in many cases, with a prominent kink, or arch, approximately one-third of the way from the N-terminus that elevates the peptide main chain off the floor of the groove.

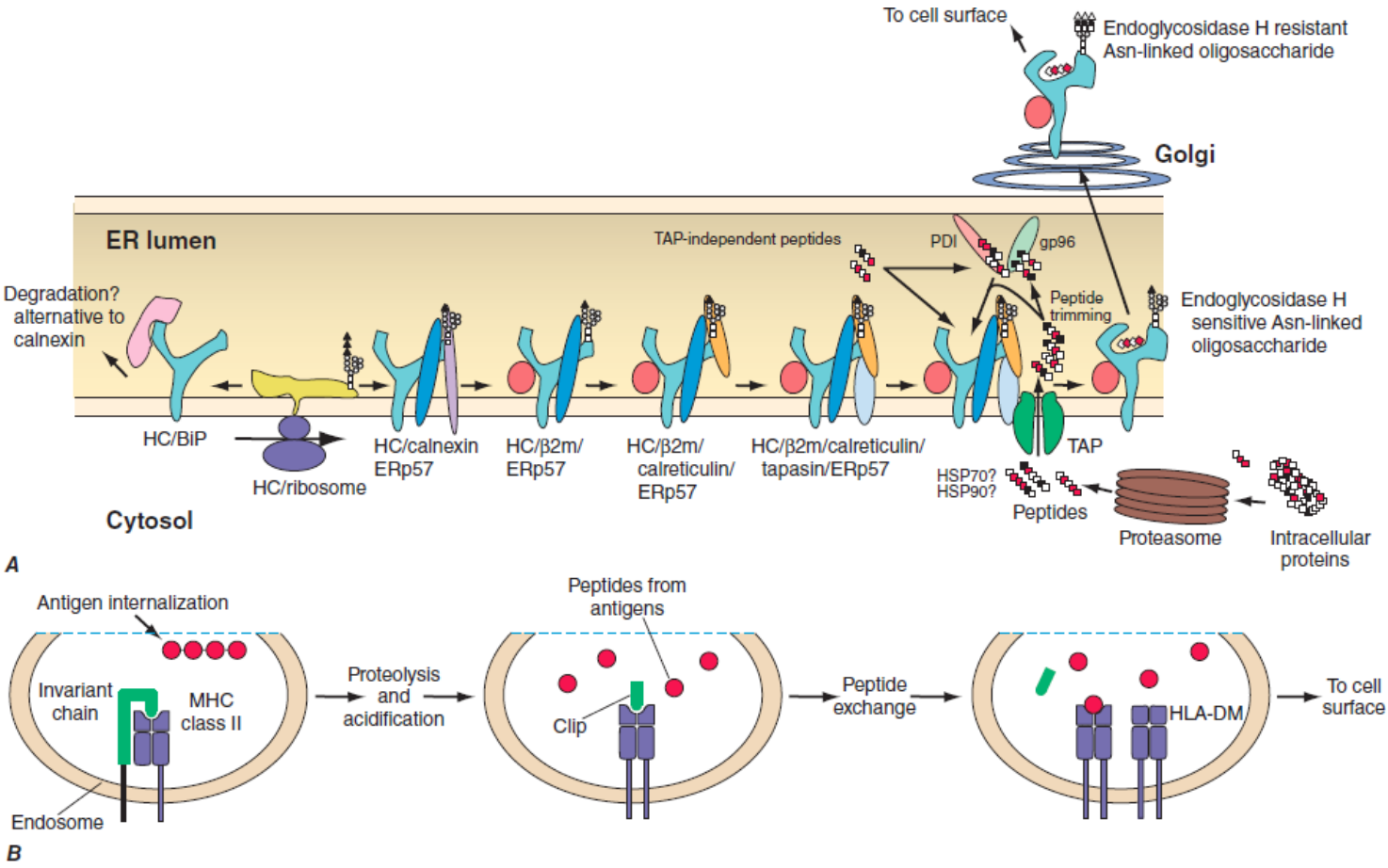
A remarkable property of peptide binding by MHC molecules is the ability to form highly stable complexes with a wide array of peptide sequences. This is accomplished by a combination of peptide sequence–independent and peptide sequence–dependent bonding. The former consists of hydrogen bond and van der Waals interactions between conserved residues in the peptide-binding groove and charged or polar atoms along the peptide backbone. The latter is dependent upon the six side pockets that are formed by the irregular surface produced by protrusion of amino acid side chains from within the binding groove. The side chains lining the pockets interact with some of the peptide side chains. The sequence polymorphism among different class I alleles and isotypes predominantly affects the residues that line these pockets, and the interactions of these residues with peptide residues constitute the sequence-dependent bonding that confers a particular sequence “motif” on the range of peptides that can bind each MHC molecule.

CLASS I BIOSYNTHESIS

(Fig. 343-3A) The biosynthesis of the classic MHC class I molecules reflects their role in presenting endogenous peptides. The heavy chain is cotranslationally inserted into the membrane of the endoplasmic reticulum (ER), where it becomes glycosylated and associates sequentially with the chaperone proteins calnexin and ERp57. It then forms a complex with β_2-microglobulin, and this complex associates with the chaperone calreticulin and the MHC-encoded molecule tapasin, which physically links the class I complex to TAP, the MHC-encoded transporter associated with antigen processing. Meanwhile, peptides generated within the cytosol from intracellular proteins by the multisubunit, multicatalytic proteasome complex are actively transported into the ER by TAP, where they are trimmed by enzymes known as ER aminopeptidases. At this point, peptides with appropriate sequence complementarity bind specific class I molecules to form complete, folded heavy chain–β_2-microglobulin–peptide trimer complexes. These are transported rapidly from the ER, through the *cis*- and *trans*-Golgi where the N-linked oligosaccharide is further processed, and thence to the cell surface.

FIGURE 343-3

Biosynthesis of class I (*A*) and class II (*B*) molecules. *A.* Nascent heavy chain (HC) becomes associated with β_2-microglobulin (β2m) and peptide through interactions with a series of chaperones. Peptides generated by the proteasome are transported into the endoplasmic reticulum (ER) by TAP. Peptides undergo N-terminal trimming in the ER and become associated with chaperones, including gp96 and PDI. Once peptide binds to HC-β2m, the HC-β2m-peptide trimeric complex exits the ER and is transported by the secretory pathway to the cell surface. In the Golgi, the N-linked oligosaccharide undergoes maturation, with addition of sialic acid residues. Molecules are not necessarily drawn to scale. ***B.*** Pathway of HLA class II molecule assembly and antigen processing. After transport through the Golgi and post-Golgi compartment, the class II–invariant chain complex moves to an acidic endosome, where the invariant chain is proteolytically cleaved into fragments and displaced by antigenic peptides, facilitated by interactions with the DMA-DMB chaperone protein. This class II molecule–peptide complex is then transported to the cell surface.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Most of the peptides transported by TAP are produced in the cytosol by proteolytic cleavage of intracellular proteins by the multisubunit, multicatalytic proteasome, and inhibitors of the proteasome dramatically reduce expression of class I–presented antigenic peptides. A thiol-dependent oxidoreductase ERp57, which mediates disulfide bond rearrangements, also appears to play an important role in folding the class I–peptide complex into a stable multicomponent molecule. The MHC-encoded proteasome subunits LMP2 and LMP7 may influence the spectrum of peptides produced but are not essential for proteasome function.

CLASS I FUNCTION

Peptide Antigen Presentation

On any given cell, a class I molecule occurs in 100,000–200,000 copies and binds several hundred to several thousand distinct peptide species. The vast majority of these peptides are self-peptides to which the host immune system is tolerant by one or more of the mechanisms that maintain tolerance (e.g., clonal deletion in the thymus or clonal anergy or clonal ignorance in the periphery [Chaps. 342 and 348]). However, class I molecules bearing foreign peptides expressed in a permissive immunologic context activate CD8 T cells, which, if naïve, will then differentiate into cytolytic T lymphocytes (CTLs). These T cells and their progeny, through their αβ TCRs, are then capable of Fas/CD95- and/or perforin-mediated cytotoxicity and/or cytokine secretion (Chap. 342) upon further encounter with the class I–peptide combination that originally activated it, or other structurally related class I–peptide complexes. As alluded to above, this phenomenon by which T cells recognize foreign antigens in the context of specific MHC alleles is termed *MHC restriction*, and the specific MHC molecule is termed the *restriction element*. The most common source of foreign peptides presented by class I molecules is viral infection, in the course of which peptides from viral proteins enter the class I pathway. The generation of a strong CTL response that destroys virally infected cells represents an important antigen-specific defense against many viral infections (Chap. 342). In the case of some viral infections—hepatitis B, for example—CTL-induced target cell apoptosis is thought to be a more important mechanism of tissue damage than any direct cytopathic effect of the virus itself. The importance of the class I pathway in the defense against viral infection is underscored by the identification of a number of viral products that interfere with the normal class I biosynthetic pathway and thus block the immunogenic expression of viral antigens.

Other examples of intracellularly generated peptides that can be presented by class I molecules in an immunogenic manner include peptides derived from nonviral intracellular infectious agents (e.g., *Listeria*, *Plasmodium*), tumor antigens, minor histocompatibility antigens, and certain

autoantigens. There are also situations in which cell surface–expressed class I molecules are thought to acquire and present exogenously derived peptides.

HLA Class I Receptors and NK Cell Recognition

(Chap. 342) NK cells, which play an important role in innate immune responses, are activated to cytotoxicity and cytokine secretion by contact with cells that lack MHC class I expression, and NK cell activation is inhibited by cells that express MHC class I. In humans, the recognition of class I molecules by NK cells is carried out by three classes of receptor families, the killer cell–inhibitory cell receptor (KIR) family, the leukocyte Ig-like receptor (LIR) family, and the CD94/NKG2 family. The KIR family, also called CD158, is encoded on chromosome 19q13.4. KIR gene nomenclature is based on the number of domains (2D or 3D) and the presence of long (L) or short (S) cytoplasmic domains. The KIR2DL1 and S1 molecules primarily recognize alleles of HLA-C, which possess a lysine at position 80 (HLA-Cw2, -4, -5, and -6), whereas the KIR2DL2/S2 and KIR2DL3/S3 families primarily recognize alleles of HLA-C with asparagine at this position (HLA-Cw1, -3, -7, and -8). The KIR3DL1 and S1 molecules predominantly recognize HLA-B alleles that fall into the HLA-Bw4 class determined by residues 77–83 in the α_1 domain of the heavy chain, whereas the KIR3DL2 molecule is an inhibitory receptor for HLA-A*03. One of the KIR products, KIR2DL4, is known to be an activating receptor for HLA-G, and KIR3DL2 and KIR2DS4 have been described as immunoregulatory ligands interacting with HLA-F. The most common KIR haplotype in whites contains one activating KIR and six inhibitory KIR genes, although there is a great deal of diversity in the population, with >100 different combinations. It appears that most individuals have at least one inhibitory KIR for a self-HLA class I molecule, providing a structural basis for NK cell target specificity, which helps prevent NK cells from attacking normal cells. The importance of KIR-HLA interactions to many immune responses is illustrated by studies associating KIR3DL1 or S1 with multiple sclerosis **(Chap. 436)**, an autoimmune disease, but also with partial protection against HIV **(Chap. 197)**, in both cases consistent with a role for HLA-KIR–mediated NK activation. Studies also show an association of KIR2DS1 with protection from relapse following allogeneic bone marrow transplantation in acute myeloid leukemia when these inhibitory receptors in the donors do not recognize the recipient HLA-C.

The LIR gene family (CD85, also called ILT) is encoded centromeric of the KIR locus on 19q13.4, and it encodes a variety of inhibitory immunoglobulin-like receptors expressed on many lymphocyte and other hematopoietic lineages. Interaction of LIR-1 (ILT2) with NK or T cells inhibits activation and cytotoxicity, mediated by many different HLA class I molecules, including HLA-G. HLA-F also appears to interact with LIR molecules, although the functional context for this is not understood.

The third family of NK receptors for HLA is encoded in the NK complex on chromosome 12p12.3-13.1 and consists of CD94 and five NKG2 genes, A/B, C, E/H, D, and F. These molecules are C-type (calcium-binding) lectins, and most of these function as disulfide-bonded heterodimers between CD94 and one of the NKG2 glycoproteins. The principal ligand of CD94/NKG2A receptors is the HLA-E molecule, complexed to a peptide derived from the signal sequence of classic HLA class I molecules and HLA-G. Thus, analogous to the way in which KIR receptors recognize HLA-C, the NKG2 receptor monitors self–class I expression, albeit indirectly through peptide recognition in the context of HLA-E. NKG2C, E, and H appear to have similar specificities but act as activating receptors. NKG2D is expressed as a homodimer and functions as an activating receptor expressed on NK cells, γδ TCR T cells, and activated CD8 T cells. When complexed with an adaptor called DAP10, NKG2D recognizes MIC-A and MIC-B molecules and activates the cytolytic response. NKG2D also binds a class of molecules known as *ULBP*, structurally related to class I molecules but not encoded in the MHC. **The function of NK cells in immune responses is discussed in Chap. 342.**

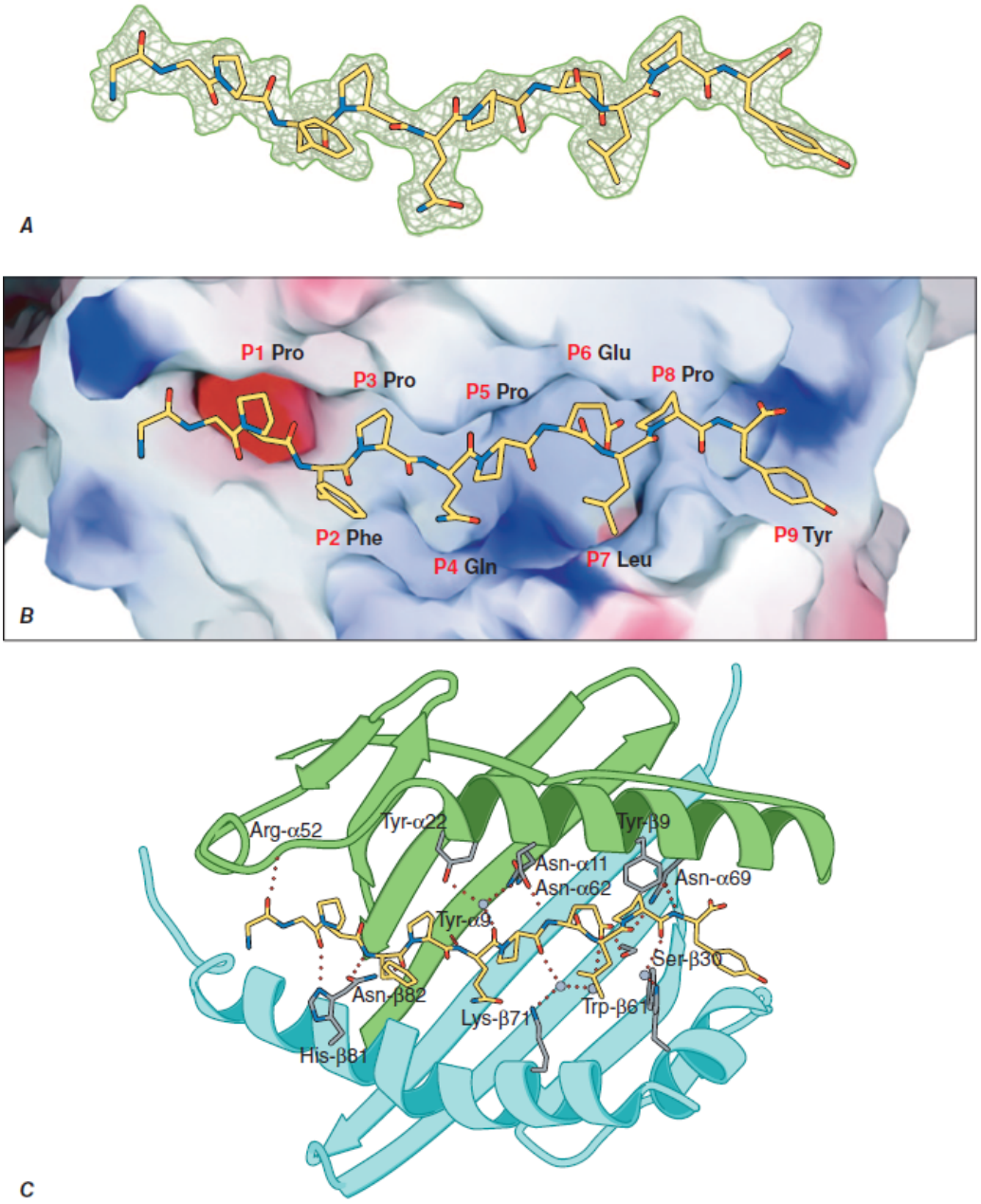
CLASS II STRUCTURE

(Fig. 343-2C) A specialized functional architecture similar to that of the class I molecules can be seen in the example of a class II molecule depicted in Fig. 343-2C, with an antigen-binding cleft arrayed above a supporting scaffold that extends the cleft toward the external cellular environment. However, in contrast to the HLA class I molecular structure, β_2-microglobulin is not associated with class II molecules. Rather, the class II molecule is a heterodimer, composed of a 29-kD α chain and a 34-kD β chain. The amino-terminal domains of each chain form the antigen-binding elements that, like the class I molecule, cradle a peptide in a groove bounded by extended α-helical loops, one encoded by the A (α chain) gene and one by the B (β chain) gene. Like the class I groove, the class II antigen-binding groove is punctuated by pockets that contact the side chains of amino acid residues of the bound peptide, but unlike the class I groove, it is open at both ends. Therefore, peptides bound by class II molecules vary greatly in length, since both the N- and C-terminal ends of the peptides can extend through the open ends of this groove. Approximately 11 amino acids within the bound peptide form intimate contacts with the class II molecule itself, with backbone hydrogen bonds and specific side chain interactions combining to provide, respectively, stability and specificity to the binding **(Fig. 343-4)**.

FIGURE 343-4

Specific intermolecular interactions determine peptide binding to MHC class II molecules. A short peptide sequence derived from alpha-gliadin (***A***) is accommodated within the MHC class II binding groove by specific interactions between peptide side chains (the P1–P9 residues illustrated in ***B***) and corresponding pockets in the MHC class II structure. The latter are determined by the genetic polymorphisms of the MHC gene, in this

case encoding an HLA-DQ2 molecule (***C***). This shows the extensive hydrogen bond and salt bridge network, which tightly constrains the pMHC complex and presents the complex of antigen and restriction element for CD4 T cell recognition. *(From C Kim et al: Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. Proc Natl Acad Sci USA 101:4175, 2004.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The genetic polymorphisms that distinguish different class II genes correspond to changes in the amino acid composition of the class II molecule, and these variable sites are clustered predominantly around the pocket structures within the antigen-binding groove. As with class I, this is a critically important feature of the class II molecule, which explains how genetically different individuals have functionally different HLA molecules.

BIOSYNTHESIS AND FUNCTION OF CLASS II MOLECULES

(Fig. 343-3B) The intracellular assembly of class II molecules occurs within a specialized compartmentalized pathway that differs dramatically from the class I pathway described above. As illustrated in Fig. 343-3B, the class II molecule assembles in the ER in association with a chaperone molecule, known as the *invariant chain*. The invariant chain performs at least two roles. First, it binds to the class II molecule and blocks the peptide-binding groove, thus preventing antigenic peptides from binding. This role of the invariant chain appears to account for one of the important differences between class I and class II MHC pathways, since it can explain why class I molecules present endogenous peptides from proteins newly synthesized in the ER but class II molecules generally do not. Second, the invariant chain contains molecular localization signals

that direct the class II molecule to traffic into post-Golgi compartments known as *endosomes*, which develop into specialized acidic compartments where proteases cleave the invariant chain, and antigenic peptides can now occupy the class II groove. The specificity and tissue distribution of these proteases appear to be an important way in which the immune system regulates access to the peptide-binding groove and T cells become exposed to specific self-antigens. Differences in protease expression in the thymus and in the periphery may in part determine which specific peptide sequences comprise the peripheral repertoire for T cell recognition. It is at this stage in the intracellular pathway, after cleavage of the invariant chain, that the MHC-encoded DM molecule catalytically facilitates the exchange of peptides within the class II groove to help optimize the specificity and stability of the MHC-peptide complex.

Once this MHC-peptide complex is deposited in the outer cell membrane, it becomes the target for T cell recognition via a specific TCR expressed on lymphocytes. Because the endosome environment contains internalized proteins retrieved from the extracellular environment, the class II–peptide complex often contains bound antigens that were originally derived from extracellular proteins. In this way, the class II peptide–loading pathway provides a mechanism for immune surveillance of the extracellular space. This appears to be an important feature that permits the class II molecule to bind foreign peptides, distinct from the endogenous pathway of class I–mediated presentation.

ROLE OF HLA IN TRANSPLANTATION

The development of modern clinical transplantation in the decades since the 1950s provided a major impetus for elucidation of the HLA system, as allograft survival is highest when donor and recipient are HLA-identical. Although many molecular events participate in transplantation rejection, allogeneic differences at class I and class II loci play a major role. Class I molecules can promote T cell responses in several different ways. In the cases of allografts in which the host and donor are mismatched at one or more class I loci, host T cells can be activated by classic *direct alloreactivity*, in which the antigen receptors on the host T cells react with the foreign class I molecule expressed on the allograft. In this situation, the response of any given TCR may be dominated by the allogeneic MHC molecule, the peptide bound to it, or some combination of the two. Another type of host anti-graft T cell response involves the uptake and processing of donor MHC antigens by host antigen-presenting cells and the subsequent presentation of the resulting peptides by host MHC molecules. This mechanism is termed *indirect alloreactivity*.

In the case of class I molecules on allografts that are shared by the host and the donor, a host T cell response may still be triggered because of peptides that are presented by the class I molecules of the graft but not of the host. The most common basis for the existence of these endogenous antigen peptides, called *minor histocompatibility antigens*, is a genetic difference between donor and host at a non-MHC locus encoding the structural gene for the protein from which the peptide is derived. These loci are termed *minor histocompatibility loci*, and nonidentical individuals typically differ at many such loci. CD4 T cells react to analogous class II variation, both direct and indirect, and class II differences alone are sufficient to drive allograft rejection.

ASSOCIATION OF HLA ALLELES WITH SUSCEPTIBILITY TO DISEASE

It has long been postulated that infectious agents provide the driving force for the allelic diversification seen in the HLA system. An important corollary of this hypothesis is that resistance to specific pathogens may differ between individuals, based on HLA genotype. Observations of specific HLA genes associated with resistance to malaria or dengue fever, persistence of hepatitis B, and to disease progression in HIV infection are consistent with this model. For example, failure to clear persistent hepatitis B or C viral infection may reflect the inability of particular HLA molecules to present viral antigens effectively to T cells. Similarly, both protective and susceptible HLA allelic associations have been described for human papilloma virus–associated cervical neoplasia, implicating the MHC as an influence in mediating viral clearance in this form of cancer.

Pathogen diversity is probably also the major selective pressure favoring HLA heterozygosity. The extraordinary scope of HLA allelic diversity increases the likelihood that most new pathogens will be recognized by some HLA molecules, helping to ensure immune fitness to the host. However, another consequence of diversification is that some alleles may become capable of recognition of “innocent bystander” molecules, including drugs, environmental molecules, and tissue-derived self-antigens. In a few instances, single HLA alleles display a strong selectivity for binding of a particular agent that accounts for a genetically determined response: Hypersensitivity to abacavir, an antiretroviral therapeutic, is directly linked to binding of abacavir in the antigen-binding pockets of *HLA-B*57:01*, where it is buried underneath antigenic peptides and distorts the landscape, changing T cell recognition specificity; an adverse drug reaction to abacavir is >500 times more likely to occur in persons with *HLA-B*57:01* than in individuals without this HLA allele. Other examples include chronic beryllium toxicity, which is linked to binding of beryllium by HLA-DP molecules with a specific glutamic acid polymorphic residue on the class II beta chain, clindamycin-related cutaneous drug reactions which are more common in individuals with *HLA-B*51:01*, and dapsone hypersensitivity in patients with leprosy who express *HLA-B*13:01*. Even in the case of more complex diseases, particular HLA alleles are strongly associated with certain inappropriate immune-mediated disease states, particularly for some common autoimmune disorders (**Chap. 348**). By comparing allele frequencies in patients with

any particular disease and in control populations, >100 such associations have been identified, some of which are listed in **Table 343-1**. The strength of genetic association is reflected in the term *relative risk*, which is a statistical odds ratio representing the risk of disease for an individual carrying a particular genetic marker compared with the risk for individuals in that population without that marker. The nomenclature shown in Table 343-1 reflects both the HLA serotype (e.g., DR3, DR4) and the HLA genotype (e.g., *DRB1** 03:01*, *DRB1** 04:01*). It is very likely that the class I and class II alleles themselves are the true susceptibility alleles for most of these associations. However, because of the extremely strong linkage disequilibrium between the DR and DQ loci, in some cases it has been difficult to determine the specific locus or combination of class II loci involved. In some cases, the susceptibility gene may be one of the HLA-linked genes located near the class I or class II region, but not the HLA gene itself, and in other cases, the susceptibility gene may be a non-HLA gene such as TNF-α, which is nearby. Indeed, since linkage disequilibrium of some haplotypes extends across large segments of the MHC region, it is quite possible that combinations of genes may account for the particular associations of HLA haplotypes with disease. For example, on some haplotypes associated with rheumatoid arthritis (RA), both HLA-DRB1 alleles and a particular polymorphism associated with the TNF locus may be contributory to disease risk. Other candidates for similar epistatic effects include the IKBL gene and the MICA locus, potentially in combination with classic HLA class II risk alleles.

TABLE 343-1

Significant HLA Class I and Class II Associations with Disease

	MARKER	GENE	STRENGTH OF ASSOCIATION
Spondyloarthropathies			
Ankylosing spondylitis	B27	*B*27:02,−04, −05*	++++
Reactive arthritis (Reiter's)	B27		++++
Acute anterior uveitis	B27		+++
Reactive arthritis (*Yersinia, Salmonella, Shigella, Chlamydia)*	B27		+++
Psoriatic spondylitis	B27		+++
Collagen-Vascular Diseases			
Juvenile arthritis, pauciarticular	DR8	*DRB1*04:01, −04, −05*	++
Rheumatoid arthritis	DR5		++
Sjögren's syndrome	DR4		+++
Systemic lupus erythematosus	DR3		++
White	DR3		+
Japanese	DR2		++
Autoimmune Gut and Skin			
Gluten-sensitive enteropathy (celiac disease)	DQ2	*DQA1*05:01*	+++
Chronic active hepatitis	DR3	*DQB1*02:01*	++
Dermatitis herpetiformis	DR3	*DRB1*04:02*	+++
Psoriasis vulgaris	Cw6	*DQB1*05:03*	++
Pemphigus vulgaris	DR4	*DQB1*03:01*	+++
Bullous pemphigoid variant	DQ1		+
	DQ7		
Autoimmune Endocrine			
Type 1 diabetes mellitus	DQ8	*DQB1*03:02*	+++
Hyperthyroidism (Graves')	DR4	*DRB1*04:01,−04*	++
Hyperthyroidism (Japanese)	DR3	*DQB1*06:02*	—
Adrenal insufficiency	DR2		+[a]
	B8		+
	DR3		+
	B35		++
	DR3		
Autoimmune Neurologic			
Myasthenia gravis	B8	*DRB1*15:01*	+
Multiple sclerosis	DR2	*DRB5*01:01*	+
	DR2		++
Other			

	MARKER	GENE	STRENGTH OF ASSOCIATION
Behçet's disease	B51	*21•OH (Cyp21B)*	++
Congenital adrenal hyperplasia	B47	*DQB1*06:02*	+++
Narcolepsy	DR2	*B*57:01*	++++
Goodpasture's syndrome (anti-GBM)	DR2		++
Abacavir hypersensitivity	B57		++++

[a]Strong negative association, that is, genetic association with protection from diabetes.

Abbreviation: GBM, glomerular basement membrane.

As might be predicted from the known function of the class I and class II gene products, almost all of the diseases associated with specific HLA alleles have an immunologic component to their pathogenesis. The recent development of soluble HLA-peptide recombinant molecules as biological probes of T cell function, often in multivalent complexes referred to as "MHC tetramers," represents an opportunity to use HLA genetic associations to develop biomarkers for detection of early disease progression. However, it should be stressed that even the strong HLA associations with disease (those associations with relative risk of ≥10) implicate normal, rather than defective, alleles. Most individuals who carry these susceptibility genes do not express the associated disease; in this way, the particular HLA gene is permissive for disease but requires other environmental (e.g., the presence of specific antigens) or genetic factors for full penetrance. In each case studied, even in diseases with very strong HLA associations, the concordance of disease in monozygotic twins is higher than in HLA-identical dizygotic twins or other sibling pairs, indicating that non-HLA genes contribute to susceptibility and can significantly modify the risk attributable to HLA.

Another group of diseases is genetically linked to HLA, not because of the immunologic function of HLA alleles but rather because they are caused by autosomal dominant or recessive abnormal alleles at loci that happen to reside in or near the HLA region. Examples of these are 21-hydroxylase deficiency (**Chap. 379**), hemochromatosis (**Chap. 407**), and spinocerebellar ataxia (**Chap. 429**).

CLASS I ASSOCIATIONS WITH DISEASE

Although the associations of human disease with particular HLA alleles or haplotypes predominantly involve the class II region, there are also several prominent disease associations with class I alleles. These include the association of Behçet's disease (**Chap. 357**) with HLA-B51, psoriasis vulgaris (**Chap. 53**) with HLA-Cw6, and, most notably, the spondyloarthritides (**Chap. 355**) with HLA-B27. More than 150 HLA-B locus alleles, designated *HLA-B*27:01–B*27:154*, encode the family of B27 class I molecules. All of the subtypes share a common B pocket in the peptide-binding groove—a deep, negatively charged pocket that shows a strong preference for binding the arginine side chain. In addition, B27 is among the most negatively charged of HLA class I heavy chains, and the overall preference is for positively charged peptides. *HLA-B*27:05* is the predominant subtype in whites and most other non-Asian populations, and this subtype is very highly associated with ankylosing spondylitis (AS) (**Chap. 355**), both in its idiopathic form and in association with chronic inflammatory bowel disease or psoriasis vulgaris. It is also associated with reactive arthritis (ReA) (**Chap. 355**), with other idiopathic forms of peripheral arthritis (undifferentiated spondyloarthropathy), and with recurrent acute anterior uveitis. B27 is found in 50–90% of individuals with these conditions, compared with a prevalence of ~7% in North American whites.

Evidence that the B27 molecule itself is involved in disease pathogenesis comes both from clinical epidemiology and on the occurrence of a spondyloarthropathy-like disease in HLA-B27 transgenic rats. The association of B27 with these diseases may derive from the specificity of a particular peptide or family of peptides bound to B27 or through another mechanism that is independent of the peptide specificity of B27. In particular, HLA-B27 has been shown to form heavy chain homodimers, utilizing the cysteine residue at position 67 of the B57 α chain, in the absence of β_2-microglobulin. These homodimers are expressed on the surface of lymphocytes and monocytes from patients with AS, and receptors including KIR3DL1, KIR3DL2, and ILT4 (LILRB2) are capable of binding to them, promoting the activation and survival of cells expressing these receptors. Alternatively, this dimerization "misfolding" of B27 may initiate an intracellular stress signaling response, called the unfolded protein response (UPR), capable of modulating immune cell function, possibly in enthesial-resident T cells that act as sensors of damage and environmental stress.

CLASS II DISEASE ASSOCIATIONS

As can be seen in Table 343-1, the majority of associations of HLA and disease are with class II alleles. Several diseases have complex HLA genetic associations.

Celiac Disease

In the case of celiac disease (**Chap. 318**), it is probable that the HLA-DQ genes are the primary basis for the disease association. HLA-DQ genes present on both the celiac-associated DR3 and DR7 haplotypes include the *DQB1** *02:01* gene, and further detailed studies have documented a specific class II αβ dimer encoded by the *DQA1** *05:01* and *DQB1** *02:01* genes, which appears to account for most of the HLA genetic contribution to celiac disease susceptibility. This specific HLA association with celiac disease may have a straightforward explanation: Peptides derived from the wheat gluten component gliaden are bound to the molecule encoded by *DQA1** *05:01* and *DQB1** *02:01* and presented to T cells. Gliaden-derived peptides that are implicated in this immune activation bind the DQ class II dimer best when the peptide contains a glutamine to glutamic acid substitution. It has been proposed that tissue transglutaminase, an enzyme present at increased levels in the intestinal cells of celiac patients, converts glutamine to glutamic acid in gliadin, creating peptides that are capable of being bound by the DQ2 molecule and presented to T cells.

Pemphigus Vulgaris

In the case of pemphigus vulgaris (**Chap. 55**), there are two HLA genes associated with disease, *DRB1** *04:02* and *DQB1** *05:03*. Peptides derived from desmoglein-3, an epidermal autoantigen, bind to the *DRB1** *04:02*- and *DQB1** *05:03*-encoded HLA molecules, and this combination of specific peptide binding and disease-associated class II molecule is sufficient to stimulate desmoglein-specific T cells. A bullous pemphigoid clinical variant, not involving desmoglein recognition, has been found to be associated with *HLA-DQB1** *03:01*.

Juvenile Arthritis

Pauciarticular juvenile arthritis (**Chap. 351**) is an autoimmune disease associated with genes at the DRB1 locus and also with genes at the DPB1 locus. Patients with both *DPB1** *02:01* and a DRB1 susceptibility allele (usually *DRB1** *08* or * *05*) have a higher relative risk than expected from the additive effect of those genes alone. In juvenile patients with rheumatoid factor–positive polyarticular disease, heterozygotes carrying both *DRB1** *04:01* and * *04:04* have a relative risk >100, reflecting an apparent synergy in individuals inheriting both of these susceptibility genes.

Type 1 Diabetes Mellitus

Type 1 (autoimmune) diabetes mellitus (**Chap. 396**) is associated with MHC genes on more than one haplotype. The presence of both the DR3 and DR4 haplotypes in one individual confers a twentyfold increased risk for type 1 diabetes; the strongest single association is with *DQB1** *03:02*, and all haplotypes that carry a *DQB1** *03:02* gene are associated with type 1 diabetes, whereas related haplotypes that carry a different DQB1 gene are not. However, the relative risk associated with inheritance of this gene can be modified, depending on other HLA genes present either on the same or a second haplotype. For example, the presence of a DR2-positive haplotype containing a *DQB1** *06:02* gene is associated with decreased risk. This gene, *DQB1** *06:02*, is considered “protective” for type 1 diabetes. Even some DRB1 genes that can occur on the same haplotype as *DQB1** *03:02* may modulate risk, so that individuals with the DR4 haplotype that contains *DRB1** *04:03* are less susceptible to type 1 diabetes than individuals with other DR4-DQB1*03:02 haplotypes. There are some characteristic structural features of the diabetes-associated DQ molecule encoded by *DQB1** *03:02*, particularly the capability for binding peptides that have negatively charged amino acids near their C-termini. This may indicate a role for specific antigenic peptides or T cell interactions in the immune response to islet-associated proteins.

Although the presence of a DR3 haplotype in combination with the DR4-DQB1*0302 haplotype is a very high-risk combination for diabetes susceptibility, the specific gene on the DR3 haplotype that is responsible for this synergy is not yet identified.

Rheumatoid Arthritis

The HLA genes associated with RA (**Chap. 351**) encode a distinctive sequence of amino acids from codons 67 to 74 of the DRβ molecule: RA-associated class II molecules carry the sequence LeuLeuGluGlnArgArgAlaAla or LeuLeuGluGlnLysArgAlaAla in this region, whereas non-RA-associated genes carry one or more differences in this region. These residues form a portion of the molecule that lies in the middle of the α-helical portion of the DRB1-encoded class II molecule, termed the *shared epitope*.

The highest risk for susceptibility to RA comes in individuals who carry both a *DRB1* 04:01* and *DRB1* 04:04* gene. These DR4-positive RA-associated alleles with the *shared epitope* are most frequent among patients with more severe, erosive disease. Several mechanisms have been proposed that link the shared epitope to immune reactivity in RA. This portion of the class II molecule may allow preferential binding of an arthritogenic peptide, it may favor the expansion of a type of self-reactive T lymphocyte, or it may itself form part of the pMHC ligand recognized by TCR that initiates synovial tissue recognition.

MOLECULAR MECHANISMS FOR HLA-DISEASE ASSOCIATIONS

As noted above, HLA molecules play a key role in the selection and establishment of the antigen-specific T cell repertoire and a major role in the subsequent activation of those T cells during the initiation of an immune response. Precise genetic polymorphisms characteristic of individual alleles dictate the specificity of these interactions and thereby instruct and guide antigen-specific immune events. These same genetically determined pathways are therefore implicated in disease pathogenesis when specific HLA genes are responsible for autoimmune disease susceptibility.

The fate of developing T cells within the thymus is determined by the affinity of interaction between TCR and HLA molecules bearing self-peptides, and thus the particular HLA types of each individual control the precise specificity of the T cell repertoire **(Chap. 342)**. The primary basis for HLA-associated disease susceptibility may well lie within this thymic maturation pathway. The positive selection of potentially autoreactive T cells, based on the presence of specific HLA susceptibility genes, may establish the threshold for disease risk in a particular individual.

At the time of onset of a subsequent immune response, the primary role of the HLA molecule is to bind peptide and present it to antigen-specific T cells. The HLA complex can therefore be viewed as encoding genetic determinants of precise immunologic activation events. Antigenic peptides that bind particular HLA molecules are capable of stimulating T cell immune responses; peptides that do not bind are not presented to T cells and are not immunogenic. This genetic control of the immune response is mediated by the polymorphic sites within the HLA antigen–binding groove that interact with the bound peptides. In autoimmune and immune-mediated diseases, it is likely that specific tissue antigens that are targets for pathogenic lymphocytes are complexed with the HLA molecules encoded by specific susceptibility alleles. In autoimmune diseases with an infectious etiology, it is likely that immune responses to peptides derived from the initiating pathogen are bound and presented by particular HLA molecules to activate T lymphocytes that play a triggering or contributory role in disease pathogenesis. The concept that early events in disease initiation are triggered by specific HLA-peptide complexes offers some prospects for therapeutic intervention, since it may be possible to design compounds that interfere with the formation or function of specific HLA-peptide–TCR interactions.

When considering mechanisms of HLA associations with immune response and disease, it is well to remember that immune-mediated disease is a multistep process in which initial HLA-peptide recognition helps establish a repertoire of potentially reactive T cells, whereas subsequent HLA-associated antigen presentation provides the essential peptide-binding specificity for peripheral T cell recognition leading to activation. These deterministic events can occur long before clinical evidence of autoimmunity, as exemplified by the HLA genetic associations with detection of specific autoantibodies in type 1 diabetes and in rheumatoid arthritis that are present for several years before disease diagnosis.

FURTHER READING

Karlin E, Phillips E: Genotyping for severe drug hypersensitivity. Curr Allergy Asthma Rep 14:418, 2014. [PubMed: 24429903]

Liu J et al: Phenome-wide association study maps new diseases to the human major histocompatibility complex region. J Med Genet 53:681, 2016. [PubMed: 27287392]

Robinson J et al: The IPD and IPD-IMGT/HLA Database: Allele variant databases. Nucleic Acids Res 43:D423, 2015. [PubMed: 25414341]

Zheng X et al: HIBAG—HLA genotype imputation with attribute bagging. Pharmacogenomics J 14:192, 2014. [PubMed: 23712092]

Chapter 344: Primary Immune Deficiency Diseases

Alain Fischer

INTRODUCTION

Immunity is intrinsic to life and an important tool in the fight for survival against pathogenic microorganisms. The human immune system can be divided into two major components: the innate immune system and the adaptive immune system (Chap. 342). The innate immune system provides the rapid triggering of inflammatory responses based on the recognition (at the cell surface or within cells) of either molecules expressed by microorganisms or molecules that serve as "danger signals" released by cells under attack. These receptor/ligand interactions trigger signaling events that ultimately lead to inflammation. Virtually all cell lineages (not just immune cells) are involved in innate immune responses; however, myeloid cells (i.e., neutrophils and macrophages) play a major role because of their phagocytic capacity. The adaptive immune system operates by clonal recognition of antigens followed by a dramatic expansion of antigen-reactive cells and execution of an immune effector program. Most of the effector cells die off rapidly, whereas memory cells persist. Although both T and B lymphocytes recognize distinct chemical moieties and execute distinct adaptive immune responses, the latter is largely dependent on the former in generating long-lived humoral immunity. Adaptive responses utilize components of the innate immune system; for example, the antigen-presentation capabilities of dendritic cells help to determine the type of effector response. Not surprisingly, immune responses are controlled by a series of regulatory mechanisms.

Hundreds of gene products have been characterized as effectors or mediators of the immune system (Chap. 342). Whenever the expression or function of one of these products is genetically impaired (provided the function is nonredundant), a primary immunodeficiency (PID) occurs.

PIDs are genetic diseases with primarily Mendelian inheritance. More than 350 conditions have now been described, and deleterious mutations in ~346 genes have been identified. The overall prevalence of PIDs has been estimated in various countries at 5–10 per 100,000 individuals; however, given the difficulty in diagnosing these rare and complex diseases, this figure is probably an underestimate. PIDs can involve all possible aspects of immune responses, from innate through adaptive, cell differentiation, and effector function and regulation. For the sake of clarity, PIDs should be classified according to (1) the arm of the immune system that is defective and (2) the mechanism of the defect (when known). Table 344-1 classifies the most prevalent PIDs according to this manner of classification; however, one should bear in mind that the classification of PIDs sometimes involves arbitrary decisions because of overlap and, in some cases, lack of data.

TABLE 344-1

Classification of Primary Immune Deficiency Diseases

Deficiencies of the Innate Immune System	
• Phagocytic cells: - Impaired production: severe congenital neutropenia (SCN) - Asplenia - Impaired adhesion: leukocyte adhesion deficiency (LAD) - Impaired killing: chronic granulomatous disease (CGD) • Innate immunity receptors and signal transduction: - Defects in Toll-like receptor signaling - Mendelian susceptibility to mycobacterial disease • Complement deficiencies: - Classical, alternative, and lectin pathways - Lytic phase	
Deficiencies of the Adaptive Immune System	
• T lymphocytes: - Impaired development - Impaired survival, migration, function	Severe combined immune deficiencies (SCIDs) DiGeorge's syndrome Combined immunodeficiencies Hyper-IgE syndrome (autosomal dominant) DOCK8 deficiency CD40 ligand deficiency Wiskott-Aldrich syndrome Ataxia-telangiectasia and other DNA repair deficiencies
• B lymphocytes: - Impaired development - Impaired function	XL and AR agammaglobulinemia Hyper-IgM syndrome Common variable immunodeficiency (CVID) IgA deficiency
Regulatory Defects	
• Innate immunity • Adaptive immunity	Autoinflammatory syndromes (outside the scope of this chapter) Severe colitis Hemophagocytic lymphohistiocytosis (HLH) Autoimmune lymphoproliferation syndrome (ALPS) Autoimmunity and inflammatory diseases (IPEX, APECED)

Abbreviations: APECED, autoimmune polyendocrinopathy candidiasis ectodermal dysplasia; AR, autosomal recessive; IPEX, immunodysregulation polyendocrinopathy enteropathy X-linked syndrome; XL, X-linked.

The consequences of PIDs vary widely as a function of the molecules that are defective. This concept translates into multiple levels of vulnerability to infection by pathogenic and opportunistic microorganisms, ranging from extremely broad (as in severe combined immunodeficiency [SCID]) to narrowly restricted to a single microorganism (as in Mendelian susceptibility to mycobacterial disease [MSMD]). The locations of the sites of infection and the causal microorganisms involved will thus help physicians arrive at proper diagnoses. PIDs can also lead to immunopathologic responses such as allergy (as in Wiskott-Aldrich syndrome [WAS]), lymphoproliferation, and autoimmunity. A combination of recurrent infections, inflammation, and autoimmunity can be observed in a number of PIDs, thus creating obvious therapeutic challenges. Finally, some PIDs increase the risk of cancer, notably but not exclusively lymphocytic cancers, for example, lymphoma.

DIAGNOSIS OF PRIMARY IMMUNODEFICIENCIES

The most frequent symptom prompting the diagnosis of a PID is the presence of recurrent or unusually severe infections. As mentioned above, recurrent allergic or autoimmune manifestations may also alert the physician to a possible diagnosis of PID. In such cases, a detailed account of the subject's personal and family medical history should be obtained. It is of the utmost importance to gather as much medical information as possible on relatives and up to several generations of ancestors. In addition to the obvious focus on primary symptoms, the clinical examination should evaluate the size of lymphoid organs and, when appropriate, look for the characteristic signs of a number of complex syndromes that may be associated with a PID.

The performance of laboratory tests should be guided to some extent by the clinical findings. Infections of the respiratory tract (bronchi, sinuses) mostly suggest a defective antibody response. In general, invasive bacterial infections can result from complement deficiencies, signaling defects of innate immune responses, asplenia, or defective antibody responses. Viral infections, recurrent *Candida* infections, and opportunistic infections are generally suggestive of impaired T cell immunity. Skin infections and deep-seated abscesses primarily reflect innate immune defects (such as chronic granulomatous disease); however, they may also appear in the autosomal dominant hyper-IgE syndrome. Table 344-2 summarizes the laboratory tests that are most frequently used to diagnose a PID. More specific tests (notably genetic tests) are then used to make a definitive diagnosis. Genomic tools now allow us to more efficiently track genetic defects through usage of gene panel resequencing and/or whole exome sequencing.

TABLE 344-2

Tests Most Frequently Used to Diagnose a Primary Immune Deficiency (PID)

TEST	INFORMATION	PID DISEASE
• Blood cell counts and cell morphology	Neutrophil counts[a] Lymphocyte counts[a] Eosinophilia Howell-Jolly bodies	↓ Severe congenital neutropenia, ↑↑ LAD T cell ID WAS, hyper-IgE syndrome Asplenia
• Chest x-ray	Thymic shadow Costochondral junctions	SCID, DiGeorge's syndrome Adenosine deaminase deficiency
• Bone x-ray	Metaphyseal ends	Cartilage hair hypoplasia
• Immunoglobulin serum levels	IgG, IgA, IgM IgE	B cell ID Hyper-IgE syndrome, WAS, T cell ID
• Lymphocyte phenotype	T, B lymphocyte counts	T cell ID, agammaglobulinemia
• Dihydrorhodamine fluorescence (DHR) assay Nitroblue tetrazolium (NBT) assay	Reactive oxygen species production by PMNs	Chronic granulomatous disease
• CH50, AP50	Classic and alternative complement pathways	Complement deficiencies
• Ultrasonography of the abdomen	Spleen size	Asplenia

[a]Normal counts vary with age. For example, the lymphocyte count is between 3000 and 9000/μL of blood below the age of 3 months and between 1500 and 2500/μL in adults.

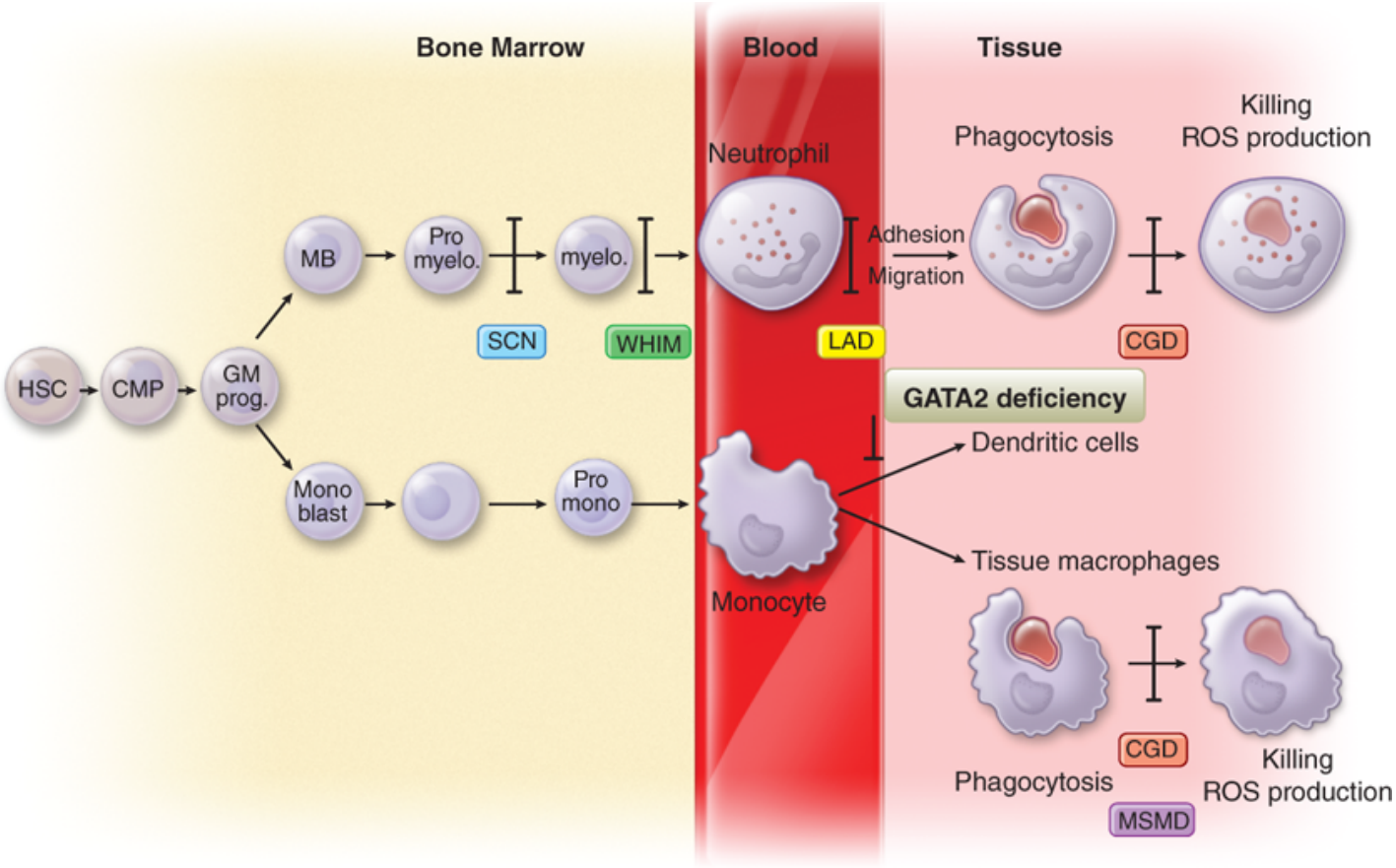
Abbreviations: ID, immunodeficiency; LAD, leukocyte adhesion deficiency; PMNs, polymorphonuclear leukocytes; SCID, severe combined immunodeficiency; WAS, Wiskott-Aldrich syndrome.

The PIDs discussed below have been grouped together according to the affected cells and the mechanisms involved (Table 344-1, **Fig. 344-1**).

FIGURE 344-1

Differentiation of phagocytic cells and related primary immunodeficiencies (PIDs). Hematopoietic stem cells (HSCs) differentiate into common myeloid progenitors (CMPs) and then granulocyte-monocyte progenitors (GM-prog.), which, in turn, differentiate into neutrophils (MB: myeloblasts; Promyelo: promyelocytes; myelo: myelocytes) or monocytes (monoblasts and promonocytes). Upon activation, neutrophils adhere to the vascular endothelium, transmigrate, and phagocytose the targets. Reactive oxygen species (ROS) are delivered to the microorganism-containing phagosomes. Macrophages in tissues kill using the same mechanism. Following activation by interferon γ (not shown here), macrophages can be armed to kill intracellular pathogens such as mycobacteria. For sake of simplicity, not all cell differentiation stages are shown. The abbreviations for PIDs are contained in boxes placed at corresponding stages of the pathway. CGD, chronic granulomatous diseases; GATA2, zinc finger

transcription factor; LAD, leukocyte adhesion deficiencies; MSMD, Mendelian susceptibility to mycobacterial disease; SCN, severe congenital neutropenia; WHIM, warts, hypogammaglobulinemia, infections, and myelokathexis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PRIMARY IMMUNODEFICIENCIES OF THE INNATE IMMUNE SYSTEM

PIDs of the innate immune system are relatively rare and account for ~10% of all PIDs.

SEVERE CONGENITAL NEUTROPENIA

Severe congenital neutropenia (SCN) consists of a group of inherited diseases that are characterized by severely impaired neutrophil counts (<500 polymorphonuclear leukocytes [PMN]/μL of blood). The condition is usually manifested from birth. SCN may also be cyclic (with a 3-week periodicity), and other neutropenia syndromes can also be intermittent. Although the most frequent inheritance pattern for SCN is autosomal dominant, autosomal recessive and X-linked recessive conditions also exist. Bacterial infections at the interface between the body and the external milieu (e.g., the orifices, wounds, and the respiratory tract) are common manifestations. Bacterial infections can rapidly progress through soft tissue and are followed by dissemination in the bloodstream. Severe visceral fungal infections can also ensue. The absence of pus is a hallmark of this condition.

Diagnosis of SCN requires examination of the bone marrow. Most SCNs are associated with a block in granulopoiesis at the promyelocytic stage (Fig. 344-1). SCN has multiple etiologies, and to date, mutations in 16 different genes have been identified. Most of these mutations result in isolated SCN, whereas others are syndromic (Chap. 60). The most frequent forms of SCN are caused by the premature cell death of granulocyte precursors, as observed in deficiencies of GFI1, HAX1, and elastase 2 (*ELANE*), with the latter accounting for 50% of SCN sufferers. Certain *ELANE* mutations cause cyclic neutropenia syndrome. A gain-of-

function mutation in the *WASP* gene (see the section on "Wiskott-Aldrich Syndrome" below) causes X-linked SCN, which is also associated with monocytopenia.

As mentioned above, SCN exposes the patient to life-threatening, disseminated bacterial and fungal infections. Treatment requires careful hygiene measures, notably in infants. Later in life, special oral and dental care is essential, along with the prevention of bacterial infection by prophylactic administration of trimethoprim/sulfamethoxazole. Subcutaneous injection of the cytokine granulocyte colony-stimulating factor (G-CSF) usually improves neutrophil development and thus prevents infection in most SCN diseases. However, there are two caveats: (1) a few cases of SCN with *ELANE* mutation are refractory to G-CSF and may require curative treatment via allogeneic hematopoietic stem cell transplantation (HSCT); and (2) a subset of G-CSF-treated patients carrying *ELANE* mutations are at a greater risk of developing acute myelogenous leukemia associated (in most cases) with somatic gain-of-function mutations of the G-CSF receptor gene.

A few SCN conditions are associated with additional immune defects involving leukocyte migration as observed in the WHIM syndrome (gain of function mutation of the chemokine CXCR4) or in moesin deficiency.

ASPLENIA

Primary failure of the development of a spleen is an extremely rare disease that can be either syndromic (in Ivemark syndrome) or isolated with an autosomal dominant expression; in the latter case, mutations in the ribosomal protein SA and the NKX2.5 genes were recently found. Due to the absence of natural filtration of microbes in the blood, asplenia predisposes affected individuals to fulminant infections by encapsulated bacteria. Although most infections occur in the first years of life, cases may also arise in adulthood. The diagnosis is confirmed by abdominal ultrasonography and the detection of Howell-Jolly bodies in red blood cells. Effective prophylactic measures (twice-daily oral penicillin and appropriate vaccination programs) usually prevent fatal outcomes.

GATA2 DEFICIENCY

Recently an immunodeficiency combining monocytopenia and dendritic and lymphoid (B and natural killer [NK]) cell deficiency (DCML), also called monocytopenia with nontuberculous mycobacterial infections (mono-MAC), has been described as a consequence of a dominant mutation in the gene *GATA2*, a transcription factor involved in hematopoiesis. This condition also predisposes to lymphedema, myelodysplasia, and acute myeloid leukemia. Infections (bacterial and viral) are life-threatening, thus indicating, together with the malignant risk, HSCT.

LEUKOCYTE ADHESION DEFICIENCY

Leukocyte adhesion deficiency (LAD) consists of three autosomal recessive conditions (LAD I, II, and III) **(Chap. 60)**. The most frequent condition (LAD I) is caused by mutations in the β2 integrin gene; following leukocyte activation, β2 integrins mediate adhesion to inflamed endothelium expressing cognate ligands. LAD III results from a defect in a regulatory protein (kindlin, also known as Fermt 3) involved in activating the ligand affinity of β2 integrins. The extremely rare LAD II condition is the end result of a defect in selectin-mediated leukocyte rolling that occurs prior to β2 integrin binding. There is a primary defect in fucose transporter such that oligosaccharide selectin ligands are missing in this syndromic condition.

Given that neutrophils are not able to reach infected tissues, LAD renders the individual susceptible to bacterial and fungal infections in a way that is similar to that of patients with SCN. LAD also causes impaired wound healing and delayed loss of the umbilical cord. A diagnosis can be suspected in cases of pus-free skin/tissue infections and massive hyperleukocytosis (>30,000/μL) in the blood (mostly granulocytes). Patients with LAD III also develop bleeding because the β2 integrin in platelets is not functional. Use of immunofluorescence and functional assays to detect β2 integrin can help form a diagnosis. Severe forms of LAD may require HSCT, although gene therapy is also now being considered. Neutrophil-specific granule deficiency (a very rare condition caused by a mutation in the gene for transcription factor C/EBPα) results in a condition that is clinically similar to LAD.

CHRONIC GRANULOMATOUS DISEASES

Chronic granulomatous diseases (CGDs) are characterized by impaired phagocytic killing of microorganisms by neutrophils and macrophages (Chap. 60). The incidence is ~1 per 200,000 live births. About 70% of cases are associated with X-linked recessive inheritance versus autosomal inheritance in the remaining 30%. CGD causes deep-tissue bacterial and fungal abscesses in macrophage-rich organs such as the lymph nodes, liver, and lungs. Recurrent skin infections (such as folliculitis) are common and can prompt an early diagnosis of CGD. The infectious agents are typically catalase-positive bacteria (such as *Staphylococcus aureus* and *Serratia marcescens*) but also include *Burkholderia cepacia*, pathogenic mycobacteria (in certain regions of the world), and fungi (mainly filamentous molds, such as *Aspergillus*).

CGD is caused by defective production of reactive oxygen species (ROS) in the phagolysosome membrane following phagocytosis of microorganisms. It results from the lack of a component of NADPH oxidase (gp91phox or p22phox) or of the associated adapter/activating proteins (p47phox, p67phox, or p40phox) that mediate the transport of electrons into the phagolysosome for creating ROS by interaction with O_2. Under normal circumstances, these ROS either directly kill engulfed microorganisms or enable the rise in pH needed to activate the phagosomal proteases that contribute to microbial killing. Diagnosis of CGD is based on assays of ROS production in neutrophils and monocytes (Table 344-2). As its name suggests, CGD is also a granulomatous disease. Macrophage-rich granulomas can often arise in the liver, spleen, and other organs. These are sterile granulomas that cause disease by obstruction (bladder, pylorus, etc.) or inflammation (colitis, restrictive lung disease).

The management of infections in patients with CGD can be a complex process. The treatment of bacterial infections is generally based on combination therapy with antibiotics that are able to penetrate into cells. The treatment of fungal infections requires aggressive, long-term use of antifungals. Inflammatory/granulomatous lesions are usually steroid-sensitive; however, glucocorticoids often contribute to the spread of infections. Hence, there is strong need for new therapeutic options in what is still a poorly understood disease.

The treatment of CGD mostly relies on preventing infections. It has been unambiguously demonstrated that prophylactic usage of trimethoprim/sulfamethoxazole is both well tolerated and highly effective in reducing the risk of bacterial infection. Daily administration of azole derivatives (notably itraconazole) also reduces the frequency of fungal complications. It has long been suggested that interferon γ administration is helpful, although medical experts continue to disagree over this controversial issue. Patients may do reasonably well with prophylaxis and careful management. However, other patients develop lifelong severe and persistent fungal infections and/or chronic inflammatory complications, leading to consideration of performing HSCT. Due to increase in reported successes, HSCT is now an established curative approach for CGD; however, the risk-versus-benefit ratio must be carefully assessed on a case-by-case basis. Gene therapy approaches are also being evaluated.

MENDELIAN SUSCEPTIBILITY TO MYCOBACTERIAL DISEASE

This group of diseases is characterized by a defect in the interleukin-12 (IL-12)–interferon (IFN) γ axis (including IL-12p40, IL-12 receptor [R] β_1, IFN-γ R_1 and R_2, STAT1, IRF8, and ISG515 deficiencies), which ultimately leads to impaired IFN-γ-dependent macrophage activation. Both recessive and dominant inheritance modes have been observed. The hallmark of this PID is a specific and narrow vulnerability to tuberculous and nontuberculous mycobacteria. The most severe phenotype (as observed in complete IFN-γ receptor deficiency) is characterized by disseminated infection that can be fatal even when aggressive and appropriate antimycobacterial therapy is applied. In addition to mycobacterial infections, MSMD patients (and particularly those with an IL-12/IL-12 R deficiency) are prone to developing *Salmonella* infections. Although MSMDs are very rare, they should be considered in any patient with persistent mycobacterial infection. Treatment with IFN-γ may efficiently bypass an IL-12/IL-12R deficiency.

TOLL-LIKE RECEPTOR (TLR) PATHWAY DEFICIENCIES

In a certain group of patients with early-onset, invasive *Streptococcus pneumoniae* infections or (less frequently) *Staphylococcus aureus* or other pyogenic infections, conventional screening for PIDs does not identify the cause of the defect in host defense. It has been established that these patients carry recessive mutations in genes that encode essential adaptor molecules (IRAK4 and MYD88) involved in the signaling pathways of the majority of known TLRs **(Chap. 342)**. Remarkably, susceptibility to infection appears to decrease after the first few years of life—perhaps an indication that adaptive immunity (once triggered by an initial microbial challenge) is then able to prevent recurrent infections.

Certain TLRs (TLR-3, 7, 8, and 9) are involved in the recognition of RNA and DNA and usually become engaged during viral infections. Very specific susceptibility to herpes simplex encephalitis has been described in patients with a deficiency in Unc93b (a molecule associated with TLR-3, 7, 8, and 9 required for correct subcellular localization), TLR-3, or associated signaling molecules TRIF, TBK1, and TRAF3, resulting in defective type I IFN production. The fact that no other TLR deficiencies have been found—despite extensive screening of patients with unexplained, recurrent infections—strongly suggests that these receptors are functionally redundant. Hypomorphic mutations in NEMO/IKK-γ (a member of the NF-κB complex, which is activated downstream of TLR receptors) lead to a complex, variable immunodeficiency, and a number of associated features. Susceptibility to both invasive, pyogenic infections and mycobacteria may be observed in this particular setting.

COMPLEMENT DEFICIENCY

The complement system is composed of a complex cascade of plasma proteins **(Chap. 342)** that leads to the deposition of C3b fragments on the surface of particles and the formation of immune complexes that can culminate in the activation of a lytic complex at the bacterial surface. C3 cleavage can be mediated via three pathways: the classic, alternate, and lectin pathways. C3b coats particles as part of the opsonization process that facilitates phagocytosis following binding to cognate receptors. A deficiency in any component of the classic pathway (C1q, C1r, C1s, C4, and C2) can predispose an individual to bacterial infections that are tissue-invasive or that occur in the respiratory tract. Likewise, a C3 deficiency or a deficiency in factor I (a protein that regulates C3 consumption, thus leading to a C3 deficiency due to its absence) also results in the same type of vulnerability to infection. It has recently been reported that a very rare deficiency in ficolin-3 predisposes affected individuals to bacterial infections. Deficiencies in the alternative pathway (factors D and properdin) are associated with the occurrence of invasive *Neisseria* infections.

Lastly, deficiencies of any complement component involved in the lytic phase (C5, C6, C7, C8, and, to a lesser extent, C9) predispose affected individuals to systemic infection by *Neisseria*. This is explained by the critical role of complement in the lysis of the thick cell wall possessed by this class of bacteria.

Diagnosis of a complement deficiency relies primarily on testing the status of the classic and alternate pathway via functional assays, that is, the CH50 and AP50 tests, respectively. When either pathway is profoundly impaired, determination of the status of the relevant components in that pathway enables a precise diagnosis. Appropriate vaccinations and daily administration of oral penicillin are efficient means of preventing recurrent infections. It is noteworthy that several complement deficiencies (in the classic pathway and the lytic phase) may also predispose affected individuals to autoimmune diseases (notably systemic lupus erythematosus; **Chap. 349**).

PRIMARY IMMUNODEFICIENCIES OF THE ADAPTIVE IMMUNE SYSTEM

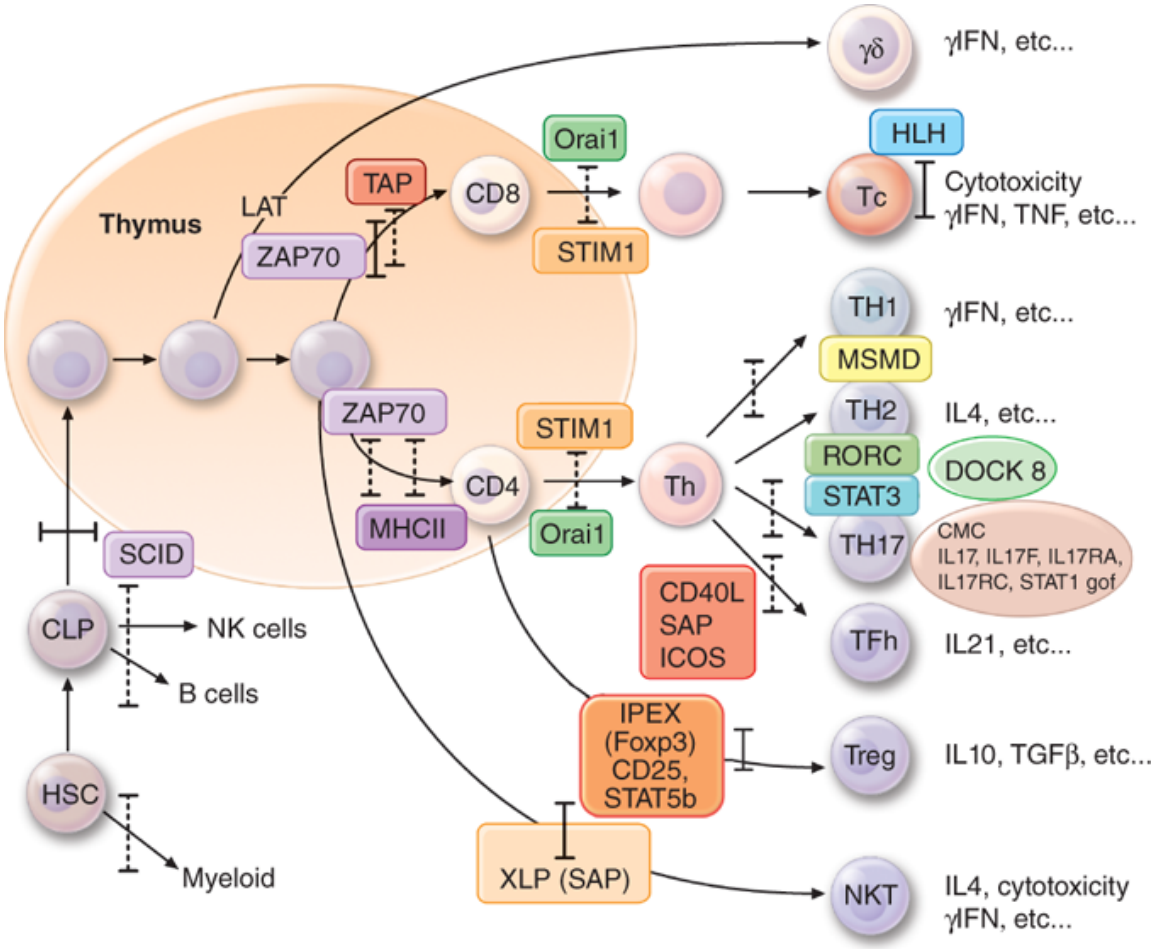
T LYMPHOCYTE DEFICIENCIES

Given the central role of T lymphocytes in adaptive immune responses **(Chap. 342)**, PIDs involving T cells generally have severe pathologic consequences; this explains the poor overall prognosis and the need for early diagnosis and the early intervention with appropriate therapy (**Table 344-1**, **Figs. 344-2** and **344-3**). Several differentiation pathways of T cell effectors have been described, one or all of which may be affected by a given PID (Fig. 344-2). Follicular helper CD4+ T cells in germinal centers are required for T-dependent antibody production, including the generation of Ig class-switched, high-affinity antibodies. CD4+ T_H1

cells provide cytokine-dependent (mostly IFN-γ-dependent) help to macrophages for intracellular killing of various microorganisms, including mycobacteria and *Salmonella*. CD4+ T_H2 cells produce IL-4, IL-5, and IL-13 and thus recruit and activate eosinophils and other cells required to fight helminth infections. CD4+ T_H17 cells produce IL-17 and IL-22 cytokines that recruit neutrophils to the skin and lungs to fight bacterial and fungal infections. Cytotoxic CD8+ T cells can kill infected cells, notably in the context of viral infections. In addition, certain T cell deficiencies predispose affected individuals to *Pneumocystis jiroveci* lung infections early in life and to chronic gut/biliary duct/liver infections by *Cryptosporidium* and related genera later on in life. Lastly, naturally occurring or induced regulatory T cells are essential for controlling inflammation (notably reactivity to commensal bacteria in the gut) and autoimmunity. The role of other T cell subsets with limited T cell receptor (TCR) diversity (such as γδTCR T cells or natural killer T [NKT] cells) in PIDs is less well known; however, these subsets can be defective in certain PIDs, and this finding can sometimes contribute to the diagnosis (e.g., NKT cell deficiency in X-linked proliferative syndrome [XLP]). T cell deficiencies account for ~20% of all cases of PID.

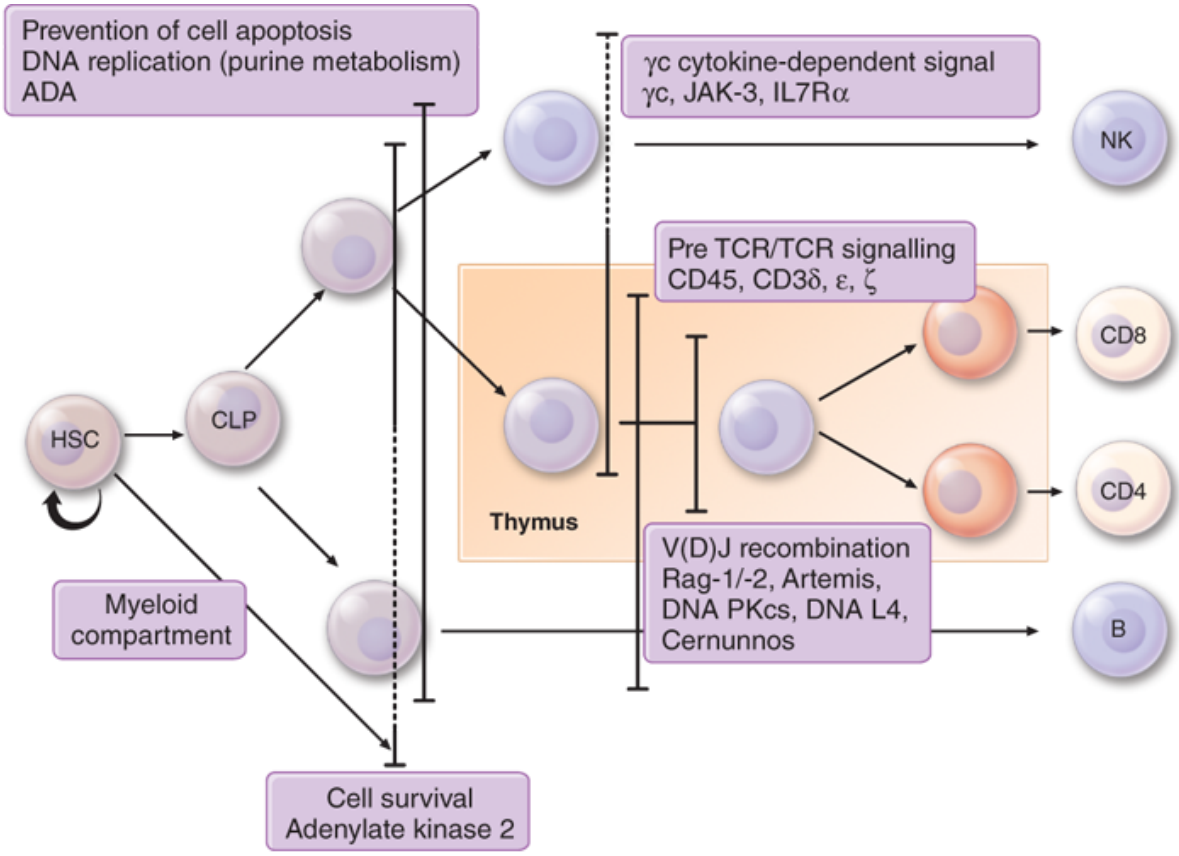
FIGURE 344-2

T cell differentiation, effector pathways, and related primary immunodeficiencies (PIDs). Hematopoietic stem cells (HSCs) differentiate into common lymphoid progenitors (CLPs), which, in turn, give rise to the T cell precursors that migrate to the thymus. The development of CD4+ and CD8+ T cells is shown. Known T cell effector pathways are indicated, that is, γδ cells, cytotoxic T cells (Tc), T_H1, T_H2, T_H17, TFh (follicular helper) CD4 effector T cells, regulatory T cells (Treg), and natural killer T cells (NKTs); abbreviations for PIDs are contained in boxes. Vertical bars indicate a complete deficiency; broken bars a partial deficiency. DOCK8, autosomal recessive form of hyper-IgE syndrome; HLH, hematopoietic lymphohistiocytosis; IL17F, IL17RA, STAT1 (gof: gain of function), CMC (chronic mucocutaneous candidiasis), CD40L, ICOS, SAP deficiencies; IPEX, immunodysregulation polyendocrinopathy enteropathy X-linked syndrome; LAT, linker for activation of T cells; MHCII, major histocompatibility complex class II deficiency; MSMD, Mendelian susceptibility to mycobacterial disease; Orai1, STIM1 deficiencies; RORC, RAR related orphan receptor C; SCID, severe combined immunodeficiency; STAT3, autosomal dominant form of hyper-IgE syndrome; TAP, TAP1 and TAP2 deficiencies; XLP, X-linked proliferative syndromes; ZAP70, zeta-associated protein deficiency.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.
FIGURE 344-3

T cell differentiation and severe combined immunodeficiencies (SCIDs). The vertical bars indicate the five mechanisms currently known to lead to SCID. The names of deficient proteins are indicated in the boxes adjacent to the vertical bars. A broken line means that deficiency is partial or involves only some of the indicated immunodeficiencies. ADA, adenosine deaminase deficiency; CLPs, common lymphoid progenitors; DNAL4, DNA ligase 4; HSCs, hematopoietic stem cells; NKs, natural killer cells; TCR, T cell receptor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Severe Combined Immunodeficiencies

SCIDs constitute a group of rare PIDs characterized by a profound block in T cell development and thus the complete absence of these cells. The developmental block is always the consequence of an intrinsic deficiency. The incidence of SCID is estimated to be 1 in 50,000 live births. Given the severity of the T cell deficiency, clinical consequences occur early in life (usually within 3 to 6 months of birth). The most frequent clinical manifestations are recurrent oral candidiasis, failure to thrive, and protracted diarrhea and/or acute interstitial pneumonitis caused by *Pneumocystis jiroveci* (although the latter can also be observed in the first year of life in children with B cell deficiencies). Severe viral infections or invasive bacterial infections can also occur. Patients may also experience complications related to infections caused by live vaccines (notably bacille Calmette-Guérin [BCG]) that may lead not only to local and regional infection but also to disseminated infection manifested by fever, splenomegaly, and skin and lytic bone lesions. A scaly skin eruption can be observed in a context of maternal T cell engraftment (see below). A diagnosis of SCID can be suspected based on the patient's clinical history and, possibly, a family history of deaths in very young children (suggestive of either X-linked or recessive inheritance). Lymphocytopenia is strongly suggestive of SCID in >90% of cases **(Table 344-2)**. The absence of a thymic shadow on a chest x-ray can also be suggestive of SCID. An accurate diagnosis relies on precise determination of the number of circulating T, B, and NK lymphocytes and their subsets. T cell lymphopenia may be masked in some patients by the presence of maternal T cells (derived from maternal-fetal blood transfers) that cannot be eliminated. Although counts are usually low (<500/μL of blood), higher maternal T cell counts may, under some circumstances, initially mask the presence of SCID. Thus, screening for maternal cells by using adequate genetic markers should be performed whenever necessary. Inheritance pattern analysis and lymphocyte phenotyping can discriminate between various forms of SCID and provide guidance in the choice of accurate molecular diagnostic tests (see below). To date, five distinct causative mechanisms for SCID (Fig. 344-3) have been identified. T cell quantification of receptor excision circles (TREC) by using the Guthrie card is a reliable diagnostic test for newborn screening. It is now operational in most of the United States and is being evaluated elsewhere. Its more widespread use will lead to the provision of therapy (see below) to uninfected patients resulting in a maximal chance of cure.

SEVERE COMBINED IMMUNODEFICIENCY CAUSED BY A CYTOKINE-SIGNALING DEFICIENCY

The most frequent SCID phenotype (accounting for 40–50% of all cases) is the absence of both T and NK cells. This outcome results from a deficiency in either the common γ chain (γc) receptor that is shared by several cytokine receptors (the IL-2, 4, 7, 9, 15, and 21 receptors) or Jak-associated kinase (JAK) 3 that binds to the cytoplasmic portion of the γc chain receptor and induces signal transduction following cytokine binding. The former form of SCID (γc deficiency) has an X-linked inheritance mode, while the second is autosomal recessive. A lack of the IL-7Rα chain (which, together with γc, forms the IL-7 receptor) induces a selective T cell deficiency.

PURINE METABOLISM DEFICIENCY

Ten to 20% of SCID patients exhibit a deficiency in adenosine deaminase (ADA), an enzyme of purine metabolism that deaminates adenosine (ado) and deoxyadenosine (dAdo). An ADA deficiency results in the accumulation of ado and dAdo metabolites that induce premature cell death of lymphocyte progenitors. The condition results in the absence of B and NK lymphocytes as well as T cells. The clinical expression of complete ADA deficiency typically occurs very early in life. Since ADA is a ubiquitous enzyme, its deficiency can also cause bone dysplasia with abnormal costochondral junctions and metaphyses (found in 50% of cases) and neurologic defects. The very rare purine nucleoside phosphorylase (PNP) deficiency causes a profound although incomplete T cell deficiency that is often associated with severe neurologic impairments.

DEFECTIVE REARRANGEMENTS OF T AND B CELL RECEPTORS

A series of SCID conditions are characterized by a selective deficiency in T and B lymphocytes with autosomal recessive inheritance. These conditions account for 20–30% of SCID cases and result from mutations in genes encoding proteins that mediate the recombination of V(D)J gene elements in T and B cell antigen receptor genes (required for the generation of diversity in antigen recognition). The main deficiencies involve RAG1, RAG2, DNA-dependent protein kinase, and Artemis. A less severe (albeit variable) immunologic phenotype can result from other deficiencies in the same pathway, that is, DNA ligase 4 and Cernunnos deficiencies. Given that these latter factors are involved in DNA repair, these deficiencies also cause develop-mental defects.

DEFECTIVE (PRE-)T CELL RECEPTOR SIGNALING IN THE THYMUS

A selective T cell defect can be caused by a series of rare deficiencies in molecules involved in signaling via the pre-TCR or the TCR. These include deficiencies in CD3 subunits associated with the (pre-)TCR (i.e., CD3δ, ε, and ζ) and CD45.

RETICULAR DYSGENESIS

Reticular dysgenesis is an extremely rare form of SCID that causes T and NK deficiencies with severe neutropenia and sensorineural deafness. It results from an adenylate kinase 2 deficiency.

Patients with SCID require appropriate care with aggressive anti-infective therapies, immunoglobulin replacement, and (when necessary) parenteral nutrition support. In most cases, curative treatment relies on HSCT. Today, HSCT provides a very high curative potential for SCID patients who are otherwise in reasonably good condition. In this regard, neonatal screening, based on quantification of TRECs on a Guthrie card sample, is being developed. Gene therapy has been found to be successful for cases of X-linked SCID (γc deficiency) and SCID caused by an ADA deficiency, although toxicity has become an issue in the treatment of the former disease that may now be overcome by use of newly generated vectors. Lastly, a third option for the treatment of ADA deficiency consists of enzyme substitution with a pegylated enzyme.

Thymic Defects

A profound T cell defect can also result from faulty development of the thymus, as is most often observed in rare cases of DiGeorge's syndrome—a relatively common condition leading to a constellation of developmental defects. In ~1% of such cases, the thymus is completely absent, leading to virtually no mature T cells. However, expansion of oligoclonal T cells can occur and is associated with skin lesions. Diagnosis (using immunofluorescence in situ hybridization) is based on the identification of a hemizygous deletion in the long arm of chromosome 22. To recover the capability for T cell differentiation, these cases require a thymic graft. CHARGE (*c*oloboma of the eye, *h*eart anomaly, choanal *a*tresia, *r*etardation, *g*enital, and *e*ar anomalies) syndrome (CHD7 deficiency) is a less frequent cause of impaired thymus development. Lastly, the very rare "nude" defect is characterized by the absence of both hair and the thymus.

Omenn Syndrome

Omenn syndrome consists of a subset of T cell deficiencies that present with a unique phenotype, including early-onset erythrodermia, alopecia, hepatosplenomegaly, and failure to thrive. These patients usually display T cell lymphocytosis, eosinophilia, and low B cell counts. It has been found that the T cells of these patients exhibit a low TCR heterogeneity. This peculiar syndrome is the consequence of hypomorphic mutations in genes usually associated with SCID, that is, *RAG-1*, *RAG-2*, or (less frequently) *Artemis* or *IL-7Rα*. The impaired homeostasis of differentiating T cells thus causes this immune system–associated disease. These patients are very fragile, requiring simultaneous anti-infective therapy, nutritional support, and immunosuppression. HSCT provides a curative approach.

Functional T Cell Defects

A subset of T cell PIDs with autosomal inheritance is characterized by partially preserved T cell differentiation but defective activation resulting in abnormal effector function **(Fig. 344-2)**. There are many causes of these defects, but all lead to susceptibility to viral and opportunistic infections, chronic diarrhea, and failure to thrive, with onset during childhood. Careful phenotyping and in vitro functional assays are required to identify these diseases, the best characterized of which are the following.

ZETA-ASSOCIATED PROTEIN 70 (ZAP70) DEFICIENCY

Zeta-associated protein 70 (ZAP70) is recruited to the TCR following antigen recognition. A ZAP70 deficiency leads typically to an almost complete absence of CD8+ T cells; CD4+ T cells are present but cannot be activated in vitro by TCR stimulation.

CALCIUM SIGNALING DEFECTS

A small number of patients have been reported who exhibit a profound defect in in vitro T and B cell activation as a result of defective antigen receptor-mediated Ca^{2+} influx. This defect is caused by a mutation in the calcium channel gene (*ORAI1*) or its activator (*STIM-1*). It is noteworthy that these patients are also prone to autoimmune manifestations (blood cytopenias) and exhibit a nonprogressive muscle disease.

HUMAN LEUKOCYTE ANTIGEN (HLA) CLASS II DEFICIENCY

Defective expression of HLA class II molecules is the hallmark of a group of four recessive genetic defects all of which affect molecules (RFX5, RFXAP, RFXANK, and CIITA) involved in the transactivation of the genes coding for HLA class II. As a result, low but variable CD4+ T cell counts are observed in addition to defective antigen-specific T and B cell responses. These patients are particularly susceptible to herpesvirus, adenovirus, and enterovirus infections and chronic gut/liver *Cryptosporidium* infections.

HLA CLASS I DEFICIENCY

Defective expression of molecules involved in antigen presentation by HLA class I molecules (i.e., TAP-1, TAP-2, and Tapasin) leads to reduced CD8+ T cell counts, loss of HLA class I antigen expression, and a particular phenotype consisting of chronic obstructive pulmonary disease and severe vasculitis.

OTHER DEFECTS

A variety of other T cell PIDs have been described, some of which are associated with a precise molecular defect (e.g., IL-2-inducible T cell kinase [ITK] deficiency, IL-21 and IL21 receptor deficiencies, CARD11 deficiency, DOCK2 deficiency, RORC deficiency). These conditions are also characterized by profound vulnerability to infections, such as severe Epstein-Barr virus (EBV)–induced B cell proliferation and autoimmune disorders in ITK deficiency. Milder phenotypes are associated with CD8 and CD3γ deficiencies.

HSCT is indicated for most of these diseases, although the prognosis is worse than in SCID because many patients are chronically infected at the time of diagnosis. Fairly aggressive immunosuppression and myeloablation may be necessary to achieve engraftment of allogeneic stem cells.

T Cell Primary Immunodeficiencies with DNA Repair Defects

This is a group of PIDs characterized by a combination of T and B cell defects of variable intensity, together with a number of nonimmunologic features resulting from DNA fragility. The autosomal recessive disorder *ataxia-telangiectasia* (AT) is the most frequently encountered condition in this group. It has an incidence of 1:40,000 live births and causes B cell defects (low IgA, IgG2 deficiency, and low antibody production), which often require immunoglobulin replacement. AT is associated with a progressive T cell immunodeficiency. As the name suggests, the hallmark features of AT are telangiectasia and cerebellar ataxia. The latter manifestations may not be detectable before the age of 3–4 years, so that AT should be considered in young children with IgA deficiency and recurrent and problematic infections. Diagnosis is based on a cytogenetic analysis showing excessive chromosomal rearrangements (mostly affecting chromosomes 7 and 14) in lymphocytes. AT is caused by a mutation in the gene encoding the ATM protein—a kinase that plays an important role in the detection and repair of DNA lesions (or cell death if the lesions are too numerous) by triggering several different pathways. Overall, AT is a progressive disease that carries a very high risk of lymphoma, leukemia, and (during adulthood) carcinomas. A variant of AT ("AT-like disease") is caused by mutation in the *MRE11* gene.

Nijmegen breakage syndrome (NBS) is a less common condition that also results from chromosome instability (with the same cytogenetic abnormalities as in AT). NBS is characterized by a severe T and B cell combined immune deficiency with autosomal recessive inheritance. Individuals with NBS exhibit microcephaly and a bird-like face, but have neither ataxia nor telangiectasia. The risk of malignancies is very high. NBS results from a deficiency in nibrin (NBSI, a protein associated with MRE11 and Rad50 that is involved in checking DNA lesions) caused by hypomorphic mutations.

Severe forms of *dyskeratosis congenita* (also known as Hoyeraal-Hreidarsson syndrome) combine a progressive immunodeficiency that can also include an absence of B and NK lymphocytes, progressive bone marrow failure, microcephaly, in utero growth retardation, and gastrointestinal disease. The disease can be X-linked or, more rarely, autosomal recessive. It is caused by the mutation of genes encoding telomere maintenance proteins, including dyskerin (DKC1).

Finally, *immunodeficiency with centromeric and facial anomalies* (ICF) is a complex syndrome of autosomal recessive inheritance that variably combines a mild T cell immune deficiency with a more severe B cell immune deficiency, coarse face, digestive disease, and mild mental retardation. A diagnostic feature is the detection by cytogenetic analysis of multiradial aspects in multiple chromosomes (most frequently 1, 9, and 16) corresponding to an abnormal DNA structure secondary to defective DNA methylation. It is the consequence of a deficiency in most cases in the DNA methyltransferase DNMT3B, ZBTB24, CDCA7, or HELLS.

T Cell Primary Immunodeficiencies with Hyper-IgE

Several T cell PIDs are associated with elevated serum IgE levels (as in Omenn syndrome). A condition sometimes referred to as *autosomal recessive hyper-IgE syndrome* is notably characterized by recurrent bacterial infections in the skin and respiratory tract and severe skin and mucosal infections by pox viruses and human papillomaviruses, together with severe allergic manifestations. T and B lymphocyte counts are low. Mutations in the *DOCK8* gene have been found in many of these patients. This condition is an indication for HSCT.

A very rare, related condition with autosomal recessive inheritance that causes a similar susceptibility to infection with various microbes (see above), including mycobacteria, reportedly results from a deficiency in Tyk-2, a JAK family kinase involved in the signaling of many different cytokine receptors.

Autosomal Dominant Hyper-IgE Syndrome

This unique condition, the *autosomal dominant hyper-IgE syndrome*, is usually diagnosed by the combination of recurrent skin and lung infections that can be complicated by pneumatoceles. Infections are caused by pyogenic bacteria and fungi. Several other manifestations characterize hyper-IgE syndrome, including facial dysmorphy, defective loss of primary teeth, hyperextensibility, scoliosis, and osteoporosis. Elevated serum IgE levels are typical of this syndrome. Defective T_H17 effector responses have been shown to account at least in part for the specific patterns of susceptibility to particular microbes. This

condition is caused by a heterozygous (dominant) mutation in the gene encoding the transcription factor STAT3 that is required in a number of signaling pathways following binding of cytokine to cytokine receptors (such as that of IL-6 and the IL-6 receptor). It also results in partially defective antibody production because of defective IL-21R signaling. Hence, immunoglobulin substitution can be considered as prophylaxis of bacterial infections.

Cartilage Hair Hypoplasia

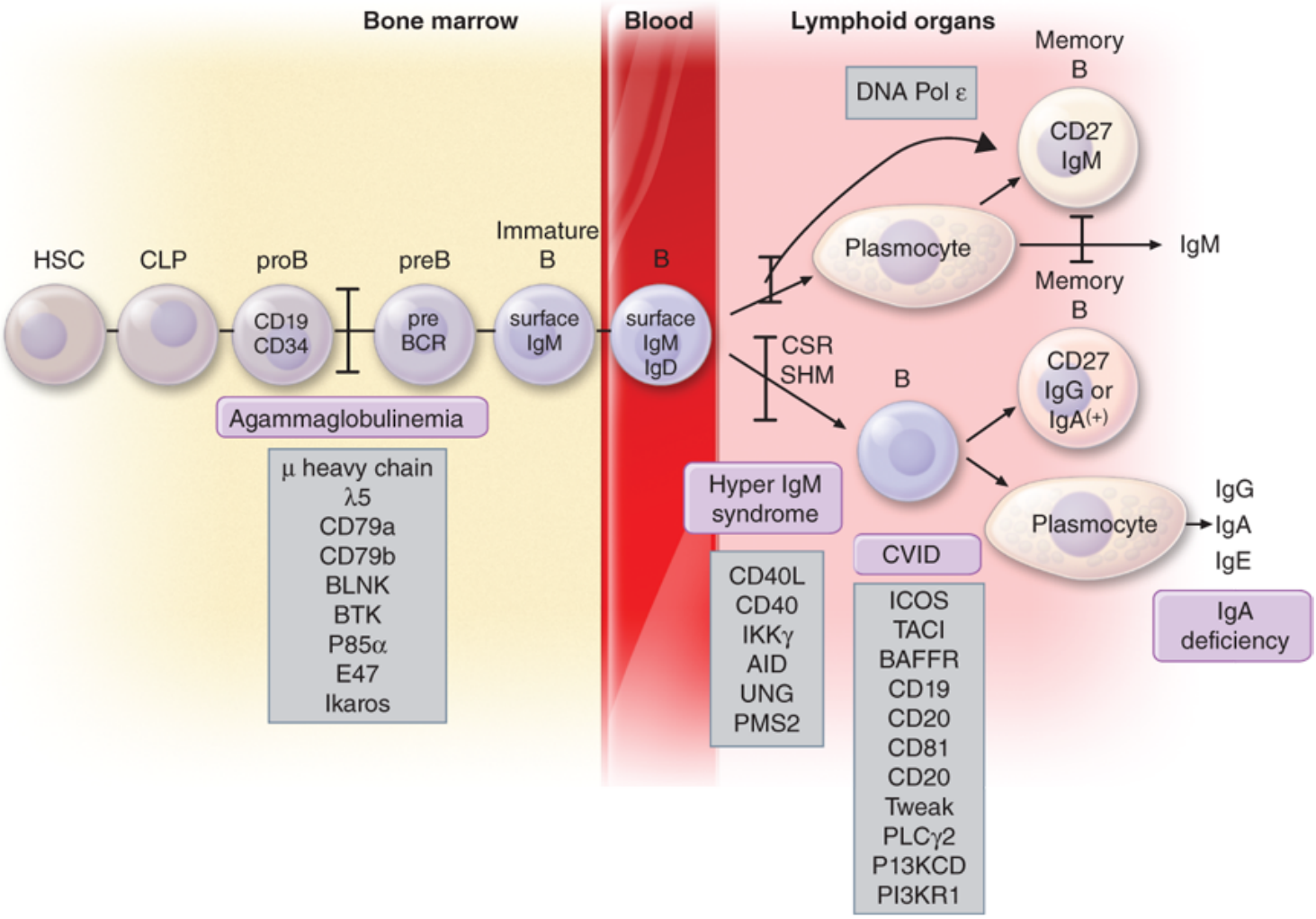
The autosomal recessive *cartilage hair hypoplasia* (CHH) disease is characterized by short-limb dwarfism, metaphyseal dysostosis, and sparse hair, together with a combined T and B cell PID of extremely variable intensity (ranging from quasi-SCID to no clinically significant immune defects). The condition can predispose to erythroblastopenia, autoimmunity, and tumors. It is caused by mutations in the *RMRP* gene for a noncoding ribosome-associated RNA.

CD40 Ligand and CD40 Deficiencies

Hyper-IgM syndrome (HIGM) is a well-known PID that is usually classified as a B cell immune deficiency (see **Fig. 344-4** and below). It results from defective immunoglobulin class switch recombination (CSR) in germinal centers and leads to profound deficiency in production of IgG, IgA, and IgE (although IgM production is maintained). Approximately half of HIGM sufferers are also prone to opportunistic infections, for example, interstitial pneumonitis caused by *Pneumocystis jiroveci* (in young children), protracted diarrhea and cholangitis caused by *Cryptosporidium*, and infection of the brain with *Toxoplasma gondii*.

FIGURE 344-4

B cell differentiation and related primary immunodeficiencies (PIDs). Hematopoietic stem cells (HSCs) differentiate into common lymphoid progenitors (CLPs), which give rise to pre-B cells. The B cell differentiation pathway goes through the pre–B cell stage (expression of the μ heavy chain and surrogate light chain), the immature B cell stage (expression of surface IgM), and the mature B cell stage (expression of surface IgM and IgD). The main phenotypic characteristics of these cells are indicated. In lymphoid organs, B cells can differentiate into plasma cells and produce IgM or undergo (in germinal centers) Ig class switch recombination (CSR) and somatic mutation of the variable region of V genes (SHM) that enable selection of high-affinity antibodies. These B cells produce antibodies of various isotypes and generate memory B cells. PIDs are indicated in the purple boxes. CVID, common variable immunodeficiency.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In the majority of cases, this condition has an X-linked inheritance and is caused by a deficiency in CD40 ligand (L). CD40L induces signaling events in B cells that are necessary for both CSR and adequate activation of other CD40-expressing cells that are involved in innate immune responses against the above-mentioned microorganisms. More rarely, the condition is caused by a deficiency in CD40 itself. The poorer prognosis of CD40L and CD40 deficiencies (relative to most other HIGM conditions) implies that (1) thorough investigations have to be performed in all cases of HIGM and (2) potentially curative HSCT should be discussed on a case-by-case basis for this group of patients.

Wiskott-Aldrich Syndrome

WAS is a complex, recessive, X-linked disease with an incidence of ~1 in 200,000 live births. It is caused by mutations in the *WASP* gene that affect not only T lymphocytes but also the other lymphocyte subsets, dendritic cells, and platelets. WAS is typically characterized by the following clinical manifestations: recurrent bacterial infections, eczema, and bleeding caused by thrombocytopenia. However, these manifestations are highly variable—mostly as a consequence of the many different *WASP* mutations that have been observed. Null mutations predispose affected individuals to invasive and bronchopulmonary infections, viral infections, severe eczema, and autoimmune manifestations. The latter include autoantibody-mediated blood cytopenia, glomerulonephritis, skin and visceral vasculitis (including brain vasculitis), erythema nodosum, and arthritis. Another possible consequence of WAS is lymphoma, which may be virally induced (e.g., by EBV or Kaposi's sarcoma–associated herpesvirus). Thrombocytopenia can be severe and compounded by the peripheral destruction of platelets associated with autoimmune disorders. Hypomorphic mutations usually lead to milder outcomes that are generally limited to thrombocytopenia. It is noteworthy that even patients with "isolated" X-linked thrombocytopenia can develop severe autoimmune disease or lymphoma later in life. The immunologic workup is not very informative; there can be a relative CD8+ T

cell deficiency, frequently accompanied by low serum IgM levels and decreased antigen-specific antibody responses. A typical feature is reduced-sized platelets on a blood smear. Diagnosis is based on intracellular immunofluorescence analysis of WAS protein (WASp) expression in blood cells. WASp regulates the actin cytoskeleton and thus plays an important role in many lymphocyte functions, including cell adhesion and migration and the formation of synapses between antigen-presenting and target cells. Predisposition to autoimmune disorders is in part related to defective regulatory T cells. The treatment of WAS should match the severity of disease expression. Prophylactic antibiotics, immunoglobulin G (IgG) supplementation, and careful topical treatment of eczema are indicated. Although splenectomy improves platelet count in a majority of cases, this intervention is associated with a significant risk of infection (both before and after HSCT). Allogeneic HSCT is curative, with fairly good results overall. Gene therapy trials are also under way. A similar condition has been reported in a girl with a deficiency in the Wiskott-Aldrich interacting protein (WIP).

A few other complex PIDs are worth mentioning. *Sp110 deficiency* causes a T cell PID with liver venoocclusive disease and hypogammaglobulinemia. *Chronic mucocutaneous candidiasis* (CMC) is a heterogeneous disease, considering the different inheritance patterns that have been observed. In some cases, chronic candidiasis is associated with late-onset bronchopulmonary infections, bronchiectasis, and brain aneurysms. Moderate forms of CMC are related to autoimmunity and AIRE deficiency (see below). In this setting, predisposition to *Candida* infection is associated with the detection of autoantibodies to T_H17 cytokines. Recently, deficiencies in IL-17A, IL-17F, and IL-17 receptor A and C and in the associated protein Act1, and above all, gain-of-function mutations in *STAT1* have been found to be associated with CMC. In all cases, CMC is related to defective T_H17 function. Innate immunodeficiency in CARD9 also predisposes to chronic invasive fungal infection.

B LYMPHOCYTE DEFICIENCIES

Deficiencies that predominantly affect B lymphocytes are the most frequent PIDs and account for 60–70% of all cases (Table 344-1, Fig. 344-4). B lymphocytes make antibodies. Pentameric IgMs are found in the vascular compartment and are also secreted at mucosal surfaces. IgG antibodies diffuse freely into extravascular spaces, whereas IgA antibodies are produced and secreted predominantly from mucosa-associated lymphoid tissues. Although Ig isotypes have distinct effector functions, including Fc receptor–mediated and (indirectly) C_3 receptor–dependent phagocytosis of microorganisms, they share the ability to recognize and neutralize a given pathogen. Defective antibody production therefore allows the establishment of invasive, pyogenic bacterial infections as well as recurrent sinus and pulmonary infections (mostly caused by *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Moraxella catarrhalis*, and, less frequently, gram-negative bacteria). If left untreated, recurrent bronchial infections lead to bronchiectasis and, ultimately, cor pulmonale and death. Parasitic infections such as caused by *Giardia lamblia* and bacterial infections caused by *Helicobacter* and *Campylobacter* of the gut are also observed. A complete lack of antibody production (namely agammaglobulinemia) can also predispose affected individuals to severe, chronic, disseminated enteroviral infections causing meningoencephalitis, hepatitis, and a dermatomyositis-like disease.

Even with the most profound of B cell deficiencies, infections rarely occur before the age of 6 months; this is because of transient protection provided by the transplacental passage of immunoglobulins during the last trimester of pregnancy. Conversely, a genetically nonimmunodeficient child born to a mother with hypogammaglobulinemia is, in the absence of maternal Ig substitution, usually prone to severe bacterial infections in utero and for several months after birth.

Diagnosis of B cell PIDs relies on the determination of serum Ig levels **(Table 344-2)**. Determination of antibody production following immunization with tetanus toxoid vaccine or nonconjugated pneumococcal polysaccharide antigens can also help diagnose more subtle deficiencies. Another useful test is B cell phenotype determination in switched μ−δ− CD27+ and nonswitched memory B cells (μ+δ+ CD27+). In agammaglobulinemic patients, examination of bone marrow B cell precursors (Fig. 344-4) can help obtain a precise diagnosis and guide the choice of genetic tests.

Agammaglobulinemia

Agammaglobulinemia is characterized by a profound defect in B cell development (<1% of the normal B cell blood count). In most patients, very low residual Ig isotypes can be detected in the serum. In 85% of cases, agammaglobulinemia is caused by a mutation in the *BTK* gene that is located on the X chromosome. The *BTK* gene product is a kinase that participates in (pre) B cell receptor signaling. When the kinase is defective, there is a block (albeit a leaky one) at the pre-B to B cell stage (Fig. 344-4). Detection of *BTK* by intracellular immunofluorescence of monocytes, and lack thereof in patients with X-linked agammaglobulinemia (XLA), is a useful diagnostic test. Not all of the mutations in *BTK* result in agammaglobulinemia, since some patients have a milder form of hypogammaglobulinemia and low but detectable B cell counts. These cases should not be confused with common variable immunodeficiency (CVID, see below). About 10% of agammaglobulinemia cases are caused by alterations in genes encoding elements of the pre-B cell receptor, i.e., the μ heavy chain, the λ5 surrogate light chain, Igα or Igβ, the scaffold protein BLNK, and the p85 α subunit of phosphatidylinositol 3 phosphate kinase (P13K) and the Ikaros transcription factor. In 5% of cases, the defect is unknown. It is noteworthy that agammaglobulinemia can be observed in patients with ICF syndrome, despite the presence of normal peripheral B cell counts. Lastly, agammaglobulinemia can be a manifestation of a myelodysplastic syndrome (associated or not with neutropenia). Treatment of agammaglobulinemic patients is based on immunoglobulin replacement (see below). Profound hypogammaglobulinemia is also observed in adults, in association with thymoma.

Hyper-IgM (HIGM) Syndromes

HIGM is a rare B cell PID characterized by defective Ig CSR. It results in very low serum levels of IgG and IgA and elevated or normal serum IgM levels. The clinical severity is similar to that seen in agammaglobulinemia, although chronic lung disease and sinusitis are less frequent and enteroviral infections are uncommon. As discussed above, a diagnosis of HIGM involves screening for an X-linked CD40L deficiency and an autosomal recessive CD40 deficiency, which affect both B and T cells. In 50% of cases affecting only B cells, these isolated HIGM syndromes result from mutations in the gene encoding activation-induced deaminase, the protein that induces CSR in B cell germinal centers. These patients usually have enlarged lymphoid organs. In the other 50% of cases, the etiology is unknown (except for rare UNG and PMS2 deficiencies). Furthermore, IgM-mediated autoimmunity and lymphomas can occur in HIGM syndrome. It is noteworthy that HIGM can result from fetal rubella syndrome or can be a predominant immunologic feature of other PIDs, such as the immunodeficiency associated with ectodermic anhydrotic hypoplasia X-linked NEMO deficiency and the combined T and B cell PIDs caused by DNA repair defects such as AT and Cernunnos deficiency.

Common Variable Immunodeficiency

CVID is an ill-defined condition characterized by low serum levels of one or more Ig isotypes. Its prevalence is estimated to be 1 in 20,000. The condition is recognized predominantly in adults, although clinical manifestations can occur earlier in life. Hypogammaglobulinemia is associated with at least partially defective antibody production in response to vaccine antigens. B lymphocyte counts are often normal but can be low. Besides infections, CVID patients may develop lymphoproliferation (splenomegaly), granulomatous lesions, colitis, antibody-mediated autoimmune disease, and lymphomas. A family history is found in 10% of cases. A clear-cut dominant inheritance pattern is found in some families, whereas recessive inheritance is observed more rarely. In most cases, no molecular cause can be identified. A small number of patients in Germany were found to carry mutations in the *ICOS* gene encoding a T cell-membrane protein that contributes to B cell activation and survival. In 10% of patients with CVID, monoallelic or biallelic mutations of the gene encoding TACI (a member of the tumor necrosis factor [TNF] receptor family that is expressed on B cells) have been found. In fact, heterozygous TACI mutations correspond to a genetic susceptibility factor, since similar heterozygous mutations are found in 1% of controls. The B-cell activating factor (BAFF) receptor was found to be defective in a kindred with CVID, although not all individuals carrying the mutation have CVID. Recently a group of patients with hypogammaglobulinemia and lymphoproliferation was shown to exhibit dominant gain of function mutations in the *PIK3CD* gene encoding the p110δ form of P13 kinase or in the *PI3KR1* gene encoding the regulatory p85α subunit of PI3 kinase. Rare cases of hypogammaglobulinemia were found to be associated with CD19 and CD81 deficiencies. These patients have B cells that can be identified by typing for other B cell markers.

A diagnosis of CVID should be made after excluding the presence of hypomorphic mutations associated with agammaglobulinemia or more subtle T cell defects; this is particularly the case in children. It is possible that many cases of CVID result from a constellation of factors, rather than a single genetic defect. Hypogammaglobulinemia can be associated with neutropenia and lymphopenia in the warts, hypogammaglobulinemia, infections, and myelokathexis syndrome (WHIM) caused by dominant gain-of-function mutation of *CXCR4*, resulting in cell retention in the bone marrow.

Selective Ig Isotype Deficiencies

IgA deficiency and CVID represent polar ends of a clinical spectrum due to the same underlying gene defect(s) in a large subset of these patients. IgA deficiency is the most common PID; it can be found in 1 in every 600 individuals. It is asymptomatic in most cases; however, individuals may present with increased numbers of acute and chronic respiratory infections that may lead to bronchiectasis. In addition, over their lifetime, these patients experience an increased susceptibility to drug allergies, atopic disorders, and autoimmune diseases. Symptomatic IgA deficiency is probably related to CVID, since it can be found in relatives of patients with CVID. Furthermore, IgA deficiency may progress to CVID. It is thus important to assess serum Ig levels in IgA-deficient patients (especially when infections occur frequently) in order to detect changes that should prompt the initiation of immunoglobulin replacement. Selective IgG2 (+G4) deficiency (which in some cases may be associated with IgA deficiency) can also result in recurrent sinopulmonary infections and should thus be specifically sought in this clinical setting. These conditions are ill-defined and often transient during childhood. A pathophysiologic explanation has not been found.

Selective Antibody Deficiency to Polysaccharide Antigens

Some patients with normal serum Ig levels are prone to *S. pneumoniae* and *H. influenzae* infections of the respiratory tract. Defective production of antibodies against polysaccharide antigens (such as those in the *S. pneumoniae* cell wall) can be observed and is probably causative. This condition may correspond to a defect in marginal zone B cells, a B cell subpopulation involved in T-independent antibody responses.

Immunoglobulin Replacement

IgG antibodies have a half-life of 21–28 days. Thus, injection of plasma-derived polyclonal IgG containing a myriad of high-affinity antibodies can provide protection against disease-causing microorganisms in patients with defective IgG antibody production. This form of therapy should not be based on laboratory data alone (i.e., IgG and/or antibody deficiency) but should be guided by the occurrence or not of infections; otherwise, patients might be subjected to unjustified IgG infusions. Immunoglobulin replacement can be performed by IV or subcutaneous routes. In the former case, injections have to be repeated every 3–4 weeks, with a residual target level of 800 mg/mL in patients who had very low IgG levels prior to therapy. Subcutaneous injections are typically performed once a week, although the frequency can be adjusted on a case-by-case basis. A trough level of 800 mg/mL is desirable. Whatever the mode of administration, the main goal is to reduce the frequency of the respiratory tract infections and prevent chronic lung and sinus disease. The two routes appear to be equally safe and efficacious, and so the choice should be left to the preference of the patient.

In patients with chronic lung disease, chest physical therapy with good pulmonary toilet and the cyclic use of antibiotics are also needed. Immunoglobulin replacement is well tolerated by most patients, although the selection of the best-tolerated Ig preparation may be necessary in certain cases. Since IgG preparations contain a small proportion of IgAs, caution should be taken in patients with residual antibody production capacity and a complete IgA deficiency, as these subjects may develop anti-IgA antibodies that can trigger anaphylactic shock. These patients should be treated with IgA-free IgG preparations. Immunoglobulin replacement is a lifelong therapy; its rationale and procedures have to be fully understood and mastered by the patient and his or her family in order to guarantee the strict observance required for efficacy.

PRIMARY IMMUNODEFICIENCIES AFFECTING REGULATORY PATHWAYS

An increasing number of PIDs have been found to cause homeostatic dysregulation of the immune system, either alone or in association with increased vulnerability to infections (Table 344-1). Defects of this type affecting the innate immune system and autoinflammatory syndromes will not be covered in this chapter. However, three specific entities (hemophagocytic lymphohistiocytosis [HLH], lymphoproliferation, and autoimmunity) will be described below.

HEMOPHAGOCYTIC LYMPHOHISTIOCYTOSIS

HLH is characterized by an unremitting activation of CD8+ T lymphocytes and macrophages that leads to organ damage (notably in the liver, bone marrow, and central nervous system). This syndrome results from a broad set of inherited diseases, most of which impair T and NK lymphocyte cytotoxicity. The manifestations of HLH are often induced by a viral infection. EBV is the most frequent trigger. In severe forms of HLH, disease onset may start during the first year of life or even (in rare cases) at birth.

Diagnosis relies on the identification of the characteristic symptoms of HLH (fever, hepatosplenomegaly, edema, neurologic diseases, blood cytopenia, increased liver enzymes, hypofibrinogenemia, high triglyceride levels, elevated markers of T cell activation, and hemophagocytic features in the bone marrow or cerebrospinal fluid). Functional assays of postactivation cytotoxic granule exocytosis (CD107 fluorescence at the cell membrane) can suggest genetically determined HLH. The conditions can be classified into three subsets:

1. Familial HLH with autosomal recessive inheritance, including perforin deficiency (30% of cases) that can be recognized by assessing intracellular perforin expression; Munc13-4 deficiency (30% of cases); syntaxin 11 deficiency (10% of cases); Munc18-2 deficiency (20% of cases); and a few residual cases that lack a known molecular defect.

2. HLH with partial albinism. Three conditions combine HLH and abnormal pigmentation, where hair examination can help in the diagnosis: Chédiak-Higashi syndrome, Griscelli syndrome, and Hermansky-Pudlak syndrome type II. Chédiak-Higashi syndrome is also characterized by the presence of giant lysosomes within leukocytes **(Chap. 60)**, in addition to a primary neurologic disorder with slow progression of symptoms over time.

3. XLP is characterized in most patients by the induction of HLH following EBV infection, while other patients develop progressive hypogammaglobulinemia similar to what is observed in CVID and/or certain lymphomas. XLP is caused by a mutation in the *SH2DIA* gene that encodes the adaptor protein SAP (associated with a SLAM family receptor). Several immunologic abnormalities have been described, including low 2B4-mediated NK cell cytotoxicity, impaired differentiation of NKT cells, defective antigen-induced T cell death, and defective T cell helper activity for B cells. A related disorder (XLP2) has recently been described. It is also X-linked and induces HLH (frequently after EBV infection), although the clinical manifestation may be less pronounced. The condition is associated with a deficiency of the antiapoptotic molecule XIAP. The pathophysiology of XLP2 remains unclear; however, it may be related to control of inflammation in macrophages as there is a functional link between XIAP and NLRC4, an inflammasome component, in which gain of function can also induce HLH. XLP2 is also frequently associated with colitis.

HLH is a life-threatening complication. The treatment of this condition requires aggressive immunosuppression with either the cytotoxic agent etoposide or anti–T cell antibodies; specific therapy targeting interferon γ, which is critical in causing HLH, is an additional option to consider. Once remission has been achieved, HSCT should be performed, since it provides the only curative form of therapy.

AUTOIMMUNE LYMPHOPROLIFERATIVE SYNDROME

Autoimmune lymphoproliferative syndrome (ALPS) is characterized by nonmalignant T and B lymphoproliferation causing splenomegaly and enlarged lymph nodes; 70% of patients also display autoimmune manifestations such as autoimmune cytopenias, Guillain-Barré syndrome, uveitis, and hepatitis **(Chaps. 62 and 342)**. A hallmark of ALPS is the presence of CD4–CD8–TCRαβ+ T cells (2–50%) in the blood of affected individuals. Hypergammaglobulinemia involving IgG and IgA is also frequently

observed. The syndrome is caused by a defect in Fas-mediated apoptosis of lymphocytes, which can thus accumulate and mediate autoimmunity. Furthermore, ALPS can lead to malignancies.

Most patients carry a heterozygous mutation in the gene encoding Fas that is characterized by dominant inheritance and variable penetrance, depending on the nature of the mutation. A rare and severe form of the disease with early onset can be observed in patients carrying a biallelic mutation of Fas, which profoundly impairs the protein's expression and/or function. Fas-ligand, caspase 10, caspase 8, and somatic neuroblastoma RAS viral oncogene homologue (NRAS) mutations have also been reported in a few cases of ALPS. Many cases of ALPS have not been precisely delineated at the molecular level. A B cell–predominant ALPS has recently been found associated with a protein kinase Cδ gene mutation. Treatment of ALPS is essentially based on the use of proapoptotic drugs, which need to be carefully administered in order to avoid toxicity.

COLITIS, AUTOIMMUNITY, AND PRIMARY IMMUNODEFICIENCIES

Several PIDs (most of which are T cell–related) can cause severe gut inflammation. The prototypic example is *immunodysregulation polyendocrinopathy enteropathy X-linked syndrome* (IPEX), characterized by a widespread inflammatory enteropathy, food intolerance, skin rashes, autoimmune cytopenias, and diabetes. The syndrome is caused by loss-of-function mutations in the gene encoding the transcription factor FOXP3, which is required for the acquisition of effector function by regulatory T cells. In most cases of IPEX, CD4+CD25+ regulatory T cells are absent from the blood. This condition has a poor prognosis and requires aggressive immunosuppression. The only possible curative approach is allogeneic HSCT. IPEX-like syndromes that lack a FOXP3 mutation have also been described. In some cases, a CD25 deficiency has been found. Defective CD25 expression also impairs regulatory cell expansion/function. This functional T cell deficiency means that CD25-deficient patients are also at increased risk of opportunistic infections. It is noteworthy that abnormalities in regulatory T cells have been described in other PID settings, such as in Omenn syndrome, STAT5b deficiency, STIM1 (Ca flux) deficiency, and WAS; these abnormalities may account (at least in part) for the occurrence of inflammation and autoimmunity. The autoimmune features observed in a small fraction of patients with DiGeorge's syndrome may have the same cause. Recently, severe inflammatory gut disease has been described in patients with a deficiency in the IL-10 receptor or IL-10.

Dominant mutations in genes encoding the regulatory molecule CTLA4, recessive mutations in the gene encoding LRBA (a molecule involved in recycling of CTLA4) as well as dominant gain of function mutation of STAT3 cause a multifaceted lymphoproliferative and autoimmune syndrome, frequently involving inflammatory bowel disease that can be associated with hypogammaglobulinemia. Molecular diagnosis is required before adapted targeted therapies are undertaken.

A distinct autoimmune entity is observed in *autoimmune polyendocrinopathy candidiasis ectodermal dysplasia* (APECED) syndrome, which is characterized by autosomal recessive inheritance. It consists of multiple autoimmune manifestations that can affect solid organs in general and endocrine glands in particular. Mild, chronic *Candida* infection is often associated with this syndrome. The condition is due to mutations in the autoimmune regulator (*AIRE*) gene and results in impaired thymic expression of self-antigens by medullary epithelial cells and impaired negative selection of self-reactive T cells that leads to autoimmune manifestations.

A combination of hypogammaglobulinemia, autoantibody production, cold-induced urticaria or skin granulomas, or autoinflammation has been reported, and has been termed the *PLCγ2-associated antibody deficiency and immune dysregulation* (PLAID or APLAID).

CONCLUSION

The variety and complexity of the clinical manifestations of the many different PIDs strongly indicate that it is important to raise awareness of these diseases. Indeed, early diagnosis is essential for establishing an appropriate therapeutic regimen. Hence, patients with suspected PIDs must always be referred to experienced clinical centers that are able to perform appropriate

molecular and genetic tests. A precise molecular diagnosis is not only necessary for initiating the most suitable treatment, but is also important for genetic counseling and prenatal diagnosis.

One pitfall that may hamper diagnosis is the high variability that is associated with many PIDs. Variable disease expression can result from the differing consequences of various mutations associated with a given condition, as exemplified by WAS and, to a lesser extent, XLA. There can also be effects of modifier genes (as also suspected in XLA) and environmental factors such as EBV infection that can be the main trigger of disease in XLP conditions. Furthermore, it has recently been established that somatic mutations in an affected gene can attenuate the phenotype of a number of T cell PIDs. This has been described for ADA deficiency, X-linked SCID, RAG deficiencies, NF-κB essential modulator (NEMO) deficiency, and, most frequently, WAS. In contrast, somatic mutations can create disease states analogous to PID, as reported for ALPS. Lastly, cytokine-neutralizing autoantibodies can mimic a PID, as shown for IFN-γ.

Many aspects of the pathophysiology of PIDs are still unknown, and the disease-causing gene mutations have not been identified in all cases (as illustrated by CVID and IgA deficiency). However, our medical understanding of PIDs has now reached the stage where scientifically based approaches to the diagnosis and treatment of these diseases can be implemented. A genetic diagnosis has become a milestone step in the care of PID patients.

FURTHER READING

Casanova JL et al: Guidelines for genetic studies in single patients: Lessons from primary immunodeficiencies. J Exp Med 211:2137, 2014. [PubMed: 25311508]

Fischer A et al: Gene therapy for primary immunodeficiencies. Clin Genet 88:507, 2015. [PubMed: 25708106]

Fischer A et al: Severe combined immunodeficiencies and related disorders. Nat Rev Dis Primers 1:15061, 2015. [PubMed: 27189259]

Holland SM: Chronic granulomatous disease. Hematol Oncol Clin North Am 2013;27:89, viii. [PubMed: 23351990]

Kwan A et al: Newborn screening for severe combined immunodeficiency in 11 screening programs in the United States. JAMA 312:729, 2014. [PubMed: 25138334]

Notarangelo LD: Functional T cell immunodeficiencies (with T cells present). Annu Rev Immunol 31:195, 2013. [PubMed: 23298211]

Ochs HD et al (eds): *Primary Immunodeficiencies: A Molecular and Genetic Approach*. New York, Oxford University Press, 2013.

Picard C et al: International Union of Immunological Societies: 2017 Primary Immunodeficiency Diseases Committee Report on Inborn Errors of Immunity. J Clin Immunol 38:96, 2018.

Picard C, Fischer A: Contribution of high-throughput DNA sequencing to the study of primary immunodeficiencies. Eur J Immunol 44:2854, 2014. [PubMed: 25154746]

Resnick ES, Cunningham-Rundles C: The many faces of the clinical picture of common variable immune deficiency. Curr Opin Allergy Clin Immunol 12:595, 2012. [PubMed: 23026770]

Chapter 345: Urticaria, Angioedema, and Allergic Rhinitis

Katherine N. Cahill; Joshua A. Boyce

INTRODUCTION

The term *atopy* implies a tendency to manifest asthma, rhinitis, urticaria, and atopic dermatitis alone or in combination, in association with the presence of allergen-specific IgE. However, individuals without an atopic background may also develop hypersensitivity reactions, particularly urticaria and anaphylaxis, associated with the presence of IgE. Since mast cells are key effector cells in allergic rhinitis and asthma, and the dominant effector in urticaria, anaphylaxis, and systemic mastocytosis, its developmental biology, activation pathway, product profile, and target tissues will be considered in the introduction to these clinical disorders. Dysregulation of mast cell development seen in mastocytosis will be covered in a separate chapter.

The binding of IgE to human mast cells and basophils, a process termed *sensitization*, prepares these cells for subsequent antigen-specific activation. The high-affinity Fc receptor for IgE, designated FcεRI, is composed of one α, one β, and two disulfide-linked γ chains, which together cross the plasma membrane seven times. The α chain is responsible for IgE binding, and the β and γ chains provide for signal transduction that follows the aggregation of the sensitized tetrameric receptors by polymeric antigen. The binding of IgE stabilizes the α chain at the plasma membrane, thus increasing the density of FcεRI receptors at the cell surface while sensitizing the cell for effector responses. This accounts for the correlation between serum IgE levels and the numbers of FcεRI receptors detected on circulating basophils. Signal transduction is initiated through the action of a Src family–related tyrosine kinase termed Lyn that is constitutively associated with the β chain. Lyn transphosphorylates the canonical immunoreceptor tyrosine-based activation motifs (ITAMs) of the β and γ chains of the receptor, resulting in recruitment of more active Lyn to the β chain and of Syk tyrosine kinase. The phosphorylated tyrosines in the ITAMs function as binding sites for the tandem *src* homology two (SH2) domains within Syk. Syk activates not only phospholipase Cγ, which associates with the linker of activated T cells at the plasma membrane, but also phosphatidylinositol 3-kinase to provide phosphatidylinositol-3,4,5-trisphosphate, which allows membrane targeting of the Tec family kinase Btk and its activation by Lyn. In addition, the Src family tyrosine kinase Fyn becomes activated after aggregation of IgE receptors and phosphorylates the adapter protein Gab2 that enhances activation of phosphatidylinositol 3-kinase. Indeed, this additional input is essential for mast cell activation, but it can be partially inhibited by Lyn, indicating that the extent of mast cell activation is in part regulated by the interplay between these Src family kinases. Activated phospholipase Cγ cleaves phospholipid membrane substrates to provide inositol-1,4,5-trisphosphate (IP_3) and 1,2-diacylglycerols (1,2-DAGs) so as to mobilize intracellular calcium and activate protein kinase C, respectively. The subsequent opening of calcium-regulated activated channels provides the sustained elevations of intracellular calcium required to recruit the mitogen-activated protein kinases, ERK, JNK, and p38 (serine/threonine kinases), which provide cascades to augment arachidonic acid release and to mediate nuclear translocation of transcription factors for various cytokines. The calcium ion–dependent activation of phospholipases cleaves membrane phospholipids to generate lysophospholipids, which, like 1,2-DAG, may facilitate the fusion of the secretory granule perigranular membrane with the cell membrane, a step that releases the membrane-free granules containing the preformed mast cell mediators.

The secretory granule of the human mast cell has a crystalline structure, unlike mast cells of lower species. IgE-dependent cell activation results in solubilization and swelling of the granule contents within the first minute of receptor perturbation; this reaction is followed by the ordering of intermediate filaments about the swollen granule, movement of the granule toward the cell surface, and fusion of the perigranular membrane with that of other granules and with the plasmalemma to form extracellular channels for mediator release while maintaining cell viability.

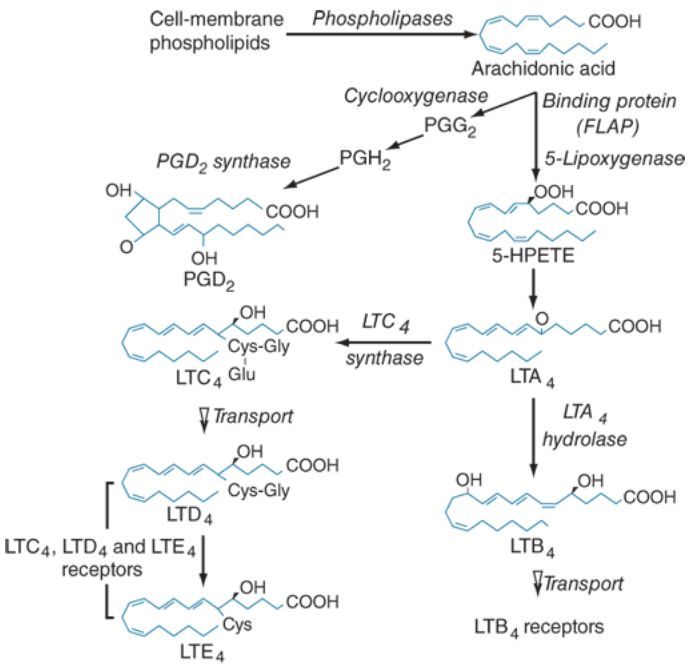
In addition to exocytosis, aggregation of FcεRI initiates two other pathways for generation of bioactive products, namely, lipid mediators and cytokines. The biochemical steps involved in expression of such cytokines as tumor necrosis factor α (TNF-α), interleukin (IL) 1, IL-6, IL-4, IL-5, IL-13, granulocyte-macrophage colony-stimulating factor (GM-CSF), and others, including an array of chemokines, have not been specifically defined for mast cells. Inhibition studies of cytokine production (IL-1β, TNF-α, and IL-6) in mouse mast cells with cyclosporine or FK506 reveal binding to the ligand-specific immunophilin and attenuation of the calcium ion- and calmodulin-dependent serine/threonine phosphatase, calcineurin.

Lipid mediator generation **(Fig. 345-1)** involves translocation of calcium ion–dependent cytosolic phospholipase A_2 to the outer nuclear membrane, with subsequent release of arachidonic acid for metabolic processing by the distinct prostanoid and leukotriene pathways. The constitutive prostaglandin endoperoxide synthase-1 (PGHS-1/cyclooxygenase-1) and the de novo inducible PGHS-2 (cyclooxygenase-2) convert released arachidonic acid to the sequential intermediates, prostaglandins G_2 and H_2. The glutathione-dependent hematopoietic prostaglandin D_2 (PGD_2) synthase then converts PGH_2 to PGD_2, the predominant mast cell prostanoid. The PGD_2 receptor DP_1 is expressed by platelets, natural killer cells, dendritic cells, and epithelial cells, whereas DP_2 is expressed by T_H2 lymphocytes, innate lymphoid type 2 cells, eosinophils, and basophils. Mast cells also generate thromboxane A_2 (TXA_2), a short lived but powerful mediator that induces bronchoconstriction and platelet activation through the T prostanoid (TP) receptor.

FIGURE 345-1

Pathways for biosynthesis and release of membrane-derived lipid mediators from mast cells. In the 5-lipoxygenase pathway, leukotriene A_4 (LTA_4) is the intermediate from which the terminal-pathway enzymes generate the distinct final products, leukotriene C_4 (LTC_4) and leukotriene B_4 (LTB_4), which leave the cell by separate

saturable transport systems. Gamma glutamyl transpeptidase and a dipeptidase then cleave glutamic acid and glycine from LTC_4 to form LTD_4 and LTE_4, respectively. The major mast cell product of the cyclooxygenase system is PGD_2.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

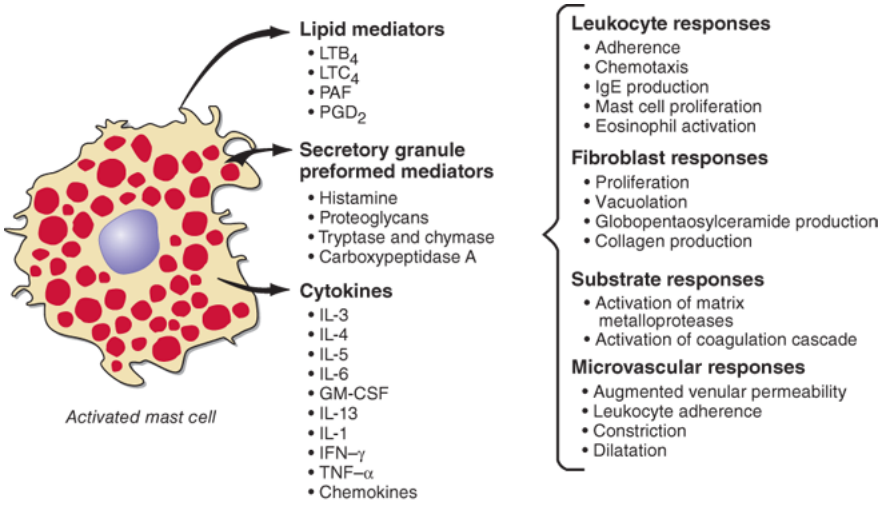
For leukotriene biosynthesis, the released arachidonic acid is metabolized by 5-lipoxygenase (5-LO) in the presence of an integral nuclear membrane protein, 5-LO activating protein (FLAP). The calcium ion–dependent translocation of 5-LO to the nuclear membrane converts the arachidonic acid to the sequential intermediates, 5-hydroperoxyeicosatetraenoic acid (5-HPETE) and leukotriene (LT) A_4. LTA_4 is conjugated with reduced glutathione by LTC_4 synthase, an integral nuclear membrane protein homologous to FLAP. Intracellular LTC_4 is released by a carrier-specific export step for extracellular metabolism to the additional cysteinyl leukotrienes, LTD_4 and LTE_4, by the sequential removal of glutamic acid and glycine. Alternatively, cytosolic LTA_4 hydrolase converts some LTA_4 to the dihydroxy leukotriene LTB_4, which also undergoes specific export. Two receptors for LTB_4, BLT_1 and BLT_2, mediate chemotaxis of human neutrophils. Two receptors for the cysteinyl leukotrienes, $CysLT_1$ and $CysLT_2$, are present on smooth muscle of the airways and the microvasculature and on hematopoietic cells such as macrophages, eosinophils, and mast cells. Whereas the $CysLT_1$ receptor has a preference for LTD_4 and is blocked by the receptor antagonists in clinical use, the $CysLT_2$ receptor is equally responsive to LTD_4 and LTC_4, is unaffected by these antagonists, and is a negative regulator of the function of the $CysLT_1$ receptor. LTD_4, acting at $CysLT_1$ receptors, is the most potent known bronchoconstrictor, whereas LTE_4 induces a vascular leak and mediates the recruitment of eosinophils to the bronchial mucosa. Recently, GPR99, CysLT3 receptor, was identified as an LTE4 receptor. The lysophospholipid formed during the release of arachidonic acid from 1-*O*-alkyl-2-acyl-*sn*-glyceryl-3-phosphorylcholine can be acetylated in the second position to form platelet-activating factor (PAF). Serum levels of PAF correlated positively with the severity of anaphylaxis to peanut in a recent study, whereas the levels of PAF acetyl hydrolase (a PAF-degrading enzyme) were inversely related to the same outcome.

Unlike most other cells of bone marrow origin, mast cells circulate as committed progenitors lacking their characteristic secretory granules. These committed progenitors express c-*kit*, the receptor for stem cell factor (SCF). Unlike most other lineages, they retain and increase c-*kit* expression with maturation. The SCF interaction with c-*kit* is an absolute requirement for the development of constitutive tissue mast cells residing in skin and connective tissue sites and for the accumulation of mast cells at mucosal surfaces during T_H2-type immune responses. Several T cell-derived cytokines (IL-3, IL-4, IL-5, and IL-9) can potentiate SCF-dependent mast cell proliferation and/or survival in vitro in mice and humans. Indeed, mast cells are absent from the intestinal mucosa in clinical T cell deficiencies, but are present in the submucosa. Based on the immunodetection of secretory granule neutral proteases, mast cells in the lung parenchyma and intestinal mucosa selectively express tryptase, and those in the intestinal and airway submucosa, perivascular spaces, skin, lymph nodes, and breast parenchyma express tryptase, chymase, and carboxypeptidase A (CPA). In the mucosal epithelium of severe asthmatics, mast cells can express tryptase and CPA without chymase. The secretory granules of mast cells selectively positive for tryptase exhibit closed scrolls with a periodicity suggestive of a crystalline structure by electron microscopy, whereas the secretory granules of mast cells with multiple proteases are scroll-poor, with an amorphous or lattice-like appearance.

Mast cells are distributed at cutaneous and mucosal surfaces and in submucosal tissues about venules and could influence the entry of foreign substances by their rapid response capability **(Fig. 345-2)**. Upon stimulus-specific activation and secretory granule exocytosis, histamine and acid hydrolases are solubilized, whereas the neutral proteases, which are cationic, remain largely bound to the anionic proteoglycans, heparin and chondroitin sulfate E, with which they function as a complex. Histamine and the various lipid mediators (PGD_2, $LTC_4/D_4/E_4$, PAF) alter venular permeability, thereby allowing influx of plasma proteins such as complement and immunoglobulins, whereas LTB_4 mediates leukocyte–endothelial cell adhesion and subsequent directed migration (chemotaxis). The accumulation of leukocytes and plasma opsonins facilitates defense of the microenvironment. The inflammatory response can also be detrimental, as in asthma, where the smooth-muscle constrictor activity of the cysteinyl leukotrienes is evident and much more potent than that of histamine.

FIGURE 345-2

Bioactive mediators of three categories generated by IgE-dependent activation of murine mast cells can elicit common but sequential target cell effects leading to acute and sustained inflammatory responses. GM-CSF, granulocyte-macrophage colony-stimulating factor; IL, interleukin; IFN, interferon; LT, leukotriene; PAF, platelet-activating factor; PGD_2, prostaglandin D_2; TNF, tumor necrosis factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The cellular component of the mast cell–mediated inflammatory response is augmented and sustained by cytokines and chemokines. IgE-dependent activation of human skin mast cells in situ elicits TNF-α production and release, which in turn induces endothelial cell responses favoring leukocyte adhesion. Similarly, activation of purified human lung mast cells or cord blood–derived cultured mast cells in vitro results in substantial production of proinflammatory (TNF-α) and immunomodulatory cytokines (IL-4, IL-5, IL-13) and chemokines. Bronchial biopsy specimens from patients with asthma reveal that mast cells are immunohistochemically positive for IL-4 and IL-5, but that the predominant localization of IL-4, IL-5, and GM-CSF is to T cells, defined as T_H2 by this profile. IL-4 modulates the T cell phenotype to the T_H2 subtype, determines the isotype switch to IgE (as does IL-13), and upregulates FcεRI-mediated expression of cytokines by mast cells based on in vitro studies.

An immediate and late cellular phase of allergic inflammation can be induced in the skin, nose, or lung of some allergic humans with local allergen challenge. The immediate phase in the nose involves pruritus and watery discharge; in the lung, it involves bronchospasm and mucus secretion; and in the skin, it involves a wheal-and-flare response with pruritus. The reduced nasal patency, reduced pulmonary function, or erythema with swelling at the skin site in a late-phase response at 6–8 h is associated with biopsy findings of infiltrating and activated T_H2 cells, eosinophils, basophils, and some neutrophils. The progression from early mast cell activation to late cellular infiltration has been used as an experimental surrogate of rhinitis or asthma. However, in asthma, there is an intrinsic hyperreactivity of the airways independent of the associated inflammation. Moreover, early- and late-phase responses (at least in the lung) are far more sensitive to blockade of IgE-dependent mast cell activation (or actions of histamine and cysteinyl leukotrienes) than are spontaneous or virally induced asthma exacerbations.

Consideration of the mechanism of immediate-type hypersensitivity diseases in the human has focused largely on the IgE-dependent recognition of otherwise innocuous substances. A region of chromosome 5 (5q23-31) contains genes implicated in the control of IgE levels including IL-4 and IL-13, as well as IL-3 and IL-9, which are involved in mucosal mast cell hyperplasia, and IL-5 and GM-CSF, which are central to eosinophil development and their enhanced tissue viability. Genes with linkage to the specific IgE response to particular allergens include those encoding the major histocompatibility complex (MHC) and certain chains of the T cell receptor (TCR-αδ). The complexity of atopy and the associated diseases includes susceptibility, severity, and therapeutic responses, each of which is among the separate variables modulated by both innate and adaptive immune stimuli.

The induction of allergic disease requires sensitization of a predisposed individual to a specific allergen. The greatest propensity for the development of atopic allergy occurs in childhood and early adolescence. The allergen is processed by antigen-presenting cells of the monocytic lineage (particularly dendritic cells) located throughout the body at surfaces that contact the outside environment, such as the nose, lungs, eyes, skin, and intestine. These antigen-presenting cells present the epitope-bearing peptides via their MHC to T helper cells and their subsets. The T cell response depends both on cognate recognition and on the cytokine microenvironment provided by the antigen-presenting dendritic cells, with IL-4 directing a T_H2 subset, interferon (IFN) γ a T_H1 profile, and IL-6 with transforming growth factor β (TGF-β) a T_H17 subset. Allergens not only present antigenic epitopes via dendritic cells but also contain pattern recognition ligands that facilitate the immune response by direct initiation of cytokine generation from innate cell types such as basophils, mast cells, eosinophils, and others. The T_H2 response is associated with activation of specific B cells that can also present allergens or that transform into plasma cells for antibody production. Synthesis and release into the plasma of allergen-specific IgE results in sensitization of FcεR1-bearing cells such as mast cells and basophils, which become activated on exposure to the specific allergen. In certain diseases, including those associated with atopy, the monocyte and eosinophil populations can express a trimeric FcεR1, which lacks the β chain, and yet respond to its aggregation. An additional recently recognized class of c-*kit*-expressing innate cells (termed group 2 innate lymphoid cells of ILC2) can generate large quantities of IL-5 and IL-13 during antihelminth responses, are prominent in nasal polyps from humans, and contribute to inflammation in allergic diseases.

URTICARIA AND ANGIOEDEMA

DEFINITION

Urticaria and angioedema represent the same pathophysiologic process occurring at different levels of the skin. Urticaria involves dilation of vascular structures in the superficial dermis, while angioedema originates from the deeper dermis and subcutaneous tissues. Not surprisingly they often appear together, with roughly 40% of patients reporting both, and affect >20% of the population at sometime during their lifespan. Urticaria can occur on any area of the body as well-circumscribed wheals with erythematous raised serpiginous borders and blanched centers that may coalesce to become giant wheals. Urticarial lesions last for <24 h, frequently migrate around the body, leave no bruising or scarring and are intensely pruritic. Angioedema is marked by dramatic swelling with more pain than pruritus and minimal erythema, which may develop with a pruritic prodrome and takes hours to days to resolve. Acute urticaria and/or angioedema are episodes that occur for <6 weeks' duration, whereas attacks persisting >6 weeks are designated chronic.

PREDISPOSING FACTORS AND ETIOLOGY

Acute or chronic urticaria and/or angioedema can occur at any point in the lifespan with the third to fifth decade the most common for chronic. Women are affected more often than men with a slight predominance for those with a history of atopy. Acute urticaria is most often the result of exposure to a food, environmental or drug allergen or viral infection while chronic urticaria is often idiopathic.

The classification of urticaria-angioedema presented in **Table 345-1** focuses on the different mechanisms for eliciting clinical disease and can be useful for differential diagnosis.

TABLE 345-1

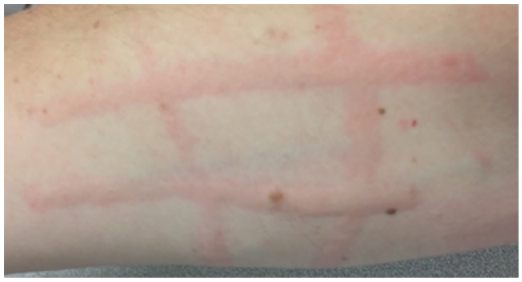
Classification of Urticaria and/or Angioedema

ACUTE	CHRONIC
Drug Reactions	Idiopathic—a subset with autoimmune component
NSAIDS, IV contrast, angiotensin-converting enzyme (ACE) inhibitors, etc.	Collagen vascular disease—urticarial vasculitis
Foods	Physical stimuli
Inhalation or contact with environmental allergens	Dermatographism
Transfusion reactions	Cholinergic urticaria
Insects	Vibration, cold, pressure, water (aquagenic)
Infections—viral, bacterial, parasitic	Sun (solar)
	Mastocytosis/Urticaria pigmentosa
	Hereditary
	Hereditary angioedema (HAE)
	Familial cold urticaria
	C3b inhibitor deficiency
	Muckle-Wells syndrome
	Schnitzler syndrome
	Hypereosinophilic syndrome
	Gleich syndrome

Additional etiologies include physical stimuli such as cold, heat, solar rays, exercise, and mechanical irritation. The physical urticarias can be distinguished by the precipitating event and other aspects of the clinical presentation. *Dermographism*, which occurs in 1–4% of the population, is defined by the appearance of a linear wheal with surrounding erythema at the site of a brisk stroke with a firm object **(Fig. 345-3)**. Dermographism has a prevalence that peaks in the second to third decades. It is not influenced by atopy and has a duration generally of <5 years. *Pressure urticaria*, which often accompanies chronic idiopathic urticaria, presents in response to a sustained stimulus such as a shoulder strap or belt, running (feet), or manual labor (hands). *Cholinergic urticaria* is distinctive in that the pruritic wheals are of small size (1–2 mm) and are surrounded by a large area of erythema; attacks are precipitated by fever, a hot bath or shower, or exercise and are presumptively attributed to a rise in core body temperature. *Exercise-induced anaphylaxis* can be precipitated by exertion alone or can be dependent on prior food ingestion. There is an association with the presence of IgE specific for α-5 gliadin, a component of wheat. The clinical presentation can be limited to flushing, erythema, and pruritic urticaria but may progress to angioedema of the face, oropharynx, larynx, or intestine or to vascular collapse; it is distinguished from cholinergic urticaria by presenting with wheals of conventional size and by not occurring with fever or a hot bath. *Cold urticaria* is local at body areas exposed to low ambient temperature or cold objects but can progress to vascular collapse with immersion in cold water (swimming). *Solar urticaria* is subdivided into six groups by the response to specific portions of the light spectrum. *Vibratory angioedema* may occur after years of occupational exposure or can be idiopathic; it may be accompanied by cholinergic urticaria. Other rare forms of physical allergy, always defined by stimulus-specific elicitation, include *local heat urticaria*, *aquagenic urticaria* from contact with water of any temperature (sometimes associated with polycythemia vera), and *contact urticaria* from direct interaction with some chemical substance (such as latex).

FIGURE 345-3

Dermographic urticarial lesion induced by stroking the forearm lightly with the edge of a tongue blade. The photograph, taken after 10 min, demonstrates a prominent wheal-and-flare reaction in the shape of a *hashtag. (Photograph provided by Katherine N. Cahill, MD, Harvard Medical School.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Isolated Angioedema

Angioedema without urticaria can be idiopathic or due to the generation of bradykinin in the setting of C1 inhibitor (C1INH) deficiency that may be inborn as an autosomal dominant characteristic or may be acquired through the appearance of an autoantibody in the setting of malignancy. The angiotensin-converting enzyme (ACE) inhibitors can provoke a similar clinical presentation in 0.2–0.7% of exposed patients due to delayed degradation of bradykinin. Black race, organ transplant, female gender, smoking, and increasing age are known risk factors for ACE-inhibitor related angioedema.

CLINICAL PRESENTATION AND PATHOPHYSIOLOGY

Urticarial eruptions are distinctly pruritic, may involve any area of the body from the scalp to the soles of the feet, and appear in crops of 12- to 36-h duration, with old lesions fading as new ones appear. Most of the physical urticarias (cold, cholinergic, dermatographism) are an exception, with individual lesions lasting <2 h. Neither urticaria nor angioedema lesions are symmetric or dependent in distribution. The most common sites for angioedema are often periorbital and perioral. Angioedema of the upper respiratory tract may be life-threatening due to transient laryngeal obstruction, whereas gastrointestinal involvement may present with abdominal colic, with or without nausea and vomiting, and can result in unnecessary surgical intervention. No residual scaring occurs with either urticaria or angioedema unless there is an underlying vasculitic process.

The pathology is characterized by edema of the superficial dermis in urticaria and of the subcutaneous tissue and deep dermis in angioedema. Collagen bundles in affected areas are widely separated, and the venules are sometimes dilated. Any perivenular infiltrate consists of lymphocytes, monocytes, eosinophils, and neutrophils that are present in varying combination and numbers.

The best evidence for IgE- and mast cell–involvement in urticaria and angioedema is *cold urticaria*. Cryoglobulins or cold agglutinins are present in up to 5% of these patients. Immersion of an extremity in an ice bath precipitates angioedema of the distal portion with urticaria at the air interface within minutes of the challenge. Histologic studies reveal marked mast cell degranulation with associated edema of the dermis and subcutaneous tissues. Elevated levels of histamine have been found in the plasma of venous effluent and in the fluid of suction blisters at experimentally induced lesional sites in patients with cold urticaria, dermographism, pressure urticaria, vibratory angioedema, light urticaria, and heat urticaria. By ultrastructural analysis, the pattern of mast cell degranulation in cold urticaria resembles an IgE-mediated response with solubilization of granule contents, fusion of the perigranular and cell membranes, and discharge of granule contents, whereas in a dermatographic lesion, there is additional superimposed zonal (piecemeal) degranulation. Elevations of plasma histamine levels with biopsy-proven mast cell degranulation have also been demonstrated with generalized attacks of *cholinergic urticaria*.

Up to 45% of patients with chronic urticaria have an autoimmune cause for their disease including autoantibodies to IgE or to the α chain of FcεRI. In some patients, autologous serum injected into their own skin can induce a wheal-and-flare reaction involving mast cell activation. The presence of these antibodies can also be recognized by their capacity to release histamine or induce activation markers such as CD63 or CD203 on basophils. An association with antibodies to microsomal peroxidase and/or thyroglobulin has been observed with both clinically significant Hashimoto's thyroiditis as well as a euthyroid state. In vitro studies reveal that these autoantibodies can mediate basophil degranulation with enhancement by serum as a source of the anaphylatoxic fragment, C5a.

The urticaria and angioedema associated with classic serum sickness or with hypocomplementemic cutaneous necrotizing angiitis (urticarial vasculitis) are believed to be immune-complex diseases. Reactions to mast cell granule–releasing agents (opioids, contrast media) and to non-steroidal anti-inflammatory drugs are most often limited to urticaria and/or angioedema, but may be systemic.

Hereditary angioedema (HAE) is a fully penetrant, autosomal dominant disease due to a mutation in the *SERPING1* gene leading to a deficiency of C1INH (type 1) in about 85% of patients or to a dysfunctional protein (type 2) in the remainder affecting 1:30,000–80,000 in the general population. A third less common type of HAE has been described in which C1INH function is normal, and the causal lesion is a mutant form of factor XII, which leads to generation of excessive bradykinin. C1INH deficiency can also develop in a sporadic acquired form as a result of excessive consumption of C1INH due either to formation of immune complexes or to the generation of an autoantibody directed to C1INH in the setting of lymphoproliferative disease. C1INH blocks the catalytic function of activated factor XII (Hageman factor) and of kallikrein, as well as the C1r/C1s components of C1, with the common result of degrading bradykinin. During clinical attacks of angioedema, C1INH function or levels fall, patients develop elevated plasma levels of bradykinin leading to angioedema and excessive activation of C1 results in a decline in C4 and C2 levels.

The use of ACE inhibitors results in impaired bradykinin degradation and explains the angioedema that occurs idiosyncratically in ACE inhibitor-exposed patients with a normal C1INH. Bradykinin-mediated angioedema, whether caused by ACE inhibitors or by C1INH deficiency, is noteworthy for the conspicuous absence of concomitant urticaria or pruritus, the frequent involvement of the gastrointestinal tract, and the duration of symptoms >24 h.

DIAGNOSIS

The classification of urticarial and angioedematous states as presented in Table 345-1 in terms of duration can facilitate identification of possible mechanisms. History alone of self-limited urticarial and/or angioedema episodes can be sufficient to make a diagnosis in the setting of acute disease triggered by drug, environmental or food allergen with history-directed confirmatory skin testing or assay for serum allergen-specific IgE. Direct reproduction of the lesion in physical urticarias is particularly valuable because it so often establishes the cause of the lesion. Even with chronic urticaria/angioedema, initial diagnostic testing should be limited and expanded testing guided by history. Complete blood count with assessment for eosinophilia, erythrocyte sedimentation rate and thyroid stimulation hormone level are recommended by consensus guidelines even though the vast majority of chronic urticaria is associated with no laboratory abnormality. Urticarial lesions that last longer than 36 h result in scarring and are reported as painful and not pruritic warrant biopsy to evaluate for cellular infiltration, nuclear debris, and fibrinoid necrosis of the venules consistent with urticarial vasculitis. Chronic angioedema without urticaria warrants assessment of complement levels. Concomitant flushing and hyperpigmented papules that urticate with stroking in the absence of angioedema raise the question of mastocytosis. An appropriate travel history should trigger an evaluation for parasites.

The diagnosis of HAE is suggested not only by family history but also by the lack of pruritus and of urticarial lesions, the prominence of recurrent gastrointestinal attacks of colic, and episodes of laryngeal edema. Laboratory diagnosis depends on demonstrating a deficiency of C1INH antigen (type 1) or a nonfunctional protein (type 2) by a catalytic inhibition assay. While levels of C1 are normal, its substrates, C4 and C2, are chronically depleted and fall further during attacks due to the activation of additional C1. Patients with the acquired forms of C1INH deficiency have the same clinical manifestations but differ in the lack of a familial element. Furthermore, their sera exhibit a reduction of C1 function and C1q protein as well as C1INH, C4, and C2. Inborn C1INH deficiency and ACE inhibitor–elicited angioedema are associated with elevated levels of bradykinin. Lastly, type 3 HAE is associated with normal levels of complement proteins and a factor XII gene mutation.

TREATMENT

TREATMENT

Urticaria and Angioedema

For most forms of urticaria, H_1 antihistamines such as chlorpheniramine or diphenhydramine effectively attenuate both urtication and pruritus, but because of their side effects and short half-life, long-acting, non-sedating agents such as loratadine, desloratadine, and fexofenadine, or low-sedating agents such as cetirizine or levocetirizine generally are used first and increased to four times daily (QID) dosing. The addition of an H_2 antagonist such as cimetidine, ranitidine, or famotidine in conventional dosages may add benefit when H_1 antihistamines are inadequate. A $CysLT_1$ receptor antagonist such as montelukast, 10 mg daily, or zafirlukast, 20 mg twice a day, can be an important add-on therapy. For chronic urticaria which has failed to respond to a combination of long-acting H_1 antihistamines QID and a $CysLT_1$ receptor antagonist or cold urticaria, monoclonal anti-IgE antibodies such as omalizumab are now the next line of therapy. Older agents with antihistamine properties such as doxepin, cyproheptadine, and hydroxyzine have proven effective when H_1 antihistamines fail but are less effective than omalizumab and are sedating.

Topical glucocorticoids are of no value, and systemic glucocorticoids are generally avoided in idiopathic, allergen-induced, or physical urticarias due to their long-term toxicity. Systemic glucocorticoids are useful in the management of patients with pressure urticaria, vasculitic urticaria (especially with eosinophil prominence), idiopathic angioedema with or without urticaria, or chronic urticaria that responds poorly to conventional treatment and should be considered in any patient with debilitating disease. With persistent vasculitic urticaria, hydroxychloroquine, dapsone, or colchicine may be added to the regimen after hydroxyzine and before or along with systemic glucocorticoids. Cyclosporine is efficacious for patients with chronic idiopathic urticaria that is severe and poorly responsive to other modalities and/or where glucocorticoids are a requirement.

Infusion of isolated or recombinant C1INH protein is approved for prophylaxis of and acute HAE attacks while administration of a bradykinin 2 receptor antagonist (Icatibant) or a kallikrein inhibitor (Ecallantide) may be used for treatment of an acute attack of HAE. Older, less expensive preventative options include attenuated androgens, which stimulate production by the normal gene of an amount of functional C1INH sufficient to control the spontaneous activation of C1. The antifibrinolytic agent ε-aminocaproic acid may be used for preoperative prophylaxis, but is contraindicated in patients with thrombotic tendencies or ischemia due to arterial atherosclerosis. Fresh frozen plasma infusion can be used for acute attacks in a setting which lacks access to the newer agents. Bradykinin 2 receptor antagonist and C1INH protein are being studied for ACE inhibitor-induced angioedema. Treatment of the underlying hematologic malignancy is indicated for acquired C1INH deficiency.

ALLERGIC RHINITIS

DEFINITION

Rhinitis is characterized by sneezing; rhinorrhea; obstruction of the nasal passages; conjunctival, nasal, and pharyngeal itching; and lacrimation and can be classified as allergic or non-allergic. A clinical history of rhinitis symptoms occurring in a temporal relationship to allergen exposure and documentation of sensitization to an environmental allergen are required for a diagnosis of allergic rhinitis. Although commonly seasonal due to elicitation by airborne pollens, it can be perennial in an environment of chronic exposure to house dust mites, animal danders, or insect (cockroach) products. The overall prevalence in North America has increased in the past 20 years and is 10–30%, with the peak prevalence of >30% occurring in the fifth decade.

PREDISPOSING FACTORS AND ETIOLOGY

Allergic rhinitis generally occurs in atopic individuals, often in association with atopic dermatitis, food allergy, urticaria, and/or asthma (**Chap. 281**). Up to 50% of patients with allergic rhinitis manifest asthma, whereas 70–80% of individuals with asthma and 80% of individuals with chronic bilateral sinusitis experience allergic rhinitis. Female sex, particulate air pollution exposure, and maternal tobacco smoking increase the risk of developing allergic rhinitis over the life span.

Trees, grasses, and weeds that depend on wind rather than insects for pollination produce sufficient quantities of pollen suitable for wide distribution by air currents to elicit seasonal allergic rhinitis. The dates of pollination of these species historically varied little from year to year in a particular locale, but may be quite different in another climate. In the temperate areas of North America, trees typically pollinate from March through May, grasses in June and early July, and ragweed from mid-August to early October. Molds, which are widespread in nature because they occur in soil or decaying organic matter, propagate spores in a pattern that depends on climatic conditions. Climate change is impacting these patterns with early tree pollination and prolonged ragweed season with the delay of the first frost. Perennial allergic rhinitis occurs in response to allergens that are present throughout the year, including animal dander, cockroach-derived proteins, mold spores, or dust mites such as *Dermatophagoides farinae* and *Dermatophagoides pteronyssinus*. Dust mites are scavengers of human skin and excrete cysteine protease allergens in their feces. In up to 40% of patients with perennial rhinitis, no clear-cut allergen can be demonstrated as causative.

PATHOPHYSIOLOGY AND MANIFESTATIONS

Episodic rhinorrhea, sneezing, obstruction of the nasal passages with lacrimation, and pruritus of the conjunctiva, nasal mucosa, and oropharynx are the hallmarks of allergic rhinitis. The nasal mucosa is pale and boggy, the conjunctiva congested and edematous, and the pharynx generally unremarkable. Swelling of the turbinates and mucous membranes with obstruction of the sinus ostia and eustachian tubes precipitates secondary infections of the sinuses and middle ear, respectively. A growing number of patients with seasonal allergic rhinitis demonstrate pollen-associated food allergen syndrome characterized by oropharyngeal pruritus and/or mild swelling following the ingestion of raw plant-based foods which contain cross-reacting pollen-related allergens.

Nasal polyps, representing mucosal protrusions containing edema fluid with variable numbers of eosinophils and degranulated mast cells, can increase obstructive symptoms with anosmia as a defining feature and can concurrently arise within the nasopharynx or sinuses. Atopy is not a risk factor for nasal polyps, which instead may occur in the setting of cystic fibrosis, aspirin-exacerbated respiratory disease characterized by the triad of asthma, rhinosinusitis, and respiratory reactions to all cyclooxygenase-1 inhibitors, and in patients with chronic staphylococcal colonization, which produces superantigens leading to an intense T_H2 inflammatory response.

The nose presents a large mucosal surface area through the folds of the turbinates and serves to adjust the temperature and moisture content of inhaled air and to filter out particulate materials >10 μm in size by impingement in a mucous blanket; ciliary action moves the entrapped particles toward the pharynx. Entrapment of pollen and digestion of the outer coat by mucosal enzymes such as lysozymes release protein allergens. The initial interaction occurs between the allergen and intraepithelial mast cells and then proceeds to involve deeper perivenular mast cells, both of which are sensitized with specific IgE. During the symptomatic season when the mucosae are already swollen and hyperemic, there is enhanced adverse reactivity to the seasonal pollen as well as irritants such as tobacco smoke and fragrances. Biopsy specimens of nasal mucosa during seasonal rhinitis show submucosal edema with infiltration by eosinophils, along with some basophils and neutrophils.

The mucosal surface fluid contains IgA that is present because of its secretory piece and also IgE, which apparently arrives by diffusion from plasma cells in proximity to mucosal surfaces. IgE fixes to mucosal and submucosal mast cells, and the intensity of the clinical response to inhaled allergens is quantitatively related to the naturally occurring pollen dose. In sensitive individuals, the introduction of allergen into the nose is associated with sneezing, nasal obstruction, and discharge, and the fluid contains histamine, PGD_2, and leukotrienes. Thus the mast cells of the nasal mucosa and submucosa generate and release mediators through IgE-dependent reactions that are capable of producing tissue edema and eosinophilic infiltration.

DIAGNOSIS

The diagnosis of seasonal allergic rhinitis depends largely on an accurate history of occurrence coincident with the pollination of the offending weeds, grasses, or trees. The continuous character of perennial allergic rhinitis due to contamination of the home or place of work makes historic analysis difficult, but there may be variability in symptoms that can be related to exposure to animal dander, dust mite and/or cockroach allergens, fungal spores, or work-related allergens such as latex. Patients with perennial rhinitis commonly develop the problem in adult life, and manifest nasal congestion and a postnasal discharge, often associated with thickening of the sinus membranes demonstrated by radiography. Perennial nonallergic rhinitis with eosinophilia syndrome (NARES) occurs in the middle decades of life and is characterized by nasal obstruction, anosmia, chronic sinusitis, and prominent eosinophilic nasal discharge in the absence of allergen sensitization. The term *vasomotor rhinitis* or *perennial nonallergic rhinitis* designates a condition of enhanced reactivity of the nasopharynx in which a symptom complex resembling perennial allergic rhinitis occurs with nonspecific stimuli, including chemical odors, temperature and humidity variations, and position changes but occurs without tissue eosinophilia or an allergic etiology. Other entities to be excluded are structural abnormalities of the nasopharynx; exposure to irritants; gustatory rhinitis associated with cholinergic activation that occurs while eating or ingesting alcohol; hypothyroidism; upper respiratory tract infection; pregnancy with prominent nasal mucosal edema; prolonged topical use of α-adrenergic agents in the form of nasal sprays (rhinitis medicamentosa); and the use of certain systemic agents such as β-adrenergic antagonists, ACE inhibitors, direct vasodilators (hydralazine), α1-adrenergic receptor antagonists, estrogens, progesterone, NSAIDS, gabapentin, phosphodiesterase-5 inhibitors, and psychotropics (Risperidone, chlorpromazine, amitriptyline).

The nasal secretions of allergic patients are rich in eosinophils, and a modest peripheral eosinophilia can be observed. Local or systemic neutrophilia implies infection. Total serum IgE is frequently elevated, but the demonstration of immunologic specificity for IgE is critical to an etiologic diagnosis. A skin test by the intracutaneous route (puncture or prick) with the allergens of interest provides a rapid and reliable approach to identifying allergen-specific IgE that has sensitized cutaneous mast cells. A positive intracutaneous skin test with 1:10–1:20 weight/volume of extract has a high predictive value for the presence of allergy. An intradermal test with a 1:500–1:1000 dilution of 0.05 mL may follow if indicated by history when the intracutaneous test is negative, but while more sensitive, it is less reliable due to the reactivity of some asymptomatic individuals at the test dose.

Newer methodology for detecting total IgE, including the development of enzyme-linked immunosorbent assays (ELISA) employing anti-IgE bound to either a solid-phase or a liquid-phase particle, provides rapid and cost-effective determinations. Measurements of specific anti-IgE in serum are obtained by its binding to an allergen and quantitation by subsequent uptake of labeled anti-IgE. As compared to the skin test, the assay of specific IgE in serum is less sensitive but has high specificity.

TREATMENT

TREATMENT

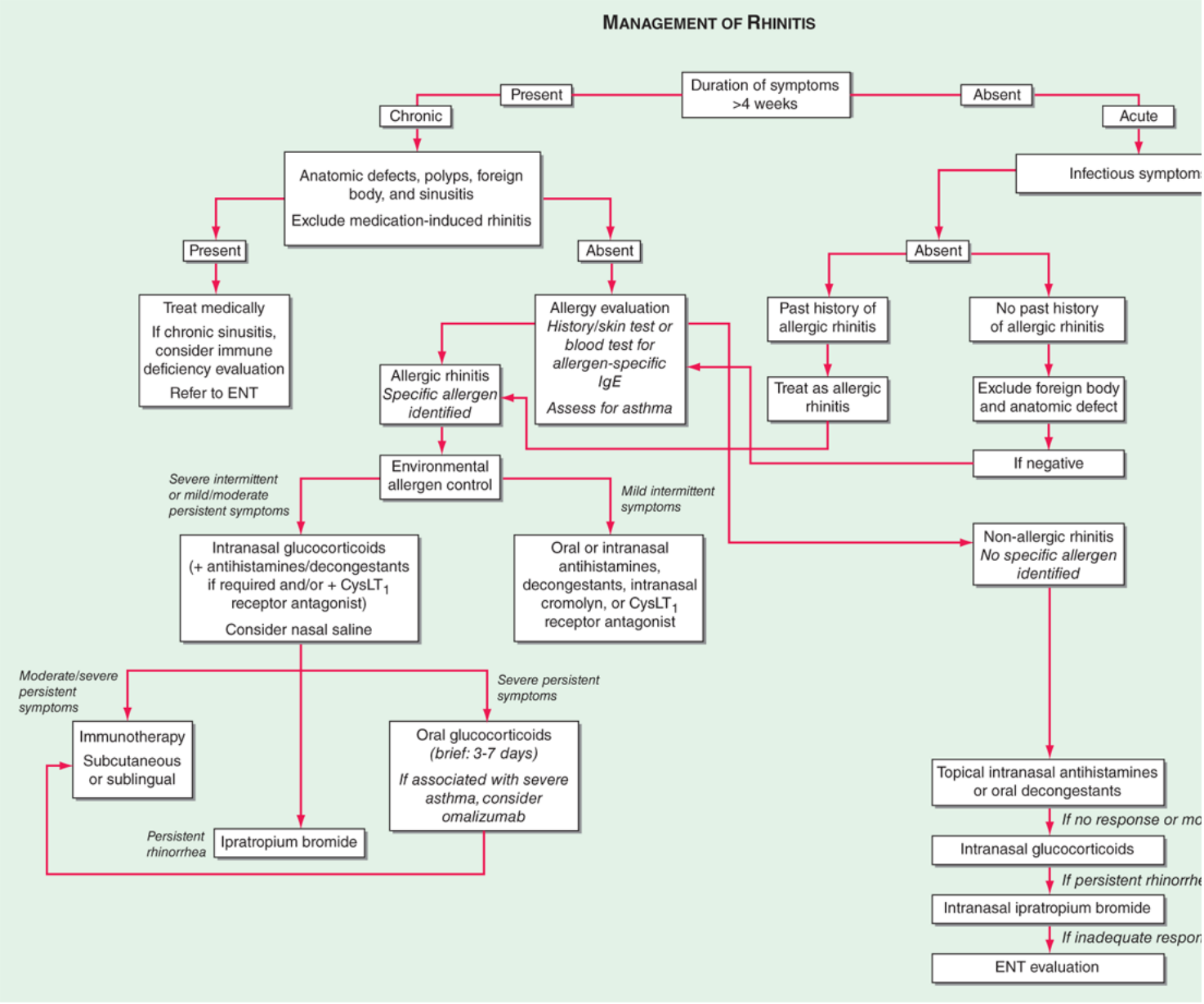
Allergic Rhinitis

Although allergen avoidance is the most cost-effective means of managing allergic rhinitis, only in the case of animal dander and possibly dust mites is it really feasible. Treatment with pharmacologic agents represents the standard approach to seasonal or perennial allergic rhinitis. Oral long-acting H_1 antihistamines are effective for nasopharyngeal itching, sneezing, and watery rhinorrhea and for such ocular manifestations as itching, tearing, and erythema, but they are less efficacious for the nasal congestion. The older antihistamines are sedating, and they induce psychomotor impairment, including reduced eye-hand coordination and impaired automobile driving skills. Their anticholinergic (muscarinic) effects include visual disturbance, urinary retention, and constipation. Because the newer H_1 antihistamines such as fexofenadine, loratadine, desloratadine, cetirizine, levocetirizine, olopatadine, bilastine, and azelastine are less lipophilic and more H_1 selective, their ability to cross the blood-brain barrier is reduced, and thus their sedating and anticholinergic side effects are minimized. These newer antihistamines do not differ appreciably in efficacy for relief of rhinitis and/or sneezing. Intranasal high-potency glucocorticoids are the most potent drugs available for the relief of established rhinitis, seasonal or perennial, and are effective in relieving nasal congestion as well as ocular symptoms. They provide efficacy with substantially reduced side effects as compared with this same class of agent administered orally. Their most frequent side effect is local irritation, with *Candida* overgrowth being a rare occurrence. The currently available intranasal glucocorticoids—beclomethasone, flunisolide, triamcinolone, budesonide, fluticasone propionate, fluticasone furoate, ciclosenide, and mometasone furoate—are equally effective for nasal symptom relief, including nasal congestion; these agents all achieve up to 70% overall symptom relief with some variation in the time period for onset of benefit. Azelastine nasal spray may benefit individuals with nonallergic vasomotor rhinitis as well as additive benefit to intranasal steroids in allergic rhinitis, but it has an adverse effect of dysgeusia (taste perversion) in some patients. Alternative nasal decongestants include α-adrenergic agents such as phenylephrine or oxymetazoline; however, the duration of their efficacy is limited because of rebound rhinitis (i.e., 7- to 14-day use can lead to rhinitis medicamentosa) and such systemic responses as hypertension. Oral α-adrenergic agonist decongestants containing pseudoephedrine are standard for the management of nasal congestion, generally in combination with an antihistamine. While oral antihistamines typically reduce nasal and ocular symptoms by about one-third, pseudoephedrine must be added to achieve a similar reduction in nasal congestion. These pseudoephedrine combination products can cause insomnia and are precluded from use in patients with narrow angle glaucoma, urinary retention, severe hypertension, marked coronary artery disease, or a first-trimester pregnancy. The $CysLT_1$ blocker montelukast is approved for treatment of both seasonal and perennial rhinitis, and it reduces both nasal and ocular symptoms by about 20%. Cromolyn sodium nasal spray inhibits mast cell degranulation, and can be used prophylactically on a continuous basis during the season. Topical ipratropium is an anticholinergic agent effective in reducing rhinorrhea, including that of patients with perennial non-allergic symptoms, and it can be additionally efficacious when combined with intranasal glucocorticoids. For concomitant allergic conjunctivitis, topical treatment with cromolyn sodium is effective in treating mild allergic symptoms and topical antihistamines such as olopatadine, azelastine, ketotifen, or epinastine administered to the eye provide rapid relief of itching and redness and are more effective than oral antihistamines.

Immunotherapy Immunotherapy consists of repeated exposure to gradually increasing concentrations of the allergen(s) considered to be specifically responsible for the symptom complex. Two forms of immunotherapy, subcutaneous (SCIT) and sublingual (SLIT), are currently available. Controlled studies of ragweed, grass, dust mite, and cat dander allergens administered via SCIT for treatment of allergic rhinitis have demonstrated improved symptom control over medications alone with the advantage of providing a durable benefit. The duration of SCIT is 3–5 years, with discontinuation being based on minimal symptoms over two consecutive seasons of exposure to the allergen. Clinical benefit appears related to the administration of a high dose of relevant allergen, advancing from weekly to monthly intervals. Patients should remain at the treatment site for at least 30 min after allergen administration so that any systemic reactions including anaphylaxis can be managed. Two to three percent of SCIT patients experience a systemic reaction over a 12-month period. Local reactions with erythema and induration are not uncommon and may persist for 1–3 days. SLIT is prepared as a tablet to be dissolved under the tongue at home after the first dose. The efficacy of SLIT is comparable to SCIT but only for the three allergens formulations available, dust mite, timothy/northern grasses and ragweed. Systemic reactions are less frequent with SLIT but transient oral pruritus is common. Immunotherapy is contraindicated in patients with significant cardiovascular disease or unstable asthma and should be conducted with particular caution in any patient requiring β-adrenergic blocking therapy because of the difficulty in managing an anaphylactic complication. The response to immunotherapy is associated with a complex of cellular and humoral effects that includes a modulation in T cell cytokine production and allergen-specific IgG_4 expansion. Immunotherapy should be reserved for clearly documented seasonal or perennial rhinitis that is clinically related to defined allergen exposure with confirmation by the presence of allergen-specific IgE through skin or in vitro specific IgE testing. Systemic treatment with a monoclonal antibody to IgE (omalizumab) that blocks mast cell and basophil sensitization has efficacy for allergic rhinitis and can be used with immunotherapy to enhance safety and efficacy. However, current approval is only for treatment of patients with persistent allergic asthma not controlled by inhaled glucocorticoid therapy. A sequence for the management of allergic or perennial rhinitis based on an allergen-specific diagnosis and stepwise management as required for symptom control would include the following: (1) identification of the offending allergen(s) by history with confirmation of the presence of allergen-specific IgE by skin test and/or serum assay; (2) avoidance of the offending allergen; and (3) medical management in a stepwise fashion **(Fig. 345-4)**. Mild intermittent symptoms of allergic rhinitis are treated with oral antihistamines, oral $CysLT_1$ receptor antagonists, intranasal antihistamines, or intranasal cromolyn prophylaxis. Moderate to more severe allergic rhinitis is managed with intranasal glucocorticoids plus oral antihistamines, oral $CysLT_1$ receptor antagonists, or antihistamine-decongestant combinations. Persistent or seasonal allergic rhinitis, rhinoconjunctivitis, or asthma which remains uncontrolled with maximal medical therapy merit consideration of allergen-specific immunotherapy.

FIGURE 345-4

Algorithm for the diagnosis and management of rhinitis. Persistent defined as >4 days/week for >4 weeks. Moderate/severe defined as abnormal sleep, impaired daily activities (school, work, sport, leisure) and/or troublesome symptoms. CysLT, cysteinyl leukotriene; ENT, ear, nose, and throat; IgE, immunoglobulin E.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Bernstein DI et al: Allergic rhinitis: Mechanisms and treatment. Immunol Allergy Clin North Am 36:261, 2016. [PubMed: 27083101]

Cho SH et al: Chronic rhinosinusitis without nasal polyps. J Allergy Clin Immunol Pract 4:575, 2016. [PubMed: 27393771]

Cicardi M et al: Classification, diagnosis, and approach to treatment for angioedema: Consensus report from the Hereditary Angioedema International Working Group. Allergy 69:602, 2014. [PubMed: 24673465]

Corren J et al: Allergic and nonallergic rhinitis, in *Middleton's Allergy: Principles and Practice*, 8th ed. NF Adkinson et al (eds). Philadelphia, Saunders, 2014, pp 664–685.

Jutel M et al: International consensus on allergen immunotherapy II: Mechanisms, standardization, and pharmacoeconomics. J Allergy Clin Immunol 137:358, 2016. [PubMed: 26853128]

Maurer M et al: Omalizumab for the treatment of chronic idiopathic or spontaneous urticaria. N Engl J Med 368:924, 2013. [PubMed: 23432142]

Saini SS: Urticaria and angioedema, in *Middleton's Allergy: Principles and Practice*, 8th ed. NF Adkinson et al (eds). Philadelphia, Saunders, 2014, pp 575–587.

Stevens WW et al: Chronic rhinosinusitis with nasal polyps. J Allergy Clin Immunol Pract 4:565, 2016. [PubMed: 27393770]

Chapter 346: Anaphylaxis

David Hong; Joshua A. Boyce

DEFINITION

Anaphylaxis is a potentially life-threatening systemic allergic reaction involving one or more organ systems that typically occurs within seconds to minutes of exposure to the anaphylactic trigger, most often a drug, food, or hymenoptera sting. Other triggers of anaphylaxis include radiocontrast administration or latex exposure. The term “anaphylaxis” was first described in 1902 by Charles Richet and Paul Portier who attempted to immunize dogs against sea anemone toxin in the same way Pasteur was able to vaccinate individuals against the smallpox virus. To their surprise, repeated administration of small, sub-lethal doses of sea anemone toxin reliably induced acute-onset death when re-administered 2–3 weeks after initial “vaccination” to the toxin. The phenomenon was termed ana (anti)-phylaxis (“protection or guarding”) because vaccination with anemone toxin resulted in the opposite intended immune effect. Charles Richet was awarded the Nobel Prize in Physiology or Medicine in 1913 for this work which led to further insights into hypersensitivity and mast cell biology.

Clinical Manifestations

While 80–90% of anaphylactic episodes are uniphasic, about 10–20% of cases are biphasic in which anaphylactic symptoms return about an hour or longer after resolution of initial symptoms. Anaphylactic reactions are particularly dangerous when hypotension or hypoxia occurs, leading potentially to cardiovascular collapse or respiratory failure, respectively. There may be upper or lower airway obstruction or both. Laryngeal edema may be experienced as a “lump” in the throat, hoarseness, or stridor, whereas bronchial obstruction is associated with a feeling of tightness in the chest and/or audible wheezing. Patients with underlying asthma are predisposed to severe involvement of the lower airways and increased mortality associated with anaphylaxis. In fatal cases with clinical bronchial obstruction, the lungs show marked hyperinflation on gross and microscopic examination. The microscopic findings in the bronchi, however, are limited to luminal secretions, peribronchial congestion, submucosal edema, and eosinophilic infiltration, and the acute emphysema is attributed to intractable bronchospasm that subsides with death. Angioedema resulting in death by mechanical obstruction occurs in the epiglottis and larynx; however, the process also is evident in the hypopharynx and to some extent in the trachea. On microscopic examination, there is wide separation of the collagen fibers and the glandular elements; vascular congestion and eosinophilic infiltration also are present. Patients dying of vascular collapse without antecedent hypoxia from respiratory insufficiency have visceral congestion with a presumptive loss of intravascular fluid volume. The associated

electrocardiographic abnormalities, with or without infarction, in some patients may reflect a primary cardiac event mediated by mast cells (which are prominent near the coronary vessels) or may be secondary to a critical reduction in blood volume.

Gastrointestinal manifestations represent another severe presentation of anaphylaxis, and include nausea, vomiting, crampy abdominal pain, and/or fecal incontinence. Angioedema of the bowel wall may also cause sufficient intravascular volume depletion to precipitate cardiovascular collapse.

Cutaneous manifestations are among the most common presentations of anaphylaxis (>90% of cases). Symptoms include urticarial eruptions, flushing with diffuse erythema, and/or a feeling of generalized warmth. Urticarial eruptions are intensely pruritic and may be localized or disseminated. They may coalesce to form giant hives but seldom persist beyond 48 h.

PREDISPOSING FACTORS AND ETIOLOGY

Because the most dangerous manifestations of anaphylaxis involve the cardiovascular and/or respiratory systems, preexisting asthma and underlying cardiovascular disease could lead to more rapid decompensation from anaphylaxis. Atopy is not generally thought to be a risk factor for anaphylaxis from drug reactions or hymenoptera stings, but is associated with radiocontrast sensitivity, exercise-induced anaphylaxis, idiopathic anaphylaxis, and allergy to foods or latex. Severe hymenoptera-induced anaphylaxis (generally with prominent hypotension) can be a presenting feature of underlying systemic mastocytosis. Hymenoptera allergy is also more likely in patients whose occupations (i.e., beekeepers, trash haulers, and landscape workers) place them in regular proximity to stinging insects. Most commonly, allergen-induced cross-linking of IgE-bound FcεRI receptors on mast cells and basophils initiates the signal transduction events leading to hypersensitivity syndromes including anaphylaxis. The generation of allergen-specific IgE is the end result of sensitization via the adaptive immune system. The mechanisms underlying sensitization are beyond the scope of this topic; however, environmental factors, innate immune responses, and cytokines are among the many variables leading to antigen-specific IgE production by B cells and plasma cells. IgE-mediated drug allergies are most common with antibiotics and certain chemotherapy drugs, though theoretically, they can occur with almost any medication. As is the case with environmental allergies, repeated exposure to the allergy-causing antigen is an important risk factor to keep in mind when evaluating patients with anaphylaxis. In the case of allergy to carboplatin, the incidence of hypersensitivity is 27% in patients who have had ≥7 lifetime infusions and as high as 46% in patients who have had ≥15 lifetime infusions. Similarly, patients with cystic fibrosis have a relatively high incidence of allergic reactions to IV antibiotics that they receive periodically to treat exacerbations of bronchiectasis. Drugs can also function as haptens that form immunogenic conjugates with host proteins. The conjugating hapten may be the parent compound, a nonenzymatically derived storage product, or a metabolite formed in the host. Recombinant biologics can also induce the formation of IgE against the proteins or against glycosylated structures that serve as immunogens. More recently, outbreaks of anaphylaxis to the EGFR antibody, cetuximab, were reported in association with elevated titers of serum IgE to alpha-1,3-galactose (alpha-gal), an oligosaccharide found in non-primate mammals. Cetuximab is derived from a mouse cell line expressing a

transferase that tags the Fab' portion of the cetuximab heavy chain with alpha-gal. Interestingly, patients with a history of multiple bites from *Amblyomma americanum* ticks commonly found in the Carolinas, Arkansas, and Tennessee are more likely to have anti-alpha-gal IgE as compared to control patients living outside those states. Such individuals who become sensitized to alpha-gal can develop episodes of delayed anaphylaxis to beef, lamb, and pork.

PATHOPHYSIOLOGY

Many of the important early mediators of anaphylaxis are derived from mast cells, basophils, and eosinophils. Mast cells and basophils contain preformed granules comprised of histamine, proteases (tryptase, chymase), proteoglycans (heparin, chondroitin sulfate), and TNF-α, which are rapidly released into surrounding tissue upon cell activation, a process known as degranulation. Mast cells, basophils, and eosinophils are also sources of arachidonic acid-derived products which include cysteinyl leukotrienes, prostaglandins, and platelet activating factor (PAF). Histamine release results in flushing, urticaria, pruritus, and, in high concentrations, hypotension and tachycardia. Cysteinyl leukotrienes and prostaglandin D_2 cause bronchoconstriction and increased microvascular permeability. Prostaglandin D_2 causes cutaneous flushing, and attracts eosinophils and basophils to the site of mast cell activation. Serum PAF levels correlate with anaphylaxis severity and are inversely proportional to the constitutive level of PAF acetylhydrolase, which is necessary for PAF inactivation. Tryptase and chymase can activate complement and coagulation pathways. Activation of these pathways results in production of the anaphylotoxins, C3a and C5a, and activation of the kallikrein-kinin system which regulates blood pressure and vascular permeability. The actions of these anaphylactic mediators are likely additive or synergistic at the target tissues.

Non-IgE-mediated reactions to certain drugs (which may occur upon the first exposure) can mimic the pathophysiology of IgE-dependent anaphylaxis due to a similar profile of mediators. For example, paclitaxel is a chemotherapy agent derived from yew tree bark and needles that requires polyethoxylated castor oil (Cremophor) to be solubilized into aqueous solution. Cremophor directly activates the complement cascade, resulting in complement-dependent induced histamine release from mast cells and basophils. A version of paclitaxel that is solubilized by being bound to albumin nanoparticles, Abraxane, has a far lower rate of hypersensitivity, especially for patients who have had infusion reactions to Cremophor-solubilized paclitaxel. Reactions to radiocontrast and vancomycin are other examples of non-IgE-mediated hypersensitivity. Opiates and NSAIDs are other drug categories that can have similar adverse reactions.

DIAGNOSIS

The diagnosis of an anaphylactic reaction depends primarily on a history revealing the onset of symptoms and signs within seconds to minutes after the putative trigger is encountered. An exception is delayed anaphylaxis to meats in alpha-gal sensitized patients. Every attempt to identify the specific cause or causes should be made so as to minimize the risk of recurrent anaphylaxis. If a particular drug or food is suspected, skin or serum specific IgE testing is useful to confirm clinical suspicions. If a specific trigger cannot be

identified, a workup of underlying atopic diatheses may be useful to identify risk factors that could play a potential contributory role. In the acute setting, laboratory biomarkers of mast cell degranulation may be useful to document the severity of an anaphylactic episode. The most obvious serum biomarker to assay, histamine, has an extremely short half-life with a measurable time-window that expires <1 h from the onset of anaphylaxis. A more practical and useful biomarker is serum tryptase which peaks 60–90 min after the onset of anaphylaxis and can be measured as long as 5 h after the onset of anaphylaxis. It may be useful to follow-up an elevated tryptase measurement in the acute setting with another measurement when the patient is clinically stable to establish a baseline reference. An elevated baseline tryptase level may warrant further workup for mastocytosis, especially if the presenting reaction occurred in the setting of hymenoptera sting.

TREATMENT

Early recognition of an anaphylactic reaction is mandatory since severe, even fatal, complications, can occur within minutes after symptoms first appear. The treatment of first choice is intramuscular administration of 0.3–0.5 mL of 1:1000 (1 mg/mL) epinephrine, with repeated doses at 5–20 min intervals as needed for a severe reaction. The failure to use epinephrine within the first 20 min of symptoms is a risk factor for poor clinical outcomes in various studies of anaphylaxis. Another important variable that may affect anaphylaxis survival is body posture, as an upright or sitting posture may lead to the “empty heart syndrome” in which there is insufficient venous return to the heart from sudden onset hypotension secondary to intravascular volume depletion. Epinephrine can further accelerate empty heart syndrome due to its chronotropic effects. For this reason, it is recommended that patients who suffer from anaphylaxis be placed in the supine position before receiving epinephrine. IV fluids and vasopressor agents may be administered in the acute medical setting if intractable hypotension occurs. Epinephrine provides both α- and β-adrenergic effects, resulting in vasoconstriction, bronchial smooth-muscle relaxation, and attenuation of enhanced venular permeability. Beta blockers may attenuate this response; therefore, an alternative anti-hypertensive may be considered in patients at high risk of needing emergency epinephrine. Oxygen alone via a nasal catheter or with nebulized albuterol may be helpful; however, either endotracheal intubation or a tracheostomy is mandatory for oxygen delivery if progressive hypoxia develops. Ancillary agents such as antihistamines, glucocorticoids, and bronchodilators are also useful therapeutics to treat urticaria/angioedema and bronchospasm once the patient is hemodynamically stable.

PREVENTION

Avoidance

The simplest, most straightforward approach to the long-term management of a patient with a history of anaphylaxis is strict avoidance of known anaphylactic triggers and education on acute management, that is, instructing the patient on the proper use and indications for use of self-administered epinephrine. Lifelong avoidance is not easy if the trigger is an occupational exposure, hymenoptera sting, a common food (i.e.,

peanut), or a drug representing the sole or best therapeutic option for the patient. Special management options may exist for these patients.

Specific Immunotherapy

Patients with large local reactions to hymenoptera stings are unlikely to have anaphylaxis with subsequent stings; however, patients of any age who have had documented anaphylaxis should be formally evaluated and started on venom immunotherapy (VIT) if skin or serologic IgE testing confirms the history. Immunotherapy is a means of "tolerizing" patients to allergen by means of serial subcutaneous administration of escalating doses of extract containing relevant allergen until a target maintenance dose is achieved. As in the case of Richet's unfortunate dogs, anaphylaxis can sometimes occur during the course of administering immunotherapy extracts, so formulating extracts and administering them is typically done under the care of a specialist familiar with this type of treatment. In the case of hymenoptera allergy, patients receive VIT extracts containing actual hymenoptera venom with a maintenance dose equivalent to 2–5 stings. The recommended duration of treatment is 3–5 years; however, patients who have experienced severe respiratory of cardiovascular anaphylaxis are often on lifelong therapy.

Tolerance Induction

IgE sensitization to foods occurs most frequently in infants and young children, especially those with atopic dermatitis, and is a risk factor for anaphylaxis (although detection of specific IgE through skin or serum testing has relatively poor predictive value). While most allergy to egg, milk, soy, and/or wheat resolves spontaneously during childhood, ~80% of children with peanut allergy remain sensitive for life. A sharp rise in the prevalence of peanut allergy was also observed in the late 1990s–early 2000s especially in countries with Western diets where the average age of peanut introduction was age ≥3 years. Curiously, in cultures where peanut was introduced much earlier into children's diets, the prevalence of peanut allergy remained low. The landmark "Learning Early About Peanut Allergy" (LEAP) study demonstrated that early introduction of peanut protein to the diet of high risk infants (4–11 months of age with atopic dermatitis and/or egg allergy) can prevent the development of most (80% or more) peanut allergy compared with children who did not consume peanuts (avoidance group), even when IgE sensitization (based on positive skin test) had already developed at the time of study entry. While the induction of tolerance at an early age seems to be key to preventing clinical reactivity later in life, it is not yet clear if this principle holds true for other foods commonly associated with hypersensitivity reactions.

Desensitization

For patients who have suffered anaphylaxis from drug allergy and whose treatment regimen requires the administration of the offending drug, desensitization may be a short-term treatment option to prevent reactions. Desensitization elicits a temporary state of tolerance to the drug in sensitized, clinically reactive patients. While it has been a proven technique for penicillin-allergic patients for decades, desensitization has more recently been proven to be effective for certain chemotherapy agents, especially platin-based chemotherapy agents which can induce IgE-mediated sensitization with repeated exposures. The exact

mechanisms underlying desensitization are not fully understood; however, temporary tolerance can be achieved through the serial administration of gradually escalating doses of drug, starting from extremely low doses, over the course of hours. So long as the patient continues to receive the drug in question at regular intervals based on drug half-life, a “desensitized” state can also be maintained until the drug is no longer needed. Drug desensitization works best for IgE-mediated reactions; however, it has been performed in cases of non-IgE-mediated anaphylaxis from Cremophor-solubilized paclitaxel as described earlier in this chapter. Other non-IgE-mediated anaphylactic reactions can often be prevented with premedication regimens. A typical premedication regimen for radiocontrast, for example, will have the patient receive prednisone 0.5 mg/kg at 13, 6, and 1 h prior to contrast administration. Diphenhydramine 25 mg is also given 1 h prior to contrast. Flushing reactions from vancomycin can often be alleviated with antihistamine premedication and down titrating the infusion rate.

FURTHER READING

Castells MC et al: Hypersensitivity reactions to chemotherapy: Outcomes and safety of rapid desensitization in 413 cases. J Allergy Clin Immunol 122:574, 2008. [PubMed: 18502492]

Chung CH et al: Cetuximab-induced anaphylaxis and IgE specific for galactose-alpha-1,3-galactose. N Engl J Med 358:1109, 2008. [PubMed: 18337601]

Du Toit G et al: LEAP Study Team. Randomized trial of peanut consumption in infants at risk for peanut allergy. N Engl J Med 373:803, 2015.

Du Toit G et al: Immune Tolerance Network LEAP-On Study Team. Effect of avoidance on peanut allergy after early peanut consumption. N Engl J Med 374:1435, 2016. [PubMed: 26942922]

Lieberman P et al: Anaphylaxis—A practice parameter update 2015. Ann Allergy Asthma Immunol 115:341, 2015. [PubMed: 26505932]

Chapter 347: Mastocytosis

Cem Akin; Joshua A. Boyce

DEFINITION

Mastocytosis is defined by accumulation of clonally expanded mast cells in tissues such as skin, bone marrow, liver, spleen, and gut. The mast cell expansion is generally recognized in skin and/or bone marrow. Mastocytosis occurs at any age and has a slight preponderance in males. Mastocytosis is a rare disorder and its exact prevalence is not known; however, it is estimated to occur in ~1 in 20,000 people. Familial occurrence is rare, and atopy is not increased compared to the general population.

CLASSIFICATION AND PATHOPHYSIOLOGY

A consensus classification for mastocytosis recognizes cutaneous mastocytosis with variants, five systemic forms, and rare mast cell sarcoma **(Table 347-1)**.

TABLE 347-1

Classification of Mastocytosis

Cutaneous mastocytosis (CM)
Maculopapular cutaneous mastocytosis (MPCM)
Solitary mastocytoma of skin
Diffuse cutaneous mastocytosis
Indolent systemic mastocytosis (ISM)
Smoldering systemic mastocytosis
Systemic mastocytosis with an associated clonal hematologic non–mast cell lineage disease (SM-AHNMD)
Aggressive systemic mastocytosis (ASM)
Mast cell leukemia (MCL)
Mast cell sarcoma (MCS)

Source: Modified from H-P Horny et al: Mastocytosis. In: *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues,* revised 4th ed. SH Swerdlow et al (eds). Lyon, France, IARC Press, 2017, pp 61–69.

Cutaneous mastocytosis is the most common diagnosis of mastocytosis in children and indicates disease limited to skin with absence of pathologic infiltrates in internal organs. It is usually diagnosed within the first year of life with demonstration of fixed, maculopapular, and hyperpigmented lesions (maculopapular cutaneous mastocytosis [MPCM], formerly known as urticaria pigmentosa), mastocytoma(s) or diffuse cutaneous mastocytosis. Systemic mastocytosis (SM) refers to involvement of a non-cutaneous site (usually bone marrow). There are five distinct variants of SM; the form designated as *indolent systemic mastocytosis* (ISM) accounts for the majority of adult patients. ISM is diagnosed when there is no evidence of an associated hematologic disorder, mast cell leukemia or tissue dysfunction due to mast cell infiltration and is not known to alter life expectancy. Systemic smoldering mastocytosis (formerly considered a subvariant of ISM) is characterized by high mast cell burden as evidenced by a bone marrow infiltration of >30% and a baseline serum tryptase >200 ng/ml (B-findings), but absence of SM-AHNMD or ASM **(Table 347-2)**. In *systemic mastocytosis associated with clonal hematologic non–mast cell lineage disease* (SM-AHNMD, or SM-AHN for short), the prognosis is determined by the nature of the associated disorder, which can range from dysmyelopoiesis to leukemias usually of myeloid origin. In *aggressive systemic mastocytosis* (ASM), mast cell infiltration/proliferation in multiple organs such as liver, spleen, gut, bone, and bone marrow resulting in 1 or more C findings and a poor prognosis (Table 347-2). *Mast cell leukemia* (MCL) is the rarest form of SM and is invariably fatal at present; the peripheral blood contains circulating, metachromatically staining, and atypical mast cells. An aleukemic form of MCL is recognized without circulating mast cells when the percentage of high-grade immature mast cells in bone marrow smears exceeds 20% in a nonspicular area. Mast cell sarcoma is a rare solid mast cell tumor with malignant invasive features.

TABLE 347-2

B and C Findings for Diagnosis of SSM and ASM

B-Findings (2 or more in the absence of any C findings are required for a diagnosis of SSM): 1. MC infiltration in bone marrow biopsy of >30% and the basal serum tryptase level >200 ng/mL 2. Hypercellular bone marrow with signs of dysmyelopoiesis but without cytopenias meeting C criteria or WHO criteria for an MDS or MPN 3. Palpable hepatomegaly, palpable splenomegaly, or lymphadenopathy (on CT or ultrasound: >2 cm) without impaired liver function or hypersplenism
C-Findings: (1 or more required for a diagnosis of ASM). C finding should be reasonably attributable to high tissue mast cell infiltration. 1. Cytopenia(s): ANC<1,000/μL or Hb<10 g/dL or PLT<100,000/μL 2. Hepatomegaly with ascites and impaired liver function 3. Palpable splenomegaly with associated hypersplenism 4. Malabsorption with hypoalbuminemia and weight loss 5. Skeletal lesions: large area(s) of osteolyses with pathologic fractures (presence of osteoporosis alone without osteolytic lesions does not satisfy this criterion)

Abbreviations: ANC, absolute neutrophil count; ASM, aggressive systemic mastocytosis; CT, computed tomography; Hb, hemoglobin; MC, mast cells; MDS, myelodysplastic syndromes; MPN, myeloproliferative disorders; PLT, platelets; SSM, smoldering systemic mastocytosis; WHO, World Health Organization.

A point mutation of A to T at codon 816 of *KIT* that causes an aspartic acid to valine substitution, resulting in a somatic gain-of-function mutation, is found in mast cells and sometimes in multiple other cell lineages in patients with mastocytosis. This substitution, as well as other rare mutations of *KIT*, is characteristic of patients with all forms of SM, but is also present in some children with cutaneous mastocytosis in lesional skin, as might be anticipated because mast cells are of bone marrow lineage. Additional mutations in genes such as *TET2, SRSF2, ASLX1*, and *RUNX1* known to be associated with other hematologic neoplastic disorders can be detected in patients usually with advanced (non-ISM) forms of SM. The prognosis for patients with cutaneous mastocytosis and for almost all patients with ISM is a normal life expectancy, whereas that for patients with SM-AHNMD is determined by the non–mast cell component. ASM and MCL carry a poorer prognosis, while patients with SSM carry an intermediate prognosis. Progression from ISM to a more advanced form is rare (approximately 3% overall); however, patients should be monitored for emergence of hematologic disease and end organ manifestations of ASM. In infants and children with cutaneous manifestations, namely, maculopapular cutaneous mastocytosis, mastocytoma(s), or bullous lesions, visceral involvement is usually lacking, and spontaneous resolution is common prior to adolescence. Progression from CM to ISM may occur in ~10% of children, especially in those with high mast cell burden

(diffuse cutaneous mastocytosis), hematologic abnormalities and those who present with smaller uniform lesions with diameters measuring <2 cm.

CLINICAL MANIFESTATIONS

The clinical manifestations of SM, distinct from a leukemic complication, are due to the release of bioactive substances acting at both local and distal sites, tissue occupancy by the mast cell mass, and the tissue response to that mass. The pharmacologically induced manifestations are intermittent flushing, tachycardia and vascular collapse, gastric distress, lower abdominal crampy pain, and diarrhea. The increase in local cell burden is evidenced by the lesions of MPCM (urticaria pigmentosa) at skin sites and internal organ biopsies such as bone marrow and gastrointestinal tract and may be a direct local cause of bone pain and/or malabsorption. Mast cell–mediated fibrotic changes may occur in liver, spleen, and bone marrow but not in gastrointestinal tissue or skin. Immunofluorescent analysis of bone marrow and skin lesions in ISM and of spleen, lymph node, and skin in ASM has revealed only one mast cell phenotype, namely, scroll-poor cells expressing tryptase, chymase, and CPA.

The cutaneous lesions of MPCM (formerly known as urticaria pigmentosa) are reddish-brown macules, papules, or plaques that respond to trauma with urtication and erythema (Darier's sign). Children with CM may present with MPCM, mastocytomas, or diffuse cutaneous mastocytosis (DCM). Mastocytomas are generally solitary elevated lesions that are yellow, brown, or red in color. Their size may vary from a few millimeters to several centimeters. Rubbing or irritation of the mastocytoma lesion may lead to systemic symptoms such as flushing and urticaria. Children with DCM present without distinct lesions, but rather a generalized thickening of skin (pachydermia) due to diffuse mast cell infiltration. DCM may be associated with bullae formation and more severe systemic symptoms including upper GI irritation and vascular collapse in the first few years of life. Maculopapular skin lesions of mastocytosis may be present in patients with adult-onset systemic disease. The apparent incidence of cutaneous lesions is ≥80% in patients with ISM and <50% in those with SM-AHNMD or ASM. In the upper gastrointestinal tract, gastritis and peptic ulcer are significant problems. In the lower intestinal tract, the occurrence of diarrhea and abdominal pain is attributed to increased motility due to mast cell mediators; this problem can be aggravated by malabsorption, which can also cause secondary nutritional insufficiency and osteomalacia. The periportal fibrosis associated with mast cell infiltration and a prominence of eosinophils may lead to portal hypertension and ascites. In some patients, anaphylaxis with rapid and life threatening vascular collapse can be induced by hymenoptera stings. These patients often have evidence of venom specific IgE. The neuropsychiatric disturbances are clinically most evident as impaired recent memory, decreased attention span, and "migraine-like" headaches. Patients may experience exacerbation of a specific clinical sign or symptom variably with alcohol ingestion, temperature changes, stress, use of mast cell–interactive opioids, or ingestion of NSAIDs.

DIAGNOSIS

Cutaneous mastocytosis is diagnosed by observing the characteristic lesions of MPCM or mastocytoma(s). A skin biopsy can be obtained to confirm these subvariants of CM, whereas patients with suspected DCM and bullous mastocytosis usually require a skin biopsy to confirm the diagnosis. Although the diagnosis of SM is generally suspected on the basis of the clinical history and physical findings, and can be supported by laboratory procedures, it can be established only by a tissue diagnosis. By convention, the diagnosis of SM depends heavily on bone marrow biopsy to meet the criteria of one major plus one minor or three minor findings **(Table 347-3)**. The bone marrow provides the major criterion by revealing aggregates of mast cells, often in paratrabecular and perivascular locations with lymphocytes and eosinophils, as well as the minor criteria of abnormal mast cell morphology, aberrant mast cell membrane immunophenotype, or a codon 816 mutation in an extracutaneous tissue. A basal serum total tryptase level is a noninvasive approach to consider before bone marrow biopsy. The pro-β and α forms of tryptase are elevated in more than one-half of patients with SM and provide a minor criterion; the fully processed ("mature") β form is increased in patients undergoing an anaphylactic reaction. A rare histolopathologic subvariant called "well differentiated systemic mastocytosis" (WDSM) is characterized by clusters of mature appearing fully granulated and round mast cells, lack of aberrant CD25 and CD2 expression, and lack of D816V KIT mutation in most patients. These patients often have a history of childhood onset cutaneous disease and their mast cells may display aberrant CD30 expression and other markers of clonality such as atypical (non-D816V) KIT mutations. Additional studies directed by the presentation include a bone densitometry, bone scan, or skeletal survey; computed tomography scan, or endoscopy; and a neuropsychiatric evaluation. Osteoporosis is increased in mastocytosis and may lead to pathologic fractures.

TABLE 347-3

Diagnostic Criteria for Systemic Mastocytosis[a]

Major: Multifocal dense infiltrates of mast cells (>15 mast cells per aggregate) in bone marrow or other extracutaneous tissues
Minor: Abnormal mast cell morphology (spindle shape, bi- or multi-lobed or eccentric nucleus, hypogranulated cytoplasm) Aberrant mast cell surface phenotype with expression of CD25 (IL-2 receptor alpha chain) and/or CD2 Detection of codon 816 mutation in peripheral blood cells, bone marrow cells, or an extracutaneous lesional tissue Total serum tryptase >20 ng/mL

[a]Diagnosis requires either the major criterion and one minor criterion or three minor criteria.

Some patients presenting with recurrent mast cell activation symptoms (particularly hypotensive syncopal anaphylactic episodes) have been found to have underlying mastocytosis. A subset of these patients may be

found to have the D816V *KIT* mutation or aberrant mast cells displaying CD25, but lack other diagnostic criteria for SM. Such patients are termed to have “monoclonal mast cell activation syndrome.”

The differential diagnosis requires the exclusion of other flushing disorders. The 24-h urine assessment of 5-hydroxy-indoleacetic acid and metanephrines should exclude a carcinoid tumor or a pheochromocytoma, respectively. Some patients presenting with recurrent mast cell activation symptoms without an obvious increase in mast cell burden in skin or bone marrow have been shown to carry aberrant mast cells with clonality markers of D816V *KIT* mutation or surface CD25 expression. Most patients with recurrent IgE-induced or idiopathic anaphylaxis present with urticaria, angioedema, and/or wheezing, which are not manifestations of SM.

TREATMENT

TREATMENT

Mastocytosis

The management of SM uses a stepwise and symptom/sign–directed approach that includes an H_1 antihistamine for flushing and pruritus, an H_2 antihistamine or proton pump inhibitor for gastric acid hypersecretion, oral cromolyn sodium for diarrhea and abdominal pain, and occasionally aspirin (in those who are known to be tolerant of NSAIDs) for severe flushing with or without associated vascular collapse, despite use of H_1 and H_2 antihistamines, to block biosynthesis of PGD_2. Systemic glucocorticoids appear to alleviate the malabsorption. Mast cell cytoreductive therapy consisting of midostaurin, IFN-α or cladribine is generally reserved for advanced, nonindolent variants of SM. Midostaurin is a multi-kinase inhibitor with activity against D816V mutated and wild type KIT, and was recently approved by Food and Drug Administration for treatment of advanced systemic mastocytosis (SM-AHNMD, ASM, and MCL), and should be considered as a first-line therapy of these disease variants. The efficacy of cytoreductive therapy in mastocytosis is variable, perhaps because of dosage limitations due to side effects. Imatinib is not effective in most cases as D816V *KIT* mutation provides resistance against it. Combination chemotherapy is appropriate for the frank leukemias. Stem cell transplantation has shown to be effective in a small subset of patients with advanced mastocytosis. A self-injectable epinephrine prescription is recommended for most patients due to increased incidence of anaphylaxis. Patients with a history of systemic hymenoptera venom reaction should be tested for venom specific IgE and placed on lifelong venom immunotherapy. Other investigational tyrosine kinase inhibitors with a capacity to inhibit D816V *KIT* mutation are currently in clinical trials.

FURTHER READING

Arock M et al: Current treatment options in patients with mastocytosis: Status in 2015 and future perspectives. Eur J Haematol 94:474, 2014.

Gotlib J et al: Efficacy and safety of midostaurin in advanced systemic mastocytosis. N Engl J Med 374:2530, 2016.

Hartmann K et al: Cutaneous manifestations in patients with mastocytosis: Consensus report of the European Competence Network on Mastocytosis; the American Academy of Allergy, Asthma & Immunology; and the European Academy of Allergology and Clinical Immunology. J Allergy Clin Immunol 137:35, 2016. [PubMed: 26476479]

Horny H-P et al: Mastocytosis (mast cell disease). In: *WHO Classification of Tumours. Pathology & Genetics. Tumours of Haematopoietic and Lymphoid Tissues*. SH Swerdlow, et al (eds). Lyon, France, IARC Press, 2008, pp 54–63.

Theoharides TC et al: Mast cells, mastocytosis, and related disorders. N Engl J Med 373:163, 2015. [PubMed: 26154789]

Ustun C et al: Consensus opinion on allogeneic hematopoietic cell transplantation in advanced systemic Mastocytosis. Biol Blood Marrow Transplant 22:1348, 2016. [PubMed: 27131865]

Valent P et al: Standards and standardization in mastocytosis: Consensus statements on diagnostics, treatment recommendations and response criteria. Eur J Clin Invest 37:435, 2007. [PubMed: 17537151]

Valent P et al: European Competence Network on Mastocytosis. Proposed diagnostic algorithm for patients with suspected mastocytosis: A proposal of the European Competence Network on Mastocytosis. Allergy 69:1267, 2014. [PubMed: 24836395]

Chapter 348: Autoimmunity and Autoimmune Diseases

Betty Diamond; Peter E. Lipsky

INTRODUCTION

One of the central features of the immune system is the capacity to mount an inflammatory response to potentially harmful foreign materials while avoiding damage to self-tissues. Whereas recognition of self plays an important role in shaping the repertoires of immune receptors on both T and B cells and in clearing apoptotic and other tissue debris from sites throughout the body, the development of potentially harmful immune responses to self-antigens is, in general, prohibited. The essential feature of an *autoimmune disease* is that tissue injury is caused by the immunologic reaction of the organism against its own tissues. *Autoimmunity*, on the other hand, refers merely to the presence of antibodies or T lymphocytes that react with self-antigens and does not necessarily imply that the self-reactivity has pathogenic consequences. Autoimmunity is present in all individuals and increases with age; however, autoimmune disease occurs only in those individuals in whom the breakdown of one or more of the basic mechanisms regulating immune tolerance results in self-reactivity that can cause tissue damage.

Polyreactive autoantibodies that recognize many host antigens are present throughout life. These antibodies are usually of the IgM heavy chain isotype and are encoded by nonmutated germline immunoglobulin variable region genes. These antibodies are essential, as they remove apoptotic debris through non inflammatory pathways. Expression of these autoantibodies may be increased after some inciting events. When autoimmunity is induced by an inciting event, such as infection or tissue damage from trauma or ischemia, the autoreactivity is in general self-limited. When such autoimmunity does persist, however, pathology may or may not result. Even in the presence of organ pathology, it may be difficult to determine whether the damage is mediated by autoreactivity. After an inciting event, the development of self-reactivity may be the consequence of an ongoing pathologic process, may be nonpathogenic, or may exacerbate tissue inflammation and damage. Individuals with autoimmune disease may have numerous autoantibodies, only some or even none of which may be pathogenic. For example, patients with systemic sclerosis may have a wide array of antinuclear antibodies that are important in disease classification but are not clearly pathogenic; in contrast, patients with pemphigus may also exhibit a wide array of autoantibodies, one of which (antibody to desmoglein 1 and 3) is known to be pathogenic.

MECHANISMS OF AUTOIMMUNITY

Since Ehrlich first postulated the existence of mechanisms to prevent the generation of self-reactivity in the early 1900s, there has been a progressive increase in understanding of this prohibition in parallel with a progressive increase in understanding of the immune system. Burnet's clonal selection theory included the idea that interaction of lymphoid cells with their specific antigens during fetal or early postnatal life would lead to elimination of such "forbidden clones." This idea was refuted, however, when it was shown that autoimmune diseases could be induced in experimental animals by simple immunization procedures, that autoantigen-binding cells could be demonstrated easily in the circulation of normal individuals, and that self-limited autoimmune phenomena frequently developed after tissue damage from infection or trauma. These observations indicated that clones of cells capable of responding to autoantigens were present in the repertoire of antigen-reactive cells in normal adults and suggested that mechanisms in addition to clonal deletion were responsible for preventing their activation.

Currently, three general processes are thought to be involved in the maintenance of selective unresponsiveness to autoantigens **(Table 348-1)**: (1) sequestration of self-antigens, rendering them inaccessible to the immune system; (2) specific unresponsiveness (tolerance or anergy) of relevant T or B cells; and (3) limitation of potential reactivity by regulatory mechanisms. Derangements of these normal processes may predispose to the development of autoimmunity **(Table 348-2)**. In general, these abnormal responses require both an exogenous trigger, such as bacterial or viral infection or cigarette smoking, and the presence of endogenous abnormalities in the cells of the immune system. Microbial superantigens, such as staphylococcal protein A and staphylococcal enterotoxins, are substances that can stimulate a broad range of T and B cells through specific interactions with selected families of immune receptors, irrespective of their antigen specificity. If autoantigen-reactive T and/or B cells express these receptors, autoimmunity may develop. Alternatively, molecular mimicry or cross-reactivity between a microbial product and a self-antigen may lead to activation of autoreactive lymphocytes. One of the best examples of autoreactivity and autoimmune disease resulting from molecular mimicry is rheumatic fever, in which antibodies to the M protein of streptococci cross-react with myosin, laminin, and other matrix proteins as well as with neuronal antigens. Deposition of these autoantibodies in the heart initiates an inflammatory response, whereas their penetration into the brain can result in Sydenham's chorea. Molecular mimicry between microbial proteins and host tissues has been reported in type 1 diabetes mellitus, rheumatoid arthritis, systemic lupus erythematosus (SLE), celiac disease, and multiple sclerosis. It is presumed that infectious agents may be able to overcome self-tolerance because they possess *pathogen-associated molecular patterns* (PAMPs). These molecules (e.g., bacterial endotoxin, RNA, or DNA) exert adjuvant-like effects on the immune system by interacting with *Toll-like receptors* (TLRs) and other *pattern recognition receptors* (PRRs) that increase the immunogenicity and immunostimulatory capacity of the microbial material. The adjuvants activate dendritic cells, which in turn stimulate the activation of previously quiescent lymphocytes that recognize both microbial antigens and self-antigens. Similarly, cellular and tissue damage due to the release of *damage-associated molecular patterns* (DAMPs), including DNA, RNA nucleosomes, and other tissue debris, may activate cells of the inflammatory and immune systems through engagement of the same array of PRRs. This pathway may lead to autoimmune disease in individuals who have impairments in mechanisms for clearance of tissue debris.

TABLE 348-1

Mechanisms Preventing Autoimmunity

1. Sequestration of self-antigens
2. Generation and maintenance of tolerance
    a. Central deletion of autoreactive lymphocytes
    b. Peripheral anergy of autoreactive lymphocytes
    c. Receptor replacement in autoreactive lymphocytes
3. Regulatory mechanisms
    a. Regulatory T cells
    b. Regulatory B cells
    c. Regulatory mesenchymal cells
    d. Regulatory cytokines
    e. Idiotype network

TABLE 348-2

Mechanisms of Autoimmunity

- I. Exogenous
  - A. Molecular mimicry
  - B. Superantigenic stimulation
  - C. Microbial and tissue damage–associated adjuvanticity
- II. Endogenous
  - A. Altered antigen presentation
    - 1. Loss of immunologic privilege
    - 2. Presentation of novel or cryptic epitopes (epitope spreading)
    - 3. Alteration of self-antigen
    - 4. Enhanced function of antigen-presenting cells
      - a. Costimulatory molecule expression
      - b. Cytokine production
  - B. Increased T cell help
    - 1. Cytokine production
    - 2. Costimulatory molecules
  - C. Increased B cell function
    - 1. B cell activating factor
    - 2. Costimulatory molecules
  - D. Apoptotic defects or defects in clearance of apoptotic material
  - E. Cytokine imbalance
  - F. Altered immunoregulation

Although previous work focused on the role of pathogenic microorganisms in triggering autoimmunity, more recent studies have focus on the role of the microbiome, the collection of non-pathogenic microorganisms that reside on various body surfaces. It has become clear that the interaction between specific constituents of these microbiota and the immune system can shape the nature of the immune response to either favor or discourage immune/inflammatory responses. Thus, some genera within the microbiome may favor a nonresponsive state dominated by regulatory T cells whereas others may favor the development of T effector cells and a proinflammatory state. Gender bias in autoimmune conditions may also be favored by differences in the dominant organisms within the microbiome.

Endogenous derangements of the immune system may also contribute to the loss of immunologic tolerance to self-antigens and the development of autoimmunity (Table 348-2). Some autoantigens reside in immunologically privileged sites, such as the brain or the anterior chamber of the eye. These sites are characterized by the inability of engrafted tissue to elicit immune responses. Immunologic privilege results from a number of events, including the limited entry of proteins from those sites into lymphatics, the local

production of immunosuppressive cytokines such as transforming growth factor β, and the local expression of molecules (including Fas ligand) that can induce apoptosis of activated T cells. Lymphoid cells remain in a state of immunologic ignorance (neither activated nor anergized) with regard to proteins expressed uniquely in immunologically privileged sites. If the privileged site is damaged by trauma or inflammation or if T cells are activated elsewhere, proteins expressed at this site can become immunogenic and also be the targets of immunologic assault. In multiple sclerosis and sympathetic ophthalmia, for example, antigens uniquely expressed in the brain and eye, respectively, become the target of activated T cells.

Alterations in antigen presentation may also contribute to autoimmunity. Peptide determinants (*epitopes*) of a self-antigen that are not routinely presented to lymphocytes may be recognized as a result of altered proteolytic processing of the molecule and the ensuing presentation of novel peptides (*cryptic epitopes*). When B cells rather than dendritic cells present self-antigen, they may also present cryptic epitopes that can activate autoreactive T cells. These cryptic epitopes will not previously have been available to effect the silencing of autoreactive lymphocytes. Furthermore, once there is immunologic recognition of one protein component of a multimolecular complex, reactivity may be induced to other components of the complex after internalization and presentation of all molecules within the complex (epitope spreading). Finally, inflammation, environmental agents, drug exposure, or normal senescence may cause a post-translational alteration in proteins, resulting in the generation of immune responses that cross-react with normal self-proteins. For example, the induction and/or release of protein arginine deiminase enzymes results in the conversion of arginine residues to citrullines in a variety of proteins, thereby altering their capacity to induce immune responses. Production of antibodies to citrullinated proteins has been observed in rheumatoid arthritis and chronic lung disease as well as in normal smokers. These antibodies may be the target of autoantibodies that contribute to organ pathology. Alterations in the availability and presentation of autoantigens may be important components of immunoreactivity in certain models of organ-specific autoimmune diseases. In addition, these factors may be relevant to an understanding of the pathogenesis of various drug-induced autoimmune conditions. However, the diversity of autoreactivity manifesting in non-organ-specific systemic autoimmune diseases suggests that these conditions may result from a more general activation of the immune system rather than from an alteration in individual self-antigens.

Many autoimmune diseases are characterized by the presence of antibodies that react with antigens present in apoptotic material. Defects in the clearance of apoptotic material have been shown to elicit autoimmunity and autoimmune disease in a number of animal models. Moreover, such defects have been found in patients with SLE. Apoptotic debris that is not cleared quickly by the immune system can function as endogenous ligands for a number of PRRs on dendritic cells and B cells. Under such circumstances, dendritic cells and/or B cells are activated, and an immune response to apoptotic debris can develop. In addition, the presence of uncleared extracellular apoptotic material within germinal centers of secondary lymphoid organs in patients with SLE may facilitate the direct activation of autoimmune B cell clones or may function to select such clones during immune responses.

Deficiency in C1q, likewise, can predispose or exacerbate autoimmunity. C1q assists in the clearance of apoptotic debris binding to IgM autoantibodies and to inhibitory receptors on monocytes and dendritic cells.

If C1q is not present, a mechanism of immune suppression is lost. Moreover, if antibodies have undergone class switch recombination to IgG, the apoptotic debris containing immune complexes will engage activating Fc receptors on myeloid cells to induce an inflammatory response. Studies in a number of experimental models have suggested that intense stimulation of T lymphocytes can produce nonspecific signals that bypass the need for antigen-specific helper T cells and lead to polyclonal B cell activation with the formation of multiple autoantibodies. For example, antinuclear, antierythrocyte, and antilymphocyte antibodies are produced during the chronic graft-versus-host reaction. In addition, true autoimmune diseases, including autoimmune hemolytic anemia and immune complex–mediated glomerulonephritis, can be induced in this manner. While such diffuse activation of helper T cell activity clearly can cause autoimmunity, nonspecific stimulation of B lymphocytes can also lead to the production of autoantibodies. Thus, the administration of polyclonal B cell activators, such as bacterial endotoxin, to normal mice leads to the production of a number of autoantibodies, including those to DNA and IgG (rheumatoid factor). A variety of genetic modifications resulting in hyperresponsiveness of B cells also can lead to the production of autoantibodies and, in animals of appropriate genetic background, a lupus-like syndrome. Moreover, excess B cell activating factor (BAFF), a B cell survival promoting cytokine, can impair B cell tolerance, cause T cell–independent B cell activation, and lead to the development of autoimmunity. SLE can also be induced in mice through exuberant dendritic cell activation, through a redundancy of TLR7 on the Y chromosome (as in BXSB-Yaa mice), or through exposure to CpG, a ligand for TLR9. The ensuing induction of inflammatory mediators can cause a switch from the production of nonpathogenic IgM autoantibodies to the production of pathogenic IgG autoantibodies in the absence of antigen-specific T cell help. Aberrant selection of the B or T cell repertoire at the time of antigen receptor expression can also predispose to autoimmunity. For example, B cell immunodeficiency caused by an absence of the B cell receptor–associated kinase (Bruton's tyrosine kinase) leads to X-linked agammaglobulinemia. This syndrome is characterized by reduced B cell numbers. This leads to high levels of BAFF which alter B cell selection and results in greater survival of autoreactive B cells. Likewise, negative selection of autoreactive T cells in the thymus requires expression of the autoimmune regulator (AIRE) gene that enables the expression of tissue-specific proteins in thymic medullary epithelial cells. Peptides from these proteins are expressed in the context of major histocompatibility complex (MHC) molecules and mediate the central deletion of autoreactive T cells. The absence of AIRE gene expression leads to a failure of negative selection of autoreactive cells, autoantibody production, and severe inflammatory destruction of multiple organs. Individuals deficient in AIRE gene expression develop autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy (APECED).

Primary alterations in the activity of T and/or B cells, cytokine imbalances, or defective immunoregulatory circuits may also contribute to the emergence of autoimmunity. Diminished production of tumor necrosis factor (TNF) and interleukin (IL) 10 has been reported to be associated with the development of autoimmunity. Overproduction or therapeutic administration of type 1 interferon has also been associated with autoimmunity. Overexpression of costimulatory molecules on T cells similarly can lead to autoantibody production.

Autoimmunity may also result from an abnormality of immunoregulatory mechanisms. Observations made in both human autoimmune disease and animal models suggest that defects in the generation and

expression of regulatory T cell (Treg) activity may allow the production of autoimmunity. It has been appreciated that the IPEX (immunodysregulation, polyendocrinopathy, enteropathy X-linked) syndrome results from the failure to express the FOXP3 gene, which encodes a molecule critical in the differentiation of Tregs. Administration of normal Tregs or of factors derived from them can prevent the development of autoimmune disease in rodent models of autoimmunity, and allogeneic stem cell transplantation ameliorates human IPEX. Abnormalities in the function of Tregs have been noted in a number of human autoimmune diseases, including rheumatoid arthritis and SLE, although it remains uncertain whether these functional abnormalities are causative or are secondary to inflammation. One of the mechanisms by which Tregs control immune/inflammatory responses is by the production of the cytokine IL-10. In this regard, children with a deficiency in the expression of IL-10 or the IL-10 receptor develop inflammatory bowel disease that mimics Crohn's disease and that can be cured by allogeneic stem cell transplantation. Finally, recent data indicate that B cells may also exert regulatory function, largely through the production of IL-10. Deficiency of IL-10-producing regulatory B cells can prolong the course of multiple sclerosis in an animal model, and such cells are thought to be functionally diminished in human SLE.

It should be apparent that no single mechanism can explain all the varied manifestations of autoimmunity or autoimmune disease. Furthermore, genetic evaluation has shown that convergence of a number of abnormalities is often required for the induction of an autoimmune disease. Additional factors that appear to be important determinants in the induction of autoimmunity include age, sex (many autoimmune diseases are far more common in women), exposure to infectious agents, and environmental contacts. How all of these disparate factors affect the capacity to develop self-reactivity is currently being investigated intensively.

GENETIC CONSIDERATIONS



Evidence in humans that there are susceptibility genes for autoimmunity comes from family studies and especially from studies of twins. Studies in type 1 diabetes mellitus, rheumatoid arthritis, multiple sclerosis, and SLE have shown that ~15–30% of pairs of monozygotic twins show disease concordance, whereas the figure is <5% for dizygotic twins. The occurrence of different autoimmune diseases within the same family has suggested that certain susceptibility genes may predispose to a variety of autoimmune diseases. Genome-wide association studies have begun to identify polymorphisms in individual genes that are associated with specific autoimmune diseases. More than 100 genetic polymorphisms associated with one or more autoimmune diseases have been identified to date. It is notable that some genes are associated with multiple autoimmune diseases, whereas others are specifically associated with only one autoimmune condition. Moreover, recent genetic evidence suggests that clusters of genetic risk factors can commonly be found in groups of autoimmune diseases. Four general clusters have been identified: one group most frequently associated with Crohn's disease, psoriasis, and multiple sclerosis; a second cluster most strongly associated with celiac disease, rheumatoid arthritis, and SLE; a third cluster most strongly associated with type 1 diabetes, multiple sclerosis, and rheumatoid arthritis; and a fourth cluster most strongly associated with type 1 diabetes, rheumatoid arthritis, celiac disease, Crohn's disease, and SLE. These results imply that

autoimmune diseases with widely different clinical presentations and patterns of organ involvement could involve similar immunopathogenic pathways or endophenotypes. For example, the same allele of the gene encoding PTPN22 is associated with multiple autoimmune diseases. Its product is a phosphatase expressed by a variety of hematopoietic cells that downregulates antigen receptor–mediated stimulation of T and B cells. The risk allele is associated with type 1 diabetes mellitus, rheumatoid arthritis, and SLE in some populations. In recent years, genome-wide association studies have demonstrated a variety of other genes that are involved in human autoimmune diseases. Most genes individually confer a relatively low risk for autoimmune diseases and are found in normal individuals. In addition, most polymorphisms associated with autoimmune diseases are in non-coding regions of DNA, implying that expression levels rather that altered function might convey most genetic risk for autoimmune diseases. Abnormalities in epigenetics or the mechanisms controlling and influencing gene expression has also been implicated in contributing to autoimmune diseases. No single gene or epigenetic modification has been identified that is essential for autoimmune diseases. In addition to this evidence from humans, certain inbred mouse strains reproducibly develop specific spontaneous or experimentally induced autoimmune diseases, whereas others do not. These findings have led to an extensive search for genes that determine susceptibility to autoimmune disease and for genes that might be protective.

The strongest consistent association for susceptibility to autoimmune disease is with particular MHC alleles. It has been suggested that the association of MHC genotype with autoimmune disease relates to differences in the ability of different allelic variations of MHC molecules to present autoantigenic peptides to autoreactive T cells. An alternative hypothesis involves the role of MHC alleles in shaping the T cell receptor repertoire during T cell ontogeny in the thymus. In addition, specific MHC gene products may themselves be the source of peptides that can be recognized by T cells. Cross-reactivity between such MHC peptides and peptides derived from proteins produced by common microbes may trigger autoimmunity by molecular mimicry. However, MHC genotype alone does not determine the development of autoimmunity. Identical twins are far more likely to develop the same autoimmune disease than MHC-identical nontwin siblings; this observation suggests that genetic factors other than the MHC affect disease susceptibility. Studies of the genetics of type 1 diabetes mellitus, SLE, rheumatoid arthritis, and multiple sclerosis in humans and mice have identified several independently segregating disease susceptibility loci in addition to the MHC. Genes that encode molecules of the innate immune response are also involved in autoimmunity. In humans, inherited homozygous deficiency of the early proteins of the classic pathway of complement (C1q, C4, or C2) as well as genes involved in the type 1 interferon pathway are very strongly associated with the development of SLE.

IMMUNOPATHOGENIC MECHANISMS IN AUTOIMMUNE DISEASES

The mechanisms of tissue injury in autoimmune diseases can be divided into antibody-mediated and cell-mediated processes. Representative examples are listed in Table 348-3.

TABLE 348-3

Mechanisms of Tissue Damage in Autoimmune Disease

EFFECTOR	MECHANISM	TARGET	DISEASE
Autoantibody	Blocking or inactivation	α Chain of the nicotinic acetylcholine receptor	Myasthenia gravis
		Phospholipid–β_2-glycoprotein I complex	Antiphospholipid syndrome
		Insulin receptor	Insulin-resistant diabetes mellitus
		Intrinsic factor	Pernicious anemia
	Stimulation	TSH receptor (LATS)	Graves' disease
		Proteinase-3 (ANCA)	Granulomatosis with polyangiitis
		Epidermal cadherin	Pemphigus vulgaris
		Desmoglein 3	
	Complement activation	α_3 Chain of collagen IV	Goodpasture's syndrome
	Immune complex formation	Double-stranded DNA	Systemic lupus erythematosus
		Immunoglobulin	Rheumatoid arthritis
	Opsonization	Platelet GpIIb:IIIa	Autoimmune thrombocytopenic purpura
		Rh antigens, I antigen	Autoimmune hemolytic anemia
	Antibody-dependent cellular cytotoxicity	Thyroid peroxidase, thyroglobulin	Hashimoto's thyroiditis
T cells	Cytokine production		Rheumatoid arthritis, multiple sclerosis, type 1 diabetes mellitus

EFFECTOR	MECHANISM	TARGET	DISEASE
	Cellular cytotoxicity		Type 1 diabetes mellitus

Abbreviations: ANCA, antineutrophil cytoplasmic antibody; LATS, long-acting thyroid stimulator; TSH, thyroid-stimulating hormone.

The pathogenicity of autoantibodies can be mediated through several mechanisms, including opsonization of soluble factors or cells, activation of an inflammatory cascade via the complement system, and interference with the physiologic function of soluble molecules or cells.

In autoimmune thrombocytopenic purpura, opsonization of platelets targets them for elimination by phagocytes. Likewise, in autoimmune hemolytic anemia, binding of immunoglobulin to red cell membranes leads to phagocytosis and lysis of the opsonized cell. Goodpasture's syndrome, a disease characterized by lung hemorrhage and severe glomerulonephritis, represents an example of antibody binding leading to local activation of complement and neutrophil accumulation and activation. The autoantibody in this disease binds to the α_3 chain of type IV collagen in the basement membrane. In SLE, activation of the complement cascade at sites of immunoglobulin deposition in renal glomeruli is considered to be a major mechanism of renal damage. Moreover, the DNA- and RNA-containing immune complexes in SLE activate TLR9 and TLR7, respectively, in plasmacytoid dendritic cells and promote the production of type 1 interferon and proinflammatory, cytokines conducive to amplification of the autoimmune response.

Autoantibodies can also interfere with normal physiologic functions of cells or soluble factors. Autoantibodies to hormone receptors can lead to stimulation of cells or to inhibition of cell function through interference with receptor signaling. For example, long-acting thyroid stimulators—autoantibodies that bind to the receptor for thyroid-stimulating hormone (TSH)—are present in Graves' disease and function as agonists, causing the thyroid to respond as if there were an excess of TSH. Alternatively, antibodies to the insulin receptor can cause insulin-resistant diabetes mellitus through receptor blockade. In myasthenia gravis, autoantibodies to the acetylcholine receptor can be detected in 85–90% of patients and are responsible for muscle weakness. The exact location of the antigenic epitope, the valence and affinity of the antibody, and perhaps other characteristics determine whether activation or blockade results from antibody binding.

Antiphospholipid antibodies are associated with thromboembolic events in primary and secondary antiphospholipid syndrome and have also been associated with fetal wastage. The major antibody is directed to the phospholipid–β_2-glycoprotein I complex and appears to exert a procoagulant effect. In pemphigus vulgaris, autoantibodies bind to desmoglein 1 and 3, components of the epidermal cell desmosome, and play a role in the induction of the disease. These antibodies exert their pathologic effect by disrupting cell–cell junctions through stimulation of the production of epithelial proteases, with consequent blister formation. Cytoplasmic antineutrophil cytoplasmic antibody (c-ANCA), found in granulomatosis with

polyangiitis, is an antibody to an intracellular antigen, the 29-kDa serine protease (proteinase-3). In vitro experiments have shown that IgG anti-c-ANCA causes cellular activation and degranulation of primed neutrophils.

It is important to note that autoantibodies of a given specificity may cause disease only in genetically susceptible hosts, as has been shown in experimental models of myasthenia gravis, SLE, rheumatic fever, and rheumatoid arthritis. Furthermore, once organ damage is initiated, new inflammatory cascades are initiated that can sustain and amplify the autoimmune process. Finally, some autoantibodies seem to be markers for disease but have, as yet, no known pathogenic potential.

AUTOIMMUNE DISEASES

Manifestations of autoimmunity are found in a large number of pathologic conditions. However, their presence does not necessarily imply that the pathologic process is an autoimmune disease. A number of attempts to establish formal criteria for the classification of diseases as autoimmune have been made, but none is universally accepted. One set of criteria is shown in **Table 348-4**; however, this scheme should be viewed merely as a guide in consideration of the problem.

TABLE 348-4

Human Autoimmune Disease: Presumptive Evidence for Immunologic Pathogenesis

Major Criteria
1. Presence of autoantibodies or evidence of cellular reactivity to self 2. Documentation of relevant autoantibody or lymphocytic infiltrate in the pathologic lesion 3. Demonstration that relevant autoantibody or T cells can cause tissue pathology a. Transplacental transmission b. Adaptive transfer into animals c. In vitro impact on cellular function
Supportive Evidence
1. Reasonable animal model 2. Beneficial effect from immunosuppressive agents 3. Association with other evidence of autoimmunity 4. No evidence of infection or other obvious cause

To classify a disease as autoimmune, it is necessary to demonstrate that the immune response to a self-antigen causes the observed pathology. Initially, the detection of antibodies to the affected tissue in the

serum of patients suffering from various diseases was taken as evidence that these diseases had an autoimmune basis. However, such autoantibodies are also found when tissue damage is caused by trauma or infection and in these cases are secondary to tissue damage. Thus, autoimmunity must be shown to be pathogenic before a disease is categorized as autoimmune.

To confirm autoantibody pathogenicity, it may be possible to transfer disease to experimental animals by the administration of autoantibodies from a patient, with the subsequent development of pathology in the recipient similar to that seen in the patient. This scenario has been documented, for example, in Graves' disease. Some autoimmune diseases can be transferred from mother to fetus and are observed in the newborn babies of diseased mothers. The symptoms of the disease in the newborn usually disappear as the levels of maternal antibody decrease. An exception, however, is congenital heart block, in which damage to the developing conducting system of the heart follows in utero transfer of anti-Ro antibody from the mother to the fetus. This antibody transfer can result in a permanent developmental defect in the heart.

In most situations, the critical factors that determine when the development of autoimmunity results in autoimmune disease have not been delineated. The relationship of autoimmunity to the development of autoimmune disease may be associated with the fine specificity of the antibodies or T cells or their specific effector capabilities. In many circumstances, a mechanistic understanding of the pathogenic potential of autoantibodies has not been established. In some autoimmune diseases, biased production of cytokines by helper T (T_H) cells may play a role in pathogenesis. In this regard, T cells can differentiate into specialized effector cells that predominantly produce interferon γ (T_H1), IL-4 (T_H2), or IL-17 (T_H17) or that provide help to B cells (T follicular helper, T_{FH}) (**Chap. 342**). T_H1 cells facilitate macrophage activation and classic cell–mediated immunity, whereas T_H2 cells are thought to have regulatory functions and are involved in the resolution of normal immune responses as well as in the development of responses to a variety of parasites. T_H17 cells produce a number of inflammatory cytokines, including IL-17 and IL-22, and seem to be prominently involved in host resistance to certain fungal infections. T_{FH} cells help B cells by constitutively producing IL-21. In a number of autoimmune diseases, such as rheumatoid arthritis, multiple sclerosis, type 1 diabetes mellitus, and Crohn's disease, there appears to be biased differentiation of T_H1 and T_H17 cells, with resultant organ damage. Studies suggest an accentuated differentiation of T_H17 cells associated with animal models of inflammatory arthritis, whereas increased differentiation of T_{FH} cells has been associated with animal models of SLE. Importantly, genetically determined or environmentally induced features of the target organ may determine susceptibility of the target organ to autoantibodies or autoreactive T cell-mediated damage.

ORGAN-SPECIFIC VERSUS SYSTEMIC AUTOIMMUNE DISEASES

The spectrum of autoimmune diseases ranges from conditions specifically affecting a single organ to systemic disorders that involve many organs (**Table 348-5**). Hashimoto's autoimmune thyroiditis is an example of an organ-specific autoimmune disease (**Chap. 375**). In this disorder, a specific lesion in the

thyroid is associated with infiltration of mononuclear cells and damage to follicular cells. Antibody to thyroid constituents can be demonstrated in nearly all cases. Other organ- or tissue-specific autoimmune disorders include pemphigus vulgaris, autoimmune hemolytic anemia, idiopathic thrombocytopenic purpura, Goodpasture's syndrome, myasthenia gravis, and sympathetic ophthalmia. One important feature of some organ-specific autoimmune diseases is the tendency for overlap, such that an individual with one specific syndrome is more likely to develop a second syndrome. For example, there is a high incidence of pernicious anemia in individuals with autoimmune thyroiditis. More striking is the tendency for individuals with an organ-specific autoimmune disease to develop multiple other manifestations of autoimmunity without the development of associated organ pathology. Thus, as many as 50% of individuals with pernicious anemia have non-cross-reacting antibodies to thyroid constituents, whereas patients with myasthenia gravis may develop antinuclear antibodies, antithyroid antibodies, rheumatoid factor, antilymphocyte antibodies, and polyclonal hypergammaglobulinemia. Part of the explanation may relate to the genetic elements shared by individuals with these different diseases.

TABLE 348-5

Diseases on the Autoimmune Spectrum

Organ Specific	
Graves' disease	Vitiligo
Hashimoto's thyroiditis	Autoimmune hemolytic anemia
Autoimmune polyglandular syndrome	Autoimmune thrombocytopenic purpura
Type 1 diabetes mellitus	Pernicious anemia
Insulin-resistant diabetes mellitus	Myasthenia gravis
Immune-mediated infertility	Multiple sclerosis
Autoimmune Addison's disease	Guillain-Barré syndrome
Pemphigus vulgaris	Stiff-man syndrome
Pemphigus foliaceus	Acute rheumatic fever
Dermatitis herpetiformis	Sympathetic ophthalmia
Autoimmune alopecia	Goodpasture's syndrome
Primary biliary cirrhosis	
Organ Nonspecific (Systemic)	
Systemic lupus erythematosus	Granulomatosis with polyangiitis
Rheumatoid arthritis	Antiphospholipid syndrome
Systemic necrotizing vasculitis	Sjögren's syndrome

Systemic autoimmune diseases differ from organ-specific diseases in that pathologic lesions are found in multiple diverse organs and tissues. The hallmark of these conditions is the demonstration of associated relevant autoimmune manifestations that are likely to have an etiologic role in organ pathology. SLE represents the prototype of these disorders because of its abundant autoimmune manifestations. SLE is a

disease of protean manifestations that characteristically involves the kidneys, joints, skin, serosal surfaces, blood vessels, and central nervous system (Chap. 349). The disease is associated with a vast array of autoantibodies whose production appears to be a part of a generalized hyperreactivity of the humoral immune system. Other features of SLE include generalized B cell hyperresponsiveness and polyclonal hypergammaglobulinemia. Current evidence suggests that both hypo- and hyperresponsiveness to antigen can lead to survival and activation of autoreactive B cells in SLE. The autoantibodies in SLE are thought to arise as part of an accentuated T cell–dependent B cell response since most pathogenic anti-DNA autoantibodies exhibit evidence of extensive somatic hypermutation.

TREATMENT

TREATMENT

Autoimmune Diseases

Treatment of autoimmune diseases can focus on suppressing the induction of autoimmunity, restoring normal regulatory mechanisms, or inhibiting the effector mechanisms. To decrease the number or function of autoreactive cells, immunosuppressive or ablative therapies are most commonly used. In recent years, cytokine blockade has been demonstrated to be effective in preventing immune activation in some diseases or in inhibiting the extensive inflammatory effector mechanisms characteristic of these diseases. New therapies have also been developed to target lymphoid cells more specifically by blocking a costimulatory signal needed for T or B cell activation, by blocking the migratory capacity of lymphocytes, or by eliminating the effector T cells or B cells. The efficacy of these therapies in some diseases—e.g., SLE (belimumab), rheumatoid arthritis (TNF neutralization, IL-6 receptor blockade, CD28 competition, B cell depletion, IL-1 competition), psoriasis (IL-12/23 depletion, TNF neutralization), and inflammatory bowel disease (TNF neutralization, IL-12/23 neutralization)—has been demonstrated. One major advance in inhibiting effector mechanisms has been the introduction of cytokine blockade that appears to limit organ damage in some diseases, including rheumatoid arthritis, inflammatory bowel disease, psoriasis, and the spondyloarthritides. Small molecules that block cytokine signaling pathways by blocking the Janus kinase (JAK) family of kinases have recently been introduced into the clinic. Biologicals that delete B cells (anti-CD20 antibody) have recently been approved for the treatment of rheumatoid arthritis and have demonstrated efficacy in other autoimmune diseases as well. Their efficacy in diseases characterized by pathogenic effector T cells has highlighted the importance of B cells as antigen presenting cells in autoimmune diseases. Finally, there is renewed interest in cellular therapies in autoimmune diseases, including hematopoietic stem cell reconstitutions and treatment with immunosuppressive mesenchymal stem cells. Therapies that prevent target-organ damage or support target-organ function also remain important in the management of autoimmune disease.

FURTHER READING

Caielli S et al: Oxidized mitochondrial nucleoids release by neutrophils drive type I interferon production in human lupus. J Exp Med 5:697, 2016.

Jackson SW et al: B cells take the front seat; Dysregulated B cell signals orchestrate loss of tolerance and autoantibody production. Curr Opin Immunology 33:70, 2015.

Teruel M, Alacron-Riquelme ME: Genetics of systemic lupus erythematosus and Sjögren's syndrome: An update. Curr Opin Rheumatol 28:506, 2016. [PubMed: 27227345]

Tsokos GC et al: New insights into the immunopathogenesis of systemic lupus erythematosus. Nat Rev Rheumatol 22:716, 2016.

Ueno H: T follicular helper cells in human autoimmunity. Curr Opin Immunol 43:24, 2016. [PubMed: 27588918]

Yin Y et al: Normalization of CD4+ T cell metabolism reverses lupus. Science Transl Med 7:274, 2015.

Chapter 349: Systemic Lupus Erythematosus

Bevra Hannahs Hahn

DEFINITION AND PREVALENCE

Systemic lupus erythematosus (SLE) is an autoimmune disease in which organs and cells undergo damage initially mediated by tissue-binding autoantibodies and immune complexes. In most patients, autoantibodies are present for a few years before the first clinical symptom appears. Ninety percent of patients are women of child-bearing years; people of all genders, ages, and ethnic groups are susceptible. Prevalence of SLE in the United States is 20–150 per 100,000 women depending on race and gender; highest prevalence is in African-American and Afro-Caribbean women, and lowest prevalence is in white men.

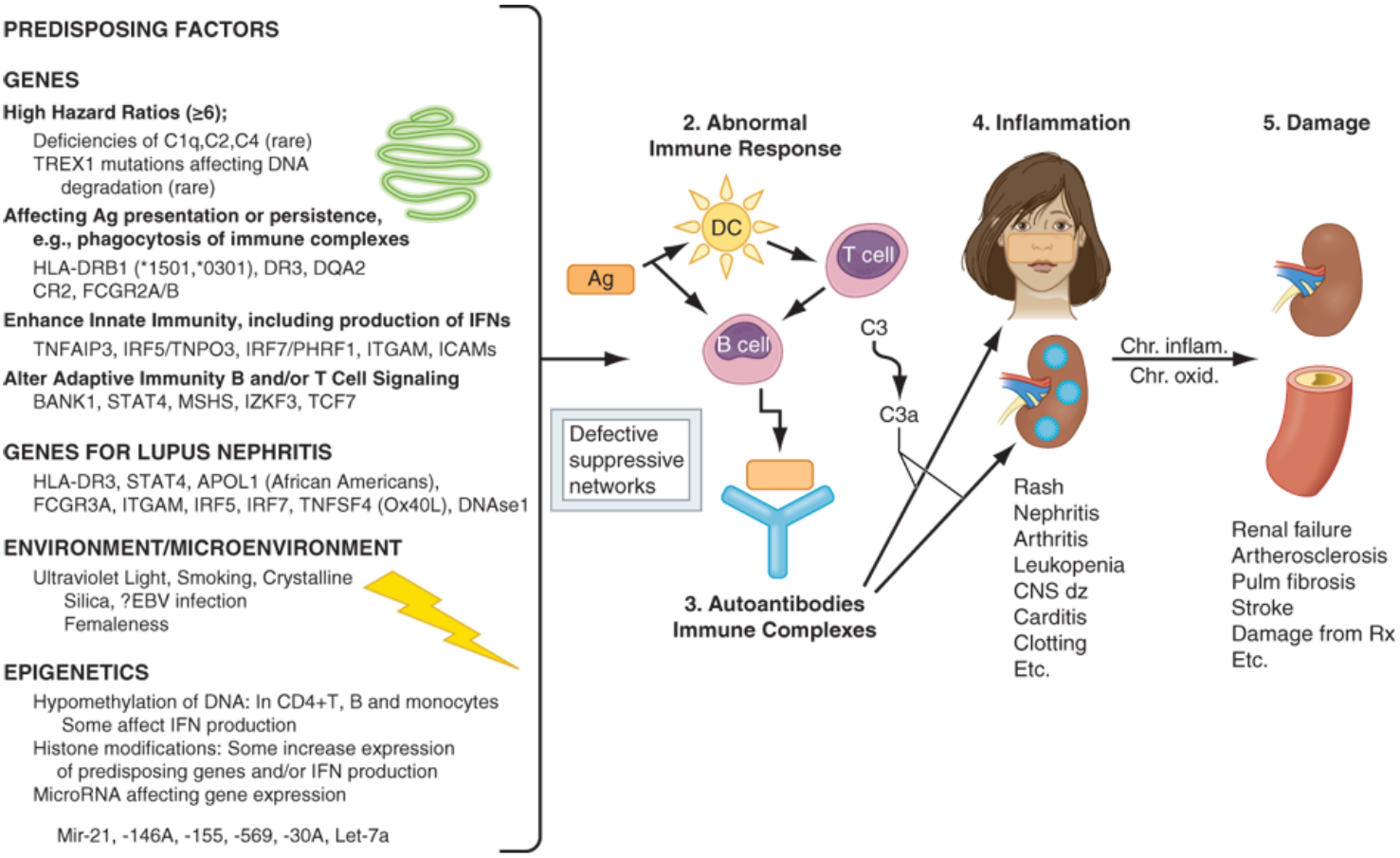
PATHOGENESIS AND ETIOLOGY

The proposed pathogenic mechanisms of SLE are illustrated in **Fig. 349-1**. The abnormal immune responses underlying SLE may be summarized as leading to production of increased quantities and immunogenic forms of nucleic acids, their accompanying proteins, and other self-antigens. The process may begin with autoimmunity-inducing activation of innate immunity, partly through binding of DNA/RNA/proteins by toll-like receptors in those cells. The changes include dendritic cells producing interferon α (IFNα), activated macrophages producing inflammatory cytokines/chemokines such as interleukin (IL12), tumor necrosis factor α (TNFα), and the B cell maturation/survival factor BLys/BAFF, neutrophils releasing DNA/protein-containing nets, and natural killer (NK) cells unable to kill autoreactive T and B cells or to produce the transforming growth factor β (TGFβ) needed for development of regulatory T cells. Upregulation of genes induced by IFNs is a genetic "signature" in peripheral blood cells of 50–80% of SLE patients. The innate immune system interacts with the B and T cells of adaptive immunity, which further drive autoimmune responses. T lymphocytes have altered metabolism (abnormal mitochondrial electron transport, membrane potential, and oxidative stress), increased glucose utilization, increased pyruvate production, activation of mTOR, and increased autophagy. T and B cells are more easily activated and driven into apoptosis than are normal cells, probably due to autoantibodies binding them plus abnormal signaling after engagement of surface molecules resulting in abnormally low production of IL2, which is required for T cell survival. B cells present antigen and secrete IL6 and IL10, further promoting autoreactive B cell survival (which is also favored by estrogen). Lupus phagocytic cells have reduced capacity to clear immune complexes, apoptotic cells, and their DNA/RNA/Ro/La and phospholipid containing surface blebs. The result is persistence of large quantities of autoantigens and resultant large quantities of autoantibodies with increased numbers of activated B cells and plasmablasts/plasma cells, and autoreactive T cells with shifts away from regulatory populations toward increased numbers and functions of Th1, T17, and Tfh cells, all of which promote production of autoantibodies and tissue damage. This damage begins with deposition of autoantibodies and/or immune complexes, followed by destruction mediated by complement activation and release of cytokines/chemokines. Non-immune tissue-fixed cells are then activated to produce more inflammation and damage, such as basal cells of the dermis, synovial fibroblasts, renal mesangial cells, podocytes and tubular epithelium, and endothelial cells throughout the body. Meanwhile, the initial immune attack is attracting into the target tissues additional B and T cells, monocytes/macrophages, dendritic cells, and plasma cells. Inflammation also causes release of vasoactive peptides, oxidative damage, growth factors and fibrosing factors. Sclerosis/fibrosis with irreversible tissue damage can occur in multiple tissues including kidneys, lungs, blood vessels, and skin. Each of these processes depends on the individual's genetic background, environmental influences, and epigenetics. Autoantibodies of SLE are referred to in Fig. 349-1 and described in **Table 349-1**.

FIGURE 349-1

Pathogenesis of systemic lupus erythematosus (SLE). Pathogenesis is related in large part to production of increased quantities and immunogenic forms of nucleic acids and other self-antigens, which drive autoimmune-inducing activation of innate immunity, autoantibodies, and T cells. Interactions between genes, environment, and epigenetic changes drive increased autophagy, Ag presentation, neutrophil netosis, autoantibody formation with increased plasma cells, and production of pathogenic effector T cells in Th1, Th17, and Tfh subsets, with ineffective regulatory networks. Genes confirmed in more than one genome-wide association analysis in multiple racial groups that increase susceptibility to SLE or lupus nephritis (HR ≥1.5) are listed (*reviewed in Teruel M, Alarcon-Riquelme ME: The genetic basis of systemic lupus erythematosus: What are the risk factors and what have we learned. J Autoimmun 74:161, 2016; and Iwamoto T, Niewold TB:*

Genetics of human lupus nephritis. Clin Immunol 2016. Epub ahead of print). Epigenetics are reviewed in *Long H et al: The critical role of epigenetics in systemic lupus erythematosus and autoimmunity, J Autoimmun 2016. Epub ahead of print.* Gene-environment interactions (reviewed in *Barbhaiya M, Costenbader KH: Environmental exposures and the development of systemic lupus erythematosus. Curr Opin Rheumatol 2016. Epub ahead of print*) result in abnormal immune responses that generate pathogenic autoantibodies and immune complexes that deposit in tissue, activate complement, induce cytokine and chemokine release causing inflammation, and over time lead to irreversible organ damage (reviewed in *Anders HJ, Rovin B: A pathophysiology-based approach to the diagnosis and treatment of lupus nephritis. Kidney Int 90:493, 2016;* and *Hahn BH: Pathogenesis of SLE, in Dubois Lupus Erythematosus, 8th ed, DJ Wallace, BH Hahn, (eds). Philadelphia, Elsevier, 2013*). Ag, antigen; C1q, complement system; C3, complement component; CNS, central nervous system; DC, dendritic cell; EBV, Epstein-Barr virus; HLA, human leukocyte antigen; FcR, immunoglobulin Fc-binding receptor; IL, interleukin; MCP, monocyte chemotactic protein; PTPN, phosphotyrosine phosphatase; UV, ultraviolet.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 349-1

Autoantibodies in Systemic Lupus Erythematosus (SLE)

Antibody	Prevalence, %	Antigen Recognized	Clinical Utility
Antinuclear antibodies	98	Multiple nuclear	Best screening test; repeated negative tests by immunofluorescence make SLE unlikely
Anti-dsDNA	70	DNA (double-stranded)	High titers are SLE-specific and in some patients correlate with disease activity, nephritis, vasculitis. Crithidia immunofluorescence is more specific for SLE than ELISA methods.
Anti-Sm	25	Protein complexed to 6 species of nuclear U1 RNA	Specific for SLE; no definite clinical correlations; most patients also have anti-RNP; more common in blacks and Asians than whites
Anti-RNP	40	Protein complexed to U1 RNA	Not specific for SLE; high titers associated with syndromes that have overlap features of several rheumatic syndromes including SLE; more common in blacks than whites; correlates with high IFN-induced gene signature
Anti-Ro (SS-A)	30	Protein complexed to hY RNA, primarily 60 kDa and 52 kDa	Not specific for SLE; associated with sicca syndrome, predisposes to subacute cutaneous lupus, and to neonatal lupus with congenital heart block; associated with decreased risk for nephritis
Anti-La (SS-B)	10	47-kDa protein complexed to hY RNA	Usually associated with anti-Ro; associated with decreased risk for nephritis
Antihistone	70	Histones associated with DNA (in nucleosome, chromatin)	More frequent in drug-induced lupus than in SLE
Antiphospholipid	50	Phospholipids, β_2 glycoprotein 1 (β_2G1) cofactor, prothrombin	Three tests available—ELISAs for cardiolipin and β_2G1, sensitive prothrombin time (DRVVT) for lupus anticoagulant; predisposes to clotting, fetal loss, thrombocytopenia
Antierythrocyte	60	Erythrocyte membrane	Measured as direct Coombs test; a small proportion develops overt hemolysis
Antiplatelet	30	Surface and altered cytoplasmic antigens on platelets	Associated with thrombocytopenia, but sensitivity and specificity are not good; this is not a useful clinical test
Antineuronal (includes antiglutamate receptor 2)	60	Neuronal and lymphocyte surface antigens	In some series, a positive test in CSF correlates with active CNS lupus
Antiribosomal P	20	Protein in ribosomes	In some series, a positive test in serum correlates with depression or psychosis due to CNS lupus

Abbreviations: CNS, central nervous system; CSF, cerebrospinal fluid; DRVVT, dilute Russell viper venom time; ELISA, enzyme-linked immunosorbent assay.

SLE is a multigenic disease. Rare single-gene defects confer high hazard ratios (HRs) for SLE (5–25), including homozygous deficiencies of early components of complement (C1q,r,s; C2; C4) and a mutation in *TREX1* (encoding a DNAase) on the X chromosome. In most genetically susceptible individuals, normal alleles of multiple genes each contribute a small amount to abnormal immune/inflammation/tissue damage responses; if enough predisposing variations are present, disease results. Approximately 60 genes with alleles increasing risk for SLE and/or lupus nephritis (examples listed in Fig. 349-1 which includes most with HR ≥1.5) have been identified in recent genome-wide association studies in different racial groups. Individually, they confer an HR for SLE of 1.5–3 and even in combination account for only 18% of disease susceptibility, suggesting that environmental exposures and epigenetics play major roles. Predisposing, antigen-presenting human leukocyte antigen (HLA) molecules are most commonly found, in multiple ethnic groups (HLA DRB1 *0301 and *1501 and DR3), as well as multiple genes across the major histocompatibility complex (MHC) 120-gene region. Non-HLA genetic factors are listed in Fig. 349-1 and include polymorphisms that affect innate and adaptive immunity pathways. Note the large number that influences IFN production—the most characteristic gene expression pattern of SLE patients. Other genes affect clearance of apoptotic cells or immune complexes, influence neutrophil adherence (*ITGAM*), and DNA repair (*TREX-1*). Some polymorphisms influence clinical manifestations; such as single nucleotide polymorphisms (SNPs) of *STAT4* that associate with severe disease, anti-DNA, nephritis, and antiphospholipid syndrome (APS), and an allele of *FCGRIIA* encoding a receptor that binds immune complexes poorly and predisposes to nephritis. Some gene effects are in promoter regions (e.g., IL-10), and others are conferred by copy numbers (e.g., C4A, TLR7). In addition, multiple epigenetic changes characterize SLE, including hypomethylation of DNA in CD4+ T cells, B cells, and monocytes, including genes that control production of type 1 interferons, and histone modifications. Some of these changes are mediated by microRNAs associated with SLE including some that control DNA Methyl transferases (DNMTs), such as mIR-146a, that control methylation of DNA in CD4+ T cells and IFN production. Some gene polymorphisms contribute to several autoimmune diseases, such as *STAT4* and *CTLA4*. All of these gene polymorphisms/transcription/epigenetic combinations influence immune responses to the external and internal environment; when such responses are too high and/or too prolonged and/or inadequately regulated, autoimmune disease is favored.

Female sex is permissive for SLE with evidence for hormone effects, genes on the X chromosome, and epigenetic differences between genders playing a role. Females of many mammalian species make higher antibody responses than males. Women exposed to estrogen-containing oral contraceptives or hormone replacement have an increased risk of developing SLE (HR 1.2–2). Estradiol binds to receptors on T and B lymphocytes, increasing activation and survival of those cells, especially autoreactive subsets, thus favoring prolonged immune responses. Genes on the X chromosome that influence SLE, such as *TREX-1*, may play a role in gender predisposition, possibly because some genes on the second X in females are not silent. People with XXY karyotype (Klinefelter's syndrome) have a significantly increased risk for SLE.

Several environmental stimuli may influence SLE (Fig. 349-1). Exposure to ultraviolet light causes flares of SLE in ~70% of patients, possibly by increasing apoptosis in skin cells or by altering DNA and intracellular proteins to make them antigenic. Some infections and lupus-inducing drugs activate autoreactive T and B cells; if such cells are not appropriately regulated, prolonged autoantibody production occurs. Most SLE patients have autoantibodies for 3 years or more before the first symptoms of disease, suggesting that regulation controls the degree of autoimmunity for years before quantities and qualities of autoantibodies, pathogenic B and T cells, and activated tissue-fixed cells such as macrophages cause clinical disease. Epstein-Barr virus (EBV) may be one infectious agent that can trigger SLE in susceptible individuals. Children and adults with SLE are more likely to be infected by EBV than age-, sex-, and ethnicity-matched controls. EBV contains amino acid sequences that mimic sequences on human spliceosomes (RNA/protein antigens) often recognized by autoantibodies in people with SLE. Current tobacco smoking increases risk for SLE (HR 1.5). Prolonged occupational exposure to crystalline silica (e.g., inhalation of soap powder dust or soil in farming activities) increases risk (HR 4.3) in African-American women. Drinking alcohol (2 glasses of wine a week or ½ of an alcoholic drink daily) reduces the risk of SLE. Thus, interplay between genetic susceptibility, environment, gender, race, and abnormal immune responses results in autoimmunity **(Chap. 348)**.

PATHOLOGY

In SLE, biopsies of affected skin show deposition of Ig at the dermal-epidermal junction (DEJ), injury to basal keratinocytes, and inflammation dominated by T lymphocytes in the DEJ and around blood vessels and dermal appendages. Clinically unaffected skin may also show Ig deposition at the DEJ. These patterns are not specific for dermatologic SLE; however, they are highly suggestive.

In renal biopsies, the pattern and severity of injury are important in diagnosis and in selecting the best therapy. Most recent clinical studies of lupus nephritis have used the International Society of Nephrology (ISN) and the Renal Pathology Society (RPS) classification **(Table 349-2)**. In the ISN/RPS classification, the addition of "a" for active and "c" for chronic changes gives physicians information regarding the potential

reversibility of disease. The system focuses on glomerular disease, although the presence of tubular interstitial and vascular disease, as well as the chronicity score in both glomeruli and interstitium, are important in predicting clinical outcomes. In general, class III and IV disease, as well as class V accompanied by III or IV disease, should be treated with aggressive immunosuppression if possible, because there is a high risk for end-stage renal disease (ESRD) if patients are untreated or undertreated. In contrast, treatment for lupus nephritis is not recommended in patients with class I or II disease or with extensive irreversible changes. In the recent Systemic Lupus International Collaborating Clinic (SLICC) criteria for classification of SLE, a diagnosis can be established on the basis of renal histology in the presence of lupus autoantibodies, without meeting additional criteria totaling 4 (Table 349-3).

TABLE 349-2

Classification of Lupus Nephritis (International Society of Nephrology and Renal Pathology Society)

Class I: Minimal Mesangial Lupus Nephritis
Normal glomeruli by light microscopy, but mesangial immune deposits by immunofluorescence.
Class II: Mesangial Proliferative Lupus Nephritis
Purely mesangial hypercellularity of any degree or mesangial matrix expansion by light microscopy, with mesangial immune deposits. A few isolated subepithelial or subendothelial deposits may be visible by immunofluorescence or electron microscopy, but not by light microscopy.
Class III: Focal Lupus Nephritis
Active or inactive focal, segmental or global endo- or extracapillary glomerulonephritis involving ≤50% of all glomeruli, typically with focal subendothelial immune deposits, with or without mesangial alterations.
Class III (A): Active lesions—focal proliferative lupus nephritis
Class III (A/C): Active and chronic lesions—focal proliferative and sclerosing lupus nephritis
Class III (C): Chronic inactive lesions with glomerular scars—focal sclerosing lupus nephritis
Class IV: Diffuse Lupus Nephritis
Active or inactive diffuse, segmental or global endo- or extracapillary glomerulonephritis involving ≥50% of all glomeruli, typically with diffuse subendothelial immune deposits, with or without mesangial alterations. This class is divided into diffuse segmental (IV-S) lupus nephritis when ≥50% of the involved glomeruli have segmental lesions, and diffuse global (IV-G) lupus nephritis when ≥50% of the involved glomeruli have global lesions. Segmental is defined as a glomerular lesion that involves less than one-half of the glomerular tuft. This class includes cases with diffuse wire loop deposits but with little or no glomerular proliferation.
Class IV-S (A): Active lesions—diffuse segmental proliferative lupus nephritis
Class IV-G (A): Active lesions—diffuse global proliferative lupus nephritis
Class IV-S (A/C): Active and chronic lesions—diffuse segmental proliferative and sclerosing lupus nephritis
Class IV-G (A/C): Active and chronic lesions—diffuse global proliferative and sclerosing lupus nephritis
Class IV-S (C): Chronic inactive lesions with scars—diffuse segmental sclerosing lupus nephritis
Class IV-G (C): Chronic inactive lesions with scars—diffuse global sclerosing lupus nephritis
Class V: Membranous Lupus Nephritis
Global or segmental subepithelial immune deposits or their morphologic sequelae by light microscopy and by immunofluorescence or electron microscopy, with or without mesangial alterations. Class V lupus nephritis may occur in combination with class III or IV, in which case both will be diagnosed. Class V lupus nephritis may show advanced sclerosis.
Class VI: Advanced Sclerotic Lupus Nephritis
≥90% of glomeruli globally sclerosed without residual activity.

Note: Indicate and grade (mild, moderate, severe) tubular atrophy, interstitial inflammation and fibrosis, and severity of arteriosclerosis or other vascular lesions.

Source: JJ Weening et al: Kidney Int 65:521, 2004. Reprinted by permission from Macmillan Publishers Ltd., Copyright 2004.

TABLE 349-3

Systemic Lupus International Collaborating Clinic Criteria for Classification of Systemic Lupus Erythematosus

Clinical Manifestations	Immunologic Manifestations
Skin Acute, subacute cutaneous LE (photosensitive, malar, maculopapular, bullous) Chronic cutaneous LE (discoid lupus, panniculitis, lichen planus-like, hypertrophic verrucous, chillblains) Oral or nasal ulcers Nonscarring Alopecia Synovitis involving ≥2 joints Serositis (pleurisy, pericarditis) Renal Prot/Cr ≥0.5 RBC casts Biopsy[a] Neurologic Seizures, psychosis, mononeuritis, myelitis, peripheral or cranial neuropathies, acute confusional state Hemolytic anemia Leukopenia (<4000/μL) or Lymphopenia (<1000/μL) Thrombocytopenia (<100,000/μL)	ANA > reference negative value Anti-dsDNA >reference, if by ELISA 2x reference Anti-Sm Antiphospholipid (any of lupus anticoagulant, false-positive RPR, anti-cardiolipin, anti-β glycoprotein I Low serum complement (C3, C4 or CH50) Positive direct Coombs test in absence of hemolytic anemia

[a]Renal biopsy read as systemic lupus qualifies for classification as SLE if any lupus autoantibodies are present, even if total criteria are fewer than 4.

Interpretation: Presence of any four criteria (must have at least 1 in each category) qualifies patient to be classified as having SLE with 93% specificity and 92% sensitivity. American College of Rheumatology is developing new criteria for SLE. For update see website *Rheumatology.org*.

Abbreviations: ANA, antinuclear antibody; Cr, creatinine; LE, lupus erythematosus; Prot, protein.

Source: M Petri et al: Arthritis Rheum 64:2677, 2012. Because these criteria are relatively new, some currently ongoing clinical studies use prior American College of Rheumatology Criteria; see EM Tan et al: Arthritis Rheum 25:1271, 1982; update MC Hochberg: Arthritis Rheum 40:1725, 1997.

Histologic abnormalities in blood vessels may also determine therapy. Patterns of vasculitis are not specific for SLE but may indicate active disease: leukocytoclastic vasculitis is most common **(Chap. 356)**.

Lymph node biopsies are usually performed to rule out infection or malignancies. In SLE, they show nonspecific diffuse chronic inflammation.

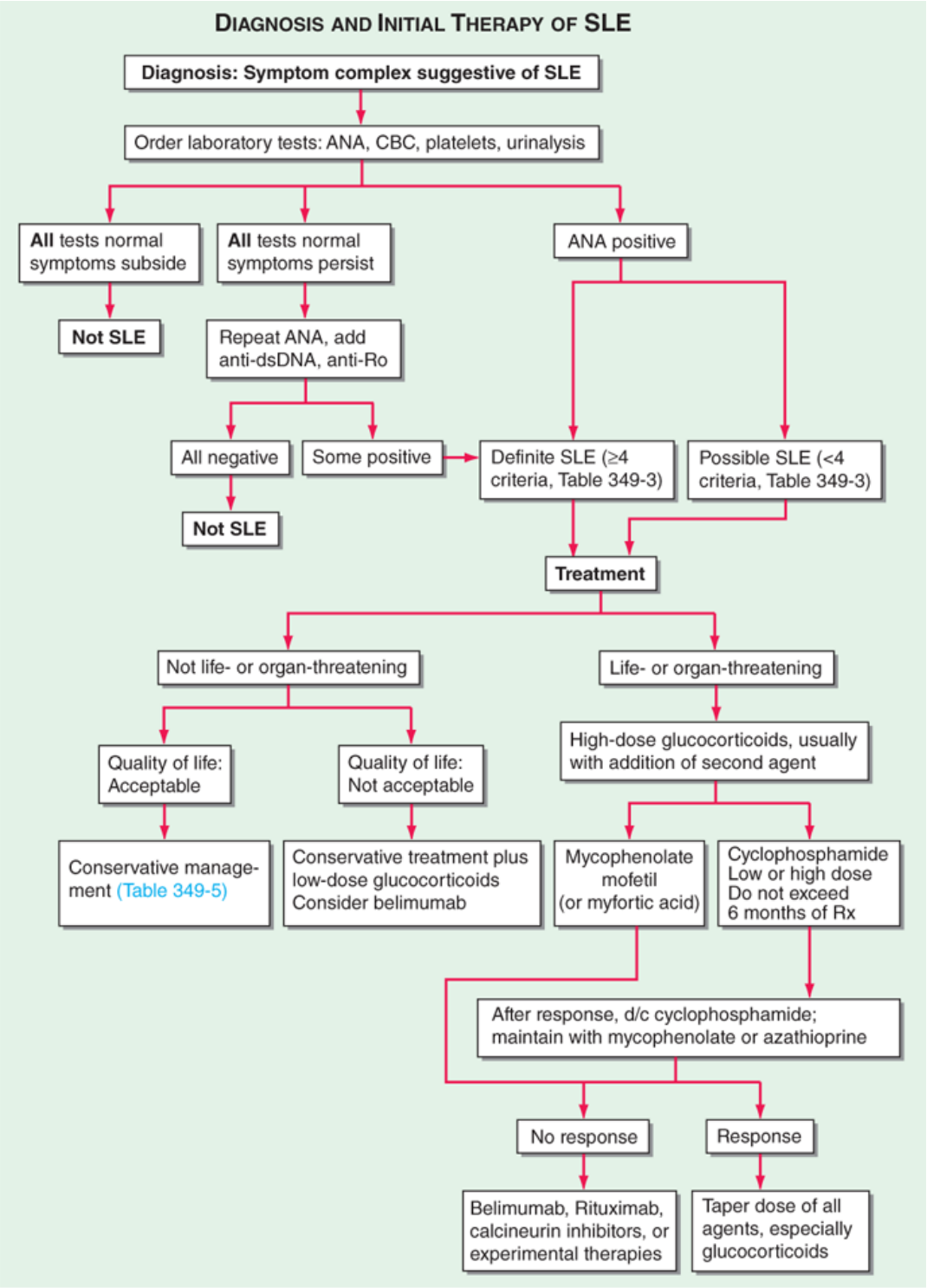
DIAGNOSIS

The diagnosis of SLE is based on characteristic clinical features and autoantibodies. Current criteria for classification are listed in Table 349-3, and an algorithm for diagnosis and initial therapy is shown in **Fig. 349-2**. The criteria are intended for confirming the diagnosis of SLE in patients included in studies; the author uses them in individual patients for estimating the probability that a disease is SLE. Any combination of four or more criteria, with at least one in the clinical and one in the immunologic category, well documented at any time during an individual's history, makes it likely that the patient has SLE. (Specificity and sensitivity are ~93% and ~92%, respectively.) In many patients,

criteria accrue over time. Antinuclear antibodies (ANA) are positive in >98% of patients during the course of disease; repeated negative tests by immunofluorescent methods suggest that the diagnosis is not SLE, unless other autoantibodies are present (Fig. 349-2). High-titer IgG antibodies to double-stranded DNA and antibodies to the Sm antigen are both specific for SLE and, therefore, favor the diagnosis in the presence of compatible clinical manifestations. The presence in an individual of multiple autoantibodies without clinical symptoms should not be considered diagnostic for SLE, although such persons are at increased risk.

FIGURE 349-2

Algorithm for diagnosis and initial therapy of systemic lupus erythematosus (SLE). For guidelines on management of lupus and lupus nephritis, see Hahn BH et al: Arthritis Care Res (Hoboken) 64:797, 2012; Bertsias GK et al: Ann Rheum Dis 71:1771, 2012; Anders HJ, Rovin B: Kidney Int 2016 Epub ahead of print. For details on mycophenolate and cyclophosphamide induction and maintenance therapies, see Henderson L et al: Cochrane Database Syst Rev 12:CD002922, 2012; Ginzler EM et al: Arthritis Rheum 62:211, 2010; Houssiau FA et al: Ann Rheum Dis 69:61, 2010; and Dooley MA et al: N Engl J Med 365:1886, 2011. For belimumab in treatment, see Bruce IN et al: Long term organ damage accumulation and safety in patients with SLE treated with belimumab plus standard of care. Lupus 25:699, 2016; Hahn BH: N Eng J Med 368:1528, 2013. For rituximab, see Lightstone L: Lupus 22:390, 2013 and Rovin BH et al: Arthritis Rheum 64:1215, 2012. For tacrolimus, see Liu Z et al: Multitarget therapy for induction treatment of lupus nephritis: A randomized trial. Ann Intern Med 162:18, 2015. ANA, antinuclear antibodies; CBC, complete blood count.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

INTERPRETATION OF CLINICAL MANIFESTATIONS

When a diagnosis of SLE is made, it is important to establish the severity and potential reversibility of the illness and to estimate the possible consequences of various therapeutic interventions. In the following paragraphs, descriptions of some disease manifestations begin with relatively mild problems and progress to those more life-threatening.

OVERVIEW AND SYSTEMIC MANIFESTATIONS

At its onset, SLE may involve one or several organ systems; over time, additional manifestations may occur (**Tables** 349-3 and **349-4**). Most of the autoantibodies characteristic of each person are present at the time clinical manifestations appear (Tables 349-1 and 349-3). Severity of SLE varies from mild and intermittent to severe and fulminant. Approximately 85% of patients have either continuing active disease (on current treatment) or one or more flares of active disease annually. Permanent complete remissions (absence of symptoms with no treatment) are rare; however, low-level disease activity on treatments such as hydroxychloroquine and/or low dose prednisone is achievable in ~35% of patients. Systemic symptoms, particularly fatigue and myalgias/arthralgias, are present most of the time. Severe systemic illness requiring high dose glucocorticoid therapy can occur with fever, prostration, weight loss, and anemia with or without other organ-targeted manifestations.

TABLE 349-4

Clinical Manifestations of SLE and Prevalence over the Entire Course of Disease[a]

Manifestation	Prevalence, %
Systemic: Fatigue, malaise, fever, anorexia, weight loss	95
Musculoskeletal	95
Arthralgias/myalgias	95
Nonerosive polyarthritis	60
Hand deformities	10
Myopathy/myositis	25/5
Ischemic necrosis of bone	15
Cutaneous	80
Photosensitivity	70
Malar rash	50
Oral ulcers	40
Alopecia	40
Discoid rash	20
Vasculitis rash	20
Other (e.g., urticaria, subacute cutaneous lupus)	15
Hematologic	85
Anemia (chronic disease)	70
Leukopenia (<4000/μL)	65
Lymphopenia (<1500/μL)	50
Thrombocytopenia (<100,000/μL)	15
Lymphadenopathy	15
Splenomegaly	15
Hemolytic anemia	10
Neurologic	60
Cognitive disorder	50

Manifestation	Prevalence, %
Mood disorder	40
Depression	25
Headache	25
Seizures	20
Mono-, polyneuropathy	15
Stroke, TIA	10
Acute confusional state or movement disorder	2–5
Aseptic meningitis, myelopathy	<1
Cardiopulmonary	60
Pleurisy, pericarditis, effusions	30–50
Myocarditis, endocarditis	10
Lupus pneumonitis	10
Coronary artery disease	10
Interstitial fibrosis	5
Pulmonary hypertension, ARDS, hemorrhage	<5
Shrinking lung syndrome	<5
Renal	30–50
Proteinuria ≥500 mg/24 h, cellular casts	30–60
Nephrotic syndrome	25
End-stage renal disease	5–10
Gastrointestinal	40
Nonspecific (nausea, mild pain, diarrhea)	30
Abnormal liver enzymes	40
Vasculitis	5
Thrombosis	15
Venous	10
Arterial	5

Manifestation	Prevalence, %
Ocular	15
Sicca syndrome	15
Conjunctivitis, episcleritis	10
Vasculitis	5

[a]Numbers indicate percentage of patients who have the manifestation at some time during the course of illness.

Abbreviations: ARDS, acute respiratory distress syndrome; SLE, Systemic Lupus Erythematosus; TIA, transient ischemic attack.

MUSCULOSKELETAL MANIFESTATIONS

Most people with SLE have intermittent polyarthritis, varying from mild to disabling, characterized by soft tissue swelling and tenderness in joints and/or tendons, most commonly in hands, wrists, and knees. Joint deformities (hands and feet) develop in only 10%. Erosions on joint x-rays are rare but can be identified by ultrasound in 10–50% of patients. Some individuals have rheumatoid-like arthritis with erosions and fulfill criteria for both RA and SLE ("rhupus"). Joint pain is the most common reason that patients increase their dose of glucocorticoids. If pain persists in a single joint, such as knee, shoulder, or hip, a diagnosis of ischemic necrosis of bone (INB) should be considered, particularly if there are no other manifestations of active SLE, because INB prevalence is increased in SLE, especially in patients treated with systemic glucocorticoids. Myositis with clinical muscle weakness, elevated creatine kinase levels, positive magnetic resonance imaging (MRI) scan, and muscle necrosis and inflammation on biopsy can occur, although most patients have myalgias without frank myositis. Glucocorticoid therapies (commonly) and antimalarial therapies (rarely) can cause muscle weakness; these adverse effects must be distinguished from active inflammatory disease.

CUTANEOUS MANIFESTATIONS

Lupus dermatitis can be classified as acute, subacute, or chronic, and there are many different types of lesions encompassed within these groups. Discoid lupus erythematosus (DLE) is the most common chronic dermatitis in lupus; lesions are roughly circular with slightly raised, scaly hyperpigmented erythematous rims and depigmented, atrophic centers in which all dermal appendages are permanently destroyed. Lesions can be disfiguring, particularly on the face and scalp. Treatment consists primarily of topical or locally injected glucocorticoids and systemic antimalarials. Only 5% of people with DLE have SLE (although half have positive ANA); however, among individuals with SLE, as many as 20% have DLE. The most common acute SLE rash is a photosensitive, slightly raised erythema, occasionally scaly, on the face (particularly the cheeks and nose—the "butterfly" rash), ears, chin, V region of the neck and chest, upper back, and extensor surfaces of the arms. Worsening of this rash often accompanies flare of systemic disease. Subacute cutaneous lupus erythematosus (SCLE) consists of scaly red patches similar to psoriasis, or circular flat red-rimmed lesions. Patients with these manifestations are exquisitely photosensitive; most have antibodies to Ro (SS-A). Other SLE rashes include recurring urticaria, lichen planus-like dermatitis, bullae, and panniculitis ("lupus profundus"). Rashes can be minor or severe; they may be the major disease manifestation. Small ulcerations on the oral or nasal mucosa are common in SLE; the lesions resemble aphthous ulcers and may or may not be painful.

RENAL MANIFESTATIONS

Nephritis is usually the most serious manifestation of SLE, particularly because nephritis and infection are the leading causes of mortality in the first decade of disease. Because nephritis is asymptomatic in most lupus patients, urinalysis should be ordered in any person suspected of having SLE. The classification of lupus nephritis is primarily histologic (see "Pathology," above, and Table 349-2). Renal biopsy is recommended for every SLE patient with any clinical evidence of nephritis; results are used to plan current and near-future therapies. Patients with dangerous proliferative forms of glomerular damage (ISN III and IV) usually have microscopic hematuria and proteinuria (>500 mg per 24 h); approximately one-half develop nephrotic syndrome, and most develop hypertension. If diffuse proliferative glomerulonephritis (DPGN) is inadequately treated, virtually all patients develop ESRD within 2 years of diagnosis. Therefore, aggressive immunosuppression is indicated (usually systemic glucocorticoids plus another immunosuppressive drug), unless damage is irreversible

(Fig. 349-2, **Table 349-5**). African Americans are more likely to develop ESRD than are whites, even with the most current therapies. Overall in the United States, ~20% of individuals with lupus DPGN die or develop ESRD within 10 years of diagnosis. Such individuals require aggressive control of SLE and of the complications of renal disease and of therapy. Approximately 20% of SLE patients with proteinuria (usually nephrotic) have membranous glomerular changes without proliferative changes on renal biopsy. Their outcome is better than for those with DPGN, but patients with class V and nephrotic range proteinuria should be treated in the same way as those with classes III or IV proliferative disease. Lupus nephritis tends to be an ongoing disease, with flares requiring re-treatment or increased treatment over many years. For most people with lupus nephritis, accelerated atherosclerosis becomes important after several years of disease; attention must be given to control of systemic inflammation, blood pressure, hyperlipidemia, and hyperglycemia.

TABLE 349-5

Medications for the Management of SLE

Medication	Dose Range	Drug Interactions	Serious or Common Adverse Effects
NSAIDs, salicylates (Ecotrin[a] and St. Joseph's aspirin[a] approved by FDA for use in SLE)	Doses toward upper limit of recommended range usually required	A2R/ACE inhibitors, glucocorticoids, fluconazole, methotrexate, thiazides	NSAIDs: Higher incidence of aseptic meningitis, elevated liver enzymes, decreased renal function, vasculitis of skin; entire class, especially COX-2-specific inhibitors, may increase risk for myocardial infarction
			Salicylates: ototoxicity, tinnitus
			Both: GI events and symptoms, allergic reactions, dermatitis, dizziness, acute renal failure, edema, hypertension
Topical glucocorticoids	Mid potency for face; mid to high potency for other areas	None known	Atrophy of skin, contact dermatitis, folliculitis, hypopigmentation, infection
Topical sunscreens	SPF 15 at least; 30+ preferred	None known	Contact dermatitis
Hydroxychloroquine[a] (quinacrine can be added or substituted)	200–400 mg qd (100 mg qd); do not exceed 6.5 mg/kg dry weight	None known	Retinal damage, agranulocytosis, aplastic anemia, ataxia, cardiomyopathy, dizziness, myopathy, ototoxicity, peripheral neuropathy, pigmentation of skin, seizures, thrombocytopenia; Quinacrine usually causes diffuse yellow skin coloration
DHEA (dehydroepiandrosterone)	200 mg qd	Unclear	Acne, menstrual irregularities, high serum levels of testosterone
Methotrexate (for dermatitis, arthritis)	10–25 mg once a week, PO or SC, with folic acid; decrease dose if CrCl <60 mL/min	Acitretin, leflunomide, NSAIDs and salicylates, penicillins, probenecid, sulfonamides, trimethoprim	Anemia, bone marrow suppression, leukopenia, thrombocytopenia, hepatotoxicity, nephrotoxicity, infections, neurotoxicity, pulmonary fibrosis, pneumonitis, severe dermatitis, seizures, pseudolymphoma
Glucocorticoids, oral[a] (several specific brands are approved by FDA for use in SLE)	Prednisone, prednisolone: 0.5–1 mg/kg per day for severe SLE 0.07–0.3 mg/kg per day or qod for milder disease	A2R/ACE antagonists, antiarrhythmics class III, cyclosporine, NSAIDs and salicylates, phenothiazines, phenytoins, quinolones, rifampin, risperidone, thiazides, sulfonylureas, warfarin	Infection, VZV infection, hypertension, hyperglycemia, hypokalemia, acne, allergic reactions, anxiety, aseptic necrosis of bone, cushingoid changes, CHF, fragile skin, insomnia, menstrual irregularities, mood swings, osteoporosis, psychosis

Medication	Dose Range	Drug Interactions	Serious or Common Adverse Effects
Methylprednisolone sodium succinate, IV[a] (FDA approved for lupus nephritis)	For severe disease, 0.5-1 g IV qd × 3 days	As for oral glucocorticoids	As for oral glucocorticoids (if used repeatedly); anaphylaxis
Cyclophosphamide[b] IV	Low dose (for whites of northern European backgrounds): 500 mg every 2 weeks for 6 doses, then begin maintenance with MMF or AZA. High dose: 7–25 mg/kg q month × 6; consider mesna administration with dose	Allopurinol, bone marrow suppressants, colony-stimulating factors, doxorubicin, rituximab, succinylcholine, zidovudine	Infection, VZV infection, bone marrow suppression, leukopenia, anemia, thrombocytopenia, hemorrhagic cystitis (less with IV), carcinoma of the bladder, alopecia, nausea, diarrhea, malaise, malignancy, ovarian and testicular failure. Ovarian failure is probably not a problem with low dose.
Oral	1.5–3 mg/kg per day; decrease dose for CrCl <25 mL/min		
Mycophenolate mofetil (MMF)[b] or mycophenolic acid (MPA)	MMF: 2–3 g/d PO total given bid for induction therapy, 1–2 g/d total given bid for maintenance therapy; max 1 g bid if CrCl <25 mL/min. Begin with low dose and increase every 1–2 weeks to minimize GI side effects. Start treatment at 0.5 g bid. MPA: 360–1080 mg bid; caution if CrCl <25 mL/min	Acyclovir, antacids, azathioprine, bile acid-binding resins, ganciclovir, iron, salts, probenecid, oral contraceptives	Infection, leukopenia, anemia, thrombocytopenia, lymphoma, lymphoproliferative disorders, malignancy, alopecia, cough, diarrhea, fever, GI symptoms, headache, hypertension, hypercholesterolemia, hypokalemia, insomnia, peripheral edema, elevated liver enzymes, tremor, rash. Limited date suggests Asians should begin treatment with doses not exceeding 2 g daily to reduce adverse events.
Azathioprine (AZA)[b]	2–3 mg/kg per day PO for induction; 1–2 mg/kg per day for maintenance; decrease frequency of dose if CrCl <50 mL/min	ACE inhibitors, allopurinol, bone marrow suppressants, interferons, mycophenolate mofetil, rituximab, warfarin, zidovudine	Infection, VZV infection, bone marrow suppression, leukopenia, anemia, thrombocytopenia, pancreatitis, hepatotoxicity, malignancy, alopecia, fever, flulike illness, GI symptoms
Belimumab	10 mg/kg IV wks 0, 2, and 4, then monthly OR subcutaneous 200 mg each week	IVIg	Infusion reactions, allergy, infections. Headache and diffuse body aching.
Rituximab (for patients resistant to above therapies)	375 mg/m^2 q wk × 4 or 1 g q 2 wks × 2	IVIg	Infection (including PML), infusion reactions, headache, arrhythmias, allergic responses
Tacrolimus	Trough blood level should not exceed 5.5 ng/mL to minimize toxicity. Begin dose at 2 mg bid		Infection, nephrotoxicity, neural toxicity

[a]Indicates medication is approved for use in SLE by the U.S. Food and Drug Administration. [b]Indicates the medication has been used with glucocorticoids in the trials showing efficacy.

Abbreviations: A2R, angiotensin II receptor; ACE, angiotensin-converting enzyme; CHF, congestive heart failure; CrCl, creatinine clearance; FDA, U.S. Food and Drug Administration; GI, gastrointestinal; IVIg, intravenous immunoglobulin; NSAIDs, nonsteroidal anti-inflammatory drugs; PML, progressive multifocal leukoencephalopathy; SLE, systemic lupus erythematosus; SPF, sun protection factor; VZV, varicella-zoster virus.

NERVOUS SYSTEM MANIFESTATIONS

There are many central nervous system (CNS) and peripheral nervous system manifestations of SLE; in some patients, these are the major cause of morbidity and mortality. It is useful to approach this diagnostically by asking first whether the symptoms result from SLE or another condition (such as infection in immunosuppressed individuals or side effects of therapies). If symptoms are related to SLE, it should be determined whether they are caused by a diffuse process (requiring immunosuppression) or vascular occlusive disease (requiring anticoagulation). The most common manifestation of diffuse CNS lupus is cognitive dysfunction, including difficulties with memory and reasoning. Headaches are also common. When excruciating, they often indicate SLE flare; when milder, they are difficult to distinguish from migraine or tension headaches. Seizures of any type may be caused by lupus; treatment often requires both antiseizure and immunosuppressive therapies. Psychosis can be the dominant manifestation of SLE; it must be distinguished from glucocorticoid-induced psychosis. The latter usually occurs in the first weeks of glucocorticoid therapy, at daily doses of ≥40 mg of prednisone or equivalent; psychosis resolves over several days after glucocorticoids are decreased or stopped. Myelopathy is not rare and is often disabling; rapid initiation of immunosuppressive therapy starting with high-dose glucocorticoids is standard of care.

VASCULAR OCCLUSIONS INCLUDING STROKE AND MYOCARDIAL INFARCTIONS

The prevalence of transient ischemic attacks, strokes, and myocardial infarctions is increased in patients with SLE. These vascular events are increased in, but not exclusive to, SLE patients with antibodies to phospholipids (antiphospholipid antibodies), which are associated with hypercoagulability and acute thrombotic events (Chap. 350). Ischemia in the brain can be caused by focal occlusion (either noninflammatory or associated with vasculitis) or by embolization from carotid artery plaque or from fibrinous vegetations of Libman-Sacks endocarditis. Appropriate tests for antiphospholipid antibodies (see below) and for sources of emboli should be ordered in such patients to estimate the need for, intensity of, and duration of anti-inflammatory and/or anticoagulant therapies. When it is most likely that a cerebral event results from clotting, long-term anticoagulation is the therapy of choice. Two processes can occur at once—vasculitis plus bland vascular occlusions—in which case it is appropriate to treat with anticoagulation plus immunosuppression.

In SLE, myocardial infarctions are primarily manifestations of accelerated atherosclerosis. The increased risk for vascular events is three- to tenfold overall, and is highest in women aged <49 years. Characteristics associated with increased risk for atherosclerosis include male gender, older age, hypertension, dyslipidemia, diabetes, dysfunctional proinflammatory high-density lipoproteins, repeated high scores for disease activity, high cumulative or daily doses of glucocorticoids, and high serum levels of homocysteine and leptin. Statin therapies reduce levels of low-density lipoproteins (LDL) in SLE patients; significant reduction of cardiac events by statins has been shown in SLE patients with renal transplants and recently in an epidemiologic study of large number of patients in Taiwan.

PULMONARY MANIFESTATIONS

The most common pulmonary manifestation of SLE is pleuritis with or without pleural effusion. This manifestation, when mild, may respond to treatment with nonsteroidal anti-inflammatory drugs (NSAIDs); when more severe, patients require a brief course of glucocorticoid therapy. Pulmonary infiltrates also occur as a manifestation of active SLE and are difficult to distinguish from infection on imaging studies. Life-threatening pulmonary manifestations include interstitial inflammation leading to fibrosis, shrinking lung syndrome, and intraalveolar hemorrhage; all of these probably require early aggressive immunosuppressive therapy as well as supportive care. Pulmonary arterial hypertension occurs in a small proportion of SLE patients and should be treated in the same way as idiopathic pulmonary hypertension.

CARDIAC MANIFESTATIONS

Pericarditis is the most frequent cardiac manifestation; it usually responds to anti-inflammatory therapy and infrequently leads to tamponade. More serious cardiac manifestations are myocarditis and fibrinous endocarditis of Libman-Sacks. The endocardial involvement can lead to valvular insufficiencies, most commonly of the mitral or aortic valves, or to embolic events. It has not been proven that glucocorticoid or other immunosuppressive therapies lead to improvement of lupus myocarditis or endocarditis, but it is usual practice to

administer a trial of high-dose steroids along with appropriate supportive therapy for heart failure, arrhythmia, or embolic events. As discussed above, patients with SLE are at increased risk for myocardial infarction, usually due to accelerated atherosclerosis, which probably results from immune attack, chronic inflammation, and/or chronic oxidative damage to arteries.

HEMATOLOGIC MANIFESTATIONS

The most frequent hematologic manifestation of SLE is anemia, usually normochromic normocytic, reflecting chronic illness. Hemolysis can be rapid in onset and severe, requiring high-dose glucocorticoid therapy, which is effective in most patients. Leukopenia is also common and almost always consists of lymphopenia, not granulocytopenia; lymphopenia rarely predisposes to infections and by itself usually does not require therapy. Thrombocytopenia may be a recurring problem. If platelet counts are >40,000/μL and abnormal bleeding is absent, therapy may not be required. High-dose glucocorticoid therapy (e.g., 1 mg/kg per day of prednisone or equivalent) is usually effective for the first few episodes of severe thrombocytopenia. Recurring or prolonged hemolytic anemia or thrombocytopenia, or disease requiring an unacceptably high dose of daily glucocorticoids, should be treated with additional strategies such as rituximab, platelet growth factors, and/or splenectomy (see "Management of Systemic Lupus Erythematosus" below).

GASTROINTESTINAL MANIFESTATIONS

Nausea, sometimes with vomiting, and diarrhea can be manifestations of an SLE flare, as can diffuse abdominal pain probably caused by autoimmune peritonitis and/or intestinal vasculitis. Increases in serum aspartate aminotransferase (AST) and alanine aminotransferase (ALT) are common when SLE is active. These manifestations usually improve promptly during systemic glucocorticoid therapy. Vasculitis involving the intestine may be life-threatening; perforations, ischemia, bleeding, and sepsis are frequent complications. Aggressive immunosuppressive therapy with high-dose glucocorticoids is recommended for short-term control; evidence of recurrence is an indication for additional therapies.

OCULAR MANIFESTATIONS

Sicca syndrome (Sjögren's syndrome; **Chap. 354**) and nonspecific conjunctivitis are common in SLE and rarely threaten vision. In contrast, retinal vasculitis and optic neuritis are serious manifestations: blindness can develop over days to weeks. Aggressive immunosuppression is recommended, although there are no controlled trials to prove effectiveness. Complications of systemic and intraorbital glucocorticoid therapy include cataracts (common) and glaucoma.

LABORATORY TESTS

Laboratory tests serve (1) to establish or rule out the diagnosis; (2) to follow the course of disease, particularly to suggest that a flare is occurring or organ damage is developing; and (3) to identify adverse effects of therapies.

TESTS FOR AUTOANTIBODIES

Diagnostically, the most important autoantibodies to detect are ANA because the test is positive in >95% of patients, usually at the onset of symptoms (Tables 349-1 AND 349-3). A few patients develop ANA within 1 year of symptom onset; repeated testing may thus be useful. ANA tests using immunofluorescent methods are more reliable than enzyme-linked immunosorbent assays (ELISAs) and/or bead assays, which have less specificity. ANA-negative lupus exists but is rare in adults and is usually associated with other autoantibodies (anti-Ro or anti-DNA). High-titer IgG antibodies to double-stranded DNA (dsDNA) (but not to single-stranded DNA) are specific for SLE. ELISA and immunofluorescent reactions of sera with the dsDNA in the flagellate *Crithidia luciliae* have ~60% sensitivity for SLE. Titers of anti-dsDNA vary over time. In some patients, increases in quantities of anti-dsDNA herald a flare, particularly of nephritis or vasculitis, especially when associated with declining levels of C3 or C4 complement. Antibodies to Sm are also specific for SLE and assist in diagnosis; anti-Sm antibodies do not usually correlate with disease activity or clinical manifestations. Antiphospholipid antibodies are not specific for SLE, but their presence fulfills one classification criterion, and they identify patients at increased risk for venous or arterial clotting, thrombocytopenia, and fetal loss. There are three widely accepted tests that measure different antibodies (anticardiolipin, anti-β_2-glycoprotein, and the lupus anticoagulant). ELISA is used for anticardiolipin and anti-β_2-glycoprotein (both internationally standardized with good reproducibility); a sensitive phospholipid-based activated prothrombin time such as the dilute Russell venom viper test is used to identify the lupus anticoagulant. The higher the titers of IgG anticardiolipin (>40 IU is considered high), and the greater the number of different antiphospholipid antibodies that are detected, the greater is the risk for a clinical episode of clotting. Quantities of antiphospholipid

antibodies may vary markedly over time; repeated testing is justified if clinical manifestations of the APS appear (Chap. 350). To classify a patient as having APS, with or without SLE, by international criteria requires the presence of one or more clotting episodes and/or repeated fetal losses plus at least two positive tests for antiphospholipid antibodies, at least 12 weeks apart; however, many patients with APS do not meet these stringent criteria, which are intended for inclusion of patients into studies.

An additional autoantibody test with predictive value (not used for diagnosis) detects anti-Ro/SS-A, which indicates increased risk for neonatal lupus, sicca syndrome, and SCLE. Women with child-bearing potential and SLE should be screened for antiphospholipid antibodies and anti-Ro, because both antibodies have the potential to cause fetal harm.

STANDARD TESTS FOR DIAGNOSIS

Screening tests for complete blood count, platelet count, and urinalysis may detect abnormalities that contribute to the diagnosis and influence management decisions.

TESTS FOR FOLLOWING DISEASE COURSE

It is useful to follow tests that indicate the status of organ involvement known to be present during SLE flares. These might include urinalysis for hematuria and proteinuria, hemoglobin levels, platelet counts, and serum levels of creatinine or albumin. There is great interest in identification of additional markers of disease activity. Candidates include levels of anti-DNA and anti-C1q antibodies, several components of complement (C3 is most widely available), activated complement products (an assay is commercially available that measures binding to the C4d receptor on erythrocytes and B cells), IFN-inducible gene expression in peripheral blood cells, serum levels of BLyS (B lymphocyte stimulator, also called BAFF), and urinary levels of TNF-like weak inducer of apoptosis (TWEAK), neutrophil gelatinase-associated lipocalin (NGAL), or monocyte chemotactic protein 1 (MCP-1). None is uniformly agreed upon as a reliable indicator of flare or of response to therapeutic interventions. It is likely that a panel of multiple proteins and nuclear products (and possibly levels of selected miRNAs and methylation profiles of DNA) will be developed to predict both impending flare and response to recently instituted therapies. Increased quantities of plasma cells, and increased expression of their gene signatures in whole blood, are associated with active disease and flares, but measurements are not commercially available. For now, the physician should determine for each patient whether certain available laboratory test changes predict flare (falling complement, rising anti-DNA, increased proteinuria, worsening anemia, etc.). If so, altering therapy in response to these changes may be advisable (30 mg of prednisone daily for 2 weeks has been shown to prevent flares in patients with rising anti-DNA plus falling complement). In addition, given the increased prevalence of atherosclerosis in SLE, it is advisable to follow the recommendations of the National Cholesterol Education Program for testing and treatment, including scoring of SLE as an independent risk factor, similar to diabetes mellitus.

MANAGEMENT OF SYSTEMIC LUPUS ERYTHEMATOSUS

There is no cure for SLE, and complete sustained remissions are rare. There is an international effort to encourage practitioners and patients to aim for low-level disease activity (mild symptoms on the lowest possible doses of medications) which can be achieved for at least a year in 30–50% of SLE patients. Therefore, the physician should plan to induce remissions of acute flares and then maintain improvements with strategies that suppress symptoms to an acceptable level and prevent organ damage. Therapeutic choices depend on (1) whether disease manifestations are life-threatening or likely to cause organ damage, justifying aggressive therapies; (2) whether manifestations are potentially reversible; and (3) the best approaches to preventing complications of disease and its treatments. Therapies, doses, and adverse effects are listed in Table 349-5.

CONSERVATIVE THERAPIES FOR MANAGEMENT OF NON-LIFE-THREATENING DISEASE

Among patients with fatigue, pain, and autoantibodies indicative of SLE, but without major organ involvement, management can be directed to suppression of symptoms. Analgesics and antimalarials are mainstays. NSAIDs are useful analgesics/anti-inflammatories, particularly for arthritis/arthralgias. However, two major issues indicate caution in using NSAIDs. First, SLE patients compared with the general population are at increased risk for NSAID-induced aseptic meningitis, elevated serum transaminases, hypertension, and renal dysfunction. Second, all NSAIDs, particularly those that inhibit cyclooxygenase-2 specifically, may increase risk for myocardial infarction. Acetaminophen to control pain may be a good strategy, but NSAIDs are more effective in some patients. The relative hazards of NSAIDs compared with low-dose glucocorticoid therapy have not been established. Antimalarials (hydroxychloroquine, chloroquine, and quinacrine) often reduce dermatitis, arthritis, and fatigue. A randomized, placebo-controlled, prospective trial has shown that withdrawal of hydroxychloroquine results in increased numbers of disease flares; hydroxychloroquine also reduces accrual of tissue damage, including renal damage, over time. Some

experts recommend a hydroxychloroquine blood level of ≥750 ng/mL to optimize responses in active SLE; after achieving response doses should be reduced. Because of potential retinal toxicity (occurring in 6% of patients after cumulative doses of 1000 g, ~5 years of continuing therapy), patients receiving antimalarials should undergo ophthalmologic examinations annually. A placebo-controlled prospective trial suggests that administration of dehydroepiandrosterone may reduce disease activity. If quality of life is inadequate despite these conservative measures, treatment with low doses of systemic glucocorticoids may be necessary. Belimumab is effective for 50% of patients with fatigue, rash, and/or the arthritis of SLE; it is expensive and should be considered after other approaches fail or are not tolerated. SLE patients most likely to respond to belimumab have robust clinical activity (a Systemic Lupus Erythematosus Disease Activity Index [SLEDAI] score of ≥10), positive anti-DNA, and low serum complement. SLEDAI is a widely used measure of SLE disease activity; scores >3 reflect clinically active disease. Lupus dermatitis should be managed with topical sunscreens, antimalarials, topical glucocorticoids and/or tacrolimus, and if severe or unresponsive, systemic glucocorticoids with or without mycophenolate mofetil, azathioprine, or belimumab.

LIFE-THREATENING SLE: PROLIFERATIVE FORMS OF LUPUS NEPHRITIS

Guidelines for management of lupus nephritis have been published recently by the American College of Rheumatology and the European League Against Rheumatism (encompassed and referenced in Fig. 349-2 and Table 349-5). The mainstay of treatment for any inflammatory life-threatening or organ-threatening manifestations of SLE is systemic glucocorticoids (0.5–1 mg/kg per day PO or 500–1000 mg of methylprednisolone sodium succinate IV daily for 3 days followed by 0.5–1 mg/kg of daily prednisone or equivalent). Evidence that glucocorticoid therapy is life-saving comes from retrospective studies from the predialysis era; survival was significantly better in people with DPGN treated with high-dose daily glucocorticoids (40–60 mg of prednisone daily for 4–6 months) versus lower doses. Currently, high doses are recommended for much shorter periods; recent trials of interventions for severe SLE use 4–6 weeks of 0.5–1 mg/kg per day of prednisone or equivalent. Thereafter, doses are tapered as rapidly as the clinical situation permits, usually to a maintenance dose ranging from 5 to 10 mg of prednisone or equivalent per day. Most patients with an episode of severe SLE require many years of maintenance therapy with low-dose glucocorticoids, which can be increased to prevent or treat disease flares. Frequent attempts to gradually reduce the glucocorticoid requirement are recommended because virtually everyone develops important adverse effects (Table 349-5). High-quality clinical studies regarding initiating therapy for severe, active SLE with IV pulses of high-dose glucocorticoids are not available. Most recent clinical trials in lupus nephritis have initiated therapy with high-dose IV glucocorticoid pulses (500–1000 mg daily for 3–5 days). This approach must be tempered by safety considerations, such as the presence of conditions adversely affected by glucocorticoids (e.g., infection, hyperglycemia, hypertension, osteoporosis). A current clinical trial is evaluating mycophenolate mofetil plus rituximab without maintenance daily glucocorticoids to treat lupus nephritis: if results are positive, the paradigm for short-term and long-term management of SLE is likely to change.

Cytotoxic/immunosuppressive agents added to glucocorticoids are recommended to treat serious SLE. Almost all prospective controlled trials in SLE involving cytotoxic agents have been conducted in combination with glucocorticoids in patients with lupus nephritis. Therefore, the following recommendations apply to treatment of nephritis. Either cyclophosphamide (an alkylating agent) or mycophenolate mofetil (a relatively lymphocyte-specific inhibitor of inosine monophosphatase and therefore of purine synthesis) is an acceptable choice for induction of improvement in severely ill patients; azathioprine (a purine analogue and cycle-specific antimetabolite) may be effective but is associated with more flares. In patients whose renal biopsies show ISN grade III or IV disease, early treatment with combinations of glucocorticoids and cyclophosphamide reduces progression to ESRD and death. Shorter-term studies with glucocorticoids plus mycophenolate mofetil (prospective randomized trials of 6 months, follow-up studies of 5 years) show that this regimen is similar to cyclophosphamide in achieving improvement. Comparisons are complicated by effects of race, since higher proportions of African Americans (and other non-Asian, nonwhite races) respond to mycophenolate than to cyclophosphamide, whereas similar proportions of whites and Asians respond to each drug. Regarding toxicity, diarrhea is more common with mycophenolate mofetil; amenorrhea, leukopenia, and nausea are more common with high dose cyclophosphamide. Importantly, rates of severe infections and death are similar in meta-analyses. Two different regimens of IV cyclophosphamide are available. For white patients with northern European backgrounds, low doses of cyclophosphamide (500 mg every 2 weeks for six total doses, followed by azathioprine or mycophenolate maintenance) are as effective as standard high doses, with less toxicity. Ten-year follow-up has shown no differences between the high-dose and low-dose groups (death or ESRD in 9–20% of patients in each group). It is not clear whether the data apply to U.S. populations, especially African Americans and Latinas. High-dose cyclophosphamide (500–1000 mg/m^2 body surface area given monthly IV for 6 months, followed by azathioprine or mycophenolate maintenance) is an acceptable approach for patients with severe nephritis (e.g., multiple cellular crescents and/or fibrinoid necrosis on renal biopsy, or rapidly progressive glomerulonephritis). Cyclophosphamide and mycophenolate responses begin 3–16 weeks after treatment is initiated, whereas glucocorticoid responses may begin within 24 h.

For maintenance therapy, mycophenolate and azathioprine probably are similar in efficacy and toxicity; both are safer than cyclophosphamide. In a recently published multicenter study, mycophenolate was superior to azathioprine in maintaining renal function and survival in patients who responded to induction therapy with either cyclophosphamide or mycophenolate. The incidence of ovarian failure, a common effect of high-dose cyclophosphamide therapy (but probably not of low-dose therapy), can be reduced by treatment with a gonadotropin-releasing hormone agonist (e.g., leuprolide 3.75 mg intramuscularly) prior to each monthly cyclophosphamide dose. Patients with high serum creatinine levels (e.g., ≥265 μmol/L [≥3.0 mg/dL]) many months in duration and high chronicity scores on renal biopsy are not likely to respond to any of these therapies. In general, it may be better to induce improvement in an African-American or Hispanic patient with proliferative glomerulonephritis with mycophenolate mofetil (2–3 g daily) rather than cyclophosphamide, with the option to switch if no evidence of response is detectable after 3–6 months of treatment. For whites and Asians, induction with either mycophenolate mofetil or cyclophosphamide is acceptable. Cyclophosphamide may be discontinued when it is clear that a patient is improving. The number of SLE flares is reduced by maintenance therapy with mycophenolate mofetil (1.5–2 g daily) or azathioprine (1–2.5 mg/kg per day). Both cyclophosphamide and mycophenolate mofetil are potentially teratogenic; patients should be off either medication for at least 3 months before attempting to conceive. Azathioprine can be used if necessary to control active SLE in patients who are pregnant. If azathioprine is used either for induction or maintenance therapy, patients may be prescreened for homozygous deficiency of the TMPT enzyme (which is required to metabolize the 6-mercaptopurine product of azathioprine) because they are at higher risk for bone marrow suppression.

Good improvement occurs in ~80% of lupus nephritis patients receiving either cyclophosphamide or mycophenolate at 1–2 years of follow-up. However, in some studies, at least 50% of these individuals have flares of nephritis over the next 5 years, and re-treatment is required; such individuals are more likely to progress to ESRD. Long-term outcome of lupus nephritis to most interventions is better in whites than in African Americans. Methotrexate (a folinic acid antagonist) may have a role in the treatment of arthritis and dermatitis but probably not in nephritis or other life-threatening disease. Small controlled trials (in Asia) of leflunomide, a relatively lymphocyte-specific pyrimidine antagonist licensed for use in rheumatoid arthritis, have suggested it can suppress disease activity in some SLE patients. Cyclosporine and tacrolimus, which inhibit calcium flux and therefore production of IL-2 and T lymphocyte functions, have not been studied in prospective controlled trials in SLE in the United States; several studies in Asia have shown they are effective in lupus nephritis. A recent trial in China showed that a combination of low dose mycophenolate mofetil (one gram daily) plus tacrolimus (4 mg daily) plus prednisone (pulse followed by 0.6 mg/kg/d) had a better response rate than high dose intravenous cyclophosphamide. Because calcineurin blockers have potential nephrotoxicity but little bone marrow toxicity, the author uses them for periods of a few months in patients with steroid-resistant cytopenias of SLE, in steroid-resistant patients who have developed bone marrow suppression from standard cytotoxic agents, or in patients with active SLE in spite of treatment with mycophenolate or cyclosphosphamide.

Most patients with SLE of any type should be treated with hydroxychloroquine since it prevents damage in skin and kidney and reduces overall damage scores. Patients with proteinuria > 500 mg daily should receive ACE inhibitors or ARBs, as they reduce the chance for ESRD.

Use of biologics directed against B cells for active SLE is under intense study. Use of anti-CD20 (rituximab), particularly in patients with SLE who are resistant to the more standard combination therapies discussed above, is controversial. Several open trials have shown efficacy in a majority of such patients, both for nephritis and for extrarenal lupus. However, recent prospective placebo-controlled randomized trials, one in renal and one in nonrenal SLE, did not show a difference between anti-CD20 and placebo when added to standard combination therapies. Belimumab, which is approved by the FDA for use in SLE without active renal disease (indication is serologically positive SLE which has failed standard treatments), is in clinical trials for active lupus nephritis. Drugs that kill plasma cells, used in multiple myeloma, are being studied in SLE, as are molecules and antibodies that prevent activation of B cells and/or T cells, such as Jak/Stat inhibitors.

SPECIAL CONDITIONS IN SLE THAT MAY REQUIRE ADDITIONAL OR DIFFERENT THERAPIES

Crescentic Lupus Nephritis

The presence of cellular or fibrotic crescents in glomeruli with proliferative glomerulonephritis indicates a worse prognosis than in patients without this feature. There are no large prospective multinational controlled trials showing efficacy of cyclophosphamide, mycophenolate, cyclosporine, or tacrolimus in such cases. Most authorities recommend high-dose cyclophosphamide as the induction therapy of choice; there is some evidence that high dose mycophenolate mofetil is equally effective.

Membranous Lupus Nephritis

Most SLE patients with membranous (INS-V) nephritis also have proliferative changes and should be treated for proliferative disease. However, some have pure membranous changes. Treatment for this group is less well defined. Some authorities do not recommend

immunosuppression unless proteinuria is in the nephrotic range (although treatment with angiotensin-converting enzyme inhibitors or angiotensin II receptor blockers is recommended). In those patients, recent prospective controlled trials suggest that alternate-day glucocorticoids plus cyclophosphamide or mycophenolate mofetil or cyclosporine are all effective in the majority of patients in reducing proteinuria. It is more controversial whether they preserve renal function over the long term.

Pregnancy and Lupus

Fertility rates for men and women with SLE are probably normal. However, rate of fetal loss is increased (approximately two- to threefold) in women with SLE. Fetal demise is higher in mothers with high disease activity, antiphospholipid antibodies (especially the lupus anticoagulant), hypertension, and/or active nephritis. Suppression of disease activity can be achieved by administration of systemic glucocorticoids. A placental enzyme, 11-β-dehydrogenase 2, deactivates glucocorticoids; it is more effective in deactivating prednisone and prednisolone than the fluorinated glucocorticoids dexamethasone and betamethasone. Glucocorticoids are listed by the FDA as pregnancy category A (no evidence of teratogenicity in human studies); cyclosporine, tacrolimus, and rituximab are listed as category C (may be teratogenic in animals but no good evidence in humans); azathioprine, hydroxychloroquine, mycophenolate mofetil, and cyclophosphamide are category D (there is evidence of teratogenicity in humans, but benefits might outweigh risks in certain situations); and methotrexate is category X (risks outweigh benefits). Therefore, active SLE in pregnant women should be controlled with hydroxychloroquine and, if necessary, prednisone/prednisolone at the lowest effective doses for the shortest time required. Azathioprine may be added if these treatments do not suppress disease activity. Adverse effects of prenatal glucocorticoid exposure (primarily betamethasone, which is not recommended) on offspring may include low birth weight, developmental abnormalities in the CNS, and predilection toward adult metabolic syndrome. It is likely that each of these glucocorticoids and immunosuppressive medications gets into breast milk, at least in low levels; patients should consider not breastfeeding if they need therapy for SLE. In SLE patients with antiphospholipid antibodies and prior fetal losses, treatment with heparin (usually low-molecular-weight) plus low-dose aspirin has been shown in prospective controlled trials to increase significantly the proportion of live births. Aspirin alone may be used, although most consider it less effective than heparin-plus-aspirin. Warfarin is teratogenic. Studies with oral thrombin and Factor Xa inhibitors are in progress for APSs; however, their role in preventing fetal loss is not established. An additional potential problem for the fetus is the presence of antibodies to Ro, sometimes associated with neonatal lupus consisting of rash and congenital heart block with or without cardiomyopathy. The cardiac manifestations can be life-threatening; therefore the presence of anti-Ro requires vigilant monitoring of fetal heart rates with prompt intervention (delivery if possible) if distress occurs. Recent evidence shows that hydroxychloroquine treatment of an anti-Ro-positive mother whose infant develops congenital heart block significantly reduces the chance that subsequent fetuses will develop heart block. There is some evidence that dexamethasone treatment of a mother in whom first- or second-degree heart block is detected in utero sometimes prevents progression of heart block. Women with SLE usually tolerate pregnancy without disease flares. However, a small proportion develops severe flares requiring aggressive glucocorticoid therapy or early delivery.

Lupus and Antiphospholipid Syndrome

Patients with SLE who have venous or arterial clotting and/or repeated fetal losses and at least two positive tests for antiphospholipid antibodies have APS and should be managed with long-term anticoagulation **(Chap. 350)**. With warfarin, a target international normalized ratio (INR) of 2.0–2.5 is recommended for patients with one episode of venous clotting; an INR of 3.0–3.5 is recommended for patients with recurring clots or arterial clotting, particularly in the CNS. Recommendations are based on both retrospective and prospective studies of posttreatment clotting events and adverse effects from anticoagulation. Thrombin and Factor Xa inhibitors are under study.

Microvascular Thrombotic Crisis (Thrombotic Thrombocytopenic Purpura, Hemolytic-Uremic Syndrome)

This syndrome of hemolysis, thrombocytopenia, and microvascular thrombosis in kidneys, brain, and other tissues carries a high mortality rate and occurs most commonly in young individuals with lupus nephritis. The most useful laboratory tests are identification of schistocytes on peripheral blood smears, elevated serum levels of lactate dehydrogenase, and antibodies to ADAMS13. Plasma exchange or extensive plasmapheresis is usually life-saving; most authorities recommend concomitant glucocorticoid therapy; there is no evidence that cytotoxic drugs are effective.

Lupus Dermatitis

Patients with any form of lupus dermatitis should minimize exposure to ultraviolet light, using appropriate clothing and sunscreens with a sun protection factor of at least 30. Topical glucocorticoids and antimalarials (such as hydroxychloroquine) are effective in reducing lesion severity in most patients and are relatively safe. Systemic treatment with retinoic acid is a useful strategy in patients with inadequate

improvement on topical glucocorticoids and antimalarials; adverse effects are potentially severe (particularly fetal abnormalities), and there are stringent reporting requirements for its use in the United States. Extensive, pruritic, bullous, or ulcerating dermatitides usually improve promptly after institution of systemic glucocorticoids; tapering may be accompanied by flare of lesions, thus necessitating use of a second medication such as hydroxychloroquine, retinoids, or belimumab. Cytotoxic medications such as methotrexate, azathioprine, or mycophenolate mofetil may also be effective. In therapy-resistant lupus dermatitis there are reports of success with topical tacrolimus (caution must be exerted because of the possible increased risk for malignancies) or with systemic dapsone or thalidomide (the extreme danger of fetal deformities from thalidomide requires permission from and supervision by the supplier; peripheral neuropathy is also common).

PREVENTIVE THERAPIES

Prevention of complications of SLE and its therapy include providing appropriate vaccinations (the administration of influenza and pneumococcal vaccines has been studied in patients with SLE; flare rates are similar to those receiving placebo) and suppressing recurrent urinary tract infections. In patients receiving glucocorticoids, the higher the daily dose the lower the immune response to vaccination; however, the great majority of patients achieve protective levels. Vaccination with attenuated live viruses is generally discouraged in patients who are immunosuppressed; however, a recent study of vaccination of a small number of SLE patient with zostavax showed safety and efficacy. Strategies to prevent osteoporosis should be initiated in most patients likely to require long-term glucocorticoid therapy and/or with other predisposing factors. Postmenopausal women can be partially protected from steroid-induced osteoporosis with calcium supplementation, Vitamin D, and either bisphosphonates or denosumab. Safety of long-term use of these strategies in premenopausal women is not well established. Control of hypertension and appropriate prevention strategies for atherosclerosis, including monitoring and treatment of dyslipidemias, management of hyperglycemia, and management of obesity, are recommended. There is increasing evidence that statin therapies can reduce deaths from cardiac events in SLE patients. Finally, the physician must keep in mind that some cancers are increased in SLE patients including non-Hodgkin lymphomas and cancers of thyroid, lung, liver, and vulvar/vaginal tissues.

EXPERIMENTAL THERAPIES

Studies of highly targeted experimental therapies for SLE are in progress. They include (1) inhibition of IFN-α, which was promising in phase II clinical trials, (2) inhibition of IL12 and IL23 signaling; (3) inhibition of IL17; (4) inhibition of IL-6; (5) elimination of plasma cells; (6) inhibition of B/T cell second signal coactivation with CTLA-Ig or anti-CD40L; (7) inhibition of innate immune activation via TLR7 or TLR7 and 9; (8) induction of regulatory T cells with peptides from immunoglobulins or autoantigens or with low doses if IL2; and (9) inhibition of lymphocyte activation by blockade of Jak/Stat. A few studies have used vigorous untargeted immunosuppression with high-dose cyclophosphamide plus anti-T cell strategies, with rescue by transplantation of autologous hematopoietic stem cells for the treatment of severe and refractory SLE. One U.S. report showed an estimated mortality rate over 5 years of 15% and sustained remission in 50%. It is hoped that in the next edition of this text, we will be able to recommend more effective and less toxic approaches to treatment of SLE based on some of these strategies.

PATIENT OUTCOMES, PROGNOSIS, AND SURVIVAL


Survival in patients with SLE in the United States, Canada, Europe, and China is ~95% at 5 years, 90% at 10 years, and 78% at 20 years. In the United States, African Americans and Hispanic Americans with a mestizo heritage have a worse prognosis than whites, whereas Africans in Africa and Hispanic Americans with a Puerto Rican origin do not. The relative importance of gene mixtures and environmental differences accounting for ethnic differences is not known. Poor prognosis (~50% mortality in 10 years) in most series is associated with (at the time of diagnosis) high serum creatinine levels (>124 µmol/L [>1.4 mg/dL]), hypertension, nephrotic syndrome (24-h urine protein excretion >2.6 g), anemia (hemoglobin <124 g/L [<12.4 g/dL]), hypoalbuminemia, hypocomplementemia, antiphospholipid antibodies, male sex, ethnicity (African American, Hispanic with mestizo heritage), and low socioeconomic status. Data regarding outcomes in SLE patients with renal transplants show mixed results: some series show a twofold increase in graft rejection compared to patients with other causes of ESRD, whereas others show no differences. Overall patient survival is comparable (85% at 2 years). Lupus nephritis occurs in ~5% of transplanted kidneys. Disability in patients with SLE is common due primarily to chronic fatigue, arthritis, and pain, as well as renal disease. As many as 30–50% of patients may achieve low disease activity (defined as mild activity on hydroxychloroquine with or without low dose glucocorticoids); fewer than 10% experience remissions (defined as no disease activity on no medications). Both of these conditions may persist for a few years, but are usually not permanent, as flares of SLE occur. The leading causes of death in the first decade of disease are systemic disease activity, renal failure, and infections; subsequently, thromboembolic events become increasingly frequent causes of mortality.

DRUG-INDUCED LUPUS

This is a syndrome of positive ANA associated with symptoms such as fever, malaise, arthritis or intense arthralgias/myalgias, serositis, and/or rash. The syndrome appears during therapy with certain medications and biologic agents, is predominant in whites, has less female predilection than SLE, rarely involves kidneys or brain, is rarely associated with anti-dsDNA, is commonly associated with antibodies to histones, and usually resolves over several weeks after discontinuation of the offending medication. The list of substances that can induce lupus-like disease is long. Among the most frequent are the antiarrhythmics procainamide, disopyramide, and propafenone; the anti-hypertensive hydralazine; several angiotensin-converting enzyme inhibitors and beta blockers; the antithyroid propylthiouracil; the antipsychotics chlorpromazine and lithium; the anticonvulsants carbamazepine and phenytoin; the antibiotics isoniazid, minocycline, and nitrofurantoin (Macrodantin); the antirheumatic sulfasalazine; the diuretic hydrochlorothiazide; the antihyperlipidemics lovastatin and simvastatin. Biologics that can cause drug-induced lupus (DIL) include inhibitors of IFNs and TNF. In DIL, ANA usually appears before symptoms; however, many of the medications mentioned above induce ANA in patients who never develop symptoms of drug-induced lupus. It is appropriate to test for ANA at the first hint of relevant symptoms and to use test results to help decide whether to withdraw the suspect agent.

FURTHER READING

Anders HJ, Rovin B: A pathophysiology-based approach to the diagnosis and treatment of lupus nephritis. Kidney Int 90:493, 2016. [PubMed: 27378475]

Hahn BH et al: American College of Rheumatology guidelines for screening, treatment, and management of lupus nephritis. Arthritis Care Res (Hoboken) 64:797, 2012. [PubMed: 22556106]

Hahn BH: Pathogenesis of SLE, in *Dubois Lupus Erythematosus*, 9th ed, DJ Wallace, BH Hahn, (eds). Philadelphia, Elsevier, 2018.

Narain S, Furie R: Update on clinical trials in systemic lupus erythematosus. Curr Opin Rheumatol 28:477, 2016. [PubMed: 27314466]

Teruel M, Alarcon-Riquelme ME: The genetic basis of systemic lupus erythematosus: What are the risk factors and what have we learned. J Autoimmun 74:161, 2016. [PubMed: 27522116]

Chapter 350: Antiphospholipid Syndrome

Haralampos M. Moutsopoulos

DEFINITIONS

Antiphospholipid syndrome (APS) is an autoantibody-mediated acquired thrombophilia characterized by recurrent arterial or venous thrombosis and/or pregnancy morbidity. It affects primarily females. APS may occur alone (primary) or in association with other autoimmune diseases, mainly systemic lupus erythematosus (SLE) (secondary). Catastrophic APS (CAPS) is a life-threatening rapidly progressive thromboembolic disease involving simultaneously three or more organs.

The major autoantibodies detected in patients' sera are directed against phospholipids and/or phospholipid (PL)-binding plasma proteins such as prothrombin and β2 glycoprotein I (β2GPI). PLs are components of the cytoplasmic membrane of all living cells. The antibodies are directed against negatively charged PLs including among others cardiolipin, phosphocholine, and phosphatidylserine. The plasma protein β2GPI is a 43-kDa plasma apolipoprotein, which consists of 326 amino acids arranged in five domains (I through V). Domain V forms a positively charged patch, suitable to interact with negatively charged PLs. In plasma, β2GPI has a circular conformation with domain V binding to and concealing the B cell epitopes lying on domain I. Another group of antibodies termed *lupus anticoagulant* (LA) prolongs clotting times in vitro, which are not corrected by adding normal plasma **(Table 350-1)**. Patients with APS often possess antibodies recognizing *Treponema pallidum* PL/cholesterol complexes, detected by Venereal Disease Research Laboratory (VDRL) tests and characterized as biologic false-positive serologic tests for syphilis (BFP-STS).

TABLE 350-1

Classification and Nomenclature of Antiphospholipid Antibodies

Name	Assay for Their Detection	Comments
Antibodies against cardiolipin (aCL)	Enzyme-linked immunosorbent assay (ELISA) using as antigen cardiolipin (CL), a negatively charged phospholipid	aCL from patients with APS recognize β2GPI existing in the human serum as well as in bovine serum, which is used to block the nonspecific binding sites on the ELISA plate. CL simply stabilizes β2GPI at high concentration on the polystyrene surface.
Antibodies against β2GPI (anti-β2GPI)	ELISA using as antigen affinity purified or recombinant β2GPI in the absence of PL	Antibodies recognize β2GPI bound in the absence of CL to an oxidized polystyrene surface, where oxygen atoms in the moieties C–O or C=O were introduced by γ-irradiation.
Lupus anticoagulant (LA)	Activated partial thromboplastin time (aPTT) Kaolin clotting time (KCT) Dilute Russel viper venom test (DRVVT)	Antibodies recognize β2GPI or prothrombin (PT) and elongate aPTT, implying that they interfere with the generation of thrombin by prothrombin. Prolongation of the clotting times is an in vitro phenomenon, and LA induces thromboses in vivo.

Abbreviations: APL, antiphospholipid syndrome; β2GPI, β2 glycoprotein I; PL, phospholipid.

EPIDEMIOLOGY

The incidence of APS is estimated to be around 5 cases per 100,000 persons per year. Anti-PL antibodies occur in 1–5% of the general population. Their prevalence increases with age; however, it is questionable whether they are able to induce thrombotic events in elderly individuals. Moreover, one-third of patients with SLE and other autoimmune diseases (Chap. 349) possess these antibodies, with only 5–10% of them developing APS.

PATHOGENESIS

The initiating event for the induction of antibodies to PL-binding proteins seems to be infections, oxidative stress, and major physical stresses such as surgery or trauma. All these factors appear to induce increased apoptosis of the vessel endothelial cells and subsequent exposure of PLs. The latter, bound with serum

proteins such as β2GPI or prothrombin, lead to neoantigen formation, which in turn triggers the induction of anti-PLs. The binding of anti-PLs to the disrupted endothelial cells leads to initiation of intravascular coagulation and thrombus formation. Complement activation has been also proposed as a mechanism of APS-related fetal injury.

CLINICAL MANIFESTATIONS AND LABORATORY FINDINGS

Clinical manifestations represent the consequences of venous or arterial thrombosis and/or pregnancy morbidity **(Table 350-2)**. Venous thrombosis, superficial or deep, occurs primarily in the lower extremities often leading to pulmonary emboli. Thrombosis of the pulmonary arteries leads to pulmonary hypertension and thrombosis of the inferior vena cava to Budd-Chiari syndrome. Cerebral venous thrombosis, presents with signs and symptoms of intracranial hypertension and retinal vein thrombosis. Arterial thrombosis affects more commonly the arteries of the brain and is manifested as migraines, cognitive dysfunction, transient ischemic attacks, stroke, and retinal artery occlusion. Arterial thrombosis of the extremities presents with ischemic leg ulcers, digital gangrene, avascular bone necrosis, while thrombosis of other arteries leads to myocardial infarction, renal artery stenosis, glomerular lesions, and infarcts of spleen, pancreas, and adrenals.

Livedo reticularis consists of a mottled reticular vascular pattern that appears as a lace-like, purplish discoloration of the skin. It is probably caused by swelling of the venules due to obstruction of capillaries by thrombi. This clinical manifestation usually occurs together with vascular lesions in the central nervous system and with aseptic bone necrosis. Libman-Sacks endocarditis consists of very small vegetations, histologically characterized by organized platelet-fibrin microthrombi surrounded by growing fibroblasts and macrophages. Glomerular involvement is manifested with hypertension, mildly elevated serum creatinine levels, as well as mild proteinuria/hematuria. Histologically, in an acute phase, thrombotic microangiopathy is present in the glomerular capillaries. In a chronic phase, fibrous intima hyperplasia, fibrous and/or fibrocellular arteriolar occlusions, and focal cortical atrophy are present (Table 350-2). Pregnancy morbidity manifests with increased risk of recurrent miscarriages, intrauterine growth retardation, preeclampsia, eclampsia, and preterm birth. The major causes of these complications are infarctions of the placenta.

TABLE 350-2

Clinical Features of Antiphospholipid Syndrome

Manifestation	%
Venous Thrombosis and Related Consequences	
Deep vein thrombosis	39
Livedo reticularis	24
Pulmonary embolism	14
Superficial thrombophlebitis	12
Thrombosis in various other sites	11
Arterial Thrombosis and Related Consequences	
Stroke	20
Cardiac valve thickening/dysfunction and/or Libman-Sacks vegetations	14
Transient ischemic attack	11
Myocardial ischemia (infarction or angina) and coronary bypass graft thrombosis	10
Leg ulcers and/or digital gangrene	9
Arterial thrombosis in the extremities	7
Retinal artery thrombosis/amaurosis fugax	7
Ischemia of visceral organs or avascular necrosis of bone	6
Multi-infarct dementia	3
Neurologic Manifestations of Uncertain Etiology	
Migraine	20
Epilepsy	7
Chorea	1
Cerebellar ataxia	1
Transverse myelopathy	0.5
Renal Manifestations Due to Various Reasons (Renal Artery/Renal Vein/Glomerular Thrombosis, Fibrous Intima Hyperplasia)	3
Musculoskeletal Manifestations	
Arthralgias	39
Arthritis	27

Manifestation	%
Obstetric Manifestations (Referred to the Number of Pregnancies)	
Preeclampsia Eclampsia	10 4
Fetal Manifestations (Referred to the Number of Pregnancies)	
Early fetal loss (<10 weeks) Late fetal loss (≥10 weeks) Premature birth among the live births	35 17 11
Hematologic Manifestations	
Thrombocytopenia Autoimmune hemolytic anemia	30 10

Source: Adapted from R Cervera et al: Arthritis Rheum 46:1019, 2002.

Premature atherosclerosis has been also recognized as a feature of APS. Musculoskeletal manifestations include in addition to bone necrosis, arthralgia/arthritis, bone marrow necrosis, muscle infarction, non-traumatic fractures, and osteoporosis. Coombs-positive hemolytic anemia and thrombocytopenia are laboratory findings associated with APS.

DIAGNOSIS AND DIFFERENTIAL DIAGNOSIS

The diagnosis of APS should be seriously considered in cases of thrombosis, cerebral vascular accidents in individuals <55 years of age, or pregnancy morbidity in the presence of livedo reticularis or thrombocytopenia. In these cases, aPL antibodies should be measured. The presence of at least one clinical and one laboratory criterion is compatible with the diagnosis, in the absence of other thrombophilia causes. Clinical criteria include: (1) vascular thrombosis defined as one or more clinical episodes of arterial, venous, or small vessel thrombosis in any tissue or organ; and (2) pregnancy morbidity, defined as (a) one or more unexplained deaths of a morphologically normal fetus at or beyond the tenth week of gestation; (b) one or more premature births of a morphologically normal neonate before the thirty-fourth week of gestation because of eclampsia, severe preeclampsia, or placental insufficiency; or (c) three or more unexplained consecutive spontaneous abortions before the tenth week of gestation. Laboratory criteria include (1) LA, (2) anticardiolipin (aCL), and/or (3) anti-β2GPI antibodies, at intermediate or high titers on two occasions, 12 weeks apart.

Differential diagnosis is based on the exclusion of other inherited or acquired causes of thrombophilia **(Chap. 112)**, Coombs-positive hemolytic anemia **(Chap. 96)**, and thrombocytopenia **(Chap. 111)**. Livedo reticularis with or without a painful ulceration on the lower extremities may be also a manifestation of disorders affecting (1) the vascular wall, such as atherosclerosis, polyarteritis nodosa, SLE, cryoglobulinemia, and lymphomas; or (2) the vascular lumen, such as myeloproliferative disorders, hypercholesterolemia, or other causes of thrombophilia.

TREATMENT

TREATMENT

Antiphospholipid Syndrome

After the first thrombotic event, APS patients should be placed on warfarin for life, aiming to achieve an international normalized ratio (INR) ranging from 2.5 to 3.5, alone or in combination with 80 mg of aspirin daily. Pregnancy morbidity is prevented by administering low-molecular-weight heparin with aspirin 80 mg daily. IV immunoglobulin (IVIg) 400 mg/kg every day for 5 days may also prevent abortions, whereas glucocorticoids are ineffective. Patients with aPL in the absence of any clinical event who are simultaneously positive for aCL, anti-β2GPI, and LA or have SLE are at risk of developing thrombotic events which can be prevented by taking aspirin 80 mg and hydroxychloroquine 200 mg daily.

Some patients with APS and patients with CAPS have recurrent thrombotic events despite appropriate anticoagulation. In these cases, IVIg 400 mg/kg every day for 5 days may be of benefit. Patients with CAPS, who are treated in the intensive care unit, are unable to receive warfarin; in this situation, therapeutic doses of low-molecular-weight heparin should be administered. In cases of heparin-induced thrombocytopenia and thrombosis syndrome, inhibitors of phospholipid-bound activated factor X (FXa), such as fondaparinux 7.5 mg SC daily or rivaroxaban 10 mg PO daily, are effective. The above drugs are administered by fixed doses and do not require close monitoring; their safety during the first trimester of pregnancy has not been clearly established.

ACKNOWLEDGMENT

I would like to thank Dr. P. G. Vlachoyiannopoulos for his contribution to the previous edition of the chapter.

FURTHER READING

Andrade D, Tektonidou M: Emerging therapies in antiphospholipid syndrome. Curr Rheumatol Rep 18:22, 2016. [PubMed: 26995745]

Merashli M et al: Antiphospholipid syndrome: An update. Eur J Clin Invest 45:653, 2015. [PubMed: 25851448]

Chapter 351: Rheumatoid Arthritis

Ankoor Shah; E. William St. Clair

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic inflammatory disease of unknown etiology characterized by a symmetric polyarthritis, the most common form of chronic inflammatory arthritis. Since persistently active RA often results in articular cartilage and bone destruction and functional disability, it is vital to diagnose and treat this disease early and aggressively before damage ensues. RA, a systemic disease, may also lead to a variety of extraarticular manifestations, including fatigue, subcutaneous nodules, lung involvement, pericarditis, peripheral neuropathy, vasculitis, and hematologic abnormalities, which must be managed accordingly.

Insights gained by a wealth of basic and clinical research over the past two decades have revolutionized the contemporary paradigms for the diagnosis and management of RA. Serum antibodies to cyclic citrullinated peptides (anti-CCPs) are routinely included with rheumatoid factor in the diagnostic evaluation of patients with suspected RA, and serve as biomarkers of prognostic significance. Advances in imaging modalities have assisted clinical decision-making by improving the detection of joint inflammation and damage. The science of RA has taken major leaps forward by illuminating new disease-related genes and environmental interactions and elucidating in more detail the molecular components and pathways of disease pathogenesis. The relative contribution of these molecular components and pathways has been further brought to light by the observed benefits of targeted biologic and small-molecule therapies. Despite this progress, incomplete understanding of the initiating events of RA and the factors perpetuating the chronic inflammatory response remains a sizable barrier to its cure and prevention.

The last two decades have witnessed a remarkable improvement in the outcomes of RA. The crippling arthritis of year's past is encountered much less frequently today. Much of this progress can be traced to the expanded therapeutic armamentarium and the adoption of early treatment intervention. The shift in treatment strategy dictates a new mind-set for primary care practitioners—namely, one that demands early referral of patients with inflammatory arthritis to a rheumatologist for prompt diagnosis and initiation of therapy. Only then will patients achieve their best outcomes.

CLINICAL FEATURES

The incidence of RA increases between 25 and 55 years of age, after which it plateaus until the age of 75 and then decreases. The presenting symptoms of RA typically result from inflammation of the joints, tendons, and bursae. Patients often complain of early morning joint stiffness lasting more than 1 h that eases with physical activity. The earliest involved joints are typically the small joints of the hands and feet. The initial pattern of joint involvement may be monoarticular, oligoarticular (≤4 joints), or polyarticular (>5 joints), usually in a symmetric distribution. Some patients with inflammatory arthritis will present with too few affected joints to be classified as having RA—so-called undifferentiated inflammatory arthritis. Those with an undifferentiated arthritis who are most likely to be diagnosed later with RA have a higher number of tender and swollen joints, test positive for serum rheumatoid factor (RF) or anti-CCP antibodies, and have higher scores for physical disability.

Once the disease process of RA is established, the wrists, metacarpophalangeal (MCP), and proximal interphalangeal (PIP) joints stand out as the most frequently involved joints **(Fig. 351-1)**. Distal interphalangeal (DIP) joint involvement may occur in RA, but it usually is a manifestation of coexistent osteoarthritis. Flexor tendon tenosynovitis is a frequent hallmark of RA and leads to decreased range of motion, reduced grip strength, and "trigger" fingers. Progressive destruction of the joints and soft tissues may lead to chronic, irreversible deformities. Ulnar deviation results from subluxation of the MCP joints, with subluxation, or partial dislocation, of the proximal phalanx to the volar side of the hand. Hyperextension of the PIP joint with flexion of the DIP joint ("swan-neck deformity"), flexion of the PIP joint with hyperextension of the DIP joint ("boutonnière deformity"), and subluxation of the first MCP joint with hyperextension of the first interphalangeal (IP) joint ("Z-line deformity") also may result from damage to the tendons, joint capsule, and other soft tissues in these small joints. Inflammation about the ulnar styloid and tenosynovitis of the extensor carpi ulnaris may cause subluxation of the distal ulna, resulting in a "piano-key movement" of the ulnar styloid. Although metatarsophalangeal (MTP) joint involvement in the feet is an early feature of disease, chronic inflammation of the ankle and midtarsal regions usually comes later and may lead to pes planovalgus ("flat feet"). Large joints, including the knees and shoulders, are often affected in established disease, although these joints may remain asymptomatic for many years after onset.

FIGURE 351-1

Metacarpophalangeal and proximal interphalangeal joint swelling in rheumatoid arthritis. *(© 2018 American College of Rheumatology. Used with permission.)*



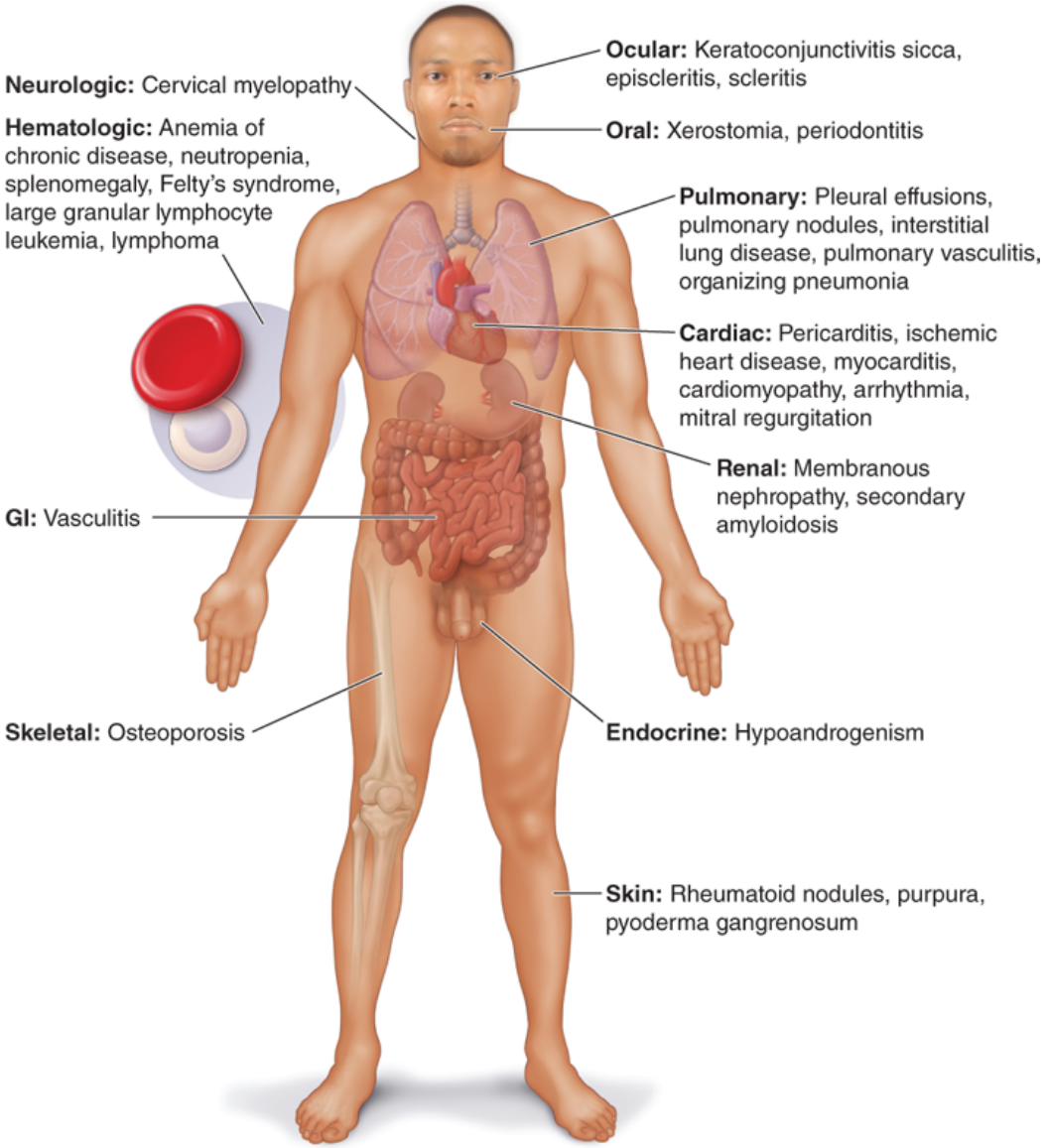
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Atlantoaxial involvement of the cervical spine is clinically noteworthy because of its potential to cause compressive myelopathy and neurologic dysfunction. Neurologic manifestations are rarely a presenting sign or symptom of atlantoaxial disease, but they may evolve over time with progressive instability of C1 on C2. The prevalence of atlantoaxial subluxation has been declining in recent years, and occurs now in <10% of patients. Unlike the spondyloarthritides (Chap. 355), RA rarely affects the thoracic and lumbar spine. Radiographic abnormalities of the temporomandibular joint occur commonly in patients with RA, but they are generally not associated with significant symptoms or functional impairment.

Extraarticular manifestations may develop during the clinical course of RA in up to 40% of patients, even prior to the onset of arthritis (Fig. 351-2). Patients most likely to develop extraarticular disease have a history of cigarette smoking, have early onset of significant physical disability, and test positive for serum RF or anti-CCP antibodies. Subcutaneous nodules, secondary Sjögren's syndrome, interstitial lung disease (ILD), pulmonary nodules, and anemia are among the most frequently observed extraarticular manifestations. Recent studies have shown a decrease in the incidence and severity of at least some extraarticular manifestations, particularly Felty's syndrome and vasculitis.

FIGURE 351-2

Extraarticular manifestations of rheumatoid arthritis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The most common systemic and extraarticular features of RA are described in more detail in the sections below.

CONSTITUTIONAL

These signs and symptoms include weight loss, fever, fatigue, malaise, depression, and in the most severe cases, cachexia; they generally reflect a high degree of inflammation and may even precede the onset of joint symptoms. In general, the presence of a fever of >38.3°C (101°F) at any time during the clinical course should raise suspicion of systemic vasculitis (see below) or infection.

NODULES

Subcutaneous nodules have been reported to occur in 30–40% of patients and more commonly in those with the highest levels of disease activity, the disease-related shared epitope (SE) (see below), a positive test for serum RF, and radiographic evidence of joint erosions. However, more recent cohort studies suggest a declining prevalence of subcutaneous nodules, perhaps, related to early and more aggressive disease-modifying therapy. When palpated, the nodules are generally firm; nontender; and adherent to periosteum, tendons, or bursae; developing in areas of the skeleton subject to repeated trauma or irritation such as the forearm, sacral prominences, and Achilles tendon. They may also occur in the lungs, pleura, pericardium, and peritoneum. Nodules are typically benign, although they can be associated with infection, ulceration, and gangrene.

SJÖGREN'S SYNDROME

Secondary Sjögren's syndrome (Chap. 354) is defined by the presence of either keratoconjunctivitis sicca (dry eyes) or xerostomia (dry mouth) in association with another connective tissue disease, such as RA. Approximately 10% of patients with RA have secondary Sjögren's syndrome.

PULMONARY

Pleuritis, the most common pulmonary manifestation of RA, may produce pleuritic chest pain and dyspnea, as well as a pleural friction rub and effusion. Pleural effusions tend to be exudative with increased numbers of monocytes and neutrophils. ILD may also occur in patients with RA and is heralded by symptoms of dry cough and progressive shortness of breath. ILD can be associated with cigarette smoking and is generally found in patients with higher disease activity, although it may be diagnosed in up to 3.5% of patients prior to the onset of joint symptoms. Recent studies have shown the overall prevalence of ILD in RA to be as high as 12%. Diagnosis is readily made by high-resolution chest computed tomography (CT) scan, which shows infiltrative opacification in the periphery of both lungs. Usual interstitial pneumonia (UIP) and non-specific interstitial pneumonia (NSIP) are the main histological and radiologic patterns of ILD. UIP causes progressive scarring of the lungs that produces on chest CT scan honeycomb changes in the periphery and lower portions of the lungs. In contrast, the most common radiographic changes in NSIP are relatively symmetric and bilateral ground glass opacities with associated fine reticulations, with volume loss and traction bronchiectasis. In both cases, pulmonary function testing shows a restrictive pattern (e.g., reduced total lung capacity) with a reduced diffusing capacity for carbon monoxide (DL_{CO}). The presence of ILD confers a poor prognosis, which if present, is associated with a 10% increase in mortality. The prognosis of ILD in RA is not quite as poor as that of idiopathic pulmonary fibrosis (e.g., usual interstitial pneumonitis). ILD secondary to RA responds more favorably than idiopathic ILD to immunosuppressive therapy **(Chap. 287)**. Pulmonary nodules are also common in patients with RA and may be solitary or multiple. Caplan's syndrome is a rare subset of pulmonary nodulosis characterized by the development of nodules and pneumoconiosis following silica exposure. Respiratory bronchiolitis and bronchiectasis are other less common pulmonary disorders associated with RA.

CARDIAC

The most frequent site of cardiac involvement in RA is the pericardium. However, clinical manifestations of pericarditis occur in <10% of patients with RA despite the fact that pericardial involvement is detectable in nearly one-half of cases by echocardiogram or autopsy studies. Cardiomyopathy, another clinically important manifestation of RA, may result from necrotizing or granulomatous myocarditis, coronary artery disease, or diastolic dysfunction. This involvement too may be subclinical and only identified by echocardiography or cardiac magnetic resonance imaging (MRI). Rarely, the heart muscle may contain rheumatoid nodules or be infiltrated with amyloid. Mitral regurgitation is the most common valvular abnormality in RA, occurring at a higher frequency than the general population.

VASCULITIS

Rheumatoid vasculitis **(Chap. 356)** typically occurs in patients with long-standing disease, a positive test for serum RF or anti-CCP antibodies, and hypocomplementemia. The overall incidence has decreased significantly in the last decade to <1% of patients. The cutaneous signs vary and include petechiae, purpura, digital infarcts, gangrene, livedo reticularis, and in severe cases large, painful lower extremity ulcerations. Vasculitic ulcers, which may be difficult to distinguish from those caused by venous insufficiency, may be treated successfully with immunosuppressive agents (requiring cytotoxic treatment in severe cases) as well as skin grafting. Sensorimotor polyneuropathies, such as mononeuritis multiplex, may occur in association with systemic rheumatoid vasculitis.

HEMATOLOGIC

A normochromic, normocytic anemia often develops in patients with RA and is the most common hematologic abnormality. The degree of anemia parallels the degree of inflammation, correlating with the levels of serum C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR). Platelet counts may also be elevated in RA as an acute-phase reactant. Immune-mediated thrombocytopenia is rare in this disease.

Felty's syndrome is defined by the clinical triad of neutropenia, splenomegaly, and nodular RA and is seen in <1% of patients, although its incidence appears to be declining in the face of more aggressive treatment of the joint disease. It typically occurs in the late stages of severe RA and is more common in whites than other racial groups. T cell large granular lymphocyte leukemia (T-LGL) may have a similar clinical presentation and often occurs in association with RA. T-LGL is characterized by a chronic, indolent clonal growth of LGL cells, leading to neutropenia and splenomegaly. As opposed to Felty's syndrome, T-LGL may develop early in the course of RA. Leukopenia apart from these disorders is uncommon and most often a side effect of drug therapy.

LYMPHOMA

Large cohort studies have shown a two- to fourfold increased risk of lymphoma in RA patients compared with the general population. The most common histopathologic type of lymphoma is a diffuse large B cell lymphoma. The risk of developing lymphoma increases if the patient has high levels of disease activity or Felty's syndrome.

ASSOCIATED CONDITIONS

In addition to extraarticular manifestations, several conditions associated with RA contribute to disease morbidity and mortality rates. They are worthy of mention because they affect chronic disease management.

Cardiovascular Disease

The most common cause of death in patients with RA is cardiovascular disease. The incidence of coronary artery disease and carotid atherosclerosis is higher in RA patients than in the general population even when controlling for traditional cardiac risk factors, such as hypertension, obesity, hypercholesterolemia, diabetes, and cigarette smoking. Furthermore, congestive heart failure (including both systolic and diastolic dysfunction) occurs at an approximately twofold higher rate in RA than in the general population. The presence of elevated serum inflammatory markers appears to confer an increased risk of cardiovascular disease in this population.

Osteoporosis

Osteoporosis is more common in patients with RA than an age- and sex-matched population, with prevalence rates of 20–30%. The inflammatory milieu of the joint probably spills over into the rest of the body and promotes generalized bone loss by activating osteoclasts. Chronic use of glucocorticoids and disability-related immobility also contributes to osteoporosis. Hip fractures are more likely to occur in patients with RA and are significant predictors of increased disability and mortality rate in this disease.

Hypoandrogenism

Men and postmenopausal women with RA have lower mean serum testosterone, luteinizing hormone (LH), and dehydroepiandrosterone (DHEA) levels than control populations. It has thus been hypothesized that hypoandrogenism may play a role in the pathogenesis of RA or arise as a consequence of the chronic inflammatory response. It is also important to realize that patients receiving chronic glucocorticoid therapy may develop hypoandrogenism owing to inhibition of LH and follicle-stimulating hormone (FSH) secretion from the pituitary gland. Because low testosterone levels may lead to osteoporosis, men with hypoandrogenism should be considered for androgen replacement therapy.

EPIDEMIOLOGY

RA affects ~0.5–1% of the adult population worldwide. There is evidence that the overall incidence of RA has been decreasing in recent decades, whereas the prevalence has remained the same because individuals with RA are living longer. The incidence and prevalence of RA varies based on geographic location, both globally and among certain ethnic groups within a country **(Fig. 351-3)**. For example, the Native American Yakima, Pima, and Chippewa tribes of North America have reported prevalence rates in some studies of nearly 7%. In contrast, many population studies from Africa and Asia show lower prevalence rates for RA in the range of 0.2–0.4%.

FIGURE 351-3

Global prevalence rates of rheumatoid arthritis (RA) with genetic associations. Listed are the major genetic alleles associated with RA. Although human leukocyte antigen (HLA)-DRB1 mutations are found globally, some alleles have been associated with RA in only certain ethnic groups.

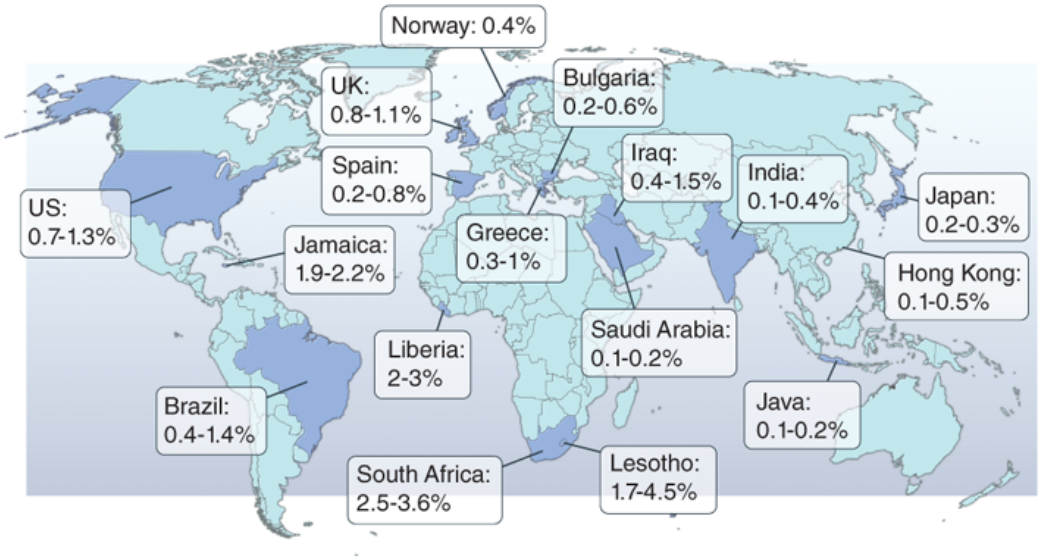
European ancestry:
HLA-DRB1:
*0401
*0404
*0301
***0101**

PTPN22: European
STAT4: North American
TNFAIP3: North American
TRAFI/CF: North American
CTLA4: European

Asian ancestry:
HLA-DRB1:
*0401 (East Asian)
*0405
*0901 (Japanese, Malaysian, Korean)

PADI4
CD244

Other:
CD40



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Like many other autoimmune diseases, RA occurs more commonly in females than in males, with a 2–3:1 ratio. Interestingly, studies of RA from some of the Latin American and African countries show an even greater predominance of disease in females compared to males, with ratios of 6–8:1. Given this preponderance of females, various theories have been proposed to explain the possible role of estrogen in disease pathogenesis. Most of the theories center on the role of estrogens in enhancing the immune response. For example, some experimental studies have shown that estrogen can stimulate production of tumor necrosis factor α (TNF-α), a major cytokine in the pathogenesis of RA.

GENETIC CONSIDERATIONS

It has been recognized for over 30 years that genetic factors contribute to the occurrence of RA as well as to its severity. The likelihood that a first-degree relative of a patient will share the diagnosis of RA is 2–10 times greater than in the general population. There remains, however, some uncertainty in the extent to which genetics plays a role in the causative mechanisms of RA. Heritability estimates range from 40 to 50% and are approximately the same for autoantibody positive and negative individuals. The estimate of genetic influence may vary across studies due to gene–environment interactions.

The alleles known to confer the greatest risk of RA are located within the major histocompatibility complex (MHC). It has been estimated that about 13% of the genetic risk for RA resides within this locus. Most, but probably not all, of this risk is associated with allelic variation in the HLA-DRB1 gene, which encodes the MHC II β-chain molecule. The disease-associated HLA-DRB1 alleles share an amino acid sequence at positions 70–74 in the third hypervariable regions of the HLA-DR β-chain, termed the *shared epitope*. Carriership of the SE alleles is associated with production of anti-CCP antibodies and worse disease outcomes. Some of these HLA-DRB1 alleles bestow a high risk of disease (*0401), whereas others confer a more moderate risk (*0101, *0404, *1001, and *0901). Additionally, there is regional variation. In Greece, for example, where RA tends to be milder than in western European countries, RA susceptibility has been associated with the *0101 SE allele. By comparison, the *0401 or *0404 alleles are found in ~50–70% of northern Europeans and are the predominant risk alleles in this group. The most common disease susceptibility SE alleles in Asians, namely the Japanese, Koreans, and Chinese, are *0405 and *0901. Lastly, disease susceptibility of Native American populations such as the Pima and Tlingit Indians, where the prevalence of RA can be as high as 7%, is associated with the SE allele *1042. The risk of RA conferred by these SE alleles is less in African and Hispanic Americans than in individuals of European ancestry.

Genome-wide association studies (GWAS) have made possible the identification of several non-MHC-related genes that contribute to RA susceptibility. GWAS are based on the detection of single-nucleotide polymorphisms (SNPs), which allow for examination of the genetic architecture of complex diseases such as RA. There are ~10 million common SNPs within a human genome consisting of 3 billion base pairs. As a rule, GWAS identify only common variants, namely, those with a frequency of >5% in the general population.

Overall, several themes have emerged from GWAS in RA. First, among the more than 100 non-MHC loci identified as risk alleles for RA, they individually have only a modest effect on risk; they also contribute to the risk for developing other autoimmune diseases, such as type 1 diabetes mellitus, systemic lupus erythematosus, and multiple sclerosis. Second, although most of the non-HLA associations are described in patients with anti-CCP antibody-positive disease, there are several risk loci that are unique to anti-CCP antibody-negative disease. Third, risk alleles vary among ethnic groups. And fourth, the risk loci mostly reside in genes encoding proteins involved in the regulation of the immune response. However, the risk alleles identified by GWAS only account at present for ~5% of the genetic risk, suggesting that rare variants or other classes of DNA variants, such as variants in copy number, may be yet found that significantly contribute to the overall risk model.

Recently, imputation of SNP data from a GWAS meta-analysis shows amino acid substitutions in the MHC locus independently associated with the risk for RA are at positions 11, 71, and 74 in HLA-DRβ1, position 9 of HLA-B, and position 9 of HLA-DPβ1. The amino acids at positions 11, 71, and 74 are located in the antigen-binding grove of the HLA-DRβ1 molecule, highlighting positions 71 and 74 that form part of the original SE.

Among the best examples of the non-MHC genes contributing to the risk of RA is the gene encoding protein tyrosine phosphatase non-receptor 22 (*PTPN22*). This gene varies in frequency among patients from different parts of Europe (e.g., 3–10%), but is absent in patients of East Asian ancestry. *PTPN22* encodes lymphoid tyrosine phosphatase, a protein that regulates T and B cell function. Inheritance of the risk allele for *PTPN22* produces a gain-of-function in the protein that is hypothesized to result in the abnormal thymic selection of autoreactive T and B cells and appears to be associated exclusively with anti-CCP-positive disease. The peptidyl arginine deiminase type IV (*PADI4*) gene is another risk allele that encodes an enzyme involved in the conversion of arginine to citrulline and is postulated to play a role in the development of antibodies to citrullinated antigens. A polymorphism in *PADI4* has been associated with RA only in Asian populations. Recently, polymorphisms in apolipoprotein M (APOM) have been demonstrated in an East Asian population to confer an increased risk for RA as well as risk for dyslipidemia, independent of RA disease activity.

Epigenetics is the study of heritable traits that affect gene expression but do not modify DNA sequence. It may provide a link between environmental exposure and predisposition to disease. The best-studied mechanisms include posttranslational histone modifications and DNA methylation. Although studies of epigenetic phenomena are limited, DNA methylation patterns have been shown to differ between RA patients and healthy controls, as well as patients with osteoarthritis. MicroRNAs, which are non-coding RNAs that function as post-transcriptional regulators of gene expression, represent an additional epigenetic mechanism that may potentially influence cellular responses. Many microRNAs have been identified as contributing to the activated phenotype of synovial fibroblasts such as miR146a or miR155.

ENVIRONMENTAL FACTORS

In addition to genetic predisposition, a host of environmental factors have been implicated in the pathogenesis of RA. The most reproducible of these environmental links is cigarette smoking. Numerous cohort and case control studies have demonstrated that smoking confers a relative risk for developing RA of 1.5–3.5. In particular, women who smoke cigarettes have a nearly 2.5 times greater risk of RA, a risk that persists even 15 years after smoking cessation. A twin who smokes will have a significantly higher risk for RA than his or her monozygotic co-twin, theoretically with the same genetic risk, who does not smoke. Interestingly, the risk from smoking is almost exclusively related to RF and anti-CCP antibody-positive disease. However, it has not been shown that smoking cessation, while having many health benefits, improves disease activity.

Researchers began to aggressively seek an infectious etiology for RA after the discovery in 1931 that sera from patients with this disease could agglutinate strains of streptococci. Certain viruses such as Epstein-Barr virus (EBV) have garnered the most interest over the past 30 years given their ubiquity, ability to persist for many years in the host, and frequent association with arthritic complaints. For example, titers of IgG antibodies against EBV antigens in the peripheral blood and saliva are significantly higher in patients with RA than the general population. EBV DNA has also been found in synovial fluid and synovial cells of RA patients. Because the evidence for these links is largely circumstantial, it has not been possible to directly implicate infection as a causative factor in RA.

Recent studies suggest that periodontitis may play a role in the pathogenesis of RA. Multiple studies provide evidence for a link between anti-CCP positive RA and cigarette smoking, periodontal disease, and the oral microbiome, specifically *Porphyromonas gingivalis*. It has been hypothesized that the immune response to *P. gingivalis* may trigger the development of RA and that induction of anti-CCP antibodies results from citrullination of arginine residues in human tissues by the enzyme peptidyl arginine deiminase (PAD). Interestingly, *P. gingivalis* is the only oral bacterial species known to harbor this enzyme. Some studies have shown a relationship between circulating antibodies to *P. gingivalis* and RA, as well as these antibodies and first-degree relatives at risk for this disease.

PATHOLOGY

RA affects the synovial tissue and underlying cartilage and bone. The synovial membrane, which covers most articular surfaces, tendon sheaths, and bursae, normally is a thin layer of connective tissue. In joints, it faces the bone and cartilage, bridging the opposing bony surfaces and inserting at periosteal regions close to the articular cartilage. It consists primarily of two cell types—type A synoviocytes (macrophage-derived) and type B synoviocytes (fibroblast-derived). The synovial fibroblasts are the most abundant and produce the structural components of joints, including

collagen, fibronectin, and laminin, as well as other extracellular constituents of the synovial matrix. The sublining layer consists of blood vessels and a sparse population of mononuclear cells within a loose network of connective tissue. Synovial fluid, an ultrafiltrate of blood, diffuses through the subsynovial lining tissue across the synovial membrane and into the joint cavity. Its main constituents are hyaluronan and lubricin. Hyaluronan is a glycosaminoglycan that contributes to the viscous nature of synovial fluid, which along with lubricin, lubricates the surface of the articular cartilage.

The pathologic hallmarks of RA are synovial inflammation and proliferation, focal bone erosions, and thinning of articular cartilage. Chronic inflammation leads to synovial lining hyperplasia and the formation of pannus, a thickened cellular membrane containing fibroblast-like synoviocytes and granulation-reactive fibrovascular tissue that invades the underlying cartilage and bone. The inflammatory infiltrate is made up of no less than six cell types: T cells, B cells, plasma cells, dendritic cells, mast cells, and, to a lesser extent, granulocytes. The T cells comprise 30–50% of the infiltrate, with the other cells accounting for the remainder. The topographical organization of these cells is complex and may vary among individuals with RA. Most often, the lymphocytes are diffusely organized among the tissue resident cells; however, in some cases, the B cells, T cells, and dendritic cells may form higher levels of organization, such as lymphoid follicles and germinal center–like structures. Growth factors secreted by synovial fibroblasts and macrophages promote the formation of new blood vessels in the synovial sublining that supply the increasing demands for oxygenation and nutrition required by the infiltrating leukocytes and expanding synovial tissue.

The structural damage to the mineralized cartilage and subchondral bone is mediated by the osteoclast. Osteoclasts are multinucleated giant cells that can be identified by their expression of CD68, tartrate-resistant acid phosphatase, cathepsin K, and the calcitonin receptor. They appear at the pannus-bone interface where they eventually form resorption lacunae. These lesions typically localize where the synovial membrane inserts into the periosteal surface at the edges of bones close to the rim of articular cartilage and at the attachment sites of ligaments and tendon sheaths. This process most likely explains why bone erosions usually develop at the radial sites of the MCP joints juxtaposed to the insertion sites of the tendons, collateral ligaments, and synovial membrane. Another form of bone loss is periarticular osteopenia that occurs in joints with active inflammation. It is associated with substantial thinning of the bony trabeculae along the metaphyses of bones, and likely results from inflammation of the bone marrow cavity. These lesions can be visualized on MRI scans, where they appear as signal alterations in the bone marrow adjacent to inflamed joints. Their signal characteristics show they are water-rich with a low fat content and are consistent with highly vascularized inflammatory tissue. These bone marrow lesions are often the forerunner of bone erosions.

The cortical bone layer that separates the bone marrow from the invading pannus is relatively thin and susceptible to penetration by the inflamed synovium. The bone marrow lesions seen on MRI scans are associated with an endosteal bone response characterized by the accumulation of osteoblasts and deposition of osteoid. Thus, in recent years, the concept of joint pathology in RA has been extended to include the bone marrow cavity. Finally, generalized osteoporosis, which results in the thinning of trabecular bone throughout the body, is a third form of bone loss found in patients with RA.

Articular cartilage is an avascular tissue comprised of a specialized matrix of collagens, proteoglycans, and other proteins. It is organized in four distinct regions (superficial, middle, deep, and calcified cartilage zones)—chondrocytes constitute the unique cellular component in these layers. Originally, cartilage was considered to be an inert tissue, but it is now known to be a highly responsive tissue that reacts to inflammatory mediators and mechanical factors, which in turn, alter the balance between cartilage anabolism and catabolism. In RA, the initial areas of cartilage degradation are juxtaposed to the synovial pannus. The cartilage matrix is characterized by a generalized loss of proteoglycan, most evident in the superficial zones adjacent to the synovial fluid. Degradation of cartilage may also take place in the perichondrocytic zone and in regions adjacent to the subchondral bone.

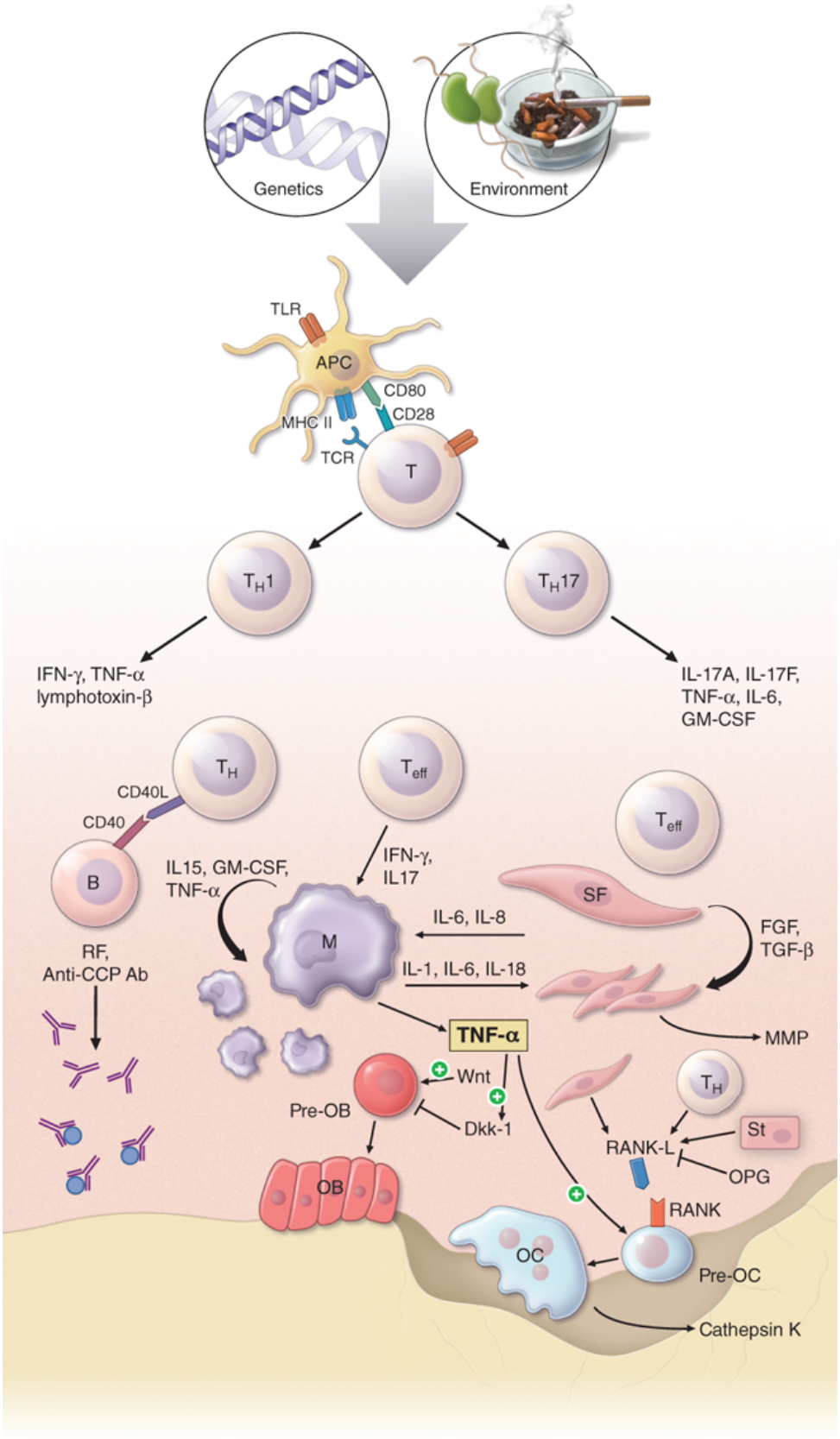
PATHOGENESIS

The pathogenic mechanisms of synovial inflammation are likely to result from a complex interplay of genetic, environmental, and immunologic factors that produces dysregulation of the immune system and a breakdown in self-tolerance **(Fig. 351-4)**. Precisely what triggers these initiating events and what genetic and environmental factors disrupt the immune system remains a mystery. However, a detailed molecular picture is emerging of the mechanisms underlying the chronic inflammatory response and the destruction of the articular cartilage and bone.

FIGURE 351-4

Pathophysiologic mechanisms of inflammation and joint destruction. Genetic predisposition along with environmental factors may trigger the development of rheumatoid arthritis (RA), with subsequent synovial T cell activation. CD4+ T cells become activated by antigen-presenting cells (APCs) through interactions between the T cell receptor and class II MHC-peptide antigen (signal 1) with co-stimulation through the CD28-CD80/86 pathway, as well as other pathways (signal 2). In theory, ligands binding Toll-like receptors (TLRs) may further stimulate activation of APCs inside the joint. Synovial CD4+ T cells differentiate into T_H1 and T_H17 cells, each with their distinctive cytokine profile. CD4+ T_H cells in turn activate B cells, some of which are destined to differentiate into autoantibody-producing plasma cells. Immune complexes, possibly comprised of rheumatoid factors (RFs) and anti–cyclic citrullinated peptides (CCP) antibodies, may form inside the joint, activating the complement pathway and amplifying inflammation. T effector cells stimulate synovial macrophages (M) and fibroblasts (SF) to secrete proinflammatory mediators, among which is tumor necrosis factor α (TNF-α). TNF-α upregulates adhesion molecules on endothelial cells, promoting leukocyte influx into the joint. It also stimulates the

production of other inflammatory mediators, such as interleukin 1 (IL-1), IL-6, and granulocyte-macrophage colony-stimulating factor (GM-CSF). TNF-α has a critically important function in regulating the balance between bone destruction and formation. It upregulates the expression of dickkopf-1 (DKK-1), which can then internalize Wnt receptors on osteoblast precursors. Wnt is a soluble mediator that promotes osteoblastogenesis and bone formation. In RA, bone formation is inhibited through the Wnt pathway, presumably due to the action of elevated levels of DKK-1. In addition to inhibiting bone formation, TNF-α stimulates osteoclastogenesis. However, it is not sufficient by itself to induce the differentiation of osteoclast precursors (Pre-OC) into activated osteoclasts capable of eroding bone. Osteoclast differentiation requires the presence of macrophage colony-stimulating factor (M-CSF) and receptor activator of nuclear factor-κB (RANK) ligand (RANKL), which binds to RANK on the surface of Pre-OC. Inside the joint, RANKL is mainly derived from stromal cells, synovial fibroblasts, and T cells. Osteoprotegerin (OPG) acts as a decoy receptor for RANKL, thereby inhibiting osteoclastogenesis and bone loss. FGF, fibroblast growth factor; IFN, interferon; TGF, transforming growth factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In RA, the preclinical stage appears to be characterized by a breakdown in self-tolerance. This idea is supported by the finding that autoantibodies, such as RF and anti-CCP antibodies, may be found in sera from patients many years before onset of clinical disease. However, the antigenic targets of anti-CCP antibodies and RF are not restricted to the joint, and their role in disease pathogenesis remains speculative. Anti-CCP antibodies are directed against deaminated peptides, which result from posttranslational modification by the enzyme PADI4. They recognize citrulline-containing regions of several different matrix proteins, including filaggrin, keratin, fibrinogen, and vimentin, and are present at higher levels in the joint fluid

compared to the serum. Other autoantibodies have been found in a minority of patients with RA, but they also occur in the setting of other types of arthritis. They bind to a diverse array of autoantigens, including type II collagen, human cartilage gp-39, aggrecan, calpastatin, immunoglobulin binding protein (BiP), and glucose-6-phosphate isomerase.

In theory, environmental stimulants may synergize with other factors to bring about inflammation in RA. People who smoke display higher citrullination of proteins in bronchoalveolar fluid than those who do not smoke. Thus, it has been speculated that long-term exposure to tobacco smoke might induce citrullination of cellular proteins in the lung and stimulate the expression of a neoepitope capable of inducing self-reactivity, which in turns, leads to formation of immune complexes and joint inflammation. Exposure to silicone dust and mineral oil, which has adjuvant effects, has also been linked to an increased risk for anti-CCP antibody-positive RA. Similarly, periodontal pathogens, such as *P. gingivalis*, may play a pathogenic role and contribute to the citrullination of cellular proteins in the oral cavity. In addition to the possible link between the oral microbiome and RA, investigators are turning their attention toward the intestinal microbiota and whether its altered composition may predispose to disease.

How might microbes or their products be involved in the initiating events of RA? The immune system is alerted to the presence of microbial infections through Toll-like receptors (TLRs). There are 10 TLRs in humans that recognize a variety of microbial products, including bacterial cell-surface lipopolysaccharides and heat-shock proteins (TLR4), lipoproteins (TLR2), double-strand RNA viruses (TLR3), and unmethylated CpG DNA from bacteria (TLR9). TLR2, 3, and 4 are abundantly expressed by synovial fibroblasts in early RA and, when bound by their ligands, upregulate production of proinflammatory cytokines. Although TLR ligands may theoretically amplify inflammatory pathways in RA, their specific role in disease pathogenesis remains uncertain.

The pathogenesis of RA is built upon the concept that self-reactive T cells drive the chronic inflammatory response. In theory, self-reactive T cells might arise in RA from abnormal central (thymic) selection or intrinsic defects lowering the threshold in the periphery for T cell activation. Either mechanism might result in abnormal expansion of the self-reactive T cell repertoire and a breakdown in T cell tolerance. The support for these theories comes mainly from studies of arthritis in mouse models. It has not been shown that patients with RA have abnormal thymic selection of T cells or defective apoptotic pathways regulating cell death. At least some antigen stimulation inside the joint seems likely, owing to the fact that T cells in the synovium express a cell-surface phenotype indicating prior antigen exposure and show evidence of clonal expansion. Of interest, peripheral blood T cells from patients with RA have been shown to display a fingerprint of premature aging that mostly affects inexperienced naïve T cells. In these studies, the most glaring findings have been the loss of telomeric sequences and a decrease in the thymic output of new T cells. Although intriguing, it is not clear how generalized T cell abnormalities might provoke a systemic disease with a predominance of synovitis.

There is substantial evidence of a role for CD4+ T cells in the pathogenesis of RA. First, the co-receptor CD4 on the surface of T cells binds to invariant sites on MHC class II molecules, stabilizing the MHC-peptide–T cell receptor complex during T cell activation. Because the SE on MHC class II molecules is a risk factor for RA, it follows that CD4+ T cell activation may play a role in the pathogenesis of this disease. Second, CD4+ memory T cells are enriched in the synovial tissue from patients with RA and can be implicated through “guilt by association.” Third, CD4+ T cells have been shown to be important in the initiation of arthritis in animal models. Fourth, some, but not all, T cell–directed therapies have shown clinical efficacy in this disease. Taken together, these lines of evidence suggest that CD4+ T cells play an important role in orchestrating the chronic inflammatory response in RA. However, other cell types, such as CD8+ T cells, natural killer (NK) cells, and B cells are present in synovial tissue and may also influence pathogenic responses.

In the rheumatoid joint, by mechanisms of cell-cell contact and release of soluble mediators, activated T cells stimulate macrophages and fibroblast-like synoviocytes to generate proinflammatory mediators and proteases that drive the synovial inflammatory response and destroy the cartilage and bone. CD4+ T cell activation is dependent on two signals: (1) T cell receptor binding to peptide-MHC on antigen-presenting cells; and (2) CD28 binding to CD80/86 on antigen-presenting cells. CD4+ T cells also provide help to B cells, which in turn, produce antibodies that may promote further inflammation in the joint. The previous T cell–centric model for the pathogenesis of RA was based on a T_H1-driven paradigm, which came from studies indicating that CD4+ T helper (T_H) cells differentiated into T_H1 and T_H2 subsets, each with their distinctive cytokine profiles. T_H1 cells were found to mainly produce interferon γ (IFN-γ), lymphotoxin β, and TNF-α, whereas T_H2 cells predominately secreted interleukin (IL)-4, IL-5, IL-6, IL-10, and IL-13. The recent discovery of another subset of T_H cells, namely the T_H17 lineage, has revolutionized our concepts concerning the pathogenesis of RA. In humans, naïve T cells are induced to differentiate into T_H17 cells by exposure to transforming growth factor β (TGF-β), IL-1, IL-6, and IL-23. Upon activation, T_H17 cells secrete a variety of proinflammatory mediators such as IL-17, IL-21, IL-22, TNF-α, IL-26, IL-6, and granulocyte-macrophage colony-stimulating factor (GM-CSF). Substantial evidence now exists from studies in both animal models and humans that IL-17 plays an important role not only in promoting joint inflammation, but also in destroying cartilage and subchondral bone. Nevertheless, secukinumab, an anti-IL17 receptor antibody, failed to show significant clinical benefit in a phase II trial involving patients with RA, raising new questions about the importance of IL-17 in perpetuating joint inflammation in this disease.

The immune system has evolved mechanisms to counterbalance the potential harmful immune-mediated inflammatory responses provoked by infectious agents and other triggers. Among these negative regulators are regulatory T (T_{reg}) cells, which are produced in the thymus and induced in

the periphery to suppress immune-mediated inflammation. They are characterized by the surface expression of CD25 and the expression of the transcription factor forkhead box P3 (FOXP3) and the absence of CD127, the IL-7 receptor. T_{regs} orchestrate dominant tolerance through contact with other immune cells and secretion of inhibitory cytokines, such as TGF-β, IL-10, and IL-35. They are heterogeneous and capable of suppressing distinct classes (T_H1, T_H2, T_H17) of the immune response. In RA, the data that T_{reg} numbers are deficient compared to normal healthy controls are contradictory and inconclusive. Although some experimental evidence suggests that T_{reg} suppressive activity is lost due to dysfunctional expression of cytotoxic T lymphocyte antigen 4 (CTLA-4), the nature of T_{reg} defects in RA and their role in disease mechanisms remains unclear.

Cytokines, chemokines, antibodies, and endogenous danger signals bind to receptors on the surface of immune cells and stimulate a cascade of intracellular signaling events that can amplify the inflammatory response. Signaling molecules and their binding partners in these pathways are the target of small-molecule drugs designed to interfere with signal transduction and in turn, block these reinforcing inflammatory loops. Examples of signaling molecules in these critical inflammatory pathways include Janus kinase (JAK)/signal transducers and activators of transcription (STAT), spleen tyrosine kinase (Syk), mitogen-activated protein kinases (MAPKs), and nuclear factor-κB (NF-κB). These pathways exhibit significant cross-talk and are found in many cell types. Some signal transducers, such as JAK3, are primarily expressed in hematopoietic cells and play an important role in the inflammatory response in RA.

Activated B cells are also important players in the chronic inflammatory response. B cells give rise to plasma cells, which in turn, produce antibodies, including RF and anti-CCP antibodies. RFs may form large immune complexes inside the joint that contribute to the pathogenic process by fixing complement and promoting the release of proinflammatory cytokines and chemokines. In mouse models of arthritis, RF-containing immune complexes and anti-CCP-containing immune complexes synergize with other mechanisms to exacerbate the synovial inflammatory response.

RA is often considered to be a macrophage-driven disease because this cell type is the predominant source of proinflammatory cytokines inside the joint. Key proinflammatory cytokines released by synovial macrophages include TNF-α, IL-1, IL-6, IL-12, IL-15, IL-18, and IL-23. Synovial fibroblasts, the other major cell type in this microenvironment, produce the cytokines IL-1 and IL-6 as well as TNF-α. TNF-α is a pivotal cytokine in the pathobiology of synovial inflammation. It upregulates adhesion molecules on endothelial cells, promoting the influx of leukocytes into the synovial microenvironment; activates synovial fibroblasts; stimulates angiogenesis; promotes pain receptor sensitizing pathways; and drives osteoclastogenesis. Fibroblasts secrete matrix metalloproteinases (MMPs) as well as other proteases that are chiefly responsible for the breakdown of articular cartilage.

Osteoclast activation at the site of the pannus is closely tied to the presence of focal bone erosion. Receptor activator of nuclear factor-κB ligand (RANKL) is expressed by stromal cells, synovial fibroblasts, and T cells. Upon binding to its receptor RANK on osteoclast progenitors, RANKL stimulates osteoclast differentiation and bone resorption. RANKL activity is regulated by osteoprotegerin (OPG), a decoy receptor of RANKL that blocks osteoclast formation. Monocytic cells in the synovium serve as the precursors of osteoclasts and, when exposed to macrophage colony-stimulating factor (M-CSF) and RANKL, fuse to form polykaryons termed *preosteoclasts*. These precursor cells undergo further differentiation into osteoclasts with the characteristic ruffled membrane. Cytokines such as TNF-α, IL-1, IL-6, and IL-17 increase the expression of RANKL in the joint and thus promote osteoclastogenesis. Osteoclasts also secrete cathepsin K, a cysteine protease that degrades the bone matrix by cleaving collagen. Stimulation of osteoclasts also contributes to generalized bone loss and osteoporosis.

Increased bone loss is only part of the story in RA, as decreased bone formation plays a crucial role in bone remodeling at sites of inflammation. Recent evidence shows that inflammation suppresses bone formation. The proinflammatory cytokine TNF-α plays a key role in actively suppressing bone formation by enhancing the expression of dickkopf-1 (DKK-1). DKK-1 is an important inhibitor of the Wnt pathway, which acts to promote osteoblast differentiation and bone formation. The Wnt system is a family of soluble glycoproteins that binds to cell-surface receptors known as frizzled (fz) and low-density lipoprotein (LDL) receptor–related proteins (LRPs) and promotes cell growth. In animal models, increased levels of DKK-1 are associated with decreased bone formation, whereas inhibition of DKK-1 protects against structural damage in the joint. Wnt proteins also induce the formation of OPG and thereby shut down bone resorption, emphasizing their key role in tightly regulating the balance between bone resorption and formation.

DIAGNOSIS

The clinical diagnosis of RA is largely based on signs and symptoms of a chronic inflammatory arthritis, with laboratory and radiographic results providing important corroborating information. In 2010, a collaborative effort between the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR) revised the 1987 ACR classification criteria for RA in an effort to improve early diagnosis with the goal of identifying patients who would benefit from early introduction of disease-modifying therapy **(Table 351-1)**. Application of the newly revised criteria yields a score of 0–10, with a score of ≥6 fulfilling the requirements for definite RA. The new classification criteria differ in several ways from the older criteria set. The new criteria include as an item a positive test for serum anti-CCP antibodies (also termed ACPA, anti-citrullinated peptide antibodies), which carries greater specificity for the diagnosis of RA than a positive test for RF. The newer classification criteria also do not take into account whether the patient has rheumatoid nodules or radiographic joint damage because these findings occur rarely in early RA. It is important to

emphasize that the new 2010 ACR-EULAR criteria are “classification criteria” as opposed to “diagnostic criteria” and serve to distinguish patients at the onset of disease who have a high likelihood of evolution to chronic disease with persistent synovitis and joint damage. The presence of radiographic joint erosions or subcutaneous nodules may inform the diagnosis in the later stages of the disease.

TABLE 351-1

Classification Criteria for Rheumatoid Arthritis

		Score
Joint involvement	1 large joint (shoulder, elbow, hip, knee, ankle) 2–10 large joints 1–3 small joints (MCP, PIP, thumb IP, MTP, wrists) 4–10 small joints >10 joints (at least 1 small joint)	0 1 2 3 5
Serology	Negative RF and negative ACPA Low-positive RF or low-positive anti-CCP antibodies (≤3 times ULN) High-positive RF or high-positive anti-CCP antibodies (>3 times ULN)	0 2 3
Acute-phase reactants	Normal CRP and normal ESR Abnormal CRP or abnormal ESR	0 1
Duration of symptoms	<6 weeks ≥6 weeks	0 1

Note: These criteria are aimed at classification of newly presenting patients who have at least one joint with definite clinical synovitis that is not better explained by another disease. A score of ≥6 fulfills requirements for definite RA.

Abbreviations: ACPA, anti-citrullinated peptide antibodies; CCP, cyclic citrullinated peptides; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; IP, interphalangeal joint; MCP, metacarpophalangeal joint; MTP, metatarsophalangeal joint; PIP, proximal interphalangeal joint; RF, rheumatoid factor; ULN, upper limit of normal.

Source: D Aletaha et al: Arthritis Rheum 62:2569, 2010.

LABORATORY FEATURES

Patients with systemic inflammatory diseases such as RA will often present with elevated nonspecific inflammatory markers such as an ESR or CRP. Detection of serum RF and anti-CCP antibodies is important in differentiating RA from other polyarticular diseases, although RF lacks diagnostic specificity and may be found in association with other chronic inflammatory diseases in which arthritis figures in the clinical manifestations.

IgM, IgG, and IgA isotypes of RF occur in sera from patients with RA, although the IgM isotype is the one most frequently measured by commercial laboratories. Serum IgM RF has been found in 75–80% of patients with RA; therefore, a negative result does not exclude the presence of this disease. It is also found in other connective tissue diseases, such as primary Sjögren’s syndrome, systemic lupus erythematosus, and type II mixed essential cryoglobulinemia, as well as chronic infections such as subacute bacterial endocarditis and hepatitis B and C. Serum RF may also be detected in 1–5% of the healthy population.

The presence of serum anti-CCP antibodies has about the same sensitivity as serum RF for the diagnosis of RA. However, its diagnostic specificity approaches 95%, so a positive test for anti-CCP antibodies in the setting of an early inflammatory arthritis is useful for distinguishing RA from other forms of arthritis. There is some incremental value in testing for the presence of both RF and anti-CCP, as some patients with RA are positive for RF but negative for anti-CCP and vice versa. The presence of RF or anti-CCP antibodies also has prognostic significance, with anti-CCP antibodies showing the most value for predicting worse outcomes.

SYNOVIAL FLUID ANALYSIS

Typically, the cellular composition of synovial fluid from patients with RA reflects an acute inflammatory state. Synovial fluid white blood cell (WBC) counts can vary widely, but generally range between 5000 and 50,000 WBC/μL compared to <2000 WBC/μL for a noninflammatory condition such as

osteoarthritis. In contrast to the synovial tissue, the overwhelming cell type in the synovial fluid is the neutrophil. Clinically, the analysis of synovial fluid is most useful for confirming an inflammatory arthritis (as opposed to osteoarthritis), while at the same time excluding infection or a crystal-induced arthritis such as gout or pseudogout **(Chap. 365)**.

JOINT IMAGING

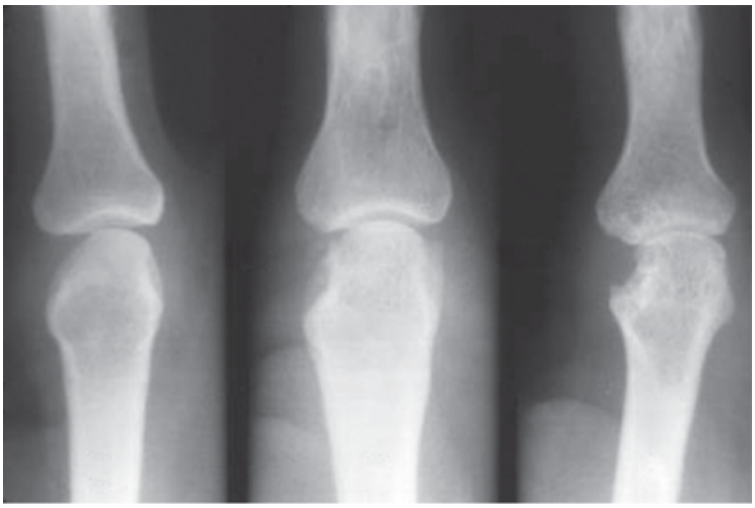
Joint imaging is a valuable tool not only for diagnosing RA, but also for tracking progression of any joint damage. Plain x-ray is the most common imaging modality, but it is limited to visualization of the bony structures and inferences about the state of the articular cartilage based on the amount of joint space narrowing. MRI and ultrasound techniques offer the added value of detecting changes in the soft tissues such as synovitis, tenosynovitis, and effusions, as well as providing greater sensitivity for identifying bony abnormalities. Plain radiographs are usually relied upon in clinical practice for the purpose of diagnosis and monitoring of affected joints. However, in selected cases, MRI and ultrasound can provide additional diagnostic information that may guide clinical decision making. Musculoskeletal ultrasound with power Doppler is increasingly used in rheumatology clinical practice for detecting synovitis and bone erosion.

Plain Radiography

Classically in RA, the initial radiographic finding is periarticular osteopenia. Practically speaking, however, this finding is difficult to appreciate on plain films and, in particular, on the newer digitalized x-rays. Other findings on plain radiographs include soft tissue swelling, symmetric joint space loss, and subchondral erosions, most frequently in the wrists and hands (MCPs and PIPs) and the feet (MTPs). In the feet, the lateral aspect of the fifth MTP is often targeted first, but other MTP joints may be involved at the same time. X-ray imaging of advanced RA may reveal signs of severe destruction, including joint subluxation and collapse **(Fig. 351-5)**.

FIGURE 351-5

X-ray demonstrating progression of erosions on the proximal interphalangeal joint. *(© 2018 American College of Rheumatology. Used with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MRI

MRI offers the greatest sensitivity for detecting synovitis and joint effusions, as well as early bone and bone marrow changes. These soft tissue abnormalities often occur before osseous changes are noted on x-ray. Presence of bone marrow edema has been recognized to be an early sign of inflammatory joint disease and can predict the subsequent development of erosions on plain radiographs as well as MRI scans. Cost and availability of MRI are the main factors limiting its routine clinical use.

Ultrasound

Ultrasound, including power color Doppler, has the ability to detect more erosions than plain radiography, especially in easily accessible joints. It can also reliably detect synovitis, including increased joint vascularity indicative of inflammation. The usefulness of ultrasound is dependent on the experience of the sonographer; however, it does offer the advantages of portability, lack of radiation, and low expense relative to MRI, factors that make it attractive as a clinical tool.

CLINICAL COURSE

The natural history of RA is complex and affected by a number of factors including age of onset, gender, genotype, phenotype (i.e., extraarticular manifestations or variants of RA), and comorbid conditions, which make for a truly heterogeneous disease. There is no simple way to predict the clinical course. It is important to realize that as many as 10% of patients with inflammatory arthritis fulfilling ACR classification criteria for RA will undergo a spontaneous remission within 6 months (particularly seronegative patients). However, the vast majority of patients will exhibit a pattern of persistent and progressive disease activity that waxes and wanes in intensity over time. A minority of patients will show intermittent and recurrent explosive attacks of inflammatory arthritis interspersed with periods of disease quiescence. Finally, an aggressive form of RA may occur in an unfortunate few with inexorable progression of severe erosive joint disease, although this highly destructive course is less common in the modern treatment era.

Disability, as measured by the Health Assessment Questionnaire (HAQ), shows gradual worsening of disability over time in the face of poorly controlled disease activity and disease progression. Disability may result from both a disease activity–related component that is potentially reversible with therapy and a joint damage–related component owing to the cumulative and largely irreversible effects of soft tissue, cartilage and bone breakdown. Early in the course of disease, the extent of joint inflammation is the primary determinant of disability, while in the later stages of disease, the amount of joint damage is the dominant contributing factor. Previous studies have shown that more than one-half of patients with RA are unable to work 10 years after the onset of their disease; however, increased employability and less work absenteeism has been reported recently with the use of newer therapies and earlier treatment intervention.

The overall mortality rate in RA is two times greater than the general population, with ischemic heart disease being the most common cause of death followed by infection. Median life expectancy is shortened by an average of 7 years for men and 3 years for women compared to control populations. Patients at higher risk for shortened survival are those with systemic extraarticular involvement, low functional capacity, low socioeconomic status, low education, and chronic prednisone use.

TREATMENT

TREATMENT

Rheumatoid Arthritis

The amount of clinical disease activity in patients with RA reflects the overall burden of inflammation and is the variable most influencing treatment decisions. Joint inflammation is the main driver of joint damage and is the most important cause of functional disability in the early stages of disease. Several composite indices have been developed to assess clinical disease activity. The ACR 20, 50, and 70 improvement criteria (which corresponds to a 20, 50, and 70% improvement, respectively, in joint counts, physician/patient assessment of disease severity, pain scale, serum levels of acute-phase reactants [ESR or CRP], and a functional assessment of disability using a self-administered patient questionnaire) are a composite index with a dichotomous response variable. The ACR improvement criteria are commonly used in clinical trials as an endpoint for comparing the proportion of responders between treatment groups. In contrast, the Disease Activity Score (DAS), Simplified Disease Activity Index (SDAI), the Clinical Disease Activity Index (CDAI), and the Routine Assessment of Patient Index Data 3 (RAPID3) are continuous measures of disease activity. These scales are increasingly used in clinical practice for tracking disease status and, in particular, for documenting treatment response.

Several developments during the past two decades have changed the therapeutic landscape in RA. They include (1) the emergence of methotrexate as the disease-modifying antirheumatic drug (DMARD) of first choice for the treatment of early RA; (2) the development of novel highly efficacious biologicals that can be used alone or in combination with methotrexate; and (3) the proven superiority of combination DMARD regimens over methotrexate alone. The medications used for the treatment of RA may be divided into broad categories: nonsteroidal anti-inflammatory drugs (NSAIDs); glucocorticoids, such as prednisone and methylprednisolone; conventional DMARDs; and biologic DMARDs **(Table 351-2)**. Although disease for some patients with RA is managed adequately with a single DMARD, such as methotrexate, it demands in most cases the use of a combination DMARD regimen that may vary in its components over the treatment course depending on fluctuations in disease activity and emergence of drug-related toxicities and comorbidities.

NSAIDs

NSAIDs were formerly viewed as the core of RA therapy, but they are now considered to be adjunctive agents for management of symptoms uncontrolled by other measures. NSAIDs exhibit both analgesic and anti-inflammatory properties. The anti-inflammatory effects of NSAIDs derive from their ability to nonselectively inhibit cyclooxygenase (COX)-1 and COX-2. Although the results of clinical trials suggest that NSAIDs are roughly equivalent in their efficacy, experience suggests that some individuals may preferentially respond to a particular NSAID. Chronic use should be minimized due to the possibility of side effects, including gastritis and peptic ulcer disease as well as impairment of renal function.

GLUCOCORTICOIDS

Glucocorticoids may serve in several ways to control disease activity in RA. First, they may be administered in low to moderate doses to achieve rapid disease control before the onset of fully effective DMARD therapy, which often takes several weeks or even months. Second, a 1- to 2-week burst of glucocorticoids may be prescribed for the management of acute disease flares, with dose and duration guided by the severity of the exacerbation. Chronic administration of low doses (5–10 mg/d) of prednisone (or its equivalent) may also be warranted to control disease activity in patients with an inadequate response to DMARD therapy. Low-dose prednisone therapy has been shown in prospective studies to retard radiographic progression of joint disease; however, the benefits of this approach must be carefully weighed against the risks. Best practices minimize chronic use of low-dose prednisone therapy owing to the risk of osteoporosis and other long-term complications; however, the use of chronic prednisone therapy is unavoidable in some cases. High-dose glucocorticoids may be necessary for treatment of severe extraarticular manifestations of RA, such as ILD. Finally, if a patient exhibits one or a few actively inflamed joints, the clinician may consider intraarticular injection of an intermediate-acting glucocorticoid such as triamcinolone acetonide. This approach may allow for rapid control of inflammation in a limited number of affected joints. Caution must be exercised to appropriately exclude joint infection, as it often mimics an RA flare.

Osteoporosis ranks as an important long-term complication of chronic prednisone use. Based on a patient's risk factors, including total prednisone dosage, length of treatment, gender, race and bone density, treatment with a bisphosphonate may be appropriate for primary prevention of glucocorticoid-induced osteoporosis. Other agents, including teriparatide and denosomab, have been approved for the treatment of osteoporosis and may be indicated in certain cases. Although prednisone use is known to increase the risk of peptic ulcer disease, especially with concomitant NSAID use, no evidence-based guidelines have been published regarding the use of gastrointestinal ulcer prophylaxis in this situation.

DMARDs

DMARDs are so named because of their ability to slow or prevent structural progression of RA. The conventional DMARDs include hydroxychloroquine, sulfasalazine, methotrexate, and leflunomide; they exhibit a delayed onset of action of ~6–12 weeks. Methotrexate is the DMARD of choice for the treatment of RA and is the anchor drug for most combination therapies. It was approved for the treatment of RA in 1988 and remains the benchmark for the efficacy and safety of new disease-modifying therapies. At the dosages used for the treatment of RA, methotrexate has been shown to stimulate adenosine release from cells, producing an anti-inflammatory effect. The clinical efficacy of leflunomide, an inhibitor of pyrimidine synthesis, appears similar to that of methotrexate; it has been shown in well-designed trials to be effective for the treatment of RA as monotherapy or in combination with methotrexate and other DMARDs.

Although similar to the other DMARDs in its slow onset of action, hydroxychloroquine has not been shown to delay radiographic progression of disease and thus is not considered to be a true DMARD. In clinical practice, hydroxychloroquine is generally used for treatment of early, mild disease or as adjunctive therapy in combination with other DMARDs. Sulfasalazine is used in a similar manner and has been shown in randomized, controlled trials to reduce radiographic progression of disease. Minocycline, gold salts, penicillamine, azathioprine, and cyclosporine have all been used for the treatment of RA with varying degrees of success; however, they are used sparingly now due to their inconsistent clinical efficacy or unfavorable toxicity profile.

BIOLOGICALS

Biologic DMARDs have revolutionized the treatment of RA over the past decade (Table 351-2). They are protein therapeutics designed mostly to target cytokines and cell-surface molecules. The TNF inhibitors were the first biologicals approved for the treatment of RA. Anakinra, an IL-1 receptor antagonist, was approved shortly thereafter; however, its benefits have proved to be relatively modest compared with the other biologicals and therefore this biological is rarely used for the treatment of RA with the availability of other more effective agents. Abatacept, rituximab, and tocilizumab are the newest members of this class.

Anti-TNF Agents

The development of TNF inhibitors was originally spurred by the experimental finding that TNF is a critical upstream mediator of joint inflammation. Currently, five agents that inhibit TNF-α are approved for the treatment of RA. There are three different anti-TNF monoclonal antibodies. Infliximab is a chimeric (part mouse and human) monoclonal antibody, whereas adalimumab and golimumab are humanized monoclonal antibodies. Certolizumab pegol is a pegylated Fc-free fragment of a humanized monoclonal antibody with binding specificity for TNF-α. Lastly, etanercept is a soluble fusion protein comprising the TNF receptor 2 in covalent linkage with the Fc portion of IgG1. All of the TNF inhibitors have been shown in randomized controlled clinical trials to reduce the signs and symptoms of RA, slow radiographic progression of joint damage, and improve physical function and quality of life. Anti-TNF drugs are typically used in combination with background methotrexate therapy. This combination regimen, which affords maximal benefit in many cases, is often the next step for treatment of patients with an inadequate response to methotrexate therapy. Etanercept, adalimumab, certolizumab pegol, and golimumab have also been approved for use as monotherapy.

Anti-TNF agents should be avoided in patients with active infection or a history of hypersensitivity to these agents and are contraindicated in patients with chronic hepatitis B infection or class III/IV congestive heart failure. The major concern is the increased risk for infection, including serious bacterial infections, opportunistic fungal infection, and reactivation of latent tuberculosis. For this reason, all patients are screened for latent tuberculosis according to national guidelines prior to starting anti-TNF therapy **(Chap. 173)**. In the United States, patients are skin-tested using an intradermal injection of purified protein derivative (PPD); individuals with skin reactions of >5 mm are presumed to have had previous exposure to

tuberculosis and are evaluated for active disease and treated accordingly. Use of an IFN-γ release assay may also be appropriate for screening as some data suggest a lower rate of false-negative and false-positive tests with an IFN-γ release assay compared to skin testing with PPD in patients treated with corticosteroids. While a combination of PPD skin test and IFN-γ release assay may offer the highest sensitivity for screening purposes, no consensus guidelines exist.

Anakinra

Anakinra is the recombinant form of the naturally occurring IL-1 receptor antagonist. Although anakinra has seen limited use for the treatment of RA, it has enjoyed a resurgence of late as an effective therapy of some rare inherited syndromes dependent on IL-1 production, including neonatal-onset inflammatory disease, Muckle-Wells syndrome, and familial cold urticaria, as well as systemic juvenile-onset inflammatory arthritis and adult-onset Still's disease. Anakinra should not be combined with an anti-TNF drug due to the high rate of serious infections observed with this regimen in a clinical trial.

Abatacept

Abatacept is a soluble fusion protein consisting of the extracellular domain of human CTLA-4 linked to the modified portion of human IgG. It inhibits the co-stimulation of T cells by blocking CD28-CD80/86 interactions and may also inhibit the function of antigen-presenting cells by reverse signaling through CD80 and CD86. Abatacept has been shown in clinical trials to reduce disease activity, slow radiographic progression of damage, and improve functional disability. Many patients receive abatacept in combination with methotrexate or another DMARD such as leflunomide. Abatacept therapy has been associated with an increased risk of infection.

Rituximab

Rituximab is a chimeric monoclonal antibody directed against CD20, a cell-surface molecule expressed by most mature B lymphocytes. It works by depleting B cells, which in turn, leads to a reduction in the inflammatory response by unknown mechanisms. These mechanisms may include a reduction in autoantibodies, inhibition of T cell activation, and alteration of cytokine production. Rituximab has been approved for the treatment of refractory RA in combination with methotrexate and has been shown to be more effective for patients with seropositive than seronegative disease. Rituximab therapy has been associated with mild to moderate infusion reactions as well as an increased risk of infection. Notably, there have been rare isolated reports of a potentially lethal brain disorder, progressive multifocal leukoencephalopathy (PML), in association with rituximab therapy, although the absolute risk of this complication appears to be very low in patients with RA. Most of these cases have occurred on a background of previous or current exposure to other potent immunosuppressive drugs.

Tocilizumab

Tocilizumab is a humanized monoclonal antibody directed against the membrane and soluble forms of the IL-6 receptor. IL-6 is a proinflammatory cytokine implicated in the pathogenesis of RA, with effects on both joint inflammation and damage. IL-6 binding to its receptor activates intracellular signaling pathways that affect the acute-phase response, cytokine production, and osteoclast activation. Clinical trials attest to the clinical efficacy of tocilizumab therapy for RA, both as monotherapy and in combination with methotrexate and other DMARDs. Tocilizumab has been associated with an increased risk of infection, neutropenia, and thrombocytopenia; the hematologic abnormalities appear to be reversible upon stopping the drug. In addition, this agent has been shown to increase LDL cholesterol. However, it is not known as yet if this effect on lipid levels increases the risk for development of atherosclerotic disease.

SMALL-MOLECULE INHIBITORS

Because some patients do not adequately respond to conventional DMARDs or biologic therapy, other therapeutic targets have been investigated to fill this gap. Recently, drug development in RA has focused attention on the intracellular signaling pathways that transduce the positive signals of cytokines and other inflammatory mediators that create the positive feedback loops in the immune response. These synthetic DMARDs aim to provide the same efficacy as biological therapies in an oral formulation.

Tofacitinib

Tofacitinib is a small-molecule inhibitor that primarily inhibits JAK1 and JAK3, which mediate signaling of the receptors for the common γ-chain-related cytokines IL-2, 4, 7, 9, 15, and 21 as well as IFN-γ and IL-6. These cytokines all play roles in promoting T and B cell activation as well as inflammation. Tofacitinib, an oral agent, has been shown in randomized, placebo-controlled clinical trials to improve the signs and symptoms of RA significantly over placebo. Possible side effects include elevated serum transaminases indicative of liver injury, neutropenia, increased cholesterol levels, and elevation in serum creatinine. Its use is also associated with an increased risk of infections. Tofacitinib can be used as monotherapy or in combination with methotrexate.

TREATMENT OF EXTRAARTICULAR MANIFESTATIONS

In general, treatment of the underlying RA favorably modifies extraarticular manifestations, and it appears that aggressive management of early disease can potentially prevent their occurrence in the first place. RA-ILD, however, can be particularly challenging to treat because some of the DMARDs used for the treatment of RA are associated with pulmonary toxicity, such as methotrexate and leflunomide. High doses of corticosteroids and adjunctive immunosuppressive agents, such as azathioprine, mycophenolate mofetil, and rituximab have been used for treatment of RA-ILD.

TABLE 351-2

DMARDs Used for the Treatment of Rheumatoid Arthritis

Drug	Dosage	Serious Toxicities	Other Common Side Effects	Initial Evaluation	Monitoring
Hydroxychloroquine	200–400 mg/d orally (≤5 mg/kg)	Irreversible retinal damage Cardiotoxicity Blood dyscrasia	Nausea Diarrhea Headache Rash	Eye examination if >40 years old or prior ocular disease	Optical coherence tomography and visual field testing every 12 months
Sulfasalazine	Initial: 500 mg orally twice daily Maintenance: 1000–1500 mg twice daily	Granulocytopenia Hemolytic anemia (with G6PD deficiency)	Nausea Diarrhea Headache	CBC, LFTs G6PD level	CBC every 2–4 weeks for first 3 months, then every 3 months
Methotrexate	10–25 mg/week orally or SQ Folic acid 1 mg/d to reduce toxicities	Hepatotoxicity Myelosuppression Infection Interstitial pneumonitis Pregnancy category X	Nausea Diarrhea Stomatitis/mouth ulcers Alopecia Fatigue	CBC, LFTs Viral hepatitis panel[a] Chest x-ray	CBC, creatinine, LFTs every 2–3 months
Leflunomide	10–20 mg/d	Hepatotoxicity Myelosuppression Infection Pregnancy category X	Alopecia Diarrhea	CBC, LFTs Viral hepatitis panel[a]	CBC, creatinine, LFTs every 2–3 months
TNF-α Inhibitors	Infliximab: 3 mg/kg IV at weeks 0, 2, 6, then every 8 weeks. May increase dose up to 10 mg/kg every 4 weeks	↑ Risk bacterial, fungal infections Reactivation of latent TB ↑ Lymphoma risk (controversial) Drug-induced lupus Neurologic deficits	Infusion reaction ↑ LFTs	PPD skin test	LFTs periodically
	Etanercept: 50 mg SQ weekly, or 25 mg SQ biweekly	As above	Injection site reaction	PPD skin test	Monitor for injection site reactions
	Adalimumab: 40 mg SQ every other week	As above	Injection site reaction	PPD skin test	Monitor for injection site reactions
	Golimumab: 50 mg SQ monthly	As above	Injection site reaction	PPD skin test	Monitor for injection site reactions
	Certolizumab: 400 mg SQ weeks 0, 2, 4, then 200 mg every other week	As above	Injection site reaction	PPD skin test	Monitor for injection site reactions

Drug	Dosage	Serious Toxicities	Other Common Side Effects	Initial Evaluation	Monitoring
Abatacept	Weight based: <60 kg: 500 mg 60–100 kg: 750 mg >100 kg: 1000 mg IV dose at weeks 0, 2, and 4, and then every 4 weeks OR 125 mg SQ weekly	↑ Risk bacterial, viral infections	Headache Nausea	PPD skin test	Monitor for infusion reactions
Anakinra	100 mg SQ daily	↑ Risk bacterial, viral infections Reactivation of latent TB Neutropenia	Injection site reaction Headache	PPD skin test CBC with differential	CBC every month for 3 months, then every 4 months for 1 year Monitor for injection site reactions
Rituximab	1000 mg IV × 2, days 0 and 14 May repeat course every 24 weeks or more Premedicate with methyl-prednisolone 100 mg to decrease infusion reaction	↑ Risk bacterial, viral infections Infusion reaction Cytopenia Hepatitis B reactivation	Rash Fever	CBC Viral hepatitis panel[a]	CBC at regular intervals
Tocilizumab	4–8 mg/kg 4–8 mg/kg IV monthly OR 162 mg SQ every other week (<100 kg weight) 162 mg SQ every week (≥100 kg weight)	Risk of infection Infusion reaction LFT elevation Dyslipidemia Cytopenias		PPD skin test	CBC and LFTs at regular intervals
Tofacitinib	5 mg orally BID OR 11 mg orally daily	Risk of infection LFT elevation Dyslipidemia Neutropenia	Upper respiratory tract infections Diarrhea Headache Nasopharyngitis	PPD skin test	CBC, LFTs, and lipids at regular intervals

[a]Viral hepatitis panel: hepatitis B surface antigen, hepatitis C viral antibody.

Abbreviations: CBC, complete blood count; DMARDs, disease-modifying antirheumatic drugs; G6PD, glucose-6-phosphate dehydrogenase; IV, intravenous; LFTs, liver function tests; PPD, purified protein derivative; SQ, subcutaneous; TB, tuberculosis.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Rheumatoid Arthritis

The original treatment pyramid for RA is now considered to be obsolete and has evolved into a new strategy that focuses on several goals: (1) early, aggressive therapy to prevent joint damage and disability; (2) frequent modification of therapy with utilization of combination therapy where appropriate; (3) individualization of therapy in an attempt to maximize response and minimize side effects; and (4) achieving, whenever possible, remission of clinical disease activity. A considerable amount of evidence supports this intensive treatment approach.

As mentioned earlier, methotrexate is the DMARD of first choice for initial treatment of moderate to severe RA. Failure to achieve adequate improvement with methotrexate therapy calls for a change in DMARD therapy, usually a transition to an effective combination regimen. Effective

combinations include: methotrexate, sulfasalazine, and hydroxychloroquine (oral triple therapy); methotrexate and leflunomide; and methotrexate plus a biological. The combination of methotrexate and an anti-TNF agent, for example, has been shown in randomized, controlled trials to be superior to methotrexate alone not only for reducing signs and symptoms of disease, but also for retarding the progression of structural joint damage. Predicting which patients are at higher risk for developing radiologic joint damage is imprecise at best, although some factors such as an elevated serum level of acute-phase reactants, high burden of joint inflammation, and the presence of erosive disease are associated with increased likelihood of developing structural injury.

In 2015, the American College of Rheumatology updated and published their guidelines for the treatment of RA. They do make a distinction in the treatment of patients with early (<6 months of disease duration) and established disease and highlight the use of a treat-to-target approach and the need to switch or add therapies for worsening or persistent moderate/high disease activity. For example, in patients with early RA who have persistent moderate/high disease activity on DMARD monotherapy, providers should consider escalation to combination DMARD therapy or switching to an anti-TNF +/– methotrexate or a non-TNF biologic +/– methotrexate. Since a more intensive initial approach (e.g., combination DMARD therapy) has been shown to produce superior long-term outcomes compared with starting methotrexate alone, the usual approach is to begin with methotrexate and rapidly step-up (e.g., after 3–6 months) to combination of DMARD therapy or an anti-TNF or non-TNF biologic agent in the absence of an inadequate therapeutic response.

Some patients may not respond to an anti-TNF drug or may be intolerant of its side effects. Initial responders to an anti-TNF agent that later worsen may benefit from switching to another anti-TNF agent or an alternative biologic with a different mechanism of action. Indeed, some studies suggest that switching to an alternative biologic such as abatacept is more effective than switching to another anti-TNF drug. Unacceptable toxicity from an anti-TNF agent may also call for switching to another biological with a different mechanism of action or a conventional DMARD regimen.

Studies have also shown that oral triple therapy (hydroxychloroquine, methotrexate, and sulfasalazine) may be used effectively for the treatment of early RA. Treatment may be initiated with methotrexate alone and lacking an adequate treatment response followed within 6 months by a step-up to oral triple therapy.

A clinical state defined as low disease activity or remission is the optimal goal of therapy, although most patients never achieve complete remission despite every effort to achieve it. Composite indices, such as the Disease Activity Score-28 (DAS-28), are useful for classifying states of low disease activity and remission; however, they are imperfect tools due to the limitations of the clinical joint examination in which low-grade synovitis may escape detection. Complete remission has been stringently defined as the total absence of all articular and extraarticular inflammation and immunologic activity related to RA. However, evidence for this state can be difficult to demonstrate in clinical practice. In an effort to standardize and simplify the definition of remission for clinical trials, the ACR and EULAR developed two provisional operational definitions of remission in RA **(Table 351-3)**. A patient may be considered in remission if he or she (1) meets all of the clinical and laboratory criteria listed in Table 351-3 or (2) has a composite SDAI score of <3.3. The SDAI is calculated by taking the sum of a tender joint and swollen joint count (using 28 joints), patient global assessment (0–10 scale), physician global assessment (0–10 scale), and CRP (in mg/dL). This definition of remission does not take into account the possibility of subclinical synovitis or that damage alone may produce a tender or swollen joint. Ignoring the semantics of these definitions, the aforementioned remission criteria are nonetheless useful for setting a level of disease control that will likely result in minimal or no progression of structural damage and disability.

PHYSICAL THERAPY AND ASSISTIVE DEVICES

In principle, all patients with RA should receive a prescription for exercise and physical activity. Dynamic strength training, community-based comprehensive physical therapy, and physical-activity coaching (emphasizing 30 min of moderately intensive activity most days a week) have all been shown to improve muscle strength and perceived health status. Foot orthotics for painful valgus deformity decrease foot pain and may reduce disability and functional limitations. Judicious use of wrist splints can also decrease pain; however, their benefits may be offset by decreased dexterity and variably curb grip strength.

SURGERY

Surgical procedures may improve pain and disability in RA with varying degrees of reported long-term success—most notably the hands, wrists, and feet. For large joints, such as the knee, hip, shoulder, or elbow, the preferred option for advanced joint disease may be total joint arthroplasty. A few surgical options exist for dealing with the smaller hand joints. Silicone implants are the most common prosthetic for MCP arthroplasty and are generally implanted in patients with severe decreased arc of motion, marked flexion contractures, MCP joint pain with radiographic abnormalities, and severe ulnar drift. Arthrodesis and total wrist arthroplasty are reserved for patients with severe disease who have substantial pain and functional impairment. These two procedures appear to have equal efficacy in terms of pain control and patient satisfaction. Numerous surgical options exist for correction of hallux valgus in the forefoot, including arthrodesis and arthroplasty, as well as primarily arthrodesis for refractory hindfoot pain.

OTHER MANAGEMENT CONSIDERATIONS

Pregnancy

Up to 75% of female RA patients will note overall improvement in symptoms during pregnancy, but often will flare after delivery. Flares during pregnancy are generally treated with low doses of prednisone; hydroxychloroquine and sulfasalazine are probably the safest DMARDs to use during pregnancy. Methotrexate and leflunomide therapy are contraindicated during pregnancy due to their teratogenicity in animals and humans. The experience with biologic agents has been insufficient to make specific recommendations for their use during pregnancy. Ideally, their use should be avoided, but controlling active RA during pregnancy may take precedence in some cases.

Elderly Patients

RA presents in up to one-third of patients after the age of 60; however, older individuals may receive less aggressive treatment due to concerns about increased risks of drug toxicity. Studies suggest that conventional DMARDs and biologic agents are equally effective and safe in younger and older patients. Due to comorbidities, many elderly patients have an increased risk of infection. Aging also leads to a gradual decline in renal function that may raise the risk for side effects from NSAIDs and some DMARDS, such as methotrexate. Renal function must be taken into consideration before prescribing methotrexate, which is mostly cleared by the kidneys. To reduce the risks of side effects, methotrexate doses may need to be adjusted downward for the drop in renal function that usually comes with the seventh and eighth decades of life. Methotrexate is usually not prescribed for patients with a serum creatinine >2 mg/dL.

TABLE 351-3

ACR/EULAR Provisional Definition of Remission in Rheumatoid Arthritis

At any time point, patient must satisfy all of the following:
Tender joint count ≤1
Swollen joint count ≤1
C-reactive protein ≤1 mg/dL
Patient global assessment ≤1 (on a 0–10 scale)
OR
At any time point, patient must have a Simplified Disease Activity Index score of ≤3.3

Source: Adapted from DT Felson et al: Arthritis Rheum 63:573, 2011.

GLOBAL CHALLENGES

Developing countries are finding an increase in the incidence of noncommunicable, chronic diseases such as diabetes, cardiovascular disease, and RA in the face of ongoing poverty, rampant infectious disease, and poor access to modern health care facilities. In these areas, patients tend to have a greater delay in diagnosis and limited access to specialists, and thus greater disease activity and disability at presentation. In addition, infection risk remains a significant issue for the treatment of RA in developing countries because of the immunosuppression associated with the use of glucocorticoids and most DMARDs. For example, in some developing countries, patients undergoing treatment for RA have a substantial increase in the incidence of tuberculosis, which demands the implementation of far more comprehensive screening practices and liberal use of isoniazid prophylaxis than in developed countries. The increased prevalence of hepatitis B and C, as well as human immunodeficiency virus (HIV), in these developing countries also poses challenges. Reactivation of viral hepatitis has been observed in association with some of the DMARDs, such as rituximab. Also, reduced access to antiretroviral therapy may limit the control of HIV infection and therefore the choice of DMARD therapies.

Despite these challenges, one should attempt to initiate early treatment of RA in the developing countries with the resources at hand. Hydroxychloroquine, sulfasalazine and methotrexate are all reasonably accessible throughout the world where they can be used as both monotherapy and in combination with other drugs. The use of biologic agents is increasing in the developed countries as well as in other areas around the world, although their use is limited by high cost; national protocols restrict their use, and concerns remain about the risk for opportunistic infections.

SUMMARY

Improved understanding of the pathogenesis of RA and its treatment has dramatically revolutionized the management of this disease. The outcomes of patients with RA are vastly superior to those of the prebiologic modifier era; more patients than in years past are able to avoid significant disability and continue working, albeit with some job modifications in many cases. The need for early and aggressive treatment of RA as well as frequent follow-up visits for monitoring of drug therapy has implications for our health care system. Primary care physicians and rheumatologists must be prepared to work together as a team to reach the ambitious goals of best practice. In many settings, rheumatologists have reengineered their practice in a way that places high priority on consultations for any new patient with early inflammatory arthritis.

The therapeutic regimens for RA are becoming increasingly complex with the rapidly expanding armamentarium. Patients receiving these therapies must be carefully monitored by both the primary care physician and the rheumatologist to minimize the risk of side effects and identify quickly any complications of chronic immunosuppression. Also, prevention and treatment of RA-associated conditions such as ischemic heart disease and osteoporosis will likely benefit from a team approach owing to the value of multidisciplinary care.

Research will continue to search for new therapies with superior efficacy and safety profiles and investigate treatment strategies that can bring the disease under control more rapidly and nearer to remission. However, prevention and cure of RA will likely require new breakthroughs in our understanding of disease pathogenesis. These insights may come from genetic studies illuminating critical pathways in the mechanisms of joint inflammation. Equally ambitious is the lofty goal of biomarker discovery that will open the door to personalized medicine for the care of patients with RA.

FURTHER READING

Catrina AI et al: Lungs, joints and immunity against citrullinated proteins in rheumatoid arthritis. Nature Rev Rheumatol 10:645, 2014.

Erickson AR et al: Clinical features of rheumatoid arthritis, in *Kelley and Firestein's Textbook of Rheumatology*, 10th ed, Firestein GS et al (eds). Philadelphia, Elsevier, 2017, pp 1167–1186.

McInnes IB, Schett G: The pathogenesis of rheumatoid arthritis. N Engl J Med 365:2205, 2011. [PubMed: 22150039]

Moreland LW et al: A randomized comparative effectiveness study of oral triple therapy versus methotrexate plus etanercept in early aggressive rheumatoid arthritis: The Treatment of Early Aggressive Rheumatoid Arthritis Trial. Arthritis Rheum 64:2824, 2012. [PubMed: 22508468]

Singh JA et al: 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. Arthritis Rheumatol 68:1, 2016. [PubMed: 26545940]

Yarwood A, Huizinga TW, Worthington J: The genetics of rheumatoid arthritis: Risk and protection in different stages of the evolution of RA. Rheumatology (Oxford) 55:199, 2016. [PubMed: 25239882]

Chapter 352: Acute Rheumatic Fever

Jonathan R. Carapetis

INTRODUCTION

Acute rheumatic fever (ARF) is a multisystem disease resulting from an autoimmune reaction to infection with group A streptococcus. Although many parts of the body may be affected, almost all of the manifestations resolve completely. The major exception is cardiac valvular damage (rheumatic heart disease [RHD]), which may persist after the other features have disappeared.

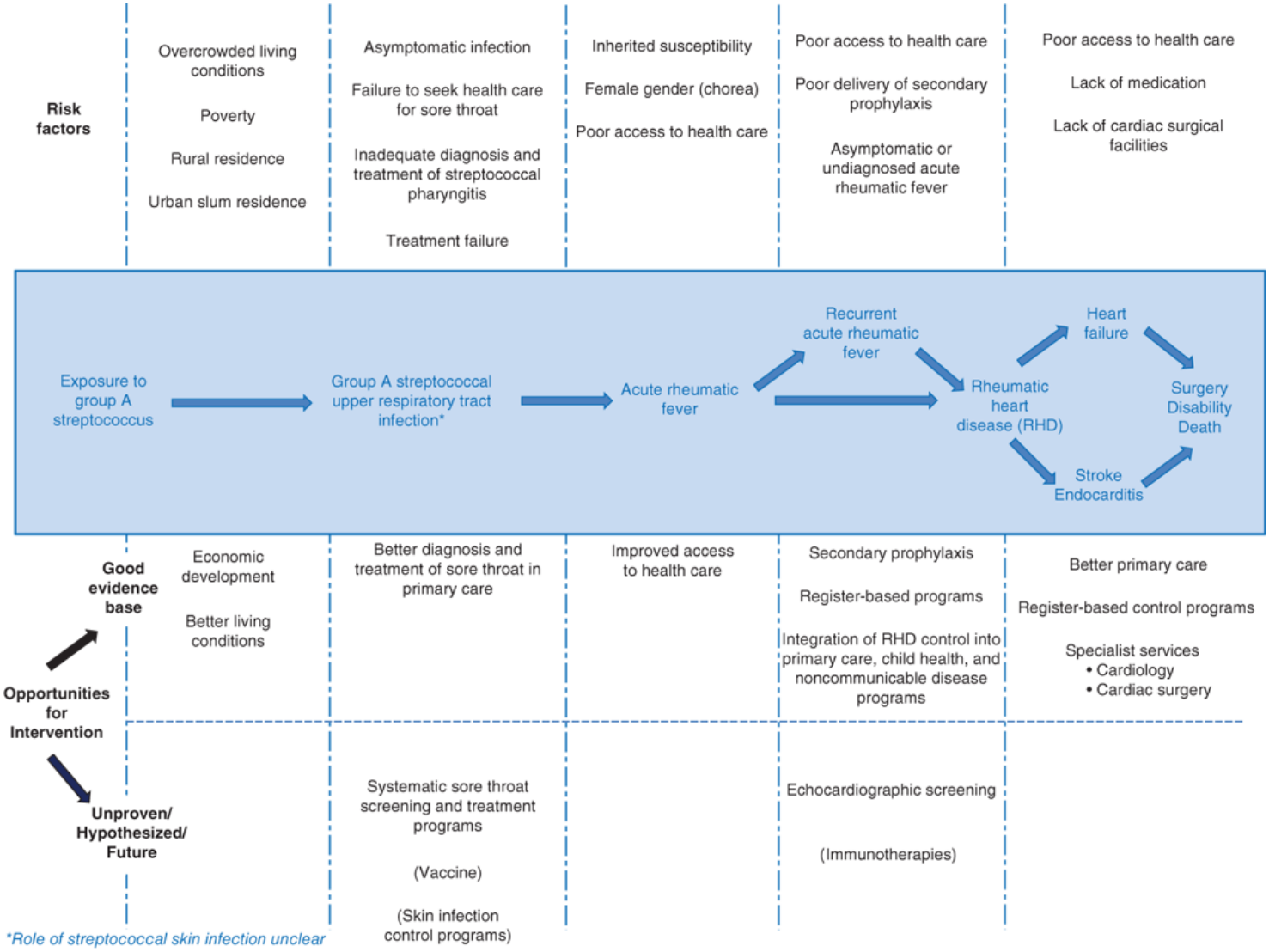
GLOBAL CONSIDERATIONS

ARF and RHD are diseases of poverty. They were common in all countries until the early twentieth century, when their incidence began to decline in industrialized nations. This decline was largely attributable to improved living conditions—particularly less crowded housing and better hygiene—which resulted in reduced transmission of group A streptococci. The introduction of antibiotics and improved systems of medical care had a supplemental effect.

The virtual disappearance of ARF and reduction in the incidence of RHD in industrialized countries during the twentieth century unfortunately was not replicated in developing countries, where these diseases continue unabated. RHD is the most common cause of heart disease in children in developing countries and is a major cause of mortality and morbidity in adults as well. It has been estimated that between 15 and 19 million people worldwide are affected by RHD, with approximately one-quarter of a million deaths occurring each year. Some 95% of ARF cases and RHD deaths now occur in developing countries, with particularly high rates in sub-Saharan Africa, Pacific nations, Australasia, and South and Central Asia. The pathogenetic pathway from exposure to group A streptococcus followed by pharyngeal infection and subsequent development of ARF, ARF recurrences, and development of RHD and its complications is associated with a range of risk factors and, therefore, potential interventions at each point (Fig. 352-1). In affluent countries, many of these risk factors are well controlled, and where needed, interventions are in place. Unfortunately, the greatest burden of disease is found in developing countries, most of which do not have the resources, capacity, and/or interest to tackle this multifaceted disease. In particular, almost none of the developing countries has a coordinated, register-based RHD control program, which is proven to be cost-effective in reducing the burden of RHD. Enhancing awareness of RHD and mobilizing resources for its control in developing countries are issues requiring international attention.

FIGURE 352-1

Pathogenetic pathway for acute rheumatic fever and rheumatic heart disease, with associated risk factors and opportunities for intervention at each step. Interventions in *parentheses* are either unproven or currently unavailable.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EPIDEMIOLOGY

ARF is mainly a disease of children age 5–14 years. Initial episodes become less common in older adolescents and young adults and are rare in persons aged >30 years. By contrast, recurrent episodes of ARF remain relatively common in adolescents and young adults. This pattern contrasts with the prevalence of RHD, which peaks between 25 and 40 years. There is no clear gender association for ARF, but RHD more commonly affects females, sometimes up to twice as frequently as males.

PATHOGENESIS

ORGANISM FACTORS

Based on currently available evidence, ARF is exclusively caused by infection of the upper respiratory tract with group A streptococci (Chap. 143). Although classically, certain M-serotypes (particularly types 1, 3, 5, 6, 14, 18, 19, 24, 27, and 29) were associated with ARF, in high-incidence regions, it is now thought that any strain of group A streptococcus has the potential to cause ARF. The potential role of skin infection and of groups C and G streptococci is currently being investigated.

HOST FACTORS

Approximately 3–6% of any population may be susceptible to ARF, and this proportion does not vary dramatically between populations. Findings of familial clustering of cases and concordance in mono-zygotic twins—particularly for chorea—confirm that susceptibility to ARF is an inherited characteristic, with 44% concordance in mono-zygotic twins compared to 12% in dizygotic twins, and heritability more recently estimated at 60%. Most evidence for host factors focuses on immunologic determinants. Some human leukocyte antigen (HLA) class II alleles, particularly HLA-DR7 and HLA-DR4,

appear to be associated with susceptibility, whereas other class II alleles have been associated with protection (HLA-DR5, HLA-DR6, HLA-DR51, HLA-DR52, and HLA-DQ). Associations have also been described with polymorphisms at the tumor necrosis factor α locus (TNF-α-308 and TNF-α-238), high levels of circulating mannose-binding lectin, and Toll-like receptors.

THE IMMUNE RESPONSE

The most widely accepted theory of rheumatic fever pathogenesis is based on the concept of molecular mimicry, whereby an immune response targeted at streptococcal antigens (mainly thought to be on the M protein and the *N*-acetylglucosamine of group A streptococcal carbohydrate) also recognizes human tissues. In this model, cross-reactive antibodies bind to endothelial cells on the heart valve, leading to activation of the adhesion molecule VCAM-1, with resulting recruitment of activated lymphocytes and lysis of endothelial cells in the presence of complement. The latter leads to release of peptides including laminin, keratin, and tropomyosin, which, in turn, activates cross-reactive T cells that invade the heart, amplifying the damage and causing epitope spreading. An alternative hypothesis proposes that the initial damage is due to streptococcal invasion of epithelial surfaces, with binding of M protein to type IV collagen allowing it to become immunogenic, but not through the mechanism of molecular mimicry.

CLINICAL FEATURES

There is a latent period of ~3 weeks (1–5 weeks) between the precipitating group A streptococcal infection and the appearance of the clinical features of ARF. The exceptions are chorea and indolent carditis, which may follow prolonged latent periods lasting up to 6 months. Although many patients report a prior sore throat, the preceding group A streptococcal infection is commonly subclinical; in these cases, it can only be confirmed using streptococcal antibody testing. The most common clinical features are polyarthritis (present in 60–75% of cases) and carditis (50–60%). The prevalence of chorea in ARF varies substantially between populations, ranging from <2 to 30%. Erythema marginatum and subcutaneous nodules are now rare, being found in <5% of cases.

HEART INVOLVEMENT

Up to 60% of patients with ARF progress to RHD. The endocardium, pericardium, or myocardium may be affected. Valvular damage is the hallmark of rheumatic carditis. The mitral valve is almost always affected, sometimes together with the aortic valve; isolated aortic valve involvement is rare. Damage to the pulmonary or tricuspid valves is usually secondary to increased pulmonary pressures resulting from left-sided valvular disease. Early valvular damage leads to regurgitation. Over ensuing years, usually as a result of recurrent episodes, leaflet thickening, scarring, calcification, and valvular stenosis may develop **(Fig. 352-2)**. **See Videos 352-1 and 352-2.** Therefore, the characteristic manifestation of carditis in previously unaffected individuals is mitral regurgitation, sometimes accompanied by aortic regurgitation. Myocardial inflammation may affect electrical conduction pathways, leading to P-R interval prolongation (first-degree atrioventricular block or rarely higher level block) and softening of the first heart sound.

VIDEO 352-1A: Transthoracic echocardiographic images of a 9-year-old girl with first episode of acute rheumatic fever.
Images demonstrate the typical echocardiographic findings of acute rheumatic carditis. The valve leaflets are relatively thin and highly mobile. The failure of coaptation of the mitral valve leaflets is the result of chordal elongation and annular dilatation. The mitral valve regurgitation is moderate with a typical posterolaterally directed regurgitant jet of rheumatic carditis. ***A.*** Acute rheumatic carditis (apical four-chamber view echocardiogram).

Image not available.
Play Video

VIDEO 352-1B: Transthoracic echocardiographic images of a 9-year-old girl with first episode of acute rheumatic fever.
Images demonstrate the typical echocardiographic findings of acute rheumatic carditis. The valve leaflets are relatively thin and highly mobile. The failure of coaptation of the mitral valve leaflets is the result of chordal elongation and annular dilatation. The mitral valve regurgitation is moderate with a typical posterolaterally directed regurgitant jet of rheumatic carditis. ***B.*** Acute rheumatic carditis (apical four-chamber view color Doppler echocardiogram).

Image not available.
Play Video

VIDEO 352-1C: Transthoracic echocardiographic images of a 9-year-old girl with first episode of acute rheumatic fever.
Images demonstrate the typical echocardiographic findings of acute rheumatic carditis. The valve leaflets are relatively thin and highly mobile. The failure of coaptation of the mitral valve leaflets is the result of chordal elongation and annular dilatation. The mitral valve regurgitation is moderate with a typical posterolaterally directed regurgitant jet of rheumatic carditis. ***C.*** Acute rheumatic carditis (parasternal long-axis view echocardiogram).

Image not available.
Play Video

VIDEO 352-1D: Transthoracic echocardiographic images of a 9-year-old girl with first episode of acute rheumatic fever.
Images demonstrate the typical echocardiographic findings of acute rheumatic carditis. The valve leaflets are relatively thin and highly mobile. The failure of coaptation of the mitral valve leaflets is the result of chordal elongation and annular dilatation. The mitral valve regurgitation is moderate with a typical posterolaterally directed regurgitant jet of rheumatic carditis. ***D.*** Acute rheumatic carditis (parasternal long-axis view color Doppler echocardiogram).

Image not available.
Play Video

VIDEO 352-2A: Transthoracic echocardiographic images are from a 5-year-old boy with chronic rheumatic heart disease with severe mitral valve regurgitation and moderate mitral valve stenosis.

Images demonstrate the typical echocardiographic findings in advanced chronic rheumatic heart disease. Both the anterior and posterior mitral valve leaflets are markedly thickened. During diastole, the motion of the anterior mitral valve leaflet tip is restricted with doming of the body of the leaflet toward the interventricular septum. This appearance is commonly described as a “hockey stick” or an “elbow” deformity. ***A.*** Chronic rheumatic heart disease (parasternal long-axis view).

Image not available.
Play Video

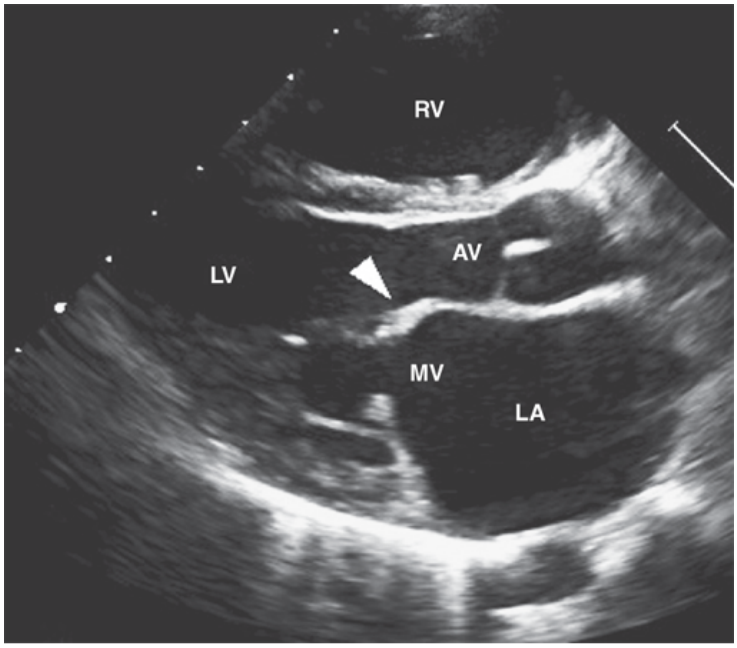
VIDEO 352-2B: Transthoracic echocardiographic images are from a 5-year-old boy with chronic rheumatic heart disease with severe mitral valve regurgitation and moderate mitral valve stenosis.

Images demonstrate the typical echocardiographic findings in advanced chronic rheumatic heart disease. Both the anterior and posterior mitral valve leaflets are markedly thickened. During diastole, the motion of the anterior mitral valve leaflet tip is restricted with doming of the body of the leaflet toward the interventricular septum. This appearance is commonly described as a “hockey stick” or an “elbow” deformity. ***B.*** Chronic rheumatic heart disease (apical two-chamber view echocardiogram).

Image not available.
Play Video

FIGURE 352-2

Transthoracic echocardiographic image from a 5-year-old boy with chronic rheumatic heart disease. This diastolic image demonstrates leaflet thickening, restriction of the anterior mitral valve leaflet tip and doming of the body of the leaflet toward the interventricular septum. This appearance (marked by the *arrowhead*) is commonly described as a “hockey stick” or an “elbow” deformity. AV, aortic valve; LA, left atrium; LV, left ventricle; MV, mitral valve; RV, right ventricle. *(Courtesy of Dr. Bo Remenyi, Department of Paediatric and Congenital Cardiac Services, Starship Children’s Hospital, Auckland, New Zealand.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

People with RHD are often asymptomatic for many years before their valvular disease progresses to cause cardiac failure. Moreover, particularly in resource-poor settings, the diagnosis of ARF is often not made, so children, adolescents, and young adults may have RHD but not know it. These cases can be diagnosed using echocardiography; auscultation is poorly sensitive and specific for RHD diagnosis in asymptomatic patients. Echocardiographic screening of school-aged children in populations with high rates of RHD is becoming more widespread and has been facilitated by improving technologies in portable echocardiography and the availability of consensus guidelines for the diagnosis of RHD on echocardiography **(Table 352-1)**. Although a diagnosis of definite RHD on screening echocardiography should lead to commencement of secondary prophylaxis, the clinical significance of borderline RHD has yet to be determined.

TABLE 352-1

World Heart Federation Criteria for Echocardiographic Diagnosis of Rheumatic Heart Disease (RHD) in Individuals <20 Years of Age[a]

Definite RHD (either A, B, C, or D)
(A) Pathologic MR and at least two morphologic features of RHD of the mitral valve
(B) MS mean gradient ≥4 mmHg (note: congenital MV anomalies must be excluded)
(C) Pathologic AR and at least two morphologic features of RHD of the AV (note: bicuspid AV and dilated aortic root must be excluded)
(D) Borderline disease of both the MV and AV
Borderline RHD (either A, B, or C)
(A) At least two morphologic features of RHD of the MV without pathologic MR or MS
(B) Pathologic MR
(C) Pathologic AR
Normal Echocardiographic Findings (all of A, B, C, and D)
(A) MR that does not meet all four Doppler criteria (physiologic MR)
(B) AR that does not meet all four Doppler criteria (physiologic AR)
(C) An isolated morphologic feature of RHD of the MV (e.g., valvular thickening), without any associated pathologic stenosis or regurgitation
(D) Morphologic feature of RHD of the AV (e.g., valvular thickening), without any associated pathologic stenosis or regurgitation
Definitions of Pathologic Regurgitation and Morphologic Features of RHD
Pathologic MR: All of the following: seen in two views; in at least one view, jet length 2 cm; peak velocity ≥3 m/s; pansystolic jet in at least one envelope
Pathologic AR: All of the following: seen in two views; in at least one view, jet length ≥1 cm; peak velocity ≥3 m/s; pandiastolic jet in at least one envelope
Morphologic features of RHD in MV: anterior MV leaflet thickening ≥3 mm (age specific); chordal thickening; restricted leaflet motion; excessive leaflet tip motion during systole
Morphologic features of RHD in AV: irregular or focal thickening; coaptation defect; restricted leaflet motion; prolapse

[a]For criteria in individuals >20 years of age, see source document.

Abbreviations: AR, aortic regurgitation; AV, aortic valve; MR, mitral regurgitation; MS, mitral stenosis; MV, mitral valve.

Source: Adapted from B Remenyi et al: World Heart Federation criteria for echocardiographic diagnosis of rheumatic heart disease-an evidence-based guideline. Nat Rev Cardiol 9:297–309, 2012.

JOINT INVOLVEMENT

The most common form of joint involvement in ARF is arthritis, i.e., objective evidence of inflammation, with hot, swollen, red, and/or tender joints, and involvement of more than one joint (i.e., polyarthritis). Polyarthritis is typically migratory, moving from one joint to another over a period of hours. ARF almost always affects the large joints—most commonly the knees, ankles, hips, and elbows—and is asymmetric. The pain is severe and usually disabling until anti-inflammatory medication is commenced.

Less severe joint involvement is also relatively common and has been recognized as a potential major manifestation in high-risk populations in the most recent revision of the Jones criteria. Arthralgia without objective joint inflammation usually affects large joints in the same migratory pattern as

polyarthritis. In some populations, aseptic monoarthritis may be a presenting feature of ARF, which may, in turn, result from early commencement of anti-inflammatory medication before the typical migratory pattern is established.

The joint manifestations of ARF are highly responsive to salicylates and other nonsteroidal anti-inflammatory drugs (NSAIDs). Indeed, joint involvement that persists for more than 1 or 2 days after starting salicylates is unlikely to be due to ARF.

CHOREA

Sydenham's chorea commonly occurs in the absence of other manifestations, follows a prolonged latent period after group A streptococcal infection, and is found mainly in females. The choreiform movements affect particularly the head (causing characteristic darting movements of the tongue) and the upper limbs **(Chap. 428)**. They may be generalized or restricted to one side of the body (hemi-chorea). In mild cases, chorea may be evident only on careful examination, whereas in the most severe cases, the affected individuals are unable to perform activities of daily living. There is often associated emotional lability or obsessive-compulsive traits, which may last longer than the choreiform movements (which usually resolve within 6 weeks but sometimes may take up to 6 months).

SKIN MANIFESTATIONS

The classic rash of ARF is *erythema marginatum* **(Chap. 16)**, which begins as pink macules that clear centrally, leaving a serpiginous, spreading edge. The rash is evanescent, appearing and disappearing before the examiner's eyes. It occurs usually on the trunk, sometimes on the limbs, but almost never on the face.

Subcutaneous nodules occur as painless, small (0.5–2 cm), mobile lumps beneath the skin overlying bony prominences, particularly of the hands, feet, elbows, occiput, and occasionally the vertebrae. They are a delayed manifestation, appearing 2–3 weeks after the onset of disease, last for just a few days up to 3 weeks, and are commonly associated with carditis.

OTHER FEATURES

Fever occurs in most cases of ARF, although rarely in cases of pure chorea. Although high-grade fever (≥39°C) is the rule, lower grade temperature elevations are not uncommon. Elevated acute-phase reactants are also present in most cases.

EVIDENCE OF A PRECEDING GROUP A STREPTOCOCCAL INFECTION

With the exception of chorea and low-grade carditis, both of which may become manifest many months later, evidence of a preceding group A streptococcal infection is essential in making the diagnosis of ARF. Because most cases do not have a positive throat swab culture or rapid antigen test, serologic evidence is usually needed. The most common serologic tests are the anti-streptolysin O (ASO) and anti-DNase B (ADB) titers. Where possible, age-specific reference ranges should be determined in a local population of healthy people without a recent group A streptococcal infection.

CONFIRMING THE DIAGNOSIS

Because there is no definitive test, the diagnosis of ARF relies on the presence of a combination of typical clinical features together with evidence of the precipitating group A streptococcal infection, and the exclusion of other diagnoses. This uncertainty led Dr. T. Duckett Jones in 1944 to develop a set of criteria (subsequently known as the *Jones criteria*) to aid in the diagnosis. The most recent revision of the Jones criteria **(Table 352-2)** require the clinician to determine if the patient is from a setting or population known to experience low rates of ARF. For this group, there is a set of "low-risk" criteria; for all others, there is a set of more sensitive criteria.

TABLE 352-2

Jones Criteria

A. For All Patient Populations with Evidence of Preceding Group A Streptococcal Infection	
Diagnosis: initial ARF	2 major manifestations or 1 major plus 2 minor manifestations
Diagnosis: recurrent ARF	2 major or 1 major and 2 minor or 3 minor
B. Major Criteria	
Low-risk populations[a]	Moderate- and high-risk populations
Carditis[b]	Carditis
• Clinical and/or subclinical	• Clinical and/or subclinical
Arthritis	Arthritis
• Polyarthritis only	• Monoarthritis or polyarthritis • Polyarthralgia[c]
Chorea	Chorea
Erythema marginatum	Erythema marginatum
SC nodules	SC nodules
C. Minor Criteria	
Low-risk populations[a]	Moderate- and high-risk populations
Polyarthralgia	Monoarthralgia
Fever (≥38.5°C)	Fever (≥38°C)
ESR ≥60 mm in the first hour and/or CRP ≥3.0 mg/dL[d]	ESR ≥30 mm/h and/or CRP ≥3.0 mg/dL[d]
Prolonged PR interval, after accounting for age variability (unless carditis is a major criterion)	Prolonged PR interval, after accounting for age variability (unless carditis is a major criterion)

[a]Low-risk populations are those with ARF incidence ≤2 per 100,000 school-age children or all-age rheumatic heart disease prevalence of ≤1 per 1000 population per year.

[b]Subclinical carditis indicates echocardiographic valvulitis. (See source document.)

[c]Polyarthralgia should only be considered as a major manifestation in moderate- to high-risk populations after exclusion of other causes. As in past versions of the criteria, erythema marginatum and SC nodules are rarely “stand-alone” major criteria. Additionally, joint manifestations can only be considered in either the major or minor categories but not both in the same patient. (See source document for more information.)

[d]CRP value must be greater than upper limit of normal for laboratory. Also, because ESR may evolve during the course of ARF, peak ESR values should be used.

Abbreviations: ARF, acute rheumatic fever; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate.

Source: From MH Gewitz et al: Revision of the Jones Criteria for the diagnosis of acute rheumatic fever in the era of Doppler echocardiography: A scientific statement from the American Heart Association. Circulation 131:1806, 2015.

TREATMENT

TREATMENT

Acute Rheumatic Fever

Patients with possible ARF should be followed closely to ensure that the diagnosis is confirmed, treatment of heart failure and other symptoms is undertaken, and preventive measures including commencement of secondary prophylaxis, inclusion on an ARF registry, and health education are commenced. Echocardiography should be performed on all possible cases to aid in making the diagnosis and to determine the severity at baseline of any carditis. Other tests that should be performed are listed in **Table 352-3**.

There is no treatment for ARF that has been proven to alter the likelihood of developing, or the severity of, RHD. With the exception of treatment of heart failure, which may be life-saving in cases of severe carditis, the treatment of ARF is symptomatic.

ANTIBIOTICS

All patients with ARF should receive antibiotics sufficient to treat the precipitating group A streptococcal infection **(Chap. 143)**. Penicillin is the drug of choice and can be given orally (as phenoxymethyl penicillin, 500 mg [250 mg for children ≤27 kg] PO twice daily, or amoxicillin, 50 mg/kg [maximum, 1 g] daily, for 10 days) or as a single dose of 1.2 million units (600,000 units for children ≤27 kg) IM benzathine penicillin G.

SALICYLATES AND NSAIDs

These may be used for the treatment of arthritis, arthralgia, and fever, once the diagnosis is confirmed. They are of no proven value in the treatment of carditis or chorea. Aspirin is the drug of choice, delivered at a dose of 50–60 mg/kg per day, up to a maximum of 80–100 mg/kg per day (4–8 g/d in adults) in 4–5 divided doses. At higher doses, the patient should be monitored for symptoms of salicylate toxicity such as nausea, vomiting, or tinnitus; if symptoms appear, lower doses should be used. When the acute symptoms are substantially resolved, usually within the first 2 weeks, patients on higher doses can have the dose reduced to 50–60 mg/kg per day for a further 2–4 weeks. Fever, joint manifestations, and elevated acute-phase reactants sometimes recur up to 3 weeks after the medication is discontinued. This does not indicate a recurrence and can be managed by recommencing salicylates for a brief period. Naproxen at a dose of 10–20 mg/kg per day is a suitable alternative to aspirin and has the advantage of twice-daily dosing.

CONGESTIVE HEART FAILURE

Glucocorticoids

The use of glucocorticoids in ARF remains controversial. Two meta-analyses have failed to demonstrate a benefit of glucocorticoids compared to placebo or salicylates in improving the short- or longer-term outcome of carditis. However, the studies included in these meta-analyses all took place >40 years ago and did not use medications in common usage today. Many clinicians treat cases of severe carditis (causing heart failure) with glucocorticoids in the belief that they may reduce the acute inflammation and result in more rapid resolution of failure. However, the potential benefits of this treatment should be balanced against the possible adverse effects. If used, prednisone or prednisolone is recommended at a dose of 1–2 mg/kg per day (maximum, 80 mg), usually for a few days or up to a maximum of 3 weeks.

MANAGEMENT OF HEART FAILURE

See Chap. 253.

BED REST

Traditional recommendations for long-term bed rest, once the cornerstone of management, are no longer widely practiced. Instead, bed rest should be prescribed as needed while arthritis and arthralgia are present and for patients with heart failure. Once symptoms are well controlled, gradual mobilization can commence as tolerated.

CHOREA

Medications to control the abnormal movements do not alter the duration or outcome of chorea. Milder cases can usually be managed by providing a calm environment. In patients with severe chorea, carbamazepine or sodium valproate is preferred to haloperidol. A response may not be seen for 1–2 weeks, and medication should be continued for 1–2 weeks after symptoms subside. There is recent evidence that corticosteroids are effective and lead to more rapid symptom reduction in chorea. They should be considered in severe or refractory cases. Prednisone or prednisolone may be commenced at 0.5 mg/kg daily, with weaning as early as possible, preferably after 1 week if symptoms are reduced, although slower weaning or temporary dose escalation may be required if symptoms worsen.

INTRAVENOUS IMMUNOGLOBULIN (IVIG)

Small studies have suggested that IVIg may lead to more rapid resolution of chorea but have shown no benefit on the short- or long-term outcome of carditis in ARF without chorea. In the absence of better data, IVIg is *not* recommended except in cases of severe chorea refractory to other treatments.

TABLE 352-3

Recommended Tests in Cases of Possible Acute Rheumatic Fever

Recommended for All Cases
White blood cell count
Erythrocyte sedimentation rate
C-reactive protein
Blood cultures if febrile
Electrocardiogram (if prolonged P-R interval or other rhythm abnormality, repeat in 2 weeks and again at 2 months if still abnormal)
Chest x-ray if clinical or echocardiographic evidence of carditis
Echocardiogram (consider repeating after 1 month if negative)
Throat swab (preferably before giving antibiotics)—culture for group A streptococcus
Antistreptococcal serology: both anti-streptolysin O and anti-DNase B titers, if available (repeat 10–14 days later if first test not confirmatory)
Tests for Alternative Diagnoses, Depending on Clinical Features
Repeated blood cultures if possible endocarditis
Joint aspirate (microscopy and culture) for possible septic arthritis
Copper, ceruloplasmin, antinuclear antibody, drug screen for choreiform movements
Serology and autoimmune markers for arboviral, autoimmune, or reactive arthritis

Source: Reprinted with permission from Menzies School of Health Research. Complete information located at *https://www.rhdaustralia.org.au/arf-rhd-guideline*.

PROGNOSIS

Untreated, ARF lasts on average 12 weeks. With treatment, patients are usually discharged from hospital within 1–2 weeks. Inflammatory markers should be monitored every 1–2 weeks until they have normalized (usually within 4–6 weeks), and an echocardiogram should be performed after 1 month to determine if there has been progression of carditis. Cases with more severe carditis need close clinical and echocardiographic monitoring in the longer term.

Once the acute episode has resolved, the priority in management is to ensure long-term clinical follow-up and adherence to a regimen of secondary prophylaxis. Patients should be entered onto the local ARF registry (if present) and contact made with primary care practitioners to ensure a plan for follow-up and administration of secondary prophylaxis before the patient is discharged. Patients and their families should also be educated about their disease, emphasizing the importance of adherence to secondary prophylaxis.

PREVENTION

PRIMARY PREVENTION

Ideally, primary prevention would entail elimination of the major risk factors for streptococcal infection, particularly overcrowded housing. This is difficult to achieve in most places where ARF is common.

Therefore, the mainstay of primary prevention for ARF remains primary prophylaxis (i.e., the timely and complete treatment of group A streptococcal sore throat with antibiotics). If commenced within 9 days of sore throat onset, a course of penicillin (as outlined above for treatment of ARF) will prevent almost all cases of ARF that would otherwise have developed. In settings where ARF and RHD are common but microbiologic diagnosis of group A

streptococcal pharyngitis is not available, such as in resource-poor countries, primary care guidelines often recommend that all patients with sore throat be treated with penicillin or, alternatively, that a clinical algorithm be used to identify patients with a higher likelihood of group A streptococcal pharyngitis. Although imperfect, such approaches recognize the importance of ARF prevention at the expense of overtreating many cases of sore throat that are not caused by group A streptococcus.

SECONDARY PREVENTION

The mainstay of controlling ARF and RHD is secondary prevention. Because patients with ARF are at dramatically higher risk than the general population of developing a further episode of ARF after a group A streptococcal infection, they should receive long-term penicillin prophylaxis to prevent recurrences. The best antibiotic for secondary prophylaxis is benzathine penicillin G (1.2 million units, or 600,000 units if ≤27 kg) delivered every 4 weeks. It can be given every 3 weeks, or even every 2 weeks, to persons considered to be at particularly high risk, although in settings where good compliance with an every-4-week dosing schedule can be achieved, more frequent dosing is rarely needed. Oral penicillin V (250 mg) can be given twice daily instead but is less effective than benzathine penicillin G. Penicillin-allergic patients can receive erythromycin (250 mg) twice daily.

The duration of secondary prophylaxis is determined by many factors, in particular the duration since the last episode of ARF (recurrences become less likely with increasing time), age (recurrences are less likely with increasing age), and the severity of RHD (if severe, it may be prudent to avoid even a very small risk of recurrence because of the potentially serious consequences) **(Table 352-4)**. Secondary prophylaxis is best delivered as part of a coordinated RHD control program, based around a registry of patients. Registries improve the ability to follow patients and identify those who default from prophylaxis and to institute strategies to improve adherence.

TABLE 352-4

American Heart Association Recommendations for Duration of Secondary Prophylaxis[a]

Category of Patient	Duration of Prophylaxis
Rheumatic fever without carditis	For 5 years after the last attack or 21 years of age (whichever is longer)
Rheumatic fever with carditis but no residual valvular disease	For 10 years after the last attack, or 21 years of age (whichever is longer)
Rheumatic fever with persistent valvular disease, evident clinically or on echocardiography	For 10 years after the last attack, or 40 years of age (whichever is longer); sometimes lifelong prophylaxis

[a]These are only recommendations and must be modified by individual circumstances as warranted. Note that some organizations recommend a minimum of 10 years of prophylaxis after the most recent episode, or until 21 years of age (whichever is longer), regardless of the presence of carditis with the initial episode.

Source: Adapted from AHA Scientific Statement Prevention of Rheumatic Fever and Diagnosis and Treatment of Acute Streptococcal Pharyngitis. Circulation 119:1541, 2009.

FURTHER READING

Carapetis JR et al: Acute rheumatic fever and rheumatic heart disease. Nat Rev Dis Primers 14:15084, 2016.

Gewitz MH et al: Revision of the Jones Criteria for the diagnosis of acute rheumatic fever in the era of Doppler echocardiography: A scientific statement from the American Heart Association. Circulation 131:1806, 2015. [PubMed: 25908771]

RHDAUSTRALIA (ARF/RHD Writing Group): The Australian guideline for prevention, diagnosis and management of acute rheumatic fever and rheumatic heart disease (2nd edition). Darwin: Menzies School of Health Research; 2012. Available at *https://www.rhdaustralia.org.au/arf-rhd-guideline*.

Steer AC et al: Status of research and development of vaccines for Streptococcus pyogenes. Vaccine 34:2953, 2016. [PubMed: 27032515]

Zühlke L et al: Clinical outcomes in 3343 children and adults with rheumatic heart disease from 14 low- and middle-income countries: Two-year follow-up of the Global Rheumatic Heart Disease Registry (the REMEDY Study). Circulation 134:1456, 2016. [PubMed: 27702773]

Chapter 353: Systemic Sclerosis (Scleroderma) and Related Disorders

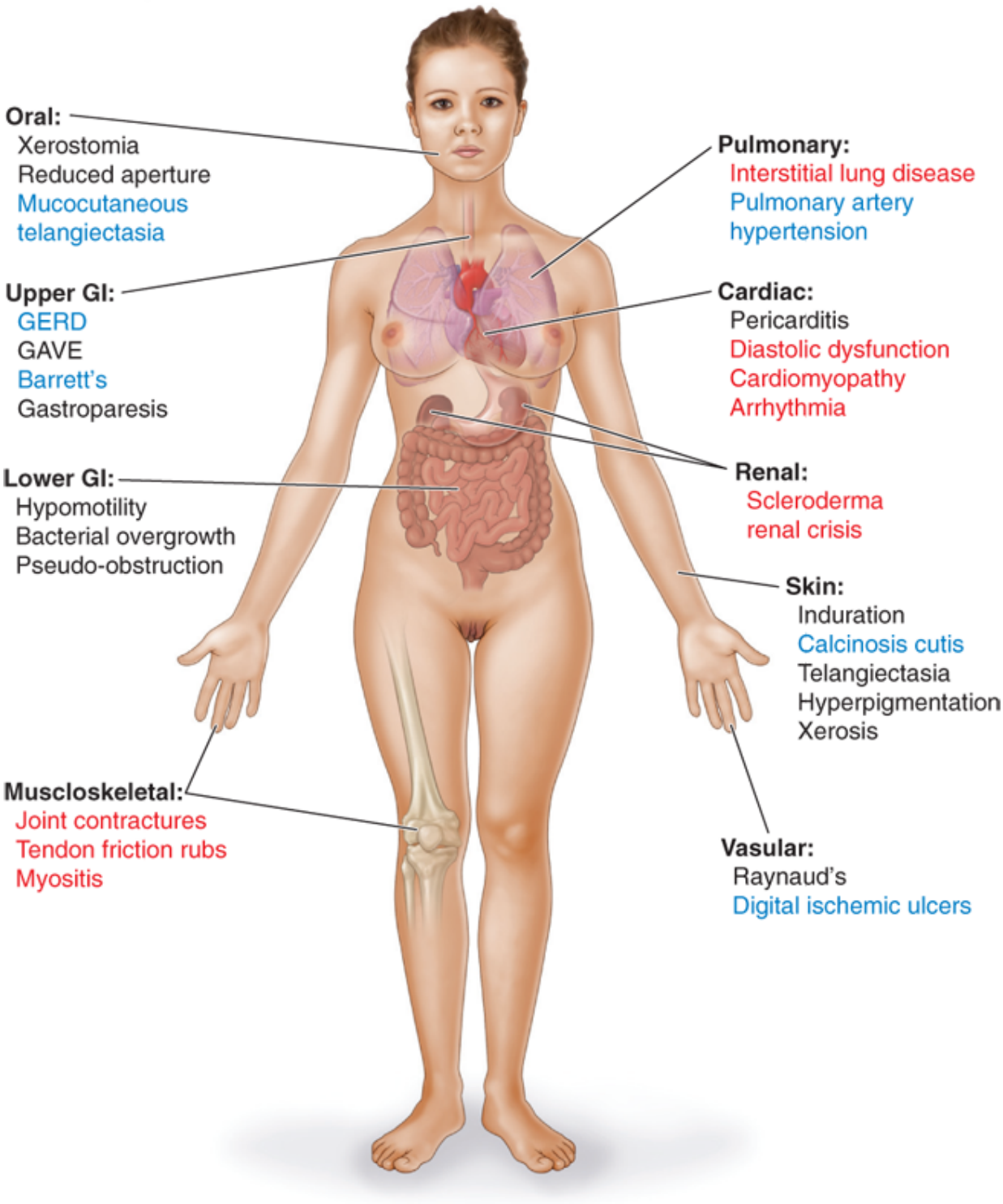
John Varga

DEFINITION AND CLASSIFICATION

Systemic sclerosis (SSc) is a complex and clinically heterogeneous orphan disease with protean clinical manifestations, a chronic and frequently progressive course, and significant disability, disfigurement and mortality. Virtually every organ can be affected (Fig. 353-1).

FIGURE 353-1

Multi-organ involvement in systemic sclerosis. Prominent complications more common in diffuse cutaneous SSc are shown in red; more common in limited cutaneous SSc in blue; and common in both forms of SSc shown in black.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

There is marked variability among SSc patients in patterns of skin involvement, organ complications, rates of disease progression, response to treatment, and survival. The early stages of SSc are associated with prominent inflammatory features; however, over time, structural alterations in multiple vascular beds and progressive visceral organ dysfunction due to fibrosis and atrophy come to dominate the clinical picture. Classification criteria for diagnosis of SSc are shown in Table 353-1.

TABLE 353-1

Classification Criteria for Diagnosis of Systemic Sclerosis

ITEM	SUB-ITEM	WEIGHT/SCORE
Skin thickening (bilateral)—fingers extending proximal to MCP joints		9
Skin thickening of fingers only	Puffy fingers Sclerodactyly (skin thickened distal to MCP joints)	2 4
Fingertip lesions	Digital tip ulcer or pitting scar	2 3
Mucocutaneous telangiectasia		2
Abnormal nails capillary pattern		2
Lung involvement	PAH Interstitial lung disease	2 2
Raynaud's phenomenon		3
SSc-specific autoantibodies	ACA Scl-70 RNA polymerase III	3

Abbreviations: ACA, anterior cerebral artery; MCP, metacarpophalangeal joint; PAH, pulmonary arterial hypertension.

Although thick and indurated skin (*scleroderma*) is the distinguishing hallmark of SSc, skin changes also occur in localized forms of scleroderma, along with multiple metabolic, inherited and autoimmune disorders (Table 353-2). Patients with SSc can be broadly segregated into two major subsets defined by the pattern of skin involvement, clinical and laboratory features, and natural history (Table 353-3). Diffuse cutaneous SSc (dcSSc) is typically associated with extensive skin induration starting in the fingers (sclerodactyly) and ascending from distal to proximal limbs and the trunk. In these patients, interstitial lung disease (ILD) and acute renal involvement develop relatively early. In contrast, in patients with limited cutaneous SSc (lcSSc), Raynaud's phenomenon generally precedes other disease manifestations, sometimes by years. In these patients, skin involvement remains confined to the fingers, distal limbs, and face, while the trunk is spared. The constellation of calcinosis cutis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia, was historically termed the *CREST syndrome*. In lcSSc, visceral organ involvement tends to show insidious progression, and digital ischemic ulcers, pulmonary arterial hypertension (PAH), hypothyroidism, and primary biliary cirrhosis may occur as late

complications. In some patients, Raynaud's phenomenon and characteristic clinical and laboratory features of SSc occur in the absence of detectable skin thickening. This syndrome has been termed *SSc sine scleroderma.*

TABLE 353-2

Conditions Associated with Skin Induration

Systemic sclerosis (SSc)
Limited cutaneous SSc
Diffuse cutaneous SSc
Localized scleroderma
Guttate (plaque) morphea, diffuse (pansclerotic) morphea, bullous morphea
Linear scleroderma, coup de sabre, hemifacial atrophy
Pansclerotic morphea
Overlap syndromes
Mixed connective tissue disease
SSc/polymyositis
Diabetic scleredema and scleredema of Buschke
Scleromyxedema (papular mucinosis)
Chronic graft-versus-host disease
Diffuse fasciitis with eosinophilia (Shulman's disease, eosinophilic fasciitis)
Stiff skin syndrome
Pachydermatoperiostosis (Primary hypertrophic osteoarthropathy)
Chemically induced and drug-associated scleroderma-like conditions
Vinyl chloride–induced disease
Eosinophilia-myalgia syndrome (associated with L-tryptophan contaminant exposure)
Nephrogenic systemic fibrosis (associated with gadolinium exposure)
Paraneoplastic syndrome

TABLE 353-3

Subsets of Systemic Sclerosis (SSc): Features of Limited Cutaneous SSc versus Diffuse Cutaneous Disease

CHARACTERISTIC FEATURE	LIMITED CUTANEOUS SSc	DIFFUSE CUTANEOUS SSc
Skin involvement	Indolent onset. Limited to fingers, distal to elbows, face; slow progression	Rapid onset. Diffuse: fingers, extremities, face, trunk; rapid progression
Raynaud's phenomenon	Antedates skin involvement, sometimes by years; may be associated with critical ischemia in the digits	Onset coincident with skin involvement; critical ischemia less common
Musculoskeletal	Mild arthralgia	Severe arthralgia, carpal tunnel syndrome, tendon friction rubs; small and large joint contractures
Interstitial lung disease	Slowly progressive, generally mild	Frequent, early onset and progression, can be severe
Pulmonary arterial hypertension	Frequent, late, may occur as an isolated complication	Often occurs in association with interstitial lung disease
Scleroderma renal crisis	Very rare	Occurs in 15%; onset may be fulminant; generally early (<4 years from disease onset)
Calcinosis cutis	Frequent, prominent	Less common, mild
Characteristic autoantibodies	Anti-centromere	Anti–topoisomerase I (Scl-70), anti-RNA polymerase III

INCIDENCE AND PREVALENCE

SSc is an acquired sporadic disease with a worldwide distribution and affecting all races. In the United States, the incidence is 9–46 cases per million per year. There are an estimated 100,000 U.S. cases, although this number may be significantly higher if patients who do not meet classification criteria are also included. There are large regional variations in incidence rates, potentially reflecting differences in case definition, environmental exposures or susceptibility genes in populations with different ancestries. Prevalence rates in England, Europe, and Japan appear to be lower than in North America and Australia. Age, sex, and ethnicity influence disease susceptibility, and blacks have higher age-specific incidence rates. In common with other connective tissue diseases, SSc shows a strong

female predominance (4.6:1), which is most pronounced in the childbearing years and declines after menopause. An additional risk factor is having an affected first-degree family member, which increases disease risk 13-fold. Although SSc can present at any age, the peak age of onset in women with both lcSSc and dcSSc is 65–74 years, although in blacks, disease onset occurs at an earlier age. Furthermore, blacks with SSc are more likely to have dcSSc, ILD, and a worse prognosis.

GENETIC CONTRIBUTION TO DISEASE PATHOGENESIS



SSc is a polygenic disease. In general, the genetic associations of SSc identified to date make only a small contribution to disease susceptibility. Disease concordance rates are low (4.7%) in monozygotic twins, although concordance for antinuclear antibody (ANA) positivity is significantly higher. On the other hand, evidence for genetic contribution to disease susceptibility is provided by the observation that 1.6% of SSc patients have a first-degree relative with SSc, a prevalence rate markedly increased compared to the general population. The risk of Raynaud's phenomenon, ILD, and other autoimmune diseases, including systemic lupus erythematosus (SLE) **(Chap. 349)**, rheumatoid arthritis **(Chap. 351)**, and autoimmune thyroiditis **(Chap. 375)**, is also increased in first-degree relatives. Current approaches to uncover genetic factors in SSc include DNA sequencing and single nucleotide polymorphism (SNP) analysis of candidate genes, and SNP analysis of the entire genome in a hypothesis-free manner. Genome-wide association studies (GWASs) involve large multi-center and multi-national cohorts. A majority of the robustly validated susceptibility loci for SSc are genes involved in innate and adaptive immune responses, highlighting the importance of autoimmunity as the initial trigger for the disease. Genetic studies have shown associations with common (small effect size) variants related to B and T lymphocyte activation and signaling (*BANK1*, *BLK*, *CD247*, *STAT4*, *IL2RA*, *CCR6, IDO1, TNFSF4/OX40L, PTPN22*, and *TNIP1*). In addition, candidate gene studies and GWASs identified a strong association with human leukocyte antigen (HLA)-Class II haplotypes on chromosome 6, including *HLA-DRB1*11:04, DQA1*05:01,* and *DQB1*03:01*, and the non-HLA genes histocompatibility complex (MHC) genes *NOTCH4* and *PSORSC1*. Other genetic variants associated with SSc are involved in innate immunity and the interferon pathways (*IRF5, IRF7, STAT4, TNFAIP3/A20, GSDMA, PRDM1 (BLIMP1)*, *TNFAIP3*, and *TLR2*). Additional associations with *IL12RB2, IL-21*, the apoptosis-related genes *DNASE1L3* and *SOX5,* and the fibrosis-related genes *CSK, CAV1, PPARG*, and *GRB10* have been reported. In addition to disease susceptibility, some of these genetic loci are associated with particular disease manifestations or serologic subsets, including ILD (*CTGF*, *CD226*), PAH (*TNIP1*), and scleroderma renal crisis (*HLA-DRB1**). While the functional consequences of these gene variants and their potential roles in pathogenesis are currently not well understood, it seems likely that in combination they cause a state of altered immune regulation, leading to increased susceptibility to autoimmunity and persistent inflammation. Of note, many of the genetic variants associated with SSc are also implicated in other autoimmune disorders, including SLE, rheumatoid arthritis, and psoriasis, suggesting common pathogenic pathways shared among these phenotypically dissimilar conditions. The genetic associations identified to date only explain a fraction of the heritability of SSc, and GWASs, and whole exome sequencing to identify additional genetic susceptibility factors in SSc, particularly rare (and potentially causal) variants, are currently ongoing.

ENVIRONMENTAL AND OCCUPATIONAL RISK FACTORS

Given the relatively modest genetic contribution to disease susceptibility in SSc, environmental factors, such as infectious agents, intestinal microbiota, and occupational, dietary, lifestyle, and drug exposures, are likely to play a

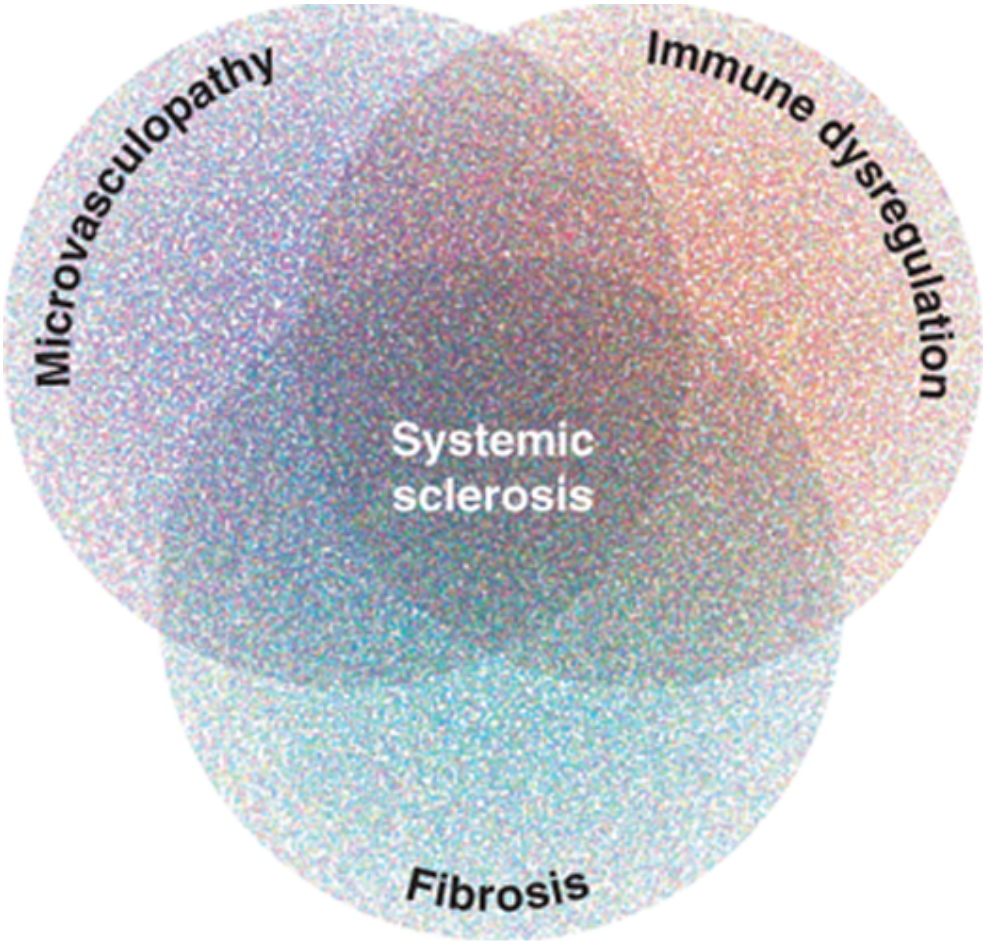
major role. Some evidence suggests potential roles for parvovirus B19, Epstein-Barr virus (EBV), cytomegalovirus (CMV), and *Rhodotorula glutinis* and other microorganisms. An epidemic of a novel syndrome with features suggestive of SSc occurred in Spain in the 1980s. The outbreak, termed *toxic oil syndrome*, was linked to use of contaminated rapeseed oil for cooking. Another epidemic outbreak, termed *eosinophilia-myalgia syndrome* (EMS), was linked to consumption of l-tryptophan-containing dietary supplements. Exposure to gadolinium contrast material in patients with renal compromise undergoing magnetic resonance scanning has been associated with nephrogenic systemic fibrosis. While each of these novel toxic-epidemic syndromes was characterized by chronic indurative skin changes and variable visceral organ involvement, the constellation of associated clinical, pathologic, and laboratory features distinguishes them from SSc. Occupational exposures tentatively linked with SSc include particulate silica (quartz), polyvinyl chloride, epoxy resins, welding fumes, and organic solvents and aromatic hydrocarbons including pain thinners, toluene, xylene, and trichloroethylene. These exposures might elicit stable and heritable epigenetic changes such as DNA methylation and histone modification underlying pathogenic alterations in gene expression. Drugs implicated in SSc-like illnesses include bleomycin, pentazocine, and cocaine, and appetite suppressants linked with PAH. Radiation therapy for cancer has been linked with *de novo* onset of SSc as well as with exacerbation of pre-existing SSc. In contrast to rheumatoid arthritis, cigarette smoking does not increase the risk of SSc. Although case reports and series of SSc in women with silicone breast implants had raised concern regarding a possible causal role of silicone in SSc, large-scale epidemiologic investigations found no evidence of increased prevalence of SSc.

PATHOGENESIS

Three cardinal pathomechanistic processes underlie the protean clinical manifestations of SSc: (1) diffuse microangiopathy, (2) inflammation and autoimmunity, and (3) visceral and vascular fibrosis in multiple organs **(Fig. 353-2)**. While all three processes are concurrently operative in SSc patients, their activity, relative severity, and contribution to the overall clinical picture vary among individual patients and over time. In general, autoimmunity and altered vascular reactivity occur early, while fibrosis and atrophy occur later in the disease. Complex and dynamic interplay among these processes initiates and sustains the fibrotic process and tissue damage.

FIGURE 353-2

The characteristic constellation of vasculopathy, autoimmunity/inflammation and fibrosis underlies the protean clinical manifestations of systemic sclerosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

ANIMAL MODELS OF DISEASE

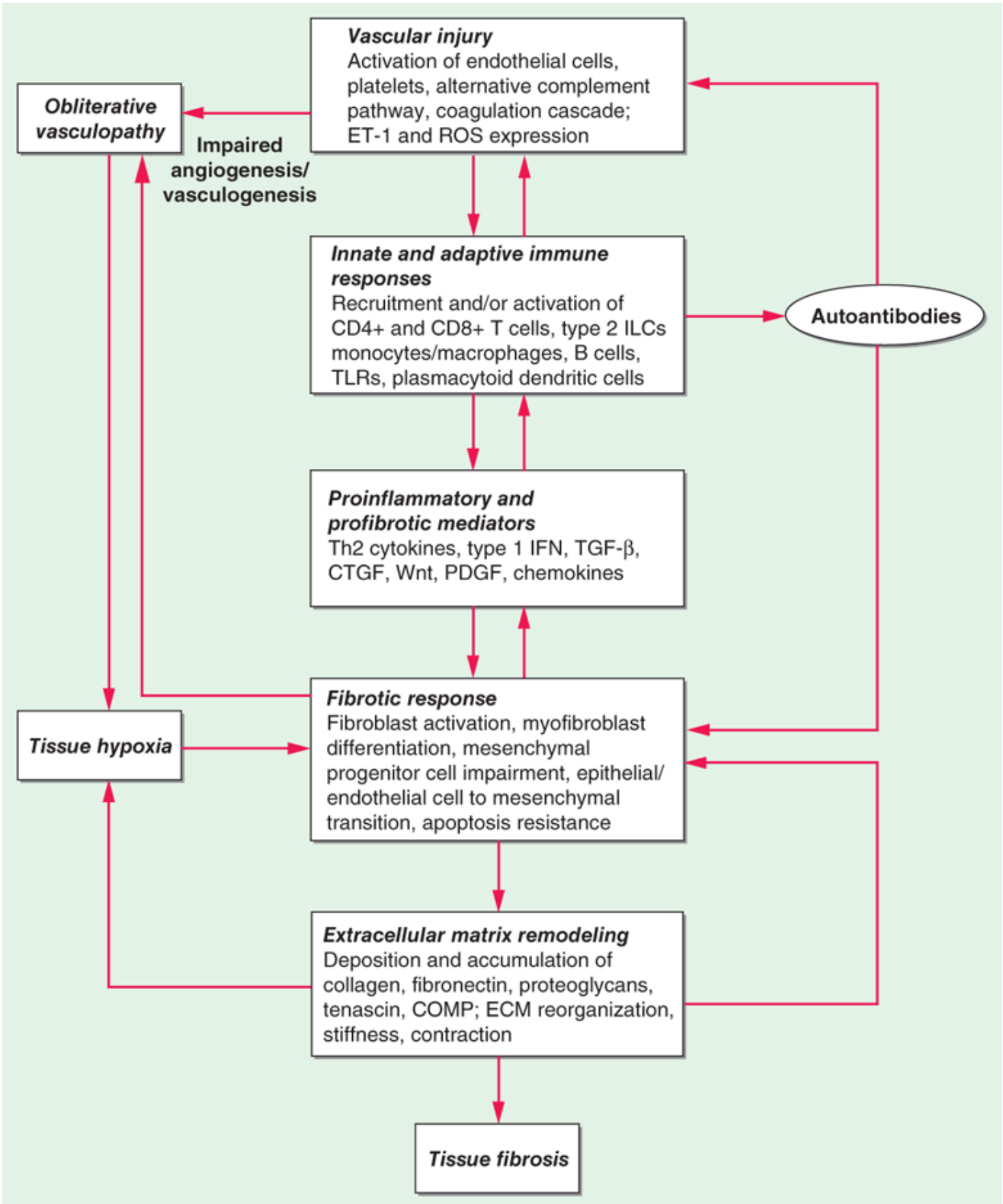
No single animal model of SSc fully reproduces the three cardinal processes that underlie pathogenesis, but some recapitulate selected aspects of the human disease. Tight-skin mice (Tsk1/+) spontaneously develop skin fibrosis due to a mutation in the fibrillin-1 gene. Mutant fibrillin-1 protein disrupts extracellular matrix assembly and causes aberrant activation of transforming growth factor β (TGF-β). Fibrillin-1 mutations in humans are associated with Marfan's disease and stiff skin syndrome, but have not been reported in SSc. Skin and lung fibrosis accompanied by variable vasculopathy and autoimmunity can be elicited in mice by injection of bleomycin or Angiotensin II, or by transplantation of HLA-mismatched bone marrow or spleen cells. Targeted genetic modifications in mice give rise to new disease models for investigating the pathogenetic roles of individual molecules, pathways, and cell types. For example, mice lacking IRF5, Smad3, uPAR, or peroxisome proliferator-activated receptor (PPAR)-γ, or constitutively overexpressing β-catenin, Wnt10b, sirtuin 3, Fra-2, TGFB1, PDGFRα, or adiponectin are either resistant or hypersensitive to experimental scleroderma, or spontaneously develop fibrosis. These disease models can contribute to understanding specific aspects of SSc pathogenesis, and to discovery and validation of novel targets for therapy.

MICROANGIOPATHY

Vascular injury is an early and possibly primary pathogenic event in SSc that leads to protean clinical manifestations of small vessel vasculopathy (**Fig. 353-3**).

FIGURE 353-3

Initial vascular injury in a genetically susceptible individual triggers functional and structural vascular alterations, inflammation and autoimmunity, culminating in fibrosis. Inflammatory and immune responses initiate and sustain fibroblast activation and differentiation, resulting in pathologic fibrogenesis and irreversible tissue damage. Vascular damage results in tissue ischemia that further contributes to progressive fibrosis and atrophy. COMP, cartilage oligomeric matrix protein; CTGF, connective tissue growth factor; PDGF, platelet-derived growth factor; ROS, reactive oxygen species; TGF-β, transforming growth factor β; TLR, toll-like receptor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Prominent microangiopathy in multiple vascular beds has important clinical sequelae including mucocutaneous telangiectasiae, Raynaud's phenomenon, ischemic digital ulcers, scleroderma renal crisis, myocardial involvement, and PAH. Raynaud's phenomenon is characterized by altered blood-flow response to cold challenge in small digital arteries. This initially reversible functional abnormality is associated with autonomic and peripheral nervous system alterations, including impaired production of the neuropeptide calcitonin gene-related peptide from sensory afferent nerves and heightened sensitivity of α_2-adrenergic receptors on vascular smooth-muscle cells. Isolated (primary) Raynaud's disease is common, generally benign and non-progressive. In contrast, secondary Raynaud's phenomenon in SSc is often progressive and complicated by irreversible structural changes, culminating in ischemic digital ulcers, necrosis, and amputation.

Viruses, cytotoxic factors, and chemokines thrombogenic microparticles, alternate complement pathway activation and autoantibodies targeting endothelial cells, phospholipids, and β2 glycoprotein I (β2GPI) are implicated as potential triggers of endothelial cell injury. Endothelial damage results in dysregulated production of vasodilatory (nitric oxide and prostacyclin) and vasoconstricting (endothelin-1) substances, as well as upregulation of intercellular adhesion molecule 1 (ICAM-1) and other surface adhesion molecules. Microvessels show enhanced permeability and transendothelial leukocyte diapedesis, abnormal activation of coagulation cascades, elevated thrombin production, and impaired fibrinolysis. Spontaneous platelet aggregation causes release of serotonin, platelet-derived growth factor (PDGF), and platelet alpha granules including thromboxane, a potent vasoconstrictor. Smooth-muscle cell–like myointimal cells in the media proliferate, the basement membrane is thickened and reduplicated, and perivascular adventitial fibrosis develops. The vasculopathic process affects capillaries, as well as arterioles, and less commonly even large vessels in many organs, resulting in reduced blood flow and tissue ischemia. Progressive luminal occlusion due to intimal and medial hypertrophy, combined with persistent endothelial cell damage and adventitial fibrosis, establish a vicious cycle that culminates in the striking absence of small blood vessels (rarefaction) in late-stage disease. Recurrent ischemia-reperfusion generates reactive oxygen species (ROS) that further damages the endothelium through peroxidation of membrane lipids. Paradoxically, the process of revascularization that normally reestablishes blood flow to ischemic tissue is defective in SSc despite elevated levels of other angiogenic factors. Moreover, bone marrow–derived circulating endothelial progenitor cells are reduced in number and impaired in function. Widespread capillary loss, obliterative vasculopathy of small and medium-sized arteries, and impaired ability to repair and replace damaged vessels are hallmarks of SSc.

INFLAMMATION AND AUTOIMMUNITY

Cellular Immunity

The following observations provide support for the inflammatory/autoimmune nature of SSc: near-universal presence of circulating autoantibodies with defined specificities; familial clustering of SSc with other autoimmune diseases; detection of activated immune cells, including T cells with oligoclonal antigen receptors, in target organs; prominent type I interferon (IFN) signatures, characterized by elevated expression of IFN-regulated genes, in a variety of cell types; elevated circulating levels and spontaneous secretion from mononuclear cells of cytokines and chemokines such as interleukin-6 (IL-6); tumor necrosis factor, IL-4, IL-10, IL-17, IL-33, CCL2, and CXCL4; genetic association of SSc with variants of MHC and other genes functionally implicated in the immune response; and the rapid clinical response, fibrosis resolution, and vascular regeneration observed in some SSc patients treated with immunomodulatory or immunoablative therapies. Genetic studies reveal strong associations with MHC locus

alleles, as well as non-HLA-linked genes encoding mediators of both adaptive and innate immune responses (*CD247, STAT4, IRF5, CD226*, *TNFAIP3/A20*, and *TNFSF4*).

Circulating monocytes from SSc patients overexpress IFN-regulated genes such as Siglec-1, have reduced levels of caveolin-1, and exhibit an inherently profibrotic phenotype. In early (edematous) stage SSc, mononuclear cell infiltrates comprised of activated T cells, monocytes/macrophages, and dendritic cells can be seen in skin, lungs, and other affected organs prior to appearance of fibrosis or vascular damage. Dendritic cells can be found in close proximity to activated fibroblasts and myofibroblasts and express toll-like receptors (TLR) and secrete IFN, IL-10, thymic stromal lymphopoietin (TSLP), and CXCL4, shaping the adaptive immune response and contributing to loss of immune tolerance. Tissue-infiltrating T cells express CD45 and HLA-DR activation markers and display restricted T cell receptor signatures indicative of oligoclonal expansion in response to recognition of as-yet unknown antigen. Of note, in patients diagnosed with SSc in close temporal association with cancer who are RNA polymerase III antibody-positive, the tumor may show mutations in RNApol3 autoantigen, which results in the generation of mutant-specific T cell immunity and cross-reactive antibodies. These findings support the premise that an abnormal antigen might act as initial trigger for the autoimmune response in SSc.

Circulating T cells in SSc express chemokine receptors and α_1 integrin, accounting for their enhanced binding to endothelium and to fibroblasts, while endothelial cells express ICAM-1 and other adhesion molecules that facilitate leukocyte diapedesis. Activated T cells show a T_H2-polarized immune response driven by dendritic cells. The T_H2 cytokines IL-4, IL-13, IL-33, and TSLP induce fibroblast activation, whereas the T_H1 cytokine interferon γ (IFN-γ) blocks cytokine-mediated fibroblast activation and exhibits anti-fibrotic properties. Evidence for altered Th17 and regulatory T cell (Treg) numbers and function in SSc has been reported. Type 2 innate lymphoid cells (iLCs), a recently discovered lymphoid cell population implicated in type 2 immunity and tissue remodeling, are also elevated in SSc skin biopsies. Alternately activated M2 macrophages, which produce TGF-β and promote angiogenesis and tissue remodeling, are increased in the skin in SSc. Although the frequency of regulatory T cells that enforce immune tolerance is elevated in the circulation and tissues, their immunosuppressive function appears to be defective. Some evidence implicates altered B cell homeostasis and function in SSc. Circulating B cells show elevated CD19 and co-stimulatory molecules CD80 and CD86, suggesting B cell chronic activation. Serum levels of a proliferation-inducing ligand (APRIL) and B cell activating factor (BAFF), members of the TNF superfamily with potent effects on B cell activation, are elevated in SSc, and associate with extent of skin and lung involvement. B cells secrete IL-6, TGF-β, and other profibrotic cytokines implicated in pathogenesis. Thus, B cell hyperactivity in SSc might directly contribute to the inflammatory and fibrotic processes, as well as generation of autoantibodies. Microarray analysis identifies a distinct subset of SSc skin biopsies with elevated expression of inflammation-related genes. Evidence of innate immune and TLR signaling, reflecting activation by type 1 IFN from plasmacytoid dendritic cells, is prominent in peripheral blood cells and target organs.

Humoral Autoimmunity

Circulating ANAs can be detected by indirect immunofluorescence in virtually all patients with SSc, even in early stages of disease. In addition, several SSc-specific autoantibodies with distinct patterns of immunofluorescence show strong associations with unique disease endophenotypes (**Table 353-4**). These antibodies are directed mostly against intracellular proteins associated with transcription, DNA repair, and RNA processing. Owing to their high specificity, mutual exclusivity and association with unique disease manifestations, SSc-associated autoantibodies

have substantial utility in clinical practice as diagnostic and prognostic markers, while their role in monitoring disease activity remains uncertain. Moreover, antibodies directed against fibrillin-1, matrix metalloproteinases, cell surface markers Angiotensin II receptor, endothelin-1 receptor, muscarinic 3 receptor, or the PDGF receptor, have been described in patients with SSc, although their clinical relevance is not yet established. These antibodies manifest functional receptor agonist activity and might have direct pathogenic roles.

TABLE 353-4

Major Systemic Sclerosis-Specific Autoantibodies and Principal Associated Features

TARGET ANTIGEN	SSc SUBSET	PROMINENT CHARACTERISTIC CLINICAL ASSOCIATION
DNA Topoisomerase I (Scl-70) Speckled pattern	dcSSc	Tendon friction rubs, digital ischemic ulcers, scleroderma, extensive skin involvement, early ILD, cardiac involvement, scleroderma renal crisis
Centromere proteins Discreet speckled (centromere) pattern	lcSSc	Digital ischemic ulcers, calcinosis cutis, isolated PAH; renal crisis rare
RNA polymerase III Speckled pattern	dcSSc	Rapidly progressive skin, tendon friction rubs, joint contractures, GAVE, renal crisis, contemporaneous cancers; digital ulcers rare
U3-RNP (fibrillarin) Nucleolar pattern	dc/lcSSc	PAH, ILD, scleroderma renal crisis, GI tract involvement, myositis
Th/T_0 Nucleolar pattern	lcSSc	ILD, PAH
PM/Scl Nucleolar pattern	lcSSc	Calcinosis cutis, ILD, myositis overlap
Ku Speckled pattern	Overlap	SLE, myositis overlap
U1-RNP Speckled pattern	MCTD	PAH, inflammatory arthritis, myositis overlap
U11/U12 RNP Speckled pattern	dc/lcSSc	ILD

Abbreviations: dcSSc, diffuse cutaneous SSc; GAVE, gastric antral vascular ectasia; ILD, interstitial lung disease; lcSSc, limited cutaneous SSc; MCTD, mixed connective tissue disease; PAH, pulmonary arterial hypertension; SLE, systemic lupus erythematosus.

A variety of mechanisms have been proposed to account for the generation of SSc-associated autoantibodies. Proteolytic cleavage, increased expression or altered subcellular localization of normal proteins, or their alterations due to mutation in the case of certain tumors, could lead to immune recognition as neoepitopes, resulting in the breaking of immune tolerance.

FIBROSIS

Fibrosis affecting multiple organs is a distinguishing feature of SSc. The process is characterized by replacement of normal tissue architecture with dense, rigid, avascular, and relatively acellular connective tissue. Fibrosis in SSc follows, and is a consequence of, inflammation, autoimmunity, and microvascular damage (Fig. 353-3). Fibroblasts are mesenchymal cells primarily responsible for the functional and structural integrity of connective tissue. Upon their activation by extracellular cues, fibroblasts proliferate, migrate, secrete collagens and other matrix molecules, growth factors, chemokines, and cytokines, and transdifferentiate into contractile myofibroblasts. Under normal conditions, these self-limited responses accomplish physiologic repair and regeneration of tissue. In contrast, when these responses become sustained and amplified, pathologic fibrosis results. Stimulatory signaling by endogenous TGF-β and paracrine fibrotic mediators including IL-6, IL-13, Wnt ligands, connective tissue growth factor (CTGF), PDGF, lysophosphatidic acid, endothelin-1, hypoxia, ROS, thrombin, and mechanical forces are responsible for sustained fibroblast activation underlying non-resolving fibrosis in SSc. Buildup of damage-associated endogenous ligands for TLR4 (EDA-fibronectin, high mobility group B1 (HMGB1) and Tenascin-C) and for TLR9 (mitochondrial DNA) within the fibrotic microenvironment further contributes to non-resolving fibrosis.

In addition to tissue-resident fibroblasts and transformed myofibroblasts, bone marrow–derived circulating mesenchymal progenitor cells also contribute to fibrosis. The factors that regulate the differentiation of mesenchymal progenitor cells and their trafficking from the circulation into lesional tissue are unknown. Epithelial and endothelial cells, mesenchymal progenitor cells, preadipocytes and tissue fibroblasts have all been proposed as sources of myofibroblasts in fibrosis. Although myofibroblasts are transiently found in normal wound healing, their persistence in fibrotic tissue, possibly due to resistance to apoptosis, contributes to scar formation.

Explanted SSc fibroblasts display an abnormally activated phenotype *ex vivo*, with variably increased rates of collagen production, spontaneous ROS generation, prominent stress fibers, and constitutive expression of alpha smooth-muscle actin. Persistence of the “scleroderma phenotype” during serial *ex vivo* passage of SSc fibroblasts may reflect autocrine TGF-β stimulatory loops, deregulated microRNA expressions, or stable acquired epigenetic modifications in these cells.

PATHOLOGY

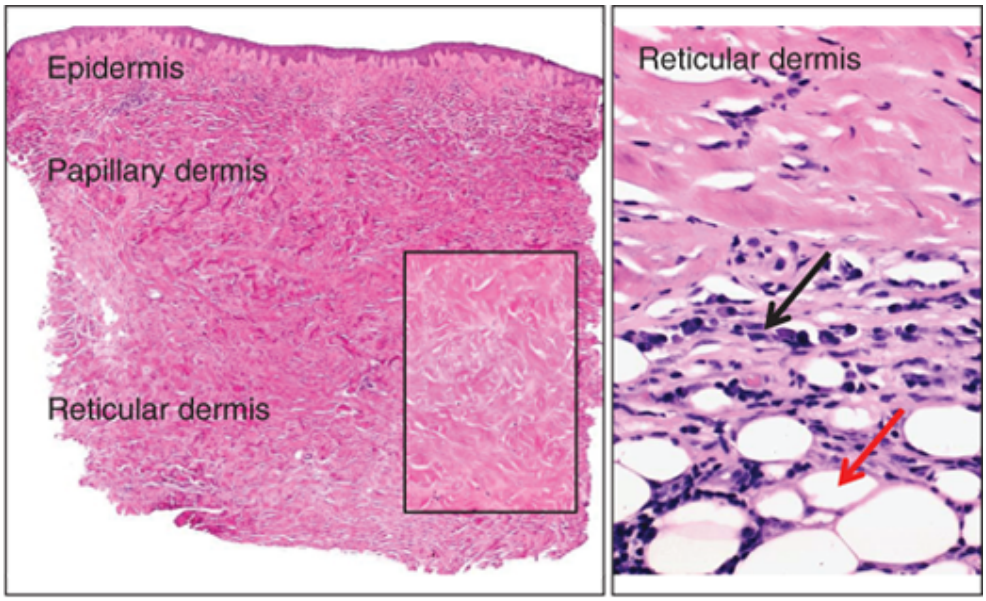
While pathological findings in SSc vary across anatomic sites, the distinguishing hallmark of SSc irrespective of the organ system is the triad of widespread capillary loss and obliterative microangiopathy, combined with fibrosis in the skin and internal organs. In early-stage disease, perivascular inflammatory cell infiltrates composed of T and B lymphocytes, activated monocytes and macrophages and mast cells may be detected in multiple organs. A non-inflammatory obliterative microangiopathy is a prominent late finding in the heart, lungs, kidneys, and gastrointestinal tract. Fibrosis is found in the skin, lungs, cardiovascular and gastrointestinal systems, tendon sheaths, perifascicular tissue surrounding skeletal muscle, and some endocrine organs. Excessive accumulation of collagens, proteoglycans, COMP and other structural matrix macromolecules progressively disrupts normal architecture, resulting in impaired function and failure of affected organs.

SKIN

The dermis is thickened, and accumulation of broad bundles of homogenized collagen oriented parallel to the epithelium is seen (**Fig. 353-4A**). Adnexal glands are atrophic, and loss of periadnexal and intradermal white adipose tissue and its replacement with collagen can be striking. While perivascular mononuclear cell infiltrates may be seen early, established skin fibrosis generally shows absence of inflammation. These findings are histologically indistinguishable from those in localized scleroderma.

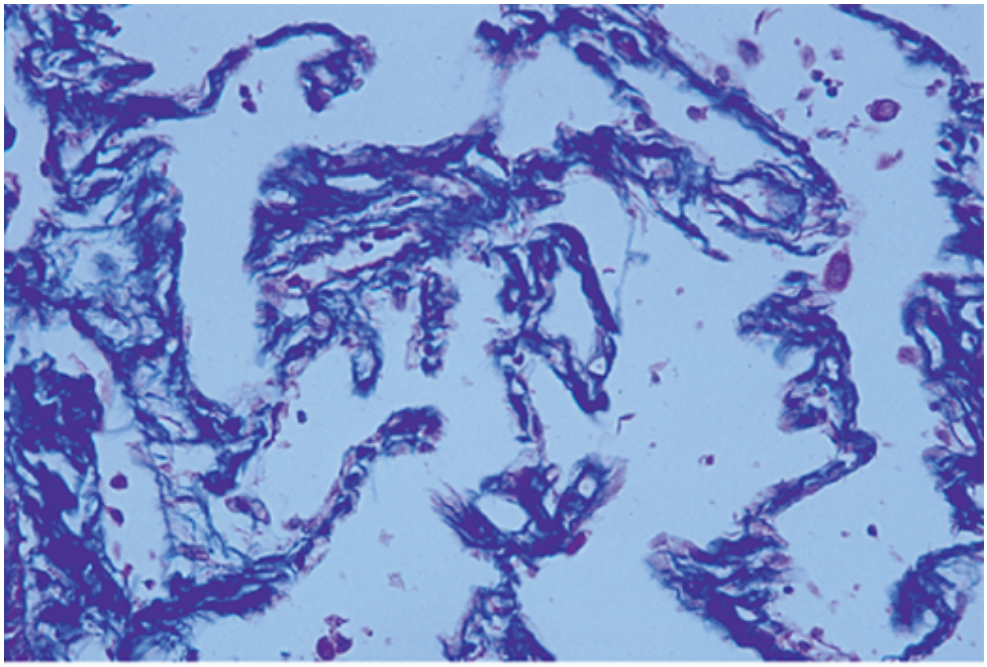
FIGURE 353-4

Pathologic findings in systemic sclerosis (SSc). ***A. Left panel:*** The skin is thickened due to expansion of the dermis. Inset, higher magnification showing thick hyalinized collagen bundles replacing skin appendages. ***Right panel:*** Mononuclear inflammatory cells within the intradermal adipose tissue. Black arrow, collagen; red arrow, dermal adipocytes. ***B.*** Early SSc-ILD. Diffuse fibrosis of the alveolar septae and a chronic inflammatory cell infiltrate. Trichrome stain. ***C.*** Pulmonary arterial obliterative vasculopathy. Striking intimal hyperplasia and luminal narrowing of small pulmonary artery, with little inflammation and minimal interstitial fibrosis, in a patient with SSc-PAH.



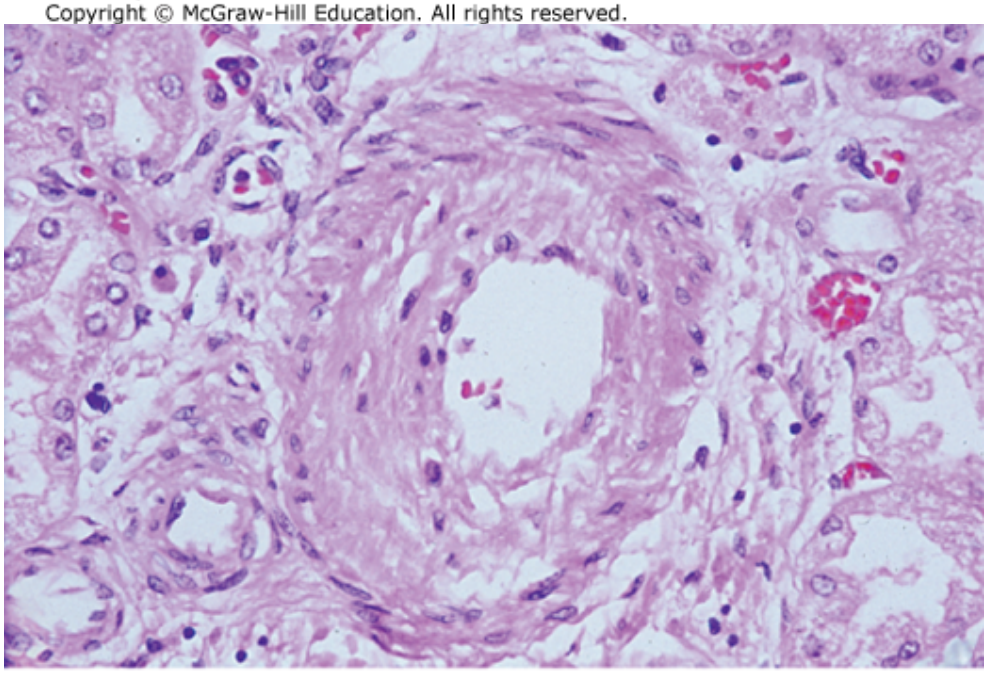
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

LUNGS

Autopsy studies in SSc universally show evidence of lung involvement. Most common is a nonspecific interstitial pneumonia (NSIP) pattern characterized by variable interstitial fibrosis and mild chronic inflammation. Patchy infiltration of the alveolar walls with T lymphocytes, macrophages, and eosinophils may occur in early disease. With progression, interstitial fibrosis and vascular damage dominate, often coexisting within the same biopsy. The usual interstitial pneumonia (UIP) pattern of spatial/temporal heterogeneity of inflammation, fibrosis and fibrotic foci

seen in idiopathic pulmonary fibrosis is less common in SSc (Fig. 353-4B). Fibrosis of the alveolar septae results in obliteration of the airspaces and loss of pulmonary blood vessels. This process impairs gas exchange and contributes to pulmonary hypertension. Intimal thickening of the pulmonary arteries, best seen with elastin stain, underlies SSc-associated PAH (Fig. 353-4C) and, at autopsy, is often associated with multiple pulmonary emboli and myocardial fibrosis. Patients may also show fibrosis and intimal proliferation in preseptal venules and veins in the lung, accounting for veno-occlusive disease. Lymphocytic bronchiolitis involving the submucosa of the terminal bronchioles may also be seen.

GASTROINTESTINAL TRACT

Pathologic changes can be found at any level from the mouth to the rectum. Atrophy and fibrosis of the muscularis propria and characteristic vascular lesions are prominent in the lower esophagus, while striated muscle in the upper third of the esophagus is generally spared. Collagenous replacement of the normal intestinal tract architecture results in impaired smooth muscle contractility and diminished peristaltic activity, with dysmotility, bacterial overgrowth, small-bowel obstruction, and perforation. Chronic gastroesophageal reflux is associated with esophageal inflammation, mucosal ulceration, and stricture formation and may lead to Barrett's metaplasia with attendant risk of adenocarcinoma. Esophageal dilation and reflux are associated with ILD due to chronic microaspiration.

KIDNEYS

In the kidneys, vascular lesions affecting the interlobular and arcuate arteries predominate. Chronic renal ischemia is associated with shrunken glomeruli. Patients with scleroderma renal crisis show acute fibrinoid necrosis of afferent arterioles, followed by intimal proliferation (onion-skin pattern), and ischemic collapse of glomeruli. These changes are reminiscent of thrombotic microangiopathies such as atypical hemolytic-uremic syndrome (see Chap. 304), and are accompanied by complement deposition, thrombosis, thrombocytopenia due to platelet consumption, and intravascular hemolysis. Extensive vascular thrombosis, glomerular collapse and sclerosis, and peritubular capillary deposits in renal biopsy are associated with irreversible renal failure.

HEART

Subclinical cardiac pathology is common, with prominent involvement of the myocardium and pericardium. The characteristic arteriolar lesions are concentric intimal hypertrophy and luminal narrowing, accompanied by patchy contraction band necrosis, loss of cardiac myocytes, and myocardial fibrosis due to microvascular involvement and ischemia-reperfusion injury. Fibrosis of the conduction system is common, especially at the sinoatrial node. The frequency of epicardial atherosclerotic coronary artery disease may be increased compared to the general population, similar to other systemic inflammatory diseases. Pericardial involvement with chronic inflammatory infiltrates and fibrinous exudates is common and may be associated with pericardial effusions.

PATHOLOGY IN OTHER ORGANS

Synovitis may be found in early SSc; with disease progression, the synovium becomes fibrotic, and in contrast to rheumatoid disease, pannus formation or bone resorption are uncommon. Fibrosis of tendon sheaths and fascia, sometimes accompanied by calcifications, produces palpable and sometimes audible tendon friction rubs.

Inflammation and, in later stages, atrophy and fibrosis of skeletal muscles are common findings, and are similar to those in polymyositis. Fibrosis of the thyroid gland and of the minor salivary glands may be seen. Placentas from SSc pregnancies show decidual vasculopathy, which is associated with poor perinatal outcomes and fetal death.

CLINICAL FEATURES

OVERVIEW

SSc can affect virtually any organ (Fig. 353-1 and **Table 353-5**). Although a dichotomous approach stratifying SSc into diffuse and limited cutaneous subsets (Table 353-2) is useful, disease expression is far more complex, and multiple distinct endophenotypes with unique patterns of manifestations can be recognized within each subset. Unique endophenotypes associate with autoantibodies with distinct and mutually exclusive specificities (Table 353-4). Patients with SSc "overlap" have typical features coexisting with clinical and laboratory evidence of another autoimmune disease, most commonly polymyositis, Sjögren's syndrome, polyarthritis, autoimmune liver disease, or SLE.

TABLE 353-5

Frequency of Clinical Organ Involvement in Limited Cutaneous and Diffuse Cutaneous Systemic Sclerosis (SSc)

FEATURES	LIMITED CUTANEOUS SSc (%)	DIFFUSE CUTANEOUS SSc (%)
Skin involvement	90[a]	100
Raynaud's phenomenon	99	98
Ischemic digital ulcers	50	25
Esophageal involvement	90	80
Interstitial lung disease	35	65
Pulmonary arterial hypertension	15	15
Myopathy	11	23
Clinical cardiac involvement	9	12
Scleroderma renal crisis	2	15
Calcinosis cutis	—	—

[a]Approximately 10% of patients have SSc s*ine* scleroderma.

INITIAL CLINICAL PRESENTATION

Characteristic initial presentation is quite different in patients with the diffuse (dcSSc) versus limited (lcSSc) cutaneous forms of the disease. In dcSSc, the interval between Raynaud's phenomenon and onset of other disease manifestations is brief (weeks to months). Soft tissue swelling, puffy fingers, and intense pruritus are signs of the early inflammatory "edematous" phase. The fingers, distal limbs, and face are usually affected first. Diffuse hyperpigmentation of the skin, carpal tunnel syndrome arthralgias, muscle weakness, fatigue, and decreased joint mobility are common. During the ensuing weeks to months, the inflammatory edematous phase evolves into the "fibrotic" phase, with skin induration associated with hair loss, reduced production of skin oils, and decline in sweating capacity. Progressive flexion contractures of the fingers ensue. The wrists, elbows, shoulders, hip girdles, knees, and ankles become stiff due to fibrosis of the supporting joint structures. While advancing skin involvement is the most visible manifestation of early dcSSc, important and clinically silent internal organ involvement commonly occurs during this stage. The initial 4 years from disease onset is the period of most rapidly evolving pulmonary and renal damage. If organ failure does not occur during this phase of dcSSc, the systemic process may stabilize.

Compared to dcSSc, the course of lcSSc tends to be more indolent. The interval between onset of Raynaud's phenomenon and disease manifestations such as GERD, cutaneous telangiectasia, or soft tissue calcifications can be as long as years. Scleroderma renal crisis, significant ILD, and tendon friction rubs occur rarely in lcSSc, while PAH, and overlap with keratoconjunctivitis sicca, polyarthritis, cutaneous vasculitis, and biliary cirrhosis can develop many years after disease onset.

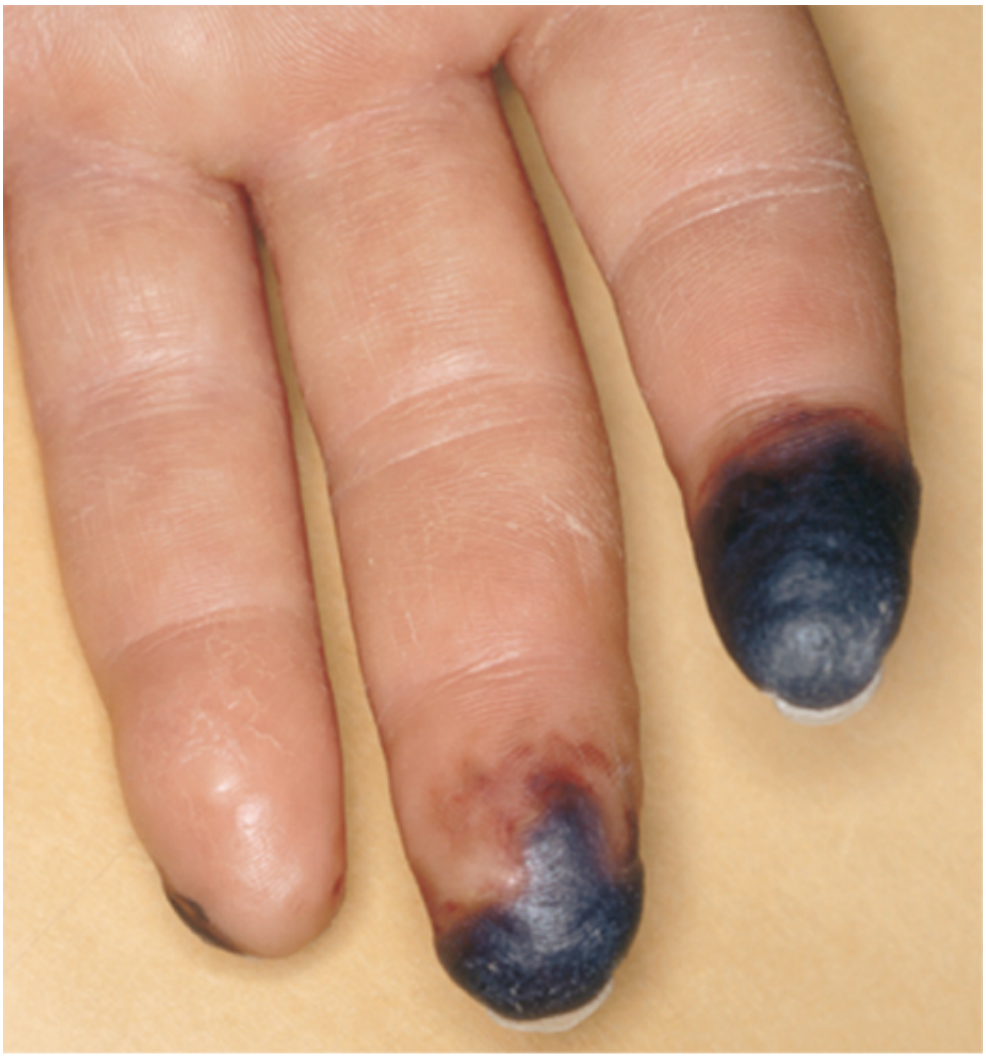
ORGAN INVOLVEMENT

RAYNAUD'S PHENOMENON

Raynaud's phenomenon, the most frequent extracutaneous complication of SSc, is characterized by episodes of reversible vasoconstriction in the fingers and toes, sometimes also affecting the tip of the nose and earlobes. Attacks, triggered by a decrease in temperature, as well as emotional stress and vibration, typically start with pallor, followed by cyanosis of variable duration. Hyperemia ensues spontaneously or with rewarming of the digit. The progression of the three color phases reflects the underlying vasoconstriction, ischemia, and reperfusion. Up to 5% of the general population has Raynaud's phenomenon. In the absence of signs or symptoms of an underlying condition, Raynaud's phenomenon is classified as primary (Raynaud's disease), which represents an exaggerated physiologic response to cold. Secondary Raynaud's phenomenon occurs in SSc and other connective tissue diseases, hematologic and endocrine conditions, and occupational disorders, and can complicate treatment with beta blockers and anticancer drugs such as cisplatin and bleomycin. Distinguishing primary Raynaud's disease from secondary Raynaud's phenomenon can present a diagnostic challenge. Raynaud's disease is supported by the following: absence of an underlying cause, a family history of Raynaud's phenomenon, absence of digital tissue necrosis or ulceration, and a negative ANA test. Secondary Raynaud's phenomenon tends to occur at an older age (>30 years), is more severe (episodes more frequent, prolonged, and painful), and is associated with ischemic digital ulcers and loss of digits (**Fig. 353-5**).

FIGURE 353-5

Digital necrosis. Sharply demarcated necrosis of the fingertip secondary to ischemia in a patient with limited cutaneous systemic sclerosis (SSc) associated with severe Raynaud's phenomenon.

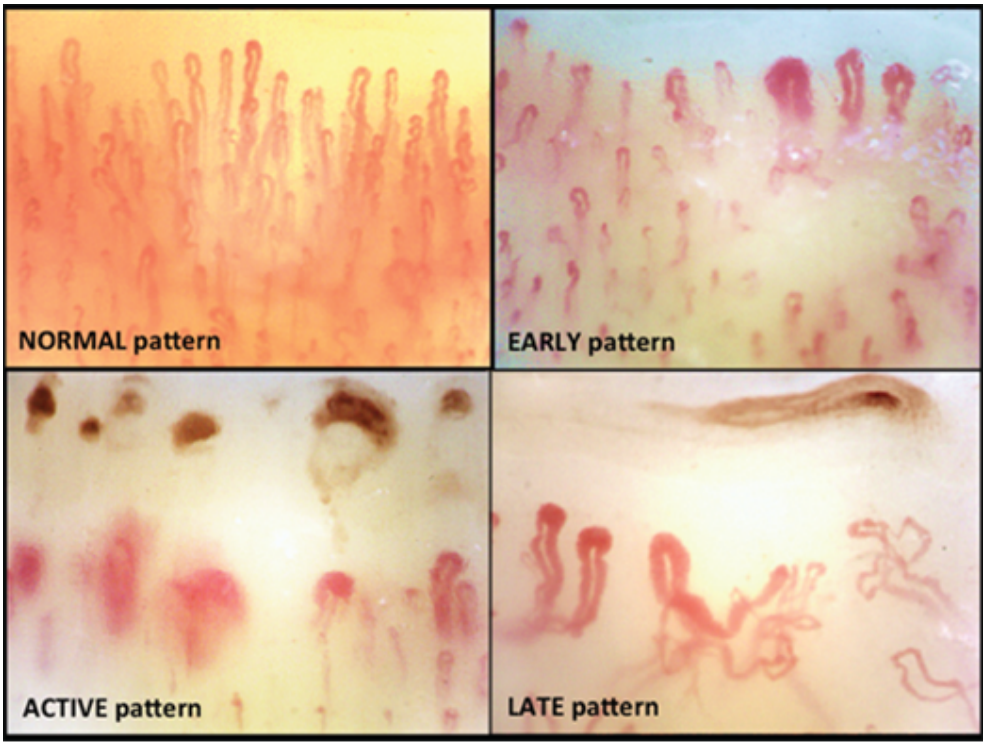


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Nailfold capillaroscopy using a low-power stereoscopic microscope or opthalmoscope permits visualization of nailbed cutaneous capillaries under immersion oil (**Fig. 353-6**). Raynaud's disease is associated with evenly spaced parallel vascular loops, whereas in secondary Raynaud's phenomenon, nailfold capillaries are distorted with widened and irregular loops, dilated lumen, microhemorrhages, and areas of vascular "dropout." Thus, nailfold capillaroscopy can be helpful in differentiating primary from secondary Raynaud's phenomenon and in establishing the early diagnosis of SSc.

FIGURE 353-6

SSc-associated nailfold capillary alterations. Normal nailfold pattern in healthy subjects. Note regularly-arrayed and uniform-size "hairpin" microvessels; "early pattern" showing dilations of microvessels and symmetrically increased microvessels (giant capillaries) representing the first morphological sign of systemic sclerosis; "active pattern" with giant capillaries, collapse with microhemorrages and loss of capillaries; "late pattern" showing massive loss of capillaries, fibrosis (white/yellow background) and neoangogenesis with secondary dilations (nailfold videocapillaroscope VIDEOCAP; magnification 220×). *(Courtesy of Professor Maurizio Cutolo, University of Genoa.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

SKIN FEATURES

Bilateral symmetrical skin thickening is the hallmark of SSc that distinguishes it from other connective tissue diseases. Skin involvement starts in the fingers and characteristically advances from distal to proximal extremities in an ascending fashion. Some patients note diffuse tanning in the absence of sun exposure as a very early manifestation. In dark-skinned individuals, vitiligo-like hypopigmentation may occur. Because pigment loss spares the perifollicular areas, the skin may have a “salt-and-pepper” appearance, most prominently on the scalp, upper back, and chest. Dermal sclerosis obliterating hair follicles, sweat glands, and eccrine and sebaceous glands cause hair loss, decreased sweating, and dry and itchy skin on the extremities. Transverse creases on the dorsum of the fingers disappear (Fig. 353-7). Fixed flexion contractures of the fingers cause reduced hand mobility and lead to muscle atrophy. Skin and subjacent tendon fibrosis accounts for fixed contractures of the wrists, elbows, and knees. Thick ridges at the neck due to firm adherence of skin to the underlying platysma muscle interfere with neck extension.

FIGURE 353-7

Sclerodactyly. Note skin induration on the fingers, and fixed flexion contractures of proximal interphalangeal joints, in a patient with limited cutaneous systemic sclerosis (lcSSc).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

In established SSc, the face assumes a characteristic "mauskopf" appearance with taut and shiny skin, loss of wrinkles, and occasionally an expressionless facies due to reduced mobility of the eyelids, cheeks, and mouth. Thinning of the lips with accentuation of the central incisor teeth and prominent perioral radial furrowing (rhytides) complete the picture. Reduced oral aperture (microstomia) interferes with eating and oral hygiene. The nose assumes a pinched, beak-like appearance. In late-stage disease, the skin becomes thin and atrophic, and is firmly bound to the subcutaneous fat (tethering). Dilated skin capillaries 2–20 mm in diameter (telangiectasiae), reminiscent of hereditary hemorrhagic telangiectasia, are frequently seen on the face, hands, lips, and oral mucosa (Fig. 353-8). The number of telangiectasias correlates with the severity of microvascular disease, including PAH. Breakdown of atrophic skin leads to chronic ulcerations at the extensor surfaces of the proximal interphalangeal joints, the volar pads of the fingertips, and bony prominences such as elbows and malleoli. Ulcers are often painful, heal slowly, and become secondarily infected, resulting in osteomyelitis. Healing of ischemic fingertip ulcerations leaves characteristic fixed digital "pits." Loss of soft tissue at the fingertips due to ischemia may be associated with striking resorption of the terminal phalanges (acro-osteolysis) (Fig. 353-9).

FIGURE 353-8

Cutaneous vascular changes. ***A.*** Vascular changes at the nailfold in lcSSc. ***B.*** Telangiectasia on the face.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 353-9

Acro-osteolysis. Note dissolution of distal terminal phalanges, commonly associated with ischemia, in a patient with long-standing limited cutaneous systemic sclerosis (lcSSc) and Raynaud's phenomenon.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Dystrophic calcifications in the skin, subcutaneous, and soft tissues (calcinosis cutis) in the presence of normal serum calcium and phosphate levels occur in up to 40% of patients, most commonly in those with long-standing anti-centromere antibody-positive lcSSc. Calcific deposits, composed of calcium hydroxyapatite crystals, vary in size from tiny punctate lesions to large conglomerate masses can be readily visualized on plain radiographs, or dual-energy CT. These deposits occur when calcium precipitates in tissue damaged by inflammation, hypoxia, or local trauma. Common locations include the finger pads, palms, extensor surfaces of the forearms, and the olecranon and prepatellar bursae (Fig. 353-10). They can cause pain and nerve compression, ulcerate through the overlying skin with drainage of chalky white material, and secondary infections. Paraspinal sheet calcifications may cause neurologic complications.

FIGURE 353-10

Calcinosis cutis. Note soft tissue calcific deposit breaking through the skin in a patient with limited cutaneous systemic sclerosis (lcSSc).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

PULMONARY FEATURES

The two principal forms of lung involvement in SSc, ILD, and pulmonary vascular disease are frequent and account for a majority of SSc-related deaths. Survival is particularly poor in SSc patients with concurrent presence of these two processes. Less common pulmonary complications of SSc include aspiration pneumonitis complicating chronic gastroesophageal reflux, pulmonary hemorrhage due to endobronchial telangiectasia, obliterative bronchiolitis, pleural reactions, restrictive physiology due to chest wall fibrosis, spontaneous pneumothorax, and drug-induced lung toxicity. The incidence of lung cancer is increased in SSc.

Interstitial Lung Disease

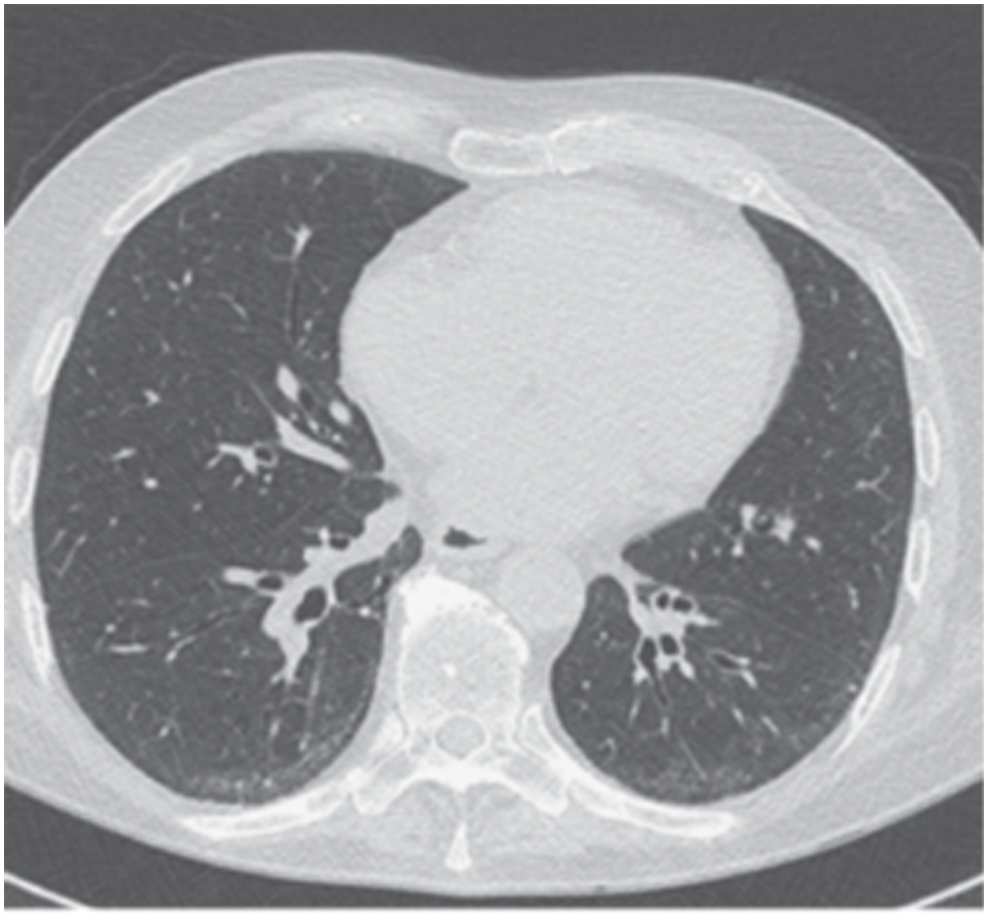
While evidence of ILD can be found in up to 65% of SSc patients by high-resolution computed tomography (HRCT), clinically significant ILD develops in 16–43%; the frequency varies depending on the detection method used. Risk factors for ILD include male sex, African-American race, diffuse skin involvement, severe gastroesophageal reflux, and the presence of topoisomerase I autoantibodies; in contrast, anti-centromere antibody-positive patients have a reduced risk of ILD. Additional risk factors include low forced vital capacity (FVC) or single-breath diffusing capacity of the lung for carbon monoxide (DLco) at initial presentation. Esophageal dilation with chronic acid reflux in SSc cause micro-aspiration, a risk factor for the development and progression of ILD. The most rapid progression in ILD generally occurs early in the disease course (within the first 3 years), when the FVC can decline by 30% per year.

Pulmonary involvement can remain asymptomatic until it is advanced. The most common presenting respiratory symptoms—exertional dyspnea, fatigue, and reduced exercise tolerance—are subtle and slowly progressive. A chronic dry cough may be present. Physical examination may reveal fine inspiratory “Velcro” crackles at the lung bases. Pulmonary function testing (PFT) is relatively sensitive for detecting early pulmonary involvement, and typically shows a restrictive ventilatory defect (FV<70% predicted and/or FEV1/FVC ratio >0.8), reduced total lung capacity (TLC) and diffusing capacity (DLco). A reduction in DLco that is significantly out of proportion to the reduction in lung volumes should raise suspicion for pulmonary vascular disease, but may also be due to anemia. Oxygen desaturation with exercise is common.

Chest radiography can be used as an initial screening tool to rule out infection and other causes of pulmonary involvement; however, compared to HRCT, it is relatively insensitive for detection of early ILD. It may demonstrate lower lobe subpleural reticular linear opacities and ground-glass opacifications, even in asymptomatic patients with normal PFTs **(Fig. 353-11)**. Additional HRCT findings include mediastinal lymphadenopathy, pulmonary nodules, traction bronchiectasis, and uncommonly, honeycomb changes. The extent of interstitial changes on chest HRCT is a predictor of ILD progression and mortality. Bronchoalveolar lavage (BAL) can demonstrate inflammatory cells in the lower respiratory tract, and may be useful for ruling out tuberculosis and other infections. However, BAL does not appear to be useful for SSc diagnosis or for identifying reversible alveolitis, and is used primarily for research. Lung biopsy is indicated only in patients with atypical findings on chest radiographs. The histologic pattern on lung biopsy may predict the risk of progression of ILD, with NSIP, carrying a better prognosis than UIP.

FIGURE 353-11

Chest CT in systemic sclerosis. ***Top panel:*** Early interstitial lung disease with subpleural reticulations and ground glass opacities in the lower lobes. Patient in supine position. ***Bottom panel:*** Extensive lung fibrosis with coarse reticular honeycombing, and traction bronchiectasis. Note dilated esophagus. *(Courtesy of Rishi Agrawal, Northwestern University.)*





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Pulmonary Arterial Hypertension PAH resulting from vascular remodeling of small (<500 µm) pulmonary arteries develops in 8–12% of patients with SSc, and occurs as an isolated abnormality or in association with ILD. PAH is defined hemodynamically as a mean pulmonary artery pressure ≥25 mmHg with a pulmonary capillary wedge pressure ≤15 mmHg and pulmonary vascular resistance >3 Wood units. The natural history of SSc-associated PAH is variable, but often follows a downhill course with onset of right heart failure. The 3-year survival of SSc patients with untreated PAH is <50%. Risk factors include lcSSc, high numbers of cutaneous telangiectasia, older age at disease onset, and the presence of antibodies to centromere, U1-RNP, U3-RNP (fibrillarin), and B23. Mutations in the BMPR2 gene associated with idiopathic PAH are not found in patients with SSc-PAH.

Although patients with PAH are often asymptomatic in early stages, they may present with nonspecific symptoms of exertional dyspnea and reduced exercise capacity. With progression, angina, near-syncope, and symptoms and signs of right-sided heart failure appear. Physical examination may show tachypnea, a loud pulmonic component of the S_2 heart sound, pulmonic/tricuspid regurgitation murmur, palpable right ventricular heave, elevated jugular venous pressure, and dependent edema. Doppler echocardiography provides a noninvasive screening method for estimating the pulmonary arterial pressure. In light of the poor prognosis of untreated PAH and better therapeutic response in patients with early diagnosis, all SSc patients should be screened for PAH at initial evaluation, followed by annual evaluation. Estimated pulmonary artery systolic pressure >40 mmHg at rest or tricuspid regurgitation jet velocities >3 m/sec suggest PAH. PFT may show a reduced DLco in isolation or out of proportion with the severity of restriction. Because echocardiography can over- or underestimate pulmonary artery pressures, cardiac catheterization is the gold standard required to confirm the diagnosis of suspected PAH, to assess its severity, including the degree of right heart dysfunction, to rule out veno-occlusive disease and other cardiac (post-capillary) causes of pulmonary hypertension, and to provide prognostic parameters. Yearly echocardiographic screening for PAH is recommended in most patients; an isolated decline in DLco may also be indicative of developing PAH. Distinguishing PAH from pulmonary hypertension secondary to pulmonary fibrosis and hypoxia in SSc can be difficult. Serum levels of N-terminal pro-brain natriuretic peptide (NT proBNP) correlate with the presence and severity of PAH in SSc, as well as survival. While NT proBNP measurements can be useful in screening for PAH and in monitoring the response to treatment, elevated levels are not specific for PAH and also occur in other forms of right and left heart disease. Despite more favorable hemodynamics, the prognosis of SSc-associated PAH is worse, and treatment response poorer, than that of idiopathic PAH, most likely due to frequent concurrence of ILD and cardiac complications in these patients.

GASTROINTESTINAL INVOLVEMENT

Involvement of the gastrointestinal tract, which can affect any level, occurs in up to 90% of SSc patients with both lcSSc and dcSSc disease (**Table 353-6**). The pathologic findings of fibrosis, smooth muscle atrophy, and obliterative small-vessel vasculopathy are similar throughout the length of the gastrointestinal tract, and contribute to reduced quality of life, malnutrition, and increased mortality.

TABLE 353-6

Prominent Gastrointestinal Manifestations of SSc and Their Management

SITE	PRINCIPAL MANIFESTATION	MANAGEMENT
Oropharynx	Diminished oral aperture Dry mouth Periodontitis, gingivitis swallowing	Periodontal care Artificial saliva Swallowing therapy
Esophagus	Reflux Dysphagia Strictures Barret's metaplasia	Lifestyle modifications Prokinetic drugs proton pump inhibitors Endoscopic procedures
Stomach	Gastroparesis Gastric antral vascular ectasia (GAVE, watermelon stomach)	Prokinetic agents Endoscopic laser cryotherapy
Small and large intestines	Bacterial overgrowth Diarrhea/constipation Pseudo-obstruction Pneumatosis intestinalis Malabsorption Colonic pseudodiverticula	Laxatives Prokinetic agents Rotating antibiotics Octreotide Parenteral nutritional support
Anorectum	Sphincter incompetence	Biofeedback, sacral nerve stimulation, surgery

Upper Gastrointestinal Tract Involvement

Decreased oral aperture interferes with regular dental hygiene. Teeth are loosened due to loss of periodontal ligament attaching teeth to the alveolar bone. Additional oropharyngeal manifestations due to a combination of xerostomia, shortened frenulum, and resorption of the mandibular condyles are frequent and cause much distress. Most patients have symptoms of gastroesophageal reflux disease (GERD): heartburn, regurgitation, and dysphagia. A combination of reduced lower esophageal sphincter pressure resulting in reflux, impaired esophageal clearance of refluxed gastric contents due to diminished motility, and delayed gastric emptying accounts for GERD. Calcium channel antagonists and phosphodiesterase inhibitors used to treat Raynaud's phenomenon can further aggravate reflux. Esophageal manometry shows abnormal motility in most patients, even in the absence of symptoms. Extra-esophageal manifestations of GERD include hoarseness, chronic cough, and microaspiration, which can result in infections and may aggravate underlying ILD. Chest CT characteristically shows a dilated patulous esophagus with

intraluminal air. Endoscopy may be necessary to rule out opportunistic infections with *Candida*, herpes virus, and CMV. Severe erosive esophagitis may be found on endoscopy in patients with minimal symptoms. Esophageal strictures and Barrett's esophagus may complicate chronic GERD. Because Barrett's metaplasia is associated with increased risk of adenocarcinoma, SSc patients with Barrett's require regular surveillance endoscopy with biopsy.

Gastroparesis with early satiety, abdominal distention, and aggravated reflux symptoms are common. Barium contrast studies are neither sensitive nor specific for evaluation of gastric involvement in SSc. Gastric antral vascular ectasia (GAVE) in the antrum may occur. These subepithelial lesions, reflecting the diffuse small-vessel vasculopathy of SSc, are described as "watermelon stomach" due to their endoscopic appearance. Patients with GAVE can have recurrent episodes of gastrointestinal bleeding, resulting in chronic unexplained anemia.

Lower Gastrointestinal Tract and Anorectal Involvement

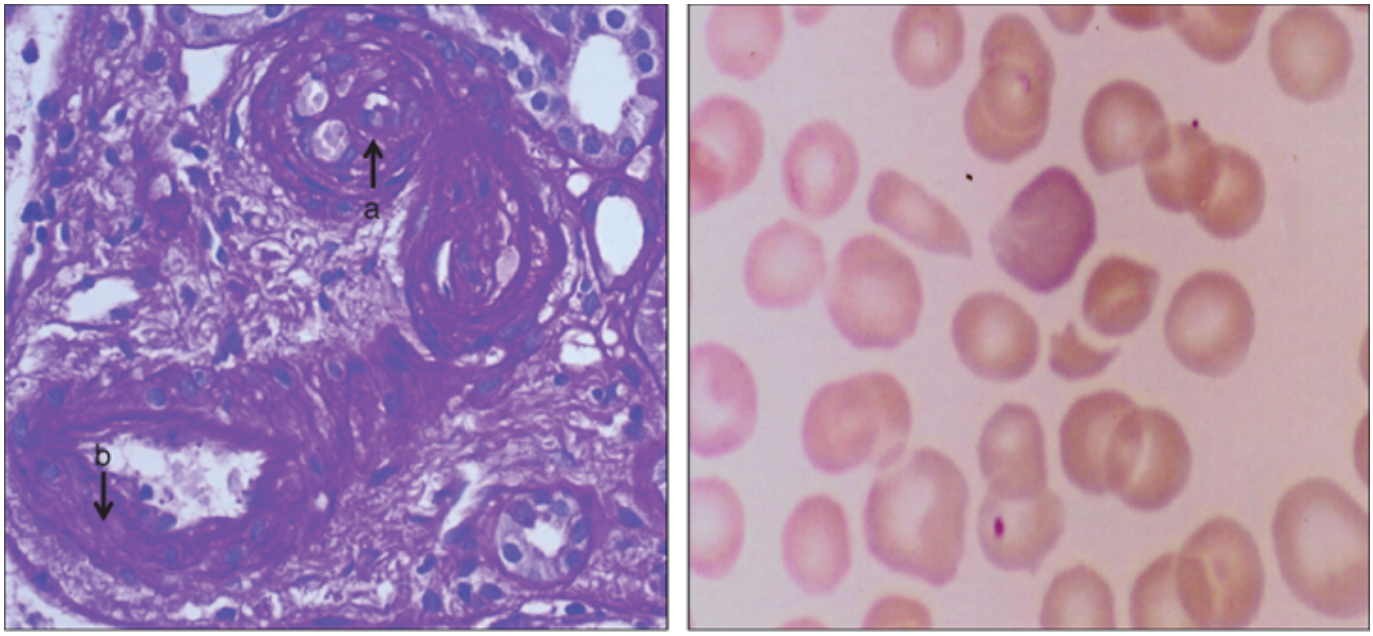
Weight loss and malnutrition due to impaired intestinal motility, malabsorption, and chronic diarrhea secondary to bacterial overgrowth are common. Fat and protein malabsorption and vitamin B_{12} and vitamin D deficiencies ensue, and may be further exacerbated by pancreatic insufficiency. Disturbed intestinal motor function can also lead to intestinal pseudo-obstruction, with symptoms that are indistinguishable from those of delayed gastric emptying. Patients present with recurrent episodes of acute abdominal pain, nausea, and vomiting, and radiographic studies show acute intestinal obstruction. A major diagnostic challenge is differentiating pseudo-obstruction, which responds to supportive care and intravenous nutritional supplementation, from mechanical obstruction. Colonic involvement may result in severe constipation, occasionally complicated by sigmoid volvulus. Fecal incontinence, gastrointestinal bleeding from telangiectasia, and rectal prolapse, can occur. In late-stage SSc, wide-mouth sacculations or diverticula occur in the colon, occasionally causing perforation and bleeding. An occasional radiologic finding is pneumatosis cystoides intestinalis due to air trapping in the bowel wall that may rarely rupture and cause benign pneumoperitoneum. Although the liver is rarely affected, primary biliary cirrhosis may coexist with SSc.

RENAL INVOLVEMENT: SCLERODERMA RENAL CRISIS

Scleroderma renal crisis presents with accelerated hypertension accompanied by acute kidney injury and progressive failure. This acute life-threatening complication of SSc occurs in 10–15% of patients, generally within 4 years of disease onset. Rarely, scleroderma renal crisis can be the initial presenting manifestation of SSc. Prior to the advent of angiotensin-converting enzyme (ACE) inhibitors, short-term survival in scleroderma renal crisis was <10%. The pathogenesis involves obliterative vasculopathy and luminal narrowing of the renal arcuate and interlobular arteries, with consequent intravascular hemolysis, along with evidence of activation of the complement pathways (**Fig. 353-12**). Progressive reduction in renal blood flow, aggravated by vasospasm, leads to juxtaglomerular renin secretion and activation of Angiotensin II, with further renal vasoconstriction resulting in a vicious cycle that culminates in accelerated hypertension. Risk factors for scleroderma renal crisis include African-American race, male sex, and diffuse or progressive skin involvement. Up to 50% of patients with scleroderma renal crisis have anti-RNA polymerase III antibodies, whereas patients with anti-centromere antibodies appear to be protected from this complication. Palpable tendon friction rubs, pericardial effusion, new unexplained anemia, and thrombocytopenia may be harbingers of impending scleroderma renal crisis. High-risk patients with early SSc should monitor their blood pressure daily. Because glucocorticoid use is associated with scleroderma renal crisis, prednisone in high-risk SSc patients should be taken only when absolutely required and at low doses (<10 mg/d).

FIGURE 353-12

Renal changes in scleroderma renal crisis. ***A.*** Renal biopsy demonstrating intimal proliferation and myxoid changes in medium-sized renal arteries. ***B.*** Film demonstrating fragmentation of red blood cells due to intravascular hemolysis in scleroderma renal crisis. (*Courtesy of Drs. Edward Stern and Christopher Denton, Royal Free Hospital, London, UK.*)



A ***B***

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

Patients characteristically present with accelerated hypertension (generally >150/90 mmHg) and progressive oliguric renal insufficiency. However, ~10% of patients with scleroderma renal crisis present with normal blood pressure. Normotensive renal crisis is generally associated with a poor outcome. Headache, blurred vision, congestive heart failure, and pulmonary edema may accompany elevation of blood pressure. Urinalysis typically shows mild proteinuria, granular casts, and microscopic hematuria; moderate thrombocytopenia and microangiopathic hemolysis with fragmented red blood cells can be seen. Progressive oliguric renal failure over several days generally follows. Scleroderma renal crisis is occasionally misdiagnosed as thrombotic thrombocytopenic purpura (TTP) or other forms of thrombotic microangiopathy. In such cases, renal biopsy and measuring vWF-cleaving protease activity may be of some benefit. Oliguria or a creatinine >3 mg/dL at presentation predicts poor outcome (permanent hemodialysis and mortality), as do biopsy findings of vascular thrombosis and glomerular ischemic collapse. Rarely, crescentic glomerulonephritis occurs in the setting of SSc and may be associated with myeloperoxidase-specific antineutrophil cytoplasmic antibodies. Membranous glomerulonephritis may occur in patients treated with d-penicillamine. Asymptomatic renal function impairment occurs in up to half of SSc patients. Such subclinical renal involvement is associated with other vascular manifestations of SSc and rarely progresses.

CARDIAC INVOLVEMENT

Although it is often silent, variable cardiac involvement in SSc is detected in 10–50% of patients screened with sensitive diagnostic tools. Clinical cardiac involvement, more frequent in dcSSc than in lcSSc, may be primary or secondary to PAH, ILD, or renal involvement, and is associated with poor outcomes. The endocardium,

myocardium, and pericardium may each be affected separately or together. Pericardial involvement is manifested as pericarditis, pericardial effusions, constrictive pericarditis, and rarely, cardiac tamponade. Conduction system fibrosis occurs commonly and may be silent or manifested by heart block. Arrythmias including premature ventricular contractions, atrial fibrillation, and supraventricular tachycardia are common. Microvascular involvement, recurrent vasospasm, and ischemia-reperfusion injury contribute to patchy myocardial fibrosis, resulting in asymptomatic systolic or diastolic left ventricular dysfunction that may progress to overt heart failure. Acute or subacute myocarditis leading to left ventricular dysfunction may occur, and diagnosis requires cardiac magnetic resonance imaging (MRI) or endomyocardial biopsy. While conventional echocardiography has low sensitivity for detecting preclinical heart involvement in SSc, newer modalities such as tissue Doppler echocardiography (TDE), cMRI, and nuclear imaging (single photon emission CT [SPECT]) reveal a high prevalence of abnormal myocardial function or perfusion. The serum levels of N-terminal pro-BNP, a ventricular hormone elevated in SSc-PAH, may also have utility as markers of primary cardiac involvement.

Musculoskeletal Complications

Musculoskeletal complications are very common in SSc. Carpal tunnel syndrome may be a presenting disease manifestation. Generalized arthralgia and stiffness are prominent in early disease. Mobility of both small and large joints is progressively impaired, and fixed contractures develop at the proximal interphalangeal joints and wrists. Large joint contractures, seen in patients with dcSSc, are frequently accompanied by tendon friction rubs characterized by coarse leathery crepitation heard or palpated upon passive joint movement, that are due to extensive fibrosis and adhesion of the tendon sheaths and fascial planes at the affected joint. Tendon friction rubs are associated with increased risk for renal and cardiac complications and reduced survival. Synovitis detected by ultrasound or MRI is common; occasional SSc patients develop erosive polyarthritis in the hands, and some have a seropositive rheumatoid arthritis overlap. Muscle weakness is common and multifactorial: deconditioning, disuse atrophy, malnutrition, inflammation, and fibrosis may all contribute. A chronic non-inflammatory myopathy characterized by atrophy and fibrosis with mildly elevated muscle enzymes can be seen in late-stage SSc. Bone resorption in the terminal phalanges causes loss of the distal tufts (acro-osteolysis) (Fig. 353-9). Resorption of the mandibular condyles can lead to bite difficulties. Osteolysis can also affect the ribs and distal clavicles.

LESS RECONGIZED DISEASE MANIFESTATIONS

Dry eyes and dry mouth (sicca complex) are common in SSc. Biopsy of the minor salivary glands shows fibrosis rather than focal lymphocytic infiltration characteristic of primary Sjögren's syndrome (**Chap. 354**). Hypothyroidism resulting from Graves' or Hashimoto's disease is common, particularly in lcSSc, and may be under-recognized. Whereas the central nervous system is generally spared, unilateral or bilateral sensory trigeminal neuropathy can occur. Erectile dysfunction is a frequent, and occasionally initial, disease manifestation. Inability to attain or maintain penile erection is due to vascular insufficiency and fibrosis of corporeal smooth muscle. Sexual performance is also adversely affected in women. While fertility is not impaired in SSc, pregnancy is associated with higher risk of adverse fetal outcomes. Furthermore, cardiopulmonary involvement may worsen during pregnancy, and new onset of scleroderma renal crisis has been described.

Cancer

Epidemiologic studies indicate an increased cancer risk in SSc. Lung cancer and esophageal adenocarcinoma typically occur in the setting of long-standing ILD or GERD and may be caused by chronic inflammation and repair. In contrast, breast, lung, and ovarian carcinomas and lymphomas tend to occur in close temporal association with the onset of SSc, particularly in patients who have autoantibodies to RNA polymerase III. In this scenario, SSc may represent a paraneoplastic syndrome triggered by the anti-tumor immune response.

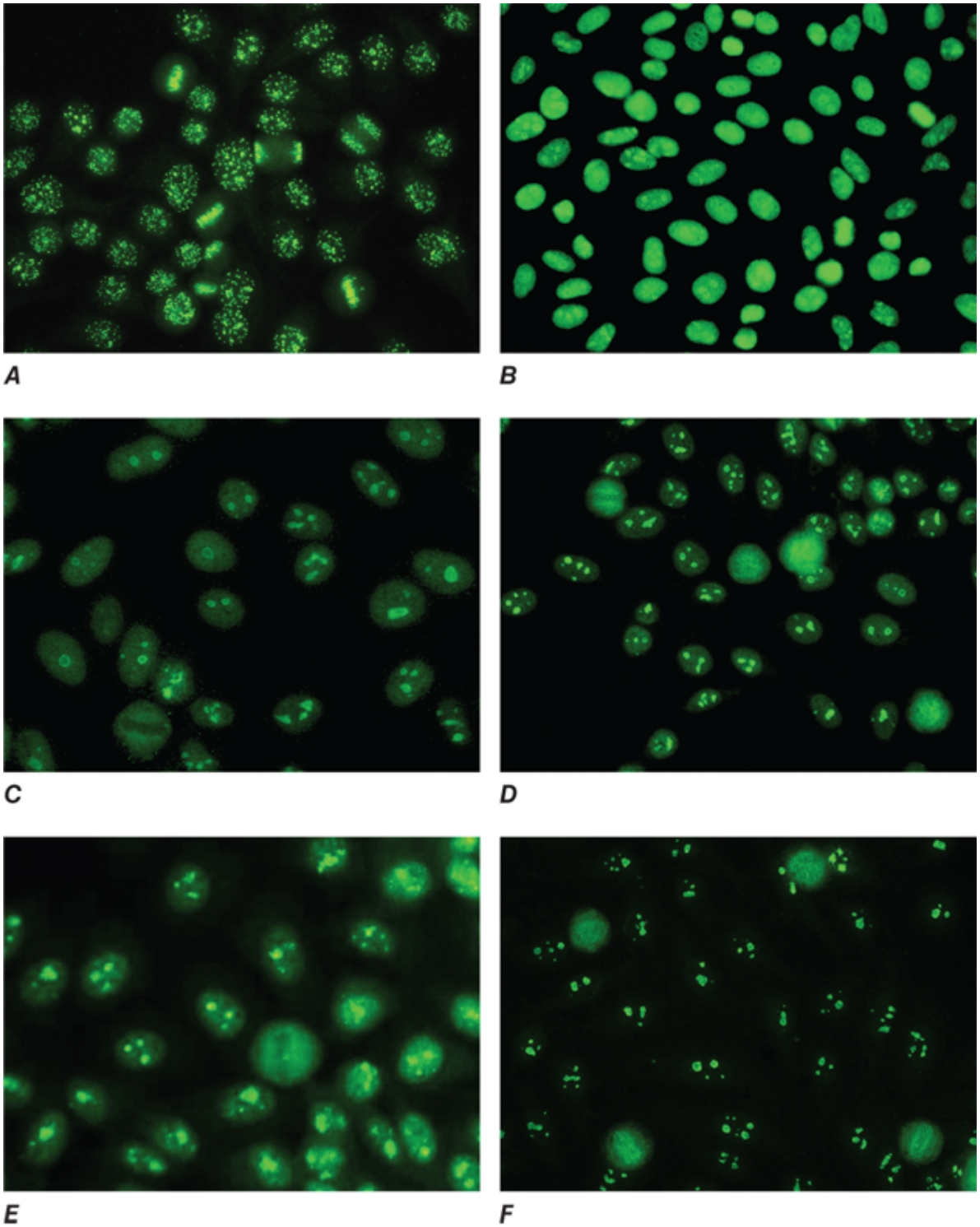
LABORATORY EVALUATION AND BIOMARKERS

Mild microcytic anemia is frequent and may indicate gastrointestinal bleeding caused by GAVE or chronic esophagitis. Macrocytic anemia may be caused by folate and vitamin B_{12} deficiency due to small-bowel bacterial overgrowth and malabsorption or by drugs such as methotrexate. Microangiopathic hemolytic anemia caused by mechanical fragmentation of red blood cells during their passage through microvessels coated with fibrin or platelet thrombi is a hallmark of scleroderma renal crisis. The erythrocyte sedimentation rate (ESR) is generally normal; an elevation may signal coexisting myositis or malignancy.

Antinuclear autoantibodies are detected in almost all patients with SSc. Anti-topoisomerase I (Scl-70) and anti-centromere antibodies are mutually exclusive and each is highly specific for SSc. Topoisomerase I antibodies are associated with increased risk of ILD and poor outcomes. Anti-centromere antibodies are associated with PAH, but only infrequently with significant cardiac, pulmonary, or renal involvement. Nucleolar immunofluorescence pattern may indicate antibodies to U3-RNP (fibrillarin), Th/To, or PM/Scl, whereas speckled immunofluorescence indicates antibodies to RNA polymerase III **(Fig. 353-13)**.

FIGURE 353-13

SSc-associated autoantibodies: immunofluorescence patterns. Indirect immunofluorescence on HEp-2 substrate shows distinct patterns: ***A.*** anti-centromere; ***B.*** anti-Scl-70/topoisomerase I; ***C.*** anti-PM/Scl; ***D.*** anti-Th/To; ***E.*** anti-RNA polymerase III; ***F.*** anti-fibrillarin/U3RNP antibodies. Except for anti-centromere (discrete dots in metaphase nucleus), variations of nucleolar staining are clues to autoantibody specificity. However, immunoassays employing purified autoantigens are recommended to confirm specificity of these autoantibodies. *(Courtesy of Marvin Fritzler and Susan Copple, Inova Diagnostics Inc., San Diego, California.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition: www.accessmedicine.com
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS, STAGING, AND MONITORING

The diagnosis of SSc is made primarily on clinical grounds and is generally straightforward in patients with established disease. The presence of skin induration with a characteristic symmetric distribution pattern associated with typical visceral organ manifestations establishes the diagnosis with a high degree of certainty. In lcSSc, a history of Raynaud's phenomenon and GERD symptoms, coupled with sclerodactyly and nailfold capillary changes, often in combination with cutaneous telangiectasia and calcinosis cutis, help to establish the diagnosis. Primary

Raynaud's disease is a benign condition that must be differentiated from early or limited SSc. Nailfold microscopy is particularly helpful in this situation, because in contrast to SSc, nailfold capillaries are normal. Diagnosing SSc at an early stage may be a challenge. In dcSSc, initial symptoms are often nonspecific, Raynaud's phenomenon may be absent, and physical examination may only show upper extremity edema and puffy fingers. Patients with early SSc might be diagnosed as arthritis, SLE, myositis, or, most commonly, undifferentiated connective tissue disease. Within weeks to months, Raynaud's phenomenon and advancing skin induration appear. SSc-specific autoantibodies provide a high degree of diagnostic certainty. Raynaud's phenomenon with fingertip ulcerations or other evidence of digital ischemia, coupled with telangiectasia, distal esophageal dysmotility, unexplained ILD or PAH, or accelerated hypertension with renal failure in the absence of clinically evident skin induration, suggests the diagnosis of SSc *sine* scleroderma.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT: Management of Systemic Sclerosis

OVERVIEW: GENERAL PRINCIPLES

To date, with the possible exception of hematopoietic stem cell therapy (HSCT), no therapy has been shown to significantly alter the natural history of SSc. In contrast, multiple interventions are highly effective in alleviating the symptoms, slowing the progression of the cumulative organ damage, and reducing disability. A significant reduction in disease-related mortality has been noted during the past 25 years. In light of the marked heterogeneity in disease manifestations, and natural history, the management of SSc mandates a "personalized medicine" approach that is specifically tailored to each individual patient's unique needs.

The following general principles should guide management (**Table 353-7**): prompt and accurate diagnosis; classification and risk stratification based on clinical and laboratory evaluation, including prognostic and predictive biomarkers; early recognition of organ-based complications and assessment of their extent, severity, and likelihood of deterioration; regular monitoring for disease progression, new complications, and response to therapy; adjusting therapy; and patient education. In order to minimize irreversible organ damage, management should be proactive, with regular screening and initiation of appropriate, intervention at the earliest possible opportunity. In light of the complex and multisystemic nature of the SSc, a team-oriented management approach integrating appropriate specialists should be pursued. Generally, a combination of drugs that impact different aspects of the disease is used. Patients should be encouraged to become familiar with potential complications and understand therapeutic options, including interventional trials, and natural history, and empowered to partner with their treating physicians. This requires a long-term relationship between patient and physician, with ongoing counseling, encouragement, and two-way dialogue.

DISEASE-MODIFYING THERAPY: IMMUNOSUPPRESSIVE AGENTS

Immuno-suppressive agents used in other autoimmune diseases have generally shown modest or no benefit in SSc. Glucocorticoids alleviate stiffness and aching in early inflammatory-stage dcSSc, but do not influence the progression of skin or internal organ involvement. Since their use is associated with an increased risk of scleroderma renal crisis, glucocorticoids should be given only when absolutely necessary, at the lowest dose possible, and for brief periods only.

Cyclophosphamide has been extensively studied in light of its efficacy in the treatment of vasculitis (Chap. 356), SLE (Chap. 349), and other autoimmune diseases (Chap. 348). Both oral and intravenous cyclophosphamide have been shown to reduce the progression of SSc-associated ILD, with stabilization and, rarely, modest improvement of pulmonary function, HRCT findings, respiratory symptoms, and skin induration. The benefits of cyclophosphamide need to be balanced against its potential toxicity, including bone marrow suppression, opportunistic infections, hemorrhagic cystitis and bladder cancer, premature ovarian failure, and late secondary malignancies.

Methotrexate had modest effect on SSc skin involvement in small studies. Mycophenolate mofetil was evaluated in both open label and randomized control trials. Both skin induration and ILD improved in patients treated with MMF, and the drug was well tolerated. Tocilizumab, a monoclonal antibody directed against the IL-6 receptor that blocks IL-6 signaling, also showed benefit in randomized SSc trials. Open-label studies and small trials provide support for the use of rituximab, a monoclonal antibody directed against the mature B cell marker CD20, along with extracorporeal photopheresis and IV immunoglobulin. Randomized trials in SSc evaluating the efficacy of abatacept, a fusion protein that inhibits T cell co-stimulation and function, are on-going. The use of cyclosporine, azathioprine, plaquenil, thalidomide, and rapamycin is currently not well supported by the literature. Intensive immune ablation using high-dose chemotherapy, followed by autologous stem cell reconstitution, in selected patients was associated with durable remission and improved long-term survival in randomized clinical trials; however, this regimen carries potential morbidity and mortality, as well as significant cost.

Antifibrotic Therapy

Because tissue fibrosis underlies organ damage in SSc, drugs that interfere with the fibrotic process represent a rational therapeutic approach. In older retrospective studies, d-penicillamine was shown to stabilize skin induration, prevent new internal organ involvement, and improve survival. However, a randomized-controlled clinical trial in early active SSc found no difference in the extent of skin involvement between patients treated with standard-dose (750 mg/d) or very low-dose (125 mg every other day) d-penicillamine. Recent clinical trials show benefit of pirfenidone and of nintedanib in patients with idiopathic pulmonary fibrosis, with significant slowing of the loss of lung function. Whether these anti-fibrotic drugs have comparable efficacy and tolerability in patients with SSc-associated ILD and other fibrotic manifestations of the disease is under investigation.

Vascular Therapy

The goal of Raynaud's therapy is to control episodes, prevent and enhance the healing of ischemic complications, and slow the progression of obliterative vasculopathy. Patients should dress warmly, minimize cold exposure, and avoid drugs that precipitate or exacerbate vasospastic episodes. Extended-release dihydropyridine calcium channel blockers such as amlodipine and diltiazem ameliorate Raynaud's phenomenon, but their use is often limited by side effects (palpitations, dependent edema, worsening gastroesophageal reflux). While ACE inhibitors do not reduce the frequency or severity of episodes, Angiotensin II receptor blockers such as losartan are effective and well tolerated. Patients with Raynaud's phenomenon unresponsive to these therapies may require the addition of α_1-adrenergic receptor blockers (e.g., prazosin), 5-phosphodiesterase inhibitors (e.g., sildenafil), topical nitroglycerine, and intermittent IV infusions of prostaglandins. Low-dose aspirin and dipyridamole prevent platelet aggregation and may have a role as adjunctive agents. In patients with ischemic digital tip ulcerations, the endothelin-1 receptor antagonist bosentan reduces the risk of new ulcers. Digital sympathectomy and intradigital injections of botulinum type A (Botox) may be considered in patients with severe on-going ischemia. Empirical long-term therapy with statins and antioxidants may retard the progression of vascular damage and obliteration. There is limited evidence-

based information for the treatment of cardiac complications of SSc, which should be guided by specialists experienced in their diagnosis and management. While selective beta blockers such as metorpolol can precipitate vasospasm, non-dihydropyridine calcium channel blockers can be used for rate control in atrial arrhythmias, and non-selective alpha/beta blockers such as carvediol for improving myocardial perfusion and left ventricular systolic function.

TABLE 353-7

Key Principles in Management

- Establish early and accurate diagnosis.
- Detect and evaluate internal organ involvement.
- Define clinical disease stage and activity.
- Tailor individualized therapy to each patient's unique needs.
- Assess treatment response, and adjust therapy as needed; monitor for disease activity, progression and new complications.

TREATMENT

TREATMENT

TREATMENT OF SSc-ASSOCIATED ILD

ILD is a leading cause of death in patients with SSc. However, as SSc-associated ILD is not necessarily progressive, it is important to identify patients who are at high risk for disease progression in the absence of treatment. The extent of ILD on HRCT and the FVC at initial evaluation, and decline in PFTs during the preceding 12-month period, are helpful in identifying these patients. Patients at high risk for ILD should be monitored by performing PFTs every 6 months; serial HRCT imaging is not recommended. Cyclophosphamide, given IV or orally for 6 to 12 months, and mycophenolate mofetil slow the decline in lung function and improve respiratory symptoms; however, cyclophosphamide is associated with more frequent side effects. The safety and efficacy of anti-fibrotic drugs recently approved for idiopathic pulmonary fibrosis in the treatment of SSc-associated ILD are currently under investigation. In certain patients who show continued progression of ILD despite medical therapy, lung transplantation might be considered as a life-prolonging procedure, although significant GERD is a concern in SSc. Recurrence of SSc-ILD in transplanted lung allografts has not been reported.

TREATMENT OF GASTROINTESTINAL COMPLICATIONS

Because oral problems including decreased oral aperture, decreased saliva production, gum recession, periodontal disease, and teeth loss are common, regular dental care is recommended. Gastroesophageal reflux is very common and may occur in the absence of symptoms. Patients should be instructed to elevate the head of the bed, eat frequent small meals, and avoid alcohol, caffeine, and known reflux exacerbants, or meals before bedtime. Proton pump inhibitors reduce acid reflux and in patients with SSc may need to be given in relatively high doses. Prokinetic agents such as metoclopramide, erythromycin (a motilin agonist), and domperidone may occasionally be helpful, but are frequently associated with side effects. Botulinum toxin injection sometimes ameliorates impaired gastric

emptying. Anti-reflux procedures such as Nissen fundoplication can result in secondary achalasia and generally should be avoided. Episodic bleeding from GAVE (watermelon stomach) may be amenable to treatment with endoscopic ablation using laser or argon plasma photocoagulation, although bleeding frequently recurs. Some patients may require enteral feeding and/or decompression via percutaneous gastrostomy or jejunostomy. Small bowel bacterial overgrowth secondary to dysmotility causes abdominal bloating and diarrhea, and may lead to malabsorption and severe malnutrition. Treatment with short courses of rotating broad-spectrum antibiotics such as metronidazole, erythromycin, and rifaximin can eradicate bacterial overgrowth. Small bowel hypomotility may respond to octreotide; however, pseudo-obstruction is difficult to treat. Fecal incontinence, a frequent and under-reported complication, may respond to anti-diarrheal medication, biofeedback therapy, sphincter augmentation, and sacral neuromodulation. Potential malnutrition should be routinely assessed.

TREATMENT OF PAH

In SSc, PAH carries an extremely poor prognosis and accounts for 30% of deaths. Because PAH is asymptomatic until advanced, patients with SSc should be screened at initial evaluation, and regularly thereafter. Treatment is generally started with an oral endothelin-1 receptor antagonist such as bosentan or a phosphodiesterase 5 inhibitor such as sildenafil. Recently, the soluble guanylate cyclase stimulator riociguat, which acts by increasing the production of nitric oxide, and the selective IP prostacyclin receptor agonist selexipag, were shown to improve PAH symptoms and survival. Patients may also require diuretics and digoxin. If hypoxemia is documented, supplemental oxygen should be prescribed in order to avoid secondary pulmonary vasoconstriction. Prostacyclin analogues such as epoprostenol or treprostinil can be given by continuous IV or SC infusion, or via intermittent nebulized inhalations. Combination therapy with different classes of agents acting additively or synergistically is often necessary. Lung transplantation remains an option for selected SSc patients with PAH who fail medical therapy, and 2-year survival rates (64%) are comparable to those of idiopathic ILD or PAH.

MANAGEMENT OF RENAL CRISIS

Scleroderma renal crisis is a medical emergency. Since the outcome is largely determined by the extent of renal damage at the time that aggressive therapy is initiated, prompt recognition of impending or early scleroderma renal crisis is essential, and efforts should be made to avoid its occurrence. High-risk SSc patients with early disease, extensive and progressive skin involvement, tendon friction rubs, and anti-RNA polymerase III antibodies should be instructed to monitor their blood pressure daily and report significant alterations immediately. Potentially nephrotoxic drugs should be avoided, and glucocorticoids should be used only when absolutely necessary and at low doses. Patients presenting with scleroderma renal crisis should be immediately hospitalized. Once other causes of renal disease are excluded, treatment should be started promptly with titration of short-acting ACE inhibitors, with the goal of achieving rapid normalization of the blood pressure. In patients with persistent hypertension, addition of angiotensin II receptor blockers, calcium channel blockers, endothelin-1 receptor blockers, prostacyclins, and direct renin inhibitors should be considered. Up to two-thirds of patients with scleroderma renal crisis will require dialysis. Substantial renal recovery can occur, and dialysis can be discontinued in 30–50% of the patients. Kidney transplantation is appropriate for patients unable to discontinue dialysis after 2 years. Survival of transplanted SSc patients is comparable to that of other diseases, and recurrence of renal crisis is rare.

SKIN CARE

Because skin involvement in SSc is never life-threatening and it stabilizes and may even regress spontaneously, disease management should not be dictated by its cutaneous manifestations. The inflammatory symptoms of early skin involvement can be controlled with antihistamines and short-term use of low-dose glucocorticoids (<5 mg/d of

prednisone). Cyclophosphamide and methotrexate have modest effects on skin induration. Because the skin is dry, the use of hydrophilic ointments and bath oils is encouraged, and regular skin massage is helpful. Telangiectasia, which presents a cosmetic problem, especially on the face, can be treated with pulsed dye laser. Ischemic digital ulcerations should be protected by occlusive dressing to promote healing and prevent infection. Infected skin ulcers are treated with topical antibiotics and surgical debridement. While no therapy has been shown to be effective in preventing soft tissue calcific deposits or promoting their dissolution, reports support the use of diltiazem, minocycline, bisphosphonates, and topical or IV sodium thiosulfate (STS). Other therapies that have been used for calcinosis include carbon dioxide laser, extracorporeal shock-wave lithotripsy, and surgical high-speed microdrilling.

TREATMENT OF MUSCULOSKELETAL COMPLICATIONS

Arthralgia and joint stiffness are very common and distressing manifestations in early-stage disease. Short courses of nonsteroidal anti-inflammatory agents, methotrexate, and cautious use of low-dose glucocorticoids alleviate symptoms. Physical and occupational therapy can be effective for preventing loss of musculoskeletal function and joint contractures, and should be initiated early.

COURSE

The natural history of SSc is highly variable and difficult to predict, especially in early stages of the disease. Patients with dcSSc tend to have a more rapidly progressive course and worse prognosis than those with lcSSc. Inflammatory symptoms of early dcSSc, such as fatigue, edema, joint pain and pruritus subside, and skin thickening reach a plateau at 2–4 years after disease onset. It is during the early edematous/inflammatory stage that life-threatening visceral organ involvement may develop. While existing visceral organ involvement, such as ILD, may progress even after skin involvement peaks, new organ involvement is rare. Scleroderma renal crisis generally occurs within the first 4 years of disease. In late-stage disease (>6 years), the skin is usually soft and atrophic. Skin regression characteristically occurs in an order that is the reverse of initial involvement, with softening on the trunks followed by proximal and finally distal extremities; however, sclerodactyly and fixed finger contractures generally persist. Relapse or recurrence of skin thickening after peak skin involvement has been reached is uncommon. Patients with lcSSc follow a clinical course that is markedly different than that of dcSSc. Raynaud's phenomenon typically precedes other disease manifestations by years or even decades. Visceral organ complications such as PAH generally develop late and progress slowly.

PROGNOSIS

SSc confers a substantial increase in the risk of premature death. Age- and gender-adjusted mortality rates are fivefold to eightfold higher compared to the general population, and more than half of all patients with SSc die from their disease. In one population-based study of SSc, the median survival was 11 years. In patients with dcSSc, 5- and 10-year survival rates are 70% and 55%, respectively, whereas in patients with lcSSc, 5- and 10-year survival rates are 90% and 75%, respectively. The prognosis correlates with the extent of skin involvement, which itself is a surrogate for visceral organ involvement. Major causes of death are PAH, pulmonary fibrosis, gastrointestinal involvement, and cardiac disease. Scleroderma renal crisis is associated with a 30% 3-year mortality. Lung cancer and excess cardiovascular deaths also contribute to increased mortality. Markers of poor prognosis include male gender, African-American race, older age at disease onset, extensive skin thickening with truncal involvement,

palpable tendon friction rubs, and evidence of significant or progressive visceral organ involvement. Laboratory predictors of increased mortality at initial evaluation include an elevated ESR, anemia, proteinuria, and anti–topoisomerase I antibodies. In one study, SSc patients with extensive skin involvement, vital capacity <55% predicted, significant gastrointestinal involvement (pseudo-obstruction or malabsorption), clinical evidence of cardiac involvement, or scleroderma renal crisis had a 9-year survival of <40%. The severity of PAH predicts mortality, and patients with mean pulmonary arterial pressure ≥45 mmHg had a 33% 3-year survival. The advent of ACE inhibitors in scleroderma renal crisis had a dramatic impact on survival, increasing from <10% at 1 year in the pre–ACE inhibitor era to >70% 3-year survival at the present time. Moreover, 10-year survival in SSc has improved from <60% in the 1970s to >66–78% in the 1990s, a trend that reflects both earlier detection and better management of complications.

LOCALIZED SCLERODERMA

The term *scleroderma* describes a group of localized skin disorders (Table 353-1). These occur more commonly in children than in adults, and in marked contrast to SSc, are generally not complicated by Raynaud's phenomenon or significant internal organ involvement. Morphea presents as solitary or multiple circular patches of thick skin or, rarely, as widespread induration (generalized or pansclerotic morphea); the fingers are generally spared. Linear scleroderma may affect subcutaneous tissues, leading to fibrosis and atrophy of supporting structures, tendons, muscle, and even bone. In children, the growth of affected long bones can be retarded. When linear scleroderma crosses large joints, significant contractures can develop.

MIXED CONNECTIVE TISSUE DISEASE

Patients who have lcSSc coexisting with features of SLE, polymyositis, and rheumatoid arthritis may have mixed connective tissue disease (MCTD). This overlap syndrome is generally associated with the presence of high titers of autoantibodies to U1-RNP. The characteristic initial presentation is Raynaud's phenomenon associated with puffy fingers and myalgia. Over time, sclerodactyly, soft tissue calcinosis, and cutaneous telangiectasia may appear. Skin rash suggestive of SLE (malar erythema, photosensitivity) or dermatomyositis (heliotrope rash on the eyelids, erythematous rash on knuckles) occur. Arthralgia is common, and some patients develop erosive polyarthritis. Pulmonary fibrosis and isolated or secondary PAH may develop. Other manifestations include esophageal dysmotility, pericarditis, Sjögren's syndrome, and renal disease, especially membranous glomerulonephritis. Laboratory evaluation shows elevated ESR and hypergammaglobulinemia. While anti-U1RNP antibodies are detected in high titers, SSc-specific autoantibodies are absent. In contrast to SSc, MCTD often responds to glucocorticoids, and the long-term prognosis is better than that of SSc. Whether MCTD is truly a distinct entity or is a subset of SLE or SSc, remains controversial.

EOSINOPHILIC FASCIITIS (DIFFUSE FASCIITIS WITH EOSINOPHILIA)

Eosinophilic fasciitis is a rare idiopathic disorder of adults associated with abrupt skin induration. The skin characteristically shows a coarse cobblestone "peau d'orange" appearance. In contrast to SSc, Raynaud's phenomenon and SSc-associated internal organ involvement and autoantibodies are absent. Furthermore, skin involvement spares the fingers. Full-thickness biopsy of the lesional skin reveals fibrosis of the subcutaneous fascia, with variable inflammation and eosinophil infiltration. In the acute phase of the illness, peripheral blood

eosinophilia may be prominent. MRI appears to be a sensitive tool for the diagnosis of eosinophilic fasciitis. Eosinophilic fasciitis can occur in association with, or preceding, various myelodysplastic syndromes or multiple myeloma. Although glucocorticoids cause prompt resolution of eosinophilia, the skin shows slow and variable improvement. The prognosis of patients with eosinophilic fasciitis who do not develop hematological complications is generally good.

FURTHER READING

Allanore Y et al: Systemic sclerosis. Nat Rev Dis Primers 1:15002, 2015.

Herzog EL et al: Interstitial lung disease associated with systemic sclerosis and idiopathic pulmonary fibrosis: How similar or distinct? Arthritis Rheum 66:1967, 2014.

Joseph CG et al: Association of the autoimmune disease scleroderma with an immunologic response to cancer. Science 343:152, 2014. [PubMed: 24310608]

Martyanov V, Whitfield ML: Molecular stratification and precision medicine in systemic sclerosis from genomic and proteomic data. Curr Opin Rheumatol 28:83, 2016. [PubMed: 26555452]

Tashkin DP et al: Mycophenolate mofetil versus oral cyclophosphamide in scleroderma-related interstitial lung disease (SLS II): A randomised controlled, double-blind, parallel group trial. Lancet Respir Med 4:708, 2016. [PubMed: 27469583]

Chapter 354: Sjögren's Syndrome

Haralampos M. Moutsopoulos

FIGURE 354-1

DEFINITION, INCIDENCE, AND PREVALENCE

Sjögren's syndrome is a chronic, slowly progressing autoimmune disease characterized by lymphocytic infiltration of the exocrine glands resulting in xerostomia and dry eyes (keratoconjunctivitis sicca). The syndrome has unique features since it presents with a wide clinical spectrum from organ-specific autoimmune exocrinopathy to systemic disease. A small but significant number of patients develop malignant lymphoma. The disease can present as an entity alone or in association with other autoimmune diseases (Table 354-1). Finally, the histopathologic lesion in the labial minor salivary glands is easily accessible aiding the diagnosis, prognosis and disease pathogenesis.

TABLE 354-1

Association of Sjögren's Syndrome with Other Autoimmune Diseases

Rheumatoid arthritis
Systemic lupus erythematosus
Scleroderma
Mixed connective tissue disease
Primary biliary cirrhosis
Autoimmune thyroid disease
Chronic active hepatitis

Middle-aged women (female-to-male ratio, 9:1) are primarily affected, although Sjögren's syndrome may occur at any age, including childhood. The prevalence of primary Sjögren's syndrome is ~0.5–1%, while 5–20% of patients with other autoimmune diseases suffer from Sjögren's syndrome (secondary).

PATHOGENESIS

Sjögren's syndrome is characterized by both lymphocytic infiltration of the exocrine glands and B lymphocyte hyperreactivity. An oligomonoclonal B cell process, which is characterized by cryoprecipitable

monoclonal immunoglobulins (IgMκ or IgAκ) with rheumatoid factor activity, is evident in up to 10% of patients.

Sera from patients with Sjögren's syndrome often contain autoantibodies to non-organ-specific antigens such as immunoglobulins (rheumatoid factors) and extractable nuclear and cytoplasmic antigens (Ro/SS-A, La/SS-B). Ro/SS-A autoantigen consists of two polypeptides (52 and 60 kDa, respectively) in conjunction with cytoplasmic RNAs, whereas the 48-kDa La/SS-B protein is bound to RNA III polymerase transcripts. Autoantibodies to Ro/SS-A and La/SS-B antigens are usually present prior to diagnosis and are associated with earlier disease onset, longer disease duration, salivary gland enlargement, extraglandular (systemic) manifestations, and more intense lymphocytic infiltration of minor salivary glands.

The major infiltrating cells in the affected exocrine glands are activated T lymphocytes in mild lesions, whereas B cells prevail in severe lesions. Macrophages and dendritic cells are also found. The number of macrophages positive for interleukin (IL) 18 has been shown to be associated with parotid gland enlargement and low serum levels of the C4 component of complement, both of which are adverse predictors for lymphoma development.

Ductal and acinar epithelial cells appear to play a significant role in the initiation and perpetuation of autoimmune injury. These cells (1) express costimulatory molecules, and inappropriately the intracellular autoantigens Ro/SS-A and La/SS-B on their membranes, acquiring the capacity to provide signals essential for lymphocytic activation; (2) produce proinflammatory cytokines and lymphocyte attracting chemokines necessary for sustaining the autoimmune lesion and allowing the formation of ectopic germinal centers, a finding predicting lymphoma development; and (3) express functional receptors of innate immunity, particularly Toll-like receptors (TLRs) 3, 7, and 9, molecules which may account for the initiation of the autoimmune reactivity.

Both infiltrating T and B cells have a tendency to be resistant to apoptosis. Levels of B cell–activating factor (BAFF) have been found to be elevated in the serum and tissues of Sjögren's syndrome patients, especially those with hypergammaglobulinemia, and probably accounts for the anti-apoptotic effect on B lymphocytes. Glandular epithelial cells seem to have an active role in the production of BAFF, which may be expressed and secreted after stimulation with type I and II interferons. The latter have been detected in ductal epithelial cells and T cells. The triggering factor for epithelial activation appears to be enteroviral infection.

Molecular analysis of human leukocyte antigen (HLA) class II genes has revealed that Sjögren's syndrome, regardless of the patient's ethnic origin, is highly associated with the HLA DQA1*0501 allele. Genome-wide association studies have disclosed an increased prevalence of single-nucleotide polymorphisms in genes of IRF-5 and STAT-4, which participate in the activation of the type I interferon pathway.

CLINICAL MANIFESTATIONS

The majority of patients with Sjögren's syndrome have symptoms related to impaired lacrimal and salivary gland function. The disease evolution is slow and in the majority of patients runs a benign course. Studies

have shown that prior to disease onset, patients with Sjögren's syndrome experience major stressful life events with which they cannot cope adequately.

The principal oral symptom of Sjögren's syndrome is dryness (xerostomia). Patients report difficulty in swallowing dry food, a burning mouth sensation, an increase in dental caries, and problems in wearing complete dentures. Physical examination shows a dry, erythematous, sticky oral mucosa. There is atrophy of the filiform papillae on the dorsum of the tongue, and saliva from the major glands is either not expressible or cloudy. Enlargement of the parotid or other major salivary glands occurs in two-thirds of patients with primary Sjögren's syndrome but is uncommon in those in association with rheumatoid arthritis. Diagnostic tests include sialometry and newer imaging techniques, including ultrasound, MRI, and magnetic resonance sialography of the major salivary glands. Biopsy of the labial minor salivary gland permits histopathologic confirmation of focal lymphocytic infiltrates.

Ocular involvement is the other major manifestation of Sjögren's syndrome. Patients usually describe a sandy or gritty feeling under the eyelids. Other ocular symptoms include burning, accumulation of secretions in thick strands at the inner canthi, decreased tearing, redness, itching, eye fatigue, and increased photosensitivity. These symptoms, which define *keratoconjunctivitis sicca*, are attributed to the destruction of corneal and bulbar conjunctival epithelium. Diagnostic evaluation of keratoconjunctivitis sicca includes measurement of tear flow by Schirmer's I test and determination of tear composition, with assessment of tear breakup time or tear lysozyme content. Slit-lamp examination of the cornea and conjunctiva after lissamine green or Rose Bengal staining reveals punctuate corneal ulcerations and attached filaments of corneal epithelium.

Involvement of other exocrine glands, which occurs less frequently, includes a decrease in mucous gland secretions of the upper and lower respiratory tree, resulting in dry nose, throat, and trachea (xerotrachea). In addition, diminished secretion of the exocrine glands of the gastrointestinal tract leads to esophageal mucosal atrophy and atrophic gastritis. Dyspareunia due to dryness of the external genitalia and dry skin also may occur.

Extraglandular (systemic) manifestations are seen in one-third of patients with Sjögren's syndrome **(Table 354-2)** but are very rare in patients whose Sjögren's syndrome is associated with rheumatoid arthritis. They can be categorized as follows: Non-specific, involvement of parenchymal organs by lymphocytes (peri-epithelial), immune complex-mediated pathology, and lymphoma development. In the first category easy fatigability, low-grade fever, Raynaud's phenomenon, myalgias, arthralgias, and arthritis are included. Arthritis in patients with primary Sjögren's syndrome is non-erosive. Involvement of parenchymal organs such as the lungs, kidneys and the liver is due to peri-epithelial accumulation of lymphocytes. On the basis of this observation the term **autoimmune epithelitis** has been coined. Lung involvement is usually manifested with dry cough and rarely with dyspnea. The underlying lung pathology includes peribronchial infiltrates and rarely lymphocyte interstitial pneumonitis. Renal involvement includes interstitial nephritis, clinically manifested by hyposthenuria and renal tubular dysfunction with or without acidosis. Untreated acidosis may lead to nephrocalcinosis. Immune complex-mediated disease is expressed with vasculitis affecting primarily small-sized vessels, mainly manifested with purpura and rarely with urticarial rash, skin ulcerations,

mononeuritis multiplex, and membranoproliferative glomerulonephritis associated with mixed cryoglobulinemia. Central nervous system involvement is rarely recognized. A few cases of myelitis associated with antibody to aquaporin 4 have been described.

TABLE 354-2

Prevalence of Extraglandular Manifestations in Primary Sjögren's Syndrome

Clinical Manifestation	Percent	Remarks
Non-Specific		
Fatigability/Myalgias	25	Fibromyalgia
Arthralgias/Arthritis	60	Usually non-erosive, leading to Jaccoud's arthropathy
Raynaud's phenomenon	37	In one-third of patients, precedes sicca manifestations
Peri-Epithelial		
Lung involvement	14	Small airway disease/lymphocyte interstitial pneumonitis
Kidney involvement	9	Interstitial kidney disease is usually asymptomatic
Liver involvement	6	Primary biliary cirrhosis stage I
Immune-Complex mediated		
Small vessel vasculitis		Purpura, urticarial lesions
Peripheral neuropathy	2	Polyneuropathy, either sensory or sensorimotor
Glomerulonephritis		Membranoproliferative
Lymphoma		
Lymphoma	6	Glandular MALT[a] lymphoma is most common

[a]Mucosa-associated lymphoid tissue.

Patients with Sjögren's syndrome associated with rheumatoid arthritis and systemic lupus erythematosus have an increased cardiovascular risk.

Lymphoma in Sjögren's syndrome usually presents later in the disease course. Persistent parotid gland enlargement, purpura, leukopenia, cryoglobulinemia, low serum C4 complement levels, autoantibodies (anti-Ro/SS-A, anti-La/SS-B), and ectopic germinal center formation in minor salivary glands are manifestations predicting the development of lymphoma. Most lymphomas are extranodal, low-grade marginal-zone B cell lymphomas and are usually detected incidentally during evaluation of the labial minor salivary gland biopsy. The affected lymph nodes are usually peripheral. Survival rates are decreased in patients with B symptoms, lymph node mass >7 cm in diameter, and high or intermediate histologic grade.

Routine laboratory tests in Sjögren's syndrome reveal mild normochromic, normocytic anemia. An elevated erythrocyte sedimentation rate is found in ~70% of patients. Certain autoantibodies may determine different disease phenotypes. Patients positive for anticentromere autoantibody present with a clinical picture similar to that of limited scleroderma **(Chap. 353)**. Antimitochondrial antibodies may connote liver involvement in the form of primary biliary cirrhosis **(Chap. 339)**. Autoantibodies to 21-hydroxylase are found in almost 20% of patients in association with a blunted adrenal response.

DIAGNOSIS AND DIFFERENTIAL DIAGNOSIS

Primary Sjögren's syndrome is diagnosed if (1) the patient presents with eye and/or mouth dryness, (2) eye tests disclose keratoconjunctivitis sicca, (3) mouth evaluation reveals dry oral mucosa, and/or (4) the patient's serum reacts with immunoglobulins (rheumatoid factors), Ro/SS-A, and/or La/SS-B autoantigens. Labial biopsy is needed for diagnostic and prognostic purposes as well as to rule out other conditions that may cause dry mouth or eyes or parotid gland enlargement **(Tables 354-3 and 354-4)**. Enlargement of major salivary glands, particularly in patients without autoantibodies, should raise the suspicion of IgG4-related syndrome. Validated methods of disease activity and classification criteria have been established **(Table 354-5)**.

TABLE 354-3

Differential Diagnosis of Sicca Symptoms

Xerostomia	Dry Eye	Bilateral Parotid Gland Enlargement
Viral infections (HCV, HIV) Drugs Psychotherapeutic Parasympatholytic Antihypertensive Psychogenic origin Irradiation Diabetes mellitus Trauma Sjögren's syndrome Amyloidosis	Inflammation Stevens-Johnson syndrome Pemphigoid Chronic conjunctivitis Chronic blepharitis Sjögren's syndrome Toxicity Burns Drugs Neurologic conditions Impaired lacrimal gland function Impaired eyelid function Miscellaneous Trauma Hypovitaminosis A Blink abnormality Anesthetic cornea Lid scarring Epithelial irregularity	Viral infections Mumps Influenza Epstein-Barr virus Coxsackievirus A Cytomegalovirus HIV, HCV Sarcoidosis, Tuberculosis IgG4 syndrome Sjögren's syndrome Metabolic disorders Diabetes mellitus Hyperlipoproteinemias (types IV and V) Chronic pancreatitis Hepatic cirrhosis Endocrine Acromegaly Gonadal hypofunction

TABLE 354-4

Differential Diagnosis of Sjögren's Syndrome

HIV Infection and Sicca Syndrome	Sjögren's Syndrome	Sarcoidosis
Predominant in young males	Predominant in middle-aged women	No age or sex preference
Lack of autoantibodies to Ro/SS-A and/or La/SS-B	Presence of autoantibodies	Lack of autoantibodies to Ro/SS-A and/or La/SS-B
Lymphoid infiltrates of salivary glands by CD8+ T lymphocytes	Lymphoid infiltrates of salivary glands by CD4+ T lymphocytes	Granulomas in salivary glands
Association with HLA-DR5	Association with HLA-DR3 and DRw52	Unknown
Positive serologic tests for HIV	Negative serologic tests for HIV	Negative serologic tests for HIV

TABLE 354-5

Revised International Classification Criteria for Sjögren's Syndrome[a,b,c]

I. Ocular symptoms: a positive response to at least one of three validated questions.
   1. Have you had daily, persistent, troublesome dry eyes for >3 months?
   2. Do you have a recurrent sensation of sand or gravel in the eyes?
   3. Do you use tear substitutes more than three times a day?

II. Oral symptoms: a positive response to at least one of three validated questions.
   1. Have you had a daily feeling of dry mouth for >3 months?
   2. Have you had recurrent or persistently swollen salivary glands as an adult?
   3. Do you frequently drink liquids to aid in swallowing dry foods?

III. Ocular signs: objective evidence of ocular involvement defined as a positive result to at least one of the following two tests:
   1. Shirmer's I test, performed without anesthesia (≤5 mm in 5 min)
   2. Rose Bengal score or other ocular dye score (≥4 according to van Bijsterveld's scoring system)

IV. Histopathology: In minor salivary glands focal lymphocytic sialoadenitis, with a focus score ≥1.

V. Salivary gland involvement: objective evidence of salivary gland involvement defined by a positive result to at least one of the following diagnostic tests:
   1. Unstimulated whole salivary flow (≤1.5 mL in 15 min)
   2. Parotid sialography
   3. Salivary scintigraphy

VI. Antibodies in the serum to Ro/SS-A or La/SS-B antigens, or both.

[a]Exclusion criteria: past head and neck radiation treatment, hepatitis C infection, AIDS, preexisting lymphoma, sarcoidosis, graft-versus-host disease, use of anticholinergic drugs. [b]Primary Sjögren's syndrome: any four of the six items, as long as item IV (histopathology) or VI (serology) is positive; or any three of the four objective-criteria items (III, IV, V, VI). [c]In patients with a potentially associated disease (e.g., another well-defined connective tissue disease), the presence of item I or item II plus any two from among items III, IV, and V may be considered indicative of secondary Sjögren's syndrome.

Source: From C Vitali et al: Ann Rheum Dis 61:554, 2002. ©2002 with permission from BMJ Publishing Group Ltd.

TREATMENT

TREATMENT

Sjögren's Syndrome

Treatment of Sjögren's syndrome aims to relieve symptoms and limit the damage from chronic xerostomia and keratoconjunctivitis sicca through substitution or stimulation of impaired secretions **(Fig. 354-1)**.

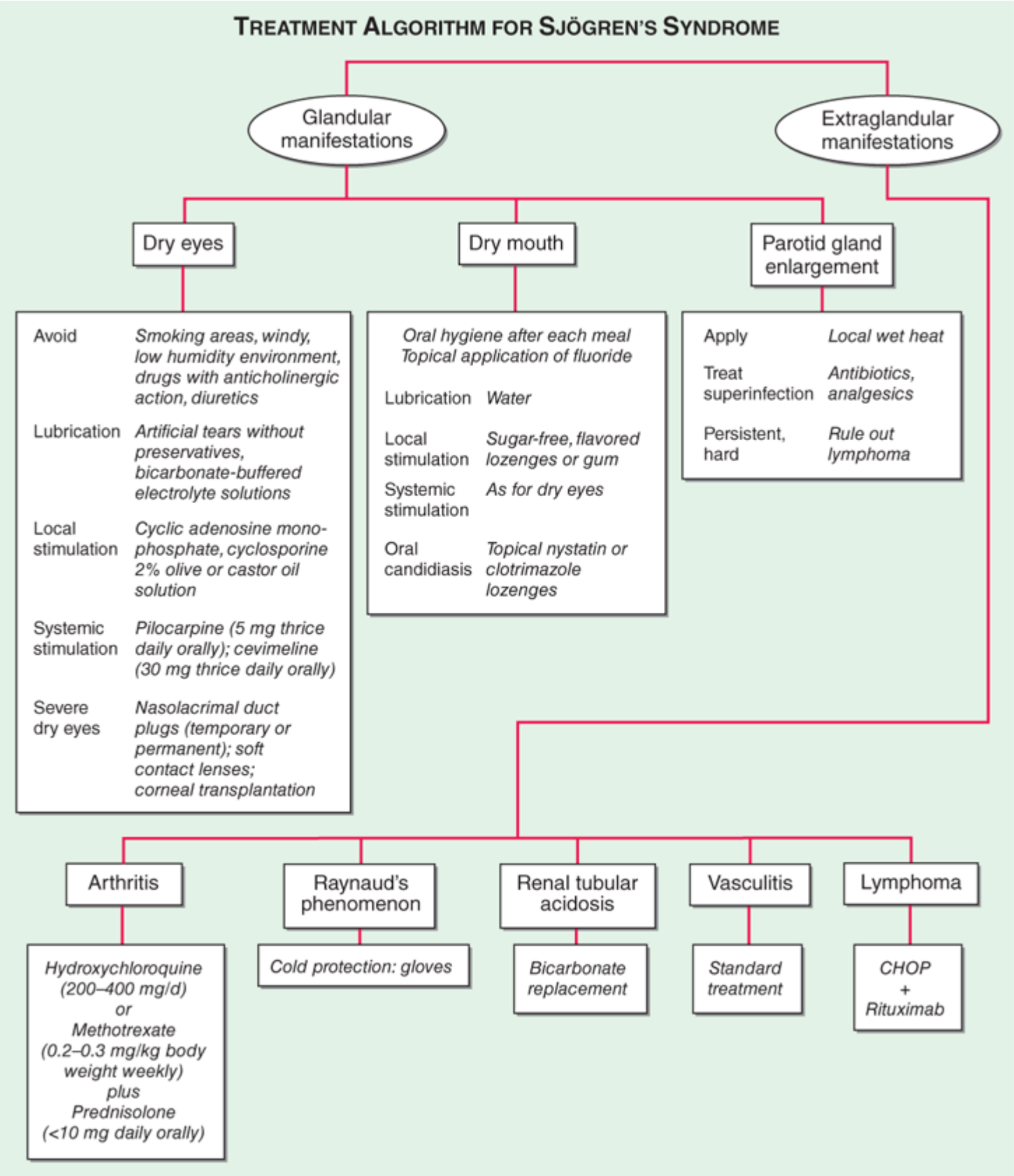
To replace deficient tears, several ophthalmic preparations are readily available (hydroxypropyl methylcellulose; polyvinyl alcohol; 0.5% methylcellulose; Hypo Tears). If corneal ulcerations are present, eye patching and boric acid ointments are recommended. Certain drugs that may decrease lacrimal and salivary secretions, such as diuretics, antihypertensive drugs, anticholinergics, and antidepressants, should be avoided.

For xerostomia, the best replacement is water. Propionic acid gels may be used to treat vaginal dryness. To stimulate secretions, orally administered pilocarpine (5 mg thrice daily) or cevimeline (30 mg thrice daily) appears to improve sicca manifestations, and both are well tolerated. Hydroxychloroquine (200 mg daily) is helpful for arthralgias and mild arthritis.

Patients with renal tubular acidosis should receive sodium bicarbonate by mouth (0.5–2 mmol/kg in four divided doses). Glucocorticoids and monoclonal antibody to CD20 (Rituximab) appear to be effective in patients with systemic disease, particularly in those with purpura, arthritis, and fatigability. Combination of anti-CD-20 with a classic CHOP regimen (cyclosporine, adriamycin [hydroxydaunorubicin], vincristine [oncovin], and prednisone) leads to increased survival rates among patients with high-grade lymphomas.

FIGURE 354-1

Treatment algorithm for Sjögren's syndrome. CHOP, cyclophosphamide, adriamycin (hydroxydaunorubicin), vincristine (oncovin), and prednisone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ACKNOWLEDGMENT

I would like to thank Dr. Athanasios G. Tzioufas for his contribution in the previous edition of the chapter.

FURTHER READING

Mavragani CP, Fragoulis GE, Moutsopoulos HM: Endocrine alterations in primary Sjogren's syndrome: An overview. J Autoimmun 39:354, 2012. [PubMed: 22695186]

Mavragani CP, Moutsopoulos HM: Sjögren's syndrome. CMAJ 186:579, 2014.

Mavragani CP, Nezos A, Moutsopoulos HM: New advances in the classification, pathogenesis and treatment of Sjogren's syndrome. Curr Opin Rheumatol 25:623, 2013. [PubMed: 23846338]

Moutsopoulos HM: Sjögren's syndrome: A forty-year scientific journey. J Autoimmun 51:1, 2014.

Chapter 355: The Spondyloarthritides

Joel D. Taurog

INTRODUCTION

The spondyloarthritides are a group of overlapping disorders that share certain clinical features, genetic associations, and pathogenic mechanisms. The classic designations include ankylosing spondylitis (AS), reactive arthritis (ReA), psoriatic arthritis (PsA) and spondylitis, enteropathic arthritis and spondylitis, juvenile-onset spondyloarthritis (JSpA), and undifferentiated SpA. More recently, these disorders have been broadly classified as predominantly axial SpA, affecting the spine, pelvis, and thoracic cage, or predominantly peripheral SpA, affecting the extremities.

ANKYLOSING SPONDYLITIS AND AXIAL SPONDYLOARTHRITIS

AS is an inflammatory disorder of unknown cause that primarily affects the axial skeleton; peripheral joints and extraarticular structures are also frequently involved. The disease usually begins in the second or third decade. The term *axial spondyloarthritis* (ax-SpA) is now in common use, supported by criteria formulated in 2009 **(Table 355-1)**. This classification includes definite AS, early stages that will progress to meet classical criteria for AS, and one or more nonprogressing phenotypes. The estimated prevalence of ax-SpA in the US adult population is from 0.9 to 1.4%, similar to that of rheumatoid arthritis (RA).

TABLE 355-1

ASAS Criteria for Classification of Axial Spondyloarthritis (to be applied for patients with back pain ≥3 months and age of onset <45 years)[a]

Sacroiliitis on Imaging Plus ≥1 SpA Feature	OR	HLA-B27 Plus ≥2 Other SpA Features
Sacroiliitis on imaging • Active (acute) inflammation on MRI highly suggestive of SpA-associated sacroiliitis[b] and/or • Definite radiographic sacroiliitis according to modified New York criteria[c]		SpA features • Inflammatory back pain[d] • Arthritis[e] • Enthesitis (heel)[f] • Anterior uveitis[g] • Dactylitis[e] • Psoriasis[e] • Crohn's disease or ulcerative colitis[e] • Good response to NSAIDs[h] • Family history of SpA[i] • HLA-B27 • Elevated CRP[j]

[a]Sensitivity 83%, specificity 84%. The imaging arm (sacroiliitis) alone has a sensitivity of 66% and a specificity of 97%. [b]Bone marrow edema and/or osteitis on short tau inversion recovery (STIR) or gadolinium-enhanced T1 image. [c]Bilateral grade ≥2 or unilateral grade 3 or 4. [d]See text for criteria. [e]Past or present, diagnosed by a physician. [f]Past or present pain or tenderness on examination at calcaneus insertion of Achilles tendon or plantar fascia. [g]Past or present, confirmed by an ophthalmologist. [h]Substantial relief of back pain at 24–48 h after a full dose of NSAID. [i]First- or second-degree relatives with ankylosing spondylitis (AS), psoriasis, uveitis, reactive arthritis (ReA), or inflammatory bowel disease (IBD). [j]After exclusion of other causes of elevated CRP.

Abbreviations: ASAS, Assessment of Spondyloarthritis International Society; CRP, C-reactive protein; MRI, magnetic resonance imaging; NSAIDs, nonsteroidal anti-inflammatory drugs; SpA, spondyloarthritis.

Source: From M Rudwaleit et al: Ann Rheum Dis 68:777, 2009. Copyright 2009, with permission from BMJ Publishing Group Ltd.

EPIDEMIOLOGY

AS shows a striking correlation with the histocompatibility antigen HLA-B27 and occurs worldwide roughly in proportion to the prevalence of B27 (Chap. 343). In North American whites, the prevalence of B27 is 7%, whereas it is 75–90% in patients with AS.

In population surveys, AS is present in 1–6% of adults inheriting B27, whereas the prevalence is 10–30% among B27+ adult first-degree relatives of AS probands. Concordance rate in identical twins is about 65%. Susceptibility to AS is determined largely by genetic factors, with B27 estimated to comprise about 20% of the genetic component. Genome-wide single-nucleotide polymorphism (SNP) analysis has identified over 100 additional non-HLA susceptibility alleles.

Patients with ax-SpA that do not have radiologic criteria for AS (see below) are said to have *non-radiographic axial SpA* (nr-ax-SpA). The prevalence of HLA-B27 in these patients is similar to that in AS; however, the proportion of females is much higher (>50% vs ~30%). No information is yet available about other susceptibility loci in nr-ax-SpA.

PATHOLOGY

Sacroiliitis is often an early manifestation of AS and nr-ax-SpA. Knowledge of the pathology comes from both biopsy and autopsy studies that cover a range of disease durations, all from patients with AS. Synovitis and myxoid marrow represent the earliest changes, followed by pannus and subchondral granulation tissue. Marrow edema, enthesitis, and chondroid differentiation are also found. Macrophages, T cells, plasma cells, and osteoclasts are prevalent. If the process continues to progress, eventually the eroded joint margins are gradually replaced by fibrocartilage regeneration and then by ossification.

In the spine, surgically resected or autopsy specimens show inflammatory granulation tissue in the paravertebral connective tissue at the junction of annulus fibrosus and vertebral bone, and in some cases along the entire outer annulus. The outer annular fibers are eroded and eventually replaced by bone, forming the beginning of a syndesmophyte, which then grows by continued endochondral ossification, ultimately bridging the adjacent vertebral bodies. Ascending progression of this process can lead to the "bamboo spine." Other lesions in the spine include diffuse osteoporosis (loss of trabecular bone despite accretion of periosteal bone), erosion of vertebral bodies at the disk margin, and inflammation and destruction of the disk-bone border. Inflammatory arthritis of the apophyseal (facet) joints is common, with synovitis, inflammation at the bony attachment of the joint capsule, and subchondral bone marrow granulation tissue. Erosion of joint cartilage by pannus is often followed by bony ankylosis. This may precede formation of syndesmophytes bridging the adjacent disks. Bone mineral density is diminished in the spine and proximal femur early in the disease course.

Peripheral synovitis in AS shows marked vascularity, evident as tortuous macrovascularity seen during arthroscopy. Lining layer hyperplasia, lymphoid infiltration, and pannus formation are also found. Central cartilaginous erosions caused by proliferation of subchondral granulation tissue are common. The characteristics of peripheral arthritis in AS and other forms of SpA are similar, and distinct from those of RA.

Inflammation in the fibrocartilaginous *enthesis*, the region where a tendon, ligament, or joint capsule attaches to bone, is a characteristic lesion in AS and other SpAs, both at axial and peripheral sites. Enthesitis is associated with prominent edema of the adjacent bone marrow and is often characterized by erosive lesions that eventually undergo ossification.

Subclinical intestinal inflammation has been found in the colon or distal ileum in a majority of patients with SpA. The histology is described below under "Enteropathic Arthritis."

PATHOGENESIS

The pathogenesis of AS is immune-mediated, but there is little direct evidence for antigen-specific autoimmunity, and there is increasing evidence to suggest more of an autoinflammatory pathogenesis. Uncertainty remains regarding the primary site of disease initiation. The dramatic response of the disease to therapeutic blockade of tumor necrosis factor α (TNF-α) or IL-17A indicates that these cytokines play a central role in the immunopathogenesis of AS. Other genes related to TNF pathways show association with AS, including *TNFRSF1A*, *LTBR*, and *TBKBP1*. At least five genes in IL-23/IL-17 pathway show association with AS, including *IL23R*, *PTER4*, *IL12B*, *CARD9*, and *TYK2*. All of these genes are also associated with inflammatory bowel disease (IBD), and three of them are associated with psoriasis. Serum levels of IL-23 and IL-17 are elevated in AS patients. In mice, a population of thymus-dependent CD3+CD4–CD8–T cells expressing γ/δ T cell receptors and IL-23 receptors has been found to reside at entheses, in the aortic root, and near the ciliary body in the eye. These cells express abundant IL-17 and IL-22 upon exposure to systemic IL-23. This finding suggests that site-specific innate immune cells play a critical role in the anatomic specificity of these lesions. In other murine studies, mechanical strain was shown to induce inflammation and new bone formation at enthesial sites.

Mast cells and, to a lesser extent, neutrophils appear to be the major IL-17-producing cells in peripheral arthritis, whereas neutrophils producing IL-17 are prominent in apophyseal joints. High levels of circulating γδ T cells expressing IL-23 receptors and producing IL-17 have been found in AS patients.

Other associated genes encode other cytokines or cytokine receptors (*IL6R*, *IL1R1*, *IL1R2*, *IL7R*, *IL27*), transcription factors involved in the differentiation of immune cells (*RUNX3*, *EOMES*, *BACH2*, *NKX2-3*, *TBX21*), or other molecules involved in activation or regulation of immune or inflammatory responses (*FCGR2A*, *ZMIZ1*, *NOS2*, *ICOSLG*).

The inflamed sacroiliac joint is infiltrated with CD4+ and CD8+ T cells and macrophages and shows high levels of TNF-α, particularly early in the disease. Abundant transforming growth factor β (TGF-β) is found in more advanced lesions. Peripheral synovitis in AS and the other spondyloarthritides is characterized by neutrophils, macrophages expressing CD68 and CD163, CD4+ and CD8+ T cells, and B cells. There is prominent staining for intercellular adhesion molecule 1 (ICAM-1), vascular cell adhesion molecule 1 (VCAM-1), matrix metalloproteinase 3 (MMP-3), and myeloid-related proteins 8 and 14 (MRP-8 and MRP-14). Unlike RA synovium, citrullinated proteins and cartilage gp39 peptide–major histocompatibility complexes (MHCs) are absent.

No specific event or exogenous agent that triggers the onset of disease has been identified, although overlapping features with ReA and IBD and the involvement of the IL-23/IL-17 pathway suggest that enteric bacteria may play a role, and microdamage from mechanical stress at enthesial sites has also been implicated.

It is firmly established that HLA-B27 plays a direct role in AS pathogenesis, but its precise molecular role remains unresolved. Rats transgenic for HLA-B27 develop arthritis and spondylitis, and this is unaffected by the absence of CD8. It thus appears that classical peptide antigen presentation to CD8+ T cells may not be the primary disease mechanism. However, the association of AS with ERAP1, which strongly influences the MHC class I peptide repertoire, suggests that peptide binding to B27 is nonetheless important. The B27 heavy chain has an unusual tendency

to misfold, a process that can be proinflammatory. Genetic and functional studies in humans have suggested a role for natural killer (NK) cells in AS, possibly through interaction with B27 heavy chain homodimers. SpA-prone B27 rats show defective dendritic cell function and share with AS patients a characteristic “reverse interferon” gene expression signature in antigen-presenting cells.

New bone formation in AS appears to be largely based on enchondral bone formation and occurs only in the periosteal compartment. It correlates with lack of regulation of the Wnt signaling pathway, which controls the differentiation of mesenchymal cells into osteophytes, by the inhibitors DKK-1 and sclerostin. Indirect evidence and data from animal models also implicate bone morphogenic proteins, hedgehog proteins, and prostaglandin E_2. There is controversy as to whether vertebral new bone formation in AS is a sequela of inflammation or whether it arises independently of inflammation. The second hypothesis is based on the observation that syndesmophyte formation is not suppressed by anti-TNF-α therapy that potently suppresses inflammation. TNF-α is also a known inducer of DKK-1, which inhibits bone formation. Magnetic resonance imaging (MRI) studies suggest that vertebral inflammatory lesions that undergo metaplasia to fat (increased T1-weighted signal) are a preferential site of subsequent syndesmophyte formation despite anti-TNF-α therapy, whereas early acute inflammatory lesions resolve. The rate of syndesmophyte formation appears to decrease after >4 years of anti-TNF-α therapy.

CLINICAL MANIFESTATIONS

In patients eventually diagnosed with AS, symptoms are usually first noticed in late adolescence or early adulthood, at a median age in the mid-twenties. In 5% of patients, symptoms begin after age 40. The initial symptom is usually dull pain, insidious in onset, felt deep in the lower lumbar or gluteal region, accompanied by low-back morning stiffness of up to a few hours’ duration that improves with activity and returns following inactivity. Within a few months, the pain has usually become persistent and bilateral. Nocturnal exacerbation of pain often forces the patient to rise and move around.

In some patients, bony tenderness (presumably reflecting enthesitis or osteitis) may accompany back pain or stiffness, whereas in others it may be the predominant complaint. Common sites include the costosternal junctions, spinous processes, iliac crests, greater trochanters, ischial tuberosities, tibial tubercles, and heels. Hip and shoulder (“root” joint) arthritis is considered part of axial disease. Hip arthritis occurs in 25–35% of patients. Shoulder involvement may be at least as common, but is usually less symptomatic. Severe isolated hip arthritis or bony chest pain may be the presenting complaint, and symptomatic hip disease can dominate the clinical picture. Arthritis of peripheral joints other than the hips and shoulders, usually asymmetric, may occur at any point in the disease course. Neck pain and stiffness from involvement of the cervical spine are usually relatively late manifestations, but are occasionally dominant symptoms.

In juvenile onset spondyloarthritis, peripheral arthritis and enthesitis predominate, with axial symptoms supervening in late adolescence.

Initially, axial physical findings mirror the inflammatory process. The most specific findings involve loss of spinal mobility, with limitation of anterior and lateral flexion and extension of the lumbar spine and of chest expansion. Limitation of motion is usually out of proportion to the degree of bony ankylosis and is thought to possibly reflect muscle spasm secondary to pain and inflammation. Pain in the sacroiliac joints may be elicited either with direct pressure or with stress on the joints. In addition, there is commonly tenderness upon palpation of the posterior spinous processes and other sites of symptomatic bony tenderness.

The modified Schober test is a useful measure of lumbar spine flexion. The patient stands erect, with heels together, and marks are made on the spine at the lumbosacral junction (identified by a horizontal line between the posterosuperior iliac spines) and 10 cm above. The patient then bends forward maximally with knees fully extended, and the distance between the two marks is measured. This distance increases by ≥5 cm in the case of normal mobility and by <4 cm in the case of decreased mobility. Chest expansion is measured as the difference between maximal inspiration and maximal forced expiration in the fourth intercostal space in males or just below the breasts in females, with the patient’s hands resting on or just behind the head. Normal chest expansion is ≥5 cm. Lateral bending measures the distance the patient’s middle finger travels down the leg with maximal lateral bending. Normal is >10 cm.

Limitation or pain with motion of the hips or shoulders is usually present if these joints are involved. It should be emphasized that in early mild, or atypical cases, the symptoms and/or physical findings may be subtle and/or nonspecific.

The course of ax-SpA is extremely variable, ranging from the individual with mild stiffness and normal radiographs to the patient with a totally fused spine and severe bilateral hip arthritis, accompanied by severe peripheral arthritis and extraarticular manifestations. Most of the available data on natural history are from observations of patients with AS, although the prevalence of peripheral arthritis, enthesitis, psoriasis, and IBD appears to be similar in nr-ax-SpA and AS. Pain tends to be persistent early in the disease and intermittent later, with alternating exacerbations and quiescent periods. In a typical severe untreated case with progression of the spondylitis to syndesmophyte formation, the patient’s posture undergoes characteristic changes, with obliterated lumbar lordosis, buttock atrophy, and accentuated thoracic kyphosis. There may be a forward stoop of the neck or flexion contractures at the hips, compensated by flexion at the knees. Disease progression can be estimated clinically from

loss of height, limitation of chest expansion and spinal flexion, and occiput-to-wall distance. Occasional individuals are encountered with advanced deformities who report having never had significant symptoms.

The factors most predictive of radiographic progression (see below) are the presence of existing syndesmophytes, high inflammatory markers, and smoking. In some but not all studies, onset of AS in adolescence and early hip involvement correlate with a worse prognosis. In women, AS tends to progress less frequently to total spinal ankylosis, although there may be an increased prevalence of isolated cervical ankylosis and peripheral arthritis. Peripheral arthritis (distal to hips and shoulders) occurs in up to 30% of patients in the United States, Canada, and Western Europe, usually as a late manifestation. In Eastern Europe, Latin America, and Asia, the prevalence is over half, with onset more commonly early in the disease course. Pregnancy has no consistent effect on AS, with symptoms improving, remaining the same, or deteriorating in one-third of pregnant patients, respectively.

The most serious complication of the spinal disease is spinal fracture, which can occur with even minor trauma to the rigid, osteoporotic spine. The lower cervical spine is most commonly involved. These fractures are often displaced, causing spinal cord injury. A recent survey suggested a >10% lifetime risk of fracture. Occasionally, fracture through a diskovertebral junction and adjacent neural arch, termed *pseudoarthrosis*, most common in the thoracolumbar spine, can be an unrecognized source of persistent localized pain and/or neurologic dysfunction. Wedging of thoracic vertebrae is common and correlates with accentuated kyphosis.

The most common extraarticular manifestation is acute anterior uveitis, which occurs in up to 40% of patients and can antedate the spondylitis. Attacks are typically unilateral, causing pain, photophobia, and increased lacrimation. These tend to recur, often in the opposite eye. Cataracts and secondary glaucoma may ensue. Up to 60% of patients with AS have inflammation in the colon or ileum. This is usually asymptomatic, but frank IBD occurs in 5–10% of patients with AS (see "Enteropathic Arthritis," below). About 10% of patients meeting criteria for AS have psoriasis (see "Psoriatic Arthritis," below). Occasional patients are seen with AS in association with skin manifestations seen in SAPHO syndrome (see below), such as acne fulminans or hidradenitis suppurativa. There is an apparently increased risk of ischemic heart disease. Aortic insufficiency occurs in a small percentage of patients. Third-degree heart block may occur alone or together with aortic insufficiency, and association with lesser degrees of heart block has been described. Cauda equina syndrome and upper pulmonary lobe fibrosis are rare late complications. Prostatitis has been reported to have an increased prevalence. Amyloidosis is rare **(Chap. 108)**.

Several validated measures of disease activity and functional outcome are in widespread use in the study and management of AS, particularly the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) and the Ankylosing Spondylitis Disease Activity Score (ASDAS), both measures of disease activity; the Bath Ankylosing Spondylitis Functional Index (BASFI), a measure of limitation in activities of daily living; and several measures of radiographic changes. The Harris hip score, although not specific for AS, can be helpful. Despite persistence of the disease, most patients remain gainfully employed. Some but not all studies of survival in AS have suggested that AS shortens life span, compared with the general population. Mortality attributable to AS is largely the result of spinal trauma, aortic insufficiency, respiratory failure, amyloid nephropathy, or complications of therapy such as upper gastrointestinal hemorrhage. The impact of anti-TNF therapy on outcome and mortality is not yet known, except for significantly improved work productivity.

LABORATORY FINDINGS

No laboratory test is diagnostic of AS. In most ethnic groups, HLA-B27 is present in 75–90% of patients. Erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) are often, but not always, elevated. Mild anemia may be present. Patients with severe disease may show an elevated alkaline phosphatase level. Elevated serum IgA levels are common. Rheumatoid factor, anti-cyclic citrullinated peptide (CCP), and antinuclear antibodies (ANAs) are largely absent unless caused by a coexistent disease, although ANAs may appear with anti-TNF therapy. Circulating levels of CD8+ T cells tend to be low, and serum matrix metalloproteinase 3 levels correlate with disease activity. Synovial fluid from peripheral joints in AS is nonspecifically inflammatory. Restricted chest wall motion causes decreased vital capacity, but ventilatory function is usually well maintained.

RADIOGRAPHIC FINDINGS

By definition, the diagnosis of AS is associated with advanced radiographically demonstrable sacroiliitis, usually symmetric. The earliest changes by standard radiography are blurring of the cortical margins of the subchondral bone, followed by erosions and sclerosis. Progression of the erosions leads to "pseudowidening" of the joint space; as fibrous and then bony ankylosis supervene, the joints may become obliterated.

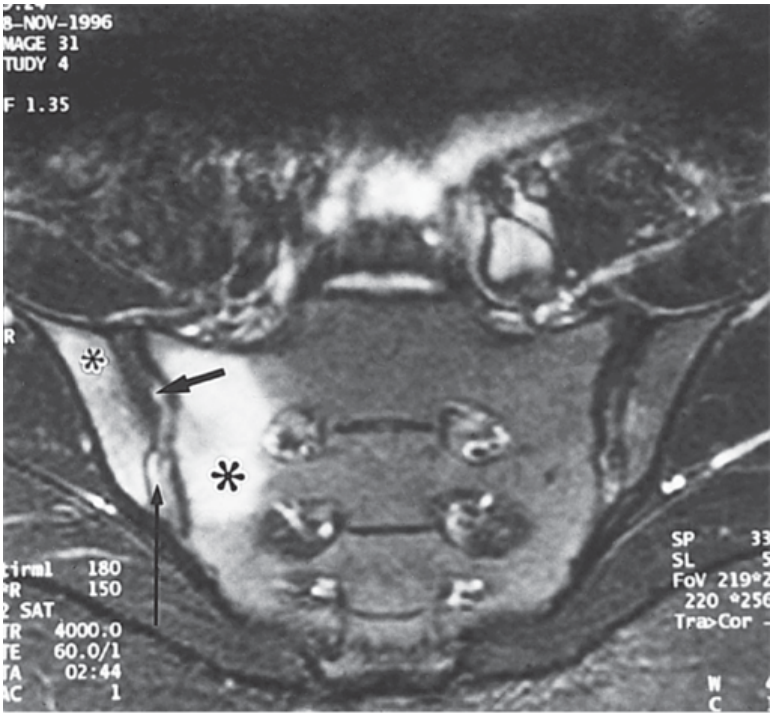
In the lumbar spine, progression of the disease can lead to loss of lordosis, and osteitis of the anterior corners of the vertebral bodies with subsequent erosion, leading to "squaring" or even "barreling" of one or more vertebral bodies. Progressive ossification leads to eventual formation of marginal syndesmophytes, visible on plain films as bony bridges connecting successive vertebral bodies anteriorly and laterally.

A recent study showed that only a quarter of patients meeting criteria for nr-ax-SpA developed radiographic sacroiliitis within 15 years. MRI is thus much more useful for the timely diagnosis of ax-SpA. Active sacroiliitis is best visualized by dynamic MRI on tilted coronal slices with fat

saturation, either T2-weighed turbo spin-echo sequence or short tau inversion recovery (STIR) with high resolution, or T1-weighted images with contrast enhancement. These techniques identify early intraarticular inflammation, cartilage changes, and underlying bone marrow edema in sacroiliitis (Fig. 355-1). The presence of erosions enhances specificity and is best detected on conventional T1-weighted images. These protocols are also sensitive for evaluation of acute and chronic spinal changes. MRI protocols routinely used to evaluate low back pain have low sensitivity for detecting inflammation and often give false-negative results in ax-SpA. Optimal results require a high index of suspicion, an appropriate protocol, an experienced radiologist, and close communication between radiologist and clinician.

FIGURE 355-1

Early sacroiliitis in a patient with ankylosing spondylitis, indicated by prominent edema in the juxtaarticular bone marrow (*asterisks*), synovium and joint capsule (*thin arrow*), and interosseous ligaments (*thick arrow*) on a short tau inversion recovery (STIR) magnetic resonance image. *(From M Bollow et al: Zeitschrift für Rheumatologie 58:61, 1999. Reproduced with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Reduced bone mineral density can be detected by dual-energy x-ray absorptiometry of the femoral neck and the lumbar spine. Use of a lateral projection of the L3 vertebral body can prevent falsely elevated readings related to spinal ossification.

DIAGNOSIS

It is important to recognize ax-SpA before the development of irreversible deformity. This goal is challenging for several reasons: (1) only a minority of back pain patients have ax-SpA; (2) an early diagnosis often relies on clinical grounds and/or an appropriate MRI protocol requiring considerable expertise; (3) young individuals with symptoms of ax-SpA often do not seek medical care; (4) reliance on definite radiographic sacroiliitis causes early or mild cases to be missed. The classification criteria for axial SpA proposed by the Assessment of Spondyloarthritis International Society (ASAS) are shown in Table 355-1. They are applicable to individuals with ≥3 months of back pain with age of onset <45 years. Active inflammation of the sacroiliac joints as determined by dynamic MRI is considered equivalent to definite radiographic sacroiliitis (see below).

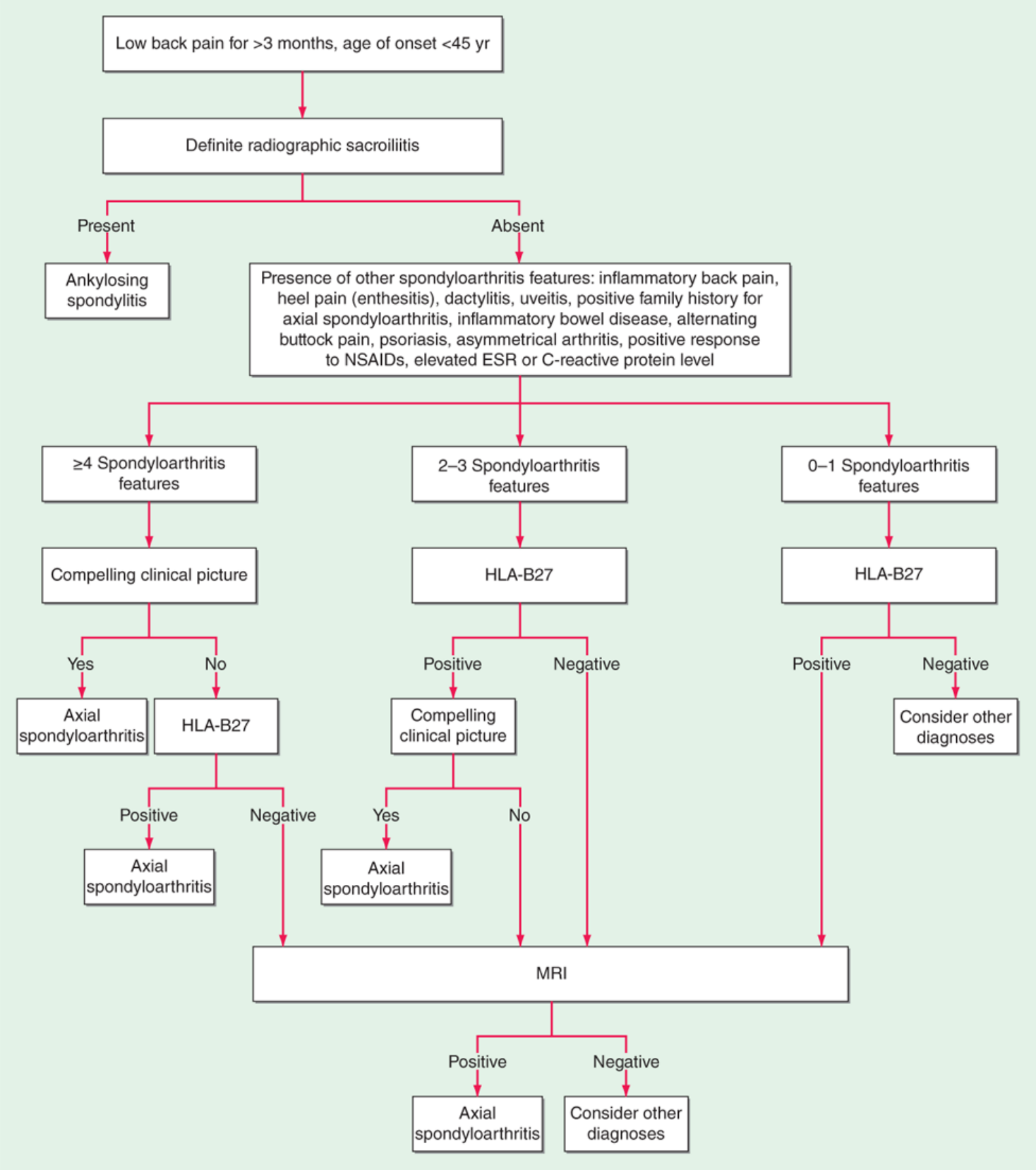
ax-SpA must be differentiated from numerous other causes of low-back pain, some substantially more common than ax-SpA. Increased specificity is obtained when the nature and pattern of the pain and the age of the patient are considered. The most typical symptom is inflammatory back pain (IBP), present in 70–80% of patients with ax-SpA, but relatively uncommon otherwise. In chronic (≥3 months) back pain, IBP has the following characteristic features: (1) age of onset <40 years; (2) insidious onset; (3) improvement with exercise; (4) no improvement with rest; and (5) pain at night with improvement upon getting up; (6) morning stiffness >30 min; (7) awakening from back pain during only the second half of the night; and (8) alternating buttock pain. The presence of two or more of these features should arouse suspicion for IBP, and four or more can be considered diagnostic. The most common causes of back pain other than SpA are primarily mechanical or degenerative rather than primarily inflammatory and tend not to show clustering of these features.

Less-common metabolic, infectious, and malignant causes of back pain must also be differentiated from AS, including infectious spondylitis, spondylodiskitis, and sacroiliitis, and primary or metastatic tumor. Ochronosis can produce a phenotype similar to AS. Calcification and ossification of paraspinous ligaments occur in *diffuse idiopathic skeletal hyperostosis* (DISH), which occurs in the middle-aged and elderly and is usually not symptomatic. Ligamentous calcification gives the appearance of "flowing wax" on the anterior bodies of the vertebrae. Intervertebral disk spaces are preserved, and sacroiliac and apophyseal joints appear normal, helping to differentiate DISH from spondylosis and from AS, respectively. Both primary and secondary hyperparathyroidism can cause subchondral bone resorption around the SI joints, with bilateral widened and ill-defined joints on radiographs, but without joint space narrowing.

An algorithm for making or excluding the diagnosis of ax-SpA in patients with chronic back pain starting before age 45 is shown in **Fig. 355-2**.

FIGURE 355-2

Algorithm for the diagnosis or exclusion of axial spondyloarthritis. The algorithm is designed for use in patients with at least a 3-month history of unexplained chronic low back pain that started before the age of 45 years. Definite radiographic sacroiliitis is based on the modified New York criteria for ankylosing spondylitis (*van der Linden S et al: Arthritis Rheum 27:361, 1984*). The algorithm is adapted from *van den Berg R et al: Ann Rheum Dis 72:1646, 2013*. The determination of whether or not a clinical picture is compelling is based on the relative weights of the spondyloarthritis features (*Feldtkeller E et al: Rheumatology [Oxford] 52:1648, 2013*) and on clinical judgment. The list of clinical features includes features of both axial and peripheral spondyloarthritis. *(From Taurog JD et al: N Engl J Med 374:2563, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Axial Spondyloarthritis

All management of ax-SpA should include an exercise program to maintain posture and range of motion. Nonsteroidal anti-inflammatory drugs (NSAIDs) are the first line of pharmacologic therapy. They reduce pain and tenderness and increase mobility in many patients. Continuous high-dose NSAID therapy may slow radiographic progression, particularly in patients who are at higher risk for progression. However, many patients have continued symptoms despite NSAID therapy and are likely to benefit from anti-TNF-α therapy. Patients with AS treated with infliximab (chimeric human/mouse anti-TNF-α monoclonal antibody), etanercept (soluble p75 TNF-α receptor–IgG fusion protein), adalimumab, or golimumab (human anti-TNF-α monoclonal antibodies, or certolizumab pegol [humanized mouse anti-TNF-α monoclonal antibody]) have shown rapid, profound, and sustained reductions in all clinical and laboratory measures of disease activity. In a good response, there is significant improvement in both objective and subjective indicators of disease activity and function, including morning stiffness, pain, spinal mobility, peripheral joint swelling, CRP, ESR, and bone mineral density. MRI studies indicate substantial resolution of bone marrow edema, enthesitis, and joint effusions in the sacroiliac joints, spine, and peripheral joints. Similar results have been obtained in large randomized controlled trials of all five agents and many open-label studies. About one-half of the patients achieve a ≥50% reduction in the BASDAI. The response tends to persist over time, and partial or full remissions are common. Predictors of the best responses include younger age, shorter disease duration, higher baseline inflammatory markers, and lower baseline functional disability. Nonetheless, some patients with long-standing disease and even spinal ankylosis can obtain significant benefit. Syndesmophyte formation may continue despite the therapy, but this may apply mainly during the early years of therapy. Although less well studied, the response of patients with nr-ax-SpA to anti-TNF therapy is generally similar to that of patients with AS.

Typically, infliximab is given intravenously, 5 mg/kg body weight, and then repeated 2 weeks later, again 6 weeks later, and then at 6- to 8-week intervals. Etanercept is given by subcutaneous injection, 50 mg once weekly. Adalimumab is given by subcutaneous injection, 40 mg biweekly. Golimumab is given by subcutaneous injection, 50 or 100 mg every 4 weeks. Certolizumab pegol is given by subcutaneous injection, 400 mg every 4 weeks. Dosage adjustments can be considered in selected cases.

These potent immunosuppressive agents are relatively safe, but patients are at increased risk for serious infections, including disseminated tuberculosis. Hypersensitivity infusion or injection site reactions are not uncommon. Cases of anti-TNF-induced psoriasis have been increasingly recognized. Rare cases of systemic lupus erythematosus (SLE)–related disease have been reported, as have hematologic disorders such as pancytopenia, demyelinating disorders, exacerbation of congestive heart failure, and severe liver disease. The overall incidence of malignancy does not appear to be increased in AS patients treated with anti-TNF therapy, but isolated cases of hematologic malignancy have occurred shortly after the start of treatment.

Because of the expense, potentially serious side effects, and unknown long-term effects of these agents, their use should be restricted to patients with a definite diagnosis and active disease (BASDAI ≥4 out of 10 and expert opinion) that is inadequately responsive to therapy with at least two different NSAIDs. Before initiation of anti-TNF therapy, all patients should be tested for tuberculin (TB) reactivity, and reactors (≥5 mm on PPD testing or a positive quantiferon test) should be treated with anti-TB agents. Contraindications include active infection or high risk of infection; malignancy or premalignancy; and history of SLE, multiple sclerosis, or related autoimmunity. Pregnancy and breast-feeding are no longer considered contraindications if appropriate precautions are taken. Infants exposed to anti-TNF in utero should not be given live vaccines before age 6 months. Continuation beyond 12 weeks of therapy requires either a 50% reduction in BASDAI or absolute reduction of ≥2 out of 10, and favorable expert opinion. Switching to a second anti-TNF agent may be effective, especially if there was a response to the first that was lost rather than primary failure.

Secukinumab, a human monoclonal antibody to IL-17A, shows dramatic efficacy in AS, similar to that seen with TNF inhibitors, and is effective in some patients who have failed or not tolerated anti-TNF therapy. The recommended dose is 150 mg subcutaneously weekly for 4 weeks, and then at 4 week intervals. Precautions regarding infection are similar to those for anti-TNF agents. An additional concern is potential exacerbation of underlying IBD, whether previously recognized or not, and careful monitoring is advised.

Sulfasalazine, in doses of 2–3 g/d, has modest benefit, primarily for peripheral arthritis. Methotrexate, although widely used, has not been shown to be of benefit in AS, nor has any therapeutic role for gold or oral glucocorticoids been documented. Potential benefit in AS has been reported for thalidomide, 200 mg/d, perhaps acting through inhibition of TNF-α. The oral Jak inhibitor, tofacitinib, showed efficacy in AS, with reduction of inflammation evident on MRI, in a 16 week phase 2 study.

The most common indication for surgery in patients with AS is severe hip joint arthritis, the pain and stiffness of which are usually dramatically relieved by total hip arthroplasty. Rare patients may benefit from surgical correction of extreme flexion deformities of the spine or of atlantoaxial subluxation.

Attacks of uveitis are usually managed effectively with local glucocorticoids and mydriatic agents, although systemic glucocorticoids, immunosuppressive drugs, or anti-TNF therapy may be required. TNF inhibitors reduce the frequency of attacks of uveitis in patients with ax-SpA,

and adalimumab has recently been approved by the FDA for treating uveitis. Cases of new or recurrent uveitis after use of a TNF inhibitor have been observed, especially with etanercept. Anti-IL-17 does not appear as effective for uveitis as anti-TNF therapy.

Management of axial osteoporosis is at present similar to that used for primary osteoporosis, since data specific for AS are not available.

REACTIVE ARTHRITIS

ReA refers to acute nonpurulent arthritis complicating an infection elsewhere in the body. In recent years, the term has been used primarily to refer to SpA following enteric or urogenital infections.

Other forms of reactive and infection-related arthritis not associated with B27 and showing a spectrum of clinical features different from SpA, such as Lyme disease, rheumatic fever, and poststreptococcal reactive arthritis, are discussed in Chaps. 181 and 352.

HISTORIC BACKGROUND

The association of acute arthritis with episodes of diarrhea or urethritis has been recognized for centuries. A large number of cases during World Wars I and II focused attention on the triad of arthritis, urethritis, and conjunctivitis, often with additional mucocutaneous lesions, which became widely known by eponyms that are now of historic interest only.

The identification of bacterial species triggering the clinical syndrome and the finding of an association with HLA-B27 led to the unifying concept of ReA as a clinical syndrome triggered by specific etiologic agents in a genetically susceptible host. A similar spectrum of clinical manifestations can be triggered by enteric infection with any of several *Shigella*, *Salmonella*, *Yersinia*, and *Campylobacter* species; by genital infection with *Chlamydia trachomatis*; by many other agents as well, apparently in some cases via nasopharyngeal infection with *Chlamydia pneumoniae* or other agents. The “classic triad” represents a small part of the spectrum of the clinical manifestations of ReA and is present only in a small minority of patients. For the purposes of this chapter, the use of the term *ReA* will be restricted to those cases of SpA in which there is at least presumptive evidence for a related symptomatic antecedent infection.

EPIDEMIOLOGY

In early reports, 60–85% of patients who developed ReA triggered by *Shigella*, *Yersinia*, or *Chlamydia* were HLA-B27-positive, but the true figure is probably lower. Other studies demonstrated a lower prevalence of B27 in ReA triggered by *Salmonella*, and little or no association in *Campylobacter*-induced ReA. More recent community-based or common-source epidemic studies showed a prevalence of B27 in ReA <50%. The most common age range is 18–40 years, but ReA can occur rarely in children and occasionally in older adults.

The attack rate of postenteric ReA generally ranges from 1 to about 30% depending on the study and causative organism, whereas the attack rate of postchlamydial ReA is about 4–8%. The gender ratio in ReA following enteric infection is nearly 1:1, whereas venereally acquired ReA occurs mainly in men. The overall prevalence and incidence of ReA are difficult to assess because of the lack of validated diagnostic criteria, variable prevalence and arthritogenic potential of the triggering microbes, and inconstant genetic susceptibility factors in different populations. In Scandinavia, an annual incidence of 10–28:100,000 has been reported. The spondyloarthritides were formerly almost unknown in sub-Saharan Africa. However, ReA and other peripheral SpAs have become common in black Africans in the wake of the AIDS epidemic, without association to B27, which is very rare in these populations. ReA is often the first manifestation of HIV infection and often remits with disease progression. In contrast, Western white patients with HIV and SpA are usually B27-positive, and the arthritis flares as AIDS advances.

PATHOLOGY

Synovial histology is similar to that of other SpAs. Enthesitis shows increased vascularity and macrophage infiltration of fibrocartilage. Microscopic histopathologic evidence of inflammation mimicking IBD has routinely been demonstrated in the colon and ileum of patients with postenteric ReA and less commonly in postvenereal ReA. The skin lesions of keratoderma blennorrhagica, associated mainly with venereally acquired ReA, are histologically indistinguishable from pustular psoriasis.

ETIOLOGY AND PATHOGENESIS

Definite bacterial triggers of ReA include several *Salmonella* spp., *Shigella* spp., *Yersinia enterocolitica*, *Yersinia pseudotuberculosis*, *Campylobacter jejuni*, and *Chlamydia trachomatis*. These triggering microbes are gram-negative bacteria with a lipopolysaccharide (LPS) component to their cell walls. All *Shigella* species have been implicated in cases of ReA, with *S. flexneri* and *S. sonnei* being the most common. After *Salmonella* infection, individuals of Caucasian descent may be more likely than those of Asian descent to develop ReA. Children may be less susceptible to ReA caused by *Salmonella* and *Campylobacter*. *Yersinia* species in Europe and Scandinavia may have greater arthritogenic

potential than in other parts of the world, and *C. trachomatis* appears to be a common cause worldwide. The ocular serovars of *C. trachomatis* appear to be particularly, perhaps uniquely, arthritogenic.

There is also evidence implicating *Clostridium difficile*, *Campylobacter coli*, certain toxigenic *Escherichia coli*, and possibly *Ureaplasma urealyticum* and *Mycoplasma genitalium* as potential triggers of ReA. *Chlamydia pneumoniae* is also a trigger of ReA, but far less commonly so than *C. trachomatis*. There have been numerous isolated reports of acute arthritis preceded by many other bacterial, viral, or parasitic infections, and arthritis following intravesicular bacillus Calmette-Guérin (BCG) treatment for bladder cancer is well documented.

It is not known whether there is a common pathogenic mechanism for triggering of ReA that is shared by all of these microorganisms, nor has the mechanism been elucidated in the case of any one of the known triggers. Many of the well established triggers share a capacity to attack mucosal surfaces, to invade host cells, and to survive intracellularly. Antigens from *Chlamydia*, *Yersinia*, *Salmonella*, and *Shigella* have been shown to be present in the synovium and/or synovial fluid leukocytes of patients with ReA for long periods following the acute attack. In ReA triggered by *Y. enterocolitica*, bacterial LPS and heat-shock protein antigens have been found in peripheral blood cells years after the triggering infection. *Yersinia* DNA and *C. trachomatis* DNA and RNA have been detected in synovial tissue from ReA patients, suggesting the presence of viable organisms despite uniform failure to culture the organism from these specimens. In *C. trachomatis*–induced ReA specifically, the bacterial load in synovial tissue of patients with remitting disease is lower than that of active disease, but mRNAs encoding proinflammatory proteins are equal to or higher than those of active disease. The specificity of these findings is unclear, however, since chromosomal bacterial DNA and 16S rRNA from a wide variety of bacteria have also been found in synovium in other rheumatic diseases, albeit less frequently.

Synovial T cells specific for antigens of the inciting organism were reported in the 1980s and early 1990s. More recent work has documented high levels of IL-17 in ReA synovial fluid, but the source has not been identified. HLA-B27 seems to be associated with more severe and chronic ReA, but its pathogenic role remains to be determined. HLA-B27 significantly prolongs the intracellular survival of *Y. enterocolitica* and *Salmonella enteritidis* in human and mouse cell lines. Prolonged intracellular bacterial survival may permit trafficking of infected leukocytes from the site of primary infection to joints, where an innate and/or adaptive immune response to persistent bacterial antigens may then promote arthritis.

CLINICAL FEATURES

The clinical manifestations of ReA range from an isolated, transient monoarthritis or enthesitis to severe multisystem disease. A careful history will often elicit evidence of an antecedent infection 1–4 weeks before onset of symptoms of the reactive disease, particularly in postenteric ReA. However, in a sizable minority, no clinical or laboratory evidence of an antecedent infection can be found, particularly in the case of postchlamydial ReA. In cases of presumed venereally acquired reactive disease, there is often a history of a recent new sexual partner, even without laboratory evidence of infection.

Constitutional symptoms are common, including fatigue, malaise, fever, and weight loss. The musculoskeletal symptoms are usually acute in onset. Arthritis is usually asymmetric and additive, with involvement of new joints occurring over a few days to 1–2 weeks. The joints of the lower extremities, especially the knee, ankle, subtalar, metatarsophalangeal, and toe interphalangeal joints, are most commonly involved, but the wrist and fingers may be involved. The arthritis is usually quite painful, and tense joint effusions are not uncommon, especially in the knee. Dactylitis, or “sausage digit,” a diffuse swelling of a solitary finger or toe, is a distinctive feature of ReA and other peripheral spondyloarthritides but can be seen in polyarticular gout and sarcoidosis. Tendinitis and fasciitis are particularly characteristic lesions, producing pain at multiple insertion sites (entheses), especially the Achilles insertion, the plantar fascia, and sites along the axial skeleton. Back and buttock pain are quite common and may be caused by insertional inflammation, muscle spasm, acute sacroiliitis, or, presumably, arthritis in intervertebral joints.

Urogenital lesions may occur throughout the course of the disease. In males, urethritis may be marked or relatively asymptomatic and may be either an accompaniment of the triggering infection or a result of the reactive phase of the disease; interestingly, it occurs in both postvenereal and postenteric ReA. Prostatitis is also common. Similarly, in females, cervicitis or salpingitis may be caused either by the infectious trigger or by the sterile reactive process.

Ocular disease is common, ranging from transient, asymptomatic conjunctivitis to an aggressive anterior uveitis that occasionally proves refractory to treatment and may result in blindness.

Mucocutaneous lesions are frequent. Oral ulcers tend to be superficial, transient, and often asymptomatic. The characteristic skin lesions, *keratoderma blennorrhagica*, consist of vesicles and/or pustules that become hyperkeratotic, ultimately forming a crust before disappearing. They are most common on the palms and soles but may occur elsewhere as well. In patients with HIV infection, these lesions are often severe and extensive, sometimes dominating the clinical picture **(Chap. 197)**. Lesions on the glans penis, termed *circinate balanitis*, consist of vesicles that quickly rupture to form painless superficial erosions, which in circumcised individuals can form crusts similar to those of keratoderma blennorrhagica. Nail changes are common and consist of onycholysis, distal yellowish discoloration, and/or heaped-up hyperkeratosis.

Less-frequent or rare manifestations of ReA include cardiac conduction defects, aortic insufficiency, central or peripheral nervous system lesions, and pleuropulmonary infiltrates.

Arthritis typically persists for 3–5 months, but more chronic courses do occur. Chronic joint symptoms persist in about 15% of patients and in up to 60% of patients in hospital-based series, but these tend to be less severe than in the acute stage. Recurrences of the acute syndrome may occur. Work disability or forced change in occupation is common in those with persistent joint symptoms. Chronic heel pain is often particularly distressing. Low-back pain, sacroiliitis, and frank AS are also common sequelae. In most studies, HLA-B27–positive patients have shown a worse outcome than B27-negative patients. Patients with *Yersinia*- or *Salmonella*-induced arthritis have less chronic disease than those whose initial episode follows epidemic shigellosis.

LABORATORY AND RADIOGRAPHIC FINDINGS

The ESR and acute-phase reactants are usually elevated during the acute phase of the disease, often markedly so. Mild anemia may be present. Synovial fluid is nonspecifically inflammatory. In most ethnic groups, 30–50% of the patients are B27-positive. The triggering infection usually does not persist at the site of primary mucosal infection through the time of onset of the reactive disease, but it may be possible to culture the organism, for example, in the case of *Yersinia*- or *Chlamydia*-induced disease. Serologic evidence of exposure to one of the causative organisms with elevation of antibodies is nonspecific and of questionable utility. Polymerase chain reaction (PCR) for chlamydial DNA in first-voided urine specimens may have high sensitivity in the acute stage but is less useful with chronic disease.

In early or mild disease, radiographic changes may be absent or confined to juxtaarticular osteoporosis. With long-standing disease, radiographic features share those of PsA; marginal erosions and loss of joint space can be seen in affected joints. Periostitis with reactive new bone formation is characteristic, as in all the SpAs. Spurs at the insertion of the plantar fascia are common.

Sacroiliitis and spondylitis may be seen as late sequelae. Sacroiliitis is more commonly asymmetric than in AS, and spondylitis, rather than ascending symmetrically, can begin anywhere along the lumbar spine. The syndesmophytes are described as nonmarginal; they are coarse, asymmetric, and “comma”-shaped, arising from the middle of a vertebral body, a pattern less commonly seen in primary AS. Progression to spinal fusion is uncommon.

DIAGNOSIS

ReA is a clinical diagnosis with no definitively diagnostic laboratory test or radiographic finding. The diagnosis should be entertained in any patient with an acute inflammatory, asymmetric, additive arthritis or tendinitis. The evaluation should include thorough but tactful questioning regarding possible triggering events. On physical examination, attention must be paid to the distribution of the joint and tendon involvement and to possible sites of extraarticular involvement, including the eyes, mucous membranes, skin, nails, and genitalia. Synovial fluid analysis is usually necessary to exclude septic or crystal-induced arthritis. Culture, serology, or molecular methods may help identify a triggering infection, but they cannot be relied upon.

Although typing for B27 has low negative predictive value in ReA, it may have prognostic significance in terms of severity, chronicity, and the propensity for spondylitis and uveitis. Furthermore, if positive, it can be helpful diagnostically in atypical cases. HIV testing is often indicated and may be necessary in selecting therapy.

Both ReA and disseminated gonococcal disease **(Chap. 151)** can be venereally acquired and associated with urethritis. Unlike ReA, gonococcal arthritis and tenosynovitis tend to involve both upper and lower extremities equally, spare the axial skeleton, and be associated with characteristic vesicular skin lesions. A positive gonococcal culture from the urethra or cervix does not exclude a diagnosis of ReA; however, culturing gonococci from blood, skin lesion, or synovium establishes the diagnosis of disseminated gonococcal disease. PCR assay for *Neisseria gonorrhoeae* and *C. trachomatis* may be helpful. Occasionally, only a therapeutic trial of antibiotics can distinguish the two.

ReA shares many features in common with psoriatic arthropathy. However, PsA is usually gradual in onset; the arthritis tends to affect primarily the upper extremities; and there are usually no associated mouth ulcers, urethritis, or bowel symptoms.

TREATMENT

TREATMENT

Reactive Arthritis

Most patients with ReA benefit to some degree from high-dose NSAIDs, although acute symptoms are rarely completely ameliorated, and some patients fail to respond at all.

Prompt, appropriate antibiotic treatment of acute chlamydial urethritis or enteric infection may prevent the emergence of ReA, but is not universally successful. Data regarding the potential benefit of antibiotic therapy that is initiated after onset of arthritis are conflicting; however, a recent systematic review and meta-analysis of 10 controlled trials suggested no benefit. The only one of these trials to use combination antibiotics showed that a majority of patients with chronic ReA associated with *C. trachomatis* or *C. pneumoniae* benefited significantly from a 6-month course of rifampin 300 mg daily plus azithromycin 500 mg daily for 5 days, then twice weekly, or 6 months of rifampin 300 mg daily plus doxycycline 100 mg twice daily. This study awaits further confirmation.

Multicenter trials have suggested that sulfasalazine, up to 3 g/d in divided doses, may be beneficial to patients with persistent ReA.[1] Patients with persistent disease may respond to azathioprine, 1–2 mg/kg per day, or to methotrexate, up to 20 mg per week; however, these therapeutic regimens have never formally been studied. Although no controlled trials of anti-TNF-α in ReA have been reported, anecdotal evidence supports the use of these agents in severe chronic cases, although lack of response has also been observed.[1]

Tendinitis and other enthesitic lesions may benefit from intralesional glucocorticoids. Uveitis may require aggressive treatment to prevent serious sequelae (see above). Skin lesions ordinarily require only symptomatic topical treatment. In patients with HIV infection and ReA, many of whom have severe skin lesions, the skin lesions in particular respond to antiretroviral therapy. Cardiac complications are managed conventionally; management of neurologic complications is symptomatic.

Comprehensive management includes counseling of patients in the avoidance of sexually transmitted disease and exposure to enteropathogens, as well as appropriate use of physical therapy, vocational counseling, and continued surveillance for long-term complications such as AS. Patients with a history of ReA are at increased risk for recurrent attacks following repeated exposures.

[1]Azathioprine, methotrexate, sulfasalazine, pamidronate, thalidomide, and anti-TNFα agents have not been approved for this purpose by the FDA at the time of publication.

PSORIATIC ARTHRITIS

Psoriatic arthritis refers to an inflammatory musculoskeletal disease that has both autoimmune and autoinflammatory features characteristically occurring in individuals with psoriasis.

HISTORIC BACKGROUND

The association between arthritis and psoriasis was noted in the nineteenth century. In the 1960s, it became clear that unlike RA, the arthritis associated with psoriasis was usually seronegative, often involved the distal interphalangeal (DIP) joints of the fingers and the spine and sacroiliac joints, had distinctive radiographic features, and showed considerable familial aggregation. In the 1970s, PsA was included in the broader category of the spondyloarthritides because of features similar to those of AS and ReA.

EPIDEMIOLOGY

The prevalence of PsA appears to be increasing in parallel with disease awareness. Recent data suggest that up to 30% of patients with psoriasis develop PsA. The duration and severity of psoriasis increase the likelihood of developing PsA. In white populations, psoriasis is estimated to have a prevalence of 1–3%. Psoriasis and PsA are less common in other races in the absence of HIV infection, and the prevalence of PsA in individuals with psoriasis may be less common. First-degree relatives of PsA patients have an elevated risk for psoriasis, for PsA, and for other forms of SpA. Of patients with psoriasis, up to 30% have an affected first-degree relative. In monozygotic twins, the reported concordance for psoriasis varies from 35 to 72%, and for PsA from 10 to 30%. A variety of HLA associations have been found. HLA-Cw0602 is directly associated with psoriasis, particularly familial juvenile-onset (type I) psoriasis. HLA-B27 is associated with psoriatic spondylitis (see below). HLA-DR7, -DQ3, and -B57 are associated with PsA because of linkage disequilibrium with Cw6. A recent study found additive associations of PsA with haplotypes containing HLA-B08, HLA-Cw0602, HLA-B27, -B38, and -B39. A correlation was also found between different haplotype combinations and entheseal, synovial, or axial predominant phenotypes. Genome-wide analyses have identified associations of PsA with polymorphisms in the IL-23 receptor (*IL23R*), molecules involved in nuclear factor κB gene expression (*TNIP1*) and signaling (*TNFAIP3*), and cytokines *TNF*, *IL12A*, and *IL12B*. A specific *IL23R* SNP is associated with PsA distinct from psoriasis without arthritis.

PATHOLOGY

The inflamed synovium in PsA resembles that of RA, although with somewhat less hyperplasia and cellularity than in RA. As noted with AS above, the synovial vascular pattern in PsA is generally greater and more tortuous than in RA, independent of disease duration. Some studies have indicated a higher tendency to synovial fibrosis in PsA. Unlike RA, PsA shows prominent enthesitis, with histology similar to that of the other spondyloarthritides.

PATHOGENESIS

PsA is almost certainly immune-mediated and presumably shares pathogenic mechanisms with psoriasis. PsA synovium is characterized by lining layer hyperplasia; diffuse infiltration with T cells, B cells, macrophages, and NK receptor–expressing cells, with upregulation of leukocyte homing receptors; and neutrophil proliferation with angiogenesis. Clonally expanded T cell subpopulations are frequent and have been demonstrated both in the synovium and the skin. Plasmacytoid dendritic cells are thought to play a key role in psoriasis, and there is some evidence for their participation in PsA. There is abundant synovial overexpression of proinflammatory cytokines, and synovial tissue staining has identified an overexpression of monocyte-derived cytokines, such as myeloid-related protein (S100A8/A9). Interferon γ, TNF-α, and IL-1β, 2, 6, 8, 10, 12, 13, and 15 are found in PsA synovium or synovial fluid. T_H17-derived cytokines are important in PsA, given the genetic association with genes in the IL-23/17 pathway and the therapeutic response to agents targeting this pathway. T_H17 cells have been identified from the dermal extracts of psoriatic lesions and the synovial fluid of PsA patients. The majority of these CD4+ IL-17+ T cells are of memory phenotype (CD4RO[+]CD45RA[–]CD11a[+]). Consistent with the extensive bone remodeling in PsA, patients with PsA have been found to have a marked increase in osteoclastic precursors in peripheral blood and upregulation of receptor activator of nuclear factor κB ligand (RANKL) in the synovial lining layer. Increased serum levels of TNF-α, RANKL, leptin, and omentin positively correlate with these osteoclastic precursors.

CLINICAL FEATURES

In 60–70% of cases, psoriasis precedes joint disease. In 15–20% of cases, the two manifestations appear within 1 year of each other. In about 15–20% of cases, the arthritis precedes the onset of psoriasis and can present a diagnostic challenge. The frequency in men and women is almost equal, although the frequency of disease patterns differs somewhat in the two sexes. The disease can begin in childhood or late in life but typically begins in the fourth or fifth decade, at an average age of 37 years.

Many classification schemes have been proposed for the broad spectrum of arthropathy in PsA. Wright and Moll described five patterns: (1) arthritis of the DIP joints; (2) asymmetric oligoarthritis; (3) symmetric polyarthritis similar to RA; (4) axial involvement (spine and sacroiliac joints); and (5) arthritis mutilans, a highly destructive form of the disease. These patterns frequently coexist, and the pattern that persists chronically often differs from that of the initial presentation. A simpler scheme in recent use contains three patterns: oligoarthritis, polyarthritis, and axial arthritis.

Nail changes in the fingers or toes occur in most patients with PsA, compared with only a minority of psoriatic patients without arthritis, and pustular psoriasis is said to be associated with more severe arthritis. Dactylitis and enthesitis are common in PsA and help to distinguish it from other joint disorders. Dactylitis occurs in >30%; enthesitis and tenosynovitis are probably present in most patients, although often not appreciated on physical examination. Shortening of digits because of underlying osteolysis is particularly characteristic of PsA **(Fig. 355-3)**, and there is a much greater tendency than in RA for both fibrous and bony ankylosis of small joints. Rapid ankylosis of one or more proximal interphalangeal (PIP) joints early in the course of disease is not uncommon. Joint involvement tends to follow a “ray” distribution, with all of the joints of one finger involved, while sparing adjacent fingers entirely. Back and neck pain and stiffness are also common in PsA.

FIGURE 355-3

Characteristic lesions of psoriatic arthritis. Inflammation is prominent in the distal interphalangeal joints (left 5th, 4th, 2nd; right 2nd, 3rd, and 5th) and proximal interphalangeal joints (left 2nd, right 2nd, 4th, and 5th). There is dactylitis in the left 2nd finger and thumb, with pronounced telescoping of the left 2nd finger. Nail dystrophy (hyperkeratosis and onycholysis) affects each of the fingers except the left 3rd finger, the only finger without arthritis. *(Courtesy of Donald Raddatz, MD; with permission.*)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Arthropathy confined to the DIP joints occurs in about 5% of cases. Accompanying nail changes in the affected digits are almost always present. These joints are also often affected in the other patterns of PsA. Approximately 30% of patients have asymmetric oligoarthritis. This pattern commonly involves a knee or another large joint with a few small joints in the fingers or toes, often with dactylitis. Symmetric polyarthritis occurs in about 40% of PsA patients at presentation. It may be indistinguishable from RA in terms of the joints involved, but other features characteristic of PsA are usually also present. Almost any peripheral joint can be involved. Axial arthropathy without peripheral involvement is found in about 5% of PsA patients. It may be clinically indistinguishable from idiopathic AS, although more neck involvement and less thoracolumbar spinal involvement are characteristic, and nail changes are not found in idiopathic AS. A small percentage of PsA patients have arthritis mutilans, in which there can be widespread shortening of digits ("telescoping"), sometimes coexisting with ankylosis and contractures in other digits.

Six patterns of nail involvement are identified: pitting, horizontal ridging, onycholysis, yellowish discoloration of the nail margins, dystrophic hyperkeratosis, and combinations of these findings. Extraarticular and extradermal manifestations are common. Eye involvement, either conjunctivitis or uveitis, is reported in 7–33% of PsA patients. Unlike the uveitis associated with AS, the uveitis in PsA is more often bilateral, chronic, and/or posterior. Aortic valve insufficiency has been found in <4% of patients, usually after long-standing disease.

Widely varying estimates of clinical outcome have been reported in PsA. At its worst, severe PsA with arthritis mutilans is potentially at least as crippling and ultimately fatal as severe RA. Unlike RA, however, many patients with PsA experience temporary remissions. Overall, erosive disease develops in the majority of patients, progressive disease with deformity and disability is common, and in some large series, mortality was found to be significantly increased compared with the general population. There appears to be a greater incidence of cardiovascular death in psoriatic disease.

The psoriasis and associated arthropathy seen with HIV infection both tend to be severe and can occur in populations with low prevalence of psoriasis. Severe enthesopathy, dactylitis, and rapidly progressive joint destruction are seen, but axial involvement is very rare. This condition is prevented by or responds well to antiretroviral therapy.

LABORATORY AND RADIOGRAPHIC FINDINGS

There are no laboratory tests diagnostic of PsA. ESR and CRP are often elevated. A small percentage of patients may have low titers of rheumatoid factor or ANAs. About 10% of patients have anti-CCP antibodies. Uric acid may be elevated in the presence of extensive psoriasis. HLA-B27 is found in 50–70% of patients with axial disease, but in ≤20% of patients with only peripheral joint involvement.

The peripheral and axial arthropathies in PsA show a number of radiographic features that distinguish them from RA and AS, respectively. Characteristics of peripheral PsA include DIP involvement, including the classic "pencil-in-cup" deformity; marginal erosions with adjacent bony proliferation ("whiskering"); small-joint ankylosis; osteolysis of phalangeal and metacarpal bone, with telescoping of digits; periostitis and proliferative new bone at sites of enthesitis, and a "ray" distribution of lesions. Characteristics of axial PsA that differ from idiopathic AS include asymmetric sacroiliitis; less zygapophyseal joint arthritis; nonmarginal, bulky, "comma"-shaped syndesmophytes that tend to be fewer, less symmetric, and less delicate than the marginal syndesmophytes of AS; fluffy hyperperiostosis on anterior vertebral bodies; severe cervical spine involvement, with a tendency to atlantoaxial subluxation but relative sparing of the thoracolumbar spine; and paravertebral ossification.

Ultrasound and MRI both readily demonstrate enthesitis and tendon sheath effusions that can be difficult to assess on physical examination. A recent MRI study of 68 PsA patients found sacroiliitis in 35%, unrelated to B27 but correlated with restricted spinal movement.

DIAGNOSIS

Classification criteria for PsA were published in 2006 (Classification of Psoriatic Arthritis [CASPAR] criteria) (Table 355-2). The sensitivity and specificity of these criteria exceed 90%, and they are useful for early diagnosis in clinical practice. Diagnosis can be challenging when the arthritis precedes psoriasis, the psoriasis is undiagnosed or obscure, or the joint involvement closely resembles another form of arthritis. A high index of suspicion is needed in any patient with an undiagnosed inflammatory arthropathy. The history should include inquiry about psoriasis in the patient and family members. Patients should disrobe for the physical examination, and psoriasiform lesions should be sought in the scalp, ears, umbilicus, and gluteal folds in addition to more accessible sites; the finger and toe nails should also be carefully examined. Axial symptoms or signs, dactylitis, enthesitis, ankylosis, the pattern of joint involvement, and characteristic radiographic changes can be helpful clues. The differential diagnosis includes all other forms of arthritis, which can occur coincidentally in individuals with psoriasis. The differential diagnosis of isolated DIP involvement is short. Osteoarthritis (Heberden's nodes) is usually not inflammatory; gout involving more than one DIP joint often involves other sites and may be accompanied by tophi; the very rare entity multicentric reticulohistiocytosis involves other joints and has characteristic small pearly periungual skin nodules; and the uncommon entity inflammatory osteoarthritis, like the others, lacks the nail changes of PsA. Radiography can be helpful in all of these cases and in distinguishing between psoriatic spondylitis and idiopathic AS. A history of trauma to an affected joint preceding the onset of arthritis may occur more frequently in PsA than in other types of arthritis, perhaps reflecting the Koebner phenomenon in which psoriatic skin lesions arise at sites of skin trauma.

TABLE 355-2

The CASPAR (*Cl*assification Criteria for *P*soriatic *Ar*thritis) Criteria[a]

To meet the CASPAR criteria, a patient must have inflammatory articular disease (joint, spine, or entheseal) with ≥3 points from any of the following five categories:
1. Evidence of current psoriasis,[b,c] a personal history of psoriasis, or a family history of psoriasis[d]
2. Typical psoriatic nail dystrophy[e] observed on current physical examination
3. A negative test result for rheumatoid factor
4. Either current dactylitis[f] or a history of dactylitis recorded by a rheumatologist
5. Radiographic evidence of juxtaarticular new bone formation[g] in the hand or foot

[a]Specificity of 99% and sensitivity of 91%. [b]Current psoriasis is assigned 2 points; all other features are assigned 1 point. [c]Psoriatic skin or scalp disease present at the time of examination, as judged by a rheumatologist or dermatologist. [d]History of psoriasis in a first- or second-degree relative. [e]Onycholysis, pitting, or hyperkeratosis. [f]Swelling of an entire digit. [g]Ill-defined ossification near joint margins, excluding osteophyte formation.

Source: From W Taylor et al: Arthritis Rheum, 54:2665, 2006.

TREATMENT

TREATMENT

Psoriatic Arthritis

Ideally, coordinated therapy is directed at both the skin and joints in PsA. Use of the anti-TNF-α agents has revolutionized the treatment of PsA. Prompt and dramatic resolution of both arthritis and skin lesions has been observed in large, randomized controlled trials of all five agents. Many of the responding patients had long-standing disease that was resistant to all previous therapy, as well as extensive skin disease. The clinical response is often more dramatic than in RA, and delay of disease progression has been demonstrated radiographically. The potential additive effect of methotrexate to anti-TNF-α agents in PsA remains uncertain. As noted above, anti-TNF therapy, paradoxically, has been reported to

trigger exacerbation or de novo appearance of psoriasis, typically the palmoplantar pustular variety. In some cases, the therapy can nevertheless be continued.

The anti-IL-17A monoclonal antibody secukinumab is effective in treating both psoriasis and PsA. Ixekizumab, another IL-17 antagonist, is approved for treatment of psoriasis and PsA. Ustekinumab, a monoclonal antibody to the shared IL-23/IL-12p40 subunit, is an efficacious treatment for psoriasis and has some efficacy for PsA. Apremilast, an oral phosphodiesterase-4 inhibitor, is approved for both psoriasis and PsA. Although not quite as effective for PsA as the biologics, apremilast has a more favorable safety profile. It is not indicated in patients with radiographically evident joint damage or axial involvement. The oral Jak inhibitor, tofacitinib, is approved for treatment of PsA in patients with an inadequate response to another disease-modifying agent. In a clinical trial it efficacy in PsA comparable to adalimumab.

Other treatment for PsA has been based on drugs that have efficacy in RA and/or in psoriasis. Methotrexate in doses of 15–25 mg/week has moderate efficacy for psoriasis, but limited efficacy for PsA. Sulfasalazine (usually given in doses of 2–3 g/d) may not have efficacy for either, and neither regimen halts progression of erosive joint disease. Other agents with efficacy in psoriasis reported to benefit PsA are cyclosporine, retinoic acid derivatives, and psoralens plus ultraviolet A light (PUVA). The pyrimidine synthetase inhibitor leflunomide has been shown to be beneficial in PsA, with modest benefit for psoriasis. There is controversy regarding the efficacy in PsA of gold and antimalarials, which have been widely used in RA.

All of these treatments require careful monitoring. Immuno-suppressive therapy may be used cautiously in HIV-associated PsA if the HIV infection is well controlled.

UNDIFFERENTIATED AND JUVENILE-ONSET SPONDYLOARTHRITIS

Many patients present with some features of one or more of the spondyloarthritides discussed above. Until recently, these patients were said to have *undifferentiated spondyloarthritis*, or simply *spondyloarthritis*, as defined by the 1991 European Spondyloarthropathy Study Group criteria. For example, a patient may present with inflammatory synovitis of one knee, Achilles tendinitis, and dactylitis of one digit. Some of these patients may have ReA in which the triggering infection remains clinically silent. In some other cases, the patient subsequently develops IBD or psoriasis. The diagnosis of undifferentiated SpA was also commonly applied to patients with IBP who did meet modified New York criteria for AS. Most of these would now be classified as nr-ax-SpA (Table 355-1).

Comparable to the classification criteria for axial symptoms, the ASAS has formulated criteria for peripheral SpA. This is intended to exclude patients with axial symptoms and thus to divide the universe of patients with SpA into predominantly axial and predominantly peripheral subsets. These criteria are shown in **Table 355-3**.

TABLE 355-3

ASAS Criteria for Peripheral Spondyloarthritis[a]

ARTHRITIS[b] *OR* ENTHESITIS *OR* DACTYLITIS
PLUS EITHER One or more of the following SpA features:
• Psoriasis • Crohn's disease or ulcerative colitis • Preceding infection • Uveitis • HLA-B27 • Sacroiliitis on imaging (radiographs or MRI)
OR two or more of the following SpA features:
• Arthritis • Enthesitis • Dactylitis • Inflammatory back pain ever • Family history for SpA

[a]Sensitivity 78%, specificity 82%. [b]Peripheral arthritis, usually predominantly lower limb and/or asymmetric. The various SpA features are as defined in Table 355-1. Preceding infection refers to preceding gastrointestinal or urogenital infection.

Source: M Rudawaleit et al: Ann Rheum Dis 70:25, 2011.

At most only one-half of the patients with undifferentiated SpA are HLA-B27-positive.

In juvenile-onset SpA, which begins between ages 7 and 16 years, an asymmetric, predominantly lower-extremity oligoarthritis and enthesitis without extraarticular features is the typical mode of presentation. There is male predominance (60–80%), and the prevalence of B27 in this condition, which has been termed the *seronegative enthesopathy and arthropathy* (*SEA*) *syndrome*, is approximately 80%. Despite the absence of axial symptoms, active sacroiliitis by MRI has been found to be common at diagnosis. Many, but not all, of these patients go on to develop AS in late adolescence or adulthood.

Management of undifferentiated SpA is similar to that of the other spondyloarthritides. Anti-TNF-α therapy is indicated in severe, persistent cases not responsive to other treatment.

Current pediatric textbooks and journals should be consulted for information on management of juvenile-onset SpA.

ENTEROPATHIC ARTHRITIS

HISTORIC BACKGROUND

A relationship between arthritis and IBD was observed in the 1930s. The relationship was further defined by epidemiologic studies in the 1950s and 1960s and included in the concept of the spondyloarthritides in the 1970s.

EPIDEMIOLOGY

Both of the common forms of IBD, ulcerative colitis (UC) and Crohn's disease (CD) **(Chap. 319)**, are associated with SpA. UC and CD both have an estimated prevalence of 0.1–0.2%, and the incidence of each is thought to have increased in recent decades. AS, nr-ax-SpA, and peripheral arthritis are all associated with UC and CD. Wide variations have been reported in the estimated frequencies of these associations. In recent series,

AS was diagnosed in 1–10%, and peripheral arthritis in 10–50% of patients with IBD. IBP and enthesopathy are common, and many patients have sacroiliitis on imaging studies.

The prevalence of UC or CD in patients with AS is thought to be 5–10%, and a recent meta-analysis found the prevalence in patients with nr-ax-SpA to be 6.4%. However, investigation of unselected SpA patients by ileocolonoscopy has revealed that from one-third to two-thirds of patients with AS have subclinical intestinal inflammation that is evident either macroscopically or histologically. These lesions have also been found in patients with undifferentiated SpA or ReA (both enterically and urogenitally acquired).

Both UC and CD have a tendency to familial aggregation, more so for CD. HLA associations have been weak and inconsistent. HLA-B27 is found in up to 70% of patients with IBD and AS, but in ≤15% of patients with IBD and peripheral arthritis or IBD alone. Three alleles of the *NOD2/CARD15* gene on chromosome 16 have been found in approximately one-half of patients with CD. These alleles are not associated with SpA per se. However, they are found significantly more often in (1) CD patients with sacroiliitis than in those without sacroiliitis, and (2) SpA patients with chronic inflammatory gut lesions than in those with normal gut histology. These associations are independent of HLA-B27. In addition to *NOD2*, over 200 other genes have been found to be associated with CD, UC, or both. Many of the SNPs associated with AS are also associated with IBD.

PATHOLOGY

Available data for IBD-associated peripheral arthritis suggest a synovial histology similar to other spondyloarthritides. Association with arthropathy does not affect the gut histology of UC or CD **(Chap. 319)**. The subclinical inflammatory lesions in the colon and distal ileum associated with SpA have been classified as either acute or chronic. The former resemble acute bacterial enteritis, with largely intact architecture and neutrophilic infiltration in the lamina propria. The latter resemble the lesions of CD, with distortion of villi and crypts, aphthoid ulceration, and mononuclear cell infiltration in the lamina propria.

PATHOGENESIS

Both IBD and SpA are immune-mediated, but the specific pathogenic mechanisms are poorly understood, and the connection between the two is obscure. The shared genetics evidently reflects shared pathogenic mechanisms. A number of rodent models showing various immune perturbations manifest both IBD and arthritis. Resident innate immune cells and intestinal dysbiosis have been implicated in both conditions. Several lines of evidence indicate trafficking of leukocytes between the gut and the joint. Mucosal leukocytes from IBD patients have been shown to bind avidly to synovial vasculature through several different adhesion molecules. Macrophages expressing CD163 are prominent in the inflammatory lesions of both gut and synovium in the spondyloarthritides.

CLINICAL FEATURES

AS associated with IBD is clinically indistinguishable from idiopathic AS. It runs a course independent of the bowel disease, and in some patients, it precedes the onset of IBD, sometimes by many years. Peripheral arthritis may also begin before onset of overt bowel disease. The spectrum of peripheral arthritis includes acute self-limited attacks of oligoarthritis that often coincide with relapses of IBD, and more chronic and symmetric polyarticular arthritis that runs a course independent of IBD activity. The patterns of joint involvement are similar in UC and CD. In general, erosions and deformities are infrequent in IBD-associated peripheral arthritis. Isolated destructive hip arthritis is a rare complication of CD, apparently distinct from osteonecrosis and septic arthritis. Dactylitis and enthesopathy are occasionally found. In addition to the ~20% of IBD patients with SpA, a comparable percentage have arthralgias or fibromyalgia symptoms.

Other extraintestinal manifestations of IBD are seen in addition to arthropathy, including uveitis, pyoderma gangrenosum, erythema nodosum, and finger clubbing, all somewhat more commonly in CD than UC. The uveitis shares the features described above for PsA-associated uveitis.

LABORATORY AND RADIOGRAPHIC FINDINGS

Laboratory findings reflect the inflammatory and metabolic manifestations of IBD. Joint fluid is usually at least mildly inflammatory. Of patients with AS and IBD, 30–70% carry the HLA-B27 gene, compared with 75–90% of patients with AS alone and 50–70% of those with AS and psoriasis. Hence, definite or probable AS in a B27-negative individual in the absence of psoriasis should prompt a search for occult IBD. Radiographic changes in the axial skeleton are the same as in uncomplicated AS. Erosions are uncommon in peripheral arthritis but may occur, particularly in the metatarsophalangeal joints.

DIAGNOSIS

Diarrhea and arthritis are both common conditions that can coexist for a variety of reasons. When etiopathogenically related, ReA and IBD-associated arthritis are the most common causes. Rare causes include celiac disease, blind loop syndromes, and Whipple's disease. In most cases,

diagnosis depends on investigation of the bowel disease.

TREATMENT

TREATMENT

Enteropathic Arthritis

Treatment of CD has been improved by therapy with anti-TNF agents. Infliximab, adalimumab, and certolizumab pegol are effective for induction and maintenance of clinical remission in CD, and infliximab has been shown to be effective in fistulizing CD. IBD-associated arthritis also responds to these agents. Other treatment for IBD, including sulfasalazine and related drugs, systemic glucocorticoids, and immunosuppressive drugs, is also usually of benefit for associated peripheral arthritis. NSAIDs are generally helpful and well tolerated, but they can precipitate flares of IBD. As noted above for psoriasis, rare cases of IBD, either CD or UC, have apparently been precipitated by anti-TNF therapy, usually etanercept, given for any of several rheumatic diseases.

SAPHO SYNDROME

The syndrome of synovitis, acne, pustulosis, hyperostosis, and osteitis (SAPHO) is characterized by a variety of skin and musculoskeletal manifestations. Dermatologic manifestations include palmoplantar pustulosis, acne conglobata, acne fulminans, and hidradenitis suppurativa. The main musculoskeletal findings are sternoclavicular and spinal hyperostosis, chronic recurrent foci of sterile osteomyelitis, and axial or peripheral arthritis. Cases with one or a few manifestations are probably the rule. The ESR and/or CRP are usually mildly to moderately elevated, occasionally dramatically. In some cases, bacteria, most often *Propionibacterium acnes*, have been cultured from bone biopsy specimens and occasionally other sites. IBD was coexistent in 8% of patients in one large series. B27 is not associated. Either bone scan or computed tomography scan is helpful diagnostically. An MRI report described characteristic vertebral body corner cortical erosions in 12 of 12 patients. High-dose NSAIDs may provide relief from bone pain. A number of uncontrolled series and case reports describe successful therapy with pamidronate or other bisphosphonates. Response to anti-TNF-α therapy has also been observed, although in a few cases this has been associated with a flare of skin manifestations. Successful prolonged antibiotic therapy has also been reported. Recent reports suggest a possible autoinflammatory pathogenesis and successful treatment with the IL-1 receptor antagonist anakinra.

FURTHER READING

Bowness P: HLA-B27. Annu Rev Immunol 33:29, 2015. [PubMed: 25861975]

Carter JD, Hudson AP: Recent advances and future directions in understanding and treating Chlamydia-induced reactive arthritis. Expert Rev Clin Immunol. 13:197, 2017. [PubMed: 27627462]

de Winter JJ et al: Prevalence of peripheral and extra-articular disease in ankylosing spondylitis versus non-radiographic axial spondyloarthritis: A meta-analysis. Arthritis Res Ther 18:196, 2016. [PubMed: 27586785]

Ellinghaus D et al: Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet 48:510, 2016. [PubMed: 26974007]

Gladman DD: Editorial: What is peripheral spondyloarthritis? Arthritis Rheumatol 67:865, 2015. [PubMed: 25545846]

Gmuca S, Weiss PF: Juvenile spondyloarthritis. Curr Opin Rheumatol 27:364, 2015. [PubMed: 26002028]

Maksymowych WP: The role of MRI in the evaluation of spondyloarthritis: A clinician's guide. Clin Rheumatol 35:1447, 2016. [PubMed: 27094946]

Reinhardt A et al: Interleukin-23-dependent gamma/delta T cells produce interleukin-17 and accumulate in the enthesis, aortic valve, and ciliary body in mice. Arthritis Rheumatol 68:2476, 2016. [PubMed: 27111864]

Ritchlin CT et al: Psoriatic arthritis. N Engl J Med 376:957, 2017. [PubMed: 28273019]

Sieper J, et al: Secukinumab efficacy in anti-TNF-naive and anti-TNF-experienced subjects with active ankylosing spondylitis: Results from the MEASURE 2 Study. Ann Rheum Dis 76:571, 2017. [PubMed: 27582421]

Sieper J, Poddubnyy D: Axial spondyloarthritis. Lancet 390:73, 2017. [PubMed: 28110981]

Taurog JD et al: Ankylosing spondylitis and axial spondyloarthritis. New Engl J Med 374:2563, 2016. [PubMed: 27355535]

Ward MM et al: American College of Rheumatology/Spondylitis Association of America/Spondyloarthritis Research and Treatment Network 2015 recommendations for the treatment of ankylosing spondylitis and nonradiographic axial spondyloarthritis. Arthritis Rheumatol 68:282, 2016. [PubMed: 26401991]

Chapter 356: The Vasculitis Syndromes

Carol A. Langford; Anthony S. Fauci

DEFINITION

Vasculitis is a clinicopathologic process characterized by inflammation of and damage to blood vessels. The vessel lumen is usually compromised, and this is associated with ischemia of the tissues supplied by the involved vessel. A broad and heterogeneous group of syndromes may result from this process, since any type, size, and location of blood vessel may be involved. Vasculitis and its consequences may be the primary or sole manifestation of a disease; alternatively, vasculitis may be a secondary component of another disease. Vasculitis may be confined to a single organ, such as the skin, or it may simultaneously involve several organ systems.

CLASSIFICATION

A major feature of the vasculitic syndromes as a group is the fact that there is a great deal of heterogeneity at the same time as there is considerable overlap among them. Table 356-1 lists the major vasculitis syndromes. The distinguishing and overlapping features of these syndromes are discussed below.

TABLE 356-1

Vasculitis Syndromes

Primary Vasculitis Syndromes	Secondary Vasculitis Syndromes
Granulomatosis with polyangiitis (Wegener's) Microscopic polyangiitis Eosinophilic granulomatosis with polyangiitis (Churg-Strauss) IgA vasculitis (Henoch-Schönlein) Cryoglobulinemic vasculitis Polyarteritis nodosa Kawasaki disease Giant cell arteritis Takayasu arteritis Behçet's disease Cogan's syndrome Single-organ vasculitis Cutaneous leukocytoclastic angiitis Cutaneous arteritis Primary central nervous system vasculitis Isolated aortitis	Vasculitis associated with probable etiology Drug-induced vasculitis Hepatitis C virus–associated cryoglobulinemic vasculitis Hepatitis B virus–associated vasculitis Cancer-associated vasculitis Vasculitis associated with systemic disease Lupus vasculitis Rheumatoid vasculitis Sarcoid vasculitis

Source: Adapted from JC Jennette et al: Arthritis Rheum 65:1, 2013.

PATHOPHYSIOLOGY AND PATHOGENESIS

Generally, most of the vasculitic syndromes are assumed to be mediated at least in part by immunopathogenic mechanisms that occur in response to certain antigenic stimuli. However, evidence supporting this hypothesis is for the most part indirect and may reflect epiphenomena as opposed to true causality. Furthermore, it is unknown why some individuals might develop vasculitis in response to certain antigenic stimuli, whereas others do not. It is likely that a number of factors are involved in the ultimate expression of a vasculitic syndrome. These include the genetic predisposition, environmental exposures, and the regulatory mechanisms associated with immune response to certain antigens. Although immune complex formation, antineutrophil cytoplasmic antibodies (ANCAs), and pathogenic T lymphocyte responses (Table 356-2) have been among the prominent hypothesized mechanisms, it is likely that the pathogenesis of individual forms of vasculitis is complex and varied.

TABLE 356-2

Potential Mechanisms of Vessel Damage in Vasculitis Syndromes

Pathogenic immune-complex formation and/or deposition
IgA vasculitis (Henoch-Schönlein)
Lupus vasculitis
Serum sickness and cutaneous vasculitis syndromes
Hepatitis C virus–associated cryoglobulinemic vasculitis
Hepatitis B virus–associated vasculitis
Production of antineutrophilic cytoplasmic antibodies
Granulomatosis with polyangiitis (Wegener's)
Microscopic polyangiitis
Eosinophilic granulomatosis with polyangiitis (Churg-Strauss)
Pathogenic T lymphocyte responses and granuloma formation
Giant cell arteritis
Takayasu arteritis
Granulomatosis with polyangiitis (Wegener's)
Eosinophilic granulomatosis with polyangiitis (Churg-Strauss)

Source: Adapted from MC Sneller, AS Fauci: Med Clin North Am 81:221, 1997.

PATHOGENIC IMMUNE-COMPLEX FORMATION

Deposition of immune complexes was the first and most widely accepted pathogenic mechanism of vasculitis. However, the causal role of immune complexes has not been clearly established in most of the vasculitic syndromes. Circulating immune complexes need not result in deposition of the complexes in blood vessels with ensuing vasculitis, and many patients with active vasculitis do not have demonstrable circulating or deposited immune complexes. The actual antigen contained in the immune complex has only rarely been identified in vasculitic syndromes. In this regard, hepatitis B antigen has been identified in both the circulating and deposited immune complexes in a subset of patients who have features of a systemic vasculitis, most notably in polyarteritis nodosa (see "Polyarteritis Nodosa"). Cryoglobulinemic vasculitis is strongly associated with hepatitis C virus infection; hepatitis C virions and hepatitis C virus antigen-antibody complexes have been identified in the cryoprecipitates of these patients (see "Cryoglobulinemic Vasculitis").

The mechanisms of tissue damage in immune complex–mediated vasculitis resemble those described for serum sickness. In this model, antigen-antibody complexes are formed in antigen excess and are deposited in vessel walls whose permeability has been increased by vasoactive amines such as histamine, bradykinin, and leukotrienes released from platelets or from mast cells as a result of IgE-triggered mechanisms. The deposition of complexes results in activation of complement components, particularly C5a, which is strongly

chemotactic for neutrophils. These cells then infiltrate the vessel wall, phagocytose the immune complexes, and release their intracytoplasmic enzymes, which damage the vessel wall. As the process becomes subacute or chronic, mononuclear cells infiltrate the vessel wall. The common denominator of the resulting syndrome is compromise of the vessel lumen with ischemic changes in the tissues supplied by the involved vessel. Several variables may explain why only certain types of immune complexes cause vasculitis and why only certain vessels are affected in individual patients. These include the ability of the reticuloendothelial system to clear circulating complexes from the blood, the size and physicochemical properties of immune complexes, the relative degree of turbulence of blood flow, the intravascular hydrostatic pressure in different vessels, and the preexisting integrity of the vessel endothelium.

ANTINEUTROPHIL CYTOPLASMIC ANTIBODIES

Antineutrophil cytoplasmic antibodies (ANCA) are antibodies directed against certain proteins in the cytoplasmic granules of neutrophils and monocytes. These autoantibodies are present in a high percentage of patients with active granulomatosis with polyangiitis (Wegener's) and microscopic polyangiitis, and in a lower percentage of patients with eosinophilic granulomatosis with polyangiitis (Churg-Strauss). Because these diseases share the presence of ANCA and small-vessel vasculitis, some investigators have come to refer to them collectively as "ANCA-associated vasculitis." However, as these diseases possess unique clinical phenotypes in which ANCA may be absent, it remains our opinion that granulomatosis with polyangiitis (Wegener's), microscopic polyangiitis and eosinophilic granulomatosis with polyangiitis (Churg-Strauss) should continue to be viewed as separate entities.

There are two major categories of ANCA based on different targets for the antibodies. The terminology of *cytoplasmic ANCA* (cANCA) refers to the diffuse, granular cytoplasmic staining pattern observed by immunofluorescence microscopy when serum antibodies bind to indicator neutrophils. Proteinase-3, a 29-kDa neutral serine proteinase present in neutrophil azurophilic granules, is the major cANCA antigen. More than 90% of patients with typical active granulomatosis with polyangiitis (Wegener's) have detectable antibodies to proteinase-3 (see below). The terminology of *perinuclear ANCA* (pANCA) refers to the more localized perinuclear or nuclear staining pattern of the indicator neutrophils. The major target for pANCA is the enzyme myeloperoxidase; other targets that can produce a pANCA pattern of staining include elastase, cathepsin G, lactoferrin, lysozyme, and bactericidal/permeability-increasing protein. However, only antibodies to myeloperoxidase have been convincingly associated with vasculitis. Antimyeloperoxidase antibodies have been reported to occur in variable percentages of patients with microscopic polyangiitis, eosinophilic granulomatosis with polyangiitis (Churg-Strauss), isolated necrotizing crescentic glomerulonephritis, and granulomatosis with polyangiitis (Wegener's) (see below). A pANCA pattern of staining that is not due to antimyeloperoxidase antibodies has been associated with nonvasculitic entities such as rheumatic and nonrheumatic autoimmune diseases, inflammatory bowel disease, certain drugs, and infections such as endocarditis and bacterial airway infections in patients with cystic fibrosis.

It is unclear why patients with these vasculitis syndromes develop antibodies to myeloperoxidase or proteinase-3 or what role these antibodies play in disease pathogenesis. There are a number of in vitro

observations that suggest possible mechanisms whereby these antibodies can contribute to the pathogenesis of the vasculitis syndromes. Proteinase-3 and myeloperoxidase reside in the azurophilic granules and lysosomes of resting neutrophils and monocytes, where they are apparently inaccessible to serum antibodies. However, when neutrophils or monocytes are primed by tumor necrosis factor α (TNF-α) or interleukin 1 (IL-1), proteinase-3 and myeloperoxidase translocate to the cell membrane, where they can interact with extracellular ANCA. The neutrophils then degranulate and produce reactive oxygen species that can cause tissue damage. Furthermore, ANCA-activated neutrophils can adhere to and kill endothelial cells in vitro. Activation of neutrophils and monocytes by ANCA also induces the release of proinflammatory cytokines such as IL-1 and IL-8. Adoptive transfer experiments in genetically engineered mice provide further evidence for a direct pathogenic role of ANCA in vivo. In contradiction, however, a number of clinical and laboratory observations argue against a primary pathogenic role for ANCA. Patients may have active granulomatosis with polyangiitis (Wegener's) in the absence of ANCA; the absolute height of the antibody titers does not correlate well with disease activity; and patients with granulomatosis with polyangiitis (Wegener's) in remission may continue to have high antiproteinase-3 (cANCA) titers for years (see below).

PATHOGENIC T LYMPHOCYTE RESPONSES AND GRANULOMA FORMATION

The histopathologic feature of granulomatous vasculitis has provided evidence to support a role of pathogenic T lymphocyte responses and cell-mediated immune injury. Vascular endothelial cells can express human leukocyte antigen (HLA) class II molecules following activation by cytokines such as interferon (IFN) γ. This allows these cells to participate in immunologic reactions such as interaction with CD4+ T lymphocytes in a manner similar to antigen-presenting macrophages. Endothelial cells can secrete IL-1, which may activate T lymphocytes and initiate or propagate in situ immunologic processes within the blood vessel. In addition, IL-1 and TNF-α are potent inducers of endothelial-leukocyte adhesion molecule 1 (ELAM-1) and vascular cell adhesion molecule 1 (VCAM-1), which may enhance the adhesion of leukocytes to endothelial cells in the blood vessel wall.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

General Principles of Diagnosis

The diagnosis of vasculitis should be considered in any patient with an unexplained systemic illness. However, there are certain clinical abnormalities that when present alone or in combination should suggest a diagnosis of vasculitis. These include palpable purpura, pulmonary infiltrates and microscopic hematuria, chronic inflammatory sinusitis, mononeuritis multiplex, unexplained ischemic events, and glomerulonephritis with evidence of multisystem disease. A number of nonvasculitic diseases may also produce some or all of these abnormalities. Thus, the first step in the workup of a patient with suspected vasculitis is to exclude other diseases that produce clinical manifestations that can mimic vasculitis **(Table 356-3)**. It is particularly important to exclude infectious diseases with features that overlap those of

vasculitis, especially if the patient's clinical condition is deteriorating rapidly and empirical immunosuppressive treatment is being contemplated.

Once diseases that mimic vasculitis have been excluded, the workup should follow a series of progressive steps that establish the diagnosis of vasculitis and determine, where possible, the category of the vasculitis syndrome **(Fig. 356-1)**. This approach is of considerable importance since several of the vasculitis syndromes require aggressive therapy with glucocorticoids and other immunosuppressive agents, whereas other syndromes usually resolve spontaneously and require symptomatic treatment only. The definitive diagnosis of vasculitis is usually made based on biopsy of involved tissue. The yield of "blind" biopsies of organs with no subjective or objective evidence of involvement is very low and should be avoided. When syndromes such as polyarteritis nodosa, Takayasu's arteritis, or primary central nervous system (CNS) vasculitis are suspected, arteriogram of organs with suspected involvement should be performed.

GENERAL PRINCIPLES OF TREATMENT

Once a diagnosis of vasculitis has been established, a decision regarding therapeutic strategy must be made (Fig. 356-1). If an offending antigen that precipitates the vasculitis is recognized, the antigen should be removed where possible. If the vasculitis is associated with an underlying disease such as an infection, neoplasm, or connective tissue disease, the underlying disease should be treated. If the syndrome represents a primary vasculitic disease, treatment should be initiated according to the category of the vasculitis syndrome. Specific therapeutic regimens are discussed below for the individual vasculitis syndromes; however, certain general principles regarding therapy should be considered. Decisions regarding treatment should be based on the use of regimens for which there has been published literature supporting efficacy for that particular vasculitic disease. Since the potential toxic side effects of certain therapeutic regimens may be substantial, the risk-versus-benefit ratio of any therapeutic approach should be weighed carefully. On the one hand, glucocorticoids and/or other immunosuppressive agents should be instituted immediately in diseases where irreversible organ system dysfunction and high morbidity and mortality rates have been clearly established. Granulomatosis with polyangiitis (Wegener's) is the prototype of a severe systemic vasculitis requiring such a therapeutic approach (see below). On the other hand, when feasible, aggressive therapy should be avoided for vasculitic manifestations that rarely result in irreversible organ system dysfunction and that usually do not respond to such therapy. For example, isolated idiopathic cutaneous vasculitis usually resolves with symptomatic treatment, and prolonged courses of glucocorticoids uncommonly result in clinical benefit. Cytotoxic agents have not proved to be beneficial in idiopathic cutaneous vasculitis, and their toxic side effects generally outweigh any potential beneficial effects. Glucocorticoids should be initiated in those systemic vasculitides that cannot be specifically categorized or for which there is no established standard therapy, or other immunosuppressive therapy should be added in these diseases only if an adequate response does not result or if remission can only be achieved and maintained with an unacceptably toxic regimen of glucocorticoids. When remission is achieved, one should continually attempt to taper glucocorticoids and discontinue when possible. When using other immunosuppressive regimens, one should base the choice of agent upon the available therapeutic data supporting efficacy in that disease, the site and severity of organ involvement, and the toxicity profile of the drug.

Physicians should be thoroughly aware of the toxic side effects of therapeutic agents employed that can include both acute and long-term complications (Table 356-4). Morbidity and mortality can occur as a result of treatment, and strategies to monitor for and prevent toxicity represent an essential part of patient care. Glucocorticoids are an important part of treatment for most vasculitides but are associated with substantial toxicities. Monitoring and prevention of glucocorticoid-induced bone loss are important in all patients. With the use of daily cyclophosphamide, strategies are particularly important and are directed toward minimization of bladder toxicity and prevention of leukopenia. Instructing the patient to take cyclophosphamide all at once in the morning with a large amount of fluid throughout the day in order to maintain dilute urine can reduce the risk of bladder injury. Bladder cancer can occur several years after discontinuation of cyclophosphamide therapy; therefore, monitoring for bladder cancer should continue indefinitely in patients who have received cyclophosphamide. Bone marrow suppression is an important toxicity of cyclophosphamide and can be observed during glucocorticoid tapering or over time, even after periods of stable measurements. Monitoring of the complete blood count every 1–2 weeks for as long as the patient receives cyclophosphamide can effectively prevent cytopenias. Maintaining the white blood cell (WBC) count at >3000/μL and the neutrophil count at >1500/μL is essential to reduce the risk of life-threatening infections.

Methotrexate and azathioprine are also associated with bone marrow suppression, and complete blood counts should be obtained every 1–2 weeks for the first 1–2 months after their initiation and once a month thereafter. To lessen toxicity, methotrexate is often given together with folic acid, 1 mg daily, or folinic acid, 5–10 mg once a week 24 h following methotrexate. Prior to initiation of azathioprine, thiopurine methyltransferase (TPMT), an enzyme involved in the metabolism of azathioprine, should be assayed because inadequate levels may result in severe cytopenia.

Rituximab (anti-CD20) can be associated with infusion reactions. In addition to administering this within in skilled infusion center, these reactions can be lessened by the use of pre-medications. There is a risk of hepatitis B reactivation with rituximab such that all patients should be screened for this infection prior to its use.

Infection represents a significant toxicity for all vasculitis patients treated with immunosuppressive therapy. Infections with *Pneumocystis jirovecii* and certain fungi can be seen even in the face of WBCs that are within normal limits, particularly in patients receiving glucocorticoids. All vasculitis patients who are receiving daily glucocorticoids in combination with another immunosuppressive agent should receive trimethoprim-sulfamethoxazole (TMP-SMX) or another prophylactic therapy to prevent *P. jirovecii* infection.

Finally, it should be emphasized that each patient is unique and requires individual decision-making. The above outline should serve as a framework to guide therapeutic approaches; however, flexibility should be practiced to provide maximal therapeutic efficacy with minimal toxic side effects in each patient.

TABLE 356-3

Conditions That Can Mimic Vasculitis

Infectious Diseases
Bacterial endocarditis Disseminated gonococcal infection Pulmonary histoplasmosis Coccidioidomycosis Syphilis Lyme disease Rocky Mountain spotted fever Whipple's disease
Coagulopathies/Thrombotic Microangiopathies
Antiphospholipid syndrome Thrombotic thrombocytopenic purpura
Neoplasms
Atrial myxoma Lymphoma Carcinomatosis
Drug Toxicity
Cocaine Levamisole Amphetamines Ergot alkaloids Methysergide Arsenic
Other
Sarcoidosis Atheroembolic disease Antiglomerular basement membrane disease (Goodpasture's syndrome) Amyloidosis Migraine

Fibromuscular dysplasia

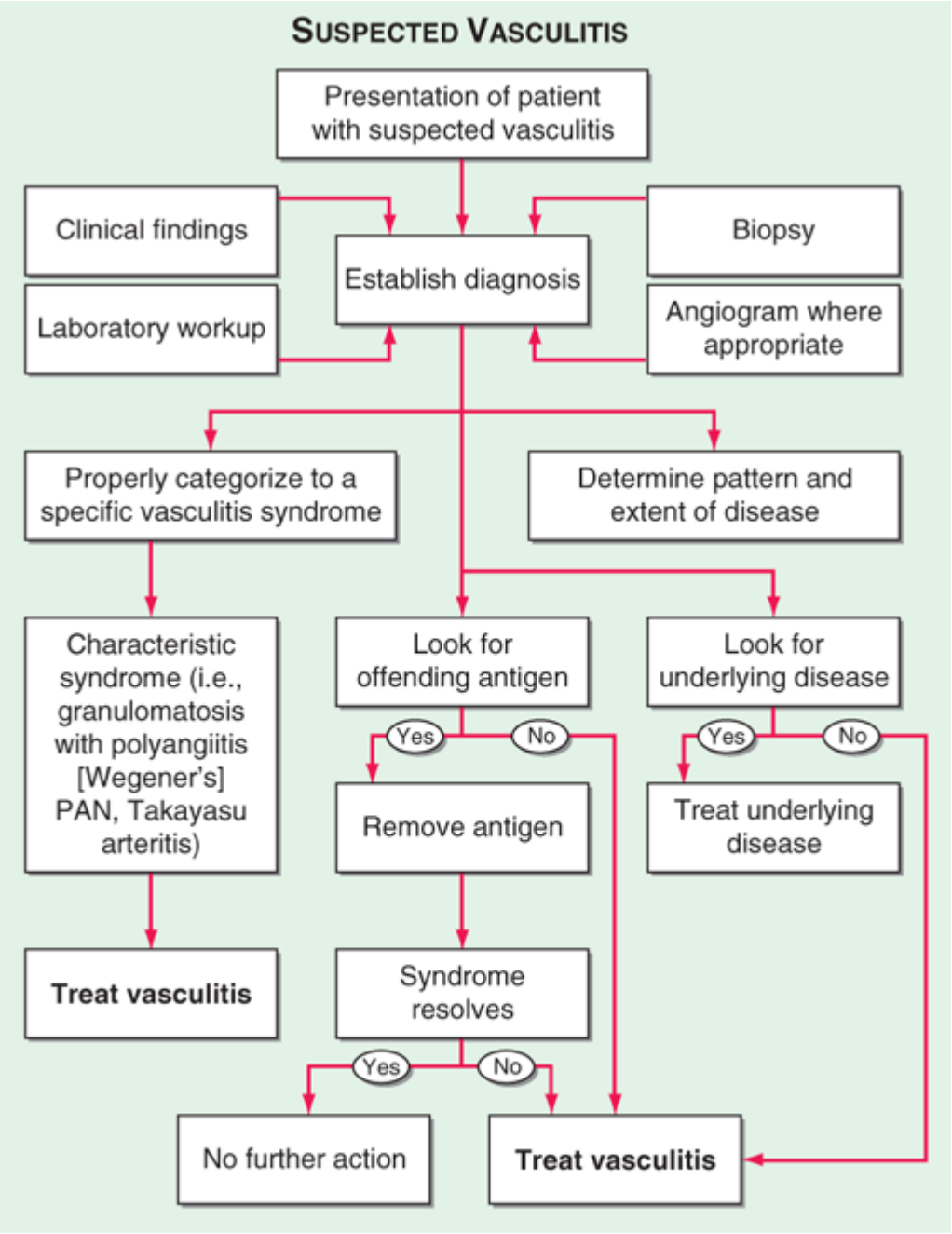
Heritable disorders of connective tissue

Segmental arterial mediolysis (SAM)

Reversible cerebral vasoconstrictive syndrome

FIGURE 356-1

Algorithm for the approach to a patient with suspected diagnosis of vasculitis. PAN, polyarteritis nodosa.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 356-4

Major Toxic Side Effects of Drugs Used in the Treatment of Systemic Small-Vessel Vasculitis

Glucocorticoids	
Osteoporosis Cataracts Glaucoma Diabetes mellitus Electrolyte abnormalities Metabolic abnormalities Suppression of inflammatory and immune responses leading to opportunistic infections Cushingoid features	Growth suppression in children Hypertension Avascular necrosis of bone Myopathy Alterations in mood Psychosis Pseudotumor cerebri Peptic ulcer diathesis Pancreatitis
Cyclophosphamide	
Bone marrow suppression Cystitis Bladder carcinoma Gonadal suppression Gastrointestinal intolerance	Hypogammaglobulinemia Pulmonary fibrosis Myelodysplasia Oncogenesis Teratogenicity Opportunistic infections
Methotrexate	
Gastrointestinal intolerance Stomatitis Bone marrow suppression Hepatotoxicity (may lead to fibrosis or cirrhosis)	Pneumonitis Teratogenicity Opportunistic infections
Azathioprine	
Gastrointestinal intolerance Bone marrow suppression Hepatotoxicity	Opportunistic infections Hypersensitivity
Rituximab	
Infusion reactions	Opportunistic infections

Progressive multifocal leuko-encephalopathy Mucocutaneous reactions	Hepatitis B reactivation Tumor lysis syndrome Late-onset neutropenia

GRANULOMATOSIS WITH POLYANGIITIS (WEGENER'S)

DEFINITION

Granulomatosis with polyangiitis (Wegener's) is a distinct clinicopathologic entity characterized by granulomatous vasculitis of the upper and lower respiratory tracts together with glomerulonephritis. In addition, variable degrees of disseminated vasculitis involving both small arteries and veins may occur.

INCIDENCE AND PREVALENCE

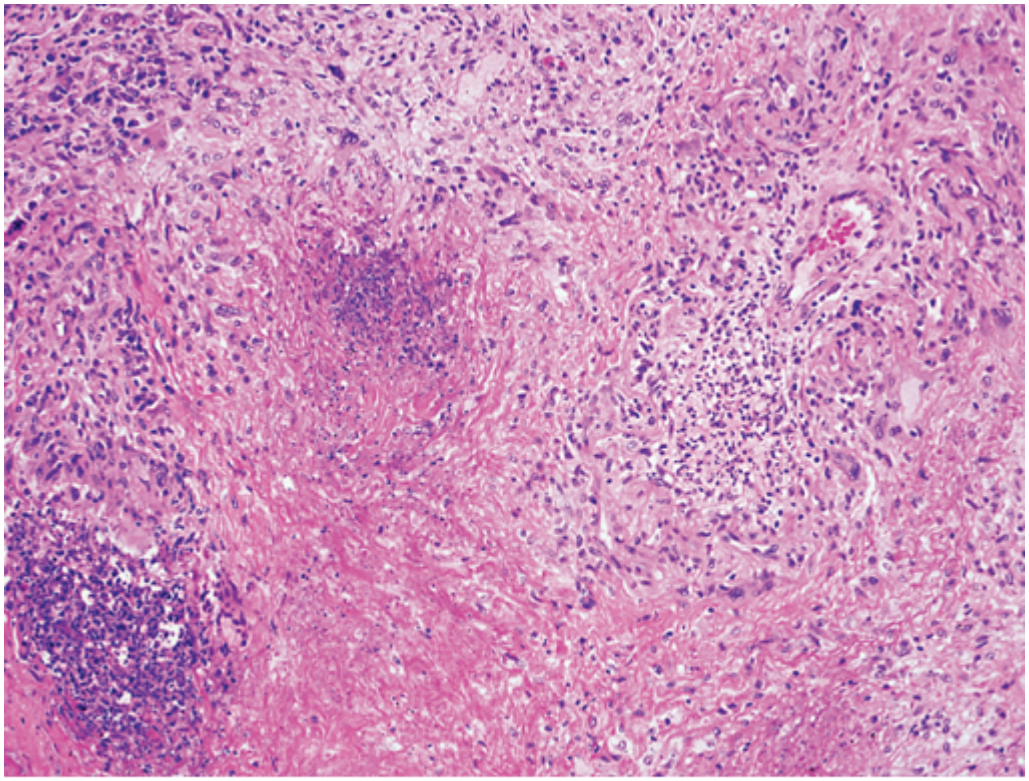
Granulomatosis with polyangiitis (Wegener's) is an uncommon disease with an estimated prevalence of 3 per 100,000. It is extremely rare in blacks compared with whites; the male-to-female ratio is 1:1. The disease can be seen at any age; ~15% of patients are <19 years of age, but only rarely does the disease occur before adolescence; the mean age of onset is ~40 years.

PATHOLOGY AND PATHOGENESIS

The histopathologic hallmarks of granulomatosis with polyangiitis (Wegener's) are necrotizing vasculitis of small arteries and veins together with granuloma formation, which may be either intravascular or extravascular **(Fig. 356-2)**. Lung involvement typically appears as multiple, bilateral, nodular cavitary infiltrates **(Fig. 356-3)**, which on biopsy almost invariably reveal the typical necrotizing granulomatous vasculitis. Upper airway lesions, particularly those in the sinuses and nasopharynx, typically reveal inflammation, necrosis, and granuloma formation, with or without vasculitis.

FIGURE 356-2

Lung histology in granulomatosis with polyangiitis (Wegener's). This area of geographic necrosis has a serpiginous border of histiocytes and giant cells surrounding a central necrotic zone. Vasculitis is also present with neutrophils and lymphocytes infiltrating the wall of a small arteriole (*upper right*). *(Courtesy of William D. Travis, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 356-3

Computed tomography scan of a patient with granulomatosis with polyangiitis (Wegener's). The patient developed multiple, bilateral, and cavitary infiltrates.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In its earliest form, renal involvement is characterized by a focal and segmental glomerulitis that may evolve into a rapidly progressive crescentic glomerulonephritis. Granuloma formation is only rarely seen on renal biopsy. In contrast to other forms of glomerulonephritis, evidence of immune complex deposition is not found in the renal lesion of granulomatosis with polyangiitis (Wegener's). In addition to the classic triad of

disease of the upper and lower respiratory tracts and kidney, virtually any organ can be involved with vasculitis, granuloma, or both.

The immunopathogenesis of this disease is unclear, although the involvement of upper airways and lungs with granulomatous vasculitis suggests an aberrant cell-mediated immune response to an exogenous or even endogenous antigen that enters through or resides in the upper airway. Chronic nasal carriage of *Staphylococcus aureus* has been reported to be associated with a higher relapse rate of granulomatosis with polyangiitis (Wegener's); however, there is no evidence for a role of this organism in the pathogenesis of the disease.

Peripheral blood mononuclear cells obtained from patients with granulomatosis with polyangiitis (Wegener's) manifest increased secretion of IFN-γ but not of IL-4, IL-5, or IL-10 compared to normal controls. In addition, TNF-α production from peripheral blood mononuclear cells and CD4+ T cells is elevated. Furthermore, monocytes from patients with granulomatosis with polyangiitis (Wegener's) produce increased amounts of IL-12. These findings indicate an unbalanced T_H1-type T cell cytokine pattern in this disease that may have pathogenic and perhaps ultimately therapeutic implications.

A high percentage of patients with granulomatosis with polyangiitis (Wegener's) develop ANCA, and these autoantibodies may play a role in the pathogenesis of this disease (see above).

CLINICAL AND LABORATORY MANIFESTATIONS

Involvement of the upper airways occurs in 95% of patients with granulomatosis with polyangiitis (Wegener's). Patients often present with severe upper respiratory tract findings such as paranasal sinus pain and drainage and purulent or bloody nasal discharge, with or without nasal mucosal ulceration **(Table 356-5)**. Nasal septal perforation may follow, leading to saddle nose deformity. Serous otitis media may occur as a result of eustachian tube blockage. Subglottic tracheal stenosis resulting from active disease or scarring occurs in ~16% of patients and may result in severe airway obstruction.

TABLE 356-5

Granulomatosis with Polyangiitis (Wegener's): Frequency of Clinical Manifestations in 158 Patients Studied at the National Institutes of Health

Manifestation	Percentage at Disease Onset	Percentage Throughout Course of Disease
Kidney		
Glomerulonephritis	18	77
Ear/Nose/Throat	73	92
Sinusitis	51	85
Nasal disease	36	68
Otitis media	25	44
Hearing loss	14	42
Subglottic stenosis	1	16
Ear pain	9	14
Oral lesions	3	10
Lung	45	85
Pulmonary infiltrates	25	66
Pulmonary nodules	24	58
Hemoptysis	12	30
Pleuritis	10	28
Eyes		
Conjunctivitis	5	18
Dacryocystitis	1	18
Scleritis	6	16
Proptosis	2	15
Eye pain	3	11
Visual loss	0	8
Retinal lesions	0	4
Corneal lesions	0	1
Iritis	0	2
Other[a]		

Manifestation	Percentage at Disease Onset	Percentage Throughout Course of Disease
Arthralgias/arthritis	32	67
Fever	23	50
Cough	19	46
Skin abnormalities	13	46
Weight loss (>10% body weight)	15	35
Peripheral neuropathy	1	15
Central nervous system disease	1	8
Pericarditis	2	6
Hyperthyroidism	1	3

[a]Fewer than 1% had parotid, pulmonary artery, breast, or lower genitourinary (urethra, cervix, vagina, testicular) involvement.

Source: GS Hoffman et al: Ann Intern Med 116:488, 1992.

Pulmonary involvement may be manifested as asymptomatic infiltrates or may be clinically expressed as cough, hemoptysis, dyspnea, and chest discomfort. It is present in 85–90% of patients. Endobronchial disease, either in its active form or as a result of fibrous scarring, may lead to obstruction with atelectasis.

Eye involvement (52% of patients) may range from a mild conjunctivitis to dacryocystitis, episcleritis, scleritis, granulomatous sclerouveitis, ciliary vessel vasculitis, and retroorbital mass lesions leading to proptosis.

Skin lesions (46% of patients) appear as papules, vesicles, palpable purpura, ulcers, or subcutaneous nodules; biopsy reveals vasculitis, granuloma, or both. Cardiac involvement (8% of patients) manifests as pericarditis, coronary vasculitis, or, rarely, cardiomyopathy. Nervous system manifestations (23% of patients) include cranial neuritis, mononeuritis multiplex, or, rarely, cerebral vasculitis and/or granuloma.

Renal disease (77% of patients) generally dominates the clinical picture and, if left untreated, accounts directly or indirectly for most of the mortality rate in this disease. Although it may smolder in some cases as a mild glomerulitis with proteinuria, hematuria, and red blood cell casts, it is clear that once clinically detectable renal functional impairment occurs, rapidly progressive renal failure usually ensues unless appropriate treatment is instituted.

While the disease is active, most patients have nonspecific symptoms and signs such as malaise, weakness, arthralgias, anorexia, and weight loss. Fever may indicate activity of the underlying disease but more often reflects secondary infection, usually of the upper airway.

Characteristic laboratory findings include a markedly elevated erythrocyte sedimentation rate (ESR), mild anemia and leukocytosis, mild hypergammaglobulinemia (particularly of the IgA class), and mildly elevated

rheumatoid factor. Thrombocytosis may be seen as an acute-phase reactant. Approximately 90% of patients with active granulomatosis with polyangiitis (Wegener's) have a positive antiproteinase-3 ANCA. However, in the absence of active disease, the sensitivity drops to ~60–70%. A small percentage of patients with granulomatosis with polyangiitis (Wegener's) may have antimyeloperoxidase rather than antiproteinase-3 antibodies, and up to 20% may lack ANCA.

Patients with granulomatosis with polyangiitis (Wegener's) have been found to have an increased incidence of venous thrombotic events. Although routine anticoagulation for all patients is not recommended, a heightened awareness for any clinical features suggestive of deep venous thrombosis or pulmonary emboli is warranted.

DIAGNOSIS

The diagnosis of granulomatosis with polyangiitis (Wegener's) is made by the demonstration of necrotizing granulomatous vasculitis on tissue biopsy in a patient with compatible clinical features. Pulmonary tissue offers the highest diagnostic yield, almost invariably revealing granulomatous vasculitis. Biopsy of upper airway tissue usually reveals granulomatous inflammation with necrosis but may not show vasculitis. Renal biopsy can confirm the presence of pauci-immune glomerulonephritis.

The specificity of a positive antiproteinase-3 ANCA for granulomatosis with polyangiitis (Wegener's) is very high, especially if active glomerulonephritis is present. However, the presence of ANCA should be adjunctive and, with rare exceptions, should not substitute for a tissue diagnosis. False-positive ANCA titers have been reported in certain infectious and neoplastic diseases.

In its typical presentation, the clinicopathologic complex of granulomatosis with polyangiitis (Wegener's) usually provides ready differentiation from other disorders. However, if all the typical features are not present at once, it needs to be differentiated from the other vasculitides, antiglomerular basement membrane disease (Goodpasture's syndrome) **(Chap. 308)**, relapsing polychondritis **(Chap. 359)**, tumors of the upper airway or lung, and infectious diseases such as histoplasmosis **(Chap. 207)**, mucocutaneous leishmaniasis **(Chap. 221)**, and rhinoscleroma **(Chap. 31)** as well as noninfectious granulomatous diseases.

Of particular note is the differentiation from other *midline destructive diseases*. These diseases lead to extreme tissue destruction and mutilation localized to the midline upper airway structures including the sinuses; erosion through the skin of the face commonly occurs, a feature that is extremely rare in granulomatosis with polyangiitis (Wegener's). Although blood vessels may be involved in the intense inflammatory reaction and necrosis, primary vasculitis is not seen. *Upper airway neoplasms* and specifically *extranodal natural killer (NK)/T cell lymphoma (nasal type)* are important causes of midline destructive disease. These lesions are diagnosed based on histology, which reveals polymorphous atypical lymphoid cells with an NK cell immunophenotype, typically Epstein-Barr virus **(Chap. 104)**. Such cases are treated based on their degree of dissemination, and localized lesions have responded to irradiation. Upper airway lesions should never be irradiated in granulomatosis with polyangiitis (Wegener's). Cocaine-induced tissue injury can be another important mimic of granulomatosis with polyangiitis (Wegener's) in patients who

present with isolated midline destructive disease. ANCA that target human neutrophil elastase can be found in patients with cocaine-induced midline destructive lesions and can complicate the differentiation from granulomatosis with polyangiitis (Wegener's). This has been further confounded by the high frequency of levamisole adulteration of cocaine, which can result in cutaneous infarction and serologic changes that may mimic vasculitis. Granulocytopenia is a common finding in levamisole-induced disease that would not be associated with granulomatosis with polyangiitis (Wegener's).

Granulomatosis with polyangiitis (Wegener's) must also be differentiated from *lymphomatoid granulomatosis*, which is an Epstein-Barr virus–positive B cell proliferation that is associated with an exuberant T cell reaction. Lymphomatoid granulomatosis is characterized by lung, skin, CNS, and kidney involvement in which atypical lymphocytoid and plasmacytoid cells infiltrate nonlymphoid tissue in an angioinvasive manner. In this regard, it clearly differs from granulomatosis with polyangiitis (Wegener's) in that it is not an inflammatory vasculitis in the classic sense but an angiocentric perivascular infiltration of atypical mononuclear cells. Up to 50% of patients may develop a true malignant lymphoma.

TREATMENT

TREATMENT

Granulomatosis with Polyangiitis (Wegener's)

Prior to the introduction of effective therapy, granulomatosis with polyangiitis (Wegener's) was universally fatal within a few months of diagnosis. Glucocorticoids alone led to some symptomatic improvement, with little effect on the ultimate course of the disease. The development of treatment with cyclophosphamide dramatically changed patient outcome such that marked improvement was seen in >90% of patients, complete remission in 75% of patients, and 5-year patient survival was seen in >80%.

Despite the ability to successfully induce remission, 50–70% of remissions are later associated with one or more relapses. The determination of relapse should be based on objective evidence of disease activity, taking care to rule out other features that may have a similar appearance such as infection, medication toxicity, or chronic disease sequelae. The ANCA titer can be misleading and should not be used to assess disease activity. Many patients who achieve remission continue to have elevated titers for years. Results from a large prospective study found that increases in ANCA were not associated with relapse and that only 43% relapsed within 1 year of an increase in ANCA levels. Thus, a rise in ANCA by itself is not a harbinger of immediate disease relapse and should not lead to reinstitution or increase in immunosuppressive therapy. Reinduction of remission after relapse is almost always achieved; however, a high percentage of patients ultimately have some degree of damage from irreversible features of their disease, such as varying degrees of renal insufficiency, hearing loss, tracheal stenosis, saddle nose deformity, and chronically impaired sinus function. Patients who developed irreversible renal failure but who achieved subsequent remission have undergone successful renal transplantation.

Treatment of granulomatosis with polyangiitis (Wegener's) is currently viewed as having two phases: *induction*, where active disease is put into remission, followed by *maintenance*. The decision regarding

which agents to use for induction and maintenance is based on disease severity together with individual patient factors that include contraindication, relapse history, and comorbidities.

CYCLOPHOSPHAMIDE INDUCTION FOR SEVERE DISEASE

For patients with severe disease, daily cyclophosphamide combined with glucocorticoids has been proven to effectively induce remission and prolong survival. At the initiation of therapy, glucocorticoids are usually given as prednisone, 1 mg/kg/d for the first month, followed by gradual tapering on an alternate-day or daily schedule with discontinuation after ~6–9 months.

Cyclophosphamide is given in doses of 2 mg/kg/d orally, but as it is renally eliminated, dosage reduction should be considered in patients with renal insufficiency. Some reports have indicated therapeutic success with less frequent and severe toxic side effects using IV cyclophosphamide. In a randomized trial, IV cyclophosphamide 15 mg/kg, three infusions given every 2 weeks, then every 3 weeks thereafter, was compared to cyclophosphamide 2 mg/kg daily given for 3 months followed by 1.5 mg/kg daily. Although IV cyclophosphamide was found to have a comparable rate of remission with a lower cumulative cyclophosphamide dose and occurrence of leukopenia, the use of a consolidation phase and an insufficient frequency of blood count monitoring may have negatively influenced the results in those who received daily cyclophosphamide. Of note in this study was that relapse occurred in 19% of those who received IV cyclophosphamide as compared to 9% who received daily oral administration. We continue to strongly favor daily rather than intermittent cyclophosphamide with utilization of blood count monitoring every 1–2 weeks (as discussed above) and limiting the duration of induction exposure to 3–6 months.

In patients with imminently life-threatening disease, such as rapidly progressive glomerulonephritis with a creatinine >4.0 mg/dL or pulmonary hemorrhage requiring mechanical ventilation, a regimen of daily cyclophosphamide and glucocorticoids is favored to induce remission. Adjunctive plasmapheresis has been used in fulminant disease, but as its role remains uncertain, this is currently being investigated in an international trial.

RITUXIMAB INDUCTION FOR SEVERE DISEASE

Rituximab is a chimeric monoclonal antibody directed against CD20 present on normal and malignant B lymphocytes that is U.S. Food and Drug Administration (FDA) approved for the treatment of granulomatosis with polyangiitis (Wegener's) and microscopic polyangiitis. In two randomized trials that enrolled ANCA-positive patients with severe active granulomatosis with polyangiitis (Wegener's) or microscopic polyangiitis, rituximab 375 mg/m^2 once a week for 4 weeks in combination with glucocorticoids was found to be as effective as cyclophosphamide with glucocorticoids for inducing disease remission. In the trial that also enrolled patients with relapsing disease, rituximab was found to be statistically superior to cyclophosphamide. Although rituximab does not have the bladder toxicity or infertility concerns, as can occur with cyclophosphamide, in both of the randomized trials, the rate of adverse events was similar in the rituximab and cyclophosphamide arms. In addition, there are no long-term safety data with rituximab in granulomatosis with polyangiitis (Wegener's) or microscopic polyangiitis.

The decision about whether to utilize cyclophosphamide or rituximab for remission induction must be individually based. Factors to consider include the severity of the disease, whether the patient has newly diagnosed or relapsing disease, medication contraindications, and individual patient factors particularly including fertility concerns.

REMISSION MAINTENANCE

The approach to remission maintenance is influenced by a number of elements that include the medication that is used for remission induction, the presence of prior relapses, the disease characteristics, medication contraindications, and individual patient factors. When cyclophosphamide is given for induction, it should be stopped after 3–6 months and switched to another agent for remission maintenance. The agents with which there has been the greatest published experience are methotrexate and azathioprine and most recently rituximab. Methotrexate is administered orally or subcutaneously starting at a dosage of 0.3 mg/kg as a single weekly dose, not to exceed 15 mg/week. If the treatment is well tolerated after 1–2 weeks, the dosage should be increased by 2.5 mg weekly up to a dosage of 20–25 mg/week and maintained at that level. Azathioprine, 2 mg/kg/d, has also proved effective in maintaining remission following induction with daily cyclophosphamide. In a randomized trial comparing methotrexate to azathioprine for remission maintenance, comparable rates of toxicity and relapse were seen. Therefore, the choice of agent is often based on toxicity profile, because methotrexate cannot be given to patients with renal insufficiency or chronic liver disease, as well as on other individual patient factors. In patients who are unable to receive methotrexate or azathioprine or who have relapsed through such treatment, mycophenolate mofetil, 1000 mg twice a day, may also sustain remission following cyclophosphamide induction. Rituximab 500 mg given intravenously every 6 months recently was compared to azathioprine given after intravenous cyclophosphamide induction in a randomized trial. Overall, a lower rate of relapse was observed with rituximab compared to azathioprine. However, the short duration of the trial and the chronic relapsing nature of these diseases continue to raise many questions as to the long-term role of rituximab for maintenance. Nonetheless, these data demonstrate that rituximab is an effective maintenance option which can be considered within the armamentarium.

For patients who receive rituximab for remission induction, the maintenance approach has not yet clearly been established. The options include to clinically observe the patient and retreat with rituximab should a relapse occur, or to pursue maintenance after rituximab with methotrexate, azathioprine, mycophenolate mofetil, or rituximab. Until further data become available, this decision is determined between the patient and physician.

The optimal duration of continuing maintenance therapy is uncertain. In the absence of toxicity, maintenance therapy is usually given for a minimum of 2 years past remission, after which time consideration can be given for tapering over a 6–12 month period until discontinuation. Patients with significant organ damage or a history of relapse may benefit from longer-term continuation of a maintenance agent.

OTHER BIOLOGIC THERAPIES

Etanercept, a dimeric fusion protein containing the 75-kDa TNF receptor bound to human IgG1, was not found to sustain remission when used adjunctively to standard therapy and should not be used in the treatment of granulomatosis with polyangiitis (Wegener's). Abatacept (CTLA4-Ig) was examined in an open-label pilot study of nonsevere relapsing disease with favorable results, but further investigation is needed before application to clinical practice. Blockade of the activity of complement C5a is also being investigated.

METHOTREXATE INDUCTION FOR NONSEVERE DISEASE

For selected patients whose disease is not immediately life threatening, methotrexate together with glucocorticoids given at the dosages described above may be considered as an alternative for induction therapy, which is then continued for maintenance.

TRIMETHOPRIM-SULFAMETHOXAZOLE

Although certain reports have indicated that trimethoprim-sulfamethoxazole (TMP-SMX) may be of benefit in the treatment of granulomatosis with polyangiitis (Wegener's) isolated to the sinonasal tissues, it should never be used alone to treat active granulomatosis with polyangiitis (Wegener's) outside of the upper airway such as in patients with renal or pulmonary disease. In a study examining the effect of TMP-SMX on relapse, decreased relapses were shown only with regard to upper airway disease, and no differences in major organ relapses were observed.

ORGAN-SPECIFIC TREATMENT

Not all manifestations of granulomatosis with polyangiitis (Wegener's) require or respond to immunosuppressive therapy. In managing non–major organ disease, such as that isolated to the sinus, joints, or skin, the risks of treatment should be carefully weighed against the benefits. Treatment with cyclophosphamide is rare if ever justified for the treatment of isolated sinus disease in granulomatosis with polyangiitis (Wegener's). Differentiation of active disease from damage is also important. Subglottic stenosis is an example of a disease manifestation that can often scar and responds optimally to nonmedical intervention rather than systemic immunosuppressive treatment.

MICROSCOPIC POLYANGIITIS

DEFINITION

The term *microscopic polyarteritis* was introduced into the literature by Davson in 1948 in recognition of the presence of glomerulonephritis in patients with polyarteritis nodosa. In 1992, the Chapel Hill Consensus Conference on the Nomenclature of Systemic Vasculitis adopted the term *microscopic polyangiitis* to connote a necrotizing vasculitis with few or no immune complexes affecting small vessels (capillaries, venules, or arterioles). Glomerulonephritis is very common in microscopic polyangiitis, and pulmonary capillaritis often occurs. The absence of granulomatous inflammation in microscopic polyangiitis is said to differentiate it from granulomatosis with polyangiitis (Wegener's).

INCIDENCE AND PREVALENCE

The incidence of microscopic polyangiitis is estimated to be 3–5/100,000. The mean age of onset is ~57 years, and males are slightly more frequently affected than females.

PATHOLOGY AND PATHOGENESIS

The vasculitis seen in microscopic polyangiitis has a predilection to involve capillaries and venules in addition to small- and medium-sized arteries. Immunohistochemical staining reveals a paucity of immunoglobulin deposition in the vascular lesion of microscopic polyangiitis, suggesting that immune-complex formation does not play a role in the pathogenesis of this syndrome. The renal lesion seen in microscopic polyangiitis is identical to that of granulomatosis with polyangiitis (Wegener's). Like granulomatosis with polyangiitis (Wegener's), microscopic polyangiitis is highly associated with the presence of ANCA, which may play a role in pathogenesis of this syndrome (see above).

CLINICAL AND LABORATORY MANIFESTATIONS

Because of its predilection to involve the small vessels, microscopic polyangiitis and granulomatosis with polyangiitis (Wegener's) share similar clinical features. Disease onset may be gradual, with initial symptoms of fever, weight loss, and musculoskeletal pain; however, it is often acute. Glomerulonephritis occurs in at least 79% of patients and can be rapidly progressive, leading to renal failure. Hemoptysis may be the first symptom of alveolar hemorrhage, which occurs in 12% of patients. Other manifestations include mononeuritis multiplex and gastrointestinal tract and cutaneous vasculitis. Upper airway disease and pulmonary nodules are not typically found in microscopic polyangiitis and, if present, suggest granulomatosis with polyangiitis (Wegener's).

Features of inflammation may be seen, including an elevated ESR, anemia, leukocytosis, and thrombocytosis. ANCA are present in 75% of patients with microscopic polyangiitis, with antimyeloperoxidase antibodies being the predominant ANCA associated with this disease.

DIAGNOSIS

The diagnosis is based on histologic evidence of vasculitis or pauci-immune glomerulonephritis in a patient with compatible clinical features of multisystem disease. Although microscopic polyangiitis is strongly ANCA-associated, no studies have as yet established the sensitivity and specificity of ANCA in this disease.

TREATMENT

TREATMENT

Microscopic Polyangiitis

The 5-year survival rate for patients with treated microscopic polyangiitis is 74%, with disease-related mortality occurring from alveolar hemorrhage or gastrointestinal, cardiac, or renal disease. Studies on treatment have come from trials that have included patients with granulomatosis with polyangiitis

(Wegener's) or microscopic polyangiitis. Currently, the treatment approach for microscopic polyangiitis is the same as is used for granulomatosis with polyangiitis (Wegener's) (see "Granulomatosis with Polyangiitis [Wegener's]" for a detailed description of this therapeutic regimen), and patients with immediately life-threatening disease should be treated with the combination of prednisone and daily cyclophosphamide or rituximab. Disease relapse has been observed in at least 34% of patients. Treatment for such relapses would be similar to that used at the time of initial presentation and based on site and severity of disease.

EOSINOPHILIC GRANULOMATOSIS WITH POLYANGIITIS (CHURG-STRAUSS)

DEFINITION

Eosinophilic granulomatosis with polyangiitis (Churg-Strauss) was described in 1951 by Churg and Strauss and is characterized by asthma, peripheral and tissue eosinophilia, extravascular granuloma formation, and vasculitis of multiple organ systems.

INCIDENCE AND PREVALENCE

Eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is an uncommon disease with an estimated annual incidence of 1–3 per million. The disease can occur at any age with the possible exception of infants. The mean age of onset is 48 years, with a female-to-male ratio of 1.2:1.

PATHOLOGY AND PATHOGENESIS

The necrotizing vasculitis of eosinophilic granulomatosis with polyangiitis (Churg-Strauss) involves small- and medium-sized muscular arteries, capillaries, veins, and venules. A characteristic histopathologic feature of eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is granulomatous reactions that may be present in the tissues or even within the walls of the vessels themselves. These are usually associated with infiltration of the tissues with eosinophils. This process can occur in any organ in the body; lung involvement is predominant, with skin, cardiovascular system, kidney, peripheral nervous system, and gastrointestinal tract also commonly involved. Although the precise pathogenesis of this disease is uncertain, its strong association with asthma and its clinicopathologic manifestations, including eosinophilia, granuloma, and vasculitis, point to aberrant immunologic phenomena.

CLINICAL AND LABORATORY MANIFESTATIONS

Patients with eosinophilic granulomatosis with polyangiitis (Churg-Strauss) often exhibit nonspecific manifestations such as fever, malaise, anorexia, and weight loss, which are characteristic of a multisystem disease. The pulmonary findings in eosinophilic granulomatosis with polyangiitis (Churg-Strauss) clearly dominate the clinical picture with severe asthmatic attacks and the presence of pulmonary infiltrates. Mononeuritis multiplex is the second most common manifestation and occurs in up to 72% of patients.

Allergic rhinitis and sinusitis develop in up to 61% of patients and are often observed early in the course of disease. Clinically recognizable heart disease occurs in ~14% of patients and is an important cause of mortality. Skin lesions occur in ~51% of patients and include purpura in addition to cutaneous and subcutaneous nodules. The renal disease in eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is less common and generally less severe than that of granulomatosis with polyangiitis and microscopic polyangiitis.

The characteristic laboratory finding in virtually all patients with eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is a striking eosinophilia, which reaches levels >1000 cells/μL in >80% of patients. Evidence of inflammation as evidenced by elevated ESR, fibrinogen, or α_2-globulins can be found in 81% of patients. The other laboratory findings reflect the organ systems involved. Approximately 48% of patients with eosinophilic granulomatosis with polyangiitis (Churg-Strauss) have circulating ANCA that is usually antimyeloperoxidase.

DIAGNOSIS

Although the diagnosis of eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is optimally made by biopsy in a patient with the characteristic clinical manifestations (see above), histologic confirmation can be challenging because the pathognomonic features often do not occur simultaneously. In order to be diagnosed with eosinophilic granulomatosis with polyangiitis (Churg-Strauss), a patient should have evidence of asthma, peripheral blood eosinophilia, and clinical features consistent with vasculitis.

TREATMENT

TREATMENT

Eosinophilic Granulomatosis with Polyangiitis (Churg-Strauss)

The prognosis of untreated eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is poor, with a reported 5-year survival of 25%. With treatment, prognosis is favorable, with one study finding a 78-month actuarial survival rate of 72%. Myocardial involvement is the most frequent cause of death and is responsible for 39% of patient mortality. Echocardiography should be performed in all newly diagnosed patients because this may influence therapeutic decisions.

Glucocorticoids alone appear to be effective in many patients. Dosage tapering is often limited by asthma, and many patients require low-dose prednisone for persistent asthma many years after clinical recovery from vasculitis. In patients who present with fulminant multisystem disease, particularly cardiac involvement, the treatment of choice is a combined regimen of daily cyclophosphamide and prednisone followed by azathioprine or methotrexate (see "Granulomatosis with Polyangiitis [Wegener's]" for a detailed description of this therapeutic regimen).

Mepolizumab (anti-IL-5 antibody) was studied in a randomized trial and found to be more effective than placebo in eosinophilic granulomatosis with polyangiitis (Churg-Strauss). Patients with life-threatening

eosinophilic granulomatosis with polyangiitis (Churg-Strauss) were excluded from the mepolizumab trial and should continue to be treated with cyclophosphamide and glucocorticoids. Mepoluzimab is FDA approved for the treatment of severe eosinophilic asthma and may particularly have a role in the setting of relapsing or resistant asthma in eosinophilic granulomatosis with polyangiitis (Churg-Strauss). Rituximab has been examined only in small retrospective series, primarily in patients who have active disease despite conventional agents or who are intolerant of these medications.

POLYARTERITIS NODOSA

DEFINITION

Polyarteritis nodosa was described in 1866 by Kussmaul and Maier. It is a multisystem, necrotizing vasculitis of small- and medium-sized muscular arteries in which involvement of the renal and visceral arteries is characteristic. Polyarteritis nodosa does not involve pulmonary arteries, although bronchial vessels may be involved; granulomas, significant eosinophilia, and an allergic diathesis are not observed.

INCIDENCE AND PREVALENCE

It is difficult to establish an accurate incidence of polyarteritis nodosa because previous reports have included polyarteritis nodosa and microscopic polyangiitis as well as other related vasculitides. Polyarteritis nodosa, as currently defined, is felt to be a very uncommon disease.

PATHOLOGY AND PATHOGENESIS

The vascular lesion in polyarteritis nodosa is a necrotizing inflammation of small- and medium-sized muscular arteries. The lesions are segmental and tend to involve bifurcations and branchings of arteries. They may spread circumferentially to involve adjacent veins. However, involvement of venules is not seen in polyarteritis nodosa and, if present, suggests microscopic polyangiitis (see below). In the acute stages of disease, polymorphonuclear neutrophils infiltrate all layers of the vessel wall and perivascular areas, which results in intimal proliferation and degeneration of the vessel wall. Mononuclear cells infiltrate the area as the lesions progress to the subacute and chronic stages. Fibrinoid necrosis of the vessels ensues with compromise of the lumen, thrombosis, infarction of the tissues supplied by the involved vessel, and, in some cases, hemorrhage. As the lesions heal, there is collagen deposition, which may lead to further occlusion of the vessel lumen. Aneurysmal dilations up to 1 cm in size along the involved arteries are characteristic of polyarteritis nodosa. Granulomas and substantial eosinophilia with eosinophilic tissue infiltrations are not characteristically found and suggest eosinophilic granulomatosis with polyangiitis (Churg-Strauss) (see above).

Multiple organ systems are involved, and the clinicopathologic findings reflect the degree and location of vessel involvement and the resulting ischemic changes. As mentioned above, pulmonary arteries are not involved in polyarteritis nodosa, and bronchial artery involvement is uncommon. The pathology in the kidney in classic polyarteritis nodosa is that of arteritis without glomerulonephritis. In patients with

significant hypertension, typical pathologic features of glomerulosclerosis may be seen. In addition, pathologic sequelae of hypertension may be found elsewhere in the body.

The presence of a polyarteritis nodosa–like vasculitis in patients with hepatitis B together with the isolation of circulating immune complexes composed of hepatitis B antigen and immunoglobulin and the demonstration by immunofluorescence of hepatitis B antigen, IgM, and complement in the blood vessel walls strongly suggest the role of immunologic phenomena in the pathogenesis of this disease. A polyarteritis nodosa–like vasculitis has also been reported in patients with hepatitis C. Hairy cell leukemia can be associated with polyarteritis nodosa; the pathogenic mechanisms of this association are unclear.

CLINICAL AND LABORATORY MANIFESTATIONS

Nonspecific signs and symptoms are the hallmarks of polyarteritis nodosa. Fever, weight loss, and malaise are present in over one-half of cases. Patients usually present with vague symptoms such as weakness, malaise, headache, abdominal pain, and myalgias that can rapidly progress to a fulminant illness. Specific complaints related to the vascular involvement within a particular organ system may also dominate the presenting clinical picture as well as the entire course of the illness **(Table 356-6)**. In polyarteritis nodosa, renal involvement most commonly manifests as hypertension, renal insufficiency, or hemorrhage due to microaneurysms.

TABLE 356-6

Clinical Manifestations Related to Organ System Involvement in Polyarteritis Nodosa

Organ System	Percent Incidence	Clinical Manifestations
Renal	60	Renal failure, hypertension
Musculoskeletal	64	Arthritis, arthralgia, myalgia
Peripheral nervous system	51	Peripheral neuropathy, mononeuritis multiplex
Gastrointestinal tract	44	Abdominal pain, nausea and vomiting, bleeding, bowel infarction and perforation, cholecystitis, hepatic infarction, pancreatic infarction
Skin	43	Rash, purpura, nodules, cutaneous infarcts, livedo reticularis, Raynaud's phenomenon
Cardiac	36	Congestive heart failure, myocardial infarction, pericarditis
Genitourinary	25	Testicular, ovarian, or epididymal pain
Central nervous system	23	Cerebral vascular accident, altered mental status, seizure

Source: From TR Cupps, AS Fauci: *The Vasculitides.* Philadelphia, Saunders, 1981.

There are no diagnostic serologic tests for polyarteritis nodosa. In >75% of patients, the leukocyte count is elevated with a predominance of neutrophils. Eosinophilia is seen only rarely and, when present at high levels, suggests the diagnosis of eosinophilic granulomatosis with polyangiitis (Churg-Strauss). The anemia of chronic disease may be seen, and an elevated ESR is almost always present. Other common laboratory findings reflect the particular organ involved. Hypergammaglobulinemia may be present, and all patients should be screened for hepatitis B and C. Antibodies against myeloperoxidase or proteinase-3 (ANCA) are rarely found in patients with polyarteritis nodosa.

DIAGNOSIS

The diagnosis of polyarteritis nodosa is based on the demonstration of characteristic findings of vasculitis on biopsy material of involved organs. In the absence of easily accessible tissue for biopsy, the arteriographic demonstration of involved vessels, particularly in the form of aneurysms of small- and medium-sized arteries

in the renal, hepatic, and visceral vasculature, is sufficient to make the diagnosis. This should consist of a catheter-directed dye arteriogram because magnetic resonance and computed tomography arteriograms do not have sufficient resolution at the current time to visualize the vessels affected in polyarteritis nodosa. Aneurysms of vessels are not pathognomonic of polyarteritis nodosa; furthermore, aneurysms need not always be present, and arteriographic findings may be limited to stenotic segments and obliteration of vessels. Biopsy of symptomatic organs such as nodular skin lesions, painful testes, and nerve/muscle provides the highest diagnostic yields.

TREATMENT

TREATMENT

Polyarteritis Nodosa

The prognosis of untreated polyarteritis nodosa is extremely poor, with a reported 5-year survival rate between 10 and 20%. Death usually results from gastrointestinal complications, particularly bowel infarcts and perforation, and cardiovascular causes. Intractable hypertension often compounds dysfunction in other organ systems, such as the kidneys, heart, and CNS, leading to additional late morbidity and mortality in polyarteritis nodosa. With the introduction of treatment, survival rate has increased substantially. Favorable therapeutic results have been reported in polyarteritis nodosa with the combination of prednisone and cyclophosphamide (see "Granulomatosis with Polyangiitis [Wegener's]" for a detailed description of this therapeutic regimen). In less severe cases of polyarteritis nodosa, glucocorticoids alone have resulted in disease remission. In patients with hepatitis B who have a polyarteritis nodosa–like vasculitis, antiviral therapy represents an important part of therapy and has been used in combination with glucocorticoids and plasma exchange. Careful attention to the treatment of hypertension can lessen the acute and late morbidity and mortality rates associated with renal, cardiac, and CNS complications of polyarteritis nodosa. Following successful treatment, relapse of polyarteritis nodosa has been estimated to occur in 10–20% of patients.

GIANT CELL ARTERITIS AND POLYMYALGIA RHEUMATICA

DEFINITION

Giant cell arteritis, historically referred to as *temporal arteritis*, is an inflammation of medium- and large-sized arteries. It characteristically involves one or more branches of the carotid artery, particularly the temporal artery. However, it is a systemic disease that can involve arteries in multiple locations, particularly the aorta and its main branches.

Giant cell arteritis is closely associated with *polymyalgia rheumatica*, which is characterized by stiffness, aching, and pain in the muscles of the neck, shoulders, lower back, hips, and thighs. Most commonly, polymyalgia rheumatica occurs in isolation, but it may be seen in 40–50% of patients with giant cell arteritis. In addition, ~10–20% of patients who initially present with features of isolated polymyalgia rheumatica later go on to develop giant cell arteritis. This strong clinical association together with data from pathophysiologic

studies has increasingly supported that giant cell arteritis and polymyalgia rheumatica represent differing clinical spectrums of a single disease process.

INCIDENCE AND PREVALENCE

Giant cell arteritis occurs almost exclusively in individuals aged >50 years. It is more common in women than in men and is rare in blacks. The incidence of giant cell arteritis varies widely in different studies and in different geographic regions. A high incidence has been found in Scandinavia and in regions of the United States with large Scandinavian populations, compared to a lower incidence in southern Europe. The annual incidence rates in individuals aged ≥50 years range from 6.9 to 32.8 per 100,000 population. Familial aggregation has been reported, as has an association with HLA-DR4. In addition, genetic linkage studies have demonstrated an association of giant cell arteritis with alleles at the HLA-DRB1 locus, particularly HLA-DRB1*04 variants. In Olmsted County, Minnesota, the annual incidence of polymyalgia rheumatica in individuals aged ≥50 years is 58.7 per 100,000 population.

PATHOLOGY AND PATHOGENESIS

Although the temporal artery is most frequently involved in giant cell arteritis, patients often have a systemic vasculitis of multiple medium- and large-sized arteries, which may go undetected. Histopathologically, the disease is a panarteritis with inflammatory mononuclear cell infiltrates within the vessel wall with frequent giant cell formation. There is proliferation of the intima and fragmentation of the internal elastic lamina. Pathophysiologic findings in organs result from the ischemia related to the involved vessels.

Experimental data support that giant cell arteritis is an antigen-driven disease in which activated T lymphocytes, macrophages, and dendritic cells play a critical role in the disease pathogenesis. Sequence analysis of the T cell receptor of tissue-infiltrating T cells in lesions of giant cell arteritis indicates restricted clonal expansion, suggesting the presence of an antigen residing in the arterial wall. Giant cell arteritis is believed to be initiated in the adventitia where CD4+ T cells enter through the vasa vasorum, become activated, and orchestrate macrophage differentiation. T cells recruited to vasculitic lesions in patients with giant cell arteritis produce predominantly IL-2 and IFN-γ, and the latter has been suggested to be involved in the progression to overt arteritis. Laboratory-based data demonstrate that at least two separate lineages of CD4 T cells––IFN-γ-producing T_H1 cells and IL-17-producing T_H17 cells—participate in vascular inflammation and may have differing levels of responsiveness to glucocorticoids.

CLINICAL AND LABORATORY MANIFESTATIONS

Giant cell arteritis is most commonly characterized clinically by the complex of fever, anemia, high ESR, and headaches in a patient aged >50 years. Other phenotypic manifestations include features of systemic inflammation, including malaise, fatigue, anorexia, weight loss, sweats, arthralgias, polymyalgia rheumatica, or large-vessel disease.

In patients with involvement of the cranial arteries, headache is the predominant symptom and may be associated with a tender, thickened, or nodular artery, which may pulsate early in the disease but may become occluded later. Scalp pain and claudication of the jaw and tongue may occur. A well-recognized and dreaded complication of giant cell arteritis, particularly in untreated patients, is ischemic optic neuropathy, which may lead to serious visual symptoms, even sudden blindness in some patients. However, most patients have complaints relating to the head or eyes before visual loss. Attention to such symptoms with institution of appropriate therapy (see below) will usually avoid this complication. Other cranial ischemic complications include strokes and scalp or tongue infarction.

Up to one-third of patients can have large-vessel disease that can be the primary presentation of giant cell arteritis or can emerge at a later point in patients who have had previous cranial arteritis features or polymyalgia rheumatica. Manifestations of large-vessel disease can include subclavian artery stenosis that can present as arm claudication or aortic aneurysms involving the thoracic and to a lesser degree the abdominal aorta, which carry risks of rupture or dissection.

Characteristic laboratory findings in addition to the elevated ESR include a normochromic or slightly hypochromic anemia. Liver function abnormalities are common, particularly increased alkaline phosphatase levels. Increased levels of IgG and complement have been reported. Levels of enzymes indicative of muscle damage such as serum creatine kinase are not elevated.

DIAGNOSIS

The diagnosis of giant cell arteritis and its associated clinicopathologic syndrome can often be suggested clinically by the demonstration of the complex of fever, anemia, and high ESR with or without symptoms of polymyalgia rheumatica in a patient >50 years. The diagnosis can be confirmed by biopsy of the temporal artery but may not be positive in all patients due to patchy histologic findings. Since involvement of the vessel may be segmental, positive yield is increased by obtaining a biopsy segment of 3–5 cm together with serial sectioning of biopsy specimens. Ultrasonography of the temporal artery has been reported to be helpful in diagnosis and has been increasingly used by some physicians. Therapy should not be delayed pending the performance of diagnostic studies. In this regard, it has been reported that temporal artery biopsies may show vasculitis even after ~14 days of glucocorticoid therapy. A dramatic clinical response to a trial of glucocorticoid therapy can further support the diagnosis.

Large-vessel disease may be suggested by symptoms and findings on physical examination such as diminished pulses or bruits. It is confirmed by vascular imaging, most commonly through magnetic resonance or computed tomography.

Isolated polymyalgia rheumatica is a clinical diagnosis made by the presence of typical symptoms of stiffness, aching, and pain in the muscles of the hip and shoulder girdle, an increased ESR, the absence of clinical features suggestive of giant cell arteritis, and a prompt therapeutic response to low-dose prednisone.

TREATMENT

TREATMENT

Giant Cell Arteritis and Polymyalgia Rheumatica

Acute disease-related mortality directly from giant cell arteritis is uncommon, with fatalities occurring from cerebrovascular events or myocardial infarction. However, patients are at risk of late mortality from aortic aneurysm rupture or dissection as patients with giant cell arteritis are 18 times more likely to develop thoracic aortic aneurysms than the general population.

The goals of treatment in giant cell arteritis are to reduce symptoms and, most importantly, to prevent visual loss. The treatment approach for cranial and large-vessel disease in giant cell arteritis is currently the same. Giant cell arteritis and its associated symptoms are exquisitely sensitive to glucocorticoid therapy. Treatment should begin with prednisone, 40–60 mg/d for ~1 month, followed by a gradual tapering. When ocular signs and symptoms occur, consideration should be given for the use of methylprednisolone 1000 mg daily for 3 days to protect remaining vision. Although the optimal duration of glucocorticoid therapy has not been established, most series have found that patients require treatment for ≥2 years. Symptom recurrence during prednisone tapering develops in 60–85% of patients with giant cell arteritis, requiring a dosage increase. The ESR can serve as a useful indicator of inflammatory disease activity in monitoring and tapering therapy and can be used to judge the pace of the tapering schedule. However, minor increases in the ESR can occur as glucocorticoids are being tapered and do not necessarily reflect an exacerbation of arteritis, particularly if the patient remains symptom-free. Under these circumstances, the tapering should continue with caution. Glucocorticoid toxicity occurs in 35–65% of patients and represents an important cause of patient morbidity. Aspirin 81 mg daily has been found to reduce the occurrence of cranial ischemic complications in giant cell arteritis and should be given in addition to glucocorticoids in patients who do not have contraindications. The use of weekly methotrexate as a glucocorticoid-sparing agent has been examined in two randomized placebo-controlled trials that reached conflicting conclusions. Infliximab, a monoclonal antibody to TNF, was studied in a randomized trial and was not found to provide benefit.

Tocilizumab (anti-IL-6 receptor) was found to be more effective than prednisone alone in a recent large-scale randomized trial of giant cell arteritis and was FDA approved for this indication. It is used adjunctively to glucocorticoids and its optimal role in patient management will continue to be defined over time. The side effect profile of tocilizumab which includes leukopenia, thrombocytopenia, transaminnase elevation, and hyperlipidemia must be weighed. Because of the risk of gastrointestinal perforation, patients with prior diverticulitis were excluded from the giant cell arteritis trial. By nature of its mechanism, tocilizumab impacts laboratory parameters of the acute phase response which will eliminate the ability to utilize these in disease activity assessment.

Abatacept (CTLA4-Ig) was examined in a small randomized trial in giant cell arteritis and demonstrated greater efficacy than glucocorticoids alone.

Patients with isolated polymyalgia rheumatica respond promptly to prednisone, which can be started at a lower dose of 10–20 mg/d. Similar to giant cell arteritis, the ESR can serve as a useful indicator in monitoring and prednisone reduction. Recurrent polymyalgia symptoms develop in the majority of patients during

prednisone tapering. One study of weekly methotrexate found that the use of this drug reduced the prednisone dose on average by only 1 mg and did not decrease prednisone-related side effects. A randomized trial in polymyalgia rheumatica did not find infliximab to lessen relapse or glucocorticoid requirements.

TAKAYASU ARTERITIS

DEFINITION

Takayasu arteritis is an inflammatory and stenotic disease of medium- and large-sized arteries characterized by a strong predilection for the aortic arch and its branches.

INCIDENCE AND PREVALENCE

Takayasu arteritis is an uncommon disease with an estimated annual incidence rate of 1.2–2.6 cases per million. It is most prevalent in adolescent girls and young women. Although it is more common in Asia, it is neither racially nor geographically restricted.

PATHOLOGY AND PATHOGENESIS

The disease involves medium- and large-sized arteries, with a strong predilection for the aortic arch and its branches; the pulmonary artery may also be involved. The most commonly affected arteries seen by arteriography are listed in Table 356-7. The involvement of the major branches of the aorta is much more marked at their origin than distally. The disease is a panarteritis with inflammatory mononuclear cell infiltrates and occasionally giant cells. There are marked intimal proliferation and fibrosis, scarring and vascularization of the media, and disruption and degeneration of the elastic lamina. Narrowing of the lumen occurs with or without thrombosis. The vasa vasorum are frequently involved. Pathologic changes in various organs reflect the compromise of blood flow through the involved vessels.

TABLE 356-7

Frequency of Arteriographic Abnormalities and Potential Clinical Manifestations of Arterial Involvement in Takayasu Arteritis

Artery	Percentage of Arteriographic Abnormalities	Potential Clinical Manifestations
Subclavian	93	Arm claudication, Raynaud's phenomenon
Common carotid	58	Visual changes, syncope, transient ischemic attacks, stroke
Abdominal aorta[a]	47	Abdominal pain, nausea, vomiting
Renal	38	Hypertension, renal failure
Aortic arch or root	35	Aortic insufficiency, congestive heart failure
Vertebral	35	Visual changes, dizziness
Coeliac axis[a]	18	Abdominal pain, nausea, vomiting
Superior mesenteric[a]	18	Abdominal pain, nausea, vomiting
Iliac	17	Leg claudication
Pulmonary	10–40	Atypical chest pain, dyspnea
Coronary	<10	Chest pain, myocardial infarction

[a]Arteriographic lesions at these locations are usually asymptomatic but may potentially cause these symptoms.

Source: G Kerr et al: Ann Intern Med 120:919, 1994.

Immunopathogenic mechanisms, the precise nature of which is uncertain, are suspected in this disease. As with several of the vasculitis syndromes, circulating immune complexes have been demonstrated, but their pathogenic significance is unclear.

CLINICAL AND LABORATORY MANIFESTATIONS

Takayasu arteritis is a systemic disease with generalized as well as vascular symptoms. The generalized symptoms include malaise, fever, night sweats, arthralgias, anorexia, and weight loss, which may occur months before vessel involvement is apparent. These symptoms may merge into those related to vascular compromise and organ ischemia. Pulses are commonly absent in the involved vessels, particularly the subclavian artery. The frequency of arteriographic abnormalities and the potentially associated clinical manifestations are listed in Table 356-7. Hypertension occurs in 32–93% of patients and contributes to renal, cardiac, and cerebral injury.

Characteristic laboratory findings include an elevated ESR, mild anemia, and elevated immunoglobulin levels.

DIAGNOSIS

The diagnosis of Takayasu arteritis should be suspected strongly in a young woman who develops a decrease or absence of peripheral pulses, discrepancies in blood pressure, and arterial bruits. The diagnosis is confirmed by the characteristic pattern on arteriography, which includes irregular vessel walls, stenosis, poststenotic dilation, aneurysm formation, occlusion, and evidence of increased collateral circulation. Complete aortic arteriography by catheter-directed dye arteriography or magnetic resonance arteriography should be obtained to fully delineate the distribution and degree of arterial disease. Histopathologic demonstration of vessel wall inflammation that is predominantly lymphocytic with granuloma formation and giant cells involving the media and adventitia adds confirmatory data; however, tissue is rarely readily available for examination. IgG4-related disease is a potential cause of aortitis and periaortitis that is histologically differentiated from Takayasu arteritis by a dense lymphoplasmacytic infiltrate rich in IgG4-positive plasma cells, a storiform pattern of fibrosis, and obliterative phlebitis.

TREATMENT

TREATMENT

Takayasu Arteritis

The long-term outcome of patients with Takayasu arteritis has varied widely between studies. Although two North American reports found overall survival to be ≥94%, the 5-year mortality rate from other studies has ranged from 0 to 35%. Disease-related mortality most often occurs from congestive heart failure, cerebrovascular events, myocardial infarction, aneurysm rupture, or renal failure. Even in the absence of life-threatening disease, Takayasu arteritis can be associated with significant morbidity. The course of the disease is variable, and although spontaneous remissions may occur, Takayasu arteritis is most often chronic and relapsing. Although glucocorticoid therapy in doses of 40–60 mg prednisone per day alleviates symptoms, there are no convincing studies that indicate that it increases survival. The combination of glucocorticoid therapy for acute signs and symptoms and an aggressive surgical and/or arterioplastic approach to stenosed vessels has markedly improved outcome and decreased morbidity by lessening the

risk of stroke, correcting hypertension due to renal artery stenosis, and improving blood flow to ischemic viscera and limbs. Unless it is urgently required, surgical correction of stenosed arteries should be undertaken only when the vascular inflammatory process is well controlled with medical therapy. In individuals who are refractory to or unable to taper glucocorticoids, methotrexate in doses up to 25 mg per week has yielded encouraging results. Preliminary results with anti-TNF therapies and tocilizumab have been encouraging, but will require further study through randomized trials to determine efficacy.

Abatacept was recently examined in the first randomized trial to be conducted in Takayasu arteritis but did not demonstrate efficacy beyond glucocorticoids alone. There have been a number of published retrospective studies with the use of tocilizumab in Takayasu arteritis that have supported benefit but further investigation is necessary to fully determine efficacy.

IgA VASCULITIS (HENOCH-SCHÖNLEIN)

DEFINITION

IgA vasculitis (Henoch-Schönlein) is a small-vessel vasculitis characterized by palpable purpura (most commonly distributed over the buttocks and lower extremities), arthralgias, gastrointestinal signs and symptoms, and glomerulonephritis.

INCIDENCE AND PREVALENCE

IgA vasculitis (Henoch-Schönlein) is usually seen in children; most patients range in age from 4 to 7 years; however, the disease may also be seen in infants and adults. It is not a rare disease; in one series it accounted for between 5 and 24 admissions per year at a pediatric hospital. The male-to-female ratio is 1.5:1. A seasonal variation with a peak incidence in spring has been noted.

PATHOLOGY AND PATHOGENESIS

The presumptive pathogenic mechanism for IgA (Henoch-Schönlein) vasculitis is immune-complex deposition. A number of inciting antigens have been suggested including upper respiratory tract infections, various drugs, foods, insect bites, and immunizations. IgA is the antibody class most often seen in the immune complexes and has been demonstrated in the renal biopsies of these patients.

CLINICAL AND LABORATORY MANIFESTATIONS

In pediatric patients, palpable purpura is seen in virtually all patients; most patients develop polyarthralgias in the absence of frank arthritis. Gastrointestinal involvement, which is seen in almost 70% of pediatric patients, is characterized by colicky abdominal pain usually associated with nausea, vomiting, diarrhea, or constipation, and is frequently accompanied by the passage of blood and mucus per rectum; bowel intussusception may occur. Renal involvement occurs in 10–50% of patients and is usually characterized by mild glomerulonephritis leading to proteinuria and microscopic hematuria, with red blood cell casts in the

majority of patients; it usually resolves spontaneously without therapy. Rarely, a progressive glomerulonephritis will develop. In adults, presenting symptoms are most frequently related to the skin and joints, while initial complaints related to the gut are less common. Although certain studies have found that renal disease is more frequent and more severe in adults, this has not been a consistent finding. However, the course of renal disease in adults may be more insidious and thus requires close follow-up. Myocardial involvement can occur in adults but is rare in children.

Laboratory studies generally show a mild leukocytosis, a normal platelet count, and occasionally eosinophilia. Serum complement components are normal, and IgA levels are elevated in about one-half of patients.

DIAGNOSIS

The diagnosis of IgA vasculitis (Henoch-Schönlein) is based on clinical signs and symptoms. Skin biopsy specimen can be useful in confirming leukocytoclastic vasculitis with IgA and C3 deposition by immunofluorescence. Renal biopsy is rarely needed for diagnosis but may provide prognostic information in some patients.

TREATMENT

TREATMENT

IgA Vasculitis (Henoch-Schönlein)

The prognosis of IgA vasculitis (Henoch-Schönlein) is excellent. Mortality is exceedingly rare, and 1–5% of children progress to end-stage renal disease. Most patients recover completely, and some do not require therapy. Treatment is similar for adults and children. When glucocorticoid therapy is required, prednisone, in doses of 1 mg/kg/d and tapered according to clinical response, has been shown to be useful in decreasing tissue edema, arthralgias, and abdominal discomfort; however, it has not proved beneficial in the treatment of skin or renal disease and does not appear to shorten the duration of active disease or lessen the chance of recurrence. Patients with rapidly progressive glomerulonephritis have been anecdotally reported to benefit from intensive plasma exchange combined with cytotoxic drugs. Disease recurrences have been reported in 10–40% of patients.

CRYOGLOBULINEMIC VASCULITIS

DEFINITION

Cryoglobulins are cold-precipitable monoclonal or polyclonal immunoglobulins. Cryoglobulinemia may be associated with a systemic vasculitis characterized by palpable purpura, arthralgias, weakness, neuropathy, and glomerulonephritis. Although this can be observed in association with a variety of underlying disorders including multiple myeloma, lymphoproliferative disorders, connective tissue diseases, infection, and liver disease, in many instances it appears to be idiopathic. Because of the apparent absence of an underlying

disease and the presence of cryoprecipitate containing oligoclonal/polyclonal immunoglobulins, this entity was referred to as *essential mixed cryoglobulinemia*. Since the discovery of hepatitis C, it has been established that the vast majority of patients who were considered to have essential mixed cryoglobulinemia have cryoglobulinemic vasculitis related to hepatitis C infection.

INCIDENCE AND PREVALENCE

The incidence of cryoglobulinemic vasculitis has not been established. It has been estimated, however, that 5% of patients with chronic hepatitis C will develop cryoglobulinemic vasculitis.

PATHOLOGY AND PATHOGENESIS

Skin biopsies in cryoglobulinemic vasculitis reveal an inflammatory infiltrate surrounding and involving blood vessel walls, with fibrinoid necrosis, endothelial cell hyperplasia, and hemorrhage. Deposition of immunoglobulin and complement is common. Abnormalities of uninvolved skin including basement membrane alterations and deposits in vessel walls may be found. Membranoproliferative glomerulonephritis is responsible for 80% of all renal lesions in cryoglobulinemic vasculitis.

The association between hepatitis C and cryoglobulinemic vasculitis has been supported by the high frequency of documented hepatitis C infection, the presence of hepatitis C RNA and anti–hepatitis C antibodies in serum cryoprecipitates, evidence of hepatitis C antigens in vasculitic skin lesions, and the effectiveness of antiviral therapy (see below). Current evidence suggests that in the majority of cases, cryoglobulinemic vasculitis occurs when an aberrant immune response to hepatitis C infection leads to the formation of immune complexes consisting of hepatitis C antigens, polyclonal hepatitis C–specific IgG, and monoclonal IgM rheumatoid factor. The deposition of these immune complexes in blood vessel walls triggers an inflammatory cascade that results in cryoglobulinemic vasculitis.

CLINICAL AND LABORATORY MANIFESTATIONS

The most common clinical manifestations of cryoglobulinemic vasculitis are cutaneous vasculitis, arthritis, peripheral neuropathy, and glomerulonephritis. Renal disease develops in 10–30% of patients. Life-threatening rapidly progressive glomerulonephritis or vasculitis of the CNS, gastrointestinal tract, or heart occurs infrequently.

The presence of circulating cryoprecipitates is the fundamental finding in cryoglobulinemic vasculitis. Rheumatoid factor is almost always found and may be a useful clue to the disease when cryoglobulins are not detected. Hypocomplementemia occurs in 90% of patients. An elevated ESR and anemia occur frequently. Evidence for hepatitis C infection must be sought in all patients by testing for hepatitis C antibodies and hepatitis C RNA.

TREATMENT

TREATMENT

Cryoglobulinemic Vasculitis

Acute mortality directly from cryoglobulinemic vasculitis is uncommon, but the presence of glomerulonephritis is a poor prognostic sign for overall outcome. In such patients, 15% progress to end-stage renal disease, with 40% later experiencing fatal cardiovascular disease, infection, or liver failure. As indicated above, the majority of cases are associated with hepatitis C infection. In such patients, treatment with antiviral therapy (Chap. 332) is first-line therapy for hepatitis C–associated cryoglobulinemic vasculitis, particularly given the efficacy of current hepatitis C therapies. Clinical improvement with antiviral therapy is dependent on the virologic response. Patients who clear hepatitis C from the blood have objective improvement in their vasculitis along with significant reductions in levels of circulating cryoglobulins, IgM, and rheumatoid factor. While transient improvement can be observed with glucocorticoids, a complete response is seen in only 7% of patients. Plasmapheresis and cytotoxic agents have been used in anecdotal reports. These observations have not been confirmed, and such therapies carry significant risks. Randomized trials with rituximab (anti-CD20) in hepatitis C–associated cryoglobulinemic vasculitis have provided evidence of benefit such that this agent should be considered in patients with active vasculitis either in combination with antiviral therapy or alone in patients who have relapsed through, are intolerant to, or have contraindications to antiviral agents.

SINGLE-ORGAN VASCULITIS

The potential for vasculitis to affect single organs has become increasingly recognized. This has been defined as vasculitis in arteries or veins of any size in a single organ that has no features that indicate that it is a limited expression of a systemic vasculitis. Examples include isolated aortitis, testicular vasculitis, vasculitis of the breast, isolated cutaneous vasculitis, and primary CNS vasculitis. In some instances, this may be discovered at the time of surgery such as orchiectomy for a testicular mass where there is concern for neoplasm that is found instead to be vasculitis. Some patients originally diagnosed with single-organ vasculitis may later develop additional manifestations of a more systemic disease. In instances where there is no evidence of systemic vasculitis and the affected organ has been removed in its entirety, the patient may be followed closely without immunosuppressive therapy. In other instances, such as primary CNS vasculitis or some patients with isolated cutaneous vasculitis, medical intervention is warranted.

IDIOPATHIC CUTANEOUS VASCULITIS

DEFINITION

The term *cutaneous vasculitis* is defined broadly as inflammation of the blood vessels of the dermis. Because of its heterogeneity, cutaneous vasculitis has been described by a variety of terms including *hypersensitivity vasculitis* and *cutaneous leukocytoclastic angiitis*. However, cutaneous vasculitis is not one specific disease but a manifestation that can be seen in a variety of settings. In >70% of cases, cutaneous vasculitis occurs either as part of a primary systemic vasculitis or as a secondary vasculitis related to an inciting agent or an

underlying disease (see "Secondary Vasculitis," below). In the remaining 30% of cases, cutaneous vasculitis occurs idiopathically.

INCIDENCE AND PREVALENCE

Cutaneous vasculitis represents the most commonly encountered vasculitis in clinical practice. The exact incidence of idiopathic cutaneous vasculitis has not been determined due to the predilection for cutaneous vasculitis to be associated with an underlying process and the variability of its clinical course.

PATHOLOGY AND PATHOGENESIS

The typical histopathologic feature of cutaneous vasculitis is the presence of vasculitis of small vessels. Postcapillary venules are the most commonly involved vessels; capillaries and arterioles may be involved less frequently. This vasculitis is characterized by a *leukocytoclasis*, a term that refers to the nuclear debris remaining from the neutrophils that have infiltrated in and around the vessels during the acute stages. In the subacute or chronic stages, mononuclear cells predominate; in certain subgroups, eosinophilic infiltration is seen. Erythrocytes often extravasate from the involved vessels, leading to palpable purpura. *Cutaneous arteritis* can also occur, which involves slightly larger-sized vessels within the dermis.

CLINICAL AND LABORATORY MANIFESTATIONS

The hallmark of idiopathic cutaneous vasculitis is the predominance of skin involvement. Skin lesions may appear typically as palpable purpura; however, other cutaneous manifestations of the vasculitis may occur, including macules, papules, vesicles, bullae, subcutaneous nodules, ulcers, and recurrent or chronic urticaria. The skin lesions may be pruritic or even quite painful, with a burning or stinging sensation. Lesions most commonly occur in the lower extremities in ambulatory patients or in the sacral area in bedridden patients due to the effects of hydrostatic forces on the postcapillary venules. Edema may accompany certain lesions, and hyperpigmentation often occurs in areas of recurrent or chronic lesions.

There are no specific laboratory tests diagnostic of idiopathic cutaneous vasculitis. A mild leukocytosis with or without eosinophilia is characteristic, as is an elevated ESR. Laboratory studies should be aimed toward ruling out features to suggest an underlying disease or a systemic vasculitis.

DIAGNOSIS

The diagnosis of cutaneous vasculitis is made by the demonstration of vasculitis on biopsy. An important diagnostic principle in patients with cutaneous vasculitis is to search for an etiology of the vasculitis—be it an exogenous agent, such as a drug or an infection, or an endogenous condition, such as an underlying disease (Fig. 356-1). In addition, a careful physical and laboratory examination should be performed to rule out the possibility of systemic vasculitis. This should start with the least invasive diagnostic approach and proceed to the more invasive only if clinically indicated.

TREATMENT

TREATMENT

Idiopathic Cutaneous Vasculitis

When an antigenic stimulus is recognized as the precipitating factor in the cutaneous vasculitis, it should be removed; if this is a microbe, appropriate antimicrobial therapy should be instituted. If the vasculitis is associated with another underlying disease, treatment of the latter often results in resolution of the former. In situations where disease is apparently self-limited, no therapy, except possibly symptomatic therapy, is indicated. When cutaneous vasculitis persists and when there is no evidence of an inciting agent, an associated disease, or an underlying systemic vasculitis, the decision to treat should be based on weighing the balance between the degree of symptoms and the risk of treatment. Some cases of idiopathic cutaneous vasculitis resolve spontaneously, whereas others remit and relapse. In patients with persistent vasculitis, a variety of therapeutic regimens have been tried with variable results. In general, the treatment of idiopathic cutaneous vasculitis has not been satisfactory. Fortunately, since the disease is generally limited to the skin, this lack of consistent response to therapy usually does not lead to a life-threatening situation. Agents with which there have been anecdotal reports of success include dapsone, colchicine, hydroxychloroquine, and nonsteroidal anti-inflammatory agents. Glucocorticoids are often used in the treatment of idiopathic cutaneous vasculitis. Therapy is usually instituted as prednisone, 1 mg/kg/d, with rapid tapering where possible, either directly to discontinuation or by conversion to an alternate-day regimen followed by ultimate discontinuation. In cases that prove refractory to glucocorticoids, a trial of another immunosuppressive agent may be indicated. Patients with chronic vasculitis isolated to cutaneous venules rarely respond dramatically to any therapeutic regimen, and cytotoxic agents should be used only as a last resort in these patients. Methotrexate and azathioprine have been used in such situations in anecdotal reports. Although cyclophosphamide is the most effective therapy for the systemic vasculitides, it should almost never be used for idiopathic cutaneous vasculitis because of the potential toxicity.

PRIMARY CENTRAL NERVOUS SYSTEM VASCULITIS

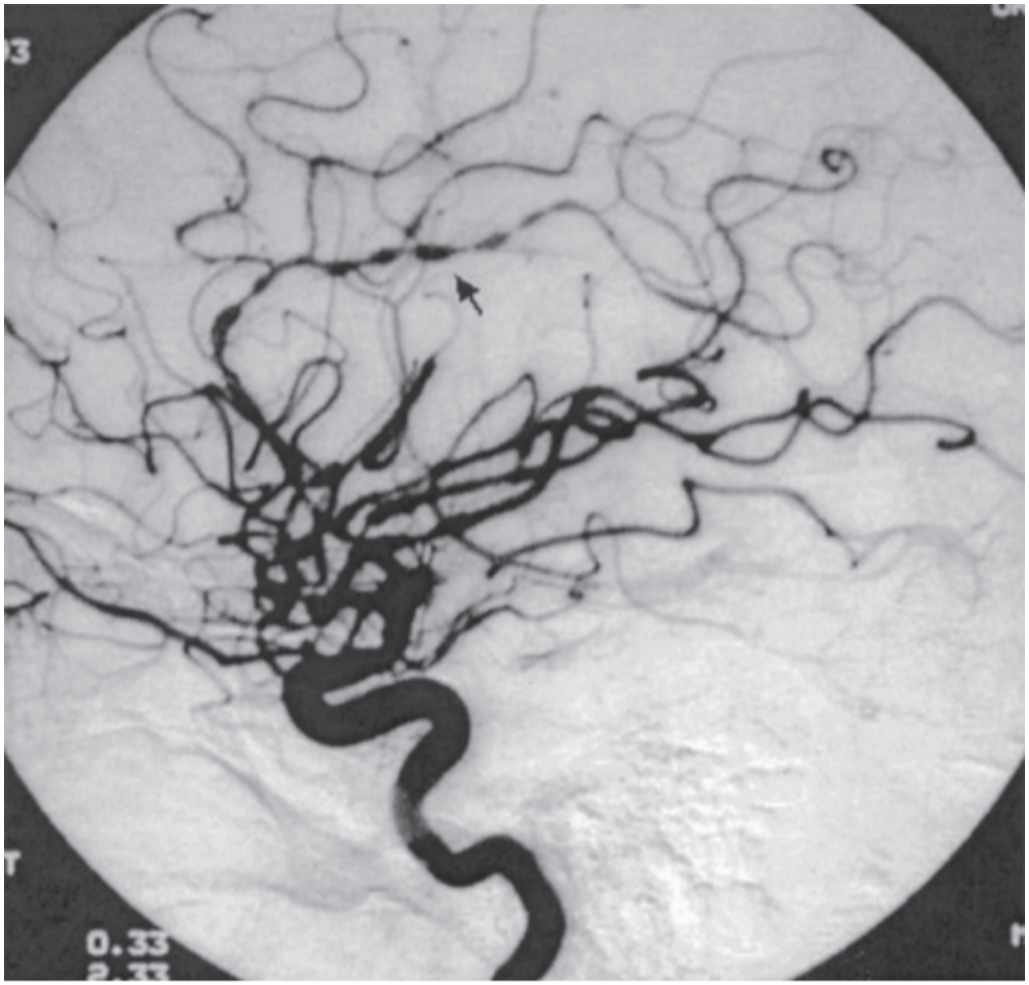
Primary central nervous system (CNS) vasculitis is an uncommon clinicopathologic entity characterized by vasculitis restricted to the vessels of the CNS without other apparent systemic vasculitis. The inflammatory process is usually composed of mononuclear cell infiltrates with or without granuloma formation.

Patients may present with headaches, altered mental function, and focal neurologic defects. Systemic symptoms are generally absent. Devastating neurologic abnormalities may occur depending on the extent of vessel involvement. The diagnosis can be suggested by abnormal magnetic resonance imaging of the brain, an abnormal lumbar puncture, and/or demonstration of characteristic vessel abnormalities on arteriography **(Fig. 356-4)**, but it is confirmed by biopsy of the brain parenchyma and leptomeninges. In the absence of a brain biopsy, care should be taken not to misinterpret as true primary vasculitis arteriographic abnormalities that might actually be related to another cause. An important entity in the differential diagnosis is reversible cerebral vasoconstrictive syndrome, which typically presents with “thunderclap” headache and is associated with arteriographic abnormalities that mimic primary CNS vasculitis that are reversible. Other diagnostic considerations include infection, atherosclerosis, emboli, connective tissue disease, sarcoidosis, malignancy,

and drug-associated causes. The prognosis of granulomatous primary CNS vasculitis is poor; however, some reports indicate that glucocorticoid therapy, alone or together with cyclophosphamide administered as described above, has induced clinical remissions.

FIGURE 356-4

Cerebral arteriogram from a 32-year-old man with primary central nervous system vasculitis. Dramatic beading (*arrow*) typical of vasculitis is seen.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

BEHÇET'S DISEASE

Behçet's disease is a clinicopathologic entity characterized by recurrent episodes of oral and genital ulcers, iritis, and cutaneous lesions. The underlying pathologic process is a leukocytoclastic venulitis, although vessels of any size and in any organ can be involved. **This disorder is described in detail in Chap. 357.**

COGAN'S SYNDROME

Cogan's syndrome is characterized by interstitial keratitis together with vestibuloauditory symptoms. It may be associated with a systemic vasculitis, particularly aortitis with involvement of the aortic valve.

Glucocorticoids are the mainstay of treatment. Initiation of treatment as early as possible after the onset of hearing loss improves the likelihood of a favorable outcome.

KAWASAKI'S DISEASE

Kawasaki's disease is an acute, febrile, multisystem disease of children. Some 80% of cases occur prior to the age of 5, with the peak incidence occurring at ≤2 years. It is characterized by nonsuppurative cervical adenitis and changes in the skin and mucous membranes such as edema; congested conjunctivae; erythema of the oral cavity, lips, and palms; and desquamation of the skin of the fingertips. Although the disease is generally benign and self-limited, it is associated with coronary artery aneurysms in ~25% of cases, with an overall case fatality rate of 0.5–2.8%. These complications usually occur between the third and fourth weeks of illness during the convalescent stage. Vasculitis of the coronary arteries is seen in almost all the fatal cases that have been autopsied. There is typical intimal proliferation and infiltration of the vessel wall with mononuclear cells. Beadlike aneurysms and thromboses may be seen along the artery. Other manifestations include pericarditis, myocarditis, myocardial ischemia and infarction, and cardiomegaly.

Apart from the up to 2.8% of patients who develop fatal complications, the prognosis of this disease for uneventful recovery is excellent. High-dose IV γ-globulin (2 g/kg as a single infusion over 10 h) together with aspirin (100 mg/kg/d for 14 days followed by 3–5 mg/kg/d for several weeks) have been shown to be effective in reducing the prevalence of coronary artery abnormalities when administered early in the course of the disease. Surgery may be necessary for Kawasaki disease patients who have giant coronary artery aneurysms or other coronary complications. Surgical treatment most commonly includes thromboendarterectomy, thrombus clearing, aneurysmal reconstruction, and coronary artery bypass grafting.

POLYANGIITIS OVERLAP SYNDROMES

Some patients with systemic vasculitis manifest clinicopathologic characteristics that do not fit precisely into any specific disease but have overlapping features of different vasculitides. Active systemic vasculitis in such settings has the same potential for causing irreversible organ system damage as when it occurs in one of the defined syndromes listed in Table 356-1. The diagnostic and therapeutic considerations as well as the prognosis for these patients depend on the sites and severity of active vasculitis. Patients with vasculitis that could potentially cause irreversible damage to a major organ system should be treated as described under "Granulomatosis with Polyangiitis (Wegener's)."

SECONDARY VASCULITIS

DRUG-INDUCED VASCULITIS

Vasculitis associated with drug reactions usually presents as palpable purpura that may be generalized or limited to the lower extremities or other dependent areas; however, urticarial lesions, ulcers, and hemorrhagic blisters may also occur **(Chap. 56)**. Signs and symptoms may be limited to the skin, although

systemic manifestations such as fever, malaise, and polyarthralgias may occur. Although the skin is the predominant organ involved, systemic vasculitis may result from drug reactions. Drugs that have been implicated in vasculitis include allopurinol, thiazides, gold, sulfonamides, phenytoin, and penicillin (Chap. 56).

An increasing number of drugs have been reported to cause vasculitis associated with antimyeloperoxidase ANCA. Of these, the best evidence of causality exists for hydralazine and propylthiouracil. The clinical manifestations in ANCA-positive drug-induced vasculitis can range from cutaneous lesions to glomerulonephritis and pulmonary hemorrhage. Outside of drug discontinuation, treatment should be based on the severity of the vasculitis. Patients with immediately life-threatening small-vessel vasculitis should initially be treated with glucocorticoids and cyclophosphamide as described for granulomatosis with polyangiitis (Wegener's). Following clinical improvement, consideration may be given for tapering such agents along a more rapid schedule.

SERUM SICKNESS AND SERUM SICKNESS–LIKE REACTIONS

These reactions are characterized by the occurrence of fever, urticaria, polyarthralgias, and lymphadenopathy 7–10 days after primary exposure and 2–4 days after secondary exposure to a heterologous protein (classic serum sickness) or a nonprotein drug such as penicillin or sulfa (serum sickness–like reaction). Most of the manifestations are not due to a vasculitis; however, occasional patients will have typical cutaneous venulitis that may progress rarely to a systemic vasculitis.

VASCULITIS ASSOCIATED WITH OTHER UNDERLYING DISEASES

Certain *infections* may directly trigger an inflammatory vasculitic process. For example, rickettsias can invade and proliferate in the endothelial cells of small blood vessels causing a vasculitis (Chap. 182). In addition, the inflammatory response around blood vessels associated with certain systemic fungal diseases such as histoplasmosis (Chap. 207) may mimic a primary vasculitic process. A leukocytoclastic vasculitis predominantly involving the skin with occasional involvement of other organ systems may be a minor component of many other infections. These include *subacute bacterial endocarditis*, *Epstein-Barr virus infection*, *HIV infection*, and a number of other infections.

Vasculitis can be associated with certain *malignancies*, particularly lymphoid or reticuloendothelial neoplasms. Leukocytoclastic venulitis confined to the skin is the most common finding; however, widespread systemic vasculitis may occur. Of particular note is the association of *hairy cell leukemia* (Chap. 106) with polyarteritis nodosa.

A number of *connective tissue diseases* have vasculitis as a secondary manifestation of the underlying primary process. Foremost among these are *systemic lupus erythematosus* (Chap. 349), *rheumatoid arthritis* (Chap. 351), *inflammatory myositis* (Chap. 358), *relapsing polychondritis* (Chap. 359), and *Sjögren's syndrome* (Chap. 354). The most common form of vasculitis in these conditions is the small-vessel venulitis isolated to the skin. However, certain patients may develop a fulminant systemic necrotizing vasculitis.

Secondary vasculitis has also been observed in association with *ulcerative colitis*, *congenital deficiencies of various complement components*, *sarcoidosis*, *primary biliary cirrhosis*, *α_1-antitrypsin deficiency*, and *intestinal bypass surgery*.

FURTHER READING

Buttgereit F et al: Polymyalgia rheumatica and giant cell arteritis: A systematic review. JAMA 315:2442, 2016. [PubMed: 27299619]

Fauci AS et al: Wegener's granulomatosis: Prospective clinical and therapeutic experience with 85 patients for 21 years. Ann Intern Med 98:76, 1983. [PubMed: 6336643]

Finkielman JD et al: Antiproteinase 3 antineutrophil cytoplasmic antibodies and disease activity in Wegener granulomatosis. Ann Intern Med 147:611, 2007. [PubMed: 17975183]

Guillevin L et al: Churg-Strauss syndrome. Clinical study and long-term follow-up of 96 patients. Medicine (Baltimore) 78:26, 1999. [PubMed: 9990352]

Hoffman GS et al: Wegener granulomatosis: An analysis of 158 patients. Ann Intern Med 16:488, 1992.

Jayne D et al: A randomized trial of maintenance therapy for vasculitis associated with antineutrophil cytoplasmic autoantibodies. N Engl J Med 349:36, 2003. [PubMed: 12840090]

Kerr GS et al: Takayasu arteritis. Ann Intern Med 120:919, 1994. [PubMed: 7909656]

Langford CA et al: A randomized, double-blind trial of abatacept (CTLA-4Ig) for the treatment of giant cell arteritis. Arthritis Rheumatol 69:837, 2017. [PubMed: 28133925]

Langford CA et al: A randomized, double-blind trial of abatacept (CTLA-4Ig) for the treatment of Takayasu arteritis. Arthritis Rheumatol 69:846, 2017. [PubMed: 28133931]

Pagnoux C et al: Clinical features and outcomes in 348 patients with polyarteritis nodosa: A systematic retrospective study of patients diagnosed between 1963 and 2005 and entered into the French Vasculitis Study Group Database. Arthritis Rheum 62:616, 2010. [PubMed: 20112401]

Stone JH et al: Rituximab versus cyclophosphamide for ANCA-associated vasculitis. N Engl J Med 363:221, 2010. [PubMed: 20647199]

Stone JH et al: Trial of tocilizumab in giant-cell arteritis. N Engl J Med 377:317, 2017. [PubMed: 28745999]

Wechsler ME et al: Mepolizumab or placebo for eosinophilic granulomatosis with polyangiitis. N Engl J Med 376:1921, 2017. [PubMed: 28514601]

Weyand CM, Goronzy JJ: Clinical practice. Giant-cell arteritis and polymyalgia rheumatica. N Engl J Med 371:50, 2014. [PubMed: 24988557]

Chapter 357: Behçet's Syndrome

Haralampos M. Moutsopoulos

DEFINITION, INCIDENCE, AND PREVALENCE

Behçet's syndrome is a multisystem disorder presenting with recurrent oral and genital ulcerations as well as ocular involvement. The diagnosis is clinical and based on internationally agreed diagnostic criteria **(Table 357-1)**.

TABLE 357-1

Diagnostic Criteria of Behçet's Syndrome

Recurrent oral ulceration plus two of the following:
Recurrent genital ulceration
Ocular lesions
Skin lesions
Positive Pathergy test

The syndrome affects young males and females from the Mediterranean region, the Middle East, and the Far East. The frequency of Behçet's syndrome increases from north to south Europe. Males and females are affected equally. Males often have more severe disease. The syndrome is rare in Sub-Saharan Africa.

PATHOGENESIS

The etiology and pathogenesis of this syndrome remain obscure. The disease appears to be in the crossroads of autoinflammatory and autoimmune disorders. The main pathologic lesion is systemic perivasculitis with early neutrophil infiltration and endothelial swelling. In some patients, diffuse inflammatory disease, involving all layers of large vessels and resulting to formation of pseudoaneurysms suggests vasculitis of vasa vasorum. Apart from activated neutrophils, increased numbers of infiltrating T_H1, T_H17, cytotoxic CD8+, and γδ T cells are observed, supporting a link between innate and adaptive autoreactive immune response. Circulating autoantibodies against α-enolase of endothelial cells, selenium binding protein, and anti-*Saccharomyces cerevisiae* antibodies have been observed, but their pathogenic role remains unclear. A recent genome-wide association study confirmed the known association of Behçet's syndrome with HLA-

B*51 and identified a second, independent association within the major histocompatibility complex (MHC) class I region. In addition, an association with interleukin (IL) 10 and the IL-23R–IL-12RB2 locus was also observed. Interestingly, the disease-associated IL-10 variant was correlated with diminished mRNA expression and low protein production.

CLINICAL FEATURES

The recurrent oral aphthous ulcerations are a sine qua non for the diagnosis. The ulcers are usually painful, shallow or deep with a central yellowish necrotic base, appear singly or in crops, and can be located anywhere in the oral cavity. Small ulcers, <10 mm in diameter, are seen in 85% of patients, whereas large or herpetiform lesions are less frequent. The ulcers persist for 1–2 weeks and subside without leaving scars. The genital ulcers are less common but more specific, are painful, do not affect the glans penis or urethra, and produce scrotal and vulvar scars.

Skin involvement is observed in 80% of patients and includes folliculitis, erythema nodosum, acne-like rashes, and, infrequently, vasculitis, Sweet syndrome, and pyoderma gangrenosum. Nonspecific skin inflammatory reactivity to scratch or intradermal saline injection (pathergy test) is a specific manifestation.

Eye involvement with scarring and bilateral panuveitis is the most dreaded complication, since it occasionally progresses rapidly to blindness. Uveitis occurs in 10–15% of patients, primarily in males. It is usually present at the onset but may also develop within the first few years. In addition to iritis, posterior uveitis, retinal vessel occlusions, and optic neuritis can be rarely seen in some patients.

Non-deforming arthritis or arthralgias are seen in 50% of patients and affect mostly the knees and ankles. Enthesopathy, avascular necrosis, myalgia, and myositis can be also seen.

Superficial or deep peripheral vein thrombosis is seen in 30% of patients. Pulmonary emboli are a rare complication. The superior vena cava is obstructed occasionally. Arterial involvement occurs in <1–5% of patients and presents with aortitis or peripheral arterial aneurysm and arterial thrombosis. Pulmonary artery vasculitis presenting with dyspnea, cough, chest pain, hemoptysis, and infiltrates on chest roentgenograms has been reported in <1% of patients and should be differentiated from thromboembolic disease since it warrants immunosuppressive and not thrombolytic therapy.

Neurologic involvement (5–10%) appears mainly in the parenchymal form (80%); it is associated with brainstem involvement and has a grave prognosis (*central nervous system [CNS]-Behçet's syndrome*). Cerebral brain thrombosis is more common in female patients. IL-6 is persistently raised in cerebrospinal fluid of these patients. Cerebral venous thrombosis is most frequently observed in the superior sagittal and transverse sinuses and is associated with headache and increased intracranial pressure. Magnetic resonance imaging (MRI) and/or proton magnetic resonance spectroscopy (MRS) are very sensitive and should be employed if CNS-Behçet's syndrome is suspected.

Gastrointestinal involvement is seen more frequently in patients from Japan and consists of mucosal ulcerations of the gut, resembling Crohn's disease.

Epididymitis is seen in 5% of patients, whereas AA amyloidosis and glomerulonephritis are uncommon.

Laboratory findings are mainly nonspecific indices of inflammation, such as leukocytosis and elevated erythrocyte sedimentation rate, as well as C-reactive protein levels.

TREATMENT

TREATMENT

Behçet's Syndrome

The severity of the syndrome usually abates with time. Apart from the patients with CNS-Behçet's syndrome and major vessel disease, the life expectancy seems to be normal and the only serious complication is blindness.

Mucous membrane involvement may respond to topical glucocorticoids in the form of mouthwash or paste. In more serious cases, thalidomide (100 mg/d) is effective. Recently it was shown that Apremilast, an inhibitor of phosphodiesterase-4, was effective for oral ulcers; however, the efficacy did not persist after withdrawal. Colchicine can be beneficial for the mucocutaneous manifestations and arthritis. Uveitis and CNS-Behçet's syndrome require systemic glucocorticoid therapy (prednisone, 1 mg/kg per day), azathioprine (2–3 mg/kg per day) and cyclosporine (2–5 mg/kg). Anti–tumor necrosis factor therapy is recommended in panuveitis refractory to immunosuppressive agents. Thrombophlebitis is treated with immunosuppressives. Pulse doses of cyclophosphamide are useful early in the course of the disease for pulmonary or peripheral arterial aneurysms.

FURTHER READING

Hatemi G et al: Behçet's syndrome: A critical digest of the 2014–2015 literature. Clin Exp Rheumatol 33(6 Suppl 94):S3, 2015. [PubMed: 26487500]

Chapter 358: Inflammatory Myopathies

Steven A. Greenberg; Anthony A. Amato

INTRODUCTION

This chapter focuses on the major types of inflammatory myopathies (IM), including dermatomyositis (DM), polymyositis (PM), immune-mediated necrotizing myopathy (IMNM), antisynthetase syndrome (ASS), and inclusion body myositis (IBM) **(Table 358-1)**. Other IM include those caused by infection, eosinophilic myositis and granulomatous myositis. Of note, inflammatory cell infiltrates can also be occasionally seen in muscle biopsies in hereditary (e.g., muscular dystrophies, metabolic myopathies) and toxic myopathies.

TABLE 358-1

Inflammatory Myopathies: Clinical and Laboratory Features

Disorder	Sex	Age of Onset	Rash	Pattern of Weakness	Laboratory Features	Muscle Biopsy	Cellular Infiltrate	Response to IS Therapy	Common Associated Conditions
DM	F > M	Childhood and Adult	Yes	Proximal > distal	Normal or increased CK (up to 50× normal or higher); Various MSAs (anti-MDA5, anti-TIF1, anti-Mi-2, anti-NXP2)	Perimysial and Perivascular Inflammation; IFN-1 regulated proteins (MHC-1, MxA), MAC deposition on capillaries	CD4+ Dendritic cells; B cells; macrophages	Yes	Myocarditis, ILD, Malignancy, Vasculitis, Other CTDs
PM	F > M	Adult	No	Proximal > distal	Increased CK (up to 50× normal or higher)	Endomysial and perivascular inflammation; ubiquitous expression of MHC-1	CD8+ T-cells; Macrophages; plasma cells	Yes	Myocarditis, ILD, Other CTDs
NM	M = F	Children and adults	No	Proximal > distal	Elevated CK (> 10× normal or higher); anti-HMGCR or anti-SRP antibodies	Necrotic muscle fibers; minimal inflammatory infiltrate	Macrophages in necrotic fibers undergoing phagocytosis	Yes	Malignancy, CTD, HMGCR antibody cases can be triggered by statin use
ASS	F > M	Children and adults	Sometimes	Proximal > distal	Elevated CK (>10× normal or higher); antisynthetase antibodies	Perimysial and Perivascular Inflammation; perimysial fragmentation with alkaline phosphatase staining; perimysial muscle damage with necrosis	CD4+ Dendritic cells; B cells; macrophages	Yes	Non-erosive arthritis, ILD, Raynaud phenomenon, mechanic hands, and fever

Disorder	Sex	Age of Onset	Rash	Pattern of Weakness	Laboratory Features	Muscle Biopsy	Cellular Infiltrate	Response to IS Therapy	Common Associated Conditions
IBM	M > F	Older adults (>50 yrs)	No	Proximal and distal; predilection for: finger/wrist flexors, knee extensors	Normal or mildly increased CK (usually <10× normal); anti-cN-1A antibodies; large granular lymphocytes on flow cytometry and reduced CD4/CD8 ratio with increased CD8 count	Endomysial and perivascular inflammation; ubiquitous expression of MHC-1; Rimmed Vacuoles; p62, LC3, TDP-43 aggregates; EM: 15–18 nm tubulofilaments; ragged red and COX negative fibers	CD8+ T-cells; Macrophages; plasma cells; myeloid dendritic cells; large granular lymphocytes	None or Minimal	Granular lymphocytic leukemia/lymphocytosis, sarcoidosis, SICCA or Sjogren syndrome

Abbreviations: CK, creatine kinase; cN-1A, cytosolic 5′-nucleotidase 1A; CTDs, connective tissue diseases; COX, cytochrome oxidase; DM, dermatomyositis; F, female; g, immunoglobulin; IBM, inclusion body myositis; IFN-1, type 1 interferon; ILD, interstitial lung disease; IS, immunosuppressive; M, male; MAC, membrane attack complex; MDA5, melanoma differentiation antigen; MHC-1, major histocompatibility antigen 1; NCP2, nuclear matrix protein 2 (NXP2); NM, necrotizing myopathy; PM, polymyositis; TIF1, transcriptional intermediary factor 1.

Source: Modified from AA Amato, JA Russell (eds): *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill Education; 2016, Table 33-1, p. 824, with permission.

Epidemiological studies suggest that the incidence of IM grouped together is greater than 4 cases per 100,000 with prevalence in the range of 14–32 per 100,000. Defining the actual incidence and prevalence of the individual myositides is limited however by the different diagnostic criteria employed in various epidemiological studies, increasing recognition of ASS, and frequent misdiagnosis of IBM and IMNM. Idiopathic PM without signs of an overlap syndrome is quite rare, while DM, IBM, and IMNM occur in roughly similar frequencies. DM can occur in children (juvenile DM), while IBM always occurs in adults and is the most common cause of myopathy in those aged >50. DM, PM, and ASS are more common in women, while IBM is more common in men.

DIAGNOSTIC APPROACH AND DIFFERENTIAL DIAGNOSIS

The approach to patients with suspected myopathy is detailed in **Chap. 441**. In any patient presenting with weakness, the first step is to localize the site of the lesion by history and clinical findings **(Chap. 21)**. Weakness could be caused by a process in the cerebral hemispheres, spinal cord **(Chap. 434)**, anterior horn cell **(Chap. 429)**, peripheral nerve **(Chaps. 438–439)**, neuromuscular junction **(Chap. 440)**, or muscle **(Chap. 441)**. Past medical history, medication use, and family history, combined with a detailed clinical examination and an appreciation for the pattern of muscle involvement (e.g., what muscles are weak and atrophic or hypertrophic as well as the presence of scapular winging, early contractures, sensory abnormalities, fasciculations, or rash) help differentiate myopathies from other neuromuscular disorders and the different types of myopathies from each other **(see Chap. 441)**. For example, atrophy with fasciculations suggest a neurogenic process such as amyotrophic lateral sclerosis, fatigable weakness on examination points to a neuromuscular junction defect such as myasthenia gravis, and concomitant sensory symptoms suggest a central process such as a spinal cord disorder or a polyneuropathy. Scapular winging, calf hypertrophy or atrophy, and early contractures before significant weakness develops would strongly suggest a muscular dystrophy, particularly if there is a positive family history. A heliotrope rash combined with Gottron papules and dilated nailfold capillaries is diagnostic for DM. The presence of atrophy and weakness of the flexor forearm muscles and quadriceps in a person aged >50 years is most likely IBM.

When the site of the lesion cannot be localized based on history and clinical examination alone, laboratory testing is required. Serum creatine kinase (CK) is the most sensitive laboratory marker of muscle destruction. Not all myopathies are associated with elevated CKs, but a markedly elevated CK (e.g., >2000 U/L) is almost always due to a myopathy. A slightly elevated CK can also be seen in neurogenic disorders, however. Myositis-associated and myositis-specific antibodies (MSA) help to distinguish subtypes of IM, as discussed below. Electromyography and nerve conductions studies (EMG/NCS) are useful in localizing the site of the lesion, but are less specific in helping to determine the actual cause of a myopathy. EMG can be useful

at times in guiding what muscle to biopsy, especially if muscles typically biopsied are normal on clinical examination. Imaging skeletal muscle can be helpful in assessing muscle involvement, revealing fatty replacement, atrophy, or edema within muscle or surround fascia.

A muscle biopsy is usually required to definitively distinguish one myopathy from another. The different forms of IM can have distinctive histopathological abnormalities as discussed below. In a patient with a classic DM rash, a muscle or skin biopsy can be performed, but an argument can also be made that biopsy is unnecessary—particularly if the patient also has an MSA specific for DM. However, a muscle biopsy should be performed in every case of suspected PM to exclude IBM (if not clinically apparent) and other causes of myopathy. Diagnosis of IMNM is by definition based upon histological findings. It is important to biopsy a muscle that is clinically affected but not too weak (e.g., Medical Research Council grade 4 out of 5 in strength), otherwise one may just see end-stage muscle. A biopsy should always be coordinated with an experienced muscle histopathology laboratory.

Patients with severe muscle pain, subjective weakness, and fatigue with normal strength and function on examination are not likely to have an IM. Polymyalgia rheumatica should be considered in older individuals with an elevated ESR or CRP, but normal CK and EMG. Fibromyalgia is likely in patients with a normal laboratory workup. In general, a muscle biopsy is not indicated unless there is objective weakness, an abnormal EMG, or elevated CK.

SPECIFIC DISORDERS

DERMATOMYOSITIS

Clinical Features

DM manifests with symmetric, proximal greater than distal weakness along with a characteristic rash that includes the heliotrope rash (erythematous discoloration of eyelids with periorbital edema), Gottron sign (erythematous rash over the extensor surfaces of joints such as the knuckles, elbows, knees, and ankles), Gottron papules (raised erythematous rash over knuckles) **(Fig. 358-1)**, V-sign (rash on the sun-exposed anterior neck and chest), shawl sign over the back of the neck and shoulders, nail bed telengiectasiae, and subcutaneous calcium deposits. The weakness and rash usually accompany one another but can be separated by several months. Furthermore, there is a spectrum of involvement such that some patients continue to manifest only with a rash (amyopathic DM), while others may present mainly with weakness and little or no visible skin changes. Patients can also complain of myalgias, arthralgias, dysphagia, and dysarthria. Cutaneous disease activity is highly relevant in DM; in comparison to other debilitating skin diseases including cutaneous lupus erythematosus, psoriasis, and atopic dermatitis, skin symptoms in DM patients are associated with an overall reduction in life quality. Pruritus can be especially debilitating. Dyspnea can occur from ventilatory muscle weakness or intrinsic pulmonary problems including interstitial lung disease (ILD), bronchopneumonia and alveolitis. Pulmonary manifestations are often associated with antisynthetase antibodies; myositis associated with the ASS can be considered a distinct disorder (discussed below). DM can present in children (juvenile DM) or in adults. There is a higher risk for malignancy in adult onset cases, ~15% within the first 2–3 years.

FIGURE 358-1

Cutaneous manifestations of dermatomyositis. ***A.*** Macular erythema plaques (Gottron sign) and erythematous papules (Gottron papules) on extensor surface of fingers and ***B.*** elbow. ***C.*** Macular erythema plaques over anterior neck and chest (V-sign) and ***D.*** the posterior neck, shoulder and upper back (Shawl sign). ***E.*** Nail bed changes with dilated capillaries.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Laboratory Features

Serum CK levels are elevated in 70–80% of patients; in 10% of those with normal CK, serum aldolase may be increased. Antinuclear antibodies can be positive but are a non-specific finding. DM is associated with several MSA targeting melanoma differentiation antigen 5 (MDA5), transcriptional intermediary factor 1 (TIF1), Mi-2 and nuclear matrix protein 2 (NXP2). These antibodies are usually associated with characteristic clinical features. For example, anti-MDA5 antibodies are associated with amyopathic DM with severe palmar rash, digital ulcers, and rapidly progressive ILD. Anti-TIF1 (or p155) antibodies and anti-NXP2 antibodies are associated with an increased risk of cancer, while anti-Mi-2 antibodies are often associated with more benign DM and a favorable response to treatment.

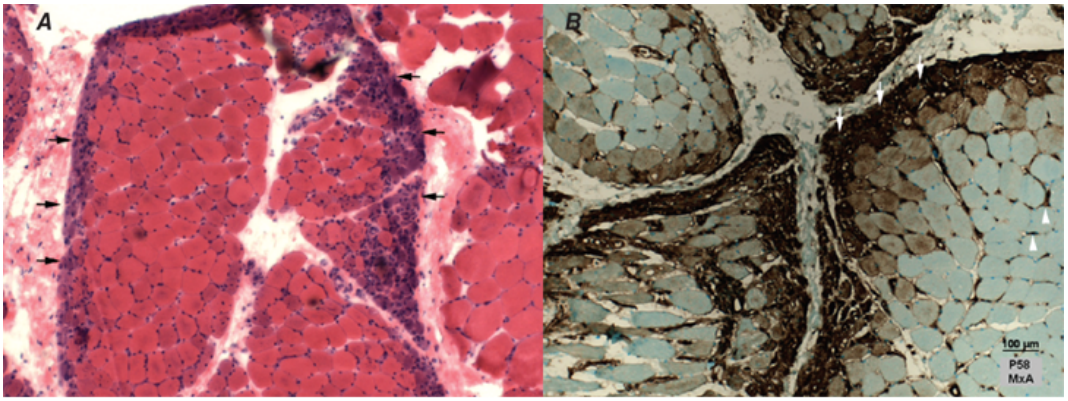
EMG of weak muscles shows increased insertional and spontaneous activity in the form of positive sharp waves and fibrillation potentials, or complex repetitive discharges along with early recruitment of small amplitude, short duration, polyphasic motor units. These findings are non-specific and can be seen in other myopathies. Skeletal muscle magnetic resonance imaging (MRI muscle) reveals edema in affected muscles, and sometimes more specific findings of abnormalities of fascia suggesting fasciitis.

Histopathology and Pathogenesis

The characteristic histopathological abnormality on muscle biopsy is perifasicular atrophy **(Fig. 358-2A)**; however, this is present in perhaps only 50% of patients. Immunohistochemical staining for myxovirus resistance protein A (MxA) is diagnostically more sensitive and highly specific **(Fig. 358-2B)**. The inflammatory cell infiltrate is predominantly perivascular and in the perimysium and is composed primarily of macrophages, B cells, and plasmacytoid dendritic cells (PDCs). Skin biopsies reveal cell-poor interface dermatitis, which is analogous to the perifasicular atrophy in that the basal layer of kerotinocytes are most damaged; the inflammatory infiltrate is typically absent or minimal, and when present is located mainly at the border zone of the dermis and epidermis.

FIGURE 358-2
Perifascicular atrophy and myxovirus resistance protein A (MxA) expression in dermatomyositis. ***A.*** Perifascicular myofibers (*black arrows*) bordering on disrupted perimysial connective tissue are atrophic and basophilic on hematoxylin and eosin (H&E) stains. ***B.*** Perifascicular myofibers (*white*

arrows) show intense staining for MxA protein along a gradient from superficial to deep; all capillaries show intense MxA expression (*white arrowheads*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The pathogenesis of DM was traditionally attributed to an antibody-mediated attack on endothelial cells, followed by complement-mediated destruction of capillaries and watershed ischemia of muscle fibers. However, recent studies suggest that this is not likely the case. Immunoglobulin deposition is largely absent on endothelial cells, and complement deposition may be a secondary phenomenon. There is increasing evidence that the microvasculopathy, skin, and muscle damage associated with DM is primarily due to toxicity from type I interferon-mediated pathways, most likely IFN-β.

Prognosis

In the absence of malignancy, prognosis is generally favorable in patients with DM, with 5-year survival rates ranging from 70 to 93%. Poor prognostic features are increased age, associated ILD, cardiac disease, and late or previous inadequate treatment.

POLYMYOSITIS

Clinical Features

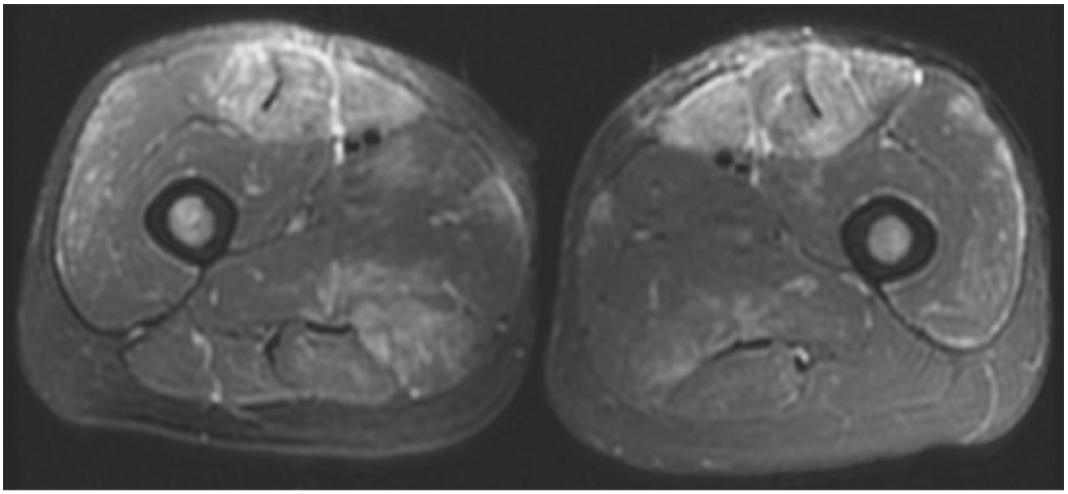
PM is a heterogenous group of disorders that usually presents with symmetric and proximal weakness that worsens over several weeks to months. As with DM, there can be associated heart, lung, and joint involvement as well as an increased risk of cancer. Some epidemiological studies suggest that the risk of cancer in PM is less than that in DM, but these older series likely included patients with IBM and dystrophies with inflammation who were misdiagnosed as having PM.

Laboratory Features

CK levels are always elevated in uncontrolled PM. A normal CK should alert clinicians to the possibility of IBM. As in DM, EMG and skeletal muscle imaging can be abnormal, but the findings are not specific **(Fig. 358-3)**.

FIGURE 358-3

Skeletal muscle MRI with short T1 inversion recovery (STIR) imaging in polymyositis. MRI of the thigh demonstrates bright signal indicative of edema/inflammation, particularly in the rectus femoris muscle. This contrasts with MRI in IBM in which there is more selective involvement of the vastus lateralis and medialis with relative sparing of the rectus femoris (see Figure 358-7 ***F*** and ***G***).

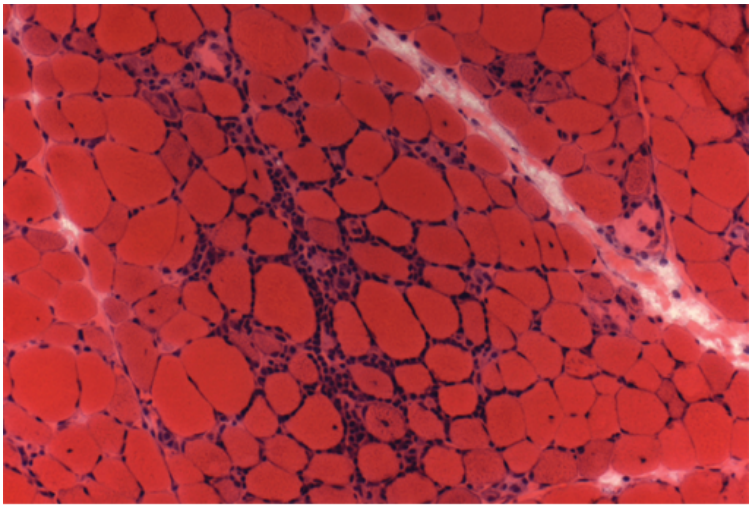


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Histopathology and Pathogenesis

Because PM is a heterogeneous category, muscle pathology varies substantially. Most often, patients with non-specific inflammatory cells present in perimysial more often than endomysial locations have been categorized as PM. A small minority of patients have mononuclear inflammatory infiltrate that surround fibers with sarcolemmal major histocompatibility (MHC-I) expression **(Fig. 358-4)**. There is debate as to whether true invasion of myofibers occurs in PM, or rather always indicates IBM. The inflammatory infiltrate predominantly consists of CD8+ T cells and macrophages located in the endomysial, perimysial and perivascular regions. As PM is heterogeneous, its varied forms of pathogenesis are poorly understood.

FIGURE 358-4
Pathology of polymyositis. Muscle biopsy demonstrates endomysial infiltrates surrounding non-necrotic muscle fibers.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prognosis

Most patients with PM improve with immunotherapies, but usually require life-long treatment. Some retrospective studies suggest that PM does not respond as well as DM to these therapies. However, many of these older series of “PM” likely included patients who actually had IMNM, IBM, or other myopathies (including muscular dystrophies) that do not respond to immunotherapies. As in DM, poor prognostic features are cancer, increased age, lung or cardiac involvement, and late or previously inadequate treatment.

OVERLAP SYNDROMES

The term “overlap syndrome” is applied when DM or PM is associated with other well-defined connective tissue diseases (CTDs) such as scleroderma, mixed connective tissue disease (MCTD), Sjögren syndrome, systemic lupus erythematosus (SLE), or rheumatoid arthritis. As in DM and PM, the myositis associated with these overlap syndromes is usually responsive to immunotherapies.

IMMUNE-MEDIATED NECROTIZING MYOPATHY

Clinical Features

IMNM is characterized by the acute or insidious onset of symmetric, proximal more than distal weakness. Dysphagia, dysarthria, or myalgia may occur. Patients may have an underlying CTD (usually scleroderma or MCTD), cancer (paraneoplastic necrotizing myopathy), or it may be idiopathic. There are at least two distinct forms of IMNM associated with specific autoantibodies (anti-HMGCR and anti-signal recognition particle [SRP]). HMGCR myopathy can be seen in patients receiving statins, inhibitors of 3-hydroxy-3-methyl-glutaryl-coenzyme reductase (HMGCR), particularly in patients aged >50 years. However, HMGCR myopathy can develop in children and young adults without a history of statin use and can mimic a limb girdle muscular dystrophy. Unlike the more common “toxic” myopathy associated with statin use, HMGCR myopathy does not improve when statins are discontinued. SRP myopathies are notable for the presence of anti-SRP antibodies and a typically subacute, aggressive, and relatively refractory course.

Laboratory Features

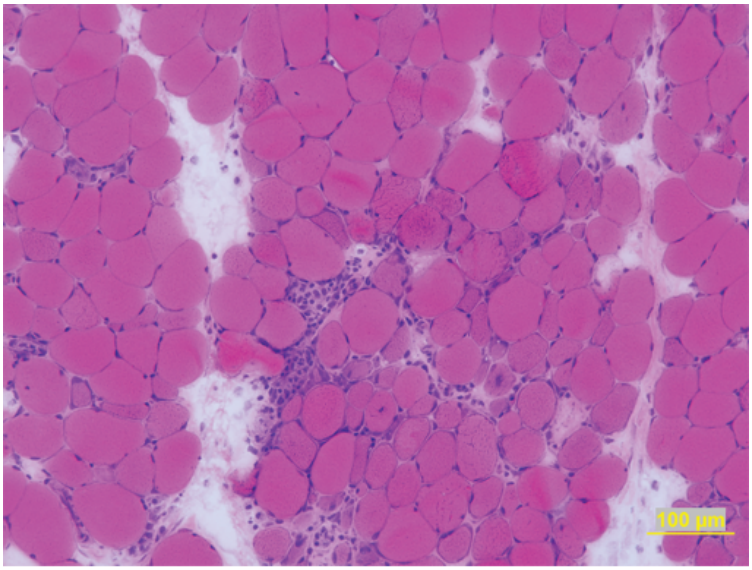
CK levels are markedly elevated (usually >10 × normal) in IMNM. As mentioned, IMNM can be associated with anti-HMGCR or anti-SRP antibodies. EMG and skeletal muscle imaging findings are non-specifically abnormal.

Histopathology and Pathogenesis

Muscle biopsies reveal multifocal necrotic and regenerating muscle fibers with a paucity of inflammatory cells **(Fig. 358-5)**. However, some patients with HMGCR myopathy have endomysial, macrophage-predominant infiltrates similar to what is seen in PM. Overexpression of MHC-I and membrane attack complex (MAC) may be evident on sarcolemma of non-necrotic fibers and MAC deposition on capillaries. The pathogenesis of IMNM is poorly understood.

FIGURE 358-5

Pathology of immune-mediated necrotizing myopathy. Muscle biopsy demonstrates scattered necrotic fibers with inflammatory infiltrate confined to those fibers undergoing myophagocytosis along with a few regenerating fibers.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prognosis

IMNM is generally much more difficult to treat than either DM or PM and aggressive immunotherapy is usually required. The progressive course despite immunotherapy and marked weakness with atrophy can lead to a misdiagnosis of a limb-girdle muscular dystrophy. There may be an increased incidence of cancer in patients with HMGCR myopathy, thus patients should undergo a malignancy workup.

ANTISYNTHETASE SYNDROME

Clinical Features

The presence of myositis, non-erosive arthritis, ILD, Raynaud phenomenon, mechanic hands, and fever associated with antibodies against aminoacyl-tRNA synthetase constitute the ASS. Some patients have an erythematous rash and muscle biopsies share histopathological features of DM, which

likely accounts for many of these patients being classified as having DM.

Laboratory Features

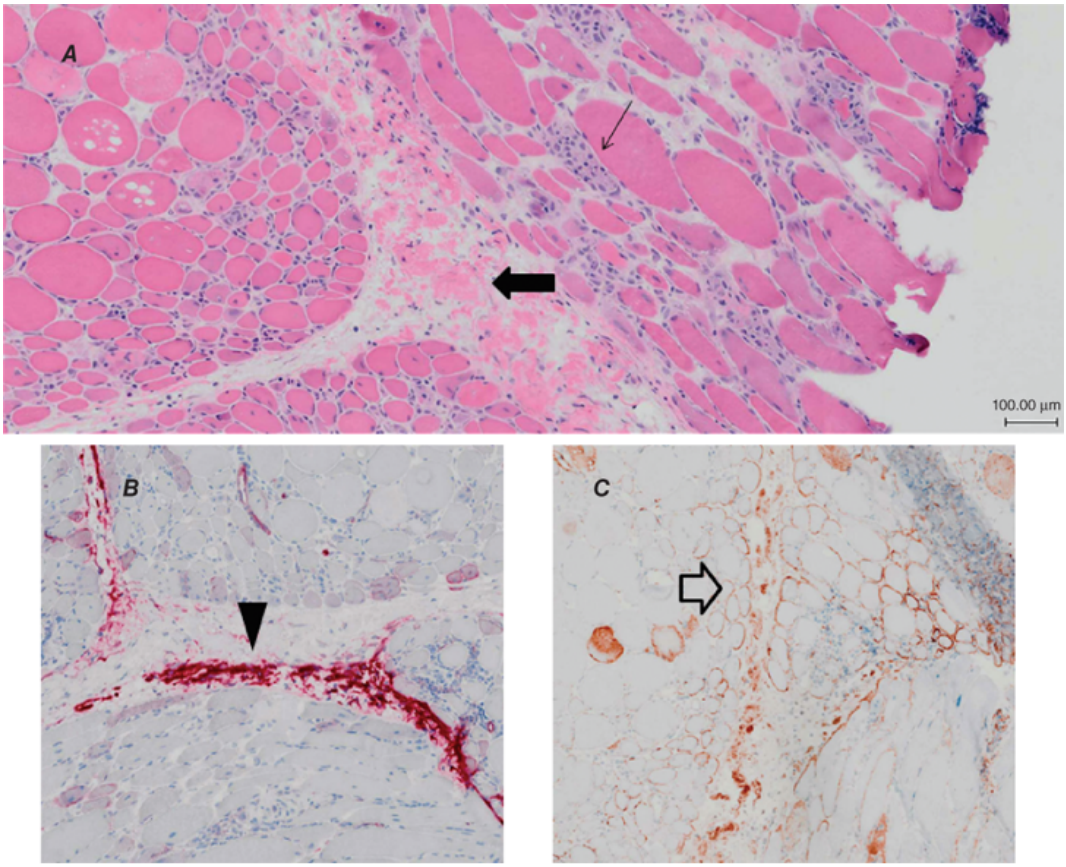
Antibodies against aminoacyl-tRNA synthetases are the most common MSA, present in 25–35% of patients with myositis. The most common aminoacyl-tRNA synthetase antibody is anti-Jo-1. CK is usually elevated in patients with ASS and myositis. Those with ILD demonstrate reduced forced vital capacity and diffusion capacity on pulmonary function tests. Spiral chest CT scans are best at demonstrating the honeycomb pattern of ILD. Skeletal muscle MRI and EMG show abnormalities similar to DM, PM, and IMNM.

Histopathology and Pathogenesis

Muscle biopsies demonstrate a predilection for perimysial damage including perimysial fragmentation and staining with alkaline phosphatase **(Fig. 358-6)**, PDCs and macrophages in the perimysium and around blood vessels, and MAC deposition on capillaries. Also similar to DM there is perifascicular muscle fiber damage, but with AAS there is more perifascicular muscle fiber necrosis compared to DM in which perifascicular atrophy is more prominent. MHC-1 and MAC deposits on muscle fibers may be seen on sarcolemma of perifascicular muscle fibers.

FIGURE 358-6

Pathology of myositis with anti-Jo-1 antibodies (antisynthetase syndrome). ***A.*** Perifascicular/perimysial muscle fiber atrophy and necrosis (*thin arrow*) associated with perimysial connective tissue is edematous and fragmented in appearance (*thick arrow*), H&E stain. ***B.*** The perimysial connective tissue intensely stains red with alkaline phosphatase stain (*arrowhead*). ***C.*** Immunostaining demonstrates deposition of membrane attack complex (MAC) deposits on the sarcolemma of non-necrotic perifascicular muscle fibers (*open arrow*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prognosis

Most patients respond to treatment although responses are less complete than for DM and PM; ILD can be particularly refractory to treatment. Unlike DM, PM, and IMNM, there does not appear to be an increased risk of malignancy.

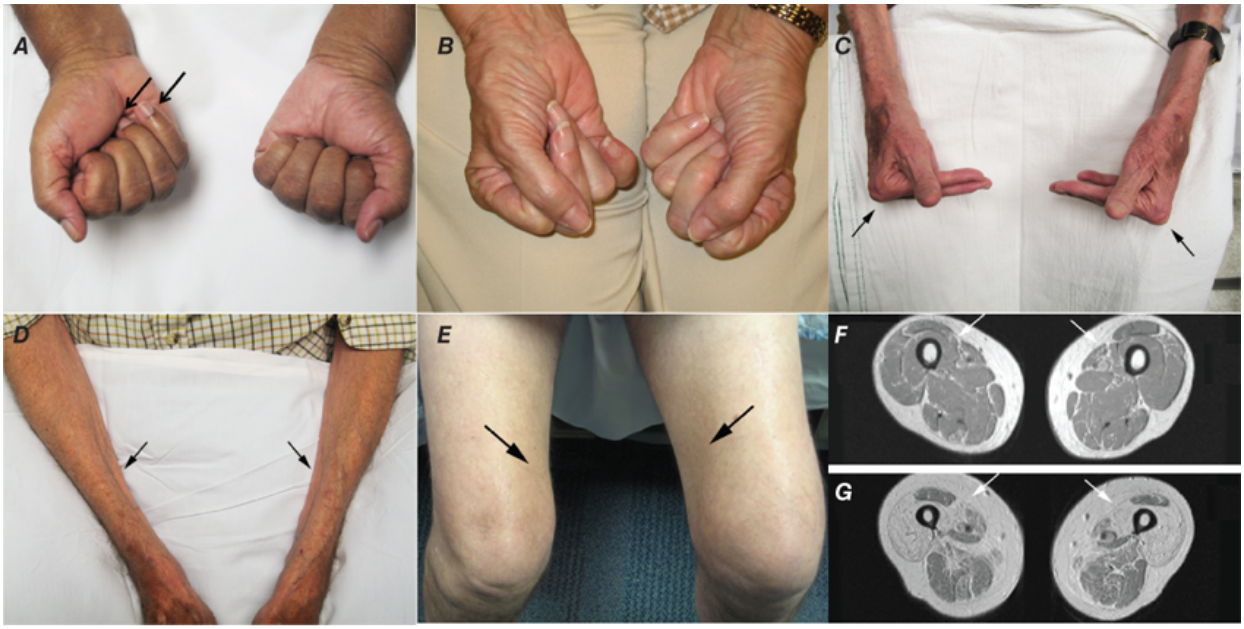
INCLUSION BODY MYOSITIS

Clinical Features

IBM usually manifests in patients over the age of 50 years and is slightly more common in men than women. It is associated with slowly progressive weakness and muscle atrophy that has a predilection for early involvement of the wrist and finger flexors in the arms and quadriceps in the legs (Fig. 358-7). Weakness is often asymmetric. Dysphagia is common and rarely can be the presenting feature. These clinical features can help distinguish IBM from PM and other forms of myopathy. The mean duration from onset of symptoms to use of wheelchair or scooter is ~15 years. There is no known increased risk of malignancy.

FIGURE 358-7

Muscle manifestations of inclusion body myositis (*A-C*). Finger flexor weakness can be (***A***) subtle and multifocal (*black arrows*), (***B***) moderate, or (***C***) severe. Note even with complete paralysis of deep and superficial finger flexors, metacarpophalangeal joint flexion (*arrows*) is often maintained due to preservation of lumbricals. ***D***. Ventral forearm atrophy (*arrows*). ***E.*** Atrophy of medial thighs due to loss of vastus medialis (*arrows*). ***F.*** Early IBM, with relatively preserved vastus medialis (*arrows*), in contrast to (***G***) advanced IBM with marked fibrous replacement of vastus medialis (*arrows*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Laboratory Features

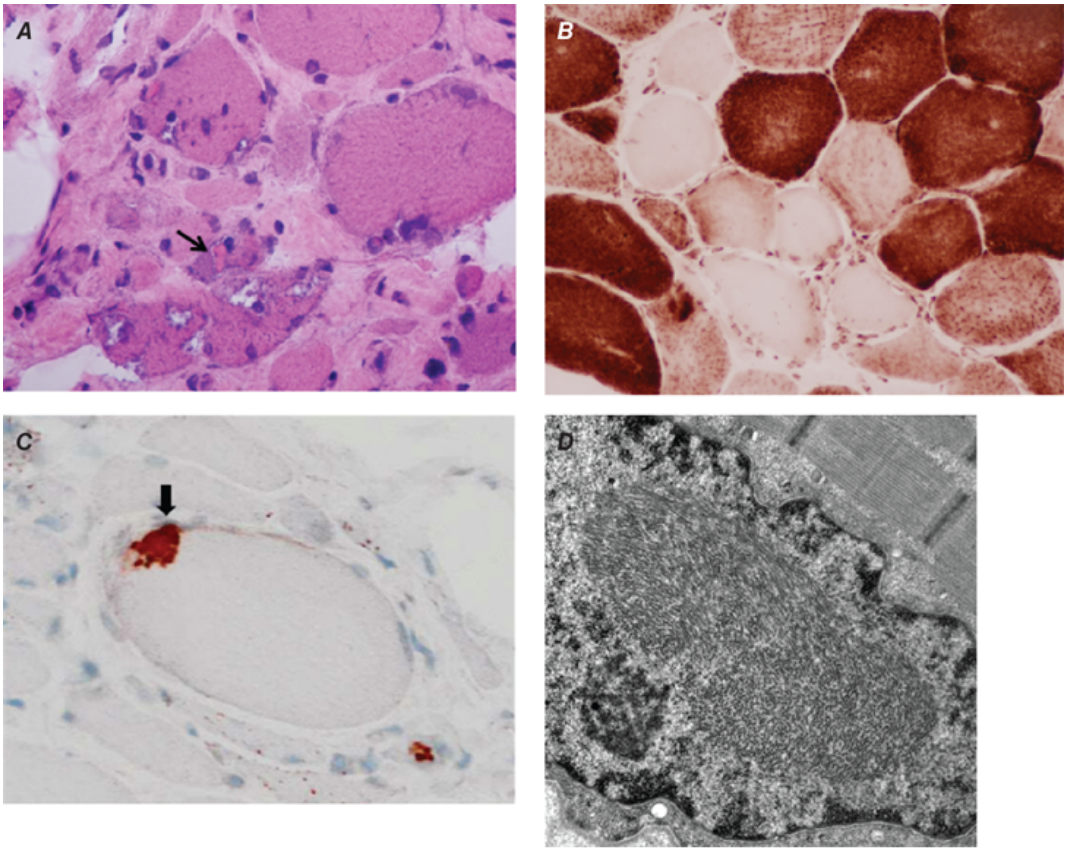
CK levels can be normal or only slightly elevated (usually <10 times normal). Antibodies targeting cytosolic 5′-nucleotidase 1A (cN-1A) are detected in the blood in a third to more than two-thirds of IBM patients, and is a highly specific diagnostic biomarker for IBM among patients with myopathy. Other blood biomarkers for IBM include the presence of an abnormal population of large granular lymphocytes on flow cytometry and a reduced CD4/CD8 ratio with an increased CD8 count. Needle EMG may demonstrate large amplitude, long duration motor unit potentials that can be misinterpreted as neurogenic but reflect the chronicity of the myopathy. Muscle MRI may show a predilection for involvement of the flexor digitorum profundus in the arms and the vastus medialis and lateralis muscles with sparing of the rectus femoris muscle.

Histopathology and Pathogenesis

Muscle biopsies demonstrate endomysial inflammatory infiltrates predominantly composed of CD8+ T cells and macrophages surrounding and invading non-necrotic muscle fibers, MHC-1 expression on the sarcolemma, fibers with rimmed vacuoles, cytochrome oxidase (COX) negative fibers, and inclusions on light or electron microscopy (Fig. 358-8). The inclusions contain beta-sheet misfolded proteins (amyloid), but are difficult to appreciate with routine Congo red stain (they are seen on frozen but not paraffin sections). Immunostaining for p62 appears to be the most sensitive stain for detection of these inclusions. Importantly, rimmed vacuoles may not be seen in as many as 20–30% of muscle biopsies. In such cases, the presence of mitochondrial abnormalities (ragged red and COX negative fibers) and immunostaining demonstrating p62 inclusions are helpful in distinguishing IBM from PM (aside from the clinical pattern of muscle weakness).

FIGURE 358-8

Pathology of inclusion body myositis. *A.* Scattered muscle fibers with rimmed vacuoles and rare fibers with eosinophilic inclusions (*arrow*), H&E stain. ***B.*** Cytochrome oxidase stain demonstrates an increased number of pale staining or COX negative muscle fibers. ***C.*** Cytoplasmic inclusions stain positive with p62 within a muscle fiber (*thick arrow*). ***D.*** Electromicroscopy reveals 15–21 nm tubulofilamentous inclusions within a myonucleus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The pathogenesis of IBM is poorly understood. The marked adaptive immune system abnormalities related to T cell inflammation and the presence of a relatively specific autoantibody against a muscle protein indicate an autoimmune attack on muscle. The chronic and highly inflammatory environment within muscles in IBM may alter protein synthesis and degradation pathways in part via aberrant immunoproteasome expression. Additional histologic features, typically referred to as “degenerative,” include aggregation of various proteins including markers of endoplasmic reticulum (ER) stress and autophagy (e.g., p62 and LC3). Involvement of ER stress and autophagy have also been observed in other autoimmune diseases, such as primary biliary cholangitis (PBC), inflammatory bowel disease, and ankylosing spondylitis, some of which can be highly refractory to immunotherapy.

Prognosis

The myopathy is slowly progressive, and is not typically responsive to immunotherapies. Most patients require a scooter or wheelchair within 10–15 years of onset of symptoms. Life expectancy is not significantly altered in IBM.

TREATMENT OF THE IM

DM, PM, ASS, and IMNM are typically responsive to immunotherapy (Table 358-2). High dose glucocorticoids (i.e., starting dose of prednisone 0.75–1.0 mg/kg per day) is considered the first-line treatment. There is uncertainty regarding when to start second-line agents (e.g., methotrexate, azathioprine, mycophenolate, immunoglobulin, or rituximab). The clinician must weigh with the patient the increased risks of immunosuppression versus possible benefits (e.g., faster improvement, steroid-sparing effect and/or avoidance of the morbidities associated with long-term glucocorticoid use). We usually start a second line agent (usually methotrexate) with glucocorticoids in patients with severe weakness or other organ system involvement (e.g., myocarditis, ILD), those with increased risk of steroid complications (e.g., diabetics, osteoporosis, or postmenopausal women), and patients with IMNM who are known to have difficult to treat myositis. In those in whom we initiate treatment with prednisone alone, a second-line agent is added in patients who fail to significantly improve after 2–4 months of treatment or in those who cannot be tapered to a low dose of prednisone.

TABLE 358-2

Immunotherapies for Inflammatory Myopathies

Therapy	Route	Dose	Side Effects	Monitor
Prednisone	Oral	0.75–1.5 mg kg/d to start	Hypertension, fluid and weight gain, hyperglycemia, hypokalemia, cataracts, gastric irritation, osteoporosis, infection, aseptic femoral necrosis	Weight, blood pressure, serum glucose/potassium, cataract formation
Methylprednisone	Intravenous	1 g in 100 mL/normal saline over 1–2 h, daily or every other day for 3–6 doses	Arrhythmia, flushing, dysgeusia, anxiety, insomnia, fluid and weight gain, hyperglycemia, hypokalemia, infection	Heart rate, blood pressure, serum glucose/potassium
Azathioprine	Oral	2–3 mg/kg per day; single a.m. dose	Flu-like illness, hepatotoxicity, pancreatitis, leukopenia, macrocytosis, neoplasia, infection, teratogenicity	Blood count, liver enzymes
Methotrexate	Oral	7.5–20 mg weekly, single or divided doses; one day a week dosing	Hepatotoxicity, pulmonary fibrosis, infection, neoplasia, infertility, leukopenia, alopecia, gastric irritation, stomatitis, teratogenicity	Liver enzymes, blood count
	Subcutaneously	20–50 mg weekly; one day a week dosing	Same as oral	Same as p.o.
Cyclophosphamide	Oral Intravenous	1.5–2 mg/kg per day; single a.m. dose 0.5 to 1.0 g/m^2 per month × 6–12 months	Bone marrow suppression, infertility, hemorrhagic cystitis, alopecia, infections, neoplasia, teratogenicity	Blood count, urinalysis
Cyclosporine	Oral	4–6 mg/kg per day, split into two daily doses	Nephrotoxicity, hypertension, infection, hepatotoxicity, hirsutism, tremor, gum hyperplasia, teratogenicity	Blood pressure, creatinine/BUN, liver enzymes, cyclosporine levels
Tacrolimus	Oral	0.1–0.2 mg/kg per d in two divided doses	Nephrotoxicity, hypertension, infection, hepatotoxicity, hirsutism, tremor, gum hyperplasia, teratogenicity	Blood pressure, creatinine/BUN, liver enzymes, tacrolimus levels
Mycophenolate mofetil	Oral	Adults (1 g BID to 1.5 g BID) Children (600 mg/m^2 per dose BID (no >1 g/d in patients with renal failure)	Bone marrow suppression, hypertension, tremor, diarrhea, nausea, vomiting, headache, sinusitis, confusion, amblyopia, cough, teratogenicity, infection, neoplasia	Blood count
Intravenous Immunoglobulin	Intravenous	2 g/kg over 2–5 days; then 1-g/kg every 4–8 weeks as needed	Hypotension, arrhythmia, diaphoresis, flushing, nephrotoxicity, headache, aseptic meningitis, anaphylaxis, stroke	Heart rate, blood pressure, creatinine/BUN

Therapy	Route	Dose	Side Effects	Monitor
Rituximab	Intravenous	A course is typically 750 mg-meter squared (up to 1 g) and repeated in 2 weeks Courses are then repeated usually every 6–18 months	Infusion reactions (as per IVIG), infection, progressive multifocal leukoencephalopathy	Some check B-cell count prior to subsequent courses (but this may not be warranted)

Abbreviations: BUN, blood urea nitrogen; CBC, complete blood count; Cr, creatinine; ECG, electrocardiogram; GI, gastrointestinal; IgA, immunoglobulin A; IVIG, intravenous immunoglobulin; LFT, liver function test; m^2, body surface area; mo: month(s); PFT, pulmonary function test; PRES, posterior reversible encephalopathy syndrome.

Source: Modified from AA Amato, JA Russell (eds): *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill Education; 2016, Table 33-8, p. 859, with permission.

Most patients with IMNM do not respond to prednisone alone or even prednisone plus a second-line agent in combination. Many require triple therapy with prednisone, methotrexate, and intravenous immunoglobulin (IVIG), and if this fails, rituximab. Recent reports suggest that anti-HMGCR myopathy may respond to monotherapy with IVIG, and a large multicenter clinical trial to test this approach is currently being organized.

Unfortunately, IBM does not typically respond to any known immunotherapies. The mainstay of treatment is physical and occupational therapy to improve function and swallowing therapy (and sometimes esophageal dilation or cricopharyngeal myotomy) in those with dysphagia.

GENERAL GUIDELINES FOR USE OF SPECIFIC IMMUMOTHERAPIES

Glucocorticoids

Treatment is initiated with prednisone (0.75–1.5 mg/kg up to 100 mg) administered as a daily morning single-dose (the most common dose used in adults is 60 mg daily). In patients with severe weakness or comorbidities (e.g., ILD, myocarditis), treatment with a short course of intravenous methylprednisolone (1 g daily for 3 days) is recommended prior to starting oral glucocorticoids. Patients are generally maintained on high-dose prednisone until strength normalizes or until improvement in strength has reached a plateau (usually 3–6 months). Subsequently, prednisone can be tapered by 5 mg every 2–4 weeks. Once the dose is reduced to 20 mg every day or every other day, the taper is slowed to 2.5 mg every 2–4 weeks. The goal is to taper prednisone to ≤10 mg daily. Although most patients improve, the response may not be complete and many will require at least a small dose of prednisone or a second-line agent to have a sustained remission. Serum CK levels are monitored; however, dose adjustments of prednisone and other immunotherapies are primarily based on the objective clinical examination and not the CK levels or the patients' subjective response. When no response is noted after an adequate trial of high-dose prednisone, alternative diagnoses (e.g., IBM or an inflammatory muscular dystrophy) and a repeat muscle biopsy should be considered.

Relapse of the myositis needs to be distinguished from steroid myopathy. Features suggesting a steroid myopathy include weakness developing while on high dosage, a normal serum CK, clinical features of steroid excess such as ecchymoses and "moon facies," and absence of muscle membrane irritability on EMG. By contrast, patients experiencing relapse of myositis may become weaker during the prednisone taper, have increasing serum CK levels, and display abnormal spontaneous activity on EMG.

SECOND-LINE THERAPIES

Methotrexate

Methotrexate is usually the second-line treatment of choice because most authorities believe it works faster than other agents. An oral dose of 5 or 7.5 mg/week is initiated, and then gradually increased as needed up to 25 mg/week. If there is no improvement after 1 month of 25 mg/week of oral methotrexate, a switch to weekly parenteral (usually subcutaneous) methotrexate is the next step, with dose escalation by 5 mg weekly; only rarely is a dose >35 mg/week used. The major side effects of methotrexate are alopecia, stomatitis, ILD, teratogenicity, oncogenicity, risk of infection, and pulmonary fibrosis, along with bone marrow, renal, and liver toxicity. Patients are concomitantly treated with folate or folinic acid.

Azathioprine

A recommended initial dose is 50 mg/day in adults, which can be increased by 50 mg every 2 weeks up to 2–3 mg/kg per d. Approximately 12% of patients develop a systemic reaction characterized by fever, abdominal pain, nausea, vomiting, and anorexia that requires discontinuation of the

drug. The major practical limitation of azathioprine is that 6–18 months of treatment is usually required before benefit can be seen. Patients can be prescreened for thiopurine methyltransferase (TPMT) deficiency that is associated with severe bone marrow toxicity from this drug.

Mycophenolate Mofetil

This drug inhibits the proliferation of T and B lymphocytes by blocking purine synthesis. It appears to be effective in different forms of myositis, and is the second-line treatment of choice for myositis patients with ILD. The starting dose is 1.0 g twice daily and can be increased to 3 g daily in divided doses, if necessary. Mycophenylate is excreted through the kidneys; therefore, the dose should be decreased (no >1 g/d total dose) in patients with renal insufficiency. An advantage of mycophenylate compared to other immunosuppressive agents is the lack of renal or hepatic toxicity.

Intravenous Immunoglobulin

IVIG is used in patients refractory to prednisone and at least one second-line immunosuppressive agent, although recent reports suggest that it may be the treatment of choice and effective as a monotherapy in anti-HMCCR myopathy. A dose of 2 g/kg is divided over 2–5 days, and repeat infusions given at monthly intervals for at least 3 months. Subsequently, intervals can be lengthened or dosage decreased: 2 g/kg every 2 months or 1 g/kg per month.

Rituximab

Rituximab is a monoclonal antibody directed against CD20+ B-cells. A large randomized controlled trial found no benefit, but there were flaws in the study design. Most authorities feel that rituximab can be beneficial in some patients who are refractory to prednisone and at least one of the other second-line agents. The typical dosage is 750 mg/m^2 (up to 1 g) IV with a second infusion 2 weeks later, and repeat courses (375 mg/m^2 as a single infusion or with a second infusion 2 weeks apart) every 6–18 months as needed.

GLOBAL ISSUES

 There is a lack of epidemiological data regard to the incidence and prevalence of various subtypes of IM throughout the world. Complicating the issue is disease awareness and the inability to obtain and process muscle biopsies and MSAs, particularly in less developed countries. Nevertheless each of these disorders occurs throughout the world. The specific environmental triggers and genetic risk factors are likely variable. Interestingly, a report out of Japan found that 28% of IBM patients had evidence of exposure to hepatitis C, which was much higher than seen in the western hemisphere and also more common than seen in PM and healthy population controls in Japan. HIV-associated PM and IBM are more commonly encountered in areas endemic for HIV and recent studies suggest most of these “PM” cases turn out to have IBM and can develop symptoms at an earlier age (e.g., in the 30s). Pyomyositis and parasitic myositis are clearly more common in the tropics. The prevalence of different types of cancers vary in different parts of the world, an important consideration with respect to paraneoplastic myositis seen in DM, PM, and IMNM. For example, nasopharyngeal cancer is particularly common in Asia, thus assessment for this type of cancer should be considered in the workup of patients from high-risk regions.

FURTHER READING

Amato AA, Russell JA (eds): *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill Education, 2016, pp. 827–871.

Greenberg SA et al: Association of inclusion body myositis with T cell large granular lymphocytic leukaemia. Brain 139:1348, 2016. [PubMed: 26920676]

Huard C et al: Correlation of cutaneous disease activity with type 1 interferon gene signature and interferon beta in dermatomyositis. Br J Dermatol 176:1224, 2017. [PubMed: 27564228]

Larman HB et al: Cytosolic 5’-nucleotidase 1A autoimmunity in sporadic inclusion body myositis. Ann Neurol 73:408, 2013. [PubMed: 23596012]

Lundberg IE et al: Diagnosis and classification of idiopathic inflammatory myopathies. J Intern Med 280:39, 2016. [PubMed: 27320359]

Mammen AL et al: Autoantibodies against 3-hydroxy-3-methylglutaryl-coenzyme A reductase in patients with statin-associated autoimmune myopathy. Arthritis Rheum 63:713, 2011. [PubMed: 21360500]

Pluk H et al: Autoantibodies to cytosolic 5’-nucleotidase 1A in inclusion body myositis. Ann Neurol 73:397, 2013. [PubMed: 23460448]

Rose MR, ENMC IBM Working Group: 188th ENMC International Workshop: Inclusion Body Myositis, 2–4 December 2011, Naarden, The Netherlands. Neuromuscul Disord 23:1044, 2013. [PubMed: 24268584]

Watanabe Y et al: Clinical features and prognosis in anti-SRP and anti-HMGCR necrotising myopathy. J Neurol Neurosurg Psychiatry 87:1038, 2016. [PubMed: 27147697]

Chapter 359: Relapsing Polychondritis

Carol A. Langford

INTRODUCTION

Relapsing polychondritis is an uncommon disorder of unknown cause characterized by inflammation of cartilage predominantly affecting the ears, nose, and laryngotracheobronchial tree. Other manifestations include scleritis, neurosensory hearing loss, polyarthritis, cardiac abnormalities, skin lesions, and glomerulonephritis. Relapsing polychondritis has been estimated to have an incidence of 3.5 per million population per year. The peak age of onset is between the ages of 40 and 50 years, but relapsing polychondritis may affect children and the elderly. It is found in all races, and both sexes are equally affected. No familial tendency is apparent. A significantly higher frequency of HLA-DR4 has been found in patients with relapsing polychondritis than in healthy individuals. A predominant subtype allele(s) of HLA-DR4 was not found. Approximately 30% of patients with relapsing polychondritis will have another rheumatologic disorder, the most frequent being systemic vasculitis, followed by rheumatoid arthritis, and systemic lupus erythematosus (SLE). Nonrheumatic disorders have also been associated with relapsing polychondritis (Table 359-1). In most cases, these disorders antedate the appearance of relapsing polychondritis, usually by months or years; however, in other instances, the onset of relapsing polychondritis can accompany disease presentation.

TABLE 359-1

Disorders Associated with Relapsing Polychondritis[a]

Systemic vasculitis
Rheumatoid arthritis
Systemic lupus erythematosus
Overlapping connective tissue disease
Spondyloarthritides
Behçet's disease
Polymyalgia rheumatica
Primary biliary cirrhosis
Pulmonary fibrosis
Hashimoto's thyroiditis
Grave's disease
Crohn's disease
Ulcerative colitis
Myelodysplastic syndrome

[a]Systemic vasculitis is the most common association, followed by rheumatoid arthritis and systemic lupus erythematosus.

Source: Modified from CJ Michet et al: Ann Intern Med 104:74, 1986.

PATHOLOGY AND PATHOPHYSIOLOGY

The earliest abnormality of hyaline and elastic cartilage noted histologically is a focal or diffuse loss of basophilic staining indicating depletion of proteoglycan from the cartilage matrix. Inflammatory infiltrates

are found adjacent to involved cartilage and consist predominantly of mononuclear cells and occasional plasma cells. In acute disease, polymorphonuclear white cells may also be present. Destruction of cartilage begins at the outer edges and advances centrally. There is lacunar breakdown and loss of chondrocytes. Degenerating cartilage is replaced by granulation tissue and later by fibrosis and focal areas of calcification. Small loci of cartilage regeneration may be present. Immunofluorescence studies have shown immunoglobulins and complement at sites of involvement. Extracellular granular material observed in the degenerating cartilage matrix by electron microscopy has been interpreted to be enzymes, immunoglobulins, or proteoglycans.

The accumulating data strongly suggest that both humoral and cell-mediated immunity play an important role in the pathogenesis of relapsing polychondritis. Immunoglobulin and complement deposits are found at sites of inflammation. In addition, antibodies to type II collagen and to matrilin-1 and immune complexes are detected in the sera of some patients. The possibility that an immune response to type II collagen may be important in the pathogenesis is supported experimentally by the occurrence of auricular chondritis in rats immunized with type II collagen. Antibodies to type II collagen are found in the sera of these animals, and immune deposits are detected at sites of ear inflammation. Humoral immune responses to type IX and type XI collagen, matrilin-1, and cartilage oligomeric matrix protein have been demonstrated in some patients. In a study, rats immunized with matrilin-1 were found to develop severe inspiratory stridor and swelling of the nasal septum. The rats had severe inflammation with erosions of the involved cartilage, which was characterized by increased numbers of CD4+ and CD8+ T cells in the lesions. The cartilage of the joints and ear pinna was not involved. All had IgG antibodies to matrilin-1. Matrilin-1 is a noncollagenous protein present in the extracellular matrix in cartilage. It is present in high concentrations in the trachea and is also present in the nasal septum but not in articular cartilage. A subsequent study demonstrated serum anti-matrilin-1 antibodies in ~13% of patients with relapsing polychondritis; ~70% of these patients had respiratory symptoms. Cell-mediated immunity may also be operative in causing tissue injury, since lymphocyte transformation can be demonstrated when lymphocytes of patients are exposed to cartilage extracts. T cells specific for type II collagen have been found in some patients, and CD4+ T cells have been observed at sites of cartilage inflammation.

CLINICAL MANIFESTATIONS

The onset of relapsing polychondritis is frequently abrupt, with the appearance of one or two sites of cartilaginous inflammation. The pattern of cartilaginous involvement and the frequency of episodes vary widely among patients. Noncartilaginous presentations may also occur. Systemic inflammatory features such as fever, fatigue, and weight loss occur and may precede the clinical signs of relapsing polychondritis by several weeks. Relapsing polychondritis may go unrecognized for several months or even years in patients who only initially manifest intermittent joint pain and/or swelling, or who have unexplained eye inflammation, hearing loss, valvular heart disease, or pulmonary symptoms.

Auricular chondritis is the most frequent presenting manifestation of relapsing polychondritis, occurring in 40% of patients and eventually affecting about 85% of patients **(Table 359-2)**. One or both ears are involved,

either sequentially or simultaneously. Patients experience the sudden onset of pain, tenderness, and swelling of the cartilaginous portion of the ear **(Fig. 359-1)**. This typically involves the pinna of the ears, sparing the earlobes because they do not contain cartilage. The overlying skin has a beefy red or violaceous color. Prolonged or recurrent episodes lead to cartilage destruction and result in a flabby or droopy ear. Swelling may close off the eustachian tube or the external auditory meatus, either of which can impair hearing. Inflammation of the internal auditory artery or its cochlear branch produces hearing loss, vertigo, ataxia, nausea, and vomiting. Vertigo is almost always accompanied by hearing loss.

TABLE 359-2

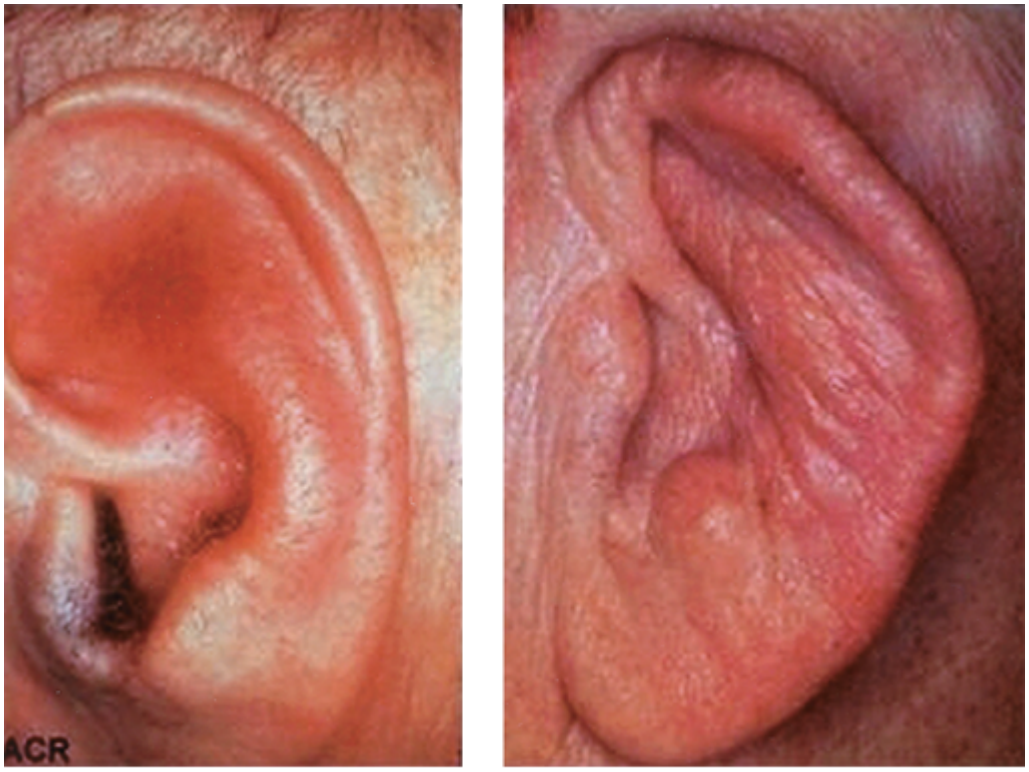
Clinical Manifestations of Relapsing Polychondritis

Clinical Feature	Presenting	Cumulative
	Frequency, %	
Auricular chondritis	43	89
Arthritis	32	72
Nasal chondritis	21	61
Ocular inflammation	18	59
Laryngotracheal symptoms	23	55
Reduced hearing	7	40
Saddle nose deformity	11	25
Cutaneous	4	25
Laryngotracheal stricture	15	23
Vasculitis	2	14
Elevated creatinine	7	13
Aortic or mitral regurgitation	0	12

Source: Modified from PD Kent et al: Curr Opin Rheumatol 16:56, 2004.

FIGURE 359-1

Left. The pinna is erythematous, swollen, and tender. Not shown is the ear lobule that is spared as there is no underlying cartilage. ***Right.*** The pinna is thickened and deformed. The destruction of the underlying cartilage results in a floppy ear. *(Reprinted from the Clinical Slide Collection on the Rheumatic Diseases, © 2018 American College of Rheumatology. Used by permission of the American College of Rheumatology.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Approximately 61% of patients will develop nasal involvement, with 21% having this at the time of presentation. Patients may experience nasal stuffiness, rhinorrhea, and epistaxis. The bridge of the nose and surrounding tissue become red, swollen, and tender and may collapse, producing a saddle nose deformity (Fig. 359-2). In some patients, nasal deformity develops insidiously without overt inflammation. Saddle nose is observed more frequently in younger patients, especially in women.

FIGURE 359-2

Saddle nose results from destruction and collapse of the nasal cartilage. *(Reprinted from the Clinical Slide Collection on the Rheumatic Diseases, © 2018 American College of Rheumatology. Used by permission of the American College of Rheumatology.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Joint involvement is the presenting manifestation in relapsing polychondritis in approximately one-third of patients and may be present for several months before other features appear. Eventually, more than one-half of the patients will have arthralgias or arthritis. The arthritis is usually asymmetric and oligo- or polyarticular, and it involves both large and small peripheral joints. An episode of arthritis lasts from a few days to several weeks and resolves spontaneously without joint erosion or deformity. Attacks of arthritis may not be temporally related to other manifestations of relapsing polychondritis. Joint fluid has been reported to be noninflammatory. In addition to peripheral joints, inflammation may involve the costochondral, sternomanubrial, and sternoclavicular cartilages. Destruction of these cartilages may result in a pectus excavatum deformity or even a flail anterior chest wall.

Eye manifestations occur in more than one-half of patients and include conjunctivitis, episcleritis, scleritis, iritis, uveitis, and keratitis. Ocular inflammation can be severe and visually threatening. Other manifestations include eyelid and periorbital edema, proptosis, optic neuritis, extraocular muscle palsies, retinal vasculitis, and renal vein occlusion.

Laryngotracheobronchial involvement occurs in ~50% of patients and is among the most serious manifestations of relapsing polychondritis. Symptoms include hoarseness, a nonproductive cough, and tenderness over the larynx and proximal trachea. Mucosal edema, strictures, and/or collapse of laryngeal or tracheal cartilage may cause stridor and life-threatening airway obstruction necessitating tracheostomy.

Involvement can extend into the lower airways resulting in tracheobronchomalacia. Collapse of cartilage in bronchi leads to pneumonia and, when extensive, to respiratory insufficiency.

Cardiac valvular regurgitation occurs in about 5–10% of patients and is due to progressive dilation of the valvular ring or to destruction of the valve cusps. Aortic regurgitation occurs in about 7% of patients, with the mitral and other heart valves being affected less often. Other cardiac manifestations include pericarditis, myocarditis, coronary vasculitis, and conduction abnormalities. Aneurysms of the proximal, thoracic, or abdominal aorta may occur even in the absence of active chondritis and occasionally rupture.

Renal disease occurs in about 10% of patients. The most common renal lesions include mesangial expansion or segmental necrotizing glomerulonephritis, which have been reported to have small amounts of electron-dense deposits in the mesangium where there is also faint deposition of C3 and/or IgG or IgM. Tubulointerstitial disease and IgA nephropathy have also been reported.

Approximately 25% of patients have skin lesions, which can include purpura, erythema nodosum, erythema multiforme, angioedema/urticaria, livedo reticularis, and panniculitis.

Features of vasculitis are seen in up to 25% of patients and can affect any size vessel. Large vessel vasculitis may present with aortic aneurysms, and medium vessel disease may affect the coronary, hepatic, mesenteric, or renal arteries or vessel supplying nerves. Skin vessel disease and involvement of the postcapillary venules can also occur. A variety of primary vasculitides have also been reported to occur in association with relapsing polychondritis **(Chap. 356)**. One specific overlap is the “MAGIC” syndrome (*m*outh *a*nd *g*enital ulcers with *i*nflamed *c*artilage) in which patients present with features of both relapsing polychondritis and Behçet’s disease **(Chap. 357)**.

LABORATORY FINDINGS AND DIAGNOSTIC IMAGING

There are no laboratory features that are diagnostic for relapsing polychondritis. Mild leukocytosis and normocytic, normochromic anemia are often present. Eosinophilia is observed in 10% of patients. The erythrocyte sedimentation rate and C-reactive protein are usually elevated. Rheumatoid factor and antinuclear antibody tests are occasionally positive in low titers, and complement levels are normal. Antibodies to type II collagen are present in fewer than one-half of the patients and are not specific. Circulating immune complexes may be detected, especially in patients with early active disease. Elevated levels of γ globulin may be present. Antineutrophil cytoplasmic antibodies (ANCA), either cytoplasmic (cANCA) or perinuclear (pANCA), are found in some patients with active disease. However, on target antigen–specific testing, there are only occasional reports of positive myeloperoxidase-ANCA, and proteinase 3-ANCA are very rarely found in relapsing polychondritis.

The upper and lower airways can be evaluated by imaging techniques such as computed tomography and magnetic resonance imaging (MRI). Bronchoscopy provides direct visualization of the airways but can be a high-risk procedure in patients with airway compromise. Pulmonary function testing with flow-volume loops can show inspiratory and/or expiratory obstruction. Imaging can also be useful to detect extracartilaginous

disease. The chest film may show widening of the ascending or descending aorta due to an aneurysm, and cardiomegaly when aortic insufficiency is present. MRI can assess aortic aneurysmal dilatation. Electrocardiography and echocardiography can be useful in further evaluating for cardiac features of disease.

DIAGNOSIS

Diagnosis is based on recognition of the typical clinical features. Biopsies of the involved cartilage from the ear, nose, or respiratory tract will confirm the diagnosis but are only necessary when clinical features are not typical. Diagnostic criteria were suggested in 1976 by McAdam et al and modified by Damiani and Levine in 1979. These criteria continue to be generally used in clinical practice. McAdam et al proposed the following: (1) recurrent chondritis of both auricles; (2) nonerosive inflammatory arthritis; (3) chondritis of nasal cartilage; (4) inflammation of ocular structures, including conjunctivitis, keratitis, scleritis/episcleritis, and/or uveitis; (5) chondritis of the laryngeal and/or tracheal cartilages; and (6) cochlear and/or vestibular damage manifested by neurosensory hearing loss, tinnitus, and/or vertigo. The diagnosis is certain when three or more of these features are present along with a positive biopsy from the ear, nasal, or respiratory cartilage. Damiani and Levine later suggested that the diagnosis could be made when one or more of the above features and a positive biopsy were present, when two or more separate sites of cartilage inflammation were present that responded to glucocorticoids or dapsone, or when three or more of the above features were present.

The differential diagnosis of relapsing polychondritis is centered around its sites of clinical involvement. Patients with granulomatosis with polyangiitis (Wegener's) may have a saddle nose and tracheal involvement but can be distinguished by the primary inflammation occurring in the mucosa at these sites, the absence of auricular involvement, and the presence of pulmonary parenchymal disease. Patients with Cogan's syndrome have interstitial keratitis and vestibular and auditory abnormalities, but this syndrome does not involve the respiratory tract or ears. Reactive arthritis may initially resemble relapsing polychondritis because of oligoarticular arthritis and eye involvement, but it is distinguished in time by the appearance of urethritis and typical mucocutaneous lesions and the absence of nose or ear cartilage involvement. Rheumatoid arthritis may initially suggest relapsing polychondritis because of arthritis and eye inflammation. The arthritis in rheumatoid arthritis, however, is erosive and symmetric. In addition, rheumatoid factor titers are usually high compared with those in relapsing polychondritis, and anti-cyclic citrullinated peptide is usually not seen. Bacterial infection of the pinna may be mistaken for relapsing polychondritis but differs by usually involving only one ear, including the earlobe. Auricular cartilage may also be damaged by trauma or frostbite. Nasal destructive disease and auricular abnormalities can also be seen in patients using cocaine adulterated with levamisole. Ear involvement in this setting differs from relapsing polychondritis by typically manifesting as purpuric plaques with necrosis extending to the pinna, which does not contain cartilage.

TREATMENT

TREATMENT

In patients with active chondritis, prednisone, 40–60 mg/d, is often effective in suppressing disease activity; it is tapered gradually once disease is controlled. In some patients, prednisone can be stopped, whereas in others, low doses in the range of 5–10 mg/d are required for continued suppression of disease. Dapsone 50–100 mg/d has been effective for cartilage inflammation and joint features in some patients; however, its use can be limited by the complication of hemolytic anemia and other side effects as well as contraindications in glucose-6-phosphate dehydrogenase (G6PD) deficiency and pregnancy. Other immunosuppressive drugs such as cyclophosphamide, methotrexate, azathioprine, or cyclosporine should be reserved for patients who have severe organ-threatening disease, fail to respond to prednisone, or require high doses to control disease activity. Patients with significant ocular inflammation often require intraocular glucocorticoids as well as high doses of prednisone. There are a small number of reports on the use of tumor necrosis factor antagonists, rituximab (anti-CD20), and tocilizumab (anti-interleukin 6 receptor), which are too few in number to assess efficacy. Heart valve replacement or repair of an aortic aneurysm may be necessary. When airway obstruction is severe, tracheostomy is required. Stents may be necessary in patients with tracheobronchial collapse.

PATIENT OUTCOME, PROGNOSIS, AND SURVIVAL

The course of relapsing polychondritis is highly variable. Some patients experience inflammatory episodes lasting from a few days to several weeks that then subside spontaneously or with treatment. Attacks may recur at intervals varying from weeks to months. In other patients, the disease has a chronic, smoldering course that may be severe. In one study, the 5-year estimated survival rate was 74% and the 10-year survival rate was 55%. About one-half of the deaths could be attributed to relapsing polychondritis or complications of treatment. Airway complications accounted for 10% of all fatalities although higher rates have been reported in other series. In general, patients with more widespread disease have a worse prognosis.

FURTHER READING

Chopra R et al: Relapsing polychondritis. Rheum Dis Clin North Am 39:263, 2013. [PubMed: 23597963]

Ernst A et al: Relapsing polychondritis and airway involvement. Chest 135:1024, 2009. [PubMed: 19017885]

Kent PD et al: Relapsing polychondritis. Curr Opin Rheumatol 16:56, 2004. [PubMed: 14673390]

Longo L et al: Relapsing polychondritis: A clinical update. Autoimmun Rev 15:539, 2016. [PubMed: 26876384]

Trentham DE, Le CH: Relapsing polychondritis. Ann Intern Med 129:114, 1998. [PubMed: 9669970]

Chapter 360: Sarcoidosis

Robert P. Baughman; Elyse E. Lower

DEFINITION

Sarcoidosis is an inflammatory disease characterized by the presence of noncaseating granulomas. The disease is often multisystem and requires the presence of involvement in two or more organs for a specific diagnosis. The finding of granulomas is not specific for sarcoidosis, and other conditions known to cause granulomas must be ruled out. These conditions include mycobacterial and fungal infections, malignancy, and environmental agents such as beryllium. Although sarcoidosis can affect virtually every organ of the body, the lung is most commonly affected. Other organs commonly affected are the liver, skin, and eye. The clinical outcome of sarcoidosis varies, with remission occurring in over one-half of patients within a few years of diagnosis; however, the remaining patients may develop a chronic disease that lasts for decades.

ETIOLOGY

Despite multiple investigations, the cause of sarcoidosis remains unknown. Currently, the most likely etiology is an infectious or noninfectious environmental agent that triggers an inflammatory response in a genetically susceptible host. Among the possible infectious agents, careful studies have shown a much higher incidence of *Propionibacter acnes* in the lymph nodes of sarcoidosis patients compared to controls. An animal model has shown that *P. acnes* can induce a granulomatous response in mice similar to sarcoidosis. Others have demonstrated the presence of a mycobacterial protein (*Mycobacterium tuberculosis* catalase-peroxidase [mKatG]) in the granulomas of some sarcoidosis patients. This protein is very resistant to degradation and may represent the persistent antigen in sarcoidosis. Immune response to this and other mycobacterial proteins has been documented by another laboratory. These studies suggest that a mycobacterium similar to *M. tuberculosis* could be responsible for sarcoidosis. The mechanism exposure/infection with such agents has been the focus of other studies. Environmental exposures to insecticides and mold have been associated with an increased risk for disease. In addition, health care workers appear to have an increased risk. Also, sarcoidosis in a donor organ has occurred after transplantation into a sarcoidosis patient. Some authors have suggested that sarcoidosis is not due to a single agent but represents a particular host response to multiple agents. Some studies have been able to correlate the environmental exposures to genetic markers. These studies have supported the hypothesis that a genetically susceptible host is a key factor in the disease.

INCIDENCE, PREVALENCE, AND GLOBAL IMPACT

Sarcoidosis is seen worldwide, with the highest prevalence reported in the Nordic population. In the United States, the disease has been reported more commonly in African Americans than whites, with the ratio of African Americans to whites ranging from 3:1 to 17:0. In the United Sates, women are more susceptible than men. The higher incidence in African Americans may have been influenced by the fact that African Americans seem to develop more extensive and chronic pulmonary disease. Because most sarcoidosis clinics are run by pulmonologists, a selection bias may have occurred. Worldwide, the prevalence of the disease varies from 20–60 per 100,000 for many groups such as Japanese, Italians, and American whites. Higher rates occur in Ireland and Nordic countries. In one closely observed community in Sweden, the lifetime risk for developing sarcoidosis was 3%.

Sarcoidosis often occurs in young, otherwise healthy adults. It is uncommon to diagnose the disease in someone aged <18 years. However, it has become clear that a second peak in incidence develops around age 60. In a study of nearly 30,000 sarcoidosis patients in the United States, the median age at diagnosis was 55.

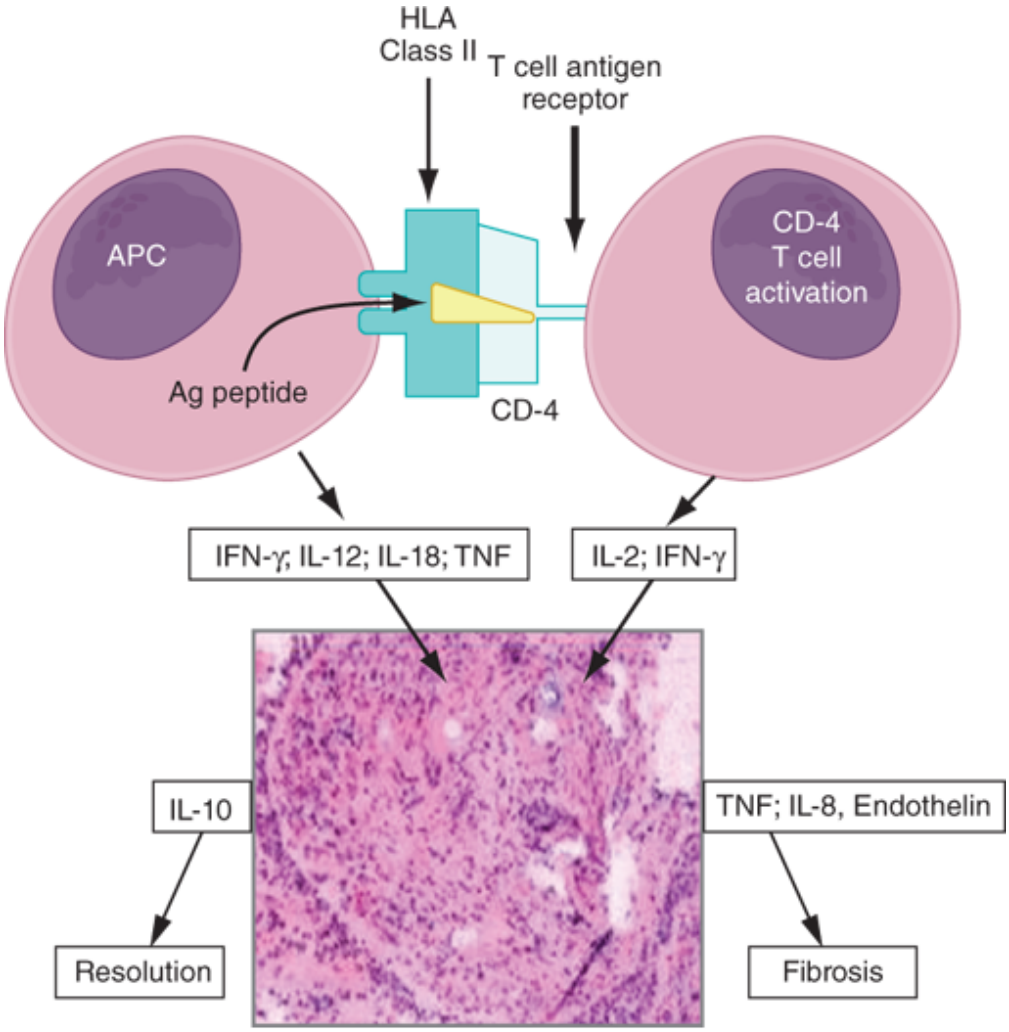
Although most cases of sarcoidosis are sporadic, a familial form of the disease exists. At least 5% of patients with sarcoidosis will have a family member with sarcoidosis. Sarcoidosis patients who are Irish or African American seem to have a two to three times higher rate of familial disease.

PATHOPHYSIOLOGY AND IMMUNOPATHOGENESIS

The granuloma is the pathologic hallmark of sarcoidosis. A distinct feature of sarcoidosis is the local accumulation of inflammatory cells. Extensive studies in the lung using bronchoalveolar lavage (BAL) have demonstrated that the initial inflammatory response is an influx of T helper cells. In addition, there is an accumulation of activated monocytes. Figure 360-1 is a proposed model for sarcoidosis. Using the HLA-CD4 complex, antigen-presenting cells present an unknown antigen to the helper T cell. Studies have clarified that specific HLA haplotypes such as HLA-DRB1*1101 are associated with an increased risk for developing sarcoidosis. In addition, different HLA haplotypes are associated with different clinical outcomes.

FIGURE 360-1

Schematic representation of initial events of sarcoidosis. The antigen-presenting cell and helper T cell complex leads to the release of multiple cytokines. This forms a granuloma. Over time, the granuloma may resolve or lead to chronic disease, including fibrosis. APC, antigen-presenting cell; HLA, human leukocyte antigen; IFN, interferon; IL, interleukin; TNF, tumor necrosis factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The macrophage/helper T cell cluster leads to activation with the increased release of several cytokines. These include interleukin (IL)-2 released from the T cell and interferon γ and tumor necrosis factor (TNF) released by the macrophage. The T cell is a necessary part of the initial inflammatory response. In advanced, untreated HIV infection, patients who lack helper T cells rarely develop sarcoidosis. In contrast, several reports confirm that sarcoidosis becomes unmasked as HIV-infected individuals receive antiretroviral therapy, with subsequent restoration of their immune system. In contrast, treatment of established pulmonary sarcoidosis with cyclosporine, a drug that downregulates helper T cell responses, seems to have little impact on sarcoidosis.

The granulomatous response of sarcoidosis can resolve with or without therapy. However, in at least 20% of patients with sarcoidosis, a chronic form of the disease develops. This persistent form of the disease is associated with increased levels in blood and/or BAL of IL-8, IL-17, and CXCL9. Also, studies have reported that patients with this chronic form of disease release excessive amounts of TNF in areas of inflammation. Specific gene signatures have been associated with more severe disease, such as cardiac, neurologic, and fibrotic pulmonary disease.

At diagnosis the natural history of the disease may be difficult to predict. One form of the disease, *Löfgren's syndrome*, consists of erythema nodosum and hilar adenopathy on chest roentgenogram. In some cases, periarticular arthritis may be identified without erythema nodosum. Löfgren's syndrome is associated with a

good prognosis, with >90% of patients experiencing disease resolution within 2 years. Recent studies have demonstrated that the HLA-DRB1*03 was found in two-thirds of Scandinavian patients with Löfgren's syndrome. More than 95% of those patients who were HLA-DRB1*03 positive had resolution of their disease within 2 years, whereas nearly one-half of the remaining patients had disease for >2 years. It remains to be determined whether these observations can be applied to a non-Scandinavian population.

CLINICAL MANIFESTATIONS

The presentation of sarcoidosis ranges from patients who are asymptomatic to those with organ failure. It is unclear how often sarcoidosis is asymptomatic. In countries where routine chest roentgenogram screening is performed, 20–30% of pulmonary cases are detected in asymptomatic individuals. The inability to screen for other asymptomatic forms of the disease would suggest that as many as one-third of sarcoidosis patients are asymptomatic.

Respiratory complaints including cough and dyspnea are the most common presenting symptoms. In many cases, the patient presents with a 2- to 4-week history of these symptoms. Unfortunately, due to the nonspecific nature of pulmonary symptoms, the patient may see physicians for up to a year before a diagnosis is confirmed. For these patients, the diagnosis of sarcoidosis is usually only suggested when a chest roentgenogram is performed.

Symptoms related to cutaneous and ocular disease are the next two most common complaints. Skin lesions are often nonspecific. However, because these lesions are readily observed, the patient and treating physician are often led to a diagnosis. In contrast to patients with pulmonary disease, patients with cutaneous lesions are more likely to be diagnosed within 6 months of symptoms.

Nonspecific constitutional symptoms include fatigue, fever, night sweats, and weight loss. Fatigue is perhaps the most common constitutional symptom that affects these patients. Given its insidious nature, patients are usually not aware of the association with their sarcoidosis until their disease resolves.

The overall incidence of sarcoidosis at the time of diagnosis and eventual common organ involvement are summarized in Table 360-1. Over time, skin, eye, and neurologic involvement seem more apparent. In the United States, the frequency of specific organ involvement appears to be affected by age, race, and gender. For example, eye disease is more common among African Americans. Under the age of 40, it occurs more frequently in women. However, in those diagnosed over the age of 40, eye disease is more common in men.

TABLE 360-1

Frequency of Common Organ Involvement and Lifetime Risk[a]

	Presentation, %[b]	Follow-Up, %[c]
Lung	95	94
Skin	24	43
Eye	12	29
Extrathoracic lymph node	15	16
Liver	12	14
Spleen	7	8
Neurologic	5	16
Cardiac	2	3

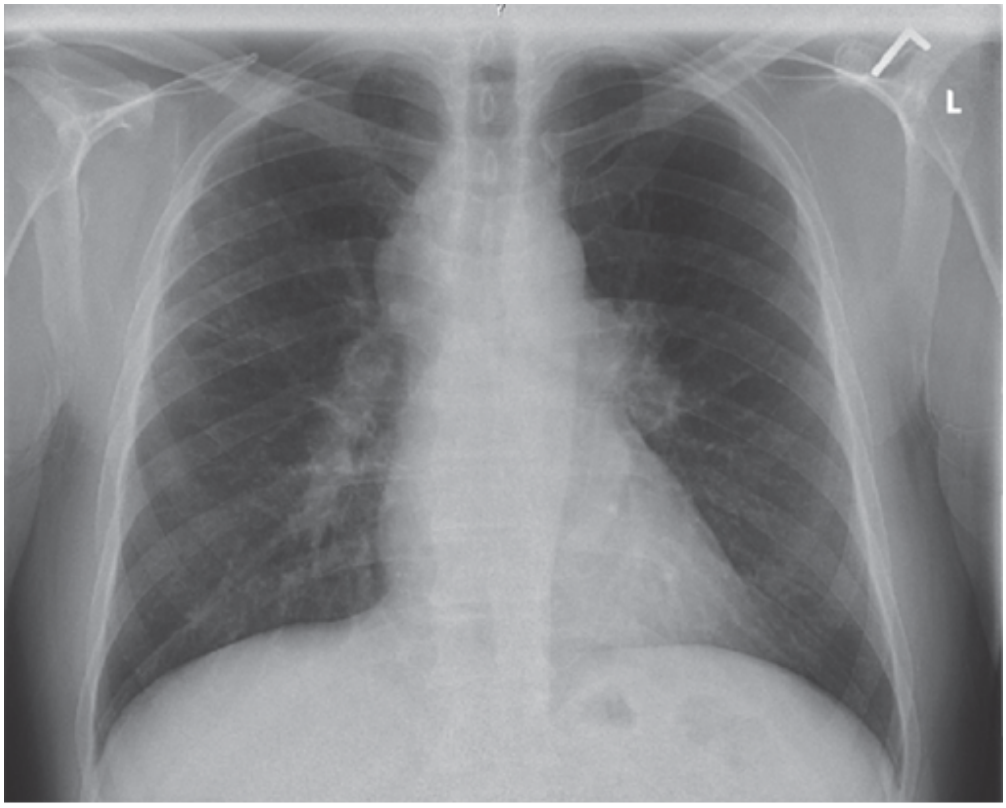
[a]Patients could have more than one organ involved. [b]From ACCESS study of 736 patients evaluated within 6 months of diagnosis. [c]From follow-up of 1024 sarcoidosis patients seen at the University of Cincinnati Interstitial Lung Disease and Sarcoidosis Clinic from 2002 to 2006.

LUNG

Lung involvement occurs in >90% of sarcoidosis patients. The most commonly used method for detecting lung disease is still the chest roentgenogram. Figure 360-2 illustrates the chest roentgenogram from a sarcoidosis patient with bilateral hilar adenopathy. Although the computed tomography (CT) scan has changed the diagnostic approach to interstitial lung disease, the CT scan is not usually considered a monitoring tool for patients with sarcoidosis. Figure 360-3 demonstrates some of the characteristic CT features, including peribronchial thickening and reticular nodular changes, which are predominantly subpleural. The peribronchial thickening seen on CT scan seems to explain the high yield of granulomas from bronchial biopsies performed for diagnosis.

FIGURE 360-2

Posterior-anterior chest roentgenogram demonstrating bilateral hilar adenopathy, stage 1 disease.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 360-3

High-resolution computed tomography scan of chest demonstrating patchy reticular nodularity, including areas of confluence.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although the CT scan is more sensitive, the standard scoring system described by Scadding in 1961 for chest roentgenograms remains the preferred method of characterizing the chest involvement. Stage 1 is hilar adenopathy alone (Fig. 360-2), often with right paratracheal involvement. Stage 2 is a combination of adenopathy plus infiltrates, whereas stage 3 reveals infiltrates alone. Stage 4 consists of fibrosis. Usually the infiltrates in sarcoidosis are predominantly an upper lobe process. Only in a few noninfectious diseases is an upper lobe predominance noted. In addition to sarcoidosis, the differential diagnosis of upper lobe disease includes hypersensitivity pneumonitis, silicosis, and Langerhans cell histiocytosis. For infectious diseases, tuberculosis and *Pneumocystis* pneumonia can often present as upper lobe diseases.

Lung volumes, mechanics, and diffusion are all useful in evaluating interstitial lung diseases such as sarcoidosis. The diffusion of carbon monoxide (DL_{CO}) is the most sensitive test for an interstitial lung disease.

Reduced lung volumes are a reflection of the restrictive lung disease seen in sarcoidosis. However, a third of the patients presenting with sarcoidosis still have lung volumes within the normal range, despite abnormal chest roentgenograms and dyspnea.

Approximately one-half of sarcoidosis patients present with obstructive disease, reflected by a reduced ratio of forced vital capacity expired in 1 second (FEV_1/FVC). Cough is a very common symptom. Airway involvement causing varying degrees of obstruction underlies the cough in most sarcoidosis patients. Airway hyperreactivity, as determined by methacholine challenge, will be positive in some of these patients. A few patients with cough will respond to traditional bronchodilators as the only form of treatment. In some cases, high-dose inhaled

glucocorticoids alone are useful. Airway obstruction can be due to large airway stenosis, which can become fibrotic and unresponsive to anti-inflammatory therapy.

Pulmonary arterial hypertension is reported in at least 5% of sarcoidosis patients. Either direct vascular involvement or the consequence of fibrotic changes in the lung can lead to pulmonary arterial hypertension. In sarcoidosis patients with end-stage fibrosis awaiting lung transplant, 70% will have pulmonary arterial hypertension. This is a much higher incidence than that reported for other fibrotic lung diseases. In less advanced, but still symptomatic, patients, pulmonary arterial hypertension has been noted in up to 50% of the cases. Because sarcoidosis-associated pulmonary arterial hypertension may respond to therapy, evaluation for this should be considered in persistently dyspneic patients.

SKIN

Skin involvement is eventually identified in over a third of patients with sarcoidosis. The classic cutaneous lesions include erythema nodosum, maculopapular lesions, hyper- and hypopigmentation, keloid formation, and subcutaneous nodules. A specific complex of involvement of the bridge of the nose, the area beneath the eyes, and the cheeks is referred to as *lupus pernio* (Fig. 360-4) and is diagnostic for a chronic form of sarcoidosis.

FIGURE 360-4

Chronic inflammatory lesions around nose, eyes, and cheeks, referred to as lupus pernio.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In contrast, erythema nodosum is a transient rash that can be seen in association with hilar adenopathy and uveitis (Löfgren's syndrome). Erythema nodosum is more common in women and in certain self-described demographic groups including whites and Puerto Ricans. In the United States, the other manifestations of skin sarcoidosis, especially lupus pernio, are more common in African Americans than whites.

The maculopapular lesions from sarcoidosis are the most common chronic form of the disease **(Fig. 360-5)**. These are often overlooked by the patient and physician, because they are chronic and not painful. Initially, these lesions are usually purplish papules and are often indurated. They can become confluent and infiltrate large areas of the skin. With treatment, the color and induration may fade. Because these lesions are caused by noncaseating granulomas, the diagnosis of sarcoidosis can be readily made by a skin biopsy.

FIGURE 360-5
Maculopapular lesions on the trunk of a sarcoidosis patient.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

EYE

The frequency of ocular manifestations for sarcoidosis varies depending on race. In Japan, >70% of sarcoidosis patients develop ocular disease, whereas in the United States only 30% have eye disease, with problems more common in African Americans than whites. Although the most common manifestation is an anterior uveitis, over a quarter of patients will have inflammation at the posterior of the eye, including retinitis and pars planitis. Although symptoms such as photophobia, blurred vision, and increased tearing can occur, some asymptomatic patients still have active inflammation. Initially asymptomatic patients with ocular sarcoidosis can eventually develop blindness. Therefore, it is recommended that all patients with sarcoidosis receive a dedicated ophthalmologic examination. Sicca is seen in over one-half of the chronic sarcoidosis patients. Dry eyes appear to be a reflection of prior lacrimal gland disease. Although the patient may no longer have active inflammation, the dry eyes may require natural tears or other lubricants.

LIVER

Using biopsies to detect granulomatous disease, liver involvement can be identified in over one-half of sarcoidosis patients. However, using liver function studies, only 20–30% of patients will have evidence of liver involvement. The most common abnormality of liver function is an elevation of the alkaline phosphatase level, consistent with an obstructive pattern. In addition, elevated transaminase levels can occur. An elevated bilirubin level is a marker for more advanced liver disease. Overall, only 5% of sarcoidosis patients have sufficient symptoms from their liver disease to require specific therapy. Although symptoms can be due to hepatomegaly, more frequently symptoms result from extensive intrahepatic cholestasis leading to portal

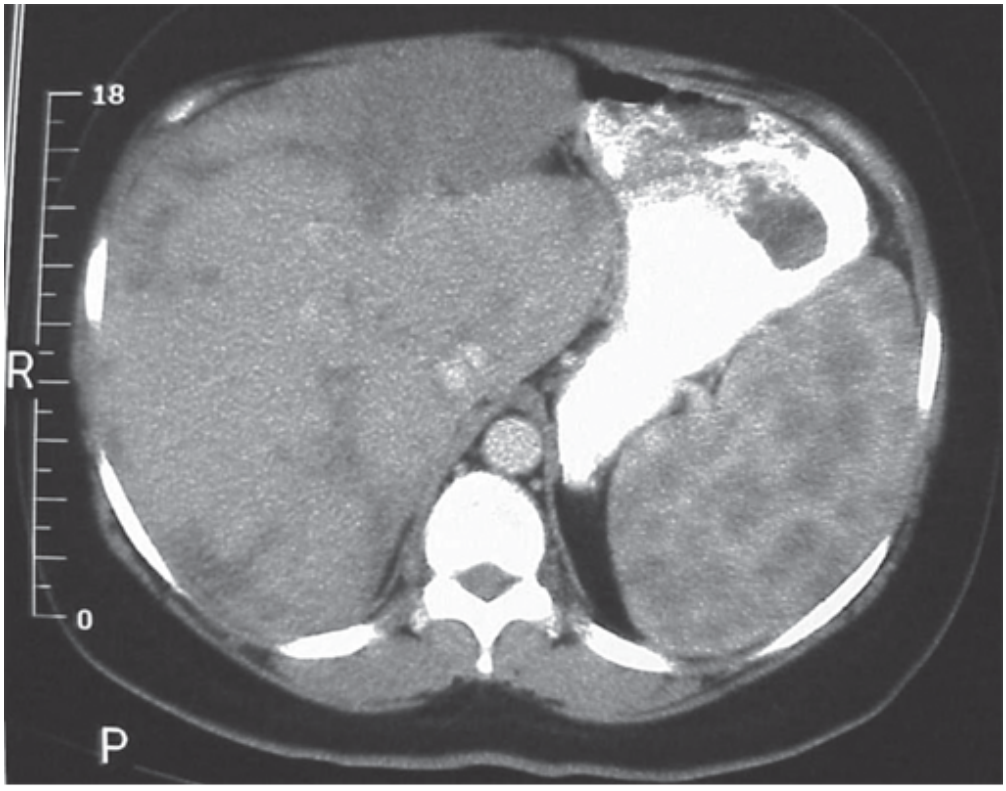
hypertension. In this case, ascites and esophageal varices can occur. It is rare that a sarcoidosis patient will require a liver transplant, because even the patient with cirrhosis due to sarcoidosis can respond to systemic therapy.

BONE MARROW AND SPLEEN

One or more bone marrow manifestations can be identified in many sarcoidosis patients. The most common hematologic problem is lymphopenia, which is a reflection of sequestration of the lymphocytes into the areas of inflammation. Anemia occurs in 20% of patients, and leukopenia is less common. Bone marrow examination will reveal granulomas in about a third of patients. Although splenomegaly can be detected in 5–10% of patients, splenic biopsy reveals granulomas in 60% of patients. The CT scan can be relatively specific for sarcoidosis involvement of the spleen (Fig. 360-6). Both bone marrow and spleen involvement are more common in African Americans than whites. Although these manifestations alone are rarely an indication for therapy, on rare occasion, splenectomy may be indicated for massive symptomatic splenomegaly or profound pancytopenia. Nonthoracic lymphadenopathy can occur in up to 20% of patients.

FIGURE 360-6

Computed tomography scan of the abdomen after oral and intravenous contrast. The stomach is compressed by the enlarged spleen. Within the spleen, areas of hypo- and hyperdensity are identified.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CALCIUM METABOLISM

Hypercalcemia and/or hypercalciuria occur in about 10% of sarcoidosis patients. It is more common in whites than African Americans and in men. The mechanism of abnormal calcium metabolism is increased production of 1,25-dihydroxyvitamin D by the granuloma itself. The 1,25-dihydroxyvitamin D causes increased intestinal absorption of calcium, leading to hypercalcemia with a suppressed parathyroid hormone (PTH) level **(Chap. 403)**. Increased exogenous vitamin D from diet or sunlight exposure may exacerbate this problem. Serum calcium should be determined as part of the initial evaluation of all sarcoidosis patients, and a repeat determination may be useful during the summer months with increased sun exposure. In patients with a history of renal calculi, a 24-h urine calcium measurement should be obtained. If a sarcoidosis patient with a history of renal calculi is to be placed on calcium supplements, a follow-up 24-h urine calcium level should be measured.

RENAL DISEASE

Direct kidney involvement occurs in <5% of sarcoidosis patients. It is associated with granulomas in the kidney itself and can lead to nephritis. However, hypercalcemia is the most likely cause of sarcoidosis-associated renal disease. In 1–2% of sarcoidosis patients, acute renal failure may develop as a result of hypercalcemia. Successful treatment of hypercalcemia with glucocorticoids and other therapies often improves but usually does not totally resolve the renal dysfunction.

NERVOUS SYSTEM

Neurologic disease is reported in 5–10% of sarcoidosis patients and appears to be of equal frequency across all ethnic groups. Any part of the central or peripheral nervous system can be affected. The presence of granulomatous inflammation is often visible on magnetic resonance imaging (MRI) studies. The MRI with gadolinium enhancement may demonstrate space-occupying lesions, but the MRI can be negative due to small lesions or the effect of systemic therapy in reducing the inflammation. The cerebral spinal fluid (CSF) findings include lymphocytic meningitis with a mild increase in protein. The CSF glucose is usually normal but can be low. Certain areas of the nervous system are more commonly affected in neurosarcoidosis. These include cranial nerve involvement, basilar meningitis, myelopathy, and anterior hypothalamic disease with associated diabetes insipidus **(Chap. 374)**. Seizures and cognitive changes also occur. Of the cranial nerves, seventh nerve paralysis can be transient and mistaken for Bell's palsy (idiopathic seventh nerve paralysis). Because this form of neurosarcoidosis often resolves within weeks and may not recur, it may have occurred prior to a definitive diagnosis of sarcoidosis. Optic neuritis is another cranial nerve manifestation of sarcoidosis. This manifestation is more chronic and usually requires long-term systemic therapy. It can be associated with both anterior and posterior uveitis. Differentiating between neurosarcoidosis and multiple sclerosis can be difficult at times. Optic neuritis can occur in both diseases. In some patients with sarcoidosis, multiple enhancing white matter abnormalities may be detected by MRI, suggesting multiple sclerosis. In such cases, the presence of meningeal enhancement or hypothalamic involvement suggests neurosarcoidosis, as does evidence of extraneurologic disease such as pulmonary or skin involvement, which also suggests sarcoidosis. Because the response of neurosarcoidosis to glucocorticoids and cytotoxic therapy is different from that of multiple sclerosis, differentiating between these disease entities is important.

CARDIAC

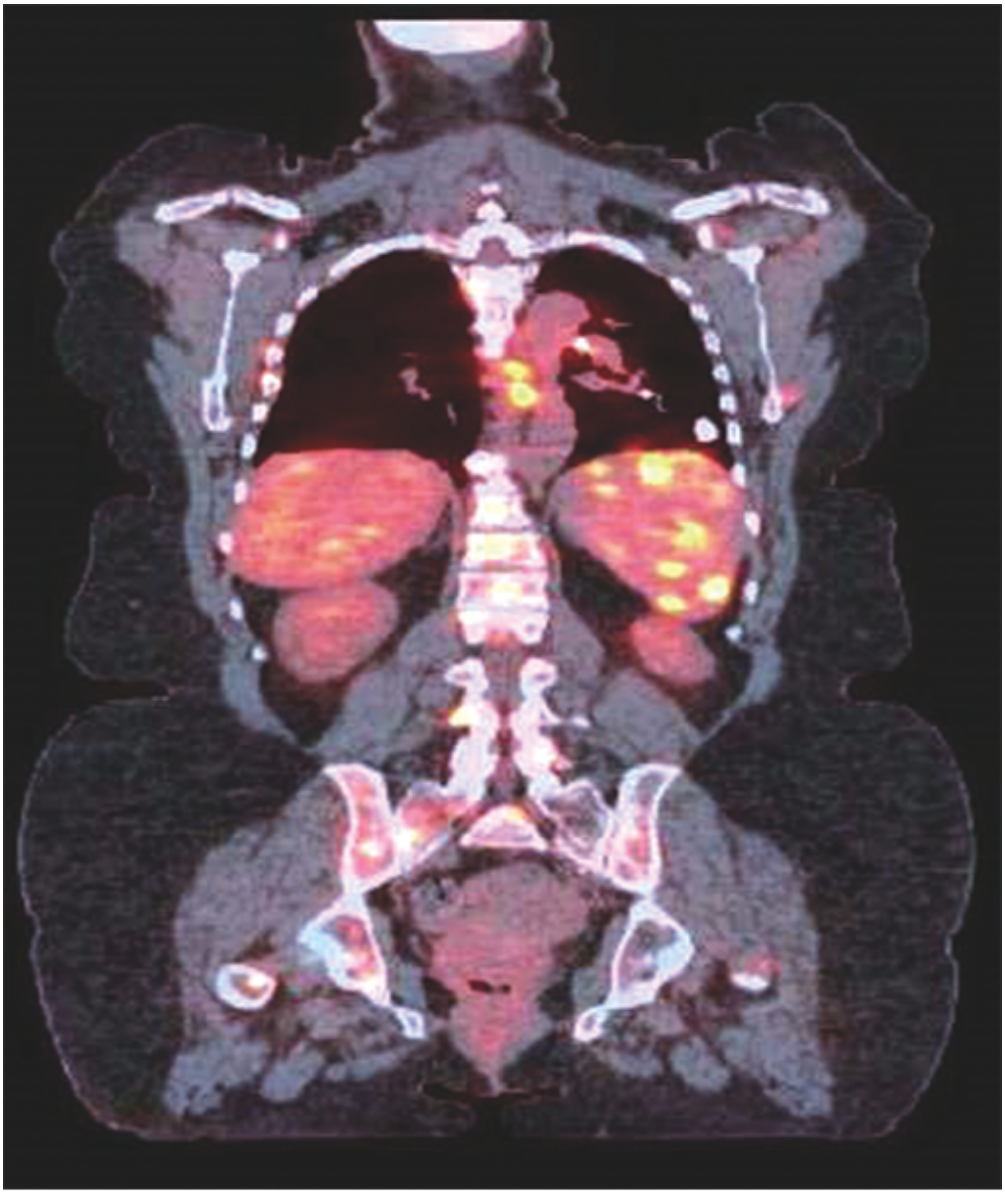
The presence of cardiac involvement is influenced by race. Although over a quarter of Japanese sarcoidosis patients develop cardiac disease, only 5% of sarcoidosis patients in the United States and Europe develop symptomatic cardiac disease. However, there is no apparent racial predilection between whites and African Americans. Cardiac disease, which usually presents as either congestive heart failure or cardiac arrhythmias, results from infiltration of the heart muscle by granulomas. Diffuse granulomatous involvement of the heart muscle can lead to profound dysfunction with left ventricular ejection fractions <10%. Even in this situation, improvement in the ejection fraction can occur with systemic therapy. Arrhythmias can also occur with diffuse infiltration or with more patchy cardiac involvement. If the atrioventricular (AV) node is infiltrated, heart block can occur, which can be detected by routine electrocardiography. Ventricular arrhythmias and sudden death due to ventricular tachycardia are common causes of death. Arrhythmias are best detected using 24-h ambulatory monitoring, and electrophysiology studies may be negative. Other screening tests for cardiac disease include routine electrocardiography and echocardiography. The confirmation of cardiac sarcoidosis is usually performed with either MRI or positron emission tomography (PET) scanning. Because ventricular arrhythmias are usually multifocal due to patchy multiple granulomas in the heart, ablation therapy is not useful. Patients with significant ventricular arrhythmias should be considered for an implanted defibrillator, which appears to have reduced the rate of death in cardiac sarcoidosis. Although systemic therapy can be useful in treating the arrhythmias, patients may still have malignant arrhythmias up to 6 months after starting successful treatment, and the risk for recurrent arrhythmias occurs whenever medications are tapered.

MUSCULOSKELETAL SYSTEM

Direct granulomatous involvement of bone and muscle can be documented by radiography (x-ray, MRI, PET scan **[Fig. 360-7]**, or gallium scan) or confirmed by biopsy in about 10% of sarcoidosis patients. However, a larger percentage of sarcoidosis patients complain of myalgias and arthralgias. These complaints are similar to those reported by patients with other inflammatory diseases, including chronic infections such as mononucleosis. Fatigue associated with sarcoidosis may be overwhelming for many patients. Recent studies have demonstrated a link between fatigue and small peripheral nerve fiber disease in sarcoidosis.

FIGURE 360-7

Positron emission tomography and computed tomography scan merged demonstrating increased activity in spleen, ribs, and spine of patient with sarcoidosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

OTHER ORGAN INVOLVEMENT

Although sarcoidosis can affect any organ of the body, rarely does it involve the breast, testes, ovary, or stomach. Because of the rarity of involvement, a mass in one of these areas requires a biopsy to rule out other diseases including cancer. For example, in a study of breast problems in female sarcoidosis patients, a breast lesion was more likely to be a granuloma from sarcoidosis than from breast cancer. However, findings on the physical examination or mammogram cannot reliably differentiate between these lesions. More importantly, as women with sarcoidosis age, breast cancer becomes more common. Therefore, it is recommended that routine screening including mammography be performed along with other imaging studies (ultrasound, MRI) or biopsy as clinically indicated.

COMPLICATIONS

Sarcoidosis is usually a self-limited, non-life-threatening disease. However, organ-threatening disease can occur. These complications can include blindness, paraplegia, or renal failure. Death from sarcoidosis occurs in about 5% of patients seen in sarcoidosis referral clinics. The usual causes of death related to sarcoidosis are from lung, cardiac, neurologic, or liver involvement. In respiratory failure, an elevation of the right atrial pressure is a poor prognostic finding. Lung complications can also include infections such as mycetoma, which can subsequently lead to massive bleeding. In addition, the use of immunosuppressive agents can increase the incidence of serious infections.

LABORATORY FINDINGS

The chest roentgenogram remains the most commonly used tool to assess lung involvement in sarcoidosis. As noted above, the chest roentgenogram classifies involvement into four stages, with stages 1 and 2 having hilar and paratracheal adenopathy. The CT scan has been used increasingly in evaluating interstitial lung disease. In sarcoidosis, the presence of adenopathy and a nodular infiltrate is not specific for sarcoidosis. Adenopathy up to 2 cm can be seen in other inflammatory lung diseases such as idiopathic pulmonary fibrosis. However, adenopathy >2 cm in the short axis supports the diagnosis of sarcoidosis over other interstitial lung diseases.

The PET scan has increasingly replaced gallium-67 scanning to identify areas of granulomatous disease in the chest and other parts of the body (Fig. 360-7). Both tests can be used to identify potential areas for biopsy. Cardiac PET scanning has also proved useful in assessing cardiac sarcoidosis. The identification of hypermetabolic activity may be due to the granulomas from sarcoidosis and not to disseminated malignancy.

MRI has also proved useful in the assessment of extrapulmonary sarcoidosis. Gadolinium enhancement has been demonstrated in areas of inflammation in the brain, heart, and bone. MRI scans may detect asymptomatic lesions. Like PET scan, MRI changes appear similar to those seen with malignancy and infection. In some cases, biopsy may be necessary to determine the cause of the radiologic abnormality.

Serum levels of angiotensin-converting enzyme (ACE) can be helpful in the diagnosis of sarcoidosis. However, the test has somewhat low sensitivity and specificity. Elevated levels of ACE are reported in 60% of patients with acute disease and only 20% of patients with chronic disease. Although there are several causes for mild elevation of ACE, including diabetes, elevations of >50% of the upper limit of normal are seen in only a few conditions including sarcoidosis, leprosy, Gaucher's disease, hyperthyroidism, and disseminated granulomatous infections such as miliary tuberculosis. Because the ACE level is determined by a biologic assay, the concurrent use of an ACE inhibitor such as lisinopril will lead to a very low ACE level.

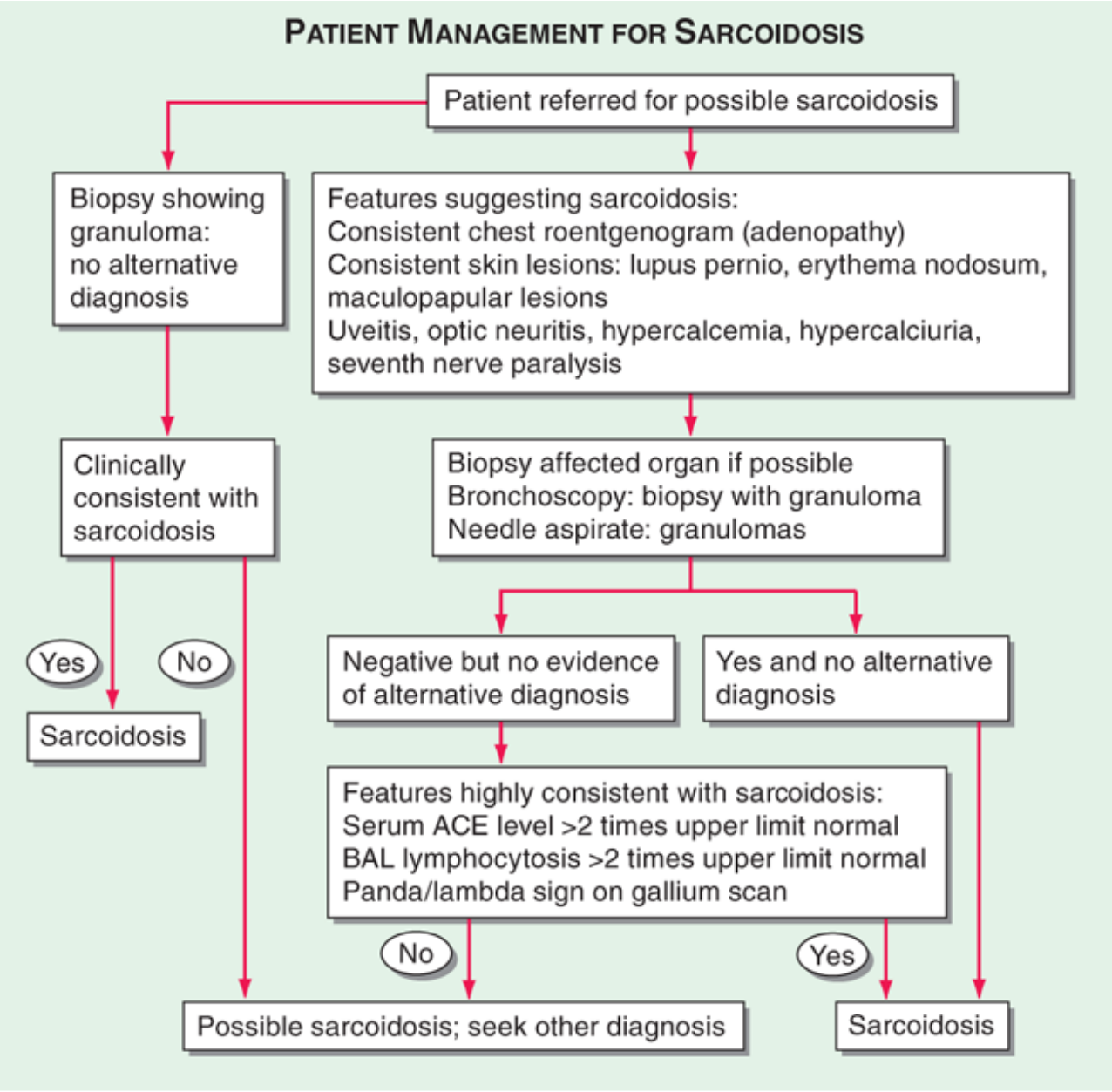
DIAGNOSIS

The diagnosis of sarcoidosis requires both compatible clinical features and pathologic findings. Because the cause of sarcoidosis remains elusive, the diagnosis cannot be made with 100% certainty. Nevertheless, the diagnosis can be made with reasonable certainty based on history and physical features along with laboratory and pathologic findings.

Patients are usually evaluated for possible sarcoidosis based on two scenarios (Fig. 360-8). In the first scenario, a patient may undergo a biopsy revealing a noncaseating granuloma in either a pulmonary or an extrapulmonary organ. If the clinical presentation is consistent with sarcoidosis and there is no alternative cause for the granulomas identified, then the patient is felt to have sarcoidosis.

FIGURE 360-8

Proposed approach to management of patient with possible sarcoidosis. Presence of one or more of these features supports the diagnosis of sarcoidosis: uveitis, optic neuritis, hypercalcemia, hypercalciuria, seventh cranial nerve paralysis, diabetes insipidus. ACE, angiotensin-converting enzyme; BAL, bronchoalveolar lavage.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In the second scenario, signs or symptoms suggesting sarcoidosis such as the presence of bilateral adenopathy may be present in an otherwise asymptomatic patient or a patient with uveitis or a rash consistent with sarcoidosis. At this point, a diagnostic procedure should be performed. For the patient with a compatible skin

lesion, a skin biopsy should be considered. Other biopsies to consider could include liver, extrathoracic lymph node, or muscle. In some cases, a biopsy of the affected organ may not be easy to perform (such as a brain or spinal cord lesion). In other cases, such as an endomyocardial biopsy, the likelihood of a positive biopsy is low. Because of the high rate of pulmonary involvement in these cases, the lung may be easier to approach by bronchoscopy. During the bronchoscopy, a transbronchial biopsy, bronchial biopsy, or transbronchial needle aspirate can be performed. The endobronchial ultrasonography-guided (EBUS) transbronchial needle aspirate can assist in diagnosing sarcoidosis in patients with mediastinal adenopathy (stage 1 or 2 radiographic pulmonary disease), whereas transbronchial biopsy has a higher diagnostic yield for those with only parenchymal lung disease (stage 3). These tests are complementary and may be performed together.

If the biopsy reveals granulomas, an alternative diagnosis such as infection or malignancy must be excluded. Bronchoscopic washings can be sent for cultures for fungi and tuberculosis. For the pathologist, the more tissue that is provided, the more comfortable is the diagnosis of sarcoidosis. A needle aspirate may be adequate in an otherwise classic case of sarcoidosis, but may be insufficient in a patient in whom lymphoma or fungal infection is a likely alternative diagnosis. Because granulomas can be seen on the edge of a lymphoma, the presence of a few granulomas from a needle aspirate may not be sufficient to clarify the diagnosis. Mediastinoscopy provides a larger sample to confirm the presence or absence of lymphoma in the mediastinum. Alternatively, for most patients, evidence of extrathoracic disease (e.g., eye involvement) may further support the diagnosis of sarcoidosis.

For patients with negative pathology, positive supportive tests may increase the likelihood of the diagnosis of sarcoidosis. These tests include an elevated ACE level, which can also be elevated in other granulomatous diseases but not in malignancy. A positive PET scan can support the diagnosis if multiple organs are affected. A BAL is often performed during the bronchoscopy. An increase in the percentage of lymphocytes supports the diagnosis of sarcoidosis. The use of the lymphocyte markers CD4 and CD8 can be used to determine the CD4/CD8 ratio of these increased lymphocytes in the BAL fluid. A ratio of >3.5 is strongly supportive of sarcoidosis but is less sensitive than an increase in lymphocytes alone. Although in general, an increase in BAL lymphocytes is supportive of the diagnosis, other conditions must be considered.

Supportive findings, when combined with commonly associated but nondiagnostic clinical features of the disease, improve the diagnostic probability of sarcoidosis. These clinical features include uveitis, renal stones, hypercalcemia, seventh cranial nerve paralysis, or erythema nodosum. The presence of one or more of these features in a patient suspected of having sarcoidosis increases the probability of sarcoidosis.

The *Kviem-Siltzbach procedure* is a specific diagnostic test for sarcoidosis. An intradermal injection of specially prepared tissue derived from the spleen of a known sarcoidosis patient is biopsied 4–6 weeks after injection. If noncaseating granulomas are seen, this is highly specific for the diagnosis of sarcoidosis. Unfortunately, there is no commercially available Kviem-Siltzbach reagent, and some locally prepared batches have lower specificity. Thus, this test is of historic interest and is rarely used in current clinical practice.

Because the diagnosis of sarcoidosis can never be certain, over time other features may arise that lead to an alternative diagnosis. Conversely, evidence for new organ involvement may eventually confirm the diagnosis of sarcoidosis.

PROGNOSIS

The risk of death or loss of organ function remains low in sarcoidosis. Poor outcomes usually occur in patients who present with advanced disease in whom treatment seems to have little impact. In these cases, irreversible fibrotic changes have frequently occurred. Over the past 20 years, the reported mortality from sarcoidosis has increased in the United States and England. Whether this is due to heightened awareness of the chronic nature of this disease or to other factors such as more widespread immunosuppressive therapy usage remains unclear.

For the majority of patients, initial presentation occurs during the granulomatous phase of the disease as depicted in Fig. 360-1. It is clear that many patients resolve their disease within 2–5 years. These patients are felt to have acute, self-limiting sarcoidosis. However, there is a form of the disease that does not resolve within the first 2–5 years. These chronic patients can be identified at presentation by certain risk factors at presentation such as fibrosis on chest roentgenogram, presence of lupus pernio, bone cysts, cardiac or neurologic disease (except isolated seventh nerve paralysis), and presence of renal calculi due to hypercalciuria. In several studies, patients who required glucocorticoids for any manifestation of their disease in the first 6 months of presentation had a >50% chance of having chronic disease. In contrast, <10% of patients who require no systemic therapy in the first 6 months required chronic therapy.

TREATMENT

TREATMENT

Sarcoidosis

Indications for therapy should be based on symptoms or presence of organ- or life-threatening disease, including disease involving the eye, heart, or nervous system. The patient with asymptomatic elevated liver function tests or an abnormal chest roentgenogram probably does not benefit from treatment. However, these patients should be monitored for evidence of progressive, symptomatic disease.

One approach to therapy is summarized in **Figs. 360-9 and 360-10**. We have divided the approach into treating acute versus chronic disease. For acute disease, no therapy remains a viable option for patients with no or mild symptoms. For symptoms confined to only one organ, topical therapy is preferable. For multi-organ disease or disease too extensive for topical therapy, an approach to systemic therapy is outlined. Glucocorticoids remain the drugs of choice for this disease. However, the decision to continue to treat with glucocorticoids or to add steroid-sparing agents depends on the tolerability, duration, and dosage of glucocorticoids. **Table 360-2** summarizes the dosage and monitoring of several commonly used drugs. According to the available trials, evidence-based recommendations are made. Most of these recommendations are for pulmonary disease because most of the trials were performed only in pulmonary disease. Treatment recommendations for extrapulmonary disease are usually similar with a few modifications. For example, the dosage of glucocorticoids is usually higher for neurosarcoidosis and lower for cutaneous disease. There was some suggestion that higher doses would be beneficial for cardiac sarcoidosis, but one study found that initial prednisone doses >40 mg/d were associated with a worse outcome because of toxicity.

Systemic therapies for sarcoidosis are usually immunosuppressive including glucocorticoids, cytotoxics, or biologics. Although most patients receive glucocorticoids as their initial systemic therapy, toxicity associated with prolonged therapy often leads to steroid-sparing alternatives. The antimalarial drugs such as hydroxychloroquine are more effective for skin than pulmonary disease. Minocycline may also be useful for cutaneous sarcoidosis. For pulmonary and other extrapulmonary disease, cytotoxic agents that include methotrexate, azathioprine, leflunomide, mycophenolate, and cyclophosphamide are often used. The most widely studied cytotoxic agent has been methotrexate. This agent works in approximately two-thirds of sarcoidosis patients, regardless of the disease manifestation. In one retrospective study comparing methotrexate to azathioprine, both drugs were equally effective. However, methotrexate was associated with significantly less toxicity. As noted in Table 360-2, specific guidelines for monitoring therapy have been recommended. Cytokine modulators such as thalidomide and pentoxifylline have also been used in a limited number of cases.

The biologic anti-TNF agents have recently been studied in sarcoidosis, with prospective randomized trials completed for etanercept, golibmumab, and infliximab. Etanercept has a limited role as a steroid-sparing agent. Golibmumab was not significantly different than placebo in treating chronic pulmonary disease. However, this may have been due to the relatively low dose of golibmumab studied. Infliximab significantly improved lung function when administered to glucocorticoid and cytotoxic pretreated patients with chronic disease. The difference in response between etanercept and infliximab is similar to that observed in Crohn's disease, where infliximab is effective and etanercept is not. However, there is a higher risk for reactivation of tuberculosis with infliximab compared to etanercept. The differential response rate could be explained by differences in mechanism of action because etanercept is a TNF receptor antagonist and infliximab is a monoclonal antibody against TNF. In contrast to etanercept, infliximab also binds to TNF on the surface of some cells that release TNF, which leads to cell lysis. This effect has been documented in Crohn's disease. Adalimumab is a humanized monoclonal anti-TNF antibody that also appears effective for sarcoidosis when dosed at higher strengths, as recommended for the treatment of Crohn's disease. The role of the newer therapeutic agents for sarcoidosis is still evolving. However, these targeted therapies confirm that TNF may be an important target, especially in the treatment of chronic disease. However, these agents are not a panacea, because sarcoidosis-like disease has occurred in patients treated with anti-TNF agents for nonsarcoidosis indications.

TABLE 360-2

Commonly Used Drugs to Treat Sarcoidosis

Drug	Initial Dose	Maintenance Dose	Monitoring	Toxicity	Support Therapy[a]	Support Monitoring[a]
Prednisone	20–40 mg qd	Taper to 5–10 mg	Glucose, blood pressure, bone density	Diabetes, osteoporosis	A: Acute pulmonary D: Extrapulmonary	
Hydroxychloroquine	200–400 mg qd	400 mg qd	Eye examination q6–12 mo	Ocular	B: Some forms of disease	D: Routine eye examination
Methotrexate	10 mg qwk	2.5–15 mg qwk	CBC, renal, hepatic q2mo	Hematologic, nausea, hepatic, pulmonary	B: Steroid sparing C: Some forms chronic disease	D: Routine hematologic, renal, and hepatic monitoring
Azathioprine	50–150 mg qd	50–200 mg qd	CBC, renal q2mo	Hematologic, nausea	C: Some forms chronic disease	D: Routine hematologic monitoring
Infliximab	3–5 mg/kg q2wk for 2 doses	3–10 mg/kg q4–8 wk	Initial PPD	Infections, allergic reaction, carcinogen	A: Chronic pulmonary disease	B: Caution in patients with latent tuberculosis or advanced congestive heart failure

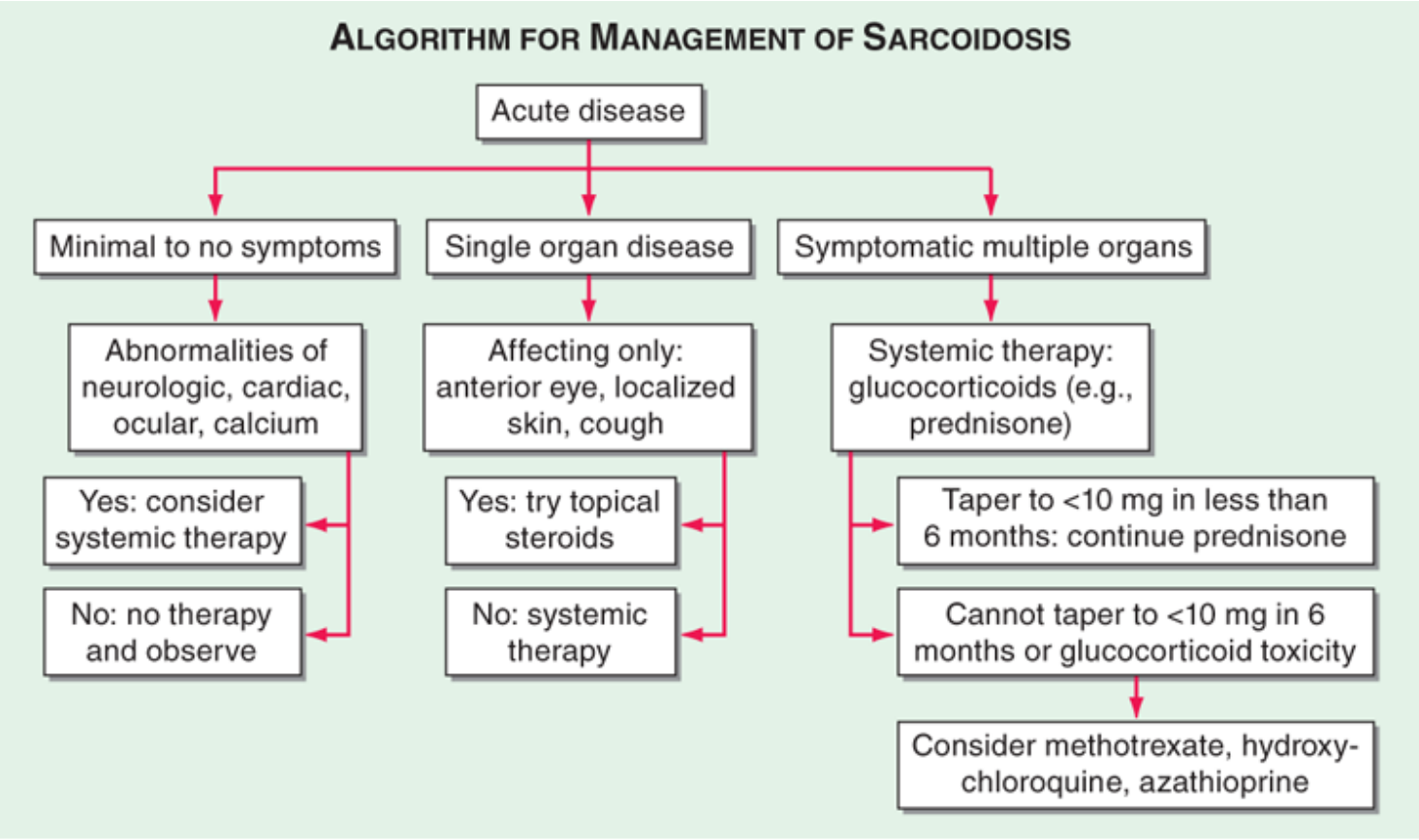
[a]Grade A: supported by at least two double-blind randomized control trials; grade B: supported by prospective cohort studies; grade C: supported primarily by two or more retrospective studies; grade D: only one retrospective study or based on experience in other diseases.

Abbreviations: CBC, complete blood count; PPD, purified protein derivative test for tuberculosis.

Source: Adapted from RP Baughman, O Selroos: Evidence-based approach to treatment of sarcoidosis, in PG Gibson et al (eds): *Evidence-Based Respiratory Medicine*. Oxford, BMJ Books Blackwell, 2005, pp 491–508.

FIGURE 360-9

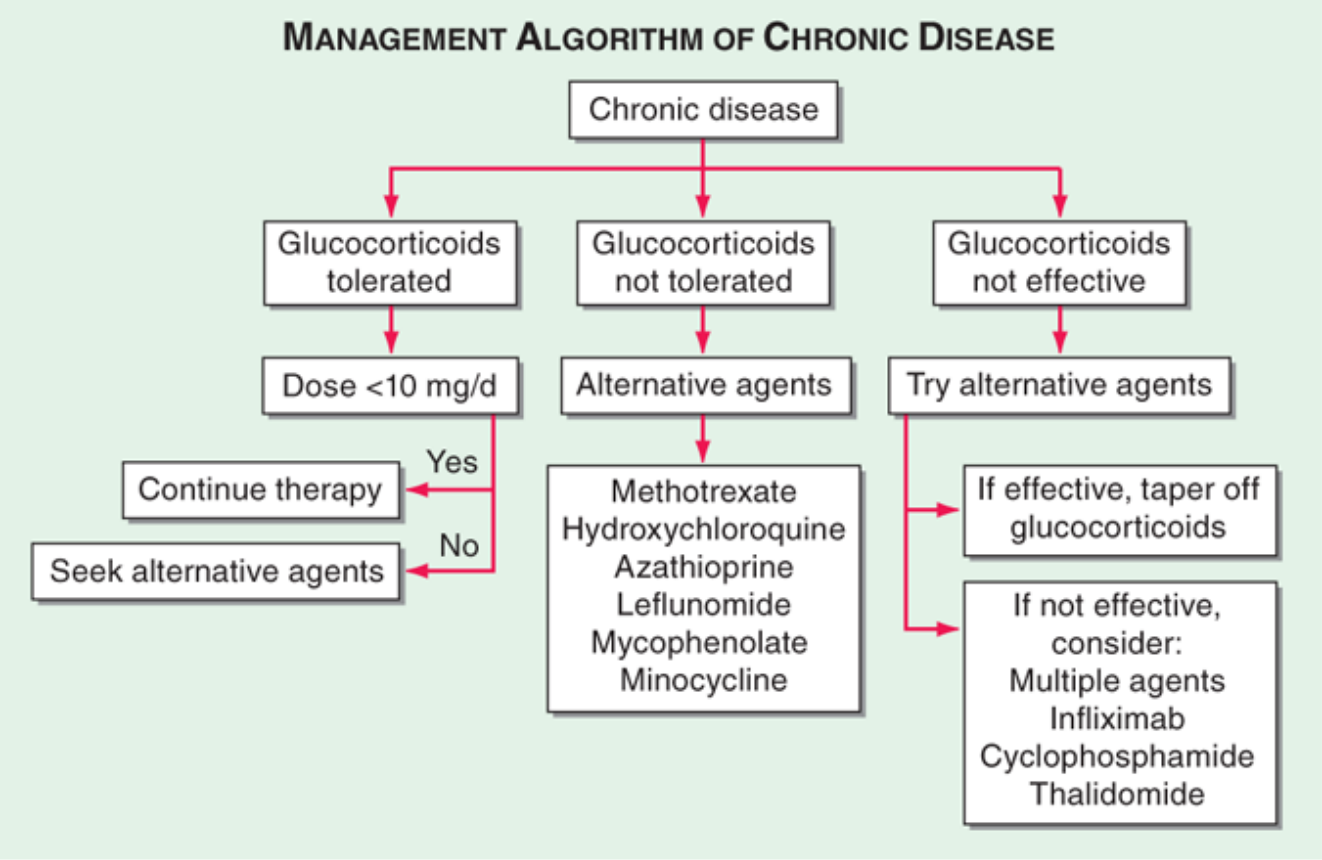
The management of acute sarcoidosis is based on level of symptoms and extent of organ involvement. In patients with mild symptoms, no therapy may be needed unless specified manifestations are noted.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 360-10

Approach to chronic disease is based on whether glucocorticoid therapy is tolerated or not.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Baughman RP, Selroos O: Evidence-based approach to treatment of sarcoidosis, in *Evidence-Based Respiratory Medicine*, PG Gibson et al (eds). Oxford, BMJ Books Blackwell, 2005, pp 491–508.

Baughman RP et al: Established and experimental medical therapy of pulmonary sarcoidosis. Eur Respir J 41:1424, 2013. [PubMed: 23397302]

Iannuzzi MC et al: Sarcoidosis. N Engl J Med 357:2153, 2007. [PubMed: 18032765]

Judson MA et al: Safety and efficacy of ustekinumab or golimumab in patients with chronic sarcoidosis. Eur Respir J 44:1296, 2014. [PubMed: 25034562]

Newman LS et al: A case control etiologic study of sarcoidosis: Environmental and occupational risk factors. Am J Respir Crit Care Med 170:1324, 2004. [PubMed: 15347561]

Valeyre D et al: Sarcoidosis. Lancet 383:1155, 2013. [PubMed: 24090799]

Chapter 361: IgG4-Related Disease

John H. Stone

INTRODUCTION

IgG4-related disease (IgG4-RD) is a fibroinflammatory condition characterized by a tendency to form tumefactive lesions. The clinical manifestations of this disease, however, are protean, as IgG4-RD can affect virtually any organ system. Commonly affected organs are the biliary tree, major salivary glands, periorbital tissues, kidneys, lungs, lymph nodes, and retroperitoneum. In addition, IgG4-RD involvement of the meninges, aorta, prostate, thyroid, pericardium, skin, and other organs is well described. The disease is believed to affect the brain parenchyma, the joints, the bone marrow, and the bowel mucosa only rarely.

The clinical features of IgG4-RD are numerous, but the pathologic findings are consistent across all affected organs. These findings include a lymphoplasmacytic infiltrate with a high percentage of IgG4-positive plasma cells; a characteristic pattern of fibrosis termed “storiform” (from the Latin *storea*, for “woven mat”); a tendency to target blood vessels, particularly veins, through an obliterative process (“obliterative phlebitis”); and a mild to moderate tissue eosinophilia.

IgG4-RD encompasses a number of conditions previously regarded as separate, organ-specific entities. A condition once known as “lymphoplasmacytic sclerosing pancreatitis” became the paradigm of IgG4-RD in 2000, when Japanese investigators recognized that these patients had elevated serum concentrations of IgG4. This form of sclerosing pancreatitis is now termed type 1 (IgG4-related) autoimmune pancreatitis (AIP). By 2003, extrapancreatic disease manifestations had been identified in patients with type 1 AIP, and descriptions of IgG4-RD other organs followed. *Mikulicz’s disease*, once considered to be a subset of Sjögren’s syndrome that affected the lacrimal, parotid, and submandibular glands, is now known to be one of the most common presentations of IgG4-RD. Similarly, the steroid-responsive subset of primary sclerosing cholangitis is explained by the fact that such patients actually have a separate disease, that is, IgG4-related sclerosing cholangitis. In this manner, the understanding of IgG4-RD has extended to include nearly every specialty of medicine.

CLINICAL FEATURES

The major organ lesions are summarized in Table 361-1. IgG4-RD usually presents subacutely, and even in the setting of multi-organ disease most patients do not have fevers or dramatic elevations of C-reactive protein levels. Some patients, however, experience substantial weight loss over periods of months. Clinically

apparent disease can evolve over months, years, or even decades before the manifestations within a given organ becomes sufficiently severe to bring the patient to medical attention. Some patients have disease that is marked by the appearance and then resolution or temporary improvement in symptoms within a particular organ. Other patients accumulate new organ involvement as their disease persists in previously affected organs. Many patients with IgG4-RD are misdiagnosed as having other conditions, particularly malignancies, or their findings are attributed initially to nonspecific inflammation. The disorder is often identified incidentally through radiologic findings or unexpectedly in pathology specimens.

TABLE 361-1

Organ Manifestations of IgG4-Related Disease

Organ	Major Clinical Features
Orbits and periorbital tissues	Painless eyelid or periocular tissue swelling; orbital pseudotumor; dacryoadenitis; dacryocystitis; orbital myositis; and mass lesions extending into the pterygopalatine fossa and infiltrating along the trigeminal nerve
Ears, nose, and sinuses	Allergic phenomena (nasal polyps, asthma, allergic rhinitis, peripheral eosinophilia); nasal obstruction, rhinorrhea, anosmia, chronic sinusitis; occasional bone-destructive lesions
Salivary glands	Submandibular and/or parotid gland enlargement (isolated bilateral submandibular gland involvement more common); minor salivary glands sometimes involved
Meninges	Headache, radiculopathy, cranial nerve palsies, or other symptoms resulting from spinal cord compression; tendency to form mass lesions; magnetic resonance imaging shows marked thickening and enhancement of dura
Hypothalamus and pituitary	Clinical syndromes resulting from involvement of the hypothalamus and pituitary, e.g., anterior pituitary hormone deficiency, central diabetes insipidus, or both; imaging reveals thickened pituitary stalk or mass formation on the stalk, swelling of the pituitary gland, or mass formation within the pituitary
Lymph nodes	Generalized lymphadenopathy or localized disease adjacent to a specific affected organ; the lymph nodes involved are generally 1–2 cm in diameter and nontender
Thyroid gland	Riedel's thyroiditis; fibrosing variant of Hashimoto's thyroiditis
Lungs	Asymptomatic finding on lung imaging; cough, hemoptysis, dyspnea, pleural effusion, or chest discomfort; associated with parenchymal lung involvement, pleural disease, or both; four main clinical lung syndromes: inflammatory pseudotumor, paravertebral mass often extending over several vertebrae, central airway disease, localized or diffuse interstitial pneumonia; pleural lesions have severe, nodular thickening of the visceral or parietal pleura with diffuse sclerosing inflammation, sometimes associated with pleural effusion

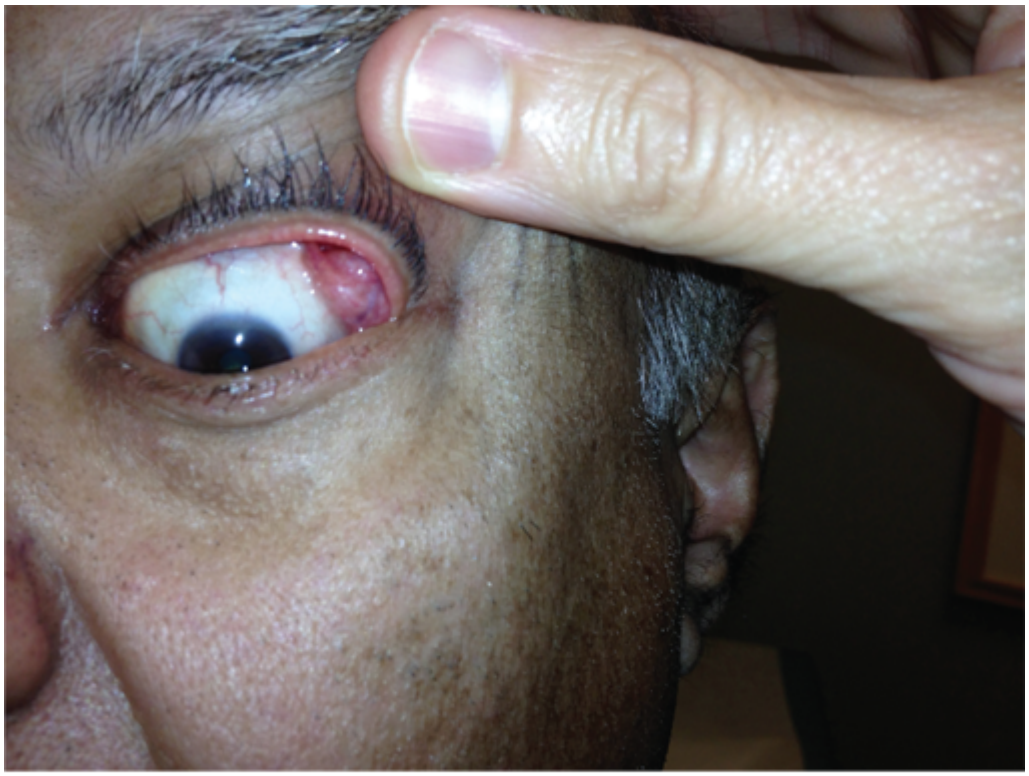
Organ	Major Clinical Features
Aorta	Asymptomatic finding on radiologic studies; surprise finding at elective aortic surgery; aortic dissection; clinicopathologic syndromes described include lymphoplasmacytic aortitis of thoracic or abdominal aorta, aortic dissection, periaortitis and periarteritis, and inflammatory abdominal aneurysm
Retroperitoneum	Backache, lower abdominal pain, lower extremity edema, hydronephrosis from ureteral involvement, asymptomatic finding on radiologic studies. Classic radiologic appearance is peri-aortic inflammation extending caudally to involve the iliac vessels.
Kidneys	Tubulointerstitial nephritis; membranous glomerulonephritis in a small minority; asymptomatic tumoral lesions, typically multiple and bilateral, are sometimes detected on radiologic studies; renal involvement strongly associated with hypocomplementemia
Pancreas	Type 1 autoimmune pancreatitis, presenting as mild abdominal pain; weight loss; acute, obstructive jaundice, mimicking adenocarcinoma of the pancreas (including a pancreatic mass); between 20 and 50% of patients present with acute glucose intolerance; imaging shows diffuse (termed “sausage-shaped pancreas”) or segmental pancreatic enlargement, with loss of normal lobularity; a mass often raises the suspicion of malignancy
Biliary tree and liver	Obstructive jaundice associated with autoimmunity in most cases; weight loss; steatorrhea; abdominal pain; and new-onset diabetes mellitus; mimicker of primary sclerosing cholangitis and cholangiocarcinoma
Other organs involved	Gallbladder, liver (mass), breast (pseudotumor), prostate (prostatism), pericardium (constrictive pericarditis), mesentery (sclerosing mesenteritis), mediastinum (fibrosing mediastinitis), skin (erythematous or flesh-colored papules), peripheral nerve (perineural inflammation)

Multiorgan disease may be evident at diagnosis but can also evolve over months to years. Some patients have disease confined to a single organ for many years. Others have either known or subclinical organ involvement at the same time as the major clinical feature. Patients with type 1 AIP may have their major disease focus in the pancreas; however, thorough evaluations by history, physical examination, blood tests, and cross-sectional imaging may demonstrate lacrimal gland enlargement, sialoadenitis, lymphadenopathy, a variety of pulmonary findings, tubulointerstitial nephritis, hepatobiliary disease, aortitis, retroperitoneal fibrosis, or other organ involvement. Spontaneous improvement, sometimes leading to clinical resolution of certain organ system manifestations, is reported in a small percentage of patients.

Two common characteristics of IgG4-RD are allergic disease and the tendency to form tumefactive lesions that mimic malignancies **(Fig. 361-1)**. Many IgG4-RD patients have allergic features such as atopy, eczema, asthma, nasal polyps, sinusitis, and modest peripheral eosinophilia. IgG4-RD also appears to account for a significant proportion of tumorous swellings—pseudotumors—in many organ systems. Some patients undergo major surgeries (e.g., Whipple procedures or thyroidectomy) for the purpose of resecting malignancies before the correct diagnosis is identified. Frequent sites of pseudotumors are the major salivary glands, lacrimal glands, lungs, and kidneys; however, nearly all organs have been affected with this manifestation.

FIGURE 361-1

A major clinical feature of IgG4-related disease is its tendency to form tumefactive lesions. Shown here are mass lesions of the lacrimal glands (***A***) and the submandibular glands (***B***).



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IgG4-RD often causes major morbidity and can lead to organ failure; however, its general pattern is to cause damage in a subacute manner. Destructive bone lesions in the sinuses, head, and middle ear spaces that mimic granulomatous polyangiitis (formerly Wegener's granulomatosis) occur rarely in IgG4-RD, but less aggressive lesions are the rule in most organs. In regions such as the retroperitoneum, substantial fibrosis often occurs before the diagnosis is established, leading to ureteral entrapment, hydronephrosis, postobstructive uropathy, and renal atrophy. The chronic pain often associated with IgG4-related retroperitoneal fibrosis probably, results from the encasement of peripheral nerves by the inflammatory process. Undiagnosed or undertreated IgG4-related cholangitis can lead to hepatic failure within months. Similarly, IgG4-related aortitis, believed to be the cause of a substantial minority of inflammatory aortitis cases, can cause aneurysms and dissections. Substantial renal dysfunction and even renal failure can ensue from IgG4-related tubulointerstitial nephritis, and renal atrophy is a frequent sequel to this disease complication. IgG4-related membranous glomerulonephropathy, a less common renal manifestation than tubulointerstitial nephritis, must be distinguished from idiopathic membranous glomerulonephropathy.

SEROLOGIC FINDINGS

The majority of patients with IgG4-RD have elevated serum IgG4 concentrations; however, the range of elevation varies widely. Serum concentrations of IgG4 as high as 30 or 40 times the upper limit of normal sometimes occur, usually in patients with disease that affects multiple organ systems simultaneously. Approximately 30% of patients have normal serum IgG4 concentrations despite classic histopathologic and immunohistochemical findings. Such patients tend to have disease that affects fewer organs. Patients with

IgG4-related retroperitoneal fibrosis have a high likelihood of normal serum IgG4 concentrations, perhaps because the process has advanced to a fibrotic stage by the time the diagnosis is considered.

The correlation between serum IgG4 concentrations and disease activity and the need for treatment is imperfect. Serum IgG4 concentrations typically decline swiftly with the institution of therapy but often do not normalize completely. Patients can achieve clinical remissions yet have persistently elevated serum IgG4 concentrations. Rapidly rising serum IgG4 concentrations may identify patients at greatest risk for clinical flares. Monitoring of serial IgG4 concentrations identifies early relapse in some patients; however, the temporal relationship between modest IgG4 elevations and the need for clinical treatment is poor. Clinical relapses occur in some patients despite persistently normal IgG4 concentrations.

IgG4 concentrations in serum are usually measured by nephelometry assays. These assays can lead to reports of spuriously low IgG4 values because of the prozone effect. This effect can be corrected by dilution of the serum sample in the laboratory. The prozone effect should be considered when the results of serologic testing for IgG4 concentrations are at odds with clinical features that strongly suggest IgG4-RD.

EPIDEMIOLOGY

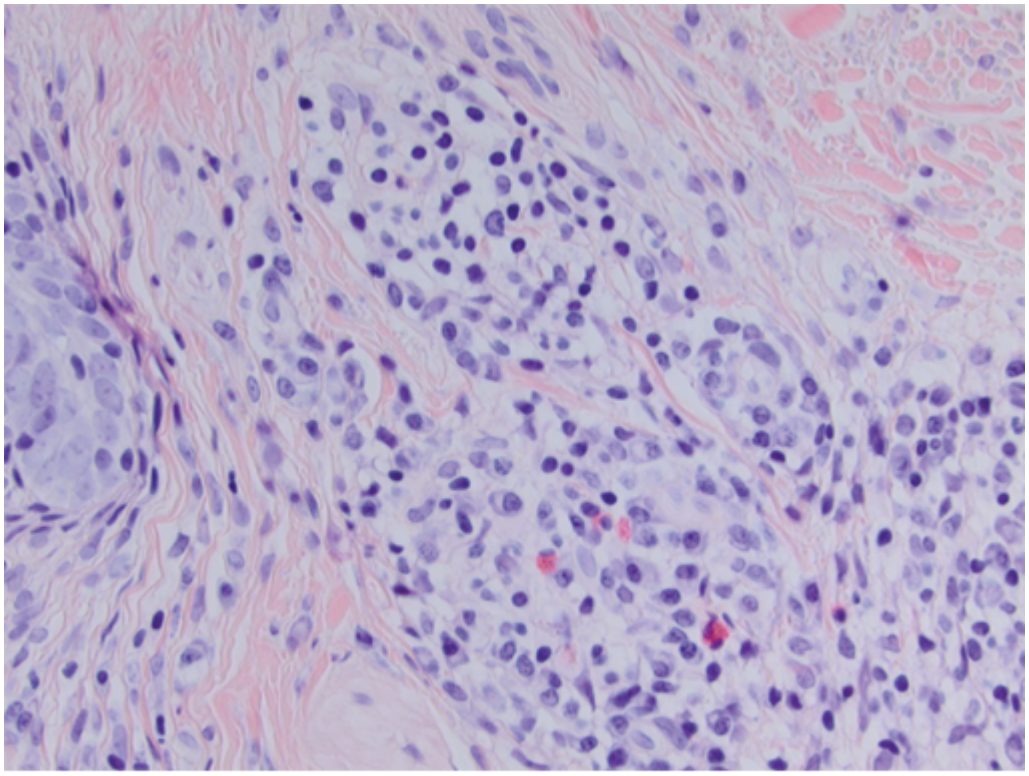
The typical patient with IgG4-RD is a middle-aged to elderly man. This epidemiology stands in stark contrast to that of many classic autoimmune conditions, which tend to affect young women. Studies of AIP patients in Japan indicate that the male-to-female ratio in that disease subset is on the order of 3:1. Even more striking, male predominance has been reported in IgG4-related tubulointerstitial nephritis and IgG4-related retroperitoneal fibrosis. Among IgG4-RD manifestations that involve organs of the head and neck, however, the sex ratio may be closer to 1:1.

PATHOLOGY

The key histopathology characteristics of IgG4-RD are a dense lymphoplasmacytic infiltrate **(Fig. 361-2)** that is organized in a storiform pattern, obliterative phlebitis, and a mild to moderate eosinophilic infiltrate. Lymphoid follicles and germinal centers are frequently observed. The infiltrate tends to aggregate around ductal structures when it affects glands such as the lacrimal, submandibular, and parotid glands or the pancreas. The inflammatory lesion often aggregates into tumefactive masses that destroy the involved tissue.

FIGURE 361-2

Hallmark histopathology characteristics of IgG4-related disease (IgG4-RD) are a dense lymphoplasmacytic infiltrate and a mild to moderate eosinophilic infiltrate. The cellular inflammation is often encased in a distinctive type of fibrosis termed “storiform,” which often has a basket weave pattern. Abundant fibroblasts and strands of fibrosis accompany the lymphoplasmacytic infiltrate and eosinophils in this figure. This biopsy was taken from a nodular lesion on the cheek; however, the findings are identical to the pathology found in the pancreas, kidneys, lungs, salivary glands, and other organs affected by IgG4-RD.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Obliterative arteritis is observed in some organs, particularly the lung; however, venous involvement is more common (and is indeed a hallmark of IgG4-RD). Several histopathology features are uncommon in IgG4-RD and, when detected, mitigate against the diagnosis of IgG4-RD. These include intense neutrophilic infiltration, leukocytoclasis, granulomatous inflammation, multinucleated giant cells, and fibrinoid necrosis.

The inflammatory infiltrate is composed of an admixture of B and T lymphocytes. B cells are typically organized in germinal centers. Plasma cells staining for CD19, CD138, and IgG4 appear to radiate from the germinal centers. In contrast, the T cells, usually CD4+, are distributed more diffusely throughout the lesion and generally represent the most abundant cell type. Fibroblasts, histiocytes, and eosinophils can all be observed in moderate numbers. Some biopsy samples are particularly enriched with eosinophils. In other samples, particularly from long-standing cases, fibrosis predominates.

The histologic appearance of IgG4-RD, although highly characteristic, requires immunohistochemical confirmation of the diagnosis with IgG4 immunostaining. IgG4-positive plasma cells predominate within the lesion, but plasma cells containing immunoglobulins from each subclass can be found. The number of IgG4-positive plasma cells can be quantified by either counting the number of cells per high-power field (HPF) or by calculating the ratio of IgG4- to IgG-bearing plasma cells. Tissue fibrosis predominates in the latter phases of organ involvement, and in this relatively acellular phase of inflammation, both the IgG4:total IgG ratio and the pattern of tissue fibrosis are more important than the number of IgG4-positive cells per HPF in establishing the diagnosis.

PATHOPHYSIOLOGY

Despite the emphasis of IgG4 in the name of this disease, the IgG4 molecule is not believed to play a direct role in the pathophysiology of disease within most organs. The IgG4 molecule can undergo Fab exchange, a phenomenon in which the two halves of the molecule dissociate from each other and reassociate with dissimilar hemi-molecules originating from other IgG4 molecules. Partly as a result of this Fab exchange, IgG4 antibodies do not bind antigen tightly. Moreover, the molecules have low affinities for Fc receptors and C1q and are regarded generally as noninflammatory immunoglobulins. The low affinities for Fc receptors and C1q impair the ability of IgG4 antibodies to induce phagocyte activation, antibody-dependent cellular cytotoxicity, and complement-mediated damage. It is possible that the role of IgG4 in this disease is actually as a counterregulatory mechanism rather than part of the primary inflammatory process.

Next-generation sequencing studies of CD4+ effector T cells have demonstrated a unique CD4+ cytotoxic T cell. This cell, also found in abundance at tissue sites of disease, makes interferon gamma, T-cell growth factor-beta, and interleukin-1, all of which may contribute to the storiform fibrosis found in this condition. The cells also elaborate perforin, granzyme A and B, and granulysin, products capable of inducing cytotoxicity. The pronounced oligoclonal expansion of this CD4+ cytotoxic T cell at tissue sites suggests that this cell is a major disease driver.

Oligoclonal expansions of plasmablasts are also present within the blood of patients with IgG4-RD. Continuous antigen presentation by B cells and plasmablasts may support this cell, which in turn produces profibrotic cytokines and other molecules, thereby directly mediating tissue injury.

TREATMENT

Not every disease manifestation of IgG4-RD requires immediate treatment because the disease takes an indolent form in many patients. IgG4-related lymphadenopathy, for example, can be asymptomatic for years, without evolution to other disease manifestations. Thus, watchful waiting is prudent in some cases. Vital organ involvement must be treated aggressively, however, because IgG4-RD can lead to serious organ dysfunction and failure. Aggressive disease can lead quickly to end-stage liver disease, permanent impairment of pancreatic function, renal atrophy, aortic dissection or aneurysms, and destructive lesions in the sinuses and nasopharynx.

Glucocorticoids are the first line of therapy. Treatment regimens, extrapolated from experience with the management of type 1 AIP, generally begin with 40 mg/d of prednisone, with tapering to discontinuation or maintenance doses of 5 mg/d within 2 or 3 months. Although the clinical response to glucocorticoids is usually swift and striking, prolonged steroid-free remissions are uncommon and the risk of steroid-induced morbidity in this middle-aged to elderly patient population is high, particularly those with baseline comorbidities and pancreatic involvement by IgG4-RD. Few data exist to support the utility of conventional steroid-sparing agents in this disease.

For patients with relapsing or glucocorticoid-resistant disease, B cell depletion with rituximab is an excellent second-line therapy. Rituximab treatment (two doses of 1 g IV, separated by approximately 15 days) leads to a swift decline in serum IgG4 concentrations, suggesting that rituximab achieves its effects in part by

preventing the repletion of short-lived plasma cells that produce IgG4. More important than its effects on IgG4 concentrations, however, may be the effect of B cell depletion on T cell function. Specific effects of rituximab on the CD4+ cytotoxic T cell described above have been documented in IgG4-RD. Rituximab may be an appropriate first-line therapy for some patients, particularly those at high risk for glucocorticoid toxicity and patients with immediately organ-threatening disease. The rapidly evolving understanding of the pathophysiology of IgG4-RD suggests several novel targeted approaches to treating the disease, some of which are in clinical trials. These novel targets generally involve interference with either B or T cell directly for the purpose of depletion or functional inhibition.

FURTHER READING

Khosroshahi A et al: International Consensus Guidance Statement on the Management and Treatment of IgG4-Related Disease. Arthritis Rheumatol 67:1688, 2015. [PubMed: 25809420]

Mahajan V et al: IgG4-related disease. Ann Rev Pathol 9:315, 2014.

Mattoo H et al: *De novo* oligoclonal expansions of circulating plasmablasts in active and relapsing IgG4-related disease. J Allergy Clin Immunol 134:679, 2014. [PubMed: 24815737]

Mattoo H et al: Clonal expansion of CD4+ cytotoxic T lymphocytes in IgG4-related disease. J Allergy Clin Immunol 138:825, 2016. [PubMed: 26971690]

Wallace ZS et al: IgG4-related disease: Baseline clinical and laboratory features in 125 patients with biopsy-proven disease. Arthritis Rheumatol 67:2466, 2015. [PubMed: 25988916]

Chapter 362: Familial Mediterranean Fever and Other Hereditary Autoinflammatory Diseases

Daniel L. Kastner

INTRODUCTION

Familial Mediterranean fever (FMF) is the prototype of a group of inherited diseases (Table 362-1) that are characterized by recurrent episodes of fever with serosal, synovial, or cutaneous inflammation and, in some individuals, the eventual development of systemic AA amyloidosis (Chap. 108). Because of the relative infrequency of high-titer autoantibodies or antigen-specific T cells, the term *autoinflammatory* has been proposed to describe these disorders, rather than autoimmune. The innate immune system, with its myeloid effector cells and germline receptors for pathogen-associated molecular patterns and endogenous danger signals, plays a predominant role in the pathogenesis of the autoinflammatory diseases. Although the hereditary recurrent fevers comprise a major category of the autoinflammatory diseases, other inherited disorders of inflammation in which recurrent fever plays a less prominent role are now also considered to be autoinflammatory.

TABLE 362-1

The Hereditary Recurrent Fever Syndromes

	FMF	TRAPS	HIDS	MWS	FCAS	NOMID
Ethnicity	Jewish, Arab, Turkish, Armenian, Italian	Any ethnic group	Predominantly Dutch, northern European	Any ethnic group	Any ethnic group	Any ethnic group
Inheritance	Recessive[a]	Dominant	Recessive	Dominant	Dominant	Most commonly de novo mutations; somatic mosaicism in a significant minority
Gene/chromosome	*MEFV*/16p13.3	*TNFRSF1A*/12p13	*MVK*/12q24	*NLRP3*/1q44	*NLRP3*/1q44	*NLRP3*/1q44
Protein	Pyrin	p55 TNF receptor	Mevalonate kinase	NLRP3 (cryopyrin)	NLRP3 (cryopyrin)	NLRP3 (cryopyrin)
Attack length	1–3 days	Often >7 days	3–7 days	1–2 days	Minutes–3 days	Continuous, with flares
Serosa	Pleurisy, peritonitis; asymptomatic pericardial effusions	Pleurisy, peritonitis, pericarditis	Abdominal pain, but seldom peritonitis; pleurisy, pericarditis uncommon	Abdominal pain; pleurisy, pericarditis rare	Rare	Rare
Skin	Erysipeloid erythema	Centrifugally migrating erythema	Diffuse maculopapular rash; oral ulcers	Diffuse urticaria-like rash	Cold-induced urticaria-like rash	Diffuse urticaria-like rash
Joints	Acute monoarthritis; chronic hip arthritis (rare)	Acute monoarthritis, arthralgia	Arthralgia, oligoarthritis	Arthralgia, large joint oligoarthritis	Polyarthralgia	Epiphyseal, patellar overgrowth, clubbing

	FMF	TRAPS	HIDS	MWS	FCAS	NOMID
Muscle	Exercise-induced myalgia common; protracted febrile myalgia rare	Migratory myalgia	Uncommon	Myalgia common	Sometimes myalgia	Sometimes myalgia
Eyes, ears	Uncommon	Periorbital edema, conjunctivitis, rarely uveitis	Uncommon	Conjunctivitis, episcleritis, optic disc edema; sensorineural hearing loss	Conjunctivitis	Conjunctivitis, uveitis, optic disc edema, blindness, sensorineural hearing loss
CNS	Aseptic meningitis rare	Headache	Headache	Headache	Headache	Aseptic meningitis, seizures
Amyloidosis	Most common in M694V homozygotes	~15% of cases, most often cysteine mutations, T50M	Uncommon	~25% of cases	Uncommon	Late complication
Treatment	Oral colchicine prophylaxis, IL-1 inhibitors for refractory cases	Glucocorticoids, etanercept, IL-1 inhibitors	NSAIDs for fever; IL-1 inhibitors	Anakinra, rilonacept, canakinumab	Anakinra, rilonacept, canakinumab	Anakinra

[a]A substantial percentage of patients with clinical FMF have only a single demonstrable *MEFV* mutation on DNA sequencing.

Abbreviations: CNS, central nervous system; FCAS, familial cold autoinflammatory syndrome; FMF, familial Mediterranean fever; HIDS, hyperimmunoglobulinemia D with periodic fever syndrome; IL, interleukin; MWS, Muckle-Wells syndrome; NOMID, neonatal-onset multisystem inflammatory disease; NSAIDs, nonsteroidal anti-inflammatory drugs; TNF, tumor necrosis factor; TRAPS, TNF receptor-associated periodic syndrome.

BACKGROUND AND PATHOPHYSIOLOGY

FMF was first recognized among Armenians, Arabs, Turks, and non-Ashkenazi (primarily North African and Iraqi) Jews. With the advent of genetic testing, FMF has been documented with increasing frequency among Ashkenazi Jews, Italians,

and other Mediterranean populations, and occasional cases have been confirmed even in the absence of known Mediterranean ancestry. FMF is generally regarded as recessively inherited, but there is an increasing awareness of clear-cut clinical cases with only a single demonstrable genetic mutation, and, for certain relatively rare FMF mutations, there is strong evidence for dominant inheritance. Particularly in countries where families are small, a positive family history can only be elicited in ~50% of cases. DNA testing demonstrates carrier frequencies as high as 1:3 among affected populations, suggesting a heterozygote advantage.

The FMF gene encodes a 781-amino acid, ~95 kDa protein denoted *pyrin* (or *marenostrin*) that is expressed in granulocytes, eosinophils, monocytes, dendritic cells, and synovial and peritoneal fibroblasts. The N-terminal 92 amino acids of pyrin define a motif, the PYRIN domain, that mediates homotypic protein-protein interactions and has been found in several other proteins, including cryopyrin (NLRP3), which is mutated in three other recurrent fever syndromes. Through the interaction of its PYRIN domain with an intermediary adaptor protein, pyrin nucleates the formation of a macromolecular *pyrin inflammasome* to activate caspase-1 (interleukin [IL] 1β-converting enzyme), and thereby IL-1β secretion. Certain bacterial toxins that block leukocyte cytoskeletal assembly by inactivating RhoA GTPase trigger pyrin inflammasome activation as a part of the normal host defense; in FMF patients the threshold for pyrin inflammasome activation is reduced.

ACUTE ATTACKS

Febrile episodes in FMF may begin even in early infancy; 90% of patients have had their first attack by age 20. Typical FMF episodes generally last 24–72 h, with arthritic attacks tending to last somewhat longer. In some patients, the episodes occur with great regularity, but more often, the frequency of attacks varies over time, ranging from as often as once every few days to remissions lasting several years. Attacks are often unpredictable, although some patients relate them to physical exertion, emotional stress, or menses; pregnancy may be associated with remission.

If measured, fever is nearly always present throughout FMF attacks. Severe hyperpyrexia and even febrile seizures may be seen in infants, and fever is sometimes the only manifestation of FMF in young children.

Over 90% of FMF patients experience abdominal attacks at some time. Episodes range in severity from dull, aching pain and distention with mild tenderness on direct palpation to severe generalized pain with absent bowel sounds, rigidity, rebound tenderness, and air-fluid levels on upright radiographs. Computed tomography (CT) scanning may demonstrate a small amount of fluid in the abdominal cavity. If such patients undergo exploratory laparotomy, a sterile, neutrophil-rich peritoneal exudate is present, sometimes with adhesions from previous episodes. Ascites is rare.

Pleural attacks are usually manifested by unilateral, sharp, stabbing chest pain. Radiographs may show atelectasis and sometimes an effusion. If performed, thoracentesis demonstrates an exudative fluid rich in neutrophils. After repeated attacks, pleural thickening may develop.

FMF arthritis is most frequent among individuals homozygous for the M694V mutation, which is especially common in the non-Ashkenazi Jewish population. Acute arthritis in FMF is usually monoarticular, affecting the knee, ankle, or hip, although other patterns can be seen. Large sterile effusions rich in neutrophils are frequent, without commensurate erythema or warmth. Even after repeated arthritic attacks, radiographic changes are rare. Before the advent of colchicine prophylaxis, chronic arthritis of the knee or hip was seen in ~5% of FMF patients with arthritis. Chronic sacroiliitis can occur in FMF irrespective of the HLA-B27 antigen, even in the face of colchicine therapy. In the United States, FMF patients are much more likely to have arthralgia than arthritis.

The most characteristic cutaneous manifestation of FMF is erysipelas-like erythema, a raised erythematous rash that most commonly occurs on the dorsum of the foot, ankle, or lower leg alone or in combination with abdominal pain, pleurisy, or arthritis. Biopsy demonstrates perivascular infiltrates of granulocytes and monocytes. This rash is seen most often in M694V homozygotes and is relatively rare in the United States.

Exercise-induced (nonfebrile) myalgia is common in FMF, and a small percentage of patients develop a protracted febrile myalgia that can last several weeks. Symptomatic pericardial disease is rare, although small pericardial effusions may be noted on echocardiography. Unilateral acute scrotal inflammation may occur in prepubertal boys. Aseptic meningitis has been reported in FMF, but the causal connection is controversial. Vasculitis, including Henoch-Schönlein purpura and polyarteritis nodosa **(Chap. 356)**, may be seen at increased frequency in FMF. The M694V FMF mutation has recently been shown to be a risk factor for Behçet's disease.

Laboratory features of FMF attacks are consistent with acute inflammation and include an elevated erythrocyte sedimentation rate, leukocytosis, thrombocytosis (in children), and elevations in C-reactive protein, fibrinogen, haptoglobin, and serum immunoglobulins. Transient albuminuria and hematuria may also be seen.

AMYLOIDOSIS

Before the advent of colchicine prophylaxis, systemic amyloidosis was a common complication of FMF. It is caused by deposition of a fragment of serum amyloid A, an acute-phase reactant, in the kidneys, adrenals, intestine, spleen, lung, and testes **(Chap. 108)**. Amyloidosis should be suspected in patients who have proteinuria between attacks; renal or rectal biopsy is used most often to establish the diagnosis. Risk factors include the M694V homozygous genotype, positive family history (independent of FMF mutational status), the *SAA1* genotype, male gender, noncompliance with colchicine therapy, and having grown up in the Middle East.

DIAGNOSIS

For typical cases, physicians experienced with FMF can often make the diagnosis on clinical grounds alone. Clinical criteria sets for FMF have been shown to have high sensitivity and specificity in parts of the world where the pretest probability of FMF is high. Genetic testing can provide a useful adjunct in ambiguous cases or for physicians not experienced in FMF. Most of the more severe disease-associated FMF mutations are in exon 10 of the gene. An updated list of mutations for FMF and other hereditary recurrent fevers can be found online at *http://fmf.igh.cnrs.fr/infevers/*.

Genetic testing has permitted a broadening of the clinical spectrum and geographic distribution of FMF and may be of prognostic value. Most studies indicate that M694V homozygotes have an earlier age of onset and a higher frequency of arthritis, rash, and amyloidosis. In contrast, the E148Q variant in exon 2 is quite common in certain Asian populations and is more likely to affect overall levels of inflammation than to cause clinical FMF. E148Q is sometimes found in *cis* with exon 10 mutations, which may complicate the interpretation of genetic test results. Only ~70% of patients with clinically typical FMF have two identifiable mutations in *trans*, consistent with the concept that FMF mutations are gain-of-function with regard to inflammasome activation, with a gene dosage effect. In those cases in which only a single mutation is identified, clinical judgment is very important, and sometimes a therapeutic trial of colchicine may help to confirm the diagnosis.

If a patient is seen during his or her first attack, the differential diagnosis may be broad, although delimited by the specific organ involvement. After several attacks the differential diagnosis may include the other hereditary recurrent fever syndromes (Table 362-1); the syndrome of periodic fever with aphthous ulcers, pharyngitis, and cervical adenopathy (PFAPA); systemic-onset juvenile rheumatoid arthritis or adult Still's disease; porphyria; hereditary angioedema; inflammatory bowel disease; and, in women, gynecologic disorders.

TREATMENT

TREATMENT

Familial Mediterranean Fever

The treatment of choice for FMF is daily oral colchicine, which decreases the frequency and intensity of attacks and prevents the development of amyloidosis in compliant patients. Intermittent dosing at the onset of attacks is not as effective as daily prophylaxis and is of unproven value in preventing amyloidosis. The usual adult dose of colchicine is 1.2–1.8 mg/d, which causes substantial reduction in symptoms in two-thirds of patients and some improvement in >90%. Children may require lower doses, although not proportionately to body weight.

Common side effects of colchicine include bloating, abdominal cramps, lactose intolerance, and diarrhea. They can be minimized by starting at a low dose and gradually advancing as tolerated, splitting the dose, use of simethicone for flatulence, and avoidance of dairy products. If taken by either parent at the time of conception, colchicine may cause a small increase in the risk of trisomy 21 (Down's syndrome). In elderly patients with renal insufficiency, colchicine can cause a myoneuropathy characterized by proximal muscle weakness and elevation of the creatine kinase. Cyclosporine inhibits hepatic excretion of colchicine by its effects on the multidrug resistance 1 (MDR1) transport system, sometimes leading to colchicine toxicity in patients who have undergone renal transplantation for amyloidosis. Intravenous colchicine should generally not be administered to patients already taking oral colchicine, because severe, sometimes fatal, toxicity can occur in this setting.

For FMF patients who do not respond to colchicine or cannot tolerate therapeutic doses, injectable IL-1 inhibitors may be used. Based on a randomized, placebo-controlled phase III trial, the monoclonal anti-IL-1β antibody canakinumab recently received the Food and Drug Administration (FDA) approval for this indication. In a small randomized placebo-controlled trial, weekly subcutaneous rilonacept, a recombinant interleukin 1 (IL-1) receptor fusion protein, significantly reduced the frequency of attacks. There is also substantial anecdotal experience with daily subcutaneous anakinra, a recombinant IL-1 receptor antagonist, in preventing the acute attacks of FMF and, in some cases, reducing established amyloid deposits. Bone marrow transplantation has been suggested for refractory FMF, but the risk-benefit ratio is currently regarded as unacceptable.

OTHER HEREDITARY RECURRENT FEVERS

TNF RECEPTOR-ASSOCIATED PERIODIC SYNDROME

TNF receptor–associated periodic syndrome (TRAPS) is caused by dominantly inherited mutations in the extracellular domains of the 55-kDa TNF receptor (TNFR1, p55). Although originally described in a large Irish family (and hence the name *familial Hibernian fever*), TRAPS has a broad ethnic distribution. TRAPS episodes often begin in childhood. The duration of attacks ranges from 1 to 2 days to as long as several weeks, and in severe cases symptoms may be nearly continuous. In addition to peritoneal, pleural, and synovial attacks similar to FMF, TRAPS patients frequently have ocular inflammation (most often conjunctivitis and/or periorbital edema), and a distinctive migratory myalgia with overlying painful erythema may be present. TRAPS patients generally respond better to glucocorticoids than to prophylactic colchicine. Untreated, ~15% develop amyloidosis. The diagnosis of TRAPS is based on the demonstration of a *TNFRSF1A* mutation in the presence of characteristic symptoms. Two particular variants, R92Q and P46L, are common in certain populations and may act more as functional polymorphisms than as disease-causing mutations. In contrast, pathogenic *TNFRSF1A* mutations, including a number of substitutions at highly conserved cysteine residues, are associated with intracellular TNFR1 misfolding, aggregation, and retention, with consequent ligand-independent kinase activation,

mitochondrial reactive oxygen species production, and proinflammatory cytokine release. Etanercept, a TNF inhibitor, ameliorates TRAPS attacks, but the long-term experience with this agent has been less favorable. IL-1 inhibition has been beneficial in a large percentage of the patients in whom it has been used, and canakinumab recently received FDA approval for the treatment of TRAPS. Monoclonal anti-TNF antibodies should be avoided, because they may exacerbate TRAPS attacks.

HYPERIMMUNOGLOBULINEMIA D WITH PERIODIC FEVER SYNDROME

Hyperimmunoglobulinemia D with periodic fever syndrome (HIDS) is a recessively inherited recurrent fever syndrome found primarily in individuals of northern European ancestry. It is caused by mutations in mevalonate kinase (*MVK*), encoding an enzyme involved in the synthesis of cholesterol and nonsterol isoprenoids, including geranylgeranyl pyrophosphate. The latter compound is essential for proper localization of RhoA GTPase to the cell membrane, and the mislocalization of RhoA leads to its inactivation and the consequent activation of the pyrin inflammasome. HIDS attacks usually begin in infancy and last 3–5 days. Clinically distinctive features include painful cervical adenopathy, a diffuse maculopapular rash sometimes affecting the palms and soles, and aphthous ulcers; pleurisy is rare, as is amyloidosis. Although originally defined by the persistent elevation of serum IgD, disease activity is not related to IgD levels, and some patients with FMF or TRAPS may have modestly increased serum IgD. Moreover, occasional patients with *MVK* mutations and recurrent fever have normal IgD levels, while all patients with mutations have markedly elevated urinary mevalonate levels during their attacks. For these reasons, some have proposed renaming this disorder *mevalonate kinase deficiency* (MKD). Canakinumab was recently FDA-approved for the treatment of HIDS/MKD.

THE CRYOPYRINOPATHIES OR CRYOPYRIN-ASSOCIATED PERIODIC SYNDROMES

Three hereditary febrile syndromes, familial cold autoinflammatory syndrome (FCAS), Muckle-Wells syndrome (MWS), and neonatal-onset multisystem inflammatory disease (NOMID), are all caused by mutations in *NLRP3* (formerly known as *CIAS1*), the gene encoding cryopyrin (or NLRP3), and represent a clinical spectrum of disease. FCAS patients develop chills, fever, headache, arthralgia, conjunctivitis, and an urticaria-like rash in response to generalized cold exposure. In MWS, an urticarial rash is noted, but it is not usually induced by cold; MWS patients also develop fevers, abdominal pain, limb pain, arthritis, conjunctivitis, and, over time, sensorineural hearing loss. NOMID is the most severe of the three disorders, with chronic aseptic meningitis, a characteristic arthropathy, and rash. Like the FMF protein, pyrin, cryopyrin has an N-terminal PYRIN domain, allowing the formation of an *NLRP3 inflammasome* that mediates caspase-1 activation and IL-1β release. Peripheral blood leukocytes from patients with FCAS, MWS, and NOMID release increased amounts of IL-1β upon in vitro stimulation, relative to healthy controls. Macrophages from cryopyrin-deficient mice exhibit decreased IL-1β production in response to certain gram-positive bacteria, bacterial RNA, and monosodium urate crystals. Patients with all three cryopyrinopathies or cryopyrin-associated periodic syndromes (CAPS) show a dramatic response to injections of IL-1 inhibitors. Canakinumab and rilonacept are FDA-approved for the treatment of FCAS and MWS, while anakinra is approved for the treatment of NOMID.

Approximately one-third of patients with clinical manifestations of NOMID do not have germline mutations in *NLRP3*, but they have been found to be mosaic for somatic *NLRP3* mutations. Such patients also respond dramatically to IL-1 inhibition. Similarly, somatic mosaicism in *NLRP3* has been found in Schnitzler's syndrome, which presents in middle age with recurrent fever, urticarial rash, elevated acute phase reactants, monoclonal IgM gammopathy, and abnormal bone remodeling. IL-1 inhibition is the treatment of choice for Schnitzler's syndrome.

OTHER INHERITED AUTOINFLAMMATORY DISEASES

There are a number of other Mendelian autoinflammatory diseases in which recurrent fevers are not a prominent clinical sign but that involve abnormalities of innate immunity. The syndrome of pyogenic arthritis with pyoderma gangrenosum and acne (PAPA) is a dominantly inherited disorder that presents with episodes of sterile pyogenic monoarthritis often induced by trauma, severe pyoderma gangrenosum, and severe cystic acne usually beginning in puberty. It is caused by mutations in *PSTPIP1*, which encodes a pyrin-binding protein, and the arthritic manifestations often respond to IL-1 inhibition. Patients with the recessively inherited deficiency of the IL-1 receptor antagonist (DIRA) present with a generalized pustular rash and multifocal sterile osteomyelitis and show dramatic clinical responses to anakinra, the recombinant form of the protein they lack. IL-36 is another member of the IL-1 family of cytokines that is regulated by an endogenous receptor antagonist. The recessively inherited deficiency of the IL-36 receptor antagonist (DITRA) presents with episodes of generalized pustular psoriasis and dramatic systemic inflammation. Dominantly inherited gain-of-function mutations in *NLRC4* lead to increased IL-1 and IL-18 production and potentially life-threatening recurrent macrophage activation syndrome.

Whereas the aforementioned disorders all involve mutations in IL-1-related molecules, other autoinflammatory diseases are caused by mutations in other components of innate immunity. *Blau's syndrome* is caused by mutations in *CARD15* (also known as *NOD2*), which regulates nuclear factor κB activation. Blau's syndrome is characterized by granulomatous dermatitis, uveitis, and arthritis; distinct *CARD15* variants predispose to Crohn's disease. Recessive mutations in one or more components of the proteasome lead to excessive interferon signaling and the syndrome of chronic atypical neutrophilic dermatosis with lipodystrophy and elevated temperature (CANDLE), a severe form of generalized panniculitis. De novo gain-of-function mutations in *TMEM173*, encoding the stimulator of interferon genes (STING), cause severe vasculopathy and pulmonary fibrosis. Recessive loss-of-function mutations in *CERCR1*, encoding adenosine deaminase 2 (ADA2), cause a vasculopathy that can manifest as livedoid rash, early-onset lacunar strokes, or polyarteritis nodosa. Mutations in the gene encoding the A20 ubiquitin-modifying enzyme cause a Behçet's-like monogenic illness ("HA20") while mutations in a different deubiquitinase (OTULIN) cause a form of panniculitis ("otulipenia").

Finally, it should be noted that a number of common, genetically complex disorders are now sometimes considered autoinflammatory, because of evidence that components of the innate immune system, such as the inflammasome, may play a role in the pathogenesis. Two prominent examples are gout and atherosclerosis.

GLOBAL CONSIDERATIONS

 All the disorders discussed in this chapter have been observed in multiple populations. However, as noted herein, FMF is most frequently observed in Mediterranean and Middle-Eastern populations and HIDS in Northern European populations, particularly the Dutch. A recessive founder mutation in *CECR1* is particularly common in the Georgian Jewish population and is associated with polyarteritis nodosa.

FURTHER READING

Canna SW et al: An activating *NLRC4* inflammasome mutation causes autoinflammation with recurrent macrophage activation syndrome. Nat Genet 46:1140, 2014. [PubMed: 25217959]

de Jesus AA, Goldbach-Mansky R: Newly recognized Mendelian disorders with rheumatic manifestations. Curr Opin Rheumatol 27:511, 2015. [PubMed: 26196376]

Park YH et al: Pyrin inflammasome activation and RhoA signaling in the autoinflammatory diseases FMF and HIDS. Nat Immunol 17:914, 2016. [PubMed: 27270401]

Stoffels M, Kastner DL: Old dogs, new tricks: Monogenic autoinflammatory disease unleashed. Annu Rev Genom Hum Genet 17:245, 2016.

Zhou Q et al: Early-onset stroke and vasculopathy associated with mutations in ADA2. N Engl J Med 370:911, 2014. [PubMed: 24552284]

Chapter 363: Approach to Articular and Musculoskeletal Disorders

John J. Cush

INTRODUCTION

Musculoskeletal complaints account for >315 million outpatient visits per year and >20% of all outpatient visits in the United States. The Centers for Disease Control and Prevention estimate that 54.4 million, or 1 in 5 adults) of the U.S. population has physician-diagnosed arthritis. While many patients will have self-limited conditions requiring minimal evaluation, reassurance, and symptomatic therapy, specific musculoskeletal presentations or their persistence may herald a more serious condition that requires further evaluation or laboratory testing to establish a diagnosis. The goal of the musculoskeletal evaluation is to formulate a differential diagnosis that leads to an accurate diagnosis and timely therapy, while avoiding excessive diagnostic testing and unnecessary treatment **(Table 363-1)**. There are several urgent conditions that must be diagnosed promptly to avoid significant morbid or mortal sequelae. These "red flag" diagnoses include septic arthritis, acute crystal-induced arthritis (e.g., gout), and fracture. Each may be suspected by its acute onset and monarticular or focal musculoskeletal pain.

TABLE 363-1

Evaluation of Patients with Musculoskeletal Complaints

Goals
Accurate diagnosis
Timely provision of therapy
Avoidance of unnecessary diagnostic testing
Identification of acute, focal/monarticular "red flag" conditions
Approach
Determine the chronology (acute vs chronic)
Determine the nature of the pathologic process (inflammatory vs noninflammatory)
Determine the extent of involvement (monarticular, polyarticular, focal, widespread)
Anatomic localization of complaint (articular vs nonarticular)
Consider the most common disorders first
Consider the need for diagnostic testing
Formulate a differential diagnosis

The majority of individuals with musculoskeletal complaints can be diagnosed with a thorough history and a comprehensive physical and musculoskeletal examination. The initial encounter should determine whether the musculoskeletal complaint signals a red flag condition (septic arthritis, gout, or fracture) or not. The evaluation should ascertain if the complaint is (1) *articular* or *nonarticular* in origin, (2) *inflammatory* or *noninflammatory* in nature, (3) *acute* or *chronic* in duration, and (4) *localized (monarticular)* or *widespread* (*polyarticular*) in distribution.

With this approach, the musculoskeletal presentation can be characterized (e.g., acute inflammatory monarthritis or a chronic noninflammatory, nonarticular widespread pain) to narrow the diagnostic possibilities. However, some patients will not fit immediately into an established diagnostic category. Many musculoskeletal disorders resemble each other at the outset, and some may take weeks or months (but not years) to evolve into a recognizable diagnostic entity. This consideration should temper the desire to establish a definitive diagnosis at the first encounter.

ARTICULAR VERSUS NONARTICULAR

The musculoskeletal evaluation must discriminate the anatomic origin(s) of the patient's complaint. For example, ankle pain can result from a variety of pathologic conditions involving disparate anatomic structures, including gouty arthritis, calcaneal fracture, Achilles tendinitis, plantar fasciitis, cellulitis, and peripheral or entrapment neuropathy. Distinguishing between articular and nonarticular conditions requires a careful and detailed examination. Articular structures include the synovium, synovial fluid, articular cartilage, intraarticular ligaments, joint capsule, and juxtaarticular bone. Nonarticular (or periarticular) structures, such as supportive extraarticular ligaments, tendons, bursae, muscle, fascia, bone, nerve, and overlying skin, may be involved in the pathologic process. Although musculoskeletal complaints are often ascribed to the joints, nonarticular disorders more frequently underlie such complaints. Distinguishing between these potential sources of pain may be challenging to the unskilled examiner. Articular disorders may be characterized by deep or diffuse pain, pain or limited range of motion on active and passive movement, and swelling (caused by synovial proliferation, effusion, or bony enlargement), crepitation, instability, "locking," or deformity. By contrast, nonarticular disorders tend to be painful on active, but not passive (or assisted), range of motion. Periarticular conditions often demonstrate point or focal tenderness in regions adjacent to articular structures, may radiate or be elicited with a specific movement or position, and have physical findings remote from the joint capsule. Moreover, nonarticular disorders seldom demonstrate swelling, crepitus, instability, or deformity of the joint itself.

INFLAMMATORY VERSUS NONINFLAMMATORY DISORDERS

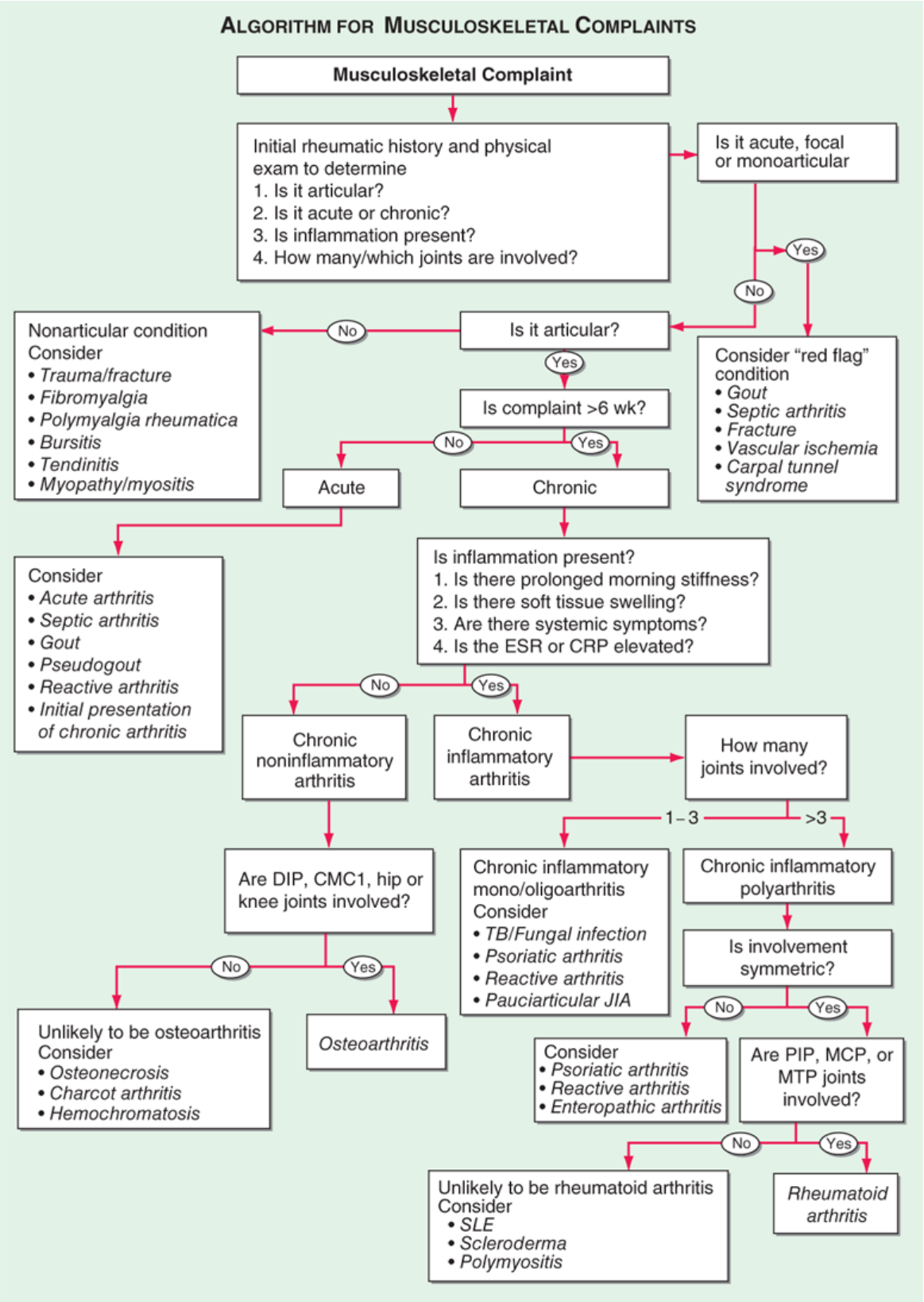
In the course of a musculoskeletal evaluation, the examiner should determine the nature of the underlying pathologic process and whether inflammatory or noninflammatory findings exist. Inflammatory disorders may be infectious (*Neisseria gonorrhoeae* or *Mycobacterium tuberculosis*), crystal-induced (gout, pseudogout), immune-related (rheumatoid arthritis [RA], systemic lupus erythematosus [SLE]), reactive (rheumatic fever, reactive arthritis), or idiopathic. Inflammatory disorders may be identified by any of the four cardinal signs of inflammation (erythema, warmth, pain, or swelling), systemic symptoms (fatigue, fever, rash, weight loss), or laboratory evidence of inflammation (elevated erythrocyte sedimentation rate [ESR] or C-reactive protein [CRP], thrombocytosis, anemia of chronic disease, or hypoalbuminemia). Articular stiffness commonly accompanies chronic musculoskeletal disorders. The duration of stiffness may be prolonged (hours) with inflammatory disorders (such as RA or polymyalgia rheumatica) and improves with activity. By contrast, intermittent stiffness (also known as gel phenomenon) is typical of noninflammatory conditions (such as osteoarthritis [OA]), shorter in duration (<60 min), and is exacerbated by activity. Fatigue may be profound with inflammation (as seen in RA and polymyalgia rheumatica) but may also be a consequence of fibromyalgia (a noninflammatory disorder), chronic pain, poor sleep, depression, anemia, cardiac failure, endocrinopathy, or malnutrition.

Noninflammatory disorders may be related to trauma (rotator cuff tear), repetitive use (bursitis, tendinitis), degeneration or ineffective repair (OA), neoplasm (pigmented villonodular synovitis), or pain amplification (fibromyalgia). Noninflammatory disorders are often characterized by pain without synovial swelling or warmth, absence of inflammatory or systemic features, daytime, intermittent gel phenomena rather than prolonged morning stiffness, and normal (for age) or negative laboratory investigations.

Identification of the nature of the underlying process and the site of the complaint will enable the examiner to characterize the musculoskeletal presentation (e.g., acute inflammatory monarthritis, chronic noninflammatory, nonarticular widespread pain). By narrowing the diagnostic considerations, the examiner can assess the need for immediate diagnostic or therapeutic intervention or for continued observation. Figure 363-1 presents an algorithmic approach to the evaluation of patients with musculoskeletal complaints. This approach relies on clinical and historic features, rather than laboratory testing, to diagnose many common rheumatic disorders.

FIGURE 363-1

Algorithm for the diagnosis of musculoskeletal complaints. An approach to formulating a differential diagnosis (shown in italics). CMC, carpometacarpal; CRP, C-reactive protein; DIP, distal interphalangeal; ESR, erythrocyte sedimentation rate; JIA, juvenile idiopathic arthritis; MCP, metacarpophalangeal; MTP, metatarsophalangeal; PIP, proximal interphalangeal; PMR, polymyalgia rheumatica; SLE, systemic lupus erythematosus.



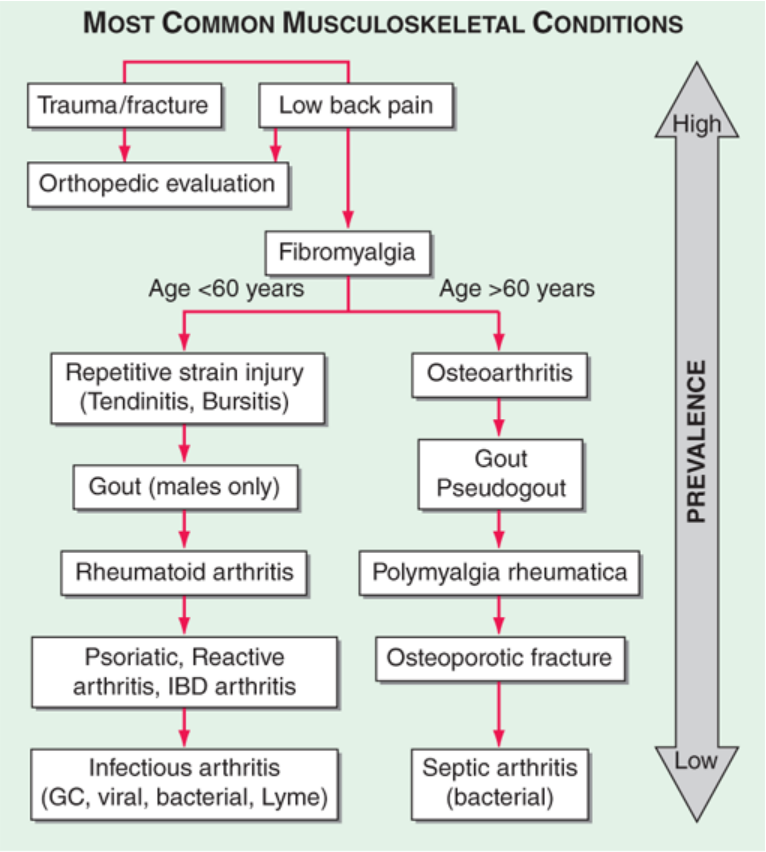
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A simpler, alternative approach would consider the most commonly encountered complaints first, based on frequency in younger versus older populations. The most prevalent causes of musculoskeletal complaints are shown in Fig. 363-2. Because trauma, fracture, overuse syndromes, and fibromyalgia are among the most common causes of musculoskeletal pain, these should be considered during the initial encounter. If excluded, other frequently occurring disorders should be considered according to the patient's age. Hence, those aged <60 years are commonly affected by repetitive use/strain disorders, gout (men only), RA, spondyloarthritis, and uncommonly, infectious arthritis. Patients aged >60 years are frequently affected by OA, crystal (gout and pseudogout) arthritis, polymyalgia rheumatica, osteoporotic fracture, and uncommonly, septic

arthritis. These conditions are between 10 and 100 times more prevalent than other serious autoimmune conditions, such as SLE, scleroderma, polymyositis, and vasculitis.

FIGURE 363-2

Algorithm for consideration of the most common musculoskeletal conditions. GC, gonococcal; IBD, inflammatory bowel disease.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CLINICAL HISTORY

Historic features may reveal important clues to the diagnosis. Aspects of the patient profile, complaint chronology, extent of joint involvement, and precipitating factors can provide important information. Certain diagnoses are more frequent in different *age* groups. SLE and reactive arthritis occur more frequently in the young, whereas fibromyalgia and RA are frequent in middle age, and OA and polymyalgia rheumatica are more prevalent among the elderly. Diagnostic clustering is also evident when *sex* and *race* are considered. Gout, spondyloarthritis, and ankylosing spondylitis are more common in men, whereas RA, fibromyalgia, osteoporosis and lupus are more frequent in women. *Racial predilections* may be evident. Thus, polymyalgia rheumatica, giant cell arteritis, and granulomatosis with polyangiitis (GPA; formerly called Wegener's granulomatosis) commonly affect whites, whereas sarcoidosis and SLE more commonly affect African Americans. *Familial aggregation* is most common with ankylosing spondylitis, gout, and Heberden's nodes of OA.

The chronology of the complaint is an important diagnostic feature and can be divided into the *onset, evolution*, and *duration*. The onset of disorders such as septic arthritis or gout tends to be abrupt, whereas OA, RA, and fibromyalgia may have more indolent presentations. The patients' complaints may evolve differently and be classified as chronic (OA), intermittent (crystal or Lyme arthritis), migratory (rheumatic fever, gonococcal or viral arthritis), or additive (RA, psoriatic arthritis). Musculoskeletal disorders are typically classified as acute or chronic based on a symptom duration that is either <6 weeks or >6 weeks, respectively. Acute arthropathies tend to be infectious, crystal-induced, or reactive. Chronic conditions include noninflammatory or immunologic arthritides (e.g., OA, RA) and nonarticular disorders (e.g., fibromyalgia).

The *extent* or *distribution* of articular involvement is often informative. Articular disorders are classified based on the number of joints involved, as either *monarticular* (one joint), *oligoarticular* or *pauciarticular* (two or three joints), or *polyarticular* (four or more joints). Although crystal and infectious arthritis are often mono- or oligoarticular, OA and RA are polyarticular disorders. Nonarticular disorders may be classified as either focal or widespread. Complaints secondary to tendinitis or carpal tunnel syndrome are typically focal, whereas weakness and myalgia, caused by polymyositis or fibromyalgia, are more widespread in their presentation. Joint involvement in RA tends to be symmetric and polyarticular. By

contrast, spondyloarthritis, reactive arthritis, gout, and sarcoid are often asymmetric and oligoarticular. OA and psoriatic arthritis may be either symmetric or asymmetric and oligo- or polyarticular. The upper extremities are frequently involved in RA and OA, whereas lower extremity arthritis is characteristic of reactive arthritis and gout at their onset. Involvement of the axial skeleton is common in OA and ankylosing spondylitis but is infrequent in RA, with the notable exception of the cervical spine.

The clinical history should also identify *precipitating events*, such as trauma (osteonecrosis, meniscal tear), drug administration **(Table 363-2)**, antecedent or intercurrent infection (rheumatic fever, reactive arthritis, hepatitis), or illnesses that may have contributed to the patient's complaint. Certain comorbidities may have musculoskeletal consequences. This is especially so for diabetes mellitus (carpal tunnel syndrome), renal insufficiency (gout), depression or insomnia (fibromyalgia), myeloma (low back pain), cancer (myositis), and osteoporosis (fracture) or when using certain drugs such as glucocorticoids (osteonecrosis, septic arthritis), diuretics or chemotherapy (gout) (Table 363-2).

TABLE 363-2

Drug-Induced Musculoskeletal Conditions

Arthralgias
Quinidine, cimetidine, beta blockers, quinolones, chronic acyclovir, interferons, IL-2, nicardipine, vaccines, rifabutin, aromatase inhibitors, HIV protease inhibitors, DPP-4 inhibitors (sitagliptin, linagliptin, alogliptin), checkpoint inhibitors (ipilimumab, pembrolizumab, nivolumab)
Myalgias/myopathy
Glucocorticoids, penicillamine, hydroxychloroquine, AZT, lovastatin, simvastatin, atorvastatin, pravastatin, clofibrate, amiodarone, interferon, IL-2, alcohol, cocaine, paclitaxel, docetaxel, imatinib mesylate, colchicine, quinolones, cyclosporine, tacrolimus, protease inhibitors, checkpoint inhibitors
Tendon rupture/tendinitis
Quinolones, glucocorticoids, isotretinoin, statins, aromatase inhibitors, collagenase injections
Gout
Diuretics, aspirin, cytotoxics, cyclosporine, alcohol, moonshine, ethambutol, fructose-containing soft drinks
Drug-induced lupus
Hydralazine, procainamide, quinidine, phenytoin, carbamazepine, methyldopa, isoniazid, chlorpromazine, lithium, penicillamine, tetracyclines, TNF inhibitors, ACE inhibitors, ticlopidine, aromatase inhibitors
Drug-induced subacute lupus
Proton pump inhibitors, calcium channel blockers (diltiazem), ACE inhibitors, TNF inhibitors, terbinafine, interferons (α and β-1a), paclitaxel, docetaxel, gemcitabine, capecitabine, aromatase inhibitors, HCTZ
Osteonecrosis/Atypical fractures
Glucocorticoids, alcohol, radiation, bisphosphonates
Osteopenia
Glucocorticoids, chronic heparin, phenytoin, aromatase inhibitors, anti-androgen therapy, thiazolinediones
Psoriasis
TNF inhibitors, beta blockers, lithium, hydroxychloroquine, chloroquine, minocycline, ACE inhibitors, terbinafine
Scleroderma
Vinyl chloride, bleomycin, baricitinib, pentazocine, organic solvents, carbidopa, tryptophan, rapeseed oil
Raynaud's phenomenon
Cisplatin, bleomycin, beta blockers, clonidine, bromocriptine, ergot alkaloids, cocaine, methylphenidate, dextroamphetamine, phenteramine, interferon therapy
Vasculitis
Allopurinol, amphetamines, cocaine (often levamisole adulterated), cannabis, thiazides, penicillamine, propylthiouracil, montelukast, TNF inhibitors, hepatitis B vaccine, trimethoprim/sulfamethoxazole, minocycline, hydralazine

Abbreviations: ACE, angiotensin-converting enzyme; AZT, zidovudine; HCTZ, hydrochlorothiazide; IL-2, interleukin 2; TNF, tumor necrosis factor.

Lastly, a thorough *rheumatic review of systems* may disclose useful diagnostic information. A variety of musculoskeletal disorders may be associated with systemic features such as fever (SLE, infection), rash (SLE, psoriatic arthritis), nail abnormalities (psoriatic or reactive arthritis), myalgias (fibromyalgia, statin- or drug-induced myopathy), or weakness (polymyositis, neuropathy). In addition, some conditions are associated with involvement of other organ systems including the eyes (Behçet's disease, sarcoidosis, spondyloarthritis), gastrointestinal tract (scleroderma, inflammatory bowel disease), genitourinary tract (reactive arthritis, gonococcemia), or nervous system (Lyme disease, vasculitis).

FIBROMYALGIA

Historically, syphilis and tuberculosis were labeled as the "great masqueraders" as their protean symptoms and potential for multi-organ involvement may result in delays in diagnosis and treatment. In the modern era, other serious diagnoses (including lupus, sarcoidosis, vasculitis and lymphoma) have also been labeled as great masqueraders. All of these are either uncommon or rare, and are overshadowed by the most common masquerader with musculoskeletal complaints—fibromyalgia. Fibromyalgia **(see Chap. 366)** is a pain amplification disorder unified by sleep disturbance, exaggerated pain and sensitivity (owing to lowered pain thresholds), and a multiplicity of symptoms with a paucity of abnormalities on clinical examination or laboratory testing. Tender "trigger points" are often found and include tenderness over the epicondyles, trochanteric bursae, anserine bursae, and muscles (gluteal, trapezius, supraspinatus) that often are misdiagnosed as other nonarticular conditions. Although fibromyalgia classically manifests as widespread aches and pains, presenting symptoms tend to be less specific, and only on further evaluation will the widespread noninflammatory features be disclosed. Fibromyalgia has numerous comorbidities including irritable bowel syndrome, dysmenorrhea, migraine, depression, anxiety, memory loss, non-anatomic paresthesia or dysesthesia, fatigue, myalgias, temporomandibular joint pain, and multiple chemical sensitivities. Fibromyalgia patients often see multiple specialists, are twice as likely to be hospitalized, and are plagued by polypharmacy. Fibromyalgia affects nearly 5 million Americans. Yet, fibromyalgia is frequently underrecognized or misdiagnosed as arthritis, lupus, multiple sclerosis, autoimmune disease, etc. Hence, patients are often referred to multiple consultants and are subjected to multiple investigations and even surgical interventions. Early consideration of this very common disorder can avert needless investigation, therapy, and concern for those afflicted (Fig. 363-2).

RHEUMATOLOGIC EVALUATION OF THE ELDERLY

The incidence of rheumatic diseases rises with age, such that 58% of those >65 years will have joint complaints. Musculoskeletal disorders in elderly patients are often not diagnosed because the signs and symptoms may be insidious, overlooked, or overshadowed by comorbidities. These difficulties are compounded by the diminished reliability of laboratory testing in the elderly, who often manifest nonpathologic abnormal results. For example, the ESR may be misleadingly elevated, and low-titer positive tests for rheumatoid factor (RF) and antinuclear antibodies (ANAs) may be seen in up to 15% of elderly patients. Although nearly all rheumatic disorders afflict the elderly, geriatric patients are particularly prone to OA, osteoporosis, osteoporotic fractures, gout, pseudogout, polymyalgia rheumatica, vasculitis, and drug-induced disorders (Table 363-2). The elderly should be approached in the same manner as other patients with musculoskeletal complaints, but with an emphasis on identifying the potential rheumatic consequences of medical comorbidities and therapies. The physical examination should identify the nature of the musculoskeletal complaint as well as coexisting diseases that may influence diagnosis and choice of treatment.

RHEUMATOLOGIC EVALUATION OF THE HOSPITALIZED PATIENT

Evaluation of a hospitalized patient with rheumatic complaints is often more complex owing to greater symptom severity, more acute presentations, and greater interplay of comorbidities. In patients with rheumatic disorders tend to be admitted for one of several reasons: (1) acute onset of inflammatory arthritis (possibly gout or septic arthritis); (2) undiagnosed systemic or febrile illness; (3) musculoskeletal trauma; (4) exacerbation or deterioration of an existing autoimmune disorder (e.g., SLE); or (5) new medical comorbidities (e.g., thrombotic event, lymphoma, infection) arising in patients with an established rheumatic disorder. Notably, rheumatic patients are seldom if ever admitted because of widespread pain or serologic abnormalities or for the initiation of new therapies.

Acute monarticular inflammatory arthritis may be a "red flag" condition (e.g., septic arthritis, gout, pseudogout) that will require arthrocentesis and, on occasion, hospitalization if infection is suspected. However, new-onset inflammatory polyarthritis will have a wider differential diagnosis (e.g., RA, hepatitis-related arthritis, chikungunya arthritis, serum sickness, drug-induced lupus, polyarticular septic arthritis) and may require targeted laboratory investigations rather than synovial fluid analyses. Patients with febrile, multisystem disorders will require exclusion of crystal, infectious, or neoplastic etiologies and an evaluation driven by the dominant symptom/finding with the greatest specificity. Conditions worthy of consideration may include gout or pseudogout, vasculitis (giant cell arteritis in the elderly or polyarteritis nodosa in younger patients), adult-onset Still's disease, SLE, antiphospholipid antibody syndrome, IgG4-related disease, and sarcoidosis. A preexisting rheumatic diagnosis (e.g., SLE, RA,

ankylosing spondylitis) should be confirmed by careful history, examination and review of medical records, as this will influence the ensuing in-patient evaluation. It is important to note that when established rheumatic disease patients are admitted to the hospital, it is usually not for a medical problem related to their autoimmune disease, but rather because of either a comorbid condition or complication of drug therapy. Patients with chronic inflammatory disorders (e.g., RA, SLE, psoriasis) have an augmented risk of infection, cardiovascular events, and neoplasia.

Certain conditions, such as acute gout, can be precipitated in hospitalized patients by surgery, dehydration, or medications and should be considered when hospitalized patients are evaluated for the acute onset of a musculoskeletal condition. Lastly, overly aggressive and unfocused laboratory testing will often yield abnormal findings that are better explained by the patient's preexisting condition (chronic lung, renal, or liver disease) rather than a new inflammatory or autoimmune disorder (lupus, vasculitis).

PHYSICAL EXAMINATION

The goal of the physical examination is to ascertain the structures involved, the nature of the underlying pathology, the functional consequences of the process, and the presence of systemic or extraarticular manifestations. A knowledge of topographic anatomy is necessary to identify the primary site(s) of involvement and differentiate articular from nonarticular disorders. The musculoskeletal examination depends largely on careful inspection, palpation, and a variety of specific physical maneuvers to elicit diagnostic signs (Table 363-3). Although most articulations of the appendicular skeleton can be examined in this manner, adequate inspection and palpation are not possible for many axial (e.g., zygapophyseal) and inaccessible (e.g., sacroiliac or hip) joints. For such joints, there is a greater reliance on specific maneuvers and imaging for assessment.

TABLE 363-3

Glossary of Musculoskeletal Terms

Crepitus
A palpable (less commonly audible) vibratory or crackling sensation elicited with joint motion; fine joint crepitus is common and often insignificant in large joints; coarse joint crepitus indicates advanced cartilaginous and degenerative changes (as in osteoarthritis)
Subluxation
Alteration of joint alignment such that articulating surfaces incompletely approximate each other
Dislocation
Abnormal displacement of articulating surfaces such that the surfaces are not in contact
Range of motion
For diarthrodial joints, the arc of measurable movement through which the joint moves in a single plane
Contracture
Loss of full movement resulting from a fixed resistance caused either by tonic spasm of muscle (reversible) or by fibrosis of periarticular structures (permanent)
Deformity
Abnormal shape or size resulting from bony hypertrophy, malalignment of articulating structures, or damage to periarticular supportive structures
Enthesitis
Inflammation of the entheses (tendinous or ligamentous insertions on bone)
Epicondylitis
Infection or inflammation involving an epicondyle

Examination of involved and uninvolved joints will determine whether *pain, warmth, erythema*, or *swelling* is present. The locale and level of pain elicited by palpation or movement should be quantified. One standard would be to count the number of tender joints on palpation of 28 easily examined joints (proximal interphalangeals [PIPs], metacarpophalangeals [MCPs], wrists, elbows, shoulders, and knees). Similarly, the number of swollen joints (0–28) can be counted and recorded. Careful examination should distinguish between true articular swelling (caused by bony hypertrophy, synovial effusion or proliferation), and nonarticular (or periarticular) involvement, which usually extends beyond the normal joint margins. Synovial effusion can be distinguished from synovial hypertrophy or bony hypertrophy by palpation or specific maneuvers. For example, small to moderate knee effusions may be identified by the “bulge sign” or “ballottement of the patellae.” Bursal effusions (e.g., effusions of the olecranon or prepatellar bursa) are often focal, periarticular, overlie bony prominences, and are fluctuant with defined borders. Joint *stability* can be assessed by stabilizing the proximal joint, by palpation, and by the application of manual stress to the distal appendage. *Subluxation* or *dislocation*, which may be secondary to traumatic, mechanical, or inflammatory causes, can be assessed by inspection and palpation. Joint *swelling* or *volume* can be assessed by palpation. Distention of the articular capsule usually causes pain and evident enlargement or fluctuance. The patient will attempt to minimize the pain by maintaining the joint in the position of least intraarticular pressure and greatest volume, usually partial flexion. For this reason, inflammatory effusions may give rise to flexion contractures. Clinically, this may be detected as fluctuant or “squishy” swelling in larger joints and grape-like compressibility in smaller joints. Inflammation may result in fixed flexion deformities or diminished range of motion—especially on extension, when intraarticular pressure is increased. Active and passive *range of motion* should be assessed in all planes, with contralateral comparison. A goniometer may be used to quantify the arc of movement. Each joint should be passively manipulated through its full range of motion (including, as appropriate, flexion, extension, rotation, abduction, adduction, lateral bending,

inversion, eversion, supination, pronation, medial/lateral deviation, and plantar- or dorsiflexion). Extreme range of motion may be seen with hypermobility syndrome, with joint pain and connective tissue laxity, often associated with Ehlers-Danlos or Marfan's syndrome. Limitation of motion is frequently caused by inflammation, effusion, pain, deformity, contracture, or restriction from neuromyopathic causes. If passive motion exceeds active motion, a periarticular process (e.g., tendinitis, tendon rupture, or myopathy) should be considered. *Contractures* may reflect antecedent synovial inflammation or trauma. Minor joint *crepitus* is common during joint palpation and maneuvers but may indicate significant cartilage degeneration as it becomes coarser (e.g., OA). Joint *deformity* usually indicates a long-standing or aggressive pathologic process. Deformities may result from ligamentous destruction, soft tissue contracture, bony enlargement, ankylosis, erosive disease, subluxation, trauma, or loss of proprioception. Examination of the musculature will document strength, atrophy, pain, or spasm. Appendicular muscle weakness should be characterized as proximal or distal. Muscle strength should be assessed by observing the patient's performance (e.g., walking, rising from a chair, grasping, writing). Strength may also be graded on a 5-point scale: 0 for no movement; 1 for trace movement or twitch; 2 for movement with gravity eliminated; 3 for movement against gravity only; 4 for movement against gravity and resistance; and 5 for normal strength. The examiner should assess for often-overlooked nonarticular or periarticular involvement, especially when articular complaints are not supported by objective findings referable to the joint capsule. The identification of soft tissue/nonarticular pain will prevent unwarranted and often expensive additional evaluations. Specific maneuvers may reveal common nonarticular abnormalities, such as a carpal tunnel syndrome (which can be identified by Tinel's or Phalen's sign). Other examples of soft tissue abnormalities include olecranon bursitis, epicondylitis (e.g., tennis elbow), enthesitis (e.g., Achilles tendinitis), and tender trigger points associated with fibromyalgia.

APPROACH TO REGIONAL RHEUMATIC COMPLAINTS

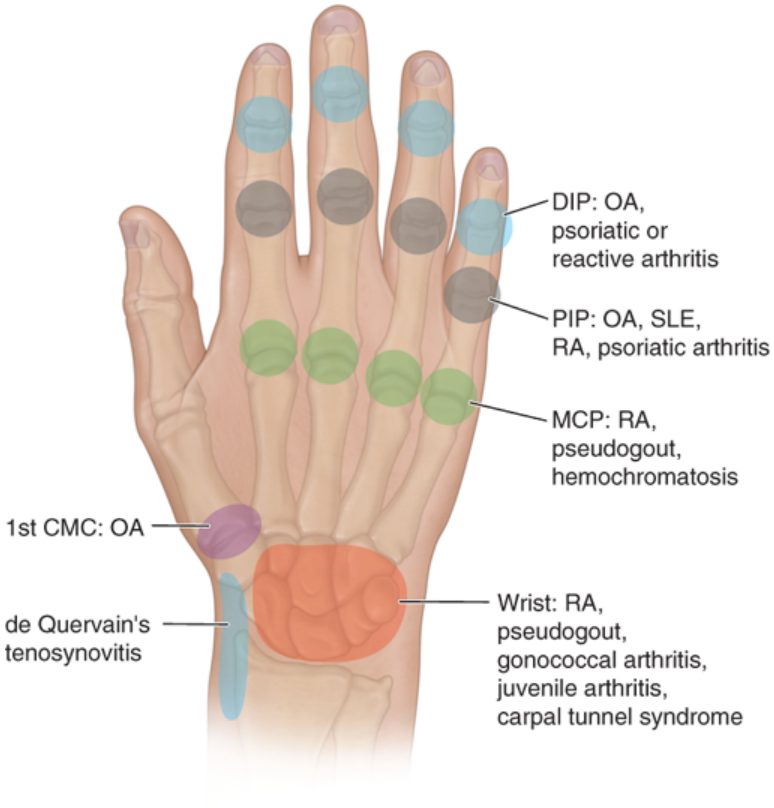
Although all patients should be evaluated in a logical and thorough manner, many cases with focal musculoskeletal complaints are caused by commonly encountered disorders that exhibit a predictable pattern of onset, evolution, and localization; they can often be diagnosed immediately on the basis of limited historic information and selected maneuvers or tests. Although nearly every musculoskeletal complaint could be approached in this manner, the evaluation of four common involved anatomic regions—the hand, shoulder, hip, and knee—are reviewed here.

HAND PAIN

Focal or unilateral hand pain may result from trauma, overuse, infection, or a reactive or crystal-induced arthritis. By contrast, bilateral hand complaints commonly suggest a degenerative (e.g., OA), systemic, or inflammatory/immune (e.g., RA) etiology. The distribution or pattern of joint involvement is highly suggestive of certain disorders **(Fig. 363-3)**. Thus, OA (or degenerative arthritis) may manifest as distal interphalangeal (DIP) and PIP joint pain with bony hypertrophy sufficient to produce Heberden's and Bouchard's nodes, respectively. Pain, with or without bony swelling, involving the base of the thumb (first carpometacarpal joint) is also highly suggestive of OA. By contrast, RA tends to cause symmetric, polyarticular involvement of the PIP, MCP, intercarpal, and carpometacarpal joints (wrist) with pain and palpable synovial tissue hypertrophy. Psoriatic arthritis may mimic the pattern of joint involvement seen in OA (DIP and PIP joints), but can be distinguished by the presence of inflammatory signs (erythema, warmth, synovial swelling), with or without carpal involvement, nail pitting, or onycholysis. Whereas lateral or medial subluxations at the PIP or DIP joints are most likely due to inflammatory OA or psoriatic arthritis, dorsal or ventral deformities (swan neck or boutonnière deformities) are typical of RA. Hemochromatosis should be considered when degenerative changes (bony hypertrophy) are seen at the second and third MCP joints with associated radiographic chondrocalcinosis or episodic, inflammatory wrist arthritis.

FIGURE 363-3

Sites of hand or wrist involvement and their potential disease associations. CMC, carpometacarpal; DIP, distal interphalangeal; MCP, metacarpophalangeal; OA, osteoarthritis; PIP, proximal interphalangeal; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus. *(From JJ Cush et al: Evaluation of musculoskeletal complaints, in Rheumatology: Diagnosis and Therapeutics, 2nd ed, JJ Cush et al [eds]. Philadelphia, Lippincott Williams & Wilkins, 2005, pp 3–20. Used with permission from Dr. John J. Cush.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Dactylitis manifests as soft tissue swelling of the whole digit and may have a sausage-like appearance. Common causes of dactylitis include psoriatic arthritis, spondyloarthritis, juvenile spondylitis, mixed connective tissue disease, scleroderma, sarcoidosis, and sickle cell disease. Soft tissue swelling over the dorsum of the hand and wrist may suggest an inflammatory extensor tendon tenosynovitis possibly caused by gonococcal infection, gout, or inflammatory arthritis (e.g., RA). Tenosynovitis is suggested by localized warmth, swelling, or pitting edema and may be confirmed when the soft tissue swelling tracks with tendon movement during flexion and extension of fingers, or when pain is induced while stretching the extensor tendon sheaths (flexing the digits distal to the MCP joints and maintaining the wrist in a fixed, neutral position).

Focal wrist pain localized to the radial aspect may be caused by de Quervain's tenosynovitis resulting from inflammation of the tendon sheath(s) involving the abductor pollicis longus or extensor pollicis brevis (Fig. 363-3). This commonly results from overuse or follows pregnancy and may be diagnosed with Finkelstein's test. A positive result is present when radial wrist pain is induced after the thumb is flexed and placed inside a clenched fist and the patient actively deviates the hand downward with ulnar deviation at the wrist. Carpal tunnel syndrome is another common disorder of the upper extremity and results from compression of the median nerve within the carpal tunnel. Manifestations include pain in the wrist that may radiate with paresthesia to the thumb, second and third fingers, and radial half of the fourth finger and, at times, atrophy of thenar musculature. Carpal tunnel syndrome is commonly associated with pregnancy, edema, trauma, OA, inflammatory arthritis, and infiltrative disorders (e.g., amyloidosis). The diagnosis may be suggested by a positive Tinel's or Phalen's sign. With each test, paresthesia in a median nerve distribution is induced or increased by either "thumping" the volar aspect of the wrist (Tinel's sign) or pressing the extensor surfaces of both flexed wrists against each other (Phalen's sign). The low sensitivity and moderate specificity of these tests may require nerve conduction velocity testing to confirm a suspected diagnosis.

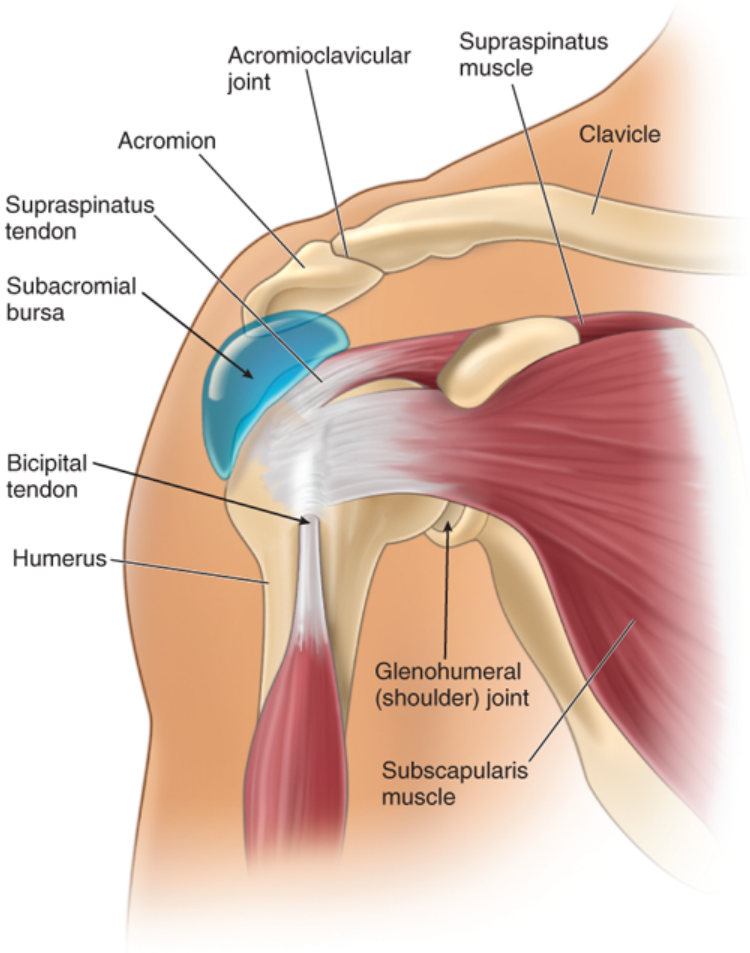
SHOULDER PAIN

During the evaluation of shoulder disorders, the examiner should carefully note any history of trauma, fibromyalgia, infection, inflammatory disease, occupational hazards, or previous cervical disease. In addition, the patient should be questioned as to the activities or movement(s) that elicit shoulder pain. While arthritis is suggested by pain on movement in all planes, pain with specific active motion suggests a periarticular (nonarticular) process. Shoulder pain may originate in the glenohumeral or acromioclavicular joints, subacromial (subdeltoid) bursa, periarticular soft tissues (e.g., fibromyalgia, rotator cuff tear/tendinitis), or cervical spine **(Fig. 363-4)**. Shoulder pain is referred frequently from the cervical spine but may also be referred from intrathoracic lesions (e.g., a Pancoast tumor) or from gallbladder, hepatic, or diaphragmatic disease. These same visceral causes may also manifest as focal scapular pain. Fibromyalgia should be suspected when glenohumeral pain is accompanied by diffuse periarticular (i.e., subacromial, bicipital) pain, tender points (i.e., trapezius or supraspinatus), and a sleep disturbance. The shoulder should be put through its full range of motion both actively and passively (with examiner assistance): forward flexion, extension, abduction, adduction,

and internal and external rotation. Manual inspection of the periarticular structures will often provide important diagnostic information. Glenohumeral involvement is best detected by placing the thumb over the glenohumeral joint just medial and inferior to the coracoid process and applying pressure anteriorly while internally and externally rotating the humeral head. Pain localized to this region is indicative of glenohumeral pathology. Synovial effusion or tissue is seldom palpable but, if present, may suggest infection, RA, amyloidosis, or an acute tear of the rotator cuff. The examiner should apply direct manual pressure over the subacromial bursa that lies lateral to and immediately beneath the acromion (Fig. 363-4). Subacromial bursitis is a frequent cause of shoulder pain. Anterior to the subacromial bursa, the bicipital tendon traverses the bicipital groove. This tendon is best identified by palpating it in its groove as the patient rotates the humerus internally and externally. Direct pressure over the tendon may reveal pain indicative of bicipital tendinitis. Palpation of the acromioclavicular joint may disclose local pain, bony hypertrophy, or, uncommonly, synovial swelling. Whereas OA and RA commonly affect the acromioclavicular joint, OA seldom involves the glenohumeral joint, unless there is a traumatic or occupational cause.

FIGURE 363-4

Origins of shoulder pain. The schematic diagram of the shoulder indicates, with arrows, the anatomic origins of shoulder pain.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Rotator cuff tendinitis or tear is a very common cause of shoulder pain. Nearly 30% of the elderly will have shoulder pain, with rotator cuff tendinitis or tear as a primary cause. The rotator cuff is formed by four tendons that attach the scapula to the proximal humerus (supraspinatus, infraspinatus, teres minor, and subscapularis tendons). Of these, the supraspinatus muscle is the most commonly damaged. Rotator cuff tendinitis is suggested by pain on active abduction (but not passive abduction), pain over the lateral deltoid muscle, night pain, and evidence of the impingement signs (pain with overhead arm activities). The Neer test for impingement is performed by the examiner raising the patient's arm into forced flexion while stabilizing and preventing rotation of the scapula. A positive sign is present if pain develops before 180° of forward flexion. Tear of the rotator cuff is common in the elderly and often results from trauma; it may manifest in the same manner as tendinitis. The drop arm test is abnormal with supraspinatus pathology and is demonstrated by passive abduction of the arm to 90° by the examiner. If the patient is unable to hold the arm up actively or unable to lower the arm slowly without dropping, the test is positive. Tendinitis or tear of the rotator cuff is best confirmed by magnetic resonance imaging (MRI) or ultrasound.

KNEE PAIN

Knee pain may result from intraarticular (OA, RA) or periarticular (anserine bursitis, collateral ligament strain) processes or be referred from hip pathology. A careful history should delineate the chronology of the knee complaint and whether there are predisposing conditions, trauma, or medications that might underlie the complaint. For example, patellofemoral disease (e.g., OA) may cause anterior knee pain that worsens with climbing stairs. Observation of the patient's gait is also important. The knee should be carefully inspected in the upright (weight-bearing) and supine positions for swelling, erythema, malalignment, visible trauma, muscle wasting, and leg length discrepancy. The most common malalignment in the knee is *genu varum* (bowlegs) or *genu valgum* (knock-knees) resulting from asymmetric cartilage loss medially or laterally. Bony swelling of the knee joint commonly results from hypertrophic osseous changes seen with disorders such as OA and neuropathic arthropathy. Swelling caused by hypertrophy of the synovium or synovial effusion may manifest as a fluctuant, ballotable, or soft tissue enlargement in the suprapatellar pouch (suprapatellar reflection of the synovial cavity) or regions lateral and medial to the patella. Synovial effusions may also be detected by balloting the patella downward toward the femoral groove or by eliciting a "bulge sign." With the knee extended, the examiner should manually compress, or "milk," synovial fluid down from the suprapatellar pouch and lateral to the patellae. The application of manual pressure lateral to the patella may cause an observable shift in synovial fluid (bulge) to the medial aspect. The examiner should note that this maneuver is only effective in detecting small to moderate effusions (<100 mL). Inflammatory disorders such as RA, gout, pseudogout, and psoriatic arthritis may involve the knee joint and produce significant pain, stiffness, swelling, or warmth. A popliteal or *Baker's cyst* may be palpated with the knee partially flexed and is best viewed posteriorly with the patient standing and knees fully extended to visualize isolated or unilateral popliteal swelling or fullness.

Anserine bursitis is an often missed periarticular cause of knee pain in adults. The pes anserine bursa underlies the insertion of the conjoined tendons (sartorius, gracilis, semitendinosus) on the anteromedial proximal tibia and may be painful following trauma, overuse, or inflammation. It is often tender in patients with fibromyalgia, obesity, and knee OA. Other forms of bursitis may also present as knee pain. The prepatellar bursa is superficial and is located over the inferior portion of the patella. The infrapatellar bursa is deeper and lies beneath the patellar ligament before its insertion on the tibial tubercle.

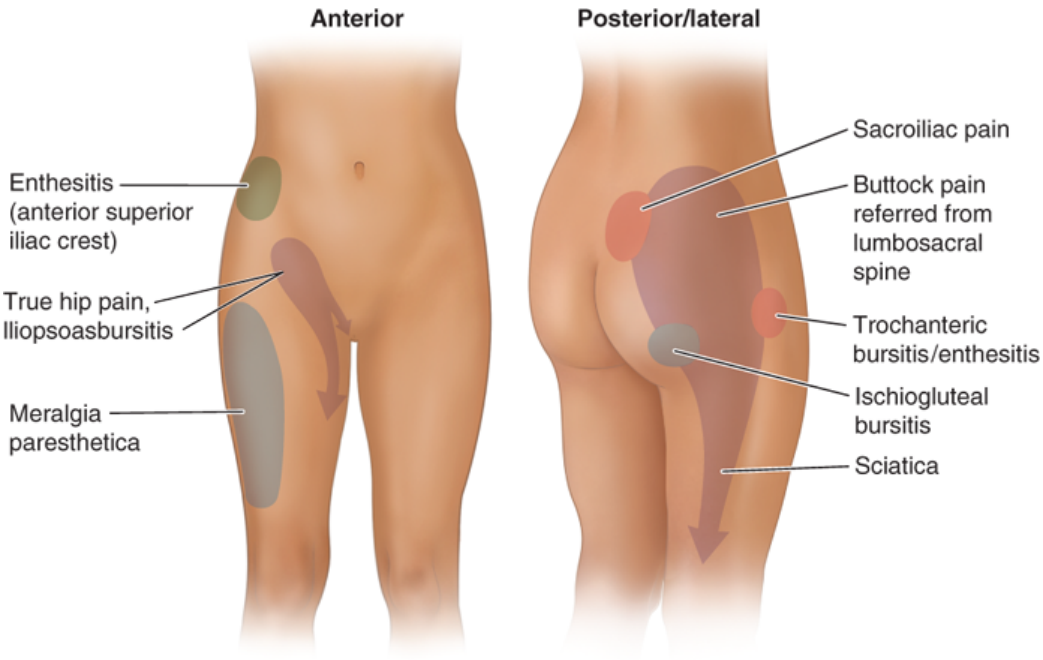
Internal derangement of the knee may result from trauma or degenerative processes. Damage to the meniscal cartilage (medial or lateral) frequently presents as chronic or intermittent knee pain. Such an injury should be suspected when there is a history of trauma, athletic activity, or chronic knee arthritis, and when the patient relates symptoms of "locking" or "giving way" of the knee. With the knee flexed 90° and the patient's foot on the table, pain elicited during palpation over the joint line or when the knee is stressed laterally or medially may suggest a meniscal tear. A positive McMurray test may also indicate a meniscal tear. To perform this test, the knee is first flexed at 90°, and the leg is then extended while the lower extremity is simultaneously torqued medially or laterally. A painful click during inward rotation may indicate a lateral meniscus tear, and pain during outward rotation may indicate a tear in the medial meniscus. Lastly, damage to the cruciate ligaments should be suspected with acute onset of pain, possibly with swelling, a history of trauma, or a synovial fluid aspirate that is grossly bloody. Examination of the cruciate ligaments is best accomplished by eliciting a drawer sign. With the patient recumbent, the knee should be partially flexed and the foot stabilized on the examining surface. The examiner should manually attempt to displace the tibia anteriorly or posteriorly with respect to the femur. If anterior movement is detected, then anterior cruciate ligament damage is likely. Conversely, significant posterior movement may indicate posterior cruciate damage. Contralateral comparison will assist the examiner in detecting significant anterior or posterior movement.

HIP PAIN

The hip is best evaluated by observing the patient's gait and assessing range of motion. The vast majority of patients reporting "hip pain" localize their pain unilaterally to the posterior gluteal musculature **(Fig. 363-5)**. Such pain tends to radiate down the posterolateral aspect of the thigh and may or may not be associated with complaints of low back pain. This presentation frequently results from degenerative arthritis of the lumbosacral spine or disks and commonly follows a dermatomal distribution with involvement of nerve roots between L4 and S1. Sciatica is caused by impingement of the L4, L5, or S1 nerve (i.e., from a herniated disk) and manifests as unilateral neuropathic pain extending from the gluteal region down the posterolateral leg to the foot. Some individuals instead localize their "hip pain" laterally to the area overlying the trochanteric bursa. Because of the depth of this bursa, swelling and warmth are usually absent. Diagnosis of trochanteric bursitis or enthesitis can be confirmed by inducing point tenderness over the trochanteric bursa. Gluteal and trochanteric pain are common findings in fibromyalgia. Range of movement may be limited by pain. Pain in the hip joint is less common and tends to be located anteriorly, over the inguinal ligament; it may radiate medially to the groin. Uncommonly, iliopsoas bursitis may mimic true hip joint pain. Diagnosis of iliopsoas bursitis may be suggested by a history of trauma or inflammatory arthritis. Pain associated with iliopsoas bursitis is localized to the groin or anterior thigh and tends to worsen with hyperextension of the hip; many patients prefer to flex and externally rotate the hip to reduce the pain from a distended bursa.

FIGURE 363-5

Origins of hip pain and dysesthesias. *(From JJ Cush et al: Evaluation of musculoskeletal complaints, in Rheumatology: Diagnosis and Therapeutics, 2nd ed, JJ Cush et al [eds]. Philadelphia, Lippincott Williams & Wilkins, 2005, pp 3–20. Used with permission from Dr. John J. Cush.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

LABORATORY INVESTIGATIONS

The vast majority of musculoskeletal disorders can be logically diagnosed by a complete history and physical examination. An additional objective of the initial encounter is to determine whether additional investigations or immediate therapy is required. Additional evaluation is indicated with: (1) monarticular conditions; (2) traumatic or inflammatory conditions; (3) the presence of neurologic findings; (4) systemic manifestations; or (5) chronic symptoms (>6 weeks) and a lack of response to symptomatic measures. The extent and nature of the additional investigation should be dictated by the clinical features and suspected pathologic process. Laboratory tests should be used to confirm a specific clinical diagnosis and not be used to screen or evaluate patients with vague rheumatic complaints. Indiscriminate use of broad batteries of diagnostic tests and radiographic procedures is rarely a useful or cost-effective means to establish a diagnosis.

Besides a complete blood count, including a white blood cell (WBC) and differential count, the routine evaluation should include a determination of an acute-phase reactant such as the ESR or CRP, which can be useful in discriminating inflammatory from noninflammatory disorders. Both are inexpensive, easily obtained, and may be elevated with infection, inflammation, autoimmune disorders, neoplasia, pregnancy, renal insufficiency, advanced age, or hyperlipidemia. Extreme elevation of the acute-phase reactants (CRP, ESR) is seldom seen without evidence of serious illness (e.g., sepsis, pleuropericarditis, polymyalgia rheumatica, giant cell arteritis, adult Still's disease).

Serum uric acid determinations are useful in the diagnosis of gout and in monitoring the response to urate-lowering therapy. Uric acid, the end product of purine metabolism, is primarily excreted in the urine. Serum values range from 238 to 516 µmol/L (4.0–8.6 mg/dL) in men; the lower values (178–351 µmol/L [3.0–5.9 mg/dL]) seen in women are caused by the uricosuric effects of estrogen. Urinary uric acid levels are normally <750 mg per 24 h. Although hyperuricemia (especially levels >535 µmol/L [>9 mg/dL]) is associated with an increased incidence of gout and nephrolithiasis, levels may not correlate with the severity of articular disease. Uric acid levels (and the risk of gout) may be increased by inborn errors of metabolism (Lesch-Nyhan syndrome), disease states (renal insufficiency, myeloproliferative disease, psoriasis), or drugs (alcohol, cytotoxic therapy, thiazides). Although nearly all patients with gout will demonstrate hyperuricemia at some time during their illness, up to 50% of patients with an acute gouty attack will have normal serum uric acid levels. Monitoring serum uric acid is useful in assessing the response to urate-lowering therapy or chemotherapy, with the target goal being a serum urate <6 mg/dL.

Serologic tests for RF, cyclic anticitrullinated peptide (CCP or ACPA) antibodies, ANAs, complement levels, Lyme and antineutrophil cytoplasmic antibodies (ANCA), or antistreptolysin O (ASO) titer should be carried out only when there is clinical evidence to specifically suggest an associated diagnosis because these have poor predictive value when used for screening, especially when the pretest probability is low. For most of these, there is no value to repeated or serial serologic testing. Although 4–5% of a healthy population will have positive tests for RF and ANAs, only 1% and <0.4% of the population will have RA or SLE, respectively. IgM RF (autoantibodies against the Fc portion of IgG) is found in 80% of patients with RA and may also be seen in low titers in patients with chronic infections (tuberculosis, leprosy, hepatitis); other autoimmune diseases (SLE, Sjögren's syndrome); and chronic pulmonary, hepatic, or renal diseases. When considering RA, both serum RF and anti-CCP antibodies should be

obtained as these are complementary. Both are comparably sensitive, but CCP antibodies are more specific than RF. In RA, the presence of anti-CCP and RF antibodies may indicate a greater risk for more severe, erosive polyarthritis. ANAs are found in nearly all patients with SLE and may also be seen in patients with other autoimmune diseases (polymyositis, scleroderma, antiphospholipid syndrome, Sjögren's syndrome), drug-induced lupus (Table 363-2), chronic liver or renal disorders, and advanced age. Positive ANAs are found in 5% of adults and in up to 14% of elderly or chronically ill individuals. The ANA test is very sensitive but poorly specific for lupus, as only 1–2% of all positive results will be caused by lupus alone. The interpretation of a positive ANA test may depend on the magnitude of the titer and the pattern observed by immunofluorescence microscopy **(Table 363-4)**. Diffuse and speckled patterns are least specific, whereas a peripheral, or rim, pattern (related to autoantibodies against double-strand [native] DNA) is highly specific and suggestive of lupus. Centromeric patterns are seen in patients with limited scleroderma (*c*alcinosis, *R*aynaud's phenomenon, *e*sophageal involvement, *s*clerodactyly, *t*elangiectasia [CREST] syndrome), primary biliary sclerosis, Sjögren's syndrome or thyroiditis and nucleolar patterns may be seen in patients with diffuse systemic sclerosis or inflammatory myositis.

TABLE 363-4

Antinuclear Antibody (ANA) Patterns and Clinical Associations

ANA Pattern	Antigen Identified	Clinical Correlate
Diffuse	Deoxyribonucleoprotein	Nonspecific
	Histones	Drug-induced lupus, lupus
Peripheral (rim)	ds-DNA	50% of SLE (specific)
Speckled	U1-RNP	>90% of MCTD
	Sm	30% of SLE (specific)
	Ro (SS-A)	Sjögren's 60%, SCLE, neonatal lupus, ANA(–) lupus
	La (SS-B)	50% of Sjögren's, 15% lupus
	Scl-70	40% of diffuse scleroderma
	PM-1	Polymyositis (PM), dermatomyositis
	Jo-1	PM w/pneumonitis + arthritis
Nucleolar	RNA polymerase I, others	40% of PSS
Centromere	Kinetochore	75% CREST (limited scleroderma), PBC, Sjogren's, thyroiditis

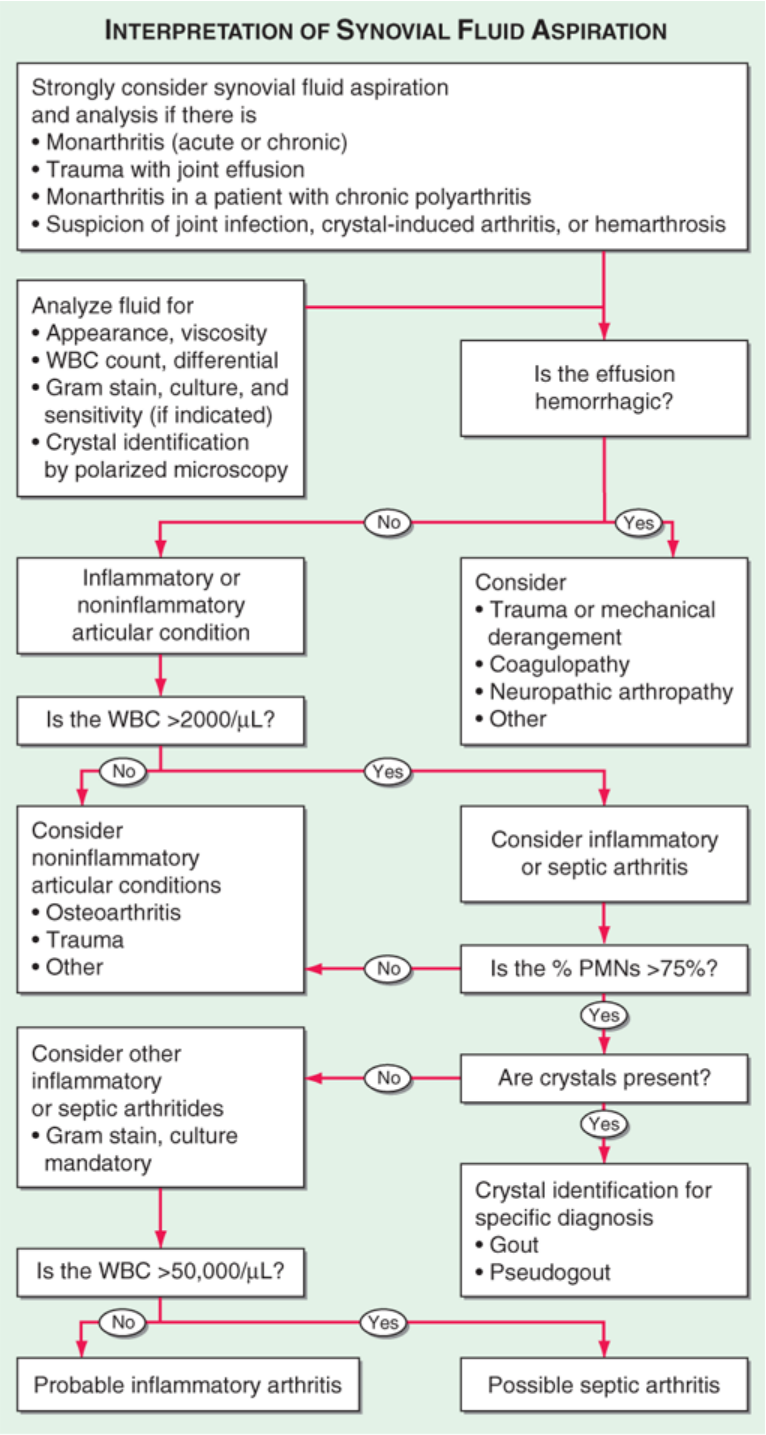
Abbreviations: ANA, antinuclear antibody; CREST, *c*alcinosis, *R*aynaud phenomenon, *e*sophageal involvement, *s*clerodactyly, and *t*elangiectasia; MCTD, mixed connective tissue disease; PBC, primary biliary cirrhosis; PSS, progressive systemic sclerosis; SCLE, subacute cutaneous lupus erythematosus; SLE, systemic lupus erythematosus.

Aspiration and analysis of synovial fluid are always indicated in acute monarthritis or when an infectious or crystal-induced arthropathy is suspected. Synovial fluid may distinguish between noninflammatory and inflammatory processes by analysis of the appearance, viscosity, and cell count. Tests for synovial fluid glucose, protein, lactate dehydrogenase, lactic acid, or autoantibodies are not recommended because they have no diagnostic value. Normal synovial fluid is clear or a pale straw color and is viscous, primarily because of the high levels of hyaluronate. Noninflammatory synovial fluid is clear, viscous, and amber-colored, with a WBC count of <2000/μL and a predominance of mononuclear cells. The viscosity of synovial fluid is assessed by expressing fluid from the syringe one drop at a time. Normally, there is a stringing effect, with a long tail behind each synovial drop. Effusions caused by OA or trauma will have normal viscosity. Inflammatory fluid is turbid and yellow, with an increased WBC count (2000–50,000/μL) and a polymorphonuclear leukocyte predominance. Inflammatory fluid has reduced viscosity (no stringing), diminished hyaluronate, and little or no tail following each drop of synovial fluid. Such effusions are found in RA, gout, and other inflammatory arthritides. Septic fluid is opaque and purulent, with a WBC count usually >50,000/μL, a predominance of polymorphonuclear leukocytes (>75%), and low viscosity. Such effusions are typical of septic arthritis but may also occur with RA or gout. In addition, hemorrhagic

synovial fluid may be seen with trauma, hemarthrosis, or neuropathic arthritis. An algorithm for synovial fluid aspiration and analysis is shown in Fig. 363-6. Synovial fluid should be analyzed immediately for appearance, viscosity, and cell count. Monosodium urate crystals (observed in gout) are seen by polarized microscopy and are long, needle-shaped, negatively birefringent, and usually intracellular. In chondrocalcinosis and pseudogout, calcium pyrophosphate dihydrate crystals are usually short, rhomboid-shaped, and positively birefringent. Whenever infection is suspected, synovial fluid should be Gram stained and cultured appropriately. If gonococcal arthritis is suspected, nucleic acid amplification tests should be used to detect either *Chlamydia trachomatis* or *N. gonorrhoeae* infection. Synovial fluid from patients with chronic monarthritis should also be cultured for *M. tuberculosis* and fungi. Last, it should be noted that crystal-induced arthritis and septic arthritis occasionally occur together in the same joint.

FIGURE 363-6

Algorithmic approach to the use and interpretation of synovial fluid aspiration and analysis. PMNs, polymorphonuclear (leukocytes); WBC, white blood cell count.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSTIC IMAGING IN JOINT DISEASES

Conventional radiography has been a valuable tool in the diagnosis and staging of articular disorders. Plain x-rays are most appropriate and cost effective when there is a history of trauma, suspected chronic infection, progressive disability, or monarticular involvement; when therapeutic alterations are considered; or when a baseline assessment is desired for what appears to be a chronic process. However, in acute inflammatory arthritis, early radiography is rarely helpful in establishing a diagnosis and may only reveal soft tissue swelling or juxtaarticular demineralization. As the disease progresses, calcification (of soft tissues, cartilage, or bone), joint space narrowing, erosions, bony ankylosis, new bone formation (sclerosis, osteophytes, or periostitis), or subchondral cysts may develop and suggest specific clinical entities. Consultation with a radiologist will help define the optimal imaging modality, technique, or positioning and prevent the need for further studies.

Additional imaging techniques may possess greater diagnostic sensitivity and facilitate early diagnosis in a limited number of articular disorders and in selected circumstances and are indicated when conventional radiography is inadequate or nondiagnostic **(Table 363-5)**. *Ultrasonography* is useful in the detection of soft tissue abnormalities, such as tendinitis, tenosynovitis, enthesitis, bursitis, and entrapment neuropathies. Wider use, lower cost, better technology, and enhanced site-specific transducers now allow for wider use in outpatient care, especially for the evaluation of synovial (Baker's) cysts, rotator cuff tears, tendinitis and tendon injury, and crystal deposition on cartilage. Use of power Doppler allows for early detection of synovitis and bony erosions. *Radionuclide scintigraphy* is a very sensitive, but poorly specific, means of detecting inflammatory or metabolic alterations in bone or periarticular soft tissue structures (Table 363-5). Scintigraphy is best suited for total-body assessment (extent and distribution) of skeletal involvement (neoplasia, Paget's disease) and the assessment of patients with undiagnosed polyarthralgias, looking for occult arthritis. The use of scintigraphy has declined with greater use and declining cost of ultrasound and MRI. MRI has largely replaced scintigraphy in diagnosing osseous infection, neoplasia, inflammation, increased blood flow, bone remodeling, heterotopic bone formation, or avascular necrosis. Gallium scanning uses ^{67}Ga, which binds serum and cellular transferrin and lactoferrin and is preferentially taken up by neutrophils, macrophages, bacteria, and tumor tissue (e.g., lymphoma). As such, it is primarily used in the identification of occult infection or malignancy. Scanning with ^{111}In-labeled WBCs has been used to detect osteomyelitis and infectious or inflammatory arthritis. Despite their utility, ^{111}In-labeled WBC or ^{67}Ga scanning has largely been replaced by MRI, except when there is a suspicion of septic joint or prosthetic joint infections.

TABLE 363-5

Diagnostic Imaging Techniques for Musculoskeletal Disorders

Method	Imaging Time, h	Cost[a]	Current Indications
Ultrasound	<1	++	Synovial (Baker's) cysts
			Rotator cuff tears Bursitis, tendinitis, tendon injury Enthesitis Carpal tunnel syndrome Urate or calcium pyrophosphate deposition on cartilage Early detection of synovial inflammation or erosions Ultrasound-guided injection/arthrocentesis
Radionuclide scintigraphy			Metastatic bone survey Evaluation of Paget's disease
^{99m}Tc	1–4	++	Identifying occult arthritis in patients with undiagnosed polyarthralgia
^{111}In-WBC	24	+++	Acute infection Prosthetic infection Acute osteomyelitis
^{67}Ga	24–48	++++	Acute and chronic infection Acute osteomyelitis
Computed tomography (CT)	<1	+++	Herniated intervertebral disk Sacroiliitis Spinal stenosis Spinal trauma Osteoid osteoma Stress fracture
Dual-energy CT	<1	NA	Uric acid deposition Tophus localization
Magnetic resonance imaging	1/2–2	++++	Avascular necrosis Osteomyelitis Septic arthritis, infected prosthetic joints Early sacroiliitis Intraarticular derangement and soft tissue injury Derangements of axial skeleton and spinal cord Herniated intervertebral disk Pigmented villonodular synovitis Inflammatory and metabolic muscle pathology

[a]Relative cost for imaging study.

Abbreviations: NA, not commercially available; WBC, white blood cell.

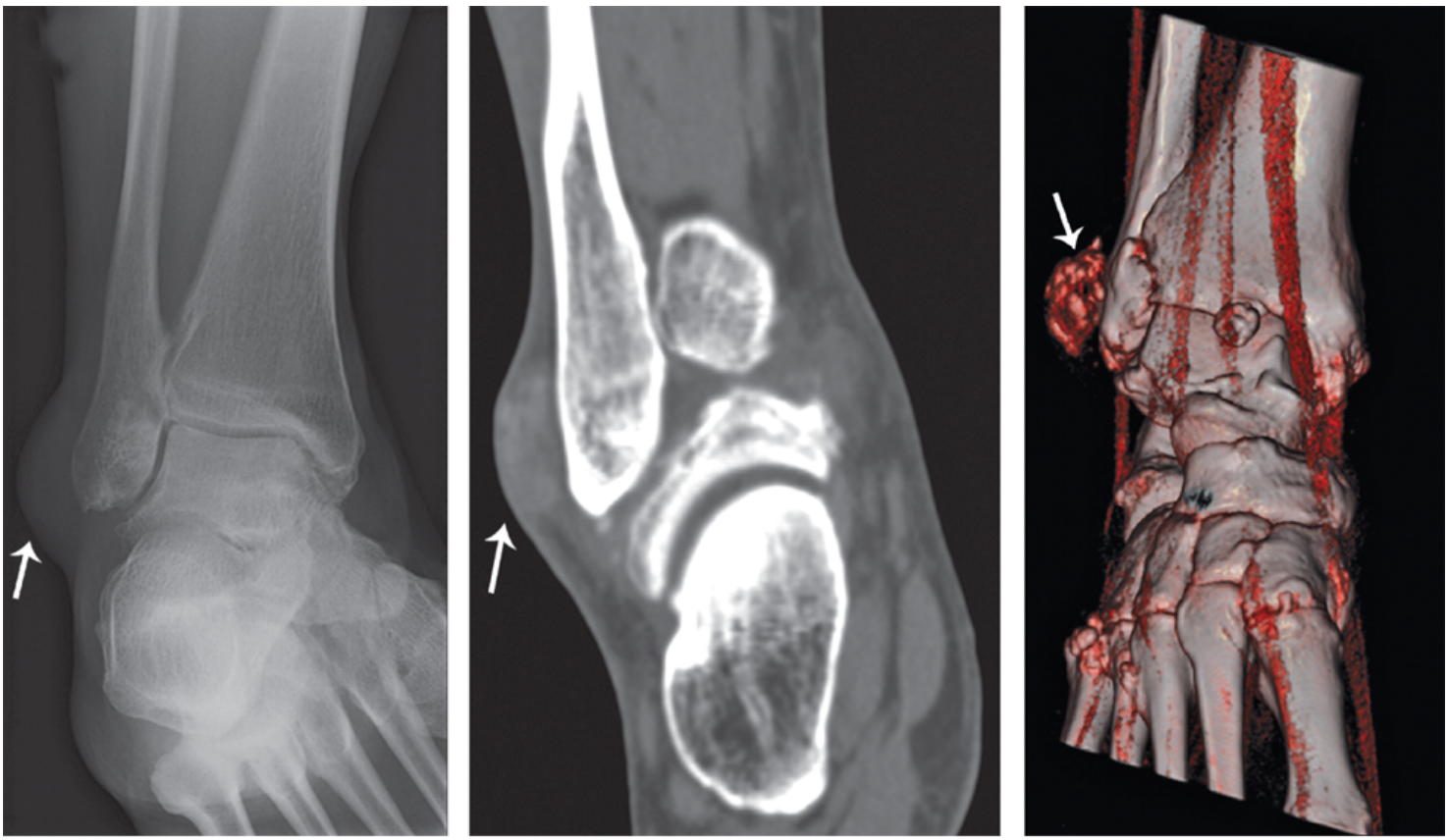
Computed tomography (*CT*) provides detailed visualization of the axial skeleton. Articulations previously considered difficult to visualize by radiography (e.g., zygapophyseal, sacroiliac, sternoclavicular, hip joints) can be effectively evaluated using CT. CT has been demonstrated to be useful in the diagnosis of low back pain syndromes (e.g., spinal stenosis vs herniated disk), sacroiliitis, osteoid osteoma, and stress fractures.

Helical or spiral CT (with or without contrast angiography) is a novel technique that is rapid, cost effective, and sensitive in diagnosing pulmonary embolism or obscure fractures, often in the setting of initially equivocal findings. High-resolution CT can be advocated in the evaluation of suspected or established infiltrative lung disease (e.g., scleroderma or rheumatoid lung). The recent use of hybrid (positron emission tomography [PET] or single-photon emission CT [SPECT]) CT scans in metastatic evaluations has incorporated CT to provide better anatomic localization of scintigraphic abnormalities.

^{18}F-Fluorodeoxyglucose (FDG) is the most commonly used radiopharmaceutical in PET scanning. FDG-PET/CT scans have been seldom used in the evaluation of septic or inflammatory arthritis, but have also been useful in the evaluation of patients with fever of unknown origin or suspected large vessel vasculitis. Dual-energy CT (DECT) scanning, developed in urology to identify urinary calculi, has been a highly sensitive and specific method used to identify and quantify uric acid deposition in tissues **(Fig. 363-7)**.

FIGURE 363-7

Dual-energy computed tomography (DECT) scan from a 45-year-old woman with right ankle swelling around the lateral malleolus. Three-dimensional volume-rendered coronal reformatted DECT image shows that the mass is composed of monosodium urate (*red*) in keeping with tophus (*arrow*). *(Used with permission from S Nicolaou et al: AJR 194:1072, 2010.)*

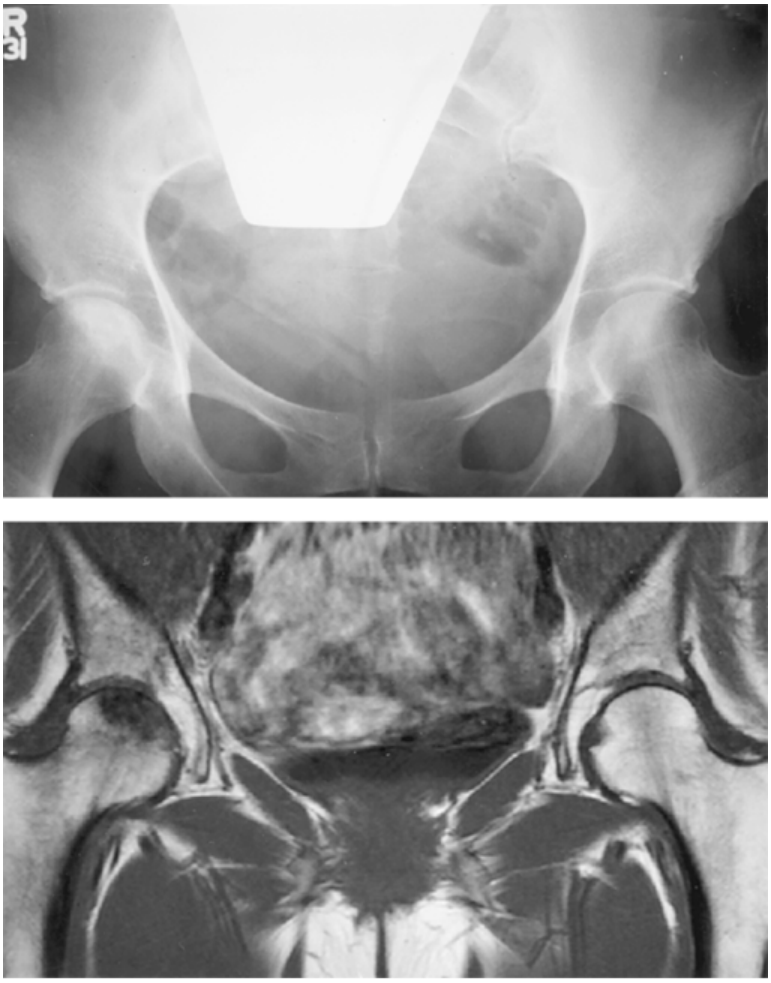


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MRI has significantly advanced the ability to image musculoskeletal structures. MRI has the advantages of providing multiplanar images with fine anatomic detail and contrast resolution **(Fig. 363-8)** that allows for the superior ability to visualize bone marrow and soft tissue periarticular structures. Although more costly with a longer procedural time than CT, the MRI has become the preferred technique when evaluating complex musculoskeletal disorders.

FIGURE 363-8

Superior sensitivity of magnetic resonance imaging (MRI) in the diagnosis of osteonecrosis of the femoral head. A 45-year-old woman receiving high-dose glucocorticoids developed right hip pain. Conventional x-rays (*top*) demonstrated only mild sclerosis of the right femoral head. T1-weighted MRI (*bottom*) demonstrated low-density signal in the right femoral head, diagnostic of osteonecrosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MRI can image fascia, vessels, nerve, muscle, cartilage, ligaments, tendons, pannus, synovial effusions, and bone marrow. Visualization of particular structures can be enhanced by altering the pulse sequence to produce either T1- or T2-weighted spin echo, gradient echo, or inversion recovery (including short tau inversion recovery [STIR]) images. Because of its sensitivity to changes in marrow fat, MRI is a sensitive but nonspecific means of detecting osteonecrosis, osteomyelitis, and marrow inflammation indicating overlying synovitis or osteitis (Fig. 363-8). Because of its enhanced soft tissue resolution, MRI is more sensitive than arthrography or CT in the diagnosis of soft tissue injuries (e.g., meniscal and rotator cuff tears); intraarticular derangements; marrow abnormalities (osteonecrosis, myeloma); and spinal cord or nerve root damage, synovitis, or cartilage damage or loss.

ACKNOWLEDGMENT

The author acknowledges the insightful contributions of Dr. Peter E. Lipsky to this chapter in previous editions.

FURTHER READING

Bhagat M et al: Utility of immunologic testing in suspected rheumatologic disease. Curr Allergy Asthma Rep 14:405, 2014. [PubMed: 24370946]

Cozzani E, Drosera M: Serology of lupus erythematosus: Correlation between immunopathological features and clinical aspects. Autoimmune Dis 2014:321359, 2014.

Cush JJ et al: Evaluation of musculoskeletal complaints. Available from *http://www.rheumaknowledgy.com/evaluation-of-musculoskeletal-complaints*. Accessed April 6, 2017.

DeFroda SF et al: Differentiating radiculopathy from lower extremity arthropathy. Am J Med 129:1124. e1, 2016.

Rudwaleit M et al: How to diagnose axial spondyloarthritis early. Ann Rheum Dis 63:535, 2004. [PubMed: 15082484]

Simpfendorfer CS: Radiologic approach to musculoskeletal infections. Infect Dis Clin North Am 31:299, 2017. [PubMed: 28366223]

Chapter 364: Osteoarthritis

David T. Felson; Tuhina Neogi

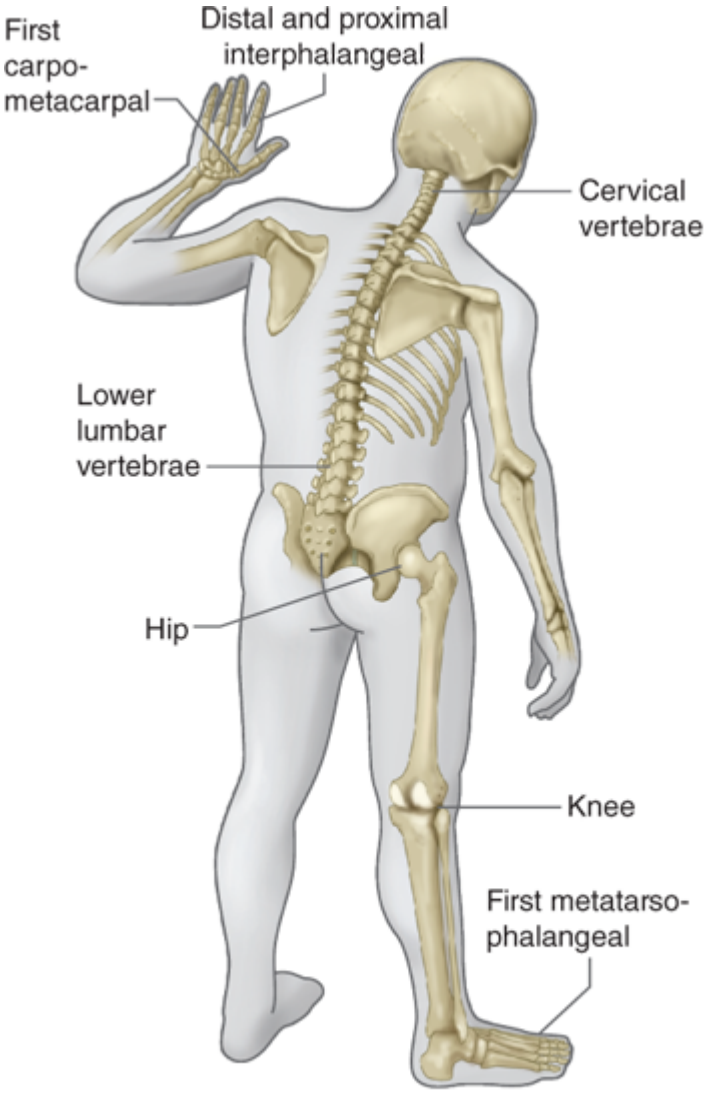
INTRODUCTION

Osteoarthritis (OA) is the most common type of arthritis. Its high prevalence, especially in the elderly, and its negative impact on physical function make it a leading cause of disability in the elderly. Because of the aging of Western populations and because obesity, a major risk factor, is increasing in prevalence, the occurrence of OA is on the rise.

OA affects certain joints, yet spares others **(Fig. 364-1)**. Commonly affected joints include the hip, knee, and first metatarsal phalangeal joint (MTP) and cervical and lumbosacral spine. In the hands, the distal and proximal interphalangeal joints and the base of the thumb are often affected. Usually spared are the wrist, elbow, and ankle. Our joints were designed, in an evolutionary sense, for brachiating apes, animals that still walked on four limbs. We thus develop OA in joints that were ill designed for human tasks such as pincer grip (OA in the thumb base) and walking upright (OA in knees and hips). Some joints, like the ankles, may be spared because their articular cartilage may be uniquely resistant to loading stresses.

FIGURE 364-1

Joints commonly affected by osteoarthritis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

OA can be diagnosed based on structural abnormalities or on the symptoms these abnormalities evoke. According to cadaveric studies, by elderly years, structural changes of OA are nearly universal. These include cartilage loss (seen as joint space loss on x-rays) and osteophytes. Many persons with x-ray evidence of OA have no joint symptoms, and although the prevalence of structural abnormalities is of interest in understanding disease pathogenesis, what matters more from a clinical perspective is the prevalence of symptomatic OA. Symptoms, usually joint pain, determine disability, visits to clinicians, and disease costs.

Symptomatic OA of the knee (pain on most days of a recent month plus x-ray evidence of OA in that knee) occurs in ~12% of persons age ≥60 in the United States and 6% of all adults age ≥30. Symptomatic hip OA is roughly one-third as common as disease in the knee. Although radiographic hand OA and the appearance of bony enlargement in affected hand joints (Fig. 364-2) are extremely common in older persons, most cases are often not symptomatic. Even so, symptomatic hand OA occurs in ~10% of elderly individuals and often produces measurable limitation in function.

FIGURE 364-2

Severe osteoarthritis of the hands affecting the distal interphalangeal joints (Heberden's nodes) and the proximal interphalangeal joints (Bouchard's nodes). There is no clear bony enlargement of the other common site in the hands, the thumb base.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The prevalence of OA rises strikingly with age, being uncommon in adults aged <40 and highly prevalent in those aged >60. It is also a disease that, at least in middle-aged and elderly persons, is much more common in women than in men.

X-ray evidence of OA is common in the lower back and neck, but back pain and neck pain have not been tied to findings of OA on x-ray. Thus, back pain and neck pain are treated separately **(Chap. 14)**.

DEFINITION

OA is joint failure, a disease in which all structures of the joint have undergone pathologic change, often in concert. The pathologic sine qua non of disease is hyaline articular cartilage loss, present in a focal and, initially, nonuniform manner. This is accompanied by increasing thickness and sclerosis of the subchondral bony plate, by outgrowth of osteophytes at the joint margin, by stretching of the articular capsule, by variable degrees of synovitis, and by weakness of muscles bridging the joint. In knees, meniscal degeneration is part of the disease. There are numerous pathways that lead to joint failure, but the initial step is often joint injury in the setting of a failure of protective mechanisms.

JOINT PROTECTIVE MECHANISMS AND THEIR FAILURE

Joint protectors include joint capsule and ligaments, muscle, sensory afferents, and underlying bone. Joint capsule and ligaments serve as joint protectors by providing a limit to excursion, thereby fixing the range of

joint motion.

Synovial fluid reduces friction between articulating cartilage surfaces, thereby serving as a protector against friction-induced cartilage wear. This lubrication function depends on *hyaluronic acid* and on *lubricin*, a mucinous glycoprotein secreted by synovial fibroblasts whose concentration diminishes after joint injury and in the face of synovial inflammation.

The ligaments, along with overlying skin and tendons, contain mechanoreceptor sensory nerves. These mechanoreceptors fire at different frequencies throughout a joint's range of motion, providing feedback by way of the spinal cord to muscles and tendons. As a consequence, these muscles and tendons can assume the right tension at appropriate points in joint excursion to act as optimal joint protectors, anticipating joint loading.

Muscles and tendons that bridge the joint are key joint protectors. Focal stress across the joint is minimized by muscle contraction that decelerates the joint before impact and assures that when joint impact arrives, it is distributed broadly across the joint surface.

Failure of these joint protectors increases the risk of joint injury and OA. For example, in animals, OA develops rapidly when a sensory nerve to the joint is sectioned and joint injury induced. Similarly, in humans, Charcot's arthropathy, a severe and rapidly progressive OA, develops when minor joint injury occurs in the presence of posterior column peripheral neuropathy. Another example of joint protector failure is rupture of ligaments, a well-known cause of the early development of OA.

CARTILAGE AND ITS ROLE IN JOINT FAILURE

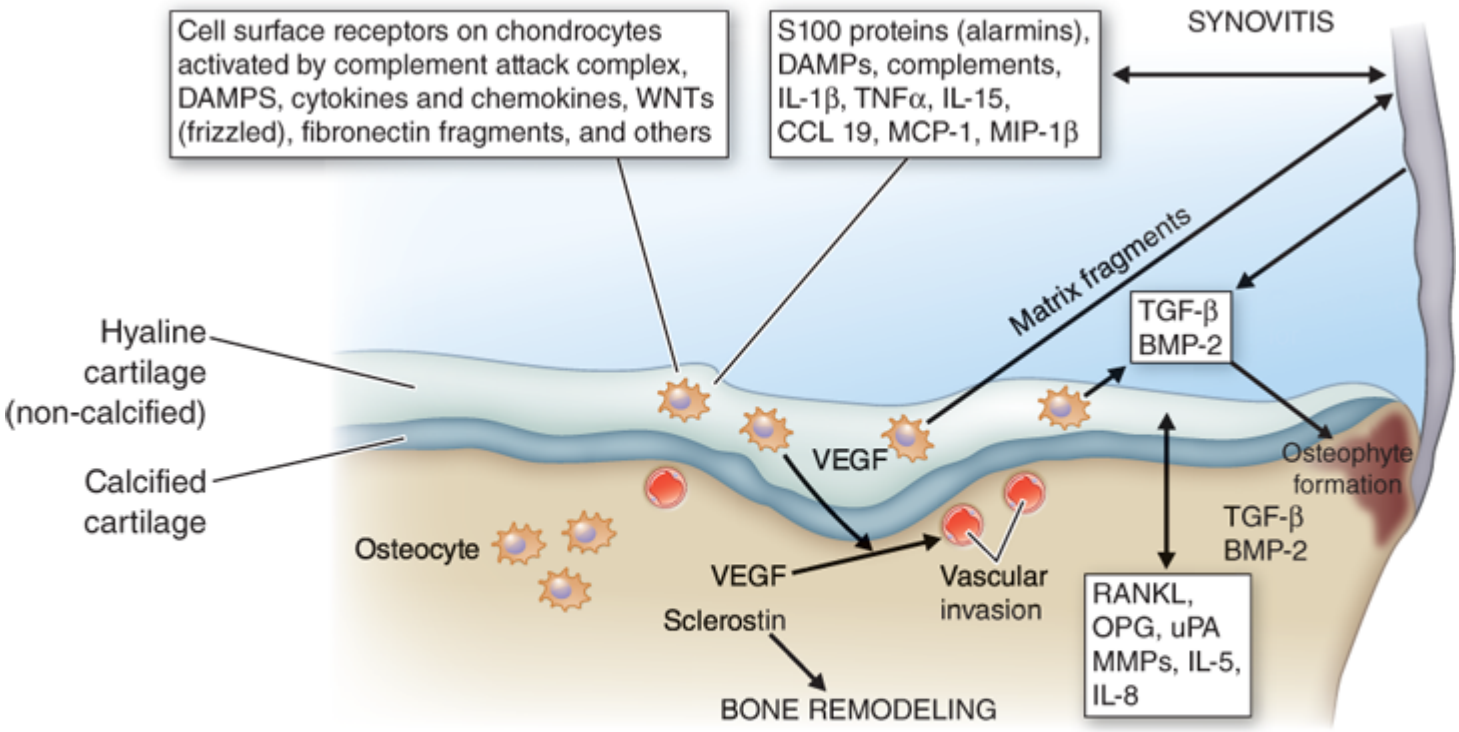
In addition to being a primary target tissue for disease, cartilage also functions as a joint protector. A thin rim of tissue at the ends of two opposing bones, cartilage is lubricated by synovial fluid to provide an almost frictionless surface across which these two bones move. The compressible stiffness of cartilage compared to bone provides the joint with impact-absorbing capacity.

The earliest changes of OA may occur in cartilage, and abnormalities there can accelerate disease development. The two major macromolecules in cartilage are type 2 collagen, which provides cartilage its tensile strength, and aggrecan, a proteoglycan macromolecule linked with hyaluronic acid, which consists of highly negatively charged glycosaminoglycans. In normal cartilage, type 2 collagen is woven tightly, constraining the aggrecan molecules in the interstices between collagen strands, forcing these highly negatively charged molecules into close proximity with one another. The aggrecan molecule, through electrostatic repulsion of its negative charges, gives cartilage its compressive stiffness. Chondrocytes, the cells within this avascular tissue, synthesize all elements of the matrix and produce enzymes that break down the matrix. Synovium and chondrocytes synthesize and release cytokines and growth factors, which provide feedback that modulates synthesis of matrix molecules **(Fig. 364-3)**. Cartilage matrix synthesis and catabolism are in a dynamic equilibrium influenced by the cytokine and growth factor environment. Mechanical and osmotic stress on chondrocytes induces these cells to alter gene expression and increase

production of inflammatory cytokines and matrix-degrading enzymes. While chondrocytes synthesize numerous enzymes, matrix metalloproteinases (MMP) (especially collagenases and ADAMTS-5) are critical enzymes in the breakdown of cartilage matrix. Both collagenase and aggrecanases act primarily in the territorial matrix surrounding chondrocytes; however, as the osteoarthritic process develops, their activities and effects spread throughout the matrix, especially in the superficial layers of cartilage.

FIGURE 364-3

Selected factors involved in the osteoarthritic process including chondrocytes, bone, and synovium. Synovitis causes release of cytokines, alarmins, damage-associated molecular pattern (DAMP) molecules, and complement, which activate chondrocytes through cell surface receptors. Chondrocytes produce matrix molecules (collagen type 2, aggrecan) and the enzymes responsible for the degradation of the matrix (e.g., ADAMTS-5 and matrix metalloproteinases [MMPs]). Bone invasion occurs through the calcified cartilage, triggered by vascular endothelial growth factor (VEGF) and other molecules. IL, interleukin; TGF, transforming growth factor; TNF, tumor necrosis factor. *(From RF Loeser et al: Arthritis Rheum 64:1697, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The synovium, cartilage, and bone all influence disease development through cytokines, chemokines, and even complement activation (Fig. 364-3). These act on chondrocyte cell surface receptors and ultimately have transcriptional effects. Matrix fragments released from cartilage stimulate synovitis. Inflammatory cytokines such as interleukin 1β (IL-1β) and tumor necrosis factor α (TNF-α) induce chondrocytes to synthesize prostaglandin E_2 and nitric oxide, which have complex effects on matrix synthesis and degradation. At early stages in the matrix response to injury the net effect of cytokine stimulation may be matrix synthesis, but ultimately, the combination of effects on chondrocytes triggers matrix degradation. Enzymes in the matrix are held in check by activation inhibitors, including tissue inhibitor of

metalloproteinase (TIMP). Growth factors are also part of this complex network, with Bone Morphogenetic Protein 2 (BMP-2) and transforming growth factor β (TGF-β) playing prominent roles in stimulating the development of osteophytes. Whereas healthy articular cartilage is avascular in part due to angiogenesis inhibitors present in cartilage, disease is characterized by the invasion of blood vessels into cartilage from underlying bone. This is influenced by vascular endothelial growth factor (VEGF) synthesis in the cartilage and bone. With these blood vessels come nerves that may bring nociceptive innervation.

Probably as a result of chronic oxidative damage, articular chondrocytes exhibit an age-related decline in synthetic capacity while maintaining the ability to produce proinflammatory mediators and matrix-degrading enzymes, findings characteristic of a senescent secretory phenotype. These chondrocytes are unable to maintain tissue homeostasis (such as after insults of a mechanical or inflammatory nature). Thus, with age, cartilage is easily damaged by minor sometimes unnoticed injuries, including those that are part of daily activities.

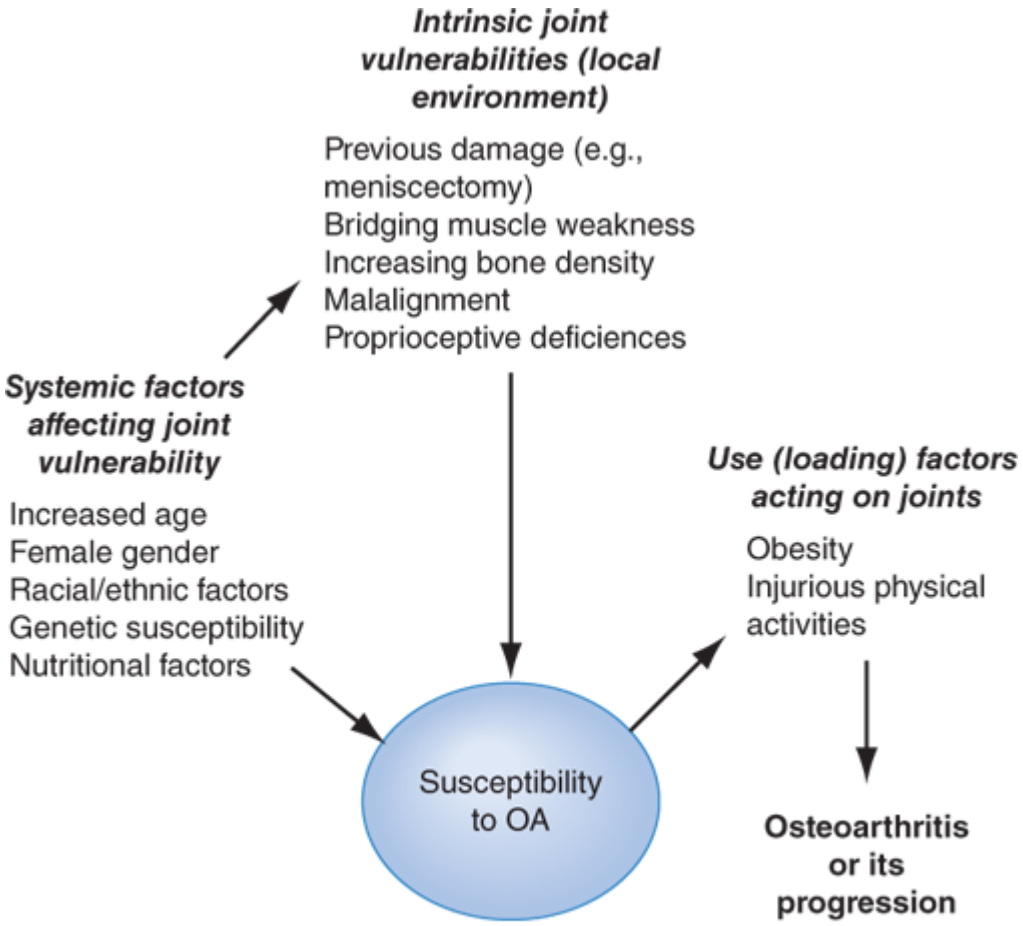
OA cartilage is characterized by gradual depletion of aggrecan, an unfurling of the tightly woven collagen matrix, and loss of type 2 collagen. With these changes comes increasing vulnerability of cartilage, which loses its compressive stiffness.

RISK FACTORS

Joint vulnerability and joint loading are the two major factors contributing to the development of OA. On the one hand, a vulnerable joint whose protectors are dysfunctional can develop OA with minimal levels of loading, perhaps even levels encountered during everyday activities. On the other hand, in a young joint with competent protectors, a major acute injury or long-term overloading is necessary to precipitate disease. Risk factors for OA can be understood in terms of their effect either on joint vulnerability or on loading **(Fig. 364-4)**.

FIGURE 364-4

Risk factors for osteoarthritis (OA) either contribute to the susceptibility of the joint (systemic factors or factors in the local joint environment) or increase risk by the load they put on the joint. Usually a combination of loading and susceptibility factors is required to cause disease or its progression.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SYSTEMIC RISK FACTORS

Age is the most potent risk factor for OA. Radiographic evidence of OA is rare in individuals aged <40; however, in some joints, such as the hands, OA occurs in >50% of persons aged >70. Aging increases joint vulnerability through several mechanisms. Whereas dynamic loading of joints stimulates cartilage matrix synthesis by chondrocytes in young cartilage, aged cartilage is less responsive to these stimuli. Partly because of this failure to synthesize matrix with loading, cartilage thins with age, and thinner cartilage experiences higher shear stress and is at greater risk of cartilage damage. Also, joint protectors fail more often with age. Muscles that bridge the joint become weaker with age and also respond less quickly to oncoming impulses. Sensory nerve input slows with age, retarding the feedback loop of mechanoreceptors to muscles and tendons related to their tension and position. Ligaments stretch with age, making them less able to absorb impulses. These factors work in concert to increase the vulnerability of older joints to OA.

Older women are at high risk of OA in all joints, a risk that emerges as women reach their sixth decade. Although hormone loss with menopause may contribute to this risk, there is little understanding of the unique vulnerability of older women versus men to OA.

HERITABILITY AND GENETICS


OA is a highly heritable disease, but its heritability is joint specific. Fifty percent of the hand and hip OA in the community is attributable to inheritance, that is, to disease present in other members of the family. However, the heritable proportion of knee OA is at most 30%, with some studies suggesting no heritability at all. Whereas many people with OA have disease in multiple joints, this “generalized OA” phenotype is rarely inherited and is more often a consequence of aging.

Emerging evidence has identified genetic mutations that confer a high risk of OA, the best replicated is a polymorphism within the growth differentiation factor 5 (GDF5) gene. This polymorphism diminishes the quantity of GDF5; GDF5 has its main influence on joint shape which is likely to be the mechanism by which genes predisposing to OA increase risk of disease.

GLOBAL CONSIDERATIONS


With the aging of the populations, both the prevalence of OA and the amount of disability worldwide related to OA have been increasing especially in developed countries where many are living into old age. Hip OA is rare in China and in immigrants from China to the United States. However, OA in the knees is at least as common, if not more so, in Chinese than in whites from the United States, and knee OA represents a major cause of disability in China, especially in rural areas. Anatomic differences between Chinese and white hips may account for much of the difference in hip OA prevalence, with white hips having a higher prevalence of anatomic predispositions to the development of OA.

RISK FACTORS IN THE JOINT ENVIRONMENT

Some risk factors increase vulnerability of the joint through local effects on the joint environment. With changes in joint anatomy, for example, load across the joint is no longer distributed evenly across the joint surface, but rather shows an increase in focal stress. In the hip, three uncommon developmental abnormalities occurring in utero or in childhood, congenital dysplasia, Legg-Perthes disease, and slipped capital femoral epiphysis, leave a child with distortions of hip joint anatomy that often lead to OA later in life. Girls are predominantly affected by acetabular dysplasia, a mild form of congenital dislocation, whereas the other abnormalities more often affect boys. Depending on the severity of the anatomic abnormalities, hip OA occurs either in young adulthood (severe abnormalities) or middle age (mild abnormalities). Femoroacetabular impingement can develop during adolescence. It is a clinical syndrome in which anatomic abnormalities of the femoral head and/or the acetabulum result in abnormal contact between the two bones especially during hip flexion and rotation, leading to cartilage and labral damage and hip pain and ultimately in later life to possible hip OA.

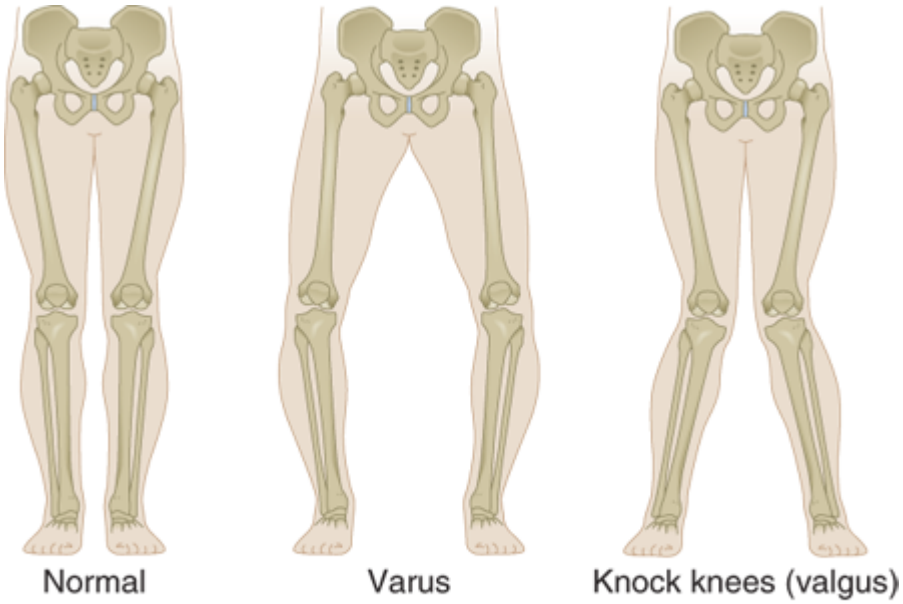
Major injuries to a joint also can produce anatomic abnormalities that leave the joint susceptible to OA. For example, a fracture through the joint surface often causes OA in joints in which the disease is otherwise rare such as the ankle and the wrist. Avascular necrosis can lead to collapse of dead bone at the articular surface, producing anatomic irregularities and subsequent OA.

Tears of ligamentous and fibrocartilaginous structures that protect the joints, such as the meniscus in the knee and the labrum in the hip, can lead to premature OA. Meniscal tears increase with age and when chronic are often asymptomatic but lead to adjacent cartilage damage and accelerated OA. Even injuries in which the affected person never received a diagnosis may increase risk of OA. For example, in the Framingham Study subjects, men with a history of major knee injury, but no surgery, had a 3.5-fold increased risk for subsequent knee OA.

Another source of anatomic abnormality is malalignment across the joint **(Fig. 364-5)**. This factor has been best studied in the knee, which is the fulcrum of the longest lever arm in the body. Varus (bowlegged) knees with OA are at exceedingly high risk of cartilage loss in the medial or inner compartment of the knee, whereas valgus (knock-kneed) malalignment predisposes to rapid cartilage loss in the lateral compartment. Malalignment causes this effect by increasing stress on a focal area of cartilage, which then breaks down. There is evidence that malalignment in the knee not only causes cartilage loss but leads to underlying bone damage, producing bone marrow lesions seen on magnetic resonance imaging (MRI). Malalignment in the knee often produces such a substantial increase in focal stress within the knee (as evidenced by its destructive effects on subchondral bone) that severely malaligned knees may be destined to progress regardless of the status of other risk factors.

FIGURE 364-5

The two types of limb malalignment in the frontal plane: varus, in which the stress is placed across the medial compartment of the knee joint, and valgus, which places excess stress across the lateral compartment of the knee.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Weakness in the quadriceps muscles bridging the knee increases the risk of the development of painful OA in the knee.

The role of bone in serving as a shock absorber for impact load is not well understood, but persons with increased bone density are at high risk of OA, suggesting that the resistance of bone to impact during joint use may play a role in disease development.

LOADING FACTORS

Obesity

Three to six times body weight is transmitted across the knee during single-leg stance. Any increase in weight may be multiplied by this factor to reveal the excess force across the knee in overweight persons during walking. Obesity is a well-recognized and potent risk factor for the development of knee OA and, less so, for hip OA. Obesity precedes the development of disease and is not just a consequence of the inactivity present in those with disease. It is a stronger risk factor for disease in women than in men, and in women, the relationship of weight to the risk of disease is linear, so that with each pound increase in weight, there is a commensurate increase in risk. Weight loss in women lowers the risk of developing symptomatic disease. Not only is obesity a risk factor for OA in weight-bearing joints, but obese persons have more severe symptoms from the disease.

Obesity's effect on the development and progression of disease is mediated mostly through the increased loading in weight-bearing joints that occurs in overweight persons. However, a modest association of obesity with an increased risk of hand OA suggests that systemic products of adipose tissue such as adipokines may affect disease risk also.

Repeated Use of Joint and Exercise

There are two categories of repetitive joint use, occupational use and leisure time physical activities. Workers performing repetitive tasks as part of their occupations for many years are at high risk of developing OA in the joints they use repeatedly. For example, farmers are at high risk for hip OA, and miners have high rates of OA in knees and spine. Workers whose jobs require regular knee bending or lifting or carrying heavy loads have a high rate of knee OA. One reason why workers may get disease is that during long days at work, their muscles may gradually become exhausted, no longer serving as effective joint protectors.

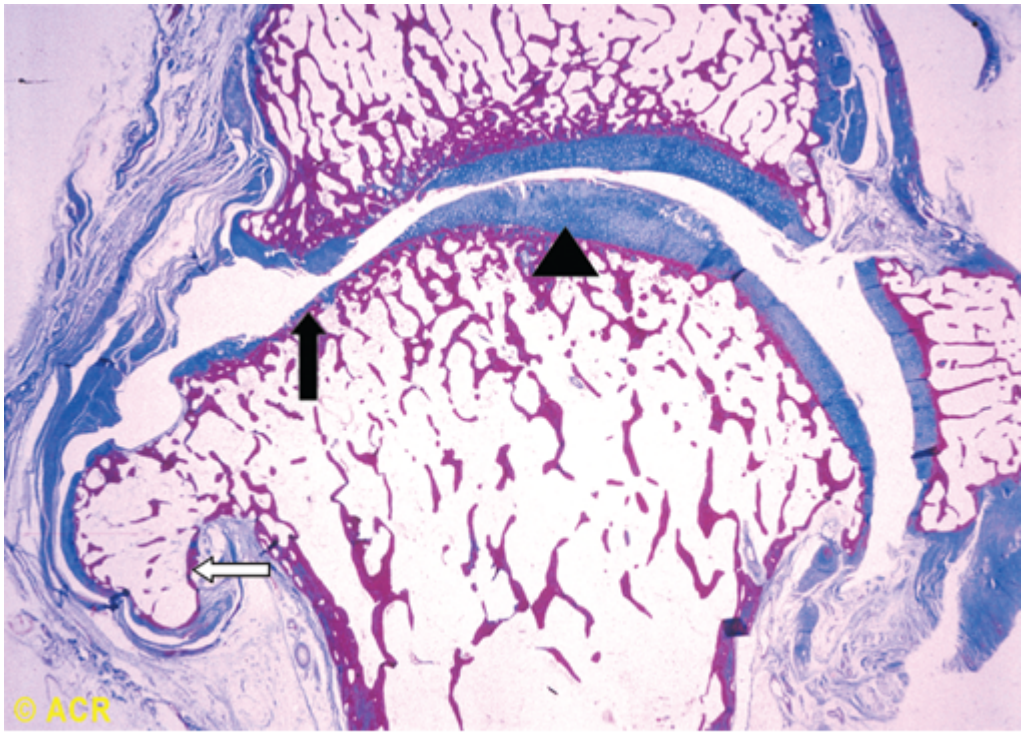
It is widely recommended for people to adopt an exercise-filled lifestyle, and long-term studies of exercise suggest no consistent association of exercise with OA risk in the majority of persons. However, persons who already have injured joints may put themselves at greater risk by engaging in certain types of exercise. For example, persons who have already sustained major knee injuries are at increased risk of progressive knee OA as a consequence of running. In addition, compared to nonrunners, elite runners (professional runners and those on Olympic teams) have high risks of both knee and hip OA. Lastly, although recreational runners are not at increased risk of knee OA, studies suggest that they have a modest increased risk of disease in the hip.

PATHOLOGY

The pathology of OA provides evidence of the involvement of many joint structures in disease. Cartilage initially shows surface fibrillation and irregularity. As disease progresses, focal erosions develop there, and these eventually extend down to the subjacent bone. With further progression, cartilage erosion down to bone expands to involve a larger proportion of the joint surface, even though OA remains a focal disease with nonuniform loss of cartilage **(Fig. 364-6)**.

FIGURE 364-6

Pathologic changes of osteoarthritis in a toe joint. Note the nonuniform loss of cartilage (*arrowhead* vs *solid arrow*), the increased thickness of the subchondral bone envelope (*solid arrow*), and the osteophyte (*open arrow*). *(© 2018 American College of Rheumatology. Used with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

After an injury to cartilage, chondrocytes undergo mitosis and clustering. Although the metabolic activity of these chondrocyte clusters is high, the net effect of this activity is to promote proteoglycan depletion in the matrix surrounding the chondrocytes. This is because the catabolic activity is greater than the synthetic activity. As disease develops, collagen matrix becomes damaged, the negative charges of proteoglycans get exposed, and cartilage swells from ionic attraction to water molecules. Because in damaged cartilage proteoglycans are no longer forced into close proximity, cartilage does not bounce back after loading as it did when healthy, and cartilage becomes vulnerable to further injury. Chondrocytes at the basal level of cartilage undergo apoptosis.

With loss of cartilage comes alteration in subchondral bone. Stimulated by growth factors and cytokines, osteoclasts and osteoblasts in the subchondral bony plate, just underneath cartilage, become activated. Bone formation produces a thickening and stiffness of the subchondral plate that occurs even before cartilage ulcerates. Trauma to bone during joint loading may be the primary factor driving this bone

response, with healing from injury (including microcracks) producing stiffness. Small areas of osteonecrosis usually exist in joints with advanced disease. Bone death may also be caused by bone trauma with shearing of microvasculature, leading to a cutoff of vascular supply to some bone areas.

At the margin of the joint, near areas of cartilage loss, osteophytes form. These start as outgrowths of new cartilage, and with neurovascular invasion from the bone, this cartilage ossifies. Osteophytes are an important radiographic hallmark of OA.

The synovium produces lubricating fluids that minimize shear stress during motion. In healthy joints, the synovium consists of a single discontinuous layer filled with fat and containing two types of cells, macrophages and fibroblasts, but in OA, it can sometimes become edematous and inflamed. There is a migration of macrophages from the periphery into the tissue, and cells lining the synovium proliferate. Inflammatory cytokines and alarmins secreted by the synovium activate chondrocytes to produce enzymes which accelerate destruction of matrix.

Additional pathologic changes occur in the capsule, which stretches, becomes edematous, and can become fibrotic.

The pathology of OA is not identical across joints. In hand joints with severe OA, for example, there are often cartilage erosions in the center of the joint probably produced by bony pressure from the opposite side of the joint.

Basic calcium phosphate and calcium pyrophosphate dihydrate crystals are present microscopically in most joints with end-stage OA. Their role in osteoarthritic cartilage is unclear, but their release from cartilage into the joint space and joint fluid likely triggers synovial inflammation, which can, in turn, produce release of cytokines and trigger nociceptive stimulation.

SOURCES OF PAIN

Because cartilage is aneural, cartilage loss in a joint is not accompanied by pain. Thus, pain in OA likely arises from structures outside the cartilage. Innervated structures in the joint include the synovium, ligaments, joint capsule, muscles, and subchondral bone. Most of these are not visualized by the x-ray, and the severity of x-ray changes in OA correlates poorly with pain severity. However, in later stages of OA, loss of cartilage integrity that is accompanied by neurovascular invasion may contribute to pain.

Based on MRI studies in osteoarthritic knees comparing those with and without pain and on studies mapping tenderness in unanesthetized joints, likely sources of pain include synovial inflammation, joint effusions, and bone marrow edema. Modest synovitis develops in many but not all osteoarthritic joints. The presence of synovitis on MRI is correlated with the presence and severity of knee pain. Capsular stretching from fluid in the joint stimulates nociceptive fibers there, inducing pain. Increased focal loading as part of the disease not only damages cartilage but probably also injures the underlying bone. As a consequence, bone marrow

edema appears on the MRI; histologically, this edema signals the presence of microcracks and scar, which are the consequences of trauma. These lesions may stimulate bone nociceptive fibers.

Pain may arise from outside the joint also, including bursae near the joints. Common sources of pain near the knee are anserine bursitis and iliotibial band syndrome.

The pathologic changes of OA may eventually lead to alterations in nervous system signaling. Specifically, peripheral nociceptors can become more responsive to sensory input, known as peripheral sensitization, and there can also be an increase in central ascending nociceptive pathway activity, known as central sensitization. Individuals with OA may also have insufficient descending inhibitory modulation. Some individuals may be genetically predisposed to becoming sensitized; however, regardless of the etiology, these nervous system alterations are associated with more severe pain severity, and may contribute to the presence of allodynia and hyperalgesia in patients with OA.

CLINICAL FEATURES

Joint pain from OA is primarily activity-related in the early stages of the disease. Pain comes on either during or just after joint use and then gradually resolves. Examples include knee or hip pain with going up or down stairs, pain in weight-bearing joints when walking, and, for hand OA, pain when cooking. Early in disease, pain is episodic, triggered often by overactive use of a diseased joint, such as a person with knee OA taking a long run and noticing a few days of pain thereafter. As disease progresses, the pain becomes continuous and even begins to be bothersome at night. Stiffness of the affected joint may be prominent, but morning stiffness is usually brief (<30 min).

In knees, buckling may occur, in part, from weakness of muscles crossing the joint. Mechanical symptoms, such as buckling, catching, or locking, could also signify internal derangement, like an anterior cruciate ligament or meniscal tear; however, these symptoms, which are common in persons with knee OA need to be further evaluated only if they develop after an acute knee injury. In the knee, pain with activities requiring knee flexion, such as stair climbing and arising from a chair, often emanates from the patellofemoral compartment of the knee, which does not actively articulate until the knee is bent ~35°.

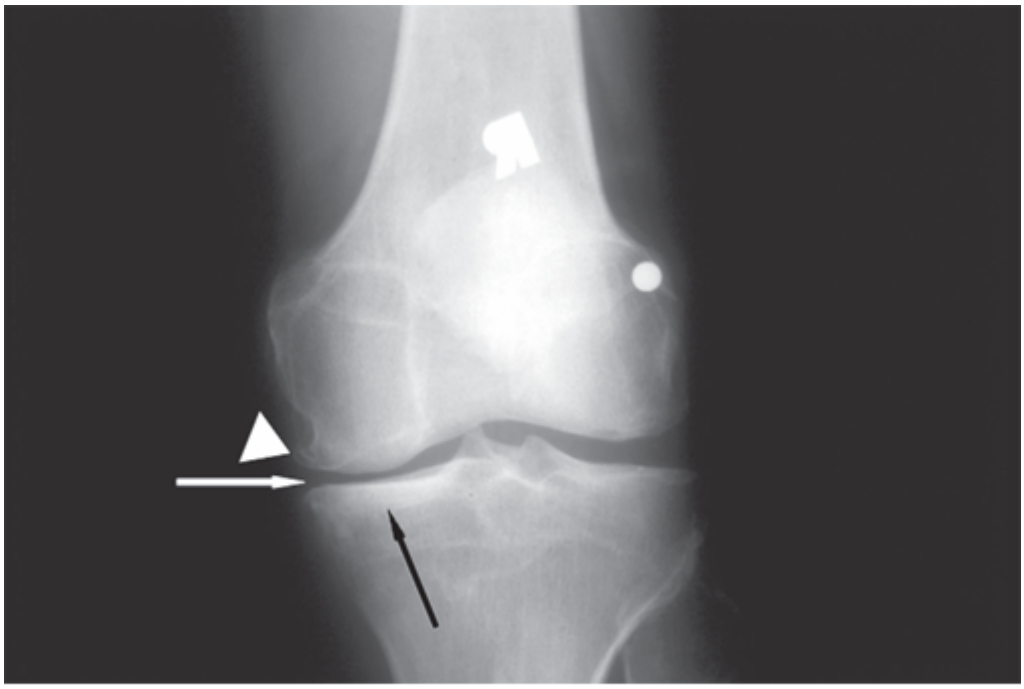
OA is the most common cause of chronic knee pain in persons aged >45, but the differential diagnosis is long. Inflammatory arthritis is likely if there is prolonged morning stiffness and many other joints are affected. Bursitis occurs commonly around knees and hips. A physical examination should focus on whether tenderness is over the joint line (at the junction of the two bones around which the joint is articulating) or outside of it. Anserine bursitis, medial and distal to the knee, is an extremely common cause of chronic knee pain that may respond to a glucocorticoid injection. Prominent nocturnal pain in the absence of end-stage OA merits a distinct workup. For hip pain, OA can be detected by loss of internal rotation on passive movement, and pain isolated to an area lateral to the hip joint usually reflects the presence of trochanteric bursitis.

No blood tests are routinely indicated for workup of patients with OA unless symptoms and signs suggest inflammatory arthritis. Examination of the synovial fluid is often more helpful diagnostically than an x-ray. If the synovial fluid white count is >1000/μL, inflammatory arthritis or gout or pseudogout is likely, the latter two being also identified by the presence of crystals.

X-rays are indicated to evaluate the possibility of OA only when joint pain and physical findings are not typical of OA or if pain persists after inauguration of treatment effective for OA. In OA, radiographic findings (Fig. 364-7) correlate poorly with the presence and severity of pain. Further, in both knees and hips, radiographs may be normal in early disease as they are insensitive to cartilage loss and other early findings.

FIGURE 364-7

X-ray of knee with medial osteoarthritis. Note the narrowed joint space on medial side of the joint only (*white arrow*), the sclerosis of the bone in the medial compartment providing evidence of cortical thickening (*black arrow*), and the osteophytes in the medial femur (*white wedge*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although MRI may reveal the extent of pathology in an osteoarthritic joint, it is not indicated as part of the diagnostic workup. Findings such as meniscal tears and cartilage and bone lesions occur not only in most patients with OA in the knee, but also in most older persons without joint pain. MRI findings almost never warrant a change in therapy.

TREATMENT

TREATMENT

Osteoarthritis

The goals of the treatment of OA are to alleviate pain and minimize loss of physical function. To the extent that pain and loss of function are consequences of inflammation, of weakness across the joint, and of laxity and instability, the treatment of OA involves addressing each of these impairments. Comprehensive therapy consists of a multimodality approach including nonpharmacologic and pharmacologic elements.

Patients with mild and intermittent symptoms may need only reassurance or nonpharmacologic treatments. Patients with ongoing, disabling pain are likely to need both nonpharmacotherapy and pharmacotherapy.

Treatments for knee OA have been more completely evaluated than those for hip and hand OA or for disease in other joints. Thus, although the principles of treatment are identical for OA in all joints, we shall focus below on the treatment of knee OA, noting specific recommendations for disease in other joints, especially when they differ from those for the knee.

NONPHARMACOTHERAPY

Because OA is a mechanically driven disease, the mainstay of treatment involves altering loading across the painful joint and improving the function of joint protectors, so they can better distribute load across the joint. Ways of lessening focal load across the joint include:

1. avoiding painful activities as these are usually activities that overload the joint;
2. improving the strength and conditioning of muscles that bridge the joint, so as to optimize their function; and
3. unloading the joint, either by redistributing load within the joint with a brace or a splint or by unloading the joint during weight bearing with a cane or a crutch.

The simplest treatment for many patients is to avoid activities that precipitate pain. For example, for the middle-aged patient whose long-distance running brings on symptoms of knee OA, a less demanding form of weight-bearing activity may alleviate all symptoms. For an older person whose daily walks up and down hills bring on knee pain, routing these away from hills might eliminate symptoms.

Since the loading effect of each pound of weight is multiplied across the knee three- to sixfold, each pound of weight loss may have a commensurate multiplier effect, unloading both knees and hips and probably relieving pain in those joints.

In hand joints affected by OA, splinting, by limiting motion, often minimizes pain for patients with involvement especially in the base of the thumb. Weight-bearing joints such as knees and hips can be unloaded by using a cane in the hand opposite the affected joint for partial weight bearing. A physical therapist can help teach the patient how to use the cane optimally, including ensuring that its height is optimal for unloading. Crutches or walkers can serve a similar beneficial function.

Exercise

Osteoarthritic pain in knees or hips during weight bearing results in lack of activity and poor mobility, and because OA is so common, the inactivity that results increases the risk of cardiovascular disease and obesity.

Aerobic capacity is poor in most elders with symptomatic knee OA, worse than others of the same age.

Weakness in muscles that bridge osteoarthritic joints is multifactorial in etiology. First, there is a decline in strength with age. Second, with limited mobility comes disuse muscle atrophy. Third, patients with painful knee or hip OA alter their gait so as to lessen loading across the affected joint, and this further diminishes muscle use. Fourth, “arthrogenous inhibition” may occur, whereby contraction of muscles bridging the joint is inhibited by a nerve afferent feedback loop emanating in a swollen and stretched joint capsule; this prevents maximal attainment of voluntary maximal strength. Because adequate muscle strength and conditioning are critical to joint protection, weakness in a muscle that bridges a diseased joint makes the joint more susceptible to further damage and pain. The degree of weakness correlates strongly with the severity of joint pain and the degree of physical limitation. One of the cardinal elements of the treatment of OA is to improve the functioning of muscles surrounding the joint.

Trials in knee and hip OA have shown that exercise lessens pain and improves physical function. Most effective exercise regimens consist of aerobic and/or resistance training, the latter of which focuses on strengthening muscles across the joint. Exercises are likely to be effective especially if they train muscles for the activities a person performs daily. Activities that increase pain in the joint should be avoided, and the exercise regimen needs to be individualized to optimize effectiveness. Range-of-motion exercises, which do not strengthen muscles, and isometric exercises that strengthen muscles, but not through range of motion, are unlikely to be effective by themselves. Low-impact exercises, including water aerobics and water resistance training, are often better tolerated by patients than exercises involving impact loading, such as running or treadmill exercises. A patient should be referred to an exercise class or to a therapist who can create an individualized regimen. In addition to conventional exercise regimens, tai chi may be effective for knee OA. However, there is no strong evidence that patients with hand OA benefit from therapeutic exercise.

Adherence over the long term is the major challenge to an exercise prescription. In trials involving patients with knee OA, who are engaged in exercise treatment, from a third to over half of patients stopped exercising by 6 months. Less than 50% continued regular exercise at 1 year. The strongest predictor of a patient’s continued exercise is a previous personal history of successful exercise. Physicians should reinforce the exercise prescription at each clinic visit, help the patient recognize barriers to ongoing exercise, and identify convenient times for exercise to be done routinely. The combination of exercise with calorie restriction and weight loss is especially effective in lessening pain.

Correction of Malalignment

Malalignment in the frontal plane (varus-valgus) markedly increases the stress across the joint, which can lead to progression of disease and to pain and disability (Fig. 364-5). Correcting varus-valgus malalignment, either surgically or with bracing, may relieve pain in persons whose knees are malaligned. However, correcting malalignment is often very challenging. Fitted braces that straighten varus knees by putting valgus stress across the knee can be effective. Unfortunately, many patients are unwilling to wear a realigning knee brace; in addition, in patients with obese legs, braces may slip with usage and lose their realigning effect. Braces are indicated for willing patients who can learn to put them on correctly and on whom they do not slip.

Pain from the patellofemoral compartment of the knee can be caused by tilting of the patella or patellar malalignment with the patella riding laterally in the femoral trochlear groove. Using a patellar brace to realign the patella, or tape to pull the patella back into the trochlear sulcus or reduce its tilt, has been shown, when compared to control in clinical trials, to lessen patellofemoral pain. However, patients may find it difficult to apply tape, and skin irritation from the tape is common and like realigning braces, patellar braces may slip.

Although their effect on malalignment is questionable, neoprene sleeves pulled up to cover the knee lessen pain and are easy to use and popular among patients. The explanation for their therapeutic effect on pain is unclear.

In patients with knee OA, acupuncture produces modest pain relief compared to placebo needles and may be an adjunctive treatment.

PHARMACOTHERAPY

Although nonpharmacologic approaches to therapy constitute its mainstay, pharmacotherapy serves an important adjunctive role in OA treatment for symptom management. Available drugs are administered using oral, topical, and intraarticular routes. To date, there are no available drugs that alter the disease process itself.

Acetaminophen, Nonsteroidal Anti-Inflammatory Drugs (NSAIDs), and Cyclooxygenase-2 (Cox-2) Inhibitors

Acetaminophen (paracetamol) is the initial analgesic of choice for patients with OA in knees, hips, or hands, even though its treatment effect in OA is small **(Table 364-1)**. For a minority of patients, it is adequate to control symptoms, in which case more toxic drugs such as NSAIDs can be avoided.

NSAIDs are the most popular drugs to treat osteoarthritic pain. They can be administered either topically or orally. In clinical trials, oral NSAIDs produce ~30% greater improvement in pain than high-dose acetaminophen. Occasional patients treated with NSAIDs experience dramatic pain relief, whereas others experience little improvement. Initially, NSAIDs should be administered topically or taken orally on an “as needed” basis because side effects are less frequent with low intermittent doses. If occasional medication use is insufficiently effective, then daily treatment may be indicated, with an anti-inflammatory dose selected (Table 364-1). Patients should be reminded to take low-dose aspirin and ibuprofen or naproxen at different times to eliminate a drug interaction.

NSAIDs taken orally have substantial and frequent side effects, the most common of which is upper gastrointestinal (GI) toxicity, including dyspepsia, nausea, bloating, GI bleeding, and ulcer disease. Thirty to forty percent of patients experience upper GI side effects so severe as to require discontinuation of medication. To minimize the risk of nonsteroidal-related GI side effects, patients should take NSAIDs after food; if risk is high, patients should take a gastroprotective agent, such as a proton pump inhibitor. Certain oral agents are safer to the stomach than others, including nonacetylated salicylates and nabumetone. Major NSAID-related GI side effects can occur in patients who do not complain of upper GI symptoms. In one study of patients hospitalized for GI bleeding, 81% had no premonitory symptoms.

Because of the increased rates of cardiovascular events associated with conventional NSAIDs such as diclofenac, many of these drugs are not appropriate long-term treatment choices for older persons with OA, especially those at high risk of heart disease or stroke. The American Heart Association has identified rofecoxib and all other COX-2 inhibitors as putting patients at high risk, although low doses of celecoxib (≤200 mg/d of celecoxib) are not associated with an elevation of risk. The only conventional NSAID that appears safe from a cardiovascular perspective is naproxen, but it does have GI toxicity.

There are other common side effects of NSAIDs, including the tendency to develop edema because of prostaglandin inhibition of afferent blood supply to glomeruli in the kidneys and, for similar reasons, a predilection toward reversible renal insufficiency. Blood pressure may increase modestly in some NSAID-treated patients. Oral NSAIDs should not be used in patients with stage IV or V renal disease and should be used with caution in those with stage III disease.

NSAIDs can be placed into a gel or topical solution with another chemical modality that enhances penetration of the skin barrier creating a topical NSAID. When absorbed through the skin, plasma concentrations are an order of magnitude lower than with the same amount of drug administered orally or parenterally. However, when these drugs are administered topically in proximity to a superficial joint (knees, hands, but not hips), the drug can be found in joint tissues such as the synovium and cartilage. Trial results have varied but generally have found that topical NSAIDs are slightly less efficacious than oral agents, but have far fewer GI and systemic side effects. Unfortunately, topical NSAIDs often cause local skin irritation where the medication is applied, inducing redness, burning, or itching (see Table 364-1).

Intraarticular Injections: Glucocorticoids and Hyaluronic Acid

Because synovial inflammation is likely to be a major cause of pain in patients with OA, local anti-inflammatory treatments administered intraarticularly may be effective in ameliorating pain, at least temporarily. Glucocorticoid injections provide such efficacy, but response is variable, with some patients having little relief of pain, whereas others experience pain relief lasting several months. Synovitis, a major cause of joint pain in OA, may abate after an injection, and this correlates with the reduction in knee pain severity. Glucocorticoid injections are useful to get patients over acute flares of pain, but their effects usually last less than 3 months. Repeated injections may cause minor amounts of cartilage loss with probably unimportant clinical consequences.

Hyaluronic acid injections can be given for treatment of symptoms in knee and hip OA, but there is controversy as to whether they have efficacy versus placebo (Table 364-1).

Other Classes of Drugs and Nutraceuticals

For patients with symptomatic knee or hip OA who have not had an adequate response to the treatments above and are either unwilling to undergo or are not candidates for total joint arthroplasty, opioid analgesics have shown modest efficacy and can be tried. Opioid management plans and patient selection are critical. Another option is the use of duloxetine, which has demonstrated modest efficacy in OA.

Recent guidelines recommend against the use of glucosamine or chondroitin for OA. Large publicly supported trials have failed to show that, compared with placebo, these compounds relieve pain in persons with disease.

Optimal nonsurgical therapy for OA is often achieved by trial and error, with each patient having idiosyncratic responses to specific treatments. Placebo (or contextual) effects may account for 50% of more of treatment effects in OA and certain modes of treatment delivery including intraarticular injections have greater contextual effects than others such as pills. When medical therapies have failed and the patient has an unacceptable reduction in their quality of life and ongoing pain and disability, then at least for knee and hip OA, total joint arthroplasty is indicated.

SURGERY

For knee OA, several operations are available. Arthroscopic debridement and lavage have diminished in popularity after randomized trials evaluating this operation have showed that its efficacy is no greater than that of sham surgery for relief of pain or disability. Although arthroscopic meniscectomy is indicated for acute meniscal tears in which symptoms such as locking and acute pain are clearly related temporally to a knee injury that produced the tear, recent trials show that doing a partial meniscectomy in persons with OA and a symptomatic meniscal tear does not relieve knee pain or improve function or even lead to resolution of catching or locking of the knee.

For patients with knee OA isolated to the medial compartment, operations to realign the knee to lessen medial loading can relieve pain. These include a high tibial osteotomy, in which the tibia is broken just below the tibial plateau and realigned so as to load the lateral, nondiseased compartment, or a unicompartmental replacement with realignment. Each surgery may provide the patient with years of pain relief before a total knee replacement is required.

Ultimately, when the patient with knee or hip OA has failed nonsurgical treatment modalities with limitations of pain or function that compromise the quality of life, the patients with reasonable expectations and readiness for surgery should be referred for total knee or hip arthroplasty. These are highly efficacious operations that relieve pain and improve function in the vast majority of patients, although rates of success are higher for hip than knee replacement. Currently, failure rates for both are ~1% per year, although these rates are higher in obese patients. The chance of surgical success is greater in centers where at least 25 such operations are performed yearly or with surgeons who perform multiple operations annually. The timing of knee or hip replacement is critical. If the patient suffers for many years until their functional status has declined substantially, with considerable muscle weakness, postoperative functional status may not improve to a level achieved by others who underwent operation earlier in their disease course.

Cartilage Regeneration

Chondrocyte transplantation has not been found to be efficacious in OA, perhaps because OA includes pathology of joint mechanics, which is not corrected by chondrocyte transplants. Similarly, abrasion arthroplasty (chondroplasty) has not been well studied for efficacy in OA, but it produces fibrocartilage in place of damaged hyaline cartilage. Both of these surgical attempts to regenerate and reconstitute articular

cartilage may be more likely to be efficacious early in disease when joint malalignment and many of the other noncartilage abnormalities that characterize OA have not yet developed.

TABLE 364-1

Pharmacologic Treatment for Osteoarthritis

Treatment	Dosage	Comments
Acetaminophen	Up to 1 g tid	Prolongs half-life of warfarin. Make sure patient is not taking other treatments containing acetaminophen to avoid hepatic toxicity.
Oral NSAIDs and COX-2 inhibitors Naproxen Salsalate Ibuprofen Celecoxib	375–500 mg bid 1500 mg bid 600–800 mg 3–4 times a day 100–200 mg qd	Take with food. Increased risk of myocardial infarction and stroke for some NSAIDs and especially COX-2 inhibitors. High rates of gastrointestinal side effects, including ulcers and bleeding, occur. Patients at high risk for gastrointestinal side effects should also take either a proton pump inhibitor or misoprostol.[a] There is an increase in gastrointestinal side effects or bleeding when taken with acetylsalicylic acid. Can also cause edema and renal insufficiency.
Topical NSAIDs		Rub onto joint. Few systemic side effects. Skin irritation common.
Diclofenac Na 1% gel	4 g qid (for knees, hands)	
Opiates	Various	Common side effects include dizziness, sedation, nausea or vomiting, dry mouth, constipation, urinary retention, and pruritus. Respiratory and central nervous system depression can occur.
Capsaicin	0.025–0.075% cream 3–4 times a day	Can irritate mucous membranes.
Intraarticular injections		
Steroids		

Treatment	Dosage	Comments
Hyaluronans	Varies from 3 to 5 weekly injections depending on preparation	Mild to moderate pain at injection site. Controversy exists regarding efficacy.

[a]Patients at high risk include those with previous gastrointestinal events, persons ≥60 years, and persons taking glucocorticoids. Trials have shown the efficacy of proton pump inhibitors and misoprostol in the prevention of ulcers and bleeding. Misoprostol is associated with a high rate of diarrhea and cramping; therefore, proton pump inhibitors are more widely used to reduce NSAID-related gastrointestinal symptoms.

Abbreviations: COX-2, cyclooxygenase-2; NSAIDs, nonsteroidal anti-inflammatory drugs.

Source: Adapted from DT Felson: N Engl J Med 354:841, 2006.

FURTHER READING

Felson D: Safety of nonsteroidal antiinflammatory drugs. N Engl J Med 375:2595, 2016. [PubMed: 28029923]

Glyn-Jones S et al: Osteoarthritis. Lancet 386:376, 2015. [PubMed: 25748615]

McAlindon TE et al: Effect of intra-articular triamcinolone vs saline on knee cartilage volume and pain in patients with knee osteoarthritis: A randomized clinical trial. JAMA 317:1967, 2017.

McAlindon TE et al: OARSI guidelines for the non-surgical management of knee osteoarthritis. Osteoarthritis Cartilage 22:363, 2014. [PubMed: 24462672]

Neogi T et al: Sensitivity and sensitisation in relation to pain severity in knee osteoarthritis: Trait or state? Ann Rheum Dis 74:682, 2015. [PubMed: 24351516]

O'Neill TW et al: Synovial tissue volume: A treatment target in knee osteoarthritis (OA). Ann Rheum Dis 75:84, 2016. [PubMed: 26116548]

Chapter 365: Gout and Other Crystal-Associated Arthropathies

H. Ralph Schumacher; Lan X. Chen

INTRODUCTION

The use of polarizing light microscopy during synovial fluid analysis in 1961 by McCarty and Hollander and the subsequent application of other crystallographic techniques, such as electron microscopy, energy-dispersive elemental analysis, and x-ray diffraction, have allowed investigators to identify the roles of different microcrystals, including monosodium urate (MSU), calcium pyrophosphate (CPP), calcium apatite (apatite), and calcium oxalate (CaOx), in inducing acute or chronic arthritis or periarthritis[1]. The clinical events that result from deposition of MSU, CPP, apatite, and CaOx have many similarities but also have important differences. Because of often similar clinical presentations, the need to perform synovial fluid analysis to distinguish the type of crystal involved must be emphasized. Polarized light microscopy alone can identify most typical crystals; apatite, however, is an exception. Aspiration and analysis of effusions are also important to assess the possibility of infection. Apart from the identification of specific microcrystalline materials or organisms, synovial fluid characteristics in crystal-associated diseases are nonspecific, and synovial fluid can be inflammatory or noninflammatory. Without crystal identification, these diseases can be confused with rheumatoid or other types of arthritis. A list of possible musculoskeletal manifestations of crystal-associated arthritis is shown in **Table 365-1**.

TABLE 365-1

Musculoskeletal Manifestations of Crystal-Induced Arthritis

Acute mono- or polyarthritis	Destructive arthropathies
Bursitis	Chronic inflammatory arthritis
Tendinitis	Spinal arthritis
Enthesitis	Peculiar type of osteoarthritis
Tophaceous deposits	Carpal tunnel syndrome

[1]Deceased

GOUT

Gout is a metabolic disease that most often affects middle-aged to elderly men and postmenopausal women. It results from an increased body pool of urate with hyperuricemia. It typically is characterized by episodic acute arthritis or chronic arthritis caused by deposition of MSU crystals in joints and connective tissue tophi and the risk for deposition in kidney interstitium or uric acid nephrolithiasis (Chap. 410).

ACUTE AND CHRONIC ARTHRITIS

Acute arthritis is the most common early clinical manifestation of gout. Usually, only one joint is affected initially, but polyarticular acute gout can occur in subsequent episodes. The metatarsophalangeal joint of the first toe often is involved, but tarsal joints, ankles, and knees also are affected commonly. Especially in elderly patients or in advanced disease, finger joints may be involved. Inflamed Heberden's or Bouchard's nodes may be a first manifestation of gouty arthritis. The first episode of acute gouty arthritis frequently begins at night with dramatic joint pain and swelling. Joints rapidly become warm, red, and tender, with a clinical appearance that often mimics that of cellulitis. Early attacks tend to subside spontaneously within 3–10 days, and most patients have intervals of varying length with no residual symptoms until the next episode. Several events may precipitate acute gouty arthritis: dietary excess, trauma, surgery, excessive ethanol ingestion, hypouricemic therapy, and serious medical illnesses such as myocardial infarction and stroke.

After many acute mono- or oligoarticular attacks, a proportion of gouty patients may present with a chronic nonsymmetric synovitis, causing potential confusion with rheumatoid arthritis (Chap. 351). Less commonly, chronic gouty arthritis will be the only manifestation, and, more rarely, the disease will manifest only as periarticular tophaceous deposits in the absence of synovitis. Women represent only 5–20% of all patients with gout. Most women with gouty arthritis are postmenopausal and elderly, have osteoarthritis and arterial

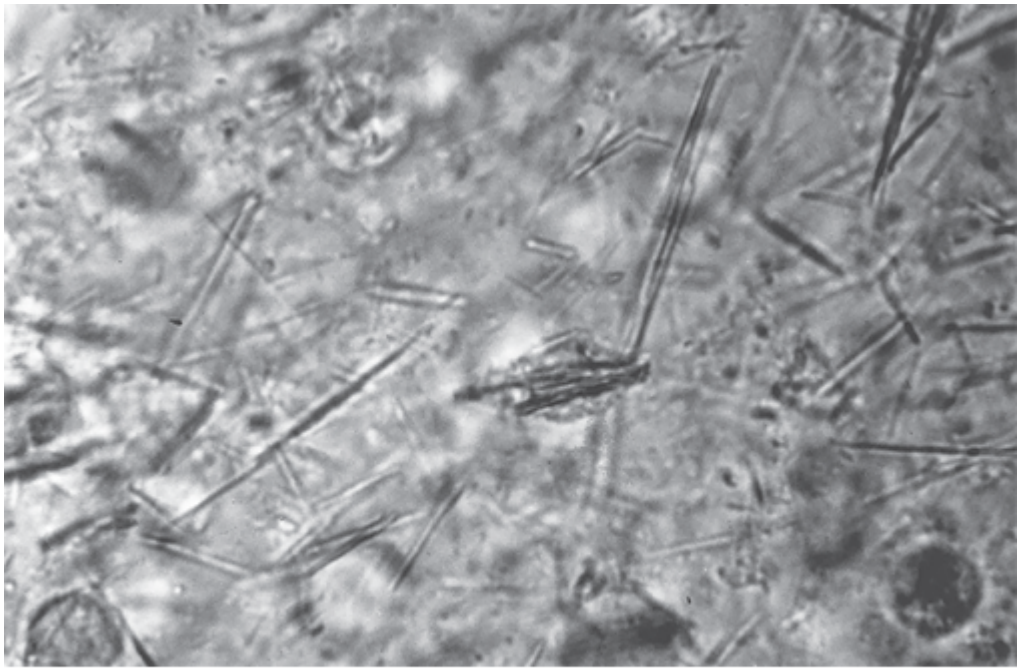
hypertension that causes mild renal insufficiency, and usually are receiving diuretics. Premenopausal gout is rare. Kindreds of precocious gout in young women caused by decreased renal urate clearance and renal insufficiency have been described.

Laboratory Diagnosis

Even if the clinical appearance strongly suggests gout, the presumptive diagnosis ideally should be confirmed by needle aspiration of acutely or chronically involved joints or tophaceous deposits. Acute septic arthritis, several of the other crystalline- associated arthropathies, palindromic rheumatism, and psoriatic arthritis may present with similar clinical features. During acute gouty attacks, needle-shaped MSU crystals typically are seen both intracellularly and extracellularly **(Fig. 365-1)**. With compensated polarized light, these crystals are brightly birefringent with negative elongation. Synovial fluid leukocyte counts are elevated from 2000 to 60,000/μL. Effusions appear cloudy due to the increased numbers of leukocytes. Large amounts of crystals occasionally produce a thick pasty or chalky joint fluid. Bacterial infection can coexist with urate crystals in synovial fluid; if there is any suspicion of septic arthritis, joint fluid must be cultured.

FIGURE 365-1

Extracellular and intracellular monosodium urate crystals, as seen in a fresh preparation of synovial fluid, illustrate needle- and rod-shaped crystals. These crystals are strongly negative birefringent crystals under compensated polarized light microscopy; 400×.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MSU crystals also can often be demonstrated in the first metatarsophalangeal joint and in knees not acutely involved with gout. Arthrocentesis of these joints is a useful technique to establish the diagnosis of gout between attacks.

Serum uric acid levels can be normal or low at the time of an acute attack, as inflammatory cytokines can be uricosuric and effective initiation of hypouricemic therapy can precipitate attacks. This limits the value of serum uric acid determinations for the diagnosis of gout. Nevertheless, serum urate levels are almost always elevated at some time and are important to use to follow the course of hypouricemic therapy. A 24-h urine collection for uric acid can, in some cases, be useful in assessing the risk of stones, elucidating overproduction or underexcretion of uric acid, and deciding whether it may be appropriate to use a uricosuric therapy **(Chap. 410)**. Excretion of >800 mg of uric acid per 24 h on a regular diet suggests that causes of overproduction of purine should be considered. Urinalysis, serum creatinine, hemoglobin, white blood cell (WBC) count, liver function tests, and serum lipids should be obtained because of possible pathologic sequelae of gout and other associated diseases requiring treatment and as baselines because of possible adverse effects of gout treatment.

Radiographic Features

Cystic changes, well-defined erosions with sclerotic margins (often with overhanging bony edges), and soft tissue masses are characteristic radiographic features of advanced chronic tophaceous gout. Ultrasound may aid earlier diagnosis by showing a double contour sign overlying the articular cartilage. Dual-energy computed tomography (CT) can show specific features establishing the presence of urate crystals.

TREATMENT

TREATMENT

Gout

ACUTE GOUTY ARTHRITIS

The mainstay of treatment during an acute attack is the administration of anti-inflammatory drugs such as nonsteroidal anti-inflammatory drugs (NSAIDs), colchicine, or glucocorticoids. NSAIDs are used most often in individuals without complicating comorbid conditions. Both colchicine and NSAIDs may be poorly tolerated and dangerous in the elderly and in the presence of renal insufficiency and gastrointestinal disorders. Ice pack applications and rest of the involved joints can be helpful. Colchicine given orally is a traditional and effective treatment if used early in an attack. Useful regimens are one 0.6-mg tablet given every 8 h with subsequent tapering or 1.2 mg followed by 0.6 mg in 1 h with subsequent day dosing depending on response. This is generally better tolerated than the formerly advised higher dose regimens. The drug must be at least temporarily discontinued promptly at the first sign of loose stools, and symptomatic treatment must be given for the diarrhea. Intravenous colchicine has been taken off the market. NSAIDs given in full anti-inflammatory doses are effective in ~90% of patients, and the resolution of signs and symptoms usually occurs in 5–8 days. The most effective drugs are any of those with a short half-life and include indomethacin, 25–50 mg tid; naproxen, 500 mg bid; ibuprofen, 800 mg tid; diclofenac, 50 mg tid; and celecoxib 800 mg followed by 400 mg 12 h later, then 400 mg bid.

Glucocorticoids given as an intramuscular injection or orally, for example, prednisone, 30–50 mg/d as the initial dose and gradually tapered with the resolution of the attack, can be effective in polyarticular gout. For a single joint or a few involved joints, intraarticular triamcinolone acetonide, 20–40 mg, or methylprednisolone, 25–50 mg, have been effective and well tolerated. Based on recent evidence on the essential role of the inflammasome and interleukin 1β (IL-1β) in acute gout, daily anakinra has been used when other treatments have failed or were contraindicated.

HYPOURICEMIC THERAPY

Ultimate control of gout requires correction of the basic underlying defect: the hyperuricemia. Attempts to normalize serum uric acid to <300–360 µmol/L (5.0–6.0 mg/dL) to prevent recurrent gouty attacks and eliminate tophaceous deposits are critical and entail a commitment to hypouricemic regimens and medications that generally are required for life. Hypouricemic drug therapy should be considered when, as in most patients, the hyperuricemia cannot be corrected by simple means (control of body weight, low-purine diet, increase in liquid intake, limitation of ethanol use, decreased use of fructose-containing foods and beverages, and avoidance of diuretics). The decision to initiate hypouricemic therapy usually is made taking into consideration the number of acute attacks (urate lowering may be cost-effective after two attacks), serum uric acid levels (progression is more rapid in patients with serum uric acid >535 µmol/L [>9.0 mg/dL]), the patient's willingness to commit to lifelong therapy, or the presence of uric acid stones. Urate-lowering therapy should be initiated in any patient who already has tophi or chronic gouty arthritis. Uricosuric agents such as probenecid can be used in patients with good renal function who underexcrete uric acid, with <600 mg in a 24-h urine sample. Urine volume should be maintained by ingestion of 1500 mL of water every day. Probenecid can be started at a dose of 250 mg twice daily and increased gradually as needed up to 3 g per day to achieve and maintain a serum uric acid level of <6 mg/dL. Probenecid is generally not effective in patients with serum creatinine levels >177 µmol/L (2 mg/dL). These patients may require allopurinol or benzbromarone (not available in the United States). Benzbromarone is another uricosuric drug that is more effective in patients with chronic kidney disease. Lesinurad is a newer uricosuric; however, it is approved only in patients already on a xanthine oxidase inhibitor as an adjuvant at 200 mg per day. Some agents used to treat common comorbidities, including losartan, fenofibrate, and amlodipine, have some mild uricosuric effects.

The xanthine oxidase inhibitor allopurinol is by far the most commonly used hypouricemic agent and is the best drug to lower serum urate in overproducers, urate stone formers, and patients with renal disease. It can be given in a single morning dose, usually 100 mg initially and increasing up to 800 mg if needed. In patients with chronic renal disease, the initial allopurinol dose should be lower and adjusted depending on the serum creatinine concentration; for example, with a creatinine clearance of 10 mL/min, one generally would use 100 mg daily. Doses can be increased gradually to reach the target urate level of less than 6 mg/dL. Toxicity of allopurinol has been recognized increasingly in patients who use thiazide diuretics, in patients allergic to penicillin and ampicillin, and in Asians expressing *HLA-B*58:01*. The most serious side effects include life-threatening toxic epidermal necrolysis, systemic vasculitis, bone marrow suppression, granulomatous hepatitis, and renal failure. Patients with mild cutaneous reactions to allopurinol can reconsider the use of a uricosuric agent, undergo an attempt at desensitization to allopurinol, or take febuxostat, a new, chemically

unrelated specific xanthine oxidase inhibitor. Febuxostat is approved in the United States at 40 or 80 mg once a day and does not require dose adjustment in mild to moderate renal disease. Pegloticase is a pegylated uricase, available for patients who do not tolerate or fail full doses of other treatments. It is given intravenously usually at 8 mg every 2 weeks and can dramatically lower serum uric acid in up to 50% of such patients.

Urate-lowering drugs are generally not initiated during acute attacks, but after the patient is stable and low-dose colchicine has been initiated to decrease the risk of the flares that often, without anti-inflammatory treatment, occur with urate lowering. Colchicine anti-inflammatory prophylaxis in doses of 0.6 mg one to two times daily should be given along with the hypouricemic therapy until the patient is normouricemic and without gouty attacks for 6 months or as long as tophi are present. Colchicine should not be used in dialysis patients and is given in lower doses to the patients with renal disease or with P glycoprotein or CYP3A4 inhibitors such as clarithromycin that can increase toxicity of colchicine.

CALCIUM PYROPHOSPHATE DEPOSITION (CPPD) DISEASE

PATHOGENESIS

The deposition of CPP crystals in articular tissues is most common in the elderly, occurring in 10–15% of persons age 65–75 years and 30–50% of those >85 years. In most cases, this process is asymptomatic, and the cause of CPPD is uncertain. Because >80% of patients are >60 years and 70% have preexisting joint damage from other conditions, it is likely that biochemical changes in aging or diseased cartilage favor crystal nucleation. In patients with CPPD arthritis, there is increased production of inorganic pyrophosphate and decreased levels of pyrophosphatases in cartilage extracts. Mutations in the *ANKH* gene, as described in both familial and sporadic cases, can increase elaboration and extracellular transport of pyrophosphate. The increase in pyrophosphate production appears to be related to enhanced activity of ATP pyrophosphohydrolase and 5′-nucleotidase, which catalyze the reaction of ATP to adenosine and pyrophosphate. This pyrophosphate could combine with calcium to form CPP crystals in matrix vesicles or on collagen fibers. There are decreased levels of cartilage glycosaminoglycans that normally inhibit and regulate crystal nucleation. High activities of transglutaminase enzymes also may contribute to the deposition of CPP crystals.

Release of CPP crystals into the joint space is followed by the phagocytosis of those crystals by monocyte-macrophages and neutrophils, which respond by releasing chemotactic and inflammatory substances and, as with MSU crystals, activating the inflammasome.

A minority of patients with CPPD arthropathy have metabolic abnormalities or hereditary CPP disease **(Table 365-2)**. These associations suggest that a variety of different metabolic products may enhance CPP crystal deposition either by directly altering cartilage or by inhibiting inorganic pyrophosphatases. Included among these conditions are hyperparathyroidism, hemochromatosis, hypophosphatasia, and hypomagnesemia. The presence of CPPD arthritis in individuals aged <50 years should lead to consideration of these metabolic disorders (Table 365-2) and inherited forms of disease, including those identified in a variety of ethnic

groups. Genomic DNA studies performed on different kindreds have shown a possible location of genetic defects on chromosome 8q or on chromosome 5p in a region that expresses the gene of the membrane pyrophosphate channel (*ANKH* gene). Investigation of younger patients with CPPD should include inquiry for evidence of familial aggregation and evaluation of serum calcium, phosphorus, alkaline phosphatase, magnesium, iron, and transferrin.

TABLE 365-2

Conditions Associated with Calcium Pyrophosphate Crystal Deposition Disease

Aging
Disease-associated
Primary hyperparathyroidism
Hemochromatosis
Hypophosphatasia
Hypomagnesemia
Chronic gout
Postmeniscectomy
Gitelman's syndrome
Epiphyseal dysplasias

CLINICAL MANIFESTATIONS

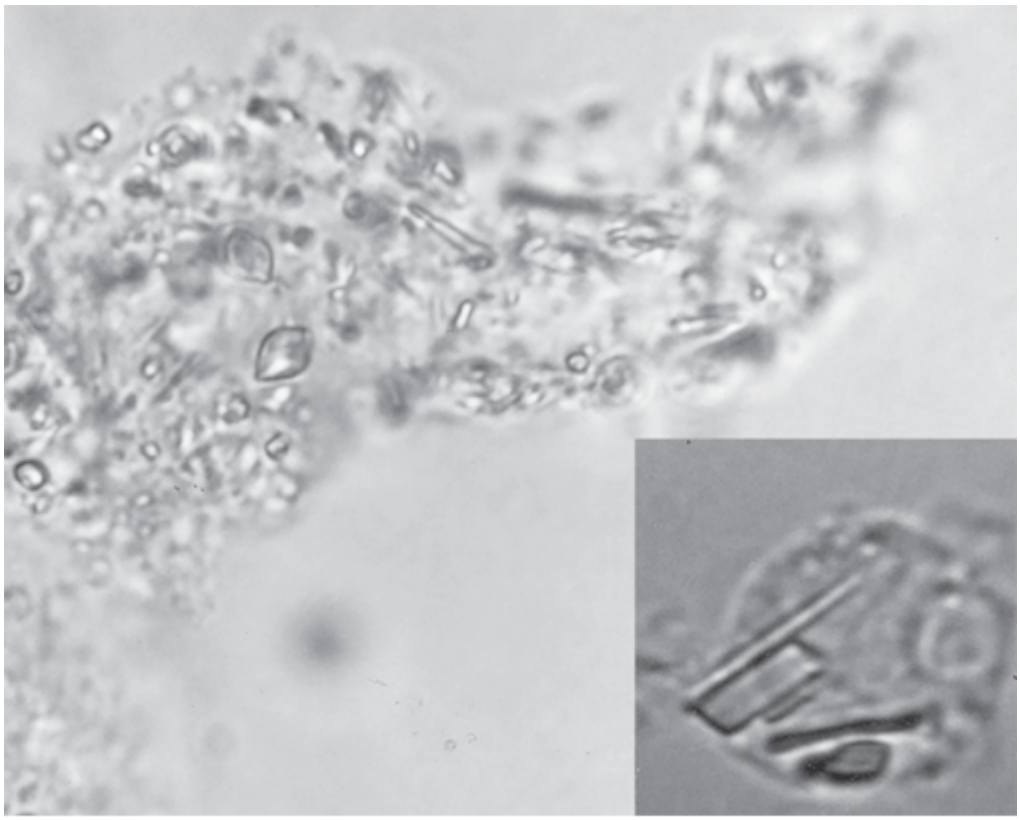
CPPD arthropathy may be asymptomatic, acute, subacute, or chronic or may cause acute synovitis superimposed on chronically involved joints. Acute CPPD arthritis originally was termed *pseudogout* by McCarty and co-workers because of its striking similarity to gout. Other clinical manifestations of CPPD include (1) association with or enhancement of peculiar forms of osteoarthritis; (2) induction of severe destructive disease that may radiographically mimic neuropathic arthritis; (3) production of chronic symmetric synovitis that is clinically similar to rheumatoid arthritis; (4) intervertebral disk and ligament calcification with restriction of spine mobility, the crowned dens syndrome, or spinal stenosis (most commonly seen in the elderly); and (5) rarely periarticular tophus-like nodules.

The knee is the joint most frequently affected in CPPD arthropathy. Other sites include the wrist, shoulder, ankle, elbow, and hands. The temporomandibular joint may be involved. Clinical and radiographic evidence indicates that CPPD deposition is polyarticular in at least two-thirds of patients. When the clinical picture resembles that of slowly progressive osteoarthritis, diagnosis may be difficult. Joint distribution may provide important clues suggesting CPPD disease. For example, primary osteoarthritis less often involves metacarpophalangeal, wrist, elbow, shoulder, or ankle joints. If radiographs or ultrasound reveal punctate and/or linear radiodense deposits within fibrocartilaginous joint menisci or articular hyaline cartilage (*chondrocalcinosis*), the diagnostic likelihood of CPPD disease is further increased. *Definitive diagnosis*

requires demonstration of typical rhomboid or rodlike crystals (generally weakly positively birefringent or nonbirefringent with polarized light) in synovial fluid or articular tissue **(Fig. 365-2)**. In the absence of joint effusion or indications to obtain a synovial biopsy, chondrocalcinosis is presumptive of CPPD. One exception is chondrocalcinosis due to CaOx in some patients with chronic renal failure.

FIGURE 365-2

Intracellular and extracellular calcium pyrophosphate (CPP) crystals, as seen in a fresh preparation of synovial fluid, illustrate rectangular, rod-shaped, and rhomboid crystals that are weakly positively or nonbirefringent crystals (compensated polarized light microscopy; 400×).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Acute attacks of CPPD arthritis may be precipitated by trauma. Rapid diminution of serum calcium concentration, as may occur in severe medical illness or after surgery (especially parathyroidectomy), can also lead to attacks.

In as many as 50% of cases, episodes of CPPD-induced inflammation are associated with low-grade fever and, on occasion, temperatures as high as 40°C (104°F). In such cases, synovial fluid analysis with microbial cultures is essential to rule out the possibility of infection. In fact, infection in a joint with any microcrystalline deposition process can lead to crystal shedding and subsequent synovitis from both crystals and microorganisms. The leukocyte count in synovial fluid in acute CPPD can range from several thousand cells to 100,000 cells/μL, with the mean being about 24,000 cells/μL and the predominant cell being the neutrophil. CPP crystals may be seen inside tissue fragments and fibrin clots and in neutrophils (Fig. 365-2). CPP crystals may coexist with MSU and apatite in some cases.

TREATMENT

TREATMENT

CPPD Disease

Untreated acute attacks may last a few days to as long as a month. Treatment by rest, joint aspiration, and NSAIDs or by intraarticular glucocorticoid injection may result in more rapid return to prior status. For patients with frequent recurrent attacks, daily prophylactic treatment with low doses of colchicine may be helpful in decreasing the frequency of the attacks. Severe polyarticular attacks usually require short courses of glucocorticoids or an IL-1β antagonist, anakinra. Unfortunately, there is no effective way to remove CPP deposits from cartilage and synovium. Uncontrolled studies suggest that the administration of NSAIDS (with a gastric protective agent if required), hydroxychloroquine, or even methotrexate may be helpful in controlling persistent synovitis. Patients with progressive destructive large-joint arthropathy may require joint replacement.

CALCIUM APATITE DEPOSITION DISEASE

PATHOGENESIS

Apatite is the primary mineral of normal bone and teeth. Abnormal accumulation of basic calcium phosphates, largely carbonate substituted apatite, can occur in areas of tissue damage (dystrophic calcification), hypercalcemic or hyperparathyroid states (metastatic calcification), and certain conditions of unknown cause (Table 365-3). In chronic renal failure, hyperphosphatemia can contribute to extensive apatite deposition both in and around joints. Familial aggregation is rarely seen; no association with *ANKH* mutations has been described thus far. Apatite crystals are deposited primarily on matrix vessels. Incompletely understood alterations in matrix proteoglycans, phosphatases, hormones, and cytokines probably can influence crystal formation.

TABLE 365-3

Conditions Associated with Apatite Deposition Disease

Aging
Osteoarthritis
Hemorrhagic shoulder effusions in the elderly (Milwaukee shoulder)
Destructive arthropathy
Tendinitis, bursitis
Tumoral calcinosis (sporadic cases)
Disease-associated
Hyperparathyroidism
Milk-alkali syndrome
Renal failure/long-term dialysis
Connective tissue diseases (e.g., systemic sclerosis, dermatomyositis, SLE)
Heterotopic calcification after neurologic catastrophes (e.g., stroke, spinal cord injury)
Heredity
Bursitis, arthritis
Tumoral calcinosis
Fibrodysplasia ossificans progressiva

Abbreviation: SLE, systemic lupus erythematosus.

Apatite aggregates are commonly present in synovial fluid in an extremely destructive chronic arthropathy of the elderly that occurs most often in the shoulders (Milwaukee shoulder) and in a similar process in hips, knees, and erosive osteoarthritis of fingers. Joint destruction is associated with damage to cartilage and supporting structures, leading to instability and deformity. Progression tends to be indolent. Symptoms range from minimal to severe pain and disability that may lead to joint replacement surgery. Whether severely affected patients represent an extreme synovial tissue response to the apatite crystals that are so common in osteoarthritis is uncertain. Synovial lining cell or fibroblast cultures exposed to apatite (or CPP) crystals can undergo mitosis and markedly increase the release of prostaglandin E_2, various cytokines, and also collagenases and neutral proteases, underscoring the destructive potential of abnormally stimulated synovial lining cells.

CLINICAL MANIFESTATIONS

Periarticular or articular deposits may occur and may be associated with acute reversible inflammation and/or chronic damage to the joint capsule, tendons, bursa, or articular surfaces. The most common sites of apatite deposition include bursae and tendons in and/or around the knees, shoulders, hips, and fingers. Clinical manifestations include asymptomatic radiographic abnormalities, acute synovitis, bursitis,

tendinitis, and chronic destructive arthropathy. Although the true incidence of apatite arthritis is not known, 30–50% of patients with osteoarthritis have apatite microcrystals in their synovial fluid. Such crystals frequently can be identified in clinically stable osteoarthritic joints, but they are more likely to come to attention in persons experiencing acute or subacute worsening of joint pain and swelling. The synovial fluid leukocyte count in apatite arthritis is usually low (<2000/μL) despite dramatic symptoms, with predominance of mononuclear cells.

DIAGNOSIS

Intra- and/or periarticular calcifications with or without erosive, destructive, or hypertrophic changes may be seen on radiographs (Fig. 365-3). They should be distinguished from the linear calcifications typical of CPPD.

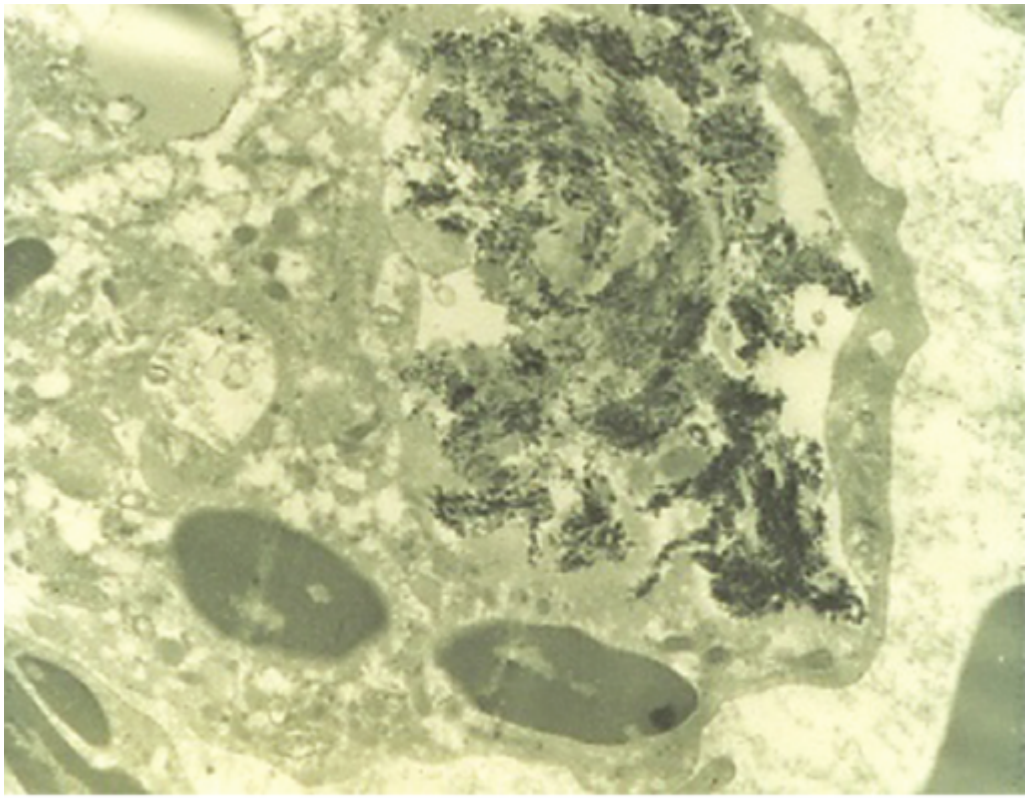
FIGURE 365-3

A. Radiograph showing calcification due to apatite crystals surrounding an eroded joint. ***B.*** An electron micrograph demonstrates dark needle-shaped apatite crystals within a vacuole of a synovial fluid mononuclear cell (30,000×).



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Definitive diagnosis of apatite arthropathy, also called basic calcium phosphate disease, depends on identification of crystals from synovial fluid or tissue (Fig. 365-3). Individual crystals are very small and can be seen only by electron microscopy. Clumps of crystals may appear as 1- to 20-µm shiny intra- or extracellular nonbirefringent globules or aggregates that stain purplish with Wright's stain and bright red with alizarin red S. Tetracycline binding and other investigative techniques are under consideration as labeling alternatives. Absolute identification depends on electron microscopy with energy-dispersive elemental analysis, x-ray diffraction, infrared spectroscopy, or Raman microspectroscopy, but these techniques usually are not required in clinical diagnosis.

TREATMENT

TREATMENT

Calcium Apatite Deposition Disease

Treatment of apatite arthritis or periarthritis is nonspecific. Acute attacks of bursitis or synovitis may be self-limiting, resolving in days to several weeks. Aspiration of effusions and the use of either NSAIDs or oral colchicine for 2 weeks or intra- or periarticular injection of a depot glucocorticoid appear to shorten the duration and intensity of symptoms. Local injection of disodium ethylenediaminetetraacetic acid (EDTA) and SC anakinra have been suggested as effective in single studies of acute calcific tendinitis at the shoulder. Other reports have described that IV gamma globulin, rituximab, calcium channel blockers, or bisphosphonates may help diffuse calcinosis. Periarticular apatite deposits may be resorbed with resolution

of attacks. Agents to lower serum phosphate levels may lead to resorption of deposits in renal failure patients receiving hemodialysis. In patients with underlying severe destructive articular changes, response to medical therapy is usually less rewarding.

CaOx DEPOSITION DISEASE

PATHOGENESIS

Primary oxalosis is a rare hereditary metabolic disorder **(Chap. 413)**. Enhanced production of oxalic acid may result from at least two different enzyme defects, leading to hyperoxalemia and deposition of CaOx crystals in tissues. Nephrocalcinosis and renal failure are typical results. Acute and/or chronic CaOx arthritis, periarthritis, and bone disease may complicate primary oxalosis during later years of illness.

Secondary oxalosis is more common than the primary disorder. In chronic renal disease, CaOx deposits have long been recognized in visceral organs, blood vessels, bones, and cartilage and are now known to be one of the causes of arthritis in chronic renal failure. Thus far, reported patients have been dependent on long-term hemodialysis or peritoneal dialysis **(Chap. 306)**, and many had received ascorbic acid supplements. Ascorbic acid is metabolized to oxalate, which is inadequately cleared in uremia and by dialysis. Such supplements and foods high in oxalate content usually are avoided in dialysis programs because of the risk of enhancing hyperoxalosis and its sequelae.

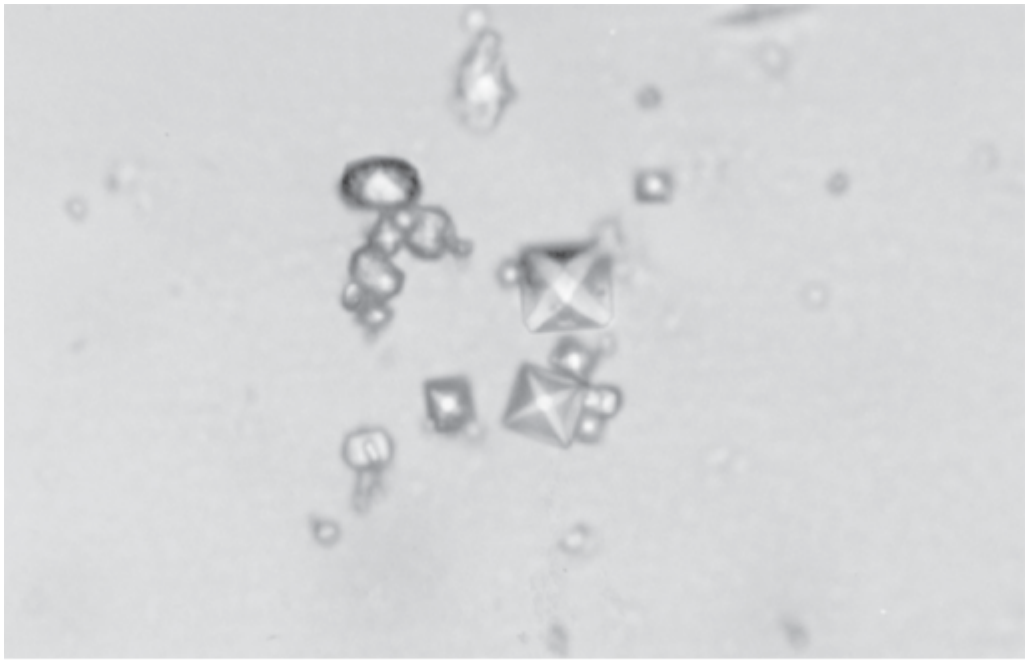
CLINICAL MANIFESTATIONS AND DIAGNOSIS

CaOx aggregates can be found in bone, articular cartilage, synovium, and periarticular tissues. From these sites, crystals may be shed, causing acute synovitis. Persistent aggregates of CaOx can, like apatite and CPP, stimulate synovial cell proliferation and enzyme release, resulting in progressive articular destruction. Deposits have been documented in fingers, wrists, elbows, knees, ankles, and feet.

Clinical features of acute CaOx arthritis may not be distinguishable from those due to urate, CPP, or apatite. Radiographs may reveal chondrocalcinosis or soft tissue calcifications. CaOx-induced synovial effusions are usually noninflammatory, with <2000 leukocytes/μL, or mildly inflammatory. Neutrophils or mononuclear cells can predominate. CaOx crystals have a variable shape and variable birefringence to polarized light. The most easily recognized forms are bipyramidal, have strong birefringence **(Fig. 365-4)**, and stain with alizarin red S.

FIGURE 365-4

Bipyramidal and small polymorphic calcium oxalate crystals from synovial fluid are a classic finding in calcium oxalate arthropathy (ordinary light microscopy; 400×).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Calcium Oxalate Deposition Disease

Treatment of CaOx arthropathy with NSAIDs, colchicine, intra-articular glucocorticoids, and/or an increased frequency of dialysis has produced only slight improvement. In primary oxalosis, liver transplantation has induced a significant reduction in crystal deposits **(Chap. 413)**.

FURTHER READING

Dalbeth N et al: Survey definitions of gout for epidemiological studies: Comparison with crystal identification as the gold standard. Arthritis Care Res (Hoboken) doi: 10.1002/acr.22896 2016 March 25 (Epub ahead of print).

Ogdie A et al: Imaging modalities for the classification of gout: Systematic literature review and meta-analysis. Ann Rheum Dis 74:1868, 2015. [PubMed: 24915980]

Roddy E et al: Concordance of the management of chronic gout in a UK primary-care population with the EULAR gout recommendations. Ann Rheum Dis 66:1311, 2007. [PubMed: 17504843]

Schumacher HR Jr, Reginato AJ: *Atlas of Synovial Fluid Analysis and Crystal Identification*. Philadelphia: Lea & Febiger, 1991.

Chapter 366: Fibromyalgia

Leslie J. Crofford

FIGURE 366-1

DEFINITION

Fibromyalgia (FM) is characterized by chronic widespread musculoskeletal pain and tenderness. Although FM is defined primarily as a pain syndrome, patients also commonly report associated neuro-psychological symptoms of fatigue, unrefreshing sleep, cognitive dysfunction, anxiety, and depression. Patients with FM have an increased prevalence of other syndromes associated with pain and fatigue, including chronic fatigue syndrome (Chap. 442), temporomandibular disorder, chronic headaches, irritable bowel syndrome, interstitial cystitis/painful bladder syndrome, and other pelvic pain syndromes. Available evidence implicates the central nervous system as key to maintaining pain and other core symptoms of FM and related conditions. The presence of FM is associated with substantial negative consequences for physical and social functioning.

EPIDEMIOLOGY

In clinical settings, a diagnosis of FM is made in ~2% of the population and is far more common in women than in men, with a ratio of ~9:1. However, in population-based survey studies worldwide, the prevalence rate is ~2–5%, with a female-to-male ratio of only 2–3:1 and with some variability depending on the method of ascertainment. The prevalence data are similar across socioeconomic classes. Cultural factors may play a role in determining whether patients with FM symptoms seek medical attention; however, even in cultures in which secondary gain is not expected to play a significant role, the prevalence of FM remains in this range.

CLINICAL MANIFESTATIONS

Pain and Tenderness

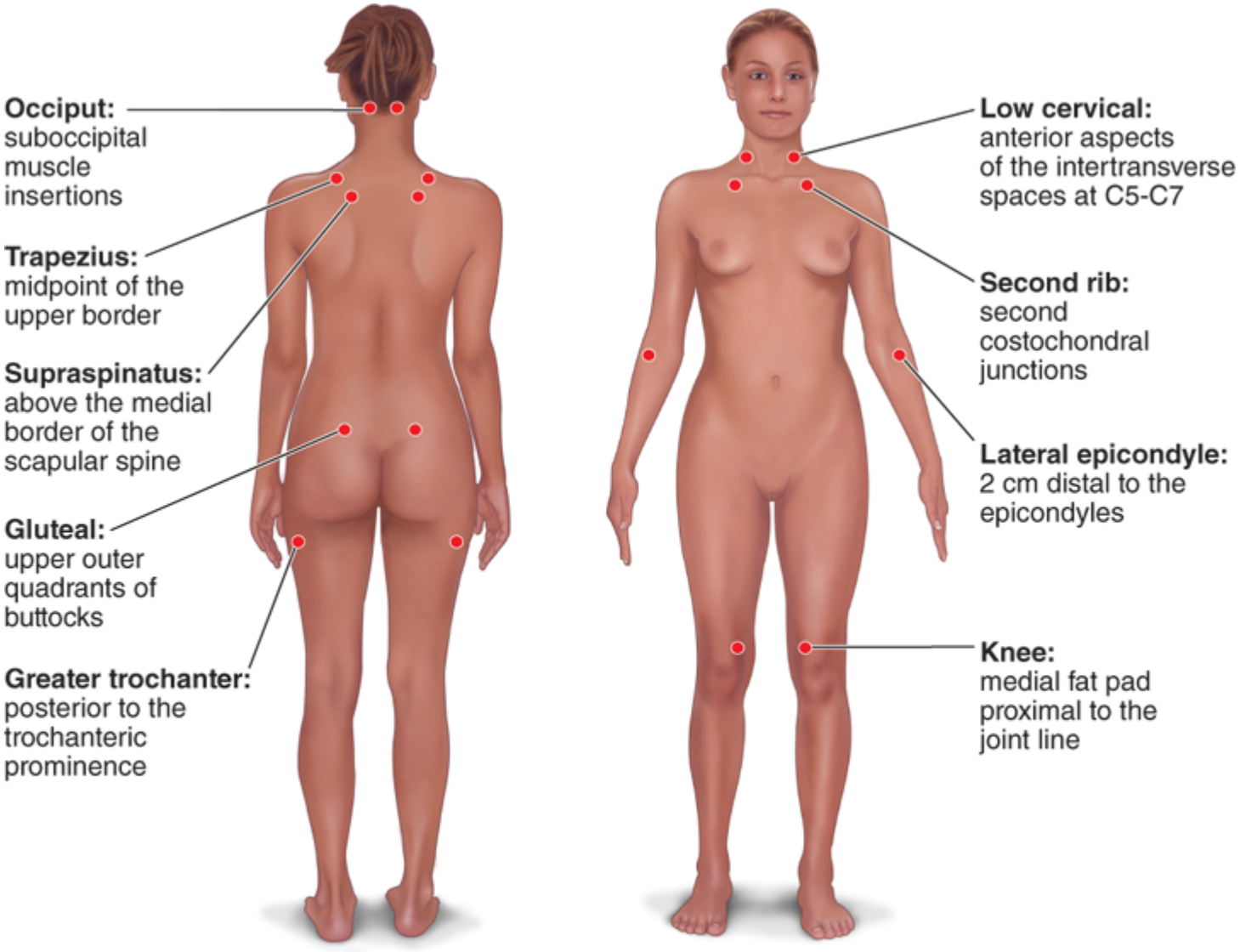
At presentation, patients with FM most commonly report “pain all over.” These patients have pain that is typically both above and below the waist on both sides of the body and involves the axial skeleton (neck, back, or chest). The pain attributable to FM is poorly localized, difficult to ignore, severe in its intensity, and

associated with a reduced functional capacity. For a diagnosis of FM, pain should have been present most of the day on most days for at least 3 months.

The pain of FM is associated with tenderness and increased evoked pain sensitivity. In clinical practice, this elevated sensitivity may be identified by pain induced by the pressure of a blood pressure cuff or skin roll tenderness. More formally, an examiner may complete a tender-point examination in which the examiner uses the thumbnail to exert pressure of ~4 kg/m^2 (or the amount of pressure leading to blanching of the tip of the thumbnail) on well-defined musculotendinous sites **(Fig. 366-1)**. Previously, the classification criteria of the American College of Rheumatology required that 11 of 18 sites be perceived as painful for a diagnosis of FM. In practice, tenderness is a continuous variable, and strict application of a categorical threshold for diagnostic specifics is not necessary. Newer criteria eliminate the need for identification of tender points and focus instead on clinical symptoms of widespread or multi-site pain and neuropsychological symptoms. The newer criteria perform well in a clinical setting in comparison to the older, tender-point criteria. However, it appears that when the new criteria are applied to populations, the result is an increase in prevalence of FM and a change in the sex ratio (see "Epidemiology," earlier).

FIGURE 366-1

Tender-point assessment in patients with fibromyalgia. *(Figure created using data from F Wolfe et al: Arthritis Care Res 62:600, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Patients with FM often have peripheral pain generators that are thought to serve as triggers for the more widespread pain attributed to central nervous system factors. Potential pain generators such as arthritis, bursitis, tendinitis, neuropathies, and other inflammatory or degenerative conditions should be identified by history and physical examination. More subtle pain generators may include joint hypermobility and scoliosis. In addition, patients may have chronic myalgias triggered by infectious, metabolic, or psychiatric conditions that can serve as triggers for the development of FM. These conditions are often identified in the differential diagnosis of patients with FM, and a major challenge is to distinguish the ongoing activity of a triggering condition from FM that is occurring as a consequence of a comorbid condition and that should itself be treated.

Neuropsychological Symptoms

In addition to widespread pain, FM patients typically report fatigue, stiffness, sleep disturbance, cognitive dysfunction, anxiety, and depression. These symptoms are present to varying degrees in most FM patients but are not present in every patient or at all times in a given patient. Relative to pain, such symptoms may,

however, have an equal or even greater impact on function and quality of life. Fatigue is highly prevalent in patients under primary care who ultimately are diagnosed with FM. Pain, stiffness, and fatigue often are worsened by exercise or unaccustomed activity (postexertional malaise). The sleep complaints include difficulty falling asleep, difficulty staying asleep, and early-morning awakening. Regardless of the specific complaint, patients awake feeling unrefreshed. Patients with FM may meet criteria for restless legs syndrome and sleep-disordered breathing; frank sleep apnea can also be documented. Cognitive issues are characterized as difficulties with attention or concentration, problems with word retrieval, and short-term memory loss. Studies have demonstrated altered cognitive function in these domains in patients with FM, though speed of processing is age-appropriate. Symptoms of anxiety and depression are common, and the lifetime prevalence of mood disorders in patients with FM approaches 80%. Although depression is neither necessary nor sufficient for the diagnosis of FM, it is important to screen for major depressive disorders by querying for depressed mood and anhedonia. Analysis of genetic factors that are likely to predispose to FM reveals shared neurobiologic pathways with mood disorders, providing the basis for comorbidity (see later in this chapter).

Overlapping Syndromes

Because FM can overlap in presentation with other chronic pain conditions, review of systems often reveals headaches, facial/jaw pain, regional myofascial pain particularly involving the neck or back, and arthritis. Visceral pain involving the gastrointestinal tract, bladder, and pelvic or perineal region is often present as well. Patients may or may not meet defined criteria for specific syndromes. It is important for patients to understand that shared pathways may mediate symptoms and treatment strategies effective for one condition may help with global symptom management.

Comorbid Conditions

FM is often comorbid with chronic musculoskeletal, infectious, metabolic, or psychiatric conditions. Whereas FM affects only 2–5% of the general population, it occurs in ≥20% of patients with degenerative or inflammatory rheumatic disorders, likely because these conditions serve as peripheral pain generators to alter central pain-processing pathways. Similarly, chronic infectious, metabolic, or psychiatric diseases associated with musculoskeletal pain can mimic FM and/or serve as a trigger for the development of FM. It is particularly important for clinicians to be sensitive to pain management of these comorbid conditions so that when FM emerges—characterized by pain outside the boundaries of what could reasonably be explained by the triggering condition, development of neuropsychological symptoms, or tenderness on physical examination—treatment of central pain processes will be undertaken as opposed to a continued focus on treatment of peripheral or inflammatory causes of pain.

Psychosocial Considerations

Symptoms of FM often have their onset and are exacerbated during periods of perceived stress. This pattern may reflect an interaction among central stress physiology, vigilance or anxiety, and central pain-processing pathways. An understanding of current psychosocial stressors will aid in patient management, as many

factors that exacerbate symptoms cannot be addressed by pharmacologic approaches. Furthermore, there is a high prevalence of exposure to previous interpersonal and other forms of violence in patients with FM and related conditions. If posttraumatic stress disorder is an issue, the clinician should be aware of it and consider treatment options.

Functional Impairment

It is crucial to evaluate the impact of FM symptoms on function and role fulfillment. In defining the success of a management strategy, improved function is a key measure. Functional assessment should include physical, mental, and social domains. Recognition of the ways in which role functioning falls short will be helpful in the establishing treatment goals.

DIFFERENTIAL DIAGNOSIS

Because musculoskeletal pain is such a common complaint, the differential diagnosis of FM is broad. **Table 366-1** lists some of the more common conditions that should be considered. Patients with inflammatory causes for widespread pain should be identifiable on the basis of specific history, physical findings, and laboratory or radiographic tests.

TABLE 366-1

Common Conditions in the Differential Diagnosis of Fibromyalgia

Inflammatory
Polymyalgia rheumatica Inflammatory arthritis: rheumatoid arthritis, spondyloarthritides Connective tissue diseases: systemic lupus erythematosus, Sjögren's syndrome
Infectious
Hepatitis C HIV infection Lyme disease Parvovirus B19 infection Epstein-Barr virus infection
Noninflammatory
Degenerative joint/spine/disk disease Myofascial pain syndromes Bursitis, tendinitis, repetitive strain injuries
Endocrine
Hypo- or hyperthyroidism Hyperparathyroidism
Neurologic Diseases
Multiple sclerosis Neuropathic pain syndromes
Psychiatric Disease
Major depressive disorder
Drugs
Statins Aromatase inhibitors

LABORATORY OR RADIOGRAPHIC TESTING

Routine laboratory and radiographic tests yield normal results in FM. Thus diagnostic testing is focused on exclusion of other diagnoses and evaluation for pain generators or comorbid conditions **(Table 366-2)**. Most patients with new chronic widespread pain should be assessed for the most common entities in the differential diagnosis. Radiographic testing should be used sparingly and only for diagnosis of inflammatory arthritis. After the patient has been evaluated thoroughly, repeat testing is discouraged unless the symptom complex changes. Particularly to be discouraged is magnetic resonance imaging (MRI) of the spine unless there are features suggesting inflammatory spine disease or neurologic symptoms.

TABLE 366-2

Laboratory and Radiographic Testing in Patients with Fibromyalgia Symptoms

Routine
Erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP) Complete blood count (CBC) Thyroid-stimulating hormone (TSH)
Guided by History and Physical Examination
Complete metabolic panel Antinuclear antibody (ANA) Anti-SSA (anti–Sjögren's syndrome A) and anti-SSB Rheumatoid factor and anti–cyclic citrullinated peptide (anti-CCP) Creatine phosphokinase (CPK) Viral (e.g., Hepatitis C, HIV) and bacterial (e.g., Lyme) serologies Spine and joint radiographs

Source: LM Arnold et al: J Women's Health 21:231, 2012; MA Fitzcharles et al: J Rheumatol 40:1388, 2013.

GENETICS AND PHYSIOLOGY

 As in most complex diseases, it is likely that a number of genes contribute to vulnerability to the development of FM. To date, these genes appear to be in pathways controlling pain and stress responses. Some of the genetic underpinnings of FM are shared across other chronic pain conditions. Genes associated

with metabolism, transport, and receptors of serotonin and other monoamines have been implicated in FM and overlapping conditions. Genes associated with other pathways involved in pain transmission have also been described as vulnerability factors for FM. Taken together, the pathways in which polymorphisms have been identified in FM patients further implicate central factors in mediation of the physiology that leads to the clinical manifestations of FM.

Psychophysical testing of patients with FM has demonstrated altered sensory afferent pain processing and impaired descending noxious inhibitory control leading to hyperalgesia and allodynia. Functional MRI and other research imaging procedures clearly demonstrate activation of the brain regions involved in the experience of pain in response to stimuli that are innocuous in study participants without FM. Pain perception in FM patients is influenced by the emotional and cognitive dimensions, such as catastrophizing and perceptions of control, providing a solid basis for recommendations for cognitive and behavioral treatment strategies.

Studies have indicated that some patients meeting criteria for FM may have a small fiber neuropathy. Other studies have identified alterations in expressed gene signatures in peripheral blood. These early studies raise the possibility that confirmatory diagnostic testing could be developed to assist in the diagnosis of FM.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Fibromyalgia

FM is common and has an extraordinary impact on the patient's function and health-related quality of life. Optimal management requires prompt diagnosis and assessment of pain, function, and psychosocial context. Physicians and other health professionals can be helpful in managing some of the symptoms and impact of FM. Developing a partnership with patients is essential for improving the outcome of FM, with a goal of understanding the factors involved, implementing a treatment strategy, and choosing appropriate nonpharmacologic and pharmacologic treatments.

TREATMENT

TREATMENT

Fibromyalgia

NONPHARMACOLOGIC TREATMENT

Patients with chronic pain, fatigue, and other neuropsychological symptoms require a framework for understanding the symptoms that have such an important impact on their function and quality of life. Explaining the genetics, triggers, and physiology of FM can be an important adjunct in relieving associated anxiety and in reducing the overall cost of health care resources. In addition, patients must be educated regarding expectations for treatment. The physician should focus on improved function and quality of life

rather than elimination of pain. Illness behaviors, such as frequent physician visits, should be discouraged and behaviors that focus on improved function strongly encouraged.

Treatment strategies should include physical conditioning, with encouragement to begin at low levels of aerobic exercise and to proceed with slow but consistent advancement. Patients who have been physically inactive or who report postexertional malaise may do best in supervised or water-based programs at the start. Strength training may be recommended after patients reach their aerobic goals. Meditative movement therapies, such as qigong, yoga, or Tai Chi, may also be helpful. Other defined physical therapies such as acupuncture or hydrotherapy may also be considered. Exercise programs are helpful in reducing tenderness and enhancing self-efficacy. Cognitive-behavioral strategies to improve sleep hygiene and reduce illness behaviors can also be helpful in management.

PHARMACOLOGIC APPROACHES

It is essential for the clinician to treat any comorbid triggering condition and to clearly delineate for the patient the treatment goals for each medication. For example, glucocorticoids or nonsteroidal anti-inflammatory drugs may be useful for management of inflammatory triggers but are not effective against FM-related symptoms. At present, the treatment approaches that have proved most successful in FM patients target afferent or descending pain pathways. Table 366-3 lists the drugs with demonstrated effectiveness. It should be emphasized that strong opioid analgesics are to be avoided in patients with FM. These agents have no demonstrated efficacy in FM and are associated with adverse effects that can worsen both symptoms and function. Tramadol, an opioid with mild serotonin-noradrenaline reuptake inhibitor activity has been studied in this population with indication of efficacy. Use of single agents to treat multiple symptom domains is strongly encouraged. For example, if a patient’s symptom complex is dominated by pain and sleep disturbance, use of an agent that exerts both analgesic and sleep-promoting effects is desirable. These agents include cyclobenzaprine, sedating antidepressants such as amitriptyline, and alpha-2-delta ligands such as gabapentin and pregabalin. For patients whose pain is associated with fatigue, anxiety, or depression, drugs that have both analgesic and antidepressant/anxiolytic effects, such as duloxetine or milnacipran, may be the best first choice.

TABLE 366-3

Pharmacologic Agents Effective for Treatment of Fibromyalgia

Muscle relaxant
Cyclobenzaprine
Antidepressants: balanced serotonin–norepinephrine reuptake inhibitors
Amitryptiline[a]
Duloxetine[b,c]
Milnacipran[b,c]
Anticonvulsants: ligand of the alpha-2-delta subunit of voltage-gated calcium channels
Pregabalin[b]
Analgesic
Tramadol

[a]RA Moore et al: Cochrane Database Syst Rev 12:CD008242, 2012. [b]Approved by the U.S. Food and Drug Administration.

[c]W Hauser et al: Cochrane Database Syst Rev 1: CD010292, 2013.

Source: GJ Macfarlane et al: EULAR revised recommendations for the management of fibromyalgia. Ann Rheum Dis 76:318, 2017.

FURTHER READING

Bennett RM et al: Criteria for the diagnosis of fibromyalgia: Validation of the modified 2010 preliminary American College of Rheumatology criteria and the development of alternative criteria. Arthritis Care Res 66:1364, 2014.

Clauw DJ: Fibromyalgia: A clinical review. JAMA 311:1547, 2014. [PubMed: 24737367]

Macfarlane GJ et al: EULAR revised recommendations for the management of fibromyalgia. Ann Rheum Dis 76:318, 2017. [PubMed: 27377815]

Chapter 367: Arthritis Associated with Systemic Disease, and Other Arthritides

Carol A. Langford; Brian F. Mandell

ARTHRITIS ASSOCIATED WITH SYSTEMIC DISEASE

ARTHROPATHY OF ACROMEGALY

Acromegaly is the result of excessive production of growth hormone by an adenoma in the anterior pituitary gland **(Chap. 373)**. The excessive secretion of growth hormone along with insulin-like growth factor I stimulates proliferation of cartilage, periarticular connective tissue, and bone, resulting in several musculoskeletal problems, including osteoarthritis, back pain, muscle weakness, and carpal tunnel syndrome.

Osteoarthritis is a common feature, most often affecting the knees, shoulders, hips, and hands. Single or multiple joints may be affected. Hypertrophy of cartilage initially produces radiographic widening of the joint space. The newly synthesized cartilage is abnormally susceptible to fissuring, ulceration, and destruction. Ligamental laxity of joints further contributes to the development of osteoarthritis. Cartilage degrades, the joint space narrows, and subchondral sclerosis and osteophytes develop. Joint examination reveals crepitus and laxity. Joint fluid is noninflammatory. Calcium pyrophosphate dihydrate crystals are found in the cartilage in some cases of acromegaly arthropathy and, when shed into the joint, can elicit attacks of pseudogout. Chondrocalcinosis may be observed on radiographs. Back pain is extremely common, perhaps as a result of spine hypermobility. Spine radiographs show normal or widened intervertebral disk spaces, hypertrophic anterior osteophytes, and ligamental calcification. The latter changes are similar to those observed in patients with diffuse idiopathic skeletal hyperostosis. Dorsal kyphosis in conjunction with elongation of the ribs contributes to the development of the barrel chest seen in acromegalic patients. The hands and feet become enlarged as a result of soft tissue proliferation. The fingers are thickened and have spadelike distal tufts. One-third of patients have a thickened heel pad. Approximately 25% of patients exhibit Raynaud’s phenomenon. Carpal tunnel syndrome occurs in about half of patients. The median nerve is compressed by excess connective tissue in the carpal tunnel. Patients with acromegaly may develop proximal muscle weakness, which is thought to be caused by the effect of growth hormone on muscle. Serum muscle enzyme levels and electromyographic findings are normal. Muscle biopsy specimens contain muscle fibers of varying size without inflammation.

ARTHROPATHY OF HEMOCHROMATOSIS

Hemochromatosis is a disorder of iron storage. Absorption of excessive amounts of iron from the intestine leads to iron deposition in parenchymal cells, which results in impairment of organ function **(Chap. 407)**. Symptoms of hemochromatosis usually begin between the ages of 40 and 60 but can appear earlier. Arthropathy, which occurs in 20–40% of patients, usually begins after the age of 50 and may be the first clinical feature of hemochromatosis. The arthropathy is an osteoarthritis-like disorder affecting the small joints of the hands and later the larger joints, such as knees, ankles, shoulders, and hips. The second and third metacarpophalangeal joints of both hands are often the first and most

prominent joints affected; this clinical picture may provide an important clue to the possibility of hemochromatosis because these joints are not predominantly affected by “routine” osteoarthritis. Patients experience some morning stiffness and pain with use of involved joints. The affected joints are enlarged and mildly tender. Radiographs show narrowing of the joint space, subchondral sclerosis, subchondral cysts, and juxtaarticular proliferation of bone. Hooklike osteophytes are seen in up to 20% of patients; although they are regarded as a characteristic feature of hemochromatosis, they can also occur in osteoarthritis and are not disease specific. The synovial fluid is noninflammatory. The synovium shows mild to moderate proliferation of iron-containing lining cells, fibrosis, and some mononuclear cell infiltration. In approximately half of patients, there is evidence of calcium pyrophosphate deposition disease, and some patients late in the course of disease experience episodes of acute pseudogout **(Chap. 365)**. An early diagnosis is suggested by high serum transferrin saturation, which is more sensitive than ferritin elevation.

Iron may damage the articular cartilage in several ways. Iron catalyzes superoxide-dependent lipid peroxidation, which may play a role in joint damage. In animal models, ferric iron has been shown to interfere with collagen formation and increase the release of lysosomal enzymes from cells in the synovial membrane. Iron inhibits synovial tissue pyrophosphatase in vitro and therefore may inhibit pyrophosphatase in vivo, resulting in chondrocalcinosis.

TREATMENT

TREATMENT Arthropathy of Hemochromatosis

The treatment of hemochromatosis is repeated phlebotomy. Unfortunately, this treatment has little effect on established arthritis, which, along with chondrocalcinosis, may progress. Symptom-based treatment of the arthritis consists of administration of acetaminophen and nonsteroidal anti-inflammatory drugs (NSAIDs), as tolerated. Acute pseudogout attacks are treated with high doses of an NSAID or a short course of glucocorticoids. Hip or knee total joint replacement has been successful in advanced disease.

HEMOPHILIC ARTHROPATHY

Hemophilia is a sex-linked recessive genetic disorder characterized by the absence or deficiency of factor VIII (hemophilia A, or classic hemophilia) or factor IX (hemophilia B, or Christmas disease) **(Chap. 112)**. Hemophilia A constitutes 85% of cases. Spontaneous hemarthrosis is a common problem with both types of hemophilia and can lead to a deforming arthritis. The frequency and severity of hemarthrosis are related to the degree of clotting factor deficiency. Hemarthrosis is not common in other disorders of coagulation such as von Willebrand disease, factor V deficiency, warfarin therapy, or thrombocytopenia.

Hemarthrosis occurs after 1 year of age, when a child begins to walk and run. In order of frequency, the joints most commonly affected are the knees, ankles, elbows, shoulders, and hips. Small joints of the hands and feet are occasionally involved.

In the initial stage of arthropathy, hemarthrosis produces a warm, tensely swollen, and painful joint. The patient holds the affected joint in flexion and guards against any movement. Blood in the joint remains liquid because of the absence of intrinsic clotting factors and the absence of tissue thromboplastin in the synovium. The synovial blood is resorbed over a period of ≥1 week, with the precise interval depending on the size of the hemarthrosis. Joint function usually returns to normal or baseline in ~2 weeks. Low-grade temperature elevation may accompany hemarthrosis, but a fever >38.3°C (101°F) warrants concern about infection.

Recurrent hemarthrosis may result in chronic arthritis. The involved joints remain swollen, and flexion deformities develop. Joint motion may be restricted and function severely limited. Restricted joint

motion or laxity with subluxation is a feature of end-stage disease.

Bleeding into muscle and soft tissue also causes musculoskeletal dysfunction. When bleeding into the iliopsoas muscle occurs, the hip is held in flexion because of the pain, resulting in a hip flexion contracture. Rotation of the hip is preserved, which distinguishes this problem from hemarthrosis or other causes of hip synovitis. Expansion of the hematoma may place pressure on the femoral nerve, resulting in femoral neuropathy. Hemorrhage into a closed compartment space, such as the calf or the volar compartment in the forearm, can result in muscle necrosis, neuropathy, and flexion deformities of the ankles, wrists, and fingers. When bleeding involves periosteum or bone, a painful pseudotumor forms. These pseudotumors occur distal to the elbows or knees in children and improve with treatment of hemophilia. Surgical removal is indicated if the pseudotumor continues to enlarge. In adults, pseudotumors develop in the femur and pelvis and are usually refractory to treatment. When bleeding occurs in muscle, cysts may develop within the muscle. Needle aspiration of a cyst is contraindicated because this procedure can induce further bleeding; however, if the cyst becomes secondarily infected, drainage may be necessary (after factor repletion).

Septic arthritis is rare in hemophilia and is difficult to distinguish from acute hemarthrosis on physical examination. If there is serious suspicion of an infected joint, the joint should be aspirated immediately, the fluid cultured, and treatment with broad-spectrum antibiotics administered, with coverage for microorganisms including *Staphylococcus*, until culture results become available. Clotting-factor deficiency should be corrected before arthrocentesis to minimize the risk of traumatic bleeding.

Radiographs of joints reflect the stage of disease. In early stages, there is only capsule distention; later, juxtaarticular osteopenia, marginal erosions, and subchondral cysts develop. Late in the disease, the joint space is narrowed and there is bony overgrowth similar to that in osteoarthritis.

TREATMENT

TREATMENT Hemarthrosis

The treatment of musculoskeletal bleeding is initiated with the immediate infusion of factor VIII or IX at the first sign of joint or muscle hemorrhage. Patients who have developed factor inhibitors are at elevated risk for joint damage and may benefit from receiving recombinant activated factor VII or activated prothrombin complex concentrate. The joint should be rested in a position of forced extension, as tolerated, to avoid contracture. Analgesia should be provided; nonselective NSAIDs, which can diminish platelet function, should be avoided if possible. Selective cyclooxygenase-2 inhibitors do not interfere with platelet function, although cardiovascular and gastrointestinal risks must still be weighed. Synovectomy—open or arthroscopic—may be attempted in patients with chronic symptomatic synovial proliferation and recurrent hemarthrosis, although hypertrophied synovium is highly vascular and subject to bleeding. Both types of synovectomy reduce the number of hemarthroses. Open surgical synovectomy, however, is associated with some loss of range of motion. Both require aggressive prophylaxis against bleeding. Radiosynovectomy with either yttrium 90 silicate or phosphorus 31 colloid has been effective and may be attempted when surgical synovectomy is not practical. Total joint replacement is indicated for severe joint destruction and incapacitating pain.

ARTHROPATHIES ASSOCIATED WITH HEMOGLOBINOPATHIES

Sickle Cell Disease

Sickle cell disease **(Chap. 94)** is associated with several musculoskeletal abnormalities **(Table 367-1)**. Children aged <5 years may develop diffuse swelling, tenderness, and warmth of the hands and feet lasting 1–3 weeks. This condition, referred to as *sickle cell dactylitis* or *hand-foot syndrome*, has also

been observed in sickle cell thalassemia. Dactylitis is believed to result from infarction of the bone marrow and cortical bone leading to periostitis and soft tissue swelling. Radiographs show periosteal elevation, subperiosteal new-bone formation, and areas of radiolucency and increased density involving the metacarpals, metatarsals, and proximal phalanges. These bone changes disappear after several months. The syndrome leaves little or no residual damage. Because hematopoiesis ceases in the small bones of the hands and feet with age, the syndrome is rarely seen after age 5.

TABLE 367-1 Musculoskeletal Abnormalities in Sickle Cell Disease

Sickle cell dactylitis	
	Avascular necrosis
Joint effusions in sickle cell crises	
	Bone changes secondary to marrow hyperplasia
Osteomyelitis	
	Septic arthritis
Infarction of bone	
	Gouty arthritis
Infarction of bone marrow	

Sickle cell crisis is associated with periarticular pain and occasionally with joint effusions. The joint and periarticular area are warm and tender. Knees and elbows are most often affected, but other joints can be involved. Joint effusions are usually noninflammatory. Acute synovial infarction can cause a sterile effusion with high neutrophil counts in synovial fluid. Synovial biopsies have shown mild lining-cell proliferation and microvascular thrombosis with infarctions. Scintigraphic studies have shown decreased marrow uptake adjacent to the involved joint. The treatment for sickle cell crisis is detailed in **Chap. 94**.

Patients with sickle cell disease seem predisposed to osteomyelitis, which commonly involves the long tubular bones **(Chap. 126)**; *Salmonella* is a particularly common cause **(Chap. 160)**. Radiographs of the involved site initially show periosteal elevation, with subsequent disruption of the cortex. Treatment of the infection results in healing of the bone lesion. In addition, sickle cell disease is associated with bone infarction resulting from vasoocclusion secondary to the sickling of red cells. Bone infarction also occurs in hemoglobin sickle cell disease and sickle cell thalassemia **(Chap. 94)**. The bone pain in sickle cell crisis is due to infarction of bone and bone marrow. In children, infarction of the epiphyseal growth plate interferes with normal growth of the affected extremity. Radiographically, infarction of the bone cortex results in periosteal elevation and irregular thickening of the bone cortex. Infarction in the bone marrow leads to lysis, fibrosis, and new bone formation. Clinical distinction between osteomyelitis and bone infarctions can be difficult; imaging can be helpful.

Avascular necrosis of the head of the femur occurs in ~5% of patients. It also occurs in the humeral head and less commonly in the distal femur, tibial condyles, distal radius, vertebral bodies, and other juxtaarticular sites. Irregularity of the femoral head and other articular surfaces often results in degenerative joint disease. Radiography of the affected joint may show patchy radiolucency and density followed by flattening of the bone. MRI is a sensitive technique for detecting early avascular necrosis as well as bone infarction elsewhere. Total hip replacement and placement of prostheses in other joints may improve function and relieve joint pain in these patients.

Septic arthritis is occasionally encountered in sickle cell disease **(Chap. 125)**. Multiple joints may be infected. Joint infection may result from bacteremia due to splenic dysfunction or from contiguous osteomyelitis. The more common microorganisms include *Staphylococcus aureus*, *Streptococcus*, and *Salmonella*. *Salmonella* does not cause septic arthritis as frequently as it causes osteomyelitis. Acute gouty arthritis is uncommon in sickle cell disease, even though 40% of patients are hyperuricemic. However, it may occur in patients generally not expected to get gout (young patients, female patients). Hyperuricemia is due to overproduction of uric acid secondary to increased red cell turnover as well as suboptimal renal excretion. Attacks may be polyarticular, and diagnostic arthrocentesis should be performed to distinguish infection from gout or synovial infarction.

The bone marrow hyperplasia in sickle cell disease results in widening of the medullary cavities, thinning of the cortices, and coarse trabeculations and central cupping of the vertebral bodies. These changes are also seen to a lesser degree in hemoglobin sickle cell disease and sickle cell thalassemia. In normal individuals red marrow is located mostly in the axial skeleton, but in sickle cell disease red marrow is found in the bones of the extremities and even in the tarsal and carpal bones. Vertebral compression may lead to dorsal kyphosis, and softening of the bone in the acetabulum may result in protrusio acetabuli.

Thalassemia

A congenital disorder of hemoglobin synthesis, β thalassemia is characterized by impaired production of β chains **(Chap. 94)**. Bone and joint abnormalities occur in β thalassemia, being most common in the major and intermedia groups. In one study, ~50% of patients with β thalassemia had evidence of symmetric ankle arthropathy characterized by a dull aching pain that was aggravated by weight bearing. The onset came most often in the second or third decade of life. The degree of ankle pain in these patients varied. Some patients experienced self-limited ankle pain that occurred only after strenuous physical activity and lasted several days or weeks. Other patients had chronic ankle pain that became worse with walking. Symptoms eventually abated in a few patients. Compression of the ankle, calcaneus, or forefoot was painful in some patients. Synovial fluid from two patients was noninflammatory. Radiographs of the ankle showed osteopenia, widened medullary spaces, thin cortices, and coarse trabeculations—findings that are largely the result of bone marrow expansion. The joint space was preserved. Specimens of bone from three patients revealed osteomalacia, osteopenia, and microfractures. Increased numbers of osteoblasts as well as increased foci of bone resorption were present on the bone surface. Iron staining was found in the bone trabeculae, in osteoid, and in the cement line. Synovium showed hyperplasia of lining cells, which contained deposits of hemosiderin. This arthropathy was considered to be related to the underlying bone pathology. The role of iron overload or abnormal bone metabolism in the pathogenesis of this arthropathy is not known. The arthropathy was treated with analgesics and splints. Patients also received transfusions to decrease hematopoiesis and bone marrow expansion.

In patients with β-thalassemia major and β-thalassemia intermedia, other joints are also involved, including the knees, hips, and shoulders. Acquired hemochromatosis with arthropathy has been described in a patient with thalassemia. Gouty arthritis and septic arthritis can occur. Avascular necrosis is not a feature of thalassemia because there is no sickling of red cells leading to thrombosis and infarction.

β-Thalassemia minor (also known as *β-thalassemia trait*) is likewise associated with joint manifestations. Chronic seronegative oligoarthritis affecting predominantly ankles, wrists, and elbows has been described; the affected patients had mild persistent synovitis without large effusions or joint erosions. Recurrent episodes of acute asymmetric arthritis have also been reported; episodes last <1 week and may affect the knees, ankles, shoulders, elbows, wrists, and metacarpal phalangeal joints. The mechanism underlying this arthropathy is unknown. Treatment with NSAIDs is not particularly effective.

MUSCULOSKELETAL DISORDERS ASSOCIATED WITH HYPERLIPIDEMIA

(See also Chap. 400) Musculoskeletal or cutaneous manifestations may be the first clinical indication of a specific hereditary disorder of lipoprotein metabolism. Patients with familial hypercholesterolemia (previously referred to as *type II hyperlipoproteinemia*) may have recurrent migratory polyarthritis involving the knees and other large peripheral joints and, to a lesser degree, peripheral small joints. Pain ranges from moderate to incapacitating. The involved joints can be warm, erythematous, swollen, and tender. Arthritis usually has a sudden onset, lasts from a few days to 2 weeks, and does not cause joint damage. Episodes may suggest acute gout attacks. Several attacks occur per year. Synovial fluid from involved joints is not inflammatory and contains few white cells and no crystals.

Joint involvement may actually represent inflammatory periarthritis or peritendinitis and not true arthritis. The recurrent, transient nature of the arthritis may suggest rheumatic fever, especially because patients with hyperlipoproteinemia may have an elevated erythrocyte sedimentation rate and elevated antistreptolysin O titers (the latter being quite common). Attacks of tendinitis, including the large Achilles and patellar tendons, may come on gradually and last only a few days or may be acute as described above. Patients may be asymptomatic between attacks. Achilles tendinitis and other joint manifestations often precede the appearance of xanthomas and may be the first clinical indication of hyperlipoproteinemia. Attacks of tendinitis may follow treatment with a lipid-lowering drug. Over time, patients may develop tendinous xanthomas in the Achilles, patellar, and extensor tendons of the hands and feet. Xanthomas have also been reported in the peroneal tendon, the plantar aponeurosis, and the periosteum overlying the distal tibia. These xanthomas are located within tendon fibers. Tuberous xanthomas are soft subcutaneous masses located over the extensor surfaces of the elbows, knees, and hands as well as on the buttocks. They appear during childhood in homozygous patients and after the age of 30 in heterozygous patients. Patients with elevated plasma levels of very-low-density lipoprotein (VLDL) and triglycerides (previously referred to as *type IV hyperlipoproteinemia*) may also have a mild inflammatory arthritis affecting large and small peripheral joints, usually in an asymmetric pattern, with only a few joints involved at a time. The onset of arthritis usually comes in middle age. Arthritis may be persistent or recurrent, with episodes lasting a few days or weeks. Some patients may experience severe joint pain or morning stiffness. Joint tenderness and periarticular hyperesthesia may also be present, as may synovial thickening. Joint fluid is usually noninflammatory and without crystals but may have increased white blood cell counts with predominantly mononuclear cells. Radiographs may show juxtaarticular osteopenia and cystic lesions. Large bone cysts have been noted in a few patients. Xanthoma and bone cysts are also observed in other lipoprotein disorders. The pathogenesis of arthritis in patients with familial hypercholesterolemia or with elevated levels of VLDL and triglycerides is not well understood. NSAIDs or analgesics usually provide adequate relief of symptoms when used on an as-needed basis.

Patients may improve clinically as they are treated with lipid-lowering agents; however, patients treated with an HMG-CoA reductase inhibitor may experience myalgias, and a few patients develop myopathy, myositis, or even rhabdomyolysis. Patients who develop myositis during statin therapy may be susceptible to this adverse effect because of an underlying muscle disorder and should be reevaluated after discontinuation of the drug. Testing for anti-HMGCR autoantibodies in patients with elevated muscle enzymes on treatment may identify patients with statin-induced necrotizing autoimmune myopathy. Myositis has also been reported with the use of niacin **(Chap. 358)** but is less common than myalgias.

Musculoskeletal syndromes have not clearly been associated with the more common mixed hyperlipidemias seen in general practice.

OTHER ARTHRITIDES

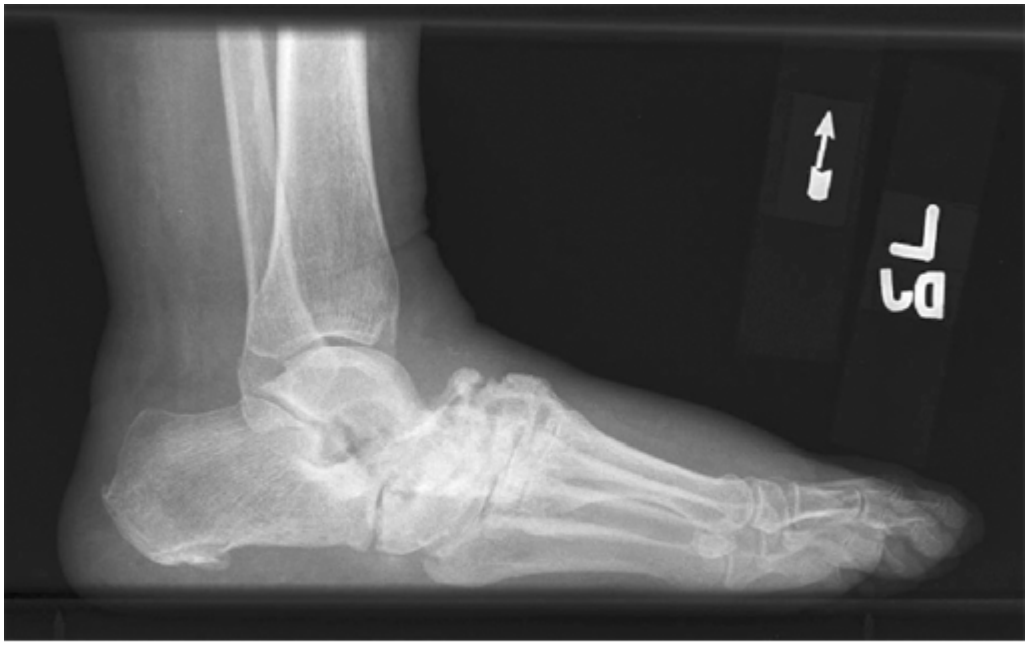
NEUROPATHIC JOINT DISEASE

Neuropathic joint disease (Charcot joint) is a progressive destructive arthritis associated with loss of pain sensation, proprioception, or both. Normal muscular reflexes that modulate joint movement are impaired. Without these protective mechanisms, joints are subjected to repeated trauma, resulting in progressive cartilage and bone damage. Today, diabetes mellitus is the most frequent cause of neuropathic joint disease **(Fig. 367-1)**. A variety of other disorders are associated with neuropathic arthritis, including tabes dorsalis, leprosy, yaws, syringomyelia, meningomyelocele, congenital indifference to pain, peroneal muscular atrophy (Charcot-Marie-Tooth disease), and amyloidosis. An arthritis resembling neuropathic joint disease has been reported in patients who have received intraarticular glucocorticoid injections, but this is a rare complication and was not observed in one series of patients with knee osteoarthritis who received intraarticular glucocorticoid injections every 3

months for 2 years. The distribution of joint involvement depends on the underlying neurologic disorder **(Table 367-2)**. In tabes dorsalis, the knees, hips, and ankles are most commonly affected; in syringomyelia, the glenohumeral joint, elbow, and wrist; and in diabetes mellitus, the tarsal and tarsometatarsal joints.

FIGURE 367-1

Charcot arthropathy associated with diabetes mellitus. Lateral foot radiograph demonstrating complete loss of the arch due to bony fragmentation and dislocation in the midfoot. *(Courtesy of Andrew Neckers, MD, and Jean Schils, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 367-2 Disorders Associated with Neuropathic Joint Disease

Diabetes mellitus	Amyloidosis
Tabes dorsalis	Leprosy
Meningomyelocele	Congenital indifference to pain
Syringomyelia	Peroneal muscular atrophy

PATHOLOGY AND PATHOPHYSIOLOGY

The pathologic changes in the neuropathic joint are similar to those found in the severe osteoarthritic joint. There is fragmentation and eventual loss of articular cartilage with eburnation of the underlying bone. Osteophytes are found at the joint margins. With more advanced disease, erosions are present on the joint surface. Fractures, devitalized bone, intraarticular loose bodies, and microscopic fragments of cartilage and bone may be present.

At least two underlying mechanisms are believed to be involved in the pathogenesis of neuropathic arthritis. An abnormal autonomic nervous system is thought to be responsible for the dysregulated blood flow to the joint with subsequent resorption of bone. Loss of bone, particularly in the diabetic foot, may be the initial finding. With the loss of deep pain, proprioception, and protective neuromuscular reflexes, the joint is subjected to repeated microtrauma, resulting in ligamental tears

and bone fractures. The injury that follows frequent intraarticular glucocorticoid injections is thought to be due to the analgesic effect of glucocorticoids, leading to overuse of an already damaged joint; the result is accelerated cartilage damage, although steroid-induced cartilage damage is more common in some other animal species than in humans. It is not understood why only a few patients with neuropathy develop clinically evident neuropathic arthritis.

CLINICAL MANIFESTATIONS

Neuropathic joint disease usually begins in a single joint and then becomes apparent in other joints, depending on the underlying neurologic disorder. The involved joint becomes progressively enlarged as a result of bony overgrowth and synovial effusion. Loose bodies may be palpated in the joint cavity. Joint instability, subluxation, and crepitus occur as the disease progresses. Neuropathic joints may develop rapidly, and a totally disorganized joint with multiple bony fragments may evolve within weeks or months. The amount of pain experienced by the patient is less than would be anticipated from the degree of joint damage. Patients may experience sudden joint pain from intraarticular fractures of osteophytes or condyles.

Neuropathic arthritis is encountered most often in patients with diabetes mellitus, with an incidence of ~0.5%. The onset of disease usually comes at an age of ≥50 years in a patient who has had diabetes for several years, but exceptions occur. The tarsal and tarsometatarsal joints are most often affected, with the metatarsophalangeal and talotibial joints next most commonly involved. The knees and spine are occasionally involved. Patients often attribute the onset of foot pain to antecedent trauma such as twisting of the foot. Neuropathic changes may develop rapidly after a foot fracture or dislocation. The foot and ankle are often swollen. Downward collapse of the tarsal bones leads to convexity of the sole, referred to as a “rocker foot.” Large osteophytes may protrude from the top of the foot. Calluses frequently form over the metatarsal heads and may lead to infected ulcers and osteomyelitis. The value of protective inserts and orthotics, as well as regular foot examination, cannot be overstated. Radiographs may show resorption and tapering of the distal metatarsal bones. The term *Lisfranc fracture-dislocation* is sometimes used to describe the destructive changes at the tarsometatarsal joints.

DIAGNOSIS

The diagnosis of neuropathic arthritis is based on the clinical features and characteristic radiographic findings in a patient with underlying sensory neuropathy. The differential diagnosis of neuropathic arthritis depends upon the severity of the process and includes osteomyelitis, avascular necrosis, advanced osteoarthritis, stress fractures, and calcium pyrophosphate deposition disease. Radiographs in neuropathic arthritis initially show changes of osteoarthritis with joint space narrowing, subchondral bone sclerosis, osteophytes, and joint effusions; marked destructive and hypertrophic changes follow later. The radiographic findings of neuropathic arthritis may be difficult to differentiate from those of osteomyelitis, especially in the diabetic foot. The joint margins in a neuropathic joint tend to be distinct, while in osteomyelitis they are blurred. Imaging studies may be helpful, but cultures of tissue from the joint are often required to exclude osteomyelitis. MRI and bone scans using indium 111–labeled white blood cells or indium 111–labeled immunoglobulin G, which will show increased uptake in osteomyelitis but not in a neuropathic joint, may be useful. A technetium bone scan will not distinguish osteomyelitis from neuropathic arthritis, as increased uptake is observed in both. The joint fluid in neuropathic arthritis is noninflammatory; it may be xanthochromic or even bloody and may contain fragments of synovium, cartilage, and bone. The finding of calcium pyrophosphate dihydrate crystals supports the diagnosis of crystal-associated arthropathy. In the absence of such crystals, an increased number of leukocytes may indicate osteomyelitis.

TREATMENT

TREATMENT Neuropathic Joint Disease

The primary focus of treatment is to stabilize the joint. Treatment of the underlying disorder, even if successful, does not usually affect established joint disease. Braces and splints are helpful. Their use requires close surveillance, because patients may be unable to appreciate pressure from a poorly adjusted brace. In the diabetic patient, early recognition of Charcot foot and its treatment—prohibition of weight bearing by the foot for at least 8 weeks—may possibly prevent severe disease from developing. Fusion of an unstable joint may improve function and reduce pain, but nonunion is frequent, especially when immobilization of the joint is inadequate.

HYPERTROPHIC OSTEOARTHROPATHY AND CLUBBING

Hypertrophic osteoarthropathy (HOA) is characterized by clubbing of digits and, in more advanced stages, by periosteal new-bone formation and synovial effusions. HOA may be primary or familial and may begin in childhood. Secondary HOA is associated with intrathoracic malignancies, suppurative and some hypoxemic lung diseases, congenital heart disease, and a variety of other disorders. Clubbing is almost always a feature of HOA but can occur as an isolated manifestation **(Fig. 367-2)**. The presence of clubbing in isolation may be congenital or represent either an early stage or one element in the spectrum of HOA. Isolated acquired clubbing has the same clinical significance as clubbing associated with periostitis.

FIGURE 367-2

Clubbing of the fingers. *(Reprinted from the Clinical Slide Collection on the Rheumatic Diseases, © 2018 American College of Rheumatology. Used by permission of the American College of Rheumatology.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pathology and Pathophysiology of Acquired HOA

In HOA, bone changes in the distal extremities begin as periostitis followed by new bone formation. At this stage, a radiolucent area may be observed between the new periosteal bone and the subjacent cortex. As the process progresses, multiple layers of new bone are deposited and become contiguous

with the cortex, with consequent cortical thickening. The outer portion of the bone is laminated in appearance, with an irregular surface. Initially, the process of periosteal new-bone formation involves the proximal and distal diaphyses of the tibia, fibula, radius, and ulna and, less frequently, the femur, humerus, metacarpals, metatarsals, and phalanges. Occasionally, scapulae, clavicles, ribs, and pelvic bones are also affected. The adjacent interosseous membranes may become ossified. The distribution of bone manifestations is usually bilateral and symmetric. The soft tissue overlying the distal third of the arms and legs may be thickened. Proliferation of connective tissue occurs in the nail bed and volar pad of digits, giving the distal phalanges a clubbed appearance. Small blood vessels in the clubbed digits are dilated and have thickened walls. In addition, the number of arteriovenous anastomoses is increased.

Several theories have been suggested for the pathogenesis of HOA, but many have been disproved or have not explained the condition's development in all clinical disorders with which it is associated. Previously proposed neurogenic and humoral theories are no longer considered likely explanations for HOA. Studies have suggested a role for platelets in the development of HOA. It has been observed that megakaryocytes and large platelet particles present in the venous circulation are fragmented in their passage through normal lung. In patients with cyanotic congenital heart disease and in other disorders associated with right-to-left shunts, these large platelet particles bypass the lung and reach the distal extremities, where they can interact with endothelial cells. Platelet–endothelial cell activation in the distal portion of the extremities may result in the release of platelet-derived growth factor (PDGF) and other factors leading to the proliferation of connective tissue and periosteum. Stimulation of fibroblasts by PDGF and transforming growth factor β results in cell growth and collagen synthesis. Elevated plasma levels of von Willebrand factor antigen have been found in patients with both primary and secondary forms of HOA, indicating endothelial activation or damage. Abnormalities of collagen synthesis have been demonstrated in the involved skin of patients with primary HOA. Other factors are undoubtedly involved in the pathogenesis of HOA, and further studies are needed to elucidate this disorder.

Clinical Manifestations

Primary or familial HOA, also referred to as *pachydermoperiostitis* or *Touraine-Solente-Golé syndrome*, usually begins insidiously at puberty. In a smaller proportion of patients, the onset comes in the first year of life. The disorder is inherited as an autosomal dominant trait with variable expression and is nine times more common among boys than among girls. Approximately one-third of patients have a family history of primary HOA.

Primary HOA is characterized by clubbing, periostitis, and unusual skin features. A small number of patients with this syndrome do not express clubbing. The skin changes and periostitis are prominent features of this syndrome. The skin becomes thickened and coarse. Deep nasolabial folds develop, and the forehead may become furrowed. Patients may have heavy-appearing eyelids and ptosis. The skin is often greasy, and there may be excessive sweating of the hands and feet. Patients may also experience acne vulgaris, seborrhea, and folliculitis. In a few patients, the skin over the scalp becomes very thick and corrugated, a feature that has been descriptively termed *cutis verticis gyrata*. The distal extremities, particularly the legs, become thickened as a consequence of the proliferation of new bone and soft tissue; when the process is extensive, the distal lower extremities resemble those of an elephant. The periostitis usually is not painful, which it can be in secondary HOA. Clubbing of the fingers may be extensive, producing large, bulbous deformities, and clumsiness. Clubbing also affects the toes. Patients may experience articular and periarticular pain, especially in the ankles and knees, and joint motion may be mildly restricted by periarticular bone overgrowth. Noninflammatory effusions occur in the wrists, knees, and ankles. Synovial hypertrophy is not found. Associated abnormalities observed in patients with primary HOA include hypertrophic gastropathy, bone marrow failure, female escutcheon, gynecomastia, and cranial suture defects. In patients with primary HOA, the symptoms disappear when adulthood is reached.

HOA secondary to an underlying disease occurs more frequently than primary HOA. It accompanies a variety of disorders and may precede clinical features of the associated disorder by months. Clubbing is more frequent than the full syndrome of HOA in patients with associated illnesses. Because clubbing evolves over months and is usually asymptomatic, it is often recognized first by the physician and not the patient. Patients may experience a burning sensation in their fingertips. Clubbing is characterized by widening of the fingertips, enlargement of the distal volar pad, convexity of the nail contour, and the loss of the normal 15° angle between the proximal nail and cuticle. The thickness of the digit at the base of the nail is greater than the thickness at the distal interphalangeal joint. An objective measurement of finger clubbing can be made by determining the diameter at the base of the nail and at the distal interphalangeal joint of all 10 digits. Clubbing is present when the sum of the individual digit ratios is >10. At the bedside, clubbing can be appreciated by having the patient place the dorsal surface of the distal phalanges of the fourth fingers together with the nails opposing each other. Normally, an open area is visible between the bases of the opposing fingernails; when clubbing is present, this open space is no longer visible. The base of the nail feels spongy when compressed, and the nail can be easily rocked on its bed. When clubbing is advanced, the finger may have a drumstick appearance, and the distal interphalangeal joint can be hyperextended. Periosteal involvement in the distal extremities may produce a burning or deep-seated aching pain. The pain, which can be quite incapacitating, is aggravated by dependency and relieved by elevation of the affected limbs. Pressure applied over the distal forearms and legs or gentle percussion of distal long bones like the tibia may be quite painful.

Patients may experience joint pain, most often in the ankles, wrists, and knees. Joint effusions may be present; usually, they are small and noninflammatory. The small joints of the hands are rarely affected. Severe joint or long bone pain may be the presenting symptom of an underlying lung malignancy and may precede the appearance of clubbing. In addition, the progression of HOA tends to be more rapid when associated with malignancies, most notably bronchogenic carcinoma. Noninflammatory but variably painful knee effusions may occur prior to the appearance of clubbing and symptoms of distal periostitis. Unlike primary HOA, secondary HOA does not commonly include excessive sweating and oiliness of the skin or thickening of the facial skin.

HOA occurs in 5–10% of patients with intrathoracic malignancies, the most common being bronchogenic carcinoma and pleural tumors **(Table 367-3)**. Lung metastases infrequently cause HOA. HOA is also seen in patients with intrathoracic infections, including lung abscesses, empyema, and bronchiectasis, but is uncommon in pulmonary tuberculosis. HOA may accompany chronic interstitial pneumonitis, sarcoidosis, and cystic fibrosis. In cystic fibrosis, clubbing is more common than the full syndrome of HOA. Other causes of clubbing include congenital heart disease with right-to-left shunts, bacterial endocarditis, Crohn’s disease, ulcerative colitis, sprue, and neoplasms of the esophagus, liver, and small and large bowel. In patients who have congenital heart disease with right-to-left shunts, clubbing alone occurs more often than the full syndrome of HOA.

TABLE 367-3 Disorders Associated with Hypertrophic Osteoarthropathy

Pulmonary
Bronchogenic carcinoma and other neoplasms
Lung abscesses, empyema, bronchiectasis
Chronic interstitial pneumonitis
Cystic fibrosis
Sarcoidosis
Gastrointestinal
Inflammatory bowel disease
Sprue
Neoplasms: esophagus, liver, bowel
Cardiovascular
Cyanotic congenital heart disease
Subacute bacterial endocarditis
Infected arterial grafts[a]
Aortic aneurysm[b]
Aneurysm of major extremity artery[a]
Patent ductus arteriosus[b]
Arteriovenous fistula of major extremity vessel[a]
Thyroid (thyroid acropachy)
Hyperthyroidism (Graves' disease)

[a]Unilateral involvement. [b]Bilateral lower-extremity involvement.

Unilateral clubbing has been found in association with aneurysms of major extremity arteries, with infected arterial grafts, and with arteriovenous fistulas of brachial vessels. Clubbing of the toes but not the fingers has been associated with an infected abdominal aortic aneurysm and patent ductus arteriosus. Clubbing of a single digit may follow trauma and has been reported in tophaceous gout and sarcoidosis. While clubbing occurs more commonly than the full syndrome in most diseases, periostitis in the absence of clubbing has been observed in the affected limb of patients with infected arterial grafts.

Hyperthyroidism (Graves' disease), treated or untreated, is occasionally associated with clubbing and periostitis of the bones of the hands and feet. This condition is referred to as *thyroid acropachy*. Periostitis may be asymptomatic and occurs in the midshaft and diaphyseal portion of the metacarpal and phalangeal bones. Significant hand-joint pain may occur; this pain may respond to successful

therapy for thyroid dysfunction. The long bones of the extremities are seldom affected. Elevated levels of long-acting thyroid stimulator are found in the sera of these patients.

Laboratory Findings

The laboratory abnormalities reflect the underlying disorder. The synovial fluid of involved joints has <500 white cells/μL, and the cells are predominantly mononuclear. Radiographs show a faint radiolucent line beneath the new periosteal bone along the shaft of long bones at their distal end. These changes are observed most frequently at the ankles, wrists, and knees. The ends of the distal phalanges may show osseous resorption. Radionuclide studies show pericortical linear uptake along the cortical margins of long bones that may precede any radiographic changes.

TREATMENT

TREATMENT Hypertrophic Osteoarthropathy

The treatment of HOA aims to identify the associated disorder and treat it appropriately. The symptoms and signs of HOA may disappear completely with removal of or effective chemotherapy for a tumor or with antibiotic therapy for a chronic pulmonary infection and drainage of the infected site. Vagotomy or percutaneous block of the vagus nerve leads to symptomatic relief in some patients. NSAIDs or analgesics may help control symptoms of HOA.

REFLEX SYMPATHETIC DYSTROPHY SYNDROME

The reflex sympathetic dystrophy syndrome is now referred to as *complex regional pain syndrome, type 1*, according to the new classification system of the International Association for the Study of Pain. This syndrome is characterized by pain and swelling, usually of a distal extremity, accompanied by vasomotor instability, trophic skin changes, and the rapid development of bony demineralization. Reflex sympathetic dystrophy syndrome, including its treatment, is covered in greater detail in **Chap. 432**.

TIETZE SYNDROME AND COSTOCHONDRITIS

Tietze syndrome is manifested by painful swelling of one or more costochondral articulations. The age of onset is usually before 40, and both sexes are affected equally. In most patients, only one joint is involved, usually the second or third costochondral joint. The onset of anterior chest pain may be sudden or gradual. The pain may radiate to the arms or shoulders and is aggravated by sneezing, coughing, deep inspirations, or twisting motions of the chest. The term *costochondritis* is often used interchangeably with *Tietze syndrome*, but some workers restrict the former term to pain of the costochondral articulations without swelling. Costochondritis is observed in patients aged >40 years; it tends to affect the third, fourth, and fifth costochondral joints, and occurs more often in women. Both syndromes may mimic cardiac or upper abdominal causes of pain. Rheumatoid arthritis, ankylosing spondylitis, and reactive arthritis may involve costochondral joints but are distinguished easily by their other clinical features. Other skeletal causes of anterior chest wall pain are xiphoidalgia and the slipping rib syndrome, which usually involves the tenth rib. Malignancies such as breast cancer, prostate cancer, plasma cell cytoma, and sarcoma can invade the ribs, thoracic spine, or chest wall and produce symptoms suggesting Tietze's syndrome. Patients with osteomalacia may have significant rib pain, with or without documented microfractures. These conditions should be distinguishable by radiography, bone scanning, vitamin D measurement, or biopsy. Analgesics, anti-inflammatory drugs, and local glucocorticoid injections usually relieve symptoms of costochondritis/Tietze's syndrome. Care should be taken to avoid overdiagnosing these syndromes in patients with acute chest pain syndromes; many patients will be tender to overly vigorous palpation of the costochondral joints.

MYOFASCIAL PAIN SYNDROME

Myofascial pain syndrome is characterized by multiple areas of localized musculoskeletal pain and tenderness in association with tender points. The pain is deep and aching and may be accompanied by a burning sensation. Myofascial pain may be regional and follow trauma, overuse, or prolonged static contraction of a muscle or muscle group, which may occur when an individual is reading or writing at a desk or working at a computer. In addition, this syndrome may be associated with underlying osteoarthritis of the neck or low back. Pain may be referred from tender points to defined areas distant from the area of original tenderness. Palpation of the tender point reproduces or accentuates the pain. The tender points are usually located in the center of a muscle belly, but they can occur at other sites such as costosternal junctions, the xiphoid process, ligamentous and tendinous insertions, fascia, and fatty areas. Tender point sites in muscle have been described as feeling indurated and taut, and palpation may cause the muscle to twitch. These findings, however, have been shown not to be unique to myofascial pain syndrome: in a controlled study, they were also present in some “normal” subjects. Myofascial pain most often involves the posterior neck, low back, shoulders, and chest. Chronic pain in the muscles of the posterior neck may involve referral of pain from a tender point in the erector neck muscle or upper trapezius to the head, leading to persistent headaches that may last for days. Tender points in the paraspinal muscles of the low back may refer pain to the buttock. Pain may be referred down the leg from a tender point in the gluteus medius and can mimic sciatica. A tender point in the infraspinatus muscle may produce local and referred pain over the lateral deltoid and down the outside of the arm into the hand. Injection of a local anesthetic such as 1% lidocaine into the tender point site often results in at least transient pain relief. Another useful technique is first to spray an agent such as ethyl chloride from the tender point toward the area of referred pain and then to stretch the muscle. This maneuver may need to be repeated several times. Massage and application of ultrasound to the affected area also may be beneficial. Patients should be instructed in methods to prevent muscle stresses related to work and recreation. Posture and resting positions are important in preventing muscle tension. The prognosis in most patients is good. In some patients, regionally localized myofascial pain syndrome may seem to evolve into more generalized fibromyalgia **(Chap. 366)**. Abnormal or nonrestorative sleep is a common accompaniment in these patients and may need to be specifically addressed.

NEOPLASIAS AND ARTHRITIS

Primary tumors and tumor-like disorders of synovium are uncommon but should be considered in the differential diagnosis of monarticular joint disease. In addition, metastases to bone and primary bone tumors adjacent to a joint may produce joint symptoms.

Pigmented villonodular synovitis (*PVNS*) is characterized by the slowly progressive, exuberant, benign proliferation of synovial tissue, usually involving a single joint. The most common age of onset is in the third decade, and women are affected slightly more often than men. The cause of this disorder is unknown.

The synovium has a brownish color and numerous large, fingerlike villi that fuse to form pedunculated nodules. There is marked hyperplasia of synovial cells in the stroma of the villi. Hemosiderin granules and lipids are found in the cytoplasm of macrophages and in the interstitial tissue. Multinucleated giant cells may be present. The proliferative synovium grows into the subsynovial tissue and invades adjacent cartilage and bone.

The clinical picture of PVNS is characterized by the insidious onset of persistent swelling and pain in affected joints, most commonly the knee. Other joints affected include the hips, ankles, calcaneocuboid joints, elbows, and small joints of the fingers or toes. The disease may also involve the common flexor sheath of the hands or fingers. Less often, tendon sheaths in the wrist, ankle, or foot may be involved. Symptoms of pain, a catching sensation, or stiffness may initially be mild and intermittent and may be present for years before the patient seeks medical attention. Radiographs

may show joint space narrowing, erosions, and subchondral cysts. The diagnosis of PVNS is strongly suggested by gradient echo MRI, which reveals a synovial mass lesion of low signal intensity typical of tissue containing hemosiderin **(Fig. 367-3)**. The joint fluid contains blood and is dark red or almost black in color. Lipid-containing macrophages may be present in the fluid. The joint fluid may be clear if hemorrhage has not occurred. Some patients have polyarticular involvement.

FIGURE 367-3

Pigmented villonodular synovitis. MRI gradient echo sagittal image showing a mass that abuts the neck of the talus with marked low signal typical of tissue containing hemosiderin. *(Courtesy of Donald Flemming, MD; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The treatment for PVNS is complete synovectomy. With incomplete synovectomy, the villonodular synovitis recurs, and the rate of tissue growth may be faster than it was originally. Irradiation of the involved joint has been successful in some patients.

Synovial chondromatosis is a disorder characterized by multiple focal metaplastic growths of normal-appearing cartilage in the synovium or tendon sheath. Segments of cartilage break loose and continue to grow as loose bodies. When calcification and ossification of loose bodies occur, the disorder is referred to as *synovial osteochondromatosis*. The disorder is usually monarticular and affects young to middle-aged individuals. The knee is most often involved, followed by hip, elbow, and shoulder. Symptoms are pain, swelling, and decreased motion of the joint. Radiographs may show several rounded calcifications within the joint cavity. Treatment is synovectomy; however, as in PVNS, the tumor may recur.

Synovial sarcoma is a malignant neoplasm often found near a large joint of both upper and lower extremities, being more common in the lower extremity. It seldom arises within the joint itself. Synovial sarcomas constitute 10% of soft tissue sarcomas. The tumor is believed to arise from primitive

mesenchymal tissue that differentiates into epithelial cells and/or spindle cells. Small foci of calcification may be present in the tumor mass. Synovial sarcoma occurs most often in young adults and is more common in men. The tumor presents as a slowly growing deep-seated mass near a joint, without much pain. The area of the knee is the most common site, followed by the foot, ankle, elbow, and shoulder. Other primary sites include the buttocks, abdominal wall, retroperitoneum, and mediastinum. The tumor spreads along tissue planes. The most common site of visceral metastasis is the lung. The diagnosis is made by biopsy. Treatment consists of wide resection of the tumor, including adjacent muscle and regional lymph nodes, followed by chemotherapy and radiation therapy. Amputation of the involved distal extremity may be required. Chemotherapy may be beneficial in some patients with metastatic disease. Isolated sites of pulmonary metastasis can be surgically removed. The 5-year survival rate with treatment is variable and depends on the staging of the tumor, ranging from ~25% to ≥60%. Synovial sarcomas tend to recur locally and metastasize to regional lymph nodes, lungs, and skeleton.

In addition to the rare direct metastases of solid cell tumors to the highly vascular synovium, neoplasia arising from nonarticular organ sites can affect joints in other ways. Acute leukemias in children can mimic juvenile inflammatory arthritis with severe joint pain and fever. In adults, chronic and acute myeloid leukemia can infiltrate the synovium in rare instances. The rarely occurring hairy cell leukemia has a peculiar tendency to cause episodic inflammatory oligoarthritis and tenosynovitis; these episodes are dramatic and mimic acute gout attacks. They respond to potent anti-inflammatory therapy with glucocorticoids; with remission of the leukemia, they may abate. Carcinomas can be associated with several paraneoplastic articular syndromes, including HOA (discussed above). Acute palmar fasciitis with polyarthritis is a well-described but rare condition associated with certain cancers, mainly adenocarcinomas. Clinically, this syndrome is fairly abrupt in onset, with pain in the metacarpophalangeal and proximal interphalangeal joints of the hands and rapidly evolving contractures of the fingers due to thickening of the palmar (flexor) tendons. A similar syndrome can be seen in diabetics. Paraneoplastic arthritis has been described and may occur in several patterns: asymmetric disease predominantly affecting the lower extremity joints and symmetric polyarthritis with hand joint involvement. Tumors are often found after the onset of the arthritis, and many patients have a preceding period of malaise or weight loss. The onset is often acute, and patients tend to be older men. These features should raise the specter of an underlying malignancy (or a viral infection such as hepatitis C) as the cause of the arthritis. In one series, the symptoms resolved with successful therapy for the malignancy and did not recur with relapse of the malignancy. Dermatomyositis has a well-described association with neoplasms and may include joint pain and arthritis. Malignancy-associated arthritis may be responsive to NSAIDs and to treatment of the primary neoplasm.

ACKNOWLEDGMENT

This chapter represents a revised version of the chapter authored by Dr. Bruce C. Gilliland that appeared in previous editions of Harrison's. Dr. Gilliland passed away on February 17, 2007. He had been a contributor to Harrison's Principles of Internal Medicine since the 11th edition.

FURTHER READING

Aguilar C et al: Bone and joint disease in sickle cell disease. Hematol Oncol Clin North Am 19:929, 2005. [PubMed: 16214653]
Botek G et al: Charcot neuroarthropathy: An often overlooked complication of diabetes. Cleve Clin J Med 77:593, 2010. [PubMed: 20810870]
Guggenbuhl P et al: Miscellaneous non-inflammatory musculoskeletal conditions. Haemochromatosis: The bone and the joint. Best Pract Res Clin Rheumatol 25:649, 2011. [PubMed: 22142745]
Kedar E, Gardner GC: Lipid-associated rheumatologic syndromes. Rheum Dis Clin North Am 39:481, 2013. [PubMed: 23597975]

Killinger Z et al: Arthropathy in acromegaly. Rheum Dis Clin North Am 36:713, 2010. [PubMed: 21092848]
Ottaviani S et al: Pigmented villonodular synovitis: A retrospective single-center study of 122 cases and review of the literature. Semin Arthritis Rheum 40:539, 2011. [PubMed: 20884045]
Pineda C, Martínez-Lavín M: Hypertrophic osteoarthropathy: What a rheumatologist should know about this uncommon condition. Rheum Dis Clin North Am 39:383, 2013. [PubMed: 23597970]
Vanderhave KL et al: Musculoskeletal care of the hemophiliac patient. J Am Acad Orthop Surg 20:553, 2012. [PubMed: 22941798]

Chapter 368: Periarticular Disorders of the Extremities

Carol A. Langford

FIGURE 368-1

INTRODUCTION

Periarticular disorders are common musculoskeletal abnormalities that can affect people throughout a wide range of ages. This chapter discusses some of the more common periarticular disorders.

BURSITIS

Bursitis is inflammation of a bursa, which is a thin-walled sac lined with synovial tissue. The function of the bursa is to facilitate movement of tendons and muscles over bony prominences. Excessive frictional forces from overuse, trauma, systemic disease (e.g., rheumatoid arthritis, gout), or infection may cause bursitis. *Subacromial bursitis* (subdeltoid bursitis) is the most common form of bursitis. The subacromial bursa, which is contiguous with the subdeltoid bursa, is located between the undersurface of the acromion and the humeral head and is covered by the deltoid muscle. Bursitis often accompanies rotator cuff tendinitis. Another frequently encountered form is *trochanteric bursitis*, which involves the bursa around the insertion of the gluteus medius onto the greater trochanter of the femur. Patients experience pain over the lateral aspect of the hip and upper thigh and have tenderness over the posterior aspect of the greater trochanter. External rotation and resisted abduction of the hip elicit pain as will direct pressure applied to the bursa. *Olecranon bursitis* occurs over the posterior elbow, and when the area is acutely inflamed, infection or gout should be excluded by aspirating the bursa and performing a Gram stain and culture on the fluid as well as examining the fluid for urate crystals. *Achilles bursitis* involves the bursa located above the insertion of the tendon to the calcaneus and results from overuse and wearing tight shoes. *Retrocalcaneal bursitis* involves the bursa that is located between the calcaneus and posterior surface of the Achilles tendon. The pain is experienced at the back of the heel, and swelling appears on the medial and/or lateral side of the tendon. It occurs in association with spondyloarthritides, rheumatoid arthritis, gout, or trauma. *Ischial bursitis* affects the bursa separating the gluteus medius from the ischial tuberosity and develops from prolonged sitting and pivoting on hard surfaces. *Iliopsoas bursitis* affects the bursa that lies between the iliopsoas muscle and hip joint and is lateral to the femoral vessels. Pain is experienced over this area and is made worse by hip extension and flexion. *Anserine bursitis* is an inflammation of the sartorius bursa located over the medial side of the tibia just below the knee and under the conjoint tendon and is manifested by pain on climbing stairs. Tenderness is present over the insertion of the conjoint tendon of the sartorius, gracilis, and semitendinosus.

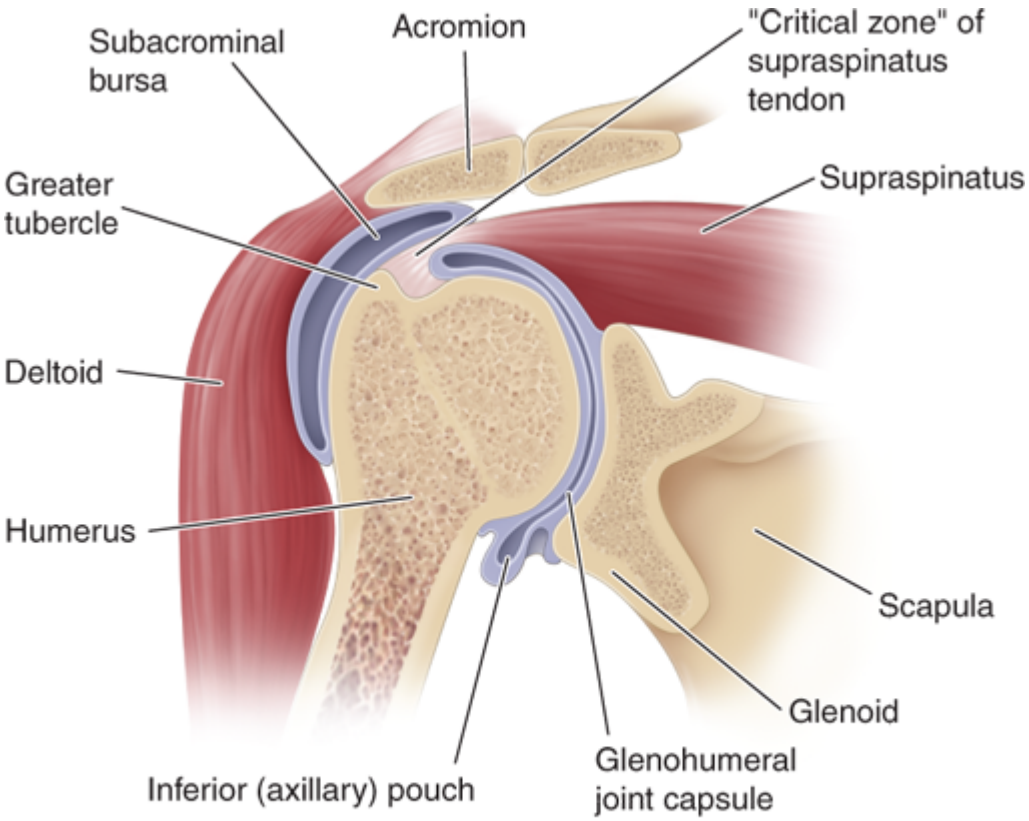
Prepatellar bursitis occurs in the bursa situated between the patella and overlying skin and is caused by kneeling on hard surfaces. Gout or infection may also occur at this site. Bursitis is typically diagnosed by history and physical examination, but visualization by ultrasound may play a useful role in selected instances for diagnosis and directed guidance of glucocorticoid injection. Treatment of bursitis consists of prevention of any aggravating situation, rest of the involved part, administration of a nonsteroidal anti-inflammatory drug (NSAID) where appropriate for an individual patient, or local glucocorticoid injection.

ROTATOR CUFF TENDINITIS AND IMPINGEMENT SYNDROME

Tendinitis of the rotator cuff is the major cause of a painful shoulder and is currently thought to be caused by inflammation of the tendon(s). The rotator cuff consists of the tendons of the supraspinatus, infraspinatus, subscapularis, and teres minor muscles, and inserts on the humeral tuberosities. Of the tendons forming the rotator cuff, the supraspinatus tendon is the most often affected, probably because of its repeated impingement (*impingement syndrome*) between the humeral head and the undersurface of the anterior third of the acromion and coracoacromial ligament above as well as the reduction in its blood supply that occurs with abduction of the arm **(Fig. 368-1)**. The tendon of the infraspinatus and that of the long head of the biceps are less commonly involved. Subacromial bursitis also accompanies this syndrome. Symptoms can appear without a triggering cause or after injury or overuse, especially with activities involving elevation of the arm with some degree of forward flexion. Impingement syndrome occurs in persons participating in baseball, tennis, swimming, or occupations that require repeated elevation of the arm. Those aged >40 years are particularly susceptible. Patients complain of a dull aching in the shoulder, which may interfere with sleep. Severe pain is experienced when the arm is actively abducted into an overhead position. The arc between 60° and 120° is especially painful. Tenderness is present over the lateral aspect of the humeral head just below the acromion. NSAIDs, local glucocorticoid injection, and physical therapy may relieve symptoms. Surgical decompression of the subacromial space may be necessary in patients refractory to conservative treatment.

FIGURE 368-1

Coronal section of the shoulder illustrating the relationships of the glenohumeral joint, the joint capsule, the subacromial bursa, and the rotator cuff (supraspinatus tendon). *(From F Kozin, in Arthritis and Allied Conditions, 13th ed, WJ Koopman [ed]. Baltimore, Williams & Wilkins, 1997, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Patients may tear the supraspinatus tendon acutely by falling on an outstretched arm or lifting a heavy object. Symptoms are pain along with weakness of abduction and external rotation of the shoulder. Atrophy of the supraspinatus muscles develops. The diagnosis is established by ultrasound, magnetic resonance imaging (MRI), or arthrogram. Surgical repair may be necessary in patients who fail to respond to conservative measures. In patients with moderate-to-severe tears and functional loss, surgery is indicated.

CALCIFIC TENDINITIS

This condition is characterized by deposition of calcium salts, primarily hydroxyapatite, within a tendon. The exact mechanism of calcification is not known but may be initiated by ischemia or degeneration of the tendon. The supraspinatus tendon is most often affected because it is frequently impinged on and has a reduced blood supply when the arm is abducted. The condition usually develops after age 40. Calcification within the tendon may evoke acute inflammation, producing sudden and severe pain in the shoulder. However, it may be asymptomatic or not related to the patient's symptoms. Diagnosis of calcific tendonitis can be made by ultrasound or radiograph. Most cases are self-limited and respond to conservative therapy with physical therapy and/or NSAIDs. A subset of patients is refractory and requires ultrasound-guided percutaneous needle aspiration and lavage or surgery.

BICIPITAL TENDINITIS AND RUPTURE

Bicipital tendinitis, or tenosynovitis, is produced by friction on the tendon of the long head of the biceps as it passes through the bicipital groove. When the inflammation is acute, patients experience anterior shoulder pain that radiates down the biceps into the forearm. Abduction and external rotation of the arm are painful and limited. The bicipital groove is very tender to palpation. Pain may be elicited along the course of the tendon by resisting supination of the forearm with the elbow at 90° (Yergason's supination sign). Acute rupture of the tendon may occur with vigorous exercise of the arm and is often painful. In a young patient, it should be repaired surgically. Rupture of the tendon in an older person may be associated with little or no pain and is recognized by the presence of persistent swelling of the biceps produced by the retraction of the long head of the biceps. Surgery is usually not necessary in this setting.

DE QUERVAIN'S TENOSYNOVITIS

In this condition, inflammation involves the abductor pollicis longus and the extensor pollicis brevis as these tendons pass through a fibrous sheath at the radial styloid process. The usual cause is repetitive twisting of the wrist. It may occur in pregnancy, and it also occurs in mothers who hold their babies with the thumb outstretched. Patients experience pain on grasping with their thumb, such as with pinching. Swelling and tenderness are often present over the radial styloid process. The Finkelstein sign is positive, which is elicited by having the patient place the thumb in the palm and close the fingers over it. The wrist is then ulnarly deviated, resulting in pain over the involved tendon sheath in the area of the radial styloid. Treatment consists initially of splinting the wrist and an NSAID. When severe or refractory to conservative treatment, glucocorticoid injections can be very effective.

PATELLAR TENDINITIS

Tendinitis involves the patellar tendon at its attachment to the lower pole of the patella. Patients may experience pain when jumping during sports, going up stairs, or doing deep knee squats. Tenderness is noted on examination over the lower pole of the patella. Treatment consists of rest, icing, and NSAIDs, followed by strengthening and increasing flexibility.

DRUG-INDUCED TENDINOPATHIES

With the broadening range of available pharmacologic agents, the potential for drug-induced tendinopathies has become increasingly recognized. The drug classes most associated with tendinopathies include quinolones, glucocorticoids, aromatase inhibitors, and statins. Although any tendon can be affected, the tendons of the lower extremities are most often impacted, particularly the Achilles tendon. The pathophysiological mechanisms responsible for drug-induced tendinopathies remain unknown. Presenting features include pain and potentially swelling over the tendon, although some patients may first present with tendon rupture. Ultrasound and MRI can provide information on tendon structure and integrity in support of the diagnosis. When suspected, the potential agent should be withdrawn and not reintroduced where possible in the overall medical management of the patient. Tendon ruptures may require surgery.

ILIOTIBIAL BAND SYNDROME

The iliotibial band is a thick connective tissue that runs from the ilium to the fibula. Patients with iliotibial band syndrome most commonly present with aching or burning pain at the site where the band courses over the lateral femoral condyle of the knee; pain may also radiate up the thigh, toward the hip. Predisposing factors for iliotibial band syndrome include a varus alignment of the knee, excessive running distance, poorly fitted shoes, or continuous running on uneven terrain. Treatment consists of rest, NSAIDs, physical therapy, and addressing risk factors such as shoes and running surface. Glucocorticoid injection into the area of tenderness can provide relief, but running must be avoided for at least 2 weeks after the injection. Surgical release of the iliotibial band has been helpful in rare patients for whom conservative treatment has failed.

ADHESIVE CAPSULITIS

Often referred to as “frozen shoulder,” adhesive capsulitis is characterized by pain and restricted movement of the shoulder, usually in the absence of intrinsic shoulder disease. Adhesive capsulitis may follow bursitis or tendinitis of the shoulder or be associated with systemic disorders such as chronic pulmonary disease, myocardial infarction, and diabetes mellitus. Prolonged immobility of the arm contributes to the development of adhesive capsulitis. Pathologically, the capsule of the shoulder is thickened, and a mild chronic inflammatory infiltrate and fibrosis may be present.

Adhesive capsulitis occurs more commonly in women aged >50 years. Pain and stiffness usually develop gradually but progress rapidly in some patients. Night pain is often present in the affected shoulder, and pain may interfere with sleep. The shoulder is tender to palpation, and both active and passive movements are restricted. Radiographs of the shoulder show osteopenia. The diagnosis is typically made by physical examination but can be confirmed if necessary by arthrography, in that only a limited amount of contrast material, usually <15 mL, can be injected under pressure into the shoulder joint.

In most patients, the condition improves spontaneously 1–3 years after onset. While pain usually improves, many patients are left with some limitation of shoulder motion. Early mobilization of the arm following an injury to the shoulder may prevent the development of this disease. Physical therapy provides the foundation of treatment for adhesive capsulitis. Local injections of glucocorticoids and NSAIDs may also provide relief of symptoms. Slow but forceful injection of contrast material into the joint may lyse adhesions and stretch the capsule, resulting in improvement of shoulder motion. Manipulation under anesthesia may be helpful in some patients.

LATERAL EPICONDYLITIS

Lateral epicondylitis, also known as tennis elbow, is a painful condition involving the soft tissue over the lateral aspect of the elbow. The pain originates at or near the site of attachment of the common extensors to the lateral epicondyle and may radiate into the forearm and dorsum of the wrist. The pain usually appears after work or recreational activities involving repeated motions of wrist extension and supination against

resistance. Most patients with this disorder injure themselves in activities other than tennis, such as pulling weeds, carrying suitcases or briefcases, or using a screwdriver. The injury in tennis usually occurs when hitting a backhand with the elbow flexed. Shaking hands and opening doors can reproduce the pain. Striking the lateral elbow against a solid object may also induce pain.

The treatment is usually rest along with administration of an NSAID. Ultrasound, icing, and friction massage may also help relieve pain. When pain is severe, the elbow is placed in a sling or splinted at 90° of flexion. When the pain is acute and well localized, injection of a glucocorticoid using a small-gauge needle may be effective. Following injection, the patient should be advised to rest the arm for at least 1 month and avoid activities that would aggravate the elbow. Once symptoms have subsided, the patient should begin rehabilitation to strengthen and increase flexibility of the extensor muscles before resuming physical activity involving the arm. A forearm band placed 2.5–5.0 cm (1–2 in.) below the elbow may help to reduce tension on the extensor muscles at their attachment to the lateral epicondyle. The patient should be advised to restrict activities requiring forcible extension and supination of the wrist. Improvement may take several months. The patient may continue to experience mild pain but, with care, can usually avoid the return of debilitating pain. Occasionally, surgical release of the extensor aponeurosis may be necessary.

MEDIAL EPICONDYLITIS

Medial epicondylitis is an overuse syndrome resulting in pain over the medial side of the elbow with radiation into the forearm. The cause of this syndrome is considered to be repetitive resisted motions of wrist flexion and pronation, which lead to microtears and granulation tissue at the origin of the pronator teres and forearm flexors, particularly the flexor carpi radialis. This overuse syndrome is usually seen in patients aged >35 years and is much less common than lateral epicondylitis. It occurs most often in work-related repetitive activities and also occurs with recreational activities such as swinging a golf club or throwing a baseball. On physical examination, there is tenderness just distal to the medial epicondyle over the origin of the forearm flexors. Pain can be reproduced by resisting wrist flexion and pronation with the elbow extended. Radiographs are usually normal. The differential diagnosis of patients with medial elbow symptoms includes tears of the pronator teres, acute medial collateral ligament tear, and medial collateral ligament instability. Ulnar neuritis has been found in 25–50% of patients with medial epicondylitis and is associated with tenderness over the ulnar nerve at the elbow as well as hypesthesia and paresthesia on the ulnar side of the hand.

The initial treatment of medial epicondylitis is conservative, involving rest, NSAIDs, friction massage, ultrasound, and icing. Some patients may require splinting. Injections of glucocorticoids at the painful site may also be effective. Patients should be instructed to rest for at least 1 month. Also, patients should start physical therapy once the pain has subsided. In patients with chronic debilitating medial epicondylitis that remains unresponsive after at least a year of treatment, surgical release of the flexor muscle at its origin may be necessary and is often successful.

PLANTAR FASCIITIS

Plantar fasciitis is a common cause of foot pain in adults, with the peak incidence occurring in people between the ages of 40 and 60 years. The pain originates at or near the site of the plantar fascia attachment to the medial tuberosity of the calcaneus. Several factors that increase the risk of developing plantar fasciitis include obesity, pes planus (flat foot or absence of the foot arch when standing), pes cavus (high-arched foot), limited dorsiflexion of the ankle, prolonged standing, walking on hard surfaces, and faulty shoes. In runners, excessive running and a change to a harder running surface may precipitate plantar fasciitis.

The diagnosis of plantar fasciitis can usually be made on the basis of history and physical examination alone. Patients experience severe pain with the first steps on arising in the morning or following inactivity during the day. The pain usually lessens with weight-bearing activity during the day, only to worsen with continued activity. Pain is made worse on walking barefoot or up stairs. On examination, maximal tenderness is elicited on palpation over the inferior heel corresponding to the site of attachment of the plantar fascia.

Imaging studies may be indicated when the diagnosis is not clear. Plain radiographs may show heel spurs, which are of little diagnostic significance. Ultrasonography in plantar fasciitis can demonstrate thickening of the fascia and diffuse hypoechogenicity, indicating edema at the attachment of the plantar fascia to the calcaneus. MRI is a sensitive method for detecting plantar fasciitis, but it is usually not required for establishing the diagnosis.

Resolution of symptoms occurs within 12 months in >80% of patients with plantar fasciitis. Initial treatment consists of ice, heat, massage, stretching, and eliminating activities that can exacerbate plantar fasciitis. Orthotics provide medial arch support and can be effective. Some patients may benefit from foot strapping or taping or by wearing a night splint designed to keep the ankle in a neutral position. A short course of NSAIDs can be given to patients when the benefits outweigh the risks. Local glucocorticoid injections have also been shown to be efficacious but may carry an increased risk for plantar fascia rupture. Plantar fasciotomy is reserved for those patients who have failed to improve after at least 6–12 months of conservative treatment.

FURTHER READING

Buchbinder R: Plantar fasciitis. N Engl J Med 350:2159, 2004.

Greis AC et al: Evaluation and nonsurgical management of rotator cuff calcific tendinopathy. Orthop Clin North Am 46:293, 2015. [PubMed: 25771323]

Harrison AK, Flatow EL: Subacromial impingement syndrome. J Am Acad Orthop Surg 19:701, 2011. [PubMed: 22052646]

Kirchgesner T et al: Drug-induced tendinopathy: From physiology to clinical applications. Joint Bone Spine 81:485, 2014. [PubMed: 24962977]

Neviaser AS, Neviaser RJ: Adhesive capsulitis of the shoulder. J Am Acad Orthop Surg 19:536, 2011. [PubMed: 21885699]

Chapter 369: Approach to the Patient with Endocrine Disorders

J. Larry Jameson

INTRODUCTION

The management of endocrine disorders requires a broad understanding of intermediary metabolism, reproductive physiology, bone metabolism, and growth. Accordingly, the practice of endocrinology is intimately linked to a conceptual framework for understanding hormone secretion, hormone action, and principles of feedback control (Chap. 370). The endocrine system is evaluated primarily by measuring hormone concentrations, arming the clinician with valuable diagnostic information. Most disorders of the endocrine system are amenable to effective treatment once the correct diagnosis is established. Endocrine deficiency disorders are treated with physiologic hormone replacement; hormone excess conditions, which usually are caused by benign glandular adenomas, are managed by removing tumors surgically or reducing hormone levels medically.

SCOPE OF ENDOCRINOLOGY

The specialty of endocrinology encompasses the study of glands and the hormones they produce. The term *endocrine* was coined by Starling to contrast the actions of hormones secreted internally (*endocrine*) with those secreted externally (*exocrine*) or into a lumen, such as the gastrointestinal tract. The term *hormone*, derived from a Greek phrase meaning “to set in motion,” aptly describes the dynamic actions of hormones as they elicit cellular responses and regulate physiologic processes through feedback mechanisms.

Unlike many other specialties in medicine, it is not possible to define endocrinology strictly along anatomic lines. The classic endocrine glands—pituitary, thyroid, parathyroid, pancreatic islets, adrenals, and gonads—communicate broadly with other organs through the nervous system, hormones, cytokines, and growth factors. In addition to its traditional synaptic functions, the brain produces a vast array of peptide hormones, and this has led to the discipline of neuroendocrinology. Through the production of hypothalamic releasing factors, the central nervous system (CNS) exerts a major regulatory influence over pituitary hormone secretion (Chap. 371). The peripheral nervous system stimulates the adrenal medulla. The immune and endocrine systems are also intimately intertwined. The adrenal hormone cortisol is a powerful immunosuppressant. Cytokines and interleukins (ILs) have profound effects on the functions of the pituitary, adrenal, thyroid, and gonads. Common endocrine diseases such as autoimmune thyroid disease and type 1

diabetes mellitus are caused by dysregulation of immune surveillance and tolerance. Less common diseases such as polyglandular failure, Addison's disease, and lymphocytic hypophysitis also have an immunologic basis.

The interdigitation of endocrinology with physiologic processes in other specialties sometimes blurs the role of hormones. For example, hormones play an important role in maintenance of blood pressure, intravascular volume, and peripheral resistance in the cardiovascular system. Vasoactive substances such as catecholamines, angiotensin II, endothelin, and nitric oxide are involved in dynamic changes of vascular tone in addition to their multiple roles in other tissues. The heart is the principal source of atrial natriuretic peptide, which acts in classic endocrine fashion to induce natriuresis at a distant target organ (the kidney). Erythropoietin, a traditional circulating hormone, is made in the kidney and stimulates erythropoiesis in bone marrow **(Chap. 59)**. The kidney is also integrally involved in the renin-angiotensin axis **(Chap. 379)** and is a primary target of several hormones, including parathyroid hormone (PTH), mineralocorticoids, and vasopressin. The gastrointestinal tract produces a vast array of peptide hormones, such as cholecystokinin, ghrelin, gastrin, secretin, and vasoactive intestinal peptide, among many others. Carcinoid and islet tumors can secrete excessive amounts of these hormones, leading to specific clinical syndromes **(Chap. 80)**. Many of these gastrointestinal hormones are also produced in the CNS, where their functions are poorly understood. Adipose tissue produces leptin, which acts centrally to control appetite, along with adiponectin, resistin, and other hormones that regulate metabolism. As hormones such as inhibin, ghrelin, and leptin are discovered, they become integrated into the science and practice of medicine on the basis of their functional roles rather than their tissues of origin.

Characterization of hormone receptors frequently reveals unexpected relationships to factors in nonendocrine disciplines. The growth hormone (GH) and leptin receptors, for example, are members of the cytokine receptor family. The G protein–coupled receptors (GPCRs), which mediate the actions of many peptide hormones, are used in numerous physiologic processes, including vision, smell, and neurotransmission.

PATHOLOGIC MECHANISMS OF ENDOCRINE DISEASE

Endocrine diseases can be divided into three major types of conditions: (1) hormone excess, (2) hormone deficiency, and (3) hormone resistance **(Table 369-1)**.

TABLE 369-1

Causes of Endocrine Dysfunction

TYPE OF ENDOCRINE DISORDER	EXAMPLES
Hyperfunction	
Neoplastic	
Benign	Pituitary adenomas, hyperparathyroidism, autonomous thyroid or adrenal nodules, pheochromocytoma
Malignant	Adrenal cancer, medullary thyroid cancer, carcinoid
Ectopic	Ectopic ACTH, SIADH secretion
Multiple endocrine neoplasia (MEN)	MEN1, MEN2
Autoimmune	Graves' disease
Iatrogenic	Cushing's syndrome, hypoglycemia
Infectious/inflammatory	Subacute thyroiditis
Activating receptor mutations	LH, TSH, Ca^{2+}, PTH receptors, $G_s\alpha$
Hypofunction	
Autoimmune	Hashimoto's thyroiditis, type 1 diabetes mellitus, Addison's disease, polyglandular failure
Iatrogenic	Radiation-induced hypopituitarism, hypothyroidism, surgical
Infectious/inflammatory	Adrenal insufficiency, hypothalamic sarcoidosis
Hormone mutations	GH, LHβ, FSHβ, vasopressin
Enzyme defects	21-Hydroxylase deficiency
Developmental defects	Kallmann's syndrome, Turner's syndrome, transcription factors
Nutritional/vitamin deficiency	Vitamin D deficiency, iodine deficiency
Hemorrhage/infarction	Sheehan's syndrome, adrenal insufficiency
Hormone Resistance	
Receptor mutations	
Membrane	GH, vasopressin, LH, FSH, ACTH, GnRH, GHRH, PTH, leptin, Ca^{2+}
Nuclear	AR, TR, VDR, ER, GR, PPARγ
Signaling pathway mutations	Albright's hereditary osteodystrophy
Postreceptor	Type 2 diabetes mellitus, leptin resistance

Abbreviations: ACTH, adrenocorticotropic hormone; AR, androgen receptor; ER, estrogen receptor; FSH, follicle-stimulating hormone; GHRH, growth hormone–releasing hormone; GnRH, gonadotropin-releasing hormone; GR, glucocorticoid receptor; LH, luteinizing hormone; PPAR, peroxisome proliferator activated receptor; PTH, parathyroid hormone; SIADH, syndrome of inappropriate antidiuretic hormone; TR, thyroid hormone receptor; TSH, thyroid-stimulating hormone; VDR, vitamin D receptor.

CAUSES OF HORMONE EXCESS

Syndromes of hormone excess can be caused by neoplastic growth of endocrine cells, autoimmune disorders, and excess hormone administration. Benign endocrine tumors, including parathyroid, pituitary, and adrenal adenomas, often retain the capacity to produce hormones, reflecting the fact that these tumors are relatively well differentiated. Many endocrine tumors exhibit subtle defects in their "set points" for feedback regulation. For example, in Cushing's disease, impaired feedback inhibition of adrenocorticotropic hormone (ACTH) secretion is associated with autonomous function. However, the tumor cells are not completely resistant to feedback, as evidenced by ACTH suppression by higher doses of dexamethasone (e.g., high-dose dexamethasone test) **(Chap. 379)**. Similar set point defects are also typical of parathyroid adenomas and autonomously functioning thyroid nodules.

The molecular basis of some endocrine tumors, such as the multiple endocrine neoplasia (MEN) syndromes (MEN1, 2A, 2B), has provided important insights into tumorigenesis **(Chap. 381)**. MEN1 is characterized primarily by the triad of parathyroid, pancreatic islet, and pituitary tumors. MEN2 predisposes to medullary thyroid carcinoma, pheochromocytoma, and hyperparathyroidism. The *MEN1* gene, located on chromosome 11q13, encodes a putative tumor-suppressor gene, menin. Analogous to the paradigm first described for retinoblastoma, the affected individual inherits a mutant copy of the *MEN1* gene, and tumorigenesis ensues after a somatic "second hit" leads to loss of function of the normal *MEN1* gene (through deletion or point mutations).

In contrast to inactivation of a tumor-suppressor gene, as occurs in MEN1 and most other inherited cancer syndromes, MEN2 is caused by activating mutations in a single allele. In this case, activating mutations of the *RET* protooncogene, which encodes a receptor tyrosine kinase, leads to thyroid C cell hyperplasia in childhood before the development of medullary thyroid carcinoma. Elucidation of this pathogenic mechanism has allowed early genetic screening for *RET* mutations in individuals at risk for MEN2, permitting identification of those who may benefit from prophylactic thyroidectomy and biochemical screening for pheochromocytoma and hyperparathyroidism.

Mutations that activate hormone receptor signaling have been identified in several GPCRs. For example, activating mutations of the luteinizing hormone (LH) receptor cause a dominantly transmitted form of male-limited precocious puberty, reflecting premature stimulation of testosterone synthesis in Leydig cells **(Chap. 384)**. Activating mutations in these GPCRs are located predominantly in the transmembrane domains and induce receptor coupling to $G_s\alpha$ even in the absence of hormone. Consequently, adenylate cyclase is

activated, and cyclic adenosine monophosphate (AMP) levels increase in a manner that mimics hormone action. A similar phenomenon results from activating mutations in $G_s\alpha$. When these mutations occur early in development, they cause McCune-Albright syndrome. When they occur only in somatotropes, the activating $G_s\alpha$ mutations cause GH-secreting tumors and acromegaly **(Chap. 373)**.

In autoimmune Graves' disease, antibody interactions with the thyroid-stimulating hormone (TSH) receptor mimic TSH action, leading to hormone overproduction **(Chap. 375)**. Analogous to the effects of activating mutations of the TSH receptor, these stimulating autoantibodies induce conformational changes that release the receptor from a constrained state, thereby triggering receptor coupling to G proteins.

CAUSES OF HORMONE DEFICIENCY

Most examples of hormone deficiency states can be attributed to glandular destruction caused by autoimmunity, surgery, infection, inflammation, infarction, hemorrhage, or tumor infiltration (Table 369-1). Autoimmune damage to the thyroid gland (Hashimoto's thyroiditis) and pancreatic islet β cells (type 1 diabetes mellitus) is a prevalent cause of endocrine disease. Mutations in a number of hormones, hormone receptors, transcription factors, enzymes, and channels can also lead to hormone deficiencies.

HORMONE RESISTANCE

Most severe hormone resistance syndromes are due to inherited defects in membrane receptors, nuclear receptors, or the pathways that transduce receptor signals. These disorders are characterized by defective hormone action despite the presence of increased hormone levels. In complete androgen resistance, for example, mutations in the androgen receptor result in a female phenotypic appearance in genetic (XY) males, even though LH and testosterone levels are increased **(Chap. 381)**. In addition to these relatively rare genetic disorders, more common acquired forms of functional hormone resistance include insulin resistance in type 2 diabetes mellitus, leptin resistance in obesity, and GH resistance in catabolic states. The pathogenesis of functional resistance involves receptor downregulation and postreceptor desensitization of signaling pathways; functional forms of resistance are generally reversible.

CLINICAL EVALUATION OF ENDOCRINE DISORDERS

Because most glands are relatively inaccessible, the physical examination usually focuses on the manifestations of hormone excess or deficiency as well as direct examination of palpable glands, such as the thyroid and gonads. For these reasons, it is important to evaluate patients in the context of their presenting symptoms, review of systems, family and social history, and exposure to medications that may affect the endocrine system. Astute clinical skills are required to detect subtle symptoms and signs suggestive of underlying endocrine disease. For example, a patient with Cushing's syndrome may manifest specific findings, such as central fat redistribution, skin striae, and proximal muscle weakness, in addition to features seen commonly in the general population, such as obesity, plethora, hypertension, and glucose intolerance. Similarly, the insidious onset of hypothyroidism—with mental slowing, fatigue, dry skin, and other features—

can be difficult to distinguish from similar, nonspecific findings in the general population. Clinical judgment that is based on knowledge of disease prevalence and pathophysiology is required to decide when to embark on more extensive evaluation of these disorders. Laboratory testing plays an essential role in endocrinology by allowing quantitative assessment of hormone levels and dynamics. Radiologic imaging tests such as computed tomography (CT) scan, magnetic resonance imaging (MRI), thyroid scan, and ultrasound are also used for the diagnosis of endocrine disorders. However, these tests generally are employed only after a hormonal abnormality has been established by biochemical testing.

HORMONE MEASUREMENTS AND ENDOCRINE TESTING

Immunoassays are the most important diagnostic tool in endocrinology, as they allow sensitive, specific, and quantitative determination of steady-state and dynamic changes in hormone concentrations. Immunoassays use antibodies to detect specific hormones. For many peptide hormones, these measurements are now configured to use two different antibodies to increase binding affinity and specificity. There are many variations of these assays; a common format involves using one antibody to capture the antigen (hormone) onto an immobilized surface and a second antibody, coupled to a chemiluminescent (immunochemiluminescent assay [ICMA]) or radioactive (immunoradiometric assay [IRMA]) signal, to detect the antigen. These assays are sensitive enough to detect plasma hormone concentrations in the picomolar to nanomolar range, and they can readily distinguish structurally related proteins, such as PTH from PTH-related peptide (PTHrP). A variety of other techniques are used to measure specific hormones, including mass spectroscopy, various forms of chromatography, and enzymatic methods; bioassays are now used rarely.

Most hormone measurements are based on plasma or serum samples. However, urinary hormone determinations remain useful for the evaluation of some conditions. Urinary collections over 24 h provide an integrated assessment of the production of a hormone or metabolite, many of which vary during the day. It is important to ensure complete collections of 24-h urine samples; simultaneous measurement of creatinine provides an internal control for the adequacy of collection and can be used to normalize some hormone measurements. A 24-h urine-free cortisol measurement largely reflects the amount of unbound cortisol, thus providing a reasonable index of biologically available hormone. Other commonly used urine determinations include 17-hydroxycorticosteroids, 17-ketosteroids, vanillylmandelic acid, metanephrine, catecholamines, 5-hydroxyindoleacetic acid, and calcium.

The value of quantitative hormone measurements lies in their correct interpretation in a clinical context. The normal range for most hormones is relatively broad, often varying by a factor of two- to tenfold. The normal ranges for many hormones are sex- and age-specific. Thus, using the correct normative database is an essential part of interpreting hormone tests. The pulsatile nature of hormones and factors that can affect their secretion, such as sleep, meals, and medications, must also be considered. Cortisol values increase fivefold between midnight and dawn; reproductive hormone levels vary dramatically during the female menstrual cycle.

For many endocrine systems, much information can be gained from basal hormone testing, particularly when different components of an endocrine axis are assessed simultaneously. For example, low testosterone and elevated LH levels suggest a primary gonadal problem, whereas a hypothalamic-pituitary disorder is likely if both LH and testosterone are low. Because TSH is a sensitive indicator of thyroid function, it is generally recommended as a first-line test for thyroid disorders. An elevated TSH level is almost always the result of primary hypothyroidism, whereas a low TSH is most often caused by thyrotoxicosis. These predictions can be confirmed by determining the free thyroxine level. In the less common circumstance when free thyroxine and TSH are both low, it is important to consider secondary hypopituitarism caused by hypothalamic-pituitary disease. Elevated calcium and PTH levels suggest hyperparathyroidism, whereas PTH is suppressed in hypercalcemia caused by malignancy or granulomatous diseases. A suppressed ACTH in the setting of hypercortisolemia, or increased urine-free cortisol, is seen with hyperfunctioning adrenal adenomas.

It is not uncommon, however, for baseline hormone levels associated with pathologic endocrine conditions to overlap with the normal range. In this circumstance, dynamic testing is useful to separate the two groups further. There are a multitude of dynamic endocrine tests, but all are based on principles of feedback regulation, and most responses can be rationalized based on principles that govern the regulation of endocrine axes. *Suppression tests* are used in the setting of suspected endocrine hyperfunction. An example is the dexamethasone suppression test used to evaluate Cushing's syndrome **(Chaps. 373 and 379)**. *Stimulation tests* generally are used to assess endocrine hypofunction. The ACTH stimulation test, for example, is used to assess the adrenal gland response in patients with suspected adrenal insufficiency. Other stimulation tests use hypothalamic-releasing factors such as corticotropin-releasing hormone (CRH) and growth hormone–releasing hormone (GHRH) to evaluate pituitary hormone reserve **(Chap. 373)**. Insulin-induced hypoglycemia evokes pituitary ACTH and GH responses. Stimulation tests based on reduction or inhibition of endogenous hormones are now used infrequently. Examples include metyrapone inhibition of cortisol synthesis and clomiphene inhibition of estrogen feedback.

SCREENING AND ASSESSMENT OF COMMON ENDOCRINE DISORDERS

Many endocrine disorders are prevalent in the adult population **(Table 369-2)** and can be diagnosed and managed by general internists, family practitioners, or other primary health care providers. The high prevalence and clinical impact of certain endocrine diseases justifies vigilance for features of these disorders during routine physical examinations; laboratory screening is indicated in selected high-risk populations.

TABLE 369-2

Examples of Prevalent Endocrine and Metabolic Disorders in the Adult

DISORDER	APPROX. PREVALENCE IN ADULTS[a]	SCREENING/TESTING RECOMMENDATIONS[b]	CHAPTER(S)
Obesity	36% BMI ≥30 70% BMI ≥25	Calculate BMI Measure waist circumference Exclude secondary causes Consider comorbid complications	**395**
Type 2 diabetes mellitus	>8%	Beginning at age 45, screen every 3 years, or earlier in high-risk groups: FPG >126 mg/dL Random plasma glucose >200 mg/dL An elevated HbA_{1C} Consider comorbid complications	**396**
Hyperlipidemia	20–25%	Cholesterol screening at least every 5 years; more often in high-risk groups Lipoprotein analysis (LDL, HDL) for increased cholesterol, CAD, diabetes Consider secondary causes	**400**
Metabolic syndrome	35%	Measure waist circumference, FPG, BP, lipids	**401**
Hypothyroidism	5–10%, women 0.5–2%, men	TSH; confirm with free T_4 Screen women after age 35 and every 5 years thereafter	**377**
Graves' disease	1–3%, women 0.1%, men	TSH, free T_4	**376**
Thyroid nodules and neoplasia	2–5% palpable >25% by ultrasound	Physical examination of thyroid Fine-needle aspiration biopsy	**378**

DISORDER	APPROX. PREVALENCE IN ADULTS[a]	SCREENING/TESTING RECOMMENDATIONS[b]	CHAPTER(S)
Osteoporosis	5–10%, women 2–5%, men	Bone mineral density measurements in women >65 years or in postmenopausal women or men at risk Exclude secondary causes	**404**
Hyperparathyroidism	0.1–0.5%, women > men	Serum calcium PTH, if calcium is elevated Assess comorbid conditions	**403**
Infertility	10%, couples	Investigate both members of couple Semen analysis in male Assess ovulatory cycles in female Specific tests as indicated	**384, 385**
Polycystic ovarian syndrome	5–10%, women	Free testosterone, DHEAS Consider comorbid conditions	**385**
Hirsutism	5–10%	Free testosterone, DHEAS Exclude secondary causes Additional tests as indicated	**387**
Menopause	Median age, 51	FSH	**388**
Hyperprolactinemia	15% in women with amenorrhea or galactorrhea	PRL level MRI, if not medication-related	**373**
Erectile dysfunction	10–25%	Careful history, PRL, testosterone Consider secondary causes (e.g., diabetes)	**390**
Hypogonadism, male	1–2%	Testosterone, LH	**384**
Gynecomastia	15%	Often, no tests are indicated Consider Klinefelter's syndrome Consider medications, hypogonadism, liver disease	**384**

DISORDER	APPROX. PREVALENCE IN ADULTS[a]	SCREENING/TESTING RECOMMENDATIONS[b]	CHAPTER(S)
Klinefelter's syndrome	0.2%, men	Karyotype Testosterone	**383**
Vitamin D deficiency	10%	Measure serum 25-OH vitamin D Consider secondary causes	**402**
Turner's syndrome	0.03%, women	Karyotype Consider comorbid conditions	**383**

[a]The prevalence of most disorders varies among ethnic groups and with aging. Data based primarily on U.S. population. [b]See individual chapters for additional information on evaluation and treatment. Early testing is indicated in patients with signs and symptoms of disease and in those at increased risk.

Abbreviations: BMI, body mass index; BP, blood pressure; CAD, coronary artery disease; DHEAS, dehydroepiandrosterone; FPG, fasting plasma glucose; FSH, follicle-stimulating hormone; HbA_{1C}, hemoglobin A_{1C}; HDL, high-density lipoprotein; LDL, low-density lipoprotein; LH, luteinizing hormone; MRI, magnetic resonance imaging; PRL, prolactin; PTH, parathyroid hormone; TSH, thyroid-stimulating hormone.

FURTHER READING

Golden SH et al: Prevalence and incidence of endocrine and metabolic disorders in the United States: A comprehensive review. J Clin Endocrinol Metab 94:1853, 2009. [PubMed: 19494161]

Golden SH et al: Health disparities in endocrine disorders: Biological, clinical, and nonclinical factors—an Endocrine Society Scientific Statement. J Clin Endocrinol Metab 97:E1579, 2012. [PubMed: 22730516]

Jameson JL, DeGroot LJ (eds): *Endocrinology: Adult and Pediatric*, 7th ed. Philadelphia, Elsevier, 2016.

Loriaux DL: *A Biographical History of Endocrinology*. Hoboken, Wiley Blackwell, 2016.

Melmed S, Polonsky KS, Larsen PR, Kronenberg HM (eds): *Williams Textbook of Endocrinology*, 13th ed. Philadelphia, Elsevier, 2016.

Chapter 370: Mechanisms of Hormone Action

J. Larry Jameson

INTRODUCTION

Hormones function as a communication system within the body. The endocrine system, composed of various glands and the hormones they produce, interacts with essentially all other physiologic systems to regulate growth, metabolism, homeostasis, and reproduction. Because hormones circulate and act via receptors in target tissues, they serve to integrate physiologic responses to external or internal cues. For example, the light-dark cycle, sensed through the visual system, modulates hypothalamic corticotopin-releasing hormone (CRH), which increases pituitary adrenocorticotropin hormone (ACTH) production, leading to increased adrenal cortisol production before the time of waking in the morning. Increased cortisol, in turn, circulates throughout the body, acting via the nuclear glucocorticoid receptor, to activate numerous genetic programs that influence metabolism, the cardiovascular system, behavior, and the immune system. This chapter provides an overview of the different types of hormones and how they function at the cellular level to control myriad physiologic processes.

CLASSES OF HORMONES

Hormones can be divided into five major types: (1) *amino acid derivatives* such as dopamine, catecholamine, and thyroid hormone; (2) *small neuropeptides* such as gonadotropin-releasing hormone (GnRH), thyrotropin-releasing hormone (TRH), somatostatin, and vasopressin; (3) *large proteins* such as insulin, luteinizing hormone (LH), and parathyroid hormone (PTH); (4) *steroid hormones* such as cortisol and estrogen that are synthesized from cholesterol-based precursors; and (5) *vitamin derivatives* such as retinoids (vitamin A) and vitamin D. A variety of *peptide growth factors*, most of which act locally, share actions with hormones. As a rule, amino acid derivatives and peptide hormones interact with cell-surface membrane receptors. Steroids, thyroid hormones, vitamin D, and retinoids are lipid-soluble and interact with intracellular nuclear receptors, although many also interact with membrane receptors or intracellular signaling proteins as well.

HORMONE AND RECEPTOR FAMILIES

Hormones and receptors can be grouped into families, reflecting structural similarities and evolutionary origins (Table 370-1). The evolution of these families generates diverse but highly selective pathways of

hormone action. Recognition of these relationships has proven useful for extrapolating information gleaned from one hormone or receptor to other family members.

TABLE 370-1

Examples of Membrane Receptor Families and Signaling Pathways

RECEPTORS	EFFECTORS	SIGNALING PATHWAYS
G Protein–Coupled Seven-Transmembrane Receptor (GPCR)		
β-Adrenergic, LH, FSH, TSH Glucagon, PTH, PTHrP, ACTH, MSH, GHRH, CRH α-Adrenergic, somatostatin TRH, GnRH	$G_s\alpha$, adenylate cyclase Ca^{2+} channels $G_i\alpha$ G_q, G_{11}	Stimulation of cyclic AMP production, protein kinase A Calmodulin, Ca^{2+}-dependent kinases Inhibition of cyclic AMP production Activation of K^+, Ca^{2+} channels Phospholipase C, diacyl-glycerol, IP_3, protein kinase C, voltage-dependent Ca^{2+} channels
Receptor Tyrosine Kinase		
Insulin, IGF-I EGF, NGF	Tyrosine kinases, IRS Tyrosine kinases, ras	MAP kinases, PI 3-kinase; AKT Raf, MAP kinases, RSK
Cytokine Receptor–Linked Kinase		
GH, PRL	JAK, tyrosine kinases	STAT, MAP kinase, PI 3-kinase, IRS-1
Serine Kinase		
Activin, TGF-β, MIS	Serine kinase	Smads

Abbreviations: IP_3, inositol triphosphate; IRS, insulin receptor substrates; MAP, mitogen-activated protein; MSH, melanocyte-stimulating hormone; NGF, nerve growth factor; PI, phosphatidylinositol; RSK, ribosomal S6 kinase; TGF-β, transforming growth factor β. For all other abbreviations, see text. Note that most receptors interact with multiple effectors and activate networks of signaling pathways.

The glycoprotein hormone family, consisting of thyroid-stimulating hormone (TSH), follicle-stimulating hormone (FSH), LH, and human chorionic gonadotropin (hCG), illustrates many features of evolutionarily related hormones. The glycoprotein hormones are heterodimers that share the α subunit in common; the β subunits are distinct and confer specific biologic actions. The overall three-dimensional architecture of the β subunits is similar, reflecting the locations of conserved disulfide bonds that restrain protein conformation. The cloning of the β-subunit genes from multiple species suggests that this family arose from a common ancestral gene, probably by gene duplication and subsequent divergence to evolve new biologic functions.

As hormone families enlarge and diverge, their receptors must co-evolve to derive new biologic functions. Related G protein–coupled receptors (GPCRs), for example, have evolved for each of the glycoprotein hormones. These receptors are also structurally similar, and each is coupled predominantly to the $G_s\alpha$ signaling pathway. However, there is minimal overlap of hormone binding. For example, TSH binds with high specificity to the TSH receptor but interacts minimally with the LH or FSH receptors. Nonetheless, there can be subtle physiologic consequences of hormone cross-reactivity with other receptors. Very high levels of hCG during pregnancy stimulate the TSH receptor and increase thyroid hormone levels, resulting via feedback inhibition in a compensatory decrease in TSH.

Insulin and insulin-like growth factor I (IGF-I) and IGF-II have structural similarities that are most apparent when precursor forms of the proteins are compared. In contrast to the high degree of specificity seen with the glycoprotein hormones, there is moderate cross-talk among the members of the insulin/IGF family. High concentrations of an IGF-II precursor produced by certain tumors (e.g., sarcomas) can cause hypoglycemia, partly because of binding to insulin and IGF-I receptors **(Chap. 403)**. High concentrations of insulin also bind to the IGF-I receptor, perhaps accounting for some of the clinical manifestations seen in conditions with chronic hyperinsulinemia.

Another important example of receptor cross-talk is seen with PTH and parathyroid hormone–related peptide (PTHrP) **(Chap. 403)**. PTH is produced by the parathyroid glands, whereas PTHrP is expressed at high levels during development and by a variety of tumors **(Chap. 89)**. These hormones have amino acid sequence similarity, particularly in their amino-terminal regions. Both hormones bind to the PTH1R receptor that is expressed in bone and kidney. Hypercalcemia and hypophosphatemia therefore may result from excessive production of either hormone, making it difficult to distinguish hyperparathyroidism from hypercalcemia of malignancy solely on the basis of serum chemistries. However, sensitive and specific assays for PTH and PTHrP now allow these disorders to be distinguished more readily.

Based on their specificities for DNA-binding sites, the nuclear receptor family can be subdivided into type 1 receptors (glucocorticoid receptor, mineralocorticoid receptor, androgen receptor, estrogen receptor, progesterone receptor) that bind steroids and type 2 receptors (thyroid hormone receptor, vitamin D receptor, retinoic acid receptor, peroxisome proliferator activated receptor) that bind thyroid hormone, vitamin D, retinoic acid, or lipid derivatives, respectively. Certain functional domains in nuclear receptors, such as the zinc finger DNA-binding domains, are highly conserved. However, selective amino acid differences within this domain confer DNA sequence specificity. The hormone-binding domains are more

variable, providing great diversity in the array of small molecules that bind to different nuclear receptors. With few exceptions, hormone binding is highly specific for a single type of nuclear receptor. One exception involves the glucocorticoid and mineralocorticoid receptors. Because the mineralocorticoid receptor also binds glucocorticoids with high affinity, an enzyme (11β-hydroxysteroid dehydrogenase) in renal tubular cells inactivates glucocorticoids, allowing selective responses to mineralocorticoids such as aldosterone. However, when very high glucocorticoid concentrations occur, as in Cushing's syndrome, the glucocorticoid degradation pathway becomes saturated, allowing excessive cortisol levels to bind mineralocorticoid receptors leading to sodium retention and potassium wasting. This phenomenon is particularly pronounced in ectopic adrenocorticotropic hormone (ACTH) syndromes **(Chap. 379)**. Another example of relaxed nuclear receptor specificity involves the estrogen receptor, which can bind an array of compounds, some of which have little apparent structural similarity to the high-affinity ligand estradiol. This feature of the estrogen receptor makes it susceptible to activation by "environmental estrogens" such as resveratrol, octylphenol, and many other aromatic hydrocarbons. However, this lack of specificity provides an opportunity to synthesize a remarkable series of clinically useful antagonists (e.g., tamoxifen) and selective estrogen response modulators (SERMs) such as raloxifene. These compounds generate distinct conformations that alter receptor interactions with components of the transcription machinery (see below), thereby conferring their unique actions.

HORMONE SYNTHESIS AND PROCESSING

The synthesis of peptide hormones and their receptors occurs through a classic pathway of gene expression: transcription → mRNA → protein → posttranslational protein processing → intracellular sorting, followed by membrane integration or secretion.

Many hormones are embedded within larger precursor polypeptides that are proteolytically processed to yield the biologically active hormone. Examples include proopiomelanocortin (POMC) → ACTH; proglucagon → glucagon; proinsulin → insulin; and pro-PTH → PTH, among others. In many cases, such as POMC and proglucagon, these precursors generate multiple biologically active peptides. It is provocative that hormone precursors are typically inactive, presumably adding an additional level of regulatory control. Prohormone conversion occurs not only for peptide hormones but also for certain steroids (testosterone → dihydrotestosterone) and thyroid hormone ($T_4 \rightarrow T_3$).

Peptide precursor processing is intimately linked to intracellular sorting pathways that transport proteins to appropriate vesicles and enzymes, resulting in specific cleavage steps, followed by protein folding and translocation to secretory vesicles. Hormones destined for secretion are translocated across the endoplasmic reticulum under the guidance of an amino-terminal signal sequence that subsequently is cleaved. Cell-surface receptors are inserted into the membrane via short segments of hydrophobic amino acids that remain embedded within the lipid bilayer. During translocation through the Golgi and endoplasmic reticulum, hormones and receptors are subject to a variety of posttranslational modifications, such as glycosylation and phosphorylation, which can alter protein conformation, modify circulating half-life, and alter biologic activity.

Synthesis of most steroid hormones is based on modifications of the precursor, cholesterol. Multiple regulated enzymatic steps are required for the synthesis of testosterone **(Chap. 384)**, estradiol **(Chap. 385)**, cortisol **(Chap. 379)**, and vitamin D **(Chap. 402)**. This large number of synthetic steps predisposes to multiple genetic and acquired disorders of steroidogenesis.

Endocrine genes contain regulatory DNA elements similar to those found in many other genes, but their exquisite control by hormones reflects the presence of specific hormone response elements. For example, the TSH genes are repressed directly by thyroid hormones acting through the thyroid hormone receptor (TR), a member of the nuclear receptor family. Steroidogenic enzyme gene expression requires specific transcription factors, such as steroidogenic factor-1 (SF-1), acting in conjunction with signals transmitted by trophic hormones (e.g., ACTH or LH). Once activated, SF-1 functions as a master regulator, inducing a large array of genes required for steroidogenic and metabolic pathways required for steroid synthesis. For some hormones, substantial regulation occurs at the level of translational efficiency. Insulin biosynthesis, although it requires ongoing gene transcription, is regulated primarily at the translational and secretory levels in response to elevated levels of glucose or amino acids.

HORMONE SECRETION, TRANSPORT, AND DEGRADATION

The circulating level of a hormone is determined by its rate of secretion and its half-life. After protein processing, peptide hormones (e.g., GnRH, insulin, growth hormone [GH]) are stored in secretory granules. As these granules mature, they are poised beneath the plasma membrane for imminent release into the circulation. In most instances, the stimulus for hormone secretion is a releasing factor or neural signal that induces rapid changes in intracellular calcium concentrations, leading to secretory granule fusion with the plasma membrane and release of its contents into the extracellular environment and bloodstream. Steroid hormones, in contrast, diffuse into the circulation as they are synthesized. Thus, their secretory rates are closely aligned with rates of synthesis. For example, ACTH and LH induce steroidogenesis by stimulating the activity of the *st*eroidogenic *a*cute *r*egulatory (StAR) protein (transports cholesterol into the mitochondrion) along with other rate-limiting steps (e.g., cholesterol side-chain cleavage enzyme, CYP11A1) in the steroidogenic pathway.

Hormone transport and degradation dictate the rapidity with which a hormonal signal decays. Some hormone signals are evanescent (e.g., somatostatin), whereas others are longer-lived (e.g., TSH). Because somatostatin exerts effects in virtually every tissue, a short half-life allows its concentrations and actions to be controlled locally. Structural modifications that impair somatostatin degradation have been useful for generating long-acting therapeutic analogues such as octreotide **(Chap. 373)**. In contrast, the actions of TSH are highly specific for the thyroid gland. Its prolonged half-life accounts for relatively constant serum levels even though TSH is secreted in discrete pulses.

An understanding of circulating hormone half-life is important for achieving physiologic hormone replacement, as the frequency of dosing and the time required to reach steady state are intimately linked to rates of hormone decay. T_4, for example, has a circulating half-life of 7 days. Consequently, >1 month is

required to reach a new steady state, and single daily doses are sufficient to achieve constant hormone levels. T_3, in contrast, has a half-life of 1 day. Its administration is associated with more dynamic serum levels, and it must be administered two to three times per day. Similarly, synthetic glucocorticoids vary widely in their half-lives; those with longer half-lives (e.g., dexamethasone) are associated with greater suppression of the hypothalamic-pituitary-adrenal (HPA) axis. Most protein hormones (e.g., ACTH, GH, prolactin [PRL], PTH, LH) have relatively short half-lives (<20 min), leading to sharp peaks of secretion and decay. The only accurate way to profile the pulse frequency and amplitude of these hormones is to measure levels in frequently sampled blood (every 10 min or less) over long durations (8–24 h). Because this is not practical in a clinical setting, an alternative strategy is to pool three to four samples drawn at about 30-min intervals, or interpret the results in the context of a relatively wide normal range. Rapid hormone decay is useful in certain clinical settings. For example, the short half-life of PTH allows the use of intraoperative PTH determinations to confirm successful removal of a parathyroid adenoma. This is particularly valuable diagnostically when there is a possibility of multicentric disease or parathyroid hyperplasia, as occurs with multiple endocrine neoplasia (MEN) or renal insufficiency.

Many hormones circulate in association with serum-binding proteins. Examples include (1) T_4 and T_3 binding to thyroxine-binding globulin (TBG), albumin, and thyroxine-binding prealbumin (TBPA); (2) cortisol binding to cortisol-binding globulin (CBG); (3) androgen and estrogen binding to sex hormone–binding globulin (SHBG); (4) IGF-I and II binding to multiple IGF-binding proteins (IGFBPs); (5) GH interactions with GH-binding protein (GHBP), a circulating fragment of the GH receptor extracellular domain; and (6) activin binding to follistatin. These interactions provide a hormonal reservoir, prevent otherwise rapid degradation of unbound hormones, restrict hormone access to certain sites (e.g., IGFBPs), and modulate the unbound, or “free,” hormone concentrations. Although a variety of binding protein abnormalities have been identified, most have little clinical consequence aside from creating diagnostic problems. For example, TBG deficiency can reduce total thyroid hormone levels greatly but the free concentrations of T_4 and T_3 remain normal. Liver disease and certain medications can also influence binding protein levels (e.g., estrogen increases TBG) or cause displacement of hormones from binding proteins (e.g., salsalate displaces T_4 from TBG). In general, only unbound hormone is available to interact with receptors and thus elicit a biologic response. Short-term perturbations in binding proteins change the free hormone concentration, which in turn induces compensatory adaptations through feedback loops. SHBG changes in women are an exception to this self-correcting mechanism. When SHBG decreases because of insulin resistance or androgen excess, the unbound testosterone concentration is increased, potentially contributing to hirsutism in women with polycystic ovary syndrome (PCOS) **(Chap. 387)**. The increased unbound testosterone level does not result in an adequate compensatory feedback correction because estrogen, not testosterone, is the primary regulator of the reproductive axis.

An additional exception to the unbound hormone hypothesis involves megalin, a member of the low-density lipoprotein (LDL) receptor family that serves as an endocytotic receptor for thyroglobulin, carrier-bound vitamins A and D and SHBG-bound androgens and estrogens. After internalization, the carrier proteins are

degraded in lysosomes and release their bound ligands within the cells. Membrane transporters have also been identified for thyroid hormones.

Hormone degradation can be an important mechanism for regulating concentrations locally. As noted above, 11β-hydroxysteroid dehydrogenase inactivates glucocorticoids in renal tubular cells, preventing actions through the mineralocorticoid receptor. Thyroid hormone deiodinases convert T_4 to T_3 and can inactivate T_3. During development, degradation of retinoic acid by Cyp26b1 prevents primordial germ cells in the male from entering meiosis, as occurs in the female ovary.

HORMONE ACTION THROUGH RECEPTORS

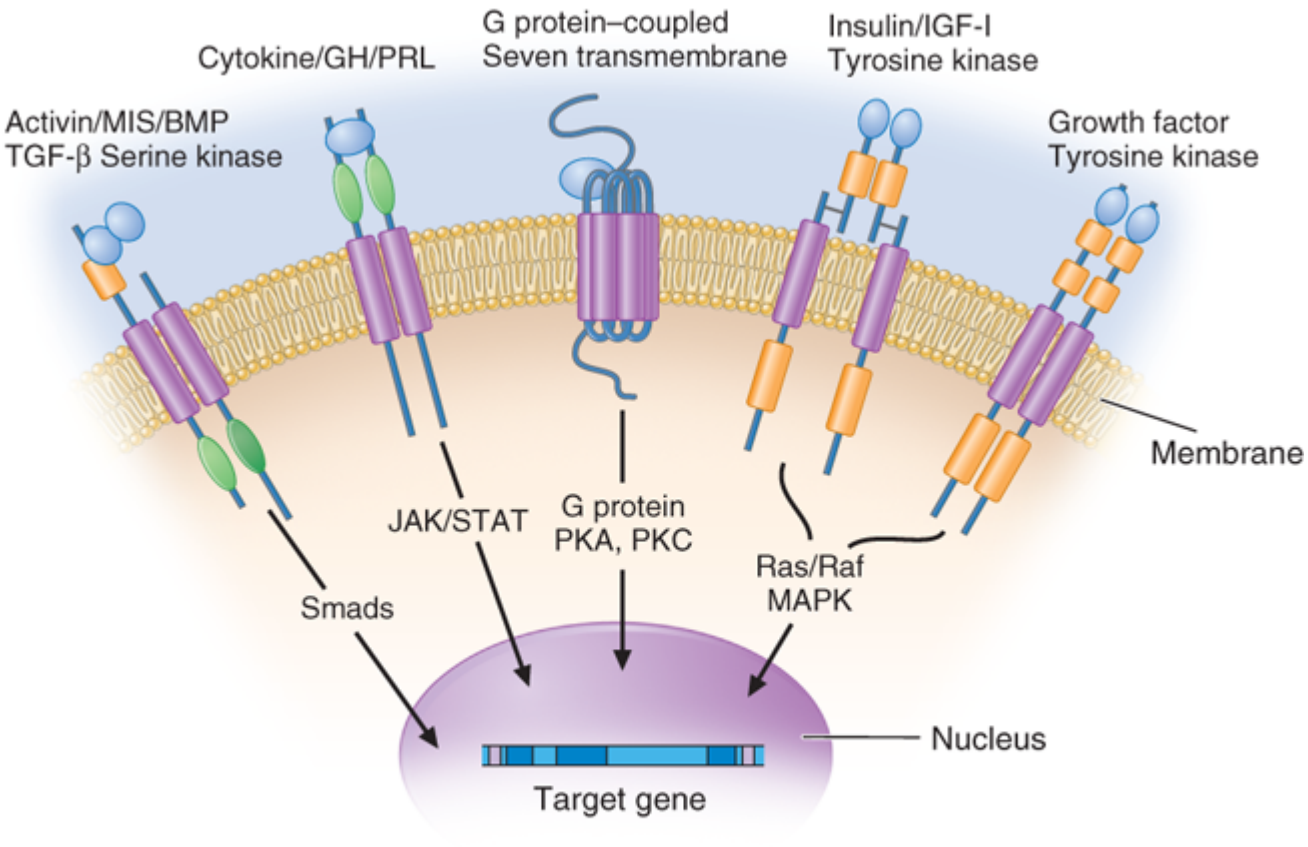
Receptors for hormones are divided into two major classes: membrane and nuclear. *Membrane receptors* primarily bind peptide hormones and catecholamines. *Nuclear receptors* bind small molecules that can diffuse across the cell membrane, such as steroids and vitamin D. Certain general principles apply to hormone-receptor interactions regardless of the class of receptor. Hormones bind to receptors with specificity and an affinity that generally coincides with the dynamic range of circulating hormone concentrations. Low concentrations of free hormone (usually 10^{-12} to 10^{-9} *M*) rapidly associate and dissociate from receptors in a bimolecular reaction such that the occupancy of the receptor at any given moment is a function of hormone concentration and the receptor's affinity for the hormone. Receptor numbers vary greatly in different target tissues, providing one of the major determinants of specific tissue responses to circulating hormones. For example, ACTH receptors are located almost exclusively in the adrenal cortex, and FSH receptors are found predominantly in the gonads. In contrast, insulin and TRs are widely distributed, reflecting the need for metabolic responses in all tissues.

MEMBRANE RECEPTORS

Membrane receptors for hormones can be divided into several major groups: (1) seven transmembrane GPCRs, (2) tyrosine kinase receptors, (3) cytokine receptors, and (4) serine kinase receptors **(Fig. 370-1)**. The *seven transmembrane GPCR family* binds a remarkable array of hormones, including large proteins (e.g., LH, PTH), small peptides (e.g., TRH, somatostatin), catecholamines (epinephrine, dopamine), and even minerals (e.g., calcium). The extracellular domains of GPCRs vary widely in size and are the major binding site for large hormones. The transmembrane-spanning regions are composed of hydrophobic α-helical domains that traverse the lipid bilayer. Like some channels, these domains are thought to circularize and form a hydrophobic pocket into which certain small ligands fit. Hormone binding induces conformational changes in these domains, transducing structural changes to the intracellular domain, which is a docking site for G proteins.

FIGURE 370-1

Membrane receptor signaling. MAPK, mitogen-activated protein kinase; PKA, C, protein kinase A, C; TGF, transforming growth factor. For other abbreviations, see text.



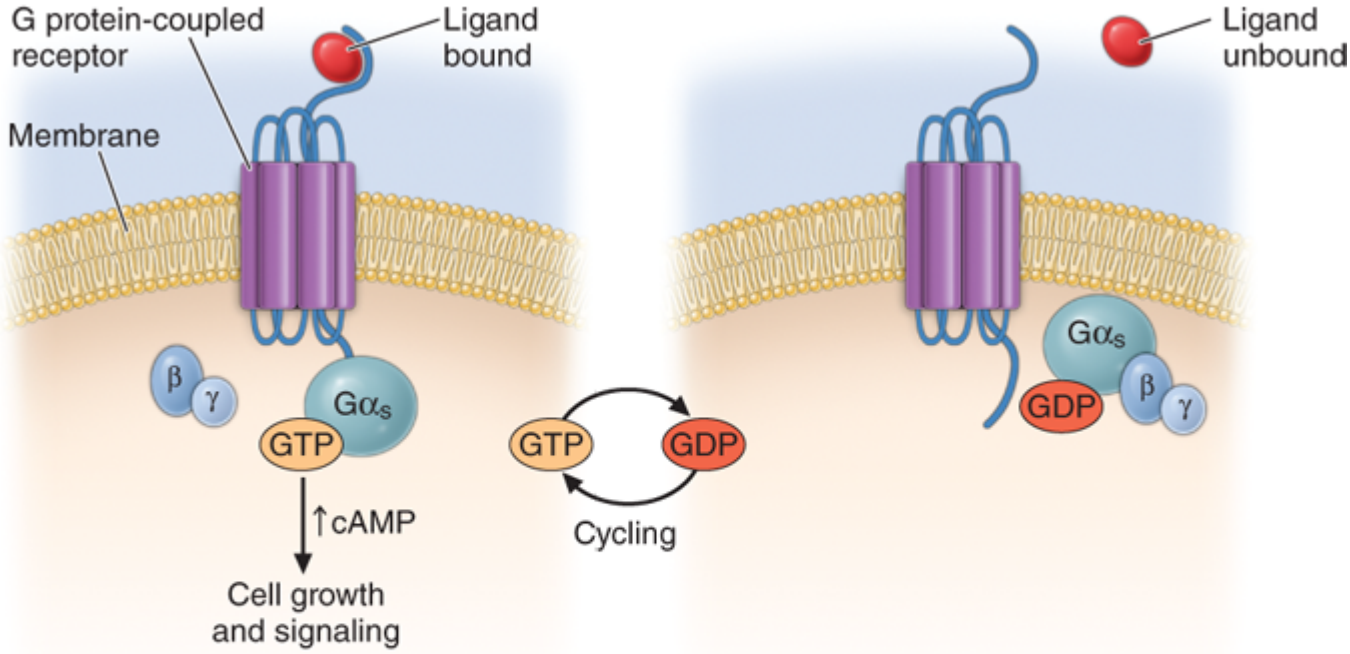
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The large family of *G proteins*, so named because they bind guanine nucleotides (guanosine triphosphate [GTP], guanosine diphosphate [GDP]), provides great diversity for coupling receptors to different signaling pathways. G proteins form a heterotrimeric complex that is composed of various α and βγ subunits **(Fig. 370-2)**. The α subunit contains the guanine nucleotide–binding site and an intrinsic GTPase that hydrolyzes GTP → GDP. The βγ subunits are tightly associated and modulate the activity of the α subunit as well as mediating their own effector signaling pathways. G protein activity is regulated by a cycle that involves GTP hydrolysis and dynamic interactions between the α and βγ subunits. Hormone binding to the receptor induces GDP dissociation, allowing Gα to bind GTP and dissociate from the βγ complex. Under these conditions, the Gα subunit is activated and mediates signal transduction through various enzymes, such as adenylate cyclase and phospholipase C. GTP hydrolysis to GDP allows reassociation with the βγ subunits and restores the inactive state. G proteins interact with other cellular proteins, including kinases, channels, G protein–coupled receptor kinases (GRKs), and arrestins, that mediate signaling as well as receptor desensitization and recycling.

FIGURE 370-2

G protein signaling. G protein-coupled receptors signal via the family of G proteins, so-named because they bind guanylyl nucleotides. In the example shown, a G-protein-coupled receptor (GPCR) bound to a ligand induces GDP dissociation, allowing Gsα to bind GTP and dissociate from the βγ complex. GTP-bound Gsα increases cAMP production by adenylyl cyclase and activates the protein kinase A pathway. Not shown are separate signaling pathways activated by the βγ complex. When GTP is converted to GDP by an intrinsic GTPase, the βγ subunits reassociate with GDP-bound Gsα and the complex returns to an inactive state. As

noted in the text, mutations in Gsα that eliminate GTPase activity result in constitutive activation of receptor signaling pathways because GTP-bound Gsα cannot be converted to its GDP-bound inactive state. cAMP, cyclic adenosine 5′-monophosphate; GDP, guanosine diphosphate; Gsα, G protein α; GTP, guanosine triphosphate.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A variety of endocrinopathies result from mutations in GPCR receptors that alter their interactions with G proteins (Table 370-2). Loss-of-function mutations are generally recessive and inactivate the relevant hormone signaling pathway. Because many of these receptors are important for development as well as signaling, patient presentations often resemble glandular failure syndromes (e.g., mutations in LH-R, FSH-R, TSH-R). Gain-of-function mutations involve a more complex mechanism. Selected mutations induce conformational changes in the GPCR that mimic the activated state normally induced by hormone binding. These mutations result in a constitutively active state in which G protein coupling stimulates cell signaling pathways, most commonly via cyclic adenosine 5′-monophosphate (cAMP) and protein kinase A. When mutations occur in the germline, the conditions are heritable and present in early life (e.g., LH-R, TSH-R). Somatic mutations can also occur and result in clonal expansion of hyperfunctioning cells.

TABLE 370-2

Genetic Causes of G protein Receptor Disorders

RECEPTOR	DISORDER	GENETICS
LH	Leydig cell hypoplasia (male) Primary amenorrhea, resistance to LH (female) Familial male precocious puberty (male) Leydig cell adenoma, precocious puberty (male)	AR, inactivating AR, inactivating AD, activating Sporadic, activating
FSH	Hypergonadotropic ovarian failure (female) Hypospermia (male) Ovarian hyperstimulation (female)	AR, inactivating AR, inactivating AD, activating
TSH	Congenital hypothyroidism, TSH resistance Nonautoimmune familial hyperthyroidism Hyperfunctioning thyroid adenoma	AR, AD, inactivating AD, activating Sporadic, activating
GnRH	Hypogonadotropic hypogonadism	AR, inactivating
Kisspeptin	Hypogonadotropic hypogonadism Precocious puberty	AR, inactivating AD, activating
TRH	Central hypothyroidism	AR, inactivating
GHRH	GH deficiency	AR, inactivating
PTH	Blomstrand chondrodysplasia Jansen metaphyseal chondrodysplasia	AR, inactivating AD, activating
Calcium sensing receptor	Familial hypocalciuric hypercalcemia Neonatal severe hyperparathyroidism Familial hypocalcemic hypercalciura	AD, inactivating AR, inactivating AD, activating
Arginine vasopressin receptor 2	Nephrogenic diabetes insipidus Nephrogenic SIADH	XL, inactivating XL, activating
ACTH	Familial ACTH resistance	AR, inactivating

RECEPTOR	DISORDER	GENETICS
Melanocortin 4	Severe obesity	Codominant, inactivating

Abbreviations: ACTH, adrenocorticotropin hormone; AD, autosomal dominant; AR, autosomal recessive; FSH, follicle-stimulating hormone; GH, growth hormone; GHRH, growth hormone–releasing hormone; GnRH, gonadotropin-releasing hormone; LH, luteinizing hormone; PTH, parathyroid hormone; SIADH, syndrome of inappropriate antidiuretic hormone secretion; TRH, thyrotropin-releasing hormone; TSH, thyroid-stimulating hormone; XL, X-linked.

Mutations in the TSH-R illustrate the range of possible clinical consequences of GPCR mutations. Recessive inactivating mutations in the TSH-R cause congenital hypothyroidism with thyroid gland hypoplasia and resistance to TSH. Clinically, the hormone profile resembles primary hypothyroidism with low T4 and high TSH. On the other hand, germline activating mutations cause congenital hyperthyroidism. The disorder is autosomal dominant because an activating mutation of one TSH-R allele is sufficient to induce cellular hyperfunction and disease. Because the TSH-R is activated in every cell of the thyroid, there is hyperplastic growth and hyperfunction that resembles the pathology seen in Graves' disease. This unusual disorder presents in infancy and must be distinguished from the more common clinical circumstance in which maternal antibodies in women with active or previously treated Graves' disease cross the placenta and stimulate the thyroid gland of the fetus. If an activating TSH-R mutation occurs later in life, in the somatic tissue, there is clonal expansion of the thyrocyte harboring the mutation, ultimately leading to an autonomous hyperfunctioning thyroid nodule. Of note, a similar condition can be caused by somatic mutations in Gsα. In this case, the Gsα GTPase is inactivated and GTP cannot be converted to GDP. Consequently, the Gsα signaling pathway in this particular cell is constitutively active, mimicking chronic TSH stimulation and again leading to clonal expansion and an autonomous hyperfunctioning thyroid nodule. About one-third of hyperfunctioning "hot" thyroid nodules harbor mutations in either the TSH-R or Gsα (TSH-R mutations are more common).

Gsα mutations in tissues other than the thyroid can also cause endocrine disease. For example, Gsα mutations in pituitary somatotropes mimic activation of the growth hormone–releasing hormone (GHRH) pathway and lead to GH-producing adenomas and acromegaly. Rarely, mutations in other components of the protein kinase A pathway in somatotropes can also cause GH-producing adenomas. Gsα mutations that occur early in development (typically mosaic) cause McCune-Albright syndrome **(Chap. 405)** and the clinical features are manifest because the activated G protein pathway mimics the actions of various hormones (PTH, melanocyte stimulating hormone [MSH], TSH, GHRH) in different tissues. Germline inactivating Gsα mutations cause a range of disorders that are transmitted and expressed in a complex manner because the locus is imprinted **(Chap. 403)**. These conditions include Albright's hereditary osteodystrophy (AHO), pseudopseudohypoparathyroidism (PPHP), and pseudohypoparathyroidism types 1b, 1c, and 2.

The *tyrosine kinase receptors* transduce signals for insulin and a variety of growth factors, such as IGF-I, epidermal growth factor (EGF), nerve growth factor, platelet-derived growth factor, and fibroblast growth factor. The cysteine-rich extracellular domains contain binding sites for the growth factors. After ligand binding, this class of receptors undergoes autophosphorylation, inducing interactions with intracellular adaptor proteins such as Shc and insulin receptor substrates (IRS). In the case of the insulin receptor, multiple kinases are activated, including the Raf-Ras-MAPK and the Akt/protein kinase B pathways. The tyrosine kinase receptors play a prominent role in cell growth and differentiation as well as in intermediary metabolism.

The GH and PRL receptors belong to the *cytokine receptor* family. Analogous to the tyrosine kinase receptors, ligand binding induces receptor interaction with intracellular kinases—the Janus kinases (JAKs), which phosphorylate members of the signal transduction and activators of transcription (STAT) family—as well as with other signaling pathways (Ras, PI3-K, MAPK). The activated STAT proteins translocate to the nucleus and stimulate expression of target genes.

The *serine kinase receptors* mediate the actions of activins, transforming growth factor β, müllerian-inhibiting substance (MIS, also known as anti-müllerian hormone, AMH), and bone morphogenic proteins (BMPs). This family of receptors (consisting of type I and II subunits) signals through proteins termed *smads* (fusion of terms for *Caenorhabditis elegans* sma + mammalian mad). Like the STAT proteins, the smads serve a dual role of transducing the receptor signal and acting as transcription factors. The pleomorphic actions of these growth factors dictate that they act primarily in a local (paracrine or autocrine) manner. Binding proteins such as follistatin (which binds activin and other members of this family) function to inactivate the growth factors and restrict their distribution.

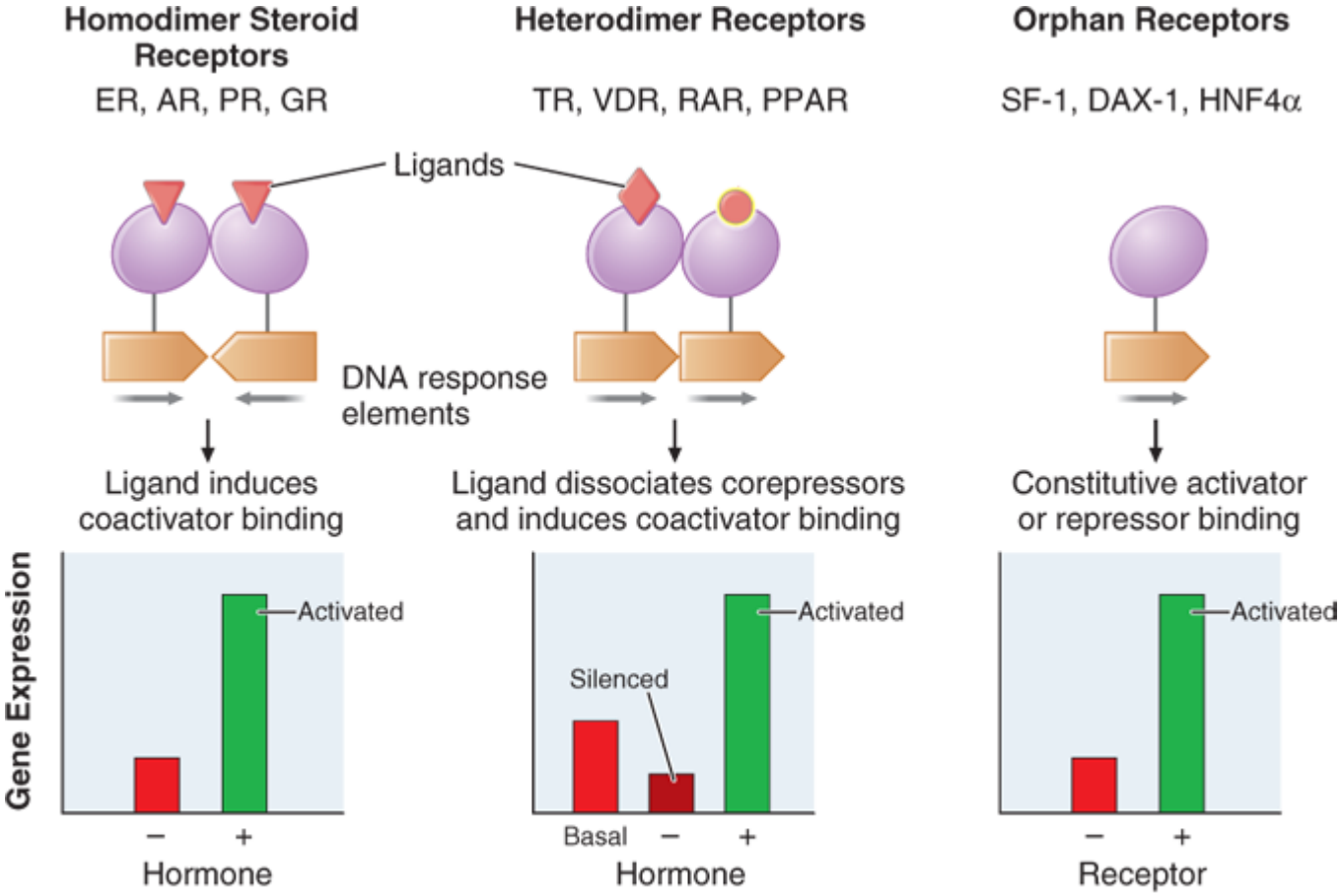
Disease-causing mutations also occur in each of these classes of receptors. For example, insulin receptor mutations cause an extreme form of insulin resistance. GH receptor mutations cause Laron-type dwarfism, characterized by low IGF-1 and high GH. AMH receptor mutations cause persistent Müllerian duct syndrome. These hormone resistance syndromes are autosomal recessive and relatively uncommon. Unlike the GPCRs, activating mutations are unusual, although they do occur for the RET tyrosine kinase receptor, which causes the autosomal dominant disorder multiple endocrine neoplasia type 2 (MEN-2) **(Chap. 381)**.

NUCLEAR RECEPTORS

The family of nuclear receptors has grown to nearly 100 members, many of which are still classified as orphan receptors because their ligands, if they exist, have not been identified **(Fig. 370-3)**. Otherwise, most nuclear receptors are classified on the basis of their ligands. Although all nuclear receptors ultimately act to increase or decrease gene transcription, some (e.g., glucocorticoid receptor) reside primarily in the cytoplasm, whereas others (e.g., TR) are located in the nucleus. After ligand binding, the cytoplasmically localized receptors translocate to the nucleus. There is growing evidence that certain nuclear receptors (e.g., glucocorticoid, estrogen) can also act at the membrane or in the cytoplasm to activate or repress signal transduction pathways, providing a mechanism for cross-talk between membrane and nuclear receptors.

FIGURE 370-3

Nuclear receptor signaling. AR, androgen receptor; DAX, *d*osage-sensitive sex-reversal, *a*drenal hypoplasia congenita, *X*-chromosome; ER, estrogen receptor; GR, glucocorticoid receptor; HNF4α, hepatic nuclear factor 4α; PPAR, peroxisome proliferator activated receptor; PR, progesterone receptor; RAR, retinoic acid receptor; SF-1, steroidogenic factor-1; TR, thyroid hormone receptor; VDR, vitamin D receptor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The structures of nuclear receptors have been studied extensively, including by x-ray crystallography. The DNA-binding domain, consisting of two zinc fingers, contacts-specific DNA recognition sequences in target genes. Most nuclear receptors bind to DNA as dimers. Consequently, each monomer recognizes an individual DNA motif, referred to as a "half-site." The steroid receptors, including the glucocorticoid, estrogen, progesterone, and androgen receptors, bind to DNA as homodimers. Consistent with this twofold symmetry, their DNA recognition half-sites are palindromic. The thyroid, retinoid, peroxisome proliferator activated, and vitamin D receptors bind to DNA preferentially as heterodimers in combination with retinoid X receptors (RXRs). Their DNA half-sites are typically arranged as direct repeats.

The carboxy-terminal hormone-binding domain mediates transcriptional control. For type II receptors such as TR and retinoic acid receptor (RAR), co-repressor proteins bind to the receptor in the absence of ligand and silence gene transcription. Hormone binding induces conformational changes, triggering the release of co-repressors and inducing the recruitment of coactivators that stimulate transcription. Thus, these receptors are capable of mediating dramatic changes in the level of gene activity. Disease states can be associated with defective regulation of these events. For example, in promyelocytic leukemia, fusion of RARα to other nuclear

proteins causes aberrant gene silencing that prevents normal cellular differentiation. Treatment with retinoic acid reverses this repression and allows cellular differentiation and apoptosis to occur. Most type 1 steroid receptors interact weakly with co-repressors, but ligand binding still induces interactions with an array of coactivators. X-ray crystallography shows that various SERMs induce distinct estrogen receptor conformations. The tissue-specific responses caused by these agents in breast, bone, and uterus appear to reflect distinct interactions with coactivators. The receptor-coactivator complex stimulates gene transcription by several pathways, including (1) recruitment of enzymes (histone acetyl transferases) that modify chromatin structure, (2) interactions with additional transcription factors on the target gene, and (3) direct interactions with components of the general transcription apparatus to enhance the rate of RNA polymerase II–mediated transcription. Studies of nuclear receptor-mediated transcription show that these are dynamic events that involve relatively rapid (e.g., 30–60 min) cycling of transcription complexes on any specific target gene.

Nuclear receptor mutations are an important cause of endocrine disease. Androgen receptor mutations cause androgen insensitivity syndrome (AIS) (Chap. 383). Because the androgen receptor is located on the X-chromosome, mutations are more commonly manifest than with other nuclear receptor disorders. Affected individuals with AIS are XY phenotypic females with retained testes and male-range testosterone levels. Tissue insensitivity to androgens varies based on the severity of the mutation. Müllerian structures are absent because Sertoli cells of the testis produce AMH during development. Female carriers of androgen receptor mutations are phenotypically normal. Recessive mutations of the estrogen, glucocorticoid, and vitamin D receptors are rare.

Thyroid hormone receptor β (TRβ) mutations have an unusual pathophysiology. They are autosomal dominant and function via a “dominant negative” mechanism to cause resistance to thyroid hormone (RTH) (Chap. 375). The mutations occur in selected regions of the TRβ hormone-binding domain and preserve the ability of the mutant receptor to heterodimerize with RXR and bind to DNA regulatory sites. The mutant receptors function as antagonists of receptors from the normal copy of the TRβ gene. Affected patients have high T_4 and T_3 and inappropriately elevated (unsuppressed) TSH, reflecting impaired feedback regulation of the hypothalamic-pituitary-thyroid axis. Organ systems are variably resistant to thyroid hormones based upon the relative expression of TRβ and TRα proteins. Mutations in the genes encoding TRα and PPARγ can also cause disease by functioning in an analogous dominant negative manner.

FUNCTIONS OF HORMONES

The functions of individual hormones are described in detail in subsequent chapters. Nevertheless, it is useful to illustrate how most biologic responses require integration of several different hormone pathways. The physiologic functions of hormones can be divided into three general areas: (1) growth and differentiation, (2) maintenance of homeostasis, and (3) reproduction.

GROWTH

Multiple hormones and nutritional factors mediate the complex phenomenon of growth **(Chap. 371)**. Short stature may be caused by GH deficiency, hypothyroidism, Cushing's syndrome, precocious puberty, malnutrition, chronic illness, or genetic abnormalities that affect the epiphyseal growth plates (e.g., *FGFR3* and *SHOX* mutations). Many factors (GH, IGF-I, thyroid hormones) stimulate growth, whereas others (sex steroids) lead to epiphyseal closure. Understanding these hormonal interactions is important in the diagnosis and management of growth disorders. For example, delaying exposure to high levels of sex steroids may enhance the efficacy of GH treatment.

MAINTENANCE OF HOMEOSTASIS

Although virtually all hormones affect homeostasis, the most important among them are the following:

1. Thyroid hormone—controls about 25% of basal metabolism in most tissues.
2. Cortisol—exerts a permissive action for many hormones in addition to its own direct effects.
3. PTH—regulates calcium and phosphorus levels.
4. Vasopressin—regulates serum osmolality by controlling renal free-water clearance.
5. Mineralocorticoids—control vascular volume and serum electrolyte (Na^+, K^+) concentrations.
6. Insulin—maintains euglycemia in the fed and fasted states.

The defense against hypoglycemia is an impressive example of integrated hormone action **(Chap. 399)**. In response to the fasting state and falling blood glucose, insulin secretion is suppressed, resulting in decreased glucose uptake and enhanced glycogenolysis, lipolysis, proteolysis, and gluconeogenesis to mobilize fuel sources. If hypoglycemia develops (usually from insulin administration or sulfonylureas), an orchestrated counterregulatory response occurs—glucagon and epinephrine rapidly stimulate glycogenolysis and gluconeogenesis, whereas GH and cortisol act over several hours to raise glucose levels and antagonize insulin action.

Although free-water clearance is controlled primarily by vasopressin, cortisol and thyroid hormone are also important for facilitating renal tubular responses to vasopressin **(Chap. 374)**. PTH and vitamin D function in an interdependent manner to control calcium metabolism **(Chap. 402)**. PTH stimulates renal synthesis of 1,25-dihydroxyvitamin D, which increases calcium absorption in the gastrointestinal tract and enhances PTH action in bone. Increased calcium, along with vitamin D, feeds back to suppress PTH, thus maintaining calcium balance.

Depending on the severity of a specific stress and whether it is acute or chronic, multiple endocrine and cytokine pathways are activated to mount an appropriate physiologic response. In severe acute stress such as trauma or shock, the sympathetic nervous system is activated and catecholamines are released, leading to increased cardiac output and a primed musculoskeletal system. Catecholamines also increase mean blood pressure and stimulate glucose production. Multiple stress-induced pathways converge on the

hypothalamus, stimulating several hormones, including vasopressin and corticotropin-releasing hormone (CRH). These hormones, in addition to cytokines (tumor necrosis factor α, interleukin [IL] 2, IL-6) increase ACTH and GH production. ACTH stimulates the adrenal gland, increasing cortisol, which in turn helps sustain blood pressure and dampen the inflammatory response. Increased vasopressin acts to conserve free water.

REPRODUCTION

The stages of reproduction include (1) sex determination during fetal development **(Chap. 383)**; (2) sexual maturation during puberty **(Chaps. 384 and 385)**; (3) conception, pregnancy, lactation, and child rearing **(Chap. 385)**; and (4) cessation of reproductive capability at menopause **(Chap. 388)**. Each of these stages involves an orchestrated interplay of multiple hormones, a phenomenon well illustrated by the dynamic hormonal changes that occur during each 28-day menstrual cycle. In the early follicular phase, pulsatile secretion of LH and FSH stimulates the progressive maturation of the ovarian follicle. This results in gradually increasing estrogen and progesterone levels, leading to enhanced pituitary sensitivity to GnRH, which, when combined with accelerated GnRH secretion, triggers the LH surge and rupture of the mature follicle. Inhibin, a protein produced by the granulosa cells, enhances follicular growth and feeds back to the pituitary to selectively suppress FSH without affecting LH. Growth factors such as EGF and IGF-I modulate follicular responsiveness to gonadotropins. Vascular endothelial growth factor and prostaglandins play a role in follicle vascularization and rupture.

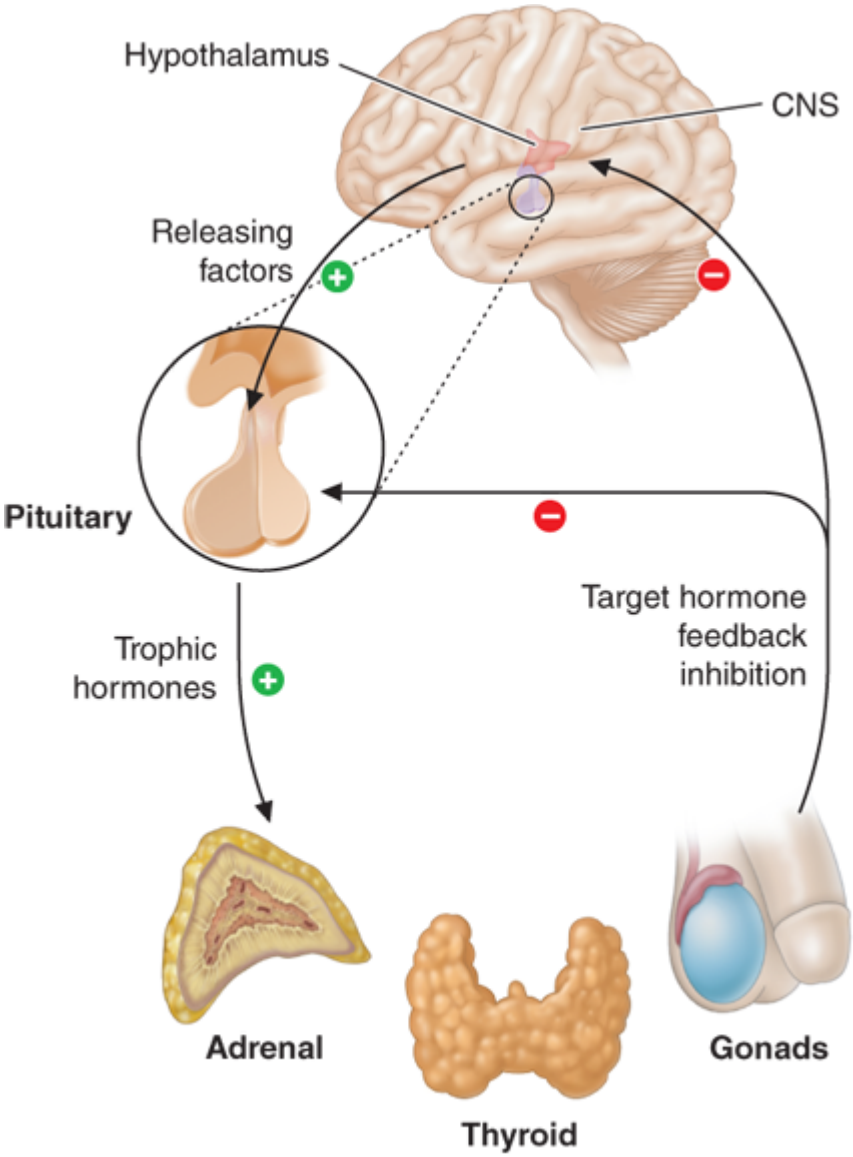
During pregnancy, the increased production of prolactin, in combination with placentally derived steroids (e.g., estrogen and progesterone), prepares the breast for lactation. Estrogens induce the production of progesterone receptors, allowing for increased responsiveness to progesterone. In addition to these and other hormones involved in lactation, the nervous system and oxytocin mediate the suckling response and milk release.

HORMONAL FEEDBACK REGULATORY SYSTEMS

Feedback control, both negative and positive, is a fundamental feature of endocrine systems. Each of the major hypothalamic-pituitary-hormone axes is governed by negative feedback, a process that maintains hormone levels within a relatively narrow range **(Chap. 371)**. Examples of hypothalamic-pituitary negative feedback include (1) thyroid hormones on the TRH-TSH axis, (2) cortisol on the CRH-ACTH axis, (3) gonadal steroids on the GnRH-LH/FSH axis, and (4) IGF-I on the GHRH-GH axis **(Fig. 370-4)**. These regulatory loops include both positive (e.g., TRH, TSH) and negative (e.g., T_4, T_3) components, allowing for exquisite control of hormone levels. As an example, a small reduction of thyroid hormone triggers a rapid increase of TRH and TSH secretion, resulting in thyroid gland stimulation and increased thyroid hormone production. When thyroid hormone reaches a normal level, it feeds back to suppress TRH and TSH, and a new steady state is attained. Feedback regulation also occurs for endocrine systems that do not involve the pituitary gland, such as calcium feedback on PTH, glucose inhibition of insulin secretion, and leptin feedback on the hypothalamus. An understanding of feedback regulation provides important insights into endocrine testing paradigms (see below).

FIGURE 370-4

Feedback regulation of endocrine axes. CNS, central nervous system.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Positive feedback control also occurs but is not well understood. The primary example is estrogen-mediated stimulation of the midcycle LH surge. Although chronic low levels of estrogen are inhibitory, gradually rising estrogen levels stimulate LH secretion. This effect, which is illustrative of an endocrine rhythm (see below), involves activation of the hypothalamic GnRH pulse generator. In addition, estrogen-primed gonadotropes are extraordinarily sensitive to GnRH, leading to amplification of LH release.

PARACRINE AND AUTOCRINE CONTROL

The previously mentioned examples of feedback control involve classic endocrine pathways in which hormones are released by one gland and act on a distant target gland. However, local regulatory systems, often involving growth factors, are increasingly recognized. *Paracrine regulation* refers to factors released by one cell that act on an adjacent cell in the same tissue. For example, somatostatin secretion by pancreatic islet δ cells inhibits insulin secretion from nearby β cells. *Autocrine regulation* describes the action of a factor

on the same cell from which it is produced. IGF-I acts on many cells that produce it, including chondrocytes, breast epithelium, and gonadal cells. Unlike endocrine actions, paracrine and autocrine control are difficult to document because local growth factor concentrations cannot be measured readily.

Anatomic relationships of glandular systems also greatly influence hormonal exposure: the physical organization of islet cells enhances their intercellular communication; the portal vasculature of the hypothalamic-pituitary system exposes the pituitary to high concentrations of hypothalamic releasing factors; testicular seminiferous tubules gain exposure to high testosterone levels produced by the interdigitated Leydig cells; the pancreas receives nutrient information and local exposure to peptide hormones (incretins) from the gastrointestinal tract; and the liver is the proximal target of insulin action because of portal drainage from the pancreas.

HORMONAL RHYTHMS

The feedback regulatory systems described above are superimposed on hormonal rhythms that are used for adaptation to the environment. Seasonal changes, the daily occurrence of the light-dark cycle, sleep, meals, and stress are examples of the many environmental events that affect hormonal rhythms. The *menstrual cycle* is repeated on average every 28 days, reflecting the time required to follicular maturation and ovulation (Chap. 385). Essentially all pituitary hormone rhythms are entrained to sleep and to the *circadian cycle*, generating reproducible patterns that are repeated approximately every 24 h. The HPA axis, for example, exhibits characteristic peaks of ACTH and cortisol production in the early morning, with a nadir during the night. Recognition of these rhythms is important for endocrine testing and treatment. Patients with Cushing's syndrome characteristically exhibit increased midnight cortisol levels compared with normal individuals (Chap. 379). In contrast, morning cortisol levels are similar in these groups, as cortisol is normally high at this time of day in normal individuals. The HPA axis is more susceptible to suppression by glucocorticoids administered at night as they blunt the early-morning rise of ACTH. Understanding these rhythms allows glucocorticoid replacement that mimics diurnal production by administering larger doses in the morning than in the afternoon. Disrupted sleep rhythms can alter hormonal regulation. For example, sleep deprivation causes mild insulin resistance, food craving, and hypertension, which are reversible, at least in the short term. Emerging evidence indicates that circadian clock pathways not only regulate sleep-wake cycles but also play important roles in virtually every cell type. For example, tissue-specific deletion of clock genes alters rhythms and levels of gene expression, as well as metabolic responses in liver, adipose, and other tissues.

Other endocrine rhythms occur on a more rapid time scale. Many peptide hormones are secreted in discrete bursts every few hours. LH and FSH secretion are exquisitely sensitive to GnRH pulse frequency. Intermittent pulses of GnRH are required to maintain pituitary sensitivity, whereas continuous exposure to GnRH causes pituitary gonadotrope desensitization. This feature of the hypothalamic-pituitary-gonadotrope axis forms the basis for using long-acting GnRH agonists to treat central precocious puberty or to decrease testosterone levels in the management of prostate cancer. It is important to be aware of the pulsatile nature of hormone secretion and the rhythmic patterns of hormone production in relating serum hormone measurements to

normal values. For some hormones, integrated markers have been developed to circumvent hormonal fluctuations. Examples include 24-h urine collections for cortisol, IGF-I as a biologic marker of GH action, and HbA_{1c} as an index of long-term (weeks to months) blood glucose control.

Often, one must interpret endocrine data only in the context of other hormones. For example, PTH levels typically are assessed in combination with serum calcium concentrations. A high serum calcium level in association with elevated PTH is suggestive of hyperparathyroidism, whereas a suppressed PTH in this situation is more likely to be caused by hypercalcemia of malignancy or other causes of hypercalcemia. Similarly, TSH should be elevated when T_4 and T_3 concentrations are low, reflecting reduced feedback inhibition. When this is not the case, it is important to consider secondary hypothyroidism, which is caused by a defect at the level of the pituitary.

FURTHER READING

Evans RM, Mangelsdorf DJ: Nuclear receptors, RXR, and the big bang. Cell 157:255, 2014. [PubMed: 24679540]

Jameson JL, DeGroot LJ (eds): *Endocrinology: Adult and Pediatric*, 7th ed. Philadelphia, Elsevier, 2016.

Melmed S et al (eds): *Williams Textbook of Endocrinology*, 13th ed. Philadelphia, Elsevier, 2016.

Moran C, Chatterjee K: Resistance to thyroid hormone α-emerging definition of a disorder of the thyroid hormone action. J Clin Endocrinol Metab 101:2317, 2016.

Vassart G, Costagliola S: G protein-coupled receptors: Mutations and endocrine diseases. Nature Rev Endocrinol 7:362, 2011.

Chapter 371: Physiology of Anterior Pituitary Hormones

Shlomo Melmed; J. Larry Jameson

INTRODUCTION

The anterior pituitary often is referred to as the “master gland” because, together with the hypothalamus, it orchestrates the complex regulatory functions of many other endocrine glands. The anterior pituitary gland produces six major hormones: (1) prolactin (PRL), (2) growth hormone (GH), (3) adrenocorticotropic hormone (ACTH), (4) luteinizing hormone (LH), (5) follicle-stimulating hormone (FSH), and (6) thyroid-stimulating hormone (TSH) **(Table 371-1)**. Pituitary hormones are secreted in a pulsatile manner, reflecting regulation by an array of specific hypothalamic releasing factors. Each of these pituitary hormones elicits specific trophic responses in peripheral target tissues. The hormonal products of those peripheral glands, in turn, exert feedback control at the level of the hypothalamus and pituitary to modulate pituitary function **(Fig. 371-1)**. Pituitary tumors cause characteristic hormone excess syndromes. Hormone deficiency may be inherited or acquired. Fortunately, there are efficacious treatments for many pituitary hormone excess and deficiency syndromes. Nonetheless, these diagnoses are often elusive; this emphasizes the importance of recognizing subtle clinical manifestations and performing the correct laboratory diagnostic tests. **For discussion of disorders of the posterior pituitary, or neurohypophysis, see Chap. 374.**

TABLE 371-1

Anterior Pituitary Hormone Expression and Regulation

CELL	CORTICOTROPE	SOMATOTROPE	LACTOTROPE	THYROTROPE	GONADOTROPE
Tissue-specific transcription factor	T-Pit	Prop-1, Pit-1	Prop-1, Pit-1	Prop-1, Pit-1, TEF	SF-1, DAX-1
Fetal appearance	6 weeks	8 weeks	12 weeks	12 weeks	12 weeks
Hormone	POMC	GH	PRL	TSH	FSH, LH
Protein	Polypeptide	Polypeptide	Polypeptide	Glycoprotein α, β subunits	Glycoprotein α, β subunits
Amino acids	266 (ACTH 1–39)	191	198	211	210, 204
Stimulators	CRH, AVP, gp-130 cytokines	GHRH, ghrelin	Estrogen, TRH, VIP	TRH	GnRH, activins, estrogen
Inhibitors	Glucocorticoids	Somatostatin, IGF-I	Dopamine	T_3, T_4, dopamine, somatostatin, glucocorticoids	Sex steroids, inhibin
Target gland	Adrenal	Liver, bone, other tissues	Breast, other tissues	Thyroid	Ovary, testis
Trophic effect	Steroid production	IGF-I production, growth induction, insulin antagonism	Milk production	T_4 synthesis and secretion	Sex steroid production, follicle growth, germ cell maturation

CELL	CORTICOTROPE	SOMATOTROPE	LACTOTROPE	THYROTROPE	GONADOTROPE
Normal range	ACTH, 4–22 pg/L	<0.5 µg/L[a]	M <15 µg/L; F <20 µg/L	0.1–5 mU/L	M, 5–20 IU/L; F (basal), 5–20 IU/L

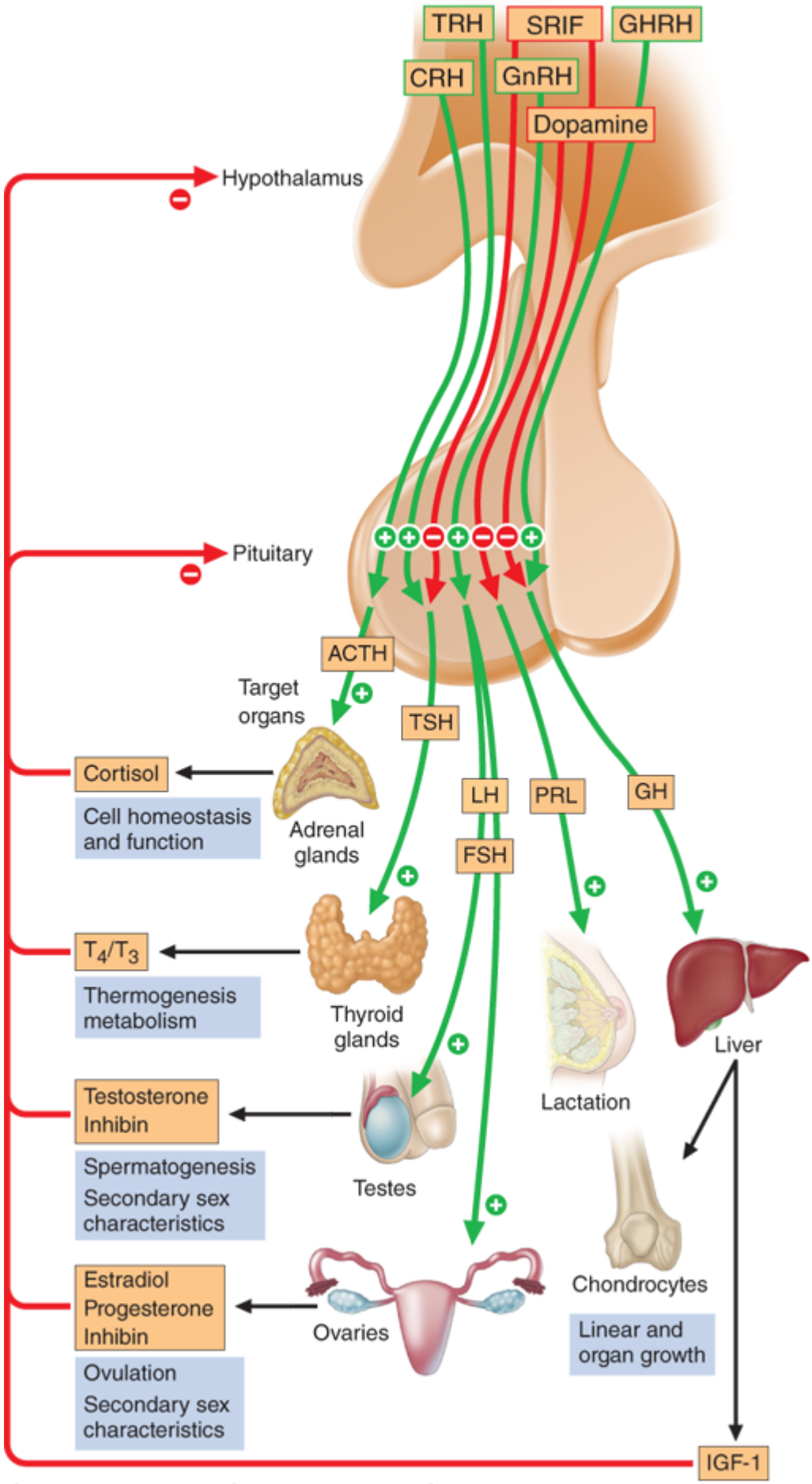
[a]Hormone secretion integrated over 24 h.

Abbreviations: M, male; F, female. For other abbreviations, see text.

Source: Adapted from S Melmed: Hypothalmic-pituitary regulation, in P Conn (ed): Conn's Translational Neuroscience. San Diego, CA, Elsevier, 2017.

FIGURE 371-1

Diagram of pituitary axes. Hypothalamic hormones regulate anterior pituitary trophic hormones that in turn determine target gland secretion. Peripheral hormones feed back to regulate hypothalamic and pituitary hormones. For abbreviations, see text.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ANATOMY AND DEVELOPMENT

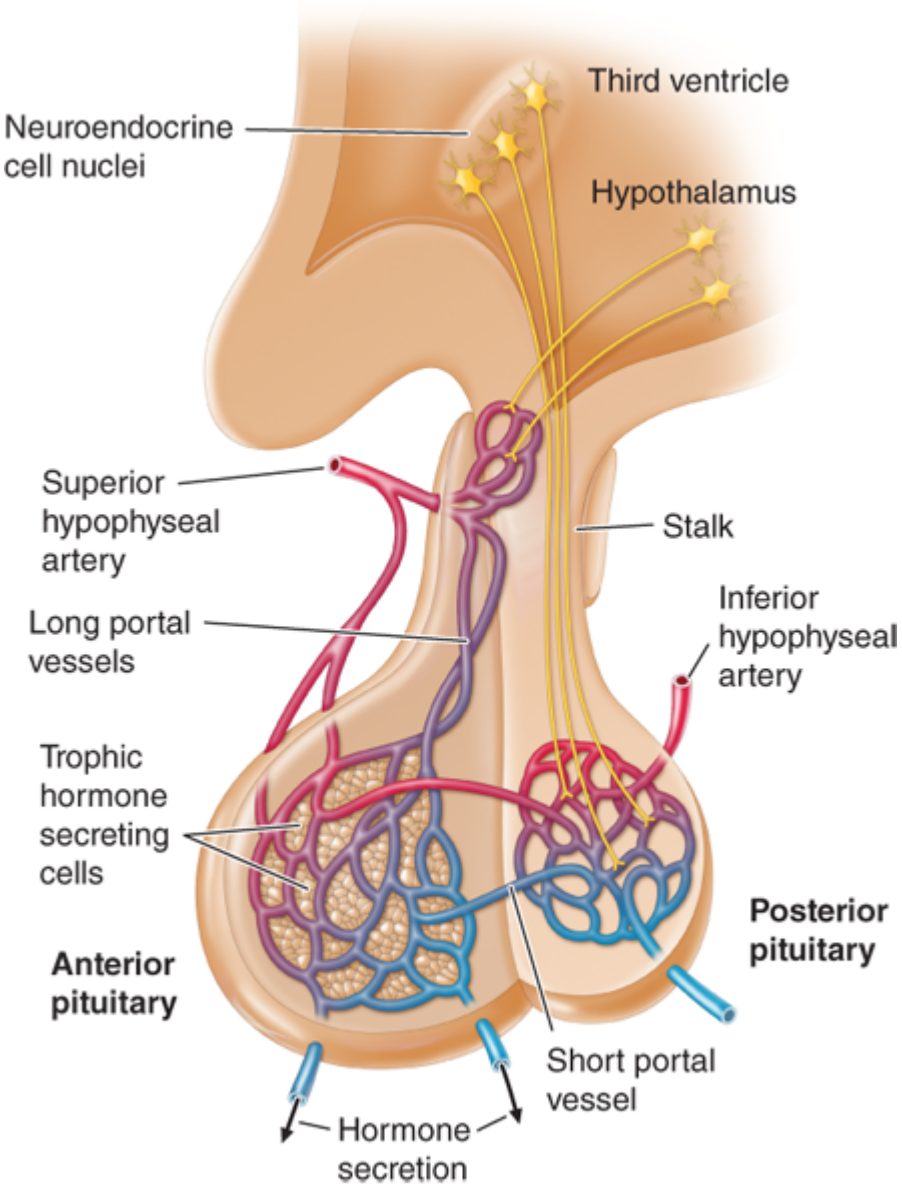
ANATOMY

The pituitary gland weighs ~600 mg and is located within the sella turcica ventral to the diaphragma sella; it consists of anatomically and functionally distinct anterior and posterior lobes. The bony sella is contiguous to vascular and neurologic structures, including the cavernous sinuses, cranial nerves, and optic chiasm. Thus, expanding intrasellar pathologic processes may have significant central mass effects in addition to their endocrinologic impact.

Hypothalamic neural cells synthesize specific releasing and inhibiting hormones that are secreted directly into the portal vessels of the pituitary stalk. Blood supply of the pituitary gland comes from the superior and inferior hypophyseal arteries **(Fig. 371-2)**. The hypothalamic-pituitary portal plexus provides the major blood source for the anterior pituitary, allowing reliable transmission of hypothalamic peptide pulses without significant systemic dilution; consequently, pituitary cells are exposed to releasing or inhibiting factors and in turn release their respective hormones as discrete pulses into the systemic circulation **(Fig. 371-3)**.

FIGURE 371-2

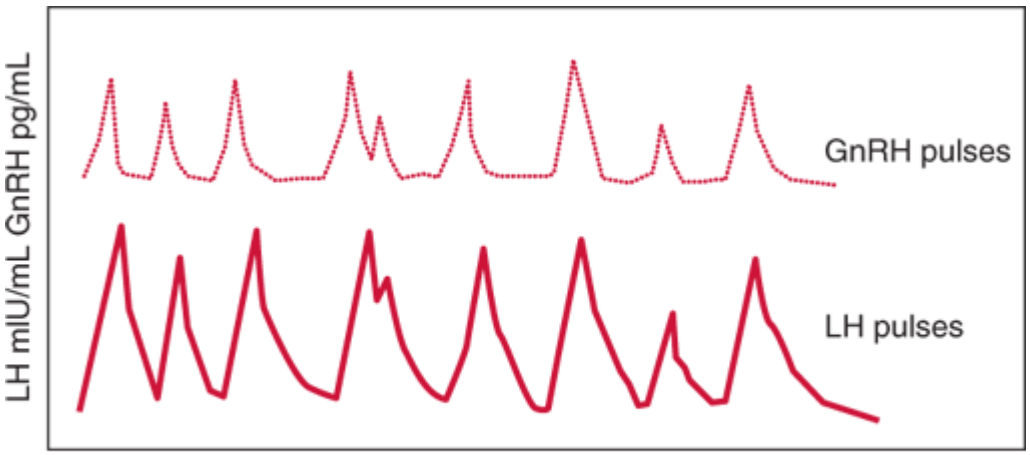
Diagram of hypothalamic-pituitary vasculature. The hypothalamic nuclei produce hormones that traverse the portal system and impinge on anterior pituitary cells to regulate pituitary hormone secretion. Posterior pituitary hormones are derived from direct neural extensions.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 371-3

Hypothalamic gonadotropin-releasing hormone (GnRH) pulses induce secretory pulses of luteinizing hormone (LH).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The posterior pituitary is supplied by the inferior hypophyseal arteries. In contrast to the anterior pituitary, the posterior lobe is directly innervated by hypothalamic neurons (supraopticohypophyseal and tuberohypophyseal nerve tracts) via the pituitary stalk **(Chap. 374)**. Thus, posterior pituitary production of vasopressin (antidiuretic hormone [ADH]) and oxytocin is particularly sensitive to neuronal damage by lesions that affect the pituitary stalk or hypothalamus.

PITUITARY DEVELOPMENT

The embryonic differentiation and maturation of anterior pituitary cells have been elucidated in considerable detail. Pituitary development from Rathke's pouch involves a complex interplay of lineage-specific transcription factors expressed in pluripotent precursor cells and gradients of locally produced growth factors (Table 371-1). The transcription factor Prop-1 induces pituitary development of Pit-1-specific lineages as well as gonadotropes. The transcription factor Pit-1 determines cell-specific expression of GH, PRL, and TSH in somatotropes, lactotropes, and thyrotropes. Expression of high levels of estrogen receptors in cells that contain Pit-1 favors PRL expression, whereas thyrotrope embryonic factor (TEF) induces TSH expression. Pit-1 binds to GH, PRL, and TSH gene regulatory elements as well as to recognition sites on its own promoter, providing a mechanism for determining specific pituitary hormone phenotypic stability. Gonadotrope cell development is further defined by the cell-specific expression of the nuclear receptors steroidogenic factor (SF-1) and *d*osage-sensitive sex reversal, *a*drenal hypoplasia critical region, on chromosome *X*, gene *1* (DAX-1). Development of corticotrope cells, which express the proopiomelanocortin (*POMC*) gene, requires the T-Pit transcription factor. Abnormalities of pituitary development caused by mutations of Pit-1, Prop-1, SF-1, DAX-1, and T-Pit result in rare, selective or combined pituitary hormone deficit syndromes.

ANTERIOR PITUITARY HORMONES

Each anterior pituitary hormone is under unique control, and each exhibits highly specific normal and dysregulated secretory characteristics.

PROLACTIN

Synthesis

PRL consists of 198 amino acids and has a molecular mass of 21,500 kDa; it is weakly homologous to GH and human placental lactogen (hPL), reflecting the duplication and divergence of a common GH-PRL-hPL precursor gene. PRL is synthesized in lactotropes, which constitute about 20% of anterior pituitary cells. Lactotropes and somatotropes are derived from a common precursor cell that may give rise to a tumor that secretes both PRL and GH. Marked lactotrope cell hyperplasia develops during pregnancy and the first few months of lactation. These transient functional changes in the lactotrope population are induced by estrogen.

Secretion

Normal adult serum PRL levels are about 10–25 µg/L in women and 10–20 µg/L in men. PRL secretion is pulsatile, with the highest secretory peaks occurring during rapid eye movement sleep. Peak serum PRL levels (up to 30 µg/L) occur between 4:00 and 6:00 A.M. The circulating half-life of PRL is about 50 min.

PRL is unique among the pituitary hormones in that the predominant central control mechanism is inhibitory, reflecting tonic dopamine-mediated suppression of PRL release. This regulatory pathway accounts for the spontaneous PRL hypersecretion that occurs with pituitary stalk section, often a consequence of head trauma or compressive mass lesions at the skull base. Pituitary dopamine type 2 (D_2) receptors mediate inhibition of PRL synthesis and secretion. Targeted disruption (gene knockout) of the murine D_2 receptor in mice results in hyperprolactinemia and lactotrope proliferation. As discussed below, dopamine agonists play a central role in the management of hyperprolactinemic disorders.

Thyrotropin-releasing hormone (TRH) (pyro Glu-His-Pro-NH_2) is a hypothalamic tripeptide that elicits PRL release within 15–30 min after intravenous injection. TRH primarily regulates TSH, and the physiologic relevance of TRH for PRL regulation is unclear **(Chap. 375)**. *Vasoactive intestinal peptide* (VIP) also induces PRL release, whereas glucocorticoids and thyroid hormone weakly suppress PRL secretion.

Serum PRL levels rise transiently after exercise, meals, sexual intercourse, minor surgical procedures, general anesthesia, chest wall injury, acute myocardial infarction, and other forms of acute stress. PRL levels increase markedly (about tenfold) during pregnancy and decline rapidly within 2 weeks of parturition. If breast-feeding is initiated, basal PRL levels remain elevated; suckling stimulates transient reflex increases in PRL levels that last for about 30–45 min. Breast suckling activates afferent neural pathways in the hypothalamus that induce PRL release. With time, suckling-induced responses diminish and interfeeding PRL levels return to normal.

Action

The PRL receptor is a member of the type I cytokine receptor family that also includes GH and interleukin (IL) 6 receptors. Ligand binding induces receptor dimerization and intracellular signaling by Janus kinase (JAK), which stimulates translocation of the signal transduction and activators of transcription (STAT) family to

activate target genes. Heterozygous mutations of the PRL receptor result in PRL insensitivity, hyperprolactinemia, and oligomenorrhea. In the breast, the lobuloalveolar epithelium proliferates in response to PRL, placental lactogens, estrogen, progesterone, and local paracrine growth factors, including insulin-like growth factor I (IGF-I).

PRL acts to induce and maintain lactation, decrease reproductive function, and suppress sexual drive. These functions are geared toward ensuring that maternal lactation is sustained and not interrupted by pregnancy. PRL inhibits reproductive function by suppressing hypothalamic gonadotropin-releasing hormone (GnRH) and pituitary gonadotropin secretion and by impairing gonadal steroidogenesis in both women and men. In the ovary, PRL blocks folliculogenesis and inhibits granulosa cell aromatase activity, leading to hypoestrogenism and anovulation. PRL also has a luteolytic effect, generating a shortened, or inadequate, luteal phase of the menstrual cycle. In men, attenuated LH secretion leads to low testosterone levels and decreased spermatogenesis. These hormonal changes decrease libido and reduce fertility in patients with hyperprolactinemia.

GROWTH HORMONE

Synthesis

GH is the most abundant anterior pituitary hormone, and GH-secreting somatotrope cells constitute up to 50% of the total anterior pituitary cell population. Mammosomatotrope cells, which coexpress PRL with GH, can be identified by using double immunostaining techniques. Somatotrope development and GH transcription are determined by expression of the cell-specific Pit-1 nuclear transcription factor. Five distinct genes encode GH and related proteins. The pituitary GH gene (*hGH-N*) produces two alternatively spliced products that give rise to 22-kDa GH (191 amino acids) and a less abundant 20-kDa GH molecule with similar biologic activity. Placental syncytiotrophoblast cells express a GH variant (*hGH-V*) gene; the related hormone human chorionic somatotropin (HCS) is expressed by distinct members of the gene cluster.

Secretion

GH secretion is controlled by complex hypothalamic and peripheral factors. *GH-releasing hormone* (GHRH) is a 44-amino-acid hypothalamic peptide that stimulates GH synthesis and release. Ghrelin, an octanoylated gastric-derived peptide, and synthetic agonists of the *GHS-R* induce GHRH and also directly stimulate GH release. *Somatostatin* (somatotropin-release inhibiting factor [SRIF]) is synthesized in the medial preoptic area of the hypothalamus and inhibits GH secretion. GHRH is secreted in discrete spikes that elicit GH pulses, whereas SRIF sets basal GH secretory tone. SRIF also is expressed in many extrahypothalamic tissues, including the central nervous system (CNS), gastrointestinal tract, and pancreas, where it also acts to inhibit islet hormone secretion. *IGF-I*, the peripheral target hormone for GH, feeds back to inhibit GH; estrogen induces GH, whereas chronic glucocorticoid excess suppresses GH release.

Surface receptors on the somatotrope regulate GH synthesis and secretion. The GHRH receptor is a G protein–coupled receptor (GPCR) that signals through the intracellular cyclic AMP pathway to stimulate

somatotrope cell proliferation as well as GH production. Inactivating mutations of the GHRH receptor cause profound dwarfism. A distinct surface receptor for ghrelin, the gastric-derived GH secretagogue, is expressed in both the hypothalamus and pituitary. Somatostatin binds to five distinct receptor subtypes (SST1 to SST5); SST2 and SST5 subtypes preferentially suppress GH (and TSH) secretion and SST5 signals to suppress ACTH secretion.

GH secretion is pulsatile, with highest peak levels occurring at night, generally correlating with sleep onset. GH secretory rates decline markedly with age so that hormone levels in middle age are about 15% of pubertal levels. These changes are paralleled by an age-related decline in lean muscle mass. GH secretion is also reduced in obese individuals, although IGF-I levels may not be suppressed, suggesting a change in the setpoint for feedback control. Elevated GH levels occur within an hour of deep sleep onset as well as after exercise, physical stress, and trauma and during sepsis. Integrated 24-h GH secretion is higher in women and is also enhanced by estrogen replacement, likely reflective of increased peripheral GH resistance. Using standard assays, random GH measurements are undetectable in ~50% of daytime samples obtained from healthy subjects and are also undetectable in most obese and elderly subjects. Thus, single random GH measurements do not distinguish patients with adult GH deficiency from normal persons.

GH secretion is profoundly influenced by nutritional factors. Using newer ultrasensitive GH assays with a sensitivity of 0.002 μg/L, a glucose load suppresses GH to <0.7 μg/L in women and to <0.07 μg/L in men. Increased GH pulse frequency and peak amplitudes occur with chronic malnutrition or prolonged fasting. GH is stimulated by intravenous L-arginine, dopamine, and apomorphine (a dopamine receptor agonist), as well as by α-adrenergic pathways. β-Adrenergic blockade induces basal GH and enhances GHRH- and insulin-evoked GH release.

Action

The pattern of GH secretion may affect tissue responses. The higher GH pulsatility observed in men compared with the relatively continuous basal GH secretion in women may be an important biologic determinant of linear growth patterns and liver enzyme induction.

The 70-kDa peripheral GH receptor protein has structural homology with the cytokine/hematopoietic superfamily. A fragment of the receptor extracellular domain generates a soluble GH binding protein (GHBP) that interacts with GH in the circulation. The liver and cartilage express the greatest number of GH receptors. GH binding to preformed receptor dimers is followed by internal rotation and subsequent signaling through the JAK/STAT pathway. Activated STAT proteins translocate to the nucleus, where they modulate expression of GH-regulated target genes. GH analogues that bind to the receptor but are incapable of mediating receptor signaling are potent antagonists of GH action. A GH receptor antagonist (pegvisomant) is approved for treatment of acromegaly.

GH induces protein synthesis and nitrogen retention and also impairs glucose tolerance by antagonizing insulin action. GH also stimulates lipolysis, leading to increased circulating fatty acid levels, reduced omental fat mass, and enhanced lean body mass. GH promotes sodium, potassium, and water retention and elevates

serum levels of inorganic phosphate. Linear bone growth occurs as a result of complex hormonal and growth factor actions, including those of IGF-I. GH stimulates epiphyseal prechondrocyte differentiation. These precursor cells produce IGF-I locally, and their proliferation is also responsive to the growth factor.

Insulin-Like Growth Factors

Although GH exerts direct effects in target tissues, many of its physiologic effects are mediated indirectly through IGF-I, a potent growth and differentiation factor. The liver is the major source of circulating IGF-I. In peripheral tissues, IGF-I also exerts local paracrine actions that appear to be both dependent on and independent of GH. Thus, GH administration induces circulating IGF-I as well as stimulating local IGF-I production in multiple tissues.

Both IGF-I and IGF-II are bound to high-affinity circulating IGF-binding proteins (IGFBPs) that regulate IGF availability and bioactivity. Levels of IGFBP3 are GH-dependent, and it serves as the major carrier protein for circulating IGF-I. GH deficiency and malnutrition usually are associated with low IGFBP3 levels. IGFBP1 and IGFBP2 regulate local tissue IGF action but do not bind appreciable amounts of circulating IGF-I.

Serum IGF-I concentrations are profoundly affected by physiologic factors. Levels increase during puberty, peak at 16 years, and subsequently decline by >80% during the aging process. IGF-I concentrations are higher in women than in men. Because GH is the major determinant of hepatic IGF-I synthesis, abnormalities of GH synthesis or action (e.g., pituitary failure, GHRH receptor defect, GH receptor defect or pharmacologic GH receptor blockade) reduce IGF-I levels. Hypocaloric states are associated with GH resistance; IGF-I levels are therefore low with cachexia, malnutrition, and sepsis. In acromegaly, IGF-I levels are invariably high and reflect a log-linear relationship with circulating GH concentrations.

IGF-I PHYSIOLOGY

Injected IGF-I (100 μg/kg) induces hypoglycemia, and lower doses improve insulin sensitivity in patients with severe insulin resistance and diabetes. In cachectic subjects, IGF-I infusion (12 μg/kg per h) enhances nitrogen retention and lowers cholesterol levels. Longer-term subcutaneous IGF-I injections enhance protein synthesis and are anabolic. Although bone formation markers are induced, bone turnover also may be stimulated by IGF-I. IGF-I is approved for use in patients with GH-resistance syndromes.

IGF-I side effects are dose-dependent, and overdose may result in hypoglycemia, hypotension, fluid retention, temporomandibular jaw pain, and increased intracranial pressure, all of which are reversible. Avascular femoral head necrosis has been reported. Chronic excess IGF-I administration presumably would result in features of acromegaly.

ADRENOCORTICOTROPIC HORMONE

(See also Chap. 379)

Synthesis

ACTH-secreting corticotrope cells constitute about 20% of the pituitary cell population. ACTH (39 amino acids) is derived from the POMC precursor protein (266 amino acids) that also generates several other peptides, including β-lipotropin, β-endorphin, met-enkephalin, α-melanocyte-stimulating hormone (α-MSH), and corticotropin-like intermediate lobe protein (CLIP). The *POMC* gene is potently suppressed by glucocorticoids and induced by corticotropin-releasing hormone (CRH), arginine vasopressin (AVP), and proinflammatory cytokines, including IL-6, as well as leukemia inhibitory factor.

CRH, a 41-amino-acid hypothalamic peptide synthesized in the paraventricular nucleus as well as in higher brain centers, is the predominant stimulator of ACTH synthesis and release. The CRH receptor is a GPCR that is expressed on the corticotrope and signals to induce *POMC* transcription.

Secretion

ACTH secretion is pulsatile and exhibits a characteristic circadian rhythm, peaking at about 6:00 A.M. and reaching a nadir about midnight. Adrenal glucocorticoid secretion, which is driven by ACTH, follows a parallel diurnal pattern. ACTH circadian rhythmicity is determined by variations in secretory pulse amplitude rather than changes in pulse frequency. Superimposed on this endogenous rhythm, ACTH levels are increased by physical and psychological stress, exercise, acute illness, and insulin-induced hypoglycemia.

Glucocorticoid-mediated negative regulation of the hypothalamic-pituitary-adrenal (HPA) axis occurs as a consequence of both hypothalamic CRH suppression and direct attenuation of pituitary *POMC* gene expression and ACTH release. In contrast, loss of cortisol feedback inhibition, as occurs in primary adrenal failure, results in extremely high ACTH levels.

Acute inflammatory or septic insults activate the HPA axis through the integrated actions of proinflammatory cytokines, bacterial toxins, and neural signals. The overlapping cascade of ACTH-inducing cytokines (tumor necrosis factor [TNF]; IL-1, -2, and -6; and leukemia inhibitory factor) activates hypothalamic CRH and AVP secretion, pituitary *POMC* gene expression, and local pituitary paracrine cytokine networks. The resulting cortisol elevation restrains the inflammatory response and enables host protection. Concomitantly, cytokine-mediated central glucocorticoid receptor resistance impairs glucocorticoid suppression of the HPA. Thus, the neuroendocrine stress response reflects the net result of highly integrated hypothalamic, intrapituitary, and peripheral hormone and cytokine signals acting to regulate cortisol secretion.

Action

The major function of the HPA axis is to maintain metabolic homeostasis and mediate the neuroendocrine stress response. ACTH induces adrenocortical steroidogenesis by sustaining adrenal cell proliferation and function. The receptor for ACTH, designated *melanocortin-2 receptor*, is a GPCR that induces steroidogenesis by stimulating a cascade of steroidogenic enzymes **(Chap. 379)**.

GONADOTROPINS: FSH AND LH

Synthesis and Secretion

Gonadotrope cells constitute about 10% of anterior pituitary cells and produce two gonadotropin hormones—LH and FSH. Like TSH and hCG, LH and FSH are glycoprotein hormones that comprise α and β subunits. The α subunit is common to these glycoprotein hormones; specificity of hormone function is conferred by the β subunits, which are expressed by separate genes.

Gonadotropin synthesis and release are dynamically regulated. This is particularly true in women, in whom rapidly fluctuating gonadal steroid levels vary throughout the menstrual cycle. Hypothalamic GnRH, a 10-amino-acid peptide, regulates the synthesis and secretion of both LH and FSH. Brain kisspeptin, a product of the *KISS1* gene, regulates hypothalamic GnRH release. GnRH is secreted in discrete pulses every 60–120 min, and the pulses in turn elicit LH and FSH pulses (Fig. 371-3). The pulsatile mode of GnRH input is essential to its action; pulses prime gonadotrope responsiveness, whereas continuous GnRH exposure induces desensitization. Based on this phenomenon, long-acting GnRH agonists are used to suppress gonadotropin levels in children with precocious puberty and in men with prostate cancer **(Chap. 83)** and are used in some ovulation-induction protocols to reduce levels of endogenous gonadotropins **(Chap. 385)**. Estrogens act at both the hypothalamus and the pituitary to modulate gonadotropin secretion. Chronic estrogen exposure is inhibitory, whereas rising estrogen levels, as occur during the preovulatory surge, exert positive feedback to increase gonadotropin pulse frequency and amplitude. Progesterone slows GnRH pulse frequency but enhances gonadotropin responses to GnRH. Testosterone feedback in men also occurs at the hypothalamic and pituitary levels and is mediated in part by its conversion to estrogens.

Although GnRH is the main regulator of LH and FSH secretion, FSH synthesis is also under separate control by the gonadal peptides inhibin and activin, which are members of the transforming growth factor β (TGF-β) family. Inhibin selectively suppresses FSH, whereas activin stimulates FSH synthesis **(Chap. 385)**.

Action

The gonadotropin hormones interact with their respective GPCRs expressed in the ovary and testis, evoking germ cell development and maturation and steroid hormone biosynthesis. In women, FSH regulates ovarian follicle development and stimulates ovarian estrogen production. LH mediates ovulation and maintenance of the corpus luteum. In men, LH induces Leydig cell testosterone synthesis and secretion, and FSH stimulates seminiferous tubule development and regulates spermatogenesis.

THYROID-STIMULATING HORMONE

Synthesis and Secretion

TSH-secreting thyrotrope cells constitute 5% of the anterior pituitary cell population. TSH shares a common α subunit with LH and FSH but contains a specific TSH β subunit. TRH is a hypothalamic tripeptide (pyroglutamyl histidylprolinamide) that acts through a pituitary GPCR to stimulate TSH synthesis and secretion; it also stimulates the lactotrope cell to secrete PRL. TSH secretion is stimulated by TRH, whereas thyroid hormones, dopamine, somatostatin, and glucocorticoids suppress TSH by overriding TRH induction.

Thyrotrope cell proliferation and TSH secretion are both induced when negative feedback inhibition by thyroid hormones is removed. Thus, thyroid damage (including surgical thyroidectomy), radiation-induced hypothyroidism, chronic thyroiditis, and prolonged goitrogen exposure are associated with increased TSH levels. Long-standing untreated hypothyroidism can lead to elevated TSH levels, which may be associated with thyrotrope hyperplasia and pituitary enlargement and may sometimes be evident on magnetic resonance imaging.

Action

TSH is secreted in pulses, although the excursions are modest in comparison to other pituitary hormones because of the low amplitude of the pulses and the relatively long half-life of TSH. Consequently, single determinations of TSH suffice to precisely assess its circulating levels. TSH binds to a GPCR on thyroid follicular cells to stimulate thyroid hormone synthesis and release **(Chap. 375)**.

FURTHER READING

Bernard V et al: New insights in prolactin: Pathological implications. Nat Rev Endocrinol 11:265, 2015. [PubMed: 25781857]

Brooks AJ, Waters MJ: The growth hormone receptor: Mechanism of activation and clinical implications. Nat Rev Endocrinol 6:515, 2010. [PubMed: 20664532]

Ho K: The year in pituitary 2014. J Clin Endocrinol Metab 99:4449, 2014. [PubMed: 25322276]

Kelberman D et al: Genetic regulation of pituitary gland development in human and mouse. Endocr Rev 30:790, 2009. [PubMed: 19837867]

Langlais D et al: Adult pituitary cell maintenance: Lineage-specific contribution of self-duplication. Mol Endocrinol 27:1103, 2013. [PubMed: 23754801]

Murray PG, Higham CE, Clayton PE: 60 years of neuroendocrinology: The hypothalamo-GH axis: The past 60 years. J Endocrinol 226:T123, 2015. [PubMed: 26040485]

Newey PJ et al: Mutant prolactin receptor and familial hyperprolactinemia. N Engl J Med 369:2012, 2013.

Chapter 372: Hypopituitarism

Shlomo Melmed; J. Larry Jameson

FIGURE 372-1

INTRODUCTION

Inadequate production of anterior pituitary hormones leads to features of hypopituitarism. Impaired production of one or more of the anterior pituitary trophic hormones can result from inherited disorders; more commonly, adult hypopituitarism is acquired and reflects the compressive mass effects of tumors or the consequences of local pituitary or hypothalamic traumatic, inflammatory, or vascular damage. These processes also may impair synthesis or secretion of hypothalamic hormones, with resultant pituitary failure (Table 372-1).

TABLE 372-1

Etiology of Hypopituitarism[a]

Development/structural
Transcription factor defect
Pituitary dysplasia/aplasia
Congenital central nervous system mass, encephalocele
Primary empty sella
Congenital hypothalamic disorders (septo-optic dysplasia, Prader-Willi syndrome, Bardet-Biedl syndrome, Kallmann syndrome)
Traumatic
Surgical resection
Radiation damage
Head injuries
Neoplastic
Pituitary adenoma
Parasellar mass (germinoma, ependymoma, glioma)
Rathke's cyst
Craniopharyngioma
Hypothalamic hamartoma, gangliocytoma
Pituitary metastases (breast, lung, colon carcinoma)
Lymphoma and leukemia
Meningioma
Infiltrative/inflammatory
Lymphocytic hypophysitis
Hemochromatosis
Sarcoidosis
Histiocytosis X
Granulomatous hypophysitis
Transcription factor antibodies
Immunotherapy
Vascular
Pituitary apoplexy
Pregnancy-related (infarction with diabetes; postpartum necrosis)
Sickle cell disease
Arteritis
Infections
Fungal (histoplasmosis)
Parasitic (toxoplasmosis)
Tuberculosis

Pneumocystis jirovecii

[a]Trophic hormone failure associated with pituitary compression or destruction usually occurs sequentially: growth hormone > follicle-stimulating hormone > luteinizing hormone > thyroid-stimulating hormone > adrenocorticotropic hormone. During childhood, growth retardation is often the presenting feature, and in adults, hypogonadism is the earliest symptom.

DEVELOPMENTAL AND GENETIC CAUSES OF HYPOPITUITARISM

Pituitary Dysplasia

Pituitary dysplasia may result in aplastic, hypoplastic, or ectopic pituitary gland development. Because pituitary development follows midline cell migration from the nasopharyngeal Rathke's pouch, midline craniofacial disorders may be associated with pituitary dysplasia. Acquired pituitary failure in the newborn also can be caused by birth trauma, including cranial hemorrhage, asphyxia, and breech delivery.

SEPTO-OPTIC DYSPLASIA

Hypothalamic dysfunction and hypopituitarism may result from dysgenesis of the septum pellucidum or corpus callosum. Affected children have mutations in the *HESX1* gene, which is involved in early development of the ventral prosencephalon. These children exhibit variable combinations of cleft palate, syndactyly, ear deformities, hypertelorism, optic nerve hypoplasia, micropenis, and anosmia. Pituitary dysfunction leads to diabetes insipidus, growth hormone (GH) deficiency and short stature, and, occasionally, thyroid-stimulating hormone (TSH) deficiency.

Tissue-Specific Factor Mutations

Several pituitary cell–specific transcription factors, such as Pit-1 and Prop-1, are critical for determining the development and committed function of differentiated anterior pituitary cell lineages. Autosomal dominant or recessive Pit-1 mutations cause combined GH, prolactin (PRL), and TSH deficiencies. These patients usually present with growth failure and varying degrees of hypothyroidism. The pituitary may appear hypoplastic on magnetic resonance imaging (MRI).

Prop-1 is expressed early in pituitary development and appears to be required for Pit-1 function. Familial and sporadic *PROP1* mutations result in combined GH, PRL, TSH, and gonadotropin deficiency. Over 80% of these patients have growth retardation; by adulthood, all are deficient in TSH and gonadotropins, and a small minority later develop adrenocorticotropic hormone (ACTH) deficiency. Because of gonadotropin deficiency, these individuals do not enter puberty spontaneously. In some cases, the pituitary gland appears enlarged on MRI. *TPIT* mutations result in ACTH deficiency associated with hypocortisolism.

Developmental Hypothalamic Dysfunction

KALLMANN SYNDROME

Kallmann syndrome results from defective hypothalamic gonadotropin-releasing hormone (GnRH) synthesis and is associated with anosmia or hyposmia due to olfactory bulb agenesis or hypoplasia **(Chap. 384)**. Classically, the syndrome may also be associated with color blindness, optic atrophy, nerve deafness, cleft palate, renal abnormalities, cryptorchidism, and neurologic abnormalities such as mirror movements. The initial genetic cause was identified in the X-linked *KAL* gene, mutations of which impair embryonic migration of GnRH neurons from the hypothalamic olfactory placode to the hypothalamus. Since then, at least a dozen additional genetic abnormalities, in addition to *KAL* mutations, have been found to cause isolated GnRH deficiency. Autosomal recessive (i.e., *GPR54, KISS1*) and dominant (i.e., *FGFR1*) modes of transmission have been described, and there is a growing list of genes associated with GnRH deficiency (including *GNRH1, PROK2, PROKR2, CHD7, PCSK1, FGF8, NELF, WDR11, TAC3, TACR3*, and *SEMA3E*). Some patients have oligogenic mutations. Associated clinical features, in addition to GnRH deficiency, vary depending on the genetic cause. GnRH deficiency prevents progression through puberty. Males present with delayed puberty and pronounced hypogonadal features, including micropenis, probably the result of low testosterone levels during infancy. Females present with primary amenorrhea and failure of secondary sexual development.

Kallmann syndrome and other causes of congenital GnRH deficiency are characterized by low luteinizing hormone (LH) and follicle-stimulating hormone (FSH) levels and low concentrations of sex steroids (testosterone or estradiol). In sporadic cases of isolated gonadotropin deficiency, the diagnosis is often one of exclusion after other known causes of hypothalamic-pituitary dysfunction have been eliminated. Repetitive GnRH administration restores normal pituitary gonadotropin responses, pointing to a hypothalamic defect in these patients.

Long-term treatment of males with human chorionic gonadotropin (hCG) or testosterone restores pubertal development and secondary sex characteristics; women can be treated with cyclic estrogen and progestin. Fertility may be restored by the administration of gonadotropins or by using a portable infusion pump to deliver subcutaneous, pulsatile GnRH.

BARDET-BIEDL SYNDROME

This very rare genetically heterogeneous disorder is characterized by mental retardation, renal abnormalities, obesity, and hexadactyly, brachydactyly, or syndactyly. Central diabetes insipidus may or may not be associated. GnRH deficiency occurs in 75% of males and half of affected females. Retinal degeneration begins in early childhood, and most patients are blind by age 30. Numerous subtypes of Bardet-Biedl syndrome (BBS) have been identified, with genetic linkage to at least nine different loci. Several of the loci encode genes involved in basal body cilia function, and this may account for the diverse clinical manifestations.

LEPTIN AND LEPTIN RECEPTOR MUTATIONS

Deficiencies of leptin or its receptor cause a broad spectrum of hypothalamic abnormalities, including hyperphagia, obesity, and central hypogonadism **(Chap. 394)**. Decreased GnRH production in these patients results in attenuated pituitary FSH and LH synthesis and release.

PRADER-WILLI SYNDROME

This is a contiguous gene syndrome that results from deletion of the paternal copies of the imprinted *SNRPN* gene, the *NECDIN* gene, and possibly other genes on chromosome 15q. Prader-Willi syndrome is associated with hypogonadotropic hypogonadism, hyperphagia-obesity, chronic muscle hypotonia, mental retardation, and adult-onset diabetes mellitus. Multiple somatic defects also involve the skull, eyes, ears, hands, and feet. Diminished hypothalamic oxytocin- and vasopressin-producing nuclei have been reported. Deficient GnRH synthesis is suggested by the observation that chronic GnRH treatment restores pituitary LH and FSH release.

ACQUIRED HYPOPITUITARISM

Hypopituitarism may be caused by accidental or neurosurgical trauma; vascular events such as apoplexy; pituitary or hypothalamic neoplasms, craniopharyngioma, lymphoma, or metastatic tumors; inflammatory disease such as lymphocytic hypophysitis; autoimmune hypophysitis associated with checkpoint inhibitor cancer immunotherapy; infiltrative disorders such as sarcoidosis, hemochromatosis (Chap. 407), and tuberculosis; or irradiation.

Increasing evidence suggests that patients with brain injury, including contact sports trauma, subarachnoid hemorrhage, and irradiation, have transient hypopituitarism and require intermittent long-term endocrine follow-up, because permanent hypothalamic or pituitary dysfunction will develop in 25–40% of these patients.

Hypothalamic Infiltration Disorders

These disorders—including sarcoidosis, histiocytosis X, amyloidosis, and hemochromatosis—frequently involve both hypothalamic and pituitary neuronal and neurochemical tracts. Consequently, diabetes insipidus occurs in half of patients with these disorders. Growth retardation is seen if attenuated GH secretion occurs before puberty. Hypogonadotropic hypogonadism and hyperprolactinemia are also common.

Inflammatory Lesions

Pituitary damage and subsequent secretory dysfunction can be seen with chronic site infections such as tuberculosis, with opportunistic fungal infections associated with AIDS, and in tertiary syphilis. Other inflammatory processes, such as granulomas and sarcoidosis, may mimic the features of a pituitary adenoma. These lesions may cause extensive hypothalamic and pituitary damage, leading to trophic hormone deficiencies.

Cranial Irradiation

Cranial irradiation may result in long-term hypothalamic and pituitary dysfunction, especially in children and adolescents, as they are more susceptible to damage after whole-brain or head and neck therapeutic irradiation. The development of hormonal abnormalities correlates strongly with irradiation dosage and the time interval after completion of radiotherapy. Up to two-thirds of patients ultimately develop hormone

insufficiency after a median dose of 50 Gy (5000 rad) directed at the skull base. The development of hypopituitarism occurs over 5–15 years and usually reflects hypothalamic damage rather than primary destruction of pituitary cells. Although the pattern of hormone loss is variable, GH deficiency is most common, followed by gonadotropin and ACTH deficiency. When deficiency of one or more hormones is documented, the possibility of diminished reserve of other hormones is likely. Accordingly, anterior pituitary function should be continually evaluated over the long term in previously irradiated patients, and replacement therapy instituted when appropriate (see below).

Lymphocytic Hypophysitis

This occurs most often in postpartum women; it usually presents with hyperprolactinemia and MRI evidence of a prominent pituitary mass that often resembles an adenoma, with mildly elevated PRL levels. Pituitary failure caused by diffuse lymphocytic infiltration may be transient or permanent but requires immediate evaluation and treatment. Rarely, isolated pituitary hormone deficiencies have been described, suggesting a selective autoimmune process targeted to specific cell types. Most patients manifest symptoms of progressive mass effects with headache and visual disturbance. The erythrocyte sedimentation rate often is elevated. Because the MRI image may be indistinguishable from that of a pituitary adenoma, hypophysitis should be considered in a postpartum woman with a newly diagnosed pituitary mass before an unnecessary surgical intervention is undertaken. The inflammatory process often resolves after several months of glucocorticoid treatment, and pituitary function may be restored, depending on the extent of damage.

Immunotherapy and Hypophysitis

Pituitary cells express cytotoxic T lymphocyte antigen-4 (CTLA-4) and up to 20% of patients receiving cancer immunotherapy with CTLA-4 blockers (e.g., ipilimumab) may develop hypophysitis with associated thyroid adrenal and gonadal failure. Pituitary hormone replacement, with or without high-dose glucocorticoids, may be safely tolerated with continued immunotherapy.

Pituitary Apoplexy

Acute intrapituitary hemorrhagic vascular events can cause substantial damage to the pituitary and surrounding sellar structures. Pituitary apoplexy may occur spontaneously in a preexisting adenoma; postpartum (Sheehan's syndrome); or in association with diabetes, hypertension, sickle cell anemia, or acute shock. The hyperplastic enlargement of the pituitary, which occurs normally during pregnancy, increases the risk for hemorrhage and infarction. Apoplexy is an endocrine emergency that may result in severe hypoglycemia, hypotension and shock, central nervous system (CNS) hemorrhage, and death. Acute symptoms may include severe headache with signs of meningeal irritation, bilateral visual changes, ophthalmoplegia, and, in severe cases, cardiovascular collapse and loss of consciousness. Pituitary computed tomography (CT) or MRI may reveal signs of intratumoral or sellar hemorrhage, with pituitary stalk deviation and compression of pituitary tissue.

Patients with no evident visual loss or impaired consciousness can be observed and managed conservatively with high-dose glucocorticoids. Those with significant or progressive visual loss, cranial nerve palsy, or loss of consciousness require urgent surgical decompression. Visual recovery after sellar surgery is inversely correlated with the length of time after the acute event. Therefore, severe ophthalmoplegia or visual deficits are indications for early surgery. Hypopituitarism is common after apoplexy.

Empty Sella

A partial or apparently totally empty sella is often an incidental MRI finding, and may be associated with intracranial hypertension. These patients usually have normal pituitary function, implying that the surrounding rim of pituitary tissue is fully functional. Hypopituitarism, however, may develop insidiously. Pituitary masses also may undergo clinically silent infarction and involution with development of a partial or totally empty sella by cerebrospinal fluid (CSF) filling the dural herniation. Rarely, small but functional pituitary adenomas may arise within the rim of normal pituitary tissue, and they are not always visible on MRI.

PRESENTATION AND DIAGNOSIS

The clinical manifestations of hypopituitarism depend on which hormones are lost and the extent of the hormone deficiency. GH deficiency causes growth disorders in children and leads to abnormal body composition in adults (see below). Gonadotropin deficiency causes menstrual disorders and infertility in women and decreased sexual function, infertility, and loss of secondary sexual characteristics in men. TSH and ACTH deficiencies usually develop later in the course of pituitary failure. TSH deficiency causes growth retardation in children and features of hypothyroidism in children and adults. The secondary form of adrenal insufficiency caused by ACTH deficiency leads to hypocortisolism with relative preservation of mineralocorticoid production. PRL deficiency causes failure of lactation. When lesions involve the posterior pituitary, polyuria and polydipsia reflect loss of vasopressin secretion. In patients with long-standing pituitary damage, epidemiologic studies document an increased mortality rate, primarily from increased cardiovascular and cerebrovascular disease. Previous head or neck irradiation is also a determinant of increased mortality rates in patients with hypopituitarism, especially from cerebrovascular disease.

LABORATORY INVESTIGATION

Biochemical diagnosis of pituitary insufficiency is made by demonstrating low levels of respective pituitary trophic hormones in the setting of low levels of target hormones. For example, low free thyroxine in the setting of a low or inappropriately normal TSH level suggests secondary hypothyroidism. Similarly, a low testosterone level without elevation of gonadotropins suggests hypogonadotropic hypogonadism. Provocative tests may be required to assess pituitary reserve **(Table 372-2)**. GH responses to insulin-induced hypoglycemia, arginine, L-dopa, growth hormone–releasing hormone (GHRH), or growth hormone–releasing peptides (GHRPs) can be used to assess GH reserve. Corticotropin-releasing hormone (CRH) administration induces ACTH release, and administration of synthetic ACTH (cosyntropin) evokes adrenal cortisol release as

an indirect indicator of pituitary ACTH reserve **(Chap. 379)**. ACTH reserve is most reliably assessed by measuring ACTH and cortisol levels during insulin-induced hypoglycemia. However, this test should be performed cautiously in patients with suspected adrenal insufficiency because of enhanced susceptibility to hypoglycemia and hypotension. Administering insulin to induce hypoglycemia is contraindicated in patients with active coronary artery disease or known seizure disorders.

TABLE 372-2

Tests of Pituitary Sufficiency

HORMONE	TEST	BLOOD SAMPLES	INTERPRETATION
Growth hormone (GH)	Insulin tolerance test: Regular insulin (0.05–0.15 U/kg IV)	–30, 0, 30, 60, 120 min for glucose and GH	Glucose <40 mg/dL; GH should be >3 µg/L
	GHRH test: 1 µg/kg IV	0, 15, 30, 45, 60, 120 min for GH	Normal response is GH >3 µg/L
	L-Arginine test: 30 g IV over 30 min	0, 30, 60, 120 min for GH	Normal response is GH >3 µg/L
	L-Dopa test: 500 mg PO	0, 30, 60, 120 min for GH	Normal response is GH >3 µg/L
Prolactin	TRH test: 200–500 µg IV	0, 20, and 60 min for TSH and PRL	Normal prolactin is >2 µg/L and increase >200% of baseline
ACTH	Insulin tolerance test: regular insulin (0.05–0.15 U/kg IV)	–30, 0, 30, 60, 90 min for glucose and cortisol	Glucose <40 mg/dL Cortisol should increase by >7 µg/dL or to >20 µg/dL
	CRH test: 1 µg/kg ovine CRH IV at 8 A.M.	0, 15, 30, 60, 90, 120 min for ACTH and cortisol	Basal ACTH increases 2- to 4-fold and peaks at 20–100 pg/mL Cortisol levels >20–25 µg/dL
	Metyrapone test: Metyrapone (30 mg/kg) at midnight	Plasma 11-deoxycortisol and cortisol at 8 A.M.; ACTH can also be measured	Plasma cortisol should be <4 g/dL to assure an adequate response Normal response is 11-deoxycortisol >7.5 µg/dL or ACTH >75 pg/mL
	Standard ACTH stimulation test: ACTH 1-24 (cosyntropin), 0.25 mg IM or IV	0, 30, 60 min for cortisol and aldosterone	Normal response is cortisol >21 g/dL and aldosterone response of >4 ng/dL above baseline

HORMONE	TEST	BLOOD SAMPLES	INTERPRETATION
	Low-dose ACTH test: ACTH 1-24 (cosyntropin), 1 μg IV	0, 30, 60 min for cortisol	Cortisol should be >21 g/dL
	3-day ACTH stimulation test consists of 0.25 mg ACTH 1-24 given IV over 8 h each day		Cortisol >21 g/dL
TSH	Basal thyroid function tests: T_4, T_3, TSH	Basal measurements	Low free thyroid hormone levels in the setting of TSH levels that are not appropriately increased indicate pituitary insufficiency
	TRH test: 200–500 μg IV	0, 20, 60 min for TSH and PRL[a]	TSH should increase by >5 mU/L unless thyroid hormone levels are increased
LH, FSH	LH, FSH, testosterone, estrogen	Basal measurements	Basal LH and FSH should be increased in postmenopausal women
			Low testosterone levels in the setting of low LH and FSH indicate pituitary insufficiency
	GnRH test: GnRH (100 μg) IV	0, 30, 60 min for LH and FSH	In most adults, LH should increase by 10 IU/L and FSH by 2 IU/L
			Normal responses are variable
Multiple hormones	Combined anterior pituitary test: GHRH (1 g/kg), CRH (1 μg/kg), GnRH (100 g), TRH (200 μg) are given IV	–30, 0, 15, 30, 60, 90, 120 min for GH, ACTH, cortisol, LH, FSH, and TSH	Combined or individual releasing hormone responses must be elevated in the context of basal target gland hormone values and may not be uniformly diagnostic (see text)

[a]Evoked PRL response indicates lactotrope integrity.

Abbreviations: T_3, triiodothyronine; T_4, thyroxine; TRH, thyrotropin-releasing hormone. For other abbreviations, see text.

TREATMENT

TREATMENT

Hypopituitarism

Hormone replacement therapy, including glucocorticoids, thyroid hormone, sex steroids, GH, and vasopressin, is usually safe and free of complications. Treatment regimens that mimic physiologic hormone production allow for maintenance of satisfactory clinical homeostasis. Effective dosage schedules are outlined in Table 372-3. Patients in need of glucocorticoid replacement require careful dose adjustments during stressful events such as acute illness, dental procedures, trauma, and acute hospitalization.

TABLE 372-3

Hormone Replacement Therapy for Adult Hypopituitarism[a]

TROPHIC HORMONE DEFICIT	HORMONE REPLACEMENT
ACTH	Hydrocortisone (10–20 mg/d in divided doses)
	Cortisone acetate (15–25 mg/d in divided doses)
	Prednisone (5 mg A.M.)
TSH	L-Thyroxine (0.075–0.15 mg daily)
FSH/LH	Males
	Testosterone gel (5–10 g/d)
	Testosterone skin patch (5 mg/d) Testosterone enanthate (200 mg IM every 2 weeks)
	Females
	Conjugated estrogen (0.65–1.25 mg qd for 25 days)
	Progesterone (5–10 mg qd) on days 16–25
	Estradiol skin patch (0.025–0.1 mg every week), adding progesterone on days 16–25 if uterus intact
	For fertility: menopausal gonadotropins, human chorionic gonadotropins
GH	Adults: Somatotropin (0.1–1.25 mg SC qd)
	Children: Somatotropin (0.02–0.05 mg/kg per day)
Vasopressin	Intranasal desmopressin (5–20 g twice daily) Oral 300–600 µg qd

[a]All doses shown should be individualized for specific patients and should be reassessed during stress, surgery, or pregnancy. Male and female fertility requirements should be managed as discussed in **Chaps. 384 and 385**.

Note: For abbreviations, see text.

DISORDERS OF GROWTH AND DEVELOPMENT

Skeletal Maturation and Somatic Growth

The growth plate is dependent on a variety of hormonal stimuli, including GH, insulin-like growth factor (IGF)-I, sex steroids, thyroid hormones, paracrine growth factors, and cytokines. The growth-promoting process also requires caloric energy, amino acids, vitamins, and trace metals and consumes about 10% of normal energy production. Malnutrition impairs chondrocyte activity, increases GH resistance, and reduces circulating IGF-I and IGF binding protein (IGBP)-3 levels.

Linear bone growth rates are very high in infancy and are pituitary-dependent. Mean growth velocity is ~6 cm/year in later childhood and usually is maintained within a given range on a standardized percentile chart. Peak growth rates occur during midpuberty when bone age is 12 (girls) or 13 (boys). Secondary sexual development is associated with elevated sex steroids that cause progressive epiphyseal growth plate closure. *Bone age* is delayed in patients with all forms of true GH deficiency or GH receptor defects that result in attenuated GH action.

Short stature may occur as a result of constitutive intrinsic growth defects or because of acquired extrinsic factors that impair growth. In general, delayed bone age in a child with short stature is suggestive of a hormonal or systemic disorder, whereas normal bone age in a short child is more likely to be caused by a genetic cartilage dysplasia or growth plate disorder **(Chap. 406)**.

GH Deficiency in Children

GH DEFICIENCY

Isolated GH deficiency is characterized by short stature, micropenis, increased fat, high-pitched voice, and a propensity to hypoglycemia due to relatively unopposed insulin action. Familial modes of inheritance are seen in at least one-third of these individuals and may be autosomal dominant, recessive, or X-linked. About 10% of children with GH deficiency have mutations in the *GH-N* gene, including gene deletions and a wide range of point mutations. Mutations in transcription factors Pit-1 and Prop-1, which control somatotrope development, result in GH deficiency in combination with other pituitary hormone deficiencies, which may become manifest only in adulthood. The diagnosis of *idiopathic GH deficiency* (IGHD) should be made only after known molecular defects have been rigorously excluded.

GHRH RECEPTOR MUTATIONS

Recessive mutations of the GHRH receptor gene in subjects with severe proportionate dwarfism are associated with low basal GH levels that cannot be stimulated by exogenous GHRH, GHRP, or insulin-induced

hypoglycemia, as well as anterior pituitary hypoplasia. The syndrome exemplifies the importance of the GHRH receptor for somatotrope cell proliferation and hormonal responsiveness.

GH INSENSITIVITY

This is caused by defects of GH receptor structure or signaling. Homozygous or heterozygous mutations of the GH receptor are associated with partial or complete GH insensitivity and growth failure (*Laron syndrome*). The diagnosis is based on normal or high GH levels, with decreased circulating GH-binding protein (GHBP), and low IGF-I levels. Very rarely, defective IGF-I, IGF-I receptor, or IGF-I signaling defects are also encountered. *STAT5B* mutations result in both immunodeficiency as well as abrogated GH signaling, leading to short stature with normal or elevated GH levels and low IGF-I levels. Circulating GH receptor antibodies may rarely cause peripheral GH insensitivity.

NUTRITIONAL SHORT STATURE

Caloric deprivation and malnutrition, uncontrolled diabetes, and chronic renal failure represent secondary causes of abrogated GH receptor function. These conditions also stimulate production of proinflammatory cytokines, which act to exacerbate the block of GH-mediated signal transduction. Children with these conditions typically exhibit features of acquired short stature with normal or elevated GH and low IGF-I levels.

PSYCHOSOCIAL SHORT STATURE

Emotional and social deprivation lead to growth retardation accompanied by delayed speech, discordant hyperphagia, and an attenuated response to administered GH. A nurturing environment restores growth rates.

PRESENTATION AND DIAGNOSIS

Short stature is commonly encountered in clinical practice, and the decision to evaluate these children requires clinical judgment in association with auxologic data and family history. Short stature should be evaluated comprehensively if a patient's height is >3 standard deviations (SD) below the mean for age or if the growth rate has decelerated. Skeletal maturation is best evaluated by measuring a radiologic bone age, which is based mainly on the degree of wrist bone growth plate fusion. Final height can be predicted using standardized scales (Bayley-Pinneau or Tanner-Whitehouse) or estimated by adding 6.5 cm (boys) or subtracting 6.5 cm (girls) from the midparental height.

LABORATORY INVESTIGATION

Because GH secretion is pulsatile, GH deficiency is best assessed by examining the response to provocative stimuli, including exercise, insulin-induced hypoglycemia, and other pharmacologic tests that normally increase GH to >7 μg/L in children. Random GH measurements do not distinguish normal children from those with true GH deficiency. Adequate adrenal and thyroid hormone replacement should be assured before testing. Age- and sex-matched IGF-I levels are not sufficiently sensitive or specific to make the diagnosis but can be useful to confirm GH deficiency. Pituitary MRI may reveal pituitary mass lesions or structural defects.

Molecular analyses for known mutations should be undertaken when the cause of short stature remains cryptic, or when additional clinical features suggest a genetic cause.

TREATMENT

TREATMENT

Disorders of Growth and Development

Replacement therapy with recombinant GH (0.02–0.05 mg/kg per day SC) restores growth velocity in GH-deficient children to ~10 cm/year. If pituitary insufficiency is documented, other associated hormone deficits should be corrected, especially adrenal steroids. GH treatment is also moderately effective for accelerating growth rates in children with Turner syndrome and chronic renal failure.

In patients with GH insensitivity and growth retardation due to mutations of the GH receptor, treatment with IGF-I bypasses the dysfunctional GH receptor.

ADULT GH DEFICIENCY (AGHD)

This disorder usually is caused by acquired hypothalamic or pituitary somatotrope damage. Acquired pituitary hormone deficiency follows a typical pattern in which loss of adequate GH reserve foreshadows subsequent hormone deficits. The sequential order of hormone loss is usually GH → FSH/LH → TSH → ACTH. Patients previously diagnosed with childhood-onset GH deficiency should be retested as adults to affirm the diagnosis.

PRESENTATION AND DIAGNOSIS

The clinical features of AGHD include changes in body composition, lipid metabolism, and quality of life and cardiovascular dysfunction (Table 372-4). Body composition changes are common and include reduced lean body mass, increased fat mass with selective deposition of intraabdominal visceral fat, and increased waist-to-hip ratio. Hyperlipidemia, left ventricular dysfunction, hypertension, and increased plasma fibrinogen levels also may be present. Bone mineral content is reduced, with resultant increased fracture rates. Patients may experience social isolation, depression, and difficulty maintaining gainful employment. Adult hypopituitarism is associated with a threefold increase in cardiovascular mortality rates in comparison to age- and sex-matched controls, and this may be due to GH deficiency, as patients in these studies were replaced with other deficient pituitary hormones.

TABLE 372-4

Features of Adult Growth Hormone Deficiency

Clinical
Impaired quality of life Decreased energy and drive Poor concentration Low self-esteem Social isolation Body composition changes Increased body fat mass Central fat deposition Increased waist-to-hip ratio Decreased lean body mass Reduced exercise capacity Reduced maximum O_2 uptake Impaired cardiac function Reduced muscle mass Cardiovascular risk factors Impaired cardiac structure and function Abnormal lipid profile Decreased fibrinolytic activity Atherosclerosis Omental obesity
Imaging
Pituitary: mass or structural damage Bone: reduced bone mineral density Abdomen: excess omental adiposity
Laboratory
Evoked GH <3 ng/mL IGF-I and IGFBP3 low or normal Increased LDL cholesterol Concomitant gonadotropin, TSH, and/or ACTH reserve deficits may be present

Abbreviation: LDL, low-density lipoprotein. For other abbreviations, see text.

LABORATORY INVESTIGATION

AGHD is rare, and in light of the nonspecific nature of associated clinical symptoms, patients appropriate for testing should be selected carefully on the basis of well-defined criteria. With few exceptions, testing should be restricted to patients with the following predisposing factors: (1) pituitary surgery, (2) pituitary or hypothalamic tumor or granulomas, (3) history of cranial irradiation, (4) radiologic evidence of a pituitary lesion, and (5) childhood requirement for GH replacement therapy. The transition of a GH-deficient adolescent to adulthood requires retesting to document subsequent AGHD. Up to 20% of patients previously treated for childhood-onset GH deficiency are found to be GH-sufficient on repeat testing as adults.

A significant proportion (~25%) of truly GH-deficient adults have low-normal IGF-I levels. Thus, as in the evaluation of GH deficiency in children, valid age- and sex-matched IGF-I measurements provide a useful index of therapeutic responses but are not sufficiently sensitive for diagnostic purposes. The most validated test to distinguish pituitary-sufficient patients from those with AGHD is insulin-induced (0.05–0.1 U/kg) hypoglycemia. After glucose reduction to ~40 mg/dL, most individuals experience neuroglycopenic symptoms **(Chap. 399)**, and peak GH release occurs at 60 min and remains elevated for up to 2 h. About 90% of healthy adults exhibit GH responses >5 μg/L; AGHD is defined by a peak GH response to hypoglycemia of <3 μg/L. Although insulin-induced hypoglycemia is safe when performed under appropriate supervision, it is contraindicated in patients with diabetes, ischemic heart disease, cerebrovascular disease, or epilepsy and in elderly patients. Alternative stimulatory tests include intravenous arginine (30 g), GHRH (1 μg/kg), GHRP-6 (90 μg), and glucagon (1 mg). Combinations of these tests may evoke GH secretion in subjects who are not responsive to a single test.

TREATMENT

TREATMENT

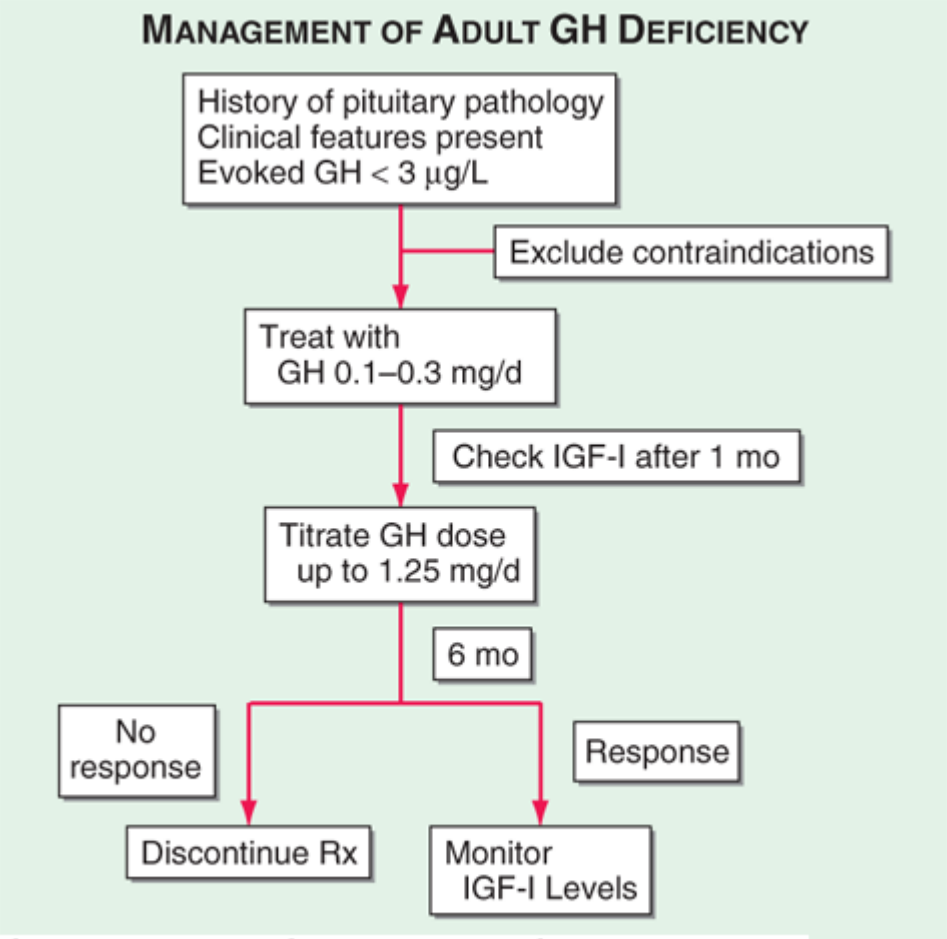
Adult GH Deficiency

Once the diagnosis of AGHD is unequivocally established, replacement of GH may be indicated. Contraindications to therapy include the presence of an active neoplasm, intracranial hypertension, and uncontrolled diabetes and retinopathy. The starting adult dose of 0.1–0.2 mg/d should be titrated (up to a maximum of 1.25 mg/d) to maintain IGF-I levels in the mid-normal range for age- and sex-matched controls **(Fig. 372-1)**. Women require higher doses than men, and elderly patients require less GH. Long-term GH maintenance sustains normal IGF-I levels and is associated with persistent body composition changes (e.g., enhanced lean body mass and lower body fat). High-density lipoprotein cholesterol increases, but total cholesterol and insulin levels may not change significantly. Lumbar spine bone mineral density increases, but this response is gradual (>1 year). Many patients note significant improvement in quality of life when evaluated by standardized questionnaires. The effect of GH replacement on mortality rates in GH-deficient patients is currently the subject of long-term prospective investigation.

About 30% of patients exhibit reversible dose-related fluid retention, joint pain, and carpal tunnel syndrome, and up to 40% exhibit myalgias and paresthesia. Patients receiving insulin require careful monitoring for dosing adjustments, as GH is a potent counterregulatory hormone for insulin action. Patients with type 2 diabetes mellitus may initially develop further insulin resistance. However, glycemic control usually improves with the sustained loss of abdominal fat associated with long-term GH replacement. Headache, increased intracranial pressure, hypertension, and tinnitus occur rarely. Pituitary tumor regrowth and progression of skin lesions or other tumors have not been encountered in long-term surveillance programs with appropriate replacement doses.

FIGURE 372-1

Management of adult growth hormone (GH) deficiency. IGF, insulin-like growth factor; Rx, Treatment.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ACTH DEFICIENCY

PRESENTATION AND DIAGNOSIS

Secondary adrenal insufficiency occurs as a result of pituitary ACTH deficiency. It is characterized by fatigue, weakness, anorexia, nausea, vomiting, and, occasionally, hypoglycemia. In contrast to primary adrenal

failure, hypocortisolism associated with pituitary failure usually is not accompanied by hyperpigmentation or mineralocorticoid deficiency.

ACTH deficiency is commonly due to glucocorticoid withdrawal after treatment-associated suppression of the hypothalamic-pituitary-adrenal (HPA) axis. Isolated ACTH deficiency may occur after surgical resection of an ACTH-secreting pituitary adenoma that has suppressed the HPA axis; this phenomenon is in fact suggestive of a surgical cure. The mass effects of other pituitary adenomas or sellar lesions may lead to ACTH deficiency, usually in combination with other pituitary hormone deficiencies. Partial ACTH deficiency may be unmasked in the presence of an acute medical or surgical illness, when clinically significant hypocortisolism reflects diminished ACTH reserve. Rarely, *TPIT* or *POMC* mutations result in primary ACTH deficiency.

LABORATORY DIAGNOSIS

Inappropriately low ACTH levels in the setting of low cortisol levels are characteristic of diminished ACTH reserve. Low basal serum cortisol levels are associated with blunted cortisol responses to ACTH stimulation and impaired cortisol response to insulin-induced hypoglycemia, or testing with metyrapone or CRH. **For a description of provocative ACTH tests, see Chap. 379.**

TREATMENT

TREATMENT

ACTH Deficiency

Glucocorticoid replacement therapy improves most features of ACTH deficiency. The total daily dose of hydrocortisone replacement preferably should generally not exceed 20 mg daily, divided into two or three doses. Prednisone (5 mg each morning) is longer acting and has fewer mineralocorticoid effects than hydrocortisone. Some authorities advocate lower maintenance doses in an effort to avoid cushingoid side effects. Doses should be increased severalfold during periods of acute illness or stress. Patients should wear medialert bracelets and/or carry identification cards with information about their glucocorticoid requirements.

GONADOTROPIN DEFICIENCY

Hypogonadism is the most common presenting feature of adult hypopituitarism even when other pituitary hormones are also deficient. It is often a harbinger of hypothalamic or pituitary lesions that impair GnRH production or delivery through the pituitary stalk. As noted below, hypogonadotropic hypogonadism is a common presenting feature of hyperprolactinemia.

A variety of inherited and acquired disorders are associated with *isolated hypogonadotropic hypogonadism* (IHH) **(Chap. 384)**. Hypothalamic defects associated with GnRH deficiency include Kallmann syndrome and mutations in more than a dozen genes that regulate GnRH neuron migration, development, and function (see above). Mutations in *GPR54, DAX1,* kisspeptin, the GnRH receptor, and the LHβ or FSHβ subunit genes also

cause pituitary gonadotropin deficiency. Acquired forms of GnRH deficiency leading to hypogonadotropism are seen in association with anorexia nervosa, stress, starvation, and extreme exercise but also may be idiopathic. Hypogonadotropic hypogonadism in these disorders is reversed by removal of the stressful stimulus or by caloric replenishment.

PRESENTATION AND DIAGNOSIS

In premenopausal women, hypogonadotropic hypogonadism presents as diminished ovarian function leading to oligomenorrhea or amenorrhea, infertility, decreased vaginal secretions, decreased libido, and breast atrophy. In hypogonadal adult men, secondary testicular failure is associated with decreased libido and potency, infertility, decreased muscle mass with weakness, reduced beard and body hair growth, soft testes, and characteristic fine facial wrinkles. Osteoporosis occurs in both untreated hypogonadal women and men.

LABORATORY INVESTIGATION

Central hypogonadism is associated with low or inappropriately normal serum gonadotropin levels in the setting of low sex hormone concentrations (testosterone in men, estradiol in women). Because gonadotropin secretion is pulsatile, valid assessments may require repeated measurements or the use of pooled serum samples. Men have reduced sperm counts.

Intravenous GnRH (100 μg) stimulates gonadotropes to secrete LH (which peaks within 30 min) and FSH (which plateaus during the ensuing 60 min). Normal responses vary according to menstrual cycle stage, age, and sex of the patient. Generally, LH levels increase about threefold, whereas FSH responses are less pronounced. In the setting of gonadotropin deficiency, a normal gonadotropin response to GnRH indicates intact pituitary gonadotrope function and suggests a hypothalamic abnormality. An absent response, however, does not reliably distinguish pituitary from hypothalamic causes of hypogonadism. For this reason, GnRH testing usually adds little to the information gained from baseline evaluation of the hypothalamic-pituitary-gonadotrope axis except in cases of isolated GnRH deficiency (e.g., Kallmann syndrome).

MRI examination of the sellar region and assessment of other pituitary functions usually are indicated in patients with documented central hypogonadism.

TREATMENT

TREATMENT

Gonadotropin Deficiency

In males, testosterone replacement is necessary to achieve and maintain normal growth and development of the external genitalia, secondary sex characteristics, male sexual behavior, and androgenic anabolic effects, including maintenance of muscle function and bone mass. Testosterone may be administered by intramuscular injections every 1–4 weeks or by using skin patches or testosterone gels **(Chap. 384)**.

Gonadotropin injections (hCG or human menopausal gonadotropin [hMG]) over 12–18 months are used to restore fertility. Pulsatile GnRH therapy (25–150 ng/kg every 2 h), administered by a subcutaneous infusion pump, is also effective for treatment of hypothalamic hypogonadism when fertility is desired.

In premenopausal women, cyclical replacement of estrogen and progesterone maintains secondary sexual characteristics and integrity of genitourinary tract mucosa and prevents premature osteoporosis **(Chap. 385)**. Gonadotropin therapy is used for ovulation induction. Follicular growth and maturation are initiated using hMG or recombinant FSH; hCG or human luteinizing hormone (hLH) is subsequently injected to induce ovulation. As in men, pulsatile GnRH therapy can be used to treat hypothalamic causes of gonadotropin deficiency.

DIABETES INSIPIDUS

See Chap. 374 for diagnosis and treatment of diabetes insipidus.

FURTHER READING

Alatzoglou KS, Dattani MT: Genetic causes and treatment of isolated growth hormone deficiency—An update. Nat Rev Endocrinol 6:562, 2010. [PubMed: 20852587]

Clemmons DR: The diagnosis and treatment of growth hormone deficiency in adults. Curr Opin Endocrinol Diabetes Obes 17:377, 2010. [PubMed: 20588115]

Dekkers OM et al: Multisystem morbidity and mortality in Cushing's syndrome: A cohort study. J Clin Endocrinol Metab 98:2277, 2013. [PubMed: 23533241]

Fleseriu M et al: Hormonal replacement in hypopituitarism in adults: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 101:3888, 2016. [PubMed: 27736313]

Honegger J et al: Diagnosis of primary hypophysitis in Germany. J Clin Endocrinol Metab 100:3841, 2015. [PubMed: 26262437]

Romijn JA: The chronic syndromes after previous treatment of pituitary tumours. Nat Rev Endocrinol 12:547, 2016. [PubMed: 27259177]

Sherlock M et al: Mortality in patients with pituitary disease. Endocr Rev 31:301, 2010. [PubMed: 20086217]

Tanriverdi F et al: Pituitary dysfunction after traumatic brain injury: A clinical and pathophysiological approach. Endocr Rev 36:305, 2015. [PubMed: 25950715]

Chapter 373: Pituitary Tumor Syndromes

Shlomo Melmed; J. Larry Jameson

HYPOTHALAMIC, PITUITARY, AND OTHER SELLAR MASSES

EVALUATION OF SELLAR MASSES

Local Mass Effects

Clinical manifestations of sellar lesions vary, depending on the anatomic location of the mass and the direction of its extension (Table 373-1). The dorsal sellar diaphragm presents the least resistance to soft tissue expansion from the sella; consequently, pituitary adenomas frequently extend in a suprasellar direction. Bony invasion may occur as well.

TABLE 373-1

Features of Sellar Mass Lesions[a]

IMPACTED STRUCTURE	CLINICAL IMPACT
Pituitary	Hypogonadism Hypothyroidism Growth failure and adult hyposomatotropism Hypoadrenalism
Optic chiasm	Loss of red perception Bitemporal hemianopia Superior or bitemporal field defect Scotoma Blindness
Hypothalamus	Temperature dysregulation Appetite and thirst disorders Obesity Diabetes insipidus Sleep disorders Behavioral dysfunction Autonomic dysfunction
Cavernous sinus	Ophthalmoplegia with or without ptosis or diplopia Facial numbness
Frontal lobe	Personality disorder Anosmia
Brain	Headache Hydrocephalus Psychosis Dementia Laughing seizures

[a]As the intrasellar mass expands, it first compresses intrasellar pituitary tissue, then usually invades dorsally through the dura to lift the optic chiasm or laterally to the cavernous sinuses. Bony erosion is rare, as is direct brain compression. Microadenomas may present with headache.

Headaches are common features of small intrasellar tumors, even with no demonstrable suprasellar extension. Because of the confined nature of the pituitary, small changes in intrasellar pressure stretch the dural plate; however, headache severity correlates poorly with adenoma size or extension.

Suprasellar extension can lead to visual loss by several mechanisms, the most common being compression of the optic chiasm, but rarely, direct invasion of the optic nerves or obstruction of cerebrospinal fluid (CSF) flow leading to secondary visual disturbances can occur. Pituitary stalk compression by a hormonally active or inactive intrasellar mass may compress the portal vessels, disrupting pituitary access to hypothalamic hormones and dopamine; this results in early hyperprolactinemia and later concurrent loss of other pituitary hormones. This "stalk section" phenomenon may also be caused by trauma, whiplash injury with posterior clinoid stalk compression, or skull base fractures. Lateral mass invasion may impinge on the cavernous sinus and compress its neural contents, leading to cranial nerve III, IV, and VI palsies as well as effects on the ophthalmic and maxillary branches of the fifth cranial nerve **(Chap. 433)**. Patients may present with diplopia, ptosis, ophthalmoplegia, and decreased facial sensation, depending on the extent of neural damage. Extension into the sphenoid sinus indicates that the pituitary mass has eroded through the sellar floor. Aggressive tumors rarely invade the palate roof and cause nasopharyngeal obstruction, infection, and CSF leakage. Temporal and frontal lobe involvement may rarely lead to uncinate seizures, personality disorders, and anosmia. Direct hypothalamic encroachment by an invasive pituitary mass may cause important metabolic sequelae, including precocious puberty or hypogonadism, diabetes insipidus, sleep disturbances, dysthermia, and appetite disorders.

Magnetic Resonance Imaging

Sagittal and coronal T1-weighted magnetic resonance imaging (MRI) before and after administration of gadolinium allows precise visualization of the pituitary gland with clear delineation of the hypothalamus, pituitary stalk, pituitary tissue and surrounding suprasellar cisterns, cavernous sinuses, sphenoid sinus, and optic chiasm. Pituitary gland height ranges from 6 mm in children to 8 mm in adults; during pregnancy and puberty, the height may reach 10–12 mm. The upper aspect of the adult pituitary is flat or slightly concave, but in adolescent and pregnant individuals, this surface may be convex, reflecting physiologic pituitary enlargement. The stalk should be midline and vertical. Computed tomography (CT) scan is reserved to define the extent of bony erosion or the presence of calcification.

Anterior pituitary gland soft tissue consistency is slightly heterogeneous on MRI, and signal intensity resembles that of brain matter on T1-weighted imaging **(Fig. 373-1)**. Adenoma density is usually lower than that of surrounding normal tissue on T1-weighted imaging, and the signal intensity increases with T2-weighted images. The high phospholipid content of the posterior pituitary results in a "pituitary bright spot."

FIGURE 373-1

Pituitary adenoma. Coronal T1-weighted postcontrast magnetic resonance image shows a homogeneously enhancing mass (*arrowheads*) in the sella turcica and suprasellar region compatible with a pituitary adenoma; the *small arrows* outline the carotid arteries.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Sellar masses are encountered commonly as incidental findings on MRI, and most of them are pituitary adenomas (incidentalomas). In the absence of hormone hypersecretion, these small intrasellar lesions can be monitored safely with MRI, which is performed annually and then less often if there is no evidence of further growth. Resection should be considered for incidentally discovered larger macroadenomas, because about one-third become invasive or cause local pressure effects. If hormone hypersecretion is evident, specific therapies are indicated as described below. When larger masses (>1 cm) are encountered, they should also be distinguished from nonadenomatous lesions. Meningiomas often are associated with bony hyperostosis; craniopharyngiomas may be calcified and are usually hypodense, whereas gliomas are hyperdense on T2-weighted images.

Ophthalmologic Evaluation

Because optic tracts may be contiguous to an expanding pituitary mass, reproducible visual field assessment using perimetry techniques should be performed on all patients with sellar mass lesions that impinge the optic chiasm (Chap. 28). Bitemporal hemianopia, often more pronounced superiorly, is observed classically. It occurs because nasal ganglion cell fibers, which cross in the optic chiasm, are especially vulnerable to compression of the ventral optic chiasm. Occasionally, homonymous hemianopia occurs from postchiasmal compression or monocular temporal field loss from prechiasmal compression. Invasion of the cavernous sinus can produce diplopia from ocular motor nerve palsy. Early diagnosis reduces the risk of optic atrophy, vision loss, or eye misalignment.

Laboratory Investigation

The presenting clinical features of functional pituitary adenomas (e.g., acromegaly, prolactinomas, or Cushing syndrome) should guide the laboratory studies **(Table 373-2)**. However, for a sellar mass with no obvious clinical features of hormone excess, laboratory studies are geared toward determining the nature of the tumor and assessing the possible presence of hypopituitarism. When a pituitary adenoma is suspected based on MRI, initial hormonal evaluation usually includes (1) basal prolactin (PRL); (2) insulin-like growth factor (IGF)-I; (3) 24-h urinary free cortisol (UFC) and/or overnight oral dexamethasone (1 mg) suppression test; (4) α subunit, follicle-stimulating hormone (FSH), and luteinizing hormone (LH); and (5) thyroid function tests. Additional hormonal evaluation may be indicated based on the results of these tests. Pending more detailed assessment of hypopituitarism, a menstrual history, measurement of testosterone and 8 A.M. cortisol levels, and thyroid function tests usually identify patients with pituitary hormone deficiencies that require hormone replacement before further testing or surgery.

TABLE 373-2

Screening Tests for Functional Pituitary Adenomas

	TEST	COMMENTS
Acromegaly	Serum IGF-I Oral glucose tolerance test with GH obtained at 0, 30, and 60 min	Interpret IGF-I relative to age- and sex-matched controls Normal subjects should suppress growth hormone to <1 μ/L
Prolactinoma	Serum PRL	Exclude medications MRI of the sella should be ordered if PRL is elevated
Cushing's disease	24-h urinary free cortisol Dexamethasone (1 mg) at 11 P.M. and fasting plasma cortisol measured at 8 A.M. ACTH assay	Ensure urine collection is total and accurate Normal subjects suppress to <5 μ/dL Distinguishes adrenal adenoma (ACTH suppressed) from ectopic ACTH or Cushing's disease (ACTH normal or elevated)

Abbreviations: ACTH, adrenocorticotropin hormone; GH, growth hormone; IGF-I, insulin-like growth factor I; MRI, magnetic resonance imaging; PRL, prolactin.

Histologic Evaluation

Immunohistochemical staining of pituitary tumor specimens obtained at transsphenoidal surgery confirms clinical and laboratory studies and provides a histologic diagnosis when hormone studies are equivocal and in cases of clinically nonfunctioning tumors.

TREATMENT

TREATMENT

Hypothalamic, Pituitary, and Other Sellar Masses

OVERVIEW

Successful management of sellar masses requires accurate diagnosis as well as selection of optimal therapeutic modalities. Most pituitary tumors are benign and slow-growing. Clinical features result from local mass effects and hormonal hyper- or hyposecretion syndromes caused directly by the adenoma or occurring as a consequence of treatment. Thus, lifelong management and follow-up are necessary for these patients.

MRI with gadolinium enhancement for pituitary visualization, new advances in transsphenoidal surgery and in stereotactic radiotherapy (including gamma-knife radiotherapy), and novel therapeutic agents have improved pituitary tumor management. The goals of pituitary tumor treatment include normalization of excess pituitary secretion, amelioration of symptoms and signs of hormonal hypersecretion syndromes, and shrinkage or ablation of large tumor masses with relief of adjacent structure compression. Residual anterior pituitary function should be preserved during treatment and sometimes can be restored by removing the tumor mass. Ideally, adenoma recurrence should be prevented.

TRANSSPHENOIDAL SURGERY

Transsphenoidal rather than transfrontal resection is the desired surgical approach for pituitary tumors, except for the rare invasive suprasellar mass surrounding the frontal or middle fossa or the optic nerves or invading posteriorly behind the clivus. Intraoperative microscopy facilitates visual distinction between adenomatous and normal pituitary tissue as well as microdissection of small tumors that may not be visible by MRI **(Fig. 373-2)**. Transsphenoidal surgery also avoids the cranial invasion and manipulation of brain tissue required by subfrontal surgical approaches. Endoscopic techniques with three-dimensional intraoperative localization have improved visualization and access to tumor tissue. Individual surgical experience is a major determinant of outcome efficacy with these techniques.

In addition to correction of hormonal hypersecretion, pituitary surgery is indicated for mass lesions that impinge on surrounding structures. Surgical decompression and resection are required for an expanding pituitary mass, which may be asymptomatic or accompanied by persistent headache, progressive visual field defects, cranial nerve palsies, hydrocephalus, and, occasionally, intrapituitary hemorrhage and apoplexy. Transsphenoidal surgery sometimes is used for pituitary tissue biopsy to establish a histologic diagnosis. Whenever possible, the pituitary mass lesion should be selectively excised; normal pituitary tissue should be manipulated or resected only when critical for effective mass dissection. Nonselective hemihypophysectomy or total hypophysectomy may be indicated if no hypersecreting mass lesion is clearly discernible, multifocal lesions are present, or the remaining nontumorous pituitary tissue is obviously necrotic. This strategy, however, increases the likelihood of postoperative hypopituitarism and the need for lifelong hormone replacement.

Preoperative mass effects, including visual field defects and compromised pituitary function, may be reversed by surgery, particularly when the deficits are not long-standing. For large and invasive tumors, it is necessary to determine the optimal balance between maximal tumor resection and preservation of anterior pituitary function, especially for preserving growth and reproductive function in younger patients. Similarly, tumor invasion outside the sella is rarely amenable to surgical cure; the surgeon must judge the risk-versus-benefit ratio of extensive tumor resection.

Side Effects

Tumor size, the degree of invasiveness, and experience of the surgeon largely determine the incidence of surgical complications. Operative mortality rate is ~1%. Transient diabetes insipidus and hypopituitarism occur in up to 20% of patients. Permanent diabetes insipidus, cranial nerve damage, nasal septal perforation, or visual disturbances may be encountered in up to 10% of patients. CSF leaks occur in 4% of patients. Less common complications include carotid artery injury, loss of vision, hypothalamic damage, and meningitis. Permanent side effects are rare after surgery for microadenomas.

RADIATION

Radiation is used either as a primary therapy for pituitary or parasellar masses or, more commonly, as an adjunct to surgery or medical therapy. Focused megavoltage irradiation is achieved by precise MRI localization, using a high-voltage linear accelerator and accurate isocentric rotational arcing. A major determinant of accurate irradiation is reproduction of the patient's head position during multiple visits and maintenance of absolute head immobility. A total of <50 Gy (5000 rad) is given as 180-cGy (180-rad) fractions divided over ~6 weeks. Stereotactic radiosurgery delivers a large single high-energy dose from a cobalt-60 source (gamma knife), linear accelerator, or cyclotron. Long-term effects of gamma-knife surgery are unclear but appear to be similar to those encountered with conventional radiation. Proton beam therapy is available in some centers and provides concentrated radiation doses within a localized region.

The role of radiation therapy in pituitary tumor management depends on multiple factors, including the nature of the tumor, the age of the patient, and the availability of surgical and radiation expertise. Because of its relatively slow onset of action, radiation therapy is usually reserved for postsurgical management. As an adjuvant to surgery, radiation is used to treat residual tumor and in an attempt to prevent regrowth. Irradiation offers the only means for potentially ablating significant postoperative residual nonfunctioning tumor tissue. In contrast, PRL-, growth hormone (GH)–, and adrenocorticotropin hormone (ACTH)–secreting residual tumor tissues are amenable to medical therapy.

Side Effects

In the short term, radiation may cause transient nausea and weakness. Alopecia and loss of taste and smell may be more long-lasting. Failure of pituitary hormone synthesis is common in patients who have undergone head and neck or pituitary-directed irradiation. More than 50% of patients develop loss of GH, ACTH, thyroid-stimulating hormone (TSH), and/or gonadotropin secretion within 10 years, usually due to hypothalamic damage. Lifelong follow-up with testing of anterior pituitary hormone reserve is therefore required after radiation treatment. Optic nerve damage with impaired vision due to optic neuritis is reported in ~2% of patients who undergo pituitary irradiation. Cranial nerve damage is uncommon now that radiation doses are

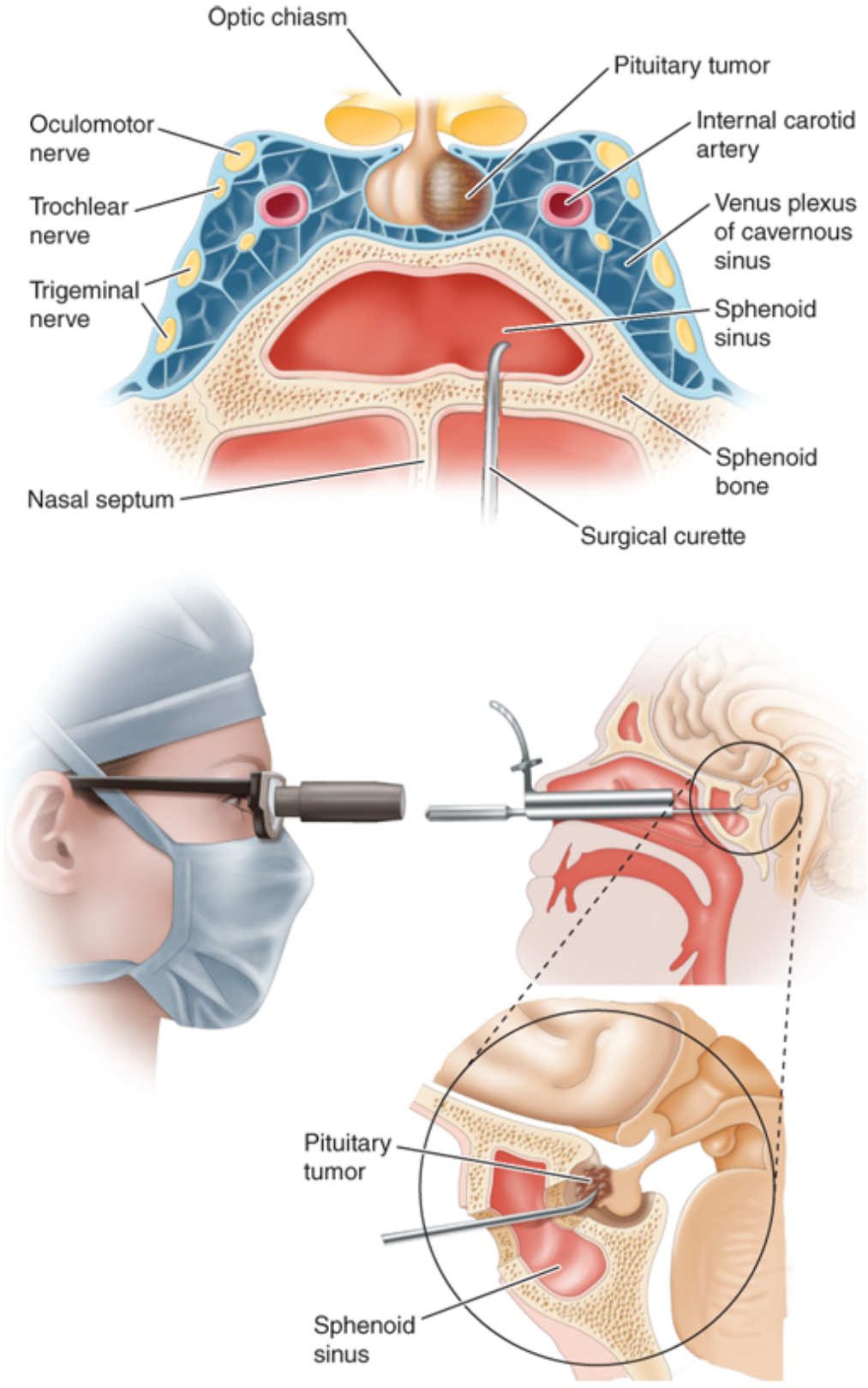
<2 Gy (200 rad) at any one treatment session and the maximum dose is <50 Gy (5000 rad). The use of stereotactic radiotherapy may reduce damage to adjacent structures. Radiotherapy for pituitary tumors has been associated with adverse mortality rates, mainly from cerebrovascular disease. The cumulative risk of developing a secondary tumor after conventional radiation is 1.3% after 10 years and 1.9% after 20 years.

MEDICAL

Medical therapy for pituitary tumors is highly specific and depends on tumor type. For prolactinomas, dopamine agonists are the treatment of choice. For acromegaly, somatostatin analogues and a GH receptor antagonist are indicated. For TSH-secreting tumors, somatostatin analogues and occasionally dopamine agonists are indicated. ACTH-secreting tumors may respond to somatostatin analogues, and adrenal-directed therapy may also be of benefit. Nonfunctioning tumors are generally not responsive to medications and require surgery and/or irradiation.

FIGURE 373-2

Transsphenoidal resection of pituitary mass via the endonasal approach. *(Adapted from R Fahlbusch: Endocrinol Metab Clin 21:669, 1992.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SELLAR MASSES

Sellar masses other than pituitary adenomas may arise from brain, hypothalamic, or pituitary tissues. Each exhibit features related to the lesion location but also unique to the specific etiology.

Hypothalamic Lesions

Lesions involving the anterior and preoptic hypothalamic regions cause paradoxical vasoconstriction, tachycardia, and hyperthermia. Acute hyperthermia usually is due to a hemorrhagic insult, but poikilothermia may also occur. Central disorders of thermoregulation result from posterior hypothalamic damage. The *periodic hypothermia syndrome* is characterized by episodic attacks of rectal temperatures <30°C (86°F), sweating, vasodilation, vomiting, and bradycardia **(Chap. 454)**. Damage to the ventromedial hypothalamic nuclei by craniopharyngiomas, hypothalamic trauma, or inflammatory disorders may be associated with *hyperphagia* and *obesity*. This region appears to contain an energy-satiety center where melanocortin receptors are influenced by leptin, insulin, pro-opiomelanocortin (POMC) products, and gastrointestinal peptides **(Chap. 394)**. Polydipsia and hypodipsia are associated with damage to central osmoreceptors located in preoptic nuclei **(Chap. 374)**. Slow-growing hypothalamic lesions can cause increased somnolence and disturbed sleep cycles as well as obesity, hypothermia, and emotional outbursts. Lesions of the central hypothalamus may stimulate sympathetic neurons, leading to elevated serum catecholamine and cortisol levels. These patients are predisposed to cardiac arrhythmias, hypertension, and gastric erosions.

Craniopharyngiomas are benign, suprasellar cystic masses that present with headaches, visual field deficits, and variable degrees of hypopituitarism. They are derived from Rathke's pouch and arise near the pituitary stalk, commonly extending into the suprasellar cistern. Craniopharyngiomas are often large, cystic, and locally invasive. Many are partially calcified, exhibiting a characteristic appearance on skull x-ray and CT images. More than half of all patients present before age 20, usually with signs of increased intracranial pressure, including headache, vomiting, papilledema, and hydrocephalus. Associated symptoms include visual field abnormalities, personality changes and cognitive deterioration, cranial nerve damage, sleep difficulties, and weight gain. Hypopituitarism can be documented in ~90%, and diabetes insipidus occurs in ~10% of patients. About half of affected children present with growth retardation. MRI is generally superior to CT for evaluating cystic structure and tissue components of craniopharyngiomas. CT is useful to define calcifications and evaluate invasion into surrounding bony structures and sinuses.

Treatment usually involves transcranial or transsphenoidal surgical resection followed by postoperative radiation of residual tumor. Surgery alone is curative in less than half of patients because of recurrences due to adherence to vital structures or because of small tumor deposits in the hypothalamus or brain parenchyma. The goal of surgery is to remove as much tumor as possible without risking complications associated with efforts to remove firmly adherent or inaccessible tissue. In the absence of radiotherapy, ~75% of craniopharyngiomas recur, and 10-year survival is <50%. In patients with incomplete resection, radiotherapy improves 10-year survival to 70–90% but is associated with increased risk of secondary malignancies. Most patients require lifelong pituitary hormone replacement.

Developmental failure of Rathke's pouch obliteration may lead to *Rathke's cysts*, which are small (<5 mm) cysts entrapped by squamous epithelium and are found in ~20% of individuals at autopsy. Although Rathke's

cleft cysts do not usually grow and are often diagnosed incidentally, about a third present in adulthood with compressive symptoms, diabetes insipidus, and hyperprolactinemia due to stalk compression. Rarely, hydrocephalus develops. The diagnosis is suggested preoperatively by visualizing the cyst wall on MRI, which distinguishes these lesions from craniopharyngiomas. Cyst contents range from CSF-like fluid to mucoid material. *Arachnoid cysts* are rare and generate an MRI image that is isointense with CSF.

Sella chordomas usually present with bony clival erosion, local invasiveness, and, on occasion, calcification. Normal pituitary tissue may be visible on MRI, distinguishing chordomas from aggressive pituitary adenomas. Mucinous material may be obtained by fine-needle aspiration.

Meningiomas arising in the sellar region may be difficult to distinguish from nonfunctioning pituitary adenomas. Meningiomas typically enhance on MRI and may show evidence of calcification or bony erosion. Meningiomas may cause compressive symptoms.

Histiocytosis X includes a variety of syndromes associated with foci of eosinophilic granulomas. Diabetes insipidus, exophthalmos, and punched-out lytic bone lesions (*Hand-Schüller-Christian disease*) are associated with granulomatous lesions visible on MRI, as well as a characteristic axillary skin rash. Rarely, the pituitary stalk may be involved.

Pituitary metastases occur in ~3% of cancer patients. Bloodborne metastatic deposits are found almost exclusively in the posterior pituitary. Accordingly, diabetes insipidus can be a presenting feature of lung, gastrointestinal, breast, and other pituitary metastases. About half of pituitary metastases originate from breast cancer; ~25% of patients with metastatic breast cancer have such deposits. Rarely, pituitary stalk involvement results in anterior pituitary insufficiency. The MRI diagnosis of a metastatic lesion may be difficult to distinguish from an aggressive pituitary adenoma; the diagnosis may require histologic examination of excised tumor tissue. Primary or metastatic lymphoma, leukemias, and plasmacytomas also occur within the sella.

Hypothalamic hamartomas and *gangliocytomas* may arise from astrocytes, oligodendrocytes, and neurons with varying degrees of differentiation. These tumors may overexpress hypothalamic neuropeptides, including gonadotropin-releasing hormone (GnRH), growth hormone–releasing hormone (GHRH), and corticotropin-releasing hormone (CRH). With GnRH-producing tumors, children present with precocious puberty, psychomotor delay, and laughing-associated seizures. Medical treatment of GnRH-producing hamartomas with long-acting GnRH analogues effectively suppresses gonadotropin secretion and controls premature pubertal development. Rarely, hamartomas also are associated with craniofacial abnormalities; imperforate anus; cardiac, renal, and lung disorders; and pituitary failure as features of *Pallister-Hall syndrome*, which is caused by mutations in the carboxy terminus of the *GLI3* gene. Hypothalamic hamartomas are often contiguous with the pituitary, and preoperative MRI diagnosis may not be possible. Histologic evidence of hypothalamic neurons in tissue resected at transsphenoidal surgery may be the first indication of a primary hypothalamic lesion.

Hypothalamic gliomas and *optic gliomas* occur mainly in childhood and usually present with visual loss. Adults have more aggressive tumors; about a third are associated with neurofibromatosis.

Brain germ cell tumors may arise within the sellar region. They include *dysgerminomas*, which frequently are associated with diabetes insipidus and visual loss. They rarely metastasize. *Germinomas, embryonal carcinomas*, *teratomas*, and *choriocarcinomas* may arise in the parasellar region and produce hCG. These germ cell tumors present with precocious puberty, diabetes insipidus, visual field defects, and thirst disorders. Many patients are GH-deficient with short stature.

PITUITARY ADENOMAS AND HYPERSECRETION SYNDROMES

Pituitary adenomas are the most common cause of pituitary hormone hypersecretion and hyposecretion syndromes in adults. They account for ~15% of all intracranial neoplasms and have been identified with a population prevalence of ~80/100,000. At autopsy, up to one-quarter of all pituitary glands harbor an unsuspected microadenoma (<10 mm diameter). Similarly, pituitary imaging detects small clinically inapparent pituitary lesions in at least 10% of individuals.

Pathogenesis

Pituitary adenomas are benign neoplasms that arise from one of the five anterior pituitary cell types. The clinical and biochemical phenotypes of pituitary adenomas depend on the cell type from which they are derived. Thus, tumors arising from lactotrope (PRL), somatotrope (GH), corticotrope (ACTH), thyrotrope (TSH), or gonadotrope (LH, FSH) cells hypersecrete their respective hormones (Table 373-3). Plurihormonal tumors express various combinations of GH, PRL, TSH, ACTH, or the glycoprotein hormone α or β subunits. They may be diagnosed by careful immunocytochemistry or may manifest as clinical syndromes that combine features of these hormonal hypersecretory syndromes. Morphologically, these tumors may arise from a single polysecreting cell type or include cells with mixed function within the same tumor.

TABLE 373-3

Classification of Pituitary Adenomas[a]

ADENOMA CELL ORIGIN	HORMONE PRODUCT	CLINICAL SYNDROME
Lactotrope	PRL	Hypogonadism, galactorrhea
Gonadotrope	FSH, LH, subunits	Silent or hypogonadism
Somatotrope	GH	Acromegaly/gigantism
Corticotrope	ACTH/none	Cushing's disease or silent
Mixed growth hormone and prolactin cell	GH, PRL	Acromegaly, hypogonadism, galactorrhea
Other plurihormonal cell	Any	Mixed
Acidophil stem cell	PRL, GH	Hypogonadism, galactorrhea, acromegaly
Mammosomatotrope	PRL, GH	Hypogonadism, galactorrhea, acromegaly
Thyrotrope	TSH	Thyrotoxicosis
Null cell	None	Pituitary failure/none
Oncocytoma	None	Pituitary failure/none

[a]Hormone-secreting tumors are listed in decreasing order of frequency. All tumors may cause local pressure effects, including visual disturbances, cranial nerve palsy, and headache.

Note: For abbreviations, see text.

Source: Adapted from S Melmed: Nat Rev Endocrinol 7:257, 2011.

Hormonally active tumors are characterized by autonomous hormone secretion with diminished feedback responsiveness to physiologic inhibitory pathways. Hormone production does not always correlate with tumor size. Small hormone-secreting adenomas may cause significant clinical perturbations, whereas larger adenomas that produce less hormone may be clinically silent and remain undiagnosed (if no central compressive effects occur). About one-third of all adenomas are clinically nonfunctioning and produce no distinct clinical hypersecretory syndrome. Most of them arise from gonadotrope cells and may secrete small

amounts of α- and β-glycoprotein hormone subunits or, very rarely, intact circulating gonadotropins. True pituitary carcinomas with documented extracranial metastases are exceedingly rare.

Almost all pituitary adenomas are monoclonal in origin, implying the acquisition of one or more somatic mutations that confer a selective growth advantage. Consistent with their clonal origin, complete surgical resection of small pituitary adenomas usually cures hormone hypersecretion. Nevertheless, hypothalamic hormones such as GHRH and CRH also enhance mitotic activity of their respective pituitary target cells in addition to their role in pituitary hormone regulation. Thus, patients who harbor rare abdominal or chest tumors that elaborate ectopic GHRH or CRH may present with somatotrope or corticotrope hyperplasia with GH or ACTH hypersecretion.

Several etiologic genetic events have been implicated in the development of pituitary tumors. The pathogenesis of sporadic forms of acromegaly has been particularly informative as a model of tumorigenesis. GHRH, after binding to its G protein–coupled somatotrope receptor, uses cyclic adenosine monophosphate (AMP) as a second messenger to stimulate GH secretion and somatotrope proliferation. A subset (~35%) of GH-secreting pituitary tumors contains sporadic mutations in Gsα (Arg 201 → Cys or His; Gln 227 → Arg). These mutations attenuate intrinsic GTPase activity, resulting in constitutive elevation of cyclic AMP, Pit-1 induction, and activation of cyclic AMP response element binding protein (CREB), thereby promoting somatotrope cell proliferation and GH secretion.

Characteristic loss of heterozygosity (LOH) in various chromosomes has been documented in large or invasive macroadenomas, suggesting the presence of putative tumor suppressor genes at these loci in up to 20% of sporadic pituitary tumors, including GH-, PRL-, and ACTH-producing adenomas and some nonfunctioning tumors. Lineage-specific cell cycle disruptions with elevated levels of CDK inhibitors are present in most pituitary adenomas.

Compelling evidence also favors growth factor promotion of pituitary tumor proliferation. Basic fibroblast growth factor (bFGF) is abundant in the pituitary and stimulates pituitary cell mitogenesis, whereas epithelial growth factor receptor (EGFR) signaling induces both hormone synthesis and cell proliferation. Mutations of *USP8* may result in overexpressed EGFR in a subset of ACTH-secreting tumors. Other factors involved in initiation and promotion of pituitary tumors include loss of negative-feedback inhibition (as seen with primary hypothyroidism or hypogonadism) and estrogen-mediated or paracrine angiogenesis. Growth characteristics and neoplastic behavior also may be influenced by several activated oncogenes, including *RAS* and pituitary tumor transforming gene (*PTTG*), or inactivation of growth suppressor genes, including *MEG3*.

Genetic Syndromes Associated with Pituitary Tumors

Several familial syndromes are associated with pituitary tumors, and the genetic mechanisms for some of them have been unraveled **(Table 373-4)**.

TABLE 373-4

Familial Pituitary Tumor Syndromes

	GENE MUTATED	CLINICAL FEATURES
Multiple endocrine neoplasia 1 (MEN 1)	*MEN1* (11q13)	Hyperparathyroidism Pancreatic neuroendocrine tumors Foregut carcinoids Adrenal adenomas Skin lesions Pituitary adenomas (40%)
Multiple endocrine neoplasia 4 (MEN 4)	*CDKNIB* (12p13)	Hyperparathyroidism Pituitary adenomas Other tumors
Carney complex	*PRKAR1A* (17q23-24)	Pituitary hyperplasia and adenomas (10%) Atrial myxomas Schwannomas Adrenal hyperplasia Lentigines
Familial pituitary adenomas	*AIP* (11q.13.2)	Acromegaly/gigantism (~15% of afflicted families)

Multiple endocrine neoplasia (MEN) 1 is an autosomal dominant syndrome characterized primarily by a genetic predisposition to parathyroid, pancreatic islet, and pituitary adenomas **(Chap. 381)**. MEN1 is caused by inactivating germline mutations in *MENIN*, a constitutively expressed tumor-suppressor gene located on chromosome 11q13. Loss of heterozygosity or a somatic mutation of the remaining normal *MENIN* allele leads to tumorigenesis. About half of affected patients develop prolactinomas; acromegaly and Cushing syndrome are less commonly encountered.

Carney complex is characterized by spotty skin pigmentation, myxomas, and endocrine tumors, including testicular, adrenal, and pituitary adenomas. Acromegaly occurs in ~20% of these patients. A subset of patients have mutations in the R1α regulatory subunit of protein kinase A (*PRKAR1A*).

McCune-Albright syndrome consists of polyostotic fibrous dysplasia, pigmented skin patches, and a variety of endocrine disorders, including acromegaly, adrenal adenomas, and autonomous ovarian function **(Chap. 405)**. Hormonal hypersecretion results from constitutive cyclic AMP production caused by inactivation of the

GTPase activity of Gsα. The Gsα mutations occur postzygotically, leading to a mosaic pattern of mutant expression.

Familial acromegaly is a rare disorder in which family members may manifest either acromegaly or gigantism. A subset of families with a predisposition for familial pituitary tumors, especially acromegaly, have been found to harbor germline mutations in the *AIP* gene, which encodes the aryl hydrocarbon receptor interacting protein.

HYPERPROLACTINEMIA

Etiology

Hyperprolactinemia is the most common pituitary hormone hypersecretion syndrome in both men and women. PRL-secreting pituitary adenomas (prolactinomas) are the most common cause of PRL levels >200 µg/L (see below). Less pronounced PRL elevation can also be seen with microprolactinomas but is more commonly caused by drugs, pituitary stalk compression, hypothyroidism, or renal failure **(Table 373-5)**.

TABLE 373-5

Etiology of Hyperprolactinemia

- **I. Physiologic hypersecretion**
  - Pregnancy
  - Lactation
  - Chest wall stimulation
  - Sleep
  - Stress
- **II. Hypothalamic–pituitary stalk damage**
  - Tumors
    - Craniopharyngioma
    - Suprasellar pituitary mass
    - Meningioma
    - Dysgerminoma
    - Metastases
  - Empty sella
  - Lymphocytic hypophysitis
  - Adenoma with stalk
  - Compression
  - Granulomas
  - Rathke cyst
  - Irradiation
  - Trauma
    - Pituitary stalk section
    - Suprasellar surgery
- **III. Pituitary hypersecretion**
  - Prolactinoma
  - Acromegaly
- **IV. Systemic disorders**
  - Chronic renal failure
  - Hypothyroidism
  - Cirrhosis
  - Pseudocyesis
  - Epileptic seizures
- **V. Drug-induced hypersecretion**
  - Dopamine receptor blockers
    - Atypical antipsychotics: risperidone
    - Phenothiazines: chlorpromazine, perphenazine

- Butyrophenones: haloperidol
- Thioxanthenes
- Metoclopramide

Dopamine synthesis inhibitors

- α-Methyldopa

Catecholamine depletors

- Reserpine

Opiates

H_2 antagonists

- Cimetidine, ranitidine

Imipramines

- Amitriptyline, amoxapine

Serotonin reuptake inhibitors

- Fluoxetine

Calcium channel blockers

- Verapamil
- Estrogens
- Thyrotropin-releasing hormone

Note: Hyperprolactinemia >200 µg/L almost invariably is indicative of a prolactin-secreting pituitary adenoma. Physiologic causes, hypothyroidism, and drug-induced hyperprolactinemia should be excluded before extensive evaluation.

Pregnancy and lactation are the important physiologic causes of hyperprolactinemia. Sleep-associated hyperprolactinemia reverts to normal within an hour of awakening. Nipple stimulation and sexual orgasm also may increase PRL. Chest wall stimulation or trauma (including chest surgery and herpes zoster) invoke the reflex suckling arc with resultant hyperprolactinemia. Chronic renal failure elevates PRL by decreasing peripheral clearance. Primary hypothyroidism is associated with mild hyperprolactinemia, probably because of compensatory TRH secretion. Mutation of the PRL receptor is a rare cause of hyperprolactinemia.

Lesions of the hypothalamic-pituitary region that disrupt hypothalamic dopamine synthesis, portal vessel delivery, or lactotrope responses are associated with hyperprolactinemia. Thus, hypothalamic tumors, cysts, infiltrative disorders, and radiation-induced damage cause elevated PRL levels, usually in the range of 30–100 µg/L. Plurihormonal adenomas (including GH and ACTH tumors) may hypersecrete PRL directly. Pituitary masses, including clinically nonfunctioning pituitary tumors, may compress the pituitary stalk to cause hyperprolactinemia.

Drug-induced inhibition or disruption of dopaminergic receptor function is a common cause of hyperprolactinemia (Table 373-5). Thus, antipsychotics and antidepressants are a relatively common cause of mild hyperprolactinemia. Most patients receiving risperidone have elevated prolactin levels, sometimes exceeding 200 µg/L. Methyldopa inhibits dopamine synthesis, and verapamil blocks dopamine release, also

leading to hyperprolactinemia. Hormonal agents that induce PRL include estrogens and thyrotropin-releasing hormone (TRH).

Presentation and Diagnosis

Amenorrhea, galactorrhea, and infertility are the hallmarks of hyperprolactinemia in women. If hyperprolactinemia develops before menarche, primary amenorrhea results. More commonly, hyperprolactinemia develops later in life and leads to oligomenorrhea and ultimately to amenorrhea. If hyperprolactinemia is sustained, vertebral bone mineral density can be reduced compared with age-matched controls, particularly when it is associated with pronounced hypoestrogenemia. Galactorrhea is present in up to 80% of hyperprolactinemic women. Although usually bilateral and spontaneous, it may be unilateral or expressed only manually. Patients also may complain of decreased libido, weight gain, and mild hirsutism.

In men with hyperprolactinemia, diminished libido, infertility, and visual loss (from optic nerve compression) are the usual presenting symptoms. Gonadotropin suppression leads to reduced testosterone, impotence, and oligospermia. True galactorrhea is uncommon in men with hyperprolactinemia. If the disorder is long-standing, secondary effects of hypogonadism are evident, including osteopenia, reduced muscle mass, and decreased beard growth.

The diagnosis of idiopathic hyperprolactinemia is made by exclusion of known causes of hyperprolactinemia in the setting of a normal pituitary MRI. Some of these patients may harbor small microadenomas below visible MRI sensitivity (~2 mm).

GALACTORRHEA

Galactorrhea, the inappropriate discharge of milk-containing fluid from the breast, is considered abnormal if it persists longer than 6 months after childbirth or discontinuation of breast-feeding. Postpartum galactorrhea associated with amenorrhea is a self-limiting disorder usually associated with moderately elevated PRL levels. Galactorrhea may occur spontaneously, or it may be elicited by nipple pressure. In both men and women, galactorrhea may vary in color and consistency (transparent, milky, or bloody) and arise either unilaterally or bilaterally. Mammography or ultrasound is indicated for bloody discharges (particularly from a single nipple), which may be caused by breast cancer. Galactorrhea is commonly associated with hyperprolactinemia caused by any of the conditions listed in Table 373-5. Acromegaly is associated with galactorrhea in about one-third of patients. Treatment of galactorrhea usually involves managing the underlying disorder (e.g., replacing T_4 for hypothyroidism, discontinuing a medication, treating prolactinoma).

Laboratory Investigation

Basal, fasting morning PRL levels (normally <20 µg/L) should be measured to assess hypersecretion. Both false-positive and false-negative results may be encountered. In patients with markedly elevated PRL levels (>1000 µg/L), reported results may be falsely lowered because of assay artifacts; sample dilution is required to measure these high values accurately. Falsely elevated values may be caused by aggregated forms of

circulating PRL, which are usually biologically inactive (macroprolactinemia). Hypothyroidism should be excluded by measuring TSH and T_4 levels.

TREATMENT

TREATMENT

Hyperprolactinemia

Treatment of hyperprolactinemia depends on the cause of elevated PRL levels. Regardless of the etiology, however, treatment should be aimed at normalizing PRL levels to alleviate suppressive effects on gonadal function, halt galactorrhea, and preserve bone mineral density. Dopamine agonists are effective for most causes of hyperprolactinemia (see the treatment section for prolactinoma, below) regardless of the underlying cause.

If the patient is taking a medication known to cause hyperprolactinemia, the drug should be withdrawn, if possible. For psychiatric patients who require neuroleptic agents, supervised dose titration or the addition of a dopamine agonist can help restore normoprolactinemia and alleviate reproductive symptoms. However, dopamine agonists may worsen the underlying psychiatric condition, especially at high doses. Hyperprolactinemia usually resolves after adequate thyroid hormone replacement in hypothyroid patients or after renal transplantation in patients undergoing dialysis. Resection of hypothalamic or sellar mass lesions can reverse hyperprolactinemia caused by stalk compression and reduced dopamine tone. Granulomatous infiltrates occasionally respond to glucocorticoid administration. In patients with irreversible hypothalamic damage, no treatment may be warranted. In up to 30% of patients with hyperprolactinemia—usually without a visible pituitary microadenoma—the condition may resolve spontaneously.

PROLACTINOMA

Etiology and Prevalence

Tumors arising from lactotrope cells account for about half of all functioning pituitary tumors, with a population prevalence of ~10/100,000 in men and ~30/100,000 in women. Mixed tumors that secrete combinations of GH and PRL, ACTH and PRL, and rarely TSH and PRL are also seen. These plurihormonal tumors are usually recognized by immunohistochemistry, sometimes without apparent clinical manifestations from the production of additional hormones. Microadenomas are classified as <1 cm in diameter and usually do not invade the parasellar region. Macroadenomas are >1 cm in diameter and may be locally invasive and impinge on adjacent structures. The female-to-male ratio for microprolactinomas is 20:1, whereas the sex ratio is near 1:1 for macroadenomas. Tumor size generally correlates directly with PRL concentrations; values >250 µg/L usually are associated with macroadenomas. Men tend to present with larger tumors than women, possibly because the features of male hypogonadism are less readily evident. PRL levels remain stable in most patients, reflecting the slow growth of these tumors. About 5% of microadenomas progress in the long term to macroadenomas.

Presentation and Diagnosis

Women usually present with amenorrhea, infertility, and galactorrhea. If the tumor extends outside the sella, visual field defects or other mass effects may be seen. Men often present with impotence, loss of libido, infertility, or signs of central nervous system (CNS) compression, including headaches and visual defects. Assuming that physiologic and medication-induced causes of hyperprolactinemia are excluded (Table 373-5), the diagnosis of prolactinoma is likely with a PRL level >200 µg/L. PRL levels <100 µg/L may be caused by microadenomas, other sellar lesions that decrease dopamine inhibition, or nonneoplastic causes of hyperprolactinemia. For this reason, an MRI should be performed in all patients with hyperprolactinemia. It is important to remember that hyperprolactinemia caused secondarily by the mass effects of nonlactotrope lesions is also corrected by treatment with dopamine agonists despite failure to shrink the underlying mass. Consequently, PRL suppression by dopamine agonists does not necessarily indicate that the underlying lesion is a prolactinoma.

TREATMENT

TREATMENT

Prolactinoma

Because microadenomas rarely progress to become macroadenomas, no treatment may be needed if patients are asymptomatic and fertility is not desired; these patients should be monitored by regular serial PRL measurements and MRI scans. For symptomatic microadenomas, therapeutic goals include control of hyperprolactinemia, reduction of tumor size, restoration of menses and fertility, and resolution of galactorrhea. Dopamine agonist doses should be titrated to achieve maximal PRL suppression and restoration of reproductive function **(Fig. 373-3)**. A normalized PRL level does not ensure reduced tumor size. However, tumor shrinkage usually is not seen in those who do not respond with lowered PRL levels. For macroadenomas, formal visual field testing should be performed before initiating dopamine agonists. MRI and visual fields should be assessed at 6- to 12-month intervals until the mass shrinks and annually thereafter until maximum size reduction has occurred.

MEDICAL

Oral dopamine agonists (cabergoline and bromocriptine) are the mainstay of therapy for patients with micro- or macroprolactinomas. Dopamine agonists suppress PRL secretion and synthesis as well as lactotrope cell proliferation. In patients with microadenomas who have achieved normoprolactinemia and significant reduction of tumor mass, the dopamine agonist may be withdrawn after 2 years. These patients should be monitored carefully for evidence of prolactinoma recurrence. About 20% of patients (especially males) are resistant to dopaminergic treatment; these adenomas may exhibit decreased D_2 dopamine receptor numbers or a postreceptor defect. D_2 receptor gene mutations in the pituitary have not been reported.

Cabergoline

An ergoline derivative, cabergoline is a long-acting dopamine agonist with high D_2 receptor affinity. The drug effectively suppresses PRL for >14 days after a single oral dose and induces prolactinoma shrinkage in most patients. Cabergoline (0.5–1.0 mg twice weekly) achieves normoprolactinemia and resumption of normal gonadal function in ~80% of patients with microadenomas; galactorrhea improves or resolves in 90% of patients. Cabergoline normalizes PRL and shrinks ~70% of macroprolactinomas. Mass effect symptoms, including headaches and visual disorders, usually improve dramatically within days after cabergoline initiation; improvement of sexual function requires several weeks of treatment but may occur before complete normalization of PRL levels. After initial control of PRL levels has been achieved, cabergoline should be reduced to the lowest effective maintenance dose. In ~5% of treated patients harboring a microadenoma, hyperprolactinemia may resolve and not recur when dopamine agonists are discontinued after long-term treatment. Cabergoline also may be effective in patients resistant to bromocriptine. Adverse effects and drug intolerance are encountered less commonly than with bromocriptine.

Bromocriptine

The ergot alkaloid bromocriptine mesylate is a dopamine receptor agonist that suppresses PRL secretion. Because it is short-acting, the drug is preferred when pregnancy is desired. In microadenomas, bromocriptine rapidly lowers serum PRL levels to normal in up to 70% of patients, decreases tumor size, and restores gonadal function. In patients with macroadenomas, PRL levels are also normalized in 70% of patients, and tumor mass shrinkage (≥50%) is achieved in most patients.

Therapy is initiated by administering a low bromocriptine dose (0.625–1.25 mg) at bedtime with a snack, followed by gradually increasing the dose. Most patients are controlled with a daily dose of <7.5 mg (2.5 mg tid).

SIDE EFFECTS

Side effects of dopamine agonists include constipation, nasal stuffiness, dry mouth, nightmares, insomnia, and vertigo; decreasing the dose usually alleviates these problems. Nausea, vomiting, and postural hypotension with faintness may occur in ~25% of patients after the initial dose. These symptoms may persist in some patients. In general, fewer side effects are reported with cabergoline. For the ~15% of patients who are intolerant of oral bromocriptine, cabergoline may be better tolerated. Intravaginal administration of bromocriptine is often efficacious in patients with intractable gastrointestinal side effects. Auditory hallucinations, delusions, and mood swings have been reported in up to 5% of patients and may be due to the dopamine agonist properties or to the lysergic acid derivative of the compounds. Rare reports of leukopenia, thrombocytopenia, pleural fibrosis, cardiac arrhythmias, and hepatitis have been described. Patients with Parkinson disease who receive at least 3 mg of cabergoline daily have been reported to be at risk for development of cardiac valve regurgitation. Studies analyzing >500 prolactinoma patients receiving recommended doses of cabergoline (up to 2 mg weekly) have shown no evidence for an increased incidence of valvular disorders. Nevertheless, because no controlled prospective studies in pituitary tumor patients are available, it is prudent to perform echocardiograms before initiating standard-dose cabergoline therapy.

Surgery

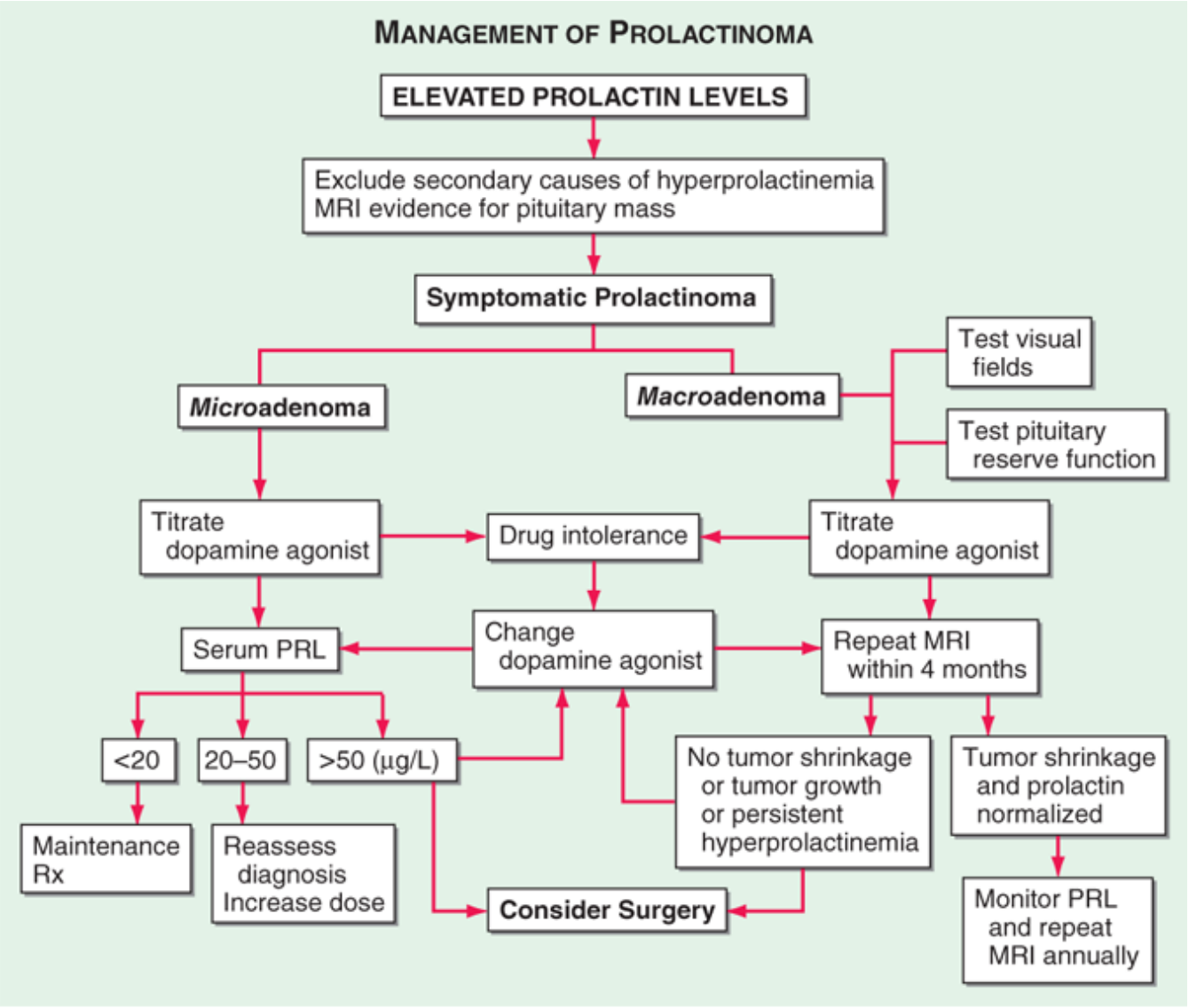
Indications for surgical adenoma debulking include dopamine resistance or intolerance and the presence of an invasive macroadenoma with compromised vision that fails to improve after drug treatment. Initial PRL normalization is achieved in ~70% of microprolactinomas after surgical resection, but only 30% of macroadenomas can be resected successfully. Follow-up studies have shown that hyperprolactinemia recurs in up to 20% of patients within the first year after surgery; long-term recurrence rates exceed 50% for macroadenomas. Radiotherapy for prolactinomas is reserved for patients with aggressive tumors that do not respond to maximally tolerated dopamine agonists and/or surgery.

PREGNANCY

The pituitary increases in size during pregnancy, reflecting the stimulatory effects of estrogen and perhaps other growth factors on pituitary vascularity and lactotrope cell hyperplasia. About 5% of microadenomas significantly increase in size, but 15–30% of macroadenomas grow during pregnancy. Bromocriptine has been used for >30 years to restore fertility in women with hyperprolactinemia, without evidence of teratogenic effects. Nonetheless, most authorities recommend strategies to minimize fetal exposure to the drug. For women taking bromocriptine who desire pregnancy, mechanical contraception should be used through three regular menstrual cycles to allow for conception timing. When pregnancy is confirmed, bromocriptine should be discontinued and PRL levels followed serially, especially if headaches or visual symptoms occur. For women harboring macroadenomas, regular visual field testing is recommended, and the drug should be reinstituted if tumor growth is apparent. Although pituitary MRI may be safe during pregnancy, this procedure should be reserved for symptomatic patients with severe headache and/or visual field defects. Surgical decompression may be indicated if vision is threatened. Although comprehensive data support the efficacy and relative safety of bromocriptine-facilitated fertility, patients should be advised of potential unknown deleterious effects and the risk of tumor growth during pregnancy. Because cabergoline is long-acting with a high D_2-receptor affinity, it is not recommended for use in women when fertility is desired.

FIGURE 373-3

Management of prolactinoma. MRI, magnetic resonance imaging; PRL, prolactin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ACROMEGALY

Etiology

GH hypersecretion is usually the result of a somatotrope adenoma but may rarely be caused by extrapituitary lesions (Table 373-6). In addition to the more common GH-secreting somatotrope adenomas, mixed mammosomatotrope tumors and acidophilic stem-cell adenomas secrete both GH and PRL. In patients with acidophilic stem-cell adenomas, features of hyperprolactinemia (hypogonadism and galactorrhea) predominate over the less clinically evident signs of acromegaly. Occasionally, mixed plurihormonal tumors are encountered that also secrete ACTH, the glycoprotein hormone α subunit, or TSH in addition to GH. Patients with partially empty sellae may present with GH hypersecretion due to a small GH-secreting adenoma within the compressed rim of pituitary tissue; some of these may reflect the spontaneous necrosis of tumors that were previously larger. GH-secreting tumors rarely arise from ectopic pituitary tissue remnants in the nasopharynx or midline sinuses.

TABLE 373-6

Causes of Acromegaly

	PREVALENCE, %
Excess Growth Hormone Secretion	
Pituitary	98
Densely or sparsely granulated GH cell adenoma	60
Mixed GH cell and PRL cell adenoma	25
Mammosomatotrope cell adenoma	10
Plurihormonal adenoma	
GH cell carcinoma or metastases	
Multiple endocrine neoplasia 1 (GH cell adenoma)	
McCune-Albright syndrome	
Ectopic sphenoid or parapharyngeal sinus pituitary adenoma	
Extrapituitary tumor	<1
Pancreatic islet cell tumor	
Lymphoma	
Excess Growth Hormone–Releasing Hormone Secretion	
Central	<1
Hypothalamic hamartoma, choristoma, ganglioneuroma	
Peripheral	<1
Bronchial carcinoid, pancreatic islet cell tumor, small cell lung cancer, adrenal adenoma, medullary thyroid carcinoma, pheochromocytoma	

Abbreviations: GH, growth hormone; PRL, prolactin.

Source: Adapted from S Melmed: N Engl J Med 355:2558, 2006.

There are case reports of ectopic GH secretion by tumors of pancreatic, ovarian, lung, or hematopoietic origin. Rarely, excess GHRH production may cause acromegaly because of chronic stimulation of somatotropes. These patients present with classic features of acromegaly, elevated GH levels, pituitary enlargement on MRI, and pathologic characteristics of pituitary hyperplasia. The most common cause of GHRH-mediated acromegaly is a chest or abdominal carcinoid tumor. Although these tumors usually express positive GHRH immunoreactivity, clinical features of acromegaly are evident in only a minority of patients with carcinoid disease. Excessive GHRH also may be elaborated by hypothalamic tumors, usually choristomas or neuromas.

Presentation and Diagnosis

Protean manifestations of GH and IGF-I hypersecretion are indolent and often are not clinically diagnosed for 10 years or more. Acral bony overgrowth results in frontal bossing, increased hand and foot size, mandibular enlargement with prognathism, and widened space between the lower incisor teeth. In children and adolescents, initiation of GH hypersecretion before epiphyseal long bone closure is associated with development of pituitary gigantism **(Fig. 373-4)**. Soft tissue swelling results in increased heel pad thickness, increased shoe or glove size, ring tightening, characteristic coarse facial features, and a large fleshy nose. Other commonly encountered clinical features include hyperhidrosis, a deep and hollow-sounding voice, oily skin, arthropathy, kyphosis, carpal tunnel syndrome, proximal muscle weakness and fatigue, acanthosis nigricans, and skin tags. Generalized visceromegaly occurs, including cardiomegaly, macroglossia, and thyroid gland enlargement.

FIGURE 373-4

Features of acromegaly/gigantism. A 22-year-old man with gigantism due to excess growth hormone is shown to the left of his identical twin. The increased height and prognathism (***A***) and enlarged hand (***B***) and foot (***C***) of the affected twin are apparent. Their clinical features began to diverge at the age of ~13 years. *(Reproduced from R Gagel, IE McCutcheon: N Engl J Med 324:524, 1999; with permission.)*



A



B



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The most significant clinical impact of GH excess occurs with respect to the cardiovascular system. Cardiomyopathy with arrhythmias, left ventricular hypertrophy, decreased diastolic function, and hypertension ultimately occur in most patients if untreated. Upper airway obstruction with sleep apnea occurs in >60% of patients and is associated with both soft tissue laryngeal airway obstruction and central sleep dysfunction. Diabetes mellitus develops in 25% of patients with acromegaly, and most patients are intolerant of a glucose load (as GH counteracts the action of insulin). Acromegaly is associated with an increased risk of colon polyps and mortality from colonic malignancy; polyps are diagnosed in up to one-third of patients. Overall mortality is increased about threefold and is due primarily to cardiovascular and cerebrovascular disorders and respiratory disease. Unless GH levels are controlled, survival is reduced by an average of 10 years compared with an age-matched control population.

Laboratory Investigation

Age-matched serum IGF-I levels are elevated in acromegaly. Consequently, an IGF-I level provides a useful laboratory screening measure when clinical features raise the possibility of acromegaly. Owing to the pulsatility of GH secretion, measurement of a single random GH level is not useful for the diagnosis or

exclusion of acromegaly and does not correlate with disease severity. The diagnosis of acromegaly is confirmed by demonstrating the failure of GH suppression to <0.4 μg/L within 1–2 h of an oral glucose load (75 g). When newer ultrasensitive GH assays are used, normal nadir GH levels are even lower (<0.05 μg/L). About 20% of patients exhibit a paradoxical GH rise after glucose. PRL should be measured, as it is elevated in ~25% of patients with acromegaly. Thyroid function, gonadotropins, and sex steroids may be attenuated because of tumor mass effects. Because most patients will undergo surgery with glucocorticoid coverage, tests of ACTH reserve in asymptomatic patients are more efficiently deferred until after surgery.

TREATMENT

TREATMENT

Acromegaly

The goal of treatment is to control GH and IGF-I hypersecretion, ablate or arrest tumor growth, ameliorate comorbidities, restore mortality rates to normal, and preserve pituitary function.

Surgical resection of GH-secreting adenomas is the initial treatment for most patients **(Fig. 373-5)**. Somatostatin analogues are used as adjuvant treatment for preoperative shrinkage of large invasive macroadenomas, immediate relief of debilitating symptoms, and reduction of GH hypersecretion; in frail patients experiencing morbidity; and in patients who decline surgery or, when surgery fails, to achieve biochemical control. Irradiation or repeat surgery may be required for patients who cannot tolerate or do not respond to adjunctive medical therapy. The high rate of late hypopituitarism and the slow rate (5–15 years) of biochemical response are the main disadvantages of radiotherapy. Irradiation is also relatively ineffective in normalizing IGF-I levels. Stereotactic ablation of GH-secreting adenomas by gamma-knife radiotherapy is promising, but long-term results and side effects appear similar to those observed with conventional radiation. Somatostatin analogues may be required while awaiting the full benefits of radiotherapy. Systemic comorbid sequelae of acromegaly, including cardiovascular disease, diabetes, and arthritis, should be managed aggressively. Mandibular surgical repair may be indicated.

SURGERY

Transsphenoidal surgical resection by an experienced surgeon is the preferred primary treatment for both microadenomas (remission rate ~70%) and macroadenomas (<50% in remission). Soft tissue swelling improves immediately after tumor resection. GH levels return to normal within an hour, and IGF-I levels are normalized within 3–4 days. In ~10% of patients, acromegaly may recur several years after apparently successful surgery; hypopituitarism develops in up to 15% of patients after surgery.

SOMATOSTATIN ANALOGUES

Somatostatin analogues exert their therapeutic effects through SSTR2 and SSTR5 receptors, both of which are expressed by GH-secreting tumors. Octreotide acetate is an eight-amino-acid synthetic somatostatin analogue. In contrast to native somatostatin, the analogue is relatively resistant to plasma degradation. It has a 2-h serum half-life and possesses 40-fold greater potency than native somatostatin to suppress GH. Octreotide is administered by subcutaneous injection, beginning with 50 μg tid; the dose can be increased

gradually up to 1500 µg/d. Octreotide suppresses integrated GH levels and normalizes IGF-I levels in ~60% of treated patients.

The long-acting somatostatin depot formulations, octreotide and lanreotide, are the preferred medical treatment for patients with acromegaly. *Octreotide LAR* is a sustained-release, long-acting formulation of octreotide incorporated into microspheres that sustain drug levels for several weeks after intramuscular injection. GH suppression occurs for as long as 6 weeks after a 30-mg intramuscular injection; long-term monthly treatment sustains GH and IGF-I suppression and also reduces pituitary tumor size in ~50% of patients. *Lanreotide* Autogel, a slow-release depot somatostatin preparation, is a cyclic somatostatin octapeptide analogue that suppresses GH and IGF-I hypersecretion after a 60-mg subcutaneous injection. Long-term (every 4–6 weeks) administration controls GH hypersecretion in about two-thirds of treated patients and improves patient compliance because of the long interval required between drug injections. Rapid relief of headache and soft tissue swelling occurs in ~75% of patients within days to weeks of somatostatin analogue initiation. Most patients report symptomatic improvement, including amelioration of headache, perspiration, obstructive apnea, and cardiac failure. For those resistant to octreotide, pasireotide, with preferential SST5 binding, has been shown to exhibit efficacy.

Side Effects

Somatostatin analogues are well tolerated in most patients. Adverse effects are short-lived and mostly relate to drug-induced suppression of gastrointestinal motility and secretion. Transient nausea, abdominal discomfort, fat malabsorption, diarrhea, and flatulence occur in one-third of patients, and these symptoms usually remit within 2 weeks. Octreotide suppresses postprandial gallbladder contractility and delays gallbladder emptying; up to 30% of patients develop long-term echogenic sludge or asymptomatic cholesterol gallstones. Other side effects include mild glucose intolerance due to transient insulin suppression, asymptomatic bradycardia, hypothyroxinemia, and local injection site discomfort. Pasireotide is associated with a higher prevalence of glucose intolerance or new-onset diabetes mellitus.

GH RECEPTOR ANTAGONIST

Pegvisomant antagonizes endogenous GH action by blocking peripheral GH binding to its receptor. Consequently, serum IGF-I levels are suppressed, reducing the deleterious effects of excess endogenous GH. Pegvisomant is administered by daily subcutaneous injection (10–20 mg) and normalizes IGF-I in ~70% of patients. GH levels, however, remain elevated as the drug does not target the pituitary adenoma. Side effects include reversible liver enzyme elevation, lipodystrophy, and injection site pain. Tumor size should be monitored by MRI.

Combined treatment with monthly somatostatin analogues and weekly or biweekly pegvisomant injections has been used effectively in resistant patients.

DOPAMINE AGONISTS

Bromocriptine and cabergoline may modestly suppress GH secretion in some patients. Very high doses of bromocriptine (≥20 mg/d) or cabergoline (0.5 mg/d) are usually required to achieve modest GH therapeutic

efficacy. Combined treatment with octreotide and cabergoline may induce additive biochemical control compared with either drug alone.

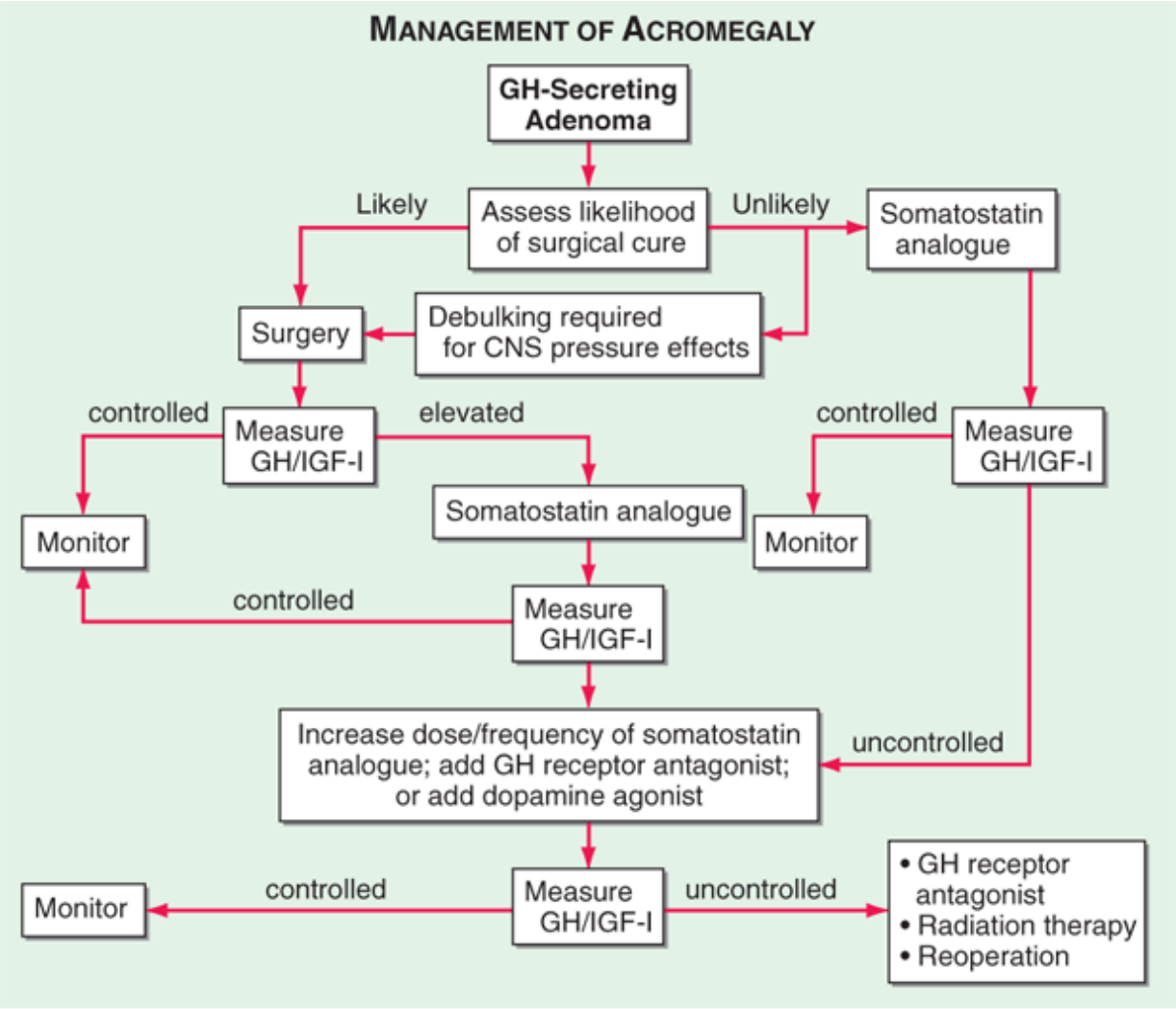
RADIATION

External radiation therapy or high-energy stereotactic techniques are used as adjuvant therapy for acromegaly. An advantage of radiation is that patient compliance with long-term treatment is not required. Tumor mass is reduced, and GH levels are attenuated over time. However, 50% of patients require at least 8 years for GH levels to be suppressed to <5 μg/L; this level of GH reduction is achieved in ~90% of patients after 18 years but represents suboptimal GH suppression. Patients may require interim medical therapy for several years before attaining maximal radiation benefits. Most patients also experience hypothalamic-pituitary damage, leading to gonadotropin, ACTH, and/or TSH deficiency within 10 years of therapy.

In summary, surgery is the preferred primary treatment for GH-secreting microadenomas (Fig. 373-5). The high frequency of GH hypersecretion after macroadenoma resection usually necessitates adjuvant or primary medical therapy for these larger tumors. Patients unable to receive or respond to unimodal medical treatment may benefit from combined treatments, or they can be offered radiation.

FIGURE 373-5

Management of acromegaly. CNS, central nervous system; IGF, insulin-like growth factor; GH, growth hormone. *(Adapted from S Melmed et al: J Clin Endocrinol Metab 94:1509–1517, 2009; © The Endocrine Society.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CUSHING'S DISEASE (ACTH-PRODUCING ADENOMA)

(See also Chap. 379)

Etiology and Prevalence

Pituitary corticotrope adenomas (Cushing's disease) account for 70% of patients with endogenous causes of Cushing's syndrome. However, it should be emphasized that iatrogenic hypercortisolism is the most common cause of cushingoid features. Ectopic tumor ACTH production, cortisol-producing adrenal adenomas, adrenal carcinoma, and adrenal hyperplasia account for the other causes; rarely, ectopic tumor CRH production is encountered.

ACTH-producing adenomas account for ~10–15% of all pituitary tumors. Because the clinical features of Cushing's syndrome often lead to early diagnosis, most ACTH-producing pituitary tumors are relatively small microadenomas. However, macroadenomas also are seen and some ACTH-expressing adenomas are clinically silent. Cushing's disease is 5–10 times more common in women than in men. These pituitary adenomas exhibit unrestrained ACTH secretion, with resultant hypercortisolemia. However, they retain partial

suppressibility in the presence of high doses of administered glucocorticoids, providing the basis for dynamic testing to distinguish pituitary from nonpituitary causes of Cushing's syndrome.

Presentation and Diagnosis

The diagnosis of Cushing's syndrome presents two great challenges: (1) to distinguish patients with pathologic cortisol excess from those with physiologic or other disturbances of cortisol production and (2) to determine the etiology of pathologic cortisol excess.

Typical features of chronic cortisol excess include thin skin, central obesity, hypertension, plethoric moon facies, purple striae and easy bruisability, glucose intolerance or diabetes mellitus, gonadal dysfunction, osteoporosis, proximal muscle weakness, signs of hyperandrogenism (acne, hirsutism), and psychological disturbances (depression, mania, and psychoses) (Table 373-7). Hematopoietic features of hypercortisolism include leukocytosis, lymphopenia, and eosinopenia. Immune suppression includes delayed hypersensitivity and infection propensity. These protean yet commonly encountered manifestations of hypercortisolism make it challenging to decide which patients mandate formal laboratory evaluation. Certain features make pathologic causes of hypercortisolism more likely; they include characteristic central redistribution of fat, thin skin with striae and bruising, and proximal muscle weakness. In children and young females, early osteoporosis may be particularly prominent. The primary cause of death is cardiovascular disease, but life-threatening infections and risk of suicide are also increased.

TABLE 373-7

Clinical Features of Cushing's Syndrome (All Ages)

SYMPTOMS/SIGNS	FREQUENCY, %
Obesity or weight gain (>115% ideal body weight)	80
Thin skin	80
Moon facies	75
Hypertension	75
Purple skin striae	65
Hirsutism	65
Menstrual disorders (usually amenorrhea)	60
Plethora	60
Abnormal glucose tolerance	55
Impotence	55
Proximal muscle weakness	50
Truncal obesity	50
Acne	45
Bruising	45
Mental changes	45
Osteoporosis	40
Edema of lower extremities	30
Hyperpigmentation	20
Hypokalemic alkalosis	15

SYMPTOMS/SIGNS	FREQUENCY, %
Diabetes mellitus	15

Source: Adapted from MA Magiokou et al, in ME Wierman (ed): Diseases of the Pituitary. Totowa, NJ, Humana, 1997.

Rapid development of features of hypercortisolism associated with skin hyperpigmentation and severe myopathy suggests an ectopic tumor source of ACTH. Hypertension, hypokalemic alkalosis, glucose intolerance, and edema are also more pronounced in these patients. Serum potassium levels <3.3 mmol/L are evident in ~70% of patients with ectopic ACTH secretion but are seen in <10% of patients with pituitary-dependent Cushing's syndrome.

Laboratory Investigation

The diagnosis of Cushing's disease is based on laboratory documentation of endogenous hypercortisolism. Measurement of 24-h urine free cortisol (UFC) is a precise and cost-effective screening test. Alternatively, the failure to suppress plasma cortisol after an overnight 1-mg dexamethasone suppression test can be used to identify patients with hypercortisolism. As nadir levels of cortisol occur at night, elevated midnight serum or salivary samples of cortisol are suggestive of Cushing's disease. Basal plasma ACTH levels often distinguish patients with ACTH-independent (adrenal or exogenous glucocorticoid) from those with ACTH-dependent (pituitary, ectopic ACTH) Cushing's syndrome. Mean basal ACTH levels are about eightfold higher in patients with ectopic ACTH secretion than in those with pituitary ACTH-secreting adenomas. However, extensive overlap of ACTH levels in these two disorders precludes using ACTH measurements to make the distinction. Preferably, dynamic testing based on differential sensitivity to glucocorticoid feedback or ACTH stimulation in response to CRH or cortisol reduction is used to distinguish ectopic from pituitary sources of excess ACTH **(Table 373-8)**. Very rarely, circulating CRH levels are elevated, reflecting ectopic tumor-derived secretion of CRH and often ACTH. **For further discussion of dynamic testing for Cushing's syndrome, see Chap. 379.**

TABLE 373-8

Differential Diagnosis of ACTH-Dependent Cushing's Syndrome[a]

	ACTH-SECRETING PITUITARY TUMOR	ECTOPIC ACTH SECRETION
Etiology	Pituitary corticotrope adenoma Plurihormonal adenoma	Bronchial, abdominal carcinoid Small cell lung cancer Thymoma
Sex	F > M	M > F
Clinical features	Slow onset	Rapid onset Pigmentation Severe myopathy
Serum potassium <3.3 μg/L	<10%	75%
24-h UFC	High	High
Basal ACTH level	Inappropriately high	Very high
Dexamethasone suppression 1 mg overnight Low-dose(0.5 mg q6h)	Cortisol >5 μg/dL	Cortisol >5 μg/dL
High-dose (2 mg q6h)	Cortisol <5 μg/dL	Cortisol >5 μg/dL
UFC >80% suppressed	Microadenomas: 90% Macroadenomas: 50%	10%
Inferior petrosal sinus sampling (IPSS)		
Basal IPSS: peripheral	>2	<2
CRH-induced IPSS: peripheral	>3	<3

[a]ACTH-independent causes of Cushing's syndrome are diagnosed by suppressed ACTH levels and an adrenal mass in the setting of hypercortisolism. Iatrogenic Cushing's syndrome is excluded by history.

Abbreviations: ACTH, adrenocorticotropic hormone; CRH, corticotropin-releasing hormone; F, female; M, male; IPSS, inferior petrosal sinus sampling; UFC, urinary free cortisol.

Most ACTH-secreting pituitary tumors are <5 mm in diameter, and about half are undetectable by sensitive MRI. The high prevalence of incidental pituitary microadenomas diminishes the ability to distinguish ACTH-secreting pituitary tumors accurately from nonsecreting incidentalomas.

Inferior Petrosal Venous Sampling

Because pituitary MRI with gadolinium enhancement is insufficiently sensitive to detect small (<2 mm) pituitary ACTH-secreting adenomas, bilateral inferior petrosal sinus ACTH sampling before and after CRH administration may be required to distinguish these lesions from ectopic ACTH-secreting tumors that may have similar clinical and biochemical characteristics. Simultaneous assessment of ACTH in each inferior petrosal vein and in the diagnosis of peripheral circulation provides a strategy for confirming and localizing pituitary ACTH production. Sampling is performed at baseline and 2, 5, and 10 min after intravenous bovine CRH (1 μg/kg) injection. An increased ratio (>2) of inferior petrosal:peripheral vein ACTH confirms pituitary Cushing's syndrome. After CRH injection, peak petrosal:peripheral ACTH ratios ≥3 confirm the presence of a pituitary ACTH-secreting tumor. The sensitivity of this test is >95%, with very rare false-positive results. False-negative results may be encountered in patients with aberrant venous drainage. Petrosal sinus catheterizations are technically difficult, and ~0.05% of patients develop neurovascular complications. The procedure should not be performed in patients with hypertension, in patients with known cerebrovascular disease, or in the presence of a well-visualized pituitary adenoma on MRI.

TREATMENT

TREATMENT

Cushing's Disease

Selective transsphenoidal resection is the treatment of choice for Cushing's disease **(Fig. 373-6)**. The remission rate for this procedure is ~80% for microadenomas but <50% for macroadenomas. However, surgery is rarely successful when the adenoma is not visible on MRI. After successful tumor resection, most patients experience a postoperative period of symptomatic ACTH deficiency that may last up to 12 months. This usually requires low-dose cortisol replacement, as patients experience both steroid withdrawal symptoms and have a suppressed hypothalamic-pituitary-adrenal axis. Biochemical recurrence occurs in ~5% of patients in whom surgery was initially successful.

When initial surgery is unsuccessful, repeat surgery is sometimes indicated, particularly when a pituitary source for ACTH is well documented. In older patients, in whom issues of growth and fertility are less important, hemi- or total hypophysectomy may be necessary if a discrete pituitary adenoma is not recognized. Pituitary irradiation may be used after unsuccessful surgery, but it cures only ~15% of patients. Because the

effects of radiation are slow and only partially effective in adults, steroidogenic inhibitors are used in combination with pituitary irradiation to block adrenal effects of persistently high ACTH levels.

Pasireotide (600 or 900 µg/d subcutaneously), a somatostatin analogue with high affinity for SST5 > SST2 receptors, may control hypercortisolemia in a subset of patients with ACTH-secreting pituitary tumors when surgery is not an option or has not been successful. In clinical trials, the drug lowered plasma ACTH levels and normalized 24-h UFC levels in ~20% of patients, and resulted in up to 40% mean pituitary tumor shrinkage. Side effects include development of hyperglycemia and diabetes in up to 70% of patients, likely due to suppressed pancreatic secretion of insulin and incretins. Because patients with hypercortisolism are insulin-resistant, hyperglycemia should be rigorously managed. Other side effects are similar to those encountered for other somatostatin analogues and include transient abdominal discomfort, diarrhea, nausea, and gallstones (20% of patients). The drug requires consistent long-term administration.

Ketoconazole, an imidazole derivative antimycotic agent, inhibits several P450 enzymes and effectively lowers cortisol in most patients with Cushing's disease when administered twice daily (600–1200 mg/d). Elevated hepatic transaminases, gynecomastia, impotence, gastrointestinal upset, and edema are common side effects.

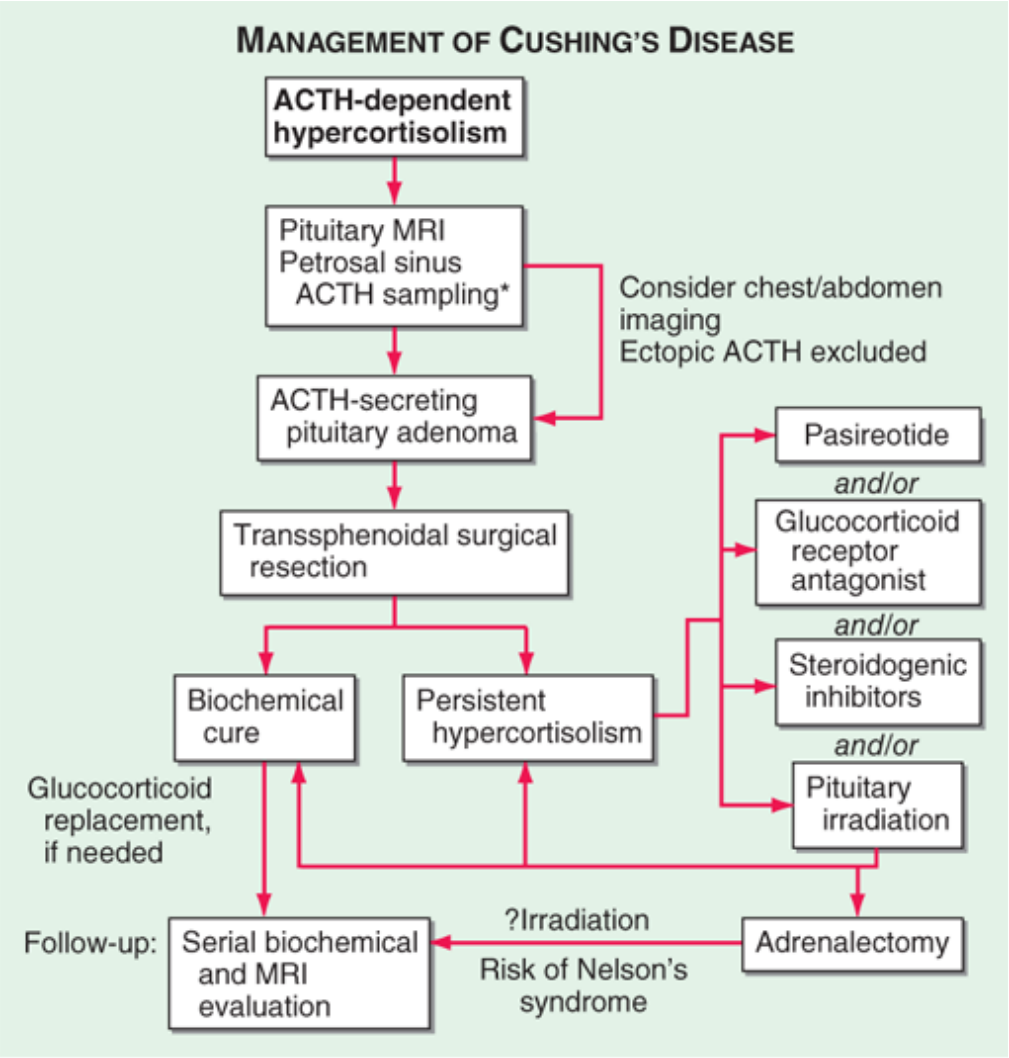
Mifepristone (300–1200 mg/d), a glucocorticoid receptor antagonist, blocks peripheral cortisol action and is approved to treat hyperglycemia in Cushing's disease. Because the drug does not target the pituitary tumor, both ACTH and cortisol levels remain elevated, thus obviating a reliable circulating biomarker. Side effects are largely due to general antagonism of other steroid hormones and include hypokalemia, endometrial hyperplasia, hypoadrenalism, and hypertension.

Metyrapone (2–4 g/d) inhibits 11β-hydroxylase activity and normalizes plasma cortisol in up to 75% of patients. Side effects include nausea and vomiting, rash, and exacerbation of acne or hirsutism. *Mitotane* (*o,p'*-DDD; 3–6 g/d orally in four divided doses) suppresses cortisol hypersecretion by inhibiting 11β-hydroxylase and cholesterol side-chain cleavage enzymes and by destroying adrenocortical cells. Side effects of mitotane include gastrointestinal symptoms, dizziness, gynecomastia, hyperlipidemia, skin rash, and hepatic enzyme elevation. It also may lead to hypoaldosteronism. Other agents include *aminoglutethimide* (250 mg tid), *trilostane* (200–1000 mg/d), *cyproheptadine* (24 mg/d), and IV *etomidate* (0.3 mg/kg/h). Glucocorticoid insufficiency is a potential side effect of agents used to block steroidogenesis.

The use of steroidogenic inhibitors has decreased the need for bilateral adrenalectomy. Surgical removal of both adrenal glands corrects hypercortisolism but may be associated with significant morbidity rates and necessitates permanent glucocorticoid and mineralocorticoid replacement. Adrenalectomy in the setting of residual corticotrope adenoma tissue predisposes to the development of *Nelson's syndrome*, a disorder characterized by rapid pituitary tumor enlargement and increased pigmentation secondary to high ACTH levels. Prophylactic radiation therapy may be indicated to prevent the development of Nelson's syndrome after adrenalectomy.

FIGURE 373-6

Management of Cushing's disease. ACTH, adrenocorticotropin hormone; MRI, magnetic resonance imaging; *, Not usually required.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

NONFUNCTIONING AND GONADOTROPIN-PRODUCING PITUITARY ADENOMAS

Etiology and Prevalence

Nonfunctioning pituitary adenomas include those that secrete little or no pituitary hormones as well as tumors that produce too little hormone to result in recognizable clinical features. They are the most common type of pituitary adenoma and are usually macroadenomas at the time of diagnosis because clinical features are not apparent until tumor mass effects occur. Based on immunohistochemistry, most clinically nonfunctioning adenomas can be shown to originate from gonadotrope cells. These tumors typically produce small amounts of intact gonadotropins (usually FSH) as well as uncombined α, LH β, and FSH β subunits. Tumor secretion may lead to elevated α and FSH β subunits and, rarely, to increased LH β subunit levels. Some adenomas express α subunits without FSH or LH. TRH administration often induces an atypical increase of tumor-derived gonadotropins or subunits.

Presentation and Diagnosis

Clinically nonfunctioning tumors often present with optic chiasm pressure and other symptoms of local expansion or may be incidentally discovered on an MRI performed for another indication (incidentaloma). Rarely, menstrual disturbances or ovarian hyperstimulation occur in women with large tumors that produce FSH and LH. More commonly, adenoma compression of the pituitary stalk or surrounding pituitary tissue leads to attenuated LH and features of hypogonadism. PRL levels are usually slightly increased, also because of stalk compression. It is important to distinguish this circumstance from true prolactinomas, as nonfunctioning tumors do not shrink in response to treatment with dopamine agonists.

Laboratory Investigation

The goal of laboratory testing in clinically nonfunctioning tumors is to classify the type of the tumor, identify hormonal markers of tumor activity, and detect possible hypopituitarism. Free α subunit levels may be elevated in 10–15% of patients with nonfunctioning tumors. In female patients, peri- or postmenopausal basal FSH concentrations are difficult to distinguish from tumor-derived FSH elevation. Premenopausal women have cycling FSH levels, also preventing clear-cut diagnostic distinction from tumor-derived FSH. In men, gonadotropin-secreting tumors may be diagnosed because of slightly increased gonadotropins (FSH > LH) in the setting of a pituitary mass. Testosterone levels are usually low despite the normal or increased LH level, perhaps reflecting reduced LH bioactivity or the loss of normal LH pulsatility. Because this pattern of hormone test results is also seen in primary gonadal failure and, to some extent, with aging (Chap. 384), the finding of increased gonadotropins alone is insufficient for the diagnosis of a gonadotropin-secreting tumor. In the majority of patients with gonadotrope adenomas, TRH administration stimulates LH β subunit secretion; this response is not seen in normal individuals. GnRH testing, however, is not helpful for making the diagnosis. For nonfunctioning and gonadotropin-secreting tumors, the diagnosis usually rests on immunohistochemical analyses of surgically resected tumor tissue, as the mass effects of these tumors usually necessitate resection.

Although acromegaly or Cushing's syndrome usually presents with unique clinical features, clinically inapparent (silent) somatotrope or corticotrope adenomas may only be diagnosed by immunostaining of resected tumor tissue. If PRL levels are <100 μg/L in a patient harboring a pituitary mass, a nonfunctioning adenoma causing pituitary stalk compression should be considered.

TREATMENT

TREATMENT

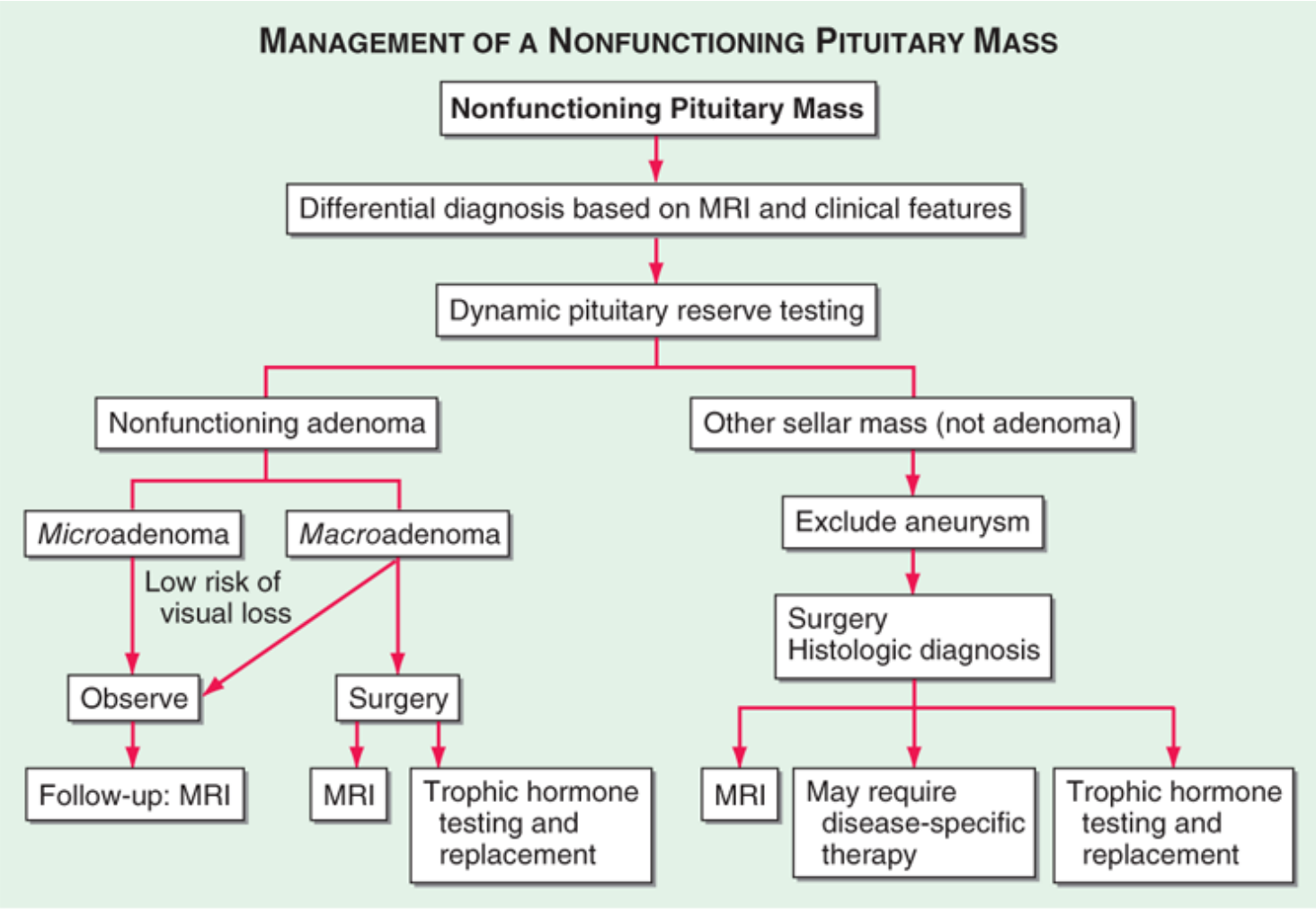
Nonfunctioning and Gonadotropin-Producing Pituitary Adenomas

Asymptomatic small nonfunctioning microadenomas with no threat to vision may be followed with regular MRI and visual field testing without immediate intervention. However, for macroadenomas, transsphenoidal surgery is indicated to reduce tumor size and relieve mass effects (Fig. 373-7). Although it is not usually possible to remove all adenoma tissue surgically, vision improves in 70% of patients with preoperative visual field defects. Preexisting hypopituitarism that results from tumor mass effects may improve or resolve completely. Beginning ~6 months postoperatively, MRI scans should be performed yearly to detect tumor regrowth. Within 5–6 years after successful surgical resection, ~15% of nonfunctioning tumors recur. When

substantial tumor remains after transsphenoidal surgery, adjuvant radiotherapy may be indicated to prevent tumor regrowth. Radiotherapy may be deferred if no postoperative residual mass is evident. Nonfunctioning pituitary tumors respond poorly to dopamine agonist treatment and somatostatin analogues are largely ineffective for shrinking these tumors. The selective GnRH antagonist Nal-Glu GnRH suppresses FSH hypersecretion but has no effect on adenoma size.

FIGURE 373-7

Management of a nonfunctioning pituitary mass. MRI, magnetic resonance imaging.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TSH-SECRETING ADENOMAS

TSH-producing macroadenomas are very rare but are often large and locally invasive when they occur. Patients usually present with thyroid goiter and hyperthyroidism, reflecting overproduction of TSH. Diagnosis is based on demonstrating elevated serum-free T_4 levels, inappropriately normal or high TSH secretion, and MRI evidence of a pituitary adenoma. Elevated uncombined α subunits are seen in many patients.

It is important to exclude other causes of inappropriate TSH secretion, such as resistance to thyroid hormone, an autosomal dominant disorder caused by mutations in the thyroid hormone β receptor (Chap. 375). The

presence of a pituitary mass and elevated β subunit levels are suggestive of a TSH-secreting tumor. Dysalbuminemic hyperthyroxinemia syndromes, caused by mutations in serum thyroid hormone binding proteins, are also characterized by elevated thyroid hormone levels, but with normal rather than suppressed TSH levels. Moreover, free thyroid hormone levels are normal in these disorders, most of which are familial.

TREATMENT

TREATMENT

TSH-Secreting Adenomas

The initial therapeutic approach is to remove or debulk the tumor mass surgically, usually using a transsphenoidal approach. Total resection is not often achieved as most of these adenomas are large and locally invasive. Normal circulating thyroid hormone levels are achieved in about two-thirds of patients after surgery. Thyroid ablation or antithyroid drugs (methimazole and propylthiouracil) can be used to reduce thyroid hormone levels. Somatostatin analogue treatment effectively normalizes TSH and α subunit hypersecretion, shrinks the tumor mass in 50% of patients, and improves visual fields in 75% of patients; euthyroidism is restored in most patients. Because somatostatin analogues markedly suppress TSH, biochemical hypothyroidism often requires concomitant thyroid hormone replacement, which may also further control tumor growth.

FURTHER READING

Di Ieva A et al: Aggressive pituitary adenomas—diagnosis and emerging treatments. Nat Rev Endocrinol 10:423, 2014. [PubMed: 24821329]

Famini P, Maya MM, Melmed S: Pituitary magnetic resonance imaging for sellar and parasellar masses: Ten-year experience in 2598 patients. J Clin Endocrinol Metab 96:1633, 2011. [PubMed: 21470998]

Freda PU et al: Pituitary incidentaloma: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 96:894, 2011. [PubMed: 21474686]

Gittleman H et al: Descriptive epidemiology of pituitary tumors in the United States, 2004–2009. J Neurosurg 121:527, 2014. [PubMed: 24926650]

Hannah-Shmouni F, Trivellin G, Stratakis CA: Genetics of gigantism and acromegaly. Growth Horm IGF Res 30–31:37–41, 2016. [PubMed: 27657986]

Melmed S: Medical progress: acromegaly. N Engl J Med 355:2558, 2006. [PubMed: 17167139]

Melmed S: Pathogenesis of pituitary tumors. Nat Rev Endocrinol 7:257, 2011. [PubMed: 21423242]

Melmed S et al: Diagnosis and treatment of hyperprolactinemia: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 96:273, 2011. [PubMed: 21296991]

Nieman LK: Cushing's syndrome: Update on signs, symptoms and biochemical screening. Eur J Endocrinol 173:M33, 2015.

Ntali G et al: Clinical review: Functioning gonadotroph adenomas. J Clin Endocrinol Metab 99:4423, 2014. [PubMed: 25166722]

Chapter 374: Disorders of the Neurohypophysis

Gary L. Robertson

INTRODUCTION

The neurohypophysis, or posterior pituitary, is formed by axons that originate in large cell bodies in the supraoptic and paraventricular nuclei of the hypothalamus. It produces two hormones: (1) arginine vasopressin (AVP), also known as antidiuretic hormone, and (2) oxytocin. AVP acts on the renal tubules to reduce water loss by concentrating the urine. Oxytocin stimulates postpartum milk letdown in response to suckling. A deficiency of AVP secretion or action causes diabetes insipidus (DI), a syndrome characterized by the production of large amounts of dilute urine. Excessive or inappropriate AVP production impairs urinary water excretion and predisposes to hyponatremia if water intake is not reduced in parallel with urine output.

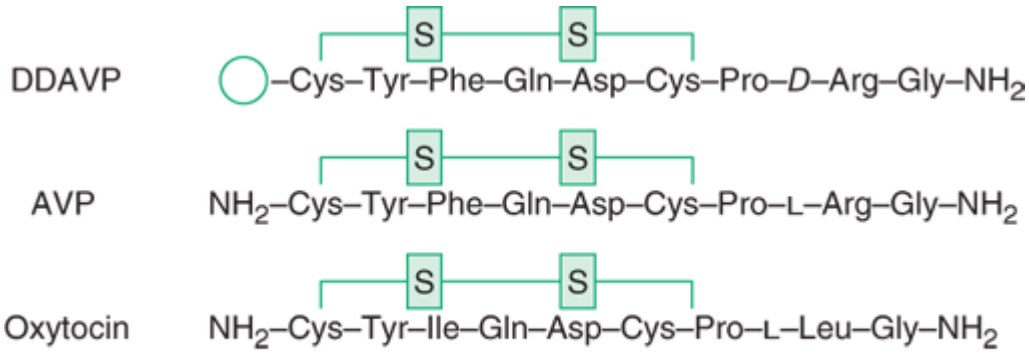
VASOPRESSIN

SYNTHESIS AND SECRETION

AVP is a nonapeptide composed of a six-member disulfide ring and a tripeptide tail **(Fig. 374-1)**. It is synthesized via a polypeptide precursor that includes AVP, neurophysin, and copeptin, all encoded by a single gene on chromosome 20. After preliminary processing and folding, the precursor is packaged in neurosecretory vesicles, where it is transported down the axon; further processed to AVP, neurophysin, and copeptin; and stored in neurosecretory vesicles until released by exocytosis into peripheral blood.

FIGURE 374-1

Primary structures of arginine vasopressin (AVP), oxytocin, and desmopressin (DDAVP).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

AVP secretion is regulated primarily by the “effective” osmotic pressure of body fluids. This control is mediated by specialized hypothalamic cells known as *osmoreceptors*, which are extremely sensitive to small changes in the plasma concentration of sodium and its anions but normally are insensitive to other solutes such as urea and glucose. The osmoreceptors appear to include inhibitory as well as stimulatory components that function in concert to form a threshold, or set point, control system. Below this threshold, plasma AVP is suppressed to levels that permit the development of a maximum water diuresis. Above it, plasma AVP rises steeply in direct proportion to plasma osmolarity, quickly reaching levels sufficient to effect a maximum antidiuresis. The absolute levels of plasma osmolarity/sodium at which minimally and maximally effective levels of plasma AVP occur, vary appreciably from person to person, apparently due to genetic influences on the set and sensitivity of the system. However, the average threshold, or set point, for AVP release corresponds to a plasma osmolarity or sodium of about 280 mosmol/L or 135 meq/L, respectively; levels only 2–4% higher normally result in maximum antidiuresis.

Although it is relatively stable in a healthy adult, the set point of the osmoregulatory system can be lowered by pregnancy, the menstrual cycle, estrogen, and relatively large, acute reductions in blood pressure or volume. Those reductions are mediated largely by neuronal afferents that originate in transmural pressure receptors of the heart and large arteries and project via the vagus and glossopharyngeal nerves to the brainstem, from which postsynaptic projections ascend to the hypothalamus. These pathways maintain a tonic inhibitory tone that decreases when blood volume or pressure falls by >10–20%. This baroregulatory system is probably of minor importance in the physiology of AVP secretion because the hemodynamic changes required to affect it usually do not occur during normal activities. However, the baroregulatory system undoubtedly plays an important role in AVP secretion in patients with disorders that produce large, acute disturbances of hemodynamic function. AVP secretion also can be stimulated by nausea, acute hypoglycemia, glucocorticoid deficiency, smoking, and, possibly, hyperangiotensinemia. The emetic stimuli are extremely potent since they typically elicit immediate, 50- to 100-fold increases in plasma AVP even when the nausea is transient and is not associated with vomiting or other symptoms. They appear to act via the emetic center in the medulla and can be blocked completely by treatment with antiemetics such as fluphenazine. There is no evidence that pain or other noxious stresses have any effect on AVP unless they elicit a vasovagal reaction with its associated nausea and hypotension.

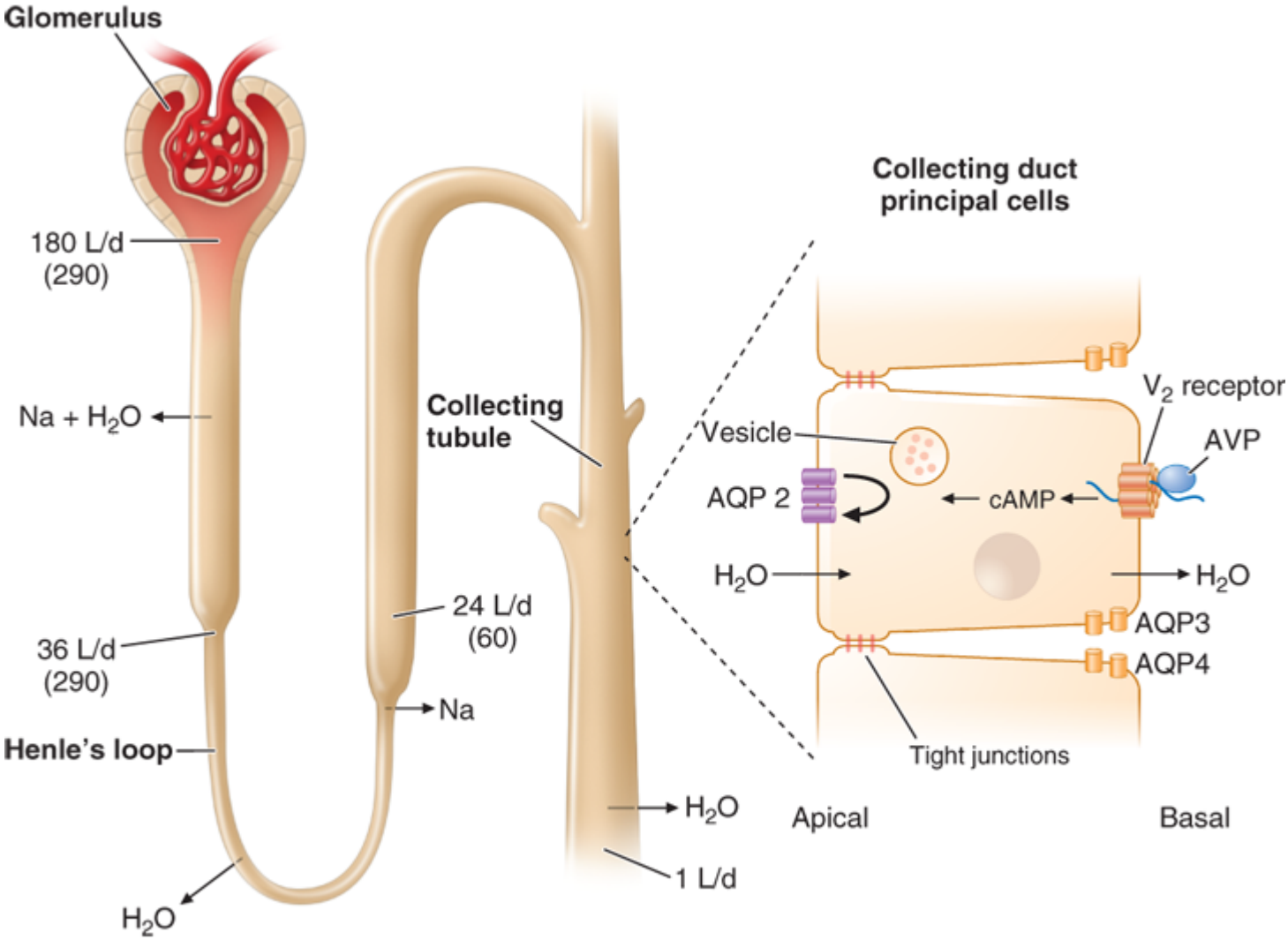
ACTION

The most important, if not the only, physiologic action of AVP is to reduce water excretion by promoting concentration of urine. This antidiuretic effect is achieved by increasing the hydroosmotic permeability of cells that line the distal tubule and medullary collecting ducts of the kidney **(Fig. 374-2)**. In the absence of AVP, these cells are impermeable to water and reabsorb little, if any, of the relatively large volume of dilute filtrate that enters from the proximal nephron. The lack of reabsorption results in the excretion of very large volumes (as much as 0.2 mL/kg per min) of maximally dilute urine (specific gravity and osmolarity ~1.000 and 50 mosmol/L, respectively), a condition known as *water diuresis*. In the presence of AVP, these cells become selectively permeable to water, allowing the water to diffuse back down the osmotic gradient created by the hypertonic renal medulla. As a result, the dilute fluid passing through the tubules is concentrated and the

rate of urine flow decreases. The magnitude of this effect varies in direct proportion to the plasma AVP concentration and the rate of solute excretion. At maximum levels of AVP and normal rates of solute excretion, it approximates a urine flow rate as low as 0.35 mL/min and a urine osmolarity as high as 1200 mosmol/L. This effect is reduced by a solute diuresis such as glucosuria in diabetes mellitus.

FIGURE 374-2

Antidiuretic effect of arginine vasopressin (AVP) in the regulation of urine volume. In a typical 70-kg adult, the kidney filters ~180 L/d of plasma. Of this, ~144 L (80%) is reabsorbed isosmotically in the proximal tubule and another 8 L (4–5%) is reabsorbed without solute in the descending limb of Henle's loop. The remainder is diluted to an osmolarity of ~60 mmol/kg by selective reabsorption of sodium and chloride in the ascending limb. In the absence of AVP, the urine issuing from the loop passes largely unmodified through the distal tubules and collecting ducts, resulting in a maximum water diuresis. In the presence of AVP, solute-free water is reabsorbed osmotically through the principal cells of the collecting ducts, resulting in the excretion of a much smaller volume of concentrated urine. This antidiuretic effect is mediated via a G protein–coupled V_2 receptor that increases intracellular cyclic AMP, thereby inducing translocation of aquaporin 2 (AQP 2) water channels into the apical membrane. The resultant increase in permeability permits an influx of water that diffuses out of the cell through AQP 3 and AQP 4 water channels on the basal-lateral surface. The net rate of flux across the cell is determined by the number of AQP 2 water channels in the apical membrane and the strength of the osmotic gradient between tubular fluid and the renal medulla. Tight junctions on the lateral surface of the cells serve to prevent unregulated water flow. The V2 receptors and aquaporin 2 are encoded by genes on chromosome Xq28 and 12q13, respectively.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

At high concentrations, AVP also causes contraction of smooth muscle in blood vessels in the skin and gastrointestinal tract, induces glycogenolysis in the liver, and potentiates adrenocorticotropic hormone (ACTH) release by corticotropin-releasing factor. These effects are mediated by V_{1a} or V_{1b} receptors that are coupled to phospholipase C. Their role, if any, in human physiology/pathophysiology is uncertain.

METABOLISM

AVP distributes rapidly into a space roughly equal to the extracellular fluid volume. It is cleared irreversibly with a half-life ($t_{1/2}$) of 10–30 min. Most AVP clearance is due to degradation in the liver and kidneys. During pregnancy, the metabolic clearance of AVP is increased three- to fourfold due to placental production of an N-terminal peptidase.

THIRST

Because AVP cannot reduce water loss below a certain minimum level obligated by urinary solute load and evaporation from skin and lungs, a mechanism for ensuring adequate intake is essential for preventing

dehydration. This vital function is performed by the thirst mechanism. Like AVP, thirst is regulated primarily by an osmostat that is situated in the anteromedial hypothalamus and is able to detect very small changes in the plasma concentration of sodium and its anions. The thirst osmostat appears to be “set” about 3% higher than the AVP osmostat. This arrangement ensures that thirst, polydipsia, and dilution of body fluids do not occur until plasma osmolarity/sodium starts to exceed the defensive capacity of the antidiuretic mechanism.

OXYTOCIN

Oxytocin is also a nonapeptide that differs from AVP only at positions 3 and 8 (Fig. 374-1). However, it has relatively little antidiuretic effect and seems to act mainly on mammary ducts to facilitate milk letdown during nursing. It also may help initiate or facilitate labor by stimulating contraction of uterine smooth muscle, but it is not clear if this action is physiologic or necessary for normal delivery.

DEFICIENCIES OF AVP SECRETION AND ACTION

DIABETES INSIPIDUS

Clinical Characteristics

A decrease of 75% or more in the secretion or action of AVP usually results in DI, a syndrome characterized by the production of abnormally large volumes of dilute urine. The 24-h urine volume exceeds 40 mL/kg body weight, and the osmolarity is <300 mosmol/L. The polyuria produces symptoms of urinary frequency, enuresis, and/or nocturia, which may disturb sleep and cause mild daytime fatigue or somnolence. It also results in a slight rise in plasma osmolarity that stimulates thirst and a commensurate increase in fluid intake (polydipsia). Overt clinical signs of dehydration are uncommon unless thirst and/or the compensatory increase of fluid intake are also impaired.

Etiology



A primary deficiency of AVP secretion usually results from agenesis or irreversible destruction of the neurohypophysis. It is referred to variously as *neurohypophyseal DI*, *neurogenic DI*, *pituitary DI*, *cranial DI*, or *central DI*. It can be caused by a variety of congenital, acquired, or genetic disorders, but in about one-half of all adult patients, it is idiopathic **(Table 374-1)**. Pituitary DI caused by surgery in or around the neurohypophysis usually appears within 24 h. After a few days, it may transition to a 2- to 3-week period of inappropriate antidiuresis, after which the DI may or may not recur permanently. Five genetic forms of pituitary DI are now known. By far, the most common is transmitted in an autosomal dominant mode and is caused by diverse mutations in the coding region of one allele of the AVP–neurophysin II (or *AVP-NPII*) gene. All the mutations alter one or more amino acids known to be critical for correct processing and/or folding of the prohormone, thus interfering with its trafficking through the endoplasmic reticulum. The misfolded mutant precursor accumulates and interferes with production of AVP by the normal allele, eventually

destroying the magnocellular neurons in which it is produced. The AVP deficiency and DI are usually not present at birth but develop gradually over a period of several months to years, progressing from partial to severe at different rates depending on the mutation. Once established, the deficiency of AVP is permanent, but for unknown reasons, the DI occasionally improves or remits spontaneously in late middle age. The parvocellular neurons that make AVP and the magnocellular neurons that make oxytocin appear to be unaffected. There are also rare autosomal recessive forms of pituitary DI. One is due to an inactivating mutation in the AVP portion of the gene; another is due to a large deletion involving the majority of the AVP gene and regulatory sequences in the intergenic region. A third form is caused by mutations of the *WFS 1* gene responsible for Wolfram's syndrome (DI, diabetes mellitus, optic atrophy, and neural deafness [DIDMOAD]). An X-linked recessive form linked to a region on Xq28 has also been described but the causative gene has not yet been identified.

TABLE 374-1

Causes of Diabetes Insipidus

Pituitary diabetes insipidus

Acquired

- Head trauma (closed and penetrating) including pituitary surgery
- Neoplasms
  - Primary
    - Craniopharyngioma
    - Pituitary adenoma (suprasellar)
    - Dysgerminoma
    - Meningioma
  - Metastatic (lung, breast)
  - Hematologic (lymphoma, leukemia)
- Granulomas
  - Sarcoidosis
  - Histiocytosis
  - Xanthoma disseminatum
- Infectious
  - Chronic meningitis
  - Viral encephalitis
  - Toxoplasmosis
- Inflammatory
  - Lymphocytic infundibuloneurohypophysitis
  - Granulomatosis with polyangiitis (Wegener's)
  - Lupus erythematosus
  - Scleroderma
- Chemical toxins
  - Tetrodotoxin
  - Snake venom
- Vascular
  - Sheehan's syndrome
  - Aneurysm (internal carotid)
  - Aortocoronary bypass
  - Hypoxic encephalopathy
- Idiopathic

Congenital malformations

- Septo-optic dysplasia
- Midline craniofacial defects

Gestational diabetes insipidus

Pregnancy (second and third trimesters)

Nephrogenic diabetes insipidus

Acquired

- Drugs
  - Lithium
  - Demeclocycline
  - Methoxyflurane
  - Amphotericin B
  - Aminoglycosides
  - Cisplatin
  - Rifampin
  - Foscarnet
- Metabolic
  - Hypercalcemia, hypercalciuria
  - Hypokalemia
- Obstruction (ureter or urethra)
- Vascular
  - Sickle cell disease and trait
  - Ischemia (acute tubular necrosis)
- Granulomas
  - Sarcoidosis
- Neoplasms
  - Sarcoma
- Infiltration
  - Amyloidosis
- Idiopathic

Genetic

- X-linked recessive (*AVP receptor-2* gene)
- Autosomal recessive (*AQP2* gene)
- Autosomal dominant (*AQP2* gene)

Primary polydipsia

Acquired

- Psychogenic
  - Schizophrenia
  - Obsessive compulsive disorder

Holoprosencephaly	Dipsogenic (abnormal thirst)
Hypogenesis, ectopia of pituitary	Granulomas (sarcoidosis)
Genetic	Infectious (tuberculous meningitis)
Autosomal dominant (*AVP-neurophysin* gene)	Head trauma (closed and penetrating)
Autosomal recessive	Demyelination (multiple sclerosis)
Type A (*AVP-neurophysin* gene)	Drugs
Type B (*AVP-neurophysin* gene)	Idiopathic
Type C (*Wolfram's [4p-WFS 1]* gene)	Iatrogenic
X-linked recessive (Xq28)	

A primary deficiency of plasma AVP also can result from increased metabolism by an N-terminal aminopeptidase produced by the placenta. It is referred to as *gestational DI* because the signs and symptoms manifest during pregnancy and usually remit several weeks after delivery.

Secondary deficiencies of AVP secretion result from inhibition by excessive intake of fluids. They are referred to as *primary polydipsia* and can be divided into three subcategories. One of them, *dipsogenic DI*, is characterized by inappropriate thirst caused by a reduction in the set of the osmoregulatory mechanism. It sometimes occurs in association with multifocal diseases of the brain such as neurosarcoid, tuberculous meningitis, and multiple sclerosis but is often idiopathic. The second subtype, *psychogenic polydipsia*, is not associated with thirst, and the polydipsia seems to be a feature of psychosis or obsessive compulsive disorder. The third subtype, *iatrogenic polydipsia*, results from recommendations to increase fluid intake for its presumed health benefits.

Primary deficiencies in the antidiuretic action of AVP result in *nephrogenic DI*. The causes can be genetic, acquired, or drug induced (Table 374-1). The most common genetic form is transmitted in a semirecessive X-linked manner. It is caused by mutations in the coding region of the V_2 receptor gene that impair trafficking and/or ligand binding of the mutant receptor. There are also autosomal recessive or dominant forms of nephrogenic DI. They are caused by *AQP2* gene mutations that result in complete or partial defects in trafficking and function of the water channels that mediate antidiuresis in the distal and collecting tubules of the kidney.

Secondary deficiencies in the antidiuretic response to AVP result from polyuria per se. They are caused by washout of the medullary concentration gradient and/or suppression of aquaporin function. They usually resolve 24–48 h after the polyuria is corrected but can complicate interpretation of some acute tests used for differential diagnosis.

Pathophysiology

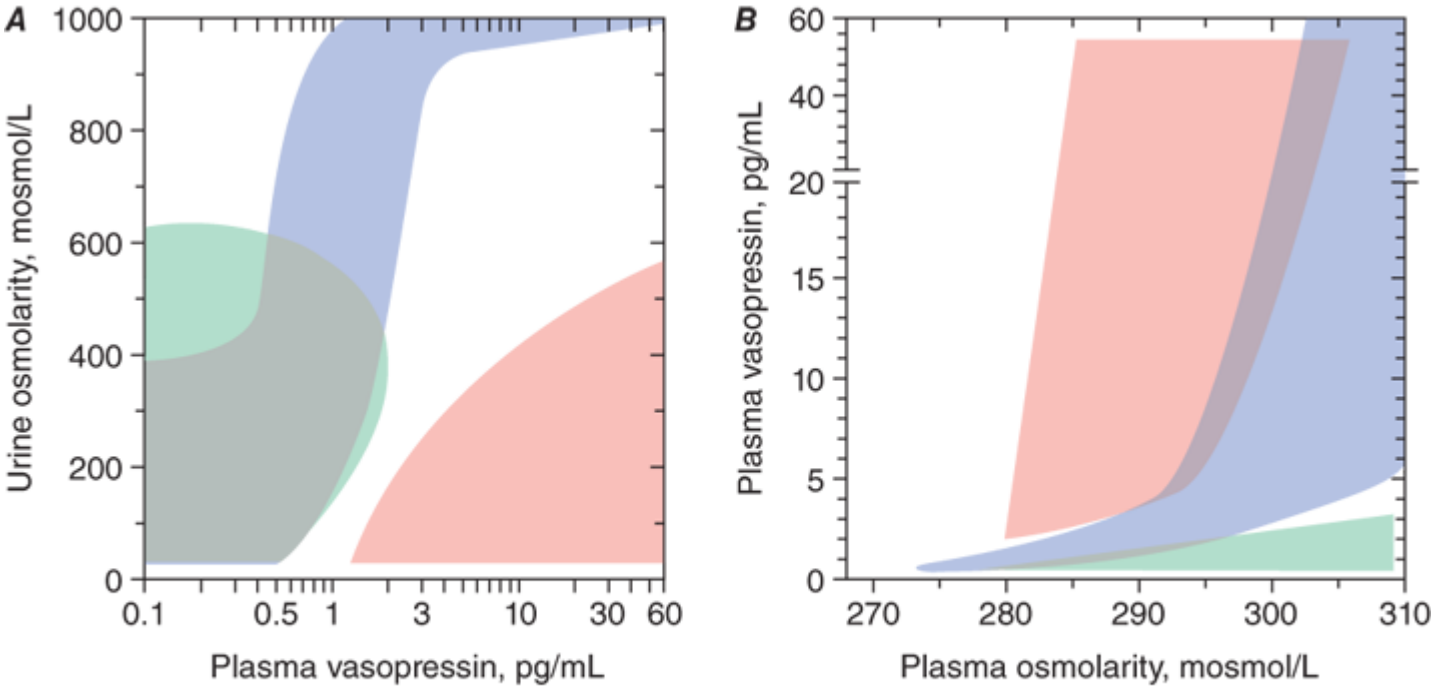
In pituitary, gestational, or nephrogenic DI, the polyuria results in a small (1–2%) decrease in body water and a commensurate increase in plasma osmolarity and sodium that stimulates thirst and a compensatory

increase in water intake. As a result, *hypernatremia and other overt physical or laboratory signs of dehydration do not develop unless the patient also has a defect in thirst or fails to increase fluid intake for some other reason*.

In pituitary and nephrogenic DI, the severity of the defect in AVP secretion or action varies significantly from patient to patient. In some, the defect is so severe that it cannot be overcome by even an intense stimulus such as nausea or severe dehydration. In others, the defect in AVP secretion or action is incomplete, and a modest stimulus such as a few hours of fluid deprivation, smoking, or a vasovagal reaction can raise urine osmolarity as high as 800 mosmol/L. However, even when the defects are partial, the relation of urine osmolarity to plasma AVP in patients with nephrogenic DI **(Fig. 374-3A)** or of plasma AVP to plasma osmolarity and sodium in patients with pituitary DI **(Fig. 374-3B)** is subnormal.

FIGURE 374-3

Relationship of plasma AVP to urine osmolarity (*A*) **and plasma osmolarity** (*B*) before and during fluid deprivation–hypertonic saline infusion test in patients who are normal or have primary polydipsia (blue zones), pituitary diabetes insipidus (green zones), or nephrogenic diabetes insipidus (pink zones).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In primary polydipsia, the pathogenesis of the polydipsia and polyuria is the reverse of that in pituitary, nephrogenic, and gestational DI. In primary polydipsia, an abnormality in cognition or thirst causes excessive intake of fluids and an increase in body water that reduces plasma osmolarity/sodium, AVP secretion, and urinary concentration. Dilution of the urine, in turn, results in a compensatory increase in urinary free-water excretion that usually offsets the increase in intake and stabilizes plasma osmolarity/sodium at a level only 1–2% below basal. Thus, hyponatremia or clinically appreciable overhydration is uncommon unless the polydipsia is very severe or the compensatory water diuresis is impaired by a drug or disease that stimulates or mimics the antiduretic effect of endogenous AVP. A rise in plasma osmolarity and sodium produced by

fluid deprivation or hypertonic saline infusion increases plasma AVP normally. However, the resultant increase in urine concentration is often subnormal because polyuria per se temporarily reduces the capacity of the kidney to concentrate the urine. Thus, the maximum level of urine osmolarity achieved during fluid deprivation is often indistinguishable from that in patients with partial pituitary or partial nephrogenic DI.

Differential Diagnosis

When symptoms of urinary frequency, enuresis, nocturia, and/or persistent thirst are present in the absence of glucosuria, the possibility of DI should be evaluated by collecting a 24-h urine on ad libitum fluid intake. If the osmolarity is <300 mosmol/L and the volume >50 mL/kg per day, the patient has DI and should be evaluated further to determine the type and select appropriate therapy. If the volume and osmolarity are not concordant, the possibility of inaccurate collection can be evaluated by determining if total urinary creatinine is normal for the size of the patient (20–30 mg/kg/day).

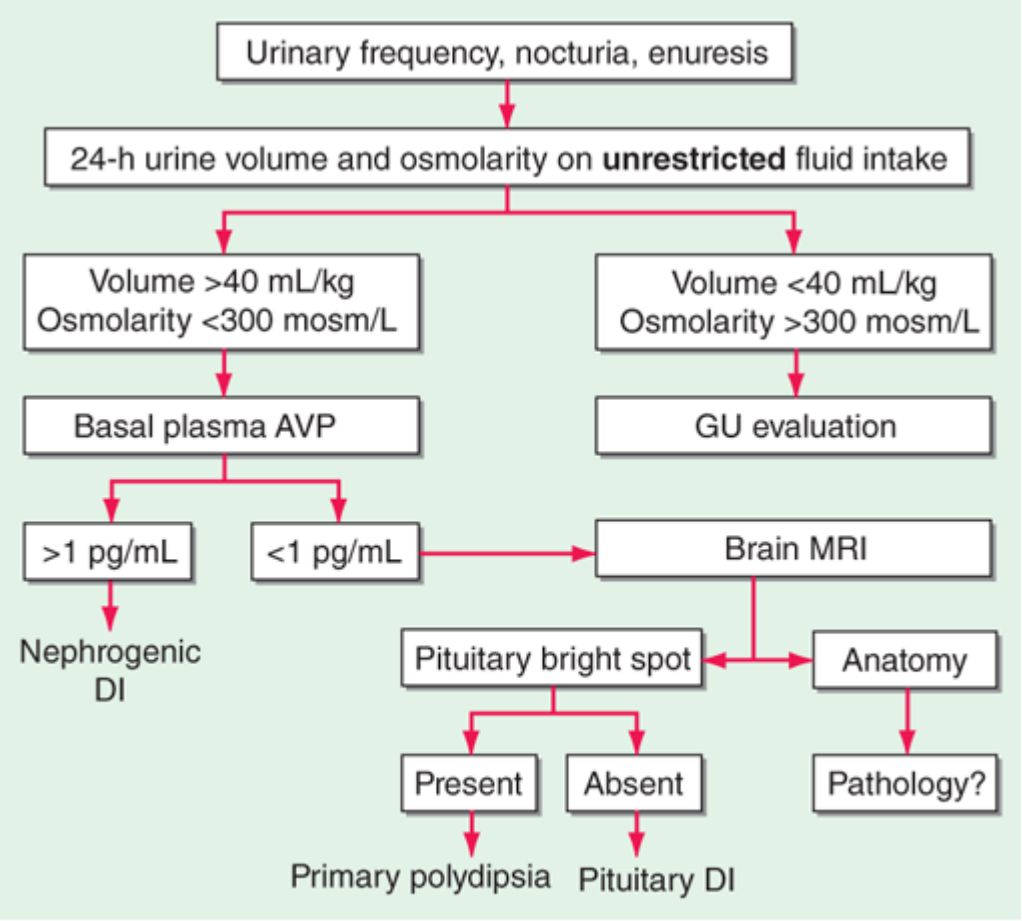
The type of DI can sometimes be inferred from the clinical setting or medical history. Often, however, such information is lacking, ambiguous, or misleading, and other approaches to differential diagnosis are needed. If basal plasma osmolarity and sodium are within normal limits, the traditional approach is to determine the effect of fluid deprivation and injection of antidiuretic hormone on urine osmolarity. This approach suffices for differential diagnosis *if* fluid deprivation raises plasma osmolarity and sodium above the normal range *without* inducing concentration of the urine. In that event, primary polydipsia and partial defects in AVP secretion and action are excluded, and the effect on urine osmolarity of injecting 2 μg of the AVP analogue, desmopressin indicates whether the patient has severe pituitary DI or severe nephrogenic DI. However, this approach is of little or no diagnostic value if fluid deprivation results in concentration of the urine because the increases in urine osmolarity achieved both before and after the injection of desmopressin are similar in patients with *partial* pituitary DI, *partial* nephrogenic DI, and primary polydipsia. These disorders can be differentiated by measuring plasma AVP during fluid deprivation and relating it to the concurrent level of plasma and urine osmolarity (Fig. 374-3). However, this approach does not always differentiate clearly between partial pituitary DI and primary polydipsia unless the measurement is made when plasma osmolarity and sodium are at or above the normal range. This level is difficult to achieve by fluid deprivation alone once urinary concentration occurs. Therefore, it is usually necessary to give a short infusion of 3% saline condition (0.1 mL/kg body weight per minute for 60–90 min) and repeat the measurement of plasma AVP. This approach is highly reliable for differential diagnosis but it is often stressful for the patient and requires special facilities and staff to perform safely and accurately.

A simpler, and less stressful, but equally reliable way to differentiate between pituitary DI, nephrogenic DI, and primary polydipsia is to start by measuring basal plasma AVP and urine osmolarity under conditions of unrestricted fluid intake **(Fig. 374-4)**. If AVP is normal or elevated (>1 pg/mL) and the concurrent urine osmolarity is low (<300 mosm/L), the patient has nephrogenic DI and the only additional evaluation required is to determine the cause. If, however, basal plasma AVP is low or undetectable (<1 pg/mL), nephrogenic DI is very unlikely and MRI of the brain can be performed to differentiate pituitary DI from primary polydipsia. In most healthy adults and children, the posterior pituitary emits a hyperintense signal visible in T1-weighted

midsagittal images. This “bright spot” is almost always present in patients with primary polydipsia but is always absent or abnormally small in patients with pituitary DI, even if their AVP deficiency is partial. The MRI is also useful in searching for pathology responsible for pituitary DI or the dipsogenic form of primary polydipsia (Fig. 374-2). The principal caveat is that MRI is not reliable for differential diagnosis of DI in patients with empty sella because they typically lack a bright spot even when their AVP secretion and action are normal. MRI also cannot be used to differentiate pituitary from nephrogenic DI because many patients with nephrogenic DI also lack a posterior pituitary bright spot, probably because they have an abnormally high rate of AVP secretion and turnover.

FIGURE 374-4

Simplified approach to the differential diagnosis of diabetes insipidus. When symptoms suggest diabetes insipidus (DI), the syndrome should be differentiated from a genitourinary (GU) abnormality by measuring the 24-h urine volume and osmolarity on unrestricted fluid intake. If DI is confirmed, basal plasma arginine vasopressin (AVP) should be measured on unrestricted fluid intake. If AVP is normal or elevated (>1 pg/mL), the patient probably has nephrogenic DI. However, if plasma AVP is low or undetectable, the patient has either pituitary DI or primary polydipsia. In that case, magnetic resonance imaging (MRI) of the brain can be performed to differentiate between these two conditions by determining whether or not the normal posterior pituitary bright spot is visible on T1-weighted midsagittal images. In addition, the MRI anatomy of the pituitary hypothalamic area can be examined to look for evidence of pathology that sometimes causes pituitary DI or the dipsogenic form of primary polydipsia. MRI is not reliable for differential diagnosis unless nephrogenic DI has been excluded because the bright spot is also absent, small, or faint in this condition.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

If MRI and/or AVP assays with the requisite sensitivity and specificity are unavailable and a fluid deprivation test is impractical or undesirable, a third way to differentiate between pituitary DI, nephrogenic DI, and primary polydipsia is a trial of desmopressin therapy. Such a trial should be conducted with very close monitoring of serum sodium as well as urine output, preferably in hospital, because desmopressin will produce hyponatremia in 8–24 h if the patient has primary polydipsia.

TREATMENT

TREATMENT

Diabetes Insipidus

The signs and symptoms of uncomplicated pituitary DI can be eliminated by treatment with desmopressin (DDAVP), a synthetic analogue of AVP (Fig. 374-1). DDAVP acts selectively at V2 receptors to increase urine concentration and decrease urine flow in a dose-dependent manner. It is also more resistant to degradation than is AVP and has a three- to fourfold longer duration of action. DDAVP can be given by IV or SC injection, nasal inhalation, or orally by means of a tablet or melt. The doses required to control pituitary DI vary widely, depending on the patient and the route of administration. However, among adults, they usually range from 1–2 μg qd or bid by injection, 10–20 μg bid or tid by nasal spray, or 100–400 μg bid or tid orally. The onset of antidiuresis is rapid, ranging from as little as 15 min after injection to 60 min after oral administration. When given in a dose that normalizes 24-h urinary osmolarity (400–800 mosmol/L) and volume (15–30 mL/kg body

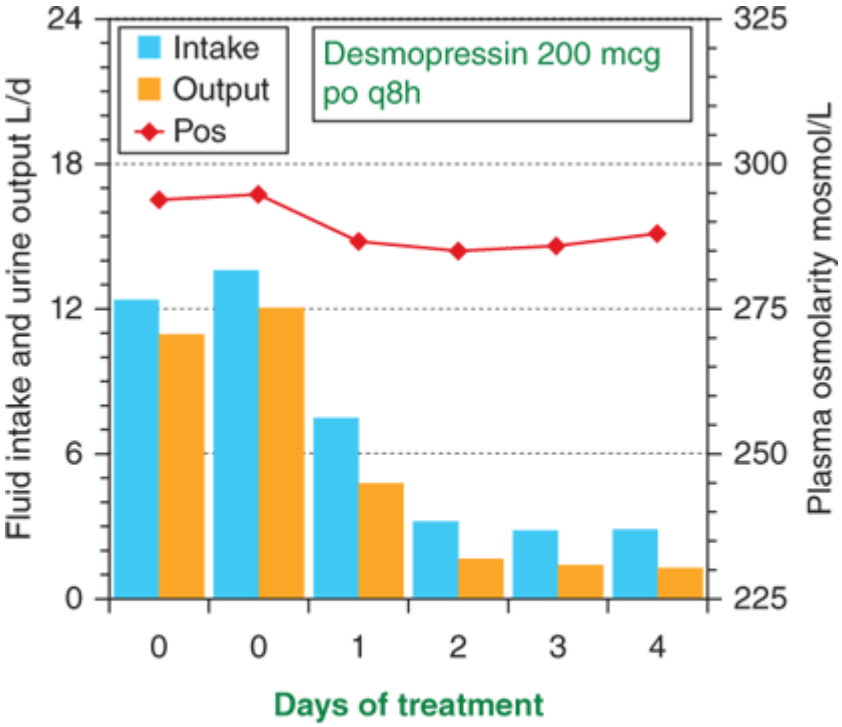
weight), DDAVP produces a slight (1–3%) increase in total body water and a decrease in plasma osmolarity/sodium that rapidly eliminates thirst and polydipsia **(Fig. 374-5)**. Consequently, water balance is maintained within the normal range. Hyponatremia rarely develops unless urine volume is reduced too far (to <10 mL/kg per day) or fluid intake is excessive due to an associated abnormality in thirst or cognition. Fortunately, thirst abnormalities are rare, and if the patient is taught to drink only when truly thirsty, DDAVP can be given safely in doses sufficient to normalize urine output without the need for allowing intermittent escape to prevent water intoxication.

Primary polydipsia cannot be treated safely with DDAVP or any other antidiuretic drug because eliminating the polyuria does not eliminate the urge to drink. Therefore, it invariably produces hyponatremia and/or other signs of water intoxication, usually within 8–24 h if urine output is normalized completely. There is no consistently effective way to correct dipsogenic or psychogenic polydipsia, but the iatrogenic form may respond to patient education. To minimize the risk of water intoxication, all patients should be warned about the use of other drugs such as thiazide diuretics or carbamazepine (Tegretol) that can impair urinary free-water excretion directly or indirectly.

The polyuria and polydipsia of nephrogenic DI are not affected by treatment with standard doses of DDAVP. If resistance is partial, it may be overcome by tenfold higher doses, but this treatment is too expensive and inconvenient for long-term use. However, treatment with conventional doses of a thiazide diuretic and/or amiloride in conjunction with a low-sodium diet and a prostaglandin synthesis inhibitor (e.g., indomethacin) usually reduces the polyuria and polydipsia by 30–70% and may eliminate them completely in some patients. Side effects such as hypokalemia and gastric irritation can be minimized by the use of amiloride or potassium supplements and by taking medications with meals.

FIGURE 374-5

Effect of desmopressin therapy on fluid intake (*blue bars*), urine output (*orange bars*), and plasma osmolarity (*red line*) in a patient with uncomplicated pituitary diabetes insipidus. Note that treatment rapidly reduces fluid intake and urine output to normal, with only a slight increase in body water as evidenced by the slight decrease in plasma osmolarity.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HYPODIPSIC HYPERNATREMIA

An increase in plasma osmolarity/sodium above the normal range (hypertonic hypernatremia) can be caused by either a decrease in total body water or an increase in total body sodium. The former results from a failure to drink enough to replace normal or increased urinary and insensible water loss. The deficient intake can be due either to water deprivation or a lack of thirst (hypodipsia). The most common cause of an increase in total body sodium is primary hyperaldosteronism (Chap. 379). Rarely, it can also result from ingestion of hypertonic saline in the form of sea water or incorrectly prepared infant formula. However, even in these forms of hypernatremia, inadequate intake of water also contributes. This chapter focuses on hypodipsic hypernatremia, the form of hypernatremia due to a primary defect in the thirst mechanism.

Clinical Characteristics

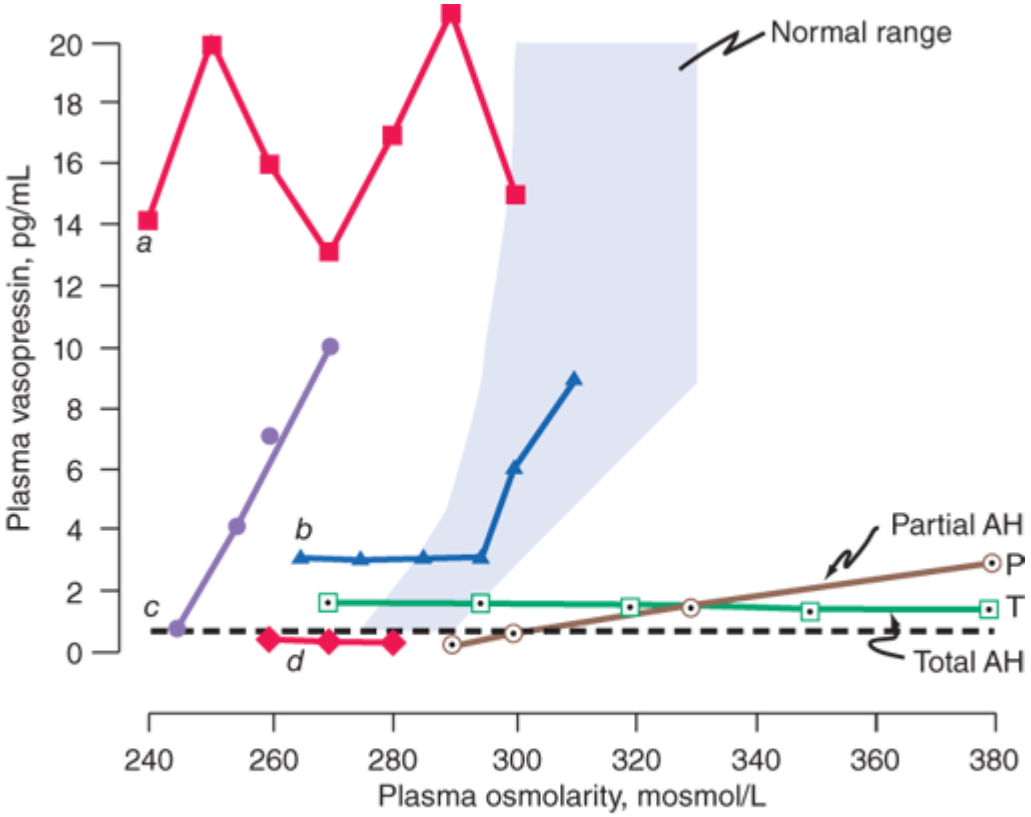
Hypodipsic hypernatremia is a syndrome characterized by chronic or recurrent hypertonic dehydration. The hypernatremia varies widely in severity and usually is associated with signs of hypovolemia such as tachycardia, postural hypotension, azotemia, hyperuricemia, and hypokalemia due to secondary hyperaldosteronism. Muscle weakness, pain, rhabdomyolysis, hyperglycemia, hyperlipidemia, and acute renal failure may also occur. Obtundation or coma may be present but are often absent. Despite inappropriately low levels of plasma AVP, DI usually is not evident at presentation but may develop during rehydration as blood volume, blood pressure, and plasma osmolarity/sodium return toward normal, further reducing plasma AVP.

Etiology

Hypodipsia is usually due to hypogenesis or destruction of the osmoreceptors in the anterior hypothalamus that regulate thirst. These defects can result from various congenital malformations of midline brain structures or may be acquired due to diseases such as occlusions of the anterior communicating artery, primary, or metastatic tumors in the hypothalamus, head trauma, surgery, granulomatous diseases such as sarcoidosis and histiocytosis, AIDS, and cytomegalovirus encephalitis. Because of their proximity, the osmoreceptors that regulate AVP secretion also are usually impaired. Thus, AVP secretion responds poorly or not at all to hyperosmotic stimulation **(Fig. 374-6)** but, in most cases, increases normally to nonosmotic stimuli such as nausea or large reductions in blood volume or blood pressure, indicating that the neurohypophysis is intact.

FIGURE 374-6

Heterogeneity of osmoregulatory dysfunction in adipsic hypernatremia (AH) and the syndrome of inappropriate antidiuresis (SIAD). Each line depicts schematically the relationship of plasma arginine vasopressin (AVP) to plasma osmolarity during water loading and/or infusion of 3% saline in a patient with either AH (*open symbols*) or SIAD (*closed symbols*). The shaded area indicates the normal range of the relationship. The horizontal broken line indicates the plasma AVP level below which the hormone is undetectable and urinary concentration usually does not occur. Lines P and T represent patients with a selective deficiency in the osmoregulation of thirst and AVP that is either partial ⊙ or total ⊡. In the latter, plasma AVP does not change in response to increases or decreases in plasma osmolarity but remains within a range sufficient to concentrate the urine even if overhydration produces hypotonic hyponatremia. In contrast, if the osmoregulatory deficiency is partial ⊙ , rehydration of the patient suppresses plasma AVP to levels that result in urinary dilution and polyuria before plasma osmolarity and sodium are reduced to normal. Lines *a*–*d* represent different defects in the osmoregulation of plasma AVP observed in patients with SIADH or SIAD. In *a* ■, plasma AVP is markedly elevated and fluctuates widely without relation to changes in plasma osmolarity, indicating complete loss of osmoregulation. In *b* ▲, plasma AVP remains fixed at a slightly elevated level until plasma osmolarity reaches the normal range, at which point it begins to rise appropriately, indicating a selective defect in the inhibitory component of the osmoregulatory mechanism. In *c* ● , plasma AVP rises in close correlation with plasma osmolarity before the latter reaches the normal range, indicating downward resetting of the osmostat. In *d* ◆ , plasma AVP appears to be osmoregulated normally, suggesting that the inappropriate antidiuresis is caused by some other abnormality.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pathophysiology

Hypodipsia results in a failure to drink enough water to replenish obligatory renal and extrarenal losses. Consequently, plasma osmolarity and sodium rise often to extremely high levels before the disorder is recognized. In most cases, urinary loss of water contributes little, if any, to the dehydration because AVP continues to be secreted in the small amounts necessary to concentrate the urine. In some patients this appears to be due to hypovolemic stimulation and/or incomplete destruction of AVP osmoreceptors because plasma AVP declines and DI develops during rehydration (Fig. 374-6). In others, however, plasma AVP does not decline during rehydration even if they are overhydrated. Consequently, they develop a hyponatremic syndrome indistinguishable from inappropriate antidiuresis. This suggests that the AVP osmoreceptors normally provide inhibitory and stimulatory input to the neurohypophysis and the patients can no longer osmotically stimulate or suppress tonic secretion of the hormone because both inputs have been totally eliminated by the same pathology that destroyed the osmoregulation of thirst. In a few patients, the neurohypophysis is also destroyed, resulting in a combination of chronic pituitary DI and hypodipsia that is particularly difficult to manage.

Differential Diagnosis

Hypodipsic hypernatremia usually can be distinguished from other causes of inadequate fluid intake (e.g., coma, paralysis, restraints, absence of fresh water) by the clinical history and setting. Previous episodes and/or denial of thirst and failure to drink spontaneously when the patient is conscious, unrestrained, and

hypernatremic are virtually diagnostic. The hypernatremia caused by excessive retention or intake of sodium can be distinguished by the presence of thirst as well as the physical and laboratory signs of hypervolemia rather than hypovolemia.

TREATMENT

TREATMENT

Hypodipsic Hypernatremia

Hypodipsic hypernatremia should be treated by administering water orally if the patient is alert and cooperative or by infusing hypotonic fluids (0.45% saline or 5% dextrose and water) if the patient is not. The amount of free water in liters required to correct the deficit (ΔFW) can be estimated from body weight in kg (BW) and the serum sodium concentration in mmol/L (S_{Na}) by the formula $\Delta FW = 0.5BW \times ([S_{Na} - 140]/140)$. If serum glucose ($S_{Glu}$) is elevated, the measured S_{Na} should be corrected (S_{Na}^*) by the formula $S_{Na}^* = S_{Na} + ([S_{Glu} - 90]/36)$. This amount plus an allowance for continuing insensible and urinary losses should be given over a 24- to 48-h period. Close monitoring of serum sodium as well as fluid intake and urinary output is essential because, depending on the extent of osmoreceptor deficiency, some patients will develop AVP-deficient DI, requiring DDAVP therapy to complete rehydration; others will develop hyponatremia and a syndrome of inappropriate antidiuresis (SIAD)-like picture if overhydrated. If hyperglycemia and/or hypokalemia are present, insulin and/or potassium supplements should be given with the expectation that both can be discontinued soon after rehydration is complete. Plasma urea/creatinine should be monitored closely for signs of acute renal failure caused by rhabdomyolysis, hypovolemia, and hypotension.

Once the patient has been rehydrated, an MRI of the brain and tests of anterior pituitary function should be performed to look for the cause and collateral defects in other hypothalamic functions. A long-term management plan to prevent or minimize recurrence of the fluid and electrolyte imbalance also should be developed. This should include a practical method to regulate fluid intake in accordance with variations in water balance as indicated by changes in body weight or serum sodium determined by home monitoring analyzers. Prescribing a constant fluid intake is ineffective and potentially dangerous because it does not take into account the large, uncontrolled variations in insensible loss that inevitably result from changes in ambient temperature and physical activity.

HYPONATREMIA DUE TO INAPPROPRIATE ANTIDIURESIS

A decrease in plasma osmolarity/sodium below the normal range (hypotonic hyponatremia) can be due to any of three different types of salt and water imbalance: (1) an increase in total body water that exceeds the increase in total body sodium (hypervolemic hyponatremia); (2) a decrease in body sodium greater than the decrease in body water (hypovolemic hyponatremia); or (3) an increase in body water with little or no change in body sodium (euvolemic hyponatremia) (Chap. 49). All three forms are associated with a failure to fully dilute the urine and mount a water diuresis in the face of hypotonic hyponatremia. However, the disorders with which they are associated and the types of salt and water imbalance that result differ. The hypervolemic

form typically occurs in disorders like severe congestive heart failure or cirrhosis in which water is retained in excessive of sodium. The hypovolemic form typically occurs in disorders such as severe diarrhea, diuretic abuse, or mineralocorticoid deficiency in which sodium is lost in excess of water. Euvolemic hyponatremia, however, is due mainly to expansion of total body water caused by excessive intake in the face of a failure to dilute the urine in response to excessive water intake. The impaired dilution is usually caused by a defect in the osmotic suppression of AVP that can have either of two causes. One is a nonhemodynamic stimulus such as nausea or a cortisol deficiency, which can be corrected quickly by treatment with antiemetics or cortisol. The other is a primary defect in osmoregulation caused by another disorder such as malignancy, stroke, or pneumonia that cannot be easily or quickly corrected. The latter is commonly known as the syndrome of inappropriate antidiuretic hormone (SIADH). Much less often, euvolemic hyponatremia can also result from AVP-independent activation of renal V_2 receptors, a variant known as nephrogenic inappropriate antidiuresis or NSIAD. Both of the latter will be discussed in this chapter.

Clinical Characteristics

Antidiuresis of any cause decreases the volume and increases the concentration of urine. If not accompanied by a commensurate reduction in fluid intake or an increase in insensible loss, the reduction in urine output results in excess water retention which expands and dilutes body fluids. If the hyponatremia develops gradually or has been present for more than a few days, it may be largely asymptomatic. However, if it develops acutely, it is usually accompanied by symptoms and signs of water intoxication that may include mild headache, confusion, anorexia, nausea, vomiting, coma, and convulsions. Severe acute hyponatremia may be lethal. Other clinical signs and symptoms vary greatly, depending on the type of hyponatremia. The hypervolemic form is characterized by generalized edema and other signs of marked volume expansion. The opposite is evident in the hypovolemic form. However, overt signs of volume expansion or contraction are absent in SIADH, SIAD, NSIAD, and other forms of euvolemic hyponatremia.

Etiology

In SIADH, the inappropriate secretion of AVP can have many different causes. They include ectopic production of AVP by lung cancer or other neoplasms; eutopic release induced by various diseases or drugs; and exogenous administration of AVP, DDAVP, or large doses of oxytocin **(Table 374-2)**. The ectopic forms result from abnormal expression of the *AVP-NPII* gene by primary or metastatic malignancies. The eutopic forms occur most often in patients with acute infections or strokes but have also been associated with many other neurologic diseases and injuries. The mechanisms by which these diseases interfere with osmotic suppression of AVP are not known. The defect in osmoregulation can take any of four distinct forms (Fig. 374-6). In one of the most common (reset osmostat), AVP secretion remains fully responsive to changes in plasma osmolarity/sodium, but the threshold, or set point, of the osmoregulatory system is abnormally low. These patients differ from those with the other types of SIADH in that they are able to maximally suppress plasma AVP and dilute their urine if their fluid intake is high enough to reduce their plasma osmolarity and/or sodium to the lower set point. In most patients, SIADH is self-limited and remits spontaneously within 2–3 weeks, but about 10% of cases are chronic. Another, smaller subgroup (~10% of the total) has inappropriate

antidiuresis without a demonstrable defect in the osmoregulation of plasma AVP (Fig. 374-6). In some of them, all young boys, the inappropriate antidiuresis has been traced to a constitutively activating mutation of the V_2 receptor gene. This unusual variant may be referred to as familial nephrogenic SIAD (NSIAD) to distinguish it from other possible causes of the syndrome. The inappropriate antidiuresis in these patients appears to be permanent, although the hyponatremia is variable owing presumably to individual differences in fluid intake.

TABLE 374-2

Causes of Syndrome of Inappropriate Antidiuretic Hormone (SIADH)

Neoplasms	Neurologic
Carcinomas	Guillain-Barré syndrome
Lung	Multiple sclerosis
Duodenum	Delirium tremens
Pancreas	Amyotrophic lateral sclerosis
Ovary	Hydrocephalus
Bladder, ureter	Psychosis
Other neoplasms	Peripheral neuropathy
Thymoma	Congenital malformations
Mesothelioma	Agenesis corpus callosum
Bronchial adenoma	Cleft lip/palate
Carcinoid	Other midline defects
Gangliocytoma	Metabolic
Ewing's sarcoma	Acute intermittent porphyria
Head trauma (closed and penetrating)	Pulmonary
Infections	Asthma
Pneumonia, bacterial or viral	Pneumothorax
Abscess, lung or brain	Positive-pressure respiration
Cavitation (aspergillosis)	Drugs
Tuberculosis, lung or brain	Vasopressin or desmopressin
Meningitis, bacterial or viral	Serotonin reuptake inhibitors
Encephalitis	Oxytocin, high dose
AIDS	Vincristine
Vascular	Carbamazepine
Cerebrovascular occlusions, hemorrhage	Nicotine
Cavernous sinus thrombosis	Phenothiazines
	Cyclophosphamide
	Tricyclic antidepressants
	Monoamine oxidase inhibitors

Pathophysiology

Impaired osmotic suppression of antidiuresis results in excessive retention of water and dilution of body fluids only if water intake exceeds insensible and urinary losses. The excess intake is sometimes due to an associated defect in the osmoregulation of thirst (dipsogenic) but can also be psychogenic or iatrogenic,

including excessive IV administration of hypotonic fluids. In SIADH and other forms of euvolemic hyponatremia, the decrease in plasma osmolarity/sodium and the increase in extracellular and intracellular volume are proportional to the amount of water retained. Thus, an increase in body water of 10% (~4 L in a 70-kg adult) reduces plasma osmolarity and sodium by ~10% (~28 mosmol/L or 14 meq/L). An increase in body water of this magnitude is rarely detectable on physical examination but will be reflected in a weight gain of about 4 kg. It also increases glomerular filtration and atrial natriuretic hormone and suppresses plasma renin activity, thereby increasing urinary sodium excretion. The resultant reduction in total body sodium decreases the expansion of extracellular volume but aggravates the hyponatremia and further expands intracellular volume. The latter increases brain swelling and intracranial pressure, which probably produces most of the symptoms of acute water intoxication. Within a few days, this swelling may be counteracted by inactivation or elimination of intracellular solutes, resulting in the remission of symptoms even though the hyponatremia persists.

In type I (hypervolemic) or type II (hypovolemic) hyponatremia, osmotic suppression of AVP secretion appears to be counteracted by a hemodynamic stimulus resulting from a large reduction in cardiac output and/or effective blood volume. The resultant antidiuresis is enhanced by decreased distal delivery of glomerular filtrate that results from increased reabsorption of sodium in proximal nephron. If the reduction in urine output is not associated with a commensurate reduction in water intake or an increase in insensible loss, body fluids are expanded and diluted, resulting in hyponatremia despite an increase in body sodium. Unlike SIADH and other forms of euvolemic hyponatremia, however, glomerular filtration is reduced and plasma renin activity and aldosterone are elevated. Thus, the rate of urinary sodium excretion is low (unless sodium reabsorption is impaired by a diuretic), and the hyponatremia is usually accompanied by edema, hypokalemia, azotemia, and hyperuricemia. In type II (hypovolemic) hyponatremia, sodium and water are also retained as an appropriate compensatory response to the severe depletion.

Differential Diagnosis

SIADH is a diagnosis of exclusion that usually can be made from the history, physical examination, and basic laboratory data. If hyperglycemia is present, its contribution to the reduction in plasma sodium can be estimated either by measuring plasma osmolarity for a more accurate estimate of the true “effective” tonicity of body fluids or by correcting the measured plasma sodium for the reduction caused by the hyperglycemia using the simplified formula

$$\text{corrected } P_{na} = \text{measured } P_{na} + (P_{glu} - 90)/36$$

where P_{na} = plasma sodium in meq/L and P_{glu} = plasma glucose in mg/dL.

If the plasma osmolarity and/or corrected plasma sodium are below normal limits, hypotonic hyponatremia is present and further evaluation to determine the type should be undertaken in order to administer safe and effective treatment. This differentiation is usually possible by evaluating standard clinical indicators of the extracellular fluid volume **(Table 374-3)**. If these findings are ambiguous or contradictory, measuring plasma

renin activity or the *rate* of urinary sodium excretion may be helpful *provided* that the hyponatremia is not in the recovery phase or is due to a primary defect in renal conservation of sodium, diuretic abuse, or hyporeninemic hypoaldosteronism. The latter may be suspected if serum potassium is elevated instead of low, as it usually is in types I and II hyponatremia. Measurements of plasma AVP are currently of no value in differentiating SIADH from the other types of hyponatremia since the plasma levels are elevated similarly in all. In patients who fulfill the clinical criteria for type III (euvolemic) hyponatremia, morning plasma cortisol should also be measured to exclude secondary adrenal insufficiency. If it is normal and there is no history of nausea/vomiting, the diagnosis of SIADH is confirmed, and a careful search for occult lung cancer or other common causes of the syndrome (Table 374-2) should be undertaken.

TABLE 374-3

Differential Diagnosis of Hyponatremia Based on Clinical Assessment of Extracellular Fluid Volume (ECFV)

CLINICAL FINDINGS	TYPE I, HYPERVOLEMIC	TYPE II, HYPOVOLEMIC	TYPE III, EUVOLEMIC	SIADH AND SIAD EUVOLEMIC
History				
CHF, cirrhosis, or nephrosis	Yes	No	No	No
Salt and water loss	No	Yes	No	No
ACTH–cortisol deficiency and/or nausea and vomiting	No	No	Yes	No
Physical examination				
Generalized edema, ascites	Yes	No	No	No
Postural hypotension	Maybe	Maybe	Maybe[a]	No
Laboratory				
BUN, creatinine	High-normal	High-normal	Low-normal	Low-normal
Uric acid	High-normal	High-normal	Low-normal	Low-normal
Serum potassium	Low-normal	Low-normal[b]	Normal[c]	Normal
Serum urate	High	High	Low	Low
Serum albumin	Low-normal	High-normal	Normal	Normal
Serum cortisol	Normal-high	Normal-high[d]	Low	Normal
Plasma renin activity	High	High	Normal	Low
Urinary sodium (meq per unit of time)[g]	Low	Low[h]	Low[e] Low[f] High[i]	High[i]

[a]Postural hypotension may occur in secondary (ACTH-dependent) adrenal insufficiency even though extracellular fluid volume and aldosterone are usually normal. [b]Serum potassium may be high if hypovolemia is due to aldosterone deficiency. [c]Serum potassium may be low if vomiting causes alkalosis. [d]Serum cortisol is low if hypovolemia is due to primary adrenal insufficiency (Addison's disease). [e]Serum cortisol will be normal or high if the cause is nausea and vomiting rather than secondary (ACTH-dependent) adrenal insufficiency. [f]Plasma renin activity may be high if the cause is secondary (ACTH) adrenal insufficiency. [g]Urinary sodium should be expressed as the *rate of excretion* rather than the concentration. In a hyponatremic adult, an excretion rate >25 meq/d (or 25 μeq/mg of creatinine) could be considered high. [h]The rate of urinary sodium excretion may be high if the hypovolemia is due to diuretic abuse,

primary adrenal insufficiency, or other causes of renal sodium wasting. [i]The rate of urinary sodium excretion may be low if intake is curtailed by symptoms or treatment.

Abbreviations: ACTH, adrenocorticotropic hormone; BUN, blood urea nitrogen; CHF, congestive heart failure; SIAD, syndrome of inappropriate antidiuresis.

SIAD due to an activating mutation of the V_2 receptor gene should be suspected if the hyponatremia occurs in a child or several members of the family or is refractory to treatment with a vaptan (see below). In that case, plasma AVP should be measured to confirm that it is appropriately suppressed while the hyponatremia and antidiuresis are present, and the V_2 receptor gene should be sequenced, if possible.

TREATMENT

TREATMENT

Hyponatremia

The management of hyponatremia differs depending on the type and the severity and duration of symptoms. In acute symptomatic SIADH, the aim should be to raise plasma osmolarity and/or plasma sodium at a rate ~1% an hour until they reach levels of ~270 mosmol/L or 130 meq/L, respectively. This can be accomplished in either of two ways. One is to infuse hypertonic (3%) saline at a rate of about 0.05 mL/kg body weight per min. This treatment often produces a solute diuresis that serves to remove some of the excess water. The other treatment for acute, symptomatic SIADH is to reduce body water by giving an AVP receptor-2 antagonist (vaptan) to block the antidiuretic effect of AVP and increase urine output **(Fig. 374-7)**. One of the vaptans, a combined V_2/V_{1a} antagonist (Conivaptan), has been approved for short-term, in-hospital IV treatment of SIADH. It should be given as a loading dose of 20 mg IV over 30 min followed by a continuous infusion of 20 mg over 24 h. Another vaptan (Tolvaptan) can be given orally starting at a dose of 15 mg po and increasing to 30 mg or 60 mg at 24 h intervals depending on the effect. With either approach, fluid intake should be restricted to less than urine output. Because the aquaretic effect of the vaptans varies in magnitude from patient to patient, the rate of rise in serum sodium also varies if fluid intake is fixed at a constant rate. This variability in effect can be reduced or eliminated by continuously monitoring the rate of urine output and adjusting the rate of IV or oral fluid intake so as to reduce body water at a constant rate. Regulating fluid intake so that it under replaces urine output by 5mL/kg body weight/h will raise serum sodium at a rate of about 1% an hour. In any event, serum sodium should be checked every 2–4 h to ensure it is not raised faster than 1mEq/L per hour or above the lower limit of normal. Doing so may result in central pontine myelinolysis, an acute, potentially fatal neurologic syndrome characterized by quadriparesis, ataxia, and abnormal extraocular movements.

In chronic and/or minimally symptomatic SIADH, the hyponatremia can and should be corrected more gradually. This can be achieved by restricting total fluid intake to less than the sum of urinary and insensible losses. Because the water derived from food (300–700 mL/d) usually approximates basal insensible losses in adults, the aim should be to reduce total discretionary intake (all liquids) to ~500 mL less than urinary

output. Adherence to this regimen is often problematic and, even if achieved, usually reduces body water and increases serum sodium by only about 1–2% per day. Therefore, it is often necessary to add a treatment that increases urinary water excretion. The oral AVP2 antagonist, tolvaptan, is best suited for this purpose. The best approach for treatment of chronic SIADH is the administration of an oral vaptan, tolvaptan, a selective V_2 antagonist that also increases urinary water excretion by blocking the antidiuretic effect of AVP.

Some restriction of fluid intake may also be necessary to achieve satisfactory control of the hyponatremia. It is approved for treatment of nonemergent SIADH with initial in-hospital dosing. Other approaches include demeclocycline, 150–300 mg PO tid or qid, which induces a reversible form of nephrogenic DI in 1–2 weeks, or fludrocortisone, 0.05–0.2 mg PO bid. The effect of the demeclocycline manifests in 7–14 days and is due to induction of a reversible form of nephrogenic DI. Fludrocortisone, 0.05–0.2 mg po bid, also raises serum sodium gradually over 1–2 weeks. Its mechanism of action is unclear but probably involves increased retention of sodium. It also increases urinary potassium excretion, which may require replacement through dietary adjustments or supplements and may induce hypertension, occasionally necessitating discontinuation of the treatment.

In the type of euvolemic hyponatremia caused by protracted nausea and vomiting or isolated glucocorticoid deficiency (type III), all abnormalities can be corrected quickly and completely by giving an antiemetic or stress doses of hydrocortisone (for glucocorticoid deficiency). As with other treatments, care must be taken to ensure that serum sodium does not rise too quickly or too far.

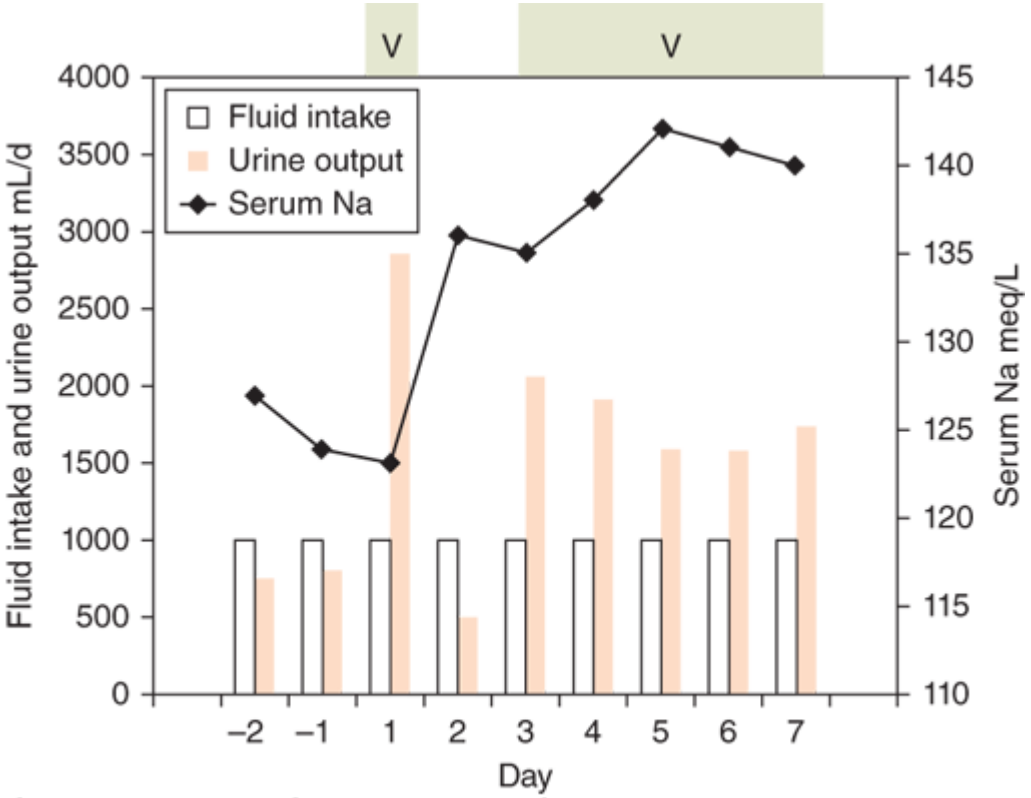
In SIAD due to an activating mutation of the V_2 receptor, the V_2 antagonists may not block the antidiuresis or raise plasma osmolarity/sodium. In that condition, use of an osmotic diuretic such as urea is reported to be effective in preventing or correcting hyponatremia. However, some vaptans may be effective in patients with a different type of activating mutation so the response to this therapy may be neither predictable nor diagnostic.

In hypervolemic hyponatremia, fluid restriction is also appropriate and somewhat effective if it can be maintained. The infusion of hypertonic saline is contraindicated because it further increases total body sodium and edema and may precipitate cardiovascular decompensation. However, as in SIADH, the V_2 receptor antagonists are also safe and effective in the treatment of hypervolemic hyponatremia caused by congestive heart failure. Tolvaptan is approved by the Food and Drug Administration for this indication with the caveat that treatment should be initiated or reinitiated in hospital. Its use should also be limited to 30 days at a time because of reports that longer periods may be associated with abnormal liver chemistries.

In hypovolemic hyponatremia, the imbalance can be corrected easily and quickly by stopping the loss of sodium and water and/or replacing the deficits by mouth or IV infusion of normal or hypertonic saline. As with the treatment of other forms of hyponatremia, care must be taken to ensure that plasma sodium does not increase too rapidly or too far. Fluid restriction and administration of AVP antagonists are contraindicated in type II hyponatremia because they would only aggravate the underlying volume depletion and could result in hemodynamic collapse.

FIGURE 374-7

The effect of vaptan therapy on water balance in a patient with chronic syndrome of inappropriate antidiuretic hormone (SIADH). The periods of vaptan (V) therapy are indicated by the green shaded boxes at the top. Urine output is indicated by orange bars. Fluid intake is shown by the open bars. Intake was restricted to 1 L/d throughout. Serum sodium is indicated by the black line. Note that sodium increased progressively when vaptan increased urine output to levels that clearly exceeded fluid intake.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GLOBAL PERSPECTIVES

The incidence, clinical characteristics, etiology, pathophysiology, differential diagnosis, and treatments of fluid and electrolyte disorders in tropical and nonindustrialized countries differ in some respects from those in the United States and other industrialized parts of the world. Hyponatremia, for example, appears to be more common and is more likely to be due to infectious diseases such as cholera, shigellosis, and other diarrheal disorders. In these circumstances, hyponatremia is probably due to gastrointestinal losses of salt and water (hypovolemia type II), but other abnormalities, including undefined infectious toxins, also may contribute. The causes of DI are similar worldwide except that malaria and venoms from snake or insect bites are much more common in some tropical climates.

FURTHER READING

Robertson GL: Vaptans for the treatment of hyponatremia. Nature Rev Endocrin 7:151, 2011.
CrossRef

Robertson GL: Diabetes insipidus: Differential diagnosis and management. Best Prac Res Clin Endocrin Metabol 30:205, 2016.
CrossRef

Chapter 375: Thyroid Gland Physiology and Testing

J. Larry Jameson; Susan J. Mandel; Anthony P. Weetman

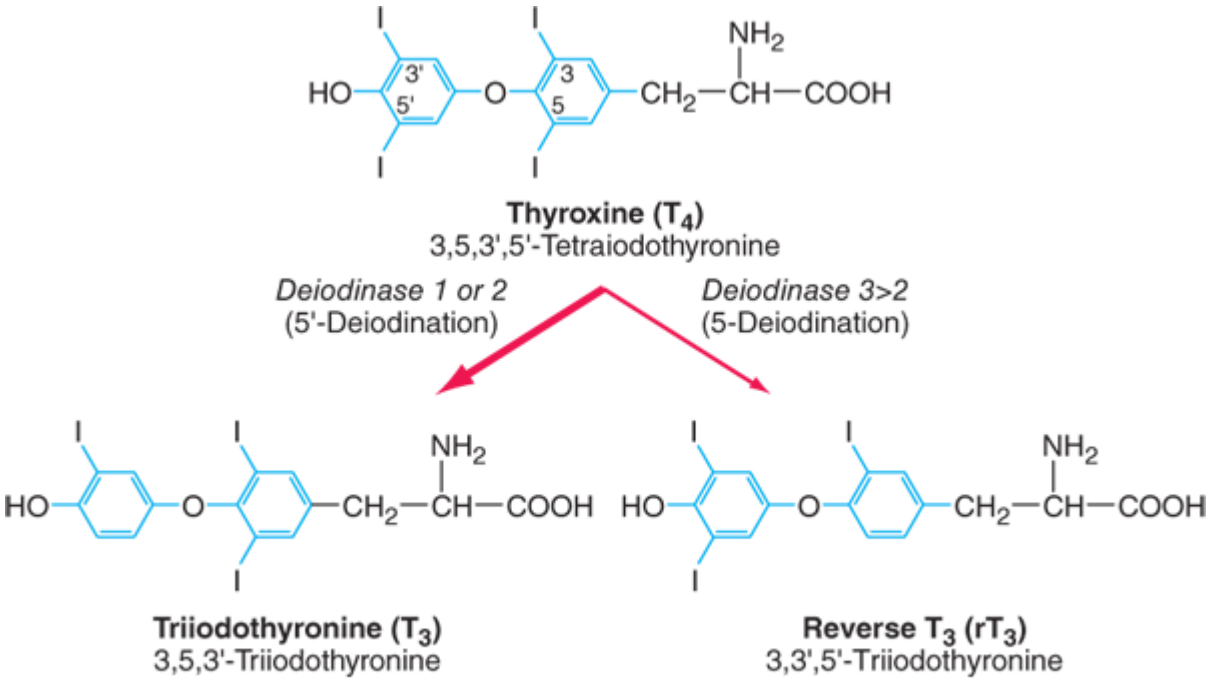
INTRODUCTION

The thyroid gland produces two related hormones, thyroxine (T_4) and triiodothyronine (T_3) **(Fig. 375-1)**.

Acting through thyroid hormone receptors α and β, these hormones play a critical role in cell differentiation and organogenesis during development and help maintain thermogenic and metabolic homeostasis in the adult. Autoimmune disorders of the thyroid gland can stimulate overproduction of thyroid hormones (*thyrotoxicosis*) or cause glandular destruction and hormone deficiency (*hypothyroidism*). Benign nodules and various forms of thyroid cancer are relatively common and amenable to detection by physical examination or various imaging techniques.

FIGURE 375-1

Structures of thyroid hormones. Thyroxine (T_4) contains four iodine atoms. Deiodination leads to production of the potent hormone triiodothyronine (T_3) or the inactive hormone reverse T_3.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ANATOMY AND DEVELOPMENT

The thyroid (Greek *thyreos*, shield, plus *eidos*, form) consists of two lobes connected by an isthmus. It is located anterior to the trachea between the cricoid cartilage and the suprasternal notch. The normal thyroid is 12–20 g in size, highly vascular, and soft in consistency. Four parathyroid glands, which produce parathyroid hormone **(Chap. 403)**, are located posterior to each pole of the thyroid. The recurrent laryngeal nerves traverse the lateral borders of the thyroid gland and must be identified during thyroid surgery to avoid injury and vocal cord paralysis.

The thyroid gland develops from the floor of the primitive pharynx during the third week of gestation. The developing gland migrates along the thyroglossal duct to reach its final location in the neck. This feature accounts for the rare ectopic location of thyroid tissue at the base of the tongue (lingual thyroid) as well as the occurrence of thyroglossal duct cysts along this developmental tract. Thyroid hormone synthesis normally begins at about 11 weeks' gestation.

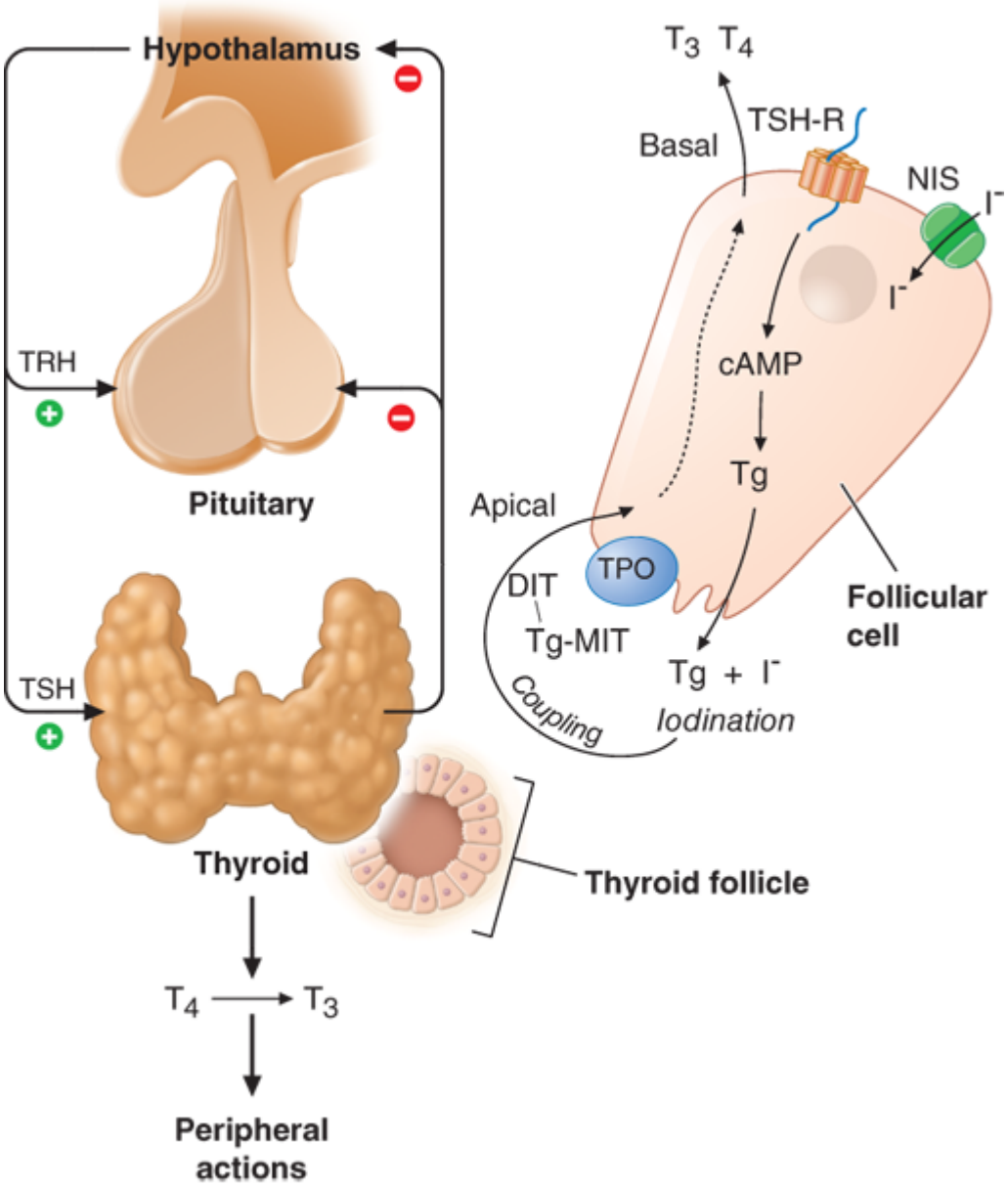
Neural crest derivatives from the ultimobranchial body give rise to thyroid medullary C cells that produce calcitonin, a calcium-lowering hormone. The C cells are interspersed throughout the thyroid gland, although their density is greatest in the juncture of the upper one-third and lower two-thirds of the gland. Calcitonin plays a minimal role in calcium homeostasis in humans but the C-cells are important because of their involvement in medullary thyroid cancer.

Thyroid gland development is orchestrated by the coordinated expression of several developmental transcription factors. Thyroid transcription factor (TTF)-1, TTF-2, NKX2-1, and paired homeobox-8 (PAX-8) are expressed selectively, but not exclusively, in the thyroid gland. In combination, they dictate thyroid cell development and the induction of thyroid-specific genes such as thyroglobulin (Tg), thyroid peroxidase (TPO), the sodium iodide symporter (Na^+/I^-, NIS), and the thyroid-stimulating hormone receptor (TSH-R). Mutations in these developmental transcription factors or their downstream target genes are rare causes of thyroid agenesis or dyshormonogenesis, although the causes of most forms of congenital hypothyroidism remain unknown **(see Chap. 376, Table 376-1)**. Because congenital hypothyroidism occurs in ~1 in 4000 newborns, neonatal screening is now performed in most industrialized countries. Transplacental passage of maternal thyroid hormone occurs before the fetal thyroid gland begins to function and provides significant hormone support to a fetus with congenital hypothyroidism. Early thyroid hormone replacement in newborns with congenital hypothyroidism prevents potentially severe developmental abnormalities.

The thyroid gland consists of numerous spherical follicles composed of thyroid follicular cells that surround secreted colloid, a proteinaceous fluid containing large amounts of thyroglobulin, the protein precursor of thyroid hormones **(Fig. 375-2)**. The thyroid follicular cells are polarized—the basolateral surface is apposed to the bloodstream and an apical surface faces the follicular lumen. Increased demand for thyroid hormone is regulated by TSH, which binds to its receptor on the basolateral surface of the follicular cells. This binding leads to Tg reabsorption from the follicular lumen and proteolysis within the cytoplasm, yielding thyroid hormones for secretion into the bloodstream.

FIGURE 375-2

Regulation of thyroid hormone synthesis. ***Left.*** Thyroid hormones T_4 and T_3 feedback to inhibit hypothalamic production of thyrotropin-releasing hormone (TRH) and pituitary production of thyroid-stimulating hormone (TSH). TSH stimulates thyroid gland production of T_4 and T_3. ***Right.*** Thyroid follicles are formed by thyroid epithelial cells surrounding proteinaceous colloid, which contains thyroglobulin. Follicular cells, which are polarized, synthesize thyroglobulin and carry out thyroid hormone biosynthesis (see text for details). DIT, diiodotyrosine; MIT, monoiodotyrosine; NIS, sodium iodide symporter; Tg, thyroglobulin; TPO, thyroid peroxidase; TSH-R, thyroid-stimulating hormone receptor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

REGULATION OF THE THYROID AXIS

TSH, secreted by the thyrotrope cells of the anterior pituitary, plays a pivotal role in control of the thyroid axis and serves as the most useful physiologic marker of thyroid hormone action. TSH is a 31-kDa hormone composed of α and β subunits; the α subunit is common to the other glycoprotein hormones (luteinizing hormone, follicle-stimulating hormone, human chorionic gonadotropin [hCG]), whereas the TSH β subunit is unique to TSH. The extent and nature of carbohydrate modification are modulated by thyrotropin-releasing hormone (TRH) stimulation and influence the biologic activity of the hormone.

The thyroid axis is a classic example of an endocrine feedback loop (Chap. 370). Hypothalamic TRH stimulates pituitary production of TSH, which, in turn, stimulates thyroid hormone synthesis and secretion. Thyroid hormones act via negative feedback predominantly through thyroid hormone receptor β2 (TRβ2) to inhibit TRH and TSH production (Fig. 375-2). The "set-point" in this axis is established by TSH. TRH is the major positive regulator of TSH synthesis and secretion. Peak TSH secretion occurs ~15 min after administration of exogenous TRH. Dopamine, glucocorticoids, and somatostatin suppress TSH but are not of major physiologic importance except when these agents are administered in pharmacologic doses. Reduced levels of thyroid hormone increase basal TSH production and enhance TRH-mediated stimulation of TSH. High thyroid hormone levels rapidly and directly suppress TSH gene expression secretion and inhibit TRH stimulation of TSH, indicating that thyroid hormones are the dominant regulator of TSH production. Like other pituitary hormones, TSH is released in a pulsatile manner and exhibits a diurnal rhythm; its highest levels occur at night. However, these TSH excursions are modest in comparison to those of other pituitary hormones, in part, because TSH has a relatively long plasma half-life (50 min). Consequently, single measurements of TSH are adequate for assessing its circulating level. TSH is measured using immunoradiometric assays that are highly sensitive and specific. These assays readily distinguish between normal and suppressed TSH values; thus, TSH can be used for the diagnosis of primary hyperthyroidism (low TSH) or primary hypothyroidism (high TSH).

THYROID HORMONE SYNTHESIS, METABOLISM, AND ACTION

THYROID HORMONE SYNTHESIS

Thyroid hormones are derived from Tg, a large iodinated glycoprotein. After secretion into the thyroid follicle, Tg is iodinated on tyrosine residues that are subsequently coupled via an ether linkage. Reuptake of Tg into the thyroid follicular cell allows proteolysis and the release of newly synthesized T_4 and T_3.

Iodine Metabolism and Transport

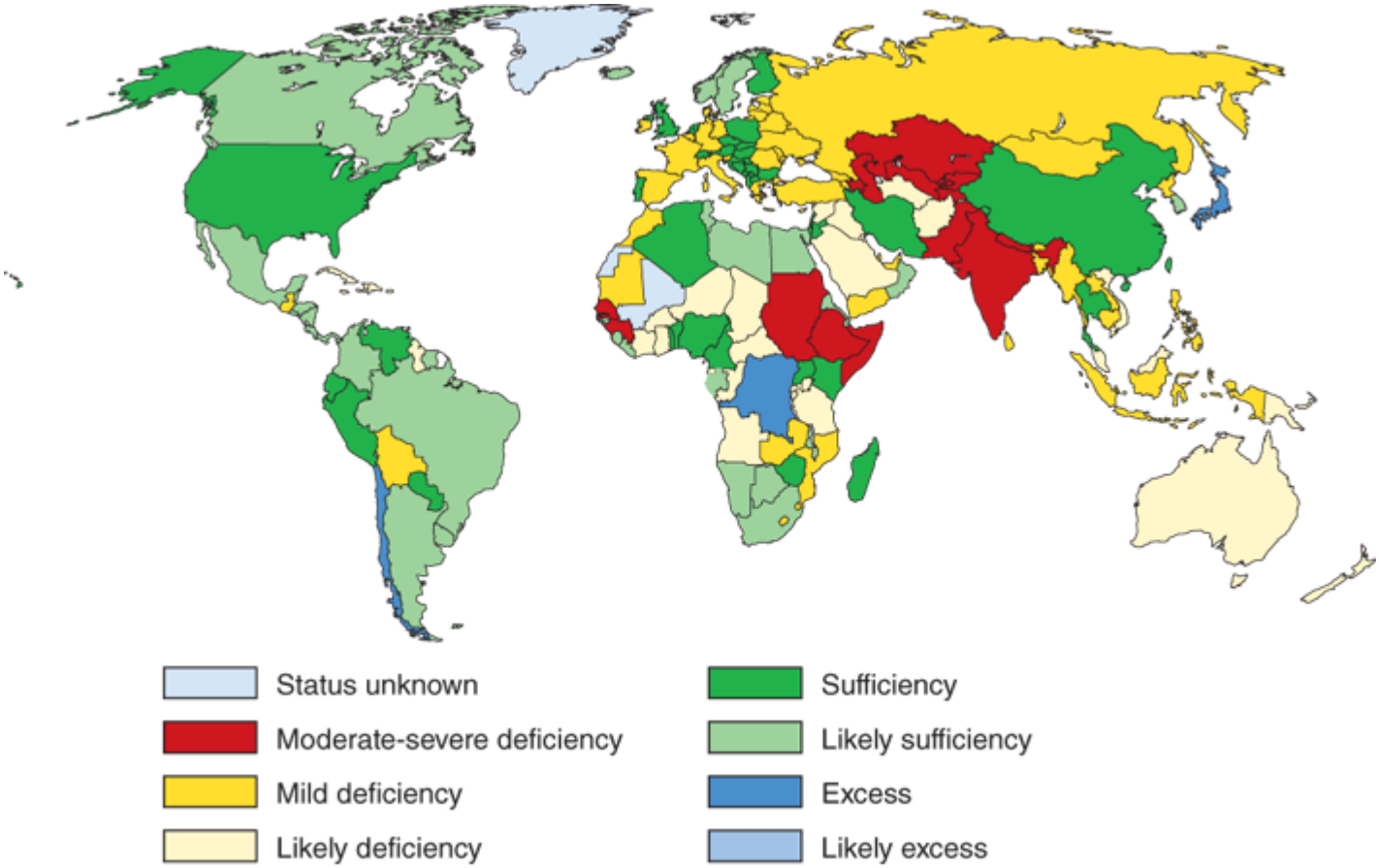
Iodide uptake is a critical first step in thyroid hormone synthesis. Ingested iodine is bound to serum proteins, particularly albumin. Unbound iodine is excreted in the urine. The thyroid gland extracts iodine from the circulation in a highly efficient manner. For example, 10–25% of radioactive tracer (e.g., ^{123}I) is taken up by the normal thyroid gland over 24 h in an iodine-replete state; this value can rise to 70–90% in Graves' disease. Iodide uptake is mediated by NIS, which is expressed at the basolateral membrane of thyroid follicular cells. NIS is most highly expressed in the thyroid gland, but low levels are present in the salivary glands, lactating

breast, and placenta. The iodide transport mechanism is highly regulated, allowing adaptation to variations in dietary supply. Low iodine levels increase the amount of NIS and stimulate uptake, whereas high iodine levels suppress NIS expression and uptake. The selective expression of NIS in the thyroid allows isotopic scanning, treatment of hyperthyroidism, and ablation of thyroid cancer with radioisotopes of iodine, without significant effects on other organs. Mutation of the *NIS* gene is a rare cause of congenital hypothyroidism, underscoring its importance in thyroid hormone synthesis. Another iodine transporter, pendrin, is located on the apical surface of thyroid cells and mediates iodine efflux into the lumen. Mutation of the *pendrin* gene causes *Pendred syndrome*, a disorder characterized by defective organification of iodine, goiter, and sensorineural deafness.

Iodine deficiency is prevalent in many mountainous regions and in central Africa, central South America, and northern Asia **(Fig. 375-3)**. Europe remains mildly iodine-deficient, and health surveys indicate that iodine intake has been falling in the United States and Australia. The World Health Organization (WHO) estimates that about 2 billion people are iodine-deficient, based on urinary excretion data. In areas of relative iodine deficiency, there is an increased prevalence of goiter and, when deficiency is severe, hypothyroidism and cretinism. *Cretinism* is characterized by mental and growth retardation and occurs when children who live in iodine-deficient regions are not treated with iodine or thyroid hormone to restore normal thyroid hormone levels during early life. These children are often born to mothers with iodine deficiency, and it is likely that maternal thyroid hormone deficiency worsens the condition. Concomitant selenium deficiency may also contribute to the neurologic manifestations of cretinism. Iodine supplementation of salt, bread, and other food substances has markedly reduced the prevalence of cretinism. Unfortunately, however, iodine deficiency remains the most common cause of preventable mental deficiency, often because of societal resistance to food additives or the cost of supplementation. In addition to overt cretinism, mild iodine deficiency can lead to subtle reduction of IQ. Oversupply of iodine, through supplements or foods enriched in iodine (e.g., shellfish, kelp), is associated with an increased incidence of autoimmune thyroid disease. The RDA is 220 µg iodine per day for pregnant women and 290 µg iodine per day for breastfeeding women. Because the effects of iodine deficiency are most severe in pregnant women and their babies, the American Thyroid Association has recommended that all pregnant and breastfeeding women in the United States and Canada take a prenatal multivitamin containing 150 µg iodine per day. Urinary iodine is >10 µg/dL in iodine-sufficient populations.

FIGURE 375-3

Worldwide iodine nutrition. Data are from the World Health Organization and the International Council for the Control of Iodine Deficiency Disorders *(http://indorgs.virginia.edu/iccidd/mi/cidds.html)*.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Organification, Coupling, Storage, and Release

After iodide enters the thyroid, it is trapped and transported to the apical membrane of thyroid follicular cells, where it is oxidized in an organification reaction that involves TPO and hydrogen peroxide produced by dual oxidase (DUOX) and DUOX maturation factor (DUOXA). The reactive iodine atom is added to specific tyrosyl residues within Tg, a large (660 kDa) dimeric protein that consists of 2769 amino acids. The iodotyrosines in Tg are then coupled via an ether linkage in a reaction that is also catalyzed by TPO. Either T_4 or T_3 can be produced by this reaction, depending on the number of iodine atoms present in the iodotyrosines. After coupling, Tg is taken back into the thyroid cell, where it is processed in lysosomes to release T_4 and T_3. Uncoupled mono- and diiodotyrosines (MIT, DIT) are deiodinated by the enzyme dehalogenase, thereby recycling any iodide that is not converted into thyroid hormones.

Disorders of thyroid hormone synthesis are rare causes of congenital hypothyroidism (Chap. 376). The vast majority of these disorders are due to recessive mutations in TPO or Tg, but defects have also been identified in the TSH-R, NIS, pendrin, hydrogen peroxide generation, and dehalogenase, as well as genes involved in thyroid gland development. Because of the biosynthetic defect, the gland is incapable of synthesizing adequate amounts of hormone, leading to increased TSH and a large goiter.

TSH Action

TSH regulates thyroid gland function through the TSH-R, a seven-transmembrane G protein–coupled receptor (GPCR). The TSH-R is coupled to the α subunit of stimulatory G protein ($G_{S\alpha}$), which activates adenylyl cyclase, leading to increased production of cyclic adenosine monophosphate (AMP). TSH also stimulates phosphatidylinositol turnover by activating phospholipase C. The functional role of the TSH-R is exemplified by the consequences of naturally occurring mutations. Recessive loss-of-function mutations cause thyroid hypoplasia and congenital hypothyroidism. Dominant gain-of-function mutations cause sporadic or familial hyperthyroidism that is characterized by goiter, thyroid cell hyperplasia, and autonomous function (Chap. 377). Most of these activating mutations occur in the transmembrane domain of the receptor. They mimic the conformational changes induced by TSH binding or the interactions of thyroid-stimulating immunoglobulins (TSI) in Graves' disease. Activating TSH-R mutations also occur as somatic events, leading to clonal selection and expansion of the affected thyroid follicular cell and autonomously functioning thyroid nodules.

Other Factors That Influence Hormone Synthesis and Release

Although TSH is the dominant hormonal regulator of thyroid gland growth and function, a variety of growth factors, most produced locally in the thyroid gland, also influence thyroid hormone synthesis. These include insulin-like growth factor I (IGF-I), epidermal growth factor, transforming growth factor β (TGF-β), endothelins, and various cytokines. The quantitative roles of these factors are not well understood, but they are important in selected disease states. In acromegaly, for example, increased levels of growth hormone and IGF-I are associated with goiter and predisposition to multinodular goiter (MNG). Certain cytokines and interleukins (ILs) produced in association with autoimmune thyroid disease induce thyroid growth, whereas others lead to apoptosis. Iodine deficiency increases thyroid blood flow and upregulates the NIS, stimulating more efficient iodine uptake. Excess iodide transiently inhibits thyroid iodide organification, a phenomenon known as the *Wolff-Chaikoff effect*. In individuals with a normal thyroid, the gland escapes from this inhibitory effect and iodide organification resumes; the suppressive action of high iodide may persist, however, in patients with underlying autoimmune thyroid disease.

THYROID FUNCTION IN PREGNANCY

Five factors alter thyroid function in pregnancy: (1) the transient increase in hCG during the first trimester, which weakly stimulates the TSH-R; (2) the estrogen-induced rise in TBG during the first trimester, which is sustained during pregnancy; (3) alterations in the immune system, leading to the onset, exacerbation, or amelioration of an underlying autoimmune thyroid disease; (4) increased thyroid hormone metabolism by the placenta; and (5) increased urinary iodide excretion, which can cause impaired thyroid hormone production in areas of marginal iodine sufficiency. Women with a precarious iodine intake (<50 μg/d) are most at risk of developing a goiter during pregnancy or giving birth to an infant with a goiter and hypothyroidism. The World Health Organization recommends a daily iodine intake of 250 μg during pregnancy and lactation and prenatal vitamins should contain 150 μg per tablet.

The rise in circulating hCG levels during the first trimester is accompanied by a reciprocal fall in TSH that persists into the middle of pregnancy. This reflects the weak binding of hCG, which is present at very high levels, to the TSH-R. Rare individuals have variant TSH-R sequences that enhance hCG binding and TSH-R activation. hCG-induced changes in thyroid function can result in transient gestational hyperthyroidism that may be associated with *hyperemesis gravidarum*, a condition characterized by severe nausea and vomiting and risk of volume depletion. However, since the hyperthyroidism is not causal, antithyroid drugs are not indicated unless concomitant Graves' disease is suspected. Parenteral fluid replacement usually suffices until the condition resolves.

Normative values for most thyroid function tests differ during pregnancy and, if available, trimester specific reference ranges should be used when diagnosing thyroid dysfunction during pregnancy. TSH levels decrease during the first trimester and then rise as gestation progresses. Total T_4 and T_3 levels are about 1.5× higher throughout pregnancy but the free T_4 progressively decreases so that third trimester values in healthy pregnancies are often below the nonpregnant lower reference cutoff.

During pregnancy, subclinical hypothyroidism occurs in 2% of women, but overt hypothyroidism is present in only 1 in 500. Prospective randomized controlled trials have not shown a benefit for universal thyroid disease screening in pregnancy. Targeted TSH testing for hypothyroidism is recommended for women planning a pregnancy if they have a strong family history of autoimmune thyroid disease, other autoimmune disorders (e.g., type 1 diabetes), infertility, prior preterm delivery or recurrent miscarriage, signs or symptoms of thyroid disease, or are older than 30 years. Thyroid hormone requirements are increased by up to 45% during pregnancy in levothyroxine-treated hypothyroid women.

THYROID HORMONE TRANSPORT AND METABOLISM

Serum-Binding Proteins

T_4 is secreted from the thyroid gland in about twentyfold excess over T_3 **(Table 375-1)**. Both hormones are bound to plasma proteins, including thyroxine-binding globulin (TBG), transthyretin (TTR, formerly known as thyroxine-binding prealbumin, or TBPA), and albumin. The plasma-binding proteins increase the pool of circulating hormone, delay hormone clearance, and may modulate hormone delivery to selected tissue sites. The concentration of TBG is relatively low (1–2 mg/dL), but because of its high affinity for thyroid hormones ($T_4 > T_3$), it carries about 80% of the bound hormones. Albumin has relatively low affinity for thyroid hormones but has a high plasma concentration (~3.5 g/dL), and it binds up to 10% of T_4 and 30% of T_3. TTR carries about 10% of T_4 but little T_3.

TABLE 375-1

Characteristics of Circulating T_4 and T_3

HORMONE PROPERTY	T_4	T_3
Serum concentrations		
Total hormone	8 µg/dL	0.14 µg/dL
Fraction of total hormone in the unbound form	0.02%	0.3%
Unbound (free) hormone	$21 \times 10^{-12} M$	$6 \times 10^{-12} M$
Serum half-life	7 d	2 d
Fraction directly from the thyroid	100%	20%
Production rate, including peripheral conversion	90 µg/d	32 µg/d
Intracellular hormone fraction	~20%	~70%
Relative metabolic potency	0.3	1
Receptor binding	$10^{-10} M$	$10^{-11} M$

When the effects of the various binding proteins are combined, ~99.98% of T_4 and 99.7% of T_3 are protein-bound. Because T_3 is less tightly bound than T_4, the fraction of unbound T_3 is greater than unbound T_4, but there is less unbound T_3 in the circulation because it is produced in smaller amounts and cleared more rapidly than T_4. The unbound or "free" concentrations of the hormones are ~2×10^{-11} *M* for T_4 and ~6×10^{-12} *M* for T_3, which roughly correspond to the thyroid hormone receptor-binding constants for these hormones (see below). The unbound hormone is thought to be biologically available to tissues. Nonetheless, the homeostatic mechanisms that regulate the thyroid axis are directed toward maintenance of normal concentrations of unbound hormones.

Abnormalities of Thyroid Hormone-Binding Proteins

A number of inherited and acquired abnormalities affect thyroid hormone-binding proteins. X-linked TBG deficiency is associated with very low levels of total T_4 and T_3. However, because unbound hormone levels

are normal, patients are euthyroid and TSH levels are normal. It is important to recognize this disorder to avoid efforts to normalize total T_4 levels, because this leads to thyrotoxicosis and is futile because of rapid hormone clearance in the absence of TBG. TBG levels are elevated by estrogen, which increases sialylation and delays TBG clearance. Consequently, in women who are pregnant or taking estrogen-containing contraceptives, elevated TBG increases total T_4 and T_3 levels; however, unbound T_4 and T_3 levels are normal.

These features are part of the explanation for why women with hypothyroidism require increased amounts of L-thyroxine replacement as TBG levels are increased by pregnancy or estrogen treatment. Mutations in TBG, TTR, and albumin may increase the binding affinity for T_4 and/or T_3 and cause disorders known as *euthyroid hyperthyroxinemia* or *familial dysalbuminemic hyperthyroxinemia* (FDH) **(Table 375-2)**. These disorders result in increased total T_4 and/or T_3, but unbound hormone levels are normal. The familial nature of the disorders, and the fact that TSH levels are normal rather than suppressed, should suggest this diagnosis. Unbound hormone levels (ideally measured by dialysis) are normal in FDH. The diagnosis can be confirmed by using tests that measure the affinities of radiolabeled hormone binding to specific transport proteins or by performing DNA sequence analyses of the abnormal transport protein genes.

TABLE 375-2

Conditions Associated with Euthyroid Hyperthyroxinemia

DISORDER	CAUSE	TRANSMISSION	CHARACTERISTICS
Familial dysalbuminemic hyperthyroxinemia (FDH)	Albumin mutations, usually R218H	AD	Increased T_4 Normal unbound T_4 Rarely increased T_3
TBG			
Familial excess	Increased TBG production	XL	Increased total T_4, T_3 Normal unbound T_4, T_3
Acquired excess	Medications (estrogen), pregnancy, cirrhosis, hepatitis	Acquired	Increased total T_4, T_3 Normal unbound T_4, T_3
Transthyretin[a]			
Excess	Islet tumors	Acquired	Usually normal T_4, T_3
Mutations	Increased affinity for T_4 or T_3	AD	Increased total T_4, T_3 Normal unbound T_4, T_3
Medications: propranolol, ipodate, iopanoic acid, amiodarone	Decreased $T_4 \rightarrow T_3$ conversion	Acquired	Increased T_4 Decreased T_3 Normal or increased TSH

DISORDER	CAUSE	TRANSMISSION	CHARACTERISTICS
Resistance to thyroid hormone (RTH)	Thyroid hormone receptor β mutations	AD	Increased unbound T_4, T_3 Normal or increased TSH Some patients clinically thyrotoxic

[a]Also known as thyroxine-binding prealbumin (TBPA).

Abbreviations: AD, autosomal dominant; TBG, thyroxine-binding globulin; TSH, thyroid-stimulating hormone; XL, X-linked.

Certain medications, such as salicylates and salsalate, can displace thyroid hormones from circulating binding proteins. Although these drugs transiently perturb the thyroid axis by increasing free thyroid hormone levels, TSH is suppressed until a new steady state is reached, thereby restoring euthyroidism. Circulating factors associated with acute illness may also displace thyroid hormone from binding proteins (Chap. 377).

Deiodinases

T_4 may be thought of as a precursor for the more potent T_3. T_4 is converted to T_3 by the deiodinase enzymes (Fig. 375-1). Type I deiodinase, which is located primarily in thyroid, liver, and kidneys, has a relatively low affinity for T_4. Type II deiodinase has a higher affinity for T_4 and is found primarily in the pituitary gland, brain, brown fat, and thyroid gland. Expression of type II deiodinase allows it to regulate T_3 concentrations locally, a property that may be important in the context of levothyroxine (T_4) replacement. Type II deiodinase is also regulated by thyroid hormone; hypothyroidism induces the enzyme, resulting in enhanced $T_4 \rightarrow T_3$ conversion in tissues such as brain and pituitary. $T_4 \rightarrow T_3$ conversion is impaired by fasting, systemic illness or acute trauma, oral contrast agents, and a variety of medications (e.g., propylthiouracil, propranolol, amiodarone, glucocorticoids). Type III deiodinase inactivates T_4 and T_3 and is the most important source of reverse T_3 (rT_3), including in the sick euthyroid syndrome. This enzyme is expressed in the human placenta but is not active in healthy individuals. In the sick euthyroid syndrome, especially with hypoperfusion, the type III deiodinase is activated in muscle and liver. Massive hemangiomas that express type III deiodinase are a rare cause of consumptive hypothyroidism in infants.

THYROID HORMONE ACTION

Thyroid Hormone Transport

Circulating thyroid hormones enter cells by passive diffusion and via specific transporters such as the monocarboxylate 8 transporter (MCT8), MCT10, and organic anion-transporting polypeptide 1C1. Mutations in the *MCT8* gene have been identified in patients with X-linked psychomotor retardation and thyroid function abnormalities (low T_4, high T_3, and high TSH). After entering cells, thyroid hormones act primarily through nuclear receptors, although they also have nongenomic actions through stimulating mitochondrial enzymatic responses and may act directly on blood vessels and the heart through integrin receptors.

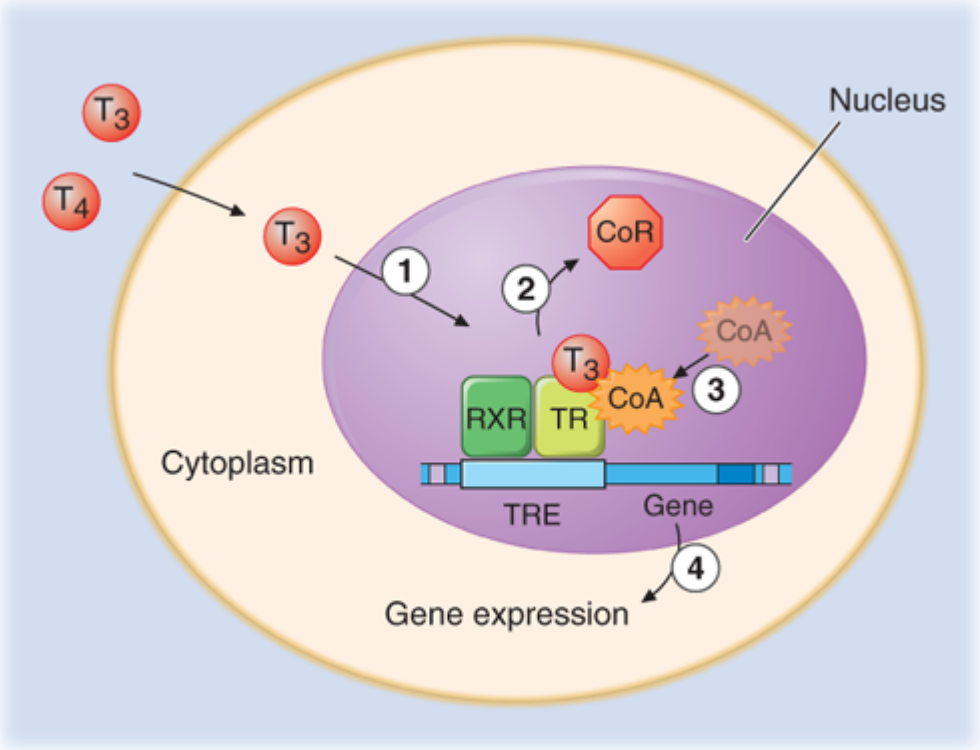
Nuclear Thyroid Hormone Receptors

Thyroid hormones bind with high affinity to nuclear *thyroid hormone receptors* (TRs) α and β. Both TRα and TRβ are expressed in most tissues, but their relative expression levels vary among organs; TRα is particularly abundant in brain, kidneys, gonads, muscle, and heart, whereas TRβ expression is relatively high in the pituitary and liver. Both receptors are variably spliced to form unique isoforms. The TRβ2 isoform, which has a unique amino terminus, is selectively expressed in the hypothalamus and pituitary, where it plays a role in feedback control of the thyroid axis (see above). The TRα2 isoform contains a unique carboxy terminus that precludes thyroid hormone binding; it may function to inhibit the action of other TR isoforms.

The TRs contain a central DNA-binding domain and a C-terminal ligand-binding domain. They bind to specific DNA sequences, termed *thyroid response elements* (TREs), in the promoter regions of target genes **(Fig. 375-4)**. The receptors bind as homodimers or, more commonly, as heterodimers with retinoic acid X receptors (RXRs) **(Chap. 370)**. The activated receptor can either stimulate gene transcription (e.g., myosin heavy chain α) or inhibit transcription (e.g., TSH β-subunit gene), depending on the nature of the regulatory elements in the target gene.

FIGURE 375-4

Mechanism of thyroid hormone receptor action. The thyroid hormone receptor (TR) and retinoid X receptor (RXR) form heterodimers that bind specifically to thyroid hormone response elements (TRE) in the promoter regions of target genes. In the absence of hormone, TR binds co-repressor (CoR) proteins that silence gene expression. The numbers refer to a series of ordered reactions that occur in response to thyroid hormone: (1) T_4 or T_3 enters the nucleus; (2) T_3 binding dissociates CoR from TR; (3) co-activators (CoA) are recruited to the T_3-bound receptor; and (4) gene expression is altered.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Thyroid hormones (T_3 and T_4) bind with similar affinities to TRα and TRβ. However, structural differences in the ligand-binding domains provide the potential for developing receptor-selective agonists or antagonists, and these are under investigation. T_3 is bound with 10–15 times greater affinity than T_4, which explains its increased hormonal potency. Although T_4 is produced in excess of T_3, receptors are occupied mainly by T_3, reflecting $T_4 \rightarrow T_3$ conversion by peripheral tissues, T_3 bioavailability in the plasma, and the greater affinity of receptors for T_3. After binding to TRs, thyroid hormone induces conformational changes in the receptors that modify its interactions with accessory transcription factors. Importantly, in the absence of thyroid hormone binding, the aporeceptors bind to co-repressor proteins that inhibit gene transcription. Hormone binding dissociates the co-repressors and allows the recruitment of co-activators that enhance transcription. The discovery of TR interactions with co-repressors explains the fact that TR silences gene expression in the absence of hormone binding. Consequently, hormone deficiency has a profound effect on gene expression because it causes gene repression as well as loss of hormone-induced stimulation. This concept has been corroborated by the finding that targeted deletion of the TR genes in mice has a less pronounced phenotypic effect than hormone deficiency.

Thyroid Hormone Resistance

Resistance to thyroid hormone (RTH) is an autosomal dominant disorder characterized by elevated thyroid hormone levels and inappropriately normal or elevated TSH. Individuals with RTH do not, in general, exhibit signs and symptoms that are typical of hypothyroidism because hormone resistance is partial and is compensated by increased levels of thyroid hormone. The clinical features of RTH can include goiter,

attention deficit disorder, mild reduction in IQ, delayed skeletal maturation, tachycardia, and impaired metabolic responses to thyroid hormone.

Classical forms of RTH are caused by mutations in the TRβ gene. These mutations, located in restricted regions of the ligand-binding domain, cause loss of receptor function. However, because the mutant receptors retain the capacity to dimerize with RXRs, bind to DNA, and recruit co-repressor proteins, they function as antagonists of the remaining normal TRβ and TRα receptors. This property, referred to as "dominant negative" activity, explains the autosomal dominant mode of transmission. The diagnosis is suspected when unbound thyroid hormone levels are increased without suppression of TSH. Similar hormonal abnormalities are found in other affected family members, although the TRβ mutation arises de novo in about 20% of patients. DNA sequence analysis of the TRβ gene provides a definitive diagnosis. RTH must be distinguished from other causes of euthyroid hyperthyroxinemia (e.g., FDH) and inappropriate secretion of TSH by TSH-secreting pituitary adenomas **(Chap. 373)**. In most patients, no treatment is indicated; the importance of making the diagnosis is to avoid inappropriate treatment of mistaken hyperthyroidism and to provide genetic counseling.

A distinct form of RTH is caused by mutations in the TRα gene. Affected patients have many clinical features of congenital hypothyroidism including growth retardation, skeletal dysplasia, and severe constipation. In contrast to RTH caused by mutations in TRβ, thyroid function tests include normal TSH, low or normal T_4, and normal or elevated T_3 levels. These distinct clinical and laboratory features underscore the different tissue distribution and functional roles of TRβ and TRα. Thyroxine treatment appears to alleviate some of the clinical manifestations of patients with RTH caused by TRα mutations.

PHYSICAL EXAMINATION

In addition to the examination of the thyroid itself, the physical examination should include a search for signs of abnormal thyroid function and the extrathyroidal features of ophthalmopathy and dermopathy (Chap. 377). Examination of the neck begins by inspecting the seated patient from the front and side and noting any surgical scars, obvious masses, or distended veins. The thyroid can be palpated with both hands from behind or while facing the patient, using the thumbs to palpate each lobe. It is best to use a combination of these methods, especially when nodules are small. The patient's neck should be slightly flexed to relax the neck muscles. After locating the cricoid cartilage, the isthmus, which is attached to the lower one-third of the thyroid lobes, can be identified and then followed laterally to locate either lobe (normally, the right lobe is slightly larger than the left). By asking the patient to swallow sips of water, thyroid consistency can be better appreciated as the gland moves beneath the examiner's fingers.

Features to be noted include thyroid size, consistency, nodularity, and any tenderness or fixation. An estimate of thyroid size (normally 12–20 g) should be made, and a drawing is often the best way to record findings. Ultrasound imaging provides the most accurate measurement of thyroid volume and nodularity and is useful for assessment of goiter prevalence in iodine deficient regions. However, ultrasound is not indicated if the thyroid physical examination is normal. The size, location, and consistency of any nodules

should also be defined. A bruit or thrill over the gland, located over the insertion of the superior and inferior thyroid arteries (supero- or inferolaterally), indicates increased vascularity, associated with turbulent rather than laminar blood flow, as occurs in hyperthyroidism. If the lower borders of the thyroid lobes are not clearly felt, a goiter may be retrosternal. Large retrosternal goiters can cause venous distention over the neck and difficulty breathing, especially when the arms are raised (Pemberton's sign). With any central mass above the thyroid, the tongue should be extended, as thyroglossal cysts then move upward. The thyroid examination is not complete without assessment for lymphadenopathy in the supraclavicular and cervical regions of the neck.

LABORATORY EVALUATION

Measurement of Thyroid Hormones

The enhanced sensitivity and specificity of *TSH assays* have greatly improved laboratory assessment of thyroid function. Because TSH levels change dynamically in response to alterations of T_4 and T_3, a logical approach to thyroid testing is to first determine whether TSH is suppressed, normal, or elevated. With rare exceptions (see below), a normal TSH level excludes a primary abnormality of thyroid function. This strategy depends on the use of immunochemiluminometric assays (ICMAs) for TSH that are sensitive enough to discriminate between the lower limit of the reference interval and the suppressed values that occur with thyrotoxicosis. Extremely sensitive assays can detect TSH levels ≤0.004 mIU/L, but, for practical purposes, assays sensitive to ≤0.1 mIU/L are sufficient. The widespread availability of the TSH ICMA has rendered the TRH stimulation test obsolete, because the failure of TSH to rise after an intravenous bolus of 200–400 µg TRH has the same implications as a suppressed basal TSH measured by ICMA.

The finding of an abnormal TSH level must be followed by measurements of circulating thyroid hormone levels to confirm the diagnosis of hyperthyroidism (suppressed TSH) or hypothyroidism (elevated TSH). Automated immunoassays are widely available for serum *total* T_4 and *total* T_3. T_4 and T_3 are highly protein-bound, and numerous factors (illness, medications, genetic factors) can influence protein binding. It is useful, therefore, to measure the free, or unbound, hormone levels, which correspond to the biologically available hormone pool. Two direct methods are used to measure *unbound thyroid hormones*: (1) unbound thyroid hormone competition with radiolabeled T_4 (or an analogue) for binding to a solid-phase antibody, and (2) physical separation of the unbound hormone fraction by ultracentrifugation or equilibrium dialysis. Although early unbound hormone immunoassays suffered from artifacts, newer assays correlate well with the results of the more technically demanding and expensive physical separation methods. An indirect method that is now less commonly used to estimate unbound thyroid hormone levels is to calculate the free T_3 or free T_4 index from the total T_4 or T_3 concentration and the *thyroid hormone binding ratio* (THBR). The latter is derived from the *T_3-resin uptake test*, which determines the distribution of radiolabeled T_3 between an absorbent resin and the unoccupied thyroid hormone binding proteins in the sample. The binding of the labeled T_3 to the resin is increased when there is reduced unoccupied protein binding sites (e.g., TBG deficiency) or increased total thyroid hormone in the sample; it is decreased under the opposite

circumstances. The product of THBR and total T_3 or T_4 provides the *free T_3 or T_4 index*. In effect, the index corrects for anomalous total hormone values caused by variations in hormone-protein binding.

Total thyroid hormone levels are *elevated* when TBG is increased due to estrogens (pregnancy, oral contraceptives, hormone therapy, tamoxifen, selective estrogen receptor modulators, inflammatory liver disease) and *decreased* when TBG binding is reduced (androgens, nephrotic syndrome). Genetic disorders and acute illness can also cause abnormalities in thyroid hormone-binding proteins, and various drugs (phenytoin, carbamazepine, salicylates, and nonsteroidal anti-inflammatory drugs [NSAIDs]) can interfere with thyroid hormone binding. Because unbound thyroid hormone levels are normal and the patient is euthyroid in all of these circumstances, assays that measure unbound hormone are preferable to those for total thyroid hormones.

For most purposes, the unbound T_4 level is sufficient to confirm thyrotoxicosis, but 2–5% of patients have only an elevated T_3 level (T_3 toxicosis). Thus, unbound T_3 levels should be measured in patients with a suppressed TSH but normal unbound T_4 levels.

There are several clinical conditions in which the use of TSH as a screening test may be misleading, particularly without simultaneous unbound T_4 determinations. Any severe nonthyroidal illness can cause abnormal TSH levels. Although hypothyroidism is the most common cause of an elevated TSH level, rare causes include a TSH-secreting pituitary tumor **(Chap. 373)**, thyroid hormone resistance, and assay artifact. Conversely, a suppressed TSH level, particularly <0.01 mIU/L, usually indicates thyrotoxicosis. However, subnormal TSH levels between 0.01 and 0.1 mIU/L may be seen during the first trimester of pregnancy (due to hCG secretion), after treatment of hyperthyroidism (because TSH can remain suppressed for several months), and in response to certain medications (e.g., high doses of glucocorticoids or dopamine). TSH levels measured by immunoassay may also be suppressed in patients ingesting biotin supplements <18 hours prior to a blood draw because the TSH capture antibodies are biotinylated and the exogenous biotin can interfere with the subsequent streptavidin capture. Importantly, secondary hypothyroidism, caused by hypothalamic-pituitary disease, is associated with a variable (low to high-normal) TSH level, which is inappropriate for the low T_4 level. Thus, *TSH should not be used as an isolated laboratory test to assess thyroid function in patients with suspected or known pituitary disease*.

Tests for the end-organ effects of thyroid hormone excess or depletion, such as estimation of basal metabolic rate, tendon reflex relaxation rates, or serum cholesterol, are relatively insensitive and are not useful as clinical determinants of thyroid function.

Tests to Determine the Etiology of Thyroid Dysfunction

Autoimmune thyroid disease is detected most easily by measuring circulating antibodies against TPO and Tg. Because antibodies to Tg alone are uncommon, it is reasonable to measure only TPO antibodies. About 5–15% of euthyroid women and up to 2% of euthyroid men have thyroid antibodies; such individuals are at

increased risk of developing thyroid dysfunction. Almost all patients with autoimmune hypothyroidism, and up to 80% of those with Graves' disease, have TPO antibodies, usually at high levels.

TSIs are antibodies that stimulate the TSH-R in Graves' disease. They are most commonly measured by commercially available tracer displacement assays called TRAb (TSH receptor antibody) with the assumption that elevated levels in the setting of clinical hyperthyroidism reflect stimulatory effects on the TSH receptor. A bioassay is less commonly used. Remission rates in patients with Graves' disease after antithyroid drug cessation are higher with disappearance rather than persistence of TRAb. Furthermore, the TRAb assay is used to predict both fetal and neonatal thyrotoxicosis caused by transplacental passage of high maternal levels of TRAb or TSI (>3× upper limit of normal) in the last trimester of pregnancy.

Serum Tg levels are increased in all types of thyrotoxicosis except *thyrotoxicosis factitia* caused by self-administration of thyroid hormone. Tg levels are particularly increased in thyroiditis, reflecting thyroid tissue destruction and release of Tg. The main role for Tg measurement, however, is in the follow-up of thyroid cancer patients. After total thyroidectomy and radioablation, Tg levels should be undetectable; in the absence of anti-Tg antibodies, measurable levels indicate incomplete ablation or recurrent cancer.

Radioiodine Uptake and Thyroid Scanning

The thyroid gland selectively transports radioisotopes of iodine (^{123}I, ^{125}I, ^{131}I) and ^{99m}Tc pertechnetate, allowing thyroid imaging and quantitation of radioactive tracer fractional uptake.

Nuclear imaging of Graves' disease is characterized by an enlarged gland and increased tracer uptake that is distributed homogeneously. Toxic adenomas appear as focal areas of increased uptake, with suppressed tracer uptake in the remainder of the gland. In toxic MNG, the gland is enlarged—often with distorted architecture—and there are multiple areas of relatively increased (functioning nodules) or decreased tracer uptake (suppressed thyroid parenchyma or nonfunctioning nodules). Subacute, viral, and postpartum thyroiditis are associated with very low uptake because of follicular cell damage and TSH suppression. Thyrotoxicosis factitia is also associated with low uptake. In addition, if there is excessive circulating exogenous iodine (e.g., from dietary sources of iodinated contrast dye), the radionuclide uptake is low even in the presence of increased thyroid hormone production.

Thyroid scintigraphy is not used in the routine evaluation of patients with thyroid nodules, but should be performed if the serum TSH level is subnormal to determine if functioning thyroid nodules are present. Functioning or "hot" nodules are almost never malignant, and fine-needle aspiration (FNA) biopsy is not indicated. The vast majority of thyroid nodules do not produce thyroid hormone ("cold" nodules), and these are more likely to be malignant (~5–10%). Whole-body and thyroid scanning is also used in the treatment and surveillance of thyroid cancer. After thyroidectomy for thyroid cancer, the TSH level is raised by either using a thyroid hormone withdrawal protocol or recombinant human TSH injection (Chap. 378). Administration of either ^{131}I or ^{123}I (in higher activities than used to image the thyroid gland alone) allows whole-body scanning (WBS) to confirm remnant ablation and to detect any functioning metastases. In addition, WBS may be helpful in surveillance of patients at risk for recurrence.

Thyroid Ultrasound

Ultrasonography is valuable for the diagnosis and evaluation of patients with nodular thyroid disease **(Chap. 378)**. Evidence-based guidelines recommend thyroid ultrasonography for all patients suspected of having thyroid nodules by either physical examination or another imaging study. Using 10- to 12-MHz linear transducers, resolution and image quality are excellent, allowing the characterization of nodules and cysts >3 mm. Sonographic patterns that combine suspicious sonographic features are highly suggestive of malignancy (e.g., hypoechoic solid nodules with infiltrative borders and microcalcifications, >90% cancer risk), whereas other patterns correlate with a lower likelihood of cancer (isoechoic solid nodules, 5–10% cancer risk). Some patterns suggest benignity (e.g., spongiform nodules, defined as those with multiple small internal cystic areas, or simple cysts <3% cancer risk) **(see Chap. 378, Fig. 378-1)**. In addition to evaluating thyroid nodules, ultrasound is useful for monitoring nodule size and for the aspiration of nodules or cystic lesions. Ultrasound-guided FNA biopsy of thyroid lesions lowers the rate of inadequate sampling and decreases sample error, thereby reducing both the nondiagnostic and false-negative rates of FNA cytology. Ultrasonography of the central and lateral cervical lymph node compartments is indispensable in the evaluation thyroid cancer patients, preoperatively and during follow-up. In addition, the American College of Radiology recommends a survey of the cervical lymph nodes as part of every diagnostic thyroid sonographic examination.

FURTHER READING

Alexander EK et al: 2017 Guidelines of the American Thyroid Association for the diagnosis and management of thyroid disease during pregnancy and postpartum. Thyroid 27:315, 2017.

Ortiga-Cavalho TM, Sidhaye AR, Wondisford FE: Thyroid hormone receptors and resistance to thyroid hormone disorders. Nat Rev Endocrinol 10:582, 2014. [PubMed: 25135573]

Rugge JB, Bougatsos C, Chou R: Screening and treatment of thyroid dysfunction: An evidence review for the U. S. Preventive Services Task Force. Ann Intern Med 162:35, 2015. [PubMed: 25347444]

Stoupa A et al: Update of thyroid developmental genes. Endocrinol Metab Clin North Am 45:243, 2016. [PubMed: 27241962]

Zimmermann MB, Boelaert K: Iodine deficiency and thyroid disorders. Lancet Diabetes Endocrinol 3: 286, 2015. [PubMed: 25591468]

Chapter 376: Hypothyroidism

J. Larry Jameson; Susan J. Mandel; Anthony P. Weetman

HYPOTHYROIDISM

Iodine deficiency remains a common cause of hypothyroidism worldwide. In areas of iodine sufficiency, autoimmune disease (Hashimoto's thyroiditis) and iatrogenic causes (treatment of hyperthyroidism) are most common (Table 376-1).

TABLE 376-1

Causes of Hypothyroidism

Primary
Autoimmune hypothyroidism: Hashimoto's thyroiditis, atrophic thyroiditis
Iatrogenic: ^{131}I treatment, subtotal or total thyroidectomy, external irradiation of neck for lymphoma or cancer
Drugs: iodine excess (including iodine-containing contrast media and amiodarone), lithium, antithyroid drugs, *p*-aminosalicylic acid, interferon α and other cytokines, aminoglutethimide, tyrosine kinase inhibitors (e.g., sunitinib)
Congenital hypothyroidism: absent or ectopic thyroid gland, dyshormonogenesis, TSH-R mutation
Iodine deficiency
Infiltrative disorders: amyloidosis, sarcoidosis, hemochromatosis, scleroderma, cystinosis, Riedel's thyroiditis
Overexpression of type 3 deiodinase in infantile hemangioma and other tumors
Transient
Silent thyroiditis, including postpartum thyroiditis
Subacute thyroiditis
Withdrawal of supraphysiologic thyroxine treatment in individuals with an intact thyroid
After ^{131}I treatment or subtotal thyroidectomy for Graves' disease
Secondary
Hypopituitarism: tumors, pituitary surgery or irradiation, infiltrative disorders, Sheehan's syndrome, trauma, genetic forms of combined pituitary hormone deficiencies
Isolated TSH deficiency or inactivity
Bexarotene treatment
Hypothalamic disease: tumors, trauma, infiltrative disorders, idiopathic

Abbreviations: TSH, thyroid-stimulating hormone; TSH-R, TSH receptor.

CONGENITAL HYPOTHYROIDISM

Prevalence

Hypothyroidism occurs in about 1 in 4000 newborns and neonatal screening is performed in most industrialized countries. It may be transient, especially if the mother has TSH-R blocking antibodies or has received antithyroid drugs, but permanent hypothyroidism occurs in the majority. Neonatal hypothyroidism is due to thyroid gland dysgenesis in 80–85%, to inborn errors of thyroid hormone synthesis in 10–15%, and is TSH-R antibody-mediated in 5% of affected newborns. The developmental abnormalities are twice as common in girls. Mutations that cause congenital hypothyroidism are being increasingly identified, but most remain idiopathic **(Table 376-2)**. Transplacental passage of maternal thyroid hormone occurs before the fetal thyroid gland begins to function and provides partial hormone support to a fetus with congenital hypothyroidism.

TABLE 376-2

Genetic Causes of Congenital Hypothyroidism

DEFECTIVE GENE PROTEIN	INHERITANCE	CONSEQUENCES
PROP-1	Autosomal recessive	Combined pituitary hormone deficiencies with preservation of adrenocorticotropic hormone
PIT-1	Autosomal recessive Autosomal dominant	Combined deficiencies of growth hormone, prolactin, thyroid-stimulating hormone (TSH)
TSHβ	Autosomal recessive	TSH deficiency
TTF-1 (TITF-1)	Autosomal dominant	Variable thyroid hypoplasia, choreoathetosis, pulmonary problems
TTF-2 (FOXE-1)	Autosomal recessive	Thyroid agenesis, choanal atresia, spiky hair
PAX-8	Autosomal dominant	Thyroid dysgenesis, kidney abnormalities
NKX2-1	Autosomal dominant	Thyroid dysgenesis, brain, lung abnormalities
NKX2-5	Autosomal dominant	Thyroid dysgenesis, heart abnormalities
TSH-receptor	Autosomal recessive	Resistance to TSH
$G_{S\alpha}$ (Albright hereditary osteodystrophy)	Autosomal dominant	Resistance to TSH
Na^+/I^- symporter (SLC5A5)	Autosomal recessive	Inability to transport iodide

DEFECTIVE GENE PROTEIN	INHERITANCE	CONSEQUENCES
DUOX2 (THOX2)	Autosomal dominant	Organification defect
DUOXA2	Autosomal recessive	Organification defect
Thyroid peroxidase	Autosomal recessive	Defective organification of iodide
Thyroglobulin	Autosomal recessive	Defective synthesis of thyroid hormone
Pendrin (SLC26A4)	Autosomal recessive	Pendred syndrome: sensorineural deafness and partial organification defect in thyroid
Dehalogenase 1 (IYD)	Autosomal recessive	Loss of iodide reutilization

Clinical Manifestations

The majority of infants appear normal at birth, and with the use of biochemical screening, few cases are now diagnosed based on clinical features, which include prolonged jaundice, feeding problems, hypotonia, enlarged tongue, delayed bone maturation, and umbilical hernia. Importantly, permanent neurologic damage results if treatment is delayed. Typical features of adult hypothyroidism may also be present **(Table 376-3)**. Other congenital malformations, especially cardiac, are four times more common in congenital hypothyroidism.

TABLE 376-3

Signs and Symptoms of Hypothyroidism (Descending Order of Frequency)

SYMPTOMS	SIGNS
Tiredness, weakness	Dry coarse skin; cool peripheral extremities
Dry skin	Puffy face, hands, and feet (myxedema)
Feeling cold	Diffuse alopecia
Hair loss	Bradycardia
Difficulty concentrating and poor memory	Peripheral edema
Constipation	Delayed tendon reflex relaxation
Weight gain with poor appetite	Carpal tunnel syndrome
Dyspnea	Serous cavity effusions
Hoarse voice	
Menorrhagia (later oligomenorrhea or amenorrhea)	
Paresthesia	
Impaired hearing	

Diagnosis and Treatment

Because of the severe neurologic consequences of untreated congenital hypothyroidism, neonatal screening programs have been established. These are generally based on measurement of TSH or T_4 levels in heel-prick blood specimens. When the diagnosis is confirmed, T_4 is instituted at a dose of 10–15 μg/kg per day, and the dose is adjusted by close monitoring of TSH levels. T_4 requirements are relatively great during the first year of life, and a high circulating T_4 level is usually needed to normalize TSH. Early treatment with T_4 results in normal IQ levels, but subtle neurodevelopmental abnormalities may occur in those with the most severe hypothyroidism at diagnosis or when treatment is delayed or suboptimal. If transient hypothyroidism is suspected, or the diagnosis is unclear, treatment can be stopped safely after the age of 3 years followed by further evaluation.

AUTOIMMUNE HYPOTHYROIDISM

Classification

Autoimmune hypothyroidism may be associated with a goiter (Hashimoto's, or *goitrous thyroiditis*) or, at the later stages of the disease, minimal residual thyroid tissue (*atrophic thyroiditis*). Because the autoimmune process gradually reduces thyroid function, there is a phase of compensation when normal thyroid hormone levels are maintained by a rise in TSH. Although some patients may have minor symptoms, this state is called

subclinical hypothyroidism. Later, unbound T_4 levels fall and TSH levels rise further; symptoms become more readily apparent at this stage (usually TSH >10 mIU/L), which is referred to as *clinical hypothyroidism* or *overt hypothyroidism*.

Prevalence

The mean annual incidence rate of autoimmune hypothyroidism is up to 4 per 1000 women and 1 per 1000 men. It is more common in certain populations, such as the Japanese, probably because of genetic factors and chronic exposure to a high-iodine diet. The mean age at diagnosis is 60 years, and the prevalence of overt hypothyroidism increases with age. Subclinical hypothyroidism is found in 6–8% of women (10% over the age of 60) and 3% of men. The annual risk of developing clinical hypothyroidism is about 4% when subclinical hypothyroidism is associated with positive thyroid peroxidase (TPO) antibodies.

Pathogenesis

In Hashimoto's thyroiditis, there is a marked lymphocytic infiltration of the thyroid with germinal center formation, atrophy of the thyroid follicles accompanied by oxyphil metaplasia, absence of colloid, and mild to moderate fibrosis. In atrophic thyroiditis, the fibrosis is much more extensive, lymphocyte infiltration is less pronounced, and thyroid follicles are almost completely absent. Atrophic thyroiditis usually represents the end stage of Hashimoto's thyroiditis rather than a separate disorder, although a distinct form of marked fibrosis occurs in which the gland is infiltrated with IgG4-positive plasma cells.

As with most autoimmune disorders, susceptibility to autoimmune hypothyroidism is determined by a combination of genetic and environmental factors, and the risk of either autoimmune hypothyroidism or Graves' disease is increased among siblings. HLA-DR polymorphisms are the best documented genetic risk factors for autoimmune hypothyroidism, especially HLA-DR3, DR4, and DR5 in Caucasians. A weak association also exists between polymorphisms in *CTLA-4*, a T cell–regulatory gene, and autoimmune hypothyroidism. Both of these genetic associations are shared by other autoimmune diseases, which may explain the relationship between autoimmune hypothyroidism and other autoimmune diseases, especially type 1 diabetes mellitus, Addison's disease, pernicious anemia, and vitiligo. HLA-DR and *CTLA-4* polymorphisms account for approximately half of the genetic susceptibility to autoimmune hypothyroidism and the role of other contributory loci remains to be clarified. A gene on chromosome 21 may be responsible for the association between autoimmune hypothyroidism and Down's syndrome. The female preponderance of thyroid autoimmunity is most likely due to sex steroid effects on the immune response, but an X chromosome–related genetic factor is also possible and may account for the high frequency of autoimmune hypothyroidism in Turner's syndrome. Environmental susceptibility factors are poorly defined at present. A high iodine or low selenium intake and decreased exposure to microorganisms in childhood increase the risk of autoimmune hypothyroidism. Smoking cessation transiently increases incidence whereas alcohol intake seems protective. These factors may account for the increase in prevalence over the last two to three decades.

The thyroid lymphocytic infiltrate in autoimmune hypothyroidism is composed of activated T cells as well as B cells. Thyroid cell destruction is primarily mediated by the CD8+ cytotoxic T cells but local production of cytokines, such as tumor necrosis factor (TNF), interleukin-1 (IL-1), and interferon γ (IFN-γ), derived from the inflammatory infiltrate may render thyroid cells more susceptible to apoptosis mediated by death receptors, such as Fas, and by oxidative stress. These cytokines also impair thyroid cell function directly and induce the expression of other proinflammatory molecules by the thyroid cells themselves, such as cytokines, HLA class I and class II molecules, adhesion molecules, CD40, and nitric oxide. Administration of high concentrations of cytokines for therapeutic purposes (especially IFN-α) is associated with increased autoimmune thyroid disease, possibly through mechanisms similar to those in sporadic disease. Novel anticancer and immunomodulatory treatments, such as tyrosine kinase inhibitors and alemtuzumab, can also induce thyroid autoimmunity via their effects on T cell regulation.

Antibodies to TPO and thyroglobulin (Tg) are clinically useful markers of thyroid autoimmunity, but any pathogenic effect is restricted to a secondary role in amplifying an ongoing autoimmune response. TPO antibodies fix complement, and complement membrane-attack complexes are present in the thyroid in autoimmune hypothyroidism. However, transplacental passage of Tg or TPO antibodies has no effect on the fetal thyroid, which suggests that T cell–mediated injury is required to initiate autoimmune damage to the thyroid.

Up to 20% of patients with autoimmune hypothyroidism have antibodies against the TSH-R, which, in contrast to thyroid-stimulating immunoglobulin (TSI), do not stimulate the receptor but prevent the binding of TSH. These TSH-R-blocking antibodies, therefore, cause hypothyroidism and, especially in Asian patients, thyroid atrophy. Their transplacental passage may induce transient neonatal hypothyroidism. Rarely, patients have a mixture of TSI and TSH-R-blocking antibodies, and thyroid function can oscillate between hyperthyroidism and hypothyroidism as one or the other antibody becomes dominant. Predicting the course of disease in such individuals is difficult, and they require close monitoring of thyroid function. Bioassays can be used to document that TSH-R-blocking antibodies reduce the cyclic AMP–inducing effect of TSH on cultured TSH-R-expressing cells, but these assays are difficult to perform. Thyrotropin-binding inhibitory immunoglobulin (TBII) assays that measure the binding of antibodies to the receptor by competition with labeled TSH do not distinguish between TSI and TSH-R-blocking antibodies, but a positive result in a patient with spontaneous hypothyroidism is strong evidence for the presence of blocking antibodies. The use of these assays does not generally alter clinical management, although it may be useful to confirm the cause of transient neonatal hypothyroidism.

Clinical Manifestations

The main clinical features of hypothyroidism are summarized in Table 376-3. The onset is usually insidious, and the patient may become aware of symptoms only when euthyroidism is restored. Patients with Hashimoto's thyroiditis may present because of goiter rather than symptoms of hypothyroidism. The goiter may not be large, but it is usually irregular and firm in consistency. Rarely uncomplicated Hashimoto's thyroiditis is associated with pain.

Patients with atrophic thyroiditis or the later stage of Hashimoto's thyroiditis present with symptoms and signs of hypothyroidism. The skin is dry, and there is decreased sweating, thinning of the epidermis, and hyperkeratosis of the stratum corneum. Increased dermal glycosaminoglycan content traps water, giving rise to skin thickening without pitting (*myxedema*). Typical features include a puffy face with edematous eyelids and nonpitting pretibial edema **(Fig. 376-1)**. There is pallor, often with a yellow tinge to the skin due to carotene accumulation. Nail growth is retarded, and hair is dry, brittle, difficult to manage, and falls out easily. In addition to diffuse alopecia, there is thinning of the outer third of the eyebrows, although this is not a specific sign of hypothyroidism.

FIGURE 376-1

Facial appearance in hypothyroidism. Note puffy eyes and thickened skin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Other common features include constipation and weight gain (despite a poor appetite). In contrast to popular perception, the weight gain is usually modest and due mainly to fluid retention in the myxedematous tissues. Libido is decreased in both sexes, and there may be oligomenorrhea or amenorrhea

in long-standing disease, but menorrhagia may occur at an early stage. Fertility is reduced, and the incidence of miscarriage is increased. Prolactin levels are often modestly increased (Chap. 373) and may contribute to alterations in libido and fertility and cause galactorrhea.

Myocardial contractility and pulse rate are reduced, leading to a reduced stroke volume and bradycardia. Increased peripheral resistance may be accompanied by hypertension, particularly diastolic. Blood flow is diverted from the skin, producing cool extremities. Pericardial effusions occur in up to 30% of patients but rarely compromise cardiac function. Although alterations in myosin heavy chain isoform expression have been documented, cardiomyopathy is rare. Fluid may also accumulate in other serous cavities and in the middle ear, giving rise to conductive deafness. Pulmonary function is generally normal, but dyspnea may be caused by pleural effusion, impaired respiratory muscle function, diminished ventilatory drive, or sleep apnea.

Carpal tunnel and other entrapment syndromes are common, as is impairment of muscle function with stiffness, cramps, and pain. On examination, there may be slow relaxation of tendon reflexes and pseudomyotonia. Memory and concentration are impaired. Experimentally, positron emission tomography (PET) scans examining glucose metabolism in hypothyroid subjects show lower regional activity in the amygdala, hippocampus, and perigenual anterior cingulated cortex, among other regions, and this activity corrects after thyroxine replacement. Rare neurologic problems include reversible cerebellar ataxia, dementia, psychosis, and myxedema coma. *Hashimoto's encephalopathy* has been defined as a steroid-responsive syndrome associated with TPO antibodies, myoclonus, and slow-wave activity on electroencephalography, but the relationship with thyroid autoimmunity or hypothyroidism is not established. The hoarse voice and occasionally clumsy speech of hypothyroidism reflect fluid accumulation in the vocal cords and tongue.

The features described above are the consequence of thyroid hormone deficiency. However, autoimmune hypothyroidism may be associated with signs or symptoms of other autoimmune diseases, particularly vitiligo, pernicious anemia, Addison's disease, alopecia areata, and type 1 diabetes mellitus. Less common associations include celiac disease, dermatitis herpetiformis, chronic active hepatitis, rheumatoid arthritis, systemic lupus erythematosus (SLE), myasthenia gravis, and Sjögren's syndrome. Thyroid-associated ophthalmopathy usually occurs in Graves' disease (see below), but in about 5% of patients it is associated with autoimmune hypothyroidism.

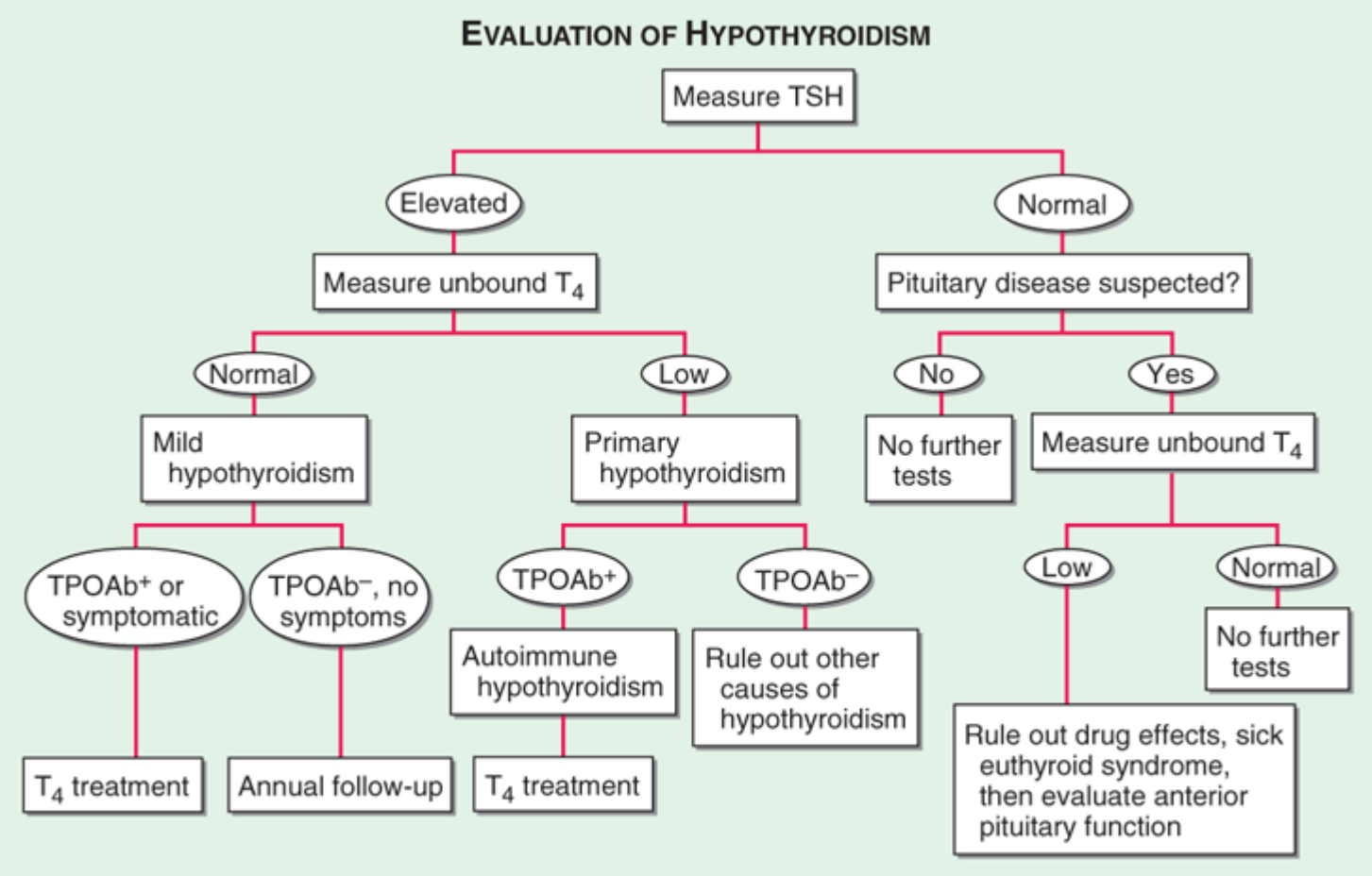
Autoimmune hypothyroidism is uncommon in children and usually presents with slow growth and delayed facial and dental maturation. The pituitary may be enlarged due to thyrotroph hyperplasia. Myopathy, with muscle swelling, is more common in children than in adults. In most cases, puberty is delayed, but precocious puberty sometimes occurs. There may be intellectual impairment if the onset is before 3 years and the hormone deficiency is severe.

Laboratory Evaluation

A summary of the investigations used to determine the existence and cause of hypothyroidism is provided in Fig. 376-2. A normal TSH level excludes primary (but not secondary) hypothyroidism. If the TSH is elevated, an unbound T_4 level is needed to confirm the presence of clinical hypothyroidism, but T_4 is inferior to TSH when used as a screening test, because it will not detect subclinical hypothyroidism. Circulating unbound T_3 levels are normal in about 25% of patients, reflecting adaptive deiodinase responses to hypothyroidism. T_3 measurements are, therefore, not indicated.

FIGURE 376-2

Evaluation of hypothyroidism. $TPOAb^+$, thyroid peroxidase antibodies present; $TPOAb^-$, thyroid peroxidase antibodies not present; TSH, thyroid-stimulating hormone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Once clinical or subclinical hypothyroidism is confirmed, the etiology is usually easily established by demonstrating the presence of TPO and Tg antibodies, which are present in >95% of patients with autoimmune hypothyroidism. TBII can be found in 10–20% of patients, but measurement is not needed routinely. Other abnormal laboratory findings in hypothyroidism may include increased creatine phosphokinase, elevated cholesterol and triglycerides, and anemia (usually normocytic or macrocytic). Except when accompanied by iron deficiency, the anemia and other abnormalities gradually resolve with thyroxine replacement.

Differential Diagnosis

An asymmetric goiter in Hashimoto's thyroiditis may be confused with a multinodular goiter (MNG) or thyroid carcinoma, in which thyroid antibodies may also be present. Ultrasound can be used to show the presence of a solitary lesion or an MNG rather than the heterogeneous thyroid enlargement typical of Hashimoto's thyroiditis. FNA biopsy is useful in the investigation of focal nodules. Other causes of hypothyroidism are discussed below and in Table 376-1 but rarely cause diagnostic confusion.

OTHER CAUSES OF HYPOTHYROIDISM

Iatrogenic hypothyroidism is a common cause of hypothyroidism and can often be detected by screening before symptoms develop. In the first 3–4 months after radioiodine treatment for Graves' disease, transient hypothyroidism may occur due to reversible radiation damage. Low-dose thyroxine treatment can be withdrawn if recovery occurs. Because TSH levels are suppressed by hyperthyroidism, unbound T_4 levels are a better measure of thyroid function than TSH in the months following radioiodine treatment. Mild hypothyroidism after subtotal thyroidectomy may also resolve after several months, as the gland remnant is stimulated by increased TSH levels.

Iodine deficiency is responsible for endemic goiter and cretinism but is an uncommon cause of adult hypothyroidism unless the iodine intake is very low or there are complicating factors, such as the consumption of thiocyanates in cassava or selenium deficiency. Although hypothyroidism due to iodine deficiency can be treated with thyroxine, public health measures to improve iodine intake should be advocated to eliminate this problem. Iodized salt or bread or a single bolus of oral or intramuscular iodized oil have all been used successfully.

Paradoxically, chronic iodine excess can also induce goiter and hypothyroidism. The intracellular events that account for this effect are unclear, but individuals with autoimmune thyroiditis are especially susceptible. Iodine excess is responsible for the hypothyroidism that occurs in up to 13% of patients treated with amiodarone (see below). Other drugs, particularly lithium, may also cause hypothyroidism. Transient hypothyroidism caused by thyroiditis is discussed below.

Secondary hypothyroidism is usually diagnosed in the context of other anterior pituitary hormone deficiencies; isolated TSH deficiency is very rare (Chap. 372). TSH levels may be low, normal, or even slightly increased in secondary hypothyroidism; the latter is due to secretion of immunoactive but bioinactive forms of TSH. The diagnosis is confirmed by detecting a low unbound T_4 level. The goal of treatment is to maintain T_4 levels in the upper half of the reference interval, because TSH levels cannot be used to monitor therapy.

TREATMENT

TREATMENT

Hypothyroidism

CLINICAL HYPOTHYROIDISM

If there is no residual thyroid function, the daily replacement dose of levothyroxine is usually 1.6 µg/kg body weight (typically 100–150 µg), ideally taken at least 30 min before breakfast. In many patients, however, lower doses suffice until residual thyroid tissue is destroyed. In patients who develop hypothyroidism after the treatment of Graves' disease, there is often underlying autonomous function, necessitating lower replacement doses (typically 75–125 µg/d).

Adult patients under 60 years old without evidence of heart disease may be started on 50–100 µg levothyroxine (T_4) daily. The dose is adjusted on the basis of TSH levels, with the goal of treatment being a normal TSH, ideally in the lower half of the reference range. TSH responses are gradual and should be measured about 2 months after instituting treatment or after any subsequent change in levothyroxine dosage. The clinical effects of levothyroxine replacement are slow to appear. Patients may not experience full relief from symptoms until 3–6 months after normal TSH levels are restored. Adjustment of levothyroxine dosage is made in 12.5- or 25-µg increments if the TSH is high; decrements of the same magnitude should be made if the TSH is suppressed. Patients with a suppressed TSH of any cause, including T_4 overtreatment, have an increased risk of atrial fibrillation and reduced bone density.

Although desiccated animal thyroid preparations (thyroid extract USP) are available, they are not recommended because the ratio of T_3 to T_4 is nonphysiologic. The use of levothyroxine combined with liothyronine (triiodothyronine, T_3) has been investigated, but benefit has not been confirmed in prospective studies. There is no place for liothyronine alone as long-term replacement, because the short half-life necessitates three or four daily doses and is associated with fluctuating T_3 levels.

Once full replacement is achieved and TSH levels are stable, follow-up measurement of TSH is recommended at annual intervals. It is important to ensure ongoing adherence as patients do not feel any symptomatic difference after missing a few doses of levothyroxine, and this sometimes leads to self-discontinuation.

In patients of normal body weight who are taking ≥200 µg of levothyroxine per day, an elevated TSH level is often a sign of poor adherence to treatment. This is also the likely explanation for fluctuating TSH levels, despite a constant levothyroxine dosage. Such patients often have normal or high unbound T_4 levels, despite an elevated TSH, because they remember to take medication for a few days before testing; this is sufficient to normalize T_4, but not TSH levels. It is important to consider variable adherence, because this pattern of thyroid function tests is otherwise suggestive of disorders associated with inappropriate TSH secretion (Chap. 375). Because T_4 has a long half-life (7 days), patients who miss a dose can be advised to take two doses of the skipped tablets at once. Other causes of increased levothyroxine requirements must be excluded, particularly malabsorption (e.g., celiac disease, small-bowel surgery, atrophic or *Helicobacter pylori*–related gastritis,), oral estrogen containing medications or selective estrogen receptor modulator therapy, ingestion with a meal, and drugs that interfere with T_4 absorption or metabolism such as bile acid sequestrants, ferrous sulfate, calcium supplements, selevamer, sucralfate, proton pump inhibitors,

lovastatin, aluminum hydroxide, rifampicin, amiodarone, carbamazepine, phenytoin, and tyrosine kinase inhibitors.

SUBCLINICAL HYPOTHYROIDISM

By definition, subclinical hypothyroidism refers to biochemical evidence of thyroid hormone deficiency in patients who have few or no apparent clinical features of hypothyroidism. There are no universally accepted recommendations for the management of subclinical hypothyroidism, but levothyroxine is recommended if the patient is a woman who wishes to conceive or is pregnant, or when TSH levels are above 10 mIU/L. Otherwise, when TSH levels are below 10 mIU/L, a trial of treatment may be considered when patients have suggestive symptoms of hypothyroidism, positive TPO antibodies, or any evidence of heart disease. It is important to confirm that any elevation of TSH is sustained over a 3-month period before treatment is given. Treatment is administered by starting with a low dose of levothyroxine (25–50 µg/d) with the goal of normalizing TSH. If levothyroxine is not given, thyroid function should be evaluated annually.

SPECIAL TREATMENT CONSIDERATIONS

Rarely, levothyroxine replacement is associated with pseudotumor cerebri in children. Presentation appears to be idiosyncratic and occurs months after treatment has begun.

Because maternal hypothyroidism may both adversely affect fetal neural development and be associated with adverse gestational outcomes (miscarriage, preterm delivery), thyroid function should be monitored to preserve euthyroidism in women with a history or high risk of hypothyroidism. The presence of thyroid autoantibodies alone, in a euthyroid patient, is also associated with miscarriage and preterm delivery; large-scale trials are underway to establish whether levothyroxine therapy improves outcomes in this group. Prior to conception, levothyroxine therapy should be targeted to maintain a serum TSH in the normal range but <2.5 mIU/L for hypothyroid women. Subsequently, thyroid function should be evaluated immediately after pregnancy is confirmed and every 4 weeks during the first half of the pregnancy, with less frequent testing after 20 weeks' gestation (every 6–8 weeks depending on whether levothyroxine dose adjustment is ongoing). The levothyroxine dose may need to be increased by up to 45% during pregnancy. Women should increase levothyroxine from once daily dosing to nine doses per week as soon as pregnancy is confirmed, to anticipate this change. Thereafter dosage should be closely monitored with a goal TSH in the lower half of the trimester-specific normative range, if available, or <2.5 mIU/L. After delivery, levothyroxine doses typically return to prepregnancy levels. Pregnant women should be counseled to separate ingestion of prenatal vitamins and iron supplements from levothyroxine.

Elderly patients may require 20% less thyroxine than younger patients. In the elderly, especially patients with known coronary artery disease, the starting dose of levothyroxine is 12.5–25 µg/d with similar increments every 2–3 months until TSH is normalized. In some patients, it may be impossible to achieve full replacement despite optimal antianginal treatment. *Emergency surgery* is generally safe in patients with untreated hypothyroidism, although routine surgery in a hypothyroid patient should be deferred until euthyroidism is achieved.

Myxedema coma still has a 20–40% mortality rate, despite intensive treatment, and outcomes are independent of the T_4 and TSH levels. Clinical manifestations include reduced level of consciousness, sometimes associated with seizures, as well as the other features of hypothyroidism (Table 376-3). Hypothermia can reach 23°C (74°F). There may be a history of treated hypothyroidism with poor compliance, or the patient may be previously undiagnosed. Myxedema coma almost always occurs in the elderly and is usually precipitated by factors that impair respiration, such as drugs (especially sedatives, anesthetics, and antidepressants), pneumonia, congestive heart failure, myocardial infarction, gastrointestinal bleeding, or cerebrovascular accidents. Sepsis should also be suspected. Exposure to cold may also be a risk factor. Hypoventilation, leading to hypoxia and hypercapnia, plays a major role in pathogenesis; hypoglycemia and dilutional hyponatremia also contribute to the development of myxedema coma.

Levothyroxine can initially be administered as a single IV bolus of 200–400 µg, which serves as a loading dose, followed by a daily oral dose of 1.6 µg/kg/d, reduced by 25% if administered IV. If suitable IV preparation is not available, the same initial dose of levothyroxine can be given by nasogastric tube (although absorption may be impaired in myxedema). Because $T_4 \rightarrow T_3$ conversion is impaired in myxedema coma, there is a rationale for adding liothyronine (T_3) intravenously or via nasogastric tube to levothyroxine treatment, although excess liothyronine has the potential to provoke arrhythmias. An initial loading dose of 5–20 µg liothyronine should be followed by 2.5–10 µg 8 hourly, with lower doses for those at cardiovascular risk.

Supportive therapy should be provided to correct any associated metabolic disturbances. External warming is indicated only if the temperature is <30°C, as it can result in cardiovascular collapse **(Chap. 454)**. Space blankets should be used to prevent further heat loss. Parenteral hydrocortisone (50 mg every 6 h) should be administered, because there is impaired adrenal reserve in profound hypothyroidism. Any precipitating factors should be treated, including the early use of broad-spectrum antibiotics, pending the exclusion of infection. Ventilatory support with regular blood gas analysis is usually needed during the first 48 h. Hypertonic saline or IV glucose may be needed if there is severe hyponatremia or hypoglycemia; hypotonic IV fluids should be avoided because they may exacerbate water retention secondary to reduced renal perfusion and inappropriate vasopressin secretion. The metabolism of most medications is impaired, and sedatives should be avoided if possible or used in reduced doses. Medication blood levels should be monitored, when available, to guide dosage.

FURTHER READING

Hanley P et al: Thyroid disorders in children and adolescents: A review. JAMA Pediatr 170:1008, 2016. [PubMed: 27571216]

Jonklaas J et al: Guidelines for the treatment of hypothyroidism: Prepared by the American Thyroid Association Task Force on thyroid hormone replacement. Thyroid 24:1670, 2014. [PubMed: 25266247]

Rugge JB, Bougatsos C, Chou R: Screening and treatment of thyroid dysfunction: An evidence review for the U. S. Preventive Services Task Force. Ann Intern Med 162:35, 2015. [PubMed: 25347444]

Chapter 377: Hyperthyroidism

J. Larry Jameson; Susan J. Mandel; Anthony P. Weetman

THYROTOXICOSIS

Thyrotoxicosis is defined as the state of thyroid hormone excess and is not synonymous with *hyperthyroidism*, which is the result of excessive thyroid function. However, the major etiologies of thyrotoxicosis are hyperthyroidism caused by Graves' disease, toxic multinodular goiter (MNG), and toxic adenomas. Other causes are listed in **Table 377-1**.

TABLE 377-1

Causes of Thyrotoxicosis

Primary Hyperthyroidism
Graves' disease
Toxic multinodular goiter
Toxic adenoma
Functioning thyroid carcinoma metastases
Activating mutation of the TSH receptor
Activating mutation of $G_{S\alpha}$ (McCune-Albright syndrome)
Struma ovarii
Drugs: iodine excess (Jod-Basedow phenomenon)
Thyrotoxicosis without Hyperthyroidism
Subacute thyroiditis
Silent thyroiditis
Other causes of thyroid destruction: amiodarone, radiation, infarction of adenoma
Ingestion of excess thyroid hormone (thyrotoxicosis factitia) or thyroid tissue
Secondary Hyperthyroidism
TSH-secreting pituitary adenoma
Thyroid hormone resistance syndrome: occasional patients may have features of thyrotoxicosis
Chorionic gonadotropin-secreting tumors[a]
Gestational thyrotoxicosis[a]

[a]Circulating TSH levels are low in these forms of secondary hyperthyroidism.

Abbreviation: TSH, thyroid-stimulating hormone.

GRAVES' DISEASE

Epidemiology

Graves' disease accounts for 60–80% of thyrotoxicosis. The prevalence varies among populations, reflecting genetic factors and iodine intake (high iodine intake is associated with an increased prevalence of Graves' disease). Graves' disease occurs in up to 2% of women but is one-tenth as frequent in men. The disorder rarely begins before adolescence and typically occurs between 20 and 50 years of age; it also occurs in the elderly.

Pathogenesis

As in autoimmune hypothyroidism, a combination of environmental and genetic factors, including polymorphisms in HLA-DR, the immunoregulatory genes *CTLA-4*, *CD25*, *PTPN22*, *FCRL3*, and *CD226*, as well as the gene encoding the thyroid-stimulating hormone receptor (TSH-R), contributes to Graves' disease susceptibility. The concordance for Graves' disease in monozygotic twins is 20–30%, compared to <5% in dizygotic twins. Indirect evidence suggests that stress is an important environmental factor, presumably operating through neuroendocrine effects on the immune system. Smoking is a minor risk factor for Graves' disease and a major risk factor for the development of ophthalmopathy. Sudden increases in iodine intake may precipitate Graves' disease, and there is a threefold increase in the occurrence of Graves' disease in the postpartum period. Graves' disease may occur during the immune reconstitution phase after highly active antiretroviral therapy (HAART) or alemtuzumab treatment.

The hyperthyroidism of Graves' disease is caused by thyroid-stimulating immunoglobulin (TSI) that are synthesized in the thyroid gland as well as in bone marrow and lymph nodes. Such antibodies can be detected by bioassays or by using the more widely available thyrotropin-binding inhibitory immunoglobulin (TBII) assays. The presence of TBII in a patient with thyrotoxicosis implies the existence of TSI, and these assays are useful in monitoring pregnant Graves' patients in whom high levels of TSI can cross the placenta and cause neonatal thyrotoxicosis. Other thyroid autoimmune responses, similar to those in autoimmune hypothyroidism (see above), occur concurrently in patients with Graves' disease. In particular, thyroid peroxidase (TPO) and thyroglobulin (Tg) antibodies occur in up to 80% of cases. Because the coexisting thyroiditis can also affect thyroid function, there is no direct correlation between the level of TSI and thyroid hormone levels in Graves' disease.

Cytokines appear to play a major role in thyroid-associated ophthalmopathy. There is infiltration of the extraocular muscles by activated T cells; the release of cytokines such as interferon γ (IFN-γ), tumor necrosis factor (TNF), and interleukin-1 (IL-1) results in fibroblast activation and increased synthesis of glycosaminoglycans that trap water, thereby leading to characteristic muscle swelling. Late in the disease,

there is irreversible fibrosis of the muscles. Though the pathogenesis of thyroid-associated ophthalmopathy remains unclear, there is mounting evidence that the TSH-R is a shared autoantigen that is expressed in the orbit; this would explain the close association with autoimmune thyroid disease. Increased fat is an additional cause of retrobulbar tissue expansion. The increase in intraorbital pressure can lead to proptosis, diplopia, and optic neuropathy.

Clinical Manifestations

Signs and symptoms include features that are common to any cause of thyrotoxicosis (Table 377-2) as well as those specific for Graves' disease. The clinical presentation depends on the severity of thyrotoxicosis, the duration of disease, individual susceptibility to excess thyroid hormone, and the patient's age. In the elderly, features of thyrotoxicosis may be subtle or masked, and patients may present mainly with fatigue and weight loss, a condition known as *apathetic thyrotoxicosis*.

TABLE 377-2

Signs and Symptoms of Thyrotoxicosis (Descending Order of Frequency)

SYMPTOMS	SIGNS[a]
Hyperactivity, irritability, dysphoria	Tachycardia; atrial fibrillation in the elderly
Heat intolerance and sweating	Tremor
Palpitations	Goiter
Fatigue and weakness	Warm, moist skin
Weight loss with increased appetite	Muscle weakness, proximal myopathy
Diarrhea	Lid retraction or lag
Polyuria	Gynecomastia
Oligomenorrhea, loss of libido	

[a]Excludes the signs of ophthalmopathy and dermopathy specific for Graves' disease.

Thyrotoxicosis may cause unexplained weight loss, despite an enhanced appetite, due to the increased metabolic rate. Weight gain occurs in 5% of patients, however, because of increased food intake. Other prominent features include hyperactivity, nervousness, and irritability, ultimately leading to a sense of easy fatigability in some patients. Insomnia and impaired concentration are common; apathetic thyrotoxicosis may be mistaken for depression in the elderly. Fine tremor is a frequent finding, best elicited by having patients stretch out their fingers while feeling the fingertips with the palm. Common neurologic manifestations include hyperreflexia, muscle wasting, and proximal myopathy without fasciculation. Chorea is rare. Thyrotoxicosis is sometimes associated with a form of hypokalemic periodic paralysis; this disorder is particularly common in Asian males with thyrotoxicosis, but it occurs in other ethnic groups as well.

The most common cardiovascular manifestation is sinus tachycardia, often associated with palpitations, occasionally caused by supraventricular tachycardia. The high cardiac output produces a bounding pulse, widened pulse pressure, and an aortic systolic murmur and can lead to worsening of angina or heart failure in the elderly or those with preexisting heart disease. Atrial fibrillation is more common in patients >50 years of age. Treatment of the thyrotoxic state alone converts atrial fibrillation to normal sinus rhythm in about half of patients, suggesting the existence of an underlying cardiac problem in the remainder.

The skin is usually warm and moist, and the patient may complain of sweating and heat intolerance, particularly during warm weather. Palmar erythema, onycholysis, and, less commonly, pruritus, urticaria, and diffuse hyperpigmentation may be evident. Hair texture may become fine, and a diffuse alopecia occurs in up to 40% of patients, persisting for months after restoration of euthyroidism. Gastrointestinal transit time is decreased, leading to increased stool frequency, often with diarrhea and occasionally mild steatorrhea. Women frequently experience oligomenorrhea or amenorrhea; in men, there may be impaired sexual function and, rarely, gynecomastia. The direct effect of thyroid hormones on bone resorption leads to osteopenia in long-standing thyrotoxicosis; mild hypercalcemia occurs in up to 20% of patients, but hypercalciuria is more common. There is a small increase in fracture rate in patients with a previous history of thyrotoxicosis.

In Graves' disease, the thyroid is usually diffusely enlarged to two to three times its normal size. The consistency is firm, but not nodular. There may be a thrill or bruit, best detected at the inferolateral margins of the thyroid lobes, due to the increased vascularity of the gland and the hyperdynamic circulation.

Lid retraction, causing a staring appearance, can occur in any form of thyrotoxicosis and is the result of sympathetic overactivity. However, Graves' disease is associated with specific eye signs that comprise *Graves' ophthalmopathy* **(Fig. 377-1A).** This condition is also called *thyroid-associated ophthalmopathy*, because it occurs in the absence of hyperthyroidism in 10% of patients. Most of these individuals have autoimmune hypothyroidism or thyroid antibodies. The onset of Graves' ophthalmopathy occurs within the year before or after the diagnosis of thyrotoxicosis in 75% of patients but can sometimes precede or follow thyrotoxicosis by several years, accounting for some cases of euthyroid ophthalmopathy.

FIGURE 377-1

Features of Graves' disease. ***A.*** Ophthalmopathy in Graves' disease; lid retraction, periorbital edema, conjunctival injection, and proptosis are marked. ***B.*** Thyroid dermopathy over the lateral aspects of the shins. ***C.*** Thyroid acropachy.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Some patients with Graves' disease have little clinical evidence of ophthalmopathy. However, the enlarged extraocular muscles typical of the disease, and other subtle features, can be detected in most patients when investigated by ultrasound or computed tomography (CT) imaging of the orbits. Unilateral signs are found in up to 10% of patients. The earliest manifestations of ophthalmopathy are usually a sensation of grittiness, eye discomfort, and excess tearing. About one-third of patients have proptosis, best detected by visualization of the sclera between the lower border of the iris and the lower eyelid, with the eyes in the primary position. Proptosis can be measured using an exophthalmometer. In severe cases, proptosis may cause corneal exposure and damage, especially if the lids fail to close during sleep. Periorbital edema, scleral injection, and chemosis are also frequent. In 5–10% of patients, the muscle swelling is so severe that diplopia results, typically, but not exclusively, when the patient looks up and laterally. The most serious manifestation is compression of the optic nerve at the apex of the orbit, leading to papilledema; peripheral field defects; and, if left untreated, permanent loss of vision.

The "NO SPECS" scoring system to evaluate ophthalmopathy is an acronym derived from the following changes:

0 = **N**o signs or symptoms

1 = **O**nly signs (lid retraction or lag), no symptoms

2 = **S**oft tissue involvement (periorbital edema)

3 = **P**roptosis (>22 mm)

4 = **E**xtraocular muscle involvement (diplopia)

5 = **C**orneal involvement

6 = **S**ight loss

Although useful as a mnemonic, the NO SPECS scheme is inadequate to describe the eye disease fully, and patients do not necessarily progress from one class to another; alternative scoring systems (e.g., the EUGOGO system developed by the European Group On Graves' Orbitopathy) that assess disease activity are preferable for monitoring and treatment purposes. When Graves' eye disease is active and severe, referral to an ophthalmologist is indicated and objective measurements are needed, such as lid-fissure width; corneal staining with fluorescein; and evaluation of extraocular muscle function (e.g., Hess chart), intraocular pressure and visual fields, acuity, and color vision.

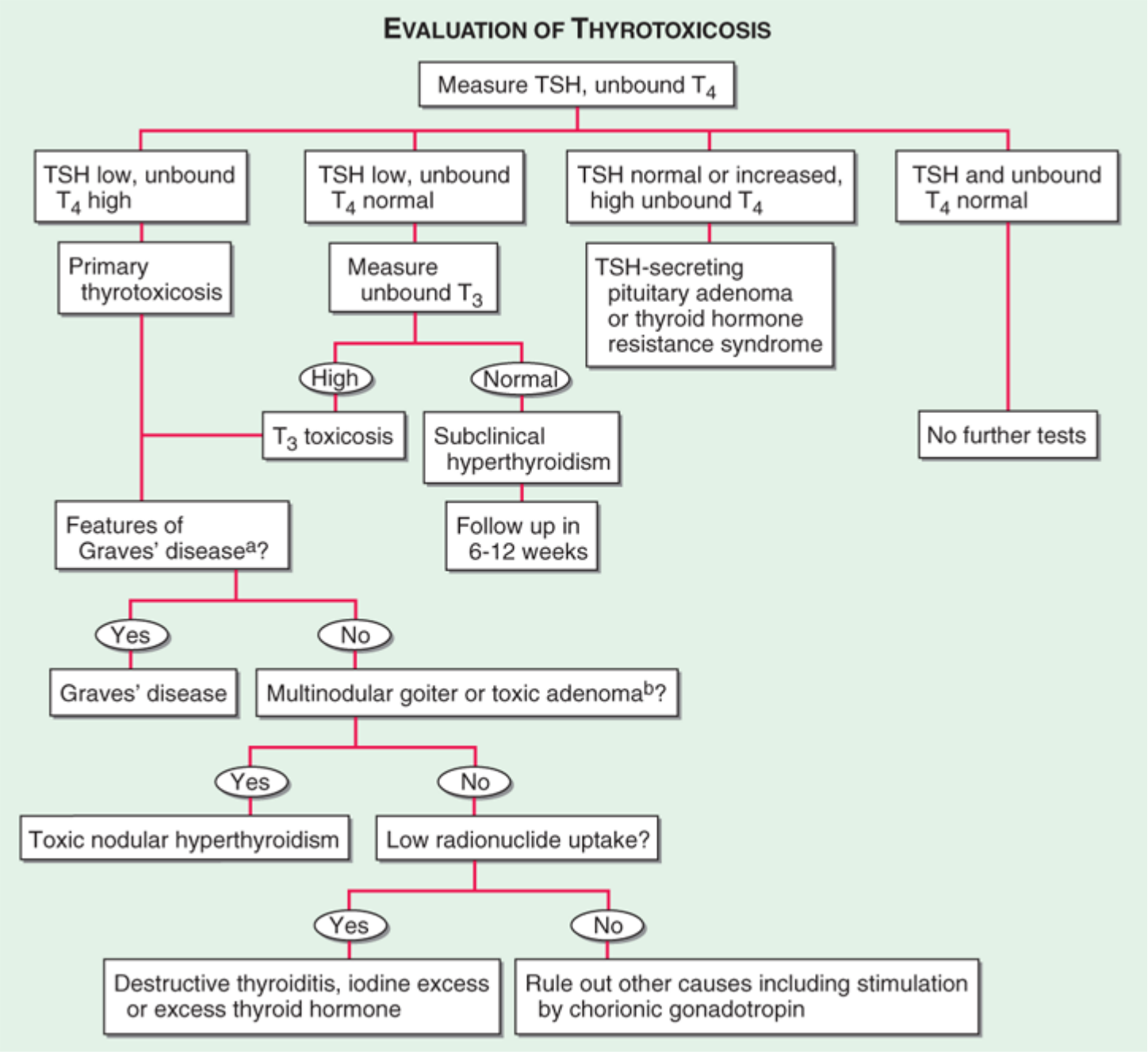
Thyroid dermopathy occurs in <5% of patients with Graves' disease **(Fig. 377-1B)**, almost always in the presence of moderate or severe ophthalmopathy. Although most frequent over the anterior and lateral aspects of the lower leg (hence the term *pretibial myxedema*), skin changes can occur at other sites, particularly after trauma. The typical lesion is a noninflamed, indurated plaque with a deep pink or purple color and an "orange skin" appearance. Nodular involvement can occur, and the condition can rarely extend over the whole lower leg and foot, mimicking elephantiasis. *Thyroid acropachy* refers to a form of clubbing found in <1% of patients with Graves' disease **(Fig. 377-1C)**. It is so strongly associated with thyroid dermopathy that an alternative cause of clubbing should be sought in a Graves' patient without coincident skin and orbital involvement.

Laboratory Evaluation

Investigations used to determine the existence and cause of thyrotoxicosis are summarized in **Fig. 377-2**. In Graves' disease, the TSH level is suppressed, and total and unbound thyroid hormone levels are increased. In 2–5% of patients (and more in areas of borderline iodine intake), only T_3 is increased (T_3 toxicosis). The converse state of T_4 toxicosis, with elevated total and unbound T_4 and normal T_3 levels, is occasionally seen when hyperthyroidism is induced by excess iodine, providing surplus substrate for thyroid hormone synthesis. Measurement of TPO antibodies or TBII may be useful if the diagnosis is unclear clinically but is not needed routinely. Associated abnormalities that may cause diagnostic confusion in thyrotoxicosis include elevation of bilirubin, liver enzymes, and ferritin. Microcytic anemia and thrombocytopenia may occur.

FIGURE 377-2

Evaluation of thyrotoxicosis. [a]Diffuse goiter, positive TPO antibodies or TRAb, ophthalmopathy, dermopathy. [b]Can be confirmed by radionuclide scan. TSH, thyroid-stimulating hormone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Differential Diagnosis

Diagnosis of Graves' disease is straightforward in a patient with biochemically confirmed thyrotoxicosis, diffuse goiter on palpation, ophthalmopathy, and often a personal or family history of autoimmune disorders. For patients with thyrotoxicosis who lack these features, the diagnosis is generally established by a radionuclide (^{99m}Tc, ^{123}I, or ^{131}I) scan and uptake of the thyroid, which will distinguish the diffuse, high uptake of Graves' disease from destructive thyroiditis, ectopic thyroid tissue, and factitious thyrotoxicosis, as

well as diagnosing a toxic adenoma or toxic MNG. Alternatively, TRAb measurement can be used to diagnose Graves' disease and color-flow Doppler ultrasonography may distinguish between hyperthyroidism (with increased blood flow) and destructive thyroiditis. In secondary hyperthyroidism due to a TSH-secreting pituitary tumor, there is also a diffuse goiter. The presence of a nonsuppressed TSH level and the finding of a pituitary tumor on CT or magnetic resonance imaging (MRI) scan suggest this diagnosis.

Clinical features of thyrotoxicosis can mimic certain aspects of other disorders, including panic attacks, mania, pheochromocytoma, and weight loss associated with malignancy. The diagnosis of thyrotoxicosis can be easily excluded if the TSH and unbound T_4 and T_3 levels are normal. A normal TSH also excludes Graves' disease as a cause of diffuse goiter.

Clinical Course

Clinical features generally worsen without treatment; mortality was 10–30% before the introduction of satisfactory therapy. Some patients with mild Graves' disease experience spontaneous relapses and remissions. Rarely, there may be fluctuation between hypo- and hyperthyroidism due to changes in the functional activity of TSH-R antibodies. About 15% of patients who enter remission after treatment develop hypothyroidism 10–15 years later as a result of the destructive autoimmune process.

The clinical course of ophthalmopathy does not follow that of the thyroid disease, although thyroid dysfunction can worsen eye signs. Ophthalmopathy typically worsens over the initial 3–6 months, followed by a plateau phase over the next 12–18 months, and then some spontaneous improvement, particularly in the soft tissue changes. However, the course is more fulminant in up to 5% of patients, requiring intervention in the acute phase if there is optic nerve compression or corneal ulceration. Diplopia may appear late in the disease due to fibrosis of the extraocular muscles. Radioiodine treatment for hyperthyroidism worsens the eye disease in a small proportion of patients (especially smokers). Antithyroid drugs or surgery have no adverse effects on the clinical course of ophthalmopathy. Thyroid dermopathy, when it occurs, usually appears 1–2 years after the development of Graves' hyperthyroidism; it may improve spontaneously.

TREATMENT

TREATMENT

Graves' Disease

The *hyperthyroidism* of Graves' disease is treated by reducing thyroid hormone synthesis, using an antithyroid drug, or reducing the amount of thyroid tissue with radioiodine (^{131}I) treatment or by thyroidectomy. Antithyroid drugs are the predominant therapy in many centers in Europe, Latin America, and Japan, whereas radioiodine is more often the first line of treatment in North America. These differences reflect the fact that no single approach is optimal and that patients may require multiple treatments to achieve remission.

The main *antithyroid drugs* are thionamides; propylthiouracil, carbimazole (not available in the United States), and the active metabolite of the latter, methimazole. All inhibit the function of TPO, reducing oxidation and organification of iodide. These drugs also reduce thyroid antibody levels by mechanisms that remain unclear, and they appear to enhance spontaneous rates of remission. Propylthiouracil inhibits deiodination of $T_4 \rightarrow T_3$. However, this effect is of minor benefit, except in the most severe thyrotoxicosis, and is offset by the much shorter half-life of this drug (90 min) compared to methimazole (6 h). Due to the hepatotoxicity of propylthiouracil, the U.S. Food and Drug Administration (FDA) has limited indications for its use to the first trimester of pregnancy, the treatment of thyroid storm, and patients with minor adverse reactions to methimazole. If propylthiouracil is used, monitoring of liver function tests is recommended.

There are many variations of antithyroid drug regimens. The initial dose of carbimazole or methimazole is usually 10–20 mg every 8 or 12 h, but once-daily dosing is possible after euthyroidism is restored. Propylthiouracil is given at a dose of 100–200 mg every 6–8 h, and divided doses are usually given throughout the course. Lower doses of each drug may suffice in areas of low iodine intake. The starting dose of an antithyroid drug can be gradually reduced (titration regimen) as thyrotoxicosis improves. Less commonly, high doses may be given combined with levothyroxine supplementation (block-replace regimen) to avoid drug-induced hypothyroidism. The titration regimen is preferred to minimize the dose of antithyroid drug and provide an index of treatment response.

Thyroid function tests and clinical manifestations are reviewed 4–6 weeks after starting treatment, and the dose is titrated based on unbound T_4 levels. Most patients do not achieve euthyroidism until 6–8 weeks after treatment is initiated. TSH levels often remain suppressed for several months and therefore do not provide a sensitive index of treatment response. The usual daily maintenance doses of antithyroid drugs in the titration regimen are 2.5–10 mg of carbimazole or methimazole and 50–100 mg of propylthiouracil. In the block-replace regimen, the initial dose of antithyroid drug is held constant, and the dose of levothyroxine is adjusted to maintain normal unbound T_4 levels. When TSH suppression is alleviated, TSH levels can also be used to monitor therapy.

Maximum remission rates (up to 30–60% in some populations) are achieved by 12–18 months for the titration regimen and are higher in patients where TRAb levels are no longer detected, than in those with TRAb persistence. For unclear reasons, remission rates appear to vary in different geographic regions. Younger patients, males, smokers, and patients with a history of allergy, severe hyperthyroidism or large goiters are most likely to relapse when treatment stops, but outcomes are difficult to predict. All patients should be followed closely for relapse during the first year after treatment and at least annually thereafter.

The common minor side effects of antithyroid drugs are rash, urticaria, fever, and arthralgia (1–5% of patients). These may resolve spontaneously or after substituting an alternative antithyroid drug; rashes may respond to an antihistamine. Rare but major side effects include hepatitis (especially with propylthiouracil; avoid use in children) and cholestasis (methimazole and carbimazole); vasculitis; and, most important, agranulocytosis (<1%). It is essential that antithyroid drugs are stopped and not restarted if a patient develops major side effects. Written instructions should be provided regarding the symptoms of possible

agranulocytosis (e.g., sore throat, fever, mouth ulcers) and the need to stop treatment pending an urgent complete blood count to confirm that agranulocytosis is not present. Management of agranulocytosis is described in Chap. 98. It is not useful to monitor blood counts prospectively, because the onset of agranulocytosis is idiosyncratic and abrupt.

Propranolol (20–40 mg every 6 h) or longer-acting selective β_1 receptor blockers such as atenolol may be helpful to control adrenergic symptoms, especially in the early stages before antithyroid drugs take effect. Beta blockers are also useful in patients with thyrotoxic periodic paralysis, pending correction of thyrotoxicosis. In consultation with a cardiologist, anticoagulation with warfarin should be considered in all patients with atrial fibrillation; there is often spontaneous reversion to sinus rhythm with control of hyperthyroidism, and long-term anticoagulation is not usually needed. Decreased warfarin doses are required when patients are thyrotoxic. If digoxin is used, increased doses are often needed in the thyrotoxic state.

Radioiodine causes progressive destruction of thyroid cells and can be used as initial treatment or for relapses after a trial of antithyroid drugs. There is a small risk of thyrotoxic crisis (see below) after radioiodine, which can be minimized by pretreatment with antithyroid drugs for at least a month before treatment. Antecedent treatment with an antithyroid drug and a beta blocker should be considered for all elderly patients or for those with cardiac problems. Carbimazole or methimazole must be stopped 2–3 days before radioiodine administration to achieve optimum iodine uptake, and can be restarted 3–7 days after radioiodine in those at risk of complications from worsening thyrotoxicosis. Propylthiouracil appears to have a prolonged radioprotective effect and should be stopped for a longer period before radioiodine is given, or a larger dose of radioiodine will be necessary.

Efforts to calculate an optimal dose of radioiodine that achieves euthyroidism without a high incidence of relapse or progression to hypothyroidism have not been successful. Some patients inevitably relapse after a single dose because the biologic effects of radiation vary between individuals, and hypothyroidism cannot be uniformly avoided even using accurate dosimetry. A practical strategy is to give a fixed dose based on clinical features, such as the severity of thyrotoxicosis, the size of the goiter (increases the dose needed), and the level of radioiodine uptake (decreases the dose needed). ^{131}I dosage generally ranges between 370 MBq (10 mCi) and 555 MBq (15 mCi). Most authorities favor an approach aimed at thyroid ablation (as opposed to euthyroidism), given that levothyroxine replacement is straightforward and most patients ultimately progress to hypothyroidism over 5–10 years, frequently with some delay in the diagnosis of hypothyroidism.

Certain radiation safety precautions are necessary in the first few days after radioiodine treatment, but the exact guidelines vary depending on local protocols. In general, patients need to avoid close, prolonged contact with children and pregnant women for 5–7 days because of possible transmission of residual isotope and exposure to radiation emanating from the gland. Rarely, there may be mild pain due to radiation thyroiditis 1–2 weeks after treatment. Hyperthyroidism can persist for 2–3 months before radioiodine takes full effect. For this reason, β-adrenergic blockers or antithyroid drugs can be used to control symptoms during this interval. Persistent hyperthyroidism can be treated with a second dose of radioiodine, usually 6

months after the first dose. The risk of hypothyroidism after radioiodine depends on the dosage but is at least 10–20% in the first year and 5% per year thereafter. Patients should be informed of this possibility before treatment and require close follow-up during the first year followed by annual thyroid function testing.

Pregnancy and breast-feeding are absolute contraindications to radioiodine treatment, but patients can conceive safely 6 months after treatment. The presence of ophthalmopathy, especially in smokers, requires caution. Prednisone, 30 mg/d, at the time of radioiodine treatment, tapered over 6–8 weeks may prevent exacerbation of ophthalmopathy, but radioiodine should generally be avoided in those with active moderate to severe eye disease. The overall risk of cancer after radioiodine treatment in adults is not increased. Although many physicians avoid radioiodine in children and adolescents because of the theoretical risks of malignancy, emerging evidence suggests that radioiodine can be used safely in older children.

Total or near-total thyroidectomy is an option for patients who relapse after antithyroid drugs and prefer this treatment to radioiodine. Some experts recommend surgery in young individuals, particularly when the goiter is very large. Careful control of thyrotoxicosis with antithyroid drugs, followed by potassium iodide (1–2 drops SSKI orally tid for 10 days), is needed prior to surgery to avoid thyrotoxic crisis and to reduce the vascularity of the gland. The major complications of surgery—bleeding, laryngeal edema, hypoparathyroidism, and damage to the recurrent laryngeal nerves—are unusual when the procedure is performed by highly experienced surgeons. Recurrence rates in the best series are <2%, but the rate of hypothyroidism is similar to that following radioiodine treatment, especially with the current trend away from subtotal thyroidectomy.

Antithyroid drugs should be used to manage Graves' disease in *pregnancy.* Because transplacental passage of these drugs may produce fetal hypothyroidism and goiter if the maternal dose is excessive, maternal antithyroid dose titration should target serum free or total T_4 levels at or just above the pregnancy reference range. If available, propylthiouracil should be used until 14–16 weeks' gestation because of the association of rare cases of methimazole/carbimazole embryopathy, including *aplasia cutis* and other defects, such as choanal atresia and tracheoesophageal fistulae. Because of the potential for teratogenic effects, recent recommendations suggest discontinuation of antithyroid medication in a newly pregnant woman with Graves' disease, who is euthyroid on a low dose of methimazole (<5–10 mg/day) or PTU (<100–200 mg/day), after evaluating recent thyroid function tests, disease history, goiter size, duration of therapy, and TRAb measurement. Following cessation, careful monitoring of maternal thyroid function tests is essential. On the other hand, for women at high risk of developing thyrotoxicosis if antithyroid drugs are discontinued (large goiter, requirement for higher antithyroid drug dosage), continued therapy is necessary, with PTU (if available) administration in the first trimester. But, because of its rare association with hepatotoxicity, propylthiouracil should be limited to the first trimester and then maternal therapy should be converted to methimazole (or carbimazole) at a ratio of 15–20 mg of propylthiouracil to 1 mg of methimazole. It is often possible to stop treatment in the last trimester because TSIs tend to decline in pregnancy. Nonetheless, the transplacental transfer of these antibodies if present at levels 3 times higher than the normative range rarely causes *fetal* or *neonatal thyrotoxicosis*. Poor intrauterine growth, a fetal heart rate of >160 beats/min,

advanced bone age, fetal goiter, and high levels of maternal TSI after 26 weeks gestation may herald this complication. Antithyroid drugs given to the mother can be used to treat the fetus and may be needed for 1–3 months after delivery, until the maternal antibodies disappear from the baby's circulation. The postpartum period is a time of major risk for relapse of Graves' disease. Breast-feeding is safe with low doses of antithyroid drugs. Graves' disease in *children* is usually managed initially with methimazole or carbimazole (avoid propylthiouracil), often given as a prolonged course of the titration regimen. Surgery or radioiodine may be indicated for severe or relapsing disease.

Thyrotoxic crisis, or *thyroid storm*, is rare and presents as a life-threatening exacerbation of hyperthyroidism, accompanied by fever, delirium, seizures, coma, vomiting, diarrhea, and jaundice. The mortality rate due to cardiac failure, arrhythmia, or hyperthermia is as high as 30%, even with treatment. Thyrotoxic crisis is usually precipitated by acute illness (e.g., stroke, infection, trauma, diabetic ketoacidosis), surgery (especially on the thyroid), or radioiodine treatment of a patient with partially treated or untreated hyperthyroidism. Management requires intensive monitoring and supportive care, identification and treatment of the precipitating cause, and measures that reduce thyroid hormone synthesis. Large doses of propylthiouracil (500–1000 mg loading dose and 250 mg every 4 h) should be given orally or by nasogastric tube or per rectum; the drug's inhibitory action on $T_4 \rightarrow T_3$ conversion makes it the antithyroid drug of choice. If not available, methimazole can be used in doses of 20 mg every 6 h. One hour after the first dose of propylthiouracil, stable iodide (5 drops SSKI every 6 h) is given to block thyroid hormone synthesis via the Wolff-Chaikoff effect (the delay allows the antithyroid drug to prevent the excess iodine from being incorporated into new hormone). Propranolol should also be given to reduce tachycardia and other adrenergic manifestations (60–80 mg PO every 4 h; or 2 mg IV every 4 h). Although other β-adrenergic blockers can be used, high doses of propranolol decrease $T_4 \rightarrow T_3$ conversion, and the doses can be easily adjusted. Caution is needed to avoid acute negative inotropic effects, but controlling the heart rate is important, as some patients develop a form of high-output heart failure. Short-acting IV esmolol can be used to decrease heart rate while monitoring for signs of heart failure. Additional therapeutic measures include glucocorticoids (e.g., hydrocortisone 300 mg IV bolus, then 100 mg every 8 h), antibiotics if infection is present, cholestyramine to sequester thyroid hormones, cooling, oxygen, and IV fluids.

Ophthalmopathy requires no active treatment when it is mild or moderate, because there is usually spontaneous improvement. General measures include meticulous control of thyroid hormone levels, cessation of smoking, and an explanation of the natural history of ophthalmopathy. Discomfort can be relieved with artificial tears (e.g., hypromellose 0.3% or carbomer 0.2% ophthalmic gel) paraffin-based eye ointment, and the use of dark glasses with side frames. Periorbital edema may respond to a more upright sleeping position or a diuretic. Corneal exposure during sleep can be avoided by using patches or taping the eyelids shut. Minor degrees of diplopia improve with prisms fitted to spectacles. Some authorities also advocate selenium 100 µg bd. Severe ophthalmopathy, with optic nerve involvement or chemosis resulting in corneal damage, is an emergency requiring joint management with an ophthalmologist. Pulse therapy with IV methylprednisolone (e.g., 500 mg of methylprednisolone once weekly for 6 weeks, then 250 mg once weekly for 6 weeks) is preferable to oral glucocorticoids, which are used for moderately active disease. When glucocorticoids are ineffective, orbital decompression can be achieved by removing bone from any wall of

the orbit, thereby allowing displacement of fat and swollen extraocular muscles. The transantral route is used most often because it requires no external incision. Proptosis recedes an average of 5 mm, but there may be residual or even worsened diplopia. Once the eye disease has stabilized, surgery may be indicated for relief of diplopia and correction of the appearance. External beam radiotherapy of the orbits has been used for many years, but the efficacy of this therapy remains unclear, and it is best reserved for those with moderately active disease who have failed or are not candidates for glucocorticoid therapy. Other immunosuppressive agents such as rituximab have shown some benefit, but their role is yet to be established.

Thyroid dermopathy does not usually require treatment, but it can cause cosmetic problems or interfere with the fit of shoes. Surgical removal is not indicated. If necessary, treatment consists of topical, high-potency glucocorticoid ointment under an occlusive dressing. Octreotide may be beneficial in some cases.

OTHER CAUSES OF THYROTOXICOSIS

Destructive thyroiditis (subacute or silent thyroiditis) typically presents with a short thyrotoxic phase due to the release of preformed thyroid hormones and catabolism of Tg (see "Subacute Thyroiditis," below). True hyperthyroidism is absent, as demonstrated by a low radionuclide uptake. Circulating Tg levels are typically increased. Other causes of thyrotoxicosis with low or absent thyroid radionuclide uptake include *thyrotoxicosis factitia*, iodine excess, and, rarely, ectopic thyroid tissue, particularly teratomas of the ovary (*struma ovarii*) and functional metastatic follicular carcinoma. Whole-body radionuclide studies can demonstrate ectopic thyroid tissue, and thyrotoxicosis factitia can be distinguished from destructive thyroiditis by the clinical features and low levels of Tg. Amiodarone treatment is associated with thyrotoxicosis in up to 10% of patients, particularly in areas of low iodine intake (see below).

TSH-secreting pituitary adenoma is a rare cause of thyrotoxicosis. It is characterized by the presence of an inappropriately normal or increased TSH level in a patient with hyperthyroidism, diffuse goiter, and elevated T_4 and T_3 levels **(Chap. 373)**. Elevated levels of the α-subunit of TSH, released by the TSH-secreting adenoma, support this diagnosis, which can be confirmed by demonstrating the pituitary tumor on MRI or CT scan. A combination of transsphenoidal surgery, sella irradiation, and octreotide may be required to normalize TSH, because many of these tumors are large and locally invasive at the time of diagnosis. Radioiodine or antithyroid drugs can be used to control thyrotoxicosis.

Thyrotoxicosis caused by *toxic MNG* and *hyperfunctioning solitary nodules* is discussed below.

THYROIDITIS

A clinically useful classification of thyroiditis is based on the onset and duration of disease **(Table 377-3)**.

TABLE 377-3

Causes of Thyroiditis

Acute
Bacterial infection: especially *Staphylococcus, Streptococcus,* and *Enterobacter*
Fungal infection: *Aspergillus, Candida, Coccidioides, Histoplasma,* and *Pneumocystis*
Radiation thyroiditis after ^{131}I treatment
Amiodarone (may also be subacute or chronic)
Subacute
Viral (or granulomatous) thyroiditis
Silent thyroiditis (including postpartum thyroiditis)
Mycobacterial infection
Drug induced (interferon, amiodarone)
Chronic
Autoimmunity: focal thyroiditis, Hashimoto's thyroiditis, atrophic thyroiditis
Riedel's thyroiditis
Parasitic thyroiditis: echinococcosis, strongyloidiasis, cysticercosis
Traumatic: after palpation

ACUTE THYROIDITIS

Acute thyroiditis is rare and due to suppurative infection of the thyroid. In children and young adults, the most common cause is the presence of a piriform sinus, a remnant of the fourth branchial pouch that connects the oropharynx with the thyroid. Such sinuses are predominantly left-sided. A long-standing goiter and degeneration in a thyroid malignancy are risk factors in the elderly. The patient presents with thyroid pain, often referred to the throat or ears, and a small, tender goiter that may be asymmetric. Fever,

dysphagia, and erythema over the thyroid are common, as are systemic symptoms of a febrile illness and lymphadenopathy.

The differential diagnosis of *thyroid pain* includes subacute or, rarely, chronic thyroiditis; hemorrhage into a cyst; malignancy including lymphoma; and, rarely, amiodarone-induced thyroiditis or amyloidosis. However, the abrupt presentation and clinical features of acute thyroiditis rarely cause confusion. The erythrocyte sedimentation rate (ESR) and white cell count are usually increased, but thyroid function is normal. Fine-needle aspiration (FNA) biopsy shows infiltration by polymorphonuclear leukocytes; culture of the sample can identify the organism. Caution is needed in immunocompromised patients as fungal, mycobacterial, or *Pneumocystis* thyroiditis can occur in this setting. Antibiotic treatment is guided initially by Gram stain and, subsequently, by cultures of the FNA biopsy. Surgery may be needed to drain an abscess, which can be localized by CT scan or ultrasound. Tracheal obstruction, septicemia, retropharyngeal abscess, mediastinitis, and jugular venous thrombosis may complicate acute thyroiditis but are uncommon with prompt use of antibiotics.

SUBACUTE THYROIDITIS

This is also termed *de Quervain's thyroiditis*, *granulomatous thyroiditis*, or *viral thyroiditis*. Many viruses have been implicated, including mumps, coxsackie, influenza, adenoviruses, and echoviruses, but attempts to identify the virus in an individual patient are often unsuccessful and do not influence management. The diagnosis of subacute thyroiditis is often overlooked because the symptoms can mimic pharyngitis. The peak incidence occurs at 30–50 years, and women are affected three times more frequently than men.

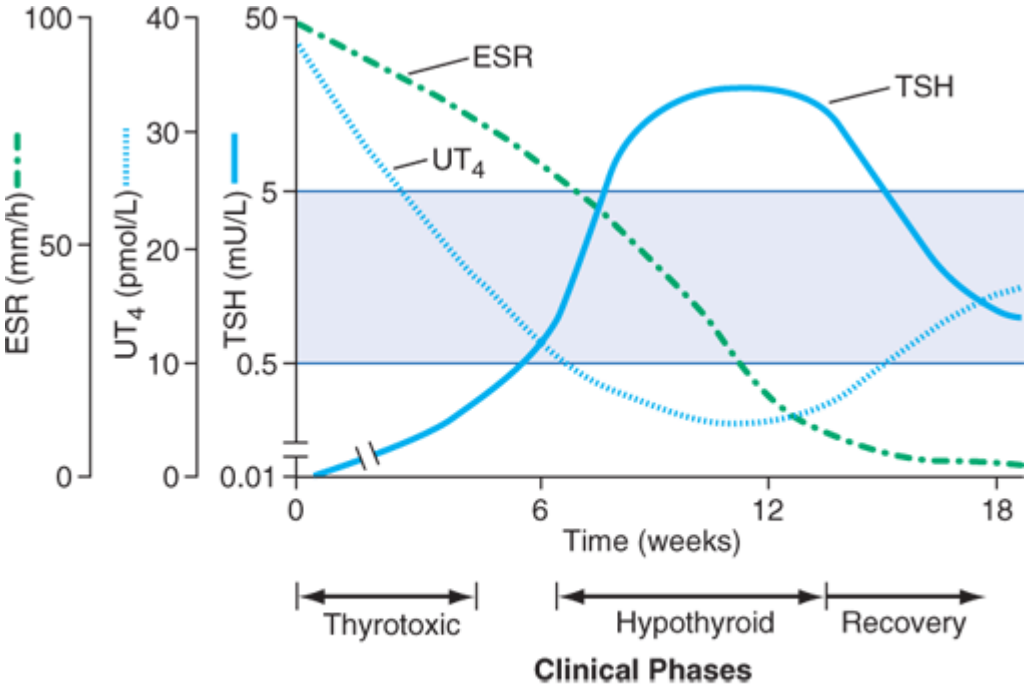
Pathophysiology

The thyroid shows a characteristic patchy inflammatory infiltrate with disruption of the thyroid follicles and multinucleated giant cells within some follicles. The follicular changes progress to granulomas accompanied by fibrosis. Finally, the thyroid returns to normal, usually several months after onset. During the initial phase of follicular destruction, there is release of Tg and thyroid hormones, leading to increased circulating T_4 and T_3 and suppression of TSH **(Fig. 377-3)**. During this destructive phase, radioactive iodine uptake is low or undetectable. After several weeks, the thyroid is depleted of stored thyroid hormone and a phase of hypothyroidism typically occurs, with low unbound T_4 (and sometimes T_3) and moderately increased TSH levels. Radioactive iodine uptake returns to normal or is even increased as a result of the rise in TSH. Finally, thyroid hormone and TSH levels return to normal as the disease subsides.

FIGURE 377-3

Clinical course of subacute thyroiditis. The release of thyroid hormones is initially associated with a thyrotoxic phase and suppressed thyroid-stimulating hormone (TSH). A hypothyroid phase then ensues, with low T_4 and TSH levels that are initially low but gradually increase. During the recovery phase, increased TSH levels combined with resolution of thyroid follicular injury lead to normalization of thyroid function, often

several months after the beginning of the illness. ESR, erythrocyte sedimentation rate; UT_4, free or unbound T_4.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical Manifestations

The patient usually presents with a painful and enlarged thyroid, sometimes accompanied by fever. There may be features of thyrotoxicosis or hypothyroidism, depending on the phase of the illness. Malaise and symptoms of an upper respiratory tract infection may precede the thyroid-related features by several weeks. In other patients, the onset is acute, severe, and without obvious antecedent. The patient typically complains of a sore throat, and examination reveals a small goiter that is exquisitely tender. Pain is often referred to the jaw or ear. Complete resolution is the usual outcome, but late-onset permanent hypothyroidism occurs in 15% of cases, particularly in those with coincidental thyroid autoimmunity. A prolonged course over many months, with one or more relapses, occurs in a small percentage of patients.

Laboratory Evaluation

As depicted in Fig. 377-3, thyroid function tests characteristically evolve through three distinct phases over about 6 months: (1) thyrotoxic phase, (2) hypothyroid phase, and (3) recovery phase. In the thyrotoxic phase, T_4 and T_3 levels are increased, reflecting their discharge from the damaged thyroid cells, and TSH is suppressed. The T_4/T_3 ratio is greater than in Graves' disease or thyroid autonomy, in which T_3 is often disproportionately increased. The diagnosis is confirmed by a high ESR and low uptake of radioiodine (<5%) or ^{99m}Tc pertechnetate (as compared to salivary gland pertechnetate concentration). The white blood cell

count may be increased, and thyroid antibodies are negative. If the diagnosis is in doubt, FNA biopsy may be useful, particularly to distinguish unilateral involvement from bleeding into a cyst or neoplasm.

TREATMENT

TREATMENT

Subacute Thyroiditis

Relatively large doses of aspirin (e.g., 600 mg every 4–6 h) or nonsteroidal anti-inflammatory drugs (NSAIDs) are sufficient to control symptoms in many cases. If this treatment is inadequate, or if the patient has marked local or systemic symptoms, glucocorticoids should be given. The usual starting dose is 15–40 mg of prednisone, depending on severity. The dose is gradually tapered over 6–8 weeks, in response to improvement in symptoms and the ESR. If a relapse occurs during glucocorticoid withdrawal, the dosage should be increased and then withdrawn more gradually. Thyroid function should be monitored every 2–4 weeks using TSH and unbound T_4 levels. Symptoms of thyrotoxicosis improve spontaneously but may be ameliorated by β-adrenergic blockers; antithyroid drugs play no role in treatment of the thyrotoxic phase. Levothyroxine replacement may be needed if the hypothyroid phase is prolonged, but doses should be low enough (50–100 µg daily) to allow TSH-mediated recovery.

SILENT THYROIDITIS

Painless thyroiditis, or "*silent*" *thyroiditis*, occurs in patients with underlying autoimmune thyroid disease and has a clinical course similar to that of subacute thyroiditis. The condition occurs in up to 5% of women 3–6 months after pregnancy and is then termed *postpartum thyroiditis*. Typically, patients have a brief phase of thyrotoxicosis lasting 2–4 weeks, followed by hypothyroidism for 4–12 weeks, and then resolution; often, however, only one phase is apparent. The condition is associated with the presence of TPO antibodies antepartum, and it is three times more common in women with type 1 diabetes mellitus. As in subacute thyroiditis, the uptake of ^{99m}Tc pertechnetate or radioactive iodine is initially suppressed. In addition to the painless goiter, silent thyroiditis can be distinguished from subacute thyroiditis by a normal ESR and the presence of TPO antibodies. Glucocorticoid treatment is not indicated for silent thyroiditis. Severe thyrotoxic symptoms can be managed with a brief course of propranolol, 20–40 mg three or four times daily. Thyroxine replacement may be needed for the hypothyroid phase but should be withdrawn after 6–9 months, as recovery is the rule. Annual follow-up thereafter is recommended, because a proportion of these individuals develop permanent hypothyroidism. The condition may recur in subsequent pregnancies.

DRUG-INDUCED THYROIDITIS

Patients receiving cytokines, such as IFN-α or IL-2, or tyrosine kinase inhibitors may develop painless thyroiditis. IFN-α, which is used to treat chronic hepatitis B or C and hematologic and skin malignancies, causes thyroid dysfunction in up to 5% of treated patients. It has been associated with painless thyroiditis,

hypothyroidism, and Graves' disease, and is most common in women with TPO antibodies prior to treatment. For discussion of amiodarone, see "Amiodarone Effects on Thyroid Function," below.

CHRONIC THYROIDITIS

Focal thyroiditis is present in 20–40% of euthyroid autopsy cases and is associated with serologic evidence of autoimmunity, particularly the presence of TPO antibodies. The most common clinically apparent cause of chronic thyroiditis is *Hashimoto's thyroiditis*, an autoimmune disorder that often presents as a firm or hard goiter of variable size (see above). *Riedel's thyroiditis* is a rare disorder that typically occurs in middle-aged women. It presents with an insidious, painless goiter with local symptoms due to compression of the esophagus, trachea, neck veins, or recurrent laryngeal nerves. Dense fibrosis disrupts normal gland architecture and can extend outside the thyroid capsule. Despite these extensive histologic changes, thyroid dysfunction is uncommon. The goiter is hard, nontender, often asymmetric, and fixed, leading to suspicion of a malignancy. Diagnosis requires open biopsy as FNA biopsy is usually inadequate. Treatment is directed to surgical relief of compressive symptoms. Tamoxifen may also be beneficial. There is an association between Riedel's thyroiditis and IgG4-related disease causing idiopathic fibrosis at other sites (retroperitoneum, mediastinum, biliary tree, lung, and orbit).

SICK EUTHYROID SYNDROME (NONTHYROIDAL ILLNESS)

Any acute, severe illness can cause abnormalities of circulating TSH or thyroid hormone levels in the absence of underlying thyroid disease, making these measurements potentially misleading. The major cause of these hormonal changes is the release of cytokines such as IL-6. Unless a thyroid disorder is strongly suspected, the routine testing of thyroid function should be avoided in acutely ill patients.

The most common hormone pattern in sick euthyroid syndrome (SES), also called nonthyroidal illness (NTI), is a decrease in total and unbound T_3 levels (low T_3 syndrome) with normal levels of T_4 and TSH. The magnitude of the fall in T_3 correlates with the severity of the illness. T_4 conversion to T_3 via peripheral 5′ (outer ring) deiodination is impaired, leading to increased reverse T_3 (rT_3). Since rT_3 is metabolized by 5′ deiodination, its clearance is also reduced. Thus, decreased clearance rather than increased production is the major basis for increased rT_3. Also, T_4 is alternately metabolized to the hormonally inactive T_3 sulfate. It is generally assumed that this low T_3 state is adaptive, because it can be induced in normal individuals by fasting. Teleologically, the fall in T_3 may limit catabolism in starved or ill patients.

Very sick patients may exhibit a dramatic fall in total T_4 and T_3 levels (low T_4 syndrome). With decreased tissue perfusion, muscle and liver expression of the type 3 deiodinase leads to accelerated T_4 and T_3 metabolism. This state has a poor prognosis. Another key factor in the fall in T_4 levels is altered binding to thyroxine-binding globulin (TBG). The commonly used free T_4 assays are subject to artifact when serum binding proteins are low and underestimate the true free T_4 level. Fluctuation in TSH levels also creates

challenges in the interpretation of thyroid function in sick patients. TSH levels may range from <0.1 mIU/L in very ill patients, especially with dopamine or glucocorticoid therapy, to >20 mIU/L during the recovery phase of SES. The exact mechanisms underlying the subnormal TSH seen in 10% of sick patients and the increased TSH seen in 5% remain unclear but may be mediated by cytokines including IL-12 and IL-18.

Any severe illness can induce changes in thyroid hormone levels, but certain disorders exhibit a distinctive pattern of abnormalities. Acute liver disease is associated with an initial rise in total (but not unbound) T_3 and T_4 levels due to TBG release; these levels become subnormal with progression to liver failure. A transient increase in total and unbound T_4 levels, usually with a normal T_3 level, is seen in 5–30% of acutely ill psychiatric patients. TSH values may be transiently low, normal, or high in these patients. In the early stage of HIV infection, T_3 and T_4 levels rise, even if there is weight loss. T_3 levels fall with progression to AIDS, but TSH usually remains normal. Renal disease is often accompanied by low T_3 concentrations, but with normal rather than increased rT_3 levels, due to an unknown factor that increases uptake of rT_3 into the liver.

The diagnosis of SES is challenging. Historic information may be limited, and patients often have multiple metabolic derangements. Useful features to consider include previous history of thyroid disease and thyroid function tests, evaluation of the severity and time course of the patient's acute illness, documentation of medications that may affect thyroid function or thyroid hormone levels, and measurements of rT_3 together with unbound thyroid hormones and TSH. The diagnosis of SES is frequently presumptive, given the clinical context and pattern of laboratory values; only resolution of the test results with clinical recovery can clearly establish this disorder. Treatment of SES with thyroid hormone (T_4 and/or T_3) is controversial, but most authorities recommend monitoring the patient's thyroid function tests during recovery, without administering thyroid hormone, unless there is historic or clinical evidence suggestive of hypothyroidism. Sufficiently large randomized controlled trials using thyroid hormone are unlikely to resolve this therapeutic controversy in the near future, because clinical presentations and outcomes are highly variable.

AMIODARONE EFFECTS ON THYROID FUNCTION

Amiodarone is a commonly used type III antiarrhythmic agent **(Chap. 247)**. It is structurally related to thyroid hormone and contains 39% iodine by weight. Thus, typical doses of amiodarone (200 mg/d) are associated with very high iodine intake, leading to greater than fortyfold increases in plasma and urinary iodine levels. Moreover, because amiodarone is stored in adipose tissue, high iodine levels persist for >6 months after discontinuation of the drug. Amiodarone inhibits deiodinase activity, and its metabolites function as weak antagonists of thyroid hormone action. Amiodarone has the following effects on thyroid function: (1) acute, transient suppression of thyroid function; (2) hypothyroidism in patients susceptible to the inhibitory effects of a high iodine load; and (3) thyrotoxicosis that may be caused by either a Jod-Basedow effect from the iodine load, in the setting of MNG or incipient Graves' disease, or a thyroiditis-like condition.

The initiation of amiodarone treatment is associated with a transient decrease of T_4 levels, reflecting the inhibitory effect of iodine on T_4 release. Soon thereafter, most individuals escape from iodide-dependent suppression of the thyroid (Wolff-Chaikoff effect), and the inhibitory effects on deiodinase activity and thyroid hormone receptor action become predominant. These events lead to the following pattern of thyroid function tests: increased T_4, decreased T_3, increased rT_3, and a transient TSH increase (up to 20 mIU/L). TSH levels normalize or are slightly suppressed within 1–3 months.

The incidence of hypothyroidism from amiodarone varies geographically, apparently correlating with iodine intake. Hypothyroidism occurs in up to 13% of amiodarone-treated patients in iodine-replete countries, such as the United States, but is less common (<6% incidence) in areas of lower iodine intake, such as Italy or Spain. The pathogenesis appears to involve an inability of the thyroid gland to escape from the Wolff-Chaikoff effect in autoimmune thyroiditis. Consequently, amiodarone-associated hypothyroidism is more common in women and individuals with positive TPO antibodies. It is usually unnecessary to discontinue amiodarone for this side effect, because levothyroxine can be used to normalize thyroid function. TSH levels should be monitored, because T_4 levels are often increased for the reasons described above.

The management of amiodarone-induced thyrotoxicosis (AIT) is complicated by the fact that there are different causes of thyrotoxicosis and because the increased thyroid hormone levels exacerbate underlying arrhythmias and coronary artery disease. Amiodarone treatment causes thyrotoxicosis in 10% of patients living in areas of low iodine intake and in 2% of patients in regions of high iodine intake. There are two major forms of AIT, although some patients have features of both. Type 1 AIT is associated with an underlying thyroid abnormality (preclinical Graves' disease or nodular goiter). Thyroid hormone synthesis becomes excessive as a result of increased iodine exposure (Jod-Basedow phenomenon). Type 2 AIT occurs in individuals with no intrinsic thyroid abnormalities and is the result of drug-induced lysosomal activation leading to destructive thyroiditis with histiocyte accumulation in the thyroid; the incidence rises as cumulative amiodarone dosage increases. Mild forms of type 2 AIT can resolve spontaneously or can occasionally lead to hypothyroidism. Color-flow Doppler ultrasonography shows increased vascularity in type 1 AIT but decreased vascularity in type 2 AIT. Thyroid scintiscans are difficult to interpret in this setting because the high endogenous iodine levels diminish tracer uptake. However, the presence of normal or rarely increased uptake favors type 1 AIT.

In AIT, the drug should be stopped, if possible, although this is often impractical because of the underlying cardiac disorder. Discontinuation of amiodarone will not have an acute effect because of its storage and prolonged half-life. High doses of antithyroid drugs can be used in type 1 AIT but are often ineffective. Potassium perchlorate, 200 mg every 6 h, has been used to reduce thyroidal iodide content. Perchlorate treatment has been associated with agranulocytosis, although the risk appears relatively low with short-term use. Glucocorticoids, as administered for subacute thyroiditis, have modest benefit in type 2 AIT. Lithium blocks thyroid hormone release and can also provide some benefit. Near-total thyroidectomy rapidly decreases thyroid hormone levels and may be the most effective long-term solution if the patient can undergo the procedure safely.

FURTHER READING

Biondi B et al: The 2015 European Thyroid Association Guidelines on diagnosis and treatment of endogenous subclinical hyperthyroidism. Eur Thyroid J 4:149, 2015. [PubMed: 26558232]

De Leo S et al: Hyperthyroidism. Lancet 388:906, 2016. [PubMed: 27038492]

Hanley P et al: Thyroid disorders in children and adolescents: A review. JAMA Pediatr 170:1008, 2016. [PubMed: 27571216]

Ross DS et al: 2016 American Thyroid Association guidelines for diagnosis and management of hyperthyroidism and other causes of thyrotoxicosis. Thyroid 26:1343, 2016.

Wiersinga WM: Advances in treatment of active, moderate-to-severe Graves' ophthalmopathy. Lancet Diabetes Endocrinol 5:134, 2017.

Chapter 378: Thyroid Nodular Disease and Thyroid Cancer

J. Larry Jameson; Susan J. Mandel; Anthony P. Weetman

GOITER AND NODULAR THYROID DISEASE

Goiter refers to an enlarged thyroid gland. Biosynthetic defects, iodine deficiency, autoimmune disease, and nodular diseases can each lead to goiter, although by different mechanisms. Biosynthetic defects and iodine deficiency are associated with reduced efficiency of thyroid hormone synthesis, leading to increased thyroid-stimulating hormone (TSH), which stimulates thyroid growth as a compensatory mechanism to overcome the block in hormone synthesis. Graves' disease and Hashimoto's thyroiditis are also associated with goiter. In Graves' disease, the goiter results mainly from the TSH-R–mediated effects of thyroid-stimulating immunoglobulins. The goitrous form of Hashimoto's thyroiditis occurs because of acquired defects in hormone synthesis, leading to elevated levels of TSH and its consequent growth effects. Lymphocytic infiltration and immune system–induced growth factors also contribute to thyroid enlargement in Hashimoto's thyroiditis.

Thyroid nodular disease is characterized by the disordered growth of thyroid cells, which can be either hyperplastic or neoplastic. A patient may have a multinodular goiter (MNG) in which thyroid nodules (generally hyperplastic) replace the majority of the normal thyroid parenchyma; this presentation is more common in areas of borderline iodine deficiency. Or, the thyroid gland may be normal in size and contain discrete thyroid nodules. Because the management of goiter depends on the etiology, the detection of thyroid enlargement on physical examination should prompt further evaluation to identify its cause.

Nodular thyroid disease is common, occurring in about 3–7% of adults when assessed by physical examination. Using ultrasound, nodules are present in up to 50% of adults, with the majority being <1 cm in diameter. Thyroid nodules may be solitary or multiple, and they may be functional or nonfunctional.

DIFFUSE NONTOXIC (SIMPLE) GOITER

Etiology and Pathogenesis

When diffuse enlargement of the thyroid occurs in the absence of nodules and hyperthyroidism, it is referred to as a *diffuse nontoxic goiter*. This is sometimes called *simple goiter*, because of the absence of nodules, or *colloid goiter*, because of the presence of uniform follicles that are filled with colloid. Worldwide, diffuse goiter is most commonly caused by iodine deficiency and is termed *endemic goiter* when it affects >5% of the population. In nonendemic regions, *sporadic goiter* occurs, and the cause is usually unknown. Thyroid enlargement in teenagers is sometimes referred to as *juvenile goiter*. In general, goiter is more common in women than men, probably because of the greater prevalence of underlying autoimmune disease and the increased iodine demands associated with pregnancy.

In *iodine-deficient areas*, thyroid enlargement reflects a compensatory effort to trap iodide and produce sufficient hormone under conditions in which hormone synthesis is relatively inefficient. Somewhat surprisingly, TSH levels are usually normal or only slightly increased, suggesting increased sensitivity to TSH or activation of other pathways that lead to thyroid growth. Iodide appears to have direct actions on thyroid vasculature and may indirectly affect growth through vasoactive substances such as endothelins and nitric oxide. Endemic goiter may also be caused by exposure to environmental *goitrogens* such as cassava root, which contains a thiocyanate; vegetables of the Cruciferae family (known as cruciferous vegetables) (e.g., Brussels sprouts, cabbage, and cauliflower); and milk from regions where goitrogens are present in grass. Although relatively rare, inherited defects in thyroid hormone synthesis lead to a diffuse nontoxic goiter. Abnormalities at each step of hormone synthesis, including iodide transport (NIS), transgenic (Tg) synthesis, organification and coupling (thrombopoietin [TPO]), and the regeneration of iodide (dehalogenase), have been described.

CLINICAL MANIFESTATIONS AND DIAGNOSIS

If thyroid function is preserved, most goiters are asymptomatic. Examination of a diffuse goiter reveals a symmetrically enlarged, nontender, generally soft gland without palpable nodules. Goiter is defined, somewhat arbitrarily, as a lateral lobe with a volume greater than the thumb of the individual being examined. On ultrasound, total thyroid volume exceeding 30 mL is considered abnormal. If the thyroid is markedly enlarged, it can cause tracheal or esophageal compression. These features are unusual, however, in the absence of nodular disease and fibrosis. *Substernal goiter* may obstruct the thoracic inlet. *Pemberton's sign* refers to facial and neck congestion due to jugular venous obstruction when the arms are raised above the head, a maneuver that draws the thyroid into the thoracic inlet. Respiratory flow measurements and CT or MRI should be used to evaluate substernal goiter in patients with obstructive signs or symptoms.

Thyroid function tests should be performed in all patients with goiter to exclude thyrotoxicosis or hypothyroidism. It is not unusual, particularly in iodine deficiency, to find a low total T_4, with normal T_3 and TSH, reflecting enhanced $T_4 \rightarrow T_3$ conversion. A low TSH with a normal free T_3 and free T_4, particularly in older patients, suggests the possibility of thyroid autonomy or undiagnosed Graves' disease, and is termed *subclinical thyrotoxicosis*. The benefit of treatment (typically with radioiodine) in subclinical thyrotoxicosis, versus follow-up and implementing treatment if free T_3 or free T_4 levels become abnormal, is unclear, but treatment is increasingly recommended in the elderly to reduce the risk of atrial fibrillation and bone loss. TPO antibodies may be useful to identify patients at increased risk of autoimmune thyroid disease. Low urinary iodine levels (<50 µg/L) support a diagnosis of iodine deficiency. Thyroid scanning is not generally necessary but will reveal increased uptake in iodine deficiency and most cases of dyshormonogenesis.

TREATMENT

TREATMENT

Diffuse Nontoxic (Simple) Goiter

Iodine replacement induces variable regression of goiter in iodine deficiency, depending on duration and the degree of hyperplasia, with accompanying fibrosis, and autonomous function that may have developed. Surgery is rarely indicated for diffuse goiter. Exceptions include documented evidence of tracheal compression or obstruction of the thoracic inlet, which are more likely to be associated with substernal MNGs (see below). Subtotal or near-total thyroidectomy for these or cosmetic reasons should be performed by an experienced surgeon to minimize

complication rates. Surgery should be followed by replacement with levothyroxine, with the aim of keeping the TSH level at the lower end of the reference interval to prevent regrowth of the goiter.

NONTOXIC MULTINODULAR GOITER

Etiology and Pathogenesis

Depending on the population studied, MNG or the presence of nodules in a thyroid of normal size occurs in up to 12% of adults. MNG should be distinguished from the presence of nodules in normal size thyroid gland (see "Approach to the Patient with Thyroid Nodules"). MNG is more common in women than men and increases in prevalence with age. It is more common in iodine-deficient regions but also occurs in regions of iodine sufficiency, reflecting multiple genetic, autoimmune, and environmental influences on the pathogenesis.

There is typically wide variation in nodule size. Histology reveals a spectrum of morphologies ranging from hypercellular, hyperplastic regions to cystic areas filled with colloid. Fibrosis is often extensive, and areas of hemorrhage or lymphocytic infiltration may be seen. Using molecular techniques, most nodules within an MNG are polyclonal in origin, suggesting a hyperplastic response to locally produced growth factors and cytokines. TSH, which is usually not elevated, may play a permissive or contributory role. Monoclonal neoplastic lesions may also occur, reflecting mutations in genes that confer a selective growth advantage to the progenitor cell.

Clinical Manifestations

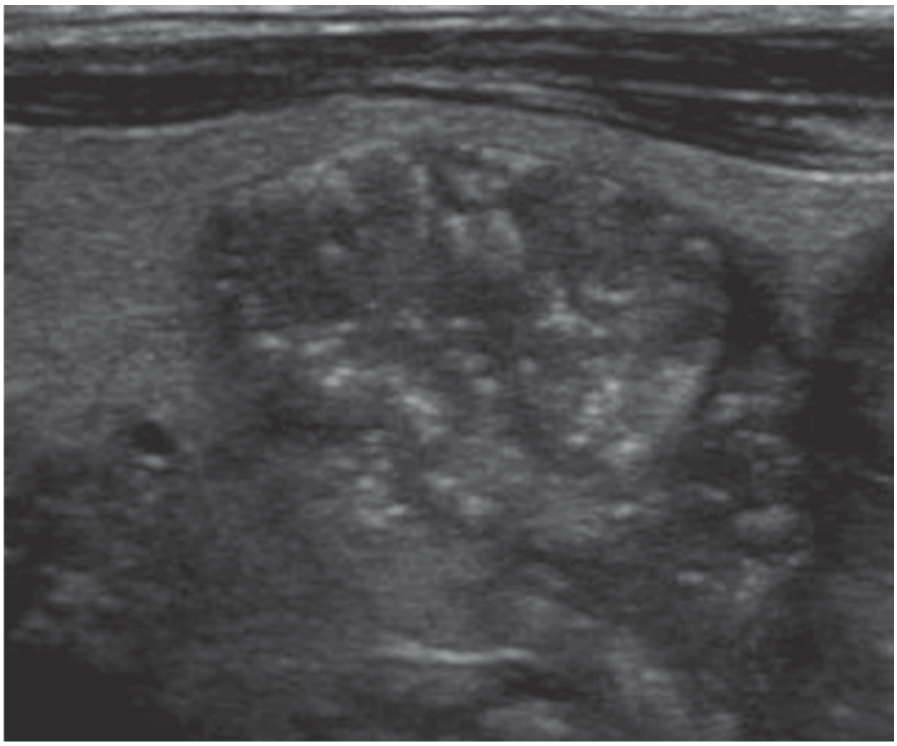
Most patients with nontoxic MNG are asymptomatic and euthyroid. MNG typically develops over many years and is detected on routine physical examination, when an individual notices an enlargement in the neck, or as an incidental finding on imaging. If the goiter is large enough, it can ultimately lead to compressive symptoms including difficulty swallowing, respiratory distress (tracheal compression), or plethora (venous congestion), but these symptoms are uncommon. Symptomatic MNGs are usually extraordinarily large and/or develop fibrotic areas that cause compression. Sudden pain in an MNG is usually caused by hemorrhage into a nodule but should raise the possibility of invasive malignancy. Hoarseness, reflecting laryngeal nerve involvement, also suggests malignancy.

Diagnosis

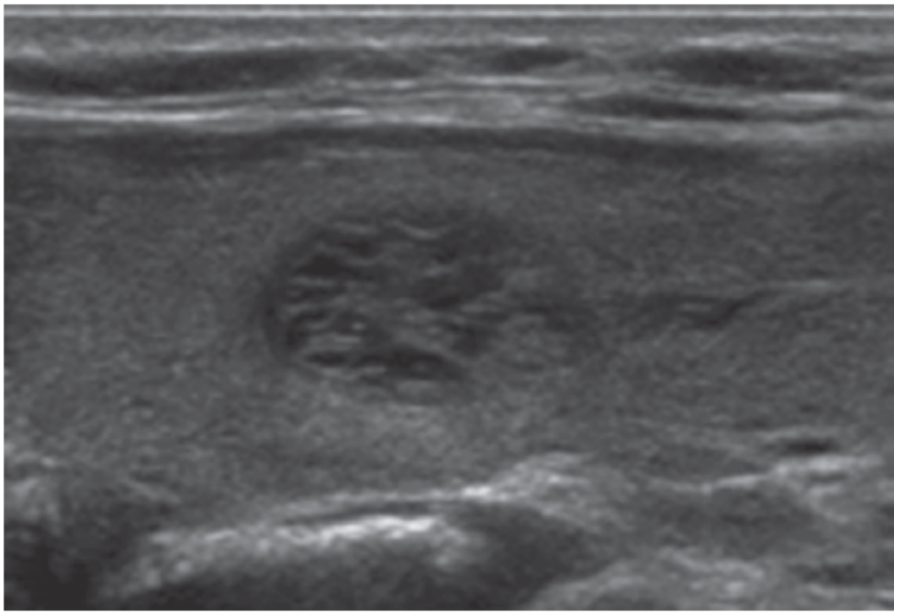
On examination, thyroid architecture is distorted, and multiple nodules of varying size can be appreciated. Because many nodules are deeply embedded in thyroid tissue or reside in posterior or substernal locations, it is not possible to palpate all nodules. Pemberton's sign, characterized by facial suffusion when the patient's arms are elevated above the head, suggests that the goiter has increased pressure in the thoracic inlet. A TSH level should be measured to exclude subclinical hyper- or hypothyroidism, but thyroid function is usually normal. Tracheal deviation is common, but compression must usually exceed 70% of the tracheal diameter before there is significant airway compromise. Pulmonary function testing can be used to assess the functional effects of compression, which characteristically causes inspiratory stridor. CT or MRI can be used to evaluate the anatomy of the goiter and the extent of substernal extension or tracheal narrowing. A barium swallow may reveal the extent of esophageal compression. The risk of malignancy in MNG is similar to that in solitary nodules. Ultrasonography can be used to identify which nodules should be biopsied based on a combination of size and sonographic features (Table 378-1) (Chap. 375). For nodules with more suspicious sonographic patterns (e.g., hypoechoic solid nodules with infiltrative borders), biopsy is recommended at a lower size cutoff than those with less suspicious imaging features (Fig. 378-1).

FIGURE 378-1

Sonographic patterns of thyroid nodules. ***A.*** High suspicion ultrasound pattern for thyroid malignancy (hypoechoic solid nodule with irregular borders and microcalcifications). ***B.*** Very low suspicion ultrasound pattern for thyroid malignancy (spongiform nodule with microcystic areas comprising over >50% of nodule volume).



A



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 378-1

Grayscale Sonographic Features Associated with Thyroid Cancer

	MEDIAN SENSITIVITY [RANGE]	MEDIAN SPECIFICITY [RANGE]
Hypoechoic compared with surrounding thyroid	81% [48–90%]	53% [36–92%]
Marked hypoechogenicity	41% [27–59%]	94% [92–94%]
Microcalcifications	44% [26–73%]	89% [69–98%]
Irregular, microlobulated margins	55% [17–84%]	79% [62–85%]
Solid consistency	86% [78–91%]	48% [30–58%]
Taller than wide shape on transverse view	48% [33–84%]	92% [82–93%]

TREATMENT

TREATMENT

Nontoxic Multinodular Goiter

Most nontoxic MNGs can be managed conservatively. T_4 suppression is rarely effective for reducing goiter size and introduces the risk of subclinical or overt thyrotoxicosis, particularly if there is underlying autonomy or if it develops during treatment. Contrast agents and other iodine-containing substances should be avoided because of the risk of inducing the *Jod-Basedow effect*, characterized by enhanced thyroid hormone production by autonomous nodules. Radioiodine has been used when surgery is contraindicated in areas where large nodular goiters are more prevalent (e.g., some areas of Europe and Brazil) because it can decrease MNG volume and may selectively ablate regions of autonomy. Dosage of ^{131}I depends on the size of the goiter and radioiodine uptake but is usually about 3.7 MBq (0.1 mCi) per gram of tissue, corrected for uptake (typical dose 370–1070 MBq [10–29 mCi]). Repeat treatment may be needed and effectiveness may be increased by concurrent administration of low-dose recombinant TSH (0.1 mg IM). It is possible to achieve a 40–50% reduction in goiter size in most patients. Earlier concerns about radiation-induced thyroid swelling and tracheal compression have diminished, as studies have shown this complication to be rare. When acute compression occurs, glucocorticoid treatment or surgery may be needed. Radiation-induced hypothyroidism is less common than after treatment for Graves' disease. However, posttreatment autoimmune thyrotoxicosis may occur

in up to 5% of patients treated for nontoxic MNG. Surgery remains highly effective but is not without risk, particularly in older patients with underlying cardiopulmonary disease.

TOXIC MULTINODULAR GOITER

The pathogenesis of toxic MNG appears to be similar to that of nontoxic MNG; the major difference is the presence of functional autonomy in toxic MNG. The molecular basis for autonomy in toxic MNG remains unknown. As in nontoxic goiters, many nodules are polyclonal, whereas others are monoclonal and vary in their clonal origins. Genetic abnormalities known to confer functional autonomy, such as activating TSH-R or $G_{s\alpha}$ mutations (see below), are not usually found in the autonomous regions of toxic MNG goiter.

In addition to features of goiter, the clinical presentation of toxic MNG includes subclinical or mild overt hyperthyroidism. The patient is usually elderly and may present with atrial fibrillation or palpitations, tachycardia, nervousness, tremor, or weight loss. Recent exposure to iodine, from contrast dyes or other sources, may precipitate or exacerbate thyrotoxicosis. The TSH level is low. The uncombined T_4 level may be normal or minimally increased; T_3 is often elevated to a greater degree than T_4. Thyroid scan shows heterogeneous uptake with multiple regions of increased and decreased uptake; 24-h uptake of radioiodine may not be increased but is usually in the upper normal range.

Prior to definitive treatment of the hyperthyroidism, ultrasound imaging should be performed to assess the presence of discrete nodules corresponding to areas of decreased uptake (“cold” nodules). If present, fine-needle aspiration (FNA) may be indicated based on sonographic patterns and size cutoffs. The cytology results, if indeterminate or suspicious, may direct the therapy to surgery.

TREATMENT

TREATMENT

Toxic Multinodular Goiter

Antithyroid drugs normalize thyroid function and are particularly useful in the elderly or ill patients with limited lifespan. In contrast to Graves’ disease, spontaneous remission does not occur and so treatment is long-term. Radioiodine is generally the treatment of choice; it treats areas of autonomy as well as decreasing the mass of the goiter by ablating the functioning nodules. Sometimes, however, a degree of autonomy may persist, presumably because multiple autonomous regions may emerge after others are treated, and further radioiodine treatment may be necessary. Surgery provides definitive treatment of underlying thyrotoxicosis as well as goiter. Patients should be rendered euthyroid using an antithyroid drug before operation.

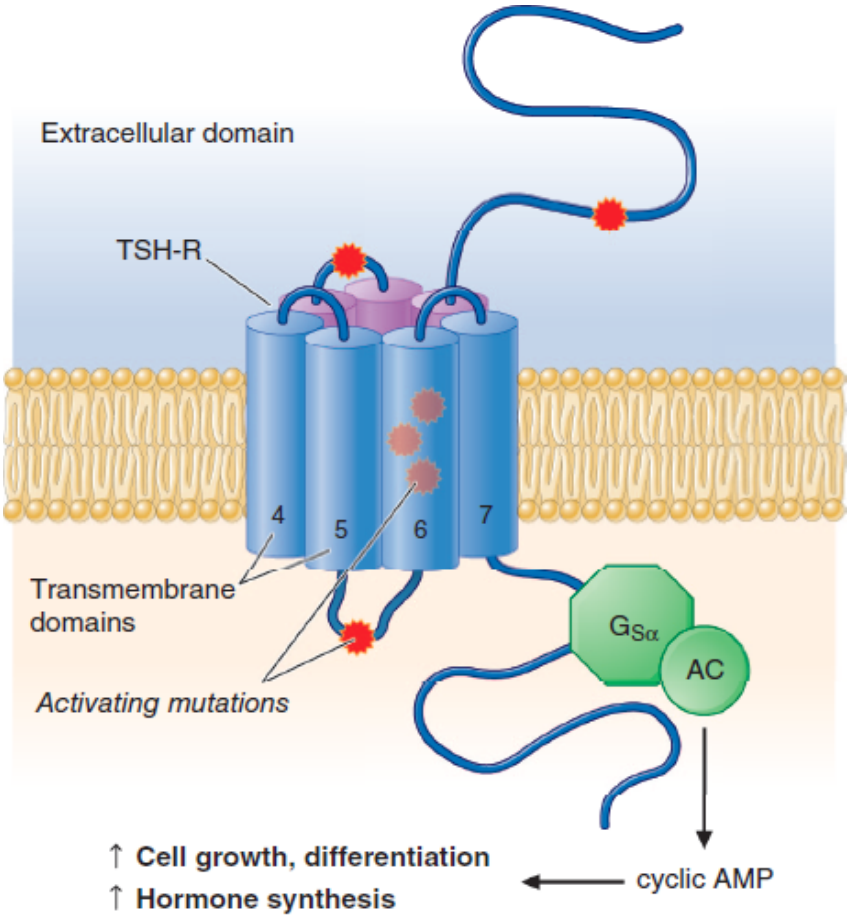
HYPERFUNCTIONING SOLITARY NODULE

A solitary, autonomously functioning thyroid nodule is referred to as *toxic adenoma*. The pathogenesis of this disorder has been unraveled by demonstrating the functional effects of mutations that stimulate the TSH-R signaling pathway. Most patients with solitary hyperfunctioning nodules have acquired somatic, activating mutations in the TSH-R (**Fig. 378-2**). These mutations, located primarily in the receptor transmembrane domain, induce constitutive receptor coupling to $G_{S\alpha}$, increasing cyclic adenosine monophosphate (AMP) levels and leading to enhanced thyroid follicular

cell proliferation and function. Less commonly, somatic mutations are identified in $G_{S\alpha}$. These mutations, which are similar to those seen in McCune-Albright syndrome (**Chap. 405**) or in a subset of somatotrope adenomas (**Chap. 373**), impair guanosine triphosphate (GTP) hydrolysis, causing constitutive activation of the cyclic AMP signaling pathway. In most series, activating mutations in either the TSH-R or the $G_{S\alpha}$ subunit genes are identified in >90% of patients with solitary hyperfunctioning nodules.

FIGURE 378-2

Activating mutations of the thyroid-stimulating hormone receptor (TSH-R). Mutations (*) that activate TSH-R reside mainly in transmembrane 5 and intracellular loop 3, although mutations have occurred in a variety of different locations. The effect of these mutations is to induce conformational changes that mimic TSH binding, thereby leading to coupling to stimulatory G protein ($G_{S\alpha}$) and activation of adenylate cyclase (AC), an enzyme that generates cyclic AMP.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Thyrotoxicosis is usually mild and is generally only detected when a nodule is >3 cm. The disorder is suggested by a subnormal TSH level; the presence of the thyroid nodule, often large enough to be palpable; and the absence of clinical features suggestive of Graves' disease or other causes of thyrotoxicosis. A thyroid scan provides a definitive diagnostic test, demonstrating focal uptake in the hyperfunctioning nodule and diminished uptake in the remainder of the gland, as activity of the normal thyroid is suppressed.

TREATMENT

TREATMENT

Hyperfunctioning Solitary Nodule

Radioiodine ablation is usually the treatment of choice. Because normal thyroid function is suppressed, ^{131}I is concentrated in the hyperfunctioning nodule with minimal uptake and damage to normal thyroid tissue. Relatively large radioiodine doses (e.g., 370–1110 MBq [10–29.9 mCi] ^{131}I) have been shown to correct thyrotoxicosis in about 75% of patients within 3 months. Hypothyroidism occurs in <10% of those patients over the next 5 years. Surgical resection is also effective and is usually limited to lobectomy, thereby preserving thyroid function and minimizing risk of hypoparathyroidism or damage to the recurrent laryngeal nerves. Medical therapy using antithyroid drugs and beta blockers can normalize thyroid function but is not an optimal long-term treatment. Using ultrasound guidance, repeated percutaneous radiofrequency thermal ablation has been used successfully in some centers to ablate hyperfunctioning nodules, and this technique has also been used to reduce the size of nonfunctioning thyroid nodules.

BENIGN LESIONS

The various types of benign thyroid nodules are listed in **Table 378-2**. Benign nodules may be hyperplastic and reflect a combination of both macro- and microfollicular architecture or they may be neoplastic, encapsulated adenomas that generally have a more monotonous microfollicular pattern. If the adenoma is composed of oncocytic follicular cells arranged in a follicular pattern, this is termed a Hürthle cell adenoma. Hyperplastic nodules generally appear as mixed cystic/solid or spongiform lesions on ultrasound. The definition of spongiform requires the presence of microcystic areas comprising >50% of the nodule volume, with the concept that this microcystic sonographic pattern recapitulates the histology of macrofollicles containing colloid (Fig. 378-1B). However, the majority of solid nodules (whether hypo-, iso-, or hyperechoic) are also benign. FNA, usually performed with ultrasound guidance, is the diagnostic procedure of choice to evaluate thyroid nodules (see the "Approach to the Patient with Thyroid Nodules" section). Pure thyroid cysts, <1% of all thyroid growths, consist of colloid and are benign as well. Cysts frequently recur, even after repeated aspiration, and may require surgical excision if they are large. Ethanol ablation to sclerose the cyst has been used successfully for patients who are symptomatic.

TABLE 378-2

Classification of Thyroid Growths

Benign	
Hyperplasia Colloid nodule **Follicular epithelial cell adenomas**	
Conventional Oncocytic (Hürthle cell)	
Malignant	**Approximate Prevalence, %**
Follicular epithelial cell	
Papillary carcinomas	80–85
Classic variant	
Follicular variant	
Diffuse sclerosing variant	
Tall cell, columnar cell variants	
Follicular carcinomas	2-5–7
Conventional	
Oncocytic (Hürthle cell)	
Poorly differentiated carcinomas	3–5
Anaplastic (undifferentiated) carcinomas	1
C cell origin(calcitonin-producing)	
Medullary thyroid cancer	<10
Sporadic	
Familial	
MEN 2	

Other malignancies	
Lymphomas	1
Metastases	
Breast, melanoma, lung, kidney	
Others	

Abbreviation: MEN, multiple endocrine neoplasia.

TSH suppression with levothyroxine therapy does not decrease thyroid nodule size in iodine-sufficient populations. However, if there is relative iodine deficiency, both iodine and levothyroxine therapy have been demonstrated to decrease nodule volume. If levothyroxine is administered in this situation, the TSH should be maintained at or just below the lower limit of normal, but not frankly suppressed. If the nodule has not decreased in size after 6–12 months of therapy, treatment should be discontinued because little benefit is likely to accrue from long-term treatment; the risk of iatrogenic subclinical thyrotoxicosis should also be considered.

THYROID CANCER

Thyroid carcinoma is the most common malignancy of the endocrine system. Malignant tumors derived from the follicular epithelium are classified according to histologic features. Differentiated tumors, such as papillary thyroid cancer (PTC) or follicular thyroid cancer (FTC), are often curable, and the prognosis is good for patients identified with early-stage disease. In contrast, anaplastic thyroid cancer (ATC) is aggressive, responds poorly to treatment, and is associated with a bleak prognosis.

Over the last 30 years, the incidence of thyroid cancer has increased from 4.9 to 14.3 cases per 100,000 individuals in the United States, with over 65,000 cases diagnosed in 2015. However, disease-specific mortality has not changed. The increased incidence is predominantly attributable to small T1 papillary cancer tumors (<2 cm), and has led experts to consider that thyroid cancer is being overdiagnosed, suggesting that cancers are being detected that would otherwise be unlikely to harm a patient. The concept of cancer overdiagnosis is predicated upon the presence of a disease reservoir (the autopsy prevalence of PTC is ~25%), activities leading to disease detection (increased diagnostic imaging with incidental detection of nodules), and a mismatch in the directional rate between diagnosis and mortality (thyroid cancer disease-specific mortality not changed in 40 years). Similar trends have been observed worldwide, especially in those countries with higher proportion of privately financed healthcare, leading to increased resource utilization including imaging. The 20-year disease-specific mortality for low risk thyroid cancer is 1%.

Current trends in thyroid cancer care focus on: (1) avoiding overdiagnosis by limiting FNA by sonographic risk stratification with size cut offs; (2) limiting surgery, radioiodine, and subsequent surveillance for low risk tumors; (3) identifying patients at higher recurrence risk for more aggressive treatment and monitoring. Prognosis is worse in older persons (>65 years). Thyroid cancer is twice as common in women as men, but male gender is associated with a worse prognosis. Additional important risk factors include a history of childhood (before age 18) head or neck irradiation, evidence for local tumor fixation or gross metastatic involvement of lymph nodes, and the presence of distant metastases (**Table 378-3**).

TABLE 378-3

Risk Factors for Thyroid Carcinoma in Patients with Thyroid Nodule from History and Physical Examination

History of head and neck irradiation before the age of 18, including, mantle radiation for Hodgkin's disease, and brain radiation for childhood leukemia or other cranial malignancies Exposure to ionizing radiation from fallout in childhood or adolescence Age <20 or >65 years Rapidly enlarging neck mass Male gender	Family history of papillary thyroid cancer in 2 or more first degree relatives, MEN 2, or other genetic syndromes associated with thyroid malignancy (e.g., Cowden's syndrome, familial polyposis, Carney complex, PTEN [phosphatase and tensin homolog] hamartoma tumor) Vocal cord paralysis, hoarse voice Nodule fixed to adjacent structures Lateral cervical lymphadenopathy

Abbreviation: MEN, multiple endocrine neoplasia.

Several unique features of thyroid cancer facilitate its management: (1) thyroid nodules are amenable to biopsy by FNA; (2) iodine radioisotopes can be used to diagnose (^{123}I and ^{131}I) and potentially treat (^{131}I) differentiated thyroid cancer, reflecting the unique uptake of this anion by the thyroid gland; and (3) serum markers allow the detection of residual or recurrent disease, including the use of Tg levels for PTC and FTC, and calcitonin for medullary thyroid cancer (MTC).

CLASSIFICATION

Thyroid neoplasms can arise in each of the cell types that populate the gland, including thyroid follicular cells, calcitonin-producing C cells, lymphocytes, and stromal and vascular elements, as well as metastases from other sites (Table 378-2). The American Joint Committee on Cancer (AJCC) has designated a staging system using the tumor, node, metastasis (TNM) classification, which is most commonly used **(Table 378-4)**. Revised guidelines released in 2018 changed the age cutoffs from <45 to <55 years, as well as some of the staging criteria (see Further Reading).

TABLE 378-4

Thyroid Cancer Classification[a]

Papillary or Follicular Thyroid Cancers		
	<45 years	>45 years
Stage I	Any T, any N, M0	T1, N0, M0
Stage II	Any T, any N, M1	T2, N0, M0
Stage III	—	T3, N0, M0
		T1–T3, N1a, M0
Stage IVA	—	T4a, any N, M0 T1–T3, N1b, M0
Stage IVB		T4b, any N, M0
Stage IVC		Any T, any N, M1
Anaplastic Thyroid Cancer		
Stage IV	All cases are stage IV	
Medullary Thyroid Cancer		
Stage I	T1, N0, M0	
Stage II	T2 or T3, N0, M0	
Stage III	T1–T3, N1a, M0	
Stage IVA	T4a, any N, M0 T1–T3, N1b, M0	
Stage IVB	T4b, any N, M0	
Stage IVC	Any T, any N, M1	

[a]Criteria include: T, the size and extent of the primary tumor (T1a ≤1 cm; T1b >1 cm but ≤2 cm; T2 >2 cm but ≤4 cm; T3 >4 cm or any tumor with extension into perithyroidal soft tissue or sternothyroid muscle; T4a invasion into subcutaneous soft tissues, larynx, trachea, esophagus, or recurrent laryngeal nerve; T4b invasion into prevertebral fascia or encasement of carotid artery or mediastinal vessels); N, the absence (N0) or presence (N1a level IV central compartment; N1b levels II–V lateral compartment,

upper mediastinal or retro/parapharyngeal) of regional node involvement; M, the absence (M0) or presence (M1) of distant metastases.

Note that updated TNM classification was released in early 2018.

Source: American Joint Committee on Cancer staging system for thyroid cancers using the TNM classification, 7th edition.

PATHOGENESIS AND GENETIC BASIS

Radiation

Early studies of the pathogenesis of thyroid cancer focused on the role of external radiation, which predisposes to chromosomal breaks, leading to genetic rearrangements and loss of tumor-suppressor genes. External radiation of the mediastinum, face, head, and neck region was administered in the past to treat an array of conditions, including acne and enlargement of the thymus, tonsils, and adenoids. Radiation exposure increases the risk of benign and malignant thyroid nodules, is associated with multicentric cancers, and shifts the incidence of thyroid cancer to an earlier age group. Radiation from nuclear fallout also increases the risk of thyroid cancer. Children seem more predisposed to the effects of radiation than adults.

TSH and Growth Factors

Many differentiated thyroid cancers express TSH receptors and, therefore, remain responsive to TSH. Higher serum TSH levels, even within normal range, are associated with increased thyroid cancer risk in patients with thyroid nodules. These observations provide the rationale for T_4 suppression of TSH in patients with thyroid cancer. Residual expression of TSH receptors also allows TSH-stimulated uptake of ^{131}I therapy (see below).

Oncogenes and Tumor-Suppressor Genes

Thyroid cancers are monoclonal in origin, consistent with the idea that they originate as a consequence of mutations that confer a growth advantage to a single cell. In addition to increased rates of proliferation, some thyroid cancers exhibit impaired apoptosis and features that enhance invasion, angiogenesis, and metastasis. Thyroid neoplasms have been analyzed for a variety of genetic alterations, but without clear evidence of an ordered acquisition of somatic mutations as they progress from the benign to the malignant state. On the other hand, certain mutations, such as *RET/PTC* and *PAX8-PPARγ1* rearrangements, are relatively specific for thyroid neoplasia.

As described above, activating mutations of the TSH-R and the $G_{S\alpha}$ subunit are associated with autonomously functioning nodules. Although these mutations induce thyroid cell growth, this type of nodule is almost always benign.

Activation of the RET-RAS-BRAF signaling pathway is seen in up to 70% of PTCs, although the types of mutations are heterogeneous. A variety of rearrangements involving the *RET* gene on chromosome 10 bring this receptor tyrosine kinase under the control of other promoters, leading to receptor overexpression. *RET* rearrangements occur in 20–40% of PTCs in different series and were observed with increased frequency in tumors developing after the Chernobyl radiation accident. Rearrangements in PTC have also been observed for another tyrosine kinase gene, *TRK1*, which is located on chromosome 1. To date, the identification of PTC with *RET* or *TRK1* rearrangements has not proven useful for predicting prognosis or treatment responses. *BRAF V600E* mutations appear to be the most common genetic alteration in PTC. These mutations activate the kinase, which stimulates the mitogen-activated protein kinase (MAPK)

cascade. *RAS* mutations, which also stimulate the MAPK cascade, are found in about 20–30% of thyroid neoplasms (*NRAS > HRAS > KRAS*), including both PTC follicular variant and FTC. Of note, simultaneous *RET*, *BRAF*, and *RAS* mutations rarely occur in the same tumor, suggesting that activation of the MAPK cascade is critical for tumor development, independent of the step that initiates the cascade.

RAS mutations also occur in FTCs. In addition, a rearrangement of the thyroid developmental transcription factor PAX8 with the nuclear receptor PPARγ is identified in a significant fraction of FTCs. Overall, about 70% of follicular cancers have mutations or genetic rearrangements. Loss of heterozygosity of 3p or 11q, consistent with deletions of tumor-suppressor genes, is also common in FTCs.

Most of the mutations seen in differentiated thyroid cancers have also been detected in ATCs. *TERT* promoter mutations occur in <10% of differentiated PTC but are more common in ATC. *BRAF* mutations are seen in up to 50% of ATCs. Mutations in *CTNNB1,* which encodes β-catenin, occur in about two-thirds of ATCs, but not in PTC or FTC. Mutations of the tumor-suppressor *P53* also play an important role in the development of ATC. Because *P53* plays a role in cell cycle surveillance, DNA repair, and apoptosis, its loss may contribute to the rapid acquisition of genetic instability as well as poor treatment responses **(Chap. 68)**.

The role of molecular diagnostics in the clinical management of thyroid cancer is under investigation. In principle, analyses of specific mutations might aid in classification, prognosis, or choice of treatment. Although *BRAF V600E* mutations are associated with loss of iodine uptake by tumor cells, there is no clear evidence to date that this information alters clinical decision-making. Higher recurrence rates have been variably reported in patients with *BRAF*-positive PTC, but the impact on survival rates is unclear.

MTC, when associated with multiple endocrine neoplasia (MEN) type 2, harbors an inherited mutation of the *RET* gene. Unlike the rearrangements of *RET* seen in PTC, the mutations in MEN 2 are point mutations that induce constitutive activity of the tyrosine kinase **(Chap. 381)**. MTC is preceded by hyperplasia of the C cells, raising the likelihood that as-yet-unidentified "second hits" lead to cellular transformation. A subset of sporadic MTC contains somatic mutations that activate *RET*.

WELL-DIFFERENTIATED THYROID CANCER

Papillary

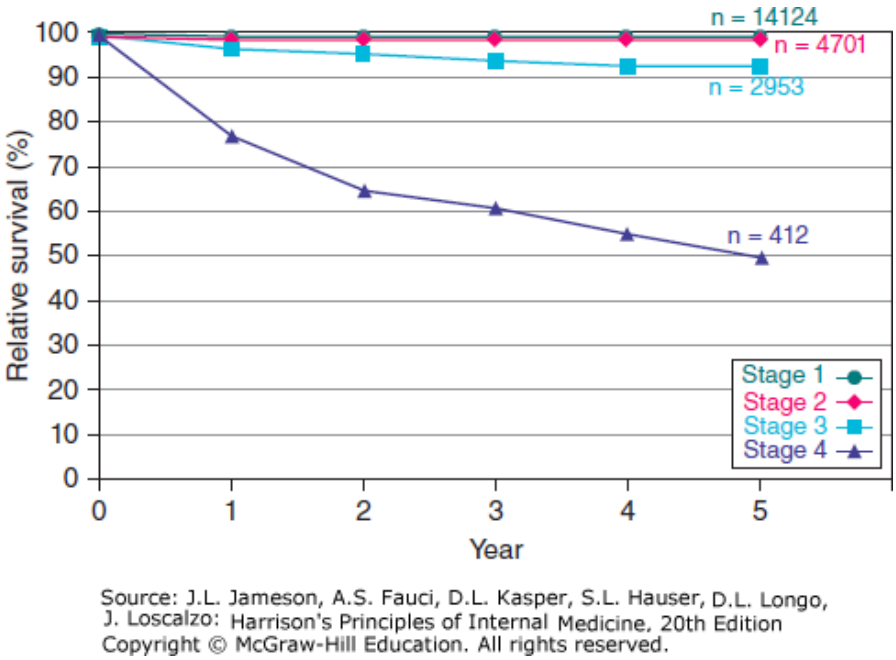
PTC is the most common type of thyroid cancer, accounting for 80–85% of well-differentiated thyroid malignancies. Microscopic PTC is present in up to 25% of thyroid glands at autopsy, but most of these lesions are very small (several millimeters) and are not clinically significant. Characteristic cytologic features of PTC help make the diagnosis by FNA or after surgical resection; these include, large, clear nuclei with powdery chromatin (described as an "orphan Annie eye" appearance) with nuclear grooves and prominent nucleoli. The histologic finding of these cells arranged in either papillary structures versus follicles distinguishes the classic and follicular variants of PTC, respectively.

PTC may be multifocal and invade locally within the thyroid gland as well as through the thyroid capsule and into adjacent structures in the neck. It has a propensity to spread via the lymphatic system but can metastasize hematogenously as well, particularly to bone and lung. Because of the relatively slow growth of the tumor, a significant burden of pulmonary metastases may accumulate, sometimes with remarkably few symptoms. The prognostic implication of lymph node spread depends upon the volume of metastatic disease. Micrometastases, defined as <2 mm of cancer in a lymph node, do not affect prognosis. However, gross metastatic involvement of multiple 2–3 cm lymph

nodes indicates a 25–30% chance of recurrence, and may increase mortality in older patients. The staging of PTC by the TNM system is outlined in Table 378-4. Most papillary cancers are identified in the early stages (>80% stages I or II) and have an excellent prognosis, with survival curves similar to expected survival **(Fig. 378-3)**. Mortality is markedly increased in stage IV disease, especially in the presence of distant metastases (stage IV-C), but this group comprises only about 1% of patients. The treatment of PTC is described below.

FIGURE 378-3

Survival rates of patients with different stages of papillary cancer. *(Adapted with permission from SB Edge, DR Byrd: Thyroid, in CC Compton, AB Fritz, FL Greene, A Trotti [eds]: AJCC Cancer Staging Manual, 7th ed. New York, Springer, 2010, pp. 87–92.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Follicular

The incidence of FTC varies widely in different parts of the world; it is more common in iodine-deficient regions. Currently, FTC accounts for only about 5% of all thyroid cancers diagnosed in the United States. FTC is difficult to diagnose by FNA because the distinction between benign and malignant follicular neoplasms requires histology because the nuclear features of follicular adenomas and carcinomas do not differ. Rather, follicular carcinoma is diagnosed by the presence of capsular and/or vascular invasion. FTC tends to spread by hematogenous routes leading to bone, lung, and central nervous system metastases. Mortality rates associated with angioinvasive FTC are less favorable than for PTC, in part because a larger proportion of patients present with stage IV disease. Poor prognostic features include distant metastases, age >50 years, primary tumor size >4 cm, Hürthle cell histology, and the presence of marked vascular invasion.

TREATMENT

TREATMENT

Well-Differentiated Thyroid Cancer Surgery

All well-differentiated thyroid cancers >1cm (T1b or larger) should be surgically excised although active surveillance may be an option for small intrathyroidal micropapillary thyroid cancers (T1a) without metastases. In addition to

removing the primary lesion, surgery allows accurate histologic diagnosis and staging. Because there is no compelling evidence that bilateral thyroid surgery improves survival, the initial surgical procedure may be either a unilateral (lobectomy) or bilateral (near total thyroidectomy) procedure for patients with intrathyroidal cancers >1 cm and <4 cm (T1b and T2 tumors) in the absence of metastatic disease. For patients at high risk for recurrence, bilateral surgery allows administration of radioiodine for remnant ablation and potential treatment of iodine-avid metastases, if indicated, as well as for monitoring of serum Tg levels. Therefore, near-total thyroidectomy is appropriate for tumors >4 cm or in the presence of metastases or clinical evidence of extrathyroidal invasion. In addition, for patients found to have a high risk tumor after lobectomy based upon aggressive pathology features (e.g., vascular invasion or a less differentiated subtype), completion surgery should be performed. Surgical complication rates are acceptably low if the surgeon is highly experienced in the procedure. Preoperative sonography should be performed in all patients to assess the central and lateral cervical lymph node compartments for suspicious adenopathy, which if present, should undergo FNA and be removed, as indicated, at surgery.

TSH SUPPRESSION THERAPY

Because most tumors are still TSH-responsive, levothyroxine suppression of TSH is a mainstay of thyroid cancer treatment. Although TSH suppression clearly provides therapeutic benefit, there are no prospective studies that define the optimal level of TSH suppression. The degree of TSH suppression should be individualized based on a patient's risk of recurrence. It should be adjusted over time as surveillance blood tests and imaging confirm absence of disease or, alternatively, indicate possible residual/recurrent cancer. For patients at low risk of recurrence, TSH should be maintained in the lower normal limit (0.5–2.0 mIU/L). For patients either at intermediate or high risk of recurrence, TSH levels should be kept to 0.1 to 0.5 mIU/L and <0.1 mIU/L, respectively, if there are no strong contraindications to mild thyrotoxicosis. TSH should be <0.1 mIU/L for those with known metastatic disease.

RADIOIODINE TREATMENT

After near-total thyroidectomy, <1 gm of thyroid tissue remains in the thyroid bed. Postsurgical radioablation of the remnant thyroid eliminates residual normal thyroid, facilitating the use of Tg determinations. In addition, well-differentiated thyroid cancer often incorporates radioiodine, although less efficiently than normal thyroid follicular cells. Radioiodine uptake is determined primarily by expression of the NIS and is stimulated by TSH, requiring expression of the TSH-R. The retention time for radioactivity is influenced by the extent to which the tumor retains differentiated functions such as iodide trapping and organification. Consequently, for patients at higher risk of recurrence and for those with known distant metastatic disease, ^{131}I therapy may provide an adjuvant role and potentially treat residual tumor cells.

Indications

Not all patients benefit from radioiodine therapy. Neither recurrence nor survival rates are improved in stage I patients with T1 tumors (≤2 cm) confined to the thyroid. No benefit has been demonstrated for larger (>2 cm but <4 cm) low-risk tumors, such as minimally invasive follicular cancer or encapsulated PTC follicular variant. However, in higher risk patients (larger tumors, more aggressive variants of papillary cancer, tumor vascular invasion, extrathyroidal invasion, presence of large-volume lymph node metastases), radioiodine reduces recurrence and may increase survival for older patients.

^{131}I Thyroid Ablation and Treatment

As noted above, the decision to use ^{131}I for thyroid ablation should be coordinated with the surgical approach, because radioablation is much more effective when there is minimal remaining normal thyroid tissue. Radioiodine is administered after iodine depletion (patient follows a low-iodine diet for 1–2 weeks) and in the presence of elevated

serum TSH levels to stimulate uptake of the isotope into both the remnant and potentially any residual tumor. To achieve high serum TSH levels, there are two approaches. A patient may be withdrawn from thyroid hormone so that endogenous TSH is secreted and, ideally, the serum TSH level is >25 mIU/L at the time of ^{131}I therapy. A typical strategy is to treat the patient for several weeks postoperatively with liothyronine (25 μg qd or bid), followed by thyroid hormone withdrawal for 2 weeks. Alternatively, recombinant human TSH (rhTSH) is administered as two daily consecutive injections (0.9 mg) with administration of ^{131}I 24 h after the second injection. The patient can continue to take levothyroxine and remains euthyroid. Both approaches have equal success in achieving remnant ablation.

A pretreatment scanning dose of ^{131}I (usually 111 MBq [3 mCi]) or ^{123}I (74 MBq [2 mCi]) can reveal the amount of residual tissue and provides guidance about the dose needed to accomplish ablation. However, because of concerns about radioactive “stunning” that impairs subsequent treatment, there is a trend to avoid pretreatment scanning with ^{131}I and use either ^{123}I or proceed directly to ablation, unless there is suspicion that the amount of residual tissue will alter therapy or that there is distant metastatic disease. In the United States, outpatient doses of up to 6475 MBq (175 mCi) can be given at most centers. The administered dose depends on the indication for therapy with lower doses of 1100 MBq (30 mCi) given for remnant ablation but higher doses of up to 5500 MBq (150 mCi) used as adjuvant therapy when residual disease is suspected or present. A whole-body scanning (WBS) following radioiodine treatment is used to confirm the ^{131}I uptake in the remnant and to identify possible metastatic disease.

Surveillance Testing

Serum thyroglobulin is a sensitive marker of residual/recurrent thyroid cancer after ablation of the residual postsurgical thyroid tissue. Current Tg assays have functional sensitivities as low as 0.1 ng/mL, as opposed to older assays with functional sensitivities of 1–2 ng/mL, reducing the number of patients with truly undetectable serum Tg levels. Because the vast majority of PTC recurrences are in cervical lymph nodes, a neck ultrasound should be performed about 6 months after thyroid ablation; ultrasound has been shown to be more sensitive than WBS in this scenario.

In low-risk patients who have no clinical evidence of residual disease after ablation, negative cervical sonography, and a basal Tg <0.2 ng/mL on levothyroxine, the risk of structural recurrence is <3% at 5 years, and the frequency of follow-up testing can be decreased to annual TSH and Tg testing, with only periodic ultrasound examination.

The use of WBS is reserved for patients with known iodine-avid metastases or those with elevated serum thyroglobulin levels and negative imaging with ultrasound, chest CT, neck cross-sectional imaging and positron emission tomography (PET) CT who may require additional ^{131}I therapy.

In addition to radioiodine, external beam radiotherapy is also used to treat gross residual neck disease or specific metastatic lesions, particularly when they cause bone pain or threaten neurologic injury (e.g., vertebral metastases).

New Potential Therapies

Kinase inhibitors are being explored as a means to target pathways known to be active in thyroid cancer, including the RAS, BRAF, RET, EGFR, VEGFR, and angiogenesis pathways. A multicenter randomized controlled trial of the multikinase inhibitor sorafenib in 417 patients with progressive metastatic thyroid cancer reported a doubling of progression-free survival to 10.8 months in the treatment group compared with the placebo group. Ongoing trials are exploring whether differentiation protocols with kinase inhibitors or other approaches might enhance radioiodine uptake and efficacy.

ANAPLASTIC AND OTHER FORMS OF THYROID CANCER

Anaplastic Thyroid Cancer

As noted above, ATC is a poorly differentiated and aggressive cancer. The prognosis is poor, and most patients die within 6 months of diagnosis. Because of the undifferentiated state of these tumors, the uptake of radioiodine is usually negligible, but it can be used therapeutically if there is residual uptake. Chemotherapy has been attempted with multiple agents, including anthracyclines and paclitaxel, but it is usually ineffective. External beam radiation therapy can be attempted and continued if tumors are responsive. Recent data demonstrate survival benefit with immune checkpoint inhibition therapy.

Thyroid Lymphoma

Lymphoma in the thyroid gland often arises in the background of Hashimoto's thyroiditis. A rapidly expanding thyroid mass suggests the possibility of this diagnosis. Diffuse large-cell lymphoma is the most common type in the thyroid. Biopsies reveal sheets of lymphoid cells that can be difficult to distinguish from small-cell lung cancer or ATC. These tumors are often highly sensitive to external radiation. Surgical resection should be avoided as initial therapy because it may spread disease that is otherwise localized to the thyroid. If staging indicates disease outside of the thyroid, treatment should follow guidelines used for other forms of lymphoma (**Chap. 104**).

MEDULLARY THYROID CARCINOMA

MTC can be sporadic or familial and accounts for about 5% of thyroid cancers. There are three familial forms of MTC: MEN 2A, MEN 2B, and familial MTC without other features of MEN (**Chap. 381**). In general, MTC is more aggressive in MEN 2B than in MEN 2A, and familial MTC is more aggressive than sporadic MTC. Elevated serum calcitonin provides a marker of residual or recurrent disease. All patients with MTC should be tested for *RET* mutations, because genetic counseling and testing of family members can be offered to those individuals who test positive for mutations.

The management of MTC is primarily surgical. Prior to surgery, pheochromocytoma should be excluded in all patients with a RET mutation. Unlike tumors derived from thyroid follicular cells, these tumors do not take up radioiodine. External radiation treatment and targeted kinase inhibitors may provide palliation in patients with advanced disease (**Chap. 381**).

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Thyroid Nodules

Palpable thyroid nodules are found in about 5% of adults, but the prevalence varies considerably worldwide. Given this high prevalence rate, practitioners may identify thyroid nodules on physical examination. However, the increased usage of diagnostic medical imaging (e.g., carotid ultrasound, cervical spine MRI) has led to an increased frequency of incidental nodule detection, accounting for the majority of patients currently presenting for nodule evaluation. The main goal of this evaluation is to identify, in a cost-effective manner, the small subgroup of individuals with malignant lesions that have the potential to be clinically significant.

Nodules are more common in iodine-deficient areas, in women, and with aging. Most palpable nodules are >1 cm in diameter, but the ability to feel a nodule is influenced by its location within the gland (superficial versus deeply embedded), the anatomy of the patient's neck, and the experience of the examiner. More sensitive methods of

detection, such as CT, thyroid ultrasound, and pathologic studies, reveal thyroid nodules in up to 50% of glands in individuals aged >50 years. The presence of these thyroid incidentalomas has led to much debate about how to detect nodules and which nodules to investigate further.

An approach to the evaluation of a solitary nodule is outlined in **Fig. 378-4**. Most patients with thyroid nodules have normal thyroid function tests. Nonetheless, thyroid function should be assessed by measuring a TSH level, which may be suppressed by one or more autonomously functioning nodules. If the TSH is suppressed, a radionuclide scan is indicated to determine if the identified nodule is "hot," as lesions with increased uptake are almost never malignant and FNA is unnecessary. Otherwise, the next step in evaluation is performance of a thyroid ultrasound for three reasons: (1) Ultrasound will confirm if the palpable nodule is indeed a nodule. About 15% of "palpable" nodules are not confirmed on imaging, and therefore, no further evaluation is required. (2) Ultrasound will assess if there are additional nonpalpable nodules for which FNA may be recommended based on imaging features and size. (3) Ultrasound will characterize the imaging pattern of the nodule, which, combined with the nodule's size, facilitate decision-making about FNA. Numerous studies have demonstrated consistent risk estimates for thyroid cancer based upon certain sonographic patterns. For example, a spongiform nodule has a <3% chance of cancer and observation rather than FNA is reasonable, whereas 10–20% of solid hypoechoic nodules with smooth borders are malignant and FNA is recommended at a size cutoff of 1 cm (Table 378-1, Fig. 378-1). Evidence-based guidelines from both the American Thyroid Association (ATA) and the American Association of Clinical Endocrinologists provide recommendations for nodule FNA based on sonographic patterns and size cut offs, with lower size cut offs for nodules with more suspicious ultrasound patterns. Given what is known about the prevalence and generally indolent behavior of small thyroid cancers <1 cm, the 2015 ATA guidelines do not recommend FNA for any nodule <1 cm unless metastatic cervical lymph nodes are present.

FNA biopsy, ideally performed with ultrasound guidance, is the best diagnostic test when performed by physicians familiar with the procedure and when the results are interpreted by experienced cytopathologists. The technique is particularly useful for detecting PTC. However, the distinction between benign and malignant follicular lesions is often not possible using cytology alone because of the absence of characteristic nuclear features in follicular carcinoma. In several large studies, FNA biopsies yielded the following cytology diagnoses: 65% benign, 5% malignant or suspicious for malignancy, 10% nondiagnostic or yielding insufficient material for diagnosis, and 20% indeterminate. The Bethesda System is now widely used to provide more uniform terminology for reporting thyroid nodule FNA cytology results. This six-tiered classification system with the respective estimated malignancy rates is shown in **Table 378-5**. Specifically, the Bethesda System subcategorized cytology specimens previously labeled as indeterminate into three categories: atypia or follicular lesion of undetermined significance (AUS/FLUS), follicular neoplasm, and suspicious for malignancy.

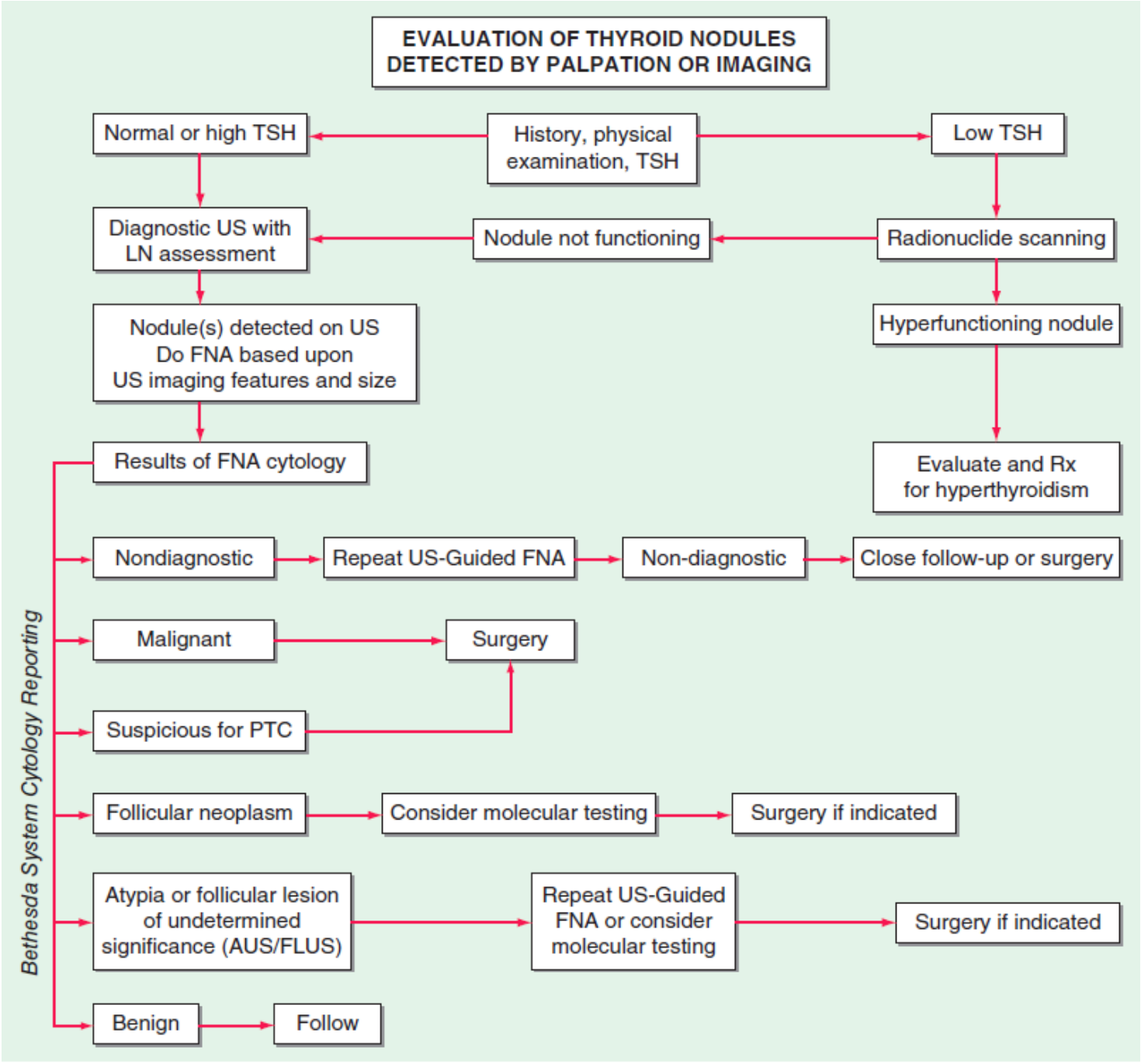
Cytology results indicative of malignancy generally mandate surgery, after performing preoperative sonography to evaluate the cervical lymph nodes. Nondiagnostic cytology specimens most often result from cystic lesions but may also occur in fibrous long-standing nodules. Ultrasound-guided FNA is indicated when a repeat FNA is necessary. Repeat FNA will yield a diagnostic cytology in about 50% of cases. Benign nodules may be monitored by ultrasound for growth, and repeat FNA may be considered if the nodule enlarges. The use of levothyroxine to suppress serum TSH is not effective in shrinking nodules in iodine-replete populations, and therefore, levothyroxine should not be used. The three indeterminate cytology classifications introduced by the Bethesda System are associated with different risks of malignancy (Table 378-5). For nodules with suspicious for malignancy cytology, surgery is recommended after ultrasound assessment of cervical lymph nodes. Options to be discussed with the patient include lobectomy versus total thyroidectomy.

On the other hand, the majority of nodules with AUS/FLUS and follicular neoplasm cytology results are benign; only 10–30% are malignant. The traditional approach for these patients is diagnostic lobectomy for histopathologic diagnosis. Therefore, up to 85% of patients undergo surgery for benign nodules. A high-sensitivity (~90%) novel molecular test using gene expression profiling technology may reduce the need for unnecessary surgery in these two groups. In a multicenter trial of over 265 such nodules, a negative gene expression classifier test reduced the risk of malignancy to about 6%, leading to clinical recommendations for follow-up rather than surgery. In addition, based upon results from next generation sequencing, molecular diagnostic panels, which include point mutations, small insertions/deletions, and gene fusions, are currently under investigation with the two goals: (1) identification and risk stratification of thyroid cancers based upon a positive result; (2) reduction in cancer risk to an acceptable level for nonsurgical surveillance based upon a negative result.

The evaluation of a thyroid nodule is stressful for most patients. They are concerned about the possibility of thyroid cancer, whether verbalized or not. It is constructive, therefore, to review the diagnostic approach and to reassure patients when no malignancy is found. When a suspicious lesion or thyroid cancer is identified, the generally favorable prognosis and available treatment options can be reassuring.

FIGURE 378-4

Approach to the patient with a thyroid nodule. See text and references for details. FNA, fine-needle aspiration; LN, lymph node; PTC, papillary thyroid cancer; TSH, thyroid-stimulating hormone; US, ultrasound.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 378-5

Bethesda Classification for Thyroid Cytology

DIAGNOSTIC CATEGORY	RISK OF MALIGNANCY
I. Nondiagnostic or unsatisfactory	1–5%
II. Benign	2–4%
III. Atypia or follicular lesion of unknown significance (AUS/FLUS)	5–15%
IV. Follicular neoplasm	15–30%
V. Suspicious for malignancy	60–75%
VI. Malignant	97–100%

FURTHER READING

Davies L, Welch HG: Current thyroid cancer trends in the United States. JAMA Otolaryngol Head Neck Surg 140:317, 2014. [PubMed: 24557566]

Fagin JA, Wells SA: Biologic and clinical perspectives on thyroid cancer. N Engl J Med 375:1054, 2016. [PubMed: 27626519]

Haugen BR et al: 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer. Thyroid 26:1, 2016. [PubMed: 26462967]

Tuttle RM, Haugen B, Perrier ND: Updated American Joint Committee and Cancer/Tumor-Node-Metastasis Staging System for differentiated and anaplastic thyroid cancer (eighth edition): What changed and why? Thyroid 27:751, 2017. [PubMed: 28463585]

Chapter 379: Disorders of the Adrenal Cortex

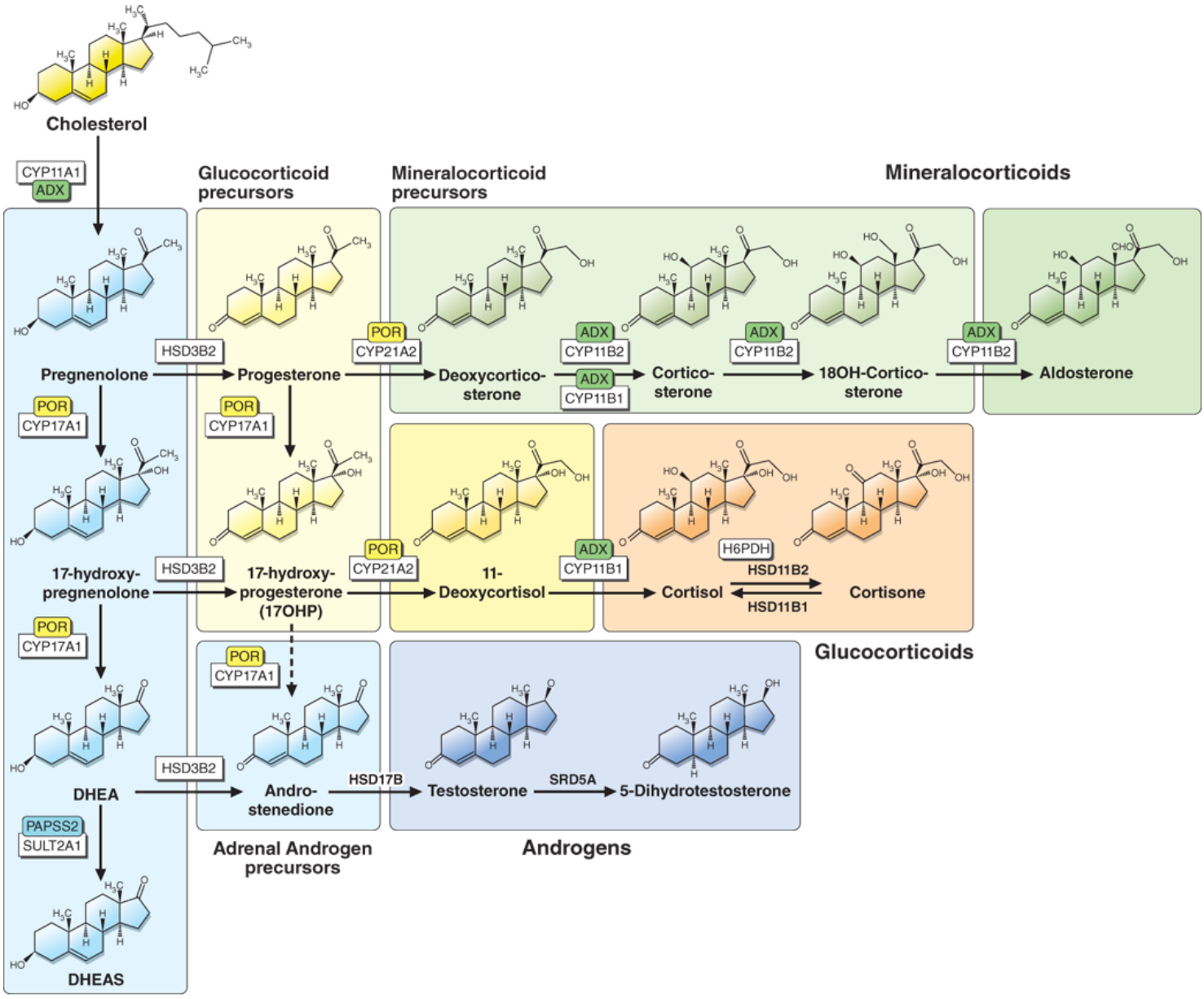
Wiebke Arlt

INTRODUCTION

The adrenal cortex produces three classes of corticosteroid hormones: glucocorticoids (e.g., cortisol), mineralocorticoids (e.g., aldosterone), and adrenal androgen precursors (e.g., dehydroepiandrosterone [DHEA]) **(Fig. 379-1)**. Glucocorticoids and mineralocorticoids act through specific nuclear receptors, regulating aspects of the physiologic stress response as well as blood pressure and electrolyte homeostasis. Adrenal androgen precursors are converted in the gonads and peripheral target cells to sex steroids that act via nuclear androgen and estrogen receptors.

FIGURE 379-1

Adrenal steroidogenesis. ADX, adrenodoxin; CYP11A1, side chain cleavage enzyme; CYP11B1, 11β-hydroxylase; CYP11B2, aldosterone synthase; CYP17A1, 17α-hydroxylase/17,20 lyase; CYP21A2, 21-hydroxylase; DHEA, dehydroepiandrosterone; DHEAS, dehydroepiandrosterone sulfate; H6PDH, hexose-6-phosphate dehydrogenase; HSD11B1, 11β-hydroxysteroid dehydrogenase type 1; HSD11B2, 11β-hydroxysteroid dehydrogenase type 2; HSD17B, 17β-hydroxysteroid dehydrogenase; HSD3B2, 3β-hydroxysteroid dehydrogenase type 2; PAPSS2, PAPS synthase type 2; POR, P450 oxidoreductase; SRD5A, 5α-reductase; SULT2A1, DHEA sulfotransferase.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Disorders of the adrenal cortex are characterized by deficiency or excess of one or several of the three major corticosteroid classes. Hormone deficiency can be caused by inherited glandular or enzymatic disorders or by destruction of the pituitary or adrenal gland by autoimmune disorders, infection, infarction, or iatrogenic events such as surgery or hormonal suppression. Hormone excess is usually the result of neoplasia, leading to increased production of adrenocorticotropic hormone (ACTH) by the pituitary or neuroendocrine cells (ectopic ACTH) or increased production of glucocorticoids, mineralocorticoids, or adrenal androgen precursors by adrenal nodules. Adrenal nodules are increasingly identified incidentally during abdominal imaging performed for other reasons.

ADRENAL ANATOMY AND DEVELOPMENT

The normal adrenal glands weigh 6–11 g each. They are located above the kidneys and have their own blood supply. Arterial blood flows initially to the subcapsular region and then meanders from the outer cortical zona glomerulosa through the intermediate zona fasciculata to the inner zona reticularis and eventually to the adrenal medulla. The right suprarenal vein drains directly into the vena cava, while the left suprarenal vein drains into the left renal vein.

During early embryonic development, the adrenals originate from the urogenital ridge and then separate from gonads and kidneys at about the sixth week of gestation. Concordant with the time of sexual differentiation (seventh to ninth week of gestation, Chap. 383), the adrenal cortex starts to produce cortisol and the adrenal sex steroid precursor DHEA. The orphan nuclear receptors SF1 (steroidogenic factor 1; encoded by the gene *NR5A1*) and DAX1 (dosage-sensitive sex reversal gene 1; encoded by the gene *NR0B1*), among others, play a crucial role during this period of development, as they regulate a multitude of adrenal genes involved in steroidogenesis.

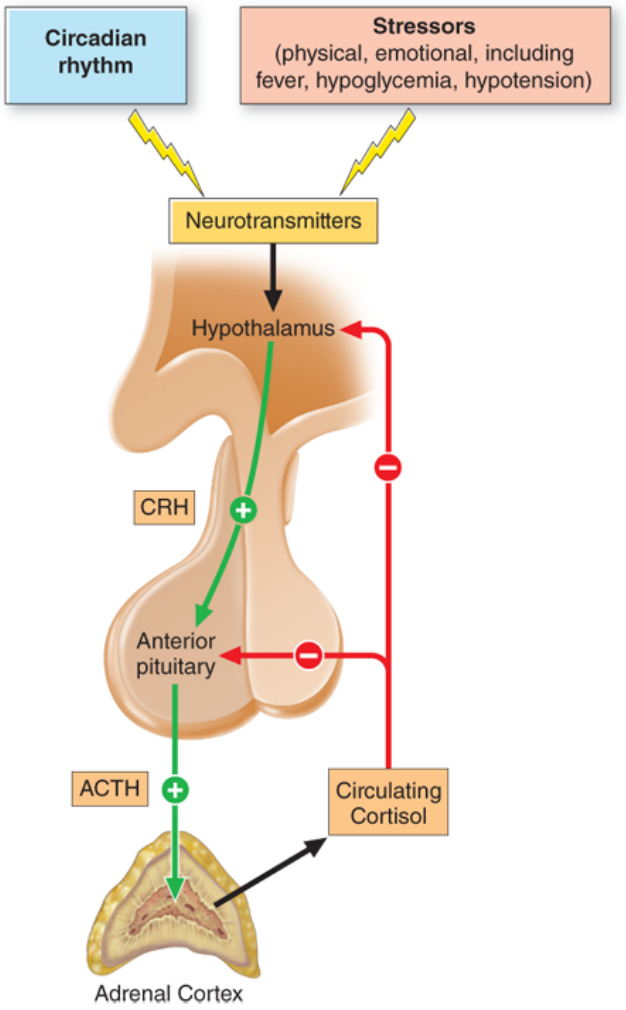
REGULATORY CONTROL OF STEROIDOGENESIS

Production of glucocorticoids and adrenal androgens is under the control of the hypothalamic-pituitary-adrenal (HPA) axis, whereas mineralocorticoids are regulated by the renin-angiotensin-aldosterone (RAA) system.

Glucocorticoid synthesis is under inhibitory feedback control by the hypothalamus and the pituitary **(Fig. 379-2)**. Hypothalamic release of corticotropin-releasing hormone (CRH) occurs in response to endogenous or exogenous stress. CRH stimulates the cleavage of the 241–amino acid polypeptide proopiomelanocortin (POMC) by pituitary-specific prohormone convertase 1 (PC1), yielding the 39–amino acid peptide ACTH. ACTH is released by the corticotrope cells of the anterior pituitary and acts as the pivotal regulator of adrenal cortisol synthesis, with additional short-term effects on mineralocorticoid and adrenal androgen synthesis. The release of CRH, and subsequently ACTH, occurs in a pulsatile fashion that follows a circadian rhythm under the control of the hypothalamus, specifically its suprachiasmatic nucleus (SCN), with additional regulation by a complex network of cell-specific clock genes. Reflecting the pattern of ACTH secretion, adrenal cortisol secretion exhibits a distinct circadian rhythm, starting to rise in the early morning hours prior to awakening, with peak levels in the morning and low levels in the evening **(Fig. 379-3)**.

FIGURE 379-2

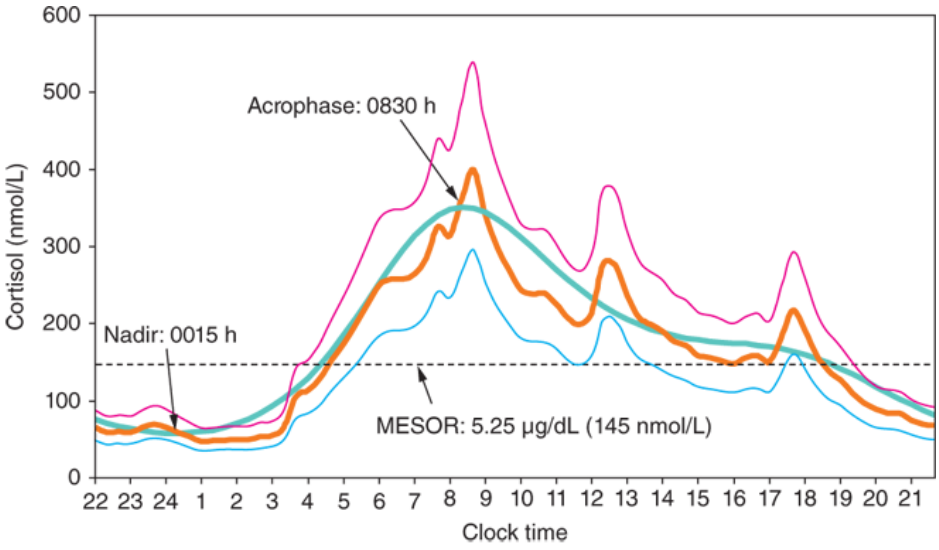
Regulation of the hypothalamic-pituitary-adrenal (HPA) axis. ACTH, adrenocorticotropic hormone; CRH, corticotropin-releasing hormone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 379-3

Physiologic cortisol circadian rhythm. Circulating cortisol concentrations (geometrical mean ± standard deviation values and fitted cosinor) drop under the rhythm-adjusted mean (MESOR) in the early evening hours, with nadir levels around midnight and a rise in the early morning hours; peak levels are observed ~8:30 AM (acrophase). *(Modified after M Debono et al: Modified-release hydrocortisone to provide circadian cortisol profiles. J Clin Endocrinol Metab 94:1548, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

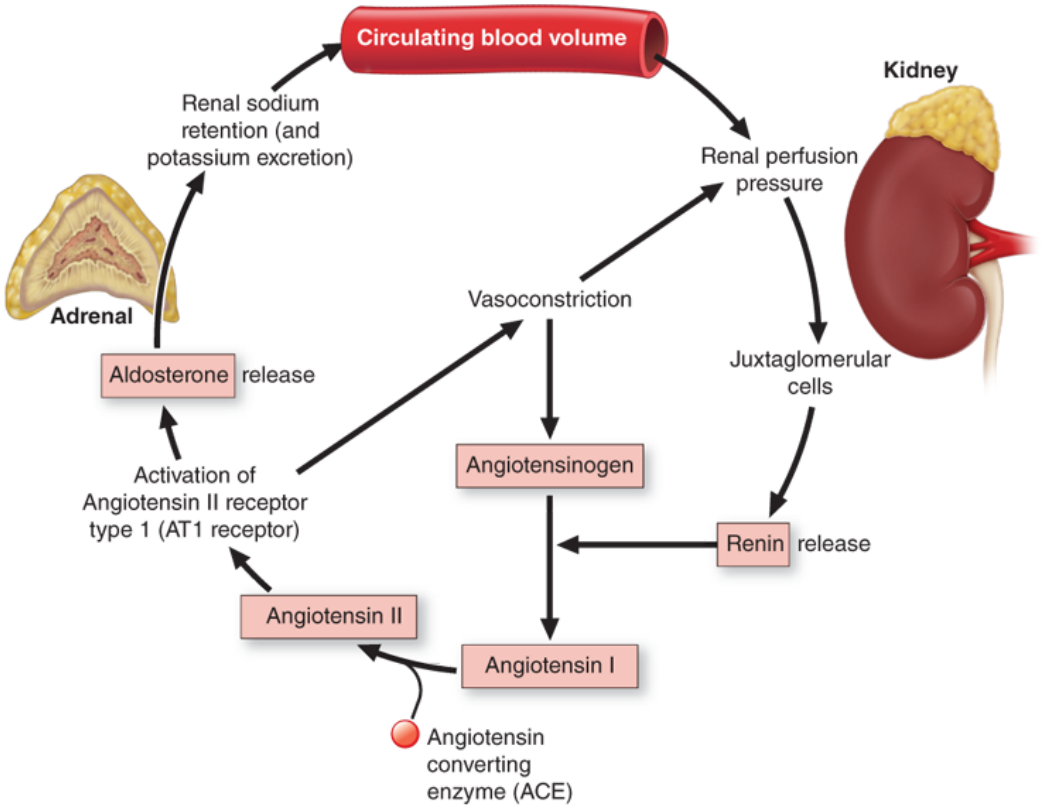
Diagnostic tests assessing the HPA axis make use of the fact that it is regulated by negative feedback. Glucocorticoid excess is diagnosed by employing a dexamethasone suppression test. Dexamethasone, a potent synthetic glucocorticoid, suppresses CRH/ACTH by binding hypothalamic-pituitary glucocorticoid receptors (GRs) and, therefore, results in downregulation of endogenous cortisol synthesis. Various versions of the dexamethasone suppression test are described in detail in **Chap. 373**. If cortisol production is autonomous (e.g., adrenal nodule), ACTH is already suppressed and dexamethasone has little additional effect. If cortisol production is driven by an ACTH-producing pituitary adenoma, dexamethasone suppression is ineffective at low doses but usually induces suppression at high doses. If cortisol production is driven by an ectopic source of ACTH, the tumors are usually resistant to dexamethasone suppression. Thus, the dexamethasone suppression test is useful to establish the diagnosis of Cushing's syndrome and to assist with the differential diagnosis of cortisol excess.

Conversely, to assess glucocorticoid deficiency, ACTH stimulation of cortisol production is used. The ACTH peptide contains 39 amino acids but the first 24 are sufficient to elicit a physiologic response. The standard ACTH stimulation test involves administration of cosyntropin (ACTH 1-24), 0.25 mg IM or IV, and collection of blood samples at 0, 30, and 60 min for cortisol. A normal response is defined as a cortisol level >20 µg/dL (>550 nmol/L) 30–60 min after cosyntropin stimulation. A low-dose (1 µg cosyntropin IV) version of this test has been advocated; however, it has no superior diagnostic value and is more cumbersome to carry out. Alternatively, an insulin tolerance test (ITT) can be used to assess adrenal function. It involves injection of insulin to induce hypoglycemia, which represents a strong stress signal that triggers hypothalamic CRH release and activation of the entire HPA axis. The ITT involves administration of regular insulin 0.1 U/kg IV (dose should be lower if hypopituitarism is likely) and collection of blood samples at 0, 30, 60, and 120 min for glucose, cortisol, and growth hormone (GH), if also assessing the GH axis. Oral or IV glucose is administered after the patient has achieved symptomatic hypoglycemia (usually glucose <40 mg/dL). A normal response is defined as a cortisol >20 µg/dL and GH >5.1 µg/L. The ITT requires careful clinical monitoring and sequential measurements of glucose. It is contraindicated in patients with coronary disease, cerebrovascular disease, or seizure disorders, which has made the short cosyntropin test the commonly accepted first-line test.

Mineralocorticoid production is controlled by the RAA regulatory cycle, which is initiated by the release of renin from the juxtaglomerular cells in the kidney, resulting in cleavage of angiotensinogen to angiotensin I in the liver **(Fig. 379-4)**. Angiotensin-converting enzyme (ACE) cleaves angiotensin I to angiotensin II, which binds and activates the angiotensin II receptor type 1 (AT1 receptor [AT1R]), resulting in increased adrenal aldosterone production and vasoconstriction. Aldosterone enhances sodium retention and potassium excretion, and increases the arterial perfusion pressure, which in turn regulates renin release. Because mineralocorticoid synthesis is primarily under the control of the RAA system, hypothalamic-pituitary damage does not significantly impact the capacity of the adrenal to synthesize aldosterone.

FIGURE 379-4

Regulation of the renin-angiotensin-aldosterone (RAA) system.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

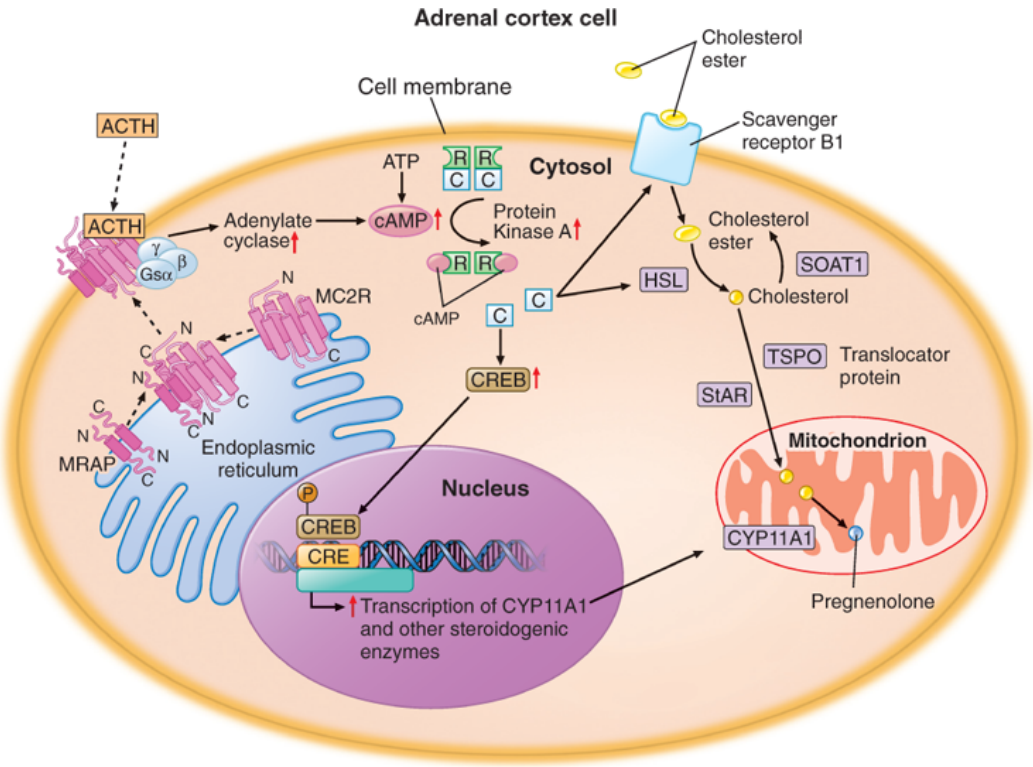
Similar to the HPA axis, the assessment of the RAA system can be used for diagnostic purposes. If mineralocorticoid excess is present, there is a counter-regulatory downregulation of plasma renin (see below for testing). Conversely, in mineralocorticoid deficiency, plasma renin is markedly increased. Physiologically, oral or IV sodium loading results in suppression of aldosterone, a response that is attenuated or absent in patients with autonomous mineralocorticoid excess.

STEROID HORMONE SYNTHESIS, METABOLISM, AND ACTION

ACTH stimulation is required for the initiation of steroidogenesis. The ACTH receptor MC2R (melanocortin 2 receptor) interacts with the MC2R-accessory protein MRAP, and the complex is transported to the adrenocortical cell membrane, where it binds to ACTH (Fig. 379-5). ACTH stimulation generates cyclic AMP (cAMP), which upregulates the protein kinase A (PKA) signaling pathway. Inactive PKA is a tetramer of two regulatory and two catalytic subunits that is dissociated by cAMP into a dimer of two regulatory subunits bound to cAMP and two free and active catalytic subunits. PKA activation impacts steroidogenesis in three distinct ways: (1) increases the import of cholesterol esters; (2) increases the activity of hormone-sensitive lipase, which cleaves cholesterol esters to cholesterol for import into the mitochondrion; and (3) increases the availability and phosphorylation of CREB (cAMP response element binding), a transcription factor that enhances transcription of CYP11A1 and other enzymes required for glucocorticoid synthesis.

FIGURE 379-5

ACTH effects on adrenal steroidogenesis. ACTH, adrenocorticotropic hormone; binding protein; HSL, hormone-sensitive lipase; MRAP, MC2R-accessory protein; protein kinase A catalytic subunit (C; *PRKACA*), PKA regulatory subunit (R; *PRKAR1A*); SOAT1, sterol O-acyltransferase 1; StAR, steroidogenic acute regulatory (protein); TSPO, translocator protein.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

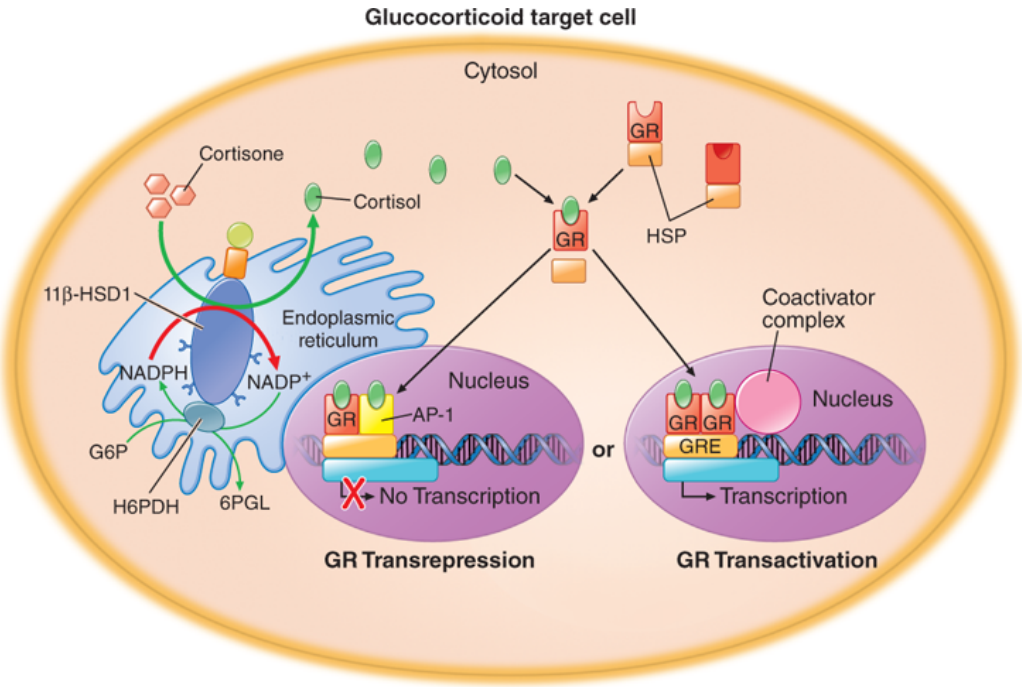
Adrenal steroidogenesis occurs in a zone-specific fashion, with mineralocorticoid synthesis occurring in the outer zona glomerulosa, glucocorticoid synthesis in the zona fasciculata, and adrenal androgen synthesis in the inner zona reticularis (Fig. 379-1). All steroidogenic pathways require cholesterol import into the mitochondrion, a process initiated by the action of the steroidogenic acute regulatory (StAR) protein, which shuttles cholesterol from the outer to the inner mitochondrial membrane. The majority of steroidogenic enzymes are cytochrome P450 (CYP) enzymes, which are either located in the mitochondrion (side chain cleavage enzyme, CYP11A1; 11β-hydroxylase, CYP11B1; aldosterone synthase, CYP11B2) or in the endoplasmic reticulum membrane (17α-hydroxylase, CYP17A1; 21-hydroxylase, CYP21A2; aromatase, CYP19A1). These enzymes require electron donation via specific redox cofactor enzymes, P450 oxidoreductase (POR), and adrenodoxin/adrenodoxin reductase (ADX/ADR) for the microsomal and mitochondrial CYP enzymes, respectively. In addition, the short-chain dehydrogenase 3β-hydroxysteroid dehydrogenase type 2 (3β-HSD2), also termed Δ4, Δ5 isomerase, plays a major role in adrenal steroidogenesis.

The cholesterol side chain cleavage enzyme CYP11A1 generates pregnenolone. Glucocorticoid synthesis requires conversion of pregnenolone to progesterone by 3β-HSD2, followed by conversion to 17-hydroxyprogesterone (17OHP) by CYP17A1, further hydroxylation at carbon 21 by CYP21A2, and eventually, 11β-hydroxylation by CYP11B1 to generate active cortisol (Fig. 379-1). Mineralocorticoid synthesis also requires progesterone, which is first converted to deoxycorticosterone (DOC) by CYP21A2 and then converted via corticosterone and 18-hydroxycorticosterone to aldosterone in three steps catalyzed by CYP11B2. For adrenal androgen synthesis, pregnenolone undergoes conversion by CYP17A1, which uniquely catalyzes two enzymatic reactions. Via its 17α-hydroxylase activity, CYP17A1 converts pregnenolone to 17-hydroxypregnenolone, followed by generation of the universal sex steroid precursor DHEA via CYP17A1 17,20 lyase activity. The majority of DHEA is secreted by the adrenal in the form of its sulfate ester, DHEAS, generated by DHEA sulfotransferase (SULT2A1).

Following its release from the adrenal, cortisol circulates in the bloodstream mainly bound to cortisol-binding globulin (CBG) and to a lesser extent to albumin, with only a minor fraction circulating as free, unbound hormone. Free cortisol is thought to enter cells directly, not requiring active transport. In addition, in a multitude of peripheral target tissues of glucocorticoid action, including adipose, liver, muscle, and brain, cortisol is generated from inactive cortisone within the cell by the enzyme 11β-hydroxysteroid dehydrogenase type 1 (11β-HSD1) **(Fig. 379-6)**. Thereby, 11β-HSD1 functions as a tissue-specific prereceptor regulator of glucocorticoid action. For the conversion of inactive cortisone to active cortisol, 11β-HSD1 requires nicotinamide adenine dinucleotide phosphate (NADPH [reduced form]), which is provided by the enzyme hexose-6-phosphate dehydrogenase (H6PDH). Like the catalytic domain of 11β-HSD1, H6PDH is located in the lumen of the endoplasmic reticulum, and converts glucose-6-phosphate (G6P) to 6-phosphogluconate (6PGL), thereby regenerating NADP+ to NADPH, which drives the activation of cortisol from cortisone by 11β-HSD1.

FIGURE 379-6

Prereceptor activation of cortisol and glucocorticoid receptor (GR) action. AP-1, activator protein-1; G6P, glucose-6-phosphate; GREs, glucocorticoid response elements; HSPs, heat shock proteins; NADPH, nicotinamide adenine dinucleotide phosphate (reduced form); 6PGL, 6-phosphogluconate.



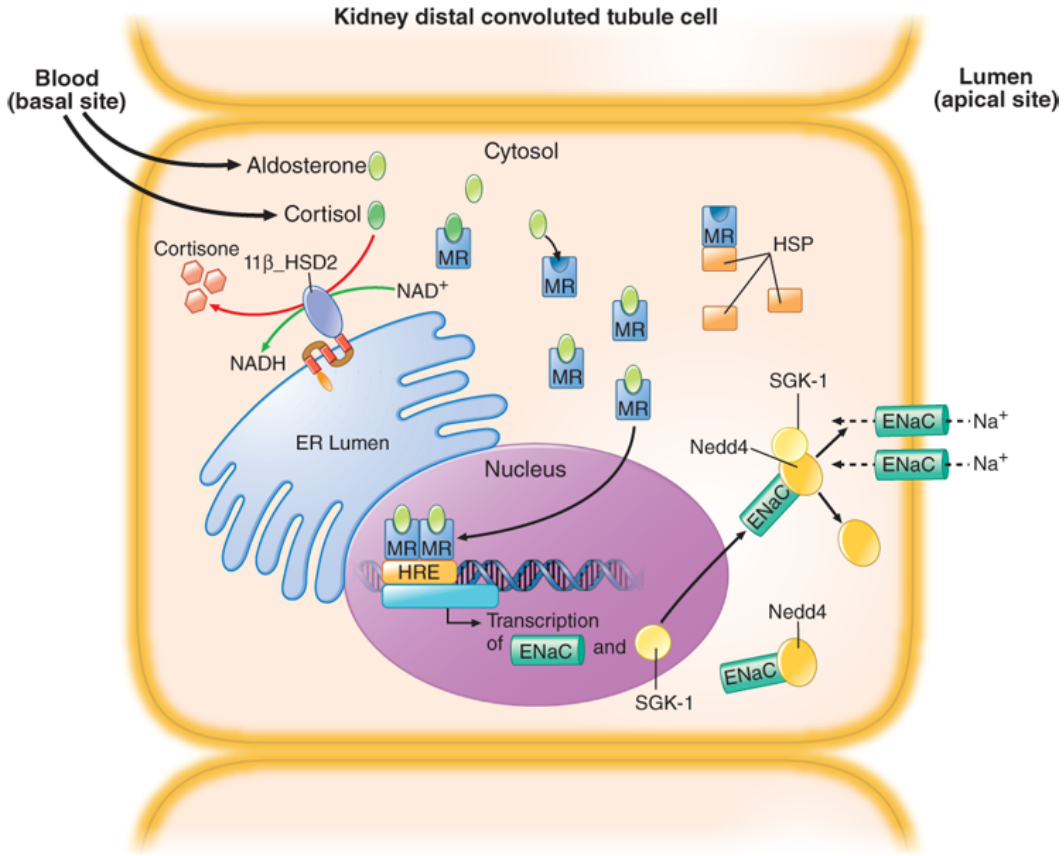
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In the cytosol of target cells, cortisol binds and activates the GR, which results in dissociation of heat shock proteins (HSPs) from the receptor and subsequent dimerization (Fig. 379-6). Cortisol-bound GR dimers translocate to the nucleus and activate glucocorticoid response elements (GREs) in the DNA sequence, thereby enhancing transcription of glucocorticoid-regulated genes (GR transactivation). However, cortisol-bound GR can also form heterodimers with transcription factors such as AP-1 or NF-κB, resulting in transrepression of proinflammatory genes, a mechanism of major importance for the anti-inflammatory action of glucocorticoids. It is important to note that corticosterone also exerts glucocorticoid activity, albeit much weaker than cortisol itself. However, in rodents, corticosterone is the major glucocorticoid, and in patients with 17-hydroxylase deficiency, lack of cortisol can be compensated for by higher concentrations of corticosterone that accumulates as a consequence of the enzymatic block.

Cortisol is inactivated to cortisone by the microsomal enzyme 11β-hydroxysteroid dehydrogenase type 2 (11β-HSD2) **(Fig. 379-7)**, mainly in the kidney, but also in the colon, salivary glands, and other target tissues. Cortisol and aldosterone bind the mineralocorticoid receptor (MR) with equal affinity; however, cortisol circulates in the bloodstream at about a 1000-fold higher concentration. Thus, only rapid inactivation of cortisol to cortisone by 11β-HSD2 prevents MR activation by excess cortisol, thereby acting as a tissue-specific modulator of the MR pathway. In addition to cortisol and aldosterone, deoxycorticosterone (DOC) (Fig. 379-1) also exerts mineralocorticoid activity. DOC accumulation due to 11β-hydroxylase deficiency or due to tumor-related excess production can result in mineralocorticoid excess.

FIGURE 379-7

Prereceptor inactivation of cortisol and mineralocorticoid receptor action. ENaC, epithelial sodium channel; HRE, hormone response element; NADH, nicotinamide adenine dinucleotide; SGK-1, serum glucocorticoid-inducible kinase-1.

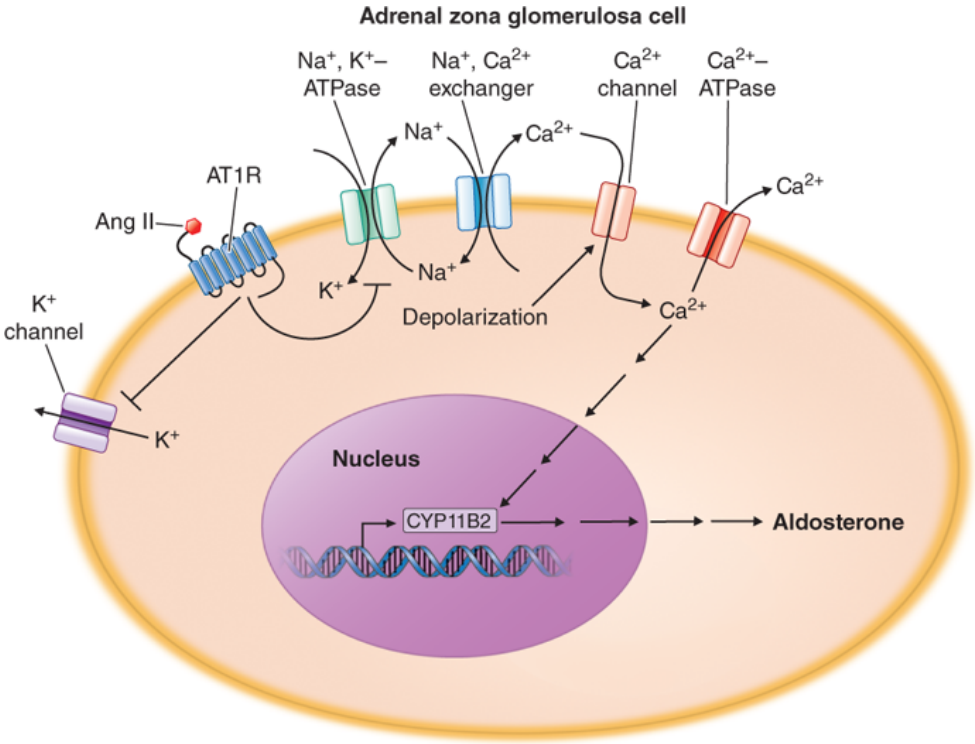


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Aldosterone synthesis in the adrenal zona glomerulosa cells is driven by the enzyme aldosterone synthase (CYP11B2). The binding of angiotensin II to the AT1 receptor causes glomerulosa cell membrane depolarization by increasing intracellular sodium through inhibition of sodium potassium (Na^+/K^+) ATPase enzymes as well as potassium channels. This drives an increase in intracellular calcium by opening of voltage-dependent calcium channels or inhibition of calcium (Ca^{2+}) ATPase enzymes. Consequently, the calcium signaling pathway is triggered, resulting in upregulation of CYP11B2 transcription **(Fig. 379-8)**.

FIGURE 379-8

Regulation of adrenal aldosterone synthesis. AngII, angiotensin II; AT1R, angiotensin II receptor type 1; CYP11B2, aldosterone synthase. *(Modified after F Beuschlein: Regulation of aldosterone secretion: from physiology to disease. Eur J Endocrinol 168:R85, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Analogous to cortisol action via the GR, aldosterone (or cortisol) binding to the MR in the kidney tubule cell dissociates the heat shock protein (HSP)–receptor complex, allowing homodimerization of the MR, and translocation of the hormone-bound MR dimer to the nucleus (Fig. 379-7). The activated MR enhances transcription of the epithelial sodium channel (ENaC) and serum glucocorticoid-inducible kinase 1 (SGK-1). In the cytosol, interaction of ENaC with Nedd4 prevents cell surface expression of ENaC. However, SGK-1 phosphorylates serine residues within the Nedd4 protein, reduces the interaction between Nedd4 and ENaC, and consequently, enhances the trafficking of ENaC to the cell surface, where it mediates sodium retention.

CUSHING'S SYNDROME

(See also Chap. 373) Cushing's syndrome reflects a constellation of clinical features that result from chronic exposure to excess glucocorticoids of any etiology. The disorder can be ACTH-dependent (e.g., pituitary corticotrope adenoma, ectopic secretion of ACTH by nonpituitary tumor) or ACTH-independent (e.g., adrenocortical adenoma, adrenocortical carcinoma [ACC], nodular adrenal hyperplasia), as well as iatrogenic (e.g., administration of exogenous glucocorticoids to treat various inflammatory conditions). The term *Cushing's disease* refers specifically to Cushing's syndrome caused by a pituitary corticotrope adenoma.

Epidemiology

Cushing's syndrome is generally considered a rare disease. It occurs with an incidence of 1–2 per 100,000 population per year. However, it is debated whether mild cortisol excess may be more prevalent among patients with features of Cushing's such as centripetal obesity, type 2 diabetes, and osteoporotic vertebral fractures, recognizing that these are relatively nonspecific and common in the population.

In the overwhelming majority of patients, Cushing's syndrome is caused by an ACTH-producing corticotrope adenoma of the pituitary (Table 379-1), as initially described by Harvey Cushing in 1912. Cushing's disease more frequently affects women, with the exception of prepubertal cases, where it is more common in boys. By contrast, ectopic ACTH syndrome is more frequently identified in men. Only 10% of patients with Cushing's syndrome have a primary, adrenal cause of their disease (e.g., autonomous cortisol excess independent of ACTH), and most of these patients are women. However, overall, the medical use of glucocorticoids for immunosuppression, or for the treatment of inflammatory disorders, is the most common cause of Cushing's syndrome.

TABLE 379-1

Causes of Cushing's Syndrome

Causes of Cushing's Syndrome	Female:Male Ratio	%
ACTH-Dependent Cushing's		**90**
Cushing's disease (= ACTH-producing pituitary adenoma)	4:1	75
Ectopic ACTH syndrome (due to ACTH secretion by bronchial or pancreatic carcinoid tumors, small-cell lung cancer, medullary thyroid carcinoma, pheochromocytoma, and others)	1:1	15
ACTH-Independent Cushing's	**4:1**	**10**
Adrenocortical adenoma		5–10
Adrenocortical carcinoma		1
Rare causes: macronodular adrenal hyperplasia; primary pigmented nodular adrenal disease (micro- and/or macronodular); McCune-Albright syndrome		<1

Abbreviation: ACTH, adrenocorticotropic hormone.

Etiology

In at least 90% of patients with Cushing's disease, ACTH excess is caused by a corticotrope pituitary microadenoma, often only a few millimeters in diameter. Pituitary macroadenomas (i.e., tumors >1 cm in size) are found in only 5–10% of patients. Pituitary corticotrope adenomas usually occur sporadically but very rarely can be found in the context of multiple endocrine neoplasia type 1 (MEN 1) **(Chap. 381)**. Pituitary adenomas causative of Cushing's disease frequently harbor mutations in the deubiquitinase USP8, which leads to constitutive activation of EGF signaling and consequent upregulated expression of the ACTH precursor POMC. USP8 mutations are found more frequently in adults (41 vs 17% in children) and in women (43 vs 17% in men) with Cushing's disease.

Ectopic ACTH production is predominantly caused by occult carcinoid tumors, most frequently in the lung, but also in thymus or pancreas. Because of their small size, these tumors are often difficult to locate. Advanced small-cell lung cancer can cause ectopic ACTH production. In rare cases, ectopic CRH and/or ACTH production has been found to originate from medullary thyroid carcinoma or pheochromocytoma, the latter co-secreting catecholamines and ACTH.

The majority of patients with ACTH-independent cortisol excess harbor a cortisol-producing adrenal adenoma; intratumor, i.e., somatic mutations in the PKA catalytic subunit PRKACA have been identified as cause of disease in 40% of these tumors. ACCs may also cause ACTH-independent disease and are often large, with excess production of several corticosteroid classes.

A rare but notable cause of adrenal cortisol excess is macronodular adrenal hyperplasia with low circulating ACTH, but with evidence for autocrine stimulation of cortisol production via intraadrenal ACTH production. These hyperplastic nodules are often also characterized by ectopic expression of G protein–coupled receptors not usually found in the adrenal, including receptors for luteinizing hormone, vasopressin, serotonin, interleukin 1, catecholamines, or gastric inhibitory peptide (GIP), the cause of food-dependent Cushing's. Activation of these receptors results in upregulation of PKA signaling, as physiologically occurs with ACTH, with a subsequent increase in cortisol production. A combination of germline and somatic mutations in the tumor-suppressor gene *ARMC5* have been identified as a prevalent cause of Cushing's due to bilateral macronodular adrenal hyperplasia; these patients often present with biochemical evidence of Cushing's but lack specific clinical signs, which develop slowly over decades and accelerate cardiovascular risk. Constitutively activating mutations in the PKA catalytic subunit PRKACA are found as somatic mutations in one-third of cortisol-producing adrenocortical adenomas and, as germline mutations, can also represent a rare cause of macronodular adrenal hyperplasia associated with cortisol excess.

Germline mutations in one of the regulatory subunits of PKA, PRKAR1A, are found in patients with primary pigmented nodular adrenal disease (PPNAD) as part of *Carney's complex*, an autosomal dominant multiple neoplasia condition associated with cardiac myxomas, hyperlentiginosis, Sertoli cell tumors, and PPNAD. PPNAD can present as micronodular or macronodular hyperplasia, or both. Phosphodiesterases can influence intracellular cAMP and can thereby impact PKA activation. Mutations in PDE11A and PDE8B have been identified in patients with bilateral adrenal hyperplasia and Cushing's, with and without evidence of PPNAD.

Another rare cause of ACTH-independent Cushing's is *McCune-Albright syndrome*, also associated with polyostotic fibrous dysplasia, unilateral café-au-lait spots, and precocious puberty. McCune-Albright syndrome is caused by activating mutations in the stimulatory G protein alpha subunit 1, GNAS-1 (guanine nucleotide-binding protein alpha stimulating activity polypeptide 1), and such mutations have also been found in bilateral macronodular hyperplasia without other McCune-Albright features and, in rare instances, also in isolated cortisol-producing adrenal adenomas (Table 379-1; **Chap. 405**).

Clinical Manifestations

Glucocorticoids affect almost all cells of the body, and thus signs of cortisol excess impact multiple physiologic systems **(Table 379-2)**, with upregulation of gluconeogenesis, lipolysis, and protein catabolism causing the most prominent features. In addition, excess glucocorticoid secretion overcomes the ability of 11β-HSD2 to rapidly inactivate cortisol to cortisone in the kidney, thereby exerting mineralocorticoid actions, manifest as diastolic hypertension, hypokalemia, and edema. Excess glucocorticoids also interfere with central regulatory systems, leading to suppression of gonadotropins with subsequent hypogonadism and amenorrhea, and suppression of the hypothalamic-pituitary-thyroid axis, resulting in decreased thyroid-stimulating hormone (TSH) secretion.

TABLE 379-2

Signs and Symptoms of Cushing's Syndrome

Body Compartment/System	Signs and Symptoms
Body fat	Weight gain, central obesity, rounded face, fat pad on back of neck ("buffalo hump")
Skin	Facial plethora, thin and brittle skin, easy bruising, broad and purple stretch marks, acne, hirsutism
Bone	Osteopenia, osteoporosis (vertebral fractures), decreased linear growth in children
Muscle	Weakness, proximal myopathy (prominent atrophy of gluteal and upper leg muscles with difficulty climbing stairs or getting up from a chair)
Cardiovascular system	Hypertension, hypokalemia, edema, atherosclerosis
Metabolism	Glucose intolerance/diabetes, dyslipidemia
Reproductive system	Decreased libido, in women amenorrhea (due to cortisol-mediated inhibition of gonadotropin release)
Central nervous system	Irritability, emotional lability, depression, sometimes cognitive defects; in severe cases, paranoid psychosis
Blood and immune system	Increased susceptibility to infections, increased white blood cell count, eosinopenia, hypercoagulation with increased risk of deep vein thrombosis and pulmonary embolism

The majority of clinical signs and symptoms observed in Cushing's syndrome are relatively nonspecific and include features such as obesity, diabetes, diastolic hypertension, hirsutism, and depression that are commonly found in patients who do not have Cushing's. Therefore, careful clinical assessment is an important aspect of evaluating suspected cases. A diagnosis of Cushing's should be considered when several clinical features are found in the same patient, in particular when more specific features are found. These include fragility of the skin, with easy bruising and broad (>1 cm), purplish striae **(Fig. 379-9)**, and signs of proximal myopathy, which becomes most obvious when trying to stand up from a chair without the use of hands or when climbing stairs. Clinical manifestations of Cushing's do not differ substantially among the different causes of Cushing's. In ectopic ACTH syndrome, hyperpigmentation of the knuckles, scars, or skin areas exposed to increased friction can be observed (Fig. 379-9) and is caused by stimulatory effects of excess ACTH and other POMC cleavage products on melanocyte pigment production. Furthermore, patients with ectopic ACTH syndrome, and some with ACC as the cause of Cushing's, may have a more brisk onset and rapid progression of clinical signs and symptoms.

FIGURE 379-9

Clinical features of Cushing's syndrome. ***A.*** Note central obesity and broad, purple stretch marks (***B.*** close-up). ***C.*** Note thin and brittle skin in an elderly patient with Cushing's syndrome. ***D.*** Hyperpigmentation of the knuckles in a patient with ectopic adrenocorticotropic hormone (ACTH) excess.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

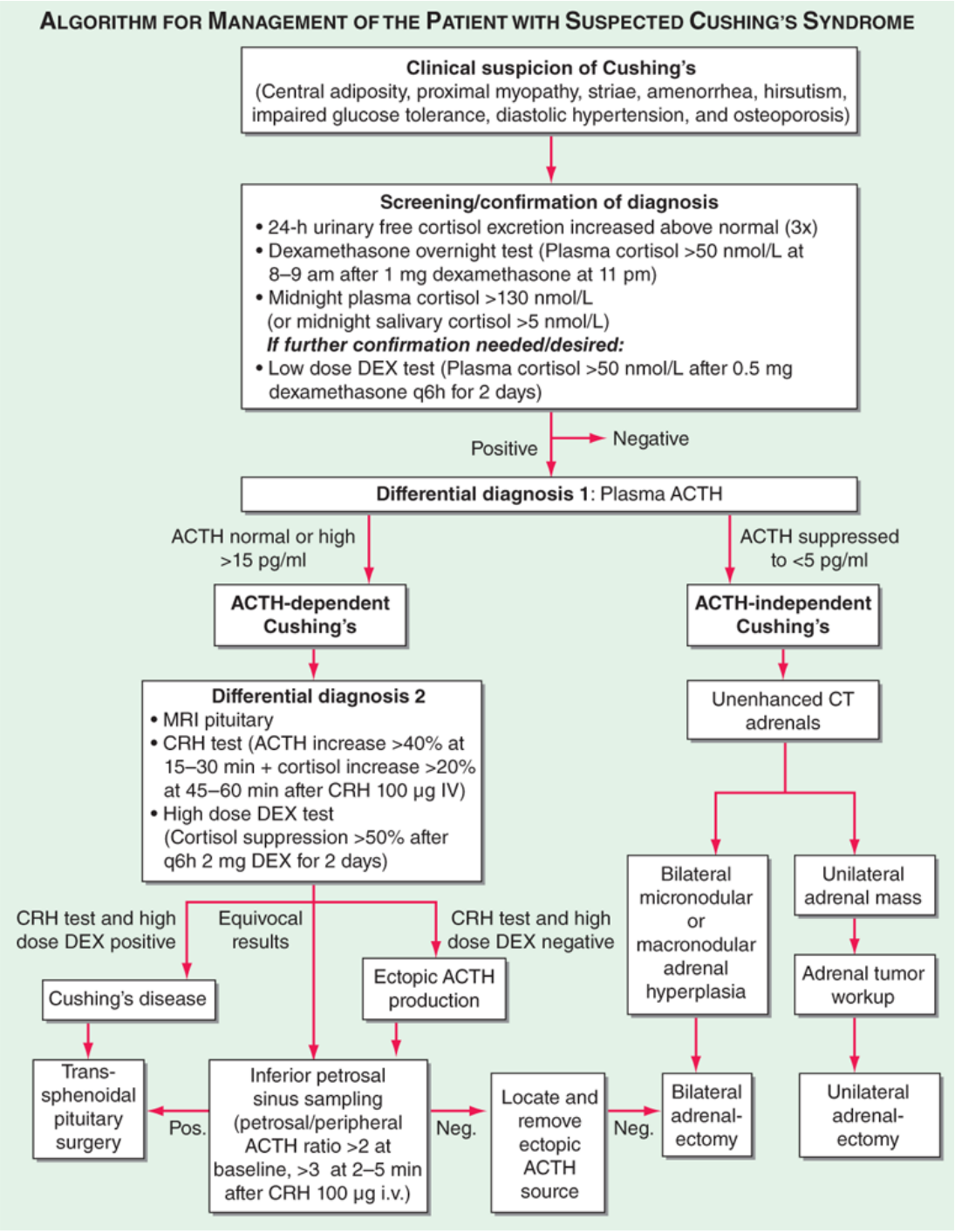
Patients with Cushing's syndrome can be acutely endangered by deep vein thrombosis, with subsequent pulmonary embolism due to a hypercoagulable state associated with Cushing's. The majority of patients also experience psychiatric symptoms, mostly in the form of anxiety or depression, but acute paranoid or depressive psychosis may also occur. Even after cure, long-term health may be affected by persistently impaired health-related quality of life and increased risk of cardiovascular disease and osteoporosis with vertebral fractures, depending on the duration and degree of exposure to significant cortisol excess.

Diagnosis

The most important first step in the management of patients with suspected Cushing's syndrome is to establish the correct diagnosis. Most mistakes in clinical management, leading to unnecessary imaging or surgery, are made because the diagnostic protocol is not followed **(Fig. 379-10)**. This protocol requires establishing the diagnosis of Cushing's beyond doubt prior to employing any tests used for the differential diagnosis of the condition. In principle, after excluding exogenous glucocorticoid use as the cause of clinical signs and symptoms, suspected cases should be tested if there are multiple and progressive features of Cushing's, particularly features with a potentially higher discriminatory value. Exclusion of Cushing's is also indicated in patients with incidentally discovered adrenal masses.

FIGURE 379-10

Management of the patient with suspected Cushing's syndrome. ACTH, adrenocorticotropic hormone; CRH, corticotropin-releasing hormone; CT, computed tomography; DEX, dexamethasone; MRI, magnetic resonance imaging.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

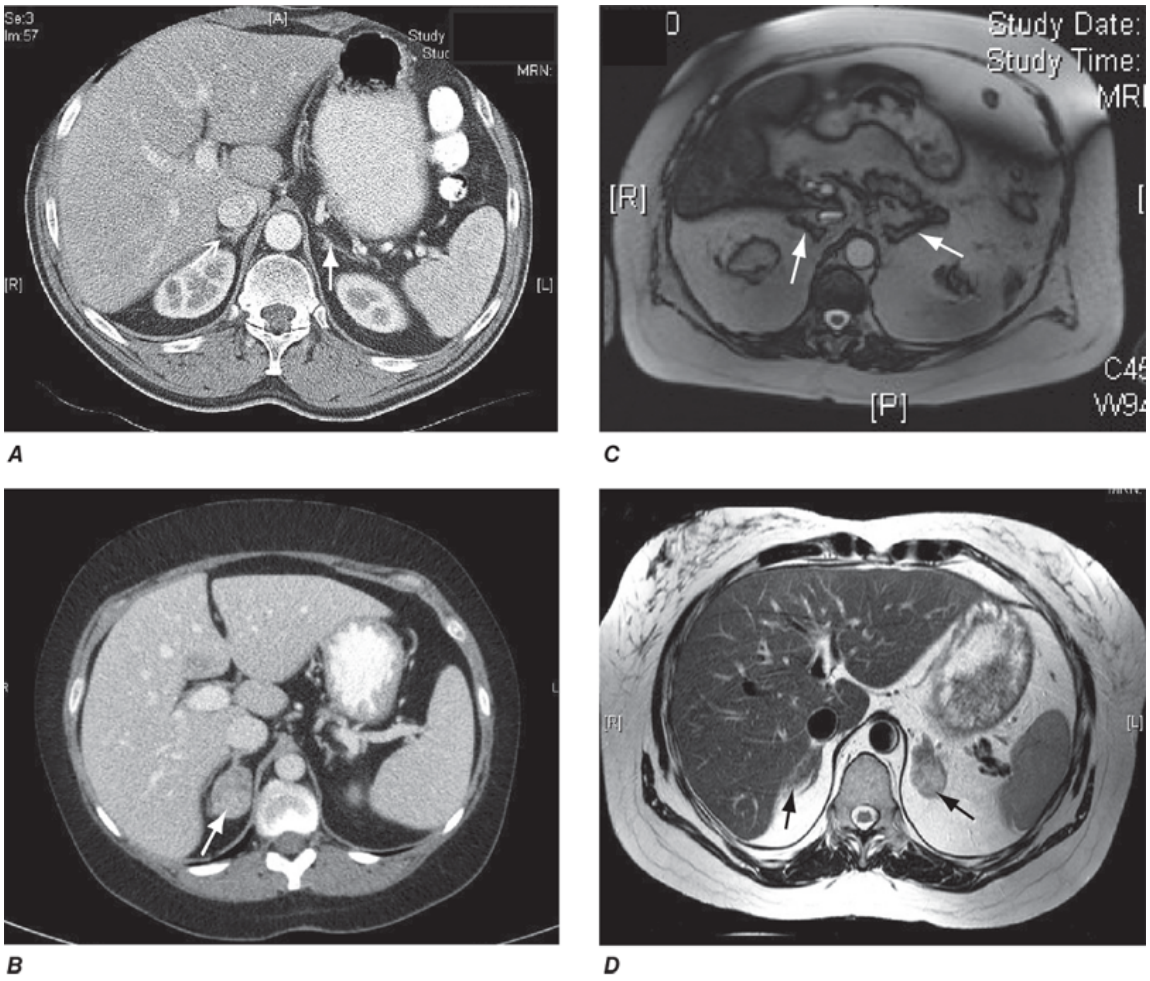
A diagnosis of Cushing's can be considered as established if the results of several tests are consistently suggestive of Cushing's. These tests may include increased 24-h urinary free cortisol excretion in three separate collections, failure to appropriately suppress morning cortisol after overnight exposure to dexamethasone, and evidence of loss of diurnal cortisol secretion with high levels at midnight, the time of the physiologically lowest secretion (Fig. 379-10). Factors potentially affecting the outcome of these diagnostic tests have to be excluded such as incomplete 24-h urine collection or rapid inactivation of dexamethasone due to concurrent intake of CYP3A4-inducing drugs (e.g., antiepileptics, rifampicin). Concurrent intake of oral contraceptives that raise CBG and thus total cortisol can cause failure to suppress after dexamethasone. If in doubt, testing should be repeated after 4–6 weeks off estrogens. Patients with pseudo-Cushing states, i.e., alcohol-related, and those with cyclic Cushing's may require further testing to safely confirm or exclude the diagnosis of Cushing's. In addition, the biochemical assays employed can affect the test results, with specificity representing a common problem with antibody-based assays for the measurement of urinary free cortisol. These assays have been greatly improved by the introduction of highly specific tandem mass spectrometry.

Differential Diagnosis

The evaluation of patients with confirmed Cushing's should be carried out by an endocrinologist and begins with the differential diagnosis of ACTH-dependent and ACTH-independent cortisol excess (Fig. 379-10). Generally, plasma ACTH levels are suppressed in cases of autonomous adrenal cortisol excess, as a consequence of enhanced negative feedback to the hypothalamus and pituitary. By contrast, patients with ACTH-dependent Cushing's have normal or increased plasma ACTH, with very high levels being found in some patients with ectopic ACTH syndrome. Importantly, imaging should only be used after it is established whether the cortisol excess is ACTH-dependent or ACTH-independent, because nodules in the pituitary or the adrenal are a common finding in the general population. In patients with confirmed ACTH-independent excess, adrenal imaging is indicated **(Fig. 379-11)**, preferably using an unenhanced computed tomography (CT) scan. This allows assessment of adrenal morphology and determination of precontrast tumor density in Hounsfield units (HUs), which helps to distinguish between benign and malignant adrenal lesions.

FIGURE 379-11

Adrenal imaging in Cushing's syndrome. ***A.*** Adrenal computed tomography (CT) showing normal bilateral adrenal morphology (*arrows*). ***B.*** CT scan depicting a right adrenocortical adenoma (*arrow*) causing Cushing's syndrome. ***C.*** Magnetic resonance imaging (MRI) showing bilateral adrenal hyperplasia due to excess adrenocorticotropic hormone stimulation in Cushing's disease. ***D.*** MRI showing bilateral macronodular hyperplasia causing Cushing's syndrome.



A C

B D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

For ACTH-dependent cortisol excess **(Chap. 373)**, a magnetic resonance image (MRI) of the pituitary is the investigation of choice, but it may not show an abnormality in up to 40% of cases because of small tumors below the sensitivity of detection. Characteristically, pituitary corticotrope adenomas fail to enhance following gadolinium administration on T1-weighted MRI images. In all cases of confirmed ACTH-dependent Cushing's, further tests are required for the differential diagnosis of pituitary Cushing's disease and ectopic ACTH syndrome. These tests exploit the fact that most pituitary corticotrope adenomas still display regulatory features, including residual ACTH suppression by high-dose glucocorticoids and CRH responsiveness. In contrast, ectopic sources of ACTH are typically resistant to dexamethasone suppression and unresponsive to CRH (Fig. 379-10). However, it should be noted that a small minority of ectopic ACTH-producing tumors exhibit dynamic responses similar to pituitary corticotrope tumors. If the two tests show discordant results, or if there is any other reason for doubt, the differential diagnosis can be further clarified by performing bilateral inferior petrosal sinus sampling (IPSS) with concurrent blood sampling for ACTH in the right and left inferior petrosal sinus and a peripheral vein. An increased central/peripheral plasma ACTH ratio >2 at baseline and >3 at 2–5 min after CRH injection is indicative of Cushing's disease (Fig. 379-10), with very high sensitivity and specificity. Of note, the results of the IPSS cannot be reliably used for lateralization (i.e., prediction of the location of the tumor within the pituitary), because there is broad interindividual variability in the venous drainage of the pituitary region. Importantly, no cortisol-lowering agents should be used prior to IPSS.

If the differential diagnostic testing indicates ectopic ACTH syndrome, then further imaging should include high-resolution, fine-cut CT scanning of the chest and abdomen for scrutiny of the lung, thymus, and pancreas. If no lesions are identified, an MRI of the chest can be considered because carcinoid tumors usually show high signal intensity on T2-weighted images. Furthermore, octreotide scintigraphy can be helpful in some cases because ectopic ACTH-producing tumors often express somatostatin receptors. Depending on the suspected cause, patients with ectopic ACTH syndrome should also undergo blood sampling for fasting gut hormones, chromogranin A, calcitonin, and biochemical exclusion of pheochromocytoma.

TREATMENT

TREATMENT

Cushing's Syndrome

Overt Cushing's is associated with a poor prognosis if left untreated. In ACTH-independent disease, treatment consists of surgical removal of the adrenal tumor. For smaller tumors, a minimally invasive approach can be used, whereas for larger tumors and those suspected of malignancy, an open approach is preferred.

In Cushing's disease, the treatment of choice is selective removal of the pituitary corticotrope tumor, usually via an endoscopic transsphenoidal approach. This results in an initial cure rate of 70–80% when performed by a highly experienced surgeon. However, even after initial remission following surgery, long-term follow-up is important because late relapse occurs in a significant number of patients. If pituitary disease recurs, there are several options, including second surgery, radiotherapy, stereotactic radiosurgery, and bilateral adrenalectomy. These options need to be applied in a highly individualized fashion.

In some patients with very severe, overt Cushing's (e.g., difficult to control hypokalemic hypertension or acute psychosis), it may be necessary to introduce medical therapy to rapidly control the cortisol excess during the period leading up to surgery. Similarly, patients with metastasized, glucocorticoid-producing carcinomas may require long-term antiglucocorticoid drug treatment. In case of ectopic ACTH syndrome, in which the tumor cannot be located, one must carefully weigh whether drug treatment or bilateral adrenalectomy is the most appropriate choice, with the latter facilitating immediate cure but requiring life-long corticosteroid replacement. In this instance, it is paramount to ensure regular imaging follow-up for identification of the ectopic ACTH source.

Oral agents with established efficacy in Cushing's syndrome are metyrapone and ketoconazole. Metyrapone inhibits cortisol synthesis at the level of 11β-hydroxylase (Fig. 379-1), whereas the antimycotic drug ketoconazole inhibits the early steps of steroidogenesis. Typical starting doses are 500 mg tid for metyrapone (maximum dose, 6 g) and 200 mg tid for ketoconazole (maximum dose, 1200 mg). Mitotane, a derivative of the insecticide o,p'DDD, is an adrenolytic agent that is also effective for reducing cortisol. Because of its side effect profile, it is most commonly used in the context of ACC, but low-dose treatment (500–1000 mg/d) has also been used in benign Cushing's. In severe cases of cortisol excess, etomidate, an agent that potently blocks 11β-hydroxylase and aldosterone synthase, can be used to lower cortisol. It is administered by continuous IV infusion in low, nonanesthetic doses.

After the successful removal of an ACTH- or cortisol-producing tumor, the HPA axis will remain suppressed. Thus, hydrocortisone replacement needs to be initiated at the time of surgery and slowly tapered following recovery, to allow physiologic adaptation to normal cortisol levels. Depending on degree and duration of cortisol excess, the HPA axis may require many months or even years to resume normal function and sometimes does not recover. Generally, ectopic ACTH syndrome shows the best recovery rate (80%) and adrenal Cushing's has the lowest (40%), with Cushing's disease intermediate (60%).

MINERALOCORTICOID EXCESS

Epidemiology

Following the first description of a patient with an aldosterone-producing adrenal adenoma (*Conn's syndrome*), mineralocorticoid excess was thought to represent a rare cause of hypertension. However, in studies systematically screening all patients with hypertension, a much higher prevalence is now recognized, ranging from 5 to 12%. The prevalence is higher when patients are preselected for hypokalemic hypertension.

Etiology

The most common cause of mineralocorticoid excess is primary aldosteronism, reflecting excess production of aldosterone by the adrenal zona glomerulosa. Bilateral micronodular hyperplasia is somewhat more common than unilateral adrenal adenomas **(Table 379-3)**. Somatic mutations in channels and enzymes responsible for increasing sodium and calcium influx in adrenal zona glomerulosa cells have been identified as prevalent causes of aldosterone-producing adrenal adenomas (Table 379-3) and, in the case of germline mutations, also of primary aldosteronism due to bilateral macronodular adrenal hyperplasia. However, bilateral adrenal hyperplasia as a cause of mineralocorticoid excess is usually micronodular, but can also contain larger nodules that might be mistaken for a unilateral adenoma. In rare instances, primary aldosteronism is caused by an ACC. Carcinomas should be considered in younger patients and in those with larger tumors, because benign aldosterone-producing adenomas usually measure <2 cm in diameter.

TABLE 379-3

Causes of Mineralocorticoid Excess

Causes of Mineralocorticoid Excess	Mechanism	%
Primary Aldosteronism		
Adrenal (Conn's) adenoma	Autonomous aldosterone excess can be caused by somatic (intratumor) mutations in the potassium channel GIRK4 (encoded by *KCNJ5*; identified as cause of disease in 40% of aldosterone-producing adenomas; rare germline mutations can cause bilateral macronodular adrenal hyperplasia). Further causes include somatic mutations affecting the α-subunit of the Na^+/K^+-ATPase (encoded by *ATP1A1*), the plasma membrane calcium-transporting ATPase 3 (encoded by *ATP2B3*), and somatic mutations in *CACNA1D* or *CACNA1H* encoding the voltage-gated calcium channel CaV 1.3 and CaV3.2, respectively. All mutations result in upregulation of CYP11B2 and hence aldosterone synthesis.	60
Bilateral (micronodular) adrenal hyperplasia	Autonomous aldosterone excess, mostly micronodular and rarely macronodular, with germline *KCNJ5* mutations being a rare cause.	60
Glucocorticoid-remediable hyperaldosteronism (dexamethasone-suppressible hyperaldosteronism)	Crossover between the *CYP11B1* and *CYP11B2* genes results in ACTH-driven aldosterone production	<1
Other Causes (Rare)		**<1**
Syndrome of apparent mineralocorticoid excess (SAME)	Mutations in *HSD11B2* result in lack of renal inactivation of cortisol to cortisone, leading to excess activation of the MR by cortisol	
Cushing's syndrome	Cortisol excess overcomes the capacity of HSD11B2 to inactivate cortisol to cortisone, consequently flooding the MR	
Glucocorticoid resistance	Upregulation of cortisol production due to GR mutations results in flooding of the MR by cortisol	
Adrenocortical carcinoma	Autonomous aldosterone and/or DOC excess	
Congenital adrenal hyperplasia	Accumulation of DOC due to mutations in *CYP11B1* or *CYP17A1*	
Progesterone-induced hypertension	Progesterone acts as an abnormal ligand due to mutations in the MR gene	
Liddle's syndrome	Mutant ENaC β or γ subunits resulting in reduced degradation of ENaC keeping the membrane channel in open conformation for longer, enhancing mineralocorticoid action	

Abbreviations: ACTH, adrenocorticotropic hormone; DOC, deoxycorticosterone; ENaC, epithelial sodium channel; GR, glucocorticoid receptor; HSD11B2, 11β-hydroxysteroid dehydrogenase type 2; MR, mineralocorticoid receptor.

A rare cause of aldosterone excess is glucocorticoid-remediable aldosteronism (GRA), which is caused by a chimeric gene resulting from crossover of promoter sequences between the *CYP11B1* and *CYP11B2* genes that are involved in glucocorticoid and mineralocorticoid synthesis, respectively (Fig. 379-1). This rearrangement brings *CYP11B2* transcription under the control of ACTH receptor signaling; consequently, aldosterone production is regulated by

ACTH rather than by renin. The family history can be helpful because there may be evidence for dominant transmission of hypertension. Recognition of the disorder is important because it can be associated with early-onset hypertension and strokes. In addition, glucocorticoid suppression can reduce aldosterone production.

Other rare causes of mineralocorticoid excess are listed in Table 379-3. An important cause is excess binding and activation of the MR by a steroid other than aldosterone. Cortisol acts as a potent mineralocorticoid if it escapes efficient inactivation to cortisone by 11β-HSD2 in the kidney (Fig. 379-7). This can be caused by inactivating mutations in the *HSD11B2* gene resulting in the syndrome of apparent mineralocorticoid excess (SAME) that characteristically manifests with severe hypokalemic hypertension in childhood. However, milder mutations may cause normokalemic hypertension manifesting in adulthood (type II SAME). Inhibition of 11β-HSD2 by excess licorice ingestion also results in hypokalemic hypertension, as does overwhelming of 11β-HSD2 conversion capacity by cortisol excess in Cushing's syndrome. DOC also binds and activates the MR and can cause hypertension if its circulating concentrations are increased. This can arise through autonomous DOC secretion by an ACC, but also when DOC accumulates as a consequence of an adrenal enzymatic block, as seen in congenital adrenal hyperplasia (CAH) due to CYP11B1 (11β-hydroxylase) or CYP17A1 (17α-hydroxylase) deficiency (Fig. 379-1). Progesterone can cause hypokalemic hypertension in rare individuals who harbor a MR mutation that enhances binding and activation by progesterone; physiologically, progesterone normally exerts antimineralocorticoid activity. Finally, excess mineralocorticoid activity can be caused by mutations in the β or γ subunits of the ENaC, disrupting its interaction with Nedd4 (Fig. 379-7), and thereby decreasing receptor internalization and degradation. The constitutively active ENAC drives hypokalemic hypertension, resulting in an autosomal dominant disorder termed *Liddle's syndrome*.

Clinical Manifestations

Excess activation of the MR leads to potassium depletion and increased sodium retention, with the latter causing an expansion of extracellular and plasma volume. Increased ENaC activity also results in hydrogen depletion that can cause metabolic alkalosis. Aldosterone also has direct effects on the vascular system, where it increases cardiac remodeling and decreases compliance. Aldosterone excess may cause direct damage to the myocardium and the kidney glomeruli, in addition to secondary damage due to systemic hypertension.

The clinical hallmark of mineralocorticoid excess is hypokalemic hypertension; serum sodium tends to be normal due to the concurrent fluid retention, which in some cases can lead to peripheral edema. Hypokalemia can be exacerbated by thiazide drug treatment, which leads to increased delivery of sodium to the distal renal tubule, thereby driving potassium excretion. Severe hypokalemia can be associated with muscle weakness, overt proximal myopathy, or even hypokalemic paralysis. Severe alkalosis contributes to muscle cramps and, in severe cases, can cause tetany.

Of note, patients with primary aldosteronism show increased rates of osteoporosis, type 2 diabetes, and cognitive dysfunction. A significant proportion of patients suffer from concurrent mild autonomous cortisol excess (MACE), termed *Connshing syndrome*.

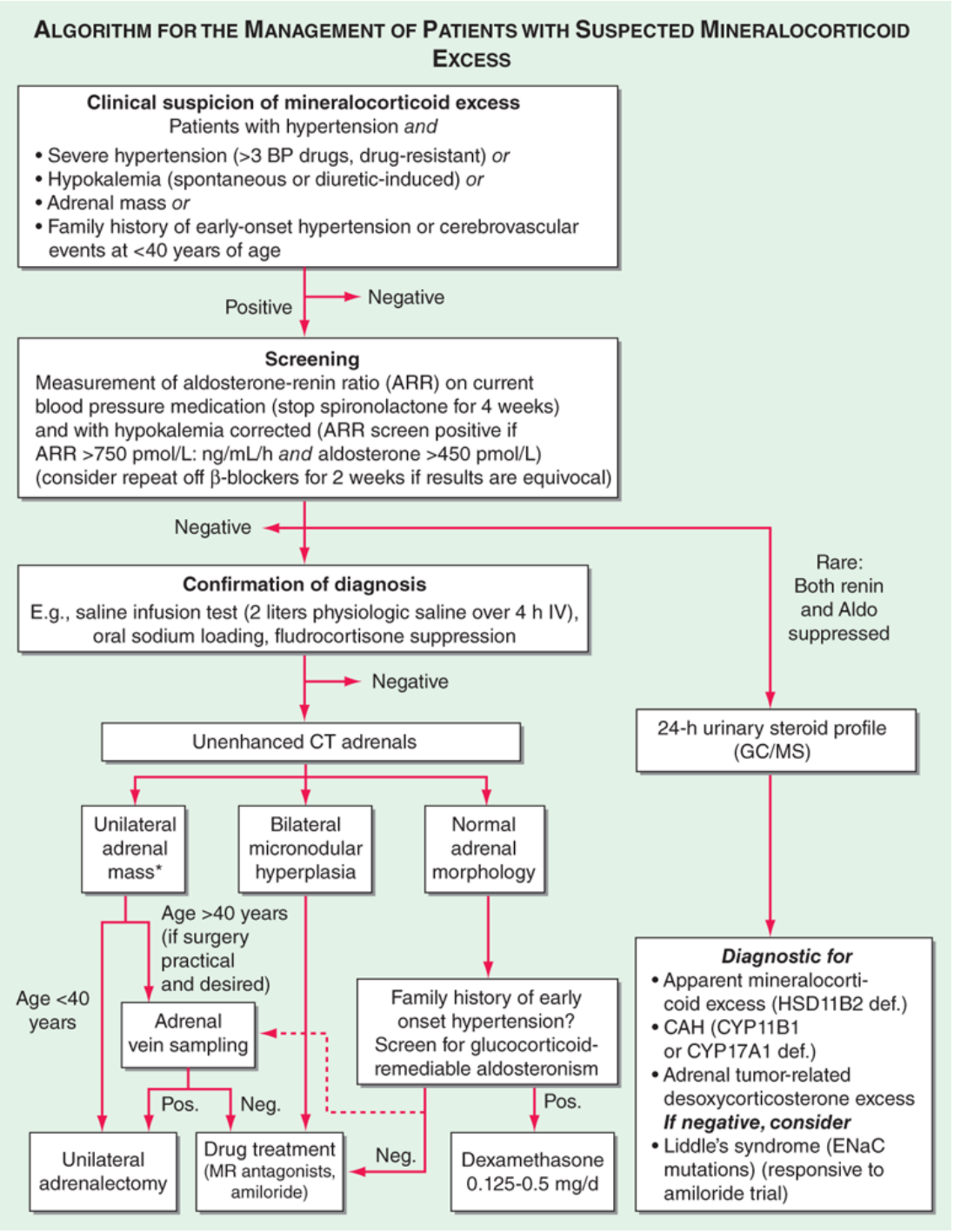
Diagnosis

Diagnostic screening for mineralocorticoid excess is not currently recommended for all patients with hypertension, but should be restricted to those who exhibit hypertension associated with drug resistance, hypokalemia, an adrenal mass, or onset of disease before the age of 40 years **(Fig. 379-12)**. The accepted screening test is concurrent measurement of plasma renin and aldosterone with subsequent calculation of the aldosterone-renin ratio (ARR) (Fig. 379-12); serum potassium needs to be normalized prior to testing. Stopping antihypertensive medication can be cumbersome, particularly in patients with severe hypertension. Thus, for practical purposes, in the first instance the patient can remain on the usual antihypertensive medications, with the exception that MR antagonists need to be ceased at least 4 weeks prior to ARR measurement. The remaining antihypertensive drugs usually do not affect the outcome of ARR testing, except that beta blocker treatment can cause false-positive results and ACE/AT1R inhibitors can cause false-negative results in milder cases **(Table 379-4)**.

FIGURE 379-12

Management of patients with suspected mineralocorticoid excess. *Perform adrenal tumor workup (see Fig. 379-13). BP, blood pressure; CAH, congenital adrenal hyperplasia; CT, computed tomography; GC/MS, gas chromatography/mass spectrometry; PRA, plasma renin activity.

ALGORITHM FOR THE MANAGEMENT OF PATIENTS WITH SUSPECTED MINERALOCORTICOID EXCESS



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 379-4

Effects of Antihypertensive Drugs on the Aldosterone-Renin Ratio (ARR)

Drug	Effect on Renin	Effect on Aldosterone	Net Effect on ARR
β Blockers	↓	↑	↑
α_1 Blockers	→	→	→
α_2 Sympathomimetics	→	→	→
ACE inhibitors	↑	↓	↓
AT1R blockers	↑	↓	↓
Calcium antagonists	→	→	→
Diuretics	(↑)	(↑)	→/(↓)

Abbreviations: ACE, angiotensin-converting enzyme; AT1R, angiotensin II receptor type 1.

ARR screening is positive if the ratio is >750 pmol/L per ng/mL per hour, with a concurrently high normal or increased aldosterone (Fig. 379-12). If one relies on the ARR only, the likelihood of a false-positive ARR becomes greater when renin levels are very low. The characteristics of the biochemical assays are also important. Some labs measure plasma renin activity, whereas others measure plasma renin concentrations. Antibody-based assays for the measurement of serum aldosterone lack the reliability of tandem mass spectrometry assays, but these are not yet ubiquitously available.

Diagnostic confirmation of mineralocorticoid excess in a patient with positive ARR screening result should be undertaken by an endocrinologist as the tests lack optimized validation. The most straightforward is the saline infusion test, which involves the IV administration of 2 L of physiologic saline over a 4-h period. Failure of aldosterone to suppress <140 pmol/L (5 ng/dL) is indicative of autonomous mineralocorticoid excess. Alternative tests are the oral sodium loading test (300 mmol NaCl/d for 3 days) or the fludrocortisone suppression test (0.1 mg q6h with 30 mmol NaCl q8h for 4 days); the latter can be difficult because of the risk of profound hypokalemia and increased hypertension. In patients with overt hypokalemic hypertension, strongly positive ARR, and concurrently increased aldosterone levels, confirmatory testing is usually not necessary.

Differential Diagnosis and Treatment

After the diagnosis of hyperaldosteronism is established, the next step is to use adrenal imaging to further assess the cause. Fine-cut CT scanning of the adrenal region is the method of choice because it provides excellent visualization of adrenal morphology. CT will readily identify larger tumors suspicious of malignancy but may miss lesions <5 mm. The differentiation between bilateral micronodular hyperplasia and a unilateral adenoma is only required if a surgical approach is feasible and desired. Consequently, selective adrenal vein sampling (AVS) should only be carried out in surgical candidates with either no obvious lesion on CT or evidence of a unilateral lesion in patients >40 years, because the latter patients have a high likelihood of harboring a coincidental, endocrine-inactive adrenal adenoma (Fig. 379-12). AVS is used to compare aldosterone levels in the inferior vena cava and between the right and left adrenal veins. AVS requires concurrent measurement of cortisol to document correct placement of the catheter in the adrenal veins and should demonstrate a cortisol gradient >3 between the vena cava and each adrenal vein. Lateralization is confirmed by an aldosterone/cortisol ratio that is at least twofold higher on one side than the other. AVS is a complex procedure that requires a highly skilled interventional radiologist. Even then, the right adrenal vein can be difficult to cannulate correctly, which, if not achieved, invalidates the procedure. There is also no agreement as to whether the two adrenal veins should be cannulated simultaneously or successively and whether ACTH stimulation enhances the diagnostic value of AVS.

Patients >40 years with confirmed mineralocorticoid excess and a unilateral lesion on CT can go straight to surgery, which is also indicated in patients with confirmed lateralization documented by a valid AVS procedure. Laparoscopic adrenalectomy is the preferred approach. Patients who are not surgical candidates, or with evidence of bilateral hyperplasia based on CT or AVS, should be treated medically (Fig. 379-12). Medical treatment, which can also be considered prior to surgery to avoid postsurgical hypoaldosteronism, consists primarily of the MR antagonist spironolactone. It can be started at 12.5–50 mg bid and titrated up to a maximum of 400 mg/d to control blood pressure and normalize potassium. Side effects include menstrual irregularity, decreased libido, and gynecomastia. The more selective MR antagonist eplerenone can also be used. Doses start at 25 mg bid, and it can be titrated up to 200 mg/d. Another useful drug is the sodium channel blocker amiloride (5–10 mg bid).

In patients with normal adrenal morphology and family history of early-onset, severe hypertension, a diagnosis of GRA should be considered and can be evaluated using genetic testing. Treatment of GRA consists of administering dexamethasone, using the lowest dose possible to control blood pressure. Some patients also require additional MR antagonist treatment.

The diagnosis of nonaldosterone-related mineralocorticoid excess is based on documentation of suppressed renin and suppressed aldosterone in the presence of hypokalemic hypertension. This testing is best carried out by employing urinary steroid metabolite profiling by gas chromatography/mass spectrometry (GC/MS). An increased free cortisol over free cortisone ratio is suggestive of SAME and can be treated with dexamethasone. Steroid profiling by GC/MS also detects the steroids associated with CYP11B1 and CYP17A1 deficiency or the irregular steroid secretion pattern in a DOC-producing ACC (Fig. 379-12). If the GC/MS profile is normal, then Liddle's syndrome should be considered. It is very sensitive to amiloride treatment but will not respond to MR antagonist treatment, because the defect is due to a constitutively active ENaC.

APPROACH TO THE PATIENT: INCIDENTALLY DISCOVERED ADRENAL MASS

Epidemiology

Incidentally discovered adrenal masses, commonly termed adrenal "incidentalomas," are common, with a prevalence of 2–5% in the general population as documented in CT and autopsy series. The prevalence increases with age, with 1% of 40-year-olds and 7% of 70-year-olds harboring an adrenal mass. The widespread use of cross-sectional imaging has also increased the recognized prevalence.

Etiology

Most solitary adrenal tumors are monoclonal neoplasms. Several genetic syndromes, including MEN 1 (*MEN1*), MEN 2 (*RET*), Carney's complex (*PRKAR1A*), and McCune-Albright (*GNAS1*), can have adrenal tumors as one of their features. Somatic mutations in *MEN1, GNAS1*, and *PRKAR1A* have been identified in a small proportion of sporadic adrenocortical adenomas. Aberrant expression of membrane receptors (GIP, α- and β-adrenergic, luteinizing hormone, vasopressin V1, and interleukin 1 receptors) has been identified in some sporadic cases of macronodular adrenocortical hyperplasia.

The majority of adrenal nodules are endocrine-inactive adrenocortical adenomas. However, larger series suggest that up to 25% of adrenal nodules are hormonally active, due to a cortisol- or aldosterone-producing adrenocortical adenoma or a pheochromocytoma associated with catecholamine excess **(Table 379-5)**. ACC is rare but is the cause of an adrenal mass in 5% of patients. However, metastases originating from another solid tissue tumor are an additional cause of adrenal incidentaloma, and have a higher incidence in patients undergoing imaging for tumor staging or follow-up monitoring (Table 379-5).

TABLE 379-5

Classification of Unilateral Adrenal Masses

Mass	Approximate Prevalence (%)
Benign	
Adrenocortical adenoma	
Endocrine-inactive	60–85
Cortisol-producing	5–10
Aldosterone-producing	2–5
Pheochromocytoma	5–10
Adrenal myelolipoma	<1
Adrenal ganglioneuroma	<0.1
Adrenal hemangioma	<0.1
Adrenal cyst	<1
Adrenal hematoma/hemorrhagic infarction	<1
Indeterminate	
Adrenocortical oncocytoma	<1
Malignant	
Adrenocortical carcinoma	2–5
Malignant pheochromocytoma	<1
Adrenal neuroblastoma	<0.1
Lymphomas (including primary adrenal lymphoma)	<1
Metastases (most frequent: breast, lung)	1–2

Note: Bilateral adrenal enlargement/masses may be caused by congenital adrenal hyperplasia, bilateral macronodular hyperplasia, bilateral hemorrhage (due to antiphospholipid syndrome or sepsis-associated Waterhouse-Friderichsen syndrome), granuloma, amyloidosis, or infiltrative disease including tuberculosis.

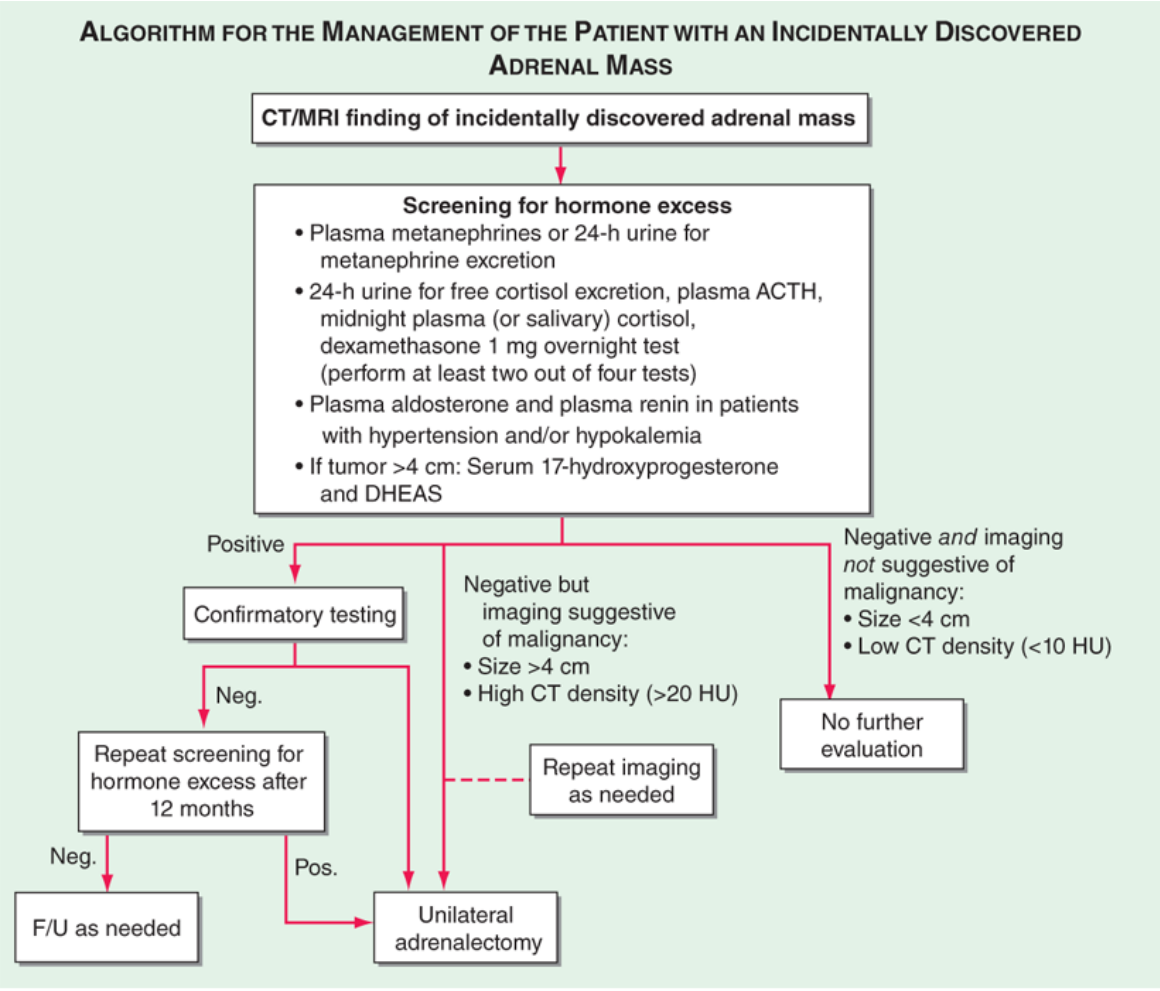
Differential Diagnosis and Treatment

Patients with an adrenal mass >1 cm require a diagnostic evaluation. Two key questions need to be addressed: (1) Does the tumor autonomously secrete hormones that could have a detrimental effect on health? (2) Is the adrenal mass benign or malignant?

Hormone secretion by an adrenal mass occurs along a continuum, with a gradual increase in clinical manifestations in parallel with hormone levels. Exclusion of catecholamine excess from a pheochromocytoma arising from the adrenal medulla is a mandatory part of the diagnostic workup **(Fig. 379-13)**. Furthermore, autonomous cortisol resulting in Cushing's syndrome requires exclusion and, in patients with hypertension or low serum potassium, also primary aldosteronism. Adrenal incidentalomas can be associated with MACE, and patients usually lack overt clinical features of Cushing's syndrome. Nonetheless, they may exhibit one or more components of the metabolic syndrome (e.g., obesity, type 2 diabetes, or hypertension). There is ongoing debate about the optimal treatment for these patients. Overproduction of adrenal androgen precursors, DHEA and its sulfate, is rare and most frequently seen in the context of ACC, as are increased levels of steroid precursors such as 17OHP.

FIGURE 379-13

Management of the patient with an incidentally discovered adrenal mass. CT, computed tomography; F/U, follow-up; MRI, magnetic resonance imaging.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

For the differentiation of benign from malignant adrenal masses, imaging is relatively sensitive, although specificity is suboptimal. Unenhanced CT is the procedure of choice for imaging the adrenal glands (Fig. 379-11). A diagnosis of ACC, pheochromocytoma, and benign adrenal myelolipoma becomes more likely with increasing diameter of the adrenal mass. However, size alone is of poor predictive value, with only 80% sensitivity and 60% specificity for the differentiation of benign from malignant masses when using a 4-cm cut-off. Metastases are rare but are found with similar frequency in adrenal masses of all sizes. Tumor density on unenhanced CT is of additional diagnostic value, as many adrenocortical adenomas are lipid rich and thus present with low attenuation values (i.e., densities of <10 Hounsfield Units [HUs]). However, similar numbers of adrenocortical adenomas are lipid poor and present with higher HU, making it difficult to differentiate them from ACCs, as well as also pheochromocytomas, both of which invariably have high attenuation values (i.e., densities >20 HU on precontrast scans). Generally, benign lesions are rounded and homogenous, whereas most malignant lesions appear lobulated and inhomogeneous. Pheochromocytoma and adrenomyelolipoma may also exhibit lobulated and inhomogeneous features. MRI also allows for the visualization of the adrenal glands with somewhat lower resolution than CT. However, because it does not involve exposure to ionizing radiation, it is preferred in children, young adults, and during pregnancy. MRI has a valuable role in the characterization of indeterminate adrenal lesions using chemical shift analysis, with malignant tumors rarely showing loss of signal on opposed-phase MRI; however, this may also be observed in a proportion of benign adrenocortical adenomas.

Fine-needle aspiration (FNA) or CT-guided biopsy of an adrenal mass is very rarely indicated. FNA of a pheochromocytoma can cause a life-threatening hypertensive crisis. FNA of an ACC violates the tumor capsule and can cause needle track metastasis. FNA should only be considered in a patient with a history of nonadrenal malignancy and a newly detected adrenal mass, after careful exclusion of pheochromocytoma, and if the outcome will influence therapeutic management. It is important to recognize that in 25% of patients with a previous history of nonadrenal malignancy, a newly detected mass on CT is not a metastasis. While FNA can diagnose extra-adrenal malignancies, it has very limited ability to differentiate between benign and malignant adrenocortical lesions, and hence should not be used for diagnosis of ACC.

Adrenal masses associated with confirmed hormone excess or suspected malignancy are usually treated surgically (Fig. 379-13) or, if adrenalectomy is not feasible or desired, with medication. Preoperative exclusion of glucocorticoid excess is particularly important for the prediction of postoperative suppression of the contralateral adrenal gland, which requires glucocorticoid replacement peri- and postoperatively. If the initial decision is for

observation, imaging and biochemical testing should be repeated 6–12 months after the first assessment. However, this may be performed earlier in patients with borderline imaging or hormonal findings. Adrenal masses associated with normal biochemistry at diagnosis and a tumor radiodensity of <10 HU on unenhanced CT can be safely assumed to represent a benign adenoma and do not require further follow-up.

ADRENOCORTICAL CARCINOMA

ACC is a rare malignancy with an annual incidence of 1–2 per million population. ACC is generally considered a highly malignant tumor; however, it presents with broad interindividual variability with regard to biologic characteristics and clinical behavior. Somatic mutations in the tumor-suppressor gene *TP53* are found in 25% of apparently sporadic ACC. Germline *TP53* mutations are the cause of the Li-Fraumeni syndrome associated with multiple solid organ cancers including ACC and are found in 25% of pediatric ACC cases; the *TP53* mutation R337H is found in almost all pediatric ACC in Brazil. Other genetic changes identified in ACC include alterations in the Wnt/β-catenin pathway and in the insulin-like growth factor 2 (IGF2) cluster; IGF2 overexpression is found in 90% of ACC.

Patients with large adrenal tumors suspicious of malignancy should be managed by a multidisciplinary specialist team, including an endocrinologist, an oncologist, a surgeon, a radiologist, and a histopathologist. FNA is not indicated in suspected ACC: first, cytology and also histopathology of a core biopsy cannot differentiate between benign and malignant primary adrenal masses; second, FNA violates the tumor capsule and may even cause needle canal metastasis. Even when the entire tumor specimen is available, the histopathologic differentiation between benign and malignant lesions is a diagnostic challenge. The most common histopathologic classification is the Weiss score, taking into account high nuclear grade; mitotic rate (>5/HPF); atypical mitosis; <25% clear cells; diffuse architecture; and presence of necrosis, venous invasion, and invasion of sinusoidal structures and tumor capsule. The presence of three or more elements suggests ACC.

Although 60–70% of ACCs show biochemical evidence of steroid overproduction, in many patients, this is not clinically apparent due to the relatively inefficient steroid production by the adrenocortical cancer cells. Excess production of glucocorticoids and adrenal androgen precursors are most common. Mixed excess production of several corticosteroid classes by an adrenal tumor is generally indicative of malignancy.

Tumor staging at diagnosis **(Table 379-6)** has important prognostic implications and requires scanning of the chest and abdomen for local organ invasion, lymphadenopathy, and metastases. Intravenous contrast medium is necessary for maximum sensitivity for hepatic metastases. An adrenal origin may be difficult to determine on standard axial CT imaging if the tumors are large and invasive, but CT reconstructions and MRI are more informative **(Fig. 379-14)** using multiple planes and different sequences. Vascular and adjacent organ invasion is diagnostic of malignancy. 18-Fluoro-2-deoxy-D-glucose positron emission tomography (18-FDG PET) is highly sensitive for the detection of malignancy and can be used to detect small metastases or local recurrence that may not be obvious on CT (Fig. 379-14). However, FDG PET has limited specificity and therefore cannot be used for differentiating benign from malignant adrenal lesions. Metastasis in ACC most frequently occurs to liver and lung.

TABLE 379-6

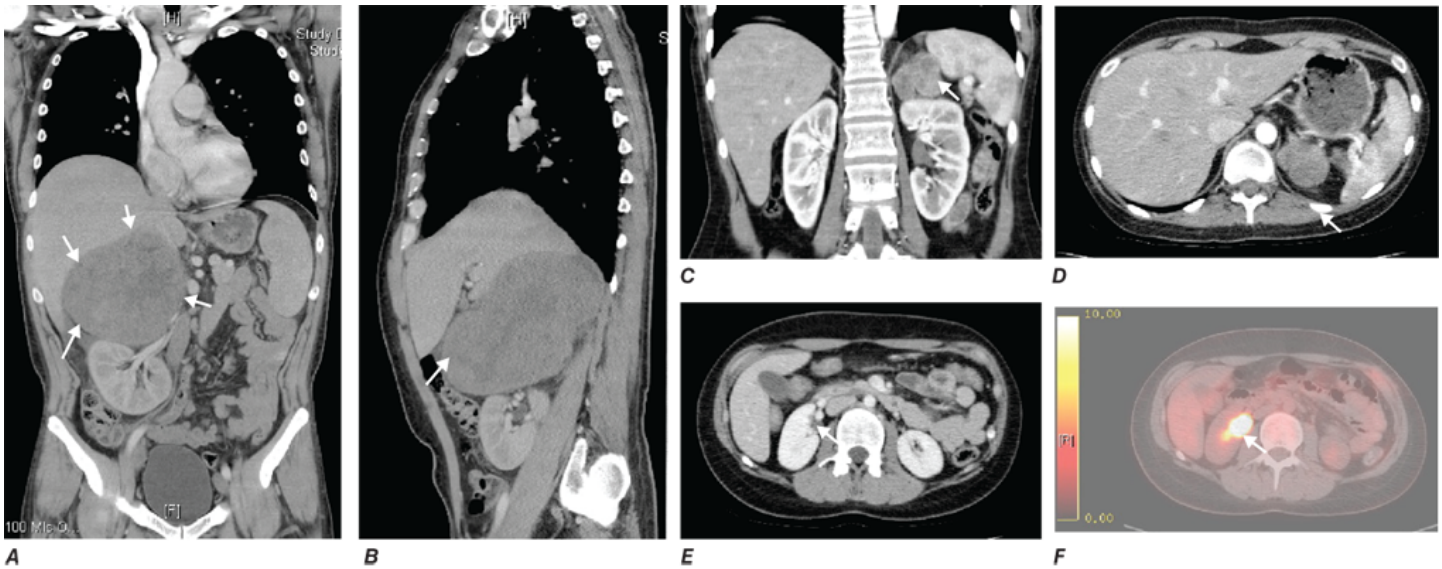
Classification System for Staging of Adrenocortical Carcinoma

ENSAT Stage	TNM Stage	TNM Definitions
I	**T1,N0,M0**	T1, tumor ≤5 cm
		N0, no positive lymph node
		M0, no distant metastases
II	**T2,N0,M0**	T2, tumor >5 cm
		N0, no positive lymph node
		M0, no distant metastases
III	**T1–T2,N1,M0**	N1, positive lymph node(s)
	T3–T4,N0–N1,M0	M0, no distant metastases
		T3, tumor infiltration into surrounding tissue
		T4, tumor invasion into adjacent organs *or* venous tumor thrombus in vena cava or renal vein
IV	**T1–T4,N0–N1,M1**	M1, presence of distant metastases

Abbreviations: ENSAT, European Network for the Study of Adrenal Tumors; TNM, tumor, node, metastasis.

FIGURE 379-14

Imaging in adrenocortical carcinoma (ACC). Magnetic resonance imaging scan with (***A***) frontal and (***B***) lateral views of a right ACC that was detected incidentally. Computed tomography (CT) scan with (***C***) coronal and (***D***) transverse views depicting a right-sided ACC. Note the irregular border and inhomogeneous structure. CT scan (***E***) and positron emission tomography/CT (***F***) visualizing a peritoneal metastasis of an ACC in close proximity to the right kidney (*arrow*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

There is no established grading system for ACC, and the Weiss score carries no prognostic value; the most important prognostic histopathologic parameter is the Ki67 proliferation index, with Ki67 <10% indicative of slow to moderate growth velocity, whereas a Ki67 ≥10% is associated with poor prognosis including high risk of recurrence and rapid progression.

Cure of ACC can only be achieved by early detection and complete surgical removal. Capsule violation during primary surgery, metastasis at diagnosis, and primary treatment in a nonspecialist center and by a nonspecialist surgeon are major determinants of poor survival. If the primary tumor invades adjacent organs, en bloc removal of kidney and spleen should be considered to reduce the risk of recurrence and regional lymph node dissection may further reduce this risk. Surgery can also be considered in a patient with metastases if there is severe tumor-related hormone excess. This indication needs to be carefully weighed against surgical risk, including thromboembolic complications, and the resulting delay in the introduction of other therapeutic options. Patients with confirmed ACC and successful removal of the primary tumor should receive adjuvant treatment with mitotane (o,p'DDD), particularly in patients with a high risk of recurrence as determined by tumor size >8 cm, histopathologic signs of vascular invasion, capsule invasion or violation, and a Ki67 proliferation index ≥10%. Adjuvant mitotane should be continued for at least 2 years, if the patient can tolerate side effects. Regular monitoring of plasma mitotane levels is mandatory (therapeutic range 14–20 mg/L; neurotoxic complications more frequent at >20 mg/L). Mitotane is usually started at 500 mg tid, with stepwise increases to a maximum dose of 2000 mg tid in days (high-dose saturation) or weeks (low-dose saturation) as tolerated. Once therapeutic range plasma mitotane levels are achieved, the dose can be tapered to maintenance doses mostly ranging from 1000 to 1500 mg tid. Mitotane treatment results in disruption of cortisol synthesis and thus requires glucocorticoid replacement; glucocorticoid replacement dose should be at least double of that usually used in adrenal insufficiency (i.e., 20 mg tid) because mitotane induces hepatic CYP3A4 activity resulting in rapid inactivation of glucocorticoids. Mitotane also increases circulating CBG, thereby decreasing the available free cortisol fraction. Single metastases can be addressed surgically or with radiofrequency ablation as appropriate. If the tumor recurs or progresses during mitotane treatment, cytotoxic chemotherapy should be considered; the established first-line chemotherapy regimen is the combination of cisplatin, etoposide, and doxorubicin plus continuing mitotane. Painful bone metastasis responds to irradiation. Overall survival in ACC is still poor, with 5-year survival rates of 30–40% and a median survival of 15 months in metastatic ACC.

ADRENAL INSUFFICIENCY

Epidemiology

The prevalence of well-documented, permanent adrenal insufficiency is 5 in 10,000 in the general population. Hypothalamic-pituitary origin of disease is most frequent, with a prevalence of 3 in 10,000, whereas primary adrenal insufficiency has a prevalence of 2 in 10,000. Approximately one-half of the latter cases are acquired, mostly caused by autoimmune destruction of the adrenal glands; the other one-half are genetic, most commonly caused by distinct enzymatic blocks in adrenal steroidogenesis affecting glucocorticoid synthesis (i.e., CAH.)

Adrenal insufficiency arising from suppression of the HPA axis as a consequence of exogenous glucocorticoid treatment is much more common, occurring in 0.5–2% of the population in developed countries.

Etiology

Primary adrenal insufficiency is most commonly caused by autoimmune adrenalitis. Isolated autoimmune adrenalitis accounts for 30–40%, whereas 60–70% develop adrenal insufficiency as part of autoimmune polyglandular syndromes (APSs) **(Chap. 381) (Table 379-7)**. APS1, also termed APECED (autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy), is the underlying cause in 10% of patients affected by APS. APS1 is transmitted in an autosomal recessive manner and is caused by mutations in the autoimmune regulator gene *AIRE*. Associated autoimmune conditions overlap with those seen in APS2, but may also include total alopecia, primary hypoparathyroidism, and, in rare cases, lymphoma. APS1 patients invariably develop chronic mucocutaneous candidiasis, usually manifest in childhood, and preceding adrenal insufficiency by years or decades. The much more prevalent APS2 is of polygenic inheritance, with confirmed associations with the *HLA-DR3* gene region in the major histocompatibility complex and distinct gene regions involved in immune regulation (*CTLA-4, PTPN22, CLEC16A*). Coincident autoimmune disease most frequently includes thyroid autoimmune disease, vitiligo, and premature ovarian failure. Less commonly, additional features may include type 1 diabetes and pernicious anemia caused by vitamin B_{12} deficiency.

TABLE 379-7

Causes of Primary Adrenal Insufficiency

Diagnosis	Gene	Associated Features
Autoimmune polyglandular syndrome 1 (APS1)	*AIRE*	Hypoparathyroidism, chronic mucocutaneous candidiasis, other autoimmune disorders, rarely lymphomas
Autoimmune polyglandular syndrome 2 (APS2)	Associations with HLA-DR3, CTLA-4	Hypothyroidism, hyperthyroidism, premature ovarian failure, vitiligo, type 1 diabetes mellitus, pernicious anemia
Isolated autoimmune adrenalitis	Associations with HLA-DR3, CTLA-4	
Congenital adrenal hyperplasia (CAH)	*CYP21A2, CYP11B1, CYP17A1, HSD3B2, POR*	See Table 379-10 (see also **Chap. 383**)
Congenital lipoid adrenal hyperplasia (CLAH)	*STAR, CYP11A1*	46,XY DSD, gonadal failure (see also **Chap. 383**)
Adrenal hypoplasia congenita (AHC)	*NR0B1 (DAX-1), NR5A1 (SF-1)*	46,XY DSD, gonadal failure (see also **Chap. 383**)
Adrenoleukodystrophy (ALD), adrenomyeloneuropathy (AMN)	*ABCD1*	Demyelination of central nervous system (ALD) or spinal cord and peripheral nerves (AMN)
Familial glucocorticoid deficiency	*MC2R*	Tall stature
	MRAP	None
	STAR	None
	NNT	None
	TXNRD2	None
	MCM4	Growth retardation, natural killer cell deficiency
Triple A syndrome	*AAAS*	Alacrima, achalasia, neurologic impairment
Smith-Lemli-Opitz syndrome	*SLOS*	Cholesterol synthesis disorder associated with mental retardation, craniofacial malformations, growth failure
Kearns-Sayre syndrome	Mitochondrial DNA deletions	Progressive external ophthalmoplegia, pigmentary retinal degeneration, cardiac conduction defects, gonadal failure, hypoparathyroidism, type 1 diabetes,
IMAGe syndrome	*CDKN1C*	Intrauterine growth retardation, metaphyseal dysplasia, genital anomalies
MIRAGE syndrome	*SAMD9*	Myelodysplasia, infection, restriction of growth, genital phenotypes, and enteropathy
Sphingosine-1-phosphate lyase deficiency	*SGPL1*	Steroid-resistant nephrotic syndrome, immunodeficiency, neurological defects, ichthyosis, primary hypothyroidism, cryptorchidism
Adrenal infections		Tuberculosis, HIV, CMV, cryptococcosis, histoplasmosis, coccidioidomycosis
Adrenal infiltration		Metastases, lymphomas, sarcoidosis, amyloidosis, hemochromatosis
Adrenal hemorrhage		Meningococcal sepsis (Waterhouse-Friderichsen syndrome), primary antiphospholipid syndrome

Diagnosis	Gene	Associated Features
Drug-induced		Mitotane, aminoglutethimide, abiraterone, trilostane, etomidate, ketoconazole, suramin, RU486
Bilateral adrenalectomy		E.g., in the management of Cushing's or after bilateral nephrectomy

Abbreviations: AIRE, autoimmune regulator; CMV, cytomegalovirus; DSD, disordered sex development; MC2R, ACTH receptor; MCM4, mini chromosome maintenance-deficient 4 homologue; MRAP, MC2R-accessory protein; NNT, nicotinamide nucleotide transhydrogenase.

X-linked adrenoleukodystrophy has an incidence of 1:20,000 males and is caused by mutations in the *X-ALD* gene encoding the peroxisomal membrane transporter protein ABCD1; its disruption results in accumulation of very long chain (>24 carbon atoms) fatty acids. Approximately 50% of cases manifest in early childhood with rapidly progressive white matter disease (cerebral ALD); 35% present during adolescence or in early adulthood with neurologic features indicative of myelin and peripheral nervous system involvement (adrenomyeloneuropathy [AMN]). In the remaining 15%, adrenal insufficiency is the sole manifestation of disease. Of note, distinct mutations manifest with variable penetrance and phenotypes within affected families.

Rarer causes of adrenal insufficiency involve destruction of the adrenal glands as a consequence of infection, hemorrhage, or infiltration (Table 379-7); tuberculous adrenalitis is still a frequent cause of disease in developing countries. Adrenal metastases rarely cause adrenal insufficiency, and this occurs only with bilateral, bulky metastases.

Inborn causes of primary adrenal insufficiency other than CAH are rare, causing <1% of cases. However, their elucidation provides important insights into adrenal gland development and physiology. Mutations causing primary adrenal insufficiency (Table 379-7) include factors regulating adrenal development and steroidogenesis (DAX-1, SF-1), cholesterol synthesis, import and cleavage (DHCR7, StAR, CYP11A1), elements of the adrenal ACTH response pathway (MC2R, MRAP) (Fig. 379-5), and factors involved in redox regulation (NNT, TXNRD2) and DNA repair (MCM4, CDKN1C).

Secondary adrenal insufficiency is the consequence of dysfunction of the hypothalamic-pituitary component of the HPA axis **(Table 379-8)**. Excluding iatrogenic suppression, the overwhelming majority of cases are caused by pituitary or hypothalamic tumors or their treatment by surgery or irradiation **(Chap. 373)**. Rarer causes include pituitary apoplexy, either as a consequence of an infarcted pituitary adenoma or transient reduction in the blood supply of the pituitary during surgery or after rapid blood loss associated with parturition, also termed Sheehan's syndrome. Isolated ACTH deficiency is rarely caused by autoimmune disease or pituitary infiltration (Table 379-8). Mutations in the ACTH precursor POMC or in factors regulating pituitary development are genetic causes of ACTH deficiency (Table 379-8).

TABLE 379-8

Causes of Secondary Adrenal Insufficiency

Diagnosis	Gene	Associated Features
Pituitary tumors (endocrine active and inactive adenomas, very rare: carcinoma)		Depending on tumor size and location: visual field impairment (bilateral hemianopia), hyperprolactinemia, secondary hypothyroidism, hypogonadism, growth hormone deficiency
Other mass lesions affecting the hypothalamic-pituitary region		Craniopharyngioma, meningioma, ependymoma, metastases
Pituitary irradiation		Radiotherapy administered for pituitary tumors, brain tumors, or craniospinal irradiation in leukemia
Autoimmune hypophysitis		Often associated with pregnancy; may present with panhypopituitarism or isolated ACTH deficiency; can be associated with autoimmune thyroid disease, more rarely with vitiligo, premature ovarian failure, type 1 diabetes, pernicious anemia
Pituitary apoplexy/hemorrhage		Hemorrhagic infarction of large pituitary adenomas or pituitary infarction consequent to traumatic major blood loss (e.g., surgery or pregnancy: Sheehan's syndrome)
Pituitary infiltration		Tuberculosis, actinomycosis, sarcoidosis, histiocytosis X, granulomatosis with polyangiitis (Wegener's), metastases
Drug-induced		Chronic glucocorticoid excess (endogenous or exogenous)
Congenital isolated ACTH deficiency	*TBX19* (Tpit)	
Combined pituitary hormone deficiency (CPHD)	*PROP-1*	Progressive development of CPHD in the order GH, PRL, TSH, LH/FSH, ACTH
	HESX1	CPHD and septo-optic dysplasia
	LHX3	CPHD and limited neck rotation, sensorineural deafness
	LHX4	CPHD and cerebellar abnormalities
	SOX3	CPHD and variable mental retardation
Proopiomelanocortin (POMC) deficiency	*POMC*	Early-onset obesity, red hair pigmentation

Abbreviations: ACTH, adrenocorticotropic hormone; GH, growth hormone; LH/FSH, luteinizing hormone/follicle-stimulating hormone; PRL, prolactin; TSH, thyroid-stimulating hormone.

Clinical Manifestations

In principle, the clinical features of primary adrenal insufficiency (Addison's disease) are characterized by the loss of both glucocorticoid and mineralocorticoid secretion **(Table 379-9)**. In secondary adrenal insufficiency, only glucocorticoid deficiency is present, as the adrenal itself is intact and thus still amenable to regulation by the RAA system. Adrenal androgen secretion is disrupted in both primary and secondary adrenal insufficiency (Table 379-9). Hypothalamic-pituitary disease can lead to additional clinical manifestations due to involvement of other endocrine axes (thyroid, gonads, growth hormone, prolactin) or visual impairment with bitemporal hemianopia caused by chiasmal compression. It is important to recognize that iatrogenic adrenal insufficiency caused by exogenous glucocorticoid suppression of the HPA axis may result in all symptoms associated with glucocorticoid deficiency (Table 379-9), if exogenous glucocorticoids are stopped abruptly. However, patients will appear clinically cushingoid as a result of the preceding overexposure to glucocorticoids.

TABLE 379-9

Signs and Symptoms of Adrenal Insufficiency

Signs and Symptoms Caused by Glucocorticoid Deficiency
Fatigue, lack of energy
Weight loss, anorexia
Myalgia, joint pain
Fever
Normochromic anemia, lymphocytosis, eosinophilia
Slightly increased TSH (due to loss of feedback inhibition of TSH release)
Hypoglycemia (more frequent in children)
Low blood pressure, postural hypotension
Hyponatremia (due to loss of feedback inhibition of AVP release)
Signs and Symptoms Caused by Mineralocorticoid Deficiency (Primary Adrenal Insufficiency Only)
Abdominal pain, nausea, vomiting
Dizziness, postural hypotension
Salt craving
Low blood pressure, postural hypotension
Increased serum creatinine (due to volume depletion)
Hyponatremia
Hyperkalemia
Signs and Symptoms Caused by Adrenal Androgen Deficiency
Lack of energy
Dry and itchy skin (in women)
Loss of libido (in women)
Loss of axillary and pubic hair (in women)
Other Signs and Symptoms
Hyperpigmentation (primary adrenal insufficiency only) (due to excess of proopiomelanocortin [POMC]-derived peptides)
Alabaster-colored pale skin (secondary adrenal insufficiency only) (due to deficiency of POMC-derived peptides)

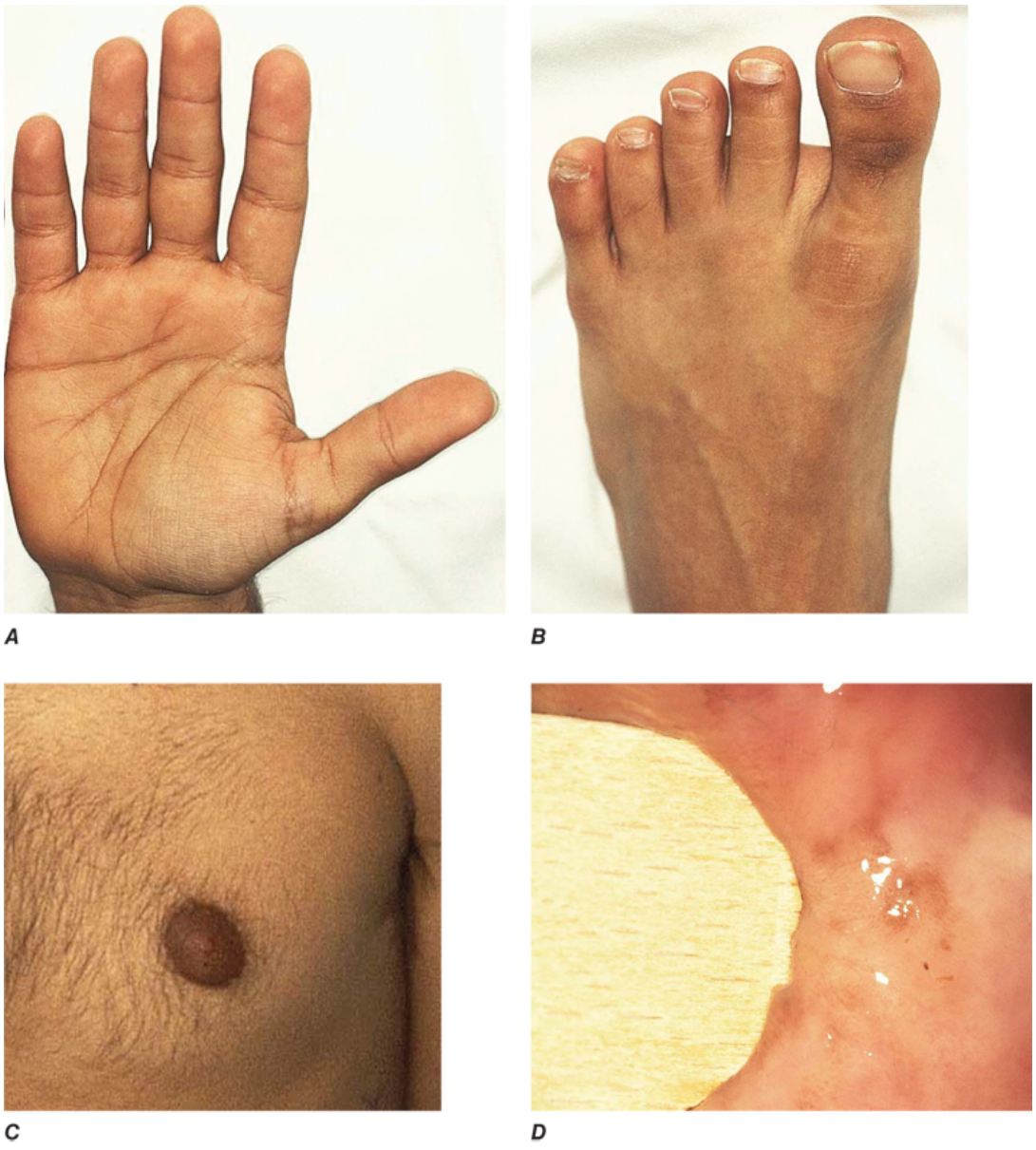
Abbreviations: AVP, arginine vasopressin; TSH, thyroid-stimulating hormone.

Chronic adrenal insufficiency manifests with relatively nonspecific signs and symptoms, such as fatigue and loss of energy, often resulting in delayed or missed diagnoses (e.g., as depression or anorexia). A distinguishing feature of primary adrenal insufficiency is hyperpigmentation, which is caused by

excess ACTH stimulation of melanocytes. Hyperpigmentation is most pronounced in skin areas exposed to increased friction or shear stress and is increased by sunlight (Fig. 379-15). Conversely, in secondary adrenal insufficiency, the skin has an alabaster-like paleness due to lack of ACTH secretion.

FIGURE 379-15

Clinical features of Addison's disease. Note the hyperpigmentation in areas of increased friction including (*A*) palmar creases, (*B*) dorsal foot, (*C*) nipples and axillary region, and (*D*) patchy hyperpigmentation of the oral mucosa.



A *B* *C* *D*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Hyponatremia is a characteristic biochemical feature in primary adrenal insufficiency and is found in 80% of patients at presentation. Hyperkalemia is present in 40% of patients at initial diagnosis. Hyponatremia is primarily caused by mineralocorticoid deficiency but can also occur in secondary adrenal insufficiency due to diminished inhibition of antidiuretic hormone (ADH) release by cortisol, resulting in mild syndrome of inappropriate secretion of antidiuretic hormone (SIADH). Glucocorticoid deficiency also results in slightly increased TSH concentrations that normalize within days to weeks after initiation of glucocorticoid replacement.

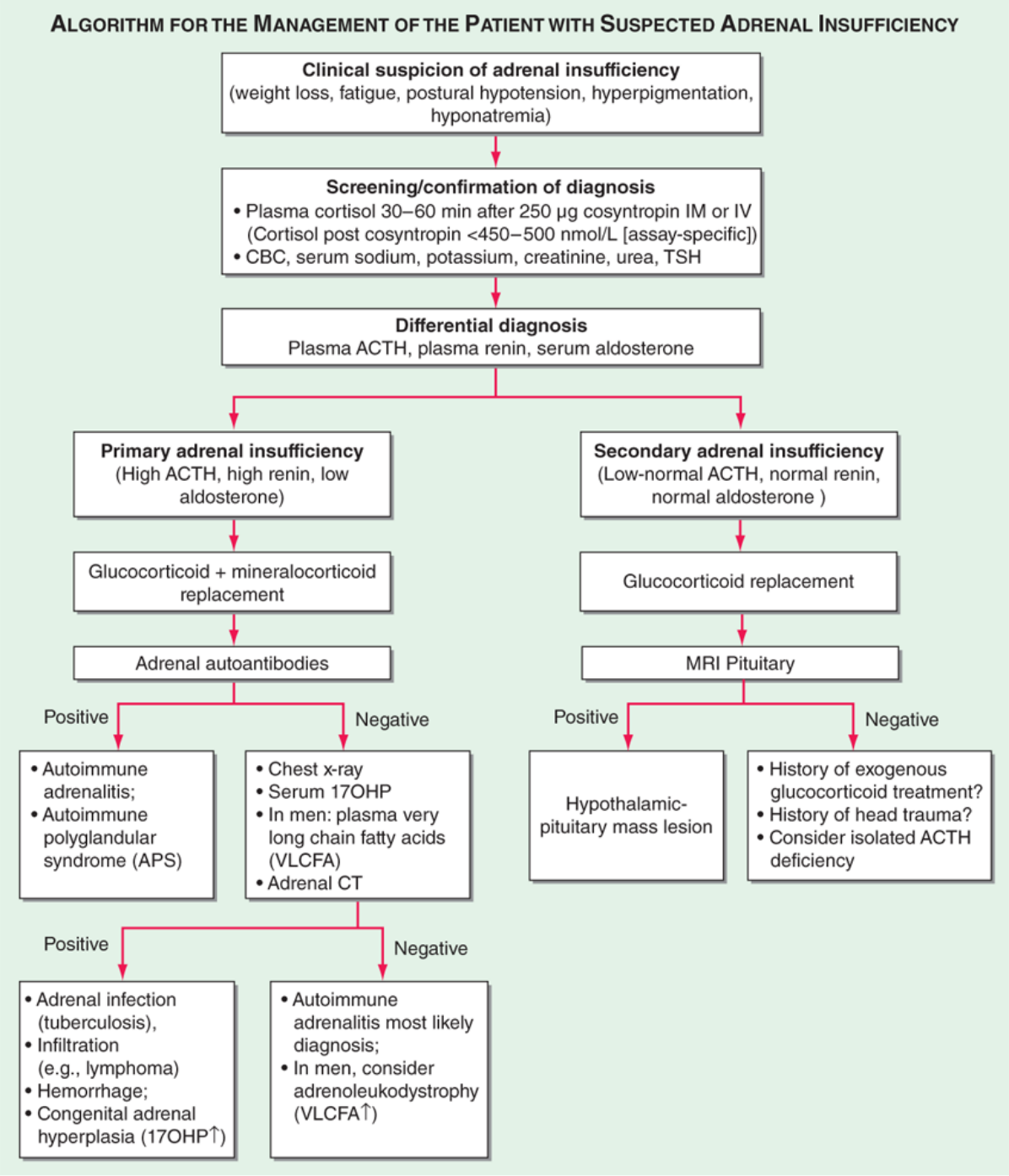
Acute adrenal insufficiency, also termed adrenal crisis, usually occurs after a prolonged period of nonspecific complaints and is more frequently observed in patients with primary adrenal insufficiency, due to the loss of both glucocorticoid and mineralocorticoid secretion. Postural hypotension may progress to hypovolemic shock. Adrenal insufficiency may mimic features of acute abdomen with abdominal tenderness, nausea, vomiting, and fever. In some cases, the primary presentation may resemble neurologic disease, with decreased responsiveness, progressing to stupor and coma. An adrenal crisis can be triggered by an intercurrent illness, surgical or other stress, or increased glucocorticoid inactivation (e.g., hyperthyroidism). Prospective data indicate 8.3 adrenal crises and 0.5 adrenal crisis-related deaths per 100 patient years.

Diagnosis

The diagnosis of adrenal insufficiency is established by the short cosyntropin test, a safe and reliable tool with excellent predictive diagnostic value **(Fig. 379-16)**. The cut-off for failure is usually defined at cortisol levels of <450–500 nmol/L (16–18 μg/dL) sampled 30–60 min after ACTH stimulation; the exact cut-off is dependent on the locally available assay, with generally lower cut-offs for mass spectrometry-based assays. During the early phase of HPA disruption (e.g., within 4 weeks of pituitary insufficiency), patients may still respond to exogenous ACTH stimulation. In this circumstance, the ITT is an alternative choice but is more invasive and should be carried out only under a specialist's supervision (see above). Induction of hypoglycemia is contraindicated in individuals with diabetes mellitus, cardiovascular disease, or history of seizures. Random serum cortisol measurements are of limited diagnostic value, because baseline cortisol levels may be coincidentally low due to the physiologic diurnal rhythm of cortisol secretion (Fig. 379-3). Similarly, many patients with secondary adrenal insufficiency have relatively normal baseline cortisol levels but fail to mount an appropriate cortisol response to ACTH, which can only be revealed by stimulation testing. Importantly, tests to establish the diagnosis of adrenal insufficiency should never delay treatment. Thus, in a patient with suspected adrenal crisis, it is reasonable to draw baseline cortisol levels, provide replacement therapy, and defer formal stimulation testing until a later time.

FIGURE 379-16

Management of the patient with suspected adrenal insufficiency. ACTH, adrenocorticotropic hormone; CBC, complete blood count; MRI, magnetic resonance imaging; PRA, plasma renin activity; TSH, thyroid-stimulating hormone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Once adrenal insufficiency is confirmed, measurement of plasma ACTH is the next step, with increased or inappropriately low levels defining primary and secondary origin of disease, respectively (Fig. 379-16). In primary adrenal insufficiency, increased plasma renin will confirm the presence of mineralocorticoid deficiency. At initial presentation, patients with primary adrenal insufficiency should undergo screening for steroid autoantibodies as a marker of autoimmune adrenalitis. If these tests are negative, adrenal imaging by CT is indicated to investigate possible hemorrhage, infiltration, or masses. In male patients with negative autoantibodies in the plasma, very long-chain fatty acids should be measured to exclude X-ALD. Patients with inappropriately low ACTH, in the presence of confirmed cortisol deficiency, should undergo hypothalamic-pituitary imaging by MRI. Features suggestive of preceding pituitary apoplexy, such as sudden-onset severe headache or history of previous head trauma, should be carefully explored, particularly in patients with no obvious MRI lesion.

TREATMENT

TREATMENT

Acute Adrenal Insufficiency

Acute adrenal insufficiency requires immediate initiation of rehydration, usually carried out by saline infusion at initial rates of 1 L/h with continuous cardiac monitoring. Glucocorticoid replacement should be initiated by bolus injection of 100 mg hydrocortisone, followed by the administration of 200 mg hydrocortisone over 24 h, preferably by continuous infusion or alternatively by bolus IV or IM injections. Mineralocorticoid replacement can be initiated once the daily hydrocortisone dose has been reduced to <50 mg because at higher doses hydrocortisone provides sufficient stimulation of MRs.

Glucocorticoid replacement for the treatment of chronic adrenal insufficiency should be administered at a dose that replaces the physiologic daily cortisol production, which is usually achieved by the oral administration of 15–25 mg hydrocortisone in two to three divided doses. Pregnancy may require an increase in hydrocortisone dose by 50% during the last trimester. In all patients, at least one-half of the daily dose should be administered in the morning. Currently available glucocorticoid preparations fail to mimic the physiologic cortisol secretion rhythm (Fig. 379-3). Long-acting glucocorticoids such as prednisolone or dexamethasone are not preferred because they result in increased glucocorticoid exposure due to extended GR activation at times of physiologically low cortisol secretion. There are no well-established dose equivalencies, but as a guide, equipotency can be assumed for 1 mg hydrocortisone, 1.6 mg cortisone acetate, 0.2 mg prednisolone, 0.25 mg prednisone, and 0.025 mg dexamethasone.

Monitoring of glucocorticoid replacement is mainly based on the history and examination for signs and symptoms suggestive of glucocorticoid over- or underreplacement, including assessment of body weight and blood pressure. Plasma ACTH, 24-h urinary free cortisol, or serum cortisol day curves reflect whether hydrocortisone has been taken or not, but do not convey reliable information about replacement quality. In patients with isolated primary adrenal insufficiency, monitoring should include screening for autoimmune thyroid disease, and female patients should be made aware of the possibility of premature ovarian failure. Supraphysiologic glucocorticoid treatment with doses equivalent to 30 mg hydrocortisone or more will affect bone metabolism, and these patients should undergo regular bone mineral density evaluation. All patients with adrenal insufficiency need to be instructed about the requirement for stress-related glucocorticoid dose adjustments. These generally consist of doubling the routine oral glucocorticoid dose in the case of intercurrent illness with fever and bed rest and the need for IV hydrocortisone injection at a daily dose of 100 mg in cases of prolonged vomiting, surgery, or trauma. All patients, but in particular those living or traveling in regions with delayed access to acute health care, should carry a hydrocortisone self-injection emergency kit, in addition to their usual steroid emergency cards and bracelets, and should receive training in its use.

Mineralocorticoid replacement in primary adrenal insufficiency should be initiated at a dose of 100–150 μg fludrocortisone. The adequacy of treatment can be evaluated by measuring blood pressure, sitting and standing, to detect a postural drop indicative of hypovolemia. In addition, serum sodium, potassium, and plasma renin should be measured regularly. Renin levels should be kept in the upper normal reference range. Changes in glucocorticoid dose may also impact on mineralocorticoid replacement as cortisol also binds the MR; 40 mg hydrocortisone is equivalent to 100 μg fludrocortisone. In patients living or traveling in areas with hot or tropical weather conditions, the fludrocortisone dose should be increased by 50–100 μg during the summer. Mineralocorticoid dose may also need to be adjusted during pregnancy, due to the antimineralocorticoid activity of progesterone, but this is less often required than hydrocortisone dose adjustment. Plasma renin cannot serve as a monitoring tool during pregnancy, because renin rises physiologically during gestation.

Adrenal androgen replacement is an option in patients with lack of energy, despite optimized glucocorticoid and mineralocorticoid replacement. It may also be indicated in women with features of androgen deficiency, including loss of libido. Adrenal androgen replacement can be achieved by once-daily administration of 25–50 mg DHEA. Treatment is monitored by measurement of DHEAS, androstenedione, testosterone, and sex hormone–binding globulin (SHBG) 24 h after the last DHEA dose.

CONGENITAL ADRENAL HYPERPLASIA

(See also Chap. 383) CAH is caused by mutations in genes encoding steroidogenic enzymes involved in glucocorticoid synthesis (*CYP21A2, CYP17A1, HSD3B2, CYP11B1*) or in the cofactor enzyme P450 oxidoreductase that serves as an electron donor to CYP21A2 and CYP17A1 (Fig. 379-1). Invariably, patients affected by CAH exhibit glucocorticoid deficiency. Depending on the exact step of enzymatic block, they may also have excess production of mineralocorticoids or deficient production of sex steroids **(Table 379-10)**. The diagnosis of CAH is readily established by measurement of the steroids accumulating before the distinct enzymatic block, either in serum or in urine, preferably by the use of mass spectrometry–based assays (Table 379-10).

TABLE 379-10

Variants of Congenital Adrenal Hyperplasia

Variant	Gene	Impact on Steroid Synthesis	Diagnostic Marker Steroids in Serum (and Urine)
21-Hydroxylase deficiency (21OHD)	*CYP21A2*	Glucocorticoid deficiency, mineralocorticoid deficiency, adrenal androgen excess	17-Hydroxyprogesterone, 21-deoxycortisol (pregnanetriol, 17-hydroxypregnanolone, pregnanetriolone)
11β-Hydroxylase deficiency (11OHD)	*CYP11B1*	Glucocorticoid deficiency, mineralocorticoid excess, adrenal androgen excess	11-Deoxycortisol, 11-deoxycorticosterone (tetrahydro-11-deoxycortisol, tetrahydro-11-deoxycorticosterone)
17α-Hydroxylase deficiency (17OHD)	*CYP17A1*	(Glucocorticoid deficiency), mineralocorticoid excess, androgen deficiency	11-Deoxycorticosterone, corticosterone, pregnenolone, progesterone (tetrahydro-11-deoxycorticosterone, tetrahydrocorticosterone, pregnenediol, pregnanediol)
3β-Hydroxysteroid dehydrogenase deficiency (3bHSDD)	*HSD3B2*	Glucocorticoid deficiency, (mineralocorticoid deficiency), adrenal androgen excess (females and males), gonadal androgen deficiency (males)	17-Hydroxypregnanolone (pregnanetriol)
P450 oxidoreductase deficiency (PORD)	*POR*	Glucocorticoid deficiency, (mineralocorticoid excess), prenatal androgen excess and postnatal androgen deficiency, skeletal malformations	Pregnenolone, progesterone, 17-hydroxyprogesterone (pregnanediol, pregnanetriol)

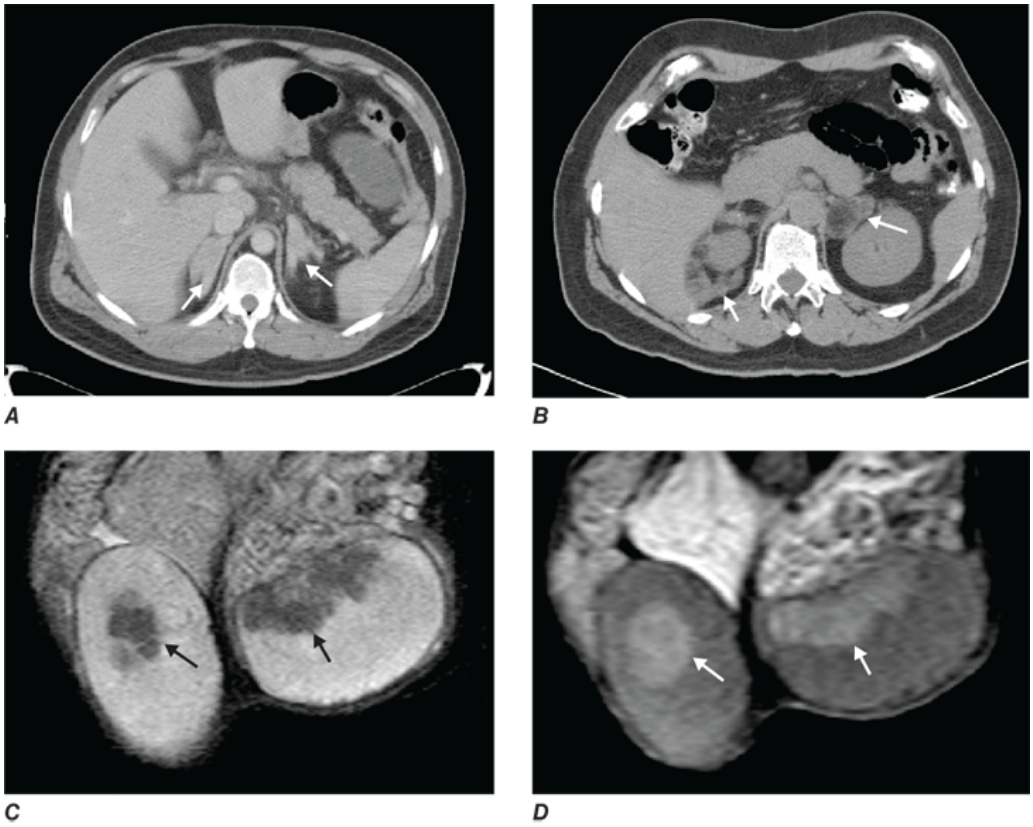
Mutations in *CYP21A2* are the most prevalent cause of CAH, responsible for 90–95% of cases. 21-Hydroxylase deficiency disrupts glucocorticoid and mineralocorticoid synthesis (Fig. 379-1), resulting in diminished negative feedback via the HPA axis. This leads to increased pituitary ACTH release, which drives increased synthesis of adrenal androgen precursors and subsequent androgen excess. The degree of impairment of glucocorticoid and mineralocorticoid secretion depends on the severity of mutations. Major loss-of-function mutations result in combined glucocorticoid and mineralocorticoid deficiency (classic CAH, neonatal presentation), whereas less severe mutations affect glucocorticoid synthesis only (simple virilizing CAH, neonatal or early childhood presentation). The mildest mutations result in the least severe clinical phenotype, nonclassic CAH, usually presenting during adolescence and early adulthood and with preserved glucocorticoid production.

Androgen excess is present in all patients and manifests with broad phenotypic variability, ranging from severe virilization of the external genitalia in neonatal girls (e.g., 46,XX disordered sex development [DSD]) to hirsutism and oligomenorrhea resembling a polycystic ovary syndrome phenotype in young women with nonclassic CAH. In countries without neonatal screening for CAH, boys with classic CAH usually present with life-threatening adrenal crisis in the first few weeks of life (salt-wasting crisis); a simple-virilizing genotype manifests with precocious pseudo-puberty and advanced bone age in early childhood, whereas men with nonclassic CAH are usually detected only through family screening.

Glucocorticoid treatment is more complex than for other causes of primary adrenal insufficiency as it not only needed to replace missing glucocorticoids but also to control the increased ACTH drive and subsequent androgen excess. Current treatment is hampered by the lack of glucocorticoid preparations that mimic the diurnal cortisol secretion profile, resulting in a prolonged period of ACTH stimulation and subsequent androgen production during the early morning hours. In childhood, optimization of growth and pubertal development are important goals of glucocorticoid treatment, in addition to prevention of adrenal crisis and treatment of 46,XX DSD. In adults, the focus shifts to preserving fertility and preventing side effects of glucocorticoid overtreatment, namely, the metabolic syndrome and osteoporosis. Fertility can be compromised in women due to oligomenorrhea/amenorrhea with chronic anovulation as a consequence of androgen excess. Men may develop so-called testicular adrenal rest tumors **(Fig. 379-17)**. These consist of hyperplastic cells with shared adrenal and gonadal characteristics located in the rete testis and should not be confused with testicular tumors. Testicular adrenal rest tissue can compromise sperm production and induce testicular fibrosis that may be irreversible.

FIGURE 379-17

Imaging in congenital adrenal hyperplasia (CAH). Adrenal computed tomography scans showing homogenous bilateral hyperplasia in a young patient with classic CAH (***A***) and macronodular bilateral hyperplasia (***B***) in a middle-aged patient with classic CAH with longstanding poor disease control. Magnetic resonance imaging scan with T1-weighted (***C***) and T2-weighted (***D***) images showing bilateral testicular adrenal rest tumors (*arrows*) in a young patient with salt-wasting CAH. *(Courtesy of N. Reisch.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Congenital Adrenal Hyperplasia

Hydrocortisone is a good treatment option for the prevention of adrenal crisis, but longer acting prednisolone may be needed to control androgen excess. In children, hydrocortisone is given in divided doses at 1–1.5 times the normal cortisol production rate (about 10–13 mg/m^2 per day). In adults, if hydrocortisone does not suffice, intermediate-acting glucocorticoids (e.g., prednisone) may be given, using the lowest dose necessary to suppress excess androgen production. For achieving fertility, dexamethasone treatment may be required, but should be only given for the shortest possible time period to limit adverse metabolic side effects. Biochemical monitoring should include androstenedione and testosterone, aiming for the normal sex-specific reference range. 17OHP is a useful marker of overtreatment, indicated by 17OHP levels within the normal range of healthy controls. Glucocorticoid overtreatment may suppress the hypothalamic-pituitary-gonadal axis. Thus, treatment needs to be carefully titrated against clinical features of disease control. Stress dose glucocorticoids should be given at double or triple the daily dose for surgery, acute illness, or severe trauma. Poorly controlled CAH can result in adrenocortical hyperplasia, which gave the disease its name, and may present as macronodular hyperplasia subsequent to long-standing ACTH excess (Fig. 379-17). The nodular areas can develop autonomous adrenal androgen production and may be unresponsive to glucocorticoid treatment.

Mineralocorticoid requirements change during life and are higher in children, explained by relative mineralocorticoid resistance that diminishes with ongoing maturation of the kidney. Children with CAH usually receive mineralocorticoid and salt replacement. However, young adults with CAH should undergo reassessment of their mineralocorticoid reserve. Plasma renin should be regularly monitored and kept within the upper half of the normal reference range.

FURTHER READING

Arlt W et al: Steroid metabolome analysis reveals prevalent glucocorticoid excess in primary aldosteronism. JCI Insight 2017 April 20;2(8). pii: 93136. doi: 10.1172/jci.insight.93136. [Epub ahead of print]

Bancos I et al: Diagnosis and management of adrenal insufficiency. Lancet Diabetes Endocrinol 3:216, 2015. [PubMed: 25098712]

Berr CM et al: Time to recovery of adrenal function after curative surgery for Cushing's syndrome depends on etiology. J Clin Endocrinol Metab 100:1300, 2015. [PubMed: 25546155]

Bornstein SR et al: Diagnosis and treatment of primary adrenal insufficiency: An Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab 101:364, 2016. [PubMed: 26760044]

Dinnes J et al: Management of endocrine disease: Imaging for the diagnosis of malignancy in incidentally discovered adrenal masses: A systematic review and meta-analysis. Eur J Endocrinol 175:R51, 2016. [PubMed: 27257145]

El-Maouche D et al: Congenital adrenal hyperplasia. Lancet 2017 May 30. pii: S0140-6736(17)31431-9. doi: 10.1016/S0140-6736(17)31431-9. [Epub ahead of print]

Fassnacht M et al: Management of adrenal incidentalomas: European Society of Endocrinology Clinical Practice Guideline in collaboration with the European Network for the Study of Adrenal Tumors. Eur J Endocrinol 175:G1, 2016. [PubMed: 27390021]

Lodish M, Stratakis CA: A genetic and molecular update on adrenocortical causes of Cushing syndrome. Nat Rev Endocrinol 12:255, 2016. [PubMed: 26965378]

Loriaux DL: Diagnosis and differential diagnosis of Cushing's syndrome. N Engl J Med 376:1451, 2017. [PubMed: 28402781]

Mallappa A et al: A phase 2 study of Chronocort, a modified-release formulation of hydrocortisone, in the treatment of adults with classic congenital adrenal hyperplasia. J Clin Endocrinol Metab 100:1137, 2015. [PubMed: 25494662]

Nieman LK et al: Treatment of Cushing's syndrome: An Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab 100:2807, 2015. [PubMed: 26222757]

Turcu AF et al: Single-dose study of a corticotropin-releasing factor receptor-1 antagonist in women with 21-hydroxylase deficiency. J Clin Endocrinol Metab 101:1174, 2016. [PubMed: 26751191]

Williams TA et al: Outcomes after adrenalectomy for unilateral primary aldosteronism: An international consensus on outcome measures and analysis of remission rates in an international cohort. Lancet Diabetes Endocrinol 5:689, 2017. [PubMed: 28576687]

Young WF Jr: The management of primary aldosteronism: Case detection, diagnosis, and treatment: An endocrine society clinical practice guideline. J Clin Endocrinol Metab 101:1889, 2016. [PubMed: 26934393]

Chapter 380: Pheochromocytoma

Hartmut P. H. Neumann

INTRODUCTION

Pheochromocytomas and paragangliomas are catecholamine-producing tumors derived from the sympathetic or parasympathetic nervous system. These tumors may arise sporadically or be inherited as features of multiple endocrine neoplasia type 2 (MEN 2), von Hippel–Lindau (VHL) disease, or several other pheochromocytoma-associated syndromes. The diagnosis of pheochromocytomas identifies a potentially correctable cause of hypertension, and their removal can prevent hypertensive crises that can be lethal. The clinical presentation is variable, ranging from an adrenal incidentaloma to a hypertensive crisis with associated cerebrovascular or cardiac complications.

EPIDEMIOLOGY

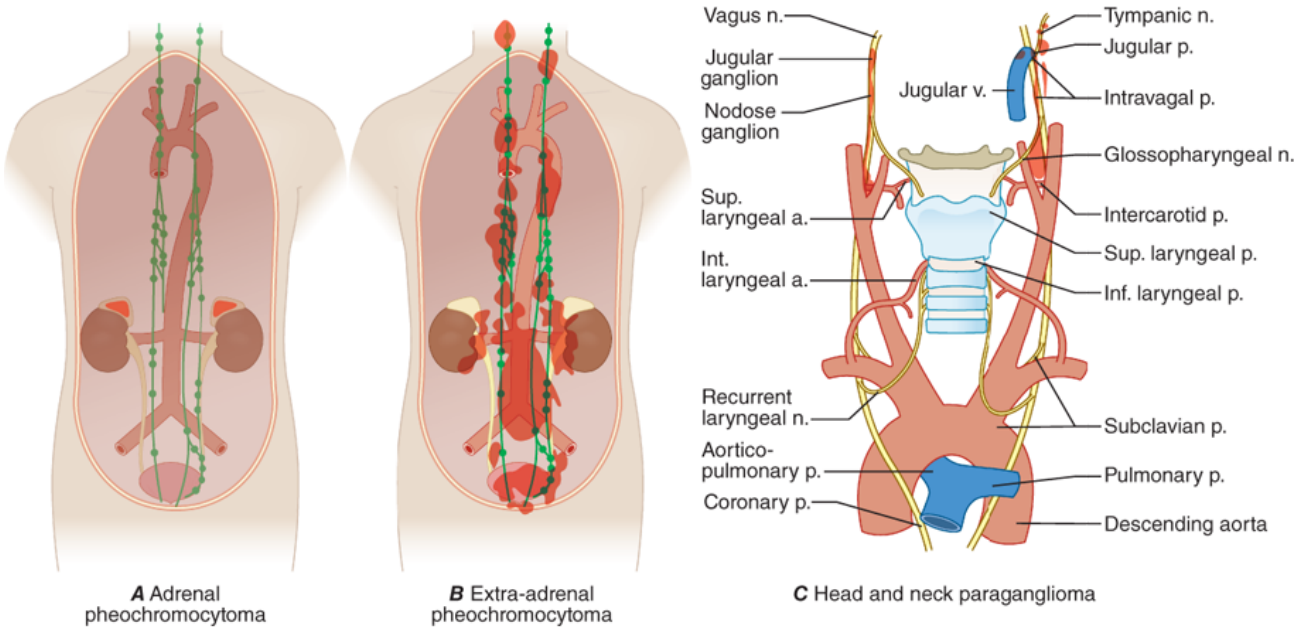
Pheochromocytoma was first described in 1800 by Charles Sugrue from Cork, Ireland, and the histological findings were first reported by Felix Fraenkel and Max Schottelius from Freiburg, Germany, in 1886. Pheochromocytoma is estimated to occur in 2–8 of 1 million persons per year, and ~0.1% of hypertensive patients harbor a pheochromocytoma. The mean age at diagnosis is ~40 years, although the tumors can occur from early childhood until late in life. The classic “rule of tens” for pheochromocytomas states that ~10% are bilateral, 10% are extra-adrenal, and 10% are malignant.

ETIOLOGY AND PATHOGENESIS

Pheochromocytomas and paragangliomas are well-vascularized tumors that arise from cells derived from the sympathetic (e.g., adrenal medulla or sympathetic trunk) or parasympathetic (e.g., carotid body, glomus tympanicum, glomus jugulare, glomus vagale) paraganglia **(Fig. 380-1)**. The name *pheochromocytoma* reflects the formerly used black-colored staining caused by chromaffin oxidation of catecholamines; although a variety of terms have been used to describe these tumors, most clinicians use this designation to describe symptomatic catecholamine-producing tumors, including those in extra-adrenal retroperitoneal, pelvic, and thoracic sites. The term *paraganglioma* is used to describe catecholamine-producing tumors in the skull base and neck; these tumors may secrete little or no catecholamine. In contrast to common clinical parlance, the World Health Organization (WHO) restricts the term *pheochromocytoma* to adrenal tumors and applies the term *paraganglioma* to tumors at all other sites.

FIGURE 380-1

The paraganglial system and topographic sites (in red) of pheochromocytomas and paragangliomas. *(Parts A and B from WM Manger, RW Gifford: Clinical and experimental pheochromocytoma. Cambridge, Blackwell Science, 1996; Part C from GG Glenner, PM Grimley: Tumors of the Extra-adrenal Paraganglion System [Including Chemoreceptors], Atlas of Tumor Pathology, 2nd Series, Fascicle 9. Washington, DC, AFIP, 1974.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The etiology of sporadic pheochromocytomas and paragangliomas is unknown. However, 25–33% of patients have an inherited condition, including germline mutations in the classically recognized *RET* (*re*arranged during *t*ransfection), *VHL, NF1* (neurofibromatosis type 1), *SDHB, SDHC,* and *SDHD* (subunits of SDH) genes or in the more recently recognized *SDHA, SDHAF2*, *TMEM127* (transmembrane protein 127), *MAX* (myc-associated factor X), *FH* (fumarate hydratase), *PDH1, PDH2* (pyruvate dehydrogenase), *HIF1alpha* (hypoxia-inducible factor), *MDH2* (malate dehydrogenase), and *KIF1Bß* (kinesin family member) genes. Biallelic gene inactivation, a characteristic of tumor suppressor genes has been demonstrated for the *VHL, NF1*, *SDHx, TMEM127, MAX, FH, PDH1, PDH2, MDH2*, and *KIF1Bß* genes. In contrast, *RET* is a protooncogene, and mutations activate receptor tyrosine kinase activity. Succinate dehydrogenase (SDH) is an enzyme of the Krebs cycle and the mitochondrial respiratory chain. The VHL protein is a component of an ubiquitin E3 ligase. *VHL* mutations reduce protein degradation, resulting in upregulation of components involved in cell cycle progression, glucose metabolism, and oxygen sensing. Currently, pheochromocytoma and paraganglioma susceptibility genes are attributed to two clusters. Cluster 1 mutations are associated with pseudohypoxia and aberrant VEGF signaling (*VHL, PHD, FH*, or *SDHx* [subunits of SDH]) whereas cluster 2 mutations are associated with abnormal activation of kinase signaling pathways (*RET, NF1, TMEM127, MAX*, or *KIF1Bβ*).

CLINICAL FEATURES

Its clinical presentation is so variable that pheochromocytoma has been termed "the great masquerader" (Table 380-1). Among the presenting manifestations, episodes of palpitation, headache, and profuse sweating are typical, and these manifestations constitute a classic triad. The presence of all three manifestations in association with hypertension makes pheochromocytoma a likely diagnosis. However, a pheochromocytoma can be asymptomatic for years, and some tumors grow to a considerable size, before patients note symptoms.

TABLE 380-1

Clinical Features Associated with Pheochromocytoma, Listed by Frequency of Occurrence

1. Headaches	10. Weight loss
2. Profuse sweating	11. Paradoxical response to antihypertensive drugs
3. Palpitations and tachycardia	12. Polyuria and polydipsia
4. Hypertension, sustained or paroxysmal	13. Constipation
5. Anxiety and panic attacks	14. Orthostatic hypotension
6. Pallor	15. Dilated cardiomyopathy
7. Nausea	16. Erythrocytosis
8. Abdominal pain	17. Elevated blood sugar
9. Weakness	18. Hypercalcemia

The dominant sign is hypertension. Classically, patients have episodic hypertension, but sustained hypertension is also common. Catecholamine crises can lead to heart failure, pulmonary edema, arrhythmias, and intracranial hemorrhage. During episodes of hormone release, which can occur at widely divergent intervals, patients are anxious and pale, and they experience tachycardia and palpitations. These paroxysms generally last <1 h and may be

precipitated by surgery, positional changes, exercise, pregnancy, urination (particularly with bladder pheochromocytomas), and various medications (e.g., tricyclic antidepressants, opiates, metoclopramide).

DIAGNOSIS

The diagnosis is based on documentation of catecholamine excess by biochemical testing and localization of the tumor by imaging. These two criteria are of equal importance, although measurement of catecholamines or metanephrines (their methylated metabolites) is traditionally the first step in diagnosis.

Biochemical Testing

Pheochromocytomas and paragangliomas synthesize and store catecholamines, which include norepinephrine (noradrenaline), epinephrine (adrenaline), and dopamine. Elevated plasma and urinary levels of catecholamines and metanephrines form the cornerstone of diagnosis. The characteristic fluctuations in the hormonal activity of tumors results in considerable variation in serial catecholamine measurements. However, most tumors continuously leak O-methylated metabolites, which are detected by measurement of metanephrines.

Catecholamines and metanephrines can be measured by different methods, including high-performance liquid chromatography, enzyme-linked immunosorbent assay, and liquid chromatography/mass spectrometry. When pheochromocytoma is suspected on clinical grounds (i.e., when values are three times the upper limit of normal), this diagnosis is highly likely regardless of the assay used. However, as summarized in **Table 380-2**, the sensitivity and specificity of available biochemical tests vary greatly, and these differences are important in assessing patients with borderline elevations of different compounds. Urinary tests for metanephrines (total or fractionated) and catecholamines are widely available and are used commonly for initial evaluation. Among these tests, those for the fractionated metanephrines and catecholamines are the most sensitive. Plasma tests are more convenient and include measurements of catecholamines and metanephrines. Measurements of plasma metanephrine are the most sensitive and are less susceptible to false-positive elevations from stress, including venipuncture. Although the incidence of false-positive test results has been reduced by the introduction of newer assays, physiologic stress responses and medications that increase catecholamine levels still can confound testing. Because the tumors are relatively rare, borderline elevations are likely to represent false-positive results. In this circumstance, it is important to exclude dietary or drug-related factors (withdrawal of levodopa or use of sympathomimetics, diuretics, tricyclic antidepressants, alpha and beta blockers) that might cause false-positive results and then to repeat testing or perform a clonidine suppression test (i.e., the measurement of plasma normetanephrine 3 h after oral administration of 300 μg of clonidine). Other pharmacologic tests, such as the phentolamine test and the glucagon provocation test, are of relatively low sensitivity and are not recommended.

TABLE 380-2

Biochemical and Imaging Methods Used for Diagnosis of Pheochromocytoma and Paraganglioma

Diagnostic Method	Sensitivity	Specificity
24-h urinary tests		
Catecholamines	+++	+++
Fractionated metanephrines	++++	++
Total metanephrines	+++	++++
Plasma tests		
Catecholamines	+++	++
Free metanephrines	++++	+++
Imaging		
CT	++++	+++
MRI	++++	+++
MIBG scintigraphy	+++	++++
Somatostatin receptor scintigraphy[a]	++	++
^{18}Fluoro-DOPA PET/CT	+++	++++

[a]Values are particularly high in head and neck paragangliomas.

Abbreviations: MIBG, metaiodobenzylguanidine; PET/CT, positron emission tomography plus CT. For the biochemical tests, the ratings correspond globally to sensitivity and specificity rates as follows: ++, <85%; +++, 85–95%; and ++++, >95%.

Diagnostic Imaging

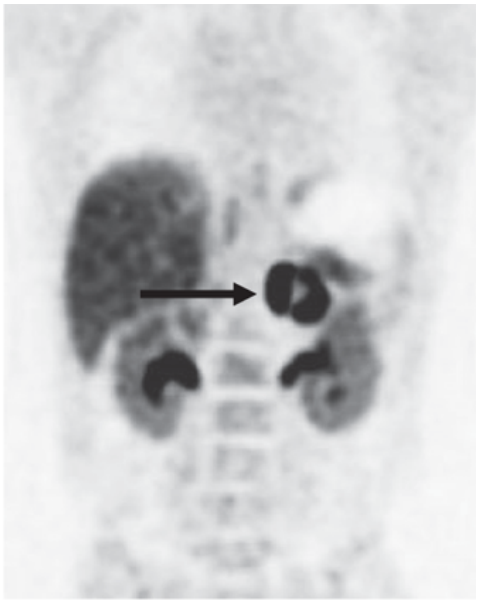
A variety of methods have been used to localize pheochromocytomas and paragangliomas (Table 380-2, **Figs. 380-2, 380-3, and 380-4**). CT and MRI are similar in sensitivity and should be performed with contrast. T2-weighted MRI with gadolinium contrast is optimal for detecting pheochromocytomas and is somewhat better than CT for imaging extraadrenal pheochromocytomas and paragangliomas. About 5% of adrenal incidentalomas, which usually are detected by CT or MRI, prove to be pheochromocytomas upon endocrinologic evaluation.

FIGURE 380-2

Typical pheochromocytoma (adrenal unilateral). ***A.*** MRI ***B.*** ^{18}F-DOPA positron emission tomography (PET), tumor marked by arrows. *(Part A was provided courtesy of Dr. Tobias Krauss, Freiburg. Part B was provided courtesy of Dr. Juri Ruf, Freiburg.)*



A

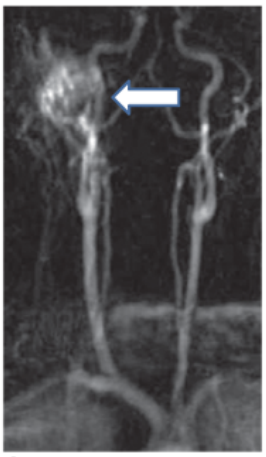


B

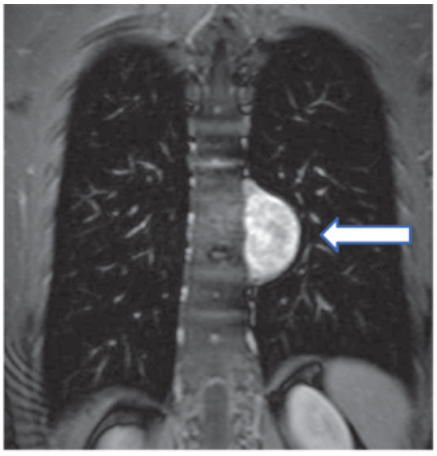
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 380-3

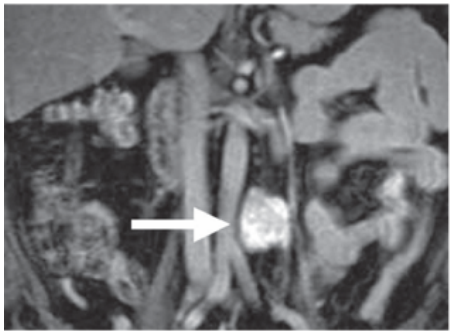
Paragangliomas (Extraadrenal pheochromocytomas). ***A.*** Carotid body tumor. ***B.*** Thoracic tumor. ***C.*** Interaorto-caval tumor. ***D.*** Pelvic tumor at the anterior wall of the urinary bladder. Tumors marked by arrows. *(Part A was provided courtesy of Dr. Carsten Boedeker, Stralsund. Parts B and D were provided courtesy of Dr Tobias Krauss, Freiburg. Part C was provided courtesy of Dr Martin Walz, Essen.)*



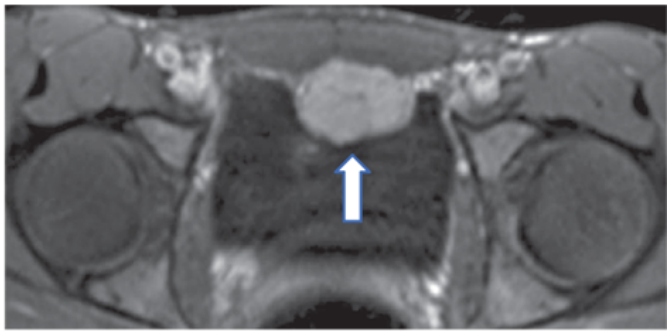
A



B



C



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 380-4

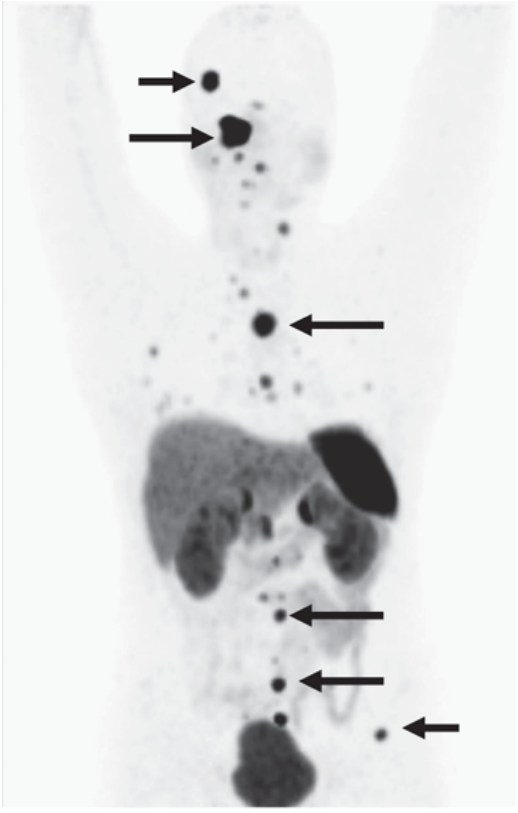
Multiple and metastatic pheochromocytoma. ***A.*** Paraganglioma syndrome**.** A patient with the SDHD W5X mutation and PGL1 ^{68}Ga-DOTATATE positron emission tomography (PET) demonstrating tumor uptake in the right jugular glomus, the right and left carotid body, both adrenal glands and an interaorto-

caval paraganglion (arrows). Note the physiologic accumulation of the radiopharmaceutical agent in the kidneys and the liver. ***B.*** ^{18}F-DOPA PET of a patient with metastatic pheochromocytoma. Several metastases marked by arrows. *(Parts A and B were provided courtesy of Dr. Juri Ruf, Freiburg.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

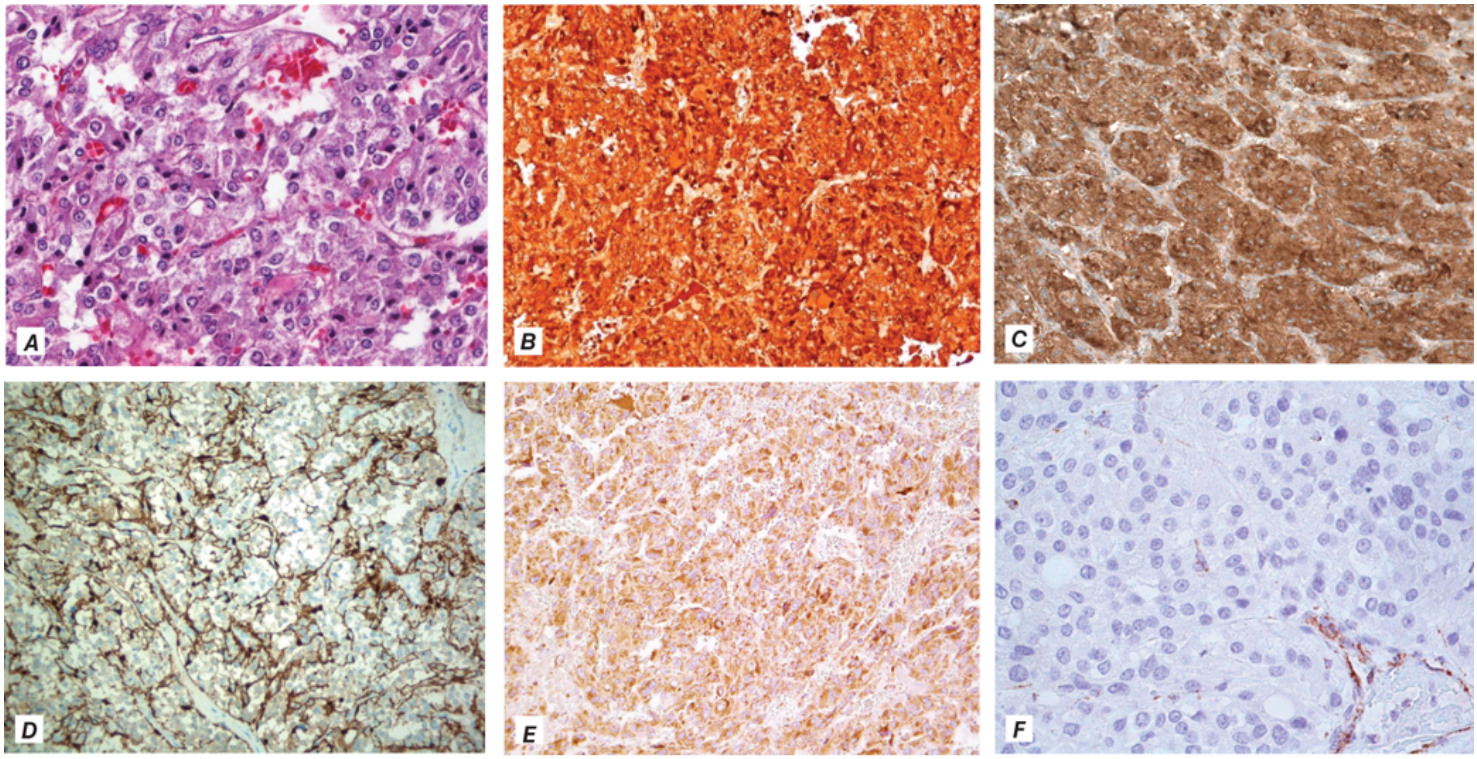
Tumors also can be localized by procedures using radioactive tracers, including ^{131}I- or ^{123}I-metaiodobenzylguanidine (MIBG) scintigraphy, ^{111}In-somatostatin analogue scintigraphy, ^{18}F-DOPA positron emission tomography (PET), ^{68}Ga-DOTATATE PET, or ^{18}F-fluorodeoxyglucose (FDG) PET (**Fig. 380-2B** and **380-4A and *B***). These agents are particularly useful in the documentation of hereditary syndromes but also in malignant pheochromocytoma, because uptake is exhibited also in paragangliomas and metastases.

Pathology

Pheochromocytomas and paragangliomas are found at the classical sites of the adrenal medulla (Fig. 380-2) and paraganglia (Fig. 380-3). Histologically the tumors often show a characteristic "Zellballen" pattern, consisting of nests of neuroendocrine chief cells with peripheral glial-like sustentacular cells. However, a broad spectrum of architectural and cytological features can be seen. Immunohistochemistry is positive for chromogranin and synaptophysin in the chief cells and S-100 in the sustentacular cells **(Fig. 380-5A-D)**. Increasingly, staining with antibodies against the proteins encoded by susceptibility genes for hereditary pheochromocytomas, such as SDHB, is used to histologically demonstrate defects of these proteins, thereby making germline mutations more likely **(Fig. 380-5E and F)**.

FIGURE 380-5

Histology and immunohistochemistry of pheochromocytoma. ***A.*** hematoxylin & eosin, ***B.*** chromogranin, ***C.*** synaptophysin, ***C*** and ***B*** stain chief cells ***D***: S-100 stains sustentacular cells. ***E***, ***F.*** Immunohistochemistry with SDHB antibody: positive staining (granular cytoplasmic staining) indicates intact SDHB (***E***), whereas negative staining (endothelial cells positive as internal control) (***F***) indicates structurally changed or absent SDHB due to a germline mutation in the *SDHB* gene, which was confirmed by molecular genetic analysis of a blood sample. *(Parts A–D and F were provided courtesy of Dr Helena Leijon, Helsinki. Part E was provided courtesy of Dr. Kurt Werner Schmid, Essen.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Differential Diagnosis

When the possibility of a pheochromocytoma is being entertained, other disorders to consider include essential hypertension, anxiety attacks, use of cocaine or amphetamines, mastocytosis or carcinoid syndrome (usually without hypertension), intracranial lesions, clonidine withdrawal, autonomic epilepsy, and factitious crises (usually from use of sympathomimetic amines). When an asymptomatic adrenal mass is identified, likely diagnoses other than pheochromocytoma include a nonfunctioning adrenal adenoma, an aldosteronoma, and a cortisol-producing adenoma (Cushing's syndrome).

TREATMENT

TREATMENT

Pheochromocytoma

Complete tumor removal, the ultimate therapeutic goal, can be achieved by partial or total adrenalectomy. It is important to preserve the normal adrenal cortex in order to prevent Addison's disease, particularly in hereditary disorders in which bilateral pheochromocytomas are most likely. Preoperative preparation of the patient has to be considered, and blood pressure should be consistently <160/90 mmHg. Classically, blood pressure has been controlled by α-adrenergic blockers (oral phenoxybenzamine, 0.5–4 mg/kg of body weight). Because patients are volume-constricted, liberal salt intake and hydration are necessary to avoid severe orthostasis. Oral prazosin or intravenous phentolamine can be used to manage paroxysms while adequate alpha blockade is awaited. Beta blockers (e.g., 10 mg of propranolol three or four times per day) can then be added. Other antihypertensives, such as calcium channel blockers or angiotensin-converting enzyme inhibitors, have also been used effectively.

Surgery should be performed by teams of surgeons and anesthesiologists with experience in the management of pheochromocytomas. Blood pressure can be labile during surgery, particularly at the outset of intubation or when the tumor is manipulated. Nitroprusside infusion is useful for intraoperative hypertensive crises, and hypotension usually responds to volume infusion. The latter side effect can, however, be avoided in normotensive pheochromocytoma patients by having on stand-by intraoperative nitroprusside, which has been shown to be safe and avoids postoperative hypotension often caused by alpha blockers; the long-lasting guideline for obligatory preoperative treatment with alpha blockers is being reconsidered.

Minimally invasive techniques (laparoscopy or retroperitoneoscopy) have become the standard approaches in pheochromocytoma surgery. They are associated with fewer complications, a faster recovery, and optimal cosmetic results. Extraadrenal abdominal and most thoracic pheochromocytomas can also be removed endoscopically. Postoperatively, catecholamine normalization should be documented. An adrenocorticotropic hormone (ACTH) test should be used to exclude cortisol deficiency when bilateral adrenal cortex–sparing surgery has been performed.

Head and neck paragangliomas are a challenge for surgeons, since damage of adjacent tissue, mainly vessels or cranial nerves is a frequent permanent side effect. Careful consideration of best management is important, and radiotherapy may be an alternative, especially for large head and neck paragangliomas.

MALIGNANT PHEOCHROMOCYTOMA

About 5–10% of pheochromocytomas and paragangliomas are malignant. The diagnosis of malignant pheochromocytoma is problematic. The typical histologic criteria of cellular atypia, presence of mitoses, and invasion of vessels or adjacent tissues are insufficient for the diagnosis of malignancy in pheochromocytoma. Thus, the term *malignant pheochromocytoma* is restricted to tumors with lymph node or distant metastases, the latter most commonly found by nuclear medicine imaging in lungs, bone, or liver locations suggesting a vascular pathway of spread **(Fig. 380-4B)**. Because hereditary syndromes are associated with multifocal tumor sites, these features should be anticipated in patients with germline mutations, especially of *RET, VHL, SDHD*, or *SDHB*. However, distant metastases also occur in these syndromes, especially in carriers of *SDHB* mutations.

Treatment of malignant pheochromocytoma or paraganglioma is challenging. Options include tumor mass reduction, alpha blockers for symptoms, chemotherapy, nuclear medicine radiotherapy and stereotactic radiation. The first-line choice is nuclear medicine therapy for scintigraphically documented metastases, preferably with ^{131}I-MIBG in 100–300 mCi doses over 3–6 cycles. Other options for radionuclide treatment are somatostatin receptor ligands, e.g., DOTATOC labeled with ^{90}Yttrium or ^{177}Lutetium, both for palliative outcomes. Averbuch's chemotherapy protocol includes dacarbazine (600 mg/m^2 on days 1 and 2), cyclophosphamide (750 mg/m^2 on day 1), and vincristine (1.4 mg/m^2 on day 1), all repeated every 21 days for 3–6 cycles. Palliation (stable disease to shrinkage) is achieved in about one-half of patients. Due to increasing insights in the genetics of pheochromocytoma, and their molecular pathways, new targeted chemotherapeutic options such as sunitinib and temozolomide/thalidomide are under development. The prognosis of metastatic pheochromocytoma or paraganglioma is variable, with 5-year survival rates of 30–60%.

PHEOCHROMOCYTOMA IN PREGNANCY

Pheochromocytomas occasionally are diagnosed in pregnancy. Endoscopic removal, preferably in the fourth to sixth month of gestation, is possible and can be followed by uneventful childbirth. Regular screening in families with inherited pheochromocytomas provides an opportunity to identify and remove asymptomatic tumors in women of reproductive age.

PHEOCHROMOCYTOMA-ASSOCIATED SYNDROMES

About 25–33% of patients with a pheochromocytoma or paraganglioma have an inherited syndrome. At diagnosis, patients with inherited syndromes are a mean of ~15 years younger than patients with sporadic tumors.

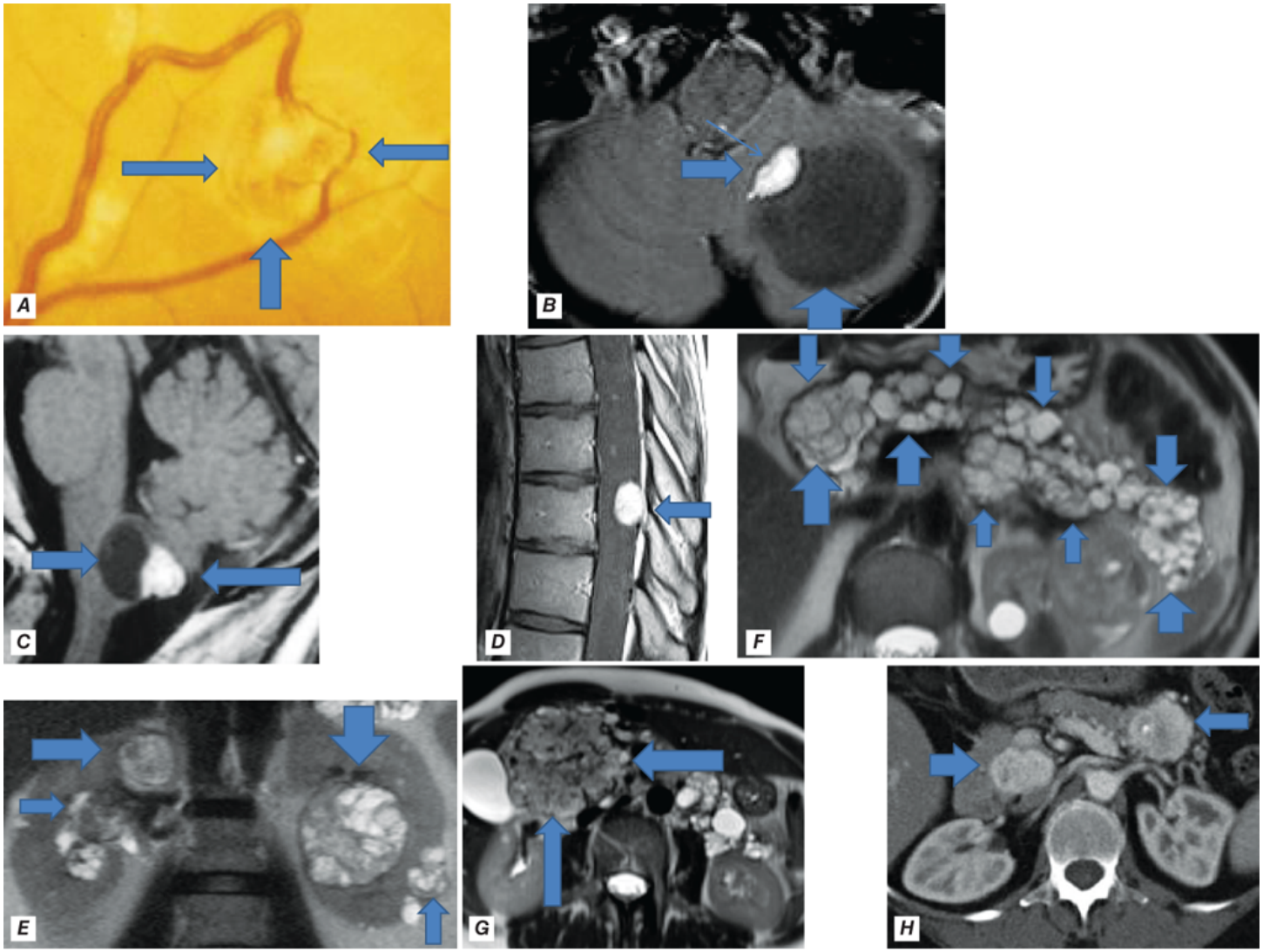
The best-known pheochromocytoma-associated syndrome is the autosomal dominant disorder MEN 2 **(see Chap. 381)**. Both types of MEN 2 (2A and 2B) are caused by mutations in *RET*, which encodes a tyrosine kinase. The locations of *RET* mutations correlate with the severity of disease and the type of MEN 2 **(see Chap. 381)**. MEN 2A is characterized by medullary thyroid carcinoma (MTC), pheochromocytoma, and hyperparathyroidism; MEN 2B also includes MTC and pheochromocytoma as well as multiple mucosal neuromas, marfanoid habitus, and other developmental disorders, though it typically lacks hyperparathyroidism. MTC is found in virtually all patients with MEN 2, but pheochromocytoma occurs in only ~50% of these patients. Nearly all pheochromocytomas in MEN 2 are benign and located in the adrenals, often bilaterally. Pheochromocytoma may be symptomatic before MTC. Prophylactic thyroidectomy is being performed in many carriers of *RET* mutations; pheochromocytomas should be excluded before any surgery in these patients.

The *paraganglioma syndromes* (PGLs) have been classified by genetic analyses of families with head and neck paragangliomas. The susceptibility genes encode subunits of the enzyme SDH, a component in the Krebs cycle and the mitochondrial electron transport chain. SDH is formed by four subunits (A–D). Mutations of *SDHA* (PGL5), *SDHB* (PGL4), *SDHC* (PGL3), *SDHD* (PGL1), and *SDHAF2* (PGL2) predispose to the PGLs. The transmission of the disease in carriers of *SDHA, SDHB*, and *SDHC* germline mutations is autosomal dominant. In contrast, in virtually all *SDHD* and *SDHAF2* families, only the progeny of affected fathers develop tumors if they inherit the mutation. PGL1 is most common, followed by PGL4; PGL2, PGL3, and PGL5 are rare. Adrenal, extraadrenal abdominal, and thoracic pheochromocytomas, which are components of PGL1, PGL4, and PGL5, are rare in PGL3 and absent in PGL2 **(Fig. 380-4A)**. About one-third of patients with PGL4 develop metastases.

VHL is an autosomal dominant disorder that predisposes to retinal and cerebellar hemangioblastomas, which also occur in the brainstem and spinal cord **(Fig. 380-6)**. Other important features of VHL are clear cell renal carcinomas, pancreatic neuroendocrine tumors, endolymphatic sac tumors of the inner ear, cystadenomas of the epididymis and broad ligament, and multiple pancreatic or renal cysts. Although the *VHL* gene can be inactivated by all types of mutations, patients with pheochromocytoma predominantly have missense mutations. About 20–30% of patients with VHL have pheochromocytomas, but in some families the incidence can reach 90%. The recognition of pheochromocytoma as a VHL-associated feature provides an opportunity to diagnose retinal, central nervous system, renal, and pancreatic tumors at a stage when effective treatment may still be possible.

FIGURE 380-6

Von Hippel–Lindau disease. Tumors and cysts marked by arrows. ***A.*** Retinal angioma (*arrows* with a pair of feeding vessels). All subsequent panels show findings on MRI: ***B–D.*** Hemangioblastomas of the cerebellum (large cyst and a solid mural tumor) (***B***) in brainstem (in part cystic) (***C***) and spinal cord (thoracic) (***D***). ***E.*** Bilateral renal clear cell carcinomas with two tumors on each side ***F.*** Multiple pancreatic cysts. ***G.*** Microcystic serous pancreatic cystadenoma (with multiple tiny spaces). ***H.*** Two pancreatic islet cell tumors. *(Part A was provided courtesy of Dr. Dieter Schmidt. Part B was provided courtesy of Dr. Christian Taschner, Freiburg. Part C was provided courtesy of Dr. Sven Glaesker, Brussels. Part D was provided courtesy of Dr. Jan-Helge Klingler, Freiburg. Part E was provided courtesy of Dr. Cordula Jilg, Freiburg. Parts F–H were provided courtesy of Dr Tobias Krauss, Freiburg.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

NF1 was the first described pheochromocytoma-associated syndrome. The *NF1* gene functions as a tumor suppressor by regulating the Ras signaling cascade. Classic features of neurofibromatosis include multiple neurofibromas, café au lait spots, axillary freckling of the skin, and Lisch nodules of the iris. Pheochromocytomas occur in only ~1% of these patients and are located predominantly in the adrenals. Malignant pheochromocytoma is not uncommon.

Other familial pheochromocytoma has been attributed to hereditary, mainly adrenal tumors in patients with germline mutations in the genes *TMEM127* and *MAX*. Transmission is also autosomal dominant, and mutations of *MAX*, like those of *SDHD,* cause tumors only if inherited from the father.

GUIDELINES FOR GENETIC SCREENING OF PATIENTS WITH PHEOCHROMOCYTOMA OR PARAGANGLIOMA

Effective preventive medicine for pheochromocytoma and pheochromocytoma-associated diseases requires management according to identified germline mutations in susceptibility genes. In addition to family history, general features suggesting an inherited syndrome include young age, multifocal tumors, extraadrenal tumors, and malignant tumors (Table 380-3 and Fig. 380-7). Because of the relatively high prevalence of familial syndromes among patients who present with pheochromocytoma or paraganglioma, it is useful to identify germline mutations even in patients without a known family history. A first step is to search for clinical features of inherited syndromes and to obtain an in-depth, multigenerational family history. Each of these syndromes exhibits autosomal dominant transmission with variable penetrance, but a proband with a mother affected by paraganglial tumors is not predisposed to PLG1 and PGL2 (*SDHD* and *SDHAF2* mutation carrier). Cutaneous neurofibromas, café au lait spots, and axillary freckling suggest neurofibromatosis. Germline mutations in *NF1* have not been reported in patients with sporadic pheochromocytomas. Thus, *NF1* testing does not need to be performed in the absence of other clinical features of neurofibromatosis. A personal or family history of MTC or an elevation of serum calcitonin strongly suggests MEN 2 and should prompt testing for *RET* mutations. A history of visual impairment or tumors of the cerebellum, kidney, brainstem, or spinal cord suggests the possibility of VHL. A personal and/or family history of head and neck paraganglioma suggests PGL1 or PGL4.

TABLE 380-3

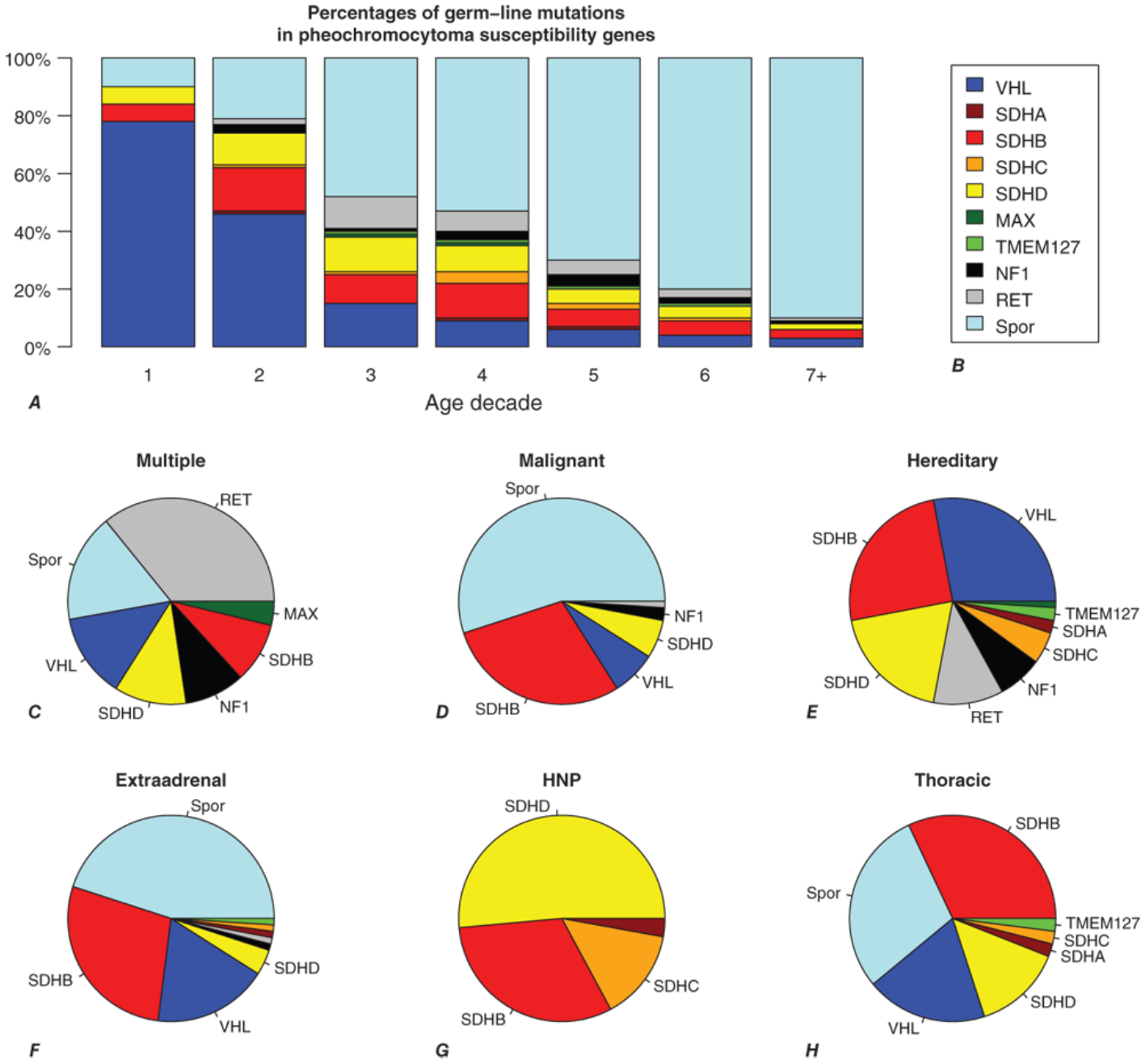
Patterns of Occurrence in Inherited Pheochromocytoma and Paraganglioma–Associated Syndromes

Mutated Gene	Adrenal Tumors	Head and Neck Tumors	Extraadrenal Retroperitoneal or Pelvic Tumors	Thoracic Tumors	Multiple Tumors	Bilateral Adrenal Tumors	Metastatic Tumors	Family History in Probands for Components of the Given Syndrome
MAX	100	<1	9	<1	82	73	9	25
NF1	98	<1	4	<1	26	26	2	15
RET	100	<1	<1	<1	61	61	<1	37
SDHA	31	44	27	3	9	4	12	3
SDHB	26	44	38	6	14	2	27	13
SDHC	3	94	<1	<1	11	<1	<1	11
SDHD	20	87	17	10	71	7	8	46
VHL	94	<1	20	3	50	57	4	37
TMEM127	95	8	1	<1	41	37	10	10

Note: Frequencies in percent of clinical characteristics of pheochromocytomas/paragangliomas of patients with germline mutations of the genes MAX, NF1, RET, SDHA, SDHB, SDHC, SDHD, VHL, and TMEM127.

FIGURE 380-7

Mutation distribution in the *VHL, RET, SDHB*, *SDHC, SDHD*, and *NF1* genes in 2796 patients with pheochromocytomas and paragangliomas from the European-American Pheochromocytoma-Paraganglioma Registry based in Freiburg, Germany, and updated as of January 1, 2017. ***A.*** Correlation with age. The bars depict the frequency of sporadic (spor) or various inherited forms of pheochromocytoma in different age groups. The inherited disorders are much more common among younger individuals presenting with pheochromocytoma. ***B.*** Percentages of mutated genes in hereditary pheochromocytomas and paragangiomas. ***C–G.*** Germline mutations according to multiple (***C***), malignant (***D***), hereditary (***E***), extraadrenal retroperitoneal (***F***), head and neck paragangliomas (**G**), thoracic (***H***). *(Data from the Freiburg International Pheochromocytoma and Paraganglioma Registry, 2017. Figures courtesy of Dr. Charis Eng, Cleveland; Dr. Ulrich Wellner, Luebeck; Dr. Birke Bausch, Freiburg; Dr. Giuseppe Opcher, Padova; and Frederic Castinetti, Marseille.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A single adrenal pheochromocytoma in a patient with an otherwise unremarkable history may still be associated with mutations of *VHL, RET, SDHB*, or *SDHD* (in decreasing order of frequency). Two-thirds of extraadrenal tumors are associated with one of these syndromes, and multifocal tumors occur with decreasing frequency in carriers of *RET, SDHD, VHL*, *SDHB*, and *MAX* mutations. About 30% of head and neck paragangliomas are associated with germline mutations of one of the SDH subunit genes (most often *SDHD*) and are rare in carriers of *VHL, RET, MAX*, and *TMEM127* mutations. Immunohistochemistry is helpful in the preselection of hereditary pheochromocytoma. Negative immunostaining with antibodies to SDHB (Fig. 380-5F), TMEM127, and MAX may predict mutations of the *SDHx* (PGL1-5), *TMEM127*, and *MAX* genes, respectively.

New technologies with whole genome sequence analysis are expected to replace targeted Sanger sequencing. It should soon be possible to search for germline mutations in a set of genes, such that all susceptibility genes for pheochromocytoma associated syndromes could be analyzed in one procedure. Of note, many sequencing protocols do not detect large deletions of one or more exons.

Once the underlying syndrome is diagnosed, the benefit of genetic testing can be extended to relatives. For this purpose, it is necessary to identify the germline mutation in the proband and, after genetic counseling, to perform DNA sequence analyses of the responsible gene in relatives to determine whether they are affected. Other family members may benefit when individuals who carry a germline mutation are biochemically screened for paraganglial tumors.

Asymptomatic paraganglial tumors, now often detected in patients with hereditary tumors and their relatives, are challenging to manage. Watchful waiting strategies have been introduced. Head and neck paragangliomas—mainly carotid body, jugular, and vagal tumors—are increasingly treated by radiation, since surgery is frequently associated with permanent palsy of cranial nerves II, VII, IX, X, XI, and XII. Nevertheless, tympanic paragangliomas are symptomatic early, and most of these tumors can easily be resected, with subsequent improvement of hearing and alleviation of tinnitus.

FURTHER READING

Bausch B et al: Characterization of endolymphatic sac tumors and von Hippel-Lindau disease in the International Endolymphatic Sac Tumor Registry. Head Neck 38: Suppl 1:E673, 2016. [PubMed: 25867206]

Bausch B et al: Systematic clinical characterization of the pheochromocytoma and paraganglioma susceptibility genes *SDHA*, *TMEM127*, *MAX* and *SDHAF2*—towards gene-informed prevention. JAMA Oncol 2017 Apr 6. doi: 10.1001/jamaoncol.2017.0223. [Epub ahead of print].

Castinetti F et al: Outcomes of adrenal-sparing surgery or total adrenalectomy in phaeochromocytoma associated with multiple endocrine neoplasia type 2: An international retrospective population-based study. Lancet Oncol 15:648, 2014. [PubMed: 24745698]

Groeben H et al: A closed case series of phaeochromotyoma and paraganglioma surgery—Time to rethink the concept of a perioperative α–receptor blockade? Br J Anaesth 118:182, 2017. [PubMed: 28100521]

Groeben H et al: Perioperative α-receptor blockade in phaeochromocytoma surgery: An observational case series. Br J Anaesth 118:182, 2017. [PubMed: 28100521]

Suarez C et al: Modern trends in the management of head and neck paragangliomas. Eur Arch Otorhinolaryngol 272:3595, 2015. [PubMed: 26475331]

Von Dobschuetz E et al: A registry-based study of thyroid paraganglioma: Histological and genetic characteristics. Endocrine-Related Cancer 22:191, 2015. [PubMed: 25595276]

Chapter 381: Multiple Endocrine Neoplasia

R. V. Thakker

INTRODUCTION

Multiple endocrine neoplasia (MEN) is characterized by a predilection for tumors involving two or more endocrine glands. Four major forms of MEN are recognized and referred to as MEN types 1–4 (MEN 1–4) **(Table 381-1)**. Each type of MEN is inherited as an autosomal dominant syndrome or may occur sporadically, that is, without a family history. However, this distinction between familial and sporadic forms is often difficult because family members with the disease may have died before symptoms developed. In addition to MEN 1–4, at least six other syndromes are associated with multiple endocrine and other organ neoplasias (MEONs) **(Table 381-2)**. These MEONs include the hyperparathyroidism-jaw tumor (HPT-JT) syndrome, Carney complex, von Hippel-Lindau disease **(Chap. 380)**, neurofibromatosis type 1 **(Chap. 86)**, Cowden's syndrome (CWS), and McCune-Albright syndrome (MAS) **(Chap. 405)**; all of these are inherited as autosomal dominant disorders, except for MAS, which is caused by mosaic expression of a postzygotic somatic cell mutation (Table 381-2).

TABLE 381-1

Multiple Endocrine Neoplasia (MEN) Syndromes

Type (Chromosomal Location)	Tumors (Estimated Penetrance)	Gene and Most Frequently Mutated Codons
MEN 1 (11q13)	Parathyroid adenoma (90%) Enteropancreatic tumor (30–70%) • Gastrinoma (>50%) • Insulinoma (10–30%) • Nonfunctioning and PPoma (20–55%) • Glucagonoma (<3%) • VIPoma (<1%) Pituitary adenoma (15–50%) • Prolactinoma (60%) • Somatotrophinoma (25%) • Corticotropinoma (<5%) • Nonfunctioning (<5%) Associated tumors • Adrenal cortical tumor (20–70%) • Pheochromocytoma (<1%) • Bronchopulmonary NET (2%) • Thymic NET (2%) • Gastric NET (10%) • Lipomas (>33%) • Angiofibromas (85%) • Collagenomas (70%) • Meningiomas (8%)	*MEN1* 83/84, 4-bp del (≈4%) 119, 3-bp del (≈3%) 209-211, 4-bp del (≈8%) 418, 3-bp del (≈4%) 514-516, del or ins (≈7%) Intron 4 ss (≈10%)

Type (Chromosomal Location)	Tumors (Estimated Penetrance)	Gene and Most Frequently Mutated Codons
MEN 2 (10 cen-10q11.2) MEN 2A MTC only MEN 2B (also known as MEN 3)	MTC (90%) Pheochromocytoma (>50%) Parathyroid adenoma (10–25%) MTC (100%) MTC (>90%) Pheochromocytoma (>50%) Associated abnormalities (40–50%) • Mucosal neuromas • Marfanoid habitus • Medullated corneal nerve fibers • Megacolon	*RET* 634, e.g., Cys → Arg (~85%) *RET* 618, missense (>50%) *RET* 918, Met → Thr (>95%)
MEN 4 (12p13)	Parathyroid adenoma[a] Pituitary adenoma[a] Reproductive organ tumors[a] (e.g., testicular cancer, neuroendocrine cervical carcinoma) ?Adrenal + renal tumors[a]	*CDKN1B*; no common mutations identified to date

[a]Insufficient numbers reported to provide prevalence information.

Note: Autosomal dominant inheritance of the MEN syndromes has been established.

Abbreviations: del, deletion; ins, insertion; MTC, medullary thyroid cancer; NET, neuroendocrine tumor; PPoma, pancreatic polypeptide–secreting tumor; VIPoma, vasoactive intestinal polypeptide–secreting tumor.

Source: Reproduced from RV Thakker et al: J Clin Endocrinol Metab 97:2990, 2012.

TABLE 381-2

Multiple Endocrine and Other Organ Neoplasia Syndromes (MEONs)

Disease[a]	Gene Product	Chromosomal Location
Hyperparathyroidism-jaw tumor (HPT-JT)	Parafibromin	1q31.2
Carney complex		
CNC1	PPKAR1A	17q24.2
CNC2	?[b]	2p16
von Hippel-Lindau disease (VHL)	pVHL (elongin)	3p25
Neurofibromatosis type 1 (NF1)	Neurofibromin	17q11.2
Cowden's syndrome (CWS)		
CWS1	PTEN	10q23.31
CWS2	SDHB	1p36.13
CWS3	SDHD	11q23.1
CWS4	KLLN	10q23.31
CWS5	PIK3CA	3q26.32
CWS6	AKT1	14q32.33
CWS7	SEC23B	20p11.23
McCune-Albright syndrome (MAS)	Gsα	20q13.32

[a]The inheritance for these disorders is autosomal dominant, except MAS, which is due to mosaicism that results from the postzygotic somatic cell mutation of the *GNAS1* gene, encoding Gsα. [b]?, unknown.

A diagnosis of a MEN or MEON syndrome may be established in an individual by one of three criteria: (1) clinical features (two or more of the associated tumors [or lesions] in an individual); (2) familial pattern (one of the associated tumors [or lesions] in a first-degree relative of a patient with a clinical diagnosis of the

syndrome); and (3) genetic analysis (a germline mutation in the associated gene in an individual, who may be clinically affected or asymptomatic). Mutational analysis in MEN and MEON syndromes is helpful in clinical practice to: (1) confirm the clinical diagnosis; (2) identify family members who harbor the mutation and require screening for relevant tumor detection and early/appropriate treatment; and (3) identify the ~50% of family members who do not harbor the germline mutation and can, therefore, be alleviated of the anxiety of developing associated tumors. This latter aspect also helps to reduce health care costs by reducing the need for unnecessary biochemical and radiologic investigations.

MULTIPLE ENDOCRINE NEOPLASIA TYPE 1

Clinical Manifestations

MEN type 1 (MEN 1), which is also referred to as Wermer's syndrome, is characterized by the triad of tumors involving the parathyroids, pancreatic islets, and anterior pituitary. In addition, adrenal cortical tumors, carcinoid tumors usually of the foregut, meningiomas, facial angiofibromas, collagenomas, and lipomas may also occur in some patients with MEN 1. Combinations of the affected glands and their pathologic features (e.g., hyperplastic adenomas of the parathyroid glands) may differ in members of the same family and even between identical twins. In addition, a nonfamilial (e.g., sporadic) form occurs in 8–14% of patients with MEN 1, and molecular genetic studies have confirmed the occurrence of de novo mutations of the *MEN1* gene in ~10% of patients with MEN 1. The prevalence of MEN 1 is ~0.25% based on randomly chosen postmortem studies but is 1–18% among patients with primary hyperparathyroidism, 16–38% among patients with pancreatic islet tumors, and <3% among patients with pituitary tumors. The disorder affects all age groups, with a reported age range of 5–81 years, with clinical and biochemical manifestations developing in the vast majority by the fifth decade. The clinical manifestations of MEN 1 are related to the sites of tumors and their hormonal products. In the absence of treatment, endocrine tumors are associated with an earlier mortality in patients with MEN 1, with a 50% probability of death by the age of 50 years. The cause of death is usually a malignant tumor, often from a pancreatic neuroendocrine tumor (NET) or foregut carcinoid. In addition, the treatment outcomes of patients with MEN 1–associated tumors are not as successful as those in patients with non–MEN 1 tumors. This is because MEN 1–associated tumors, with the exception of pituitary NETs, are usually multiple, making it difficult to achieve a successful surgical cure. Occult metastatic disease is also more prevalent in MEN 1, and the tumors may be larger, more aggressive, and resistant to treatment.

Parathyroid Tumors

Primary hyperparathyroidism occurs in ~90% of patients and is the most common feature of MEN 1 **(See also Chap. 403)**. Patients may have asymptomatic hypercalcemia or vague symptoms associated with hypercalcemia (e.g., polyuria, polydipsia, constipation, malaise, or dyspepsia). Nephrolithiasis and osteitis fibrosa cystica (less commonly) may also occur. Biochemical investigations reveal hypercalcemia, usually in association with elevated circulating parathyroid hormone (PTH) **(Table 381-3)**. The hypercalcemia is usually mild, and severe hypercalcemia or parathyroid cancer is a rare occurrence. Additional differences in the primary hyperparathyroidism of patients with MEN 1, as opposed to those without MEN 1, include an earlier

age at onset (20–25 vs 55 years) and an equal male-to-female ratio (1:1 vs 1:3). Preoperative imaging (e.g., neck ultrasound with ^{99m}Tc-sestamibi parathyroid scintigraphy) is of limited benefit because all parathyroid glands may be affected, and neck exploration may be required irrespective of preoperative localization studies.

TABLE 381-3

Biochemical and Radiological Screening in Multiple Endocrine Neoplasia Type 1

Tumor	Age to Begin (Years)	Biochemical Test (Plasma or Serum) Annually	Imaging Test (Time Interval)
Parathyroid	8	Calcium, PTH	None
Pancreatic NETs			
Gastrinoma	20	Gastrin (± gastric pH)	None
Insulinoma	5	Fasting glucose, insulin	None
Other pancreatic NET	<10	Chromogranin A; pancreatic polypeptide, glucagon, vasoactive intestinal peptide	MRI, CT, or EUS (annually)
Anterior pituitary	5	Prolactin, IGF-I	MRI (every 3 years)
Adrenal	<10	None unless symptoms or signs of functioning tumor and/or tumor >1 cm identified on imaging	MRI or CT (annually with pancreatic imaging)
Thymic and bronchial carcinoid	15	None	CT or MRI (every 1–2 years)

Abbreviations: CT, computed tomography; EUS, endoscopic ultrasound; IGF-I, insulin-like growth factor I; MRI, magnetic resonance imaging; PTH, parathyroid hormone.

Source: Reproduced from RV Thakker et al: J Clin Endocrinol Metab 97:2990, 2012.

TREATMENT

TREATMENT

Parathyroid Tumors

Surgical removal of the abnormally overactive parathyroids in patients with MEN 1 is the definitive treatment. However, it is controversial whether to perform subtotal (e.g., removal of 3.5 glands) or total parathyroidectomy with or without autotransplantation of parathyroid tissue in the forearm, and whether surgery should be performed at an early or late stage. Minimally invasive parathyroidectomy is not recommended because all four parathyroid glands are usually affected with multiple adenomas or hyperplasia. Surgical experience should be taken into account given the variability in pathology in MEN 1. Calcimimetics (e.g., cinacalcet), which act via the calcium-sensing receptor, have been used to treat primary hyperparathyroidism in some patients when surgery is unsuccessful or contraindicated.

Pancreatic Tumors

The incidence of pancreatic islet cell tumors, which are NETs, in patients with MEN 1 ranges from 30 to 80% in different series **(See also Chap. 80)**. Most of these tumors (Table 381-1) produce excessive amounts of hormone (e.g., gastrin, insulin, glucagon, vasoactive intestinal polypeptide [VIP]) and are associated with distinct clinical syndromes, although some are nonfunctioning or nonsecretory. These pancreatic islet cell tumors have an earlier age at onset in patients with MEN 1 than in patients without MEN 1.

Gastrinoma

Gastrin-secreting tumors (gastrinomas) are associated with marked gastric acid production and recurrent peptic ulcerations, a combination referred to as the Zollinger-Ellison syndrome. Gastrinomas occur more often in patients with MEN 1 who are aged >30 years. Recurrent severe multiple peptic ulcers, which may perforate, and cachexia are major contributors to the high mortality. Patients with Zollinger-Ellison syndrome may also suffer from diarrhea and steatorrhea. The diagnosis is established by demonstration of an elevated fasting serum gastrin concentration in association with increased basal gastric acid secretion (Table 381-3). However, the diagnosis of Zollinger-Ellison syndrome may be difficult in hypercalcemic MEN 1 patients, because hypercalcemia can also cause hypergastrinemia. Ultrasonography, endoscopic ultrasonography, computed tomography (CT), nuclear magnetic resonance imaging (MRI), selective abdominal angiography, venous sampling, and somatostatin receptor scintigraphy are helpful in localizing the tumor prior to surgery. Gastrinomas represent >50% of all pancreatic NETs in patients with MEN 1, and ~20% of patients with gastrinomas will be found to have MEN 1. Gastrinomas, which may also occur in the duodenal mucosa, are the major cause of morbidity and mortality in patients with MEN 1. Most MEN 1 gastrinomas are malignant and metastasize before a diagnosis is established.

TREATMENT

TREATMENT

Gastrinoma

Medical treatment of patients with MEN 1 and Zollinger-Ellison syndrome is directed toward reducing basal acid output to <10 mmol/L. Parietal cell H^+-K^+-adenosine triphosphatase (ATPase) inhibitors (e.g., omeprazole or lansoprazole) reduce acid output and are the drugs of choice for gastrinomas. Some patients may also require additional treatment with the histamine H_2 receptor antagonists, cimetidine or ranitidine.

The role of surgery in the treatment of gastrinomas in patients with MEN 1 is controversial. The goal of surgery is to reduce the risk of distant metastatic disease and improve survival. For a nonmetastatic gastrinoma situated in the pancreas, surgical excision is often effective. However, the risk of hepatic metastases increases with tumor size, such that 25–40% of patients with pancreatic NETs >4 cm develop hepatic metastases, and 50–70% of patients with tumors 2–3 cm in size have lymph node metastases. Survival in MEN 1 patients with gastrinomas <2.5 cm in size is 100% at 15 years, but 52% at 15 years, if metastatic disease is present. The presence of lymph node metastases does not appear to adversely affect survival. Surgery for gastrinomas that are >2–2.5 cm has been recommended, because the disease-related survival in these patients is improved following surgery. In addition, duodenal gastrinomas, which occur more frequently in patients with MEN 1, have been treated successfully with surgery. However, in most patients with MEN 1, gastrinomas are multiple or extrapancreatic, and with the exception of duodenal gastrinomas, surgery is rarely successful. For example, the results of one study revealed that only ~15% of patients with MEN 1 were free of disease immediately after surgery, and at 5 years, this number had decreased to ~5%; the respective outcomes in patients without MEN 1 were better, at 45% and 40%. Given these findings, most specialists recommend a nonsurgical management for gastrinomas in MEN 1, except as noted earlier for smaller, isolated lesions. Treatment of disseminated gastrinomas is difficult. Chemotherapy with streptozotocin and 5-fluorouracil; hormonal therapy with octreotide or lanreotide, which are human somatostatin analogues; hepatic artery embolization; administration of human leukocyte interferon; and removal of all resectable tumor have been successful in some patients.

Insulinoma

These β islet cell insulin-secreting tumors represent 10–30% of all pancreatic tumors in patients with MEN 1. Patients with an insulinoma present with hypoglycemic symptoms (e.g., weakness, headaches, sweating, faintness, seizures, altered behavior, weight gain) that typically develop after fasting or exertion and improve after glucose intake. The most reliable test is a supervised 72-h fast. Biochemical investigations reveal increased plasma insulin concentrations in association with hypoglycemia (Table 381-3). Circulating concentrations of C peptide and proinsulin, which are also increased, are useful in establishing the diagnosis. It also is important to demonstrate the absence of sulfonylureas in plasma and urine samples obtained during the investigation of hypoglycemia (Table 381-3). Surgical success is greatly enhanced by preoperative localization by endoscopic ultrasonography, CT scanning, or celiac axis angiography. Additional localization methods may include preoperative and perioperative percutaneous transhepatic portal venous sampling, selective intraarterial stimulation with hepatic venous sampling, and intraoperative direct pancreatic ultrasonography. Insulinomas occur in association with gastrinomas in 10% of patients with MEN 1, and the two tumors may arise at different times. Insulinomas occur more often in patients with MEN 1 who are aged

<40 years, and some arise in individuals aged <20 years. In contrast, in patients without MEN 1, insulinomas generally occur in those aged >40 years. Insulinomas may be the first manifestation of MEN 1 in 10% of patients, and ~4% of patients with insulinomas will have MEN 1.

TREATMENT

TREATMENT

Insulinoma

Medical treatment, which consists of frequent carbohydrate meals and diazoxide or octreotide, is not always successful, and surgery is the optimal treatment. Surgical treatment, which ranges from enucleation of a single tumor to a distal pancreatectomy or partial pancreatectomy, has been curative in many patients. Chemotherapy may include streptozotocin, 5-fluorouracil, and doxorubicin. Hepatic artery embolization has been used for metastatic disease.

Glucagonoma

These glucagon-secreting pancreatic NETs occur in <3% of patients with MEN 1. The characteristic clinical manifestations of a skin rash (necrolytic migratory erythema), weight loss, anemia, and stomatitis may be absent. The tumor may have been detected in an asymptomatic patient with MEN 1 undergoing pancreatic imaging or by the finding of glucose intolerance and hyperglucagonemia.

TREATMENT

TREATMENT

Glucagonoma

Surgical removal of the glucagonoma is the treatment of choice. However, treatment may be difficult because ~50–80% of patients have metastases at the time of diagnosis. Medical treatment with somatostatin analogues (e.g., octreotide or lanreotide) or chemotherapy with streptozotocin and 5-fluorouracil has been successful in some patients, and hepatic artery embolization has been used to treat metastatic disease.

Vasoactive Intestinal Peptide (VIP) Tumors (VIPomas)

VIPomas have been reported in only a few patients with MEN 1. This clinical syndrome is characterized by watery diarrhea, hypokalemia, and achlorhydria (WDHA syndrome), which is also referred to as the Verner-Morrison syndrome, or the VIPoma syndrome. The diagnosis is established by excluding laxative and diuretic abuse, confirming a stool volume in excess of 0.5–1.0 L/d during a fast, and documenting a markedly increased plasma VIP concentration.

TREATMENT

TREATMENT

Vipomas

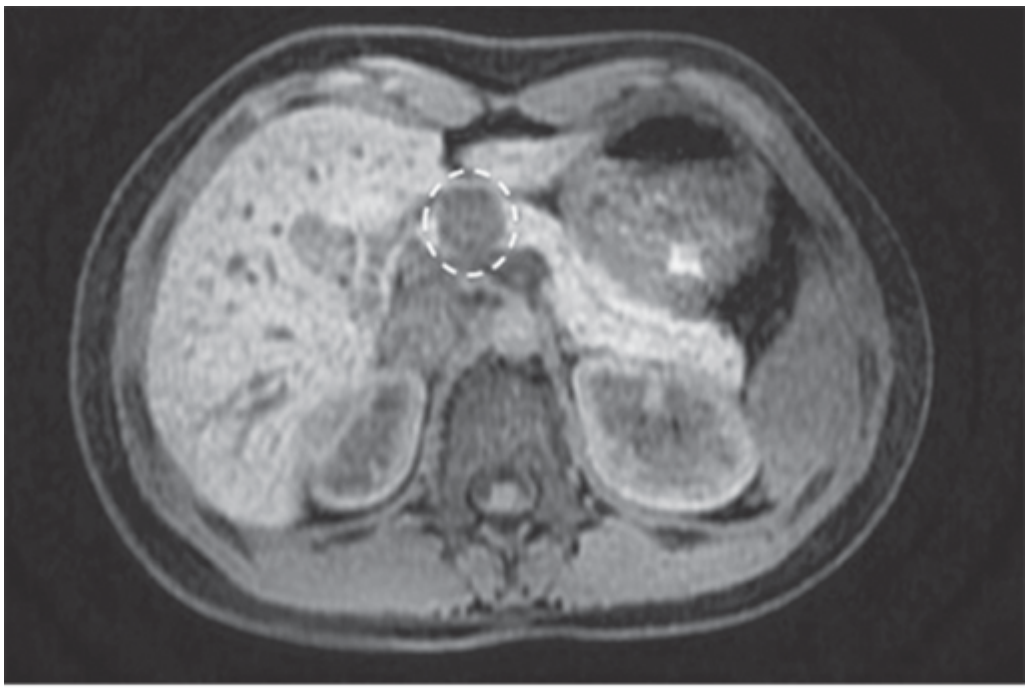
Surgical management of VIPomas, which are mostly located in the tail of the pancreas, can be curative. However, in patients with unresectable tumor, somatostatin analogues, such as octreotide and lanreotide, may be effective. Streptozotocin with 5-fluorouracil may be beneficial, along with hepatic artery embolization for the treatment of metastases.

Pancreatic Polypeptide-Secreting Tumors (PPomas) and Nonfunctioning Pancreatic NETs

PPomas are found in a large number of patients with MEN 1. No pathologic sequelae of excessive polypeptide (PP) secretion are apparent, and the clinical significance of PP is unknown. Many PPomas may have been unrecognized or classified as nonfunctioning pancreatic NETs, which likely represent the most common enteropancreatic NET associated with MEN 1 **(Fig. 381-1)**. The absence of both a clinical syndrome and specific biochemical abnormalities may result in a delayed diagnosis of nonfunctioning pancreatic NETs, which are associated with a worse prognosis than other functioning tumors, including insulinoma and gastrinoma. The optimum screening method and its timing interval for nonfunctioning pancreatic NETs remain to be established. At present, endoscopic ultrasound likely represents the most sensitive method of detecting small pancreatic tumors, but somatostatin receptor scintigraphy is the most reliable method for detecting metastatic disease (Table 381-3).

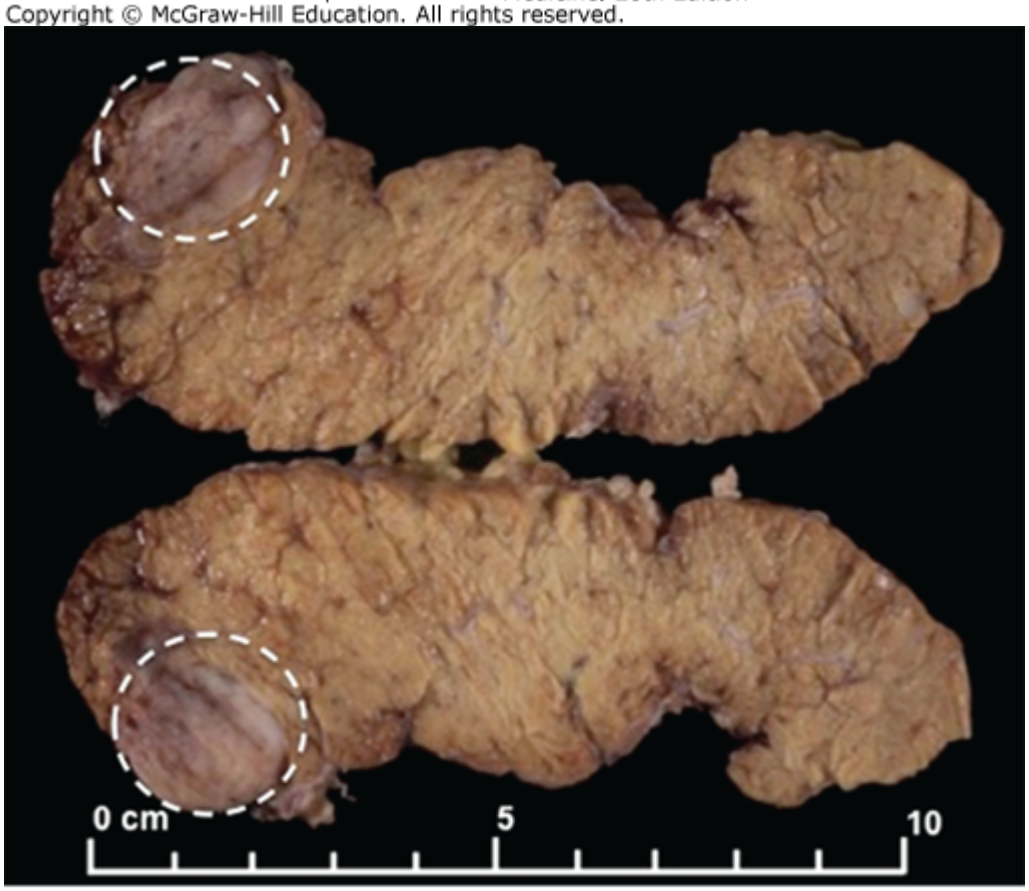
FIGURE 381-1

Pancreatic nonfunctioning neuroendocrine tumor (NET) in a 14-year-old patient with multiple endocrine neoplasia type 1 (MEN 1). ***A.*** An abdominal magnetic resonance imaging scan revealed a low-intensity >2.0 cm (anteroposterior maximal diameter) tumor within the neck of pancreas. There was no evidence of invasion of adjacent structures or metastases. The tumor is indicated by *white dashed circle*. ***B.*** The pancreatic NET was removed by surgery, and macroscopic examination confirmed the location of the tumor (*white dashed circles*) in the neck of the pancreas. Immunohistochemistry showed the tumor to immunostain for chromogranin A, but not gastrointestinal peptides or menin, thereby confirming that it was a nonsecreting NET due to loss of menin expression. *(Part A adapted with permission from PJ Newey et al: J Clin Endocrinol Metab 10:3640, 2009.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

PPomas and Nonfunctioning Pancreatic NETs

The management of nonfunctioning pancreatic NETs in the asymptomatic patient is controversial. One recommendation is to undertake surgery irrespective of tumor size after biochemical assessment is complete. Alternatively, other experts recommend surgery based on tumor size, using either >1 cm or >2 cm at different centers. Pancreatoduodenal surgery is successful in removing the tumors in 80% of patients, but >40% of patients develop complications, including diabetes mellitus, frequent steatorrhea, early and late dumping syndromes, and other gastrointestinal symptoms. However, ~50–60% of patients treated surgically survive >5 years. When considering these recommendations, it is important to consider that occult metastatic disease (e.g., tumors not detected by imaging investigations) is likely to be present in a substantial proportion of these patients at the time of presentation. Inhibitors of tyrosine kinase receptors (TKRs) and of the mammalian target of rapamycin (mTOR) signaling pathway have been reported to be effective in treating pancreatic NETs and in doubling the progression-free survival time.

Other Pancreatic NETs

NETs secreting growth hormone–releasing hormone (GHRH), GHRHomas, have been reported rarely in patients with MEN 1. It is estimated that ~33% of patients with GHRHomas have other MEN 1–related tumors. GHRHomas may be diagnosed by demonstrating elevated serum concentrations of growth hormone and GHRH. More than 50% of GHRHomas occur in the lung, 30% occur in the pancreas, and 10% are found in the small intestine. Somatostatinomas secrete somatostatin, a peptide that inhibits the secretion of a variety of hormones, resulting in hyperglycemia, cholelithiasis, low acid output, steatorrhea, diarrhea, abdominal pain, anemia, and weight loss. Although 7% of pancreatic NETs secrete somatostatin, the clinical features of somatostatinoma syndrome are unusual in patients with MEN 1.

Pituitary Tumors

Pituitary tumors occur in 15–50% of patients with MEN 1 (Table 381-1) **(See also Chap. 373)**. These occur as early as 5 years of age or as late as the ninth decade. MEN 1 pituitary adenomas are more frequent in women than men and significantly are macroadenomas (i.e., diameter >1 cm). Moreover, about one-third of these pituitary tumors show invasive features such as infiltration of tumor cells into surrounding normal juxtatumoral pituitary tissue. However, no specific histologic parameters differentiate between MEN 1 and non–MEN 1 pituitary tumors. Approximately 60% of MEN 1–associated pituitary tumors secrete prolactin, <25% secrete growth hormone, 5% secrete adrenocorticotropic hormone (ACTH), and the remainder appear to be nonfunctioning, with some secreting glycoprotein subunits (Table 381-1). However, pituitary tumors derived from MEN 1 patients may exhibit immunoreactivity to several hormones. In particular, there is a greater frequency of somatolactotrope tumors. Prolactinomas are the first manifestation of MEN 1 in ~15% of patients, whereas somatotrope tumors occur more often in patients aged >40 years. Fewer than 3% of

patients with anterior pituitary tumors will have MEN 1. Clinical manifestations are similar to those in patients with sporadic pituitary tumors without MEN 1 and depend on the hormone secreted and the size of the pituitary tumor. Thus, patients may have symptoms of hyperprolactinemia (e.g., amenorrhea, infertility, and galactorrhea in women, or impotence and infertility in men) or have features of acromegaly or Cushing's disease. In addition, enlarging pituitary tumors may compress adjacent structures such as the optic chiasm or normal pituitary tissue, causing visual disturbances and/or hypopituitarism. In asymptomatic patients with MEN 1, periodic biochemical monitoring of serum prolactin and insulin-like growth factor I (IGF-I) levels, as well as MRI of the pituitary, can lead to early identification of pituitary tumors (Table 381-3). In patients with abnormal results, hypothalamic-pituitary testing should characterize the nature of the pituitary lesion and its effects on the secretion of other pituitary hormones.

TREATMENT

TREATMENT

Pituitary Tumors

Treatment of pituitary tumors in patients with MEN 1 consists of therapies similar to those used in patients without MEN 1 and includes appropriate medical therapy (e.g., bromocriptine or cabergoline for prolactinoma; or octreotide or lanreotide for somatotrope tumors) or selective transsphenoidal adenomectomy, if feasible, with radiotherapy reserved for residual unresectable tumor tissue. Pituitary tumors in MEN 1 patients may be more aggressive and less responsive to medical or surgical treatments.

Associated Tumors

Patients with MEN 1 may also develop carcinoid tumors, adrenal cortical tumors, facial angiofibromas, collagenomas, thyroid tumors, and lipomatous tumors.

Carcinoid Tumors

Carcinoid tumors occur in >3% of patients with MEN 1 (Table 381-1) **(See also Chap. 80)**. The carcinoid tumor may be located in the bronchi, gastrointestinal tract, pancreas, or thymus. At the time of diagnosis, most patients are asymptomatic and do not have clinical features of the carcinoid syndrome. Importantly, no hormonal or biochemical abnormality (e.g., plasma chromogranin A) is consistently observed in individuals with thymic or bronchial carcinoid tumors. Thus, screening for these tumors is dependent on radiologic imaging. The optimum method for screening has not been established. CT and MRI are sensitive for detecting thymic and bronchial tumors (Table 381-3), although repeated CT scanning raises concern about exposure to repeated doses of ionizing radiation. Octreotide scintigraphy may also reveal some thymic and bronchial carcinoids, although there is insufficient evidence to recommend its routine use. Gastric carcinoids, of which the type II gastric enterochromaffin-like (ECL) cell carcinoids (ECLomas) are associated with MEN 1 and Zollinger-Ellison syndrome, may be detected incidentally at the time of gastric endoscopy for dyspeptic symptoms in MEN 1 patients. These tumors, which may be found in >10% of MEN 1 patients, are usually multiple and sized <1.5 cm. Bronchial carcinoids in patients with MEN 1 occur predominantly in women

(male-to-female ratio, 1:4). In contrast, thymic carcinoids in European patients with MEN 1 occur predominantly in men (male-to-female ratio, 20:1), with cigarette smokers having a higher risk for these tumors; thymic carcinoids in Japanese patients with MEN 1 have a less marked sex difference (male-to-female ratio 2:1). The course of thymic carcinoids in MEN 1 appears to be particularly aggressive. The presence of thymic tumors in patients with MEN 1 is associated with a median survival after diagnosis of ~9.5 years, with 70% of patients dying as a direct result of the tumor.

TREATMENT

TREATMENT

TREATMENT

Carcinoid Tumors

If resectable, surgical removal of carcinoid tumors is the treatment of choice. For unresectable tumors and those with metastatic disease, treatment with radiotherapy or chemotherapeutic agents (e.g., cisplatin, etoposide) may be used. In addition, somatostatin analogues, such as octreotide or lanreotide, have resulted in symptom improvement and regression of some tumors. Little is known about the malignant potential of gastric type II ECLomas, but treatment with somatostatin analogues, such as octreotide or lanreotide, has resulted in regression of these ECLomas.

Adrenocortical Tumors

Asymptomatic adrenocortical tumors occur in 20–70% of patients with MEN 1 depending on the radiologic screening methods used **(Table 381-1)** (See also Chap. 379). Most of these tumors, which include cortical adenomas, hyperplasia, multiple adenomas, nodular hyperplasia, cysts, and carcinomas, are nonfunctioning. Indeed, <10% of patients with enlarged adrenal glands have hormonal hypersecretion, with primary hyperaldosteronism and ACTH-independent Cushing's syndrome being encountered most commonly. Occasionally, hyperandrogenemia may occur in association with adrenocortical carcinoma. Pheochromocytoma in association with MEN 1 is rare. Biochemical investigation (e.g., plasma renin and aldosterone concentrations, low-dose dexamethasone suppression test, urinary catecholamines, and/or metanephrines) should be undertaken in those with symptoms or signs suggestive of functioning adrenal tumors or in those with tumors >1 cm. Adrenocortical carcinoma occurs in ~1% of MEN 1 patients but increases to >10% for adrenal tumors >1 cm.

TREATMENT

TREATMENT

TREATMENT

Adrenocortical Tumors

Consensus has not been reached about the management of MEN 1–associated nonfunctioning adrenal tumors, because the majority are benign. However, the risk of malignancy increases with size, particularly for tumors with a diameter >4 cm. Indications for surgery for adrenal tumors include size >4 cm in diameter, atypical or suspicious radiologic features (e.g., increased Hounsfield unit on unenhanced CT scan) and size of

1–4 cm in diameter, or significant measurable growth over a 6-month period. The treatment of functioning (e.g., hormone-secreting) adrenal tumors is similar to that for tumors occurring in non–MEN 1 patients.

Meningioma

Central nervous system (CNS) tumors, including ependymomas, schwannomas, and meningiomas, have been reported in MEN 1 patients (Table 381-1). Meningiomas are found in <10% of patients with other clinical manifestations of MEN 1 (e.g., primary hyperparathyroidism) for >15 years. The majority of meningiomas are not associated with symptoms, and 60% do not enlarge. The treatment of MEN 1–associated meningiomas is similar to that in non–MEN 1 patients.

Lipomas

Subcutaneous lipomas occur in >33% of patients with MEN 1 (Table 381-1) and are frequently multiple. In addition, visceral, pleural, or retroperitoneal lipomas may occur in patients with MEN 1. Management is conservative. However, when surgically removed for cosmetic reasons, they typically do not recur.

Facial Angiofibromas and Collagenomas

The occurrence of multiple facial angiofibromas in patients with MEN 1 may range from >20 to >90%, and occurrence of collagenomas may range from 0 to >70% (Table 381-1). These cutaneous findings may allow presymptomatic diagnosis of MEN 1 in the relatives of a patient with MEN 1. Treatment for these cutaneous lesions is usually not required.

Thyroid Tumors

Thyroid tumors, including adenomas, colloid goiters, and carcinomas, have been reported to occur in >25% of patients with MEN 1. However, the prevalence of thyroid disorders in the general population is high, and it has been suggested that the association of thyroid abnormalities in patients with MEN 1 may be incidental. The treatment of thyroid tumors in MEN 1 patients is similar to that for non–MEN 1 patients.

Genetics and Screening



The *MEN1* gene is located on chromosome 11q13 and consists of 10 exons, which encode a 610–amino acid protein, menin, that regulates transcription, genome stability, cell division, and proliferation. The pathophysiology of MEN 1 follows the Knudson two-hit hypothesis with a tumor-suppressor role for menin. Inheritance of a germline *MEN1* mutation predisposes an individual to developing a tumor that arises following a somatic mutation, which may be a point mutation or more commonly a deletion, leading to loss of heterozygosity (LOH) in the tumor DNA. The germline mutations of the *MEN1* gene are scattered throughout the entire 1830-bp coding region and splice sites, and there is no apparent correlation between the location of *MEN1* mutations and clinical manifestations of the disorder, in contrast with the situation in

patients with MEN 2 (Table 381-1). More than 10% of *MEN1* germline mutations arise de novo and may be transmitted to subsequent generations. Some families with MEN 1 mutations develop parathyroid tumors as the sole endocrinopathy, and this condition is referred to as familial isolated hyperparathyroidism (FIHP). However, between 5 and 25% of patients with MEN 1 do not harbor germline mutations or deletions of the *MEN1* gene. Such patients with MEN 1–associated tumors but without *MEN1* mutations may represent phenocopies or have mutations involving other genes. Other genes associated with MEN 1–like features include *CDC73*, which encodes parafibromin, whose mutations result in the HPT-JT syndrome; the calcium-sensing receptor gene (*CaSR*), whose mutations result in familial benign hypocalciuric hypercalcemia (FBHH); and the aryl hydrocarbon receptor interacting protein gene (*AIP*), a tumor suppressor located on chromosome 11q13 whose mutations are associated with familial isolated pituitary adenomas (FIPA). Genetic testing to determine the *MEN1* mutation status in symptomatic family members within a MEN 1 kindred, as well as to all index cases (e.g., patients) with two or more endocrine tumors, is advisable. If an *MEN1* mutation is not identified in the index case with two or more endocrine tumors, clinical and genetic tests for other disorders such as HPT-JT syndrome, FBHH, FIPA, MEN 2, or MEN 4 should be considered, because these patients may represent phenocopies for MEN 1.

The current guidelines recommend that *MEN1* mutational analysis should be undertaken in: (1) an index case with two or more MEN 1–associated endocrine tumors (e.g., parathyroid, pancreatic, or pituitary tumors); (2) asymptomatic first-degree relatives of a known *MEN1* mutation carrier; and (3) first-degree relatives of a *MEN1* mutation carrier with symptoms, signs, or biochemical or radiologic evidence for one or more MEN 1–associated tumors. In addition, *MEN1* mutational analysis should be considered in patients with suspicious or atypical MEN 1. This would include individuals with parathyroid adenomas before the age of 30 years or multigland parathyroid disease; individuals with gastrinoma or multiple pancreatic NETs at any age; or individuals who have two or more MEN 1–associated tumors that are not part of the classical triad of parathyroid, pancreatic islet, and anterior pituitary tumors (e.g., parathyroid tumor plus adrenal tumor). Family members, including asymptomatic individuals who have been identified to harbor a *MEN1* mutation, will require biochemical and radiologic screening (Table 381-3). In contrast, relatives who do not harbor the *MEN1* mutation have a risk of developing MEN 1–associated endocrine tumors that is similar to that of the general population; thus, relatives without the *MEN1* mutation do not require repeated screening.

Mutational analysis in asymptomatic individuals should be undertaken at the earliest opportunity and, if possible, in the first decade of life because tumors have developed in some children by the age of 5 years. Appropriate biochemical and radiologic investigations (Table 381-3) aimed at detecting the development of tumors should then be undertaken in affected individuals. Mutant gene carriers should undergo biochemical screening at least once per annum and also have baseline pituitary and abdominal imaging (e.g., MRI or CT), which should then be repeated at 1- to 3-year intervals (Table 381-3). Screening should commence after 5 years of age and should continue for life because the disease may develop as late as the eighth decade. The screening history and physical examination elicit the symptoms and signs of hypercalcemia; nephrolithiasis; peptic ulcer disease; neuroglycopenia; hypopituitarism; galactorrhea and amenorrhea in women; acromegaly; Cushing's disease; and visual field loss and the presence of subcutaneous lipomas, angiofibromas, and collagenomas. Biochemical screening should include measurements of serum calcium,

PTH, gastrointestinal hormones (e.g., gastrin, insulin with a fasting glucose, glucagon, VIP, PP), chromogranin A, prolactin, and IGF-I in all individuals. More specific endocrine function tests should be undertaken in individuals who have symptoms or signs suggestive of a specific clinical syndrome. Biochemical screening for the development of MEN 1 tumors in asymptomatic members of families with MEN 1 is of great importance to reduce morbidity and mortality from the associated tumors.

MULTIPLE ENDOCRINE NEOPLASIA TYPE 2 AND TYPE 3

Clinical Manifestations

MEN type 2 (MEN 2), which is also called Sipple's syndrome, is characterized by the association of medullary thyroid carcinoma (MTC), pheochromocytomas, and parathyroid tumors (Table 381-1). Three clinical variants of MEN 2 are recognized: MEN 2A, MEN 2B, and MTC only. MEN 2A, which is often referred to as MEN 2, is the most common variant. In MEN 2A, MTC is associated with pheochromocytomas in 50% of patients (may be bilateral) and with parathyroid tumors in 20% of patients. MEN 2A may rarely occur in association with Hirschsprung's disease, caused by the absence of autonomic ganglion cells in the terminal hindgut, resulting in colonic dilatation, severe constipation, and obstruction. MEN 2A may also be associated with cutaneous lichen amyloidosis, which is a pruritic lichenoid lesion that is usually located on the upper back. MEN 2B, which is also referred to as MEN 3, represents 5% of all cases of MEN 2 and is characterized by the occurrence of MTC and pheochromocytoma in association with a Marfanoid habitus; mucosal neuromas of the lips, tongue, and eyelids; medullated corneal fibers; and intestinal autonomic ganglion dysfunction leading to multiple diverticulae and megacolon. Parathyroid tumors do not usually occur in MEN 2B. MTC only (FMTC) is a variant in which MTC is the sole manifestation of the syndrome. However, the distinction between FMTC and MEN 2A is difficult and should only be considered if there are at least four family members aged >50 years who are affected by MTC but not pheochromocytomas or primary hyperparathyroidism. All of the MEN 2 variants are due to mutations of the rearranged during transfection (*RET*) protooncogene, which encodes a TKR. Moreover, there is a correlation between the locations of *RET* mutations and MEN 2 variants. Thus, ~95% of MEN 2A patients have mutations involving the cysteine-rich extracellular domain, with mutations of codon 634 accounting for ~85% of MEN 2A mutations; FMTC patients also have mutations of the cysteine-rich extracellular domain, with most mutations occurring in codon 618. In contrast, ~95% of MEN 2B/MEN 3 patients have mutations of codon 918 of the intracellular tyrosine kinase domain (Table 381-1 and **Table 381-4**).

TABLE 381-4

Recommendations for Tests and Surgery in Men 2 and Men 3[a]

		Recommended Age (years) for Test/Intervention				
RET Mutation, Exon (Ex) Location, and Codon Involved	Risk[b]	*RET* Mutational Analysis	First Serum Calcitonin and Neck Ultrasound	Prophylactic Thyroidectomy	Screening for Pheochromocytoma	Screening for PHPT
Ex8 (533)[c]; Ex10 (609, 611, 618, 620)[c]; Ex11 (630, 631, 666)[c]; Ex13 (768, 790)[c]; Ex14 (804)[c]; Ex15 (891)[c]; EX16 (912)[c]	+	<3–5	5	<5[d]	16[e]	16
Ex11 (634)[c]; Ex15 (883)[c]	++	<3	<3	<5[f]	11[e]	11
Ex15 (883)[g]; Ex16 (918)[g]	+++	ASAP and by <1	ASAP and by <0.5–1	ASAP and by <1	11[e]	—[h]

[a]Adapted from American Thyroid Association Guidelines, RT Kloos et al: Thyroid 6:565, 2009; and revised American Thyroid Association Guidelines, S Wells et al: Thyroid 25:567, 2015. [b]Risk for early development of metastasis and aggressive growth of medullary thyroid cancer: +++, highest; ++, high; and + moderate. [c]Mutations associated with MEN 2A (or medullary thyroid carcinoma only). [d] Timing of surgery to be based on elevation of serum calcitonin and/or joint discussion with pediatrician, surgeon and parent/family. Later surgery may be appropriate if serum calcitonin and neck ultrasound are normal. [e]Prescence of pheochromocytoma must be excluded prior to any surgical intervention, and also in women with *RET* mutation, who are planning pregnancy or a pregnant. [f]Surgery earlier than 5 years based on elevation of serum calcitonin. Optimal timing of surgery should be decided by the surgeon and pediatrician, in

consultation with the child's parent. [g]Mutations associated with MEN 2B (MEN 3). [h]Not required because PHPT is not a feature of MEN 2B (MEN 3).

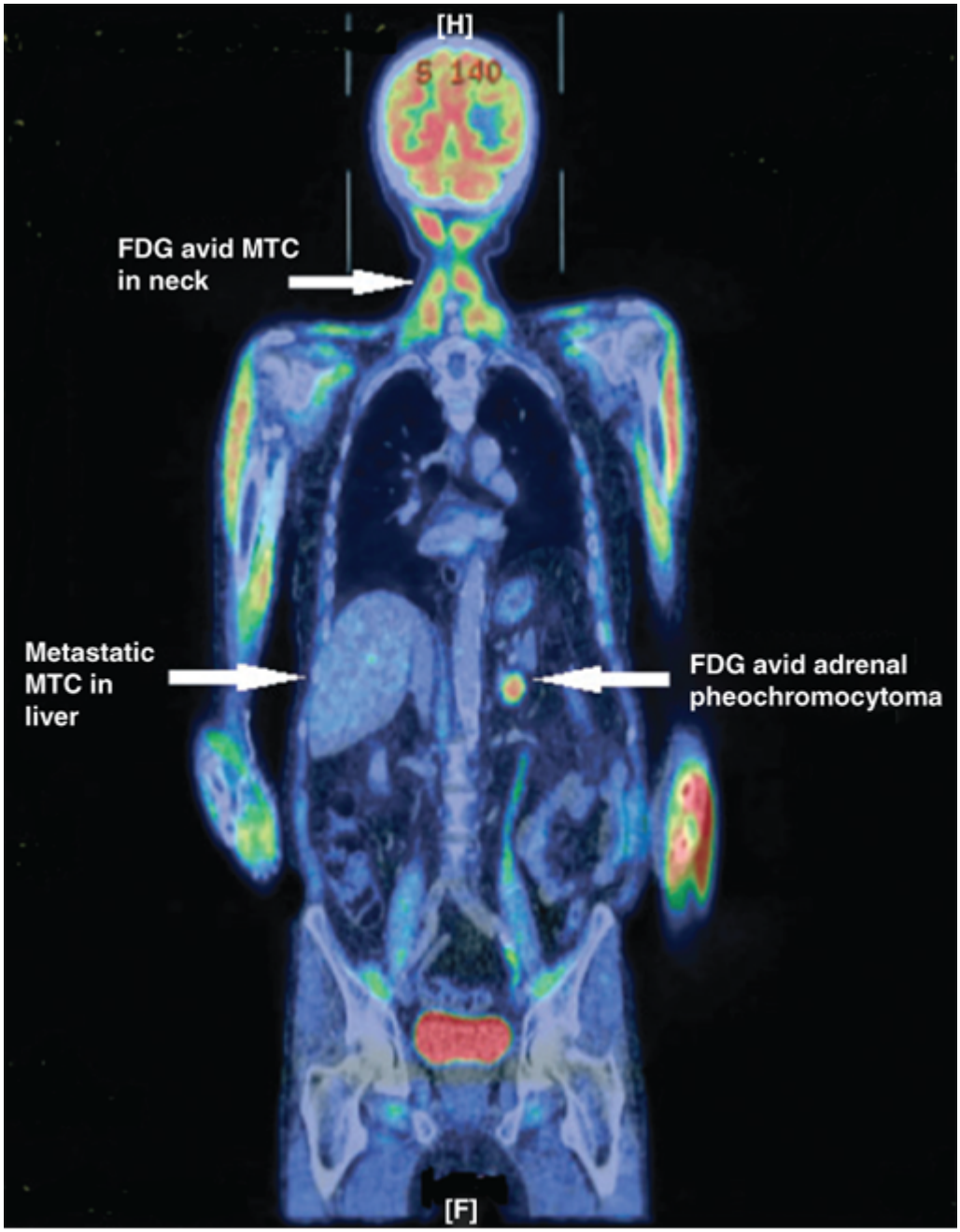
Abbreviations: ASAP, as soon as possible; MEN, multiple endocrine neoplasia; PHPT, primary hyperparathyroidism.

Medullary Thyroid Carcinoma

MTC is the most common feature of MEN 2A and MEN 2B and occurs in almost all affected individuals. MTC represents 5–10% of all thyroid gland carcinomas, and 20% of MTC patients have a family history of the disorder. The use of *RET* mutational analysis to identify family members at risk for hereditary forms of MTC has altered the presentation of MTC from that of symptomatic tumors to a preclinical disease for which prophylactic thyroidectomy (Table 381-4) is undertaken to improve the prognosis and ideally result in cure. However, in patients who do not have a known family history of MEN 2A, FMTC, or MEN 2B, and therefore have not had *RET* mutational analysis, MTC may present as a palpable mass in the neck, which may be asymptomatic or associated with symptoms of pressure or dysphagia in >15% of patients. Diarrhea occurs in 30% of patients and is associated either with elevated circulating concentrations of calcitonin or tumor-related secretion of serotonin and prostaglandins. Some patients may also experience flushing. In addition, ectopic ACTH production by MTC may cause Cushing's syndrome. The diagnosis of MTC relies on the demonstration of hypercalcitoninemia (>90 pg/mL in the basal state); stimulation tests using IV pentagastrin (0.5 mg/kg) and or calcium infusion (2 mg/kg) are rarely used now, reflecting improvements in the assay for calcitonin. Neck ultrasonography with fine-needle aspiration of the nodules can confirm the diagnosis. Radionucleotide thyroid scans may reveal MTC tumors as "cold" nodules. Radiography may reveal dense irregular calcification within the involved portions of the thyroid gland and in lymph nodes involved with metastases. Positron emission tomography (PET) may help to identify the MTC and metastases **(Fig. 381-2)**. Metastases of MTC usually occur to the cervical lymph nodes in the early stages and to the mediastinal nodes, lung, liver, trachea, adrenal, esophagus, and bone in later stages. Elevations in serum calcitonin concentrations are often the first sign of recurrence or persistent disease, and the serum calcitonin doubling time is useful for determining prognosis. MTC can have an aggressive clinical course, with early metastases and death in ~10% of patients. A family history of aggressive MTC or MEN 2B may be elicited.

FIGURE 381-2

Fluorodeoxyglucose (FDG) positron emission tomography scan in a patient with multiple endocrine neoplasia type 2A, showing medullary thyroid cancer (MTC) with hepatic and skeletal (left arm) metastasis and a left adrenal pheochromocytoma. Note the presence of excreted FDG compound in the bladder. *(Reproduced with permission from A Naziat et al: Clin Endocrinol [Oxf] 78:966, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Medullary Thyroid Carcinoma

Individuals with *RET* mutations who do not have clinical manifestations of MTC should be offered prophylactic surgery between the ages of <1 and 5 years. The timing of surgery will depend on the type of *RET* mutation and its associated risk for early development, metastasis, and aggressive growth of MTC (Table 381-4). Such patients should have a total thyroidectomy with a systematic central neck dissection to remove occult nodal metastasis, although the value of undertaking a central neck dissection has been subject to

debate. Prophylactic thyroidectomy, with lifelong thyroxine replacement, has dramatically improved outcomes in patients with MEN 2 and MEN 3, such that ~90% of young patients with *RET* mutations who had a prophylactic thyroidectomy have no evidence of persistent or recurrent MTC at 7 years after surgery. In patients with clinically evident MTC, a total thyroidectomy with bilateral central resection is recommended, and an ipsilateral lateral neck dissection should be undertaken if the primary tumor is >1 cm in size or there is evidence of nodal metastasis in the central neck. Surgery is the only curative therapy for MTC. The 10-year survival in patients with metastatic MTC is ~20%. For inoperable MTC or metastatic disease, the TKR inhibitors, such as vandetanib and cabozantinib, have improved the progression-free survival times. Other types of chemotherapy are of limited efficacy, but radiotherapy may help to palliate local disease.

Pheochromocytoma

These noradrenaline- and adrenaline-secreting tumors occur in >50% of patients with MEN 2A and MEN 2B and are a major cause of morbidity and mortality **(See also Chap. 380)**. Patients may have symptoms and signs of catecholamine secretion (e.g., headaches, palpitations, sweating, poorly controlled hypertension), or they may be asymptomatic with detection through biochemical screening based on a history of familial MEN 2A, MEN 2B, or MTC. Pheochromocytomas in patients with MEN 2A and MEN 2B differ significantly in distribution when compared with patients without MEN 2A and MEN 2B. Extra-adrenal pheochromocytomas, which occur in 10% of patients without MEN 2A and MEN 2B, are observed rarely in patients with MEN 2A and MEN 2B. Malignant pheochromocytomas are much less common in patients with MEN 2A and MEN 2B. The biochemical and radiologic investigation of pheochromocytoma in patients with MEN 2A and MEN 2B is similar to that in non–MEN 2 patients and includes the measurement of plasma (obtained from supine patients) and urinary free fractionated metanephrines (e.g., normetanephrine and metanephrines measured separately), CT or MRI scanning, radionuclide scanning with meta-iodo-(^{123}I or ^{131}I)-benzyl guanidine (MIBG), and PET using (^{18}F)-fluorodopamine or (^{18}F)-fluoro-2-dexoxy-D-glucose (Fig. 381-2).

TREATMENT

TREATMENT

Pheochromocytoma

Surgical removal of pheochromocytoma, using α and β adrenoreceptor blockade before and during the operation, is the recommended treatment. Other antihypertensive agents, including calcium channel blockers, are sometimes required for adequate blood pressure control. Endoscopic adrenal-sparing surgery, which decreases postoperative morbidity, hospital stay, and expense, as opposed to open surgery, has become the method of choice.

Parathyroid Tumors

Parathyroid tumors occur in 10–25% of patients with MEN 2A. However, >50% of these patients do not have hypercalcemia **(See also Chap. 403)**. The presence of abnormally enlarged parathyroids, which are unusually

hyperplastic, is often seen in the normocalcemic patient undergoing thyroidectomy for MTC. The biochemical investigation and treatment of hypercalcemic patients with MEN 2A is similar to that of patients with MEN 1.

Genetics and Screening



To date, ~50 different *RET* mutations have been reported, and these are located in exons 5, 8, 10, 11, 13, 14, 15, and 16. *RET* germline mutations are detected in >95% of MEN 2A, FMTC, and MEN 2B families, with Cys634Arg being most common in MEN 2A, Cys618Arg being most common in FMTC, and Met918Thr being most common in MEN 2B (Tables 381-1 and 381-4). Between 5 and 10% of patients with MTC or MEN 2A–associated tumors have de novo *RET* germline mutations, and ~50% of patients with MEN 2B have de novo *RET* germline mutations. These de novo *RET* germline mutations always occur on the paternal allele. Approximately 5% of patients with sporadic pheochromocytoma have a germline *RET* mutation, but such germline *RET* mutations do not appear to be associated with sporadic primary hyperparathyroidism. Thus, *RET* mutational analysis should be performed in: (1) all patients with MTC who have a family history of tumors associated with MEN 2, FMTC, or MEN 3, such that the diagnosis can be confirmed and genetic testing offered to asymptomatic relatives; (2) all patients with MTC and pheochromocytoma without a known family history of MEN 2 or MEN 3; (3) all patients with MTC, but without a family history of MEN 2, FMTC, or MEN 3, because these patients may have a de novo germline *RET* mutations; (4) all patients with bilateral pheochromocytoma; and (5) patients with unilateral pheochromocytoma, particularly if this occurs with increased calcitonin levels.

Screening for MEN 2/MEN 3–associated tumors in patients with *RET* germline mutations should be undertaken annually and include serum calcitonin measurements, a neck ultrasound for MTC, plasma and 24-h urinary fractionated metanephrines for pheochromocytoma, and albumin-corrected serum calcium or ionized calcium with PTH for primary hyperparathyroidism. In patients with MEN 2–associated *RET* mutations, screening for MTC should begin by 1–5 years, for pheochromocytoma by 11–16 years, and for primary hyperparathyroidism by 11–16 years of age (Table 381-4).

MULTIPLE ENDOCRINE NEOPLASIA TYPE 4

Clinical Manifestations

Patients with MEN 1–associated tumors, such as parathyroid adenomas, pituitary adenomas, and pancreatic NETs, occurring in association with gonadal, adrenal, renal, and thyroid tumors have been reported to have mutations of the gene encoding the 196–amino acid cyclin-dependent kinase inhibitor (CK1) p27 kip1 (*CDNKIB*). Such families with MEN 1–associated tumors and *CDNKIB* mutations are designated to have MEN 4 (Table 381-1). The investigations and treatments for the MEN 4–associated tumors are similar to those for MEN 1 and non–MEN 1 tumors.

Genetics and Screening

 To date, 13 different MEN 4–associated mutations of *CDKNIB*, which is located on chromosome 12p13, have been reported, and all of these are associated with a loss of function. These MEN 4 patients may represent ~3% of the 5–10% of patients with MEN 1 who do not have mutations of the *MEN1* gene. Germline *CDKNIB* mutations may rarely be found in patients with sporadic (i.e., nonfamilial) forms of primary hyperparathyroidism.

HYPERPARATHYROIDISM-JAW TUMOR SYNDROME

(See also Chap. 403)

Clinical Manifestations

Hyperparathyroidism-jaw tumor (HPT-JT) syndrome is an autosomal dominant disorder characterized by the development of parathyroid tumors (15% are carcinomas) and fibro-osseous jaw tumors. In addition, some patients may also develop Wilms' tumors, renal cysts, renal hamartomas, renal cortical adenomas, papillary RCCs, pancreatic adenocarcinomas, uterine tumors, testicular mixed germ cell tumors with a major seminoma component, and Hürthle cell thyroid adenomas. The parathyroid tumors may occur in isolation and without any evidence of jaw tumors, and this may cause confusion with other hereditary hypercalcemic disorders, such as MEN 1. However, genetic testing to identify the causative mutation will help to establish the correct diagnosis. The investigation and treatment for HPT-JT–associated tumors are similar to those in non-HPT-JT patients, except that early parathyroidectomy is advisable because of the increased frequency of parathyroid carcinoma.

Genetics and Screening

 The gene that causes HPT-JT is located on chromosome 1q31.2 and encodes a 531–amino acid protein, parafibromin (Table 381-2). Parafibromin is also referred to as cell division cycle protein 73 (CDC73) and has a role in transcription. Genetic testing in families helps to identify mutation carriers who should be periodically screened for the development of tumors **(Table 381-5)**.

TABLE 381-5

HPT-JT Screening Guidelines

Tumor[a]	Test	Frequency[b]
Parathyroid	Serum Ca, PTH	6–12 months
Ossifying jaw fibroma	Panoramic jaw x-ray with neck shielding[c]	5 years
Renal	Abdominal MRI[c,d]	5 years
Uterine	Ultrasound (transvaginal or transabdominal) and additional imaging ± D&C if indicated[e]	Annual

[a]Screening for most common HPT-JT–associated tumors is considered. Assessment for other reported tumor types may be indicated (e.g., pancreatic, thyroid, testicular tumors). [b]Frequency of repeating test after baseline tests performed. [c]X-rays and imaging involving ionizing radiation should ideally be avoided to minimize risk of generating subsequent mutations. [d]Ultrasound scan recommended if MRI unavailable. [e]Such selective pelvic imaging should be considered after obtaining a detailed menstrual history.

Abbreviations: Ca, calcium; D&C, dilation and curettage; HPT-JT, hyperparathyroidism-jaw tumor syndrome; MRI, magnetic resonance imaging; PTH, parathyroid hormone.

Source: Reproduced from PJ Newey et al: Hum Mutat 31:295, 2010.

VON HIPPEL-LINDAU DISEASE

(See also Chap. 380)

Clinical Manifestations

von Hippel-Lindau (VHL) disease is an autosomal dominant disorder characterized by hemangioblastomas of the retina and CNS; cysts involving the kidneys, pancreas, and epididymis; renal cell carcinoma (RCC); pheochromocytomas; and pancreatic islet cell tumors. The retinal and CNS hemangioblastomas are benign vascular tumors that may be multiple; those in the CNS may cause symptoms by compressing adjacent structures and/or increasing intracranial pressure. In the CNS, the cerebellum and spinal cord are the most frequently involved sites. The renal abnormalities consist of cysts and carcinomas, and the lifetime risk of

RCC in VHL is 70%. The endocrine tumors in VHL consist of pheochromocytomas and pancreatic islet cell tumors. The clinical presentation of pheochromocytoma in VHL disease is similar to that in sporadic cases, except that there is a higher frequency of bilateral or multiple tumors, which may involve extra-adrenal sites in VHL disease. The most frequent pancreatic lesions in VHL are multiple cyst-adenomas, which rarely cause clinical disease. However, nonsecreting pancreatic islet cell tumors occur in <10% of VHL patients, who are usually asymptomatic. The pancreatic tumors in these patients are often detected by regular screening using abdominal imaging. Pheochromocytomas should be investigated and treated as described earlier for MEN 2. The pancreatic islet cell tumors frequently become malignant, and early surgery is recommended.

Genetics and Screening



The *VHL* gene, which is located on chromosome 3p26-p25, is widely expressed in human tissues and encodes a 213–amino acid protein (pVHL) (Table 381-2). A wide variety of germline *VHL* mutations have been identified. *VHL* acts as a tumor-suppressor gene. A correlation between the type of mutation and the clinical phenotype has been reported; large deletions and protein-truncating mutations are associated with a low incidence of pheochromocytomas, whereas some missense mutations in VHL patients are associated with pheochromocytoma (referred to as VHL type 2C). Other missense mutations may be associated with hemangioblastomas and RCC but not pheochromocytoma (referred to as VHL type 1), whereas distinct missense mutations are associated with hemangioblastomas, RCC, and pheochromocytoma (VHL type 2B). VHL type 2A, which refers to the occurrence of hemangioblastomas and pheochromocytoma without RCC, is associated with rare missense mutations. The basis for these complex genotype-phenotype relationships remains to be elucidated. One major function of pVHL, which is also referred to as elongin, is to downregulate the expression of vascular endothelial growth factor (VEGF) and other hypoxia-inducible mRNAs. Thus, pVHL, in complex with other proteins, regulates the expression of hypoxia-inducible factors (HIF-1 and HIF-2) such that loss of functional pVHL leads to a stabilization of the HIF protein complexes, resulting in VEGF overexpression and tumor angiogenesis. Screening for the development of pheochromocytomas and pancreatic islet cell tumors is as described earlier for MEN 2 and MEN 1, respectively (Tables 381-3 and 381-4).

NEUROFIBROMATOSIS

Clinical Manifestations

Neurofibromatosis type 1 (NF1), which is also referred to as von Recklinghausen's disease, is an autosomal dominant disorder characterized by the following manifestations: neurologic (e.g., peripheral and spinal neurofibromas); ophthalmologic (e.g., optic gliomas and iris hamartomas such as Lisch nodules); dermatologic (e.g., café au lait macules); skeletal (e.g., scoliosis, macrocephaly, short stature, and pseudoarthrosis); vascular (e.g., stenoses of renal and intracranial arteries); and endocrine (e.g., pheochromocytoma, carcinoid tumors, and precocious puberty). Neurofibromatosis type 2 (NF2) is also an autosomal dominant disorder but is characterized by the development of bilateral vestibular schwannomas (acoustic neuromas) that lead to deafness, tinnitus, or vertigo. Some patients with NF2 also develop

meningiomas, spinal schwannomas, peripheral nerve neurofibromas, and café au lait macules. Endocrine abnormalities are not found in NF2 and are associated solely with NF1. Pheochromocytomas, carcinoid tumors, and precocious puberty occur in ~1% of patients with NF1, and growth hormone deficiency has also been reported. The features of pheochromocytomas in NF1 are similar to those in non-NF1 patients, with 90% of tumors being located within the adrenal medulla and the remaining 10% at an extra-adrenal location, which often involves the para-aortic region. Primary carcinoid tumors are often periampullary and may also occur in the ileum but rarely in the pancreas, thyroid, or lungs. Hepatic metastases are associated with symptoms of the carcinoid syndrome, which include flushing, diarrhea, bronchoconstriction, and tricuspid valve disease. Precocious puberty is usually associated with the extension of an optic glioma into the hypothalamus with resultant early activation of gonadotropin-releasing hormone secretion. Growth hormone deficiency has also been observed in some NF1 patients, who may or may not have optic chiasmal gliomas, but it is important to note that short stature is frequent in the absence of growth hormone deficiency in patients with NF1. The investigation and treatment for tumors are similar to those undertaken for each respective tumor type in non-NF1 patients.

Genetics and Screening


The *NF1* gene, which is located on chromosome 17q11.2 and acts as a tumor suppressor, consists of 60 exons that span more than 350 kb of genomic DNA (Table 381-2). Mutations in *NF1* are of diverse types and are scattered throughout the exons. The NF1 gene product is the protein neurofibromin, which has homologies to the p120GAP (GTPase activating protein) and acts on p21ras by converting the active GTP bound form to its inactive GDP form. Mutations of *NF1* impair this downregulation of the p21ras signaling pathways, which in turn results in abnormal cell proliferation. Screening for the development of pheochromocytomas and carcinoid tumors is as described earlier for MEN 2 and MEN 1, respectively (Tables 381-3 and 381-4).

CARNEY COMPLEX

Clinical Manifestations

Carney complex (CNC) is an autosomal dominant disorder characterized by spotty skin pigmentation (usually of the face, labia, and conjunctiva), myxomas (usually of the eyelids and heart, but also the tongue, palate, breast, and skin), psammomatous melanotic schwannomas (usually of the sympathetic nerve chain and upper gastrointestinal tract), and endocrine tumors that involve the adrenals, Sertoli cells, somatotropes, thyroid, and ovary. Cushing's syndrome, the result of primary pigmented nodular adrenal disease (PPNAD), is the most common endocrine manifestation of CNC and may occur in one-third of patients. Patients with CNC and Cushing's syndrome often have an atypical appearance by being thin (as opposed to having truncal obesity). In addition, they may have short stature, muscle and skin wasting, and osteoporosis. These patients often have levels of urinary free cortisol that are normal or increased only marginally. Cortisol production may fluctuate periodically with days or weeks of hypercortisolism; this pattern is referred to as "periodic

Cushing's syndrome." Patients with Cushing's syndrome usually have loss of the circadian rhythm of cortisol production. Acromegaly, the result of a somatotrope tumor, affects ~10% of patients with CNC. Testicular tumors may also occur in one-third of patients with CNC. These may either be large-cell calcifying Sertoli cell tumors, adrenocortical rests, or Leydig cell tumors. The Sertoli cell tumors occasionally may be estrogen-secreting and lead to precocious puberty or gynecomastia. Some patients with CNC have been reported to develop thyroid follicular tumors, ovarian cysts, or breast duct adenomas.

Genetics and Screening

 CNC type 1 (CNC1) is due to mutations of the protein kinase A (PKA) regulatory subunit 1 α (R1α) (*PPKAR1A*), a tumor suppressor, whose gene is located on chromosome 17q.24.2 (Table 381-2). The gene causing CNC type 2 (CNC2) is located on chromosome 2p16 and has not yet been identified. It is interesting to note, however, that some tumors do not show LOH of 2p16 but instead show genomic instability, suggesting that this CNC gene may not be a tumor suppressor. Screening and treatment of these endocrine tumors are similar to those described earlier for patients with MEN 1 and MEN 2 (Tables 381-3 and 381-4).

COWDEN'S SYNDROME

Clinical Manifestations

Multiple hamartomatous lesions, especially of the skin, mucous membranes (e.g., buccal, intestinal, and colonic), breast, and thyroid, are characteristic of Cowden's syndrome (CWS), which is an autosomal dominant disorder. Thyroid abnormalities occur in two-thirds of patients with CWS, and these usually consist of multinodular goiters or benign adenomas, although <10% of patients may have a follicular thyroid carcinoma. Breast abnormalities occur in >75% of patients and consist of either fibrocystic disease or adenocarcinomas. The investigation and treatment for CWS tumors are similar to those undertaken for non-CWS patients.

Genetics and Screening

 CWS is genetically heterogenous, and seven types (CWS1–7) are recognized (Table 381-2). CWS1 is due to mutations of the phosphate and tensin homologue deleted on chromosome 10 (*PTEN*) gene, located on chromosome 10q23.31. CWS2 is caused by mutations of the succinate dehydrogenase subunit B (*SDHB*) gene, located on chromosome 1p36.13; and CWS3 is caused by mutations of the *SDHD* gene, located on chromosome 11q13.1. *SDHB* and *SDHD* mutations are also associated with pheochromocytoma. CWS4 is caused by hypermethylation of the Killin (*KLLN*) gene, the promoter of which shares the same transcription site as *PTEN* on chromosome 10q23.31. CWS5 is caused by mutations of the phosphatidylinositol 3-kinase catalytic alpha (*PIK3CA*) gene on chromosome 3q26.32. CWS6 is caused by mutations of the V-Akt murine thymoma viral oncogene homolog 1 (*AKT1*) gene on chromosome 14q32.33, and CWS7 is caused by

mutations of the saccharomyces cerevisiae homology of B (SEC23B) gene and chromosome 20p11.23. Screening for thyroid abnormalities entails neck ultrasonography and fine-needle aspiration with analysis of cell cytology.

MCCUNE-ALBRIGHT SYNDROME

(See also Chap. 405)

Clinical Manifestations

McCune-Albright syndrome (MAS) is characterized by the triad of polyostotic fibrous dysplasia, which may be associated with hypophosphatemic rickets; café au lait skin pigmentation; and peripheral precocious puberty. Other endocrine abnormalities include thyrotoxicosis, which may be associated with a multinodular goiter, somatotrope tumors, and Cushing's syndrome (due to adrenal tumors). Investigation and treatment for each endocrinopathy are similar to those used in patients without MAS.

Genetics and Screening



MAS is a disorder of mosaicism that results from postzygotic somatic cell mutations of the G protein α-stimulating subunit (Gsα), encoded by the *GNAS1* gene, located on chromosome 20q13.32 (Table 381-2). The Gsα mutations, which include Arg201Cys, Arg201His, Glu227Arg, or Glu227His, are activating and are found only in cells of the abnormal tissues. Screening for hyperfunction of relevant endocrine glands and development of hypophosphatemia, which may be associated with elevated serum fibroblast growth factor 23 (FGF23) concentrations, is undertaken in MAS patients.

ACKNOWLEDGMENT

The author is grateful to the Medical Research Council (UK) for support and to Mrs. Tracey Walker for typing the manuscript.

FURTHER READING

American Thyroid Association Guidelines Task Force et al: Medullary thyroid cancer: management guidelines of the American Thyroid Association. Thyroid 19:565, 2009. [PubMed: 19469690]

Christakis I et al: Clinical features, treatments, and outcomes of patients with thymic carcinoids and multiple endocrine neoplasia type 1 syndrome at MD Anderson Cancer Center. Horm Cancer 7:279, 2016. [PubMed: 27311764]

Mulligan LM: RET revisited: Expanding the oncogenic portfolio. Nat Rev Cancer 14:173, 2014. [PubMed: 24561444]

Newey PJ et al: Asymptomatic children with multiple endocrine neoplasia type 1 (*MEN1*) mutations may harbor non-functioning pancreatic neuroendocrine tumors. J Clin Endocrinol Metab 10: 3640, 2009.

Pilarski R et al: Cowden syndrome and the PTEN hamartoma tumor syndrome: Systematic review and revised diagnostic criteria. J Natl Cancer Inst 105:1607, 2013. [PubMed: 24136893]

Salpea P, Stratakis CA: Carney complex and McCune Albright syndrome: An overview of clinical manifestations and human molecular genetics. Mol Cell Endocrinol 386:85, 2014. [PubMed: 24012779]

Thakker RV: Multiple endocrine neoplasia type 1 (MEN1) and type 4 (MEN4). Mol Cell Endocrinol 5:2, 2014.

Thakker RV et al: Clinical practice guidelines for multiple endocrine neoplasia type 1 (MEN1). J Clin Endocrinol Metab 97:2990, 2012. [PubMed: 22723327]

Wells SA Jr et al: Revised American Thyroid Association guidelines for the management of medullary thyroid carcinoma. Thyroid 25:567, 2015. [PubMed: 25810047]

Yates CJ, Newey PJ, Thakker RV: Challenges and controversies in management of pancreatic neuroendocrine tumours in patients with MEN1. Lancet Diabetes Endocrinol 3:895, 2015. [PubMed: 26165399]

Chapter 382: Autoimmune Polyendocrine Syndromes

Peter A. Gottlieb; Aaron W. Michels

INTRODUCTION

Polyglandular deficiency syndromes have been given many different names, reflecting the wide spectrum of disorders that have been associated with these syndromes and the heterogeneity of their clinical presentations. The name used in this chapter for this group of disorders is *autoimmune polyendocrine syndrome* (APS). In general, these disorders are divided into two major categories, APS type 1 (APS-1) and APS type 2 (APS-2). Some groups have further subdivided APS-2 into APS type 3 (APS-3) and APS type 4 (APS-4) depending on the type of autoimmunity involved. For the most part, this additional classification does not clarify our understanding of disease pathogenesis or prevention of complications in individual patients. Importantly, there are many nonendocrine disease associations included in these syndromes, suggesting that although the underlying autoimmune disorder predominantly involves endocrine targets, it does not exclude other tissues. The disease associations found in APS-1 and APS-2 are summarized in **Table 382-1**. Understanding these syndromes and their disease manifestations can lead to early diagnosis and treatment of additional disorders in patients and their family members.

TABLE 382-1

Disease Associations with Autoimmune Polyendocrine Syndromes

Autoimmune Polyendocrine Syndrome Type 1	**Autoimmune Polyendocrine Syndrome Type 2**	**Other Autoimmune Polyendocrine Disorders**
Endocrine	**Endocrine**	IPEX (immune dysfunction polyendocrinopathy X-linked)
Addison's disease	Addison's disease	Thymic tumors
Hypoparathyroidism	Type 1 diabetes	Anti-insulin receptor antibodies
Hypogonadism	*Graves' disease or autoimmune thyroiditis*	POEMS syndrome
Graves' disease or autoimmune thyroiditis	*Hypogonadism*	Insulin autoimmune syndrome (Hirata's syndrome)
Type 1 diabetes		Adult combined pituitary hormone deficiency (CPHD) with anti-Pit1 autoantibodies
		Kearns-Sayre syndrome
		DIDMOAD syndrome
Nonendocrine	**Nonendocrine**	Congenital rubella associated with thyroiditis and/or diabetes
Mucocutaneous candidiasis	Celiac disease, dermatitis herpetiformis	
Chronic active hepatitis	Pernicious anemia	
Pernicious anemia	Vitiligo	
Vitiligo	*Alopecia*	
Asplenism	*Myasthenia gravis*	

Autoimmune Polyendocrine Syndrome Type 1	Autoimmune Polyendocrine Syndrome Type 2	Other Autoimmune Polyendocrine Disorders
Ectodermal dysplasia	*IgA deficiency*	
Alopecia	*Parkinson's disease*	
Malabsorption syndromes	*Idiopathic thrombocytopenia*	
IgA deficiency		

Abbreviations: DIDMOAD, *d*iabetes *i*nsipidus, *d*iabetes *m*ellitus, progressive bilateral *o*ptic *a*trophy, and sensorineural *d*eafness; POEMS, *p*olyneuropathy, *o*rganomegaly, *e*ndocrinopathy, *M*-protein, and *s*kin changes.

Note: Italics denote less common disorders.

APS-1

APS-1 (Online Mendelian Inheritance in Man [OMIM] 240300) has also been called autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy (APECED). Mucocutaneous candidiasis, hypoparathyroidism, and Addison's disease form the three major components of this disorder. However, as summarized in Table 382-1, many other organ systems can be involved over time. APS-1 is rare, with fewer than 500 cases reported in the literature.

The classical form of APS-1 is an autosomal recessive disorder caused by mutations in the *AIRE* gene (autoimmune regulator gene) found on chromosome 21. This gene is most highly expressed in thymic medullary epithelial cells (mTECs) where it appears to control the expression of tissue-specific self-antigens (e.g., insulin). Deletion of this regulator leads to decreased expression of tissue-specific self-antigens and is hypothesized to allow autoreactive T cells to avoid central deletion, which normally occurs during T cell maturation in the thymus. The *AIRE* gene is also expressed in epithelial cells found in peripheral lymphoid organs, but its role in these extrathymic cells remains controversial. To date, over 100 mutations have been described in this gene, and there is a higher frequency within certain ethnic groups including Iranian Jews, Sardinians, Finns, Norwegians, and Irish. Recently, several autosomal dominant mutations have been identified, and are localized primarily in the PHD1 domain of the *AIRE* gene, rather than the CARD region where the autosomal recessive mutations have been found. Individuals with this non-classical form of APS-1 may have a later onset of symptoms, and less aggressive disease, without the full spectrum of autoimmune components being expressed.

Clinical Manifestations

Classical APS-1 develops very early in life, often in infancy (Table 382-2). Chronic mucocutaneous candidiasis without signs of systemic disease is often the first manifestation. It affects the mouth and nails more frequently than the skin and esophagus. Chronic oral candidiasis can result in atrophic disease with areas suggestive of leukoplakia, which can pose a risk for future carcinoma. The etiology is associated with anticytokine autoantibodies (anti-IL-17A, IL-17F, and IL-22) related to T helper (T_H) 17 T cells and depressed production of these cytokines by peripheral blood mononuclear cells. Hypoparathyroidism usually develops next, followed by adrenal insufficiency. The time from development of one component of the disorder to the next can be many years, and the order of disease appearance is variable.

TABLE 382-2

Comparison of APS-1 and APS-2

APS-1	APS-2
Early onset: infancy	Later onset
Siblings often affected and at risk	Multigenerational
Equivalent sex distribution	Females > males affected
Monogenic: *AIRE* gene, chromosome 21, autosomal recessive	Polygenic: *HLA, MICA, PTNP22, CTLA4*
Not HLA associated for entire syndrome, some specific component risk	DR3/DR4 associated; other HLA class III gene associations noted
Autoantibodies to type 1 interferons and IL-17 and IL-22	No autoantibodies to cytokines
Autoantibodies to specific target organs	Autoantibodies to specific target organs
Asplenism	No defined immunodeficiency
Mucocutaneous candidiasis	Association with other nonendocrine immunologic disorders like myasthenia gravis and idiopathic thrombocytopenic purpura

Abbreviations: APS, autoimmune polyendocrine syndrome; IL, interleukin.

Chronic candidiasis is nearly always present and is not very responsive to treatment. Hypoparathyroidism is found in >85% of cases, and Addison's disease is found in nearly 80%. Gonadal failure appears to affect women more than men (70% vs 25%, respectively), and hypoplasia of the dental enamel also occurs frequently (77% of patients). Other endocrine disorders that occur less frequently include type 1 diabetes (23%) and autoimmune thyroid disease (18%). Nonendocrine manifestations that present less frequently include alopecia (40%), vitiligo (26%), intestinal malabsorption (18%), pernicious anemia (31%), chronic active hepatitis (17%), and nail dystrophy. An unusual and debilitating manifestation of the disorder is the development of refractory diarrhea/obstipation that may be related to autoantibody-mediated destruction of enterochromaffin or enterochromaffin-like cells. The incidence rates for many of these disorders peak in

the first or second decade of life, but the individual disease components continue to emerge over time. Therefore, prevalence rates may be higher than originally reported.

Diagnosis

The diagnosis of APS-1 is usually made clinically when two of the three major component disorders are found in an individual patient. Siblings of individuals with APS-1 should be considered affected even if only one component disorder has been detected due to the known inheritance of the syndrome. Genetic analysis of the *AIRE* gene should be undertaken to identify mutations. Detection of anti–interferon α and anti–interferon omega antibodies can identify nearly 100% of cases with APS-1. The autoantibody arises independent of the type of *AIRE* gene mutation and is not found in other autoimmune disorders.

Diagnosis of each underlying disorder should be done based on their typical clinical presentations **(Table 382-3)**. Mucocutaneous candidiasis may present throughout the gastrointestinal tract, and it may be detected in the oral mucosa or from stool samples. Evaluation by a gastroenterologist to examine the esophagus for candidiasis or secondary stricture may be merited based on symptoms. Other gastrointestinal manifestations of APS-1, including malabsorption and obstipation, may also bring these young patients to the attention of gastroenterologists for first evaluation. Specific physical examination findings of hyperpigmentation, vitiligo, alopecia, tetany, and signs of hyper- or hypothyroidism should be considered as signs of development of component disorders.

TABLE 382-3

Clinical Features and Recommended Follow-Up for APS-1 and APS-2

Component Disease	Recommended Evaluation
APS-1	
Addison's disease	Sodium, potassium, ACTH, cortisol, 21- and 17-hydroxylase autoantibodies
Diarrhea	History
Ectodermal dysplasia	Physical examination
Hypoparathyroidism	Serum calcium, phosphate, PTH
Hepatitis	Liver function tests
Hypothyroidism/Graves' disease	TSH; thyroid peroxidase and/or thyroglobulin autoantibodies and anti-TSH receptor Ab
Male hypogonadism	FSH/LH, testosterone
Malabsorption	Physical examination, anti-IL-17 and anti-IL-22 autoantibodies
Mucocutaneous candidiasis	Physical examination, mucosal swab, stool samples
Obstipation	History
Ovarian failure	FSH/LH, estradiol
Pernicious anemia	CBC, vitamin B_{12} levels
Splenic atrophy	Blood smear for Howell-Jolly bodies; platelet count; ultrasound if positive
Type 1 diabetes	Glucose, hemoglobin A_{1c}, diabetes-associated autoantibodies (insulin, GAD65, IA-2, ZnT8)
APS-2	

Component Disease	Recommended Evaluation
Addison's disease	21-Hydroxylase autoantibodies, ACTH stimulation testing if positive
Alopecia	Physical examination
Autoimmune hyper- or hypothyroidism	TSH; thyroid peroxidase and/or thyroglobulin autoantibodies, anti-TSH receptor Ab
Celiac disease	Transglutaminase autoantibodies; small intestine biopsy if positive
Cerebellar ataxia	Dictated by signs and symptoms of disease
Chronic inflammatory demyelinating polyneuropathy	Dictated by signs and symptoms of disease
Hypophysitis	Dictated by signs and symptoms of disease, anti-Pit1 autoantibody
Idiopathic heart block	Dictated by signs and symptoms of disease
IgA deficiency	IgA level
Myasthenia gravis	Dictated by signs and symptoms of disease, antiacetylcholinesterase Ab
Myocarditis	Dictated by signs and symptoms of disease
Pernicious anemia	Anti–parietal cell autoantibodies
	CBC, vitamin B_{12} levels if positive
Serositis	Dictated by signs and symptoms of disease
Stiff man syndrome	Dictated by signs and symptoms of disease
Vitiligo	Physical examination, NALP-1 polymorphism

Abbreviations: Ab, antibody; ACTH, adrenocorticotropic hormone; APS, autoimmune polyendocrine syndrome; CBC, complete blood count; FSH, follicle-stimulating hormone; IL, interleukin; LH, luteinizing hormone; PTH, parathyroid hormone; TSH, thyroid-stimulating hormone.

The development of disease-specific autoantibody assays can help confirm disease and also detect risk for future disease. For example, where possible, detection of anti-cytokine antibodies to interleukin (IL) 17 and IL-22 would confirm the diagnosis of mucocutaneous candidiasis due to APS-1. The presence of anti-21-hydroxylase antibody or anti-17-hydroxylase antibody (which may be found more commonly in adrenal insufficiency associated with APS-1) would confirm the presence or risk for Addison's disease. Other autoantibodies found in type 1 diabetes (e.g., anti-GAD65), pernicious anemia, and other component conditions should be screened for on a regular basis (6- to 12-month intervals depending on the age of the subject).

Laboratory tests, including a complete metabolic panel, phosphorous and magnesium, thyroid-stimulating hormone (TSH), adrenocorticotropic hormone (ACTH; morning), hemoglobin A_{1c}, plasma vitamin B_{12} level, and complete blood count with peripheral smear looking for Howell-Jolly bodies (asplenism), should also be performed at these time points. Detection of abnormal physical findings or test results should prompt subsequent examinations of the relevant organ system (e.g., presence of Howell-Jolly bodies indicates need for ultrasound of spleen).

TREATMENT

TREATMENT

APS-1

Therapy of individual disease components is carried out as outlined in other relevant chapters. Replacement of deficient hormones (e.g., adrenal, pancreas, ovaries/testes) will treat most of the endocrinopathies noted. Several unique issues merit special emphasis. Adrenal insufficiency can be masked by primary hypothyroidism by prolonging the half-life of cortisol. The caveat therefore is that replacement therapy with thyroid hormone can precipitate an adrenal crisis in an undiagnosed individual. Hence, all patients with hypothyroidism and the possibility of APS should be screened for adrenal insufficiency to allow treatment with glucocorticoids prior to the initiation of thyroid hormone replacement. Treatment of mucocutaneous candidiasis with ketoconazole in an individual with subclinical adrenal insufficiency may also precipitate adrenal crisis. Furthermore, mucocutaneous candidiasis may be difficult to eradicate entirely. Severe cases of disease involvement may require systemic immunomodulatory therapy, but this is not commonly needed.

APS-2

APS-2 (OMIM 269200) is more common than APS-1 with a prevalence of 1–2 in 100,000. It has a gender bias and occurs more often in female patients with a ratio of at least 3:1 compared to male patients. In contrast to APS-1, APS-2 often has its onset in adulthood with a peak incidence between 20 and 60 years of age. It shows

a familial, multigenerational heritage (Table 382-2). The presence of two or more of the following endocrine deficiencies in the same patient defines the presence of APS-2: primary adrenal insufficiency (Addison's disease; 50–70%), Graves' disease or autoimmune thyroiditis (15–69%), type 1 diabetes mellitus (T1D; 40–50%), and primary hypogonadism. Frequently associated autoimmune conditions include celiac disease (3–15%), myasthenia gravis, vitiligo, alopecia, serositis, and pernicious anemia. These conditions occur with increased frequency in affected patients but are also are found in their family members (Table 382-3).

Genetic Considerations



The overwhelming risk factor for APS-2 has been localized to the genes in the human lymphocyte antigen (HLA) complex on chromosome 6. Primary adrenal insufficiency in APS-2, but not APS-1, is strongly associated with both HLA-DR3 and HLA-DR4. Other class I and class II genes and alleles, such as HLA-B8, HLA-DQ2 and HLA-DQ8, and HLA-DR subtype such as DRB1*04:04, appear to contribute to organ-specific disease susceptibility **(Table 382-4)**. HLA-B8- and HLA-DR3-associated illnesses include selective IgA deficiency, juvenile dermatomyositis, dermatitis herpetiformis, alopecia, scleroderma, autoimmune thrombocytopenia purpura, hypophysitis, metaphyseal osteopenia, and serositis.

TABLE 382-4

APS-2 and Other Polyendocrine Disorder Associations

Disease	HLA Association	Initiating Factor	Mechanism	Autoantigen
Graves' Disease	DR3	Iodine Anti-CD52	Antibody	TSH receptor
Myasthenia gravis	DR3, DR7	Thymoma Penicillamine	Antibody	Acetylcholine receptor
Anti-insulin receptor	?	SLE or other autoimmune disease	Antibody	Insulin receptor
Hypoparathyroidism	?	?	Antibody	Cell surface inhibitor
Insulin autoimmune syndrome	DR4, DRB1*0406	Methimazole Sulfhydryl-containing drugs	Antibody	Insulin
Celiac disease	DQ2/DQ8	Gluten diet	T cell	Transglutaminase
Type 1 diabetes	DR3/DR4 DQ2/DQ8	? Congenital rubella	T cell	Insulin, GAD65, IA-2, ZnT8, IGRP
Addison's disease	DR3/DR4 DRB1*0404	Unknown	T cell	21-Hydroxylase P450-5cc
Thyroiditis	DR3/DQB1*0201 DQA1*0301	Iodine Interferon α	T cell	Thyroglobulin Thyroid peroxidase
Pernicious anemia	?	?	T cell	Intrinsic factor H+/K+ ATPase

Disease	HLA Association	Initiating Factor	Mechanism	Autoantigen
Vitiligo	?	Melanoma Antigen Immunization	?	Melanocyte
Chromosome dysgenesis–trisomy 21 and Turner's syndrome	DQA1*0301	?	?	Thyroid, islet, transglutaminase
Hypophysitis	?	Pit-1, TDRD6	?	Pituitary, Pit-1

Abbreviations: APS, autoimmune polyendocrine syndrome; SLE, systemic lupus erythematosus; TSH, thyroid-stimulating hormone.

Several other immune genes have been proposed to be associated with Addison's disease and therefore with APS-2 (Table 382-3). The "5.1" allele of a major histocompatibility complex (MHC) gene is an atypical class I HLA molecule MIC-A. The MIC-A5.1 allele has a very strong association with Addison's disease that is not accounted for by linkage disequilibrium with DR3 or DR4. Its role is complicated because certain HLA class I genes can offset this effect. PTPN22 codes for a polymorphism in a protein tyrosine phosphatase, which acts on intracellular signaling pathways in both T and B lymphocytes. It has been implicated in T1D, Addison's disease, and other autoimmune conditions. CTLA4 is a receptor on the T cell surface that modulates the activation state of the cell as part of the signal 2 pathway (i.e., binding to CD80/86 on antigen presenting cells). Polymorphisms of this gene appear to cause downregulation of the cell surface expression of the receptor, leading to decreased T cell activation and proliferation. This appears to contribute to Addison's disease and potentially other components of APS-2. Allelic variants of the IL-2Rα are linked to development of T1D and autoimmune thyroid disease and could contribute to the phenotype of APS-2 in certain individuals.

Diagnosis

When one of the component disorders is present, a second associated disorder occurs more commonly than in the general population (Table 382-3). There is controversy as to which tests to use and how often to screen individuals for disease. A strong family history of autoimmunity should raise suspicion in an individual with an initial component diagnosis. The development of a rarer form of autoimmunity, such as Addison's disease, should prompt more extensive screening for other linked disorders, as ~50% of Addison's disease patients develop another autoimmune diseases during their lifetime.

Circulating autoantibodies, as previously discussed, can precede the development of clinical disease by many years but would allow the clinician to follow the patient and identify the disease onset at its earliest time point (Tables 382-3 and **382-4**). For each of the endocrine components of the disorder, appropriate

autoantibody assays are listed and, if positive, should prompt physiologic testing to diagnose clinical or subclinical disease. For Addison's disease, antibodies to 21-hydroxylase antibodies are highly diagnostic for risk of adrenal insufficiency. However, individuals may take many years to develop overt symptoms of hypoadrenalism. Screening of 21-hydroxylase antibody–positive patients can be performed measuring morning ACTH and cortisol on a yearly basis. Rising ACTH values over time or low morning cortisol in association with signs or symptoms of adrenal insufficiency should prompt testing via the cosyntropin stimulation test (**Chap. 379**). T1D can be screened for by measuring autoantibodies directed against insulin, GAD65, IA-2, and ZnT8. Risk for progression to disease is based on the number of antibodies (≥2 islet autoantibodies with normal glucose tolerance is now defined as stage 1 of T1D as the lifetime risk for developing clinical symptoms is nearly 100%), and metabolic factors (impaired oral glucose tolerance test). National Institutes of Health–sponsored trial groups such as Type 1 Diabetes TrialNet are screening first- and second-degree family members for these autoantibodies and identifying prediabetic individuals who may qualify for intervention trials to change the course of the disease prior to onset. Efforts are now underway to screen the general population for T1D risk with islet autoantibodies.

Screening tests for thyroid disease can include anti–thyroid peroxidase (TPO) or anti-thyroglobulin autoantibodies or anti-TSH receptor antibodies for Graves' disease. Yearly measurements of TSH can then be used to follow these individuals. Celiac disease can be screened for using the anti–tissue transglutaminase (tTg) antibody test. For those <20 years of age, testing every 1–2 years should be performed, whereas less frequent testing is indicated after the age of 20 because the majority of individuals who develop celiac disease have the antibody earlier in life. Positive tTg antibody test results should be confirmed on repeat testing, followed by small-bowel biopsy to document pathologic changes of celiac disease. Many patients have asymptomatic celiac disease that is nevertheless associated with osteopenia and impaired growth. If left untreated, symptomatic celiac disease has been reported to be associated with an increased risk of gastrointestinal malignancy, especially lymphoma, and osteoporosis later in life.

The knowledge of the particular disease associations should guide other autoantibody or laboratory testing. A complete history and physical examination should be performed every 1–3 years including CBC, metabolic panel, TSH, and vitamin B_{12} levels to screen for most of the possible abnormalities. More specific tests should be based on specific findings from the history and physical examination.

TREATMENT

TREATMENT

APS-2

With the exception of Graves' disease, the management of each endocrine component of APS-2 involves hormone replacement and is covered in detail in the chapters on adrenal (**Chap. 379**), thyroid (**Chap. 375**), gonadal (**Chaps. 384 and 385**), and parathyroid diseases (**Chap. 403**). As noted for APS-1, adrenal insufficiency can be masked by primary hypothyroidism and should be considered and treated as discussed above. In patients with T1D, decreasing insulin requirements or hypoglycemia, without obvious secondary causes,

may indicate the emergence of adrenal insufficiency. Hypocalcemia in APS-2 patients is more likely due to malabsorption, potentially from undiagnosed Celiac disease, than hypoparathyroidism.

Immunotherapy for autoimmune endocrine disease has been reserved for T1D, for the most part, reflecting the lifetime burden of the disease for the individual patient and society. Although several immunotherapies (e.g., modified anti-CD3, rituximab, abatacept, alefacept) can prolong the honeymoon phase of T1D, none has achieved long-term success. Active research using new approaches and combination therapy may change the treatment of this disease or other autoimmune conditions that share similar pathways. Furthermore, treatment of subclinical disease diagnosed by the presence of autoantibodies may provide a mechanism to preempt the development of overt disease and is the subject of active basic and clinical research.

IPEX

*I*mmune dysregulation, *p*olyendocrinopathy, *e*nteropathy, and *X*-linked disease (IPEX; OMIM 304790) is a rare X-linked recessive disorder. The disease onset is in infancy and is characterized by severe enteropathy, T1D, and skin disease, as well as variable association with several other autoimmune disorders. Many infants die within the first days of life, but the course is variable, with some children surviving for 12–15 years. Early onset of T1D, often at birth, is highly suggestive of the diagnosis because nearly 80% of IPEX patients develop T1D. Although treatment of the individual disorders can temporarily improve the situation, treatment of the underlying immune deficiency is required and includes immunosuppressive therapy generally followed by hematopoietic stem cell transplantation. Transplantation is the only life-saving form of therapy and can be fully curative by normalizing the imbalanced immune system found in this disorder.

IPEX is caused by mutations in the *FOXP3* gene, which is also mutated in the Scurfy mouse, an animal model that shares much of the phenotype of IPEX patients. The FOXP3 transcription factor is expressed in regulatory T cells designated CD4+CD25+FOXP3+ (Treg). Lack of this factor causes a profound deficiency of this Treg population and results in rampant autoimmunity due to the lack of peripheral tolerance normally provided by these cells. Certain mutations may lead to varying forms of expression of the full syndrome, and there are rare cases where the *FOXP3* gene is intact but other genes involved in this pathway (e.g., CD25, IL-2Rα) may be causative. Future therapy with autologous CD4+ T cells transfected with a functioning *FOXP3* gene may offer a better long-term outcome than has been seen in those treated with stem cell transplantation.

THYMIC TUMORS

Thymomas and thymic hyperplasia are associated with several autoimmune diseases, with the most common being myasthenia gravis (44%) and red cell aplasia (20%). Graves' disease, T1D, and Addison's disease may also be associated with thymic tumors. Patients with myasthenia gravis and thymoma may have unique anti–acetylcholine receptor autoantibodies. Most thymomas lack AIRE expression within the thymoma, and this could be a potential factor in the development of autoimmunity. In support of this concept, thymoma is the one other disease with "frequent" development of anticytokine antibodies and

mucocutaneous candidiasis in adults. The majority of tumors are malignant, and temporary remissions of the autoimmune condition can occur with resection of the tumor.

ANTI-INSULIN RECEPTOR ANTIBODIES

This is a very rare disorder where severe insulin resistance (type B) is caused by the presence of anti-insulin receptor antibodies. It is associated with acanthosis nigricans, which can also be associated with other forms of less severe insulin resistance. About one-third of patients have an associated autoimmune illness such as systemic lupus erythematosus or Sjögren's syndrome. Therefore, the presence of antinuclear antibodies, elevated erythrocyte sedimentation rate, hyperglobulinemia, leukopenia, and hypocomplementemia may accompany the presentation. The presence of anti-insulin receptor autoantibodies leads to marked insulin resistance, requiring >100,000 units of insulin to be given daily with only partial control of hyperglycemia. Patients can also have severe hypoglycemia due to partial activation of the insulin receptor by the antibody. The course of the disease is variable, and several patients have had spontaneous remissions. A therapeutic approach that targets B lymphocytes, including rituximab, cyclophosphamide, and pulse steroids has been validated in follow-on case reports to induce remission of the disease.

INSULIN AUTOIMMUNE SYNDROME (HIRATA'S SYNDROME)

The insulin autoimmune syndrome, associated with Graves' disease and methimazole therapy (or other sulfhydryl-containing medications), is of particular interest due to a remarkably strong association with a specific HLA haplotype. Such patients with elevated titers of anti-insulin antibodies frequently present with hypoglycemia. In Japan, the disease is restricted to HLA-DR4-positive individuals with DRB1*04:06. Curiously, a recent report demonstrated that five out of six Caucasian patients taking lipoic acid (sulfhydryl group) who developed insulin autoimmune syndrome were primarily DRB1*04:03 (which is related to DRB1*04:06); the sixth was DRB1*04:06. In Hirata's syndrome the anti-insulin antibodies are often polyclonal. Discontinuation of the medication generally leads to resolution of the syndrome over time. There are very rare cases of insulin autoimmune syndrome not associated with sulfhydryl-containing medications that result in profound, life-threatening hypoglycemia. Treatment involves treating the underlying condition that causes anti-insulin antibodies, such as a B lymphocyte lymphoma (tend to have monoclonal insulin antibodies) or systemic lupus erythematosus. As hypoglycemia is profound when elevated titers of high affinity insulin antibodies bind secreted insulin and then release it into circulation, treatment begins with high dose glucocorticoids and potentially rituximab to target B lymphocytes.

POEMS SYNDROME

POEMS (*p*olyneuropathy, *o*rganomegaly, *e*ndocrinopathy, *M*-protein, and *s*kin changes; also known as Crow-Fukase syndrome; OMIM 192240) patients usually present with a progressive sensorimotor polyneuropathy, diabetes mellitus (50%), primary gonadal failure (70%), and a plasma cell dyscrasia with sclerotic bony lesions. Associated findings can be hepatosplenomegaly, lymphadenopathy, and hyperpigmentation.

Patients often present in the fifth to sixth decade of life and have a median survival after diagnosis of <3 years. The syndrome is assumed to be secondary to circulating immunoglobulins, but patients have excess vascular endothelial growth factor as well as elevated levels of other inflammatory cytokines such as IL1-β, IL-6, and tumor necrosis factor α. Patients have been treated with thalidomide, and more recently lenalidomide, leading to a decrease in vascular endothelial growth factor. Hyperglycemia responds to small, subcutaneous doses of insulin. The hypogonadism is due to primary gonadal disease with elevated plasma levels of follicle-stimulating hormone and luteinizing hormone. Temporary resolution of the features of POEMS, including normalization of blood glucose, may occur after radiotherapy for localized plasma cell lesions of bone or after chemotherapy, lenalidomide and dexamethasone, or autologous stem cell transplantation.

OTHER DISORDERS

Other diseases can exhibit polyendocrine deficiencies, including Kearns-Sayre syndrome, DIDMOAD syndrome (*di*abetes *i*nsipidus, *di*abetes *m*ellitus, progressive bilateral *o*ptic *a*trophy, and sensorineural *d*eafness; also termed Wolfram's syndrome), Down's syndrome or trisomy 21 (OMIM 190685), Turner's syndrome (monosomy X, 45,X0), and congenital rubella.

Kearns-Sayre syndrome (OMIM 530000) is a rare mitochondrial DNA disorder characterized by myopathic abnormalities leading to ophthalmoplegia and progressive weakness in association with several endocrine abnormalities, including hypoparathyroidism, primary gonadal failure, diabetes mellitus, and hypopituitarism. Crystalline mitochondrial inclusions are found in muscle biopsy specimens, and such inclusions have also been observed in the cerebellum. Antiparathyroid antibodies have not been described; however, antibodies to the anterior pituitary gland and striated muscle have been identified, and the disease may have autoimmune components. These mitochondrial DNA mutations occur sporadically and do not appear to be associated with a familial syndrome.

Wolfram's syndrome (OMIM 222300, chromosome 4; OMIM 598500, mitochondrial) is a rare autosomal recessive disease that is also called DIDMOAD. Neurologic and psychiatric disturbances are prominent in most patients and can cause severe disability. The disease is caused by defects in *Wolfram Syndrome 1* (WFS1) gene, which encodes a 100-kDa transmembrane protein that has been localized to the endoplasmic reticulum and is found in neuronal and neuroendocrine tissue. Its expression induces ion channel activity with a resultant increase in intracellular calcium and may play an important role in intracellular calcium homeostasis. Wolfram's syndrome appears to be a slowly progressive neurodegenerative process, and there is nonautoimmune selective destruction of the pancreatic beta cells. Diabetes mellitus with an onset in childhood is usually the first manifestation. Diabetes mellitus and optic atrophy are present in all reported cases, but expression of the other features is variable. Treatments targeting endoplasmic reticulum dysfunction are being tested and may be a bridge till gene therapy can be developed to treat the most severely affected cases.

Down's syndrome, or trisomy 21 (OMIM 190685), is associated with the development of T1D, thyroiditis, and celiac disease. Patients with Turner's syndrome also appear to be at increased risk for the development of thyroid disease and celiac disease. It is recommended to screen patients with trisomy 21 and Turner's syndrome for associated autoimmune diseases on a regular basis.

GLOBAL CONSIDERATIONS



Identification of these syndromes requires access to central laboratories with the ability to detect unique autoantibodies and to sequence the specific genes that may underlie these disorders. Early recognition of the clinical features of these disorders and timely referral and/or consultation with tertiary care centers to confirm the diagnosis and initiate therapy is important to improving outcomes. The *AIRE* recessive gene mutations found in APS-1 were originally described in high frequency in several populations including Finns, Iranian Jews, Sardinians, Norwegians, and Irish. Although individuals from many other countries have now been found to have these mutations, and the newly identified dominant *AIRE* gene mutations, understanding the frequency in the background population may raise the clinician's level of suspicion for these rare disorders. Hirata syndrome was originally reported in Japanese populations, but also may be found in other populations as noted.

FURTHER READING

Anderson MS, Su MA: AIRE expands: New roles in immune tolerance and beyond. Nat Rev Immunol 16:247, 2016. [PubMed: 26972725]

Bruserud O et al: A longitudinal follow-up of autoimmune polyendocrine syndrome type 1. J Clin Endocrinol Metab 101:2975, 2016. [PubMed: 27253668]

Oftedal BE et al: Dominant mutations in the autoimmune regulator *AIRE* are associated with common organ-specific autoimmune diseases. Immunity 42:1185, 2015. [PubMed: 26084028]

Chapter 383: Disorders of Sex Development

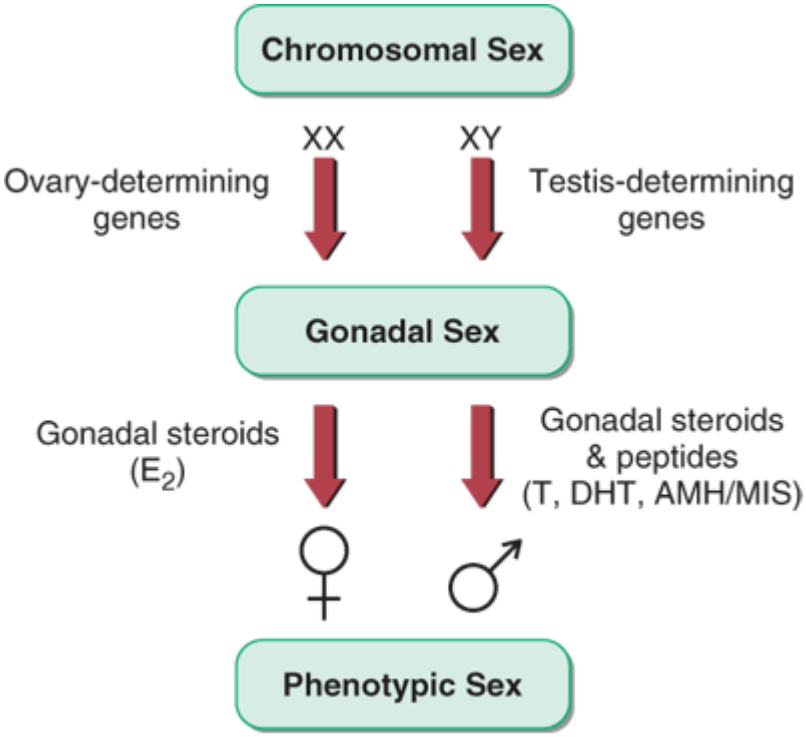
John C. Achermann; J. Larry Jameson

INTRODUCTION

Sex development begins in utero but continues into young adulthood with the achievement of sexual maturity and reproductive capability. The major determinants of sex development can be divided into three components: chromosomal sex, gonadal sex (sex determination), and phenotypic sex (sex differentiation) (Fig. 383-1). Variations at each of these stages can result in disorders (or differences) of sex development (DSDs) (Table 383-1). In the newborn period, ~1 in 4000 babies undergo investigation because of ambiguous (atypical) genitalia. Urgent assessment is indicated, because some causes such as congenital adrenal hyperplasia (CAH) can be associated with life-threatening adrenal crises. Support for the parents and clear communication about the diagnosis and management options are essential. The involvement of an experienced multidisciplinary team is important for counseling, planning appropriate investigations, and discussing long-term well-being. DSDs can also present at other ages and to a range of health professionals. Subtler forms of gonadal dysfunction (e.g., Klinefelter's syndrome [KS], Turner's syndrome [TS]) often are diagnosed later in life by internists. Because these conditions are associated with a variety of psychological, reproductive, and potential medical consequences, an open dialogue must be established between the patient and health care providers to ensure continuity and attention to these issues across the life span (Table 383-2). Support groups also have an important role to play for many of these conditions.

FIGURE 383-1

Sex development can be divided into three major components: chromosomal sex, gonadal sex, and phenotypic sex. DHT, dihydrotestosterone; MIS, müllerian-inhibiting substance also known as anti-müllerian hormone, AMH; T, testosterone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 383-1

Classification of Disorders of Sex Development (DSDs)

Sex Chromosome DSD	46,XY DSD (see Table 383-3)	46,XX DSD (see Table 383-4)
47,XXY (Klinefelter's syndrome and variants) 45,X (Turner's syndrome and variants) 45,X/46,XY mosaicism (mixed gonadal dysgenesis) 46,XX/46,XY (chimerism/mosaicism)	**Disorders of gonadal (testis) development** Complete or partial gonadal dysgenesis (e.g., *SRY, SOX9, SF1, WT1, DHH, GATA4/ZFPM2, MAP3K1*) Impaired fetal Leydig cell function (e.g., *SF1/NR5A1, CXorf6/MAMLD1, HHAT, SAMD9*) Ovotesticular DSD Testis regression **Disorders in androgen synthesis or action** Disorders of androgen biosynthesis LH receptor (*LHCGR*) Smith-Lemli-Opitz syndrome Steroidogenic acute regulatory (*StAR*) protein Cholesterol side-chain cleavage (*CYP11A1*) 3β-Hydroxysteroid dehydrogenase II (*HSD3B2*) 17α-Hydroxylase/17,20-lyase (*CYP17A1*) P450 oxidoreductase (*POR*) Cytochrome b5 (*CYB5A*) 17β-Hydroxysteroid dehydrogenase III (*HSD17B3*) 5α-Reductase II (*SRD5A2*) Aldo-keto reductase 1C2 (*AKR1C2*) Disorders of androgen action Androgen insensitivity syndrome Drugs and environmental modulators **Other** Syndromic associations of male genital development Persistent müllerian duct syndrome Vanishing testis syndrome Isolated hypospadias Congenital hypogonadotropic hypogonadism Cryptorchidism Environmental influences	**Disorders of gonadal (ovary) development** Gonadal dysgenesis Ovotesticular DSD Testicular DSD (e.g., *SRY+*, dup *SOX9, RSPO1, NR5A1*) **Androgen excess** Fetal 3β-Hydroxysteroid dehydrogenase II (*HSD3b2*) 21-Hydroxylase (*CYP21A2*) P450 oxidoreductase (*POR*) 11β-Hydroxylase (*CYP11B1*) Glucocorticoid receptor mutations Fetoplacental Aromatase deficiency (*CYP19*) Oxidoreductase deficiency (*POR*) Maternal Maternal virilizing tumors (e.g., luteomas) Androgenic drugs **Other** Syndromic associations (e.g., cloacal anomalies) Müllerian agenesis/hypoplasia (e.g., MRKH)

Sex Chromosome DSD	**46,XY DSD (see Table 383-3)**	**46,XX DSD (see Table 383-4)**
		Uterine abnormalities (e.g., MODY5) Vaginal atresia (e.g., McKusick-Kaufman) Labial adhesions

Abbreviations: MODY, maturity-onset diabetes of the young; MRKH, Mayer-Rokitansky-Kuster-Hauser syndrome.

Source: Modified from IA Hughes: Arch Dis Child 91:554, 2006.

TABLE 383-2

Presentation of DSD at Different Stages of Life

Presentation	Features	Professional	Examples
Prenatal	Karyotype-phenotype discordance	Obstetrician; fetal medicine	Many
Neonatal	Atypical genitalia	Obstetrician; neonatal medicine	Many
	Salt losing crisis	Pediatrician	CAH (CYP21)
Childhood	Hernia	Surgeon	CAIS
	Androgenization	Endocrinologist	CAH (CYP21, CYP11B1)
	Poor growth	Pediatrician	Turner's, 45,X/46,XY
	Associated features	Oncologist/nephrologist	Wilms' tumor
Puberty	Androgenization	Endocrinologist	17β-HSD, 5α-reductase, SF1
	Absent puberty	Endocrinologist	Gonadal dysgenesis, CAH (CYP17), Turner's
Post-puberty	Amenorrhea	Gynecologist	CAIS
Adult	Infertility	Andrologist	Klinefelter's, 45,X/46,XY, SF1

Abbreviations: CAH, congenital adrenal hyperplasia; CAIS, complete androgen insensitivity syndrome; 17β-HSD, 17β-hydroxysteroid dehydrogenase deficiency; SF1, steroidogenic factor 1 (NR5A1).

SEX DEVELOPMENT

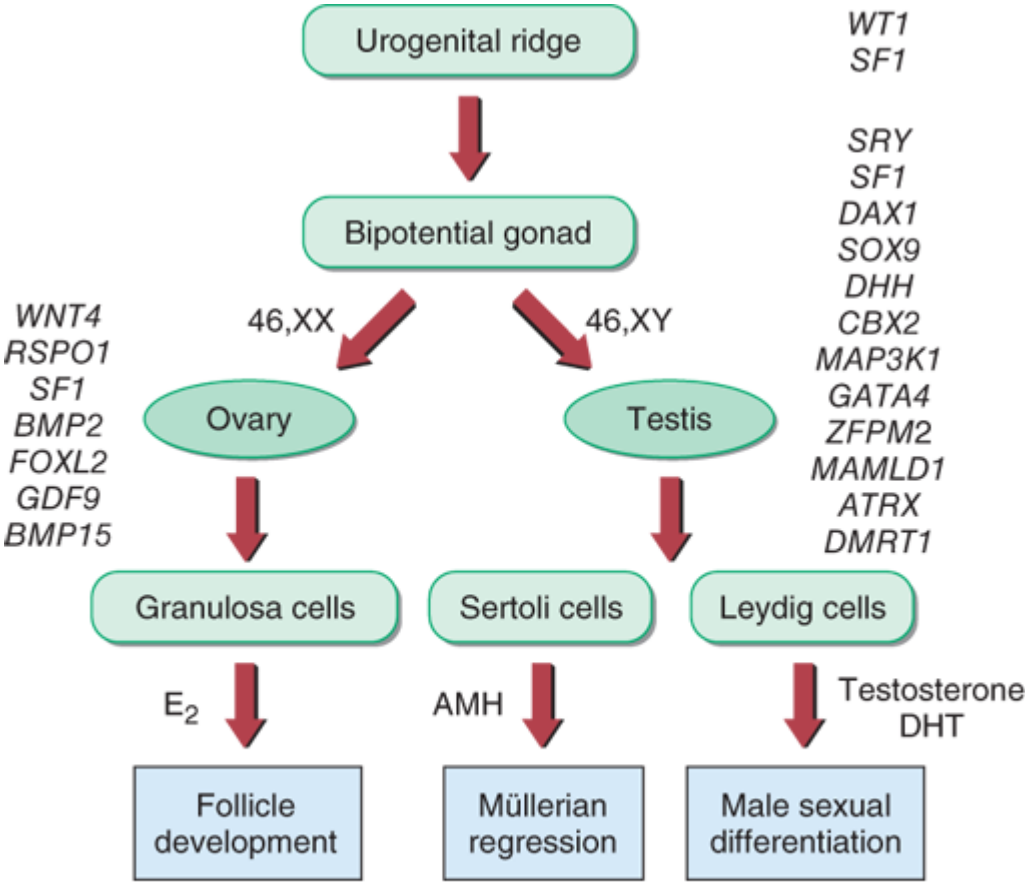
Chromosomal sex, defined by a karyotype, describes the X and/or Y chromosome complement (46,XY; 46,XX) that is established at the time of fertilization. The presence of a normal Y chromosome determines that testis development will occur even in the presence of multiple X chromosomes (e.g., 47,XXY or 48,XXXY). The loss of

an X chromosome impairs gonad development (45,X or 45,X/46,XY mosaicism). Fetuses with no X chromosome (45,Y) are not viable.

Gonadal sex refers to the histologic and functional characteristics of gonadal tissue as testis or ovary. The embryonic gonad is bipotential and can develop (from ~42 days after conception) into either a testis or an ovary, depending on which genes are expressed **(Fig. 383-2)**. Testis development is initiated by expression of the Y chromosome gene *SRY* (sex-determining region on the Y chromosome) that encodes an HMG box transcription factor. *SRY* is expressed transiently in cells destined to become Sertoli cells and serves as a pivotal switch to establish the testis lineage. Mutation of *SRY* prevents testis development in 46,XY individuals, whereas translocation of *SRY* in 46,XX individuals is sufficient to induce testis development and a male phenotype. Other genes are necessary to continue testis development. *SOX9* (*SRY*-related HMG-box gene 9) is upregulated by *SRY* in the developing testis but is suppressed in the ovary. *WT1* (Wilms' tumor–related gene 1) acts early in the genetic pathway and regulates the transcription of several genes, including *SFI* (*NR5A1*), *DAX1* (*NR0B1*), and *AMH* (encoding müllerian-inhibiting substance [MIS]). *SF1* encodes steroidogenic factor 1, a nuclear receptor that functions in cooperation with other transcription factors to regulate a large array of adrenal and gonadal genes, including *SOX9* and many genes involved in steroidogenesis. *SF1* mutations causing loss of function are found in ~10% of XY patients with gonadal dysgenesis and impaired androgenization. In contrast, duplication of a related gene *DAX1* also impairs testis development, revealing the exquisite sensitivity of the testis-determining pathway to gene dosage effects. In addition to the genes mentioned above, at least 30 other genes are also involved in gonad development (Fig. 383-2). These genes encode an array of signaling molecules and paracrine growth factors in addition to transcription factors.

FIGURE 383-2

The genetic regulation of gonadal development. AMH, anti-müllerian hormone (müllerian-inhibiting substance); *ATRX*, α-thalassemia, mental retardation on the X; BMP2 and 15, bone morphogenic factors 2 and 15; *CBX2*, chromobox homologue 2; *DAX1*, dosage sensitive sex-reversal, adrenal hypoplasia congenita on the X chromosome, gene 1 (also known as NR0B1); *DHH*, desert hedgehog; DHT, dihydrotestosterone; *DMRT 1,2*, doublesex MAB3-related transcription factor 1,2; *FOXL2*, forkhead transcription factor L2; *GATA4*, GATA binding protein 4; GDF9, growth differentiation factor 9; *MAMLD1*, mastermind-like domain containing 1; *MAP3K1*, mitogen-activated protein kinase kinase kinase 1; *RSPO1*, R-spondin 1; *SF1*, steroidogenic factor 1 (also known as NR5A1); *SOX9*, *SRY*-related HMG-box gene 9; *SRY*, sex-determining region on the Y chromosome; *WNT4*, wingless-type MMTV integration site 4; *WT1*, Wilms' tumor–related gene 1; ZFPM2, zinc finger protein, multitype 2.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although ovarian development once was considered a less active process, it is now clear that specific genes are expressed during the earliest stages of ovary development. Some of these factors may repress testis development (e.g., WNT4, R-spondin-1) (Fig. 383-2). Once the ovary has formed, additional factors are required for normal follicular development (e.g., follicle-stimulating hormone [FSH] receptor, *GDF9*). Steroidogenesis in the ovary requires the development of follicles that contain granulosa cells and theca cells surrounding the oocytes **(Chap. 385)**. Thus, there is relatively limited ovarian steroidogenesis until puberty.

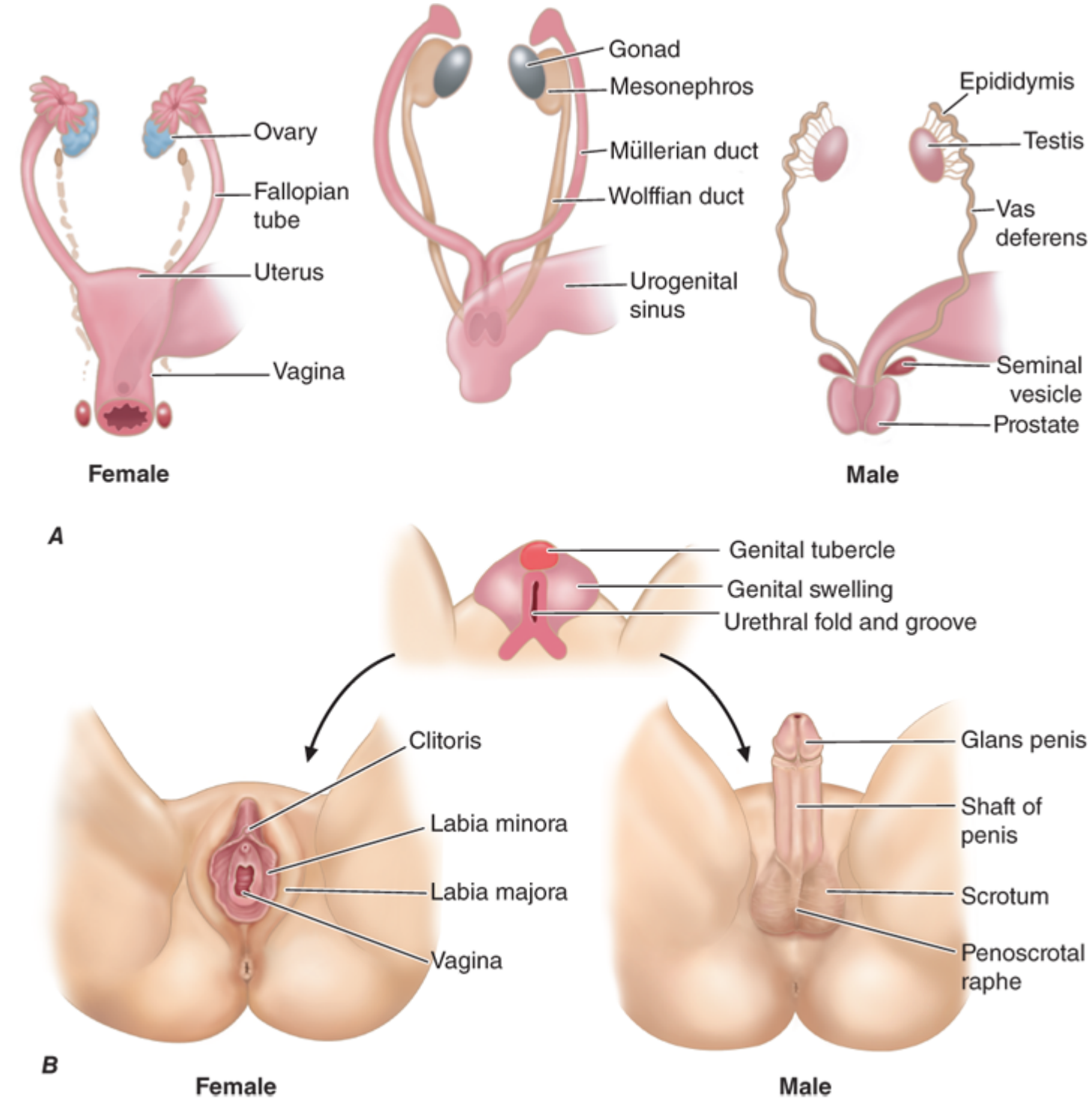
Germ cells also develop in a sex dimorphic manner. In the developing ovary, primordial germ cells (PGCs) proliferate and enter meiosis, whereas they proliferate and then undergo mitotic arrest in the developing testis. PGC entry into meiosis is initiated by retinoic acid. The developing testis produces high levels of CYP26B1, an enzyme that degrades retinoic acid, preventing PGC entry into meiosis. Approximately 7 million germ cells are present in the fetal ovary in the second trimester, and 1 million remain at birth. Only 400 are ovulated during a woman's reproductive life span **(Chap. 385)**.

Phenotypic sex refers to the structures of the external and internal genitalia and secondary sex characteristics. The developing testis releases anti-müllerian hormone (AMH; also known as müllerian-inhibiting substance [MIS]) from Sertoli cells and testosterone from Leydig cells. AMH acts through specific receptors to cause regression of the müllerian structures from 60–80 days after conception. At ~60–140 days after conception, testosterone supports the development of wolffian structures, including the epididymides, vasa deferentia, and seminal vesicles. Testosterone is the precursor for dihydrotestosterone (DHT), a potent

androgen that promotes development of the external genitalia, including the penis and scrotum (60–100 days, and thereafter) **(Fig. 383-3)**. The urogenital sinus develops into the prostate and prostatic urethra in the male and into the urethra and lower portion of the vagina in the female. The genital tubercle becomes the glans penis in the male and the clitoris in the female. The urogenital swellings form the scrotum or the labia majora, and the urethral folds fuse to form the shaft of the penis and the male urethra or the labia minora. In the female, wolffian ducts regress and the müllerian ducts form the fallopian tubes, uterus, and upper segment of the vagina. A female phenotype will develop in the absence of the gonad, but estrogen is needed for maturation of the uterus and breast at puberty.

FIGURE 383-3

Sex development. ***A.*** Internal urogenital tract. ***B.*** External genitalia. *(After E Braunwald et al [eds]: Harrison's Principles of Internal Medicine, 15th ed. New York, McGraw-Hill, 2001.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The prenatal hormone environment likely influences aspects of gender identity and behavior. This is an area of ongoing research and is beyond the scope of this chapter.

DISORDERS OF CHROMOSOMAL SEX

Variations in sex chromosome number and structure can present as DSDs (e.g., 45,X/46,XY). KS (47,XXY) and TS (45,X) do not usually present with genital ambiguity but are associated with gonadal dysfunction (Table

383-3).

TABLE 383-3

Clinical Features of Chromosomal Disorders of Sex Development (DSDs)

			Genitalia		
Disorder	**Common Chromosomal Complement**	**Gonad**	**External**	**Internal**	**Breast Development**
Klinefelter's syndrome	47,XXY or 46,XY/47,XXY	Hyalinized testes	Male	Male	Gynecomastia
	Clinical Features				
	Small testes, azoospermia, decreased facial and axillary hair, decreased libido, tall stature and increased leg length, decreased penile length, increased risk of breast tumors, thromboembolic disease, learning difficulties, speech delay and decreased verbal IQ, obesity, diabetes mellitus, metabolic syndrome, varicose veins, hypothyroidism, systemic lupus erythematosus, epilepsy				
Turner's syndrome	45,X or 45,X/46,XX	Streak gonad or immature ovary	Female	Hypoplastic female	Immature female
	Clinical Features				
	Infancy: lymphedema, web neck, shield chest, low-set hairline, cardiac defects and coarctation of the aorta, urinary tract malformations, and horseshoe kidney				
	Childhood: short stature, cubitus valgus, short neck, short fourth metacarpals, hypoplastic nails, micrognathia, scoliosis, otitis media and sensorineural hearing loss, ptosis and amblyopia, multiple nevi and keloid formation, autoimmune thyroid disease, visuospatial learning difficulties				
	Adulthood: pubertal failure and primary amenorrhea, hypertension, obesity, dyslipidemia, impaired glucose tolerance and insulin resistance, autoimmune thyroid disease, cardiovascular disease, aortic root dilation, osteoporosis, inflammatory bowel disease, chronic hepatic dysfunction, increased risk of colon cancer, hearing loss				
45,X/46,XY mosaicism	45,X/46,XY	Testis or streak gonad	Variable	Variable	Usually male

			Genitalia		
Disorder	**Common Chromosomal Complement**	**Gonad**	**External**	**Internal**	**Breast Development**
	Clinical Features				
	Short stature, increased risk of gonadal tumors, some Turner's syndrome features				
Ovotesticular DSD (true hermaphroditism)	46,XX/46,XY	Testis and ovary or ovotestis	Variable	Variable	Gynecomastia
	Clinical Features				
	Possible increased risk of gonadal tumors				

KLINEFELTER'S SYNDROME (47,XXY)

Pathophysiology

The classic form of KS (47,XXY) occurs after meiotic nondisjunction of the sex chromosomes during gametogenesis (40% during spermatogenesis, 60% during oogenesis). Mosaic forms of KS (46,XY/47,XXY) result from chromosomal mitotic nondisjunction within the zygote and occur in at least 10% of individuals with this condition. Other chromosomal variants of KS (e.g., 48,XXYY, 48,XXXY) are less common.

Clinical Features

KS is characterized by small testes, infertility, gynecomastia, tall stature/increased leg length, and hypogonadism in phenotypic males. It has an incidence of at least 1 in 1000 men, but ~75% of cases are not diagnosed. Of those who are diagnosed, only 10% are identified prepubertally, usually because of small genitalia or cryptorchidism. Others are diagnosed after puberty, usually based on low androgens and/or gynecomastia. Developmental delay, speech difficulties, and poor motor skills may be features but are variable, especially in adolescence. Later in life, body habitus or infertility leads to the diagnosis. Testes are small and firm (median length 2.5 cm [4 mL volume]; almost always <3.5 cm [12 mL]) and typically seem inappropriately small for the degree of androgenization. Biopsies are not usually necessary but typically reveal seminiferous tubule hyalinization and azoospermia. Other clinical features of KS are listed in Table 383-3. Plasma concentrations of FSH and luteinizing hormone (LH) are increased in most adults with 47,XXY, and plasma testosterone is decreased (50–75%), reflecting primary gonadal insufficiency. Estradiol is often

increased resulting in gynecomastia (Chap. 384). Patients with mosaic forms of KS have less severe clinical features, have larger testes, and sometimes achieve spontaneous fertility.

TREATMENT

TREATMENT

Klinefelter's Syndrome

Growth, endocrine function, and bone mineralization should be monitored, especially from adolescence. Educational and psychological support is important for many individuals with KS. Androgen supplementation improves virilization, libido, energy, hypofibrinolysis, and bone mineralization in men with low testosterone levels but may occasionally worsen gynecomastia (Chap. 384). Gynecomastia can be treated by surgical reduction if it causes concern (Chap. 384). Fertility has been achieved by using in vitro fertilization in men with oligospermia or with intracytoplasmic sperm injection (ICSI) after retrieval of spermatozoa by testicular sperm extraction techniques. In specialized centers, successful spermatozoa retrieval using this technique is possible in >50% of men with nonmosaic KS. Results may be better in younger men. After ICSI and embryo transfer, successful pregnancies can be achieved in ~50% of these cases. The risk of transmitting chromosomal anomalies needs to be considered, although this outcome is much less common than originally predicted, and the role of preimplantation screening is debated. Long-term monitoring of men with KS is important given the increased risk of breast cancer, cardiovascular disease, metabolic syndrome, and autoimmune disorders. Because most men with KS are never diagnosed, it is important that all internists consider this diagnosis in men with these features who might be seeking medical advice for other conditions.

TURNER'S SYNDROME (GONADAL DYSGENESIS; 45,X)

Pathophysiology

Approximately one-half of women with TS have a 45,X karyotype, about 20% have 45,X/46,XX mosaicism, and the remainder have structural abnormalities of the X chromosome such as X fragments, isochromosomes, or rings. The clinical features of TS result from haploinsufficiency of multiple X chromosomal genes (e.g., short stature homeobox, *SHOX*). However, imprinted genes also may be affected when the inherited X has different parental origins.

Clinical Features

TS is characterized by bilateral streak gonads, primary amenorrhea, short stature, and other phenotypic features (Table 383-3). It affects ~1 in 2500 women and is diagnosed at different ages. Prenatally, a diagnosis of TS usually is made incidentally after chorionic villus sampling or amniocentesis for unrelated reasons such as advanced maternal age. Prenatal ultrasound findings include increased nuchal translucency. The postnatal diagnosis of TS should be considered in female neonates or infants with lymphedema, nuchal folds, low hairline, or left-sided cardiac defects and in girls with unexplained growth failure or pubertal delay.

Although limited spontaneous pubertal development occurs in up to 30% of girls with TS (10%, 45,X; 30–40%, 45,X/46,XX) and ~2% have menarche, the vast majority of women with TS develop complete ovarian insufficiency. Therefore, this diagnosis should be considered in all women who present with primary or secondary amenorrhea and elevated gonadotropin levels.

TREATMENT

TREATMENT

Turner's Syndrome

The management of girls and women with TS requires a multidisciplinary approach because many different organ systems can be affected. Detailed cardiac and renal evaluation should be performed at the time of diagnosis. Individuals with congenital heart defects (CHDs) (30%) (bicuspid aortic valve, 30–50%; coarctation of the aorta, 30%; aortic root dilation, 5%) require long-term follow-up by an experienced cardiologist, antibiotic prophylaxis for dental or surgical procedures, and serial magnetic resonance imaging (MRI) of aortic root dimensions, because progressive aortic root dilation is associated with increased risk of aortic dissection. Individuals found to have congenital renal and urinary tract malformations (30%) are at risk for urinary tract infections, hypertension, and nephrocalcinosis. Hypertension can occur independently of cardiac and renal malformations and should be monitored and treated as in other patients with essential hypertension. Clitoral enlargement or other evidence of virilization suggests the presence of covert Y chromosomal material and is associated with increased risk of gonadoblastoma. Regular assessment of thyroid function, weight, dentition, hearing, speech, vision, and educational issues should be performed during childhood. Otitis media and middle-ear disease are prevalent in childhood (50–85%), and sensorineural hearing loss becomes progressively common with age (70–90%). Autoimmune hypothyroidism (15–30%) can occur in childhood but has a mean age of onset in the third decade. Counseling about long-term growth and fertility issues should be provided. Patient support groups are active throughout the world and can play an invaluable role.

Short stature can be an issue for some girls because untreated final height rarely exceeds 150 cm in nonmosaic 45,X TS. Recombinant growth hormone stimulates growth rate in children with TS and is occasionally combined with low doses of the nonaromatizable anabolic steroid oxandrolone (up to 0.05 mg/kg per day) in an older child (>9 years). However, final height increments are often about 5–10 cm, and individualization of treatment response to regimens may be beneficial. Girls with evidence of ovarian insufficiency require estrogen replacement to induce breast and uterine development, support growth, and maintain bone mineralization. Most physicians now initiate low-dose estrogen therapy (one-tenth to one-eighth of the adult replacement dose) to induce puberty at an age-appropriate time (~11 years). Doses of estrogen are increased gradually to allow development over a 2- to 4-year period. Progestins are added later to regulate withdrawal bleeds. Some women with TS have achieved successful pregnancy after ovum donation and in vitro fertilization but the risks of cardiac complications are high, and expert counseling and management are needed. Long-term follow-up of women with TS involves careful surveillance of sex hormone replacement and reproductive function, bone mineralization, cardiac function and aortic root

dimensions, blood pressure, weight and glucose tolerance, hepatic and lipid profiles, thyroid function, and hearing. This service is provided by a dedicated TS clinic in some centers.

45,X/46,XY MOSAICISM (MIXED GONADAL DYSGENESIS)

The phenotype of individuals with 45,X/46,XY mosaicism (sometimes called *mixed gonadal dysgenesis*) can vary considerably. Some have a predominantly female phenotype with somatic features of TS, streak gonads, and müllerian structures, and are managed as TS with a Y chromosome. Most 45,X/46,XY individuals have a male phenotype and testes, and the diagnosis is made incidentally after amniocentesis or during investigation of infertility. In practice, most newborns referred for assessment have atypical genitalia and variable somatic features. Management is complex and needs to be individualized. A female sex-of-rearing is often assigned if uterine structures are present, gonads are intraabdominal, and the phallus is very small. In such situations, gonadectomy usually is considered to prevent further androgen secretion at puberty and prevent risk of gonadoblastoma (up to 25%). Individuals raised as males usually have reconstructive surgery for hypospadias and removal of dysgenetic or streak gonads if the gonads cannot be brought down into the scrotum. Scrotal testes can be preserved but require regular examination for tumor development and sonography at the time of puberty. Biopsy for carcinoma in situ is recommended in adolescence, and testosterone supplementation may be required to support androgenization in puberty or if low testosterone is detected in adulthood. Height potential is usually reduced; some children receive recombinant growth hormone using TS protocols. Screening for cardiac, renal, and other TS features should be considered, and psychological support offered for the family and young person.

OVOTESTICULAR DSD

Ovotesticular DSD (formerly called *true hermaphroditism*) occurs when both an ovary and a testis—or when an ovotestis—are found in one individual. Most individuals with this diagnosis have a 46,XX karyotype, especially in sub-Saharan Africa, and present with ambiguous genitalia at birth or with breast development and phallic development at puberty. A 46,XX/46,XY chimeric karyotype is less common and has a variable phenotype.

DISORDERS OF GONADAL AND PHENOTYPIC SEX

Disorders of gonadal and phenotypic sex can result in reduced androgen production or action in individuals with a 46,XY karyotype (46,XY DSD), or excess androgen production in individuals with a 46,XX karyotype (46,XX DSD) (Table 383-1). These conditions cover a spectrum of phenotypes ranging from phenotypic females with a Y-chromosome to phenotypic males with a 46,XX karyotype to individuals with atypical genitalia. Karyotype is a useful starting investigation for diagnosis, but does not define an individual's gender.

46,XY DSD

Underandrogenization of the 46,XY fetus (formerly called *male pseudohermaphroditism*) reflects defects in androgen production or action. It can result from disorders of testis development, defects of androgen synthesis, or resistance to testosterone and DHT (Table 383-1).

Disorders of Testis Development

TESTICULAR DYSGENESIS

Pure (or *complete*) *gonadal dysgenesis* (*Swyer's syndrome*) is associated with streak gonads, müllerian structures (due to insufficient AMH/MIS secretion), and a complete absence of androgenization. Phenotypic females with this condition often present because of absent pubertal development and are found to have a 46,XY karyotype. Serum sex steroids, AMH/MIS, and inhibin B are low, and LH and FSH are elevated. Patients with *partial gonadal dysgenesis* (*dysgenetic testes*) may produce enough MIS to regress the uterus and sufficient testosterone for partial androgenization, and therefore usually present in the newborn period with atypical genitalia. Gonadal dysgenesis can result from mutations or deletions of testis-promoting genes (*WT1*, *CBX2*, *SF1*, *SRY*, *SOX9*, *MAP3K1*, *DHH*, *GATA4/ZFPM2*, *ATRX*, *ARX*, *DMRT*) or duplication of chromosomal loci containing "antitestis" genes (e.g., *WNT4/RSPO1*, *DAX1*) **(Table 383-4)**. Among these, deletions or mutations of *SRY* and heterozygous mutations of *SF1* (*NR5A1*) appear to be most common but still account collectively for <25% of cases. Associated clinical features may be present, reflecting additional functional roles for these genes. For example, renal dysfunction occurs in patients with specific *WT1* mutations (Denys-Drash and Frasier's syndromes), primary adrenal failure occurs in some patients with *SF1* mutations, and severe cartilage abnormalities (campomelic dysplasia) are the predominant clinical feature of *SOX9* mutations. A family history of DSD, infertility, or early menopause is important because mutations in *SF1/NR5A1* can be inherited from a mother in a sex-limited dominant manner (which can mimic X-linked inheritance). In some situations, a woman may later develop primary ovarian insufficiency because of the effect of *SF1* on the ovary. Intraabdominal dysgenetic testes should be removed to prevent malignancy, and estrogens can be used to induce secondary sex characteristics and uterine development in 46,XY individuals raised as females, if person feels that female gender is appropriate for them. *Absent (vanishing) testis syndrome* (*bilateral anorchia*) reflects regression of the testis during development. The etiology is unknown, but the absence of müllerian structures indicates adequate secretion of AMH early in utero. Usually, androgenization of the external genitalia is normal. These individuals can be offered testicular prostheses and should receive androgen replacement in adolescence.

TABLE 383-4

Selected Genetic Causes of 46,XY Disorders of Sex Development (DSDs)

Gene	Inheritance	Gonad	Uterus	External Genitalia	Associated Features
Disorders of Testis Development					
WT1	AD	Dysgenetic testis	+/–	Female or ambiguous	Wilms' tumor, renal abnormalities, gonadal tumors (WAGR, Denys-Drash and Frasier's syndromes)
CBX2	AD	Ovary	+	Female	
SF1	AR/AD (SL)	Dysgenetic testis/Leydig dysfunction	+/–	Female or ambiguous	Primary adrenal failure; primary ovarian insufficiency in female (46,XX) relatives
SRY	Y	Dysgenetic testis or ovotestis	+/–	Female or ambiguous	
SOX9	AD	Dysgenetic testis or ovotestis	+/–	Female or ambiguous	Campomelic dysplasia
MAP3K1	AD (SL)	Dysgenetic testis	+/–	Female or ambiguous	
GATA4	AD	Dysgenetic testis	–	Female, ambiguous or male	Congenital heart disease
ZFPM2	AD	Dysgenetic testis		Female, ambiguous or male	
ARX	X	Dysgenetic testis	–	Male or ambiguous	Developmental delay; X-linked lissencephaly

Gene	Inheritance	Gonad	Uterus	External Genitalia	Associated Features
SAMD9	AD	Dysgenetic testis/Leydig dysfunction	+	Female, ambiguous or male	Myelodysplasia, infection, growth restriction, adrenal hypoplasia, enteropathy
DHH	AR	Dysgenetic testis/Leydig dysfunction	+	Female	Minifascicular neuropathy
MAMLD1	X	Dysgenetic testis/Leydig dysfunction	–	Hypospadias	
DAX1	dupXp21	Dysgenetic testis	+/–	Female or ambiguous	
WNT4/RSPO1	dup1p35	Dysgenetic testis	+	Ambiguous	
Disorders of Androgen Synthesis					
LHR	AR	Testis	–	Female, ambiguous or micropenis	Leydig cell hypoplasia
DHCR7	AR	Testis	–	Variable	Smith-Lemli-Opitz syndrome: coarse facies, second-third toe syndactyly, failure to thrive, developmental delay, cardiac and visceral abnormalities
StAR	AR	Testis	–	Female or ambiguous	Congenital lipoid adrenal hyperplasia (primary adrenal failure)
CYP11A1	AR	Testis	–	Ambiguous	Primary adrenal failure

Gene	Inheritance	Gonad	Uterus	External Genitalia	Associated Features
HSD3B2	AR	Testis	–	Ambiguous	CAH, primary adrenal failure ± salt loss, partial androgenization due to ↑ DHEA
CYP17	AR	Testis	–	Female or ambiguous	CAH, hypertension due to ↑ corticosterone and 11-deoxycorticosterone, except in isolated 17,20-lyase deficiency
CYB5A	AR	Testis	–	Ambiguous	Apparent isolated 17,20-lyase deficiency; methemoglobinemia
POR	AR	Testis	–	Ambiguous or male	Mixed features of 21-hydroxylase deficiency and 17α-hydroxylase/17,20-lyase deficiency, sometimes associated with Antley-Bixler craniosynostosis
HSD17B3	AR	Testis	–	Female or ambiguous	Partial androgenization at puberty, ↑ androstenedione-to-testosterone ratio
SRD5A2	AR	Testis	–	Ambiguous or micropenis	Partial androgenization at puberty, ↑ testosterone-to-dihydrotestosterone ratio
AKR1C2 (*AKR1C4*)	AR	Testis	–	Female or ambiguous	Decreased fetal DHT production
Disorders of Androgen Action					

Gene	Inheritance	Gonad	Uterus	External Genitalia	Associated Features
Androgen receptor	X	Testis	–	Female, ambiguous, micropenis or normal male	Phenotypic spectrum from complete androgen insensitivity syndrome (female external genitalia) and partial androgen insensitivity (ambiguous) to normal male genitalia and infertility

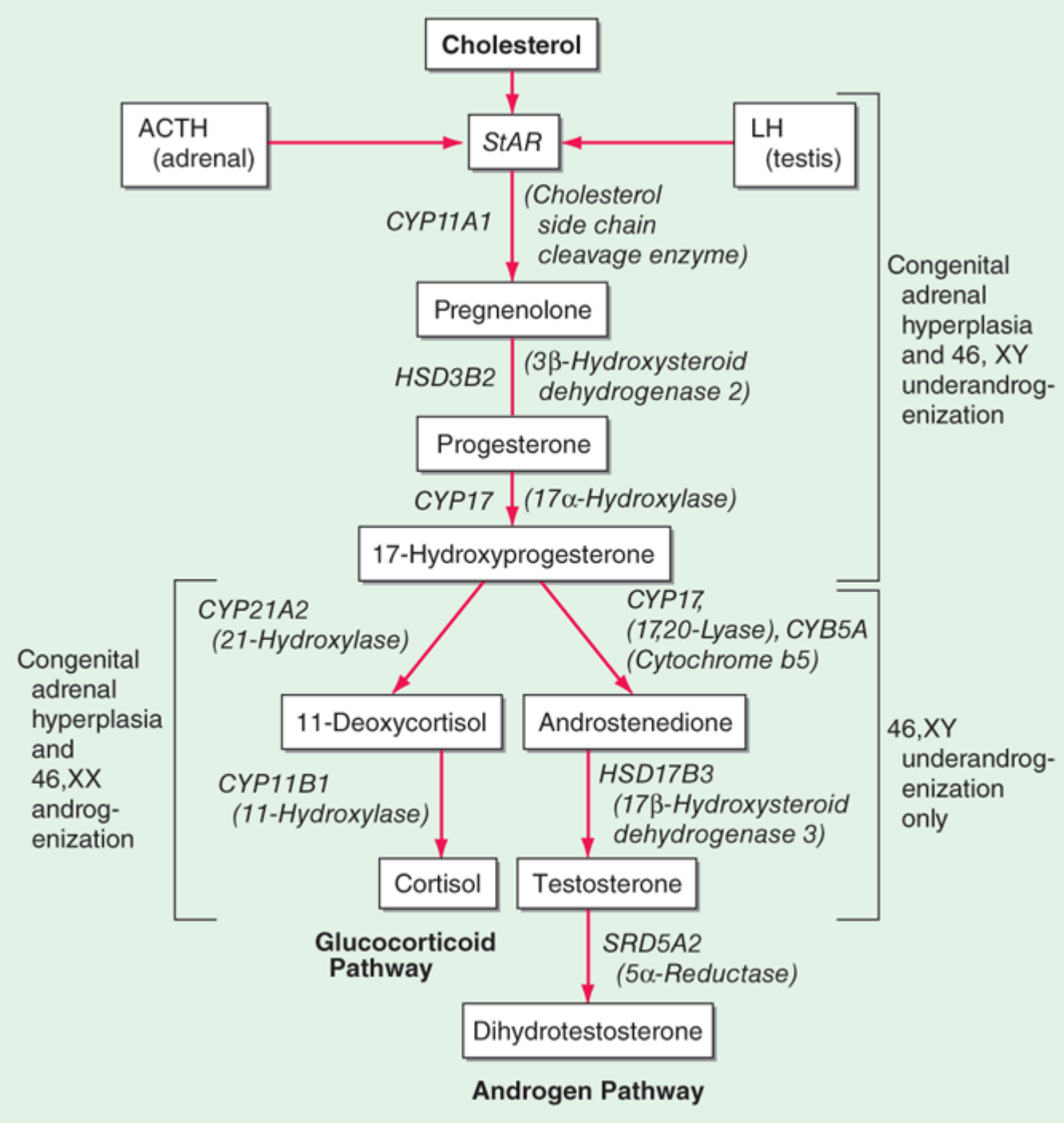
Abbreviations: AD, autosomal dominant; *AKR1C2*, aldo-keto reductase family 1 member 2; AR, autosomal recessive; *ARX*, aristaless related homeobox, X-linked; CAH, congenital adrenal hyperplasia; *CBX2*, chromobox homologue 2; *CYB5A*, cytochrome b5 POR, P450 oxidoreductase; *CYP11A1*, P450 cholesterol side-chain cleavage; *CYP17*, 17α-hydroxylase and 17,20-lyase; *DAX1*, dosage sensitive sex-reversal, adrenal hypoplasia congenita on the X chromosome, gene 1; DHEA, dehydroepiandrosterone; *DHCR7*, sterol 7 δ reductase; *DHH*, desert hedgehog; *GATA4*, GATA binding protein 4; *HSD17B3,* 17β-hydroxysteroid dehydrogenase type 3; *HSD3B2*, 3β-hydroxysteroid dehydrogenase type 2; *LHR*, LH receptor; *MAP3K1*, mitogen-activated protein kinase kinase kinase 1; *SF1,* steroidogenic factor 1; SL, sex-limited; *SOX9, SRY*-related HMG-box gene 9; *SRD5A2*, 5α-reductase type 2; *SRY,* sex-related gene on the Y chromosome; *StAR*, steroidogenic acute regulatory protein; WAGR, Wilms' tumor, aniridia, genitourinary anomalies, and mental retardation; *WNT4*, wingless-type mouse mammary tumor virus integration site, 4; *WT1,* Wilms' tumor–related gene 1; ZFPM2, zinc finger protein, multitype 2.

Disorders of Androgen Synthesis

Defects in the pathway that regulates androgen synthesis **(Fig. 383-4)** cause underandrogenization of the 46,XY fetus (Table 383-1). Müllerian regression is unaffected because Sertoli cell function is preserved. These conditions can present with a spectrum of genital appearances, ranging from female-typical external genitalia or clitoromegaly in some individuals to penoscrotal hypospadias or a small phallus in others.

FIGURE 383-4

Simplified overview of glucocorticoid and androgen synthesis pathways. Defects in *CYP21A2* and *CYP11B1* shunt steroid precursors into the androgen pathway and cause androgenization of the 46,XX fetus. Testosterone is synthesized in the testicular Leydig cells and converted to dihydrotestosterone peripherally. Defects in enzymes involved in androgen synthesis result in underandrogenization of the 46,XY fetus. *StAR*, steroidogenic acute regulatory protein. *(After E Braunwald et al [eds]: Harrison's Principles of Internal Medicine, 15th ed. New York, McGraw-Hill, 2001.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

LH RECEPTOR

Mutations in the LH receptor (LHCGR) cause Leydig cell hypoplasia and androgen deficiency, due to impaired actions of human chorionic gonadotropin in utero and LH late in gestation and during the neonatal period. As a result, testosterone and DHT synthesis are reduced.

STEROIDOGENIC ENZYME PATHWAYS

Mutations in *steroidogenic acute regulatory protein* (*StAR*) and *CYP11A1* affect both adrenal and gonadal steroidogenesis (Fig. 383-4) **(Chap. 379)**. Affected individuals (46,XY) usually have severe early-onset salt-losing adrenal failure and a female phenotype, although later-onset milder variants have been reported. Defects in *3β-hydroxysteroid dehydrogenase type 2* (*HSD3*β2) also cause adrenal insufficiency in severe

cases, but the accumulation of dehydroepiandrosterone (DHEA) has a mild androgenizing effect, resulting in ambiguous genitalia or hypospadias. Salt loss occurs in many but not all children. Patients with CAH due to *17α-hydroxylase (CYP17) deficiency* have variable underandrogenization and develop hypertension and hypokalemia due to the potent salt-retaining effects of corticosterone and 11-deoxycorticosterone. Patients with complete loss of 17α-hydroxylase function often present as phenotypic females who do not enter puberty and are found to have inguinal testes and hypertension in adolescence. Some mutations in *CYP17* selectively impair 17,20-lyase activity without altering 17α-hydroxylase activity, leading to underandrogenization without mineralocorticoid excess and hypertension. Disruption of the coenzyme, *cytochrome b5* (*CYB5A*), can present similarly, and methemoglobinemia is usually present. Mutations in *P450 oxidoreductase* (*POR*) affect multiple steroidogenic enzymes, leading to reduced androgen production and a biochemical pattern of apparent combined 21-hydroxylase and 17α-hydroxylase deficiency, sometimes with skeletal abnormalities (Antley-Bixler craniosynostosis). Defects in *17β-hydroxysteroid dehydrogenase type 3* (*HSD17β3*) and *5α-reductase type 2* (*SRD5A2*) interfere with the synthesis of testosterone and DHT, respectively. These conditions are characterized by minimal or absent androgenization in utero, but some phallic development can occur during adolescence due to the action of other enzyme isoforms. Individuals with 5α-reductase type 2 deficiency have normal wolffian structures and usually do not develop breast tissue. At puberty, the increase in testosterone induces muscle mass and other virilizing features despite DHT deficiency. Some individuals change gender from female to male at puberty. Thus, the management of this disorder requires expert support. DHT cream can improve prepubertal phallic growth in patients raised as male. Gonadectomy before adolescence and estrogen replacement at puberty can be considered in individuals raised as females who feel they have a female gender identity. Disruption of alternative pathways to fetal DHT production might also present with 46,XY DSD (*AKR1C2/AKR1C4*).

Disorders of Androgen Action

ANDROGEN INSENSITIVITY SYNDROME

Mutations in the androgen receptor cause resistance to androgen (testosterone, DHT) action or the *androgen insensitivity syndrome* (*AIS*). AIS is a spectrum of disorders that affects at least 1 in 100,000 46,XY individuals. Because the androgen receptor is X-linked, only 46,XY offspring are affected if the mother is a carrier of a mutation. XY individuals with *complete AIS* (formerly called *testicular feminization syndrome*) have a female phenotype, normal breast development (due to aromatization of testosterone), a short vagina but no uterus (because MIS production is normal), scanty pubic and axillary hair, and a female gender identity and sex role behavior. Gonadotropins and testosterone levels can be low, normal, or elevated, depending on the degree of androgen resistance and the contribution of estradiol to feedback inhibition of the hypothalamic-pituitary-gonadal axis. AMH/MIS levels in childhood are normal or high. CAIS sometimes presents as inguinal hernias (containing testes) in childhood or more often with primary amenorrhea in late adolescence. Because there is a low risk of malignancy, gonadectomy has been recommended for girls diagnosed in childhood, with estrogen replacement prescribed at puberty. Alternatively, the gonads can be left in situ until breast development is complete and then removed to avoid tumor risk. Some adults with complete AIS decline gonadectomy, but should be counseled about the risk of malignancy, especially because early

detection of premalignant changes by imaging or biomarkers is currently not possible. The use of graded dilators in adolescence is usually sufficient to dilate the vagina for sexual intercourse.

Partial AIS (Reifenstein's syndrome) results from androgen receptor mutations that maintain residual function. Patients often present in infancy with penoscrotal hypospadias and undescended testes and with gynecomastia at the time of puberty. Those individuals raised as males usually have hypospadias repair in childhood and may request breast reduction in adolescence. Some boys enter puberty spontaneously. High-dose testosterone has been given to support development if puberty does not progress, but long-term data are limited. Other patients present with clitoral enlargement and labial fusion and may be raised as females. The surgical and psychosexual management of these patients is complex and requires active involvement of the parents and the patient during the appropriate stages of development. *Azoospermia* and male-factor infertility also have been described in association with mild loss-of-function mutations in the androgen receptor.

OTHER DISORDERS AFFECTING 46,XY MALES

Persistent müllerian duct syndrome is the presence of a uterus in an otherwise phenotypic male. This condition can result from mutations in AMH or its receptor (AMHR2). The uterus may be removed, but only if damage to the vasa deferentia and blood supply can be avoided. *Isolated hypospadias* occurs in ~1 in 250 males. Most cases are idiopathic, although evidence of penoscrotal hypospadias, poor phallic development, and/or bilateral cryptorchidism requires investigation for an underlying DSD (e.g., partial gonadal dysgenesis, mild defect in testosterone action, or even severe forms of 46,XX CAH). Unilateral undescended testes (cryptorchidism) affect >3% of boys at birth. Orchidopexy should be considered if the testis has not descended by 6–9 months of age. Bilateral cryptorchidism occurs less frequently and should raise suspicion of gonadotropin deficiency or DSD. *Syndromic associations* and *intrauterine growth retardation* also occur relatively frequently in association with impaired testicular function or target tissue responsiveness, but the underlying etiology of many of these conditions is unknown.

46,XX DSD

Inappropriate androgenization of the 46,XX fetus (formerly called *female pseudohermaphroditism*) occurs when the gonad (ovary) contains androgen-secreting testicular tissue or after increased androgen exposure, which is usually adrenal in origin (Table 383-1).

46,XX Testicular/Ovotesticular DSD

Testicular tissue can develop in 46,XX testicular DSD (46,XX males) after translocation of *SRY*, duplication of *SOX9*, or defects in *RSPO1* or *SF1*/*NR5A1* **(Table 383-5)**.

TABLE 383-5

Selected Genetic Causes of 46,XX Disorders of Sex Development (DSDs)

Gene	Inheritance	Gonad	Uterus	External Genitalia	Associated Features
Testicular/Ovotesticular DSD					
SRY	Translocation	Testis or ovotestis	–	Male or ambiguous	
SOX9	dup17q24	Unknown	–	Male or ambiguous	
SF1 (codon 92)	AD	Testis or ovotestis	±	Male or ambiguous	
RSPO1	AR	Testis or ovotestis	±	Male or ambiguous	Palmar plantar hyperkeratosis, squamous cell skin carcinoma
WNT4	AR	Testis or ovotestis	–	Male or ambiguous	SERKAL syndrome (renal dysgenesis, adrenal and lung hypoplasia)
Increased Androgen Synthesis					
HSD3B2	AR	Ovary	+	Clitoromegaly	CAH, primary adrenal failure, mild androgenization due to ↑ DHEA
CYP21A2	AR	Ovary	+	Ambiguous	CAH, phenotypic spectrum from severe salt-losing forms associated with adrenal failure to simple virilizing forms with compensated adrenal function, ↑ 17-hydroxyprogesterone

Gene	Inheritance	Gonad	Uterus	External Genitalia	Associated Features
POR	AR	Ovary	+	Ambiguous or female	Mixed features of 21-hydroxylase deficiency and 17α-hydroxylase/17,20-lyase deficiency, sometimes associated with Antley-Bixler craniosynostosis
CYP11B1	AR	Ovary	+	Ambiguous	CAH, hypertension due to ↑ 11-deoxycortisol and 11-deoxycorticosterone
CYP19	AR	Ovary	+	Ambiguous	Maternal virilization during pregnancy, absent breast development at puberty
Glucocorticoid receptor	AR	Ovary	+	Ambiguous	↑ ACTH, 17-hydroxyprogesterone and cortisol; failure of dexamethasone suppression

Abbreviations: ACTH, adrenocorticotropin; AR, autosomal recessive; CAH, congenital adrenal hyperplasia; *CYP11B1*, 11β-hydroxylase; *CYP19*, aromatase; *CYP21A2*, 21-hydroxylase; DHEA, dehydroepiandrosterone; *HSD3B2*, 3β-hydroxysteroid dehydrogenase type 2; *POR*, P450 oxidoreductase; *RSPO1*, R-spondin 1; *SF1,* steroidogenic factor 1; *SOX9*, *SRY*-related HMG-box gene 9; *SRY*, sex-related gene on the Y chromosome.

Increased Androgen Exposure

21-HYDROXYLASE DEFICIENCY (CONGENITAL ADRENAL HYPERPLASIA)

The *classic form* of 21-hydroxylase deficiency (21-OHD) is the most common cause of CAH **(Chap. 379)**. It has an incidence between 1 in 10,000 and 1 in 15,000 and is the most common cause of androgenization in chromosomal 46,XX females (Table 383-5). Affected individuals are homozygous or compound heterozygous for severe mutations in the enzyme 21-hydroxylase (*CYP21A2*). This mutation causes a block in adrenal glucocorticoid and mineralocorticoid synthesis, increasing 17-hydroxyprogesterone and shunting steroid precursors into the androgen synthesis pathway (Fig. 383-4). Glucocorticoid insufficiency causes a compensatory elevation of adrenocorticotropin (ACTH), resulting in adrenal hyperplasia and additional synthesis of steroid precursors proximal to the enzymatic block. Increased androgen synthesis in utero causes androgenization of the 46,XX fetus in the first trimester. Ambiguous genitalia are seen at birth, with

varying degrees of clitoral enlargement and labial fusion. Excess androgen production causes gonadotropin-independent precocious puberty in males with 21-OHD.

The *salt-wasting* form of 21-OHD results from severe combined glucocorticoid and mineralocorticoid deficiency. A salt-wasting crisis usually manifests between 5 and 21 days of life and is a potentially life-threatening event that requires urgent fluid resuscitation and steroid treatment. Thus, a diagnosis of 21-OHD should be considered in any baby with atypical genitalia with bilateral nonpalpable gonads. Males (46,XY) with 21-OHD have no genital abnormalities at birth but are equally susceptible to adrenal insufficiency and salt-losing crises.

Females with the *classic simple virilizing* form of 21-OHD also present with genital ambiguity. They have impaired cortisol biosynthesis but do not develop salt loss. Patients with *nonclassic 21-OHD* produce normal amounts of cortisol and aldosterone but at the expense of producing excess androgens. Hirsutism (60%), oligomenorrhea (50%), and acne (30%) are the most common presenting features. This is one of the most common recessive disorders in humans, with an incidence as high as 1 in 100 to 500 in many populations and 1 in 27 in Ashkenazi Jews of Eastern European origin.

Biochemical features of acute salt-wasting 21-OHD are hyponatremia, hyperkalemia, hypoglycemia, inappropriately low cortisol and aldosterone, and elevated 17-hydroxyprogesterone, ACTH, and plasma renin activity. Presymptomatic diagnosis of classic 21-OHD is now made by neonatal screening tests for increased 17-hydroxyprogesterone in many centers. In most cases, 17-hydroxyprogesterone is markedly increased. In adults, ACTH stimulation (0.25 mg of cosyntropin IV) with assays for 17-hydroxyprogesterone at 0 and 30 min can be useful for detecting nonclassic 21-OHD and heterozygotes **(Chap. 379)**.

TREATMENT

TREATMENT

Congenital Adrenal Hyperplasia

Acute salt-wasting crises require fluid resuscitation, IV hydrocortisone, and correction of hypoglycemia. Once the patient is stabilized, glucocorticoids must be given to correct the cortisol insufficiency and suppress ACTH stimulation, thereby preventing further virilization, rapid skeletal maturation, and the development of polycystic ovaries. Typically, hydrocortisone (10–15 mg/m^2 per day in three divided doses) is used in childhood with a goal of partially suppressing 17-hydroxyprogesterone. The aim of treatment is to use the lowest glucocorticoid dose that adequately suppresses adrenal androgen production without causing signs of glucocorticoid excess such as impaired growth and obesity. Salt-wasting conditions are treated with mineralocorticoid replacement. Infants usually need salt supplements up to the first year of life. Plasma renin activity and electrolytes are used to monitor mineralocorticoid replacement, remembering that normal ranges are age-dependent. Some patients with simple virilizing 21-OHD also benefit from mineralocorticoid supplements. Parents and patients should be educated about the need for increased doses of steroids during sickness, and patients should carry medic alert systems.

Steroid treatment for older adolescents and adults varies depending on lifestyle, age, and factors such as a desire to optimize fertility. Hydrocortisone remains a useful approach, but treatment with prednisolone at night may provide more complete ACTH suppression. Steroid doses should be adjusted to individual requirements because overtreatment can result in iatrogenic Cushing's-like features, including weight gain, insulin resistance, hypertension, and osteopenia. Because it is long acting, dexamethasone given at night is useful for ACTH suppression but is often associated with more side effects, making hydrocortisone or prednisolone preferable for most patients. Androstenedione and testosterone may be useful measurements of long-term control, with less fluctuation than 17-hydroxyprogesterone. Mineralocorticoid requirements often decrease in adulthood, and doses should be reassessed and reduced to avoid hypertension in adults. In very severe cases, adrenalectomy has been advocated but incurs the risks of surgery and total adrenal insufficiency. Newer approaches to treatment under investigation include more physiological cortisol replacement strategies and specifically targeting androgen excess.

Girls with significant genital androgenization due to classic 21-OHD usually undergo vaginal reconstruction and sometimes clitoral reduction (maintaining the glans and nerve supply), but the optimal timing of these procedures is debated, as is the need for the individual to be able to consent. There is a higher threshold for undertaking clitoral surgery in some centers because long-term sensation and ability to achieve orgasm can be affected, but the long-term results of newer techniques are not yet known. Full information about all options should be provided, with appropriate support. Good endocrine control to reduce testosterone levels is also important. If surgery is performed in infancy, surgical revision or regular vaginal dilatation may be needed in adolescence or adulthood, and long-term psychological support and psychosexual counseling may be appropriate. Women with 21-OHD frequently develop polycystic ovaries and have reduced fertility, especially when control is poor. Fecundity is achieved in 60–90% of women with good metabolic control, but ovulation induction (or even adrenalectomy) may be required. Dexamethasone should be avoided in pregnancy. Men with poorly controlled 21-OHD may develop testicular adrenal rests and are at risk for reduced fertility. Prenatal treatment of 21-OHD by the administration of dexamethasone to mothers is still under evaluation. Treatment must be started early in pregnancy, but has the risk that both affected and nonaffected fetuses are exposed. The long-term effects of prenatal dexamethasone exposure on fetal development are still under evaluation, and current guidelines recommend full informed consent before treatment, ideally in a study protocol that allows long-term follow-up of all children treated. Newer techniques such as cell-free fetal DNA testing and early genotyping may potentially reduce treatment of nonaffected fetuses.

The treatment of other forms of CAH includes mineralocorticoid and glucocorticoid replacement for salt-losing conditions (e.g., *StAR*, *CYP11A1*, *HSD3β2*), suppression of ACTH drive with glucocorticoids in disorders associated with hypertension (e.g., *CYP17*, *CYP11B1*), and appropriate sex hormone replacement in adolescence and adulthood, when necessary.

OTHER CAUSES

Increased androgen synthesis can also occur in CAH due to defects in *POR*, *11β-hydroxylase* (*CYP11B1*), and *3β-hydroxysteroid dehydrogenase type 2* (*HSD3B2*) and with mutations in the genes encoding *aromatase*

(*CYP19*) and the glucocorticoid receptor. Increased androgen exposure in utero can occur with maternal virilizing tumors and with ingestion of androgenic compounds.

OTHER DISORDERS AFFECTING 46,XX FEMALES

Congenital absence of the vagina occurs in association with *müllerian agenesis* or *hypoplasia* as part of the Mayer-Rokitansky-Kuster-Hauser (MRKH) syndrome (rarely caused by *WNT4* mutations). This diagnosis should be considered in otherwise phenotypically normal females with primary amenorrhea. Associated features include renal (agenesis) and cervical spinal abnormalities.

GLOBAL CONSIDERATIONS

The approach to a child or adolescent with ambiguous genitalia or another DSD requires cultural sensitivity, as the concepts of sex and gender vary widely. Rare genetic DSDs can occur more frequently in specific populations (e.g., *5α-reductase type 2* in the Dominican Republic). Different forms of CAH also show ethnic and geographic variability. In many countries, appropriate biochemical tests may not be readily available, and access to appropriate forms of treatment and support may be limited.

FURTHER READING

Achermann JC et al: Disorders of sex development: Effect of molecular diagnostics. Nat Rev Endocrinol 11:478, 2015. [PubMed: 25942653]

Ahmed SF et al: Society for Endocrinology UK guidance on the initial evaluation of an infant or an adolescent with a suspected disorder of sex development (Revised 2015). Clin Endocrinol (Oxf) 84:771, 2016. [PubMed: 26270788]

Backeljauw PF et al: Proceedings from the Turner Resource Network Symposium: The crossroads of health care research and health care delivery. Am J Med Genet 167:1962, 2015.

Cools M et al: Gonadal pathology and tumor risk in relation to clinical characteristics in patients with 45,X/46,XY mosaicism. J Clin Endocrinol Metab 96:E1171, 2011. [PubMed: 21508138]

Davis SM et al: Testis Development and Fertility Potential in Boys with Klinefelter Syndrome. Endocrinol Metab Clin North Am 44:843, 2015. [PubMed: 26568497]

Gravholt CH et al: Clinical practice guidelines for the care of girls and women with Turner syndrome: Proceedings from the 2016 Cincinnati International Turner Syndrome Meeting. Eur J Endocrinol 177:G1, 2017. [PubMed: 28705803]

Hughes IA et al: Androgen insensitivity syndrome. Lancet 380:1490, 2012.

Lee PA et al: Global disorders of sex development update since 2006: Perceptions, approach and care. Horm Res Paediatr 85:158, 2016. [PubMed: 26820577]

Turcu AF, Auchus RJ: The next 150 years of congenital adrenal hyperplasia. J Steroid Biochem Mol Biol 153:63, 2015. [PubMed: 26047556]

Warne GL, Raza J: Disorders of sex development (DSDs), their presentation and management in different cultures. Rev Endocr Metab Disord 9:227, 2008. [PubMed: 18633712]

Chapter 384: Disorders of the Testes and Male Reproductive System

Shalender Bhasin; J. Larry Jameson

INTRODUCTION

The male reproductive system regulates sex differentiation, androgenization, and the hormonal changes that accompany puberty, ultimately leading to spermatogenesis and fertility. Under the control of the pituitary hormones—luteinizing hormone (LH) and follicle-stimulating hormone (FSH)—the Leydig cells of the testes produce testosterone and germ cells are nurtured by Sertoli cells to divide, differentiate, and mature into sperm. During embryonic development, testosterone and dihydrotestosterone (DHT) induce the wolffian duct and virilization of the external genitalia. During puberty, testosterone promotes somatic growth and the development of secondary sex characteristics. In the adult, testosterone is necessary for spermatogenesis, libido and normal sexual function, and maintenance of muscle and bone mass. This chapter focuses on the physiology of the testes and disorders associated with decreased androgen production, which may be caused by gonadotropin deficiency or by primary testis dysfunction. A variety of testosterone formulations now allow more physiologic androgen replacement. Infertility occurs in ~5% of men and is increasingly amenable to treatment by hormone replacement or by using sperm transfer techniques. **For further discussion of sexual dysfunction, disorders of the prostate, and testicular cancer, see Chaps. 390, 83, 84, respectively.**

DEVELOPMENT AND STRUCTURE OF THE TESTIS

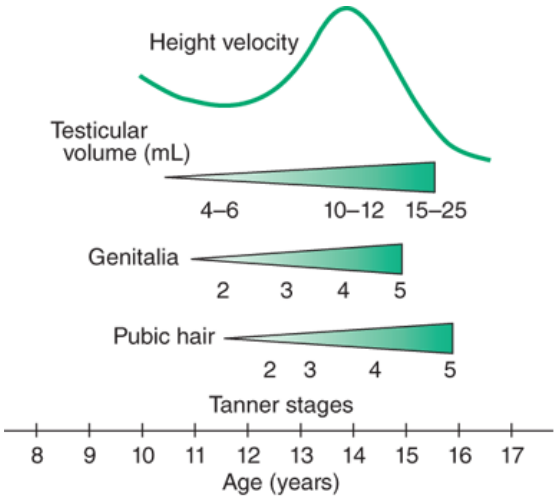
The fetal testis develops from the undifferentiated gonad after expression of a genetic cascade that is initiated by the SRY (sex-related gene on the Y chromosome) **(Chap. 383)**. SRY induces differentiation of Sertoli cells, which surround germ cells and, together with peritubular myoid cells, form testis cords that will later develop into seminiferous tubules. Fetal Leydig cells and endothelial cells migrate into the gonad from the adjacent mesonephros but may also arise from interstitial cells that reside between testis cords. Fetal Leydig cells atrophy after birth and do not contribute to the origin of adult Leydig cells, which originate from undifferentiated progenitor cells that appear in the testis after birth and acquire full steroidogenic function during puberty. Testosterone produced by the fetal Leydig cells supports the growth and differentiation of wolffian duct structures that develop into the epididymis, vas deferens, and seminal vesicles. Testosterone is also converted to DHT (see below), which induces formation of the prostate and the external male genitalia, including the penis, urethra, and scrotum. Testicular descent through the inguinal canal is controlled in part by Leydig cell production of insulin-like factor 3 (INSL3), which acts via a receptor termed *Great* (*G* protein–coupled *re*ceptor *a*ffecting *t*estis descent). Sertoli cells produce müllerian inhibiting substance (MIS), which causes regression of the müllerian structures, including the fallopian tube, uterus, and upper segment of the vagina.

NORMAL MALE PUBERTAL DEVELOPMENT

Puberty commonly refers to the maturation of the reproductive axis and the development of secondary sex characteristics. In addition to reproductive hormones, it requires a coordinated response of multiple hormonal systems including metabolic signals (e.g., leptin), as well as the adrenal and growth hormone (GH) axes **(Fig. 384-1)**. The development of secondary sex characteristics is initiated by *adrenarche*, which usually occurs between 6 and 8 years of age when the adrenal gland begins to produce greater amounts of androgens from the zona reticularis, the principal site of dehydroepiandrosterone (DHEA) production. The sex maturation process is greatly accelerated by the activation of the hypothalamic-pituitary axis and the production of gonadotropin-releasing hormone (GnRH). The GnRH pulse generator in the hypothalamus is active during fetal life and early infancy, but is restrained until the early stages of puberty by a neuroendocrine brake imposed by the inhibitory actions of glutamate and γ–amino butyric acid (GABA) in the mediobasal hypothalamus, and neuropeptide Y. Although the pathways that initiate reactivation of the GnRH pulse generator at the onset of puberty remain incompletely understood, mounting evidence supports involvement of GPR54, a G protein–coupled receptor that binds an endogenous ligand, kisspeptin. Individuals with mutations of GPR54 fail to enter puberty, and experiments in primates demonstrate that infusion of the ligand is sufficient to induce premature puberty. Kisspeptin signaling plays an important role in mediating the feedback action of sex steroids on gonadotropin secretion and in regulating the tempo of sexual maturation at puberty. Leptin, a hormone produced by adipose cells, plays a permissive role in the resurgence of GnRH secretion at the onset of puberty, as leptin-deficient individuals also fail to enter puberty **(Chap. 394)**. Adipocyte hormone leptin, gut hormone ghrelin, neuropeptide Y, and kisspeptin integrate the signals originating in energy stores and metabolic tissues with mechanisms that control onset of puberty through regulation of GnRH secretion. Energy deficit and excess, and metabolic stress are associated with disturbed reproductive maturation and timing of puberty.

FIGURE 384-1

Pubertal events in males. Sexual maturity ratings for genitalia and pubic hair and divided into five stages. *(From WA Marshall, JM Tanner: Variations in the pattern of pubertal changes in boys. Arch Dis Child 45:13, 1970.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The early stages of puberty are characterized by nocturnal surges of LH and FSH. Growth of the testes is usually the first clinical sign of puberty, reflecting an increase in seminiferous tubule volume. Increasing levels of testosterone deepen the voice and stimulate muscle growth. Conversion of testosterone to DHT leads to growth of the external genitalia and pubic hair. DHT also stimulates prostate and facial hair growth and initiates recession of the temporal hairline. The growth spurt occurs at a testicular volume of about 10–12 mL. GH increases early in puberty and is stimulated in part by the rise in gonadal steroids. GH increases the level of insulin-like growth factor 1 (IGF-1), which enhances linear bone growth. The prolonged pubertal exposure to gonadal steroids (mainly estradiol) ultimately induces epiphyseal closure and limits further bone growth.

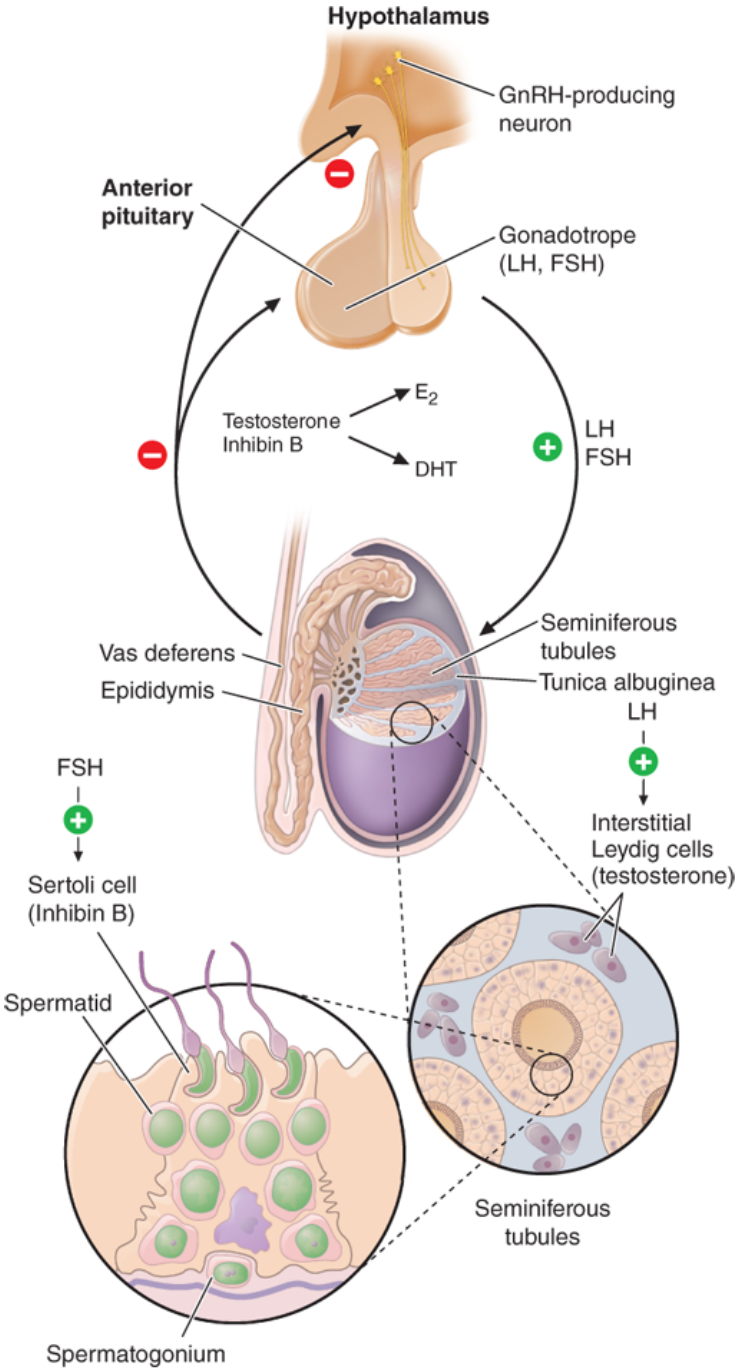
REGULATION OF TESTICULAR FUNCTION

REGULATION OF THE HYPOTHALAMIC-PITUITARY-TESTIS AXIS IN ADULT MAN

Hypothalamic GnRH regulates the production of the pituitary gonadotropins, LH and FSH **(Fig. 384-2)**. GnRH is released in discrete pulses approximately every 2 h, resulting in corresponding pulses of LH and FSH. These dynamic hormone pulses account in part for the wide variations in LH and testosterone, even within the same individual. LH acts primarily on the Leydig cell to stimulate testosterone synthesis. The regulatory control of androgen synthesis is modulated by dynamic integration of the feedforward elements exerted on the testis by LH and FSH, and the feedback exerted by testosterone and estrogen on both the hypothalamus and the pituitary. FSH acts on the Sertoli cell to regulate spermatogenesis and the production of Sertoli products such as inhibin B, which acts to selectively suppress pituitary FSH. Despite these somewhat distinct Leydig and Sertoli cell–regulated pathways, testis function is integrated at several levels: GnRH regulates both gonadotropins; spermatogenesis requires high levels of testosterone; numerous paracrine interactions between Leydig and Sertoli cells are necessary for normal testis function.

FIGURE 384-2

Hypothalamic pituitary gonadotropin axis, structure of testis, seminiferous tubule. E_2, 17β-estradiol; DHT, dihydrotestosterone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

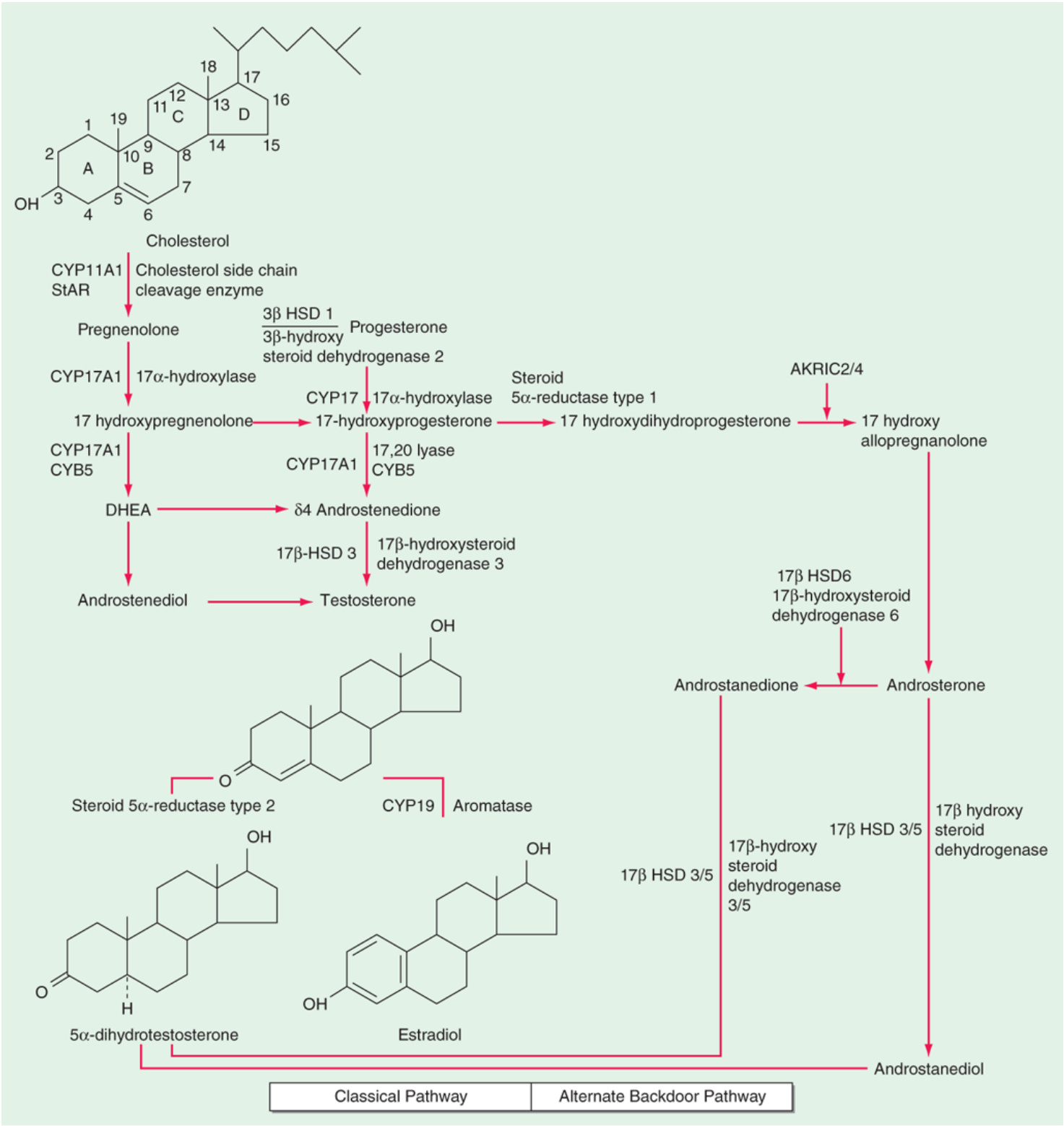
THE LEYDIG CELL: ANDROGEN SYNTHESIS

LH binds to its seven transmembrane, G protein–coupled receptor to activate the cyclic AMP pathway. Stimulation of the LH receptor induces *st*eroid *a*cute *r*egulatory (StAR) protein, along with several steroidogenic enzymes involved in androgen synthesis. LH receptor mutations cause Leydig cell hypoplasia or agenesis, underscoring the importance of this pathway for Leydig cell development and function. The rate-limiting process in testosterone synthesis is the transport of intracellular cholesterol by the StAR protein to the inner mitochondrial membrane. Mutations of the StAR protein are associated with Congenital Lipoid Adrenal Hyperplasia, a rare form of congenital adrenal hyperplasia (CAH) characterized by very low adrenal and gonadal steroids. Peripheral benzodiazepine receptor, a mitochondrial cholesterol-binding protein, is also an acute regulator of Leydig cell steroidogenesis. The major enzymatic steps involved in testosterone synthesis are summarized in **Fig. 384-3**. After cholesterol transport into the mitochondrion, the formation of pregnenolone by CYP11A1 (side chain cleavage enzyme) is a limiting enzymatic step. The 17α-hydroxylase and the 17,20-lyase reactions are catalyzed by a single enzyme, CYP17; posttranslational modification (phosphorylation) of this enzyme and the presence of specific enzyme cofactors, such as cytochrome B, confer 17,20-lyase activity selectively in the testis and zona reticularis of the adrenal gland. Abiraterone is a dual inhibitor of 17 α-hydroxylase and 17,20-lyase activities, which play an important role in androgen synthesis in castration-resistant prostate cancers. Testosterone can be converted to the more potent DHT by a family of steroid 5α-reductase enzymes, or it can be aromatized to estradiol by CYP19

(aromatase). At least two isoforms of steroid 5α-reductase, SRD5A1 and SRD5A2, have been described; all known patients with 5α-reductase deficiency have had mutations in SRD5A2, the predominant form in the prostate and the skin. Finasteride predominantly inhibits SRD5A2, while dutasteride is a dual inhibitor of both SRD5A1 and SRD5A2. DHT can also be derived directly through the backdoor pathway in which 17 α-hydroxyprogesterone is converted through a series of 5 α and 3 α reductions to androsterone and eventually to DHT. Recent reports of mutations in AKR1C2/4 genes in undervirilized 46, XY individuals suggest that the backdoor pathway for DHT formation, which was originally described in the tammar wallaby, is active in the human fetal testis.

FIGURE 384-3

The biochemical pathway in the conversion of 27-carbon sterol cholesterol to androgens and estrogens.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

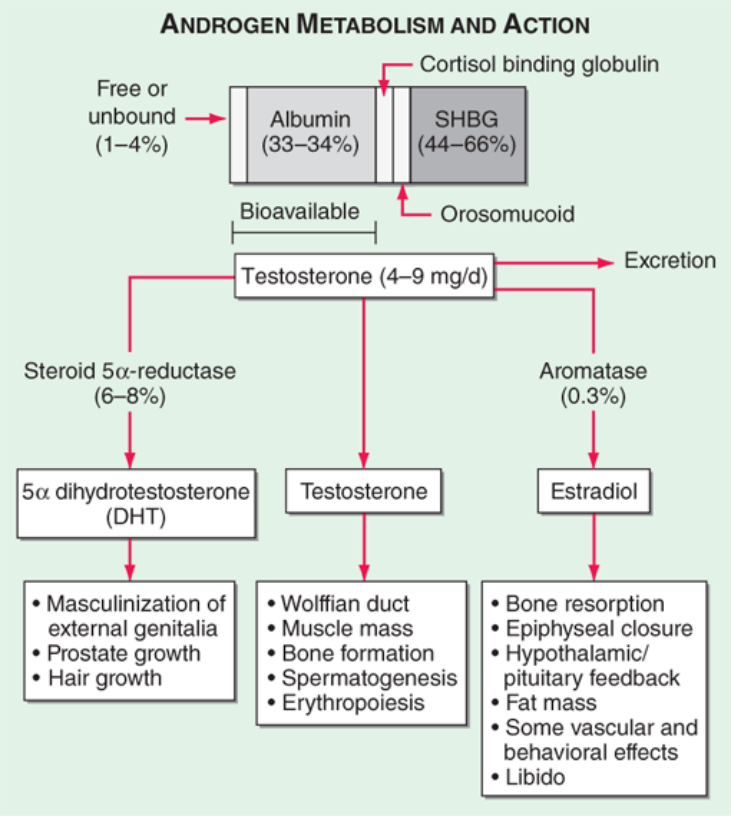
Testosterone Transport and Metabolism

In males, 95% of circulating testosterone is derived from testicular production (3–10 mg/d). Direct secretion of testosterone by the adrenal and the peripheral conversion of androstenedione to testosterone collectively account for another 0.5 mg/d of testosterone. Only a small amount of DHT (70 µg/d) is secreted directly by the testis; most circulating DHT is derived from peripheral conversion of testosterone. Most of the daily production of estradiol (~45 µg/d) in men is derived from aromatase-mediated peripheral conversion of testosterone and androstenedione.

Circulating testosterone is bound predominantly to sex hormone–binding globulin (SHBG) and albumin **(Fig. 384-4)**, and to a lesser extent to cortisol binding globulin (CBG), and orosomucoid. SHBG binds testosterone with much greater affinity than albumin, CBP, and orosomucoid. Only 1.0–4.0% of testosterone is unbound. According to the "free hormone" hypothesis, only the unbound fraction is biologically active. The term "bioavailable testosterone" refers to unbound testosterone plus testosterone bound loosely to albumin, and reflects the concept that albumin-bound testosterone can dissociate at the capillary level, especially in tissues with long transit time, such as the liver and the brain. SHBG-bound testosterone also may be internalized through endocytic pits by binding to a protein called megalin. SHBG concentrations are decreased by androgens, obesity, diabetes mellitus, hypothyroidism, nephrotic syndrome, and genetic factors. Conversely, estrogen administration, hyperthyroidism, many chronic inflammatory illnesses, infections such as HIV or hepatitis B and C, aging, and the use of some anticonvulsants are associated with high SHBG concentrations.

FIGURE 384-4

Androgen metabolism and actions. SHBG, sex hormone–binding globulin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Testosterone is metabolized predominantly in the liver, although some degradation occurs in peripheral tissues, particularly the prostate and the skin. In the liver, testosterone is converted by a series of enzymatic steps that involve 5α- and 5β-reductases, 3α- and 3β-hydroxysteroid dehydrogenases, and 17β-hydroxysteroid dehydrogenase into androsterone, etiocholanolone, DHT, and 3-α-androstanediol. These compounds undergo glucuronidation or sulfation before being excreted by the kidneys.

Mechanism of Androgen Action

Testosterone exerts some of its biologic effects by binding to androgen receptor (AR), either directly or after its conversion to DHT by the steroid 5-α reductase. The actions of testosterone on the Wolffian structures, skeletal muscle, erythropoiesis, and bone in men do not require its obligatory conversion to DHT. However, the conversion of testosterone to DHT is necessary for the masculinization of the urogenital sinus and genital tubercle. Aromatization of testosterone to estradiol mediates additional effects of testosterone on the bone resorption, epiphyseal closure, sexual desire, vascular endothelium, and fat. DHT can also be converted in some tissues by 3-keto reductase/3β-hydroxysteroid dehydrogenase enzymes to 5α-androstane-3β,17β-diol, which is a high-affinity ligand and agonist of estrogen receptor ER β.

The AR is structurally related to the nuclear receptors for estrogen, glucocorticoids, and progesterone (Chap. 370). The AR is encoded by a gene on the long arm of the X chromosome and has a molecular mass of about 110 kDa. A polymorphic region in the amino terminus of the receptor, which contains a variable number of glutamine repeats, modifies the transcriptional activity of the receptor. The AR protein is distributed in both the cytoplasm and the nucleus. The ligand binding to the AR induces conformational changes that allow the recruitment and assembly of tissue-specific cofactors, and causes it to translocate into the nucleus, where it binds to specific androgen response elements in the DNA or other transcription factors already bound to DNA. Thus, the AR is a ligand-regulated transcription factor that regulates the expression of androgen-dependent genes in a tissue-specific manner. Some androgen effects, such as those on the smooth muscle, may be mediated by nongenomic AR signal transduction pathways. Testosterone binds to AR with half the affinity of DHT. The DHT-AR complex also has greater thermostability and a slower dissociation rate than the testosterone-AR complex. However, the molecular basis for selective testosterone versus DHT actions remains incompletely explained.

THE SEMINIFEROUS TUBULES: SPERMATOGENESIS

The seminiferous tubules are convoluted, closed loops with both ends emptying into the rete testis, a network of progressively larger efferent ducts that ultimately form the epididymis (Fig. 384-2). The seminiferous tubules total about 600 m in length and comprise about two-thirds of testis volume. The walls of the tubules are formed by polarized Sertoli cells that are apposed to peritubular myoid cells. Tight junctions between Sertoli cells create the blood-testis barrier. Germ cells comprise the majority of the seminiferous epithelium (~60%) and are intimately embedded within the cytoplasmic extensions of the Sertoli cells, which function as “nurse cells.” Germ cells progress through characteristic stages of mitotic and meiotic divisions. A pool of type A spermatogonia serve as stem cells capable of self-renewal. Primary spermatocytes are derived from type B spermatogonia and undergo meiosis before progressing to spermatids that undergo spermiogenesis (a differentiation process involving chromatin condensation, acquisition of an acrosome, elongation of cytoplasm, and formation of a tail) and are released from Sertoli cells as mature spermatozoa. The complete differentiation process into mature sperm requires 74 days. Peristaltic-type action by peritubular myoid cells transports sperm into the efferent ducts. The spermatozoa spend an additional 21 days in the epididymis, where they undergo further maturation and capacitation. The normal adult testes produce >100 million sperm per day.

Naturally occurring mutations in FSHβ or in the FSH receptor confirm an important, but not essential, role for this pathway in spermatogenesis. Females with mutations in FSHβ or the FSH receptor are hypogonadal and infertile because ovarian follicles do not mature; males with these mutations exhibit variable degrees of reduced spermatogenesis, presumably because of impaired Sertoli cell function. Because Sertoli cells produce inhibin B, an inhibitor of FSH, seminiferous tubule damage (e.g., by radiation) causes a selective increase of FSH. Testosterone reaches very high concentrations locally in the testis and is essential for spermatogenesis. The cooperative actions of FSH and testosterone are important in the progression of meiosis and spermiation. FSH and testosterone regulate germ cell survival via the intrinsic and the extrinsic apoptotic mechanisms. FSH may also play an important role in supporting spermatogonia. Gonadotropin-regulated testicular RNA helicase (GRTH/DDX25), a testis-specific gonadotropin/androgen-regulated RNA helicase, is present in germ cells and Leydig cells and may be an important factor in the paracrine regulation of germ cell development. Several cytokines and growth factors are also involved in the regulation of spermatogenesis by paracrine and autocrine mechanisms. A number of knockout mouse models exhibit impaired germ cell development or spermatogenesis, presaging possible mutations associated with male infertility.

The human Y chromosome contains a small pseudoautosomal region that can recombine with homologous regions of the X chromosome. Most of the Y chromosome does not recombine with the X chromosome and is referred to as the male-specific region of the Y (MSY). The MSY contains 156 transcription units that encode for 26 proteins, including nine families of Y-specific multicopy genes; many of these Y-specific genes are also testis-specific and necessary for spermatogenesis. Microdeletions in several nonoverlapping subregions of the Y chromosome—AZFa, AZFb, AZFc and AZFd, which contain many spermatogenic genes (e.g., RNA-binding motif, *RBM*; deleted in azoospermia, *DAZ*)—are associated with oligospermia or azoospermia. Approximately 15% of infertile men with azoospermia and about 6% of men with severe oligozoospermia harbor a Y microdeletion. Microdeletions of the AZFa and AZFb subregions are typically associated with Sertoli cell only or maturation arrest histology, azoospermia, and poor prognosis for sperm retrieval. In contrast, AZFc subregion microdeletions are typically associated with oligozoospermia and higher success rates for sperm retrieval. Microdeletion involving the DAZ gene in the AZFc region is one of the commonest Y chromosome microdeletions in infertile men. A partial deletion of the AZFc region called the *gr/gr deletion* is associated with infertility among Caucasian men in Europe and the Western Pacific region.

TREATMENT

TREATMENT

Male Factor Infertility

Treatment options for male factor infertility have expanded greatly in recent years. Secondary hypogonadism is highly amenable to treatment with pulsatile GnRH or gonadotropins (see below). Assisted reproductive technologies, such as the in vitro fertilization (IVF) and intracytoplasmic sperm injection (ICSI), have provided new opportunities for patients with primary testicular failure and disorders of sperm transport. Choice of initial treatment options depends on sperm concentration and motility. Expectant management should be attempted initially in men with mild male factor infertility (sperm count of 15–20 × 10^6/mL and normal motility). Moderate male factor infertility (10–15 x 10^6/mL and 20–40% motility) should begin with intrauterine insemination alone or in combination with treatment of the female partner with clomiphene or gonadotropins, but it may require IVF with or

without ICSI. For men with a severe defect (sperm count of $<10 \times 10^6$/mL, 10% motility), IVF with ICSI or donor sperm has become the treatment of choice. Yq microdeletions will be transmitted through ICSI from the affected father to his male offspring if sperm carrying Yq microdeletion is used.

CLINICAL AND LABORATORY EVALUATION OF MALE REPRODUCTIVE FUNCTION

HISTORY AND PHYSICAL EXAMINATION

The history should focus on developmental stages such as puberty and growth spurts, as well as androgen-dependent events such as early morning erections, frequency and intensity of sexual thoughts, and frequency of masturbation or intercourse. Although libido and the overall frequency of sexual acts are decreased in androgen-deficient men, young hypogonadal men can achieve erections in response to visual erotic stimuli. Men with acquired androgen deficiency often report decreased energy and low mood.

The physical examination should focus on secondary sex characteristics such as hair growth, gynecomastia, testicular volume, prostate, and height and body proportions. *Eunuchoid proportions* are defined as an arm span >2 cm greater than height and suggest that androgen deficiency occurred before epiphyseal fusion. Hair growth in the face, axilla, chest, and pubic regions is androgen-dependent; however, changes may not be noticeable unless androgen deficiency is severe and prolonged. Ethnicity also influences the intensity of hair growth **(Chap. 387)**. Testicular volume is best assessed by using a Prader orchidometer. Testes range from 3.5 to 5.5 cm in length, which corresponds to a volume of 12–25 mL. Advanced age does not influence testicular size, although the consistency becomes less firm. Asian men generally have smaller testes than western Europeans, independent of differences in body size. Because of its possible role in infertility, the presence of varicocele should be sought by palpation while the patient is standing; it is more common on the left side. Patients with Klinefelter syndrome have markedly reduced testicular volumes (1–2 mL). In congenital hypogonadotropic hypogonadism, testicular volumes provide a good index for the degree of gonadotropin deficiency and the likelihood of response to therapy.

GONADOTROPIN AND INHIBIN MEASUREMENTS

LH and FSH are measured using two-site immunoradiometric, immunofluorometric, or chemiluminescent assays, which have very low cross-reactivity with other pituitary glycoprotein hormones and human chorionic gonadotropin (hCG) and have sufficient sensitivity to measure the low levels present in patients with hypogonadotropic hypogonadism. In men with a low testosterone level, an LH level can distinguish primary (high LH) versus secondary (low or inappropriately normal LH) hypogonadism. An elevated LH level indicates a primary defect at the testicular level, whereas a low or inappropriately normal LH level suggests a defect at the hypothalamic-pituitary level. LH pulses occur about every 1–3 h in normal men. Thus, gonadotropin levels fluctuate, and samples should be pooled or repeated when results are equivocal. FSH is less pulsatile than LH because it has a longer half-life. Selective increase in FSH suggests damage to the seminiferous tubules. Inhibin B, a Sertoli cell product that suppresses FSH, is reduced with seminiferous tubule damage. Inhibin B is a dimer with α-β_B subunits and is measured by two-site immunoassays.

GnRH Stimulation Testing

The GnRH test is performed by measuring LH and FSH concentrations at baseline and at 30 and 60 min after intravenous administration of 100 μg of GnRH. A minimally acceptable response is a twofold LH increase and a 50% FSH increase. In the prepubertal period or with severe GnRH deficiency, the gonadotrope may not respond to a single bolus of GnRH because it has not been primed by endogenous hypothalamic GnRH; in these patients, GnRH responsiveness may be restored by chronic, pulsatile GnRH administration. With the availability of sensitive and specific LH assays, GnRH stimulation testing is used rarely.

TESTOSTERONE ASSAYS

Total Testosterone

Total testosterone includes both unbound and protein-bound testosterone and is measured by radioimmunoassays, immunometric assays, or liquid chromatography tandem mass spectrometry (LC-MS/MS). LC-MS/MS involves extraction of serum by organic solvents, separation of testosterone from other steroids by high-performance liquid chromatography and mass spectrometry, and quantitation of unique testosterone fragments by mass spectrometry. LC-MS/MS provides accurate and sensitive measurements of testosterone levels even in the low range and has emerged as the method of choice for testosterone measurement. The use of laboratories that have been certified by the Centers for Disease Control's Hormone Standardization Program for Testosterone (HoST) can ensure that testosterone measurements are accurate and calibrated to an international standard. A single fasting morning sample provides a good approximation of the average testosterone concentration with the realization that testosterone levels fluctuate in response to pulsatile LH. Testosterone is generally lower in the late afternoon and is reduced by acute illness. The harmonized normal range for total testosterone, measured using LC-MS/MS in nonobese populations of European and American men, 19–39 years, is 264–916 ng/dL. This harmonized reference range can be applied to values from laboratories that are certified by the CDC's Hormone Standardization Program for Testosterone.

Alterations in SHBG levels due to aging, obesity, diabetes mellitus, hyperthyroidism, some types of medications, chronic illness, or on a congenital basis, can affect total testosterone levels. Heritable factors contribute substantially to the population level variation in testosterone levels and genome wide association studies have revealed polymorphisms in SHBG gene as important contributors to variation in testosterone levels.

Measurement of Unbound Testosterone Levels

Most circulating testosterone is bound to SHBG and to albumin; only 1.0–4% of circulating testosterone is unbound, or "free." Free testosterone should ideally be measured by equilibrium dialysis under standardized conditions using an accurate and reliable assay for total testosterone. The unbound testosterone concentration also can be calculated from total testosterone, SHBG, and albumin concentrations. Recent research has shown that testosterone binding to SHBG is a multi-step process that involves complex allosteric interactions between the two binding sites within the SHBG dimer; a novel ensemble allosteric model of testosterone's binding to SHBG dimers provides good estimates of free testosterone concentrations. The previous law-of-mass action equations based on linear models of testosterone binding to SHBG used assumptions that have been shown to be erroneous. Tracer analogue methods are relatively inexpensive and convenient, but they are inaccurate. *Bioavailable testosterone* refers to unbound testosterone plus testosterone that is loosely bound to albumin; it can be determined by the ammonium sulfate precipitation method. However, the measurements of bioavailable testosterone using the ammonium sulfate precipitation are technically challenging, susceptible to imprecision, and are not recommended.

hCG Stimulation Test

The hCG stimulation test is performed by administering a single injection of 1500–4000 IU of hCG intramuscularly and measuring testosterone levels at baseline and 24, 48, 72, and 120 h after hCG injection. An alternative regimen involves three injections of 1500 units of hCG on successive days and measuring testosterone levels 24 h after the last dose. An acceptable response to hCG is a doubling of the testosterone concentration in adult men. In prepubertal boys, an increase in testosterone to >150 ng/dL indicates the presence of testicular tissue. No response may indicate an absence of testicular tissue or marked impairment of Leydig cell function. Measurement of MIS, a Sertoli cell product, is also used to detect the presence of testes in prepubertal boys with cryptorchidism.

SEMEN ANALYSIS

Semen analysis is the most important step in the evaluation of male infertility. Samples are collected by masturbation following a period of abstinence for 2–3 days. Semen volumes and sperm concentrations vary considerably among fertile men, and several samples may be needed before concluding that the results are abnormal. Analysis should be performed within an hour of collection. Using semen samples from over 4500 men in 14 countries, whose partners had a time-to-pregnancy of <12 months, WHO has generated the following one-sided reference limits for semen parameters: semen volume, 1.5 mL; total sperm number, 39 million per ejaculate; sperm concentration, 15 million per mL; vitality, 58% live; progressive motility, 32%; total (progressive + non-progressive) motility, 40%; morphologically normal forms, 4.0%. Some men with low sperm counts are nevertheless fertile. A variety of tests for sperm function can be performed in specialized laboratories, but these add relatively little to the treatment options.

TESTICULAR BIOPSY

Testicular biopsy is useful in some patients with oligospermia or azoospermia as an aid in diagnosis and indication for the feasibility of treatment. Using fine-needle aspiration biopsy is performed under local anesthesia to aspirate tissue for histology. Alternatively, open biopsies can be performed under local or general anesthesia when more tissue is required. A normal biopsy in an azoospermic man with a normal FSH level suggests obstruction of the vas deferens, which may be correctable surgically. Biopsies are also used to harvest sperm for ICSI and to classify disorders such as hypospermatogenesis (all stages present but in reduced numbers), germ cell arrest (usually at primary spermatocyte stage), and Sertoli cell–only syndrome (absent germ cells) or hyalinization (sclerosis with absent cellular elements).

Testing for Y Chromosome Microdeletions

Y chromosome microdeletions are detected by extracting DNA from peripheral blood leukocytes and using polymerase chain reaction (PCR) amplification using primers for some 300 sequence-tagged sites on the Y chromosome, followed by gel electrophoresis to determine whether the DNA sequences corresponding to the selected Y chromosome markers are present. However, because these ~300 Y chromosome markers account for only a small fraction of the 23 million base pairs on the Y chromosome, a negative test does not exclude microdeletions in other subregions of the Y chromosome.

DISORDERS OF SEXUAL DIFFERENTIATION

See Chap. 383.

DISORDERS OF PUBERTY

The onset and tempo of puberty varies greatly in the general population and is affected by genetic, nutritional, and environmental factors. Although a substantial fraction of the variance in the timing of puberty is explained by heritable factors, the genes involved remain unknown.

PRECOCIOUS PUBERTY

Puberty in boys aged <9 years is considered precocious. *Isosexual precocity* refers to premature sexual development consistent with phenotypic sex and includes features such as the development of facial hair and phallic growth. Isosexual precocity is divided into gonadotropin-dependent and gonadotropin-independent causes of androgen excess **(Table 384-1)**. *Heterosexual precocity* refers to the premature development of estrogenic features in boys, such as breast development.

TABLE 384-1

Causes of Precocious or Delayed Puberty in Boys

- I. Precocious puberty
  - A. Gonadotropin-dependent
    - 1. Idiopathic
    - 2. Hypothalamic hamartoma or other lesions
    - 3. CNS tumor or inflammatory state
  - B. Gonadotropin-independent
    - 1. Congenital adrenal hyperplasia
    - 2. hCG-secreting tumor
    - 3. McCune-Albright syndrome
    - 4. Activating LH receptor mutation
    - 5. Exogenous androgens
    - 6. Androgen producing tumors of the adrenal or the testis
- II. Delayed puberty
  - A. Constitutional delay of growth and puberty
  - B. Systemic disorders
    - 1. Chronic disease
    - 2. Malnutrition
    - 3. Anorexia nervosa
  - C. CNS tumors and their treatment (radiotherapy and surgery)
  - D. Hypothalamic-pituitary causes of pubertal failure (low gonadotropins)
    - 1. Congenital disorders (Table 384-2)
    - 2. Acquired disorders
      - a. Pituitary tumors
      - b. Hyperprolactinemia
      - c. Infiltrative disorders, such as hemachromatosis
  - E. Gonadal causes of pubertal failure (elevated gonadotropins)
    - 1. Klinefelter syndrome
    - 2. Bilateral undescended testes
    - 3. Orchitis
    - 4. Chemotherapy or radiotherapy
    - 5. Anorchia
  - F. Androgen insensitivity

Abbreviations: CNS, central nervous system; GnRH, gonadotropin-releasing hormone; hCG, human chronic gonadotropin; LH, luteinizing hormone.

Gonadotropin-Dependent Precocious Puberty

This disorder, called *central precocious puberty* (CPP), is less common in boys than in girls. It is caused by premature activation of the GnRH pulse generator, sometimes because of central nervous system (CNS) lesions such as hypothalamic hamartomas, but it is often idiopathic. CPP is characterized by gonadotropin levels that are inappropriately elevated for age. Because pituitary priming has occurred, GnRH elicits LH and FSH responses typical of those seen in puberty or in adults. MRI should be performed to exclude a mass, structural defect, infection, or inflammatory process. Mutations in kisspeptin, kisspeptin receptor, and MKRN3, an imprinted gene encoding makorin RING-finder protein 3, which is expressed only from the paternally inherited allele, have been associated with CPP. Loss of function mutations in MKRN3 remove the brake that restrains pulsatile GnRH, resulting in precocious puberty.

Gonadotropin-Independent Precocious Puberty

Androgens from the testis or the adrenal are increased but gonadotropins are low. This group of disorders includes hCG-secreting tumors; CAH; sex steroid–producing tumors of the testis, adrenal, and ovary; accidental or deliberate exogenous sex steroid administration; hypothyroidism; and activating mutations of the LH receptor or $G_s\alpha$ subunit.

Familial Male-Limited Precocious Puberty

Also called *testotoxicosis*, familial male-limited precocious puberty is an autosomal dominant disorder caused by activating mutations in the LH receptor, leading to constitutive stimulation of the cyclic AMP pathway and testosterone production. Clinical features include premature androgenization in boys, growth acceleration in early childhood, and advanced bone age followed by premature epiphyseal fusion. Testosterone is elevated, and LH is suppressed. Treatment options include inhibitors of testosterone synthesis (e.g., ketoconazole, medroxyprogesterone acetate), AR antagonists (e.g., flutamide and bicalutamide), and aromatase inhibitors (e.g., anastrozole).

MCCUNE-ALBRIGHT SYNDROME

This is a sporadic disorder caused by somatic (postzygotic) activating mutations in the $G_s\alpha$ subunit that links G protein–coupled receptors to intracellular signaling pathways **(Chap. 405)**. The mutations impair the guanosine triphosphatase activity of the $G_s\alpha$ protein, leading to constitutive activation of adenylyl cyclase. Like activating LH receptor mutations, this stimulates testosterone production and causes gonadotropin-independent precocious puberty. In addition to sexual precocity, affected individuals may have autonomy in the adrenals, pituitary, and thyroid glands. *Café au lait* spots are characteristic skin lesions that reflect the onset of the somatic mutations in melanocytes during embryonic development. Polyostotic fibrous dysplasia is caused by activation of the parathyroid hormone receptor pathway in bone. Treatment is similar to that in patients with activating LH receptor mutations. Bisphosphonates have been used to treat bone lesions.

CONGENITAL ADRENAL HYPERPLASIA

Boys with CAH who are not well controlled with glucocorticoid suppression of adrenocorticotropic hormone (ACTH) can develop premature virilization because of excessive androgen production by the adrenal gland **(Chaps. 379 and 383)**. LH is low, and the testes are small. Adrenal rests may develop within the testis of poorly controlled patients with CAH because of chronic ACTH stimulation; adrenal rests do not require surgical removal and regress with effective glucocorticoid therapy. Some children with CAH may develop gonadotropin-dependent precocious puberty with early maturation of the hypothalamic-pituitary-gonadal axis, elevated gonadotropins, and testicular growth.

Heterosexual Sexual Precocity

Breast enlargement in prepubertal boys can result from familial aromatase excess, estrogen-producing tumors in the adrenal gland, Sertoli cell tumors in the testis, marijuana smoking, or exogenous estrogens or androgens. Occasionally, germ cell tumors that secrete hCG can be associated with breast enlargement due to excessive stimulation of estrogen production (see "Gynecomastia," below).

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Precocious Puberty

After verification of precocious development, serum testosterone, LH and FSH levels should be measured to determine whether gonadotropins are increased in relation to chronologic age (gonadotropin-dependent) or whether sex steroid secretion is occurring independent of LH and FSH (gonadotropin-independent). In children with gonadotropin-dependent precocious puberty, CNS lesions should be excluded by history, neurologic examination, and MRI scan of the head. If organic causes are not found, one is left with the diagnosis of idiopathic central precocity. Patients with high testosterone but suppressed LH concentrations have gonadotropin-independent sexual precocity; in these patients, DHEA sulfate (DHEAS) and 17α-hydroxyprogesterone should be measured. High levels of testosterone and 17α-hydroxyprogesterone suggest the possibility of CAH due to 21α-hydroxylase or 11β-hydroxylase deficiency. If testosterone and DHEAS are elevated, adrenal tumors should be excluded by obtaining a CT scan of the adrenal glands. Patients with elevated testosterone but without increased 17α-hydroxyprogesterone or DHEAS should undergo careful evaluation of the testis by palpation and ultrasound to exclude a Leydig cell neoplasm. Activating mutations of the LH receptor should be considered in children with gonadotropin-independent precocious puberty in whom CAH, androgen abuse, and adrenal and testicular neoplasms have been excluded.

TREATMENT

TREATMENT

Precocious Puberty

In patients with a known cause (e.g., a CNS lesion or a testicular tumor), therapy should be directed towards the underlying disorder. In patients with idiopathic CPP, long-acting GnRH analogues can be used to suppress gonadotropins and decrease testosterone, halt early pubertal development, delay accelerated bone maturation, prevent early epiphyseal closure, promote final height gain, and mitigate the psychosocial consequences of early pubertal

development without causing osteoporosis. The treatment is most effective for increasing final adult height if it is initiated before age 6. Puberty resumes after discontinuation of the GnRH analogue. Counseling is an important aspect of the overall treatment strategy.

In children with gonadotropin-independent precocious puberty, inhibitors of steroidogenesis, such as ketoconazole, AR antagonists, and aromatase inhibitors have been used empirically. Long-term treatment with spironolactone (a weak androgen antagonist) and ketoconazole has been reported to normalize growth rate and bone maturation and to improve predicted height in small, nonrandomized trials in boys with familial male-limited precocious puberty. Aromatase inhibitors, such as testolactone and letrozole, have been used as adjuncts to antiandrogen therapy for children with familial male-limited precocious puberty, CAH, and McCune-Albright syndrome. More potent novel inhibitors of testosterone synthesis, such as abiraterone, have not been evaluated in boys with gonadotropin-independent precocious puberty.

DELAYED PUBERTY

Puberty is delayed in boys if it has not ensued by age 14, an age that is 2–2.5 standard deviations above the mean for healthy children. Delayed puberty is more common in boys than in girls. There are four main categories of delayed puberty: (1) constitutional delay of growth and puberty (~60% of cases); (2) functional hypogonadotropic hypogonadism caused by systemic illness or malnutrition (~20% of cases); (3) hypogonadotropic hypogonadism caused by genetic or acquired defects in the hypothalamic-pituitary region (~10% of cases); and (4) hypergonadotropic hypogonadism secondary to primary gonadal failure (~15% of cases) (Table 384-1). The constitutional delay of growth and puberty clusters in families displays an autosomal dominant pattern of inheritance, and has been linked in some families with a locus on pericentromeric region of chromosome 2. Functional hypogonadotropic hypogonadism is more common in girls than in boys. Permanent causes of hypogonadotropic or hypergonadotropic hypogonadism are identified in <25% of boys with delayed puberty.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Delayed Puberty

History of systemic illness, eating disorders, excessive exercise, social and psychological problems, and abnormal patterns of linear growth during childhood should be verified. Boys with pubertal delay may have accompanying emotional and physical immaturity relative to their peers, which can be a source of anxiety. Physical examination should focus on height; arm span; weight; visual fields; and secondary sex characteristics, including hair growth, testicular volume, phallic size, and scrotal reddening and thinning. Testicular size >2.5 cm generally indicates that the child has entered puberty.

The main diagnostic challenge is to distinguish those with constitutional delay, who will progress through puberty at a later age, from those with an underlying pathologic process. Constitutional delay should be suspected when there is a family history and when there are delayed bone age and short stature. Pituitary priming by pulsatile GnRH is required before LH and FSH are synthesized and secreted normally. Thus, blunted responses to exogenous GnRH can be seen in patients with constitutional delay, GnRH deficiency, or pituitary disorders. On the other hand, low-normal basal gonadotropin levels or a normal response to exogenous GnRH is consistent with an early stage of puberty, which is often heralded by nocturnal GnRH secretion. Thus, constitutional delay is a diagnosis of exclusion that requires ongoing evaluation until the onset of puberty and the growth spurt.

TREATMENT

TREATMENT

Delayed Puberty

If therapy is considered appropriate, it can begin with 25–50 mg testosterone enanthate or testosterone cypionate every 2 weeks, or by using a 2.5-mg testosterone patch or 25-mg testosterone gel. Because aromatization of testosterone to estrogen is obligatory for mediating androgen effects on epiphyseal fusion, concomitant treatment with aromatase inhibitors may allow attainment of greater final adult height. Testosterone treatment should be interrupted after 6 months to determine if endogenous LH and FSH secretion have ensued. Other causes of delayed puberty should be considered when there are associated clinical features or when boys do not enter puberty spontaneously after a year of observation or treatment.

Reassurance without hormonal treatment is appropriate for many individuals with presumed constitutional delay of puberty. However, the impact of delayed growth and pubertal progression on a child's social relationships and school performance should be weighed. The boys with constitutional delay of puberty are less likely to achieve their full genetic height potential and have reduced total body bone mass as adults, mainly due to narrow limb bones and vertebrae as a result of impaired periosteal expansion during puberty. Furthermore, the time of onset of puberty is negatively associated with bone mineral content and density in boys at skeletal maturity. Judicious use of androgen therapy in carefully selected boys with constitutional delay can induce pubertal induction and progression, and promote short-term growth without compromising final height, and when administered with an aromatase inhibitor, it may improve final height.

DISORDERS OF THE MALE REPRODUCTIVE AXIS DURING ADULTHOOD

HYPOGONADOTROPIC HYPOGONADISM

Because LH and FSH are trophic hormones for the testes, impaired secretion of these pituitary gonadotropins results in secondary hypogonadism, which is characterized by low testosterone in the setting of low or inappropriately normal LH and FSH. Those with the most severe gonadotropin deficiency have complete absence of pubertal development, sexual infantilism, and, in some cases, hypospadias and undescended testes. Patients with partial gonadotropin deficiency have delayed or arrested sex development. The 24-h LH secretory profiles are heterogeneous in patients with hypogonadotropic hypogonadism, reflecting variable abnormalities of LH pulse frequency or amplitude. In severe cases, basal LH is low and there are no LH pulses. A smaller subset of patients has low-amplitude LH pulses or markedly reduced pulse frequency. Occasionally, only sleep-entrained LH pulses occur, reminiscent of the pattern seen in the early stages of puberty. Hypogonadotropic hypogonadism can be classified into congenital and acquired disorders. Congenital disorders most commonly involve GnRH deficiency, which leads to gonadotropin deficiency. Acquired disorders are much more common than congenital disorders and may result from a variety of sellar mass lesions or infiltrative diseases of the hypothalamus or pituitary, or due to the effects of drugs, nutritional or psychiatric disorders, or systemic diseases.

Congenital Disorders Associated with Gonadotropin Deficiency

Congenital hypogonadotropic hypogonadism is a heterogeneous group of disorders characterized by decreased gonadotropin secretion and testicular dysfunction either due to impaired function of the GnRH pulse generator or the gonadotrope. The disorders characterized by GnRH deficiency represent a family of oligogenic disorders whose phenotype spans a wide spectrum. Some individuals with GnRH deficiency may suffer from complete absence of pubertal development, while others may manifest varying degrees of gonadotropin deficiency and pubertal delay, and a subset that carries the same mutations as their affected family members may even have normal reproductive function. In ~10% of men with idiopathic hypogonadotropic hypogonadism (IHH), reversal of gonadotropin deficiency may occur in adult life after sex steroid therapy. Also, a small fraction of men with IHH may present with androgen deficiency and infertility in adult life after having gone through apparently normal pubertal development. Nutritional, emotional, or metabolic stress may unmask gonadotropin deficiency and reproductive dysfunction (e.g., hypothalamic amenorrhea) in some patients who harbor mutations in the candidate genes but who previously had normal reproductive function. The clinical phenotype may include isolated anosmia or hyposmia. Oligogenicity, and gene-gene and gene-environment interactions may contribute to variations in clinical phenotype.

Mutations in a number of genes involved in the development and migration of GnRH neurons, or in the regulation of GnRH secretion have been linked to GnRH deficiency, although the genetic defect remains elusive in nearly two thirds of cases. Familial hypogonadotropic hypogonadism can be transmitted as an X-linked (20%), autosomal recessive (30%), or autosomal dominant (50%) trait. Some individuals with IHH have sporadic mutations in the same genes that cause inherited forms of the disorder. The genetic defects associated with GnRH deficiency can ben conveniently classified as anosmic (Kallmann syndrome) or normosmic (Table 384-2), although the occurrence of both anosmic and normosmic forms of GnRH deficiency in the same families suggests commonality of pathophysiologic mechanisms. *Kallmann syndrome*, the anosmic form of GnRH deficiency, can result from mutations in one or more genes associated with olfactory bulb morphogenesis and the migration of GnRH neurons from their origin in the region of the olfactory placode, along the scaffold established by the olfactory nerves, through the cribriform plate into their final location into the pre-optic region of the hypothalamus. Thus, mutations in *KAL1,* genes involved in fibroblast growth factor (FGF) signaling (*FGF8*, *FGFR1*, *FGF17, IL17RD, DUSP6, SPRY4, and FLRT3), NELF, genes involved in PROK signaling (PROK2* and *PROK2R*), *WDR11, SEMA3, HS6ST1, CHD7, and FEZF1* have been described in patients with Kallmann syndrome. An X-linked form of IHH is caused by mutations in the *KAL1* gene, which encodes anosmin, a protein that mediates the migration of neural progenitors of the olfactory bulb and GnRH-producing neurons. These individuals have GnRH deficiency and variable combinations of anosmia or hyposmia, renal defects, and neurologic abnormalities including mirror movements. Proteins such as those involved in FGF and prokineticin signaling, and *KAL1*, which account for the great majority of Kallmann syndrome cases, interact with heparin sulfate glycosominoglycan compounds within the extracellular matrix in promoting GnRH neuronal migration. Mutations in the *FGFR1* gene cause an autosomal dominant form of hypogonadotropic hypogonadism that clinically resembles Kallmann syndrome; mutations in its putative ligand, the *FGF8* gene product have also been associated with IHH. Craniofacial tissues and olfactory ensheathing cells also play important roles in neurogenesis and migration of the GnRH neurons, and additional proteins that regulate these cell types may also be involved in the pathogenesis of Kallmann syndrome. The co-occurrence of tooth anomalies, cleft palate, craniofacial anomalies, pigmentation, and other neurological defects in patients with Kallmann Syndrome suggest that the syndrome may be a part of the spectrum of neurocristopathies.

TABLE 384-2

Causes of Congenital Hypogonadotropic Hypogonadism

Gene	Locus	Inheritance	Associated Features
A. Hypogonadotropic Hypogonadism due to GnRH Deficiency			
A1. GnRH Deficiency Associated with Hyposmia or Anosmia			
KAL1	Xp22	X-linked	Anosmia, renal agenesis, synkinesia, cleft lip/palate, oculomotor/visuospatial defects, gut malformations
NELF	9q34.3	AR	Anosmia, hypogonadotropic hypogonadism
FGF8	10q24	AR	Anosmia (some patients may be normosmic), skeletal abnormalities
FGFR1	8p11-p12	AD	Anosmia, cleft lip/palate, synkinesia, syndactyly
PROK2	3p21	AR	Anosmia/ sleep dysregulation
PROK2R	20p12.3	AR	Variable
CHD7	8q12.1		Anosmia, other features of CHARGE syndrome
FEZ1	8p22	AR	Anosmia, olfactory bulb aplasia
WDR11	10q26	AD	Anosmia
SOX10	22q13		Deafness
SEMA3A	7q21		Anosmia; some persons with mutations are normal
HS6ST1	2q14	complex	Anosmia
A2. GnRH Deficiency with Normal Sense of Smell			
GNRHR	4q21	AR	None
GnRH1	8p21	AR	None
KISS1R	19p13	AR	None
TAC3	12q13	AR	Microphallus, cryptorchidism, reversal of GnRH deficiency
TAC3R	4q25	AR	Microphallus, cryptorchidism, reversal of GnRH deficiency
LEPR	1p31	AR	Obesity
LEP	7q31	AR	Obesity
B. Hypogonadotropic Hypogonadism not due to GnRH Deficiency			
PC1	5q15-21	AR	Obesity, diabetes mellitus, ACTH deficiency
HESX1	3p21	AR AD	Septooptic dysplasia, CPHD Isolated GH insufficiency
LHX3	9q34	AR	CPHD (ACTH spared), cervical spine rigidity
PROP1	5q35	AR	CPHD (ACTH usually spared)

Gene	Locus	Inheritance	Associated Features
FSHβ	11p13	AR	↑ LH
LHβ	19q13	AR	↑ FSH
SF1 (NR5A1)	9p33	AD/AR	Primary adrenal failure, XY sex reversal

Abbreviations: ACTH, adrenocorticotropic hormone; AD, autosomal dominant; AR, autosomal recessive; CHARGE syndrome: eye coloboma, heart defects, choanal atresia, growth and developmental retardation, genitourinary anomalies, ear anomalies; CPHD, combined pituitary hormone deficiency; DAX1, *d*osage-sensitive sex-reversal, *a*drenal hypoplasia congenita, *X*-chromosome; FGFR1, fibroblast growth factor receptor 1; FSHβ, follicle-stimulating hormone β-subunit; GNRHR, gonadotropin-releasing hormone receptor; GPR54, G protein–coupled receptor 54; HESX1, homeo box gene expressed in embryonic stem cells 1; KAL1, Kallmann Syndrome Interval Gene 1, also known as anosmin 1; LEP, leptin; LEPR, leptin receptor; LHX3, LIM homeobox gene 3; LHβ, luteinizing hormone β-subunit; NELF, nasal embryonic LHRH factor; PC1, prohormone convertase 1; PROK2, prokineticin 2; PROP1, Prophet of Pit 1; SF1, steroidogenic factor 1.

Normosmic GnRH deficiency results from defects in pulsatile GnRH secretion, its regulation, or its action on the gonadotrope and has been associated with mutations in GnRHR, GNRH1, KISS1R, TAC3, TACR3, NROB1 (DAX1). Some mutations, such as those in PROK2, PROKR2, and CHD7 have been associated with both anosmic as well as normosmic form of IHH. *GnRH receptor* mutations, the most frequent identifiable cause of normosmic IHH, account for ~40% of autosomal recessive and 10% of sporadic cases of hypogonadotropic hypogonadism. These patients have decreased LH response to exogenous GnRH. Some receptor mutations alter GnRH binding affinity, allowing apparently normal responses to pharmacologic doses of exogenous GnRH, whereas other mutations may alter signal transduction downstream of hormone binding. Mutations of the *GnRH1* gene have also been reported in patients with hypogonadotropic hypogonadism, although they are rare. G protein–coupled receptor *KISS1R* (GPR54) and its cognate ligand, kisspeptin (*KISS1*), are important regulators of sexual maturation in primates. Recessive mutations in GPR54 cause gonadotropin deficiency without anosmia. Patients retain responsiveness to exogenous GnRH, suggesting an abnormality in the neural pathways controlling GnRH release. The genes encoding neurokinin B (TAC3), which is involved in preferential activation of GnRH release in early development, and its receptor (TAC3R) have been implicated in some families with normosmic IHH. X-linked hypogonadotropic hypogonadism also occurs in *adrenal hypoplasia congenita*, a disorder caused by mutations in the *DAX1* gene, which encodes a nuclear receptor in the adrenal gland and reproductive axis. Adrenal hypoplasia congenita is characterized by absent development of the adult zone of the adrenal cortex, leading to neonatal adrenal insufficiency. Puberty usually does not occur or is arrested, reflecting variable degrees of gonadotropin deficiency. Although sexual differentiation is normal, some patients have testicular dysgenesis and impaired spermatogenesis despite gonadotropin replacement. Less commonly, adrenal hypoplasia congenita, sex reversal, and hypogonadotropic hypogonadism can be caused by mutations of steroidogenic factor 1 (SF1). Rarely, recessive mutations in the *LHβ* or *FSHβ* genes have been described in patients with selective deficiencies of these gonadotropins.

A number of homeodomain transcription factors are involved in the development and differentiation of the specialized hormone-producing cells within the pituitary gland (Table 384-2). Patients with mutations of *PROP1* have combined pituitary hormone deficiency that includes GH, prolactin (PRL) thyroid-stimulating hormone (TSH), LH, and FSH, but not ACTH. *LHX3* mutations cause combined pituitary hormone deficiency in association with cervical spine rigidity. *HESX1* mutations cause septooptic dysplasia and combined pituitary hormone deficiency. Mutations of *ARNT1*, inherited as an autosomal recessive disorder, are associated with diabetes insipidus, ACTH deficiency, GH, LH, FSH deficiency, anterior pituitary hypoplasia, hypoplastic frontal and temporal lobes, thin corpus callosum, prominent forehead, and retrognathia. Patients with *SOX2* mutations can have gonadotropin deficiency, variable deficiencies of TSH and ACTH, pituitary hypoplasia, microphthalmia, and intellectual disability.

Prader-Willi syndrome is characterized by obesity, hypotonic musculature, mental retardation, hypogonadism, short stature, and small hands and feet. Prader-Willi syndrome is a genomic imprinting disorder caused by deletions of the proximal portion of paternally derived chromosome 15q11-15q13 region, which contains a bipartite imprinting center; uniparental disomy of the maternal alleles; or mutations of the genes/loci involved in imprinting **(Chap. 456)**. *Laurence-Moon syndrome* is an autosomal recessive disorder characterized by obesity, hypogonadism, mental retardation, polydactyly, and retinitis pigmentosa. Recessive mutations of leptin, or its receptor, cause severe obesity and pubertal arrest, apparently because of hypothalamic GnRH deficiency **(Chap. 394)**.

Acquired Hypogonadotropic Disorders

SEVERE ILLNESS, STRESS, MALNUTRITION, AND EXERCISE

These may cause reversible gonadotropin deficiency. Although gonadotropin deficiency and reproductive dysfunction are well documented in these conditions in women, men exhibit similar but less-pronounced responses. Unlike women, most male runners and other endurance athletes have normal gonadotropin and sex steroid levels, despite low body fat and frequent intensive exercise. Testosterone levels fall at the onset of illness and recover during recuperation. The magnitude of gonadotropin suppression generally correlates with the severity of illness. Although hypogonadotropic hypogonadism is the most common cause of androgen deficiency in patients with acute illness, some have elevated levels of LH and FSH, which suggest primary gonadal dysfunction. The pathophysiology of reproductive dysfunction during acute illness is unknown but likely involves a combination of cytokine and/or glucocorticoid effects. There is a high frequency of low testosterone levels in patients with chronic illnesses such as HIV infection, end-

stage renal disease, chronic obstructive lung disease, and many types of cancer and in patients receiving glucocorticoids. About 20% of HIV-infected men with low testosterone levels have elevated LH and FSH levels; these patients presumably have primary testicular dysfunction. The remaining 80% have either normal or low LH and FSH levels; these men have a central hypothalamic-pituitary defect or a dual defect involving both the testis and the hypothalamic-pituitary centers. Muscle wasting is common in chronic diseases associated with hypogonadism, which also leads to debility, poor quality of life, and adverse outcome of disease. There is great interest in exploring strategies that can reverse androgen deficiency or attenuate the sarcopenia associated with chronic illness.

Men using opioids for relief of cancer or noncancerous pain or because of addiction often have suppressed testosterone and LH levels and high prevalence of sexual dysfunction and osteoporosis; the degree of suppression is dose-related and particularly severe with long acting opioids such as methadone. Opioids suppress GnRH secretion and alter the sensitivity to feedback inhibition by gonadal steroids. Men who are heavy users of marijuana have decreased testosterone secretion and sperm production. The mechanism of marijuana-induced hypogonadism is decreased GnRH secretion. Gynecomastia observed in marijuana users can also be caused by plant estrogens in crude preparations. Androgen deprivation therapy in men with prostate cancer has been associated with increased risk of bone fractures, diabetes mellitus, cardiovascular events, fatigue, sexual dysfunction, tender gynecomastia, and poor quality of life.

OBESITY

In men with mild to moderate obesity, SHBG levels decrease in proportion to the degree of obesity, resulting in lower total testosterone levels. However, free testosterone levels usually remain within the normal range. SHBG production in the liver is inhibited by hepatic lipids, and by TNF-α and interleukin-1, but it is not affected by insulin. Thus, the low SHBG levels seen in obesity and diabetes are likely the result of low grade inflammation and the increased amount of hepatic lipids rather than high insulin levels. Estradiol levels are higher in obese men compared to healthy, nonobese controls, because of aromatization of testosterone to estradiol in adipose tissue. Weight loss is associated with reversal of these abnormalities including an increase in total and free testosterone levels and a decrease in estradiol levels. A subset of obese men with moderate to severe obesity may have a defect in the hypothalamic-pituitary axis as suggested by low free testosterone in the absence of elevated gonadotropins. Weight gain in adult men can accelerate the rate of age-related decline in testosterone levels.

HYPERPROLACTINEMIA

(See also Chap. 373) Elevated PRL levels are associated with hypogonadotropic hypogonadism. PRL inhibits hypothalamic GnRH secretion either directly or through modulation of tuberoinfundibular dopaminergic pathways. A PRL-secreting tumor may also destroy the surrounding gonadotropes by invasion or compression of the pituitary stalk. Treatment with dopamine agonists reverses gonadotropin deficiency, although there may be a delay relative to PRL suppression.

SELLAR MASS LESIONS

Neoplastic and nonneoplastic lesions in the hypothalamus or pituitary can directly or indirectly affect gonadotrope function. In adults, pituitary adenomas constitute the largest category of space-occupying lesions affecting gonadotropin and other pituitary hormone production. Pituitary adenomas that extend into the suprasellar region can impair GnRH secretion and mildly increase PRL secretion (usually <50 µg/L) because of impaired tonic inhibition by dopaminergic pathways. These tumors that cause hyperprolactinemia by stalk compression should be distinguished from prolactinomas, which typically are associated with higher PRL levels. The presence of diabetes insipidus suggests the possibility of a craniopharyngioma, infiltrative disorder, or other hypothalamic lesions **(Chap. 374)**.

HEMOCHROMATOSIS

(See also Chap. 407) Both the pituitary and testis can be affected by excessive iron deposition. However, the pituitary defect is the predominant lesion in most patients with hemochromatosis and hypogonadism. The diagnosis of hemochromatosis is suggested by the association of characteristic skin discoloration, hepatic enlargement or dysfunction, diabetes mellitus, arthritis, cardiac conduction defects, and hypogonadism.

PRIMARY TESTICULAR CAUSES OF HYPOGONADISM

Common causes of primary testicular dysfunction include Klinefelter syndrome, uncorrected cryptorchidism, cancer chemotherapy, radiation to the testes, trauma, torsion, infectious orchitis, HIV infection, anorchia syndrome, and myotonic dystrophy. Primary testicular disorders may be associated with impaired spermatogenesis, decreased androgen production, or both. **See Chap. 383 for disorders of testis development, androgen synthesis, and androgen action.**

Klinefelter Syndrome

(See also Chap. 383) Klinefelter syndrome is the most common chromosomal disorder associated with testicular dysfunction and male infertility. It occurs in about 1 in 600 live-born males. Azoospermia is the rule in men with Klinefelter syndrome who have the 47,XXY karyotype; however, men with mosaicism may have germ cells, especially at a younger age. The clinical phenotype of Klinefelter syndrome can be variable, possibly because of mosaicism, polymorphisms in AR gene, the parental origin of the X chromosome, X-linked copy number variations, gene-dosage effects in conjunction with X chromosome inactivation, variable testosterone levels, or other genetic factors. Testicular histology shows hyalinization of seminiferous tubules and absence of spermatogenesis. Although their function is impaired, the number of Leydig cells appears to increase. Testosterone is decreased and estradiol is increased, leading to clinical features of undervirilization and gynecomastia. Men with Klinefelter syndrome are at increased risk of systemic lupus erythematosus, Sjögren's syndrome, breast cancer, diabetes mellitus, osteoporosis, non-Hodgkin's lymphoma, and some types of lung cancer, and

reduced risk of prostate cancer. Periodic mammography for breast cancer surveillance is recommended for men with Klinefelter syndrome. Fertility can be achieved by intracytoplasmic injection of sperm retrieved surgically from the testes of men with Klinefelter syndrome, including some men with nonmosaic form of Klinefelter syndrome. The karyotypes 48,XXXY and 49,XXXXY are associated with a more severe phenotype, increased risk of congenital malformations, and lower intelligence than 47,XXY individuals.

Cryptorchidism

Cryptorchidism occurs when there is incomplete descent of the testis from the abdominal cavity into the scrotum. About 3% of full-term and 30% of premature male infants have at least one undescended testis at birth, but descent is usually complete by the first few weeks of life. The incidence of cryptorchidism is <1% by 9 months of age. Androgens regulate predominantly the inguinoscrotal descent of the testes through degeneration of the cranio-suspensory ligament and a shortening of the gubernaculums, respectively. Mutations in INSL3 and leucine-rich repeat family of G-protein-coupled receptor 8 (LGR8), which regulate transabdominal portion of testicular descent, have been found in some patients with cryptorchidism.

Cryptorchidism is associated with increased risk of malignancy, infertility, inguinal hernia, and torsion. Unilateral cryptorchidism, even when corrected before puberty, is associated with decreased sperm count, possibly reflecting unrecognized damage to the fully descended testis or other genetic factors. Epidemiologic, clinical, and molecular evidence supports the idea that cryptorchidism, hypospadias, impaired spermatogenesis, and testicular cancer may be causally related to common genetic and environment perturbations, and are components of the testicular dysgenesis syndrome.

Acquired Testicular Defects

Viral orchitis may be caused by the mumps virus, echovirus, lymphocytic choriomeningitis virus, and group B arboviruses. Orchitis occurs in as many as one-fourth of adult men with mumps; the orchitis is unilateral in about two-thirds and bilateral in the remainder. Orchitis usually develops a few days after the onset of parotitis but may precede it. The testis may return to normal size and function or undergo atrophy. Semen analysis returns to normal for three-fourths of men with unilateral involvement but normal for only one-third of men with bilateral orchitis. *Trauma*, including testicular torsion, can also cause secondary atrophy of the testes. The exposed position of the testes in the scrotum renders them susceptible to both thermal and physical trauma, particularly in men with hazardous occupations.

The testes are sensitive to *radiation damage*. Doses >200 mGy (20 rad) are associated with increased FSH and LH levels and damage to the spermatogonia. After ~800 mGy (80 rad), oligospermia or azoospermia develops, and higher doses may obliterate the germinal epithelium. Permanent androgen deficiency in adult men is uncommon after therapeutic radiation; however, most boys given direct testicular radiation therapy for acute lymphoblastic leukemia have permanently low testosterone levels. Sperm banking should be considered before patients undergo radiation treatment or chemotherapy.

Drugs interfere with testicular function by several mechanisms, including inhibition of testosterone synthesis (e.g., ketoconazole), blockade of androgen action (e.g., spironolactone), increased estrogen (e.g., marijuana), or direct inhibition of spermatogenesis (e.g., chemotherapy).

Combination chemotherapy for acute leukemia, Hodgkin's disease, and testicular and other cancers may impair Leydig cell function and cause infertility. The degree of gonadal dysfunction depends on the type of chemotherapeutic agent and the dose and duration of therapy. Because of high response rates and the young age of these men, infertility and androgen deficiency have emerged as important long-term complications of cancer chemotherapy. Cyclophosphamide and combination regimens containing procarbazine are particularly toxic to germ cells. Thus, 90% of men with Hodgkin's lymphoma receiving MOPP (mechlorethamine, oncovin, procarbazine, prednisone) therapy develop azoospermia or extreme oligozoospermia; newer regimens that do not include procarbazine, such as ABVD (adriamycin, bleomycin, vinblastine, dacarbazine), are less toxic to germ cells.

Alcohol, when consumed in excess for prolonged periods, decreases testosterone, independent of liver disease or malnutrition. Elevated estradiol and decreased testosterone levels may occur in men taking digitalis.

The occupational and recreational history should be carefully evaluated in all men with infertility because of the toxic effects of many *chemical agents* on spermatogenesis. Known environmental hazards include pesticides (e.g., vinclozolin, dicofol, atrazine), sewage contaminants (e.g., ethinyl estradiol in birth control pills, surfactants such as octylphenol, nonyphenol), plasticizers (e.g., pthalates), flame retardants (PCBs, polybrominated diphenol ethers), industrial pollutants (e.g., heavy metals cadmium and lead, dioxins, polycyclic aromatic hydrocarbons), microwaves and ultrasound. In some populations, sperm density is said to have declined by as much as 40% in the past 50 years. Environmental estrogens or antiandrogens may be partly responsible.

Testicular failure also occurs as a part of *polyglandular autoimmune insufficiency* **(Chap. 381)**. Sperm antibodies can cause isolated male infertility. In some instances, these antibodies are secondary phenomena resulting from duct obstruction or vasectomy. Granulomatous diseases can affect the testes, and testicular atrophy occurs in 10–20% of men with lepromatous leprosy because of direct tissue invasion by the mycobacteria. The tubules are involved initially, followed by endarteritis and destruction of Leydig cells.

Systemic disease can cause primary testis dysfunction in addition to suppressing gonadotropin production. In cirrhosis, a combined testicular and pituitary abnormality leads to decreased testosterone production independent of the direct toxic effects of ethanol. Impaired hepatic extraction of

adrenal androstenedione leads to extraglandular conversion to estrone and estradiol, which partially suppresses LH. Testicular atrophy and gynecomastia are present in approximately one-half of men with cirrhosis. In chronic renal failure, androgen synthesis and sperm production decrease despite elevated gonadotropins. The elevated LH level is due to reduced clearance, but it does not restore normal testosterone production. About one-fourth of men with renal failure have hyperprolactinemia. Improvement in testosterone production with hemodialysis is incomplete, but successful renal transplantation may return testicular function to normal. Testicular atrophy is present in one-third of men with sickle cell anemia. The defect may be at either the testicular or the hypothalamic-pituitary level. Sperm density can decrease temporarily after acute febrile illness in the absence of a change in testosterone production. Infertility in men with celiac disease is associated with a hormonal pattern typical of androgen resistance, namely elevated testosterone and LH levels.

Neurologic diseases associated with altered testicular function include myotonic dystrophy, spinobulbar muscular atrophy, and paraplegia. In myotonic dystrophy, small testes may be associated with impairment of both spermatogenesis and Leydig cell function. Spinobulbar muscular atrophy is caused by an expansion of the glutamine repeat sequences in the amino-terminal region of the AR; this expansion impairs function of the AR, but it is unclear how the alteration is related to the neurologic manifestations. Men with spinobulbar muscular atrophy often have undervirilization and infertility as a late manifestation. Spinal cord injury that causes paraplegia is often associated with low testosterone levels and may cause persistent defects in spermatogenesis; some patients retain the capacity for penile erection and ejaculation.

ANDROGEN INSENSITIVITY SYNDROMES

Mutations in the AR cause resistance to the action of testosterone and DHT. These X-linked mutations are associated with variable degrees of defective male phenotypic development and undervirilization **(Chap. 383)**. Although not technically hormone-insensitivity syndromes, two genetic disorders impair testosterone conversion to active sex steroids. Mutations in the *SRD5A2* gene, which encodes 5α-reductase type 2, prevent the conversion of testosterone to DHT, which is necessary for the normal development of the male external genitalia. Mutations in the *CYP19* gene, which encodes aromatase, prevent testosterone conversion to estradiol. Males with *CYP19* mutations have delayed epiphyseal fusion, tall stature, eunuchoid proportions, visceral adiposity, and osteoporosis, consistent with evidence from an estrogen receptor–deficient individual that these testosterone actions are mediated via estrogen.

GYNECOMASTIA

Gynecomastia refers to enlargement of the male breast. It is caused by excess estrogen action and is usually the result of an increased estrogen/androgen ratio. True gynecomastia is associated with glandular breast tissue that is >4 cm in diameter and often tender. Glandular tissue enlargement should be distinguished from excess adipose tissue: glandular tissue is firmer and contains fibrous-like cords. Gynecomastia occurs as a normal physiologic phenomenon in the newborn (due to transplacental transfer of maternal and placental estrogens), during puberty (high estrogen to androgen ratio in early stages of puberty), and with aging (increased fat tissue and increased aromatase activity along with the age-related decline in testosterone levels), but it can also result from pathologic conditions associated with androgen deficiency or estrogen excess. The prevalence of gynecomastia increases with age and body mass index (BMI), likely because of increased aromatase activity in adipose tissue. Medications that alter androgen metabolism or action may also cause gynecomastia. The relative risk of breast cancer is increased in men with gynecomastia, although the absolute risk is relatively small.

PATHOLOGIC GYNECOMASTIA

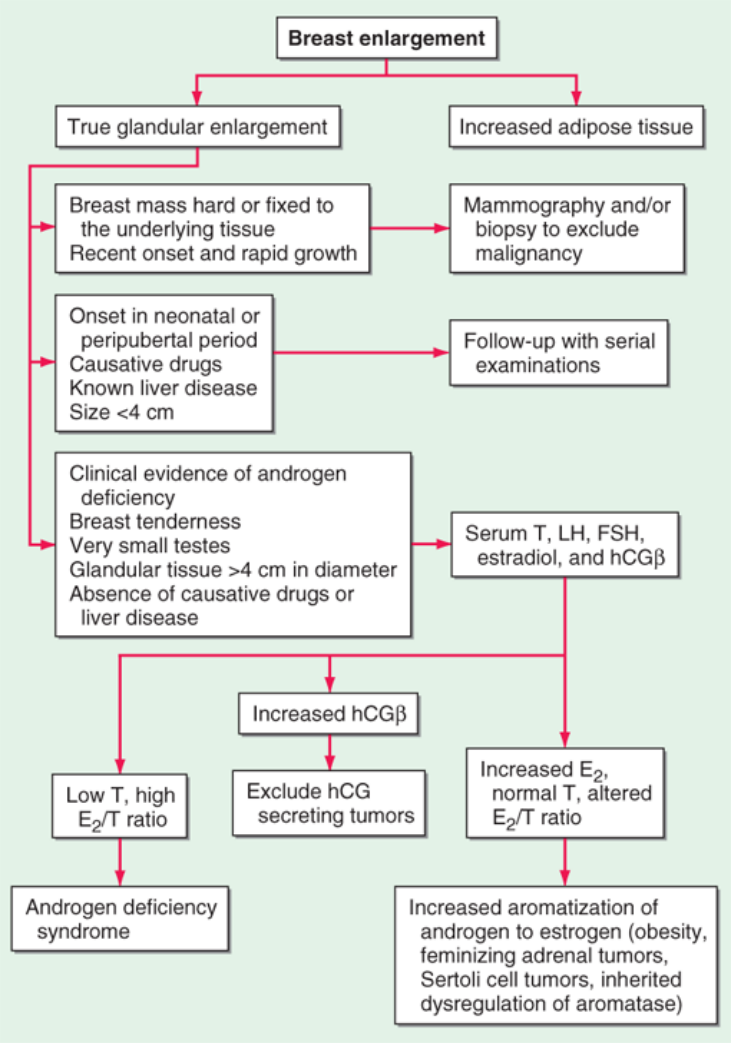
Any cause of *androgen deficiency* can lead to gynecomastia, reflecting an increased estrogen/androgen ratio, as estrogen synthesis still occurs by aromatization of residual adrenal and gonadal androgens. Gynecomastia is a characteristic feature of Klinefelter syndrome **(Chap. 383)**. *Androgen insensitivity* disorders also cause gynecomastia. *Excess estrogen production* may be caused by tumors, including Sertoli cell tumors in isolation or in association with Peutz-Jegher syndrome or Carney complex. Tumors that produce hCG, including some testicular tumors, stimulate Leydig cell estrogen synthesis. *Increased conversion of androgens to estrogens* can be a result of increased availability of substrate (androstenedione) for extraglandular estrogen formation (CAH, hyperthyroidism, and most feminizing adrenal tumors) or to diminished catabolism of androstenedione (liver disease) so that estrogen precursors are shunted to aromatase in peripheral sites. Obesity is associated with increased aromatization of androgen precursors to estrogens. Extraglandular aromatase activity can also be increased in tumors of the liver or adrenal gland or rarely as an inherited disorder. Several families with *increased peripheral aromatase activity* inherited as an autosomal dominant or as an X-linked disorder have been described. In some families with this disorder, an inversion in chromosome 15q21.2-3 causes the CYP19 gene to be activated by the regulatory elements of contiguous genes resulting in excessive estrogen production in the fat and other extragonadal tissues. The familial aromatase excess syndrome due to CYP19 mutation or chromosomal rearrangement is characterized by pre- or peripubertal onset of gynecomastia, advanced bone age, short adult height due to premature epiphyseal closure, and hypogonadotropic hypogonadism. *Drugs* can cause gynecomastia by acting directly as estrogenic substances (e.g., oral contraceptives, phytoestrogens, digitalis), inhibiting androgen synthesis (e.g., ketoconazole), or action (e.g., spironolactone); for many drugs, such as cimetidine, imatinib, or some antiretroviral drugs for HIV, the precise mechanism is unknown.

Because up to two-thirds of pubertal boys and about half of hospitalized men have palpable glandular tissue that is benign, detailed investigation or intervention is not indicated in all men presenting with gynecomastia **(Fig. 384-5)**. In addition to the extent of gynecomastia, recent onset, rapid growth, tender tissue, and occurrence in a lean subject should prompt more extensive evaluation. This should include a careful drug history, measurement and

examination of the testes, assessment of virilization, evaluation of liver function, and hormonal measurements including testosterone, estradiol, and androstenedione, LH, and hCG. A karyotype should be obtained in men with very small testes to exclude Klinefelter syndrome. In spite of extensive evaluation, the etiology is established in fewer than one-half of patients.

FIGURE 384-5

Evaluation of gynecomastia. E_2, 17β-estradiol; FSH, follicle-stimulating hormones; hCGβ, human chorionic gonadotropin β; LH, luteinizing hormone; T, testosterone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Gynecomastia

When the primary cause can be identified and corrected shortly after the onset of gynecomastia, breast enlargement usually subsides over several months. However, if gynecomastia is of long duration, surgery is the most effective therapy. Indications for surgery include severe psychological and/or cosmetic problems, continued growth or tenderness, or suspected malignancy. In patients who have painful gynecomastia and in whom surgery cannot be performed, treatment with antiestrogens such as tamoxifen (20 mg/d) can reduce pain and breast tissue size in over half the patients. Estrogen receptor antagonists, tamoxifen and raloxifen, have been reported in small trials to reduce breast size in men with pubertal gynecomastia, although complete regression of breast enlargement is unusual with the use of estrogen receptor antagonists. Aromatase inhibitors can be effective in the early proliferative phase of the disorder. However, in a randomized trial in men with established gynecomastia, anastrozole proved no more effective than placebo in reducing breast size. Tamoxifen is effective in prevention and treatment of breast enlargement and breast pain in men with prostate cancer who are receiving anti-androgen therapy.

AGING-RELATED CHANGES IN MALE REPRODUCTIVE FUNCTION

A number of cross-sectional and longitudinal studies (e.g., The Baltimore Longitudinal Study of Aging, the Framingham Heart Study, the Massachusetts Male Aging Study, and the European Male Aging Study [EMAS]) have established that testosterone concentrations decrease with advancing age. This age-related decline starts in the third decade of life and progresses slowly; the rate of decline in testosterone concentrations is greater in obese men, in men with chronic illness, and in those taking medications. Because SHBG concentrations are higher in older men than in younger men, free or bioavailable testosterone concentrations decline with aging to a greater extent than total testosterone concentrations. The age-related decline in testosterone is due to defects at all levels of the hypothalamic-pituitary-testicular axis: pulsatile GnRH secretion is attenuated, LH response to GnRH is reduced, and testicular response to LH is impaired. However, the gradual rise of LH with aging suggests that testis dysfunction is the main cause of declining androgen levels. The term *andropause* has been used to denote age-related decline in testosterone concentrations; this term is a misnomer because there is no discrete time when testosterone concentrations decline abruptly.

Several epidemiologic studies, such as the Framingham Heart Study, the EMAS, and the Study of Osteoporotic Fractures in Men (MrOS) that used mass spectrometry for measuring testosterone levels have reported ~10% prevalence of low testosterone levels in middle-aged and older men; the prevalence of unequivocally low testosterone and sexual symptoms in men aged 40–70 years in the EMAS was 2.1%, and increased with age from 0.1% for men aged 40–49 years of age to 5.1% for those aged 70–79 years. The age-related decline in testosterone should be distinguished from classical hypogonadism due to diseases of the testes, the pituitary, and the hypothalamus. Low total and bioavailable testosterone concentrations have been associated with decreased appendicular skeletal muscle mass and strength, decreased self-reported physical function, higher visceral fat mass, insulin resistance, and increased risk of coronary artery disease and mortality. An analysis of signs and symptoms in older men in the EMAS revealed a syndromic association of sexual symptoms with total testosterone levels below 320 ng/dL and free testosterone levels below 64 pg/mL in community-dwelling older men.

A series of placebo-controlled testosterone trials have provided important information about the efficacy of testosterone in improving outcomes in older men. Testosterone replacement in older men, aged ≥65, with sexual symptoms improved sexual activity, sexual desire, and erectile function more than placebo. Testosterone replacement did not improve fatigue or cognitive function, and had only a small effect on mood and mobility. Among older men with low testosterone and age-associated memory impairment, testosterone replacement did not improve memory or other measures of cognition relative to placebo. Testosterone replacement was associated with significantly greater increase in vertebral as well as femoral volumetric bone mineral density and estimated bone strength relative to placebo. Testosterone replacement was associated with a greater increase in hemoglobin levels and corrected anemia in a greater proportion of men who had unexplained anemia of aging. Testosterone administration was associated with a significantly greater increase in coronary artery noncalcified plaque volume, as measured by coronary artery computerized tomography angiography. Neither the testosterone trials nor a randomized trial of the effects of testosterone on atherosclerosis progression in aging men (TEAAM Trial) with low or low normal testosterone levels found significant differences between testosterone and placebo arms in the rates of change in either the coronary artery calcium scores or the common carotid artery intima-media thickness. Neither of the trials were long enough or large enough to determine the effects of testosterone replacement therapy on prostate or major adverse cardiovascular events. In systematic reviews of randomized controlled trials, testosterone therapy of healthy older men with low or low-normal testosterone levels was associated with greater increments in lean body mass, grip strength, and self-reported physical function than that associated with placebo. Testosterone therapy has not been shown to improve clinical depression, fracture risk, progression to dementia, progression from prediabetes to diabetes, or response to phosphodiesterase inhibitors in older men.

The long-term risks of testosterone therapy remain largely unknown. While there is no evidence that testosterone causes prostate cancer, there is concern that testosterone therapy might cause subclinical prostate cancers to grow. Testosterone therapy is associated with increased risk of detection of prostate events.

The data relating cardiovascular disease (CVD) and venous thromboembolic (VTE) risk with the use of testosterone supplementation in men with low testosterone levels and hypogonadal symptoms are few and inconclusive. The relationship of testosterone and cardiovascular events in cross-sectional and prospective cohort studies has been inconsistent. A small number of epidemiologic studies have reported an inverse relationship between testosterone concentrations and common carotid artery intima-media thickness. Low testosterone level has been associated with increased risk of all-cause mortality, especially cardiovascular mortality. It is possible that testosterone is a marker of health; older men with multiple co-morbid conditions who are at increased risk of death may have low testosterone levels as a result of comorbid conditions.

Most meta-analyses have not shown a statistically significant association between testosterone and cardiovascular events, major adverse cardiovascular events, or deaths. No adequately powered randomized trials have been conducted to determine the effects of testosterone replacement in major adverse cardiovascular events. Thus, there are insufficient data to establish a causal link between testosterone therapy and CV events.

Population screening of all older men for low testosterone levels is not recommended, and testing should be restricted to men who have symptoms or signs attributable to androgen deficiency. Testosterone therapy is not recommended for all older men with low testosterone levels. In older men with significant symptoms of androgen deficiency who have unequivocally low testosterone levels, testosterone therapy may be considered on an individualized basis and should be instituted after careful discussion of the risks and benefits (see "Testosterone Replacement," below).

Testicular morphology, semen production, and fertility are maintained up to a very old age in men. Although concern has been expressed about age-related increases in germ cell mutations and impairment of DNA repair mechanisms, there is no clear evidence that the frequency of chromosomal aneuploidy is increased in the sperm of older men. However, the incidence of autosomal dominant diseases, such as achondroplasia, polyposis coli, Marfan syndrome, and Apert syndrome, increases in the offspring of men who are advanced in age, consistent with transmission of sporadic missense

mutations. Advanced paternal age may be associated with increased rates of de novo mutations, which may contribute to an increased risk of neurodevelopmental diseases such as schizophrenia and autism. The somatic mutations in male germ cells that enhance the proliferation of germ cells could lead to within-testis expansion of mutant clonal lines, thus favoring the propagation of germ cells carrying these pathogenic mutations, and increasing the risk of mutations in the offspring of older fathers (the “selfish spermatogonial selection” hypothesis).

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

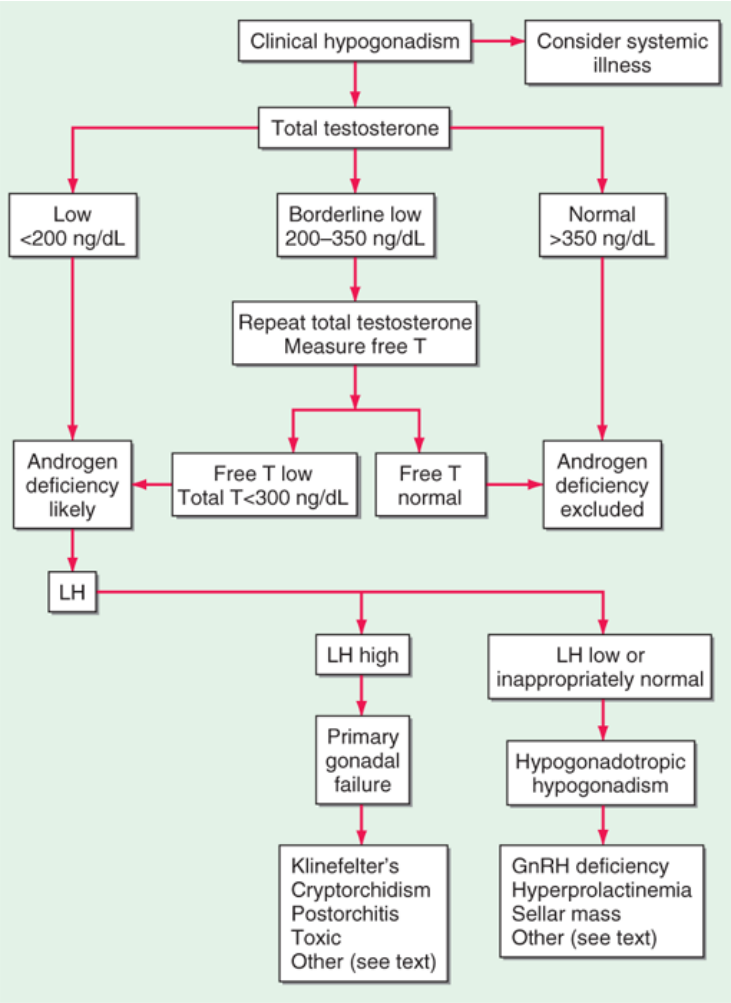
Androgen Deficiency

Hypogonadism is often characterized by decreased sex drive, reduced frequency of sexual activity, inability to maintain erections, reduced beard growth, loss of muscle mass, decreased testicular size, and gynecomastia. Erectile dysfunction and androgen deficiency are two distinct clinical disorders that can co-exist in middle-aged and older men. Less than 10% of patients with erectile dysfunction have testosterone deficiency. Thus, it is useful to evaluate men presenting with erectile dysfunction for androgen deficiency. Except when extreme, these clinical features of androgen deficiency may be difficult to distinguish from changes that occur with normal aging. Moreover, androgen deficiency may develop gradually. When symptoms or clinical features suggest possible androgen deficiency, the laboratory evaluation is initiated by the measurement of total testosterone, preferably in the morning using a reliable assay, such as liquid chromatography tandem mass spectrometry (LC-MS/MS) that has been calibrated to an international testosterone standard (Fig. 384-6). A consistently low total testosterone level >264 ng/dL measured by an LC-MS/MS assay in a Centers for Disease Control (CDC)—certified laboratory, in association with symptoms, is evidence of testosterone deficiency. An early-morning testosterone level >400 ng/dL makes the diagnosis of androgen deficiency unlikely. In men with testosterone levels between 200 and 400 ng/dL, the total testosterone level should be repeated and a free testosterone level should be measured. In older men and in patients with other clinical states that are associated with alterations in SHBG levels, a direct measurement of free testosterone level by equilibrium dialysis can be useful in unmasking testosterone deficiency.

When androgen deficiency has been confirmed by the consistently low testosterone concentrations, LH should be measured to classify the patient as having primary (high LH) or secondary (low or inappropriately normal LH) hypogonadism. An elevated LH level indicates that the defect is at the testicular level. Common causes of primary testicular failure include Klinefelter syndrome, HIV infection, uncorrected cryptorchidism, cancer chemotherapeutic agents, radiation, surgical orchiectomy, or prior infectious orchitis. Unless causes of primary testicular failure are known, a karyotype should be performed in men with low testosterone and elevated LH to diagnose Klinefelter syndrome. Men who have a low testosterone but “inappropriately normal” or low LH levels have secondary hypogonadism; their defect resides at the hypothalamic-pituitary level. Common causes of acquired secondary hypogonadism include space-occupying lesions of the sella, hyperprolactinemia, chronic illness, hemochromatosis, excessive exercise, and the use of anabolic-androgenic steroids, opiates, marijuana, glucocorticoids, and alcohol. Measurement of PRL and MRI scan of the hypothalamic-pituitary region can help exclude the presence of a space-occupying lesion. Patients in whom known causes of hypogonadotropic hypogonadism have been excluded are classified as having IHH. It is not unusual for congenital causes of hypogonadotropic hypogonadism, such as Kallmann syndrome, to be diagnosed in young adults.

FIGURE 384-6

Evaluation of hypogonadism. GnRH, gonadotropin-releasing hormone; LH, luteinizing hormone; T, testosterone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Androgen Deficiency

GONADOTROPINS

Gonadotropin therapy is used to establish or restore fertility in patients with gonadotropin deficiency of any cause. Several gonadotropin preparations are available. Human menopausal gonadotropin (hMG; purified from the urine of postmenopausal women) contains 75 IU FSH and 75 IU LH per vial. hCG (purified from the urine of pregnant women) has little FSH activity and resembles LH in its ability to stimulate testosterone production by Leydig cells. Recombinant LH is also available. Treatment is usually begun with hCG alone, and hMG is added later to promote the FSH-dependent stages of spermatid development. Recombinant human FSH (hFSH) is available and is indistinguishable from purified urinary hFSH in its biologic activity and pharmacokinetics in vitro and in vivo, although the mature β subunit of recombinant hFSH has seven fewer amino acids. Recombinant hFSH is available in ampoules containing 75 IU (~7.5 μg FSH), which accounts for >99% of protein content. Once spermatogenesis is restored using combined FSH and LH therapy, hCG alone is often sufficient to maintain spermatogenesis.

Although a variety of treatment regimens are used, 1000 IU of hCG or recombinant human LH (rhLH) administered intramuscularly three times weekly is a reasonable starting dose. Testosterone levels should be measured 6–8 weeks later and 48–72 h after the hCG or rhLH injection; the hCG/rhLH dose should be adjusted to achieve testosterone levels in the mid-normal range. Sperm counts should be monitored on a monthly basis. It may take several months for spermatogenesis to be restored; therefore, it is important to forewarn patients about the potential length and expense of the treatment and to provide conservative estimates of success rates. If testosterone levels are in the mid-normal range but the sperm concentrations are low after 6 months of therapy with hCG alone, FSH should be added. This can be done by using hMG, highly purified urinary hFSH, or recombinant hFSH. The selection of FSH dose is empirical. A common practice is to start with the addition of 75 IU FSH three times each week in conjunction with the hCG/rhLH injections. If sperm densities are still low after 3 months of combined treatment, the FSH dose should be increased to 150 IU. Occasionally, it may take ≥18–24 months for spermatogenesis to be restored.

The two best predictors of success using gonadotropin therapy in hypogonadotropic men are testicular volume at presentation and time of onset of gonadotropin deficiency. In general, men with testicular volumes >8 mL have better response rates than those who have testicular volumes <4 mL. Patients who become hypogonadotropic after puberty experience higher success rates than those who have never undergone pubertal changes. Spermatogenesis can usually be reinitiated by hCG alone, with high rates of success for men with postpubertal onset of hypogonadotropism. The presence of a primary testicular abnormality, such as cryptorchidism, will attenuate testicular response to gonadotropin therapy. Prior androgen therapy does not preclude subsequent response to gonadotropin therapy, although some studies suggest that it may attenuate response to subsequent gonadotropin therapy.

TESTOSTERONE REPLACEMENT

Androgen therapy is indicated to restore testosterone levels to normal to correct features of androgen deficiency in men with organic hypogonadism due to known diseases of the testes, pituitary, and the hypothalamus. Testosterone replacement induces secondary sex characteristics, improves libido and overall sexual activity; increases lean muscle mass, hemoglobin and hematocrit, and bone mineral density, and decreases fat mass. The benefits of testosterone replacement therapy have only been proven in men who have documented symptomatic androgen deficiency, as demonstrated by testosterone levels that are well below the lower limit of normal.

Testosterone is available in a variety of formulations with distinct pharmacokinetics **(Table 384-3)**. Testosterone serves as a prohormone and is converted to 17β-estradiol by aromatase and to 5α-dihydrotestosterone by steroid 5α-reductase. Therefore, when evaluating testosterone formulations, it is important to consider whether the formulation being used can achieve physiologic estradiol and DHT concentrations, in addition to normal testosterone concentrations. The current recommendation is to restore testosterone levels to the mid-normal range.

Oral Derivatives of Testosterone

Testosterone is well-absorbed after oral administration but is quickly degraded during the first pass through the liver. Therefore, it is difficult to achieve sustained blood levels of testosterone after oral administration of crystalline testosterone. 17α-Alkylated derivatives of testosterone (e.g., 17α-methyl testosterone, oxandrolone, fluoxymesterone) are relatively resistant to hepatic degradation and can be administered orally; however, because of the potential for hepatotoxicity, including cholestatic jaundice, peliosis, and hepatoma, these formulations should not be used for testosterone replacement. Hereditary angioedema due to C1 esterase deficiency is the only exception to this general recommendation; in this condition, oral 17α-alkylated androgens are useful because they stimulate hepatic synthesis of the C1 esterase inhibitor.

Injectable Forms of Testosterone

The esterification of testosterone at the 17β-hydroxy position makes the molecule hydrophobic and extends its duration of action. The slow release of testosterone ester from an oily depot in the muscle accounts for its extended duration of action. The longer the side chain, the greater the hydrophobicity of the ester and longer the duration of action. Thus, testosterone enanthate, cypionate, and undecanoate with longer side chains have longer duration of action than testosterone propionate. Within 24 h after intramuscular administration of 200 mg testosterone enanthate or cypionate, testosterone levels rise into the high-normal or supraphysiologic range and then gradually decline into the hypogonadal range over the next 2 weeks. A bimonthly regimen of testosterone enanthate or cypionate therefore results in peaks and troughs in testosterone levels that may be accompanied by changes in a patient's mood, sexual desire, and energy level; weekly administration of testosterone enanthate or cypionate can reduce these variations in testosterone levels during the dosing interval. The kinetics of testosterone enanthate and cypionate are similar. Estradiol and DHT levels are normal if testosterone replacement is physiologic.

A long-acting testosterone undecanoate in oil, administered at an initial priming dose of 750 mg intramuscularly followed by a second dose of 750 mg 4 weeks later, and then at a maintenance dose of 750 mg every 10 weeks, maintains serum testosterone, estradiol, and DHT in the normal male range and corrects symptoms of androgen deficiency in a majority of treated men. However, its relative drawback is the large injection volume and cough in a small proportion of patients.

Transdermal Testosterone Patch

The nongenital testosterone patch, when applied in an appropriate dose, can normalize testosterone, DHT, and estradiol levels 4–12 h after application. Sexual function and well-being are restored in androgen-deficient men treated with the nongenital patch. One 4-mg patch may not be sufficient to increase testosterone into the mid-normal male range in all hypogonadal men; many patients may need two 4-mg patches daily to achieve the targeted testosterone concentrations. The use of testosterone patches may be associated with skin irritation in some individuals.

Testosterone Gel

Several transdermal testosterone gels, Androgel, Testim, Fortesta, and Axiron, and some generic versions, when applied topically to the skin in appropriate doses (Table 384-3), can maintain total and free testosterone concentrations in the normal range in hypogonadal men. The current recommendations are to begin with an initial FDA-recommended dose and adjust the dose based on testosterone levels. The advantages of the testosterone gel include the ease of application. A major concern is the potential for inadvertent transfer of the gel to a sexual partner or to children who may come in close contact with the patient. The ratio of DHT to testosterone concentrations is higher in men treated with the testosterone gel than in healthy men. Also, there is considerable intra- and inter-individual variation in serum testosterone levels in men treated with the transdermal gel due to variations in transdermal absorption and plasma clearance of testosterone. Therefore, monitoring of serum testosterone levels and multiple dose adjustments may be required to achieve and maintain testosterone levels in the target range.

Buccal Adhesive Testosterone

A buccal testosterone tablet, which adheres to the buccal mucosa and releases testosterone as it is slowly dissolved, has been approved. After twice-daily application of 30-mg tablets, serum testosterone levels are maintained within the normal male range in a majority of treated hypogonadal men. The adverse effects include buccal ulceration and gum problems in a few subjects. The effects of food and brushing on absorption have not been studied in detail.

Pellets of crystalline testosterone can be inserted in the subcutaneous tissue through a small skin incision. Testosterone is released by surface erosion of the implant and absorbed into the systemic circulation, and testosterone levels can be maintained in the normal range for 3–4 months. Potential drawbacks include incising the skin for insertion and removal, and spontaneous extrusions and fibrosis at the site of the implant.

Testosterone Formulations Not Available in the United States

Testosterone undecanoate, when administered orally in oleic acid, is absorbed preferentially through the lymphatics into the systemic circulation and is spared the first-pass degradation in the liver. Doses of 40–80 mg orally, two or three times daily, are typically used. However, the clinical responses are variable and suboptimal. DHT-to-testosterone ratios are higher in hypogonadal men treated with oral testosterone undecanoate, as compared to eugonadal men.

An intranasal testosterone gel is now available as a metered dose pump and is administered typically at a starting dose of 11 mg testosterone in the form of 2 pump actuations, one in each nostril three times daily. Formulation-specific adverse effects include rhinorrhea, nasal discomfort, epistaxis, nasopharyngitis, and nasal scab.

Novel Androgen Formulations

A number of androgen formulations with better pharmacokinetics or more selective activity profiles are under development. Initial clinical trials have demonstrated the feasibility of administering testosterone by the sublingual, oral, or buccal routes. Long-acting biodegradable microsphere formulations have also been investigated. 7α-Methyl-19-nortestosterone is an androgen that cannot be 5α-reduced; therefore, compared to testosterone, it has relatively greater agonist activity in muscle and gonadotropin suppression but lesser activity on the prostate.

Selective Androgen Receptor Modulators (SARMs) are a class of AR ligands that bind the AR and display tissue-selective actions. A number of nonsteroidal SARMs that act as agonists on the muscle and bone and which spare the prostate to varying degrees have advanced to phase III human trials. Nonsteroidal SARMs do not serve as substrates for either the steroid 5-α reductase or the CYP19aromatase. SARM binding to AR induces specific conformational changes in the AR protein, which then modulates protein-protein interactions between AR and its coregulators, resulting in tissue-specific regulation of gene expression. SARMs that are strong agonists for the muscle, bone, and sexual function, and antagonists for the prostate may be valuable in treating men with prostate cancer, who are receiving androgen deprivation therapy.

Pharmacologic Uses of Androgens

Androgens and selective AR modulators are being evaluated as anabolic therapies for functional limitations associated with aging and chronic illness. Testosterone supplementation increases skeletal muscle mass, maximal voluntary strength, and muscle power in healthy men, hypogonadal men, older men with low testosterone levels, HIV-infected men with weight loss, and men receiving glucocorticoids. These anabolic effects of testosterone are related to testosterone dose and circulating concentrations. Systematic reviews have confirmed that testosterone therapy of HIV-infected men with weight loss promotes improvements in body weight, lean body mass, muscle strength, and depression indices, leading to the recommendation that testosterone be considered as an adjunctive therapy in HIV-infected men who are experiencing unexplained weight loss and who have low testosterone levels. It is unknown whether testosterone therapy of older men with functional limitations is safe and effective in improving physical function, vitality, and health-related quality of life, and reducing disability. Concerns about potential adverse effects of testosterone on prostate and cardiovascular event rates have encouraged the development of selective AR modulators that are preferentially anabolic and spare the prostate.

Testosterone administration induces hypertrophy of both type 1 and 2 fibers and increases satellite cell (muscle progenitor cells) and myonuclear number. Androgens promote the differentiation of mesenchymal, multipotent progenitor cells into the myogenic lineage and inhibit their differentiation into the adipogenic lineage. Testosterone binding to AR promotes the association of liganded AR with β-catenin and its translocation into the nucleus where it binds TCF-4 and activates Wnt-target genes, including follistatin, which blocks signaling through the TGFβ pathway, thereby promoting myogenic differentiation of muscle progenitor cells. Testosterone may have additional effects on satellite cell replication and muscle protein synthesis, which may contribute to an increase in skeletal muscle mass.

Other indications for androgen therapy are in selected patients with anemia due to bone marrow failure (an indication largely supplanted by erythropoietin) or for hereditary angioedema.

Male Hormonal Contraception Based on Combined Administration of Testosterone and Gonadotropin Inhibitors

Supraphysiologic doses of testosterone (200 mg testosterone enanthate weekly) suppress LH and FSH secretion and induce azoospermia in 50% of Caucasian men and >95% of Chinese men. The WHO-supported multicenter efficacy trials have demonstrated that suppression of spermatogenesis to azoospermia or severe oligozoozspermia (<3 million/mL) by administration of supraphysiologic doses of testosterone enanthate to men results in highly effective contraception. Because of concern about long-term adverse effects of supraphysiologic testosterone doses, regimens that combine other

gonadotropin inhibitors, such as GnRH antagonists and progestins with replacement doses of testosterone, have been investigated. Regimens containing an androgen plus a progestin such as depo medroxyprogesterone acetate, etonogestrelm, or norethisterone enanthate have been highly effective in inducing azoospermia or severe oligozoospermia (sperm density <1 million/mL) in nearly 99% of treated men over a 1-year period. The combined regimens of testosterone plus a progestin have been associated with weight gain, acne, mood changes including depressed mood, libido changes, and decreased plasma high-density lipoprotein (HDL) cholesterol and their long-term safety has not been demonstrated. One such trial of a combined regimen of testosterone undecanoate plus norethisterone enanthate was stopped early due to adverse events. Selective AR modulators that are more potent inhibitors of gonadotropins than testosterone and spare the prostate hold promise for their contraceptive potential.

Recommended Regimens for Androgen Replacement

Testosterone esters are administered typically at doses of 75–100 mg intramuscularly every week, or 150–200 mg every 2 weeks. Testosterone undecanoate is administered at an initial dose of 750 mg followed 4 weeks later by a second injection of 750 mg and then 750 mg every 10 weeks. Testosterone gels are typically applied over a covered area of skin at initial doses that vary with the formulation. The patients should wash their hands after gel application and keep the area of gel application covered with clothing to minimize the risk of gel transfer to another person. One or two 4-mg nongenital testosterone patches are applied daily over the skin of the back, thigh, or upper arm away from pressure areas. Bioadhesive buccal testosterone tablets at a dose of 30 mg are applied twice daily on the buccal mucosa. Intranasal testosterone is administered as a spray in each nostril 3 times a day (33 mg/day).

Establishing Efficacy of Testosterone Replacement Therapy

Because a clinically useful marker of androgen action is not available, correction of symptoms, induction and maintenance of secondary sex characteristics, and restoration of testosterone levels into the mid-normal range remain the goal of therapy. Measurements of LH and FSH are not useful in assessing the adequacy of testosterone replacement. Testosterone should be measured 3 months after initiating therapy to assess adequacy of therapy. There is substantial inter-individual variability in serum testosterone levels, especially with transdermal gels, presumably due to genetic differences in testosterone clearance and substantial variation in transdermal absorption. In patients who are treated with testosterone enanthate or cypionate, testosterone levels should be 350–600 ng/dL 1 week after the injection. If testosterone levels are outside this range, adjustments should be made either in the dose or in the interval between injections. In men on transdermal patch, gel, or buccal testosterone therapy, testosterone levels should be in the mid-normal range (400–750 ng/dL) 4–12 h after application. If testosterone levels are outside this range, the dose should be adjusted. Multiple dose adjustments are often necessary to achieve testosterone levels in the desired therapeutic range.

Restoration of sexual function, induction and maintenance of secondary sex characteristics, well-being, and maintenance of muscle and bone health are important objectives of testosterone replacement therapy. The patient should be asked about sexual desire and activity, the presence of early morning erections, and the ability to achieve and maintain erections adequate for sexual intercourse. The hair growth in response to androgen replacement is variable and depends on ethnicity. Hypogonadal men with prepubertal onset of androgen deficiency who begin testosterone therapy in their late twenties or thirties may find it difficult to adjust to their newly found sexuality and may benefit from counseling. If the patient has a sexual partner, the partner should be included in counseling because of the dramatic physical and sexual changes that occur with androgen treatment.

Contraindications for Androgen Administration

Testosterone administration is contraindicated in men with prostate or breast cancer **(Table 384-4)**. Testosterone therapy should not be administered without further urologic evaluation to men with a palpable prostate nodule or induration, or prostate-specific antigen >3 ng/mL, or with severe lower urinary tract symptoms (American Urological Association lower urinary tract symptom score >19). Testosterone replacement should not be administered to men with baseline hematocrit ≥50%, severe untreated obstructive sleep apnea, uncontrolled or poorly controlled congestive heart failure, or to men with myocardial infarction, stroke, or acute coronary syndrome in the preceding 3 months.

Monitoring Potential Adverse Experiences

The clinical effectiveness and safety of testosterone replacement therapy should be assessed 3–6 months after initiating testosterone therapy and annually thereafter **(Table 384-5)**. Potential adverse effects include acne, oiliness of skin, erythrocytosis, breast tenderness and enlargement, leg edema, and increased risk of detection of prostate events. In addition, there may be formulation-specific adverse effects such as skin irritation with transdermal patch; risk of gel transfer to a sexual partner with testosterone gels; buccal ulceration and gum problems with buccal testosterone; pain and mood fluctuation with injectable testosterone esters; cough and injection site pain with long-acting testosterone undecanoate; and, nasal irritation, epistaxis, and nasal scab with intranasal formulation.

Hemoglobin Levels

Administration of testosterone to androgen-deficient men is typically associated with a ~3% increase in hemoglobin levels, due to increased erythropoiesis, stimulation of erythropoietin, suppression of hepcidin, and increased iron availability for erythropoiesis. The magnitude of hemoglobin increase during testosterone therapy is greater in older men than younger men, and in men who have sleep apnea, a significant smoking history, or chronic obstructive lung disease, or who live at high altitude. The frequency of erythrocytosis is higher in hypogonadal men treated with injectable testosterone esters than in those treated with transdermal formulations, presumably due to the higher testosterone dose delivered by the typical regimens of testosterone esters. Erythrocytosis is the most frequent adverse event reported in testosterone trials in middle-aged and older men and is

also the most frequent cause of treatment discontinuation in these trials. If hematocrit rises above 54%, testosterone therapy should be stopped until hematocrit has fallen to <50%. After evaluation of the patient for hypoxia and sleep apnea, testosterone therapy may be reinitiated at a lower dose.

Prostate and Serum PSA Levels

Testosterone replacement therapy increases prostate volume to the size seen in age-matched controls but does not increase prostate volume beyond that expected for age. There is no evidence that testosterone therapy causes prostate cancer. However, androgen administration can exacerbate preexisting metastatic prostate cancer. Many older men harbor microscopic foci of cancer in their prostates. It is not known whether long-term testosterone administration will induce these microscopic foci to grow into clinically significant cancers.

PSA levels are lower in testosterone-deficient men and are restored to normal after testosterone replacement. There is considerable test-retest variability in PSA measurements. Increments in PSA levels after testosterone supplementation in androgen-deficient men are generally <0.5 ng/mL, and increments >1.0 ng/mL over a 3–6-month period are unusual. The 90% confidence interval for the change in PSA values in men with benign prostatic hypertrophy, measured 3–6 months apart, is 1.4 ng/mL. Therefore, the Endocrine Society expert panel suggested that an increase in PSA >1.4 ng/mL in any one year after starting testosterone therapy, if confirmed, should lead to urologic evaluation. PSA velocity criterion can be used for patients who have sequential PSA measurements for >2 years; a change of >0.40 ng/mL per year merits closer urologic follow-up.

Cardiovascular Risk

As discussed above, there is insufficient evidence to determine whether testosterone replacement therapy increases the risk of major adverse cardiovascular events in hypogonadal men. A large prospective randomized trial is being planned to determine the effects of testosterone replacement therapy on major adverse cardiovascular events in middle-aged and older men with low testosterone levels and symptoms of androgen deficiency.

Androgen Abuse by Athletes and Recreational Bodybuilders

The illicit use of androgenic-anabolic steroids (AAS) to enhance athletic performance first surfaced in the 1950s among powerlifters and spread rapidly to other sports, professional as well as high school athletes, and recreational bodybuilders. In the early 1980s, the use of AAS spread beyond the athletic community into the general population, and now, as many as 3 million Americans—most of them men—have likely used these compounds. Most AAS users are not athletes, but rather recreational weightlifters, who use these drugs to look lean and more muscular. The most commonly used AAS include testosterone esters, nandrolone, stanozolol, methandienone, and methenolol. AAS users generally use increasing doses of multiple steroids in a practice known as stacking.

The adverse effects of long-term AAS abuse remain poorly understood. Most of the information about the adverse effects of AAS has emerged from case reports, uncontrolled studies, or from clinical trials that used replacement doses of testosterone. The adverse event data from clinical trials using physiologic replacement doses of testosterone have been extrapolated unjustifiably to AAS users who may administer 10–100 times the replacement doses of testosterone over many years, to support the claim that AAS use is safe. A substantial fraction of androgenic steroid users also use other drugs that are perceived to be muscle-building or performance-enhancing, such as growth hormone; erythropoiesis stimulating agents; insulin; stimulants such as amphetamine, clenbuterol, cocaine, ephedrine, and thyroxine; and drugs perceived to reduce adverse effects such as hCG, aromatase inhibitors, or estrogen antagonists. The adverse events associated with AAS use may be due to AAS themselves, concomitant use of other drugs, high-risk behaviors, and host characteristics that may render these individuals more susceptible to AAS use or to other high risk behaviors.

The high rates of premature mortality and morbidities observed in AAS users are alarming. One Finnish study reported 4.6 times the risk of death among elite power lifters than in age-matched men from the general population. The causes of death among power lifters included suicides, myocardial infarction, and hepatic coma. A retrospective review of patient records in Sweden also reported higher standardized mortality ratios for AAS users than for nonusers. Increased death rates among AAS users include suicide, homicide, and accidents. The median age of death among AAS users—24 years—is even lower than that for heroin or amphetamine users.

Four categories of adverse events associated with AAS abuse are of particular concern: cardiovascular events, psychiatric, prolonged suppression of the hypothalamic-pituitary-testicular axis, and potential neurotoxicity. Numerous reports of premature cardiac death among young AAS users raise concerns about the adverse cardiovascular effects of AAS. High doses of AAS may induce proatherogenic dyslipidemia, accelerate atherogenesis, increase thrombosis risk via effects on clotting factors and platelets, and induce vasospasm through their effects on vascular nitric oxide. Recent studies of AAS users using tissue Doppler and strain imaging, and magnetic resonance imaging have reported diastolic and systolic dysfunction, including significantly lower early and late diastolic tissue velocities, reduced E/A ratio, and reduced peak systolic strain in AAS users than in nonusers. Power athletes using AAS often have short QT intervals but increased QT dispersion, which may predispose them to ventricular arrhythmias. Long-term AAS use may be associated with myocardial hypertrophy and fibrosis. Myocardial tissue of power lifters using AAS has been shown to be infiltrated with fibrous tissue and fat droplets. The finding of AR on myocardial cells suggests that AAS might be directly toxic to myocardial cells. Studies of long-term AAS users using computerized tomography angiography have revealed accelerated atherogenesis.

Unlike replacement doses of testosterone, which are associated with only a small decrease in HDL cholesterol and little or no effect on total cholesterol, LDL cholesterol and triglyceride levels, supraphysiologic doses of testosterone and orally administered, 17-α-alylated, nonaromatizable AAS are associated with marked reductions in HDL cholesterol and increases in LDL cholesterol.

Some AAS users develop hypomanic and manic symptoms (irritability, aggressiveness, reckless behavior, and occasional psychotic symptoms, sometimes associated with violence) during AAS exposure, and major depression (sometimes associated with suicidality) during AAS withdrawal. Users may also be susceptible to other forms of illicit drug use, which may be potentiated or exacerbated by AAS.

Long-term AAS use suppresses LH and FSH secretion and inhibits endogenous testosterone production and spermatogenesis. Men, who have used AAS for more than a few months, experience marked suppression of the hypothalamic-pituitary-testicular (HPT) axis after stopping AAS that may be associated with sexual dysfunction, fatigue, infertility, depressed mood, and even suicidality. In some long-term AAS users, HPT suppression may last more than a year, and in a few individuals, recovery of the HPT axis may be incomplete or may never occur. The symptoms of androgen deficiency caused by androgen withdrawal may cause some men to revert back to using AAS, leading to continued use and AAS dependence. As many as 30% of AAS users develop a syndrome of AAS dependence, characterized by long-term AAS use despite adverse medical and psychiatric effects. AAS withdrawal hypogonadism has emerged as an important cause of androgen deficiency accounting for a substantial fraction of testosterone prescriptions in many men's health clinics.

Supraphysiologic doses of testosterone may also impair insulin sensitivity. Orally administered androgens also have been associated with insulin resistance and diabetes.

Unsafe injection practices, high-risk behaviors, and increased rates of incarceration render AAS users at increased risk of HIV, and hepatitis B and C. In one survey, nearly 1 in 10 gay men had injected AAS or other substances, and AAS users were more likely to report high-risk unprotected anal sex than other men.

Elevated liver enzymes, cholestatic jaundice, hepatic neoplasms, and peliosis hepatis have been reported with oral, 17-α-alkylated AAS. AAS use may cause muscle hypertrophy without compensatory adaptations in tendons, ligaments, and joints, thus increasing the risk of tendon and joint injuries. Upper extremity tendon ruptures are observed almost exclusively among weightlifters who use AAS. AAS use is associated with acne, baldness, as well as increased body hair.

The suspicion of AAS use should be raised by the increased hemoglobin and hematocrit, suppressed LH and FSH and testosterone levels, low high-density lipoproteins cholesterol, and low testicular volume and sperm density in a person who looks highly muscular. In most AAS users seeking medical attention, direct nonjudgmental questioning is sufficient to uncover AAS use and formal testing for AAS usually is not needed. However, if needed, accredited laboratories use gas chromatography-mass spectrometry or liquid chromatography-mass spectrometry to detect anabolic steroid abuse. In recent years, the availability of high-resolution mass spectrometry and tandem mass spectrometry has further improved the sensitivity of detecting androgen abuse. Illicit testosterone use is detected generally by the application of the measurement of urinary testosterone to epitestosterone ratio and further confirmed by the use of the ^{13}C:^{12}C ratio in testosterone by the use of isotope ratio combustion mass spectrometry. Exogenous testosterone administration increases urinary testosterone glucuronide excretion and consequently the testosterone to epitestosterone ratio. Ratios above 4 suggest exogenous testosterone use but can also reflect genetic variation. Genetic variations in the uridine diphospho-glucuronyl transferase 2B17 (*UGT2B17*), the major enzyme for testosterone glucuronidation, affect testosterone to epitestosterone ratio. Synthetic testosterone has a lower ^{13}C:^{12}C ratio than endogenously produced testosterone and these differences in ^{13}C:^{12}C ratio can be detected by isotope ratio combustion mass spectrometry, which is used to confirm exogenous testosterone use in individuals with a high testosterone to epitestosterone ratio.

TABLE 384-3

Clinical Pharmacology of Some Testosterone Formulations

Formulation	Regimen	Pharmacokinetic profile	DHT and E2	Advantages	Disadvantages
T enanthate or cypionate	150–200 mg IM q 2 wk or 75–100 mg/wk	After a single IM injection, serum T levels rise into the supraphysiological range, then decline gradually into the low normal or the hypogonadal range by the end of the dosing interval	DHT and E2 levels rise in proportion to the increase in T levels; T:DHT and T:E2 ratios do not change	Corrects symptoms of androgen deficiency; relatively inexpensive, if self-administered; flexibility of dosing	Requires IM injection; peaks and valleys in serum T levels
Topical testosterone gels and axillary testosterone solution	Available in sachets, tubes and pumps	When used in appropriate doses, these topical formulations restore serum T and E2 levels to the physiological male range	Serum DHT levels and DHT to T ratio are higher in hypogonadal men treated with the transdermal gels than in healthy eugonadal men	Corrects symptoms of androgen deficiency, ease of application, good skin tolerability	Potential of transfer to a female partner or child by direct skin-to-skin contact; skin irritation in a small proportion of treated men; moderately high DHT levels; considerable interindividual and intra-indivudal variation in on-treatment testosterone levels
Transdermal testosterone patch	1 or 2 patches, designed to nominally deliver 4–8 mg T over 24 h applied daily on nonpressure areas	Restores serum T, DHT, and E2 levels to the physiological male range	T:DHT and T:E2 levels are in the physiological male range	Ease of application, corrects symptoms of androgen deficiency	Serum T levels in some androgen-deficient men may be in the low-normal range; these men may need application of 2 patches daily; skin irritation at the application site occurs frequently in many patients
Buccal, bioadhesive, T tablets	30 mg controlled release, bioadhesive tablets bid	Absorbed from the buccal mucosa	Normalizes serum T and DHT levels in hypogonadal men	Corrects symptoms of androgen deficiency	Gum-related adverse events in 16% of treated men
T pellets	Several pellets implanted sc; dose and regimen vary with formulation	Serum T peaks at 1 mo and then is sustained in normal range for 3–4 mo, depending on formulation	T:DHT and T:E2 ratios do not change	Corrects symptoms of androgen deficiency	Requires surgical incision for insertions; pellets may extrude spontaneously
17-α-methyl T	This 17-α-alkylated compound should **not** be used because of potential for liver toxicity.	Orally active			Clinical responses are variable; potential for liver toxicity; should **not** be used for treatment of androgen deficiency

Formulation	Regimen	Pharmacokinetic profile	DHT and E2	Advantages	Disadvantages
Oral T undecanoate*	40–80 mg po bid or tid with meals	When administered in oleic acid, T undecanoate is absorbed through the lymphatics, bypassing the portal system; considerable variability in the same individual on different days and among individuals	High DHT to T ratio	Convenience of oral administration	Not approved in the US;variable clinical responses, variable serum T levels, high DHT:T ratio
Injectable long-acting T undecanoate in oil[1]	US regimen 750 mg IM, followed by 750 mg at 4 wk, and 750 mg every 10 weeks	When administered at the recommended dose, serum T levels are maintained in the normal range in a majority of treated men	DHT and E2 levels rise in proportion to the increase in T levels; T:DHT and T:E2 ratios do not change	Corrects symptoms of androgen deficiency; requires infrequent administration.	Requires IM injection of a large volume; cough reported immediately after injection in a small number of men
Testosterone-in-adhesive matrix patch*	2×60 cm^2 patches delivering ~4.8 mg of T/d	Restores serum T, DHT and E_2 to the physiological range	T:DHT and T:E_2 are in the physiological range.	Lasts 2 d	Some skin irritation
Intranasal Testosterone	2 actuations of the metered dose pump (11 mg) applied into the nostrils three times daily	Restores T into the normal male range	T:DHT and T:E2 ratio in the physiologic range		Requires 3x daily application; nasal irritation, epistaxis, nasophyringitis

*These formulations are not approved for clinical use in the United States, but are available outside the United States in many countries. Physicians in those countries where these formulations are available should follow the approved drug regimens.

Abbreviations: DHT, dihydrotestosterone; E2, estradiol; T, testosterone.

TABLE 384-4

Conditions in Which Testosterone Administration Is Associated with an Increased Risk of Adverse Outcomes

Conditions in which testosterone administration is associated with very high risk of serious adverse outcomes:
Metastatic prostate cancer
Breast cancer
Conditions in which testosterone administration is associated with moderate to high risk of adverse outcomes:
Undiagnosed prostate nodule or induration
PSA > 3
Erythrocytosis (hematocrit >50%)
Severe lower urinary tract symptoms associated with benign prostatic hypertrophy as indicated by American Urological Association/International prostate symptom score >19
Uncontrolled or poorly controlled congestive heart failure
Myocardial infarction, stroke, or acute coronary syndrome in the preceding 3 months

Abbreviation: PSA, prostate-specific antigen.

Source: Reproduced from the Endocrine Society Guideline for Testosterone Therapy of Androgen Deficiency Syndromes in Men (Bhasin S et al: J Clin Endocrinol Metab 95:2536, 2010).

TABLE 384-5

Monitoring Men Receiving Testosterone Therapy

1. Evaluate the patient 3–6 months after treatment initiation and then annually to assess whether symptoms have responded to treatment and whether the patient is suffering from any adverse effects.
2. Monitor testosterone level 3–6 months after initiation of testosterone therapy:

- Therapy should aim to raise serum testosterone level into the mid-normal range.
- Injectable testosterone enanthate or cypionate: Measure serum testosterone level midway between injections. If testosterone is >600 ng/dL (20.9 nmol/L) or <350 ng/dL (12.2 nmol/L), adjust dose or frequency.
- Transdermal patches: Assess testosterone level 3–12 h after application of the patch; adjust dose to achieve testosterone level in the midnormal range.
- Buccal testosterone bioadhesive tablet: Assess level immediately before application of fresh system.
- Transdermal gels and solution: Assess testosterone level 2–12 h after patient has been on treatment for at least 2 weeks; adjust dose to achieve serum testosterone level in the midnormal range.
- Testosterone pellets: Measure testosterone levels at the end of the dosing interval. Adjust the number of pellets and/or the dosing interval to achieve serum testosterone levels in the normal range.
- Injectable testosterone undecanoate: Measure serum testosterone level just prior to each subsequent injection and adjust the dosing interval to maintain serum testosterone in mid-normal range.

3. Check hematocrit at baseline, at 3–6 months, and then annually. If hematocrit is >54%, stop therapy until hematocrit decreases to a safe level; evaluate the patient for hypoxia and sleep apnea; reinitiate therapy with a reduced dose.
4. Measure bone mineral density of lumbar spine and/or femoral neck after 1–2 yr of testosterone therapy in hypogonadal men with osteoporosis or low trauma fracture, consistent with regional standard of care.
5. In men aged ≥40 years with baseline PSA >0.6 ng/mL, perform digital rectal examination and check PSA level before initiating treatment, at 3–6 months, and then in accordance with guidelines for prostate cancer screening depending on the age and race of the patient.
6. Obtain urological consultation if there is:

- An increase in serum PSA concentration >1.4 ng/mL within any 12-month period of testosterone treatment.
- A PSA velocity of >0.4 ng/mL·yr using the PSA level after 6 months of testosterone administration as the reference (only applicable if PSA data are available for a period exceeding 2 yr).
- Detection of a prostatic abnormality on digital rectal examination.
- An AUA/ IPSS prostate symptom score of >19 along with an increase in IPSS score of ≥5 points above baseline.

7. Evaluate formulation-specific adverse effects at each visit:

- Buccal testosterone tablets*: Inquire about alterations in taste and examine the gums and oral mucosa for irritation.
- Injectable testosterone esters (enanthate, cypionate, and undecanoate): Ask about fluctuations in mood or libido, and rarely cough after injections.
- Testosterone patches: Look for skin reaction at the application site.
- Testosterone gels: Advise patients to cover the application sites with a shirt and to wash the skin with soap and water before having skin-to-skin contact, because testosterone gels leave a testosterone residue on the skin that can be transferred to a woman or child who might come in close contact. Serum testosterone levels are maintained when the application site is washed 4–6 h after application of the testosterone gel.
- Testosterone pellets: Look for signs of infection, fibrosis, or pellet extrusion.
- Intranasal testosterone: Look for signs of nasal irritation or scab

*Not approved for clinical use in the United States.

Abbreviations: AUA/IPSS, American Urological Association International Prostate Symptom Score; PSA, prostate-specific antigen.

Source: Reproduced with permission from the Endocrine Society Guideline for Testosterone Therapy of Androgen Deficiency Syndromes in Adult Men (Bhasin S et al: J Clin Endocrinol Metab 95:2536, 2010).

FURTHER READING

Abreu AP et al: Central precocious puberty caused by mutations in the imprinted gene *MKRN3*. N Engl J Med 368:26, 2013.

Balasubramanian R et al: Human GnRH deficiency: A unique disease model to unravel the ontogeny of GnRH neurons. Neuroendocrinology 92:81, 2010. [PubMed: 20606386]

Basaria S et al: Effects of testosterone administration for 3 years on subclinical atherosclerosis progression in older men with low or low-normal testosterone levels: A randomized clinical trial. JAMA 314:570, 2015. [PubMed: 26262795]

bhasin S et al: Testosterone therapy in men with androgen deficiency syndromes: An endocrine society clinical practice guideline. J Clin Endocrinol Metab 95:2536, 2010. [PubMed: 20525905]

Bhasin S et al: Effect of testosterone supplementation with and without a dual 5α-reductase inhibitor on fat-free mass in men with suppressed testosterone production: a randomized controlled trial. JAMA 307:931, 2012. [PubMed: 22396515]

Budoff MJ et al: Testosterone treatment and coronary artery plaque volume in older men with low testosterone. JAMA 317:708, 2017. [PubMed: 28241355]

Finkelstein JS et al: Gonadal steroids and body composition, strength, and sexual function in men. N Engl J Med 369:1011, 2013. [PubMed: 24024838]

Forni PE, Wray S: GnRH, anosmia and hypogonadotropic hypogonadism—Where are we? Front Neuroendocrinol 36:165, 2015. [PubMed: 25306902]

Pope HG Jr et al: Adverse health consequences of performance-enhancing drugs: An Endocrine Society scientific statement. Endocr Rev 35:341, 2014. [PubMed: 24423981]

Snyder PJ et al: Effects of testosterone treatment in older men. N Engl J Med 74:611, 2016.

Spitzer M et al: Risks and benefits of testosterone therapy in older men. Nat Rev Endocrinol 9:414, 2013. [PubMed: 23591366]

Travison TG et al: Harmonized reference ranges for circulating testosterone levels in men of four cohort studies in the USA and Europe. J Clin Endocrinol Metab 2017 in press.

Zakharov MN et al: A multi-step, dynamic allosteric model of testosterone's binding to sex hormone binding globulin. Mol Cell Endocrinol 399:190, 2015. [PubMed: 25240469]

Chapter 385: Disorders of the Female Reproductive System

Janet E. Hall

INTRODUCTION

The female reproductive system regulates the hormonal changes responsible for puberty and adult reproductive function. Normal reproductive function in women requires the dynamic integration of hormonal signals from the hypothalamus, pituitary, and ovary, resulting in repetitive cycles of follicle development, ovulation, and preparation of the endometrial lining of the uterus for implantation should conception occur. It is critical to understand pubertal development in normal girls (and boys) as a yardstick for identifying precocious and delayed puberty.

For further discussion of related topics, see the following chapters: amenorrhea and pelvic pain (Chap. 386), infertility and contraception (Chap. 389), menopause (Chap. 388), disorders of sex development (Chap. 383), and disorders of the male reproductive system (Chap. 384).

DEVELOPMENT OF THE OVARY AND EARLY FOLLICULAR GROWTH

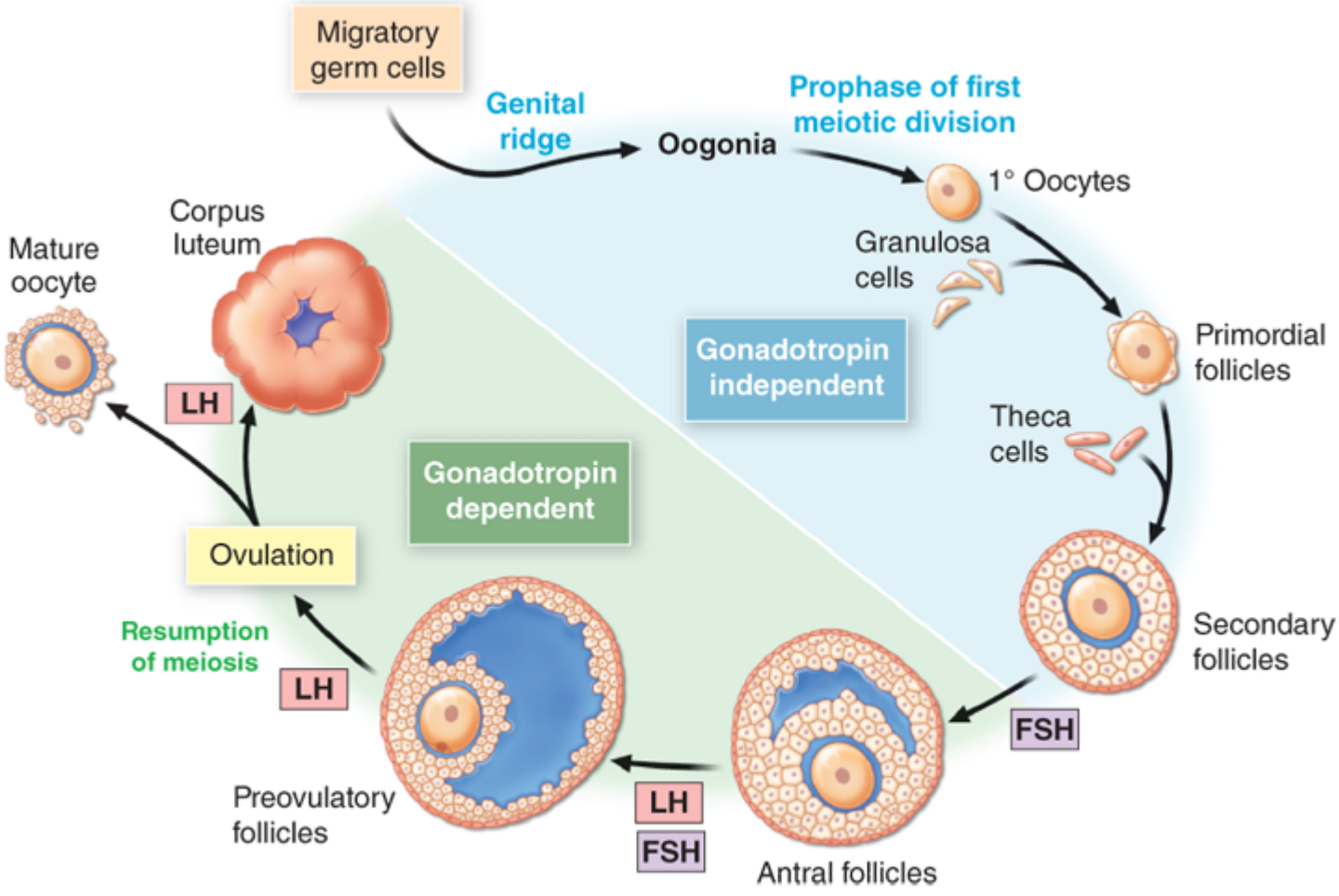
The ovary orchestrates the development and release of a mature oocyte and secretes hormones (e.g., estrogen, progesterone, inhibins A and B, relaxin) that play critical roles in a variety of target tissues, including breast, bone, and uterus, in addition to the hypothalamus and pituitary. To achieve these functions in repeated monthly cycles, the ovary undergoes some of the most dynamic changes of any organ in the body. Primordial germ cells can be identified by the third week of gestation, and their migration to the genital ridge is complete by 6 weeks of gestation. Germ cells persist within the genital ridge, are then referred to as *oogonia*, and are essential for induction of ovarian development. In patients with 45,X Turner syndrome, primordial germ cells proliferate and migrate to the genital ridge, but do not persist as their survival requires the presence of pregranulosa cells that are dependent on the presence of both X chromosomes. **(Chap. 383)**.

The germ cell population expands, and starting at ~8 weeks of gestation, oogonia begin to enter prophase of the first meiotic division and become primary oocytes. This allows the oocyte to be surrounded by a single layer of flattened granulosa cells to form a primordial follicle **(Fig. 385-1)**. Granulosa cells are derived from mesonephric cells that invade the ovary early in its development, pushing the germ cells to the periphery. Although there is evidence that both oocyte-like cells and follicle-like structures can form from embryonic

stem cells in culture, there is, as yet, no clear evidence that this occurs in vivo and thus, the ovary appears to contain a nonrenewable pool of germ cells. Through the combined processes of mitosis, meiosis, and atresia, the population of oogonia reaches its maximum of 6–7 million by 20 weeks of gestation, after which there is a progressive loss of both oogonia and primordial follicles through the process of atresia. It appears that entry into meiosis provides some degree of protection from programmed cell death. At birth, oogonia are no longer present in the ovary, and only 1–2 million germ cells remain in the form of primordial follicles **(Fig. 385-2)**. The oocyte persists in prophase of the first meiotic division until just before ovulation, when meiosis resumes.

FIGURE 385-1

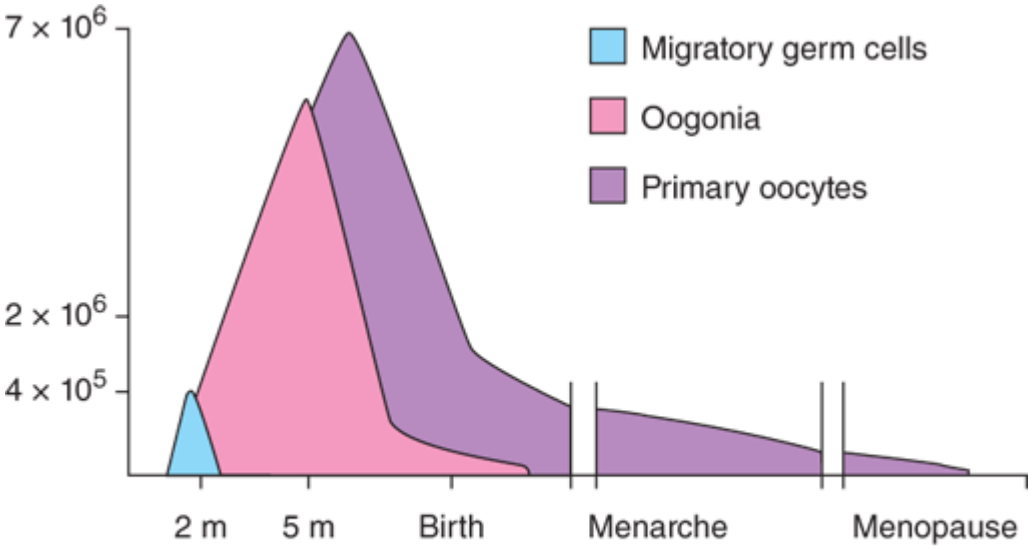
Stages of ovarian development from the arrival of the migratory germ cells at the genital ridge through gonadotropin-independent and gonadotropin-dependent phases that ultimately result in ovulation of a mature oocyte. FSH, follicle-stimulating hormone; LH, luteinizing hormone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 385-2

Ovarian germ cell number is maximal at mid-gestation and decreases precipitously thereafter.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The quiescent primordial follicles are recruited to further growth and differentiation through a highly regulated process that limits the size of the developing cohort to ensure that folliculogenesis can continue throughout the reproductive life span. This initial recruitment of primordial follicles to form primary follicles (Fig. 385-1) is characterized by growth of the oocyte and the transition from squamous to cuboidal granulosa cells. The theca interna cells that surround the developing follicle begin to form as the primary follicle grows. Acquisition of a zona pellucida by the oocyte and the presence of several layers of surrounding cuboidal granulosa cells mark the development of secondary follicles. It is at this stage that granulosa cells develop follicle-stimulating hormone (FSH), estradiol, and androgen receptors and communicate with one another through the development of gap junctions.

Bidirectional signaling between the germ cells and the somatic cells in the ovary is a necessary component underlying the maturation of the oocyte and the capacity for hormone secretion. For example, oocyte-derived growth differentiation factor 9 (GDF-9) and bone morphogenic protein-15 (BMP-15), also known as GDF-9b, are required for migration of pregranulosa and pretheca cells to the outer surface of the developing follicle and, hence, initial follicle formation. GDF-9 is also required for formation of secondary follicles, as are granulosa cell–derived KIT ligand (KITL) and the forkhead transcription factor (FOXL2). A significant number of genes have been identified that are required for development of the normal complement of oogonia in the ovary, initial follicle development and resistance to follicle loss; all are candidates for premature ovarian insufficiency (POI) and mutations in >50 genes have been identified in patients with POI with even more that have been associated with an earlier age at natural menopause.

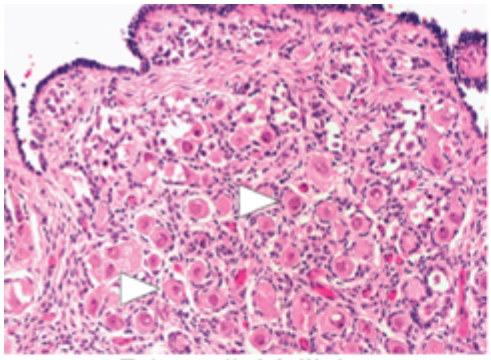
DEVELOPMENT OF A MATURE FOLLICLE

The early stages of follicle growth are primarily driven by intraovarian factors; after initial recruitment, development to the secondary follicle stage may take close to a year. Further maturation to the preovulatory stage, including the resumption of meiosis in the oocyte, requires the combined stimulus of FSH and luteinizing hormone (LH) (Fig. 385-1). Recruitment of secondary follicles from the resting follicle pool

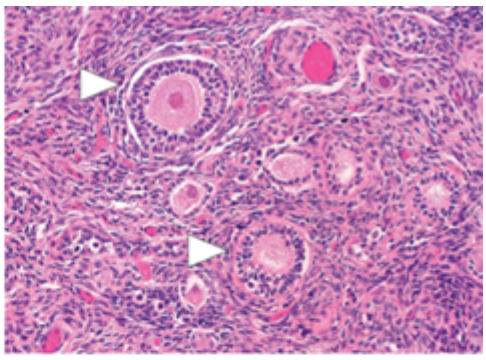
requires the direct action of FSH, whereas anti-müllerian hormone (AMH) produced from small growing follicles, restrains this effect of FSH controlling the number of follicles entering the actively growing pool. Accumulation of follicular fluid between the layers of granulosa cells creates an antrum that divides the granulosa cells into two functionally distinct groups: mural cells that line the follicle wall and cumulus cells that surround the oocyte **(Fig. 385-3)**. In addition to its role in normal development of the müllerian system, the WNT signaling pathway is required for normal antral follicle development and may also play a role in ovarian steroidogenesis. Recruitment to the small antral stage generally occurs over several cycles with further growth to follicle sizes of >4–7 mm in several waves during a single cycle. A single dominant follicle emerges from the growing follicle pool within the first 5–7 days after the onset of menses while the majority of follicles fall off their growth trajectory and become atretic. Autocrine actions of activin and BMP-6, derived from the granulosa cells, and paracrine actions of GDF-9, BMP-15, BMP-6, and Gpr149, derived from the oocyte, are involved in granulosa cell proliferation and modulation of FSH responsiveness. Differential exposure to these factors, and to vascular endothelial growth factor (VEGF), can attenuate vascular density and permeability, likely explaining the mechanism whereby a given follicle is selected for continued growth to the preovulatory stage. The dominant follicle can be distinguished by its size, evidence of granulosa cell proliferation, large number of FSH receptors, high aromatase activity, and elevated concentrations of estradiol and inhibin A in follicular fluid. In addition, secretion of estradiol and inhibin from the dominant follicle inhibits FSH and the growth of other follicles.

FIGURE 385-3

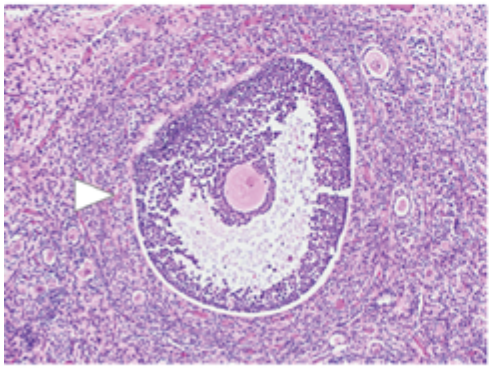
Development of ovarian follicles. The Graafian follicle is also known as a tertiary or preovulatory follicle. *(Courtesy of J.H. Eichhorn and D. Roberts, Massachusetts General Hospital; with permission.)*



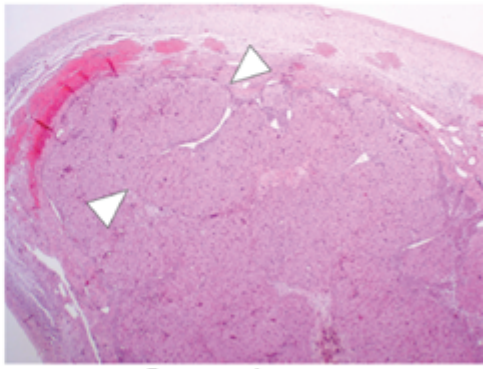
Primordial follicles



Primary follicles



Graafian follicle



Corpus luteum

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The dominant follicle undergoes rapid expansion during the 5–6 days prior to ovulation, reflecting granulosa cell proliferation and accumulation of follicular fluid. FSH induces LH receptors on the granulosa cells, and the preovulatory, or Graafian, follicle moves to the outer ovarian surface in preparation for ovulation. The LH surge triggers the resumption of meiosis, the suppression of granulosa cell proliferation, and the induction of cyclooxygenase 2 (COX-2), prostaglandins, the progesterone receptor (PR), and the epidermal growth factor (EGF)-like growth factors amphiregulin, epiregulin, betacellulin, and neuroregulin 1, all of which are required for ovulation. Ovulation requires production of extracellular matrix leading to expansion of the cumulus cell population that surrounds the oocyte and the controlled expulsion of the egg and follicular fluid. Both progesterone and prostaglandins (induced by the ovulatory stimulus) are essential for this process as are members of the matrix metalloproteinase family. After ovulation, luteinization of theca and granulosa cells is induced by LH in conjunction with the acquisition of a rich vascular network in response to VEGF and basic fibroblast growth factor (FGF). Traditional regulators of central reproductive control, gonadotropin-releasing hormone (GnRH) and its receptor (GnRHR), as well as kisspeptin, are also produced in the ovary and may be involved in corpus luteum function.

REGULATION OF OVARIAN FUNCTION

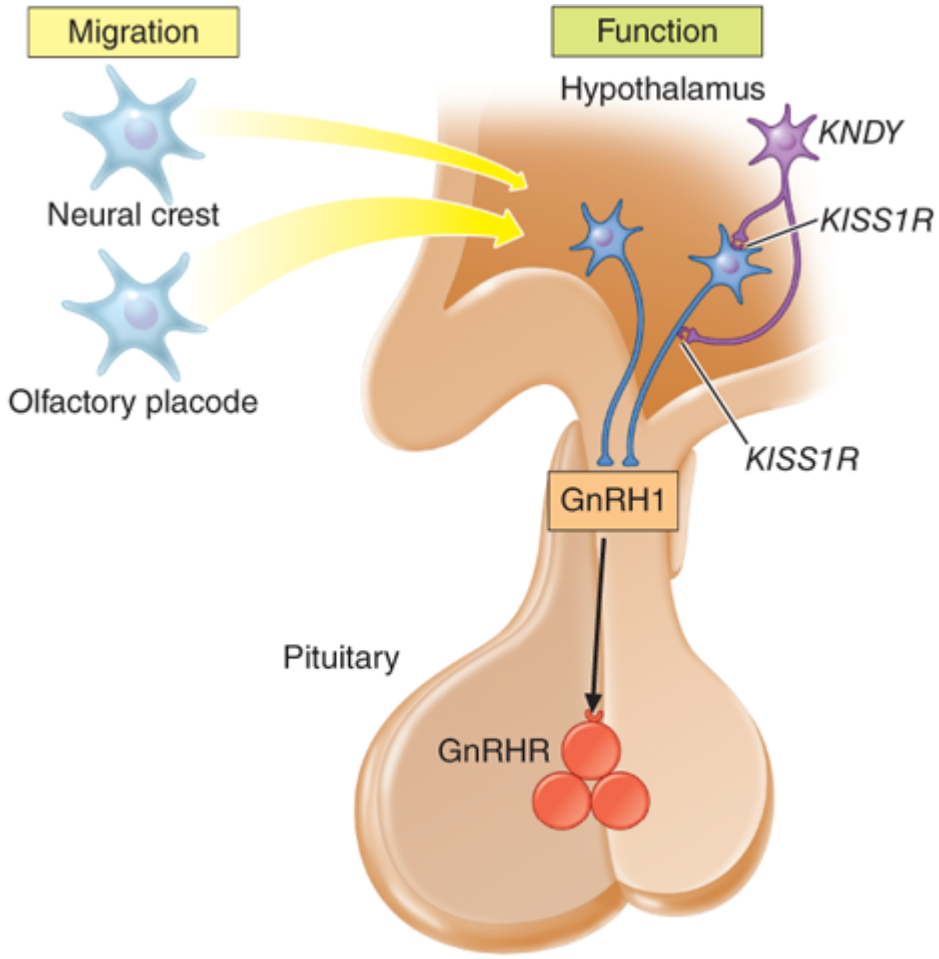
HYPOTHALAMIC AND PITUITARY SECRETION

GnRH neurons derive from cells in the olfactory placode and to a lesser extent, the neural crest. They migrate along the scaffold of the olfactory neurons across the cribiform plate to the hypothalamus where they

separate from the olfactory neurons. Studies in GnRH-deficient patients who fail to undergo puberty have provided insights into genes that control the ontogeny and function of GnRH neurons **(Fig. 385-4)**. *KAL1*, *FGF8/FGFR1*, *PROK2/PROKR2, NSMF*, *HS6SD1,* and *CDH7*, among others **(Chap. 384)**, have been implicated in the migration of GnRH neurons to the hypothalamus. Approximately 7000 GnRH neurons, scattered throughout the medial basal hypothalamus, establish contacts with capillaries of the pituitary portal system in the median eminence. GnRH is secreted into the pituitary portal system in discrete pulses to stimulate synthesis and secretion of LH and FSH from pituitary gonadotropes, which comprise ~10% of cells in the pituitary **(Chap. 371)**. Functional connections of GnRH neurons with the portal system are established by the end of the first trimester, coinciding with the production of pituitary gonadotropins. Thus, like the ovary, the hypothalamic and pituitary components of the reproductive system are present before birth. However, the high levels of estradiol and progesterone produced by the placenta suppress hypothalamic-pituitary stimulation of ovarian hormonal secretion in the fetus.

FIGURE 385-4

Genetic studies in patients with congenital forms of hypogonadotropic hypogonadism have expanded our understanding of the development and migration of gonadotropin-releasing hormone (GnRH) neurons from the olfactory placode and neural crest to the hypothalamus as well as the upstream regulation of GnRH secretion.

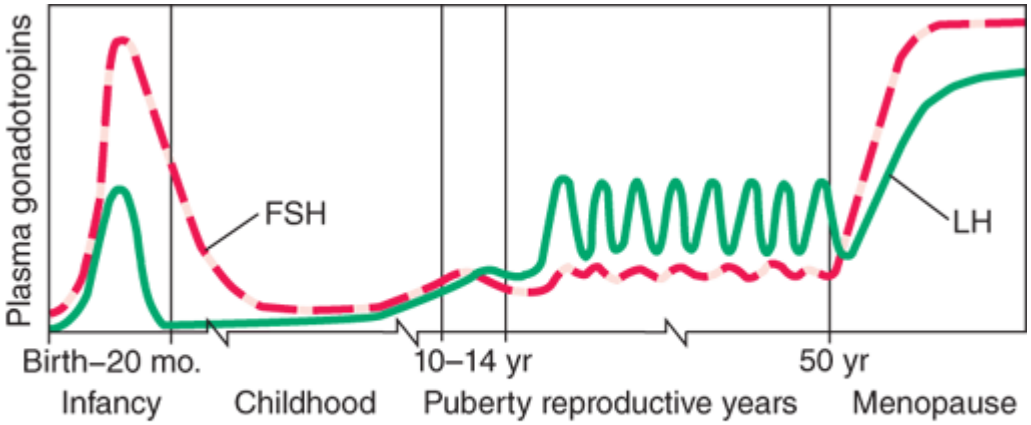


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

After birth and the loss of placenta-derived steroids, gonadotropin levels rise. FSH levels are much higher in girls than in boys. This rise in FSH results in circulating estradiol and increased inhibin B, but without terminal follicle maturation or ovulation . Studies that have identified mutations in *TAC3,* which encodes neurokinin B, and its receptor, *TAC3R,* in patients with GnRH deficiency indicate that both are involved in control of GnRH secretion and may be particularly important at this early stage of development. By 12–20 months of age, the reproductive axis is again suppressed, and a period of relative quiescence persists until puberty **(Fig. 385-5)**. At the onset of puberty, pulsatile GnRH secretion induces pituitary gonadotropin production. In the early stages of puberty, LH and FSH secretion are apparent only during sleep, but as puberty develops, pulsatile gonadotropin secretion occurs throughout the day and night.

FIGURE 385-5

Follicle-stimulating hormone (FSH) and luteinizing hormone (LH) are increased during the neonatal years but go through a period of childhood quiescence before increasing again during puberty. Gonadotropin levels are cyclic during the reproductive years and increase dramatically with the loss of negative feedback that accompanies menopause.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The mechanisms responsible for the childhood quiescence and pubertal reactivation of the reproductive axis remain incompletely understood. GnRH neurons in the hypothalamus respond to both excitatory and inhibitory factors. Increased sensitivity to the inhibitory influence of gonadal steroids has long been implicated in the inhibition of GnRH secretion during childhood but has not been definitively established in the human. Metabolic signals, including adipocyte-derived leptin, play a permissive role in reproductive function **(Chap. 394)**. Studies of patients with isolated GnRH deficiency reveal that mutations in the G protein–coupled receptor 54 (*GPR54*) gene (now known as *KISS1R*) preclude the onset of puberty. The ligand for this receptor is derived from the parent peptide, kisspeptin-1 (*KISS1*), and is a powerful stimulant for GnRH release. A potential role for kisspeptin in the onset of puberty has been suggested by upregulation of *KISS1* and *KISS1R* transcripts in the hypothalamus at the time of puberty. *TAC3,* which stimulates GnRH secretion through kisspeptin signaling, and dynorphin (*Dyn*), which plays an inhibitory role in GnRH control, are frequently co-expressed with *KISS1* in KNDy neurons of the median eminence that project to GnRH neurons. This system is intimately involved in both estrogen and progesterone negative feedback regulation of GnRH secretion.

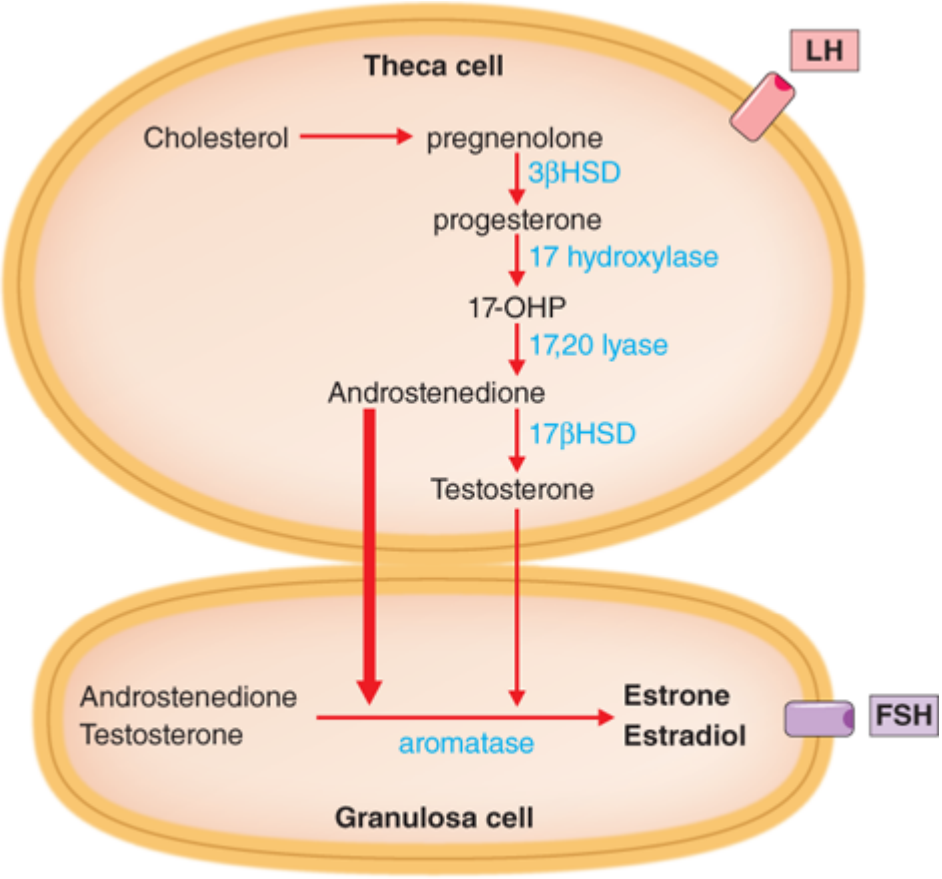
RFamide-Related peptides (RFRPs) are the mammalian orthologues of gonadotropin inhibitory hormone (GnIH) which was initially discovered in the quail. While RFRP-1 and RFRP-3 neurons send axonal projections to GnRH neurons in humans, and RFRPs are secreted into the pituitary portal system, further studies are required to determine their potential physiologic role in the human.

OVARIAN STEROIDS

Ovarian steroid-producing cells do not store hormones but produce them in response to LH and FSH during the normal menstrual cycle. The sequence of steps and the enzymes involved in the synthesis of steroid hormones are similar in the ovary, adrenal, and testis. However, the enzymes required to catalyze specific steps are compartmentalized and may not be abundant or even present in all cell types. Within the developing ovarian follicle, estrogen synthesis from cholesterol requires close integration between theca and granulosa cells—sometimes called the *two-cell model for steroidogenesis* **(Fig. 385-6)**. FSH receptors are confined to the granulosa cells, whereas LH receptors are restricted to the theca cells until the late stages of follicular development, when they are also found on granulosa cells. The theca cells surrounding the follicle are highly vascularized and use cholesterol, derived primarily from circulating lipoproteins, as the starting point for the synthesis of androstenedione and testosterone under the control of LH. These steroid precursors cross the basal lamina to the granulosa cells, which receive no direct blood supply. The mural granulosa cells are particularly rich in aromatase and, under the control of FSH, produce estradiol, the primary steroid secreted from the follicular phase ovary and the most potent estrogen. Theca cell–produced androstenedione and, to a lesser extent, testosterone are also secreted into peripheral blood, where they can be converted to dihydrotestosterone in skin and to estrogens in adipose tissue. The hilar interstitial cells of the ovary are functionally similar to Leydig cells and are also capable of secreting androgens. Stromal cells proliferate in response to androgens (as in polycystic ovarian syndrome [PCOS]), but do not secrete androgens. However, high levels of androgens may be produced by luteinized theca cells in women with hyperthecosis.

FIGURE 385-6

Estrogen production in the ovary requires the cooperative function of the theca and granulosa cells under the control of luteinizing hormone (LH) and follicle-stimulating hormone (FSH). HSD, hydroxysteroid dehydrogenase; OHP, hydroxyprogesterone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Development of the rich capillary network following rupture of the follicle at the time of ovulation makes it possible for large molecules such as low-density lipoprotein (LDL) to reach the luteinized granulosa and theca lutein cells. As in the follicle, both cell types are required for steroidogenesis in the corpus luteum. The luteinized granulosa cells are the main source of progesterone production, whereas the smaller theca lutein cells produce 17-hydroxyprogesterone and androgenic substrates for aromatization to estradiol by the luteinized granulosa cells. Production of estrogen metabolites by the corpus luteum plays a significant role in maintenance of the vascularization required for its function. LH is critical for formation and maintenance of corpus luteum structure and function. LH and human chorionic gonadotropin (hCG) bind to a common receptor; thus, in conception cycles, hCG rescues the declining function of the corpus luteum, maintaining steroid and peptide secretion for the first 10 weeks of pregnancy. HCG is commonly used for luteal phase support in the treatment of infertility.

Steroid Hormone Actions

Both estrogen and progesterone play critical roles in the expression of secondary sexual characteristics in women (**Chap. 370**). Estrogen promotes development of the ductule system in the breast, whereas progesterone is responsible for glandular development. In the reproductive tract, estrogens create a receptive environment for fertilization and support pregnancy and parturition through carefully coordinated changes in the endometrium, thickening of the vaginal mucosa, thinning of the cervical mucus, and uterine growth and contractions. Progesterone induces secretory activity in the estrogen-primed endometrium,

increases the viscosity of cervical mucus, and inhibits uterine contractions. Both gonadal steroids play critical roles in negative and positive feedback of gonadotropin secretion. Progesterone also increases basal body temperature and has therefore been used clinically as a marker of ovulation.

The vast majority of circulating estrogens and androgens are carried in the blood bound to carrier proteins, which restrain their free diffusion into cells and prolong their clearance, serving as a reservoir. High-affinity binding proteins include sex hormone–binding globulin (SHBG), which binds androgens with somewhat greater affinity than estrogens, and corticosteroid-binding globulin (CBG), which also binds progesterone. Modulations in binding protein levels by insulin, androgens, and estrogens contribute to high bioavailable testosterone levels in PCOS and to high circulating total estrogen and progesterone levels during pregnancy.

Estrogens act primarily through binding to the nuclear receptors, estrogen receptor (ER) α and β. Transcriptional coactivators and co-repressors modulate ER action **(Chap. 370)**. Both ER subtypes are present in the hypothalamus, pituitary, ovary, and reproductive tract. Although ERα and β exhibit some functional redundancy, there is also a high degree of specificity, particularly in expression within cell types. For example, ERα functions in ovarian theca cells, whereas ERβ is critical for granulosa cell function. There is also evidence for membrane-initiated signaling by estrogen. Similar signaling mechanisms pertain for progesterone with evidence of transcriptional regulation through PR A and B protein isoforms, as well as rapid membrane signaling.

OVARIAN PEPTIDES

Inhibin was initially isolated from gonadal fluids based on its ability to selectively inhibit FSH secretion from pituitary cells. Inhibin is a heterodimer composed of an α subunit and a βA or βB subunit to form inhibin A or inhibin B, both of which are secreted from the ovary. Activin is a homodimer of inhibin β subunits with the capacity to stimulate the synthesis and secretion of FSH. Inhibins and activins are members of the transforming growth factor β (TGF-β) superfamily of growth and differentiation factors. During the purification of inhibin, follistatin, an unrelated monomeric protein that inhibits FSH secretion, was discovered. Within the pituitary, follistatin inhibits FSH secretion indirectly by binding and neutralizing activin.

Inhibin B is constitutively secreted from the granulosa cells of small antral follicles and its serum levels increase in conjunction with granulosa cell proliferation during recruitment of secondary follicles under the control of FSH. In addition to its role as a marker of decreasing ovarian reserve during reproductive aging, inhibin B is an important inhibitor of FSH, independent of estradiol, during the menstrual cycle. Inhibin A is present in both granulosa and theca cells and is secreted by the dominant follicle. Inhibin A is also present in luteinized granulosa cells and is a major secretory product of the corpus luteum. Synthesis and secretion of inhibin A are directly controlled by FSH and LH. Although activin is also secreted from the ovary, the excess of follistatin in serum, combined with its nearly irreversible binding of activin, make it unlikely that ovarian activin plays an endocrine role in FSH regulation. However, there is evidence that activin plays an autocrine/paracrine role in the ovary, in addition to its intra-pituitary role in modulation of FSH production.

AMH (also known as müllerian-inhibiting substance) is important in ovarian biology in addition to the function from which it derived its name (i.e., promotion of the degeneration of the müllerian system during embryogenesis in the male). AMH is produced by granulosa cells from small follicles and is a marker of ovarian reserve with advantages over inhibin B because of its relative stability across the menstrual cycle. AMH inhibits the recruitment of primordial follicles into the follicle pool and counters FSH stimulation of aromatase expression. AMH is increased in polycystic ovarian syndrome in conjunction with the abundance of small follicles in this disorder.

Gonadotropin Surge Attenuating Factor (GnSAF) is an ovarian factor that attenuates GnRH-induced gonadotropin secretion. Its role is not yet fully understood, but there is an inverse relationship between GnSAF and follicle size suggesting that its primary role involves the early stages of follicle development rather than curtailing the gonadotropin surge as its name implies.

Relaxin is produced primarily by the theca lutein cells of the corpus luteum. Both relaxin and its receptor, RXFP1, are highly expressed in the uterus during the peri-implantation period in the marmoset and its primary role appears to be in promoting decidualization and vascularization of the endometrium prior to implantation. Relaxin was named for its ability to suppress myometrial contractility in pigs and rodents, but it does not appear to exert this activity in women.

HORMONAL INTEGRATION OF THE NORMAL MENSTRUAL CYCLE

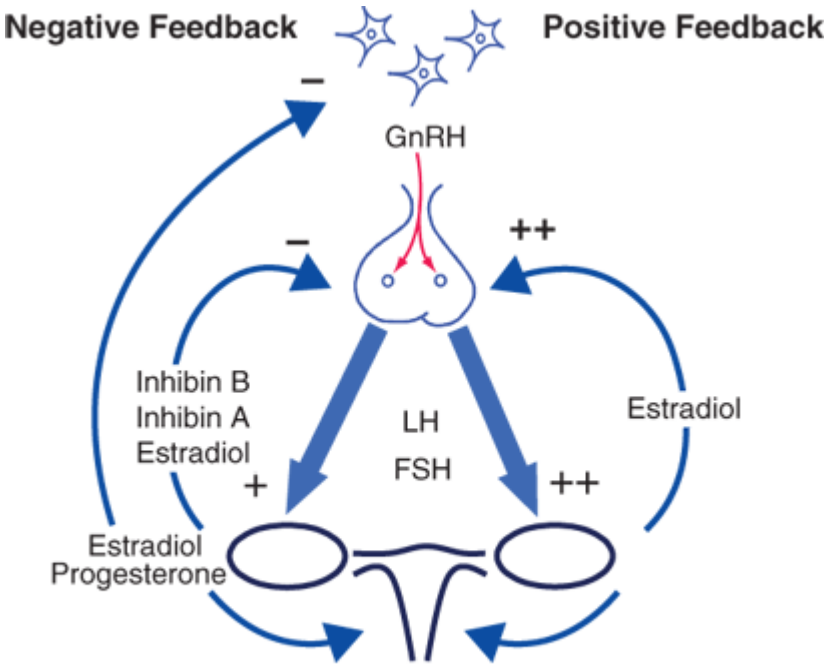
The sequence of changes responsible for mature reproductive function is coordinated through a series of negative and positive feedback loops that alter pulsatile GnRH secretion, the pituitary response to GnRH, and the relative secretion of LH and FSH from the gonadotrope. The frequency and amplitude of pulsatile GnRH secretion differentially modulate the synthesis and secretion of LH and FSH. Slow GnRH pulse frequencies favor FSH synthesis whereas increased GnRH pulse frequency and amplitude favor LH synthesis. Activin is produced in both pituitary gonadotropes and folliculostellate cells and stimulates the synthesis and secretion of FSH through autocrine-paracrine mechanisms that are modulated by follistatin. Inhibins function as potent antagonists of activins through sequestration of the activin receptors. Although inhibin is expressed in the pituitary, gonadal inhibin is the principal source of feedback inhibition of FSH.

For the majority of the cycle, the reproductive system functions in a classic endocrine negative feedback mode. Estradiol and progesterone inhibit GnRH secretion, acting through kisspeptin and dynorphin in the KNDy neurons, and the inhibins act at the pituitary to selectively inhibit FSH synthesis and secretion **(Fig. 385-7)**. Estradiol also contributes to negative feedback at the pituitary with an effect that is greater for FSH than LH. This tightly regulated negative feedback control of FSH is critical for development of the single mature oocyte that characterizes normal reproductive function in women. In addition to these negative feedback controls, the menstrual cycle is uniquely dependent on estrogen-induced positive feedback to produce an LH surge that is essential for ovulation of a mature follicle. Estrogen negative feedback in women occurs primarily at the hypothalamus with a small pituitary contribution, whereas estrogen positive feedback occurs at the pituitary in women with upregulation of GnRH signaling. In women, hypothalamic

GnRH secretion plays a permissive role in generating the preovulatory gonadotropin surge, a mechanism that differs significantly from that in rodents and other species that rely on seasonal and circadian cues, in which a surge of GnRH also occurs.

FIGURE 385-7

The reproductive system in women is critically dependent on both negative feedback of gonadal steroids and inhibin to modulate follicle-stimulating hormone (FSH) secretion and on estrogen positive feedback to generate the preovulatory luteinizing hormone (LH) surge. GnRH, gonadotropin-releasing hormone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

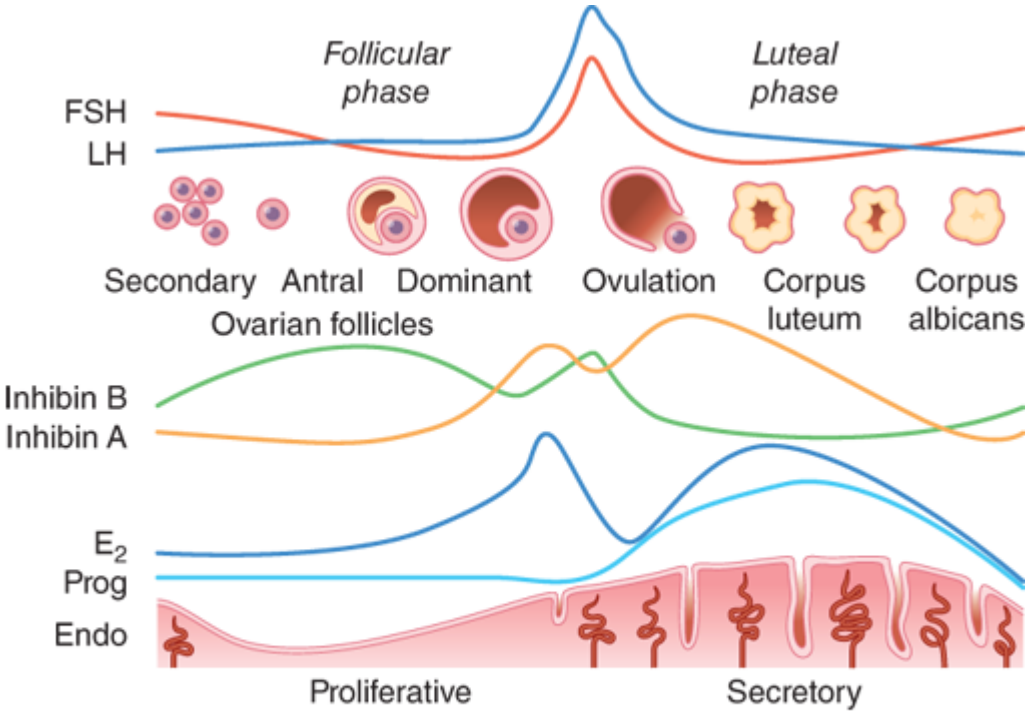
THE FOLLICULAR PHASE

The follicular phase is characterized by recruitment of a cohort of secondary follicles and the ultimate selection of a dominant preovulatory follicle **(Fig. 385-8)**. The follicular phase begins, by convention, on the first day of menses. However, follicle recruitment is initiated by the rise in FSH that begins in the late luteal phase of the previous cycle in conjunction with the loss of negative feedback of gonadal steroids and likely inhibin A. The fact that a 20–30% increase in FSH is adequate for follicular recruitment speaks to the marked sensitivity of the resting follicle pool to FSH. The resultant granulosa cell proliferation is responsible for increasing early follicular phase levels of inhibin B. Inhibin B in conjunction with rising levels of estradiol and inhibin A, restrain FSH secretion during this critical period such that only a single follicle matures in the vast majority of cycles. The increased risk of multiple gestation associated with the increased levels of FSH characteristic of advanced maternal age, or with exogenous gonadotropin administration in the treatment of infertility, attests to the importance of negative feedback regulation of FSH. With further growth of the dominant follicle, estradiol and inhibin A increase and the follicle acquires LH receptors. Increasing levels of estradiol are responsible for proliferative changes in the endometrium. The exponential rise in estradiol

results in positive feedback on the pituitary, leading to the generation of an LH surge (and a smaller FSH surge), thereby triggering ovulation and luteinization of granulosa and theca cells.

FIGURE 385-8

Relationship between gonadotropins, follicle development, gonadal secretion, and endometrial changes during the normal menstrual cycle. E_2, estradiol; Endo, endometrium; FSH, follicle-stimulating hormone; LH, luteinizing hormone; Prog, progesterone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE LUTEAL PHASE

The luteal phase begins with the formation of the corpus luteum from the ruptured follicle (Fig. 385-8). Progesterone and inhibin A are produced from the luteinized granulosa cells, which continue to aromatize theca-derived androgen precursors, producing estradiol. The combined actions of estrogen and progesterone are responsible for the secretory changes in the endometrium that are necessary for implantation. The corpus luteum is supported by LH but has a finite life span because of diminished sensitivity to LH. The demise of the corpus luteum results in a progressive decline in hormonal support of the endometrium. Inflammation or local hypoxia and ischemia result in vascular changes in the endometrium, leading to the release of cytokines, cell death, and shedding of the endometrium.

If conception occurs, hCG produced by the trophoblast binds to LH receptors on the corpus luteum, maintaining steroid hormone production and preventing involution of the corpus luteum until its hormonal function is taken over by the placenta 6–10 weeks after conception.

CLINICAL ASSESSMENT OF OVARIAN FUNCTION

Menstrual bleeding should become regular within 2–4 years of menarche, although anovulatory and irregular cycles are common before that. For the remainder of adult reproductive life, the cycle length counted from the first day of menses to the day preceding subsequent menses is ~28 days, with a range of 25–35 days. However, cycle-to-cycle variability for an individual woman is ±2 days. Luteal phase length is relatively constant between 12 and 14 days in normal cycles; thus, the major variability in cycle length is due to variations in follicular phase length. The duration of menstrual bleeding in ovulatory cycles varies between 4 and 6 days. There is a gradual shortening of cycle length with age such that women aged >35 years have cycles that are shorter than during their younger reproductive years. Anovulatory cycles increase as women approach menopause, and bleeding patterns may be erratic.

Women who report regular monthly bleeding with cycles that do not vary by >4 days generally have ovulatory cycles, but several other clinical signs can be used to assess the likelihood of ovulation. Some women experience *mittelschmerz*, described as midcycle pelvic discomfort that is thought to be caused by the rapid expansion of the dominant follicle at the time of ovulation. A constellation of premenstrual moliminal symptoms such as bloating, breast tenderness, and food cravings often occur several days before menses in ovulatory cycles, but their absence cannot be used as evidence of anovulation. Methods that can be used to determine whether ovulation is likely include a serum progesterone level >5 ng/mL ~7 days before expected menses, an increase in basal body temperature of 0.24°C (>0.5°F) in the second half of the cycle due to the thermoregulatory effect of progesterone, or the detection of the urinary LH surge using ovulation predictor kits. Because ovulation occurs ~36 h after the LH surge, urinary LH can be helpful in timing intercourse to coincide with ovulation.

Ultrasound can be used to detect the growth of the fluid-filled antrum of the developing follicle and to assess endometrial proliferation in response to increasing estradiol levels in the follicular phase. It can also be used to provide evidence of ovulation by documenting collapse of the dominant follicle and/or the presence of a corpus luteum as well as the characteristic echogenicity of the secretory endometrium of the luteal phase.

PUBERTY

NORMAL PUBERTAL DEVELOPMENT IN GIRLS

The first menstrual period (*menarche*) occurs relatively late in the series of developmental milestones that characterize normal pubertal development **(Table 385-1)**. Menarche is preceded by the appearance of pubic and then axillary hair (*adrenarche*) as a result of maturation of the zona reticularis in the adrenal gland and increased adrenal androgen secretion, particularly dehydroepiandrosterone (DHEA). The triggers for adrenarche remain unknown but may involve increases in body mass index, as well as in utero and neonatal factors. Menarche is also preceded by breast development (*thelarche*). The breast is exquisitely sensitive to the very low levels of estrogen that result from peripheral conversion of adrenal androgens and the low levels of estrogen secreted from the ovary early in pubertal maturation. Breast development precedes the appearance of pubic and axillary hair in ~60% of girls. The interval between the onset of breast development and menarche is ~2 years. There has been a gradual decline in the age of menarche over the past century,

attributed in large part to improvement in nutrition, and there is a relationship between adiposity and earlier sexual maturation in girls. In the United States, menarche occurs at an average age of 12.5 years (Table 385-1).

TABLE 385-1

Mean Age (Years) of Pubertal Milestones in Girls

	Onset of Breast/Pubic Hair Development	Age of Peak Height Velocity	Menarche	Final Breast/Pubic Hair Development	Adult Height
White	10.2	11.9	12.6	14.3	17.1
Black	9.6	11.5	12	13.6	16.5

Source: From FM Biro et al: J Pediatr 148:234, 2006.

Much of the variation in the timing of puberty is due to genetic factors. Heritability estimates from twin studies range between 50 and 80%. Adrenarche and thelarche occur ~1 year earlier in black compared with white girls, although the difference in the timing of menarche is less pronounced. Genome-wide association studies have identified over a hundred genes associated with pubertal timing in boys and girls attesting to the high degree of coordination of this reproductive and growth milestone. These findings include genes involved in GnRH secretion (e.g., *TACR3,* and the maternally imprinted gene, *MKRN3*, that has been associated with familial precocious puberty), pituitary development and function (e.g., *POU1F1*), hormone synthesis and bioactivity (e.g., *STARD4, ESR1, RXRG*), gonadal feedback (e.g., *INHBA, ESR1*), and energy homeostasis and growth including *LIN28B*, a sentinel puberty gene, which is a potent regulator of microRNA processing.

Other important hormonal changes also occur in conjunction with puberty. Growth hormone (GH) levels increase early in puberty, stimulated in part by the pubertal increase in estrogen secretion. GH increases insulin-like growth factor-I (IGF-I), which enhances linear growth. The growth spurt is generally less pronounced in girls than in boys, with a peak growth velocity of ~7 cm/year. Linear growth is ultimately limited by closure of epiphyses in the long bones as a result of prolonged exposure to estrogen. Puberty is also associated with mild insulin resistance.

DISORDERS OF PUBERTY

The differential diagnosis of precocious and delayed puberty is similar in boys **(Chap. 384)** and girls. However, there are differences in the timing of normal puberty and differences in the relative frequency of specific disorders in girls compared with boys.

Precocious Puberty

Traditionally, precocious puberty has been defined as the development of secondary sexual characteristics before the age of 8 in girls based on data from Marshall and Tanner in British girls studied in the 1960s. More recent studies led to recommendations that girls be evaluated for precocious puberty if breast development or pubic hair is present at <7 years of age for white girls or <6 years for black girls; however, these guidelines have not been widely accepted in favor of careful follow-up in girls presenting at <8 years.

Precocious puberty in girls is most often centrally mediated **(Table 385-2)**, resulting from early activation of the hypothalamic-pituitary-ovarian axis. It is characterized by pulsatile LH secretion (which is initially associated with deep sleep) and an enhanced LH and FSH response to exogenous GnRH or a GnRH agonist (two- to threefold stimulation) **(Table 385-3)**. True precocity is marked by advancement in bone age of >2 standard deviations, a recent history of growth acceleration, and progression of secondary sexual characteristics. In girls, centrally mediated precocious puberty (CPP) is idiopathic in ~85% of cases; however, neurogenic causes must be considered. Activating mutations in *KISS* and *KISS1R* have been found in a small number of patients with CPP, and loss of function mutations in *MKRN3* have been reported in familial CPP. However, the frequency of these mutations is insufficient to justify their use in routine clinical testing. GnRH agonists that induce pituitary desensitization are the mainstay of treatment to prevent premature epiphyseal closure and preserve adult height, as well as to manage psychosocial repercussions of precocious puberty.

TABLE 385-2

Differential Diagnosis of Precocious Puberty

Central (GnRH Dependent)	Peripheral (GnRH Independent)
Idiopathic	Congenital adrenal hyperplasia
CNS tumors	Estrogen-producing tumors
Hamartomas	Adrenal tumors
Astrocytomas	Ovarian tumors
Adenomyomas	Gonadotropin/hCG-producing tumors
Gliomas	Exogenous exposure to estrogen or androgen or lavender or tea-tree oil
Germinomas	McCune-Albright syndrome
CNS infection	Aromatase excess syndrome
Head trauma	
Iatrogenic	
Radiation	
Chemotherapy	
Surgical	
CNS malformation	
Arachnoid or suprasellar cysts	
Septo-optic dysplasia	
Hydrocephalus	

Abbreviations: CNS, central nervous system; GnRH, gonadotropin-releasing hormone; hCG, human chorionic gonadotropin.

TABLE 385-3

Evaluation of Precocious and Delayed Puberty

	Precocious	Delayed
Initial Screening Tests		
History and physical	×	×
Assessment of growth velocity	×	×
Bone age	×	×
LH, FSH	×	×
Estradiol, testosterone	×	×
DHEAS	×	×
17-Hydroxyprogesterone	×	
TSH, T_4	×	×
Complete blood count		×
Sedimentation rate, C-reactive protein		×
Electrolytes, renal function		×
Liver enzymes		×
IGF-I, IGFBP-3		×
Urinalysis		×
Secondary Tests		
Pelvic ultrasound	×	×
Cranial MRI	×	×
β-hCG	×	

	Precocious	Delayed
GnRH/agonist stimulation test	×	×
ACTH stimulation test	×	
Inflammatory bowel disease panel	×	×
Celiac disease panel		×
Prolactin		×
Karyotype		×

Abbreviations: ACTH, adrenocorticotropic hormone; DHEAS, dehydroepiandrosterone sulfate; FSH, follicle-stimulating hormone; hCG, human chorionic gonadotropin; IGF-I, insulin-like growth factor-I; IGFBP-3, IGF-binding protein 3; LH, luteinizing hormone; MRI, magnetic resonance imaging; TSH, thyroid-stimulating hormone; T_4, thyroxine.

Peripherally mediated precocious puberty does not involve activation of the hypothalamic-pituitary-ovarian axis and is characterized by suppressed gonadotropins in the presence of elevated estradiol. Management of peripheral precocious puberty involves treating the underlying disorder (Table 385-2) and limiting the effects of gonadal steroids using aromatase inhibitors, inhibitors of steroidogenesis, and ER blockers. It is important to be aware that central precocious puberty can also develop in girls whose precocity was initially peripherally mediated, as in McCune-Albright syndrome and congenital adrenal hyperplasia.

Incomplete and intermittent forms of precocious puberty may also occur. For example, premature breast development may occur in girls before the age of 2 years, with no further progression and without significant advancement in bone age, estrogen production, or compromised height. Premature adrenarche can also occur in the absence of progressive pubertal development, but it must be distinguished from late-onset congenital adrenal hyperplasia and androgen-secreting tumors, in which case it may be termed *heterosexual precocity*. Premature adrenarche may be associated with obesity, hyperinsulinemia, and the subsequent predisposition to PCOS.

Delayed Puberty

Delayed puberty **(Table 385-4)** is defined as the absence of secondary sexual characteristics by age 13 in girls. The diagnostic considerations are very similar to those for primary amenorrhea **(Chap. 386)**. Between 25 and 40% of delayed puberty in girls is of ovarian origin, with Turner's syndrome accounting for the majority of such patients. Delayed puberty may occur in the setting of systemic illnesses, including celiac disease and chronic renal disease, and endocrinopathies such as diabetes and hypothyroidism. In addition, girls appear to be particularly susceptible to the adverse effects of decreased energy balance resulting from exercise,

dieting, and/or eating disorders and thus, functional hypothalamic amenorrhea (HA) can present with primary amenorrhea. Together these reversible conditions account for ~25% of delayed puberty in girls. Congenital hypogonadotropic hypogonadism in girls or boys can be caused by mutations in several different genes or combinations of genes (Fig. 385-4, **Chap. 384**, Table 385-2). Approximately 50% of girls with congenital hypogonadotropic hypogonadism, with or without anosmia, have a history of some degree of breast development, and 10% report one to two episodes of vaginal bleeding. Family studies suggest that genes identified in association with absent puberty may also cause delayed puberty, and recent reports have further suggested that a genetic susceptibility to environmental stresses such as diet and exercise may account for at least some cases of functional HA, including in girls who present with primary amenorrhea. Although neuroanatomic causes of delayed puberty are considerably less common in girls than in boys, it is always important to rule these out in the setting of hypogonadotropic hypogonadism.

TABLE 385-4

Differential Diagnosis of Delayed Puberty

Hypergonadotropic
Ovarian
Turner’s syndrome
Gonadal dysgenesis
Chemotherapy/radiation therapy
Galactosemia
Autoimmune oophoritis
Congenital lipoid hyperplasia
Steroidogenic enzyme abnormalities
17α-Hydroxylase deficiency
Aromatase deficiency
Gonadotropin/receptor mutations
*FSH*β*, LHR, FSHR*
Androgen resistance syndrome
Hypogonadotropic
Genetic
Hypothalamic syndromes
Leptin/leptin receptor
HESX1 (septo-optic dysplasia)
PC1 (prohormone convertase)

IHH and Kallmann's syndrome
KAL1, FGF8, FGFR1, NSMF, PROK2, PROKR2, SEM3A, HS6ST1, WDR11, CHD7
KISS1, KISS1R, TAC3, TAC3R, GnRH1, GnRHR, and others
Abnormalities of pituitary development/function
PROP1
CNS tumors/infiltrative disorders
Craniopharyngioma
Astrocytoma, germinoma, glioma
Prolactinomas, other pituitary tumors
Histiocytosis X
Chemotherapy/radiation
Functional
Chronic diseases
Malnutrition
Excessive exercise
Eating disorders

Abbreviations: CHD7, chromodomain-helicase-DNA-binding protein 7; CNS, central nervous system; *FGF8*, fibroblast growth factor 8; *FGFR1*, fibroblast growth factor 1 receptor; *FSH*β, follicle-stimulating hormone β chain; *FSHR*, FSH receptor; *GNRHR*, gonadotropin-releasing hormone receptor; *HESX1*, homeobox, embryonic stem cell expressed 1; *HS6ST1*, heparin sulfate 6-O sulfotransferase 1; IHH, idiopathic hypogonadotropic hypogonadism; KAL, Kallmann; *KISS1*, kisspeptin 1; *KISSR1*, KISS1 receptor; *LHR*, luteinizing hormone receptor; *NSMF*, NMDA receptor synaptonuclear signaling and neuronal migration factor; *PROK2*, prokineticin 2; *PROKR2* prokineticin receptor 2; *PROP1*, prophet of Pit1, paired-like homeodomain transcription factor *SEMA3A*, semaphorin-3A; *WDR11*, WD repeat-containing protein 11.

FURTHER READING

Devoto L et al: The significance of estradiol metabolites in human corpus luteum physiology. Steroids 123:50, 2017. [PubMed: 28502859]

Francois CM et al: A novel action of follicle-stimulating hormone in the ovary promotes estradiol production without inducing excessive follicular growth before puberty. Sci Rep 7:46222, 2017. [PubMed: 28397811]

George JT et al: Effect of gonadotropin-inhibitory hormone on luteinizing hormone secretion in humans. Clin Endocrinol (Oxf) 86:731, 2017. [PubMed: 28186349]

Kaplowitz P, Bloch C: The section on endocrinology. Evaluation and referral of children with signs of early puberty. Pediatrics 137:e20153732, 2016.

Macedo DB et al: Sexual precocity—Genetic bases of central precocious puberty and automonous gonadal activation. Endocr Dev 29:50, 2016. [PubMed: 26680572]

Marshall SA et al: The role of relaxin in normal and abnormal uterine function during the menstrual cycle and early pregnancy. Reprod Sci 24:342, 2017. [PubMed: 27365367]

Perry JR et al: Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature 514:92, 2014. [PubMed: 25231870]

Skorupskaite K et al: Interactions between neurokinin B and kisspeptin in mediating estrogen feedback in healthy women. J Clin Endocrinol Metab 101:4628, 2016. [PubMed: 27636018]

Stamou MI et al: Discovering genes essential to the hypothalamic regulation of human reproduction using a human disease model: Adjusting to life in the “-omics” era. Endocr Rev 36:603, 2015. [PubMed: 26394276]

Tsutsui K, Ubuka T: GnIH control of feeding and reproductive behaviors. Front Endocrinol (Lausanne) 7:170, 2016. [PubMed: 28082949]

Tucker EJ et al: Premature ovarian insufficiency: New perspectives on genetic cause and phenotypic spectrum. Endocr Rev 37:609, 2016. [PubMed: 27690531]

Willemsen RH, Dunger DB: Normal variation in pubertal timing: Genetic determinants in relation to growth and adiposity. Endocr Dev 29:17, 2016. [PubMed: 26680570]

Zhu J et al: A shared genetic basis for self-limited delayed puberty and idiopathic hypergonadotropic hypogonadism. J Clin Endocrinol Metab 100:E646, 2015. [PubMed: 25636053]

Chapter 386: Menstrual Disorders and Pelvic Pain

Janet E. Hall

INTRODUCTION

Menstrual dysfunction can signal an underlying abnormality that may have long-term health consequences. Although frequent or prolonged bleeding usually prompts a woman to seek medical attention, infrequent or absent bleeding may seem less troubling and the patient may not bring it to the attention of the physician. Thus, a focused menstrual history is a critical part of every encounter with a female patient. Pelvic pain is a common complaint that may relate to an abnormality of the reproductive organs but also may be of gastrointestinal, urinary tract, or musculoskeletal origin. Depending on its cause, pelvic pain may require urgent surgical attention. Recent guidelines no longer recommend routine pelvic examination in asymptomatic, average-risk women other than periodic cervical cancer screening. However, pelvic examination is an important part of the evaluation of amenorrhea, abnormal uterine bleeding, and pelvic pain.

MENSTRUAL DISORDERS

DEFINITION AND PREVALENCE

Amenorrhea refers to the absence of menstrual periods. Amenorrhea is classified as *primary* if menstrual bleeding has never occurred in the absence of hormonal treatment or *secondary* if menstrual periods cease for 3–6 months. Primary amenorrhea is a rare disorder that occurs in <1% of the female population. However, between 3 and 5% of women experience at least 3 months of secondary amenorrhea in any specific year. There is no evidence that race or ethnicity influences the prevalence of amenorrhea. However, because of the importance of adequate nutrition for normal reproductive function, both the age at menarche and the prevalence of secondary amenorrhea vary significantly in different parts of the world.

Oligomenorrhea is defined as a cycle length >35 days or <10 menses per year. Both the frequency and the amount of vaginal bleeding are irregular in oligomenorrhea, and moliminal symptoms (premenstrual breast tenderness, food cravings, mood lability), suggestive of ovulation, are variably present. Anovulation can also present with intermenstrual intervals <24 days or vaginal bleeding for >7 days. Frequent or heavy irregular bleeding is termed *dysfunctional uterine bleeding* if anatomic uterine and outflow tract lesions or a bleeding diathesis have been excluded. Oligo- or anovulation are most frequently associated with polycystic ovarian syndrome (PCOS).

Primary Amenorrhea

The absence of menarche (the first menstrual period) by age 16 has been used traditionally to define primary amenorrhea. However, other factors, such as growth, secondary sexual characteristics, and the presence of cyclic pelvic pain, also influence the age at which primary amenorrhea should be investigated. Recent studies suggest that puberty is occurring at an earlier age, particularly in obese girls. However, it is important to note that these data reflect earlier breast development alone with minimal change in the age of menarche. Thus, an evaluation for amenorrhea should be initiated by age 15 or 16 in the presence of normal growth and secondary sexual characteristics; age 13 in the absence of secondary sexual

characteristics or if height is less than the third percentile; age 12 or 13 in the presence of breast development and cyclic pelvic pain; or within 2 years of breast development if menarche, has not occurred.

Secondary Amenorrhea or Oligomenorrhea

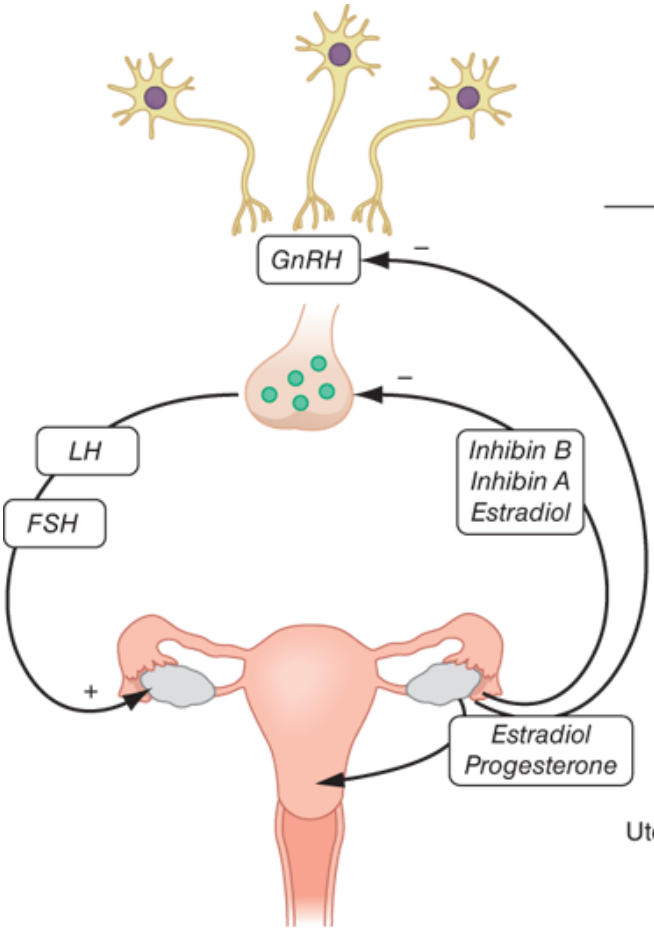
Irregular cycles are relatively common for up to 3 years after menarche and for 1–2 years before the final menstrual period. In the intervening years, menstrual cycle length is ~28 days, with an intermenstrual interval normally ranging between 25 and 35 days. Cycle-to-cycle variability in an individual woman who is ovulating consistently is generally +/− 2 days. Pregnancy is the most common cause of amenorrhea and should be excluded early in any evaluation of menstrual irregularity. However, many women occasionally miss a single period. Three months of secondary amenorrhea, or 6 months in women with previously irregular cycles, should prompt an evaluation, as should a history of intermenstrual intervals >35 or <21 days or bleeding that persists for >7 days.

DIAGNOSIS

Pregnancy is the most common cause of amenorrhea, and must be excluded in all cases, regardless of patient history. Evaluation of menstrual dysfunction depends on understanding the interrelationships between the four critical components of the reproductive tract: (1) the hypothalamus, (2) the pituitary, (3) the ovaries, and (4) the uterus and outflow tract **(Fig. 386-1; Chap. 385)**. This system is maintained by complex negative and positive feedback loops involving the ovarian steroids (estradiol and progesterone) and peptides (inhibin B and inhibin A) and the hypothalamic (gonadotropin-releasing hormone [GnRH]) and pituitary (follicle-stimulating hormone [FSH] and luteinizing hormone [LH]) components of this system (Fig. 386-1).

FIGURE 386-1

Role of the hypothalamic-pituitary-gonadal axis in the etiology of amenorrhea. Gonadotropin-releasing hormone (GnRH) secretion from the hypothalamus stimulates follicle-stimulating hormone (FSH) and luteinizing hormone (LH) secretion from the pituitary to induce ovarian folliculogenesis and steroidogenesis. Ovarian secretion of estradiol and progesterone controls the shedding of the endometrium, resulting in menses, and, in combination with the inhibins, provides feedback regulation of the hypothalamus and pituitary to control secretion of FSH and LH. The prevalence of amenorrhea resulting from abnormalities at each level of the reproductive system (hypothalamus, pituitary, ovary, uterus, and outflow tract) varies depending on whether amenorrhea is primary or secondary. PCOS, polycystic ovarian syndrome.



	Primary	Secondary
Hypothalamus	27%	36%
Pituitary	2%	15%
PCOS	7%	30%
Ovary	43%	12%
Uterus/outflow tract	19%	7%

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Disorders of menstrual function can be thought of in two main categories: disorders of the uterus and outflow tract and disorders of ovulation. Many of the conditions that cause primary amenorrhea are congenital but go unrecognized until the time of normal puberty (e.g., genetic, chromosomal, and anatomic abnormalities). All causes of secondary amenorrhea also can cause primary amenorrhea.

Disorders of the Uterus or Outflow Tract

Abnormalities of the uterus and outflow tract typically present as primary amenorrhea. In patients with normal pubertal development and a blind vagina, the differential diagnosis includes *obstruction* by a transverse vaginal septum or imperforate hymen; *müllerian agenesis* (Mayer-Rokitansky-Kuster-Hauser syndrome), which can be caused by mutations in the *WNT4* gene; and *androgen insensitivity syndrome* (AIS), which is an X-linked recessive disorder that accounts for ~10% of all cases of primary amenorrhea (Chap. 384). Patients with AIS have a 46,XY karyotype, but because of the lack of androgen receptor responsiveness, those with complete AIS lack features of androgenization and have female external genitalia. The absence of pubic and axillary hair distinguishes them clinically from patients with müllerian agenesis, as does a testosterone level in the male range. The rare patient with 5α reductase type 2 enzyme deficiency has a similar presentation, but undergoes virilization at the time of puberty. *Asherman's syndrome* presents as secondary amenorrhea or hypomenorrhea and results from partial or complete obliteration of the uterine cavity by adhesions that prevent normal growth and shedding of the endometrium. Curettage performed for pregnancy complications accounts for >90% of cases; genital tuberculosis is an important cause in regions where it is endemic.

TREATMENT

TREATMENT

Disorders of the Uterus or Outflow Tract

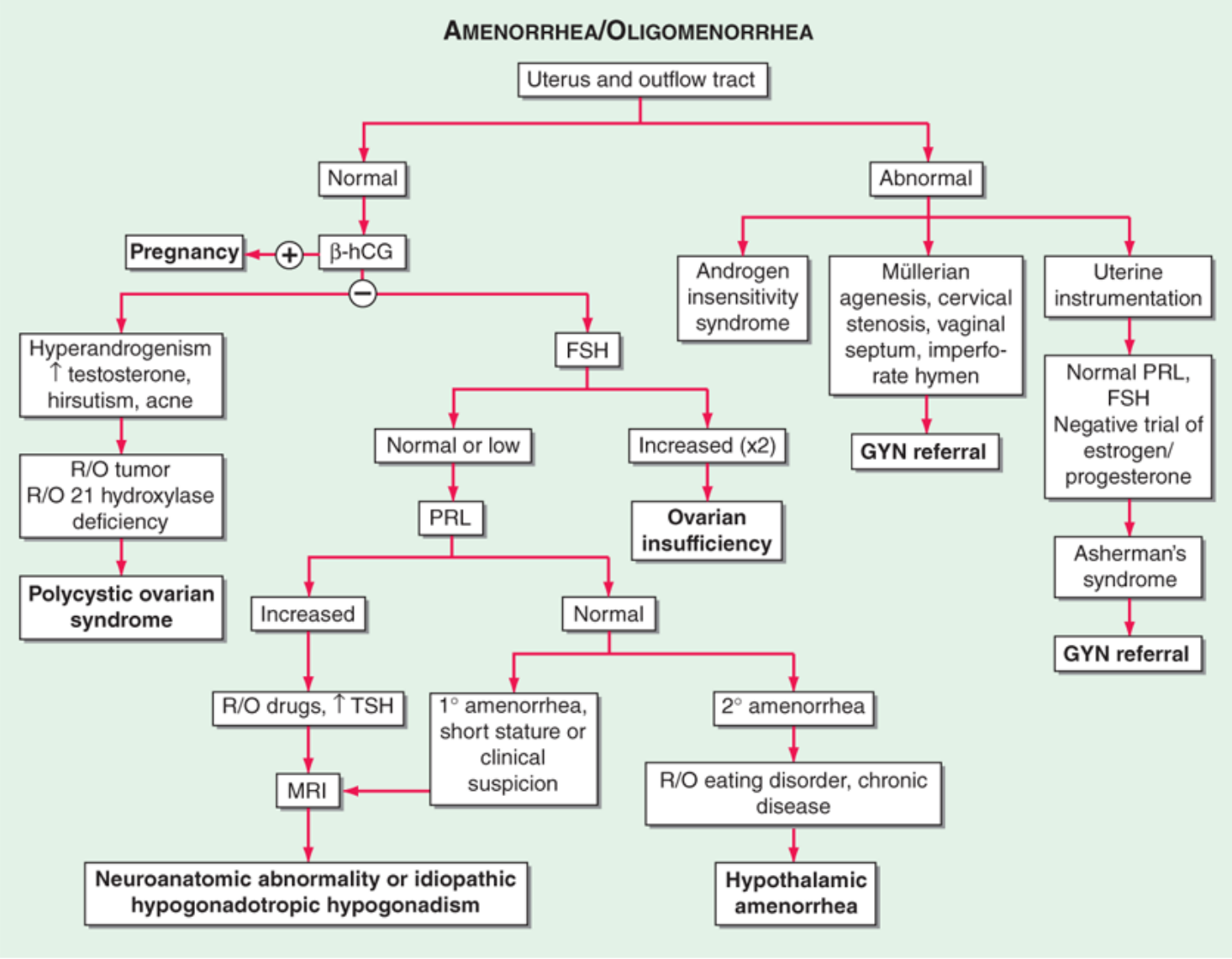
Obstruction of the outflow tract requires surgical correction. It is important that this be performed as soon as the diagnosis is made as the risk of endometriosis is increased with retrograde menstrual flow. *Müllerian agenesis* may require surgical intervention to allow sexual intercourse, although vaginal dilatation is adequate in some patients. Because ovarian function is normal, assisted reproductive techniques can be used with a surrogate carrier. *Androgen resistance syndrome* requires gonadectomy because there is risk of gonadoblastoma in the dysgenetic gonads, although surgery is generally delayed until after breast development and the pubertal growth spurt. Estrogen replacement is indicated after gonadectomy, and vaginal dilatation may be required to allow sexual intercourse.

Disorders of Ovulation

Once uterus and outflow tract abnormalities have been excluded, other causes of amenorrhea involve disorders of ovulation. The differential diagnosis is based on the results of initial tests, including a pregnancy test, an FSH level (to determine whether the cause is likely to be ovarian or central), and assessment of hyperandrogenism (Fig. 386-2).

FIGURE 386-2

Algorithm for evaluation of amenorrhea. β-hCG, human chorionic gonadotropin; FSH, follicle-stimulating hormone; GYN, gynecologist; MRI, magnetic resonance imaging; PRL, prolactin; R/O, rule out; TSH, thyroid-stimulating hormone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HYPOGONADOTROPIC HYPOGONADISM

Low estrogen levels in combination with normal or low levels of LH and FSH are seen with anatomic, genetic, or functional abnormalities that interfere with hypothalamic GnRH secretion or normal pituitary responsiveness to GnRH. Although relatively uncommon, tumors and infiltrative diseases should be considered in the differential diagnosis of hypogonadotropic hypogonadism (Chap. 373). These disorders may present with primary or secondary amenorrhea. They may occur in association with other features suggestive of hypothalamic or pituitary dysfunction, such as short stature, diabetes insipidus, galactorrhea, and headache. Hypogonadotropic hypogonadism also may be seen after cranial irradiation. In the postpartum period, amenorrhea occurs normally in association with breast feeding, but may also be caused by pituitary necrosis (Sheehan's syndrome) or lymphocytic hypophysitis. Because reproductive dysfunction is commonly associated with hyperprolactinemia from neuroanatomic lesions or medications, prolactin should be measured in all patients with hypogonadotropic hypogonadism (Chap. 373).

Isolated hypogonadotropic hypogonadism (IHH) occurs in women, although it is three times more common in men. IHH generally presents with primary amenorrhea, although 50% have some degree of breast development, and ~10% report one to two menses. IHH is associated with anosmia in half of women (termed Kallmann's syndrome). Genetic causes of IHH have been identified in ~50% of patients (Chaps. 384 and 385).

Functional hypothalamic amenorrhea (HA) is a diagnosis of exclusion of other causes of hypogonadotropic hypogonadism including chronic diseases (type 1 diabetes, celiac disease, hyperthyroidism, Cushing Syndrome) and use of opioids, glucocorticoids or psychotropic medications that increase prolactin levels. Functional HA is most commonly associated with conditions causing a mismatch between energy expenditure and energy intake and/or significant stress. Variants in genes associated with IHH may increase susceptibility to these environmental inputs, accounting in part for the clinical variability in this disorder. Metabolic and stress signaling is transduced to the reproductive axis, at least in part, through leptin signaling from the periphery and via hypothalamic kisspeptin control of GnRH. The diagnosis of HA generally can be made on the basis of a careful history, a physical examination, and the demonstration of low levels of gonadotropins and normal prolactin levels. Eating disorders and chronic disease must be specifically excluded. An atypical history, headache, signs of other hypothalamic dysfunction, or hyperprolactinemia, even if mild, necessitates cranial magnetic resonance imaging (MRI) to exclude a neuroanatomic cause. Up to 10% of women with HA may have some features of PCOS (increased ovarian volume, higher anti-mullerian hormone [AMH] levels, and slightly elevated androgen levels).

HYPERGONADOTROPIC HYPOGONADISM

Ovarian failure is considered premature when it occurs in women <40 years old and accounts for ~10% of secondary amenorrhea. *Primary ovarian insufficiency* (POI) has replaced the terms *premature menopause* and *premature ovarian failure* in recognition of the continuum of impaired ovarian function encompassed by this disorder. Ovarian insufficiency is associated with the loss of negative-feedback restraint on the hypothalamus and pituitary, resulting in increased FSH and LH levels. FSH is a better marker of ovarian failure because of loss of negative feedback effects of both estradiol and the inhibins and because its levels are less variable than those of LH. AMH levels will also be low in patients with POI, but are more frequently used in management of infertility. As with natural menopause, POI may wax and wane, and serial measurements may be necessary to establish the diagnosis.

Once the diagnosis of POI has been established, further evaluation is indicated because of other health problems that may be associated with POI. Although POI is most commonly of unknown cause, it also occurs in association with a variety of chromosomal abnormalities (most often Turner's syndrome), autoimmune polyglandular failure syndromes, and other rare disorders. Radiotherapy and chemotherapy may reduce ovarian reserve, with effects on both the oocytes and the supporting granulosa cells. New approaches, including ovarian preservation, are being developed to support long-term fertility choices in women of reproductive age prior to oncologic treatment. The recognition that early ovarian failure occurs in premutation carriers of the fragile X syndrome is important because of the increased risk of severe mental retardation in male children with *FMR1* mutations. Thus, follow-up testing should include a karyotype in all POI patients, serum anti-cortical and 21-hydroxylase antibodies (specific but not sensitive for subsequent adrenal insufficiency), thyroid function and thyroid peroxidase antibodies, *FMR1* premutation screening, and assessment of bone mineral density. Ovarian biopsy is of no diagnostic or prognostic value. Although the number of genetic causes POI is increasing, routine testing for mutations other than *FMR1* is currently not recommended.

Hypergonadotropic hypogonadism occurs rarely in other disorders, such as mutations in the FSH or LH receptors. Aromatase deficiency and 17α-hydroxylase deficiency are associated with decreased estrogen and elevated gonadotropins and with hyperandrogenism and hypertension, respectively. Gonadotropin-secreting tumors in women of reproductive age generally present with high, rather than low, estrogen levels and cause ovarian hyperstimulation or dysfunctional bleeding.

TREATMENT

TREATMENT

Hypo- and Hypergonadotropic Causes of Amenorrhea

Amenorrhea almost always is associated with chronically low levels of estrogen, whether it is caused by hypogonadotropic hypogonadism or ovarian insufficiency. Development of secondary sexual characteristics requires gradual titration of estradiol replacement with eventual addition of progestin. Hormone replacement with either low-dose

estrogen/progesterone regimens or oral contraceptive pills is recommended until the usual age of menopause for bone and cardiovascular protection. In women with functional HA or anorexia nervosa, hormone replacement alone may not be sufficient to restore or maintain bone density. Patients with hypogonadotropic hypogonadism who are interested in fertility require treatment with both exogenous FSH and LH or pulsatile GnRH. Patients with ovarian failure can consider oocyte donation, which has a high rate of success in this population, although its use in women with Turner's syndrome is limited by significant maternal cardiovascular risk.

POLYCYSTIC OVARIAN SYNDROME

PCOS is diagnosed based on a combination of clinical or biochemical evidence of hyperandrogenism, amenorrhea or oligomenorrhea, and the ultrasound appearance of polycystic ovaries. Approximately half of patients with PCOS are obese, and abnormalities in insulin dynamics are common, as is metabolic syndrome. Symptoms generally begin shortly after menarche and are slowly progressive. Lean oligo-ovulatory patients with PCOS generally have high LH levels in the presence of normal to low levels of FSH and estradiol, although these may be suppressed by undernutrition or stress. The LH/FSH ratio is less pronounced in obese patients in whom insulin resistance is a more prominent feature.

TREATMENT

TREATMENT

Polycystic Ovarian Syndrome

A major abnormality in patients with PCOS is the failure of regular, predictable ovulation. Thus, these patients are at risk for the development of dysfunctional bleeding and endometrial hyperplasia associated with unopposed estrogen exposure. Endometrial protection can be achieved with the use of oral contraceptives or progestins (medroxyprogesterone acetate, 5–10 mg, or prometrium, 200 mg daily for 10–14 days of each month). Oral contraceptives are also useful for management of hyperandrogenic symptoms, as are spironolactone and cyproterone acetate (not available in the United States), which function as weak androgen receptor blockers. Management of the associated metabolic syndrome may be appropriate for some patients (**Chap. 401**). For patients interested in fertility, weight control is a critical first step. Clomiphene citrate is highly effective as a first-line treatment, as is the aromatase inhibitor letrozole. Exogenous gonadotropins can be used by experienced practitioners; a diagnosis of polycystic ovaries increases the risk of hyperstimulation, even in women with regular, ovulatory menstrual cycles. Metformin is frequently used in patients with PCOS, and is appropriate as an adjunct with diet and exercise for obese women with PCOS, or for treatment of diabetes or impaired glucose tolerance, as in non-PCOS patients. However, metformin is not recommended for endometrial protection or treatment of hyperandrogenic symptoms, infertility, pregnancy loss or prevention of gestational diabetes.

PELVIC PAIN

The mechanisms that cause pelvic pain are similar to those that cause abdominal pain (**Chap. 12**) and include inflammation of the parietal peritoneum, obstruction of hollow viscera, vascular disturbances, and pain originating in the abdominal wall. Pelvic pain may reflect pelvic disease per se but also may reflect extrapelvic disorders that refer pain to the pelvis. In up to 60% of cases, pelvic pain can be attributed to gastrointestinal problems, including appendicitis, cholecystitis, infections, intestinal obstruction, diverticulitis, and inflammatory bowel disease. Urinary tract and musculoskeletal disorders are also common causes of pelvic pain.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Pelvic Pain

As with all types of abdominal pain, the first priority is to identify life-threatening conditions (shock, peritoneal signs) that may require emergent surgical management. The possibility of pregnancy should be identified as soon as possible by menstrual history and/or testing. A thorough history that includes the type, location, radiation, and status with respect to increasing or decreasing severity can help identify the cause of acute pelvic pain. Specific associations with vaginal bleeding, sexual activity, defecation, urination, movement, or eating should be specifically sought. Determination of whether the pain is acute versus chronic and cyclic versus noncyclic will direct further investigation **(Table 386-1)**. However, disorders that cause cyclic pain occasionally may cause noncyclic pain, and the converse is also true.

TABLE 386-1

Causes of Pelvic Pain

	Acute	Chronic
Cyclic pelvic pain		Mittelschmerz Dysmenorrhea Endometriosis
Noncyclic pelvic pain	Pelvic inflammatory disease Ruptured or hemorrhagic ovarian cyst, endometrioma, or ovarian torsion Ectopic pregnancy Endometritis Acute growth or degeneration of uterine myoma Threatened abortion	Pelvic congestion syndrome Adhesions and retroversion of the uterus Pelvic malignancy Vulvodynia Chronic pelvic inflammatory disease Tuberculous salpingitis History of sexual abuse

ACUTE PELVIC PAIN

Pelvic inflammatory disease (PID) most commonly presents with bilateral lower abdominal pain. It is generally of recent onset and is exacerbated by intercourse or jarring movements. Fever is present in about half of these patients; abnormal uterine bleeding occurs in about one-third. New vaginal discharge, urethritis, and chills may be present but are less specific signs. With public health efforts to control sexually transmitted diseases, the rate and severity of PID have declined in the United States and Europe; however, this is not the case in the developing world. Subclinical PID with its attendant risks of infertility and ectopic pregnancy remains a significant problem worldwide. Public health and professional organizations recommend annual testing for *C. trachomatis* in all sexually active women <25 and both *C. trachomatis* and *N. gonorrhea* in all women at increased risk. *Adnexal pathology* can present acutely and may be due to rupture, bleeding or torsion of cysts, or, much less commonly, the fallopian tubes. Neoplasms of the ovary or fallopian tube are much less common causes. Fever may be present with ovarian torsion. *Ectopic pregnancy* is associated with right- or left-sided lower abdominal pain, with clinical signs generally appearing 6–8 weeks after the last normal menstrual period. Vaginal bleeding occurs in ~50% of cases. Orthostatic signs and fever may be present. Risk factors include the presence of known tubal disease, previous ectopic pregnancies, a history of infertility, diethylstilbestrol (DES) exposure of the mother in utero, or a history of pelvic infections. Rupture of the fallopian tube remains a life-threatening emergency; the incidence depends on access to care but is ~18% in developed countries. *Threatened abortion* may also present with amenorrhea, abdominal pain, and vaginal bleeding. Although more common than ectopic pregnancy, it is rarely associated with systemic signs. *Uterine pathology* includes

endometritis and, less frequently, degenerating leiomyomas (fibroids). Endometritis often is associated with vaginal bleeding and systemic signs of infection. It occurs in the setting of sexually transmitted infections, uterine instrumentation, or postpartum infection.

A sensitive pregnancy test, complete blood count with differential, urinalysis, tests for chlamydial and gonococcal infections, and abdominal ultrasound aid in making the diagnosis and directing further management.

TREATMENT

TREATMENT

Acute Pelvic Pain

Treatment of acute pelvic pain depends on the suspected etiology but may require surgical or gynecologic intervention. Conservative management is an important consideration for ovarian cysts, if torsion is not suspected, to avoid unnecessary pelvic surgery and the subsequent risk of infertility due to adhesions. Surgical treatment may be required for ectopic pregnancies; however, women presenting with unruptured ectopic pregnancies may be appropriate for treatment with methotrexate, which is effective in ~90% of cases when multiple doses are used.

TREATMENT

TREATMENT

Chronic Pelvic Pain

Some women experience discomfort at the time of ovulation (*mittelschmerz*). The pain can be quite intense but is generally of short duration. The mechanism is thought to involve rapid expansion of the dominant follicle, although it also may be caused by peritoneal irritation by follicular fluid released at the time of ovulation.

DYSMENORRHEA

Dysmenorrhea refers to the crampy lower abdominal midline discomfort that begins with the onset of menstrual bleeding and gradually decreases over the next 12–72 h. It may be associated with nausea, diarrhea, fatigue, and headache and occurs in 60–93% of adolescents, beginning with the establishment of regular ovulatory cycles. Its prevalence decreases after pregnancy and with the use of oral contraceptives.

Primary dysmenorrhea results, in a majority of cases, from hormone-dependent prostaglandin (PG)-pathway mechanisms that cause intense uterine contractions, decreased blood flow, and increased peripheral nerve hypersensitivity, resulting in pain. However, variability in response to COX inhibitors suggests that PG-independent pathways, such as platelet activating factor, may also mediate inflammation.

Secondary dysmenorrhea is caused by underlying pelvic pathology. *Endometriosis* results from the presence of endometrial glands and stroma outside the uterus. These deposits of ectopic endometrium respond to hormonal stimulation and cause dysmenorrhea, which begins several days before menses. Endometriosis also may be associated with painful intercourse, painful bowel movements, and tender nodules in the uterosacral ligament. Fibrosis and adhesions can produce lateral displacement of the cervix, which is a useful sign on speculum examination. Transvaginal pelvic ultrasound is part of the initial workup and may detect an endometrioma within the ovary, rectovaginal or bladder nodules, or ureteral involvement. The CA125 level may be increased, but it has low negative predictive value. Definitive diagnosis requires laparoscopy. Symptomatology does not always predict the extent of endometriosis. The prevalence is lower in black and Hispanic women than in Caucasians and Asians. *Other secondary causes* of dysmenorrhea include adenomyosis, a condition caused by the presence of ectopic endometrial glands and stroma within the myometrium. Cervical stenosis, which may result from

trauma, infection, or surgery also may cause pain associated with menses. Pelvic congestion syndrome is associated with pelvic varicosities with low blood flow.

TREATMENT

TREATMENT

Dysmenorrhea

Local application of heat is of some benefit. Exercise, sexual activity, a vegetarian diet, use of vitamins D, B_1, B_6, and E and fish oil, acupuncture, and yoga have all been suggested to be of benefit but studies are not adequate to provide recommendations. However, nonsteroidal anti-inflammatory drugs (NSAIDs) are very effective and provide >80% sustained response rates. Ibuprofen, naproxen, ketoprofen, mefanamic acid, and nimesulide are all superior to placebo. Treatment should be started a day before expected menses and continued for 2–3 days. Oral contraceptives also reduce symptoms of dysmenorrhea. The use of tocolytics, antiphosphodiesterase inhibitors, and magnesium has been suggested, but there are insufficient data to recommend them. Failure of response to NSAIDs and/or oral contraceptives is suggestive of a pelvic disorder such as *endometriosis*, and diagnostic laparoscopy should be considered to guide further treatment.

FURTHER READING

Bloomfield H et al: Screening pelvic examinations in asymptomatic average risk adult women. WA-ESP Project #09-009; 2013.

Bouilly J et al: Identification of multiple gene mutations accounts for the new genetic architecture of ovarian insufficiency. J Clin Endocrinol Metab 101:4541, 2016. [PubMed: 27603904]

Brunham RC et al: Pelvic inflammatory disease. N Engl J Med 372:2039, 2015. [PubMed: 25992748]

Fourman LR, Fazeli PK: Neuroendocrine causes of amenorrhea—An update. J Clin Endocrinol Metab 100:812, 2015. [PubMed: 25581597]

Ju H et al: The prevalence and risk factors of dysmenorrhea. Epidem Rev 36:104, 2014.

Morley LC et al: On behalf of the Royal College of Obstetricians and Gynecologists. Metformin therapy for the management of infertility in women with polycystic ovary syndrome. Scientific Impact Paper No. 13. BJOG 124:e306, 2017. [PubMed: 28834093]

Oladosu FA et al: Nonsteroidal anti-inflammatory drug resistance in dysmenorrhea: Epidemiology, causes, and treatment. Am J Obstet Gynecol Sept 6, 2017 [Epub ahead of print].

Vercellini P et al: Endometriosis: Pathogenesis and treatment. Nat Rev Endocrinol 10:261, 2014. [PubMed: 24366116]

Chapter 387: Hirsutism

David A. Ehrmann

INTRODUCTION

Hirsutism, which is defined as androgen-dependent excessive male-pattern hair growth, affects ~10% of women. Hirsutism is most often idiopathic or the consequence of androgen excess associated with polycystic ovary syndrome (PCOS). Less frequently, it may result from adrenal androgen overproduction as occurs in nonclassic congenital adrenal hyperplasia (CAH) **(Table 387-1)**. Rarely, it is a harbinger of a serious underlying condition. Cutaneous manifestations commonly associated with hirsutism include acne and male-pattern balding (androgenic alopecia). *Virilization* refers to a condition in which androgen levels are sufficiently high to cause additional signs and symptoms, such as deepening of the voice, breast atrophy, increased muscle bulk, clitoromegaly, and increased libido. Virilization may be due to benign hyperplasia of ovarian theca and stroma cells *(*e.g., *hyperthecosis)*; it may also result from an ovarian or adrenal neoplasm.

TABLE 387-1

Causes of Hirsutism

Gonadal hyperandrogenism
Ovarian hyperandrogenism
Polycystic ovary syndrome/functional ovarian hyperandrogenism
Ovarian steroidogenic blocks
Syndromes of extreme insulin resistance
Ovarian neoplasms
Hyperthecosis
Adrenal hyperandrogenism
Premature adrenarche
Functional adrenal hyperandrogenism
Congenital adrenal hyperplasia (nonclassic and classic)
Abnormal cortisol action/metabolism
Adrenal neoplasms
Other endocrine disorders
Cushing's syndrome
Hyperprolactinemia
Acromegaly
Peripheral androgen overproduction
Obesity
Idiopathic
Pregnancy-related hyperandrogenism

Hyperreactio luteinalis
Thecoma of pregnancy
Drugs
Androgens
Oral contraceptives containing androgenic progestins
Minoxidil
Phenytoin
Diazoxide
Cyclosporine
Valproic Acid
True hermaphroditism

HAIR FOLLICLE GROWTH AND DIFFERENTIATION

Hair can be categorized as either *vellus* (fine, soft, and not pigmented) or *terminal* (long, coarse, and pigmented). The number of hair follicles does not change over an individual's lifetime, but the follicle size and type of hair can change in response to numerous factors, particularly androgens. Androgens are necessary for terminal hair and sebaceous gland development and mediate differentiation of pilosebaceous units (PSUs) into either a terminal hair follicle or a sebaceous gland. In the former case, androgens transform the vellus hair into a terminal hair; in the latter case, the sebaceous component proliferates and the hair remains vellus.

There are three phases in the cycle of hair growth: (1) *anagen* (growth phase), (2) *catagen* (involution phase), and (3) *telogen* (rest phase). Depending on the body site, hormonal regulation may play an important role in the hair growth cycle. For example, the eyebrows, eyelashes, and vellus hairs are androgen-insensitive, whereas the axillary and pubic areas are sensitive to low levels of androgens. Hair growth on the face, chest, upper abdomen, and back requires higher levels of androgens and is therefore more characteristic of the pattern typically seen in men. Androgen excess in women can lead to increased hair growth in most androgen-sensitive sites except in the scalp region, where hair loss occurs because androgens cause scalp hairs to spend less time in the anagen phase.

Although androgen excess underlies most cases of hirsutism, there is only a modest correlation between androgen levels and the quantity of hair growth. This is due to the fact that hair growth from the follicle also depends on local growth factors, and there is variability in end organ (PSU) sensitivity. Genetic factors and ethnic

background also influence hair growth. In general, dark-haired individuals tend to be more hirsute than blond or fair individuals. Asians and Native Americans have relatively sparse hair in regions sensitive to high androgen levels, whereas people of Mediterranean descent are more hirsute.

CLINICAL ASSESSMENT

Historic elements relevant to the assessment of hirsutism include the age at onset and rate of progression of hair growth and associated symptoms or signs (e.g., menstrual irregularity and acne). Depending on the cause, excess hair growth typically is first noted during the second and third decades of life. The growth is usually slow but progressive. Sudden development and rapid progression of hirsutism suggest the possibility of an androgen-secreting neoplasm, in which case virilization also may be present.

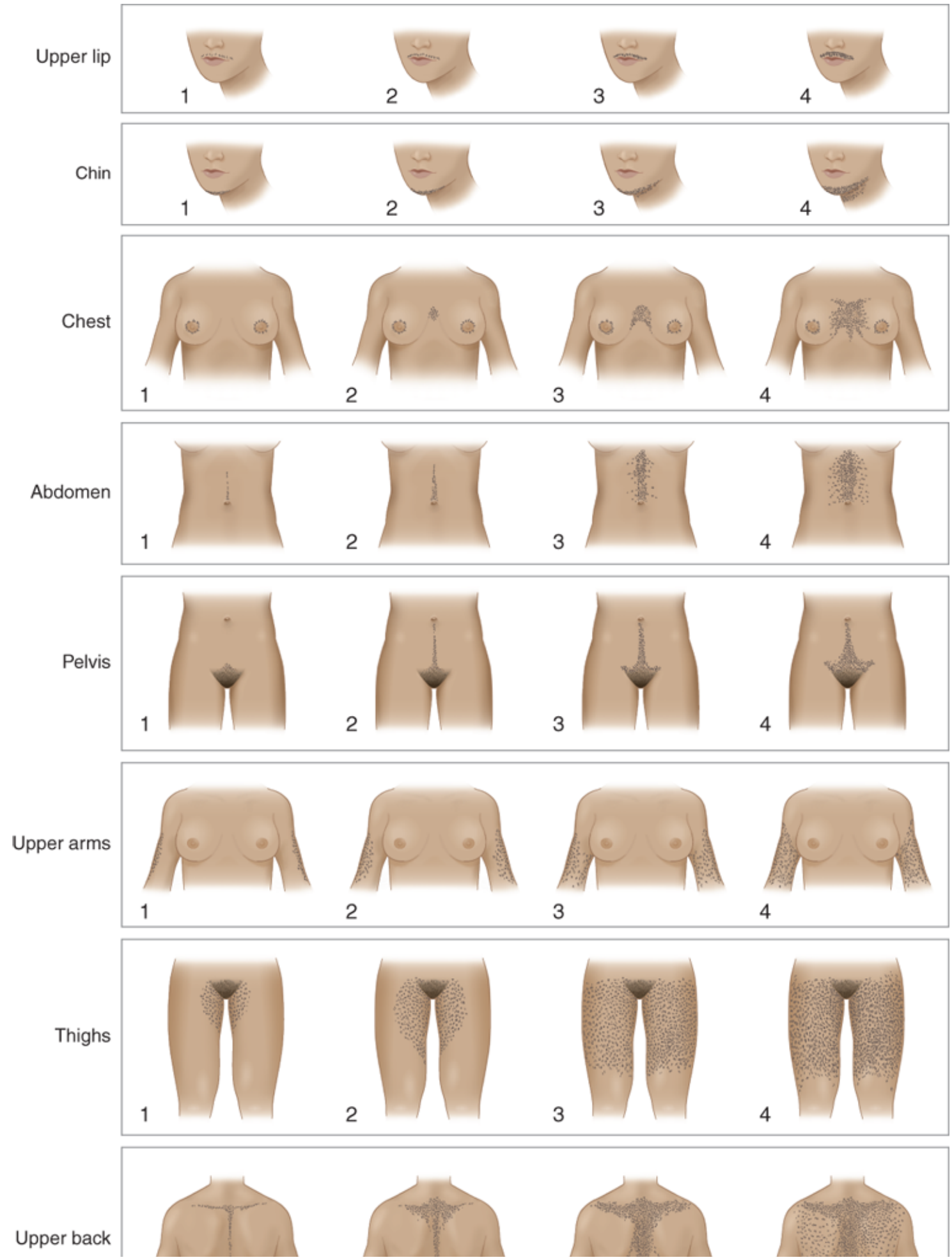
The age at onset of menstrual cycles (menarche) and the pattern of the menstrual cycle should be ascertained; irregular cycles from the time of menarche onward are more likely to result from ovarian rather than adrenal androgen excess. Associated symptoms such as galactorrhea should prompt evaluation for hyperprolactinemia **(Chap. 373)** and possibly hypothyroidism **(Chap. 375)**. Hypertension, striae, easy bruising, centripetal weight gain, and weakness suggest hypercortisolism (Cushing's syndrome; **Chap. 379**). Rarely, patients with growth hormone excess (i.e., acromegaly) present with hirsutism. Use of medications such as phenytoin, minoxidil, and cyclosporine may be associated with androgen-independent excess hair growth (i.e., hypertrichosis). A family history of infertility and/or hirsutism may indicate disorders such as nonclassic CAH **(Chap. 379)**.

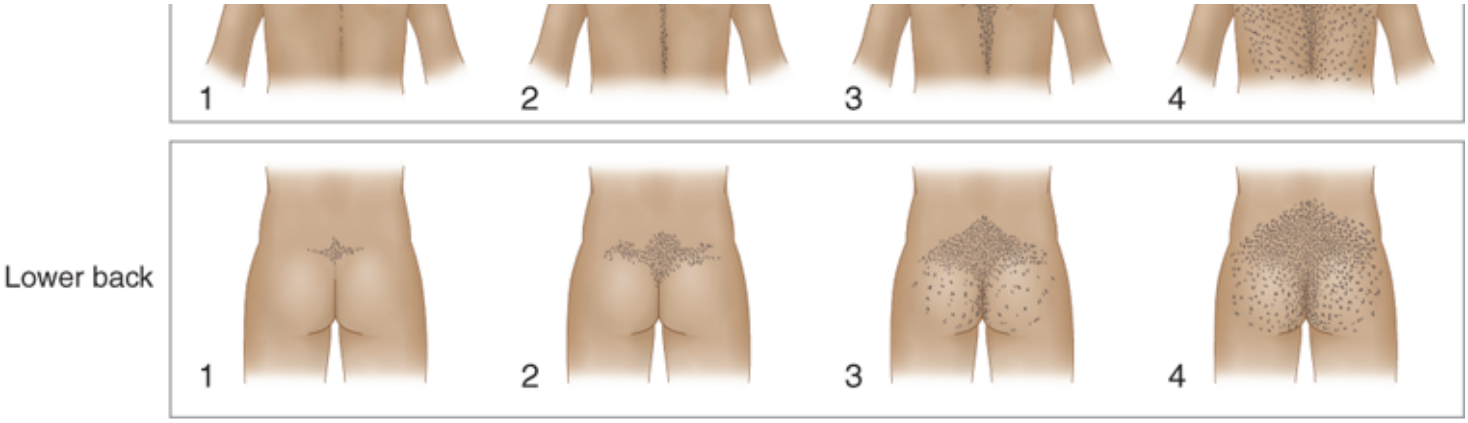
Physical examination should include measurement of height and weight and calculation of body mass index (BMI). A BMI >25 kg/m^2 is indicative of excess weight for height, and values >30 kg/m^2 are often seen in association with hirsutism, probably the result of increased conversion of androgen precursors to testosterone. Notation should be made of blood pressure, as adrenal causes may be associated with hypertension. Cutaneous signs sometimes associated with androgen excess and insulin resistance include acanthosis nigricans and skin tags.

An objective clinical assessment of hair distribution and quantity is central to the evaluation in any woman presenting with concerns about excessive hair growth. This assessment permits the distinction between hirsutism and hypertrichosis and provides a baseline reference point to gauge the response to treatment. A simple and commonly used method to grade hair growth is the modified scale of Ferriman and Gallwey **(Fig. 387-1)**, in which each of nine androgen-sensitive sites is graded from 0 to 4. Approximately 95% of white women have a score <8 on this scale; thus, it is normal for most women to have some hair growth in androgen-sensitive sites. Scores >8 suggest excess androgen-mediated hair growth, a finding that should be assessed further by means of hormonal evaluation (see below). In racial/ethnic groups that are less likely to manifest hirsutism (e.g., Asian women), additional cutaneous evidence of androgen excess should be sought, including pustular acne and thinning scalp hair.

FIGURE 387-1

Hirsutism scoring scale of Ferriman and Gallwey. The nine body areas that have androgen-sensitive areas are graded from 0 (no terminal hair) to 4 (frankly virile) to obtain a total score. A normal hirsutism score is <8. *(Modified from DA Ehrmann et al: Hyperandrogenism, hirsutism, and polycystic ovary syndrome, in LJ DeGroot and JL Jameson [eds], Endocrinology, 5th ed. Philadelphia, Saunders, 2006; with permission.)*


Upper lip
1
2
3
4
Chin
1
2
3
4
Chest
1
2
3
4
Abdomen
1
2
3
4
Pelvis
1
2
3
4
Upper arms
1
2
3
4
Thighs
1
2
3
4
Upper back



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HORMONAL EVALUATION

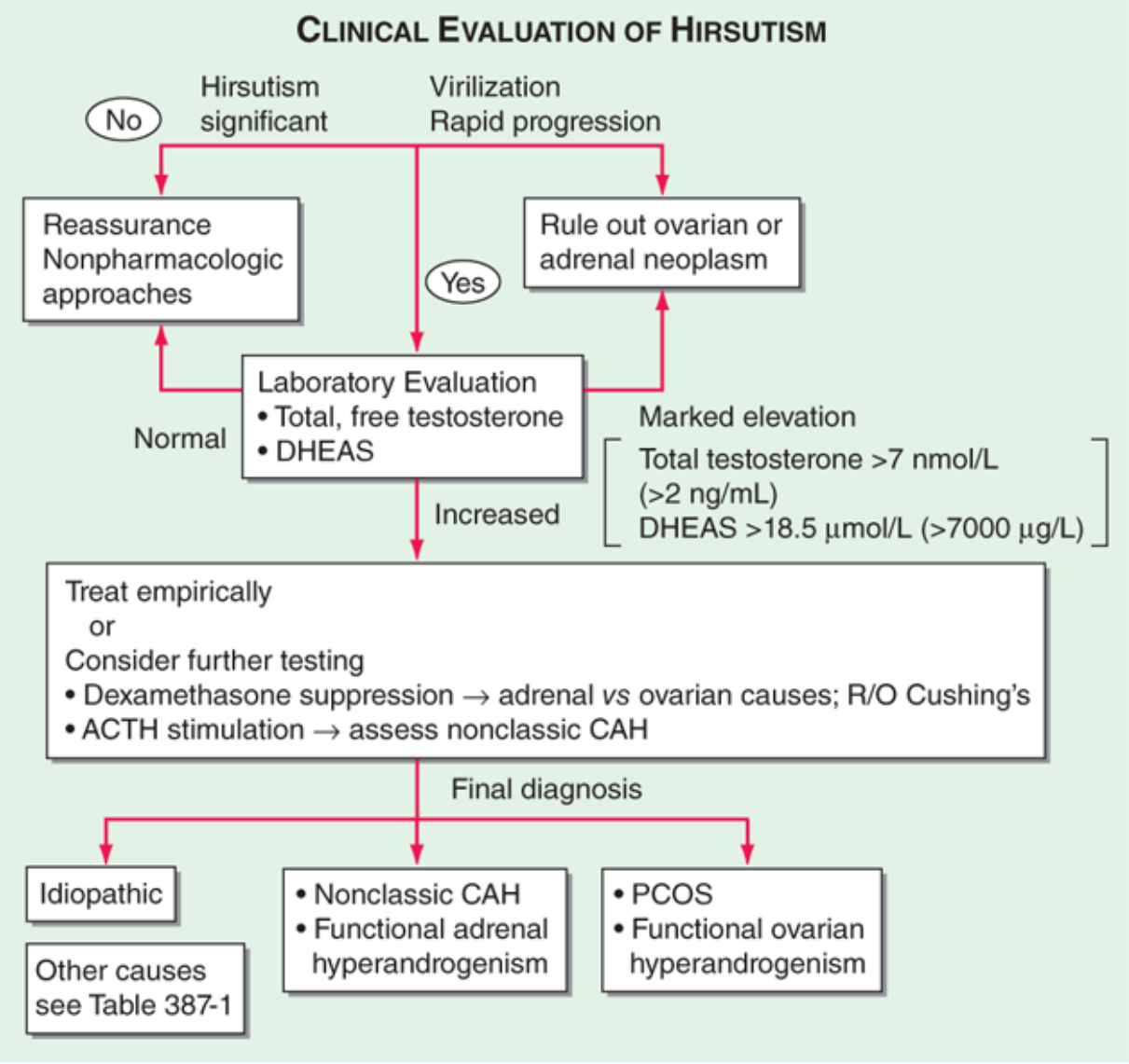
Androgens are secreted by the ovaries and adrenal glands in response to their respective tropic hormones: luteinizing hormone (LH) and adrenocorticotropic hormone (ACTH). Testosterone is the principal circulating steroid involved in the etiology of hirsutism; other steroids that may contribute to the development of hirsutism include androstenedione, dehydroepiandrosterone (DHEA) and its sulfated form (DHEAS). The ovaries and adrenal glands normally contribute about equally to testosterone production. Approximately half of the total testosterone originates from direct glandular secretion, and the remainder is derived from the peripheral conversion of androstenedione and DHEA (Chap. 384).

Although it is the most important circulating androgen, testosterone is in effect the penultimate androgen in mediating hirsutism; it is converted to the more potent dihydrotestosterone (DHT) by the enzyme 5α-reductase, which is located in the PSU. DHT has a higher affinity for, and slower dissociation from, the androgen receptor. The local production of DHT allows it to serve as the primary mediator of androgen action at the level of the pilosebaceous unit. There are two isoenzymes of 5α-reductase: type 2 is found in the prostate gland and in hair follicles, and type 1 is found primarily in sebaceous glands.

One approach to testing for hyperandrogenemia is depicted in Fig. 387-2. In addition to measuring blood levels of testosterone and DHEAS, it is important to measure the level of free (or unbound) testosterone. The fraction of testosterone that is not bound to its carrier protein, sex hormone–binding globulin (SHBG), is biologically available for conversion to DHT and binding to androgen receptors. Hyperinsulinemia and/or androgen excess decrease hepatic production of SHBG, resulting in levels of total testosterone within the high-normal range, whereas the unbound hormone is elevated more substantially. Although there is a decline in ovarian testosterone production after menopause, ovarian estrogen production decreases to an even greater extent, and the concentration of SHBG is reduced. Consequently, there is an increase in the relative proportion of unbound testosterone, and it may exacerbate hirsutism after menopause.

FIGURE 387-2

Algorithm for the evaluation and differential diagnosis of hirsutism. ACTH, adrenocorticotropic hormone; CAH, congenital adrenal hyperplasia; DHEAS, sulfated form of dehydroepiandrosterone; PCOS, polycystic ovarian syndrome.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A baseline plasma total testosterone level >12 nmol/L (>3.5 ng/mL) usually indicates a virilizing tumor, whereas a level >7 nmol/L (>2 ng/mL) is suggestive of tumor but may also be observed in women with hyperthecosis. A basal DHEAS level >18.5 µmol/L (>7000 µg/L) suggests an adrenal tumor. Although DHEAS has been proposed as a "marker" of predominant adrenal androgen excess, it is not unusual to find modest elevations in DHEAS among women with PCOS. Computed tomography (CT) or magnetic resonance imaging (MRI) should be used to localize an adrenal mass, and ultrasound usually suffices to identify an ovarian mass if clinical evaluation and hormonal levels suggest these possibilities.

PCOS is the most common cause of ovarian androgen excess (Chap. 385). An increased ratio of LH to follicle-stimulating hormone is characteristic in carefully studied patients with PCOS. However, because of the pulsatile nature of gonadotropin secretion, this finding may be absent in up to half of women with PCOS. Transvaginal ultrasound classically shows enlarged ovaries and increased stroma in women with PCOS. However, cystic ovaries also may be found in women without clinical or laboratory features of PCOS. Although usually limited to a research setting, a gonadotropin-releasing hormone agonist test can be used to make a specific diagnosis of

ovarian hyperandrogenism. A peak 17-hydroxyprogesterone level ≥7.8 nmol/L (≥2.6 µg/L) after the administration of 100 µg nafarelin (or 10 µg/kg leuprolide) subcutaneously is virtually diagnostic of ovarian hyperandrogenism.

Because adrenal androgens are readily suppressed by low doses of glucocorticoids, the dexamethasone androgen-suppression test may broadly distinguish ovarian from adrenal androgen overproduction. A blood sample is obtained before and after the administration of dexamethasone (0.5 mg orally every 6 h for 4 days). An adrenal source is suggested by suppression of unbound testosterone into the normal range; incomplete suppression suggests ovarian androgen excess. An overnight 1-mg dexamethasone suppression test, with measurement of 8:00 A.M. serum cortisol, is useful when there is clinical suspicion of Cushing's syndrome **(Chap. 379)**.

Nonclassic CAH is most commonly due to 21-hydroxylase deficiency but also can be caused by autosomal recessive defects in other steroidogenic enzymes necessary for adrenal corticosteroid synthesis **(Chap. 379)**. Because of the enzyme defect, the adrenal gland cannot secrete glucocorticoids (especially cortisol) efficiently. This results in diminished negative feedback inhibition of ACTH, leading to compensatory adrenal hyperplasia and the accumulation of steroid precursors that subsequently are converted to androgen. Deficiency of 21-hydroxylase can be reliably excluded by determining a morning 17-hydroxyprogesterone level <6 nmol/L (<2 µg/L) (drawn in the follicular phase). Alternatively, 21-hydroxylase deficiency can be diagnosed by measurement of 17-hydroxyprogesterone 1 h after the administration of 250 µg of synthetic ACTH (cosyntropin) intravenously.

TREATMENT

TREATMENT

Hirsutism

Treatment of hirsutism may be accomplished pharmacologically or by mechanical means of hair removal. Nonpharmacologic treatments should be considered in all patients either as the only treatment or as an adjunct to drug therapy.

Nonpharmacologic treatments include (1) bleaching, (2) depilatory (removal from the skin surface) such as shaving and chemical treatments, and (3) epilatory (removal of the hair including the root) such as plucking, waxing, electrolysis, laser and intense pulsed light (IPL). Despite perceptions to the contrary, shaving does not increase the rate or density of hair growth. Chemical depilatory treatments may be useful for mild hirsutism that affects only limited skin areas, though they can cause skin irritation. Wax treatment removes hair temporarily but is uncomfortable. Electrolysis is effective for more permanent hair removal, particularly in the hands of a skilled electrologist. Laser and IPL are used to treat large areas of pigmented, terminal hair. Light of specific wavelength, duration, and energy is absorbed by melanin in the hair shaft and follicle leading to photothermolysis. Properly delivered, this treatment delays hair regrowth and causes permanent hair removal in many patients.

Pharmacologic therapy is directed at interrupting one or more of the steps in the pathway of androgen synthesis and action: (1) suppression of adrenal and/or ovarian androgen production, (2) enhancement of androgen-binding to plasma-binding proteins, particularly SHBG, (3) impairment of the peripheral conversion of androgen precursors to active androgen, and (4) inhibition of androgen action at the target tissue level. Attenuation of hair

growth is typically not evident until 4–6 months after initiation of medical treatment and in most cases leads to only a modest reduction in hair growth.

Combination estrogen-progestin therapy in the form of an oral contraceptive is usually the first-line endocrine treatment for hirsutism and acne, after cosmetic and dermatologic management. The estrogenic component of most oral contraceptives currently in use is either ethinyl estradiol or mestranol. The suppression of LH leads to reduced production of ovarian androgens. The reduced androgen levels also result in a dose-related increase in SHBG, thus lowering the fraction of unbound plasma testosterone. Estrogens also have a direct, dose-dependent suppressive effect on sebaceous cell function.

The choice of a specific oral contraceptive should be predicated on the progestational component, as progestins vary in their suppressive effect on SHBG levels and in their androgenic potential. Ethynodiol diacetate has relatively low androgenic potential, whereas progestins such as norgestrel and levonorgestrel are particularly androgenic, as judged from their attenuation of the estrogen-induced increase in SHBG. Norgestimate exemplifies the newer generation of progestins that are virtually nonandrogenic. Drospirenone, an analogue of spironolactone that has both antimineralocorticoid and antiandrogenic activities, has been approved for use as a progestational agent in combination with ethinyl estradiol.

Oral contraceptives are contraindicated in women with a history of thromboembolic disease and women with increased risk of breast or other estrogen-dependent cancers **(Chap. 388)**. There is a relative contraindication to the use of oral contraceptives in smokers and those with hypertension or a history of migraine headaches. In most trials, estrogen-progestin therapy alone improves the extent of acne by a maximum of 50–70%. The effect on hair growth may not be evident for 6 months, and the maximum effect may require 9–12 months owing to the length of the hair growth cycle. Improvements in hirsutism are typically in the range of 20%, but there may be an arrest of further progression of hair growth.

Because oral contraceptives are efficacious and have fewer side effects, they are recommended over glucocorticoids as first-line treatment of hirsutism in CAH. If the response to oral contraceptives is inadequate, glucocorticoids may be used. The lowest effective dose of glucocorticoid should be used (e.g., dexamethasone [0.2–0.5 mg] or prednisone [5–10 mg]) taken at bedtime to achieve maximal suppression by inhibiting the nocturnal surge of ACTH.

Cyproterone acetate is the prototypic antiandrogen. It acts mainly by competitive inhibition of the binding of testosterone and DHT to the androgen receptor. In addition, it may enhance the metabolic clearance of testosterone by inducing hepatic enzymes. Although not available for use in the United States, cyproterone acetate is widely used in Canada, Mexico, and Europe. Cyproterone (50–100 mg) is given on days 1–15 and ethinyl estradiol (50 μg) is given on days 5–26 of the menstrual cycle. Side effects include irregular uterine bleeding, nausea, headache, fatigue, weight gain, and decreased libido.

Spironolactone, which usually is used as a mineralocorticoid antagonist, is also a weak antiandrogen. It is almost as effective as cyproterone acetate when used at high enough doses (100–200 mg daily). Patients should be monitored intermittently for hyperkalemia or hypotension, though these side effects are uncommon. Pregnancy should be avoided because of the risk of feminization of a male fetus. Spironolactone can also cause menstrual

irregularity. It often is used in combination with an oral contraceptive, which suppresses ovarian androgen production and helps prevent pregnancy.

Flutamide is a potent nonsteroidal antiandrogen that is effective in treating hirsutism, but concerns about the induction of hepatocellular dysfunction preclude its use. Finasteride is a competitive inhibitor of 5α-reductase type 2. Beneficial effects on hirsutism have been reported, but the predominance of 5α-reductase type 1 in the PSU appears to account for its limited efficacy. Finasteride would also be expected to impair sexual differentiation in a male fetus, and it should not be used in women who may become pregnant.

Eflornithine cream (Vaniqa) has been approved as a novel treatment for unwanted facial hair in women, but long-term efficacy remains to be established. It can cause skin irritation under exaggerated conditions of use. Ultimately, the choice of any specific agent(s) must be tailored to the unique needs of the patient being treated. As noted previously, pharmacologic treatments for hirsutism should be used in conjunction with nonpharmacologic approaches. It is also helpful to review the pattern of female hair distribution in the normal population to dispel unrealistic expectations.

FURTHER READING

Brown DL et al: Ovarian stromal hyperthecosis: Sonographic features and histologic associations. J Ultrasound Med 28:587, 2009. [PubMed: 19389897]

Haak CS et al: Hair removal in hirsute women with normal testosterone levels: A randomized controlled trial of long-pulsed diode laser vs. intense pulsed light. Br J Dermatol 163:1007, 2010.

Mccartney CR, Marshall JC: Polycystic Ovary Syndrome. N Engl J Med 375:1398, 2016. [PubMed: 27705264]

Rosenfield RL, Ehrmann DA: The Pathogenesis of Polycystic Ovary Syndrome (PCOS): The Hypothesis of PCOS as Functional Ovarian Hyperandrogenism Revisited. Endocr Rev 37:467, 2016. [PubMed: 27459230]

van Zuuren EJ et al: Interventions for hirsutism (excluding laser and photoepilation therapy alone). Cochrane Database Syst Rev CD010334, 2015.

Zimmerman Y et al: The effect of combined oral contraception on testosterone levels in healthy women: A systematic review and meta-analysis. Human Reprod Update 20:76, 2013.

Chapter 388: Menopause and Postmenopausal Hormone Therapy

JoAnn E. Manson; Shari S. Bassuk

INTRODUCTION

Menopause is the permanent cessation of menstruation due to loss of ovarian follicular function. It is diagnosed retrospectively after 12 months of amenorrhea. The average age at menopause is 51 years among U.S. women. *Perimenopause* refers to the time period preceding menopause, when fertility wanes and menstrual cycle irregularity increases, until the first year after cessation of menses. The onset of perimenopause precedes the final menses by 2–8 years, with a mean duration of 4 years. Smoking accelerates the menopausal transition by 2 years.

Although the peri- and postmenopausal transitions share many symptoms, the physiology and clinical management of the two differ. Low-dose oral contraceptives have become a therapeutic mainstay in perimenopause, whereas postmenopausal hormone therapy (HT) has been a common method of symptom alleviation after menstruation ceases.

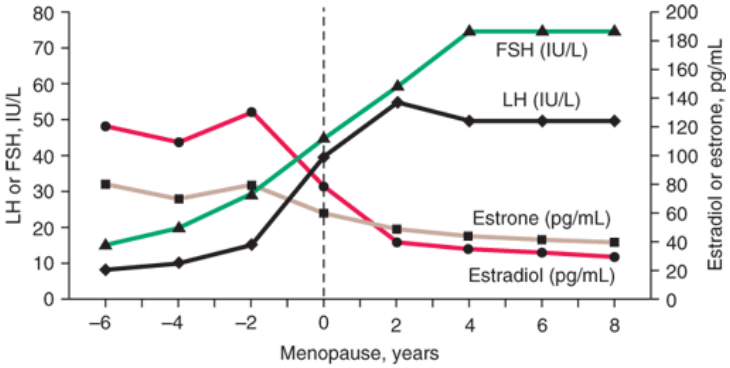
PERIMENOPAUSE

PHYSIOLOGY

Ovarian mass and fertility decline sharply after age 35 and even more precipitously during perimenopause; depletion of primary follicles, a process that begins before birth, occurs steadily until menopause **(Chap. 385)**. In perimenopause, intermenstrual intervals shorten significantly (typically by 3 days) as a result of an accelerated follicular phase. Follicle-stimulating hormone (FSH) levels rise because of altered folliculogenesis and reduced inhibin secretion. In contrast to the consistently high FSH and low estradiol levels seen in menopause, perimenopause is characterized by "irregularly irregular" hormone levels. The propensity for anovulatory cycles can produce a hyperestrogenic, hypoprogestagenic environment that may account for the increased incidence of endometrial hyperplasia or carcinoma, uterine polyps, and leiomyoma observed among women of perimenopausal age. Mean serum levels of selected ovarian and pituitary hormones during the menopausal transition are shown in **Fig. 388-1**. With transition into menopause, estradiol levels fall markedly, whereas estrone levels are relatively preserved, a pattern reflecting peripheral aromatization of adrenal and ovarian androgens. Levels of FSH increase more than those of luteinizing hormone, presumably because of the loss of inhibin as well as estrogen feedback.

FIGURE 388-1

Mean serum levels of ovarian and pituitary hormones during the menopausal transition. FSH, follicle-stimulating hormone; LH, luteinizing hormone. *(From JL Shifren, I Schiff: J Womens Health Gend Based Med 9 Suppl 1:S3, 2000. Reproduced with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSTIC TESTS

The Stages of Reproductive Aging Workshop +10 (STRAW+10) classification provides a comprehensive framework for the clinical assessment of ovarian aging. As shown in **Fig. 388-2**, menstrual cycle characteristics are the principal criteria for characterizing the menopausal transition, with biomarker measures as supportive criteria. Because of their extreme intraindividual variability, FSH and estradiol levels are imperfect diagnostic indicators of perimenopause in menstruating women. However, a consistently low FSH level in the early follicular phase (days 2–5) of the menstrual cycle does not

support a diagnosis of perimenopause, while levels >25 IU/L in a random blood sample are characteristic of the late menopause transition. FSH measurement can also aid in assessing fertility; levels of <20 IU/L, 20 to <30 IU/L, and ≥30 IU/L measured on day 3 of the cycle indicate a good, fair, and poor likelihood of achieving pregnancy, respectively. Antimüllerian hormone and inhibin B may also be useful for assessing reproductive aging.

FIGURE 388-2

The Stages of Reproductive Aging Workshop +10 (STRAW +10) staging system for reproductive aging in women. AMH, antimüllerian hormone; FSH, follicle-stimulating hormone. *(From SD Harlow et al: Menopause 14:387, 2012. Reproduced with permission.)*

Menarche — FMP (0)

Stage	−5	−4	−3b	−3a	−2	−1	+1a	+1b	+1c	+2
Terminology	Reproductive				Menopausal transition		Postmenopause			
	Early	Peak	Late		Early	Late	Early			Late
					Perimenopause					
Duration	Variable				Variable	1–3 years	2 years (1+1)		3–6 years	Remaining lifespan
Principal criteria										
Menstrual cycle	Variable to regular	Regular	Regular	Subtle changes in flow/ length	Variable Length Persistent ≥7-day difference in length of consecutive cycles	Interval of amenorrhea of ≥60 days				
Supportive criteria										
Endocrine FSH AMH Inhibin B			Low Low	Variable* Low Low	↑ Variable* Low Low	↑ >25 IU/L** Low Low	↑ Variable Low Low		Stabilizes Very low Very low	
Antral follicle count			Low	Low	Low	Low	Very low		Very low	
Descriptive characteristics										
Symptoms						Vasomotor symptoms Likely	Vasomotor symptoms Most likely			Increasing symptoms of urogenital atrophy

*Blood draw on cycle days 2–5 ↑ = elevated.
**Approximate expected level based on assays using current international pituitary standard.

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SYMPTOMS

Determining whether symptoms that develop in midlife are due to ovarian senescence or to other age-related changes is difficult. There is strong evidence that the menopausal transition can cause hot flashes, night sweats, irregular bleeding, and vaginal dryness, and there is moderate evidence that it can cause sleep disturbances in some women. There is inconclusive or insufficient evidence that ovarian aging is a major cause of mood swings, depression, impaired memory or concentration, somatic symptoms, urinary incontinence, or sexual dysfunction. In one U.S. study, nearly 60% of women reported hot flashes in the 2 years before their final menses. Symptom intensity, duration, frequency, and effects on quality of life are highly variable.

TREATMENT

TREATMENT

Perimenopause

PERIMENOPAUSAL THERAPY

For women with irregular or heavy menses or hormone-related symptoms that impair quality of life, low-dose combined oral contraceptives are a staple of therapy. Static doses of estrogen and progestin (e.g., 20 μg of ethinyl estradiol and 1 mg of norethindrone acetate daily for 21 days each month) can eliminate vasomotor symptoms and restore regular cyclicity. Oral contraceptives provide other benefits, including protection against ovarian and endometrial cancers and increased bone density, although it is not clear whether use during perimenopause decreases fracture risk later in life. Moreover, the contraceptive benefit is important, given that the unintentional pregnancy rate among women in their forties rivals that of adolescents.

Contraindications to oral contraceptive use include cigarette smoking, liver disease, a history of thromboembolism or cardiovascular disease, breast cancer, or unexplained vaginal bleeding. Progestin-only formulations (e.g., 0.35 mg of norethindrone daily) or medroxyprogesterone (Depo-Provera) injections (e.g., 150 mg IM every 3 months) may provide an alternative for the treatment of perimenopausal menorrhagia in women who smoke or have cardiovascular risk factors. Although progestins neither regularize cycles nor reduce the number of bleeding days, they reduce the volume of menstrual flow.

Nonhormonal strategies to reduce menstrual flow include the use of nonsteroidal anti-inflammatory agents such as mefenamic acid (an initial dose of 500 mg at the start of menses, then 250 mg qid for 2–3 days) or, when medical approaches fail, endometrial ablation. It should be noted that menorrhagia requires an evaluation to rule out uterine disorders. Transvaginal ultrasound with saline enhancement is useful for detecting leiomyomata or polyps, and endometrial aspiration can identify hyperplastic changes.

TRANSITION TO MENOPAUSE

For sexually active women using contraceptive hormones to alleviate perimenopausal symptoms, the question of when and if to switch to HT must be individualized. Doses of estrogen and progestogen (either synthetic progestins or natural forms of progesterone) in HT are lower than those in oral contraceptives and have not been documented to prevent pregnancy. Although a 1-year absence of spontaneous menses reliably indicates ovulation cessation, it is not possible to assess the natural menstrual pattern while a woman is taking an oral contraceptive. Women willing to switch to a barrier method of contraception should do so; if menses occur spontaneously, oral contraceptive use can be resumed. The average age of final menses among relatives can serve as a guide for when to initiate this process, which can be repeated yearly until menopause has occurred.

MENOPAUSE AND POSTMENOPAUSAL HT

One of the most complex health care decisions facing women is whether to use postmenopausal HT. Once prescribed primarily to relieve vasomotor symptoms, HT has been promoted as a strategy to forestall various disorders that accelerate after menopause, including osteoporosis and cardiovascular disease. In 2000, nearly 40% of postmenopausal women aged 50–74 in the United States had used HT. This widespread use occurred despite the paucity of conclusive data, until recently, on the health consequences of such therapy. Although many women rely on their health care providers for a definitive answer to the question of whether to use postmenopausal hormones, balancing the benefits and risks for an individual patient is challenging.

Although observational studies suggest that HT prevents cardiovascular and other chronic diseases, the apparent benefits may result at least in part from differences between women who opt to take postmenopausal hormones and women who do not. Those choosing HT tend to be healthier, have greater access to medical care, are more compliant with prescribed treatments, and maintain a more health-promoting lifestyle. Randomized trials, which eliminate these confounding factors, have not consistently confirmed the benefits found in observational studies. Indeed, the largest HT trial to date, the Women's Health Initiative (WHI), which examined more than 27,000 postmenopausal women aged 50–79 (mean age, 63) for an average of 5–7 years, was stopped early because of an overall unfavorable benefit-risk ratio in the estrogen-progestin arm and an excess risk of stroke that was not offset by a reduced risk of coronary heart disease (CHD) in the estrogen-only arm.

The following summary offers a decision-making guide based on a synthesis of currently available evidence. Prevention of cardiovascular disease is eliminated from the equation due to lack of evidence for such benefits in randomized clinical trials.

BENEFITS AND RISKS OF POSTMENOPAUSAL HT

See Table 388-1.

TABLE 388-1

Benefits and Risks of Postmenopausal Hormone Therapy in the Overall Study Population of Women aged 50–79 Years in the Intervention Phase of the Women's Health Initiative (WHI) Estrogen-Progestin and Estrogen-Alone Trials[a]

		Estrogen-Progestin		Estrogen Alone	
Outcome	**Effect**	**Relative Benefit or Risk**	**Absolute Benefit or Risk[b]**	**Relative Benefit or Risk**	**Absolute Benefit or Risk[b]**
Definite Benefits					
Symptoms of menopause	Definite improvement	↓65–90% decreased risk[c]		↓65–90% decreased risk[c]	
Osteoporosis	Definite increase in bone mineral density and decrease in fracture risk	↓33% decreased risk for hip fracture	6 fewer cases (11 vs. 17) of hip fracture	↓33% decreased risk for hip fracture	6 fewer cases (13 vs. 19) of hip fracture
Definite Risks[h]					
Endometrial cancer	Definite increase in risk with estrogen alone (see below for estrogen-progestin)	See below	See below		4.6 excess cases (observational studies)
Pulmonary embolism	Definite increase in risk	↑98% increased risk	9 excess cases (18 vs. 9)	↑35% increased risk (n.s.)	4 excess cases (14 vs. 10)
Deep-vein thrombosis	Definite increase in risk	↑87% increased risk	11.5 excess cases (25 vs. 14)	↑48% increased risk	7.5 excess cases (23 vs. 15)
Breast cancer	Definite increase in risk with long-term use (≥5 years) of estrogen-progestin	↑24% increased risk	8.5 excess cases (43 vs. 35)	↓21% decreased risk (n.s.)	7 fewer cases (28 vs. 35)
Gallbladder disease	Definite increase in risk	↑57% increased risk	47 excess cases (131 vs. 84)	↑55% increased risk	58 excess cases (164 vs. 106)
Probable or Uncertain Risks and Benefits[h]					
Coronary heart disease[d]	Probable increase in risk among older women and women many years past menopause; possible decrease in risk or no effect in younger or recently menopausal women[e]	↑18% increased risk (n.s.)	6 excess cases (41 vs. 35)	No increase in risk	No difference in risk
Myocardial infarction	Significant interaction by age group for estrogen alone, with reduced risk in younger—but not older—women (*p* for trend by age = 0.02)	↑24% increased risk (n.s.)	6 excess cases (35 vs. 29)	No increase in risk[e]	No difference in risk[e]
Stroke	Probable increase in risk	↑37% increased risk	9 excess cases (33 vs. 24)	↑35% increased risk	11 excess cases (45 vs. 34)

		Estrogen-Progestin		Estrogen Alone	
Outcome	**Effect**	**Relative Benefit or Risk**	**Absolute Benefit or Risk[b]**	**Relative Benefit or Risk**	**Absolute Benefit or Risk[b]**
Ovarian cancer	Probable increase in risk with long-term use (≥5 years)	↑41% increased risk (n.s.)	1 excess case (5 vs. 4)	Not available	Not available
Endometrial cancer	Probable decrease in risk with estrogen-progestin during long-term follow-up (see above for estrogen alone)	↓33% decreased risk[f]	3 fewer cases (7 vs. 10)	See above	See above
Urinary incontinence	Probable increase in risk	↑49% increased risk	549 excess cases (1661 vs. 1112)	↑61% increased risk	852 excess cases (2255 vs. 1403)
Colorectal cancer	Probable decrease in risk with estrogen-progestin; possible increase in risk in older women with estrogen alone (*p* for trend by age = 0.02 for estrogen alone)	↓38% decreased risk	6.5 fewer cases (10 vs. 17)	No increase or decrease in risk[e]	No difference in risk[e]
Type 2 diabetes	Probable decrease in risk	↓19% decreased risk	16 fewer cases (72 vs. 88)	↓14% decreased risk	21 fewer cases (134 vs. 155)
Dementia (age ≥65)	Increase in risk in older women (but inconsistent data from observational studies and randomized trials)	↑101% increased risk	23 excess cases (46 vs. 23)	↑47% increased risk (n.s.)	15 excess cases (44 vs. 29)
Total mortality	Possible increase in risk among older women and women many years past menopause; possible decrease in risk or no effect in younger or recently menopausal women (*p* for trend by age <0.05 for both trials combined)	No increase in risk	No difference in risk	No increase in risk[e]	No difference in risk[e]
Global index[g]	Probable increase in risk or no effect among older women and women many years past menopause; possible decrease in risk or no effect in younger or recently menopausal women (*p* for trend by age = 0.02 for estrogen alone)	↑12% increased risk	20.5 excess cases (189 vs. 168)	No increase in risk[e]	No difference in risk[e]

[a]The estrogen-progestin arm of the WHI assessed 5.6 years of conjugated equine estrogens (0.625 mg/d) plus medroxyprogesterone acetate (2.5 mg/d) versus placebo. The estrogen-alone arm of the WHI assessed 7.1 years of conjugated equine estrogens (0.625 mg/d) versus placebo. [b]Number of cases per 10, 000 women per year. [c]The WHI was not designed to assess the effect of HT on menopausal symptoms. Data from other randomized trials suggest that HT reduces risk for menopausal symptoms by 65–90%. [d]*Coronary heart disease* is defined as nonfatal myocardial infarction or coronary death. [e]There was a significant interaction by age; that is, the association between HT and the specified outcome was different in younger women and older women. [f]This is the risk reduction that was observed during a cumulative 13-year follow-up period (5.6 years of treatment plus 8.2 years of postintervention observation). [g]The *global index* is a composite outcome representing the first event for each participant from among the following: coronary heart disease, stroke, pulmonary embolism, breast cancer, colorectal cancer, endometrial cancer (estrogen-progestin arm only), hip fracture, and death. Because participants can experience more than one type of event, the global index cannot be derived by a simple summing of the component events. [h]Includes some outcomes where results were divergent between the estrogen-progestin arm and the estrogen-alone arm.

Abbreviation: n.s., not statistically significant.

Source: Data from JE Manson et al: JAMA 310:1353, 2013.

Definite Benefits

SYMPTOMS OF MENOPAUSE

Compelling evidence, including data from randomized clinical trials, indicates that estrogen therapy is highly effective for controlling vasomotor and genitourinary symptoms. Alternative approaches, including the use of antidepressants (such as paroxetine, 7.5 mg/d; or venlafaxine, 75–150 mg/d), gamma-aminobutyric acid analogues (such as gabapentin, 900–2400 mg/d [dose divided 3 times per day]; or pregabalin, 150–300 mg/d [dose divided twice per day]), or clonidine (0.1 mg/d), may also alleviate vasomotor symptoms, although they are less effective than HT. Paroxetine is the only nonhormonal drug approved by the U.S. Food and Drug Administration for treatment of vasomotor symptoms. Bazedoxifene, an estrogen agonist/antagonist, in combination with conjugated estrogens has also received approval for this use. Cognitive behavioral therapy and clinical hypnosis have been shown in randomized trials to help with vasomotor symptom management. Weight loss, mindfulness-based stress reduction, stellate ganglion block, and the consumption of S-equol soy derivatives are also promising strategies, although more trials are needed. For genitourinary syndrome of menopause, the efficacy of low-dose vaginal estrogen is similar to that of oral or transdermal estrogen; oral ospemifene or vaginal prasterone are additional options.

OSTEOPOROSIS

(See also Chap. 404)

Bone density

By reducing bone turnover and resorption rates, estrogen slows the aging-related bone loss experienced by most postmenopausal women. More than 50 randomized trials have demonstrated that postmenopausal estrogen therapy, with or without a progestogen, rapidly increases bone mineral density at the spine by 4–6% and at the hip by 2–3% and that those increases are maintained during treatment.

Fractures

Data from observational studies indicate a 50–80% lower risk of vertebral fracture and a 25–30% lower risk of hip, wrist, and other peripheral fractures among current estrogen users; addition of a progestogen does not appear to modify this benefit. In the WHI, 5–7 years of either combined estrogen-progestin or estrogen-only therapy was associated with a 33% reduction in hip fractures and 25–30% fewer total fractures among a population unselected for osteoporosis. Bisphosphonates (such as alendronate, 10 mg/d or 70 mg once per week; risedronate, 5 mg/d or 35 mg once per week; ibandronate, 2.5 mg/d or 150 mg once per month or 3 mg every 3 months IV; or zoledronic acid 5 mg once per year IV) and denosumab (60 mg twice per year SC) increase bone mass density by reducing bone resorption and have been shown in randomized trials to decrease fracture rates. Other treatment options include bazedoxifene in combination with conjugated estrogens; the selective estrogen receptor modulator (SERM) raloxifene (60 mg/d); and parathyroid hormone (teriparatide, 20 µg/d SC). Unlike estrogen, these alternative therapies do not appear to have adverse effects on the endometrium or breast. Increased weight-bearing and resistance exercise; adequate calcium intake (1000–1200 mg/d through diet or supplements in two or three divided doses); and adequate vitamin D intake (600–1000 IU/d) may also reduce the risk of osteoporosis-related fractures. According to a 2011 report by the Institute of Medicine (now the National Academy of Medicine), 25-hydroxyvitamin D blood levels of ≥50 nmol/L are sufficient for bone-density maintenance and fracture prevention. The Fracture Risk Assessment (FRAX®) score, an algorithm that combines an individual's bone-density score with age and other risk factors to predict her 10-year risk of hip and major osteoporotic fracture, may be of use in guiding decisions about pharmacologic treatment (see *www.shef.ac.uk/FRAX/*).

Definite Risks

ENDOMETRIAL CANCER (WITH ESTROGEN ALONE)

A combined analysis of 30 observational studies found a tripling of endometrial cancer risk among short-term users (1–5 years) of unopposed estrogen and a nearly tenfold increased risk among long-term users (≥10 years). These findings are supported by results from the randomized Postmenopausal Estrogen/Progestin Interventions (PEPI) trial, in which 24% of women assigned to unopposed estrogen for 3 years developed atypical endometrial hyperplasia—a premalignant lesion—as opposed to only 1% of women assigned to placebo. Use of a progestogen, which opposes the effects of estrogen on the endometrium, eliminates these risks and may even reduce risk (see later).

VENOUS THROMBOEMBOLISM

A meta-analysis of observational studies found that current oral estrogen use was associated with a 2.5-fold increase in risk of venous thromboembolism in postmenopausal women. A meta-analysis of randomized trials, including the WHI, found a 2.1-fold increase in risk. Results from the WHI indicate a nearly twofold increase in risk of pulmonary embolism and deep-vein thrombosis with estrogen-progestin and a 35–50% increase in these risks with estrogen-only therapy. Transdermal estrogen, taken alone or with certain progestogens (micronized progesterone or pregnane derivatives), appears to be a safer alternative with respect to thrombotic risk.

BREAST CANCER (WITH ESTROGEN-PROGESTIN)

An increased risk of breast cancer has been found among current or recent estrogen users in observational studies; this risk is directly related to duration of use. In a meta-analysis of 51 case-control and cohort studies, short-term use (<5 years) of postmenopausal HT did not appreciably elevate breast cancer incidence, whereas long-term use (≥5 years) was associated with a 35% increase in risk. In contrast to findings for endometrial cancer, combined estrogen-progestin regimens appear to increase breast cancer risk more than estrogen alone. Data from randomized trials also indicate that estrogen-progestin raises breast cancer risk. In the WHI, women assigned to receive combination hormones for an average of 5.6 years were 24% more likely to develop breast cancer than women assigned to placebo, but 7.1 years of estrogen-only therapy did not increase risk. Indeed, the WHI showed a trend toward a reduction in breast cancer risk with estrogen alone, although it is unclear whether this finding would pertain to formulations of estrogen other than conjugated equine estrogens or to treatment durations of >7 years. In the Heart and Estrogen/Progestin Replacement Study (HERS), 4 years of combination therapy was associated with a 27% increase in breast cancer risk. Although the latter finding was not statistically significant, the totality of evidence strongly implicates estrogen-progestin therapy in breast carcinogenesis.

Some observational data suggest that the length of the interval between menopause onset and HT initiation may influence the association between such therapy and breast cancer risk, with a "gap time" of <3–5 years conferring a higher HT-associated breast cancer risk. (This pattern of findings contrasts with that for CHD, as discussed later in this chapter.) However, this association remains inconclusive and may be a spurious finding attributable to higher rates of screening mammography and thus earlier cancer detection in HT users than in nonusers, especially in early menopause. Indeed, in the WHI trial, hazard ratios for HT and breast cancer risk did not differ among women 50–59, those 60–69, and those 70–79 years of age at trial entry. (There was insufficient power to examine finer age categories.) Additional research is needed to clarify the issue.

GALLBLADDER DISEASE

Large observational studies report a two- to threefold increased risk of gallstones or cholecystectomy among postmenopausal women taking oral estrogen. In the WHI, women randomized to estrogen-progestin or estrogen alone were ~55% more likely to develop gallbladder disease than those assigned to placebo. Risks were also increased in HERS. Transdermal HT might be a safer alternative, but further research is needed.

Probable or Uncertain Risks and Benefits

CORONARY HEART DISEASE/STROKE

Until recently, HT had been enthusiastically recommended as a possible cardioprotective agent. In the past three decades, multiple observational studies suggested, in the aggregate, that estrogen use leads to a 35–50% reduction in CHD incidence among postmenopausal women. The biologic plausibility of such an association is supported by data from randomized trials demonstrating that exogenous estrogen lowers plasma low-density lipoprotein (LDL) cholesterol levels and raises high-density lipoprotein (HDL) cholesterol levels by 10–15%. Administration of estrogen also favorably affects lipoprotein(a) levels, LDL oxidation, endothelial vascular function, fibrinogen, and plasminogen activator inhibitor 1. However, estrogen therapy has unfavorable effects on other biomarkers of cardiovascular risk: it boosts triglyceride levels; promotes coagulation via factor VII, prothrombin fragments 1 and 2, and fibrinopeptide A elevations; and raises levels of the inflammatory marker C-reactive protein.

Randomized trials of estrogen or combined estrogen-progestin in women with preexisting cardiovascular disease have not confirmed the benefits reported in observational studies. In HERS (a secondary-prevention trial designed to test the efficacy and safety of estrogen-progestin therapy with regard to clinical cardiovascular outcomes), the 4-year incidence of coronary death and nonfatal myocardial infarction was similar in the active-treatment and placebo groups, and a 50% increase in risk of coronary events was noted during the first year among participants assigned to the active-treatment group. Although it is possible that progestin may mitigate estrogen's benefits, the Estrogen Replacement and Atherosclerosis (ERA) trial indicated that angiographically determined progression of coronary atherosclerosis was unaffected by either opposed or unopposed estrogen treatment. Moreover, no cardiovascular benefit was found in the Papworth Hormone Replacement Therapy Atherosclerosis Study, a trial of transdermal estradiol with and without norethindrone; the Women's Estrogen for Stroke Trial (WEST), a trial of oral 17β-estradiol; or the Estrogen in the Prevention of Reinfarction Trial (ESPRIT), a trial of oral estradiol valerate. Thus, in clinical trials, HT has not proved effective for the secondary prevention of cardiovascular disease in postmenopausal women.

Primary-prevention trials also suggest an early increase in cardiovascular risk and an absence of cardioprotection with postmenopausal HT. In the WHI, women assigned to 5.6 years of estrogen-progestin therapy were 18% more likely to develop CHD (defined in primary analyses as nonfatal myocardial infarction or coronary death) than those assigned to placebo, although this risk elevation was not statistically significant. However, during the trial's first year, there was a significant 80% increase in risk, which diminished in subsequent years (p for trend by time = 0.03). In the estrogen-only arm of the WHI, no overall effect on CHD was observed during the 7.1 years of the trial or in any specific year of follow-up. This pattern of results was similar to that for the outcome of total myocardial infarction.

However, a closer look at available data suggests that timing of initiation of HT may critically influence the association between such therapy and CHD. Estrogen may slow early stages of atherosclerosis but have adverse effects on advanced atherosclerotic lesions. It has been hypothesized that the prothrombotic and proinflammatory effects of estrogen manifest themselves predominantly among women with subclinical lesions who initiate HT well after the menopausal transition, whereas women with less arterial damage who start HT early in menopause may derive cardiovascular benefit because they have not yet developed advanced lesions. Data from experiments in nonhuman primates and from some recent randomized trials in humans support this concept. Conjugated estrogens had no effect on the extent of coronary artery plaque in cynomolgus monkeys assigned to receive estrogen alone or combined with progestin starting 2 years (~6 years in human terms) after oophorectomy and well after the establishment of atherosclerosis. However, administration of exogenous hormones immediately after oophorectomy, during the early stages of atherosclerosis, reduced the extent of plaque by 70%. In the Early versus Late Intervention Trial with Estradiol (ELITE), a 6-year trial among 643 healthy postmenopausal women that was designed to test whether effects of estrogen on the development and progression of atherosclerosis depend on age at initiation of therapy, oral 17β-estradiol administered with or without vaginal micronized progesterone significantly slowed carotid atherosclerotic progression in women within 6 years of menopause onset (mean age, 55.4 years) but not in women more than 10 years past menopause onset (mean age, 65.4 years) (p, interaction=0.007). On the other hand, in the Kronos Early Estrogen Prevention Study (KEEPS), a 4-year trial among 729 healthy postmenopausal women within 3 years of menopause onset at trial entry (mean age, 53 years), neither oral conjugated estrogens nor transdermal estradiol, administered with oral micronized progesterone, affected carotid atherosclerotic progression. However, the low prevalence of this endpoint in the overall study population may have curtailed power to detect a treatment difference.

Lending further credence to the timing hypothesis are results of subgroup analyses of data from observational studies and large clinical trials. For example, among women who entered the WHI trial with a relatively favorable cholesterol profile, estrogen with or without progestin led to a 40% lower risk of incident CHD. Among women who entered with a worse cholesterol profile, therapy resulted in a 73% higher risk (p for interaction = 0.02). The presence or

absence of the metabolic syndrome **(Chap. 401)** also strongly influenced the relation between HT and incident CHD. Among women with the metabolic syndrome, HT more than doubled CHD risk, whereas no association was observed among women without the syndrome. Moreover, although there was no association between estrogen-only therapy and CHD in the WHI trial cohort as a whole, such therapy was associated with a CHD risk reduction of 40% among participants aged 50–59; in contrast, a risk reduction of only 5% was observed among those aged 60–69, and a risk increase of 9% was found among those aged 70–79 (p for trend by age = 0.08). For the outcome of total myocardial infarction, estrogen alone was associated with a borderline-significant 45% reduction and a nonsignificant 24% increase in risk among the youngest and oldest women, respectively (p for trend by age = 0.02). Estrogen was also associated with lower levels of coronary artery calcified plaque in the younger age group. Although age did not have a similar effect in the estrogen-progestin arm of the WHI, CHD risks increased with years since menopause (p for trend = 0.08), with a significantly elevated risk among women who were ≥20 years past menopause. For the outcome of total myocardial infarction, estrogen-progestin was associated with a 9% risk reduction among women <10 years past menopause as opposed to a 16% increase in risk among women 10–19 years past menopause and a twofold increase in risk among women >20 years past menopause (p for trend = 0.01). In the large observational Nurses' Health Study, women who chose to start HT within 4 years of menopause experienced a lower risk of CHD than did nonusers, whereas those who began therapy ≥10 years after menopause appeared to receive little coronary benefit. Observational studies include a high proportion of women who begin HT within 3–4 years of menopause, whereas clinical trials include a high proportion of women ≥12 years past menopause; this difference helps to reconcile some of the apparent discrepancies between the two types of studies.

For the outcome of stroke, WHI participants assigned to estrogen-progestin or estrogen alone were ~35% more likely to suffer a stroke than those assigned to placebo. Whether or not age at initiation of HT influences stroke risk is not well understood. In the WHI and the Nurses' Health Study, HT was associated with an excess risk of stroke in all age groups. Further research is needed on age, time since menopause, and other individual characteristics (including biomarkers) that predict increases or decreases in cardiovascular risk associated with exogenous HT. Furthermore, it remains uncertain whether different doses, formulations, or routes of administration of HT will produce different cardiovascular effects.

COLORECTAL CANCER

Observational studies have suggested that HT reduces risks of colon and rectal cancer, although the estimated magnitudes of the relative benefits have ranged from 8 to 34% in various meta-analyses. In the WHI (the sole trial to examine the issue), estrogen-progestin was associated with a significant 38% reduction in colorectal cancer over a 5.6-year period, although no benefit was seen with 7 years of estrogen-only therapy. However, a modifying effect of age was observed, with a doubling of risk with HT in women aged 70–79 but no risk elevation in younger women (p for trend by age = 0.02).

COGNITIVE DECLINE AND DEMENTIA

A meta-analysis of 10 case-control and two cohort studies suggested that postmenopausal HT is associated with a 34% decreased risk of dementia. Subsequent randomized trials (including the WHI), however, have failed to demonstrate any benefit of estrogen or estrogen-progestin therapy on the progression of mild to moderate Alzheimer's disease and/or have indicated a potential adverse effect of HT on the incidence of dementia, at least in women ≥65 years of age. Among women randomized to HT (as opposed to placebo) at age 50–55 in the WHI, no effect on cognition was observed during the postintervention phase. Determining whether timing of initiation of HT influences cognitive outcomes will require further study.

OVARIAN CANCER AND OTHER DISORDERS

On the basis of limited observational and randomized data, it has been hypothesized that HT increases the risk of ovarian cancer and reduces the risk of type 2 diabetes mellitus. Results from the WHI support these hypotheses. The WHI also found that HT use was associated with an increased risk of urinary incontinence and that estrogen-progestin was associated with increased rates of lung cancer mortality.

ENDOMETRIAL CANCER (WITH ESTROGEN-PROGESTIN)

In the WHI, use of estrogen-progestin was associated with a nonsignificant 17% reduction in risk of endometrial cancer. A significant reduction in risk emerged during the postintervention period (see later).

ALL-CAUSE MORTALITY

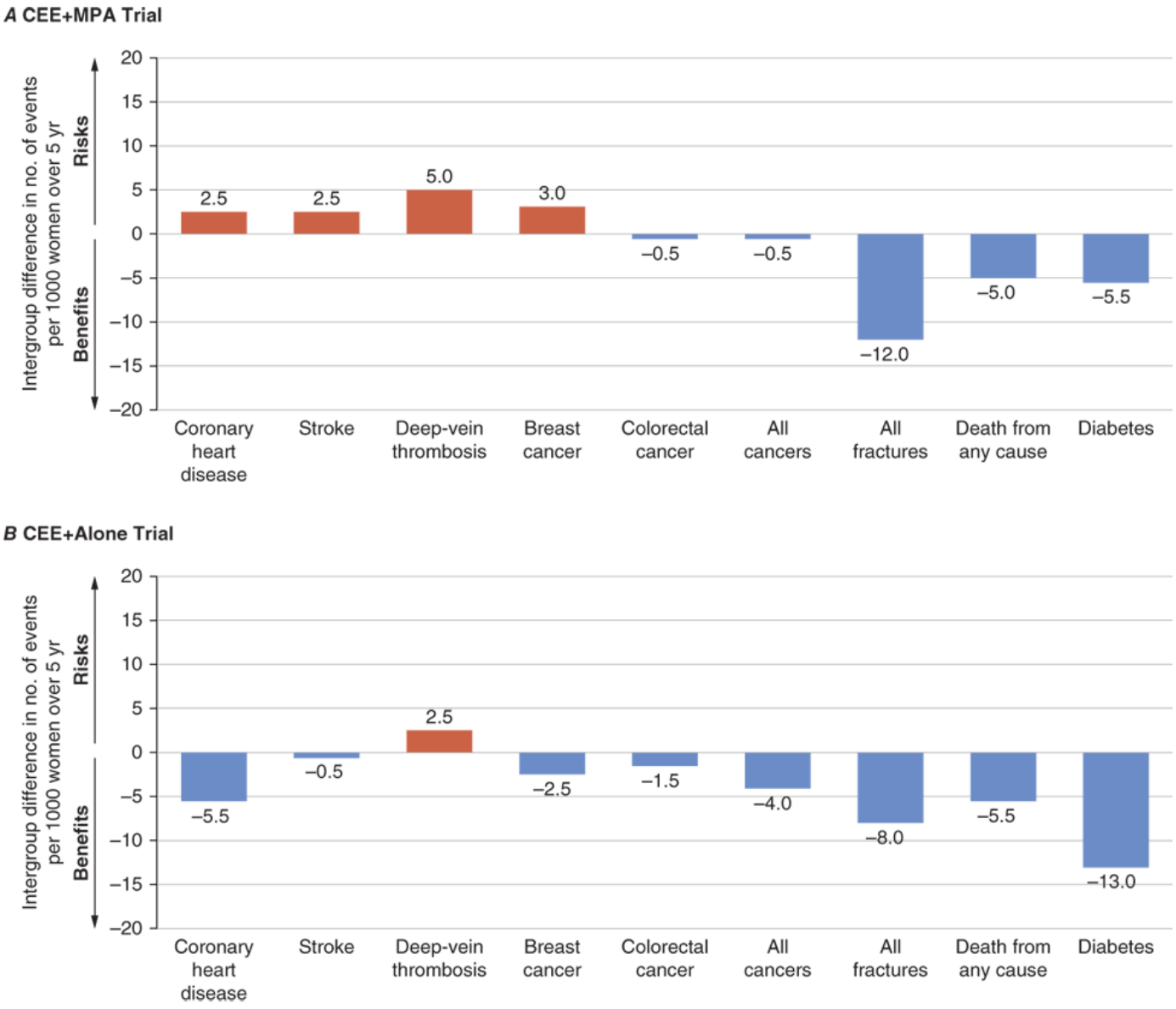
In the overall WHI cohort, estrogen with or without progestin was not associated with all-cause mortality. However, there was a trend toward reduced mortality in younger women, particularly with estrogen alone. For women aged 50–59, 60–69, and 70–79 years, relative risks (RRs) associated with estrogen-only therapy were 0.70, 1.01, and 1.21, respectively (p for trend = 0.04).

OVERALL BENEFIT-RISK PROFILE

Estrogen-progestin was associated with an unfavorable benefit-risk profile (excluding relief from menopausal symptoms) as measured by a "global index"—a composite outcome including CHD, stroke, pulmonary embolism, breast cancer, colorectal cancer, endometrial cancer, hip fracture, and death (Table 388-1)—in the WHI cohort as a whole, and this association did not vary by 10-year age group. Estrogen-only therapy was associated with a neutral benefit-risk profile in the WHI cohort as a whole. However, there was a significant trend toward a more favorable benefit-risk profile among younger women and a less favorable profile among older women, with RRs of 0.84, 0.99, and 1.17 for women aged 50–59, 60–69, and 70–79 years, respectively (p for trend by age = 0.02). The balance of benefits and risks of estrogen with and without progestin among women aged 50-59 is shown in **Fig. 388-3.**

FIGURE 388-3

Benefits and risks of the two hormone therapy formulations evaluated in the Women's Health Initiative, in women aged 50–59 years. Results are shown for the two formulations, conjugated equine estrogens (CEE) alone or in combination with medroxyprogesterone acetate (MPA). Risks and benefits are expressed as the difference in number of events (number in the HT group minus the number in the placebo group) per 1000 women over 5 years. *(Data are from JE Manson et al: JAMA 310:1353, 2013.) (Graphic display is from JE Manson, AM Kaunitz: N Engl J Med 374:803, 2016 and is reproduced with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CHANGES IN HEALTH STATUS AFTER DISCONTINUATION OF HT

In the WHI, many but not all risks and benefits associated with active use of HT dissipated within 5–7 years after discontinuation of therapy. For estrogen-progestin, an elevated risk of breast cancer persisted (RR = 1.28 [95% confidence interval, 1.11–1.48]) during a median cumulative 13-year follow-up period (5.6 years of treatment plus 8.2 years of postintervention observation), but most cardiovascular disease risks became neutral. A reduction in hip fracture risk persisted (RR = 0.81 [0.68–0.97]), and a significant reduction in endometrial cancer risk emerged (RR = 0.67 [0.49–0.91]). For estrogen alone, the reduction in breast cancer risk became statistically significant (RR = 0.79 [0.65–0.97]) during a median cumulative 13-year follow-up period (6.8 years of treatment plus 6.6 years of postintervention observation), and significant differences by age group persisted for total myocardial infarction and the global index, with more favorable results for younger women.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Postmenopausal HT

The rational use of postmenopausal HT requires balancing the potential benefits and risks. Figure 388-4 provides one approach to decision-making. The clinician should first determine whether the patient has moderate to severe menopausal symptoms—the primary indication for initiation of systemic HT. Systemic HT may also be used to prevent osteoporosis in women at high risk of fracture who cannot tolerate alternative osteoporosis therapies. (Vaginal estrogen or other medications may be used to treat genitourinary syndrome of menopause in the absence of vasomotor symptoms.) The benefits and risks of such therapy should be reviewed with the patient, giving more emphasis to absolute than to relative measures of effect and pointing out uncertainties in clinical knowledge where relevant. Because chronic disease rates generally increase with age, absolute risks tend to be greater in older women, even when RRs remain similar. Potential side effects—especially vaginal bleeding that may result from use of the combined estrogen-progestogen formulations recommended for women with an intact uterus—should be noted. The patient's own preference regarding therapy should be elicited and factored into the

decision. Contraindications should be assessed routinely and include unexplained vaginal bleeding; liver dysfunction or disease; venous thromboembolism; known blood clotting disorder or thrombophilia (transdermal estrogen may be an option); untreated hypertension; history of endometrial cancer (except stage 1 without deep invasion), breast cancer or other estrogen-dependent cancer; and history of CHD, stroke, or transient ischemic attack. Relative contraindications to systemic HT include an elevated risk of breast cancer (e.g., women who have one or more first-degree relatives with breast cancer, susceptibility genes such as *BRCA1* or *BRCA2*, a personal history of cellular atypia detected by breast biopsy [see also Breast Cancer Risk Score at *http://www.cancer.gov/bcrisktool/*]); hypertriglyceridemia (>400 mg/dL); and active gallbladder disease (transdermal estrogen may be an option in the latter two cases). Primary prevention of heart disease should not be viewed as an expected benefit of HT, and an increase in the risk of stroke as well as a small early increase in the risk of coronary artery disease should be considered. Nevertheless, such therapy may be appropriate if the noncoronary benefits of treatment clearly outweigh the risks. A woman who suffers an acute coronary event or stroke while taking HT should discontinue therapy immediately.

Short-term use (<5 years for estrogen-progestogen and <7 years for estrogen alone) is appropriate for relief of menopausal symptoms among women without contraindications to such use. However, such therapy should be avoided by women with an elevated baseline risk of future cardiovascular events. Women who have contraindications for or are opposed to HT may derive benefit from the use of certain antidepressants (including venlafaxine, fluoxetine, or paroxetine), gabapentin or pregabalin, or clonidine, and, for genitourinary symptoms, intravaginal estrogen creams or devices, ospemifene, or prasterone.

Long-term use (≥5 years for estrogen-progestogen and ≥7 years for estrogen alone) is more problematic because a heightened risk of breast cancer must be factored into the decision, especially for estrogen-progestogen. Reasonable candidates for such use include the small percentage of postmenopausal women who have persistent severe vasomotor symptoms along with an increased risk of osteoporosis (e.g., those with osteopenia, a personal or family history of nontraumatic fracture, or a weight below 125 lbs), who also have no personal or family history of breast cancer in a first-degree relative or other contraindications, and who have a strong personal preference for therapy. Poor candidates are women with elevated cardiovascular risk, those at increased risk of breast cancer, and those at low risk of osteoporosis. Even for reasonable candidates, strategies to minimize dose and duration of use should be employed. For example, women using HT to relieve intense vasomotor symptoms in early postmenopause should consider discontinuing therapy within 5 years, resuming it only if such symptoms persist. Because of the role of progestogens in increasing breast cancer risk, regimens that employ cyclic rather than continuous progestogen exposure as well as formulations other than medroxyprogesterone acetate should be considered if treatment is extended. For prevention of osteoporosis, alternative therapies such as bisphosphonates or SERMs should be considered. Research on alternative progestogens and androgen-containing preparations has been limited, particularly with respect to long-term safety. Additional research on the effects of these agents on cardiovascular disease, glucose tolerance, and breast cancer will be of particular interest.

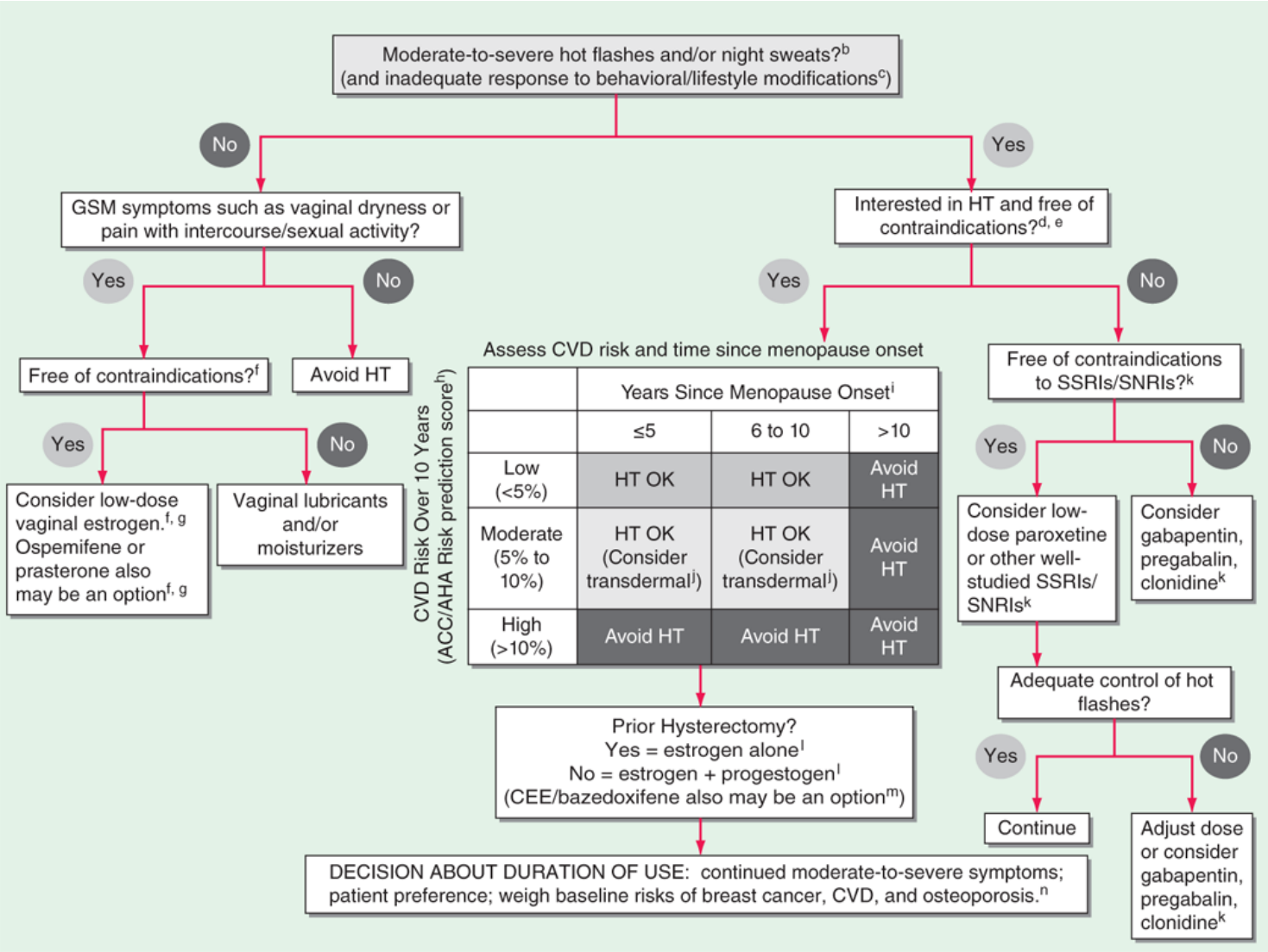
In addition to HT, lifestyle choices such as smoking abstention, adequate physical activity, and a healthy diet can play a role in controlling symptoms and preventing chronic disease. An expanding array of pharmacologic options (e.g., bisphosphonates, SERMs, and other agents for osteoporosis; cholesterol-lowering or antihypertensive agents for cardiovascular disease) should also reduce the widespread reliance on hormone use. However, short-term HT may still benefit some women.

FIGURE 388-4

Algorithm for menopausal symptom management.[a] The algorithm was developed in collaboration with the North American Menopause Society and is available in a free mobile app called MenoPro (dual mode for clinicians and patients). ACC, American College of Cardiology; AHA, American Heart Association; CEE, conjugated equine estrogens; CVD, cardiovascular disease; GSM, genitourinary syndrome of menopause; HT, hormone therapy; SNRIs, serotonin-norepinephrine reuptake inhibitors; SSRIs, selective serotonin reuptake inhibitors. *(Adapted from JE Manson et al: Menopause 22:247, 2015. Used with permission.)*

[a]Algorithm applies to women with menopausal symptoms who are aged ≥45 years and to women who have had removal of both ovaries, regardless of age. Women aged <45 years or those with uncertain menopausal status may need additional clinical evaluation before applying this algorithm. [b]Women who are at high risk of osteoporotic fracture but are unable to tolerate alternative preventive medications may also be reasonable candidates for systemic HT even if they do not have moderate to severe vasomotor symptoms. [c]Patients should try lifestyle modifications for at least 3 months before using this algorithm. (A patient handout with suggested lifestyle modifications can be found at *http://www.menopause.org/docs/for-women/mnflashes.pdf*.) [d]Reassess each step at least once every 6–12 months (assuming the patient's continued preference for HT) or if the patient's health status changes. [e]See text for contraindications to systemic HT. [f]Contraindications to low-dose vaginal estrogen include unexplained vaginal bleeding, and breast cancer, endometrial cancer, or other estrogen-dependent cancer. Contraindications to ospemifene and prasterone are similar to those for low-dose vaginal estrogen, and contraindications for ospemifene additionally include venous or arterial thromboembolic disease, severe liver disease, and use of estrogens or estrogen agonists-antagonists. MenoPro, a free mobile app from the North American Menopause Society, is available for further guidance on use of these medications for treatment of GSM symptoms. [g]Regularly updated tables of all U.S. and Canadian formulations approved by regulatory authorities for the treatment of GSM symptoms are available from the North American Menopause Society (*http://www.menopause.org/docs/default-source/professional/nams-ht-tables.pdf*). [h]Ten-year risk of cardiovascular disease, as assessed by ACC/AHA risk prediction score (DC Goff Jr et al: 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 129[Suppl 2]:149, 2014.) [i]Women >10 years past menopause are not good candidates for initiation (first use) of HT. [j]Consider avoiding oral HT. Transdermal HT may be a preferable option because it has a less adverse effect on clotting factors, triglyceride levels, and inflammation

factors than oral HT. [k]MenoPro, a free mobile app from the North American Menopause Society, is available for further guidance on use of these nonhormonal medications (as well as hormonal options) for treatment of vasomotor symptoms of menopause. [l]Regularly updated tables of all U.S. and Canadian formulations approved by regulatory authorities for the treatment of menopausal vasomotor symptoms are available from the North American Menopause Society (http://www.menopause.org/docs/default-source/professional/nams-httables.pdf). In the United States, the most commonly prescribed oral estrogens for systemic treatment of vasomotor symptoms are 17β-estradiol (1.0 mg/d, 0.5 mg/d, and other doses) and conjugated equine estrogens (CEE, 0.625 mg/d, 0.3 mg/d, and other doses). The most commonly prescribed transdermal estrogen products are 17β-estradiol skin patches (0.0375 mg/d, 0.05 mg/d, and other doses). The most commonly prescribed progestogens are medroxyprogesterone acetate (MPA, 2.5 mg/d, 5 mg/d, and 10 mg/d) and micronized progesterone (100 mg/d and 200 mg/d). Also available are oral estrogen-progestin combinations, such as oral CEE and MPA, oral 17β-estradiol or ethinyl estradiol with norethindrone acetate, and other options. [m]CEE/bazedoxifene may be an option for women with a uterus, especially those with concerns about breast tenderness, breast density, or uterine bleeding. Contraindications to CEE/bazedoxifene are similar to those for systemic HT. [n]See text for additional discussion.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

bassuk ss, manson je: Menopausal hormone therapy and cardiovascular disease risk: Utility of biomarkers and clinical factors for risk stratification. Clin Chem 60:68, 2014. [PubMed: 24379312]

canonico m et al: Hormone replacement therapy and risk of venous thromboembolism in postmenopausal women: Systematic review and meta-analysis. BMJ 336:1227, 2008.

chlebowski rt et al: Oestrogen plus progestin and lung cancer in postmenopausal women (Women's Health Initiative trial): A post-hoc analysis of a randomised controlled trial. Lancet 374:1243, 2009. [PubMed: 19767090]

kaunitz am, manson je: Management of menopausal symptoms. Obstet Gynecol 126:859, 2015. [PubMed: 26348174]

manson JE, bassuk SS: *Hot Flashes, Hormones & Your Health*. New York: McGraw-Hill, 2007.

manson je et al: Menopausal hormone therapy and health outcomes during the intervention and extended poststopping phases of the Women's Health Initiative randomized trials. JAMA 310:1353, 2013. [PubMed: 24084921]

manson je et al: Algorithm and mobile app for menopausal symptom management and hormonal/non-hormonal therapy decision making: A clinical decision-support tool from the North American Menopause Society. Menopause 22:247, 2015. [PubMed: 25314150]

north american menopause society: Nonhormonal management of menopause-associated vasomotor symptoms: 2015 position statement of the North American Menopause Society. Menopause 22:1155, 2015. [PubMed: 26382310]

North American Menopause Society: The 2017 hormone therapy position statement of the North American Menopause Society. Menopause 24:728, 2017. [PubMed: 28650869]

rossouw je et al: Postmenopausal hormone therapy and risk of cardiovascular disease by age and years since menopause. JAMA 297:1465, 2007. [PubMed: 17405972]

Chapter 389: Infertility and Contraception

Janet E. Hall

FIGURE 389-1

INFERTILITY

DEFINITION AND PREVALENCE

Infertility is the inability to conceive after 12 months of unprotected sexual intercourse or after 6 months in women ≥35. This revised definition is based on data indicating that 50% of apparently normal couples will conceive within 3 months, 75–82% within 6 months, and 85–92% within 12 months, but recognizes the age-related decrease in fertility. In the United States, the overall rate of infertility in married women aged 15–44 is 6.7% based on the recent National Survey of Family Growth. The infertility rate has remained relatively stable over the past 30 years in most countries. However, the proportion of couples without children has risen, reflecting both higher numbers of couples in childbearing years and a trend to delay childbearing. This trend has important implications because of the age-related decrease in fecundability, the ability to conceive and carrying a baby to term; the incidence of primary impaired fecundability increases from ~15% between the ages of 15 and 29 to 18% between the ages of 30 and 35, and 40% between the ages of 35 and 44. It is estimated that 12% of women in the United States have received medical assistance for infertility, although this represents <50% of women with current fertility problems. Both infertility and the use of medical services increase with age and both are affected by race and ethnicity. There is increased infertility in non-Hispanic black women and lower use of fertility services among Hispanic and non-Hispanic black women, suggesting disparities in access to care.

GLOBAL CONSIDERATIONS

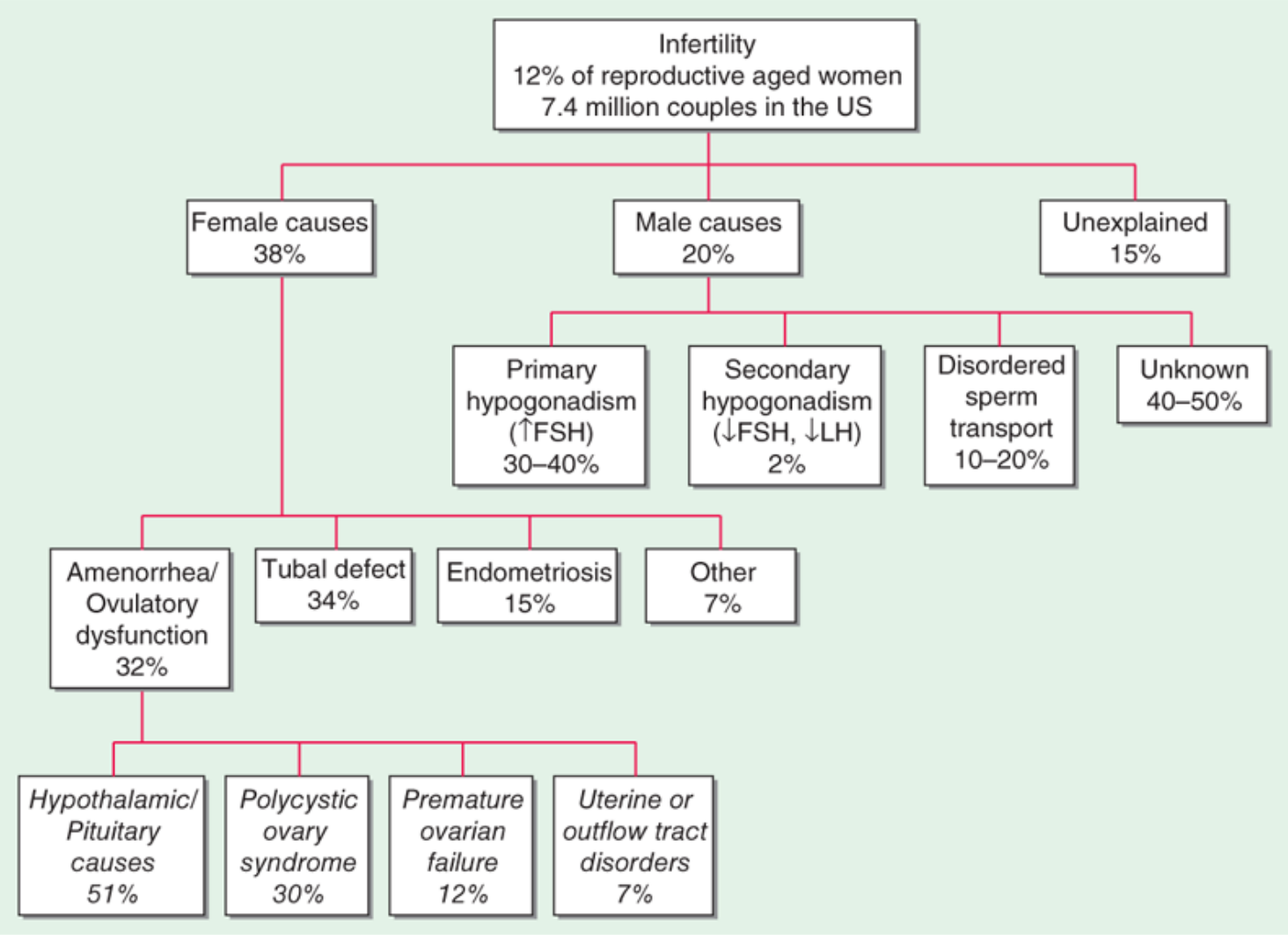
The World Health Organization (WHO) considers infertility as a disability (an impairment of function) and thus access to health care for this indication falls under the Convention on the Rights of Persons with Disability. Thirty-four million women, predominantly from developing countries, have infertility resulting from maternal sepsis and unsafe abortion. In populations <60 years old, infertility is ranked the fifth highest serious global disability.

CAUSES OF INFERTILITY

The spectrum of infertility ranges from reduced conception rates or the need for medical intervention to irreversible causes of infertility. Infertility can be attributed primarily to male factors in 20% of couples and female factors in 38% of couples and is unexplained in about 15% of couples **(Fig. 389-1)**. Both male and female factors contribute to infertility in 25% of couples. Decreases in the ability to conceive as a function of age in women has led to recommendations that not only should women ≥34 years old seek attention sooner, but that they also receive an expedited workup and approach to treatment.

FIGURE 389-1

Causes of infertility. FSH, follicle-stimulating hormone; LH, luteinizing hormone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Infertility

INITIAL EVALUATION

In all couples presenting with infertility, the initial evaluation includes discussion of the appropriate timing of intercourse and discussion of modifiable risk factors such as smoking, alcohol, caffeine, and obesity. The range of required investigations should be reviewed as well as a brief description of infertility treatment options, including adoption. Initial investigations are focused on determining whether the primary cause of the infertility is male, female, or both. These investigations include a semen analysis in the male, confirmation of ovulation in the female, and, in the majority of situations, documentation of tubal patency in the female. In some cases, after an extensive workup excluding all male and female factors, a specific cause cannot be identified, and infertility may ultimately be classified as unexplained.

PSYCHOLOGICAL ASPECTS OF INFERTILITY

Infertility is invariably associated with psychological stress related not only to the diagnostic and therapeutic procedures themselves but also to repeated cycles of hope and loss associated with each new procedure or cycle of treatment that does not result in the birth of a child. These feelings are often combined with a sense of isolation from friends and family. Counseling and stress-management techniques should be introduced early in the evaluation of infertility. Importantly, infertility and its treatment do not appear to be associated with long-term psychological sequelae.

FEMALE CAUSES

Abnormalities in menstrual function constitute the most common cause of female infertility. These disorders, which include ovulatory dysfunction and abnormalities of the uterus or outflow tract, may present as amenorrhea or as irregular or short menstrual cycles. A careful history and physical examination and a limited number of laboratory tests will help to determine whether the abnormality is (1) hypothalamic or pituitary (low follicle-stimulating hormone [FSH], luteinizing hormone [LH], and estradiol with or without an increase in prolactin), (2) polycystic ovary syndrome (PCOS; irregular cycles, hyperandrogenism and/or polycystic ovarian pathology assessed by ultrasound in the absence of other causes of androgen excess), (3) ovarian (low estradiol with increased FSH), or (4) a uterine or outflow tract abnormality. The frequency of these diagnoses depends on whether the amenorrhea is primary or occurs after normal puberty and menarche **(see Fig. 386-2)**.

The approach to further evaluation of these disorders is described in detail in Chap. 386.

Ovulatory Dysfunction

In women with a history of regular menstrual cycles, *evidence of ovulation* should be sought **(Chap. 385)**. Even in the presence of ovulatory cycles, evaluation of *ovarian reserve* is recommended for women age >35 years if they are interested in fertility. Measurement of FSH on day 3 of the cycle (an FSH level <10 IU/mL on cycle day 3 predicts adequate ovarian oocyte reserve) is the most cost-effective test. Other tests antral follicle count on ultrasound, and anti-müllerian hormone (AMH; <0.5 ng/mL predicts reduced ovarian reserve although there is variability between labs). Importantly, tests of ovarian reserve help to predict the response to exogenous gonadotropins but do not predict ability to conceive.

Tubal Disease

Tubal dysfunction may result from pelvic inflammatory disease (PID), appendicitis, endometriosis, pelvic adhesions, tubal surgery, previous use of an intrauterine device (IUD), and a previous ectopic pregnancy. However, a cause is not identified in up to 50% of patients with documented tubal factor infertility. Because of the high prevalence of tubal disease, evaluation of tubal patency by hysterosalpingogram (HSG) or laparoscopy should occur early in the majority of couples with infertility. Subclinical infections with *Chlamydia trachomatis* may be an underdiagnosed cause of tubal infertility and require the treatment of both partners.

Endometriosis

Endometriosis is defined as the presence of endometrial glands or stroma outside the endometrial cavity and uterine musculature. Its presence is suggested by a history of dyspareunia (painful intercourse), worsening dysmenorrhea that often begins before menses, or a thickened rectovaginal septum or deviation of the cervix on pelvic examination. Mild endometriosis does not appear to impair fertility; the pathogenesis of the infertility associated with moderate and severe endometriosis may be multifactorial with impairments of folliculogenesis, fertilization, and implantation,

as well as adhesions. Endometriosis is often clinically silent, however, and can only be excluded definitively by laparoscopy.

MALE CAUSES

Known causes of male infertility include primary testicular disease, genetic disorders (particularly Y chromosome microdeletions), disorders of sperm transport, and hypothalamic-pituitary disease resulting in secondary hypogonadism (See also Chap. 384). However, the etiology is not ascertained in up to one-half of men with suspected male factor infertility. The key initial diagnostic test is a *semen analysis*. Testosterone levels should be measured if the sperm count is low on repeated examination or if there is clinical evidence of hypogonadism. Gonadotropin levels will help to determine a gonadal versus a central cause of hypogonadism.

TREATMENT

TREATMENT

Infertility

In addition to addressing the negative impact of smoking on fertility and pregnancy outcome, counseling about nutrition and weight is a fundamental component of infertility and pregnancy management. Both low and increased body mass index (BMI) are associated with infertility in women and with increased morbidity during pregnancy. Obesity has also been associated with reduced fertility in men. The treatment of infertility should be tailored to the problems unique to each couple. In many situations, including unexplained infertility, mild-to-moderate endometriosis, and/or borderline semen parameters, a stepwise approach to infertility is optimal, beginning with low-risk interventions and moving to more invasive, higher risk interventions only if necessary. After determination of all infertility factors and their correction, if possible, this approach might include, in increasing order of complexity: (1) expectant management, (2) clomiphene citrate or an aromatase inhibitor (see below) with or without intrauterine insemination (IUI), (3) gonadotropins with or without IUI, and (4) in vitro fertilization (IVF). The time used for evaluation, correction of problems identified, and expectant management can be longer in women age <30 years, but this process should be advanced rapidly in women aged >35 years. In some situations, expectant management will not be appropriate.

OVULATORY DYSFUNCTION

Treatment of ovulatory dysfunction should first be directed at identification of the etiology of the disorder to allow specific management when possible. Dopamine agonists, for example, may be indicated in patients with hyperprolactinemia **(Chap. 373)**; lifestyle modification may be successful in women with low body weight, a history of intensive exercise or obesity.

Medications used for ovulation induction include agents that increase FSH through alteration of negative feedback, gonadotropins, and pulsatile GnRH. *Clomiphene citrate* is a nonsteroidal estrogen antagonist that increases FSH and LH levels by blocking estrogen negative feedback at the hypothalamus. The efficacy of clomiphene for ovulation induction is highly dependent on patient selection. It induces ovulation in ~60% of women with PCOS and has traditionally been the initial treatment of choice. Combination with agents that modify insulin levels such as metformin does not appear to improve outcomes. Clomiphene citrate is less successful in patients with hypogonadotropic hypogonadism. *Aromatase inhibitors* are also used for treatment of infertility. Studies suggest they may have advantages over clomiphene in some populations. *Estrogen receptor blockade* using Tamoxifen has been used in conjunction with gonadotropins in breast cancer patients undergoing in vitro fertilization (IVF) for embryo

banking. *Gonadotropins* are highly effective for ovulation induction in women with hypogonadotropic hypogonadism and PCOS and are used to induce the development of multiple follicles in unexplained infertility and in older reproductive-age women. Disadvantages include a significant risk of multiple gestation and the risk of ovarian hyperstimulation, particularly in women with polycystic ovaries, with or without other features of PCOS. Careful monitoring and a conservative approach to ovarian stimulation reduce these risks. Currently available gonadotropins include urinary preparations of LH and FSH, highly purified FSH, and recombinant FSH. Although FSH is the key component, LH is essential for steroidogenesis in hypogonadotropic patients, and LH or human chorionic gonadotropin (hCG) may improve results through effects on terminal differentiation of the oocyte. These methods are commonly combined with IUI.

None of these methods are effective in women with premature ovarian failure, in whom donor oocyte or adoption are the methods of choice.

TUBAL DISEASE

If hysterosalpingography suggests a tubal or uterine cavity abnormality or if a patient is aged ≥35 at the time of initial evaluation, laparoscopy with tubal lavage is recommended, often with a hysteroscopy. Although tubal reconstruction may be attempted if tubal disease is identified, it is generally being replaced by the use of IVF. These patients are at increased risk of developing an ectopic pregnancy.

ENDOMETRIOSIS

Although 60% of women with minimal or mild endometriosis may conceive within 1 year without treatment, laparoscopic resection or ablation appears to improve conception rates. Medical management of advanced stages of endometriosis is widely used for symptom control but has not been shown to enhance fertility. In moderate and severe endometriosis, conservative surgery is associated with pregnancy rates of 50 and 39%, respectively, compared with rates of 25 and 5% with expectant management alone. In some patients, IVF may be the treatment of choice.

MALE FACTOR INFERTILITY

The treatment options for male factor infertility have expanded in recent years (Chap. 384). Secondary hypogonadism is highly amenable to treatment with gonadotropins or pulsatile gonadotropin-releasing hormone (GnRH) where available. In vitro techniques have provided new opportunities for patients with primary testicular failure and disorders of sperm transport. Choice of initial treatment options depends on sperm concentration and motility. Expectant management should be attempted initially in men with mild male factor infertility (sperm count of 15 to 20 $\times 10^6$/mL and normal motility). Moderate male factor infertility (10–15 $\times 10^6$/mL and 20–40% motility) should begin with IUI alone or in combination with treatment of the female partner with ovulation induction, but it may require IVF with or without intracytoplasmic sperm injection (ICSI). For men with a severe defect (sperm count of $<10 \times 10^6$/mL, 10% motility), IVF with ICSI or donor sperm should be used. If ICSI is performed because of azoospermia due to congenital bilateral absence of the vas deferens, genetic testing for *CFTR* gene mutations and counseling should be provided because of the risk of cystic fibrosis.

ASSISTED REPRODUCTIVE TECHNOLOGIES

The development of assisted reproductive technologies (ARTs) has dramatically altered the treatment of male and female infertility. IVF is indicated for patients with many causes of infertility that have not been successfully managed with more conservative approaches. IVF or ICSI is often the treatment of choice in couples with a significant male factor or tubal disease, whereas IVF using donor oocytes is used in patients with premature ovarian failure and in women of advanced reproductive age. Success rates are influenced by cause of infertility and age, varying between

48% in women <35 to ≤10% in women >40. Success rates are highest in anovulatory women and lowest in women with decreased ovarian reserve. In the United States, success rates are higher in white than in black, Asian, or Hispanic women. Although often effective, IVF is costly and requires careful monitoring of ovulation induction and the use of invasive techniques, including the aspiration of multiple follicles. IVF is associated with a significant risk of multiple gestation, particularly in women age <35, in whom the rate can be as high as 30%. However, improved techniques and recognition of the risk associated with even twin pregnancies has led to adoption of age-specific guidelines by many clinics and a significant decline in the rate of twins (<25%) and very few higher order multiple births.

CONTRACEPTION

Only 15% of married couples in the United States report having unprotected sexual intercourse in the past 3 months. Although recent statistics indicate a decrease in unintended pregnancy in the United States, 45% of births are still the result of unintended pregnancy; approximately one-third of these result from incorrect use or failure of contraceptives, and >50% result in induced abortion. Unintended pregnancy is higher in Latina than white women with black women having the highest rates. Teenage pregnancies continue to represent a serious public health problem in the United States, with >1 million unintended pregnancies each year—a significantly greater incidence than in other industrialized nations. However, changes in teen behaviors are occurring, with an increase in contraceptive use at both first and most recent sexual encounter.

GLOBAL CONSIDERATIONS


The use of contraception in women aged 15–49 years who were married or in a union doubled worldwide from 36% in 1970 to 64% in 2015. The absolute number of married women who use contraception is projected to increase to nearly 800 million by 2030. However, there remains an unmet need for family planning in at least 10% of the population in most regions of the world.

Of the contraceptive methods available (Table 389-1), a reversible form of contraception is used by >50% of couples with a significant increase in the use of long-acting forms such IUDs in the past decade. Sterilization (male or female) is used as a permanent form of contraception by over one-third of couples. Pregnancy termination is relatively safe when directed by health care professionals but is rarely the option of choice.

TABLE 389-1

Effectiveness of Different Forms of Contraception

Method of Contraception	Theoretical Effectiveness (%)	Actual Effectiveness (%)	Continued Use at 1 Year (%)	Use of Method by U.S. Women at Risk of Unintended Pregnancy (%)
No Method	15	15		10
Fertility Awareness	96	76	47	1.2
Withdrawal	96	78	46	4.4
Barrier Methods				
Condoms	98	82	43	13.7
Diaphragm	94	82	57	2
Spermicides	82	72	43	1
Sterilization				
Female	99.5	99.5	100	22.6
Male	99.5	99.9	100	7.4
Intrauterine Device				9.3
Copper T	99.4	99.8	85	
Progestin-containing	99.8	99.8	88	
Hormonal Contraceptives				
Combined and Progestin only	99.7	91	67	23.3
Transdermal Patch	99.7	91	67	0.5

Method of Contraception	Theoretical Effectiveness (%)	Actual Effectiveness (%)	Continued Use at 1 Year (%)	Use of Method by U.S. Women at Risk of Unintended Pregnancy (%)
Vaginal Ring	99.7	91	67	1.8
Implant				1.2
Depot Provera	99.8	94	56	
Nexplanon	99.5	99.5	84	
Emergency Contraception				0.3

Adapted from J Trussell et al: Contraceptive Efficacy, in *Contraceptive Technology*, 20th revised ed, RA Hatcher et al (eds). New York, Ardent Media, 2011; CDC. NCHS National Survey of Family Growth, 2011-2013; J Jones, WD Mosher, K Daniels: Current contraceptive use in the United States, 2006-2010, and changes in patterns of use since 1995, National Health Statistics Reports, 2012, No. 60, *http://www.cdc.gov/nchs/data/nhsr/nhsr060.pdf*; and NE Birgisson et al: Preventing unintended pregnancy: The contraceptive CHOICE project in review. J Womens Health (Larchmt) 24:349, 2015.

No single contraceptive method is ideal, although all are safer than carrying a pregnancy to term. The effectiveness of a given method of contraception does not just depend on the efficacy of the method itself. Discrepancies between theoretical and actual effectiveness emphasize the importance of patient education and adherence when considering various forms of contraception (Table 389-1). Contraceptive use is stratified by race/ethnicity with higher oral contraceptive use in white women and greater use of long-acting reversible contraceptive (LARC) methods in Latina women. For oral contraceptives, discontinuation rates are highest among Black women whereas there is no racial/ethnic differences for LARC methods. Knowledge of the advantages and disadvantages of each contraceptive is essential for counseling an individual about the methods that are safest and most consistent with his or her lifestyle. The WHO has extensive family planning resources for the clinician and patient that can be accessed online. Similar resources for determining medical eligibility are available through the Centers for Disease Control and Prevention (CDC). Considerations for contraceptive use in obese patients and after bariatric surgery are discussed below.

BARRIER METHODS

Barrier contraceptives (such as condoms, diaphragms, and cervical caps) and spermicides are easily available, reversible, and have fewer side effects than hormonal methods. However, their effectiveness is highly dependent on adherence and proper use (Table 389-1). A major advantage of barrier contraceptives is the protection provided against sexually transmitted infections (STIs) **(Chap. 131)**. Consistent use is associated with a decreased risk of HIV, gonorrhea, nongonococcal urethritis, and genital herpes, probably due in part to the concomitant use of spermicides. Natural membrane condoms may be less effective than latex condoms for prevention of sexually transmitted diseases, and petroleum-based lubricants can degrade condoms and decrease their efficacy for preventing HIV infection. Barrier methods used by women include the diaphragm, cervical cap, and contraceptive sponge. There is some evidence that the diaphragm is more effective when used in conjunction with a spermicide. The cervical cap and

sponge are less effective than the diaphragm, and there have been rare reports of toxic shock syndrome with the diaphragm and contraceptive sponge.

STERILIZATION

Sterilization procedures are highly effective for both men and women (Table 389-1) and are commonly chosen by fertile men and multiparous women >30 years old. Sterilization refers to a procedure that prevents fertilization by surgical interruption of the fallopian tubes in women or the vas deferens in men. Although tubal ligation and vasectomy are potentially reversible, these procedures should be considered permanent and should not be undertaken without patient counseling.

Tubal ligation methods are highly effective with a 10-year cumulative pregnancy rate of 1.85 per 100 women with both postpartum or interval procedures. However, when pregnancy does occur, the risk of ectopic pregnancy may be as high as 30%. The success rate of tubal reanastomosis depends on the method of ligation used, but even after successful reversal, the risk of ectopic pregnancy remains high. The use of salpingectomy has increased to 33% with emerging evidence that ovarian cancer originates from dysplastic cells in the fallopian tube. Hysteroscopic sterilization has been used, particularly in women with pelvic adhesions or other co-morbidities. Essure is the most commonly used commercially available product and involves insertion of a nickel-titanium double coil which results in tubal fibrosis. Data indicate very low unintended pregnancy rates (1.5 per 1000 women) with ultrasound and/or HSG confirmation of correct placement. Although still available, the U.S. Food and Drug Administration (FDA) issued a black box warning in 2015 due to post-marketing reports of long-term pain, abnormal bleeding, and allergic reactions. Intrauterine quinacrine is also effective and has been used for many years in resource-poor settings. *Vasectomy* is a highly effective and low risk outpatient surgical procedure. The no-scalpel technique, which is used in the United States, results in fewer complications, but has not been accepted worldwide. The development of azoospermia may be delayed for 2–6 months, and other forms of contraception must be used until two sperm-free ejaculations provide proof of sterility. Current data indicate that reanastomosis may restore fertility in 50–70% of men, but the success rate declines with time after vasectomy and may be influenced by factors such as the development of antisperm antibodies which occurs in 60–80% of men.

INTRAUTERINE DEVICES

IUDs inhibit pregnancy through a spermicidal effect (copper IUDs) or by inhibiting ovulation (progestin containing devices). IUDs provide a high level of efficacy in the absence of systemic metabolic effects, and ongoing motivation is not required to ensure efficacy once the device has been placed. IUD use is greatest in Europe and Canada (33%) but is increasing in other parts of the world, including the United States. An IUD should not be used in women at high risk for development of STI or in women at high risk for bacterial endocarditis. Progestin-containing IUDs are contraindicated in women with breast cancer. The IUD may not be effective in women with uterine leiomyomas because they alter the size or shape of the uterine cavity. IUD use is associated with increased menstrual blood flow, although this is less pronounced with the progestin-releasing IUD.

HORMONAL METHODS

Oral Contraceptive Pills

Because of their ease of use and efficacy, oral contraceptive pills are the most widely used form of hormonal contraception. They act by suppressing ovulation, changing cervical mucus, and altering the endometrium. The current formulations are made from synthetic estrogens and progestins. The estrogen component of the pill consists of ethinyl estradiol or mestranol, which is metabolized to ethinyl estradiol. Multiple synthetic progestins are used. Norethindrone and its derivatives are used in many formulations. Low-dose norgestimate and the more recently developed (third-generation) progestins (desogestrel, gestodene, drospirenone) have a less androgenic profile; levonorgestrel appears to be the most androgenic of the progestins and should be avoided in patients with hyperandrogenism. The three major formulations of oral contraceptives are (1) fixed-dose estrogen-progestin combination, (2) phasic estrogen-progestin combination, and (3) progestin only. Each of these formulations is administered daily for 3 weeks followed by a week of no medication during which menstrual bleeding generally occurs. Two extended oral contraceptives are approved for use in the United States; Seasonale is a 3-month preparation with 84 days of active drug and 7 days of placebo, whereas Lybrel is a continuous preparation. Current doses of ethinyl estradiol range from 10 to 50 µg. However, indications for the 50-µg dose are rare, and the majority of formulations contain 20–35 µg of ethinyl estradiol. The reduced estrogen and progestin content in the second- and third-generation pills has decreased both side effects and risks associated with oral contraceptive use **(Table 389-2)**. At the currently used doses, patients must be cautioned not to miss pills due to the potential for ovulation and this may be particularly important in obese women. Side effects, including breakthrough bleeding, amenorrhea, breast tenderness, and weight gain, often respond to a change in formulation. Oral contraceptive use is associated with a decreased risk of endometrial, ovarian and colon cancer. However, even the lower dose oral contraceptives have been associated with an increased risk of breast cancer, cardiovascular disease (myocardial infarction, stroke, venous thromboembolism [VTE]), but the absolute excess risk is extremely low. VTE risk is higher with the third-generation than the second-generation progestins, and the risk of stroke and VTE is also higher with drospirenone (although not cyproterone). Again, the absolute excess risk is small. In addition to their use as highly effective contraceptives, estrogen-progestin combinations are used for treatment of amenorrhea and oligoamenorrhea and continuous formulations are commonly used for treatment of premenstrual syndrome and premenstrual dysphoric disorder, menstrual migraine, leiomyomas, and endometriosis.

TABLE 389-2

Oral Contraceptives: Contraindications and Disease Risk

Contraindications

Absolute

- Women age >35 years who smoke ≥15 cigarettes per day
- Known ischemic heart disease or multiple risk factors for cardiovascular disease (older age, smoking, diabetes, and hypertension)
- Previous thromboembolic event, stroke or known thrombogenic mutations
- Complicated valvular heart disease
- Complicated solid organ transplantation
- Hypertension (systolic ≥160 mmHg or diastolic ≥100 mmHg)
- Systemic lupus erythematous (positive or unknown antiphospholipid antibodies
- Cirrhosis, hepatic adenoma or hepatoma
- Pregnancy and early postpartum (<21 days)
- Undiagnosed abnormal uterine bleeding

Relative

- Hypertension (adequately controlled or systolic 140–159 or diastolic 90–99)
- Women receiving anticonvulsant drug therapy
- Women following bariatric surgery (malabsorptive procedure)

Disease Risks

Increased

- Coronary heart disease—increased in smokers >35; no relation to progestin type
- Hypertension—relative risk 1.8 (current users) and 1.2 (previous users)
- Venous thrombosis—relative risk ~4; may be higher with third-generation progestin, drospirenone, and patch; compounded by obesity (tenfold increased risk compared with nonobese, no OCP); markedly increased with factor V Leiden or prothrombin gene mutations
- Stroke—slight increase; unclear relation to migraine headache
- Cerebral vein thrombosis—relative risk ~13–15; synergistic with prothrombin gene mutation
- Cervical cancer—relative risk 2–4
- Breast cancer—may increase risk, particularly in carriers of *BRCA1* and possibly *BRCA2*

Decreased

- Ovarian cancer—50% reduction in risk
- Endometrial cancer—40% reduction in risk

Abbreviation: OCP, oral contraceptive pill.

The microdose progestin-only minipill is less effective as a contraceptive, having a pregnancy rate of 2–7 per 100 women-years. However, it may be appropriate for women at increased risk for cardiovascular disease or for women who cannot tolerate synthetic estrogens.

Alternative Methods

A *weekly contraceptive patch* (Ortho Evra) is available and has similar efficacy to oral contraceptives. Approximately 2% of patches fail to adhere, and a similar percentage of women have skin reactions. Efficacy is lower in women weighing >90 kg. The amount of estrogen delivered may be comparable to that of a 40-μg ethinyl estradiol oral contraceptive, raising the possibility of increased risk of VTE, which must be balanced against potential benefits for women not able to successfully use other methods. A *monthly contraceptive estrogen/progestin injection* (Lunelle) is highly effective, with a first-year failure rate of <0.2%, but it may be less effective in obese women. Its use is associated with bleeding irregularities that diminish over time. Fertility returns rapidly after discontinuation. A *monthly vaginal ring* (NuvaRing) that is intended to be left in place during intercourse is also available for contraceptive use. It is highly effective, with a 12-month failure rate of 0.7%. Ovulation returns within the first recovery cycle after discontinuation.

Long-Term Hormonal Contraceptives

There are two forms of long-acting hormonal contraceptives in addition to the progestin-containing IUD, both of which act by inhibiting ovulation and causing changes in the endometrium and cervical mucus that result in decreased implantation and sperm transport. Depot medroxyprogesterone acetate (Depo-Provera, DMPA) is effective for 3 months when administered SC or IM, but return of fertility after discontinuation may be delayed for up to 12–18 months. Irregular bleeding is common initially with amenorrhea in almost 70% of users at 2 years. Two percent of women discontinue use due to weight gain. Nexplanon is a hormonal implant that slowly releases the progestin, etonogestrel, and is approved by the FDA for 3 years of use. It is associated with unscheduled bleeding but has very favorable continuation rates, and is associated with rapid return of ovulation after removal. Both forms of contraception may be particularly good for women in whom an estrogen-containing contraceptive is contraindicated (e.g., migraine exacerbation, sickle cell anemia, fibroids). DMPA can induce bone loss and may be contraindicated in areas of high HIV prevalence because of potential negative effects on both immunologic regulation and genital track permeability. Neither of these side effects occur with use of the progestin implant. DMPA and Nexplanon are contraindicated in women with current breast cancer because of theoretical concerns about the adverse effects of progestins on breast cancer.

POSTCOITAL CONTRACEPTION

The probability of pregnancy without relation to time of the month is 8%, but the probability varies significantly in relation to proximity to ovulation and may be as high has 30%. In order of efficacy, methods of postcoital contraception include the following:

1. Copper IUD insertion within a maximum of 5 days has a reported efficacy of 99–100% and prevents pregnancy by its spermicidal effect; insertion is frequently available through family planning clinics, but may be associated with a higher risk of abdominal pain compared with other methods.

2. Oral antiprogestins (ulipristal acetate, 30 mg single dose, available worldwide, or mifepristone, 600 mg single dose, not available for this indication in the United States) prevent pregnancy by delaying or preventing ovulation; when administered, ideally within 72 h but up to 120 h after intercourse, they have an efficacy of 98–99%; require a prescription.

3. Levonorgestrel (1.5 mg as a single dose) delays or prevents ovulation and is not effective after ovulation; should be taken within 72 h of unprotected intercourse, and has an efficacy that varies between 60 and 94%; it is available over the counter.

Combined estrogen and progestin regimens have lower efficacy and are no longer recommended. A pregnancy test is not necessary before the use of oral methods, but pregnancy should be excluded before IUD insertion. Risk factors for failure of oral regimens include close proximity to ovulation and unprotected intercourse after use. In addition, there is an increased risk of pregnancy in obese and overweight women using levonorgestrel for postcoital contraception and an increased risk in obese women using an antiprogestin.

IMPACT OF OBESITY ON CONTRACEPTIVE CHOICE

Approximately one-third of adults in the United States are obese. Although obesity is associated with some reduction in fertility, the vast majority of obese women can conceive. The risk of pregnancy-associated complications is higher in obese women. Intrauterine contraception may be more effective than oral or transdermal methods for obese women. The WHO guidelines provide no restrictions (class 1) for the use of intrauterine contraception, DMPA, and progestin-only pills for obese women (BMI ≥30) in the absence of coexistent medical problems, whereas methods that include estrogen (pill, patch, ring) are considered class 2 (advantages generally outweigh theoretical or proven risks) due to the increased risk of thromboembolic disease. There are no restrictions to the use of any contraceptive methods following restrictive bariatric surgery procedures, but both combined and progestin-only pills are relatively less effective following procedures associated with malabsorption.

FURTHER READING

Birgisson NE et al: Preventing unintended pregnancy: The contraceptive CHOICE project in review. J Womens Health (Larchmt) 24:349, 2015. [PubMed: 25825986]

Broughton DE, Moley KH: Obesity and female infertility: Potential mediators of obesity's impact. Fertil Steril 107:840, 2017. [PubMed: 28292619]

Curtis KM, Peipert JF: Long-acting reversible contraception. N Engl J Med 376:461, 2017. [PubMed: 28146650]

Donnez J, Dolmans M-M: Fertility preservation in women. N Engl J Med 377:1657, 2017. [PubMed: 29069558]

Jackson AV et al: Racial and ethnic differences in contraception use and obstetric outcomes: A review. Semin Perinatol 41:273, 2017. [PubMed: 28651792]

Lee JK, Schwartz EB: The safety of available and emerging options for emergency contraception. Expert Opinion on Drug Safety 16:1163, 2017. [PubMed: 28730840]

Morch LS et al: Contemporary hormonal contraception and the risk of breast cancer. N Engl J Med 377:2228, 2017. [PubMed: 29211679]

Simmons KB, Elelman AB: Hormonal contraception and obesity. Fertil Steril 106:1282, 2016. [PubMed: 27565257]

Steiner AZ et al: Association between biomarkers of ovarian reserve and infertility among older women of reproductive age. JAMA 318:1367, 2017. [PubMed: 29049585]

Tepper NK et al: Update to CDC's U.S. medical eligibility Criteria for contraceptive use, 2016: Revised recommendations for the use of hormonal contraception among women at high risk for HIV infection. MMWR 66:990, 2017. [PubMed: 28934178]

Chapter 390: Sexual Dysfunction

Kevin T. McVary

INTRODUCTION

Male sexual dysfunction affects 10–25% of middle-aged and elderly men, and female sexual dysfunction occurs with a similar frequency. Demographic changes, the popularity of newer treatments, and greater awareness of sexual dysfunction by patients and society have led to increased diagnosis and associated health care expenditures for the management of this common disorder. Sexual health and satisfaction with sex life are important aspects of quality of life for many, including those in poor health. Because many patients are reluctant to initiate discussion of their sex lives, physicians should address this topic directly to elicit a history of sexual dysfunction. Specifically addressing sexual health should be a routine part of the clinical encounter.

MALE SEXUAL DYSFUNCTION

PHYSIOLOGY OF MALE SEXUAL RESPONSE

Normal male sexual function requires (1) an intact libido, (2) the ability to achieve and maintain penile erection, (3) ejaculation, and (4) detumescence. *Libido* refers to sexual desire and is influenced by a variety of visual, olfactory, tactile, auditory, imaginative, and hormonal stimuli. Sex steroids, particularly testosterone, act to increase libido. Libido can be diminished by hormonal or psychiatric disorders and by medications.

Penile tumescence leading to erection depends on an increased flow of blood into the lacunar network accompanied by complete relaxation of the arteries and corporal smooth muscle. The microarchitecture of the corpora is composed of a mass of smooth muscle (trabecula) that contains a network of endothelial-lined vessels (lacunar spaces). Subsequent compression of the trabecular smooth muscle against the fibroelastic tunica albuginea causes a passive closure of the emissary veins and accumulation of blood in the corpora. In the presence of a full erection and a competent valve mechanism, the corpora become noncompressible cylinders from which blood does not escape.

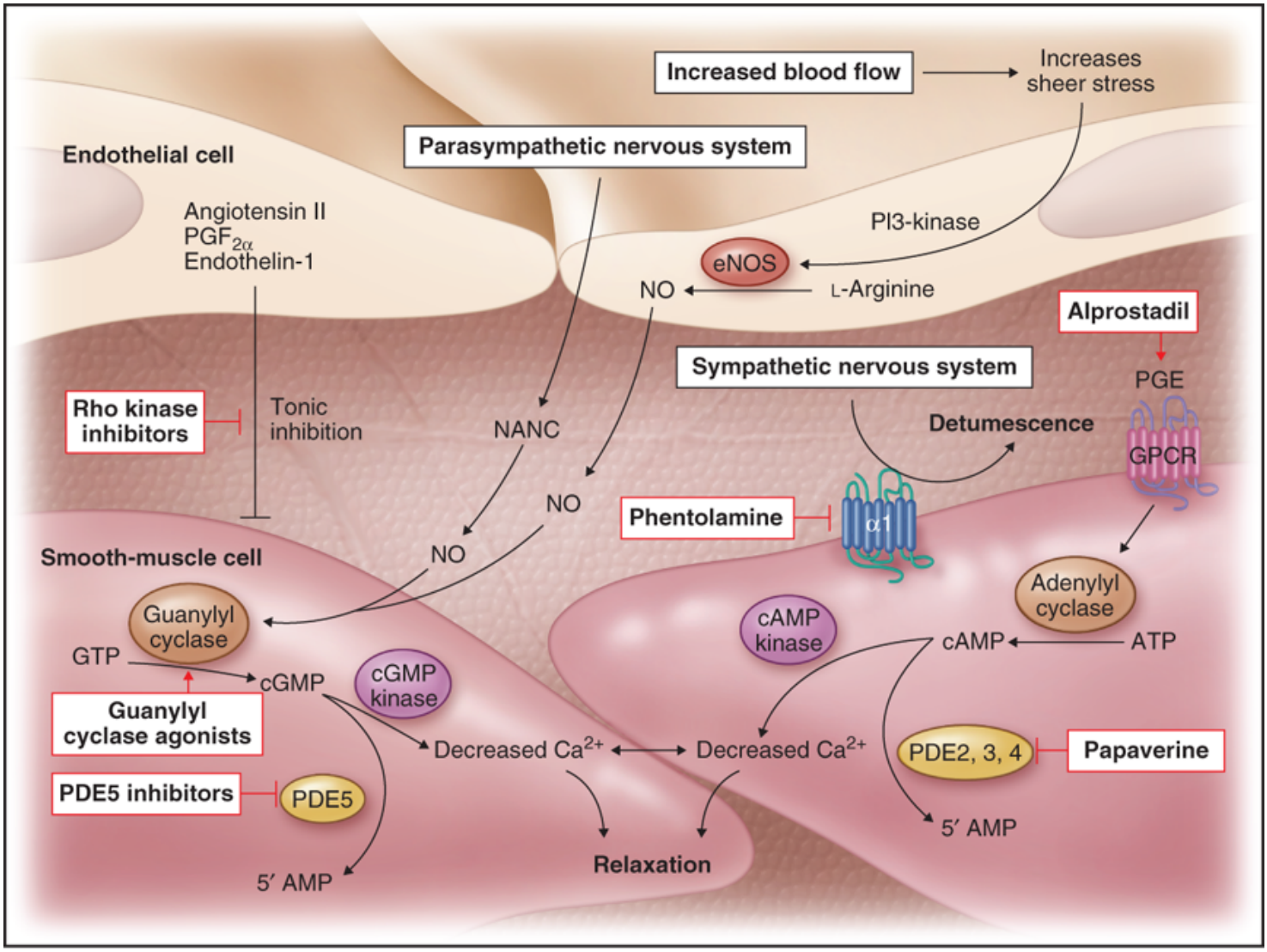
The central nervous system (CNS) exerts an important influence by either stimulating or antagonizing spinal pathways that mediate erectile function and ejaculation. The erectile response is mediated by a combination of central (psychogenic) innervation and peripheral (reflexogenic) innervation. Sensory nerves that originate from receptors in the penile skin and glans converge to form the dorsal nerve of the penis, which travels to the S2-S4 dorsal root ganglia via the pudendal nerve. Parasympathetic nerve fibers to the penis arise from neurons in the intermediolateral columns of the S2-S4 sacral spinal segments. Sympathetic innervation originates from the T-11 to the L-2 spinal segments and descends through the hypogastric plexus.

Neural input to smooth-muscle tone is crucial to the initiation and maintenance of an erection. There is also an intricate interaction between the corporal smooth-muscle cell and its overlying endothelial cell lining **(Fig. 390-1A)**. Nitric oxide, which induces vascular relaxation, promotes erection and is opposed by endothelin 1 (ET-1) and Rho kinase, which mediate vascular contraction. Nitric oxide is synthesized from L-arginine by nitric oxide synthase and is released from the nonadrenergic, noncholinergic (NANC) autonomic nerve supply to act postjunctionally on smooth-muscle cells. Nitric oxide increases the production of cyclic 3′,5′-guanosine monophosphate (cyclic GMP), which induces relaxation of smooth muscle **(Fig. 390-1B)**. Cyclic GMP is gradually broken down by phosphodiesterase type 5 (PDE-5). Inhibitors of PDE-5 such as the oral medications

sildenafil, vardenafil, and tadalafil maintain erections by reducing the breakdown of cyclic GMP **(Fig. 390-2)**. However, if nitric oxide is not produced at some level, PDE-5 inhibitors are ineffective, as these drugs facilitate, but do not initiate, the initial enzyme cascade. In addition to nitric oxide, vasoactive prostaglandins (PGE_1, $PGF_{2\alpha}$) are synthesized within the cavernosal tissue and increase cyclic AMP levels, also leading to relaxation of cavernosal smooth-muscle cells.

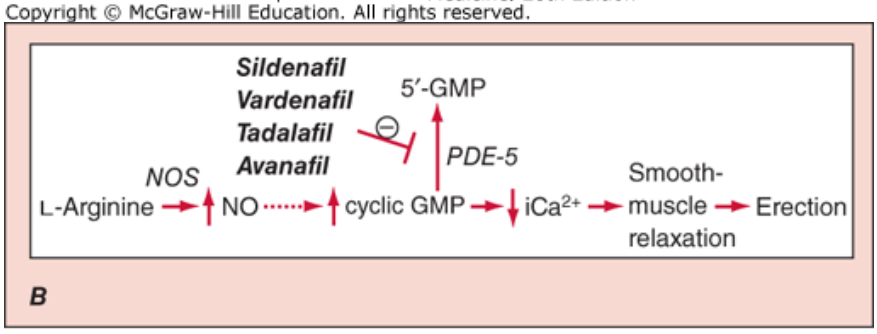
FIGURE 390-1

Pathways that control erection and detumescence. ***A.*** Outflow from the parasympathetic nervous system leads to relaxation of the cavernous sinusoids in two ways, both of which increase the concentration of nitric oxide (NO) in smooth-muscle cells. First, NO is the neurotransmitter in nonadrenergic, noncholinergic (NANC) fibers; second, stimulation of endothelial nitric oxide synthase (eNOS) through cholinergic output causes increased production of NO. The NO produced in the endothelium then diffuses into the smooth-muscle cells and decreases its intracellular calcium concentration through a pathway mediated by cyclic guanosine monophosphate (cGMP), leading to relaxation. A separate mechanism that decreases the intracellular calcium level is mediated by cyclic adenosine monophosphate (cAMP). With increased cavernosal blood flow, as well as increased levels of vascular endothelial growth factor (VEGF), the endothelial release of NO is further sustained through the phosphatidylinositol 3 (PI3) kinase pathway. Active treatments (red boxes) include drugs that affect the cGMP pathway (phosphodiesterase type 5 [PDE-5] inhibitors and guanylyl cyclase agonists), the cAMP pathway (alprostadil), or both pathways (papaverine), along with neural-tone mediators (phentolamine and Rho kinase inhibitors). Agents that are being developed include guanylyl cyclase agonists (to bypass the need for endogenous NO) and Rho kinase inhibitors (to inhibit tonic contraction of smooth-muscle cells mediated through endothelin). ***B.*** Biochemical pathways of NO synthesis and action. Sildenafil, vardenafil, tadalafil, and avanafil enhance erectile function by inhibiting PDE-5, thereby maintaining high levels of cyclic 3′,5′-guanosine monophosphate (cyclic GMP). α1, α-adrenergic receptor; GPCR, G-protein–coupled receptor; GTP, guanosine triphosphate; iCa^{2+}, intracellular calcium; NOS, nitric oxide synthase; PGE, prostaglandin E; PGF, prostaglandin F. *(Part A from K McVary: N Engl J Med 357:2472, 2007; with permission.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

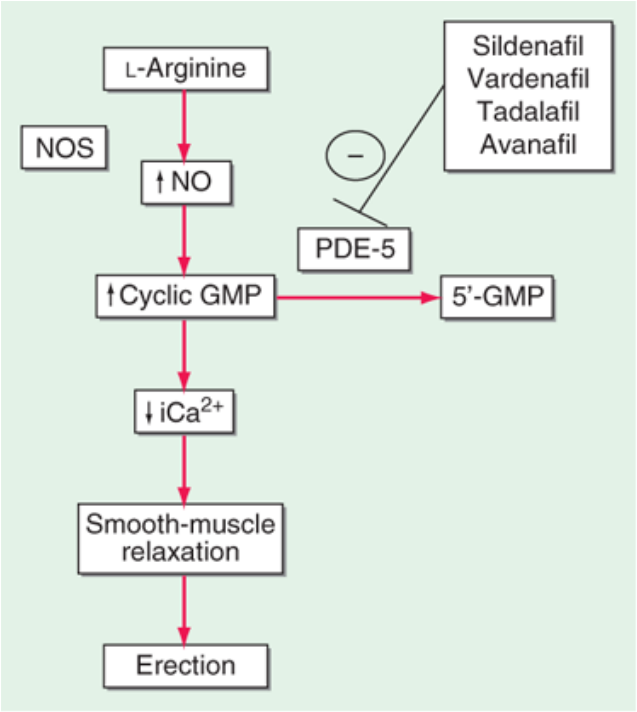


B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 390-2

Biochemical pathways modified by phosphodiesterase type 5 (PDE-5) inhibitors. Sildenafil, vardenafil, tadalafil, and avanafil enhance erectile function by inhibiting PDE-5, thereby maintaining high levels of cyclic 3′,5′-guanosine monophosphate (cyclic GMP). iCa^{2+}, intracellular calcium; NO, nitric oxide; NOS, nitric oxide synthase.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Ejaculation is stimulated by the sympathetic nervous system; this results in contraction of the epididymis, vas deferens, seminal vesicles, and prostate, causing seminal fluid to enter the urethra. Seminal fluid emission is followed by rhythmic contractions of the bulbocavernosus and ischiocavernosus muscles, leading to ejaculation. This is followed by expulsion, characterized by stereotypic rhythmic contractions of the striated perineal muscles, leading to forceful expulsion of semen with the bladder neck closed. This emission and expulsion are controlled by the autonomic (parasympathetic and sympathetic) and somatic spinal centers, respectively. The synchronization between autonomic and somatic spinal centers is orchestrated by interneurons that form a spinal ejaculation generator which is present in mammals including man.

Premature ejaculation usually is related to anxiety or a learned behavior and is amenable to behavioral therapy or treatment with medications such as selective serotonin reuptake inhibitors (SSRIs). *Retrograde ejaculation* results when the internal urethral sphincter does not close; it may occur in men with diabetes or after surgery involving the bladder neck.

Detumescence is mediated by norepinephrine from the sympathetic nerves, endothelin from the vascular surface, and smooth-muscle contraction induced by postsynaptic α-adrenergic receptors and activation of Rho kinase. These events increase venous outflow and restore the flaccid state. Venous leak can cause premature detumescence and is caused by insufficient relaxation of the corporal smooth muscle rather than a specific anatomic defect. *Priapism* refers to a persistent and painful erection and may be associated with sickle cell anemia, hypercoagulable states, spinal cord injury, or injection of vasodilator agents into the penis.

ERECTILE DYSFUNCTION

Epidemiology

Erectile dysfunction (ED) is not considered a normal part of the aging process. Nonetheless, it is associated with certain physiologic and psychological changes related to age. In the Massachusetts Male Aging Study (MMAS), a community-based survey of men aged 40–70, 52% of responders reported some degree of ED. Complete ED occurred in 10% of respondents, moderate ED in 25%, and minimal ED in 17%. The incidence of moderate or severe ED more than doubled between the ages of 40 and 70. In the National Health and Social Life Survey (NHSLS), which included a sample of men and women aged 18–59, 10% of men reported being unable to maintain an erection (corresponding to the proportion of men in the MMAS reporting severe ED).

Incidence was highest among men in the age group 50–59 (21%) and men who were poor (14%), divorced (14%), and less educated (13%).

The incidence of ED is also higher among men with certain medical disorders, such as diabetes mellitus, obesity, lower urinary tract symptoms secondary to benign prostatic hyperplasia (BPH), heart disease, hypertension, decreased high-density lipoprotein (HDL) levels and diseases associated with general systemic inflammation (e.g., rheumatoid arthritis). Cardiovascular disease and ED share etiologies as well as pathophysiology (e.g., endothelial dysfunction), and the degree of ED appears to correlate with the severity of cardiovascular disease. Consequently, ED represents a “sentinel symptom” in patients with occult cardiovascular and peripheral vascular disease.

Smoking is also a significant risk factor in the development of ED. Medications used in treating diabetes or cardiovascular disease are additional risk factors (see below). There is a higher incidence of ED among men who have undergone radiation or surgery for prostate cancer and in those with a lower spinal cord injury. Psychological causes of ED include depression, anger, stress from unemployment, anxiety (especially younger men), and other stress-related causes.

Pathophysiology

ED may result from three basic mechanisms: (1) failure to initiate (psychogenic, endocrinologic, or neurogenic), (2) failure to fill (arteriogenic), and (3) failure to store adequate blood volume within the lacunar network (venoocclusive dysfunction). These categories are not mutually exclusive, and multiple factors contribute to ED in many patients. For example, diminished filling pressure can lead secondarily to venous leak. Psychogenic factors frequently coexist with other etiologic factors and should be considered in all cases. Diabetic, atherosclerotic, and drug-related causes account for >80% of cases of ED in older men.

Vasculogenic

The most common organic cause of ED is a disturbance of blood flow to and from the penis. Atherosclerotic or traumatic arterial disease can decrease flow to the lacunar spaces, resulting in decreased rigidity and an increased time to full erection. Excessive outflow through the veins despite adequate inflow also may contribute to ED. Structural alterations to the fibroelastic components of the corpora may cause a loss of compliance and inability to compress the tunical veins. This condition may result from aging, increased cross-linking of collagen fibers induced by nonenzymatic glycosylation, hypoxemia, or altered synthesis of collagen associated with hypercholesterolemia.

Neurogenic

Disorders that affect the sacral spinal cord or the autonomic fibers to the penis preclude nervous system relaxation of penile smooth muscle, thus leading to ED. In patients with spinal cord injury, the degree of ED depends on the completeness and level of the lesion. Patients with incomplete lesions or injuries to the upper part of the spinal cord are more likely to retain erectile capabilities than are those with complete lesions or injuries to the lower part. Although 75% of patients with spinal cord injuries have some erectile capability, only 25% have erections sufficient for penetration. Other neurologic disorders commonly associated with ED include multiple sclerosis and peripheral neuropathy. The latter is often due to either diabetes or alcoholism. Pelvic surgery may cause ED through disruption of the autonomic nerve supply.

Endocrinologic

Androgens increase libido, but their exact role in erectile function is unclear. Individuals with castrate levels of testosterone can achieve erections from visual or sexual stimuli. Nonetheless, normal levels of testosterone appear to be important for erectile function, particularly in older males. Androgen replacement therapy can improve depressed erectile function when it is secondary to hypogonadism; however, it is not useful for ED when endogenous testosterone levels are normal. Increased prolactin may decrease libido by suppressing gonadotropin-releasing hormone (GnRH), and it also leads to decreased testosterone levels. Treatment of hyperprolactinemia with dopamine agonists can restore libido and testosterone.

Diabetic

ED occurs in 35–75% of men with diabetes mellitus. Pathologic mechanisms are related primarily to diabetes-associated vascular and neurologic complications. Diabetic macrovascular complications are related mainly to age, whereas microvascular complications correlate with the duration of diabetes and the degree of glycemic control (**Chap. 396**). Individuals with diabetes also have reduced amounts of nitric oxide synthase in both endothelial and neural tissues.

Psychogenic

Two mechanisms contribute to the inhibition of erections in psychogenic ED. First, psychogenic stimuli to the sacral cord may inhibit reflexogenic responses, thereby blocking activation of vasodilator outflow to the penis. Second, excess sympathetic stimulation in an anxious man may increase penile smooth-muscle tone. The most common causes of psychogenic ED are performance anxiety, depression, relationship conflict, loss of attraction, sexual inhibition, conflicts over sexual preference, sexual abuse in childhood, and fear of pregnancy or sexually transmitted disease. Almost all patients with ED, even when it has a clear-cut organic basis, develop a psychogenic component as a reaction to ED.

Medication-Related Medication-induced ED **(Table 390-1)** is estimated to occur in 25% of men seen in general medical outpatient clinics. The adverse effects related to drug therapy are additive, especially in older men. In addition to the drug itself, the underlying disease being treated is likely to contribute to sexual dysfunction. Among the antihypertensive agents, the thiazide diuretics and beta blockers have been implicated most frequently. Calcium channel blockers and angiotensin converting-enzyme inhibitors are cited less frequently. These drugs may act directly at the corporal level (e.g., calcium channel blockers) or indirectly by reducing pelvic blood pressure, which is important in the development of penile rigidity. α-Adrenergic blockers are less likely to cause ED. Estrogens, GnRH agonists, H_2 antagonists, and spironolactone cause ED by suppressing gonadotropin production or by blocking androgen action. Antidepressant and antipsychotic agents—particularly neuroleptics, tricyclics, and SSRIs—are associated with erectile, ejaculatory, orgasmic, and sexual desire difficulties. Among the SSRIs, paroxetine and escitalopram have been associated with the highest risk of sexual dysfunction. Bupropion, nefazodone, and mirtazapine appear less likely to cause sexual dysfunction. A number of molecular pathways have been implicated in antidepressant-induced sexual adverse events. Serotonin has been hypothesized to inhibit normal sexual response by decreasing dopamine-enhanced libido, arousal and erection, and by increasing prolactin release. SSRIs have also been shown to be potent inhibitors of nitric oxide synthesis.

TABLE 390- 1

Drugs Associated with Erectile Dysfunction

Classification	Drugs
Diuretics	Thiazides
	Spironolactone
Antihypertensives	Calcium channel blockers
	Methyldopa
	Clonidine
	Reserpine
	Beta blockers
	Guanethidine
Cardiac/antihyperlipidemics	Digoxin
	Gemfibrozil
	Clofibrate
Antidepressants	Selective serotonin reuptake inhibitors
	Tricyclic antidepressants
	Lithium
	Monoamine oxidase inhibitors
Tranquilizers	Butyrophenones
	Phenothiazines
H_2 antagonists	Ranitidine
	Cimetidine
Hormones	Progesterone
	Estrogens
	Corticosteroids

Classification	Drugs
	GnRH agonists
	5α-Reductase inhibitors
	Cyproterone acetate
Cytotoxic agents	Cyclophosphamide
	Methotrexate
	Roferon-A
Anticholinergics	Disopyramide
	Anticonvulsants
Recreational	Ethanol
	Cocaine
	Marijuana

Abbreviation: GnRH, gonadotropin-releasing hormone.

If there is a strong association between the institution of a drug and the onset of ED, alternative medications should be considered. Otherwise, it is often practical to treat the ED without attempting multiple changes in medications, as it may be difficult to establish a causal role for a drug.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Erectile Dysfunction

A good physician-patient relationship helps unravel the possible causes of ED, many of which require discussion of personal and sometimes embarrassing topics. For this reason, a primary care provider is often ideally suited to initiate the evaluation. However, a significant percentage of men experience ED and remain undiagnosed unless specifically questioned about this issue. By far the two most common reasons for underreporting of ED are patient embarrassment and perceptions of physicians' inattention to the disorder. Once the topic is initiated by the physician, patients are more willing to discuss their potency issues. A complete medical and sexual history should be taken in an effort to assess whether the cause of ED is organic, psychogenic, or multifactorial (Fig. 390-3).

Both the patient and his sexual partner should be interviewed regarding sexual history. ED should be distinguished from other sexual problems, such as premature ejaculation. Lifestyle factors such as sexual orientation, the patient's distress from ED, performance anxiety, and details of sexual techniques should be addressed. Standardized questionnaires are available to assess ED, including the International Index of Erectile Function (IIEF) and the more easily administered Sexual Health Inventory for Men (SHIM), a validated abridged version of the IIEF.The initial evaluation of ED begins with a review of the patient's medical, surgical, sexual, and psychosocial histories. The history should note whether the patient has experienced pelvic trauma, surgery, or

radiation. In light of the increasing recognition of the relationship between lower urinary tract symptoms and ED, it is advisable to evaluate for the presence of symptoms of bladder outlet obstruction. Questions should focus on the onset of symptoms, the presence and duration of partial erections, and the progression of ED. A history of nocturnal or early morning erections is useful for distinguishing physiologic ED from psychogenic ED. Nocturnal erections occur during rapid eye movement (REM) sleep and require intact neurologic and circulatory systems. Organic causes of ED generally are characterized by a gradual and persistent change in rigidity or the inability to sustain nocturnal, coital, or self-stimulated erections. The patient should be questioned about the presence of penile curvature or pain with coitus. It is also important to address libido, as decreased sexual drive and ED are sometimes the earliest signs of endocrine abnormalities (e.g., increased prolactin, decreased testosterone levels). It is useful to ask whether the problem is confined to coitus with one partner or also involves other partners; ED not uncommonly arises in association with new or extramarital sexual relationships. Situational ED, as opposed to consistent ED, suggests psychogenic causes. Ejaculation is much less commonly affected than erection, but questions should be asked about whether ejaculation is normal, premature, delayed, or absent. Relevant risk factors should be identified, such as diabetes mellitus, coronary artery disease (CAD), and neurologic disorders. The patient's surgical history should be explored with an emphasis on bowel, bladder, prostate, and vascular procedures. A complete drug history is also important. Social changes that may precipitate ED are also crucial to the evaluation, including health worries, spousal death, divorce, relationship difficulties, and financial concerns.

Because ED commonly involves a host of endothelial cell risk factors, men with ED report higher rates of overt and silent myocardial infarction. Therefore, ED in an otherwise asymptomatic male warrants consideration of other vascular disorders, including CAD.

Men who suffer from ED are at high risk for concomitant lower urinary tract symptoms (LUTS) from BPH and vice versa. Given that some treatments of one disorder will impact the other, the clinician should consider an assessment of LUTS in any man with ED.

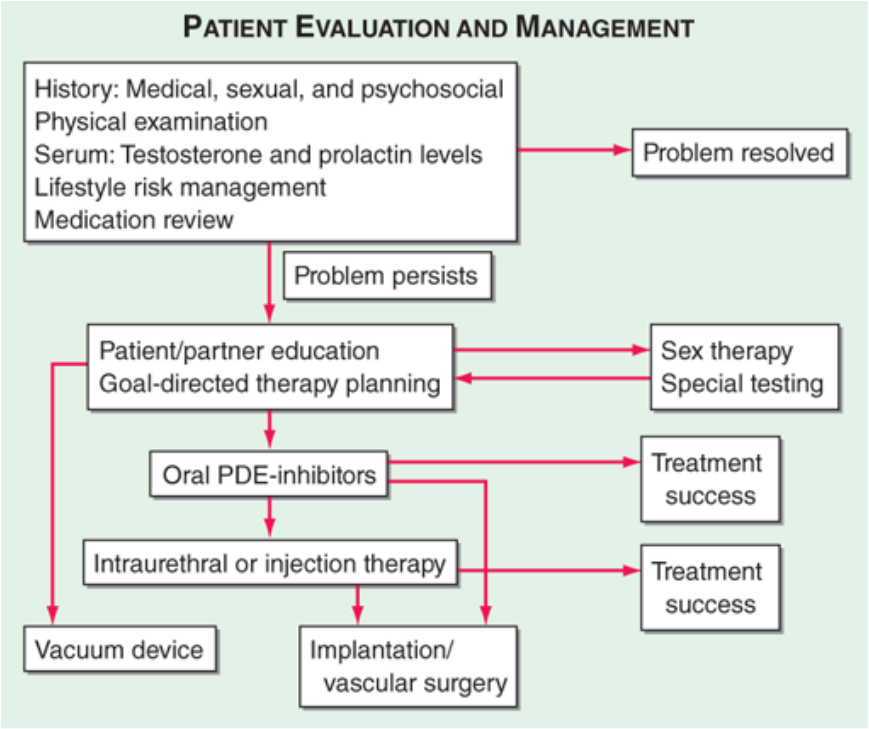
The physical examination is an essential element in the assessment of ED. Signs of hypertension as well as evidence of thyroid, hepatic, hematologic, cardiovascular, or renal diseases should be sought. An assessment should be made of the endocrine and vascular systems, the external genitalia, and the prostate gland. The penis should be palpated carefully along the corpora to detect fibrotic plaques. Reduced testicular size and loss of secondary sexual characteristics are suggestive of hypogonadism. Neurologic examination should include assessment of anal sphincter tone, investigation of the bulbocavernosus reflex, and testing for peripheral neuropathy.

Although hyperprolactinemia is uncommon, a serum prolactin level should be measured, as decreased libido and/or ED may be the presenting symptoms of a prolactinoma or another mass lesion of the sella **(Chap. 373)**. The serum testosterone level should be measured, and if it is low, gonadotropins should be measured to determine whether hypogonadism is primary (testicular) or secondary (hypothalamic-pituitary) in origin **(Chap. 384)**. If not performed recently, serum chemistries, complete blood count (CBC), and lipid profiles may be of value, as they can yield evidence of anemia, diabetes, hyperlipidemia, or other systemic diseases associated with ED. Determination of serum prostate-specific antigen (PSA) should be conducted according to recommended clinical guidelines **(Chap. 83)**.

Additional diagnostic testing is rarely necessary in the evaluation of ED. However, in selected patients, specialized testing may provide insight into pathologic mechanisms of ED and aid in the selection of treatment options. Optional specialized testing includes (1) studies of nocturnal penile tumescence and rigidity, (2) vascular testing (in-office injection of vasoactive substances, penile Doppler ultrasound, penile angiography, dynamic infusion cavernosography/cavernosometry), (3) neurologic testing (biothesiometry-graded vibratory perception, somatosensory-evoked potentials), and (4) psychological diagnostic tests. The information potentially gained from these procedures must be balanced against their invasiveness and cost.

FIGURE 390-3

Algorithm for the evaluation and management of patients with erectile dysfunction. PDE, phosphodiesterase.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Male Sexual Dysfunction

PATIENT EDUCATION

Patient and partner education is essential in the treatment of ED. In goal-directed therapy, education facilitates understanding of the disease, the results of the tests, and the selection of treatment. Discussion of treatment options helps clarify how treatment is best offered and stratify first- and second-line therapies. Patients with high-risk lifestyle issues such as obesity, smoking, alcohol abuse, and recreational drug use should be counseled on the role those factors play in the development of ED.

Therapies currently employed for the treatment of ED include oral phosphodiesterase type 5 inhibitor therapy (most commonly used), injection therapies, testosterone therapy, penile devices, and psychological therapy. In addition, limited data suggest that treatments for underlying risk factors and comorbidities—for example, weight loss, exercise, stress reduction, and smoking cessation—may improve erectile function. Decisions regarding therapy should take into account the preferences and expectations of patients and their partners.

ORAL AGENTS

Sildenafil, tadalafil, vardenafil, and avanafil are the only approved and effective oral agents for the treatment of ED. These four medications have markedly improved the management of ED because they are effective for the treatment of a broad range of causes, including psychogenic, diabetic, vasculogenic, post-radical prostatectomy (nerve-sparing procedures), and spinal cord injury. They belong to a class of medications that are selective and potent inhibitors of PDE-5, the predominant phosphodiesterase isoform found in the penis. They are administered in graduated doses and enhance erections after sexual stimulation (Fig. 390-2). The onset of action is ~30–120 min, depending on the medication used and other factors, such as recent food intake. Reduced initial doses should be considered for patients who are elderly, are taking concomitant alpha blockers, have renal insufficiency, or are taking medications that inhibit the CYP3A4 metabolic pathway in the liver (e.g., erythromycin,

cimetidine, ketoconazole, and possibly itraconazole and mibefradil), as they may increase the serum concentration of the PDE-5 inhibitors (PDE-5i) or promote hypotension.

Initially there were concerns about the cardiovascular safety of these drugs. It is known that these agents can act as mild vasodilators and warnings exist about orthostatic hypotension with concomitant use of alpha blockers. The use of PDE5i is not contraindicated in men who are also receiving alpha blockers, but they must be stabilized on this blood pressure medication prior to initiating therapy. Earlier concerns that the use of PDE5i would increase cardiovascular events have been mitigated by the results of several controlled trials showing no increase in myocardial ischemic events or overall mortality compared to the general population.

Several randomized trials have demonstrated the efficacy of this class of medications. There are no compelling data to support the superiority of one PDE-5i over another. Subtle differences between agents have variable clinical relevance **(see Table 390-2)**.

Patients may fail to respond to a PDE-5i for several reasons **(Table 390-3)**. Some patients may not tolerate PDE-5i secondary to adverse events from vasodilation in nonpenile tissues expressing PDE-5 or from the inhibition of homologous nonpenile isozymes (i.e., PDE-6 found in the retina). Abnormal vision attributed to the effects of PDE-5i on retinal PDE-6 is of short duration, reported only with sildenafil and not thought to be clinically significant. A more serious concern is the possibility that PDE-5i may cause nonarteritic anterior ischemic optic neuropathy (NAION); although data to support that association are limited, it is prudent to avoid the use of these agents in men with a prior history of nonarteritic anterior ischemic optic neuropathy.

Testosterone supplementation combined with a PDE-5i may be beneficial in improving erectile function in hypogonadal men with ED who are unresponsive to PDE-5i alone. These drugs do not affect ejaculation, orgasm, or sexual drive. Side effects associated with PDE-5i include headaches (19%), facial flushing (9%), dyspepsia (6%), and nasal congestion (4%). Approximately 7% of men using sildenafil may experience transient altered color vision (blue halo effect), and 6% of men taking tadalafil may experience loin pain. PDE-5i is contraindicated in men receiving nitrate therapy for cardiovascular disease, including agents delivered by the oral, sublingual, transnasal, and topical routes. These agents can potentiate its hypotensive effect and may result in profound shock. Likewise, amyl/butyl nitrate “poppers” may have a fatal synergistic effect on blood pressure. PDE-5i also should be avoided in patients with congestive heart failure and cardiomyopathy because of the risk of vascular collapse. Because sexual activity leads to an increase in physiologic expenditure [5–6 metabolic equivalent tasks (METs)], physicians have been advised to exercise caution in prescribing any drug for sexual activity to those with active coronary disease, heart failure, borderline hypotension, or hypovolemia and to those on complex antihypertensive regimens.

Although the various forms of PDE-5i have a common mechanism of action, there are a few differences among the four agents (Table 390-2). Tadalafil is unique in its longer half-life and avanafil appears to have the fastest onset of action. All four drugs are effective for patients with ED of all ages, severities, and etiologies. Although there are pharmacokinetic and pharmacodynamic differences among these agents, clinically relevant differences are not clear.

ANDROGEN THERAPY

Testosterone replacement is used to treat both primary and secondary causes of hypogonadism **(Chap. 384)**. Androgen supplementation in the setting of normal testosterone is rarely efficacious in the treatment of ED and is discouraged. Methods of androgen replacement include transdermal patches and gels, including cutaneous nasal and axillary gels. Parenteral administration of long-acting testosterone esters (enanthate and cypionate), long-acting subcutaneous pellets, and oral preparations (17 α-alkylated derivatives) are also available **(Chap. 384)**. Oral androgen preparations have the potential for hepatotoxicity and should be avoided.

The increased scrutiny of testosterone caused the U.S. Food and Drug Administration (FDA) to issue a warning that there is a “weak signal” that testosterone replacement therapy increases the risk of thromboembolic events and may have addictive properties. Though testosterone therapy has known risks, such as water retention in heart failure patients, and worsening sleep apnea, increasing evidence suggests that, when monitored appropriately, this therapy decreases the risk for metabolic syndrome, changes body composition by increasing lean muscle mass, and improves insulin sensitivity and average hemoglobin

A1c. This evidence, combined with the fact that hypogonadism is a known risk factor for metabolic syndrome and cardiovascular disease, has led to the conclusion that testosterone therapy for age-related hypogonadism in fact improves overall health and decreases risk of cardiovascular events. It is important to note that men with secondary hypogonadism who desire fertility should not be treated directly with testosterone, but with an alternative such as the selective estrogen-receptor modulator (SERM) clomiphene citrate, which increases gonadotropin levels, stimulating testicular T production.

Testosterone circulates in the body in two forms: free and unbound or bound to proteins such as albumin or sex hormone-binding globulin (SHBG). SHBG has a very high affinity for testosterone and, thus testosterone bound to SHBG does not bind to the androgen receptor and is not bioavailable. Bioavailable testosterone is any testosterone that is not bound to SHBG. Unfortunately, reliable assays to directly measure bioavailable testosterone or free testosterone are expensive and difficult to perform, and are thus not offered by most laboratories. However, direct measurement of SHBG is inexpensive and reliable, allowing free and bioavailable testosterone to be calculated.

Men who receive testosterone should be reevaluated after 3–6 months and at least annually thereafter for testosterone levels, erectile function, and adverse effects, which may include gynecomastia, sleep apnea, development or exacerbation of LUTS or BPH, prostate cancer, lowering of HDL, erythrocytosis, elevations of liver function tests, and reduced fertility. Periodic reevaluation should include measurement of CBC and PSA and digital rectal examination. Therapy should be discontinued in patients who do not respond within 3–6 months without an alternate explanation (e.g., elevated estradiol).

VACUUM CONSTRICTION DEVICES

Vacuum constriction devices (VCDs) are a well-established noninvasive therapy. They are a reasonable treatment alternative for select patients who cannot take sildenafil or do not desire other interventions. VCDs draw venous blood into the penis and use a constriction ring to restrict venous return and maintain tumescence. Adverse events with VCD include pain, numbness, bruising, and altered ejaculation. Additionally, many patients complain that the devices are cumbersome and that the induced erections have a nonphysiologic appearance and feel.

INTRAURETHRAL ALPROSTADIL

If a patient fails to respond to oral agents, a reasonable next choice is intraurethral or self-injection of vasoactive substances. Intraurethral prostaglandin E_1 (alprostadil), in the form of a semisolid pellet (doses of 125–1000 μg), is delivered with an applicator. Approximately 65% of men receiving intraurethral alprostadil respond with an erection when tested in the office, but only 50% achieve successful coitus at home. Intraurethral insertion is associated with a markedly reduced incidence of priapism in comparison to intracavernosal injection.

INTRACAVERNOSAL SELF-INJECTION

Injection of synthetic formulations of alprostadil is effective in 70–80% of patients with ED, but discontinuation rates are high because of the invasive nature of administration. Doses range between 1 and 40 μg. Injection therapy is contraindicated in men with a history of hypersensitivity to the drug and men at risk for priapism (hypercoagulable states, sickle cell disease). Side effects include local adverse events, prolonged erections, pain, and fibrosis with chronic use. Various combinations of alprostadil, phentolamine, and/or papaverine sometimes are used.

SURGERY

A less frequently used form of therapy for ED involves the surgical implantation of a semirigid or inflatable penile prosthesis. The choice of prosthesis is dependent on patient preference and should take into account body habitus and manual dexterity, which may affect the ability of the patient to manipulate the device. Because of the permanence of prosthetic devices, patients should be advised to first consider less invasive options for treatment. These surgical treatments are associated with a low rate of potential complications but generally are reserved for treatment of refractory ED or in men who cannot tolerate less invasive treatments. Despite their cost and the requirement for surgery, penile prostheses are associated with very high rates of patient and partner satisfaction.

SEX THERAPY

A course of sex therapy may be useful for addressing specific interpersonal factors that may affect sexual functioning. Sex therapy generally consists of in-session discussion and at-home exercises specific to the person and the relationship. Psychosexual therapy involves techniques such as sensate focus (nongenital massage), sensory awareness exercises, correction of misconceptions about sexuality, and interpersonal difficulties therapy (e.g., open communication about sexual issues, physical intimacy scheduling, and behavioral interventions). These approaches may be useful in patients who have psychogenic or social components to their ED, although data from randomized trials are scanty and inconsistent. It is preferable to include both partners in therapy if the patient is involved in an ongoing relationship.

TABLE 390-2

PDE5 Inhibitor

Drug	Onset of Action	$T_{1/2}$	Dose	Adverse Effects	Contraindications
Sildenafil	T_{max} 30–120 min Duration 4 h High-fat meal decreases absorption. Alcohol use may affect efficacy.	2–5 h	25–100 mg Starting dose 50 mg	Headache, flushing, dyspepsia, nasal congestion, altered vision	Nitrates Hypotension Cardiovascular risk factors Retinitis pigmentosa Change dose with some antiretrovirals Should be on stable dose of alpha blockers
Vardenafil	T_{max} 30–120 min Duration 4–5 h High-fat meal decreases absorption. ETOH may affect efficacy.	4.5 h	5–10 mg	Headache, flushing, rhinitis, dyspepsia	Same as sildenafil May have minor prolongation of QT interval Concomitant use of Class I anti-arrhythmic
Tadalafil	T_{max} 30–60 min Duration 12–36 h Plasma concentration not affected by food or ETOH	17.5 h	10 mg, 20 mg; 2.5 or 5 mg for daily dose	Headache, dyspepsia, backpain, nasal congestion, myalgia	Same as sildenafil
Avanafil	T_{max} 30 min Duration 2 h Plasma concentration not affected by food	3–5 h	50 mg 100 mg and 200 mg dose	Headache, flushing, nasal congestion nasopharyngitis back pain	Same as sildenafil

TABLE 390-3

Issues to Consider if Patients Report Failure of PDE-5i to Improve Erectile Dysfunction

1. A trial of medication on at least 6 different days at the maximal dose should be performed before declaring patient nonresponsive to PDE-5i use.
2. Confirm that the patient did not partake in a high-fat meal prior to taking medication.
3. Failure to include physical and psychic stimulation at the time of foreplay to induce endogenous NO.
4. Unrecognized hypogonadism

Abbreviations: NO, nitric oxide; PDE-5i, phosphodiesterase.

FEMALE SEXUAL DYSFUNCTION

Female sexual dysfunction (FSD) has traditionally included disorders of desire, arousal, pain, and muted orgasm. The associated risk factors for FSD are similar to those in males: cardiovascular disease, endocrine disorders, hypertension, neurologic disorders, and smoking **(Table 390-4)**. Women with hypertension report significantly lower sexual satisfaction (especially younger women).

TABLE 390-4

Risk Factors for Female Sexual Dysfunction

Neurologic disease: stroke, spinal cord injury, parkinsonism
Trauma, genital surgery, radiation
Endocrinopathies: diabetes, hyperprolactinemia
Liver and/or renal failure
Cardiovascular disease, especially hypertension
Psychological factors and interpersonal relationship disorders: sexual abuse, life stressors
Medications
Antiandrogens: cimetidine, spironolactone
Antidepressants, alcohol, hypnotics, sedatives
Antiestrogens or GnRH antagonists
Antihistamines, sympathomimetic amines
Antihypertensives: diuretics, calcium channel blockers
Alkylating agents
Anticholinergics

Abbreviation: GnRH, gonadotropin-releasing hormone.

EPIDEMIOLOGY

Epidemiologic data are limited, but the available estimates suggest that as many as 43% of women complain of at least one sexual problem. Despite the recent interest in organic causes of FSD, desire and arousal phase disorders (including lubrication complaints) remain the most common presenting problems when surveyed in a community-based population.

PHYSIOLOGY OF THE FEMALE SEXUAL RESPONSE

The female sexual response requires the presence of estrogens. A role for androgens is also likely but less well established. In the CNS, estrogens and androgens work synergistically to enhance sexual arousal and response. A number of studies report enhanced libido in women during preovulatory phases of the menstrual cycle, suggesting that hormones involved in the ovulatory surge (e.g., estrogens) increase desire.

Sexual motivation is heavily influenced by context, including the environment and partner factors. Once sufficient sexual desire is reached, sexual arousal is mediated by the central and autonomic nervous systems. Cerebral sympathetic outflow is thought

to increase desire, and peripheral parasympathetic activity results in clitoral vasocongestion and vaginal secretion (lubrication).

The neurotransmitters for clitoral corporal engorgement are similar to those in the male, with a prominent role for neural, smooth-muscle, and endothelial released nitric oxide (NO). A fine network of vaginal nerves and arterioles promotes a vaginal transudate. The major transmitters of this complex vaginal response are not certain, but roles for NO and vasointestinal polypeptide (VIP) are suspected. Investigators studying the normal female sexual response have challenged the long-held construct of a linear and unmitigated relationship between initial desire, arousal, vasocongestion, lubrication, and eventual orgasm. Caregivers should consider a paradigm of a positive emotional and physical outcome with one, many, or no orgasmic peak and release.

Although there are anatomic differences as well as variation in the density of vascular and neural beds in males and females, the primary effectors of sexual response are strikingly similar. Intact sensation is important for arousal. Thus, reduced levels of sexual functioning are more common in women with peripheral neuropathies (e.g., diabetes). Vaginal lubrication is a transudate of serum that results from the increased pelvic blood flow associated with arousal. Vascular insufficiency from a variety of causes may compromise adequate lubrication and result in dyspareunia. Cavernosal and arteriole smooth-muscle relaxation occurs via increased nitric oxide synthase (NOS) activity and produces engorgement in the clitoris and the surrounding vestibule. Orgasm requires an intact sympathetic outflow tract; hence, orgasmic disorders are common in female patients with spinal cord injuries.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Female Sexual Dysfunction

Many women do not volunteer information about their sexual response. Open-ended questions in a supportive atmosphere are helpful in initiating a discussion of sexual fitness in women who are reluctant to discuss such issues. Once a complaint has been voiced, a comprehensive evaluation should be performed, including a medical history, a psychosocial history, a physical examination, and limited laboratory testing.

The history should include the usual medical, surgical, obstetric, psychological, gynecologic, sexual, and social information. Past experiences, intimacy, knowledge, and partner availability should also be ascertained. Medical disorders that may affect sexual health should be delineated. They include diabetes, cardiovascular disease, gynecologic conditions, obstetric history, depression, anxiety disorders, and neurologic disease. Medications should be reviewed as they may affect arousal, libido, and orgasm. The need for counseling and recognizing life stresses should be identified. The physical examination should assess the genitalia, including the clitoris. Pelvic floor examination may identify prolapse or other disorders. Laboratory studies are needed, especially if menopausal status is uncertain. Estradiol, follicle-stimulating hormone (FSH), and luteinizing hormone (LH) are usually obtained, and dehydroepiandrosterone (DHEA) should be considered as it reflects adrenal androgen secretion. A CBC, liver function assessment, and lipid studies may be useful, if not otherwise obtained. Complicated diagnostic evaluation such as clitoral Doppler ultrasonography and biothesiometry require expensive equipment and are of uncertain utility. It is important for the patient to identify which symptoms are most distressing.

The evaluation of FSD previously occurred mainly in a psychosocial context. However, inconsistencies between diagnostic categories based only on psychosocial considerations and the emerging recognition of organic etiologies have led to a new classification of FSD. This diagnostic scheme is based on four components that are not mutually exclusive: (1) *Hypoactive sexual desire*—the persistent or recurrent lack of sexual thoughts and/or receptivity to sexual activity, which causes personal distress. Hypoactive sexual desire may result from endocrine failure or may be associated with psychological or emotional disorders, (2) *Sexual interest arousal disorder*—the persistent or recurrent inability to attain or maintain sexual excitement, which causes personal distress, (3) *Orgasmic disorder*—the persistent or recurrent loss of orgasmic potential after sufficient sexual stimulation and arousal, which causes personal distress, and (4) *Sexual pain disorder*—persistent or recurrent genital pain associated with noncoital sexual stimulation, which causes personal distress. This newer classification emphasizes “personal distress” as a

requirement for dysfunction and provides clinicians with an organized framework for evaluation before or in conjunction with more traditional counseling methods.

TREATMENT

TREATMENT

Female Sexual Dysfunction

General

An open discussion with the patient is important as couples may need to be educated about normal anatomy and physiologic responses, including the role of orgasm, in sexual encounters. Physiologic changes associated with aging and/or disease should be explained. Couples may need to be reminded that clitoral stimulation rather than coital intromission may be more beneficial.

Behavioral modification and nonpharmacologic therapies should be a first step. Patient and partner counseling may improve communication and relationship strains. Lifestyle changes involving known risk factors can be an important part of the treatment process. Emphasis on maximizing physical health and avoiding lifestyles (e.g., smoking, alcohol abuse) and medications likely to produce FSD is important (Table 390-3). The use of topical lubricants may address complaints of dyspareunia and dryness. Contributing medications such as antidepressants may need to be altered, including the use of medications with less impact on sexual function, dose reduction, medication switching, or drug holidays.

HORMONAL THERAPY

In postmenopausal women, estrogen replacement therapy may be helpful in treating vaginal atrophy, decreasing coital pain, and improving clitoral sensitivity (**Chap. 388**). Menopause and its transition represent significant risk factors for the development of vulvovaginal atrophy-related sexual dysfunction. Available vaginal estrogen preparations include conjugated equine estrogens, estradiol vaginal cream, a sustained-release intravaginal estradiol ring, or a low-dose estradiol tablet. Vaginal estrogen preparations with the lowest systemic absorption rate may be preferred in women with history of breast cancer and severe vaginal atrophy. Vaginal lubricants and moisturizers applied on a regular basis have an efficacy comparable to that of local estrogen therapy and should be offered to women wishing to avoid the use of vaginal estrogens. If a hormonal supplement is chosen, then estrogen replacement in the form of local cream is the preferred method, as it avoids systemic side effects. Androgen levels in women decline substantially before menopause. However, low levels of testosterone or DHEA are not effective predictors of a positive therapeutic outcome with androgen therapy. The widespread use of exogenous androgens is not supported by the literature except in select circumstances (premature ovarian failure or menopausal states) and in secondary arousal disorders.

Atrophic vaginitis is very common in postmenopausal women and is most commonly treated with estrogen-based treatments. However, many women are hesitant to use estrogen-based treatments due to health concerns or are unable to use them due to a history of breast cancer or endometrial cancer. Hyaluronic acid vaginal gel has been found to be efficacious in treating atrophic vaginitis.

ORAL AGENTS

Flibanserin, originally developed as an antidepressant, has been approved by the FDA as a treatment for low sexual desire in premenopausal women. Flibanserin, a postsynaptic agonist of serotonin receptor 1A and antagonist of serotonin receptor 2A, increases sexual desire and reduces resultant stress in women with HypoSexual Desire Disorder (HSDD) with few adverse effects. Flibanserin may boost sex drive in women who experience low sexual desire and who find the experience distressing. The drug should be discontinued if there is no improvement in sex drive after 8 weeks. Potentially serious side effects include low blood pressure, dizziness, and fainting, particularly if it is mixed with alcohol. Other common adverse events include dizziness, nausea, fatigue, sleepiness, and insomnia. Health care professionals and pharmacies dealing with flibanserin have to undergo a certification process, and patients need to submit a written agreement to abstain from alcohol.

The efficacy of PDE-5i in FDS has been a marked disappointment in light of the proposed role of NO-dependent physiology in the normal female sexual response. The use of PDE-5i for FSD should be discouraged pending proof that it is effective.

CLITORAL VACUUM DEVICE

In patients with arousal and orgasmic difficulties, the option of using a clitoral vacuum device may be explored. This handheld battery-operated device has a small soft plastic cup that applies a vacuum over the stimulated clitoris. This causes increased cavernosal blood flow, engorgement, and vaginal lubrication.

FURTHER READING

AUA Clinical Guideline: The management of erectile dysfunction. Available from *https://www.auanet.org/common/pdf/education/clinical-guidance/Erectile-Dysfunction.pdf*. Accessed January 12, 2017.

Bhasin S: A perspective on the evolving landscape in male reproductive medicine. J Clin Endocrinol Metab 101:827, 2016. [PubMed: 26908110]

Bhasin S et al: Sexual dysfunction in men and women with endocrine disorders. Lancet 369:597, 2007. [PubMed: 17307107]

Cappelleri JC, Rosen RC: The Sexual Health Inventory for Men (SHIM): A 5-year review of research and clinical experience. Int J Impot Res 17:307, 2005. [PubMed: 15875061]

McVary KT: Clinical practice. Erectile dysfunction. N Engl J Med 357:2472, 2007. [PubMed: 18077811]

Parish SJ, Hahn SR: Hypoactive sexual desire disorder: A review of epidemiology, biopsychology, diagnosis, and treatment. Sex Med Rev 4:103, 2016. [PubMed: 27872021]

Chapter 391: Women's Health

Andrea Dunaif

INTRODUCTION

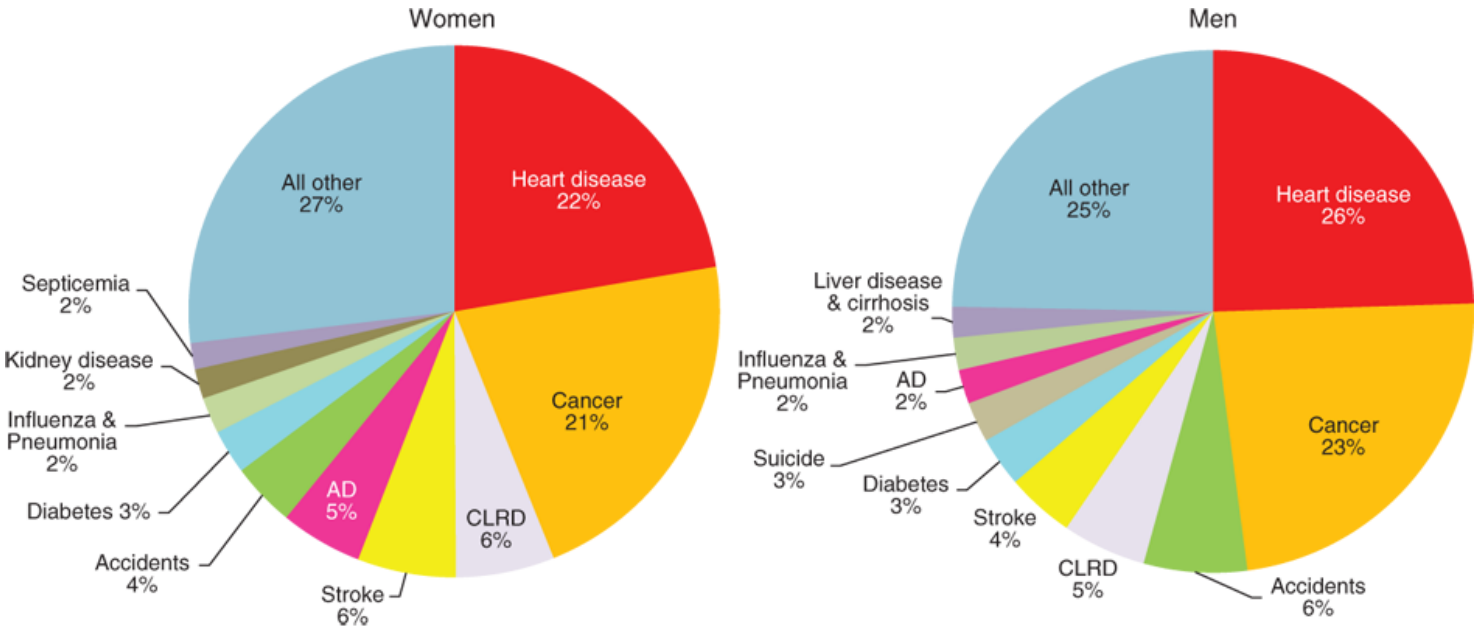
The clinical discipline of women's health is well established. Indeed, its emphasis on greater attention to patient education and medical decision-making is a paradigm for what has become known as patient-centered health care. Moreover, the recognition of sex differences in disease processes and health outcomes is an important example of precision medicine. Sex difference refers to the biologic differences conferred by sex chromosomes and hormones. In contrast, gender differences are related to psychosocial roles and cultural expectations. The study of sex differences continues to grow as a scientific discipline. In 2016, the National Institutes of Health recognized its importance by implementing the expectation that sex should be considered as a biologic variable in study designs, analyses, and reporting in not only human but also vertebrate animal research. Strong scientific justification must be provided to limit research to only one sex.

DISEASE RISK: REALITY AND PERCEPTION

The leading causes of death are the same in women and men: (1) heart disease and (2) cancer **(Fig. 391-1)**. The leading cause of cancer death, lung cancer, is the same in both sexes. Breast cancer is the second leading cause of cancer death in women, but it causes about 60% fewer deaths than does lung cancer. Men are more likely than women to die from suicide and accidents.

FIGURE 391-1

Percent distribution of 10 leading causes of death in women compared to men in the United States in 2014. In both women and men, the first and second leading causes of death are the same, heart disease and cancer, respectively. Causes of death then diverge by sex. For example, accidents are the third leading cause of death in men but the sixth leading cause of death in women. Stroke, chronic lower respiratory disease (CLRD) and Alzheimer's disease (AD) cause a larger percentage of deaths in women than in men. Suicide is among the 10 leading causes of death in men but not in women. *(Data from https://www.cdc.gov/nchs/data/nvsr/nvsr65/nvsr65_04.pdf.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Women's risk for many diseases increases at menopause. The median age of menopause in Caucasian women from industrialized countries is between 50 and 52 years, where women spend one-third of their lives in the postmenopausal period. Menopause occurs at earlier ages in

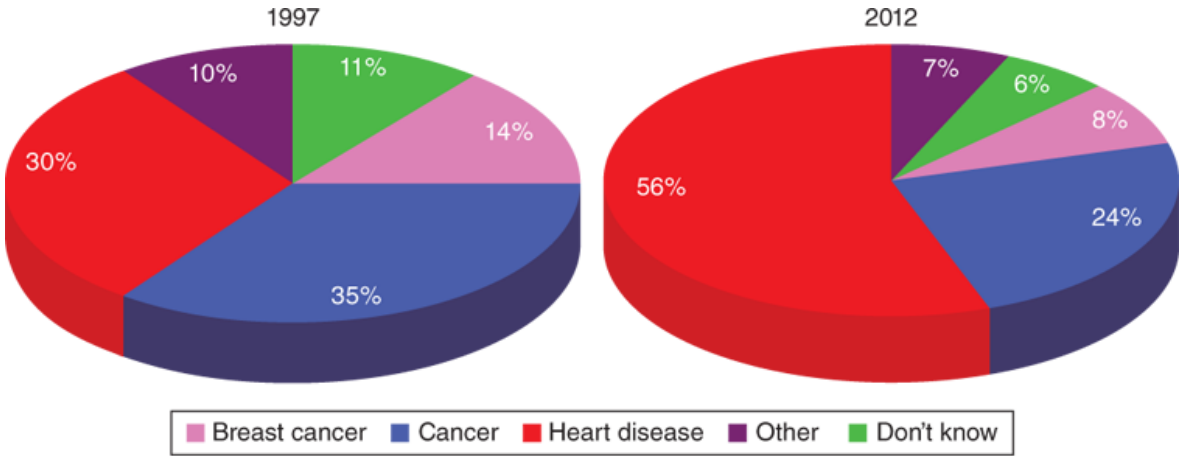
Hispanic and African-American women as well as in women of lower socioeconomic status. Estrogen levels fall abruptly at menopause, inducing a variety of physiologic and metabolic responses. Rates of cardiovascular disease (CVD) increase and bone density begins to decrease rapidly after menopause.

In the United States, women live on average 4.8 years longer than men, with a life expectancy at birth in 2014 of 81.2 years in women compared with 76.4 years in men of all races. Life expectancy was lower in African Americans of both sexes and higher in Hispanics of both sexes than their Caucasian counterparts. The difference in life expectancy between men and women has decreased since its peak of 7.8 years in 1979 but has remained unchanged at 4.8 years since 2010. Accordingly, elderly women outnumber elderly men, so that age-related conditions, such as hypertension, have a female preponderance.

Public awareness campaigns have resulted in a marked increase in the percentage of U.S. women knowing that CVD is the leading cause of death in women. In 1997, the majority of U.S. women surveyed thought that cancer (35%) rather than heart disease (30%) was the leading cause of death in women **(Fig. 391-2)**. In 2012, these perceptions were reversed, with 56% of U.S. women surveyed recognizing that heart disease rather than cancer (24%) was the leading cause of death in women (Fig. 391-2). Although awareness of heart disease has improved substantially among black and Hispanic women over this time period, these groups were 66% less likely than white women to recognize that heart disease is the leading cause of death in women.

FIGURE 391-2

Changes in perceived leading causes of death among women surveyed in 1997 compared with those surveyed in 2012. In 1997, cancer was cited as the leading cause of death in women, not heart disease. In 2012, this trend had reversed. The rate of awareness that heart disease is the leading cause of death in women was significantly higher in 2012 (56% vs 30%, $p < .001$) than in 1997. *(Data adapted from L Mosca et al: Circulation 127:1254, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Nevertheless, women aged <65 years still consider breast cancer to be their leading health risk, despite the fact that death rates from breast cancer have been falling since the 1990s. In 2011, 1 in 30.8 deaths in women was due to breast cancer, whereas 1 in 7.5 deaths was due to CVD. Although a woman's lifetime risk of developing breast cancer if she lives past 85 years is about 1 in 9, it is much more likely that she will die from CVD than from breast cancer. In other words, many elderly women have breast cancer but die from other causes. Similarly, a minority of women are aware that lung cancer is the leading cause of cancer death in women. Physicians are also less likely to recognize women's risk for CVD. Even in 2012, only 21% of U.S. women surveyed reported that their physicians had counseled them about their risk for heart disease. These misconceptions are unfortunate as they perpetuate inadequate attention to modifiable risk factors such as dyslipidemia, hypertension, and cigarette smoking.

SEX DIFFERENCES IN HEALTH AND DISEASE

ALZHEIMER'S DISEASE

(See also Chap. 423) Alzheimer's disease (AD) affects approximately twice as many women as men. Because the risk for AD increases with age, part of this sex difference is accounted for by the fact that women live longer than men. However, additional factors probably contribute to the increased risk for AD in women, including sex differences in brain size, structure, and functional organization. There is emerging evidence for sex-specific differences in gene expression, not only for genes on the X and Y chromosomes but also for some autosomal genes. Estrogens

have pleiotropic genomic and nongenomic effects on the central nervous system, including neurotrophic actions in key areas involved in cognition and memory. There are sex differences in the severity of AD with women experiencing greater deficits in cognition.

Women with AD have lower endogenous estrogen levels than do women without AD. These observations have led to the hypothesis that estrogen is neuroprotective. Some studies have suggested that estrogen administration improves cognitive function in nondemented postmenopausal women as well as in women with AD, and several observational studies have suggested that postmenopausal hormone therapy (HT) may decrease the risk of AD. The Women's Health Initiative Memory Study (WHIMS), an ancillary study in the Women's Health Initiative (WHI) in women aged ≥65 years, found significantly increased risk for both dementia and mild cognitive impairment in women receiving estrogen alone (combined continuous equine estrogen [CEE], 0.625 mg daily) or estrogen with progestin (CEE, 0.625 mg daily, and medroxyprogesterone acetate [MPA], 2.5 mg daily) compared to placebo. However, the Kronos Early Estrogen Prevention Study (KEEPS), a randomized clinical trial of early initiation of HT after menopause that compared CEE 0.45 mg daily, 50 μg of weekly transdermal estradiol (both estrogen arms included cyclic oral micronized progesterone 200 mg daily for 12 days each month), or placebo, found no adverse effects of HT on cognitive function. In summary, there is no evidence from placebo-controlled trials that HT improves cognitive function.

CVD AND STROKE

(See also Chap. 267) There are major sex differences in CVD, the leading cause of death in men and women in developed countries. However, there are also major gender differences because of perceptions by both women and their health care providers that women are at lower risk for CVD. As a result of these misconceptions, women are less likely to seek medical help when they experience symptoms of CVD. Health care providers are less likely to suspect CVD, so women receive fewer interventions for modifiable risk factors as well as fewer acute interventions than do men. Women and their health care providers are also less aware that prodromal symptoms of cardiac disease differ in women compared to men. Women are less likely than men to present with chest pain and more likely to present with fatigue, shortness of breath, indigestion/nausea, and anxiety.

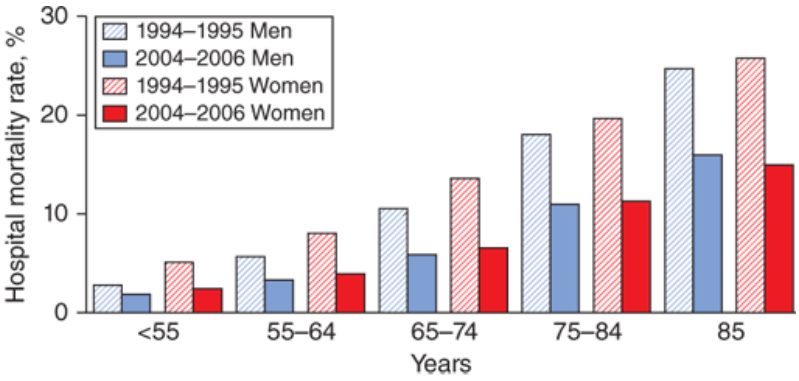
Sex steroids have major effects on the cardiovascular system and lipid metabolism. Estrogen increases high-density lipoprotein (HDL) and lowers low-density lipoprotein (LDL), whereas androgens have the opposite effect. Estrogen has direct vasodilatory effects on the vascular endothelium, enhances insulin sensitivity, and has antioxidant and anti-inflammatory properties. There is a striking increase in CVD after both natural and surgical menopause, suggesting that endogenous estrogens are cardioprotective. Women also have longer QT intervals on electrocardiograms, and this increases their susceptibility to certain arrhythmias.

CVD presents differently in women, who are usually 10–15 years older than their male counterparts and are more likely to have comorbidities such as hypertension, congestive heart failure, and diabetes mellitus (DM). In the Framingham study, angina was the most common initial symptom of CVD in women, whereas myocardial infarction (MI) was the most common initial presentation in men. Women more often have atypical symptoms such as fatigue, anxiety, nausea, indigestion, and upper back pain. Although awareness that heart disease is the leading cause of death in women has nearly doubled over the last 15 years, women remain less aware that its symptoms are often atypical, and they are less likely to contact 9-1-1 when they experience such symptoms. The recent availability of a high-specificity troponin assay with sex-specific cutoffs has increased diagnostic accuracy for MI in women but not in men.

Deaths from CVD have decreased markedly in men since 1980, whereas CVD deaths only began to decrease substantially in women beginning in 2000. Women with MI are more likely to present with cardiac arrest or cardiogenic shock, whereas men are more likely to present with ventricular tachycardia. Further, younger women with MI are more likely to die than are men of similar age. However, this mortality gap has decreased in recent years because younger women have experienced greater improvements in survival after MI than men **(Fig. 391-3)**. The improvement in survival is due largely to a reduction in comorbidities, suggesting a greater attention to modifiable risk factors in women. Nevertheless, 1 year after MI, 26% of women aged >45 years will die compared to 19% of men. Within 5 years of the first MI, 47% of women compared to 36% of men will die.

FIGURE 391-3

Hospital mortality rates in men and women for acute myocardial infarction (MI) in 1994–1995 compared with 2004–2006. Women aged <65 years had substantially greater mortality than men of similar age in 1994–1995. Mortality rates declined markedly for both sexes across all age groups in 2004–2006 compared with 1994–1995. However, there was a more striking decrease in mortality in women aged <75 years compared with men of similar age. The mortality rate reduction was largest in women <55 years (52.9%) and lowest in men of similar age (33.3%). *(Data adapted from V Vaccarino et al: Arch Intern Med 169:1767, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Physicians are less likely to suspect heart disease in women with chest pain and less likely to perform diagnostic and therapeutic cardiac procedures in women. Women are less likely to receive therapies such as angioplasty, thrombolytic therapy, coronary artery bypass grafts (CABGs), beta blockers, and aspirin. There are also sex differences in outcomes when women with CVD do receive therapeutic interventions. Women undergoing CABG surgery have more advanced disease, a higher perioperative mortality rate, less relief of angina, and less graft patency; however, 5- and 10-year survival rates are similar. Women undergoing percutaneous transluminal coronary angioplasty have lower rates of initial angiographic and clinical success than men, but they also have a lower rate of restenosis and a better long-term outcome. Women may benefit less and have more frequent serious bleeding complications from thrombolytic therapy compared with men. Factors such as older age, more comorbid conditions, smaller body size, and more severe CVD in women at the time of events or procedures account in part for the observed sex differences.

Elevated cholesterol levels, hypertension, smoking, obesity, low HDL cholesterol levels, DM, and lack of physical activity are important risk factors for CVD in both men and women. Total triglyceride levels are an independent risk factor for CVD in women but not in men. Low HDL cholesterol and DM are more important risk factors for CVD in women than in men. Smoking is an important risk factor for CVD in women—it accelerates atherosclerosis, exerts direct negative effects on cardiac function, and is associated with an earlier age of menopause. Several disorders affect women exclusively, including pregnancy-associated hypertension, preeclampsia, gestational DM, polycystic ovary syndrome, or predominantly, including rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE). Cholesterol-lowering drugs are equally effective in men and women for primary and secondary prevention of CVD. In contrast to men, randomized trials showed that aspirin was not effective in the primary prevention of CVD in women; it did significantly reduce the risk of ischemic stroke.

The sex differences in CVD prevalence, beneficial biologic effects of estrogen on the cardiovascular system, and reduced risk for CVD in observational studies led to the hypothesis that HT was cardioprotective. However, the WHI, which studied >16,000 women on CEE plus MPA or placebo and >10,000 women with hysterectomy on CEE alone or placebo, did not demonstrate a benefit of HT for the primary or secondary prevention of CVD. In addition, CEE plus MPA was associated with an increased risk for CVD, particularly in the first year of therapy, whereas CEE alone neither increased nor decreased CVD risk. Both CEE plus MPA and CEE alone were associated with an increased risk for ischemic stroke.

In the WHI, there was a suggestion of a reduction in CVD risk in women who initiated HT closer to menopause. This finding suggests that the time at which HT is initiated is critical for cardioprotection. According to this “timing” hypothesis, HT has differential effects, depending on the stage of atherosclerosis; adverse effects are seen with advanced, unstable lesions. This hypothesis was supported by data from the Danish Osteoporosis Prevention Study (DOPS), an open-label randomized trial of triphasic oral estradiol compared with no treatment in recently menopausal or perimenopausal women (a cyclic oral synthetic progestin, norethisterone acetate, was added in women who had a uterus), that found significantly reduced mortality and CVD after 10 years of HT. However, DOPS was designed to investigate HT for the primary prevention of osteoporotic bone fractures, and CVD outcomes were not prespecified endpoints.

KEEPS was designed to directly test the “timing” hypothesis. Seven hundred twenty-seven recently menopausal women aged 42–58 years (mean 52.7 years) were randomized to oral CEE (lower dose than WHI), transdermal estradiol, or placebo for 4 years; both estrogen arms included oral cyclical micronized progesterone (see above section on AD for dosing details). There were no significant beneficial or deleterious effects on the progression of atherosclerosis by computed tomography assessment of coronary artery calcification in either HT arm. Adverse events including stroke, MI, venous thromboembolism, and breast cancer were not increased in the HT arms compared with the placebo arm. There were improvements in hot flashes, night sweats, mood, sexual function, and bone density in the HT arms. This relatively

small study does not suggest that early HT administration reduces atherosclerosis. However, the study suggests that short-term HT may be safely administered for symptom relief in recently menopausal women. **HT is discussed further in Chap. 388.**

DIABETES MELLITUS

(See also Chap. 396) Women are more sensitive to insulin than men are. Despite this, the prevalence of type 2 DM is similar in men and women. There is a sex difference in the relationship between endogenous androgen levels and DM risk. Higher bioavailable testosterone levels are associated with increased risk in women, whereas lower bioavailable testosterone levels are associated with increased risk in men. Polycystic ovary syndrome, preeclampsia, pregnancy-associated hypertension, and gestational DM—common conditions in premenopausal women—are associated with a significantly increased risk for type 2 DM. Among individuals with DM, women have a greater risk for MI than do men. Women with DM have a sixfold greater risk of dying of CVD compared to women without DM.

Premenopausal women with DM lose the cardioprotective effect of female sex and have rates of CVD identical to those in males. These women have impaired endothelial function and reduced coronary vasodilatory responses, which may predispose to cardiovascular complications. Women with DM are more likely to have left ventricular hypertrophy. Women with DM receive less aggressive treatment for modifiable CVD risk factors than men with DM. In the WHI, CEE plus MPA significantly reduced the incidence of DM, whereas with CEE alone, there was only a trend toward decreased DM incidence.

HYPERTENSION

(See also Chap. 271) After age 60, hypertension is more common in U.S. women than in men, largely because of the high prevalence of hypertension in older age groups and the longer survival of women. Isolated systolic hypertension is present in 30% of women >60 years old. Sex hormones affect blood pressure. Both normotensive and hypertensive women have higher blood pressure levels during the follicular phase than during the luteal phase. In the Nurses' Health Study, the relative risk of hypertension was 1.8 in current users of oral contraceptives, but this risk is lower with the newer low-dose contraceptive preparations. HT is not associated with hypertension. Among secondary causes of hypertension, there is a female preponderance of renal artery fibromuscular dysplasia.

The benefits of treatment for hypertension have been dramatic in both women and men. A meta-analysis of the effects of hypertension treatment, the Individual Data Analysis of Antihypertensive Intervention Trial, found a reduction of risk for stroke and for major cardiovascular events in women. The effectiveness of various antihypertensive drugs appears to be comparable in women and men; however, women may experience more side effects. For example, women are more likely to develop cough with angiotensin-converting enzyme inhibitors.

AUTOIMMUNE DISORDERS

(See also Chap. 348) Most autoimmune disorders occur more commonly in women than in men; they include autoimmune thyroid and liver diseases, SLE, RA, scleroderma, multiple sclerosis (MS), and idiopathic thrombocytopenic purpura. However, there is no sex difference in the incidence of type 1 DM, and ankylosing spondylitis occurs more commonly in men. Women may be more resistant to bacterial infections than men. Sex differences in both immune responses and adverse reactions to vaccines have been reported. For example, there is a female preponderance of postvaccination arthritis.

Adaptive immune responses are more robust in women than in men; this may be explained by the stimulatory actions of estrogens and the inhibitory actions of androgens on the cellular mediators of immunity. Consistent with an important role for sex hormones, there is variation in immune responses during the menstrual cycle, and the activity of certain autoimmune disorders is altered by castration or pregnancy (e.g., RA and MS may remit during pregnancy). Nevertheless, the majority of studies show that exogenous estrogens and progestins in the form of HT or oral contraceptives do not alter autoimmune disease incidence or activity. Exposure to fetal antigens, including circulating fetal cells that persist in certain tissues, has been speculated to increase the risk of autoimmune responses. There is clearly an important genetic component to autoimmunity, as indicated by the familial clustering and HLA association of many such disorders. X chromosome genes also contribute to sex differences in immunity. Indeed, nonrandom X chromosome inactivation may be a risk factor for autoimmune diseases.

HIV INFECTION

(See also Chap. 197) Women accounted for almost 19% of the ~40,000 new HIV diagnoses in the United States in 2015. Of these newly diagnosed women, 61% were African American, 19% were Caucasian, and 15% were Hispanic. Annual HIV diagnoses declined by 20% among women from 2010 to 2014. Nevertheless, AIDS remains an important cause of death in younger women, particularly African-American women

aged 25–44 years. Heterosexual contact with an at-risk partner is the fastest-growing transmission category, and women are more susceptible to HIV infection during vaginal sex than men. This increased susceptibility is accounted for in part by an increased prevalence of sexually transmitted diseases, i.e., gonorrhea and syphilis, in women.

Some studies have suggested that hormonal contraceptives may increase the risk of HIV transmission. Progesterone has been shown to increase susceptibility to infection in nonhuman primate models of HIV. Women are also more likely to be infected by multiple variants of the virus than men. Women with HIV have more rapid decreases in their CD4 cell counts than do men. Compared with men, HIV-infected women more frequently develop candidiasis, but Kaposi's sarcoma is less common than it is in men. Women have more adverse reactions, such as lipodystrophy, dyslipidemia, and rash, with antiretroviral therapy than do men. This observation is explained in part by sex differences in the pharmacokinetics of certain antiretroviral drugs, resulting in higher plasma concentrations in women.

OBESITY

(See also Chap. 395) The prevalence of both obesity (body mass index [BMI] ≥30 kg/m^2) and abdominal obesity (waist circumference ≥88 cm in women) is higher in U.S. women than in men. Between 2005 and 2014, the prevalence of obesity (BMI ≥30 kg/m^2) and class 3 obesity (BMI ≥40 kg/m^2) increased significantly in women but not in men. In 2014, the prevalence of obesity was 40.4% and class 3 obesity 9.9% for women, and 35.0% and class 3 obesity 5.5% for men. The prevalence of abdominal obesity increased over this time period in both sexes. More than 80% of patients who undergo bariatric surgery are women. Pregnancy and menopause are risk factors for obesity.

There are major sex differences in body fat distribution. Women characteristically have gluteal and femoral or gynoid pattern of fat distribution, whereas men typically have a central or android pattern. Women have more subcutaneous fat than men. In women, endogenous androgen levels are positively associated with abdominal obesity, and androgen administration increases visceral fat. In contrast, there is an inverse relationship between endogenous androgen levels and abdominal obesity in men. Further, androgen administration decreases visceral fat in these obese men. The reasons for these sex differences in the relationship between visceral fat and androgens are unknown. Studies in humans also suggest that sex steroids play a role in modulating food intake and energy expenditure.

In men and women, abdominal obesity characterized by increased visceral fat is associated with an increased risk for CVD and DM. Obesity increases a woman's risk for certain cancers, in particular postmenopausal breast and endometrial cancer, in part because adipose tissue provides an extragonadal source of estrogen through aromatization of circulating adrenal and ovarian androgens, especially the conversion of androstenedione to estrone. Obesity increases the risk of infertility, miscarriage, and complications of pregnancy.

OSTEOPOROSIS

(See also Chap. 404) Osteoporosis is about five times more common in postmenopausal women than in age-matched men, and osteoporotic hip fractures are a major cause of morbidity in elderly women. Men accumulate more bone mass and lose bone more slowly than do women. Sex differences in bone mass are found as early as infancy. Calcium intake, vitamin D, and estrogen all play important roles in bone formation and bone loss. Particularly during adolescence, calcium intake is an important determinant of peak bone mass. Vitamin D deficiency is surprisingly common in elderly women, occurring in >40% of women living in northern latitudes. Receptors for estrogens and androgens have been identified in bone. Estrogen deficiency is associated with increased osteoclast activity and a decreased number of bone-forming units, leading to net bone loss. The aromatase enzyme, which converts androgens to estrogens, is also present in bone. Estrogen is an important determinant of bone mass in men (derived from the aromatization of androgens) as well as in women.

PHARMACOLOGY

On average, women have lower body weights, smaller organs, a higher percentage of body fat, and lower total-body water than men. There are also important sex differences in drug action and metabolism that are not accounted for by these differences in body size and composition. Sex steroids alter the binding and metabolism of a number of drugs. Further, menstrual cycle phase and pregnancy can alter drug action. Women also take more medications than men, including over-the-counter formulations and supplements. The greater use of medications combined with these biologic differences may account for the reported higher frequency of adverse drug reactions in women than in men.

Two-thirds of cases of drug-induced torsades des pointes, a rare, life-threatening ventricular arrhythmia, occur in women because they have a longer, more vulnerable QT interval. These drugs, which include certain antihistamines, antibiotics, antiarrhythmics, and antipsychotics, can prolong cardiac repolarization by blocking cardiac voltage-gated potassium channels. Women require lower doses of neuroleptics to

control schizophrenia. Women awaken from anesthesia faster than do men given the same doses of anesthetics. In 2013, the Food and Drug Administration recommended that doses of the drug zolpidem be lowered for women because of slower drug clearance than in men.

PSYCHOLOGICAL DISORDERS

(See also Chap. 444) Depression, anxiety, and affective and eating disorders (bulimia and anorexia nervosa) are more common in women than in men. Epidemiologic studies from both developed and developing nations consistently find major depression to be twice as common in women as in men, with the sex difference becoming evident in early adolescence. Depression occurs in 10% of women during pregnancy and in 10–15% of women during the postpartum period. There is a high likelihood of recurrence of postpartum depression with subsequent pregnancies. The incidence of major depression diminishes after the age of 45 years and does not increase with the onset of menopause. Depression in women appears to have a worse prognosis than does depression in men; episodes last longer, and there is a lower rate of spontaneous remission. Schizophrenia and bipolar disorders occur at equal rates in men and women, although there may be sex differences in symptoms.

Both biologic and social factors account for the greater prevalence of depressive disorders in women. Men have higher levels of the neurotransmitter serotonin. Sex steroids also affect mood, and fluctuations during the menstrual cycle have been linked to symptoms of premenstrual syndrome. Sex hormones differentially affect the hypothalamic-pituitary-adrenal responses to stress. Testosterone appears to blunt cortisol responses to corticotropin-releasing hormone. Both low and high levels of estrogen can activate the hypothalamic-pituitary-adrenal axis.

SLEEP DISORDERS

(See also Chap. 27) There are striking sex differences in sleep and its disorders. During sleep, women have an increased amount of slow-wave activity, differences in timing of delta activity, and an increase in the number of sleep spindles. Testosterone modulates neural control of breathing and upper airway mechanics. Men have a higher prevalence of sleep apnea. Testosterone administration to hypogonadal men as well as to women increases apneic episodes during sleep. Women with the hyperandrogenic disorder polycystic ovary syndrome have an increased prevalence of obstructive sleep apnea, and apneic episodes are positively correlated with their circulating testosterone levels. In contrast, progesterone accelerates breathing, and in the past, progestins were used for treatment of sleep apnea.

SUBSTANCE ABUSE AND TOBACCO

(See also Chaps. 445 and 448) Substance abuse is more common in men than in women. However, one-third of Americans who suffer from alcoholism are women. Women alcoholics are less likely to be diagnosed than men. A greater proportion of men than women seek help for alcohol and drug abuse. Men are more likely to go to an alcohol or drug treatment facility, whereas women tend to approach a primary care physician or mental health professional for help under the guise of a psychosocial problem. Late-life alcoholism is more common in women than in men. On average, alcoholic women drink less than alcoholic men but exhibit the same degree of impairment. Blood alcohol levels are higher in women than in men after drinking equivalent amounts of alcohol, adjusted for body weight. This greater bioavailability of alcohol in women is due to both the smaller volume of distribution and the slower gastric metabolism of alcohol secondary to lower activity of gastric alcohol dehydrogenase than is the case in men. In addition, alcoholic women are more likely to abuse tranquilizers, sedatives, and amphetamines. Women alcoholics have a higher mortality rate than do nonalcoholic women and alcoholic men. Women also appear to develop alcoholic liver disease and other alcohol-related diseases with shorter drinking histories and lower levels of alcohol consumption. Alcohol abuse also poses special risks to a woman, adversely affecting fertility and the health of the baby (fetal alcohol syndrome). Even moderate alcohol use increases the risk of breast cancer, hypertension, and stroke in women.

More men than women smoke tobacco, but this sex difference continues to decrease. Women have a much larger burden of smoking-related disease. Smoking markedly increases the risk of CVD in premenopausal women and is also associated with a decrease in the age of menopause. Women who smoke are more likely to develop chronic obstructive pulmonary disease and lung cancer than men and at lower levels of tobacco exposure. Postmenopausal women who smoke have lower bone density than women who never smoked. Smoking during pregnancy increases the risk of preterm deliveries and low birth weight infants.

VIOLENCE AGAINST WOMEN

More than one in three women in the United States have experienced rape, physical violence, and/or stalking by an intimate partner. Adult women are much more likely to be raped by a spouse, ex-spouse, or acquaintance than by a stranger. Domestic or intimate partner violence

is a leading cause of death among young women. Domestic violence may be an unrecognized feature of certain clinical presentations, such as chronic abdominal pain, headaches, and eating disorders, in addition to more obvious manifestations such as trauma. Intimate partner violence is an important risk factor for depression, substance abuse, and suicide in women. Screening instruments can accurately identify women experiencing intimate partner violence. Such screening by health care providers is acceptable to women in settings ensuring adequate privacy and safety.

SUMMARY

Women's health is now a mature discipline, and the importance of sex differences in biologic processes is well recognized. Nevertheless, ongoing misperceptions about disease risk, not only among women but also among their health care providers, result in inadequate attention to modifiable risk factors. Research into the fundamental mechanisms of sex differences will provide important biologic insights. Further, those insights will have an impact on both women's and men's health.

FURTHER READING

Carter CL et al: Sex and gender differences in Alzheimer's disease: Recommendations for future research. J Womens Health 21:1018, 2012.

Clayton JA, Collins FS: Policy: NIH to balance sex in cell and animal studies. Nature 509:282, 2014. [PubMed: 24834516]

Flegal KM et al: Trends in obesity among adults in the United States, 2005 to 2014. JAMA 315:2284, 2016. [PubMed: 27272580]

!!How sex/gender influence health & disease (A-Z) *http://orwh.od.nih.gov.ezproxy1.library.usyd.edu.au/resources/sexgenderhealth/*.

Mcsweeney JC et al: Preventing and experiencing ischemic heart disease as a woman: State of the science: A scientific statement from the American Heart Association. Circulation 133:1302, 2016. [PubMed: 26927362]

Ritz SA et al: First steps for integrating sex and gender considerations into basic experimental biomedical research. FASEB J 28:4, 2014. [PubMed: 24056086]

Siegel RL et al: Cancer statistics, 2017. CA Cancer J Clin 67:7, 2017. [PubMed: 28055103]

Chapter 392: Men's Health

Shalender Bhasin; Shehzad Basaria

INTRODUCTION

The emergence of men's health as a distinct discipline within internal medicine is founded on the wide consensus that men and women differ across their lifespan in their susceptibility to disease, in the clinical manifestations of the disease, and in their response to treatment. Furthermore, men and women weigh the health consequences of illness differently and have different motivation for seeking care. Men and women experience different types of disparities in access to healthcare services, and in the manner in which health care is delivered to them because of a complex array of socioeconomic and cultural factors. Attitudinal and institutional barriers to accessing care, fear, and embarrassment due to the perception that it is not manly to seek medical help, and reticence on the part of patients and physicians in discussing issues related to sexuality, drug use, and aging have heightened the need for programs tailored to address the specific health needs of men.

The sex differences in disease prevalence, susceptibility, and clinical manifestations of the disease were discussed in Chap. 391 (Women's Health) and will not be discussed here. It is notable that the two leading causes of death in both men and women—heart disease and cancer—are the same. However, men have higher prevalence of neurodevelopmental and degenerative disorders, substance abuse disorders, including the use of performance enhancing drugs and alcohol dependence, diabetes, and cardiovascular disease, and women have higher prevalence of autoimmune disorders, depression, rheumatologic disorders, and osteoporosis, The men are substantially more likely to die from accidents, suicides, and homicides than women. Among men, 15–34 years of age, unintentional injuries, homicides, and suicides account for over three-fourths of all deaths. Among men, 35–64 years of age, heart disease, cancer, and unintentional injuries are the leading causes of death. Among men ≥65 years of age, heart disease, cancer, lower respiratory tract infections, and stroke are the major causes of death.

The biologic bases of sex differences in disease susceptibility, progression, and manifestation remain incompletely understood, and are likely multifactorial. Undoubtedly, sex-specific differences in the genetic architecture and circulating sex hormones influence disease phenotype; additionally, epigenetic effects of sex hormones during fetal life, early childhood, and during pubertal development may epigenetically imprint sexual and nonsexual behaviors, body composition, and disease susceptibility. The circulating and tissue concentrations of sex hormones differ substantially in men and women, and these hormonal differences may affect gene expression in cells of males and females in all parts of the body. The presence of only one X chromosome in men renders them more susceptible to X-linked disorders than women. Due to the X inactivation of one randomly chosen X chromosome, women's bodies contain two epigenetically different cell populations. The genes that do not undergo X inactivation exhibit dosage differences between male and female cells. Expression of the Y chromosome genes in men may affect the function of somatic cells containing the Y chromosome. The differences in the

imprinting of maternally and paternally derived genes may also contribute to sex differences in the expression of disease. Reproductive load and physiologic changes during pregnancy, including profound hormonal and metabolic shifts, and microchimerism (transfer of cells from the mother to the fetus and from the fetus to the mother) may affect disease susceptibility and disease severity in women. Sociocultural norms of child-rearing practices, societal expectations of gender roles, and the long-term economic impact of these practices and gender roles influence health behaviors and disease risk. Furthermore, the trajectories of age-related changes in sex hormones during the reproductive and postreproductive years vary substantially between men and women, and influence the sex-specific patterns of the temporal evolution of age-related conditions such as osteoporosis, breast cancer, and autoimmune disease.

In a reflection of the growing attention on issues related to men's health, men's health clinics have mushroomed all over the country. Although the major threats to men's health have not changed—heart disease, cancer, and unintentional injury continue to dominate the list of major medical causes of morbidity and mortality in men—the men who attend men's health clinics do so largely for sexual, reproductive, and urologic health concerns involving common conditions, such as androgen deficiency syndromes, age-related decline in testosterone levels, sexual dysfunction, muscle dysmorphia and anabolic-androgenic steroid (AAS) use, lower urinary tract symptoms (LUTS), and medical complications of prostate cancer therapy, which are the subjects of this chapter. Additionally, we are witnessing the emergence of new categories of body image disorders in men that had not been recognized until the 1980s, such as the body dysmorphia syndrome and the use of performance enhancing drugs to increase muscularity and lean appearance. Although menopause has been the subject of intense investigation for more than five decades, these issues that are specific to men's health are just beginning to gain attention that they deserve because of their high prevalence and impact on overall health, well-being, and quality of life.

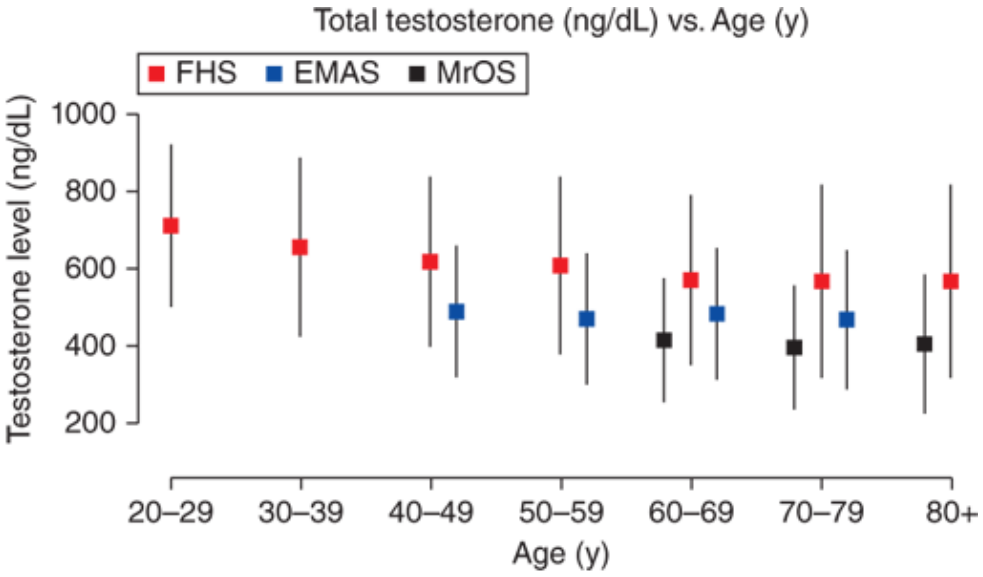
AGING-RELATED CHANGES IN MALE REPRODUCTIVE FUNCTION

A number of cross-sectional and longitudinal studies (e.g., the Baltimore Longitudinal Study of Aging, the Framingham Heart Study [FHS], the Massachusetts Male Aging Study, and the European Male Aging Study [EMAS]) have established that testosterone concentrations decrease with advancing age. This age-related decline starts in the third decade of life and progresses slowly **(Fig. 392-1)**; the rate of decline in testosterone concentrations is greater in obese men, in men with chronic illness and in those taking medications than in healthy older men. Because sex-hormone binding globulin (SHBG) concentrations are higher in older men than in younger men, free or bioavailable testosterone concentrations decline with aging to a greater extent than total testosterone concentrations. The age-related decline in testosterone is due to defects at all levels of the hypothalamic-pituitary-testicular (HPT) axis: pulsatile gonadotropin-releasing hormone (GnRH) secretion is attenuated, luteinizing hormone (LH) response to GnRH is reduced, and testicular response to LH is impaired. However, the gradual rise of LH with aging suggests that testis dysfunction is the main cause of declining androgen levels. The term *andropause* has been used to denote age-related decline in testosterone concentrations; this term is a misnomer because there is no discrete time when testosterone concentrations decline abruptly.

FIGURE 392-1

Age-related decline in total testosterone levels. Total testosterone levels measured using liquid chromatography tandem mass spectrometry in men of the Framingham Heart Study (FHS), the European Male Aging Study (EMAS),

and the Osteoporotic Fractures in Men Study (MrOS). *(Reproduced with permission from S Bhasin et al: J Clin Endocrinol Metab 96:2430, 2011.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In epidemiologic surveys, low total and bioavailable testosterone concentrations have been associated with decreased appendicular skeletal muscle mass and strength, decreased self-reported physical function, higher visceral fat mass, insulin resistance, and increased risk of coronary artery disease and mortality **(Table 392-1)**. An analysis of signs and symptoms in older men in the EMAS revealed a syndromic association of sexual symptoms with total testosterone levels <320 ng/dL and free testosterone levels <64 pg/mL in community-dwelling older men.

TABLE 392-1

Association of Testosterone Levels with Outcomes in Older Men

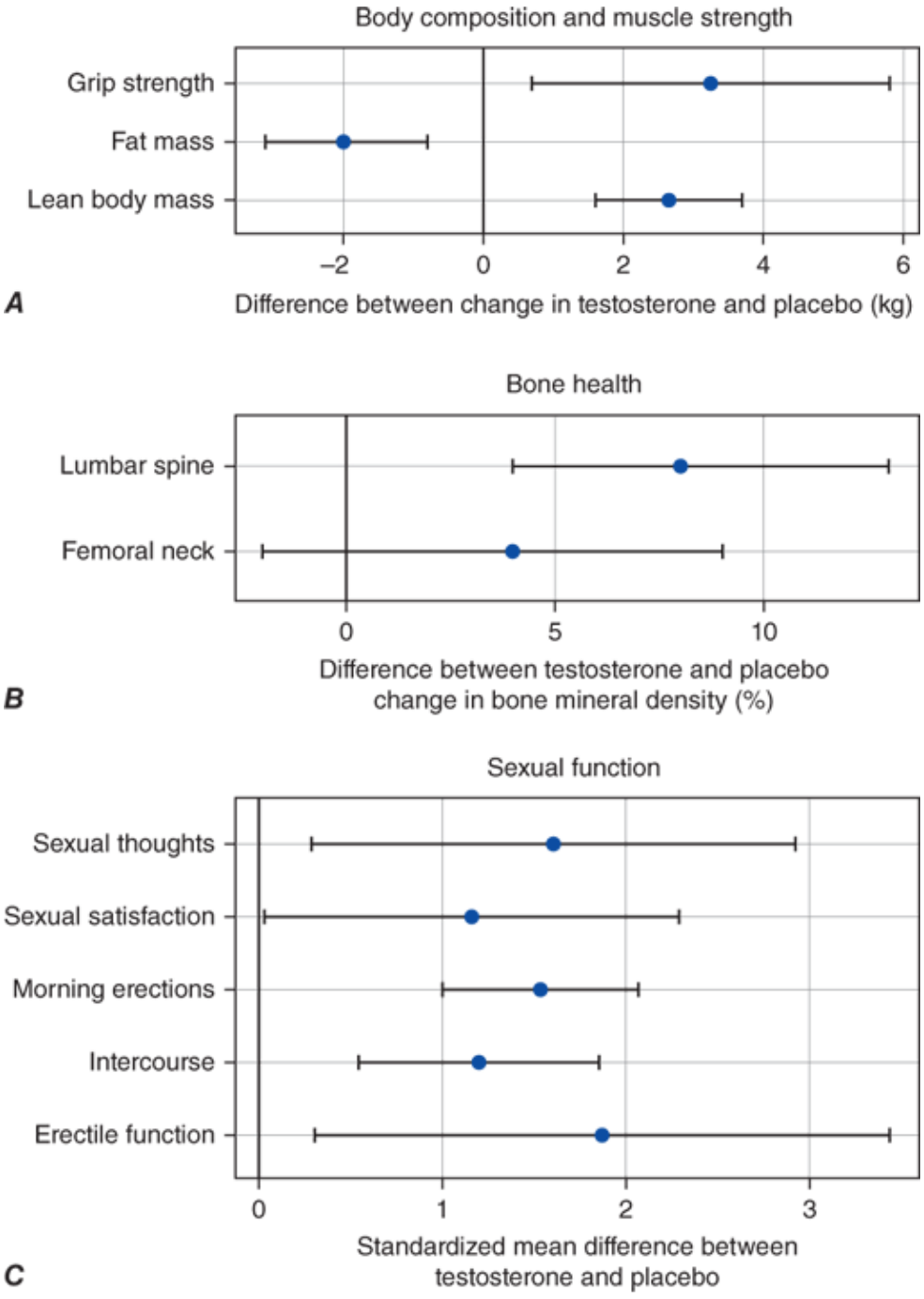
1. Positively associated with:
   - Muscle mass and muscle strength
   - Physical function
   - Sexual desire
   - Bone mineral density, bone geometry, and volumetric bone mineral density
2. Negatively associated with:
   - Coronary artery disease
   - Visceral fat
   - Diabetes mellitus
   - Metabolic syndrome
   - Mortality
   - Falls and fracture risk
   - Frailty
3. Not associated with:
   - Lower urinary tract symptoms
   - Erectile dysfunction
   - Dementia
   - Major depression

In systematic reviews of randomized controlled trials, testosterone therapy of healthy older men with low or low-normal testosterone levels was associated with greater increments in lean body mass, grip strength, and self-reported physical function than that associated with placebo (Fig. 392-2). Testosterone therapy also induced greater improvement in vertebral but not femoral bone mineral density (BMD). Testosterone therapy of older men with sexual dysfunction and unequivocally low testosterone levels improves libido, but testosterone effects on erectile function and response to selective phosphodiesterase inhibitors have been inconsistent. Testosterone therapy has not been shown to improve depression scores, fracture risk, cognitive function, response to phosphodiesterase inhibitors, or clinical outcomes in older men. Furthermore, neither the long-term risks nor clinical benefits of testosterone therapy in older men have been demonstrated in adequately powered trials. While there is no evidence that testosterone causes prostate cancer, there is concern that testosterone therapy might cause subclinical prostate cancers to grow. Testosterone therapy is associated with increased risk of detection of prostate events (Fig. 392-3).

FIGURE 392-2

The effects of testosterone therapy on body composition, muscle strength, bone mineral density (BMD), and sexual function in intervention trials. The point estimates and the associated 95% confidence intervals are shown. ***A.*** The effects of testosterone therapy on lean body mass, grip strength, and fat mass in a meta-analysis of randomized trials. *(Data derived from S Bhasin et al: Nat Clin Pract Endocrinol Metab 2:146, 2006.)* ***B.*** The effects of testosterone therapy on lumbar and femoral BMD in a meta-analysis of randomized trials. *(Data derived from a*

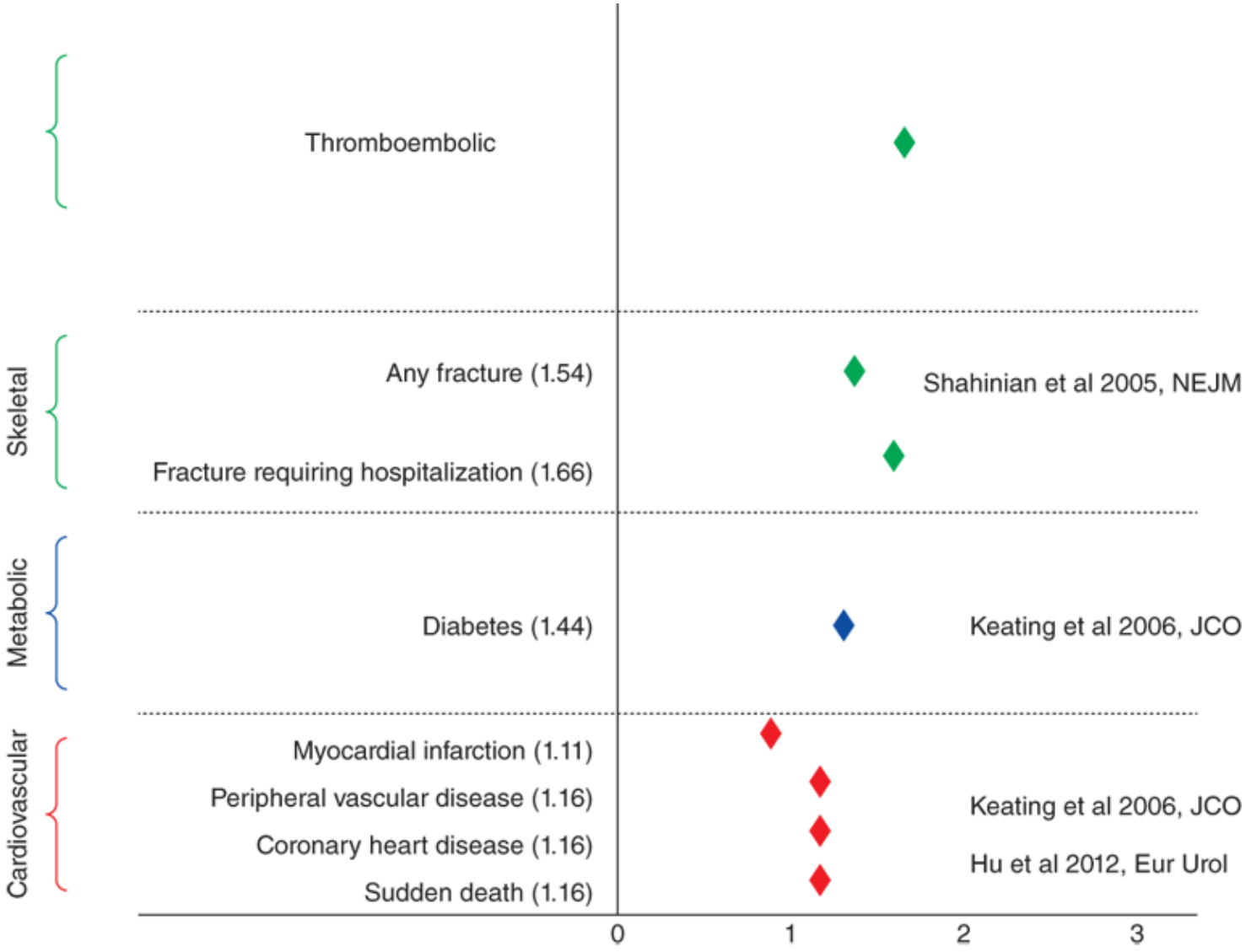
meta-analysis by MJ Tracz et al: J Clin Endocrinol Metab 91:2011, 2006.) ***C.*** The effects of testosterone therapy on measures of sexual function in men with baseline testosterone <10 nmol/L (290 ng/dL). *(Data derived from a meta-analysis by AM Isidori et al: Clin Endocrinol (Oxf) 63:381, 2005.) (Reproduced with permission from M Spitzer et al: Nat Rev Endocrinol 9:414, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 392-3

Adverse cardiometabolic and skeletal effects of androgen deprivation therapy (ADT) in men receiving ADT for prostate cancer. Administration of ADT has been associated with increased risk of thromboembolic events, fractures, and diabetes. Some, but not all, studies have reported increased risk of cardiovascular events in men receiving ADT. *(Data on relative risk were derived from VB Shahinian et al: N Engl J Med 352:154, 2005; NL Keating et al: J Clin Oncol 24:4448, 2006; and JC Hu et al: Eur Urol 61:1119, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

One randomized testosterone trial in older men with mobility limitation and high burden of chronic conditions such as diabetes, heart disease, hypertension, and hyperlipidemia, reported a greater number of cardiovascular events in men randomized to the testosterone arm of the study than in those randomized assigned to the placebo arm. Since then, two large retrospective analyses of patient databases have reported higher frequency of cardiovascular events, including myocardial infarction in older men with preexisting heart disease. A meta-analysis of randomized testosterone trials in older men found increased risk of cardiovascular events in men assigned to testosterone arms of the trials (Fig. 392-3).

Population screening of all older men for low testosterone levels is not recommended; testing should be restricted to men who have symptoms or physical features attributable to androgen deficiency. Testosterone therapy is not recommended for all older men with low testosterone levels. In older men with significant symptoms of androgen deficiency who have consistently low testosterone levels, testosterone therapy may be considered on an individualized basis and should be instituted after careful discussion of the risks and benefits.

Testicular morphology, semen production, and fertility are maintained up to a very old age in men. Although concern has been expressed about age-related increases in germ cell mutations and impairment of DNA repair

mechanisms, there is no clear evidence that the frequency of chromosomal aneuploidy is increased in the sperm of older men. However, the incidence of autosomal dominant diseases, such as achondroplasia, polyposis coli, Marfan syndrome, and Apert's syndrome, increases in the offspring of men who are advanced in age, consistent with transmission of sporadic missense mutations. Advanced paternal age may be associated with increased rates of de novo mutations, which may contribute to an increased risk of neurodevelopmental diseases such as schizophrenia and autism. The somatic mutations in male germ cells that enhance the proliferation of germ cells could lead to within-testis expansion of mutant clonal lines, thus favoring the propagation of germ cells carrying these pathogenic mutations, and increasing the risk of mutations in the offspring of older fathers (the "selfish spermatogonial selection" hypothesis).

Sexual Dysfunction

Various forms of sexual dysfunction are a major motivating factor for men seeking care at men's health clinics. The landmark descriptions of the human sexual response cycle by Masters and Johnson demonstrating that men and women display predictable physiologic responses after sexual stimulation provided the basis for rational classification of human sexual disorders. Accordingly, sexual disorders have been classified into four categories depending on phase of sexual response cycle in which the abnormality exists:

1. Hypoactive sexual desire disorder
2. Erectile dysfunction
3. Ejaculatory and orgasmic disorders
4. Disorders of pain

Classification of the patient's disorder into these categories is important as the etiologic factors, diagnostic tests, and the therapeutic strategies vary for each class of sexual disorder. Historically, the classification and nomenclature for sexual disorders were based on Diagnostic and Statistical Manual (DSM), based on the erroneous belief that sexual disorders in men are largely psychogenic in their origin. However, the recognition of erectile dysfunction as a manifestation of systemic disease and the availability of easy-to-use oral selective phosphodiesterase-5 (PDE5) inhibitors have placed sexual disorders in men within the purview of the primary care provider. These disorders have been discussed in **Chap. 390** (Sexual Dysfunction).

MUSCLE DYSMORPHIA SYNDROME IN MEN—A FORM OF BODY IMAGE DISORDER

Muscle dysmorphia is a form of body image disorder characterized by a pathological preoccupation with muscularity and leanness. The men with muscle dysmorphia express a strong desire to be more muscular and lean. These men describe shame and embarrassment about their body size and shape and often report aversive symptoms such as dissatisfaction with appearance, preoccupation with bodybuilding and muscularity, and functional impairment. Patients with muscle dysmorphia also report higher rates of mood and anxiety disorders, and obsessive and compulsive behaviors than individuals with no history of muscle dysmorphia. These men often experience impairment of social and occupational functioning.

The patients with muscle dysmorphia syndrome—nearly all men—are almost always engaged in weightlifting and body building and are more likely to use performance enhancing drugs, especially AASs than men in the general population or even weightlifters without body dysmorphia. The muscle dysmorphia disorder renders men to an increased risk of disease due to the combined interactive effects of the intensity of physical exercise, the use of performance enhancing drugs, and other lifestyle factors associated with weightlifting and the use of performance enhancing drugs. These patients are also at increased risk of functioning poorly in their occupation and social life than men without this disorder. No randomized trials of any treatment modalities have been conducted; anecdotally, behavioral and cognitive therapies have been tried with varying degrees of success.

AAS Abuse by Athletes and Recreational Bodybuilders

The illicit use of AASs to enhance athletic performance first surfaced in the 1950s among powerlifters and spread rapidly to other sports, professional as well as high school athletes, and recreational bodybuilders. In the early 1980s, the use of AAS spreads beyond the athletic community into the general population. As many as 3 million Americans—most of them men—have likely used these compounds. Most AAS users are not athletes, but rather recreational weightlifters, who use these drugs to look lean and more muscular.

The most commonly used AASs include testosterone esters, nandrolone, stanozolol, methandienone, and methenolol. AAS users generally use increasing doses of multiple steroids in a practice known as stacking.

The adverse effects of long-term AAS abuse remain poorly understood. Most of the information about the adverse effects of AAS has emerged from case reports, uncontrolled studies, or from clinical trials that used replacement doses of testosterone **(Table 392-2)**. The adverse event data from clinical trials using physiologic replacement doses of testosterone have been extrapolated unjustifiably to AAS users who may administer 10–100 times the replacement doses of testosterone over many years and to support the claim that AAS use is safe. A substantial fraction of AAS users also use other drugs that are perceived to be muscle-building or performance-enhancing, such as growth hormone; erythropoiesis stimulating agents; insulin; stimulants such as amphetamine, clenbuterol, cocaine, ephedrine, and thyroxine; and drugs perceived to reduce adverse effects such as human chorionic gonadotropin (hCG), aromatase inhibitors, or estrogen antagonists. The men who abuse AAS are more likely to engage in other high-risk behaviors than nonusers. The adverse events associated with AAS use may be due to AAS themselves, concomitant use of other drugs, high-risk behaviors, and host characteristics that may render these individuals more susceptible to AAS use or to other high-risk behaviors.

TABLE 392-2

Potential Adverse Effects Associated with the Use of Anabolic-Androgenic Steroids (AASs)

Organ System	Effect
Cardiovascular	Dyslipidemia Atherosclerotic disease Sudden death Myocardial fibrosis, cardiomyopathy Cardiac conduction abnormalities Hypertension
Neuroendocrine	HPT axis suppression Hypogonadism after AAS withdrawal Gynecomastia
Females	Virilizing effects
Neuropsychiatric	Major mood disorders (mania, hypomania, depression) Aggression, violence AAS dependence Neuronal apoptosis Cognitive deficits
Hematologic	Polycythemia Hypercoagulability and thrombosis
Hepatic	Inflammatory and cholestatic effects Peliosis hepatis (rare) Neoplasms (rare)
Musculoskeletal	Premature epiphyseal closure (in adolescents) Tendon rupture
Kidney	Renal failure secondary to rhabdomyolysis Focal segmental glomerulosclerosis
Dermatologic	Acne Striae

Abbreviation: HPT axis, hypothalamic-pituitary-testicular axis.

Source: Modified with permission from HG Pope Jr et al: Adverse health consequences of performance-enhancing drugs: an endocrine society scientific statement. *Endocr Rev* 35:341, 2014.

The high rates of mortality and morbidities observed in AAS users are alarming. One Finnish study reported 4.6 times the risk of death among elite power lifters than in age-matched men from the general population. The causes of death among power lifters included suicides, myocardial infarction, hepatic coma, and non-Hodgkin's lymphoma. A retrospective review of patient records in Sweden also reported higher standardized mortality ratios for AAS users than for nonusers. Studies indicate that 32% of deaths among AAS users were suicidal, 26% homicidal, and 35% accidental. The median age of death among AAS users—24 years—is even lower than that for heroin or amphetamine users.

Numerous reports of cardiac death among young AAS users raise concerns about the adverse cardiovascular effects of AAS. High doses of AAS may induce proatherogenic dyslipidemia, increase thrombosis risk via effects on clotting factors and platelets, induce vasospasm through their effects on vascular nitric oxide, and induce myocardial hypertrophy and fibrosis.

Replacement doses of testosterone, when administered parenterally, are associated with only a small decrease in high-density lipoprotein (HDL) cholesterol and little or no effect on total cholesterol, low-density lipoprotein (LDL) cholesterol, and triglyceride levels. In contrast, supraphysiologic doses of testosterone and orally administered, 17-α-alkylated, nonaromatizable AAS are associated with marked reductions in HDL cholesterol and increases in LDL cholesterol.

Recent studies of AAS users using tissue Doppler and strain imaging and magnetic resonance imaging have reported diastolic and systolic dysfunction, including significantly lower early and late diastolic tissue velocities, reduced E/A ratio, and reduced peak systolic strain in AAS users than in nonusers. Power athletes using AAS often have short QT intervals but increased QT dispersion, which may predispose them to ventricular arrhythmias. Long-term AAS use may be associated with myocardial hypertrophy and fibrosis. Myocardial tissue of power lifters using AAS has been shown to be infiltrated with fibrous tissue and fat droplets. The finding of androgen receptors on myocardial cells suggests that AAS might be directly toxic to myocardial cells.

Long-term AAS use suppresses LH and follicle-stimulating hormone (FSH) secretion and inhibits endogenous testosterone production and spermatogenesis. Men who have used AAS for more than a few months experience marked suppression of the HPT axis after stopping AAS that may be associated with sexual dysfunction, fatigue, infertility, and depressive symptoms. In some AAS users, HPT suppression may last more than a year, and in a few individuals, complete recovery may not occur. The symptoms of androgen deficiency during AAS withdrawal may cause some men to revert back to using AAS, leading to continued use and AAS dependence. As many as 30% of AAS users develop a syndrome of AAS dependence, characterized by long-term AAS use despite adverse medical and psychiatric effects.

Supraphysiologic doses of testosterone may also impair insulin sensitivity. Orally administered androgens have been associated with insulin resistance and diabetes.

Unsafe injection practices, high-risk behaviors, and increase rates of incarceration render AAS users at increased risk of HIV, and hepatitis B and C. In one survey, nearly one in 10 gay men had injected AAS or other substances, and AAS users were more likely to report high-risk unprotected anal sex than other men.

Some AAS users develop hypomanic and manic symptoms during AAS exposure (irritability, aggressiveness, reckless behavior, and occasional psychotic symptoms, sometimes associated with violence) and major depression (sometimes associated with suicidality) during AAS withdrawal. Users may also develop other forms of illicit drug use, which may be potentiated or exacerbated by AAS.

Elevated liver enzymes, cholestatic jaundice, hepatic neoplasms, and peliosis hepatis have been reported with oral, 17-α-alkylated AAS. AAS use may cause muscle hypertrophy without compensatory adaptations in tendons, ligaments, and joints, thus increasing the risk of tendon and joint injuries. AAS use is associated with acne, baldness, as well as increased body hair.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Detection of AAS Use

AAS users generally mistrust physicians and seek medical help infrequently; when they do seek medical help, it is often for the treatment of AAS withdrawal syndrome, infertility, gynecomastia, or other medical or psychiatric complications of AAS use. The suspicion of AAS use should be raised by the increased hemoglobin and hematocrit levels, suppressed luteinizing hormone and follicle-stimulating hormone and testosterone levels, low high-density lipoproteins cholesterol, and low testicular volume and sperm density in a person who looks highly muscular (Table 392-3). A combination of these findings along with self-report of their use by the patient—which usually can be elicited by a tactful interview—are often sufficient to establish a diagnosis in clinical practice.

Accredited laboratories use gas chromatography and mass spectrometry or liquid chromatography and mass spectrometry to detect AAS abuse. In recent years, the availability of high-resolution mass spectrometry and tandem mass spectrometry has further improved the sensitivity of detecting AAS abuse. Illicit testosterone use is detected generally by the application of the measurement of urinary testosterone to epitestosterone ratio and further confirmed by the use of the ^{13}C:^{12}C ratio in testosterone by the use of isotope ratio combustion mass spectrometry. Exogenous testosterone administration increases urinary testosterone glucuronide excretion and consequently the testosterone to epitestosterone ratio. Ratios >4 suggest exogenous testosterone use but can also reflect genetic variation. Genetic variations in the uridine diphospho-glucuronyl transferase 2B17 (*UGT2B17*), the major enzyme for testosterone glucuronidation, affect testosterone to epitestosterone ratio. Synthetic testosterone has a lower ^{13}C:^{12}C ratio than endogenously produced testosterone and these differences in the ^{13}C:^{12}C ratio can be detected by isotope ratio combustion mass spectrometry, which is used to confirm exogenous testosterone use in individuals with a high testosterone to epitestosterone ratio.

TABLE 392-3

Detection of the Use of Anabolic-Androgenic Steroids

Clinical indicators that should raise suspicion of anabolic-androgenic steroid (AAS) use
Very muscular phenotype
Reduced testicular volume (<15 mL)
Laboratory indicators
Suppressed LH and FSH levels
Increased hematocrit
Detection of AASs
LC-MS/MS analysis of urine
Detection of exogenous testosterone use
Urinary testosterone-to-epitestosterone ratio
Isotope ratio mass spectrometry analysis to detect differences in the ^{13}C:^{12}C ratio in exogenous and endogenous testosterone

Abbreviations: FSH, follicle-stimulating hormone; LC-MS/MS, liquid chromatography and tandem mass spectrometry; LH, luteinizing hormone.

TREATMENT

TREATMENT

Integrated Management of Patients with AAS Use

The nonathlete weightlifters who abuse AAS frequently do not seek medical treatment and often mistrust physicians. They also do not view these drugs and the associated lifestyle as deleterious to their health. In turn, many internists erroneously view AAS abuse as largely a problem of cheating in competitive sports, while, in fact, most AAS users are not athletes at all. Also, physicians often have a poor understanding of the factors motivating the use of these performance-enhancing drugs, the long-term health effects of AAS, and the associated psychopathologies which may affect treatment choices.

In addition to treating the underlying body dysmorphia disorder which motivates the use of these drugs, the treatment should be directed at the symptoms or the condition for which the patient seeks therapy, such as infertility, sexual dysfunction, gynecomastia, or depressive symptoms. Accordingly, therapy may include some combination of the cognitive and behavioral therapy for the muscle dysmorphia syndrome, antidepressant therapy for depression, selective PDE5 inhibitors for erectile dysfunction, or the use of selective estrogen receptor modulators or aromatase inhibitors to reactivate HPT axis or hCG to restore testosterone levels.

As discussed above, AASs suppress the male hypothalamic-pituitary-gonadal axis and men with long-term AAS use may experience symptoms of profound androgen deficiency such as sexual dysfunction, fatigue, and depressive symptoms during AAS withdrawal. Some of these patients may resume the use of AAS or start using other drugs to

combat the distressing withdrawal symptoms. There are no randomized trials of any therapies for AAS withdrawal. Case reports and clinical experience suggest that administration of selective estrogen receptor modulators, CYP19 aromatase inhibitors, or hCG may restore circulating testosterone levels. Clomiphene citrate, a partial estrogen receptor agonist, administered in a dose of 25–50 mg on alternate days can increase LH and FSH levels and restore testosterone levels in a vast majority of men with AAS withdrawal syndrome. However, the recovery of sexual function during clomiphene administration is variable in spite of improvements in testosterone levels. Anecdotally, other aromatase inhibitors such as anastrozole have also been used. hCG, administered by intramuscular injections of 750–1500 international units three times each week, can raise testosterone levels into the normal range. Some patients may not respond to either clomiphene or hCG therapy raising the possibility of irreversible long-term toxic effects of AAS on Leydig cell function.

Adjunctive cognitive and behavioral therapy or antidepressants to treat depression inadequately responsive to endocrine therapies alone may be needed. Emerging human and animal evidence suggests AAS and opioids likely promote dependence via common mechanisms. The opioid antagonist naltrexone blocks AAS dependence in animals. Therefore, treatments for human opioid dependence might also benefit AAS-dependence. Many patients who abuse AAS suffer from body-image disorder such as “muscle dysmorphia” and require psychiatric treatment for this underlying disorder.

LUTS IN MEN

LUTS in men include storage symptoms (urgency, daytime as well as nighttime frequency, and urgency incontinence), voiding disturbances (slow or intermittent stream, difficulty in initiating micturition, straining to void, pain or discomfort during the passage of urine, and terminal dribbling), or postmicturition symptoms (a sense of incomplete voiding after passing urine, and postmicturition dribble). The overactive bladder syndrome refers to urgency with or without urgency incontinence, usually with urinary frequency and nocturia, and is often due to detrusor muscle overactivity. A presumptive diagnosis of benign prostatic hyperplasia should be made only in men with LUTS, who have demonstrable evidence of prostate enlargement and obstruction based on the size of the prostate. LUTS have historically been attributed to benign prostatic hyperplasia although it has become apparent that the pathophysiologic mechanisms of LUTS are complex and multifactorial and may include structural or functional abnormalities of the bladder, bladder neck, prostate, distal sphincter mechanism, and urethra, as well as abnormalities in the neural control of the lower urinary tract. Diuretics, antihistamines, antidepressants, and other medications that have anticholinergic properties can cause or exacerbate LUTS in older men. The intensity of LUTS tends to fluctuate over time.

LUTS is highly prevalent in older men, affecting nearly 50% of >65 and 70% of men >80. The LUTS adversely affects quality of life because of its impact on sleep, ability to perform activities of daily living, and depressive symptoms. LUTS is often associated with erectile dysfunction.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Lower Urinary Tract Symptoms

Medical evaluation should include assessment of potential causes of symptoms; medications including herbal and over-the-counter products that might contribute symptoms; the symptom severity and bother using an International Prostate Symptom Score, and in some patients a frequency-volume chart. The impact of LUTS on sleep, activities of daily living, and quality of life should be evaluated. Evaluation should also include digital prostate examination, neurological examination focused on perineum and lower extremities, urinalysis, fasting blood glucose, electrolytes, creatinine, and prostate-specific antigen (PSA). Urodynamic studies are not required in most patients, but are recommended when invasive surgical therapies are being considered. A urological referral may be appropriate if the patient has hydronephrosis, renal insufficiency, recurrent urinary tract infections, hematuria, or history of acute urinary retention.

TREATMENT

TREATMENT

Patients with LUTS

Considerations of the severity of symptoms; the impact of symptoms on sleep, activities of daily living, and quality of life; the natural history of the disease; and potential adverse effects of the intervention should guide the decision of whether to intervene or not. In men with mild to moderately severe LUTS, the symptoms typically progress slowly over many years, and may remain stable or even improve in some men. The men who have mild symptoms can usually be reassured and followed. Several simple steps such as reducing caffeine and alcohol intake, especially late in the day, taking the diuretic medication early in the day, avoiding excessive water intake close to bedtime, double voiding to ensure complete emptying of the bladder may be helpful in reducing the severity of symptoms. Men with mild to moderate bothersome LUTS can be treated effectively using α-adrenergic antagonists, steroid 5α-reductase inhibitors, PDE5 inhibitors, or anticholinergic agents alone or in combination. Selective α-adrenergic antagonists are typically the first line of therapy; their side effects may include hypotension, dizziness, nasal congestion, headache, and floppy iris syndrome. In men with probable benign prostate obstruction with gland enlargement and LUTS, therapy with steroid 5α-reductase inhibitors, finasteride, or dutasteride, for one or more years improves urinary symptoms and flow rate and reduces prostatic volume. Long-term treatment with 5α-reductase inhibitors can reduce the risk of acute urinary retention and need for prostate surgery. Combined administration of steroid 5α-reductase inhibitor and α(1)-adrenergic blocker can rapidly improve urinary symptoms and reduce the relative risk of acute urinary retention and surgery. PDE5 inhibitors when administered chronically alone or in combination with α-adrenergic blockers are effective in improving LUTS and erectile dysfunction (ED) through their effects on nitric oxide—cyclic guanosine monophosphate (cGMP) in the bladder, urethra, and prostate. PDE5 inhibitors do not improve uroflow parameters, and their hypotensive effect may be potentiated by α(1)-adrenergic blockers. Anticholinergic drugs are used for the treatment of overactive bladder in men with prominent urgency symptoms and no evidence of elevated postvoid residual urine. Containment products, such as pads, can help improve social life in men who have severe storage symptoms, including incontinence. Surgery is indicated when medical therapy fails, symptoms progress in spite of medical therapy, or the patient develops acute urinary retention, hydronephrosis, renal insufficiency, or recurrent urinary tract infections, or if the patient has postvoid residual urine volume >25% of the urinary bladder volume.

MEDICAL COMPLICATIONS OF PROSTATE CANCER THERAPY

Prostate cancer is the most common malignancy in American men, accounting for 19% of all diagnosed cancers and ~8% of all cancer deaths; its incidence is on the rise, partly due to increased screening with PSA. In 2017, ~161,360 new cases of prostate cancer were diagnosed in the United States, and there were 26,730 deaths related to prostate cancer. The majority of these men have low-grade, organ-confined prostate cancer and excellent prospects of long-term survival. Substantial improvement in survival in men with prostate cancer has focused attention on the high prevalence of sexual dysfunction, physical dysfunction, and low vitality in the men, which are important contributors to poor quality of life among the patients treated for prostate cancer. The pathophysiology of these symptoms after radical prostatectomy is multifactorial, but denervation and androgen deficiency are important contributors to these symptoms.

Androgen deficiency is common in men with prostate cancer. Testosterone levels decline with age and men with prostate cancer are at risk of having low testosterone levels simply by virtue of their age. However, total and free testosterone levels are even lower in men with prostate cancer, who have undergone prostatectomy, when compared to noncancer age-matched controls. This age-related androgen deficiency in men with prostate cancer is associated with fatigue, sexual dysfunction, mobility limitation, and decreased physical function. Even with bilateral nerve-sparing procedure, >50% of men develop sexual dysfunction after surgery. Although there is some recovery of sexual function with passage of time, 40–50% of men undergoing radical prostatectomy find their sexual performance to be a moderate-to-large problem 18 months after surgery. Sexual problems are a source of psychosocial distress in men with localized prostate cancer. The men with locally advanced or metastatic prostate cancer who undergo androgen deprivation therapy (ADT) encounter even more distressing symptoms because of the profound androgen deficiency. In addition to fatigue and sexual dysfunction, and hot flushes, these men are at increased risk for diabetes, metabolic syndrome, coronary heart disease, and frailty.

Testosterone Therapy in Men with History of Prostate Cancer

A history of prostate cancer has historically been considered a contraindication for testosterone therapy. This guidance is based on observations that testosterone promotes the growth of metastatic prostate cancer. Metastatic prostate cancer generally regresses after orchidectomy and ADT. Androgen receptor signaling plays a central role in maintaining growth of normal prostate and prostate cancer. PSA levels are lower in hypogonadal men and increase after testosterone therapy. Prostate volume is lower in hypogonadal men and increases after testosterone therapy to levels seen in age-matched controls.

However, the role of testosterone in prostate cancer is complex. Epidemiological studies and their meta-analyses have not revealed a consistent relationship between serum testosterone and prostate cancer. Others have reported that low testosterone levels are associated with high-grade cancers. In a landmark randomized trial, testosterone therapy of older men with low testosterone did not affect intraprostatic androgen levels or the expression of androgen-dependent prostatic genes. The suppression of circulating testosterone levels by a GnRH antagonist also does not affect intraprostatic androgen concentrations. Open label trials and retrospective analyses of testosterone therapy in men with prostate cancer, who have undergone radical prostatectomy and have undetectable PSA levels after radical prostatectomy, have found very low rates of PSA recurrence. Even in men with high-grade prostatic intraepithelial neoplasia (HGPIN)—a group at high risk of developing prostate cancer—testosterone therapy for 1 year did not increase PSA or rates of prostate cancer.

A majority of men diagnosed with prostate cancer today has localized disease that can be potentially cured by radical prostatectomy. The men with organ-confined prostate cancer (pT2,N0,M0), Gleason score <6, are at a very low risk of disease recurrence after radical prostatectomy with 0.5% biochemical recurrence rate and 0.2% local recurrence rate >10–15 years. Similarly, preoperative PSA <10 ng/mL is associated with lower risk of disease recurrence than PSA >10 ng/mL. After radical prostatectomy, in the absence of residual cancer, PSA becomes undetectable within a month. An undetectable PSA after radical prostatectomy is a good indicator of biochemical recurrence-free survival at 5 years. Therefore, men with organ-confined prostate cancer (pT2), Gleason score <6, and a preoperative PSA of <10 ng/mL, who have had undetectable PSA levels (<0.1 ng/mL) for >2 years after radical prostatectomy, have very low risk of disease recurrence (<0.5% at 10 years) and may be considered for testosterone therapy on an individualized basis. If testosterone therapy is instituted, it should be associated with careful monitoring of PSA levels and in consultation with a urologist.

MEDICAL COMPLICATIONS OF ADT

In patients with prostate cancer and distant metastases, ADT improves survival. In patients with locally advanced disease, ADT in combination with external beam radiation or as an adjuvant therapy (postprostatectomy and pelvic lymphadenectomy) also has been shown to improve survival. However, ADT is being increasingly used as primary therapy in men with localized disease and in men encountering biochemical recurrence without clear evidence of survival advantage. The overall use of ADT in men with prostate cancer has increased in the past two decades and its use in men with localized disease and biochemical recurrence accounts for a substantial fraction of this increase. Since most men with prostate cancer die of conditions other than their primary malignancy, recognition and management of these adverse effects is paramount.

Profound hypogonadism resulting from ADT is associated with sexual dysfunction, vasomotor symptoms, gynecomastia, decreased muscle mass and strength, frailty, increased fat mass, anemia, fatigue, bone loss, loss of body hair, depressive symptoms, and reduced quality of life. Diabetes and cardiovascular disease have recently been added to the list of these complications (Fig. 392-3). Treatment with GnRH agonists in men with prostate cancer is associated with rapid induction of insulin resistance, hyperinsulinemia, and a significant increase in the risk of incident diabetes. Metabolic syndrome is prevalent in >50% of men undergoing long-term ADT when compared to age-matched men with prostate cancer not undergoing ADT (22%) and their age-matched eugonadal counterparts (20%). Some but not all studies have reported an increased risk of cardiovascular events, death due to cardiovascular events, and peripheral vascular disease in men undergoing ADT. Some reports suggest that men receiving ADT are at an increased risk of thromboembolic events and cognitive dysfunction. The rates of acute kidney injury are higher in men currently receiving ADT than in men not receiving ADT; the increased risk appears to be particularly associated with the use of combined regiments of a GnRH agonist plus and an antiandrogen. ADT also is associated with substantially increased risk of osteoporosis and bone fractures.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Men Receiving ADT

The benefits of ADT in treating nonmetastatic prostate cancer should be carefully weighed against the risks of ADT-induced adverse events **(Table 392-4)**. If ADT is medically indicated, consider whether intermittent ADT is a feasible option. Men being considered for ADT should undergo assessment of cardiovascular, diabetes, and fracture risk; this assessment may include measurement of blood glucose, plasma lipids, and BMD by dual energy x-ray absorptiometry. Institute measures to prevent bone loss, including physical activity, adequate calcium and vitamin D intake, and pharmacological therapy in men with a previous minimal trauma fracture and those with 10-year risk of a major osteoporotic fracture >20%, unless contraindicated. Bisphosphonates and denosumab have been shown to reduce fracture risk in men undergoing ADT. Men with prostate cancer who are receiving ADT should be monitored for weight gain and diabetes. Encourage lifestyle intervention, including physical activity and exercise, and attention to weight, blood pressure, lipid profile, blood glucose, and smoking cessation, to reduce the risk of cardiometabolic complications. In randomized trials, medroxyprogesterone, cyproterone acetate, and a serotonin uptake inhibitor, venlafaxine, have been shown to be more efficacious than placebo in alleviating hot flushes. The side effects of these medications—increased appetite and weight gain with medroxyprogesterone, gynecomastia with estrogenic compounds, and dry mouth with venlafaxine—should be weighed against their relative efficacy. Acupuncture, soy products, vitamin E, herbal medicines, and transdermal estradiol have been used empirically for the treatment of vasomotor symptoms without clear evidence of efficacy. Gynecomastia can be prevented by local radiation therapy or the use of an antiestrogen or an aromatase inhibitor; these therapies are effective in alleviating pain and tenderness, but are less effective in reducing established gynecomastia. For long-standing gynecomastia that persists after cessation of ADT and is bothersome, mammoplasty is an effective treatment option.

TABLE 392-4

Checklist for Men Undergoing Androgen Deprivation Therapy (ADT)

1. Weigh the risks and benefits of ADT and whether intermittent ADT is a feasible and safe option.
2. Perform a baseline assessment including fasting glucose, plasma lipids, blood pressure, bone mineral density, and FRAX® score.
3. Optimize calcium and vitamin D intake, encourage structured physical activity and exercise, and consider pharmacologic therapy in men with a previous minimal trauma fracture and those with a 10-year risk of a major osteoporotic fracture >20%, unless contraindicated.
4. Monitor body weight, fasting glucose, plasma lipids, blood pressure, and bone mineral density, and encourage smoking cessation and physical activity.
5. In men who are receiving ADT and who experience bothersome hot flushes, as indicated by sleep disturbance or interference with work or activities of daily living, consider initial therapy with venlafaxine. If in effective, add medroxyprogesterone acetate.
6. In men who experience painful breast enlargement, consider therapy with an estrogen receptor antagonist, such as tamoxifen.

PREVENTION OF SEXUALLY TRANSMITTED DISEASES

Adolescent boys and young men 15–24 years; men who have sex with men, or have multiple sex partners, or have unprotected sex without condom, or have sex with sex workers; men who use illicit drugs; men who have history of previous sexually transmitted infection (STI); and transgender men are at increased risk for STIs. STIs increase the risk of oropharyngeal and anogenital cancers, liver disease, pelvic pain, infertility, inadvertent transmission of infection to others, and emergency department visits, and are a preventable cause of excess morbidity and mortality. HIV, hepatitis B and C infections, and syphilis can have additional disease-specific complications. The prevention and treatment of STIs are discussed in **Chap. 131**. Additionally, the Centers for Disease Control (CDC) and U.S. Preventive Health Services Task Force (USPHSTF) have published guidelines on the prevention, treatment, and pre- and postexposure prophylaxis of STIs. The approach to the prevention of STIs includes a structured risk assessment; counseling about safe sex practices including condom use; immunization of individuals at risk; diagnosis and treatment of infected individuals whether or not they are symptomatic; detection and treatment of sexual partners; and targeted sex education of adolescents and young men who are at high risk for STIs. The USPHSTF recommends screening for HIV in all men, 15–65 years, regardless of risk, and for hepatitis B virus and syphilis in men at increased risk. Because more than half of STIs occur in persons, 15–24 years, the USPHSTF also recommends behavioral counseling for all sexually active adolescents and adult men at increased risk of STIs to encourage condom use and other protective behaviors, including consideration of abstinence, reducing the number of sex partners, and avoidance of unsafe sex practices. Consistent and correct condom use is the most important method of preventing STIs. Effective immunizations are available against hepatitis B, human papillomavirus (HPV), and *Neisseria meningitides.* The CDC's Advisory Committee on Immunization Practices (ACIP) recommends universal hepatitis B immunization for all unvaccinated adults presenting to an STI clinic, all HIV-infected adults, and health workers. Although ACIP recommends HPV vaccination in males 9–21 years, and in men 9–26 if they have sex with men or have an immunecompromising condition, recent data suggest that the prevalence of HPV and its complications continues to increase until middle age and some experts recommend extending the age limit for HPV vaccination. Meningococcal vaccination is indicated for men who have sex with men from an area of outbreak and for all HIV-infected men.

Because men seeking care in men's health clinics often do so for sexual and urogenital problems, these visits offer opportunities for counseling, screening, and treatment of STIs, and institution of immunization and other preventive measures for STIs.

FURTHER READING

Abrams P et al: Evaluation and treatment of lower urinary tract symptoms in older men. J Urol 189:S93, 2013. [PubMed: 23234640]

Basaria S: Cardiovascular disease associated with androgen deprivation therapy: Time to give it due respect. J Clin Oncol 33:1232, 2015. [PubMed: 25753444]

Bhasin S et al: Testosterone therapy in men with androgen deficiency syndromes: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 95:2536–59, 2010. [PubMed: 20525905]

Centers for Disease Control and Prevention. Sexually Transmitted Diseases Treatment Guidelines. MMWR 64:51, 2015.

Choi PY et al: Muscle dysmorphia: A new syndrome in weightlifters. Br J Sports Med 36:375, 2002. [PubMed: 12351338]

Jones C et al: Management of lower urinary tract symptoms in men: summary of NICE guidance. BMJ 340:c2354, 2010. [PubMed: 20484350]

López AM et al: Fracture risk in patients with prostate cancer on androgen deprivation therapy. Osteoporos Int 16:707, 2005. [PubMed: 15714259]

Mortality Tables at *http://www.cdc.gov/nchs/deaths.htm* (HHS, CDC, NCHS).

Pope HG Jr et al: Adverse health consequences of performance-enhancing drugs: An endocrine society scientific statement. Endocr Rev 35:341, 2014. [PubMed: 24423981]

Shahani S et al: Androgen deprivation therapy in prostate cancer and metabolic risk for atherosclerosis. J Clin Endocrinol Metab 93:2042, 2008. [PubMed: 18349064]

Siegel RL et al: Cancer statistics. CA Cancer J Clin 67:7, 2017. [PubMed: 28055103]

Smith MR et al: Denosumab in men receiving androgen deprivation therapy for prostate cancer. N Engl J Med 361:745, 2009. [PubMed: 19671656]

Spitzer M et al: Risks and benefits of testosterone therapy in older men. Nat Rev Endocrinol 9:414, 2013. [PubMed: 23591366]

United States Preventive Health Services Task Force. Final recommendation statement sexually transmitted infections: behavioral counseling. *https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/sexually-transmitted-infections-behavioral-counseling1.* Accessed June 21, 2017.

Chapter 393: Lesbian, Gay, Bisexual, and Transgender (LGBT) Health

Baligh R. Yehia; Harvey J. Makadon

UNDERSTANDING LGBT HEALTH DISPARITIES

Numerous studies highlight health disparities involving the care of lesbian, gay, bisexual, and transgender (LGBT) people. Lesbian and bisexual women are less likely to receive recommended preventive screenings such as breast, cervical, and colorectal cancer screenings. Among men who have sex with men, rates of human papillomavirus–associated anal cancers are 17 times higher than those of heterosexual men. In addition, gay and bisexual men accounted for 67% of all new HIV diagnoses in the United States in 2014, and they disproportionately contract sexually transmitted infections. In 2014, men who have sex with men accounted for 83% of primary and secondary syphilis infections in the United States where the sex of the sexual partner was known. Transgender individuals have a higher prevalence of HIV infection and suicide compared with other groups; 41% of transgender adults report ever attempting suicide compared with 1.6% of the general population.

Research has found that LGBT individuals are more likely to experience depression, anxiety, and alcohol and drug use than their counterparts. Most concerning are the rates of suicide attempts and ideation among the LGBT community, particularly youth. LGB youth are more than twice as likely to attempt suicide than their heterosexual peers, and approximately 30% of LGB students report having attempted suicide over a 12-month period. In addition, U.S. studies indicate that substance abuse is twice as common in LGBT youth compared with their counterparts. These findings are mirrored among LGBT adults: the prevalence of substance abuse disorders is 20–30% compared with approximately 9% in the general population.

These disparities are compounded by structural barriers to healthcare, including decreased access to medical care, lack of awareness to the unique health needs of LGBT individuals, and stigma and discrimination toward the LGBT community. Many LGB individuals perceive the healthcare setting and providers as threatening, which may lead to avoiding needed medical care or withholding important medical information. A large U.S. survey identified that 8% of LGB and 27% of transgender individuals were refused needed healthcare, and almost 11% of LGB and 21% of transgender people reported being subjected to harsh or abusive language by healthcare professionals. Apart from healthcare settings, more than two-thirds of LGB people report discrimination in their personal lives, and 90% of transgender individuals report harassment, mistreatment, or discrimination at work. Chronic exposure to high levels of stress from real or

anticipated discrimination, referred to as “minority stress,” may be an important factor contributing to the poor health outcomes experienced by LGBT populations.

While some research on LGBT health has been conducted, the Institute of Medicine has called for more study to better understand the needs and experiences of LGBT individuals. Moreover, many LGBT individuals experience health disparities across their life cycle (e.g., LGBT youth are at greater risk of suicide and homelessness, while elderly LGBT individuals face barriers to health because of isolation and fewer family supports), necessitating a longitudinal approach to examining LGBT health issues. We have limited data on the health of LGBT individuals outside the United States and Europe. However, studies demonstrate that problems are greatest where people cannot be open about their sexual orientation and gender identity. Encouraging greater LGBT acceptance and access to healthcare will be critical to improving outcomes and experiences for LGBT communities.

CREATING POSITIVE HEALTH EXPERIENCES FOR LGBT PATIENTS

Understanding Gender Identify and Sexual Orientation

Addressing health disparities and creating positive healthcare experiences requires an understanding of the diversity of cultural expression and lives of LGBT persons. Foremost, providers must be able to distinguish *gender identity* from *sexual orientation*. Gender identity is a person’s internal sense of their gender. It should not be confused with sex assigned at birth, which is based on anatomy and biology. Gender identity expands beyond the binary male and female, and includes persons who think of their gender as containing elements of both or neither. Many individuals who do not identify with the gender that correlates with their sex assigned at birth often use the terms *transgender* or *trans-male* or *trans-female* to identify themselves. Sexual orientation refers to how one thinks of their physical or emotional attraction to others. Sexual orientation has three dimensions: attraction, behavior, and identity. *Attraction* refers to one’s desire to be with someone, regardless of one’s behavior or stated identity. For example, a woman may be attracted to another woman, but this attraction may never be acted upon and may not form part of her sexual identity. *Behavior* refers to a person’s sexual and romantic partners. Although sexual identity often aligns with behavior, some individuals who identify as heterosexual may have same-gender partners and some individuals who identify as lesbian or gay may have different-gender partners. Lastly, *identity* refers to how a person defines their own sexuality. Common terms for sexual identity include *gay*, *lesbian*, *bisexual*, *straight*, *heterosexual*, *homosexual*, and *asexual* (Table 393-1). As individuals go through the process of understanding their sexuality and self-identity over time, they may change how they define their sexual identity.

TABLE 393-1

Common LGBT Terminology and Definitions

Term	Definition
Agender	Identifying as having no gender.
Asexual	Experiencing little or no sexual attraction to others.
Assigned sex at birth	The sex (male or female) assigned to a child at birth, most often based on the child's external anatomy. Also referred to as birth sex, natal sex, biological sex, or sex.
Bisexual	A sexual orientation that describes a person who is emotionally and sexually attracted to people of their own gender and people of other genders.
Cisgender	A person whose gender identity and assigned sex at birth correspond (i.e., a person who is not transgender).
Gay	A sexual orientation that describes a person who is emotionally and sexually attracted to people of their own gender. It can be used regardless of gender identity, but it is more commonly used to describe men.
Gender dysphoria	Distress experienced by some individuals whose gender identity does not correspond with their assigned sex at birth. Manifests as clinically significant distress or impairment in social, occupational, or other important areas of functioning.
Gender expression	The way a person acts, dresses, speaks, and behaves (i.e., feminine, masculine, androgynous). Gender expression does not necessarily correspond to assigned sex at birth or gender identity.
Gender identity	A person's internal sense of being a man/male, woman/female, both, neither, or another gender.
Gender nonconforming	Expressing a gender that differs from a given society's norms for males and females.
Heterosexual	A sexual orientation that describes women who are emotionally and sexually attracted to men, and men who are emotionally and sexually attracted to women.

Term	Definition
Intersex (disorders of sexual development)	A group of rare conditions where the reproductive organs and genitals do not develop as expected.
Lesbian	A sexual orientation that describes a woman who is emotionally and sexually attracted to other women.
Men who have sex with men (MSM)/Women who have sex with women (WSW)	Categories used in research and public health to describe those who engage in same-sex sexual behavior, regardless of their sexual orientation. Individuals rarely use the terms MSM or WSW to describe themselves.
Pangender	Describes a person whose gender identity comprises many genders.
Pansexual	A sexual orientation that describes a person who is emotionally and sexually attracted to people regardless of gender.
Queer	An umbrella term used by some to describe people who think of their sexual orientation or gender identity as outside of societal norms. Some people view the term as more fluid and inclusive than traditional categories for sexual orientation and gender identity. Due to its history as a derogatory term, it is not embraced or used by all members of the LGBT community.
Questioning	Describes an individual who is unsure about or is exploring their own sexual orientation and/or gender identity.
Same-sex attraction	Describes the experience of a person who is emotionally and/or sexually attracted to people of the same gender. Use of this term is not indicative of a person's sexual behavior.
Sexual orientation	Describes how a person characterizes their physical and emotional attraction to others. Sexual orientation is distinct from sex, gender identity, and gender expression.
Trans man/transgender man/female-to-male (FTM)	A transgender person whose gender identity is male may use these terms to describe themselves. Some will just use the term *man*.

Term	Definition
Trans woman/transgender woman/male-to-female (MTF)	A transgender person whose gender identity is female may use these terms to describe themselves. Some will just use the term *woman*.
Transgender	Describes a person whose gender identity and assigned sex at birth do not correspond. Also used as an umbrella term to include gender identities outside of male and female.
Transition/Affirmation	For transgender persons, the process of coming to recognize, accept, and express one's gender identity. Most often, this refers to the period when a person makes social, legal, and/or medical changes, such as changing their clothing, name, and sex designation, as well as using medical interventions.

> *Note:* It is important to note that definitions vary across communities, that they change over time, and that not all LGBT people agree with all these definitions.

The creation of a welcoming environment requires not making any assumptions about an individual's gender identity or sexual orientation. Both front-line staff and clinicians should be cognizant of patient communication. For example:

Instead of saying "How may I help you, sir?"

Say "How may I help you?"

Instead of saying "She is here for her appointment."

Say "The patient is here in the waiting room."

Instead of saying "Do you have a wife?"

Say "Are you in a relationship?

Instead of saying "What are your mother's and fathers' names?"

Say "What are your parents' names?"

Developing Comfort and Competency in Sexual Health

Developing comfort discussing sexual health and intimacy is critical to providing appropriate care. A good starting place is to ask if a patient is sexually active—and if so, with whom, how often, and what they do with

their partner(s). These discussions can allow providers to focus subsequent conversations on issues most relevant to a patient's health. For example, a gay man with multiple sexual partners who engages in receptive anal sex without condoms is at high risk for HIV and sexually transmitted infections. It will be important to recommend more frequent screenings and discuss use of pre-exposure prophylaxis (PrEP) and condoms to prevent HIV and sexually transmitted infections. If you are seeing a transgender man, it will be important to know if he still has natal female genitalia to ensure appropriate cancer screening. Notably, many if not most transgender people have not had genital affirmation surgery and retain their natal sex organs.

Creating a Welcoming and Safe Healthcare Environment

Hospitals and clinics can take a number of steps to create a welcoming and safe space for LGBT patients. This starts by establishing and communicating a nondiscrimination policy that clearly includes gender identity, gender expression, and sexual orientation protections. Additionally, hospitals and clinics can develop and implement an equal visitation policy to ensure equal visitation for LGBT patients from same-sex partners, parents, and other family and friends. Staff training in LGBT patient-centered care also is a key component of creating inclusive health environments. This includes covering LGBT cultural competency, caring for LGBT patients, creating an inclusive environment for LGBT patients and staff, and other topics important for LGBT health.

As hospitals and clinics continue to adopt electronic health records, collecting sexual orientation and gender identity information becomes increasingly important to delivering personalized care to LGBT individuals. It allows providers to monitor quality of care and track population-based outcomes. This information can be captured by three questions:

Do you think of yourself as: straight or heterosexual; lesbian, gay, or homosexual; bisexual; something else; don't know; choose not to disclose.

What is your current gender identity? Male; female; transgender male/trans man/female-to-male (FTM); transgender female/trans woman/male-to-female (MTF); genderqueer, neither exclusively male nor female; additional category, please specify; choose not to disclose.

What sex were you assigned at birth on your original birth certificate? Male; female; choose not to disclose.

The physical environment of a hospital or clinic is important, but the majority of clinical spaces do not signal that they are safe spaces for LGBT patients. Most healthcare posters, pamphlets, and materials feature heterosexual individuals or couples; adding LGBT-friendly images and text can help signal that the hospital or clinic is a safe space for sexual and gender minorities. In addition, easily identifying LGBT-competent providers by using websites, buttons, and pins can help patients select providers and feel at ease when attending appointments. Lastly, designating all-gender bathrooms is important to creating welcoming spaces, particularly for transgender and gender-nonconforming individuals.

FUTURE DIRECTION IN LGBT HEALTH

While social and cultural acceptance of the LGBT community has improved in certain parts of the world, many LGBT individuals continue to experience discrimination, stigmatization, and violence. Inequitable healthcare policies and practices, lack of awareness to LGBT health issues, and limited understanding of the unique health needs of LGBT individuals contribute to decreased access to care and disparities in health outcomes for LGBT individuals. Addressing these barriers will require improved data collection on the LGBT population; understanding of the intersectionality of gender identity, sexual orientation, race/ethnicity, and other determinants of health; and outcomes-focused research across the life course. In striving to deliver high-quality care experiences for all patients, hospitals, clinics, and providers will have to focus on meeting the needs of the LGBT community.

FURTHER READING

Centers for Disease Control and Prevention: Lesbian, Gay, Bisexual, and Transgender Health. 2016. Available at *www.cdc.gov/lgbthealth/*. Accessed September 5, 2017.

Fenway Health: Glossary of LGBT Terms for Health Care Teams. 2016. Available at *www.lgbthealtheducation.org/wp-content/uploads/LGBT-Glossary_March2016.pdf*. Accessed September 5, 2017.

Institute of Medicine: The Health of Lesbian, Gay, Bisexual, and Transgender (LGBT) People: Building a Foundation for Better Understanding. 2011. Available at *www.nap.edu/catalog.php?record_id=13128*. Accessed September 5, 2017.

Institute of Medicine: Collecting Sexual Orientation and Gender Identity Data in Electronic Health Records. 2012. Available at *www.nationalacademies.org/hmd/Reports/2012/Collecting-Sexual-Orientation-and-Gender-Identity-Data-in-Electronic-Health-Records.aspx*. Accessed September 5, 2017.

Joint Commission: Advancing Effective Communication, Cultural Competence, Patient- and Family-Centered Care for the Lesbian, Gay, Bisexual, and Transgender Community: A Field Guide; 2011. Available at *www.jointcommission.org/lgbt*. Accessed September 5, 2017.

Substance Abuse and Mental Health Services Administration: Top Health Issues for LGBT Populations Information & Resource Kit. 2012. Available at *www.store.samhsa.gov/shin/content/SMA12-4684/SMA12-4684.pdf*. Accessed September 5, 2017.

Chapter 394: Pathobiology of Obesity

Jeffrey S. Flier; Eleftheria Maratos-Flier

INTRODUCTION

In a world where food supplies are intermittent, the ability to store energy in excess of what is required for immediate use is essential for survival. Fat cells, residing within widely distributed adipose tissue depots, are adapted to store excess energy efficiently as triglyceride and, when needed, to release stored energy as free fatty acids for use at other sites. This physiologic system, orchestrated through endocrine and neural pathways, permits humans to survive starvation for as long as several months. However, in the presence of nutritional abundance and a sedentary lifestyle, and influenced importantly by genetic endowment, this system increases adipose energy stores and produces adverse health consequences.

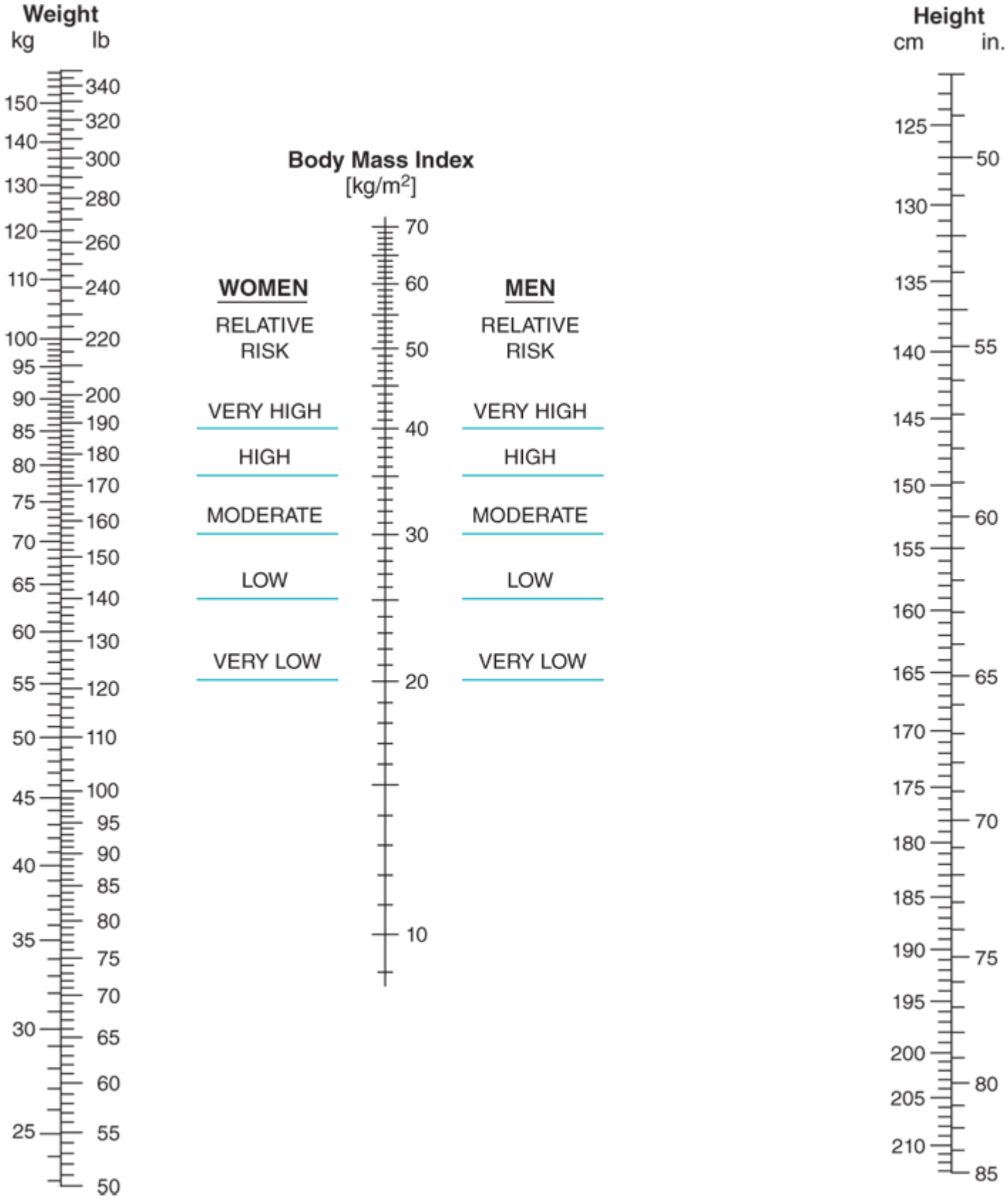
DEFINITION AND MEASUREMENT

Obesity is a state of excess adipose tissue mass. Although often viewed as equivalent to increased body weight, this need not be the case—lean but very muscular individuals may be overweight by numerical standards without having increased adiposity. Body weights are distributed continuously in populations, so that choice of a medically meaningful distinction between lean and obese is somewhat arbitrary. Obesity is therefore defined by assessing its linkage to morbidity or mortality.

Although not a direct measure of adiposity, the most widely used method to gauge obesity is the *body mass index* (BMI), which is equal to weight/height^2 (in kg/m^2) **(Fig. 394-1)**. Other approaches to quantifying obesity include anthropometry (skinfold thickness), densitometry (underwater weighing), computed tomography (CT) or magnetic resonance imaging (MRI), and electrical impedance. Using data from the Metropolitan Life Tables, BMIs for the midpoint of all heights and frames among both men and women range from 19 to 26 kg/m^2; at a similar BMI, women have more body fat than men. Based on data of substantial morbidity, a BMI of 30 is most commonly used as a threshold for obesity in both men and women. Most but not all large-scale epidemiologic studies suggest that all-cause, metabolic, cancer, and cardiovascular morbidity begin to rise (albeit at a slow rate) when BMIs are ≥25. Most authorities use the term *overweight* (rather than obese) to describe individuals with BMIs between 25 and 30. A BMI between 25 and 30 should be viewed as medically significant and worthy of therapeutic intervention in the presence of risk factors that are influenced by adiposity, such as hypertension and glucose intolerance.

FIGURE 394-1

Nomogram for determining body mass index. To use this nomogram, place a ruler or other straight edge between the body weight (without clothes) in kilograms or pounds located on the left-hand line and the height (without shoes) in centimeters or inches located on the right-hand line. The body mass index is read from the middle of the scale and is in metric units. *(Copyright 1979, George A. Bray, MD; used with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The distribution of adipose tissue in different anatomic depots also has substantial implications for morbidity. Specifically, intraabdominal and abdominal subcutaneous fat have more significance than subcutaneous fat present in the buttocks and lower extremities. This distinction is most easily made

clinically by determining the waist-to-hip ratio, with a ratio >0.9 in women and >1.0 in men being abnormal. Many of the most important complications of obesity, such as insulin resistance, diabetes, hypertension, hyperlipidemia, and hyperandrogenism in women, are linked more strongly to intraabdominal and/or upper body fat than to overall adiposity **(Chap. 401)**. The mechanism underlying this association is unknown but may relate to the fact that intraabdominal adipocytes are more lipolytically active than those from other depots. Release of free fatty acids into the portal circulation has adverse metabolic actions, especially on the liver. Adipokines and cytokines that are differentially secreted by adipocyte depots may play a role in the systemic complications of obesity.

PREVALENCE

Data from the National Health and Nutrition Examination Surveys (NHANES) show that the percentage of the American adult population with obesity (BMI >30) has increased from 14.5% (between 1976 and 1980) to 36.5% (between 2011 and 2014). As many as 70% of U.S. adults aged ≥20 years were either overweight (defined as BMI >25) or obese (BMI >30) between the years of 2013 and 2014. Extreme obesity (BMI ≥40) has also increased and affects 5.7% of the population. The increasing prevalence of medically significant obesity raises great concern. Overall, the prevalence of obesity is higher in women (38%) than in men (34%). In women, poverty is associated with increased prevalence. Obesity is more common among blacks and Hispanics, and less common in Asians. Among some Asian subgroups, health and mortality risks may begin at lower BMIs, associated with greater intraabdomnal obesity. The prevalence in children and adolescents has been rising at a worrisome rate, reaching 17.0% in 2011–2014, but may be leveling off.

GLOBAL CONSIDERATIONS

 According to the World Health Organization, global obesity almost doubled between 1980 and 2008. There were >200 million obese men and almost 300 million obese women, 11% of adults worldwide, in 2008. A majority (65%) of the world's population lives in countries where being overweight or obese kills more people than being underweight. In 2014, the U.S. led the Organization for Economic Co-operation and Development (OECD) countries in obesity prevalence (34%), with New Zealand, Australia, the UK, Mexico, and Canada close behind in the high 20% range. Worldwide, the highest prevalence of obesity is in several Pacific Island states including the Cook Islands (50%), Nauru (45.6%), and Tonga (43.3%) and Gulf States including Qatar (42%) and Kuwait (38%). In developing countries, such as India, obesity prevalence is rising (5%) with a greater tendency to harmful intraabdominal obesity at lower BMIs in this population, and the consequences for metabolic and cardiovascular health are disproportionate to obesity prevalence. The causes of increased obesity globally are complicated, and may vary from country to country, but contributory factors include rising incomes, changing food supplies, and reduced activity.

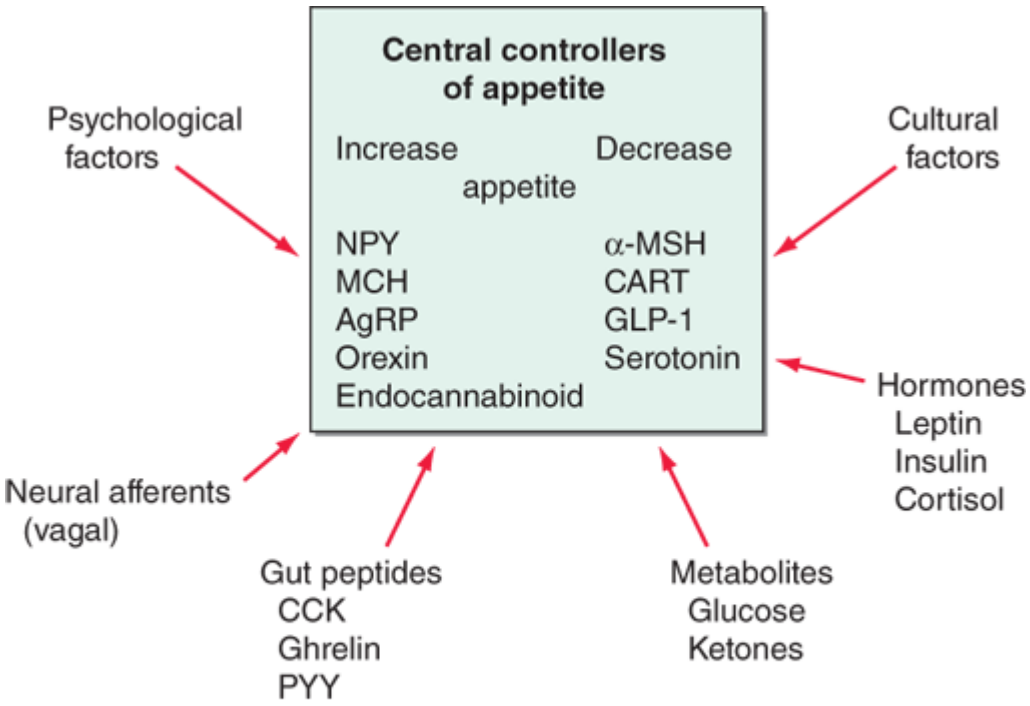
PHYSIOLOGIC REGULATION OF ENERGY BALANCE

Substantial evidence suggests that body weight is regulated by both endocrine and neural components that ultimately influence the effector arms of energy intake and expenditure. This complex regulatory system is necessary because even small imbalances between energy intake and expenditure will ultimately have large effects on body weight. For example, a 0.3% positive imbalance over 30 years would result in a 9-kg (20-lb) weight gain. This exquisite regulation of energy balance cannot be monitored easily by calorie-counting in relation to physical activity. Rather, body weight regulation or dysregulation depends on a complex interplay of hormonal and neural signals. Alterations in stable weight by forced overfeeding or food deprivation induce physiologic changes that resist these perturbations: with weight loss, appetite increases and energy expenditure falls; with overfeeding, appetite falls and energy expenditure increases. This latter compensatory mechanism frequently fails, however, permitting obesity to develop when food is abundant and physical activity is limited. A major regulator of these adaptive responses is the adipocyte-derived hormone leptin, which acts through brain circuits (predominantly in the hypothalamus) to influence appetite, energy expenditure, and neuroendocrine function (see below).

Appetite is influenced by many factors that are integrated by the brain, most importantly within the hypothalamus **(Fig. 394-2)**. Signals that impinge on the hypothalamic center include neural afferents, hormones, and metabolites. Vagal inputs are particularly important, bringing information from viscera, such as gut distention. Hormonal signals include leptin, insulin, cortisol, and gut peptides. Among the latter is ghrelin, which is made in the stomach and stimulates feeding, and peptide YY (PYY) and cholecystokinin, which is made in the small intestine and signals to the brain through direct action on hypothalamic control centers and/or via the vagus nerve. Metabolites, including glucose, can influence appetite, as seen by the effect of hypoglycemia to induce hunger; however, glucose is not normally a major regulator of appetite. These diverse hormonal, metabolic, and neural signals act by influencing the expression and release of various hypothalamic peptides (e.g., neuropeptide Y [NPY], Agouti-related peptide [AgRP], α-melanocyte-stimulating hormone [α-MSH], and melanin-concentrating hormone [MCH]) that are integrated with serotonergic, catecholaminergic, endocannabinoid, and opioid-signaling pathways (see below). Psychological and cultural factors also play a role in the final expression of appetite. Apart from rare genetic syndromes involving leptin, its receptor, and the melanocortin system, specific defects in this complex appetite control network that influence common cases of obesity are not well defined.

FIGURE 394-2

The factors that regulate appetite through effects on central neural circuits. Some factors that increase or decrease appetite are listed. AgRP, Agouti-related peptide; CART, cocaine- and amphetamine-related transcript; CCK, cholecystokinin; GLP-1, glucagon--related peptide-1; MCH, melanin-concentrating hormone; α-MSH, α-melanocyte-stimulating hormone; NPY, neuropeptide Y.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Energy expenditure includes the following components: (1) resting or basal metabolic rate; (2) the energy cost of metabolizing and storing food; (3) the thermic effect of exercise; and (4) adaptive thermogenesis, which varies in response to long-term caloric intake (rising with increased intake). Basal metabolic rate accounts for ~70% of daily energy expenditure, whereas active physical activity contributes 5–10%. Thus, a significant component of daily energy consumption is fixed.

Genetic models in mice indicate that mutations in certain genes (e.g., targeted deletion of the insulin receptor in adipose tissue) protect against obesity, apparently by increasing energy expenditure. Adaptive thermogenesis occurs in *brown adipose tissue* (BAT), which plays an important role in energy metabolism in many mammals. In contrast to white adipose tissue, which is used to store energy in the form of lipids, BAT expends stored energy as heat. A mitochondrial *uncoupling protein* (UCP-1) in BAT dissipates the hydrogen ion gradient in the oxidative respiration chain and releases energy as heat. The metabolic activity of BAT is increased by a central action of leptin, acting through the sympathetic nervous system that heavily innervates this tissue. In rodents, BAT deficiency causes obesity and diabetes; stimulation of BAT with a specific adrenergic agonist (β_3 agonist) protects against diabetes and obesity. BAT exists in humans (especially neonates), and although its physiologic role is not yet established, identification of functional BAT in many adults using positron emission tomography (PET) imaging has increased interest in the implications of the tissue for pathogenesis and therapy of obesity. Beige fat cells, recently described, resemble BAT cells in expressing UCP-1. They are scattered through white adipose tissue, and their thermogenic potential is uncertain.

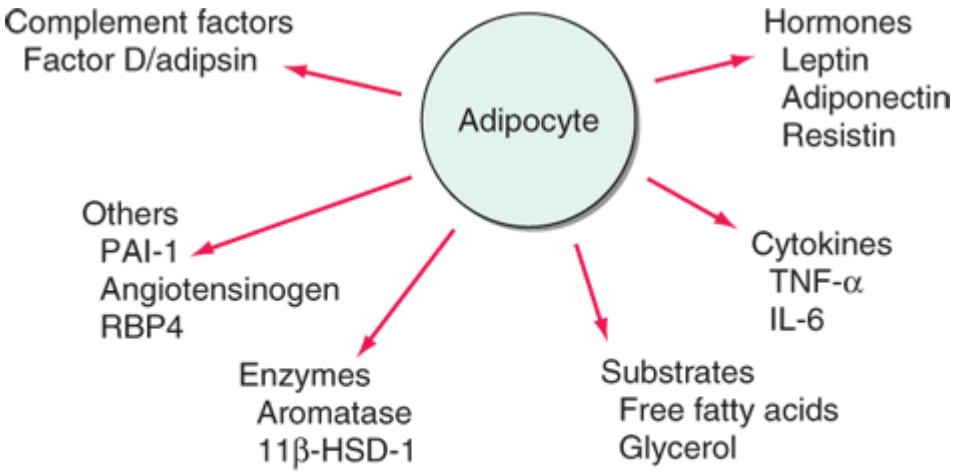
THE ADIPOCYTE AND ADIPOSE TISSUE

Adipose tissue is composed of the lipid-storing adipose cell and a stromal/vascular compartment in which cells including preadipocytes and macrophages reside. Adipose mass increases by enlargement of adipose cells through lipid deposition, as well as by an increase in the number of adipocytes. Obese adipose tissue is also characterized by increased numbers of infiltrating macrophages. The process by which adipose cells are derived from a mesenchymal preadipocyte involves an orchestrated series of differentiation steps mediated by a cascade of specific transcription factors. One of the key transcription factors is *peroxisome proliferator-activated receptor γ* (PPARγ), a nuclear receptor that binds the thiazolidinedione class of insulin-sensitizing drugs used in the treatment of type 2 diabetes **(Chap. 397)**.

Although the adipocyte has generally been regarded as a storage depot for fat, it is also an endocrine cell that releases numerous molecules in a regulated fashion **(Fig. 394-3)**. These include the energy balance–regulating hormone leptin, cytokines such as tumor necrosis factor (TNF)-α and interleukin (IL)-6, complement factors such as factor D (also known as *adipsin*), prothrombotic agents such as plasminogen activator inhibitor I, and a component of the blood pressure–regulating system, angiotensinogen. Adiponectin, an abundant adipose-derived protein whose levels are reduced in obesity, enhances insulin sensitivity and lipid oxidation and has vascular-protective effects, whereas RBP4, whose levels are increased in obesity, may induce insulin resistance. Obesity is accompanied by increased fat storage in tissues such as muscle and liver, and this ectopic lipid has been linked to metabolic disturbances. These factors, and others not yet identified, play a role in the physiology of lipid homeostasis, insulin sensitivity, blood pressure control, coagulation, and vascular health, and are likely to contribute to obesity-related pathologies.

FIGURE 394-3

Factors released by the adipocyte that can affect peripheral tissues. IL-6, interleukin 6; PAI, plasminogen activator inhibitor; RBP4, retinal binding protein 4; TNF, tumor necrosis factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ETIOLOGY OF OBESITY

Although the molecular pathways regulating energy balance are beginning to be illuminated, the causes of obesity remain elusive. In part, this reflects the fact that obesity is a heterogeneous group of disorders. At

one level, the pathophysiology of obesity seems simple: a chronic excess of nutrient intake relative to the level of energy expenditure. However, due to the complexity of the neuroendocrine and metabolic systems that regulate energy intake, storage, and expenditure, it has been difficult to quantitate all the relevant parameters (e.g., food intake and energy expenditure) over time in human subjects.

Role of Genes Versus Environment

Obesity is commonly seen in families, and the heritability of body weight is similar to that for height. Inheritance is usually not Mendelian, however, and it is difficult to distinguish the role of genes and environmental factors. Adoptees more closely resemble their biologic than adoptive parents with respect to obesity, providing strong support for genetic influences. Likewise, identical twins have very similar BMIs whether reared together or apart, and their BMIs are much more strongly correlated than those of dizygotic twins. These genetic effects appear to relate to both energy intake and expenditure. Currently, identified genetic variants, both common and rare, account for <5% of the variance of body weight.

Whatever the role of genes, it is clear that the environment plays a key role in obesity, as evidenced by the fact that famine limits obesity in even the most obesity-prone individual. In addition, the recent increase in the prevalence of obesity in the United States is far too rapid to be due to changes in the gene pool. Undoubtedly, genes influence the susceptibility to obesity in response to specific diets and availability of nutrition. Cultural factors are also important—these relate to both availability and composition of the diet and to changes in the level of physical activity. In industrial societies, obesity is more common among poor women, whereas in underdeveloped countries, wealthier women are more often obese. In children, obesity correlates to some degree with time spent watching television. Although the role of diet composition in obesity continues to generate controversy, it appears that high-fat diets may, when combined with simple, rapidly absorbed carbohydrates, promote obesity. Specific genes are likely to influence the response to specific diets, but these genes are largely unidentified.

Additional environmental factors may contribute to the increasing obesity prevalence. Both epidemiologic correlations and experimental data suggest that sleep deprivation leads to increased obesity. Changes in gut microbiome with capacity to alter energy balance are receiving experimental support from animal studies, and a possible role for obesigenic viral infections continues to receive sporadic attention.

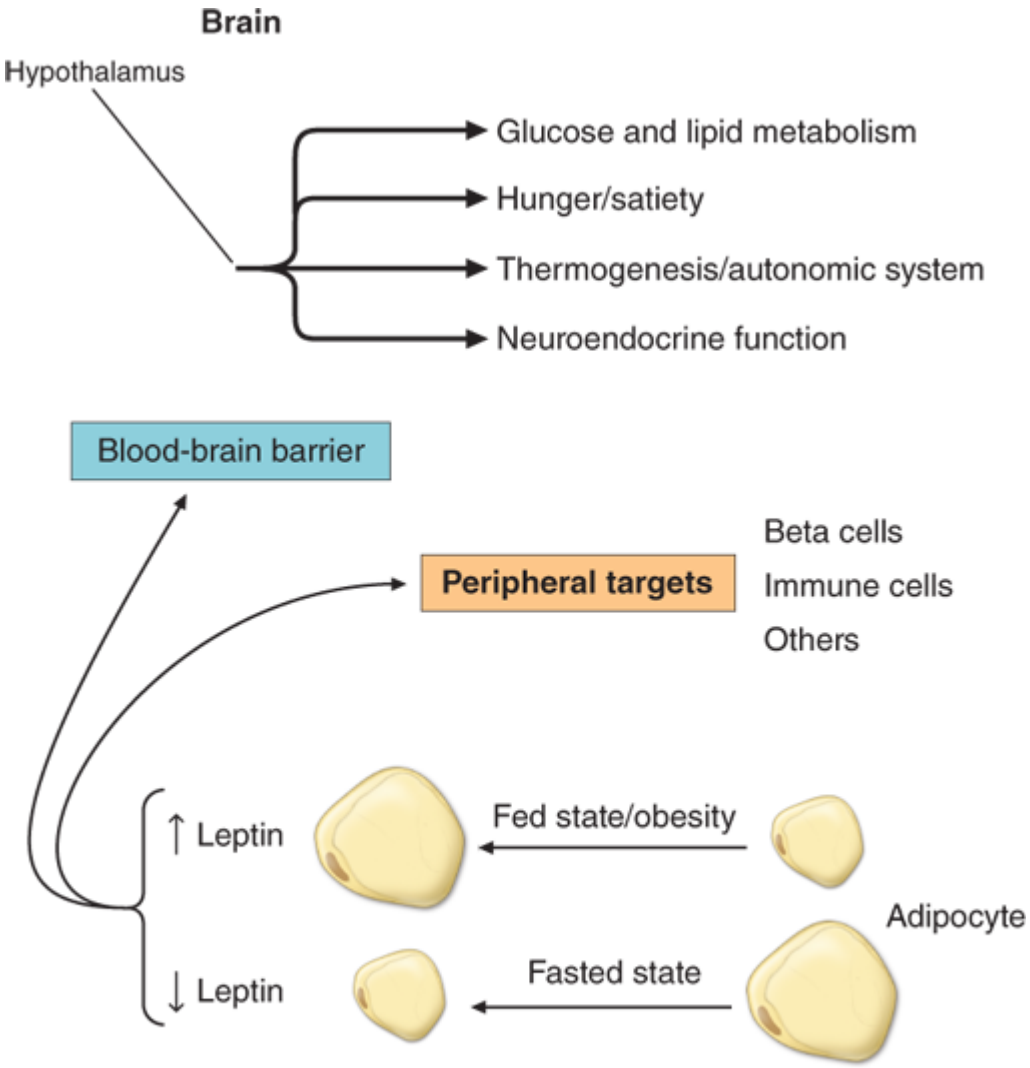
Specific Genetic Syndromes

For many years, obesity in rodents has been known to be caused by a number of distinct mutations distributed through the genome. Most of these single-gene mutations cause both hyperphagia and diminished energy expenditure, suggesting a physiologic link between these two parameters of energy homeostasis. Identification of the *ob* gene mutation in genetically obese (ob/ob) mice represented a major breakthrough in the field. The ob/ob mouse develops severe obesity, insulin resistance, and hyperphagia, as well as efficient metabolism (e.g., it gets fat even when ingesting the same number of calories as lean litter mates). The product of the *ob* gene is the peptide leptin, a name derived from the Greek root *leptos*, meaning thin. Leptin is secreted by adipose cells and acts primarily through the hypothalamus. Its level of production

provides an index of adipose energy stores (Fig. 394-4). Raising leptin levels can decrease food intake and increase energy expenditure. Another mouse mutant, db/db, which is resistant to leptin, has a mutation in the leptin receptor and develops a similar syndrome. The *ob* gene is present in humans where it is also expressed in fat. Several families with morbid, early-onset obesity caused by inactivating mutations in either leptin or the leptin receptor have been described, thus demonstrating the biologic relevance of the leptin pathway in humans. Obesity in these individuals begins shortly after birth, is severe, and is accompanied by neuroendocrine abnormalities. The most prominent of these is hypogonadotropic hypogonadism, which is reversed by leptin replacement in the leptin-deficient subset. Central hypothyroidism and growth retardation are seen in the mouse model, but their occurrence in leptin-deficient humans is less clear. Mutations in the leptin or leptin receptor genes do not play a prominent role in common forms of obesity.

FIGURE 394-4

The physiologic system regulated by leptin. Rising or falling leptin levels act through the hypothalamus to influence appetite, energy expenditure, and neuroendocrine function and through peripheral sites to influence systems such as the immune system.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Mutations in several other genes cause severe obesity in humans **(Table 394-1)**; each of these syndromes is rare. Mutations in the gene encoding proopiomelanocortin (POMC) cause severe obesity through failure to synthesize α-MSH, and potentially β-MSH, key neuropeptides that inhibit appetite in the hypothalamus. The absence of POMC also causes secondary adrenal insufficiency due to absence of adrenocorticotropic hormone (ACTH), as well as pale skin and red hair due to absence of α-MSH. Proenzyme convertase 1 (PC-1) mutations are thought to cause obesity by preventing synthesis of α-MSH from its precursor peptide, POMC. α-MSH binds to the type 4 melanocortin receptor (MC4R), a key hypothalamic receptor that inhibits eating. Heterozygous loss-of-function mutations of this receptor account for as much as 5% of severe obesity. Loss of function of MRAP2, a protein required for normal MC4R signaling, has been found in rare cases of severe obesity. These six genetic defects define a pathway through which leptin (by stimulating POMC and increasing α-MSH) restricts food intake and limits weight **(Fig. 394-5)**. The results of genomewide association studies to identify genetic loci responsible for obesity in the general population have so far been disappointing. More than 100 replicated loci linked to obesity have been identified, but together they account for <3% of interindividual variation in BMI. The most replicated of these involves a region which includes several genes including *FTO*, which is of unknown function, but like many of the other recently described candidates, is expressed in the brain. Because the heritability of obesity is estimated to be 40–70%, it is likely that many more loci remain to be identified. It is possible that epistatic interactions between causative loci or unknown gene-environment interactions explain the missing heritability of obesity.

TABLE 394-1

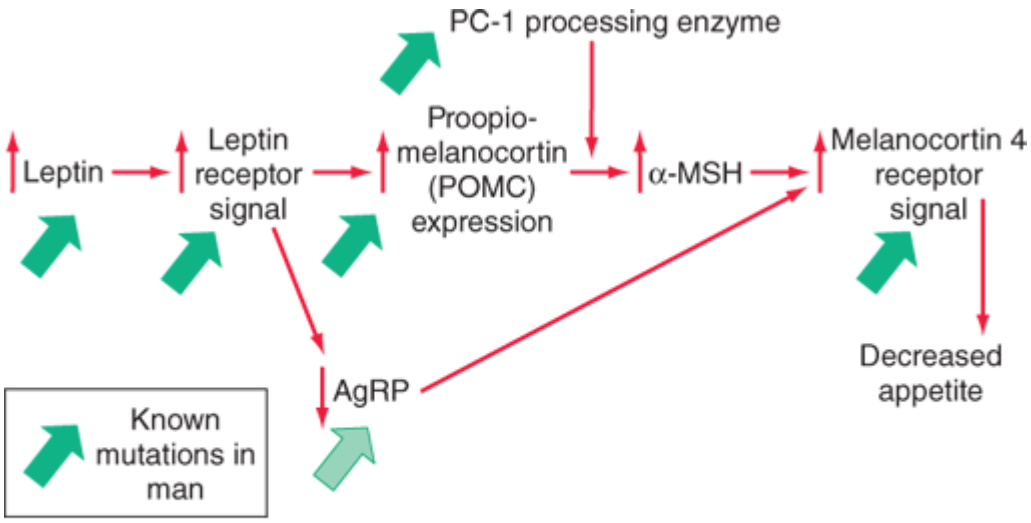
Selected Obesity Genes in Humans and Mice

Gene	Gene Product	Mechanism of Obesity	In Human	In Rodent
Lep (ob)	Leptin, a fat-derived hormone	Mutation prevents leptin from delivering satiety signal; brain perceives starvation	Yes	Yes
LepR (db)	Leptin receptor	Same as above	Yes	Yes
POMC	Proopiomelanocortin, a precursor of several hormones and neuropeptides	Mutation prevents synthesis of melanocyte-stimulating hormone (MSH), a satiety signal	Yes	Yes
MC4R	Type 4 receptor for MSH	Mutation prevents reception of satiety signal from MSH	Yes	Yes
AgRP	Agouti-related peptide, a neuropeptide expressed in the hypothalamus	Overexpression inhibits signal through *MC4R*	No	Yes
PC-1	Prohormone convertase 1, a processing enzyme	Mutation prevents synthesis of neuropeptide, probably MSH	Yes	No
Fat	Carboxypeptidase E, a processing enzyme	Same as above	No	Yes
Tub	Tub, a hypothalamic protein of unknown function	Hypothalamic dysfunction	No	Yes
TrkB	TrkB, a neurotrophin receptor	Hyperphagia due to uncharacterized hypothalamic defect	Yes	Yes

FIGURE 394-5

A central pathway through which leptin acts to regulate appetite and body weight. Leptin signals through proopiomelanocortin (POMC) neurons in the hypothalamus to induce increased production of α-melanocyte-stimulating hormone (α-MSH), requiring the processing enzyme PC-1 (proenzyme convertase 1). α-MSH acts

as an agonist on melanocortin-4 receptors to inhibit appetite, and the neuropeptide AgRp (Agouti-related peptide) acts as an antagonist of this receptor. Mutations that cause obesity in humans are indicated by the *solid green arrows*.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In addition to these human obesity genes, studies in rodents reveal several other molecular candidates for hypothalamic mediators of human obesity or leanness. The *tub* gene encodes a hypothalamic peptide of unknown function; mutation of this gene causes late-onset obesity. The *fat* gene encodes carboxypeptidase E, a peptide-processing enzyme; mutation of this gene is thought to cause obesity by disrupting production of one or more neuropeptides. Mutations in TUB and Carboxypeptidase-E (CPE) have recently been identified in humans.

AgRP is coexpressed with NPY in arcuate nucleus neurons. AgRP antagonizes α-MSH action at MC4 receptors, and its overexpression induces obesity. In contrast, a mouse deficient in the peptide MCH, whose administration causes feeding, is lean.

A number of complex human syndromes with defined inheritance are associated with obesity (Table 394-2). Although specific genes have limited definition at present, their identification may enhance our understanding of more common forms of human obesity. In the Prader-Willi syndrome, a multigenic neurodevelopmental disorder, obesity coexists with short stature, mental retardation, hypogonadotropic hypogonadism, hypotonia, small hands and feet, fish-shaped mouth, and hyperphagia. Most patients have reduced expression of imprinted paternally inherited genes encoded in the 15q11-13 chromosomal region. Reduced expression of Snord116, a small nucleolar RNA highly expressed in hypothalamus, may be an important cause of defective hypothalamic function in this disorder. Bardet-Biedl syndrome (BBS) is a genetically heterogeneous disorder characterized by obesity, mental retardation, retinitis pigmentosa, diabetes, renal and cardiac malformations, polydactyly, and hypogonadotropic hypogonadism. At least 16 genetic loci have been identified, and most of the encoded proteins form two multiprotein complexes that are involved in ciliary function and microtubule-based intracellular transport. Some evidence suggests that mutations might disrupt leptin receptor trafficking in key hypothalamic neurons, causing leptin resistance.

TABLE 394-2

A Comparison of Syndromes of Obesity—Hypogonadism and Mental Retardation

	Syndrome				
Feature	**Prader-Willi**	**Laurence-Moon-Biedl**	**Ahlstrom's**	**Cohen's**	**Carpenter's**
Inheritance	Sporadic; two-thirds have defect	Autosomal recessive	Autosomal recessive	Probably autosomal recessive	Autosomal recessive
Stature	Short	Normal; infrequently short	Normal; infrequently short	Short or tall	Normal
Obesity	Generalized	Generalized	Truncal	Truncal	Truncal, gluteal
	Moderate to severe	Early onset, 1–2 years	Early onset, 2–5 years	Mid-childhood, age 5	
	Onset 1–3 years				
Craniofacies	Narrow bifrontal diameter	Not distinctive	Not distinctive	High nasal bridge	Acrocephaly
	Almond-shaped eyes			Arched palate	Flat nasal bridge
	Strabismus			Open mouth	High-arched palate
	V-shaped mouth			Short philtrum	
	High-arched palate				

	Syndrome				
Feature	**Prader-Willi**	**Laurence-Moon-Biedl**	**Ahlstrom's**	**Cohen's**	**Carpenter's**
Limbs	Small hands and feet	Polydactyly	No abnormalities	Hypotonia	Polydactyly
	Hypotonia			Narrow hands and feet	Syndactyly
					Genu valgum
Reproductive status	1° Hypogonadism	1° Hypogonadism	Hypogonadism in males but not in females	Normal gonadal function or hypogonadotropic hypogonadism	2° Hypogonadism
Other features	Enamel hypoplasia			Dysplastic ears	
	Hyperphagia			Delayed puberty	
	Temper tantrums				
	Nasal speech				
Mental retardation	Mild to moderate		Normal intelligence	Mild	Slight

Other Specific Syndromes Associated with Obesity

CUSHING'S SYNDROME

Although obese patients commonly have central obesity, hypertension, and glucose intolerance, they lack other specific stigmata of Cushing's syndrome (Chap. 379). Nonetheless, a potential diagnosis of Cushing's syndrome is often entertained. Cortisol production and urinary metabolites (17OH steroids) may be increased in simple obesity. Unlike in Cushing's syndrome, however, cortisol levels in blood and urine in the basal state and in response to corticotropin-releasing hormone (CRH) or ACTH are normal; the overnight 1-

mg dexamethasone suppression test is normal in 90%, with the remainder being normal on a standard 2-day low-dose dexamethasone suppression test. Obesity may be associated with excessive local reactivation of cortisol in fat by 11β-hydroxysteroid dehydrogenase 1, an enzyme that converts inactive cortisone to cortisol.

HYPOTHYROIDISM

The possibility of hypothyroidism should be considered, but it is an uncommon cause of obesity; hypothyroidism is easily ruled out by measuring thyroid-stimulating hormone (TSH). Much of the weight gain that occurs in hypothyroidism is due to myxedema **(Chap. 375)**.

INSULINOMA

Patients with insulinoma often gain weight as a result of overeating to avoid hypoglycemic symptoms **(Chap. 399)**. The increased substrate plus high insulin levels promote energy storage in fat. This can be marked in some individuals but is modest in most.

CRANIOPHARYNGIOMA AND OTHER DISORDERS INVOLVING THE HYPOTHALAMUS

Whether through tumors, trauma, or inflammation, hypothalamic dysfunction of systems controlling satiety, hunger, and energy expenditure can cause varying degrees of obesity **(Chap. 372)**. It is uncommon to identify a discrete anatomic basis for these disorders. Subtle hypothalamic dysfunction is probably a more common cause of obesity than can be documented using currently available imaging techniques. Growth hormone (GH), which exerts lipolytic activity, is diminished in obesity and is increased with weight loss. Despite low GH levels, insulin-like growth factor (IGF) I (somatomedin) production is normal, suggesting that GH suppression may be a compensatory response to increased nutritional supply.

Pathogenesis of Common Obesity

Obesity can result from increased energy intake, decreased energy expenditure, or a combination of the two. Thus, identifying the etiology of obesity should involve measurements of both parameters. However, it is difficult to perform direct and accurate measurements of energy intake in free-living individuals; and the obese, in particular, often underreport intake. Measurements of chronic energy expenditure are possible using doubly labeled water or metabolic chamber/rooms. In subjects at stable weight and body composition, energy intake equals expenditure. Consequently, these techniques allow assessment of energy intake in free-living individuals. The level of energy expenditure differs in established obesity, during periods of weight gain or loss, and in the pre- or postobese state. Studies that fail to take note of this phenomenon are not easily interpreted.

There is continued interest in the concept of a body weight “set point.” This idea is supported by physiologic mechanisms centered around a sensing system in adipose tissue that reflects fat stores and a receptor, or “adipostat,” that is in the hypothalamic centers. When fat stores are depleted, the adipostat signal is low, and the hypothalamus responds by stimulating hunger and decreasing energy expenditure to conserve energy.

Conversely, when fat stores are abundant, the signal is increased, and the hypothalamus responds by decreasing hunger and increasing energy expenditure. The recent discovery of the *ob* gene, and its product leptin, and the *db* gene, whose product is the leptin receptor, provides important elements of a molecular basis for this physiologic concept (see above).

What Is the Status of Food Intake in Obesity? (Do the Obese Eat More Than the Lean?)

This question has stimulated much debate, due in part to the methodologic difficulties inherent in determining food intake. Many obese individuals believe that they eat small quantities of food, and this claim has often been supported by the results of food intake questionnaires. However, it is now established that average energy expenditure increases as individuals get more obese, due primarily to the fact that metabolically active lean tissue mass increases with obesity. Given the laws of thermodynamics, the obese person must therefore eat more than the average lean person to maintain their increased weight. It may be the case, however, that a subset of individuals who are predisposed to obesity have the capacity to become obese initially without an absolute increase in caloric consumption.

What Is the State of Energy Expenditure in Obesity?

The average total daily energy expenditure is higher in obese than lean individuals when measured at stable weight. However, energy expenditure falls as weight is lost, due in part to loss of lean body mass and to decreased sympathetic nerve activity. When reduced to near-normal weight and maintained there for a while, (some) obese individuals have lower energy expenditure than (some) lean individuals. There is also a tendency for those who will develop obesity as infants or children to have lower resting energy expenditure rates than those who remain lean. The physiologic basis for variable rates of energy expenditure (at a given body weight and level of energy intake) is essentially unknown.

Another component of thermogenesis, called *nonexercise activity thermogenesis* (NEAT), has been linked to obesity. It is the thermogenesis that accompanies physical activities other than volitional exercise such as the activities of daily living, fidgeting, spontaneous muscle contraction, and maintaining posture. NEAT accounts for about two-thirds of the increased daily energy expenditure induced by overfeeding. The wide variation in fat storage seen in overfed individuals is predicted by the degree to which NEAT is induced. The molecular basis for NEAT and its regulation is unknown.

Leptin in Typical Obesity

The vast majority of obese persons have increased leptin levels but do not have mutations of either leptin or its receptor. They appear, therefore, to have a form of functional “leptin resistance.” Data suggesting that some individuals produce less leptin per unit fat mass than others or have a form of relative leptin deficiency that predisposes to obesity are at present contradictory and unsettled. The mechanism for leptin resistance, and whether it can be overcome by raising leptin levels or combining leptin with other treatments in a subset of obese individuals, is not yet established. Some data suggest that leptin may not effectively cross the

blood-brain barrier as levels rise. It is also apparent from animal studies that leptin-signaling inhibitors, such as SOCS3 and PTP1b, are involved in the leptin-resistant state.

PATHOLOGIC CONSEQUENCES OF OBESITY

Obesity has major adverse effects on health **(See also Chap. 395)**. Obesity is associated with an increase in mortality, with a 50–100% increased risk of death from all causes compared to normal-weight individuals, mostly due to cardiovascular causes. Obesity and overweight together are the second leading cause of preventable death in the United States, accounting for 300,000 deaths per year. Mortality rates rise as obesity increases, particularly when obesity is associated with increased intraabdominal fat (see above). Life expectancy of a moderately obese individual could be shortened by 2–5 years, and a 20- to 30-year-old male with a BMI >45 may lose 13 years of life. It is likely that the degree to which obesity affects particular organ systems is influenced by susceptibility genes that vary in the population.

Insulin Resistance and Type 2 Diabetes Mellitus

Hyperinsulinemia and insulin resistance are pervasive features of obesity, increasing with weight gain and diminishing with weight loss **(Chap. 401)**. Insulin resistance is more strongly linked to intraabdominal fat than to fat in other depots. Molecular links between obesity and insulin resistance in fat, muscle, and liver have been sought for many years. Major factors include: (1) insulin itself, by inducing receptor downregulation; (2) free fatty acids that are increased and capable of impairing insulin action; (3) intracellular lipid accumulation; and (4) several circulating peptides produced by adipocytes, including the cytokines TNF-α and IL-6, RBP4, and the “adipokine” adiponectin, which have altered expression in obese adipocytes and can modify insulin action. Additional mechanisms are obesity-linked inflammation, including infiltration of macrophages into tissues including fat, and induction of the endoplasmic reticulum stress response, which can bring about resistance to insulin action in cells. Despite the prevalence of insulin resistance, most obese individuals do not develop diabetes, suggesting that diabetes requires an interaction between obesity-induced insulin resistance and other factors such as impaired insulin secretion **(Chap. 396)**. Obesity, however, is a major risk factor for diabetes, and as many as 80% of patients with type 2 diabetes mellitus are obese. Weight loss and exercise, even of modest degree, increase insulin sensitivity and often improve glucose control in diabetes.

Reproductive Disorders

Disorders that affect the reproductive axis are associated with obesity in both men and women. Male hypogonadism is associated with increased adipose tissue, often distributed in a pattern more typical of females. In men whose weight is >160% ideal body weight (IBW), plasma testosterone and sex hormone–binding globulin (SHBG) are often reduced, and estrogen levels (derived from conversion of adrenal androgens in adipose tissue) are increased **(Chap. 384)**. Gynecomastia may be seen. However, masculinization, libido, potency, and spermatogenesis are preserved in most of these individuals. Free testosterone may be decreased in morbidly obese men whose weight is >200% IBW.

Obesity has long been associated with menstrual abnormalities in women, particularly in women with upper body obesity (Chap. 385). Common findings are increased androgen production, decreased SHBG, and increased peripheral conversion of androgen to estrogen. Most obese women with oligomenorrhea have polycystic ovarian syndrome (PCOS), with its associated anovulation and ovarian hyperandrogenism; 40% of women with PCOS are obese. Most nonobese women with PCOS are also insulin-resistant, suggesting that insulin resistance, hyperinsulinemia, or the combination of the two are causative or contribute to the ovarian pathophysiology in PCOS in both obese and lean individuals. In obese women with PCOS, weight loss often restores normal menses. The increased conversion of androstenedione to estrogen, which occurs to a greater degree in women with lower body obesity, may contribute to the increased incidence of uterine cancer in postmenopausal women with obesity.

Cardiovascular Disease

The Framingham Study revealed that obesity was an independent risk factor for the 26-year incidence of cardiovascular disease in men and women (including coronary disease, stroke, and congestive heart failure). The waist-to-hip ratio may be the best predictor of these risks. When the additional effects of hypertension and glucose intolerance associated with obesity are included, the adverse impact of obesity is even more evident. The effect of obesity on cardiovascular mortality in women may be seen at BMIs as low as 25. Obesity, especially abdominal obesity, is associated with an atherogenic lipid profile; with increased low-density lipoprotein cholesterol, very-low-density lipoprotein, and triglyceride; and with decreased high-density lipoprotein cholesterol and decreased levels of the vascular protective adipokine adiponectin (Chap. 400). Obesity is also associated with hypertension. Measurement of blood pressure in the obese requires use of a larger cuff size to avoid artifactual increases. Obesity-induced hypertension is associated with increased peripheral resistance and cardiac output, increased sympathetic nervous system tone, increased salt sensitivity, and insulin-mediated salt retention; it is often responsive to modest weight loss.

Pulmonary Disease

Obesity may be associated with a number of pulmonary abnormalities. These include reduced chest wall compliance, increased work of breathing, increased minute ventilation due to increased metabolic rate, and decreased functional residual capacity and expiratory reserve volume. Severe obesity may be associated with obstructive sleep apnea and the “obesity hypoventilation syndrome” with attenuated hypoxic and hypercapnic ventilatory responses. Sleep apnea can be obstructive (most common), central, or mixed and is associated with hypertension. Weight loss (10–20 kg) can bring substantial improvement, as can major weight loss following gastric bypass or restrictive surgery. Continuous positive airway pressure has been used with some success.

Hepatobiliary Disease

Obesity is frequently associated with nonalcoholic fatty liver disease (NAFLD), and this association represents one of the most common causes of liver disease in industrialized countries. The hepatic fatty infiltration of NAFLD progresses in a subset to inflammatory nonalcoholic steatohepatitis (NASH) and more

rarely to cirrhosis and hepatocellular carcinoma. Steatosis typically improves following weight loss, secondary to diet or bariatric surgery. The mechanism for the association remains unclear. Obesity is associated with enhanced biliary secretion of cholesterol, supersaturation of bile, and a higher incidence of gallstones, particularly cholesterol gallstones (**Chap. 339**). A person 50% above IBW has about a sixfold increased incidence of symptomatic gallstones. Paradoxically, fasting increases supersaturation of bile by decreasing the phospholipid component. Fasting-induced cholecystitis is a complication of extreme diets.

Cancer

Obesity is associated with increased risk of several cancer types, and in addition can lead to poorer treatment outcomes and increased cancer mortality. Obesity in males is associated with higher mortality from cancer of the esophagus, colon, rectum, pancreas, liver, and prostate; obesity in females is associated with higher mortality from cancer of the gallbladder, bile ducts, breasts, endometrium, cervix, and ovaries. Some of the latter may be due to increased rates of conversion of androstenedione to estrone in adipose tissue of obese individuals. Other possible mechanistic links may involve hormones, growth factors, and cytokines whose levels are linked to nutritional state, including insulin, leptin, adiponectin, and IGF-I, as well as activation of signaling pathways linked to both obesity and cancer. It has been estimated that obesity accounts for 14% of cancer deaths in men and 20% in women in the United States.

Bone, Joint, and Cutaneous Disease

Obesity is associated with an increased risk of osteoarthritis, no doubt partly due to the trauma of added weight bearing, but potentially linked as well to activation of inflammatory pathways that could promote synovial pathology. The prevalence of gout may also be increased (**Chap. 365**). One of the skin problems associated with obesity is acanthosis nigricans, manifested by darkening and thickening of the skinfolds on the neck, elbows, and dorsal interphalangeal spaces. Acanthosis reflects the severity of underlying insulin resistance and diminishes with weight loss. Friability of skin may be increased, especially in skinfolds, enhancing the risk of fungal and yeast infections. Finally, venous stasis is increased in the obese.

FURTHER READING

Allison MB, Myers MG: 20 years of leptin: Connecting leptin signaling to biological function. J Endocrinol 223:T25, 2014. [PubMed: 25232147]

Farooqi IS, O'Rahilly S: 20 years of leptin: Human disorders of leptin action. J Endocrinol 223:T63, 2014. [PubMed: 25232148]

Heymsfield SB, Wadden TA: Mechanisms, pathophysiology and management of obesity. N Engl J Med 376:254, 2017. [PubMed: 28099824]

Ogden CL et al: Prevalence of obesity in the United States. JAMA 312:189, 2014. [PubMed: 25005661]

Chapter 395: Evaluation and Management of Obesity

Robert F. Kushner

INTRODUCTION

More than 66% of U.S. adults are categorized as overweight or obese, and the prevalence of obesity is increasing rapidly in most of the industrialized world. Children and adolescents also are becoming more obese, indicating that the current trends will accelerate over time. Obesity is associated with an increased risk of multiple health problems, including hypertension, type 2 diabetes, dyslipidemia, obstructive sleep apnea, nonalcoholic fatty liver disease, degenerative joint disease, and some malignancies. Thus, it is important for physicians to identify, evaluate, and treat patients for obesity and associated comorbid conditions.

EVALUATION

Physicians should screen all adult patients for obesity and offer intensive counseling and behavioral interventions to promote sustained weight loss. The five main steps in the evaluation of obesity, as described below, are (1) a focused obesity-related history, (2) a physical examination to determine the degree and type of obesity, (3) assessment of comorbid conditions, (4) determination of fitness level, and (5) assessment of the patient's readiness to adopt lifestyle changes.

The Obesity-Focused History

Information from the history should address the following seven questions:

What factors contribute to the patient's obesity?

How is the obesity affecting the patient's health?

What is the patient's level of risk from obesity?

What does the patient find difficult about managing weight?

What are the patient's goals and expectations?

Is the patient motivated to begin a weight management program?

What kind of help does the patient need?

Although the vast majority of cases of obesity can be attributed to behavioral factors that affect diet and physical activity patterns, the history may suggest secondary causes that merit further evaluation. Disorders to consider include polycystic ovarian syndrome, hypothyroidism, Cushing's syndrome, and hypothalamic disease. Drug-induced weight gain also should be considered. Common causes include medications for diabetes (insulin, sulfonylureas, thiazolidinediones); steroid hormones; antipsychotic agents (clozapine, olanzapine, risperidone); mood stabilizers (lithium); antidepressants (tricyclics, monoamine oxidase inhibitors, paroxetine, mirtazapine); and antiepileptic drugs (valproate, gabapentin, carbamazepine). Other medications, such as nonsteroidal anti-inflammatory drugs and calcium channel blockers, may cause peripheral edema but do not increase body fat.

The patient's current diet and physical activity patterns may reveal factors that contribute to the development of obesity and may identify behaviors to target for treatment. This type of historic information is best obtained by the combination of a questionnaire and an interview.

Body Mass Index (BMI) and Waist Circumference

Three key anthropometric measurements are important in evaluating the degree of obesity: weight, height, and waist circumference. The BMI, calculated as weight (kg)/height $(m)^2$ or as weight (lb)/height $(in)^2 \times 703$, is used to classify weight status and risk of disease (Table 395-1). BMI provides an estimate of body fat and is related to disease risk. Lower BMI thresholds for overweight and obesity have been proposed for the Asia-Pacific region since this population appears to be at risk for glucose and lipid abnormalities at lower body weights.

TABLE 395-1

Classification of Weight Status and Disease Risk

Classification	Body Mass Index (kg/m^2)	Obesity Class	Disease Risk
Underweight	<18.5	—	—
Healthy weight	18.5–24.9	—	—
Overweight	25.0–29.9	—	Increased
Obesity	30.0–34.9	I	High
Obesity	35.0–39.9	II	Very high
Extreme obesity	≥40	III	Extremely high

Source: Adapted from the National Institutes of Health, National Heart, Lung, and Blood Institute: *Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults.* U.S. Department of Health and Human Services, U.S. Public Health Service, 1998.

Excess abdominal fat, assessed by measurement of waist circumference or waist-to-hip ratio, is independently associated with a higher risk for diabetes mellitus and cardiovascular disease. Measurement of the waist circumference is a surrogate for visceral adipose tissue and should be performed in the horizontal plane above the iliac crest (Table 395-2).

TABLE 395-2

Ethnic-Specific Cutpoint Values for Waist Circumference

Ethnic Group	Waist Circumference
Europeans	
Men	>94 cm (>37 in)
Women	>80 cm (>31.5 in)
South Asians and Chinese	
Men	>90 cm (>35 in)
Women	>80 cm (>31.5 in)
Japanese	
Men	>85 cm (>33.5 in)
Women	>90 cm (>35 in)
Ethnic South and Central Americans	Use South Asian recommendations until more specific data are available.
Sub-Saharan Africans	Use European data until more specific data are available.
Eastern Mediterranean and Middle Eastern (Arab) populations	Use European data until more specific data are available.

Source: From KGMM Alberti et al for the IDF Epidemiology Task Force Consensus Group: Lancet 366:1059, 2005.

Physical Fitness

Several prospective studies have demonstrated that physical fitness, reported by questionnaire or measured by a maximal treadmill exercise test, is an important predictor of all-cause mortality rate independent of BMI and body composition. These observations highlight the importance of taking a physical activity and exercise history during examination as well as emphasizing physical activity as a treatment approach.

Obesity-Associated Comorbid Conditions

The evaluation of comorbid conditions should be based on presentation of symptoms, risk factors, and index of suspicion. For all patients, a fasting lipid panel should be performed (total, low-density lipoprotein, and high-density lipoprotein cholesterol and triglyceride levels) and a fasting blood glucose level and blood pressure determined. Symptoms and diseases that are directly or indirectly related to obesity are listed in **Table 395-3**. Although individuals vary, the number and severity of organ-specific comorbid conditions usually rise with increasing levels of obesity. Patients at very high absolute risk include those with the following: established coronary heart disease; presence of other atherosclerotic diseases, such as peripheral arterial disease, abdominal aortic aneurysm, and symptomatic carotid artery disease; type 2 diabetes; and sleep apnea.

TABLE 395-3

Obesity-Related Organ Systems Review

Cardiovascular	**Respiratory**
Hypertension	Dyspnea
Congestive heart failure	Obstructive sleep apnea
Cor pulmonale	Hypoventilation syndrome
Varicose veins	Pickwickian syndrome
Pulmonary embolism	Asthma
Coronary artery disease	**Gastrointestinal**
Endocrine	Gastroesophageal reflux disease
Metabolic syndrome	Nonalcoholic fatty-liver disease
Type 2 diabetes	Cholelithiasis
Dyslipidemia	Hernias
Polycystic ovarian syndrome	Colon cancer
Musculoskeletal	**Genitourinary**
Hyperuricemia and gout	Urinary stress incontinence
Immobility	Obesity-related glomerulopathy
Osteoarthritis (knees and hips)	Hypogonadism (male)
Low back pain	Breast and uterine cancer
Carpal tunnel syndrome	Pregnancy complications
Psychological	**Neurologic**
Depression/low self-esteem	Stroke
Body image disturbance	Idiopathic intracranial hypertension
Social stigmatization	Meralgia paresthetica
Integument	Dementia
Striae distensae	
Stasis pigmentation of legs	
Lymphedema	
Cellulitis	
Intertrigo, carbuncles	
Acanthosis nigricans	

Acrochordons (skin tags)	
Hidradenitis suppurativa	

Identifying the High-Risk Patient

Efforts are under way to develop more practical and useful assessments to identify patients who are at high risk in addition to using BMI alone. Analogous to other staging systems commonly used for congestive heart failure or chronic kidney disease, the American Society of Clinical Endocrinologists (AACE) and the American College of Endocrinology (ACE) have proposed an obesity disease staging system that is based on ethnic-specific BMI cutoffs in conjunction with assessment for adiposity-related complications. Stage 0 is assigned to individuals who are overweight or obese by BMI classification but have no complications, whereas stages 1 and 2 are defined as individuals who are overweight or obese by BMI classification and have one or more mild-moderate complications (stage 1) or at least one severe complication (stage 2). A different functional staging system for obesity, called the Edmonton Obesity Staging System (EOSS), classifies individuals with obesity into five graded categories (0–4), based on their morbidity and health-risk profile along three domains—medical, functional, and mental. In this system, staging occurs independent of BMI.

Assessing the Patient's Readiness to Change

An attempt to initiate lifestyle changes when the patient is not ready usually leads to frustration and may hamper future weight-loss efforts. Assessment includes patient motivation and support, stressful life events, psychiatric status, time availability and constraints, and appropriateness of goals and expectations. Readiness can be viewed as the balance of two opposing forces: (1) motivation, or the patient's desire to change; and (2) resistance, or the patient's resistance to change.

A helpful method to begin a readiness assessment is to use the motivational interviewing technique of "anchoring" the patient's interest and confidence to change on a numerical scale. With this technique, the patient is asked to rate—on a scale from 0 to 10, with 0 being not so important (or confident) and 10 being very important (or confident)—his or her level of interest in and confidence about losing weight at this time. This exercise helps establish readiness to change and also serves as a basis for further dialogue.

TREATMENT

TREATMENT

Obesity

THE GOAL OF THERAPY

The primary goals of treatment are to improve obesity-related comorbid conditions and reduce the risk of developing future comorbidities. Information obtained from the history, physical examination, and diagnostic tests is used to determine risk and develop a treatment plan **(Fig. 395-1)**. The decision of how aggressively to treat the patient and which modalities to use is determined by the patient's risk status, expectations, and available resources. Not all patients who are deemed obese by BMI alone need to be treated, as exemplified by the concepts of obesity paradox or the metabolically healthy obese. However, patients who present with obesity-related comorbidities and who would benefit from weight-loss intervention should be managed proactively. Therapy for obesity always begins with lifestyle management and may include pharmacotherapy or surgery, depending on BMI risk category **(Table 395-4)**. Setting an initial weight-loss goal of 8–10% over 6 months is a realistic target.

LIFESTYLE MANAGEMENT

Obesity care involves attention to three essential elements of lifestyle: dietary habits, physical activity, and behavior modification. Because obesity is fundamentally a disease of energy imbalance, all patients must learn how and when energy is consumed (diet), how and when energy is expended (physical activity), and how to incorporate this information into their daily lives (behavioral therapy). Lifestyle management has been shown to result in a modest (typically 3–5 kg) weight loss when compared with no treatment or usual care.

Diet Therapy

The primary focus of diet therapy is to reduce overall calorie consumption. Guidelines from the American Heart Association/American College of Cardiology/The Obesity Society (AHA/ACC/TOS) recommend initiating treatment with a calorie deficit of 500–750 kcal/d compared with the patient's habitual diet. Alternatively, a diet of 1200–1500 kcal/d for women and 1500–1800 kcal/d for men (adjusted for the individual's body weight) can be prescribed. This reduction is consistent with a goal of losing ~1–2 lb/week. The calorie deficit can be instituted through dietary substitutions or alternatives. Examples include choosing smaller portion sizes, eating more fruits and vegetables, consuming more whole-grain cereals, selecting leaner cuts of meat and skimmed dairy products, reducing consumption of fried foods and other foods with added fats and oils, and drinking water instead of sugar-sweetened beverages. It is important that dietary counseling remains patient centered and that the selected goals are SMART (specific, measurable, agreed upon, realistic, timely).

The macronutrient composition of the diet will vary with the patient's preference and medical condition. The 2015 U.S. Department of Agriculture Dietary Guidelines for Americans **(Chap. 325)**, which focus on health promotion and risk reduction, can be applied to treatment of patients who are overweight or obese. The recommendations include maintaining a diet rich in whole grains, fruits, vegetables, and dietary fiber; decreasing sodium intake to <2300 mg/d; consuming fat-free or low-fat dairy products; and keeping added sugars and saturated fat intake to <10% of daily calories. Application of these guidelines to specific calorie goals can be found on the website *www.choosemyplate.gov*. Since portion control is one of the most difficult strategies for patients to manage, the use of preprepared products such as meal replacements is a simple and convenient suggestion. Examples include frozen entrees, canned beverages, and bars. Use of meal replacements in the diet has been shown to result in a 7–8% weight loss.

Numerous randomized trials comparing diets of different macronutrient composition (e.g., low-carbohydrate, low-fat, Mediterranean) have shown that weight loss depends primarily on reduction of total caloric intake and adherence to the prescribed diet, not the specific proportions of carbohydrate, fat, and protein in the diet. The macronutrient composition will ultimately be determined by the patient's taste preferences, cooking style, and culture. However, the patient's underlying medical problems are also important in guiding the recommended dietary composition. The dietary prescription will vary according to the patient's metabolic profile and risk factors. A consultation with a registered dietitian for medical nutrition therapy is particularly useful in considering patient preference and treatment of comorbid diseases.

Another dietary approach to consider is based on the concept of *energy density*, which refers to the number of calories (i.e., amount of energy) a food contains per unit of weight. People tend to ingest a constant volume of food regardless of caloric or macronutrient content. Adding water or fiber to a food decreases its energy density by increasing weight without affecting caloric content. Examples of foods with low-energy density include soups, fruits, vegetables, oatmeal, and lean meats. Dry foods and high-fat foods such as pretzels, cheese, egg yolks, potato chips, and red meat have a high-energy density. Diets containing low-energy-dense foods have been shown to control hunger and thus to result in decreased caloric intake and weight loss.

Occasionally, very low-calorie diets (VLCDs) are prescribed as a form of aggressive dietary therapy. The primary purpose of a VLCD is to promote a rapid and significant (13- to 23-kg) short-term weight loss over a 3- to 6-month period. The proprietary formulas designed for this purpose typically supply ≤800 kcal, 50–80 g of protein, and 100% of the recommended daily intake for vitamins and minerals. According to a review by the National Task Force on the Prevention and Treatment of Obesity, indications for initiating a VLCD include the involvement of well-motivated individuals who are moderately to severely obese (BMI, >30 kg/m^2), have failed at more conservative approaches to weight loss, and have a medical condition that would be immediately improved with rapid weight loss. These conditions include poorly controlled type 2 diabetes, hypertriglyceridemia, obstructive sleep apnea, and symptomatic peripheral edema. The risk for gallstone formation increases exponentially at rates of weight loss >1.5 kg/week (3.3 lb/week). Prophylaxis against gallstone formation with ursodeoxycholic acid (600 mg/d) is effective in reducing this risk. VLCDs should be used only in limited circumstances and only when provided by trained practitioners in a medical care setting where medical monitoring and high intensity lifestyle intervention can be provided. Medical supervision is required because of the rapid rate of weight loss and potential for health complications.

Physical Activity Therapy

Although exercise alone is only moderately effective for weight loss, the combination of dietary modification and exercise is the most effective behavioral approach for the treatment of obesity. The most important role of exercise appears to be in the maintenance of the weight loss. The 2008 Physical Activity Guidelines for Americans (*www.health.gov/paguidelines*) recommend that adults should engage in 150 min of moderate-intensity or 75 min a week of vigorous-intensity aerobic physical activity per week, performed in episodes of at least 10 min and preferably spread throughout the week. Focusing on simple ways to add physical activity into the normal daily routine through leisure activities, travel, and domestic work should be suggested. Examples include brisk walking, using the stairs, doing housework and yard work, and engaging in sports. Asking the patient to wear a pedometer or accelerometer to monitor total accumulation of steps or kcal expended as part of the activities of daily living is a useful strategy. Step counts are highly correlated with activity level. Studies have demonstrated that lifestyle activities are as effective as structured exercise programs for improving cardiorespiratory fitness and weight loss. A high level of physical activity (>300 min of moderate-intensity activity per week) is often needed to lose weight and sustain weight loss. These exercise recommendations are daunting to most patients and need to be implemented gradually. Consultation with an exercise physiologist or personal trainer may be helpful.

Behavioral Therapy

Cognitive behavioral therapy is used to help change and reinforce new dietary and physical activity behaviors. Strategies include self-monitoring techniques (e.g., journaling, weighing, and measuring food and activity); stress management; stimulus control (e.g., using smaller plates, not eating in front of the television or in the car); social support; problem solving; and cognitive restructuring to help patients develop more positive and realistic thoughts about themselves. When recommending any behavioral lifestyle change, the patient should be asked to identify what, when, where, and how the behavioral change will be performed. The patient should keep a record of the anticipated behavioral change so that progress can be reviewed at the next office visit. Because these techniques are time consuming to implement, their supervision is often undertaken by ancillary office staff, such as a nurse-clinician or registered dietitian.

PHARMACOTHERAPY

Adjuvant pharmacologic treatments should be considered for patients with a BMI ≥30 kg/m^2 or for patients with a BMI ≥27 kg/m^2 who have concomitant obesity-related diseases and for whom dietary and physical activity therapy has not been successful. When an antiobesity medication is prescribed, patients should be actively engaged in a lifestyle program that provides the strategies and skills needed to use the drug effectively, since such support increases total weight loss.

Medications for obesity have traditionally fallen into two major categories: appetite suppressants (*anorexiants*) and gastrointestinal fat blockers.. Four new antiobesity medications have been approved by the U.S. Food and Drug Administration (FDA) since 2012: lorcaserin, phentermine/topiramate (PHEN/TPM) extended release, naltrexone sustained release (SR)/bupropion SR, and liraglutide. Gastrointestinal fat blockers reduce the absorption of selective macronutrients, such as fat, from the gastrointestinal tract.

Centrally Acting Anorexiant Medications

Anorexiants affect *satiety* (the absence of hunger after eating) and hunger (the biologic sensation that prompts eating). By increasing satiety and decreasing hunger, these agents help patients reduce caloric intake without a sense of deprivation. The target site for the actions of anorexiants is the ventromedial and lateral hypothalamic regions in the central nervous system **(Chap. 394)**. The biologic effect of these agents on appetite regulation is produced by augmentation of the neurotransmission of three monoamines: norepinephrine; serotonin (5-hydroxytryptamine [5-HT]); and, to a lesser degree, dopamine. The classic sympathomimetic adrenergic agents (benzphetamine, phendimetrazine, diethylpropion, mazindol, and phentermine) function by stimulating norepinephrine release or by blocking its reuptake. Among the anorexiants, phentermine is the most commonly prescribed; there are limited long-term data on its effectiveness. A 2002 review of six randomized, placebo-controlled trials of phentermine for weight control found that patients lost 0.6–6.0 additional kg of weight over 2–24 weeks of treatment. The most common side effects of the amphetamine-derived anorexiants are restlessness, insomnia, dry mouth, constipation, and increased blood pressure and heart rate.

PHEN/TPM is a combination drug that contains a catecholamine releaser (phentermine) and an anticonvulsant (topiramate). Topiramate is approved by the FDA as an anticonvulsant for the treatment of epilepsy and for the prophylaxis of migraine headaches. Weight loss was identified as an unintended side effect of topiramate during clinical trials for epilepsy. The mechanism responsible for weight loss is uncertain but is thought to be mediated through the drug's modulation of γ-aminobutyric acid receptors, inhibition of carbonic anhydrase, and antagonism of glutamate. PHEN/TPM has undergone two 1-year pivotal randomized, placebo-controlled, double-blind trials of efficacy and safety: EQUIP and CONQUER. In a third study, SEQUEL, 78% of CONQUER participants continued to receive their blinded treatment for an additional year. All participants received diet and exercise counseling. Participant numbers, eligibility, characteristics, and weight-loss outcomes are displayed in **Table 395-5**. Intention-to-treat 1-year placebo-subtracted weight loss for PHEN/TPM was 9.3% (15-mg/92-mg dose) and 6.6% (7.5-mg/46-mg dose), respectively, in the EQUIP and CONQUER trials. Clinical and statistical dose-dependent improvements were seen in selected cardiovascular and metabolic outcome measurements that were related to the weight loss. The most common adverse events experienced by the drug-randomized group were paresthesias, dry mouth, constipation, dysgeusia, and insomnia. Because of an increased risk of congenital fetal oral-cleft formation from topiramate, women of childbearing age should have a negative pregnancy test before treatment and monthly thereafter, and use effective contraception consistently during medication therapy.

Lorcaserin is a selective 5-$HT2_C$ receptor agonist with a functional selectivity ~15 times that of 5-$HT2_A$ receptors and 100 times that of 5-$HT2_B$ receptors. This selectivity is important, since the drug-induced valvulopathy documented with two other serotonergic agents that were removed from the market—fenfluramine and dexfenfluramine—was due to activation of the 5-$HT2_B$ receptors expressed on cardiac valvular interstitial cells. By activating the 5-$HT2_C$ receptor, lorcaserin is thought to decrease food intake through the pro-opiomelanocortin (POMC) system of neurons.

Lorcaserin has undergone two randomized, placebo-controlled, double-blind trials for efficacy and safety. Participants were randomized to receive lorcaserin (10 mg bid) or placebo in the BLOOM study and lorcaserin (10 mg bid or qd) or placebo in the BLOSSOM study. All participants received diet and exercise counseling. Participant numbers, eligibility, characteristics, and weight-loss outcomes are displayed in Table 395-5. Patients who were overweight or obese had at least one coexisting condition (hypertension, dyslipidemia, cardiovascular disease, impaired glucose tolerance, or sleep apnea)—medical conditions that are commonly seen in the office setting. Intention-to-treat 1-year placebo-subtracted weight loss was 3.6% and 3.0%, respectively, in the BLOOM and BLOSSOM trials. Echocardiography was performed at the screening visit and at scheduled time points over the course of the studies. There was no difference in the development of FDA-defined valvulopathy between drug-treated and placebo-treated participants at 1 or 2 years. Modest statistical improvements consistent with the weight loss were seen in selected cardiovascular and metabolic outcome measurements. The most common adverse events experienced by the drug group were headache, dizziness, and nausea.

Naltrexone SR/bupropion SR (NB) is a combination of an opioid antagonist and a mild reuptake inhibitor of dopamine and norepinephrine, respectively. Individually, naltrexone is approved by the FDA for the treatment of alcohol dependence and for the blockade of the effects of exogenously administered opioids, whereas bupropion is approved as an antidepressant and smoking cessation aid. As a combination drug, each component works in consort: bupropion stimulates secretion of α-melanocyte stimulating hormone (MSH) from POMC whereas naltrexone blocks the feedback inhibitory effects of opioid receptors activated by the β-endorphin released in the hypothalamus, thus allowing the inhibitory effects of MSH to reduce food intake.

The medication has undergone three randomized, placebo-controlled, double-blind trials for efficacy and safety. Participants were randomized to receive NB (8 mg/90 mg two tablets bid) or placebo in the three COR studies. Whereas participants received standardized nutritional and exercise counseling in COR-I and COR-II, a more intensive behavior modification program was provided in COR-BMOD (Table 395-5). Intention-to-treat 1-year placebo-subtracted weight loss was 4.8%, 5.1%, and 4.2%, respectively, in the COR-I, COR-II, and COR-BMOD trials. Clinical and statistical dose-dependent improvements were seen in selected cardiovascular and metabolic outcome measurements that were related to the weight loss. However, the medication led to slight increased or smaller decreases in blood pressure and pulse than placebo. The most common adverse events experienced by the drug-randomized groups were nausea, constipation, headache, vomiting, dizziness, diarrhea, insomnia, and dry mouth.

Liraglutide, the fourth new medication, is a glucagon-like peptide-1 (GLP-1) analogue with 97% homology to human GLP-1 that was previously approved for the treatment of type 2 diabetes at doses up to 1.8 mg once daily. In addition to its effect as an incretin hormone (glucose-induced insulin secretion), liraglutide inhibits both gastric emptying and glucagon secretion and stimulates GLP-1 receptors in the arcuate nucleus of the hypothalamus to reduce feeding.

Liraglutide has undergone three randomized, placebo-controlled, double-blind trials for efficacy and safety. Participants were randomized to receive liraglutide (3.0 mg sc daily) or placebo for initial weight loss—SCALE (patients without diabetes) and SCALE Diabetes (patients with diabetes), or for weight maintenance after initial weight loss (SCALE Maintenance) (Table 395-5). All participants received diet and exercise counseling. For SCALE and SCALE Maintenance, patients were overweight or obese and had treated or untreated hypertension or dyslipidemia. Intention-to-treat 1-year placebo-subtracted weight loss was 5.4%, and 6.1%, respectively, in the SCALE and SCALE Maintenance trials. Clinical and statistical dose-dependent improvements were seen in selected cardiovascular and metabolic outcome measurements; however, there is a small increase in heart rate. The most common adverse effects include nausea, diarrhea, constipation, and vomiting. GLP-1 agonists should not be prescribed in patients with a family or personal history of medullary thyroid cancer or multiple endocrine neoplasia.

In approving the four new antiobesity medications, the FDA introduced a new provision with important clinical relevance: a prescription trial period to assess effectiveness. Response to these medications should be assessed after 12 weeks of treatment for PHEN/TPM and lorcaserin (or 16 weeks for naltrexone SR/bupropion SR and liraglutide since these medications are uptitrated during the first month). Determining responsiveness at 3 or 4 months is based on the post hoc observed trial data that patients who did not lose a prespecified amount of weight early in treatment were less successful at 1 year. For PHEN/TPM, if the patient has not lost at least 3% of body weight at 3 months, the clinician can either escalate the dose and reassess progress at 6 months or discontinue treatment entirely. For lorcaserin and naltrexone SR/bupropion SR, the medication should be discontinued if the patient has not lost at least 5% of body weight. The corresponding responsive target for liraglutide is a 4% weight loss.

Peripherally Acting Medications

Orlistat (Xenical^{TM}) is a synthetic hydrogenated derivative of a naturally occurring lipase inhibitor, lipostatin, that is produced by the mold *Streptomyces toxytricini*. This drug is a potent, slowly reversible inhibitor of pancreatic, gastric, and carboxylester lipases and phospholipase A_2, which are required for the hydrolysis of dietary fat into fatty acids and monoacylglycerols. Orlistat acts in the lumen of the stomach and small intestine by forming a covalent bond with the active site of these lipases. Taken at a therapeutic dose of 120 mg tid, orlistat blocks the digestion and absorption of ~30% of dietary fat. After discontinuation of the drug, fecal fat content usually returns to normal within 48–72 h.

Multiple randomized, double-blind, placebo-controlled studies have shown that, after 1 year, orlistat produces a weight loss of ~9–10%, whereas placebo recipients have a 4–6% weight loss. Because orlistat is minimally (<1%) absorbed from the gastrointestinal tract, it has no systemic side effects. The drug's tolerability is related to the malabsorption of dietary fat and the subsequent passage of fat in the feces. Adverse gastrointestinal effects, including flatus with discharge, fecal urgency, fatty/oily stool, and increased defecation, are reported in at least 10% of orlistat-treated patients. These side effects generally are experienced early, diminish as patients control their dietary fat intake, and only infrequently cause patients to withdraw from clinical trials. When taken concomitantly, psyllium mucilloid is helpful in controlling orlistat-induced gastrointestinal side effects. Because serum concentrations of the fat-soluble vitamins D and E and β-carotene may be reduced by orlistat treatment, vitamin supplements are recommended to prevent potential deficiencies. Orlistat was approved for over-the-counter use in 2007.

SURGERY

Bariatric surgery **(Fig. 395-2)** can be considered for patients with severe obesity (BMI, ≥40 kg/m^2) or for those with moderate obesity (BMI, ≥35 kg/m^2) associated with a serious medical condition. Weight-loss surgeries have traditionally been classified into three categories on the basis of anatomic changes: restrictive, restrictive malabsorptive, and malabsorptive. More recently, however, the clinical benefits of bariatric surgery in achieving weight loss and alleviating metabolic comorbidities have been attributed largely to changes in the physiologic responses of gut hormones, bile acid metabolism, the microbiota, and in adipose tissue metabolism. Metabolic effects resulting from bypassing the foregut include altered responses of ghrelin, glucagon-like peptide 1, peptide YY3-36, and oxyntomodulin. Additional effects on food intake and body weight control may be attributed to changes in vagal signaling. The loss of fat mass, particularly visceral fat, is associated with multiple metabolic, adipokine, and inflammatory changes that include improved insulin sensitivity and glucose disposal; reduced free fatty acid flux; increased adiponectin levels; and decreased interleukin 6, tumor necrosis factor α, and high-sensitivity C-reactive protein levels.

Restrictive surgeries limit the amount of food the stomach can hold and slow the rate of gastric emptying. *Laparoscopic adjustable gastric banding* is the prototype of this category. The first banding device, the LAP-BAND, was approved for use in the United States in 2001 and the second, the REALIZE band, in 2007. In contrast to previous devices, these bands have diameters that are adjustable by way of their connection to a reservoir that is implanted under the skin. Injection of saline into the reservoir and removal of saline from the reservoir tighten and loosen the band's internal diameter, respectively, thus changing the size of the gastric opening. Although the mean percentage of total body weight lost at 5 years is estimated at 20–25%, longer-term follow-up has been more disappointing leading to near abandonment of the procedure. In the *laparoscopic sleeve gastrectomy*, the stomach is restricted by stapling and dividing it vertically, removing ~80% of the greater curvature and leaving a slim banana-shaped remnant stomach along the lesser curvature. Weight loss after this procedure is superior to that after laparoscopic adjustable gastric banding.

The three restrictive-malabsorptive bypass procedures combine the elements of gastric restriction and selective malabsorption: Roux-en-Y gastric bypass, biliopancreatic diversion, and biliopancreatic diversion with duodenal switch (Fig. 395-2). Roux-en-Y is the most commonly undertaken and most accepted bypass procedure. They are routinely performed by laparoscopy.

These procedures generally produce a 30–35% average total body weight loss that is maintained in ~60% of patients at 5 years. Significant improvement in multiple obesity-related comorbid conditions, including type 2 diabetes, hypertension, dyslipidemia, obstructive sleep apnea, quality of life, and long-term cardiovascular events, has been reported. A meta-analysis of controlled clinical trials comparing bariatric surgery versus no surgery showed that surgery was associated with a reduced odds ratio (OR) risk of global mortality (OR = 0.55), cardiovascular death (OR = 0.58), and all-cause mortality (OR = 0.70).

Among the observed improvements in comorbidities, the prevention and treatment of type 2 diabetes resulting from bariatric surgery has garnered the most attention. Fifteen-year data from the Swedish Obese Subjects study demonstrated a marked reduction (i.e., by 78%) in the incidence of type 2 diabetes development among obese patients who underwent bariatric surgery. Several randomized controlled studies have shown greater weight loss and more improved glycemic control at 1 and 3 years among surgical patients than among patients receiving conventional medical therapy. A retrospective cohort study of >4000 adults with diabetes found that overall 68.2% of patients experienced an initial complete type 2 diabetes remission within 5 years after surgery. However, among these patients, one-third redeveloped type 2 diabetes within 5 years. The rapid improvement seen in diabetes after restrictive-malabsorptive procedures is thought to be due to caloric restriction, reduced insulin resistance, and surgery-specific effects on glucose homeostasis brought about by alteration of gut hormones.

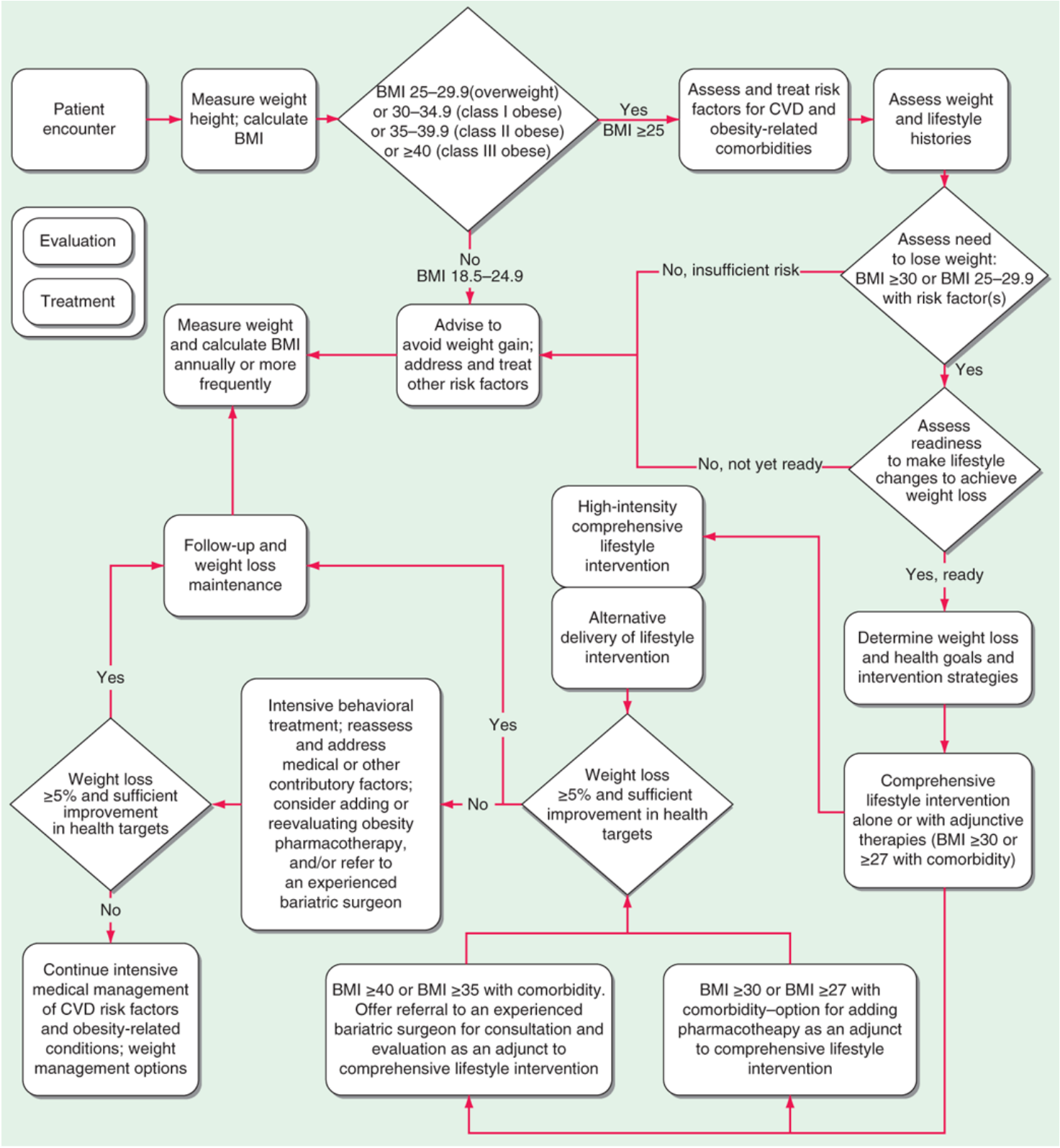
The mortality rate from bariatric surgery is generally <1% but varies with the procedure, the patient's age and comorbid conditions, and the experience of the surgical team. The most common surgical complications include stomal stenosis or marginal ulcers (occurring in 5–15% of patients) that present as prolonged nausea and vomiting after eating or inability to advance the diet to solid foods. These complications typically are treated by endoscopic balloon dilation and acid suppression therapy, respectively. For patients who undergo laparoscopic adjustable gastric banding, there are no intestinal absorptive abnormalities other than mechanical reduction in gastric size and outflow. Therefore, selective deficiencies are uncommon unless eating habits become unbalanced. In contrast, the restrictive-malabsorptive procedures carry an increased risk for micronutrient deficiencies of vitamin B_{12}, iron, folate, calcium, and vitamin D. Patients with restrictive-malabsorptive procedures require lifelong supplementation with these micronutrients.

Intraluminal Gastric Balloons

Recently, the FDA approved two gastric balloon devices for weight loss that are placed in the stomach endoscopically. The RESHAPE device consists of two silicone balloons attached to a central silicone shaft, whereas the ORBERA is a single-balloon device. Mean weight loss of 7.2 kg and 8.8 kg, respectively, was seen for these devices in short-term pivotal trials. Both systems are approved only for up to 6 months of use in adults with a BMI of 30–40 kg/m^2. Adverse effects include nausea, vomiting, and abdominal pain.

FIGURE 395-1

Treatment algorithm—chronic disease management model for primary care of patients with overweight and obesity. This algorithm applies to the assessment of overweight and obesity and subsequent decisions based on that assessment. BMI indicates body mass index; CVD, cardiovascular disease; FDA, U.S. Food and Drug Administration. *(From Jensen MD et al: 2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society. Circulation. 129[suppl 2]:S102, 2014.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 395-4

A Guide to Opting for Treatment for Obesity

	BMI Category (kg/m^2)				
Treatment	**25–26.9**	**27–29.9**	**30–34.9**	**35–39.9**	**≥40**
Diet, exercise, behavioral therapy	With comorbidities	With comorbidities	+	+	+
Pharmacotherapy	—	With comorbidities	+	+	+
Surgery	—	—	—	With comorbidities	+

Source: From the National Heart, Lung, and Blood Institute, North American Association for the Study of Obesity (2000).

TABLE 395-5

Clinical Trials for Antiobesity Medications

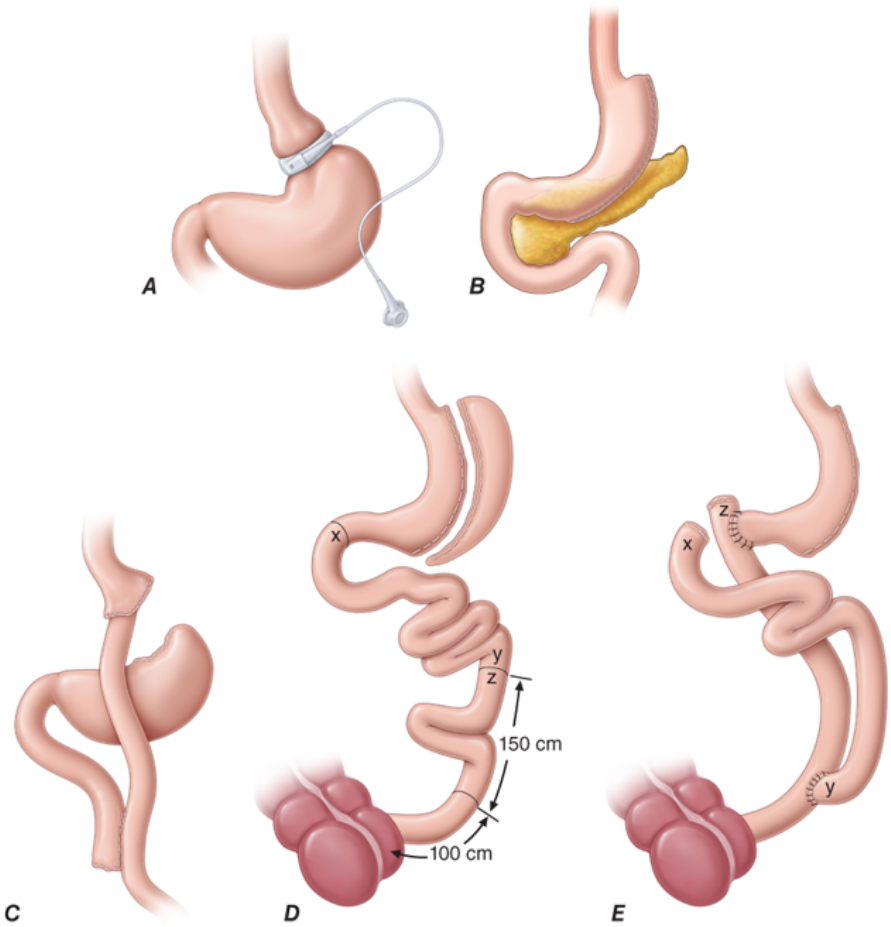
	PHEN/TPM		Lorcaserin		Naltrexone SR/bupropion SR			Liraglutide	
	EQUIP	**CONQUER**	**BLOOM**	**BLOSSOM**	**COR-I**	**COR-II**	**COR-BMOD**	**SCALE**	**SCALE Maintenance**
No. of participants (ITT-LOCF)	1230	2487	3182	4008	1742	1496	793	3731	422
BMI (kg/m^2)	≥35	27–45	27–45	30–45	30–45	30–45	30–45	≥27	≥27
Age (y)	18–70	18–70	18–65	18–65	18–65	18–65	18–65	≥18	≥18
Comorbid conditions (cardiovascular and metabolic)	≥1	≥2	≥1	≥1	≥1	≥1	≥1	≥1	≥1
Mean weight loss (%) with treatment vs placebo	10.9 vs 1.6	7.8 vs 1.2	5.8 vs 2.2	4.8 vs 2.8	6.1 vs 1.3	6.5 vs 1.9	9.3 vs 5.1	8.0 vs 2.6	6.2 vs 0.2
Placebo-subtracted weight loss (%)	9.3	6.6	3.6	3.0	4.8	4.6	4.2	5.4	6.0
Categorical change in 5% weight loss with treatment vs placebo	66.7 vs 17.3	62 vs 21	47.5 vs 20.3	47.2 vs 25	48 vs 16	50.5 vs 17.1	66.4 vs 42.5	63.2 vs 27.1	81.4 vs 48.9
Study completion rate, treatment vs placebo (%)	66.4 vs 52.9	69 vs 57	55.4 vs 45.1	57.2 vs 52	50	54	57.9 vs 58.4	71.9 vs 64.4	75 vs 69.5

Note: EQUIP (PHEN/TPM = 15/92 mg dose; CONQUER (PHEN/TPM = 7.5/46 mg dose).

Abbreviations: BMI, body mass index; ITT-LOCF, intention to treat, last observation carried forward; PHEN-TPM, phentermine-topiramate extended release.

FIGURE 395-2

Bariatric surgical procedures. Examples of operative interventions used for surgical manipulation of the gastrointestinal tract. ***A.*** Laparoscopic adjustable gastric banding. ***B.*** Laparoscopic sleeve gastrectomy. ***C.*** The Roux-en-Y gastric bypass. ***D.*** Biliopancreatic diversion with duodenal switch. ***E.*** Biliopancreatic diversion. *(From ML Kendrick, GF Dakin: Mayo Clin Proc 815:518, 2006; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Apovian CM et al: Pharmacological management of obesity: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 100:342, 2015. [PubMed: 25590212]

Garvey WT et al: American Association of Clinical Endocrinologists and American College of Endocrinology Comprehensive Clinical Practice Guidelines for Medical Care of Patients with Obesity. Endocrine Practice 22 (suppl 3):1, 2016. [PubMed: 27219496]

Jensen MD et al: 2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society. Circulation 129 (suppl 2):S102, 2014. [PubMed: 24222017]

Chapter 396: Diabetes Mellitus: Diagnosis, Classification, and Pathophysiology

Alvin C. Powers; Kevin D. Niswender; Carmella Evans-Molina

INTRODUCTION

Diabetes mellitus (DM) refers to a group of common metabolic disorders that share the phenotype of hyperglycemia. Several distinct types of DM are caused by a complex interaction of genetics and environmental factors. Depending on the etiology of the DM, factors contributing to hyperglycemia include reduced insulin secretion, decreased glucose utilization, and increased glucose production. The metabolic dysregulation associated with DM causes secondary pathophysiologic changes in multiple organ systems that impose a tremendous burden on the individual with diabetes and on the health care system. In the United States, DM is the leading cause of end-stage renal disease (ESRD), nontraumatic lower extremity amputations, and adult blindness. It also predisposes to cardiovascular diseases. With an increasing incidence worldwide, DM is likely to continue to be a leading cause of morbidity and mortality in the future.

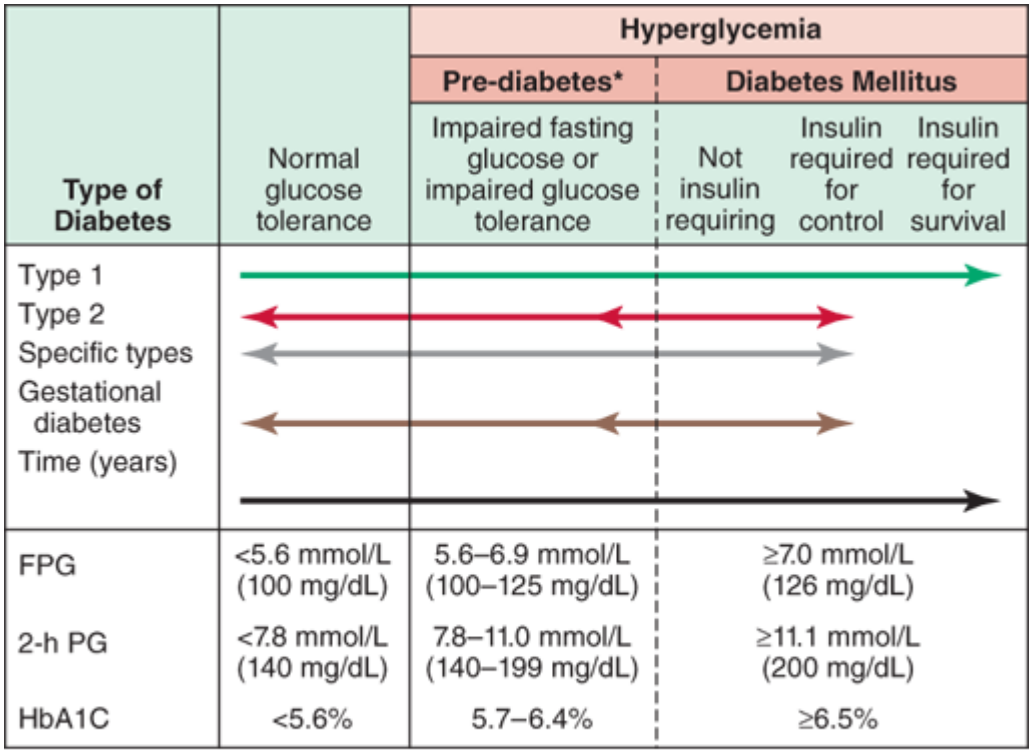
CLASSIFICATION

DM is classified on the basis of the pathogenic process leading to hyperglycemia, as opposed to earlier criteria such as age of onset or type of therapy **(Fig. 396-1)**. There are two broad categories of DM, designated as either type 1 or type 2 DM **(Table 396-1)**. However, there is increasing recognition of other forms of diabetes in which the molecular pathogenesis is better understood and may be associated with a single gene defect. These alternative forms may share features of type 1 and/or type 2 DM. Type 1 DM develops as a result of autoimmunity against the insulin-producing beta cells, resulting in complete or near-total insulin deficiency. Type 2 DM is a heterogeneous group of disorders characterized by variable degrees of insulin resistance, impaired insulin secretion, and increased hepatic glucose production. Distinct genetic and metabolic defects in insulin action and/or secretion give rise to the common phenotype of hyperglycemia in type 2 DM and have important therapeutic implications now that pharmacologic agents are available to target specific metabolic derangements. Both type 1 and type 2 diabetes are preceded by a period of progressive worsening of glucose homeostasis, followed by the development of hyperglycemia that exceeds the threshold for clinical diagnosis. In terms of type 2 diabetes, this phase is referred to as prediabetes and are more specifically classified as impaired fasting glucose (IFG) or impaired glucose tolerance (IGT) (Fig. 396-1). Recently, three distinct stages of type 1 DM have been defined based on the development of

autoantibodies against pancreatic beta cell antigens or the development of worsening dysglycemia (discussed below).

FIGURE 396-1

Spectrum of glucose homeostasis and diabetes mellitus (DM). The spectrum from normal glucose tolerance to diabetes in type 1 DM, type 2 DM, specific types of diabetes, and gestational DM is shown from left to right. In most types of DM, the individual traverses from normal glucose tolerance to impaired glucose tolerance to overt diabetes (these should be viewed not as abrupt categories but as a spectrum). *Arrows* indicate that changes in glucose tolerance may be bidirectional in some types of diabetes. For example, individuals with type 2 DM may return to the impaired glucose tolerance category with weight loss; in gestational DM, diabetes may revert to impaired glucose tolerance or even normal glucose tolerance after delivery. The fasting plasma glucose (FPG), the 2-h plasma glucose (PG) after a glucose challenge, and the hemoglobin A_{1c} (HbA_{1c}) for the different categories of glucose tolerance are shown at the lower part of the figure. These values do not apply to the diagnosis of gestational DM. *Some use the term *increased risk for diabetes* or *intermediate hyperglycemia* (World Health Organization) rather than *prediabetes*. *(Adapted from the American Diabetes Association, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 396-1

Etiologic Classification of Diabetes Mellitus

I. Type 1 diabetes (immune-mediated beta cell destruction, usually leading to absolute insulin deficiency)
II. Type 2 diabetes (may range from predominantly insulin resistance with relative insulin deficiency to a predominantly insulin secretory defect with insulin resistance)
III. Specific types of diabetes
  A. Genetic defects of beta cell development or function characterized by mutations in:
    1. Hepatocyte nuclear transcription factor (HNF) 4α (MODY 1)
    2. Glucokinase (MODY 2)
    3. HNF-1α (MODY 3)
    4. Insulin promoter factor-1, HNF-1β, NeuroD1, and others leading to other forms of MODY
    5. Insulin, subunits of ATP-sensitive potassium channel leading to permanent neonatal diabetes
    6. Mitochondrial DNA
    7. Other pancreatic islet regulators/proteins such as *KLF11, PAX4, BLK, GATA4, GATA6, SLC2A2* (GLUT2), *RFX6, GLIS3*
  B. Transient neonatal diabetes
  C. Diseases of the exocrine pancreas—pancreatitis, pancreatectomy, neoplasia, cystic fibrosis, hemochromatosis, fibrocalculous pancreatopathy, mutations in carboxyl ester lipase
  D. Genetic defects in insulin action, including type A insulin resistance, Leprechaunism, Rabson-Mendenhall syndrome, Lipodystrophy syndromes
  E. Endocrinopathies—acromegaly, Cushing's syndrome, glucagonoma, pheochromocytoma, hyperthyroidism, somatostatinoma, aldosteronoma
  F. Drug- or chemical-induced—glucocorticoids, vacor (a rodenticide), pentamidine, nicotinic acid, diazoxide, β-adrenergic agonists, thiazides, calcineurin and mTOR inhibitors, hydantoins, asparaginase, α-interferon, protease inhibitors, antipsychotics (atypicals and others), epinephrine
  G. Infections—congenital rubella, cytomegalovirus, coxsackievirus
  H. Uncommon forms of immune-mediated diabetes—"stiff-person" syndrome, anti-insulin receptor antibodies
  I. Other genetic syndromes sometimes associated with diabetes—Wolfram's syndrome, Down's syndrome, Klinefelter's syndrome, Turner's syndrome, Friedreich's ataxia, Huntington's chorea, Laurence-Moon-Biedl syndrome, myotonic dystrophy, porphyria, Prader-Willi syndrome
IV. Gestational diabetes mellitus (GDM)

Abbreviation: MODY, maturity-onset diabetes of the young.

Source: Adapted from American Diabetes Association: Diabetes Care 37(Suppl 1):S14, 2014 and updated from American Diabetes Association: Diabetes Care 40(Suppl 1):S11–24, 2017.

Two features of the current classification of DM merit emphasis from previous classifications. First, the terms *insulin-dependent diabetes mellitus* (IDDM) or *non-insulin-dependent diabetes mellitus* (NIDDM) are no longer used. Because many individuals with type 2 DM eventually require insulin treatment for control of glycemia, the use of the term NIDDM generated considerable confusion. A second difference is that age or treatment modality is not a criterion. Although type 1 DM most commonly develops before the age of 30, autoimmunity against beta cells can develop at any age. It is estimated that between 5 and 10% of individuals who develop DM after age 30 have type 1 DM. Although type 2 DM more typically develops with increasing age, it is now being diagnosed more frequently in children and young adults, particularly in obese adolescents.

OTHER TYPES OF DM

Other etiologies of DM include specific genetic defects in insulin secretion or action, metabolic abnormalities that impair insulin secretion, mitochondrial abnormalities, and a host of conditions that impair glucose tolerance (Table 396-1). *Maturity-onset diabetes of the young* (MODY) and *monogenic diabetes* are subtypes of DM characterized by autosomal dominant inheritance, early onset of hyperglycemia (usually <25 years; sometimes in neonatal period), and impaired insulin secretion (discussed below). Mutations in the insulin receptor cause a group of rare disorders characterized by severe insulin resistance.

DM may also develop as a result of cystic fibrosis or chronic pancreatitis, in which the islets become damaged from a primary pathologic process originating in the pancreatic exocrine tissue. Hormones that antagonize insulin action can lead to DM. Thus, DM is often a feature of endocrinopathies such as acromegaly and Cushing's disease. Viral infections have been implicated in pancreatic islet destruction but are an extremely rare cause of DM. A form of acute onset of type 1 diabetes, termed *fulminant diabetes*, has been noted in Japan and may be related to viral infection of the islets.

GESTATIONAL DM

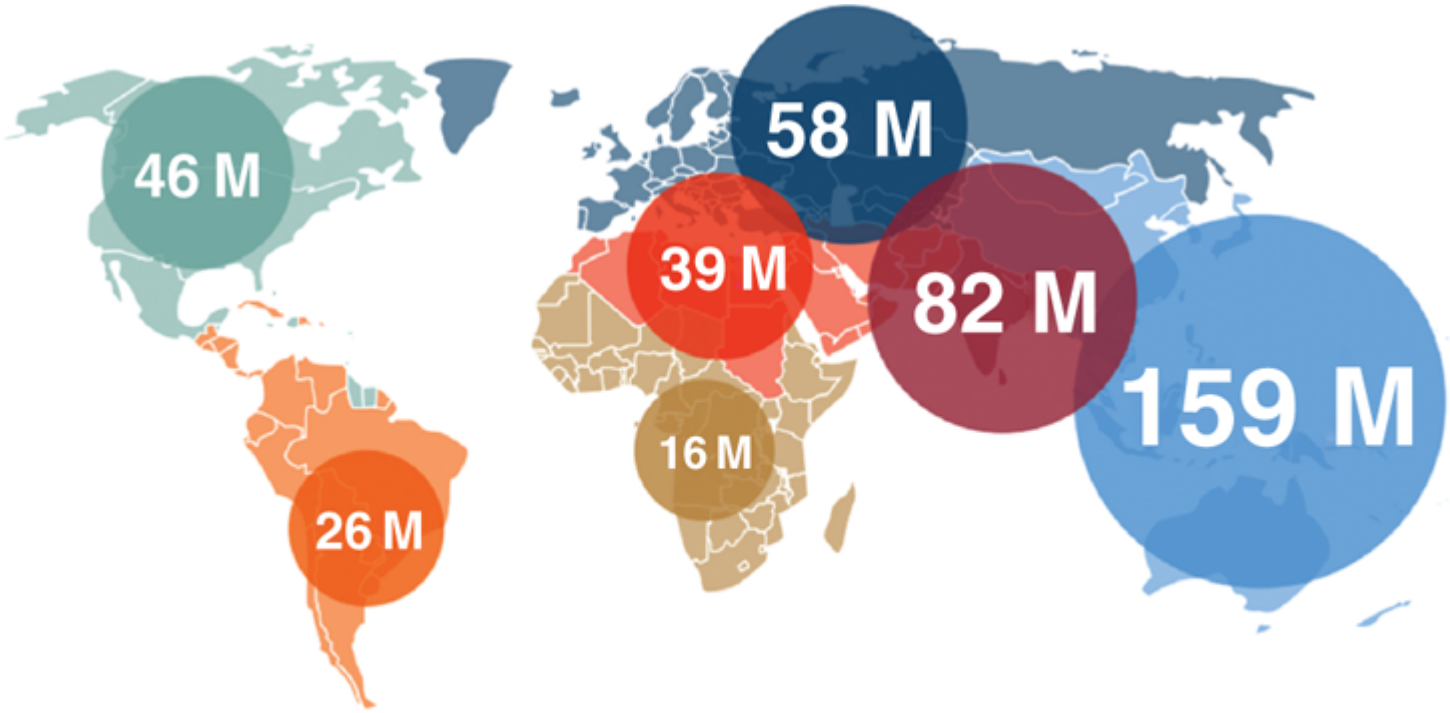
Glucose intolerance developing during the second or third trimester of pregnancy is classified as gestational diabetes mellitus (GDM). Insulin resistance is related to the metabolic changes of pregnancy, during which the increased insulin demands may lead to IGT or diabetes. The American Diabetes Association (ADA) recommends that diabetes diagnosed within the first trimester be classified as preexisting pregestational diabetes rather than GDM. In 2015, the International Diabetes Federation (IDF) estimated that one in seven pregnancies worldwide was affected by either GDM or preexisting DM. Most women with GDM revert to normal glucose tolerance postpartum but have a substantial risk (35–60%) of developing DM in the next 10–20 years. In addition, children born to a mother with GDM also have an increased risk of developing metabolic syndrome and type 2 DM later in life. Currently, the ADA recommends that women with a history of GDM undergo lifelong screening for the development of diabetes or prediabetes at least every 3 years.

EPIDEMIOLOGY AND GLOBAL CONSIDERATIONS

The worldwide prevalence of DM has risen dramatically over the past two decades, from an estimated 30 million cases in 1985 to 415 million in 2017 **(Fig. 396-2)**. Based on current trends, the IDF projects that 642 million individuals will have diabetes by the year 2040 (see *http://www.idf.org/*). Although the prevalence of both type 1 and type 2 DM is increasing worldwide, the prevalence of type 2 DM is rising much more rapidly, presumably because of increasing obesity, reduced activity levels as countries become more industrialized, and the aging of the population. The incidence of type 1 diabetes has been increasing at a rate of 3–5% per year worldwide. The cause for this increase is not well understood, but type 1 DM is increasingly being diagnosed at younger ages. In 2015, the prevalence of diabetes in individuals aged 20–79 ranged from 7.2–11.4%. The countries with the greatest number of individuals with diabetes in 2015 are China (109.6 million), India (73 million), the United States (30.3 million), Brazil (14 million), and the Russian Federation (9 million). In the most recent estimate for the United States (2017), the Centers for Disease Control and Prevention (CDC) estimated that 9.4% of the population had diabetes, and as many as 34% of U.S. adults had prediabetes. Approximately 25% of the individuals with diabetes in the United States were undiagnosed; globally, it is estimated that as many of 50% of individuals with diabetes may be undiagnosed. The prevalence of DM increases with age. In 2015, the prevalence of DM in the United Sates was estimated to be 0.25% in individuals age <20 years, 4.1% in persons aged 20–44 years, and 16.2% in persons 45–64 years old. In individuals aged >65 years, the prevalence of DM was 25.9%. Similar age-related trends have been observed worldwide. The prevalence of diabetes is similar among men and women, but diabetes-related mortality rates are higher in men compared to women.

FIGURE 396-2

Worldwide prevalence of diabetes mellitus. Global estimate is 415 million individuals with diabetes in 2017. Regional estimates of the number of individuals with diabetes (20–79 years of age) are shown (2017). *(Adapted from the IDF Diabetes Atlas, the International Diabetes Federation, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

There is considerable geographic variation in the incidence of both type 1 and type 2 DM. Currently, Scandinavia has the highest incidence of type 1 DM; the lowest incidence is in the Pacific Rim where it is twenty- to thirtyfold lower. Northern Europe and the United States have an intermediate rate. Much of the increased risk of type 1 DM is believed to reflect the frequency of high-risk human leukocyte antigen (HLA) alleles among ethnic groups in different geographic locations.

However, now populations less enriched with these classic high-risk HLA alleles are experiencing more rapid increases in type 1 DM incidence, suggesting an influence of environmental factors.

The prevalence of type 2 DM and its harbinger, IGT, is highest in certain Pacific islands and the Middle East and intermediate in countries such as India and the United States. This variability is likely due to genetic, behavioral, and environmental factors. DM prevalence also varies among different ethnic populations within a given country, with indigenous populations usually having a greater incidence of diabetes than the general population of the country. For example, the CDC estimated that the age-adjusted prevalence of DM in the United States (age >20 years; 2010–2012) was 8% in non-Hispanic whites, 9% in Asian Americans, 13% in Hispanics, 13% in non-Hispanic blacks, and 16% in American-Indian and Alaskan native populations. The onset of type 2 DM occurs, on average, at an earlier age in ethnic groups other than non-Hispanic whites. In Asia, the prevalence of diabetes is increasing rapidly, and the diabetes phenotype appears to be somewhat different from that in the United States and Europe, with an onset at a lower body mass index (BMI) and younger age, greater visceral adiposity, and reduced insulin secretory capacity.

Diabetes is a major cause of mortality. In recent years, diabetes has been listed as the seventh leading cause of death in the United States, but several studies indicate that diabetes-related deaths are likely underreported. Data from the IDF suggests that diabetes was responsible for almost 5 million deaths

worldwide, accounting for 14.5% of global all-cause mortality in adults aged 20–79 years of age. Diabetes also has important economic implications. In 2015, it was estimated that $673 billion or 12% of health care expenditures worldwide were spent on diabetes (range 5–20%). Up to 75% of individuals with diabetes live in low or middle-income countries.

DIAGNOSIS

Glucose tolerance is classified into three broad categories: normal glucose homeostasis, impaired glucose homeostasis, or DM. Glucose tolerance can be assessed using the fasting plasma glucose (FPG), the response to oral glucose challenge, or the hemoglobin A_{1c} (HbA_{1c}). An FPG <5.6 mmol/L (100 mg/dL), a plasma glucose <7.9 mmol/L (140 mg/dL) following an oral glucose challenge, and an HbA_{1c} <5.7% are considered to define normal glucose tolerance. The International Expert Committee with members appointed by the ADA, the European Association for the Study of Diabetes, and the IDF have issued diagnostic criteria for DM **(Table 396-2)** based on the following premises: (1) the FPG, the response to an oral glucose challenge (oral glucose tolerance test [OGTT]), and HbA_{1c} differ among individuals, and (2) DM is defined as the level of glycemia at which diabetes-specific complications occur rather than deviation from a population-based mean. For example, the prevalence of retinopathy in Native Americans (Pima Indian population) begins to increase at an FPG >6.4 mmol/L (116 mg/dL) **(Fig. 396-3)**.

TABLE 396-2

Criteria for the Diagnosis of Diabetes Mellitus

- Symptoms of diabetes plus random blood glucose concentration ≥11.1 mmol/L (200 mg/dL)[a] *or*
- Fasting plasma glucose ≥7.0 mmol/L (126 mg/dL)[b] *or*
- Hemoglobin A_{1c} ≥ 6.5%[c] *or*
- 2-h plasma glucose ≥11.1 mmol/L (200 mg/dL) during an oral glucose tolerance test[d]

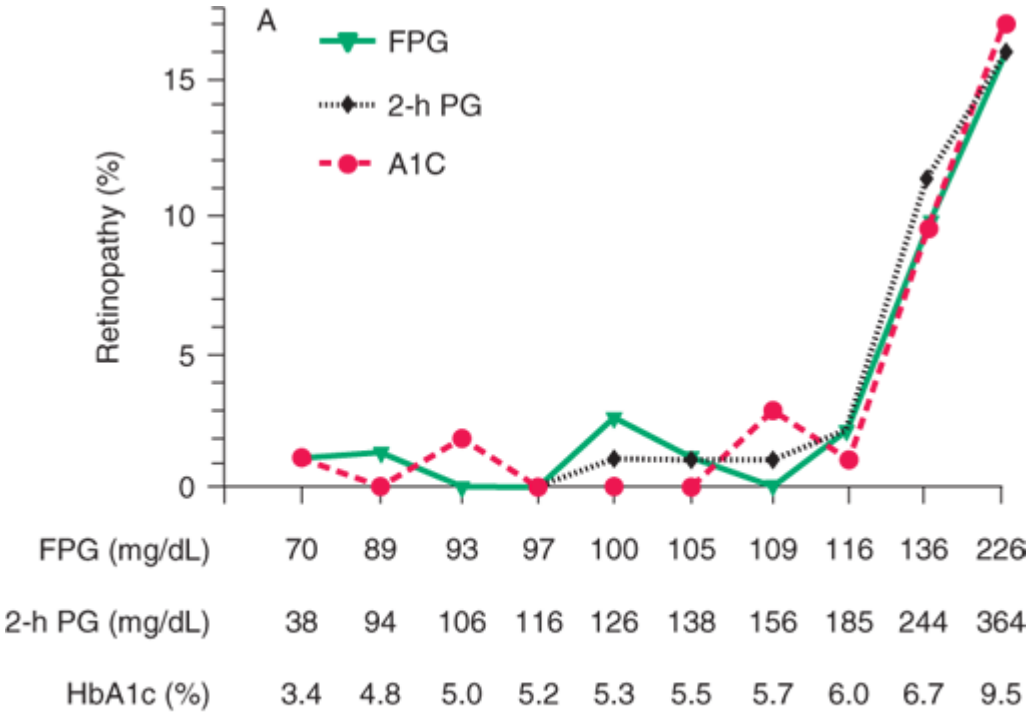
[a]Random is defined as without regard to time since the last meal. [b]Fasting is defined as no caloric intake for at least 8 h. [c]Hemoglobin A_{1c} test should be performed in a laboratory using a method approved by the National Glycohemoglobin Standardization Program and correlated to the reference assay of the Diabetes Control and Complications Trial. Point-of-care hemoglobin A_{1c} should not be used for diagnostic purposes. [d]The test should be performed using a glucose load containing the equivalent of 75 g anhydrous glucose dissolved in water, not recommended for routine clinical use.

Note: In the absence of unequivocal hyperglycemia and acute metabolic decompensation, these criteria should be confirmed by repeat testing on a different day.

Source: Adapted from American Diabetes Association: Diabetes Care 40(Suppl 1): S13-27, 2018.

FIGURE 396-3

Relationship of diabetes-specific complication and glucose tolerance. This figure shows the incidence of retinopathy in Pima Indians as a function of the fasting plasma glucose (FPG), the 2-h plasma glucose after a 75-g oral glucose challenge (2-h PG), or the hemoglobin A_{1c} (HbA_{1c}). Note that the incidence of retinopathy greatly increases at a fasting plasma glucose >116 mg/dL, a 2-h plasma glucose of 185 mg/dL, or an HbA_{1c} >6.5%. (Blood glucose values are shown in mg/dL; to convert to mmol/L, divide value by 18.) *(Copyright 2002, American Diabetes Association. From Diabetes Care 25[Suppl 1]: S5–S20, 2002.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Abnormal glucose homeostasis (Fig. 396-1) is defined as (1) FPG = 5.6–6.9 mmol/L (100–125 mg/dL), which is defined as *impaired fasting glucose* (IFG); the World Health Organization uses 6.1–6.9 mmol/L (110–125 mg/dL) for IFG; (2) plasma glucose levels between 7.8 and 11 mmol/L (140 and 199 mg/dL) following an oral glucose challenge, which is termed *impaired glucose tolerance* (IGT); or (3) HbA_{1c} of 5.7–6.4%. An HbA_{1c} of 5.7–6.4%, IFG, and IGT do not identify the same individuals, but individuals in all three groups are at greater risk of progressing to type 2 DM, have an increased risk of cardiovascular disease, and should be counseled about ways to decrease these risks (see below). Some use the terms *prediabetes*, *increased risk of diabetes*, or *intermediate hyperglycemia* (World Health Organization) and slightly different metrics for this category.

These values for the FPG, the glucose following an oral glucose challenge, and HbA_{1c} are continuous rather than discrete variables. A FPG ≥7.0 mmol/L (126 mg/dL), a glucose ≥11.1 mmol/L (200 mg/dL) 2 h after an oral glucose challenge, or an HbA_{1c} ≥6.5% meets the criteria for the diagnosis of DM (Table 396-2). A random plasma glucose concentration ≥11.1 mmol/L (200 mg/dL) accompanied by classic symptoms of DM (polyuria, polydipsia, weight loss) is also sufficient for the diagnosis of DM (Table 396-2). The current criteria for the diagnosis of DM emphasize the HbA_{1c} or the FPG as the most reliable and convenient tests for identifying DM

in asymptomatic individuals. However, some individuals may meet criteria for one test but not the other. Also, it is important to note that race and ethnicity may impact the reliability of HbA_{1c} levels. For example, African Americans have a higher HbA_{1c} value compared to non-Hispanic whites with a similar level of glycemia. An OGTT, although a valid means for diagnosing DM, is not often used in routine clinical care with the exception of pregnancy care and screening for gestational diabetes.

The diagnosis of DM has profound implications for an individual from both a medical and a financial standpoint. Thus, abnormalities on screening tests for diabetes should be repeated before making a definitive diagnosis of DM, unless acute metabolic derangements or a markedly elevated plasma glucose are present (Table 396-2). These criteria also allow for the diagnosis of DM to be withdrawn in situations when the glucose intolerance reverts to normal.

SCREENING

Widespread use of the FPG or the HbA_{1c} as a screening test for type 2 DM is recommended because (1) a large number of individuals who meet the current criteria for DM are asymptomatic and unaware that they have the disorder, (2) epidemiologic studies suggest that type 2 DM may be present for up to a decade before diagnosis, (3) some individuals with type 2 DM have one or more diabetes-specific complications at the time of their diagnosis, (4) treatment of type 2 DM may favorably alter the natural history of DM, (5) diagnosis of prediabetes should spur efforts for diabetes prevention. The ADA recommends screening all individuals aged >45 years every 3 years and screening individuals at an earlier age if they are overweight (BMI >25 kg/m^2 or ethnically relevant definition for overweight) and have one additional risk factor for diabetes **(Table 396-3)**. Although a number of immunologic markers for type 1 DM are becoming available (discussed below), their routine use outside a clinical trial is discouraged, pending the identification of clinically beneficial interventions for individuals at high risk for developing type 1 DM.

TABLE 396-3

Risk Factors for Type 2 Diabetes Mellitus

Family history of diabetes (i.e., parent or sibling with type 2 diabetes)
Overweight or obese (BMI ≥25 kg/m^2, ≥23 kg/m^2 in Asian Americans, or other ethnically relevant definition for overweight)
Physical inactivity
Race/ethnicity (e.g., African American, Latino, Native American, Asian American, Pacific Islander)
Previously identified with IFG, IGT, or an hemoglobin A_{1c} of 5.7–6.4%
History of GDM
Hypertension (blood pressure ≥140/90 mmHg)
HDL cholesterol level <35 mg/dL (0.90 mmol/L) and/or a triglyceride level >250 mg/dL (2.82 mmol/L)
Polycystic ovary syndrome or acanthosis nigricans
History of cardiovascular disease

Abbreviations: BMI, body mass index; GDM, gestational diabetes mellitus; HDL, high-density lipoprotein; IFG, impaired fasting glucose; IGT, impaired glucose tolerance.

Source: Adapted from American Diabetes Association: Diabetes Care 40(Suppl 1): S13, 2018.

REGULATION OF GLUCOSE HOMEOSTASIS

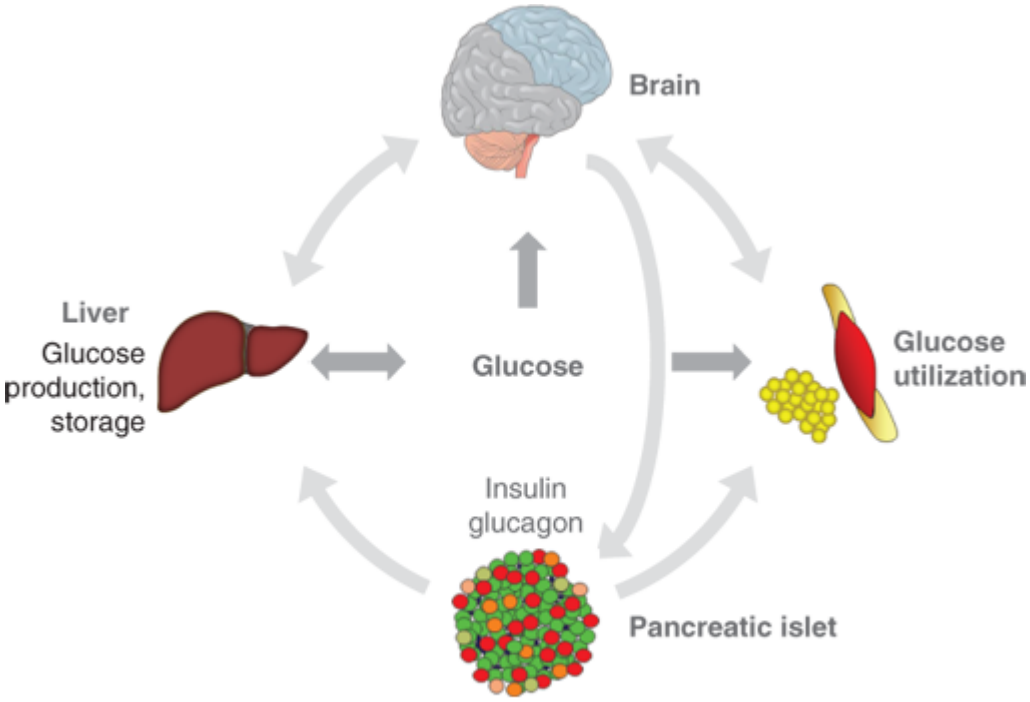
OVERALL REGULATION OF GLUCOSE HOMEOSTASIS

Glucose homeostasis reflects a balance between energy intake from ingested food, hepatic glucose production (gluconeogenesis), and peripheral tissue glucose uptake and utilization. Insulin is the most important regulator of this metabolic equilibrium, but neural input, metabolic signals, and other hormones (e.g., glucagon) result in integrated control of glucose supply and utilization **(Fig. 396-4)**. The organs that regulate glucose and lipids communicate by neural and humoral mechanisms with fat and muscle producing adipokines, myokines, and metabolites that influence liver function. In the fasting state, low insulin levels, together with modest increases in glucagon, increase glucose production by promoting hepatic gluconeogenesis and glycogen breakdown (glycogenolysis) and reducing glucose uptake in insulin-sensitive tissues (skeletal muscle and fat), thereby promoting mobilization of stored precursors such as amino acids and free fatty acids (lipolysis). Glucagon, secreted by pancreatic alpha cells normally only when blood glucose or insulin levels are low or during exercise, is increased in DM and stimulates glycogenolysis and gluconeogenesis by the liver and to a small degree by the renal medulla **(Chap. 399)**. Postprandially, the glucose load elicits a rise in insulin and fall in glucagon, leading to a reversal of these processes. Insulin, an anabolic hormone, promotes the storage of carbohydrate and fat and protein synthesis. The major portion of postprandial glucose is used by skeletal muscle, an effect of insulin-stimulated glucose uptake. Other tissues,

most notably the brain, use glucose in an insulin-independent fashion. Factors secreted by skeletal myocytes, adipocytes (leptin, resistin, adiponectin, etc.), and bone also influence glucose homeostasis.

FIGURE 396-4

Regulation of glucose homeostasis. The organs shown contribute to glucose utilization, production, or storage. See text for a description of the communications (*arrows*), which can be neural or humoral. Although not shown, the GI tract and bone produce factors that influence glucose homeostasis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

INSULIN BIOSYNTHESIS

Insulin, produced by the beta cells of the pancreatic islets, is initially synthesized as a single-chain 86-amino-acid precursor polypeptide, preproinsulin. Subsequent proteolytic processing removes the amino-terminal signal peptide, giving rise to proinsulin. Proinsulin is structurally related to insulin-like growth factors I and II, which bind weakly to the insulin receptor. Cleavage of an internal 31-residue fragment from proinsulin generates C-peptide with the A (21 amino acids) and B (30 amino acids) chains of insulin being connected by disulfide bonds. The mature insulin molecule and C-peptide are stored together and co-secreted from secretory granules in the beta cells. Because C-peptide is cleared more slowly than insulin, it is a useful marker of insulin secretion and allows discrimination of endogenous and exogenous sources of insulin in the evaluation of hypoglycemia **(Chaps. 399 and 80)**. Elevated levels of serum proinsulin have been observed in both type 1 and 2 DM and are thought to be indicative of beta cell dysfunction. Pancreatic beta cells co-secrete islet amyloid polypeptide (IAPP) or amylin, a 37-amino-acid peptide, along with insulin. The role of IAPP in normal physiology is incompletely defined, but it is the major component of the amyloid fibrils found in the islets of patients with type 2 diabetes, and an analogue is sometimes used in treating type 1 and type 2

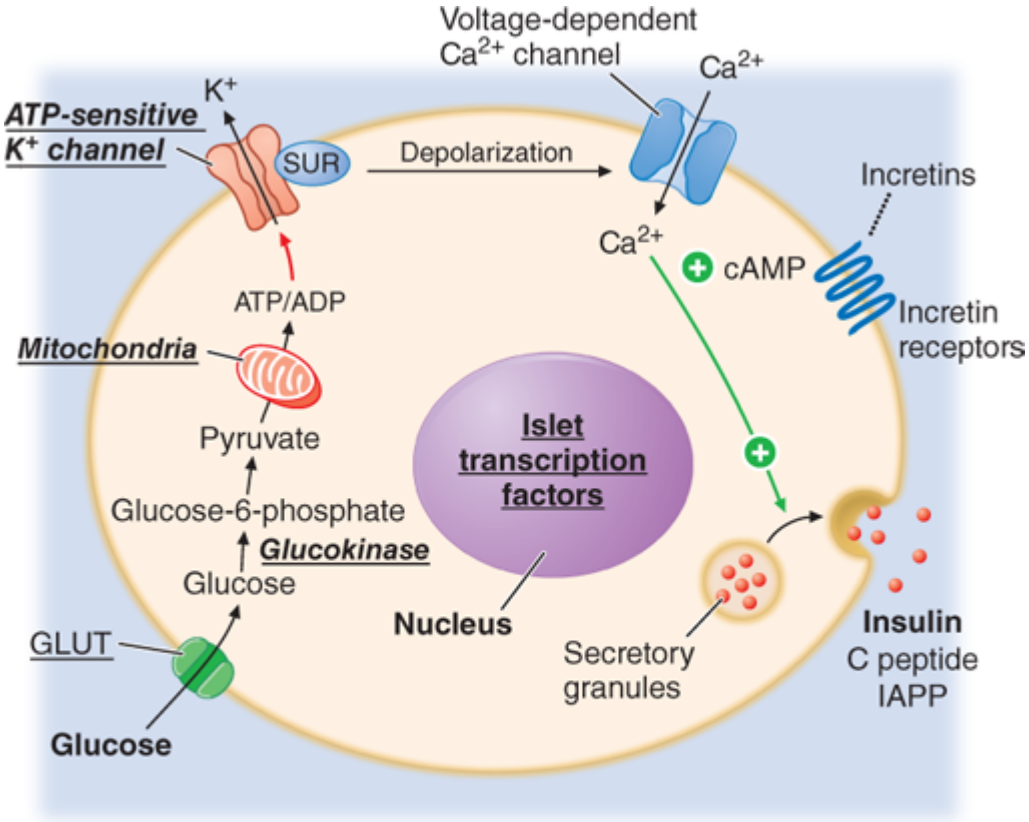
DM. Human insulin is produced by recombinant DNA technology; structural modifications at one or more amino acid residues modify insulin's physical and pharmacologic characteristics **(Chap. 397)**.

INSULIN SECRETION

Glucose is the key regulator of insulin secretion by the pancreatic beta cell, although amino acids, ketones, various nutrients, gastrointestinal peptides, and neurotransmitters also influence insulin secretion. Glucose levels >3.9 mmol/L (70 mg/dL) stimulate insulin synthesis, primarily by enhancing protein translation and processing. Glucose stimulation of insulin secretion begins with its transport into the beta cell by a facilitative glucose transporter **(Fig. 396-5)**. Glucose phosphorylation by glucokinase is the rate-limiting step that controls glucose-regulated insulin secretion. Further metabolism of glucose-6-phosphate via glycolysis generates ATP, which inhibits the activity of an ATP-sensitive K^+ channel. This channel consists of two separate proteins: one is the binding site for certain oral hypoglycemics (e.g., sulfonylureas, meglitinides); the other is an inwardly rectifying K^+ channel protein (Kir6.2). Inhibition of this K^+ channel induces beta cell membrane depolarization, which opens voltage-dependent calcium channels (leading to an influx of calcium) and stimulates insulin secretion. Insulin secretory profiles reveal a pulsatile pattern of hormone release, with small secretory bursts occurring about every 10 min, superimposed upon greater amplitude oscillations of about 80–150 min. A number of metabolic pathways internal to the beta cell as well as external hormonal cues amplify glucose-stimulated insulin secretion. Glucagon-like peptide-1 (GLP-1) and glucose-dependent insulinotropic peptide (GIP) are incretin hormones that bind specific receptors on the beta cell to stimulate insulin secretion through cyclic AMP production, but have this effect only when the blood glucose is above the fasting level. Incretin hormones also suppress glucagon production and secretion. Incretin analogues or pharmacologic agents that prolong the activity of endogenous GLP-1 are used therapeutically in type 2 DM. Classically, GLP-1 release was thought to occur solely from neuroendocrine L-cells of the gastrointestinal tract following food ingestion. However, recent pre-clinical studies suggest that intraislet production of GLP-1 from alpha cells may play a role in the regulation of insulin secretion.

FIGURE 396-5

Mechanisms of glucose-stimulated insulin secretion and abnormalities in diabetes. Glucose and other nutrients regulate insulin secretion by the pancreatic beta cell. Glucose is transported by a glucose transporter (GLUT1 and/or GLUT2 in humans, GLUT2 in rodents); subsequent glucose metabolism by the beta cell alters ion channel activity, leading to insulin secretion. The SUR receptor is the binding site for some drugs that act as insulin secretagogues. Mutations in the events or proteins underlined are a cause of monogenic forms of diabetes. ADP, adenosine diphosphate; ATP, adenosine triphosphate; cAMP, cyclic adenosine monophosphate; IAPP, islet amyloid polypeptide or amylin; SUR, sulfonylurea receptor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

INSULIN ACTION

Once insulin is secreted into the portal venous system, ~50% is removed and degraded by the liver. Unextracted insulin enters the systemic circulation where it binds to receptors in target sites. Insulin binding to its receptor stimulates intrinsic tyrosine kinase activity, leading to receptor autophosphorylation and the recruitment of intracellular signaling molecules, such as insulin receptor substrates (IRS). IRS and other adaptor proteins initiate a complex cascade of phosphorylation and dephosphorylation reactions, resulting in the widespread metabolic and mitogenic effects of insulin. As an example, activation of the phosphatidylinositol-3′-kinase (PI-3-kinase) pathway stimulates translocation of a facilitative glucose transporter (e.g., GLUT4) to the cell surface, an event that is crucial for glucose uptake by skeletal muscle and fat. Activation of other insulin receptor signaling pathways induces glycogen synthesis, protein synthesis, lipogenesis, and regulation of various genes in insulin-responsive cells.

PATHOGENESIS

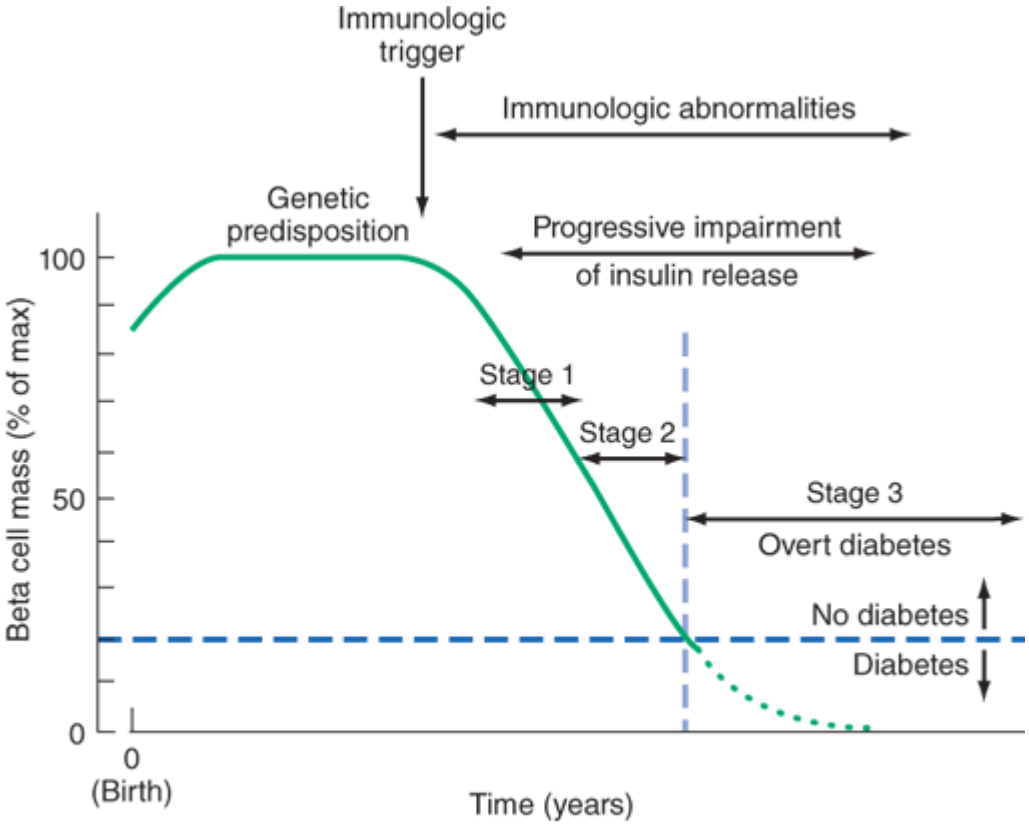
TYPE 1 DM

Type 1 DM is the result of interactions of genetic, environmental, and immunologic factors that ultimately lead to immune-mediated destruction of the pancreatic beta cells and insulin deficiency. Type 1 DM can develop at any age, but most commonly develops before 20 years of age. Most, but not all, individuals with

type 1 DM have evidence of islet-directed autoimmunity. However, some individuals who have the clinical phenotype of type 1 DM lack immunologic markers indicative of an autoimmune process involving the beta cells and the genetic markers of type 1 DM. These individuals are thought to develop insulin deficiency by unknown, nonimmune mechanisms and may be ketosis prone; many are African American or Asian in heritage. The temporal decline of beta cell function and mass preceding the development of type 1 DM is shown schematically in **Fig. 396-6**. In susceptible individuals, the autoimmune process is thought to be triggered by an infectious or environmental stimulus. In the majority of patients, autoantibodies against beta cell antigens appear after this triggering event, followed by progressive loss of insulin secretion. The rate of decline in beta cell function varies widely among individuals, with some patients progressing rapidly to clinical diabetes and others evolving to diabetes more slowly and over a period of several years. Features of diabetes do not become evident until a threshold loss of insulin secretion and beta cell mass occurs. Autopsy studies suggest the degree of loss of beta cell mass is variable at the time of disease presentation but may be as high as 70–80%. At this point, residual, functional beta cells exist but are insufficient in number and quality to maintain glucose tolerance. The events that trigger the transition from glucose intolerance to frank diabetes are often associated with increased insulin requirements, as might occur during infections or at puberty. After the initial clinical presentation of type 1 DM, a “honeymoon” phase may ensue during which time glycemic control is achieved with modest doses of insulin or, rarely, insulin is not needed. However, this fleeting phase of endogenous insulin production from residual beta cells disappears and the individual becomes insulin deficient. Many individuals with long-standing type 1 DM produce a small amount of insulin (as reflected by C-peptide production), and some individuals with 50 years of type 1 DM have insulin-positive cells in the pancreas at autopsy.

FIGURE 396-6

Temporal model for development of type 1 diabetes. Individuals with a genetic predisposition are exposed to a trigger that initiates an autoimmune process, resulting in the development of islet autoantibodies and a gradual decline in beta cell function and mass. Stage 1 disease is characterized by the development of two or more islet cell autoantibodies but the maintenance of normoglycemia. Stage 2 disease is defined by continued autoimmunity and the development of dysglycemia. Stage 3 is defined by the development of hyperglycemia that exceeds the diagnostic criteria for the diagnosis of diabetes. The downward slope of the beta cell function varies among individuals and may not be continuous. A “honeymoon” phase may be seen in the first 1 or 2 years after the onset of diabetes and is associated with reduced insulin requirements. *(Adapted and modified from ER Kaufman: Medical Management of Type 1 Diabetes, 6th ed. American Diabetes Association, Alexandria, VA, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GENETIC CONSIDERATIONS



Susceptibility to type 1 DM involves multiple genes. The concordance of type 1 DM in identical twins ranges between 40 and 60%, indicating that additional modifying factors are likely involved in determining whether diabetes develops. The major susceptibility gene for type 1 DM is located in the HLA region on chromosome 6. Polymorphisms in the HLA complex account for 40–50% of the genetic risk of developing type 1 DM. This region contains genes that encode the class II major histocompatibility complex (MHC) molecules, which present antigen to helper T cells and thus are involved in initiating the immune response (Chap. 343). The ability of class II MHC molecules to present antigen is dependent on the amino acid composition of their antigen-binding sites. Amino acid substitutions may influence the specificity of the immune response by altering the binding affinity of different antigens for class II molecules.

Most individuals with type 1 DM have the HLA DR3 and/or DR4 haplotype. Refinements in genotyping of HLA loci have shown that the haplotypes DQA1*0301, DQB1*0302, and DQB1*0201 are most strongly associated with type 1 DM. These haplotypes are present in 40% of children with type 1 DM as compared to 2% of the normal U.S. population. However, most individuals with predisposing haplotypes do not develop diabetes.

In addition to MHC class II associations, genome association studies have identified at least 20 additional genetic loci that contribute susceptibility to type 1 DM (i.e., polymorphisms in the promoter region of the insulin gene, the CTLA-4 gene, interleukin 2 receptor, and PTPN22, etc.). Among recent cohorts of individuals

with new-onset type 1 diabetes, there is a decreased representation of the highest risk HLA alleles and increasing penetrance of disease in those genotypes classically associated with lower risk. Genes that confer protection against the development of the disease also exist. The haplotype DQA1*0102, DQB1*0602 is extremely rare in individuals with type 1 DM (<1%) and appears to provide protection from type 1 DM.

Although the risk of developing type 1 DM is increased tenfold in relatives of individuals with the disease, the risk is relatively low: 3–4% if the parent has type 1 DM and 5–15% in a sibling (depending on which HLA haplotypes are shared). Hence, most individuals with type 1 DM (75%) do not have a first-degree relative with this disorder.

Pathophysiology

Although other islet cell types (alpha cells [glucagon-producing], delta cells [somatostatin-producing], or PP cells [pancreatic polypeptide-producing]) are functionally and embryologically similar to beta cells, they are spared from the autoimmune destruction. However, altered patterns of hormone secretion from these other cell types in type 1 DM likely contributes to metabolic instability. Alpha cell dysfunction as reflected by fasting hyperglucagonemia, hyperglucagonemia in the post-prandial state, and an impaired glucagon response to hypoglycemia. Pathologically, the pancreatic islets have modest infiltration of lymphocytes (a process termed *insulitis*). After beta cells are destroyed, it is thought that the inflammatory process abates and the islets become atrophic. Studies of the autoimmune process in humans and in animal models of type 1 DM (NOD mouse and BB rat) have identified the following abnormalities in the humoral and cellular arms of the immune system: (1) islet cell autoantibodies; (2) activated lymphocytes in the islets, peripancreatic lymph nodes, and systemic circulation; (3) T lymphocytes that proliferate when stimulated with islet proteins; and (4) release of cytokines within the insulitis. Beta cells seem to be particularly susceptible to the toxic effect of some cytokines (tumor necrosis factor α [TNF-α], interferon γ, and interleukin 1 [IL-1]). The precise mechanisms of beta cell death are not known but may involve formation of nitric oxide metabolites, apoptosis, and direct CD8+ T cell cytotoxicity. The islet destruction is mediated by T lymphocytes rather than islet autoantibodies, as these antibodies do not generally react with the cell surface of islet cells and are not capable of transferring DM to animals. Efforts to suppress the autoimmune process at the time of diagnosis of diabetes have largely been ineffective or only temporarily effective in slowing beta cell destruction. Thus, increased emphasis has now been placed on interventions earlier in the disease course (i.e., during Stage 1 and 2 disease; Fig. 396-6).

Pancreatic islet molecules targeted by the autoimmune process include proinsulin, insulin, glutamic acid decarboxylase (GAD; the biosynthetic enzyme for the neurotransmitter GABA), ICA-512/IA-2 (homology with tyrosine phosphatases), and a beta cell–specific zinc transporter (ZnT-8). Most of the autoantigens are not beta cell–specific, which raises the question of how the beta cells are selectively destroyed. Current theories favor initiation of an autoimmune process directed at one beta cell molecule, which then spreads to other islet molecules as the immune process destroys beta cells and creates a series of secondary autoantigens. Stress pathways and processes arising within the beta cell may exacerbate autoimmunity through the development of modified proteins or “neoantigens” that serve as additional immune targets.

Immunologic Markers

Islet cell autoantibodies (ICAs) are a composite of several different antibodies directed at pancreatic islet molecules such as GAD, insulin, IA-2/ICA-512, and ZnT-8, and serve as a marker of the autoimmune process of type 1 DM. Assays for autoantibodies to GAD-65 and insulin are commercially available. Testing for ICAs can be useful in classifying the type of DM as type 1 and in identifying nondiabetic individuals at risk for developing type 1 DM. ICAs are present in the majority of individuals (>85%) diagnosed with new-onset type 1 DM, in a significant minority of individuals with newly diagnosed type 2 DM (5–10%), and occasionally in individuals with GDM (<5%). ICAs are present in 3–4% of first-degree relatives of individuals with type 1 DM. In combination with impaired insulin secretion after IV glucose tolerance testing, they predict a >50% risk of developing type 1 DM within 5 years. Increasing numbers of autoantibodies are associated with an increased risk of diabetes development. In children with multiple autoantibodies, ~70% developed type 1 DM after 10 years of follow-up, with 80% developing diabetes after 15-year of follow-up. At present, the measurement of ICAs in nondiabetic individuals remains a research tool because no treatments have been demonstrated to prevent the occurrence or progression to type 1 DM.

Environmental Factors

Numerous environmental events have been proposed to trigger the autoimmune process in genetically susceptible individuals; however, none have been conclusively linked to diabetes. Identification of an environmental trigger has been difficult because the event may precede the onset of DM by several years (Fig. 396-6). Putative environmental triggers include viruses (coxsackie, rubella, enteroviruses most prominently), bovine milk proteins, nitrosourea compounds, vitamin D deficiency, and environmental toxins. There is increasing interest in the microbiome and type 1 diabetes **(Chap. 459)**.

Prevention of Type 1 DM

A number of interventions have prevented diabetes in animal models, but relatively few interventions have been tested in humans in Stages 1 and 2 of type 1 DM. The Diabetes Prevention Trial–Type 1 concluded that administering insulin (IV or PO) to individuals at high risk for developing type 1 DM did not prevent type 1 DM. However, this is an area of active clinical investigation with several trials evaluating interventions that target different aspects of the immune response in early stage type 1 DM.

TYPE 2 DM

Insulin resistance and abnormal insulin secretion are central to the development of type 2 DM. Although the primary defect is controversial, most studies support the view that insulin resistance precedes an insulin secretory defect but that diabetes develops only when insulin secretion becomes inadequate. Type 2 DM likely encompasses a range of disorders with the common phenotype of hyperglycemia. Most of our current understanding (and the discussion below) of the pathophysiology and genetics is based on studies of individuals of European descent. It is becoming increasing apparent that DM in other ethnic groups (Asian, African, and Latin American) has a somewhat different, but yet undefined, pathophysiology. In general,

Latinos have greater insulin resistance and East Asians and South Asians have more beta cell dysfunction, but both defects are present in both populations. East and South Asians appear to develop type 2 DM at a younger age and a lower BMI. In some groups, DM that is ketosis prone (often in obese individuals) or ketosis-resistant (often lean) is sometimes seen.

GENETIC CONSIDERATIONS



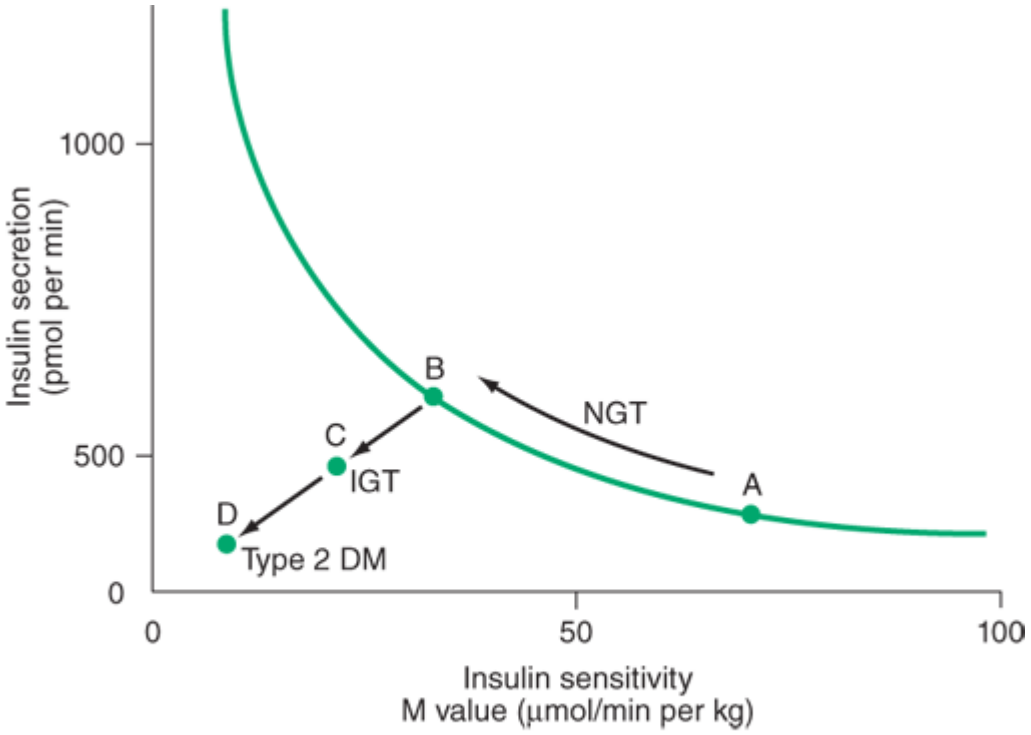
Type 2 DM has a strong genetic component. The concordance of type 2 DM in identical twins is between 70 and 90%. Individuals with a parent with type 2 DM have an increased risk of diabetes; if both parents have type 2 DM, the risk approaches 40%. Insulin resistance, as demonstrated by reduced glucose utilization in skeletal muscle, is present in many nondiabetic, first-degree relatives of individuals with type 2 DM. The disease is polygenic and multifactorial, because in addition to genetic susceptibility, environmental factors (such as obesity, poor nutrition, and physical inactivity) modulate the phenotype. The in utero environment also contributes, and either increased or reduced birth weight increases the risk of type 2 DM in adult life. Children of pregnancies complicated by gestational hyperglycemia also exhibit an increased risk of type 2 DM. The genes that predispose to type 2 DM are incompletely identified, but genome-wide association studies have identified a large number of genes that convey a relatively small risk for type 2 DM (>70 genes, each with a relative risk of 1.06–1.5). Most prominent is a variant of the transcription factor 7–like 2 gene that has been associated with type 2 DM in several populations and with IGT. Genetic polymorphisms associated with type 2 DM have also been found in the genes encoding the peroxisome proliferator–activated receptor γ, inward rectifying potassium channel, zinc transporter, IRS, and calpain 10. The mechanisms by which these genetic loci increase the susceptibility to type 2 DM are not clear, but most are predicted to alter islet function or development or insulin secretion. Although the genetic susceptibility to type 2 DM is under active investigation (it is estimated that <10% of genetic risk is determined by loci identified thus far), it is currently not possible to use a combination of known genetic loci to predict type 2 DM.

Pathophysiology

Type 2 DM is characterized by impaired insulin secretion, insulin resistance, excessive hepatic glucose production, abnormal fat metabolism, and systemic low-grade inflammation. Obesity, particularly visceral or central (as evidenced by the hip-waist ratio), is very common in type 2 DM (≥80% of patients are obese). In the early stages of the disorder, glucose tolerance remains near-normal, despite insulin resistance, because the pancreatic beta cells compensate by increasing insulin output **(Fig. 396-7)**. As insulin resistance and compensatory hyperinsulinemia progress, the pancreatic islets in certain individuals are unable to sustain the hyperinsulinemic state. IGT, characterized by elevations in postprandial glucose, then develops. A further decline in insulin secretion and an increase in hepatic glucose production lead to overt diabetes with fasting hyperglycemia. Ultimately, beta cell failure ensues. Possibly because of inadequate insulin suppression, glucagon is relatively overproduced and secreted, further augmenting hepatic glucose production. Although both insulin resistance and impaired insulin secretion contribute to the pathogenesis of type 2 DM, the relative contribution of each varies from individual to individual.

FIGURE 396-7

Metabolic changes during the development of type 2 diabetes mellitus (DM). Insulin secretion and insulin sensitivity are related, and as an individual becomes more insulin resistant (by moving from point A to point B), insulin secretion increases. A failure to compensate by increasing the insulin secretion results initially in impaired glucose tolerance (IGT; point C) and ultimately in type 2 DM (point D). NGT, normal glucose tolerance. *(Adapted from SE Kahn: J Clin Endocrinol Metab 86:4047, 2001; RN Bergman, M Ader: Trends Endocrinol Metab 11:351, 2000.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Metabolic Abnormalities

ABNORMAL MUSCLE AND FAT METABOLISM

Insulin resistance, the decreased ability of insulin to act effectively on target tissues (especially muscle, liver, and fat), is a prominent feature of type 2 DM and results from a combination of genetic susceptibility and obesity. Insulin resistance is relative, however, because supranormal levels of circulating insulin will normalize the plasma glucose. Insulin dose-response curves exhibit a rightward shift, indicating reduced sensitivity, and a reduced maximal response, indicating an overall decrease in maximum glucose utilization (30–60% lower than in normal individuals). Insulin resistance impairs glucose utilization by insulin-sensitive tissues and increases hepatic glucose output; both effects contribute to the hyperglycemia. Increased hepatic glucose output predominantly accounts for increased FPG levels, whereas decreased peripheral glucose utilization results in postprandial hyperglycemia. In skeletal muscle, there is a greater impairment in nonoxidative glucose usage (glycogen formation) than in oxidative glucose metabolism through glycolysis. Glucose metabolism in insulin-independent tissues is not altered in type 2 DM.

The precise molecular mechanism leading to insulin resistance in type 2 DM has not been elucidated. Insulin receptor levels and tyrosine kinase activity in skeletal muscle are reduced, but these alterations are most likely secondary to hyperinsulinemia and are not a primary defect. Therefore, “postreceptor” defects in insulin-regulated phosphorylation/dephosphorylation appear to play the predominant role in insulin resistance. Abnormalities include the accumulation of lipid intermediates within skeletal myocytes, which may impair mitochondrial oxidative phosphorylation and reduce insulin-stimulated mitochondrial ATP production. Impaired fatty acid oxidation and lipid accumulation within skeletal myocytes also may generate reactive oxygen species such as lipid peroxides. Of note, not all insulin signal transduction pathways are resistant to the effects of insulin (e.g., those controlling cell growth and differentiation using the mitogenic-activated protein kinase pathway). Consequently, hyperinsulinemia may increase the insulin action through these pathways, potentially accelerating diabetes-related conditions such as atherosclerosis.

The obesity accompanying type 2 DM, particularly in a central or visceral location, is thought to be part of the pathogenic process **(Chap. 394)**. In addition to these white fat depots, humans now are recognized to have brown fat, which has much greater thermogenic capacity. Efforts are under way to increase the activity or quantity of brown fat. The increased adipocyte mass leads to increased levels of circulating free fatty acids and other fat cell products. For example, adipocytes secrete a number of biologic products (nonesterified free fatty acids, retinol-binding protein 4, leptin, TNF-α, resistin, IL-6, and adiponectin). In addition to regulating body weight, appetite, and energy expenditure, adipokines also modulate insulin sensitivity. The increased production of free fatty acids and some adipokines may cause insulin resistance in skeletal muscle and liver. The venous drainage of the visceral adipose beds is the portal circulation and this likely contributes to hepatic dysfunction. Free fatty acids also impair glucose utilization in skeletal muscle, promote glucose production by the liver, and impair beta cell function. In contrast, the production by adipocytes of adiponectin, an insulin-sensitizing peptide, is reduced in obesity, and this may contribute to hepatic insulin resistance. Adipocyte products and adipokines also produce an inflammatory state and may explain why markers of inflammation such as IL-6 and C-reactive protein are often elevated in type 2 DM. In addition, inflammatory cells have been found infiltrating adipose tissue.

IMPAIRED INSULIN SECRETION

Insulin secretion and sensitivity are interrelated (Fig. 396-7). In type 2 DM, insulin secretion initially increases in response to insulin resistance to maintain normal glucose tolerance. Initially, the insulin secretory defect is mild and selectively involves glucose-stimulated insulin secretion, including a greatly reduced first secretory phase. The response to other nonglucose secretagogues, such as arginine, is preserved, but overall beta cell function is reduced by as much as 50% at the onset of type 2 DM. Abnormalities in proinsulin processing are reflected by increased secretion of proinsulin in type 2 DM. Eventually, the insulin secretory defect is progressive.

The reason(s) for the decline in insulin secretory capacity in type 2 DM is unclear. The assumption is that a second genetic defect—superimposed upon insulin resistance—leads to defects in beta cell function, mass, and potentially cellular identity and differentiation status. Beta cell mass is decreased by ~50% in individuals with long-standing type 2 DM. Islet amyloid polypeptide or amylin, co-secreted by the beta cell, forms amyloid fibrillar deposits found in the islets of individuals with long-standing type 2 DM. Whether such islet

amyloid deposits are a primary or secondary event is not known. The metabolic environment of diabetes also negatively impacts islet function. For example, chronic hyperglycemia paradoxically impairs islet function ("glucose toxicity") and leads to a worsening of hyperglycemia. Improvement in glycemic control is often associated with improved islet function. In addition, elevated levels of free fatty acids ("lipotoxicity"), and systemic and local elevations in pro-inflammatory cytokines from increased numbers of islet-associated macrophages, may also worsen islet function. Reduced GLP-1 action may contribute to the reduced insulin secretion.

INCREASED HEPATIC GLUCOSE AND LIPID PRODUCTION

In type 2 DM, insulin resistance in the liver reflects the failure of hyperinsulinemia to suppress gluconeogenesis, which results in fasting hyperglycemia and decreased glycogen storage by the liver in the postprandial state. Increased hepatic glucose production occurs early in the course of diabetes, although likely after the onset of insulin secretory abnormalities and insulin resistance in skeletal muscle. As a result of insulin resistance in adipose tissue, lipolysis and free fatty acid flux from adipocytes are increased and efficiently cleared by liver leading to increased very-low-density lipoprotein [VLDL]-triglyceride) synthesis in hepatocytes and secretion from liver. This is also responsible for the dyslipidemia found in type 2 DM (elevated triglycerides, reduced high-density lipoprotein [HDL], and increased small dense low-density lipoprotein [LDL] particles). If this lipid is retained, steatosis in the liver may lead to nonalcoholic fatty liver disease and abnormal liver function tests.

Insulin Resistance Syndromes

The insulin resistance condition comprises a spectrum of disorders, with hyperglycemia representing one of the most readily diagnosed features. The *metabolic syndrome*, the *insulin resistance syndrome*, and *syndrome X* are terms used to describe a constellation of metabolic derangements that includes insulin resistance, hypertension, dyslipidemia (decreased HDL and elevated triglycerides), central or visceral obesity, type 2 DM or IGT/IFG, and accelerated cardiovascular disease. This syndrome is discussed in **Chap. 401**.

A number of relatively rare forms of severe insulin resistance include features of type 2 DM or IGT (Table 396-1). Mutations in the insulin receptor that interfere with binding or signal transduction are a rare cause of insulin resistance. Acanthosis nigricans and signs of hyperandrogenism (hirsutism, acne, and oligomenorrhea in women) are also common physical features. Two distinct syndromes of severe insulin resistance have been described in adults: (1) type A, which affects young women and is characterized by severe hyperinsulinemia, obesity, and features of hyperandrogenism; and (2) type B, which affects middle-aged women and is characterized by severe hyperinsulinemia, features of hyperandrogenism, and autoimmune disorders. Individuals with the type A insulin resistance syndrome have an undefined defect in the insulin-signaling pathway; individuals with the type B insulin resistance syndrome have autoantibodies directed at the insulin receptor. These receptor autoantibodies may block insulin binding or may stimulate the insulin receptor, leading to intermittent hypoglycemia.

Polycystic ovary syndrome (PCOS) is a common disorder that affects premenopausal women and is characterized by chronic anovulation and hyperandrogenism **(Chap. 385)**. Insulin resistance is seen in a

significant subset of women with PCOS, and the disorder substantially increases the risk for type 2 DM, independent of the effects of obesity.

Lipodystrophies are group of heterogeneous disorders characterized by selective loss of adipose tissue, leading to severe insulin resistance and hypertriglyceridemia. Lipodystrophies can be inherited or acquired and associated with variable degrees of adipose tissue loss.

Prevention

Type 2 DM is preceded by a period of IGT or IFG, and a number of lifestyle modifications and pharmacologic agents prevent or delay the onset of DM. Individuals with prediabetes or increased risk of diabetes should be referred to a structured program to reduce body weight and increase physical activity as well as being screened for cardiovascular disease. The Diabetes Prevention Program (DPP) demonstrated that intensive changes in lifestyle (diet and exercise for 30 min/d five times/week) in individuals with IGT prevented or delayed the development of type 2 DM by 58% compared to placebo. This effect was seen in individuals regardless of age, sex, or ethnic group. In the same study, metformin prevented or delayed diabetes by 31% compared to placebo. The lifestyle intervention group lost 5–7% of their body weight during the 3 years of the study; the effects of the intervention persisted for at least 15 years. Studies in Finnish and Chinese populations noted similar efficacy of diet and exercise in preventing or delaying type 2 DM. A number of agents, including α-glucosidase inhibitors, metformin, thiazolidinediones, GLP-1 receptor pathway modifiers, and orlistat, prevent or delay type 2 DM but are not approved by the Food and Drug Administration for this purpose. Individuals with a strong family history of type 2 DM and individuals with IFG or IGT should be strongly encouraged to maintain a normal BMI and engage in regular physical activity. Pharmacologic therapy for individuals with prediabetes is currently controversial because its cost-effectiveness and safety profile are not known. The ADA suggests that metformin be considered in individuals with both IFG and IGT who are at very high risk for progression to diabetes (age <60 years, BMI ≥35 kg/m^2, and women with a history of GDM). Individuals with IFG, IGT, or an HbA_{1c} of 5.7–6.4% should be monitored annually to determine if diagnostic criteria for diabetes are present.

GENETICALLY DEFINED, MONOGENIC FORMS OF DM RELATED TO REDUCED INSULIN SECRETION

Several monogenic forms of DM have been identified. More than 10 different variants of MODY, caused by mutations in genes encoding islet-enriched transcription factors or glucokinase (Fig. 396-5; Table 396-1), are transmitted as autosomal dominant disorders. MODY 1, MODY 3, and MODY 5 are caused by mutations in hepatocyte nuclear transcription factor (HNF) 4α, HNF-1α, and HNF-1β, respectively. As their names imply, these transcription factors are expressed in the liver but also in other tissues, including the pancreatic islets and kidney. These factors most likely affect islet development or the expression of genes important in glucose-stimulated insulin secretion or the maintenance of beta cell mass. For example, individuals with an HNF-1α mutation (MODY 3) have a progressive decline in glycemic control but may respond to sulfonylureas.

In fact, some of these patients were initially thought to have type 1 DM but were later shown to respond to a sulfonylurea, and insulin was discontinued. Individuals with a HNF-1β mutation have progressive impairment of insulin secretion and hepatic insulin resistance, and require insulin treatment with minimal response to sulfonylureas. These individuals often have other abnormalities such as renal cysts, mild pancreatic exocrine insufficiency, and abnormal liver function tests. Individuals with MODY 2, the result of mutations in the glucokinase gene, have mild-to-moderate, but stable hyperglycemia that does not respond to oral hypoglycemic agents. Glucokinase catalyzes the formation of glucose-6-phosphate from glucose, a reaction that is important for glucose sensing by the beta cells (Fig. 396-5) and for glucose utilization by the liver. As a result of glucokinase mutations, higher glucose levels are required to elicit insulin secretory responses, thus altering the set point for insulin secretion. MODY 4 is a rare variant caused by mutations in pancreatic and duodenal homeobox 1, a transcription factor that regulates pancreatic development and insulin gene transcription. Homozygous inactivating mutations cause pancreatic agenesis, whereas heterozygous mutations may result in DM. Studies of populations with type 2 DM suggest that mutations in MODY-associated genes are an uncommon (<5%) cause of type 2 DM.

Transient or permanent neonatal diabetes (onset <6 months of age) occurs. Permanent neonatal diabetes is a heterogeneous group of disorders caused by genetic mutations that impact beta cell function and/or pancreatic development (Fig. 396-5). Affected individuals typically require treatment with insulin and exhibit phenotypic overlap with type 1 DM. Activating mutations in the ATP-sensitive potassium channel subunits (Kir6.2 and ABCC8) impair glucose-stimulated insulin secretion. However, these individuals may respond to sulfonylureas and can be treated with these agents. Mutations in the transcription factor *GATA6* are the most common cause of pancreatic agenesis. Homozygous glucokinase mutations cause a severe form of neonatal diabetes, while mutations in mitochondrial DNA are associated with diabetes and deafness. A number of mutations identified in the coding sequence of the insulin gene have been found to interfere with proinsulin folding, processing, and bioactivity and are designated as Mutant *Ins*-gene-induced Diabetes of Youth (MIDYs). Some of the neonatal diabetes syndromes are associated with a spectrum of neurologic dysfunction and a variety of extrapancreatic manifestations. Any individual who developed diabetes at 6 months of age or who has atypical features of type 1 or type 2 diabetes should be screened for forms of monogenic diabetes.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Diabetes Mellitus

Once the diagnosis of DM is made, attention should be directed to symptoms related to diabetes (acute and chronic) and classifying the type of diabetes. DM and its complications produce a wide range of symptoms and signs; those secondary to acute hyperglycemia may occur at any stage of the disease, whereas those related to chronic hyperglycemia typically begin to appear during the second decade of hyperglycemia (**Chap. 398**). Because of long delays in clinical recognition, individuals with previously undetected type 2 DM may present with chronic complications of DM at the time of diagnosis. The history and physical examination

should assess for symptoms or signs of acute hyperglycemia and screen for chronic microvascular and macrovascular complications and conditions associated with DM **(Chap. 398)**.

HISTORY

A complete medical history should be obtained with special emphasis on DM-relevant aspects such as current weight as well as any recent changes in weight, family history of DM and its complications, sleep history, risk factors for cardiovascular disease, exercise, smoking status, history of pancreatic disease, and ethanol use. Symptoms of hyperglycemia include polyuria, polydipsia, weight loss, fatigue, weakness, blurry vision, frequent superficial infections (vaginitis, fungal skin infections), and slow healing of skin lesions after minor trauma. Metabolic derangements relate mostly to hyperglycemia (osmotic diuresis) and to the catabolic state of the patient (urinary loss of glucose and calories, muscle breakdown due to protein degradation and decreased protein synthesis). Blurred vision results from changes in the water content of the lens and resolves as hyperglycemia is controlled.

In a patient with established DM, the initial assessment should include a review of symptoms at the time of the initial diabetes diagnosis. This is an essential part of the history that can help define whether the correct type of DM has been diagnosed. Special emphasis should be placed on prior diabetes care, including types of therapies tried, the nature of any intolerance to previous therapies, prior HbA_{1c} levels, self-monitoring blood glucose results, frequency of hypoglycemia (<3.0 mmol/L, <54 mg/dL), presence of DM-specific complications, and assessment of the patient's knowledge about diabetes, exercise, nutrition, and sleep history. Diabetes-related complications may afflict several organ systems, and an individual patient may exhibit some, all, or none of the symptoms related to the complications of DM **(Chap. 398)**. In addition, the presence of DM-related comorbidities should be established (cardiovascular disease, hypertension, dyslipidemia). Pregnancy plans should be ascertained in women of childbearing age. The American Diabetes Association recommends that all women of childbearing age be counseled about the importance of tight glycemic control (HbA_{1c} <6.5%) prior to conception.

PHYSICAL EXAMINATION

In addition to a complete physical examination, special attention should be given to DM-relevant aspects such as weight and BMI, retinal examination, orthostatic blood pressure, foot examination, peripheral pulses, and insulin injection sites. Blood pressure >130/80 mmHg is considered hypertension in individuals with diabetes. Because periodontal disease is more frequent in DM, the teeth and gums should also be examined.

An annual foot examination should (1) assess blood flow (pedal pulses), sensation (vibratory sensation [128-MHz tuning fork at the base of the great toe], the ability to sense touch with a monofilament [5.07, 10-g monofilament], pinprick sensation, ankle reflexes, and nail care; (2) look for the presence of foot deformities such as hammer or claw toes and Charcot foot; and (3) identify sites of potential ulceration. The ADA recommends annual screening for distal symmetric polyneuropathy beginning with the initial diagnosis of diabetes and annual screening for autonomic neuropathy 5 years after diagnosis of type 1 DM and at the time of diagnosis of type 2 DM. This testing is aimed at detecting loss of protective sensation (LOPS) caused by diabetic neuropathy **(Chap. 398)**.

CLASSIFICATION OF DM IN AN INDIVIDUAL PATIENT

The etiology of diabetes in an individual with new-onset disease can usually be assigned on the basis of clinical criteria. Individuals with type 1 DM tend to have the following characteristics: (1) onset of disease prior to age 30 years; (2) lean body habitus; (3) requirement of insulin as the initial therapy; (4) propensity to develop ketoacidosis; and (5) an increased risk of other autoimmune disorders such as autoimmune thyroid disease, adrenal insufficiency, pernicious anemia, celiac disease, and vitiligo. In contrast, individuals with type 2 DM often exhibit the following features: (1) diabetes onset after the age of 30 years; (2) are usually obese (80% are obese, but elderly individuals may be lean); (3) may not require insulin therapy initially; and (4) may have associated conditions such as insulin resistance, hypertension, cardiovascular disease, dyslipidemia, or polycystic ovarian syndrome. In type 2 DM, insulin resistance is often associated with abdominal obesity (as opposed to hip and thigh obesity) and hypertriglyceridemia. Although most individuals diagnosed with type 2 DM are older, the age of diagnosis is declining, and there is a marked increase among overweight children and adolescents. Some individuals with phenotypic type 2 DM present with diabetic ketoacidosis but lack autoimmune markers and may be later treated with oral glucose-lowering agents rather than insulin (this clinical picture is sometimes referred to as *ketosis-prone type 2 DM*). On the other hand, some individuals (5–10%) with the phenotypic appearance of type 2 DM do not have absolute insulin deficiency but have autoimmune markers (GAD and other ICA autoantibodies) suggestive of type 1 DM (termed *latent autoimmune diabetes of the adult*). Such individuals are more likely to be <50 years of age, thinner, and have a personal or family history of other autoimmune disease than individuals with type 2 DM. They are much more likely to require insulin treatment within 5 years. Monogenic forms of diabetes (discussed above) should be considered in those with diabetes onset in childhood or early adulthood and especially those diagnosed within the first 6 months of life, an autosomal pattern of diabetes inheritance, diabetes without typical features of type 1 or 2 diabetes, and stable mild fasting hyperglycemia. Genetic testing should be considered in individuals suspected of having a monogenic form of diabetes as this may guide therapy selection. Despite recent advances in the understanding of the pathogenesis of diabetes, it often remains difficult to categorize some patients unequivocally. Individuals who deviate from the clinical profile of type 1 and type 2 DM, or who have other associated defects such as deafness, pancreatic exocrine disease (type 3c DM), and other endocrine disorders, should be classified accordingly (Table 396-1).

LABORATORY ASSESSMENT

The laboratory assessment should first determine whether the patient meets the diagnostic criteria for DM (Table 396-2) and then assess the degree of glycemic control **(Chap. 397)**. In addition to the standard laboratory evaluation, the patient should be screened for DM-associated conditions (e.g., albuminuria, dyslipidemia, thyroid dysfunction).

The classification of the type of DM may be facilitated by laboratory assessments. Serum insulin or C-peptide measurements may be useful, but should always be interpreted with a concurrent blood glucose level. A low C-peptide in the setting of an elevated blood glucose level may confirm a patient's need for insulin. However, C-peptide levels are unable to completely distinguish type 1 from type 2 DM, as many individuals with type 1 DM retain some C-peptide production. Measurement of islet cell antibodies at the time of diabetes onset may be useful if the type of DM is not clear based on the characteristics described above.

FURTHER READING

Atkinson MA et al: Type 1 diabetes. Lancet 383:69, 2014. [PubMed: 23890997]

Classification and Diagnosis of Diabetes. Diabetes Care 41:S13, 2018. [PubMed: 29222373]

Dall TM et al: The economic burden of elevated blood glucose levels in 2012: Diagnosed and undiagnosed diabetes, gestational diabetes mellitus, and prediabetes. Diabetes Care 37:3172, 2014. [PubMed: 25414388]

Fuchsberger C et al: The genetic architecture of type 2 diabetes. Nature 536:41, 2016. [PubMed: 27398621]

Hattersley AT, Patel KA: Precision diabetes: Learning from monogenic diabetes. Diabetologia 60:769, 2017. [PubMed: 28314945]

Insel RA et al: Staging presymptomatic type 1 diabetes: A scientific statement of JDRF, the Endocrine Society, and the American Diabetes Association. Diabetes Care 38:1964, 2015. [PubMed: 26404926]

Selph S et al: Screening for type 2 diabetes mellitus: A systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 162:765, 2015. [PubMed: 25867111]

Skyler JS et al: Differentiation of diabetes by pathophysiology, natural history, and prognosis. Diabetes 66:241, 2017. [PubMed: 27980006]

Chapter 397: Diabetes Mellitus: Management and Therapies

Alvin C. Powers; Kevin D. Niswender; Michael R. Rickels

OVERALL GOALS

The goals of therapy for type 1 or type 2 diabetes mellitus (DM) are to (1) eliminate symptoms related to hyperglycemia, (2) reduce or eliminate the long-term microvascular and macrovascular complications of DM (**Chap. 398**), and (3) allow the patient to achieve as normal a lifestyle as possible. To reach these goals, the physician should identify a target level of glycemic control for each patient, provide the patient with the educational and pharmacologic resources necessary to reach this level, and monitor/treat DM-related complications. Symptoms of diabetes usually resolve when the plasma glucose is <11.1 mmol/L (200 mg/dL), and thus most DM treatment focuses on achieving the second and third goals. This chapter first reviews the ongoing treatment of diabetes in the outpatient setting and then discusses the treatment of severe hyperglycemia, as well as the treatment of diabetes in hospitalized patients.

The care of an individual with either type 1 or type 2 DM requires a multidisciplinary team. Central to the success of this team are the patient's participation, input, and enthusiasm, all of which are essential for optimal diabetes management. Members of the health care team include the primary care provider and/or the endocrinologist or diabetologist, a certified diabetes educator, a nutritionist, a psychologist, and/or social worker. In addition, when the complications of DM arise, subspecialists (including ophthalmologists, neurologists, podiatrists, nephrologists, transplant surgeons, cardiologists, and cardiovascular surgeons) with experience in DM-related complications are essential.

ONGOING ASPECTS OF COMPREHENSIVE DIABETES CARE

A number of names are sometimes applied to different approaches to diabetes care, such as intensive insulin therapy, intensive glycemic control, and "tight control." The current chapter, and other sources, uses the term *comprehensive diabetes care* to emphasize the fact that optimal diabetes therapy involves more than plasma glucose management and medications. Although glycemic control is central to optimal diabetes therapy, comprehensive diabetes care of both type 1 and type 2 DM should also detect and manage DM-specific complications (**Chap. 398**), and modify risk factors for DM-associated diseases. The key elements of comprehensive diabetes care are summarized in **Table 397-1**. The morbidity and mortality of DM can be greatly reduced by timely and consistent surveillance, including the detection, prevention, and management of DM-related complications (Table 397-1 and **Chap. 398**). Such screening procedures are indicated for all individuals with DM, but many individuals with diabetes do not receive these or comprehensive diabetes care. In addition to the physical aspects of DM, social, family, financial, cultural, and employment-related issues may impact diabetes care. The treatment goals for patients with diabetes summarized in **Table 397-2** should be individualized. The prevention and treatment of clinically significant hypoglycemia (<3.0 mmol/L or 54 mg/dL) is discussed in **Chap. 399**. The International Diabetes Federation (IDF), recognizing that resources available for diabetes care vary widely throughout the world, has issued guidelines for "recommended care" (a well-developed service base and with health care funding systems consuming a significant part of their national wealth), "limited care" (health care settings with very limited resources), and "comprehensive care" (health care settings with considerable resources). This chapter provides guidance for this comprehensive level of diabetes care.

TABLE 397-1

Guidelines for Ongoing, Comprehensive Medical Care for Patients with Diabetes

- Individualized glycemic goal and therapeutic plan
- Self-monitoring of blood glucose (individualized frequency)
- HbA_{1c} testing (2–4 times/year)
- Lifestyle management in the care of diabetes, including:
  - Diabetes-self-management education and support
  - Nutrition therapy
  - Physical activity
  - Psychosocial care, including evaluation for depression, anxiety
- Detection, prevention, or management of diabetes-related complications, including:
  - Diabetes-related eye examination (annual or biannual; **Chap. 398**)
  - Diabetes-related foot examination (1–2 times/year by provider; daily by patient; **Chap. 398**)
  - Diabetes-related neuropathy examination (annual; **Chap. 398**)
  - Diabetes-related kidney disease testing (annual; **Chap. 398**)
- Manage or treat diabetes-relevant conditions, including:
  - Blood pressure (assess quarterly; **Chap. 398**)
  - Lipids (annual; **Chap. 398**)
  - Consider antiplatelet therapy **(Chap. 398)**
  - Influenza/pneumococcal/hepatitis B immunizations **(Chap. 4)**

Abbreviation: HbA_{1c}, hemoglobin A_{1c}.

TABLE 397-2

Treatment Goals for Adults with Diabetes[a]

Index	Goal
Glycemic control[b]	
HbA_{1c}	<7.0%[c]
Preprandial capillary plasma glucose	4.4–7.2 mmol/L (80–130 mg/dL)
Postprandial capillary plasma glucose[d]	<10.0 mmol/L (<180 mg/dL)
Blood pressure	<140/90 mmHg[e]

[a]As recommended by the American Diabetes Association; goals should be individualized for each patient (see text) with a different goals for different patients. [b]HbA_{1c} is primary goal. [c]Diabetes Control and Complications Trial-based assay. [d]1–2 h after beginning of a meal. [e]The ADA also advises individualization of the BP goal with consideration of other co-morbidities and adverse events of therapy. A goal of of <130/80 mmHg may be appropriate for younger individuals or individuals with cardiovascular risk factors.

Abbreviation: HbA_{1c}, hemoglobin A_{1c}.

Source: Adapted from American Diabetes Association: Diabetes Care 41(Suppl 1):S86, 2018.

LIFESTYLE MANAGEMENT IN DIABETES CARE

The patient with type 1 or type 2 DM should receive education about nutrition, exercise, psychosocial support, care of diabetes during illness, and medications to lower the plasma glucose. The American Diabetes Association (ADA) uses the term “Lifestyle Management” to refer to aspects of diabetes care, including: (1) diabetes self-management education (DSME) and diabetes self-management support (DSMS); (2) nutrition therapy; and (3) psychosocial care. Along with improved compliance, patient education allows individuals with DM to assume greater responsibility for his/her care. Patient education should be viewed as a continuing process with regular visits for reinforcement; it should not be a process that is completed after one or two visits to a nurse educator or nutritionist. DSME and DSMS are ways to improve the patient’s knowledge, skills, and abilities necessary for diabetes self-care and should also emphasize psychosocial issues and emotional well-being. More frequent contact between the patient and the diabetes management team (e.g., electronic, telephone) improves glycemic control.

Diabetes Self-Management Education and Support

The diabetes educator is a health care professional (nurse, dietician, or pharmacist) with specialized patient education skills who is certified in diabetes education (e.g., American Association of Diabetes Educators). Education topics important for optimal diabetes self-care include self-monitoring of blood glucose (SMBG); urine ketone monitoring (type 1 DM); insulin administration; guidelines for diabetes management during illnesses; prevention and management of hypoglycemia **(Chap. 399)**; foot and skin care; diabetes management before, during, and after exercise; and risk factor-modifying activities. The focus is providing patient-centered, individualized education.

Nutrition Therapy

Medical nutrition therapy (MNT) is a term used by the ADA to describe the optimal coordination of caloric intake with other aspects of diabetes therapy (insulin, exercise, and weight loss). Primary measures of MNT are directed at preventing or delaying the onset of type 2 DM in high-risk individuals (obese or with prediabetes) by promoting weight reduction. Medical treatment of obesity is a rapidly evolving area and is discussed in **Chap. 395**. Secondary measures of MNT are directed at improving glycemic control. Tertiary measures of MNT are directed at managing diabetes-related complications (cardiovascular disease [CVD], nephropathy). Although the recommendations for all three types of MNT overlap, this chapter emphasizes secondary measures of MNT. Pharmacologic approaches that facilitate weight loss and metabolic surgery should be considered in selected patients (**Chaps. 394** and **395**).

In general, the components of optimal MNT are similar for individuals with type 1 or type 2 DM and similar to those for the general population—high quality, nutrient-dense without a specific focus on composition (Mediterranean, dietary approaches to stop hypertension, etc.; **Table 397-3**). Historically, nutrition education imposed restrictive, complicated regimens on the patient. Current practices have greatly changed, although many patients and health care providers still view the diabetic diet as monolithic and static. There is not a specific diet for individuals with diabetes, certainly not a one diet for everyone. For example, MNT now includes foods with some sucrose and seeks to modify other risk factors such as hyperlipidemia and hypertension. Using the *glycemic index*, an estimate of the postprandial rise in the blood glucose when a certain amount of that food is consumed, may reduce postprandial glucose excursions and improve glycemic control.

TABLE 397-3

Nutritional Recommendations for Adults with Diabetes or Prediabetes[a]

General dietary guidelines

- Vegetable, fruits, whole grains, legumes, low-fat dairy products in food higher in fiber and lower in glycemic content

Fat in diet (optimal % of diet is not known; should be individualized)

- Mediterranean-style diet rich in monounsaturated fatty acids
- Minimal *trans* fat consumption

Carbohydrate in diet (optimal % of diet is not known; should be individualized)

- Monitor carbohydrate intake in regard to calories
- Sucrose-containing foods may be consumed with adjustments in insulin dose, but minimize intake
- Estimate grams of carbohydrate in diet (type 1 DM)
- Consider using glycemic index to predict how consumption of a particular food may affect blood glucose
- Fructose preferred over sucrose

Protein in diet (optimal % of diet is not known; should be individualized)

Other components

- Reduced-calorie and nonnutritive sweeteners may be useful
- Routine supplements of vitamins, antioxidants, or trace elements not supported by evidence
- Sodium intake as advised for general population

[a]See text for differences for patients with type 1 or type 2 diabetes.

Source: Adapted from American Diabetes Association: Diabetes Care 37(Suppl 1):S14, 2014 with modifications from American Diabetes Association: Diabetes Care 41(Suppl 1):S36, 2018.

The goal of MNT in type 1 DM is to coordinate and match the caloric intake, both temporally and quantitatively, with the appropriate amount of insulin. MNT in type 1 DM and SMBG should be integrated to define the optimal insulin regimen. The ADA

encourages patients and providers to use carbohydrate counting to estimate the nutrient content of a meal or snack. Based on the patient's estimate of the carbohydrate content of a meal, an insulin-to-carbohydrate ratio determines the bolus insulin dose for a meal or snack. MNT must be flexible enough to allow for exercise, and the insulin regimen must allow for deviations in caloric intake. An important component of MNT in type 1 DM is to minimize the weight gain often associated with intensive insulin therapy.

The goals of MNT in type 2 DM should focus on weight loss and address the greatly increased prevalence of cardiovascular risk factors (hypertension, dyslipidemia, obesity) and disease in this population. The majority of these individuals are obese, and weight loss is strongly encouraged. Hypocaloric diets and modest weight loss (5–7%) often result in rapid and dramatic glucose lowering in individuals with new-onset type 2 DM. MNT for type 2 DM should emphasize modest caloric reduction and increased physical activity. Weight loss and exercise improve insulin sensitivity.

Fasting for religious reasons, such as during Ramadan, presents a challenge for individuals with diabetes, especially those taking medications to lower the plasma glucose. Under IDF-issued guidelines on fasting, individuals are risk-stratified as those who can safely fast with medical evaluation and supervision and those in whom fasting is not advised.

Physical Activity

Exercise has multiple positive benefits including cardiovascular risk reduction, reduced blood pressure, maintenance of muscle mass, reduction in body fat, and weight loss. For individuals with type 1 or type 2 DM, exercise is also useful for lowering plasma glucose (during and following exercise) and increasing insulin sensitivity. In patients with diabetes, the ADA recommends 150 min/week (distributed over at least 3 days) of moderate aerobic physical activity with no gaps longer than 2 days. Resistance exercise, flexibility and balance training, and reduced sedentary behavior throughout the day are advised.

Despite its benefits, exercise presents challenges for individuals with DM because they lack the normal glucoregulatory mechanisms (normally, insulin falls and glucagon rises during exercise). Skeletal muscle is a major site for metabolic fuel consumption in the resting state, and the increased muscle activity during vigorous, aerobic exercise greatly increases fuel requirements. Individuals with type 1 DM are prone to either hyperglycemia or hypoglycemia during exercise, depending on the preexercise plasma glucose, the circulating insulin level, and the level of exercise-induced catecholamines. If the insulin level is too low, the rise in catecholamines may increase the plasma glucose excessively, promote ketone body formation, and possibly lead to ketoacidosis. Conversely, if the circulating insulin level is excessive, this relative hyperinsulinemia may reduce hepatic glucose production (decreased glycogenolysis, decreased gluconeogenesis) and increase glucose entry into muscle, leading to hypoglycemia.

To avoid exercise-related hyper- or hypoglycemia, individuals with type 1 DM should (1) monitor blood glucose before, during, and after exercise; (2) delay exercise if blood glucose is >14 mmol/L (250 mg/dL) and ketones are present; (3) if the blood glucose is <5.6 mmol/L (100 mg/dL), ingest carbohydrate before exercising; (4) monitor glucose during exercise and ingest carbohydrate to prevent hypoglycemia; (5) decrease insulin doses (based on previous experience) before and after exercise and inject insulin into a nonexercising area; and (6) learn individual glucose responses to different types of exercise. In individuals with type 2 DM, exercise-related hypoglycemia is less common but can occur in individuals taking either insulin or insulin secretagogues. Untreated proliferative retinopathy is a relative contraindication to vigorous exercise, because this may lead to vitreous hemorrhage or retinal detachment (**Chap. 398**).

Psychosocial Care

Because the individual with DM faces challenges that affect many aspects of daily life, psychosocial assessment and support are a critical part of comprehensive diabetes care. Patients should view themselves as essential members of the diabetes care team and not as someone who is cared for by the diabetes management team. Even with considerable effort, normoglycemia can be an elusive goal, and solutions to worsening glycemic control may not be easily identifiable. Depression, anxiety, or "Diabetes Distress," defined by the ADA as "…negative psychological reactions related to emotional burdens…in having to manage a

chronic disease like diabetes, should be recognized and may require the care of a mental health specialist." Emotional stress may provoke a change in behavior so that individuals no longer adhere to a dietary, exercise, or therapeutic regimen. The individual with DM must accept that he or she may develop complications related to DM. Eating disorders, including binge eating disorders, bulimia, and anorexia nervosa, appear to occur more frequently in individuals with type 1 or type 2 DM.

MONITORING THE LEVEL OF GLYCEMIC CONTROL

Optimal monitoring of glycemic control involves plasma glucose measurements by the patient and an assessment of long-term control by the providers on the diabetes management team (measurement of hemoglobin A_{1c} [HbA_{1c}] and review of the patient's SMBG). These measurements are complementary: the patient's measurements provide a picture of short-term glycemic control, whereas the HbA_{1c} reflects average glycemic control over the previous 2–3 months.

Self-Monitoring of Blood Glucose

SMBG is the standard of care in diabetes management and allows the patient to monitor his or her blood glucose at any time. In SMBG, a small drop of blood and an easily detectable enzymatic reaction allow measurement of the capillary plasma glucose. Many glucose monitors can rapidly and accurately measure glucose (calibrated to provide plasma glucose value even though blood glucose is measured) in small amounts of blood (3–10 μL) obtained from the fingertip; alternative testing sites (e.g., forearm) are less reliable. By combining glucose measurements with diet and exercise history, and medication changes, the diabetes management team and patient can improve the treatment program.

The frequency of SMBG measurements must be individualized and adapted to address the goals of diabetes care. Individuals with type 1 DM or individuals with type 2 DM taking multiple insulin injections each day should routinely measure their plasma glucose three or more times per day (some measure >10 times/day) to estimate and select mealtime boluses of short-acting insulin and to modify long-acting insulin doses. Most individuals with type 2 DM require less frequent monitoring, although the optimal frequency of SMBG has not been clearly defined. Individuals with type 2 DM who are taking insulin should use SMBG more frequently than those on oral agents. Individuals with type 2 DM who are on oral medications should use SMBG as a means of assessing the efficacy of their medication and the impact of dietary choices and exercise. Because plasma glucose levels fluctuate less in these individuals, one or fewer SMBG measurements per day may be sufficient. Most measurements in individuals with type 1 or type 2 DM should be performed prior to a meal and supplemented with postprandial measurements to assist in reaching postprandial glucose targets (Table 397-2).

Devices for continuous glucose monitoring (CGM) usually do not replace the need for traditional glucose measurements and require calibration by SMBG. These rapidly evolving technologies require substantial expertise on the part of the diabetes management team and the patient. Current CGM systems measure the glucose in interstitial fluid, which is in equilibrium with the plasma glucose. These devices provide useful short-term information about the patterns of glucose changes as well as an enhanced ability to detect hypoglycemic episodes. Alerts and alarms (vibration, sound) can notify the patient if the glucose is rising or falling rapidly, or is predicted to cross a hyper- or hypoglycemic threshold. Clinical experience with these devices in type 1 DM is growing rapidly, especially in individuals who have not achieved glycemic targets, those with hypoglycemia unawareness to decrease the frequency of serious hypoglycemia (especially nocturnal hypoglycemia), and those desiring more frequent glycemic feedback. The combination of an insulin-infusion device (discussed below) and a CGM are currently open-loop, meaning the patient must adjust the insulin-infusion device, but closed-loop systems (insulin-infusion device automatically adjusted by algorithm) may soon be entering clinical practice; one system that adjusts the basal rate has been recently approved by the U.S. Food and Drug Administration (FDA).

Assessment of Long-Term Glycemic Control

Measurement of glycated hemoglobin (HbA_{1c}) is the standard method for assessing long-term glycemic control. When plasma glucose is consistently elevated, there is an increase in nonenzymatic glycation of hemoglobin; this alteration reflects the

glycemic history over the previous 2–3 months, because erythrocytes have an average life span of 120 days (glycemic level in the preceding month contributes about 50% to the HbA_{1c} value). Measurement of HbA_{1c} at the "point of care" allows for more rapid feedback and may therefore assist in adjustment of therapy.

HbA_{1c} should be measured in all individuals with DM during their initial evaluation and as part of their comprehensive diabetes care. As the primary predictor of long-term complications of DM, the HbA_{1c} should mirror, to a certain extent, the short-term measurements of SMBG. These two measurements are complementary in that recent intercurrent illnesses may impact the SMBG measurements but not the HbA_{1c}. Likewise, postprandial and nocturnal hyperglycemia may not be detected by the SMBG of fasting and preprandial capillary plasma glucose but will be reflected in the HbA_{1c}. The HbA_{1c} is an "average" and thus does not detect glycemic variability in the way SMBG and CGM can. In standardized assays, the HbA_{1c} approximates the following mean plasma glucose values: an HbA_{1c} of 6% = 7.0 mmol/L (126 mg/dL), 7% = 8.6 mmol/L (154 mg/dL), 8% = 10.2 mmol/L (183 mg/dL), 9% = 11.8 mmol/L (212 mg/dL), 10% = 13.4 mmol/L (240 mg/dL), 11% = 14.9 mmol/L (269 mg/dL), and 12% = 16.5 mmol/L (298 mg/dL). However, there is interindividual variability in the HbA_{1c} to mean glucose relationship, and in African-Americans the HbA_{1c} is on average 0.4% higher than in Caucasians for the same mean glucose. Clinical conditions leading to abnormal RBC parameters such as hemoglobinopathies, anemias, reticulocytosis, transfusions, and uremia may alter the HbA_{1c} result. In patients achieving their glycemic goal, the ADA recommends measurement of the HbA_{1c} at least twice per year. More frequent testing (every 3 months) is warranted when glycemic control is inadequate or when therapy has changed. Laboratory standards for the HbA_{1c} test have been established and should be correlated to the reference assay of the Diabetes Control and Complications Trial (DCCT). The degree of glycation of other proteins, such as albumin, or measurement of 1,5-anhydroglucitol can be used as an alternative indicator of glycemic control when the HbA_{1c} is inaccurate. The fructosamine assay (measuring glycated albumin) reflects the glycemic status over the prior 2 weeks.

PHARMACOLOGIC TREATMENT OF DIABETES

Comprehensive care of type 1 and type 2 DM requires an emphasis on nutrition, exercise, and monitoring of glycemic control but also usually involves glucose-lowering medication(s). This chapter discusses classes of such medications but does not describe every glucose-lowering agent available worldwide. The initial step is to select an individualized, glycemic goal for the patient.

ESTABLISHMENT OF TARGET LEVEL OF GLYCEMIC CONTROL

Because the complications of DM are related to glycemic control, normoglycemia or near-normoglycemia is the desired, but often elusive, goal for most patients. Normalization or near-normalization of the plasma glucose for long periods of time is extremely difficult, as demonstrated by the DCCT and United Kingdom Prospective Diabetes Study (UKPDS). Regardless of the level of hyperglycemia, improvement in glycemic control will lower the risk of diabetes-specific complications, most notably the microvascular complications (Chap. 398).

The target for glycemic control (as reflected by the HbA_{1c}) must be individualized, and the goals of therapy should be developed in consultation with the patient after considering a number of medical, social, and lifestyle issues. The ADA calls this a *patient-centered approach*, and other organizations such as the IDF and American Association of Clinical Endocrinologists (AACE) also suggest an individualized glycemic goal. Important factors to consider include the patient's age and ability to understand and implement a complex treatment regimen, presence and severity of complications of diabetes, known CVD, ability to recognize hypoglycemic symptoms, presence of other medical conditions or treatments that might affect survival or the response to therapy, lifestyle and occupation (e.g., possible consequences of experiencing hypoglycemia on the job), and level of support available from family and friends.

In general, the ADA suggests that the goal is to achieve an HbA_{1c} as close to normal as possible without significant hypoglycemia. In most individuals, the target HbA_{1c} should be <7% (Table 397-2) with a more stringent target ≤6.5% for some patients. With modern implementation of intensive insulin therapy for type 1 DM, the level of HbA_{1c} is no longer inversely related to the frequency and severity of hypoglycemia as seen in the DCCT; nevertheless, it may still be appropriate to set a higher HbA_{1c} target <7.5 or 8% for patients with impaired awareness of hypoglycemia. A higher HbA_{1c} goal may also be appropriate for the very young or old or in individuals with limited life span or comorbid conditions. For example, an appropriate HbA_{1c} goal in elderly individuals with multiple, chronic illnesses and impaired activities of daily living might be 8.0 or 8.5%.

More stringent glycemic control (HbA_{1c} of ≤6%) is not beneficial, and may be detrimental, in patients with type 2 DM and a high risk of CVD. Large clinical trials (UKPDS, Action to Control Cardiovascular Risk in Diabetes [ACCORD], Action in Diabetes and Vascular Disease: Preterax and Diamicron MR Controlled Evaluation [ADVANCE], Veterans Affairs Diabetes Trial [VADT]; **Chap. 398**) have examined glycemic control in type 2 DM in individuals with low risk of CVD, with high risk of CVD, or with established CVD and have found that more intense glycemic control is not beneficial and, in some patient populations, may have a negative impact on some outcomes. These divergent outcomes stress the need for individualized glycemic goals based on the following general guidelines: (1) early in the course of type 2 diabetes when the CVD risk is lower, improved glycemic control likely leads to improved cardiovascular outcome, but this benefit may occur more than a decade after the period of improved glycemic control; (2) intense glycemic control in individuals with established CVD or at high risk for CVD is not advantageous, and may be deleterious, over a follow-up of 3–5 years; (3) hypoglycemia in such high-risk populations (elderly, CVD) should be avoided; and (4) improved glycemic control reduces microvascular complications of diabetes **(Chap. 398)** even if it does not improve macrovascular complications like CVD.

TYPE 1 DIABETES MELLITUS

General Aspects

The ADA recommendations for fasting and bedtime glycemic goals and HbA_{1c} targets are summarized in Table 397-2. The goal is to design and implement insulin regimens that mimic physiologic insulin secretion. Because individuals with type 1 DM partially or completely lack endogenous insulin production, administration of basal insulin is essential for regulating glycogen breakdown, gluconeogenesis, lipolysis, and ketogenesis (i.e., largely fine-tuning hepatic and adipose metabolism). Likewise, insulin replacement for meals should be appropriate for the carbohydrate intake and promote normal glucose utilization and storage.

Intensive Management

Intensive insulin therapy has the goal of achieving near-normal glycemia. This approach requires multiple resources, including thorough and continuing patient education, comprehensive recording of plasma glucose measurements and nutrition intake by the patient, and a variable insulin regimen that matches carbohydrate intake and insulin dose. Insulin regimens include multiple-component insulin regimens, multiple daily injections (MDIs), or continuous subcutaneous (SC) insulin infusion (CSII) (each discussed below).

The benefits of intensive insulin therapy and improved glycemic control include a reduction in the acute metabolic and chronic microvascular complications of DM. From a psychological standpoint, the patient experiences greater control over his or her diabetes and often notes an improved sense of well-being, greater flexibility in the timing and content of meals, and the capability to alter insulin dosing with exercise. In addition, intensive insulin therapy prior to and during pregnancy reduces the risk of fetal malformations and morbidity. Intensive insulin therapy is encouraged in newly diagnosed patients with type 1 DM because it may prolong the period of C-peptide production, which may result in better glycemic control and a reduced risk of

serious hypoglycemia. Although intensive management confers impressive benefits, it is also accompanied by significant personal and financial costs and is therefore not appropriate for all individuals.

Insulin Preparations

Current insulin preparations are generated by recombinant DNA technology and consist of the amino acid sequence of human insulin or variations thereof. In the United States, most insulin is formulated as U-100 (100 units/mL); short-acting insulin formulated as U-200 (200 units/mL; lispro) and long-acting as U-300 (300 units/mL; glargine) are available in order to limit injection volumes for patients with high insulin requirements. Regular insulin formulated as U-500 (500 units/mL) is sometimes used in patients with severe insulin resistance. Human insulin has been formulated with distinctive pharmacokinetics (regular and neutral protamine Hagedorn [NPH] insulin have the native insulin amino acid sequence) or genetically modified to alter insulin absorption and hence insulin action. Insulins can be classified as short-acting or long-acting **(Table 397-4)**. For example, one short-acting insulin formulation, insulin lispro, is an insulin analogue in which the 28th and 29th amino acids (lysine and proline) on the insulin B chain have been reversed by recombinant DNA technology. Insulin aspart and insulin glulisine are genetically modified insulin analogues with properties similar to lispro. A biosimilar version of lispro has been approved. All three of these insulin analogues have full biologic activity but less tendency for self-aggregation, resulting in more rapid absorption and onset of action and a shorter duration of action. These characteristics are particularly advantageous for allowing entrainment of insulin injection and action to rising plasma glucose levels following meals. The shorter duration of action also appears to be associated with a decreased number of hypoglycemic episodes, primarily because the decay of insulin action corresponds to the decline in plasma glucose after a meal. Thus, insulin aspart, lispro, or glulisine is preferred over regular insulin for prandial coverage in many patients. Insulin glargine is a long-acting biosynthetic human insulin that differs from normal insulin in that asparagine is replaced by glycine at amino acid 21, and two arginine residues are added to the C terminus of the B chain, leading to the formation of microprecipitates at physiologic pH in subcutaneous tissue. Compared to NPH insulin, the onset of insulin glargine action is later, the duration of action is longer (~24 h), and there is a less pronounced peak. A lower incidence of hypoglycemia, especially at night, has been reported with insulin glargine when compared to NPH insulin. A biosimilar version is now available. Insulin detemir has a fatty acid side chain that reversibly binds to albumin and prolongs its action by slowing absorption and catabolism, but its duration of action may only reach 12–20 h. Twice-daily injections of glargine, or especially detemir, are sometimes required to provide optimal 24-h coverage. Because of modification and extension of the carboxy-terminal terminus of the B chain, insulin degludec forms multihexamers in subcutaneous tissue and binds albumin, prolonging its duration of action (>42 h); it provides similar glycemic control as glargine but with less frequent nocturnal and severe hypoglycemia.

TABLE 397-4

Properties of Insulin Preparations[a]

	Time of Action		
Preparation	**Onset, h**	**Peak, h**	**Effective Duration, h**
Short-acting[b]			
Aspart	<0.25	0.5–1.5	2–4
Glulisine	<0.25	0.5–1.5	2–4
Lispro[f]	<0.25	0.5–1.5	2–4
Regular[g]	0.5–1.0	2–3	3–6
Inhaled human insulin	0.5–1.0	2–3	3
Long-acting[g]			
Degludec	1–9	—[c]	42[d]
Detemir	1–4	—[c]	12–24[d]
Glargine[f]	2–4	—[c]	20–24
NPH	2–4	4–10	10–16
Examples of insulin combinations[e]			
75/25–75% protamine lispro, 25% lispro	<0.25	Dual[f]	10–16
70/30–70% protamine aspart, 30% aspart	<0.25	Dual[f]	15-18
50/50–50% protamine lispro, 50% lispro	<0.25	Dual[f]	10–16
70/30–70% NPH, 30% regular	0.5–1	Dual[f]	10–16
Combination of long-acting insulin and GLP-1 receptor agonist	See text		

[a]Injectable insulin preparations (with exception of inhaled formulation) available in the United States; others are available in the United Kingdom and Europe. [b]Formulation with niacinamide has a slightly more rapid onset and offset.[c]Degludec, determir, and glargine have minimal peak activity. [d]Duration is dose-dependent. [e]Other insulin combinations are available. [f]Dual: two peaks—one at 2–3 h and the second one several hours later. [g]Also available in concentrations >U-100.

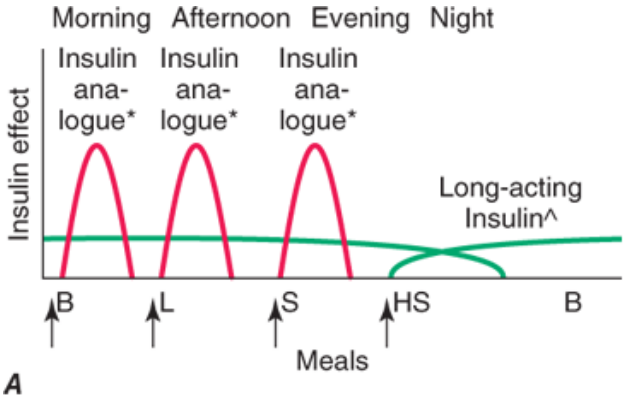
Basal insulin requirements are provided by long-acting insulin formulations (NPH insulin, insulin glargine, insulin detemir, or insulin degludec). These are usually prescribed with short-acting insulin in an attempt to mimic physiologic insulin release with meals. Although mixing of NPH and short-acting insulin formulations is common practice, this mixing may alter the insulin absorption profile (especially the short-acting insulins). For example, lispro absorption is delayed by mixing with NPH. The alteration in insulin absorption when the patient mixes different insulin formulations should not prevent mixing insulins. However, the following guidelines should be followed: (1) mix the different insulin formulations in the syringe immediately before injection (inject within 2 min after mixing); (2) do not store insulin as a mixture; (3) follow the same routine in terms of insulin mixing and administration to standardize the physiologic response to injected insulin; and (4) do not mix insulin glargine, detemir, or degludec with other insulins. The miscibility of some insulins allows for the production of combination insulins that contain 70% NPH and 30% regular (70/30), or equal mixtures of NPH and regular (50/50). By including the insulin analogue mixed with protamine, several additional combinations have a short-acting and long-acting profile (Table 397-4). Although more convenient for the patient (only two injections/day), combination insulin formulations do not allow independent adjustment of short-acting and long-acting activity. Several insulin formulations are available as insulin "pens," which are more convenient for some patients. Other novel insulins, such as one with a duration of action of several days, are in development. Insulin delivery by inhalation to provide meal-time insulin is approved, but not widely used. Prior to its use, the forced expiratory volume in one second (FEV_1) should be measured. Inhaled insulin can cause bronchospasm and cough and should be not be in individuals with lung disease or who smoke. Long-acting insulin/glucagon-like peptide-1 (GLP-1) receptor agonist combinations in fixed doses (degludec + liraglutide or glargine + lixisenatide) have recently become available, are effective, and are not associated with weight gain.

Insulin Regimens

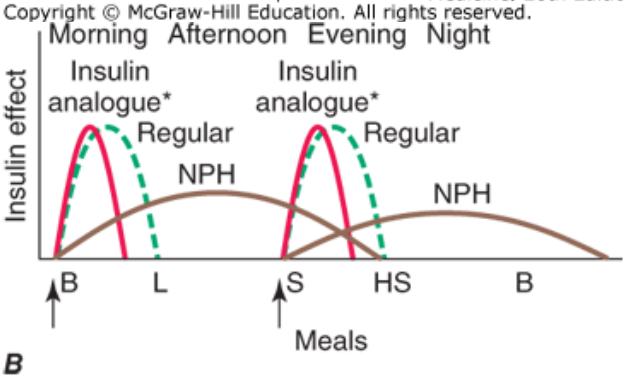
Representations of the various insulin regimens that may be used in type 1 DM are illustrated in **Fig. 397-1**. Although the insulin profiles are depicted as "smooth," symmetric curves, there is considerable patient-to-patient variation in the peak and duration. In all regimens, long-acting insulins (NPH, glargine, detemir, or degludec) supply basal insulin, whereas regular, insulin aspart, glulisine, or lispro provide prandial insulin. Short-acting insulin analogues should be injected just before (<10 min) and regular insulin 30–45 min prior to a meal. Sometimes short-acting insulin analogues are injected just after a meal (gastroparesis, unpredictable food intake).

FIGURE 397-1

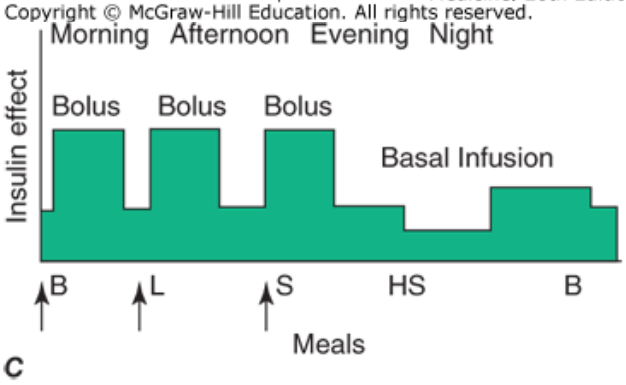
Representative insulin regimens for the treatment of diabetes. For each panel, the *y*-axis shows the amount of insulin effect and the *x*-axis shows the time of day. B, breakfast; HS, bedtime; L, lunch; S, supper. *Lispro, glulisine, or insulin aspart can be used. The time of insulin injection is shown with a *vertical arrow*. The type of insulin is noted above each insulin curve. ***A.*** Multiple-component insulin regimen consisting of long-acting insulin (degludec, determir, or glargine) to provide basal insulin coverage and three shots of glulisine, lispro, or insulin aspart to provide glycemic coverage for each meal. ***B.*** Injection of two shots of long-acting insulin (NPH) and short-acting insulin analogue (glulisine, lispro, insulin aspart [*solid red line*], or regular insulin [*green dashed line*]). Some deliver the second dose of NPH at bedtime or also use a short-acting insulin at lunch. Only one formulation of short-acting insulin is used. ***C.*** Insulin administration by insulin infusion device is shown with the basal insulin and a bolus injection at each meal. The basal insulin rate is decreased during the evening and increased slightly prior to the patient awakening in the morning. Glulisine, lispro, or insulin aspart is used in the insulin infusion device. *(Part C adapted from FR Kaufman: Medical Management of Type 1 Diabetes, 6th ed. American Diabetes Association, Alexandria, VA, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A shortcoming of current insulin regimens is that injected insulin immediately enters the systemic circulation, whereas endogenous insulin is secreted into the portal venous system. Thus, exogenous insulin administration exposes the liver to subphysiologic insulin levels. No insulin regimen reproduces the precise insulin secretory pattern of the pancreatic islet. However, the most physiologic regimens entail more frequent insulin injections, greater reliance on short-acting insulin, and more frequent capillary plasma glucose measurements (or by CGM). In general, individuals with type 1 DM require 0.4–1 units/kg per day of insulin divided into multiple doses, with ~50% of the insulin given as basal insulin.

MDI regimens refer to the combination of basal insulin and bolus insulin (preprandial short-acting insulin). The timing and dose of short-acting, preprandial insulin are altered to accommodate the SMBG results, anticipated food intake, and physical activity. Such regimens offer the patient with type 1 DM more flexibility in terms of lifestyle and the best chance for achieving near normoglycemia. One such regimen, shown in Fig. 397-1B, consists of basal insulin with glargine, detemir, or degludec and preprandial lispro, glulisine, or insulin aspart. The insulin aspart, glulisine, or lispro dose is based on individualized algorithms that integrate the preprandial glucose and the anticipated carbohydrate intake. To determine the meal component of the

preprandial insulin dose, the patient uses an insulin-to-carbohydrate ratio (a common ratio for type 1 DM is 1 unit/10–15 g of carbohydrate, but this must be determined for each individual). To this insulin, dose is added the supplemental or correcting insulin based on the preprandial blood glucose (one formula uses 1 unit of insulin for every 2.7 mmol/L [50 mg/dL] over the preprandial glucose target; another formula uses [body weight in kg] × [blood glucose – desired glucose in mg/dL]/1500). Such calculations must be adjusted based on each individual's sensitivity to insulin. An alternative multiple-component insulin regimen consists of bedtime NPH insulin, a small dose of NPH insulin at breakfast (20–30% of bedtime dose), and preprandial short-acting insulin. Other variations of this regimen are in use but have the disadvantage that NPH has a significant peak, making hypoglycemia more common. Frequent SMBG (≥4 times per day) is essential for these types of insulin regimens, although less frequent SMBG may be acceptable when used together with CGM.

In the past, one commonly used regimen consisted of twice-daily injections of NPH mixed with a short-acting insulin before the morning and evening meals (Fig. 397-1B). Such regimens usually prescribe two-thirds of the total daily insulin dose in the morning (with about two-thirds given as long-acting insulin and one-third as short-acting) and one-third before the evening meal (with approximately one-half given as long-acting insulin and one-half as short-acting). The drawback to such a regimen is that it forces a rigid schedule on the patient, in terms of daily activity and the content and timing of meals. Moreover, if the patient's meal pattern or content varies or if physical activity is increased, hyperglycemia or hypoglycemia may result. Moving the long-acting insulin from before the evening meal to bedtime may avoid nocturnal hypoglycemia and provide more insulin as glucose levels rise in the early morning as growth hormone and cortisol secretion peak (so-called dawn phenomenon). The insulin dose in such regimens should be adjusted based on SMBG results with the following general assumptions: (1) the fasting glucose is primarily determined by the prior evening long-acting insulin; (2) the prelunch glucose is a function of the morning short-acting insulin; (3) the presupper glucose is a function of the morning long-acting insulin; and (4) the bedtime glucose is a function of the presupper, short-acting insulin. This is not an optimal regimen for the patient with type 1 DM, but is sometimes used for patients with insulin-requiring type 2 DM.

CSII is a very effective insulin regimen for the patient with type 1 DM **(Fig. 397-1C)**. To the basal insulin infusion, a preprandial insulin ("bolus") is delivered by the insulin infusion device based on instructions from the patient, who uses an individualized algorithm incorporating the preprandial plasma glucose and anticipated carbohydrate intake. These sophisticated devices can accurately deliver small doses of insulin (microliters per hour) and have several advantages: (1) multiple basal infusion rates can be programmed to accommodate nocturnal versus daytime basal insulin requirement; (2) basal infusion rates can be altered during periods of exercise; (3) different waveforms of insulin infusion with meal-related bolus allow better matching of insulin depending on meal composition; and (4) programmed algorithms consider ongoing action of prior insulin administration and blood glucose values in calculating the insulin dose. These devices require instruction by a health professional with considerable experience with insulin infusion devices and very frequent patient interactions with the diabetes management team. Insulin infusion devices present unique challenges, such as infection at the infusion site, unexplained hyperglycemia because the infusion set becomes obstructed, or diabetic ketoacidosis (DKA) if the insulin infusion device becomes disconnected. Because most physicians use lispro, glulisine, or insulin aspart in CSII, the extremely short half-life of these insulins quickly leads to insulin deficiency if the delivery system is interrupted. Essential to the safe use of infusion devices is thorough patient education, frequent SMBG (or by CGM), and a backup safety plan in the event of insulin infusion device failure. CGM sensor-augmented insulin infusion devices integrate the information from the CGM to inform insulin delivery. Currently, sensor communicating functions can interrupt basal insulin delivery during hypoglycemia (threshold suspension) or when hypoglycemia is anticipated (predictive suspension [not available in the United States]), which may be particularly useful for addressing nocturnal hypoglycemia. A partial closed-loop system has recently become available that combines patient-directed preprandial boluses with automated adjustment of between meal and basal insulin delivery based on CGM. Clinical experience with closed-loop systems is limited but increasing.

Other Agents That Improve Glucose Control

The role of amylin, a 37-amino-acid peptide co-secreted with insulin from pancreatic beta cells, in normal glucose homeostasis is uncertain. However, based on the rationale that patients who are insulin deficient are also amylin deficient, an analogue of amylin (pramlintide) was created and found to reduce postprandial glycemic excursions in type 1 and type 2 diabetic patients taking insulin. Pramlintide injected just before a meal slows gastric emptying and suppresses glucagon but does not alter insulin levels. Pramlintide is approved for insulin-treated patients with type 1 and type 2 DM. Addition of pramlintide produces a modest reduction in the HbA_{1c} and seems to dampen meal-related glucose excursions. In type 1 DM, pramlintide is started as a 15-µg SC injection before each meal and titrated up to a maximum of 30–60 µg as tolerated. In type 2 DM, pramlintide is started as a 60-µg SC injection before each meal and may be titrated up to a maximum of 120 µg. The major side effects are nausea and vomiting, and dose escalations should be slow to limit these side effects. Because pramlintide slows gastric emptying, it may influence absorption of other medications and should not be used in combination with other drugs that slow gastrointestinal (GI) motility. The short-acting insulin given before the meal should initially be reduced to avoid hypoglycemia and then titrated as the effects of the pramlintide become evident. Because pramlintide suppresses glucagon, it may worsen hypoglycemia recovery and should not be used in patients with hypoglycemia unawareness.

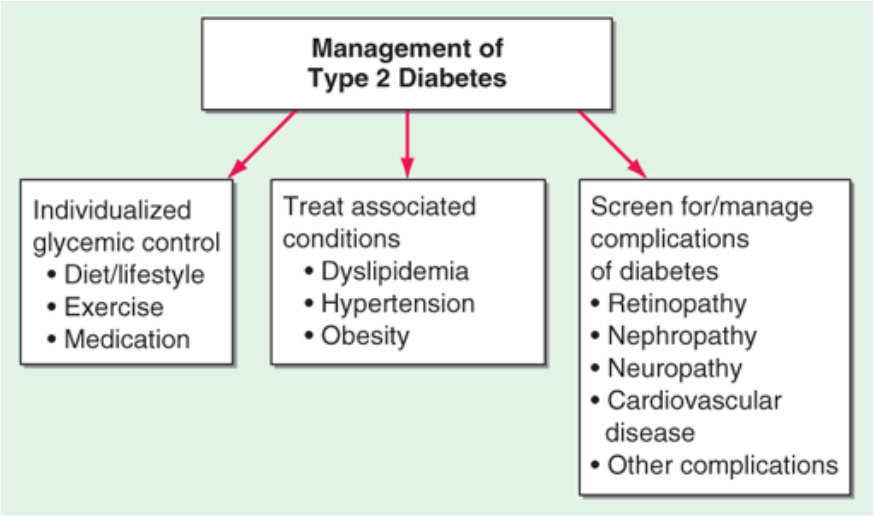
TYPE 2 DIABETES MELLITUS

General Aspects

The goals of glycemia-controlling therapy for type 2 DM are similar to those in type 1 DM. Whereas glycemic control tends to dominate the management of type 1 DM, the care of individuals with type 2 DM must also include attention to the treatment of conditions associated with type 2 DM (e.g., obesity, hypertension, dyslipidemia, CVD) and detection/management of DM-related complications **(Fig. 397-2; Chap. 398)**. Reduction in cardiovascular risk is of paramount importance because this is the leading cause of mortality in these individuals.

FIGURE 397-2

Essential elements in comprehensive care of type 2 diabetes.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Type 2 DM management should begin with MNT (discussed above). An exercise regimen to increase insulin sensitivity and promote weight loss should also be instituted. Pharmacologic approaches to the management of type 2 DM include oral glucose-lowering agents, insulin, and other agents that improve glucose control; most physicians and patients prefer oral glucose-lowering agents as the initial choice. Any therapy that improves glycemic control reduces "glucose toxicity" to beta cells and may improve endogenous insulin secretion. However, type 2 DM is a progressive disorder and ultimately requires multiple therapeutic agents and often insulin in most patients.

Glucose-Lowering Agents

Advances in the therapy of type 2 DM have generated oral glucose-lowering agents that target different pathophysiologic processes in type 2 DM. Based on their mechanisms of action, glucose-lowering agents are subdivided into agents that increase insulin secretion, reduce glucose production, increase insulin sensitivity, enhance GLP-1 action, or promote urinary excretion of glucose **(Table 397-5)**. Glucose-lowering agents other than insulin (with the exception of amylin analogue) are ineffective in type 1 DM and should not be used for glucose management of severely ill individuals with type 2 DM. Insulin is sometimes the initial glucose-lowering agent in type 2 DM.

TABLE 397-5

Agents Used for Treatment of Type 1 or Type 2 Diabetes

	Mechanism of Action	Examples[a]	HbA_{1c} Reduction (%)[b]	Agent-Specific Advantages	Agent-Specific Disadvantages	Contraindications
Oral						
Biguanides[c*]	↓ Hepatic glucose production	Metformin	1–2	Weight neutral, do not cause hypoglycemia, inexpensive, extensive experience, ↓ CV events	Diarrhea, nausea, lactic acidosis, vitamin B12 deficiency	Renal insufficiency (see text for GFR <45 mL/min), CHF, radiographic contrast studies, hospitalized patients, acidosis
α-Glucosidase inhibitors[c**]	↓ GI glucose absorption	Acarbose, miglitol, voglibose	0.5–0.8	Reduce postprandial glycemia	GI flatulence, liver function tests	Renal/liver disease
Dipeptidyl peptidase IV inhibitors[c***]	Prolong endogenous GLP-1 action; ↑ Insulin, ↓ glucagon	Alogliptin, linagliptin, saxagliptin, sitagliptin, vildagliptin	0.5–0.8	Well tolerated, do not cause hypoglycemia	Angioedema/urticarial and immune-mediated dermatologic effects	Reduced dose with renal disease
Insulin secretagogues: Sulfonylureas[c*]	↑ Insulin secretion	Glibornuride, gliclazide, glimepiride, glipizide, gliquidone, glyburide, glyclopyramide	1–2	Short onset of action, lower postprandial glucose, inexpensive	Hypoglycemia, weight gain	Renal/liver disease
Insulin secretagogues: Nonsulfonylureas[c***]	↑ Insulin secretion	Mitiglinide nateglinide, repaglinide	0.5–1.0	Short onset of action, lower postprandial glucose	Hypoglycemia	Renal/liver disease

	Mechanism of Action	Examples[a]	HbA_{1c} Reduction (%)[b]	Agent-Specific Advantages	Agent-Specific Disadvantages	Contraindications
Sodium-glucose cotransporter 2 inhibitors***	↑ renal glucose excretion	Canagliflozin, dapagliflozin, empagliflozin, ertugliflozin	0.5–1.0	do not cause hypoglycemia, ↓ weight and BP; see text for CVD effect	Urinary and genital infections, polyuria, dehydration, exacerbate tendency to hyperkalemia and DKA; see text	Moderate renal insufficiency, insulin-deficient DM
Thiazolidinediones[c]***	↓ Insulin resistance, ↑ glucose utilization	Pioglitazone, rosiglitazone	0.5–1.4	Lower insulin requirements	Peripheral edema, CHF, weight gain, fractures, macular edema	CHF, liver disease
Parenteral						
Amylin agonists[c,d]***	Slow gastric emptying, ↓ glucagon	Pramlintide	0.25–0.5	Reduce postprandial glycemia, weight loss	Injection, nausea, ↑ risk of hypoglycemia with insulin	Agents that also slow GI motility
GLP-1 receptor agonists[c]***	↑ Insulin, ↓ glucagon, slow gastric emptying, satiety	Albiglutide, dulaglutide, exenatide, liraglutide, lixisenatide, semaglutide	0.5–1.0	Weight loss, do not cause hypoglycemia; see text for CVD effect	Injection, nausea, ↑ risk of hypoglycemia with insulin secretagogues	Renal disease, agents that also slow GI motility; medullary carcinoma of thyroid, pancreatic disease
Insulin[c,d]****	↑ Glucose utilization, ↓ hepatic glucose production, and other anabolic actions	See text and Table 397-4	Not limited	Known safety profile	Injection, weight gain, hypoglycemia	
Medical nutrition therapy and physical activity[c]*	↓ Insulin resistance, ↑ insulin secretion	Low-calorie, low-fat diet, exercise	1–3	Other health benefits	Compliance difficult, long-term success low	

[a]Examples are approved for use in the United States; others are available in other countries. Examples may not include all agents in the class. [b]HbA*1c* reduction (absolute) depends partly on starting HbA_{1c}. [c]Used for treatment of type 2 diabetes. [d]Used in conjunction with insulin for treatment of type 1 diabetes. Cost of agent in the United States: [*]low, [**]moderate, [***]high, [****]variable.

Note: Some agents used to treat type 2 DM are not included in table (see text).

Abbreviations: ACE, angiotensin-converting enzyme; CHF, congestive heart failure; CV, cardiovascular; GI, gastrointestinal; HbA_{1c}, hemoglobin A_{1c}.

BIGUANIDES

Metformin, representative of this class of agents, reduces hepatic glucose production and improves peripheral glucose utilization slightly (Table 397-5). Metformin activates AMP-dependent protein kinase and enters cells through organic cation transporters (polymorphisms of these may influence the response to metformin). Recent evidence indicates that metformin's mechanism for reducing hepatic glucose production is to antagonize glucagon's ability to generate cAMP in hepatocytes. Metformin reduces fasting plasma glucose (FPG) and insulin levels, improves the lipid profile, and promotes modest weight loss. An extended-release form is available and may have fewer GI side effects (diarrhea, anorexia, nausea, metallic taste). Because of its relatively slow onset of action and GI symptoms with higher doses, the initial dose should be low and then escalated every 1–2 weeks based on SMBG measurements to a maximally tolerated dose of 2000 mg daily. Metformin is effective as monotherapy and can be used in combination with other oral agents or with insulin. Long-term use is associated with reduced micro- and probably macrovascular complications, but the data are less conclusive for macrovascular complications. The major toxicity of metformin, lactic acidosis, is very rare and can be prevented by careful patient selection. Vitamin B_{12} levels are lower during metformin treatment and should be monitored. Metformin should not be used in patients with moderate renal insufficiency (glomerular filtration rate [GFR] <45 mL/min), any form of acidosis, unstable congestive heart failure (CHF), liver disease, or severe hypoxemia. The National Institute for Health and Clinical Excellence in the United Kingdom suggests that metformin may be safe at a GFR >30 mL/min, with a reduced dose when the GFR is <45 mL/min. Metformin should be discontinued in hospitalized patients, in patients who can take nothing orally, and in those receiving radiographic contrast material. Insulin should be used until metformin can be restarted.

INSULIN SECRETAGOGUES—AGENTS THAT AFFECT THE ATP-SENSITIVE K^+ CHANNEL

Insulin secretagogues stimulate insulin secretion by interacting with the ATP-sensitive potassium channel on the beta cell **(Chap. 396)**. These drugs are most effective in individuals with type 2 DM of relatively recent onset (<5 years) who have residual endogenous insulin production. First-generation sulfonylureas (chlorpropamide, tolazamide, tolbutamide) have a longer half-life, a greater incidence of hypoglycemia, and more frequent drug interactions, and are no longer used. Second-generation sulfonylureas have a more rapid onset of action and better coverage of the postprandial glucose rise, but the shorter half-life of some agents may require more than once-a-day dosing. Sulfonylureas reduce both fasting and postprandial glucose and should be initiated at low doses and increased at 1- to 2-week intervals based on SMBG. In general, sulfonylureas increase insulin acutely and thus should be taken shortly before a meal; with chronic therapy, though, the insulin release is more sustained. Long-term use is associated with reduced micro- and macrovascular complications. Glimepiride and glipizide can be given in a single daily dose and are preferred over glyburide, especially in the elderly. Repaglinide, nateglinide, and mitiglinide are not sulfonylureas but also interact with the ATP-sensitive potassium channel. Because of their short half-life, these agents are given immediately before each meal to reduce meal-related glucose excursions.

Insulin secretagogues, especially the longer acting ones, have the potential to cause hypoglycemia, especially in elderly individuals. Hypoglycemia is usually related to delayed meals, increased physical activity, alcohol intake, or renal insufficiency. Individuals who ingest an overdose of some agents develop prolonged and serious hypoglycemia and should be monitored closely in the hospital **(Chap. 399)**. Most sulfonylureas are metabolized in the liver to compounds (some of which are active) that are cleared by the kidney. Thus, their use in individuals with significant hepatic or renal dysfunction is not advisable. Weight gain, a common side effect of sulfonylurea therapy, results from the increased insulin levels and improvement in glycemic control. Some sulfonylureas have significant drug interactions with alcohol and some medications including warfarin, aspirin,

ketoconazole, α-glucosidase inhibitors, and fluconazole. A related isoform of ATP-sensitive potassium channels is present in the myocardium and the brain. All of these agents except glyburide have a low affinity for this isoform. Despite concerns that this agent might affect the myocardial response to ischemia and observational studies suggesting that sulfonylureas increase cardiovascular risk, studies have not shown an increased cardiac mortality with glyburide or other agents in this class.

INSULIN SECRETAGOGUES—AGENTS THAT ENHANCE GLP-1 RECEPTOR SIGNALING

"Incretins" amplify glucose-stimulated insulin secretion **(Chap. 396)**. Agents that either act as a GLP-1 receptor agonist or enhance endogenous GLP-1 activity are approved for the treatment of type 2 DM (Table 397-5). Agents in this class do not cause hypoglycemia because of the glucose-dependent nature of incretin-stimulated insulin secretion (unless there is concomitant use of an agent that can lead to hypoglycemia—sulfonylureas, etc.). GLP-1 receptor agonists increase glucose-stimulated insulin secretion, suppress glucagon, and slow gastric emptying. These agents do not promote weight gain; in fact, most patients experience modest weight loss and appetite suppression. Short-acting GLP-1 receptor agonists are exanatide and lixisenatide. Long-acting GLP-1 receptor agonists include liraglutide, exenatide, albiglutide, dulaglutide, and lixisenatide. Short-acting GLP-1 receptor agonists provide mostly postprandial coverage whereas the long-acting GLP-1 receptor agonists reduce both the postprandial and fasting glucose.

For example, exenatide, a synthetic version of a peptide initially identified in the saliva of the Gila monster (exendin-4), is an analogue of GLP-1. Unlike native GLP-1, which has a half-life of ~2 min, differences in the exenatide amino acid sequence render it resistant to the enzyme that degrades GLP-1 (dipeptidyl peptidase IV [DPP-IV]). Thus, exenatide has prolonged GLP-1-like action and binds to GLP-1 receptors found in islets, the GI tract, and the brain. Liraglutide, another GLP-1 receptor agonist, is almost identical to native GLP-1 except for an amino acid substitution and addition of a fatty acyl group (coupled with a γ-glutamic acid spacer) that promote binding to albumin and plasma proteins and prolong its half-life. Higher doses of liraglutide than used for glucose-lowering effects have been approved for weight loss therapy for obesity. Liraglutide treatment has also been associated with a decrease in CVD events in patients with type 2 DM and established CVD and with lower rates of diabetic kidney disease. In a similar patient population, semaglutide treatment was associated with fewer CVD events and reduced diabetic kidney disease, but with an increased rate of retinopathy-related complications. Whether the effect on CVD is a drug class effect is not clear as other GLP-1 receptor agonists have not reduced CVD events. Treatment with these agents should start at a low dose to minimize initial side effects (nausea being the limiting one). GLP-1 receptor agonists can be used as combination therapy with metformin, sulfonylureas, and thiazolidinediones. Some patients taking insulin secretagogues may require a reduction in those agents to prevent hypoglycemia. The major side effects are nausea, vomiting, and diarrhea. Some formulations carry a black box warning from the FDA because of an increased risk of thyroid C-cell tumors in rodents and are contraindicated in individuals with medullary carcinoma of the thyroid, multiple endocrine neoplasia, or pancreatic disease. Because GLP-1 receptor agonists slow gastric emptying, they may influence the absorption of other drugs. Whether GLP-1 receptor agonists enhance beta cell survival or promote beta cell proliferation in humans as in rodents is not known, but these agents do not appear to alter the natural history of type 2 DM.

DPP-IV inhibitors inhibit degradation of native GLP-1 and thus enhance the incretin effect. DPP-IV, which is widely expressed on the cell surface of endothelial cells and some lymphocytes, degrades a wide range of peptides (not GLP-1 specific). DPP-IV inhibitors promote insulin secretion in the absence of hypoglycemia or weight gain and appear to have a preferential effect on postprandial blood glucose. The levels of GLP-1 action in the patient are greater with the GLP-1 receptor agonists than with DPP-IV inhibitors. DPP-IV inhibitors are used either alone or in combination with other oral agents in type 2 DM. Reduced doses should be given to patients with renal insufficiency. There is conflicting evidence concerning a potentially increased risk for acute pancreatitis with GLP-1 receptor agonists and DPP-IV inhibitors, although initial concerns about possible premalignant lesions appear to be unfounded. For now, it is reasonable to avoid these agents in patients with pancreatic disease or with other significant risk factors for acute pancreatitis (e.g., heavy alcohol use, severely elevated serum triglycerides, hypercalcemia).

ALPHA-GLUCOSIDASE INHIBITORS

α-Glucosidase inhibitors reduce postprandial hyperglycemia by delaying glucose absorption; they do not affect glucose utilization or insulin secretion (Table 397-5). Postprandial hyperglycemia, secondary to impaired hepatic and peripheral glucose disposal, contributes significantly to the hyperglycemic state in type 2 DM. These drugs, taken just before each meal, reduce

glucose absorption by inhibiting the enzyme that cleaves oligosaccharides into simple sugars in the intestinal lumen. Therapy should be initiated at a low dose with the evening meal and increased to a maximal dose over weeks to months. The major side effects (diarrhea, flatulence, abdominal distention) are related to increased delivery of oligosaccharides to the large bowel and can be reduced somewhat by gradual upward dose titration. α-Glucosidase inhibitors may increase levels of sulfonylureas and increase the incidence of hypoglycemia. Simultaneous treatment with bile acid resins and antacids should be avoided. These agents should not be used in individuals with inflammatory bowel disease, gastroparesis, or a serum creatinine >177 µmol/L (2 mg/dL). This class of agents is not as potent as other oral agents in lowering the HbA_{1c} but is unique because it reduces the postprandial glucose rise. If hypoglycemia from other diabetes treatments occurs while taking these agents, the patient should consume glucose because the degradation and absorption of complex carbohydrates will be retarded.

THIAZOLIDINEDIONES

Thiazolidinediones (Table 397-5) reduce insulin resistance by binding to the peroxisome proliferator-activated receptor γ (PPAR-γ) nuclear receptor (which forms a heterodimer with the retinoid X receptor). The PPAR-γ receptor is found at highest levels in adipocytes but is expressed at lower levels in many other tissues. Agonists of this receptor regulate a large number of genes, promote adipocyte differentiation, reduce hepatic fat accumulation, and promote fatty acid storage. Thiazolidinediones promote a redistribution of fat from central to peripheral locations. Circulating insulin levels decrease with use of the thiazolidinediones, indicating a reduction in insulin resistance. Although direct comparisons are not available, the two currently available thiazolidinediones appear to have similar efficacy. The prototype of this class of drugs, troglitazone, was withdrawn from the U.S. market after reports of hepatotoxicity and an association with an idiosyncratic liver reaction that sometimes led to hepatic failure. Although rosiglitazone and pioglitazone do not appear to induce the liver abnormalities seen with troglitazone, the FDA recommends measurement of liver function tests prior to initiating therapy. Modestly increased transaminase levels related to underlying fatty liver disease should not preclude treatment as these levels may improve with thiazolidinediones due to a reduction in hepatic fat content.

Rosiglitazone raises low-density lipoprotein (LDL), high-density lipoprotein (HDL), and triglycerides slightly. Pioglitazone raises HDL to a greater degree and LDL a lesser degree but lowers triglycerides. The clinical significance of the lipid changes with these agents is not known and may be difficult to ascertain because most patients with type 2 DM are also treated with a statin.

Thiazolidinediones are associated with weight gain (2–3 kg), a small reduction in the hematocrit, and a mild increase in plasma volume. Peripheral edema and CHF are more common in individuals treated with these agents. These agents are contraindicated in patients with liver disease or CHF (class III or IV). The FDA has issued an alert that rare patients taking these agents may experience a worsening of diabetic macular edema. An increased risk of fractures has been noted in postmenopausal women taking these agents. Thiazolidinediones have been shown to induce ovulation in premenopausal women with polycystic ovary syndrome. Women should be warned about the risk of pregnancy because the safety of thiazolidinediones in pregnancy is not established.

Concerns about increased cardiovascular risk associated with rosiglitazone led to considerable restrictions on its use and to the FDA issuing a “black box” warning in 2007. However, based on new information, the FDA has revised its guidelines and categorizes rosiglitazone similar to other drugs for type 2 DM. According to a recent FDA pronouncement, pioglitazone may be associated with an increased risk of bladder cancer. In one study, pioglitazone lowered the risk for recurrent stroke or myocardial infarction in insulin-resistant individuals without diabetes who had a prior stroke or transient ischemic attack.

Sodium-Glucose Co-Transporter 2 (SGLT2) Inhibitors

These agents (Table 397-5) lower the blood glucose by selectively inhibiting this co-transporter, which is expressed almost exclusively in the proximal, convoluted tubule in the kidney. This inhibits glucose reabsorption, lowers the renal threshold for glucose, and leads to increased urinary glucose excretion. Thus, the glucose-lowering effect is insulin independent and not related to changes in insulin sensitivity or secretion. The loss of urinary glucose may promote modest weight reduction. Since these agents also impair proximal reabsorption of sodium, their use is associated with a diuretic effect and 3–6 mm Hg

reduction in systolic blood pressure. Due to the increased urinary glucose, urinary and genital mycotic infections are more common in both men and women, and the diuretic effect can lead to reduced intravascular volume and acutely impaired kidney function. Inhibition of SGLT2 on the alpha cell may lead to increased glucagon and consequently liver production of glucose and ketones. Euglycemic DKA may occur during illness or when ongoing glucosuria masks stress-induced requirements for insulin. These agents should not be prescribed for patients with type 1 DM or pancreatogenic forms of DM associated with insulin deficiency. Empagliflozin and canagliflozin reduces CVD events and all cause cardiovascular mortality in patients with type 2 DM and established CVD, the risk for nephropathy, and the rate of hospitalization for CHF. A possible increased risk of bladder cancer has been seen with dapagliflozin; canagliflozin is associated with an increased risk of leg and foot amputation and bone fractures.

OTHER THERAPIES FOR TYPE 2 DM

Bile acid–binding resins

Evidence indicates that bile acids, by signaling through nuclear receptors, may have a role in metabolism. Bile acid metabolism is abnormal in type 2 DM. The bile acid–binding resin colesevelam has been approved for the treatment of type 2 DM (already approved for treatment of hypercholesterolemia). Because bile acid–binding resins are minimally absorbed into the systemic circulation, how bile acid–binding resins lower blood glucose is not known. The most common side effects are GI (constipation, abdominal pain, and nausea). Bile acid–binding resins can increase plasma triglycerides and should be used cautiously in patients with a tendency for hypertriglyceridemia. The role of this class of drugs in the treatment of type 2 DM is not yet defined.

Bromocriptine

A formulation of the dopamine receptor agonist bromocriptine (Cycloset) has been approved by the FDA for the treatment of type 2 DM. However, its role in the treatment of type 2 DM is uncertain.

INSULIN THERAPY IN TYPE 2 DM

Insulin should be considered as part of the initial therapy in type 2 DM, particularly in lean individuals or those with severe weight loss, in individuals with underlying renal or hepatic disease that precludes oral glucose-lowering agents, or in individuals who are hospitalized or acutely ill. Insulin therapy is ultimately required by a substantial number of individuals with type 2 DM because of the progressive nature of the disorder and the relative insulin deficiency that develops in patients with long-standing diabetes. Both physician and patient reluctance often delay the initiation of insulin therapy, but glucose control and patient well-being are improved by insulin therapy in patients who have not reached glycemic targets.

Because endogenous insulin secretion continues and is capable of providing some coverage of mealtime caloric intake, insulin is usually initiated in a single dose of long-acting insulin (0.2–0.4 U/kg per day), given in the evening or just before bedtime (NPH, glargine, detemir, or degludec). Because fasting hyperglycemia and increased hepatic glucose production are prominent features of type 2 DM, bedtime insulin is more effective in clinical trials than a single dose of morning insulin. Glargine given at bedtime has less nocturnal hypoglycemia than NPH insulin. Some physicians prefer a relatively low, fixed starting dose of long-acting insulin (5–15 units) or a weight-based dose (0.1 units/kg). The insulin dose may then be adjusted in 10% increments as dictated by SMBG results. Both morning and bedtime long-acting insulin may be used in combination with oral glucose-lowering agents. Initially, basal insulin may be sufficient, but often prandial insulin coverage with multiple insulin injections is needed as diabetes progresses (see insulin regimens used for type 1 DM). Other insulin formulations that have a combination of short-acting and long-acting insulin (Table 397-4) are sometimes used in patients with type 2 DM because of convenience but do not allow independent adjustment of short-acting and long-acting insulin dose and often do not achieve the same degree of glycemic control as basal/bolus regimens. In selected patients with type 2 DM, insulin-infusion devices may be considered.

CHOICE OF INITIAL GLUCOSE-LOWERING AGENT

The level of hyperglycemia and the patient's individualized goal (see "Establishment of Target Level of Glycemic Control") should influence the initial choice of therapy. Assuming that maximal benefit of MNT and increased physical activity has been realized, patients with mild hyperglycemia (FPG <7.0–11.0 mmol/L [126–199 mg/dL]) often respond well to a single, oral glucose-lowering agent, while those with moderate hyperglycemia (FPG 11.1–13.9 mmol/L [200–250 mg/dL]) will usually require more than one oral agent or insulin. Patients with more severe hyperglycemia (FPG >13.9 mmol/L [250 mg/dL]) may respond partially but are unlikely to achieve normoglycemia with oral therapy. Insulin can be used as initial therapy in individuals with severe

hyperglycemia (FPG <13.9–16.7 mmol/L [250–300 mg/dL]) or in those who are symptomatic from the hyperglycemia. This approach is based on the rationale that more rapid glycemic control will reduce “glucose toxicity” to the islet cells, improve endogenous insulin secretion, and possibly allow oral glucose-lowering agents to be more effective. If this occurs, the insulin may be discontinued.

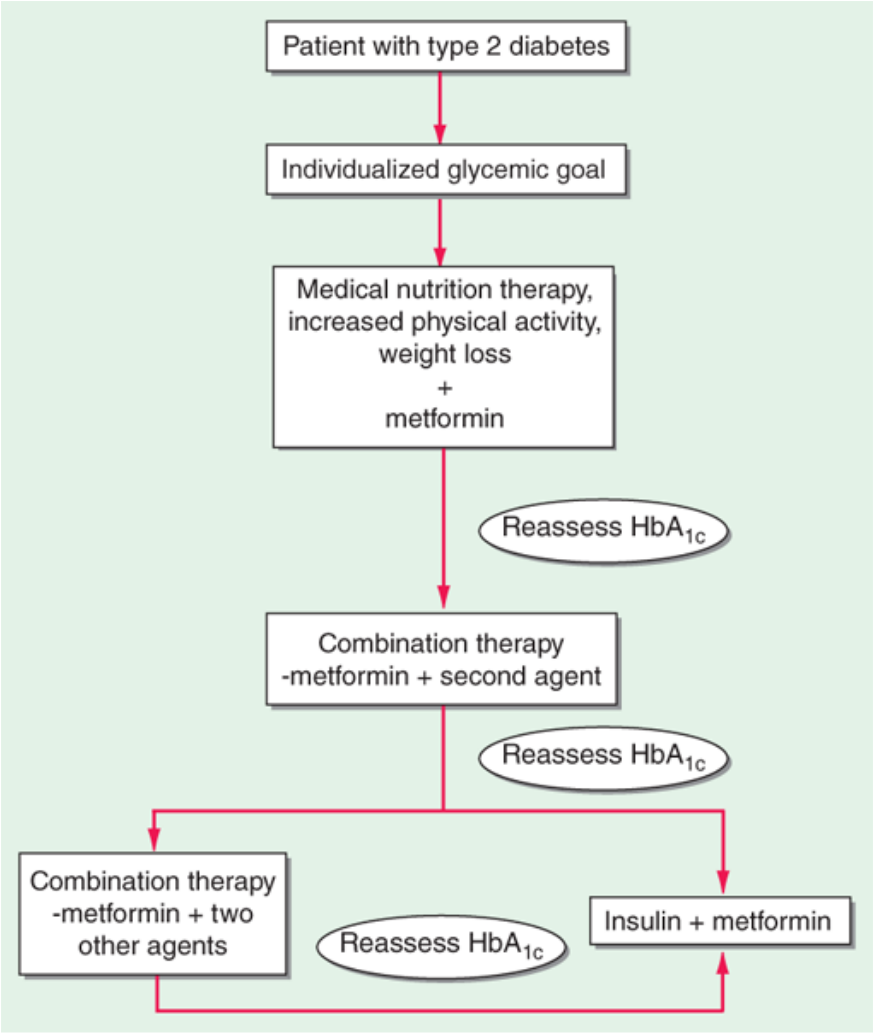
Insulin secretagogues, biguanides, α-glucosidase inhibitors, thiazolidinediones, GLP-1 receptor agonists, DPP-IV inhibitors, SLGT2 inhibitors, and insulin are approved for monotherapy of type 2 DM. Although each class of oral glucose-lowering agents has advantages and disadvantages (Table 397-5), certain generalizations apply: (1) insulin secretagogues, biguanides, GLP-1 receptor agonists, and thiazolidinediones improve glycemic control to a similar degree (1–2% reduction in HbA_{1c}) and are more effective than α-glucosidase inhibitors, DPP-IV inhibitors, and SLGT2 inhibitors; (2) assuming a similar degree of glycemic improvement, the clinical advantage of one class of drugs is not clear; any therapy that improves glycemic control is likely beneficial; (3) insulin secretagogues, GLP-1 receptor agonists, DPP-IV inhibitors, α-glucosidase inhibitors, and SLGT2 inhibitors begin to lower the plasma glucose immediately, whereas the glucose-lowering effects of the biguanides and thiazolidinediones are delayed by weeks; (4) not all agents are effective in all individuals with type 2 DM; (5) biguanides, α-glucosidase inhibitors, GLP-1 receptor agonists, DPP-IV inhibitors, thiazolidinediones, and SLGT2 inhibitors do not directly cause hypoglycemia; (6) most individuals will eventually require treatment with more than one class of oral glucose-lowering agents or insulin, reflecting the progressive nature of type 2 DM; and (7) durability of glycemic control is slightly less for sulfonylureas compared to metformin or thiazolidinediones.

Considerable clinical experience exists with metformin and sulfonylureas because they have been available for several decades. It is assumed that the α-glucosidase inhibitors, GLP-1 receptor agonists, DPP-IV inhibitors, thiazolidinediones, and SLGT2 inhibitors will reduce DM-related complications by improving glycemic control, but long-term data are not yet available. The thiazolidinediones are theoretically attractive because they target a fundamental abnormality in type 2 DM, namely insulin resistance.

Treatment algorithms by several professional societies (ADA/European Association for the Study of Diabetes [EASD], IDF, AACE) suggest metformin as initial therapy because of its efficacy, known side effect profile, and low cost **(Fig. 397-3)**. Metformin’s advantages are that it promotes mild weight loss, lowers insulin levels, and improves the lipid profile slightly. Based on SMBG results and the HbA_{1c}, the dose of metformin should be increased until the glycemic target is achieved or maximum dose is reached.

FIGURE 397-3

Glycemic management of type 2 diabetes. See text for discussion of treatment of severe hyperglycemia or symptomatic hyperglycemia. Agents that can be combined with metformin include insulin secretagogues, thiazolidinediones, α-glucosidase inhibitors, DPP-IV inhibitors, GLP-1 receptor agonists, SLGT2 inhibitors, and insulin. HbA_{1c}, hemoglobin HbA_{1c}.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

COMBINATION THERAPY WITH GLUCOSE-LOWERING AGENTS

A number of combinations of therapeutic agents are successful in type 2 DM metformin + second oral agent, metformin + GLP-1 receptor agonist, metformin + insulin, or combinations of a long-acting insulin and a GLP-1 receptor agonist. Because mechanisms of action of the first and second agents should be different, the effect on glycemic control is usually additive. There are little data to support the choice of one combination over another combination. Based on recent demonstrations of a beneficial cardiovascular effect in certain individuals with type 2 DM and CVD, or at high-risk of CVD, empagliflozin, canagliflozin, and liraglutide should now be considered in these populations. Medication costs vary considerably (Table 397-5), and this often factors into medication choice. Several fixed-dose combinations of oral agents are available, but evidence that they are superior to titration of single agent to a maximum dose and then addition of a second agent is lacking. If adequate control is not achieved with the combination of two agents (based on reassessment of the HbA_{1c} every 3 months), a third oral agent, GLP-1 receptor agonist, or basal insulin should be added (Fig. 397-3). Treatment approaches vary considerably from country to country. For example, α-glucosidase inhibitors are used commonly in South Asian patients (Indian), but infrequently in the United States or Europe. Whether this reflects an underlying difference in the disease or physician preference is not clear.

Treatment with insulin often becomes necessary as type 2 DM enters the phase of relative insulin deficiency and is signaled by inadequate glycemic control with one or two oral glucose-lowering agents. Insulin alone or in combination should be used in patients who fail to reach glycemic targets. For example, a single dose of long-acting insulin at bedtime is often effective in combination with metformin. As endogenous insulin production falls further, multiple injections of long-acting and short-acting insulin regimens are necessary to control postprandial glucose excursions. These insulin regimens are identical to the long-acting and short-acting combination regimens discussed above for type 1 DM, although usually at higher doses given insulin

resistance. Weight gain and hypoglycemia are the major adverse effects of insulin therapy. The daily insulin dose required can become quite large (1–2 units/kg per day) as endogenous insulin production falls and insulin resistance persists. Individuals who require >1 unit/kg per day of long-acting insulin should be considered for combination therapy with metformin or a thiazolidinedione. The addition of metformin or a thiazolidinedione can reduce insulin requirements in some individuals with type 2 DM, while maintaining or even improving glycemic control. Insulin plus a thiazolidinedione promotes weight gain and is associated with peripheral edema. Addition of a thiazolidinedione to a patient's insulin regimen may necessitate a reduction in the insulin dose to avoid hypoglycemia. Patients requiring large doses of insulin (>200 units/day) can be treated with a more concentrated form of insulin.

SURGICAL THERAPIES

Whole pancreas transplantation can normalize glucose control in type 1 DM and when performed simultaneously with or after kidney transplantation can prolong the life of the kidney transplant by offering protection against recurrent diabetic nephropathy. Pancreatic islet transplantation is available as a less invasive form of beta cell replacement therapy for type 1 DM, but remains investigational in the United States. Due to the risks associated with chronic immunosuppression, whole pancreas and pancreatic islet transplantation may be considered for patients with severe metabolic instability or already requiring immunosuppression in support of a kidney or other organ transplant. Patients with chronic pancreatitis and preserved islet function who require pancreatectomy for pain relief may benefit from autologous islet transplantation as this may prevent or ameliorate postsurgical DM.

Metabolic (also referred to as bariatric) surgery for obese individuals with type 2 DM has shown considerable promise, sometimes with dramatic resolution of the diabetes or major reductions in the needed dose of glucose-lowering therapies (**Chap. 395**). Several large, non-randomized clinical trials have demonstrated a much greater efficacy of metabolic surgery compared to medical management in the treatment of type 2 DM and may be considered in individuals with T2DM and a BMI >30 kg/m^2. The ADA clinical guidelines state that metabolic surgery should be considered in individuals with type 2 DM and a body mass index >30 kg/m^2 if hyperglycemia is inadequately controlled despite optimal medical therapy.

EMERGING THERAPIES

Many individuals with long-standing type 1 DM still produce very small amounts of insulin or have insulin-positive cells within the pancreas. This suggests that beta cells may slowly regenerate but are quickly destroyed by the autoimmune process. Particularly early in the disease course, efforts to suppress the autoimmune process and allow for beta cell regeneration are being tested at the time of diagnosis of type 1 DM, and for prevention in autoantibody-positive individuals at Stages 1 and 2 of type 1 DM (417-6). Closed-loop insulin infusion devices that infuse the appropriate amount of insulin in response to changing glucose levels are progressing rapidly. Bi-hormonal infusion devices that deliver both insulin and glucagon are under development. New therapies under evaluation or development for type 2 DM include activators of glucokinase, inhibitors of 11 β-hydroxysteroid dehydrogenase-1, GPR40 agonists, and agents to reduce inflammation.

Because whole pancreas and pancreatic islet transplantation are both limited by organ availability from deceased donors, stem cell-derived islet cells and xenogeneic sources of islets may eventually allow for a limitless supply of insulin-producing cells for transplantation.

ADVERSE EFFECTS OF THERAPY FOR DM

As with any therapy, the benefits of efforts directed toward glycemic control must be balanced against the risks of treatment (Table 397-5). Side effects of intensive treatment include an increased frequency of serious hypoglycemia, weight gain, increased economic costs, and greater demands on the patient. In the DCCT, quality of life was very similar in the intensive and standard therapy groups. The most serious complication of therapy for DM is hypoglycemia, and its treatment with oral glucose or glucagon injection is discussed in **Chap. 399**. Severe, recurrent hypoglycemia warrants examination of treatment regimen and

glycemic goal for the individual patient. Weight gain occurs with most (insulin, insulin secretagogues, thiazolidinediones) but not all (metformin, α-glucosidase inhibitors, GLP-1 receptor agonists, DPP-IV inhibitors) therapies. The weight gain is partially due to the anabolic effects of insulin and the reduction in glucosuria. As a result of concerns about CV safety of diabetes therapies, the FDA requires information about the cardiovascular safety profile as part of its evaluation of new medications for type 2 DM.

ACUTE DISORDERS RELATED TO SEVERE HYPERGLYCEMIA

Individuals with type 1 or type 2 DM and severe hyperglycemia (>13.9 mmol/L [250 mg/dL]) should be assessed for clinical stability, including mentation and hydration. Depending on the patient and the rapidity and duration of the severe hyperglycemia, an individual may require more intense and rapid therapy to lower the blood glucose. However, many patients with poorly controlled diabetes and hyperglycemia have few symptoms. The physician should assess if the patient is stable or if DKA or a hyperglycemic hyperosmolar state (HHS) should be considered. Ketones, an indicator of DKA, should be measured in individuals with type 1 DM when the plasma glucose is >13.9 mmol/L (250 mg/dL), during a concurrent illness, or with symptoms such as nausea, vomiting, or abdominal pain. Blood measurement of β-hydroxybutyrate is preferred over urine testing with nitroprusside-based assays that measure only acetoacetate and acetone.

DKA and HHS are acute, severe disorders directly related to diabetes. DKA was formerly considered a hallmark of type 1 DM, but also occurs in individuals with type 2 DM who can sometimes subsequently be treated with oral glucose-lowering agents (usually obese individuals of Hispanic or African-American descent). HHS is primarily seen in individuals with type 2 DM. Both disorders are associated with absolute or relative insulin deficiency, volume depletion, and acid-base abnormalities. DKA and HHS exist along a continuum of hyperglycemia, with or without ketosis. The metabolic similarities and differences in DKA and HHS are highlighted in **Table 397-6**. Both disorders are associated with potentially serious complications if not promptly diagnosed and carefully treated.

TABLE 397-6

Laboratory Values in Diabetic Ketoacidosis (DKA) and Hyperglycemic Hyperosmolar State (HHS) (Representative Ranges at Presentation)

	DKA	HHS
Glucose,[a] mmol/L (mg/dL)	13.9–33.3 (250–600)	33.3–66.6 (600–1200)
Sodium, meq/L	125–135	135–145
Potassium[a,b]	Normal to ↑	Normal
Magnesium[a]	Normal	Normal
Chloride[a]	Normal	Normal
Phosphate[a,b]	Normal	Normal
Creatinine	Slightly ↑	Moderately ↑
Osmolality (mosm/mL)	300–320	330–380
Plasma ketones[a]	++++	+/–
Serum bicarbonate,[a] meq/L	<15	Normal to slightly ↓
Arterial pH	6.8–7.3	>7.3
Arterial P_{CO_2},[a] mmHg	20–30	Normal
Anion gap[a] $(Na - [Cl + HCO_3])$	↑	Normal to slightly ↑

[a]Large changes occur during treatment of DKA. [b]Although plasma levels may be normal or high at presentation, total-body stores are usually depleted.

DIABETIC KETOACIDOSIS

Clinical Features

The symptoms and physical signs of DKA are listed in Table 397-7 and usually develop over 24 h. DKA may be the initial symptom complex that leads to a diagnosis of type 1 DM, but more frequently, it occurs in individuals with established diabetes. Nausea and vomiting are often prominent, and their presence in an individual with diabetes warrants laboratory evaluation for DKA. Abdominal pain may be severe and can resemble acute pancreatitis or ruptured viscus. Hyperglycemia leads to glucosuria, volume depletion, and tachycardia. Hypotension can occur because of volume depletion in combination with peripheral vasodilatation. Kussmaul respirations and a fruity odor on the patient's breath (secondary to metabolic acidosis and increased acetone) are classic signs of the disorder. Lethargy and central nervous system depression may evolve into coma with severe DKA but should also prompt evaluation for other reasons for altered mental status (e.g., infection, hypoxemia). Cerebral edema,

an extremely serious complication of DKA, is seen most frequently in children. Signs of infection, which may precipitate DKA, should be sought on physical examination, even in the absence of fever. Tissue ischemia (heart, brain) can also be a precipitating factor. Omission of insulin because of an eating disorder, mental health disorders, or an unstable psychosocial environment may sometimes be a factor precipitating DKA.

TABLE 397-7

Manifestations of Diabetic Ketoacidosis

Symptoms	**Physical Findings**
Nausea/vomiting	Tachycardia
Thirst/polyuria	Dehydration/hypotension
Abdominal pain	Tachypnea/Kussmaul respirations/respiratory distress
Shortness of breath	Abdominal tenderness (may resemble acute pancreatitis or surgical abdomen)
Precipitating events	Lethargy/obtundation/cerebral edema/possibly coma
Inadequate insulin administration	
Infection (pneumonia/UTI/gastroenteritis/sepsis)	
Infarction (cerebral, coronary, mesenteric, peripheral)	
Drugs (cocaine)	
Pregnancy	

Abbreviation: UTI, urinary tract infection.

Pathophysiology

DKA results from relative or absolute insulin deficiency combined with counterregulatory hormone excess (glucagon, catecholamines, cortisol, and growth hormone). Both insulin deficiency and glucagon excess, in particular, are necessary for DKA to develop. The decreased ratio of insulin to glucagon promotes gluconeogenesis, glycogenolysis, and ketone body formation in the liver, as well as increases in substrate delivery from fat and muscle (free fatty acids, amino acids) to the liver. Markers of inflammation (cytokines, C-reactive protein) are elevated in both DKA and HHS.

The combination of insulin deficiency and hyperglycemia reduces the hepatic level of fructose-2,6-bisphosphate, which alters the activity of phosphofructokinase and fructose-1,6-bisphosphatase. Glucagon excess decreases the activity of pyruvate kinase, whereas insulin deficiency increases the activity of phosphoenolpyruvate carboxykinase. These changes shift the handling of pyruvate toward glucose synthesis and away from glycolysis. The increased levels of glucagon and catecholamines in the face of low insulin levels promote glycogenolysis. Insulin deficiency also reduces levels of the GLUT4 glucose transporter, which impairs glucose uptake into skeletal muscle and fat and reduces intracellular glucose metabolism.

Ketosis results from a marked increase in free fatty acid release from adipocytes, with a resulting shift toward ketone body synthesis in the liver. Reduced insulin levels, in combination with elevations in catecholamines and growth hormone, increase lipolysis and the release of free fatty acids. Normally, these free fatty acids are converted to very low-density lipoprotein (VLDL) in the liver. However, in DKA, hyperglucagonemia alters hepatic metabolism to favor ketone body formation, through activation of the enzyme carnitine palmitoyltransferase I. This enzyme is crucial for regulating fatty acid transport into the mitochondria, where beta oxidation and conversion to ketone bodies occur. At physiologic pH, ketone bodies exist as ketoacids, which are neutralized by bicarbonate. As bicarbonate stores are depleted, metabolic acidosis ensues. Increased lactic acid production also contributes to the acidosis. The increased free fatty acids increase VLDL-triglyceride production. VLDL clearance is also reduced because the activity of insulin-sensitive lipoprotein lipase in muscle and fat is decreased.

DKA is often precipitated by increased insulin requirements, as occurs during a concurrent illness (Table 397-7). Failure to augment insulin therapy often compounds the problem. Complete omission or inadequate administration of insulin by the patient or health care team (in a hospitalized patient with type 1 DM) may precipitate DKA. Patients using insulin-infusion devices with short-acting insulin may develop DKA, because even a brief interruption in insulin delivery (e.g., mechanical malfunction) quickly leads to insulin deficiency.

Laboratory Abnormalities and Diagnosis

The timely diagnosis of DKA is crucial and allows for prompt initiation of therapy. DKA is characterized by hyperglycemia (serum glucose > 13.9 mmol/L [250 mg/dL], ketosis, and metabolic acidosis (serum bicarbonate <15 mmol/L with increased anion gap) along with a number of secondary metabolic derangements (Table 397-6). Occasionally, the serum glucose is only minimally elevated, and may even be normal (euglycemic DKA). This has been noted in individuals treated with SGLT2 inhibitors. Arterial pH ranges between 6.8 and 7.3, depending on the severity of the acidosis. Despite a total-body potassium deficit, the serum potassium at presentation may be mildly elevated, secondary to the acidosis and volume depletion. Total-body stores of sodium, chloride, phosphorus, and magnesium are also reduced in DKA but are not accurately reflected by their levels in the serum because of hypovolemia and hyperglycemia. Elevated blood urea nitrogen (BUN) and serum creatinine levels reflect intravascular volume depletion. Leukocytosis, hypertriglyceridemia, and hyperlipoproteinemia are commonly found as well. Hyperamylasemia may suggest a diagnosis of pancreatitis, especially when accompanied by abdominal pain. However, in DKA the amylase is usually of salivary origin and thus is not diagnostic of pancreatitis. Serum lipase should be obtained if pancreatitis is suspected.

The measured serum sodium is reduced as a consequence of the hyperglycemia (1.6-mmol/L [1.6-meq] reduction in serum sodium for each 5.6-mmol/L [100-mg/dL] rise in the serum glucose). A normal serum sodium in the setting of DKA indicates a more profound water deficit. In “conventional” units, the calculated serum osmolality (2 × [serum sodium + serum potassium] + plasma glucose [mg/dL]/18 + BUN/2.8) is mildly to moderately elevated, although to a lesser degree than that found in HHS (see below).

In DKA, the ketone body, β-hydroxybutyrate, is synthesized at a threefold greater rate than acetoacetate; however, acetoacetate is preferentially detected by a commonly used ketosis detection reagent (nitroprusside). Serum ketones are present at significant levels (usually positive at serum dilution of ≥1:8). The nitroprusside tablet, or stick, is often used to detect urine ketones; certain medications such as captopril or penicillamine may cause false-positive reactions. Serum or plasma assays for β-hydroxybutyrate are preferred because they more accurately reflect the true ketone body level.

The metabolic derangements of DKA exist along a spectrum, beginning with mild acidosis with moderate hyperglycemia evolving into more severe findings. The degree of acidosis and hyperglycemia do not necessarily correlate closely because a variety of factors determine the level of hyperglycemia (oral intake, urinary glucose loss). Ketonemia is a consistent finding in DKA and distinguishes it from simple hyperglycemia. The differential diagnosis of DKA includes starvation ketosis, alcoholic ketoacidosis (bicarbonate usually >15 meq/L), and other forms of increased anion-gap acidosis **(Chap. 51)**.

TREATMENT

TREATMENT

Diabetic Ketoacidosis

The management of DKA is outlined in **Table 397-8**. After initiating IV fluid replacement and insulin therapy, the agent or event that precipitated the episode of DKA should be sought and aggressively treated. If the patient is vomiting or has altered mental status, a nasogastric tube should be inserted to prevent aspiration of gastric contents. Central to successful treatment of DKA is careful monitoring and frequent reassessment to ensure that the patient and the metabolic derangements are improving. A

comprehensive flow sheet should record chronologic changes in vital signs, fluid intake and output, and laboratory values as a function of insulin administered.

After the initial bolus of normal saline, replacement of the sodium and free water deficit is carried out over the next 24 h (fluid deficit is often 3–5 L). When hemodynamic stability and adequate urine output are achieved, IV fluids should be switched to 0.45% saline depending on the calculated volume deficit. The change to 0.45% saline helps to reduce the trend toward hyperchloremia later in the course of DKA. Alternatively, initial use of lactated Ringer's IV solution may reduce the hyperchloremia that commonly occurs with normal saline.

A bolus of IV (0.1 units/kg) short-acting regular insulin is usually administered immediately (Table 397-8), and subsequent treatment should provide continuous and adequate levels of circulating insulin. IV administration is usually preferred (0.1 units/kg of regular insulin per hour) because it ensures rapid distribution and allows adjustment of the infusion rate as the patient responds to therapy. DKA can also be treated with SC short-acting insulin analogues. IV regular insulin should be continued until the acidosis resolves and the patient is metabolically stable. As the acidosis and insulin resistance associated with DKA resolve, the insulin infusion rate can be decreased (to 0.02–0.1 units/kg per hour). Long-acting insulin, in combination with SC short-acting insulin, should be administered as soon as the patient resumes eating, because this facilitates transition to an outpatient insulin regimen and reduces length of hospital stay. It is crucial to continue the insulin infusion until adequate insulin levels are achieved by administering long-acting insulin by the SC route. Even relatively brief periods of inadequate insulin administration in this transition phase may result in DKA relapse.

Hyperglycemia usually improves at a rate of 4.2–5.6 mmol/L (50–100 mg/dL) per hour as a result of insulin-mediated glucose disposal, reduced hepatic glucose release, and rehydration. The latter reduces catecholamines, increases urinary glucose loss, and expands the intravascular volume. The decline in the plasma glucose within the first 1–2 h may be more rapid and is mostly related to volume expansion. When the plasma glucose reaches 11.1–13.9 mmol/L (200–250 mg/dL), glucose should be added to the 0.45% saline infusion to maintain the plasma glucose in the 8.3–11.1 mmol/L (150–200 mg/dL) range, and the insulin infusion should be continued at a lower rate to inhibit ketogenesis. More rapid correction of the serum glucose can precipitate the development of cerebral edema. Ketoacidosis begins to resolve as insulin reduces lipolysis, increases peripheral ketone body use, suppresses hepatic ketone body formation, and promotes bicarbonate regeneration. However, the acidosis and ketosis resolve more slowly than hyperglycemia. As ketoacidosis improves, β-hydroxybutyrate is converted to acetoacetate. Ketone body levels may appear to increase if measured by laboratory assays that use the nitroprusside reaction, which only detects acetoacetate and acetone. The improvement in acidosis and anion gap, a result of bicarbonate regeneration and decline in ketone bodies, is reflected by a rise in the serum bicarbonate level and the arterial pH. Depending on the rise of serum chloride, the anion gap (but not bicarbonate) will normalize. A hyperchloremic acidosis (serum bicarbonate of 15–18 mmol/L [15–18 meq/L]) often follows successful treatment and gradually resolves as the kidneys regenerate bicarbonate and excrete chloride.

Potassium stores are depleted in DKA (estimated deficit 3–5 mmol/kg [3–5 meq/kg]). During treatment with insulin and fluids, various factors contribute to the development of hypokalemia. These include insulin-mediated potassium transport into cells, resolution of the acidosis (which also promotes potassium entry into cells), and urinary loss of potassium salts of organic acids. Thus, potassium repletion should commence as soon as adequate urine output and a normal serum potassium are documented. If the initial serum potassium level is elevated, then potassium repletion should be delayed until the potassium falls into the normal range. Inclusion of 20–40 meq of potassium in each liter of IV fluid is reasonable, but additional potassium supplements may also be required. To reduce the amount of chloride administered, potassium phosphate or acetate can be substituted for the chloride salt. The goal is to maintain the serum potassium at >3.5 mmol/L (3.5 meq/L).

Despite a bicarbonate deficit, bicarbonate replacement is not usually necessary. In fact, theoretical arguments suggest that bicarbonate administration and rapid reversal of acidosis may impair cardiac function, reduce tissue oxygenation, and promote hypokalemia. The results of most clinical trials do not support the routine use of bicarbonate replacement, and one study in children found that bicarbonate use was associated with an increased risk of cerebral edema. However, in the presence of

severe acidosis (arterial pH <7.0), the ADA advises bicarbonate (50 mmol [meq/L] of sodium bicarbonate in 200 mL of sterile water with 10 meq/L KCl per hour for 2 h until the pH is >7.0). Hypophosphatemia may result from increased glucose usage, but randomized clinical trials have not demonstrated that phosphate replacement is beneficial in DKA. If the serum phosphate is <0.32 mmol/L (1 mg/dL), then phosphate supplement should be considered and the serum calcium monitored. Hypomagnesemia may develop during DKA therapy and may also require supplementation.

With appropriate therapy, the mortality rate of DKA is low (<1%) and is related more to the underlying or precipitating event, such as infection or myocardial infarction. Venous thrombosis, upper GI bleeding, and acute respiratory distress syndrome occasionally complicate DKA. The major nonmetabolic complication of DKA therapy is cerebral edema, which most often develops in children as DKA is resolving. The etiology of and optimal therapy for cerebral edema are not well established, but overreplacement of free water and rapid normalization of serum glucose should be avoided.

Following treatment, the physician and patient should review the sequence of events that led to DKA to prevent future recurrences. Foremost is patient education about the symptoms of DKA, its precipitating factors, and the management of diabetes during a concurrent illness. During illness or when oral intake is compromised, patients should (1) frequently measure the capillary blood glucose; (2) measure urinary ketones when the serum glucose is >13.7 mmol/L (250 mg/dL); (3) drink fluids to maintain hydration; (4) continue or increase insulin; and (5) seek medical attention if dehydration, persistent vomiting, or uncontrolled hyperglycemia develop. Using these strategies, early DKA can be prevented or detected and treated appropriately on an outpatient basis.

TABLE 397-8

Management of Diabetic Ketoacidosis

1. Confirm diagnosis (↑ plasma glucose, positive serum ketones, metabolic acidosis).
2. Admit to hospital; intensive care setting may be necessary for frequent monitoring or if pH <7.00 or unconscious.
3. Assess:

   Serum electrolytes (K^+, Na^+, Mg^{2+}, Cl^-, bicarbonate, phosphate)

   Acid-base status—pH, HCO_3^-, Pco_2, β-hydroxybutyrate

   Renal function (creatinine, urine output)
4. Replace fluids: 2–3 L of 0.9% saline over first 1–3 h (10–20 mL/kg per hour); subsequently, 0.45% saline at 250–500 mL/h; change to 5% glucose and 0.45% saline at 150–250 mL/h when plasma glucose reaches 250 mg/dL (13.9 mmol/L).
5. Administer short-acting regular insulin: IV (0.1 units/kg), then 0.1 units/kg per hour by continuous IV infusion; increase two- to threefold if no response by 2–4 h. If the initial serum potassium is <3.3 mmol/L (3.3 meq/L), do not administer insulin until the potassium is corrected.
6. Assess patient: What precipitated the episode (noncompliance, infection, trauma, pregnancy, infarction, cocaine)? Initiate appropriate workup for precipitating event (cultures, CXR, ECG).
7. Measure capillary glucose every 1–2 h; measure electrolytes (especially K^+, bicarbonate, phosphate) and anion gap every 4 h for first 24 h.
8. Monitor blood pressure, pulse, respirations, mental status, fluid intake and output every 1–4 h.
9. Replace K^+: 10 meq/h when plasma K^+ <5.0–5.2 meq/L (or 20–30 meq/L of infusion fluid), ECG normal, urine flow and normal creatinine documented; administer 40–80 meq/h when plasma K^+ <3.5 meq/L or if bicarbonate is given. If initial serum potassium is >5.2 mmol/L (5.2 meq/L), do not supplement K^+ until the potassium is corrected.
10. See text about bicarbonate or phosphate supplementation.
11. Continue above until patient is stable, glucose goal is 8.3–11.1 mmol/L (150–200 mg/dL), and acidosis is resolved. Insulin infusion may be decreased to 0.02–0.1 units/kg per hour.
12. Administer long-acting insulin as soon as patient is eating. Allow for a 2–4 hour overlap in insulin infusion and SC long-acting insulin injection.

Abbreviations: CXR, chest x-ray; ECG, electrocardiogram.

Source: Adapted from M Sperling, in *Therapy for Diabetes Mellitus and Related Disorders*, American Diabetes Association, Alexandria, VA, 1998; and EA Nyenwe and AE Kitabchi: Metabolism 65:507, 2016.

HYPERGLYCEMIC HYPEROSMOLAR STATE

Clinical Features

The prototypical patient with HHS is an elderly individual with type 2 DM, with a several-week history of polyuria, weight loss, and diminished oral intake that culminates in mental confusion, lethargy, or coma. The physical examination reflects profound dehydration and hyperosmolality and reveals hypotension, tachycardia, and altered mental status. Notably absent are symptoms of nausea, vomiting, and abdominal pain and the Kussmaul respirations characteristic of DKA. HHS is often precipitated by a serious, concurrent illness such as myocardial infarction or stroke. Sepsis, pneumonia, and other serious infections are frequent precipitants and should be sought. In addition, a debilitating condition (prior stroke or dementia) or social situation that compromises water intake usually contributes to the development of the disorder.

Pathophysiology

Relative insulin deficiency and inadequate fluid intake are the underlying causes of HHS. Insulin deficiency increases hepatic glucose production (through glycogenolysis and gluconeogenesis) and impairs glucose utilization in skeletal muscle (see above discussion of DKA). Hyperglycemia induces an osmotic diuresis that leads to intravascular volume depletion, which is exacerbated by inadequate fluid replacement. The absence of ketosis in HHS is not understood. Presumably, the insulin deficiency is only relative and less severe than in DKA. Lower levels of counterregulatory hormones and free fatty acids have been found in HHS than in DKA in some studies. It is also possible that the liver is less capable of ketone body synthesis or that the insulin/glucagon ratio does not favor ketogenesis.

Laboratory Abnormalities and Diagnosis

The laboratory features in HHS are summarized in Table 397-6. Most notable are the marked hyperglycemia (plasma glucose may be >55.5 mmol/L [1000 mg/dL]), hyperosmolality (>350 mosmol/L), and prerenal azotemia. The measured serum sodium may be normal or slightly low despite the marked hyperglycemia. The corrected serum sodium is usually increased (add 1.6 meq to measured sodium for each 5.6-mmol/L [100-mg/dL] rise in the serum glucose). In contrast to DKA, acidosis and ketonemia are absent or mild. A small anion-gap metabolic acidosis may be present secondary to increased lactic acid. Moderate ketonuria, if present, is secondary to starvation.

TREATMENT

TREATMENT

Hyperglycemic Hyperosmolar State

Volume depletion and hyperglycemia are prominent features of both HHS and DKA. Consequently, therapy of these disorders shares several elements (Table 397-8). In both disorders, careful monitoring of the patient's fluid status, laboratory values, and insulin infusion rate is crucial. Underlying or precipitating problems should be aggressively sought and treated. In HHS, fluid losses and dehydration are usually more pronounced than in DKA due to the longer duration of the illness. The patient with HHS is usually older, more likely to have mental status changes, and more likely to have a life-threatening precipitating event with accompanying comorbidities. Even with proper treatment, HHS has a substantially higher mortality rate than DKA (up to 15% in some clinical series).

Fluid replacement should initially stabilize the hemodynamic status of the patient (1–3 L of 0.9% normal saline over the first 2–3 h). Because the fluid deficit in HHS is accumulated over a period of days to weeks, the rapidity of reversal of the hyperosmolar state must balance the need for free water repletion with the risk that too rapid a reversal may worsen neurologic function. If the serum sodium is >150 mmol/L (150 meq/L), 0.45% saline should be used. After hemodynamic stability is achieved, the IV fluid administration is directed at reversing the free water deficit using hypotonic fluids (0.45% saline initially, then 5% dextrose in water [D_5W]). The calculated free water deficit (which averages 9–10 L) should be reversed over the next 1–2 days (infusion rates of 200–300 mL/h of hypotonic solution). Potassium repletion is usually necessary and should be dictated by repeated measurements of the serum potassium. In patients taking diuretics, the potassium deficit can be quite large and may be accompanied by magnesium deficiency. Hypophosphatemia may occur during therapy and can be improved by using KPO_4 and beginning nutrition.

As in DKA, rehydration and volume expansion lower the plasma glucose initially, but insulin is also required. A reasonable regimen for HHS begins with an IV insulin bolus of 0.1 unit/kg followed by IV insulin at a constant infusion rate of 0.1 unit/kg per hour. If the serum glucose does not fall, increase the insulin infusion rate by twofold. As in DKA, glucose should be added to IV fluid when the plasma glucose falls to 11.1–13.9 mmol/L (200–250 mg/dL), and the insulin infusion rate should be decreased to 0.02–0.1 unit/kg per hour. The insulin infusion should be continued until the patient has resumed eating and can be transferred

to a SC insulin regimen. The patient should be discharged from the hospital on insulin, although some patients can later switch to oral glucose-lowering agents.

MANAGEMENT OF DIABETES IN A HOSPITALIZED PATIENT

Virtually all medical and surgical subspecialties are involved in the care of hospitalized patients with diabetes. Hyperglycemia, whether in a patient with known diabetes or in someone without known diabetes, appears to be a predictor of poor outcome in hospitalized patients. General anesthesia, surgery, infection, or concurrent illness raises the levels of counterregulatory hormones (cortisol, growth hormone, catecholamines, and glucagon) and cytokines that may lead to transient insulin resistance and hyperglycemia. These factors increase insulin requirements by increasing glucose production and impairing glucose utilization and thus may worsen glycemic control. The concurrent illness or surgical procedure may lead to variable insulin absorption and also prevent the patient with DM from eating normally and, thus, may promote hypoglycemia. Glycemic control should be assessed on admission using the HbA_{1c}. Electrolytes, renal function, and intravascular volume status should be assessed as well. The high prevalence of CVD in individuals with DM (especially in type 2 DM) may necessitate preoperative cardiovascular evaluation **(Chap. 398)**.

The goals of diabetes management during hospitalization are near-normoglycemia, avoidance of hypoglycemia, and transition back to the outpatient diabetes treatment regimen. Upon hospital admission, frequent glycemic monitoring should begin, as should planning for diabetes management after discharge. Glycemic control appears to improve the clinical outcomes in a variety of settings, but optimal glycemic goals for the hospitalized patient are incompletely defined. In a number of cross-sectional studies of patients with diabetes, a greater degree of hyperglycemia was associated with worse cardiac, neurologic, and infectious outcomes. In some studies, patients who do not have preexisting diabetes but who develop modest blood glucose elevations during their hospitalization appear to benefit from achieving near-normoglycemia using insulin treatment. However, a large randomized clinical trial (Normoglycemia in Intensive Care Evaluation Survival Using Glucose Algorithm Regulation [NICE-SUGAR]) of individuals in the intensive care unit (ICU) (most of whom were receiving mechanical ventilation) found an increased mortality rate and a greater number of episodes of severe hypoglycemia with very strict glycemic control (target blood glucose of 4.5–6 mmol/L or 81–108 mg/dL) compared to individuals with a more moderate glycemic goal (target blood glucose of <10 mmol/L or 180 mg/dL). Currently, most data suggest that very strict blood glucose control in acutely ill patients likely worsens outcomes and increases the frequency of hypoglycemia. The ADA suggests the following glycemic goals for hospitalized patients: (1) in critically or non-critically ill patients: glucose of 7.8–10.0 mmol/L or 140–180 mg/dL; (2) in selected patients: glucose of 6.1–7.8 mmol/L or 110–140 mg/dL with avoidance of hypoglycemia.

Critical aspects for optimal diabetes care in the hospital include the following. (1) A hospital-wide system approach to treatment of hyperglycemia and prevention of hypoglycemia is needed. Inpatient diabetes management teams consisting of nurse practitioners and physicians are increasingly common. (2) Diabetes treatment plans should focus on the transition from the ICU and the transition from the inpatient to outpatient setting. (3) Adjustment of the discharge treatment regimen of patients whose diabetes was poorly controlled on admission (as reflected by the HbA_{1c}) is important.

The physician caring for an individual with diabetes in the perioperative period, during times of infection or serious physical illness, or simply when the patient is fasting for a diagnostic procedure must monitor the plasma glucose vigilantly, adjust the diabetes treatment regimen, and provide glucose infusion as needed. Hypoglycemia is frequent in hospitalized patients, and many of these episodes are avoidable. Hospital systems should have a diabetes management protocol to avoid inpatient hypoglycemia. Measures to reduce or prevent hypoglycemia include frequent glucose monitoring and anticipating potential modifications of insulin/glucose administration because of changes in the clinical situation or treatment (e.g., tapering of glucocorticoids) or interruption of enteral or parenteral infusions or PO intake.

Depending on the severity of the patient's illness and the hospital setting, the physician can use either an insulin infusion or SC insulin. Insulin infusions are preferred in the ICU or in a clinically unstable setting because the half-life of the infused insulin is

quite short (minutes). The absorption of SC insulin may be variable in such situations. Insulin infusions can also effectively control plasma glucose in the perioperative period and when the patient is unable to take anything by mouth. Regular insulin is used rather than insulin analogues for IV insulin infusion because it is less expensive and equally effective. The physician must consider carefully the clinical setting in which an insulin infusion will be used, including whether adequate ancillary personnel are available to monitor the plasma glucose frequently and whether they can adjust the insulin infusion rate to maintain the plasma glucose within the optimal range. Insulin-infusion algorithms should integrate the insulin sensitivity of the patient, frequent blood glucose monitoring, and the trend of changes in the blood glucose to determine the insulin-infusion rate. Insulin-infusion algorithms jointly developed and implemented by nursing and physician staff are advised. Because of the short half-life of IV regular insulin, it is necessary to administer long-acting insulin prior to discontinuation of the insulin infusion (2–4 h before the infusion is stopped) to avoid a period of insulin deficiency.

In patients who are not critically ill or not in the ICU, basal or "scheduled" insulin is provided by SC, long-acting insulin supplemented by prandial and/or "corrective" insulin using a short-acting insulin (insulin analogues preferred). The use of "sliding scale," short-acting insulin alone, where no insulin is given unless the blood glucose is elevated, is inadequate for inpatient glucose management and should not be used. The short-acting, preprandial insulin dose should include coverage for food consumption (based on anticipated carbohydrate intake) plus a corrective or supplemental insulin based on the patient's insulin sensitivity and the blood glucose. For example, if the patient is thin (and likely insulin-sensitive), a corrective insulin supplement might be 1 unit for each 2.7 mmol/L (50 mg/dL) over the glucose target. If the patient is obese and insulin-resistant, then the insulin supplement might be 2 units for each 2.7 mmol/L (50 mg/dL) over the glucose target. It is critical to individualize the regimen and adjust the basal or "scheduled" insulin dose frequently, based on the corrective insulin required. A consistent carbohydrate-controlled diabetes meal plan for hospitalized patients provides a predictable amount of carbohydrate for a particular meal each day (but not necessarily the same amount for breakfast, lunch, and supper) and avoids concentrated sweets. Individuals with type 1 DM who are undergoing general anesthesia and surgery or who are seriously ill should receive continuous insulin, either through an IV insulin infusion, their insulin infusion device, or by SC administration of a reduced dose of long-acting insulin. Short-acting insulin alone is insufficient. Prolongation of a surgical procedure or delay in the recovery room is not uncommon and may result in periods of insulin deficiency leading to DKA. Insulin infusion is the preferred method for managing patients with type 1 DM over a prolonged (several hours) perioperative period or when serious concurrent illness is present (0.5–1.0 units/h of regular insulin). If the diagnostic or surgical procedure is brief (<2 h), a reduced dose of SC insulin may suffice (20–50% basal reduction, with short-acting bolus insulin withheld or reduced). This approach prevents interruption of insulin infusion device therapy, or for MDI, facilitates the transition back to long-acting insulin after the procedure. The blood glucose should be monitored frequently during the illness or in the perioperative period.

Individuals with type 2 DM can be managed with either an insulin infusion or SC long-acting insulin (20–50% reduction depending on clinical setting) plus preprandial, short-acting insulin. Oral glucose-lowering agents should be discontinued upon admission and are not useful in regulating the plasma glucose in clinical situations where the insulin requirements and glucose intake are changing rapidly. Moreover, these oral agents may be dangerous if the patient is fasting (e.g., hypoglycemia with sulfonylureas) or at risk for declining kidney function due to, for example, radiographic contrast media or unstable CHF (lactic acidosis with metformin). Once clinically stable, oral glucose-lowering agents may be resumed in anticipation of discharge.

SPECIAL CONSIDERATIONS IN DM

TOTAL PARENTERAL NUTRITION (TPN)

TPN greatly increases insulin requirements **(See also Chap. 328)**. In addition, individuals not previously known to have DM may become hyperglycemic during TPN and require insulin treatment. IV insulin infusion is the preferred treatment for hyperglycemia, and rapid titration to the required insulin dose is done most efficiently using a separate insulin infusion. After the total insulin dose has been determined, a proportion of this insulin may be added directly to the TPN solution to cover the nutritional requirements for insulin, and adjusted based on the need for modified dosing of short-acting insulin. Total enteral

nutrition (TEN) also increases insulin requirements and may lead to or worsen hyperglycemia. Hyperglycemia may be limited by using high protein formulations, but often requires insulin treatment. Short- or intermediate (i.e., NPH)-acting insulins should be used to cover bolus or continuous enteral feeding to minimize the risk for hypoglycemia should the TEN be interrupted or held. Patients with insulin deficiency (type 1 DM and pancreatogenic DM) should also receive long-acting insulin (0.1–0.2 units/kg per day) to cover basal insulin requirements should the TPN or TEN be interrupted or cycled.

GLUCOCORTICOIDS

Glucocorticoids increase insulin resistance, decrease glucose utilization, increase hepatic glucose production, and impair insulin secretion. These changes lead to a worsening of glycemic control in individuals with DM and may precipitate hyperglycemia in other individuals. If new-onset hyperglycemia remains during chronic treatment with supraphysiologic doses of glucocorticoid (>5 mg of prednisone or equivalent), the DM may be called “steroid-induced diabetes.” The effects of glucocorticoids on glucose homeostasis are dose-related, usually reversible, and most pronounced in the postprandial period. If the FPG is near the normal range, oral diabetes agents (e.g., sulfonylureas, metformin) may be sufficient to reduce hyperglycemia. If the FPG is >11.1 mmol/L (200 mg/dL), oral agents are usually not efficacious, and insulin therapy is required. Short-acting insulin may be sufficient alone or together with long-acting insulin in order to control postprandial glucose excursions.

DIABETES MANAGEMENT IN OLDER ADULTS

Diabetes is very common in older adults, being present in ~25% of individuals over the age of 65. Increasingly, individuals with many years of type 1 DM are part of the patient population. As discussed above, individualized therapeutic goals and modalities in older adults should consider biologic age, other comorbidities and risk factors (hypertension, CV disease, etc.), neurocognitive and physical functional status, living arrangements, social support, and other medications. For example, the HbA1c goal for a highly functional 80-year-old should be different than that for an individual with diabetes in long-term care (skilled nursing facilities). In the former, the HbA1c goal (<7.5%) and selected therapies may be similar to younger individuals whereas in an individual with complex/poor health or cognitive impairment, an HbA1c goal of <8.5% would be reasonable. Critical to diabetes management in all older individuals is the avoidance of hypoglycemia, which can worsen underlying cognitive impairment or CV disease. Thus, medications that can cause hypoglycemia (insulin secretagogues, insulin) should be used carefully. In choosing medications for diabetes, the adverse effects (Table 397-5) should be considered (especially heart failure, renal insufficiency, etc.). Hypertension and dyslipidemia should be treated in elderly individuals with diabetes since there is clear benefit of blood pressure control with the benefit for lipid-lowering medications being less clearly demonstrated.

REPRODUCTIVE ISSUES

Reproductive capacity in either men or women with DM appears to be normal. Menstrual cycles may be associated with alterations in glycemic control in women with DM. Pregnancy is associated with marked insulin resistance; the increased insulin requirements often precipitate DM and lead to the diagnosis of gestational diabetes mellitus (GDM). Glucose, which at high levels is a teratogen to the developing fetus, readily crosses the placenta, but insulin does not. Thus, hyperglycemia from the maternal circulation may stimulate insulin secretion in the fetus. The anabolic and growth effects of insulin may result in macrosomia. GDM complicates ~7% (range 1–14%) of pregnancies. The incidence of GDM is greatly increased in certain ethnic groups, including African Americans and Latinas, consistent with a similar increased risk of type 2 DM. Current recommendations advise screening for glucose intolerance between weeks 24 and 28 of pregnancy in women not known to have diabetes. Therapy for GDM is similar to that for individuals with pregnancy-associated diabetes and involves MNT and insulin, if hyperglycemia persists. Oral glucose-lowering agents are not approved for use during pregnancy, but studies using metformin or glyburide have shown efficacy and have not found toxicity. With current practices, the morbidity and mortality rates of the mother with GDM and the fetus are not different from those in the nondiabetic population. Individuals who develop GDM are at marked increased risk for developing type 2 DM in the future and should be screened periodically for DM (see screening recommendations in **Chap. 396**). Most individuals with GDM revert to normal glucose tolerance after delivery, but some will continue to have overt diabetes or impairment of glucose tolerance after delivery. In addition, children of women with

GDM appear to be at risk for obesity and glucose intolerance and have an increased risk of diabetes beginning in the later stages of adolescence.

Pregnancy in individuals with known DM requires meticulous planning and adherence to strict treatment regimens. Intensive insulin therapy and near-normalization of the HbA_{1c} (<6.5%) are essential for individuals with existing DM who are planning pregnancy. Consideration should be given to insulin infusion and CGM devices that may help to improve glycemic control prior to conception since the most crucial period of glycemic control is soon after fertilization. The risk of fetal malformations is increased 4–10 times in individuals with uncontrolled DM at the time of conception, and normal plasma glucose during the preconception period and throughout the periods of organ development in the fetus should be the goal, with more frequent monitoring of HbA1c every 2 months throughout gestation.

LIPODYSTROPHIC DM

Lipodystrophy, or the loss of subcutaneous fat tissue, may be generalized in certain genetic conditions such as leprechaunism, or acquired as part of an autoimmune disorder. Generalized lipodystrophy is associated with leptin deficiency and severe insulin resistance and is often accompanied by acanthosis nigricans, hepatic steatosis, and severe hypertriglyceridemia. Recombinant human leptin (metreleptin) may allow for the achievement of metabolic control in generalized lipodystrophy, but is associated with the development of neutralizing antibodies and is available only through a restricted program under a Risk Evaluation and Mitigation Strategy (REMS).

Protease Inhibitors and Lipodystrophy

Protease inhibitors used in the treatment of HIV disease **(Chap. 197)** have been associated with a centripetal accumulation of fat (visceral and abdominal area), accumulation of fat in the dorsocervical region, loss of extremity fat, decreased insulin sensitivity (elevations of the fasting insulin level and reduced glucose tolerance on IV glucose tolerance testing), and dyslipidemia. Although many aspects of the physical appearance of these individuals resemble Cushing's syndrome, increased cortisol levels do not account for this appearance. The possibility remains that this is related to HIV infection by some undefined mechanism, because some features of the syndrome were observed before the introduction of protease inhibitors. Therapy for HIV-related lipodystrophy is not well established.

FURTHER READING

American Diabetes Association: Comprehensive Medical Evaluation and Assessment of Comorbidities. Diabetes Care 41:S28, 2018. [PubMed: 29222374]

Evert AB et al: Nutrition therapy recommendations for the management of adults with diabetes. Diabetes Care 37:S120, 2014. [PubMed: 24357208]

Glycemic Targets. Diabetes care older adults. Diabetes Care 41:S119, 2018. [PubMed: 29222382]

Hassanein M et al: Diabetes and Ramadan: Practical guidelines. Diabetes Res Clin Pract 126:303, 2017. [PubMed: 28347497]

Lifestyle Management. Diabetes Care 41:S38, 2018. [PubMed: 29222375]

Nyenwe EA, Kitabchi AE: The evolution of diabetic ketoacidosis: An update of its etiology, pathogenesis and management. Metabolism 65:507, 2016. [PubMed: 26975543]

Qaseem A et al: Oral pharmacologic treatment of type 2 diabetes mellitus: A clinical practice guideline update from the American College of Physicians. Ann Intern Med 166:279, 2017. [PubMed: 28055075]

Rubino F et al: Metabolic surgery in the treatment algorithm for type 2 diabetes: A joint statement by international diabetes organizations. Diabetes Care 39:86, 2016.

Thabit H, Hovorka R: Coming of age: the artificial pancreas for type 1 diabetes. Diabetologia 59:1795, 2016. [PubMed: 27364997]

Young-Hyman D et al: Psychosocial care for people with diabetes: A position statement of the American Diabetes Association. Diabetes Care 39:2126, 2016. [PubMed: 27879358]

Chapter 398: Diabetes Mellitus: Complications

Alvin C. Powers; John M. Stafford; Michael R. Rickels

INTRODUCTION

Diabetes-related complications affect many organ systems and are responsible for the majority of morbidity and mortality associated with the disease. For many years in the United States, diabetes has been the leading cause of new blindness in adults, renal failure, and nontraumatic lower extremity amputation. More recently, diabetes has also emerged as a leading contributor to coronary heart disease (CHD). Diabetes-associated complications related to hyperglycemia usually do not appear until the second decade of hyperglycemia. In contrast, diabetes-associated CHD risk, related in part to insulin resistance, may develop before hyperglycemia is established. Because type 2 diabetes mellitus (DM) often has a long asymptomatic period of hyperglycemia before diagnosis, many individuals with type 2 DM have both glucose-related and insulin resistance-related complications at the time of diagnosis. Fortunately, many of the diabetes-related complications can be prevented or delayed with a focus on diet, fitness, early detection, aggressive glycemic control, and efforts to minimize the risks of complications. Recent studies show a decline in diabetes-related complications in individuals, but this is tempered by the increase in the number of individuals with diabetes. For example, the rate of myocardial infarction (MI) associated with diabetes declined by 67% between 1990 and 2010.

Diabetes-related complications can be divided into vascular and nonvascular complications and are similar for type 1 and type 2 DM **(Table 398-1)**. The vascular complications of DM are further subdivided into microvascular (retinopathy, neuropathy, nephropathy) and macrovascular complications (CHD, peripheral arterial disease [PAD], cerebrovascular disease). Microvascular complications are diabetes-specific, whereas macrovascular complications have pathophysiologic features that are both shared with the general population and diabetes-specific. Nonvascular complications include infections, skin changes, and hearing loss. Some studies suggest that 2 DM increases the risk of dementia and impaired cognitive function.

TABLE 398-1

Diabetes-Related Complications

Microvascular
Eye disease
Retinopathy (nonproliferative/proliferative)
Macular edema
Neuropathy
Sensory and motor (mono- and polyneuropathy)
Autonomic
Nephropathy (albuminuria and declining renal function)
Macrovascular
Coronary heart disease
Peripheral arterial disease
Cerebrovascular disease
Other
Gastrointestinal (gastroparesis, diarrhea)
Genitourinary (uropathy/sexual dysfunction)
Dermatologic
Infectious
Cataracts
Glaucoma
Cheiroarthropathy[a]

Periodontal disease
Hearing loss
Other comorbid conditions associated with diabetes (relationship to hyperglycemia is uncertain): depression, obstructive sleep apnea, fatty liver disease, hip fracture, osteoporosis (in type 1 diabetes), cognitive impairment or dementia, low testosterone in men.

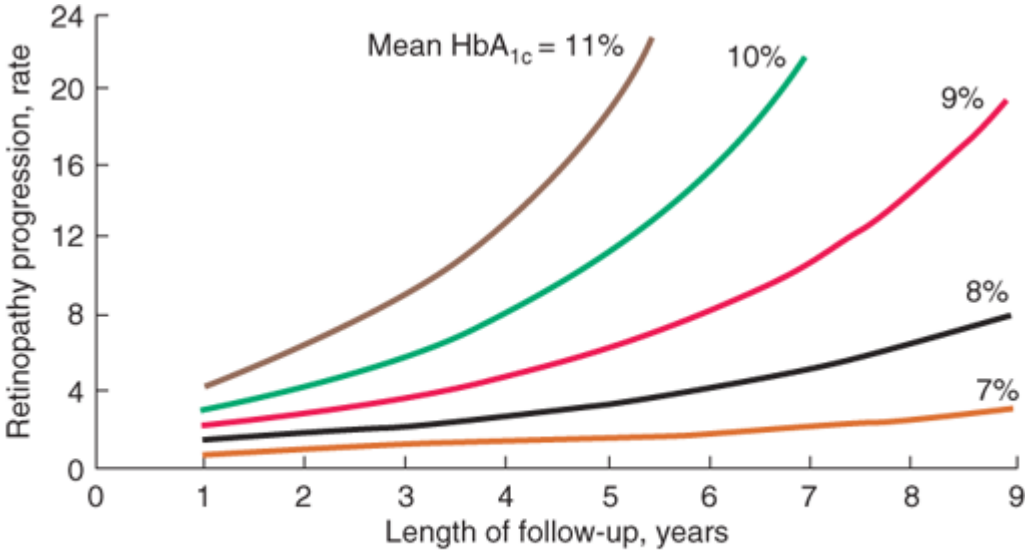
[a]Thickened skin and reduced joint mobility.

GLYCEMIC CONTROL AND COMPLICATIONS

The microvascular complications of both type 1 and type 2 DM result from chronic hyperglycemia **(Fig. 398-1)**. Evidence implicating a causative role for chronic hyperglycemia in the development of macrovascular complications is less conclusive. CHD events and mortality rate are two to four times greater in patients with type 2 DM and correlate with fasting and postprandial plasma glucose levels as well the hemoglobin A_{1c} (HbA_{1c}). Other factors such as dyslipidemia and hypertension also play important roles in macrovascular complications.

FIGURE 398-1

Relationship of glycemic control and diabetes duration to diabetic retinopathy. The progression of retinopathy in individuals in the Diabetes Control and Complications Trial is graphed as a function of the length of follow-up with different curves for different hemoglobin A_{1c} (HbA_{1c}) values. *(Adapted from The Diabetes Control and Complications Trial Research Group: Diabetes 44:968, 1995.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The Diabetes Control and Complications Trial (DCCT) provided definitive proof that reduction in chronic hyperglycemia can prevent many complications of type 1 DM (Fig. 398-1). This large multicenter clinical trial randomized >1400 individuals with type 1 DM to either intensive or conventional diabetes management and prospectively evaluated the development of diabetes-related complications during a mean follow-up of 6.5 years. Individuals in the intensive diabetes management group received multiple administrations of insulin each day (injection or pump) along with extensive educational, psychological, and medical support. Individuals in the conventional diabetes management group received twice-daily insulin injections and quarterly nutritional, educational, and clinical evaluation. The goal in the former group was normoglycemia; the goal in the latter group was prevention of symptoms of diabetes. Individuals in the intensive diabetes management group achieved a substantially lower HbA_{1c} (7.3%) than individuals in the conventional diabetes management group (9.1%). After the DCCT results were reported in 1993, study participants continue to be followed in the Epidemiology of Diabetes Intervention and Complications (EDIC) trial, which recently completed 30 years of follow-up (DCCT + EDIC). At the end of the DCCT phase, study participants in both intensive and conventional arms were offered intensive therapy. However, during the subsequent follow-up of >18 years, the initial separation in glycemic control disappeared with both arms maintaining a mean HbA_{1c} of 8.0%.

The DCCT phase demonstrated that improvement of glycemic control reduced nonproliferative and proliferative retinopathy (47% reduction), albuminuria (39% reduction), clinical nephropathy (54% reduction), and neuropathy (60% reduction). Improved glycemic control also slowed the progression of early diabetic complications. During the DCCT phase, weight gain (4.6 kg) and severe hypoglycemia (requiring assistance of another person to treat) were more common in the intensive therapy group. The benefits of an improvement in glycemic control occurred over the entire range of HbA_{1c} values (Fig. 398-1), indicating that at any HbA_{1c} level, an improvement in glycemic control is beneficial. The results of the DCCT predicted that individuals in the intensive diabetes management group would gain 7.7 additional years of vision, 5.8 additional years free from end-stage renal disease (ESRD), and 5.6 years free from lower extremity amputations. If all complications of DM were combined, individuals in the intensive diabetes management group would experience >15.3 more years of life without significant microvascular complications of DM, compared to individuals who received standard therapy. This translates into an additional 5.1 years of life expectancy for individuals in the intensive diabetes management group. The 30-year follow-up data in the intensively treated group show a continued reduction in retinopathy, nephropathy, and cardiovascular disease. For example, individuals in the intensive therapy group had a 42–57% reduction in cardiovascular events (nonfatal MI, stroke, or death from a cardiovascular event) at a mean follow-up of 18 years, even though their subsequent glycemic control was the same as those in the conventional diabetes management group from years 6.5 to 17. During the EDIC phase, <1% of the cohort had become blind, lost a limb to amputation, or required dialysis. Other complications of diabetes, including autonomic neuropathy, bladder and sexual dysfunction, and cardiac autonomic neuropathy, were reduced in the intensive therapy group. Importantly, those in the intensive therapy group did not have a difference in cognitive function either during the DCCT phase (when the frequency of hypoglycemia was greater) or during the follow-up EDIC phase (when the frequency of hypoglycemia was similar).

The United Kingdom Prospective Diabetes Study (UKPDS) studied the course of >5000 individuals with type 2 DM for >10 years. This study used multiple treatment regimens and monitored the effect of intensive glycemic control and risk factor treatment on the development of diabetic complications. Newly diagnosed individuals with type 2 DM were randomized to (1) intensive management using various combinations of insulin, a sulfonylurea, or metformin or (2) conventional therapy using dietary modification and pharmacotherapy with the goal of symptom prevention. In addition, individuals were randomly assigned to different antihypertensive regimens. Individuals in the intensive treatment arm achieved an HbA_{1c} of 7%, compared to a 7.9% HbA_{1c} in the standard treatment group. The UKPDS demonstrated that each percentage point reduction in HbA_{1c} was associated with a 35% reduction in microvascular complications. As in the DCCT, there was a continuous relationship between glycemic control and development of complications. Improved glycemic control also reduced the cardiovascular event rate in the follow-up period of >10 years.

One of the major findings of the UKPDS was that strict blood pressure control significantly reduced both macro- and microvascular complications. In fact, the beneficial effects of blood pressure control were greater than the beneficial effects of glycemic control. Lowering blood pressure to moderate goals (144/82 mmHg) reduced the risk of DM-related death, stroke, microvascular endpoints, retinopathy, and heart failure (risk reductions between 32 and 56%).

Similar reductions in the risks of retinopathy and nephropathy were also seen in a small trial of lean Japanese individuals with type 2 DM randomized to either intensive glycemic control or standard therapy with insulin (Kumamoto study). These results demonstrate the effectiveness of improved glycemic control in individuals of different ethnicity and, presumably, a different etiology of DM (i.e., phenotypically different from those in the DCCT and UKPDS). The Action to Control Cardiovascular Risk in Diabetes (ACCORD) and Action in Diabetes and Vascular Disease: Preterax and Diamicron MR Controlled Evaluation (ADVANCE) trials also found that improved glycemic control reduced microvascular complications.

Thus, these large clinical trials in type 1 and type 2 DM indicate that chronic hyperglycemia plays a causative role in the pathogenesis of diabetic microvascular complications. In both the DCCT and the UKPDS, cardiovascular events were reduced at follow-up of >10 years, even though the improved glycemic control was not maintained. The positive impact of a period of improved glycemic control on later disease has been termed a *legacy effect* or *metabolic memory*.

A summary of the features of diabetes-related complications includes the following. (1) Duration and degree of hyperglycemia correlate with complications. (2) Intensive glycemic control is beneficial in all forms of DM. (3) Blood pressure control is critical, especially in type 2 DM. (4) Survival in patients with type 1 DM is improving, and diabetes-related complications are declining. (5) Not all individuals with diabetes develop diabetes-related complications. Other incompletely defined factors appear to modulate the development of complications. For example, despite long-standing DM, some individuals never develop nephropathy or retinopathy. Many of these patients have glycemic control that is indistinguishable from those who develop microvascular complications, suggesting a genetic susceptibility for developing particular complications, especially retinopathy and nephropathy.

MECHANISMS OF COMPLICATIONS

Chronic hyperglycemia is the important etiologic factor leading to complications of DM, but the mechanism(s) by which it leads to such diverse cellular and organ dysfunction is unknown. An emerging hypothesis is that hyperglycemia leads to epigenetic changes **(Chap. 456)** that influence gene expression in affected cells. Other hypotheses are that chronic hyperglycemia leads to formation of advanced glycosylation end products (AGEs; e.g., pentosidine, glucosepane, and carboxymethyllysine) which bind to specific cell surface receptor and/or the nonenzymatic glycosylation of intra- and extracellular proteins, leading to cross-linking of proteins, accelerated atherosclerosis, glomerular dysfunction, endothelial dysfunction, and altered extracellular matrix composition. Other theories predict that hyperglycemia: (1) increases glucose metabolism via the sorbitol pathway related to the enzyme aldose reductase; (2) increases the formation of diacylglycerol, leading to activation of protein kinase C, which alters the transcription of genes for fibronectin, type IV collagen, contractile proteins, and extracellular matrix proteins in endothelial cells and neurons; and/or (3) increases the flux through the hexosamine pathway, which generates fructose-6-phosphate, a substrate for O-linked glycosylation and proteoglycan production, leading to altered function by glycosylation of proteins such as endothelial nitric oxide synthase.

Growth factors may play an important role in some diabetes-related microvascular complications, and their production is increased by most of these proposed pathways. For example, vascular endothelial growth factor A (VEGF-A) is increased locally in diabetic proliferative retinopathy and decreases after laser photocoagulation. A possible unifying mechanism is that hyperglycemia leads to increased production of reactive oxygen species or superoxide in the mitochondria and this may activate several of the pathways described above. Although hyperglycemia serves as the initial trigger for complications of diabetes, it is still unknown whether the same pathophysiologic processes are operative in all complications or whether some pathways predominate in certain organs.

The mechanisms of diabetes-related macrovascular complications including MI and stroke are glucose-related mechanisms, but also include traditional cardiovascular risk factors (dyslipidemia, hypertension), and insulin resistance. In type 2 diabetes, insulin resistance is present years prior to diagnosis and is associated with obesity and ectopic accumulation of lipids in muscle and liver. Additionally, insulin fails to appropriately suppress lipolysis from adipose tissue, which results in increased delivery of fatty acids to liver, muscle, endothelial cells, and cardiac tissues, leading to tissue accumulation of triglycerides, diacylglycerol, and ceramides.

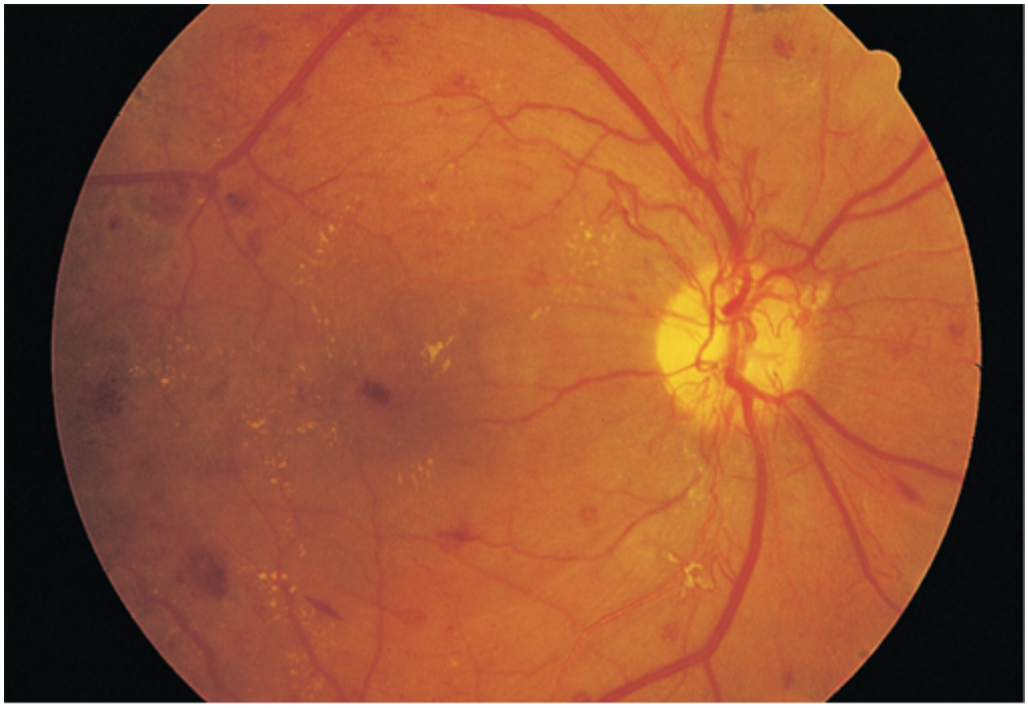
OPHTHALMOLOGIC COMPLICATIONS OF DIABETES MELLITUS

DM is the leading cause of blindness between the ages of 20 and 74 in the United States. The gravity of this problem is highlighted by the finding that individuals with DM are 25 times more likely to become legally blind than individuals without DM. Severe vision loss is primarily the result of progressive diabetic retinopathy which leads significant macular edema and new blood vessel formation. Diabetic retinopathy is classified into two stages: nonproliferative and proliferative. Nonproliferative diabetic retinopathy usually

appears late in the first decade or early in the second decade of the disease and is marked by retinal vascular microaneurysms, blot hemorrhages, and cotton-wool spots **(Fig. 398-2)**. Mild nonproliferative retinopathy may progress to more extensive disease, characterized by changes in venous vessel caliber, intraretinal microvascular abnormalities, and more numerous microaneurysms and hemorrhages. The pathophysiologic mechanisms invoked in nonproliferative retinopathy include loss of retinal pericytes, increased retinal vascular permeability, alterations in retinal blood flow, and abnormal retinal microvasculature, all of which can lead to retinal ischemia.

FIGURE 398-2

Diabetic retinopathy results in scattered hemorrhages, yellow exudates, and neovascularization. This patient has neovascular vessels proliferating from the optic disc, requiring urgent panretinal laser photocoagulation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The appearance of neovascularization in response to retinal hypoxemia is the hallmark of proliferative diabetic retinopathy (Fig. 398-2). These newly formed vessels appear near the optic nerve and/or macula and rupture easily, leading to vitreous hemorrhage, fibrosis, and ultimately retinal detachment. Not all individuals with nonproliferative retinopathy go on to develop proliferative retinopathy, but the more severe the nonproliferative disease, the greater the chance of evolution to proliferative retinopathy within 5 years. This creates an important opportunity for early detection and treatment of diabetic retinopathy. Clinically significant macular edema can occur in the context of nonproliferative or proliferative retinopathy. Fluorescein angiography and optical coherence tomography are useful to detect macular edema, which is associated with a 25% chance of moderate visual loss over the next 3 years. Duration of DM and degree of glycemic control are the best predictors of the development of retinopathy; hypertension, nephropathy, and dyslipidemia are also risk factors. Nonproliferative retinopathy is found in many individuals who have had

DM for >20 years. Although there is genetic susceptibility for retinopathy, it confers less influence than either the duration of DM or the degree of glycemic control.

TREATMENT

TREATMENT

Diabetic Retinopathy

The most effective therapy for diabetic retinopathy is prevention. Intensive glycemic and blood pressure control will delay the development or slow the progression of retinopathy in individuals with either type 1 or type 2 DM. Paradoxically, during the first 6–12 months of improved glycemic control, established diabetic retinopathy may transiently worsen. Fortunately, this progression is temporary, and in the long term, improved glycemic control is associated with less diabetic retinopathy. Individuals with known retinopathy may be candidates for prophylactic laser photocoagulation when initiating intensive therapy. Once advanced retinopathy is present, improved glycemic control imparts less benefit, although adequate ophthalmologic care can prevent most blindness. Fenofibrate, while not reducing cardiovascular events in individuals with diabetes and dyslipidemia, does reduce the progression of retinopathy.

Regular, comprehensive eye examinations are essential for all individuals with DM **(see Table 397-1)**. Most diabetic eye disease can be successfully treated if detected early. Routine, nondilated eye examinations by the primary care provider or diabetes specialist are inadequate to detect diabetic eye disease, which requires a dilated eye exam performed by an optometrist or ophthalmologist, and subsequent management by a retinal specialist for optimal care of these disorders. Treatment of proliferative retinopathy or macular edema with laser photocoagulation and/or anti-VEGF therapy (ocular injection) usually is successful in preserving vision. Aspirin therapy (650 mg/d) does not appear to influence the natural history of diabetic retinopathy.

RENAL COMPLICATIONS OF DIABETES MELLITUS

Diabetic nephropathy is the leading cause of chronic kidney disease (CKD), ESRD, and CKD requiring renal replacement therapy. Furthermore, the prognosis of individuals with diabetes on dialysis is poor. Albuminuria in individuals with DM is associated with an increased risk of cardiovascular disease. Individuals with diabetic nephropathy commonly have diabetic retinopathy.

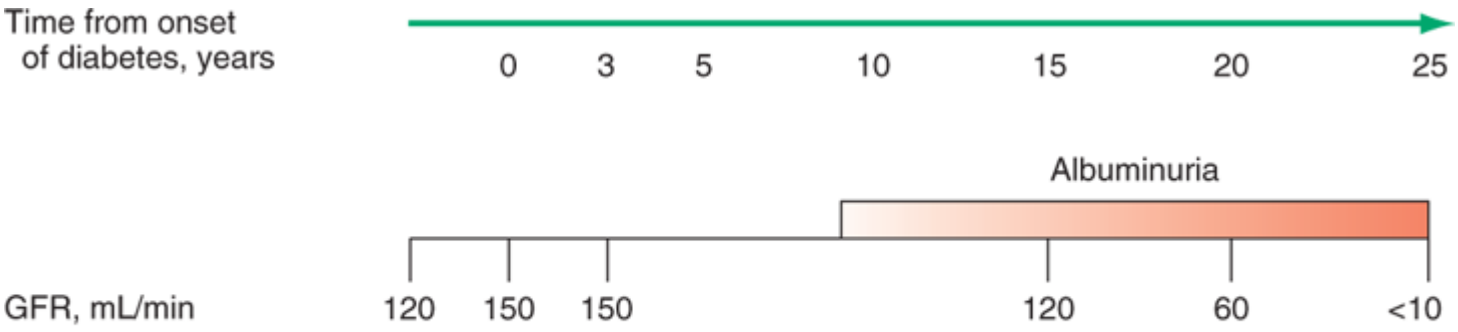
Like other microvascular complications, the pathogenesis of diabetic nephropathy is related to chronic hyperglycemia. The mechanisms by which chronic hyperglycemia leads to diabetic nephropathy, although incompletely defined, involve the effects of soluble factors (growth factors, angiotensin II, endothelin, advanced glycation end products [AGEs]), hemodynamic alterations in the renal microcirculation (glomerular hyperfiltration or hyperperfusion, increased glomerular capillary pressure), and structural changes in the glomerulus (increased extracellular matrix, basement membrane thickening, mesangial expansion, fibrosis). Some of these effects may be mediated through angiotensin II receptors. Smoking

accelerates the decline in renal function. Because only 20–40% of patients with diabetes develop diabetic nephropathy, additional genetic or environmental susceptibility factors remain unidentified. Known risk factors include race and a family history of diabetic nephropathy. Diabetic nephropathy and ESRD secondary to DM develop more commonly in African Americans, Native Americans, and Hispanic individuals with diabetes.

The natural history of diabetic nephropathy is characterized by a sequence of events that was initially defined for individuals with type 1 DM but appears similar in type 2 DM **(Fig. 398-3)**. Glomerular hyperperfusion and renal hypertrophy occur in the first years after the onset of DM and are associated with an increase of the estimated glomerular filtration rate (GFR). During the first 5 years of DM, thickening of the glomerular basement membrane, glomerular hypertrophy, and mesangial volume expansion occur as the GFR returns to normal. After 5–10 years of type 1 DM, many individuals begin to excrete small amounts of albumin in the urine. The American Diabetes Association (ADA) no longer uses the terms microalbuminuria or macroalbuminuria (previously used to refer to increased urinary protein of different levels) and instead uses the term albuminuria to refer to increased urinary protein excretion (spot urinary albumin-to-creatinine ratio >30 mg/g Cr). Likewise, this chapter uses the albuminuria, but emphasizes that this should be persistent and is a continuous variable. In some individuals with type 1 diabetes and albuminuria of short duration, the albuminuria regresses. Albuminuria is a risk factor for cardiovascular disease (CVD) and CKD. Diabetic kidney disease refers to albuminuria and reduced GFR (<60 mL/min/1.73 m^2); CKD related to diabetes, which may not be accompanied by albuminuria, is also discussed in **Chap. 305**. Most patients with CKD related to diabetes also have diabetic retinopathy. Once there is marked albuminuria and a reduction in GFR, the pathologic changes are likely irreversible.

FIGURE 398-3

Time course of development of diabetic nephropathy. The relationship of time from onset of diabetes, albuminuria, and the glomerular filtration rate (GFR) are shown.



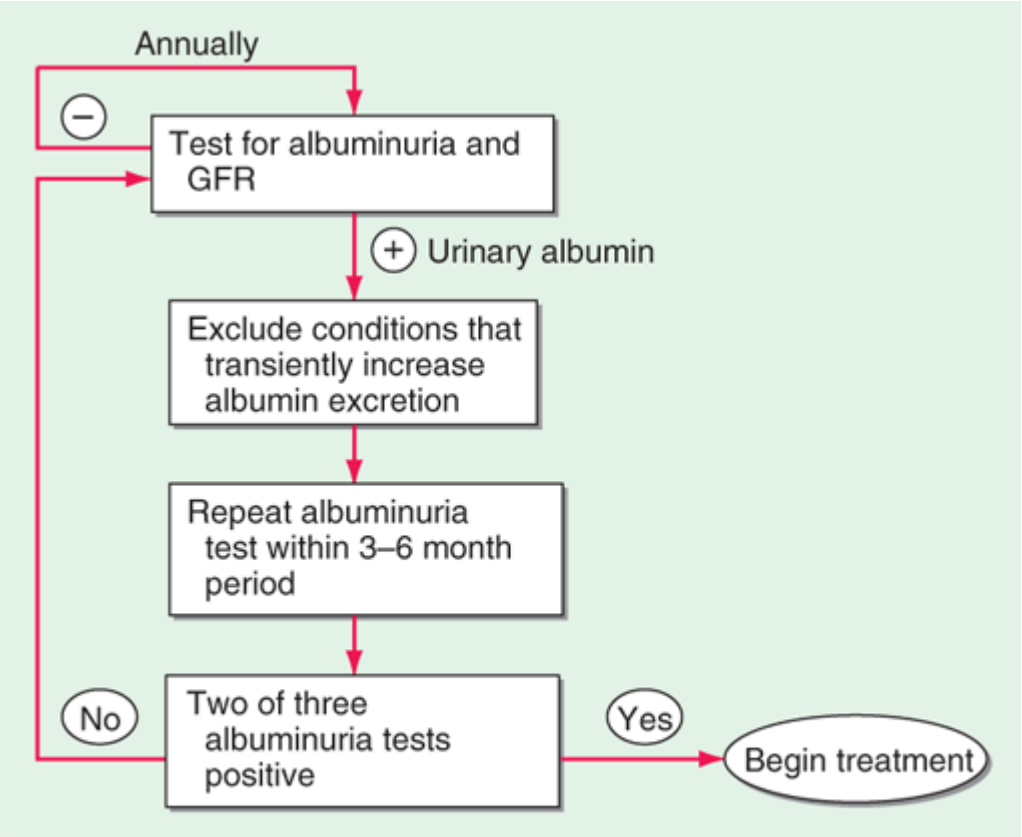
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The nephropathy that develops in type 2 DM differs from that of type 1 DM in the following respects: (1) albuminuria may be present when type 2 DM is diagnosed, reflecting its long asymptomatic period; (2) hypertension more commonly accompanies albuminuria; and (3) albuminuria may be less predictive of diabetic kidney disease. Finally, it should be noted that albuminuria in type 2 DM may be secondary to factors unrelated to DM, such as hypertension, congestive heart failure (CHF), prostate disease, or infection.

As part of comprehensive diabetes care (Chap. 397), albuminuria should be detected at an early stage when effective therapies can be instituted. Because some individuals with type 1 or type 2 DM have a decline in GFR in the absence of albuminuria, assessment should include a spot urinary albumin-to-creatinine ratio and an estimated GFR (Fig. 398-4). The urine protein measurement by routine urinalysis does not detect low levels of albumin excretion. Screening for albuminuria should commence 5 years after the onset of type 1 DM and at the time of diagnosis of type 2 DM.

FIGURE 398-4

Screening for albuminuria should be performed annually in patients with type 1 diabetes for ≥5 years, in patients with type 2 diabetes, and during pregnancy. Albuminuria = spot urinary albumin-to-creatinine ratio >30 mg/g Cr; GRD = estimated glomerular filtration rate. Nondiabetes-related conditions that might increase cause albuminuria are urinary tract infection, hematuria, heart failure, febrile illness, severe hyperglycemia, severe hypertension, and vigorous exercise.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Type IV renal tubular acidosis (hyporeninemic hypoaldosteronism) may occur in type 1 or 2 DM. These individuals develop a propensity to hyperkalemia and acidemia, which may be exacerbated by medications (especially angiotensin-converting enzyme [ACE] inhibitors, angiotensin receptor blockers [ARBs], and mineralocorticoid receptor antagonists). Patients with DM are predisposed to radiocontrast-induced nephrotoxicity. Risk factors for radiocontrast-induced nephrotoxicity are preexisting nephropathy and volume depletion. Individuals with DM undergoing radiographic procedures with contrast dye should be well hydrated before and after dye exposure, and the serum creatinine should be monitored for 24–48 h following the procedure. Metformin should be held until postintervention confirmation of preserved kidney function.

TREATMENT

TREATMENT

Diabetic Nephropathy

The optimal therapy for diabetic nephropathy is prevention by control of glycemia **(Chap. 397 outlines glycemic goals and approaches)**. Interventions effective in slowing progression of albuminuria include (1) improved glycemic control, (2) strict blood pressure control, and (3) administration of an ACE inhibitor or ARB. Dyslipidemia should also be treated.

Improved glycemic control reduces the rate at which albuminuria appears and progresses in type 1 and type 2 DM. However, once there is a large amount of albuminuria, it is unclear whether improved glycemic control will slow progression of renal disease. During the later phase of declining renal function, insulin requirements may fall as the kidney is a site of insulin degradation. As the GFR decreases with progressive nephropathy, the use and dose of glucose-lowering agents should be reevaluated **(see Table 397-5)**. Some glucose-lowering medications (sulfonylureas and metformin) are contraindicated in advanced renal insufficiency.

Many individuals with type 1 or type 2 DM develop hypertension. Numerous studies in both type 1 and type 2 DM demonstrate the effectiveness of strict blood pressure control in reducing albumin excretion and slowing the decline in renal function. Blood pressure should be maintained at <140/90 mmHg in individuals with diabetes and possibly <130/80 in individuals at increased risk for CVD and CKD progression.

Either ACE inhibitors or ARBs should be used to reduce the albuminuria and the associated decline in GFR that accompanies it in individuals with type 1 or type 2 DM (see "Hypertension," below). Most experts believe that the two classes of drugs are equivalent in patient with diabetes. ARBs can be used as an alternative in patients who develop ACE inhibitor–associated cough or angioedema. After initiation of therapy, some increase the dose and monitor the urinary albumin. There is no benefit of intervention prior to onset of albuminuria or using a combination of an ACE inhibitor and an ARB. If use of either ACE inhibitors or ARBs is not possible or the blood pressure is not controlled, then, diuretics, calcium channel blockers (nondihydropyridine class), or beta blockers should be used. Some glucagon-like peptide-1 (GLP-1) receptor agonists and sodium-glucose co-transporter 2 Inhibitors improve glycemic control and reduce the progression of diabetic kidney disease in individuals with T2DM and established CVD **(Chap 397)**.

The ADA suggests a protein intake of 0.8 mg/kg of body weight/day in individuals with diabetic kidney disease.

Nephrology consultation should be considered when albuminuria appears and when the estimated GFR is <30 mL/min per 1.743 m^2. Complications of atherosclerosis are the leading cause of death in diabetic individuals with nephropathy and hyperlipidemia should be treated aggressively. Referral for transplant evaluation should be made when the GFR approaches 20 mL/min per 1.743 m^2. Preemptive (before dialysis) renal transplantation from a living kidney donor may be a preferred therapy. Kidney transplantation can be

performed alone or as a combined pancreas-kidney transplant, which offers the promise of normoglycemia and freedom from both insulin and dialysis. As compared with nondiabetic individuals, hemodialysis in patients with DM is associated with more frequent complications, such as hypotension (due to autonomic neuropathy or loss of reflex tachycardia), more difficult vascular access, and accelerated progression of retinopathy.

NEUROPATHY AND DIABETES MELLITUS

Diabetic neuropathy, which occurs in ~50% of individuals with long-standing type 1 and type 2 DM, manifests as a diffuse neuropathy (distal symmetrical polyneuropathy and/or autonomic neuropathy), a mononeuropathy, and/or a radiculopathy/polyradiculopathy. As with other complications of DM, the development of neuropathy correlates with the duration of diabetes and glycemic control. Additional risk factors are body mass index (BMI) (the greater the BMI, the greater the risk of neuropathy) and smoking. The presence of CVD, elevated triglycerides, and hypertension is also associated with diabetic peripheral neuropathy. Both myelinated and unmyelinated nerve fibers are lost. Because the clinical features of diabetic neuropathy are similar to those of other neuropathies, the diagnosis of diabetic neuropathy should be made only after other possible etiologies are excluded **(Chap. 438)**.

Distal Symmetric Polyneuropathy (DSPN)

DSPN, the most common form of diabetic neuropathy, most frequently presents with distal sensory loss and pain, but up to 50% of patients do not have symptoms of neuropathy. Symptoms may include a sensation of numbness, tingling, sharpness, or burning that begins in the feet and spreads proximally. Hyperesthesia, paresthesia, and dysesthesia also may occur. Pain typically involves the lower extremities, is usually present at rest, and worsens at night. Both an acute (lasting <12 months) and a chronic form of painful diabetic neuropathy may occur. The acute form is sometimes treatment-related, occurring in the context of improved glycemic control. As diabetic neuropathy progresses, the pain subsides and eventually disappears, but a sensory deficit persists and motor defects may develop. Physical examination **(Chap. 396)** often reveals sensory loss (to 10-g monofilament and/or vibration), loss of ankle deep-tendon reflexes, abnormal position sense, and muscular atrophy or foot drop. Annual screening for DSPN should begin 5 years after diagnosis of type 1 DM and at the time of diagnosis of type 2 DM and is aimed at detecting loss of protective sensation (LOPS). LOPS and DSPN are major risk factors for foot ulceration and falls due to small and large nerve fiber dysfunction.

Autonomic Neuropathy

Individuals with long-standing type 1 or 2 DM may develop signs of autonomic dysfunction involving the cholinergic, noradrenergic, and peptidergic (peptides such as pancreatic polypeptide, substance P, etc.) systems. DM-related autonomic neuropathy can involve multiple systems, including the cardiovascular, gastrointestinal (GI), genitourinary, sudomotor, and metabolic systems. Cardiovascular autonomic neuropathy, reflected by decreased heart rate variability, resting tachycardia and orthostatic hypotension is

associated with an increase in CVD. Reports of sudden death in DM have also been attributed to cardiovascular autonomic neuropathy. Gastroparesis and bladder-emptying abnormalities are often caused by the autonomic neuropathy seen in DM (discussed below). Hyperhidrosis of the upper extremities and anhidrosis of the lower extremities result from sympathetic nervous system dysfunction. Anhidrosis of the feet can promote dry skin with cracking, which increases the risk of foot ulcers. Autonomic neuropathy may reduce counterregulatory hormone release (especially catecholamines), leading to an inability to sense hypoglycemia appropriately (hypoglycemia unawareness; **Chap. 399**), thereby subjecting the patient to the risk of severe hypoglycemia and complicating efforts to improve glycemic control.

Mononeuropathy and/or Radiculopathy/Polyradiculopathy

Mononeuropathy (dysfunction of isolated cranial or peripheral nerves) is less common than polyneuropathy in DM and presents with pain and motor weakness in the distribution of a single nerve. Mononeuropathies can occur at entrapment sites such as carpal tunnel or be noncompressive. Involvement of the third cranial nerve is most common and is heralded by diplopia. Physical examination reveals ptosis and ophthalmoplegia with normal pupillary constriction to light. Sometimes other cranial nerves, such as IV, VI, or VII (Bell's palsy), are affected. Peripheral mononeuropathies or simultaneous involvement of more than one nerve (mononeuropathy multiplex) may also occur. Diabetic radiculopathy or polyradiculopathy is a syndrome characterized by severe disabling pain in the distribution of one or more nerve roots. It may be accompanied by motor weakness. Intercostal or truncal radiculopathy causes pain over the thorax or abdomen. Involvement of the lumbar plexus or femoral nerve may cause severe pain in the thigh or hip and may be associated with muscle weakness in the hip flexors or extensors (diabetic amyotrophy). Fortunately, diabetic polyradiculopathies are usually self-limited and resolve over 6–12 months.

TREATMENT

TREATMENT

Diabetic Neuropathy

Prevention of diabetic neuropathy is critical through improved glycemic control. Treatment of diabetic neuropathy is less than satisfactory. Lifestyle modifications (exercise, diet) has some efficacy in DSPN in type 2 DM and hypertension and hypertriglyceridemia should be treated. Efforts to improve glycemic control in long-standing diabetes may be confounded by hypoglycemia unawareness. Avoidance of neurotoxins (alcohol) and smoking, supplementation with vitamins for possible deficiencies (B_{12}, folate; **Chap. 326**).

Patients should be educated that loss of sensation in the foot increases the risk for ulceration and its sequelae and that prevention of such problems is paramount. Patients with symptoms or signs of neuropathy or LOPS should check their feet daily and take precautions (footwear) aimed at preventing calluses or ulcerations. If foot deformities are present, a podiatrist should be involved.

Chronic, painful diabetic neuropathy is difficult to treat with only symptomatic treatment being available; evidence of the effectiveness of improved glycemic control in painful diabetic neuropathy is lacking. Two

agents, duloxetine and pregabalin, have been approved by the U.S. Food and Drug Administration (FDA) for pain associated with diabetic neuropathy. Tapentadol, a centrally acting opioid, is also FDA-approved, but has only modest efficacy and poses addiction risk, making it and other opioids less desirable and not a first-line therapy. Diabetic neuropathy may respond to tricyclic antidepressants, gabapentin, venlafaxine, carbamazepine, tramadol, and topical capsaicin, although none of these are FDA-approved for this indication. No direct comparisons of agents are available, and it is reasonable to switch agents if there is no response or if side effects develop. Referral to a pain management center may be necessary. Because the pain of acute diabetic neuropathy may resolve over time, medications may be discontinued as progressive neuronal damage from DM occurs.

Therapy of orthostatic hypotension secondary to autonomic neuropathy is also difficult. Nonpharmacologic maneuvers (adequate salt intake, avoidance of dehydration and diuretics, lower extremity support hose, and physical activity) may offer some benefit. A variety of agents have limited success (midodrine and droxidopa are FDA-approved for orthostatic hypotension of any etiology). Patients with resting tachycardia may be considered for beta-blocker therapy with caution exercised if there is hypoglycemia unawareness.

GASTROINTESTINAL/GENITOURINARY DYSFUNCTION

Long-standing type 1 and 2 DM may affect the motility and function of the GI and genitourinary systems. The most prominent GI symptoms are delayed gastric emptying (gastroparesis) and altered small- and large-bowel motility (constipation or diarrhea). Gastroparesis may present with symptoms of anorexia, nausea, vomiting, early satiety, and abdominal bloating. Microvascular complications (retinopathy and neuropathy) are usually present. Nuclear medicine scintigraphy after ingestion of a radiolabeled meal may document delayed gastric emptying, but may not correlate well with the patient's symptoms. Noninvasive "breath tests" following ingestion of a radiolabeled meal are emerging as a diagnostic tool. Although parasympathetic dysfunction secondary to chronic hyperglycemia is important in the development of gastroparesis, hyperglycemia itself also impairs gastric emptying. Nocturnal diarrhea, alternating with constipation, is a feature of DM-related GI autonomic neuropathy. In type 1 DM, these symptoms should also prompt evaluation for celiac sprue because of its increased frequency.

Diabetic autonomic neuropathy may lead to genitourinary dysfunction including cystopathy and female sexual dysfunction (reduced sexual desire, dyspareunia, reduced vaginal lubrication). Symptoms of diabetic cystopathy begin with an inability to sense a full bladder and a failure to void completely. As bladder contractility worsens, bladder capacity and the postvoid residual increase, leading to symptoms of urinary hesitancy, decreased voiding frequency, incontinence, and recurrent urinary tract infections.

Erectile dysfunction and retrograde ejaculation are very common in DM and may be one of the earliest signs of diabetic neuropathy (Chap. 390). Erectile dysfunction, which increases in frequency with the age of the patient and the duration of diabetes, may occur in the absence of other signs of diabetic autonomic neuropathy.

TREATMENT

TREATMENT

Gastrointestinal/Genitourinary Dysfunction

Current treatments for these complications of DM are inadequate and nonspecific. Improved glycemic control should be a goal, but has not clearly shown benefit. Smaller, more frequent meals that are easier to digest (liquid) and low in fat and fiber may minimize symptoms of gastroparesis. Medications that slow gastric emptying (opioids, GLP-1-receptor agonists) should be avoided. Metoclopramide may be used with severe symptoms but is restricted to short-term treatment in both the United States and Europe. Gastric electrical stimulatory devices are available but not approved. Diabetic diarrhea in the absence of bacterial overgrowth is treated symptomatically (**Chap. 318**).

Diabetic cystopathy should be treated with scheduled voiding or self-catheterization. Drugs that inhibit type 5 phosphodiesterase are effective for erectile dysfunction, but their efficacy in individuals with DM is slightly lower than in the nondiabetic population (**Chap. 390**).

CARDIOVASCULAR MORBIDITY AND MORTALITY

CVD is increased in individuals with type 1 or type 2 DM. The Framingham Heart Study revealed a marked increase in PAD, coronary artery disease, MI, and CHF (risk increase from one- to fivefold) in DM. In addition, the prognosis for individuals with diabetes who have coronary artery disease or MI is worse than for nondiabetics. CHD is more likely to involve multiple vessels in individuals with DM. In addition to CHD, cerebrovascular disease is increased in individuals with DM (threefold increase in stroke). Thus, after controlling for all known cardiovascular risk factors, type 2 DM increases the cardiovascular death rate twofold in men and fourfold in women. Congestive heart failure (CHF) is common in long-standing T2DM.

The American Heart Association considers DM as a controllable risk factor for cardiovascular disease; in some studies, type 2 DM patients without a prior MI have a similar risk for coronary artery-related events as nondiabetic individuals who have had a prior MI. Cardiovascular risk assessment in type 2 DM should encompass a more nuanced approach. Cardiovascular risk is lower and not equivalent in a younger individual with a brief duration of type 2 DM compared to an older individual with long-standing type 2 DM. In individuals without a known diagnosis of diabetes, elevated A1C is predictive not just of diabetes risk, but also risk of CHD, stroke, and all-cause mortality. Because of the extremely high prevalence of underlying CVD in individuals with diabetes (especially in type 2 DM), evidence of atherosclerotic vascular disease (e.g., cardiac stress test) should be sought in an individual with diabetes who has symptoms, even if atypical, suggestive of cardiac ischemia or peripheral or carotid arterial disease. The screening of asymptomatic individuals with diabetes for CHD is not recommended or cost-effective. The absence of chest pain (“silent ischemia”) is common in individuals with diabetes, and a thorough cardiac evaluation should be considered prior to major surgical procedures.

The increase in cardiovascular morbidity and mortality rates in diabetes appears to relate to the synergism of hyperglycemia with other cardiovascular risk factors such as dyslipidemia (elevated triglycerides, low HDL-cholesterol and small-dense LDL), hypertension, obesity, reduced physical activity, and cigarette smoking. Additional risk factors prevalent include CKD (albuminuria, reduced GFR), abnormal platelet function, increased markers of inflammation, and endothelial dysfunction. The results of the ACCORD trial and VADT trial, which demonstrated that tight glucose control had limited benefit on cardiovascular outcomes in individuals with established cardiovascular disease, suggesting the importance of insulin resistance and dyslipidemia.

TREATMENT

TREATMENT

Cardiovascular Disease

Treatment of coronary disease in the diabetic individual has substantial overlap with treatment of in individuals without diabetes **(Chap. 267)**. Revascularization procedures for CHD, including percutaneous coronary interventions (PCIs) and coronary artery bypass grafting (CABG), may be less efficacious in the diabetic individual. Initial success rates of PCI in diabetic individuals are similar to those in the nondiabetic population, but diabetic patients have higher rates of restenosis and lower long-term patency and survival rates in older studies. CABG plus optimal medical management likely has better outcomes than PCI for individuals with diabetes.

Aggressive cardiovascular risk modification in all individuals with DM and glycemic control should be individualized, as discussed in **Chap. 397**. In patients with known CHD and type 2 DM, an ACE inhibitor (or ARB), a statin, and acetylsalicylic acid (ASA; aspirin) should be considered. Beta blockers can be used in individuals with diabetes after MI. In patients with CHF, thiazolidinediones should not be used **(Chap. 397)**. However, metformin can be used in patients with stable CHF if the renal function is normal. Some newer glucose lowering therapies also have cardiovascular benefit, including the GLP-1 analogs semaglutide (SUSTAIN-6) and liraglutide (LEADER), and the SGLT2 inhibitors empagliflozin (EMPA-REG) and canagliflozin (CANVAS).

Antiplatelet therapy reduces cardiovascular events in individuals with DM who have CHD and is recommended. The ADA recommends considering the use of aspirin for primary prevention of coronary events in individuals with diabetes with an increased cardiovascular risk (>50 years with at least one risk factor such as hypertension, dyslipidemia, smoking, family history, or albuminuria). ASA is not recommended for primary prevention in those with a low cardiovascular risk (<50 years with no risk factors). The aspirin dose is the same as in nondiabetic individuals.

Cardiovascular Risk Factors

DYSLIPIDEMIA

Individuals with DM may have several forms of dyslipidemia (**Chap. 400**). Because of the additive cardiovascular risk of hyperglycemia and hyperlipidemia, lipid abnormalities should be assessed aggressively and treated as part of comprehensive diabetes care (**Chap. 397**). The most common pattern of dyslipidemia is hypertriglyceridemia and reduced high-density lipoprotein (HDL) cholesterol levels. DM itself does not increase levels of low-density lipoprotein (LDL), but the small dense LDL particles found in type 2 DM are more atherogenic because they are more easily glycated and susceptible to oxidation.

Almost all treatment studies of diabetic dyslipidemia have been performed in individuals with type 2 DM because of the greater frequency of dyslipidemia in this form of diabetes. Interventional studies have shown that the beneficial effects of LDL reduction with statins are similar in the diabetic and nondiabetic populations. Large prospective trials of primary and secondary intervention for CHD have included some individuals with type 2 DM, and subset analyses have consistently found that reductions in LDL reduce cardiovascular events and morbidity in individuals with DM. No prospective studies have addressed similar questions in individuals with type 1 DM. Because the frequency of CVD is low in children and young adults with diabetes, assessment of cardiovascular risk should be incorporated into the guidelines discussed below.

Based on the guidelines provided by the ADA, all individuals with diabetes should be advised about lifestyle modification, including diet, weight loss, and increased physical activity (**Chap. 397**). If individuals with diabetes have elevated triglyceride levels (>1.7 mmol/L [150 mg/dL]) or low HDL cholesterol (<1 mmol/L [40 mg/dL]) in men and (<1.3 mmol/L [50 mg/dL]) in women, lifestyle modification and improved glycemic control should be further emphasized. If triglycerides remain >5.7 mmol/L (500 mg/dL), treatment with fish oil and fibrate drugs may reduce the risk of pancreatitis.

In terms of the addition of pharmacologic therapy, the ADA recommends: (1) all patients with diabetes and atherosclerotic cardiovascular disease should receive high-intensity statin therapy; (2) in patients aged 40–75 years, consider using moderate-intensity statin therapy (without additional risk factors) intensity statin therapy (with additional risk factors); (3) in patients <40 years and additional risk factors, consider moderate-intensity statin therapy. The ADA recommendations for individuals with diabetes who are >75 years are similar to that for individuals aged 40–75 years. Combination therapy with a statin and a fibrate or niacin is not recommended with the exception of a statin and ezetimibe and a statin in patients with recent acute coronary syndrome. The choice of statin and dosing should be individualized based on response and side effects. If the LDL cholesterol remains > 70 mg/dL in an individual with diabetes and CVD on a statin, consider the addition of ezetimibe or PCSDK9 inhibitor (**Chap. 400**). Statin usage is associated with a mild increase in the risk of developing type 2 DM. This risk is greatest in individuals with other risk factors for type 2 DM (**Chap. 396**). However, the cardiovascular benefits of statin use outweigh the mildly increased risk of diabetes.

HYPERTENSION

Hypertension can accelerate other complications of DM, particularly CVD, nephropathy, and retinopathy. Blood pressure should be measured at every clinic visit. In targeting a goal of blood pressure of <140/90 mmHg, therapy should first emphasize lifestyle modifications such as weight loss, exercise, stress management, and sodium restriction. The BP goal should be individualized. In some younger individuals or

those with increased CV risk, the provider may target a blood pressure of <130/80 mmHg. Realizing that more than one agent is usually required to reach the blood pressure goal, the ADA recommends that all patients with diabetes and hypertension be treated with an ACE inhibitor or an ARB initially. Subsequently, agents that reduce cardiovascular risk (beta blockers, thiazide diuretics, and calcium channel blockers) should be incorporated into the regimen. ACE inhibitors and ARBs are likely equivalent in most patients with diabetes and renal disease, but should not be combined. Serum potassium and renal function should be monitored.

Because of the high prevalence of atherosclerotic disease in individuals with type 2 DM, the possibility of renovascular hypertension should be considered when the blood pressure is not readily controlled.

LOWER EXTREMITY COMPLICATIONS

DM is the leading cause of nontraumatic lower extremity amputation in the United States. Foot ulcers and infections are also a major source of morbidity in individuals with DM. The reasons for the increased incidence of these disorders in DM involve the interaction of several pathogenic factors: neuropathy, abnormal foot biomechanics, PAD, and poor wound healing. The peripheral sensory neuropathy interferes with normal protective mechanisms and allows the patient to sustain major or repeated minor trauma to the foot, often without knowledge of the injury. Disordered proprioception causes abnormal weight bearing while walking and subsequent formation of callus or ulceration. Motor and sensory neuropathy lead to abnormal foot muscle mechanics and to structural changes in the foot (hammer toe, claw toe deformity, prominent metatarsal heads, Charcot joint). Autonomic neuropathy results in anhidrosis and altered superficial blood flow in the foot, which promote drying of the skin and fissure formation. PAD and poor wound healing impede resolution of minor breaks in the skin, allowing them to enlarge and to become infected.

Many individuals with type 2 DM develop a foot ulcer (great toe or metatarsophalangeal areas are most common), and a significant subset who develop an ulceration will ultimately undergo amputation (14–24% risk with that ulcer or subsequent ulceration). Risk factors for foot ulcers or amputation include male sex, diabetes for >10 years, peripheral neuropathy, abnormal structure of foot (bony abnormalities, callus, thickened nails), PAD, smoking, history of previous ulcer or amputation, visual impairment, poor glycemic control, and diabetic nephropathy, especially dialysis. Large calluses are often precursors to or overlie ulcerations.

TREATMENT

TREATMENT

Lower Extremity Complications

The optimal therapy for foot ulcers and amputations is prevention through identification of high-risk patients, education of the patient, and institution of measures to prevent ulceration. High-risk patients should be identified during the routine, annual foot examination performed on all patients with DM (see

"Ongoing Aspects of Comprehensive Diabetes Care" in Chap. 397). If the monofilament test or one of the other tests is abnormal, the patient is diagnosed with LOPS (Chap. 396). Providers should consider screening for asymptomatic PAD in individuals >50 years of age who have diabetes and other risk factors using ankle-brachial index testing in high-risk individuals (Chap. 275). Patient education should emphasize (1) careful selection of footwear, (2) daily inspection of the feet to detect early signs of poor-fitting footwear or minor trauma, (3) daily foot hygiene to keep the skin clean and moist, (4) avoidance of self-treatment of foot abnormalities and high-risk behavior (e.g., walking barefoot), and (5) prompt consultation with a health-care provider if an abnormality arises. Patients at high risk for ulceration or amputation may benefit from evaluation by a foot care specialist. Calluses and nail deformities should be treated by a podiatrist. Interventions directed at risk factor modification include orthotic shoes and devices, callus management, nail care, and prophylactic measures to reduce increased skin pressure from abnormal bony architecture. Attention to other risk factors for vascular disease (smoking, dyslipidemia, hypertension) and improved glycemic control are also important.

Despite preventive measures, foot ulceration and infection are common and represent a serious problem. Due to the multifactorial pathogenesis of lower extremity ulcers, management of these lesions is multidisciplinary and often demands expertise in orthopedics, vascular surgery, endocrinology, podiatry, and infectious diseases. The plantar surface of the foot is the most common site of ulceration. Ulcers may be primarily neuropathic (no accompanying infection) or may have surrounding cellulitis or osteomyelitis. Cellulitis without ulceration should be treated with antibiotics that provide broad-spectrum coverage, including anaerobes (see below).

An infected ulcer is a clinical diagnosis, because superficial culture of any ulceration will likely find multiple bacterial species of unknown significance. The infection surrounding the foot ulcer is often the result of multiple organisms, with aerobic gram-positive cocci (staphylococci including MRSA, Group A and B streptococci) being most common and with aerobic gram-negative bacilli and/or obligate anaerobes as co-pathogens.

Gas gangrene may develop in the absence of clostridial infection. Cultures should be obtained from the debrided ulcer base or from purulent drainage or aspiration of the wound. Wound depth should be determined by inspection and probing with a blunt-tipped sterile instrument. A wound that probes to the bone represents clinical evidence of osteomyelitis. Plain radiographs of the foot should be performed to assess the possibility of osteomyelitis in chronic ulcers that have not responded to therapy. Magnetic resonance imaging (MRI) is the most specific modality, with nuclear medicine scans and labeled white cell studies as alternatives. Surgical debridement is often necessary.

Osteomyelitis is best treated by a combination of prolonged antibiotics and debridement of infected bone when possible. The possible contribution of vascular insufficiency should be considered in all patients. Peripheral arterial bypass procedures are often effective in promoting wound healing and in decreasing the need for amputation of the ischemic limb (Chap. 275).

Interventions with demonstrated efficacy in diabetic foot ulcers or wounds: (1) off-loading, (2) debridement, (3) wound dressings, (4) appropriate use of antibiotics, (5) revascularization, and (6) limited amputation. Off-loading is the complete avoidance of weight bearing on the ulcer, which removes the mechanical trauma that retards wound healing. Bed rest and a variety of orthotic devices or contact casting limit weight bearing on wounds or pressure points. Surgical debridement is important and effective, but the efficacy of other modalities for wound healing (enzymes, growth factors, cellular therapy, hyperbaric oxygen) is unclear. Dressings such as hydrocolloid dressings promote wound healing by creating a moist environment, controlling the exudate, and protecting the wound. Antiseptic agents should be avoided. Topical antibiotics are of limited value. Referral for physical therapy, orthotic evaluation, and rehabilitation should occur once the infection is controlled.

Mild or nonlimb-threatening infections can be treated with oral antibiotics directed predominantly at methicillin-susceptible staphylococci and streptococci (e.g., dicloxacillin, cephalosporin, amoxicillin/clavulanate). However, in patients with a prior history of MRSA or in locations with a high prevalence of MRSA, treatment with clindamycin, doxycycline, or trimethoprim-sulfamethoxazole is preferred. Trimethoprim-sulfamethoxazole exhibits less reliable coverage of streptococci than the β-lactams, and individuals with diabetes may develop adverse effects including acute kidney injury and hyperkalemia. Surgical debridement of necrotic tissue, local wound care (avoidance of weight bearing over the ulcer), and close surveillance for progression of infection are crucial. More severe infections require IV antibiotics as well as bed rest and local wound care. Urgent surgical debridement may be required. Optimization of glycemic control should be a goal. IV antibiotics should provide broad-spectrum coverage directed toward *Staphylococcus aureus*, including MRSA, streptococci, gram-negative aerobes, and anaerobic bacteria. Initial antimicrobial regimens include vancomycin plus a β-lactam/β-lactamase inhibitor or carbapenem or vancomycin plus a combination of a quinolone plus metronidazole. Daptomycin, ceftaroline, or linezolid may be substituted for vancomycin. If the infection surrounding the ulcer is not improving with IV antibiotics, reassessment of antibiotic coverage and reconsideration of the need for surgical debridement or revascularization are indicated. With clinical improvement, oral antibiotics and local wound care can be continued on an outpatient basis with close follow-up.

INFECTIONS

Individuals with DM have a greater frequency and severity of infection. The reasons for this include incompletely defined abnormalities in cell-mediated immunity and phagocyte function associated with hyperglycemia, as well as diminished vascularization. Hyperglycemia aids the colonization and growth of a variety of organisms (*Candida* and other fungal species). Many common infections are more frequent and severe in the diabetic population, whereas several rare infections are seen almost exclusively in the diabetic population. Examples of this latter category include rhinocerebral mucormycosis, emphysematous infections of the gallbladder and urinary tract, and “malignant” or invasive otitis externa. Invasive otitis externa is usually secondary to *Pseudomonas aeruginosa* infection in the soft tissue surrounding the external auditory canal, usually begins with pain and discharge, and may rapidly progress to osteomyelitis and meningitis.

These infections should be sought, in particular, in patients presenting with severe hyperglycemia (**Chap. 397**).

Pneumonia, urinary tract infections, and skin and soft tissue infections are all more common in the diabetic population. In general, the organisms that cause pulmonary infections are similar to those found in the nondiabetic population; however, gram-negative organisms, *S. aureus*, and *Mycobacterium tuberculosis* are more frequent pathogens. Urinary tract infections (either lower tract or pyelonephritis) are the result of common bacterial agents such as *Escherichia coli*, although several yeast species (*Candida albicans* and *Torulopsis glabrata*) are commonly observed. Complications of urinary tract infections include emphysematous pyelonephritis and emphysematous cystitis. Bacteriuria occurs frequently in individuals with diabetic cystopathy and does not require antibiotic therapy. Susceptibility to furunculosis, superficial candidal infections, and vulvovaginitis are increased. Poor glycemic control is a common denominator in individuals with these infections. Individuals with diabetes have an increased rate of colonization of *S. aureus* in the skinfolds and nares. Individuals with diabetes also have a greater risk of postoperative wound infections.

DERMATOLOGIC MANIFESTATIONS

The most common skin manifestations of DM are xerosis and pruritus and are usually relieved by skin moisturizers. Protracted wound healing and skin ulcerations are also frequent complications. Diabetic dermopathy, sometimes termed *pigmented pretibial papules*, or “diabetic skin spots,” begins as an erythematous macule or papule that evolves into an area of circular hyperpigmentation. These lesions result from minor mechanical trauma in the pretibial region and are more common in elderly men with DM. Bullous diseases, such as bullosa diabeticorum (shallow ulcerations or erosions in the pretibial region), are also seen. *Necrobiosis lipoidica diabeticorum* is an uncommon disorder, accompanying diabetes in predominantly young women. This usually begins in the pretibial region as an erythematous plaque or papules that gradually enlarge, darken, and develop irregular margins, with atrophic centers and central ulceration. They are often painful. Vitiligo occurs at increased frequency in individuals with type 1 DM. *Acanthosis nigricans* (hyperpigmented velvety plaques seen on the neck, axilla, or extensor surfaces) is sometimes a feature of severe insulin resistance and accompanying diabetes. Generalized or localized *granuloma annulare* (erythematous plaques on the extremities or trunk), *lichen planus* (violaceous papules on the cutaneous surface +/– erosions in the mouth and genitalia), and *scleredema* (areas of skin thickening on the back or neck at the site of previous superficial infections) are more common in the diabetic population. *Lipoatrophy* and *lipohypertrophy* can occur at insulin injection sites but are now unusual with the use of human insulin, and avoided by rotating injection sites.

FURTHER READING

American Diabetes Association: Cardiovascular disease and risk management: Diabetes Care 41:S86, 2018. [PubMed: 29222380]

Fox CS et al: Update on prevention of cardiovascular disease in adults with type 2 diabetes mellitus in light of recent evidence: A scientific statement from the American Heart Association and the American Diabetes Association. Diabetes Care 38:1777, 2015. [PubMed: 26246459]

Hingorani A et al: The management of diabetic foot: A clinical practice guideline by the Society for Vascular Surgery in collaboration with the American Podiatric Medical Association and the Society for Vascular Medicine. J Vasc Surg 63:3S, 2016. [PubMed: 26804367]

Holman RR et al: 10-Year follow-up of intensive glucose control in type 2 diabetes. N Engl J Med 359:1577, 2008. [PubMed: 18784090]

Microvascular Complications and Foot Care. Diabetes Care 41:105, 2018.

Nathan DM et al: Diabetes control and complications trial/epidemiology of diabetes interventions and complications study at 30 years: Advances and contributions. Diabetes 62:3976, 2013. [PubMed: 24264395]

Pop-Busui R et al: Diabetic neuropathy: A position statement by the American Diabetes Association. Diabetes Care 40:136, 2017. [PubMed: 27999003]

Solomon SD et al: Diabetic retinopathy: A position statement by the American Diabetes Association. Diabetes Care 40:412, 2017. [PubMed: 28223445]

Chapter 399: Hypoglycemia

Philip E. Cryer; Stephen N. Davis

INTRODUCTION

Hypoglycemia is most commonly caused by drugs used to treat diabetes mellitus or by exposure to other drugs, including alcohol. However, a number of other disorders, including critical organ failure, sepsis and inanition, hormone deficiencies, non-β-cell tumors, insulinoma, and prior gastric surgery, can cause hypoglycemia **(Table 399-1)**. Hypoglycemia may be documented by *Whipple's triad:* (1) symptoms consistent with hypoglycemia, (2) a low plasma glucose concentration measured with a precise method, and (3) relief of symptoms after the plasma glucose level is raised. The lower limit of the fasting plasma glucose concentration is normally ~70 mg/dL (~3.9 mmol/L), but lower venous glucose levels occur normally, late after a meal, during pregnancy, and during prolonged fasting (>24 h). Hypoglycemia can cause serious morbidity; if severe it can be fatal. It should be considered in any patient with episodes of confusion, an altered level of consciousness, or a seizure.

TABLE 399-1

Causes of Hypoglycemia in Adults

Ill or Medicated Individual
1. Drugs Insulin or insulin secretagogue Alcohol Others 2. Critical illness Hepatic, renal or cardiac failure Sepsis Inanition 3. Hormone deficiency Cortisol Growth hormone Glucagon and epinephrine (in insulin-deficient diabetes) 4. Non–islet cell tumor (e.g., Mesenchymal tumors)
Seemingly Well Individual
5. Endogenous hyperinsulinism Insulinoma Functional β-cell disorders (nesidioblastosis) Noninsulinoma pancreatogenous hypoglycemia Post–gastric bypass hypoglycemia Insulin autoimmune hypoglycemia Antibody to insulin Antibody to insulin receptor Insulin secretagogue Other 6. Disorders of Gluconeogenesis and fatty acid oxidation. 7. Exercise 8. Accidental, surreptitious, or malicious hypoglycemia

Source: Adapted from PE Cryer et al: J Clin Endocrinol Metab 94:709, 2009. ©The Endocrine Society, 2009.

SYSTEMIC GLUCOSE BALANCE AND GLUCOSE COUNTERREGULATION

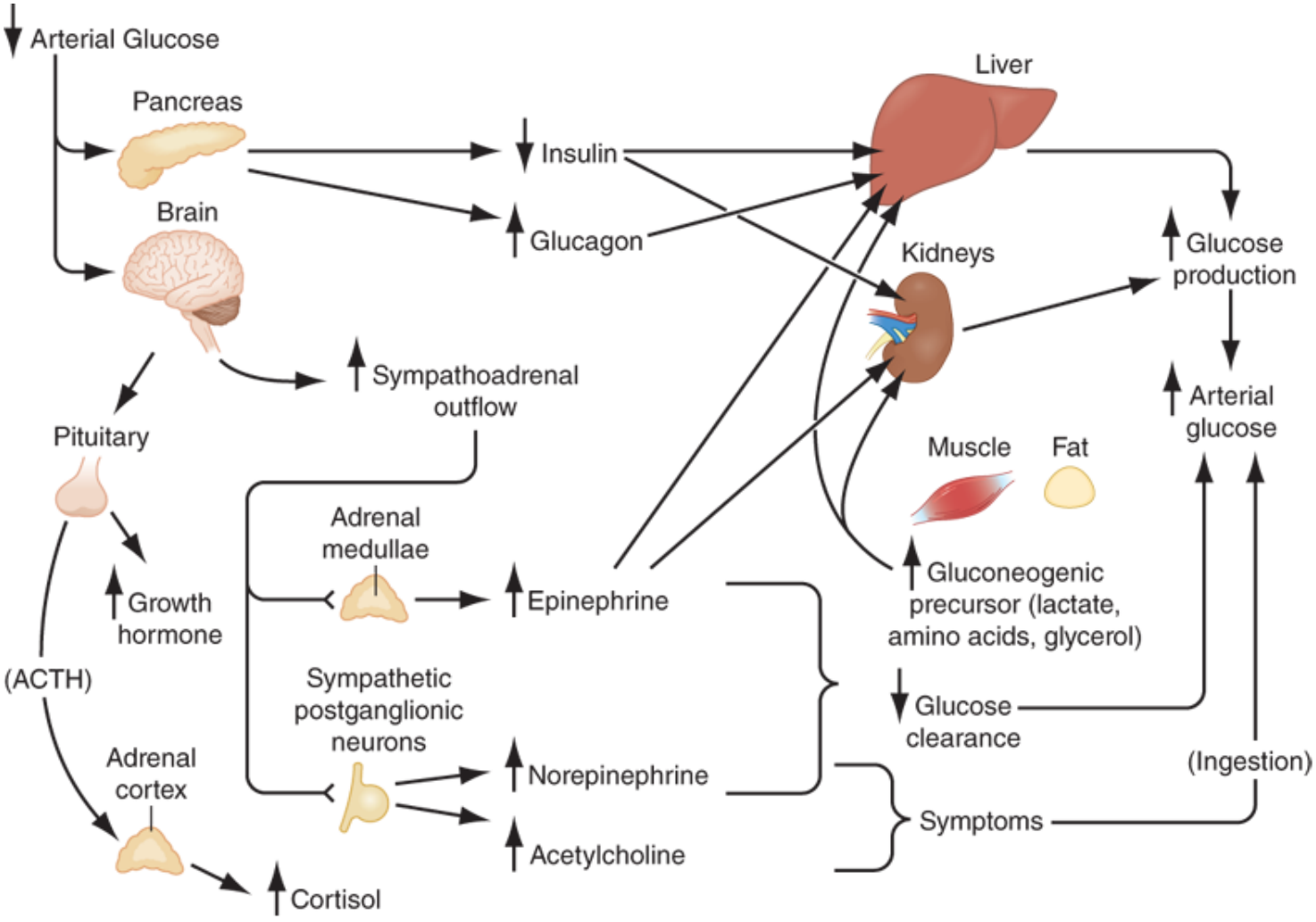
Glucose is an obligate metabolic fuel for the brain under physiologic conditions. The brain cannot synthesize glucose or store more than a few minutes' supply as glycogen and therefore requires a continuous supply of glucose from the arterial circulation. As the arterial plasma glucose concentration falls below the physiologic range, blood-to-brain

glucose transport becomes insufficient to support brain energy metabolism and function. However, multiple integrated glucose counterregulatory mechanisms normally prevent or rapidly correct hypoglycemia.

Plasma glucose concentrations are normally maintained within a relatively narrow range—roughly 70–110 mg/dL (3.9–6.1 mmol/L) in the fasting state, with transient higher excursions after a meal—despite wide variations in exogenous glucose delivery from meals and in endogenous glucose utilization by, for example, exercising muscle. Between meals and during fasting, plasma glucose levels are maintained by endogenous glucose production, hepatic glycogenolysis, and hepatic (and renal) gluconeogenesis **(Fig. 399-1)**. Although hepatic glycogen stores are usually sufficient to maintain plasma glucose levels for ~8 h, this period can be shorter if glucose demand is increased by exercise or if glycogen stores are depleted by illness or starvation.

FIGURE 399-1

Physiology of glucose counterregulation: Mechanisms that normally prevent or rapidly correct hypoglycemia. In insulin-deficient diabetes, the key counterregulatory responses–suppression of insulin and increases in glucagon–are lost, and stimulation of sympathoadrenal outflow is attenuated. ACTH, adrenocorticotropic hormone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Gluconeogenesis normally requires low insulin levels and the presence of anti-insulin (counterregulatory) hormones together with a coordinated supply of precursors from muscle and adipose tissue to the liver (and kidneys). Muscle provides lactate, pyruvate, alanine, glutamine, and other amino acids. Triglycerides in adipose tissue are broken

down into fatty acids and glycerol, which is a gluconeogenic precursor. Fatty acids provide an alternative oxidative fuel to tissues other than the brain (which requires glucose).

Systemic glucose balance, maintenance of the normal plasma glucose concentration, is accomplished by a network of hormones, neural signals, and substrate effects that regulate endogenous glucose production and glucose utilization by tissues other than the brain **(Chap. 396)**. Among the regulatory factors, insulin plays a dominant role (**Table 399-2**; Fig. 399-1). As plasma glucose levels decline within the physiologic range, pancreatic β-cell insulin secretion decreases, thereby increasing hepatic glycogenolysis and hepatic (and renal) gluconeogenesis. Low insulin levels also reduce glucose utilization in peripheral tissues, inducing lipolysis and proteolysis and consequently releasing gluconeogenic precursors. Thus, a decrease in insulin secretion is the first defense against hypoglycemia.

TABLE 399-2

Physiologic Responses to Decreasing Plasma Glucose Concentrations

Response	Glycemic Threshold, mmol/L (mg/dL)	Physiologic ↓ Effects	Role in Prevention or Correction of Hypoglycemia (Glucose Counterregulation)
↓ Insulin	4.4–4.7 (80–85)	↑ R_a (↓ R_d), increased lipolysis; ↑ FFA ↑ Glycerol	Primary glucose regulatory factor/first defense against hypoglycemia
↑ Glucagon	3.6–3.9 (65–70)	↑ R_a	Primary glucose counterregulatory factor/second defense against hypoglycemia
↑ Epinephrine	3.6–3.9 (65–70)	↑ R_a, ↓ R_c, increased lipolysis; ↑ FFA & Glycerol	Third defense against hypoglycemia, critical when glucagon is deficient
↑ Cortisol and growth hormone	3.6–3.9 (65–70)	↑ R_a, ↓ R_c	Involved in defense against prolonged hypoglycemia; not critical
Symptoms	2.8–3.1 (50–55)	Recognition of hypoglycemia	Prompt behavioral defense against hypoglycemia (food ingestion)
↓ Cognition	<2.8 (<50)	—	Compromises behavioral defense against hypoglycemia

Note: R_a, rate of glucose appearance, glucose production by the liver and kidneys; R_c, rate of glucose clearance, glucose utilization relative to the ambient plasma glucose by insulin-sensitive tissues; R_d, rate of glucose disappearance, glucose utilization by insulin-sensitive tissues such as skeletal muscle. R_d by the brain is not altered by insulin, glucagon, epinephrine, cortisol, or growth hormone. FFA, free fatty acids.

Source: From PE Cryer, in S Melmed et al (eds): *Williams Textbook of Endocrinology,* 12th ed. New York, Elsevier, 2012.

As plasma glucose levels decline just below the physiologic range, glucose counterregulatory (plasma glucose–raising) hormones are released (Table 399-2; Fig. 399-1). Among these, pancreatic α-cell glucagon, and adrenomedullary epinephrine play a primary role. Glucagon stimulates hepatic glycogenolysis and gluconeogenesis. Adrenomedullary epinephrine also stimulates hepatic glycogenolysis and gluconeogenesis (and renal gluconeogenesis), but limits peripheral uptake of glucose and stimulates lipolysis with production of glycerol and fatty acids. Epinephrine becomes critical when glucagon is deficient. When hypoglycemia is prolonged beyond ~4 h, cortisol and growth hormone also support glucose production and restrict glucose utilization to a limited amount (~20% compared to epinephrine). Thus cortisol and growth hormone play no role in defense against acute hypoglycemia.

As plasma glucose levels fall further, symptoms prompt behavioral defense against hypoglycemia, including the ingestion of food (Table 399-2; Fig. 399-1). The normal glycemic thresholds for these responses to decreasing plasma glucose concentrations are shown in Table 399-2. However, these thresholds are dynamic. They shift to higher-than-normal glucose levels in people with poorly controlled diabetes, who can experience symptoms of hypoglycemia when their glucose levels decline toward the normal range. On the other hand, thresholds shift to lower-than-normal glucose levels in people with recurrent hypoglycemia; i.e., patients with intensively treated diabetes or an insulinoma have symptoms at glucose levels lower than those that cause symptoms in healthy individuals.

Clinical Manifestations

Neuroglycopenic manifestations of hypoglycemia are the direct result of central nervous system glucose deprivation. These features include behavioral changes, confusion, fatigue, seizure, loss of consciousness, cardiac arrhythmias, and, if hypoglycemia is severe, death. Neurogenic (or autonomic) manifestations of hypoglycemia result from the perception of physiologic changes caused by the central nervous system–mediated sympathoadrenal discharge that is triggered by hypoglycemia. They include *adrenergic* symptoms (mediated largely by norepinephrine released from sympathetic postganglionic neurons but perhaps also by epinephrine released from the adrenal medullae), such as palpitations, tremor, and anxiety, as well as *cholinergic* symptoms (mediated by acetylcholine released from sympathetic postganglionic neurons), such as sweating, hunger, and paresthesiae. Clearly, these are nonspecific symptoms. Their attribution to hypoglycemia requires that the corresponding plasma glucose concentration be low and that the symptoms resolve after the glucose level is raised (as delineated by Whipple's triad).

Common signs of hypoglycemia include diaphoresis and pallor. Heart rate and systolic blood pressure are typically increased, but may not be raised in an individual who has experienced repeated, recent episodes of hypoglycemia. Neuroglycopenic manifestations are often observable. Transient focal neurologic deficits occur occasionally. Permanent neurologic deficits are rare.

Etiology and Pathophysiology

Hypoglycemia activates pro-inflammatory, pro-coagulant and pro-atherothrombotic responses in T1DM, T2DM, and non-diabetic individuals. These responses increase platelet aggregation, reduce fibrinolytic balance (↑ plasminogen activator inhibitor-1), and increase intravascular coagulation. Hypoglycemia also reduces protective nitric oxide-mediated arterial vasodilator mechanisms in healthy, T1DM, and T2DM individuals.

HYPOGLYCEMIA IN DIABETES

Impact and Frequency

Hypoglycemia is the limiting factor in the glycemic management of diabetes mellitus. First, it causes recurrent morbidity in most people with type 1 diabetes (T1DM) and in many with advanced type 2 diabetes (T2DM), and it is sometimes fatal. Second, it precludes maintenance of euglycemia over a lifetime of diabetes and thus full realization of the well-established microvascular benefits of glycemic control. Third, it causes a vicious cycle of recurrent hypoglycemia by producing hypoglycemia-associated autonomic failure—i.e., the clinical syndromes of defective glucose counterregulation and of hypoglycemia unawareness.

Hypoglycemia is a fact of life for people with T1DM if treated with insulin, sulfonylurea, or glinides. They suffer an average of two episodes of symptomatic hypoglycemia per week and at least one episode of severe, at least temporarily disabling hypoglycemia each year. An estimated 6–10% of people with T1DM die as a result of hypoglycemia. The incidence of hypoglycemia is lower in T2DM than in T1DM. However, its prevalence in insulin-requiring T2DM is surprisingly high. Recent studies have revealed a hypoglycemia prevalence approaching 70%. In fact, as patients with T2DM outnumber those with T1DM by ten- to twentyfold, the prevalence of hypoglycemia is now greater in T2DM. Insulin, sulfonylureas, or glinides can cause hypoglycemia in T2DM. Metformin, thiazolidinediones, α-glucosidase inhibitors, glucagon-like peptide 1 (GLP-1) receptor agonists, and dipeptidyl peptidase IV (DPP-IV) inhibitors do not cause hypoglycemia. However, they increase the risk when combined with a sulfonylurea, glinide, or insulin. Notably, the frequency of hypoglycemia approaches that in T1DM as persons with T2DM develop absolute insulin deficiency and require more complex treatment with insulin.

Conventional Risk Factors

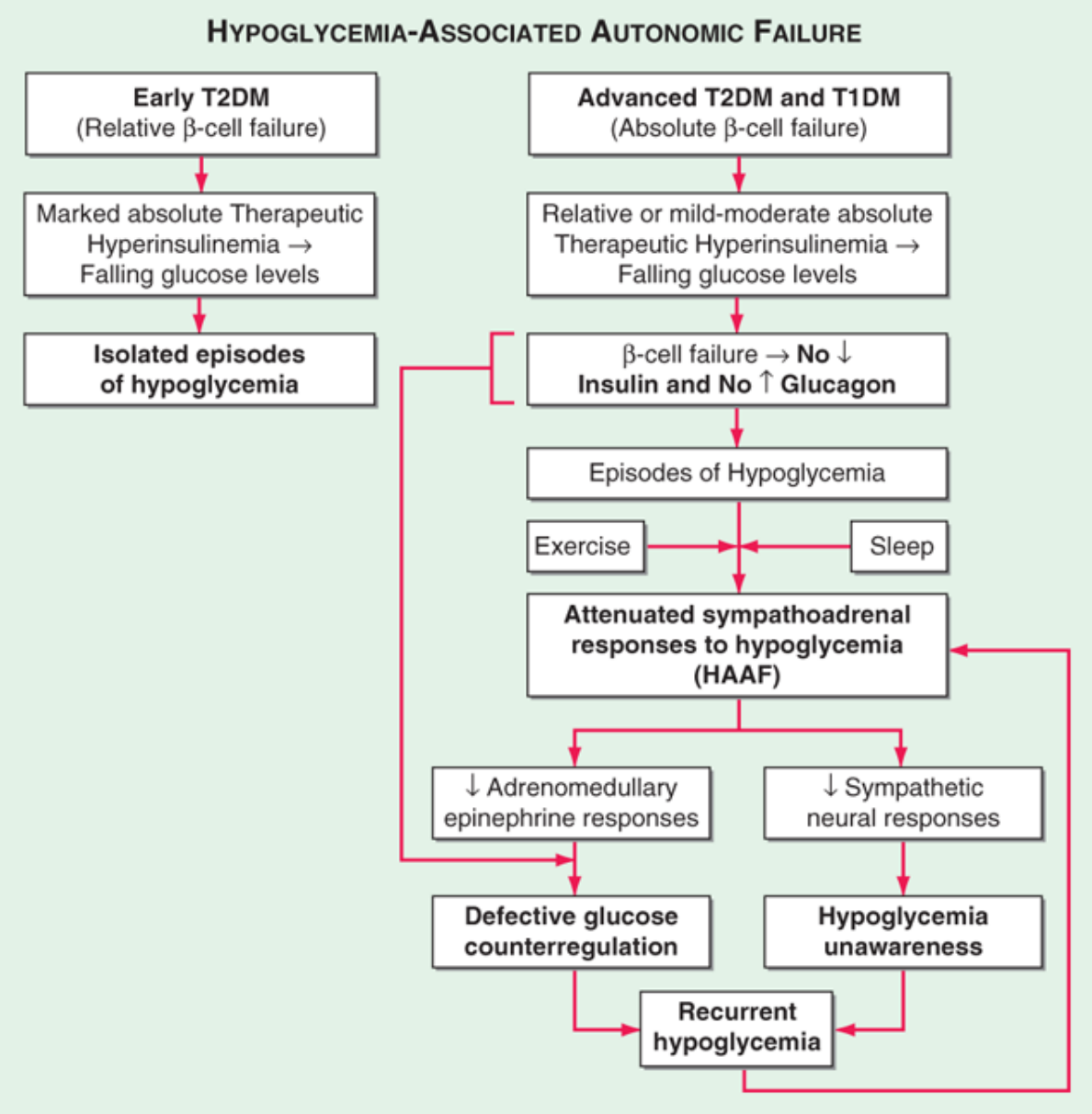
The conventional risk factors for hypoglycemia in diabetes are identified on the basis of the premise that relative or absolute insulin excess is the sole determinant of risk. Relative or absolute insulin excess occurs when (1) insulin (or insulin secretagogue) doses are excessive, ill-timed, or of the wrong type; (2) the influx of exogenous glucose is reduced (e.g., during an overnight fast, periods of temporary fasting, or after missed meals or snacks); (3) insulin-independent glucose utilization is increased (e.g., during exercise); (4) sensitivity to insulin is increased (e.g., with improved glycemic control, in the middle of the night, late after exercise, or with increased fitness or weight loss); (5) endogenous glucose production is reduced (e.g., after alcohol ingestion); and (6) insulin clearance is reduced (e.g., in renal failure). However, these conventional risk factors alone explain a minority of episodes; other factors are typically involved.

Hypoglycemia-Associated Autonomic Failure (HAAF)

While marked insulin excess alone can cause hypoglycemia, iatrogenic hypoglycemia in diabetes (either T1DM and/or T2DM) is typically the result of the interplay of relative or absolute therapeutic insulin excess and compromised physiologic and behavioral defenses against falling plasma glucose concentrations (Table 399-2; **Fig. 399-2**). Defective glucose counterregulation compromises physiologic defense (particularly decrements in insulin and increments in glucagon and epinephrine), and hypoglycemia unawareness compromises behavioral defense (ingestion of carbohydrate).

FIGURE 399-2

Hypoglycemia-associated autonomic failure (HAAF) in insulin-deficient diabetes. T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus. *(Modified from PE Cryer: Hypoglycemia in Diabetes. Pathophysiology, Prevalence, and Prevention,* 2nd *ed.* © *American Diabetes Association, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DEFECTIVE GLUCOSE COUNTERREGULATION

In the setting of absolute endogenous insulin deficiency, insulin levels do not decrease as plasma glucose levels fall; thus the first defense against hypoglycemia is lost. After a few years disease duration in T1DM, glucagon levels do not increase as plasma glucose levels fall; a second defense against hypoglycemia is lost. Reduced glucagon responses to hypoglycemia also occur in long duration T2DM. However, pancreatic alpha cells that produce glucagon are present in the same number and size in T1DM as compared to age matched non-diabetic individuals. Thus, the defect that restricts glucagon release during hypoglycemia in T1DM (and presumably in long-standing T2DM), appears to be a signaling defect, as glucagon responses to other physiologic stress in T1DM (e.g., exercise) are preserved. Finally, the increase in epinephrine levels, the third critical defense against acute hypoglycemia, is typically attenuated. The glycemic threshold for the sympathoadrenal (adrenomedullary epinephrine and sympathetic neural norepinephrine) response is shifted to lower plasma glucose concentrations. That shift is typically the result of recent antecedent iatrogenic hypoglycemia. In the setting of absent decrements in insulin and of absent increments in glucagon, the

attenuated increment in epinephrine causes the clinical syndrome of defective glucose counterregulation. Affected patients are at ≥25-fold greater risk of severe iatrogenic hypoglycemia during intensive glycemic therapy for their diabetes than are patients with normal epinephrine responses. This functional—and potentially reversible—disorder is distinct from classic diabetic autonomic neuropathy—a structural and irreversible disorder.

HYPOGLYCEMIA UNAWARENESS

The attenuated sympathoadrenal response (largely the reduced sympathetic neural response) to hypoglycemia causes the clinical syndrome of *hypoglycemia unawareness*—i.e., loss of the warning adrenergic and cholinergic symptoms that previously allowed the patient to recognize developing hypoglycemia and therefore to abort the episode by ingesting carbohydrates. Affected patients are at a sixfold increased risk of severe iatrogenic hypoglycemia during intensive glycemic therapy of their diabetes.

HAAF IN DIABETES

The concept of HAAF in diabetes posits that recent antecedent iatrogenic hypoglycemia (or sleep or prior exercise) causes both defective glucose counterregulation (by reducing the epinephrine response to a given level of subsequent hypoglycemia in the setting of absent insulin and glucagon responses) and hypoglycemia unawareness (by reducing the sympathoadrenal response to a given level of subsequent hypoglycemia). These impaired responses create a vicious cycle of recurrent iatrogenic hypoglycemia (Fig. 399-2). Hypoglycemia unawareness and, to some extent, the reduced epinephrine component of defective glucose counterregulation are reversible by as little as 2–3 weeks of scrupulous avoidance of hypoglycemia in most affected patients.

On the basis of this pathophysiology, additional risk factors for hypoglycemia in diabetes include (1) absolute insulin deficiency, indicating that insulin levels will not decrease and glucagon levels will not increase as plasma glucose levels fall; (2) a history of severe hypoglycemia or of hypoglycemia unawareness, implying recent antecedent hypoglycemia, as well as prior exercise or sleep, indicating that the sympathoadrenal response will be attenuated; (3) impaired renal function resulting in reduced clearance of exogenous and endogenous insulin; (4) classical diabetic autonomic neuropathy, and (5) lower hemoglobin A_{1C} (HBA_{1C}), or lower glycemic goals even at elevated HBA_{1C} levels (8–10%), as they represent an increased probability of recent antecedent hypoglycemia.

Hypoglycemia Risk Factor Reduction

Several recent multicenter, randomized, controlled trials investigating the potential benefits of tight glucose control in either inpatient or outpatient settings have reported a high prevalence of severe hypoglycemia. In the NICE-SUGAR study, attempts to control in-hospital plasma glucose values towards physiologic levels resulted in increased mortality risk. The ADVANCE and ACCORD studies and the Veterans Affairs Diabetes Trial (VADT) also found a significant incidence of severe hypoglycemia among T2DM patients. Severe hypoglycemia with accompanying serious cardiovascular morbidity and mortality also occurred in the standard (e.g., not receiving intensified treatment) control group in both the ACCORD study and the VADT. Thus, severe hypoglycemia can and does occur at HbA_{1c} values of 8–10% in both T1DM and T2DM. Somewhat surprisingly, all three studies found little or no benefit of intensive glucose control to reduce macrovascular events in T2DM. In fact, the ACCORD study was ended early because of the increased mortality rate in the intensive glucose control arm. Whether iatrogenic hypoglycemia was the cause of the increased mortality risk is not known. In light of these findings, some new recommendations and

paradigms have been formulated. Whereas there is little debate regarding the need to reduce hyperglycemia in the hospital, the glycemic maintenance goals have been modified to lie between 140 and 180 mg/dL. Accordingly, the benefits of insulin therapy and reduced hyperglycemia can be obtained while the prevalence of hypoglycemia is reduced.

Similarly, evidence exists that intensive glucose control can reduce the prevalence of microvascular disease in both T1DM and T2DM. These benefits need to be weighed against the increased prevalence of hypoglycemia. Certainly, the level of glucose control (i.e., the HbA_{1c} level) should be evaluated for each patient. Multicenter trials have demonstrated that individuals with recently diagnosed T1DM or T2DM can have better glycemic control with less hypoglycemia. In addition, there is still long-term benefit in reducing HbA_{1c} values from higher to lower, albeit still above recommended levels. Perhaps a reasonable therapeutic goal is the lowest HbA_{1c} level that does not cause severe hypoglycemia and that preserves awareness of hypoglycemia.

Pancreatic transplantation (both whole-organ and islet-cell) has been used in part as a treatment for severe hypoglycemia. Generally, rates of hypoglycemia are reduced after transplantation. This decrease appears to be due to increased physiologic insulin and glucagon responses during hypoglycemia.

The use of continuous glucose monitors,either alone or in combination with continuous subcutaneous infusion via a wearable pump, offers promise as a method of reducing hypoglycemia while improving HbA_{1c}. Specifically, continuous glucose monitoring coupled with temporary discontinuation of subcutaneous insulin infusion when the monitor predicts a low glucose concentration is particularly promising. Furthermore, ongoing progress utilizing a portable wearable “artificial pancreas” incorporating continuous glucose sensor modulation of either insulin alone or bi-hormonal delivery of both insulin and glucagon has been established. Additionally, stem cell-derived β-cells also offer promise of novel therapeutic interventions to reduce hypoglycemia.

Other interventions to stimulate counterregulatory responses, such as selective serotonin-reuptake inhibitors, β-adrenergic receptor antagonists, opiate receptor antagonists, and fructose, remain experimental and have not been assessed in large-scale clinical trials.

Thus, intensive glycemic therapy **(Chap. 397)** needs to be applied along with the patient’s education and empowerment, frequent self-monitoring of blood glucose, flexible insulin (and other drug) regimens (including the use of insulin analogues, both short- and longer-acting), individualized glycemic goals, and ongoing professional guidance, support, and consideration of both the conventional risk factors and those indicative of compromised glucose counterregulation. Given a history of hypoglycemia unawareness, a 2- to 3-week period of scrupulous avoidance of hypoglycemia is indicated.

HYPOGLYCEMIA WITHOUT DIABETES

There are many causes of hypoglycemia (Table 399-1). Because hypoglycemia is common in insulin- or insulin secretagogue–treated diabetes, it is often reasonable to assume that a clinically suspicious episode is the result of hypoglycemia. On the other hand, because hypoglycemia is rare in the absence of relevant drug-treated diabetes, it is reasonable to conclude that a hypoglycemic disorder is present only in patients in whom Whipple’s triad can be demonstrated.

Particularly when patients are ill or medicated, the initial diagnostic focus should be on the possibility of drug involvement and then on critical illnesses, hormone deficiency, or non–islet cell tumor hypoglycemia. In the absence of any of these etiologic factors and in a seemingly well individual, the focus should shift to possible endogenous hyperinsulinism or accidental, surreptitious, or even malicious hypoglycemia.

Drugs

Insulin and insulin secretagogues suppress glucose production and stimulate glucose utilization. Ethanol blocks gluconeogenesis but not glycogenolysis. Thus, alcohol-induced hypoglycemia typically occurs after a several-day ethanol binge during which the person eats little food, with consequent glycogen depletion. Ethanol is usually measurable in blood at the time of presentation, but its levels correlate poorly with plasma glucose concentrations. Because gluconeogenesis becomes the predominant route of glucose production during prolonged hypoglycemia, alcohol can contribute to the progression of hypoglycemia in patients with insulin-treated diabetes.

Many other drugs have been associated with hypoglycemia. These include commonly used drugs such as angiotensin-converting enzyme inhibitors and angiotensin receptor antagonists, β-adrenergic receptor antagonists, quinolone antibiotics, indomethacin, quinine, and sulfonamides.

Critical Illness

Among hospitalized patients, serious illnesses such as renal, hepatic, or cardiac failure; sepsis; and inanition are second only to drugs as causes of hypoglycemia.

Rapid and extensive hepatic destruction (e.g., toxic hepatitis) causes fasting hypoglycemia because the liver is the major site of endogenous glucose production. The mechanism of hypoglycemia in patients with cardiac failure is unknown. Hepatic congestion and hypoxia may be involved. Although the kidneys are a source of glucose production, hypoglycemia in patients with renal failure is also caused by the reduced clearance of insulin (thereby inappropriately increasing insulin relative to the prevailing glucose levels), and the reduced mobilization of gluconeogenic precursors in renal failure.

Sepsis is a relatively common cause of hypoglycemia. Increased glucose utilization is induced by cytokine production in macrophage-rich tissues such as the liver, spleen, and lung. Hypoglycemia develops if glucose production fails to keep pace. Cytokine-induced inhibition of gluconeogenesis in the setting of nutritional glycogen depletion, in combination with hepatic and renal hypoperfusion, may also contribute to hypoglycemia.

Hypoglycemia can be seen with starvation. Due to brain conversion and utilization of alternative substrates, such as lactate, pyruvate, and ketone bodies, there is only a modest counterregulatory neuroendocrine and autonomic nervous system response. During periods of prolonged starvation (fasting) plasma glucose levels are lower in women as compared to men; perhaps because of loss of whole-body fat stores and subsequent depletion of gluconeogenic precursors (e.g., amino acids), necessitating increased glucose utilization.

Hormone Deficiencies

Neither cortisol nor growth hormone is critical to the prevention of hypoglycemia, at least in adults. Nonetheless, hypoglycemia can occur with prolonged fasting in patients with primary adrenocortical failure (Addison’s disease) or hypopituitarism. Anorexia and weight loss are typical features of chronic cortisol deficiency and likely result in

glycogen depletion. Cortisol deficiency is associated with impaired gluconeogenesis and low levels of gluconeogenic precursors; these associations suggest that substrate-limited gluconeogenesis, in the setting of glycogen depletion, is the cause of hypoglycemia. Growth hormone deficiency can cause hypoglycemia in young children. In addition to extended fasting, high rates of glucose utilization (e.g., during exercise or in pregnancy) or low rates of glucose production (e.g., after alcohol ingestion) can precipitate hypoglycemia in adults with previously unrecognized hypopituitarism.

Hypoglycemia is not a feature of the epinephrine-deficient state that results from bilateral adrenalectomy when glucocorticoid replacement is adequate, nor does it occur during pharmacologic adrenergic blockade when other glucoregulatory systems are intact. Combined deficiencies of glucagon and epinephrine play a key role in the pathogenesis of iatrogenic hypoglycemia in people with insulin-deficient diabetes, as discussed earlier. Otherwise, deficiencies of these hormones are not usually considered in the differential diagnosis of a hypoglycemic disorder.

Non-β-Cell Tumors

Fasting hypoglycemia, often termed *non–islet cell tumor hypoglycemia*, occurs occasionally in patients with large mesenchymal or epithelial tumors (e.g., hepatomas, adrenocortical carcinomas, carcinoids). The glucose kinetic patterns resemble those of hyperinsulinism (see next), but insulin secretion is suppressed appropriately during hypoglycemia. In most instances, hypoglycemia is due to overproduction of an incompletely processed form of insulin-like growth factor II ("big IGF-II") that does not complex normally with circulating binding proteins and thus more readily gains access to target tissues. The tumors are usually apparent clinically, plasma ratios of IGF-II to IGF-I are high, and free IGF-II levels (and levels of pro-IGF-II [1–21]) are elevated. Curative surgery is seldom possible, but reduction of tumor bulk may ameliorate hypoglycemia. Therapy with a glucocorticoid, growth hormone, or both has also been reported to alleviate hypoglycemia. Hypoglycemia attributed to ectopic IGF-I production has been reported but is rare.

Endogenous Hyperinsulinism

Hypoglycemia due to endogenous hyperinsulinism can be caused by (1) a primary β-cell disorder—typically a β-cell tumor (*insulinoma*), sometimes multiple insulinomas, or a functional β-cell disorder with β-cell hypertrophy or hyperplasia; (2) an antibody to insulin or to the insulin receptor; (3) a β-cell secretagogue such as a sulfonylurea; or perhaps (4) ectopic insulin secretion, among other very rare mechanisms. None of these causes is common.

The fundamental pathophysiologic feature of endogenous hyperinsulinism caused by a primary β-cell disorder or an insulin secretagogue is the failure of insulin secretion to fall to very low levels during hypoglycemia. This feature is assessed by measurement of plasma insulin, C-peptide (the connecting peptide that is cleaved from proinsulin to produce insulin), proinsulin, and glucose concentrations during hypoglycemia. Insulin, C-peptide, and proinsulin levels need not be high relative to normal, euglycemic values; rather, they are inappropriately high in the setting of a low plasma glucose concentration. Critical diagnostic findings are a plasma insulin concentration ≥3 μU/mL (≥18 pmol/L), a plasma C-peptide concentration ≥0.6 ng/mL (≥0.2 nmol/L), and a plasma proinsulin concentration ≥5.0 pmol/L when the plasma glucose concentration is <55 mg/dL (<3.0 mmol/L) with symptoms of hypoglycemia. A low plasma β-hydroxybutyrate concentration (≤2.7 mmol/L) and an increment in plasma glucose level of >25 mg/dL (>1.4 mmol/L) after IV administration of glucagon (1.0 mg) indicate increased insulin (or IGF) actions.

The diagnostic strategy is (1) to measure plasma glucose, insulin, C-peptide, proinsulin, and β-hydroxybutyrate concentrations and to screen for circulating oral hypoglycemic agents during an episode of hypoglycemia and (2) to assess symptoms during the episode and seek their resolution following correction of hypoglycemia by IV injection of glucagon (i.e., to document Whipple's triad). This is straightforward if the patient is hypoglycemic when seen. Since endogenous hyperinsulinemic disorders usually, but not invariably, cause fasting hypoglycemia, a diagnostic episode may develop after a relatively short outpatient fast. Serial sampling during an inpatient diagnostic fast of up to 72 h or after a mixed meal is more problematic. An alternative is to give patients a detailed list of the required measurements and ask them to present to an emergency room, with the list, during a symptomatic episode. Obviously, a normal plasma glucose concentration during a symptomatic episode indicates that the symptoms are not the result of hypoglycemia.

An *insulinoma*—an insulin-secreting pancreatic islet β-cell tumor—is the prototypical cause of endogenous hyperinsulinism and therefore should be sought in patients with a compatible clinical syndrome. However, insulinoma is not the only cause of endogenous hyperinsulinism. Some patients with fasting endogenous hyperinsulinemic hypoglycemia have diffuse islet involvement with β-cell hypertrophy and sometimes hyperplasia. This pattern is commonly referred to as *nesidioblastosis*, although β-cells budding from ducts are not invariably found. Other patients have a similar islet pattern but with postprandial hypoglycemia, a disorder termed *noninsulinoma pancreatogenous hypoglycemia*. Postgastric bypass postprandial hypoglycemia, which most often follows Roux-en-Y gastric bypass, is also characterized by diffuse islet involvement and endogenous hyperinsulinism. Some have suggested that exaggerated GLP-1 responses to meals cause hyperinsulinemia and hypoglycemia, but the relevant pathogenesis has not been clearly established. If medical treatments with agents such as an α-glucosidase inhibitor, diazoxide, or octreotide fail, partial pancreatectomy may be required. Autoimmune hypoglycemias include those caused by an antibody to insulin that binds post-meal insulin and then gradually disassociates, with consequent late postprandial hypoglycemia. Alternatively, an insulin receptor antibody can function as an agonist. The presence of an insulin secretagogue, such as a sulfonylurea or a glinide, results in a clinical and biochemical pattern similar to that of an insulinoma but can be distinguished by the presence of the circulating secretagogue. Finally, there are reports of very rare phenomena such as ectopic insulin secretion, a gain-of-function insulin receptor mutation, and exercise-induced hyperinsulinemia.

Insulinomas are uncommon, with an estimated yearly incidence of 1 in 250,000. Because >90% of insulinomas are benign, they are a treatable cause of potentially fatal hypoglycemia. The median age at presentation is 50 years in sporadic cases, but the tumor usually presents in the third decade when it is a component of multiple endocrine neoplasia type 1 **(Chap. 381)**. More than 99% of insulinomas are within the substance of the pancreas, and the tumors are usually small (<2.0 cm in diameter in 90% of cases). Therefore, they come to clinical attention because of hypoglycemia rather than mass effects. CT or MRI detects ~70–80% of insulinomas. These methods detect metastases in the roughly 10% of patients with a malignant insulinoma. Transabdominal ultrasound often identifies insulinomas, and endoscopic ultrasound has a sensitivity of ~90%. Somatostatin receptor scintigraphy is thought to detect insulinomas in about half of patients. Selective pancreatic arterial calcium injections, with the endpoint of a sharp increase in hepatic venous insulin levels, regionalize insulinomas with high sensitivity, but this invasive procedure is seldom necessary except to confirm endogenous hyperinsulinism in the diffuse islet disorders. Intraoperative pancreatic ultrasonography almost invariably localizes insulinomas that are not readily palpable by the surgeon. Surgical resection of a solitary insulinoma is generally curative. Diazoxide, which inhibits insulin secretion, or the somatostatin analogue octreotide can be used to treat hypoglycemia in patients with unresectable tumors; everolimus, an mTOR (mammalian target of rapamycin) inhibitor, is promising.

ACCIDENTAL, SURREPTITIOUS, OR MALICIOUS HYPOGLYCEMIA

Accidental ingestion of an insulin secretagogue (e.g., as the result of a pharmacy or other medical error) or even accidental administration of insulin can occur. Factitious hypoglycemia, caused by surreptitious or even malicious administration of insulin or an insulin secretagogue, shares many clinical and laboratory features with insulinoma. It is most common among health care workers, patients with diabetes or their relatives, and people with a history of other factitious illnesses. However, it should be considered in all patients being evaluated for hypoglycemia of obscure cause. Ingestion of an insulin secretagogue causes hypoglycemia with increased C-peptide levels, whereas exogenous insulin causes hypoglycemia with low C-peptide levels reflecting suppression of insulin secretion.

Analytical error in the measurement of plasma glucose concentrations is rare. On the other hand, glucose monitors used to guide treatment of diabetes are not quantitative instruments, particularly at low glucose levels, and should not be used for the definitive diagnosis of hypoglycemia. Even with a quantitative method, low measured glucose concentrations can be artifactual—e.g., the result of continued glucose metabolism by the formed elements of the blood ex vivo, particularly in the presence of leukocytosis, erythrocytosis, or thrombocytosis or with delayed separation of the serum from the formed elements (pseudohypoglycemia).

INBORN ERRORS OF METABOLISM CAUSING HYPOGLYCEMIA

Nondiabetic hypoglycemia also results from inborn errors of metabolism. Such hypoglycemia most commonly occurs in infancy but can also occur in adulthood. Cases in adults can be classified into those resulting in fasting hypoglycemia, postprandial hypoglycemia, and exercise-induced hypoglycemia.

Fasting Hypoglycemia

Although rare, disorders of glycogenolysis can result in fasting hypoglycemia. These disorders include glycogen storage disease (GSD) of types 0, I, III, and IV and Fanconi-Bickel syndrome (**Chap. 412**). Patients with GSD types I and III characteristically have high blood lactate levels before and after meals, respectively. Both groups have hypertriglyceridemia, but ketones are high in GSD type III. Defects in fatty acid oxidation also result in fasting hypoglycemia. These defects can include (1) defects in the carnitine cycle; (2) fatty-acid β-oxidation disorders; (3) electron transfer disturbances; and (4) ketogenesis disorders. Finally, defects in gluconeogenesis (fructose-1, 6-biphosphatase) have been reported to result in recurrent hypoglycemia and lactic acidosis.

Postprandial Hypoglycemia

Inborn errors of metabolism resulting in postprandial hypoglycemia are also rare. These errors include (1) glucokinase, SUR1, and Kir6.2 potassium channel mutations; (2) congenital disorders of glycosylation; and (3) inherited fructose intolerance.

Exercise-Induced Hypoglycemia

Exercise-induced hypoglycemia, by definition, follows exercise. It results in hyperinsulinemia caused by increased activity of monocarboxylate transporter 1 in β cells.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Hypoglycemia

In addition to the recognition and documentation of hypoglycemia as well as its treatment (often on an urgent basis), diagnosis of the hypoglycemic mechanism is critical for the selection of therapy that prevents, or at least minimizes, recurrent hypoglycemia.

RECOGNITION AND DOCUMENTATION

Hypoglycemia is suspected in patients with typical symptoms; in the presence of confusion, an altered level of consciousness, or a seizure; or in a clinical setting in which hypoglycemia is known to occur. Blood should be drawn, whenever possible, before the administration of glucose to allow documentation of a low plasma glucose concentration. Convincing documentation of hypoglycemia requires the fulfillment of Whipple's triad. Thus, the ideal time to measure the plasma glucose level is during a symptomatic episode. A normal glucose level excludes hypoglycemia as the cause of the symptoms. A low glucose level confirms that hypoglycemia is the cause of the symptoms, provided the latter resolve after the glucose level is raised. When the cause of the hypoglycemic episode is obscure, additional measurements—made while the glucose level is low and before treatment—should include plasma insulin, C-peptide, proinsulin, and β-hydroxybutyrate levels; also critical are screening for circulating oral hypoglycemic agents and assessment of symptoms before and after the plasma glucose concentration is raised.

When the history suggests prior hypoglycemia and no potential mechanism is apparent, the diagnostic strategy is to evaluate the patient as just described and assess for Whipple's triad during and after an episode of hypoglycemia. On the other hand, while it cannot be ignored, a distinctly low plasma glucose concentration measured in a patient without corresponding symptoms raises the possibility of an artifact (pseudohypoglycemia).

DIAGNOSIS OF THE HYPOGLYCEMIC MECHANISM

In a patient with documented hypoglycemia, a plausible hypoglycemic mechanism can often be deduced from the history, physical examination, and available laboratory data (Table 399-1). Drugs, particularly alcohol or agents used to treat diabetes, should be the first consideration—even in the absence of known use of a relevant drug—given the possibility of surreptitious, accidental, or malicious drug administration. Other considerations include evidence of a relevant critical illness, hormone deficiencies (less commonly), and a non-β-cell tumor that can be pursued diagnostically (rarely). Absent one of these mechanisms in an otherwise seemingly well individual, the physician should consider endogenous hyperinsulinism and proceed with measurements and assessment of symptoms during spontaneous hypoglycemia or under conditions that might elicit hypoglycemia.

URGENT TREATMENT

If the patient is able and willing, oral treatment with glucose tablets or glucose-containing fluids, candy, or food is appropriate. A reasonable initial dose is 15–20 g of glucose. If the patient is unable or unwilling (because of neuroglycopenia) to take carbohydrates orally, parenteral therapy is necessary. IV administration of glucose (25 g) should be followed by a glucose infusion guided by serial plasma glucose measurements. If IV therapy is not practical, SC or IM glucagon (1.0 mg in adults) can be used, particularly in patients with T1DM. Because it acts by stimulating glycogenolysis, glucagon is ineffective in glycogen-depleted individuals (e.g., those with alcohol-induced hypoglycemia). Glucagon also stimulates insulin secretion and is therefore less useful in T2DM. The somatostatin analogue octreotide can be used to suppress insulin secretion in sulfonylurea-induced hypoglycemia. These

treatments raise plasma glucose concentrations only transiently, and patients should therefore be urged to eat as soon as is practical to replete glycogen stores.

PREVENTION OF RECURRENT HYPOGLYCEMIA

Prevention of recurrent hypoglycemia requires an understanding of the hypoglycemic mechanism. Offending drugs can be discontinued or their doses reduced. Hypoglycemia caused by a sulfonylurea can persist for hours or even days. Underlying critical illnesses can often be treated. Cortisol and growth hormone can be replaced if levels are deficient. Surgical, radiotherapeutic, or chemotherapeutic reduction of a non–islet cell tumor can alleviate hypoglycemia even if the tumor cannot be cured; glucocorticoid or growth hormone administration also may reduce hypoglycemic episodes in such patients. Surgical resection of an insulinoma is curative; medical therapy with diazoxide or octreotide can be used if resection is not possible and in patients with a nontumor β-cell disorder. Partial pancreatectomy may be necessary in the latter patients. The treatment of autoimmune hypoglycemia (e.g., with glucocorticoid or immunosuppressive drugs) is problematic, but these disorders are sometimes self-limited. Failing these treatments, frequent feedings and avoidance of fasting may be required. Administration of uncooked cornstarch at bedtime or even an overnight intragastric infusion of glucose may be necessary for some patients.

FURTHER READING

Bonds DE et al: The association between symptomatic, severe hypoglycaemia and mortality in type 2 diabetes: Retrospective epidemiological analysis of the ACCORD study. BMJ 340:b4909, 2010. [PubMed: 20061358]

Cryer PE: *Hypoglycemia in Diabetes*, 3rd ed. Alexandria, VA, American Diabetes Association, 2016.

Cryer PE: Hypoglycemia, in *Williams Textbook of Endocrinology,* 13th ed, S Melmed et al (eds). Philadelphia, Saunders, 2016, pp. 1582–1607.

Joy NG et al: Effects of acute and antecedent hypoglycemia on endothelial function and markers of atherothrombotic balance in health humans. Diabetes 64:2571, 2015. [PubMed: 25695946]

Russell SJ et al: Outpatient glycemic control with a bionic pancreas in type 1 diabetes. N Engl J Med 371:313, 2014. [PubMed: 24931572]

Chapter 400: Disorders of Lipoprotein Metabolism

Daniel J. Rader; Sekar Kathiresan

INTRODUCTION

Lipoproteins are complexes of lipids and proteins that are essential for transport of cholesterol, triglycerides (TGs), and fat-soluble vitamins in the blood. Disorders of lipoprotein metabolism include primary and secondary conditions that substantially increase or decrease specific circulating lipids (e.g., cholesterol or TGs) or lipoproteins (e.g., low density or high density lipoproteins, see below). The demonstration that cholesterol-lowering therapy significantly reduces the clinical complications of atherosclerotic cardiovascular disease (ASCVD) makes it important for clinicians to be familiar with the diagnosis and treatment of lipoprotein disorders. This chapter reviews normal lipoprotein physiology, the pathophysiology of disorders of lipoprotein metabolism, the effects of genetic and environmental factors on lipoprotein metabolism, and the clinical approaches to the diagnosis and management of lipoprotein disorders.

LIPOPROTEIN METABOLISM

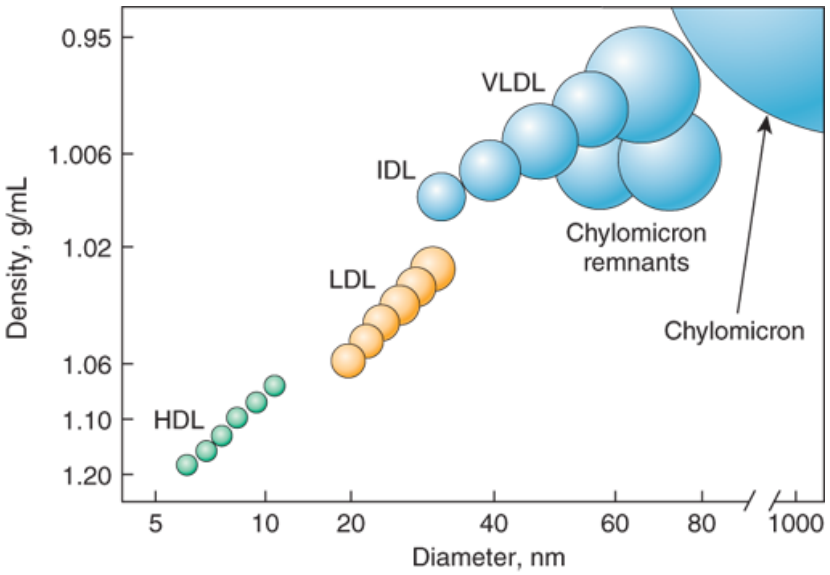
LIPOPROTEIN CLASSIFICATION AND COMPOSITION

Lipoproteins are large macromolecular complexes composed of lipids and proteins that transport poorly soluble lipids (primarily TGs, cholesterol, and fat-soluble vitamins) through body fluids (plasma, interstitial fluid, and lymph) to and from tissues. Lipoproteins play an essential role in the absorption of dietary cholesterol, long-chain fatty acids, and fat-soluble vitamins; the transport of TGs, cholesterol, and fat-soluble vitamins from the liver to peripheral tissues; and the transport of cholesterol from peripheral tissues to the liver and intestine.

Lipoproteins contain a core of hydrophobic lipids (TGs and cholesteryl esters) surrounded by a shell of hydrophilic lipids (phospholipids, unesterified cholesterol) and proteins (called apolipoproteins) that interact with body fluids. The plasma lipoproteins are divided into five major classes based on their relative density **(Fig. 400-1 and Table 400-1)**: chylomicrons, very-low-density lipoproteins (VLDLs), intermediate-density lipoproteins (IDLs), low-density lipoproteins (LDLs), and high-density lipoproteins (HDLs). Each lipoprotein class comprises a family of particles that vary in density, size, and protein composition. Because lipid is less dense than water, the density of a lipoprotein particle is primarily determined by the amount of lipid per particle. Chylomicrons are the most lipid-rich and therefore least dense lipoprotein particles, whereas HDLs have the least lipid and are therefore the most dense lipoproteins. In addition to their density, lipoprotein particles can be classified according to their size, determined either by nondenaturing gel electrophoresis or by nuclear magnetic resonance profiling. There is a strong inverse relationship between density and size, with the largest particles being the most buoyant (chylomicrons) and the smallest particles being the most dense (HDL).

FIGURE 400-1

The density and size distribution of the major classes of lipoprotein particles. Lipoproteins are classified by density and size, which are inversely related. HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; VLDL, very-low-density lipoprotein.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 400-1

Major Lipoprotein Classes

Lipoprotein	Density, g/mL[a]	Size, nm[b]	Electrophoretic Mobility[c]	Apolipoproteins		Other Constituents
				Major	Other	
Chylomicrons	0.930	75–1200	Origin	ApoB-48	A-I, A-V, C-I, C-II, C-III, E	Retinyl esters
Chylomicron remnants	0.930–1.006	30–80	Slow pre-β	ApoB-48	A-I, A-V, C-I, C-II, C-III, E	Retinyl esters
VLDL	0.930–1.006	30–80	Pre-β	ApoB-100	A-I, A-II, A-V, C-I, C-II, C-III, E	Vitamin E
IDL	1.006–1.019	25–35	Slow pre-β	ApoB-100	C-I, C-II, C-III, E	Vitamin E
LDL	1.019–1.063	18–25	β	ApoB-100		Vitamin E
HDL	1.063–1.210	5–12	α	ApoA-I	A-II, A-IV, A-V, C-III, E	LCAT, CETP, paroxonase
Lp(a)	1.050–1.120	25	Pre-β	ApoB-100	Apo(a)	Oxidized phospholipids

[a]The density of the particle is determined by ultracentrifugation. [b]The size of the particle is measured using gel electrophoresis. [c]The electrophoretic mobility of the particle on agarose gel electrophores reflects the size and surface charge of the particle, with β being the position of LDL and α being the position of HDL.

Note: All of the lipoprotein classes contain phospholipids, esterified and unesterified cholesterol, and triglycerides to varying degrees.

Abbreviations: CETP, cholesteryl ester transfer protein; HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LCAT, lecithin-cholesterol acyltransferase; LDL, low-density lipoprotein; Lp(a), lipoprotein A; VLDL, very-low-density lipoprotein.

The proteins associated with lipoproteins, called *apolipoproteins* (Table 400-2), are required for the assembly, structure, function, and metabolism of lipoproteins. Apolipoproteins provide a structural basis for lipoproteins, activate enzymes important in lipoprotein metabolism, and act as ligands for cell surface receptors. ApoB is a very large protein and is the major structural protein of chylomicrons, VLDLs, IDLs, and LDLs; one molecule of apoB, either apoB-48 (chylomicron) or apoB-100 (VLDL, IDL, or LDL), is present on each lipoprotein particle. The human liver synthesizes the full-length apoB-100, whereas the intestine makes the shorter apoB-48, which is derived from the same *APOB* gene by post-transcriptional mRNA editing. HDLs have different apolipoproteins that define this lipoprotein class, most importantly apoA-I, which is synthesized in both the liver and intestine and is found on virtually all HDL particles. ApoA-II is the second most abundant HDL apolipoprotein and is on approximately two-thirds of the HDL particles. ApoC-II, apoC-III, and apoA-V regulate the metabolism of TGs-rich lipoproteins. ApoE plays a critical role in the metabolism and clearance of TG-rich particles. Most apolipoproteins, other than apoB, exchange actively among lipoprotein particles in the blood. Apolipoprotein(a) [apo(a)] is a distinctive apolipoprotein that results in the formation of a lipoprotein known as lipoprotein(a) [Lp(a)], and is discussed more below.

TABLE 400-2

Major Apolipoproteins

Apolipoprotein	Primary Source	Lipoprotein Association	Function
ApoA-I	Intestine, liver	HDL, chylomicrons	Structural protein for HDL Activates LCAT
ApoA-II	Liver	HDL, chylomicrons	Structural protein for HDL
ApoA-IV	Intestine, liver	HDL, chylomicrons	Unknown
ApoA-V	Liver	VLDL, chylomicrons	Promotes LPL-mediated triglyceride lipolysis
Apo(a)	Liver	Lp(a)	Structural protein for Lp(a)
ApoB-48	Intestine	Chylomicrons, chylomicron remnants	Structural protein for chylomicrons
ApoB-100	Liver	VLDL, IDL, LDL, Lp(a)	Structural protein for VLDL, LDL, IDL, Lp(a) Ligand for binding to LDL receptor
ApoC-I	Liver	Chylomicrons, VLDL, HDL	Unknown
ApoC-II	Liver	Chylomicrons, VLDL, HDL	Cofactor for LPL
ApoC-III	Liver, intestine	Chylomicrons, VLDL, HDL	Inhibits LPL activity and lipoprotein binding to receptors
ApoE	Liver	Chylomicron remnants, IDL, HDL	Ligand for binding to LDL receptor and other receptors

Abbreviations: HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LCAT, lecithin-cholesterol acyltransferase; LDL, low-density lipoprotein; Lp(a), lipoprotein A; LPL, lipoprotein lipase; VLDL, very-low-density lipoprotein.

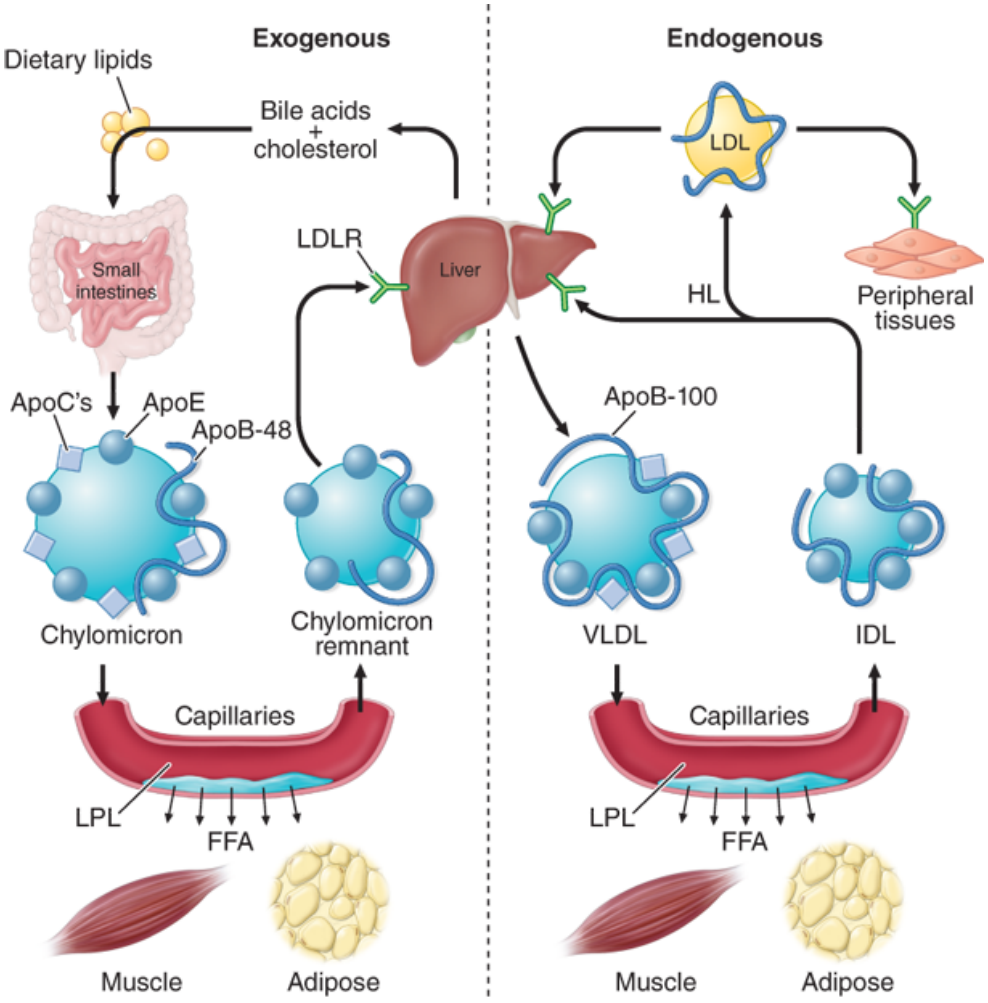
TRANSPORT OF INTESTINALLY DERIVED DIETARY LIPIDS BY CHYLOMICRONS

One critical role of lipoproteins is the efficient transport of dietary lipids from the intestine to tissues that require fatty acids for energy or store and metabolize lipids and of intestinal cholesterol to the liver **(Fig. 400-2)**. Dietary lipids are hydrolyzed by lipases within the intestinal lumen and emulsified with bile acids to form micelles. Dietary cholesterol, fatty acids, and fat-soluble vitamins are absorbed in the proximal small intestine. Cholesterol and retinol are esterified (by the addition of a fatty acid) in the enterocyte to form cholesteryl esters and retinyl esters, respectively. Longer-chain fatty acids (>12 carbons) are incorporated into TGs and packaged with apoB-48, cholesteryl esters, retinyl esters, phospholipids, in a process that requires the action of the microsomal TGs transfer protein (MTP), to form chylomicrons. Nascent chylomicrons are secreted into the intestinal lymph and delivered via the thoracic duct directly to the systemic circulation, where they are extensively processed by peripheral tissues before reaching the liver. The particles encounter lipoprotein lipase (LPL), which is anchored to a glycosylphosphatidylinositol-anchored protein, GPIHBP1, that is attached to the endothelial surfaces of capillaries in adipose tissue, heart, and skeletal muscle (Fig. 400-2). The TGs of chylomicrons are hydrolyzed by LPL, and free fatty acids are released. ApoC-II and apoA-V are apolipoproteins that are transferred to circulating chylomicrons from HDL in the post-prandial state; apoC-II acts as a required cofactor for LPL activation and apoA-V serves as a facilitator of LPL activity. The released free fatty acids are taken up by adjacent myocytes or adipocytes and either oxidized to generate energy or reesterified and stored as TG. Some of the released free fatty acids bind albumin before entering cells and are transported to other tissues, especially the liver. The chylomicron particle progressively shrinks in size as the hydrophobic TG core is hydrolyzed and the hydrophilic lipids

(cholesterol and phospholipids) and apolipoproteins on the particle surface are transferred to HDL, ultimately creating chylomicron remnants.

FIGURE 400-2

The exogenous and endogenous lipoprotein metabolic pathways. The exogenous pathway transports dietary lipids to the periphery and the liver. The endogenous pathway transports hepatic lipids to the periphery. FFA, free fatty acid; HL, hepatic lipase; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; LDLR, low-density lipoprotein receptor; LPL, lipoprotein lipase; VLDL, very-low-density lipoprotein.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Chylomicron remnants are rapidly removed from the circulation by the liver through a process that requires apoE as a ligand for receptors in the liver. Consequently, few, if any, chylomicrons or chylomicron remnants are generally present in the blood after a 12-h fast, except in patients with certain disorders of lipoprotein metabolism.

TRANSPORT OF HEPATICALLY DERIVED LIPIDS BY VLDL AND LDL

Another key role of lipoproteins is the transport of hepatic lipids from the liver to the periphery (Fig. 400-2) to provide an energy source during fasting. During the fasting state, lipolysis of adipose TGs generates fatty acids that are transported to the liver, and the liver is also capable of synthesizing fatty acids through de novo lipogenesis. These fatty acids are esterified by the liver into TGs, which are packaged into VLDL particles along with apoB-100, cholesteryl esters, phospholipids, and vitamin E in a process that, like for chylomicron assembly, requires MTP. VLDL thus resemble chylomicrons in that they are "triglyceride-rich lipoproteins," but they contain apoB-100 rather than apoB-48, are smaller and less buoyant, and have a higher ratio of cholesterol to TG (~1 mg of cholesterol for every 5 mg of TG). After secretion by the liver into the plasma, as with chylomicrons, the TGs of VLDL are hydrolyzed by LPL, especially in muscle, heart, and adipose tissue. After the relatively TG-depleted VLDL remnants dissociate from LPL, they are referred to as IDLs, which contain roughly similar amounts of cholesterol and TG. The liver removes ~40–60% of IDL by receptor–mediated endocytosis via binding to apoE,

which is acquired through transfer of this protein from HDL. The remainder of IDL is further remodeled by hepatic lipase (HL) to form LDL. During this process, phospholipids and TG in the particle are hydrolyzed, and most of the remaining apolipoproteins except apoB-100 are transferred to other lipoproteins. Approximately 70% of LDL is removed from the circulation by receptor–mediated endocytosis (primarily the LDL receptor) in the liver with apoB-100 serving as the ligand for the LDL receptor. It should be noted that apoB-48 does not contain the LDL receptor-binding ligand region and, therefore, clearance of apoB-48-containing chylomicron remnants is dependent on apoE-mediated clearance as noted above. Some LDL particles are lipolytically processed to “small dense LDL” particles.

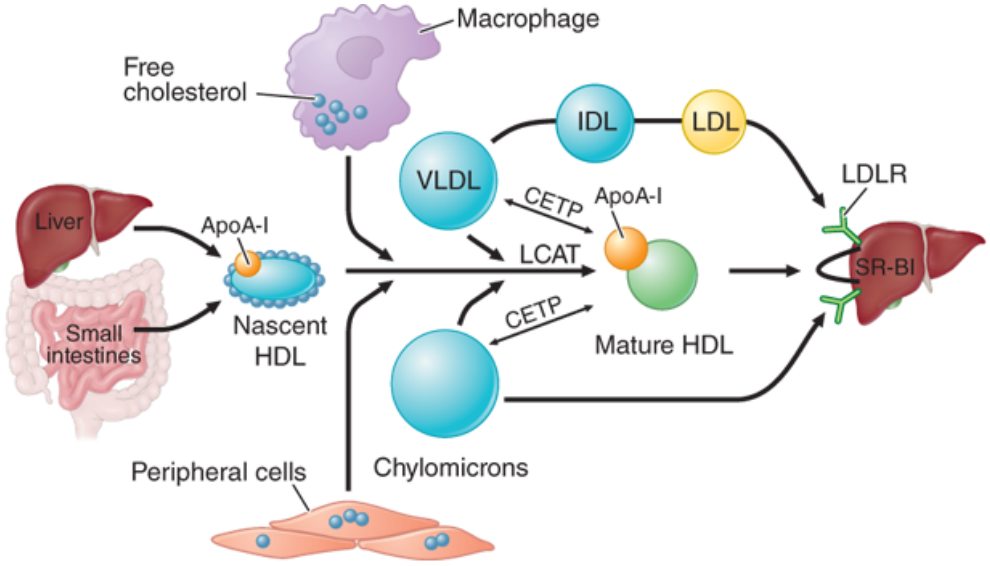
Lp(a) is a lipoprotein similar to LDL in lipid and protein composition, but it contains an additional distinctive protein called apolipoprotein(a) [apo(a)]. Apo(a) is synthesized in the liver and attached to apoB-100 by a disulfide linkage. The major site of clearance of Lp(a) is the liver, but the uptake pathway is not known.

HDL METABOLISM AND REVERSE CHOLESTEROL TRANSPORT

All nucleated cells synthesize cholesterol, but only hepatocytes and enterocytes can effectively excrete cholesterol from the body, into either the bile or the gut lumen, respectively. In the liver, cholesterol is secreted into the bile, either directly or after conversion to bile acids. Cholesterol in peripheral cells is transported from the plasma membranes of peripheral cells to the liver and intestine by a process termed “reverse cholesterol transport” that is facilitated by HDL (Fig. 400-3).

FIGURE 400-3

High-density lipoprotein (HDL) metabolism and reverse cholesterol transport. This pathway transports excess cholesterol from the periphery back to the liver for excretion in the bile. The liver and the intestine produce nascent HDLs. Free cholesterol is acquired from macrophages and other peripheral cells and esterified by lecithin-cholesterol acyltransferase (LCAT), forming mature HDLs. HDL cholesterol can be selectively taken up by the liver via SR-BI (scavenger receptor class BI). Alternatively, HDL cholesteryl ester can be transferred by cholesteryl ester transfer protein (CETP) from HDLs to very-low-density lipoproteins (VLDLs) and chylomicrons, which can then be taken up by the liver. IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; LDLR, low-density lipoprotein receptor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Nascent HDL particles are synthesized by the intestine and the liver. Newly secreted apoA-I rapidly acquires phospholipids and unesterified cholesterol from its site of synthesis (intestine or liver) via cellular efflux promoted by the membrane protein ATP-binding cassette protein A1 (ABCA1). This process results in the formation of discoidal HDL particles, which then recruit additional unesterified cholesterol from cells or circulating lipoproteins. Within the HDL particle, the cholesterol is esterified to cholesteryl ester (CE) through the addition of a fatty acid by lecithin-cholesterol acyltransferase (LCAT), a plasma enzyme associated with HDL; the hydrophobic CE forms the core of the mature HDL particle. As HDL acquires more CE, it becomes spherical, and additional apolipoproteins and lipids are transferred to the particles from the surfaces of chylomicrons and VLDLs during lipolysis.

HDL cholesterol is transported to hepatocytes by two major pathways. HDL CE can be “selectively” taken up by hepatocytes via the scavenger receptor class B1 (SR-B1), a cell surface HDL receptor that mediates the selective transfer of CE from HDL with subsequent

dissociation and “recycling” of the HDL particle. In addition, HDL CE can be transferred to apoB-containing lipoproteins in exchange for TG by the cholesteryl ester transfer protein (CETP). The CE esters are then removed from the circulation by LDL receptor–mediated endocytosis. HDL-derived CE taken up by the hepatocyte through these pathways is hydrolyzed and much of the cholesterol is ultimately excreted directly into the bile or converted to bile acids with excretion to bile, providing a biliary route into the intestinal lumen. There is also evidence that, under certain conditions, HDL cholesterol can be transported directly into the intestinal lumen without requiring a transhepatobiliary route, a process known as “’transintenstinal cholesterol excretion.”

HDL particles undergo extensive remodeling within the plasma compartment by a variety of lipid transfer proteins and lipases. The phospholipid transfer protein (PLTP) transfers phospholipids from other lipoproteins to HDL or among different classes of HDL particles and is a regulator of HDL metabolism. After CETP- and PLTP-mediated lipid exchange, the TG-enriched HDL becomes a much better substrate for HL, which hydrolyzes the TGs and phospholipids to generate smaller HDL particles. A related enzyme called endothelial lipase (EL) hydrolyzes HDL phospholipids, generating smaller HDL particles that are catabolized faster. Remodeling of HDL influences the metabolism, function, and plasma concentrations of HDL.

DISORDERS OF ELEVATED CHOLESTEROL AND TGS

Disorders of lipoprotein metabolism are collectively referred to as “dyslipidemias.” Dyslipidemias are generally characterized clinically by increased plasma levels of cholesterol, TGs, or both, variably accompanied by reduced levels of HDL cholesterol. Unusually low levels of cholesterol also fall within the broad scope of lipoprotein disorders. Because plasma lipids are commonly screened (see below), dyslipidemia is frequently seen in clinical practice. The majority of patients with dyslipidemia have some combination of genetic predisposition (often polygenic) and environmental contribution (diet, lifestyle, medical condition, or drug). Many, but not all, patients with dyslipidemia are at increased risk for ASCVD, which is the primary reason for making the diagnosis, as intervention can substantially reduce this risk. In addition, patients with markedly elevated levels of TGs may be at risk for acute pancreatitis and require intervention to reduce this risk.

Although literally hundreds of proteins influence lipoprotein metabolism and may interact to produce dyslipidemia in an individual patient, there are a limited number of discrete “nodes” or pathways that regulate lipoprotein metabolism and are dysfunctional in specific dyslipidemias. These include: (1) assembly and secretion of TG-rich VLDLs by the liver; (2) lipolysis of TG-rich lipoproteins by LPL; (3) receptor-mediated uptake of apoB-containing lipoproteins by the liver; (4) cellular cholesterol metabolism in the hepatocyte and the enterocyte; and (5) neutral lipid transfer and phospholipid hydrolysis in the plasma. The following discussion will focus on these regulatory nodes, recognizing that in many cases these nodes interact with and influence each other.

DYSLIPIDEMIA CAUSED BY EXCESSIVE HEPATIC SECRETION OF VLDL

Excessive production of VLDL by the liver is one of the most common causes of dyslipidemia. Individuals with excessive hepatic VLDL production usually have elevated fasting TGs and low levels of HDL cholesterol (HDL-C), with variable elevations in LDL cholesterol (LDL-C). A cluster of other metabolic risk factors are often found in association with VLDL overproduction, including obesity, glucose intolerance, insulin resistance, and hypertension (the so-called “metabolic syndrome,” **Chap. 401**). Some of the major factors that drive hepatic VLDL secretion include a high-carbohydrate diet, excessive alcohol use, obesity and insulin resistance, nephrotic syndrome, and genetic factors.

Secondary Causes of VLDL Overproduction

HIGH-CARBOHYDRATE DIET

Dietary carbohydrates are utilized as a substrate for fatty acid synthesis in the liver. Some of the newly synthesized fatty acids are esterified, forming TGs, and secreted in VLDL. Thus, excessive intake of calories as carbohydrates, which is frequent in Western societies, leads to increased hepatic VLDL-TG secretion.

ALCOHOL

Excessive alcohol consumption inhibits hepatic oxidation of free fatty acids, thus promoting hepatic TG synthesis and VLDL secretion. Regular alcohol use also raises plasma levels of HDL-C and should be considered in patients with the relatively unusual combination of elevated TGs and elevated HDL-C.

OBESITY AND INSULIN RESISTANCE

Obesity and insulin resistance are frequently accompanied by dyslipidemia characterized by elevated plasma levels of TG, low HDL-C, variable levels of LDL-C, and increased levels of small dense LDL **(See also Chaps. 395 and 396)**. The increase in adipocyte mass and accompanying decreased insulin sensitivity associated with obesity have multiple effects on lipid metabolism, with one of the major effects being excessive hepatic VLDL production. More free fatty acids are delivered from the expanded and insulin-resistant adipose tissue to the liver, where they are re-esterified in hepatocytes to form TGs, which are packaged into VLDLs for secretion into the circulation. In addition, the increased insulin levels promote increased fatty acid synthesis in the liver. In insulin-resistant patients who progress to type 2 diabetes mellitus, dyslipidemia remains common, even when the patient is under relatively good glycemic control. In addition to increased VLDL production, insulin resistance can also result in decreased LPL activity, resulting in reduced catabolism of chylomicrons and VLDLs and more severe hypertriglyceridemia (see below).

NEPHROTIC SYNDROME

Nephrotic syndrome is a classic cause of excessive VLDL production **(See also Chap. 305)**. The molecular mechanism of VLDL overproduction remains poorly understood but has been attributed to the effects of hypoalbuminemia leading to increased hepatic protein synthesis. Effective treatment of the underlying renal disease often normalizes the lipid profile, but most patients with chronic nephrotic syndrome require lipid-lowering drug therapy.

CUSHING'S SYNDROME

Endogenous or exogenous glucocorticoid excess is associated with increased VLDL synthesis and secretion and hypertriglyceridemia **(See also Chap. 379)**. Patients with Cushing's syndrome frequently have dyslipidemia especially characterized by hypertriglyceridemia and low HDL-C, although elevations in plasma levels of LDL-C can also be seen.

Primary (Genetic) Causes of VLDL Overproduction

Genetic variation influences hepatic VLDL production. A number of genes have been identified in which common and low-frequency variants probably contribute to increased VLDL production, likely involving interactions with diet and other environmental factors. The best recognized inherited condition associated with VLDL overproduction is familial combined hyperlipidemia.

FAMILIAL COMBINED HYPERLIPIDEMIA (FCHL)

FCHL is generally characterized by elevations in plasma levels of TGs (VLDL) and LDL-C (including small dense LDL) and reduced plasma levels of HDL-C. It is estimated to occur in ~1 in 100–200 individuals and is an important contributor to premature coronary heart disease (CHD); ~20% of patients who develop CHD under age 60 have FCHL. FCHL can manifest in childhood but is usually not fully expressed until adulthood. The disease clusters in families, and affected family members typically have one of three possible phenotypes: (1) elevated plasma levels of LDL-C, (2) elevated plasma levels of TGs due to elevation in VLDL, or (3) elevated plasma levels of both LDL-C and TG. The lipoprotein profile can switch among these three phenotypes in the same individual over time and may depend on factors such as diet, exercise, weight, and insulin sensitivity. Patients with FCHL have substantially elevated plasma levels of apoB, often disproportionately high relative to the plasma LDL-C concentration, indicating the presence of small dense LDL particles, which are characteristic of this syndrome.

Individuals with this phenotype generally share the same metabolic defect, namely overproduction of VLDL by the liver. The molecular etiology of this condition remains poorly understood, and no single gene has been identified in which mutations cause this disorder in a simple Mendelian fashion. It is likely that defects in a combination of genes can cause the condition, suggesting that a more appropriate term for the disorder might be *polygenic combined hyperlipidemia.*

The presence of a mixed dyslipidemia (plasma TG levels between 200 and 600 mg/dL and total cholesterol levels between 200 and 400 mg/dL, usually with HDL-C levels <40 mg/dL in men and <50 mg/dL in women) and a family history of dyslipidemia and/or premature CHD suggests the diagnosis. Measurement of apoB levels can help support the diagnosis if they are substantially elevated relative to the LDL-C level. Individuals with this phenotype should be treated aggressively due to significantly increased risk of premature CHD. Decreased dietary intake of simple carbohydrates, increase aerobic exercise, and weight loss can all have beneficial effects on the lipid profile. Patients with type 2 diabetes should be aggressively treated to maintain good glucose control. Most patients with FCHL require lipid-lowering drug therapy, starting with statins, to reduce apoB-containing lipoprotein levels and lower the risk of cardiovascular disease.

LIPODYSTROPHY

Lipodystrophy is a condition in which the generation of adipose tissue generally or in certain fat depots is impaired. Lipodystrophies are often associated with insulin resistance and elevated plasma levels of VLDL and chylomicrons due to increased fatty acid synthesis and VLDL production, as well as reduced clearance of TG-rich particles. Patients with congenital generalized lipodystrophy—a recessive disorder caused by mutations in the *AGPAT2* and *BSCL2* genes—are very rare. These patients have nearly complete absence of subcutaneous fat, accompanied by profound insulin resistance and leptin deficiency, severe hypertriglyceridemia, and accumulation of TGs in multiple tissues including the liver. Patients with generalized lipodystrophy can often be effectively treated with recombinant leptin administration. Partial lipodystrophy is a dominantly inherited disorder that is somewhat more common than the generalized form. It is caused by mutations in several different genes, including lamin A/C (*LMNA*), PPAR gamma (*PPARG*), perilipin (*PLIN1*), and *AKT2*. Partial lipodystrophy is characterized by markedly reduced subcutaneous fat in the extremities and buttocks, accompanied by increased facial, neck, and truncal fat. These patients generally have insulin resistance, often quite severe, accompanied by type 2 diabetes, hepatosteatosis, and dyslipidemia. The dyslipidemia, attributed mostly to increased VLDL production but also possibly due to other factors, is usually characterized by substantially elevated TGs and cholesterol and can be difficult to manage clinically. Patients with partial lipodystrophy are at substantially increased risk of atherosclerotic vascular disease and should therefore be treated aggressively for their dyslipidemia with statins and, if necessary, additional lipid-lowering therapies.

DYSLIPIDEMIA CAUSED BY IMPAIRED LIPOLYSIS OF TG-RICH LIPOPROTEINS

Impaired lipolysis of the TGs in TG-rich lipoproteins (TRLs) also commonly contributes to dyslipidemia. As noted above, LPL is the key enzyme responsible for hydrolyzing the TGs in chylomicrons and VLDL. LPL is synthesized and secreted into the extracellular space from adipocytes, skeletal myocytes, and cardiomyocytes. It is then transported from the subendothelial to the vascular endothelial surfaces by GPIHPB1, which helps dock it to the endothelial surface. Individuals with impaired LPL activity, whether secondary or due to a primary genetic disorder, have elevated fasting TGs and low levels of HDL-C, usually without elevation in LDL-C or apoB. Insulin resistance, in addition to causing excessive VLDL production, can also cause impaired LPL activity and lipolysis. A number of common, low-frequency, and rare genetic variants have been described that influence LPL activity, and single-gene Mendelian disorders that reduce LPL activity have also been described **(Table 400-3)**.

TABLE 400-3

Primary Hyperlipoproteinemias Caused by Known Single-Gene Mutations

Genetic Disorder	Protein (Gene) Defect	Lipoproteins Elevated	Clinical Findings	Genetic Transmission	Estimated Incidence
Hypertriglyceridemia					
Lipoprotein lipase deficiency	LPL (*LPL*)	Chylomicrons, VLDL	Eruptive xanthomas, hepatosplenomegaly, pancreatitis	AR	~1/1,000,000
Familial apoC-II deficiency	ApoC-II (*APOC2*)	Chylomicrons, VLDL	Eruptive xanthomas, hepatosplenomegaly, pancreatitis	AR	<1/1,000,000
ApoA-V deficiency	ApoA-V (*APOA5*)	Chylomicrons, VLDL	Eruptive xanthomas, hepatosplenomegaly, pancreatitis	AR	<1/1,000,000
GPIHBP1 deficiency	*GPIHBP1*	Chylomicrons	Eruptive xanthomas, pancreatitis	AR	<1/1,000,000
Combined Hyperlipidemia					
Familial hepatic lipase deficiency	Hepatic lipase (*LIPC*)	VLDL remnants, HDL	Pancreatitis, CHD	AR	<1/1,000,000
Familial dysbetalipoproteinemia	ApoE (*APOE*)	Chylomicron remnants, VLDL remnants	Palmar and tuberoeruptive xanthomas, CHD, PVD	AR	~1/10,000
Hypercholesterolemia					
Familial hypercholesterolemia	LDL receptor (*LDLR*)	LDL	Tendon xanthomas, CHD	AD	~1/250 to 1/500
Familial defective apoB-100	ApoB-100 (*APOB*)	LDL	Tendon xanthomas, CHD	AD	<~1/1500
Autosomal dominant hypercholesterolemia, type 3	PCSK9 (*PCSK9*)	LDL	Tendon xanthomas, CHD	AD	<1/1,000,000
Autosomal recessive hypercholesterolemia	ARH (*LDLRAP*)	LDL	Tendon xanthomas, CHD	AR	<1/1,000,000
Sitosterolemia	*ABCG5* or *ABCG8*	LDL	Tendon xanthomas, CHD	AR	<1/1,000,000

Abbreviations: AD, autosomal dominant; apo, apolipoprotein; AR, autosomal recessive; ARH, autosomal recessive hypercholesterolemia; CHD, coronary heart disease; LDL, low-density lipoprotein; LPL, lipoprotein lipase; PVD, peripheral vascular disease; VLDL, very-low density lipoprotein.

Secondary Causes of Impaired Lipolysis of TRLs

OBESITY AND INSULIN RESISTANCE

In addition to hepatic overproduction of VLDL, as discussed above, obesity, insulin resistance, and type 2 diabetes have been reported to be associated with variably reduced LPL activity **(See also Chaps. 394, 395, and 396)**. This may be due in part to the effects of tissue insulin resistance leading to reduced transcription of LPL in skeletal muscle and adipose, as well as to increased production of the LPL inhibitor apoC-III by the liver. This reduction in LPL activity often exacerbates the effects of increased VLDL production and contributes to the dyslipidemia seen in these patients.

Primary (Genetic) Causes and Genetic Predisposition to Impaired Lipolysis of TRLs

FAMILIAL CHYLOMICRONEMIA SYNDROME

As noted above, LPL is required for the hydrolysis of TGs in chylomicrons and VLDLs. Genetic deficiency or inactivity of LPL results in impaired lipolysis and profound elevations in plasma chylomicrons, causing *familial chylomicronemia syndrome*. While chylomicronemia predominates, in fact these patients often have elevated plasma levels of VLDL as well. The fasting plasma is turbid, and if left undisturbed for several hours, the chylomicrons float to the top and form a creamy supernatant layer. Fasting TG levels are almost invariably >1000 mg/dL. Fasting cholesterol levels are also elevated but to a lesser degree. The most common cause of FCS involves mutations in the *LPL* gene. *LPL deficiency* has autosomal recessive inheritance (loss of function mutations in both alleles) and has an estimated frequency of ~1 in 1 million, though its true prevalence is unknown. Heterozygotes with LPL mutations often have moderate elevations in plasma TG levels and increased risk for CHD.

Familial chylomicronemia syndrome can be caused by mutations in genes other than LPL. For example, apoC-II is a required cofactor for LPL. *ApoC-II deficiency* due to loss of function mutations in both *APOC2* alleles results in functional lack of LPL activity and severe hyperchylomicronemia that is indistinguishable from LPL deficiency. It is also recessive in inheritance pattern and much rarer than LPL deficiency. Individuals heterozygous for a mutation in *APOC2* do not generally have hypertriglyceridemia. Another apolipoprotein, apoA-V, facilitates the association of VLDL and chylomicrons with LPL and promotes hydrolysis of the TGs. Individuals harboring loss-of-function mutations in both *APOA5* alleles causing *ApoA-V deficiency* develop a form of familial chylomicronemia syndrome. Heterozygosity for variants in *APOA5* that reduce its function contributes to the polygenic basis of hypertriglyceridemia. GPIHBP1 is required for transport and tethering of LPL to the endothelial luminal surface. Homozygosity for mutations in *GPIHBP1* that interfere with its synthesis or folding cause familial chylomicronemia syndrome. Autoantibodies to GPIHBP1 have also been reported to cause severe hyperchylomicronemia.

Familial chylomicronemia syndrome can present in childhood or adulthood with recurrent episodes of severe abdominal pain due to acute pancreatitis. In this setting, the diagnosis should be suspected if a fasting TG level is >750 mg/dL. Eruptive xanthomas, which are small, yellowish-white papules, may appear in clusters on the back, buttocks, and extensor surfaces of the arms and legs. On funduscopic examination, the retinal blood vessels may be opalescent (lipemia retinalis). Hepatosplenomegaly is sometimes noted as a result of uptake of circulating chylomicrons by reticuloendothelial cells in the liver and spleen. Premature CHD is not generally a feature of familial chylomicronemia syndromes.

The diagnosis of familial chylomicronemia syndrome is a clinical diagnosis based on persistence and severity of hypertriglyceridemia in the setting of a history of proven or suspected acute pancreatitis. While LPL activity can be measured in “postheparin plasma” obtained after an IV heparin injection to release the endothelial-bound LPL, this assay is not widely available. Molecular sequencing of the candidate FCS genes can be used to confirm the diagnosis, but is not required for making the clinical diagnosis.

Because of the risk of pancreatitis, it is important to consider the diagnosis and institute therapeutic interventions in familial chylomicronemia syndrome. The goal is to prevent pancreatitis by reducing fasting TG levels to <500 mg/dL. Dietary fat intake should be markedly restricted (to as little as 15 gm/day), often with fat-soluble vitamin supplementation. Consultation with a registered dietician familiar with this disorder is essential. Usually dietary fat restriction alone is not successful in resolving the chylomicronemia, in which case fish oils have been modestly effective in some patients; fibrates (such as fenofibrate) may be tried but are also unlikely to be

effective. A new therapeutic approach involving the suppression of *APOC3* with an antisense oligonucleotide is a promising approach for patients with FCS. In patients with apoC-II deficiency, apoC-II can be provided by infusing fresh-frozen plasma to resolve the chylomicronemia in the acute setting. Management of patients with familial chylomicronemia syndrome is particularly challenging during pregnancy when VLDL production is increased.

FAMILIAL HYPERTRIGLYCERIDEMIA (FHTG)

FHTG is characterized by elevated fasting TGs without a clear secondary cause, average to below average LDL-C levels, low HDL-C levels, and a family history of hypertriglyceridemia. Plasma LDL-C levels are often reduced due to defective conversion of TG-rich lipoproteins to LDL. In contrast to FCHL, apoB levels are not elevated. The identification of other first-degree relatives with hypertriglyceridemia is useful in making the diagnosis. Unlike in FCHL, this condition is not generally associated with a significantly increased risk of CHD. However, if the hypertriglyceridemia is exacerbated by environmental factors, medical conditions, or drugs, the TGs can rise to a level at which acute pancreatitis is a risk. Indeed, management of patients with this condition is mostly focused on reduction of TGs to prevent pancreatitis.

Individuals with this phenotype generally have reduced lipolysis of TRLs, although overproduction of VLDL by the liver can also contribute. While this disorder runs in families, often with a dominant pattern of inheritance, a molecular etiology has not been established. Combinations of gene variants have been shown to cause this phenotype and therefore a more appropriate term for this condition might be *polygenic hypertriglyceridemia.*

It is important to consider and rule out secondary causes of the hypertriglyceridemia as discussed above. Increased intake of simple carbohydrates, obesity, insulin resistance, alcohol use, estrogen treatment, and certain medications can exacerbate this phenotype. Patients who are at high risk for CHD due to other risk factors should be treated with statin therapy. In patients who are otherwise not at high risk for CHD, lipid-lowering drug therapy can frequently be avoided with appropriate dietary and lifestyle changes. Patients with plasma TG levels >500 mg/dL after a trial of diet and exercise should be considered for drug therapy with a fibrate or fish oil to reduce TGs in order to prevent pancreatitis.

DYSLIPIDEMIA CAUSED BY IMPAIRED HEPATIC UPTAKE OF APOB-CONTAINING LIPOPROTEINS

Impaired uptake of LDL and remnant lipoproteins by the liver is another common cause of dyslipidemia. As discussed above, the LDL receptor is the major receptor responsible for uptake of LDL and remnant particles by the liver. Down-regulation of LDL receptor activity or genetic variation that reduces the activity of the LDL receptor pathway leads to elevations in LDL-C. One major factor that reduces LDL receptor activity is a diet high in saturated and *trans* fats. Other medical conditions that reduce LDL receptor activity include hypothyroidism and estrogen deficiency. In addition, genetic variation in a number of genes influences LDL clearance, and mutations in some of these genes cause several discrete Mendelian disorders of elevated LDL-C (Table 400-3).

Secondary Causes of Impaired Hepatic Uptake of Lipoproteins

HYPOTHYROIDISM

Hypothyroidism is associated with elevated plasma LDL-C levels due primarily to a reduction in hepatic LDL receptor function and delayed clearance of LDL **(See also Chap. 375)**. Thyroid hormone increases hepatic expression of the LDL receptor. Hypothyroid patients also frequently have increased levels of circulating IDL, and some patients with hypothyroidism also have mild hypertriglyceridemia. Because hypothyroidism is often subtle and therefore easily overlooked, all patients presenting with elevated plasma levels of LDL-C, especially if there has been an unexplained increase in LDL-C, should be screened for hypothyroidism. Thyroid replacement therapy usually ameliorates the hypercholesterolemia; if not, the patient probably has a primary lipoprotein disorder and may require lipid-lowering drug therapy with a statin.

CHRONIC KIDNEY DISEASE

Chronic kidney disease (CKD) is often associated with mild hypertriglyceridemia (150–400 mg/dL) due to the accumulation of VLDLs and remnant lipoproteins in the circulation **(See also Chap. 305)**. TG lipolysis and remnant clearance are both reduced in patients with renal failure. Because the risk of ASCVD is increased in CKD, patients should usually be treated with lipid-lowering agents, particularly statins.

Patients with solid organ transplants often have increased lipid levels due to the effect of the drugs required for immunosuppression. These patients can present a difficult clinical management problem, but statins are often indicated in these patients, with careful attention to the potential for untoward muscle-related side effects.

Primary (Genetic) Causes of Impaired Hepatic Uptake of Lipoproteins

Genetic variation contributes substantially to elevated LDL-C levels in the general population. It has been estimated that at least 50% of variation in LDL-C is genetically determined. Many patients with elevated LDL-C have *polygenic hypercholesterolemia* due to multiple genetic variants exerting modest LDL-raising effects. In patients who are genetically predisposed to higher LDL-C levels, diet plays a key role; indeed increased saturated and *trans* fats in the diet shifts the entire distribution of LDL levels in the population to the right. Importantly, single-gene (Mendelian) causes of elevated LDL-C are relatively common and should be considered in the differential diagnosis of elevated LDL-C (Table 400-3).

FAMILIAL HYPERCHOLESTEROLEMIA (FH)

FH, also known as autosomal dominant hypercholesterolemia (ADH), is an autosomal co-dominant disorder characterized by elevated plasma levels of LDL-C in the absence of hypertriglyceridemia. FH is caused by mutations that lead to reduced function of the LDL receptor, with the most common being mutations in the *LDLR* gene itself. The reduction in LDL receptor activity in the liver results in a reduced rate of clearance of LDL from the circulation. The plasma level of LDL increases to a level such that the rate of LDL production equals the rate of LDL clearance by residual LDL receptor as well as non-LDL receptor mechanisms. The elevated levels of LDL-C in FH are primarily due to delayed removal of LDL from the blood; in addition, because the removal of IDL is also delayed, the production of LDL from IDL is also increased. Individuals with two mutated *LDLR* alleles (FH homozygotes, or compound heterozygotes) have much higher LDL-C levels than those with one mutant allele (FH heterozygotes).

Although mutations in the *LDLR* are the most common cause of FH, mutations in at least two other genes, *APOB* and *PCSK9*, can also cause ADH. ApoB-100 is the critical structural protein in LDL and contains a ligand for binding to the LDL receptor. Mutations in the LDL receptor–binding domain of apoB-100 cause a form of FH, also known as ADH type 2 or familial defective apoB (FDB). The mutations reduce the affinity of LDL binding to the LDL receptor, such that LDL is removed from the circulation at a reduced rate. Of note, truncating mutations in *APOB* cause low LDL-C levels (see below). The proprotein convertase subtilisin/kexin type 9 (PCSK9) is a secreted protein that binds to the LDL receptor and targets it for lysosomal degradation. Normally, after LDL binds to the LDL receptor, it is internalized along with the receptor, and in the low pH of the endosome, the LDL receptor dissociates from the LDL and recycles to the cell surface. When circulating PCSK9 binds the receptor, the complex is internalized and the receptor is directed to the lysosome, rather than to the cell surface, reducing the number of active LDL receptors. Gain-of-function mutations in *PCSK9* that enhance the activity of PCSK9 cause a form of FH, also known as ADH type 3. Of note, loss-of-function mutations in *PCSK9* markedly lower LDL-C levels (see below).


The population frequency of heterozygous FH was originally estimated to be 1 in 500 individuals, but recent data suggest it is ~1 in 250 individuals, making it one of the most common single-gene disorders in humans. FH has a much higher prevalence in certain founder populations, such as South African Afrikaners, Christian Lebanese, French Canadians, and Lancaster County Amish. Heterozygous FH is characterized by elevated plasma levels of LDL-C (~190–400 mg/dL) and usually relatively normal levels of TGs. Patients with heterozygous FH have hypercholesterolemia from birth, and disease recognition is often based on detection of hypercholesterolemia on routine screening, or a notable family history of hypercholesterolemia, or premature coronary heart disease. Inheritance of FH is dominant, meaning that the condition is inherited from one parent, and ~50% of the patient's siblings and children can be expected to have FH. For this reason, family-based "cascade screening" can be very effective in identifying additional persons with FH. The family history is frequently positive for premature CHD on the side of the family from which the mutation was inherited. Physical findings in some, but not all, patients with heterozygous FH include corneal arcus and/or tendon xanthomas, particularly involving the dorsum of the hands and the Achilles tendons. Untreated heterozygous FH is associated with a markedly increased risk of cardiovascular disease; untreated men with heterozygous FH have an ~50% chance of having a myocardial infarction before age 60 years, and women with heterozygous FH are at substantially increased risk as well. The age of onset of cardiovascular disease is highly variable and depends on the specific molecular defect, the level of LDL-C, and coexisting cardiovascular risk factors.

The diagnosis of FH is generally a clinical diagnosis based on substantial hypercholesterolemia with LDL-C >190 mg/dL in the absence of a secondary etiology, and a family history of hypercholesterolemia and/or premature coronary disease. Secondary causes of significant hypercholesterolemia such as hypothyroidism, nephrotic syndrome, and obstructive liver disease should be excluded. Sequencing of the FH genes (*LDLR*, *APOB*, *PCSK9*) to confirm the diagnosis is available and worthy of consideration; persons with confirmed FH are at higher risk of CVD than those with similar LDL-C levels who don't have FH and therefore may benefit from more aggressive treatment of hypercholesterolemia.

FH patients should always be actively treated to lower plasma levels of LDL-C, preferably starting in childhood. Initiation of a diet low in saturated and *trans* fats is recommended, but heterozygous FH patients require pharmacologic therapy for effective control of their LDL-C levels. Statins are the initial drug class of choice, and usually a more potent member of the class. Many heterozygous FH patients cannot achieve adequate control of their LDL-C levels even with high-intensity statin therapy, and a cholesterol absorption inhibitor (ezetimibe), a PCSK9 inhibitor, or a bile acid sequestrant are the next-line classes of drugs.

Homozygous FH (hoFH) is caused by mutations in both alleles of the LDL receptor or double heterozygosity for mutations in two FH genes. Patients with homozygous FH have been classified into those with virtually no detectable LDL receptor activity (*receptor negative*) and those patients with markedly reduced but detectable LDL receptor activity (*receptor defective*). LDL-C levels in patients with homozygous FH range from about 400 to >1000 mg/dL, with receptor-defective patients at the lower end and receptor-negative patients at the higher end of the range. TGs are usually normal. Some patients with homozygous FH, particularly receptor-negative patients, present in childhood with cutaneous xanthomas on the hands, wrists, elbows, knees, heels, or buttocks. The devastating consequence of homozygous FH is accelerated ASCVD, which often presents in childhood or early adulthood. Atherosclerosis often develops first in the aortic root, where it can cause aortic valvular or supravalvular stenosis, and typically extends into the coronary ostia, which become stenotic. Symptoms can be atypical, and sudden death is not uncommon. Untreated, receptor-negative patients with homozygous FH rarely survive beyond the second decade; patients with receptor-defective LDL receptor defects have a better prognosis but almost invariably develop clinically apparent atherosclerotic vascular disease by age 30, and often much sooner. Carotid and femoral disease develops later in life and is usually not clinically significant.

Homozygous FH should be suspected in a child or young adult with LDL >400 mg/dL without secondary cause. Cutaneous xanthomas, evidence of CVD, and hypercholesterolemia in both parents all are supportive of the diagnosis. While the diagnosis is usually made on clinical grounds, specific mutations can usually be identified by DNA sequencing. Patients with homozygous FH must be treated aggressively to delay the onset and progression of CVD. Although receptor negative patients have no response to statins and PCSK9 inhibitors, receptor defective patients can have modest responses to these medicines and they should be tried in patients with hoFH. Two drugs that reduce the hepatic production of VLDL and thus LDL, a small-molecule inhibitor of the microsomal TG transfer protein (MTP) and an antisense oligonucleotide to apoB, are approved in the United States for the treatment of patients with homozygous FH and should be considered in patients who have insufficient response to statins and PCSK9 inhibitors. LDL apheresis, a physical method of purging the blood of LDL in which the LDL particles are selectively removed from the circulation, should be considered in hoFH patients who have persistently elevated LDL-C levels despite attempts at drug therapy. Liver transplantation is effective in decreasing plasma LDL-C levels in this disorder and is sometimes used as a last resort. Liver-directed gene therapy is under development for hoFH.

FH is an autosomal dominant disorder. There are a few rare conditions that cause an FH-like phenotype in an autosomal recessive manner and should be considered in patients with substantial hypercholesterolemia who do not report a dominant family history of hypercholesterolemia or premature CHD.

AUTOSOMAL RECESSIVE HYPERCHOLESTEROLEMIA (ARH)

ARH is a very rare autosomal recessive disorder that was originally reported in individuals of Sardinian descent. The disease is caused by mutations in the gene *LDLRAP1* encoding the protein LDLR adaptor protein (also called the ARH protein) which is required for LDL receptor–mediated endocytosis in the liver. LDLRAP1 binds to the cytoplasmic domain of the LDL receptor and links the receptor to the endocytic machinery. In the absence of LDLRAP1, LDL binds to the extracellular domain of the LDL receptor, but the lipoprotein-receptor complex fails to be internalized. ARH, like homozygous FH, is characterized by hypercholesterolemia, tendon xanthomas, and premature coronary artery disease (CAD). The levels of plasma LDL-C tend to be intermediate between the levels present in FH homozygotes and FH heterozygotes, and CAD is not usually symptomatic until the third decade. LDL receptor function in cultured fibroblasts is normal or only modestly reduced in ARH, whereas LDL receptor function in lymphocytes and the liver is negligible. Unlike FH homozygotes, the hyperlipidemia responds to treatment with statins, but these patients often require additional therapy to lower plasma LDL-C to acceptable levels.

SITOSTEROLEMIA

Sitosterolemia is a rare autosomal recessive disease that is caused by biallelic loss-of-function mutations in either of two members of the ATP-binding cassette (ABC) half transporter family, *ABCG5* and *ABCG8*. These genes are expressed in both enterocytes and hepatocytes. The proteins heterodimerize to form a functional complex that transports plant sterols such as sitosterol and campesterol, and animal sterols, predominantly cholesterol, across the biliary membrane of hepatocytes into the bile and across the intestinal luminal surface of enterocytes into the gut lumen, reducing their absorption and promoting their excretion. In normal individuals, <5%

of dietary plant sterols are absorbed by the proximal small intestine. The small amounts of plant sterols that enter the circulation are preferentially excreted into the bile and thus levels of plant sterols are kept very low in tissues. In sitosterolemia, the intestinal absorption of sterols is increased and biliary and fecal excretion of the sterols is reduced, resulting in increased plasma and tissue levels of both plant sterols and cholesterol. The increase in hepatic sterol levels results in transcriptional suppression of the expression of the LDL receptor, resulting in reduced uptake of LDL and substantially increased LDL-C levels. In addition to the clinical picture of severe hypercholesterolemia, often accompanied by tendon xanthomas and premature ASCVD, these patients also have anisocytosis and poikilocytosis of erythrocytes and megathrombocytes due to the incorporation of plant sterols into cell membranes. Episodes of hemolysis and splenomegaly are a distinctive clinical feature of this disease compared to other genetic forms of hypercholesterolemia and can be a clue to the diagnosis. Sitosterolemia should be suspected in a patient with severe hypercholesterolemia without a family history of such or who fails to respond to statin therapy. Sitosterolemia can be diagnosed by a laboratory finding of a substantial increase in plasma sitosterol and/or other plant sterols, and should be confirmed by gene sequencing of *ABCG5* and *ABCG8*. It is important to make the diagnosis, because diet, bile acid sequestrants, and cholesterol-absorption inhibitors are the most effective agents to reduce LDL-C and plasma plant sterol levels in these patients. Of note, heterozygosity for mutations in *ABCG5* or *ABCG8* is now recognized to cause a moderate form of hypercholesterolemia.

LYSOSOMAL ACID LIPASE DEFICIENCY (LALD)

LALD, also known as *cholesteryl ester storage disease*, is an autosomal recessive disorder caused by loss-of-function variants in both alleles of the gene *LIPA* encoding the enzyme lysosomal acid lipase (LAL). LAL is responsible for hydrolyzing neutral lipids, particularly TGs and cholesteryl esters, after delivery to the lysosome by cell-surface receptors such as the LDL receptor. It is particularly important in the liver, which clears large amounts of lipoproteins from the circulation. LALD is characterized by elevated LDL-C, usually in association with low HDL-C and with variably elevated TG levels, together with progressive fatty liver ultimately leading to hepatic fibrosis. Genetic deficiency of LAL results in accumulation of neutral lipid in the hepatocytes, leading to hepatosplenomegaly, microvesicular steatosis, and ultimately fibrosis and end-stage liver disease. The most severe form of this disorder, Wolman's disease, presents in infancy and is rapidly fatal. The etiology of the elevated LDL-C levels is primarily due to impaired LDL receptor–mediated clearance of LDL. LALD should be suspected in nonobese patients with elevated LDL-C, low HDL-C, and evidence of fatty liver in the absence of overt insulin resistance. The diagnosis can be made with a dried blood spot assay of LAL activity and confirmed by DNA genotyping for the most common mutation, followed if necessary by sequencing of the gene to find the second mutation. Liver biopsy is required to assess the degree of inflammation and fibrosis. LALD is underdiagnosed; it is critically important to suspect it and make the diagnosis because enzyme replacement therapy is now available and is highly effective in treating this condition.

The above conditions primarily cause elevations in LDL due to impaired catabolism of LDL from the blood. There are a few forms of primary dyslipidemia that impair the catabolism of "remnant" TG-rich lipoproteins (after their processing by LPL) and therefore cause elevations in both cholesterol and TGs due to remnant accumulation.

FAMILIAL DYSBETALIPOPROTEINEMIA (FDBL)

FDBL (also known as *type III hyperlipoproteinemia*) is usually a recessive disorder characterized by a mixed hyperlipidemia (elevated cholesterol and TGs) due to the accumulation of remnant lipoprotein particles (chylomicron remnants and VLDL remnants, or IDL). ApoE is present in multiple copies on chylomicron remnants and IDL, and mediates their removal via hepatic lipoprotein receptors (Fig. 400-2). FDBL is due to genetic variants of apoE, most commonly apoE2, that result in an apoE protein with reduced ability to bind lipoprotein receptors. The *APOE* gene is polymorphic in sequence, resulting in the expression of three common isoforms: apoE3, which is the most common; and apoE2 and apoE4, which both differ from apoE3 by a single amino acid. Although associated with slightly higher LDL-C levels and increased CHD risk, the apoE4 allele is not associated with FDBL. Individuals who carry one or two apoE4 alleles have an increased risk of Alzheimer's disease. ApoE2 has a lower affinity for the LDL receptor; therefore, chylomicron remnants and IDL containing apoE2 are removed from plasma at a slower rate. Individuals who are homozygous for the E2 allele (the E2/E2 genotype) comprise the most common subset of patients with FDBL.

Approximately 0.5% of the general population are apoE2/E2 homozygotes, but only a small minority of these individuals actually develop hyperlipidemia characteristic of FDBL. In most cases, an additional, sometimes identifiable, factor precipitates the development of hyperlipoproteinemia. The most common precipitating factors are a high-fat diet, diabetes mellitus, obesity, hypothyroidism, renal disease, HIV infection, estrogen deficiency, alcohol use, or certain drugs. The disease seldom presents in women before menopause. Certain "dominant negative" mutations in apoE can cause a dominant form of FDBL where the hyperlipidemia is fully manifest in the heterozygous state, but these mutations are very rare.

Patients with FDBL usually present in adulthood with hyperlipidemia, xanthomas, or premature coronary or peripheral vascular disease. In FDBL, in contrast to other disorders of elevated TGs, the plasma levels of cholesterol and TG are often elevated to a similar degree, and the level of HDL-C is usually normal or reduced. Two distinctive types of xanthomas, tuberoeruptive and palmar, are seen in FDBL patients. Tuberoeruptive xanthomas begin as clusters of small papules on the elbows, knees, or buttocks and can grow to the size of small grapes. Palmar xanthomas (alternatively called *xanthomata striata palmaris*) are orange-yellow discolorations of the creases in the palms and wrists. Both of these xanthoma types are virtually pathognomonic for FDBL. Subjects with FDBL have premature ASCVD and tend to have more peripheral vascular disease than is typically seen in FH.

The definitive diagnosis of FDBL can be made either by the documentation of very high levels of remnant lipoproteins or by identification of the apoE2/E2 genotype. A variety of methods are used to identify remnant lipoproteins in the plasma, including "β-quantification" by ultracentrifugation (ratio of directly measured VLDL-C to total plasma TG >0.30), lipoprotein electrophoresis (broad β band), or nuclear magnetic resonance lipoprotein profiling. The Friedewald formula for calculation of LDL-C is not valid in FDBL because the VLDL particles are depleted in TG and enriched in cholesterol. The plasma levels of LDL-C are actually low in this disorder due to defective metabolism of VLDL to LDL. DNA-based apoE genotyping can be performed to confirm homozygosity for apoE2. However, absence of the apoE2/E2 genotype does not strictly rule out the diagnosis of FDBL, because other mutations in apoE can (rarely) cause this condition.

Because FDBL is associated with increased risk of premature ASCVD, it should be treated aggressively. Other metabolic conditions that can worsen the hyperlipidemia (see above) should be managed. Patients with FDBL are typically diet-responsive and can respond favorably to weight reduction and to low-cholesterol, low-fat diets. Alcohol intake should be curtailed. Pharmacologic therapy is often required, and statins are the first line in management. In the event of statin intolerance or insufficient control of hyperlipidemia, cholesterol absorption inhibitors, fibrates, and PCSK9 inhibitors are also effective in the treatment of FDBL.

HEPATIC LIPASE DEFICIENCY

Hepatic lipase (HL; gene name *LIPC*) is a member of the same gene family as LPL and hydrolyzes TGs and phospholipids in remnant lipoproteins and HDL. Hydrolysis of lipids in remnant particles by HL contributes to their hepatic uptake via an apoE-mediated process. HL deficiency is a very rare autosomal recessive disorder characterized by elevated plasma levels of cholesterol and TGs (mixed hyperlipidemia) due to the accumulation of lipoprotein remnants, accompanied by elevated plasma level of HDL-C. The diagnosis is confirmed by measuring HL activity in postheparin plasma and/or confirmation of loss-of-function mutations in both alleles of HL/*LIPC*. Due to the small number of patients with HL deficiency, the association of this genetic defect with ASCVD is not entirely clear, although anecdotally patients with HL deficiency who have premature CVD have been described. As with FDBL, statin therapy is recommended to reduce remnant lipoproteins and cardiovascular risk.

Additional Secondary Causes of Dyslipidemia

Many of the secondary causes of dyslipidemia **(Table 400-4)** have been described above. Additional considerations are discussed here.

TABLE 400-4

Secondary Causes of Dyslipidemia

LDL		HDL		VLDL Elevated	IDL Elevated	Chylomicrons Elevated	Lp(a) Elevated
Elevated	**Reduced**	**Elevated**	**Reduced**				
Hypothyroidism Nephrotic syndrome Cholestasis Acute intermittent porphyria Anorexia nervosa Hepatoma Drugs: thiazides, cyclosporin, carbamazepine	Severe liver disease Malabsorption Malnutrition Gaucher's disease Chronic infectious disease Hyperthyroidism Drugs: niacin toxicity	Alcohol Exercise Exposure to chlorinated hydrocarbons Drugs: estrogen	Smoking DM type 2 Obesity Malnutrition Gaucher's disease Cholesteryl ester storage disease Drugs: anabolic steroids, beta blockers	Obesity DM type 2 Glycogen storage disease Nephrotic syndrome Hepatitis Alcohol Renal failure Sepsis Stress Cushing's syndrome Pregnancy Acromegaly Lipodystrophy Drugs: estrogen, beta blockers, glucocorticoids, bile acid binding resins, retinoic acid	Multiple myeloma Monoclonal gammopathy Autoimmune disease Hypothyroidism	Autoimmune disease DM type 2	Chronic kidney disease Nephrotic syndrome Inflammation Menopause Orchidectomy Hypothyroidism Acromegaly Drugs: growth hormone, isotretinoin

Abbreviations: DM, diabetes mellitus; HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; Lp(a), lipoprotein A; VLDL, very-low-density lipoprotein.

LIVER DISORDERS

Because the liver is the principal site of formation and clearance of lipoproteins, liver disorders can affect plasma lipid levels in a variety of ways **(See also Chap. 329)**. Hepatitis due to infection, drugs, or alcohol is often associated with increased VLDL synthesis and mild to moderate hypertriglyceridemia. Severe hepatitis and liver failure are associated with dramatic reductions in plasma cholesterol and TGs due to reduced lipoprotein biosynthetic capacity. Cholestasis is often associated with hypercholesterolemia. A major pathway by which cholesterol is excreted from the body is via secretion into bile, either directly or after conversion to bile acids, and cholestasis blocks this critical excretory pathway. In cholestasis, free cholesterol, coupled with phospholipids, is secreted into the plasma as a constituent of a lamellar particle called *LP-X*. The particles can deposit in skinfolds, producing lesions resembling those seen in patients with FDBL (xanthomata strata palmaris). Planar and eruptive xanthomas can also be seen in patients with cholestasis.

DRUGS

Many drugs have an impact on lipid metabolism and can result in significant alterations in the lipoprotein profile (Table 400-4). Estrogen administration is associated with increased VLDL and HDL synthesis, resulting in elevated plasma levels of both TGs and HDL-C. This lipoprotein pattern is distinctive because the levels of plasma TG and HDL-C are typically inversely related. Plasma TG levels should be monitored when birth control pills or postmenopausal estrogen therapy is initiated to ensure that the increase in VLDL production does not lead to severe hypertriglyceridemia. Use of low-dose preparations of estrogen or the estrogen patch can minimize the effect of exogenous estrogen on lipids.

INHERITED CAUSES OF LOW LEVELS OF APoB-CONTAINING LIPOPROTEINS

Plasma concentrations of LDL-C <60 mg/dL are unusual. Although in some cases LDL-C levels in this range may be reflective of malnutrition or serious chronic illness, LDL-C <60 mg/dL in an otherwise healthy individual suggests an inherited condition. The major inherited causes of low LDL-C are reviewed here.

Abetalipoproteinemia

The synthesis and secretion of apoB-containing lipoproteins in the enterocytes of the proximal small bowel and in the hepatocytes of the liver involve a complex series of events that coordinate the coupling of various lipids with apoB-48 and apoB-100, respectively. Abetalipoproteinemia is a rare autosomal recessive disease caused by loss-of-function mutations in the gene encoding microsomal TG transfer protein (MTP; gene name *MTTP*), a protein that transfers lipids to nascent chylomicrons and VLDLs in the intestine and liver, respectively. Plasma levels of cholesterol and TG are extremely low in this disorder, and chylomicrons, VLDLs, LDLs, and apoB are undetectable in plasma. The parents of patients with abetalipoproteinemia (obligate heterozygotes) have normal plasma lipid and apoB levels. Abetalipoproteinemia usually presents in early childhood with diarrhea and failure to thrive due to fat malabsorption. The initial neurologic manifestations are loss of deep tendon reflexes, followed by decreased distal lower extremity vibratory and proprioceptive sense, dysmetria, ataxia, and the development of a spastic gait, often by the third or fourth decade. Patients with abetalipoproteinemia also develop a progressive pigmented retinopathy presenting with decreased night and color vision, followed by reductions in daytime visual acuity and ultimately progressing to near-blindness. The presence of spinocerebellar degeneration and pigmented retinopathy in this disease has resulted in some patients with abetalipoproteinemia being misdiagnosed as having Friedreich's ataxia.

Most of the clinical manifestations of abetalipoproteinemia result from defects in the absorption and transport of fat-soluble vitamins. Vitamin E and retinyl esters are normally transported from enterocytes to the liver by chylomicrons, and vitamin E is dependent on VLDL for transport out of the liver and into the circulation. As a consequence of the inability of these patients to secrete apoB-containing particles, patients with abetalipoproteinemia are markedly deficient in vitamin E and are also mildly to moderately deficient in vitamins A and K. Patients with abetalipoproteinemia should be referred to specialized centers for confirmation of the diagnosis and appropriate therapy. Treatment consists of a low-fat, high-caloric, vitamin-enriched diet accompanied by large supplemental doses of vitamin E. It is imperative that treatment be initiated as soon as possible to prevent development of neurologic sequelae, which can progress even with appropriate therapy. New therapies for this serious disease are needed.

Familial Hypobetalipoproteinemia (FHBL)

FHBL generally refers to a condition of low total cholesterol, LDL-C, and apoB due to mutations in the *APOB* gene. Most of the mutations causing FHBL result in a truncated apoB protein, resulting in impaired assembly and secretion of chylomicrons from enterocytes and VLDL from the liver. Mutations that result in VLDL particles containing a truncated apoB protein are cleared from the circulation at an accelerated rate, which also contributes to patients with this disorder having low levels of LDL-C and apoB. Individuals heterozygous for these mutations usually have LDL-C levels <60–80 mg/dL and also tend to have lower levels of plasma TG. Many FHBL patients have elevated levels of hepatic fat (due to reduced VLDL export) and sometimes have increased levels of liver transaminases, although it appears that these patients infrequently develop associated inflammation and fibrosis.

Mutations in both apoB alleles cause homozygous FHBL, an extremely rare disorder resembling abetalipoproteinemia with nearly undetectable LDL-C and apoB. The neurologic defects in this form of hypobetalipoproteinemia tend to be less severe than is typically seen in abetalipoproteinemia. Homozygous hypobetalipoproteinemia can be distinguished from abetalipoproteinemia by examining the inheritance pattern of the plasma LDL-C level. The levels of LDL-C and apoB are normal in the parents of patients with abetalipoproteinemia and low in those of patients with homozygous hypobetalipoproteinemia.

Familial Combined Hypolipidemia

Nonsense mutations in both alleles of the gene Angiopoietin-like 3 (*ANGPTL3*) lead to low plasma levels of all three major lipid fractions —TG, LDL-C, and HDL-C, a phenotype termed familial combined hypolipidemia. ANGPTL3 is a protein synthesized by the liver and secreted into the bloodstream. It inhibits LPL, thus delaying clearance of TRLs from the blood and increasing TRL blood concentrations. Deficiency of ANGPTL3, therefore, raises LPL activity and predominantly lowers blood TG. ANGPTL3 deficiency is associated with a

reduced risk for CHD. Therapies to antagonize ANGPTL3 are in development and initial human studies show that inhibition of ANGPTL3 by either an antisense oligonucleotide or a monoclonal antibodies lower blood levels of TG and LDL-C.

PCSK9 Deficiency

Another inherited cause of low LDL-C results from loss-of-function mutations in *PCSK9*. PCSK9 is a secreted protein that binds to the extracellular domain of the LDL receptor in the liver and promotes the degradation of the receptor. Heterozygosity for nonsense mutations in PCSK9 that interfere with the synthesis of the protein are associated with increased hepatic LDL receptor activity and reduced plasma levels of LDL-C. Such mutations are more frequent in individuals of African descent. Individuals who are heterozygous for a loss-of-function mutation in PCSK9 have an ~30–40% reduction in plasma levels of LDL-C and have a substantial protection from CHD relative to those without a PCSK9 mutation, presumably due to having lower plasma cholesterol levels since birth. Homozygotes for these nonsense mutations have been reported and have extremely low LDL-C levels (<20 mg/dL) but appear otherwise healthy. A sequence variation of somewhat higher frequency (R46L) is found predominantly in individuals of European descent. This mutation impairs, but does not completely destroy, PCSK9 function. As a consequence, the plasma levels of LDL-C in individuals carrying this mutation are more modestly reduced (~15–20%); individuals with these mutations have a 45% reduction in CHD risk. The discovery of this condition led to the development of therapies that antagonize PCSK9, thus reducing LDL-C levels and risk of CHD. Two antibodies against PCSK9 are currently on the market **(Table 400-5)**.

TABLE 400-5

Summary of the Major Approved Drugs Used for the Treatment of Dyslipidemia

Drug	Major Indications	Starting Dose	Maximal Dose	Mechanism	Common Side Effects
HMG-CoA reductase inhibitors (statins)	Elevated LDL-C; increased CV risk			↓ Cholesterol synthesis, ↑ Hepatic LDL receptors, ↓ VLDL production	Myalgias and myopathy ↑ transaminases, ↑ diabetes risk
Lovastatin Pravastatin Simvastatin Fluvastatin Atorvastatin Rosuvastatin Pitavastatin		20–40 mg daily 40–80 mg daily 20–40 mg daily 20–40 mg daily 20–40 mg daily 5–20 mg daily 1–2 mg daily	80 mg daily 80 mg daily 80 mg daily 80 mg daily 80 mg daily 40 mg daily 4 mg daily		
Cholesterol absorption inhibitor Ezetimibe	Elevated LDL-C	10 mg daily	10 mg daily	↓ Cholesterol absorption, ↑ LDL receptors	Elevated transaminases
Bile acid sequestrants Cholestyramine Colestipol Colesevelam	Elevated LDL-C	4 g daily 5 g daily 3750 mg daily	32 g daily 40 g daily 4375 mg daily	↑ Bile acid excretion and ↑ LDL receptors	Bloating, constipation, elevated triglycerides
PCSK9 inhibitors Evolocumab Alirocumab	Elevated LDL-C	140 mg SQ q 2 weeks 75 mg SQ q 2 weeks	420 mg SQ q 1 month (hoFH) 150 mg SQ q 2 weeks	↓ PCSK9 activity, ↑ LDL receptors	Injection site reactions
MTP inhibitor Lomitapide	HoFH	5 mg daily	60 mg daily	↓ VLDL production	Nausea, diarrhea, increased hepatic fat
ApoB inhibitor Mipomersen	HoFH	200 mg SC weekly	200 mg SC weekly	↓ VLDL production	Injection site reactions, flu-like symptoms, increased hepatic fat

Drug	Major Indications	Starting Dose	Maximal Dose	Mechanism	Common Side Effects
Fibric acid derivatives Gemfibrozil Fenofibrate	Elevated TG	 600 mg bid 145 mg qd	 600 mg bid 145 mg qd	↑ LPL, ↓ VLDL synthesis	Dyspepsia, myalgia, gallstones, elevated transaminases
Omega-3 fatty acids Omega-3 acid ethyl esters Icosapent ethyl	Elevated TG	4 g daily 4 g daily	4 g daily 4 g daily	↑ TG catabolism	Dyspepsia, fishy odor to breath

Abbreviations: GI, gastrointestinal; HDL-C, high-density lipoprotein cholesterol; HoFH, homozygous familial hypercholesterolemia; LDL, low-density lipoprotein; LDL-C, LDL-cholesterol; LPL, lipoprotein lipase; TG, triglyceride; VLDL, very-low-density lipoprotein.

DISORDERS OF REDUCED HDL CHOLESTEROL

Low levels of HDL-C are very commonly encountered in clinical practice. Low HDL-C is an important independent predictor of increased cardiovascular risk and has been used regularly in standardized risk calculators. However, it is doubtful that low HDL-C is directly causal for the development of ASCVD. HDL metabolism is strongly influenced by TG metabolism, insulin resistance, and inflammation, among other environmental and medical factors. Thus the HDL-C measurement integrates a number of cardiovascular risk factors, potentially explaining its strong inverse association with ASCVD.

The majority of patients with low HDL-C have some combination of genetic predisposition and secondary factors. Variants in dozens of genes have been shown to influence HDL-C levels. Even more important quantitatively, obesity and insulin resistance have strong suppressive effects on HDL-C, and low HDL-C in these conditions is widely observed. Furthermore, the vast majority of patients with elevated TGs have reduced levels of HDL-C. Most patients with low HDL-C who have been studied in detail have accelerated catabolism of HDL and its associated apoA-I as the physiologic basis for the low HDL-C. Importantly, although HDL-C remains an important biomarker for assessing cardiovascular risk, it is not currently a direct target of intervention for raising the level in order to reduce cardiovascular risk.

INHERITED CAUSES OF VERY LOW LEVELS OF HDL-C

Mutations in genes encoding proteins that play critical roles in HDL synthesis and catabolism can result in reductions in plasma levels of HDL-C. Unlike the genetic forms of hypercholesterolemia, which are invariably associated with premature coronary atherosclerosis, genetic forms of hypoalphalipoproteinemia (low HDL-C) are often not associated with clearly increased risk of ASCVD.

Gene Deletions in the *APOA5-A1-C3-A4* Locus and Coding Mutations in *APOA1*

Complete genetic deficiency of apoA-I due to a complete deletion of the *APOA1* gene results in the virtual absence of circulating HDL and appears to increase the risk of premature ASCVD. The genes encoding *APOA5*, *APOA1*, *APOC3*, and *APOA4* are clustered together on chromosome 11. Some patients with no apoA-I have genomic deletions that include other genes in the cluster. ApoA-I is required for LCAT activity. In the absence of LCAT, free cholesterol levels increase in both plasma (not HDL) and in tissues. The free cholesterol can form deposits in the cornea and in the skin, resulting in corneal opacities and planar xanthomas. Premature CHD is associated with apoA-I deficiency.

Missense and nonsense mutations in the apoA-I gene are present in some patients with low plasma levels of HDL-C (usually 15–30 mg/dL), but are a rare cause of low plasma HDL-C levels. Most individuals with low plasma HDL-C levels due to missense mutations in apoA-I do not appear to have premature CHD. Patients who are heterozygous for an Arg173Cys substitution in apoA-I (so-called apoA-

I_{Milano}) have very low plasma levels of HDL-C due to impaired LCAT activation and accelerated clearance of the HDL particles containing the abnormal apoA-I. Despite having very low plasma levels of HDL-C, these individuals do not have an increased risk of premature CHD. A few selected missense mutations in apoA-I and apoA-II promote the formation of amyloid fibrils, which can cause systemic amyloidosis.

Tangier Disease (*ABCA1* Deficiency)

Tangier disease is a rare autosomal co-dominant form of extremely low plasma HDL-C levels that is caused by mutations in the gene encoding ABCA1, a cellular transporter that facilitates efflux of unesterified cholesterol and phospholipids from cells to apoA-I (Fig. 400-3). ABCA1 in the liver and intestine rapidly lipidates the apoA-I secreted from the basolateral membranes of these tissues. In the absence of ABCA1, the nascent, poorly lipidated apoA-I is immediately cleared from the circulation. Thus, patients with Tangier disease have extremely low circulating plasma levels of HDL-C (<5 mg/dL) and apoA-I (<5 mg/dL). Cholesterol accumulates in the reticuloendothelial system of these patients, resulting in hepatosplenomegaly and pathognomonic enlarged, grayish yellow or orange tonsils. An intermittent peripheral neuropathy (mononeuritis multiplex) or a sphingomyelia-like neurologic disorder can also be seen in this disorder. Tangier disease is probably associated with some increased risk of premature atherosclerotic disease, although the association is not as robust as might be anticipated, given the very low levels of HDL-C and apoA-I in these patients. Patients with Tangier disease also have low plasma levels of LDL-C, which may attenuate the atherosclerotic risk. Obligate heterozygotes for ABCA1 mutations have moderately reduced plasma HDL-C levels (15–30 mg/dL), and their risk of premature CHD remains uncertain.

Familial LCAT Deficiency

This rare autosomal recessive disorder is caused by mutations in LCAT, an enzyme synthesized in the liver and secreted into the plasma, where it circulates associated with lipoproteins (Fig. 400-3). As reviewed above, the enzyme is activated by apoA-I and mediates the esterification of cholesterol to form cholesteryl esters. Consequently, in familial LCAT deficiency, the proportion of free cholesterol in circulating lipoproteins is greatly increased (from ~25% to >70% of total plasma cholesterol). Deficiency in this enzyme interferes with the maturation of HDL particles and results in rapid catabolism of circulating apoA-I.

Two genetic forms of familial LCAT deficiency have been described in humans: complete deficiency (also called *classic LCAT deficiency*) and partial deficiency (also called *fish-eye disease*). Progressive corneal opacification due to the deposition of free cholesterol in the cornea, very low plasma levels of HDL-C (usually <10 mg/dL), and variable hypertriglyceridemia are characteristic of both disorders. In partial LCAT deficiency, there are no other known clinical sequelae. In contrast, patients with complete LCAT deficiency have hemolytic anemia and progressive renal insufficiency that eventually leads to end-stage renal disease. Remarkably, despite the extremely low plasma levels of HDL-C and apoA-I, premature ASCVD is not a consistent feature of either LCAT deficiency or fish eye disease. The diagnosis can be confirmed in a specialized laboratory by assaying plasma LCAT activity or by sequencing the *LCAT* gene.

Primary Hypoalphalipoproteinemia

The condition of low plasma levels of HDL-C (the “alpha lipoprotein”) is referred to as *hypoalphalipoproteinemia*. Primary hypoalphalipoproteinemia is defined as a plasma HDL-C level below the tenth percentile in the setting of relatively normal cholesterol and TG levels, no apparent secondary causes of low plasma HDL-C, and no clinical signs of LCAT deficiency or Tangier disease. This syndrome is often referred to as *isolated low HDL*. A family history of low HDL-C facilitates the diagnosis of an inherited condition, which may follow an autosomal dominant pattern. The metabolic etiology of this disease appears to be primarily accelerated catabolism of HDL and its apolipoproteins. Some of these patients may have ABCA1 mutations and therefore have heterozygous Tangier disease. Several kindreds with primary hypoalphalipoproteinemia and an increased incidence of premature CHD have been described, although it is not clear if the low HDL-C level is the cause of the accelerated atherosclerosis in these families. Association of hypoalphalipoproteinemia with premature CHD may depend on the specific nature of the gene defect or the underlying metabolic defect that either directly or indirectly causes the low plasma HDL-C level.

SCREENING, DIAGNOSIS, AND MANAGEMENT OF DISORDERS OF LIPOPROTEIN METABOLISM

SCREENING

Hypercholesterolemia is a cause of premature CHD that is highly treatable and therefore persons should be actively screened. Plasma lipid levels should be measured, preferably after a 12-h overnight fast, in all adults; guidelines suggest that screening of children between 9 and 11 years of age is also recommended. In most clinical laboratories, the total cholesterol and TGs in the plasma are measured enzymatically, and then the cholesterol in the supernatant is measured after precipitation of apoB-containing lipoproteins to determine the HDL-C. The LDL-C is then estimated using the following equation (the Friedewald formula):

LDL-C = total cholesterol – (TG/5) – HDL-C

(The VLDL cholesterol content is estimated by dividing the plasma TG by 5, reflecting the ratio of TG to cholesterol in VLDL particles.) This formula is reasonably accurate if test results are obtained on fasting plasma and if the TG level does not exceed ~200 mg/dL; by convention it cannot be used if the TG level is >400 mg/dL. LDL-C can be directly measured by a number of methods. Further evaluation and treatment are based primarily on the clinical assessment of absolute cardiovascular risk using risk calculators such as the AHA/ACC risk calculator based on a large amount of observational data.

DIAGNOSIS

A critical first step in managing a lipoprotein disorder is to attempt to determine the class or classes of lipoproteins that are increased or decreased in the patient. Once the dyslipidemia is accurately classified, efforts should be directed to rule out any possible secondary causes (Table 400-4). A careful social, medical, and family history should be obtained. A fasting glucose should be obtained in the initial workup of all subjects with an elevated TG level. Nephrotic syndrome and chronic renal insufficiency should be excluded by obtaining urine protein and serum creatinine. Liver function tests should be performed to rule out hepatitis and cholestasis. Hypothyroidism should be ruled out by measuring serum thyroid-stimulating hormone.

Once secondary causes have been ruled out, attempts should be made to diagnose a primary lipid disorder, because the underlying genetic defect can provide important prognostic information regarding the risk of developing CHD, the response to drug therapy, and the management of other family members. Obtaining the correct diagnosis often requires a detailed family medical history, lipid analyses in family members, and sometimes specialized testing.

Severe Hypertriglyceridemia

If the fasting plasma TG level is >750 mg/dL, the patient may have chylomicronemia. If the elevated TG levels are persistent and the total cholesterol-to-TG ratio is >8, particularly in the setting of a history of pancreatitis, familial chylomicronemia syndrome should be considered. While LPL activity measured in postheparin plasma can support diagnosis, genetic testing for FCS genes may be indicated. Most individuals with persistent severe hypertriglyceridemia do not have a single gene disorder (FCS) but instead are genetically predisposed and have secondary factors (diet, obesity, glucose intolerance, alcohol ingestion, estrogen therapy) that contribute to the hyperlipidemia. Such patients are still at risk for acute pancreatitis and should be treated to reduce their TG levels and thus their risk of pancreatitis.

Severe Hypercholesterolemia

If the levels of LDL-C are very high (greater than a ninety-fifth percentile for age and sex) in absence of secondary causes, familial hypercholesterolemia should be considered, particularly if there is a family history of hypercholesterolemia and/or premature CHD. While FH is a clinical diagnosis, genetic sequencing is now widely available as may be considered in order to determine the molecular diagnosis; a finding of a causal mutation may appropriately result in earlier and more aggressive therapy to lower LDL-C and could also promote family-based cascade screening. Recessive forms of severe hypercholesterolemia are rare, but if a patient with severe hypercholesterolemia has parents with normal cholesterol levels, ARH, sitosterolemia, and LALD should be considered. Patients with more moderate hypercholesterolemia that does not segregate in families as a monogenic trait are likely to have polygenic hypercholesterolemia.

Combined Hyperlipidemia

Elevations in the plasma levels of both cholesterol and TGs are seen in patients with increased plasma levels of both VLDL and LDL or of remnant lipoproteins. A β-quantification to determine the VLDL cholesterol/TG ratio in plasma (see discussion of FDBL), an NMR lipoprotein profile, or a direct measurement of the plasma LDL-C should be performed at least once prior to initiation of lipid-lowering therapy to determine if the hyperlipidemia is due to the accumulation of remnants or to an increase in both LDL and VLDL. Measurement of plasma apoB levels can help identify patients with FCHL who may require more aggressive treatment.

Given the prevalence of primary and secondary dyslipidemias and the clinical benefits of early diagnosis and initiation of therapy, it is essential that physicians screen lipids systematically, rule out secondary causes of dyslipidemia, suspect inherited disorders of lipoprotein metabolism where appropriate, actively promote family-based cascade screening, and be knowledgeable about the existing therapeutic options, including PCSK9 inhibitors. The field of "clinical lipidology" has matured and is moving toward a more systematic clinical application of genomic medicine. Diagnostic DNA sequencing or genotyping in patients with suspected FH, FCS, and FDBL has the potential to enhance molecular diagnosis, facilitate appropriate therapeutic interventions, and promote family-based cascade screening.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Lipoprotein Disorders

The major goals in the clinical management of lipoprotein disorders are: (1) prevention of acute pancreatitis in patients with severe hypertriglyceridemia; and (2) prevention of CVD and related cardiovascular events.

MANAGEMENT OF SEVERE HYPERTRIGLYCERIDEMIA TO PREVENT PANCREATITIS

Although the observational relationship between severe hypertriglyceridemia, particularly chylomicronemia, and acute pancreatitis is well-established, there has never been a clinical trial designed or powered to prove that intervention to reduce TGs reduces the risk of pancreatitis. Nevertheless, it is generally considered appropriate medical practice to intervene in patients with TGs >500 mg/dL in order to reduce the risk of pancreatitis. It remains controversial whether individuals with severe hypertriglyceridemia are at increased risk for ASCVD.

Lifestyle

Modifying the lifestyle of the patient with severe hypertriglyceridemia often is associated with a significant reduction in plasma TG level. Patients who drink alcohol should be encouraged to decrease or preferably eliminate their intake. Patients with severe hypertriglyceridemia often benefit from a formal dietary consultation with a dietician intimately familiar with counseling patients on the dietary management of high TGs. Dietary fat intake should be restricted to reduce the formation of chylomicrons in the intestine. The excessive intake of simple carbohydrates should be discouraged because insulin drives TG production in the liver. Aerobic exercise and even increase in regular physical activity can have a positive effect in reducing TG levels and should be strongly encouraged. For patients who are overweight, weight loss can help to reduce TG levels. In extreme cases, bariatric surgery has been shown to not only produce effective weight loss but also substantially reduce plasma TG levels.

Pharmacologic Therapy for Severe Hypertriglyceridemia

Despite the above interventions, however, many patients with severe hypertriglyceridemia require pharmacologic therapy (Table 400-5). Patients who persist in having fasting TG >500 mg/dL despite active lifestyle management are candidates for pharmacologic therapy. The two major classes of drugs used for management of these patients are fibrates and omega-3 fatty acids (fish oils). In addition, statins can reduce plasma TG levels and also reduce CVD risk, and should be used in patients with severe hypertriglyceridemia who are at increased risk of CVD.

Fibrates

Fibric acid derivatives, or fibrates, are agonists of PPARα, a nuclear receptor involved in the regulation of lipid metabolism. Fibrates stimulate LPL activity (enhancing TG hydrolysis), reduce apoC-III synthesis (enhancing lipoprotein remnant clearance), promote β-oxidation of fatty acids, and may reduce VLDL TG production. This class of therapeutic agents sometimes lowers but more often raises the plasma level of LDL-C in individuals with severe hypertriglyceridemia. Fibrates are generally well tolerated, but are associated with an increase in the incidence of gallstones. Fibrates can cause myopathy, especially when combined with other lipid-lowering therapy

(statins, niacin), and can raise creatinine. Fibrates should be used with caution in patients with CKD. Importantly, fibrates can potentiate the effect of warfarin and certain oral hypoglycemic agents, so the anticoagulation status and plasma glucose levels should be closely monitored in patients on these agents.

Omega 3 Fatty Acids (Fish Oils)

Omega-3 fatty acids, or omega-3 polyunsaturated fatty acids (n-3 PUFAs), commonly known as fish oils, are present in high concentration in fish and in flaxseed. The most widely used n-3 PUFAs for the treatment of hyperlipidemias are the two active molecules in fish oil: eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA). n-3 PUFAs have been concentrated into tablets and in doses of 3–4 g/d are effective at lowering fasting TG levels. Fish oils can cause an increase in plasma LDL-C levels in some patients. In general, fish oils are well tolerated, with the major side effect being dyspepsia. They appear to be safe, at least at doses up to 3–4 g, but can be associated with a prolongation in the bleeding time.

MANAGEMENT OF CHOLESTEROL TO PREVENT CARDIOVASCULAR DISEASE

In contrast to hypertriglyceridemia and pancreatitis, there are abundant and compelling data that intervention to reduce LDL-C substantially reduces the risk of CVD, including myocardial infarction and stroke, as well as total mortality. Thus, it is imperative that patients with hypercholesterolemia be assessed for cardiovascular risk and for the need for intervention. It is also worth noting that patients at high risk for CVD who have plasma LDL-C levels in the “normal” or average range also benefit from intervention to reduce LDL-C levels.

Lifestyle

The first approach to a patient with hypercholesterolemia and high cardiovascular risk is to make any necessary lifestyle changes. In obese patients, efforts should be made to reduce body weight to the ideal level. Patients should receive dietary counseling to reduce the content of saturated fats, *trans* fats, and cholesterol in the diet. Regular aerobic exercise has relatively little impact on reducing plasma LDL-C levels, although it has cardiovascular benefits independent of LDL lowering.

Pharmacologic Therapy for Hypercholesterolemia

The decision to use LDL-lowering drug therapy (Table 400-5)—with a statin being first-line therapy—depends on the level of LDL-C as well as the level of cardiovascular risk. In general, patients with a Mendelian disorder of elevated LDL-C such as FH must be treated to reduce the very high lifetime risk of CVD, and treatment should be initiated as early as possible in adulthood or, in some cases, during childhood.

Otherwise, the decision to initiate LDL-lowering drug therapy is generally determined by the level of cardiovascular risk. In patients with established CVD, statin therapy is well supported by clinical trial data and should be used regardless of the LDL-C level. For patients >40 years old without clinical CVD, the AHA/ACC risk calculator *(http://my.americanheart.org/professional/StatementsGuidelines/PreventionGuidelines/Prevention-Guidelines_UCM_457698_SubHomePage.jsp)* can be used to determine the 10-year absolute risk for CVD, and current guidelines suggest that a 10-year risk >7.5% merits consideration of statin therapy regardless of plasma LDL-C level. For younger patients, the assessment of lifetime risk of CVD may help inform the decision to start a statin.

HMG-C*o*A Reductase Inhibitors (Statins)

Statins inhibit HMG-CoA reductase, a key enzyme in cholesterol biosynthesis. By inhibiting cholesterol biosynthesis, statins lead to increased hepatic LDL receptor activity and accelerated clearance of circulating LDL, resulting in a dose-dependent reduction in plasma levels of LDL-C. The magnitude of LDL lowering associated with statin treatment varies widely among individuals, but once a patient is on a statin, the doubling of the statin dose produces an ~6% further reduction in the level of plasma LDL-C. The statins currently available differ in their LDL-C–reducing potency (Table 400-5). Current recommendations are to use “high-intensity” statin therapy in most patients deemed at high risk of CVD. Currently, there is no convincing evidence that any of the different statins confer an advantage that is independent of the effect on LDL-C. Statins also reduce plasma TGs in a dose-dependent fashion, which is roughly proportional to their LDL-C–lowering effects (if the TGs are <400 mg/dL). Statins have a modest HDL-raising effect (5–10%) that is not generally dose-dependent.

Statins are well tolerated and can be taken in tablet form once a day. Potential side effects include dyspepsia, headaches, fatigue, and muscle or joint pains. Severe myopathy and even rhabdomyolysis occur rarely with statin treatment. The risk of statin-associated

myopathy is increased by the presence of older age, frailty, renal insufficiency, and coadministration of drugs that interfere with the metabolism of statins, such as erythromycin and related antibiotics, antifungal agents, immunosuppressive drugs, and fibric acid derivatives (particularly gemfibrozil). Severe myopathy can usually be avoided by careful patient selection, avoidance of interacting drugs, and instructing the patient to contact the physician immediately in the event of unexplained muscle pain. In the event of muscle symptoms, the plasma creatine kinase (CK) level should be obtained to differentiate myopathy from myalgia. Serum CK levels need not be monitored on a routine basis in patients taking statins, because an elevated CK in the absence of symptoms does not predict the development of myopathy and does not necessarily suggest the need for discontinuing the drug.

Another consequence of statin therapy can be elevation in liver transaminases (alanine aminotransferase [ALT] and aspartate aminotransferase [AST]). They should be checked before starting therapy, at 2–3 months, and then annually. Substantial (greater than three times the upper limit of normal) elevation in transaminases is relatively rare, and mild-to-moderate (one to three times normal) elevation in transaminases in the absence of symptoms need not mandate discontinuing the medication. Severe clinical hepatitis associated with statins is exceedingly rare, and the trend is toward less frequent monitoring of transaminases in patients taking statins. The statin-associated elevation in liver enzymes resolves upon discontinuation of the medication.

Statins appear to be remarkably safe. Meta-analyses of large randomized controlled clinical trials with statins do not suggest an increase in any major noncardiac diseases except type 2 diabetes. A small excess percentage of those taking statins will develop diabetes but the benefits associated with the reduction in cardiovascular events outweigh the increase in incidence of diabetes. Statins are the drug class of choice for LDL-C reduction and are by far the most widely used class of lipid-lowering drugs.

Cholesterol Absorption Inhibitors

Cholesterol within the lumen of the small intestine is derived from the diet (about one-third) and the bile (about two-thirds) and is actively absorbed by the enterocyte through a process that involves the protein NPC1L1. Ezetimibe (Table 400-5) is a cholesterol absorption inhibitor that binds directly to and inhibits NPC1L1 and blocks the intestinal absorption of cholesterol. Ezetimibe (10 mg) inhibits cholesterol absorption by almost 60%, resulting in a reduction in delivery of dietary sterols in the liver and an increase in hepatic LDL receptor expression. The mean reduction in plasma LDL-C on ezetimibe (10 mg) is 18%, and the effect is additive when used in combination with a statin. Effects on TG and HDL-C levels are negligible. When used in combination with a statin, monitoring of liver transaminases is recommended. The only roles for ezetimibe in monotherapy are in patients who do not tolerate statins and in sitosterolemia.

Bile Acid Sequestrants (Resins)

Bile acid sequestrants bind bile acids in the intestine and promote their excretion rather than reabsorption in the ileum. To maintain the bile acid pool size, the liver diverts cholesterol to bile acid synthesis. The decreased hepatic intracellular cholesterol content results in upregulation of the LDL receptor and enhanced LDL clearance from the plasma. Bile acid sequestrants, including cholestyramine, colestipol, and colesevelam (Table 400-5), primarily reduce plasma LDL-C levels but can cause an increase in plasma TGs. Therefore, patients with hypertriglyceridemia generally should not be treated with bile acid–binding resins. Cholestyramine and colestipol are insoluble resins that must be suspended in liquids. Colesevelam is available as tablets but generally requires up to six to seven tablets per day for effective LDL-C lowering. Most side effects of resins are limited to the gastrointestinal tract and include bloating and constipation. Because bile acid sequestrants are not systemically absorbed, they are very safe and the cholesterol-lowering drug of choice in children and in women of childbearing age who are lactating, pregnant, or could become pregnant. They are effective in combination with statins and in combination with ezetimibe and are particularly useful with one or both of these drugs for patients with severe hypercholesterolemia or those with statin intolerance.

PCSK9 Inhibitors

PCSK9 inhibitors are antibodies that bind to circulating PCSK9 and prevent its interaction with the LDL receptor. This permits more LDL receptors to recycle back to the cell surface and functionally increases the number of LDL receptors available to remove LDL from the blood. They are highly effective in lowering LDL-C, with a mean 50–60% reduction in LDL-C. They also reduce plasma levels of Lp(a) modestly. PCSK9 inhibition has been proven to reduce cardiovascular events in patients with existing CHD. These antibodies are administered subcutaneously every 2–4 weeks. They are generally well-tolerated, with the major side effect being injection site reactions. They are indicated as second line (after statin) or third line (after statin + ezetimibe) therapy in patients with FH or CHD in whom LDL-C is not reduced to acceptable levels with statin (+/− ezetimibe) alone.

Specialized Drugs For Homozygous FH

Two “orphan” drugs are approved specifically for the management of homozygous FH. They include a small-molecule inhibitor of MTP, called lomitapide, and an antisense oligonucleotide against apoB, called mipomersen. These drugs reduce VLDL production and LDL-C levels in homozygous FH patients. Due to their mechanism of action, each drug causes an increase in hepatic fat, the long-term consequences of which are unknown. In addition, lomitapide is associated with gastrointestinal-related side effects, and mipomersen is associated with skin reactions and flu-like symptoms. One of these drugs should be considered in hoFH patients after a trial of a statin plus PCSK9 inhibitor is shown to be insufficient to reduce LDL-C levels.

LDL Apheresis

Patients who cannot reduce their LDL-C to acceptable levels despite optimally tolerated drug therapy are candidates for LDL apheresis. In this process, the patient’s plasma is passed over a column that selectively removes the LDL, and the LDL-depleted plasma is returned to the patient. LDL apheresis is indicated for patients on maximally tolerated combination drug therapy (including a PCSK9 inhibitor) who have CHD and a plasma LDL-C level >200 mg/dL or no CHD and a plasma LDL-C level >300 mg/dL; LDL apheresis could be considered in high-risk patients who have an LDL-C > 160 mg/dL on maximal therapy.

FURTHER READING

Abul-Husn NS et al: Genetic identification of familial hypercholesterolemia within a single U.S. health care system. Science 354:6319, 2016.

Cholesterol Treatment Trialists’ (CTT) Collaboration et al: Efficacy and safety of LDL-lowering therapy among men and women: Meta-analysis of individual data from 174,000 participants in 27 randomised trials. Lancet 385:1397, 2015. [PubMed: 25579834]

Dron JS, Hegele RA: Genetics of triglycerides and the risk of atherosclerosis. Curr Atheroscler Rep 19:31, 2017. [PubMed: 28534127]

Garg A (ed): *Dyslipidemias: Pathophysiology, Evaluation and Management*. Humana Press, 2015.

Gidding SS et al: The agenda for familial hypercholesterolemia: A scientific statement from the American Heart Association. Circulation 132:2167, 2015. [PubMed: 26510694]

Khavandi M et al: Treatment of dyslipidemias to prevent cardiovascular disease in patients with type 2 diabetes. Curr Cardiol Rep 19:7, 2017. [PubMed: 28132397]

Khera AV et al: Association of rare and common variation in the lipoprotein lipase gene with coronary artery disease. JAMA 317:937, 2017. [PubMed: 28267856]

Rader DJ: New Therapeutic approaches to the treatment of dyslipidemia. Cell Metab 23:405, 2016. [PubMed: 26853751]

Sabatine MS et al: Evolocumab and clinical outcomes in patients with cardiovascular disease. N Engl J Med 376:1713, 2017. [PubMed: 28304224]

Chapter 401: The Metabolic Syndrome

Robert H. Eckel

INTRODUCTION

The metabolic syndrome (syndrome X, insulin resistance syndrome) consists of a constellation of metabolic abnormalities that confer increased risk of cardiovascular disease (CVD) and diabetes mellitus. Evolution of the criteria for the metabolic syndrome since the original definition by the World Health Organization in 1998 reflects growing clinical evidence and analysis by a variety of consensus conferences and professional organizations. The major features of the metabolic syndrome include central obesity, hypertriglyceridemia, low levels of high-density lipoprotein (HDL) cholesterol, hyperglycemia, and hypertension **(Table 401-1)**.

TABLE 401-1

NCEP:ATPIII[a] 2001 and Harmonizing Definition Criteria for the Metabolic Syndrome

<table>
<tr><th>NCEP:ATPIII 2001</th><th colspan="3">Harmonizing Definition[b]</th></tr>
<tr><td rowspan="6">Three or more of the following:
• Central obesity: waist circumference >102 cm (M), >88 cm (F)
• Hypertriglyceridemia: triglyceride level ≥150 mg/dL or specific medication
• Low HDL[c] cholesterol: <40 mg/dL and <50 mg/dL for men and women, respectively, or specific medication
• Hypertension: blood pressure ≥130 mmHg systolic or ≥85 mmHg diastolic or specific medication
• Fasting plasma glucose level ≥100 mg/dL or specific medication or previously diagnosed type 2 diabetes</td><td colspan="3">Three of the following:
Waist circumference (cm)</td></tr>
<tr><td>Men</td><td>Women</td><td>Ethnicity</td></tr>
<tr><td>≥94</td><td>≥80</td><td>Europid, sub-Saharan African, Eastern and Middle Eastern</td></tr>
<tr><td>≥90</td><td>≥80</td><td>South Asian, Chinese, and ethnic South and Central American</td></tr>
<tr><td>≥85</td><td>≥90</td><td>Japanese</td></tr>
<tr><td colspan="3">• Fasting triglyceride level >150 mg/dL or specific medication
• HDL cholesterol level <40 mg/dL and <50 mg/dL for men and women, respectively, or specific medication
• Blood pressure >130 mm systolic or >85 mm diastolic or previous diagnosis or specific medication
• Fasting plasma glucose level ≥100 mg/dL (alternative indication: drug treatment of elevated glucose levels)</td></tr>
</table>

[a]National Cholesterol Education Program and Adult Treatment Panel III. [b]In this analysis, the following thresholds for waist circumference were used: white men, ≥94 cm; African-American men, ≥94 cm; Mexican-American men, ≥90 cm; white women, ≥80 cm; African-American women, ≥80 cm; Mexican-American women, ≥80 cm. For participants whose designation was “other race—including multiracial,” thresholds that were once based on Europid cutoffs (≥94 cm for men and ≥80 cm for women) and on South Asian cutoffs (≥90 cm for men and ≥80 cm for women) were used. For participants who were considered “other Hispanic,” the International Diabetes Federation thresholds for ethnic South and Central Americans were used. [c]High-density lipoprotein.

GLOBAL HEALTH/EPIDEMIOLOGY

The most challenging feature of the metabolic syndrome to define is waist circumference. Intraabdominal circumference (visceral adipose tissue) is the most strongly related to insulin resistance and risk of diabetes and CVD, and for any given waist circumference the distribution of adipose tissue

between subcutaneous (SC) and visceral depots varies substantially. Thus, within and between populations, there is a lesser vs greater risk at the same waist circumference. These differences in populations reflect the range of waist circumferences considered to confer risk in different geographic locations (Table 401-1).

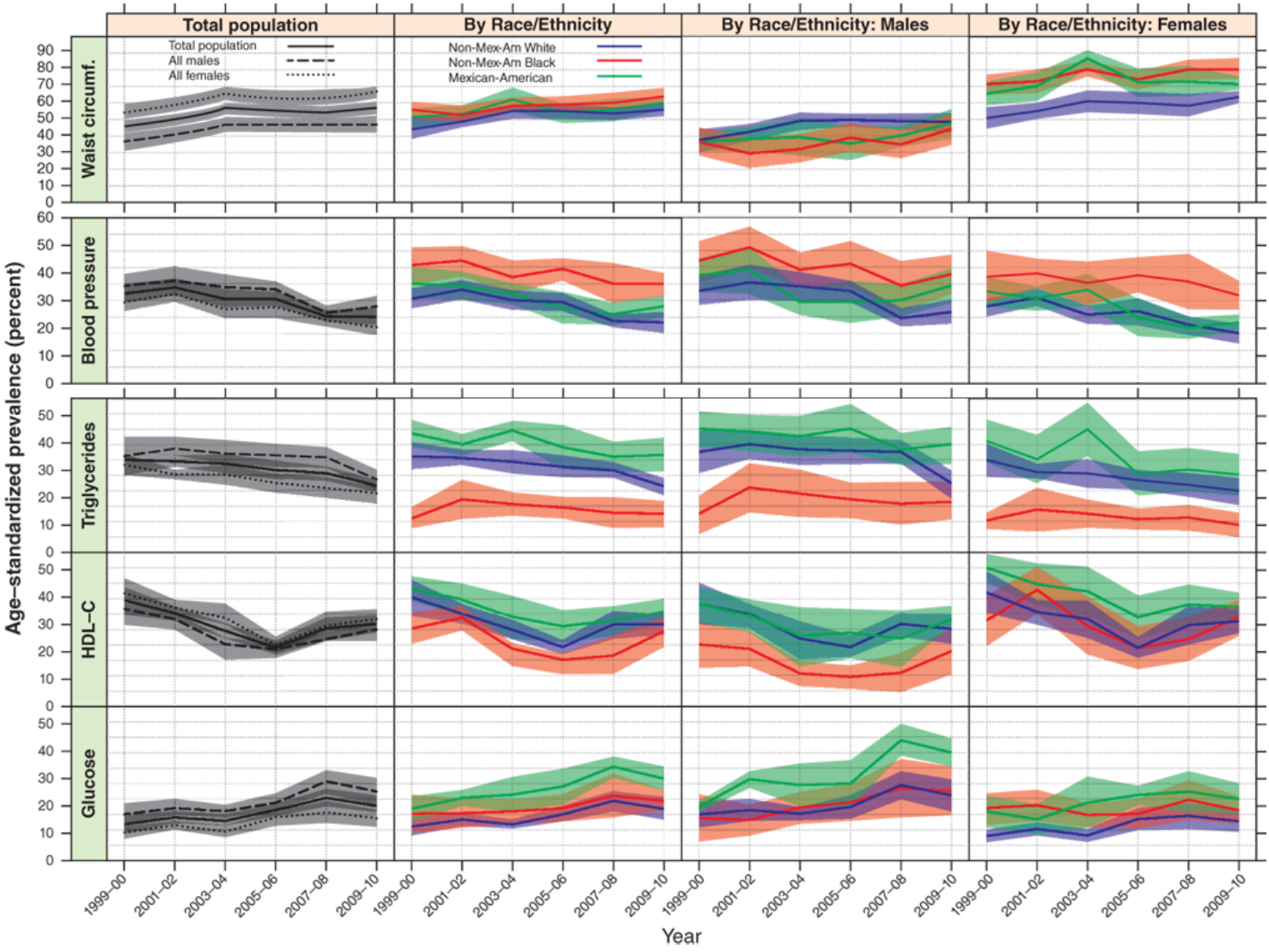
The prevalence of the metabolic syndrome varies around the world, in part reflecting the age and ethnicity of the populations studied and the diagnostic criteria applied. In general, the prevalence of the metabolic syndrome increases with age. The highest recorded prevalence worldwide is among Native Americans, with an age-adjusted 53% of women and 45% of men meeting the criteria of the National Cholesterol Education Program and Adult Treatment Panel III (NCEP:ATPIII). Greater global industrialization is associated with rising rates of obesity, and expected increase in the prevalence of the metabolic syndrome, especially as the population ages. Moreover, the rising prevalence and severity of obesity among children reflects features of the metabolic syndrome in a younger population, now estimated to be up to 23% and >60% amongst obese and overweight children.

In 2012, the overall prevalence of the metabolic syndrome in the United States was 33% with a higher prevalence in women than men (36% vs 30%, respectively). When stratified by race/ethnicity, the highest prevalence of the metabolic syndrome was 35% in Hispanics followed by 33% in non-Hispanic Caucasians and blacks. From 2003–2004 to 2011–2012, overall prevalence of the metabolic syndrome increased from 33% in 2003–2004 to 35% in 2011–2012. In France, studies of a cohort of >18-year-old adults revealed a gradual increase from <10% prevalence for each sex to 19% of men and ~10% for women. However, greater compliance with the French Nutrition and Health Program-Guideline Score was associated with a lower prevalence in both genders. A similar heart healthy nutritional impact on the prevalence of metabolic syndrome is now apparent in the United States.

The frequency distribution of the five components of the syndrome for the U.S. population (NHANES III) is summarized in **Fig. 401-1**. Increases in waist circumference predominate among women, whereas increases in fasting plasma triglyceride levels (i.e., >150 mg/dL), reductions in HDL cholesterol levels, and hyperglycemia are more likely in men.

FIGURE 401-1

Prevalence and trends of the five components of metabolic syndrome in the adult US population, 1999–2010. Prevalence and trends of the components of Metabolic Syndrome for US adults aged 20 or older for the total population by sex (first column), by race (second column), and by race and sex (third and fourth columns). Shaded areas represent 95% confidence intervals. HDL-C, high-density lipoprotein cholesterol; Non-Mex-Am, Non-Mexican American; Waist circumf, waist circumference. *(From H Beltrán-Sánchez et al: Prevalence and trends of metabolic syndrome in the adult U.S. population, 1999-2010. J Am Coll Cardiol 62:697, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

RISK FACTORS

Overweight/Obesity

The metabolic syndrome was first described in the early twentieth century; however, the worldwide overweight/obesity epidemic has recently been the force driving its increasing recognition. Central adiposity is a key feature of the syndrome, and the syndrome's prevalence reflects the strong relationship between waist circumference and increasing adiposity. However, despite the importance of obesity, patients who are of normal weight may also be insulin-resistant and may have the metabolic syndrome. This phenotype is particularly evident for populations in India, Southeast Asia, and Central America.

Sedentary Lifestyle

Physical inactivity and less cardiorespiratory fitness are a predictor of CVD events and the related risk of death. Many components of the metabolic syndrome are associated with a sedentary lifestyle, including increased adipose tissue (predominantly central), reduced HDL cholesterol, and increased triglycerides, blood pressure, and glucose in genetically susceptible persons. Compared with individuals who watch television or videos or use the computer <1 h daily, those who do so for >4 h daily have a twofold increased risk of the metabolic syndrome.

Genetics

No single gene explains the complex phenotype called the metabolic syndrome. However, using genome wide association and candidate gene approaches, a number of genetic variants are associated with the metabolic syndrome. Although many of the loci have unknown function, many others relate to body weight and composition, insulin resistance, and lipid and lipoprotein metabolism.

Aging

The metabolic syndrome affects nearly 50% of the U.S. population aged >60, and at >60 years of age women are more often affected. The age dependency of the syndrome's prevalence is seen in most populations around the world.

Diabetes Mellitus

Diabetes mellitus can be included in both the NCEP and the harmonizing definitions of the metabolic syndrome, but the greatest value of the metabolic syndrome, and especially fasting glucose, is predicting type 2 diabetes. The great majority (~75%) of patients with type 2 diabetes or impaired glucose tolerance have the metabolic syndrome. The presence of the metabolic syndrome in these populations relates to a higher prevalence of CVD than in patients who have type 2 diabetes or impaired glucose tolerance but do not have the syndrome.

Cardiovascular Disease

Individuals with the metabolic syndrome are twice as likely to die of CVD as those who do not, and their risk of an acute myocardial infarction or stroke is threefold higher. The approximate prevalence of the metabolic syndrome among patients with coronary heart disease (CHD) is 60%, with a prevalence of ~35% among patients with premature coronary artery disease (≤age 45) and a particularly high prevalence among women. With appropriate cardiac rehabilitation and changes in lifestyle (e.g., nutrition, physical activity, weight reduction, and—in some cases—pharmacologic therapy), the prevalence of the syndrome can be reduced.

Lipodystrophy

Lipodystrophic disorders in general are associated with the metabolic syndrome. Moreover, it is quite common for such patients to present with the metabolic syndrome. Both genetic lipodystrophy (e.g., Berardinelli-Seip congenital lipodystrophy, Dunnigan familial partial lipodystrophy) and acquired lipodystrophy (e.g., HIV-related lipodystrophy and in HIV patients receiving certain antiretroviral therapies) may give rise to severe insulin resistance and many of the components of the metabolic syndrome.

ETIOLOGY

Insulin Resistance

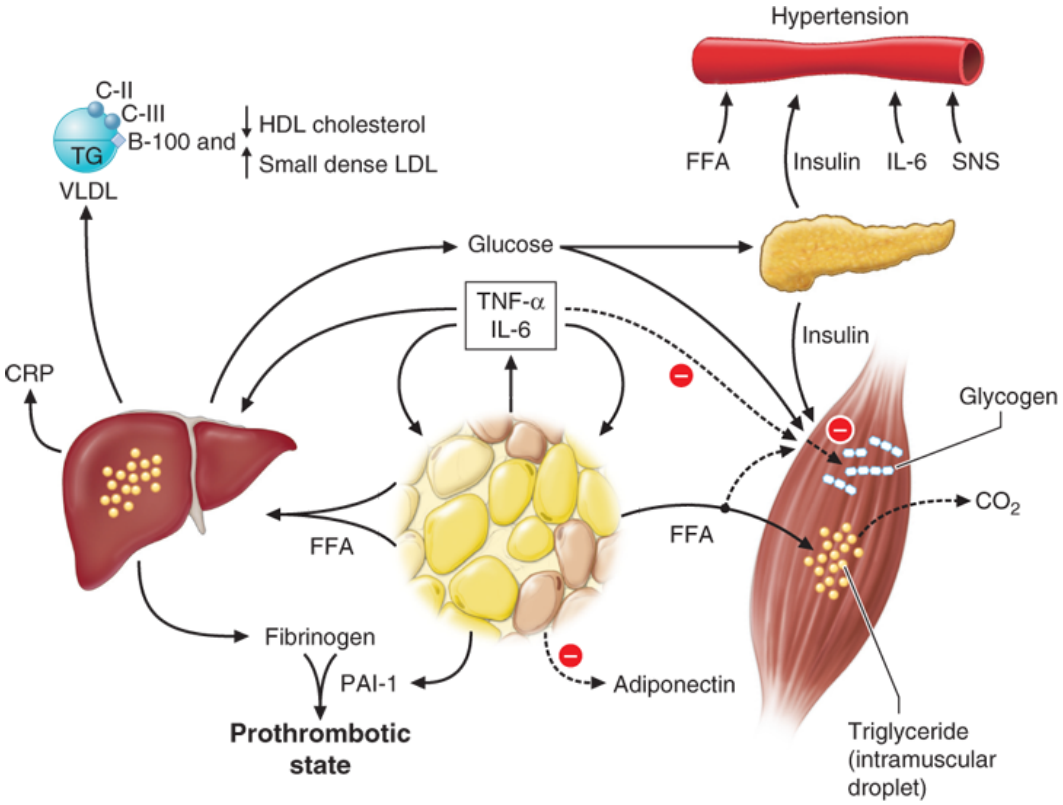
The most accepted and unifying hypothesis to describe the pathophysiology of the metabolic syndrome is insulin resistance, caused systemically by an incompletely understood defect in insulin action **(Chap. 396)**. The onset of insulin resistance is heralded by postprandial hyperinsulinemia, which is followed by fasting hyperinsulinemia and ultimately by hyperglycemia.

An early major contributor to the development of insulin resistance is an overabundance of circulating fatty acids **(Fig. 401-2)**. Plasma albumin-bound free fatty acids are derived predominantly from adipose-tissue triglyceride stores released by intracellular lipolytic enzymes. The lipolysis of triglyceride-rich lipoproteins in tissues by lipoprotein lipase also produces free fatty acids. Insulin mediates both anti-lipolysis and the stimulation of lipoprotein lipase in adipose tissue. Of note, the inhibition of lipolysis in adipose tissue is the most sensitive pathway of insulin action. Thus, when insulin resistance develops, increased lipolysis produces more fatty acids, which further decrease the anti-lipolytic effect of insulin. Excessive fatty acids enhance substrate availability and create insulin resistance by modifying downstream signaling. Fatty acids impair insulin-mediated glucose uptake and accumulate as triglycerides in both skeletal and cardiac muscle, whereas increased fatty acid flux increases glucose production and triglyceride production and accumulation in the liver.

FIGURE 401-2

Pathophysiology of the metabolic syndrome. Free fatty acids (FFAs) are released in abundance from an expanded adipose tissue mass. In the liver, FFAs result in increased production of glucose and triglycerides and secretion of very low-density lipoproteins (VLDLs). Associated lipid/lipoprotein abnormalities include reductions in high-density lipoprotein (HDL) cholesterol and an increased low-density lipoprotein (LDL) particle number (no.). FFAs also reduce insulin sensitivity in muscle by inhibiting insulin-mediated glucose uptake. Associated defects include a reduction in glucose partitioning to glycogen and increased lipid accumulation in triglyceride (TG). The increase in circulating glucose, and to some extent FFAs, increases pancreatic insulin secretion, resulting in hyperinsulinemia. Hyperinsulinemia may result in enhanced sodium reabsorption and increased sympathetic nervous system (SNS) activity and contribute to hypertension, as might higher levels of circulating FFAs. The pro-inflammatory state is superimposed and contributory to the insulin resistance produced by excessive FFAs. The enhanced secretion of interleukin 6 (IL-6) and tumor necrosis factor α (TNF-α) produced by adipocytes and monocyte-derived macrophages results in more insulin resistance and lipolysis of adipose tissue triglyceride stores to circulating FFAs. IL-6 and other cytokines also enhance hepatic glucose production, VLDL production by the liver, hypertension and insulin resistance in muscle. Insulin resistance also contributes to increased triglyceride accumulation in the liver. Cytokines and FFAs also increase hepatic production of fibrinogen and adipocyte production of plasminogen activator inhibitor 1 (PAI-1), resulting in a pro-thrombotic state. Higher levels of circulating

cytokines stimulate hepatic production of C-reactive protein (CRP). Reduced production of the anti-inflammatory and insulin-sensitizing cytokine adiponectin is also associated with the metabolic syndrome. *(Modified from RH Eckel et al: Lancet 365:1415, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Leptin resistance also may be a pathophysiologic mechanism to explain the metabolic syndrome. Physiologically, leptin reduces appetite, promotes energy expenditure, and enhances insulin sensitivity. In addition, leptin may regulate cardiac and vascular function through a nitric oxide–dependent mechanism. However, when obesity develops, hyperleptinemia ensues, with evidence of leptin resistance in the brain and other tissues resulting in inflammation, insulin resistance, hyperlipidemia, and a plethora of cardiovascular disorders, such as hypertension, atherosclerosis, CHD, and heart failure.

The oxidative stress hypothesis provides a unifying theory for aging and the predisposition to the metabolic syndrome. In studies of insulin-resistant individuals with obesity, type 2 diabetes, the offspring of patients with type 2 diabetes, and the elderly, a defect in mitochondrial oxidative phosphorylation leads to the accumulation of triglycerides and related lipid molecules in muscle, liver and perhaps other tissues, i.e., β-cells.

Recently, the gut microbiome has emerged as an important contributor to the development of obesity and related metabolic disorders, including the metabolic syndrome. Although the mechanisms remain uncertain, interaction among genetic predisposition, diet, bile acid metabolism, and the intestinal flora is important.

Increased Waist Circumference

Waist circumference is an important component of the most recent and frequently applied diagnostic criteria for the metabolic syndrome. However, measuring waist circumference does not reliably distinguish increases in adipose tissue from that in visceral fat; this distinction requires CT or MRI. With increases in visceral adipose tissue, adipose tissue–derived free fatty acids reach the liver. In contrast, increases in abdominal SC fat release lipolysis products into the systemic circulation and therefore have fewer effects on hepatic metabolism. Relative increases in visceral versus SC adipose tissue with increasing waist circumference in Asians and Asian Indians may explain the greater prevalence of the syndrome in those populations than in African-American men, in whom SC fat predominates. It is also possible that visceral fat is a marker for—but not the source of—excess postprandial free fatty acids in obesity.

Dyslipidemia

In general, free fatty acid flux to the liver results in increased production of apoB-containing, triglyceride-rich, very low-density lipoproteins (VLDLs) **(See also Chap. 400)**. The effect of insulin on this process is complex, but *hypertriglyceridemia* is an excellent marker of the insulin-resistant condition. Not only is hypertriglyceridemia a feature of the metabolic syndrome, but patients with the metabolic syndrome have elevated levels of apoC-III

carried on VLDLs and other lipoproteins. This increase in apoC-III is inhibitory to lipoprotein lipase, further contributing to hypertriglyceridemia, and confers more risk for atherosclerotic cardiovascular disease (ASCVD).

The other major lipoprotein disturbance in the metabolic syndrome is a *reduction in HDL cholesterol*. This reduction is a consequence of changes in HDL composition and metabolism. In the presence of hypertriglyceridemia, a decrease in the cholesterol content of HDL is a consequence of reduced cholesteryl ester content of the lipoprotein core in combination with cholesteryl ester transfer protein–mediated alterations in triglycerides that make the particle small and dense. This change in lipoprotein composition also results in increased clearance of HDL from the circulation. These changes in HDL have a relationship to insulin resistance that is probably indirect, occurring in concert with the changes in triglyceride-rich lipoprotein metabolism.

In addition to HDLs, low-density lipoproteins (LDLs) have alterations in composition in the metabolic syndrome. With fasting serum triglycerides at >2.0 mM (~180 mg/dL), there is usually a predominance of small, dense LDLs, which are thought to be more atherogenic although their association with hypertriglyceridemia and low HDLs make their independent contribution to ASCVD events difficult to assess. Individuals with hypertriglyceridemia often have increases in cholesterol content of both VLDL1 and VLDL2 sub-fractions and in LDL particle number. Both of these lipoprotein changes may contribute to atherogenic risk in patients with the metabolic syndrome.

Glucose Intolerance

Defects in insulin action in the metabolic syndrome lead to impaired suppression of glucose production by the liver (and kidney) and reduced glucose uptake and metabolism in insulin-sensitive tissues—i.e., muscle and adipose tissue **(See also Chap. 396)**. There is a strong relationship between impaired fasting glucose or impaired glucose tolerance and insulin resistance in studies of humans, nonhuman primates, and rodents. To compensate for defects in insulin action, insulin secretion and/or clearance increases or decreases, respectively, so that euglycemia remains. Ultimately, this compensatory mechanism fails because of defects in insulin secretion, resulting in progression from impaired fasting glucose and/or impaired glucose tolerance to type 2 diabetes mellitus.

Hypertension

The relationship between insulin resistance and hypertension is well established. Paradoxically, under normal physiologic conditions, insulin is a vasodilator with secondary effects on sodium reabsorption in the kidney. However, in the setting of insulin resistance, the vasodilatory effect of insulin is lost but the renal effect on sodium reabsorption is preserved. Sodium reabsorption is increased in Caucasians with the metabolic syndrome but not in Africans or Asians. Insulin also increases the activity of the sympathetic nervous system, an effect that is preserved in the setting of insulin resistance. Insulin resistance is also associated with pathway-specific impairment in phosphatidylinositol-3-kinase signaling. In the endothelium, this impairment may cause an imbalance between the production of nitric oxide and the secretion of endothelin 1, with a consequent decrease in blood flow. In addition, increases in angiotensinogen gene expression in adipose tissue of obese subjects results in increases in circulating angiotensin II and vasoconstriction. Although these mechanisms are provocative, the inadequate evaluation of insulin action by measurement of fasting insulin levels or by homeostasis model assessment shows that insulin resistance contributes only partially to the increased prevalence of hypertension in the metabolic syndrome.

Another possible mechanism underlying hypertension in the metabolic syndrome is the vasoactive role of perivascular adipose tissue. Reactive oxygen species released by NADPH oxidase impair endothelial function and result in local vasoconstriction. Other paracrine effects such as leptin or other pro-inflammatory cytokines released from adipose tissue, such as TNF-α may also be important.

Hyperuricemia is another consequence of insulin resistance in the metabolic syndrome. There is growing evidence not only that uric acid is associated with hypertension but also that reduction of uric acid normalizes blood pressure in hyperuricemic adolescents with hypertension. The mechanism appears to be in part related to an adverse effect of uric acid on nitric acid synthase in the macula densa of the kidney and stimulation of the renin-angiotensin aldosterone system.

Pro-Inflammatory Cytokines

The increases in pro-inflammatory cytokines—including interleukins 1, 6, and 18; resistin; TNF-α; and the systemic biomarker C-reactive protein—reflect overproduction by the expanded adipose tissue mass (Fig. 401-2). Adipose tissue–derived macrophages may be the primary source of pro-inflammatory cytokines locally and in the systemic circulation. It remains unclear, however, how much of the insulin resistance is caused by the paracrine effects of these cytokines and how much by the endocrine effects.

Adiponectin

Adiponectin is an anti-inflammatory cytokine produced exclusively by adipocytes. Adiponectin enhances insulin sensitivity and inhibits many steps in the inflammatory process. In the liver, adiponectin inhibits the expression of gluconeogenic enzymes and the rate of glucose production. In muscle, adiponectin increases glucose transport and enhances fatty acid oxidation, partially through the activation of AMP kinase. Reductions in adiponectin

levels are common in the metabolic syndrome. The relative contributions of adiponectin deficiency and overabundance of the pro-inflammatory cytokines are unclear.

CLINICAL FEATURES

Symptoms and Signs

The metabolic syndrome typically is not associated with symptoms. On physical examination, waist circumference and blood pressure are often elevated. The presence of either or both of these signs should prompt the clinician to search for other biochemical abnormalities that may be associated with the metabolic syndrome. Less frequently, lipoatrophy or acanthosis nigricans are present on examination. Because these physical findings characteristically are associated with severe insulin resistance, other components of the metabolic syndrome are much more common.

Associated Diseases

CARDIOVASCULAR DISEASE

The relative risk for new-onset CVD in patients with the metabolic syndrome who do not have diabetes averages 1.5–3 fold. However, in INTERHEART, a study of 26,903 subjects from 52 countries, the risk for acute myocardial infarction in subjects with the metabolic syndrome (WHO or IDF definition) is comparable to that conferred by some, but not all, of the component risk factors. Diabetes mellitus (OR: 2.72) and hypertension (OR: 2.60) are stronger than other risk factors. Although congestive heart failure and the metabolic syndrome can occur together, typically this consequence is secondary to metabolic syndrome-related ASCVD or hypertension. Metabolic syndrome is also associated with increases in the risk for stroke, peripheral vascular disease, and Alzheimer's disease. However, as for myocardial infarction, the risk beyond the additive role of the components of the metabolic syndrome remains debatable. In the Reasons for Geographic and Racial Differences in Stroke (REGARDS) cohort, an observational study of black and white adults ≥45 years old across the United States, there were 9741 participants and 41% had the metabolic syndrome. After adjustment for multiple confounders, metabolic syndrome was associated with increases in high sensitivity C-reactive protein (hsCRP), and this relationship was associated with a 1.34 relative risk for all-cause mortality; but <50% of deaths were from CVD. The population attributable risk was 9.5% for the metabolic syndrome alone and 14.7% for both metabolic syndrome and increased hsCRP. The relationship metabolic syndrome and hsCRP to mortality was greater for whites than blacks.

TYPE 2 DIABETES

Overall, the risk for type 2 diabetes among patients with the metabolic syndrome is increased three- to fivefold. In the Framingham Offspring Study's 8-year follow-up of middle-aged participants, the population-attributable risk of the metabolic syndrome for developing type 2 diabetes was 62% among men and 47% among women, yet increases in fasting plasma glucose explained most, if not all, of this increased risk.

Other Associated Conditions

In addition to the features specifically used to define the metabolic syndrome, other metabolic alterations are secondary to, or accompany insulin resistance. Those alterations include increases in apoB and apoC-III, uric acid, pro-thrombotic factors (fibrinogen, plasminogen activator inhibitor 1), serum viscosity, asymmetric dimethylarginine, homocysteine, white blood cell count, pro-inflammatory cytokines, C-reactive protein, increased urine albumin/creatinine ratio, non-alcoholic fatty liver disease (NAFLD), and/or non-alcoholic steatohepatitis NASH), polycystic ovary syndrome, and obstructive sleep apnea.

NONALCOHOLIC FATTY LIVER DISEASE

NAFLD has become the most common liver disease, in part a consequence of the insulin resistance of the metabolic syndrome. The mechanism relates to increases in free fatty acid flux, reductions in intrahepatic fatty acid oxidation with resultant increases in triglyceride biosynthesis and hepatocellular accumulation, with variable inflammation and oxidative stress. The more serious NASH, a consequence of NAFLD in some patients and precursor of cirrhosis and end stage liver disease, includes a more substantial pro-inflammatory contribution. NASH is now present in 3–5% of the U.S. population and other Western countries. Of patients with the metabolic syndrome, ~25–60% have NAFLD and up to 35% have NASH. As the prevalence of overweight/obesity and the metabolic syndrome increases, NASH may become one of the more common causes of end-stage liver disease and hepatocellular carcinoma.

HYPERURICEMIA

Hyperuricemia reflects defects in insulin action on the renal tubular reabsorption of uric acid and may contribute to hypertension through its effect on the endothelium **(See also Chap. 410)**. An increase in asymmetric dimethylarginine, an endogenous inhibitor of nitric oxide synthase, also relates to

endothelial dysfunction. In addition, increases in urine albumin/creatinine ratio may relate to altered endothelial pathophysiology in the insulin-resistant state.

POLYCYSTIC OVARY SYNDROME

Polycystic ovary syndrome is highly associated with insulin resistance (50–80%) and the metabolic syndrome, with a prevalence of the syndrome between 40 and 50% **(See also Chap. 385)**. Women with polycystic ovary syndrome are 2–4 times more likely to have the metabolic syndrome than are women without polycystic ovary syndrome.

OBSTRUCTIVE SLEEP APNEA

Obstructive sleep apnea is commonly associated with obesity, hypertension, increased circulating cytokines, impaired glucose tolerance, and insulin resistance **(See also Chap. 27)**. In fact, obstructive sleep apnea may predict metabolic syndrome, even in the absence of excess adiposity. Moreover, when biomarkers of insulin resistance are compared between patients with obstructive sleep apnea and weight-matched controls, insulin resistance is found to be more severe in those with apnea. Continuous positive airway pressure treatment improves insulin sensitivity in patients with obstructive sleep apnea.

DIAGNOSIS

The diagnosis of the metabolic syndrome relies on fulfillment of the criteria listed in Table 401-1, as assessed using tools at the bedside and in the laboratory. The medical history should include evaluation of symptoms for obstructive sleep apnea in all patients and polycystic ovary syndrome in premenopausal women. Family history will help determine risk for CVD and diabetes mellitus. Blood pressure and waist circumference measurements provide information necessary for the diagnosis.

Laboratory Tests

Measurement of fasting lipids and glucose is needed in determining whether the metabolic syndrome is present. The measurement of additional biomarkers associated with insulin resistance can be individualized. Such tests might include those for apoB, hsCRP, fibrinogen, uric acid, urinary albumin/creatinine ratio, and liver function. A sleep study should be performed if symptoms of obstructive sleep apnea are present. If polycystic ovary syndrome is suspected based on clinical features and anovulation, testosterone, luteinizing hormone, and follicle-stimulating hormone should be measured.

TREATMENT

TREATMENT

The Metabolic Syndrome

LIFESTYLE

Obesity, particularly abdominal, is the driving force behind the metabolic syndrome **(See also Chap. 395)**. Thus, weight reduction is the primary approach to the disorder. With at least a 5% and more so with 10% weight reduction, improvement in insulin sensitivity results in favorable modifications in many components of the metabolic syndrome. In general, recommendations for weight loss include a combination of caloric restriction, increased physical activity, and behavior modification. Caloric restriction is the most important component, whereas increases in physical activity are important for maintenance of weight loss. Some but not all evidence suggests that the addition of exercise to caloric restriction may promote greater weight loss from the visceral depot. The tendency for weight regain after successful weight reduction underscores the need for long-lasting behavioral changes.

Diet

Before prescribing a weight-loss diet, it is important to emphasize that it has taken the patient a long time to develop an expanded fat mass; thus, the correction need not occur quickly. Given that ~3500 kcal = 1 lb. of fat, ~500-kcal restriction daily equates to weight reduction of 1 lb. per week. Diets restricted in carbohydrate typically provide a more rapid initial weight loss. However, after 1 year, the amount of weight reduction is minimally more reduced or no different from that with caloric restriction alone. Thus, adherence to the diet is more important than the chosen diet. Moreover, there is concern about low-carbohydrate diets enriched in saturated fat, particularly for patients at risk for ASCVD. Therefore, a high-quality dietary pattern—i.e., a diet enriched in fruits, vegetables, whole grains, lean poultry, and fish—should be encouraged to maximize overall health benefit.

Physical Activity

Before prescribing a physical activity program to patients with the metabolic syndrome, it is important to ensure that the increased activity does not incur risk. Some high-risk patients should undergo formal cardiovascular evaluation before initiating an exercise program. For an inactive participant, gradual increases in physical activity should be encouraged to enhance adherence and avoid injury. Although increases in physical activity can lead to

modest weight reduction, 60–90 min of daily activity is required to achieve this goal. Even if an overweight or obese adult is unable to undertake this level of activity, a significant health benefit will follow from at least 30 min of moderate-intensity activity daily. The caloric value of 30 min of a variety of activities can be found at *www.heart.org/HEARTORG/GettingHealthy/WeightManagement/LosingWeight/Losing-Weight_UCM_307904_Article.jsp*. Of note, a variety of routine activities, such as gardening, walking, and housecleaning, require moderate caloric expenditure. Thus, physical activity should not be defined solely in terms of formal exercise such as jogging, swimming, or tennis.

Behavior Modification

Behavioral treatment typically includes recommendations for dietary restriction and more physical activity that predicts sufficient weight loss that benefits metabolic health. The subsequent challenge is the duration of the program because weight regain so often follows successful weight reduction. Improved long-term outcomes often follow a variety of methods, such as a personal or group counselor, the Internet, social media, and telephone follow-up to maintain contact between providers and patients.

Obesity

In some patients with the metabolic syndrome, treatment options need to extend beyond lifestyle intervention **(See also Chap. 395)**. Weight-loss drugs come in two major classes: appetite suppressants and absorption inhibitors. Appetite suppressants approved by the U.S. Food and Drug Administration include phentermine (for short-term use [3 months] only) as well as the more recent additions phentermine/topiramate, lorcaserin, naltrexone/bupropion and high dose (3.0 mg) liraglutide (rather than 1.8 mg, maximum for treatment of type 2 diabetes), which are approved without restrictions on the duration of therapy. In clinical trials, the phentermine/topiramate extended release combination has resulted in ~8% weight loss relative to placebo in 50% of patients. Side effects include palpitations, headache, paresthesias, constipation, and insomnia. Lorcaserin results in less weight loss—typically ~5% beyond placebo—but can cause headache and nasopharyngitis. Naltrexone/bupropion extended release reduces body weight ≥10% in ~20% of patients, however, the drug combination is contraindicated in patients with seizure disorders or any condition that predisposes to seizures. Naltrexone/bupropion also increases pulse and blood pressure and should not be given to patients with uncontrolled hypertension. High dose liraglutide results in ~6% weight loss relative to placebo with ~33% of patients with >10% weight loss. Common side effects are limited to the upper gastrointestinal tract, including nausea, and less frequently, emesis.

Orlistat inhibits fat absorption by ~30% and is moderately effective compared with placebo (~4% more weight loss). Moreover, orlistat reduced the incidence of type 2 diabetes, an effect that was especially evident among patients with impaired glucose tolerance at baseline. This drug is often difficult to take because of oily leakage per rectum. In general, for all weight loss drugs, greater weight reduction leads to greater improvement in metabolic syndrome components, including the conversion from prediabetes to type 2 diabetes.

Metabolic or bariatric surgery is an option for patients with the metabolic syndrome who have a body mass index >40 kg/m^2, or >35 kg/m^2 with comorbidities. An evolving application for metabolic surgery includes patients with a body mass index as low as 30 kg/m^2 and type 2 diabetes. Gastric bypass or vertical sleeve gastrectomy results in dramatic weight reduction and improvement in most features of the metabolic syndrome. A survival benefit with gastric bypass has also been realized.

LDL CHOLESTEROL

The rationale for the NCEP: ATPIII's development of criteria for the metabolic syndrome was to go beyond LDL cholesterol in identifying and reducing the risk of ASCVD **(See also Chap. 400)**. The working assumption by the panel was that LDL cholesterol goals had already been achieved and that increasing evidence supports a linear reduction in ASCVD events because of progressive lowering of LDL cholesterol with statins. The 2013 ACC/AHA Cholesterol Guidelines have no specific recommendations for patients with the metabolic syndrome; however, a statin should be prescribed in all patients with diabetes age 40–79 with an LDL cholesterol between 60 and 189 mg/dL. For those patients with diabetes and known ASCVD, the current evidence supports a high intensity-statin dose (e.g., atorvastatin 40–80 mg or rosuvastatin 20–40 mg daily). For those patients with the metabolic syndrome but without diabetes, the 10–year ASCVD risk estimator should be employed and a risk ≥7.5% should lead to a discussion between provider and patient about initiating statin therapy for primary prevention of ASCVD.

Diets restricted in saturated fats (<6% of calories) and *trans*-fats (as few as possible) should be applied aggressively. Although evidence is controversial, dietary cholesterol can also be restricted. If LDL cholesterol remains elevated, pharmacologic intervention is needed. Based on substantial evidence, treatment with statins, which lower LDL cholesterol by 15–60%, is the first-choice medication intervention. Of note, for each doubling of the statin dose, LDL cholesterol is further lowered by only ~6%. Hepatotoxicity (more than a threefold increase in hepatic aminotransferases) is rare, but myopathy occurs in ~10–20% of patients. The cholesterol absorption inhibitor ezetimibe is well tolerated and should be the second-choice medication intervention. Ezetimibe typically reduces LDL cholesterol by 15–20%. Proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibitors are potent LDL cholesterol lowering drugs (~45–60%) but are not needed for most patients with the metabolic syndrome. Of course, if these patients also have familial hypercholesterolemia, or insufficient LDL cholesterol lowering on statins ± ezetimibe, a PCSK9 inhibitor should be considered. The bile acid sequestrants cholestyramine, colestipol, and colesevalam may be more effective than ezetimibe but, because they can increase triglyceride levels, must be used with caution in patients with the metabolic syndrome when fasting triglycerides are >300 mg/dL. Side effects include gastrointestinal symptoms (palatability, bloating, belching, constipation, anal irritation). Nicotinic acid has similar LDL cholesterol–lowering capabilities (<20%).

Fibrates are best employed to lower LDL cholesterol when triglycerides are not elevated. Fenofibrate may be more effective than gemfibrozil in this setting.

TRIGLYCERIDES

The NCEP: ATPIII focused on non-HDL cholesterol rather than on triglycerides whereas the 2013 ACA/AHA Cholesterol Guidelines stated that fasting triglycerides >500 mg/dL should be treated to prevent more serious hypertriglyceridemia and pancreatitis **(See also Chap. 400)**. Although a fasting triglyceride value of >150 mg/dL is a component of the metabolic syndrome, post hoc analyses of multiple fibrate trials have suggested reduction in the primary ASCVD outcome in patients (with or without concomitant statin therapy) with fasting triglycerides >200 mg/dL, often in the setting of reduced levels of HDL cholesterol.

A fibrate (gemfibrozil or fenofibrate) is the drug of choice to lower fasting triglyceride levels, which are typically reduced by 30–45%. Concomitant administration with drugs metabolized by the 3A4 cytochrome P450 system (including some statins) increases the risk of myopathy. In these cases, fenofibrate may be preferable to gemfibrozil. In the Veterans Affairs HDL Intervention Trial, gemfibrozil was administered to men with known CHD and levels of HDL cholesterol <40 mg/dL. A coronary disease event and mortality rate benefit was experienced predominantly among men with hyperinsulinemia and/or diabetes, many of whom were identified retrospectively as having the metabolic syndrome. Of note, the degree of triglyceride lowering in this trial did not predict benefit. Although levels of LDL cholesterol did not change, a decrease in LDL particle number correlated with benefit.

Other drugs that lower triglyceride levels include statins, nicotinic acid, and prescription omega-3 fatty acids. For this purpose, an intermediate or high dose of the “more potent” statins (atorvastatin, rosuvastatin) is needed. The effect of nicotinic acid on fasting triglycerides is dose related and ~20–35%, an effect that is less pronounced than that of fibrates. In patients with the metabolic syndrome and diabetes, nicotinic acid may increase fasting glucose levels and clinical trials with nicotinic acid + statin have failed to reduce ASCVD events. Prescriptions of omega-3 fatty acid preparations that include high doses of eicosapentaenoic acid ± docosahexaenoic acid (~1.5–4.5 g/d) lower fasting triglyceride levels by ~25–40%. Here, no drug interactions with fibrates or statins occur, and the main side effect of their use is eructation with a fishy taste. Freezing the nutraceutical can partially block this unpleasant side effect. Clinical trials of high-dose omega-3 fatty acids in patients with and without the metabolic syndrome are ongoing.

HDL CHOLESTEROL

Very few lipid-modifying compounds increase HDL cholesterol levels **(See also Chap. 400)**. Statins, fibrates, and bile acid sequestrants have modest effects (5–10%), whereas ezetimibe and omega-3 fatty acids have no effect. Nicotinic acid is the only currently available drug with predictable HDL cholesterol-raising properties. The response is dose related, and nicotinic acid can increase HDL cholesterol by up to 30% above baseline. After several trials of nicotinic acid versus placebo in statin-treated patients, there is no evidence that raising HDL cholesterol with nicotinic acid beneficially affects ASCVD events in patients with or without the metabolic syndrome.

BLOOD PRESSURE

The direct relationship between blood pressure and all-cause mortality rate has been well established in studies comparing patients with hypertension (>140/90 mmHg), patients with pre-hypertension (>120/80 mmHg but <140/90 mmHg), and individuals with normal blood pressure (<120/80 mmHg) **(See also Chap. 271)**. In patients who have the metabolic syndrome without diabetes, the best choice for the initial antihypertensive medication is an angiotensin-converting enzyme (ACE) inhibitor or an angiotensin II receptor blocker, as these two classes of drugs are effective and well tolerated. In all patients with hypertension, a sodium-restricted dietary pattern enriched in fruits and vegetables, whole grains, and low-fat dairy products should be advocated. Home monitoring of blood pressure may assist in maintaining good blood-pressure control.

IMPAIRED FASTING GLUCOSE

In patients with the metabolic syndrome and type 2 diabetes, aggressive glycemic control may favorably modify fasting levels of triglycerides and/or HDL cholesterol **(See also Chap. 396)**. In patients with impaired fasting glucose who do not have diabetes, a lifestyle intervention that includes weight reduction, dietary saturated fat restriction, and increased physical activity has been shown to reduce the incidence of type 2 diabetes. Metformin also reduces the incidence of diabetes, although the effect is less pronounced than that of lifestyle intervention.

INSULIN RESISTANCE

Several drug classes (biguanides, thiazolidinediones [TZDs]) increase insulin sensitivity **(See also Chap. 397)**. Because insulin resistance is the primary pathophysiologic mechanism for the metabolic syndrome, representative drugs in these classes reduce its prevalence. Both metformin and TZDs enhance insulin action in the liver and suppress endogenous glucose production. TZDs, but not metformin, also improve insulin-mediated glucose uptake in muscle and adipose tissue. In a meta-analysis of nine trials involving 12,026 participants, the TZD pioglitazone versus placebo was associated with reduction in ASCVD events in patients with insulin resistance (metabolic syndrome), prediabetes and type 2 diabetes. However, adverse effects including weight gain, bone fracture, and congestive heart failure with/or without edema were seen. Benefit of TZDs has been seen in patients with NAFLD, and with metformin in women with polycystic ovary syndrome, and both drug classes have been shown to reduce markers of inflammation.

FURTHER READING

Alberti KG et al: Harmonizing the metabolic syndrome: A joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation 120:1640, 2009. [PubMed: 19805654]

Beltran-Sanchez H et al: Prevalence and trends of metabolic syndrome in the adult U.S. population, 1999–2010. J Am Coll Cardiol 62:693, 2013.

Brown AE, Walker M: Genetics of insulin resistance and the metabolic syndrome. Curr Cardiol Rep 18:75, 2016. [PubMed: 27312935]

Eckel RH et al: The metabolic syndrome. Lancet 365:1415, 2005. [PubMed: 15836891]

Ferland A, Eckel RH: Does sustained weight loss reverse the metabolic syndrome? Curr Hypertens Rep. 13:456, 2011. [PubMed: 21809180]

Genser L et al: Obesity, type 2 diabetes, and the metabolic syndrome: Pathophysiologic relationships and guidelines for surgical intervention. Surg Clin North Am 96:681, 2016. [PubMed: 27473795]

Grundy SM et al: Diagnosis and management of the metabolic syndrome: An American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. Circulation 112:2735, 2005. [PubMed: 16157765]

Lim S, Eckel RH: Pharmacological treatment and therapeutic perspective of metabolic syndrome. Rev Endocr Metab Disord 15:329, 2014. [PubMed: 25342235]

O'Neill S, O'Driscoll L: Metabolic syndrome: A closer look at the growing epidemic and its associated pathologies. Obes Rev 16:1, 2015. [PubMed: 25407540]

Tune JD et al: Cardiovascular consequences of metabolic syndrome. Transl Res 183:57, 2017. [PubMed: 28130064]

Chapter 402: Bone and Mineral Metabolism in Health and Disease

F. Richard Bringhurst; Marie B. Demay; Henry M. Kronenberg

BONE STRUCTURE AND METABOLISM

Bone is a dynamic tissue that is remodeled constantly throughout life. The arrangement of compact and cancellous bone provides strength and density suitable for both mobility and protection. Compact or cortical bone forms the roughly cylindrical shell of long bones; cancellous or trabecular bone forms the plate-like meshwork that internally supports the cortical shell. In addition, bone provides a reservoir for calcium, magnesium, phosphorus, sodium, and other ions necessary for homeostatic functions. Bone also hosts and regulates hematopoiesis by providing niches for hematopoietic cell proliferation and differentiation. The skeleton is highly vascular and receives about 10% of the cardiac output. Remodeling of bone is accomplished by two distinct cell types: osteoblasts produce bone matrix, and osteoclasts resorb the matrix. The activities of these cells are coordinated by osteocytes, long-lived regulatory cells embedded within bone matrix.

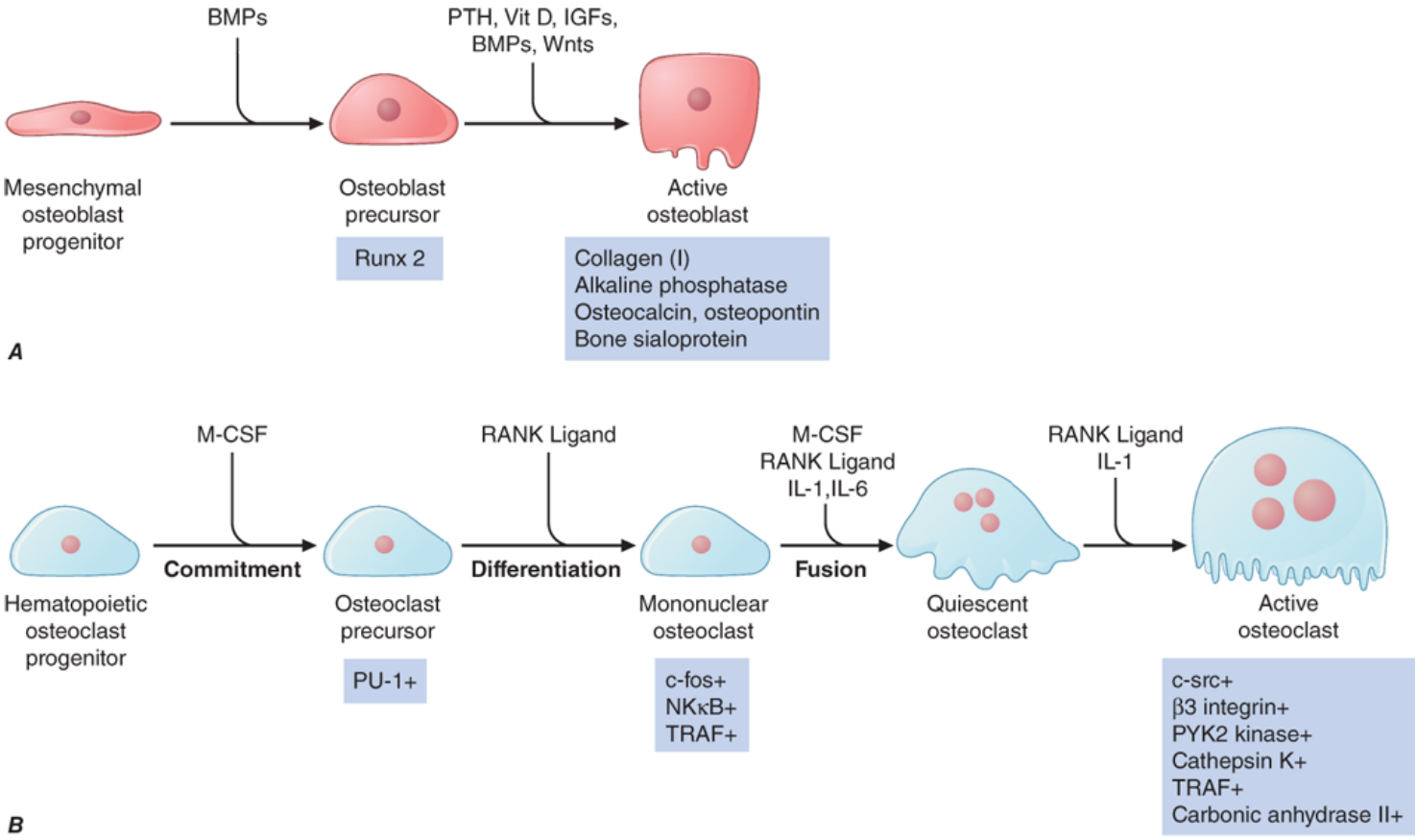
The extracellular components of bone consist of a solid mineral phase in close association with an organic matrix, of which 90–95% is type I collagen **(Chap. 406)**. The noncollagenous portion of the organic matrix is heterogeneous and contains serum proteins such as albumin as well as many locally produced proteins, whose functions are incompletely understood. Those proteins include cell attachment/signaling proteins such as thrombospondin, osteopontin, and fibronectin; calcium-binding proteins such as matrix gla protein and osteocalcin; and proteoglycans such as biglycan and decorin. Some of the proteins organize collagen fibrils; others influence mineralization and binding of the mineral phase to the matrix.

The mineral phase is made up of calcium and phosphate and is best characterized as a poorly crystalline hydroxyapatite. The mineral phase of bone is deposited initially in intimate relation to the collagen fibrils and is found in specific locations in the “holes” between the collagen fibrils. This architectural arrangement of mineral and matrix results in a two-phase material well suited to withstand mechanical stresses. The organization of collagen influences the amount and type of mineral phase formed in bone. Although the primary structures of type I collagen in skin and bone tissues are similar, there are differences in posttranslational modifications and distribution of intermolecular cross-links. The holes in the packing structure of the collagen are larger in mineralized collagen of bone and dentin than in unmineralized collagens such as those in tendon. Single amino acid substitutions in the helical portion of either the α1 (*COL1A1*) or α2 (*COL1A2*) chains of type I collagen disrupt the organization of bone in osteogenesis imperfecta. The severe skeletal fragility associated with this group of disorders highlights the importance of the fibrillar matrix in the structure of bone **(Chap. 406)**.

Osteoblasts synthesize and secrete the organic matrix and regulate its mineralization. They are derived from cells of mesenchymal origin **(Fig. 402-1A)**. Active osteoblasts are found on the surface of newly forming bone. As an osteoblast secretes matrix, which then is mineralized, the cell may become an *osteocyte*, still connected with its blood supply through a series of canaliculi. Osteocytes account for the vast majority of the cells in bone. They are thought to be the mechanosensors in bone that communicate signals to surface osteoblasts and osteoclasts and their progenitors through the canalicular network and thereby serve as master regulators of bone formation and resorption. Remarkably, osteocytes also secrete fibroblast growth factor 23 (FGF23), a major hormonal regulator of phosphate metabolism (see below). Mineralization of the matrix, both in trabecular bone and in osteones of compact cortical bone (*Haversian systems*), begins soon after the matrix is secreted (primary mineralization) but is not completed for several weeks or even longer (secondary mineralization). Although this mineralization takes advantage of the high concentrations of calcium and phosphate, already near saturation in serum, mineralization is a carefully regulated process that is dependent on the activity of osteoblast-derived alkaline phosphatase, which probably works by hydrolyzing inhibitors of mineralization.

FIGURE 402-1

Pathways regulating development of (A) osteoblasts and (B) osteoclasts. Hormones, cytokines, and growth factors that control cell proliferation and differentiation are shown above the arrows. Transcription factors and other markers specific for various stages of development are depicted below the arrows. BMPs, bone morphogenic proteins; wnts, wingless-type mouse mammary tumor virus integration site; PTH, parathyroid hormone; Vit D, vitamin D; IGFs, insulin-like growth factors; Runx2, Runt-related transcription factor 2; M-CSF, macrophage colony-stimulating factor; PU-1, a monocyte- and B lymphocyte–specific ets family transcription factor; NFκB, nuclear factor κB; TRAF, tumor necrosis factor receptor–associated factor; RANK ligand, receptor activator of NFκB ligand; IL-1, interleukin 1; IL-6, interleukin 6. *(Modified from T Suda et al: Endocr Rev 20:345, 1999, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Genetic studies in humans and mice have identified several key genes that control osteoblast development. *Runx2* is a transcription factor expressed specifically in chondrocyte (cartilage cells) and osteoblast progenitors as well as in hypertrophic chondrocytes and mature osteoblasts. *Runx2* regulates the expression of several important osteoblast proteins, including osterix (another transcription factor needed for osteoblast maturation), osteopontin, bone sialoprotein, type I collagen, osteocalcin, and receptor-activator of NFκB (RANK) ligand. *Runx2* expression is regulated in part by bone morphogenic proteins (BMPs). *Runx2*-deficient mice are devoid of osteoblasts, whereas mice with a deletion of only one allele (*Runx2*+/–) exhibit a delay in formation of the clavicles and some cranial bones. The latter abnormalities are similar to those in the human disorder *cleidocranial dysplasia*, which is also caused by heterozygous inactivating mutations in *Runx2*.

The paracrine signaling molecule, Indian hedgehog (Ihh), also plays a critical role in osteoblast development, as evidenced by Ihh-deficient mice that lack osteoblasts in the type of bone formed on a cartilage mold (endochondral ossification). Signals originating from members of the wnt (wingless-type mouse mammary tumor virus integration site) family of paracrine factors are also important for osteoblast proliferation and differentiation. Osteocytes regulate osteoblasts partly by secreting a potent inhibitor of wnt signaling called sclerostin. Numerous other growth-regulatory factors affect osteoblast function, including the three closely related transforming growth factor βs, fibroblast growth factors (FGFs) 2 and 18, platelet-derived growth factor, and insulin-like growth factors (IGFs) I and II. Hormones such as parathyroid hormone (PTH) and 1,25-dihydroxyvitamin D [1,25$(OH)_2$D] activate receptors expressed by osteoblasts to assure mineral homeostasis and influence a variety of bone cell functions.

Resorption of bone is carried out mainly by *osteoclasts*, multinucleated cells that are formed by fusion of cells derived from the common precursor of macrophages and osteoclasts. Thus, these cells derive from the hematopoietic lineage, quite different from the mesenchymal cells that become osteoblasts. Multiple factors that regulate osteoclast development have been identified **(Fig. 402-1B)**. Factors produced by osteocytes, osteoblasts and marrow stromal cells allow cells of the osteoblast lineage to control osteoclast development and activity. Macrophage colony-stimulating factor (M-CSF) plays a critical role during several steps in the pathway and ultimately leads to fusion of osteoclast progenitor cells to form multinucleated, active osteoclasts. RANK ligand, a member of the tumor necrosis factor (TNF) family, is expressed on the surface of osteocytes, osteoblasts, and stromal fibroblasts. In a process involving cell-cell interactions, RANK ligand binds to the RANK receptor on osteoclast progenitors, stimulating osteoclast differentiation and activation. Alternatively, a soluble decoy receptor, referred to as osteoprotegerin, can bind RANK ligand and inhibit osteoclast differentiation. Several growth factors and cytokines (including interleukins 1, 6, and 11; TNF; and interferon γ) modulate osteoclast differentiation and function. Most hormones that influence osteoclast function do not target these cells directly but instead target cells of the osteoblast lineage to increase production of M-CSF and RANK. Both PTH and 1,25$(OH)_2$D increase osteoclast number and activity by this indirect mechanism. Calcitonin, in contrast, binds to its receptor on the basal surface of osteoclasts and directly inhibits osteoclast function. Estradiol has multiple cellular targets in bone, including osteoclasts, immune cells, and osteoblasts; actions on all these cells serve to decrease osteoclast number and decreased bone resorption.

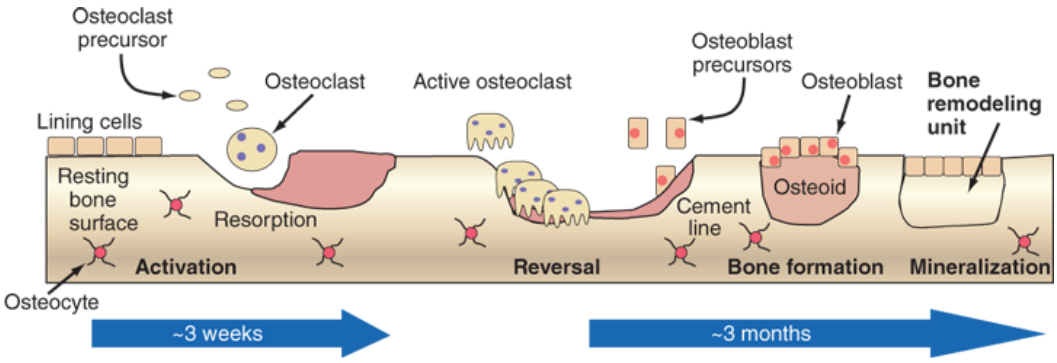
Osteoclast-mediated resorption of bone takes place in scalloped spaces (*Howship's lacunae*) where the osteoclasts are attached through a specific αvβ3 integrin to components of the bone matrix such as osteopontin. The osteoclast forms a tight seal to the underlying matrix and secretes protons, chloride, and proteinases into a confined space that has been likened to an extracellular lysosome. The active osteoclast surface forms a ruffled border that contains a specialized proton-pump ATPase that secretes acid that solubilizes the mineral phase. Carbonic anhydrase (type II isoenzyme) within the osteoclast generates the needed protons. The bone matrix is resorbed in the acid environment adjacent to the ruffled border by proteases, such as cathepsin K, that act at low pH.

In the embryo and the growing child, bone develops mostly by replacing previously calcified cartilage (endochondral bone formation) with subsequent remodeling, or, in a few bones, is formed without a cartilage matrix (intramembranous bone formation). During endochondral bone formation, chondrocytes proliferate, secrete and mineralize a matrix, enlarge (hypertrophy), and then die, enlarging bone and providing the matrix and factors that stimulate endochondral bone formation. This program is regulated by both local factors such as IGF-I and -II, Ihh, parathyroid hormone–related peptide (PTHrP), BMPs, and FGFs and by systemic hormones such as growth hormone, glucocorticoids, and estrogen.

New bone, whether formed in infants or in adults during repair, has a relatively high ratio of cells to matrix and is characterized by coarse fiber bundles of collagen that are interlaced and randomly dispersed (woven bone). In adults, the more mature bone is organized with fiber bundles regularly arranged in parallel or concentric sheets (lamellar bone). In long bones, deposition of lamellar bone in a concentric arrangement around blood vessels forms the Haversian systems. Growth in length of bones is dependent on proliferation of cartilage cells and the endochondral sequence at the growth plate. Growth in width and thickness is accomplished by formation of bone at the periosteal surface and by resorption at the endosteal surface, with the rate of formation exceeding that of resorption. In adults, after the growth plates of cartilage close through the actions of estrogen, growth in length and endochondral bone formation cease. Even in adults, however, remodeling of bone (within Haversian systems as well as along the surfaces of trabecular bone) continues throughout life. In adults, ~4% of the surface of trabecular bone (such as iliac crest) is involved in active resorption, whereas 10–15% of trabecular surfaces are covered with osteoid, unmineralized new bone formed by osteoblasts. Radioisotope studies indicate that as much as 18% of the total skeletal calcium is deposited and removed each year. Thus, bone is an active metabolizing tissue that requires an intact blood supply. The cycle of bone resorption and formation is a highly orchestrated process, directed by osteocytes and carried out by the basic multicellular unit, which is composed of a group of osteoclasts and osteoblasts **(Fig. 402-2)**.

FIGURE 402-2

Schematic representation of bone remodeling. The cycle of bone remodeling is carried out by the basic multicellular unit (BMU), which consists of a group of osteoclasts and osteoblasts. In cortical bone, the BMUs tunnel through the tissue, whereas in cancellous bone, they move across the trabecular surface. The process of bone remodeling is initiated by contraction of the lining cells and the recruitment of osteoclast precursors. These precursors fuse to form multinucleated, active osteoclasts that mediate bone resorption. Osteoclasts adhere to bone and subsequently remove it by acidification and proteolytic digestion. As the BMU advances, osteoclasts leave the resorption site and osteoblasts move in to cover the excavated area and begin the process of new bone formation by secreting osteoid, which eventually is mineralized into new bone. After osteoid mineralization, osteoblasts flatten and form a layer of lining cells over new bone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The response of bone to fractures, infection, and interruption of blood supply and to expanding lesions is relatively limited. Dead bone must be resorbed, and new bone must be formed, a process carried out in association with growth of new blood vessels into the involved area. In injuries that disrupt the organization of the tissue such as a fracture in which apposition of fragments is poor or when motion exists at the fracture site, progenitor stromal cells recapitulate the endochondral bone formation of early development and form cartilage that is replaced by bone and, variably, fibrous tissue. When there is good apposition with fixation and little motion at the fracture site, repair occurs predominantly by formation of new bone without other mediating tissue.

Remodeling of bone occurs along lines of force generated by mechanical stress. The signals from these mechanical stresses are sensed by osteocytes, which transmit signals to osteoclasts and osteoblasts or their precursors. One such signal made by osteocytes is sclerostin, an inhibitor of wnt signaling. Mechanical forces suppress sclerostin production and thus increase bone formation by osteoblasts. Expanding lesions in bone such as

tumors induce resorption at the surface in contact with the tumor by producing ligands such as PTHrP that stimulate osteoclast differentiation and function. Thus, bone plasticity reflects the interaction of cells with each other and with the environment.

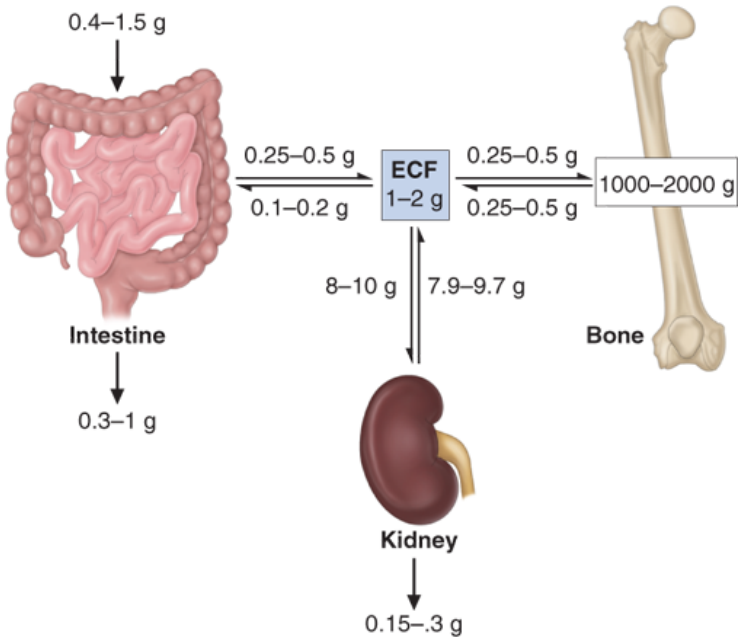
Measurement of the products of osteoblast and osteoclast activity can assist in the diagnosis and management of bone diseases. Osteoblast activity can be assessed by measuring serum bone-specific alkaline phosphatase. Similarly, osteocalcin, a protein secreted from osteoblasts, is made virtually only by osteoblasts. Osteoclast activity can be assessed by measurement of products of collagen degradation. Collagen molecules are covalently linked to each other in the extracellular matrix through the formation of hydroxypyridinium cross-links **(Chap. 406)**. After digestion by osteoclasts, these cross-linked peptides can be measured both in urine and in blood.

CALCIUM METABOLISM

Over 99% of the 1–2 kg of calcium present normally in the adult human body resides in the skeleton, where it provides mechanical stability and serves as a reservoir sometimes needed to maintain extracellular fluid (ECF) calcium concentration **(Fig. 402-3)**. Skeletal calcium accretion first becomes significant during the third trimester of fetal life, accelerates throughout childhood and adolescence, reaches a peak in early adulthood, and gradually declines thereafter at rates that rarely exceed 1–2% per year. These slow changes in total skeletal calcium content contrast with relatively high daily rates of closely matched fluxes of calcium into and out of bone (~250–500 mg each), a process mediated by coupled osteoblastic and osteoclastic activity. Another 0.5–1% of skeletal calcium is freely exchangeable (e.g., in chemical equilibrium) with that in the ECF.

FIGURE 402-3

Calcium homeostasis. Schematic illustration of calcium content of extracellular fluid (ECF) and bone as well as of diet and feces; magnitude of calcium flux per day as calculated by various methods is shown at sites of transport in intestine, kidney, and bone. Ranges of values shown are approximate and were chosen to illustrate certain points discussed in the text. In conditions of calcium balance, rates of calcium release from and uptake into bone are equal.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The concentration of ionized calcium in the ECF must be maintained within a narrow range because of the critical role calcium plays in a wide array of cellular functions, especially those involved in neuromuscular activity, secretion, and signal transduction. Intracellular cytosolic free calcium levels are ~100 nmol/L and are 10,000-fold lower than ionized calcium concentration in the blood and ECF (1.1–1.3 mmol/L). Cytosolic calcium does not play the structural role played by extracellular calcium; instead, it serves a signaling function. The steep chemical gradient of calcium from outside to inside the cell promotes rapid calcium influx through various membrane calcium channels that can be activated by hormones, metabolites, or neurotransmitters, swiftly changing cellular function. In blood, total calcium concentration is normally 2.2–2.6 m*M* (8.5–10.5 mg/dL), of which ~50% is ionized. The remainder is bound ionically to negatively charged proteins (predominantly albumin and immunoglobulins) or loosely complexed with phosphate, citrate, sulfate, or other anions. Alterations in serum protein concentrations directly affect the total blood calcium concentration even if the ionized calcium concentration remains normal. An algorithm to correct for protein changes adjusts the total serum calcium (in mg/dL) upward by 0.8 times the deficit in serum albumin (g/dL) or by 0.5 times the deficit in serum immunoglobulin (in g/dL). Such corrections provide only rough approximations of actual free calcium concentrations, however, and may be misleading, particularly during acute illness. Acidosis also alters ionized calcium by reducing its association with proteins. The best practice is to measure blood ionized calcium directly by a method that employs calcium-selective electrodes in acute settings during which calcium abnormalities might occur.

Control of the ionized calcium concentration in the ECF ordinarily is accomplished by adjusting the rates of calcium movement across intestinal and renal epithelia. These adjustments are mediated mainly via changes in blood levels of the hormones, PTH and $1,25(OH)_2D$. Blood ionized calcium directly suppresses PTH secretion by activating calcium-sensing receptors (CaSRs) in parathyroid cells. Also, ionized calcium indirectly affects PTH secretion by lowering $1,25(OH)_2D$ production. This active vitamin D metabolite inhibits PTH production by an incompletely understood mechanism of negative feedback (Chap. 403).

Normal dietary calcium intake in the United States varies widely, ranging from 10–37 mmol/d (400–1500 mg/d). A National Academy of Medicine (formerly, Institute of Medicine) analysis recommends a daily allowance of 25–30 mmol (1000–1200 mg) for most adults. Intestinal absorption of ingested calcium involves both active (transcellular) and passive (paracellular) mechanisms. Passive calcium absorption is nonsaturable and approximates 5% of daily calcium intake, whereas active absorption involves apical calcium entry via specific ion channels (TRPV5 and TRPV6), whose expression is controlled principally by $1,25(OH)_2D$, and normally ranges from 20 to 70%. Active calcium transport occurs mainly in the proximal small bowel (duodenum and proximal jejunum), although some active calcium absorption occurs in most segments of the small intestine. Optimal rates of calcium absorption require gastric acid. This is especially true for weakly dissociable calcium supplements such as calcium carbonate. In fact, large boluses of calcium carbonate are poorly absorbed because of their neutralizing effect on gastric acid. In achlorhydric subjects and for those taking drugs that inhibit gastric acid secretion, supplements should be taken with meals to optimize their absorption. Use of calcium citrate may be preferable in these circumstances. Calcium absorption may also be blunted in disease states such as pancreatic or biliary insufficiency, in which ingested calcium remains bound to unabsorbed fatty acids or other food constituents. At high levels of calcium intake, synthesis of $1,25(OH)_2D$ is reduced; this decreases the rate of active intestinal calcium absorption. The opposite occurs with dietary calcium restriction. Some calcium, ~2.5–5 mmol/d (100–200 mg/d), is excreted as an obligate component of intestinal secretions and is not regulated by calciotropic hormones.

The feedback-controlled hormonal regulation of intestinal absorptive efficiency results in a relatively constant daily net calcium absorption of ~5–7.5 mmol/d (200–400 mg/d) despite large changes in daily dietary calcium intake. This daily load of absorbed calcium is excreted by the kidneys in a manner that is also tightly regulated by the concentration of ionized calcium in the blood. Approximately 8–10 g/d of calcium is filtered by the glomeruli, of which only 2–3% appears in the urine. Most filtered calcium (65%) is reabsorbed in the proximal tubules via a passive, paracellular route that is coupled to concomitant NaCl reabsorption and not specifically regulated. The cortical thick ascending limb of Henle's loop (cTAL) reabsorbs roughly another 20% of filtered calcium, also via a paracellular mechanism. Calcium reabsorption in the cTAL requires a tight-junctional protein called paracellin-1 and is inhibited by increased blood concentrations of calcium or magnesium, acting via the CaSR, which is highly expressed on basolateral membranes in this nephron segment. Operation of the renal CaSR provides a mechanism, independent of those engaged directly by PTH or $1,25(OH)_2D$, by which serum ionized calcium can control renal calcium reabsorption. Finally, ~10% of filtered calcium is reabsorbed in the distal convoluted tubules (DCTs) by a transcellular mechanism. Calcium enters the luminal surface of the cell through specific apical calcium channels (TRPV5), whose number is regulated. It then moves across the cell in association with a specific calcium-binding protein (calbindin-D28k) that buffers cytosolic calcium concentrations from the large mass of transported calcium. Ca^{2+}-ATPases and Na^+/Ca^{2+} exchangers actively extrude calcium across the basolateral surface and thereby maintain the transcellular calcium gradient. All these processes are stimulated directly or indirectly by PTH. The DCT is also the site of action of thiazide diuretics, which lower urinary calcium excretion by inducing sodium depletion and thereby augmenting proximal calcium reabsorption. Conversely, dietary sodium loads, or increased distal sodium delivery caused by loop diuretics or saline infusion, induce calciuresis.

The homeostatic mechanisms that normally maintain a constant serum ionized calcium concentration may fail at extremes of calcium intake or when the hormonal systems or organs involved are compromised. Thus, even with maximal activity of the vitamin D–dependent intestinal active transport system, sustained calcium intakes <5 mmol/d (<200 mg/d) cannot provide enough net calcium absorption to replace obligate losses via the intestine, the kidney, sweat, and other secretions. In this case, increased blood levels of PTH and $1,25(OH)_2D$ activate osteoclastic bone resorption to obtain needed calcium from bone, which leads to progressive bone loss and negative calcium balance. Increased PTH and $1,25(OH)_2D$ also enhance renal calcium reabsorption, and $1,25(OH)_2D$ enhances calcium absorption in the gut. At very high calcium intakes (>100 mmol/d [>4 g/d]), passive intestinal absorption continues to deliver calcium into the ECF despite maximally downregulated intestinal active transport and renal tubular calcium reabsorption. This can cause severe hypercalciuria, nephrocalcinosis, progressive renal failure, and hypercalcemia (e.g., "milk-alkali syndrome"). Deficiency or excess of PTH or vitamin D, intestinal disease, and renal failure represent other commonly encountered challenges to normal calcium homeostasis (Chap. 403).

PHOSPHORUS METABOLISM

Although 85% of the ~600 g of body phosphorus is present in bone mineral, phosphorus is also a major intracellular constituent both as the free anion(s) and as a component of numerous organophosphate compounds, including structural proteins, enzymes, transcription factors, carbohydrate and lipid intermediates, high-energy stores (ATP [adenosine triphosphate], creatine phosphate), and nucleic acids. Unlike calcium, phosphorus exists intracellularly at concentrations close to those present in ECF (e.g., 1–2 mmol/L). In cells and in the ECF, phosphorus exists in several forms,

predominantly as $H_2PO_4^-$ or $NaHPO_4^-$, with perhaps 10% as HPO_4^{2-}. This mixture of anions will be referred to here as "phosphate." In serum, about 12% of phosphorus is bound to proteins. Concentrations of phosphates in blood and ECF generally are expressed in terms of elemental phosphorus, with the normal range in adults being 0.75–1.45 mmol/L (2.5–4.5 mg/dL). Because the volume of the intracellular fluid compartment is twice that of the ECF, measurements of ECF phosphate may not accurately reflect phosphate availability within cells that follows even modest shifts of phosphate from one compartment to the other.

Phosphate is widely available in foods and is absorbed efficiently (65%) by the small intestine even in the absence of vitamin D. However, phosphate absorptive efficiency may be enhanced (to 85–90%) via active transport mechanisms that are stimulated by $1,25(OH)_2D$. These mechanisms involve activation of Na^+/PO_4^{2-} co-transporters that move phosphate into intestinal cells against an unfavorable electrochemical gradient. Daily net intestinal phosphate absorption varies widely with the composition of the diet but is generally in the range of 500–1000 mg/d. Phosphate absorption can be inhibited by large doses of calcium salts or by sevelamer hydrochloride (Renagel), strategies commonly used to control levels of serum phosphate in renal failure. Aluminum hydroxide antacids also reduce phosphate absorption but are used less commonly because of the potential for aluminum toxicity. Low serum phosphate stimulates renal proximal tubular synthesis of $1,25(OH)_2D$, perhaps by suppressing blood levels of FGF23 (see below).

Serum phosphate levels vary by as much as 50% on a normal day. This reflects the effect of food intake but also an underlying circadian rhythm that produces a nadir between 7 and 10 A.M. Carbohydrate administration, especially as IV dextrose solutions in fasting subjects, can decrease serum phosphate by >0.7 mmol/L (2 mg/dL) due to rapid uptake into and utilization by cells. A similar response is observed in the treatment of diabetic ketoacidosis and during metabolic or respiratory alkalosis. Because of this wide variation in serum phosphate, it is best to perform measurements in the basal, fasting state.

Control of serum phosphate is determined mainly by the rate of renal tubular reabsorption of the filtered load, which is ~4–6 g/d. Because intestinal phosphate absorption is highly efficient, urinary excretion is not constant but varies directly with dietary intake. The fractional excretion of phosphate (ratio of phosphate to creatinine clearance) is generally in the range of 10–15%. The proximal tubule is the principal site at which renal phosphate reabsorption is regulated. This is accomplished by changes in the levels of apical expression and activity of specific Na^+/PO_4^{2-} co-transporters (NaPi-2a and NaPi-2c) in the proximal tubule. Levels of these transporters at the apical surface of these cells are reduced rapidly by PTH, a major hormonal regulator of renal phosphate excretion. FGF23 can impair phosphate reabsorption dramatically by a similar mechanism. Activating *FGF23* mutations cause the rare disorder autosomal dominant hypophosphatemic rickets (ADHR). In contrast to PTH, FGF23 also leads to reduced synthesis of $1,25(OH)_2D$, which may worsen the resulting hypophosphatemia by lowering intestinal phosphate absorption. Renal reabsorption of phosphate is responsive to changes in dietary intake such that experimental dietary phosphate restriction leads to a dramatic lowering of urinary phosphate within hours, preceding any decline in serum phosphate (e.g., filtered load). This physiologic renal adaptation to changes in dietary phosphate availability occurs independently of PTH and may be mediated in part by changes in levels of serum FGF23. Findings in *FGF23*-knockout mice suggest that FGF23 normally acts to lower blood phosphate and $1,25(OH)_2D$ levels. In turn, elevation of blood phosphate increases blood levels of FGF23.

Renal phosphate reabsorption is impaired by hypocalcemia, hypomagnesemia, and severe hypophosphatemia. Phosphate clearance is enhanced by ECF volume expansion and impaired by dehydration. Phosphate retention is an important pathophysiologic feature of renal insufficiency **(Chap. 305)**.

HYPOPHOSPHATEMIA

Causes

Hypophosphatemia can occur by one or more of three primary mechanisms: (1) inadequate intestinal phosphate absorption, (2) excessive renal phosphate excretion, and (3) rapid redistribution of phosphate from the ECF into bone or soft tissue **(Table 402-1)**. Because phosphate is so abundant in foods, inadequate intestinal absorption is almost never observed now that aluminum hydroxide antacids, which bind phosphate in the gut, are no longer widely used. Fasting or starvation, however, may result in depletion of body phosphate and predispose to subsequent hypophosphatemia during refeeding, especially if this is accomplished with IV glucose alone.

TABLE 402-1

Causes of Hypophosphatemia

I. Reduced renal tubular phosphate reabsorption
- A. PTH/PTHrP-dependent
  1. Primary hyperparathyroidism
  2. Secondary hyperparathyroidism
     - a. Vitamin D deficiency/resistance
     - b. Calcium starvation/malabsorption
     - c. Bartter's syndrome
     - d. Autosomal recessive renal hypercalciuria with hypomagnesemia
  3. PTHrP-dependent hypercalcemia of malignancy
  4. Familial hypocalciuric hypercalcemia
- B. PTH/PTHrP-independent
  1. Excess FGF23 or other "phosphatonins"
     - a. X-linked hypophosphatemic rickets (XLH)
     - b. Autosomal recessive hypophosphatemia (ARHP)
     - c. Autosomal dominant hypophosphatemic rickets (ADHR) (DMP1, ENPP1 deficiency)
     - d. Tumor-induced osteomalacia syndrome (TIO)
     - e. McCune-Albright syndrome (fibrous dysplasia)
     - f. Epidermal nevus syndrome
  2. Intrinsic renal disease
     - a. Fanconi's syndrome(s)
     - b. Cystinosis
     - c. Wilson's disease
     - d. NaPi-2a or NaPi-2c mutations
  3. Other systemic disorders
     - a. Poorly controlled diabetes mellitus
     - b. Alcoholism
     - c. Hyperaldosteronism
     - d. Hypomagnesemia
     - e. Amyloidosis
     - f. Hemolytic-uremic syndrome
     - g. Renal transplantation or partial liver resection
     - h. Rewarming or induced hyperthermia
  4. Drugs or toxins
     - a. Ethanol
     - b. Acetazolamide, other diuretics
     - c. High-dose estrogens or glucocorticoids
     - d. Heavy metals (lead, cadmium, saccharated ferric oxide)
     - e. Toluene, *N*-methyl formamide
     - f. Cisplatin, ifosfamide, foscarnet, rapamycin

II. Impaired intestinal phosphate absorption
- A. Aluminum-containing antacids
- B. Sevelamer

III. Shifts of extracellular phosphate into cells
- A. Intravenous glucose
- B. Insulin therapy for prolonged hyperglycemia or diabetic ketoacidosis
- C. Catecholamines (epinephrine, dopamine, albuterol)
- D. Acute respiratory alkalosis
- E. Gram-negative sepsis, toxic shock syndrome
- F. Recovery from starvation or acidosis
- G. Rapid cellular proliferation
  1. Leukemic blast crisis
  2. Intensive erythropoietin, other growth factor therapy

IV. Accelerated net bone formation
- A. After parathyroidectomy
- B. Treatment of vitamin D deficiency, Paget's disease
- C. Osteoblastic metastases

Abbreviations: PTH, parathyroid hormone; PTHrP, parathyroid hormone–related peptide.

Chronic hypophosphatemia usually signifies a persistent renal tubular phosphate-wasting disorder. Excessive activation of PTH/PTHrP receptors in the proximal tubule as a result of primary or secondary hyperparathyroidism or because of the PTHrP-mediated hypercalcemia syndrome in malignancy **(Chap. 403)** is among the more common causes of renal hypophosphatemia, especially because of the high prevalence of vitamin D deficiency in older Americans. Familial hypocalciuric hypercalcemia and Jansen's chondrodystrophy are rare examples of genetic disorders in this category **(Chap. 403)**.

Several genetic and acquired diseases cause PTH/PTHrP-independent tubular phosphate wasting with associated rickets and osteomalacia. All these diseases manifest severe hypophosphatemia; renal phosphate wasting, sometimes accompanied by aminoaciduria; inappropriately low blood levels of $1,25(OH)_2D$; low-normal serum levels of calcium; and evidence of impaired cartilage or bone mineralization. Analysis of these diseases led to the discovery of the hormone FGF23, which is an important physiologic regulator of phosphate metabolism. FGF23 decreases phosphate reabsorption in the proximal tubule and also suppresses the 1α-hydroxylase responsible for synthesis of $1,25(OH)_2D$. FGF23 is synthesized by cells of the osteoblast lineage, primarily osteocytes. High-phosphate diets increase FGF23 levels, and low-phosphate diets decrease them. ADHR was the first disease linked to abnormalities in FGF23. ADHR results from activating mutations in the gene that encodes FGF23. These mutations alter a cleavage site that ordinarily allows for inactivation of intact FGF23. Several other genetic disorders feature elevated FGF23 and hypophosphatemia. The most common of these is X-linked hypophosphatemic rickets (XLH), which results from inactivating mutations in an endopeptidase termed *PHEX* (*p*hosphate-regulating gene with *h*omologies to *e*ndopeptidases on the *X* chromosome) that is expressed most abundantly on the surface of osteocytes and mature osteoblasts. Patients with XLH usually have high FGF23 levels, and ablation of the *FGF23* gene reverses the hypophosphatemia found in the mouse version of XLH. How inactivation of PHEX leads to increased levels of FGF23 has not been determined. Two rare autosomal recessive hypophosphatemic syndromes associated with elevated FGF23 are due to inactivating mutations of dentin matrix protein-1 (DMP1) and ectonucleotide pyrophosphatase/phosphodiesterase 1 (ENPP1), respectively, both of which normally are highly expressed in bone and presumably regulate FGF23 production. An unusual hypophosphatemic disorder, tumor-induced osteomalacia (TIO), is an acquired disorder in which tumors, usually of mesenchymal origin and generally histologically benign, secrete FGF23 and/or other molecules that induce renal phosphate wasting. The hypophosphatemic syndrome resolves completely within hours to days after successful resection of the responsible tumor. Such tumors typically express large amounts of FGF23 mRNA, and patients with TIO usually exhibit elevations of FGF23 in their blood.

Dent's disease is an X-linked recessive disorder caused by inactivating mutations in CLCN5, a chloride transporter expressed in endosomes of the proximal tubule; features include hypercalciuria, hypophosphatemia, and recurrent kidney stones. Renal phosphate wasting is common among poorly controlled diabetic patients and alcoholics, who therefore are at risk for iatrogenic hypophosphatemia when treated with insulin or IV glucose, respectively. Diuretics and certain other drugs and toxins can cause defective renal tubular phosphate reabsorption (Table 402-1).

In hospitalized patients, hypophosphatemia is often attributable to massive redistribution of phosphate from the ECF into cells. Insulin therapy for diabetic ketoacidosis is a paradigm for this phenomenon, in which the severity of the hypophosphatemia is related to the extent of antecedent depletion of phosphate and other electrolytes **(Chap. 397)**. The hypophosphatemia is usually greatest at a point many hours after initiation of insulin therapy and is difficult to predict from baseline measurements of serum phosphate at the time of presentation, when prerenal azotemia can obscure significant phosphate depletion. Other factors that may contribute to such acute redistributive hypophosphatemia include antecedent starvation or malnutrition, administration of IV glucose without other nutrients, elevated blood catecholamines (endogenous or exogenous), respiratory alkalosis, and recovery from metabolic acidosis.

Hypophosphatemia also can occur transiently (over weeks to months) during the phase of accelerated net bone formation that follows parathyroidectomy for severe primary hyperparathyroidism or during treatment of vitamin D deficiency or lytic Paget's disease. This is usually most prominent in patients who preoperatively have evidence of high bone turnover (e.g., high serum levels of alkaline phosphatase). Osteoblastic metastases can also lead to this syndrome.

Clinical and Laboratory Findings

The clinical manifestations of severe hypophosphatemia reflect a generalized defect in cellular energy metabolism because of ATP depletion, a shift from oxidative phosphorylation toward glycolysis, and associated tissue or organ dysfunction. Acute, severe hypophosphatemia occurs mainly or exclusively in hospitalized patients with underlying serious medical or surgical illness and preexisting phosphate depletion due to excessive urinary losses, severe malabsorption, or malnutrition. Chronic hypophosphatemia tends to be less severe, with a clinical presentation dominated by

musculoskeletal complaints such as bone pain, osteomalacia, pseudofractures, and proximal muscle weakness or, in children, rickets and short stature.

Neuromuscular manifestations of severe hypophosphatemia are variable but may include muscle weakness, lethargy, confusion, disorientation, hallucinations, dysarthria, dysphagia, oculomotor palsies, anisocoria, nystagmus, ataxia, cerebellar tremor, ballismus, hyporeflexia, impaired sphincter control, distal sensory deficits, paresthesia, hyperesthesia, generalized or Guillain-Barré–like ascending paralysis, seizures, coma, and even death. Serious sequelae such as paralysis, confusion, and seizures are likely only at phosphate concentrations <0.25 mmol/L (<0.8 mg/dL). Rhabdomyolysis may develop during rapidly progressive hypophosphatemia. The diagnosis of hypophosphatemia-induced rhabdomyolysis may be overlooked, as up to 30% of patients with acute hypophosphatemia (<0.7 m*M*) have creatine phosphokinase elevations that peak 1–2 days after the nadir in serum phosphate, when the release of phosphate from injured myocytes may have led to a near normalization of circulating levels of phosphate.

Respiratory failure and cardiac dysfunction, which are reversible with phosphate treatment, may occur at serum phosphate levels of 0.5–0.8 mmol/L (1.5–2.5 mg/dL). Renal tubular defects, including tubular acidosis, glycosuria, and impaired reabsorption of sodium and calcium, may occur. Hematologic abnormalities correlate with reductions in intracellular ATP and 2,3-diphosphoglycerate and may include erythrocyte microspherocytosis and hemolysis; impaired oxyhemoglobin dissociation; defective leukocyte chemotaxis, phagocytosis, and bacterial killing; and platelet dysfunction with spontaneous gastrointestinal hemorrhage.

TREATMENT

TREATMENT

Hypophosphatemia

Severe hypophosphatemia (<0.75 mmol/L [<2 mg/dL]), particularly in the setting of underlying phosphate depletion, constitutes a dangerous electrolyte abnormality that should be corrected promptly. Unfortunately, the cumulative deficit in body phosphate cannot be predicted easily from knowledge of the circulating level of phosphate, and therapy must be approached empirically. The threshold for IV phosphate therapy and the dose administered should reflect consideration of renal function, the likely severity and duration of the underlying phosphate depletion, and the presence and severity of symptoms consistent with those of hypophosphatemia. In adults, phosphate may be safely administered IV as neutral mixtures of sodium or potassium phosphate salts at initial doses of 0.2–0.8 mmol/kg of elemental phosphorus over 6 h (e.g., 10–50 mmol over 6 h), with doses >20 mmol/6 h reserved for those who have serum levels <0.5 mmol/L (1.5 mg/dL) and normal renal function. A suggested approach is presented in **Table 402-2**. Serum levels of phosphate and calcium must be monitored closely (every 6–12 h) throughout treatment. It is necessary to avoid a serum calcium-phosphorus product >50 to reduce the risk of heterotopic calcification. Hypocalcemia, if present, should be corrected before administering IV phosphate. Less severe hypophosphatemia, in the range of 0.5–0.8 mmol/L (1.5–2.5 mg/dL), usually can be treated with oral phosphate in divided doses of 750–2000 mg/d as elemental phosphorus; higher doses can cause bloating and diarrhea.

Management of chronic hypophosphatemia requires knowledge of the cause(s) of the disorder. Hypophosphatemia related to the secondary hyperparathyroidism of vitamin D deficiency usually responds to treatment with vitamin D and calcium alone. XLH, ADHR, TIO, and related renal tubular disorders usually are managed with divided oral doses of phosphate, often with calcium and $1,25(OH)_2D$ supplements to bypass the block in renal $1,25(OH)_2D$ synthesis and prevent secondary hyperparathyroidism caused by suppression of ECF calcium levels. Thiazide diuretics may be used to prevent nephrocalcinosis in patients who are managed this way. Complete normalization of hypophosphatemia is generally not possible in these conditions. Optimal therapy for TIO is extirpation of the responsible tumor, which may be localized by radiographic skeletal survey or bone scan (many are located in bone) or by radionuclide scanning using sestamibi or labeled octreotide. Successful treatment of TIO-induced hypophosphatemia with octreotide has been reported in a small number of patients.

TABLE 402-2

Intravenous Therapy for Hypophosphatemia

Consider			
Likely severity of underlying phosphate depletion			
Concurrent parenteral glucose administration			
Presence of neuromuscular, cardiopulmonary, or hematologic complications of hypophosphatemia			
Renal function (reduce dose by 50% if serum creatinine >220 µmol/L [>2.5 mg/dL])			
Serum calcium level (correct hypocalcemia first; reduce dose by 50% in hypercalcemia)			
Guidelines			
Serum Phosphorus, mm (mg/dL)	**Rate of Infusion, mmol/h**	**Duration, h**	**Total Administered, mmol**
<0.8 (<2.5)	2	6	12
<0.5 (<1.5)	4	6	24
<0.3 (<1)	8	6	48

Rates shown are calculated for a 70-kg person; levels of serum calcium and phosphorus must be measured every 6–12 h during therapy; infusions can be repeated to achieve stable serum phosphorus levels >0.8 mmol/L (>2.5 mg/dL); most formulations available in the United States provide 3 mmol/mL of sodium or potassium phosphate.

HYPERPHOSPHATEMIA

Causes

When the filtered load of phosphate and glomerular filtration rate (GFR) are normal, control of serum phosphate levels is achieved by adjusting the rate at which phosphate is reabsorbed by the proximal tubular NaPi-2 co-transporters. The principal hormonal regulators of NaPi-2 activity are PTH and FGF23. Hyperphosphatemia, defined in adults as a fasting serum phosphate concentration >1.8 mmol/L (5.5 mg/dL), usually results from impaired glomerular filtration, hypoparathyroidism, excessive delivery of phosphate into the ECF (from bone, gut, or parenteral phosphate therapy), or a combination of these factors **(Table 402-3)**. The upper limit of normal serum phosphate concentrations is higher in children and neonates (2.4 mmol/L [7 mg/dL]). It is useful to distinguish hyperphosphatemia caused by impaired renal phosphate excretion from that which results from excessive delivery of phosphate into the ECF (Table 402-3).

TABLE 402-3

Causes of Hyperphosphatemia

- I. Impaired renal phosphate excretion
  - A. Renal insufficiency
  - B. Hypoparathyroidism
    - 1. Developmental
    - 2. Autoimmune
    - 3. After neck surgery or radiation
    - 4. Activating mutations of the calcium-sensing receptor
  - C. Parathyroid suppression
    - 1. Parathyroid-independent hypercalcemia
      - a. Vitamin D or vitamin A intoxication
      - b. Sarcoidosis, other granulomatous diseases
      - c. Immobilization, osteolytic metastases
      - d. Milk-alkali syndrome
    - 2. Severe hypermagnesemia or hypomagnesemia
  - D. Pseudohypoparathyroidism
  - E. Acromegaly
  - F. Tumoral calcinosis
  - G. Heparin therapy
- II. Massive extracellular fluid phosphate loads
  - A. Rapid administration of exogenous phosphate (intravenous, oral, rectal)
  - B. Extensive cellular injury or necrosis
    - 1. Crush injuries
    - 2. Rhabdomyolysis
    - 3. Hyperthermia
    - 4. Fulminant hepatitis
    - 5. Cytotoxic therapy
    - 6. Severe hemolytic anemia
  - C. Transcellular phosphate shifts
    - 1. Metabolic acidosis
    - 2. Respiratory acidosis

In chronic renal insufficiency, reduced GFR leads to phosphate retention. Hyperphosphatemia in turn further impairs renal synthesis of $1,25(OH)_2D$, increases FGF23 levels, and stimulates PTH secretion and hypertrophy both directly and indirectly (by lowering blood ionized calcium levels). Thus, hyperphosphatemia is a major cause of the secondary hyperparathyroidism of renal failure and must be addressed early in the course of the disease (**Chaps. 305** and **403**).

Hypoparathyroidism leads to hyperphosphatemia via increased expression of NaPi-2 co-transporters in the proximal tubule. Hypoparathyroidism, or parathyroid suppression, has multiple potential causes, including autoimmune disease; developmental, surgical, or radiation-induced absence of functional parathyroid tissue; vitamin D intoxication or other causes of PTH-independent hypercalcemia; cellular PTH resistance (pseudohypoparathyroidism or hypomagnesemia); infiltrative disorders such as Wilson's disease and hemochromatosis; and impaired PTH secretion caused by hypermagnesemia, severe hypomagnesemia, or activating mutations in the CaSR. Hypocalcemia may also contribute directly to impaired phosphate clearance, as calcium infusion can induce phosphaturia in hypoparathyroid subjects. Increased tubular phosphate reabsorption also occurs in acromegaly, during heparin administration, and in tumoral calcinosis. Tumoral calcinosis is caused by a rare group of genetic disorders in which FGF23 is processed in a way that leads to low levels of active FGF23 in the bloodstream. This may result from mutations in the FGF23 sequence or via inactivating mutations in the *GALNT3* gene, which encodes a galactosaminyl transferase that normally adds sugar residues to FGF23 that slow its proteolysis. A similar syndrome results from FGF23 resistance due to inactivating mutations of the FGF23 co-receptor Klotho. These abnormalities cause elevated serum $1,25(OH)_2D$, parathyroid suppression, increased intestinal calcium absorption, and focal hyperostosis with large, lobulated periarticular heterotopic ossifications (especially at shoulders or hips) and are accompanied by hyperphosphatemia. In some forms of tumoral calcinosis serum phosphorus levels are normal.

When large amounts of phosphate are delivered rapidly into the ECF, hyperphosphatemia can occur despite normal renal function. Examples include overzealous IV phosphate therapy, oral or rectal administration of large amounts of phosphate-containing laxatives or enemas (especially in children), extensive soft tissue injury or necrosis (crush injuries, rhabdomyolysis, hyperthermia, fulminant hepatitis, cytotoxic chemotherapy), extensive hemolytic anemia, and transcellular phosphate shifts induced by severe metabolic or respiratory acidosis.

Clinical Findings

The clinical consequences of acute, severe hyperphosphatemia are due mainly to the formation of widespread calcium phosphate precipitates and resulting hypocalcemia. Thus, tetany, seizures, accelerated nephrocalcinosis (with renal failure, hyperkalemia, hyperuricemia, and metabolic acidosis), and pulmonary or cardiac calcifications (including development of acute heart block) may occur. The severity of these complications relates to the elevation of serum phosphate levels, which can reach concentrations as high as 7 mmol/L (20 mg/dL) in instances of massive soft tissue injury or tumor lysis syndrome.

TREATMENT

TREATMENT

Hyperphosphatemia

Therapeutic options for management of severe hyperphosphatemia are limited. Volume expansion may enhance renal phosphate clearance. Aluminum hydroxide antacids or sevalamer may be helpful in chelating and limiting absorption of offending phosphate salts present in the intestine. Hemodialysis is the most effective therapeutic strategy and should be considered early in the course of severe hyperphosphatemia, especially in the setting of renal failure and symptomatic hypocalcemia.

MAGNESIUM METABOLISM

Magnesium is the major intracellular divalent cation. Normal concentrations of extracellular magnesium and calcium are crucial for normal neuromuscular activity. Intracellular magnesium forms a key complex with ATP and is an important cofactor for a wide range of enzymes, transporters, and nucleic acids required for normal cellular function, replication, and energy metabolism. The concentration of magnesium in serum is closely regulated within the range of 0.7–1 mmol/L (1.5–2 meq/L; 1.7–2.4 mg/dL), of which 30% is protein-bound and another 15% is loosely complexed to phosphate and other anions. One-half of the 25 g (1000 mmol) of total body magnesium is located in bone, only one-half of which is insoluble in the mineral phase. Almost all extraskeletal magnesium is present within cells, where the total concentration is 5 m*M*, 95% of which is bound to proteins and other macromolecules. Because only 1% of body magnesium resides in the ECF, measurements of serum magnesium levels may not accurately reflect the level of total body magnesium stores.

Dietary magnesium content normally ranges from 6 to 15 mmol/d (140–360 mg/d), of which 30–40% is absorbed, mainly in the jejunum and ileum. Intestinal magnesium absorptive efficiency is stimulated by $1,25(OH)_2D$ and can reach 70% during magnesium deprivation. Urinary magnesium excretion normally matches net intestinal absorption and is ~4 mmol/d (100 mg/d). Regulation of serum magnesium concentrations is achieved mainly by control of renal magnesium reabsorption. Only 20% of filtered magnesium is reabsorbed in the proximal tubule, whereas 60% is reclaimed in the cTAL and another 5–10% in the DCT. Magnesium reabsorption in the cTAL occurs via a paracellular route that requires both a lumen-positive potential, created by NaCl reabsorption, and tight-junction proteins encoded by members of the Claudin gene family. Magnesium reabsorption in the cTAL is increased by PTH but inhibited by hypercalcemia or hypermagnesemia, both of which activate the CaSR in this nephron segment.

HYPOMAGNESEMIA

Causes

Hypomagnesemia usually signifies substantial depletion of body magnesium stores (0.5–1 mmol/kg). Hypomagnesemia can result from intestinal malabsorption; protracted vomiting, diarrhea, or intestinal drainage; defective renal tubular magnesium reabsorption; or rapid shifts of magnesium from the ECF into cells, bone, or third spaces (Table 402-4). Dietary magnesium deficiency is unlikely except possibly in the setting of alcoholism. A rare genetic disorder that causes selective intestinal magnesium malabsorption has been described (primary infantile hypomagnesemia). Another rare inherited disorder (hypomagnesemia with secondary hypocalcemia) is caused by mutations in the gene encoding TRPM6, a protein that, along with TRPM7, forms a channel important for both intestinal and distal-tubular renal transcellular magnesium transport. Malabsorptive states, often compounded by vitamin D deficiency, can critically limit magnesium absorption and produce hypomagnesemia despite the compensatory effects of secondary hyperparathyroidism and of hypocalcemia and hypomagnesemia to enhance cTAL magnesium reabsorption. Diarrhea or surgical drainage fluid may contain ≥5 mmol/L of magnesium. Proton pump inhibitors (omeprazole and others) may produce hypomagnesemia by an unknown mechanism that does not involve renal wasting of magnesium.

TABLE 402-4

Causes of Hypomagnesemia

I. Impaired intestinal absorption
- A. Hypomagnesemia with secondary hypocalcemia (TRPM6 mutations)
- B. Malabsorption syndromes
- C. Vitamin D deficiency
- D. Proton pump inhibitors

II. Increased intestinal losses
- A. Protracted vomiting/diarrhea
- B. Intestinal drainage, fistulas

III. Impaired renal tubular reabsorption
- A. Genetic magnesium-wasting syndromes
  1. Gitelman's syndrome
  2. Bartter's syndrome
  3. Claudin 16 or 19 mutations
  4. Potassium channel mutations (Kv1.1, Kir4.1)
  5. Na^+,K^+-ATPase γ-subunit mutations (FXYD2)
- B. Acquired renal disease
  1. Tubulointerstitial disease
  2. Postobstruction, ATN (diuretic phase)
  3. Renal transplantation
- C. Drugs and toxins
  1. Ethanol
  2. Diuretics (loop, thiazide, osmotic)
  3. Cisplatin
  4. Pentamidine, foscarnet
  5. Cyclosporine
  6. Aminoglycosides, amphotericin B
  7. Cetuximab
- D. Other
  1. Extracellular fluid volume expansion
  2. Hyperaldosteronism
  3. SIADH
  4. Diabetes mellitus
  5. Hypercalcemia
  6. Phosphate depletion
  7. Metabolic acidosis
  8. Hyperthyroidism

IV. Rapid shifts from extracellular fluid
- A. Intracellular redistribution
  1. Recovery from diabetic ketoacidosis
  2. Refeeding syndrome
  3. Correction of respiratory acidosis
  4. Catecholamines
- B. Accelerated bone formation
  1. Post-parathyroidectomy
  2. Treatment of vitamin D deficiency
  3. Osteoblastic metastases
- C. Other
  1. Pancreatitis, burns, excessive sweating
  2. Pregnancy (third trimester) and lactation

Abbreviations: ATN, acute tubular necrosis; SIADH, syndrome of inappropriate antidiuretic hormone.

Several genetic magnesium-wasting syndromes have been described, including inactivating mutations of genes encoding the DCT NaCl co-transporter (Gitelman's syndrome), proteins required for cTAL Na-K-2Cl transport (Bartter's syndrome), claudin 16 or claudin 19 (autosomal recessive renal hypomagnesemia with hypercalciuria), a DCT Na^+,K^+-ATPase γ-subunit (autosomal dominant renal hypomagnesemia with hypocalciuria), DCT K+ channels (Kv1.1, Kir4.1) and a mitochondrial gene encoding a tRNA. Activating mutations of the CaSR can cause hypomagnesemia as well as hypocalcemia. ECF expansion, hypercalcemia, and severe phosphate depletion may impair magnesium reabsorption, as can various forms of renal injury, including those caused by drugs such as cisplatin, cyclosporine, aminoglycosides, and pentamidine as well as the EGF receptor inhibitory antibody, cetuximab (EGF action is required for normal DCT apical expression of TRPM6) (Table 402-4). A rising blood concentration of ethanol directly impairs tubular magnesium reabsorption, and persistent glycosuria with osmotic diuresis leads to magnesium wasting and probably contributes to the high frequency of hypomagnesemia in poorly controlled diabetic patients. Magnesium depletion is aggravated by metabolic acidosis, which causes intracellular losses as well.

Hypomagnesemia due to rapid shifts of magnesium from ECF into the intracellular compartment can occur during recovery from diabetic ketoacidosis, starvation, or respiratory acidosis. Less acute shifts may be seen during rapid bone formation after parathyroidectomy, with treatment of vitamin D deficiency, or with osteoblastic metastases. Large amounts of magnesium may be lost with acute pancreatitis, extensive burns, or protracted and severe sweating and during pregnancy and lactation.

Clinical and Laboratory Findings

Hypomagnesemia may cause generalized alterations in neuromuscular function, including tetany, tremor, seizures, muscle weakness, ataxia, nystagmus, vertigo, apathy, depression, irritability, delirium, and psychosis. Patients are usually asymptomatic when serum magnesium concentrations are >0.5 mmol/L (1 meq/L; 1.2 mg/dL), although the severity of symptoms may not correlate with serum magnesium levels. Cardiac arrhythmias may occur, including sinus tachycardia, other supraventricular tachycardias, and ventricular arrhythmias. Electrocardiographic abnormalities may include prolonged PR or QT intervals, T-wave flattening or inversion, and ST straightening. Sensitivity to digitalis toxicity may be enhanced.

Other electrolyte abnormalities often seen with hypomagnesemia, including hypocalcemia (with hypocalciuria) and hypokalemia, may not be easily corrected unless magnesium is administered as well. The hypocalcemia may be a result of concurrent vitamin D deficiency, although hypomagnesemia can cause impaired synthesis of $1,25(OH)_2D$, cellular resistance to PTH, and, at very low serum magnesium (<0.4 mmol/L [<0.8 meq/L; <1 mg/dL]), a defect in PTH secretion; these abnormalities are reversible with therapy.

TREATMENT

TREATMENT

Hypomagnesemia

Mild, asymptomatic hypomagnesemia may be treated with oral magnesium salts ($MgCl_2$, MgO, $Mg[OH]_2$) in divided doses totaling 20–30 mmol/d (40–60 meq/d). Diarrhea may occur with larger doses. More severe hypomagnesemia should be treated parenterally, preferably with IV $MgCl_2$, which can be administered safely as a continuous infusion of 50 mmol/d (100 meq Mg^{2+}/d) if renal function is normal. If GFR is reduced, the infusion rate should be lowered by 50–75%. Use of IM $MgSO_4$ is discouraged; the injections are painful and provide relatively little magnesium (2 mL of 50% $MgSO_4$ supplies only 4 mmol). $MgSO_4$ may be given IV instead of $MgCl_2$, although the sulfate anions may bind calcium in serum and urine and aggravate hypocalcemia. Serum magnesium should be monitored at intervals of 12–24 h during therapy, which may continue for several days because of impaired renal conservation of magnesium (only 50–70% of the daily IV magnesium dose is retained) and delayed repletion of intracellular deficits, which may be as high as 1–1.5 mmol/kg (2–3 meq/kg).

It is important to consider the need for calcium, potassium, and phosphate supplementation in patients with hypomagnesemia. Vitamin D deficiency frequently coexists and should be treated with oral or parenteral vitamin D or 25(OH)D (but not with $1,25(OH)_2D$, which may impair tubular magnesium reabsorption, possibly via PTH suppression). In severely hypomagnesemic patients with concomitant hypocalcemia and hypophosphatemia, administration of IV magnesium alone may worsen hypophosphatemia, provoking neuromuscular symptoms or rhabdomyolysis, due to rapid stimulation of PTH secretion. This is avoided by administering both calcium and magnesium.

HYPERMAGNESEMIA

Causes

Hypermagnesemia is rarely seen in the absence of renal insufficiency, as normal kidneys can excrete large amounts (250 mmol/d) of magnesium. Mild hypermagnesemia due to excessive reabsorption in the cTAL occurs with CaSR mutations in familial hypocalciuric hypercalcemia and has been

described in some patients with adrenal insufficiency, hypothyroidism, or hypothermia. Massive exogenous magnesium exposures, usually via the gastrointestinal tract, can overwhelm renal excretory capacity and cause life-threatening hypermagnesemia **(Table 402-5)**. A notable example of this is prolonged retention of even normal amounts of magnesium-containing cathartics in patients with intestinal ileus, obstruction, or perforation. Extensive soft tissue injury or necrosis can also deliver large amounts of magnesium into the ECF in patients who have suffered trauma, shock, sepsis, cardiac arrest, or severe burns.

TABLE 402-5

Causes of Hypermagnesemia

I. Excessive magnesium intake
A. Cathartics, urologic irrigants
B. Parenteral magnesium administration
II. Rapid mobilization from soft tissues
A. Trauma, shock, sepsis
B. Cardiac arrest
C. Burns
III. Impaired magnesium excretion
A. Renal failure
B. Familial hypocalciuric hypercalcemia
IV. Other
A. Adrenal insufficiency
B. Hypothyroidism
C. Hypothermia

Clinical and Laboratory Findings

The most prominent clinical manifestations of hypermagnesemia are vasodilation and neuromuscular blockade, which may appear at serum magnesium concentrations >2 mmol/L (>4 meq/L; >4.8 mg/dL). Hypotension that is refractory to vasopressors or volume expansion may be an early sign. Nausea, lethargy, and weakness may progress to respiratory failure, paralysis, and coma, with hypoactive tendon reflexes, at serum magnesium levels >4 mmol/L. Other findings may include gastrointestinal hypomotility or ileus; facial flushing; pupillary dilation; paradoxical bradycardia; prolongation of PR, QRS, and QT intervals; heart block; and, at serum magnesium levels approaching 10 mmol/L, asystole.

Hypermagnesemia, acting via the CaSR, causes hypocalcemia and hypercalciuria due to both parathyroid suppression and impaired cTAL calcium reabsorption.

TREATMENT

TREATMENT

Hypermagnesemia

Successful treatment of hypermagnesemia generally involves identifying and interrupting the source of magnesium and employing measures to increase magnesium clearance from the ECF. Use of magnesium-free cathartics or enemas may be helpful in clearing ingested magnesium from the gastrointestinal tract. Vigorous IV hydration should be attempted, if appropriate. Hemodialysis is effective and may be required in patients with significant renal insufficiency. Calcium, administered IV in doses of 100–200 mg over 1–2 h, has been reported to provide temporary improvement in signs and symptoms of hypermagnesemia.

VITAMIN D

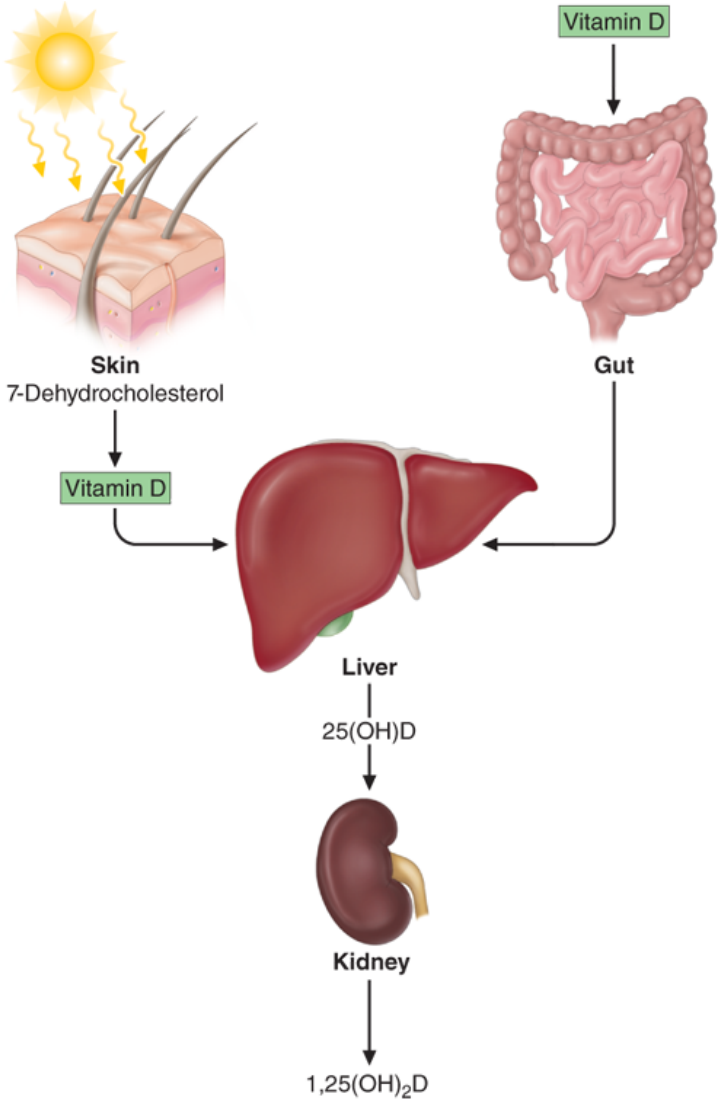
SYNTHESIS AND METABOLISM

1,25-dihydroxyvitamin D [1,25$(OH)_2$D] is the major steroid hormone involved in regulation of mineral ion homeostasis. Vitamin D and its metabolites are hormones and hormone precursors rather than vitamins, since in the proper biologic setting, they can be synthesized endogenously **(Fig. 402-4)**. In response to ultraviolet radiation of the skin, a photochemical cleavage results in the formation of vitamin D from 7-dehydrocholesterol. Cutaneous production of vitamin D is decreased by melanin and high solar protection factor sunblocks, which effectively impair skin penetration by ultraviolet

light. The increased use of sunblocks in North America and Western Europe and a reduction in the magnitude of solar exposure of the general population over the last several decades has led to an increased reliance on dietary sources of vitamin D. In the United States and Canada, these sources largely consist of fortified cereals and dairy products, in addition to fish oils and egg yolks. Vitamin D from plant sources is in the form of vitamin D_2, whereas that from animal sources is vitamin D_3. These two forms have equivalent biologic activity and are activated equally well by the vitamin D hydroxylases in humans. Vitamin D enters the circulation, whether absorbed from the intestine or synthesized cutaneously, bound to vitamin D–binding protein, an α-globulin synthesized in the liver. Vitamin D is subsequently 25-hydroxylated in the liver by a cytochrome P450 oxidase in the mitochondria and microsomes. The activity of this hydroxylase is not tightly regulated, and the resultant metabolite, 25-hydroxyvitamin D [25(OH)D], is the major circulating and storage form of vitamin D. Approximately 88% of 25(OH)D circulates bound to the vitamin D–binding protein, 0.03% is free, and the rest circulates bound to albumin. The half-life of 25(OH)D is ~2–3 weeks, with that of $25(OH)D_2$ being shorter than that of $25(OH)D_3$ due to a lower affinity of vitamin D-binding protein for the former. The half-life of 25(OH)D is also greatly shortened when vitamin D–binding protein levels are reduced, as can occur with increased urinary losses in the nephrotic syndrome.

FIGURE 402-4

Vitamin D synthesis and activation. Vitamin D is synthesized in the skin in response to ultraviolet radiation and also is absorbed from the diet. It is then transported to the liver, where it undergoes 25-hydroxylation. This metabolite is the major circulating form of vitamin D. The final step in hormone activation, 1α-hydroxylation, occurs in the kidney.



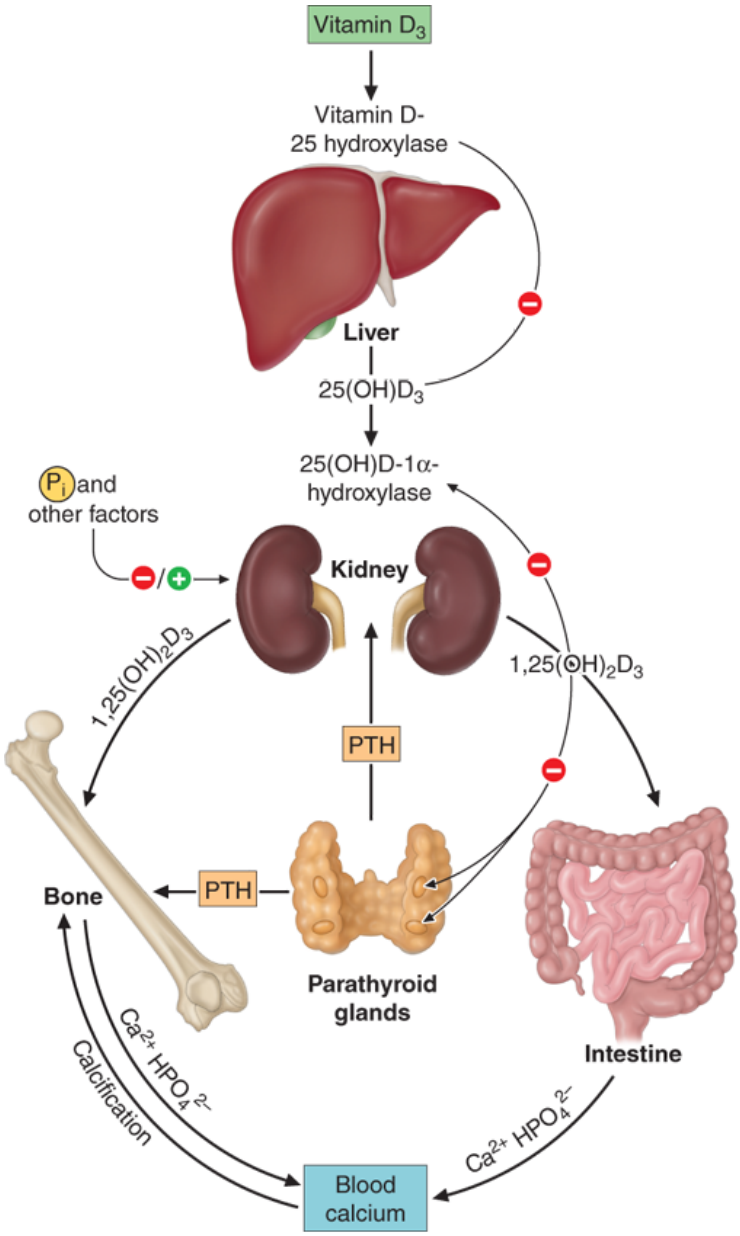
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The second hydroxylation, required for the formation of the mature hormone, occurs in the kidney (Fig. 402-5). The 25-hydroxyvitamin D-1α-hydroxylase is a tightly regulated cytochrome P450–like mixed-function oxidase expressed in the proximal convoluted tubule cells of the kidney. PTH and hypophosphatemia are the major inducers of this microsomal enzyme in the kidney, whereas calcium, FGF23, and the enzyme's product, $1,25(OH)_2D$, repress it. The 25-hydroxyvitamin D-1α-hydroxylase is also present in numerous other cell types, where it is not subject to hormonal regulation. It is expressed in epidermal keratinocytes, but keratinocyte production of $1,25(OH)_2D$ is not thought to contribute to circulating levels of

this hormone. In addition to being present in the trophoblastic layer of the placenta, the 1α-hydroxylase is produced by macrophages associated with granulomas and lymphomas. In these latter pathologic states, the activity of the enzyme is induced by interferon γ and TNF-α but is not regulated by calcium or $1,25(OH)_2D$; therefore, hypercalcemia, associated with elevated levels of $1,25(OH)_2D$, may be observed. Treatment of sarcoidosis-associated hypercalcemia with glucocorticoids, ketoconazole, or chloroquine reduces $1,25(OH)_2D$ production and effectively lowers serum calcium. In contrast, chloroquine has not been shown to lower the elevated serum $1,25(OH)_2D$ levels in patients with lymphoma.

FIGURE 402-5

Schematic representation of the hormonal control loop for vitamin D metabolism and function. A reduction in the serum calcium below ~2.2 mmol/L (8.8 mg/dL) prompts a proportional increase in the secretion of parathyroid hormone (PTH) and so mobilizes additional calcium from the bone. PTH promotes the synthesis of $1,25(OH)_2D$ in the kidney, which in turn stimulates the mobilization of calcium from bone and intestine and regulates the synthesis of PTH by negative feedback.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The major pathway for inactivation of vitamin D metabolites is an additional hydroxylation step by the vitamin D 24-hydroxylase, an enzyme that is expressed in most tissues. $1,25(OH)_2D$ is the major inducer of this enzyme; therefore, this hormone promotes its own inactivation, thereby limiting its biologic effects. FGF23 also induces this hydroxylase, thereby reducing circulating $1,25(OH)_2D$ levels by increasing its inactivation, as well as by impairing its synthesis. Mutations of the gene encoding this enzyme (CYP24 A1) can lead to infantile hypercalcemia and, in those less severely affected, long-standing hypercalciuria, nephrocalcinosis and nephrolithiasis can occur.

Polar metabolites of $1,25(OH)_2D$ are secreted into the bile and reabsorbed via the enterohepatic circulation. Impairment of this recirculation, which is seen with diseases of the terminal ileum, leads to accelerated losses of vitamin D metabolites.

ACTIONS OF $1,25(OH)_2D$

$1,25(OH)_2D$ mediates its biologic effects by binding to a member of the nuclear receptor superfamily, the vitamin D receptor (VDR). This receptor belongs to the subfamily that includes the thyroid hormone receptors, the retinoid receptors, and the peroxisome proliferator–activated receptors; however, in contrast to the other members of this subfamily, only one VDR isoform has been isolated. The VDR binds to target DNA sequences as a heterodimer with the retinoid X receptor, recruiting a series of coactivators that modify chromatin and approximate the VDR to the basal transcriptional apparatus, resulting in the induction of target gene expression. The mechanism of transcriptional repression by the VDR varies with different target genes but has been shown to involve either interference with the action of activating transcription factors or the recruitment of novel proteins to the VDR complex, resulting in transcriptional repression.

The affinity of the VDR for $1,25(OH)_2D$ is approximately three orders of magnitude higher than that for other vitamin D metabolites. In normal physiologic circumstances, these other metabolites are not thought to stimulate receptor-dependent actions. However, in states of vitamin D toxicity, the markedly elevated levels of 25(OH)D may lead to hypercalcemia by interacting directly with the VDR and by displacing $1,25(OH)_2D$ from vitamin D–binding protein, resulting in increased bioavailability of the active hormone.

The VDR is expressed in a wide range of cells and tissues. The molecular actions of $1,25(OH)_2D$ have been studied most extensively in tissues involved in the regulation of mineral ion homeostasis. This hormone is a major inducer of calbindin 9K, a calcium-binding protein expressed in the intestine, which is thought to play an important role in the active transport of calcium across the enterocyte. The two major calcium transporters expressed by intestinal epithelia, TRPV5 and TRPV6 (transient receptor potential vanilloid), are also vitamin D responsive. By inducing the expression of these and other genes in the small intestine, $1,25(OH)_2D$ increases the efficiency of intestinal calcium absorption, and it also has been shown to have several important actions in the skeleton. The VDR is expressed in osteoblasts and regulates the expression of several genes in this cell. These genes include the bone matrix proteins, osteocalcin and osteopontin, which are upregulated by $1,25(OH)_2D$, in addition to type I collagen, which is transcriptionally repressed by $1,25(OH)_2D$. Both $1,25(OH)_2D$ and PTH induce the expression of RANK ligand, which promotes osteoclast differentiation and increases osteoclast activity, by binding to RANK on osteoclast progenitors and mature osteoclasts. This is the mechanism by which $1,25(OH)_2D$ induces bone resorption. $1,25(OH)_2D$ regulates phosphate homeostasis, primarily by inducing the expression of FGF23 in osteocytes. However, the skeletal features associated with VDR-knockout mice (rickets, osteomalacia) are largely corrected by increasing calcium and phosphorus intake, underscoring the importance of vitamin D action in the gut.

The VDR is expressed in the parathyroid gland, and $1,25(OH)_2D$ has been shown to have antiproliferative effects on parathyroid cells and to suppress the transcription of the parathyroid hormone gene. These effects of $1,25(OH)_2D$ on the parathyroid gland are an important part of the rationale for current therapies directed at preventing and treating hyperparathyroidism associated with renal insufficiency.

The VDR is also expressed in tissues and organs that do not play a role in mineral ion homeostasis. Notable in this respect is the observation that $1,25(OH)_2D$ has an antiproliferative effect on several cell types, including keratinocytes, breast cancer cells, and prostate cancer cells. The effects of $1,25(OH)_2D$ and the VDR on keratinocytes are particularly intriguing, since the VDR is primarily a transcriptional repressor in these cells. Alopecia is seen in humans and mice with mutant VDRs but is not a feature of vitamin D deficiency; thus, the effects of the VDR on the hair follicle are ligand-independent.

VITAMIN D DEFICIENCY

The mounting concern about the relationship between solar exposure and the development of skin cancer has led to increased reliance on dietary sources of vitamin D. Although the prevalence of vitamin D deficiency varies, the third National Health and Nutrition Examination Survey (NHANES III) revealed that vitamin D deficiency is prevalent throughout the United States, the prevalence being >29% in obese children. The clinical syndrome of vitamin D deficiency can be a result of deficient production of vitamin D in the skin, lack of dietary intake, accelerated losses of vitamin D, impaired vitamin D activation, or resistance to the biologic effects of $1,25(OH)_2D$ **(Table 402-6)**. The elderly and nursing home residents are particularly at risk for vitamin D deficiency, since both the efficiency of vitamin D synthesis in the skin and the absorption of vitamin D from the intestine decline with age. The presence of terminal ileal disease also results in impaired enterohepatic circulation of vitamin D metabolites. While intestinal malabsorption of dietary fats and short bowel syndrome, including that associated with intestinal bypass surgery, lead to vitamin D deficiency, the cause of vitamin D deficiency in obese individuals is poorly understood. In addition to intestinal diseases, accelerated inactivation of vitamin D metabolites can be seen with drugs that induce hepatic cytochrome P450 mixed-function oxidases such as barbiturates, phenytoin, and rifampin. Impaired 25-hydroxylation, associated with severe liver disease or isoniazid, is an uncommon cause of vitamin D deficiency. A mutation in the gene responsible for 25-

hydroxylation has been identified in a few kindreds. Impaired 1α-hydroxylation is prevalent in the population with profound renal dysfunction due to an increase in circulating FGF23 levels. Thus, therapeutic interventions should be considered in patients whose creatinine clearance is <0.5 mL/s (30 mL/min). Mutations in the renal 1α-hydroxylase are the basis for the genetic disorder, pseudovitamin D–deficiency rickets. This autosomal recessive disorder presents with the syndrome of vitamin D deficiency in the first year of life. Patients present with growth retardation, rickets, and hypocalcemic seizures. Serum $1,25(OH)_2D$ levels are low despite normal 25(OH)D levels and elevated PTH levels. Treatment with vitamin D metabolites that do not require 1α-hydroxylation for activity results in disease remission, although lifelong therapy is required. A second autosomal recessive disorder, hereditary vitamin D–resistant rickets, a consequence of vitamin D receptor mutations, is a greater therapeutic challenge. These patients present in a similar fashion during the first year of life, but alopecia often accompanies the disorder, demonstrating a functional role of the VDR in the keratinocyte stem cell population required for hair follicle regeneration. Serum levels of $1,25(OH)_2D$ are dramatically elevated in these individuals both because of increased production due to stimulation of 1α-hydroxylase activity as a consequence of secondary hyperparathyroidism and because of impaired inactivation, since induction of the 24-hydroxylase by $1,25(OH)_2D$ requires an intact VDR. Since the receptor mutation results in hormone resistance, daily calcium and phosphorus infusions may be required to bypass the defect in intestinal mineral ion absorption.

TABLE 402-6

Causes of Impaired Vitamin D Action

Vitamin D deficiency	Impaired 1α-hydroxylation
Impaired cutaneous production	Hypoparathyroidism
Dietary absence	Ketoconazole
Malabsorption	1α-hydroxylase mutation
Accelerated loss of vitamin D	FGF23 excess
Increased metabolism (barbiturates, phenytoin, rifampin)	Oncogenic osteomalacia
Impaired enterohepatic circulation	X-linked hypophosphatemic rickets
Nephrotic syndrome	Renal Failure
Impaired 25-hydroxylation	Target Organ Resistance
Liver disease, isoniazid	Vitamin D receptor mutation
25-hydroxylase mutation	Phenytoin
	Other
	Obesity

Regardless of the cause, the clinical manifestations of vitamin D deficiency are largely a consequence of impaired intestinal calcium absorption. Mild to moderate vitamin D deficiency is asymptomatic, whereas long-standing vitamin D deficiency results in hypocalcemia accompanied by secondary hyperparathyroidism, impaired mineralization of the skeleton (osteopenia on X-ray or decreased bone mineral density), and proximal myopathy. Vitamin D deficiency also has been shown to be associated with an increase in overall mortality, including cardiovascular causes. In the absence of an intercurrent illness, the hypocalcemia associated with long-standing vitamin D deficiency rarely presents with acute symptoms of hypocalcemia such as numbness, tingling, and seizures. However, the concurrent development of hypomagnesemia, which impairs parathyroid function, or the administration of potent bisphosphonates, which impair bone resorption, can lead to acute symptomatic hypocalcemia in vitamin D–deficient individuals.

Rickets and Osteomalacia

In children, before epiphyseal fusion, vitamin D deficiency results in growth retardation associated with an expansion of the growth plate known as *rickets*. Three layers of chondrocytes are present in the normal growth plate: the reserve zone, the proliferating zone, and the hypertrophic zone. Rickets associated with impaired vitamin D action is characterized by expansion of the hypertrophic chondrocyte layer. The proliferation and differentiation of the chondrocytes in the rachitic growth plate are normal, and the expansion of the growth plate is a consequence of impaired apoptosis of the late hypertrophic chondrocytes, an event that precedes replacement of these cells by osteoblasts during endochondral bone formation. Investigations in murine models demonstrate that hypophosphatemia, which in vitamin D deficiency is a consequence of secondary hyperparathyroidism, is a key etiologic factor in the development of the rachitic growth plate.

The hypocalcemia and hypophosphatemia that accompany vitamin D deficiency result in impaired mineralization of bone matrix proteins, a condition known as *osteomalacia*. Osteomalacia is also a feature of long-standing hypophosphatemia, which may result from renal phosphate wasting, or chronic use of etidronate or phosphate-binding antacids. This hypomineralized matrix is biomechanically inferior to normal bone; as a result, patients with osteomalacia are prone to bowing of weight-bearing extremities and skeletal fractures. Vitamin D and calcium supplementation have been shown to decrease the incidence of hip fracture among ambulatory nursing home residents in France, suggesting that undermineralization of bone

contributes significantly to morbidity in the elderly. Proximal myopathy is a striking feature of severe vitamin D deficiency both in children and in adults. Rapid resolution of the myopathy is observed upon vitamin D treatment.

Though vitamin D deficiency is the most common cause of rickets and osteomalacia, many disorders lead to inadequate mineralization of the growth plate and bone. Calcium deficiency without vitamin D deficiency, the disorders of vitamin D metabolism previously discussed, and hypophosphatemia can all lead to inefficient mineralization. Even in the presence of normal calcium and phosphate levels, chronic acidosis and drugs such as bisphosphonates can lead to osteomalacia. The inorganic calcium/phosphate mineral phase of bone cannot form at low pH. Bisphosphonates bind to and prevent hydroxyapatite crystal growth. Since alkaline phosphatase is necessary for normal mineral deposition, probably because the enzyme can hydrolyze inhibitors of mineralization such as inorganic pyrophosphate, genetic inactivation of the alkaline phosphatase gene (hereditary hypophosphatasia) also can lead to osteomalacia in the setting of normal calcium and phosphate levels.

Diagnosis of Vitamin D Deficiency, Rickets, and Osteomalacia

The most specific screening test for vitamin D deficiency in otherwise healthy individuals is a serum 25(OH)D level. Although the normal ranges vary, levels of 25(OH)D <37 nmol/L (<15 ng/mL) are associated with increasing PTH levels and lower bone density. The National Academy of Medicine has defined vitamin D sufficiency as a vitamin D level >50 nmol/L (>20 ng/mL), although higher levels may be required to optimize intestinal calcium absorption in the elderly and those with underlying disease states, including obesity. Vitamin D deficiency leads to impaired intestinal absorption of calcium, resulting in decreased serum total and ionized calcium values. This hypocalcemia results in secondary hyperparathyroidism, a homeostatic response that initially maintains serum calcium levels at the expense of the skeleton. Due to the PTH-induced increase in bone turnover, alkaline phosphatase levels are often increased. In addition to increasing bone resorption, PTH decreases urinary calcium excretion while promoting phosphaturia. This results in hypophosphatemia, which exacerbates the mineralization defect in the skeleton. With prolonged vitamin D deficiency resulting in osteomalacia, calcium stores in the skeleton become relatively inaccessible, since osteoclasts cannot resorb unmineralized osteoid, and frank hypocalcemia ensues. Since PTH is a major stimulus for the renal 25(OH)D 1α-hydroxylase, there is increased synthesis of the active hormone, $1,25(OH)_2D$. Paradoxically, levels of this hormone are often normal in severe vitamin D deficiency. Therefore, measurements of $1,25(OH)_2D$ are not accurate reflections of vitamin D stores and should not be used to diagnose vitamin D deficiency in patients with normal renal function.

Radiologic features of vitamin D deficiency in children include a widened, expanded growth plate that is characteristic of rickets. These findings not only are apparent in the long bones but also are present at the costochondral junction, where the expansion of the growth plate leads to swellings known as the “rachitic rosary.” Impairment of intramembranous bone mineralization leads to delayed fusion of the calvarial sutures and a decrease in the radiopacity of cortical bone in the long bones. If vitamin D deficiency occurs after epiphyseal fusion, the main radiologic finding is a decrease in cortical thickness and relative radiolucency of the skeleton. A specific radiologic feature of osteomalacia, whether associated with phosphate wasting or vitamin D deficiency, is pseudofractures, or Looser’s zones. These are radiolucent lines that occur where large arteries are in contact with the underlying skeletal elements; it is thought that the arterial pulsations lead to the radiolucencies. As a result, these pseudofractures are usually a few millimeters wide, are several centimeters long, and are seen particularly in the scapula, the pelvis, and the femoral neck.

TREATMENT

TREATMENT

Vitamin D Deficiency

Based on the National Academy of Medicine 2010 report, the recommended daily intake of vitamin D is 600 IU from 1 to 70 years of age, and 800 IU for those over 70. Based on the observation that 800 IU of vitamin D, with calcium supplementation, decreases the risk of hip fractures in elderly women, this higher dose is thought to be an appropriate daily intake for prevention of vitamin D deficiency in adults. The safety margin for vitamin D is large, and vitamin D toxicity usually is observed only in patients taking doses in the range of 40,000 IU daily. Treatment of vitamin D deficiency should be directed at the underlying disorder, if possible, and also should be tailored to the severity of the condition. Vitamin D should always be repleted in conjunction with calcium supplementation since most of the consequences of vitamin D deficiency are a result of impaired mineral ion homeostasis. In patients in whom 1α-hydroxylation is impaired, metabolites that do not require this activation step are the treatment of choice. They include $1,25(OH)_2D_3$ (calcitriol [Rocaltrol], 0.25–0.5 μg/d) and 1α-hydroxyvitamin D_2 (doxercalciferol [Hectorol], 2.5–5 μg/d). If the pathway required for activation of vitamin D is intact, severe vitamin D deficiency can be treated with pharmacologic repletion initially (50,000 IU weekly for 3–12 weeks), followed by maintenance therapy (800 IU daily). Pharmacologic doses may be required for maintenance therapy in patients who are taking medications such as barbiturates or phenytoin, that accelerate metabolism of, or cause resistance to $1,25(OH)_2D$. Polymorphisms in the 25 hydroxylase and the 24 hydroxylase genes can also lead to different responses to the normal recommended daily intake of vitamin D. Calcium supplementation should include 1.5–2 g/d of elemental calcium. Normocalcemia is usually observed within 1 week of the institution of therapy, although increases in PTH and alkaline phosphatase levels may persist for 3–6 months. The most efficacious methods to monitor treatment and resolution of vitamin D deficiency are serum and urinary calcium measurements. In patients who are vitamin D replete and are taking adequate calcium supplementation, the 24-h urinary calcium excretion should be in the range of 100–250 mg/24 h. Lower levels suggest problems with adherence to the treatment regimen or

with absorption of calcium or vitamin D supplements. Levels >250 mg/24 h predispose to nephrolithiasis and should lead to a reduction in vitamin D dosage and/or calcium supplementation.

FURTHER READING

Bikle D et al: Vitamin D metabolites in captivity? Should we measure free or total 25(OH)D to assess vitamin D status? J Steroid Biochem Mol Biol 173:105, 2017. [PubMed: 28093353]

Dallas SL et al: The osteocyte: An endocrine cell and more. Endocr Rev 34:658, 2013. [PubMed: 23612223]

de Baaij JH et al: Magnesium in man: Implications for health and disease. Physiol Rev 95:1, 2015. [PubMed: 25540137]

Felsenfeld AJ, Levine BS: Approach to treatment of hypophosphatemia. Am J Kidney Dis 60:655, 2012. [PubMed: 22863286]

Feng X, Teitelbaum SL: Osteoclasts: New insights. Bone Res 1:11, 2013. [PubMed: 26273491]

Heaney RP, Holick MF: Why the IOM recommendations for vitamin D are deficient. J Bone Miner Res 26:455, 2011. [PubMed: 21337617]

Institute of Medicine (IOM): *Dietary Reference Intakes for Calcium and Vitamin D*. Washington, DC: National Academies Press, 2011.

Siddiqui JA, Partridge NC: Physiological bone remodeling: Systemic regulation and growth factor involvement. Physiology 31:233, 2016. [PubMed: 27053737]

Zofkova I: Hypercalcemia. Pathophysiological aspects. Phys Res 65:1, 2016.

Chapter 403: Disorders of the Parathyroid Gland and Calcium Homeostasis

John T. Potts, Jr.; Harald W. Jüppner

INTRODUCTION

The four parathyroid glands are located posterior to the thyroid gland. They produce parathyroid hormone (PTH), which is the primary regulator of calcium physiology. PTH acts directly on bone, where it induces calcium release, and on the kidney, where it enhances calcium reabsorption in the distal tubules, and in the proximal renal tubules the synthesis of 1,25-dihydroxyvitamin D (1,25$[OH]_2$D), a hormone that increases gastrointestinal calcium absorption. Serum PTH levels are tightly regulated by a negative feedback loop. Calcium, acting through the calcium-sensing receptor (CaSR), and vitamin D, acting through its nuclear receptor, reduce PTH release and synthesis. Additional evidence indicates that fibroblast growth factor 23 (FGF23), a phosphaturic hormone, can suppress PTH secretion. Understanding the hormonal pathways that regulate calcium levels and bone metabolism is essential for effective diagnosis and management of a wide array of hyper- and hypocalcemic disorders.

Hyperparathyroidism (HPT), characterized by excess production of PTH, is a common cause of hypercalcemia and is usually the result of autonomously functioning adenomas or hyperplasia. Surgery for this disorder is highly effective and has been shown to reverse some of the deleterious effects of long-standing PTH excess on bone density. Humoral hypercalcemia of malignancy is also common and is usually due to the overproduction of parathyroid hormone–related peptide (PTHrP) by cancer cells. The similarities in the biochemical characteristics of HPT and humoral hypercalcemia of malignancy, first noted by Albright in 1941, are now known to reflect the actions of PTH and PTHrP through the same G protein–coupled PTH/PTHrP receptor.

The genetic basis of multiple endocrine neoplasia (MEN) types 1 and 2, familial hypocalciuric hypercalcemia (FHH), different forms of pseudohypoparathyroidism (PHP), Jansen's syndrome, disorders of vitamin D synthesis and action, and the molecular events associated with parathyroid gland neoplasia have provided new insights into the regulation of calcium homeostasis. PTH and *possibly some* of its analogues are promising therapeutic agents for the treatment of postmenopausal or senile osteoporosis, and calcimimetic agents, which activate the CaSR, have provided new approaches for PTH suppression.

PARATHYROID HORMONE

PHYSIOLOGY

The primary function of PTH is to maintain the extracellular fluid (ECF) calcium concentration within a narrow normal range. The hormone acts directly on bone and kidney and indirectly on the intestine through its effects on synthesis of 1,25$(OH)_2$D to increase serum calcium concentrations; in turn, PTH production is closely regulated by the concentration of serum ionized calcium. This feedback system is the critical homeostatic mechanism for maintenance of ECF calcium. Any tendency toward hypocalcemia, as might be induced by calcium- or vitamin D-deficient diets, is counteracted by an increased secretion of PTH. This in turn (1) increases the rate of dissolution of bone mineral, thereby increasing the flow of calcium from bone into blood; (2) reduces the renal clearance of calcium, returning more of the calcium and phosphate filtered at the glomerulus into ECF; (3) increases the efficiency of calcium absorption in the intestine by stimulating the production of 1,25$(OH)_2$D. Immediate control of blood calcium is due to PTH effects on bone and, to a lesser extent, on renal calcium clearance. Maintenance of steady-state calcium balance, on the other hand, probably results from the effects of 1,25$(OH)_2$D on calcium absorption **(Chap. 402)**. The renal actions of the hormone are exerted at multiple sites and include inhibition of phosphate transport (proximal tubule), augmentation of calcium reabsorption (distal tubule), and stimulation of the renal 25(OH)D-1α-hydroxylase. As much as 12 mmol (500 mg) calcium is transferred between the ECF and bone each day (a large amount in relation to the total ECF calcium pool), and PTH has a major effect on this transfer. The homeostatic role of the hormone can preserve calcium concentration in blood at the cost of bone demineralization.

PTH has multiple actions on bone, some direct and some indirect. PTH-mediated changes in bone calcium release can be seen within minutes. The chronic effects of PTH are to increase the number of bone cells, both osteoblasts and osteoclasts, and to increase the remodeling of bone; these effects are apparent within hours after the hormone is given and persist for hours after PTH is withdrawn. Continuous exposure to elevated PTH (as in HPT or long-term infusions in animals) leads to increased osteoclast-mediated bone resorption. However, the intermittent administration of PTH, elevating hormone levels for 1–2 hours each day, leads to a net stimulation of bone formation rather than bone breakdown. Striking increases, especially in trabecular bone in the spine and hip, have been reported with the use of PTH in combination with estrogen. PTH(1–34) as monotherapy caused a highly significant reduction in fracture incidence in a worldwide placebo-controlled trial.

Osteoblasts (or stromal cell precursors), which have PTH/PTHrP receptors, are crucial to this bone-forming effect of PTH; osteoclasts, which mediate bone breakdown, lack such receptors. PTH-mediated stimulation of osteoclasts is indirect, acting in part, through cytokines released from osteoblasts to activate osteoclasts; in experimental studies of bone resorption in vitro, osteoblasts must be present for PTH to activate osteoclasts to resorb bone **(Chap. 402)**.

STRUCTURE

PTH is an 84-amino-acid single-chain peptide. The amino-terminal portion, PTH(1–34), is highly conserved and is critical for the biologic actions of the molecule. Modified synthetic fragments of the amino-terminal sequence as small as PTH(1–11) are sufficient to activate the PTH/PTHrP receptor (see below). The carboxyl-terminal regions of the full-length PTH(1–84) molecule also can bind to a separate binding protein/receptor (cPTH-R), but this receptor has been incompletely characterized. Fragments shortened at the amino-terminus possibly by binding to cPTH-R can reduce, directly or indirectly, some of the biologic actions of full-length PTH(1–84) and of PTH(1–34).

BIOSYNTHESIS, SECRETION, AND METABOLISM

Synthesis

Parathyroid cells have multiple methods of adapting to increased needs for PTH production. Most rapid (within minutes) is secretion of preformed hormone in response to hypocalcemia. Second, within hours, PTH mRNA expression is induced by sustained hypocalcemia. Finally, protracted challenge leads within days to cellular replication to increase parathyroid gland mass.

PTH is initially synthesized as a larger molecule (preproparathyroid hormone, consisting of 115 amino acids). After a first cleavage step to remove the "pre" sequence of 25 amino acid residues, a second cleavage step removes the "pro" sequence of 6 amino acid residues before secretion of the mature peptide comprising 84 residues. Mutations in the preprotein region of the gene can cause hypoparathyroidism by interfering with hormone synthesis, transport, or secretion.

Transcriptional suppression of the PTH gene by calcium is nearly maximal at physiologic calcium concentrations. Hypocalcemia increases transcriptional activity within hours. $1,25(OH)_2D$ strongly suppresses PTH gene transcription. In patients with renal failure, IV administration of supraphysiologic levels of $1,25(OH)_2D$ or analogues of this active metabolite can dramatically suppress PTH overproduction, which is sometimes difficult to control due to severe secondary HPT. Regulation of proteolytic destruction of preformed hormone (posttranslational regulation of hormone production) is an important mechanism for mediating rapidly (within minutes) changes in hormone availability. High calcium increases and low calcium inhibits the proteolytic destruction of stored hormone.

Regulation of PTH Secretion

PTH secretion increases steeply to a maximum value of about five times the basal rate of secretion as the calcium concentration falls from normal to the range of 1.9–2.0 mmol/L (7.6–8.0 mg/dL; measured as total calcium). However, the ionized fraction of blood calcium is the important determinant of hormone secretion. Severe intracellular magnesium deficiency impairs PTH secretion (see below).

ECF calcium controls PTH secretion by interaction with a CaSR, a G protein–coupled receptor (GPCR) for which Ca^{2+} ions act as the primary ligand (see below). This receptor is a member of a distinctive subgroup of the GPCR superfamily that mediates its actions through two related signaling G proteins, namely Gq and G11, is characterized by a large extracellular domain suitable for "clamping" the small-molecule ligand. Stimulation of the CaSR by high calcium levels suppresses PTH secretion. The CaSR is present in parathyroid glands and the calcitonin-secreting cells of the thyroid (C cells), as well as in multiple other sites, including brain and kidney. Genetic evidence has revealed a key biologic role for the CaSR in parathyroid gland responsiveness to calcium and in renal calcium clearance. Heterozygous loss-of-function mutations in CaSR cause the syndrome of FHH, in which the blood calcium abnormality resembles that observed in HPT but with hypocalciuria; two more recently defined variants of FHH, FHH2, and FHH3, are caused either by heterozygous mutations in G11, one of the signaling proteins down-stream of the CaSR, or by heterozygous mutations in *AP2A1*. Homozygous loss-of-function mutations in the CaSR are the cause of severe neonatal HPT, a disorder that can be lethal if not treated within the first days of life. On the other hand, heterozygous gain-of-function mutations cause a form of hypocalcemia resembling hypoparathyroidism (see below).

Metabolism

The secreted form of PTH is indistinguishable by immunologic criteria and by molecular size from the 84-amino-acid peptide (PTH[1–84]) extracted from glands. However, much of the immunoreactive material found in the circulation is smaller than the extracted or secreted hormone. The principal circulating fragments of immunoreactive hormone lack a portion of the critical amino-terminal sequence required for biologic activity and, hence, are biologically inactive fragments (so-called middle and carboxyl-terminal fragments). Much of the proteolysis of hormone occurs in the liver and kidney. Peripheral metabolism of PTH does not appear to be regulated by physiologic states (high versus low calcium, etc.); hence, peripheral metabolism of hormone, although responsible for rapid clearance of secreted hormone, appears to be a high-capacity, metabolically invariant catabolic process.

The rate of clearance of the secreted 84-amino-acid peptide from blood is more rapid than the rate of clearance of the biologically inactive fragment(s) corresponding to the middle and carboxyl-terminal regions of PTH. Consequently, the interpretation of results obtained with earlier PTH radioimmunoassays was influenced by the nature of the peptide fragments detected by the antibodies.

Although the problems inherent in PTH measurements have been largely circumvented by use of double-antibody immunometric assays, it is now known that some of these assays detect, besides the intact molecule, large amino-terminally truncated forms of PTH, which are present in normal and uremic individuals in addition to PTH(1–84). The concentration of these fragments relative to that of intact PTH(1–84) is higher with induced hypercalcemia than in eucalcemic or hypocalcemic conditions and is higher in patients with impaired renal function. PTH(7–84) has been identified as a major component of these amino-terminally truncated fragments. Growing evidence suggests that the PTH(7–84) (and probably related amino-terminally truncated fragments) can act, through yet undefined mechanisms, as an inhibitor of PTH action and may be of clinical significance, particularly in patients with chronic kidney disease (CKD). In this group of patients, efforts to prevent secondary HPT by a variety of measures (vitamin D analogues, higher calcium intake, higher dialysate calcium, phosphate-lowering strategies, and calcimetic drugs) can lead to oversuppression of the parathyroid glands since some amino-terminally truncated PTH fragments, such as PTH(7–84), react in many immunometric PTH assays (now termed second-generation assays; see below under "Diagnosis"), thus overestimating the levels of biologically active, intact PTH. Such excessive parathyroid gland suppression in CKD can lead to adynamic bone disease (see below), which has been associated in children with further impaired growth and increased bone fracture rates in adults, and can furthermore lead to significant hypercalcemia. The measurement of PTH with newer third-generation immunometric assays, which use detection antibodies directed against extreme amino-terminal PTH epitopes and thus detect only full-length PTH(1–84), may provide some advantage to prevent bone disease in CKD.

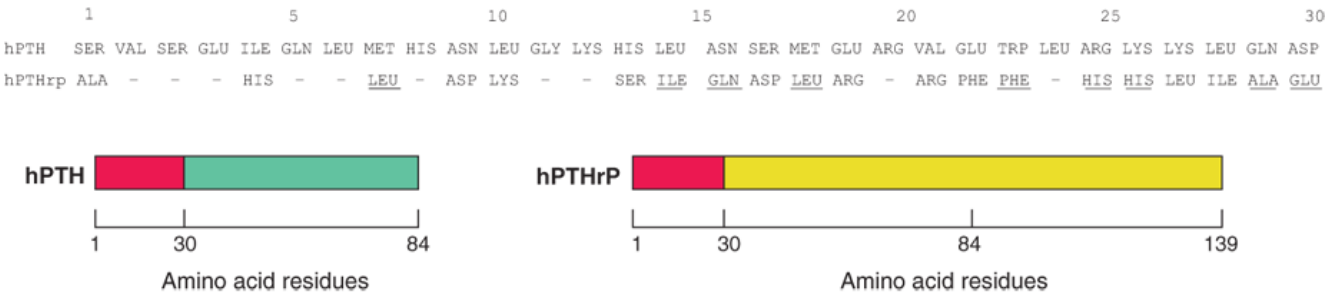
PARATHYROID HORMONE–RELATED PROTEIN (PTHrP)

PTHrP is responsible for most instances of humoral hypercalcemia of malignancy **(Chap. 89)**, a syndrome that resembles primary HPT but without elevated PTH levels. Most cell types normally produce PTHrP, including brain, pancreas, heart, lung, mammary tissue, placenta, endothelial cells, and smooth muscle. In fetal animals, PTHrP directs transplacental calcium transfer, and high concentrations of PTHrP are produced in mammary tissue and secreted into milk, but the biologic significance of the very high concentrations of this hormone in breast milk is unknown. PTHrP also plays an essential role in endochondral bone formation and in branching morphogenesis of the breast, and possibly in uterine contraction and other biologic functions.

PTH and PTHrP, although products of different genes, exhibit considerable functional and structural homology **(Fig. 403-1)** and have evolved from a shared ancestral gene. The structure of the gene encoding human PTHrP, however, is more complex than that of PTH, containing multiple additional exons, which can undergo alternate splicing patterns during formation of the mature mRNA. Protein products of 139, 141, and 173 amino acids are produced, and other molecular forms may result from tissue-specific degradation at accessible internal cleavage sites. The biologic roles of these various molecular species and the nature of the circulating forms of PTHrP are unclear. In fact, it is uncertain whether PTHrP circulates at any significant level in adults. As a paracrine factor, PTHrP may be produced, act, and be destroyed locally within tissues. In adults, PTHrP appears to have little influence on calcium homeostasis, except in disease states, when large tumors, especially of the squamous cell type as well as renal cell carcinomas, lead to massive overproduction of the hormone and hypercalcemia.

FIGURE 403-1

Schematic diagram to illustrate similarities and differences in structure of human parathyroid hormone (PTH) and human PTH-related peptide (PTHrP). Close structural (and functional) homology exists between the first 30 amino acids of hPTH and hPTHrP (red area). The PTHrP sequence may be ≥139 amino acid residues in length. PTH is only 84 residues long; after residue 30, there is little structural homology between the two. Dashed lines in the PTHrP sequence indicate identity; underlined residues, although different from those of PTH, still represent conservative changes (charge or polarity preserved). Ten amino acids are identical, and a total of 20 of 30 are homologues.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PTH AND PTHrP HORMONE ACTION

Both PTH and PTHrP bind to and activate the PTH/PTHrP receptor. The PTH/PTHrP receptor (also known as the PTH-1 receptor, PTH1R) belongs to a subfamily of GPCRs that includes the receptors for calcitonin, glucagon, secretin, vasoactive intestinal peptide, and other peptides. Although both ligands

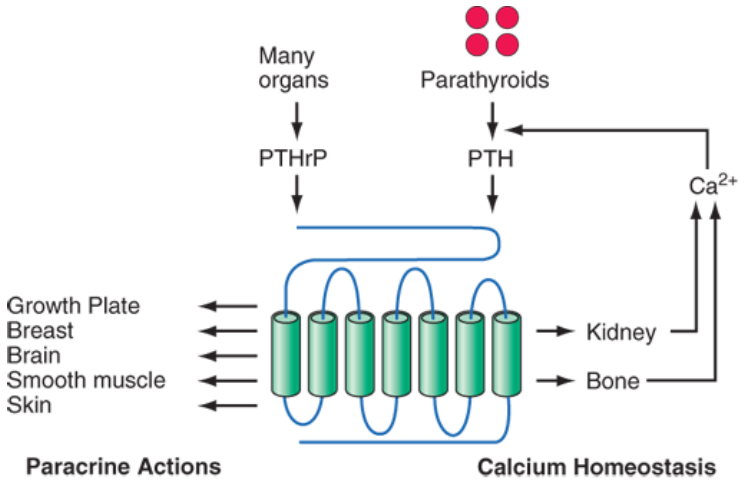
activate the PTH1R, the two peptides induce distinct responses in the receptor, which explains how a single receptor without isoforms can serve two biologic roles. The extracellular regions of the receptor are involved in hormone binding, and the intracellular domains, after hormone activation, bind G protein subunits to transduce hormone signaling into cellular responses through the stimulation of second messenger formation. A second receptor that binds PTH, originally termed the *PTH-2 receptor* (PTH2R), is primarily expressed in brain, pancreas, and testis. Different mammalian PTH1Rs respond equivalently to PTH and PTHrP, at least when tested with traditional assays, whereas only the human PTH2R responds efficiently to PTH (but not to PTHrP). PTH2Rs from other species show little or no stimulation of second-messenger formation in response to PTH or PTHrP. The endogenous ligand of the PTH2R was shown to be a hypothalamic peptide referred to as tubular infundibular peptide of 39 residues, TIP39, that is distantly related to PTH and PTHrP. The PTH1R and the PTH2R can be traced backward in evolutionary time to fish; in fact, the zebrafish genome contains, in addition to the PTH1R and the PTH2R orthologs, a third receptor, the PTH3R, that is more closely related to the fish PTH1R than to the fish PTH2R. The evolutionary conservation of structure and function suggests important biologic roles for these receptors, even in fish, which lack discrete parathyroid glands but produce two molecules that are closely related to mammalian PTH.

Studies using the cloned PTH1R confirm that it can be coupled to more than one G protein and second-messenger pathway, apparently explaining the multiplicity of pathways stimulated by PTH. Activation of protein kinases (A and C) and calcium transport channels is associated with a variety of hormone-specific tissue responses. These responses include inhibition of phosphate and bicarbonate transport, stimulation of calcium transport, and activation of renal 1α-hydroxylase in the kidney. The responses in bone include effects on collagen synthesis, alkaline phosphatase, ornithine decarboxylase, citrate decarboxylase, and glucose-6-phosphate dehydrogenase activities, phospholipid synthesis, as well as calcium and phosphate transport. Ultimately, these biochemical events lead to an integrated hormonal response in bone turnover and calcium homeostasis. PTH also activates Na^+/Ca^{2+} exchangers at renal distal tubular sites and stimulates translocation of preformed calcium transport channels, moving them from the interior to the apical surface to increase tubular uptake of calcium. PTH-dependent stimulation of phosphate excretion (reducing reabsorption—the opposite effect from actions on calcium in the kidney) involves the down-regulation of two sodium-dependent phosphate co-transporters, NPT2a and NPT2c, and their expression at the apical membrane, thereby reducing phosphate reabsorption in the proximal renal tubules. Similar mechanisms may be involved in other renal tubular transporters that are influenced by PTH. Recent studies reaffirm the critical linkage of blood phosphate lowering to net calcium entry into blood by PTH action and emphasize the participation of bone cells other than osteoclasts in the rapid calcium elevating actions of PTH.

PTHrP exerts important developmental influences on fetal bone development and in adult physiology. A homozygous ablation of the gene encoding PTHrP (or disruption of the PTH1R gene) in mice causes a lethal phenotype in which animals are born with pronounced acceleration of chondrocyte maturation that resembles a lethal form of chondrodysplasia in humans that is caused by homozygous or compound heterozygous, inactivating PTH1R mutations **(Fig. 403-2)**. Heterozygous PTH1R mutations in humans furthermore can be a cause of delayed tooth eruption and mice that are heterozygous for ablation of the PTHrP gene display reduced mineral density consistent with osteoporosis. Experiments with these mouse models point to a hitherto unappreciated role of PTHrP as a paracrine/autocrine factor that modulates bone metabolism in adults as well as during bone development.

FIGURE 403-2

Dual role for the actions of the PTH/PTHrP receptor (PTH1R). Parathyroid hormone (PTH; endocrine-calcium homeostasis) and PTH-related peptide (PTHrP; paracrine–multiple tissue actions including growth plate cartilage in developing bone) use the single receptor for their disparate functions mediated by the amino-terminal 34 residues of either peptide. Other regions of both ligands interact with other receptors (not shown).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CALCITONIN

Calcitonin is a hypocalcemic peptide hormone that in several mammalian species acts as an indirect antagonist to the calcemic actions of PTH **(See also Chap. 381)**. Calcitonin seems to be of limited physiologic significance in humans, at least with regard to calcium homeostasis. It is of medical significance

because of its role as a tumor marker in sporadic and hereditary cases of medullary carcinoma and its medical use as an adjunctive treatment in severe hypercalcemia and in Paget's disease of bone.

The hypocalcemic activity of calcitonin is accounted for primarily by inhibition of osteoclast-mediated bone resorption and secondarily by stimulation of renal calcium clearance. These effects are mediated by receptors on osteoclasts and renal tubular cells. Calcitonin exerts additional effects through receptors present in the brain, the gastrointestinal tract, and the immune system. The hormone, for example, exerts analgesic effects directly on cells in the hypothalamus and related structures, possibly by interacting with receptors for related peptide hormones such as calcitonin gene–related peptide (CGRP) or amylin. Both of these ligands have specific high-affinity receptors that share considerable structural similarity with the PTH1R and can also bind to and activate calcitonin receptors. The calcitonin receptor shares considerable structural similarity with the PTH1R.

The thyroid is the major source of the hormone, and the cells involved in calcitonin synthesis arise from neural crest tissue. During embryogenesis, these cells migrate into the ultimobranchial body, derived from the last branchial pouch. In submammalian vertebrates, the ultimobranchial body constitutes a discrete organ, anatomically separate from the thyroid gland; in mammals, the ultimobranchial gland fuses with and is incorporated into the thyroid gland.

The naturally occurring calcitonins consist of a peptide chain of 32 amino acids. There is considerable sequence variability among species. Calcitonin from salmon, which is used therapeutically, is 10–100 times more potent than mammalian forms in lowering serum calcium.

There are two calcitonin genes, α and β; the transcriptional control of these genes is complex. Two different mRNA molecules are transcribed from the α gene; one is translated into the precursor for calcitonin, and the other message is translated into an alternative product, CGRP. CGRP is synthesized wherever the calcitonin mRNA is expressed (e.g., in medullary carcinoma of the thyroid). The β, or CGRP-2, gene is transcribed into the mRNA for CGRP in the central nervous system (CNS); this gene does not produce calcitonin, however. CGRP has cardiovascular actions and may serve as a neurotransmitter or play a developmental role in the CNS.

The circulating level of calcitonin in humans is lower than that in many other species. In humans, even extreme variations in calcitonin production do not change calcium and phosphate metabolism; no definite effects are attributable to calcitonin deficiency (totally thyroidectomized patients receiving only replacement thyroxine) or excess (patients with medullary carcinoma of the thyroid, a calcitonin-secreting tumor) **(Chap. 381)**. Calcitonin has been a useful pharmacologic agent to suppress bone resorption in Paget's disease **(Chap. 405)** and osteoporosis **(Chap. 404)** and in the treatment of hypercalcemia of malignancy (see below). However, bisphosphates are usually more effective and the physiologic role, if any, of calcitonin in humans is uncertain. On the other hand, ablation of the calcitonin gene (combined because of the close proximity with ablation of the CGRP gene) in mice leads to reduced bone mineral density, suggesting that its biologic role in mammals is still not fully understood.

HYPERCALCEMIA

Hypercalcemia can be a manifestation of a serious illness such as malignancy or can be detected coincidentally by laboratory testing in a patient with no obvious illness **(See also Chap. 50)**. The number of patients recognized with asymptomatic hypercalcemia, usually HPT, increased in the late twentieth century.

Whenever hypercalcemia is confirmed, a definitive diagnosis must be established. Although HPT, a frequent cause of asymptomatic hypercalcemia, is a chronic disorder in which manifestations, if any, may be expressed only after months or years, hypercalcemia can also be the earliest manifestation of malignancy, the second most common cause of hypercalcemia in the adult. The causes of hypercalcemia are numerous **(Table 403-1)**, but HPT and cancer account for 90% of all cases.

TABLE 403-1

Classification of Causes of Hypercalcemia

I. **Parathyroid-Related**
- A. Primary hyperparathyroidism
  1. Adenoma(s)
  2. Multiple endocrine neoplasia
  3. Carcinoma
- B. Lithium therapy
- C. Familial hypocalciuric hypercalcemia

II. **Malignancy-Related**
- A. Solid tumor with metastases (breast)
- B. Solid tumor with humoral mediation of hypercalcemia (lung, kidney)
- C. Hematologic malignancies (multiple myeloma, lymphoma, leukemia)

III. **Vitamin D–Related**
- A. Vitamin D intoxication
- B. ↑ $1,25(OH)_2D$; sarcoidosis and other granulomatous diseases
- C. ↑ $1,25(OH)_2D$; impaired $1,25(OH)_2D$ metabolism due to 24-hydroxylase deficiency and inactivating mutations in the sodium-dependent phosphate co-transporters

IV. **Associated with High Bone Turnover**
- A. Hyperthyroidism
- B. Immobilization
- C. Thiazides
- D. Vitamin A intoxication
- E. Fat necrosis

V. **Associated with Renal Failure**
- A. Severe secondary hyperparathyroidism
- B. Aluminum intoxication
- C. Milk-alkali syndrome

Before undertaking a diagnostic workup, it is essential to be sure that true hypercalcemia, not a false-positive laboratory test, is present. A false-positive diagnosis of hypercalcemia is usually the result of inadvertent hemoconcentration during blood collection or elevation in serum proteins such as albumin. Hypercalcemia is a chronic problem, and it is cost-effective to obtain several serum calcium measurements; these tests need not be in the fasting state.

Clinical features are helpful in differential diagnosis. Hypercalcemia in an adult who is asymptomatic is usually due to primary HPT. In malignancy-associated hypercalcemia, the disease is usually not occult; rather, symptoms of malignancy bring the patient to the physician, and hypercalcemia is discovered during the evaluation. In such patients, the interval between detection of hypercalcemia and death, especially without vigorous treatment, is often <6 months. Accordingly, if an asymptomatic individual has had hypercalcemia or some manifestation of hypercalcemia such as kidney stones for >1 or 2 years, it is unlikely that malignancy is the cause. Nevertheless, differentiating primary HPT from occult malignancy can occasionally be difficult, and careful evaluation is required, particularly when the duration of the hypercalcemia is unknown. Hypercalcemia not due to HPT or malignancy can result from excessive vitamin D action, impaired metabolism of $1,25(OH)_2D$, high bone turnover from any of several causes, or from renal failure (Table 403-1).

Dietary history and a history of ingestion of vitamins or drugs are often helpful in diagnosing some of the less frequent causes. Immunometric PTH assays serve as the principal laboratory test in establishing the diagnosis.

Hypercalcemia from any cause can result in fatigue, depression, mental confusion, anorexia, nausea, vomiting, constipation, reversible renal tubular defects, increased urine output, a short QT interval in the electrocardiogram, and, in some patients, cardiac arrhythmias. There is a variable relation from one patient to the next between the severity of hypercalcemia and the symptoms. Generally, symptoms are more common at calcium levels >2.9–3.0 mmol/L (11.6–12.0 mg/dL), but some patients, even at this level, are asymptomatic. When the calcium level is >3.2 mmol/L (12.8 mg/dL), calcification in kidneys, skin, vessels, lungs, heart, and stomach occurs and renal insufficiency may develop, particularly if blood phosphate levels are normal or elevated due to impaired renal excretion. Severe hypercalcemia, usually defined as ≥3.7–4.5 mmol/L (14.8–18.0 mg/dL), can be a medical emergency; coma and cardiac arrest can occur.

Acute management of the hypercalcemia is usually successful. The type of treatment is based on the severity of the hypercalcemia and the nature of associated symptoms, as outlined below.

PRIMARY HYPERPARATHYROIDISM

Natural History and Incidence

Primary HPT is a generalized disorder of calcium, phosphate, and bone metabolism due to an increased secretion of PTH. The elevation of circulating hormone usually leads to hypercalcemia and hypophosphatemia. There is great variation in the manifestations. Patients may present with multiple signs and symptoms, including recurrent nephrolithiasis, peptic ulcers, mental changes, and, less frequently, extensive bone resorption. However, with greater awareness of the disease and wider use of multiphasic screening tests, including measurements of blood calcium, the diagnosis is frequently made in patients who have no symptoms and minimal, if any, signs of the disease other than hypercalcemia and elevated levels of PTH. The manifestations may be subtle, and the disease may have a benign course for many years or a lifetime. This milder form of the disease is usually termed *asymptomatic HPT*. Rarely, HPT develops or worsens abruptly and causes severe complications such as marked dehydration and coma, so-called hypercalcemic parathyroid crisis.

The annual incidence of the disease is calculated to be as high as 0.2% in patients >60, with an estimated prevalence, including undiscovered asymptomatic patients, of ≥1%; some reports suggest the incidence may be declining. If confirmed, these changing estimates may reflect less frequent routine testing of serum calcium in recent years, earlier overestimates in incidence, or unknown factors. The disease has a peak incidence between the third and fifth decades but occurs in young children and in the elderly.

Etiology

Parathyroid tumors are most often encountered as isolated adenomas without other endocrinopathy. They may also arise in hereditary syndromes such as MEN syndromes. As many as 10% of patients with HPT are found to have mutations in 1 of 11 genes (see below). Parathyroid tumors may also arise as secondary to underlying disease (excessive stimulation in secondary HPT, especially chronic renal failure), or after other forms of excessive stimulation such as lithium therapy. These etiologies are discussed below.

SOLITARY ADENOMAS

A single abnormal gland is the cause in ~80% of patients; the abnormality in the gland is usually a benign neoplasm or adenoma and rarely a parathyroid carcinoma. Some surgeons and pathologists report that the enlargement of multiple glands is common; double adenomas are reported. In ~15% of patients, all glands are hyperfunctioning; *chief cell parathyroid hyperplasia* is usually hereditary and frequently associated with other endocrine abnormalities.

HEREDITARY SYNDROMES AND MULTIPLE PARATHYROID TUMORS

Hereditary HPT can occur without other endocrine abnormalities but is usually part of a *multiple endocrine neoplasia* syndrome **(Chap. 381)**. MEN1 (Wermer's syndrome) consists of HPT and tumors of the pituitary and pancreas, often associated with gastric hypersecretion and peptic ulcer disease (Zollinger-Ellison syndrome). MEN2A is characterized by pheochromocytoma and medullary carcinoma of the thyroid, as well as HPT; MEN2B has additional associated features such as multiple neuromas but usually lacks HPT. Each of these MEN syndromes is transmitted in an apparent autosomal dominant manner, although, as noted below, the genetic basis of MEN1 involves biallelic loss of a tumor suppressor.

The *hyperparathyroidism jaw tumor* (HPT-JT) syndrome occurs in families with parathyroid tumors (sometimes carcinomas) in association with benign jaw tumors. This disorder is caused by mutations in *CDC73* (*HRPT2*) and mutations in this gene are also observed in parathyroid cancers. Some kindreds exhibit hereditary HPT without other endocrinopathies. This disorder is often termed *nonsyndromic familial isolated hyperparathyroidism* (FIHP). There is speculation that these families may be examples of variable expression of the other syndromes such as MEN 1, MEN 2, or the HPT-JT syndrome, but they may also have distinctive, still unidentified genetic causes. For example, different heterozygous GCM2 mutations co-segregate with the disease in several FIHP kindreds; some of these mutations enhanced activity of a GCM2-dependent reporter.

Pathology

Adenomas are most often located in the inferior parathyroid glands, but in 6–10% of patients, parathyroid adenomas may be located in the thymus, the thyroid, the pericardium, or behind the esophagus. Adenomas are usually 0.5–5 g in size but may be as large as 10–20 g (normal glands weigh 25 mg on average). Chief cells are predominant in both hyperplasia and adenoma. With chief cell hyperplasia, the enlargement may be so asymmetric that some involved glands appear grossly normal. If generalized hyperplasia is present, however, histologic examination reveals a uniform pattern of chief cells and disappearance of fat even in the absence of an increase in gland weight. Thus, microscopic examination of biopsy specimens of several glands can be helpful to interpret findings at surgery.

Parathyroid carcinoma is often not aggressive. Long-term survival without recurrence is common if at initial surgery the entire gland is removed without rupture of the capsule. Recurrent parathyroid carcinoma is usually slow-growing with local spread in the neck, and surgical correction of recurrent disease may be feasible. Occasionally, however, parathyroid carcinoma is more aggressive, with distant metastases (lung, liver, and bone) found at the time of initial operation. It may be difficult to appreciate initially that a primary tumor is carcinoma; increased numbers of mitotic figures and increased fibrosis of the gland stroma may precede invasion. The diagnosis of carcinoma is often made in retrospect. HPT from a parathyroid carcinoma may be indistinguishable from other forms of primary HPT but is usually more severe clinically. A potential clue to the diagnosis is offered by the degree of calcium elevation. Calcium values of 3.5–3.7 mmol/L (14–15 mg/dL) are frequent with carcinoma and may alert the surgeon to remove the abnormal

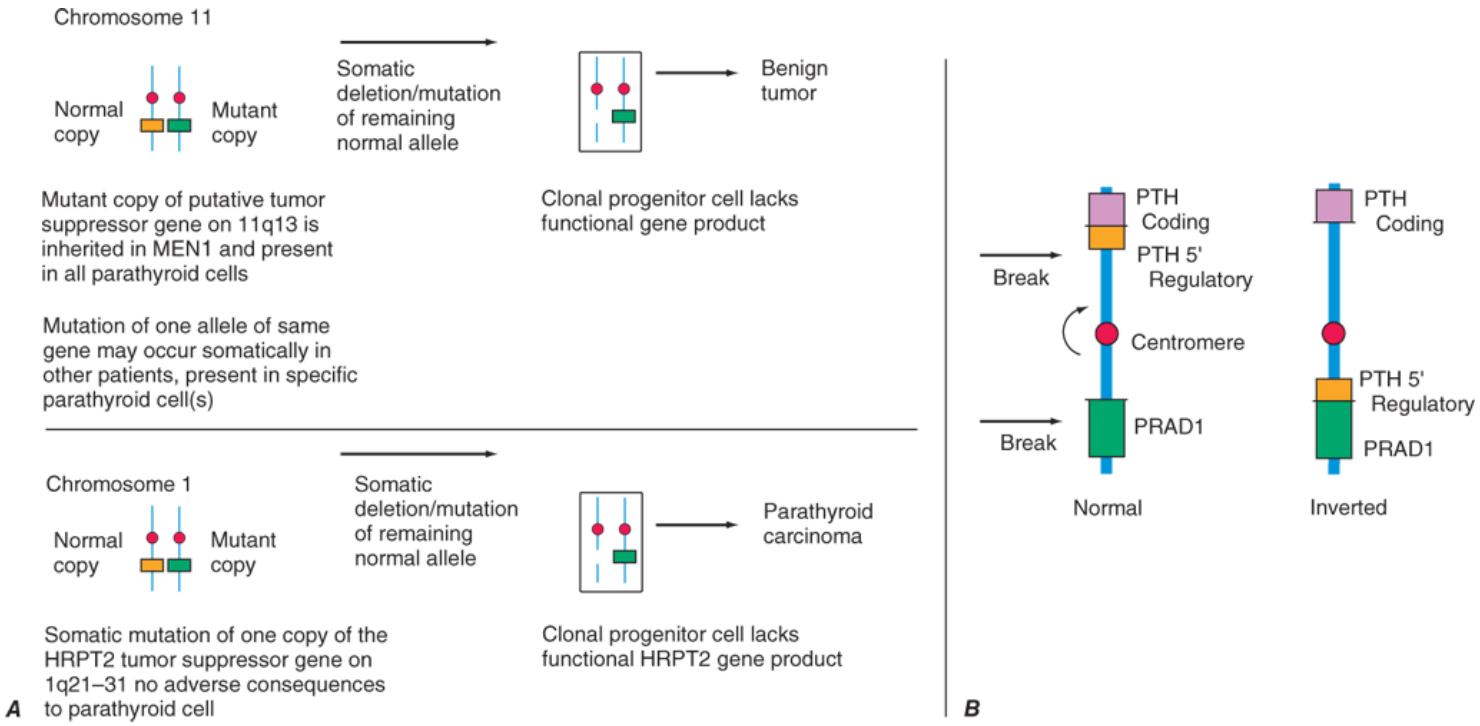
gland with care to avoid capsular rupture. Recent findings concerning the genetic basis of parathyroid carcinoma (distinct from that of benign adenomas) indicate the need, in these kindreds, for family screening (see below).

GENETIC DEFECTS ASSOCIATED WITH HPT

As in many other types of neoplasia, two fundamental types of genetic defects have been identified in parathyroid gland tumors: (1) overactivity of protooncogenes and (2) loss of function of tumor-suppressor genes. The former, by definition, can lead to uncontrolled cellular growth and function by activation (gain-of-function mutation) of a single allele of the responsible gene, whereas the latter requires loss of function of both allelic copies. Biallelic loss of function of a tumor-suppressor gene is usually characterized by a germ-line defect (all cells) and an additional somatic deletion/mutation in the tumor **(Fig. 403-3)**.

FIGURE 403-3

A. Schematic diagram indicating molecular events in tumor susceptibility. The patient with the hereditary abnormality (multiple endocrine neoplasia, or MEN) is envisioned as having one defective gene inherited from the affected parent on chromosome 11, but one copy of the normal gene is present from the other parent. In the monoclonal tumor (benign tumor), a somatic event, here partial chromosomal deletion, removes the remaining normal gene from a cell. In nonhereditary tumors, two successive somatic mutations must occur, a process that takes a longer time. By either pathway, the cell, deprived of growth-regulating influence from this gene, has unregulated growth and becomes a tumor. A different genetic locus also involving loss of a tumor-suppressor gene termed HRPT2 is involved in the pathogenesis of parathyroid carcinoma. *(From A Arnold: J Clin Endocrine Metab 77:1108, 1993. Copyright 1993, The Endocrine Society.)* ***B.*** Schematic illustration of the mechanism and consequences of gene rearrangement and overexpression of the PRAD 1 protooncogene (pericentromeric inversion of chromosome 11) in parathyroid adenomas. The excessive expression of PRAD1 (a cell cycle control protein, cyclin D1) by the highly active PTH gene promoter in the parathyroid cell contributes to excess cellular proliferation. *(From J Habener et al, in L DeGroot, JL Jameson [eds]: Endocrinology, 4th ed. Philadelphia, Saunders, 2001; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Mutations in the *MEN1* gene locus, encoding the protein MENIN, on chromosome 11q13 are responsible for causing MEN1; the normal allele of this gene fits the definition of a tumor-suppressor gene. Inheritance of one mutated allele in this hereditary syndrome, followed by loss of the other allele via somatic cell mutation, leads to monoclonal expansion and tumor development. Also, in ~15–20% of sporadic parathyroid adenomas, both alleles of the *MEN1* locus on chromosome 11 are somatically deleted, implying that the same defect responsible for MEN1 can also cause the sporadic disease **(Fig. 403-3A)**. Consistent with the Knudson hypothesis for two-step neoplasia in certain inherited cancer syndromes **(Chap. 67)**, the earlier onset of HPT in the hereditary syndromes reflects the need for only one mutational event to trigger the monoclonal outgrowth. In sporadic adenomas, typically occurring later in life, two different somatic events must occur before the *MEN1* gene is silenced.

Other presumptive anti-oncogenes involved in HPT include a still unidentified gene mapped to chromosome 1p seen in 40% of sporadic parathyroid adenomas and a gene mapped to chromosome Xp11 in patients with secondary HPT and renal failure, who progressed to “tertiary” HPT, now known to reflect monoclonal outgrowths within previously hyperplastic glands.

A more complex pattern, still incompletely resolved, arises with genetic defects and carcinoma of the parathyroids. This appears to be due to biallelic loss of a functioning copy of a gene, *HRPT2* (or CDC73), originally identified as the cause of the HPT-JT syndrome. Several inactivating mutations have been identified in *HRPT2* (located on chromosome 1q21-31), which encodes a 531-amino-acid protein called parafibromin. The responsible genetic mutations in *HRPT2* appear to be necessary, but not sufficient, for parathyroid cancer.

In general, the detection of additional genetic defects in these parathyroid tumor–related syndromes and the variations seen in phenotypic expression/penetrance indicate the multiplicity of the genetic factors responsible. Nonetheless, the ability to detect the presence of the major genetic contributors has greatly aided a more informed management of family members of patients identified in the hereditary syndromes such as MEN1, MEN2, and HPT-JT.

An important contribution from studies on the genetic origin of parathyroid carcinoma has been the realization that the mutations involve a different pathway than that involved with the benign gland enlargements. Unlike the pathogenesis of genetic alterations seen in colon cancer, where lesions evolve from benign adenomas to malignant disease by progressive genetic changes, the alterations commonly seen in most parathyroid cancers (*HRPT2* mutations) are infrequently seen in sporadic parathyroid adenomas.

Abnormalities at the *Rb* gene were the first to be noted in parathyroid cancer. The *Rb* gene, a tumor-suppressor gene located on chromosome 13q14, was initially associated with retinoblastoma but has since been implicated in other neoplasias, including parathyroid carcinoma. Early studies implicated allelic deletions of the *Rb* gene in many parathyroid carcinomas and decreased or absent expression of the Rb protein. However, because there are often large deletions in chromosome 13 that include many genes in addition to the *Rb* locus (with similar findings in some pituitary carcinomas), it remains possible that other tumor-suppressor genes on chromosome 13 may be playing a role in parathyroid carcinoma.

Study of the parathyroid cancers found in some patients with the HPT-JT syndrome has led to identification of a much larger role for mutations in the *HRPT2* gene in most parathyroid carcinomas, including those that arise sporadically, without apparent association with the HPT-JT syndrome. Mutations in the coding region have been identified in 75–80% of all parathyroid cancers analyzed, leading to the conclusion that, with addition of presumed mutations in the noncoding regions, this genetic defect may be seen in essentially all parathyroid carcinomas. Of special importance was the discovery that, in some sporadic parathyroid cancers, germ-line mutations have been found; this, in turn, has led to careful investigation of the families of these patients and a new clinical indication for genetic testing in this setting.

Hypercalcemia occurring in family members (who are also found to have the germ-line mutations) can lead to the finding, at parathyroid surgery, of premalignant parathyroid tumors.

Overall, it seems there are multiple factors in parathyroid cancer, in addition to the *HRPT2* and *Rb* gene, although the *HRPT2* gene mutation is the most invariant abnormality. *RET* encodes a tyrosine kinase type receptor; specific inherited germ-line mutations lead to a constitutive activation of the receptor, thereby explaining the autosomal dominant mode of transmission and the relatively early onset of neoplasia. In the MEN 2 syndrome, the *RET* protooncogene may be responsible for the earliest disorder detected, the polyclonal disorder (C cell hyperplasia, which then is transformed into a clonal outgrowth—a medullary carcinoma with the participation of other, still uncharacterized genetic defects).

In some parathyroid adenomas, activation of a protooncogene has been identified (**Fig. 403-3B**). A reciprocal translocation involving chromosome 11 has been identified that juxtaposes the *PTH* gene promoter upstream of a gene product termed PRAD1, encoding a cyclin D protein that plays a key role in normal cell division. This translocation plus other mechanisms that cause an equivalent overexpression of cyclin D1 are found in 20–40% of parathyroid adenomas.

Mouse models have confirmed the role of several of the major identified genetic defects in parathyroid disease and the MEN syndromes. Loss of the *MEN1* gene locus or overexpression of the PRAD1 protooncogene or the mutated *RET* protooncogene have been analyzed by genetic manipulation in mice, with the expected onset of parathyroid tumors or medullary carcinoma, respectively.

Signs and Symptoms

Many patients with HPT are asymptomatic. Manifestations of HPT involve primarily the kidneys and the skeletal system. Kidney involvement, due either to deposition of calcium in the renal parenchyma or to recurrent nephrolithiasis, was present in 60–70% of patients prior to 1970. With earlier detection, renal complications occur in <20% of patients in many large series. Renal stones are usually composed of either calcium oxalate or calcium phosphate. In occasional patients, repeated episodes of nephrolithiasis or the formation of large calculi may lead to urinary tract obstruction, infection, and loss of renal function. Nephrocalcinosis may also cause decreased renal function and phosphate retention.

The distinctive bone manifestation of HPT is *osteitis fibrosa cystica*, which occurred in 10–25% of patients in series reported 50 years ago. Histologically, the pathognomonic features are an increase in the giant multinucleated osteoclasts in scalloped areas on the surface of the bone (Howship's lacunae) and a replacement of the normal cellular and marrow elements by fibrous tissue. X-ray changes include resorption of the phalangeal tufts and replacement of the usually sharp cortical outline of the bone in the digits by an irregular outline (subperiosteal resorption). In recent years, osteitis fibrosa cystica is very rare in primary HPT, probably due to the earlier detection of the disease.

Dual-energy x-ray absorptiometry (DXA) of the spine provides reproducible quantitative estimates (within a few percent) of spinal bone density. Similarly, bone density in the extremities can be quantified by densitometry of the hip or of the distal radius at a site chosen to be primarily cortical. CT is a very sensitive technique for estimating spinal bone density, but reproducibility of standard CT is no better than 5%. Newer CT techniques (spiral, “extreme” CT) are more reproducible but are currently available in a limited number of medical centers. Cortical bone density is reduced while cancellous bone density, especially in the spine, is relatively preserved. In symptomatic patients, dysfunctions of the CNS, peripheral nerve and muscle, gastrointestinal tract, and joints also occur. It has been reported that severe neuropsychiatric manifestations may be reversed by parathyroidectomy. When present in symptomatic patients, neuromuscular manifestations may include proximal muscle weakness, easy fatigability, and atrophy of muscles and may be so striking as to suggest a primary neuromuscular disorder. The distinguishing feature is the complete regression of neuromuscular disease after surgical correction of the HPT.

Gastrointestinal manifestations are sometimes subtle and include vague abdominal complaints and disorders of the stomach and pancreas. Again, cause and effect are unclear. In MEN 1 patients with HPT, duodenal ulcer may be the result of associated pancreatic tumors that secrete excessive quantities of gastrin (Zollinger-Ellison syndrome). Pancreatitis has been reported in association with HPT, but the incidence and the mechanism are not established.

Much attention has been paid in recent years to the manifestations of and optimum management strategies for asymptomatic HPT. This is now the most prevalent form of the disease. *Asymptomatic primary hyperparathyroidism* is defined as biochemically confirmed HPT (elevated or inappropriately normal PTH levels despite hypercalcemia) with the absence of signs and symptoms typically associated with more severe HPT such as features of renal or bone disease.

Four conferences on the topic have been held in the United States over the past two decades, with the most recent in 2013. The published proceedings include discussion of more subtle manifestations of disease, its natural history (without parathyroidectomy), and guidelines both for indications for surgery and medical monitoring in non-operated patients.

Issues of concern include the potential for cardiovascular deterioration, the presence of subtle neuropsychiatric symptoms, and the longer-term status of skeletal integrity in patients not treated surgically. The current consensus is that medical monitoring rather than surgical correction of HPT may be justified in certain patients. The current recommendation is that patients who show mild disease, as defined by not meeting guidelines **(Table 403-2)**, can be safely followed under management guidelines **(Table 403-3)**. There is, however, growing uncertainty about subtle disease manifestations and whether surgery is therefore indicated in most patients. Among the issues is the evidence of eventual (>8 years) deterioration in bone mineral density after a decade of relative stability. There is concern that this late-onset deterioration in bone density in nonoperated patients could contribute significantly to the well-known age-dependent fracture risk (osteoporosis). Significant and sustained improvements in bone mineral density are seen after successful parathyroidectomy and some evidence for reduction in fractures.

TABLE 403-2

Guidelines for Surgery in Asymptomatic Primary Hyperparathyroidism[a]

Parameter	Guideline
Serum calcium (above normal)	>1 mg/dL
Renal	Creatinine clearance <60 mL/min 24-h urine for calcium >400 mg/d and increased stone risk by biochemical stone risk analysis Presence of nephrolithiasis or nephrocalcinosis by X-ray, ultrasound, or CT
Skeletal	BMD by DXA: T-score <-2.5 at lumbar spine, total hip, femoral neck, or distal 1/3 radius Vertebral fracture by X-ray, CT, MRI, or VFA
Age	<50

[a] JP Bilezikian et al: Guidelines for the management of asymptomatic primary hyperparathyroidism: Summary statement from the fourth international workshop. J Clin Endocrinol Metab 99(10):3561, 2014. Creatinine clearance calculated by Cockcroft-Gault equation or Modification of Diet in Renal Disease (MDRD) equation.

TABLE 403-3

Guidelines for Monitoring in Asymptomatic Primary Hyperparathyroidism

Parameter	Guideline
Serum calcium	Annually
Renal	eGFR, annually; serum creatinine, annually. If renal stones suspected, 24-h biochemical stone profile, renal imaging by X-ray, ultrasound, or CT
Serum creatinine	Annually
Skeletal	Every 1-2 y (3 sites), X-ray or VFA (Vertebral Fracture Assessment) of spine if clinically indicated (e.g., height loss, back pain)

Source: JP Bilezikian et al: J Clin Endocrinol Metab 99(10):3561, 2014.

Cardiovascular disease including left ventricular hypertrophy, cardiac functional defects, and endothelial dysfunction have been reported as reversible in European patients with more severe symptomatic disease after surgery, leading to numerous studies of these cardiovascular features in those with milder disease. There are reports of endothelial dysfunction in patients with mild asymptomatic HPT, but more observation is needed the expert panels concluded, especially whether there is reversibility with surgery.

A topic of considerable interest and some debate is assessment of neuropsychiatric status and health-related quality of life (QOL) status in hyperparathyroid patients both before surgery and in response to parathyroidectomy. Several observational studies suggest improvements in symptom score after surgery. Randomized studies of surgery versus observation, however, have yielded inconclusive results, especially regarding benefits of surgery. Many studies report that HPT is associated with increased neuropsychiatric symptoms, but it is not possible at present to determine which patients might improve after surgery.

DIAGNOSIS

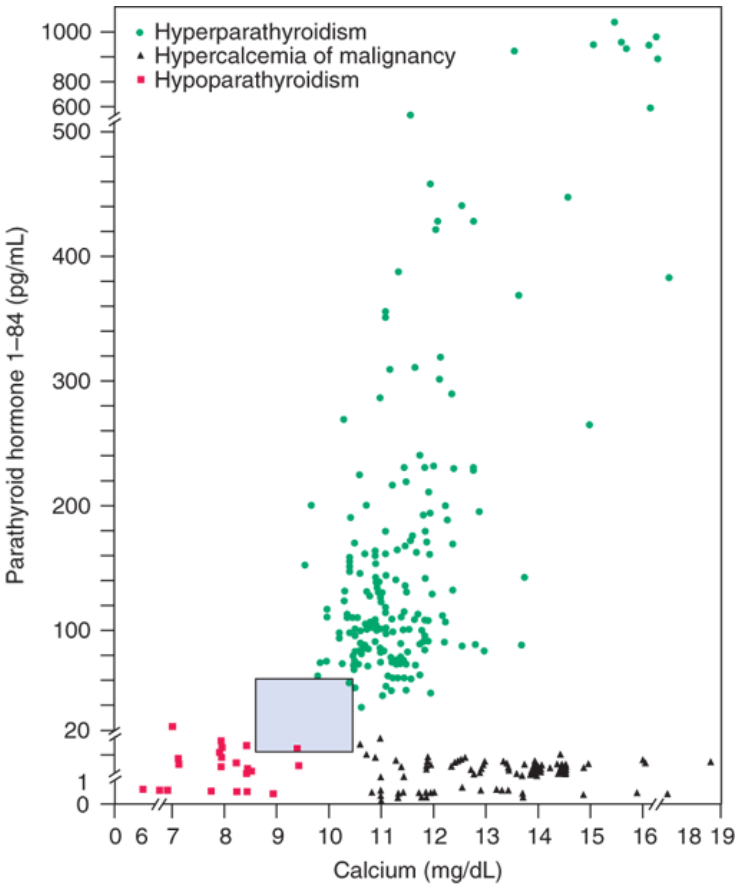
The diagnosis is typically made by detecting an elevated immunoreactive PTH level in a patient with asymptomatic hypercalcemia (see "Differential Diagnosis: Special Tests," below). Serum phosphate is usually low but may be normal, especially if renal failure has developed.

Several modifications in PTH assays have been introduced in efforts to improve their utility in light of information about metabolism of PTH (as discussed above). First-generation assays were based on displacement of radiolabeled PTH from antibodies that reacted with PTH (often also PTH fragments). Double-antibody or immunometric assays (one antibody that is usually directed against the carboxyl-terminal portion of intact PTH to capture the hormone and a second radio- or enzyme-labeled antibody that is usually directed against the amino-terminal portion of intact PTH) greatly improved the diagnostic discrimination of the tests by eliminating interference from circulating biologically inactive fragments, detected by the original first-generation assays. Double-antibody assays are now referred to as second-generation. Such PTH assays have in some centers and testing laboratories been replaced by third-generation assays after it was discovered that large PTH fragments, devoid of only the extreme amino-terminal portion of the PTH molecule, are also present in blood and are detected, incorrectly as intact PTH. These amino-terminally truncated PTH fragments were prevented from registering in the newer third-generation assays by use of a detection antibody directed against the extreme amino-terminal epitope. These assays may be useful for clinical research studies as in management of chronic renal disease, but the consensus is that either second- or third-generation assays are useful in the diagnosis of primary HPT and for the diagnosis of high-turnover bone disease in CKD.

Many tests based on renal responses to excess PTH (renal calcium and phosphate clearance; blood phosphate, chloride, magnesium; nephrogenous cyclic AMP) were used in earlier decades. These tests have low specificity for HPT and are therefore not cost-effective; they have been replaced by PTH immunometric assays combined with simultaneous blood calcium measurements **(Fig. 403-4)**.

FIGURE 403-4

Levels of immunoreactive parathyroid hormone (PTH) detected in patients with primary hyperparathyroidism, hypercalcemia of malignancy, and hypoparathyroidism. Boxed area represents the upper and normal limits of blood calcium and/or immunoreactive PTH. *(From SR Nussbaum, JT Potts, Jr, in L DeGroot, JL Jameson [eds]: Endocrinology, 4th ed. Philadelphia, Saunders, 2001; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Hyperparathyroidism

Surgical excision of the abnormal parathyroid tissue is the definitive therapy for this disease. As noted above, medical surveillance without operation for patients with mild, asymptomatic disease is, however, still preferred by some physicians and patients, particularly when the patients are more elderly. Evidence favoring surgery, if medically feasible, is growing because of concerns about skeletal, cardiovascular, and neuro-psychiatric disease, even in mild HPT.

Two surgical approaches are generally practiced. The conventional parathyroidectomy procedure was neck exploration with general anesthesia; this procedure is being replaced in many centers, whenever feasible, by an outpatient procedure with local anesthesia, termed *minimally invasive parathyroidectomy*. Parathyroid exploration is challenging and should be undertaken by an experienced surgeon. Certain features help in predicting the pathology (e.g., multiple abnormal glands in familial cases). However, some critical decisions regarding management can be made only during the operation.

With conventional surgery, one approach is still based on the view that typically only one gland (the adenoma) is abnormal. If an enlarged gland is found, a normal gland should be sought. In this view, if a biopsy of a normal-sized second gland confirms its histologic (and presumed functional) normality, no further exploration, biopsy, or excision is needed. At the other extreme is the minority viewpoint that all four glands be sought and that most of the total parathyroid tissue mass be removed. The concern with the former approach is that the recurrence rate of HPT may be high if a second abnormal gland is missed; the latter approach could involve unnecessary surgery and an unacceptable rate of hypoparathyroidism. When normal glands are found in association with one enlarged gland, excision of the single adenoma usually leads to cure or at least years free of symptoms. Long-term follow-up studies to establish true rates of recurrence are limited.

Recently, there has been growing experience with new surgical strategies that feature a minimally invasive approach guided by improved preoperative localization and intraoperative monitoring by PTH assays. Preoperative ^{99m}Tc sestamibi scans with single-photon emission CT (SPECT) are used to predict the location of an abnormal gland and intraoperative sampling of PTH before and at 5-minute intervals after removal of a suspected adenoma to confirm a rapid fall (>50%) to normal levels of PTH. In several centers, a combination of preoperative sestamibi imaging, cervical block anesthesia, minimal surgical incision, and intraoperative PTH measurements has allowed successful outpatient surgical management with a clear-cut cost benefit compared

to general anesthesia and more extensive neck surgery. The use of these minimally invasive approaches requires clinical judgment to select patients unlikely to have multiple gland disease (e.g., MEN or secondary HPT). The growing acceptance of the technique and its relative ease for the patient has lowered the threshold for surgery.

Severe hypercalcemia may provide a preoperative clue to the presence of parathyroid carcinoma. In such cases, when neck exploration is undertaken, the tissue should be widely excised; care is taken to avoid rupture of the capsule to prevent local seeding of tumor cells.

Multiple-gland hyperplasia, as predicted in familial cases, poses more difficult questions of surgical management. Once a diagnosis of hyperplasia is established, all the glands must be identified. Two schemes have been proposed for surgical management. One is to totally remove three glands with partial excision of the fourth gland; care is taken to leave a good blood supply for the remaining gland. Other surgeons advocate total parathyroidectomy with immediate transplantation of a portion of a removed, minced parathyroid gland into the muscles of the forearm, with the view that surgical excision is easier from the ectopic site in the arm if there is recurrent hyperfunction.

In a minority of cases, if no abnormal parathyroid glands are found in the neck, the issue of further exploration must be decided. There are documented cases of five or six parathyroid glands and of unusual locations for adenomas such as in the mediastinum.

When a second parathyroid exploration is indicated, the minimally invasive techniques for preoperative localization such as ultrasound, CT scan, and isotope scanning are combined with venous sampling and/or selective digital arteriography in one of the centers specializing in these procedures. Intraoperative monitoring of PTH levels by rapid PTH immunoassays may be useful in guiding the surgery. At one center, long-term cures have been achieved with selective embolization or injection of large amounts of contrast material into the end-arterial circulation feeding the parathyroid tumor.

A decline in serum calcium occurs within 24 h after successful surgery; usually blood calcium falls to low-normal values for 3–5 days until the remaining parathyroid tissue resumes full hormone secretion. Acute postoperative hypocalcemia is likely only if severe bone mineral deficits are present or if injury to all the normal parathyroid glands occurs during surgery. In general, there are few problems encountered in patients with uncomplicated disease such as a single adenoma (the clear majority), who do not have symptomatic bone disease nor a large deficit in bone mineral, who are vitamin D and magnesium sufficient, and who have good renal and gastrointestinal function. The extent of postoperative hypocalcemia varies with the surgical approach. If all glands are biopsied, hypocalcemia may be transiently symptomatic and more prolonged. Hypocalcemia is more likely to be symptomatic after second parathyroid explorations, particularly when normal parathyroid tissue was removed at the initial operation and when the manipulation and/or biopsy of the remaining normal glands are more extensive in the search for the missing adenoma.

Patients with HPT have efficient intestinal calcium absorption due to the increased levels of $1,25(OH)_2D$ stimulated by PTH excess. Once hypocalcemia signifies successful surgery, patients can be put on a high-calcium intake or be given oral calcium supplements. Despite mild hypocalcemia, most patients do not require parenteral therapy. If the serum calcium falls to <2 mmol/L (8 mg/dL), *and if the phosphate level rises simultaneously*, the possibility that surgery has caused hypoparathyroidism must be considered. With unexpected hypocalcemia, coexistent hypomagnesemia should be considered, as it interferes with PTH secretion and causes functional hypoparathyroidism **(Chap. 402)**.

Signs of hypocalcemia include symptoms such as muscle twitching, a general sense of anxiety, and positive Chvostek's and Trousseau's signs coupled with serum calcium consistently <2 mmol/L (8 mg/dL). Parenteral calcium replacement at a low level should be instituted when hypocalcemia is symptomatic. The rate and duration of IV therapy are determined by the severity of the symptoms and the response of the serum calcium to treatment. An infusion of 0.5–2 mg/kg per hour or 30–100 mL/h of a 1-mg/mL solution usually suffices to relieve symptoms. Usually, parenteral therapy is required for only a few days. If symptoms worsen or if parenteral calcium is needed for >2–3 days, therapy with a vitamin D analogue and/or oral calcium (2–4 g/d) should be started (see below). It is cost-effective to use calcitriol (doses of 0.5–1 μg/d) because of the rapidity of onset of effect and prompt cessation of action when stopped, in comparison to other forms of vitamin D. A rise in blood calcium after several months of vitamin D replacement may indicate restoration of parathyroid function to normal. It is also appropriate to monitor serum PTH serially to estimate gland function in such patients.

If magnesium deficiency was present, it can complicate the postoperative course since magnesium deficiency impairs the secretion of PTH. Hypomagnesemia should be corrected whenever detected. Magnesium replacement can be effective orally (e.g., $MgCl_2$, $MgOH_2$), but parenteral repletion is usual to ensure postoperative recovery, if magnesium deficiency is suspected due to low blood magnesium levels. Because the depressant effect of magnesium on central and peripheral nerve functions does not occur at levels <2 mmol/L (normal range 0.8–1.2 mmol/L), parenteral replacement can be given rapidly. A cumulative dose as great as 0.5–1 mmol/kg of body weight can be administered if severe hypomagnesemia is present; often, however, total doses of 20–40 mmol are sufficient.

MEDICAL MANAGEMENT

The guidelines for recommending surgical intervention, if feasible (Table 403-2), as well as for monitoring patients with asymptomatic HPT who elect not to undergo parathyroidectomy (Table 403-3), reflect the changes over time since the first conference on the topic in 1990. Medical monitoring rather than corrective surgery is still acceptable, but it is clear that surgical intervention is the more frequently recommended option for the reasons noted above. Tightened guidelines favoring surgery include lowering the recommended level of serum calcium elevation, more careful attention to skeletal integrity through reference to peak skeletal mass at baseline (T scores) rather than age-adjusted bone density (*Z* scores), as well as the presence of any fragility fracture. The other changes noted in the two guidelines (Tables 403-2 and 403-3) reflect accumulated experience and practical consideration, such as a

difficulty in quantity of urine collections. Despite the usefulness of the guidelines, the importance of individual patient and physician judgment and preference are clear in all recommendations.

When surgery is not selected, or not medically feasible, there is interest in the potential value of specific medical therapies. There is no long-term experience regarding specific clinical outcomes such as fracture prevention, but it has been established that bisphosphonates increase bone mineral density significantly without changing serum calcium (as does estrogen, but the latter is not favored because of reported adverse effects in other organ systems). Calcimimetics that lower PTH secretion lower calcium but do not affect bone mass density (BMD).

OTHER PARATHYROID-RELATED CAUSES OF HYPERCALCEMIA

Lithium Therapy

Lithium, used in the management of bipolar depression and other psychiatric disorders, causes hypercalcemia in ~10% of treated patients. The hypercalcemia is dependent on continued lithium treatment, remitting and recurring when lithium is stopped and restarted. The parathyroid adenomas reported in some hypercalcemic patients with lithium therapy may reflect the presence of an independently occurring parathyroid tumor; a permanent effect of lithium on parathyroid gland growth need not be implicated as most patients have complete reversal of hypercalcemia when lithium is stopped. However, long-standing stimulation of parathyroid cell replication by lithium may predispose to development of adenomas (as is documented in secondary HPT and renal failure).

At the levels achieved in blood in treated patients, lithium can be shown in vitro to shift the PTH secretion curve to the right in response to calcium; i.e., higher calcium levels are required to lower PTH secretion, probably acting at the calcium sensor (see below). This effect can cause elevated PTH levels and consequent hypercalcemia in otherwise normal individuals. Fortunately, there are usually alternative medications for the underlying psychiatric illness. Parathyroid surgery should not be recommended unless hypercalcemia and elevated PTH levels persist after lithium is discontinued.

GENETIC DISORDERS CAUSING HYPERPARATHYROID-LIKE SYNDROMES

Familial Hypocalciuric Hypercalcemia

FHH (also called *familial benign hypercalcemia*) is inherited as an autosomal dominant trait. Affected individuals are discovered because of asymptomatic hypercalcemia. Most cases of FHH (FHH1) are caused by an inactivating mutation in a single allele of the CaSR (see below), leading to inappropriately normal or even increased secretion of PTH, whereas another hypercalcemic disorder, namely the exceedingly rare Jansen's disease, is caused by a constitutively active PTH/PTHrP receptor in target tissues. Neither FHH1 nor Jansen's disease, however, are growth disorders of the parathyroids. Other forms of FHH are caused either by heterozygous mutations in *GNA11* (encoding Gα11), one of the signaling proteins downstream of the CaSR (FHH2), or by mutations in *AP1A1* (FHH3).

The pathophysiology of FHH1 is now understood. The primary defect is abnormal sensing of the blood calcium by the parathyroid gland and renal tubule, causing inappropriate secretion of PTH and excessive reabsorption of calcium in the distal renal tubules. The CaSR is a member of the third family of GPCRs (type C or type III). The receptor responds to increased ECF calcium concentration by suppressing PTH secretion through second-messenger signaling involving the G proteins Gα11 and Gαq, thereby providing negative-feedback regulation of PTH secretion. Many different inactivating CaSR mutations have been identified in patients with FHH1. These mutations lower the capacity of the sensor to bind calcium, and the mutant receptors function as though blood calcium levels were low; excessive secretion of PTH occurs from an otherwise normal gland. Approximately two-thirds of patients with FHH have mutations within the protein-coding region of the CaSR gene. The remaining one-third of kindreds may have mutations in the promoter of the CaSR gene or are caused by mutations in other genes.

Even before elucidation of the pathophysiology of FHH, abundant clinical evidence served to separate the disorder from primary HPT; these clinical features are still useful in differential diagnosis. Patients with primary HPT have <99% renal calcium reabsorption, whereas most patients with FHH have >99% reabsorption. The hypercalcemia in FHH is often detectable in affected members of the kindreds in the first decade of life, whereas hypercalcemia rarely occurs in patients with primary HPT or the MEN syndromes who are aged <10 years. PTH may be elevated in the different forms of FHH, but the values are usually normal or lower for the same degree of calcium elevation than is observed in patients with primary HPT. Parathyroid surgery performed in a few patients with FHH before the nature of the syndrome was understood led to permanent hypoparathyroidism; nevertheless, hypocalciuria persisted, establishing that hypocalciuria is not PTH-dependent (now known to be due to the abnormal CaSR in the kidney).

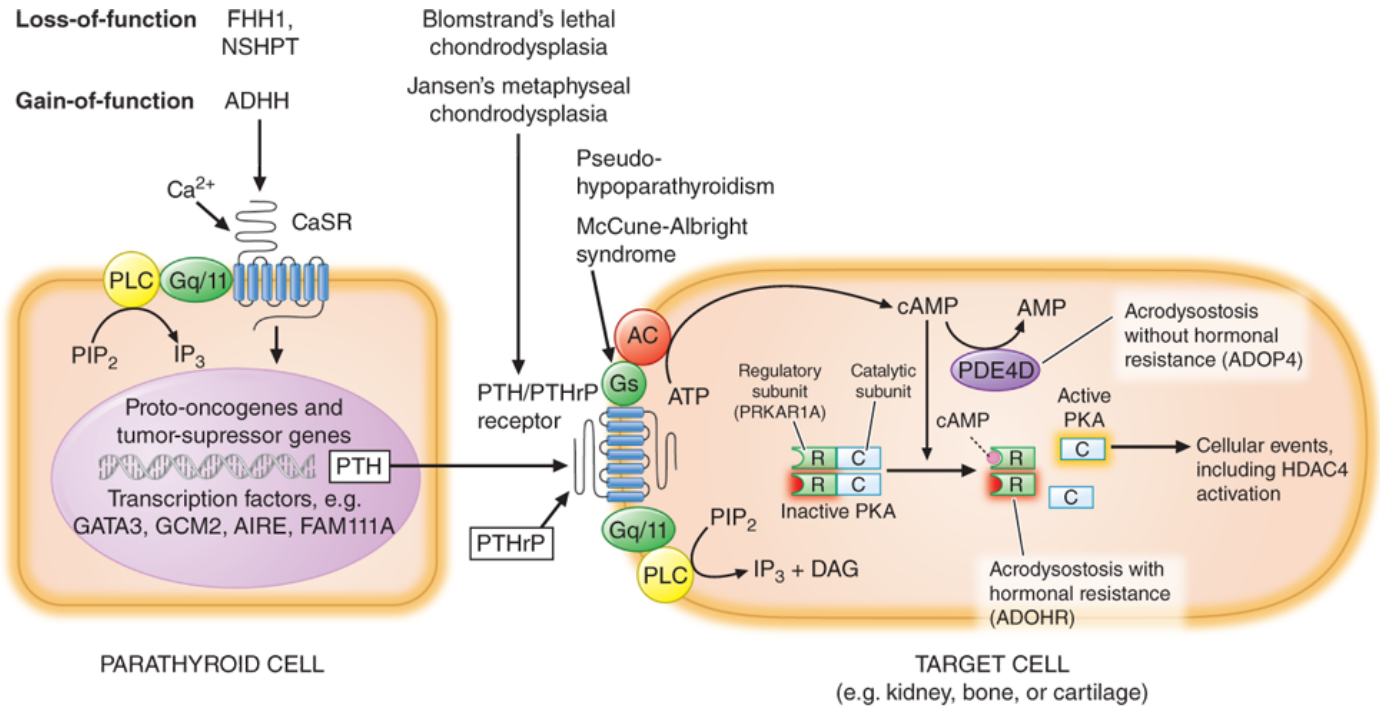
Few clinical signs or symptoms are present in patients with FHH, while other endocrine abnormalities are not. Most patients are detected as a result of family screening after hypercalcemia is detected in a proband. In those patients inadvertently operated upon for primary HPT, the parathyroids appeared normal or moderately hyperplastic. Parathyroid surgery is not appropriate, nor, in view of the lack of symptoms, does medical treatment seem needed to lower the calcium. One striking exception to the rule against parathyroid surgery in this syndrome is the occurrence, usually in consanguineous marriages (due to the rarity of the gene mutation), of a homozygous or compound heterozygote state, resulting in severe impairment of CaSR function. In this condition, neonatal severe hypercalcemia, total parathyroidectomy is mandatory, but calcimetics have been used as a temporary measure. Rare but well-documented cases of acquired hypocalciuric hypercalcemia are reported due to antibodies against the CaSR. They appear to be a complication of an underlying autoimmune disorder and respond to therapies directed against the underlying disorder.

Jansen's Disease

Activating mutations in the PTH/PTHrP receptor (PTH1R) have been identified as the cause of this rare autosomal dominant syndrome. Because the mutations lead to constitutive activation of receptor function, one abnormal copy of the mutant receptor is sufficient to cause the disease, thereby accounting for its dominant mode of transmission. The disorder leads to short-limbed dwarfism due to abnormal regulation of chondrocyte maturation in the growth plates of the bone that are formed through the endochondral process. In adult life, there are numerous abnormalities in bone, including multiple cystic resorptive areas resembling those seen in severe HPT. Hypercalcemia and hypophosphatemia with undetectable or low PTH levels are typically observed. The pathogenesis of the growth plate abnormalities in Jansen's disease has been confirmed by transgenic experiments in which targeted expression of the mutant PTH/PTHrP receptor to the proliferating chondrocyte layer of growth plate emulated several features of the human disorder. Some of these genetic mutations in the parathyroid gland or PTH target cells that affect Ca^{2+} metabolism are illustrated in **Figure 403-5**.

FIGURE 403-5

Illustration of some genetic mutations that alter calcium metabolism by effects on the parathyroid cell or target cells of parathyroid hormone (PTH) action. Alterations in PTH production by the parathyroid cell can be caused by changes in the response to extracellular fluid calcium (Ca^{2+}) that are detected by the calcium-sensing receptor (CaSR). Furthermore, PTH (or PTHrP) can show altered efficacy in target cells such as in proximal tubular cells, by altered function of its receptor (PTH/PTHrP receptor) or the signal transduction proteins, G proteins such as $G_s\alpha$ that is linked to adenylate cyclase (AC), the enzyme responsible for producing cyclic AMP (cAMP) (also illustrated are Gαq/Gα11, which activate an alternate pathway of receptor signal transmission involving the generation of inositol triphosphate [IP_3] or diacylglycerol [DAG]). Heterozygous loss-of-function mutations in the CaSR cause familial hypocalciuric hypercalcemia (FHH) and homozygous mutations (both alleles mutated) and severe neonatal hyperparathyroidism (NSHPT); heterozygous gain-of-functions causes autosomal dominant hypercalciuric hypocalcemia (ADHH). Other defects in parathyroid cell function that occur at the level of gene regulation (oncogenes or tumor suppressor genes) or transcription factors are discussed in the text. Blomstrand's lethal chondrodysplasia is due to homozygous or compound heterozygous loss-of-function mutations in the PTH/PTHrP receptor, a neonatally lethal disorder, while pseudohypoparathyroidism involves inactivation at the level of the G proteins, specifically mutations that eliminate or reduce $G_s\alpha$ activity in the kidney (see text for details). Acrodysostosis can occur with (ADOHR; mutant regulatory subunit of PKA) or without hormonal resistance (ADOP4; mutant PDE4D). Jansen's metaphyseal chondrodysplasia and McCune-Albright syndrome represent gain-of-function mutations in the PTH/PTHrP receptor and $G_s\alpha$ protein, respectively.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MALIGNANCY-RELATED HYPERCALCEMIA

Clinical Syndromes and Mechanisms of Hypercalcemia

Hypercalcemia due to malignancy is common (occurring in as many as 20% of cancer patients, especially with certain types of tumor such as lung carcinoma), often severe and difficult to manage, and, on rare occasions, difficult to distinguish from primary HPT. Although malignancy is often clinically obvious or readily detectable by medical history, hypercalcemia can occasionally be due to an occult tumor. Previously, hypercalcemia associated with

malignancy was thought to be due to local invasion and destruction of bone by tumor cells; many cases are now known to result from the elaboration by the malignant cells of humoral mediators of hypercalcemia. PTHrP is the responsible humoral agent in most solid tumors that cause hypercalcemia.

The histologic character of the tumor is more important than the extent of skeletal metastases in predicting hypercalcemia. Small cell carcinoma (oat cell) and adenocarcinoma of the lung, though the most common lung tumors associated with skeletal metastases, rarely cause hypercalcemia. By contrast, many patients with squamous cell carcinoma of the lung develop hypercalcemia. Histologic studies of bone in patients with squamous cell or epidermoid carcinoma of the lung, in sites invaded by tumor as well as areas remote from tumor invasion, reveal increased bone resorption.

Two main mechanisms of hypercalcemia are operative in cancer hypercalcemia. Many solid tumors associated with hypercalcemia, particularly squamous cell and renal tumors, produce and secrete PTHrP that causes increased bone resorption and mediate the hypercalcemia through systemic actions on the skeleton. Alternatively, direct bone marrow invasion occurs with hematologic malignancies such as leukemia, lymphoma, and multiple myeloma. Lymphokines and cytokines (including PTHrP) produced by cells involved in the marrow response to the tumors promote resorption of bone through local destruction. Several hormones, hormone analogues, cytokines, and growth factors have been implicated as the result of clinical assays, in vitro tests, or chemical isolation. The etiologic factor produced by activated normal lymphocytes and by myeloma and lymphoma cells, originally termed *osteoclast activation factor*, now appears to represent the biologic action of several different cytokines, probably interleukin 1 and lymphotoxin or tumor necrosis factor (TNF). In some lymphomas, there is a third mechanism, caused by an increased blood level of $1,25(OH)_2D$, produced by the abnormal lymphocytes.

In the more common mechanism, usually termed *humoral hypercalcemia of malignancy*, solid tumors (cancers of the lung and kidney, in particular), in which bone metastases are absent, minimal, or not detectable clinically, secrete PTHrP measurable by immunoassay. Secretion by the tumors of the PTH-like factor, PTHrP, activates the PTH1R, resulting in a pathophysiology closely resembling HPT, but with normal or suppressed PTH levels. The clinical picture resembles primary HPT (hypophosphatemia accompanies hypercalcemia), and elimination or regression of the primary tumor leads to disappearance of the hypercalcemia.

As in HPT, patients with the humoral hypercalcemia of malignancy have elevated urinary nephrogenous cyclic AMP excretion, hypophosphatemia, and increased urinary phosphate clearance. However, in humoral hypercalcemia of malignancy, immunoreactive PTH is undetectable or suppressed, making the differential diagnosis easier. Other features of the disorder differ from those of true HPT. Although the biologic actions of PTH and PTHrP are exerted through the same receptor, subtle differences in receptor activation by the two ligands must account for some of the discordance in pathophysiology, when an excess of one or the other peptide occurs. Other cytokines elaborated by the malignancy may contribute to the variations from HPT in these patients as well. Patients with humoral hypercalcemia of malignancy may have low to normal levels of $1,25(OH)_2D$ instead of elevated levels as in true HPT. In some patients with the humoral hypercalcemia of malignancy, osteoclastic resorption is unaccompanied by an osteoblastic or bone-forming response, implying inhibition of the normal coupling of bone formation and resorption.

Several different assays (single- or double-antibody, different epitopes) have been developed to detect PTHrP. Most data indicate that circulating PTHrP levels are undetectable (or low) in normal individuals except perhaps in pregnancy (high in human milk) and elevated in most cancer patients with the humoral syndrome. The etiologic mechanisms in cancer hypercalcemia may be multiple in the same patient. For example, in breast carcinoma (metastatic to bone) and in a distinctive type of T cell lymphoma/leukemia initiated by human T cell lymphotropic virus I, hypercalcemia is caused by direct local lysis of bone as well as by a humoral mechanism involving excess production of PTHrP. HPT has been reported to coexist with the humoral cancer syndrome and, rarely, ectopic HPT due to tumor elaboration of true PTH is reported.

Diagnostic Issues

Levels of PTH measured by the double-antibody technique are undetectable or extremely low in tumor hypercalcemia, as would be expected with the mediation of the hypercalcemia by a factor other than PTH (the hypercalcemia suppresses the normal parathyroid glands). In a patient with minimal symptoms referred for hypercalcemia, low or undetectable PTH levels would focus attention on a possible occult malignancy (except for very rare cases of ectopic HPT).

Ordinarily, the diagnosis of cancer hypercalcemia is not difficult because tumor symptoms are prominent when hypercalcemia is detected. Indeed, hypercalcemia may be noted incidentally during the workup of a patient with known or suspected malignancy. Clinical suspicion that malignancy is the cause of the hypercalcemia is heightened when there are other signs or symptoms of a paraneoplastic process such as weight loss, fatigue, muscle weakness, or unexplained skin rash, or when symptoms specific for a particular tumor are present. Squamous cell tumors are most frequently associated with hypercalcemia, particularly tumors of the lung, kidney, head and neck, and urogenital tract. Radiologic examinations can focus on these areas when clinical evidence is unclear. Bone scans with technetium-labeled bisphosphonate are useful for detection of osteolytic metastases; the sensitivity is high, but specificity is low; results must be confirmed by conventional x-rays to be certain that areas of increased uptake are due to osteolytic metastases per se. Bone marrow biopsies are helpful in patients with anemia or abnormal peripheral blood smears.

TREATMENT

TREATMENT

Malignancy-Related Hypercalcemia

Treatment of the hypercalcemia of malignancy is first directed to control of tumor; reduction of tumor mass usually corrects hypercalcemia. If a patient has severe hypercalcemia yet has a good chance for effective tumor therapy, treatment of the hypercalcemia should be vigorous while awaiting the results of definitive therapy (see general approach to hypercalcemic states below). If hypercalcemia occurs in the late stages of a tumor that is resistant to antitumor therapy, the treatment of the hypercalcemia should be judicious as high calcium levels can have a mild sedating effect. Standard therapies for hypercalcemia (discussed below) are applicable to patients with malignancy.

VITAMIN D–RELATED HYPERCALCEMIA

Vitamin D–mediated hypercalcemia can be due to excessive ingestion vitamin D analogs or abnormal metabolism of the vitamin. Abnormal metabolism of the vitamin is usually acquired in association with a widespread granulomatous disorder. Vitamin D metabolism is carefully regulated, particularly the activity of renal 1α-hydroxylase, the enzyme responsible for the production of $1,25(OH)_2D$ **(Chap. 402)**. The regulation of 1α-hydroxylase and the normal feedback suppression by $1,25(OH)_2D$ seem to work less well in infants than in adults and to operate poorly, if at all, in sites other than the renal tubule; these phenomena may explain the occurrence of hypercalcemia secondary to excessive $1,25(OH)_2D_3$ production in infants with Williams' syndrome (see below) and in adults with sarcoidosis or lymphoma.

Vitamin D Intoxication

Chronic ingestion of 40–100 times the normal physiologic requirement of vitamin D (amounts >40,000–100,000 U/d) is usually required to produce significant hypercalcemia in otherwise healthy individuals. The stated upper limit of safe dietary intake is 2000 U/d (50 μg/d) in adults because of concerns about potential toxic effects of cumulative supraphysiologic doses. These recommendations are now regarded as too restrictive, since some estimates are that in elderly individuals in northern latitudes, 2000 U/d or more may be necessary to avoid vitamin D insufficiency.

Hypercalcemia in vitamin D intoxication is due to an excessive biologic action of the vitamin, perhaps the consequence of increased levels of 25(OH)D rather than merely increased levels of the active metabolite $1,25(OH)_2D$ (the latter may not be elevated in vitamin D intoxication). 25(OH)D has definite, if low, biologic activity in the intestine and bone. The production of 25(OH)D is less tightly regulated than is the production of $1,25(OH)_2D$. Hence concentrations of 25(OH)D are elevated several-fold in patients with excess vitamin D intake.

The diagnosis is substantiated by documenting elevated levels of 25(OH)D >100 ng/mL. Hypercalcemia is usually controlled by restriction of dietary calcium intake and appropriate attention to hydration. These measures, plus discontinuation of vitamin D, usually lead to resolution of hypercalcemia. However, vitamin D stores in fat may be substantial, and vitamin D intoxication may persist for weeks after vitamin D ingestion is terminated. Such patients are responsive to glucocorticoids, which in doses of 100 mg/d of hydrocortisone or its equivalent usually return serum calcium levels to normal over several days; severe intoxication may require intensive therapy.

Sarcoidosis and Other Granulomatous Diseases

In patients with sarcoidosis and other granulomatous diseases, such as tuberculosis and fungal infections, excess $1,25(OH)_2D$ is synthesized in macrophages or other cells in the granulomas. Indeed, increased $1,25(OH)_2D$ levels have been reported in anephric patients with sarcoidosis and hypercalcemia. Macrophages obtained from granulomatous tissue convert 25(OH)D to $1,25(OH)_2D$ at an increased rate. There is a positive correlation in patients with sarcoidosis between 25(OH)D levels (reflecting vitamin D intake) and the circulating concentrations of $1,25(OH)_2D$, whereas normally there is no increase in $1,25(OH)_2D$ with increasing 25(OH)D levels due to multiple feedback controls on renal 1α-hydroxylase **(Chap. 402)**. The usual regulation of active metabolite production by calcium and phosphate or by PTH does not operate in these patients. Clearance of $1,25(OH)_2D$ from blood may be decreased in sarcoidosis as well. PTH levels are usually low and $1,25(OH)_2D$ levels are elevated, but primary HPT and sarcoidosis may coexist in some patients.

Management of the hypercalcemia can often be accomplished by avoiding excessive sunlight exposure and limiting vitamin D and calcium intake. Presumably, however, the abnormal sensitivity to vitamin D and abnormal regulation of $1,25(OH)_2D$ synthesis will persist as long as the disease is active. Alternatively, glucocorticoids in the equivalent of 100 mg/d of hydrocortisone or equivalent doses of glucorticoids may help control hypercalcemia. Glucocorticoids appear to act by blocking excessive production of $1,25(OH)_2D$, as well as the response to it in target organs.

Idiopathic Hypercalcemia of Infancy

This rare disorder, usually referred to as *Williams' syndrome*, is an autosomal dominant disorder characterized by multiple congenital development defects, including supravalvular aortic stenosis, mental retardation, and an elfin facies, in association with hypercalcemia due to abnormal sensitivity to vitamin D. The hypercalcemia associated with the syndrome was first recognized in England after fortification of milk with vitamin D. The cardiac and

developmental abnormalities were independently described, but the connections between these defects and hypercalcemia were not described until later. Levels of $1,25(OH)_2D$ can be elevated, ranging from 46 to 120 nmol/L (150–500 pg/mL). The mechanism of the abnormal sensitivity to vitamin D and of the increased circulating levels of $1,25(OH)_2D$ is still unclear. Studies suggest that genetic mutations involving microdeletions at the elastin locus and perhaps other genes on chromosome 7 may play a role in the pathogenesis. Other causes of hypercalcemia in infants and young children are 24-hydroxylase deficiency that impairs metabolism of $1,25(OH)_2D$, or mutations involving the sodium-dependent phosphate transporters (NPT2a or NPT2c).

HYPERCALCEMIA ASSOCIATED WITH HIGH BONE TURNOVER

Hyperthyroidism

As many as 20% of hyperthyroid patients have high-normal or mildly elevated serum calcium concentrations; hypercalciuria is even more common. The hypercalcemia is due to increased bone turnover, with bone resorption exceeding bone formation. Severe calcium elevations are not typical, and the presence of such suggests a concomitant disease such as HPT. Usually, the diagnosis is obvious, but signs of hyperthyroidism may occasionally be occult, particularly in the elderly **(Chap. 375)**. Hypercalcemia is managed by treatment of the hyperthyroidism. Reports that thyroid-stimulating hormone (TSH) itself normally has a bone-protective effect suggest that suppressed TSH levels also play a role in hypercalcemia.

Immobilization

Immobilization is a rare cause of hypercalcemia in adults in the absence of an associated disease but may cause hypercalcemia in children and adolescents, particularly after spinal cord injury and paraplegia or quadriplegia. With resumption of ambulation, the hypercalcemia in children usually returns to normal.

The mechanism appears to involve a disproportion between bone formation and bone resorption; the former decreased and the latter increased. Hypercalciuria and increased mobilization of skeletal calcium can develop in normal volunteers subjected to extensive bed rest, although hypercalcemia is unusual. Immobilization of an adult with a disease associated with high bone turnover, however, such as Paget's disease, may cause hypercalcemia.

Thiazides

Administration of benzothiadiazines (thiazides) can cause hypercalcemia in patients with high rates of bone turnover. Traditionally, thiazides are associated with aggravation of hypercalcemia in primary HPT, but this effect can be seen in other high-bone-turnover states as well. The mechanism of thiazide action is complex. Chronic thiazide administration leads to reduction in urinary calcium; the hypocalciuric effect appears to reflect the enhancement of proximal tubular resorption of sodium and calcium in response to sodium depletion. Some of this renal effect is due to augmentation of PTH action and is more pronounced in individuals with intact PTH secretion. However, thiazides cause hypocalciuria in hypoparathyroid patients on high-dose vitamin D and oral calcium replacement if sodium intake is restricted. This finding is the rationale for the use of thiazides as an adjunct to therapy in hypoparathyroid patients, as discussed below. Thiazide administration to normal individuals causes a transient increase in blood calcium (usually within the high-normal range) that reverts to preexisting levels after a week or more of continued administration. If hormonal function and calcium and bone metabolism are normal, homeostatic controls are reset to counteract the calcium-elevating effect of the thiazides. In the presence of HPT or increased bone turnover from another cause, homeostatic mechanisms are ineffective. The abnormal effects of the thiazide on calcium metabolism disappear within days of cessation of the drug.

Vitamin A Intoxication

Vitamin A intoxication is a rare cause of hypercalcemia and is most commonly a side effect of dietary faddism **(Chap. 326)**. Calcium levels can be elevated into the 3–3.5-mmol/L (12–14 mg/dL) range after the ingestion of 50,000–100,000 units of vitamin A daily (10–20 times the minimum daily requirement). Typical features of severe hypercalcemia include fatigue, anorexia, and, in some, severe muscle and bone pain. Excess vitamin A intake is presumed to increase bone resorption.

The diagnosis can be established by history and by measurement of vitamin A levels in serum. Occasionally, skeletal x-rays reveal periosteal calcifications, particularly in the hands. Withdrawal of the vitamin is usually associated with prompt disappearance of the hypercalcemia and reversal of the skeletal changes. As in vitamin D intoxication, administration of 100 mg/d hydrocortisone or its equivalent leads to a rapid return of the serum calcium to normal.

HYPERCALCEMIA ASSOCIATED WITH RENAL FAILURE

Severe Secondary HPT

The pathogenesis of secondary HPT in CKD is incompletely understood. Resistance to the normal level of PTH is a major factor contributing to the development of hypocalcemia, which, in turn, is a stimulus to parathyroid gland enlargement. However, recent findings have indicated that an increase of FGF23 production by osteocytes (and possibly osteoblasts) in bone occurs well before an elevation in PTH is detected. FGF23 is a potent inhibitor of the

renal 1-alpha hydroxylase and the FGF23-dependent reduction in $1,25(OH)_2$ vitamin D seems to be an important stimulus for the development of secondary HPT.

Secondary HPT occurs not only in patients with renal failure but also in those with osteomalacia due to multiple causes (Chap. 402), including deficiency of vitamin D action and PHP (deficient response to PTH downstream of PTHR1). For both disorders, hypocalcemia seems to be the common denominator in initiating the development of secondary HPT. Primary (1°) and secondary (2°) HPT can be distinguished conceptually by the autonomous growth of the parathyroid glands in primary HPT (presumably irreversible) and the adaptive response of the parathyroids in secondary HPT (typically reversible). In fact, reversal over weeks from an abnormal pattern of secretion, presumably accompanied by involution of parathyroid gland mass to normal, occurs in patients with osteomalacia who have been treated effectively with calcium and vitamin D. However, it is now recognized that a true clonal outgrowth (irreversible) can arise in long-standing, inadequately treated CKD (e.g., tertiary [3°] HPT; see below).

Patients with secondary HPT may develop bone pain, ectopic calcification, and pruritus. The bone disease seen in patients with secondary HPT and CKD is termed *renal osteodystrophy* and affects primarily bone turnover. However, osteomalacia is frequently encountered as well and may be related to the circulating levels of FGF23.

Two other skeletal disorders have been frequently associated in the past with CKD patients treated by long-term dialysis, who received aluminum-containing phosphate binders. Aluminum deposition in bone (see below) leads to an osteomalacia-like picture. The other entity is a low- turnover bone disease termed “aplastic” or “adynamic” bone disease; PTH levels are lower than typically observed in CKD patients with secondary HPT. It is believed that the condition is caused, at least in part, by excessive PTH suppression, which may be even greater than previously appreciated in light of evidence that some of the immunoreactive PTH detected by most commercially available PTH assays is not the full-length biologically active molecule (as discussed above) but may consist of amino-terminally truncated fragments that do not activate the PTH1R.

TREATMENT

TREATMENT

Hypercalcemia in Secondary HPT

Medical therapy to reverse secondary HPT in CKD includes reduction of excessive blood phosphate by restriction of dietary phosphate, the use of nonabsorbable phosphate binders, and careful, selective addition of calcitriol (0.25–2 μg/d) or related analogues. Calcium carbonate became preferred over aluminum-containing antacids to prevent aluminum-induced bone disease. However, synthetic gels that also bind phosphate (such as sevelamer; Chap. 305) are now widely used, with the advantage of avoiding not only aluminum retention, but excess calcium loading, which may contribute to cardiovascular calcifications. Intravenous calcitriol (or related analogues), administered as several pulses each week, helps control secondary HPT. Aggressive but carefully administered medical therapy can often, but not always, reverse HPT and its symptoms and manifestations.

Occasional patients develop severe manifestations of secondary HPT, including hypercalcemia, pruritus, extraskeletal calcifications, and painful bones, despite aggressive medical efforts to suppress the HPT. PTH hypersecretion no longer responsive to medical therapy, a state of severe HPT in patients with CKD that requires surgery, has been referred to as *tertiary hyperparathyroidism*. Parathyroid surgery is necessary to control this condition. Based on genetic evidence from examination of tumor samples in these patients, the emergence of autonomous parathyroid function is due to a monoclonal outgrowth of one or more previously hyperplastic parathyroid glands. The adaptive response has become an independent contributor to disease; this finding seems to emphasize the importance of optimal medical management to reduce the proliferative response of the parathyroid cells that enables the irreversible genetic change.

Aluminum Intoxication

Aluminum intoxication (and often hypercalcemia as a complication of medical treatment) in the past occurred in patients on chronic dialysis; manifestations included acute dementia and unresponsive and severe osteomalacia. Bone pain, multiple nonhealing fractures, particularly of the ribs and pelvis, and a proximal myopathy occur. Hypercalcemia develops when these patients are treated with vitamin D or calcitriol because of impaired skeletal responsiveness. Aluminum is present at the site of osteoid mineralization, osteoblastic activity is minimal, and calcium incorporation into the skeleton is impaired. The disorder is now rare because of the avoidance of aluminum-containing antacids or aluminum excess in the dialysis regimen (Chap. 408).

Milk-Alkali Syndrome

The milk-alkali syndrome is due to excessive ingestion of calcium and absorbable antacids such as milk or calcium carbonate. It is much less frequent since proton-pump inhibitors and other treatments became available for peptic ulcer disease. For a time, the increased use of calcium carbonate in the management of secondary HPT led to reappearance of the syndrome. Several clinical presentations—acute, subacute, and chronic—have been described, all of which feature hypercalcemia, alkalosis, and renal failure. The chronic form of the disease, termed *Burnett’s syndrome*, is associated with irreversible renal damage. The acute syndromes reverse if the excess calcium and absorbable alkali are stopped.

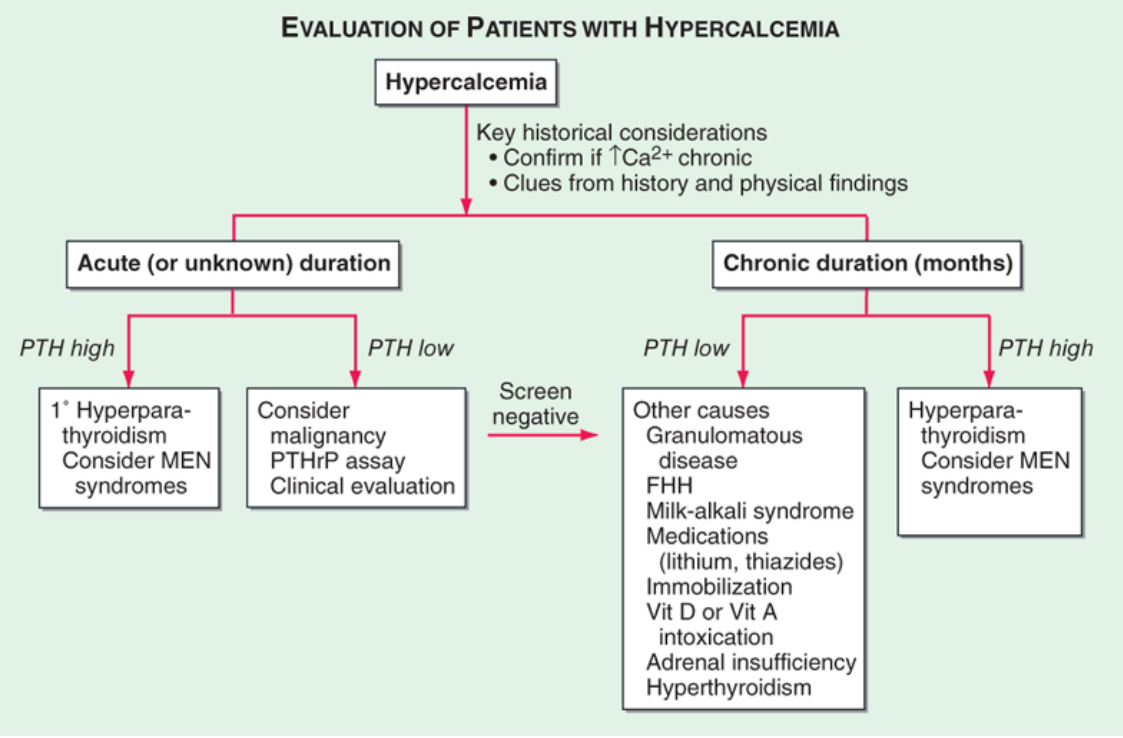
Individual susceptibility is important in the pathogenesis, as some patients are treated with calcium carbonate and alkali regimens without developing the syndrome. One variable is the fractional calcium absorption as a function of calcium intake. Some individuals absorb a high fraction of calcium, even with intakes ≥2 g of elemental calcium per day, instead of reducing calcium absorption with high intake, as occurs in most normal individuals. Resultant mild hypercalcemia after meals in such patients is postulated to contribute to the generation of alkalosis. Development of hypercalcemia causes increased sodium excretion and some depletion of total-body water. These phenomena and perhaps some suppression of endogenous PTH secretion due to mild hypercalcemia lead to increased bicarbonate resorption and to alkalosis in the face of continued calcium carbonate ingestion. Alkalosis per se selectively enhances calcium resorption in the distal nephron, thus aggravating the hypercalcemia. The cycle of mild hypercalcemia → bicarbonate retention → alkalosis → renal calcium retention → severe hypercalcemia perpetuates and aggravates hypercalcemia and alkalosis as long as calcium and absorbable alkali are ingested.

DIFFERENTIAL DIAGNOSIS: SPECIAL TESTS

Differential diagnosis of hypercalcemia is best achieved by using clinical criteria, but immunometric assays to measure PTH are especially useful in distinguishing among major causes **(Fig. 403-6)**. The clinical features that deserve emphasis are the presence or absence of symptoms or signs of disease and evidence of chronicity. If one discounts fatigue or depression, >90% of patients with primary HPT have *asymptomatic hypercalcemia*; symptoms of malignancy are usually present in cancer-associated hypercalcemia. Disorders other than HPT and malignancy cause <10% of cases of hypercalcemia, and some of the nonparathyroid causes are associated with clear-cut manifestations such as renal failure.

FIGURE 403-6

Algorithm for the evaluation of patients with hypercalcemia. See text for details. FHH, familial hypocalciuric hypercalcemia; MEN, multiple endocrine neoplasia; PTH, parathyroid hormone; PTHrP, parathyroid hormone–related peptide.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HPT is the likely diagnosis in patients with *chronic hypercalcemia*. If hypercalcemia has been manifest for >1 year, malignancy can usually be excluded as the cause. A striking feature of malignancy-associated hypercalcemia is the rapidity of the course, whereby signs and symptoms of the underlying malignancy are evident within months of the detection of hypercalcemia. Although clinical considerations are helpful in arriving at the correct diagnosis of the cause of hypercalcemia, appropriate laboratory testing is essential for definitive diagnosis. The immunoassay for PTH usually separates HPT from all other causes of hypercalcemia (exceptions are very rare reports of ectopic production of excess PTH by nonparathyroid tumors). Patients with HPT have elevated PTH levels despite hypercalcemia, whereas patients with malignancy and the other causes of hypercalcemia (except for disorders mediated by PTH such as lithium-induced hypercalcemia) have levels of hormone below normal or undetectable. Assays based on the double-antibody method for PTH exhibit very high sensitivity (especially if serum calcium is simultaneously evaluated) and specificity for the diagnosis of primary HPT (Fig. 403-4).

In summary, PTH values are elevated in >90% of parathyroid-related causes of hypercalcemia, undetectable or low in malignancy-related hypercalcemia, and undetectable or normal in vitamin D–related and high-bone-turnover causes of hypercalcemia. In view of the specificity of the PTH immunoassay and the high frequency of HPT in hypercalcemic patients, it is cost-effective to measure the PTH level in all hypercalcemic patients unless malignancy or a

specific nonparathyroid disease is obvious. False-positive PTH assay results are rare. Immunoassays for PTHrP are helpful in diagnosing certain types of malignancy-associated hypercalcemia. Although FHH is parathyroid-related, the disease should be managed distinctively from HPT. Clinical features and the low urinary calcium excretion can help make the distinction. Because the incidence of malignancy and HPT both increase with age, they can coexist as two independent causes of hypercalcemia.

$1,25(OH)_2D$ levels are elevated in many (but not all) patients with primary HPT. In other disorders associated with hypercalcemia, concentrations of $1,25(OH)_2D$ are low or, at the most, normal. However, this test is of low specificity and is not cost-effective, as not all patients with HPT have elevated $1,25(OH)_2D$ levels and not all nonparathyroid hypercalcemic patients have suppressed $1,25(OH)_2D$. Measurement of $1,25(OH)_2D$ is, however, critically valuable in establishing the cause of hypercalcemia in sarcoidosis and certain lymphomas.

A useful general approach is outlined in Fig. 403-6. If the patient is *asymptomatic* and there is evidence of *chronicity* to the hypercalcemia, HPT is almost certainly the cause. If PTH levels (usually measured at least twice) are elevated, the clinical impression is confirmed and little additional evaluation is necessary. If there is only a short history or no data as to the duration of the hypercalcemia, *occult malignancy* must be considered; if the PTH levels are not elevated, then a thorough workup must be undertaken for malignancy, including chest x-ray, CT of chest and abdomen, and bone scan. Immunoassays for PTHrP may be especially useful in such situations. Attention should also be paid to clues for underlying hematologic disorders such as anemia, increased plasma globulin, and abnormal serum immunoelectrophoresis; bone scans can be negative in some patients with metastases such as in multiple myeloma. Finally, if a patient with chronic hypercalcemia is asymptomatic and malignancy therefore seems unlikely on clinical grounds, but PTH values are not elevated, it is useful to search for other chronic causes of hypercalcemia such as occult sarcoidosis. A careful history of dietary supplements and drug use may suggest intoxication with vitamin D or vitamin A or the use of thiazides.

TREATMENT

TREATMENT

General Approach to Hypercalcemic States

The approach to medical treatment of hypercalcemia varies with its severity. Mild hypercalcemia, <3 mmol/L (12 mg/dL), can be managed by hydration. More severe hypercalcemia (levels of 3.2–3.7 mmol/L [13–15 mg/dL]) must be managed aggressively; above that level, hypercalcemia can be life-threatening and requires emergency measures **(Table 403-4)**. By using a combination of approaches in severe hypercalcemia, the serum calcium concentration can be decreased by 0.7–2.2 mmol/L (3–9 mg/dL) within 24–48 h in most patients, enough to relieve acute symptoms, prevent death from hypercalcemic crisis, and permit diagnostic evaluation. Therapy can then be directed at the underlying disorder—the second priority.

Hypercalcemia develops because of excessive skeletal calcium release, increased intestinal calcium absorption, or inadequate renal calcium excretion. Understanding the particular pathogenesis helps guide therapy. For example, hypercalcemia in patients with malignancy is primarily due to excessive skeletal calcium release and is, therefore, minimally improved by restriction of dietary calcium. On the other hand, patients with vitamin D hypersensitivity or vitamin D intoxication have excessive intestinal calcium absorption, and restriction of dietary calcium is beneficial. Decreased renal function or ECF depletion decreases urinary calcium excretion. In such situations, rehydration may rapidly reduce or reverse the hypercalcemia, even though increased bone resorption persists. As outlined below, the more severe the hypercalcemia, the greater the number of combined therapies that should be used. Rapid acting (hours) approaches—rehydration, forced diuresis, and calcitonin—can be used with the most effective antiresorptive agents such as bisphosphonates (since severe hypercalcemia usually involves excessive bone resorption).

HYDRATION, INCREASED SALT INTAKE, MILD AND FORCED DIURESIS

The first principle of treatment is to restore normal hydration. Many hypercalcemic patients are dehydrated because of vomiting, inanition, and/or hypercalcemia-induced defects in urinary concentrating ability. The resultant drop in glomerular filtration rate is accompanied by an additional decrease in renal tubular sodium and calcium clearance. Restoring a normal ECF volume corrects these abnormalities and increases urine calcium excretion by 2.5–7.5 mmol/d (100–300 mg/d). Increasing urinary sodium excretion to 400–500 mmol/d increases urinary calcium excretion even further than simple rehydration. After rehydration has been achieved, saline can be administered or furosemide or ethacrynic acid can be given twice daily to depress the tubular reabsorptive mechanism for calcium (care must be taken to prevent dehydration). The combined use of these therapies can increase urinary calcium excretion to ≥12.5 mmol/d (500 mg/d) in most hypercalcemic patients. Since this is a substantial percentage of the exchangeable calcium pool, the serum calcium concentration usually falls 0.25–0.75 mmol/L (1–3 mg/dL) within 24 h. Precautions should be taken to prevent potassium and magnesium depletion; calcium-containing renal calculi are a potential complication.

Under life-threatening circumstances, the preceding approach can be pursued more aggressively, but the availability of effective agents to block bone resorption (such as bisphosphonates) has reduced the need for extreme diuresis regimens **(Table 403-5).** Depletion of potassium and magnesium is inevitable unless replacements are given; pulmonary edema can be precipitated. The potential complications can be reduced by careful monitoring of central venous pressure and plasma or urine electrolytes; catheterization of the bladder may be necessary. Dialysis treatment may be needed when renal function is compromised.

BISPHOSPHONATES

The bisphosphonates are analogues of pyrophosphate, with high affinity for bone, especially in areas of increased bone turnover, where they are powerful inhibitors of bone resorption. These bone-seeking compounds are stable in vivo because phosphatase enzymes cannot hydrolyze the central carbon-phosphorus-carbon bond. The bisphosphonates are concentrated in areas of high bone turnover and are taken up by and inhibit osteoclast action; the mechanism of action is complex. The bisphosphonate molecules that contain amino groups in the side chain structure (see below) interfere with prenylation of proteins and can lead to cellular apoptosis. The highly active nonamino group–containing bisphosphonates are also metabolized to cytotoxic products.

The initial bisphosphonate widely used in clinical practice, etidronate, was effective but had several disadvantages, including the capacity to inhibit bone formation as well as blocking resorption. Subsequently, a number of second- or third-generation compounds have become the mainstays of antiresorptive therapy for treatment of hypercalcemia and osteoporosis. The newer bisphosphonates have a highly favorable ratio of blocking resorption versus inhibiting bone formation; they inhibit osteoclast-mediated skeletal resorption yet do not cause mineralization defects at ordinary doses. Though the bisphosphonates have similar structures, the routes of administration, efficacy, toxicity, and side effects vary. The potency of the compounds for inhibition of bone resorption varies more than 10,000-fold, increasing in the order of etidronate, tiludronate, pamidronate, alendronate, risedronate, and zolendronate. The IV use of pamidronate and zolendronate is approved for the treatment of hypercalcemia; between 30 and 90 mg pamidronate, given as a single IV dose over a few hours, returns serum calcium to normal within 24–48 h with an effect that lasts for weeks in 80–100% of patients. Zolendronate given in doses of 4 or 8 mg/5-min infusion has a more rapid and more sustained effect than pamidronate in direct comparison.

These drugs are used extensively in cancer patients. Absolute survival improvements are noted with pamidronate and zolendronate in multiple myeloma, for example. However, though still rare, there are increasing reports of jaw necrosis, especially after dental surgery, mainly in cancer patients treated with multiple doses of the more potent bisphosphonates.

DENOSUMAB

Denosumab is the most recent antiresorptive therapy to be approved for the treatment of hypercalcemia. It is a monoclonal antibody that binds to receptor activator of nuclear factor-κB (RANKL) and prevents it from binding to the receptor RANK on osteoclast precursors and mature osteoclasts. The inhibition of differentiation, activation, and function of osteoclasts leads to a reduction in bone resorption. It has a profound suppressive effect on biochemical markers of bone resorption and is the most powerful antiresorptive agent currently available. Repeated doses of denosumab, 120 mg given subcutaneously, may be effective in patients with hypercalcemia of malignancy who have lost responsiveness to bisphosphonates.

OTHER THERAPIES

Calcitonin acts within a few hours of its administration, principally through receptors on osteoclasts, to block bone resorption. Calcitonin, after 24 h of use, is no longer effective in lowering calcium. Tachyphylaxis, a known phenomenon with this drug, seems to explain this effect, since the drug is initially often effective. Therefore, in life-threatening hypercalcemia, calcitonin can be used effectively within the first 24 h in combination with rehydration and saline diuresis while waiting for more sustained effects from a simultaneously administered bisphosphonate such as pamidronate. Usual doses of calcitonin are 2–8 U/kg of body weight IV, SC, or IM every 6–12 h. *Plicamycin* (formerly mithramycin), which inhibits bone resorption and *gallium nitrate*, which exerts a hypocalcemic action also by inhibiting bone resorption, is no longer used because of superior alternatives such as bisphosphonates.

Glucocorticoids have utility, especially in hypercalcemia complicating certain malignancies. They increase urinary calcium excretion and decrease intestinal calcium absorption when given in pharmacologic doses, but they also cause negative skeletal calcium balance. In normal individuals and in patients with primary HPT, glucocorticoids neither increase nor decrease the serum calcium concentration. In patients with hypercalcemia due to certain osteolytic malignancies, however, glucocorticoids may be effective as a result of antitumor effects. The malignancies in which hypercalcemia responds to glucocorticoids include multiple myeloma, leukemia, Hodgkin's disease, other lymphomas, and carcinoma of the breast, at least early in the course of the disease. Glucocorticoids are also effective in treating hypercalcemia due to vitamin D intoxication and sarcoidosis. Glucocorticoids are also useful in the rare form of hypercalcemia, now recognized in certain autoimmune disorders in which inactivating antibodies against the receptor imitate FHH. Elevated PTH and calcium levels are effectively lowered by the glucocorticoids. In all the preceding situations, the hypocalcemic effect develops over several days, and the usual glucocorticoid dosage is 40–100 mg prednisone (or its equivalent) daily in four divided doses. The side effects of chronic glucocorticoid therapy may be acceptable in some circumstances.

Dialysis is often the treatment of choice for severe hypercalcemia complicated by renal failure, which is difficult to manage medically. Peritoneal dialysis with calcium-free dialysis fluid can remove 5–12.5 mmol (200–500 mg) of calcium in 24–48 h and lower the serum calcium concentration by 0.7–2.2 mmol/L (3–9 mg/dL). Large quantities of phosphate are lost during dialysis, and serum inorganic phosphate concentration usually falls, potentially aggravating hypercalcemia. Therefore, the serum inorganic phosphate concentration should be measured after dialysis, and phosphate supplements should be added to the diet or to dialysis fluids if necessary.

Phosphate therapy, PO or IV, has a limited role in certain circumstances (Chap. 402). Correcting hypophosphatemia lowers the serum calcium concentration by several mechanisms, including bone/calcium exchange. The usual oral treatment is 1–1.5 g phosphorus per day for several days, given in divided doses. It is generally believed, but not established, that toxicity does not occur if therapy is limited to restoring serum inorganic phosphate concentrations to normal.

Raising the serum inorganic phosphate concentration above normal decreases serum calcium levels, sometimes strikingly. Intravenous phosphate is one of the most dramatically effective treatments available for severe hypercalcemia but is toxic and even dangerous (fatal hypocalcemia). For these reasons, it is used rarely and only in severely hypercalcemic patients with cardiac or renal failure where dialysis, the preferable alternative, is not feasible or is unavailable.

SUMMARY

The various therapies for hypercalcemia are listed in Table 403-4. The choice depends on the underlying disease, the severity of the hypercalcemia, the serum inorganic phosphate level, and the renal, hepatic, and bone marrow function. Mild hypercalcemia (≤3 mmol/L [12 mg/dL)]) can usually be managed by hydration. Severe hypercalcemia (≥3.7 mmol/L [15 mg/dL]) requires rapid correction. IV pamidronate, or zolendronate, or subcutaneous denosumab should be administered. In addition, for the first 24–48 h, aggressive sodium-calcium diuresis with IV saline should be given and, following rehydration, large doses of furosemide or ethacrynic acid, but only if appropriate monitoring is available and cardiac and renal function are adequate. Intermediate degrees of hypercalcemia between 3 and 3.7 mmol/L (12 and 15 mg/dL) should be approached with vigorous hydration and then the most appropriate selection for the patient of the combinations used with severe hypercalcemia.

TABLE 403-4

Therapies for Severe Hypercalcemia

Treatment	Onset of Action	Duration of Action	Advantages	Disadvantages
Most Useful Therapies				
Hydration with normal saline	Hours	During infusion	Rehydration invariably needed	Volume overload
Forced diuresis; normal saline plus loop diuretic	Hours	During treatment	Rapid action	Volume overload, cardiac decompensation, intensive monitoring, electrolyte disturbance, inconvenience
Pamidronate	1–2 days	10–14 days to weeks	High potency; intermediate onset of action	Fever in 20%, hypophosphatemia, hypocalcemia, hypomagnesemia, rarely jaw necrosis, atypical femoral fracture
Zolendronate	1–2 days	>3 weeks	Same as for pamidronate (lasts longer)	Same as pamidronate above
Denosumab	1–2 days	>3 weeks	Strongest antiresorptive	Occasional severe hypocalcemia, rarely jaw necrosis, skin infections, atypical femoral fracture
Special Use Therapies				
Calcitonin	Hours	1–2 days	Rapid onset of action; useful as adjunct in severe hypercalcemia	Rapid tachyphylaxis
Phosphate Oral	24 h	During use	Chronic management (with hypophosphatemia); low toxicity if P <4 mg/dL	Limited use except as adjuvant or chronic therapy
Glucocorticoids	Days	Days, weeks	Oral therapy, antitumor agent	Active only in certain malignancies, vitamin D excess and sarcoidosis; glucocorticoid side effects
Dialysis	Hours	During use and 24–48 h afterward	Useful in renal failure; onset of effect in hours; can immediately reverse life-threatening hypercalcemia	Complex procedure, reserved for extreme or special circumstances

TABLE 403-5

Functional Classification of Hypocalcemia (Excluding Neonatal Conditions)

PTH Absent	
Hereditary hypoparathyroidism	Hypomagnesemia
Acquired hypoparathyroidism	
PTH Ineffective	
Chronic kidney disease	Active vitamin D ineffective
Active vitamin D lacking	Intestinal malabsorption
↓ Dietary intake or sunlight	Vitamin D–dependent rickets type II
Defective metabolism:	
Anticonvulsant therapy	Pseudohypoparathyroidism
Vitamin D–dependent rickets type I	Mutant, less active PTH
PTH Overwhelmed	
Severe, acute hyperphosphatemia	Osteitis fibrosa after parathyroidectomy
Tumor lysis	
Acute kidney injury	
Rhabdomyolysis	

Abbreviation: PTH, parathyroid hormone.

HYPOCALCEMIA

(See also Chap. 50)

PATHOPHYSIOLOGY OF HYPOCALCEMIA: CLASSIFICATION BASED ON MECHANISM

Chronic hypocalcemia is less common than hypercalcemia; causes include chronic renal failure, hereditary and acquired hypoparathyroidism, vitamin D deficiency, PHP, and hypomagnesemia.

Acute rather than chronic hypocalcemia is seen in critically ill patients or as a consequence of certain medications and often does not require specific treatment. Transient hypocalcemia is seen with severe sepsis, burns, acute kidney injury, and extensive transfusions with citrated blood. Although as many as one-half of patients in an intensive care setting are reported to have calcium concentrations of <2.1 mmol/L (8.5 mg/dL), most do not have a reduction in ionized calcium. Patients with severe sepsis may have a decrease in ionized calcium (true hypocalcemia), but in other severely ill individuals, hypoalbuminemia is the primary cause of the reduced total calcium concentration. Alkalosis increases calcium binding to proteins, and in this setting direct measurements of ionized calcium should be made.

Medications such as protamine, heparin, and glucagon may cause transient hypocalcemia. These forms of hypocalcemia are usually not associated with tetany and resolve with improvement in the overall medical condition. The hypocalcemia after repeated transfusions of citrated blood usually resolves quickly.

Patients with *acute pancreatitis* have hypocalcemia that persists during the acute inflammation and varies in degree with disease severity. The cause of hypocalcemia remains unclear. PTH values are reported to be low, normal, or elevated, and both resistance to PTH and impaired PTH secretion have been

postulated. Occasionally, a chronic low total calcium and low ionized calcium concentration are detected in an elderly patient without obvious cause and with a paucity of symptoms; the pathogenesis is unclear.

Chronic hypocalcemia, however, is usually symptomatic and requires treatment. Neuromuscular and neurologic manifestations of chronic hypocalcemia include muscle spasms, carpopedal spasm, facial grimacing, and, in extreme cases, laryngeal spasm and convulsions. Respiratory arrest may occur. Increased intracranial pressure occurs in some patients with long-standing hypocalcemia, often in association with papilledema. Mental changes include irritability, depression, and psychosis. The QT interval on the electrocardiogram is prolonged, in contrast to its shortening with hypercalcemia. Arrhythmias occur, and digitalis effectiveness may be reduced. Intestinal cramps and chronic malabsorption may occur. Chvostek's or Trousseau's sign can be used to confirm latent tetany.

The classification of hypocalcemia shown in Table 403-5 is based on an organizationally useful premise that PTH is responsible for minute-to-minute regulation of plasma calcium concentration and, therefore, that the occurrence of hypocalcemia must mean a failure of the homeostatic action of PTH. Failure of the PTH response can occur if there is hereditary or acquired parathyroid gland failure, if a mutant PTH is secreted, or if PTH is ineffective in target organs, or if the action of the hormone is overwhelmed by the loss of calcium from the ECF at a rate faster than it can be replaced.

PTH ABSENT

Whether hereditary or acquired, hypoparathyroidism has a number of common components. The disease is rare with estimates from all causes to be ~25–35 patients/100,000 of the population (based on U.S. and Danish estimates). Symptoms of untreated hypocalcemia are shared by both types of hypoparathyroidism, although the onset of hereditary hypoparathyroidism can be more gradual and associated with other developmental defects. Basal ganglia calcification and extrapyramidal syndromes are more common and earlier in onset in hereditary hypoparathyroidism. Acquired hypoparathyroidism secondary to surgery in the neck is still more common than hereditary hypoparathyroidism, but the frequency of surgically induced parathyroid failure has diminished as a result of improved surgical techniques that spare the parathyroid glands and increased use of nonsurgical therapy for hyperthyroidism. PHP, an example of ineffective PTH action rather than a failure of parathyroid gland production, may share several features with hypoparathyroidism, including extraosseous calcification and extrapyramidal manifestations such as choreoathetotic movements and dystonia.

Papilledema and raised intracranial pressure may occur in both hereditary and acquired hypoparathyroidism, as do chronic changes in fingernails and hair and lenticular cataracts, the latter usually reversible with treatment of hypocalcemia. Certain skin manifestations, including alopecia and candidiasis, are characteristic of hereditary hypoparathyroidism associated with autoimmune polyglandular failure **(Chap. 381)**.

Hypocalcemia associated with hypomagnesemia is associated with both deficient PTH release and impaired responsiveness to the hormone. Patients with hypocalcemia secondary to hypomagnesemia have absent or low levels of circulating PTH, indicative of diminished hormone release despite a maximum physiologic stimulus by hypocalcemia. Plasma PTH levels return to normal with correction of the hypomagnesemia. Thus hypoparathyroidism with low levels of PTH in blood can be due to hereditary gland failure, acquired gland failure, or acute but reversible gland dysfunction (hypomagnesemia).

Genetic Abnormalities and Hereditary Hypoparathyroidism

Hereditary hypoparathyroidism can occur as an isolated entity without other endocrine or dermatologic manifestations. More typically it is syndromic, occurring in association with other abnormalities such as defective development of the thymus or failure of other endocrine organs such as the adrenal, thyroid, or ovary **(Chap. 381)**. Hereditary hypoparathyroidism is often manifest within the first decade but may appear later.

Genetic defects associated with hypoparathyroidism serve to illuminate the complexity of organ development, hormonal biosynthesis and secretion, and tissue-specific patterns of endocrine effector function When hypoparathyroidism is associated with other developmental or organ defects, treatment of the hypocalcemia can still be effective.

A form of hypoparathyroidism associated with defective development of both the thymus and the parathyroid glands is termed the *DiGeorge syndrome*, or the *velocardiofacial syndrome*. Congenital cardiovascular, facial, and other developmental defects are present, and patients may die in early childhood with severe infections, hypocalcemia and seizures, or cardiovascular complications. Patients can survive into adulthood, and milder, incomplete forms occur. Most cases are sporadic, but an autosomal dominant form involving microdeletions of chromosome 22q11.2 has been described. Smaller deletions in chromosome 22 are seen in incomplete forms of the DiGeorge syndrome, appearing in childhood or adolescence, that are manifest primarily by parathyroid gland failure. The chromosome 22 defect is now termed *DSG1*; more recently, a defect in chromosome 10p is also recognized—now called *DSG2*. The phenotypes seem similar. Studies on the chromosome 22 defect have pinpointed a transcription factor, TBX1. Deletions of the orthologous mouse gene show a phenotype similar to the human syndrome.

Another autosomal dominant developmental defect, featuring hypoparathyroidism, deafness, and renal dysplasia (HDR) has been studied at the genetic level. Cytogenic abnormalities in some, but not all kindred, point to translocation defects on chromosome 10, as in DiGeorge syndrome. However, the lack of immunodeficiency and heart defects distinguishes the two syndromes. Mouse models, as well as deletional analysis in some HDR patients, have identified the transcription factor GATA3, which is important in embryonic development and is expressed in developing kidney, ear structures, and the parathyroids.

Another pair of linked developmental disorders involving the parathyroids is recognized. *Kenney-Caffey syndrome type I* features hypoparathyroidism, short stature, osteosclerosis, and thick cortical bones. A defect seen in Middle Eastern patients, particularly in Saudi Arabia, termed *Sanjad-Sakati syndrome*, also exhibits growth failure and other dysmorphic features. This syndrome, which is clearly autosomal recessive, involves a gene on chromosome 1q42-q43. Both syndromes apparently involve a chaperone protein, called *TBCE*, relevant to tubulin function. Recently, a defect in FAM111A was identified as the cause of *Kenney-Caffey syndrome type 2*.

Hypoparathyroidism can occur in association with a complex hereditary autoimmune syndrome involving failure of the adrenals, the ovaries, the immune system, and the parathyroids in association with recurrent mucocutaneous candidiasis, alopecia, vitiligo, and pernicious anemia **(Chap. 381)**. The responsible gene on chromosome 21q22.3 has been identified. The protein product, which resembles a transcription factor, has been termed the *autoimmune regulator*, or AIRE. A stop codon mutation occurs in many Finnish families with the disorder, commonly referred to as *polyglandular autoimmune type 1 deficiency*, while another AIRE mutation (Y85C) is typically observed in Jews of Iraqi and Iranian descent.

Hypoparathyroidism is seen in two disorders associated with mitochondrial dysfunction and myopathy, one termed the *Kearns-Sayre syndrome* (KSS), with ophthalmoplegia and pigmentary retinopathy, and the other termed the *MELAS syndrome*, *m*itochondrial *e*ncephalopathy, *l*actic *a*cidosis, and *s*troke-like episodes. Mutations or deletions in mitochondrial genes have been identified.

Several forms of hypoparathyroidism, each rare in frequency, are seen as isolated defects; the genetic mechanisms are varied. The inheritance includes autosomal dominant, autosomal recessive, and X-linked modes. Three separate autosomal defects involving the parathyroid gene have been recognized: one is dominant and the other two are recessive. The dominant form has a point mutation in the signal sequence, a critical region involved in intracellular transport of the hormone precursor. An Arg for Cys mutation interferes with processing of the precursor and is believed to trigger an apoptotic cellular response, hence acting as a dominant negative. The other two forms are recessive. A point mutation also blocks cleavage of the PTH precursor but requires both alleles to be mutated. Another involves a single-nucleotide base change that results in an exon splicing defect; the lost exon contains the promoter—hence, the gene is silenced. An autosomal recessive form of the disease was found to be caused by an Arg to Cys mutation at amino acid residue 25 of the secreted PTH. An X-linked recessive form of hypoparathyroidism has been described in males and the defect has been localized to chromosome Xq26-q27, perhaps involving the *SOX3* gene.

Abnormalities in the CaSR are detected in three distinctive hypocalcemic disorders. All are rare but more than 10 different gain-of-function mutations have been found in one form of hypocalcemia termed *autosomal dominant hypocalcemic hypercalciuria (ADHH)*. The receptor senses the ambient calcium level as excessive and suppresses PTH secretion, leading to hypocalcemia. The hypocalcemia is aggravated by constitutive receptor activity in the renal tubule causing excretion of inappropriate amounts of calcium. Recognition of the syndrome is important because efforts to treat the hypocalcemia with vitamin D analogues and increased oral calcium exacerbate the already excessive urinary calcium excretion (several grams or more per 24 h), leading to irreversible renal damage from stones and ectopic calcification.

Other causes of isolated hypoparathyroidism include homozygous, inactivating mutations in the parathyroid-specific transcription factor GCM2, which lead to an autosomal recessive form of the disease, or heterozygous point mutations in GCM2, which have a dominant negative effect on the wild-type protein and thus lead to an autosomal dominant form of hypoparathyroidism. Furthermore, heterozygous mutations in Gα11, one of the two signaling proteins downstream of the CaSR, have been identified as a cause of autosomal dominant hypoparathyroidism.

The *Bartter syndrome* is a group of disorders associated with disturbances in electrolyte and acid/base balance, sometimes with nephrocalcinosis and other features. Several types of ion channels or transporters are involved. Curiously, *Bartter syndrome type V* has the electrolyte and pH disturbances seen in the other syndromes but appears to be due to a gain-of-function in the CaSR. The defect may be more severe than in ADHH and explains the additional features seen beyond hypocalcemia and hypercalciuria.

As with autoimmune disorders that block the CaSR (discussed above under hypercalcemic conditions), there are autoantibodies that at least transiently activate the CaSR, leading to suppressed PTH secretion and hypocalcemia. This disorder may wax and wane.

Acquired Hypoparathyroidism

Acquired chronic hypoparathyroidism is usually the result of inadvertent surgical removal of all the parathyroid glands; in some instances, not all the tissue is removed, but the remainder undergoes vascular supply compromise secondary to fibrotic changes in the neck after surgery. In the past, the most frequent cause of acquired hypoparathyroidism was surgery for hyperthyroidism. Hypoparathyroidism now usually occurs after surgery for hyperparathyroidism when the surgeon, facing the dilemma of removing too little tissue and thus not curing the HPT, removes too much. Parathyroid function may not be totally absent in all patients with postoperative hypoparathyroidism.

Rare causes of acquired chronic hypoparathyroidism include radiation-induced damage subsequent to radioiodine therapy of hyperthyroidism and glandular damage in patients with hemochromatosis or hemosiderosis after repeated blood transfusions. Infection may involve one or more of the parathyroids but usually does not cause hypoparathyroidism because all four glands are rarely involved.

Transient hypoparathyroidism is frequent following surgery for HPT. After a variable period of hypoparathyroidism, normal parathyroid function may return due to hyperplasia or recovery of remaining tissue. Occasionally, recovery occurs months after surgery.

TREATMENT

TREATMENT

Acquired and Hereditary Hypoparathyroidism

Conventional treatment has involved replacement with vitamin D or $1,25(OH)_2D_3$ (calcitriol) combined with a high oral calcium intake. In most patients, blood calcium and phosphate levels are satisfactorily regulated, but some patients show resistance and a brittleness, with a tendency to alternate between hypocalcemia and hypercalcemia. For many patients, vitamin D in doses of 40,000–120,000 U/d (1–3 mg/d) combined with ≥1 g elemental calcium is satisfactory. The wide dosage range reflects the variation encountered from patient to patient; precise regulation of each patient is required. Compared to typical daily requirements in euparathyroid patients of 200 U/d (or in older patients as high as 800 U/d), the high dose of vitamin D (as much as 100-fold higher) reflects the reduced conversion of vitamin D to $1,25(OH)_2D$. Many physicians now use 0.5–1 µg of calcitriol in management of such patients, especially if they are difficult to control. Because of its storage in fat, when vitamin D is withdrawn, weeks are required for the disappearance of the biologic effects, compared with a few days for calcitriol, which has a rapid turnover.

Oral calcium and vitamin D restore the overall calcium-phosphate balance but do not reverse the lowered urinary calcium reabsorption typical of hypoparathyroidism. Therefore, care must be taken to avoid excessive urinary calcium excretion after vitamin D and calcium replacement therapy; otherwise, nephrocalcinosis and kidney stones can develop, and the risk of CKD is increased. Thiazide diuretics lower urine calcium by as much as 100 mg/d in hypoparathyroid patients on vitamin D, provided they are maintained on a low-sodium diet. Use of thiazides seems to be of benefit in mitigating hypercalciuria and easing the daily management of these patients.

Hypoparathyroidism is rare among endocrine disorders in not being treated with the missing hormone. Recent developments have changed that. Experimental use of PTH(1-34), the synthetic fragment used in treatment of osteoporosis showed promise. Subsequently, the full length molecule PTH(1-84) has been shown to be effective and is now FDA-approved for therapy of hypoparathyroidism. Published reports illustrate that its use substantially reduced the requirements for supplemental calcium and active vitamin D to maintain serum calcium. Recommendations offered by a recent conference on management of hypoparathyroidism suggest its use, particularly in patients with inadequate control of blood calcium, requirement for inconveniently/excessively high doses of calcium and active vitamin D replacement, and/or high urine calcium.

Hypomagnesemia

Severe hypomagnesemia (<0.4 mmol/L; <0.8 meq/L) is associated with hypocalcemia **(Chap. 402)**. Restoration of the total-body magnesium deficit leads to rapid reversal of hypocalcemia. There are at least two causes of the hypocalcemia—impaired PTH secretion and reduced responsiveness to PTH. **For further discussion of causes and treatment of hypomagnesemia, see Chap. 402.**

The effects of magnesium on PTH secretion are similar to those of calcium; hypermagnesemia suppresses and hypomagnesemia stimulates PTH secretion. The effects of magnesium on PTH secretion are normally of little significance, however, because the calcium effects dominate. Greater change in magnesium than in calcium is needed to influence hormone secretion. Nonetheless, hypomagnesemia might be expected to increase hormone secretion. It is therefore surprising to find that severe hypomagnesemia is associated with blunted secretion of PTH. The explanation for the paradox is that severe, chronic hypomagnesemia leads to intracellular magnesium deficiency, which interferes with secretion and peripheral responses to PTH. The mechanism of the cellular abnormalities caused by hypomagnesemia is unknown, although effects on adenylate cyclase (for which magnesium is a cofactor) have been proposed.

PTH levels are undetectable or inappropriately low in severe hypomagnesemia despite the stimulus of severe hypocalcemia, and acute repletion of magnesium leads to a rapid increase in PTH level. Serum phosphate levels are often not elevated, in contrast to the situation with acquired or idiopathic hypoparathyroidism, probably because phosphate deficiency is often seen in hypomagnesemia **(Chap. 363)**.

Diminished peripheral responsiveness to PTH also occurs in some patients, as documented by subnormal response in urinary phosphorus and urinary cyclic AMP excretion after administration of exogenous PTH to patients who are hypocalcemic and hypomagnesemic. Both blunted PTH secretion and lack of renal response to administered PTH can occur in the same patient. When acute magnesium repletion is undertaken, the restoration of PTH levels to normal or supranormal may precede restoration of normal serum calcium by several days.

TREATMENT

TREATMENT

Hypomagnesemia

Repletion of magnesium cures the condition. Repletion should be parenteral. Attention must be given to restoring the intracellular deficit, which may be considerable. After IV magnesium administration, serum magnesium may return transiently to the normal range, but unless replacement therapy is adequate, serum magnesium will again fall. If the cause of the hypomagnesemia is renal magnesium wasting, magnesium may have to be given long term to prevent recurrence **(Chap. 402)**.

PTH INEFFECTIVE

PTH is ineffective when the PTHR1–signaling protein complex is defective (as in the different forms of PHP, discussed below); when PTH action to promote calcium absorption from the diet via the synthesis of $1,25(OH)_2D$ is insufficient because of vitamin D deficiency or because vitamin D is ineffective (defects in vitamin D receptor or vitamin D synthesis); or in CKD in which the calcium-elevating action of PTH is impaired.

Typically, hypophosphatemia is more severe than hypocalcemia in vitamin D deficiency states because of the increased secretion of PTH, which, although only partly effective in elevating blood calcium, is readily capable of promoting urinary phosphate excretion.

PHP, on the other hand, has a pathophysiology that is different from the other disorders of ineffective PTH action. PHP resembles hypoparathyroidism (in which PTH synthesis is deficient) and is manifested by hypocalcemia and hyperphosphatemia, yet elevated PTH levels. The cause of the disorder is defective PTH-dependent activation of the stimulatory G protein complex or the downstream effector protein kinase A, resulting in failure of PTH to increase intracellular cAMP or to respond to elevated cAMP levels (see below).

Chronic Kidney Disease

Improved medical management of CKD now allows many patients to survive for decades and hence time enough to develop features of renal osteodystrophy, which must be controlled to avoid additional morbidity. Impaired production of $1,25(OH)_2D$ is now thought to be the principal factor that causes calcium deficiency, secondary HPT, and bone disease; hyperphosphatemia typically occurs only in the later stages of the disease. Low levels of $1,25(OH)_2D$ due to increased FGF23 production in bone are critical in the development of hypocalcemia. The uremic state also causes impairment of intestinal absorption by mechanisms other than defects in vitamin D metabolism. Nonetheless, treatment with supraphysiologic amounts of vitamin D or calcitriol can correct the impaired calcium absorption. Since increased FGF23 levels are seen even in early stages of CKD, and have been reported to correlate with increased mortality and left ventricular hypertrophy, there is current interest in approaches to lower intestinal phosphate absorption early during the course of kidney disease and to thereby lower FGF23 levels. However, there is concern as to whether vitamin D supplementation increases the circulating FGF23 levels in CKD patients. Although vitamin D analogs improve survival in this patient population, it is notable that there are often dramatic elevations of FGF23.

Hyperphosphatemia in CKD lowers blood calcium levels by several mechanisms, including extraosseous deposition of calcium and phosphate, impairment of the bone-resorbing action of PTH, and reduction in $1,25(OH)_2D$ production due to elevated FGF23 and diminished renal tissue.

TREATMENT

TREATMENT

Chronic Kidney Disease

Therapy of CKD **(Chap. 305)** involves appropriate management of patients prior to dialysis and adjustment of regimens once dialysis is initiated. Attention should be paid to restriction of phosphate in the diet; avoidance of aluminum-containing phosphate-binding antacids to prevent the problem of aluminum intoxication; provision of an adequate calcium intake by mouth, usually 1–2 g/d; and supplementation with 0.25–1 μg/d calcitriol or other activated forms of vitamin D. Each patient must be monitored closely. The aim of therapy is to restore normal calcium balance to prevent osteomalacia and severe secondary HPT (it is usually recommended to maintain PTH levels between 100 and 300 pg/mL) and, in light of evidence of genetic changes and monoclonal outgrowths of parathyroid glands in CKD patients, to prevent secondary from becoming autonomous HPT. Reduction of hyperphosphatemia and restoration of normal intestinal calcium absorption by calcitriol can improve blood calcium levels and reduce the manifestations of secondary HPT. Since adynamic bone disease can occur in association with low PTH levels, it is important to avoid excessive suppression of the parathyroid glands while recognizing the beneficial effects of controlling the secondary HPT. These patients should probably be closely monitored with PTH assays that detect only the full-length PTH(1–84) to ensure that biologically active PTH and not inactive, inhibitory PTH fragments are measured. Use of phosphate-binding agents such as sevelamer are approved only in end-stage renal disease (ESRD), but it may be necessary to initiate such treatment much earlier during the course of kidney disease to prevent the increase in FGF23 and its “off-target” effects.

Vitamin D Deficiency Due to Inadequate Diet and/or Sunlight

Vitamin D deficiency due to inadequate intake of dairy products enriched with vitamin D, lack of vitamin supplementation, and reduced sunlight exposure in the elderly, particularly during winter in northern latitudes, is more common in the United States than previously recognized. Biopsies of bone in elderly patients with hip fracture (documenting osteomalacia) and abnormal levels of vitamin D metabolites, PTH, calcium, and phosphate indicate that vitamin D deficiency may occur in as many as 25% of elderly patients, particularly in northern latitudes in the United States. Concentrations of 25(OH)D are low or low-normal in these patients. Quantitative histomorphometric analysis of bone biopsy specimens from such individuals reveals widened osteoid seams consistent with osteomalacia **(Chap. 402)**. PTH hypersecretion compensates for the tendency for the blood calcium to fall but also increases renal phosphate excretion and thus causes osteomalacia.

Treatment involves adequate replacement with vitamin D and calcium until the deficiencies are corrected. Severe hypocalcemia rarely occurs in moderately severe vitamin D deficiency of the elderly, but vitamin D deficiency must be considered in the differential diagnosis of mild hypocalcemia.

Mild hypocalcemia, secondary HPT, severe hypophosphatemia, and a variety of nutritional deficiencies occur with gastrointestinal diseases. Hepatocellular dysfunction can lead to reduction in 25(OH)D levels, as in portal or biliary cirrhosis of the liver, and malabsorption of vitamin D and its metabolites, including $1,25(OH)_2D$, may occur in a variety of bowel diseases, hereditary or acquired. Hypocalcemia itself can lead to steatorrhea, due to deficient production of pancreatic enzymes and bile salts. Depending on the disorder, vitamin D or its metabolites can be given parenterally, guaranteeing adequate blood levels of active metabolites.

Defective Vitamin D Metabolism

ANTICONVULSANT THERAPY

Anticonvulsant therapy with any of several agents induces acquired vitamin D deficiency by increasing the conversion of vitamin D to inactive compounds and/or causing resistance to its action. The more marginal the vitamin D intake in the diet, the more likely that anticonvulsant therapy will lead to abnormal mineral and bone metabolism.

VITAMIN D–DEPENDENT RICKETS TYPE I

Vitamin D–dependent rickets type I, previously termed *pseudo-vitamin D–resistant rickets*, differs from true vitamin D–resistant rickets (vitamin D–dependent rickets type II, see below) in that it is typically less severe and the biochemical and radiographic abnormalities can be reversed with appropriate doses of the vitamin's active metabolite, $1,25(OH)_2D$. Physiologic amounts of calcitriol cure the disease **(Chap. 402)**. This finding fits with the pathophysiology of the disorder, which is autosomal recessive, and is now known to be caused by mutations in the gene encoding 25(OH)D-1α-hydroxylase. Both alleles are inactivated in affected patients and compound heterozygotes, harboring distinct mutations, are common.

Clinical features include hypocalcemia, often with tetany or convulsions, hypophosphatemia, secondary HPT, and osteomalacia, often associated with skeletal deformities and increased alkaline phosphatase. Treatment involves physiologic replacement doses of $1,25(OH)_2D$ **(Chap. 402)**.

VITAMIN D–DEPENDENT RICKETS TYPE II

Vitamin D–dependent rickets type II results from end-organ resistance to the active metabolite $1,25(OH)_2D_3$. The clinical features resemble those of the type I disorder and include hypocalcemia, hypophosphatemia, secondary HPT, and rickets but also partial or total alopecia. Plasma levels of $1,25(OH)_2D$ are elevated, in keeping with the refractoriness of the end organs. This disorder is caused by mutations in the gene encoding the vitamin D receptor; treatment is difficult and requires regular, usually nocturnal calcium infusions, which dramatically improve growth, but do not restore hair growth **(Chap. 402)**.

Pseudohypoparathyroidism

PHP refers to a group of distinct inherited disorders. Patients affected by PHP type Ia (PHP-Ia) are characterized by symptoms and signs of hypocalcemia in association with distinctive skeletal and developmental defects. The hypocalcemia is due to a deficient response to PTH, which is probably restricted to the proximal renal tubules. Hyperplasia of the parathyroids, a response to hormone-resistant hypocalcemia, causes elevation of PTH levels. Studies, both clinical and basic, have clarified some aspects of these disorders, including the variable clinical spectrum, the pathophysiology, the genetic defects, and their mode of inheritance.

A working classification of the various forms of PHP is given in **Table 403-6.** The classification scheme is based on the signs of ineffective PTH action (low calcium and high phosphate), low or normal urinary cAMP response to exogenous PTH, the presence or absence of *Albright's hereditary osteodystrophy* (AHO), and assays to measure the concentration of the $G_s\alpha$ subunit of the adenylate cyclase enzyme. Using these criteria, there are four types: PHP types Ia and Ib (PHP1A and PHP1B); pseudopseudohypoparathyroidism (PPHP) and PHP-II (PHP2).

TABLE 403-6

Classification of Pseudohypoparathyroidism (PHP) and Pseudopseudohypoparathyroidism (PPHP)

Type	Hypocalcemia, Hyperphosphatemia	Response of Urinary cAMP to PTH	Serum PTH	$G_s\alpha$ Subunit Deficiency	AHO	Resistance to Hormones Other than PTH
PHP1A	Yes	↓	↑	Yes	Yes	Yes
PPHP	No	Normal	Normal	Yes	Yes	No
PHP1B	Yes	↓	↑	No	No	Yes (in some patients)
PHP2	Yes	Normal	↑	No	No	No
Acrodysostosis with hormonal resistance	Yes	Normal (but ↓ phosphaturic response)	↑	No	Yes	Yes

Abbreviations: ↓, decreased;↑, increased; AHO, Albright's hereditary osteodystrophy; PTH, parathyroid hormone.

PHP1A AND PHP1B

Individuals with PHP1, the most common of the disorders, show a deficient urinary cyclic AMP response to administration of exogenous PTH. Patients with PHP1 are divided into type Ia (PHP1A) and type Ib (PHP1B). Patients with PHP1A show evidence for AHO and reduced amounts of $G_s\alpha$ protein/activity, as determined in readily accessible tissues such as erythrocytes, lymphocytes, and fibroblasts. Only some PHP1B patients show typically AHO features, but they have normal $G_s\alpha$ activity. PHP1C, sometimes listed as a third form of PHP1, is really a variant of PHP1A, although the mutant $G_s\alpha$ shows normal activity in certain in vitro assays.

Most patients who have PHP1A reveal characteristic features of AHO, which consist of short stature, early-onset obesity, round face, obesity, skeletal anomalies (brachydactyly), intellectual impairment, and/or heterotopic calcifications. Patients have low calcium and high phosphate levels, as with true hypoparathyroidism. PTH levels, however, are elevated, reflecting resistance to hormone action.

Amorphous deposits of calcium and phosphate are found in the basal ganglia in about one-half of patients. The defects in metacarpal and metatarsal bones are sometimes accompanied by short phalanges as well, possibly reflecting premature closing of the epiphyses. The typical findings are short fourth and fifth metacarpals and metatarsals. The defects are usually bilateral. Exostoses and radius curvus are frequent.

Inheritance and Genetic Defects

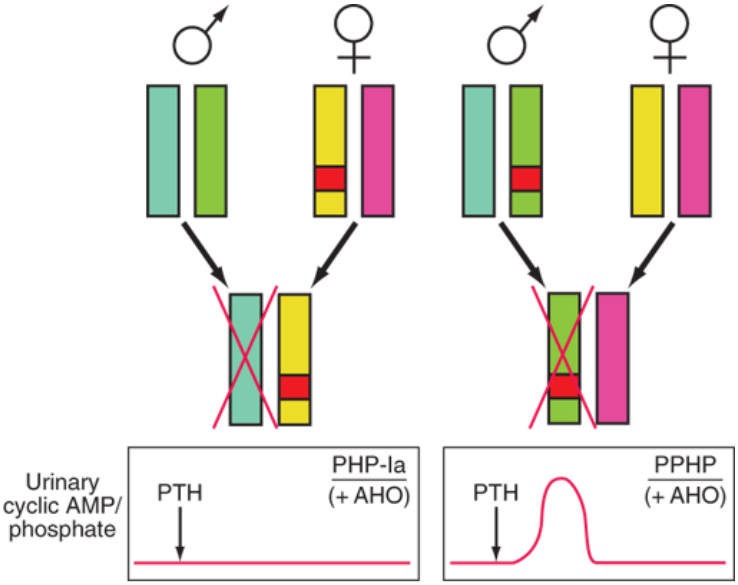
Multiple defects at the *GNAS* locus have now been identified in PHP1A, PHP1B, and PPHP patients. This gene, which is located on chromosome 20q13.3, encodes the α-subunit of the stimulatory G-protein ($G_s\alpha$), among other products (see below). Mutations include abnormalities in splice junctions associated with deficient mRNA production, point mutations, insertions, and/or deletion that all result in a protein with defective function resulting in a 50% reduction of $G_s\alpha$ activity in erythrocytes or other cells.

Detailed analyses of disease transmission in affected kindreds have clarified many features of PHP1A, PPHP, and PHP1B **(Fig. 403-7)**. The former two entities, often traced through multiple generations, have an inheritance pattern consistent with genetic imprinting. The phenomenon of gene imprinting, involving methylation of genetic loci, independent of any mutation, impairs transcription from either the maternal or the paternal allele **(Chap. 456)**. The $G_s\alpha$ transcript is biallelically expressed in most tissues; expression from paternal allele is silenced through as-of-yet unknown mechanisms in some tissues including the proximal renal tubules and the thyroid; consequently, inheritance of a defective paternal allele has no implications with regard to hormonal function. Thus, females affected by either PHP1A or PPHP will have offspring with PHP1A, if these children inherit the allele carrying the *GNAS* mutation; in contrast, if the mutant allele is inherited from a male affected by either disorder, the offspring will exhibit PPHP. Consistent with these data in humans, gene-ablation studies in mice have shown that inheritance of the mutant $G_s\alpha$ allele from the female causes much reduced $G_s\alpha$ protein in renal cortex, hypocalcemia, and resistance to PTH. Offspring inheriting the mutant allele from the male showed no evidence of PTH resistance or hypocalcemia.

FIGURE 403-7

Paternal imprinting of renal parathyroid hormone (PTH) resistance (*GNAS* gene for $G_s\alpha$ subunit) in pseudohypoparathyroidism (PHP1A). An impaired excretion of urinary cyclic AMP and phosphate is observed in patients with PHP. In the renal cortex, there is selective silencing of the paternal $G_s\alpha$ expression. The disease becomes manifest only in patients who inherit the defective gene from an obligate female carrier (*left*). If the genetic defect is inherited from an obligate male gene carrier, there is no biochemical abnormality; administration of PTH causes an appropriate increase in the urinary

cyclic AMP and phosphate concentration (pseudo-PHP [PPHP]; *right*). Both patterns of inheritance lead to Albright's hereditary osteodystrophy (AHO), perhaps because of haplotype insufficiency—i.e., both copies of $G_s\alpha$ must be active in the fetus for normal bone development.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Imprinting is tissue selective. Paternal $G_s\alpha$ expression is not silenced in most tissues. It seems likely, therefore, that the AHO phenotype recognized in PPHP as well as PHP1A reflects $G_s\alpha$ haploinsufficiency during embryonic or postnatal development.

The complex mechanisms that control the *GNAS* gene contribute to challenges involved in unraveling the pathogenesis of these disorders, especially that of PHP1B. Much intensive work with families in which multiple members are affected by PHP1B, as well as studies of the complex regulation of the *GNAS* gene locus, have now shown that autosomal dominant PHP1B is caused by microdeletions within or up-stream of the maternal *GNAS* locus, which are associated with a loss of DNA methylation at one or several loci of the maternal allele (Table 403-6). These abnormalities in methylation silence the expression of the gene. This leads in the proximal renal tubules—where $G_s\alpha$ appears to be expressed exclusively from the maternal allele—to PTH resistance.

PHP1B, lacking the AHO phenotype in most instances, shares with PHP1A the hypocalcemia and hyperphosphatemia caused by PTH resistance, and thus the blunted urinary cyclic AMP response to administered PTH, a standard test to assess the presence or absence of hormone resistance (Table 403-6). Furthermore, these endocrine abnormalities become apparent only if the disease-causing mutation is inherited maternally. Bone responsiveness may be excessive rather than blunted in PHP1B (and in PHP1A) patients, based on case reports that have emphasized an osteitis fibrosa–like pattern in several PHP1B patients.

PHP2 refers to patients with hypocalcemia and hyperphosphatemia, who have a normal urinary cyclic AMP, but an impaired urinary phosphaturic response to PTH. In a PHP2 variant, referred to as acrodysostosis with hormonal resistance (ADOHR), patients have a defect in the regulatory subunit of PKA (PRKAR1A) that mediates the response to PTH distal to cAMP production. Acrodysostosis without hormonal resistance is caused by mutations in the cAMP-selective phosphodiesterase 4 (ADOP4). It remains unclear why the PTH-resistance in some patients, labeled as PHP2 without bony abnormalities, resolves upon treatment with vitamin D supplements.

The diagnosis of these hormone-resistant states can usually be made without difficulty when there is a positive family history for features of AHO, in association with the signs and symptoms of hypocalcemia. In both categories—PHP1A and PHP1B—serum PTH levels are elevated, particularly when patients are hypocalcemic. However, patients with PHP1B or PHP2 without acrodysostosis present only with hypocalcemia and high PTH levels, as evidence for hormone resistance. In PHP1A and PHP1B, the response of urinary cyclic AMP to the administration of exogenous PTH is blunted. The diagnosis of PHP2, in the absence of acrodysostosis, is more complex and vitamin D deficiency must be excluded before such a diagnosis can be entertained.

TREATMENT

TREATMENT

Pseudohypoparathyroidism

Treatment of PHP is similar to that of hypoparathyroidism, except that calcium and activated vitamin D doses are usually higher. Patients with PHP show no PTH resistance in the distal tubules—hence, urinary calcium clearance is typically reduced and they are not at risk of developing nephrocalcinosis as patients with true hypoparathyroidism, unless overtreatment occurs, for example, after the completion of pubertal development and skeletal mutation, when calcium and $1,25(OH)_2$ treatment should be reduced. Variability in response makes it necessary to establish the optimal regimen for each patient, based on maintaining appropriate blood calcium level and urinary calcium excretion, and keeping the PTH level within or slightly above the normal range.

PTH OVERWHELMED

Occasionally, loss of calcium from the ECF is so severe that PTH cannot compensate. Such situations include acute pancreatitis and severe, acute hyperphosphatemia, often in association with renal failure, conditions in which there is rapid efflux of calcium from the ECF. Severe hypocalcemia can occur quickly; PTH rises in response to hypocalcemia but does not return blood calcium to normal.

Severe, Acute Hyperphosphatemia

Severe hyperphosphatemia is associated with extensive tissue damage or cell destruction **(Chap. 402)**. The combination of increased release of phosphate from muscle and impaired ability to excrete phosphorus because of renal failure causes moderate to severe hyperphosphatemia, the latter causing calcium loss from the blood and mild to moderate hypocalcemia. Hypocalcemia is usually reversed with tissue repair and restoration of renal function as phosphorus and creatinine values return to normal. There may even be a mild hypercalcemic period in the oliguric phase of renal function recovery. This sequence, severe hypocalcemia followed by mild hypercalcemia, reflects widespread deposition of calcium in muscle and subsequent redistribution of some of the calcium to the ECF after phosphate levels return to normal.

Other causes of hyperphosphatemia include hypothermia, massive hepatic failure, and hematologic malignancies, either because of high cell turnover of malignancy or because of cell destruction by chemotherapy.

TREATMENT

TREATMENT

Severe, Acute Hyperphosphatemia

Treatment is directed toward lowering of blood phosphate by the administration of phosphate-binding antacids or dialysis, often needed for the management of CKD. Although calcium replacement may be necessary if hypocalcemia is severe and symptomatic, calcium administration during the hyperphosphatemic period tends to increase extraosseous calcium deposition and aggravate tissue damage. The levels of $1,25(OH)_2D$ may be low during the hyperphosphatemic phase and return to normal during the oliguric phase of recovery.

Osteitis Fibrosa after Parathyroidectomy

Severe hypocalcemia after parathyroid surgery is rare now that osteitis fibrosa cystica is an infrequent manifestation of HPT. When osteitis fibrosa cystica is severe, however, bone mineral deficits can be large. After parathyroidectomy, hypocalcemia can persist for days if calcium replacement is inadequate. Treatment may require parenteral administration of calcium; addition of calcitriol and oral calcium supplementation is sometimes needed for weeks to a month or two until bone defects are filled (which, of course, is of therapeutic benefit in the skeleton), making it possible to discontinue parenteral calcium and/or reduce the amount.

DIFFERENTIAL DIAGNOSIS OF HYPOCALCEMIA

Care must be taken to ensure that true hypocalcemia is present; in addition, acute transient hypocalcemia can be a manifestation of a variety of severe, acute illnesses, as discussed above. *Chronic hypocalcemia*, however, can usually be ascribed to a few disorders associated with absent or ineffective PTH. Important clinical criteria include the duration of the illness, signs or symptoms of associated disorders, and the presence of features that suggest a hereditary abnormality. A nutritional history can be helpful in recognizing a low intake of vitamin D and calcium in the elderly, and a history of excessive alcohol intake may suggest magnesium deficiency.

Hypoparathyroidism and PHP are typically lifelong illnesses, usually (but not always) appearing by adolescence; hence, a recent onset of hypocalcemia in an adult is more likely due to nutritional deficiencies, renal failure, or intestinal disorders that result in deficient or ineffective vitamin D. Neck surgery, even long past, however, can be associated with a delayed onset of postoperative hypoparathyroidism. A history of seizure disorder raises the issue of anticonvulsive medication. Developmental defects may point to the diagnosis of PHP. Rickets and a variety of neuromuscular syndromes and deformities may indicate ineffective vitamin D action, either due to defects in vitamin D metabolism or to vitamin D deficiency.

A pattern of *low calcium with high phosphorus* in the absence of renal failure or massive tissue destruction almost invariably means hypoparathyroidism or PHP. A *low calcium and low phosphorus* pattern points to absent or ineffective vitamin D, thereby impairing the action of PTH on calcium metabolism

(but not phosphate clearance). The relative ineffectiveness of PTH in calcium homeostasis in vitamin D deficiency, anticonvulsant therapy, gastrointestinal disorders, and hereditary defects in vitamin D metabolism leads to secondary HPT as a compensation. The excess PTH on renal tubule phosphate transport accounts for renal phosphate wasting and hypophosphatemia.

Exceptions to these patterns may occur. Most forms of hypomagnesemia are due to long-standing nutritional deficiency as seen in chronic alcoholics. Despite the fact that the hypocalcemia is principally due to an acute absence of PTH, phosphate levels are usually low, rather than elevated, as in hypoparathyroidism. Chronic renal failure is often associated with hypocalcemia and hyperphosphatemia, despite secondary HPT.

Diagnosis is usually established by application of the PTH immunoassay, tests for vitamin D metabolites, and measurements of the urinary cyclic AMP response to exogenous PTH. In hereditary and acquired hypoparathyroidism and in severe hypomagnesemia, PTH is either undetectable or inappropriately in the normal range (Fig. 403-4). This finding in a hypocalcemic patient is supportive of hypoparathyroidism, as distinct from ineffective PTH action, in which even mild hypocalcemia is associated with elevated PTH levels. Hence a failure to detect elevated PTH levels establishes the diagnosis of hypoparathyroidism; elevated levels suggest the presence of secondary HPT, as found in many of the situations in which the hormone is ineffective due to associated abnormalities in vitamin D action. Assays for 25(OH)D can be helpful. Low or low-normal 25(OH)D indicates vitamin D deficiency due to lack of sunlight, inadequate vitamin D intake, or intestinal malabsorption. Recognition that mild hypocalcemia, rickets, and hypophosphatemia are due to anticonvulsant therapy is made by history.

FURTHER READING

Bastepe M, Jüppner H: Pseudohypoparathyroidism, Albright's hereditary osteodystrophy, and progressive osseous heteroplasia: Disorders caused by inactivating GNAS mutations, in *Endocrinology*, 7th ed, JL Jameson, LJ DeGroot (eds), 2016, pp. 1147–1159.

Bilezikian J et al: Guidelines for the Management of Asymptomatic Primary Hyperparathyroidism: Summary Statement from the Fourth International Workshop. J Clin Endocrinol Metab 99:3561, 2014. [PubMed: 25162665]

Brandi M et al: Management of Hypoparathyroidism: Summary Statement and Guidelines. J Clin Endocrinol Metab 101:2273, 2016. [PubMed: 26943719]

Gardella T et al: Parathyroid hormone and the parathyroid hormone receptor type 1 in the regulation of calcium and phosphate homeostasis and bone metabolism, in *Endocrinology*, 7th ed, JL Jameson, LJ DeGroot (eds), 2016, pp. 969–990.

Thakker R et al: Regulation of calcium homeostasis and genetic disorders that affect calcium metabolism, in *Endocrinology*, 7th ed, JL Jameson, LJ DeGroot (eds), 2016, pp. 1063–1089.

Chapter 404: Osteoporosis

Robert Lindsay; Felicia Cosman

INTRODUCTION

Osteoporosis, a condition characterized by decreased bone strength, is prevalent among postmenopausal women but also occurs in both women and men as a function of age and with underlying conditions or major risk factors associated with bone demineralization. Its chief clinical manifestations are vertebral and hip fractures, although fractures can occur at almost any skeletal site. Osteoporosis affects >10 million individuals in the United States, but only a small proportion are diagnosed and treated.

DEFINITION

Osteoporosis is defined as a reduction in the strength of bone that leads to an increased risk of fractures. Loss of bone tissue is associated with deterioration in skeletal microarchitecture. The World Health Organization (WHO) operationally defines osteoporosis as a bone density that falls 2.5 standard deviations (SD) below the mean for young healthy adults of the same sex and race—also referred to as a *T-score* of –2.5. Postmenopausal women at the lower end of the young normal range (a T-score <–1.0) are defined as having low bone density and are also at increased risk of osteoporosis. Although risk is lower in this group, >50% of fractures among postmenopausal women, including hip fractures, occur in this group with low bone density. As a consequence, clinical assessment has evolved to include absolute risk of fracture, which incorporates bone mineral density (BMD) with age, gender, and other clinical risk factors to calculate 10-year fracture risk.

Osteoporosis-related fractures are adulthood fractures of any bone that occur in the setting of trauma less than or equal to a fall from standing height, with the exceptions of fingers, toes, face and skull.

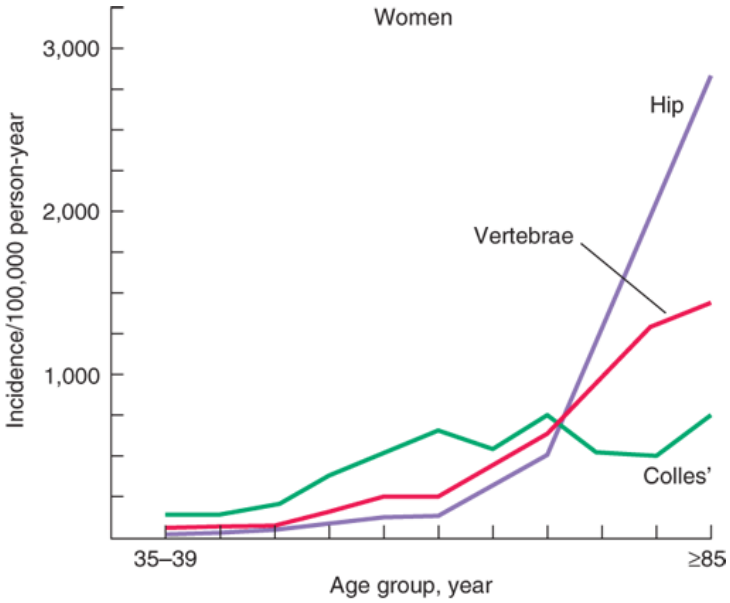
EPIDEMIOLOGY

In the United States, as many as 8 million women and 2 million men have osteoporosis (BMD T-score <–2.5 at lumbar spine, total hip or femoral neck), and >40 million individuals have bone mass levels that put them at increased risk of developing osteoporosis (e.g. BMD T-score <–1.0). Osteoporosis occurs more frequently with increasing age, as bone tissue is lost progressively. In women, the loss of ovarian function at menopause (typically around age 50) precipitates rapid bone loss so that most women meet the diagnostic criterion for osteoporosis by age 70–80. As the population ages, the number of individuals with osteoporosis and fractures will also increase, despite a recognized reduction in age specific risk. It is estimated that currently about 2 million fractures occur each year in the United States as a consequence of osteoporosis. Many of the fractures defined as related to osteoporosis occur in individuals with low bone mass. Within that population, segregation of those at high risk of facture for treatment has become an important issue in clinical management.

The epidemiology of fractures follows the trend for loss of bone density, with most fractures, especially those of the hip and vertebrae, showing exponential increases with advancing age. (Fig. 404-1). Lifetime osteoporotic fracture risk for a woman who reaches the age of 50 is about 50% and corresponding risk for a 50-year old man is about 20%.

FIGURE 404-1

Epidemiology of vertebral, hip, and Colles' fractures with age. *(Adapted from C Cooper, LJ Melton III: Trends Endocrinol Metab 3:224, 1992; with permission.)*



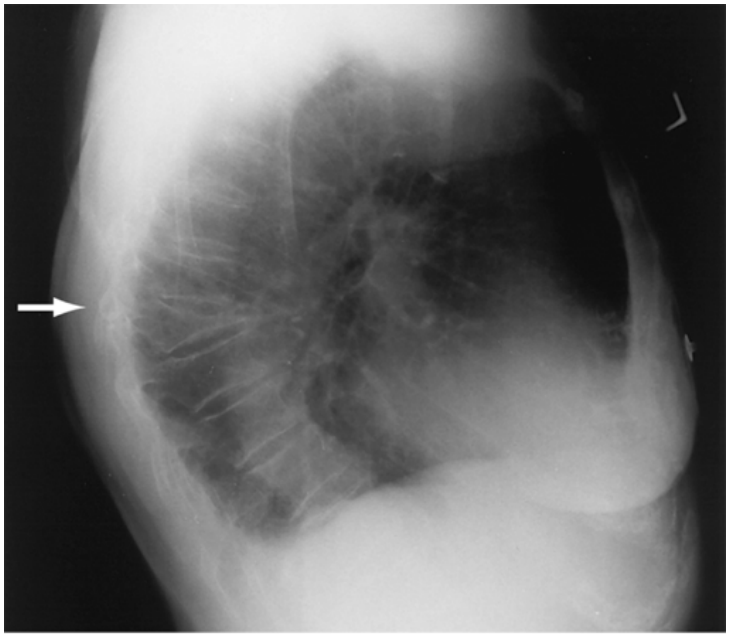
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

About 300,000 hip fractures occur each year in the United States, almost all requiring hospital admission and emergency surgical intervention. The lifetime probability that a 50-year-old white individual will have a hip fracture is 14% for women and 5% for men; the risk for African-Americans is about half of those rates, and the risk for Asians and nonblack Hispanics appears similar to that for Caucasians. Hip fractures are associated with a high incidence of deep-vein thrombosis and pulmonary embolism and a mortality rate between 5 and 20% during the year after surgery, with higher mortality rates among males and African Americans. There is also significant morbidity after hip fracture, with about 30% of survivors requiring long-term care (at least temporarily), and many never regaining the independence that they had prior to the fracture.

There are about 500,000 symptomatic vertebral fractures per year in the United States, but probably >1,000,000 vertebral fractures occur yearly that are not recognized clinically at the time of the event. Many of these initially "silent" vertebral fractures are identified incidentally during radiography for other purposes (Fig. 404-2). Even when asymptomatic, these vertebral fractures are a major sign of skeletal fragility and carry the same predictive value for subsequent fracture. Vertebral fractures rarely require hospitalization, but are associated with long-term morbidity and an increase in mortality rates, primarily related to pulmonary disease. Multiple vertebral fractures lead to height loss (often of several inches), kyphosis, and secondary pain and discomfort related to altered biomechanics of the back. Thoracic fractures can be associated with restrictive lung disease, whereas lumbar fractures are associated with abdominal symptoms that include distention, early satiety, and constipation.

FIGURE 404-2

Lateral spine x-ray showing severe osteopenia and a severe wedge-type deformity (severe anterior compression).

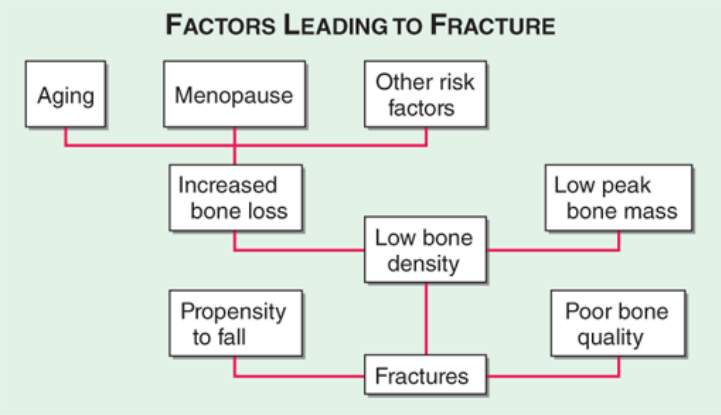


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Approximately 400,000 wrist fractures occur in the United States each year. Fractures of other bones (including about 150,000 pelvic fractures and >100,000 proximal humerus fractures) also occur with osteoporosis. Although some fractures result from major trauma, the threshold for fracture is reduced in osteoporotic bone **(Fig. 404-3)**. In addition to reduced bone density with advancing age, there are a number of risk factors for fracture; the common ones are summarized in **Table 404-1**. Prior fractures, a family history of osteoporosis-related fractures (particularly hip fractures), low body weight, cigarette consumption, and excessive alcohol consumption are all independent predictors of fracture. Chronic diseases with inflammatory components that increase skeletal remodeling, such as rheumatoid arthritis, increase the risk of osteoporosis, as do diseases associated with malabsorption. Chronic diseases that increase the risk of falling or frailty, including dementia, Parkinson's disease, and multiple sclerosis, also increase fracture risk (Table 404-1).

FIGURE 404-3

Factors leading to osteoporotic fractures.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 404-1

Risk Factors for Osteoporosis Fracture

Nonmodifiable	Potentially modifiable
Personal history of fracture as an adult	Current cigarette smoking
History of fracture in first-degree relative	Estrogen deficiency
Female gender	Early menopause (<45 years) or bilateral ovariectomy
Advanced age	Prolonged premenstrual amenorrhea (>1 year)
White race	Poor nutrition especially low calcium and vitamin D intake
Dementia	Alcoholism
	Impaired eyesight despite adequate correction
	Recurrent falls
	Inadequate physical activity
	Poor health/frailty

In the United States and Europe, osteoporosis-related fractures are more common among women than men, presumably due to a lower peak bone mass as well as postmenopausal bone loss in women. However, this gender difference in bone density and age-related increase in hip fractures is not as apparent in some other cultures, possibly due to genetics, physical activity level, or diet.

Fractures are themselves risk factors for future fractures (Table 404-1). Vertebral fractures increase the risk of other vertebral fractures as well as fractures of the peripheral skeleton such as the hip and wrist. Wrist fractures also increase the risk of vertebral and hip fractures. Among individuals aged >50, any fracture except those of the fingers, toes, face, and skull should be considered as potentially related to osteoporosis regardless of the specific circumstances of the fracture. Osteoporotic bone is more likely to fracture than is normal bone at any level of trauma, and a fracture in a person aged >50 should trigger evaluation for osteoporosis. This often does not occur since postfracture care is fragmented. Recent attempts to coordinate care with one individual assuming the responsibility for guiding patients through the system and ensuring their evaluation and treatment for osteoporosis may improve care, but is more difficult to do in the open medical care systems in the United States. In countries with single payor systems, that approach does seem to be effective, as is also the case in closed health care systems in the United States.

The risk for future fracture after a first fracture is not linear. Highest risk occurs within the first 2 years after the first fracture. A recent large Medicare Database study indicated that almost 20% of women will have a second fracture within 2 years after the first. Risk diminishes to less than half of those rates in the subsequent 3 years and declines to baseline thereafter for most fracture types, though risk after a vertebral or hip fracture may persist a bit longer.

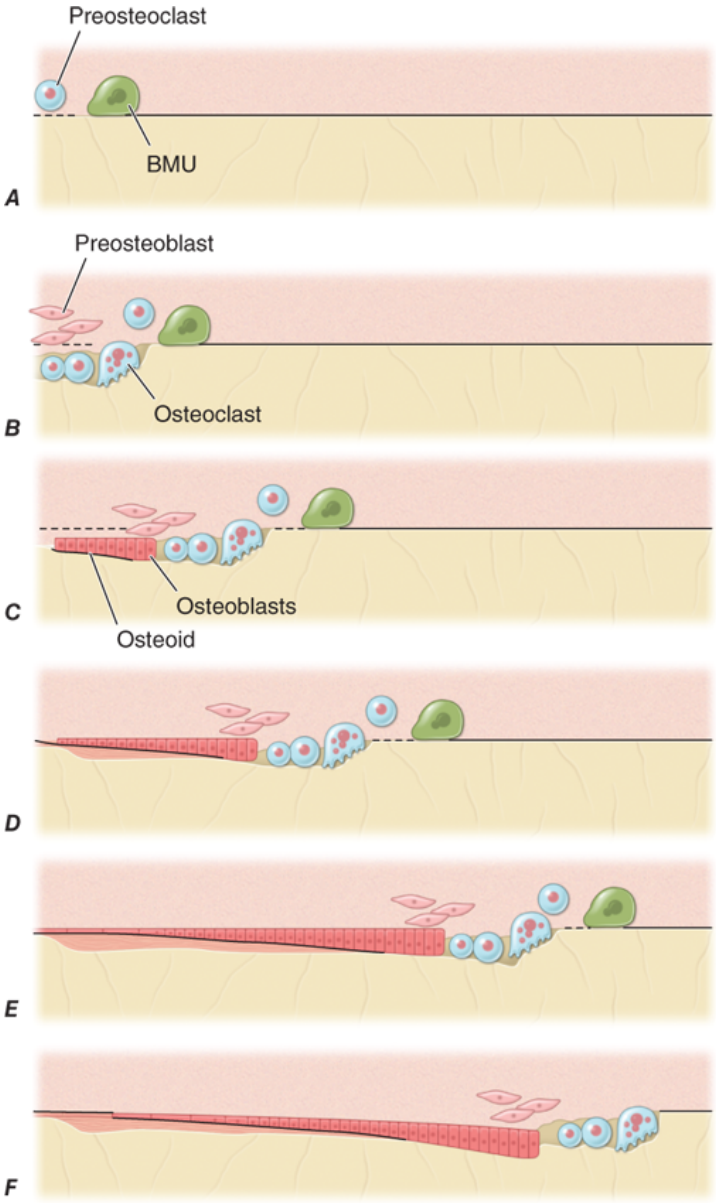
PATHOPHYSIOLOGY

BONE REMODELING

Osteoporosis results from bone loss due to age-related changes in bone remodeling as well as extrinsic and intrinsic factors that exaggerate this process. These changes may be superimposed on a low peak bone mass. Consequently, understanding the bone remodeling process is fundamental to understanding the pathophysiology of osteoporosis **(Chap. 402)**. During growth, the skeleton increases in size by linear growth and by apposition of new bone tissue on the outer surfaces of the cortex **(Fig. 404-4)**. The latter process is called *modeling*, a process that also allows the long bones to adapt in shape to the stresses placed on them. Increased sex hormone production at puberty is required for skeletal maturation, which reaches maximum mass and density in early adulthood. The sexual dimorphism in skeletal size becomes obvious after puberty, although true bone density remains similar between the sexes. Nutrition and lifestyle also play an important role in growth, though genetic factors primarily determine peak skeletal mass and density.

FIGURE 404-4

Mechanism of bone remodeling. The basic molecular unit (BMU) moves along the trabecular surface at a rate of about 10 µm/d. The figure depicts remodeling over ~120 days. ***A.*** Origination of BMU-lining cells contracts to expose collagen and attract preosteoclasts. ***B.*** Osteoclasts fuse into multinucleated cells that resorb a cavity. Mononuclear cells continue resorption, and preosteoblasts are stimulated to proliferate. ***C.*** Osteoblasts align at bottom of cavity and start forming osteoid (black). ***D.*** Osteoblasts continue formation and mineralization. Previous osteoid starts to mineralize (*horizontal lines*). ***E.*** Osteoblasts begin to flatten. ***F.*** Osteoblasts turn into lining cells; bone remodeling at initial surface (left of drawing) is now complete, but BMU is still advancing (to the right). *(Adapted from SM Ott, in JP Bilezikian et al [eds]: Principles of Bone Biology, vol. 18. San Diego, Academic Press, 1996, pp 231–241.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Numerous genes control skeletal growth, peak bone mass, and body size, as well as skeletal structure and density. Heritability estimates of 50–80% for bone density and size have been derived on the basis of twin studies. Though peak bone mass is often lower among individuals with a family history of osteoporosis, association studies of candidate genes (vitamin D receptors; type I collagen, estrogen receptors [ER], and interleukin 6 [IL-6]; and insulin-like growth factor I [IGF-I]) and bone mass, bone turnover, and fracture prevalence have been inconsistent. Linkage studies suggest that a genetic locus on chromosome 11 is associated with high bone mass. Families with high bone mass and without much apparent age-related bone loss have been shown to have a point mutation in LRP5, a low-density lipoprotein receptor–related protein. The role of this gene in the general population is not clear, although a nonfunctional mutation results in osteoporosis-pseudoglioma syndrome, and LRP5 signaling appears to be important in controlling bone formation. Genome-wide scans for low bone mass suggest multiple genes are involved, many of which are also implicated also in control of body size.

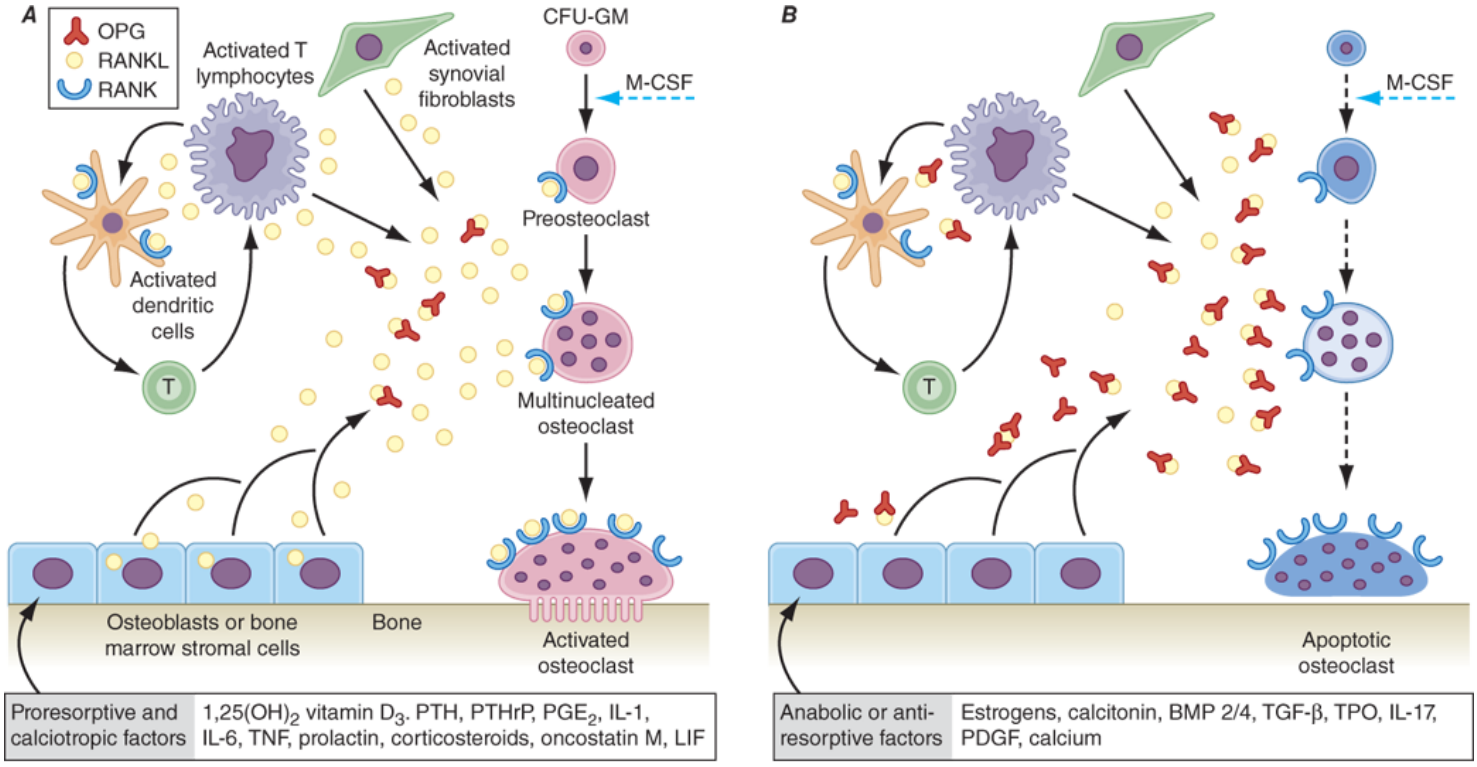
In adults, bone remodeling, not modeling, is the principal metabolic skeletal process. Bone remodeling has two primary functions: (1) to repair microdamage within the skeleton to maintain skeletal strength and ensure the relative youth of the skeleton and (2) to supply calcium when needed from the skeleton to maintain serum calcium. Remodeling may be activated by microdamage to bone as a result of excessive or accumulated stress. Acute demands for calcium involve osteoclast-mediated resorption as well as calcium transport by osteocytes. Chronic demands for calcium can result in secondary hyperparathyroidism, increased bone remodeling, and overall loss of bone tissue. Bone remodeling occurs through well coordinated activity of osteocytes, osteoblasts, and osteoclasts. Osteocytes are the terminal-differentiated cells derived from osteoblasts after incorporation into newly formed bone tissue. Osteoblasts derive from mesenchymal cell lineage and osteoclasts from monocyte/macrophage lineage.

Bone remodeling also is regulated by multiple hormones, including estrogens (in both genders), androgens, vitamin D, and parathyroid hormone (PTH), as well as locally produced growth factors, such as IGF-I, transforming growth factor β (TGF-β), PTH–related peptide (PTHrP), interleukins (ILs), prostaglandins, and members of the tumor necrosis factor (TNF) superfamily. These factors primarily modulate the rate at which new remodeling sites are

activated, a process that results initially in bone resorption by osteoclasts, followed by a period of repair during which new bone tissue is synthesized by osteoblasts (Chap. 402). The cytokine responsible for communication between the osteoblasts, other marrow cells, and osteoclasts is RANK ligand (RANKL) (receptor activator of nuclear factor-kappa-B [NFκB]; RANKL). RANKL, a member of the TNF family, is secreted by osteocytes, osteoblasts, and certain cells of the immune system. The osteoclast receptor for this protein is referred to as *RANK*. Activation of RANK by RANKL is a final common path in osteoclast development and activation. A humoral decoy for RANKL, also secreted by osteoblasts, is referred to as *osteoprotegerin* (Fig. 404-5). Modulation of osteoclast recruitment and activity appears to be related to the interplay among these three factors. Additional influences include nutrition (particularly calcium intake) and physical activity level. RANKL production is in part regulated by the canonical Wnt signaling pathway. Wnt activation through mechanical loading, or by hormonal or cytokine factors, stimulates bone formation by increasing formation and activity of osteoblasts and decreases RANKL secretion, which inhibits production and activity of osteoclasts. Sclerostin, also an osteocyte protein, is a major inhibitor of Wnt activation and bone formation. Both the RANKL and Wnt pathways have become major targets for pharmacologic treatment of osteoporosis (see below).

FIGURE 404-5

Hormonal control of bone resorption. ***A.*** Proresorptive and calciotropic factors. ***B.*** Anabolic and antiosteoclastic factors. RANKL expression is induced in osteoblasts, activated T cells, synovial fibroblasts, and bone marrow stromal cells. It binds to membrane-bound receptor RANK to promote osteoclast differentiation, activation, and survival. Conversely, osteoprotegerin (OPG) expression is induced by factors that block bone catabolism and promote anabolic effects. OPG binds and neutralizes RANKL, leading to a block in osteoclastogenesis and decreased survival of preexisting osteoclasts. CFU-GM, colony-forming units, granulocyte macrophage; M-CSF, macrophage colony-stimulating factor; RANKL, receptor activator of nuclear factor NFκB; PTH, parathyroid hormone; PGE_2, prostaglandin E_2; TNF, tumor necrosis factor; LIF, leukemia inhibitory factor; TPO, thrombospondin; PDGF, platelet-derived growth factor; OPG-L, osteoprotegerin-ligand; IL, interleukin; TGF-β, transforming growth factor β. *(From WJ Boyle et al: Nature 423: 337, 2003.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In young adults, resorbed bone is replaced by an equal amount of new bone tissue. Thus, the mass of the skeleton remains constant after peak bone mass is achieved by the age of about 20. After age 30–45, however, the resorption and formation processes become imbalanced, and resorption exceeds formation. This imbalance may begin at different ages and varies at different skeletal sites; it becomes exaggerated in women after menopause. Excessive bone loss can be due to an increase in osteoclastic activity and/or a decrease in osteoblastic activity. In addition, an increase in remodeling activation frequency, and thus the number of remodeling sites, can magnify the small imbalance seen at each remodeling unit. Increased recruitment of bone remodeling sites produces a reversible reduction in bone tissue but also can result in permanent loss of tissue and disrupted skeletal architecture. In trabecular bone, if the osteoclasts penetrate trabeculae, they leave no template for new bone formation to occur, and, consequently, rapid bone loss ensues and cancellous connectivity becomes impaired. A higher number of remodeling sites increases the likelihood of this event. In cortical bone, increased activation of remodeling creates more porous bone. The effect of this increased porosity on cortical bone strength may be modest if the overall diameter of the bone is not changed. However, decreased apposition of new bone on the periosteal surface coupled with increased endocortical resorption of bone decreases the biomechanical strength of long bones. Even a slight exaggeration in normal bone loss increases the risk of osteoporosis-related fractures because of the architectural changes that occur, and osteoporosis is largely a disease of disordered skeletal architecture, although currently the only clinical tool generally available (dual-energy x-ray absorptiometry [DXA]) measures mass (an estimate of the mineral in bone) not architecture. Several

tools are becoming available that may give more insight into the architecture of the skeleton (including Trabecular Bone Score, a noninvasive addition to DXA).

CALCIUM NUTRITION

Peak bone mass may be impaired by inadequate calcium intake during growth among other nutritional factors (calories, protein, and other minerals), leading to increased risk of osteoporosis later in life. During the adult phase of life, insufficient calcium intake contributes to secondary hyperparathyroidism and an increase in the rate of bone remodeling to assist in maintaining normal serum calcium levels. PTH stimulates the hydroxylation of vitamin D in the kidney, leading to increased levels of 1,25-dihydroxyvitamin D [1,25$(OH)_2$D] and enhanced gastrointestinal calcium absorption. PTH also reduces renal calcium loss. Although these are all appropriate compensatory homeostatic responses for adjusting calcium economy, the long-term effects are detrimental to the skeleton because the increased remodeling rates and the ongoing imbalance between resorption and formation at remodeling sites combine to accelerate loss of bone tissue.

Total daily calcium intakes <400 mg are detrimental to the skeleton, and intakes in the range of 600–800 mg, which is about the average intake among adults in the United States, are also probably suboptimal. The recommended daily required intake of 1000–1200 mg for adults accommodates population heterogeneity in controlling calcium balance **(Chap. 325)**. Such intakes should preferentially come from dietary sources and supplements used only when dietary intakes fall short, and cannot be modified easily. The supplement should be enough to bring total intake to about 1200 mg/d. Recent studies have suggested that there may be differences in safety based on calcium source; high intakes primarily from supplement sources appear to result in a greater risk of renal stones, and perhaps cardiovascular calcifications (although the literature is inconsistent and controversial).

VITAMIN D

Severe vitamin D deficiency causes rickets in children and osteomalacia in adults **(See also Chap. 402)**. However, vitamin D insufficiency may be more prevalent than previously thought, particularly among individuals at increased risk such as the elderly; those living in northern latitudes; and individuals with poor nutrition, obesity, malabsorption, or chronic liver or renal disease. Dark-skinned individuals are also at high risk of vitamin D deficiency. Although there is considerable controversy about overall optimal health targets for serum 25-hydroxyvitamin D (25[OH]D), there is evidence that for optimal skeletal health, serum 25(OH)D should be >75 nmol/L (30 ng/mL). To achieve this level for most adults requires an intake of at least 800–1000 units/d, or higher in individuals with risk factors (as above).

Vitamin D insufficiency leads to compensatory secondary hyperparathyroidism and is an important risk factor for osteoporosis and fractures. Some studies have shown that >50% of inpatients on a general medical service exhibit biochemical features of vitamin D deficiency, including increased levels of PTH and alkaline phosphatase and lower levels of ionized calcium. In women living in northern latitudes, vitamin D levels decline during the winter months. This is associated with seasonal bone loss, reflecting increased bone turnover. Even among healthy ambulatory individuals, mild vitamin D deficiency is increasing in prevalence. In part this is due to decreased exposure to sunlight coupled with increased use of potent sunscreens. Treatment with vitamin D can return levels to normal [>75 µmol/L (30 ng/mL)] and prevent the associated increase in bone remodeling, bone loss, and fractures. Reduced falls and fracture rates also have been documented among individuals in northern latitudes who have greater vitamin D intake and have higher 25(OH)D levels (though one study suggested an increased fall risk with higher 25OHD levels). Although vitamin D levels might affect risk and/or severity of other diseases, including cancers (colorectal, prostate, and breast), autoimmune diseases, multiple sclerosis, cardiovascular disease and diabetes, most controlled clinical trials have not confirmed these effects.

ESTROGEN STATUS

Estrogen deficiency causes bone loss by two distinct but interrelated mechanisms: (1) activation of new bone remodeling sites and (2) exaggeration of the imbalance between bone formation and resorption. The change in activation frequency causes a transient bone loss until a new steady state between resorption and formation is achieved. The remodeling imbalance, however, results in a permanent decrement in mass. In addition, the very presence of more remodeling sites in the skeleton increases the probability that trabeculae will be penetrated, eliminating the template on which new bone can be formed and accelerating the loss of bony tissue.

The most common estrogen-deficient state is the cessation of ovarian function at the time of menopause, which occurs on average at age 51 **(Chap. 388)**. Thus, with current life expectancy, an average woman will spend about 30 years without an ovarian supply of estrogen. Breast cancer treatment with aromatase inhibitors is an increasingly common cause of estrogen deficiency. The mechanism by which estrogen deficiency causes bone loss is summarized in Fig. 404-5. Marrow cells (macrophages, monocytes, osteoclast precursors, mast cells) as well as bone cells (osteoblasts, osteocytes, osteoclasts) express ERs α and β. Loss of estrogen increases production of RANKL and reduces production of osteoprotegerin, increasing osteoclast formation and recruitment. Estrogen also may play a role in determining the life span of bone cells by controlling the rate of apoptosis. Thus, in situations of estrogen deprivation, the life span of osteoblasts may be decreased, whereas the longevity and activity of osteoclasts are increased. The rate and duration of bone loss after menopause are heterogeneous and unpredictable. Once surfaces are lost in cancellous bone, the rate of bone loss declines. In cortical bone, loss is slower but may continue for a longer time period.

Since remodeling is initiated at the surface of bone, it follows that trabecular bone—which has a considerably larger surface area (80% of the total) than cortical bone—will be affected preferentially by estrogen deficiency. Fractures occur earliest at sites where trabecular bone contributes most to bone

strength; consequently, vertebral fractures are the most common early skeletal consequence of estrogen deficiency.

In males, estrogen may an important role in regulation of bone remodeling. In an experiment in which males were rendered estrogen and androgen deficient, restoring estrogen supply reduced remodeling rate more than restoring androgen.

PHYSICAL ACTIVITY

Inactivity, such as prolonged bed rest or paralysis, results in significant bone loss. Concordantly, athletes have higher bone mass than non-athletes. These changes in skeletal mass are most marked when the stimulus begins during growth and before the age of puberty. Adults are less capable than children of increasing bone mass after restoration of physical activity. Epidemiologic data support the beneficial effects on the skeleton of chronic high levels of physical activity. Fracture risk is lower in rural communities and in countries where physical activity is maintained into old age. However, when exercise is initiated during adult life, the effects of moderate exercise on the skeleton are modest, with a bone mass increase of 1–2% in short-term studies of <2 years' duration. It is argued that more active individuals are less likely to fall and are more capable of protecting themselves upon falling, thereby reducing fracture risk. Continuing physical activity into the later years appears to slow cognitive decline, another major reason for including exercise programs for the aging population.

CHRONIC DISEASES

Various genetic and acquired diseases are associated with an increase in the risk of osteoporosis **(Table 404-2)**. Mechanisms that contribute to bone loss are unique for each disease and typically result from multiple factors, including nutrition, reduced physical activity levels, and factors that affect rates of bone remodeling. In most, but not all circumstances, the primary diagnosis is made before osteoporosis presents clinically. Both Type I and Type II diabetes mellitus are associated with an increased fracture risk, with increased risk at higher bone density than in the non-diabetic population. This may be due to differences in the chemical composition of bone tissue that is more brittle than normal, a predilection for conversion of precursors to adipose cells rather than osteoblasts, and the sequelae of diabetes that increase the risk of falls and injury.

TABLE 404-2

Diseases Associated with an Increased Risk of Generalized Osteoporosis in Adults

Hypogonadal states	**Hematologic disorders/malignancy**
Turner's syndrome	Multiple myeloma
Klinefelter's syndrome	Lymphoma and leukemia
Anorexia nervosa	Malignancy-associated parathyroid hormone (PTHrP) production
Hypothalamic amenorrhea	Mastocytosis
Hyperprolactinemia	Hemophilia
Other primary or secondary hypogonadal states	Thalassemia
Endocrine disorders	**Selected inherited disorders**
Cushing's syndrome	Osteogenesis imperfecta
Hyperparathyroidism	Marfan's syndrome
Thyrotoxicosis	Hemochromatosis
Diabetes mellitus (both type 1 and 2)	Hypophosphatasia
Acromegaly	Glycogen storage diseases
Adrenal insufficiency	Homocystinuria
Nutritional and gastrointestinal disorders	Ehlers-Danlos syndrome
Malnutrition	Porphyria
Parenteral nutrition	Menkes' syndrome
Malabsorption syndromes	Epidermolysis bullosa
Gastrectomy	**Other disorders**
Severe liver disease, especially biliary cirrhosis	Immobilization
Pernicious anemia	Chronic obstructive pulmonary disease
Rheumatologic disorders	Pregnancy and lactation
Rheumatoid arthritis	Scoliosis
Ankylosing spondylitis	Multiple sclerosis
	Sarcoidosis
	Amyloidosis

MEDICATIONS

A large number of medications used in clinical practice have potentially detrimental effects on the skeleton **(Table 404-3)**. *Glucocorticoids* are the most common cause of medication-induced osteoporosis. It is often not possible to determine the extent to which osteoporosis is related to glucocorticoid or to other factors, as the effects of medication are superimposed on the effects of the primary disease, which in itself may be associated with bone loss (e.g., rheumatoid arthritis). Excessive doses of thyroid hormone can accelerate bone remodeling and result in bone loss.

TABLE 404-3

Drugs Associated with an Increased Risk of Generalized Osteoporosis in Adults

Glucocorticoids	Excessive thyroxine
Cyclosporine	Aluminum
Cytotoxic drugs	Gonadotropin-releasing hormone agonists
Anticonvulsants	Heparin
Aromatase inhibitors	Lithium
SSRIs	Protein Pump Inhibitors
	Thiazolidinediones
	Androgen Deprivation Therapies

Other medications have less detrimental effects on the skeleton than pharmacologic doses of glucocorticoids. *Anticonvulsants* are thought to increase the risk of osteoporosis, although many affected individuals have concomitant insufficiency of 1,25(OH)2D, as some anticonvulsants induce the cytochrome P450 system and vitamin D metabolism. Patients undergoing transplantation are at high risk for rapid bone loss and fracture not only from glucocorticoids but also from treatment with other *immunosuppressants* such as cyclosporine and tacrolimus (FK506). In addition these, patients often have underlying metabolic abnormalities such as hepatic or renal failure that predispose to bone loss. Recently the use of proton pump inhibitors has been shown in observational studies to be associated with a higher risk of fracture. Given their widespread and frequent longterm use, the skeletal effect is important from a public health perspective and when reviewing risk for fracture in individuals.

Aromatase inhibitors, which potently block the aromatase enzyme that converts androgens and other adrenal precursors to estrogen, reduce circulating postmenopausal estrogen supply dramatically. These agents, which are used in various stages for breast cancer treatment, also have been shown to have a detrimental effect on bone density and risk of fracture. Androgen deprivation therapies, used to treat men with prostate cancer, also result in rapid loss of bone and increased fracture risk. Various diabetes medications, including but not limited to thiazoledinediones, and antidepressants, including the Selective Serotonin Reuptake Inhibitor (SSRIs) increase risk of osteoporosis and fracture. It is difficult in some cases to separate the risk accrued by the underlying disease from that attributable the medication. Thus, both depression and diabetes are risk factors for fracture by themselves.

SMOKING

Smoking produces detrimental effects on bone mass mediated directly by toxic effects on osteoblasts or indirectly by modifying estrogen metabolism. On average, cigarette smokers reach menopause 1–2 years earlier than the general population. Cigarette smoking also produces secondary effects that can modulate skeletal status, including intercurrent respiratory and other illnesses, frailty, decreased exercise, poor nutrition, and the need for additional medications (e.g., glucocorticoids for lung disease).

DIAGNOSIS

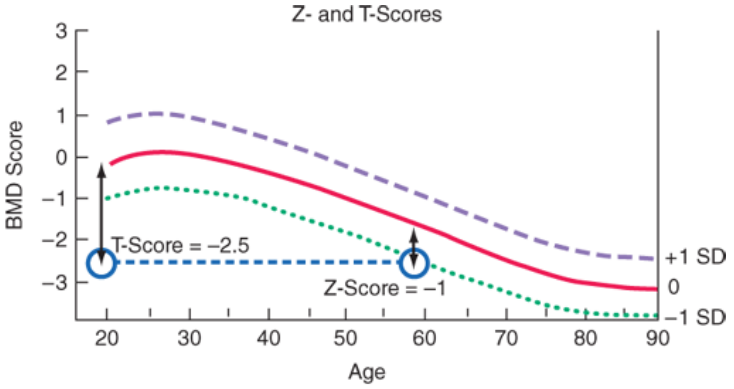
MEASUREMENT OF BONE MASS

Several noninvasive techniques are available for estimating skeletal mass or BMD. They include DXA, single-energy x-ray absorptiometry (SXA), quantitative CT, and ultrasound (US). DXA is a highly accurate x-ray technique that has become the standard for measuring bone density. Though it can be used for measurement in any skeletal site, clinical determinations usually are made of the lumbar spine and hip. DXA also can be used to measure BMD of the wrist and body composition. In the DXA technique, two x-ray energies are used to estimate the area of mineralized tissue, and the mineral content is divided by bone area, which partially corrects for body and bone size. However, this correction is only partial since DXA is a two-dimensional scanning technique and cannot estimate the depth or posteroanterior length of the bone. Thus, small slim people tend to have lower than average BMD, a feature that is important in interpreting BMD measurements. Bone spurs, which are common in osteoarthritis, tend to falsely increase bone density mostly of the spine and are a particular problem in measuring spine BMD in older individuals. Because DXA measurement devices are provided by several different manufacturers, the

output varies in absolute terms. Consequently, it has become standard practice to relate the results to "normal" values by using T-scores (a T-score of 1 equals 1 SD), which compare individual results to those in a young population that is matched for race and sex, with the average value given a score of zero and the range being +2.5 to –2.5 (i.e., 2.5 SDs above or below the mean). Z-scores (also SDs) compare individual results to those of an age and gender-matched reference population. Thus, a 60-year-old woman with a Z-score of –1 (1 SD below mean for age) has a T-score of –2.5 (2.5 SD below mean for a young control group) (Fig. 404-6). A T-score <–2.5 in the lumbar spine, femoral neck, or total hip has been defined as osteoporosis.

FIGURE 404-6

Relationship between Z-scores and T-scores in a 60-year-old woman. BMD, bone mineral density; SD, standard deviation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

As noted above, since >50% of fractures occur in individuals with low bone mass (i.e., a T-score between –1.0 and –2.5), rather than osteoporosis, attempts are ongoing to redefine the disease as a fracture risk rather than a specific BMD. To that end the absolute fracture risk assessment tool FRAX (Fracture Risk Assessment) often accompanies the report of bone density. FRAX estimates include age, gender, height, weight, fracture history, hip fracture in a parent, steroid use, rheumatoid arthritis, other secondary causes as well as bone density of the femoral neck. The program then calculates the estimated risk over a ten-year time frame for major osteoporosis-related fractures (clinical spine, hip, wrist and proximal humerus) as well as hip fracture.

CT can also be used to measure the spine and the hip, but is rarely used clinically, in part because the radiation exposure and cost are both much higher than with DXA. High resolution peripheral quantitative computed tomography (QCT) can be used to measure bone in the forearm or tibia, and is a research tool that provides information on skeletal architecture non-invasively. Magnetic resonance imaging can also be used to obtain some architectural information on the forearm and perhaps the hip, but again is primarily a research tool at present.

Ultrasound can be used to measure bone mass by calculating the attenuation of the signal as it passes through bone or the speed with which it traverses the bone. Although the ultrasound technique was purported to assesses properties of bone other than mass (e.g., quality), this has not been confirmed. Because of its relatively low cost and mobility, ultrasound bone density measurement is amenable for use as a screening procedure in stores or health fairs.

All these techniques for measuring BMD have been approved by the U.S. Food and Drug Administration (FDA) on the basis of their capacity to predict fracture risk. The hip is the preferred site of measurement in most individuals, since it predicts the risk of hip fracture, the most important consequence of osteoporosis, better than any other bone density measurement site. When hip measurements are performed by DXA, the spine can be measured at the same time. In younger individuals such as perimenopausal or early postmenopausal women, spine measurements may be the most sensitive indicator of bone loss. When the spine or hip is not measureable due to severe degenerative spine disease or scoliosis or prior spine or hip surgery, BMD of the wrist is often measured.

INDICATIONS FOR BONE MASS MEASUREMENT

Clinical guidelines have been developed for the use of bone densitometry in clinical practice (Table 404-4). The National Osteoporosis Foundation (NOF) guidelines recommend bone mass measurements in postmenopausal women, assuming they have one or more risk factors for osteoporosis in addition to age, sex, and estrogen deficiency. The guidelines further recommend that bone mass measurement be considered in *all* women by age 65, a position ratified by the U.S. Preventive Health Services Task Force. In males the use of bone density determination is not recommended until the age of 70 years in the absence of multiple risk factors or the occurrence of an osteoporosis-related fracture.

TABLE 404-4

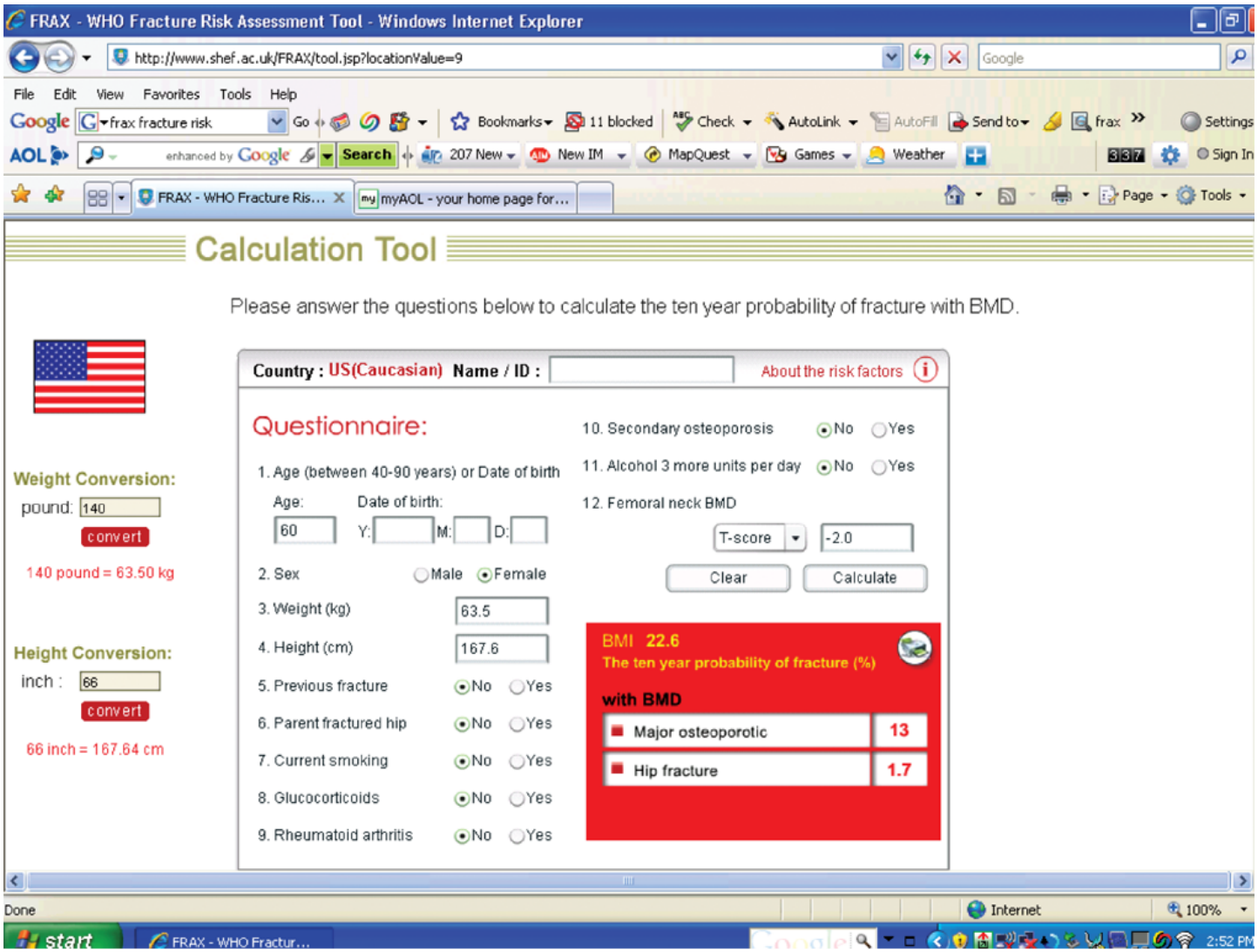
Indications for BMD Testing

- Women aged ≥65 and men aged ≥70; regardless of clinical risk factors
- Younger postmenopausal women, women in the menopausal transition, and men aged from 50 to 69 with clinical risk factors for fracture
- Adults who have a fracture at or after age 50
- Adults with a condition (e.g., rheumatoid arthritis) or taking a medication (e.g., glucocorticoids at a daily dose >5 mg prednisone or equivalent for >3 months associated with low bone mass or bone loss

Risk factors (age, prior fracture, family history of hip fracture, low body weight, cigarette consumption, excessive alcohol use, steroid use, and rheumatoid arthritis) can be combined with BMD to assess the 10-year fracture probabilities. Fracture risk probability calculators are available as part of the report from many DXA machines and also available online (*https://www.sheffield.ac.uk/FRAX/*) **(Fig. 404-7)**. In the United States it has been determined to be cost effective to treat if the 10-year fracture risk from FRAX is ≥20%, and/or the 10-year risk of hip fracture is ≥3%. FRAX is an imperfect tool, as it does not include any assessment of fall risk, and secondary causes are excluded when BMD is entered. More importantly, it does not distinguish the contribution toward of future fracture probability from an acute recent fracture versus the much lesser importance of the more remote fracture. Moreover, there is no mandate for vertebral fracture diagnosis and no additional fracture probability estimated for patients who have had multiple fractures. Nonetheless it is useful as an educational tool for patients, particularly for those who are excessively worried about BMD levels despite relative youth and health.

FIGURE 404-7

FRAX calculation tool. When the answers to the indicated questions are filled in, the calculator can be used to assess the 10-year probability of fracture. The calculator (available online at *http://www.shef.ac.uk/FRAX/tool.jsp?locationValue=9*) also can risk adjust for various ethnic groups.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

VERTEBRAL IMAGING

DXA equipment can also be used to obtain lateral images of the thoracic and lumbar spine, a technique called vertebral fracture assessment (VFA). While not as definitive as a radiograph it is an excellent screening tool for both women and men based on age and BMD even in the absence of any specific symptoms since the majority of vertebral fractures are asymptomatic for a long time. Furthermore, the VFA can be used to evaluate height loss or back pain that suggest the presence of an undiagnosed vertebral fracture.

Because vertebral fractures are often asymptomatic when they first occur, the diagnosis of vertebral fracture is rarely made at the time. Since vertebral fractures, whether symptomatic or asymptomatic, are associated with the same clinical sequelae, it is critical that patients with these fractures are identified. Vertebral fracture prevalence in the US based on the National Health and Nutrition Evaluation Studies (NHANES) population appears to be about 10% in the 1970s and 20% in the 1980s, when the strictest criteria for diagnosis are utilized. The NOF and other organizations have recommended that women by the age of 65 and men by the age of 70 undergo vertebral imaging if a T-Score is ≤–1.5 at the spine, hip, or femoral neck. For women by the age of 70 and men by the age of 80 if a T-Score is <–1.0. For younger individuals, vertebral imaging is recommended for those with an osteoporosis related fracture, height loss, or glucocorticoid use. (See **Table 404-5**.)

TABLE 404-5

Indications for Vertebral Testing

Consider vertebral imaging tests for the following individuals[a]

- All women aged ≥70 and all men aged ≥80 if BMD T-score at the spine, total hip, or femoral neck is <1.0
- Women aged from 65 to 69 and men aged from 70 to 79 if BMD T-score at the spine, total hip, or femoral neck is <1.5
- Postmenopausal women and men aged ≥50 with specific risk factors:
  - Low-trauma fracture during adulthood (aged ≥50)
  - Historical height loss of ≥1.5 in. (4 cm)[b]
  - Prospective height loss of ≥0.8 in. (2 cm)[c]
  - Recent or ongoing long-term glucocorticoid treatment

[a]If bone density testing is not available, vertebral imaging may be considered based on age alone. [b]Current height compared to peak height during childhood. [c]Cumulative height loss measured during interval medical assessment.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Osteoporosis

The perimenopausal transition is a good opportunity to initiate a discussion about risk factors for osteoporosis and consideration of indications for a BMD test. A careful history and physical examination should be performed to identify risk factors for osteoporosis. A low Z-score increases the suspicion of a secondary cause for bone loss unless the low value can be explained by body size. Height loss >2.5–3.8 cm (>1–1.5 in.) is an indication for VFA by DXA or radiography to rule out asymptomatic vertebral fractures, as is the presence of significant kyphosis or back pain, particularly if it began after menopause. For patients who present with fractures, it is important to ensure that the fractures are not caused by an underlying malignancy. Usually this is clear on routine radiography, but on occasion, CT, MRI, or radionuclide scans may be necessary. In this regard it is important not to dismiss fractures simply because they happened on significant trauma. Persons with osteoporosis fracture more readily with any level of injury, a concept that needs continual emphasis.

ROUTINE LABORATORY EVALUATION

There is no established algorithm for the evaluation of women who present with osteoporosis. A general evaluation that includes complete blood count, serum and 24-h urine calcium, and renal and hepatic function tests is useful for identifying selected secondary causes of low bone mass, particularly for women with fractures or unexpectedly low Z-scores. An elevated serum calcium level suggests hyperparathyroidism or malignancy, whereas a reduced serum calcium level may reflect malnutrition or a malabsorption disease such as celiac disease. In the presence of hypercalcemia, a serum PTH level differentiates between hyperparathyroidism (PTH↑) and malignancy (PTH↓), and a high PTHrP level can help document the presence of humoral hypercalcemia of malignancy **(Chap. 403)**. A low urine calcium (<50 mg/24 h) suggests malnutrition, or malabsorption; a high urine calcium (>300 mg/24 h) during normal calcium intake (excluding calcium supplements for at least a week before the urine collection) is indicative of hypercalciuria. Hypercalciuria occurs primarily in three situations: (1) a renal calcium leak, which is more common in males with osteoporosis; (2) absorptive hypercalciuria, which can be idiopathic or associated with increased $1,25(OH)_2D$ in granulomatous disease; or (3) hematologic malignancies or conditions associated with excessive bone turnover such as Paget's disease, hyperparathyroidism, and hyperthyroidism. Renal hypercalciuria is treated with thiazide diuretics, which lower

urine calcium and help improve calcium economy. In this setting, thiazides alone can improve bone mass and possibly reduce risk of fracture. They might also reduce renal stone risk.

Individuals who have osteoporosis-related fractures or bone density in the osteoporotic range should have a measurement of serum 25(OH)D level, since the intake of vitamin D required to achieve a target level >30 ng/mL is highly variable. Hyperthyroidism should be evaluated by measuring thyroid-stimulating hormone (TSH).

When there is clinical suspicion of Cushing's syndrome, urinary free cortisol levels or a fasting serum cortisol should be measured after overnight dexamethasone. When bowel disease, malabsorption, or malnutrition is suspected, serum albumin, cholesterol, and a complete blood count should be checked. Asymptomatic malabsorption may be heralded by anemia (macrocytic—vitamin B_{12} or folate deficiency; microcytic—iron deficiency) or low serum cholesterol or urinary calcium levels. If these or other features suggest malabsorption, further evaluation is required. Asymptomatic celiac disease with selective malabsorption is being found with increasing frequency; the diagnosis can be made by testing for transglutaminase IgA antibodies, but may require confirmation by endoscopic biopsy. A trial of a gluten-free diet can also be confirmatory **(Chap. 318)**. When osteoporosis is found associated with symptoms of rash, multiple allergies, diarrhea, or flushing, mastocytosis should be considered and excluded by using 24-h urine histamine collection or serum tryptase.

Myeloma can masquerade as generalized osteoporosis, although it more commonly presents with bone pain and characteristic "punched-out" lesions on radiography. Serum and urine electrophoresis and/or evaluation for serum free light chains light chains in urine are required to exclude this diagnosis. More commonly a monoclonal gammopathy (MGUS) is found and the patient subsequently monitored to ensure that this is not an incipient myeloma. MGUS itself may be associated with an increased risk of osteoporosis. A bone marrow biopsy may be required to rule out myeloma (in patients with equivocal electrophoretic results) and also can be used to exclude mastocytosis, leukemia, and other marrow infiltrative disorders such as Gaucher's disease.

BONE BIOPSY

Tetracycline labeling of the skeleton allows determination of the rate of remodeling as well as evaluation for other metabolic bone diseases. The current use of BMD tests, in combination with hormonal evaluation and biochemical markers of bone remodeling, has largely replaced the clinical use of bone biopsy, although it remains an important tool in the diagnosis of chronic kidney disease mineral bone disease (CKDMBD), in evaluating mechanism of action of osteoporosis pharmacologies and in clinical research.

BIOCHEMICAL MARKERS

Several biochemical tests are available that provide an index of the overall rate of bone remodeling **(Table 404-6)**. Biochemical markers usually are characterized as those related primarily to *bone formation* or *bone resorption*. These tests measure the overall state of bone remodeling at a single point in time. Clinical use of these tests has been hampered by biologic variability (in part related to circadian rhythm) as well as analytic variability, although the latter is improving.

For the most part, remodeling markers do not predict rates of bone loss well enough in individuals to make accurate assessment of potential future changes in bone density. However, they do provide adjunct information that assists in both evaluation of the patient and in assessment of treatment response. Markers of bone resorption may help in the prediction of fracture risk, independently of bone density, particularly in older individuals. In women ≥65 years, when bone density results are greater than the usual treatment thresholds noted above, a high level of bone resorption should prompt consideration of treatment. The primary use of biochemical markers is for monitoring the response to treatment. With the introduction of antiresorptive therapeutic agents, bone remodeling declines rapidly, with the fall in resorption occurring earlier than the fall in formation. Inhibition of bone resorption is maximal within 3 months or so. Thus, measurement of bone resorption (CTX is the preferred marker) before initiating therapy and 3–6 months after starting therapy provides an earlier estimate of patient response than does bone densitometry. A decline in resorptive markers can be ascertained after treatment with bisphosphonates, denosumab or estrogen; this effect is less marked after treatment with weaker agents such as raloxifene or calcitonin. Bone turnover markers are also useful in monitoring the effects of 1-34hPTH, or teriparatide, which rapidly increases bone formation (P1NP is the most sensitive but osteocalcin is also a very good formation marker) and later bone resorption. The recent suggestion of "drug holidays" (see below) has opened another use for biochemical markers, allowing evaluation of the off-effect of drugs such as bisphosphonates.

TABLE 404-6

Biochemical Markers of Bone Metabolism in Clinical Use

Bone formation
Serum bone-specific alkaline phosphatase
Serum osteocalcin
Serum propeptide of type I procollagen
Bone resorption
Urine and serum cross-linked N-telopeptide
Urine and serum cross-linked C-telopeptide

TREATMENT

TREATMENT

Osteoporosis

MANAGEMENT OF PATIENTS WITH FRACTURES

Treatment of a patient with osteoporosis frequently involves management of acute fractures as well as treatment of the underlying disease. Hip fractures almost always require surgical repair if the patient is to become ambulatory again. Depending on the location and severity of the fracture, condition of the neighboring joint, and general status of the patient, procedures may include open reduction and internal fixation with pins and plates, hemiarthroplasties, and total arthroplasties. These surgical procedures are followed by intense rehabilitation in an attempt to return patients to their pre-fracture functional level. Long bone fractures often require either external or internal fixation. Other fractures (e.g., vertebral, rib, and pelvic fractures) usually are managed with supportive care, requiring no specific orthopedic treatment.

Only ~25–30% of vertebral compression fractures present with sudden-onset back pain. For acutely symptomatic fractures, treatment with analgesics is required, including nonsteroidal anti-inflammatory agents and/or acetaminophen, sometimes with the addition of a narcotic agent (codeine or oxycodone). (A few small, randomized clinical trials suggest that calcitonin may reduce pain related to acute vertebral compression fracture). A technique that involves percutaneous injection of artificial cement (polymethylmethacrylate) into the vertebral body (vertebroplasty or kyphoplasty), may offer significant pain relief in some patients, however controlled trials of these procedures have provided some doubt of their efficacy. Furthermore, risks include acute extravasation of cement outside of the vertebral body with neurologic impairment and possibly an increased risk of vertebral fracture in adjacent vertebrae due to increased rigidity of the treated vertebral body. Short periods of bed rest may be helpful for pain management, but in general, early mobilization is recommended as it helps prevent further bone loss associated with immobilization. Occasionally, use of a soft elastic-style brace may facilitate earlier mobilization. Muscle spasms often occur with acute compression fractures and can be treated with muscle relaxants and heat treatments. Severe pain usually resolves within 6–10 weeks. More chronic severe pain might suggest the possibility of multiple myeloma or other underlying conditions.

Vertebral fractures cause height loss because of the loss of vertebral body height during compression of the vertebral body. These fractures can produce kyphotic posture, particularly when wedge shaped or just loss of thoracic height. Chronic pain following vertebral fracture is probably not bony in origin; instead, it is related to abnormal strain on muscles, ligaments, and tendons and to secondary facet-joint arthritis associated with alterations in thoracic and/or abdominal shape. Chronic pain may also be the result of ribs sitting right on top of the iliac crest bones, particularly in patients who have had multiple vertebral compression fractures. Chronic pain is difficult to treat effectively and may require analgesics, sometimes including narcotic analgesics. Frequent intermittent rest in a supine or semireclining position is often required to allow the soft tissues, which are under tension, to relax. Back and core-strengthening exercises may be beneficial. Heat treatments help relax muscles and reduce the muscular component of discomfort. Various physical modalities, such as ultrasound and transcutaneous nerve stimulation, may be beneficial in some patients. Pain also occurs in the neck region, not as a result of compression fractures (which almost never occur in the cervical spine as a result of osteoporosis) but because of chronic strain associated with trying to elevate the head in a person with a significant thoracic kyphosis.

Multiple vertebral fractures often are associated with psychological symptoms; this is not always appreciated. The changes in body configuration and back pain can lead to marked loss of self-image and a secondary depression. Altered balance, precipitated by the kyphosis and the anterior movement of the body's center of gravity, leads to a fear of falling, a consequent tendency to remain indoors, and the onset of social isolation. These symptoms sometimes can be alleviated by family support and/or psychotherapy. Medication may be necessary when depressive features are present.

Multiple studies show that patients presenting with fractures after age 50 years (even fractures traditionally linked to osteoporosis) are largely not screened or treated for osteoporosis. Estimates suggest that fewer than 25% of fracture patients receive follow-up care. Recently several studies have demonstrated the effectiveness of a relatively simple and inexpensive program that reduces the risk of subsequent fractures. In the Kaiser system it is estimated that a 20% decline in hip fracture occurrence was seen with the introduction of a fracture liaison service. This involves a health care professional (usually a nurse or physician's assistant) whose job is to educate patients, and coordinate evaluation and osteoporosis treatment as patients move from an emergency room, acute care hospital, rehabilitation hospital and/or orthopedic practice to outpatient management. If the Kaiser experience can be repeated, not only would there be significant savings of health care dollars, but also a dramatic drop in hip fracture incidence and a marked improvement in morbidity and mortality among the aging population.

MANAGEMENT OF THE UNDERLYING DISEASE

Risk Factor Reduction

After risk assessment patients should be thoroughly educated to reduce the impact of modifiable risk factors associated with bone loss and falling. Medications should be reviewed to ensure that all are necessary and taken at the lowest required dose. Glucocorticoid medication, if present, should be evaluated to determine that it is truly indicated and is being given in doses that are as low as possible. For those on thyroid hormone replacement, TSH testing should be performed to determine that an excessive dose is not being used, as iatrogenic thyrotoxicosis can be associated with increased bone loss. In patients who smoke, efforts should be made to facilitate smoking cessation. Reducing risk factors for falling also include alcohol abuse treatment and a review of the medical regimen for any drugs that might be associated with orthostatic hypotension and/or sedation, including hypnotics and anxiolytics. If nocturia occurs, the frequency should be reduced, if possible (e.g., by decreasing or modifying diuretic use), as arising in the middle of sleep is a common precipitant of a fall. Patients should be instructed about environmental safety with regard to eliminating exposed wires, curtain strings, slippery rugs, and mobile tables. Avoiding stocking feet on wood floors, checking carpet condition (particularly on stairs), and providing good light in paths to bathrooms and outside the home are important preventive measures. Treatment for impaired vision is recommended, particularly a problem with depth perception, which is specifically associated with increased falling risk. Elderly patients with neurologic impairment (e.g., stroke, Parkinson's disease, Alzheimer's disease) are particularly at risk of falling and require specialized supervision and care.

Nutritional Recommendations

Calcium

A large body of data indicates that optimal calcium intake reduces bone loss and suppresses bone turnover. Recommended intakes from an Institute of Medicine report are shown in **Table 404-7**. The National Health and Nutritional Evaluation Studies (NHANES) have consistently documented that average calcium intakes fall considerably short of these recommendations. The preferred source of calcium is diet, but many patients require calcium supplementation to bring intake to about 1200 mg/d. Best sources of calcium include dairy products (milk, yogurt, and cheese), nondaily milks (almond, rice, soy), and fortified foods such as certain cereals, waffles, snacks, juices, and crackers. Some of these fortified foods contain as much calcium per serving as milk. Various vegetables and fruits, such as kale, broccoli, and dried figs contain reasonably high calcium content, though some of it may not be fully bioavailable. Calcium intake calculators are available at *NOF.org* or *NYSOPEP.org* and will give a rough idea of total calcium intake.

If calcium supplements are required, they should be taken in doses sufficient to bring total intake to the required level (1200 mg/d). Doses of supplements should be ≤600 mg per single dose, as the calcium absorption fraction decreases at higher doses. Calcium supplements should be calculated on the basis of the elemental calcium content of the supplement, not the weight of the calcium salt **(Table 404-8)**. Calcium supplements containing carbonate are best taken with food since they require acid for solubility. Calcium citrate supplements can be taken at any time. To confirm bioavailability, calcium supplements can be placed in distilled vinegar. They should dissolve within 30 min.

Several controlled clinical trials of calcium, mostly with accompanying vitamin D, have confirmed reductions in clinical fractures, including fractures of the hip (~20–30% risk reduction), particularly in elderly individuals who are more likely to be dietarily deficient. All recent studies of pharmacologic agents have been conducted in the context of calcium replacement (± vitamin D). Thus, it is standard practice to ensure an adequate calcium and vitamin D intake in patients with osteoporosis whether they are receiving additional pharmacologic therapy or not. A systematic review confirmed a greater BMD response to antiresorptive therapy when calcium intake was adequate.

Although side effects from supplemental calcium are minimal (eructation and constipation mostly with carbonate salts), individuals with a history of kidney stones should have a 24-h urine calcium determination before starting increased calcium to avoid exacerbating hypercalciuria. A recent analysis of published data has suggested that high intakes of calcium particularly from supplements are associated with an increase in the risk of heart disease. This is an evolving story with data both confirming and refuting the finding. Since high calcium intakes also increase the risk of renal stones and confer no extra benefit to the skeleton, the recommendation that total intakes should be between 1000 and 1500 mg/d seems reasonable.

Vitamin D

Diet alone rarely contains sufficient vitamin D to maintain target circulating levels [serum 25(OH)D consistently >75 μmol/L (30 ng/mL) Vitamin D is synthesized from a precursor in the skin under the influence of heat and ultraviolet light **(Chap. 402)** but production is blocked by sunscreen and sun avoidance. Therefore, large segments of the population do not obtain sufficient vitamin D from either skin production or dietary sources. Since vitamin D supplementation at doses that would achieve these serum levels is safe and inexpensive, the National Academy of Medicine (formerly, Institute of Medicine, IOM) recommends daily intakes of 200 IU for adults <50 years of age, 400 IU for those 50–70 years, and 600 IU for those >70 years (based on obtaining a serum level of 20 ng/mL, lower than the level recommended by most other guidelines). Multivitamin tablets usually contain 400 IU, and many

calcium supplements also contain vitamin D. Some data suggest that higher doses (≥1000 IU) may be required in the elderly and chronically ill. The IOM report suggests that it is safe to take up to 4000 IU/d. For those with osteoporosis or those at risk of osteoporosis 1000–2000IU/day can usually maintain serum 25(OH)D above 30 ng/mL.

Other Nutrients

Other nutrients such as salt, high animal protein intakes, and caffeine may have modest effects on calcium excretion or absorption. Adequate vitamin K status is required for optimal carboxylation of osteocalcin. States in which vitamin K nutrition or metabolism is impaired, such as with long-term warfarin therapy, have been associated with reduced bone mass. Research concerning cola intake is controversial but suggests a possible link to reduced bone mass through factors that are independent of caffeine.

Magnesium is abundant in foods, and magnesium deficiency is quite rare in the absence of a serious chronic disease. Magnesium supplementation may be warranted in patients with inflammatory bowel disease, celiac disease, chemotherapy, severe diarrhea, malnutrition, or alcoholism. Dietary phytoestrogens, which are derived primarily from soy products and legumes (e.g., garbanzo beans [chickpeas] and lentils), exert some estrogenic activity but are insufficiently potent to justify their use in place of a pharmacologic agent in the treatment of osteoporosis.

Patients with hip fractures are often frail and relatively malnourished. Some data suggest an improved outcome in such patients when they are provided calorie and protein supplementation. Excessive protein intake can increase renal calcium excretion, but this can be corrected by an adequate calcium intake.

Exercise

Exercise in young individuals increases the likelihood that they will attain the maximal genetically determined peak bone mass. Meta-analyses of studies performed in postmenopausal women indicate that weight-bearing exercise helps prevent bone loss but does not appear to result in substantial gain of bone mass. This beneficial effect wanes if exercise is discontinued. Most of the studies are short term, and a more substantial effect on bone mass is likely if exercise is continued over a long period. Exercise also has beneficial effects on neuromuscular function, and it improves coordination, balance, and strength, thereby reducing the risk of falling. A walking program is a practical way to start. Other activities such as dancing, racquet sports, cross-country skiing, and use of gym equipment, are also recommended, depending on the patient's personal preference and general condition. Even women who cannot walk benefit from swimming or water exercises, not so much for the effects on bone, which are quite minimal, but because of effects on muscle. Exercise habits should be consistent, optimally at least three times a week. For most patients we suggest participation in exercise regimes that the patient enjoys, in order to improve adherence. We also emphasize the importance of making exercise a social activity, again to improve adherence.

PHARMACOLOGIC TREATMENT OF OSTEOPOROSIS

Patients presenting with typical osteoporosis related fractures (certainly hip and spine), in the setting of a BMD in the low bone mass or osteoporosis range should be treated with pharmacologic agents. Most guidelines also suggest that patients be considered for treatment when BMD T-Score is ≤–2.5, a level consistent with the diagnosis of osteoporosis. Treatment also should also be considered in postmenopausal women with fracture or multiple risk factors even if BMD is not in the osteoporosis range. Treatment thresholds depend on cost-effectiveness analyses but in the United States are >20% for 10-year major fracture probability and >3% 10-year hip fracture probability. It must be emphasized, however, that as with other diseases, risk assessment is an inexact science when applied to individual patients. Fractures are chance occurrences that can happen to anyone and do! Patients often accept risks that are higher than the physician might like out of concern for the (usually considerably lower) risks of adverse events of drugs.

Pharmacologic therapies for osteoporosis are either antiresorptive or anabolic. The antiresorptive agents include medications that have broad effects such as hormone/estrogen therapy and selective estrogen receptor modulators (SERMS) as well as those agents that are specific for the treatment of osteoporosis (bisphosphonates, denosumab, and calcitonin). The only currently approved anabolic agent is teriparatide, but two additional anabolic agents are currently under FDA review for treatment of osteoporosis (abaloparatide and romosozumab).

Antiresorptive Agents

Estrogens

A large body of clinical trial data indicates that various types of estrogens (conjugated equine estrogens, estradiol, estrone, esterified estrogens, ethinyl estradiol, and mestranol) reduce bone turnover, prevent bone loss, and induce small increases in bone mass of the spine, hip, and total body. The effects of estrogen are seen in women with natural or surgical menopause and in late postmenopausal women with or without established osteoporosis. Estrogens are efficacious when administered orally or transdermally. For both oral and transdermal routes of administration, combined estrogen/progestin preparations are now available in many countries, obviating the problem of taking two tablets or using a patch and oral progestin.

For oral estrogens, the standard recommended doses have been 0.3 mg/d for esterified estrogens, 0.625 mg/d for conjugated equine estrogens, and 5 µg/d for ethinyl estradiol. For transdermal estrogen, the commonly used dose supplies 50 µg estradiol per day, but a lower dose may be appropriate for some individuals. Dose-response data for conjugated equine estrogens indicate that lower doses (0.3 and 0.45 mg/d) are effective. Doses even lower have also been shown to slow bone loss.

Fracture Data

Epidemiologic databases indicate that women who take estrogen replacement have a 50% reduction, on average, of osteoporosis related fractures, including hip fractures. The beneficial effect of estrogen is greatest among those who start replacement early and continue the treatment; the benefit

declines after discontinuation to the extent that there is no residual protective effect against fracture by 10 years after discontinuation. The first clinical trial evaluating fractures as secondary outcomes, the Heart and Estrogen-progestin Replacement Study (HERS) trial, showed no effect of hormone therapy on hip or other clinical fractures in women with established coronary artery disease. These data made the results of the Women's Health Initiative (WHI) exceedingly important **(Chap. 388)**. The estrogen-progestin arm of the WHI in >16,000 postmenopausal healthy women indicated that hormone therapy reduces the risk of hip and clinical spine fracture by 34% and that of all clinical fractures by 24%.

A few smaller clinical trials have evaluated spine fracture occurrence as an outcome with estrogen therapy. They have consistently shown that estrogen treatment reduces the incidence of vertebral compression fracture.

The WHI has provided a vast amount of data on the multisystemic effects of hormone therapy. Although earlier observational studies suggested that estrogen replacement might reduce heart disease, the WHI showed that combined estrogen-progestin treatment increased risk of fatal and nonfatal myocardial infarction by ~29%, confirming data from the HERS study. Other important relative risks included a 40% increase in stroke, a 100% increase in venous thromboembolic disease, and a 26% increase in risk of breast cancer. Subsequent analyses have confirmed the increased risk of stroke and in a substudy showed a twofold increase in dementia. Benefits other than the fracture reductions noted above included a 37% reduction in the risk of colon cancer. These relative risks have to be interpreted in light of absolute risk **(Fig. 404-8)**. For example, out of 10,000 women treated with estrogen-progestin for 1 year, there will be 8 excess heart attacks, 8 excess breast cancers, 18 excess venous thromboembolic events, 5 fewer hip fractures, 44 fewer clinical fractures, and 6 fewer colorectal cancers. These numbers must be multiplied by years of hormone treatment. There was no effect of hormone treatment on the risk of uterine cancer or total mortality.

It is important to note that these WHI findings apply specifically to hormone treatment in the form of conjugated equine estrogen plus medroxyprogesterone acetate. The relative benefits and risks of unopposed estrogen in women who had hysterectomies vary somewhat. They still show benefits against fracture occurrence and increased risk of venous thrombosis and stroke, similar in magnitude to the risks for combined hormone therapy. In contrast, though, the estrogen-only arm of WHI indicated no increased risk of heart attack or breast cancer. The data suggest that at least some of the detrimental effects of combined therapy are related to the progestin component. In addition, there is the possibility, suggested by primate data that the risk accrues mainly to women who have some years of estrogen deficiency before initiating treatment. Nonetheless there is marked reluctance among women for ET/HT and the US preventive services task force has specifically suggested that ET/HT not be used for disease prevention.

Mode of Action

Two subtypes of ERs, α and β, have been identified in bone and other tissues. Cells of monocyte lineage express both ERα and ERβ, as do osteoblasts. Estrogen-mediated effects vary with the receptor type. Using ER knockout mouse models, elimination of ERα produces a modest reduction in bone mass, whereas mutation of ERβ has less of an effect on bone. A male patient with a homozygous mutation of ERα had markedly decreased bone density as well as abnormalities in epiphyseal closure, confirming the important role of ERα in bone biology. The mechanism of estrogen action in bone is an area of active investigation (Fig. 404-5). Although data are conflicting, estrogens may inhibit osteoclasts directly. However, the majority of estrogen (and androgen) effects on bone resorption are mediated indirectly through paracrine factors produced by osteoblasts. These actions include (1) increasing OPG production by osteoblasts (2) Increasing IGF-I and TGF-β and (3) suppressing IL-1 (α and β), IL-6, TNF-α, and osteocalcin synthesis. The indirect estrogen actions primarily decrease bone resorption.

Progestins

In women with a uterus, daily progestin or cyclical progestins at least 12 days per month are prescribed in combination with estrogens to reduce the risk of uterine cancer. Medroxyprogesterone acetate and norethindrone acetate blunt the high-density lipoprotein response to estrogen, but micronized progesterone does not. Neither medroxyprogesterone acetate nor micronized progesterone appears to have an independent effect on bone; at lower doses of estrogen, norethindrone acetate may have an additive benefit. On breast tissue, progestins may account for the increase the risk of breast cancer with combination treatment.

SERMs

Two SERMs are used currently in postmenopausal women: raloxifene, which is FDA-approved for the prevention and treatment of osteoporosis as well as the prevention of breast cancer, and tamoxifen, which is approved for the prevention and treatment of breast cancer. A third SERM, bazodoxiphene, is marketed in combination with conjugated estrogen for treatment of menopausal symptoms and prevention of bone loss. Bazedoxifene protects the uterus and breast from effects of estrogen and makes the use of progestin unnecessary.

Tamoxifen reduces bone turnover and bone loss in postmenopausal women compared with placebo groups. These findings support the concept that tamoxifen acts as an estrogenic agent in bone. There are limited data on the effect of tamoxifen on fracture risk, but the Breast Cancer Prevention study indicated a possible reduction in clinical vertebral, hip, and Colles' fractures. Tamoxifen is not FDA approved for prevention or treatment of osteoporosis. The major benefit of tamoxifen is on breast cancer occurrence. The breast cancer prevention trial indicated that tamoxifen administration over 4–5 years reduced the incidence of new invasive and noninvasive breast cancer by ~45% in women at increased risk of breast cancer. The incidence of ER-positive breast cancers was reduced by 65%. Tamoxifen increases the risk of uterine cancer in postmenopausal women, limiting its use for breast cancer prevention in women at low or moderate risk.

Raloxifene (60 mg/d) has effects on bone turnover and bone mass that are very similar to those of tamoxifen, indicating that this agent is also estrogenic on the skeleton. The effect of raloxifene on bone density (+1.4–2.8% versus placebo in the spine, hip, and total body) is somewhat less than that seen with

standard doses of estrogens. Raloxifene reduces the occurrence of vertebral fracture by 30–50%, depending on the population; however, there are no data confirming that raloxifene can reduce the risk of nonvertebral fractures >8 years of observation.

Raloxifene, like tamoxifen and estrogen, has effects in other organ systems. The most beneficial effect appears to be a reduction in invasive breast cancer (mainly decreased ER-positive) occurrence of ~65% in women who take raloxifene compared to placebo. In a head-to-head study raloxifene was as effective as tamoxifen in preventing breast cancer in high-risk women and raloxifene is FDA approved for this indication. In a further study raloxifene had no effect on heart disease in women with increased risk for this outcome. In contrast to tamoxifen, raloxifene is not associated with an increase in the risk of uterine cancer or benign uterine disease. Raloxifene increases the occurrence of hot flashes but reduces serum total and low-density lipoprotein cholesterol, lipoprotein(a), and fibrinogen. Raloxifene with positive effects on breast cancer and vertebral fractures has become a useful agent for the treatment of the younger asymptomatic postmenopausal woman. In some women, a recurrence of menopausal symptoms may occur. Usually this is evanescent but occasionally is sufficiently impactful on daily life and sleep, that the drug must be withdrawn. Raloxifene increases the risk of deep vein thrombosis and may increase the risk of death from stroke among older women. Consequently it is not usually recommended for women aged >70 years.

Mode of Action of SERMs

All SERMs bind to the ER, but each agent produces a unique receptor-drug conformation. As a result, specific coactivator or co-repressor proteins are bound to the receptor **(Chap. 370)**, resulting in differential effects on gene transcription that vary depending on other transcription factors present in the cell. Another aspect of selectivity is the affinity of each SERM for the different ERα and ERβ subtypes, which are expressed differentially in various tissues. These tissue-selective effects of SERMs offer the possibility of tailoring estrogen therapy to best meet the needs and risk factor profile of an individual patient.

Bisphosphonates

Alendronate, risedronate, ibandronate, and zoledronic acid are approved for the prevention and treatment of postmenopausal osteoporosis. Alendronate, risedronate, and zoledronic acid are also approved for the treatment of steroid-induced osteoporosis, and risedronate and zoledronic acid are approved for prevention of steroid-induced osteoporosis. Alendronate, risedronate, and zoledronic acid are also approved for treatment of osteoporosis in men.

Alendronate decreases bone turnover and increases bone mass in the spine by up to 8% versus placebo and by 6% versus placebo in the hip. Multiple trials have evaluated its effect on fracture occurrence. The Fracture Intervention Trial provided evidence in >2000 women with prevalent vertebral fractures that daily alendronate treatment (5 mg/d for 2 years and 10 mg/d for 9 months afterward) reduces vertebral fracture risk by about 50%, multiple vertebral fractures by up to 90%, and hip fractures by up to 50%. Several subsequent trials have confirmed these findings **(Figs. 404-9 and 404-10)**. For example, in a study of >1900 women with low bone mass treated with alendronate (10 mg/d) versus placebo, the incidence of all nonvertebral fractures was reduced by ~47% after only 1 year. In the United States the 10mg dose is approved for treatment of osteoporosis and 5 mg/d for prevention.

Trials comparing once-weekly alendronate, 70 mg, with daily 10-mg dosing have shown equivalence with regard to bone mass and bone turnover responses. Consequently, once-weekly therapy generally is preferred because of the low incidence of gastrointestinal side effects and ease of administration. Alendronate should be taken with a full glass of water before breakfast after an overnight fast, as bisphosphonates are poorly absorbed. Because of the potential for esophageal irritation, alendronate is contraindicated in patients who have stricture or inadequate emptying of the esophagus. It is recommended that patients remain upright (standing or sitting)t for at least 30 min after taking the medication to avoid esophageal irritation, and that food and fluids (other than water) be avoided for the same duration. In clinical trials, overall gastrointestinal symptomatology was no different with alendronate than with placebo, but all oral bisphosphonates have been associated with esophageal irritation and inflammation.

Risedronate also reduces bone turnover and increases bone mass. Controlled clinical trials have demonstrated 40–50% reduction in vertebral fracture risk over 3 years, accompanied by a 40% reduction in clinical non-spine fractures. The only clinical trial specifically designed to evaluate hip fracture outcome (HIP) indicated that risedronate reduced hip fracture risk in women in their seventies with confirmed osteoporosis by 40%. In contrast, risedronate was not effective at reducing hip fracture occurrence in older women (80+ years) without proven osteoporosis. Studies have shown that 35 mg of risedronate administered once weekly is therapeutically equivalent to 5 mg/d. Patients should take risedronate with a full glass of plain water to facilitate delivery to the stomach and should not lie down for 30 min after taking the drug. (There is also a preparation of risedronate that can be taken with food; it is the only bisphosphonate that has this dosing flexibility.) The incidence of gastrointestinal side effects in trials with risedronate was similar to that of placebo.

Ibandronate is the third amino-bisphosphonate approved in the United States. Ibandronate (2.5 mg/d) has been shown in clinical trials to reduce vertebral fracture risk by ~40% but with no overall effect on non-vertebral fractures. In a post hoc analysis of subjects with a femoral neck T-score of ≤–3, ibandronate reduced the risk of nonvertebral fractures by ~60%. In clinical trials, ibandronate doses of 150 mg/month PO or 3 mg every 3 months IV had greater effects on turnover and bone mass than did 2.5 mg/d. Patients should take oral ibandronate in the same way as other bisphosphonates, but with 1 h elapsing before other food or drink (other than plain water).

Zoledronic acid is a potent bisphosphonate with a unique administration regimen (5 mg by 15 min IV infusion annually). Zoledronic acid data confirm that it is highly effective in fracture risk reduction. In a study of >7000 women followed for 3 years, zoledronic acid 5 mg IV annually) reduced the risk of vertebral fractures by 70%, nonvertebral fractures by 25%, and hip fractures by 40%. These results were associated with less height loss and disability. In the treated population, there was an increased risk of almost 25% of an acute phase reaction in patients with no prior bisphosphonate exposure (fever, myalgias, headache, malaise), but effects were short-lived (2–3 days). Detailed evaluation of all bisphosphonates failed to confirm a risk of atrial fibrillation. Zoledronic acid has also been studied in a placebo controlled trial of women and men within 3 months of an acute hip fracture. The risk of recurrent fracture was reduced by 35% and there was a 28% reduction in mortality.

Common Bisphosphonate Adverse Events

All bisphosphonates have been associated with some musculoskeletal and joint pains of unclear etiology, which are occasionally severe. There is potential for renal toxicity and bisphosphonates are contraindicated in those with estimated GFR <30–35 mL/min. Hypocalcemia can occur.

Recently there has been concern about two potential side effects associated with bisphosphonate use. The first is osteonecrosis of the jaw (ONJ). ONJ usually follows a dental procedure in which bone is exposed (dental extractions and implants). It is presumed that the exposed bone becomes infected and dies. ONJ is more common among cancer victims receiving high doses of bisphosphonates for skeletal metastases. It is rare among persons with osteoporosis on usual doses of bisphosphonates. Oral antibiotic rinses and oral systemic antibiotics may be useful to prevent this rare adverse event if risk is perceived to be particularly high. The second is called atypical femoral fracture. These are unusual fractures that occur in the subtrochanteric femoral region or across the femoral shaft distal to the lesser trochanter. They are often preceded by pain in the lateral thigh or groin, that can be present for weeks, months or even years before the fracture. The fractures occur on trivial trauma, are horizontal with a medial beak and are non-comminuted. A committee put together by the American Society for Bone and Mineral Research described the major and minor criteria for these fractures, which appear to be related to duration of bisphosphonate therapy. The overall risk appears quite low, especially when compared to the number of hip fractures saved by these therapies, but they often require surgical fixation and are difficult to heal. Some evidence suggests that if the fractures are found early, when there is evidence of periosteal stress reaction or stress fracture, prior to the occurrence of overt fracture, that teriparatide can help heal the fracture and preclude the need for surgical repair. We routinely inform patients initiating bisphosphonates that if they develop thigh or groin pain they should inform us. Routine x-rays will sometimes pick up cortical thickening or even a stress fracture, but more commonly MRI or Technetium bone scan is required. The presence of an abnormality requires at minimum a period of modified weight bearing and may need prophylactic rodding of the femur. It is important to realize that these may be bilateral (about 50% of the time) and when an abnormality is found the other femur should be checked. It is unknown whether patients who have these atypical femur fractures can ever receive antiresorptive therapies again in the future, but it seems prudent to avoid their use for the majority of these individuals.

Mode of Action

Bisphosphonates are structurally related to pyrophosphates, compounds that are incorporated into bone matrix. Bisphosphonates specifically impair osteoclast function and reduce osteoclast number, in part by inducing apoptosis. Recent evidence suggests that the nitrogen-containing bisphosphonates also inhibit protein prenylation, one of the end products in the mevalonic acid pathway, by inhibiting the enzyme farnesyl pyrophosphate synthase. This effect disrupts intracellular protein trafficking and ultimately may lead to apoptosis. Some bisphosphonates have very long retention in the skeleton and may exert long-term effects. The consequences of this, if any, are unknown.

Calcitonin

Calcitonin is a polypeptide hormone produced by the thyroid gland **(Chap. 403)**. Its physiologic role is unclear as no skeletal disease has been described in association with calcitonin deficiency or excess. Calcitonin preparations are approved by the FDA for Paget's disease, hypercalcemia, and osteoporosis in women >5 years past menopause.

Injectable calcitonin produces small increments in bone mass of the lumbar spine. However, difficulty of administration and frequent reactions, including nausea and facial flushing, make general use limited. A nasal spray containing calcitonin (200 IU/d) is available for treatment of osteoporosis in postmenopausal women. One study suggests that nasal calcitonin produces small increments in bone mass and a small reduction in new vertebral fractures in calcitonin-treated patients (at one dose) versus those on calcium alone. There has been no proven effectiveness against nonvertebral fractures. Calcitonin is not indicated for prevention of osteoporosis and is not sufficiently potent to prevent bone loss in early postmenopausal women. Calcitonin might have an analgesic effect on bone pain, both in the subcutaneous and possibly in the nasal form. Recently concerns have been raised about an increase in the incidence of cancer associated with calcitonin use. Initially, the cancer noted was of the prostate, but an analysis of all data suggested a more general increase in cancer risk. In Europe the EMA have removed the osteoporosis indication, and an FDA Advisory Committee has voted for a similar change in the United States.

Mode of Action

Calcitonin suppresses osteoclast activity by direct action on the osteoclast calcitonin receptor. Osteoclasts exposed to calcitonin cannot maintain their active ruffled border, which normally maintains close contact with underlying bone.

Denosumab

A novel agent that given twice yearly by subcutaneous administration in a randomized controlled trial in postmenopausal women with osteoporosis has been shown to increase BMD in the spine, hip, and forearm and reduce vertebral, hip, and nonvertebral fractures over a 3-year period by 70, 40, and 20%, respectively **(Fig. 404-11)**. Other clinical trials indicate ability to increase bone mass in postmenopausal women with low bone mass (above osteoporosis range) and in postmenopausal women with breast cancer treated with aromatase inhibitor therapies. In the oncology literature, denosumab reduces the risk of fractures in women on aromatase inhibitors and also reduces the risk of breast cancer recurrence significantly. In a study of men with prostate cancer treated with androgen deprivation therapy denosumab increased bone mass and reduced vertebral fracture occurrence. Denosumab was approved by the FDA in 2010 for the treatment of postmenopausal women who have a high risk for osteoporotic fractures, including those with a history of fracture or multiple risk factors for fracture, and those who have failed or are intolerant to other osteoporosis therapy. Denosumab is also approved for the treatment of osteoporosis in men at high risk for fracture, and women with breast cancer on aromatase inhibitors and men with prostate cancer on androgen deprivation treatment. A very long-term observational extension of the pivotal trial in postmenopausal women has provided evidence that BMD continues

to increase in both the spine and hip with 3–10 years of denosumab treatment, with fracture rates that are at least as low as those seen during the active placebo controlled trial.

Denosumab may increase the risk of ONJ and atypical femur fractures similarly to bisphosphonates. Estimated incidence is 5/10,000 patient years for ONJ and 1/10,000 patient years for atypical femur fractures. Denosumab can cause hypersensitivity reactions, hypocalcemia and skin reactions including dermatitis, rash, and eczema. Early concerns about an imbalance in infections with denosumab have largely been allayed.

When denosumab is discontinued, there is a rebound increase in bone turnover and an apparent acceleration of bone loss. This likely reflects the maturation of osteoclast precursors that have accumulated in marrow when the drug was administered and can become mature bone resorbing cells once the drug is withdrawn. The consequences of this rebound increase in remodeling associated bone loss is a rapid increase in the risk of fracture, particularly vertebral fracture, and a specific increase in the occurrence of multiple vertebral fractures. In patients who need to stop denosumab or in patients in whom BMD and fracture risk reduction goals have been met, temporary use of bisphosphonate treatment can prevent the rebound increase in remodeling and rapid bone loss.

Mode of Action

Denosumab is a fully human monoclonal antibody to RANKL, the final common effector of osteoclast formation, activity, and survival. Denosumab binds to RANKL, inhibiting its ability to initiate formation of mature osteoclasts from osteoclast precursors and to bring mature osteoclasts to the bone surface and initiate bone resorption. Denosumab also plays a role in reducing the survival of the osteoclast. Through these actions on the osteoclast, denosumab induces potent antiresorptive action, as assessed biochemically and histomorphometrically.

Anabolic Agents

Parathyroid Hormone

Endogenous PTH is an 84-amino-acid peptide that is largely responsible for calcium homeostasis **(Chap. 403)**. Although chronic elevation of PTH, as occurs in hyperparathyroidism, is associated with bone loss (particularly cortical bone), PTH also can exert anabolic effects on bone. Consistent with this, some observational studies have indicated that mild endogenous hyperparathyroidism is associated with maintenance of trabecular bone mass, but loss of cortical bone. On the basis of these findings, early small-scale observational studies showed that PTH analogues could augment trabecular BMD. Subsequent controlled clinical trials have confirmed that PTH can increase bone mass and reduce fracture occurrence. The first randomized controlled trial in postmenopausal women showed that (1–34)PTH (teriparatide), when superimposed on ongoing estrogen therapy, produced substantial increments in bone mass (13% over a 3-year period compared with estrogen alone) and reduced the risk of vertebral compression deformity. In the pivotal study (median, 19 months' duration), 20 μgPTH(1–34) daily by subcutaneous injection (with no additional therapy) reduced vertebral fractures by 65% and nonvertebral fractures by 40–50% **(Fig. 404-12)**. Teriparatide produces rapid and robust increases in bone formation and then bone remodeling overall, resulting in substantial increases in bone mass and improvements in microarchitecture, including cancellous connectivity and cortical width. The BMD effects, particularly in the hip, are lower when patients switch from bisphosphonates to teriparatide, possibly in proportion to the potency of the antiresorptive agent. The hip BMD effect is particularly impaired when patients switch from denosumab to teriparatide. In patients on denosumab who need teriparatide treatment, there may be a role for combination therapy. In previously untreated women, teriparatide is best administered as monotherapy and followed by a potent antiresorptive agent such as denosumab or a bisphosphonate.

In women with painful acute osteoporotic vertebral fractures, teriparatide reduced subsequent vertebral fractures by about 50% compared with risedronate. There was no difference in nonvertebral fracture outcome between the two medications. A recent study comparing teriparatide with risedronate in patients with prevalent vertebral fractures showed significant benefit for teriparatide against vertebral fractures and nearly significant benefit for teriparatide against nonvertebral fractures.

Side effects of teriparatide are generally mild and can include muscle pain, weakness, dizziness, headache, and nausea. Rodents given prolonged treatment with PTH in high doses (3 to 60 times the human dose) developed osteogenic sarcomas after ~18 months of treatment. Rare cases of osteosarcoma have been described in patients treated with teriparatide consistent with the background incidence of osteosarcoma adults. Long-term surveillance studies of a high proportion of patients diagnosed with osteosarcoma as adults in both the United States and Scandinavia reveal no prior exposure to teriparatide in any of the cases.

PTH use may be limited by its mode of administration; alternative modes of delivery are being investigated. Because of the rodent osteosarcoma data, and the maximum duration of teriparatide in the pivotal trial of 2 years, the FDA has limited teriparatide treatment to 2 years.

Mode of Action

Exogenously administered PTH appears to have direct actions on osteoblast activity, with biochemical and histomorphometric evidence of de novo bone formation early in response to PTH, before activation of bone resorption. Subsequently, PTH activates bone remodeling but still appears to favor bone formation over bone resorption. PTH given by daily injection stimulates osteoblast recruitment and activity through activation of Wnt signaling. Unlike all other treatments, PTH produces a true increase in bone tissue and an apparent restoration of bone microarchitecture **(Fig. 404-13)**.

Abaloparatide

Abaloparatide is a synthetic analogue of human PTH-related peptide (PTHrP), which has significant homology to PTH and also binds the PTH Type 1 Receptor. Abaloparatide and teriparatide exert different binding affinities to the two different receptor conformations, R^0 and RG. Compared to TPTD,

abaloparatide binds with similar high affinity to the RG conformation, but with much lesser affinity to the R^0 conformation. These differences appear to result in a similar bone formation stimulus but lesser bone resorption stimulus and abaloparatide was specifically chosen for development among a large number of PTH and PTHrP analogues for what appeared to be an optimized anabolic profile.

In the Phase 3 Abaloparatide Comparator Trial in Vertebral Endpoints (ACTIVE) study, 2463 postmenopausal women with osteoporosis were randomized to blinded daily subcutaneous abaloparatide vs placebo or open label TPTD. At 18 months, spine BMD increase was similar with abalo and TPTD (11.2% abaloparatide and 10.5% Teriparatide); in the total hip, BMD increments were slightly larger with abaloparatide (4.2% vs 3.3%). New vertebral fracture incidence was reduced by 86% with abaloparatide and 80% with teriparatide compared with placebo (both $p<0.001$). Nonvertebral fractures were reduced by 43% with abaloparatide ($p=0.05$) and 28% with TPTD (NS, $p=0.22$). The ACTIVE study is in an ongoing extension where 92% of eligible participants from the abaloparatide and placebo arms transitioned to open label alendronate for a total treatment period of 24 months of alendronate.

Romosozumab

Romosozumab is a humanized antibody that blocks the osteocyte production of sclerostin, resulting in an increase in bone formation and decline in bone resorption. In the pivotal trial (FRAME), 7180 postmenopausal women with osteoporosis were randomized to receive blinded monthly subcutaneous romosozumab (210 mg) or placebo for 1 year followed by transition to open-label subcutaneous denosumab (60 mg) every 6 months for an additional year. BMD increased over 13% in the spine and almost 7% in the hip in one year with romosozumab. At 1 year, the incidence of new vertebral fractures in the romosozumab group was significantly reduced by 73% compared with placebo. Clinical fracture risk (nonvertebral fractures and clinical vertebral fractures combined) was significantly reduced by 36%. Nonvertebral fractures were also reduced but the difference just missed statistical significance perhaps due to geographical differences; in the high enrolling Latin American region, there was no significant reduction in nonvertebral fractures, probably due to a very low background incidence in that region. In the rest of the world, nonvertebral fractures were significantly reduced by >40%. During the second year of the FRAME study, both groups transitioned to denosumab. Over 24 months, women who had received romosozumab during the first 12 months and then denosumab had 75% fewer new vertebral fractures than those who had received placebo for a year followed by denosumab. There were also nearly significant trends toward reduced clinical and nonvertebral fractures in the romosozumab/denosumab group. Compared with baseline, BMD increased by 17.6% in the spine and 8.8% in the total hip in the romosozumab/denosumab group. Safety and tolerability of the two drugs was similar with a slightly higher incidence of injection site reactions in the denosumab group. The FRAME study is in an ongoing extension where all participants received continued denosumab for an additional year. A parallel trial of very high risk patients, all of whom have prevalent vertebral fractures is also ongoing which compares romosozumab to alendronate for 1 year, followed by transition to or continuation of alendronate for 2 additional years.

OTHER UNAPPROVED PHARMACOLOGIC AGENTS

Odanacatib, a cathepsin K inhibitor, inhibits the osteoclast collagenase enzyme, preventing bone resorption but not affecting osteoclast viability. This agent was in late stage drug development. In a very large controlled clinical trial (~17,000 postmenopausal women with osteoporosis), bone mass increased substantially in spine and hip and reduced vertebral, hip and all nonvertebral fractures. Unfortunately, the medication was associated with a significantly increased risk of stroke and the development of this agent was aborted in September 2016.

Testosterone has been used to treat osteoporosis associated with low testosterone levels in men. There are data that indicate that testosterone can increase bone density, but no fracture endpoints. Since there are many other effects of testosterone, especially in older men (including prostate hypertrophy), decisions to use it for treatment of osteoporosis have to take the multisystemic effects into account.

Sodium fluoride was tested in two large parallel clinical trials in the late 1980s. Although BMD increased substantially, the increase was in part due to fluoride incorporation in the hydroxyapatite crystal. Fracture risk was not reduced and in fact was increased in nonvertebral sites. Therefore, fluoride is no longer considered a viable option for osteoporosis treatment.

Strontium ranelate has never been approved for osteoporosis in the United States but is approved in Europe and other exUS countries. It increases bone mass throughout the skeleton, but some of the increase is related to strontium incorporation into hydroxyapatite. In clinical trials, the drug reduced the risk of vertebral fractures by 37% and that of nonvertebral fractures by 14%. It appears to be modestly antiresorptive while at the same time not causing as much of a decrease in bone formation (measured biochemically). In 2014, the use of strontium was restricted because of an increased risk of cardiovascular disease and severe skin reactions. Small increased risks of venous thrombosis also occur.

Several small studies of growth hormone (GH), alone or in combination with other agents, have not shown consistent or substantial positive effects on skeletal mass.

NONPHARMACOLOGIC APPROACHES

Protective pads worn around the outer thigh, which cover the trochanteric region of the hip, can prevent hip fractures in elderly residents in nursing homes. The use of hip protectors is limited largely by issues of compliance and comfort, but new devices are being developed that may circumvent these problems and provide adjunctive treatments.

Kyphoplasty and *vertebroplasty* are also useful non-pharmacologic approaches for the treatment of painful vertebral fractures. However, no long-term data are available.

TREATMENT MONITORING

There are currently no well-accepted guidelines for monitoring treatment of osteoporosis. Because most osteoporosis treatments produce small or moderate bone mass increments on average, it is reasonable to consider BMD as a monitoring tool. Changes must exceed ~4% in the spine and 6% in the hip to be considered significant in any individual. The hip is the preferred site due to larger surface area and greater reproducibility. Medication-induced increments may require several years to produce changes of this magnitude (if they do at all). Consequently, it can be argued that BMD should be repeated at intervals >2 years. Only significant BMD reductions should prompt a change in medical regimen, as it is expected that many individuals will not show responses greater than the detection limits of the current measurement techniques.

Biochemical markers of bone turnover can help in treatment monitoring, but little hard evidence currently supports this concept; it remains unclear which endpoint is most useful. If bone turnover markers are used, a determination should be made before therapy is started and repeated ≥4 months after therapy is initiated. In general, a change in bone turnover markers must be 30–40% lower than the baseline to be significant because of the biologic and technical variability in these tests. A positive change in biochemical markers and/or bone density can be useful to help patients adhere to treatment regimens. Because markers change more rapidly than bone density they are often early signs of treatment effect. Currently collagen C-telopeptide measured on a fasting serum sample in the morning is the preferred marker of bone resorption, and osteocalcin or the propeptide of type 1 collagen (P1NP) for formation.

TABLE 404-7

Adequate Calcium Intake

Life Stage Group	Estimated Adequate Daily Calcium Intake, mg/d
Young children (1–3 years)	500
Older children (4–8 years)	800
Adolescents and young adults (9–18 years)	1300
Men and women (19–50 years)	1000
Men and women (51 and older)	1200

Note: Pregnancy and lactation needs are the same as for nonpregnant women (e.g., 1300 mg/d for adolescents/young adults and 1000 mg/d for ≥19 years).

Source: Adapted from the Standing Committee on the Scientific Evaluation of Dietary Reference Intakes. Food and Nutrition Board. Institute of Medicine. Washington, DC, 1997, National Academy Press.

TABLE 404-8

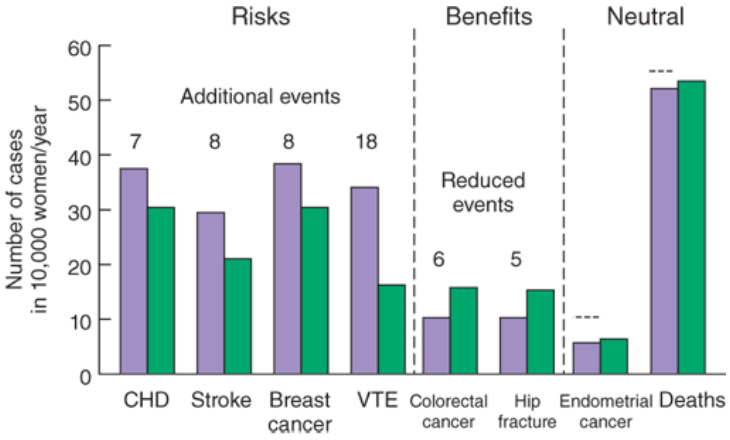
Elemental Calcium Content of Various Oral Calcium Preparations

Calcium Preparation	Elemental Calcium Content
Calcium citrate	60 mg/300 mg
Calcium lactate	80 mg/600 mg
Calcium gluconate	40 mg/500 mg
Calcium carbonate	400 mg/g
Calcium carbonate +5 μg vitamin D_2 (OsCal 250)	250 mg/tablet
Calcium carbonate (Tums 500)	500 mg/tablet

Source: Adapted from SM Krane, MF Holick, Chap. 355, in *Harrison's Principles of Internal Medicine*, 14th ed. New York, McGraw-Hill, 1998.

FIGURE 404-8

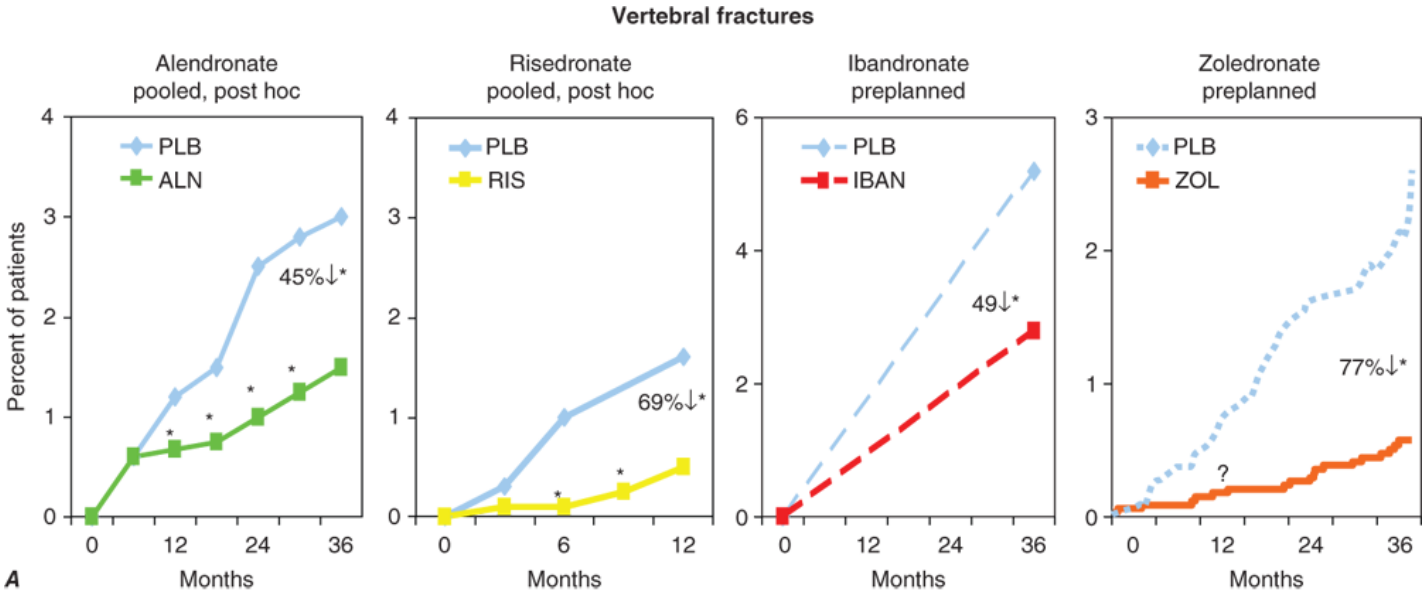
Effects of hormone therapy on event rates: green, placebo; purple, estrogen and progestin. CHD, coronary heart disease; VTE, venous thromboembolic events. *(Adapted from Women's Health Initiative. WHI HRT Update. Available at http://www.nhlbi.nih.gov.ezproxy1.library.usyd.edu.au/health/women/upd2002.htm.)*



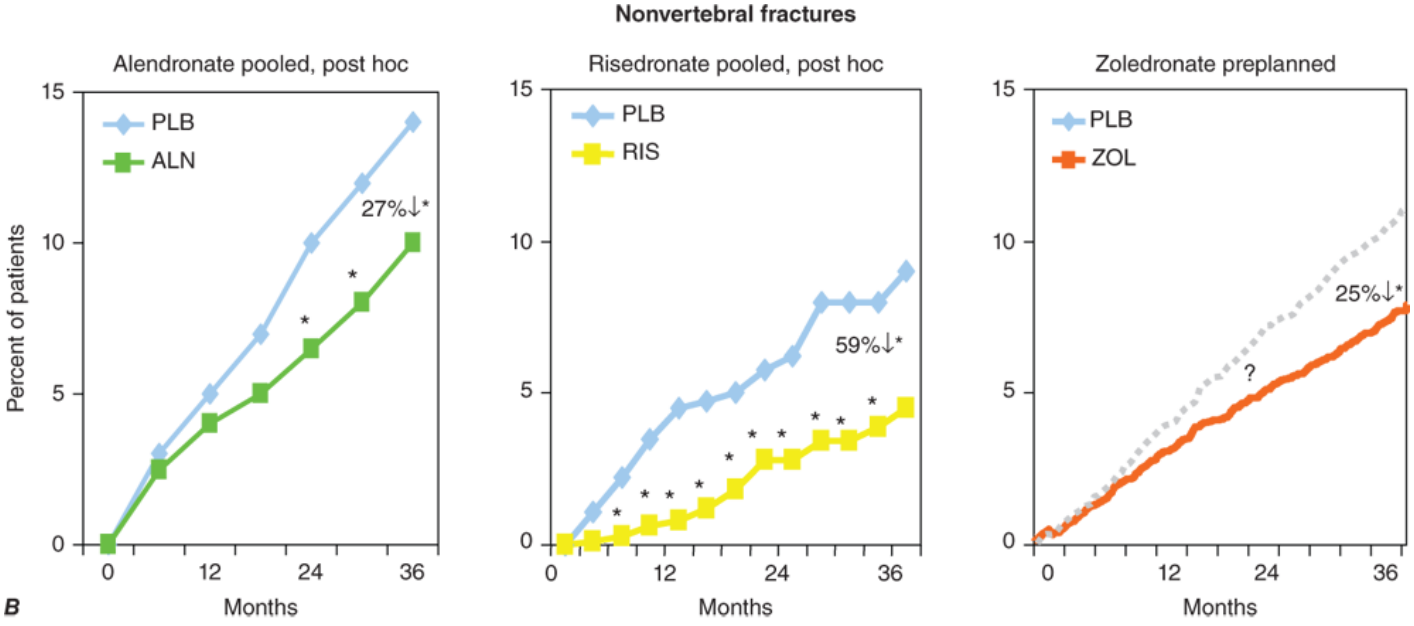
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 404-9

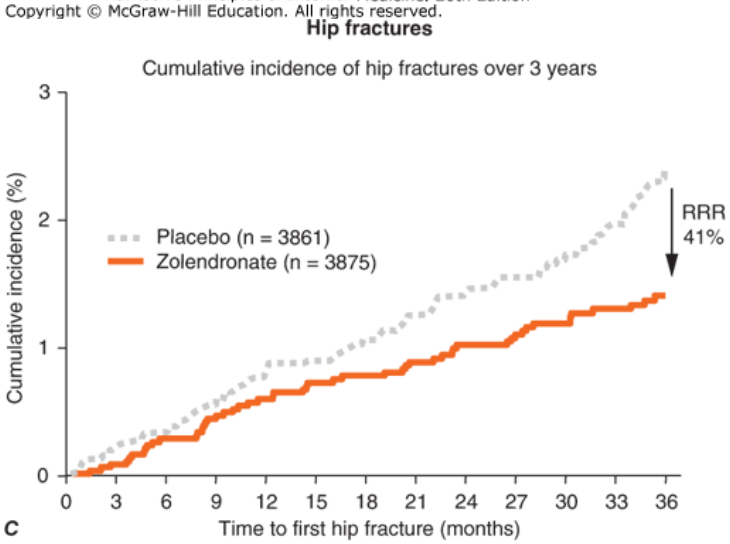
Effects of various bisphosphonates on clinical vertebral fractures ***A.*** nonvertebral fractures ***B.,*** and hip fractures ***C.*** Plb, placebo; RRR, relative risk reduction. (*After DM Black et al: J Clin Endocrinol Metab 85:4118, 2000; C Roux et al: Curr Med Res Opin 4:433, 2004; CH Chesnut et al: J Bone Miner Res 19: 1241, 2004; DM Black et al: N Engl J Med 356:1809, 2007; JT Harrington et al: Calcif Tissue Int 74:129, 2003.*)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



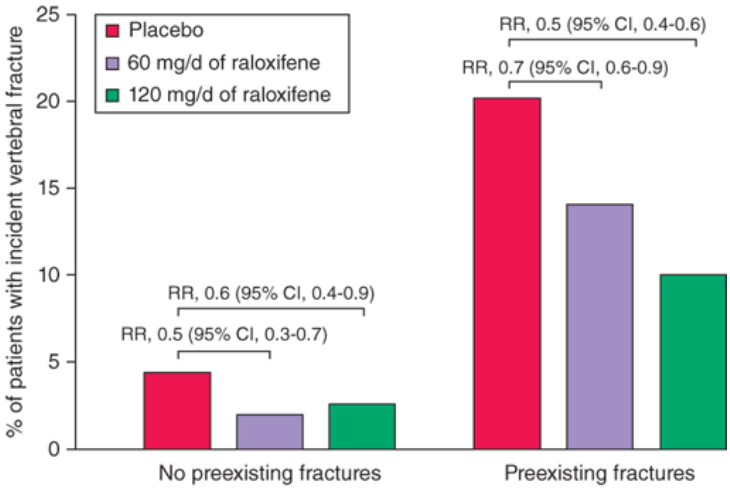
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 404-10

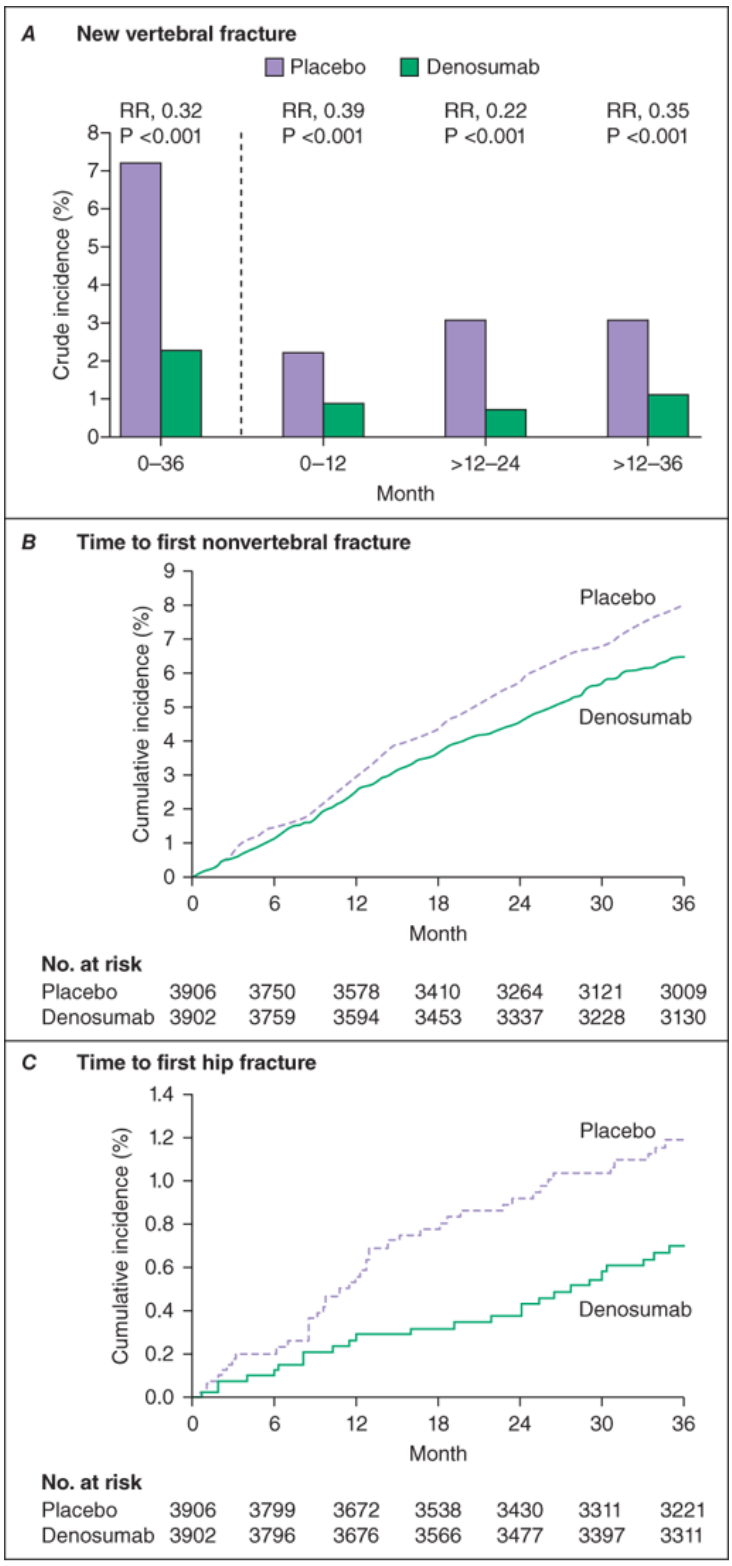
Effects of two doses of raloxifene on incident vertebral fractures in the MORE trial. *(After B Ettinger et al: JAMA:282:637, 1999.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 404-11

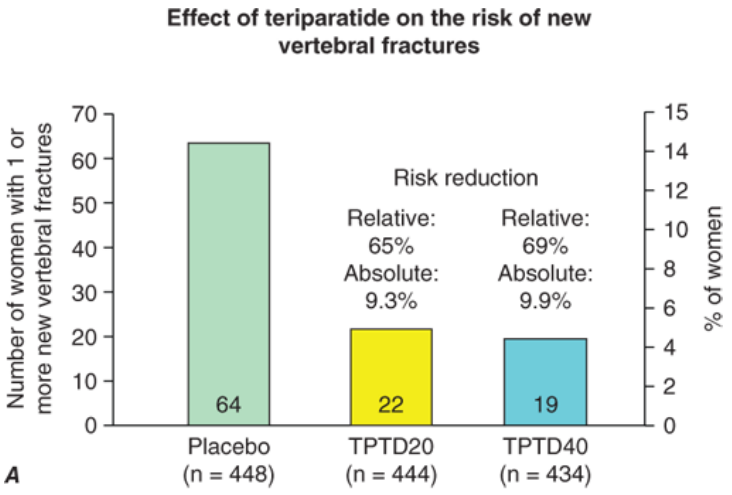
Effects of denosumab on new vertebral fractures *A.* and times to nonvertebral and hip fracture ***B.*** and ***C.*** *(After SR Cummings et al: N Engl J Med:361:756, 2009.)*



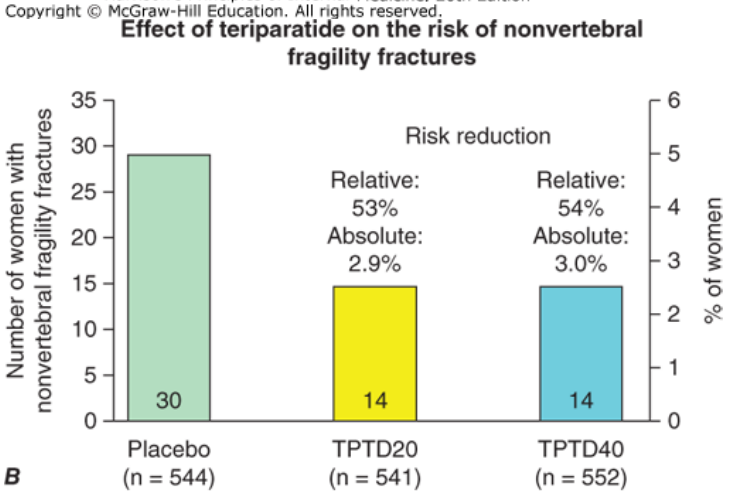
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 404-12

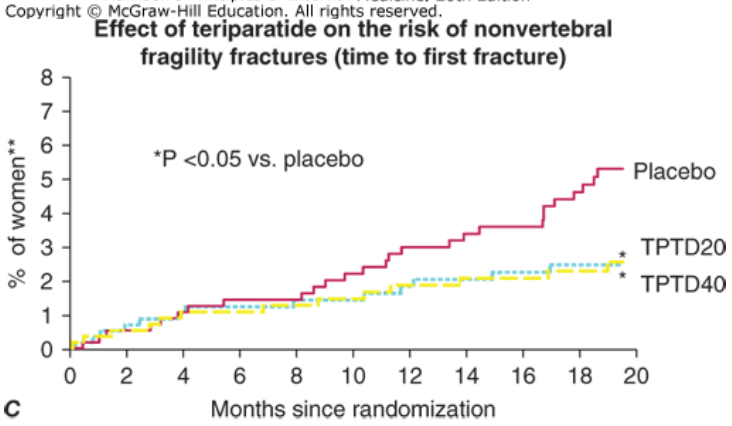
Effects of teriparatide on new vertebral fractures *A.* and nonvertebral fragility fractures ***B.*** and ***C.*** *(After RM Neer et al: N Engl J Med 344:1434, 2001.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



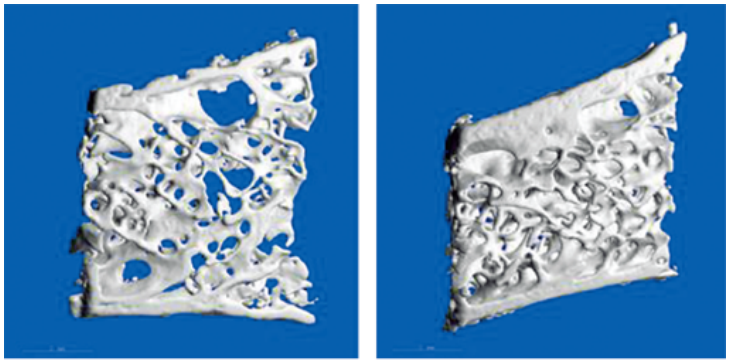
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 404-13

Effect of parathyroid hormone (PTH) treatment on bone microarchitecture. Paired biopsy specimens from a 64-year-old woman before ***A.*** and after ***B.*** treatment with PTH. *(From DW Dempster et al: J Bone Miner Res 16:1846, 2001.)*



A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GLUCOCORTICOID-INDUCED OSTEOPOROSIS

Osteoporotic fractures are a well-characterized consequence of the hypercortisolism associated with Cushing's syndrome. However, the therapeutic use of glucocorticoids is by far the most common form of glucocorticoid-induced osteoporosis (GCIO). Glucocorticoids are used widely in the treatment of a variety of disorders, including chronic lung disorders, rheumatoid arthritis, and other connective tissue diseases, inflammatory bowel disease, and after transplantation. Osteoporosis and related fractures are serious side effects of chronic glucocorticoid therapy. Because the effects of glucocorticoids on the skeleton are often superimposed on the consequences of aging and menopause, it is not surprising that women and the elderly are most frequently affected. The skeletal response to steroids is remarkably heterogeneous, however, and even young, growing individuals treated with glucocorticoids can present with fractures.

The risk of fractures depends on the dose and duration of glucocorticoid therapy, although recent data suggest that there may be no completely safe dose. Bone loss is more rapid during the early months of treatment, and trabecular bone is affected more severely than cortical bone. As a result, fractures have been shown to increase within 3 months of steroid treatment. There is an increase in fracture risk in both the axial skeleton and the appendicular skeleton, including risk of hip fracture. Bone loss can occur with any route of steroid administration, including high-dose inhaled glucocorticoids and intraarticular injections. Alternate-day delivery does not appear to ameliorate the skeletal effects of glucocorticoids.

PATHOPHYSIOLOGY

Glucocorticoids increase bone loss by multiple mechanisms, including (1) inhibition of osteoblast function and an increase in osteoblast apoptosis, resulting in impaired synthesis of new bone; (2) stimulation of bone resorption, probably as a secondary effect; (3) impairment of the absorption of calcium across the intestine, probably by a vitamin D–independent effect; (4) increase of urinary calcium loss and perhaps induction of some degree of secondary hyperparathyroidism; (5) reduction of adrenal androgens and suppression of ovarian and testicular secretion of estrogens and androgens; and (6) induction of glucocorticoid myopathy, which may exacerbate effects on skeletal and calcium homeostasis as well as increase the risk of falls.

EVALUATION OF THE PATIENT

Because of the prevalence of glucocorticoid-induced bone loss, it is important to evaluate the status of the skeleton in all patients starting or already receiving long-term glucocorticoid therapy. Modifiable risk factors should be identified, including those for falls. Examination should include testing of height and muscle strength. Laboratory evaluation should include an assessment of 24-h urinary calcium. All patients on long-term (>3 months) glucocorticoids should have measurement of bone mass at both the spine and the hip using DXA. If only one skeletal site can be measured, it is best to assess the spine in individuals <60 years and the hip in those >60 years.

PREVENTION

Bone loss caused by glucocorticoids can be prevented, and the risk of fractures significantly reduced. Strategies must include using the lowest dose of glucocorticoid for disease management. Topical and inhaled routes of administration are preferred, where appropriate. Risk factor reduction is important, including smoking cessation, limitation of alcohol consumption, and participation in weight-bearing exercise, when appropriate. All patients should receive an adequate calcium and vitamin D intake from the diet or from supplements.

TREATMENT

TREATMENT

Glucocorticoid-Induced Osteoporosis

Several bisphosphonates (alendronate, risedronate, and zoledronic acid) have been demonstrated in large clinical trials to reduce the risk of fractures in patients being treated with glucocorticoids and are FDA-approved for the treatment of GCIO. Teriparatide is also approved for treatment of glucocorticoid induced osteoporosis. In a trial comparing teriparatide to alendronate, BMD increases were much greater and vertebral fracture risk reduction far more substantial with teriparatide compared to alendronate. A study of denosumab has just been completed and indicates greater efficacy of denosumab compared with risedronate for treatment of GCIO. The American College of Rheumatology has just released new guidelines for the management of GCIO (in 2016).

FURTHER READING

Amin S et al: Trends in fracture incidence: A population-based study over 20 years. J Bone Miner Res 29:581, 2014. [PubMed: 23959594]

Black DM, Rosen CJ: Postmenopausal osteoporosis. N Engl J Med 374:2096, 2016. [PubMed: 27223157]

Cosman F et al: Treatment sequence matters: Anabolic and antiresorptive therapy for osteoporosis. J Bone Miner Res 32:198, 2017. [PubMed: 27925287]

Cosman F et al: Clinician's guide to prevention and treatment of osteoporosis. Osteoporos Int 25:2359, 2014. [PubMed: 25182228]

Cosman F et al: Spine fracture prevalence in a nationally representative sample of US women and men aged >/=40 years: results from the National Health and Nutrition Examination Survey (NHANES) 2013–2014. Osteoporos Int 28:2319, 2017.

Panday K et al: Medication-induced osteoporosis: Screening and treatment strategies. Ther Adv Musculoskel Dis 6:185, 2014.

Roux C, Briot K: Imminent fracture risk. Osteoporos Int. 28:1765, 2017. [PubMed: 28236126]

Sapir-Koren R, Livshits G: Osteocyte control of bone remodeling: Is sclerostin a key molecular coordinator of the balanced bone resorption-formation cycles? Osteoporos Int 25:2685, 2014. [PubMed: 25030653]

Wright NC et al: The recent prevalence of osteoporosis and low bone mass in the United States based on bone mineral density at the femoral neck or lumbar spine. J Bone Miner Res 29:2520, 2014. [PubMed: 24771492]

Chapter 405: Paget's Disease and Other Dysplasias of Bone

Murray J. Favus; Tamara J. Vokes

PAGET'S DISEASE OF BONE

Paget's disease is a localized bone-remodeling disorder that affects widespread, noncontiguous areas of the skeleton. The pathologic process is initiated by overactive osteoclastic bone resorption followed by a compensatory increase in osteoblastic new bone formation, resulting in a structurally disorganized mosaic of woven and lamellar bone. Pagetic bone is expanded, less compact, and more vascular; thus, it is more susceptible to deformities and fractures. Although most patients are asymptomatic, symptoms resulting directly from bony involvement (bone pain, secondary arthritis, fractures) or secondarily from the expansion of bone causing compression of surrounding neural tissue are not uncommon.

Epidemiology

There is a marked geographic variation in the frequency of Paget's disease, with high prevalence in Western Europe (Great Britain, France, and Germany, but not Switzerland or Scandinavia) and among those who have immigrated to Australia, New Zealand, South Africa, and North and South America. The disease is rare in native populations of the Americas, Africa, Asia, and the Middle East; when it does occur, the affected subjects usually have evidence of European ancestry, supporting the migration theory. For unclear reasons, the prevalence and severity of Paget's disease are decreasing, and the age of diagnosis is increasing.

The prevalence is greater in males and increases with age. Autopsy series reveal Paget's disease in about 3% of those over age 40. Prevalence of positive skeletal radiographs in patients aged >55 years is 2.5% for men and 1.6% for women. Elevated alkaline phosphatase (ALP) levels in asymptomatic patients have an age-adjusted incidence of 12.7 and 7 per 100,000 person-years in men and women, respectively.

Etiology

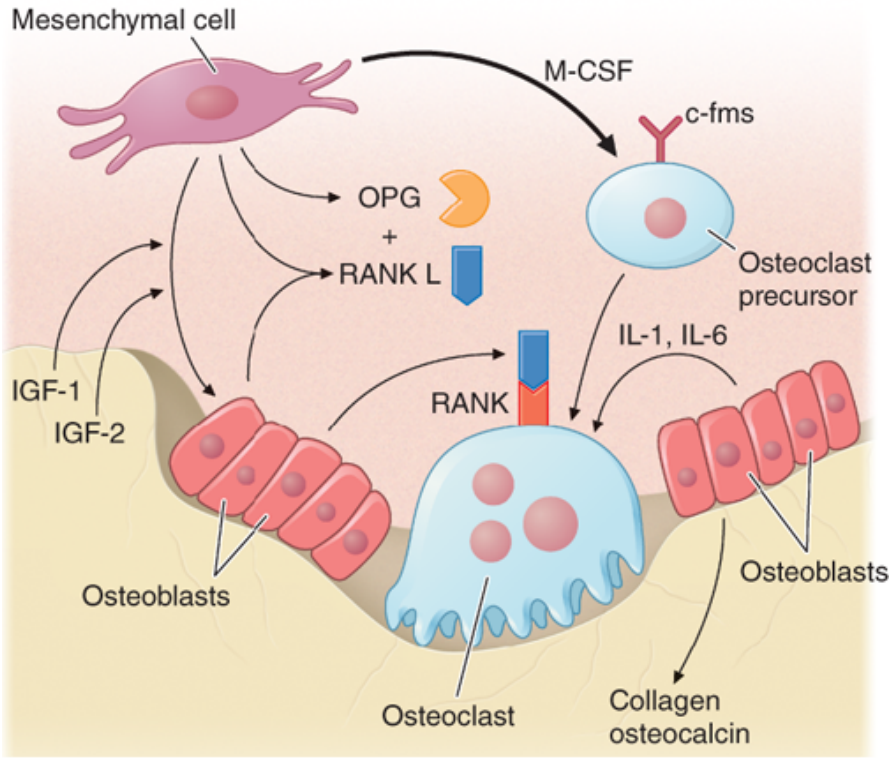
The etiology of Paget's disease of bone remains unknown, but evidence supports both genetic and viral etiologies. A positive family history is found in 15–25% of patients and, when present, raises the prevalence of the disease seven- to tenfold among first-degree relatives.

A clear genetic basis has been established for several rare familial bone disorders that clinically and radiographically resemble Paget's disease but have more severe presentation and earlier onset. A homozygous deletion of the *TNFRSF11B* gene, which encodes osteoprotegrin **(Fig. 405-1)**, causes *juvenile Paget's disease*, also known as *familial idiopathic hyperphosphatasia*, a disorder characterized by uncontrolled osteoclastic differentiation and resorption. Familial patterns of disease in several large kindred are consistent with an autosomal dominant pattern of inheritance with variable penetrance. *Familial expansile osteolysis, expansile skeletal hyperphosphatasia, and early-onset Paget's disease* are associated with mutations in *TNFRSF11A* gene, which encodes RANK (receptor activator of nuclear factor-κB), a member of the tumor necrosis factor superfamily critical for osteoclast differentiation (Fig. 405-1). Finally, mutations in the gene for

valosin-containing protein cause a rare syndrome with autosomal dominant inheritance and variable penetrance known as *inclusion body myopathy with Paget's disease and frontotemporal dementia* (*IBMPFD*). The role of genetic factors is less clear in the more common form of late-onset Paget's disease. The most common mutations identified in familial and sporadic cases of Paget's disease have been in the *SQSTM1* gene (sequestasome-1 or p62 protein) in the C-terminal ubiquitin-binding domain. The other candidate genes include: CSF1 (1p13), which encodes macrophage colony stimulating factor (M-CSF), a cytokine that is required for osteoclast differentiation; RIN3 (14q32), which encodes a guanine exchange factor called Rab and Ras interactor 3; OPTN (10p13), which is involved in regulating NFκB, TNFRSF11A (18q21) which encodes Receptor Activator of Nuclear Factor-κB (RANK), a receptor that is essential for osteoclast differentiation; and TM7SF4, which encodes Dendritic Cell-Specific Transmembrane Protein (DC-STAMP), a molecule that is essential for fusion of osteoclast. The phenotypic variability in patients with *SQSTM1* mutations suggests that additional factors, such as other genetic influences or viral infection, may influence clinical expression of the disease.

FIGURE 405-1

Diagram illustrating factors that promote differentiation and function of osteoclasts and osteoblasts and the role of the RANK pathway. Stromal bone marrow (mesenchymal) cells and differentiated osteoblasts produce multiple growth factors and cytokines, including macrophage colony-stimulating factor (M-CSF), to modulate osteoclastogenesis. RANKL (receptor activator of nuclear factor-κB ligand) is produced by osteoblast progenitors and mature osteoblasts and can bind to a soluble decoy receptor known as osteoprotegerin (OPG) to inhibit RANKL action. Alternatively, a cell-cell interaction between osteoblast and osteoclast progenitors allows RANKL to bind to its membrane-bound receptor, RANK, thereby stimulating osteoclast differentiation and function. RANK binds intracellular proteins called tumor necrosis factor receptor–associated factors (TRAFs) that mediate receptor signaling through transcription factors such as NF-κB. M-CSF binds to its receptor, c-fms, which is the cellular homologue of the *fms* oncogene. See text for the potential role of these pathways in disorders of osteoclast function such as Paget's disease and osteopetrosis. IL, interleukin; IGF, insulin-like growth factor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Several lines of evidence suggest that a viral infection may contribute to the clinical manifestations of Paget's disease, including (1) the presence of cytoplasmic and nuclear inclusions resembling paramyxoviruses (measles and respiratory

syncytial virus) in pagetic osteoclasts and (2) viral mRNA in precursor and mature osteoclasts. The viral etiology is further supported by conversion of osteoclast precursors to pagetic-like osteoclasts by vectors containing the measles virus nucleocapsid or matrix genes. The decline in the incidence of Paget's disease coincides with the widespread vaccination against measles, also consistent with the potential role of virus in the development of the disease. However, the viral etiology has been questioned by the inability to culture a live virus from pagetic bone and by failure to clone the full-length viral genes from material obtained from patients with Paget's disease.

Pathophysiology

The principal abnormality in Paget's disease is the increased number and activity of osteoclasts. Pagetic osteoclasts are large, increased 10- to 100-fold in number, and have a greater number of nuclei (as many as 100 compared to 3–5 nuclei in the normal osteoclast). The overactive osteoclasts may create a sevenfold increase in resorptive surfaces and an erosion rate of 9 µg/d (normal is 1 µg/d). Several causes for the increased number and activity of pagetic osteoclasts have been identified: (1) osteoclastic precursors are hypersensitive to $1,25(OH)_2D_3$; (2) osteoclasts are hyperresponsive to RANK ligand (RANKL), the osteoclast stimulatory factor that mediates the effects of most osteotropic factors on osteoclast formation; (3) marrow stromal cells from pagetic lesions have increased RANKL expression; (4) osteoclast precursor recruitment is increased by interleukin (IL) 6, which is increased in the blood of patients with active Paget's disease and is overexpressed in pagetic osteoclasts; (5) expression of the protooncogene c-*fos*, which increases osteoclastic activity, is increased; and (6) the antiapoptotic oncogene *Bcl*-2 in pagetic bone is overexpressed. Numerous osteoblasts are recruited to active resorption sites and produce large amounts of new bone matrix. As a result, bone turnover is high, and bone mass is normal or increased, not reduced, unless there is concomitant deficiency of calcium and/or vitamin D.

The characteristic feature of Paget's disease is increased bone resorption accompanied by accelerated bone formation. An initial osteolytic phase involves prominent bone resorption and marked hypervascularization. Radiographically, this manifests as an advancing lytic wedge, or "blade of grass" lesion. The second phase is a period of very active bone formation and resorption that replaces normal lamellar bone with haphazard (woven) bone. Fibrous connective tissue may replace normal bone marrow. In the final sclerotic phase, bone resorption declines progressively and leads to a hard, dense, less vascular pagetic or mosaic bone, which represents the so-called burned-out phase of Paget's disease. All three phases may be present at the same time at different skeletal sites.

Clinical Manifestations

Diagnosis is often made in asymptomatic patients because they have elevated ALP levels on routine blood chemistry testing or an abnormality on a skeletal radiograph obtained for another indication. The skeletal sites most commonly involved are the pelvis, vertebral bodies, skull, femur, and tibia. Familial cases with an early presentation often have numerous active sites of skeletal involvement.

The most common presenting symptom is pain, which may result from increased bony vascularity, expanding lytic lesions, fractures, bowing, or other deformities. Bowing of the femur or tibia causes gait abnormalities and abnormal mechanical stresses with secondary osteoarthritis of the hip or knee joints. Long bone bowing also causes extremity pain by stretching the muscles attached to the bone softened by the pagetic process. Back pain results from enlarged pagetic vertebrae, vertebral compression fractures, spinal stenosis, degenerative changes of the joints, and altered body mechanics with kyphosis and forward tilt of the upper back. Rarely, spinal cord compression may result from bone enlargement or from the vascular steal syndrome. Skull involvement may cause headaches, symmetric or asymmetric enlargement of the parietal or frontal bones (frontal bossing), and increased head size. Cranial expansion may narrow cranial foramens and cause neurologic complications including hearing loss from cochlear nerve damage from temporal bone involvement, cranial nerve palsies, and softening of the base of the skull (*platybasia*) with the risk of brainstem compression. Pagetic

involvement of the facial bones may cause facial deformity; loss of teeth and other dental conditions; and, rarely, airway compression.

Fractures are serious complications of Paget's disease and usually occur in long bones at areas of active or advancing lytic lesions. Common fracture sites are the femoral shaft and subtrochanteric regions. Neoplasms arising from pagetic bone are rare (<0.5%). The incidence of sarcoma appears to be decreasing, possibly because of earlier, more effective treatment with potent antiresorptive agents. The majority of tumors are osteosarcomas, which usually present with new pain in a long-standing pagetic lesion. Osteoclast-rich benign giant cell tumors may arise in areas adjacent to pagetic bone, and they respond to glucocorticoid therapy.

Cardiovascular complications may occur in patients with involvement of large (15–35%) portions of the skeleton and a high degree of disease activity (ALP four times above normal). The extensive arteriovenous shunting and marked increases in blood flow through the vascular pagetic bone lead to a high-output state and cardiac enlargement. However, high-output heart failure is relatively rare and usually develops in patients with concomitant cardiac pathology. In addition, calcific aortic stenosis and diffuse vascular calcifications have been associated with Paget's disease.

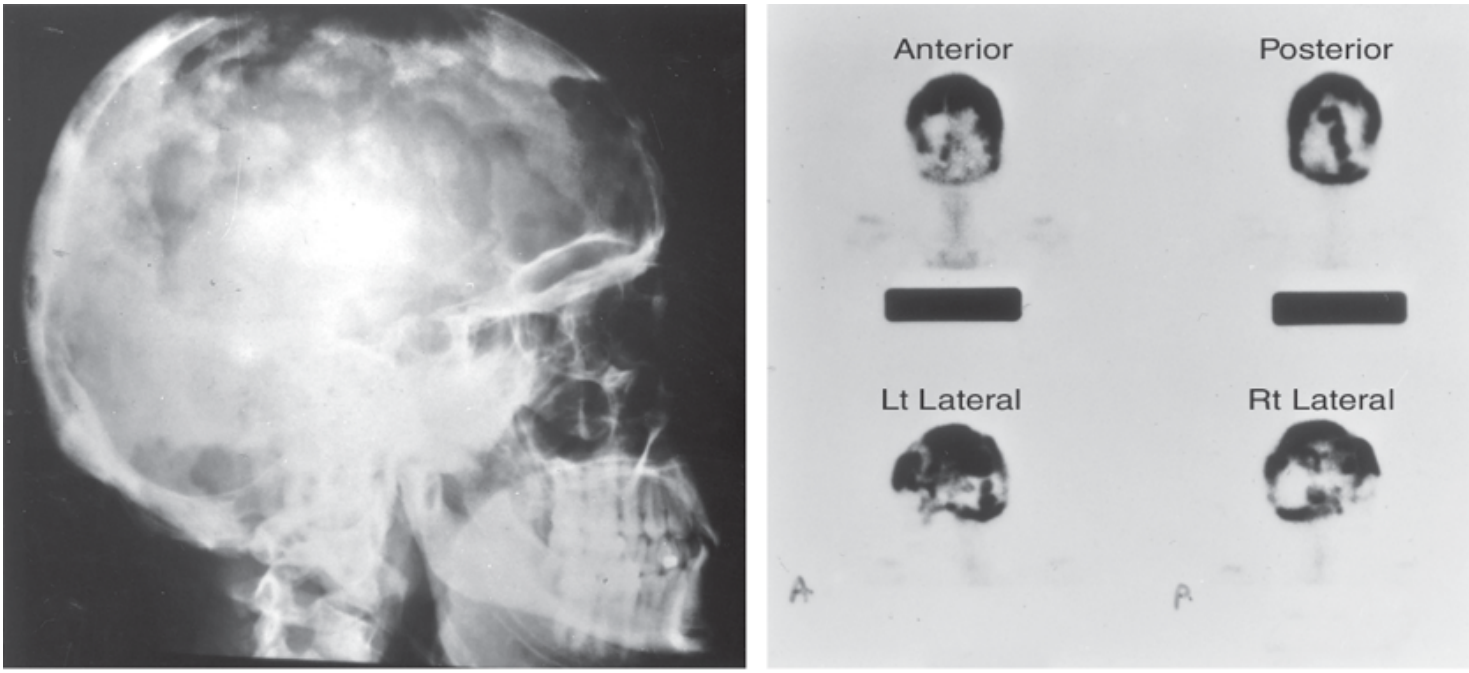
Diagnosis

The diagnosis may be suggested on clinical examination by the presence of an enlarged skull with frontal bossing, bowing of an extremity, or short stature with simian posturing. An extremity with an area of warmth and tenderness to palpation may suggest an underlying pagetic lesion. Other findings include bony deformity of the pelvis, skull, spine, and extremities; arthritic involvement of the joints adjacent to lesions; and leg-length discrepancy resulting from deformities of the long bones.

Paget's disease is usually diagnosed from radiologic and biochemical abnormalities. Radiographic findings typical of Paget's disease include enlargement or expansion of an entire bone or area of a long bone, cortical thickening, coarsening of trabecular markings, and typical lytic and sclerotic changes. Skull radiographs **(Fig. 405-2)** reveal regions of "cotton wool," or osteoporosis circumscripta, thickening of diploic areas, and enlargement and sclerosis of a portion or all of one or more skull bones. Vertebral cortical thickening of the superior and inferior end plates creates a "picture frame" vertebra. Diffuse radiodense enlargement of a vertebra is referred to as "ivory vertebra." Pelvic radiographs may demonstrate disruption or fusion of the sacroiliac joints; porotic and radiodense lesions of the ilium with whorls of coarse trabeculation; thickened and sclerotic iliopectineal line (brim sign); and softening with protrusio acetabuli, with axial migration of the hips and functional flexion contracture. Radiographs of long bones reveal bowing deformity and typical pagetic changes of cortical thickening and expansion and areas of lucency and sclerosis **(Fig. 405-3)**. Radionuclide ^{99m}Tc bone scans are less specific but are more sensitive than standard radiographs for identifying sites of active skeletal lesions. Although computed tomography (CT) and magnetic resonance imaging (MRI) studies are not necessary in most cases, CT may be useful for the assessment of possible fracture, and MRI is necessary to assess the possibility of sarcoma, giant cell tumor, or metastatic disease in pagetic bone. Definitive diagnosis of malignancy often requires bone biopsy.

FIGURE 405-2

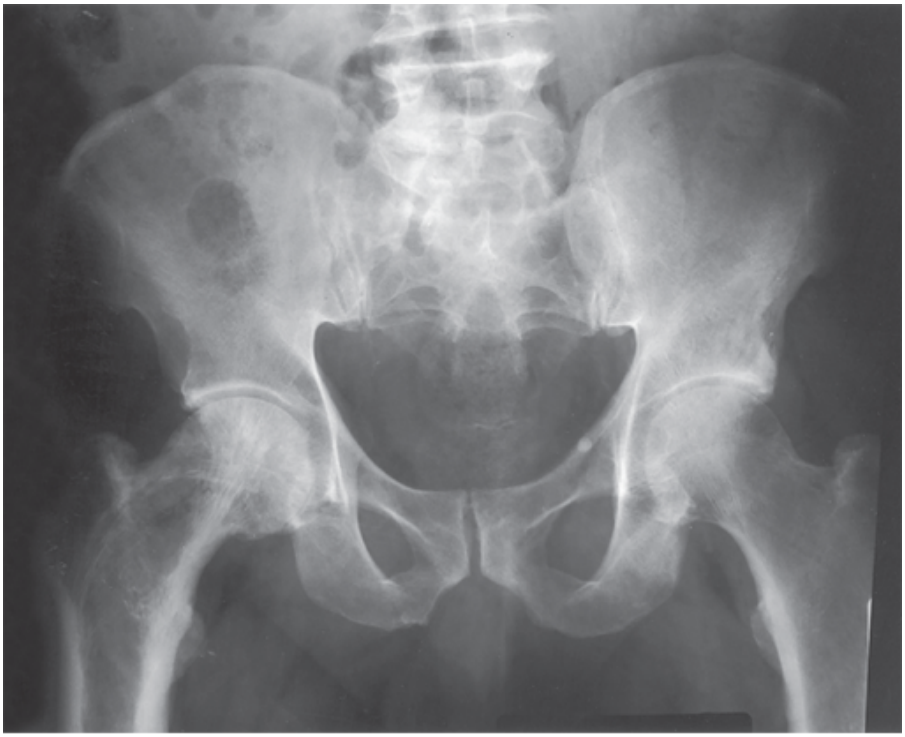
A 48-year-old woman with Paget's disease of the skull. ***Left.*** Lateral radiograph showing areas of both bone resorption and sclerosis. ***Right.*** ^{99m}Tc HDP bone scan with anterior, posterior, and lateral views of the skull showing diffuse isotope uptake by the frontal, parietal, occipital, and petrous bones.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 405-3

Radiograph of a 73-year-old man with Paget's disease of the right proximal femur. Note the coarsening of the trabecular pattern with marked cortical thickening and narrowing of the joint space consistent with osteoarthritis secondary to pagetic deformity of the right femur.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Biochemical evaluation is useful in the diagnosis and management of Paget's disease. The marked increase in bone turnover can be monitored using biochemical markers of bone formation and resorption. The parallel rise in markers of bone formation and resorption confirms the coupling of bone formation and resorption in Paget's disease. The degree of bone marker elevation reflects the extent and severity of the disease. For most patients, serum total ALP remains the test of

choice both for diagnosis and assessing response to therapy. Occasionally, a symptomatic patient with evidence of progression at a single site may have a normal total ALP level but increased bone-specific ALP. For unclear reasons, serum osteocalcin, another marker of bone formation, is not always elevated and is not recommended for use in diagnosis or management of Paget's disease. In contrast, bone formation marker P1NP does reflect the activity of the disease and can be used instead of total ALP. Bone resorption markers (serum or urine N-telopeptide or C-telopeptide measured in the blood or urine) are also elevated in active Paget's disease and decrease more rapidly in response to therapy than does ALP.

Serum calcium and phosphate levels are normal in Paget's disease. Immobilization of a patient with active Paget's disease may rarely cause hypercalcemia and hypercalciuria and increase the risk for nephrolithiasis. However, the discovery of hypercalcemia, even in the presence of immobilization, should prompt a search for another cause of hypercalcemia. In contrast, hypocalcemia or mild secondary hyperparathyroidism may develop in Paget's patients with very active bone formation and insufficient calcium and vitamin D intake, particularly during bisphosphonate therapy when bone resorption is rapidly suppressed and active bone formation continues. Therefore, adequate calcium and vitamin D intake should be instituted prior to administration of bisphosphonates.

TREATMENT

TREATMENT

Paget's Disease of Bone

The development of effective and potent pharmacologic agents **(Table 405-1)** has changed the treatment philosophy from treating only symptomatic patients to treating asymptomatic patients who are at risk for complications. According to the Endocrine Society Clinical Practice Guidelines published in 2014, pharmacologic therapy is indicated for most patients with active Paget's disease who are at risk of complications. Treatment may be initiated to control symptoms caused by metabolically active Paget's disease such as bone pain, fracture, headache, pain from pagetic radiculopathy or arthropathy, or neurologic complications; to decrease local blood flow and minimize operative blood loss in patients who need surgery at an active pagetic site; to reduce hypercalciuria that may occur during immobilization; and to decrease the risk of complications when disease activity is high (elevated ALP) and when the site of involvement involves weight-bearing bones, areas adjacent to major joints, vertebral bodies, and the skull. Whether or not early therapy prevents late complications remains to be determined. A randomized study of over 1200 patients from the United Kingdom showed no difference in bone pain, fracture rates, quality of life, and hearing loss between patients who received pharmacologic therapy to control symptoms (bone pain) and those receiving bisphosphonates to normalize serum ALP. However, the conclusions of that study are debatable since the most potent agent (zoledronic acid, the current drug of choice) was not used/available. It seems likely that the restoration of normal bone architecture following suppression of pagetic activity will prevent further deformities and complications.

Agents approved for treatment of Paget's disease suppress the very high rates of bone resorption and secondarily decrease the high rates of bone formation (Table 405-1). As a result of decreasing bone turnover, pagetic structural patterns, including areas of poorly mineralized woven bone, are replaced by more normal cancellous or lamellar bone. Reduced bone turnover can be documented by a decline in serum formation markers (ALP and P1NP) and urine or serum resorption markers (N-telopeptide, C-telopeptide).

Bisphosphonates are the mainstay of pharmacologic therapy of Paget's disease. Among them, zoledronic acid is currently recommended as the first choice, particularly for those who have severe disease or need rapid normalization of bone turnover (neurologic symptoms, severe bone pain due to a lytic lesion, risk of an impending fracture, or pretreatment prior to elective surgery in an area of active disease). Zoledronic acid normalized bone turnover faster and in a high proportion of patients (over 90%) than oral bisphosphonates with the therapeutic effect persisting for months or even years. It is given at

a dose of 5 mg as an intravenous infusion over 20 min although slower rates of infusion are recommended for elderly or those with mild impairment of renal function. More significant renal impairment (GFR <35 mL/min) is a contraindication for use of zoledronic acid due to higher risk of further deterioration of renal function. About 20–25% of patients experience a flu-like syndrome after the first infusion, which can be partly ameliorated by pretreatment with acetaminophen or nonsteroidal anti-inflammatory drugs (NSAIDs). Oral bisphosphonates, alendronate and risedronate, can be used in subjects who have mild disease or some degree of renal impairment. Oral bisphosphonates should be taken first thing in the morning on an empty stomach, followed by maintenance of upright posture with no food, drink, or other medications for 30–60 min. The first clinically useful agent, etidronate, is no longer used due to its low potency and higher risk of inducing osteomalacia. The efficacy of different agents, based on their ability to normalize or decrease ALP levels, is summarized in Table 405-1, although the response rates are not comparable because they are obtained from different studies.

The subcutaneous injectable form of salmon calcitonin is approved for the treatment of Paget's disease but is rarely used due to its low potency and should be reserved for patients who either do not tolerate bisphosphonates or have a contraindication to their use. For patients with contraindication to bisphosphonates, another alternative is denosumab, an antibody to RANKL, which has been reported to result in reduction in ALP. However, it has not been approved for this indication and has less complete and less durable effect than bisphosphonates.

TABLE 405-1

Pharmacologic Agents Approved for Treatment of Paget's Disease

Name	Dose and Mode of Delivery	Normalization of alkaline phosphatase (ALP)
Zoledronic acid	5 mg IV over 15 min	90% of patients at 6 mo
Pamidronate	30 mg IV/d over 4 h on 3 days	~50% of patients
Risedronate	30 mg PO/d for 2 mo	73% of patients
Alendronate	40 mg PO/d for 6 mo	63% of patients
Tiludronate	800 mg PO daily for 3 mo	35% of patients
Etidronate	200–400 mg PO/d × 6 mo	15% of patients
Calcitonin (Miacalcin)	100 U SC daily for 6–18 mo (may reduce to 50 U 3 × per wk)	(Reduction of ALP by up to 50%)

SCLEROSING BONE DISORDERS

OSTEOPETROSIS

Osteopetrosis refers to a group of disorders caused by severe impairment of osteoclast-mediated bone resorption. Other terms that are often used include marble bone disease, which captures the solid x-ray appearance of the involved skeleton, and Albers-Schonberg disease, which refers to the milder, adult form of osteopetrosis also known as autosomal dominant

osteopetrosis type II. The major types of osteopetrosis include malignant (severe, infantile, autosomal recessive) osteopetrosis and benign (adult, autosomal dominant) osteopetrosis types I and II. A rare autosomal recessive intermediate form has a more benign prognosis. Autosomal recessive carbonic anhydrase (CA) II deficiency produces osteopetrosis of intermediate severity associated with renal tubular acidosis and cerebral calcification.

Etiology and Genetics

Naturally occurring and gene-knockout animal models with phenotypes similar to those of the human disorders have been used to explore the genetic basis of osteopetrosis. The primary defect in osteopetrosis is the loss of osteoclastic bone resorption and preservation of normal osteoblastic bone formation. Osteoprotegerin (OPG) is a soluble decoy receptor that binds osteoblast-derived RANK ligand, which mediates osteoclast differentiation and activation (Fig. 405-1). Transgenic mice that overexpress OPG develop osteopetrosis, presumably by blocking RANK ligand. Mice deficient in RANK lack osteoclasts and develop severe osteopetrosis.

Recessive mutations of CA II prevent osteoclasts from generating an acid environment in the clear zone between its ruffled border and the adjacent mineral surface. Absence of CA II, therefore, impairs osteoclastic bone resorption. Other forms of human disease have less clear genetic defects. About one-half of the patients with malignant infantile osteopetrosis have a mutation in the *TCIRG1* gene encoding the osteoclast-specific subunit of the vacuolar proton pump, which mediates the acidification of the interface between bone mineral and the osteoclast ruffled border. Mutations in the *CICN7* chloride channel gene cause autosomal dominant osteopetrosis type II.

Clinical Presentation

The incidence of autosomal recessive severe (malignant) osteopetrosis ranges from 1 in 200,000 to 1 in 500,000 live births. As bone and cartilage fail to undergo modeling, paralysis of one or more cranial nerves may occur due to narrowing of the cranial foramens. Failure of skeletal modeling also results in inadequate marrow space, leading to extramedullary hematopoiesis with hypersplenism and pancytopenia. Hypocalcemia due to lack of osteoclastic bone resorption may occur in infants and young children. The untreated infantile disease is fatal, often before age 5.

Adult (benign) osteopetrosis is an autosomal dominant disease that is usually diagnosed by the discovery of typical skeletal changes in young adults who undergo radiologic evaluation of a fracture. The prevalence is 1 in 100,000 to 1 in 500,000 adults. The course is not always benign, because fractures may be accompanied by loss of vision, deafness, psychomotor delay, mandibular osteomyelitis, and other complications usually associated with the juvenile form. In some kindred, nonpenetrance results in skip generations, while in other families, severely affected children are born into families with benign disease. The milder form of the disease does not usually require treatment.

Radiography

Typically, there are generalized symmetric increases in bone mass with thickening of both cortical and trabecular bone. Diaphyses and metaphyses are broadened, and alternating sclerotic and lucent bands may be seen in the iliac crests, at the ends of long bones, and in vertebral bodies. The cranium is usually thickened, particularly at the base of the skull, and the paranasal and mastoid sinuses are underpneumatized.

Laboratory Findings

The only significant laboratory findings are elevated serum levels of osteoclast-derived tartrate-resistant acid phosphatase (TRAP) and the brain isoenzyme of creatine kinase. Serum calcium may be low in severe disease, and parathyroid hormone and 1,25-dihydroxyvitamin D levels may be elevated in response to hypocalcemia.

TREATMENT

TREATMENT

Osteopetrosis

Allogeneic HLA-identical bone marrow transplantation has been successful in some children. Following transplantation, the marrow contains progenitor cells and normally functioning osteoclasts. A cure is most likely when children are transplanted before age 4. Marrow transplantation from nonidentical HLA-matched donors has a much higher failure rate. Limited studies in small numbers of patients have suggested variable benefits following treatment with interferon γ-1β, 1,25-dihydroxyvitamin D (which stimulates osteoclasts directly), methylprednisolone, and a low-calcium/high-phosphate diet.

Surgical intervention is indicated to decompress optic or auditory nerve compression. Orthopedic management is required for the surgical treatment of fractures and their complications including malunion and postfracture deformity.

PYKNODYSOSTOSIS

This is an autosomal recessive form of osteosclerosis that is believed to have affected the French impressionist painter Henri de Toulouse-Lautrec. The molecular basis involves mutations in the gene that encodes cathepsin K, a lysosomal metalloproteinase highly expressed in osteoclasts and important for bone-matrix degradation. Osteoclasts are present but do not function normally. Pyknodysostosis is a form of short-limb dwarfism that presents with frequent fractures but usually a normal life span. Clinical features include short stature; kyphoscoliosis and deformities of the chest; high arched palate; proptosis; blue sclerae; dysmorphic features including small face and chin, frontooccipital prominence, pointed beaked nose, large cranium, and obtuse mandibular angle; and small, square hands with hypoplastic nails. Radiographs demonstrate a generalized increase in bone density, but in contrast to osteopetrosis, the long bones are normally shaped. Separated cranial sutures, including the persistent patency of the anterior fontanel, are characteristic of the disorder. There may also be hypoplasia of the sinuses, mandible, distal clavicles, and terminal phalanges. Persistence of deciduous teeth and sclerosis of the calvarium and base of the skull are also common. Histologic evaluation shows normal cortical bone architecture with decreased osteoblastic and osteoclastic activities. Serum chemistries are normal, and unlike osteopetrosis, there is no anemia. There is no known treatment for this condition, and there are no reports of attempted bone marrow transplant.

PROGRESSIVE DIAPHYSEAL DYSPLASIA

Also known as *Camurati-Engelmann disease*, progressive diaphyseal dysplasia is an autosomal dominant disorder that is characterized radiographically by diaphyseal hyperostosis and a symmetric thickening and increased diameter of the endosteal and periosteal surfaces of the diaphyses of the long bones, particularly the femur and tibia, and, less often, the fibula, radius, and ulna. The genetic defect responsible for the disease has been localized to the area of chromosome 19q13.2 encoding tumor growth factor (TGF)-β1. The mutation promotes activation of TGF-β1. The clinical severity is variable. The most common presenting symptoms are pain and tenderness of the involved areas, fatigue, muscle wasting, and gait disturbance. The weakness may be mistaken for muscular dystrophy. Characteristic body habitus includes thin limbs with little muscle mass yet prominent and palpable bones and, when the skull is involved, large head with prominent forehead and proptosis. Patients may also display signs of cranial nerve palsies, hydrocephalus, central hypogonadism, and Raynaud’s phenomenon. Radiographically, patchy progressive endosteal and periosteal new bone formation is observed along the diaphyses of the long bones. Bone scintigraphy shows increased radiotracer uptake in involved areas.

Treatment with low-dose glucocorticoids relieves bone pain and may reverse the abnormal bone formation. Intermittent bisphosphonate therapy has produced clinical improvement in a limited number of patients.

HYPEROSTOSIS CORTICALIS GENERALISATA

This is also known as *van Buchem's disease*; it is an autosomal recessive disorder characterized by endosteal hyperostosis in which osteosclerosis involves the skull, mandible, clavicles, and ribs. The major manifestations are due to narrowed cranial foramens with neural compressions that may result in optic atrophy, facial paralysis, and deafness. Adults may have an enlarged mandible. Serum ALP levels may be elevated, which reflect the uncoupled bone remodeling with high osteoblastic formation rates and low osteoclastic resorption. As a result, there is increased accumulation of normal bone. Endosteal hyperostosis with syndactyly, known as *sclerosteosis*, is a more severe form. The genetic defects for both sclerosteosis and van Buchem's disease have been associated with mutations in the *SOST* gene.

MELORHEOSTOSIS

Melorheostosis (Greek, "flowing hyperostosis") may occur sporadically or follow a pattern consistent with an autosomal recessive disorder. The major manifestation is progressive linear hyperostosis in one or more bones of one limb, usually a lower extremity. The name comes from the radiographic appearance of the involved bone, which resembles melted wax that has dripped down a candle. Symptoms appear during childhood as pain or stiffness in the area of sclerotic bone. There may be associated ectopic soft tissue masses, composed of cartilage or osseous tissue, and skin changes overlying the involved bone, consisting of scleroderma-like areas and hypertrichosis. The disease does not progress in adults, but pain and stiffness may persist. Laboratory tests are unremarkable. No specific etiology is known. There is no specific treatment. Surgical interventions to correct contractures are often unsuccessful.

OSTEOPOIKILOSIS

The literal translation of osteopoikilosis is "spotted bones"; it is a benign autosomal dominant condition in which numerous small, variably shaped (usually round or oval) foci of bony sclerosis are seen in the epiphyses and adjacent metaphyses. The lesions may involve any bone except the skull, ribs, and vertebrae. They may be misidentified as metastatic lesions. The main differentiating points are that bony lesions of osteopoikilosis are stable over time and do not accumulate radionucleotide on bone scanning. In some kindred, osteopoikilosis is associated with connective tissue nevi known as *dermatofibrosis lenticularis disseminata*, also known as *Buschke-Ollendorff syndrome*. Histologic inspection reveals thickened but otherwise normal trabeculae and islands of normal cortical bone. No treatment is indicated.

HEPATITIS C–ASSOCIATED OSTEOSCLEROSIS

Hepatitis C–associated osteosclerosis (HCAO) is a rare acquired diffuse osteosclerosis in adults with prior hepatitis C infection. After a latent period of several years, patients develop diffuse appendicular bone pain and a generalized increase in bone mass with elevated serum ALP. Bone biopsy and histomorphometry reveal increased rates of bone formation, decreased bone resorption with a marked decrease in osteoclasts, and dense lamellar bone. One patient had increased serum OPG levels, and bone biopsy showed large numbers of osteoblasts positive for OPG and reduced osteoclast number. Empirical therapy includes pain control, and there may be beneficial response to bisphosphonate. Long-term antiviral therapy may reverse the bone disease.

DISORDERS ASSOCIATED WITH DEFECTIVE MINERALIZATION

HYPOPHOSPHATASIA

This is a rare inherited disorder that presents as rickets in infants and children or osteomalacia in adults with paradoxically low serum levels of ALP. The frequency of the severe neonatal and infantile forms is about 1 in 100,000 live births in Canada, where the disease is most common because of its high prevalence among Mennonites and Hutterites. It is rare in African

Americans. The severity of the disease is remarkably variable, ranging from intrauterine death associated with profound skeletal hypomineralization at one extreme to premature tooth loss as the only manifestation in some adults. Severe cases are inherited in an autosomal recessive manner, but the genetic patterns are less clear for the milder forms. The disease is caused by a deficiency of tissue nonspecific (bone/liver/kidney) ALP (TNSALP), which, although ubiquitous, results only in bone abnormalities. Protein levels and functions of the other ALP isozymes (germ cell, intestinal, placental) are normal. Defective ALP permits accumulation of its major naturally occurring substrates including phosphoethanolamine (PEA), inorganic pyrophosphate (PPi), and pyridoxal 5′-phosphate (PLP). The accumulation of PPi interferes with mineralization through its action as a potent inhibitor of hydroxyapatite crystal growth.

Perinatal hypophosphatasia becomes manifest during pregnancy and is often complicated by polyhydramnios and intrauterine death. The infantile form becomes clinically apparent before the age of 6 months with failure to thrive, rachitic deformities, functional craniosynostosis despite widely open fontanels (which are actually hypomineralized areas of the calvarium), raised intracranial pressure, and flail chest with predisposition to pneumonia. Hypercalcemia and hypercalciuria are common. This form has a mortality rate of about 50%. Prognosis seems to improve for the children who survive infancy. Childhood hypophosphatasia has variable clinical presentation. Premature loss of deciduous teeth (before age 5) is the hallmark of the disease. Rickets causes delayed walking with waddling gait, short stature, and dolichocephalic skull with frontal bossing. The disease often improves during puberty but may recur in adult life. Adult hypophosphatasia presents during middle age with painful, poorly healing metatarsal stress fractures or thigh pain due to femoral pseudofractures. It is important to recognize hypophosphatasia in adults because treatment with bisphosphonates can result in increased rather than decreased bone fragility.

Laboratory investigation reveals low ALP levels and normal or elevated levels of serum calcium and phosphorus despite clinical and radiologic evidence of rickets or osteomalacia. Serum parathyroid hormone, 25-hydroxyvitamin D, and 1,25-dihydroxyvitamin D levels are normal. The elevation of PLP is specific for the disease and may even be present in asymptomatic parents of severely affected children. Because vitamin B_6 increases PLP levels, vitamin B_6 supplements should be discontinued 1 week before testing. Clinical testing is available to detect loss-of-function mutation(s) within the *ALPL* gene that encodes TNSALP.

In contrast to other forms of rickets and osteomalacia, calcium and vitamin D supplementation should be avoided because they may aggravate hypercalcemia and hypercalciuria. A low-calcium diet, glucocorticoids, and calcitonin have been used in a small number of patients with variable responses. Because fracture healing is poor, placement of intramedullary rods is best for acute fracture repair and for prophylactic prevention of fractures. In 2015, asfotase alfa, a tissue nonspecific ALP has been approved as enzyme replacement therapy perinatal/infantile- and juvenile-onset forms.

AXIAL OSTEOMALACIA

This is a rare disorder characterized by defective skeletal mineralization despite normal serum calcium and phosphate levels. Clinically, the disorder presents in middle-aged or elderly men with chronic axial skeletal discomfort. Cervical spine pain may also be present. Radiographic findings are mainly osteosclerosis due to coarsened trabecular patterns typical of osteomalacia. Spine, pelvis, and ribs are most commonly affected. Histologic changes show defective mineralization and flat, inactive osteoblasts. The primary defect appears to be an acquired defect in osteoblast function. The course is benign, and there is no established treatment. Calcium and vitamin D therapies are not effective.

FIBROGENESIS IMPERFECTA OSSIUM

This is a rare condition of unknown etiology. It presents in both sexes; in middle age or later; and with progressive, intractable skeletal pain and fractures; worsening immobilization; and a debilitating course. Radiographic evaluation reveals

generalized osteomalacia, osteopenia, and occasional pseudofractures. Histologic features include a tangled pattern of collagen fibrils with abundant osteoblasts and osteoclasts. There is no effective treatment. Spontaneous remission has been reported in a small number of patients. Calcium and vitamin D have not been beneficial.

FIBROUS DYSPLASIA AND MCCUNE-ALBRIGHT SYNDROME

Fibrous dysplasia is a sporadic disorder characterized by the presence of one (monostotic) or more (polyostotic) expanding fibrous skeletal lesions composed of bone-forming mesenchyme. The association of the polyostotic form with café au lait spots and hyperfunction of an endocrine system such as pseudoprecocious puberty of ovarian origin is known as *McCune-Albright syndrome* (MAS). A spectrum of the phenotypes is caused by activating mutations in the *GNAS1* gene, which encodes the α subunit of the stimulatory G protein ($G_s\alpha$). As the postzygotic mutations occur at different stages of early development, the extent and type of tissue affected are variable and explain the mosaic pattern of skin and bone changes. GTP binding activates the $G_s\alpha$ regulatory protein and mutations in regions of $G_s\alpha$ that selectively inhibit GTPase activity, which results in constitutive stimulation of the cyclic AMP–protein kinase A signal transduction pathway. Such mutations of the $G_s\alpha$ protein–coupled receptor may cause autonomous function in bone (parathyroid hormone receptor); skin (melanocyte-stimulating hormone receptor); and various endocrine glands including ovary (follicle-stimulating hormone receptor), thyroid (thyroid-stimulating hormone receptor), adrenal (adrenocorticotropic hormone receptor), and pituitary (growth hormone–releasing hormone receptor). The skeletal lesions are composed largely of mesenchymal cells that do not differentiate into osteoblasts, resulting in the formation of imperfect bone. In some areas of bone, fibroblast-like cells develop features of osteoblasts in that they produce extracellular matrix that organizes into woven bone. Calcification may occur in some areas. In other areas, cells have features of chondrocytes and produce cartilage-like extracellular matrix.

Clinical Presentation

Fibrous dysplasia occurs with equal frequency in both sexes, whereas MAS with precocious puberty is more common (10:1) in girls. The monostotic form is the most common and is usually diagnosed in patients between 20 and 30 years of age without associated skin lesions. The polyostotic form typically manifests in children <10 years old and may progress with age. Early-onset disease is generally more severe. Lesions may become quiescent in puberty and progress during pregnancy or with estrogen therapy. In polyostotic fibrous dysplasia, the lesions most commonly involve the maxilla and other craniofacial bones, ribs, and metaphyseal or diaphyseal portions of the proximal femur or tibia. Expanding bone lesions may cause pain, deformity, fractures, and nerve entrapment. Sarcomatous degeneration involving the facial bones or femur is infrequent (<1%). The risk of malignant transformation is increased by radiation, which has proven to be ineffective treatment. In rare patients with widespread lesions, renal phosphate wasting and hypophosphatemia may cause rickets or osteomalacia. Hypophosphatemia may be due to production of a phosphaturic factor by the abnormal fibrous tissue.

MAS patients may have café au lait spots, which are flat, hyperpigmented skin lesions that have rough borders (“coast of Maine”) in contrast to the café au lait lesions of neurofibromatosis that have smooth borders (“coast of California”). The most common endocrinopathy is isosexual pseudoprecocious puberty in girls. Other less common endocrine disorders include thyrotoxicosis, Cushing’s syndrome, acromegaly, hyperparathyroidism, hyperprolactinemia, and pseudoprecocious puberty in boys.

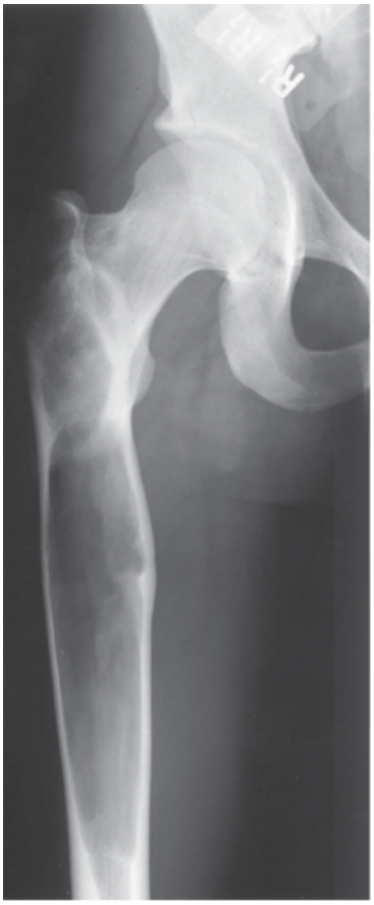
Radiographic Findings

In long bones, the fibrous dysplastic lesions are typically well-defined, radiolucent areas with thin cortices and a ground-glass appearance. Lesions may be lobulated with trabeculated areas of radiolucency **(Fig. 405-4)**. Involvement of facial bones usually presents as radiodense lesions, which may create a leonine appearance (leontiasis osea). Expansile cranial

lesions may narrow foramens and cause optic lesions, reduce hearing, and create other manifestations of cranial nerve compression.

FIGURE 405-4

Radiograph of a 16-year-old male with fibrous dysplasia of the right proximal femur. Note the multiple cystic lesions, including the large lucent lesion in the proximal midshaft with scalloping of the interior surface. The femoral neck contains two lucent cystic lesions.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Laboratory Results

Serum ALP is occasionally elevated but calcium, parathyroid hormone, 25-hydroxyvitamin D, and 1,25-dihydroxy-vitamin D levels are normal. Patients with extensive polyostotic lesions may have hypophosphatemia, hyperphosphaturia, and osteomalacia. The hypophosphatemia and phosphaturia are directly related to the levels of fibroblast growth factor 23 (FGF23). Biochemical markers of bone turnover may be elevated.

TREATMENT

TREATMENT

Fibrous Dysplasia and MAS

Spontaneous healing of the lesions does not occur, and there is no established effective treatment. Improvement in bone pain and partial or complete resolution of radiographic lesions have been reported after IV bisphosphonate therapy.

Surgical stabilization is used to prevent pathologic fracture or destruction of a major joint space and to relieve nerve root or cranial nerve compression or sinus obstruction.

OTHER DYSPLASIAS OF BONE AND CARTILAGE

PACHYDERMOPERIOSTOSIS

Pachydermoperiostosis, or hypertrophic osteoarthropathy (primary or idiopathic), is an autosomal dominant disorder characterized by periosteal new bone formation that involves the distal extremities. The lesions present as clubbing of the digits and hyperhidrosis and thickening of the skin, primarily of the face and forehead. The changes usually appear during adolescence, progress over the next decade, and then become quiescent. During the active phase, progressive enlargement of the hands and feet produces a paw-like appearance, which may be mistaken for acromegaly. Arthralgias, pseudogout, and limited mobility may also occur. The disorder must be differentiated from secondary hypertrophic osteopathy that develops during the course of serious pulmonary disorders. The two conditions can be differentiated by standard radiography of the digits in which secondary pachydermoperiostosis has exuberant periosteal new bone formation and a smooth and undulating surface. In contrast, primary hypertrophic osteopathy has an irregular periosteal surface.

There are no diagnostic blood or urine tests. Synovial fluid does not have an inflammatory profile. There is no specific therapy, although a limited experience with colchicine suggests some benefit in controlling the arthralgias.

OSTEOCHONDRODYSPLASIAS

These include several hundred heritable disorders of connective tissue. These primary abnormalities of cartilage manifest as disturbances in cartilage and bone growth. Selected growth-plate chondrodysplasias are described here. **For discussion of chondrodysplasias, see Chap. 406.**

Achondrodysplasia

This is a relatively common form of short-limb dwarfism that occurs in 1 in 15,000 to 1 in 40,000 live births. The disease is caused by a mutation of the fibroblast growth factor receptor 3 (*FGFR3*) gene that results in a gain-of-function state. Most cases are sporadic mutations. However, when the disorder appears in families, the inheritance pattern is consistent with an autosomal dominant disorder. The primary defect is abnormal chondrocyte proliferation at the growth plate that causes development of short, but proportionately thick, long bones. Other regions of the long bones may be relatively unaffected. The disorder is manifest by the presence of short limbs (particularly the proximal portions), normal trunk, large head, saddle nose, and an exaggerated lumbar lordosis. Severe spinal deformity may lead to cord compression. The homozygous disorder is more serious than the sporadic form and may cause neonatal death. Pseudoachondroplasia clinically resembles achondrodysplasia but has no skull abnormalities.

Enchondromatosis

This is also called *dyschondroplasia* or *Ollier's disease*; it is also a disorder of the growth plate in which the primary cartilage is not resorbed. Cartilage ossification proceeds normally, but it is not resorbed normally, leading to cartilage accumulation. The changes are most marked at the ends of long bones, where the highest growth rates occur. Chondrosarcoma develops infrequently. The association of enchondromatosis and cavernous hemangiomas of the skin and soft tissues is known as *Maffucci's syndrome*. Both Ollier's disease and Maffucci's syndrome are associated with various malignancies, including granulosa cell tumor of the ovary and cerebral glioma.

Multiple Exostoses

This is also called *diaphyseal aclasis* or *osteochondromatosis*; it is a genetic disorder that follows an autosomal dominant pattern of inheritance. In this condition, areas of growth plates become displaced, presumably by growing through a defect in the perichondrium. The lesion begins with vascular invasion of the growth-plate cartilage, resulting in a characteristic radiographic finding of a mass that is in direct communication with the marrow cavity of the parent bone. The underlying cortex is resorbed. The disease is caused by inactivating mutations of the *EXT1* and *EXT2* genes, whose products normally regulate processing of chondrocyte cytoskeletal proteins. The products of the *EXT* gene likely function as tumor suppressors, with the loss-of-function mutation resulting in abnormal proliferation of growth-plate cartilage. Solitary or multiple lesions are located in the metaphyses of long bones. Although usually asymptomatic, the lesions may interfere with joint or tendon function or compress peripheral nerves. The lesions stop growing when growth ceases but may recur during pregnancy. There is a small risk for malignant transformation into chondrosarcoma.

EXTRASKELETAL (ECTOPIC) CALCIFICATION AND OSSIFICATION

Deposition of calcium phosphate crystals (*calcification*) or formation of true bone (*ossification*) in nonosseous soft tissue may occur by one of three mechanisms: (1) metastatic calcification due to a supranormal calcium × phosphate concentration product in extracellular fluid; (2) dystrophic calcification due to mineral deposition into metabolically impaired or dead tissue despite normal serum levels of calcium and phosphate; and (3) ectopic ossification, or true bone formation. Disorders that may cause extraskeletal calcification or ossification are listed in **Table 405-2**.

TABLE 405-2

Diseases and Conditions Associated with Ectopic Calcification and Ossification

Metastatic calcification	Dystrophic calcification
Hypercalcemic states	Inflammatory disorders
Primary hyperparathyroidism	Scleroderma
Sarcoidosis	Dermatomyositis
Vitamin D intoxication	Systemic lupus erythematosus
Milk-alkali syndrome	Trauma-induced
Renal failure	Ectopic ossification
Hyperphosphatemia	Myositis ossificans
Tumoral calcinosis	Postsurgery
Secondary hyperparathyroidism	Burns
Pseudohypoparathyroidism	Neurologic injury
Renal failure	Other trauma
Hemodialysis	Fibrodysplasia ossificans progressiva
Cell lysis following chemotherapy	
Therapy with vitamin D and phosphate	

METASTATIC CALCIFICATION

Soft tissue calcification may complicate diseases associated with significant hypercalcemia, hyperphosphatemia, or both. In addition, vitamin D and phosphate treatments or calcium administration in the presence of mild hyperphosphatemia, such as during hemodialysis, may induce ectopic calcification. Calcium phosphate precipitation may complicate any disorder when the serum calcium × phosphate concentration product is >75. The initial calcium phosphate deposition is in the form

of small, poorly organized crystals, which subsequently organize into hydroxyapatite crystals. Calcifications that occur in hypercalcemic states with normal or low phosphate have a predilection for kidney, lungs, and gastric mucosa. Hyperphosphatemia with normal or low serum calcium may promote soft tissue calcification with predilection for the kidney and arteries. The disturbances of calcium and phosphate in renal failure and hemodialysis are common causes of soft tissue (metastatic) calcification.

TUMORAL CALCINOSIS

This is a rare genetic disorder characterized by masses of metastatic calcifications in soft tissues around major joints, most often shoulders, hips, and ankles. Tumoral calcinosis differs from other disorders in that the periarticular masses contain hydroxyapatite crystals or amorphous calcium phosphate complexes, while in fibrodysplasia ossificans progressiva (below), true bone is formed in soft tissues. About one-third of tumoral calcinosis cases are familial, with both autosomal recessive and autosomal dominant modes of inheritance reported. The disease is also associated with a variably expressed abnormality of dentition marked by short bulbous roots, pulp calcification, and radicular dentin deposited in swirls. The primary defect responsible for the metastatic calcification appears to be hyperphosphatemia resulting from the increased capacity of the renal tubule to reabsorb filtered phosphate. Spontaneous soft tissue calcification is related to the elevated serum phosphate, which, along with normal serum calcium, exceeds the concentration product of 75.

The disease usually presents in childhood and continues throughout the patient's life. The calcific masses are typically painless and grow at variable rates, sometimes becoming large and bulky. The masses are often located near major joints but remain extracapsular. Joint range of motion is not usually restricted unless the tumors are very large. Complications include compression of neural structures and ulceration of the overlying skin with drainage of chalky fluid and risk of secondary infection. Small deposits not detected by standard radiographs may be detected by ^{99m}Tc bone scanning. The most common laboratory findings are hyperphosphatemia and elevated serum 1,25-dihydroxyvitamin D levels. Serum calcium, parathyroid hormone, and ALP levels are usually normal. Renal function is also usually normal. Urine calcium and phosphate excretions are low, and calcium and phosphate balances are positive.

An acquired form of the disease may occur with other causes of hyperphosphatemia, such as secondary hyperparathyroidism associated with hemodialysis, hypoparathyroidism, pseudohypoparathyroidism, and massive cell lysis following chemotherapy for leukemia. Tissue trauma from joint movement may contribute to the periarticular calcifications. Metastatic calcifications are also seen in conditions associated with hypercalcemia, such as in sarcoidosis, vitamin D intoxication, milk-alkali syndrome, and primary hyperparathyroidism. In these conditions, however, mineral deposits are more likely to occur in proton-transporting organs such as kidney, lungs, and gastric mucosa in which an alkaline milieu is generated by the proton pumps.

TREATMENT

TREATMENT

Tumoral Calcinosis

Therapeutic successes have been achieved with surgical removal of subcutaneous calcified masses, which tend not to recur if all calcification is removed from the site. Reduction of serum phosphate by chronic phosphorus restriction may be accomplished using low dietary phosphorus intake alone or in combination with oral phosphate binders. The addition of the phosphaturic agent acetazolamide may be useful. Limited experience using the phosphaturic action of calcitonin deserves further testing.

DYSTROPHIC CALCIFICATION

Posttraumatic calcification may occur with normal serum calcium and phosphate levels and normal ion-solubility product. The deposited mineral is either in the form of amorphous calcium phosphate or hydroxyapatite crystals. Soft tissue calcification complicating connective tissue disorders such as scleroderma, dermatomyositis, and systemic lupus erythematosus may involve localized areas of the skin or deeper subcutaneous tissue and is referred to as *calcinosis circumscripta*. Mineral deposition at sites of deeper tissue injury including periarticular sites is called *calcinosis universalis*.

ECTOPIC OSSIFICATION

True extraskeletal bone formation that begins in areas of fasciitis following surgery, trauma, burns, or neurologic injury is referred to as *myositis ossificans*. The bone formed is organized as lamellar or trabecular, with normal osteoblasts and osteoclasts conducting active remodeling. Well-developed haversian systems and marrow elements may be present. A second cause of ectopic bone formation occurs in an inherited disorder, *fibrodysplasia ossificans progressiva*.

FIBRODYSPLASIA OSSIFICANS PROGRESSIVA

This is also called *myositis ossificans progressiva*; it is a rare autosomal dominant disorder characterized by congenital deformities of the hands and feet and episodic soft tissue swellings that ossify. Ectopic bone formation occurs in fascia, tendons, ligaments, and connective tissue within voluntary muscles. Tender, rubbery induration, sometimes precipitated by trauma, develops in the soft tissue and gradually calcifies. Eventually, heterotopic bone forms at these sites of soft tissue trauma. Morbidity results from heterotopic bone interfering with normal movement and function of muscle and other soft tissues. Mortality is usually related to restrictive lung disease caused by an inability of the chest to expand. Laboratory tests are unremarkable.

There is no effective medical therapy. Bisphosphonates, glucocorticoids, and a low-calcium diet have largely been ineffective in halting progression of the ossification. Surgical removal of ectopic bone is not recommended, because the trauma of surgery may precipitate formation of new areas of heterotopic bone. Dental complications including frozen jaw may occur following injection of local anesthetics.

FURTHER READING

Cundy T et al: Durability of response to zoledronate treatment and competing mortality in Paget's disease of bone. Bone Miner Res 32:753, 2017.

Langston AL et al: Randomized trial of intensive bisphosphonate treatment versus symptomatic management in Paget's disease of bone. J Bone Miner Res 25:20, 2010. [PubMed: 19580457]

Ralston SH et al: Letter to the Editor: The Endocrine Society Clinical Practice Guidelines on Paget's Disease: Many recommendations are not evidence based. J Clin Endocrinol Metab 100:L45, 2015. [PubMed: 26047080]

Reid IR et al: Treatment of Paget's disease of bone with denosumab: Case report and literature review. Calcif Tissue Int 99:322, 2016. [PubMed: 27193832]

Robinson C et al: Fibrous dysplasia/McCune-Albright syndrome: Clinical and translational perspectives. Curr Osteoporos Rep 14:178, 2016. [PubMed: 27492469]

Singer FR: Paget's disease of the bone—Genetic and environmental factors. Nat Rev Endocrinol 11:662, 2015. [PubMed: 26284446]

Singer FR et al: Paget's disease of bone: An endocrine society clinical practice guideline. J Clin Endocrinol Metab 99:4408, 2014. [PubMed: 25406796]

Whyte MP: Sclerosing bone disorders, in *Primer on the Metabolic Bone Diseases and Disorders of Mineral Metabolism*, 8th ed, CJ Rosen (ed). Wiley-Blackwell, 2013, pp 769–785.

Whyte MP: Hypophosphatasia—Aetiology, nosology, pathogenesis, diagnosis and treatment. Nat Rev Endocrinol 12:233, 2016. [PubMed: 26893260]

Chapter 406: Heritable Disorders of Connective Tissue

Darwin J. Prockop; John F. Bateman

FIGURE 406-1

CLASSIFICATION OF CONNECTIVE TISSUE DISORDERS

Some of the most common conditions that are transmitted genetically in families are disorders that produce clinically obvious changes in the skeleton, skin, or other relatively acellular tissues that have been loosely defined as connective tissues. Because of their heritability, some of the disorders were recognized as potentially traceable to mutated genes soon after the principles of genetics were introduced into medicine by Garrod and others. About half a century later, McKusick emphasized the specificity of many of the diseases for selective connective tissues and suggested that they were probably caused by mutations in genes coding for the major proteins found in those tissues. In the last several decades, many of the disorders have been linked to mutations in several hundred different genes expressed in connective tissues. However, classifying the disorders on the basis of either their clinical presentations or the mutations causing them is continuing to present a challenge for both the clinician and the molecular biologist.

The information on the disorders has continued to develop on two levels. The initial clinical classifications suggested by McKusick and many others had to be refined as more patients were examined. For example, some patients had skin changes similar to those commonly seen in Ehlers-Danlos syndrome (EDS), but this feature was over-shadowed by other features such as extreme hypotonia or sudden rupture of large blood vessels. To account for the full spectrum of presentations in patients and families, many of the disorders have been re-classified several times and each has been divided into a series of sub-types. The task was daunting. For example, a recent effort to classify all the heritable disorders that alter the skeleton defined 436 distinctive conditions that were divided into 42 major groups.

The identification of mutations causing the diseases has developed on a parallel track. The first genes cloned for connective tissues were the two genes coding for Type I collagen, the most abundant protein in bones, skin, tendons and several other tissues. Some of the first assays in patients with osteogenesis imperfecta (OI) revealed mutations in Type I collagen genes. And biochemical data developed primarily with cultures of skin fibroblasts from the patients demonstrated that the mutations dramatically altered the synthesis or structure of collagen fibers. The results stimulated efforts to identify additional mutations in genes coding for structural proteins. Genes for collagens provided an attractive paradigm to search for mutations, since a series of different types of collagens were found in different connective tissues and the collagen genes were readily isolated by their unique signature sequences. Also, the collagen genes were vulnerable to a large number of different mutations because of unusual structural requirements of the protein. The search for mutations in collagen genes proved fruitful in that mutations were found in most patients with OI, in many patients with hyper-extensible skin, in some patients with dwarfism, and in patients with other disorders, including some such as the Alport syndrome (AS) that were not initially classified as disorders of connective tissue. Also mutations in collagen genes were found in subset of patients with a diagnosis of osteoarthritis (OA) and a subset of patients with the diagnosis of osteoporosis. However, the search for mutations quickly expanded to hundreds of other genes that included genes for other structural proteins, for the post-translational processing of the structural proteins, for growth factors and their receptors, and other genes whose functions are still not fully understood.

In many instances, the mutations helped to define the clinical sub-type of the disorder. In some, however, it did not. Some patients with the same clinical presentations were found to have mutations in different genes. Also, some patients with different manifestations were found to have mutations in same genes. In addition, it was difficult to establish whether a change in the structure of a gene caused the phenotypic changes in the patients and was not simply a neutral polymorphism. Therefore, there has been a continuing debate as to whether the disorders should be classified by their clinical presentations or by the genes at fault. As an illustration of the problems, mutations in 324 genes have been found associated with the 436 defined disorders of the skeleton. The latest nosology for the disorders remains a “hybrid between a list of clinically defined disorders, waiting for molecular clarification, and an annotated database documenting the phenotypic spectrum produced by mutations in a given gene.” A simpler system of classification proved feasible for one rare heritable disorder of skin, epidermatolysis bullosa. The disorder was first defined by the presence of friction-induced blister. It was then divided into subtypes that were defined by the ultrastructural layers of the skin that cleaved and blistered. Most patients in each subtype were subsequently shown to have mutations in genes expressed in the corresponding layer of skin. Even with these patients, however, the strength of the genotype-phenotype correlation varies and mutations have not yet been found in every patient.

The best pathway through this maze of information is probably to begin by matching the signs and symptoms in a patient with the presentations that define each clinical classification. A major focus should be on the most common disorders, recognizing that the signs and symptoms may vary among different individuals and family members with the same diagnosis. Then, attempt to reach a decision, in consultation with the patient, parents and probably a specialist, as to whether a DNA analysis for the probable mutation is indicated. Among the considerations are the cost, the rigor with which

the clinical classification has been linked to mutated genes, the reassurance the diagnosis can bring to patients and their families, the use of the information for prenatal diagnosis, and the possibility that mutation-specific therapies may be developed in the future. For patients with the most severe forms, it is probably best to consult a specialist in the disease to determine a program for therapy. Patient help groups have formed for many of the diseases and are an important source of information.

Patients with the most common forms of the disorders have mutations in a limited number of genes. This chapter will focus primarily on these. Also, it will provide a brief summary of biosynthesis and structure of connective tissues that may help guide the physician from the nature of the mutations to their clinical presentations.

COMPOSITION OF CONNECTIVE TISSUES

Connective tissues such as skin, bone, cartilage, ligaments, and tendons are the critical structural frameworks of the body. They consist of a complex interacting extracellular matrix network of collagens, proteoglycans, and a large number of non-collagenous glycoproteins and proteins. While these precise combinations of up to ~500 potential extracellular matrix building blocks provide tissue-specific function, there are many overarching similarities in composition such as the role of composite collagen fibrils in providing strength and form, elastin fibrils and proteoglycans and other interacting proteins, and glycoproteins that fine-tune function **(Table 406-1)**. The most abundant components of many connective tissues are three similar fibrillar collagens (Types I, II, and III). They have a similar tensile strength that is comparable to that of steel wires. The three fibrillar collagens are distributed in a tissue-specific manner: Type I collagen accounts for most of the protein of dermis, ligaments, tendons, and demineralized bone; Type I and Type III are the most abundant proteins of large blood vessels; and Type II is the most abundant protein of cartilage.

TABLE 406-1

Constituents of Connective Tissues

Connective Tissue	Major Constituents	Approximate Amounts, % dry wt	Characteristics or Functions
Dermis, ligaments, tendons	Type I collagen	80	Large bundles of fibrils
	Type III collagen	5–15	Thin fibrils
	Type IV collagen, laminins, and nidogen	<5	Form basal laminae under epithelium
	Types V, VI, and VII collagens	<5	V modifies Type I fibrils; VI forms beaded micro-fibrils; VII forms anchoring fibrils for epidermis
	Fibrillin aggregates/elastin	<5	Provide elasticity
	Fibronectin	<5	Associated with collagen fibers and cell surfaces
	Proteoglycans[a]/hyaluronan	<0.5	Provide resiliency
Bone (demineralized)	Type I collagen	90	Complex fibril network
	Type VI collagen	1–2	Beaded micro-fibrils
	Proteoglycans[a]/hyaluronan	1	Function unclear
	Osteonectin, osteopontin, osteocalcin, α2-glycoprotein, and sialoproteins	1–5	May regulate mineralization
Aorta	Type I collagen	20–40	Fibril network
	Type III collagen	20–40	Thin fibrils
	Fibrillin aggregates/elastin	20–40	Provide elasticity
	Type IV collagen, laminins, and nidogen	<5	Form basal lamina under endothelial cells
	Types V and VI collagens	<2	V modifies Type I fibrils; VI forms beaded micro-fibrils
	Proteoglycans[a]/hyaluronan	<3	Provide resiliency
Cartilage	Type II collagen	40–50	Arcades of thin fibrils
	Type IX collagen	5–10	Links Type II fibrils and other components
	Type VI collagen	<1	Beaded micro-fibrils, largely pericellular
	Type X collagen	5–10	Forms pericellular network in hypertrophic growth plate cartilage
	Type XI collagen	<10	Incorporated into some Type II fibrils
	Proteoglycans[a]/hyaluronan	15–50	Provide resiliency

Connective Tissue	Major Constituents	Approximate Amounts, % dry wt	Characteristics or Functions
	Small leucine-rich repeat proteins (SLRPs; >6 kinds)	<5	Multiple functions in assembly and function of the tissue

[a]Over 30 proteoglycans have been identified. They differ in the structures of their core proteins and their contents of glycosaminoglycan side chains of chondroitin-4-sulfate, chondroitin-6-sulfate, dermatan sulfate, and keratin sulfate. Basal lamina contains a proteoglycan with a side chain of heparan sulfate that resembles heparin.

BIOSYNTHESIS AND TURNOVER OF CONNECTIVE TISSUES

Connective tissues are among the most stable components in living organisms, but they are not inert. During embryonic development, connective tissue membranes appear as early as the four-cell blastocyst to provide a structural scaffold for the developing embryo. With the development of blood vessels and skeleton, there is a rapid increase in the synthesis, degradation, and resynthesis of connective tissues. The turnover continues at a slower, but still rapid pace throughout postnatal development and then spikes during the growth spurt of puberty. During adulthood, the metabolic turnover of most connective tissues is slow, but it continues at a moderate pace in bone. With age, malnutrition, physical inactivity, and low gravitational stress, the rate of degradation of most connective tissues, especially in bone and skin, begins to exceed the rate of synthesis and the tissues shrink. In starvation, a large fraction of the collagen in skin and other connective tissues is degraded and provides amino acids for gluconeogenesis **(Chap. 327)**. In both OA and rheumatoid arthritis, there is extensive degradation of articular cartilage collagen. Glucocorticoids weaken most tissues by decreasing collagen synthesis. In many pathologic states, however, collagen is deposited in excess. With most injuries to tissues, inflammatory and immune responses stimulate the deposition of collagen fibrils in the form of fibrotic scars. In humans, as distinct from many other species, the deposition of the fibrils is largely irreversible and prevents regeneration of normal tissues in diseases such as hepatic cirrhosis, pulmonary fibrosis, atherosclerosis, and nephrosclerosis.

Structure and Biosynthesis of Fibrillar Collagens

The tensile strength of collagen fibers derives primarily from the self-assembly of protein monomers into large fibril structures in a process that resembles crystallization. The self-assembly requires monomers of highly uniform and relatively rigid structure. It also requires a complex series of posttranslational processing steps that maintain the solubility of the monomers until they are transported to the appropriate extracellular sites for fibril assembly. Because of the stringent requirements for correct self-assembly, it is not surprising that mutations in genes for fibrillar collagens cause many of the diseases of connective tissues.

The monomers of the three fibrillar collagens are formed from three polypeptide chains, called *α chains*, that are wrapped around each other into a rope-like triple-helical conformation. The triple helix is a unique structure among proteins, and it provides rigidity to the molecule. It also orients the side chains of amino acids in an “inside out” manner relative to most other proteins so that the charged and hydrophobic residues on the surface can direct self-assembly of the monomers into fibrils. The triple-helical conformation of the monomer is generated because each of the α chains has a repetitive amino acid sequence in which glycine (Gly) appears as every third amino acid. Each α chain contains about 1000 amino acids. Therefore, the sequence of each α chain can be designated as $(\text{-Gly-X-Y-})_n$, where X and Y represent amino acids other than glycine and n is >338. The presence of glycine, the smallest amino acid, in every third position in the sequence is critical because this residue must fit into a sterically restricted space in the middle of the helix where the three chains come together. The requirement for a glycine residue at every third position explains the severe effects of mutations that convert any of the glycine residues to an amino acid with a bulkier side chain (see below). Many of the X- and Y-position amino acids are proline and hydroxyproline, which, because of their ring structures, provide additional rigidity to the triple helix. Other X- and Y-positions are occupied by charged or hydrophobic amino acids that precisely direct lateral and longitudinal assembly of the monomers into highly ordered fibrils. Mutations that substitute amino acids in some X- and Y-positions can, in rare instances, also produce genetic diseases.

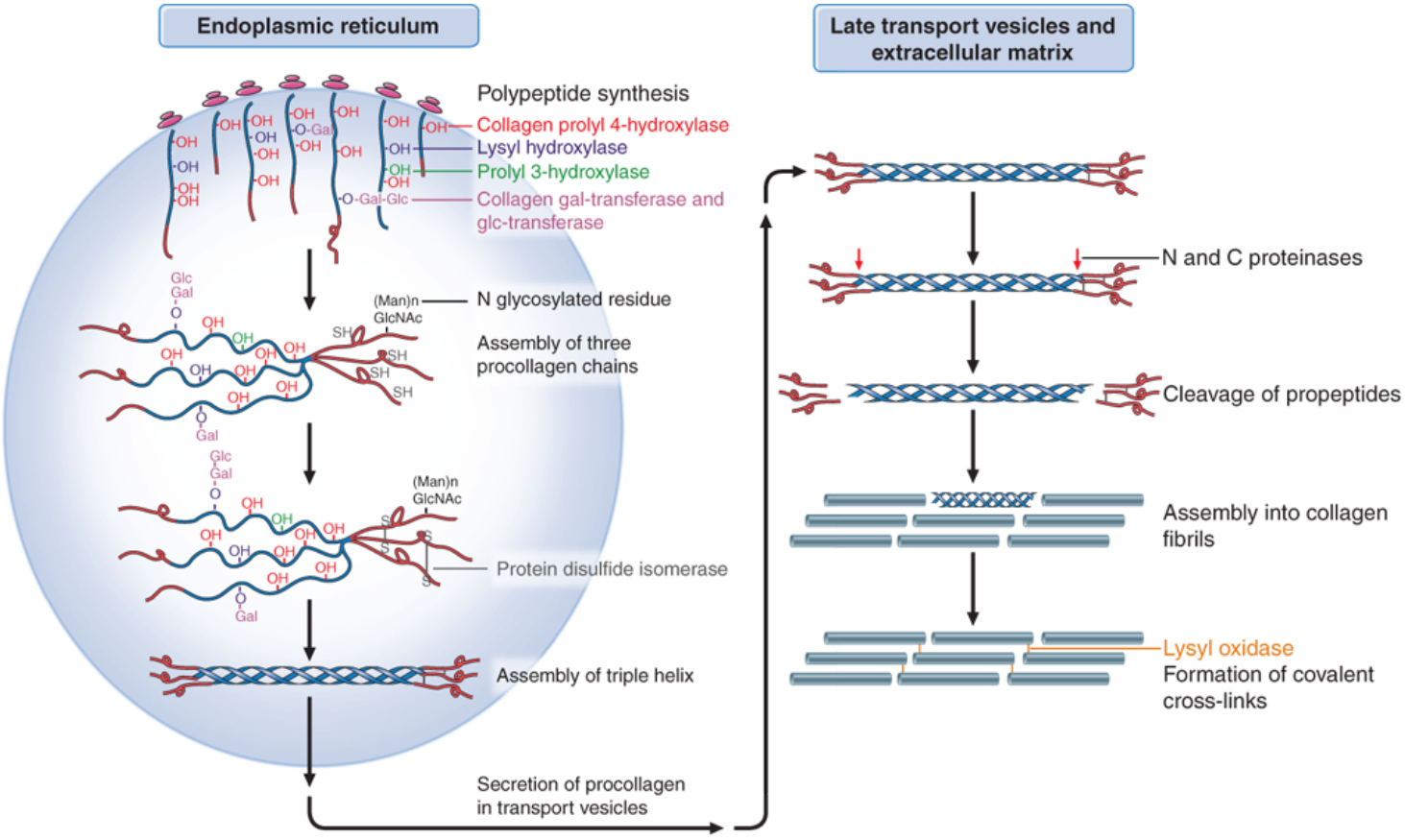
The fibers formed by the three fibrillar collagens differ in thickness and length, but they have a similar fine structure. As viewed by electron microscopy, they all have a characteristic pattern of cross-striations that are about one-quarter the length of the monomers and reflect the precise packing into fibrils. The three fibrillar collagens, however, differ in sequences found in the X- and Y-positions of the α chains and therefore in some of their physical properties. Type I collagen is composed of two identical α1(I) chains and a third α2(I) chain that differs slightly in its amino acid sequence. Type II collagen is composed of three identical α(II) chains. Type III collagen is also composed of three distinctive but identical α1(III) chains.

To deliver a monomer of the correct structure to the appropriate site of fibril assembly, the biosynthesis of fibrillar collagens involves a large number of unique processing steps **(Fig. 406-1)**. The monomer, first synthesized as a soluble precursor called *procollagen*, contains an additional globular domain at each end. As the pre-proα chains of procollagen are synthesized on ribosomes, the free N-terminal ends move into the cisternae of the rough endoplasmic reticulum (ER). Signal peptides at the N-termini are cleaved, and additional posttranslational reactions begin. Proline residues in the Y-position of the repeating -Gly-X-Y- sequences are converted to hydroxyproline by the enzyme prolyl hydroxylase. The hydroxylation of prolyl residues is

essential for the three α chains of the monomer to fold into a triple helix at body temperature. The enzyme requires ascorbic acid as one of its essential cofactors, an observation that explains why wounds fail to heal in scurvy **(Chap. 326)**. In scurvy, some of the underhydroxylated and unfolded protein accumulates in the cisternae of the rough ER and is degraded. Lysine residues in the Y-position are also hydroxylated to hydroxylysine by a separate lysyl hydroxylase. Many of the hydroxylysine residues are glycosylated with galactose or with galactose and glucose. A large mannose-rich oligosaccharide is assembled on the C-terminal propeptide of each chain. The proα chains are assembled by interactions among these C-terminal propeptides that control the selection of the appropriate partner chains to form hetero- or homotrimers and provide the correct chain registration required for subsequent formation of the collagen triple helix. After the C-terminal propeptides assemble the three proα chains, a nucleus of triple helix is formed near the C-terminus, and the helical conformation is propagated toward the N-terminus in a zipper-like manner that resembles crystallization. The folding into the triple helix is spontaneous in solution, but as discussed below, identification of rare mutations causing OI demonstrated that the folding *in cellulo* is assisted by a number of ancillary proteins which also prevent collagen fibril formation with the ER. The fully folded protein is then transported to the Golgi via a specific COPII vesicle process. After further modifications in the Golgi stack, the procollagen is secreted. After secretion, procollagen is processed to collagen by cleavage of the N-propeptides and C-propeptides by two specific proteinases. The release of the propeptides decreases the solubility of the protein about 1000-fold. The entropic energy that is released drives the self-assembly of the collagen into fibrils. Self-assembled collagen fibers have considerable tensile strength, but their strength is increased further by cross-linking reactions that form covalent bonds between α chains in one molecule and α chains in adjacent molecules. The resulting fibers, comprised of hundreds or thousands of triple-helical monomers, have some of the properties of a crystal but have innate imperfections that make them highly flexible.

FIGURE 406-1

Schematic summary of biosynthesis of fibrillar collagens. *(Modified and reproduced with permission from J Myllyharju, KI Kivirikko: Trends in Genetics 20:33, 2004.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although the assembly of collagen monomers into fibers is largely a spontaneous reaction, the process in tissues is modulated by the presence of less abundant collagens (Type V with Type I, and Type XI with Type II) and by other components such as a series of small leucine-rich proteins (SLRPs). Some of the less abundant components alter the rate of fibril assembly, whereas others change the morphology of the fibers or their interactions with cells and other molecules. The presence of these other components is one explanation for why, in some tissues, the fibers are further assembled into large tendons; in others, into sheets; and in still others into complex structures such as the hexagonal array of fibers that provide both the strength and transparency of the cornea.

Collagen fibers are resistant to most proteases, but during degradation of connective tissues, they are cleaved by specific matrix metalloproteinases (collagenases) that cause partial unfolding of the triple helices into gelatin-like structures that are further degraded by less specific proteinases.

OTHER COLLAGENS AND RELATED MOLECULES

The unique properties of the triple helix are used to define a family of at least 28 collagens that contain repetitive -Gly-X-Y- sequences and form triple helices of varying length and complexity. The proteins are heterogeneous both in structure and function, and many are the sites of mutations causing genetic diseases. For example, the Type IV collagen found in basement membranes is composed of three α chains synthesized from any of six different genes. Mutations in any of the six genes can cause AS.

Fibrillin Aggregates and Elastin

In addition to tensile strength, many tissues such as the lung, large blood vessels, and ligaments require elasticity. The elasticity was originally ascribed to an amorphous rubber-like protein named elastin. Subsequent analyses, largely sparked by discoveries of mutations causing the Marfan syndrome (MFS), demonstrated that the elasticity resided in thin fibrils composed primarily of large glycoproteins named fibrillins. The fibrillins contain large numbers of epidermal growth factor–like domains interspersed with characteristic cysteine-rich domains that are also found in latent transforming growth factor β (TGF-β) binding proteins. The fibrillins assemble into long beadlike strands that also contain numerous other components including small and variable amounts of elastin, bone morphogenic proteins (BMPs), and microfibril-associated glycoproteins (MAGPs). The principles whereby the fibrils provide elasticity to tissue and their biosynthetic assembly are still under investigation. As well as contributing to extracellular matrix structure, the fibrillins play a major role in TGF-β signaling.

Proteoglycans

The resiliency to compression of connective tissues such as cartilage or the aorta is largely explained by the presence of proteoglycans. Proteoglycans are composed of a core protein to which are attached a large series of negatively charged polymers of disaccharides (largely chondroitin sulfates). At least 30 proteoglycans have been identified. They vary in their binding to collagens and other components of matrix, but specific functions have not been assigned to most. The major proteoglycan of cartilage, called aggrecan, has a core protein of 2000 amino acids that is decorated with about 100 side chains of chondroitin sulfate and keratin sulfate. The core protein, in turn, binds to long chains of the polymeric disaccharide hyaluronan to form proteoglycan aggregates, one of the largest soluble macromolecular structures in nature. Because of its highly negative charge and extended structure, the proteoglycan aggregate binds large amounts of water and small ions to distend the three-dimensional arcade of collagen fibers found in the same tissues. It thereby makes the cartilage resilient to pressure.

SPECIFIC DISORDERS

OSTEOGENESIS IMPERFECTA

The central feature of OI is a severe decrease in bone mass that makes bones brittle. The disorder is frequently associated with blue sclerae, dental abnormalities (dentinogenesis imperfecta), progressive hearing loss, and a positive family history. Most patients have mutations in one of the two genes coding for Type I collagen.

Classification

OI was originally classified into two subtypes of congenita and tarda depending on the age of onset of the symptoms. Sillence suggested a series of subtypes based on clinical, radiological findings, and mode of inheritance. As with the other disorders discussed here, the description of rare recessive forms of OI and discovery of mutations in new genes has opened a debate as to whether the disorders should be classified by the clinical phenotypes or by the genes at fault. For the moment, the classification based on the clinical presentations seems the most useful (Table 406-2).

TABLE 406-2

Classification of Osteogenesis Imperfecta (OI)

Phenotype	Type	Typical Features	Inheritance	Gene/Protein Defect	Protein Defect
Non-deforming form	OI type 1	Mild to moderate bone fragility, normal or near normal stature, in most, blue sclerae, normal dentition in most hearing loss in ~50%.	AD	***COL1A1*** ***COL1A2***	Collagen I haploinsufficiency
Perinatally lethal form	OI type 2	Extreme bone fragility, short stature, long bone bowing	AD AR	***COL1A1*** ***COL1A2*** *CRTAP* *LEPRE1*/P3H1 *PPIB*/CYPB	Collagen I structural mutations Collagen post-translational modification and folding machinery
Progressively deforming OI	OI type 3	Moderate to severe bone deformity, blue sclerae at birth, hearing loss and abnormal dentition common.	AD AR	***COL1A1*** ***COL1A2*** *CRTAP* *LEPRE1*/P3H1 *PPIB*/CYPB *FKBP10*/FKBP65 *PLOD2*/ LH2 *SERPINH1*/HSP47 *CREB3L1*/OASIS *SEC24D* *BMP1* *WNT1* *SERPINF1*/PEDF *TMEM38B*/TRIC-B *SP7*/OSX	Collagen I structural mutations Collagen post-translational modification and folding machinery Protein folding/endoplasmic reticulum stress sensor COPII vesicle component collagen secretion Proteolytic removal of procollagen N-propeptide Wnt cell signaling pathway Signaling and collagen binding protein, important for mineralization Endoplasmic reticulum cation channel. Calcium homoestasis/collagen synthesis Transcription factor, bone formation

Phenotype	Type	Typical Features	Inheritance	Gene/Protein Defect	Protein Defect
Common Variable OI with normal sclerae	OI type 4	Mild to moderate, bone fragility, normal sclerae, variable dentition, hearing loss in <10%.	AD AR	***COL1A1*** ***COL1A2*** *WNT1* *CRTAP* *FKBP10*/FKBP65 *PPIB*/CYPB *SERPINF1*/PEDF *SPARC* *SP7*/OSX	Collagen I structural mutations Wnt cell signaling pathway Collagen post-translational modification and folding machinery Signaling and collagen binding protein, important for mineralization Collagen binding/extracellular matrix assembly Transcription factor, bone formation
OI with calcification of the interosseous membranes	OI type 5	Calcification of the interosseous membranes in forearm and legs and/or hypertrophic callus. Variable bone deformity, normal sclerae and dentition.	AD	*IFITM5*	Transcription factor, bone formation
Bruck syndrome type 1	BRKS1	Contractures with pterygia, fractures in infancy or early childhood, postnatal short stature, severe limb deformity, and progressive scoliosis	AR	*FKBP10*/FKBP65	Collagen folding machinery
Bruck syndrome type 2	BRKS2	As for Bruck syndrome type 1	AR	*PLOD2*/LH2	Collagen post-translational modification of lysine

Note: Predominant OI gene mutations (>90%) are in *COL1A1* and *COL1A2* (in bold typeface).

Abbreviations: AD, autosomal dominant; AR, autosomal recessive.

Type I is the mildest subtype and can produce either mild or no apparent deformities of the skeleton. Most patients have distinctly blue sclerae. Type II produces bone so brittle that it is lethal *in utero* or shortly after birth; it can be subclassified into Types II A, B, and C, depending on radiologic findings. Of the nonlethal forms, Type III is progressively deforming with moderate to severe bone deformity and Type IV (common variable OI with normal sclerae) has mild to moderate bone fragility.

The classifications of patients by types of OI do not consistently predict the clinical course of the disease. Some patients appear normal at birth and become progressively worse; others have multiple fractures in infancy and childhood, improve after puberty, and fracture more frequently later in life. Women are particularly prone to fracture during pregnancy and after menopause. A few women from families with mild variants of OI do not develop fractures until after menopause, and their disease may be difficult to distinguish from postmenopausal osteoporosis.

Incidence

Type I OI has a frequency of about 1 in 15,000–20,000 births. Type II OI has a reported incidence of about 1 in 60,000. Only a limited number of patients with the severe forms of OI have been reported, and the combined incidence of the severe forms that are recognizable at birth (Types II, III, and IV) may be much higher than 1 in 60,000.

Skeletal Effects

In Type I OI, the fragility of bones may be severe enough to limit physical activity or be so mild that individuals are unaware of any disability. Radiographs of the skull in patients with mild disease may show a mottled appearance because of small islands of irregular ossification. In Type II OI, ossification of many bones is frequently incomplete. Continuously beaded or broken ribs and crumpled long bones (accordina femora) may be present. For reasons that are not apparent, the long bones may be either thick or thin. In Types III and IV, multiple fractures from minor physical stress can produce severe deformities. Kyphoscoliosis can impair respiration, cause cor pulmonale, and predispose to pulmonary infections. The appearance of "popcorn-like" deposits of mineral in x-rays of the ends of long bones is an ominous sign. Progressive neurologic symptoms may result from basilar compression and communicating hydrocephalus. Type V OI is recognized by the presence of dislocated radial heads and hyperplastic callus formation.

In all forms of OI, bone mineral density is decreased. However, the degree of osteopenia may be difficult to evaluate because recurrent fractures limit exercise and thereby diminish bone mass. Surprisingly, fractures appear to heal normally.

Ocular Features

The sclerae can be normal, gray, slightly bluish, or bright blue. Blue sclerae, however, are an inherited trait in some families who do not have increased bone fragility.

Dentinogenesis

The teeth may be normal, moderately discolored, or grossly abnormal. The enamel generally appears normal, but the teeth may have a characteristic amber, yellowish brown, or translucent bluish gray color because of a deficiency of dentin that is rich in Type I collagen. The deciduous teeth are usually smaller than normal, whereas permanent teeth are frequently bell-shaped and restricted at the base. In some patients, the teeth readily fracture and need to be extracted. Similar tooth defects, however, can be inherited without any evidence of OI.

Hearing Loss

Hearing loss usually begins during the second decade of life and occurs in >50% of individuals aged >30. The loss can be conductive, sensorineural, or mixed, and it varies in severity. The middle ear usually exhibits maldevelopment, deficient ossification, persistence of cartilage in areas that are normally ossified, and abnormal calcium deposits.

Other Features

Changes in other connective tissues can include thin skin that scars extensively, joint laxity with permanent dislocations indistinguishable from those of EDS, and occasionally, cardiovascular manifestations such as aortic regurgitation, floppy mitral valves, mitral incompetence, and fragility of large blood vessels. For unknown reasons, some patients develop bouts of a hypermetabolic state with elevated serum thyroxine levels, hyperthermia, and excessive sweating.

Molecular Defects

 Of the ~1720 unique gene mutations now described in OI, >86% are heterozygous mutations in either *COL1A1* and *COL1A2*, the genes coding for proα1 or proα2 chain of Type I procollagen **(Table 406-2)**.

Most patients with Type I OI and blue sclerae have mutations that reduce the synthesis of proα1 chains to about one half. Mutations that reduce the synthesis of proα2 chains produce slightly more severe phenotypes and skin defects similar to EDS.

In contrast to the null mutations found in Type I OI, most of the severe variants (Types II, III, and IV) are caused by mutations that produce structurally abnormal proα chains that have compromised assembly or abnormal folding of the triple helix. As with collagen mutations in other connective tissue diseases, these structural mutations generally fall into two functional categories. The relatively rare mutations in the C-propeptide domain can prevent or seriously impair initial assembly of the procollagen trimers. These misfolded chains are retained in the ER and targeted for degradation by the ER-associated proteasomal pathway. Because these mutations induce an ER-stress response, the unfolded protein response (UPR) may have many downstream effects on cells. A possible role for ER stress in the pathology of OI is highlighted by the recent finding that a mutation in *CREB3L1*, the gene for OASIS, a member of the ATF6 family of ER stress sensors, produces severe recessive OI. ER stress is a new concept in the pathophysiology of connective tissue disease and has been best characterized for chondrodysplasias (CDs) (see below).

The most common Type I collagen mutations, however, are single base substitutions that introduce an amino acid with a bulky side chain for one of the glycine residues that appear as every third amino acid in the triple helix. In effect, any of the 338 glycine residues in the helical domain of either the proα1 or proα2 chain of Type I procollagen is a potential site for a disease-producing mutation. These mutations compromise the structural integrity of the triple helix, causing disruption to helix folding, retention of the mutant trimers in the ER, and increased posttranslational hydroxylation and

glycosylation of lysines. Collagen-containing helix mutations can form insoluble aggregates in the ER that are degraded by the autophagosome-endosome system, rather than the proteasomes. The precise nature of any resulting ER stress response remains to be characterized.

A similar sequence of events occurs with less common mutations that produce partial gene deletions, partial gene duplications, and splicing mutations. In addition to their intracellular effects, the structurally abnormal mutant-containing collagen that is secreted by the cell can also have important extracellular affects. For example, the presence of one abnormal proα chain in a procollagen molecule can interfere with cleavage of the N-propeptide from the protein. The persistence of the N-propeptide on a fraction of the molecules interferes with the self-assembly of normal collagen so that thin and irregular collagen fibrils are formed. Furthermore, if structurally abnormal collagens are incorporated into fibrils, they may have a destabilizing effect and be selectively degraded, or they may alter the interactions of collagen with other connective tissue components, disturbing architecture and stability.

Several generalizations can be made about mutations in Type I collagen genes. One is that unrelated patients rarely have the same mutation in the same gene. Glycine substitutions in the N-terminal region of the triple helix tend to produce milder phenotypes, apparently because they have less effect on the zipper-like propagation of the triple-helical conformation from the C terminus. Rare substitutions of charged amino acids (Asp, Arg) or a branched amino acid (Val) in X- or Y-positions produce lethal phenotypes, apparently because they are located at sites for lateral assembly of the monomers or binding of other components of the matrix.

The search for mutations causing the less common and autosomal recessive forms of OI identified mutations in genes for a series of proteins that are essential for the timely folding of the procollagen monomer: cartilage-associated protein (*CRTAP*), prolyl-3-hydroxylase (*LEPRE1*/P3H1), cyclophilin B (*PPIB*), collagen chaperone-like protein HSP47 (*SERPINH1*), and the procollagen chaperone protein FKBP65 (*FKBP10*). Recently, mutations have been characterized the collagen vesicular secretion pathway (*SEC24D*) and in additional downstream components of the collagen fibrillogenesis pathway; *BMP1*, the gene coding for a metalloproteinase that cleaves the C-propeptide of Type I procollagen; *SPARC* (osteonectin), a collagen binding protein involved in extracellular matrix assembly; and *PLOD2* (LH2, lysyl oxidase 2) which is involved in establishing collagen cross-links. In addition to these mutations that affect the collagen assembly pathway, mutations have been characterized in genes involved in the regulation of bone formation and mineralization such as *SP7* (osterix), *IFITM5*, *WNT1*, and *TMEM38B* (Table 406-2).

Inheritance and Mosaicism in Germ-Line Cells and in Somatic Cells

Type I OI is inherited as an autosomal dominant trait. However, some patients with Type I OI appear to represent sporadic new mutations or a diagnosis that was missed in earlier generations. Most lethal OI is the result of sporadic mutations that occur in the germ line in one of the parents. Because of the possibility for germ-line mosaicism for newly generated mutations, there is about a 7% probability that a second child could inherit a severe variant of OI.

Diagnosis

OI is usually diagnosed on the basis of clinical criteria. The presence of fractures together with blue sclerae, dentinogenesis imperfecta, or family history of the disease is usually sufficient to make the diagnosis. Other causes of pathologic fractures must be excluded, including battered child syndrome, nutritional deficiencies, malignancies, and other inherited disorders such as CDs and hypophosphatasia that can have overlapping presentations. The absence of superficial bruises can be helpful in distinguishing OI from battered child syndrome. X-rays usually reveal a decrease in bone density that can be verified by photon or x-ray absorptiometry. Bone microscopy can be helpful in the diagnosis. The diagnosis, like other genetic disorders, is now routinely conducted using targeted candidate gene sequencing but whole genome sequencing is becoming increasingly common.

TREATMENT

TREATMENT

Osteogenesis Imperfecta

Therapy should be directed toward decreasing the incidence of bone fractures, bone pain, and the restrictions on mobility. Physical therapy and occupational therapy are important. Diet should include adequate intake of calcium and vitamin D adjusted for the diminished weight of most patients. Orthopedic procedures are frequently required for deformities of long bone and scoliosis. Some surgeons recommend inserting rods into long bones. Drugs that have been developed for the therapy of osteoporosis are beneficial for some patients, but definitive data are difficult to obtain because of the small number of OI patients available for study. Bisphosphonates that inhibit osteoclasts are regarded as a mainstay of care in many medical centers for children with moderate to severe OI. Limited data are available on a series other therapies that are currently being tested: a monoclonal antibody that targets the receptor activator of nuclear factor-κB ligand (RANKL) to inhibit osteoclasts; inhibitors of cathepsin K to prevent digestion of proteins in the bone matrix; growth hormone to stimulate osteoblasts; a parathyroid analogue which stimulates osteoblasts; monoclonal antibodies to sclerostin that limit bone mass by inhibiting Wnt/β-catenin signaling in osteoblasts; and inhibitors to TGF-β that modulate osteoclast/osteoblast activity during bone remodeling. Cell therapy with infusion of mesenchymal stem/stromal cells improved a small cohort of children, but the improvements persisted for only a few months and the procedure has not been widely adopted.

EHLERS-DANLOS SYNDROME

EDS is characterized by hyper-extensible skin and hypermobile joints but the category includes rare patients with other distinctive features. Mutations in different types of collagen are found in many patients but other genes are at fault in rare forms. Contrary to initial expectations, no patients have been found with mutations in the gene for elastin in EDS.

Classification

Several types of EDS have been defined, based on the extent to which the skin, joints, and other tissues are involved, mode of inheritance, and by molecular and biochemical analysis **(Table 406-3)**. Classical EDS includes a severe form of the disease (Type I) and a milder form (Type II), both characterized by joint hypermobility and skin that is velvety in texture, hyper-extensible, and easily scarred. In hypermobile EDS (Type III), joint hypermobility is more prominent than skin changes. In vascular-type EDS (Type IV), the skin changes are more prominent than joint changes, and the patients are predisposed to sudden death from rupture of large blood vessels or other hollow organs. EDS V is similar to EDS II but is inherited as an X-linked trait. The ocular-scoliotic type of EDS (Type VI) is characterized by scoliosis, ocular fragility, and a cone-shaped deformity of the cornea (keratoconus). The arthrochalasic type of EDS (Type VII A and B) is characterized by marked joint hypermobility that is difficult to distinguish from EDS III except by the specific molecular defects in the processing of Type I procollagen to collagen. The periodontotic-type EDS (Type VIII) is distinguished by prominent periodontal changes. EDS IX, X, and XI were defined on the basis of preliminary biochemical and clinical data. EDS due to tenascin X deficiency has not been assigned a type; it is an autosomal recessive form of the syndrome similar to EDS II. The cardiac valvular form of EDS has similar features to EDS II, but also involves severe changes to the aorta. The progeroid form of EDS displays features of both EDS and progeria. Because of overlapping signs and symptoms, many patients and families with some of the features of EDS cannot be assigned to any of the defined types.

TABLE 406-3

Different Forms of Ehlers-Danlos Syndrome

Type	Typical Features	Inheritance	Gene Defect	Protein Defect
Classic (EDS I—severe and EDS II—mild)	Skin hyperextensibility and fragility, joint hypermobility, tissue fragility manifested by widened atrophic scarring	AD	*COL5A1* *COL5A2*	Collagen V
		AD	*COL1A1*	Proα1 (I) and proα2 (I) chains of procollagen I
		AD, AR	*COL1A2*	
Hypermobile (EDS III)	Joint hypermobility, moderate skin involvement, absence of tissue fragility	AD	*TNXB*	Tenascin X
Vascular (EDS IV)	Markedly reduced life span due to spontaneous rupture of internal organs such as arteries and intestines; skin is thin, translucent, and fragile, with extensive bruising; hypermobile minor joints; characteristic facial appearance	AD	*COL3A1*	Collagen III
Ocular-scoliotic EDS VI (EDS VIA and EDS VIB)	Features of classic EDS as well as severe muscular hypotonia after birth, progressive kyphoscoliosis, a Marfanoid habitus, osteopenia, occasionally rupture of the eye globe and great arteries	AR	*PLOD1*	Deficiency of procollagen-lysine 5-dioxygenase activity (EDS VIA)
			Unknown for EDS VIB	Unknown for EDS VIB
Arthrochalasic EDS VII (EDS VIIA and EDS VIIB)	Congenital bilateral hip dislocation, hypermobile joints, moderate skin involvement, osteopenia	AD	*COL1A1* *COL1A2*	Mutations that prevent cleavage of the N propeptides
Dermatosparactic EDS VII C	Redundant and fragile skin, prominent hernias, joint laxity, dysmorphic features	AR	*ADAMTS2*	Deficiency of procollagen I N-terminal proteinase
Periodontotic EDS VIII	Absorptive periodontosis with premature loss of permanent teeth, fragility of the skin, skin lesions	AD	C1R C1S	Components of the complement pathway
EDS due to tenascin X deficiency	Similar to EDS II	AR	*TNXB*	Tenascin X
EDS, cardiac valvular form	Similar to EDS II	AR	*COL1A2*	Type I collagen deficiency
EDS, progeroid form	Similar to EDS I-III with hair loss, hypotonia and aged appearance	AR	*B4GALT7*	Deficiency of galactosyltransferase 7 (defective synthesis of dermatan sulfate proteoglycans)

Type	Typical Features	Inheritance	Gene Defect	Protein Defect
EDS, musculocontractural form	Hyperextensible and thin skin, hypermobility and contractures of hands and feet, kyphoscoliosis	AR	*CHST14* *DSE*	Dermatan 4-*O*-sulfotransferase 1 (*CHST14*) and DS epimerase 1 (*DSE*) leading to defective synthesis of dermatan sulfate proteoglycans

Abbreviations: AD, autosomal dominant; AR, autosomal recessive.

Incidence

The overall incidence of EDS is about 1 in 5000 births, with a higher rate for blacks. Classical and hypermobile types of EDS are the most common. Patients with milder forms frequently do not seek medical attention.

Skin

Skin changes vary from thin and velvety to skin that is either dramatically hyperextensible ("rubber person" syndrome) or easily torn or scarred. Patients with classical EDS develop characteristic "cigarette-paper" scars. In vascular-type EDS, extensive scars and hyperpigmentation develop over bony prominences, and the skin may be so thin that subcutaneous blood vessels are visible. In the periodontotic type of EDS, the skin is more fragile than hyperextensible, and it heals with atrophic, pigmented scars. Easy bruisability occurs in several types of EDS.

Ligament and Joint Changes

Laxity and hypermobility of joints vary from mild to unreducible dislocations of hips and other large joints. In mild forms, patients learn to avoid dislocations by limiting physical activity. In more severe forms, surgical repair may be required. Some patients have progressive difficulty with age.

Other Features

Mitral valve prolapse and hernias occur, particularly with Type I. Pes planus and mild to moderate scoliosis are common. Extreme joint laxity and repeated dislocations may lead to degenerative arthritis. In the ocular-scoliotic type of EDS, the eye may rupture with minimal trauma, and kyphoscoliosis can cause respiratory impairment. Also, sclerae may be blue.

Molecular Defects

 Subsets of patients with different types of EDS have mutations in the structural genes for collagens (Table 406-3). These include mutations in the *COL1A1* gene in a few patients with moderately severe classical EDS (Type I); mutations in *COL1A2* in rare patients with an aortic valvular form of EDS; mutations in two of the three genes (*COL5A1* and *COL5A2*) for Type V collagen, a minor collagen found in association with Type I collagen, in about half the patients with classical EDS (Types I and II); mutations in the *COL3A1* gene for Type III collagen that is abundant in the aorta in patients with the frequently lethal vascular EDS (Type IV).

Some of the Type I collagen-related mutations alter processing of the protein or genes for the processing enzymes. Arthrochalasic EDS (Type VII) is caused by mutations in the amino acid sequence that make Type I procollagen resistant to cleavage by procollagen N-proteinase or by mutations that decrease the activity of the enzyme. The persistence of the N-propeptide causes the formation of collagen fibrils that are thin and irregular. Some of the patients have fragile bones and therefore a phenotype that overlaps with OI. The ocular-scoliotic type of EDS (Type VI) is caused by homozygous or compound heterozygous mutations in the *PLOD 1* gene, which encodes procollagen-lysine 5-dioxygenase (lysyl hydroxylase 1), an enzyme required for formation of stable cross-links in collagen fibers.

Some patients with the hypermobile EDS (Type III) and a few with mild EDS (Type II) have mutations in the *TNXB* gene, which encodes tenascin X, another minor component of connective tissue that appears to regulate the assembly of collagen fibers. Mutations in proteoglycans have been found in a few patients. The progeroid and musculocontractural forms of EDS result from mutations in the key enzymes of proteoglycan biosynthesis: *B4GALT7*, the gene for β-1, 4-galactosyltransferase 7, *CHST14*, dermatan 4-*O*-sulfotransferase 1 and *DSE*, dermatan sulfate epimerase.

Diagnosis

The diagnosis is based on clinical criteria and increasingly on DNA sequencing. Correlations between genotype and phenotype can be challenging, but gene or biochemical tests are particularly useful for the diagnosis of vascular-Type IV EDS with its dire prognosis.

As with other heritable diseases of connective tissue, there is a large degree of variability among members of the same family carrying the same mutation. Some patients have increased fractures and are difficult to distinguish from OI. A few families with heritable aortic aneurysms have mutations in the gene for Type III collagen without any evidence of EDS or OI.

TREATMENT

TREATMENT

Ehlers-Danlos Syndrome

Patients with mild forms require little special therapy. They, or their families, frequently learn how to reset dislocated joints. In severe forms, surgical repair and tightening of joint ligaments require careful evaluation of individual patients, as the ligaments frequently do not hold sutures. Patients with easy bruisability should be evaluated for bleeding disorders. Patients with Type IV EDS and members of their families should be evaluated at regular intervals for early detection of aneurysms, but surgical repair may be difficult because of friable tissues. Also, women with Type IV EDS should be counseled about the increased risk of uterine rupture, bleeding, and other complications of pregnancy.

CHONDRODYSPLASIAS

CDs, also referred to as skeletal dysplasias, are heritable skeletal disorders that are characterized by dwarfism and abnormal body proportions **(See also Chap. 405)**. The category also includes some individuals with normal stature and body proportions who have features such as ocular changes or cleft palate, which are common in more severe CDs. Many patients develop degenerative joint changes; and mild CD in adults may be difficult to differentiate from primary generalized OA. An undefined number of patients have mutations in either the most abundant collagen in cartilage (Type II) or the less abundant collagens (Types X or XI). Other patients have mutations in genes that code for other components of cartilage or for proteins required for the embryonic development of cartilage, including a common mutation in a gene for a fibroblast growth factor receptor.

Classification

Over 200 distinct types and subtypes have been defined based on criteria such as “bringing death” (thanatophoric), causing “twisted” bones (diastrophic), affecting metaphyses (metaphyseal), affecting epiphyses (epiphyseal), affecting spine (spondylo-), and producing histologic changes such as an apparent increase in the fibrous material in the epiphyses (fibrochondrogenesis). Also, a number of eponyms are based on the first or most comprehensive case reports. Severe forms of the diseases produce dwarfism with gross distortions of most cartilaginous structures and of other structures including the eye. Mild forms are more difficult to classify. Among the features are cataracts, degeneration of the vitreous, and retinal detachment, high forehead, hypoplastic facies, cleft palate, short extremities and gross distortions of the epiphyses, metaphyses, and joint surfaces. Patients with Stickler syndrome (hereditary arthro-ophthalmopathy) have been classified into three types based on a combination of the ocular phenotype and mutated genes.

Incidence

The overall incidence of all forms of CD ranges from 1 per 2500 to 1 per 4000 births. Data on the frequency of individual CDs are incomplete, but the incidence of the Stickler syndrome is 1 in 10,000. Therefore, the disease is probably among the more common heritable disorders of connective tissue.

Molecular Defects

 Mutations in the *COL2A1* gene for the Type II collagen of cartilage are found in a fraction of patients with both mild and severe CDs. For example, a mutation in the gene substituting a cysteine residue for an arginine was found in three unrelated families with spondyloepiphyseal dysplasia (SED) and precocious generalized OA. Mutations in the gene were also found in some lethal CDs characterized by gross deformities of bones and cartilage, such as those found in SED congenita, spondyloepimetaphyseal dysplasia congenita, hypochondrogenesis/achondrogenesis Type II, and Kniest syndrome. The highest incidence of *COL2A1* mutations, however, occurs in patients with the distinctive features of the Stickler syndrome, which is characterized by skeletal changes, orofacial abnormalities, and auditory abnormalities. Most of the mutations in *COL2A1* are premature stop codons that produce haploinsufficiency. In addition, some of the patients with the Stickler syndrome or a closely related syndrome have mutations in two genes specific for Type XI collagen, which is an unusual heterotrimer formed from α chains encoded by the gene for Type II collagen (*COL2A1*) and two distinctive genes for Type XI collagen (*COL11A1* and *COL11A2*). Mutations in the *COL11A1* gene are also found in patients with Marshall syndrome, which is similar to classic Stickler syndrome, but with more severe hearing loss and dysmorphic features, such as a flat or retracted midface with a flat nasal bridge, short nose, anteverted nostrils, long philtrum, and large-appearing eyes.

CDs are also caused by mutations in the less abundant collagens found in cartilage. For example, patients with Schmid metaphyseal CD have mutations in the gene for Type X collagen, a short, network-forming collagen found in the hypertrophic zone of endochondral cartilage. The syndrome is characterized by short stature, *coxa vara*, flaring metaphyses, and waddling gait. As with other collagen genes, the most common mutations are of two types: Nonsense mutations that lead to haploinsufficiency and structural mutations that compromise collagen assembly. In Type X collagen all the structural mutations detected occur in the C-terminal NC1 domain that coordinates the formation of the trimers. This NC1 domain is functionally equivalent to the C-propeptide of the fibrillar collagens. These mutations disturb the structure of the NC1 domain, leading to misfolding and initiation of cellular ER stress via the UPR. While the UPR evolved to allow cells to adjust their ER folding capacity to differing protein folding loads, it is deployed by cells when mutant misfolded proteins accumulate in the ER. Activation of the UPR attenuates protein translation and activates mutant protein degradation pathways such as ER-associated degradation. If these strategies do not sufficiently reduce the stress response, cell death may occur. In Schmid metaphyseal CD, mutant misfolded Type X collagen induces the UPR, resulting in downstream consequences that contribute to the pathophysiology. This general mechanism may also contribute to pathology in other chondrodyplasias (and in other connective tissues disorders) where gene mutations lead to protein structural abnormalities.

Some patients have mutations in genes for proteins that interact with collagens. Patients with pseudoachondroplasia or autosomal dominant multiple epiphyseal dysplasia have mutations in the gene for the cartilage oligomeric matrix protein (*COMP*), a protein that interacts with both collagens and proteoglycans in cartilage. However, some families with multiple epiphyseal dysplasia have a defect in one of the three genes for Type IX collagen (*COL9A1*, *COL9A2*, and *COL9A3*) or in matrilin-3, another extracellular protein found in cartilage. With misfolding mutations in COMP and matrilin-3, the activation of the UPR has been described, providing further evidence that the UPR is a component of pathology of these conditions.

Some CDs are caused by mutations in genes that affect early development of cartilage and related structures. The most common form of short-limbed dwarfism, achondroplasia, is caused by mutations in the gene for a receptor for a fibroblastic growth factor (*FGFR3*). The mutations in the *FGFR3* gene causing achondroplasias are unusual in several respects. The same single-base mutation in the gene that converts glycine to arginine at position 380 in the *FGFR3* gene is present in >90% of patients. Most patients harbor sporadic new mutations, and therefore this nucleotide change must be one of the most common recurring mutations in the human genome. The mutation causes unregulated signal transduction through the receptor and inappropriate development of cartilage. Mutations that alter other domains of *FGFR3* have been found in patients with the more severe disorders of hypochondroplasia and thanatophoric dysplasia and in a few families with a variant of craniosynostosis. However, most patients with craniosynostosis appear to have mutations in the related *FGFR2* gene. The similarities between the phenotypes produced by mutations in genes for FGF receptors and mutations in structural proteins of cartilage are probably explained by the observation that the activity of FGFs is regulated in part by binding of FGFs to proteins sequestered in the extracellular matrix. Therefore, the situation parallels the interactions between transforming growth factors (TGFs) and fibrillin in MFS (see below).

Other mutations involve the proteoglycans of cartilage, aggrecan (*AGC1*) and perlecan (*HSPG2*), and in the proteoglycan posttranslational sulphation pathway (*DTDST*, *PAPSS2* and *CHST3*). Mutations in >45 other genes have been defined in CDs.

Diagnosis

The diagnosis of CDs is made on the basis of the physical appearance, slit-lamp eye examinations, x-ray findings, histologic changes, and clinical course. Targeted gene and exome sequencing or more global sequencing strategies are used for molecular diagnosis. Given the wide spectrum of CD phenotypes, these genes tests are becoming critical diagnostic tools. For Stickler syndrome, more precise diagnostic criteria have made it possible to identify Type I variants with mutations in the *COL2A1* gene with a high degree of accuracy. It has been suggested that the Type II variant with mutations in the *COL11A1* gene can be identified on the basis of a “beaded” vitreous phenotype, and the Type III variant with mutations in the *COL11A2* gene can be identified on the basis of the characteristic systemic features without the ocular involvement. Prenatal diagnosis based on analysis of DNA obtained from chorionic villus or amniotic fluid is possible.

TREATMENT

TREATMENT

Chondrodysplasias

The treatment is symptomatic and is directed to secondary features such as degenerative arthritis. Many patients require joint replacement surgery and corrective surgery for cleft palate. The eyes should be monitored carefully for the development of cataracts and the need for laser therapy to prevent retinal detachment. In general, patients should be advised to avoid obesity and contact sports. Counseling for the psychological problems of short stature is critical. Several clinical trials therapeutically targeting the *FGFR3* pathway in achondroplasia are underway.

MARFAN SYNDROME (MFS)

MFS includes features that primarily affect the skeleton, the cardiovascular system, and the eyes. Most patients have mutations in the gene for fibrillin-1 (*FBN1*).

Classification

MFS was initially characterized by a triad of features: (1) skeletal changes that include long, thin extremities, frequently associated with loose joints; (2) reduced vision as the result of dislocations of the lenses (ectopia lentis); and (3) aortic aneurysms. An international panel has developed a series of revised “Ghent criteria” that are useful in classifying patients.

Incidence and Inheritance

The incidence of MFS is among the highest of any heritable disorder: about 1 in 3000/5000 births in most racial and ethnic groups. The related syndromes are less common. Mutations are generally inherited as autosomal dominant traits, but about one-fourth of patients have sporadic new mutations.

Skeletal Effects

Patients have long limbs and are usually tall compared to other members of the same family. The ratio of the upper segment (top of the head to the top of the pubic ramus) to the lower segment (top of the pubic ramus to the floor) is usually 2 SDs below mean for age, race, and sex. The fingers and hands are long and slender and have a spider-like appearance (arachnodactyly). Many patients have severe chest deformities, including depression (pectus excavatum), protrusion (pectus carinatum), or asymmetry. Scoliosis is frequent and usually accompanied by kyphosis. High-arched palate and high pedal arches or pes cavus are common. A few patients have severe joint hypermobility similar to EDS. CT or MRI examinations of the lumbar sacral region frequently reveals enlargement of the neural canal, thinning of the pedicles and laminae, widening of the foramins, or anterior meningocele (dural ectasia).

Cardiovascular Features

Cardiovascular abnormalities are the major source of morbidity and mortality (Chap. 274). Mitral valve prolapse develops early in life and progresses to mitral valve regurgitation of increasing severity in about one-quarter of patients. Dilation of the root of the aorta and the sinuses of Valsalva are characteristic and ominous features of the disease that can develop at any age. The rate of dilation is unpredictable, but it can lead to aortic regurgitation, dissection of the aorta, and rupture. Dilation is probably accelerated by physical and emotional stress, as well as by pregnancy. Patients usually differ from patients with familial aortic aneurysms who tend to develop aneurysms in the abdominal aorta. The location of the aneurysms, however, is somewhat variable, and the high incidence of aortic aneurysms in the general population (1 in 100) makes the differential diagnosis difficult unless other features of MFS are clearly present.

Ocular Features

Upward displacement of the lens is common. It is usually not progressive but may contribute to the formation of cataracts. The ocular globe is frequently elongated, and most patients are myopic, but with adequate vision. Retinal detachment can occur.

Other Features

Striae may occur over the shoulders and buttocks. A number of patients develop spontaneous pneumothorax. Inguinal and incisional hernias are common. Patients are typically thin with little subcutaneous fat, but adults may develop centripetal obesity.

Molecular Defects


More than 90% of patients clinically classified as having MFS by the “Ghent criteria” have a mutation in the gene for *FBN1*. Mutations in the same gene are found in a few patients who do not meet the Ghent criteria. Also, a few MFS patients without mutations in the *FBN1* gene have mutations in the gene for TGF-β receptor 2 (*TGFBR2*). In addition, mutations in either *TGFBR2* or *TGFBR1* are found in the related Loeys-Dietz syndrome which is characterized by aortic aneurysms, cleft palate, and hypertelorism. Mutations in the *FBN2* gene, which is structurally similar to the *FBN1* gene, are found in patients with MFS-like syndrome of congenital contractual arachnodactyly.

FBN1 gene mutations are scattered throughout its 65 coding exons. Most are private mutations, but ~10% are recurrent new mutations that are largely located in CpG sequences known to be “hot spots.” Most severe mutations are located in the central codons (24–32). About one-third of the mutations introduce premature termination codons, and about two-thirds are missense mutations that alter calcium-binding domains in the repetitive epidermal growth factor–like domains of the protein. Rarer mutations alter the processing of the protein. As in many genetic diseases, the severity of the phenotype cannot be predicted from the nature of the mutation.

The discovery that syndromes similar to MFS are caused by mutations in *TGFBR1* and *TGFBR2* refocused attention on structural similarity between *FBN1* and TGF-β binding proteins that sequester TGF-β in the extracellular matrix. As a result, some of the manifestations of MFS have been shown to arise from alterations in binding sites that modulate TGF-β bioavailability during development of the skeleton and other tissues. Likewise, *TGFBR1* and

TGFBR2 mutations in Loeys-Dietz syndrome alter TGF-β signaling. In both MFS and Loey-Dietz syndrome, the pathogenic mechanisms involve increased TGF-β signaling which contributes to aneurysm formation.

Diagnosis

All patients with a suspected diagnosis of MFS should have a slit-lamp examination and an echocardiogram. Also, homocystinuria should be ruled out by amino acid analysis of plasma **(Chap. 413)**. The diagnosis of MFS according to the international Ghent standards places emphasis on major criteria that include presence of at least four skeletal abnormalities: ectopia lentis; dilation of the ascending aorta with or without dissection; dural ectasia; and a blood relative who meets the same criteria, with or without a DNA diagnosis. A final diagnosis is based on a balanced assessment of the major criteria together with several minor criteria. The absence of ocular changes suggests the Loeys-Dietz syndrome, and the presence of contractures with some of the signs of OI suggests congential contractual arachnodactyly.

Diagnostic tests based on gene sequencing or detection of protein defects are available. These results are unlikely to alter the treatment or prognosis, but are helpful to inform the patients and families and to rapidly exclude the diagnosis in unaffected family members.

TREATMENT

TREATMENT

Marfan Syndrome

Patients should be advised that the risks are increased by severe physical exertion, emotional stress, and pregnancy. Surgical correction of the aorta, aortic valve, and mitral valve has been successful in many patients, but tissues are frequently friable. The scoliosis tends to be progressive and should be treated by mechanical bracing and physical therapy if >20° or by surgery if it progresses to >45°. Dislocated lenses rarely require surgical removal, but patients should be followed closely for retinal detachment.

Propranolol or other β-adrenergic blocking agents are used to lower blood pressure and thereby delay or prevent aortic dilation. The finding that MFS pathophysiology involves alterations in TGF-β signaling has raised the possibility of new therapeutic strategies. Attenuation of TGF-β signaling with agents such as angiotensin II receptor blockers (e.g., Losartan) was effective reducing aortic enlargement in animal studies and clinical trials are still in progress.

ELASTIN-RELATED DISEASES

Mutations in the elastin gene (*ELN*) have been found in patients with supravalvular aortic stenosis and skin that hangs in loose and redundant folds (cutis laxa). As indicated in Table 406-3, patients with several forms of EDS have similar changes in skin that were initially thought to reflect changes in elastin.

EPIDERMOLYSIS BULLOSA (EB)

EB has been defined as the category of heritable disorders involving skin that is specifically characterized by blistering as a result of friction. Using this criterion, it was possible to define subtypes by the ultrastructural layer of skin in which the cleavage and blistering occurred. These functional and anatomical criteria made it possible to establish that most patients with a specific subtype have mutations in genes coding for a structural protein or a cell adherence protein expressed in the corresponding layer of skin.

Classification and Incidence

The four major types of EB are: (1) EB simplex in which cleavage occurs within the epidermis, (2) junctional EB in which cleavage occurs within the lamina lucida, (3) dystrophic EB in which cleavage occurs within the sub-lamina densa, and (4) Kindler syndrome with a mixed level of cleavage in different layers. Patients are then separated into major and minor subtypes based on clinical features and analysis of mutations.

The incidence of EB in the United States is about 1 in 50,000.

Molecular Defects

 The distinctive anatomic locations in skin have made it possible to relate the clinical subtypes of EB to mutations for specific components. In EB simplex, mutations are found primarily in the genes for the major keratins of basal epithelial cells (keratins 5 and 14), and the cell adhesion proteins plectin, plakophilin-1, desmoplakin, and dystonin. Mutations in exophilin-5 and transglutaminase 5, both of which impact keratin filaments, have been reported. Patients with the related syndrome, epidermolytic ichthyosis, have mutations in keratin 1 and keratin 10. In junctional EB, mutations occur in Type XVII collagen, a laminin (laminin-332) and α6β4 integrin. In the severe syndrome of dystrophic EB, mutations are found in the gene that codes for

Type VII collagen that forms long loops anchoring the epidermis to the dermis. Patients with more complex features of what is classified as the Kindler syndrome have mutations in Kindlin-1, a. focal adhesion protein involved in integrin activation.

Diagnosis and Treatment

The diagnosis is based on skin that readily breaks and forms blisters from minor trauma. EB simplex is generally milder than junctional EB or dystrophic EB. Dystrophic EB variants usually have large and prominent scars. Precise classification within subtypes usually requires immunofluorescent mapping. DNA diagnostic tests have been developed as research tools but are not readily available. The treatment is symptomatic. Novel therapeutic approaches such as gene therapy, protein replacement therapy and cell therapy are being explored.

ALPORT SYNDROME

AS is an inherited disorder characterized by hematuria and several associated features. It was not initially considered as a disorder of connective tissue. However, the search for mutations in the genes coding for collagens found that most patients had mutations in collagen found in basement membranes (Type IV). Four forms of the AS are now recognized: (1) classic AS, which is inherited as an X-linked disorder with hematuria, sensorineural deafness, and conical deformation of the anterior surface of the lens (lenticonus); (2) an X-linked form associated with diffuse leiomyomatosis; (3) an autosomal recessive form; and (4) an autosomal dominant form. Both autosomal recessive and dominant forms can cause renal disease without deafness or lenticonus.

Incidence

The incidence of AS is about 1 in 10,000 births in the general population and as high as 1 in 5000 in some ethnic groups. About 80% of AS patients have the classical X-linked variant.

Molecular Defects

 Most patients have mutations in four of the six genes for the chains of Type IV collagen (*COL4A3*, *COL4A4*, *COL4A5*, and *COL4A6*). The genes for the proteins are arranged in tandem pairs on different chromosomes in an unusual head-to-head orientation and with overlapping promoters; i.e., the *COL4A1* and *COL4A2* genes are head-to-head on chromosome 13q34, the *COL4A3* and *COL4A4* genes are on chromosome 2q35–37, and the *COL4A5* and *COL4A6* genes are on chromosome Xq22. The X-linked variants are caused by either mutations in the *COL4A5* gene or by partial deletions of both of the adjacent *COL4A4* and *COL4A5* genes. The autosomal recessive variants are caused by mutations in either the *COL4A3* or *COL4A4* genes. The mutations responsible for the autosomal dominant variants are still unknown, but they have been mapped to the same locus as the *COL4A3* and *COL4A4* genes.

Diagnosis and Treatment

The diagnosis of classic AS is based on X-linked inheritance of hematuria, sensorineural deafness, and lenticonus. The lenticonus together with hematuria is pathognomonic of classic AS. The sensorineural deafness is primarily in the high-tone range. It can frequently be detected only by an audiogram and is usually not progressive. Because of the X-linked transmission, women are generally underdiagnosed and are usually less severely affected than men. The hematuria usually progresses to nephritis and may cause renal failure in late adolescence in affected males and at older ages in some women. Renal transplantation is usually successful.

FURTHER READING

Bateman JF et al: Genetic diseases of connective tissues: Cellular and extracellular effects of ECM mutations. Nat Rev Genet 10:173, 2009. [PubMed: 19204719]

Bonafe L et al: Nosology and classification of genetic skeletal disorders: 2015 revision. Am J Med Genet A 167A:2869, 2015. [PubMed: 26394607]

Forlino A, Marini JC: Osteogenesis imperfecta. Lancet 387:1657, 2016. [PubMed: 26542481]

Hynes RO: The evolution of metazoan extracellular matrix. J Cell Biol 196:671, 2012. [PubMed: 22431747]

Loeys BL et al: The revised Ghent nosology for the Marfan syndrome. J Med Genet 47:476, 2010. [PubMed: 20591885]

Malfait F, de Paepe A: The Ehlers-Danlos syndrome. Adv Exp Med Biol 802:129, 2014. [PubMed: 24443025]

Marom R et al: Pharmacological and biological therapeutic strategies for osteogenesis imperfecta. Am J Med Genet C Semin Med Genet 172:367, 2016. [PubMed: 27813341]

Prockop DJ, Kivirikko KI: Collagens: molecular biology, diseases, and potentials for therapy. Annu Rev Biochem 64:403–434, 1995. Review. [PubMed: 7574488]

Ramirez F, Dietz HC: Marfan's syndrome: From molecular pathogenesis to clinical treatment. Curr Opin Genet Dev 17:252, 2007. [PubMed: 17467262]

Ramirez F, Sakai LY: Biogenesis and function of fibrillin assemblies. Cell Tissue Res 339:71, 2010. [PubMed: 19513754]

Shapiro JR et al (eds): *Osteogenesis Imperfecta. A Translational Approach to Brittle Bone Disease*. London, Academic Press, 2014.

Uitto J et al: Molecular pathology of the basement membrane zone in heritable blistering diseases: The paradigm of epidermolysis bullosa. Matrix Biology 57–58:76, 2017.

Chapter 407: Hemochromatosis

Lawrie W. Powell

DEFINITION

Hemochromatosis results from a relatively common inherited genetic mutation in European populations. Once thought to be a single disease entity it is now known to be an iron-storage disorder with genetic heterogeneity but with a final common metabolic pathway resulting in inappropriately low production of the hormone hepcidin. This leads to an increase in intestinal iron absorption and the deposition of excessive amounts of iron in parenchymal cells with eventual tissue damage and organ failure. Thus, the term *hemochromatosis* now refers to a group of genetic diseases that predispose to iron overload, potentially leading to fibrosis and organ failure. Cirrhosis of the liver, diabetes mellitus, arthritis, cardiomyopathy, and hypogonadotropic hypogonadism are the major clinical manifestations.

The following terminology is widely accepted.

1. *Hereditary hemochromatosis* is most often caused by a mutant gene, termed *HFE*, which is tightly linked to the HLA-A locus on chromosome 6p. Persons who are homozygous for the mutation are at increased risk of iron overload and account for 80–90% of clinical hereditary hemochromatosis in persons of northern European descent. In such subjects, the presence of hepatic fibrosis, cirrhosis, arthropathy, or hepatocellular carcinoma constitutes iron overload–related disease. Rarer forms of non-*HFE* hemochromatosis are caused by mutations in other genes involved in iron metabolism **(Table 407-1)**. The disease can be recognized during its early stages when iron overload and organ damage are minimal. At this stage, the disease is best referred to as *early hemochromatosis* or *precirrhotic hemochromatosis*.

2. *Secondary iron overload* occurs as a result of an iron-loading anemia, such as thalassemia or sideroblastic anemia, in which erythropoiesis is increased but ineffective. In the acquired iron-loading disorders, massive iron deposits in parenchymal tissues can lead to the same clinical and pathologic features as in hemochromatosis.

TABLE 407-1

Classification of Iron Overload States

Hereditary Hemochromatosis	
Hemochromatosis, *HFE*-related (type 1)	
C282Y homozygosity	
C282Y/H63D compound heterozygosity	
Hemochromatosis, non-*HFE*-related	
Juvenile hemochromatosis (type 2A) (hemojuvelin mutations)	
Juvenile hemochromatosis (type 2B) (hepcidin mutation)	
Mutated transferrin receptor 2, *TFR2* (type 3)	
Mutated ferroportin 1 gene, *SLC11A3* (type 4)	
Acquired Iron Overload	
Iron-loading anemias Thalassemia major Sideroblastic anemia Chronic hemolytic anemias Transfusional and parenteral iron overload Dietary iron overload	Chronic liver disease Hepatitis C Alcoholic cirrhosis, especially when advanced Nonalcoholic steatohepatitis Porphyria cutanea tarda Dysmetabolic iron overload syndrome Post-portacaval shunting
Miscellaneous	
Iron overload in sub-Saharan Africa	
Neonatal iron overload	
Aceruloplasminemia	
Congenital atransferrinemia	

PREVALENCE

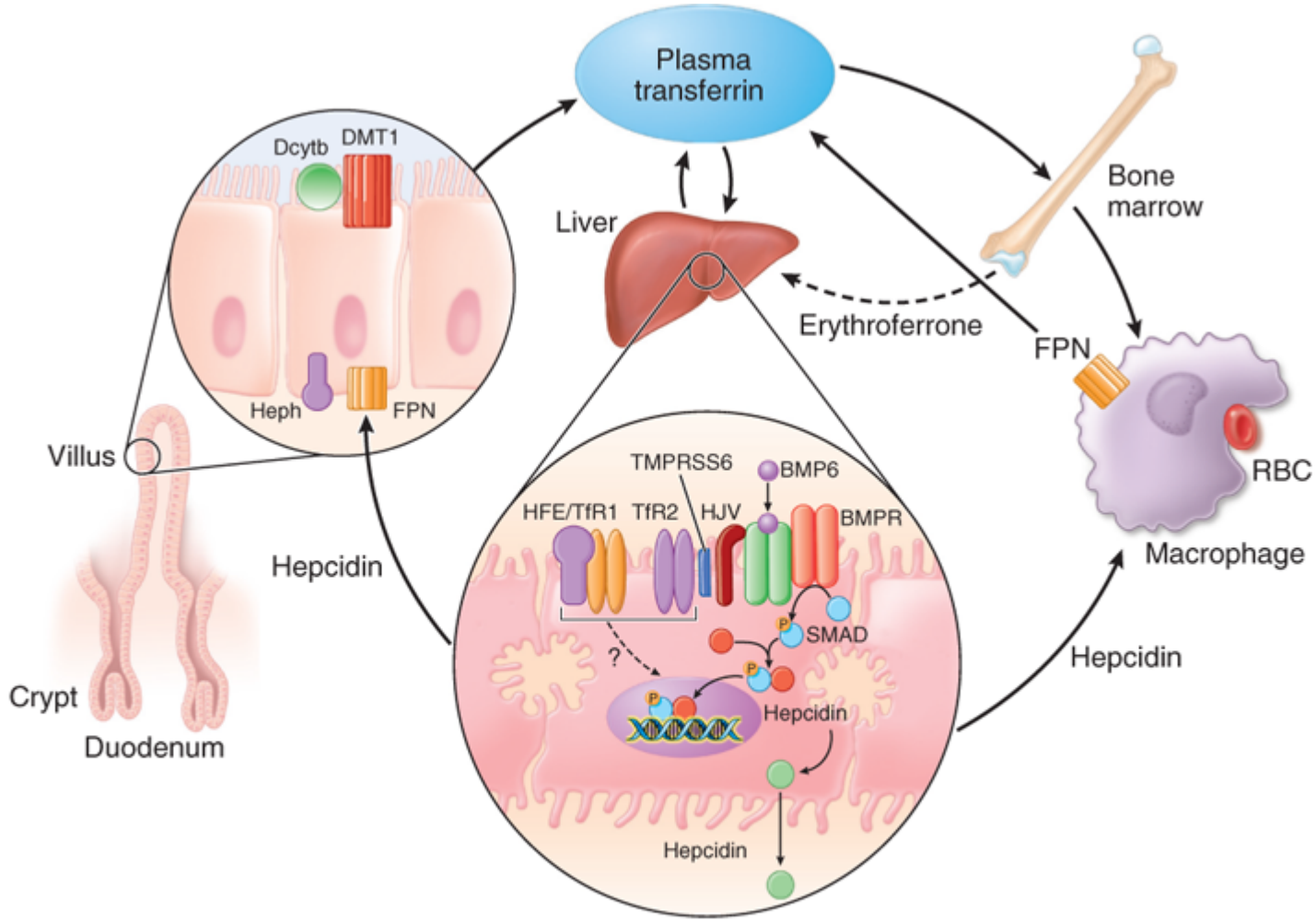
Although *HFE*-associated hemochromatosis mutations are common, the prevalence varies in different ethnic groups. It is most common in populations of northern European extraction in whom ~1 in 10 persons are heterozygous carriers and 0.3–0.5% are homozygotes. However, expression of the disease is variable and modified by several factors, especially alcohol consumption, dietary iron intake, blood loss associated with menstruation and pregnancy, and blood donation. Recent population studies indicate that ~30% of homozygous men develop iron overload–related disease and about 6% develop hepatic cirrhosis; for women, the figure is closer to 1%. Presumably there are as yet unidentified modifying genes responsible for expression. Nearly 70% of untreated patients develop the first symptoms between ages 40 and 60. The disease is rarely evident before age 20, although with family screening (see "Screening for Hemochromatosis," below) and periodic health examinations, asymptomatic subjects with iron overload can be identified, including young menstruating women.

In contrast to *HFE*-associated hemochromatosis, the non-*HFE*-associated forms of hemochromatosis (Table 407-1) are rare, but they affect all races and young people (juvenile hemochromatosis).

These result from mutations in one or more of the genes for proteins in the hepcidin pathway (**Fig. 407-1**), i.e., hemojuvelin, transferrin receptor 2 (TfR2), or ferroportin. The resultant clinical disease is very similar to HFE-related disease because they all lead to hepcidin deficiency, which is the final common pathway (Fig. 407-1).

FIGURE 407-1

Pathways of normal iron homeostasis. Dietary inorganic iron traverses the brush border membrane of duodenal enterocytes via the divalent metal-ion transporter 1 (DMT1) after reduction of ferric (Fe^{3+}) iron to the ferrous (Fe^{2+}) state by duodenal cytochrome B (DcytB). Iron then moves from the enterocyte to the circulation via a process requiring the basolateral iron exporter ferroportin (FPN) and the iron oxidase hephaestin (Heph). In the circulation, iron binds to plasma transferrin and is thereby distributed to sites of iron utilization and storage. Much of the diferric transferrin supplies iron to immature erythrocyte cells in the bone marrow for hemoglobin synthesis. At the end of their life, senescent red blood cells (RBCs) are phagocytosed by macrophages, and iron is returned to the circulation after export through ferroportin. The liver-derived peptide hepcidin represses basolateral iron transport in the gut as well as iron released from macrophages and other cells and serves as a central regulator of body-iron traffic. At least three separate signals regulate hepcidin production in response to changes in body-iron requirements. The first involves the detection of circulating diferric transferrin by HFE and TfR2. A second relies on hepatic iron stores activating the hemojuvelin (HJV)-dependent bone morphogenetic protein (BMP)/SMAD pathway. The third involves signaling molecules released from erythroid precursor cells and there is strong evidence that erythroferrone fulfills this role. TMPRSS6 is a protease that modulates HJV activity. Heme is metabolized by heme oxygenase within the enterocytes, and the released iron then follows the same pathway. Mutations in the genes encoding HFE, TfR2, hemojuvelin, and hepcidin all lead to decreased hepcidin release and increased iron absorption, resulting in hemochromatosis (Table 407-1).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A rare autosomal dominant form of hemochromatosis results from two types of mutations in the gene for the iron transporter ferroportin. Firstly, loss of function mutations decrease the cell surface localization of ferroportin, thereby reducing its ability to export iron ("ferroportin disease"). A second mutation abolishes the hepcidin-induced ferroportin internalization and degradation resulting in a "gain-of-function." Here the tissue iron distribution is similar to that in HFE-related disease (e.g., in parenchymal cells).

GENETIC BASIS

A homozygous G to A mutation in the *HFE* gene resulting in a cysteine to tyrosine substitution at position 282 (C282Y) is the most common mutation. It is identified in 85–90% of patients with hereditary hemochromatosis in populations of northern European descent but is found in only 60% of cases from Mediterranean populations. A second, relatively common *HFE* mutation (H63D) results in a substitution of histidine to aspartic acid at codon 63. Homozygosity for H63D is not associated with clinically significant iron overload. Some compound heterozygotes (e.g., one copy each of C282Y and H63D) have mild to moderately increased body-iron stores but develop clinical disease only in association with cofactors such as heavy

alcohol intake or hepatic steatosis. Thus, *HFE*-associated hemochromatosis is inherited as an autosomal recessive trait; heterozygotes have no, or minimal, increase in iron stores. However, this slight increase in hepatic iron can act as a cofactor that may modify the expression of other diseases such as porphyria cutanea tarda (PCT) or nonalcoholic steatohepatitis (NASH).

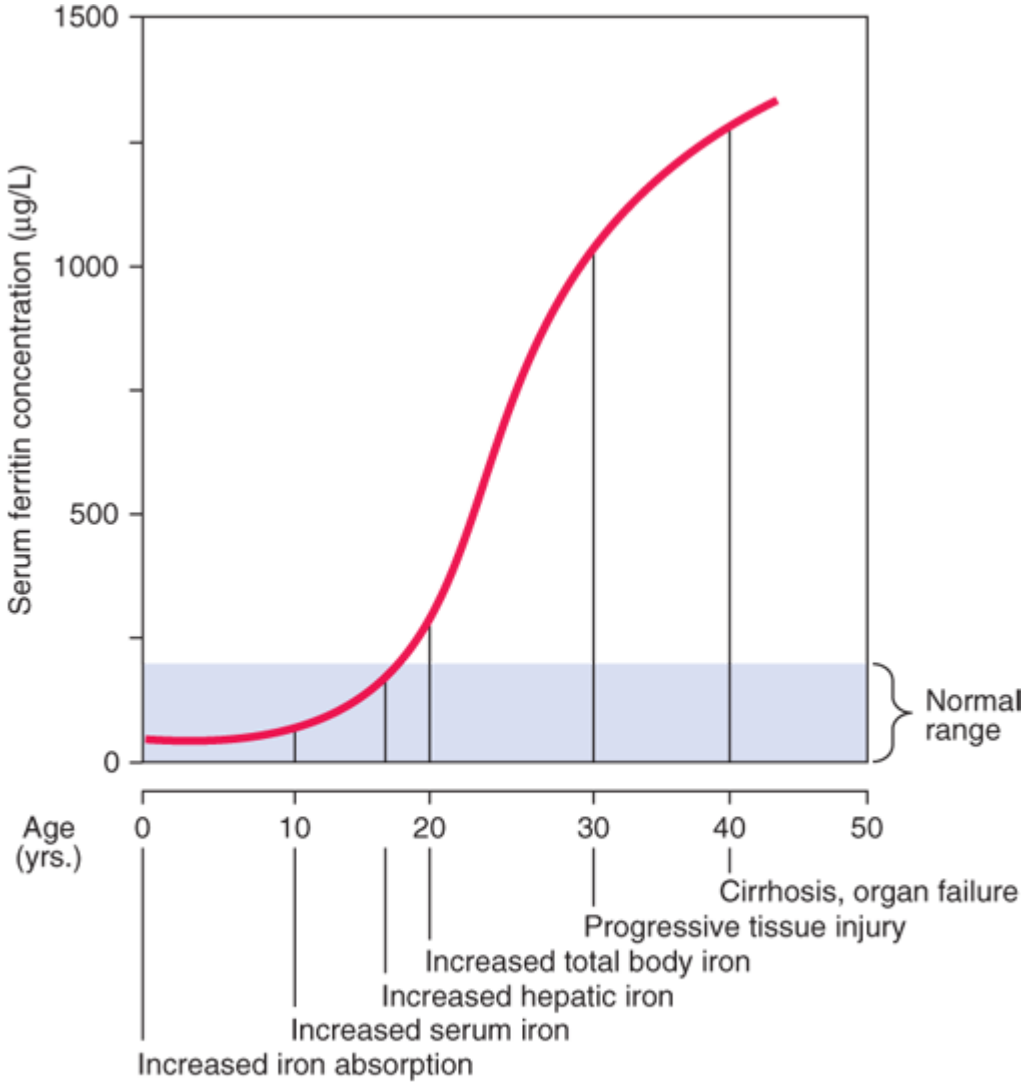
Thus, mutations in other genes involved in iron metabolism are responsible for non-*HFE*-associated hemochromatosis, including juvenile hemochromatosis, which affects persons in the second and third decades of life (Table 407-1). Mutations in the genes encoding hepcidin, transferrin receptor 2 (TfR2), and hemojuvelin (Fig. 407-1) result in clinicopathologic features that are indistinguishable from HFE-associated hemochromatosis. However, mutations in ferroportin, responsible for the efflux of iron from enterocytes and most other cell types, result in iron loading of reticuloendothelial cells and macrophages as well as parenchymal cells.

PATHOPHYSIOLOGY AND THE ROLE OF HEPCIDIN

Normally, the body-iron content of 3–4 g is maintained such that intestinal mucosal absorption of iron is equal to iron loss. This amount is ~1 mg/d in men and 1.5 mg/d in menstruating women. In hemochromatosis, mucosal absorption is greater than body requirements and amounts to ≥4 mg/d. The progressive accumulation of iron increases plasma iron and saturation of transferrin and results in a progressive increase of plasma ferritin **(Fig. 407-2)**. The key regulatory hormone that allows the liver to communicate with the bone marrow was discovered quite serendipitously and has transformed our understanding of the coordination of absorption, mobilization, and storage of iron to meet the requirements of erythropoiesis. It was called hepcidin based upon its anti-bacterial activity ("**HEPatic** bacterio**CID**al prote**IN**"). This liver-derived peptide represses basolateral iron transport in the intestine and iron release from macrophages and other cells by binding to ferroportin. Hepcidin, in turn, responds to signals in the liver mediated by HFE, TfR2, and hemojuvelin (Fig. 407-1). The development of minihepcidins, i.e., small peptides that mimic the action of hepcidin, is promising for the development of new therapeutic approaches for iron overload disorders caused by low hepcidin levels.

FIGURE 407-2

Sequence of events in genetic hemochromatosis and their correlation with the serum ferritin concentration. Increased iron absorption is present throughout life. Overt, symptomatic disease usually develops between ages 40 and 60, but latent disease can be detected long before this.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The *HFE* gene encodes a 343-amino-acid protein that is structurally related to MHC class I proteins (HFE). The basic defect in *HFE*-associated hemochromatosis is a lack of cell surface expression of HFE (due to the C282Y mutation). The normal (wild-type) HFE protein forms a complex with β_2-microglobulin and transferrin receptor 1 (TfR1). The C282Y mutation completely abrogates this interaction. As a result, the mutant HFE protein remains trapped intracellularly, reducing TfR1-mediated iron uptake by the intestinal crypt cell. This impaired TfR1-mediated iron uptake leads to upregulation of the divalent metal transporter (DMT1) on the brush border of the villus cells, causing inappropriately increased intestinal iron absorption (Fig. 407-1). In advanced disease, the body may contain 20 g or more of iron that is deposited mainly in parenchymal cells of the liver, pancreas, and heart. Iron deposition in the pituitary causes hypogonadotropic hypogonadism in both men and women. Tissue injury may result from disruption of iron-laden lysosomes, from lipid peroxidation of subcellular organelles by excess iron, or from stimulation of collagen synthesis by activated stellate cells.

Secondary iron overload with deposition in parenchymal cells occurs in chronic disorders of erythropoiesis, particularly in those due to defects in hemoglobin synthesis or ineffective erythropoiesis such as sideroblastic anemia and thalassemia **(Chap. 94)**. In these disorders, iron absorption is increased. Moreover,

these patients require blood transfusions and are frequently treated inappropriately with iron. PCT, a disorder characterized by a defect in porphyrin biosynthesis **(Chap. 409)**, can also be associated with excessive parenchymal iron deposits. The magnitude of the iron load in PCT is usually insufficient to produce tissue damage. However, some patients with PCT also have mutations in the *HFE* gene, and some have associated hepatitis C virus (HCV) infection. Although the relationship between these disorders remains to be clarified, iron overload accentuates the inherited enzyme deficiency in PCT and should be avoided along with other agents (alcohol, estrogens, haloaromatic compounds) that may exacerbate PCT. Another cause of hepatic parenchymal iron overload is hereditary aceruloplasminemia. In this disorder, impairment of iron mobilization due to deficiency of ceruloplasmin (a ferroxidase) causes iron overload in hepatocytes.

Excessive iron ingestion over many years rarely results in hemochromatosis. An important exception has been reported in South Africa among groups who brew fermented beverages in vessels made of iron (see later). Hemochromatosis has been described in apparently normal persons who have taken medicinal iron over many years, but such individuals probably had genetic disorders.

The common denominator in all patients with hemochromatosis is *excessive amounts of iron in parenchymal tissues*. Parenteral administration of iron in the form of blood transfusions or iron preparations results predominantly in reticuloendothelial cell iron overload. This appears to lead to less tissue damage than iron loading of parenchymal cells.

In the liver, parenchymal iron is in the form of ferritin and hemosiderin. In the early stages, these deposits are seen in the periportal parenchymal cells, especially within lysosomes in the pericanalicular cytoplasm of the hepatocytes. This stage progresses to perilobular fibrosis and to fibrous septa due to activation of stellate cells. In the advanced stage, a macronodular or mixed macro- and micronodular cirrhosis develops. Hepatic fibrosis and cirrhosis correlate significantly with hepatic iron concentration.

Histologically, iron is increased in many organs, particularly in the liver, heart, and pancreas, and, to a lesser extent, in the endocrine glands. The epidermis of the skin is thin, and melanin is increased in the cells of the basal layer and dermis. Deposits of iron are present around the synovial lining cells of the joints.

CLINICAL MANIFESTATIONS

C282Y homozygotes can be characterized by the stage of progression as follows: (1) a genetic predisposition without abnormalities; (2) iron overload without symptoms; (3) iron overload with symptoms (e.g., arthritis and fatigue); and (4) iron overload with organ damage—in particular, cirrhosis. Thus, many subjects with significant iron overload are asymptomatic. For example, in a study of 672 asymptomatic C282Y homozygous subjects—identified by either family screening or routine health examinations—there was hepatic iron overload (grades 2–4) in 56% and 34.5% of male and female subjects, respectively; hepatic fibrosis (stages 2–4) in 18.4% and 5.4%, respectively; and cirrhosis in 5.6% and 1.9%, respectively.

Initial symptoms are often nonspecific and include lethargy, arthralgia, skin pigmentation, loss of libido, and features of diabetes mellitus. Hepatomegaly, increased pigmentation, spider angiomas, splenomegaly,

arthropathy, ascites, cardiac arrhythmias, congestive heart failure, loss of body hair, testicular atrophy, and jaundice are prominent in advanced disease.

The *liver* is usually the first organ to be affected, and hepatomegaly is present in >95% of symptomatic patients.

Manifestations of portal hypertension and esophageal varices occur less commonly than in cirrhosis from other causes. Hepatocellular carcinoma develops in ~30% of patients with cirrhosis, and it is the most common cause of death in treated patients—hence the importance of early diagnosis and therapy. The incidence increases with age, it is more common in men, and it occurs almost exclusively in cirrhotic patients.

Excessive skin pigmentation is present in patients with advanced disease. The characteristic metallic or slate-gray hue is sometimes referred to as *bronzing* and results from increased melanin and iron in the dermis. Pigmentation usually is diffuse and generalized.

Diabetes mellitus occurs in ~65% of patients with advanced disease and is more likely to develop in those with a family history of diabetes, suggesting that direct damage to the pancreatic islets by iron deposition occurs in combination with other risk factors. The management is similar to that of other forms of diabetes.

Arthropathy develops in 25–50% of symptomatic patients. It usually occurs after age 50 but may occur as a first manifestation or long after therapy. The joints of the hands, especially the second and third metacarpophalangeal joints, are usually the first joints involved, a feature that helps to distinguish the chondrocalcinosis associated with hemochromatosis from the idiopathic form (Chap. 365). A progressive polyarthritis involving wrists, hips, ankles, and knees may also ensue. Acute brief attacks of synovitis may be associated with deposition of calcium pyrophosphate (chondrocalcinosis or pseudogout), mainly in the knees. Radiologic manifestations include cystic changes of the subchondral bones, loss of articular cartilage with narrowing of the joint space, diffuse demineralization, hypertrophic bone proliferation, and calcification of the synovium. The arthropathy tends to progress despite removal of iron by phlebotomy. Although the relation of these abnormalities to iron metabolism is not known, the fact that similar changes occur in other forms of iron overload suggests that iron is directly involved.

Cardiac involvement is the presenting manifestation in ~15% of symptomatic patients. The most common manifestation is congestive heart failure, which occurs in ~10% of young adults with the disease, especially those with juvenile hemochromatosis. Symptoms of congestive heart failure may develop suddenly, with rapid progression to death if untreated. The heart is diffusely enlarged; this may be misdiagnosed as idiopathic cardiomyopathy if other overt manifestations are absent. Cardiac arrhythmias include premature supraventricular beats, paroxysmal tachyarrhythmias, atrial flutter, atrial fibrillation, and varying degrees of atrioventricular block.

Hypogonadism occurs in both sexes and may antedate other clinical features. Manifestations include loss of libido, impotence, amenorrhea, testicular atrophy, gynecomastia, and sparse body hair. These changes are

primarily the result of decreased production of gonadotropins due to impairment of hypothalamic-pituitary function by iron deposition.

DIAGNOSIS

The association of (1) hepatomegaly, (2) skin pigmentation, (3) diabetes mellitus, (4) heart disease, (5) arthritis, and (6) hypogonadism should suggest the diagnosis. However, as stated above, significant iron overload may exist with none or only some of these manifestations. Therefore, a high index of suspicion is needed to make the diagnosis early. Treatment before permanent organ damage occurs can reverse the iron toxicity and restore life expectancy to normal.

The history should be particularly detailed in regard to disease in other family members; alcohol ingestion; iron intake; and ingestion of large doses of ascorbic acid, which promotes iron absorption **(Chap. 326)**. Appropriate tests should be performed to exclude iron deposition due to hematologic disease. The presence of liver, pancreatic, cardiac, and joint disease should be confirmed by physical examination, radiography, and standard function tests of these organs.

The degree of increase in total body iron stores can be assessed by (1) measurement of serum iron and the percent saturation of transferrin (or the unsaturated iron-binding capacity), (2) measurement of serum ferritin concentration, (3) liver biopsy with measurement of the iron concentration and calculation of the hepatic iron index **(Table 407-2)**, and (4) magnetic resonance imaging (MRI) of the liver. In addition, a retrospective assessment of body-iron storage is also provided by performing weekly phlebotomy and calculating the amount of iron removed before iron stores are exhausted (1 mL blood = ~0.5 mg iron).

TABLE 407-2

Representative Iron Values in Normal Subjects, Patients with Hemochromatosis, and Patients with Alcoholic Liver Disease

Determination	Normal	Symptomatic Hemochromatosis	Homozygotes with Early, Asymptomatic Hemochromatosis	Heterozygotes	Alcoholic Liver Disease
Plasma iron, μmol/L (μg/dL)	9–27 (50–150)	32–54 (180–300)	Usually elevated	Elevated or normal	Often elevated
Total iron-binding capacity, μmol/L (μg/dL)	45–66 (250–370)	36–54 (200–300)	36–54 (200–300)	Elevated or normal	45–66 (250–370)
Transferrin saturation, %	22–45	50–100	50–100	Normal or elevated	27–60
Serum ferritin, μg/L		1000–6000	200–500	Usually <500	10–500
Men	20–250				
Women	15–150				
Liver iron, μg/g dry wt	300–1400	6000–18,000	2000–4000	300–3000	300–2000
Hepatic iron index	<1.0	>2	1.5–2	<2	<2

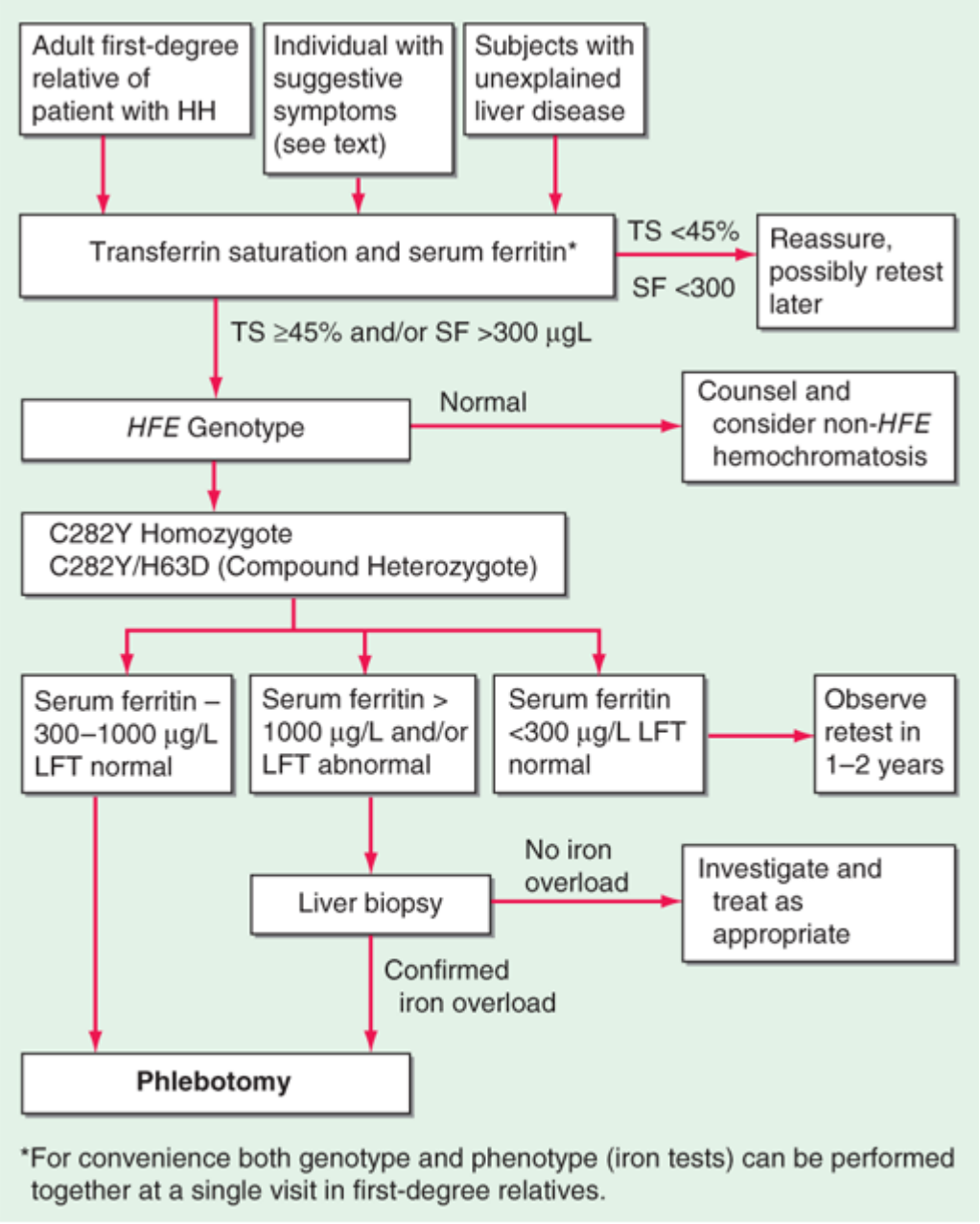
Each of these methods for assessing iron stores has advantages and limitations. The serum iron level and percent saturation of transferrin are elevated early in the course, but their specificity is reduced by significant false-positive and false-negative rates. For example, serum iron concentration may be increased in patients with alcoholic liver disease without iron overload; in this situation, however, the hepatic iron index is usually not increased as in hemochromatosis (Table 407-1). In otherwise healthy persons, a fasting serum transferrin saturation >45% is abnormal and suggests homozygosity for hemochromatosis.

The serum ferritin concentration is usually a good index of body-iron stores, whether decreased or increased. In fact, an increase of 1 μg/L in serum ferritin level reflects an increase of ~5 mg in body stores. In most

untreated patients with hemochromatosis, the serum ferritin level is significantly increased (Fig. 407-2 and Table 407-1), and a serum ferritin level >1000 µg/L is the strongest predictor of disease expression among individuals homozygous for the C282Y mutation. However, in patients with inflammation and hepatocellular necrosis, serum ferritin levels may be elevated out of proportion to body-iron stores due to increased release from tissues. Therefore, a repeat determination of serum ferritin should be carried out after acute hepatocellular damage has subsided (e.g., in alcoholic liver disease). Ordinarily, the combined measurements of the percent transferrin saturation and serum ferritin level provide a simple and reliable screening test for hemochromatosis, including the precirrhotic phase of the disease. If either of these tests is abnormal, genetic testing for hemochromatosis should be performed **(Fig. 407-3)**.

FIGURE 407-3

Algorithm for screening for *HFE*-associated hemochromatosis. HH, hereditary hemochromatosis, homozygous subject (C282Y +/+); LFT, liver function tests; SF, serum ferritin concentration; TS, transferrin saturation.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The role of liver biopsy in the diagnosis and management of hemochromatosis has been reassessed as a result of the widespread availability of genetic testing for the C282Y mutation. The absence of severe fibrosis can be accurately predicted in most patients using clinical and biochemical variables. Thus, there is virtually no risk of severe fibrosis in a C282Y homozygous subject with (1) serum ferritin level <1000 µg/L, (2) normal serum alanine aminotransferase values, (3) no hepatomegaly, and (4) no excess alcohol intake. However, it should be emphasized that liver biopsy is the only reliable method for establishing or excluding the presence of hepatic cirrhosis, which is the critical factor determining prognosis and the risk of developing hepatocellular carcinoma. Biopsy also permits histochemical estimation of tissue iron and measurement of hepatic iron concentration. Increased density of the liver due to iron deposition can be demonstrated by computed tomography (CT) or MRI, and with improved technology, MRI has become more accurate in determining hepatic iron concentration.

SCREENING FOR HEMOCHROMATOSIS

When the diagnosis of hemochromatosis is established, it is important to counsel and screen other family members **(Chap. 457)**. Asymptomatic and symptomatic family members with the disease usually have an increased saturation of transferrin and an increased serum ferritin concentration. These changes occur even before the iron stores are greatly increased (Fig. 407-2). All adult first-degree relatives of patients with hemochromatosis should be tested for the C282Y and H63D mutations and counseled appropriately (Fig. 407-3). In affected individuals, it is important to confirm or exclude the presence of cirrhosis and begin therapy as early as possible. For children of an identified proband, testing for *HFE* of the other parent is helpful because if normal, the child is merely an obligate heterozygote and at no risk. Otherwise, for practical purposes, children need not be checked before they are 18 years old.

The role of population screening for hemochromatosis is controversial. Recent studies indicate that it is highly effective for primary care physicians to screen subjects using transferrin saturation and serum ferritin levels. Such screening also detects iron deficiency. Genetic screening of the normal population is feasible but is probably not cost effective.

TREATMENT

TREATMENT

Hemochromatosis

The therapy of hemochromatosis involves removal of the excess body iron and supportive treatment of damaged organs. Iron removal is best accomplished by weekly or twice-weekly phlebotomy of 500 mL. Although there is an initial modest decline in the volume of packed red blood cells to about 35 mL/dL, the level stabilizes after several weeks. The plasma transferrin saturation remains increased until the available iron stores are depleted. In contrast, the plasma ferritin concentration falls progressively, reflecting the gradual decrease in body-iron stores. One 500-mL unit of blood contains 200–250 mg of iron, and ≥25 g of iron may have to be removed. Therefore, in patients with advanced disease, weekly phlebotomy may be required for 1–2 years, and it should be continued until the serum ferritin level is ≤100 μg/L. Thereafter, phlebotomies are performed at appropriate intervals to maintain ferritin levels at ≤100 μg/L. The transferrin saturation fluctuates and may still be elevated but should not dictate further therapy unless it is persistently at 100% when free unbound iron may circulate. Usually one phlebotomy every 3 months will suffice. It is important, however, not to overtreat and render the patient iron-deficient.

Chelating agents such as deferoxamine, when given parenterally, remove 10–20 mg of iron per day, which is much less than that mobilized by once-weekly phlebotomy. Phlebotomy is also less expensive, more convenient, and safer for most patients. However, chelating agents are indicated when anemia or hypoproteinemia is severe enough to preclude phlebotomy. Subcutaneous infusion of deferoxamine using a portable pump is the most effective means of its administration.

An effective oral iron chelating agent, deferasirox (Exjade), is now available. This agent is effective in thalassemia and secondary iron overload, but it is expensive and carries the risk of significant side effects.

Alcohol consumption should be severely curtailed or eliminated because it increases the risk of cirrhosis in hereditary hemochromatosis nearly tenfold. Dietary adjustments are unnecessary, although vitamin C and iron supplements should be avoided. The management of hepatic failure, cardiac failure, and diabetes mellitus is similar to conventional therapy for these conditions. Loss of libido and change in secondary sex characteristics are managed with testosterone replacement or gonadotropin therapy **(Chap. 384)**.

End-stage liver disease may be an indication for liver transplantation, although results are improved if the excess iron can be removed beforehand. The available evidence indicates that the fundamental metabolic abnormality in hemochromatosis is reversed by successful liver transplantation.

PROGNOSIS

The principal causes of death are cardiac failure, hepatocellular failure or portal hypertension and hepatocellular carcinoma.

Life expectancy is improved by removal of the excessive stores of iron and maintenance of these stores at near-normal levels. The 5-year survival rate with therapy increases from 33 to 89%. With repeated phlebotomy, the liver decreases in size, liver function improves, pigmentation of skin decreases, and cardiac failure may be reversed. Diabetes improves in ~40% of patients, but removal of excess iron has little effect on hypogonadism or arthropathy. Hepatic fibrosis may decrease, but established cirrhosis is irreversible. Hepatocellular carcinoma occurs as a late sequela in patients who are cirrhotic at presentation. The apparent increase in its incidence in treated patients is probably related to their increased life span. Hepatocellular carcinoma rarely develops if the disease is treated in the precirrhotic stage. Indeed, the life expectancy of homozygotes treated before the development of cirrhosis is normal.

The importance of family screening and early diagnosis and treatment cannot be overemphasized. Asymptomatic individuals detected by family studies should have phlebotomy therapy if iron stores are moderately to severely increased. Assessment of iron stores at appropriate intervals is also important. With this management approach, most manifestations of the disease can be prevented.

ROLE OF *HFE* MUTATIONS IN OTHER LIVER DISEASES



There is considerable interest in the role of *HFE* mutations and hepatic iron in several other liver diseases. Several studies have shown an increased prevalence of *HFE* mutations in PCT patients. Iron accentuates the inherited enzyme deficiency in PCT and clinical manifestations of PCT. The situation in NASH is less clear, but some studies have shown an increased prevalence of *HFE* mutations in NASH patients. The role of phlebotomy therapy, however, is unproven despite an intriguing fall in liver enzyme levels. In chronic HCV infection, *HFE* mutations are not more common, but some subjects have increased hepatic iron. Before

initiating antiviral therapy in these patients, it is reasonable to perform phlebotomy therapy to remove excess iron stores, because this reduces liver enzyme levels.

HFE mutations are not increased in frequency in alcoholic liver disease. Hemochromatosis in a heavy drinker can be distinguished from alcoholic liver disease by the presence of the C282Y mutation.

End-stage liver disease may also be associated with iron overload of the degree seen in hemochromatosis. The mechanism is uncertain, although studies have shown that alcohol suppresses hepatic hepcidin secretion. Hemolysis also plays a role. *HFE* mutations are uncommon.

Whether subjects homozygous for C282Y are at increased risk of breast and colorectal cancer is controversial.

GLOBAL CONSIDERATIONS



The *HFE* mutation is of northern European origin (Celtic or Nordic) with a heterozygous carrier rate of ~1 in 10 (1 in 8 in Ireland). Thus, *HFE*-associated hemochromatosis is quite rare in non-European populations, e.g., Asia. However, non-*HFE*-associated hemochromatosis resulting from mutations in other genes involved in iron metabolism (Fig. 407-1) is ubiquitous and should be considered when one encounters iron overload.

African iron overload occurs primarily in sub-Saharan Africa and was previously thought to be due to the consumption of an iron-rich fermented maize beverage. However, recent evidence suggests that it is primarily the result of a non-*HFE*-related genetic trait that is exacerbated by the dietary iron loading. A similar form of iron-overload has been described in African Americans. Further research is needed to clarify this condition.

FURTHER READING

Bardou-Jacquet E et al: Decreased cardiovascular and extrahepatic cancer-related mortality in treated patients with mild HFE hemochromatosis. J Hepatol 62:682, 2015. [PubMed: 25450707]

Guo S et al: Iron homeostasis: Transport, metabolism, and regulation. Curr Opin Clin Nutr Metab Care 19:276, 2016. [PubMed: 27137899]

Powell LW et al: Haemochromatosis. Lancet 388:706, 2016. [PubMed: 26975792]

Chapter 408: Wilson's Disease

George J. Brewer

FIGURE 408-1

INTRODUCTION

Wilson's disease is an autosomal recessive disorder caused by mutations in the *ATP7B* gene, which encodes a membrane-bound, copper-transporting ATPase. Clinical manifestations are caused by copper toxicity and primarily involve the liver and the brain. Because effective treatment is available, it is important to make this diagnosis early.

The frequency of Wilson's disease in most populations is about 1 in 30,000–40,000, and the frequency of carriers of *ATP7B* mutations is ~1%. Siblings of a diagnosed patient have a 1 in 4 risk of Wilson's disease, whereas children of an affected patient have about a 1 in 200 risk. Although a large number of inactivating mutations have been reported in the *ATP7B* gene, mutation screening for diagnosis is now available, and a definitive diagnosis can be obtained in 50–75% of patients, depending on the population. DNA haplotype analysis can be used to genotype siblings of an affected patient. A rare multisystem disorder of copper metabolism with features of both Menke's and Wilson's diseases has been reported. It is termed the MEDNIK (mental retardation, enteropathy, deafness, neuropathy, ichthyosis, keratodermia) syndrome and is caused by mutations in the *AP1S1* gene, which encodes an adaptor protein necessary for intracellular trafficking of copper pump proteins ATP7A (Menke's disease) and ATP7B (Wilson's disease).

PATHOGENESIS

ATP7B protein deficiency impairs biliary copper excretion, resulting in positive copper balance, hepatic copper accumulation, and copper toxicity from oxidant damage. Excess hepatic copper is initially bound to metallothionein; liver damage begins as this storage capacity is exceeded, sometimes by 3 years of age. Defective copper incorporation into apoceruloplasmin leads to excess catabolism and low blood levels of ceruloplasmin. Serum copper levels are usually lower than normal because of low blood levels of ceruloplasmin, which normally binds >90% of serum copper. As the disease progresses, nonceruloplasmin serum copper ("free" copper) levels increase, resulting in copper buildup in other parts of the body (e.g., in the brain, with consequent neurologic and psychiatric disease).

CLINICAL PRESENTATION

Hepatic Features

Wilson's disease may present as hepatitis, cirrhosis, or hepatic decompensation. Patients typically present in the mid- to late teenage years in Western countries, although the age of presentation is quite broad and extends into the fifth decade of life. An episode of hepatitis may occur—with elevated serum aminotransferase levels, with or without jaundice—and then spontaneously regress. Hepatitis often recurs, and most of these patients eventually develop cirrhosis. Hepatic decompensation is associated with elevated serum bilirubin, reduced serum albumin and coagulation factors, ascites, peripheral edema, and hepatic encephalopathy. In severe hepatic failure, hemolytic anemia may develop because large amounts of copper derived from hepatocellular necrosis are released into the bloodstream. The association of hemolysis and liver disease makes Wilson's disease a likely diagnosis.

Neurologic Features

The neurologic manifestations of Wilson's disease typically occur in patients in their early twenties, although the age of onset extends into the sixth decade of life. MRI and CT scans reveal damage in the basal ganglia and occasionally in the pons, medulla, thalamus, cerebellum, and subcortical areas. The three main movement disorders include dystonia, incoordination, and tremor. Dysarthria and dysphagia are common. In some patients, the clinical picture closely resembles that of Parkinson's disease. Dystonia can involve any part of the body and eventually leads to grotesque positions of the limbs, neck, and trunk. Autonomic disturbances may include orthostatic hypotension and sweating abnormalities as well as bowel, bladder, and sexual dysfunction. Memory loss, migraine-type headaches, and seizures may occur. Patients have difficulty focusing on tasks, but cognition usually is not grossly impaired. Sensory abnormalities and muscular weakness are not features of the disease.

Psychiatric Features

Half of patients with neurologic disease have a history of behavioral disturbances with onset in the 5 years before diagnosis. The features are diverse and may include loss of emotional control (temper tantrums, crying bouts), depression, hyperactivity, or loss of sexual inhibition.

Other Manifestations

Some female patients have repeated spontaneous abortions, and most become amenorrheic prior to diagnosis. Cholelithiasis and nephrolithiasis occur with increased frequency. Some patients have osteoarthritis, particularly of the knee. Microscopic hematuria is common, and levels of urinary excretion of phosphates, amino acids, glucose, or urates may increase; however, a full-blown Fanconi's syndrome is rare. Sunflower cataracts and Kayser-Fleischer rings (copper deposits in the outer rim of the cornea) may be seen. Electrocardiographic and other cardiac abnormalities have been reported but are not common.

DIAGNOSIS

Diagnostic tests for Wilson's disease are listed in **Table 408-1**. Serum ceruloplasmin levels should not be used for definitive diagnosis, because they are normal in up to 10% of affected patients and are reduced in 20% of carriers. Kayser-Fleischer rings **(Fig. 408-1)** can be definitively diagnosed only by an ophthalmologist using a slit lamp. They are present in >99% of patients with neurologic/psychiatric forms of the disease and have been described very rarely in the absence of Wilson's disease. Kayser-Fleischer rings are present in only ~30–50% of patients diagnosed in the hepatic or presymptomatic state; thus, the absence of rings does not exclude the diagnosis.

TABLE 408-1

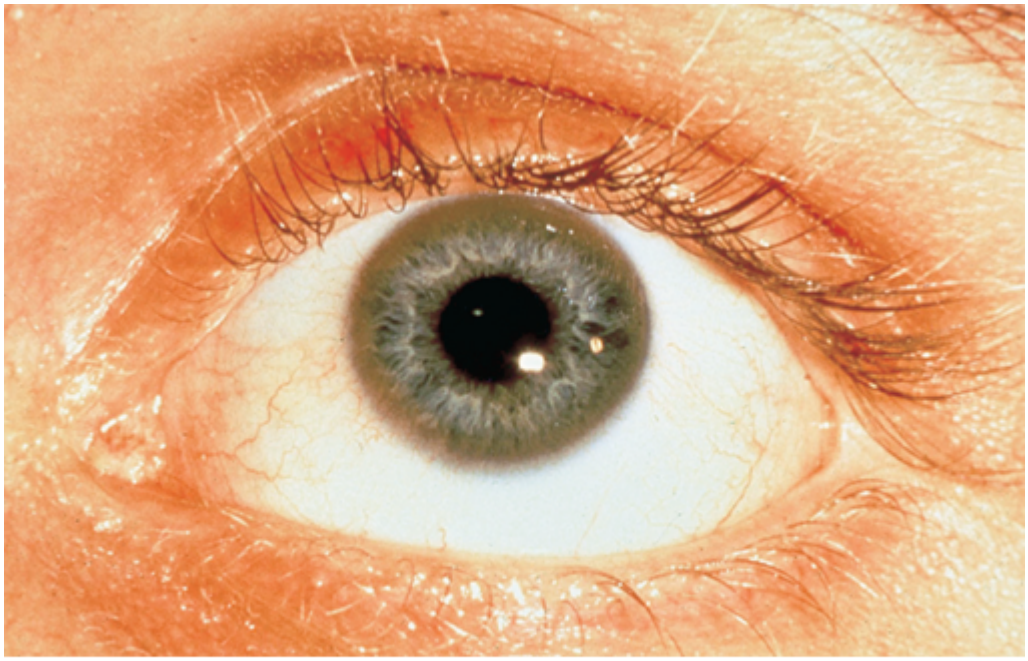
Useful Tests for Wilson's Disease

Test	Usefulness[a]	Normal Value	Heterozygous Carriers	Wilson's Disease
Serum ceruloplasmin	+	180–350 mg/L (18–35 mg/dL)	Low in 20%	Low in 90%
Kayser-Fleischer rings	++	Absent	Absent	Present in >99% if neurologic or psychiatric symptoms are present Present in 30–50% in hepatic presentation and presymptomatic state
Urine copper (24 h)	+++	0.3–0.8 µmol (20–50 µg)	Normal to 1.3 µmol (80 µg)	>1.6 µmol (>100 µg) in symptomatic patients; 0.9 to >1.6 µmol (60 to >100 µg) in presymptomatic patients
Liver copper	++++	0.3–0.8 µmol/g (20–50 µg/g of tissue)	Normal to 2.0 µmol (125 µg)	>3.1 µmol (>200 µg) (Obstructive liver disease can cause false-positive results.)
DNA testing of *ATP7B* gene	+++	No mutations	One mutation	Two mutations
Haplotype analysis	++++ (siblings only)	0 matches	1 match	2 matches

[a]Usefulness range: + (somewhat useful) to ++++ (very useful).

FIGURE 408-1

A Kayser-Fleischer ring. Although in this case, the brownish ring rimming the cornea is clearly visible to the naked eye, confirmation is usually made by slit-lamp examination.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Urine copper measurement is an important diagnostic tool, but urine must be collected carefully to avoid contamination. Symptomatic patients invariably have urine copper levels >1.6 µmol (>100 µg) per 24 h. Heterozygotes have values <1.3 µmol (<80 µg) per 24 h. About half of presymptomatic patients who are ultimately affected have diagnostically elevated urine copper values, but the other half have levels that are in an intermediate range between 0.9 and 1.6 µmol (60–100 µg) per 24 h. Because heterozygotes may have values up to 1.3 µmol (80 µg) per 24 h, patients in this range may require a liver biopsy for definitive diagnosis.

Testing for mutations in *ATP7B* can be done and is recommended for a definitive diagnosis. Several hundred inactivating mutations have been described and genetic testing laboratories can now make a definitive diagnosis in about 50–75% of patients, depending on the population.

The gold standard for diagnosis remains liver biopsy with quantitative copper assays. Affected patients have values >3.1 µmol/g (>200 µg/g [dry weight] of liver). Copper stains are not reliable. False-positive results can occur with long-standing obstructive liver disease, which can elevate hepatic and urine copper concentrations and rarely causes Kayser-Fleischer rings.

TREATMENT

TREATMENT

Wilson's Disease

Recommended anticopper treatments are listed in **Table 408-2**. Penicillamine was previously the primary anticopper treatment but now plays only a minor role because of its toxicity and because it often worsens existing neurologic disease if used as initial therapy. If penicillamine is given, it should always be

accompanied by pyridoxine (25 mg/d). Trientine is a less toxic chelator and is supplanting penicillamine when a chelator is indicated.

For patients with hepatitis or cirrhosis but without evidence of hepatic decompensation or neurologic/psychiatric symptoms, zinc is the therapy of choice although some experts advocate therapy with trientine. Zinc has proven efficacy in Wilson's disease and is essentially nontoxic. It produces a negative copper balance by blocking intestinal absorption of copper, and it induces hepatic metallothionein synthesis, thereby sequestering additional toxic copper. All presymptomatic patients should be treated prophylactically because the disease is close to 100% penetrant.

The first step in evaluating patients presenting with hepatic decompensation is to establish disease severity, which can be estimated with the Nazer prognostic index **(Table 408-3)**. Patients with scores <7 can usually be managed with medical therapy. Patients with scores >9 should be considered immediately for liver transplantation. For patients with scores between 7 and 9, clinical judgment is required in deciding whether to recommend transplantation or medical therapy. A combination of trientine and zinc has been used to treat patients with Nazer scores as high as 9, but such patients should be watched carefully for indications of hepatic deterioration, which mandates transplantation.

For initial medical treatment of patients with hepatic decompensation, the recommended regimen is a chelator (preferably trientine) plus zinc (Table 408-2). Zinc should not, however, be ingested simultaneously with trientine, which chelates zinc and forms therapeutically ineffective complexes. Administration of the two drugs should be separated by at least 1 h.

For initial neurologic therapy, tetrathiomolybdate is emerging as the drug of choice because of its rapid control of free copper, preservation of neurologic function, and low toxicity. Penicillamine and trientine should be avoided because both have a high risk of worsening the neurologic condition. Until tetrathiomolybdate is commercially available, zinc therapy is recommended. Although it is relatively slow-acting, zinc itself does not exacerbate neurologic abnormalities. Although hepatic transplantation may alleviate neurologic symptoms, it does so only by copper removal, which can be done more safely and inexpensively with anticopper drugs. Pregnant patients should be treated with zinc or trientine throughout pregnancy but without tight copper control because copper deficiency can be teratogenic.

Anticopper therapy must be lifelong. With treatment, liver function usually recovers after about a year although residual liver damage is usually present. Neurologic and psychiatric symptoms usually improve after 6–24 months of treatment.

TABLE 408-2

Recommended Anticopper Drugs for Wilson's Disease

Disease Status	First Choice	Second Choice
Initial hepatic		
Hepatitis or cirrhosis without decompensation	Zinc[a]	Trientine
Hepatic decompensation		
Mild	Trientine[b] and zinc	Penicillamine[b] and zinc
Moderate	Trientine and zinc	Hepatic transplantation
Severe	Hepatic transplantation	Trientine and zinc
Initial neurologic/psychiatric	Tetrathiomolybdate[c] and zinc	Zinc
Maintenance	Zinc	Trientine
Presymptomatic	Zinc	Trientine
Pediatric	Zinc	Trientine
Pregnant	Zinc	Trientine

[a]Zinc acetate is supplied as Galzin, manufactured by Gate Pharmaceutical. The recommended adult dose for all the above indications is 50 mg of elemental zinc three times daily, with each dose separated by at least 1 h from consumption of food and beverages other than water as well as from trientine or penicillamine doses. [b]Trientine is supplied as Syprine and penicillamine as Cuprimine, both manufactured by Merck. The recommended adult dosage for both drugs is 500 mg twice daily, with each dose at least 0.5 h before or 2 h after meals and separated by at least 1 h from zinc administration. [c]Tetrathiomolybdate is being studied in clinical trials.

TABLE 408-3

Prognostic Index of Nazer

Laboratory Measurement	Normal Value	Score (in Points)				
		0	1	2	3	4
Serum bilirubin[a]	0.2–1.2 mg/dL	<5.8	5.8–8.8	8.8–11.7	11.7–17.5	>17.5
Serum aspartate aminotransferase	10–35 IU/L	<100	100–150	151–200	201–300	>300
Prolongation of prothrombin time (s)	—	<4	4–8	9–12	13–20	>20

[a]If hemolysis is present, serum bilirubin cannot be used as a measure of liver function until the hemolysis subsides.

Source: Modified from H Nazer et al: Gut 27:1377, 1986; with permission from BMJ Publishing Group.

MONITORING ANTICOPPER THERAPY

When trientine or penicillamine is first used, it is necessary to monitor for drug toxicity, particularly bone marrow suppression and proteinuria. Complete blood counts, standard biochemical profiles, and a urinalysis should be performed at weekly intervals for 1 month, then at twice-weekly intervals for 2 or 3 months, then at monthly intervals for 3 or 4 months, and at 4- to 6-month intervals thereafter.

The anticopper effects of trientine and penicillamine can be monitored by following “free” serum copper levels. Changes in urine copper levels are more difficult to interpret because excretion reflects the effect of the drug as well as body loading with copper. Free serum copper is calculated by subtracting the ceruloplasmin copper from the total serum copper. Each 10 mg/L (1 mg/dL) of ceruloplasmin contributes 0.5 μmol/L (3 μg/dL) of serum copper. The normal serum-free copper value is 1.6–2.4 μmol/L (10–15 μg/dL); the level is often as high as 7.9 μmol/L (50 μg/dL) in untreated Wilson’s disease. With treatment, the serum-free copper should be <3.9 μmol/L (<25 μg/dL).

Zinc treatment does not require monitoring of blood or urine for toxicity. Its only significant side effect is gastric burning or nausea in ~10% of patients, usually with the first morning dose. This effect can be mitigated if the first dose is taken an hour after breakfast or if zinc is taken with a small amount of protein. Because zinc mainly affects stool copper, 24-h urine copper can be used to reflect body loading. The typical value in untreated symptomatic patients is >3.1 μmol (>200 μg) per 24 h. This level should decrease during

the first 1–2 years of therapy to <2.0 µmol (<125 µg) per 24 h. A normal value (0.3–0.8 µmol [20–50 µg]) is rarely reached during the first decade of therapy and should raise concern about overtreatment (copper deficiency), the first sign of which is anemia and/or leukopenia.

GLOBAL CONSIDERATIONS



The age of onset of clinical disease may be considerably younger in India and the Far East; in these regions, onset often occurs in children at only 5 or 6 years of age. The incidence of the disease may be increased in certain populations as a result of founder effects. For example, in Sardinia, the incidence may be 1 in 3000. In countries where penicillamine, trientine, and zinc acetate (as Galzin) are not available or are unaffordable, zinc salts such as gluconate or sulfate provide an alternative treatment option.

FURTHER READING

Askari FK et al: Treatment of Wilson's disease with zinc. XVIII. Initial treatment of the hepatic decompensation presentation with trientine and zinc. J Lab Clin Med 142:385, 2003. [PubMed: 14713890]

Brewer GJ: *Wilson's Disease: A Clinician's Guide to Recognition, Diagnosis, and Management*. Boston, Kluwer Academic Publishers, 2001.

Brewer GJ et al: Treatment of Wilson disease with ammonium tetrathiomolybdate: III. Initial therapy in a total of 55 neurologically affected patients and follow-up with zinc therapy. Arch Neurol 60:379, 2003. [PubMed: 12633149]

Brewer GJ et al: Treatment of Wilson's disease with zinc: XV long-term follow-up studies. J Lab Clin Med 132:264, 1998. [PubMed: 9794697]

Martinelli D et al: MEDNIK syndrome: A novel defect of copper metabolism treatable by zinc acetate therapy. Brain 136:872, 2013. [PubMed: 23423674]

Nazer H et al: Wilson's disease: Clinical presentation and use of prognostic index. Gut 27:1377, 1986. [PubMed: 3792921]

Scheinberg IH, Sternlieb I: *Wilson's Disease (Major Problems in Internal Medicine, Volume 23)*. LHJ Smith (ed). Philadelphia, W.B. Saunders Company, 1984.

Schilsky ML: Wilson disease: Genetic basis of copper toxicity and natural history. Semin Liver Dis 16:83, 1996. [PubMed: 8723326]

Chapter 409: The Porphyrias

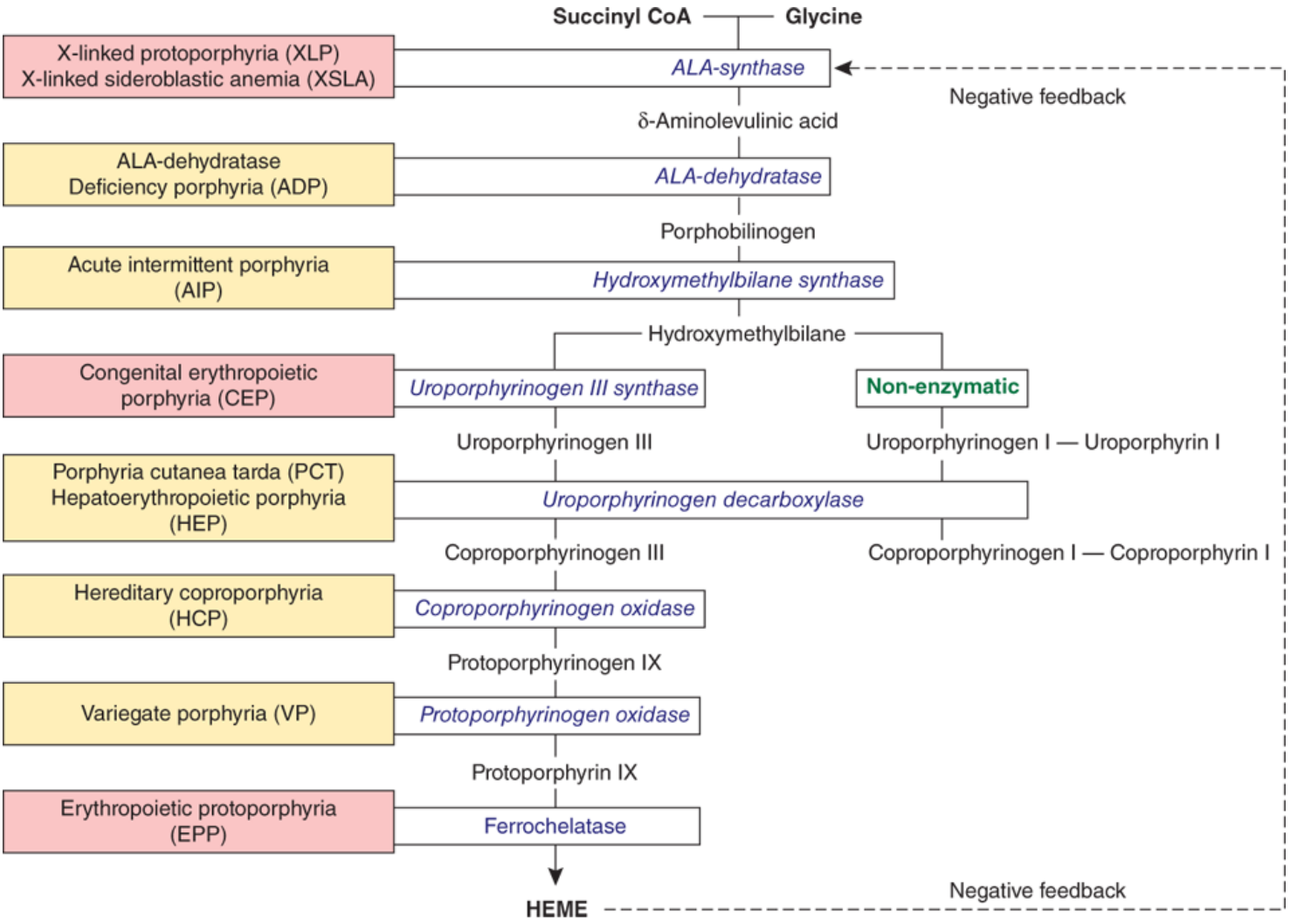
Robert J. Desnick; Manisha Balwani

THE PORPHYRIAS: INTRODUCTION

The porphyrias are metabolic disorders, each resulting from the deficiency or increased activity of a specific enzyme in the heme biosynthetic pathway (**Fig. 409-1** and **Table 409-1**). These enzyme disorders are inherited as autosomal dominant, autosomal recessive, or X-linked traits, with the exception of porphyria cutanea tarda (PCT), which is usually sporadic (Table 409-1). The porphyrias are classified as either *hepatic* or *erythropoietic*, depending on the primary site of overproduction and accumulation of their respective porphyrin precursors or porphyrins (**Tables 409-1** and **409-2**), although some have overlapping features. For example, PCT, the most common porphyria, is hepatic and presents with blistering cutaneous photosensitivity, which is typically characteristic of the erythropoietic porphyrias (EPPs).

FIGURE 409-1

The human heme biosynthetic pathway indicating in *linked boxes* the enzyme that, when deficient or overexpressed, causes the respective porphyria. Hepatic porphyrias are shown in *yellow boxes* and erythropoietic porphyrias in *pink boxes*.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 409-1

Human Porphyrias: Major Clinical and Laboratory Features

Porphyria	Deficient Enzyme	Inheritance	Principal Symptoms: NV or CP+	Enzyme Activity % of Normal	Increased Porphyrin Precursors and/or Porphyrins		
					Erythrocytes	Urine	Stool
Hepatic Porphyrias							
5-ALA-dehydratase-deficient porphyria (ADP)	ALA-dehydratase	AR	NV	~5	Zn-Protoporphyrin	ALA, coproporphyrin III	—
Acute intermittent porphyria (AIP)	HMB-synthase	AD	NV	~50	—	ALA, PBG, uroporphyrin	—
Porphyria cutanea tarda (PCT)	URO-decarboxylase	AD	CP	~20	—	Uroporphyrin, 7-carboxylate porphyrin	Isocoproporphyrin
Hereditary coproporphyria (HCP)	COPRO-oxidase	AD	NV and CP	~50	—	ALA, PBG, coproporphyrin III	Coproporphyrin III
Variegate porphyria (VP)	PROTO-oxidase	AD	NV and CP	~50	—	ALA, PBG, coproporphyrin III	Coproporphyrin III, protoporphyrin
Erythropoietic Porphyrias							
Congenital erythropoietic porphyria (CEP)	URO-synthase	AR	CP	1–5	Uroporphyrin I Coproporphyrin I	Uroporphyrin I[a] Coproporphyrin I[a]	Coproporphyrin I
Erythropoietic protoporphyria (EPP)	Ferrochelatase	AR	CP	~20–30	Protoporphyrin	—	Protoporphyrin
X-linked protoporphyria (XLP)	ALA-synthase 2	XL	CP	>100[b]	Protoporphyrin	—	Protoporphyrin

[a]Type I isomers. [b]Increased activity due to "gain-of-function" mutations in *ALAS2* exon 11.

Abbreviations: AD, autosomal dominant; ALA, 5-aminolevulinic acid; AR, autosomal recessive; COPRO I, coproporphyrin I; COPRO III, coproporphyrin III; CP, cutaneous photosensitivity; ISOCOPRO, isocoproporphyrin; + Nv, neurovisceral; PBG, porphobilinogen; PROTO, protoporphyrin IX; URO I, uroporphyrin I; URO III, uroporphyrin III; XL, X-linked.

TABLE 409-2

Human HEME Biosynthetic Enzymes and Genes

Enzyme	Gene Symbol	Chromosomal Location	cDNA (bp)	Gene		Protein (aa)	Subcellular Location	Known Mutations[b]	Three-Dimensional Structure[c]
				Size (kb)	Exons[a]				
ALA-synthase									
Housekeeping	*ALAS1*	3p21.1	2199	17	11	640	M	—	
Erythroid-specific	*ALAS2*	Xp11.2	1937	22	11	587	M	>30	—
ALA-dehydratase									
Housekeeping	*ALAD*	9q32	1149	15.9	12 (1A + 2 – 12)	330	C	12	Y
Erythroid-specific	*ALAD*	9q32	1154	15.9	12 (1B + 2 – 12)	330	C	—	
HMB-synthase									
Housekeeping	*HMBS*	11q23.3	1086	11	15 (1 + 3 – 15)	361	C	400	E
Erythroid-specific	*HMBS*	11q23.3	1035	11	15 (2 – 15)	344	C	10	
URO-synthase									
Housekeeping	*UROS*	10q26.2	1296	34	10 (1 + 2B – 10)	265	C	45	H
Erythroid-specific	*UROS*	10q26.2	1216	34	10 (2A + 2B – 10)	265	C	4	
URO-decarboxylase	*UROD*	1p34.1	1104	3	10	367	C	122	H
COPRO-oxidase	*CPOX*	3q12.1	1062	14	7	354	M	70	H
PROTO-oxidase	*PPOX*	1q23.3	1431	5.5	13	477	M	181	—
Ferrochelatase	*FECH*	18q21.31	1269	45	11	423	M	192	B

[a]Number of exons and those encoding separate housekeeping and erythroid-specific forms indicated in parentheses. [b]Number of known mutations from the Human Gene Mutation Database (*www.hgmd.org*). [c]Crystallized from human (H), murine (M), *Escherichia coli* (E), *Bacillus subtilis* (B), or yeast (Y) purified enzyme; references in Protein Data Bank (*www.rcsb.org*).

Abbreviations: C, cytoplasm; M, mitochondria.

Source: From KE Anderson et al: Disorders of heme biosynthesis: X-linked sideroblastic anemia and the porphyrias, in *The Metabolic and Molecular Bases of Inherited Diseases*, CR Scriver et al (eds). New York, McGraw-Hill, 2001, pp 2991–3062.

The major manifestations of the acute hepatic porphyrias are neurologic, including neuropathic abdominal pain, peripheral motor neuropathy, and mental disturbances, with attacks often precipitated by dieting, certain porphyrinogenic drugs, and hormonal changes. While hepatic porphyrias are symptomatic primarily in adults, rare homozygous variants of the autosomal dominant hepatic porphyrias usually manifest clinically prior to puberty. In contrast, the erythropoietic porphyrias usually present at birth or in early childhood with cutaneous photosensitivity, or in the case of congenital erythropoietic porphyria (CEP), even *in utero* as nonimmune hydrops fetalis. Cutaneous sensitivity to sunlight results from excitation of excess porphyrins in the skin by long-wave ultraviolet light, leading to cell damage, scarring, and disfigurement. Thus, the porphyrias are metabolic disorders in which environmental, physiologic, and genetic factors interact to cause disease.

Because many symptoms of the porphyrias are nonspecific, diagnosis is often delayed. Laboratory measurement of porphyrin precursors (5′-aminolevulinic acid [ALA] and porphobilinogen [PBG]) in the urine or porphyrins in the urine, plasma, erythrocytes, or feces is required to confirm or exclude the various types of porphyria (see below). However, a definite diagnosis requires demonstration of the specific gene defect **(Table 409-3)**. The genes encoding all the heme biosynthetic enzymes have been characterized, permitting identification of the mutations causing each porphyria (Table 409-2). Molecular genetic analyses now make it possible to provide precise heterozygote or homozygote identification and prenatal diagnoses in families with known mutations.

TABLE 409-3

Diagnosis of Acute and Cutaneous Porphyrias

Symptoms	First-Line Test: Abnormality	Possible Porphyria	Second-Line Testing if First-Line Testing Is Positive: To Include: Urine (U), Plasma (P), and Fecal (F) Porphyrins; for Acute Porphyrias, Add Red Blood Cell (RBC) HMB-Synthase; for Blistering Skin Lesions, Add P and RBC Porphyrins	Confirmatory Test: Enzyme Assay and/or Mutation Analysis
Neurovisceral	**Spot U:** ↑↑ALA and normal PBG	ADP	**U porphyrins:** ↑↑, mostly COPRO III **P & F porphyrins:** normal or slightly ↑ **RBC HMB-synthase:** normal	Rule out other causes of elevated ALA; ↓↓RBC ALA-dehydratase activity (<10%); ALA-dehydratase mutation analysis
	Spot U: ↑↑PBG	AIP	**U porphyrins:** ↑↑, mostly URO and COPRO **P & F porphyrins:** normal or slightly ↑ **RBC HMB-synthase:** usually ↓	HMB-synthase mutation analysis
	"	HCP	**U porphyrins:** ↑↑, mostly COPRO III **P porphyrins:** normal or slightly ↑(↑ if skin lesions present) **F porphyrins:** ↑↑, mostly COPRO III	Measure RBC HMB-synthase: normal activity COPRO-oxidase mutation analysis
	"	VP	**U porphyrins:** ↑↑, mostly COPRO III **P porphyrins:** ↑↑(characteristic fluorescence peak at neutral pH) **F porphyrins:** ↑↑, mostly COPRO and PROTOO	Measure RBC HMB-synthase: normal activity PROTO-oxidase mutation analysis
Blistering skin lesions	**P:** ↑ porphyrins	PCT and HEP	**U porphyrins:** ↑↑, mostly URO and heptacarboxylate porphyrin **P porphyrins:** ↑↑ **F porphyrins:** ↑↑, including increased isocoproporphyrin **RBC porphyrins:** ↑↑ zinc PROTO in HEP[a]	RBC URO-decarboxylase activity: half-normal in familial PCT (~20% of all PCT cases); substantially deficient in HEP URO-decarboxylase mutation analysis: mutation(s) present in familial PCT (heterozygous) and HEP (homozygous)
	"	HCP and VP	See HCP and VP above. Also, U ALA and PBG: may be ↑	
	"	CEP	**RBC and U porphyrins:** ↑↑, mostly URO I and COPRO I **F porphyrins:** ↑↑; mostly COPRO I	↓↓ RBC URO-synthase activity (<15%) URO-synthase mutation analysis
Nonblistering photosensitivity	**P:** porphyrins usually ↑	EPP	**RBC porphyrins:** ↓↓, mostly free PROTO **U porphyrins:** normal **F porphyrins:** normal or ↓, mostly PROTO	*FECH* mutation analysis
	P: porphyrins usually ↑	XLP	**RBC porphyrins:** ↑↑, approximately equal free and zinc PROTO **U porphyrins:** normal **F porphyrins:** normal or ↑, mostly PROTO	*ALAS2* mutation analysis

[a]Nonspecific increases in zinc protoporphyrins are common in other porphyrias.

Abbreviations: ADP, 5-ALA-dehydratase-deficient porphyria; AIP, acute intermittent porphyria; ALA, 5-aminolevulinic acid; CEP, congenital erythropoietic porphyria; COPRO I, coproporphyrin I; COPRO III, coproporphyrin III; EPP, erythropoietic protoporphyria; F, fecal; HCP, hereditary coporphyria; HEP, hepatoerythropoietic porphyria; ISOCOPRO, isocoproporphyrin; P, plasma; PBG, porphobilinogen; PCT, porphyria cutanea tarda; PROTO, protoporphyrin IX; RBC, erythrocytes; U, urine; URO I, uroporphyrin I; URO III, uroporphyrin III; VP, variegate porphyria; XLP, X-linked protoporphyria.

Source: Based on KE Anderson et al: Ann Intern Med 142:439, 2005.

In addition to recent reviews of the porphyrias, informative and up-to-date websites are sponsored by the American Porphyria Foundation (*www.porphyriafoundation.com*) and the European Porphyria Initiative (*www.porphyria-europe.org*). An extensive list of unsafe and safe drugs for individuals with acute porphyrias is provided at the Drug Database for Acute Porphyrias (*www.drugs-porphyria.com*).

GLOBAL CONSIDERATIONS

The porphyrias are panethnic metabolic diseases that affect individuals around the globe. The acute hepatic porphyrias—acute intermittent porphyria (AIP), hereditary coproporphyria (HCP), and variegate porphyria (VP)—are autosomal dominant disorders. The frequency of symptomatic AIP, the most common acute hepatic porphyria, is ~1 in 20,000 among Caucasian individuals of Western European ancestry, and it is particularly frequent in Scandinavians, with a frequency of ~1 in 10,000 in Sweden. However, a recent study using genomic/exomic databases showed an estimated frequency of pathogenic variants in the *HMBS* gene as ~1 in 1,700. Thus, the penetrance of AIP, and likely the other acute hepatic porphyrias, is low, about 1–10% of those with pathogenic mutations experiencing acute attacks (see below).

VP is particularly frequent in South Africa, where its high prevalence (>10,000 affected patients) is in part due to a genetic "founder effect." The autosomal recessive acute hepatic porphyria, ALA-dehydratase-deficient porphyria (ADP), is very rare, and <20 patients have been reported worldwide.

The EPPs—CEP, EPP, and X-linked protoporphyria (XLP)—also are panethnic. EPP is the most common porphyria in children, whereas CEP is very rare, with about 200 reported cases worldwide. The frequency of EPP varies globally because most patients have the common low expression ferrochelatase (*FECH*) mutation that varies in frequency in different populations. It rarely occurs in Africans, is present in about 10% of whites, and is frequent (~30%) in the Japanese. The reported prevalence of EPP in the Caucasian population ranges from 1 in 75,000 to 1 in 152,000.

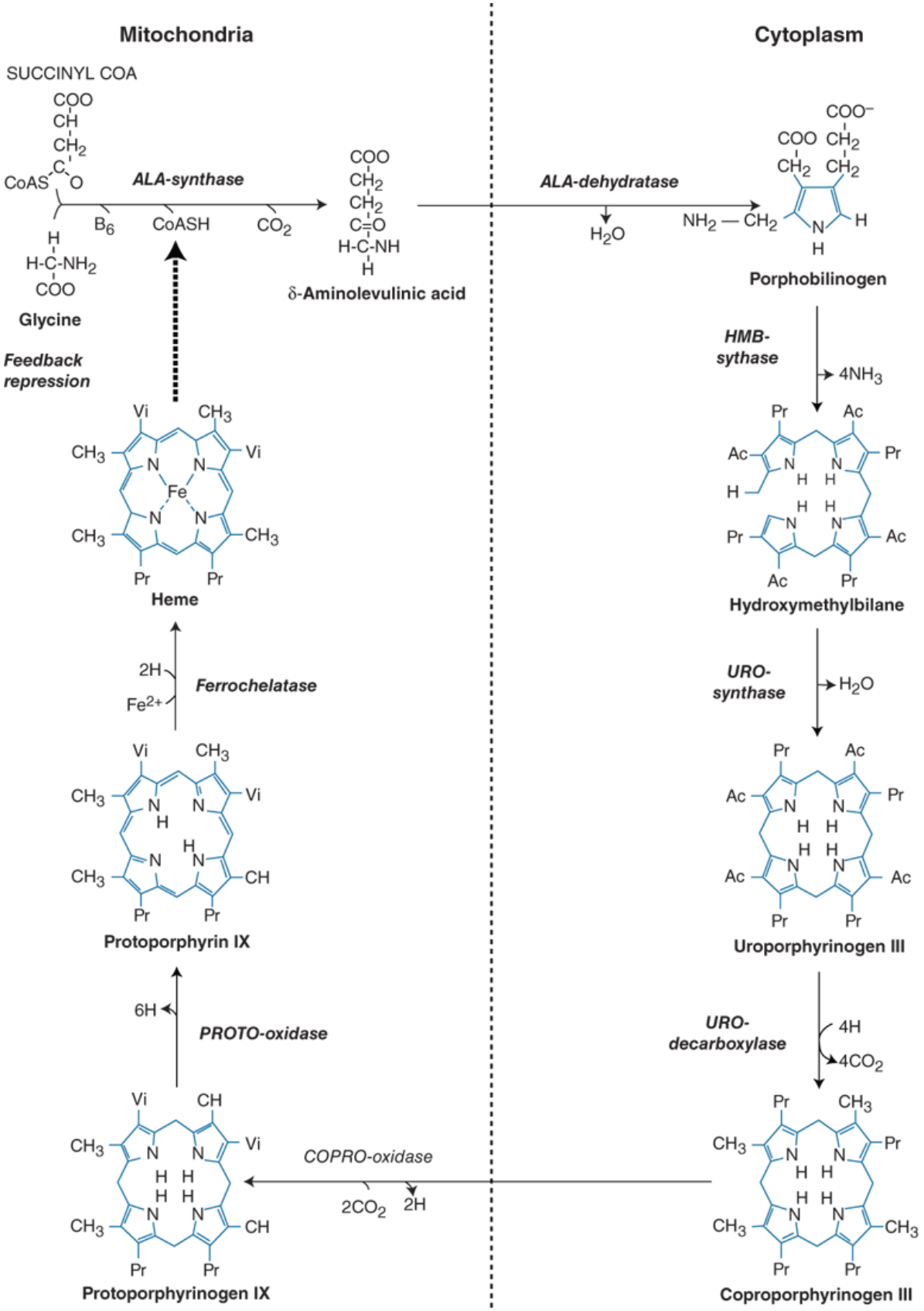
The autosomal recessive porphyrias—ADP, CEP, hepatoerythropoietic porphyria (HEP)—are more frequent in regions with high rates of consanguineous unions. PCT, which is typically sporadic, occurs more frequently in countries in which its predisposing risk factors such as hepatitis C and HIV are more prevalent.

HEME BIOSYNTHESIS

Heme biosynthesis involves eight enzymatic steps in the conversion of glycine and succinyl-CoA to heme (**Fig. 409-2** and Table 409-2). These eight enzymes are encoded by nine genes, as the first enzyme in the pathway, ALA-synthase, has two genes that encode unique housekeeping (*ALAS1*) and erythroid-specific (*ALAS2*) isozymes. The first and last three enzymes in the pathway are located in the mitochondrion, whereas the other four are in the cytosol. Heme is required for a variety of hemoproteins such as hemoglobin, myoglobin, respiratory cytochromes, and the cytochrome P450 (CYPs) enzymes. Hemoglobin synthesis in erythroid precursor cells accounts for ~85% of daily heme synthesis in humans. Hepatocytes account for most of the rest, primarily for the synthesis of CYPs, which are especially abundant in the liver endoplasmic reticulum, and turn over more rapidly than many other hemoproteins, such as the mitochondrial respiratory cytochromes. As shown in Fig. 409-2, the pathway intermediates are the porphyrin precursors, ALA and PBG, and porphyrins (mostly in their reduced forms, known as *porphyrinogens*). At least in humans, these intermediates do not accumulate in significant amounts under normal conditions or have important physiologic functions.

FIGURE 409-2

The heme biosynthetic pathway showing the eight enzymes and their substrates and products. Four of the enzymes are localized in the mitochondria and four in the cytosol.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The first enzyme, ALA-synthase, catalyzes the condensation of glycine, activated by pyridoxal phosphate and succinyl-coenzyme A, to form ALA. In the liver, this rate-limiting enzyme can be induced by a variety of drugs, steroids, and other chemicals. Distinct nonerythroid (e.g., housekeeping) and erythroid-specific forms of ALA-synthase are encoded by separate genes located on chromosome 3p21.1 (*ALAS1*) and

Xp11.2 (*ALAS2*), respectively. Defects in the erythroid gene *ALAS2* that decrease its activity cause an X-linked sideroblastic anemia (XLSA). Gain-of-function mutations in the last exon (11) of *ALAS2* that increase its activity cause an X-linked form of EPP, known as XLP.

The second enzyme, ALA-dehydratase, catalyzes the condensation of two molecules of ALA to form PBG. Hydroxymethylbilane synthase (HMB-synthase; also known as PBG-deaminase) catalyzes the head-to-tail condensation of four PBG molecules by a series of deaminations to form the linear tetrapyrrole, HMB. Uroporphyrinogen III synthase (URO-synthase) catalyzes the rearrangement and rapid cyclization of HMB to form the asymmetric, physiologic, octacarboxylate porphyrinogen, uroporphyrinogen (URO'gen) III.

The fifth enzyme in the pathway, uroporphyrinogen decarboxylase (URO-decarboxylase), catalyzes the sequential removal of the four carboxyl groups from the acetic acid side chains of URO'gen III to form coproporphyrinogen (COPRO'gen) III, a tetracarboxylate porphyrinogen. This compound then enters the mitochondrion via a specific transporter, where COPRO-oxidase, the sixth enzyme, catalyzes the decarboxylation of two of the four propionic acid groups to form the two vinyl groups of protoporphyrinogen (PROTO'gen) IX, a decarboxylate porphyrinogen. Next, PROTO-oxidase oxidizes PROTO'gen to protoporphyrin IX by the removal of six hydrogen atoms. The product of the reaction is a porphyrin (oxidized form), in contrast to the preceding tetrapyrrole intermediates, which are porphyrinogens (reduced forms). Finally, ferrous iron is inserted into protoporphyrin to form heme, a reaction catalyzed by the eighth enzyme in the pathway, FECH (also known as heme synthase or protoheme ferrolyase).

REGULATION OF HEME BIOSYNTHESIS

Regulation of heme synthesis differs in the two major heme-forming tissues, the liver and erythron. In the liver, the concentration of "free" heme regulates the synthesis and mitochondrial translocation of the housekeeping form of ALA-synthase 1. Heme represses the synthesis of the ALA-synthase 1 messenger RNA (mRNA) and interferes with the transport of the enzyme from the cytosol into mitochondria. Hepatic ALA-synthase 1 is increased by many of the same chemicals that induce the CYPs enzymes in the endoplasmic reticulum of the liver. Because most of the heme in the liver is used for the synthesis of CYPs enzymes, hepatic ALA-synthase 1 and the CYPs are regulated in a coordinated fashion, and many drugs that induce hepatic ALA-synthase 1 also induce the CYP genes. The other hepatic heme biosynthetic enzymes are presumably expressed at constant levels, although their relative activities and kinetic properties differ. For example, normal individuals have high activities of ALA-dehydratase, but low activities of HMB-synthase, the latter being the second rate-limiting step in the pathway.

In the erythron, novel regulatory mechanisms allow for the production of the very large amounts of heme needed for hemoglobin synthesis. The response to stimuli for hemoglobin synthesis occurs during cell differentiation, leading to an increase in cell number. In contrast, the erythroid-specific ALA-synthase 2 is expressed at higher levels than the housekeeping enzyme, and erythroid-specific control mechanisms regulate other pathway enzymes as well as iron transport into erythroid cells. Separate erythroid-specific and nonerythroid or "housekeeping" transcripts are known for the first four enzymes in the pathway. As noted above, housekeeping- and erythroid-specific ALA-synthases are encoded by genes on different chromosomes, but for each of the next three genes in the pathway, both erythroid and nonerythroid transcripts are transcribed by alternative promoters from their single respective genes (Table 409-2).

CLASSIFICATION OF THE PORPHYRIAS

As mentioned above, the porphyrias can be classified as either *hepatic* or *erythropoietic*, depending on whether the heme biosynthetic intermediates that accumulate arise initially from the liver or developing erythrocytes, or as *acute* or *cutaneous*, based on their clinical manifestations. Table 409-1 lists the porphyrias, their principal symptoms, and major biochemical abnormalities. Three of the five hepatic porphyrias—AIP, HCP, and VP—usually present during adult life with acute attacks of neurologic manifestations and elevated levels of one or both of the porphyrin precursors, ALA and PBG, and are thus classified as *acute porphyrias*. Patients with ADP have presented in infancy and adolescence, and typically have elevated ALA with normal or slightly elevated PBG levels. The fifth hepatic disorder, PCT, presents with blistering skin lesions. HCP and VP also may have cutaneous manifestations similar to PCT.

The erythropoietic porphyrias—CEP, EPP, and XLP—are characterized by elevations of porphyrins in bone marrow and erythrocytes and present with cutaneous photosensitivity. The skin lesions in CEP resemble PCT but are usually much more severe, whereas EPP and XLP cause a more immediate, severe, painful, and nonblistering type of photosensitivity. EPP is the most common porphyria to cause symptoms before puberty. Around 20% of EPP patients develop minor abnormalities of liver function, with up to about 5% developing hepatic complications that can become life-threatening. XLP has a clinical presentation similar to EPP causing photosensitivity and liver disease.

DIAGNOSIS OF PORPHYRIA

A few specific and sensitive first-line laboratory tests should be used whenever symptoms or signs suggest the diagnosis of porphyria (Table 409-3). If a first-line test is significantly abnormal, more comprehensive testing should follow to establish the type of porphyria, including the specific causative gene mutation.

Acute Porphyrias

An acute porphyria should be suspected in patients with neurovisceral symptoms after puberty. Symptoms include acute abdominal pain, nausea, vomiting, tachycardia, hypertension, and motor neuropathy. As these symptoms are common, other causes should be ruled out. The diagnosis is made by measuring urinary porphyrin precursors (ALA and PBG) in a spot sample of urine (Fig. 409-2). Urinary PBG is always increased during acute attacks of AIP, HCP, and VP and is not substantially increased in any other medical condition. Therefore, this measurement is both sensitive and specific. Results from spot (single-void) urine specimens are highly informative because very substantial increases in PBG are expected during acute attacks of porphyria. A 24-h collection can unnecessarily delay diagnosis. The same spot urine specimen should be saved for quantitative determination of ALA, PBG, and creatinine, in order to confirm the qualitative PBG result and also to detect patients with ADP. Urinary porphyrins may remain increased longer than porphyrin precursors in HCP and VP. Therefore, it is useful to measure total urinary porphyrins in the same sample, keeping in mind that urinary porphyrin increases are often nonspecific. Measurement of urinary porphyrins alone should be avoided for screening, because these may be increased in disorders other than porphyrias, such as chronic liver disease, and misdiagnoses of porphyria can result from minimal increases in urinary porphyrins that have no diagnostic significance. Measurement of erythrocyte HMB-synthase is not useful as a first-line test. Moreover, the enzyme activity is not decreased in all AIP patients, a borderline low normal value is not diagnostic, and the enzyme is not deficient in other acute porphyrias.

Cutaneous Porphyrias

Blistering skin lesions due to porphyria are virtually always accompanied by increases in total plasma porphyrins. A fluorometric method is preferred, because the porphyrins in plasma in VP are mostly covalently linked to plasma proteins and may be less readily detected by high-performance liquid chromatography (HPLC). The normal range for plasma porphyrins is somewhat increased in patients with end-stage renal disease.

Although a total plasma porphyrin determination will usually detect EPP and XLP, an erythrocyte protoporphyrin determination is more sensitive. Increases in erythrocyte protoporphyrin occur in many other conditions. Therefore, the diagnosis of EPP must be confirmed by showing a predominant increase in free protoporphyrin rather than zinc protoporphyrin. In XLP, both free and zinc protoporphyrin are markedly increased. Interpretation of laboratory reports can be difficult, because the term *free erythrocyte protoporphyrin* sometimes actually represents zinc protoporphyrin.

More extensive testing is justified when an initial test is positive. A substantial increase in PBG may be due to AIP, HCP, or VP. These acute porphyrias can be distinguished by measuring urinary porphyrins (using the same spot urine sample), fecal porphyrins, and plasma porphyrins. Assays for COPRO-oxidase or PROTO-oxidase are not available for clinical testing. More specifically, mutation analysis by sequencing the genes encoding HMB-synthase, COPRO-oxidase, and PROTO-oxidase will detect almost all disease-causing mutations, and will be diagnostic even when the levels of urinary ALA and PBG have returned to normal or near normal. The various porphyrias that cause blistering skin lesions can be differentiated by measuring porphyrins in urine, feces, and plasma. These porphyrias also should be confirmed at the DNA level by the demonstration of the causative gene mutation(s). It is often difficult to diagnose or “rule out” porphyria in patients who have had suggestive symptoms months or years in the past, and in relatives of patients with acute porphyrias, because porphyrin precursors and porphyrins may be normal. In those situations, detection of the specific gene mutation in the index case can make the diagnosis and facilitate the diagnosis and genetic counseling of at-risk relatives. Consultation with a specialist laboratory and physician will assist in selecting the heme biosynthetic gene or genes to be sequenced.

THE HEPATIC PORPHYRIAS

Markedly elevated plasma and urinary concentrations of the porphyrin precursors, ALA and/or PBG, which originate from the liver, are especially evident during attacks of neurologic manifestations of the four acute porphyrias—ADP, AIP, HCP, and VP. In PCT, excess porphyrins also accumulate initially in the liver and cause chronic blistering of sun-exposed areas of the skin.

ALA-DEHYDRATASE-DEFICIENT PORPHYRIA

ADP is a rare, autosomal recessive, acute hepatic porphyria caused by a severe deficiency of ALA-dehydratase activity. To date, there are only a few documented cases, some in children or young adults, in which specific gene mutations have been identified. These affected homozygotes had <10% of normal ALA-dehydratase activity in erythrocytes, but their clinically asymptomatic parents and heterozygous relatives had about half-normal levels of activity and did not excrete increased levels of ALA. The frequency of ADP is unknown, but the frequency of heterozygous individuals with <50% normal ALA-dehydratase activity was ~2% in a screening study in Sweden. Because there are multiple causes for deficient ALA-dehydratase activity, it is important to confirm the diagnosis of ADP by mutation analysis.

Clinical Features

The clinical presentation depends on the amount of residual ALA-dehydratase activity. Four of the documented patients were male adolescents with symptoms resembling those of AIP, including abdominal pain and neuropathy. One patient was an infant with more severe disease, including failure to thrive beginning at birth. The earlier age of onset and more severe manifestations in this patient reflect a more significant deficiency of ALA-dehydratase activity. Another patient developed an acute motor polyneuropathy at age 63 that was associated with a myeloproliferative disorder. He was heterozygous for an δ-aminolevulinic acid dehydratase (*ALAD*) mutation that presumably was present in erythroblasts that underwent clonal expansion due to the bone marrow malignancy.

Diagnosis

All patients had significantly elevated levels of plasma and urinary ALA and urinary coproporphyrin (COPRO) III; ALAD activities in erythrocytes were <10% of normal. Hereditary tyrosinemia type 1 (fumarylacetoacetase deficiency) and lead intoxication should be considered in the differential diagnosis because either succinylacetone (which accumulates in hereditary tyrosinemia and is structurally similar to ALA) or lead can inhibit ALA-dehydratase, increase urinary excretion of ALA and COPRO III, and cause manifestations that resemble those of the acute porphyrias. Heterozygotes are clinically asymptomatic and do not excrete increased levels of ALA but can be detected by demonstration of intermediate levels of erythrocyte ALA-dehydratase activity or a specific mutation in the *ALAD* gene. To date, molecular studies of ADP patients have identified point mutations, splice-site mutations, and a two-base deletion in the *ALAD* gene (Human Gene Mutation Database; *www.hgmd.org*). The parents in each case were not consanguineous, and the index cases had inherited a different *ALAD* mutation from each parent. Prenatal diagnosis of this disorder is possible by determination of ALA-dehydratase activity and/or gene mutations in cultured chorionic villi or amniocytes.

TREATMENT

TREATMENT

ALA-Dehydratase-Deficient Porphyria

The treatment of ADP acute attacks is similar to that of AIP (see below). The severely affected infant referred to above was supported by hyperalimentation and periodic blood transfusions but did not respond to intravenous hemin and died after liver transplantation.

ACUTE INTERMITTENT PORPHYRIA

This hepatic porphyria is an autosomal dominant condition resulting from the half-normal level of HMB-synthase activity. The disease is widespread but is especially common in Scandinavia and Great Britain. Clinical expression is highly variable, and activation of the disease is often related to environmental or hormonal factors, such as drugs, diet, and steroid hormones. Attacks can be prevented by avoiding known precipitating factors. Rare homozygous dominant AIP also has been described in children (see below).

Clinical Features

Induction of the rate-limiting hepatic enzyme ALA-synthase in heterozygotes who have half-normal HMB-synthase activity is thought to underlie the acute attacks in AIP. The disorder remains latent (or asymptomatic) in the great majority of those who are heterozygous for *HMBS* mutations, and this is almost always the case prior to puberty. In patients with no history of acute symptoms, porphyrin precursor excretion is usually normal, suggesting that half-normal hepatic HMB-synthase activity is sufficient and that hepatic ALA-synthase activity is not increased. However, under conditions where heme synthesis is increased in the liver, half-normal HMB-synthase activity may become limiting, and ALA, PBG, and other heme pathway intermediates may accumulate and be excreted in the urine. Common precipitating factors include endogenous and exogenous steroids, porphyrinogenic drugs, alcohol ingestion, and low-calorie diets, usually instituted for weight loss.

The fact that AIP is almost always latent before puberty suggests that adult levels of steroid hormones are important for clinical expression. Symptoms are more common in women, suggesting a role for estrogens or progestins. Premenstrual attacks are probably due to endogenous progesterone. Acute porphyrias are sometimes exacerbated by exogenous steroids, including oral contraceptive preparations containing progestins. Surprisingly, pregnancy is usually well tolerated, suggesting that beneficial metabolic changes may ameliorate the effects of high levels of progesterone. **Table 409-4** provides a partial list of the major drugs that are harmful in AIP (and also in HCP and VP). Extensive lists of unsafe and safe drugs are available on websites sponsored by the American Porphyria Foundation (*www.porphyriafoundation.com*) and the European Porphyria Initiative (*www.porphyria-europe.org*), and at the Drug Database for Acute Porphyrias website (*www.drugs-porphyria.com*). Reduced intake of calories and carbohydrate, as may occur with illness or attempts to lose weight, can also increase porphyrin precursor excretion and induce attacks of porphyria. Increased carbohydrate intake may ameliorate attacks. Studies in a knockout AIP mouse model indicate that the hepatic *ALAS1* gene is regulated by the peroxisome proliferator-activated receptor γ coactivator 1α (PGC-1α). Hepatic PGC-1α is induced by fasting, which in turn activates *ALAS1* transcription, resulting in increased heme biosynthesis. This finding suggests an important link between nutritional status and the attacks in acute porphyrias. Attacks also can be provoked by infections, surgery, and ethanol.

TABLE 409-4

Unsafe Drugs in Porphyria

Documented Porphyrinogenic	Probably Porphyrinogenic	Possibly Porphyrinogenic	
Carbamazepine	Altretamine	Aceclofenac	Parecoxib
Carisoprodol	Aminophylline	Acitretin	Pentifylline
Chloramphenicol	Amiodarone	Acrivastine	Pentoxyverine
Clindamycin	Amitriptyline	Alfuzosin	Phenylpropanolamine + cinnarizine
Dextropropoxyphene	Amlodipine	Anastrozole	Pizotifen
Dihydralazine	Amprenavir	Auranofin	Polidocanol
Dihydroergotamine	Aprepitant	Azelastine	Polyestradiol
Drospirenone + estrogen	Atorvastatin	Benztropine	Phosphate
Dydrogesterone	Azathioprine	Benzydamine	Potassium canrenoate
Etonogestrel	Bosentan	Betaxolol	Pravastatin
Fosphenytoin sodium	Bromocriptine	Bicalutamide	Prednisolone
Hydralazine	Buspirone	Biperiden	Prilocaine
Hydroxyzine	Busulfan	Bupropion	Proguanil
Indinavir	Butylscopolamine	Carvedilol	Propafenone
Ketamine	Cabergoline	Chlorambucil	Pseudoephedrine + dexbrompheniramine
Ketoconazole	Ceftriaxone + lidocaine	Chlorcyclizine + guaifenesin	Quillaia extract
Lidocaine	Cerivastatin	Chloroquine	Quinagolide
Lynestrenol	Cetirizine	Chlorprothixene	Quinine
Lynestrenol + estrogen	Cholinetheophyllinate	Chlorzoxazone	Quinupristin + Dalfopristin
Mecillinam	Clarithromycin	Chorionic	Reboxetine
Medroxyprogesterone	Clemastine	Gonadotropin	Repaglinide
Megestrol	Clonidine	Ciclosporin	Rizatriptan
Methylergometrine	Cyclizine	Cisapride	Rofecoxib
Methyldopa	Cyproterone	Citalopram	Ropinirole

Documented Porphyrinogenic	Probably Porphyrinogenic	Possibly Porphyrinogenic	
Mifepristone	Danazol	Clomethiazole	Ropivacaine
Nicotinic acid/meclozine/hydroxyzine	Delavirdine	Clomiphene	Roxithromycin
	Desogestrel + estrogen	Clomipramine	Sertraline
Nitrofurantoin	Diazepam	Clopidogrel	Sevoflurane
Norethisterone	Dienogest + estrogen	Clotrimazole	Sibutramine
Norgestimate + estrogen	Diclofenac	Cortisone	Sildenafil
Orphenadrine	Diltiazem	Cyclandelate	Sirolimus
Phenobarbital	Diphenhydramine	Cyclophosphamide	Sodium aurothiomalate
Phenytoin	Disopyramide	Cyproheptadine	Sodium oleate + chlorocymol
Pivampicillin	Disulfiram	Dacarbazine	Stavudine
Pivmecillinam	Drospirenone + estrogen	Daunorubicin	Sulindac
Primidone	Dydrogesterone	Desogestrel	Sumatriptan
Rifampicin	Ergoloid mesylate	Dichlorobenzyl alcohol	Tacrolimus
Ritonavir	Erythromycin	Dithranol	Tadalafil
Spironolactone	Estramustine	Docetaxel	Tegafur + uracil
Sulfadiazine + trimethoprim	Ethosuximide	Donepezil	Telmisartan
Tamoxifen	Etoposide	Doxycycline	Thioridazine
Testosterone, injection	Exemestane	Ebastine	Thioguanine
Thiopental	Felbamate	Econazole	Tolfenamic acid
Trimethoprim	Felodipine	Efavirenz	Tolterodine
Valproic acid	Fluconazole	Escitalopram	Torsemide
Venlafaxine	Flunitrazepam	Esomeprazole	Triamcinolone
Vinblastine	Fluvastatin	Estradiol/tablets	Trihexyphenidyl
Vincristine	Glibenclamide	Estriol/tablets	Trimipramine
Vindesine	Halothane	Estrio/vaginal crème, tablet	Valerian
Vinorelbine	Hyoscyamine		Venlafaxine

Documented Porphyrinogenic	Probably Porphyrinogenic	Possibly Porphyrinogenic	
Xylometazoline	Ifosfamide	Estrogen, conjugate	Vinblastine
Zaleplon	Imipramine	Finasteride	Vincristine
Ziprasidone	Irinotecan	Flecainide	Vindesine
Zolmitriptan	Isoniazid	Flucloxacillin	Vinorelbine
Zolpidem	Isradipine	Fluoxetine	Xylometazoline
Zuclopenthixol	Itraconazole	Flupentixol	Zaleplon
	Lamivudine + zidovudine	Flutamide	Ziprasidone
	Lansoprazole	Fluvoxamine	Zolmitriptan
	Lercanidipine	Follitropin alfa and beta	Zolpidem
	Levonorgestrel	Galantamine	Zuclopenthixol
	Lidocaine	Glimepiride	
	Lopinavir	Glipizide	
	Lutropin alfa	Gonadorelin	
	Lymecycline	Gramicidin	
	Meclozine	Guaifenesin	
	Medroxyprogesterone + estrogen	Hydrocortisone	
	Metoclopramide	Hydroxycarbamide	
	Metronidazole	Hydroxychloroquine	
	Metyrapone	Ibutilide	
	Moxonidine	Imatinib	
	Nandrolone	Indomethacin	
	Nefazodone	Ketobemidone + DDBA	
	Nelfinavir	Ketoconazole	
	Nevirapine	Ketorolac	
	Nifedipine	Lamotrigine	
	Nimodipine	Letrozole	

Documented Porphyrinogenic	Probably Porphyrinogenic	Possibly Porphyrinogenic	
	Nitrazepam	Levodopa + benserazide	
	Norethisterone	Levonorgestrel intra-uterine	
	Nortriptyline		
	Oxcarbazepine	Levosimendan	
	Oxytetracycline	Lidocaine	
	Paclitaxel	Linezolid	
	Paroxetine	Lofepramine	
	Phenazone + caffeine	Lomustine	
	Pioglitazone	Malathion	
	Probenecid	Maprotiline	
	Progesterone, vaginal gel	Mebendazole	
	Quinidine	Mefloquine	
	Rabeprazole	Melperone	
	Raloxifene	Melphalan	
	Rifabutin	Mepenzolate	
	Riluzole	Mepivacaine	
	Risperidone	Mercaptopurine	
	Rosiglitazone	Methadone	
	Saquinavir	Methylprednisolone	
	Selegiline	Methixene	
	Simvastatin	Metolazone	
	Sulfasalazine	Metronidazole	
	Telithromycin	Mexiletine	
	Terbinafine	Mianserin	
	Terfenadine	Midazolam	
	Testosterone, transdermal patch	Minoxidil	

Documented Porphyrinogenic	Probably Porphyrinogenic	Possibly Porphyrinogenic	
	Tetracycline	Mirtazapine	
	Theophylline	Mitomycin	
	Thiamazole	Mitoxantrone	
	Tibolone	Moclobemide	
	Ticlopidine	Montelukast	
	Tinidazole	Morphine + scopolamine	
	Thiotepa	Multivitamins	
	Topiramate	Mupirocin	
	Topotecan	Nabumetone	
	Toremifene	Nafarelin	
	Tramadol	Naltrexone	
	Trimegestone + estrogen	Nateglinide	
	Verapamil	Nilutamide	
	Voriconazole	Noscapine	
	Zidovudine/AZT	Omeprazole	
		Oxybutynin	
		Oxycodone	
		Pantoprazole	
		Papaverine	

Note: Based on list in "Patient's and Doctor's Guide to Medication in Acute Porphyria," Swedish Porphyria Association and Porphyria Centre Sweden. Also see the website Drug Database for Acute Porphyrias (*www.drugs-porphyria.com*) for a searchable list of safe and unsafe drugs.

Because the neurovisceral symptoms rarely occur before puberty and are often nonspecific, a high index of suspicion is required to make the diagnosis. The disease can be disabling but is rarely fatal. Abdominal pain, the most common symptom, is poorly localized, but may be associated with cramping, ileus, abdominal distention, and decreased bowel sounds. However, increased bowel sounds and diarrhea may occur. Abdominal tenderness, fever, and leukocytosis are usually absent or mild because the symptoms are neurologic rather than inflammatory. Nausea; vomiting; constipation; tachycardia; hypertension; mental symptoms; pain in the limbs, head, neck, or chest; muscle weakness; sensory loss; dysuria; and urinary retention are characteristic. Tachycardia, hypertension, restlessness, tremors, and excess sweating are due to sympathetic overactivity.

The peripheral neuropathy is due to axonal degeneration (rather than demyelinization) and primarily affects motor neurons. Significant neuropathy does not occur with all acute attacks; abdominal symptoms are usually more prominent. Motor neuropathy affects the proximal muscles initially, more often in the shoulders and arms. The course and degree of involvement are variable and sometimes may be focal and involve cranial nerves. Deep tendon reflexes initially may be normal or hyperactive but become decreased or absent as the neuropathy

advances. Sensory changes such as paresthesia and loss of sensation are less prominent. Progression to respiratory and bulbar paralysis and death occurs especially when the diagnosis and treatment are delayed. Sudden death may result from sympathetic overactivity and cardiac arrhythmia.

Mental symptoms such as anxiety, insomnia, depression, disorientation, hallucinations, and paranoia can occur in acute attacks. Seizures can be due to neurologic effects or to hyponatremia. Treatment of seizures is difficult because most anti-seizure drugs can exacerbate AIP (clonazepam may be safer than phenytoin or barbiturates). Hyponatremia results from hypothalamic involvement and inappropriate vasopressin secretion or from electrolyte depletion due to vomiting, diarrhea, poor intake, or excess renal sodium loss. Persistent hypertension and impaired renal function may occur. When an attack resolves, abdominal pain may disappear within hours, and paresis begins to improve within days and may continue to improve over several years.

Homozygous dominant AIP is a rare form of AIP in which patients inherit *HMBS* mutations from each of their heterozygous parents and, therefore, have very low (<2%) enzyme activity. The disease has been described in a Dutch girl, two young British siblings, and a Spanish boy. In these homozygous affected patients, the disease presented in infancy with failure to thrive, developmental delay, bilateral cataracts, and/or hepatosplenomegaly. Urinary ALA and PBG concentrations were markedly elevated. All of these patients' *HMBS* mutations (R167W, R167Q, and R172Q) were in exon 10 within five bases of each other. Studies of the brain magnetic resonance images (MRIs) of children with homozygous AIP have suggested damage primarily in white matter that was myelinated postnatally, while tracks that myelinated prenatally were normal. Most children with homozygous AIP die at an early age.

Diagnosis

ALA and PBG levels are substantially increased in plasma and urine, especially during acute attacks. For example, urinary PBG excretion during an attack is usually 50–200 mg/24 h (220–880 µmol/24 h) (normal, 0–4 mg/24 h, [0–18 µmol/24 h]), and urinary ALA excretion is 20–100 mg/24 h (150–760 µmol/24 h) (normal, 1–7 mg/24 h [8–53 µmol/24 h]). Because levels often remain high after symptoms resolve, the diagnosis of an acute attack in a patient with biochemically proven AIP is based primarily on clinical features. Excretion of ALA and PBG decreases over a few days after intravenous hemin administration. A normal urinary PBG level before hemin effectively excludes AIP as a cause for current symptoms. Fecal porphyrins are usually normal or minimally increased in AIP, in contrast to HCP and VP. Most AIP heterozygotes with no history of symptoms have normal urinary excretion of ALA and PBG and are classified as latent. Patients can also have high levels of urine PBG and ALA with no clinical symptoms. These patients may have a previous history of an acute attack. These patients are classified as "asymptomatic high excretors" (ASHE). Therefore, the detection of the family's *HMBS* mutation will diagnose asymptomatic family members.

Patients with *HMBS* mutations in the initiation of translation codon in exon 1 and in the intron 15′-splice donor site have normal enzyme levels in erythrocytes and deficient activity only in nonerythroid tissues. This occurs because the erythroid and housekeeping forms of HMB-synthase are encoded by a single gene, which has two promoters. Thus, the enzyme assay may not be diagnostic, and genetic testing should be used to confirm the diagnosis.

More than 410 *HMBS* mutations have been identified in AIP, including missense, nonsense, and splicing mutations and insertions and deletions, with most mutations found in only one or a few families (Human Gene Mutation Database, *www.hgmd.org*). The prenatal diagnosis of a fetus at risk can be made with cultured amniotic cells or chorionic villi. However, this is seldom done, because the prognosis of individuals with *HMBS* mutations is generally favorable.

TREATMENT

TREATMENT

Acute Intermittent Porphyria

During acute attacks, narcotic analgesics may be required for abdominal pain, and phenothiazines are useful for nausea, vomiting, anxiety, and restlessness. Chloral hydrate can be given for insomnia, and benzodiazepines are probably safe in low doses if a minor tranquilizer is required. Carbohydrate loading, usually with intravenous glucose (at least 300 g daily), may be effective in milder acute attacks of porphyria (without paresis, hyponatremia, etc.) if hemin is not available. Intravenous hemin is more effective and should be used as first-line therapy for all acute attacks. The standard regimen is 3–4 mg/kg of heme, in the form of lyophilized hematin (Recordati Rare Diseases), heme albumin (hematin reconstituted with human albumin), or heme arginate (Orphan Europe), infused daily for 4 days. Heme arginate and heme albumin are chemically stable and are less likely than hematin to produce phlebitis or an anticoagulant effect. Recovery depends on the degree of neuronal damage and usually is rapid if therapy is started early. Recovery from severe motor neuropathy may require months or years. Identification and avoidance of inciting factors can hasten recovery from an attack and prevent future attacks. Inciting factors are usually

multiple, and removal of one or more hastens recovery and helps prevent future attacks. Frequent attacks that occur during the luteal phase of the menstrual cycle may be prevented with a gonadotropin-releasing hormone analogue, which prevents ovulation and progesterone production, or by prophylactic hematin administration.

The long-term risk of hypertension and chronic renal disease is increased in AIP; a number of patients have undergone successful renal transplantation. Chronic, low-grade abnormalities in liver function tests are common, and the risk of hepatocellular carcinoma is increased. Hepatic imaging is recommended at least yearly for early detection of these tumors. Other long-term complications include neuropathy, chronic pain, nausea, depression, and/or anxiety.

Orthotopic liver transplantation (OLT) has been successful and is curative in patients with severe, disabling, intractable attacks that are refractory to hemin therapy. Reports from both the UK and the U.S. show a marked improvement with no subsequent attacks, an improvement in the neuropathic manifestations, and normalization of the urinary PBG and ALA levels after liver transplantation. OLT is associated with morbidity and mortality and should be considered a treatment of last resort in these patients. In addition, patients who already have advanced neuropathy are considered poor risks for transplantation. Some patients with both recurrent attacks and end-stage renal disease have benefitted from combined liver and kidney transplantation.

Liver-directed gene therapy has proven successful in the prevention of drug-induced biochemical attacks in a murine model of human AIP, and clinical trials of adeno-associated virus vector (AAV)-*HMBS* gene transfer have been initiated. Although the therapy was safe, there was essentially no biochemical evidence of its effectiveness, nor did it prevent the recurrent attacks in the treated patients. Recent data from Phase 1 clinical trials of a hepatic-targeted RNA interference (RNAi) therapy directed to inhibit the markedly elevated levels of the hepatic *ALAS1* mRNA in patients with high levels of ALA and PBG showed reduced levels of the *ALAS1* mRNA and markedly decreased urinary ALA and PBG concentrations. In AIP patients with recurrent attacks, early studies indicate that the RNAi therapy reduced the frequency of acute attacks.

PORPHYRIA CUTANEA TARDA

PCT, the most common of the porphyrias, can be either sporadic (type 1) or familial (type 2) and can also develop after exposure to halogenated aromatic hydrocarbons. Hepatic URO-decarboxylase is deficient in all types of PCT, and for clinical symptoms to manifest, this enzyme deficiency must be substantial (~20% of normal activity or less); it is currently attributed to generation of an URO-decarboxylase inhibitor in the liver, which forms uroporphomethene in the presence of iron and under conditions of oxidative stress. The majority of PCT patients (~80%) have no *UROD* mutations and are said to have sporadic (type 1) disease. PCT patients heterozygous for *UROD* mutations have familial (type 2) PCT. In these patients, inheritance of a *UROD* mutation from one parent results in half-normal enzyme activity in liver and all other tissues, which is a significant predisposing factor, but is insufficient by itself to cause symptomatic PCT. As discussed below, other genetic and environmental factors contribute to susceptibility for both types of PCT. Because penetrance of the genetic trait is low, many patients with familial (type 2) PCT have no family history of the disease. HEP is an autosomal recessive form of porphyria that results from the marked systemic deficiency of URO-decarboxylase activity with clinical symptoms in childhood.

Clinical Features

Blistering skin lesions that appear most commonly on the backs of the hands are the major clinical feature **(Fig. 409-3)**. These rupture and crust over, leaving areas of atrophy and scarring. Lesions may also occur on the forearms, face, legs, and feet. Skin friability and small white papules termed milia are common, especially on the backs of the hands and fingers. Hypertrichosis and hyperpigmentation, especially of the face, are especially troublesome in women. Occasionally, the skin over sun-exposed areas becomes severely thickened, with scarring and calcification that resembles systemic sclerosis. Neurologic features are absent.

FIGURE 409-3

Typical cutaneous lesions in a patient with porphyria cutanea tarda. Chronic, crusted lesions resulting from blistering due to photosensitivity on the dorsum of the hand of a patient with porphyria cutanea tarda. *(Courtesy of Dr. Karl E. Anderson; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A number of susceptibility factors, in addition to inherited *UROD* mutations in type 2 PCT, can be recognized clinically and can affect management. These include hepatitis C, HIV, excess alcohol, elevated iron levels, and estrogens. The importance of excess hepatic iron as a precipitating factor is underscored by the finding that the incidence of the common hemochromatosis-causing mutations, hemochromatosis gene (*HFE*) mutations C282Y and H63D, are increased in patients with types 1 and 2 PCT **(Chap. 407)**. Excess alcohol is a long-recognized contributor, as is estrogen use in women. HIV is probably an independent but less common risk factor that, like hepatitis C, does not cause PCT in isolation. Multiple susceptibility factors that appear to act synergistically can be identified in the individual PCT patient. Patients with PCT characteristically have chronic liver disease and sometimes cirrhosis and are at risk for hepatocellular carcinoma. Various chemicals can also induce PCT; an epidemic of PCT occurred in eastern Turkey in the 1950s as a consequence of wheat contaminated with the fungicide hexachlorobenzene. PCT also occurs after exposure to other chemicals, including di- and trichlorophenols and 2,3,7,8-tetrachlorodibenzo-(*p*)-dioxin (TCDD, dioxin).

Diagnosis

Porphyrins are increased in the liver, plasma, urine, and stool. The urinary ALA level may be slightly increased, but the PBG level is normal. Urinary porphyrins consist mostly of uroporphyrins and heptacarboxylate porphyrin, with lesser amounts of coproporphyrin and hexa- and pentacarboxylate porphyrins. Plasma porphyrins are also increased, and fluorometric scanning of diluted plasma at neutral pH can rapidly distinguish VP and PCT (Table 409-3). Isocoproporphyrins, which are increased in feces and sometimes in plasma and urine, are diagnostic for hepatic URO-decarboxylase deficiency.

Type 2 PCT and HEP can be distinguished from type 1 by finding decreased URO-decarboxylase in erythrocytes. URO-decarboxylase activity in liver, erythrocytes, and cultured skin fibroblasts in type 2 PCT is ~50% of normal in affected individuals and in family members with latent disease. In HEP, the URO-decarboxylase activity is markedly deficient, with typical levels of 3–10% of normal. Over 120 mutations have been identified in the *UROD* gene (Human Gene Mutation Database; *www.hgmd.org*). Of the mutations listed in the database, ~65% are missense or nonsense and ~10% are splice-site mutations. Most *UROD* mutations have been identified in only one or two families.

TREATMENT

TREATMENT

Porphyria Cutanea Tarda

Alcohol, estrogens, iron supplements, and, if possible, any drugs that may exacerbate the disease should be discontinued, but this step does not always lead to improvement. A complete response can almost always be achieved by the standard therapy, repeated phlebotomy, to reduce hepatic iron. A unit (450 mL) of blood can be removed every 1–2 weeks. The aim is to gradually reduce excess hepatic iron until the serum ferritin level reaches the lower limits of normal. Because iron overload is not marked in most cases, remission may occur after only five or six phlebotomies; however, PCT patients with hemochromatosis may require more treatments to bring their iron levels down to the normal range. To document improvement in PCT, it is most convenient to follow the total plasma porphyrin concentration, which becomes normal

sometime after the target ferritin level is reached. Hemoglobin levels or hematocrits and serum ferritin should be followed closely to prevent development of iron deficiency and anemia. After remission, continued phlebotomy may not be needed. Plasma porphyrin levels are followed at 6- to 12-month intervals for early detection of recurrences, which are treated by additional phlebotomy.

An alternative when phlebotomy is contraindicated or poorly tolerated is a low-dose regimen of chloroquine or hydroxychloroquine, both of which complex with the excess porphyrins and promote their excretion. Small doses (e.g., 125 mg chloroquine phosphate twice weekly) should be given, because standard doses can induce transient, sometimes marked increases in photosensitivity and hepatocellular damage. Recent studies indicate that low-dose hydroxychloroquine is as safe and effective as phlebotomy in PCT. Hepatic imaging can diagnose or exclude complicating hepatocellular carcinoma. Treatment of PCT in patients with end-stage renal disease is facilitated by administration of erythropoietin.

HEREDITARY COPROPORPHYRIA

HCP is an autosomal dominant hepatic porphyria that results from the half-normal activity of COPRO-oxidase. The disease presents with acute attacks, as in AIP. Cutaneous photosensitivity also may occur, but much less commonly than in VP. HCP patients may have acute attacks and cutaneous photosensitivity together or separately. HCP is less common than AIP and VP. Homozygous dominant HCP and harderoporphyria, a biochemically distinguishable variant of HCP, present with clinical symptoms in children (see below).

Clinical Features

HCP is influenced by the same factors that cause attacks in AIP. The disease is latent before puberty, and symptoms, which are virtually identical to those of AIP, are more common in women. HCP is generally less severe than AIP. Blistering skin lesions are identical to PCT and VP and begin in childhood in rare homozygous cases.

Diagnosis

COPRO III is markedly increased in the urine and feces in symptomatic patients, and often persists, especially in feces, when there are no symptoms. Urinary ALA and PBG levels are increased (but less than in AIP) during acute attacks, but may revert to normal more quickly than in AIP when symptoms resolve. Plasma porphyrins are usually normal or only slightly increased, but they may be higher in cases with skin lesions. The diagnosis of HCP is readily confirmed by increased fecal porphyrins consisting almost entirely of COPRO III, which distinguishes it from other porphyrias.

Although the diagnosis can be confirmed by measuring COPRO-oxidase activity, the assays for this mitochondrial enzyme are not available and require cells other than erythrocytes. To date, >65 mutations have been identified in the *CPOX* gene, 67% of which are missense or nonsense (Human Gene Mutation Database; *www.hgmd.org*). Detection of a *CPOX* mutation in a symptomatic individual permits the identification of asymptomatic family members.

TREATMENT

TREATMENT

Hereditary Coproporphyria

Neurologic symptoms are treated as in AIP (see above). Phlebotomy and chloroquine are not effective for the cutaneous lesions.

VARIEGATE PORPHYRIA

VP is an autosomal dominant hepatic porphyria that results from the deficient activity of PROTO-oxidase, the seventh enzyme in the heme biosynthetic pathway, and can present with neurologic symptoms, photosensitivity, or both. VP is particularly common in South Africa, where 3 of every 1000 whites have the disorder. Most are descendants of a couple who emigrated from Netherlands to South Africa in 1688. In other countries, VP is less common than AIP. Rare cases of homozygous dominant VP, presenting in childhood with cutaneous symptoms, also have been reported.

Clinical Features

VP can present with skin photosensitivity, acute neurovisceral crises, or both. In two large studies of VP patients, ~60% had only skin lesions, 20% had only acute attacks, and ~20% had both. Acute attacks are identical to those in AIP and are precipitated by the same factors as AIP (see

above). Blistering skin manifestations are similar to those in PCT, but are more difficult to treat and usually are of longer duration. Homozygous VP is associated with photosensitivity, neurologic symptoms, and developmental disturbances, including growth retardation, in infancy or childhood; all cases had increased erythrocyte levels of zinc protoporphyrin, a characteristic finding in all homozygous porphyrias so far described.

Diagnosis

Urinary ALA and PBG levels are increased during acute attacks, but may return to normal more quickly than in AIP. Increases in fecal protoporphyrin and COPRO III and in urinary COPRO III are more persistent. Plasma porphyrin levels also are increased, particularly when there are cutaneous lesions. VP can be distinguished rapidly from all other porphyrias by examining the fluorescence emission spectrum of porphyrins in plasma since VP has a unique fluorescence peak at neutral pH.

Assays of PROTO-oxidase activity in cultured fibroblasts or lymphocytes are not widely available. Over 180 mutations have been identified in the *PPOX* gene from unrelated VP patients (Human Gene Mutation Database; *www.hgmd.org*). The missense mutation R59W is the common mutation in most South Africans with VP of Dutch descent. Five missense mutations were common in English and French VP patients; however, most mutations have been found in only one or two families.

TREATMENT

TREATMENT

Variegate Porphyria

Acute attacks are treated as in AIP, and hemin should be started early in most cases. Other than avoiding sun exposure, there are few effective measures for treating the skin lesions. β-Carotene, phlebotomy, and chloroquine are not helpful.

THE ERYTHROPOIETIC PORPHYRIAS

In the erythropoietic porphyrias, excess porphyrins from bone marrow erythrocyte precursors are transported via the plasma to the skin and lead to cutaneous photosensitivity.

X-LINKED SIDEROBLASTIC ANEMIA

XLSA results from the deficient activity of the erythroid form of ALA-synthase (ALA-synthase 2) and is associated with ineffective erythropoiesis, weakness, and pallor.

Clinical Features

Typically, males with XLSA develop refractory hemolytic anemia, pallor, and weakness during infancy. They have secondary hypersplenism, become iron overloaded, and can develop hemosiderosis. The severity depends on the level of residual erythroid ALA-synthase activity and on the responsiveness of the specific mutation to pyridoxal 5′-phosphate supplementation (see below). Peripheral blood smears reveal a hypochromic, microcytic anemia with striking anisocytosis, poikilocytosis, and polychromasia; the leukocytes and platelets appear normal. Hemoglobin content is reduced, and the mean corpuscular volume and mean corpuscular hemoglobin concentration are decreased. Patients with milder, late-onset disease have been reported recently.

Diagnosis

Bone marrow examination reveals hypercellularity with a left shift and megaloblastic erythropoiesis with an abnormal maturation. A variety of Prussian blue-staining sideroblasts are observed. Levels of urinary porphyrin precursors and of both urinary and fecal porphyrins are normal. The activity of erythroid ALA-synthase 2 is decreased in bone marrow, but this enzyme is difficult to measure in the presence of the normal ALA-synthase 1 housekeeping enzyme. Definitive diagnosis requires the demonstration of mutations in the erythroid *ALAS2* gene.

TREATMENT

TREATMENT

X-Linked Sideroblastic Anemia

The severe anemia may respond to pyridoxine supplementation. This cofactor is essential for ALA-synthase activity, and mutations in the pyridoxine binding site of the enzyme have been found in several responsive patients. Cofactor supplementation may make it possible to eliminate or reduce the frequency of transfusion. Unresponsive patients may be transfusion dependent and require chelation therapy.

CONGENITAL ERYTHROPOIETIC PORPHYRIA

CEP, also known as Günther's disease, is an autosomal recessive disorder. It is due to the markedly deficient, but not absent, activity of URO-synthase and the resultant accumulation of URO I and COPRO I isomers. CEP is associated with hemolytic anemia and cutaneous lesions.

Clinical Features

Severe cutaneous photosensitivity typically begins from birth. The skin over light-exposed areas is friable, and bullae and vesicles are prone to rupture and infection. Skin thickening, focal hypo- and hyperpigmentation, and hypertrichosis of the face and extremities are characteristic. Secondary infection of the cutaneous lesions can lead to disfigurement of the face and hands. Porphyrins are deposited in teeth and in bones. As a result, the teeth are brownish and fluoresce on exposure to long-wave ultraviolet light. Hemolysis is due to the marked increase in erythrocyte porphyrins and leads to splenomegaly. Adults with a milder later-onset form of the disease also have been described.

Diagnosis

URO and COPRO (mostly type I isomers) accumulate in the bone marrow, erythrocytes, plasma, urine, and feces. The predominant porphyrin in feces is COPRO I. The diagnosis of CEP can be confirmed by demonstration of markedly deficient URO-synthase activity and/or by the identification of specific mutations in the *UROS* gene. The disease can be detected in utero by measuring porphyrins in amniotic fluid and URO-synthase activity in cultured amniotic cells or chorionic villi, or by the detection of the family's specific gene mutations. Molecular analyses of the mutant alleles from unrelated patients have revealed the presence of >50 mutations in the *UROS* gene, including four in the erythroid-specific promoter of the *UROS* gene. Genotype/phenotype correlations can predict the severity of the disease. The CEP phenotype may be modulated by sequence variations in the erythroid-specific ALA-synthase 2, mutation of which typically causes XLP. One mutation (p.ArgR216WTrp) in *GATA1*, encoding the X-linked erythroid-specific transcription factor GATA binding protein 1 (*GATA1*), has been identified in an individual with CEP, thrombocytopenia, and β-thalassemia.

TREATMENT

TREATMENT

Congenital Erythropoietic Porphyria

Severe cases often require transfusions for anemia. Chronic transfusions of sufficient blood to suppress erythropoiesis are effective in reducing porphyrin production but result in iron overload. Splenectomy may reduce hemolysis and decrease transfusion requirements. Protection from sunlight and from minor skin trauma is important. Complicating bacterial infections should be treated promptly. Recently, bone marrow and cord blood transplantation has proven curative in several transfusion-dependent children, providing the rationale for stem cell gene therapy.

ERYTHROPOIETIC PROTOPORPHYRIA

EPP is an autosomal recessive disorder resulting from the deficient activity of FECH, the last enzyme in the heme biosynthetic pathway. EPP is the most common erythropoietic porphyria in children and, after PCT, the second most common porphyria in adults. EPP patients have FECH activities as low as 15–25% of normal in lymphocytes and cultured fibroblasts. Protoporphyrin accumulates in bone marrow reticulocytes and then appears in plasma, is taken up in the liver, and is excreted in bile and feces. Protoporphyrin transported to the vessels in the skin causes the nonblistering phototoxicity. In most symptomatic patients (~90%) with this disorder, a deleterious mutation in one *FECH* allele was inherited with the relatively common (~10% of Caucasians) intronic 3 (IVS3) alteration (IVS3–48T>C) on the other allele that results in the low expression of the normal enzyme. In about 2% of EPP families, two *FECH* deleterious mutations have been found.

XLP is a less common condition with the same phenotype in affected males, including increased erythrocyte protoporphyrin levels resulting from gain-of-function mutations in the last exon of the erythroid-specific form of 5-aminolevulinate-synthase 2 (*ALAS2*). These mutations delete ALAS2 C-terminal amino acids resulting in its increased activity and the subsequent accumulation of protoporphyrin. Manifestations in female heterozygotes with XLP can range from asymptomatic to as severe as their affected male relatives. The variation in the presence and severity of manifestations in XLP heterozygotes results primarily from random X-chromosomal inactivation. XLP accounts for ~2–10% of cases with the EPP phenotype in Europe and North America.

Clinical Features

In EPP and male XLP patients, skin photosensitivity, which differs from that in other porphyrias, usually begins in childhood and consists of pain, tingling, and itching occurring within minutes of sunlight exposure **(Fig. 409-4)**. Photosensitivity is associated with substantial elevations in erythrocyte protoporphyrin and occurs only in patients with genotypes that result in FECH activities below ~35% of normal. Vesicular lesions are uncommon. Redness, swelling, burning, and itching can develop shortly after sun exposure and resemble angioedema. Pain symptoms may seem out of proportion to the visible skin involvement. Sparse vesicles and bullae occur in ~10% of cases. Chronic skin changes may include lichenification, leathery pseudovesicles, labial grooving, and nail changes. Severe scarring is rare, as are pigment changes, friability, and hirsutism. Unless hepatic or other complications develop, protoporphyrin levels and symptoms of photosensitivity remain remarkably stable over many years in most patients. Factors that exacerbate the hepatic porphyrias play little or no role in EPP or XLP.

FIGURE 409-4

Erythema and edema of the hands due to acute photosensitivity in a 10-year-old boy with erythropoietic protoporphyria. *(From P Poblette-Gutierrez et al: Eur J Dermatol 16:230, 2006.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The primary source of excess protoporphyrin is the bone marrow reticulocytes. Erythrocyte protoporphyrin is free (not complexed with zinc) and is mostly bound to hemoglobin. In plasma, protoporphyrin is bound to albumin. Hemolysis and anemia are usually absent or mild.

Although EPP is an erythropoietic porphyria, up to 20% of EPP patients may have minor abnormalities of liver function, and in about 5% of these patients the accumulation of protoporphyrins causes chronic liver disease that can progress to liver failure requiring transplantation. Protoporphyrin is insoluble, and excess amounts form crystalline structures in liver cells (Fig. 409-4) and can decrease hepatic bile flow. Studies in the mouse model of EPP have shown that the bile duct epithelium may be damaged by toxic bile, leading to biliary fibrosis. Thus, rapidly progressive liver disease appears to be related to the cholestatic effects of protoporphyrins and is associated with increasing hepatic protoporphyrin levels due to impaired hepatobiliary excretion and increased photosensitivity. The hepatic complications also are often characterized by increasing levels of protoporphyrins in erythrocytes and plasma as well as severe abdominal and back pains, especially in the right upper quadrant. Gallstones composed at least in part of protoporphyrin occur in some patients. Hepatic complications appear to be higher in autosomal recessive EPP due to two *FECH* mutations and in males with XLP.

Diagnosis

A substantial increase in erythrocyte protoporphyrin, which is predominantly free and not complexed with zinc, is the hallmark of EPP. Protoporphyrin levels are also variably increased in bone marrow, plasma, bile, and feces. Erythrocyte protoporphyrin concentrations are increased in other conditions such as lead poisoning, iron deficiency, various hemolytic disorders, all homozygous forms of other porphyrias, and sometimes even in acute porphyrias. In all these conditions, however, in contrast to EPP, protoporphyrin is complexed with zinc. Therefore, after an increase in erythrocyte protoporphyrin is found in a suspected EPP patient, it is important to confirm the diagnosis by an

assay that distinguishes free and zinc-complexed protoporphyrin. Erythrocytes in EPP also exhibit red fluorescence under a fluorescence microscopy at 620 nm. Urinary levels of porphyrins and porphyrin precursors are normal. FECH activity in cultured lymphocytes or fibroblasts is decreased (<30% of normal mean). DNA diagnosis by mutation analysis is recommended to detect the causative *FECH* mutation(s) and/or the presence of the IVS3–48T>C low expression allele. To date, >190 mutations have been identified in the *FECH* gene, many of which result in an unstable or absent enzyme protein (null alleles) (Human Gene Mutation Database; *www.hgmd.org*).

In XLP, the erythrocyte protoporphyrin levels appear to be higher than in EPP and the proportions of free and zinc protoporphyrins may reach 50%. To date, four *ALAS2* mutations, three deletions of one to four bases, and one novel nonsense mutation have been described, which markedly increase ALA-synthase 2 activity and cause XLP. XLP accounts for about 2% of patients with the EPP phenotype in Western Europe. Recent studies show that about 10% of North American patients with the EPP phenotype have XLP.

TREATMENT

TREATMENT

Erythropoietic Protoporphyria

Avoiding sunlight exposure and wearing clothing designed to provide protection for conditions with chronic phototoxicity are essential. Various other treatments, including oral β-Carotene, have proven of little benefit. Afamelanotide, an α-melanocyte-stimulating hormone (MSH) analogue, that stimulates tanning, has been approved for the treatment of EPP and XLP in the European Union by the European Medicines Agency. Approval by the U.S. Food and Drug Administration is pending at this time.

Treatment of hepatic complications, which may be accompanied by motor neuropathy, is difficult. Cholestyramine and other porphyrin absorbents such as activated charcoal may interrupt the enterohepatic circulation of protoporphyrin and promote its fecal excretion, leading to some improvement. Splenectomy may be helpful when the disease is accompanied by hemolysis and significant splenomegaly. Plasmapheresis and intravenous hemin are sometimes beneficial.

Liver transplantation has been carried out in some EPP and XLP patients with severe liver complications and is often successful in the short term. However, the disease often recurs in the transplanted liver due to continued bone marrow production of excess protoporphyrin. In a retrospective study of 17 liver-transplanted EPP patients, 11 (65%) had recurrent EPP liver disease. Posttransplantation treatment with hematin and plasmapheresis should be considered to prevent the recurrence of liver disease. However, bone marrow transplantation, which has been successful in human EPP and which prevented liver disease in a mouse model, should be considered after liver transplantation, if a suitable donor can be found.

ACKNOWLEDGMENT

The authors thank Dr. Karl E. Anderson for his review of the manuscript and helpful comments and suggestions. This work is supported in part by the Porphyrias Consortium (U54 DK083909), a part of the National Institutes of Health (NIH) Rare Disease Clinical Research Network (RDCRN), supported through collaboration between the NIH Office of Rare Diseases Research (ORDR) at the National Center for Advancing Translational Science (NCATS), and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. MB is supported by a career development award K23-DK-095946.

FURTHER READING

Anderson KE et al: Recommendations for the diagnosis and treatment of the acute porphyrias. Ann Intern Med 142:439, 2005. [PubMed: 15767622]

Balwani M, Desnick RJ: The porphyrias: Advances in diagnosis and treatment. Blood 120:4496, 2012. [PubMed: 22791288]

Bonkovsky HL et al: Acute porphyrias in the USA: Features of 108 subjects from porphyrias consortium. Am J Med 127:1233, 2015.

Chen B et al: Acute intermittent porphyria: Predicted pathogenicity of HMBS variants indicates extremely low penetrance of the autosomal dominant disease. Hum Mutat 37:1215, 2016. [PubMed: 27539938]

Puy H et al: Porphyrias. Lancet 375:924, 2010. [PubMed: 20226990]

Langendonk JG et al: Afamelanotide for erythropoietic protoporphyria. N Engl J Med 373:48, 2015. [PubMed: 26132941]

Pischik E, Kauppinen R: An update of clinical management of acute intermittent porphyria. Appl Clin Genet 8:201, 2015. [PubMed: 26366103]

Schulenburg-Brand D et al: The cutaneous porphyrias. Dermatol Clin 32:369, 2014. [PubMed: 24891059]

Stein PE et al: Update review of the acute porphyrias. Br J Haematol 176:527, 2017. [PubMed: 27982422]

Yasuda M et al: AAV8-mediated gene therapy prevents induced biochemical attacks of intermittent porphyria and improves neuromotor function. Mol Ther 18:17, 2010. [PubMed: 19861948]

Yasuda M et al: Liver transplantation for acute intermittent porphyria: Biochemical and pathologic studies of the explanted liver. Mol Med 21:487, 2015. [PubMed: 26062020]

Chapter 410: Disorders of Purine and Pyrimidine Metabolism

John N. Mecchella; Christopher M. Burns

INTRODUCTION

Purines (adenine and guanine) and pyrimidines (cytosine, thymine, uracil) serve fundamental roles in the replication of genetic material, gene transcription, protein synthesis, and cellular metabolism. Disorders that involve abnormalities of nucleotide metabolism range from relatively common diseases such as hyperuricemia and gout, in which there is increased production or impaired excretion of a metabolic end product of purine metabolism (uric acid), to rare enzyme deficiencies that affect purine and pyrimidine synthesis or degradation. Understanding these biochemical pathways has led, in some instances, to the development of specific forms of treatment, such as the use of allopurinol and febuxostat to reduce uric acid production.

URIC ACID METABOLISM

Uric acid is the final breakdown product of purine degradation in humans. It is a weak diprotic acid with pK_a values of 5.75 and 10.3. Urates, the ionized forms of uric acid, predominate in plasma, extracellular fluid, and synovial fluid, with ~98% existing as monosodium urate at pH 7.4.

Plasma is saturated with monosodium urate at a concentration of 405 µmol/L (6.8 mg/dL) at 37°C. At higher concentrations, plasma is therefore supersaturated—a situation that creates the potential for urate crystal precipitation. However, plasma urate concentrations can reach 4800 µmol/L (80 mg/dL) without precipitation, perhaps because of the presence of solubilizing substances.

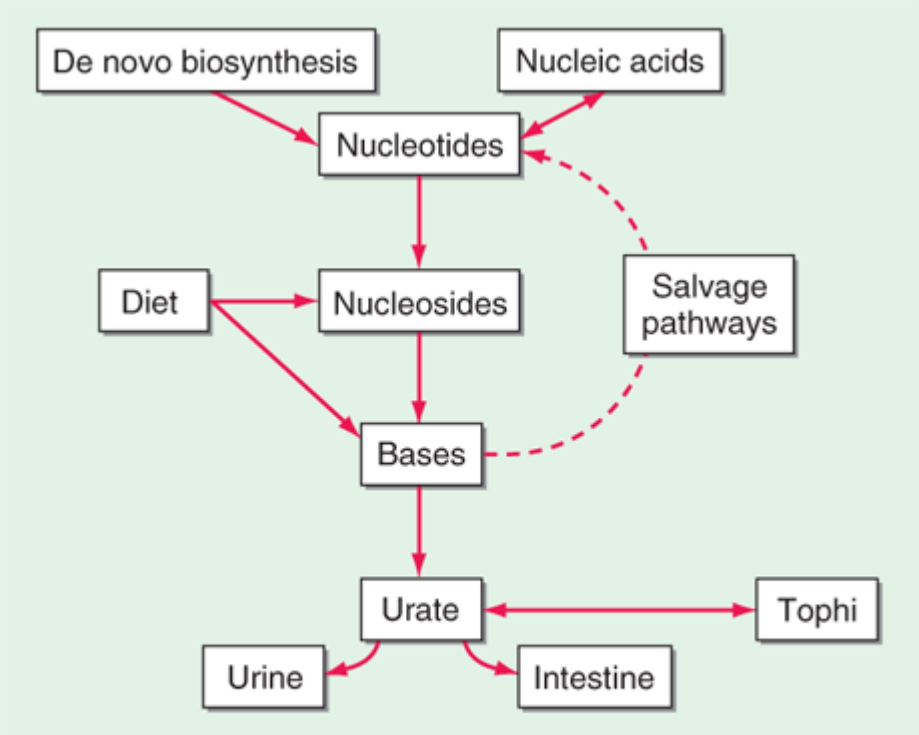
The pH of urine greatly influences the solubility of uric acid. At pH 5.0, urine is saturated with uric acid at concentrations ranging from 360 to 900 µmol/L (6–15 mg/dL). At pH 7.0, saturation is reached at concentrations from 9840 to 12,000 µmol/L (158–200 mg/dL). Ionized forms of uric acid in urine include monosodium, disodium, potassium, ammonium, and calcium urates.

Although purine nucleotides are synthesized and degraded in all tissues, urate is produced only in tissues that contain xanthine oxidase, primarily the liver and small intestine. Urate production varies with the purine content of the diet and with rates of purine biosynthesis, degradation, and salvage **(Fig. 410-1)**. Normally,

two-thirds to three-fourths of urate is excreted by the kidneys, and most of the remainder is eliminated through the intestines.

FIGURE 410-1

The total-body urate pool is the net result between urate production and excretion. Urate production is influenced by dietary intake of purines and the rates of de novo biosynthesis of purines from nonpurine precursors, nucleic acid turnover, and salvage by phosphoribosyltransferase activities. The formed urate is normally excreted by urinary and intestinal routes. Hyperuricemia can result from increased production, decreased excretion, or a combination of both mechanisms. When hyperuricemia exists, urate can precipitate and deposit in tissues as tophi.

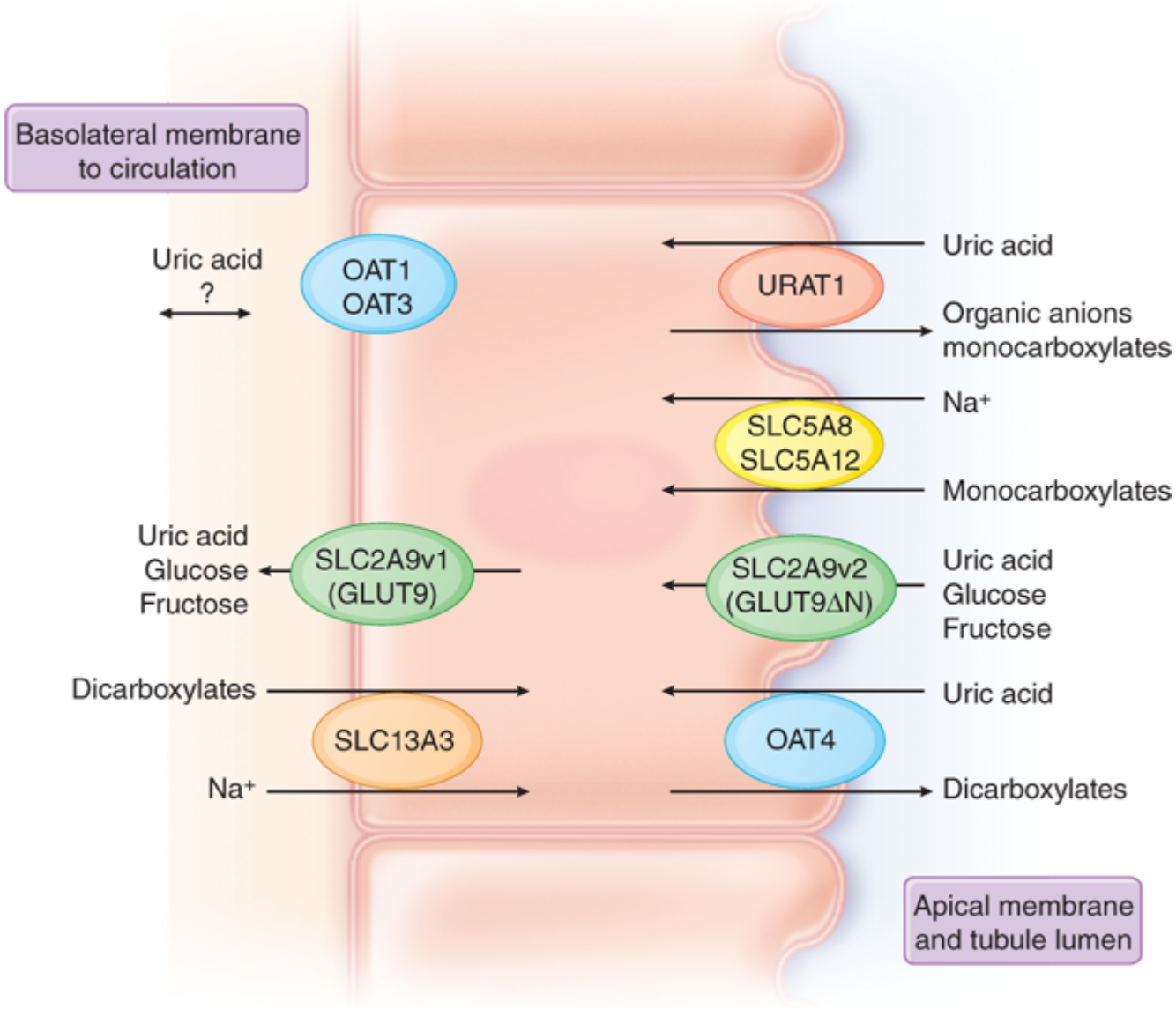


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The kidneys clear urate from the plasma and maintain physiologic balance by utilizing specific organic anion transporters (OATs), including urate transporter 1 (URAT1, SLC22A12) **(Fig. 410-2)**. In humans, OAT1 (SLC22A6), OAT2 (SLC22A7), and OAT3 (SLC22A8) are located on the basolateral membrane of renal proximal tubule cells. OAT4 (SLC22A11), OAT10 (SLC22A13), and URAT1 are located on the apical brush-border membrane of these cells. The latter transporters carry urate and other organic anions into the tubular cells from the lumen in exchange for intracellular organic anions. Once inside the cell, urate must pass to the basolateral side of the lumen in a process controlled by voltage-dependent carriers, including glucose transporter 9 (GLUT9, SLC2A9). *Uricosuric* compounds **(Table 410-1)** directly inhibit URAT1 on the apical side of the tubular cell (so-called *cis*-inhibition). In contrast, *antiuricosuric* compounds (those that promote hyperuricemia), such as nicotinate, pyrazinoate, lactate, and other aromatic organic acids, serve as the exchange anion inside the cell, thereby stimulating anion exchange and urate reabsorption (*trans*-stimulation). The activities of URAT1, other OATs, and sodium anion transporters result in excretion of 8–12% of the filtered urate as uric acid.

FIGURE 410-2

Schematic for handling of uric acid by the kidney. A complex interplay of transporters on both the apical and basolateral aspects of the renal tubule epithelial cell is involved in the reabsorption of uric acid. See text for details. Most uricosuric compounds inhibit URAT1 on the apical side, as well as OAT1, OAT3, and GLUT9 on the basolateral side.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 410-1

Medications with Uricosuric Activity

Acetohexamide	Glyceryl guaiacolate
Adrenocorticotropic hormone	Glycopyrrolate
Ascorbic acid	Halofenate
Azauridine	Losartan
Benzbromarone	Meclofenamate
Calcitonin	Phenolsulfonphthalein
Chlorprothixene	Phenylbutazone
Citrate	Probenecid
Dicumarol	Radiographic contrast agents
Diflunisal	Salicylates (>2 g/d)
Estrogens	Sulfinpyrazone
Fenofibrate	Tetracycline that is outdated
Glucocorticoids	Zoxazolamine

Most children have serum urate concentrations of 180–240 µmol/L (3–4 mg/dL). Levels begin to rise in males during puberty but remain low in females until menopause. The most recent mean serum urate values for men and premenopausal women in the United States are 415 and 360 µmol/L (6.14 and 4.87 mg/dL), respectively, according to National Health and Nutrition Evaluation Survey (NHANES) data for 2007–2008. After menopause, values for women increase to approximately those for men. In adulthood, concentrations rise steadily over time and vary with height, body weight, blood pressure, renal function, and alcohol intake.

HYPERURICEMIA

Hyperuricemia can result from increased production or decreased excretion of uric acid or from a combination of the two processes. Sustained hyperuricemia predisposes some individuals to develop clinical

manifestations including gouty arthritis **(Chap. 365)**, urolithiasis, and renal dysfunction (see below).

In general, hyperuricemia is defined as a plasma (or serum) urate concentration >405 μmol/L (>6.8 mg/dL). The risk of developing gouty arthritis or urolithiasis increases with higher urate levels and escalates in proportion to the degree of elevation. The prevalence of hyperuricemia is increasing among ambulatory adults and even more markedly among hospitalized patients. The prevalence of gout in the United States more than doubled between the 1960s and the 1990s. Based on NHANES data from 2007 to 2008, these trends continue, with an approximate prevalence of gout among men of 5.9% (6.1 million) and among women of 2.0% (2.2 million). Mean serum urate levels rose to 6.14 mg/dL among men and 4.87 mg/dL among women, with consequent hyperuricemia prevalences of 21.2 and 21.6%, respectively (with *hyperuricemia* defined as a serum urate level of >7.0 mg/dL [415 μmol/L] for men and >5.7 mg/dL [340 μmol/L] for women). These numbers represent a 1.2% increase in the prevalence of gout, a 0.15-mg/dL increase in the serum urate level, and a 3.2% increase in the prevalence of hyperuricemia over figures reported in NHANES-III (1988–1994). These rises are thought to be driven by increased obesity and hypertension and perhaps also by better medical care and increased longevity.

CAUSES OF HYPERURICEMIA

Hyperuricemia may be classified as primary or secondary, depending on whether the cause is innate or an acquired disorder. However, it is more useful to classify hyperuricemia in relation to the underlying pathophysiology—i.e., whether it results from increased production, decreased excretion, or a combination of the two (Fig. 410-1, **Table 410-2**).

TABLE 410-2

Classification of Hyperuricemia by Pathophysiology

Urate Overproduction		
Primary idiopathic HPRT deficiency PRPP synthetase overactivity Hemolytic processes Lymphoproliferative diseases	Myeloproliferative diseases Polycythemia vera Psoriasis Paget's disease Glycogenosis III, V, and VII	Rhabdomyolysis Exercise Alcohol Obesity Purine-rich diet
Decreased Uric Acid Excretion		
Primary idiopathic Renal insufficiency Polycystic kidney disease Diabetes insipidus Hypertension Acidosis Lactic acidosis Diabetic ketoacidosis	Starvation ketosis Berylliosis Sarcoidosis Lead intoxication Hyperparathyroidism Hypothyroidism Toxemia of pregnancy Bartter's syndrome Down syndrome	Drug ingestion Salicylates (<2 g/d) Diuretics Alcohol Levodopa Ethambutol Pyrazinamide Nicotinic acid Cyclosporine
Combined Mechanism		
Glucose-6-phosphatase deficiency	Fructose-1-phosphate aldolase deficiency	Alcohol Shock

Abbreviations: HPRT, hypoxanthine phosphoribosyltransferase; PRPP, phosphoribosylpyrophosphate.

Increased Urate Production

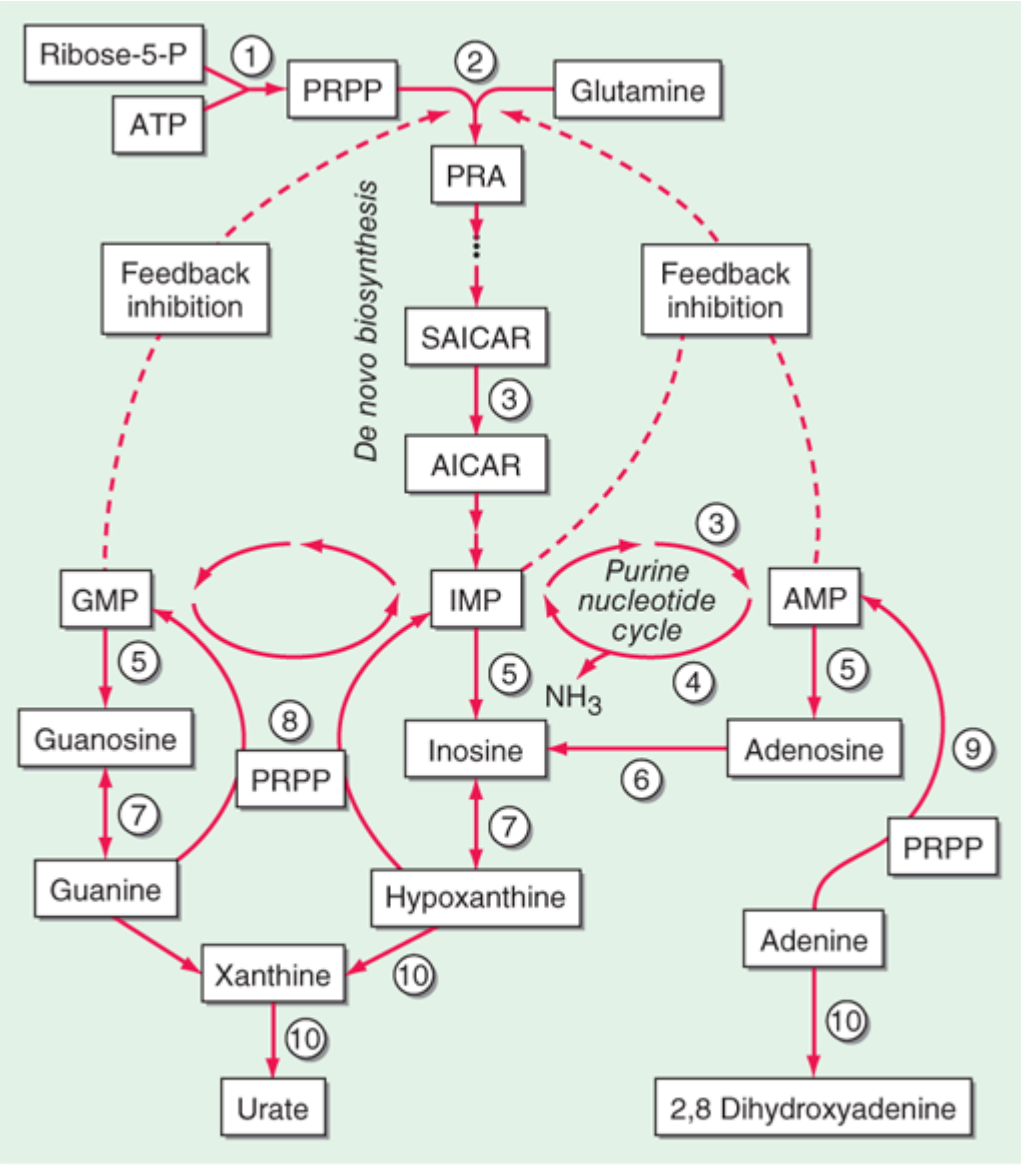
Diet contributes to the serum urate concentration in proportion to its purine content. Strict restriction of purine intake reduces the mean serum urate level by ~60 μmol/L (~1 mg/dL) and urinary uric acid excretion by ~1.2 mmol/d (~200 mg/d). Foods high in nucleic acid content include liver, "sweetbreads" (i.e., thymus and pancreas), kidney, and anchovy.

Endogenous sources of purine production also influence the serum urate level **(Fig. 410-3)**. De novo purine biosynthesis is a multistep process that forms inosine monophosphate (IMP). The rates of purine biosynthesis and urate production are predominantly determined by amidophosphoribosyltransferase

(amidoPRT), which combines phosphoribosylpyrophosphate (PRPP) and glutamine. A secondary regulatory pathway is the salvage of purine bases by hypoxanthine phosphoribosyltransferase (HPRT). HPRT catalyzes the combination of the purine bases hypoxanthine and guanine with PRPP to form the respective ribonucleotides IMP and guanosine monophosphate (GMP).

FIGURE 410-3

Abbreviated scheme of purine metabolism. (1) Phosphoribosylpyrophosphate (PRPP) synthetase, (2) amidophosphoribosyltransferase (amidoPRT), (3) adenylosuccinate lyase, (4) (myo-)adenylate (AMP) deaminase, (5) 5′-nucleotidase, (6) adenosine deaminase, (7) purine nucleoside phosphorylase, (8) hypoxanthine phosphoribosyltransferase (HPRT), (9) adenine phosphoribosyltransferase (APRT), and (10) xanthine oxidase. AICAR, aminoimidazole carboxamide ribotide; ATP, adenosine triphosphate; GMP, guanylate; IMP, inosine monophosphate; PRA, phosphoribosylamine; SAICAR, succinylaminoimidazole carboxamide ribotide.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Serum urate levels are closely coupled to the rates of de novo purine biosynthesis, which is driven in part by the level of PRPP, as evidenced by two X-linked inborn errors of purine metabolism **(Table 410-3)**. Both increased PRPP synthetase activity and HPRT deficiency are associated with overproduction of purines, hyperuricemia, and hyperuricaciduria (see below for clinical descriptions).

TABLE 410-3

Inborn Errors of Purine Metabolism

Enzyme	Activity	Inheritance	Clinical Features	Laboratory Features
Hypoxanthine phosphoribosyltransferase	Complete deficiency	X-linked	Self-mutilation, choreoathetosis, gout, and uric acid lithiasis	Hyperuricemia, hyperuricosuria
	Partial deficiency	X-linked	Gout and uric acid lithiasis	Hyperuricemia, hyperuricosuria
Phosphoribosylpyrophosphate synthetase	Overactivity	X-linked	Gout, uric acid lithiasis, and deafness	Hyperuricemia, hyperuricosuria
Adenine phosphoribosyltransferase	Deficiency	Autosomal recessive	2,8-Dihydroxyadenine lithiasis	—
Xanthine oxidase	Deficiency	Autosomal recessive	Xanthinuria and xanthine lithiasis	Hypouricemia, hypouricosuria
Adenylosuccinate lyase	Deficiency	Autosomal recessive	Autism and psychomotor retardation	—
Myoadenylate deaminase	Deficiency	Autosomal recessive	Myopathy with exercise intolerance or asymptomatic	—
Adenosine deaminase	Deficiency	Autosomal recessive	Severe combined immunodeficiency disease and chondro-osseous dysplasia	—
Purine nucleoside phosphorylase	Deficiency	Autosomal recessive	T cell–mediated immunodeficiency	—

Accelerated purine nucleotide degradation can also cause hyperuricemia—i.e., with conditions of rapid cell turnover, proliferation, or cell death, as in leukemic blast crises, cytotoxic therapy for malignancy, hemolysis,

or rhabdomyolysis. Hyperuricemia can result from excessive degradation of skeletal muscle ATP after strenuous physical exercise or status epilepticus and in glycogen storage disease types III, V, and VII **(Chap. 412)**. The hyperuricemia of myocardial infarction, smoke inhalation, and acute respiratory failure may also be related to accelerated breakdown of ATP.

Decreased Uric Acid Excretion

More than 90% of individuals with sustained hyperuricemia have a defect in the renal handling of uric acid. For any given plasma urate concentration, patients who have gout excrete ~40% less uric acid than those who do not. When plasma urate levels are raised by purine ingestion or infusion, uric acid excretion increases in patients with and without gout; however, in those with gout, plasma urate concentrations must be 60–120 µmol/L (1–2 mg/dL) higher than normal to achieve equivalent uric acid excretion rates.

Diminished uric acid excretion could theoretically result from decreased glomerular filtration, decreased tubular secretion, or enhanced tubular reabsorption. Decreased urate filtration does not appear to cause primary hyperuricemia but does contribute to the hyperuricemia of renal insufficiency. Although hyperuricemia is invariably present in chronic renal disease, the correlation among serum creatinine, urea nitrogen, and urate concentrations is poor. Extrarenal clearance of uric acid increases as renal damage becomes more severe.

Many agents that cause hyperuricemia exert their effects by stimulating reabsorption rather than inhibiting secretion. This stimulation appears to occur through a process of “priming” renal urate reabsorption through the sodium-dependent loading of proximal tubular epithelial cells with anions capable of *trans*-stimulating urate reabsorption. The sodium-coupled monocarboxyl transporters SMCT1 and 2 (SLC5A8, SLC5A12) in the brush border of the proximal tubular cells mediate sodium-dependent loading of these cells with monocarboxylates. A similar transporter, SLC13A3, mediates sodium-dependent influx of dicarboxylates into the epithelial cell from the basolateral membrane. Some of these carboxylates are well known to cause hyperuricemia, including pyrazinoate (from pyrazinamide treatment), nicotinate (from niacin therapy), and the organic acids lactate, β-hydroxybutyrate, and acetoacetate. The mono- and divalent anions then become substrates for URAT1 and OAT4, respectively, and are exchanged for uric acid from the proximal tubule. Increased blood levels of these anions result in their increased glomerular filtration and greater reabsorption by proximal tubular cells. The increased intraepithelial cell concentrations lead to increased uric acid reabsorption by promoting URAT1-, OAT4-, and OAT10-dependent anion exchange. Low doses of salicylates also promote hyperuricemia by this mechanism. Sodium loading of proximal tubular cells also provokes urate retention by reducing extracellular fluid volume and increasing angiotensin II, insulin, and parathyroid hormone release. Additional OAT1, OAT2, and OAT3 are involved in the movement of uric acid through the basolateral membrane, although the detailed mechanisms are still being elucidated.

GLUT9 (SLC2A9) is an electrogenic hexose transporter with splicing variants that mediate co-reabsorption of uric acid along with glucose and fructose at the apical membrane (GLUT9ΔN/SLC2A9v2) as well as through the basolateral membrane (SLC2A9v1) and thus into the circulation. GLUT9 has recently been identified as a high-capacity urate transporter, with rates 45–60 times faster than its glucose/fructose transport activity.

GLUT9 may be responsible for the observed association of the consumption of fructose-sweetened soft drinks with an increased risk of hyperuricemia and gout. Genome-wide association scanning suggests that polymorphisms in SLC2A9 may play an important role in susceptibility to gout in the Caucasian population. The presence of one predisposing variant allele increases the relative risk of developing gout by 30–70%, most likely by increasing expression of the shorter isoform, SLC2A9v2 (GLUT9ΔN). Notably, though, genetic polymorphisms explain only ~6% of the differences in serum uric acid levels in Caucasians. Clearly, gout is polygenic and complex, and at this time the utility of genetic testing for relevant polymorphisms remains investigational and of no clinical utility.

Alcohol promotes hyperuricemia because of increased urate production and decreased uric acid excretion. Excessive alcohol consumption accelerates hepatic breakdown of ATP to increase urate production. Alcohol consumption can also induce hyperlacticacidemia, which blocks uric acid secretion. The higher purine content in some alcoholic beverages may also be a factor. Consumption of beer confers a greater risk of gout than liquor, and moderate wine intake does not increase gout risk. Intake of red meat and fructose increases the risk of gout, whereas intake of low-fat dairy products, purine-rich vegetables, whole grains, nuts and legumes, less sugary fruits, coffee, and vitamin C reduces the risk.

EVALUATION

Hyperuricemia does not necessarily represent a disease, nor is it a specific indication for therapy. The decision to treat depends on the cause and the potential consequences of hyperuricemia in each individual.

Quantification of uric acid excretion can be used to determine whether hyperuricemia is caused by overproduction or decreased excretion. On a purine-free diet, men with normal renal function excrete <3.6 mmol/d (600 mg/d). Thus, the hyperuricemia of individuals who excrete uric acid above this level while on a purine-free diet is due to purine overproduction; for those who excrete lower amounts on the purine-free diet, it is due to decreased excretion. If the assessment is performed while the patient is on a regular diet, the level of 4.2 mmol/d (800 mg/d) can be used as the discriminating value.

COMPLICATIONS

The most recognized complication of hyperuricemia is *gouty arthritis*. NHANES 2007–2008 found a prevalence of gout among U.S. adults of 3.9%, with figures of ~6% for men and ~2% for women. The higher the serum urate level, the more likely an individual is to develop gout. In one study, the incidence of gout was 4.9% among individuals with serum urate concentrations >540 µmol/L (>9.0 mg/dL) as opposed to only 0.5% among those with values between 415 and 535 µmol/L (7.0 and 8.9 mg/dL). The complications of gout correlate with both the duration and the severity of hyperuricemia. **For further discussion of gout, see Chap. 365**.

Hyperuricemia also causes several renal problems: (1) nephrolithiasis; (2) urate nephropathy, a rare cause of renal insufficiency attributed to monosodium urate crystal deposition in the renal interstitium; and (3) uric

acid nephropathy, a reversible cause of acute renal failure resulting from deposition of large amounts of uric acid crystals in the renal collecting ducts, pelvis, and ureters.

Nephrolithiasis

Uric acid nephrolithiasis occurs most commonly, but not exclusively, in individuals with gout. In gout, the prevalence of nephrolithiasis correlates with the serum and urinary uric acid levels, reaching ~50% with serum urate levels of 770 µmol/L (13 mg/dL) or urinary uric acid excretion >6.5 mmol/d (1100 mg/d).

Uric acid stones can develop in individuals with no evidence of arthritis, only 20% of whom are hyperuricemic. Uric acid can also play a role in other types of kidney stones. Some individuals who do not have gout but have calcium oxalate or calcium phosphate stones have hyperuricemia or hyperuricaciduria. Uric acid may act as a nidus on which calcium oxalate can precipitate or lower the formation product for calcium oxalate crystallization.

Urate Nephropathy

Urate nephropathy, sometimes referred to as *urate nephrosis*, is a late manifestation of severe gout and is characterized histologically by deposits of monosodium urate crystals surrounded by a giant-cell inflammatory reaction in the medullary interstitium and pyramids. The disorder is now rare and cannot be diagnosed in the absence of gouty arthritis. The lesions may be clinically silent or cause proteinuria, hypertension, and renal insufficiency.

Uric Acid Nephropathy

This reversible cause of acute renal failure is due to precipitation of uric acid in renal tubules and collecting ducts that obstructs urine flow. Uric acid nephropathy develops following sudden urate overproduction and marked hyperuricaciduria. Factors that favor uric acid crystal formation include dehydration and acidosis. This form of acute renal failure occurs most often during an aggressive “blastic” phase of leukemia or lymphoma prior to or coincident with cytolytic therapy but has also been observed in individuals with other neoplasms, following epileptic seizures, and after vigorous exercise with heat stress. Autopsy studies have demonstrated intraluminal precipitates of uric acid, dilated proximal tubules, and normal glomeruli. The initial pathogenic events are believed to include obstruction of collecting ducts with uric acid and obstruction of the distal renal vasculature.

If recognized, uric acid nephropathy is potentially reversible. Appropriate therapy has reduced the mortality rate from ~50% to near zero. Serum levels cannot be relied on for diagnosis because this condition has developed in the presence of urate concentrations varying from 720 to 4800 µmol/L (12–80 mg/dL). The distinctive feature is the urinary uric acid concentration. In most forms of acute renal failure with decreased urine output, urinary uric acid content is either normal or reduced, and the ratio of uric acid to creatinine is <1. In acute uric acid nephropathy, the ratio of uric acid to creatinine in a random urine sample or a 24-h specimen is >1, and a value that high is essentially diagnostic.

HYPERURICEMIA AND METABOLIC SYNDROME

Metabolic syndrome **(Chap. 401)** is characterized by abdominal obesity with visceral adiposity, impaired glucose tolerance due to insulin resistance with hyperinsulinemia, hypertriglyceridemia, increased low-density lipoprotein cholesterol, decreased high-density lipoprotein cholesterol, and hyperuricemia. Hyperinsulinemia reduces the renal excretion of uric acid and sodium. Not surprisingly, hyperuricemia resulting from euglycemic hyperinsulinemia may precede the onset of type 2 diabetes, hypertension, coronary artery disease, and gout in individuals with metabolic syndrome.

TREATMENT

TREATMENT

Hyperuricemia

ASYMPTOMATIC HYPERURICEMIA

Hyperuricemia is present in ~21% of the population and in at least 25% of hospitalized individuals. The vast majority of hyperuricemic persons are at no clinical risk. In the past, the association of hyperuricemia with cardiovascular disease and renal failure led to the use of urate-lowering agents for patients with asymptomatic hyperuricemia. This practice is no longer recommended except for individuals receiving cytolytic therapy for neoplastic disease, who are treated with urate-lowering agents in an effort to prevent uric acid nephropathy. Because hyperuricemia can be a component of the metabolic syndrome, its presence is an indication to screen for and aggressively treat any accompanying obesity, hyperlipidemia, diabetes mellitus, or hypertension.

Hyperuricemic individuals, especially those with higher serum urate levels, are at risk for the development of gouty arthritis. However, most hyperuricemic persons never develop gout, and prophylactic treatment is not indicated. Furthermore, neither structural kidney damage nor tophi are identifiable before the first attack. Reduced renal function cannot be attributed to asymptomatic hyperuricemia, and treatment of asymptomatic hyperuricemia does not alter the progression of renal dysfunction in patients with renal disease. An increased risk of stone formation in those with asymptomatic hyperuricemia has not been established.

Thus, because treatment with specific antihyperuricemic agents entails inconvenience, cost, and potential toxicity, routine treatment of asymptomatic hyperuricemia cannot be justified other than for prevention of acute uric acid nephropathy. In addition, routine screening for asymptomatic hyperuricemia is not recommended. If hyperuricemia is diagnosed, however, the cause should be determined. Causal factors should be corrected if the condition is secondary, and associated problems such as hypertension, hypercholesterolemia, diabetes mellitus, and obesity should be treated.

SYMPTOMATIC HYPERURICEMIA

See Chap. 365 for treatment of gout, including urate nephrosis.

Nephrolithiasis

Antihyperuricemic therapy is recommended for the individual who has both gouty arthritis and either uric acid– or calcium-containing stones, both of which may occur in association with hyperuricaciduria. Regardless of the nature of the calculi, fluid ingestion should be sufficient to produce a daily urine volume >2 L. Alkalinization of the urine with sodium bicarbonate or acetazolamide may be justified to increase the solubility of uric acid. Specific treatment of uric acid calculi requires reducing the urine uric acid concentration with a xanthine oxidase inhibitor, such as allopurinol or febuxostat. These agents decrease the serum urate concentration and the urinary excretion of uric acid in the first 24 h, with a maximal reduction within 2 weeks. Allopurinol can be given once a day because of the long half-life (18 h) of its active metabolite, oxypurinol. In the febuxostat trials, the generally recommended dose of allopurinol (300 mg/d) was effective at achieving a target serum urate concentration <6.0 mg/dL (357 μmol/L) in <50% of patients; this result suggested that higher doses should be considered. Allopurinol is effective in patients with renal insufficiency, but the dose should be reduced. Allopurinol is also useful in reducing the recurrence of calcium oxalate stones in patients with gout and in individuals with hyperuricemia or hyperuricaciduria who do not have gout. Febuxostat (40–80 mg/d) is also taken once daily, and doses do not need to be adjusted in the presence of mild to moderate renal dysfunction. Potassium citrate (30–80 mmol/d orally in divided doses) is an alternative therapy for patients with uric acid stones alone or mixed calcium/uric acid stones. A xanthine oxidase inhibitor is also indicated for the treatment of 2,8-dihydroxyadenine kidney stones.

Uric Acid Nephropathy

Uric acid nephropathy is often preventable, and immediate appropriate therapy has greatly reduced the mortality rate. Vigorous IV hydration and diuresis with furosemide dilute the uric acid in the tubules and promote urine flow to ≥100 mL/h. The administration of acetazolamide (240–500 mg every 6–8 h) and sodium bicarbonate (89 mmol/L) IV enhances urine alkalinity and thereby solubilizes more uric acid. It is important to ensure that the urine pH remains >7.0 and to watch for circulatory overload. In addition, antihyperuricemic therapy in the form of allopurinol in a single dose of 8 mg/kg is administered to reduce the amount of urate that reaches the kidney. If renal insufficiency persists, subsequent daily doses should be reduced to 100–200 mg because oxypurinol, the active metabolite of allopurinol, accumulates in renal failure. Despite these measures, hemodialysis may be required. Urate oxidase (rasburicase) can also be administered IV to prevent or to treat tumor lysis syndrome.

HYPOURICEMIA

Hypouricemia, defined as a serum urate concentration <120 μmol/L (<2.0 mg/dL), can result from decreased production of urate, increased excretion of uric acid, or a combination of both mechanisms. This condition occurs in <0.2% of the general population and <0.8% of hospitalized individuals. Hypouricemia causes no symptoms or pathology and therefore requires no therapy.

Most hypouricemia results from increased renal uric acid excretion. The finding of normal amounts of uric acid in a 24-h urine collection from an individual with hypouricemia is evidence for a renal cause. Medications with uricosuric properties (Table 410-1) include aspirin (at doses >2.0 g/d), losartan, fenofibrate,

x-ray contrast materials, and glyceryl guaiacolate. Total parenteral hyperalimentation can also cause hypouricemia, possibly a result of the high glycine content of the infusion formula. Other causes of increased urate clearance include conditions such as neoplastic disease, hepatic cirrhosis, diabetes mellitus, and inappropriate secretion of vasopressin; defects in renal tubular transport such as primary Fanconi syndrome and Fanconi syndromes caused by Wilson's disease, cystinosis, multiple myeloma, and heavy metal toxicity; and isolated congenital defects in the bidirectional transport of uric acid. Hypouricemia can be a familial disorder that is generally inherited in an autosomal recessive manner. Most cases are caused by a loss of function mutation in *SLC22A12*, the gene that encodes URAT-1, resulting in increased renal urate clearance. Individuals with normal *SLC22A12* most likely have a defect in other urate transporters. Although hypouricemia is usually asymptomatic, some patients suffer from urate nephrolithiasis or exercise-induced renal failure.

SELECTED INBORN ERRORS OF PURINE AND PYRIMIDINE METABOLISM

More than 30 defects in human purine and pyrimidine metabolic pathways have been identified thus far **(See also** Table 410-3, **Table 410-4**, Fig. 410-3, and **Fig. 410-4)**. Many are benign, but about half are associated with clinical manifestations, some causing major morbidity and mortality. Advances in genetics, along with high-performance liquid chromatography and tandem mass spectrometry, have facilitated diagnosis.

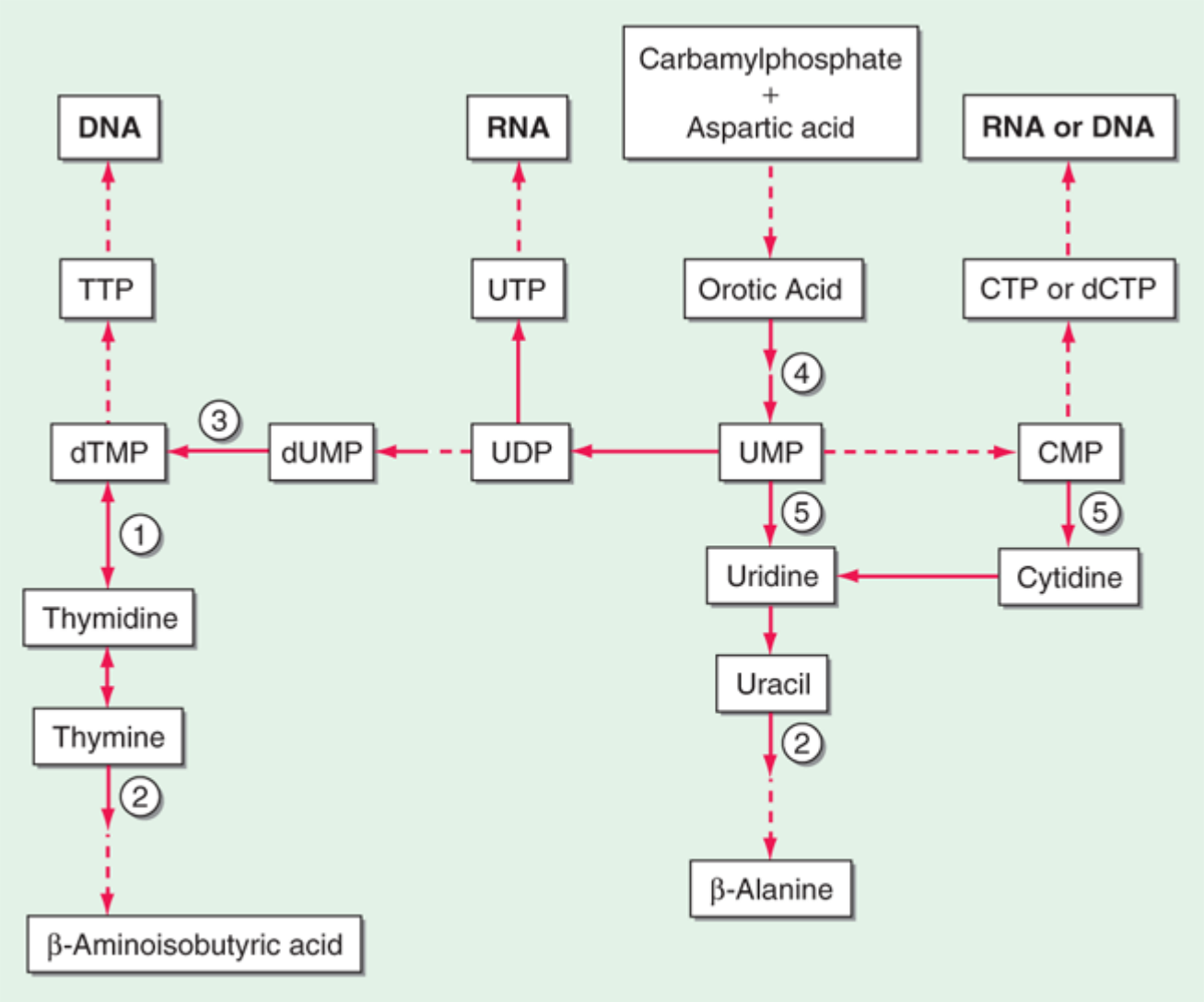
TABLE 410-4

Inborn Errors of Pyrimidine Metabolism

Enzyme	Activity	Inheritance	Clinical Features	Laboratory Features
Uridine-5′-monophosphate synthetase	Deficiency	Autosomal recessive	Orotic acid crystalluria; obstructive uropathy, hypochromic megaloblastic anemia	Orotic aciduria
Pyrimidine 5′-nucleotidase	Deficiency	Autosomal recessive	Hemolytic anemia	Basophilic stippling of erythrocytes; high levels of cytidine and uridine ribonucleotides
Pyrimidine 5′-nucleotidase	Superactivity	Uncertain	Developmental delay, seizures, ataxia, language deficit	Hypouricosuria
Thymidine phosphorylase	Deficiency	Autosomal recessive	Mitochondrial neurogastrointestinal encephalopathy	Hypouricosuria
Dihydropyrimidine dehydrogenase	Deficiency	Autosomal recessive	Seizures, motor and mental retardation	High levels of uracil, thymine, and 5-hydroxymethyluracil and low levels of dihydropyrimidines in urine
Dihydropyrimidinase	Deficiency	Uncertain	Seizures, mental retardation	Dihydropyrimidinuria
Ureidopropionase	Deficiency	Uncertain	Hypotonia, dystonia, developmental delay	High urinary excretion of *N*-carbamyl-β-alanine and *N*-carbamyl β-aminoisobutyric acid

FIGURE 410-4

Abbreviated scheme of pyrimidine metabolism. (1) Thymidine kinase, (2) dihydropyrimidine dehydrogenase, (3) thymidylate synthase, (4) UMP synthase, (5) 5′-nucleotidase. CMP, cytidine-5′-monophosphate; dTMP, deoxythymidine-5′-monophosphate; dUMP, deoxyuridine-5′-monophosphate; TTP, thymidine triphosphate; UDP, uridine-5′-diphosphate; UMP, uridine-5′-monophosphate; UTP, uridine triphosphate.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PURINE DISORDERS

HPRT Deficiency

The HPRT gene is located on the X chromosome. Affected males are hemizygous for the mutant gene; carrier females are asymptomatic. A complete deficiency of HPRT, the Lesch-Nyhan syndrome, is characterized by hyperuricemia, self-mutilative behavior, choreoathetosis, spasticity, and mental retardation. A partial deficiency of HPRT, the Kelley-Seegmiller syndrome, is associated with hyperuricemia but no central nervous system manifestations. In both disorders, the hyperuricemia results from urate overproduction and can

cause uric acid crystalluria, nephrolithiasis, obstructive uropathy, and gouty arthritis. Early diagnosis and appropriate therapy with allopurinol can prevent or eliminate all the problems attributable to hyperuricemia without affecting behavioral or neurologic abnormalities.

Increased PRPP Synthetase Activity

Like the HPRT deficiency states, PRPP synthetase overactivity is X-linked and results in gouty arthritis and uric acid nephrolithiasis. Neurologic hearing loss occurs in some families.

Adenine Phosphoribosyltransferase (APRT) Deficiency

APRT deficiency is inherited as an autosomal recessive trait. Affected individuals develop kidney stones composed of 2,8-dihydroxyadenine. Caucasians with the disorder have a complete deficiency (type I), whereas Japanese individuals have some measurable enzyme activity (type II). Expression of the defect is similar in the two populations, as is the frequency of the heterozygous state (0.4–1.1 per 100). Allopurinol treatment prevents stone formation.

Hereditary Xanthinuria

A deficiency of xanthine oxidase causes all purine in the urine to occur in the form of hypoxanthine and xanthine. About two-thirds of deficient individuals are asymptomatic. The remainder develop kidney stones composed of xanthine.

Myoadenylate Deaminase Deficiency

Primary (inherited) and secondary (acquired) forms of myoadenylate deaminase deficiency have been described. The primary form is inherited as an autosomal recessive trait. Clinically, some persons may have relatively mild myopathic symptoms with exercise or other triggers, but most individuals with this defect are asymptomatic. Therefore, another explanation for the myopathy should be sought in symptomatic patients with this deficiency. The acquired deficiency occurs in association with a wide variety of neuromuscular diseases, including muscular dystrophies, neuropathies, inflammatory myopathies, and collagen vascular diseases.

Adenylosuccinate Lyase Deficiency

Deficiency of this enzyme is due to an autosomal recessive trait and causes profound psychomotor retardation, seizures, and other movement disorders. All individuals with this deficiency are mentally retarded, and most are autistic.

Adenosine Deaminase Deficiency and Purine Nucleoside Phosphorylase Deficiency

See Chap. 344.

PYRIMIDINE DISORDERS

The pyrimidine cytidine is found in both DNA and RNA; it is a complementary base pair for guanine. Thymidine is found only in DNA, where it is paired with adenine. Uridine is found only in RNA and can pair with either adenine or guanine in RNA secondary structures. Pyrimidines can be synthesized by a de novo pathway (Fig. 410-4) or reused in a salvage pathway. Although >25 different enzymes are involved in pyrimidine metabolism, disorders of these pathways are rare. Seven disorders of pyrimidine metabolism have been discovered (Table 410-4), three of which are discussed below.

Orotic Aciduria

Hereditary orotic aciduria is caused by mutations in a bifunctional enzyme, uridine-5′-monophosphate (UMP) synthase, which converts orotic acid to UMP in the de novo synthesis pathway (Fig. 410-4). The disorder is characterized by hypochromic megaloblastic anemia that is unresponsive to vitamin B_{12} and folic acid, growth retardation, and neurologic abnormalities. Increased excretion of orotic acid causes crystalluria and obstructive uropathy. Replacement of uridine (100–200 mg/kg per day) corrects anemia, reduces orotic acid excretion, and improves the other sequelae of the disorder.

Pyrimidine 5′-nucleotidase Deficiency

Pyrimidine 5′-nucleotidase catalyzes the removal of the phosphate group from pyrimidine ribonucleoside monophosphates (cytidine-5′-monophosphate or UMP) (Fig. 410-4). An inherited deficiency of this enzyme causes hemolytic anemia with prominent basophilic stippling of erythrocytes. The accumulation of pyrimidines or cytidine diphosphate choline is thought to induce hemolysis. There is no specific treatment. Acquired pyrimidine 5′-nucleotidase deficiency has been reported in lead poisoning and in thalassemia.

Dihydropyrimidine Dehydrogenase Deficiency

Dihydropyrimidine dehydrogenase is the rate-limiting enzyme in the pathway of uracil and thymine degradation (Fig. 410-4). Deficiency of this enzyme causes excessive urinary excretion of uracil and thymine. In addition, this deficiency causes nonspecific cerebral dysfunction with convulsive disorders, motor retardation, and mental retardation. No specific treatment is available.

Medication Effect on Pyrimidine Metabolism

A variety of medications can influence pyrimidine metabolism. The anticancer agents fluorodeoxyuridine and 5-fluorouracil and the antimicrobial agent fluorocytosine cause cytotoxicity when converted to fluorodeoxyuridylate, a specific suicide inhibitor of thymidylate synthase. Fluorocytosine must be converted to 5-fluorouracil to be effective. This conversion is catalyzed by cytosine deaminase activity. Fluorocytosine's action is selective because cytosine deaminase is present in bacteria and fungi but not in human cells. Dihydropyrimidine dehydrogenase is involved in the degradation of 5-fluorouracil. Consequently, deficiency of this enzyme is associated with 5-fluorouracil neurotoxicity.

Leflunomide, which is used to treat rheumatoid arthritis, inhibits de novo pyrimidine synthesis by inhibiting dihydroorotate dehydrogenase, resulting in an antiproliferative effect on T cells. Allopurinol, which inhibits xanthine oxidase in the purine metabolic pathway, also inhibits the activity of orotidine-5′-phosphate decarboxylase, a step in UMP synthesis. Consequently, allopurinol use is associated with increased excretion of orotidine and orotic acid. There are no known clinical effects of this inhibition.

ACKNOWLEDGMENT

The authors are grateful to Robert L. Wortmann for contributions to this chapter in previous editions of the book.

FURTHER READING

Balasubramaniam S et al: Inborn errors of purine metabolism: Clinical update and therapies. J Inherit Metab Dis 37:669, 2014. [PubMed: 24972650]

Balasubramaniam S et al: Inborn errors of pyrimidine metabolism: Clinical update and therapy. J Inherit Metab Dis 37:687, 2014. [PubMed: 25030255]

Ben Salem C et al: Drug-induced hyperuricaemia and gout. Rheumatology (Oxford) 266:56:679, 2017.

Bhole V, Krishnan E: Gout and the heart. Rheum Dis Clin North Am 40:125, 2014. [PubMed: 24268013]

Burns CM, Wortman RL: Clinical features and treatment of gout, in *Kelley and Firestein's Textbook of Rheumatology*, 10th ed, GS Firestein et al (eds). Philadelphia, Elsevier, 2017, pp 1620–1644.

Hirano M, Peters GJ: Advances in purine and pyrimidine metabolism in health and diseases. Nucleosides Nucleotides Nucleic Acids 35:495, 2016. [PubMed: 27906626]

Terkeltaub R (ed): *Gout and Other Crystal Arthropathies*. Philadelphia, Elsevier Health Sciences, 2012.

Chapter 411: Lysosomal Storage Diseases

Robert J. Hopkin; Gregory A. Grabowski

INTRODUCTION

Lysosomes are heterogeneous subcellular organelles containing specific hydrolyses that allow selective processing or degradation of proteins, nucleic acids, carbohydrates, and lipids. There are more than 50 different lysosomal storage diseases (LSDs), classified according to the nature of the stored material **(Table 411-1)**. Several of the most prevalent disorders are reviewed here: Tay-Sachs disease, Fabry disease, Gaucher disease, Niemann-Pick disease, lysosomal acid lipase deficiency (LALD), the mucopolysaccharidoses, and Pompe disease. LSDs should be considered in the differential diagnosis of patients with neurologic, renal, or muscular degeneration and/or unexplained hepatomegaly, splenomegaly, cardiomyopathy, or skeletal dysplasias and deformations. Physical findings are disease specific, and enzyme assays or genetic testing can be used to make a definitive diagnosis. Although the nosology of LSDs segregates the variants into distinct phenotypes, these are heuristic; in the clinic, each disease exhibits—to varying degrees—a continuous spectrum of manifestations, from severe to attenuated variants.

TABLE 411-1

Selected Lysosomal Storage Diseases

					Clinical Features			
Disorder[a]	**Enzyme Deficiency [Specific Therapy]**	**Stored Material**	**Clinical Types (Onset)**	**Inheritance**	**Neurologic**	**Liver, Spleen Enlargement**	**Skeletal Dysplasia**	**Ophth**
Mucopolysaccharidoses CTMucopolysaccharidoses (MPS)								
MPS I H, Hurler (136)	α-L-Iduronidase [ET, HSCT]	Dermatan sulfate Heparan sulfate	Infantile Intermediate	AR	Cognitive degeneration	+++	++++	Corn cloud
MPS I H/S, Hurler/Scheie			Childhood/Adult		Cognitive degeneration			
MPS I S, Scheie					None			
MPS II, Hunter (136)	Iduronate sulfatase [ET]	Dermatan sulfate Heparan sulfate	Severe infantile Mild juvenile	X-linked	Cognitive degeneration, less in mild form	+++	++++	Retin dege no co cloud
MPS III A, Sanfilippo A (136)	Heparan-*N*-sulfatase	Heparan sulfate	Late infantile	AR	Severe Cognitive degeneration	+	+	None
MPS III B, Sanfilippo B (136)	*N*-Acetyl-α-glucosaminidase	Heparan sulfate	Late infantile	AR	Severe Cognitive degeneration	+	+	None
MPS III C, Sanfilippo C (136)	Acetyl-CoA: α-glucosaminide *N*-acetyltransferase	Heparan sulfate	Late infantile	AR	Severe Cognitive degeneration	+	+	None
MPS III D, Sanfilippo D (136)	*N*-Acetylglucosamine-6-sulfate sulfatase	Heparan sulfate	Late infantile	AR	Severe Cognitive degeneration	+	+	None
MPS IV A, Morquio A (136)	*N*-Acetylgalactosamine-6-sulfate sulfatase [ET—trials]	Keratan sulfate Chondroitin-6 sulfate	Childhood	AR	None	+	++++	Corn cloud
MPS IV B, Morquio (136)	β-Galactosidase		Childhood	AR	None	±	++++	
MPS VI, Maroteaux-Lamy (136)	Arylsulfatase B [ET, BMT]	Dermatan sulfate	Late infantile	AR	None	++	++++	Corn cloud
MPS VII (136)	β-Glucuronidase	Dermatan sulfate Heparan sulfate	Neonatal Infantile Adult	AR	Cognitive degeneration, absent in some adults	+++	+++	Corn cloud

					Clinical Features			
Disorder[a]	**Enzyme Deficiency [Specific Therapy]**	**Stored Material**	**Clinical Types (Onset)**	**Inheritance**	**Neurologic**	**Liver, Spleen Enlargement**	**Skeletal Dysplasia**	**Opht**
GM_2 Gangliosidoses								
Tay-Sachs disease (153)	β-Hexosaminidase A	GM_2 gangliosides	Infantile Juvenile	AR	Cognitive degeneration; seizures; later juvenile form	None	None	Cher in inf
Sandhoff disease (153)	β-Hexosaminidases A and B	GM_2 gangliosides	Infantile	AR	Cognitive degeneration; seizures	++	±	Cher
Neutral Glycosphingolipidoses								
Fabry disease (150)	α-Galactosidase A [ET]	Globotriaosylceramide	Childhood	X-linked	Painful acroparesthesias	None	None	Corn dystr vascu
Gaucher disease (146)	Acid β-glucosidase [ET, SRT]	Glucosylceramide, glycosylsphingosine	Type 1 Type 2 Type 3	AR	None ++++ ++	++++ +++ ++++	++++ + ++++	None Eye r Eye r
Niemann-Pick disease (144) A and B	Sphingomyelinase [ET—trials]	Sphingomyelin	Neuronopathic, type A Nonneuronopathic, type B	AR	Cognitive degeneration; seizures	++++	None Osteoporosis	Macu dege
Glycoproteinoses								
Fucosidosis (140)	α-Fucosidase	Glycopeptides; oligosaccharides	Infantile Juvenile	AR	Cognitive degeneration	++	++	None
α-Mannosidosis (140)	α-Mannosidase	Oligosaccharides	Infantile Milder variant	AR	Cognitive degeneration	+++	+++	Cata corn cloud
β-Mannosidosis (140)	β-Mannosidase	Oligosaccharides		AR	Seizures; Cognitive degeneration		++	None
Aspartylglucosaminuria (141)	Aspartylglucosaminidase	Aspartylglucosamine; glycopeptides	Young adult	AR	Cognitive degeneration	±	++	None
Sialidosis (140)	Neuraminidase	Sialyloligosaccharides	Type I, congenital Type II, infantile and juvenile	AR	Myoclonus; Cognitive degeneration	++, less in type I	++, less in type I	Cher
Mucolipidoses (ML)								

					Clinical Features			
Disorder[a]	**Enzyme Deficiency [Specific Therapy]**	**Stored Material**	**Clinical Types (Onset)**	**Inheritance**	**Neurologic**	**Liver, Spleen Enlargement**	**Skeletal Dysplasia**	**Opht**
ML-II, I-cell disease (138)	UDP-*N*-Acetylglucosamine-1-phosphotransferase	Glycoprotein; glycolipids	Infantile	AR	Cognitive degeneration	+	++++	Corn clou
ML-III, pseudo-Hurler polydystrophy (138)	UDP-*N*-Acetylglucosamine-1-phosphotransferase	Glycoprotein; glycolipids	Late infantile	AR	Mild Cognitive degeneration	None	+++	Corn clou retin hype astig
Leukodystrophies								
Krabbe disease (147)	Galactosylceramidase [BMT/HSCT]	Galactosylceramide Galactosylsphingosine	Infantile	AR	Cognitive degeneration	None	None	None
Metachromatic leukodystrophy (148)	Arylsulfatase A	Cerebroside sulfate	Infantile Juvenile Adult	AR	Cognitive degeneration; dementia; psychosis in adult	None	None	Opti
Multiple sulfatase deficiency (149)	Active site cysteine to C_{α}-formylglycine-converting enzyme	Sulfatides; mucopoly-saccharides	Late infantile	AR	Cognitive degeneration	+	++	Retir dege
Disorders of Neutral Lipids								
Infantile-onset LALD (142)	Acid lysosomal lipase [ET]	Cholesteryl esters; triglycerides	Infantile	AR	None	+++	None	None
Childhood/Adult-onset LALD (142)	Acid lysosomal lipase [ET]	Cholesteryl esters	Childhood	AR	None	Hepatomegaly	None	None
Farber disease (143)	Acid ceramidase	Ceramide	Infantile Juvenile	AR	Occasional Cognitive degeneration	±	None	Macu dege
Disorders of Glycogen								
Pompe disease (135)	Acid α-glucosidase [ET]	Glycogen	Infantile, late onset	AR	Neuromuscular	±	None	None
Late onset GAA deficiency (135)	Acid α-glucosidase [ET]	Glycogen	Variable: juvenile to adulthood	AR	Neuromuscular	None	None	None
Danon disease (154)	LAMP-2 (lysosomal associated membrane protein-2)	Glycogen	Variable: childhood to adulthood	X-linked (? Dominant)	Cardiomyopathy Neuromuscular Inconsistent Cognitive degeneration	None	None	None

[a]Numbers in parentheses refer to the chapters in CR Scriver et al: *The Online Metabolic and Molecular Bases of Inherited Disease,* New York, McGraw-Hill, *ommbid.mhmedical.com*, which provide comprehensive reviews.

Abbreviations: AR, autosomal recessive; BMT/HSCT, bone marrow or stem cell transplantation; ET, enzyme therapy; SRT, substrate reduction therapy.

PATHOGENESIS

Lysosomal biogenesis involves ongoing synthesis of lysosomal hydrolases, membrane constitutive proteins, and new membranes. Lysosomes originate from the fusion of trans-Golgi network vesicles with late endosomes. Progressive vesicular acidification accompanies the maturation of these vesicles; this gradient facilitates the pH-dependent dissociation of receptors and ligands and also activates lysosomal hydrolases. Lysosomes are components of the lysosome/autophagy/mitophagy system, which are disrupted in the LSDs.

Abnormalities at any biosynthetic step can impair enzyme activation and lead to a lysosomal storage disorder. After leader sequence clipping, remodeling of complex oligosaccharides (including the lysosomal targeting ligand mannose-6-phosphate as well as high-mannose oligosaccharide chains of many soluble lysosomal hydrolases) occurs during transit through the Golgi. Lysosomal integral or associated membrane proteins are sorted to the membrane or interior of the lysosome by several different peptide signals. Phosphorylation, sulfation, additional proteolytic processing, and macromolecular assembly of heteromers occur concurrently. Such posttranslational modifications are critical to enzyme function, and defects can result in multiple enzyme/protein deficiencies.

The final common pathway for LSDs is the accumulation of specific macromolecules within tissues and cells that normally have a high flux of these substrates. The majority of lysosomal enzyme deficiencies result from point mutations or genetic rearrangements at a locus that encodes a single lysosomal hydrolase. However, some mutations cause deficiencies of several different lysosomal hydrolases by alteration of the enzymes/proteins involved in targeting, active site modifications, or macromolecular association or trafficking. All LSDs are inherited as autosomal recessive disorders except for Hunter (mucopolysaccharidosis type II), Danon, and Fabry diseases, which are X-linked. Substrate accumulation leads to lysosomal distortion/dysfunction, which has significant pathologic consequences. In addition, abnormal amounts of metabolites may also have pharmacologic effects important to disease pathophysiology and propagation, particularly activation of the innate immune responses.

For many LSDs, the accumulated substrates are endogenously synthesized within particular tissue sites of pathology. Other diseases have greater exogenous substrate supplies. For example, substrates are delivered by low-density lipoprotein receptor–mediated uptake in Fabry and LALD or by phagocytosis in Gaucher disease type 1. The *threshold hypothesis* refers to a level of enzyme activity below which disease develops; small changes in enzyme activity near the threshold can lead to or modify disease. A critical element of this model is that enzymatic activity can be challenged by changes in substrate flux based on genetic background, cell turnover, recycling, or metabolic demands. Thus, a set level of residual enzyme may be adequate for substrate in some tissues or cells but not in others. In addition, several variants of each LSD exist at a clinical level. These disorders therefore represent a continuum of manifestations that are not easily dissociated into discrete entities. The molecular/genetic bases for such variations have not been elucidated in any detail.

SELECTED DISORDERS

TAY-SACHS DISEASE

About 1 in 30 Ashkenazi Jews is a carrier for Tay-Sachs disease (total hexosaminidase A [Hex A] deficiency), resulting from defective α-chains. The infantile form is a fatal neurodegenerative disease with macrocephaly, loss of motor skills, increased startle reaction, and a macular cherry red spot. The juvenile-onset form presents as ataxia and dementia, with death by age 10–15 years. The adult-onset disorder is characterized by clumsiness in childhood; progressive motor weakness in adolescence; and additional spinocerebellar and lower-motor-neuron signs and dysarthria in adulthood. Intelligence declines slowly, and psychiatric disorders are common. Screening for Tay-Sachs disease carriers is recommended in the Ashkenazi Jewish population. Sandhoff disease, due to a deficiency in both Hex A and Hex B resulting from defective β-chains, is phenotypically similar to Tay-Sachs disease, but also includes hepatosplenomegaly and bony dysplasias.

FABRY DISEASE

Fabry disease, an X-linked disorder, results from mutations in *GALA* that encodes α-galactosidase A. The estimated prevalence of hemizygous males ranges from 1 in 40,000 to 1 in 3500 in selected populations. Clinically, the disease manifests with angiokeratomas (telangiectatic skin lesions), hypohidrosis, corneal and lenticular opacities, acroparesthesia; and progressive small-vessel disease of the kidney, heart, and brain.

Angiokeratomas and acroparesthesias may appear in childhood. Angiokeratomas are punctate, dark red to blue-black, flat or slightly raised, and usually symmetric; they do not blanch with pressure. They are often small and can be easily overlooked. They usually are most dense between the umbilicus and the knees—the "bathing suit area"—but may occur anywhere, including the mucosal surfaces. Angiokeratomas also occur in several other very rare LSDs. Corneal and lenticular lesions, detectable on slit-lamp examination, may help in establishing a diagnosis of Fabry disease. Debilitating episodic burning pain of the hands, feet, and proximal extremities (acroparesthesia) can last from minutes to days and can be precipitated by changes in temperature, exercise, fatigue, or fever. Abdominal pain can resemble that from appendicitis or renal colic. Proteinuria, isosthenuria, and progressive renal dysfunction occur in the second to fourth decades; ~5% of male patients with idiopathic renal failure have *GALA* mutations. Hypertension, left ventricular hypertrophy, anginal chest pain, and congestive heart failure can occur in the third to fourth decades. About 1–3% of patients with idiopathic hypertrophic myocardiopathy have Fabry disease. Similarly, ~3–5% of male patients with idiopathic stroke at 35–50 years of age have *GALA* mutations. Leg lymphedema without hypoproteinemia and episodic diarrhea also occur. Death is due to renal failure or cardiovascular or cerebrovascular disease in untreated male patients. Variants with residual α-galactosidase A activity may have late-onset manifestations that are limited to the cardiovascular system and resemble hypertrophic cardiomyopathy. Variants with predominant cardiac, renal, or central nervous system (CNS)

manifestations are becoming better defined. Up to 70% of heterozygous females exhibit clinical manifestations. However, in females, heart disease is the most common life-threatening manifestation, followed in frequency by stroke and renal disease.

Gabapentin and carbamazepine diminish chronic and episodic acroparesthesia. Chronic hemodialysis or kidney transplantation can be lifesaving in patients with renal failure. Intravenous enzyme therapy clears stored lipids from a variety of cells, particularly those of the renal, cardiac, and skin vascular endothelium. Renal insufficiency appears to be irreversible. Early institution of enzyme therapy may prevent or slow the progression of life-threatening complications.

GAUCHER DISEASE

Gaucher disease, an autosomal recessive disorder, results from defective activity of acid β-glucosidase; ~600 *GBA1* mutations have been described in such patients. Disease variants are classified by the absence or presence and progression of neuronopathic involvement.

Gaucher disease type 1 is a nonneuronopathic disease (i.e., absence of early-onset or progressive CNS disease) presenting in childhood to adulthood as slowly to rapidly progressive visceral disease. About 55–60% of patients are diagnosed at <20 years of age in white populations and at even younger ages in other groups. This pattern of presentation is distinctly bimodal, with peaks at <10–15 years and at ~25 years. Younger patients tend to have greater degrees of hepatosplenomegaly and accompanying blood cytopenias. In contrast, the older patients have a greater tendency for chronic bone disease. Hepatosplenomegaly occurs in virtually all symptomatic patients and can be minor or massive. Accompanying anemia and thrombocytopenia are variable and are not directly related to liver or spleen volumes. Severe liver dysfunction is unusual. Splenic infarctions can resemble an acute abdomen. Pulmonary hypertension and alveolar Gaucher cell accumulation are uncommon but life-threatening and can occur at any age. *GBA1* mutations in the hetero- or homozygous state are a significantly increased life-time risk for developing Parkinson disease.

All patients with Gaucher disease have nonuniform infiltration of bone marrow by lipid-laden macrophages termed *Gaucher cells*. This phenomenon can lead to marrow packing with subsequent infarction, ischemia, necrosis, and cortical bone destruction. Bone marrow involvement spreads from proximal to distal in the limbs and can involve the axial skeleton extensively, causing vertebral collapse. In addition to bone marrow involvement, bone remodeling is defective, with loss of total bone calcium leading to osteopenia, osteonecrosis, avascular infarction, and vertebral compression fractures with spinal cord involvement. Aseptic necrosis of the femoral head is common, as is fracture of the femoral neck. The mechanism by which diseased bone marrow macrophages interact with osteoclasts and/or osteoblasts to cause bone disease is not well understood. Chronic, ill-defined bone pain can be debilitating and poorly correlated with radiographic findings. “Bone crises” are associated with localized excruciating pain and, on occasion, local erythema, fever, and leukocytosis. These crises represent acute infarctions of bone, as evidenced in nuclear scans by localized absent uptake of pyrophosphate agents. Decreased acid β-glucosidase activity (0–20% of normal) in nucleated cells establishes the diagnosis. The enzyme is not normally present in bodily fluids. The sensitivity of enzyme testing is poor for heterozygote detection; molecular testing by *GBA1* sequencing is preferred. The disease frequency varies from about 1 in 1000 among Ashkenazi Jews to <1 in 100,000 in other populations; ~1 in 12–15 Ashkenazi Jews carries a Gaucher disease allele. Four common mutations account for ~85% of the mutations in that population of affected patients: N370S (1226G), 84GG (a G insertion at cDNA position 84), L444P (1448C), and IVS-2 (an intron 2 splice junction mutation).

Genotype/phenotype studies indicate a significant, though not absolute, correlation between disease type and severity and the *GBA1* genotype. The most common mutation in the Ashkenazi Jewish population (N370S) shares a 100% association with nonneuronopathic or type 1 Gaucher disease. The N370S/N370S and N370S/other mutant allele genotypes are associated with later-onset/less severe disease and with earlier-onset/severe disease, respectively. As many as 50–60% of individuals with the N370S/N370S genotype are asymptomatic. Other alleles include L444P (very low activity), 84GG (null), or IVS-2 (null) and rare/private or uncharacterized alleles. The L444P/L444P patients frequently have life-threatening to very severe/early-onset disease, and many, though not all, develop CNS involvement in the first two decades of life.

Symptom-based treatment of blood cytopenias and joint replacement surgeries continue to have important roles in management. However, regular intravenous enzyme therapy has been the first-line treatment for significantly affected patients and is highly efficacious and safe in diminishing hepatosplenomegaly and improving hematologic values. An oral substrate reduction therapy (eliglustat tartrate), which inhibits glycolipid synthesis, is approved as a first-line therapy for adults. Bone disease is decreased and can be prevented, but irreversible damage cannot be reversed, by enzyme therapy. Adult patients may benefit from adjunctive treatment with bisphosphonates to improve bone density. Patients who cannot be treated with enzyme, either because it is not effective or because they have developed an allergy or other hypersensitivities to the enzyme, may receive *substrate reduction therapy* with either eliglustat tartrate or miglustat; the latter is approved as a second-line oral therapy.

Gaucher disease type 2 is a rare, severe, progressive CNS disease that leads to death by 2 years of age. Gaucher disease type 3 has highly variable manifestations in the CNS and viscera. It can present in early childhood with rapidly progressive, massive visceral disease and slowly progress to static CNS involvement; in adolescence with dementia; or in early adulthood with rapidly progressive, uncontrollable myoclonic seizures and mild visceral disease. Visceral disease in type 3 is nearly identical to that in type 1 but is generally more severe. Early CNS findings may be limited to defects in lateral gaze tracking, which may remain static for decades. Cognitive degeneration can be slowly progressive or static. Type 3 is much more frequent among individuals of non-Western World descent. Visceral—but not CNS—involvement responds to enzyme therapy.

NIEMANN-PICK DISEASE

Niemann-Pick diseases are autosomal recessive disorders that result from defects in acid sphingomyelinase. Types A and B are distinguished by the early age of onset and progressive CNS disease in type A. Type A typically has its onset in the first 6 months of life, with rapidly progressive CNS deterioration, spasticity, failure to thrive, and massive hepatosplenomegaly. Type B has a later, more variable onset and is characterized by a progression of hepatosplenomegaly, with eventual development of cirrhosis and hepatic replacement by foam cells. Affected patients develop progressive pulmonary disease with dyspnea, hypoxemia, and a reticular infiltrative pattern on chest x-ray. Foam cells are present in alveoli, lymphatic vessels, and pulmonary arteries. Progressive hepatic or lung disease can lead to death in adolescence or early adulthood.

The diagnosis is established by markedly decreased (1–10% of normal) sphingomyelinase activity in nucleated cells. There is no approved specific treatment for Niemann-Pick disease, but intravenous enzyme therapy clinical trials are in phase 3. The efficacy of hepatic or bone marrow transplantation has not been clearly established.

Niemann-Pick C diseases are progressive CNS diseases due to mutations in either *NPC1* or *NPC2*, lysosomal proteins involved in cholesterol transport out of the lysosome. They can present with liver or splenic disease, but their major manifestations are progressive CNS disease over one to two decades. Treatment with substrate inhibition agents (e.g., Miglustat) has shown some promise and substrate depletion with cyclodextrin is in clinical trials for NPC1 disease.

MUCOPOLYSACCHARIDOSES

Mucopolysaccharidosis type I (MPS I) is an autosomal recessive disorder caused by deficiency of α-L-iduronidase. The continuum of involvement traditionally has been divided into three categories: (1) Hurler disease (MPS I H) for severe deficiency with neurodegeneration, (2) Scheie disease (MPS I S) for later-onset disease without neurologic involvement and with relatively less severe disease in other organ systems, and (3) Hurler-Scheie syndrome (MPS I H/S) for patients intermediate between these extremes. MPS I H/S is characterized by severe somatic disease, usually without overt neurologic deterioration.

MPS I often presents in infancy or early childhood as chronic rhinitis, clouding of the corneas, and hepatosplenomegaly. As the disease progresses, nearly every organ system can be affected. In the more severe forms, cardiac and respiratory diseases become life threatening in childhood. Skeletal disease can be quite severe, resulting in very limited mobility.

There are two current treatments for the MPS I diseases. Hematopoietic stem cell transplantation (HSCT) is the standard treatment for patients presenting at <2 years of age who appear to have or are at risk for neurologic degeneration. HSCT results in stabilization of CNS disease and reverses hepatosplenomegaly. It also beneficially affects cardiac and respiratory disease. HSCT does not eliminate corneal disease or result in the resolution of progressive skeletal disease. Enzyme therapy effectively addresses hepatosplenomegaly and alleviates cardiac and respiratory disease. The enzyme does not effectively penetrate the CNS and does not directly affect CNS disease. Enzyme therapy and HSCT appear to have similar effects on visceral signs and symptoms. Enzyme therapy poses a lower risk of life-threatening complications and may therefore be advantageous for patients who have attenuated manifestations without CNS disease. A combination of enzyme therapy and HSCT has been used, with enzyme therapy initiated prior to transplantation in an attempt to reduce the disease burden. The experience with this approach is not well documented, but it appears to have advantages over HSCT alone.

Hunter disease (MPS II) is an X-linked disorder due to deficiency in iduronate sulfate sulfatase and has manifestations similar to those of MPS I, including neurologic degeneration. There is no corneal clouding or other eye disease. Like MPS I, MPS II is clinically variable, with CNS and non-CNS variants. HSCT has not been successful in treating CNS disease associated with MPS II. The FDA and the European Medicines Agency (EMA) have approved enzyme therapy for the visceral manifestations of MPS II.

MPS IV or Morquio syndrome is a rare autosomal recessive condition (1 to 200,000–300,000) and is different than the other mucopolysaccharidoses in presenting as a spondyloepiphyseal skeletal dysplasia. There are also important heart and respiratory complications. This disorder often presents in childhood, but the age of onset and rate of progression are quite variable. Two variants, type A and type B, are caused by deficiencies in *N*-acetyl-galactosamine-6-sulfatase (GALNS) and an acid β–galactosidase, respectively. A recombinant human GALNS enzyme replacement therapy (elosulfase alfa) is approved for the treatment of MPS IVA, making it is essential to confirm the specific diagnosis. Treatment has been shown to improve ambulatory mobility and decrease pain. There is no current specific treatment for MPS IVB.

Enzyme therapy for Maroteaux-Lamy disease (MPS VI), arylsulfatase B deficiency, has received U.S. Food and Drug Administration (FDA) approval as well as by similar agencies in other countries. This very rare autosomal recessive disorder is characterized by hepatosplenomegaly, bone disease, heart disease, and respiratory compromise. Short stature is also an important manifestation. Visceral signs and symptoms are similar to those in MPS I; however, MPS VI is not associated with neurologic degeneration.

MPS VII, Sly syndrome, is due to mutations in the *GUSB* gene, which codes for the β-glucuronidase enzyme. Severe deficiency in this enzyme may present with fetal hydrops which can lead to stillbirth or perinatal demise. Other patients with MPS VII may present later with short stature coarse facial features and hepatosplenomegaly. There is current research on enzyme replacement therapy for this disorder.

POMPE DISEASE

Acid maltase (acid α-glucosidase, GAA) deficiency, also called Pompe disease, is the only LSD leading to primary glycogen storage. The classic severe infantile form presents with hypotonia, myocardiopathy, and hepatosplenomegaly. This variant is rapidly progressive and generally results in death in the first year of life. However, as with other LSDs, there are early- and late-onset forms of this disorder. The late-onset variants may be as common as 1 in 40,000; patients typically present with a slowly progressive myopathy that may resemble limb-girdle muscular dystrophy. Respiratory insufficiency may be the presenting sign or may develop with advancing disease. In late stages of the disease, patients may require mechanical ventilation, report swallowing difficulties, and experience loss of bowel and bladder control. Myocardiopathy is not usually seen in late-onset variants of Pompe disease.

The FDA, EMA, and similar agencies have approved enzyme therapy for Pompe disease patients of all ages. This treatment clearly prolongs life in the infantile form, consistently resulting in improved cardiac function. Respiratory function is also improved in most treated infants. Some infants demonstrate marked improvement in motor functions, while others have minor changes in muscle tone or strength. Prevention of deterioration has been shown with GAA enzyme therapy in the late-onset forms. Early intervention with GAA enzyme therapy in such patients may limit or prevent deterioration, but very advanced disease will have significant irreversible components.

LYSOSOMAL ACID LIPASE DEFICIENCY

Wolman syndrome (now infantile-onset LALD) and cholesterol ester storage disease (now childhood/adult-onset LALD) are caused by deficiency of lysosomal acid lipase due to autosomal recessive mutations in the *LIPA*. The enzyme is responsible for hydrolysis of cholesterol esters and triglycerides delivered to the lysosome via the LDLR pathway. Accumulation of these in the tissues leads to progressive organ dysfunction including liver disease, intestinal malabsorption, heart dysfunction, and other manifestations. The most severe form presents in early infancy with failure to thrive, vomiting, and hepatosplenomegaly. The infantile-onset LALD patients die without specific treatment by age 1 year (median age of death 3.7 months). Childhood/adult-onset LALD can have a variable age of initial presentation with nonspecific signs, but often involves elevated liver enzymes, nonalcoholic fatty liver disease, cryptogenic cirrhosis, and varying severities of hepatosplenomegaly. Disease progresses throughout life and may result in early (adolescence) liver cirrhosis and (early adulthood) atherosclerosis or early death without treatment. Importantly, statins can decrease the hypercholesterolemia, but do not alter the basic progressive tissue, (e.g., liver) pathology. Enzyme replacement therapy for LALD has major effects in reversing disease manifestations and was approved for patients at all ages by the EMA, FDA, and several other country agencies in 2015 and 2016.

FURTHER READING

Aldenhoven M et al: Long-term outcome of Hurler syndrome patients after hematopoietic cell transplantation: An international multicenter study. Blood 125:2164, 2015. [PubMed: 25624320]

Balwani M et al: Recommendations for the use of eliglustat in the treatment of adults with Gaucher disease type 1 in the United States. Mol Genet Metab 117:95, 2016. [PubMed: 26387627]

Hopkin RJ et al: The management and treatment of children with Fabry disease: A United States-based perspective; Fabry Pediatric Expert Panel. Mol Genet Metab 117:104, 2016. [PubMed: 26546059]

Ortiz A et al: Time to treatment benefit for adult patients with Fabry disease receiving agalsidase β: Data from the Fabry Registry. J Med Genet. 53:495, 2016. [PubMed: 26993266]

Schoser B et al: Survival and long-term outcomes in late-onset Pompe disease following alglucosidase alfa treatment: A systematic review and meta-analysis. J Neurol. 264:621, 2017.

Chapter 412: Glycogen Storage Diseases and Other Inherited Disorders of Carbohydrate Metabolism

Priya S. Kishnani; Yuan-Tsong Chen

FIGURE 412-1

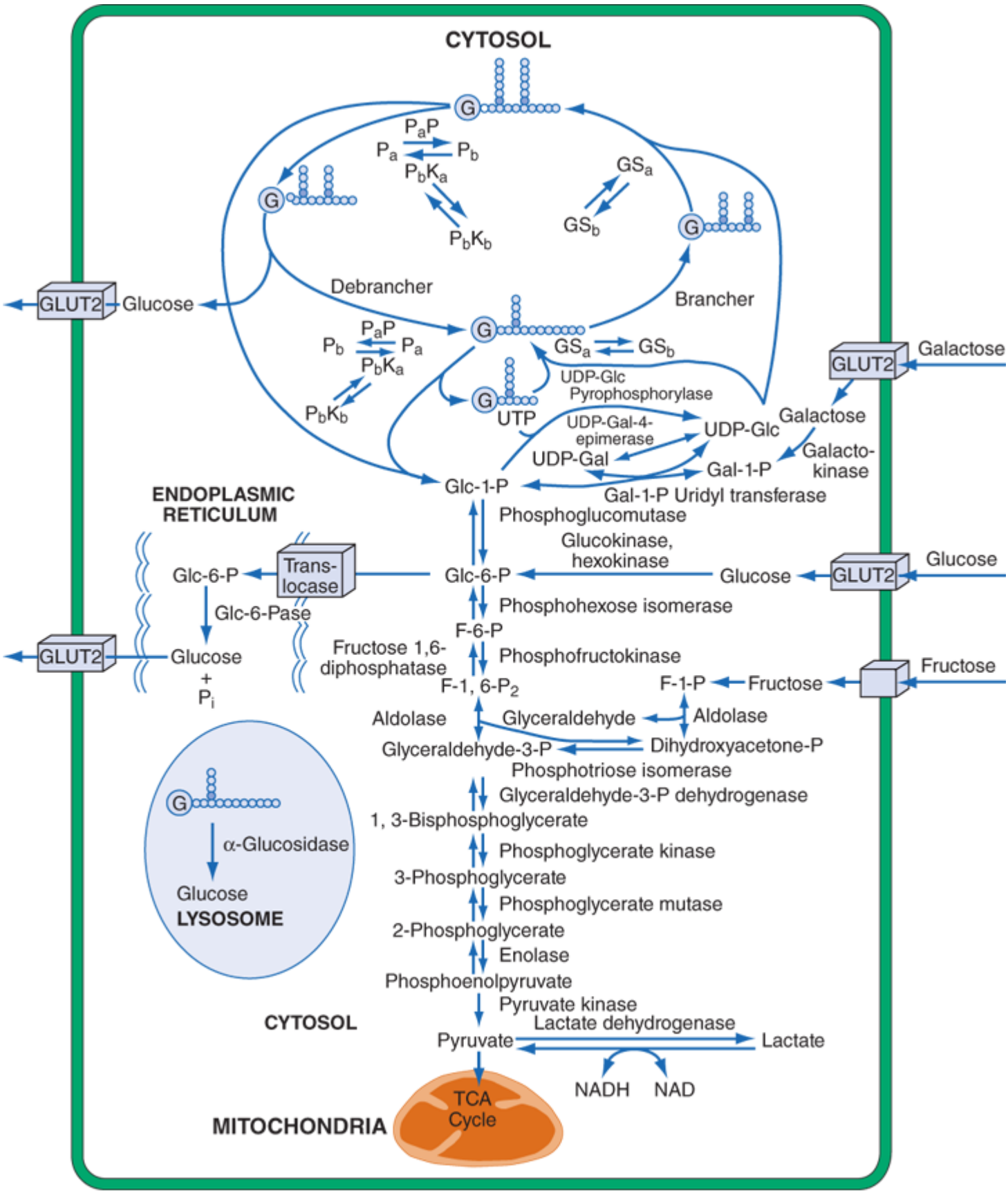
INTRODUCTION

Carbohydrate metabolism plays a vital role in cellular function by providing the energy required for most metabolic processes. The relevant biochemical pathways involved in the metabolism of these carbohydrates are shown in **Fig. 412-1**. Glucose is the principal substrate of energy metabolism in humans. Metabolism of glucose generates ATP through glycolysis and mitochondrial oxidative phosphorylation. The body obtains glucose through the ingestion of polysaccharides (primarily starch) and disaccharides (e.g., lactose, maltose, and sucrose). Lactose and fructose are two other monosaccharides that serve as sources of fuel for cellular metabolism; however, their role as fuel sources is much less significant than that of glucose. Lactose is derived from galactose + glucose, which is found in milk products, and is an important component of certain glycolipids, glycoproteins, and glycosaminoglycans. Fructose is found in fruits, vegetables, and honey. Sucrose (fructose + glucose) is another dietary source of fructose and is a commonly used sweetener.

FIGURE 412-1

Metabolic pathways related to glycogen storage diseases and galactose and fructose disorders. Nonstandard abbreviations are as follows: GS_a, active glycogen synthase; GS_b, inactive glycogen synthase; P_a, active phosphorylase; P_b, inactive phosphorylase; P_aP, phosphorylase α phosphatase; P_bK_a, active phosphorylase β kinase; P_bK_b, inactive phosphorylase β kinase; G, glycogenin, the primer protein for glycogen synthesis.

(Modified from AR Beaudet, in KJ Isselbacher et al [eds]: Harrison's Principles of Internal Medicine, 13th ed. New York, McGraw-Hill, 1994, p 1855.)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Glycogen, the storage form of glucose in animal cells, is composed of glucose residues joined in straight chains by α1-4 linkages and branched at intervals of 4–10 residues by α1-6 linkages. Glycogen forms a

treelike molecule and can have a molecular weight of many millions. Glycogen may aggregate to form structures recognizable by electron microscopy. With the exception of type 0 disease, defects in glycogen metabolism typically cause an accumulation of glycogen in the tissues—hence the designation *glycogen storage diseases* (GSDs). The structure of stored glycogen can be normal or abnormal in the various disorders. Defects in gluconeogenesis or glycolytic pathways including galactose and fructose metabolism usually do not result in glycogen accumulation.

Clinical manifestations of the various disorders of carbohydrate metabolism differ markedly. The symptoms range from minimally harmful to lethal. Unlike disorders of lipid metabolism, mucopolysaccharidoses, or other storage diseases, many carbohydrate disorders have been managed with dietary therapy, yet despite these strides, long-term complications result, and there is a need for definitive therapies. Genes responsible for inherited defects of carbohydrate metabolism have been cloned, and mutations have been identified. With the use of tools such as DNA sequencing panels, whole exome sequencing, and whole genome sequencing, new disorders of glycogen storage continue to be identified and the phenotype of known disorders continues to expand, as is seen in the case of GSD type II and III. Advances in our understanding of the molecular basis of these diseases are now being used to improve diagnosis and management. Some of these disorders are candidates for enzyme replacement therapy, substrate reduction therapy, gene therapy, and other genomic tools, such as SiRNA technology and CRISPR genome editing technology.

Historically, the GSDs were categorized numerically in the order in which the enzymatic defects were identified. They are also classified by the organs involved (liver, muscle, and/or heart) and clinical manifestations. The latter is the system followed in this chapter **(Table 412-1)**. The overall frequency of all forms of GSDs is ~1 in 20,000 live births. Most are inherited as autosomal recessive traits; however, phosphoglycerate kinase deficiency—one form of liver phosphorylase kinase (PhK) deficiency—and lysosomal-associated membrane protein 2 (LAMP2) deficiency are X-linked disorders. The most common childhood disorders are glucose-6-phosphatase deficiency (type I), lysosomal acid α-glucosidase deficiency (type II), debrancher deficiency (type III), and liver PhK deficiency (type IX). The most common adult disorder is myophosphorylase deficiency (type V, or McArdle disease).

TABLE 412-1

Features of Glycogen Storage Diseases and Galactose and Fructose Disorders

Type/Common Name	Basic Defect	Clinical Features	Comments
Liver Glycogenoses			
Disorders with Hepatomegaly and Hypoglycemia			
Ia/von Gierke	Glucose-6-phosphatase	Growth retardation, enlarged liver and kidney, hypoglycemia, elevated blood lactate, cholesterol, triglycerides, and uric acid	Common, severe hypoglycemia. Complications in adulthood include hepatic adenomas, hepatic carcinoma, osteoporosis, pulmonary hypertension and renal failure.
Ib	Glucose-6-phosphate translocase	As for Ia, with additional findings of neutropenia and neutrophil dysfunction, increased risk for mucosal ulceration, and periodontal disease, inflammatory bowel disease, hypothyroidism	~10% of type I
IIIa/Cori or Forbes	Liver and muscle debranching enzyme	*Childhood:* Hepatomegaly, growth retardation, muscle weakness, hypoglycemia, hyperlipidemia, elevated liver aminotransferases	Common, intermediate severity of hypoglycemia, yet severe cases are seen. Hepatic adenomas, liver cirrhosis, and hepatic carcinoma can occur.

Type/Common Name	Basic Defect	Clinical Features	Comments
		Adulthood: Proximal and distal muscle atrophy and weakness; peripheral neuropathy with preferential median nerve involvement; variable cardiomyopathy, liver cirrhosis, progressive liver failure, risk for HCC in some	Muscle weakness can progress to need for ambulatory aids such as wheel chair.
IIIb	Liver debranching enzyme (normal muscle debrancher activity)	Liver symptoms same as in type IIIa; no muscle symptoms	~15% of type III
IV/Andersen	Branching enzyme	Failure to thrive, hypotonia, hepatomegaly, splenomegaly, progressive liver cirrhosis and failure (death usually before fifth year); a small subset do not have liver progression. *Adult form:* Isolated myopathy, neurogenic bladder, peripheral neuropathy, cognitive impairment.	One of the rarer glycogenoses. Other neuromuscular variants exist.
VI/Hers	Liver phosphorylase	Hepatomegaly, variable hypoglycemia, hyperlipidemia, and ketosis	Often underdiagnosed, severe cases being recognized
IX/phosphorylase kinase deficiency		As for VI, progressive liver failure is seen in some patients.	Common, X-linked, typically less severe than autosomal forms; clinical variability within and between subtypes; severe cases being recognized across different subtypes

Type/Common Name	Basic Defect	Clinical Features	Comments
IXa (*PHKA2*)	Liver PhK	Hypoglycemia, hyperketosis hepatomegaly, chronic liver disease, hyperlipidemia, elevated liver enzymes, growth retardation.	X-linked
IXb(*PHKB*)	Liver and muscle PhK	Hepatomegaly, growth retardation	Autosomal recessive
IXc (*PHKG2*)	Liver PhK	more severe than IXa; marked hepatomegaly, recurrent hypoglycemia, liver cirrhosis.	Autosomal recessive
IXd (*PHKA2*)	Muscle PhK	Exercise intolerance, cramps, myalgia, myoglobinuria; no hepatomegaly.	X-linked or Autosomal recessive
0/liver glycogen synthase deficiency	Glycogen synthase	Fasting hypoglycemia and ketosis, elevated lactic acid, alanine levels and hyperglycemia after glucose load, no hepatomegaly	Decreased liver glycogen stores
XI/Fanconi-Bickel	Glucose transporter 2	Failure to thrive, short stature, hypophosphatemic rickets, metabolic acidosis, hepatomegaly, proximal renal tubular dysfunction, impaired glucose and galactose utilization	Rare, consanguinity in 70%
Muscle Glycogenoses			
Disorders with Muscle-Energy Impairment			
V/McArdle	Muscle phosphorylase	Exercise intolerance, muscle cramps, myoglobinuria on strenuous exercise, increased CK, “second-wind” phenomenon	Common, male predominance

Type/Common Name	Basic Defect	Clinical Features	Comments
VII/Tarui	Phosphofructokinase—M subunit	As for type V, with additional findings of compensated hemolysis, myalgia	Prevalent in Ashkenazi Jews and Japanese
Phosphoglycerate kinase deficiency	Phosphoglycerate kinase	As for type V, with additional findings of hemolytic anemia and CNS dysfunction	Rare, X-linked
Phosphoglycerate mutase deficiency	Phosphoglycerate mutase—M subunit	As for type V	Rare, most patients African American
Lactate dehydrogenase deficiency	Lactic acid dehydrogenase—M subunit	As for type V, with additional findings of erythematous skin eruption and uterine stiffness resulting in childbirth difficulty in females	Rare
Fructose 1,6-bisphosphate aldolase A deficiency	Fructose 1,6-bisphosphate aldolase A	As for type V, with additional finding of hemolytic anemia, splenomegaly, rhabdomyolysis, jaundice	Rare
Pyruvate kinase deficiency	Pyruvate kinase—muscle isozyme	Muscle cramps and/or fixed muscle weakness	Rare
Muscle phosphorylase kinase deficiency	Muscle-specific phosphorylase kinase	As for type V. Some patients may have muscle weakness and atrophy.	Rare, autosomal recessive
β-Enolase deficiency	Muscle β-enolase	Exercise intolerance	Rare
Disorders with Progressive Skeletal Muscle Myopathy and/or Cardiomyopathy			

Type/Common Name	Basic Defect	Clinical Features	Comments
II/Pompe	Lysosomal acid α-glucosidase	*Infantile:* Hypotonia, muscle weakness, cardiac enlargement and failure, fatal early. *Late onset (juvenile and adult):* Progressive skeletal muscle weakness and atrophy, proximal muscles and respiratory muscles seriously affected.	Common, undetectable or very low level of enzyme activity in infantile form; variable residual enzyme activity in late-onset form
PRKAG2 deficiency	AMP-activated gamma 2 protein kinase	Severe cardiomyopathy and early heart failure (9–55 years). Congenital fetal form is rapidly fatal with hypertrophic cardiomyopathy and Wolff-Parkinson-White syndrome. Other involvement includes myalgia, myopathy, and seizures.	Autosomal dominant
Danon disease	Lysosomal-associated membrane protein 2 (LAMP2)	Severe cardiomyopathy and heart failure (8–15 years)	Very rare, X-linked
Late-onset polyglucosan body myopathy	Glycogenin-1	Adult-onset proximal muscle weakness, nervous system involvement uncommon	Autosomal recessive, rare
Galactose Disorders			
Galactosemia with uridyltransferase deficiency	Galactose 1-phosphate uridyltransferase	Vomiting, hepatomegaly, jaundice, cataracts, amino aciduria, failure to thrive	Long-term complications exist despite early diagnosis and treatment.
Galactokinase deficiency	Galactokinase	Cataracts	Benign

Type/Common Name	Basic Defect	Clinical Features	Comments
Uridine diphosphate galactose 4-epimerase deficiency	Uridine diphosphate galactose 4-epimerase	Similar to transferase deficiency with additional findings of hypotonia and nerve deafness	Benign variant exists.
Fructose Disorders			
Essential fructosuria	Fructokinase	Asymptomatic, positive urine reducing substance	Benign, autosomal recessive
Hereditary fructose intolerance	Fructose 1,6-bisphosphate aldolase B	Vomiting, lethargy, failure to thrive, hepatic failure, aversion to sweets, severity of symptoms depending on age/quantity of sugar ingested	Prognosis good with early diagnosis and fructose restriction, autosomal recessive
Fructose 1,6-diphosphatase deficiency	Fructose 1,6-diphosphatase	Episodic hypoglycemia, hyperlactic acidemia, and ketoacidosis usually following illness, hepatomegaly	Avoid fasting, good prognosis.

Abbreviations: CK, creatine kinase; CNS, central nervous system; M, muscle; PhK, phosphorylase kinase.

SELECTED LIVER GLYCOGENOSES

DISORDERS WITH HEPATOMEGALY AND HYPOGLYCEMIA

Type I GSD (Glucose-6-Phosphatase or Translocase Deficiency, Von Gierke Disease)

Type I GSD is an autosomal recessive disorder caused by glucose-6-phosphatase deficiency in liver, kidney, and intestinal mucosa. There are two subtypes of GSD I: type Ia, in which the glucose-6-phosphatase enzyme is defective, and type Ib, in which the translocase that transports glucose-6-phosphate across the microsomal membrane is defective. The defects in both subtypes lead to inadequate conversion of glucose-6-phosphate to glucose in the liver and thus make affected individuals susceptible to fasting hypoglycemia.

CLINICAL AND LABORATORY FINDINGS

Persons with type I GSD may develop hypoglycemia and lactic acidosis during the neonatal period; however, more commonly, they exhibit hepatomegaly at 3–4 months of age. Hypoglycemia, hypoglycemic seizures, and lactic acidosis can develop after a short fast. These children usually have doll-like facies with fat cheeks, relatively thin extremities, short stature, and a protuberant abdomen that is due to massive hepatomegaly. The kidneys are enlarged, but the spleen and heart are of normal size. The hepatocytes are distended by glycogen and fat, with large and prominent lipid vacuoles. Despite hepatomegaly, liver enzyme levels are usually normal or near normal. Easy bruising and epistaxis are associated with prolonged bleeding time as a result of impaired platelet aggregation/adhesion. Hyperuricemia is present. Hyperlipidemia includes elevation of triglycerides, low-density lipoproteins, and phospholipids. Type Ib patients have additional findings of neutropenia and impaired neutrophil function, which result in recurrent bacterial infections and oral and intestinal mucosal ulceration. GSD I patients may experience intermittent diarrhea, which can worsen with age. In GSD Ib, diarrhea is largely due to loss of mucosal barrier function caused by inflammation.

LONG-TERM COMPLICATIONS

Gout usually becomes symptomatic at puberty as a result of long-term hyperuricemia. Puberty is often delayed. Some female patients have ultrasound findings consistent with polycystic ovaries; however, the other clinical features of polycystic ovary syndrome, such as acne and hirsutism, are not seen. Several reports of successful pregnancy in women with GSD I suggest that fertility is not affected. Increased bleeding during menstrual cycles, including life-threatening menorrhagia, has been reported. Secondary to lipid abnormalities, there is an increased risk of pancreatitis. Patients with GSD I may be at increased risk for cardiovascular disease. In adult patients, frequent fractures can occur and radiographic evidence of osteopenia/osteoporosis can be found; in prepubertal patients, radial bone mineral content is significantly reduced. Pulmonary hypertension—although rare—has been reported. By the second or third decade of life, many patients with type I GSD develop hepatic adenomas that can hemorrhage and, in some cases, become malignant. Renal disease is a serious late complication. Almost all patients aged >20 years have proteinuria, and many have hypertension, kidney stones, nephrocalcinosis, and altered creatinine clearance. In some patients, renal function deteriorates and progresses to complete failure, requiring dialysis or transplantation.

DIAGNOSIS

Clinical presentation and abnormal plasma lactate and lipid values suggest that a patient may have GSD I, and gene-based mutation analysis provides a noninvasive means of reaching a definitive diagnosis for most patients with types Ia and Ib disease. Before the glucose-6-phosphatase and glucose-6-phosphate translocase genes were cloned, a definitive diagnosis required a liver biopsy to demonstrate a deficiency.

Type III GSD (Debrancher Deficiency, Limit Dextrinosis)

Type III GSD is an autosomal recessive disorder caused by a deficiency of glycogen debranching enzymes. Debranching and phosphorylase enzyme are responsible for the complete degradation of glycogen into glucose. When debranching enzyme is defective, glycogen breakdown is incomplete, resulting in abnormal glycogen accumulation with short outer chains, resembling limit dextrin.

CLINICAL AND LABORATORY FINDINGS

Patients with GSD III present with hepatomegaly, hypoglycemia, short stature, variable skeletal myopathy, and cardiomyopathy. GSD type IIIa involves both liver and muscle. However, ~15% of patients have only liver involvement and is classified as *type IIIb*. Hypoglycemia and hyperlipidemia occur in children. In type III disease (as opposed to type I disease), fasting ketosis can be prominent, aminotransferase levels are elevated, and blood lactate and uric acid concentrations are usually normal. Serum creatine kinase (CK) levels can sometimes be used to identify patients with muscle involvement, but normal levels do not rule out muscle enzyme deficiency. In most patients with type III disease, hepatomegaly improves with age; however, liver fibrosis, cirrhosis progressing to liver failure, and hepatocellular carcinoma, are noted in many in late adulthood. Hepatic adenomas may occur, although less commonly than in GSD I. Left ventricular hypertrophy, significant scarring of the myocardium, and life-threatening arrhythmias have been reported. Patients with type IIIa disease may experience muscle weakness in childhood that can become severe after the third or fourth decade of life. The pattern of muscle weakness is variable and both proximal and distal muscle weakness are seen. Peripheral neuropathy may become discernible later in life with preferential median nerve involvement. Polycystic ovaries are common in GSD III, and some patients develop features of polycystic ovarian syndrome, such as hirsutism and irregular menstrual cycles. Reports of successful pregnancy in women with GSD III suggest that fertility is normal.

DIAGNOSIS

Deficient debranching enzyme activity can be demonstrated in liver, skeletal muscle, and heart in type IIIa GSD. In type IIIb, debranching enzyme deficiency is seen in the liver but not in muscle. The liver has distended hepatocytes due to glycogen buildup; areas of periportal fibrosis are also noted very early in the disease course. With the availability of molecular genetic testing, reliance on invasive tests such as liver and muscle biopsies is declining. DNA-based analyses now provide a noninvasive way of subtyping these disorders in most patients.

Type IX GSD (Liver Phosphorylase Kinase Deficiency)

Defects of PhK cause a heterogeneous group of glycogenoses. The PhK enzyme complex consists of four subunits (α, β, γ, and δ). Each subunit is encoded by different genes (X chromosome as well as autosomes) that are differentially expressed in various tissues. PhK deficiency can be divided into several subtypes on the basis of the gene/subunit involved, the tissues primarily affected, and the mode of inheritance. The most common subtype is GSD IXa, an X-linked liver PhK deficiency caused by mutations in the *PHKA2* gene, which is also one of the most common liver glycogenoses. PhK activity may also be deficient in erythrocytes and leukocytes but is normal in muscle. Typically, a child between the ages of 1 year and 5 years presents with growth retardation and hepatomegaly. Children tend eventually to exhibit normal growth patterns initiated by a delayed growth spurt during puberty. Liver fibrosis has been identified in some patients, including children. Levels of cholesterol, triglycerides, and liver enzymes are mildly elevated. Fasting ketosis is a feature of the disease. Lactic and uric acid levels are usually normal. Hypoglycemia may be mild in some but severe and recurrent in others. Phenotypic variability is being increasingly recognized, with significant disease involvement in some cases of the X-linked form. Liver histology shows distention of hepatocytes due to excess glycogen accumulation. There is a broad clinical spectrum of presentations identified with GSD IX. Hepatomegaly and abnormal blood chemistries gradually return to normal with age. Many adults reach a

normal final height and are practically asymptomatic, despite persistent PhK deficiency, yet liver involvement can progress to cirrhosis, fibrosis, and liver failure. Some patients have significant ketosis. It is recommended that patients be monitored for hepatic complications with regular CT or MRI scans. Though previously thought to be a mild disease, the understanding is evolving with more severe cases coming to light, even in the X-linked form. Further research is needed to completely understand the natural history and long-term complications of GSD IX.

Treatment is symptom-based. A diet rich in complex carbohydrates and proteins and in small, frequent feedings are effective in preventing hypoglycemia. Blood ketones and glucose should be evaluated during times of stress. Liver transplantation may be considered in those with severe hepatic involvement.

Other subtypes of type IX GSD include GSD IXb, an autosomal recessive form of liver and muscle PhK deficiency caused by *PHKB* mutations. GSD IXc, an autosomal recessive form of liver PhK deficiency that often develops into liver cirrhosis, is due to *PHKG2* mutations. GSD IXd, a muscle-specific PhK deficiency that causes cramps and myoglobinuria with exercise, is caused by *PHKA1* mutations. The previous reports of cardiac-specific PhK deficiency is now considered to be a secondary phenomenon, as these patients have mutations in the *PRKAG2* gene. Patients with cardiac PRKAG2 syndrome often present with cardiomyopathy during infancy. The condition is lethal because of massive glycogen deposition in the myocardium. Details about this condition are described under the section about PRKAG2 deficiency.

Type IV GSD (Branching enzyme deficiency, Amylopectinosis, Polyglucosan disease or Andersen disease) is caused by deficiency of branching enzyme activity leading to accumulation of an abnormal glycogen with poor solubility. The disease is clinically heterogeneous. Individuals typically present in the first 18 months of life with failure to thrive, hepatosplenomegaly, and progressive liver cirrhosis leading to death in early childhood. Some patients may develop hepatic adenomas and hepatocellular carcinoma. GSD IV has extrahepatic manifestations involving the central and peripheral nervous system as well as cardiac and skeletal muscles. The adult form is known as adult polyglucosan body disease (APBD) and may present as an isolated myopathy or with systemic involvement of the central and peripheral nervous system characterized by neurogenic bladder, peripheral neuropathy, leukodystrophy, and mild cognitive impairment. Definitive diagnosis requires demonstration of branching enzyme deficiency in liver, muscle, cultured skin fibroblasts or leukocytes, or genetic testing of the *GBE1* gene. It is likely that life expectancy is shortened in APBD patients though it is yet to be confirmed by long-term natural history studies. Good supportive care is crucial to improve clinical outcomes.

Treatment for the adult form of GSD IV includes symptomatic support for gait abnormalities, bladder dysfunction, as well as periodic monitoring to uncover any new neurological deficits. Liver transplantation may be performed for progressive hepatic failure. However, caution should be exercised in selecting patients for liver transplant as a nonprogressive hepatic form of the disease exists in some whereas in others, cardiac and nervous system involvement may occur after transplantation.

Other Liver Glycogenoses with Hepatomegaly and Hypoglycemia

These disorders include hepatic phosphorylase deficiency (Hers disease, type VI) and hepatic glycogenosis with renal Fanconi syndrome (type XI). Patients with GSD type VI can have growth retardation, hyperlipidemia, and hyperketosis in addition to hepatomegaly and hypoglycemia. Some patients have a less severe clinical course. GSD XI is caused by defects in the facilitative glucose transporter 2 (GLUT-2), which transports glucose and galactose in and out of hepatocytes, pancreatic cells, and the basolateral membranes of intestinal and renal epithelial cells. The disease is characterized by proximal renal tubular dysfunction, impaired glucose and galactose utilization, and accumulation of glycogen in liver and kidney.

SELECTED MUSCLE GLYCOGENOSES

DISORDERS WITH MUSCLE-ENERGY IMPAIRMENT

Type V GSD (Muscle Phosphorylase Deficiency, McArdle Disease)

Type V GSD is an autosomal recessive disorder caused by deficiency of muscle phosphorylase. McArdle disease is a prototypical muscle-energy disorder as the enzyme deficiency limits ATP generation by glycogenolysis and results in glycogen accumulation.

CLINICAL AND LABORATORY FINDINGS

There can be a broad, heterogeneous spectrum of clinical presentations with the neonatal form, which is rapidly fatal at one extreme, and the classical form with myalgia, cramps, and dark-colored urine at the other. Symptoms can be precipitated by: (1) brief, high intensity activity, such as sprinting or carrying heavy loads; and/or (2) less intense but sustained activity, such as climbing stairs or walking uphill. Most patients can engage in moderate exercise, such as walking on level ground, for long periods. Patients often exhibit the “second-wind” phenomenon, in which, after a short break from the initiation of strenuous physical effort, they are able to continue the activity without pain. Although most patients experience episodic muscle pain and cramping as a result of exercise, 35% report permanent pain that seriously affects sleep and other activities. Burgundy-colored urine is reported after exercise; resulting from myoglobinuria secondary to rhabdomyolysis. Renal failure can result from intense myoglobinuria after vigorous exercise. Symptom onset as late as the eighth decade has been reported.

Although cardiac involvement is not usually associated with muscle phosphorylase deficiency, hypertrophic cardiomyopathy has been observed in an adult patient with GSD V. In rare cases, electromyographic findings may suggest inflammatory myopathy, a diagnosis that may be confused with polymyositis. These patients may be at risk for statin-induced myopathy and rhabdomyolysis.

At rest, the serum CK level is usually elevated; after exercise, the CK level increases even more. Exercise leads to an increase in levels of blood ammonia, inosine, hypoxanthine, and uric acid; these abnormalities reflect residues of accelerated muscle purine nucleotide recycling as a result of insufficient ATP production. NADH is underproduced during physical exertion.

DIAGNOSIS

Lack of an increase in blood lactate and exaggerated blood ammonia elevations after an ischemic exercise test are indicative of a muscle glycogenosis and suggest a defect in the conversion of glycogen or glucose to lactate. This abnormal exercise response, however, can also occur with other defects in glycogenolysis or glycolysis, such as deficiencies of muscle phosphofructokinase or debranching enzyme (when the test is done after fasting). A noninvasive, nonischemic forearm exercise test has been developed. Although this test has high sensitivity, is easy to perform and is cost-effective, the abnormal exercise response does not exclude other muscle glycogenosis. The cycle test detects the hallmark heart rate observed during the second-wind phenomenon. A diagnostic confirmation is established by enzymatic assay in muscle tissue or by mutation analysis of the myophosphorylase gene.

DISORDERS WITH PROGRESSIVE SKELETAL MUSCLE MYOPATHY AND/OR CARDIOMYOPATHY

Pompe Disease, Type II GSD (Acid α-1,4 Glucosidase Deficiency)

Pompe disease is an autosomal recessive disorder caused by a deficiency of lysosomal acid α-1,4 glucosidase, an enzyme responsible for the degradation of glycogen in the lysosomes. This disease is characterized by the accumulation of glycogen in the lysosomes as opposed to accumulation in cytoplasm (as in the other glycogenoses).

CLINICAL AND LABORATORY FINDINGS

The disorder encompasses a range of phenotypes. Each includes myopathy but differs in the age of onset, extent of organ involvement, and clinical severity. The most severe is the infantile form, with cardiomegaly, hypotonia, and death before 2 years of age. Infants often present with cardiomyopathy at birth, and develop a generalized muscle weakness with feeding difficulties, macroglossia, hepatomegaly, and congestive heart failure due to the rapidly progressive hypertrophic cardiomyopathy.

The late-onset form (juvenile/late-childhood or adult form, LOPD) is characterized primarily by skeletal muscle manifestations and respiratory muscle involvement, and a more slowly progressive course. The juvenile form typically presents as delayed motor milestones (if age of onset is early enough) and difficulty in walking. With disease progression, patients often develop swallowing difficulties, proximal muscle weakness, and respiratory muscle involvement. Death may occur before the end of the second decade.

Adults typically present between the second and seventh decades with slowly progressive myopathy without overt cardiac involvement. The clinical picture is dominated by slowly progressive, predominantly proximal limb girdle muscle weakness. The pelvic girdle, paraspinal muscles, and diaphragm are most seriously affected. Respiratory symptoms include somnolence, morning headache, orthopnea, and exertional dyspnea. Respiratory failure causes significant morbidity and mortality in the late-onset form. In rare instances, patients present with respiratory insufficiency as the initial symptom. Basilar artery aneurysms and dilation of the ascending aorta have been observed in patients with Pompe disease. Ptosis, lingual weakness, gastrointestinal dysmotility, and incontinence due to poor sphincter tone are now being recognized as part of the clinical spectrum. Small-fiber neuropathy, which presents with painful paresthesia

or pins-and-needles sensations, is also seen in some patients with LOPD. Individuals with advanced disease often require some form of ventilatory support and are dependent on a walking aid or wheelchair.

Laboratory findings include elevated levels of serum CK, aspartate aminotransferase, alanine aminotransferase, and lactate dehydrogenase. Levels of urine glucose tetrasaccharide (Glc_4), a breakdown product of glycogen, are elevated, especially on the severe end of the disease spectrum and can be used as a biomarker to monitor disease progression and treatment responsiveness. In infants, chest x-ray shows massive cardiomegaly, and electrocardiographic findings include a high-voltage QRS complex and a shortened PR interval. Muscle biopsy shows vacuoles that stain positive for glycogen; the muscle acid phosphatase level is increased, presumably from a compensatory increase of lysosomal enzymes. Electromyography reveals myopathic features, with irritability of muscle fibers and pseudomyotonic discharges, which appears early in the paraspinal muscles. Serum CK is not always elevated in adults and, depending on the muscle biopsied or tested, muscle histology or electromyography may not be abnormal.

DIAGNOSIS

The confirmatory step for a diagnosis of Pompe disease is enzyme assay demonstrating deficient acid α-glucosidase or a gene sequence with two pathogenic mutations in the *GAA* gene. Enzyme activity can be measured in muscle, cultured skin fibroblasts, or blood. The latter is increasingly being used and is very reliable when performed in laboratories with experience. Prenatal diagnosis using mutation analysis of DNA extracted from fetal cells obtained by amniocentesis or by measuring GAA enzyme activity in chorionic villi or amniocytes is available. Carrier detection and prenatal diagnosis, using DNA-based targeted mutation analysis, are also possible if disease-causing family mutations are already known.

The approval of enzyme replacement therapy with alglucosidase alfa in 2006 has changed the natural history and clinical course of Pompe disease. Other adjunctive treatment options include dietary modifications, submaximal aerobic exercise, and respiratory muscle strength training. Gene therapy is under study as another treatment modality. Early diagnosis with early enzyme replacement initiation is the key to treatment efficacy.

Pompe disease is now part of the recommended uniform screening panel (RUSP) for newborns in the United States and newborn screening (NBS) has been initiated in several states. In Taiwan, where NBS for Pompe disease is performed routinely for all infants, early disease detection and treatment initiation has led to better treatment outcomes in infantile Pompe patients.

Late-Onset Polyglucosan Body Myopathy due to GYG1 Mutations

This is an autosomal recessive, slowly progressive skeletal myopathy caused by mutations in the *GYG1* gene blocking glycogenin-1 biosynthesis. GYG1 mutations results in a reduced or complete absence of glyogenin-1 which is necessary for glycogen synthesis in muscles. Affected individuals commonly present with adult-onset proximal muscle weakness prominently affecting the hip and shoulder girdles. Cardiomyopathy and cardiac failure necessitating cardiac transplantation is seen. Compared to GSD IV APBD, nervous system involvement has not been reported although both disorders cause polyglucosan deposition.

GSD Mimicking Hypertrophic Cardiomyopathy

Danon disease is an X-linked glycogen storage disorder caused by mutations in the LAMP2 gene. This results in deficiency of lysosomal-associated membrane protein 2 (LAMP2), leading to accumulation of glycogen in the heart and skeletal muscle. Patients present primarily with hypertrophic cardiomyopathy, but can be distinguished from the usual causes of hypertrophic cardiomyopathy by their electrophysiological abnormalities, particularly ventricular pre-excitation and conduction defects. The onset of cardiac symptoms such as chest pain, palpitations, syncope, and cardiac arrest may occur between the ages of 8 and 15 years. Ocular manifestations are often under-recognized and include peripheral pigmentary retinopathy, lens changes, and abnormal electroretinograms. The prognosis for LAMP2 deficiency is poor, with progressive end-stage heart failure early in adulthood. Treatment is mainly symptomatic and involves management of heart failure, correction of conduction abnormalities, and physical therapy, among others. Cardiac transplantation can be considered for refractory cases of heart failure.

AMP-ACTIVATED PROTEIN KINASE GAMMA 2 DEFICIENCY (PRKAG2 DEFICIENCY)

AMP-activated protein kinase gamma 2 (PRKAG2) deficiency is caused by mutations in the *PRKAG2* gene. This gene encodes the γ2 subunit of AMP-activated protein kinase (AMPK) which is important in many cellular ATP metabolic pathways. Affected individuals present with cardiac abnormalities including hypertrophic cardiomyopathy and conduction system abnormalities, particularly Wolff-Parkinson-White syndrome. The extent of cardiac involvement is variable and includes supraventricular tachycardia, sinus bradycardia, left ventricular dysfunction or even sudden cardiac death in some cases. In addition to cardiac involvement, there is a broad spectrum of phenotypic presentations including myalgia, myopathy and seizures. Unlike Danon disease, cardiomyopathy due to *PRKAG2* mutations is compatible with long-term survival except for a congenital form that presents in early infancy with a rapid fatal course. PRKAG2 syndrome should be considered as a differential diagnosis in infants presenting with severe hypertrophic cardiomyopathy. In rare instances, PRKAG2 cases may be misdiagnosed as infantile Pompe disease due to phenotypical similarity. Treatment is usually symptomatic and supportive, as in Danon disease.

TREATMENT

TREATMENT

Glycogen Storage Disease Mimicking Hypertrophic Cardiomyopathy

Heart transplantation has been suggested as a preventive measure for LAMP2 deficiency and noncongenital PRKAG2 deficiency.

SELECTED DISORDERS OF GALACTOSE METABOLISM

“Classic” galactosemia is caused by galactose 1-phosphate uridyltransferase (GALT) deficiency. It is a serious disease with an incidence of 1 in 60,000 and an early onset of symptoms. The newborn infant normally receives up to 40% of caloric intake as lactose (glucose + galactose). Without the transferase, the infant is unable to metabolize galactose 1-phosphate (Fig. 412-1), which consequently accumulates, resulting in injury to parenchymal cells of the kidney, liver, and brain. After the first feeding, infants can present with

vomiting, diarrhea, hypotonia, jaundice, and hepatomegaly. There is an increased risk for *Escherichia coli* neonatal sepsis in galactosemic infants; often with the onset of sepsis preceding the diagnosis of galactosemia.

Widespread newborn screening for galactosemia has identified these infants early and allowed them to be placed on dietary restriction. Elimination of galactose from the diet reverses growth failure as well as renal and hepatic dysfunction, improving the prognosis. However, on long-term follow-up, some patients still have ovarian failure manifesting as primary or secondary amenorrhea as well as developmental delays and learning disabilities that increase in severity with age. Of women with classic galactosemia, 80–90% or more report hypergonadotropic hypogonadism. While most female patients are infertile when they reach childbearing age, a few have given birth. Several mutations appear to be protective, particularly the p.Ser135Leu mutation, which is more common in the African-American population. Methods for fertility preservation, such as cryopreservation, are available. In addition, most patients have speech disorders, and a smaller proportion demonstrate poor growth, impaired motor function, and balance (with or without overt ataxia). Adults on dairy-free diets have developed cataracts, tremors, and low bone density. The treatment of galactosemia to prevent long-term complications remains a challenge.

Deficiency of *galactokinase* (Fig. 412-1) causes cataracts. Deficiency of *uridine diphosphate galactose 4-epimerase* can be benign when the enzyme deficiency is limited to blood cells, but can be as severe as classic galactosemia when the enzyme deficiency is generalized.

SELECTED DISORDERS OF FRUCTOSE METABOLISM

Fructokinase deficiency, or essential fructosemia (Fig. 412-1), causes a benign condition that is incidentally diagnosed from the presence of fructose as a reducing substance in the urine.

Deficiency of *fructose 1, 6-bisphosphate aldolase* (aldolase B; hereditary fructose intolerance) is a serious disease in infants. These patients are healthy and symptom-free until fructose or sucrose (table sugar) is ingested (usually from fruit, sweetened cereal, or sucrose-containing formula). Clinical manifestations may include jaundice, hepatomegaly, vomiting, lethargy, irritability, and convulsions. The incidence of celiac disease is higher among patients with hereditary fructose intolerance (>10%) than in the general population (1–3%). Laboratory findings show prolonged clotting time, hypoalbuminemia, elevation of bilirubin and aminotransferase levels, and proximal renal tubular dysfunction. If the disease goes undiagnosed and the deleterious intake of sugar continues, hypoglycemic episodes recur, and eventually death can occur from progressive liver and renal failure. The mainstay of treatment is the elimination of all sources of sucrose, fructose, and sorbitol from the diet. Once dietary control is established, liver and kidney dysfunction improve, and catch-up growth is common; intellectual development is usually not affected. Over time, the patient's symptom intensity improves, even after fructose ingestion. The long-term prognosis is good.

Fructose 1,6-diphosphatase deficiency is characterized by childhood life-threatening episodes of hypoglycemia, acidosis, hyperventilation, convulsions, and coma. These episodes are often triggered by foods that contain fructose, and include febrile infections and gastroenteritis when oral food intake is low.

Laboratory findings include low blood glucose levels, high lactate and uric acid levels, and metabolic acidosis. Renal tubular and liver functions are normal and aversion to sweets is usually not seen, unlike hereditary fructose intolerance. Treatment of acute episodes requires the correction of hypoglycemia and acidosis by IV infusion of dextrose. Further episodes can be prevented by avoidance of fasting and elimination of fructose and sucrose from the diet. A complex carbohydrate such as cornstarch, which provides slow and sustained levels of glucose, is useful for the long-term prevention of hypoglycemia. With proper treatment, prognosis is good, and patients who survive childhood develop normally.

GLOBAL CONSIDERATIONS



The GSDs and other inherited disorders of carbohydrate metabolism, although individually rare, are reported in most ethnic populations. The prevalent genetic mutations for each disease may vary in different ethnic populations, but clinical symptoms are remarkably similar and treatment guidelines apply to all. Symptomatic treatment is available for these disorders, and today, advances in the field have resulted in more definitive treatment approaches. Availability of NBS for Pompe disease has shown that the frequency of Pompe disease is much higher than previously estimated. This has allowed for early treatment initiation, and improved outcomes. NBS also mitigates the long diagnostic delays and misdiagnoses often associated with Pompe disease. The lessons learned from Pompe disease have bearing on the other GSDs.

ACKNOWLEDGMENT

The authors are grateful to Mrudu Herbert for her contributions to this chapter.

FURTHER READING

Bosch AM: Classical galactosaemia revisited. J Inherit Metab Dis 29:516, 2006. [PubMed: 16838075]

Bouteldja N, Timson DJ: The biochemical basis of hereditary fructose intolerance. J Inherit Metab Dis 33:105, 2010. [PubMed: 20162364]

Kishnani PS et al: Diagnosis and management of glycogen storage disease type I: A practice guideline of the American College of Medical Genetics and Genomics. Genet Med 16:e1, 2014. [PubMed: 25356975]

Kishnani PS, Beckemeyer AA, Mendelsohn NJ: The new era of Pompe disease: Advances in the detection, understanding of the phenotypic spectrum, pathophysiology, and management. Am J Med Genet C Semin Med Genet 160c:1, 2012. [PubMed: 22253049]

Kishnani PS et al: Glycogen storage disease type III diagnosis and management guidelines. ACMG. Genet Med 12:446, 2010.

Lévesque S et al: Diagnosis of late-onset Pompe disease and other muscle disorders by next-generation sequencing. Orphanet J Rare Dis 11:8, 2016. [PubMed: 26809617]

Martin MA et al: Glycogen storage disease type V, in GeneReviews® [Internet], RA Pagon et al (eds). Seattle (WA): University of Washington, Seattle, 2006.

Porto AG et al: Clinical Spectrum of PRKAG2 Syndrome. Circ Arrhythm Electrophysiol 9: e003121, 2016. [PubMed: 26729852]

Yang CF et al: Very early treatment for infantile-onset Pompe disease contributes to better outcomes. J Pediatr 169:174.e1, 2016.

Chapter 413: Inherited Disorders of Amino Acid Metabolism in Adults

Nicola Longo

INTRODUCTION

Amino acids are the building blocks of proteins and serve as neurotransmitters (glycine, glutamate, γ-aminobutyric acid) or as precursors of hormones, coenzymes, pigments, purines, or pyrimidines. Eight amino acids, referred to as *essential* (histidine, isoleucine, leucine, lysine, methionine, phenylalanine, valine, threonine, and tryptophan), cannot be synthesized by humans and must be obtained from dietary sources. The others are formed endogenously. Each amino acid has a unique degradative pathway by which its nitrogen and carbon components are used for the synthesis of other amino acids, carbohydrates, and lipids. Disorders of amino acid metabolism and transport **(Chap. 414)** are individually rare—the incidences range from 1 in 10,000 for cystinuria or phenylketonuria to 1 in 200,000 for homocystinuria or alkaptonuria—but collectively, they affect perhaps 1 in 1000 newborns. Almost all are transmitted as autosomal recessive traits.

The features of inherited disorders of amino acid catabolism are summarized in **Table 413-1**. In general, these disorders are named for the compound that accumulates to highest concentration in blood (-*emias*) or urine (-*urias*). In the aminoacidopathies, the parent amino acid is found in excess, whereas products in the catabolic pathway accumulate in organic acidemias. Which compound(s) accumulates depends on the site of the enzymatic block, the reversibility of the reactions proximal to the lesion, and the availability of alternative pathways of metabolic "runoff." Biochemical and genetic heterogeneity are common. Five distinct forms of hyperphenylalaninemia, nine forms of homocystinuria, and methylmalonic acidemia are recognized. Such heterogeneity reflects the presence of a large array of molecular defects.

TABLE 413-1

Inherited Disorders of Amino Acid Metabolism

Amino Acid(s)	Condition	Enzyme Defect	Clinical Findings	Inheritance
Phenylalanine	DNAJC12 Deficiency	Hydroxylase Co-Chaperone	Dystonia, parkinsonism, intellectual disability	AR
	Phenylketonuria	Phenylalanine hydroxylase	Intellectual disability, microcephaly, hypopigmented skin and hairs, eczema, “mousy” odor	AR
	DHPR deficiency	Dihydropteridine reductase	Intellectual disability, hypotonia, spasticity, myoclonus	AR
	PTPS deficiency	6-Pyruvoyl-tetrahydropterin synthase	Dystonia, neurologic deterioration, seizures, intellectual disability	AR
	GTP cyclohydrolase I deficiency	GTP cyclohydrolase I	Intellectual disability, seizures, dystonia, temperature instability	AR
	Carbinolamine dehydratase deficiency	Pterin-4α-carbinolamine dehydratase	Transient hyperphenylalaninemia (benign)	AR
Tyrosine	Tyrosinemia type I (hepatorenal)	Fumarylacetoacetate hydrolase	Liver failure, cirrhosis, rickets, failure to thrive, peripheral neuropathy, “boiled cabbage” odor	AR
	Tyrosinemia type II (oculocutaneous)	Tyrosine transaminase	Palmoplantar keratosis, painful corneal erosions with photophobia, learning disability	AR
	Tyrosinemia type III	4-Hydroxyphenylpyruvate dioxygenase	Hypertyrosinemia with normal liver function, occasional mental delay	AR
	Hawkinsinuria	4-Hydroxyphenylpyruvate dioxygenase	Transient failure to thrive, metabolic acidosis in infancy	AD
	Alkaptonuria	Homogentisic acid oxidase	Ochronosis, arthritis, cardiac valve involvement, coronary artery calcification	AR
	Albinism (oculocutaneous)	Tyrosinase	Hypopigmentation of hair, skin, and optic fundus; visual loss; photophobia	AR
	Albinism (ocular)	Different enzymes or transporters	Hypopigmentation of optic fundus, visual loss	AR, XL
	DOPA-responsive dystonia	Tyrosine hydroxylase	Rigidity, truncal hypotonia, tremor, intellectual disability	AR
GABA	4-Hydroxybutyric aciduria	Succinic semialdehyde dehydrogenase	Seizures, intellectual disability, hypotonia	AR
	ABAT deficiency	GABA transaminase	Seizures, intellectual disability, hypotonia	AR
Tryptophan	Hydroxykynureninuria	Kynureninase	Intellectual disability, spasticity	AR
Histidine	Histidinemia	Histidine-ammonia lyase	Benign	AR
	Urocanic aciduria	Urocanase	Occasional intellectual disability	AR
	Formiminoglutamic aciduria	Formiminotransferase	Occasional intellectual disability	AR

Amino Acid(s)	Condition	Enzyme Defect	Clinical Findings	Inheritance
Glycine	Glycine encephalopathy	Glycine cleavage (4 enzymes)	Infantile seizures, lethargy, apnea, profound intellectual disability	AR
	Sarcosinemia	Sarcosine dehydrogenase	Benign	AR
	Hyperoxaluria type I	Alanine:glyoxylate aminotransferase	Calcium oxalate nephrolithiasis, renal failure	AR
	Hyperoxaluria type II	D-Glyceric acid dehydrogenase/glyoxylate reductase	Calcium oxalate nephrolithiasis, renal failure	AR
Serine	Phosphoglycerate dehydrogenase deficiency	Phosphoglycerate dehydrogenase	Seizures, microcephaly, intellectual disability	AR
	PSAT1 deficiency	Phosphoserine aminotransferase	Seizures, microcephaly, intellectual disability	AR
	PSPHP deficiency	Phosphoserine phosphatase	Seizures, microcephaly, intellectual disability	AR
Proline	Hyperprolinemia type I	Proline oxidase	Benign	AR
	Hyperprolinemia type II	Δ^1-Pyrroline-5-carboxylate dehydrogenase	Febrile seizures, intellectual disability	AR
	Hyperhydroxyprolinemia	Hydroxyproline oxidase	Benign	AR
	Prolidase deficiency	Prolidase	Mild intellectual disability, chronic dermatitis	AR
Methionine	Hypermethioninemia	Methionine adenosyltransferase	Usually benign	AR
	S-Adenosylhomocysteine hydrolase deficiency	S-Adenosylhomocysteine hydrolase	Hypotonia, intellectual disability, absent tendon reflexes, delayed myelination	AR
	Glycine N-methyltransferase deficiency	Glycine N-methyltransferase	Elevated liver transaminases	AR
	Adenosine kinase deficiency	Adenosine kinase	Intellectual disability, seizures, liver dysfunction	AR
Homocystine	Homocystinuria	Cystathionine β-synthase	Lens dislocation, thrombotic vascular disease, intellectual disability, osteoporosis	AR
	Homocystinuria	5,10-Methylenetetrahydrofolate reductase	Intellectual disability, gait and psychiatric abnormalities, recurrent strokes	AR
	Homocystinuria	Methionine synthase (cblE, G)	Intellectual disability, hypotonia, seizures, megaloblastic anemia	AR
	Homocystinuria and methylmalonic acidemia	Vitamin B_{12} lysosomal efflux and metabolism (cblC, -D, -F, -J, -X)	Intellectual disability, lethargy, failure to thrive, hypotonia, seizures, megaloblastic anemia	AR
Cystathionine	Cystathioninuria	β-Cystathioninase	Benign	AR
Cystine	Cystinosis	Cystinosin CTNS (lysosomal efflux)	Renal Fanconi's syndrome, rickets, photophobia, hypotonia, renal failure	AR

Amino Acid(s)	Condition	Enzyme Defect	Clinical Findings	Inheritance
S-Sulfo-L-cysteine	Sulfocysteinuria	Sulfate oxidase or molybdenum cofactor deficiency	Seizures, intellectual disability, dislocated lenses	AR
Lysine	Hyperlysinemia, saccharopinuria	α-Aminoadipic semialdehyde synthase	Benign	AR
	Pyridoxine-dependent seizures	L-Δ^1-Piperideine-6-carboxilate dehydrogenase	Seizures, intellectual disability	AR
Lysine, tryptophan	α-Ketoadipic acidemia	α-Ketoadipic acid dehydrogenase	Benign	?
	Glutaric acidemia type I	Glutaryl-CoA dehydrogenase	Progressive severe dystonia and athetosis, motor delays	AR
	Glutaric acidemia type II	Electron transfer flavoproteins (ETF) or ETF:ubiquinone oxidoreductase	Hypoglycemia, metabolic acidosis, “sweaty feet” odor, hypotonia, cardiomyopathy	AR
Ornithine	Gyrate atrophy of the choroid and retina	Ornithine-Δ-aminotransferase	Myopia, night blindness, loss of peripheral vision, cataracts, chorioretinal degeneration	AR
Urea cycle	Carbamoylphosphate synthase-1 deficiency	Carbamoylphosphate synthase-1	Lethargy progressing to coma, protein aversion, intellectual disability, hyperammonemia	AR
	N-Acetylglutamate synthase deficiency	*N*-Acetylglutamate synthase	Lethargy progressing to coma, protein aversion, intellectual disability, hyperammonemia	AR
	Ornithine transcarbamylase deficiency	Ornithine transcarbamylase	Lethargy progressing to coma, protein aversion, intellectual disability, hyperammonemia	XL
	Citrullinemia type I	Argininosuccinate synthase	Lethargy progressing to coma, protein aversion, intellectual disability, hyperammonemia, liver failure	AR
	Argininosuccinic acidemia	Argininosuccinate lyase	Lethargy progressing to coma, protein aversion, intellectual disability, hyperammonemia, trichorrhexis nodosa	AR
	Arginase deficiency	Arginase	Spastic tetraparesis, microcephaly, intellectual disability, mild hyperammonemia	AR
	Hyperornithinemia, hyperammonemia, homocitrullinuria	Mitochondrial ornithine carrier ORNT1	Vomiting, lethargy, failure to thrive, intellectual disability, episodic confusion, hyperammonemia, protein intolerance	AR
	Citrullinemia type 2	Mitochondrial aspartate/glutamate carrier CTLN2	Neonatal intrahepatic cholestasis, adult presentation with sudden behavioral changes and stupor, coma, hyperammonemia	AR
Proline, ornithine, arginine	Δ^1-Pyrroline-5-carboxylate synthase deficiency	Δ^1-Pyrroline-5-carboxylate synthase	Hypotonia, seizures, neurodegeneration, peripheral neuropathy, joint laxity, skin hyperelasticity, subcapsular cataracts, hyperammonemia	AR
Glutamine	Glutamine synthase deficiency	Glutamine synthase	Brain malformations, pachygyria, seizures, hypotonia, dysmorphic features	AR
Valine	Hypervalinemia	Branched chain aminotransferase-2	Headache, memory impairment, failure to thrive, hypotonia, developmental delays	AR

Amino Acid(s)	Condition	Enzyme Defect	Clinical Findings	Inheritance
	Isobutyryl-CoA dehydrogenase deficiency	Isobutyryl-CoA dehydrogenase	Benign	AR
Valine, leucine, isoleucine	Maple syrup urine disease	Branched chain ketoacid dehydrogenase (E1α, E1β, E2, E3 deficiency)	Lethargy, vomiting, encephalopathy, seizures, intellectual disability, “maple syrup” odor, protein intolerance	AR
Leucine	Isovaleric acidemia	Isovaleryl-CoA dehydrogenase	Acidosis, ketosis, vomiting, coma, hyperammonemia, “sweaty feet” odor, protein intolerance	AR
	3-Methylcrotonyl glycinuria	3-Methylcrotonyl-CoA carboxylase	Stress-induced metabolic acidosis, hypotonia, hypoglycemia, “cat's urine” odor	AR
	3-Methylglutaconic aciduria type I	3-Methylglutaconyl-CoA hydratase deficiency	Stress-induced acidosis, leukodystrophy, hypotonia, hepatomegaly	AR
	3-Hydroxy-3-methylglutaric aciduria	3-Hydroxy-3-methylglutaryl-CoA lyase	Stress-induced hypoketotic hypoglycemia and acidosis, encephalopathy, hyperammonemia	AR
Isoleucine	2-Methylbutyryl-glycinuria	2-Methylbutyryl-CoA dehydrogenase	Benign	AR
	2-Methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency	2-Methyl-3-hydroxybutyryl-CoA dehydrogenase	Developmental regression, seizures, and rigidity sometimes triggered by illnesses	XL
	3-Oxothiolase deficiency	3-Oxothiolase	Fasting-induced acidosis and ketosis, vomiting, lethargy	AR
Valine, isoleucine, methionine, threonine	Propionic acidemia (pccA, -B, -C)	Propionyl-CoA carboxylase	Metabolic ketoacidosis, hyperammonemia, hypotonia, lethargy, coma, protein intolerance, intellectual disability, hyperglycinemia	AR
	Multiple carboxylase/biotinidase deficiency	Holocarboxylase synthase or biotinidase	Metabolic ketoacidosis, diffuse rash, alopecia, seizures, intellectual disability	AR
	Methylmalonic acidemia (mutase, cblA, B, racemase)	Methylmalonyl-CoA mutase/racemase or cobalamin reductase/adenosyltransferase	Metabolic ketoacidosis, hyperammonemia, hypertonia, lethargy, coma, protein intolerance, intellectual disability, hyperglycinemia	AR

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; Cbl, cobalamin; DOPA, dihydroxyphenylalanine; GABA, γ-aminobutyric acid; GTP, guanosine 5′-triphosphate; XL, X-linked.

The manifestations of these conditions differ widely (Table 413-1). Some, such as sarcosinemia, produce no clinical consequences. At the other extreme, complete deficiency of ornithine transcarbamylase is lethal in the untreated neonate. Central nervous system (CNS) dysfunction, in the form of developmental retardation, seizures, alterations in sensorium, or behavioral disturbances, is present in more than half the disorders. Protein-induced vomiting, neurologic dysfunction, and hyperammonemia occur in many disorders of urea cycle intermediates. Metabolic ketoacidosis, often accompanied by hyperammonemia, is frequent in organic acidemias. Some disorders produce focal tissue or organ involvement such as liver disease, renal failure, cutaneous abnormalities, or ocular lesions.

The analysis of plasma amino acids (by ion-exchange chromatography or liquid chromatography/tandem mass spectrometry), urine organic acids (by gas chromatography/mass spectrometry), and plasma acylcarnitine profile (by tandem mass spectrometry) is commonly used to diagnose and monitor most of these disorders. The diagnosis is confirmed by enzyme assay on cells or tissues from the patients or, more commonly, by DNA testing. The clinical manifestations in many of these conditions can be prevented or mitigated if a diagnosis is achieved early and appropriate treatment (e.g., dietary protein or

amino acid restriction or vitamin supplementation) is instituted promptly. For this reason, newborn screening programs seek to identify several of these disorders. Infants with a positive screening test need additional metabolic testing (usually suggested by the newborn screening program) to confirm or exclude the diagnosis. Confirmed cases should be referred to a metabolic center for initiation of therapy. The parents need to be counseled about the natural history the disease and its recurrence risk in future pregnancies. In some cases, parents need testing because they might have a disorder themselves (such as glutaric acidemia type 1, methylcrotonyl coenzyme A carboxylase deficiency, primary carnitine deficiency, or fatty acid oxidation defects) since mothers with these conditions can sometimes be identified by abnormal newborn screening results in their offspring. Some metabolic disorders can remain asymptomatic until adult age, presenting only when fasting or severe stress require full activity of affected metabolic pathways to provide energy.

Selected disorders that illustrate the principles, properties, and problems presented by the disorders of amino acid metabolism are discussed in this chapter.

THE HYPERPHENYLALANINEMIAS

The hyperphenylalaninemias (Table 413-1) result from impaired conversion of phenylalanine to tyrosine. The most common and clinically important is *phenylketonuria* (frequency 1:16,500), which is an autosomal recessive disorder characterized by an increased concentration of phenylalanine and its by-products in body fluids and by severe intellectual disability if untreated in infancy. It results from reduced activity of phenylalanine hydroxylase. The accumulation of phenylalanine inhibits the transport of other amino acids required for protein or neurotransmitter synthesis, reduces synthesis and increases degradation of myelin, and leads to inadequate formation of norepinephrine and serotonin. Phenylalanine is a competitive inhibitor of tyrosinase, a key enzyme in the pathway of melanin synthesis, and accounts for the hypopigmentation of hair and skin. Untreated children with classic phenylketonuria are normal at birth but fail to attain early developmental milestones, develop microcephaly, and demonstrate progressive impairment of cerebral function. Hyperactivity, seizures, and severe intellectual disability are major clinical problems later in life. Electroencephalographic abnormalities; “mousy” odor of skin, hair, and urine (due to phenylacetate accumulation); and a tendency to develop hypopigmentation and eczema complete the devastating clinical picture. In contrast, affected children who are detected and treated at birth show none of these abnormalities.

TREATMENT

TREATMENT

Phenylketonuria

To prevent intellectual disability, diagnosis and initiation of dietary treatment of classic phenylketonuria must occur before the child is 2 weeks of age. For this reason, newborns in North America, Australia, and Europe are screened by determinations of blood phenylalanine levels. Abnormal values are confirmed using quantitative analysis of plasma amino acids. Dietary phenylalanine restriction is usually instituted if blood phenylalanine levels are >360 µmol/L (6 mg/dL). Treatment consists of a special diet low in phenylalanine and supplemented with tyrosine, since tyrosine becomes an essential amino acid in phenylalanine hydroxylase deficiency. With therapy, plasma phenylalanine concentrations should be maintained between 120 and 360 µmol/L (2 and 6 mg/dL). Dietary restriction should be continued and monitored indefinitely. Some patients with milder forms of phenylketonuria (phenylalanine <1200 µmol/L at presentation) show increased tolerance to dietary proteins and improved metabolic control when treated with tetrahydrobiopterin (5–20 mg/kg per day), an essential cofactor of phenylalanine hydroxylase.

A number of women with phenylketonuria who have been treated since infancy will reach adulthood and become pregnant. If maternal phenylalanine levels are not strictly controlled before and during pregnancy, their offspring are at increased risk for congenital defects and microcephaly (*maternal phenylketonuria*). After birth, these children have severe intellectual disability and growth retardation. Pregnancy risks can be minimized by continuing lifelong phenylalanine-restricted diets and assuring strict phenylalanine restriction 2 months prior to conception and throughout gestation.

THE HOMOCYSTINURIAS (HYPERHOMOCYSTEINEMIAS)

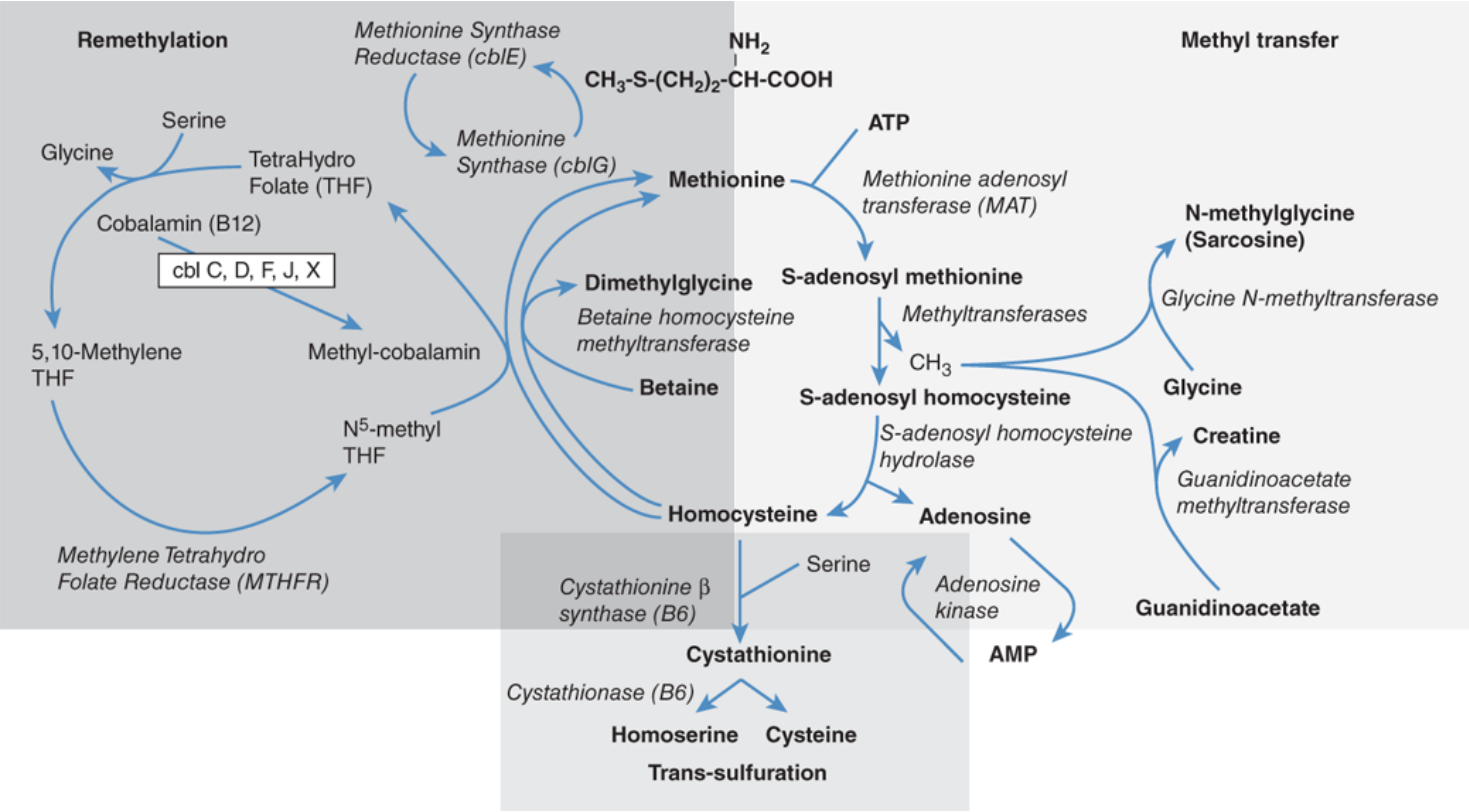
The homocystinurias are nine biochemically and clinically distinct disorders (Table 413-1) characterized by increased concentration of the sulfur-containing amino acid homocystine in blood and urine.

Classic homocystinuria, the most common (frequency 1:400,000), results from reduced activity of cystathionine β-synthase **(Fig. 413-1)**, the pyridoxal phosphate–dependent enzyme that condenses homocysteine with serine to form cystathionine. Most patients present between 3 and 5 years of age with dislocated optic lenses and intellectual disability (in about half of cases). Some patients develop a marfanoid habitus and radiologic evidence of osteoporosis.

FIGURE 413-1

Pathways, enzymes, and coenzymes involved in the homocystinurias. Methionine transfers a methyl group during its conversion to homocysteine. Defects in methyl transfer or in the subsequent metabolism of homocysteine by the pyridoxal phosphate (vitamin B_6)-dependent cystathionine β-synthase increase plasma methionine levels. Homocysteine is transformed into methionine via remethylation. This occurs through methionine synthase, a reaction requiring

methylcobalamin and folic acid. Deficiencies in these enzymes or lack of cofactors is associated with decreased or normal methionine levels. In an alternative pathway, homocysteine can be remethylated by betaine:homocysteine methyl transferase.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Life-threatening vascular complications (affecting coronary, renal, and cerebral arteries) can occur during the first decade of life and are the major cause of morbidity and mortality. Classic homocystinuria can be diagnosed with analysis of plasma amino acids, showing elevated methionine and presence of free homocystine. Total plasma homocysteine is also extremely elevated (usually >100 μ*M*). Treatment consists of a special diet restricted in protein and methionine. In approximately half of patients, oral pyridoxine (25–500 mg/d) produces a fall in plasma methionine and homocystine concentration in body fluids. Folate and vitamin B_{12} deficiency should be prevented by adequate supplementation. Betaine is also effective in reducing homocystine levels by favoring its remethylation to methionine.

The other forms of homocystinuria are the result of impaired remethylation of homocysteine to methionine. This can be caused by defective methionine synthase or reduced availability of two essential cofactors, 5-methyltetrahydrofolate and methylcobalamin (methyl-vitamin B_{12}). In contrast to cystathionine β-synthase, elevated levels of free homocystine are associated with low levels of methionine in the plasma amino acid profile in remethylation defects. Therapy in these cases requires administration of methylfolate, hydroxycobalamin (an activated form of vitamin B_{12}), and betaine.

Hyperhomocysteinemia refers to increased total plasma concentration of homocysteine with or without an increase in free homocystine (disulfide form). Hyperhomocysteinemia, in the absence of significant homocystinuria, is found in some heterozygotes for the genetic defects noted above or in homozygotes for milder variants. Changes of homocysteine levels are also observed with increasing age; with smoking; in postmenopausal women; in patients with renal failure, hypothyroidism, leukemias, inflammatory bowel disease, or psoriasis; and during therapy with drugs such as methotrexate, nitrous oxide, isoniazid, and some antiepileptic agents. Homocysteine can act as an atherogenic and thrombophilic agent and increased total plasma homocysteine have been associated with an increased risk for coronary, cerebrovascular, and peripheral arterial disease as well as for deep-vein thrombosis**.** In addition, hyperhomocysteinemia and folate and vitamin B_{12} deficiencies have been associated with an increased risk of neural tube defects in pregnant women and dementia (Alzheimer's type) in the general population. Vitamin supplements are effective in reducing plasma homocysteine levels in these cases, although there are limited effects on cardiovascular disease.

ALKAPTONURIA

Alkaptonuria is a rare (frequency 1:200,000) disorder of tyrosine catabolism in which deficiency of homogentisate 1,2-dioxygenase (also known as *homogentisic acid oxidase*) leads to excretion of large amounts of homogentisic acid in urine and accumulation of oxidized homogentisic acid pigment in connective tissues (*ochronosis*). Alkaptonuria may go unrecognized until middle life, when degenerative joint disease develops. Prior to this time, about half of patients might be diagnosed for the presence of urine that become dark with standing or addition of alkali. Foci of gray-brown scleral pigment and

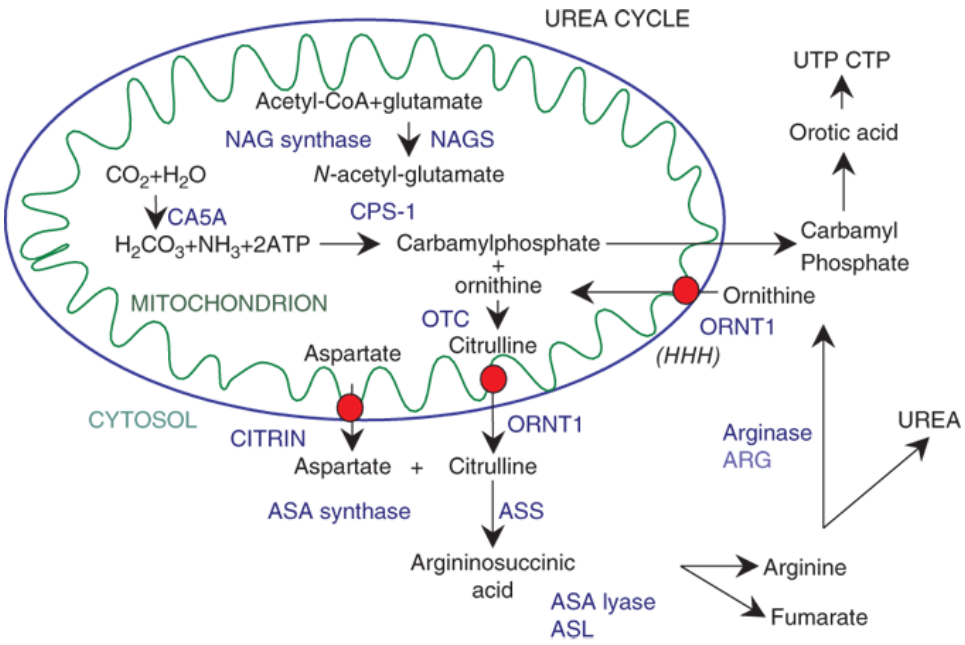
generalized darkening of the concha, anthelix, and, finally, helix of the ear usually develop after age 30. Low back pain usually starts between 30 and 40 years of age. *Ochronotic arthritis* is heralded by pain, stiffness, and some limitation of motion of the hips, knees, and shoulders. Acute arthritis may resemble rheumatoid arthritis, but small joints are usually spared. Pigmentation of heart valves, larynx, tympanic membranes, and skin occurs, and occasional patients develop pigmented renal or prostatic calculi. Pigment deposition in the heart and blood vessels leads to aortic stenosis necessitating valve replacement, especially after 60 years of age. The diagnosis should be suspected in a patient whose urine darkens to blackness. Homogentisic acid in urine is identified by urine organic acid analysis. Ochronotic arthritis is treated symptomatically with pain medications, spinal surgery, and arthroplasty **(Chap. 364)**. Ascorbic acid and protein restriction are not effective in reducing homogentisic acid production. By contrast, nitisinone (2-[2-nitro-4-trifluoromethylbenzoyl]-1,3-cyclohexanedione), a drug used in tyrosinemia type I, reduces urinary excretion of homogentisic acid and, in conjunction with a low-protein diet, might prevent the long-term complications of alkaptonuria.

UREA CYCLE DEFECTS

Excess ammonia generated from protein nitrogen is removed by the urea cycle, a process mediated by several enzymes and transporters (**Fig. 413-2**, Table 413-1). Complete absence of any of these enzymes usually causes severe hyperammonemia in newborns, while milder variants can be seen in adults. The accumulation of ammonia and glutamine leads to direct neuronal toxicity and brain edema. Deficiencies in urea cycle enzymes are individually rare, but as a group, they affect about 1:35,000 individuals. They are all transmitted as autosomal recessive traits, with the exception of ornithine transcarbamylase deficiency, which is X-linked and the most frequent urea cycle defect. Hepatocytes of females with ornithine transcarbamylase deficiency express either the normal or the mutant allele due to random X-inactivation and may be unable to remove excess ammonia if mutant cells are predominant.

FIGURE 413-2

The urea cycle. This cycle, which is fully expressed only in the liver, forms urea starting from ammonia (NH_3) derived from the nitrogen group of all amino acids. It requires many enzymes and mitochondrial transporters, any of which can be defective and may impair the function of the urea cycle. Ammonia escaping the urea cycle in periportal hepatocytes is conjugated with glutamate by glutamine synthase in perivenous hepatocytes to generate glutamine. ARG, arginase; ASA, Argininosuccinic acid; ASL, Arginino succinate lyase; ASS, argininosuccinate synthase; CA5A, carbonic anhydrase 5a; citrin (*SLC25A13*), aspartate/glutamate exchanger; CP, carbamylphosphate; CPS-1, carbamyl phosphate synthase 1; CTP, cytidine triphosphate; HHH, hyperammonemia, hyperornithinemia, homocitrullinuria syndrome; NAG, *N*-acetylglutamate; NAGS, N-acetylglutamate synthase; ORNT1 (*SLC25A15*), ornithine/citrulline mitochondrial transporter; OTC, ornithine transcarbamylase; UTP, uridine triphosphate.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Infants with classic urea cycle defects present at 1–4 days of life with refusal to eat and lethargy progressing to coma and death. Milder enzyme deficiencies present with protein avoidance, recurrent vomiting, migraine, mood swings, chronic fatigue, irritability, and disorientation that can progress to coma. Some cases have presented with acute or chronic hepatic dysfunction. Females with ornithine transcarbamylase deficiency can present at time of childbirth due to the combination of involuntary fasting and stress that favors catabolism. Administration of systemic corticosteroids can precipitate hyperammonemia and can be fatal in previously asymptomatic individuals. These patients may be misdiagnosed as having gastrointestinal disorders, food allergies, behavioral problems, or nonspecific hepatitis. The diagnosis requires measurement of plasma ammonia, plasma amino acids, and urine orotic acid, useful for differentiating ornithine transcarbamylase deficiency from carbamyl phosphate synthase-1 and *N*-acetylglutamate synthase deficiency. Increased plasma glutamine is seen with all urea cycle defects since ammonia not removed by the urea cycle in periportal hepatocytes is conjugated to glutamate by glutamine synthase in perivenous hepatocytes. Citrulline is low or undetectable in proximal defects of the urea cycle (*N*-acetylglutamate

synthase, carbamylphosphate synthase 1, and ornithine transcarbamylase deficiency), with urine orotic acid being increased only in ornithine transcarbamylase deficiency. Plasma citrulline is markedly increased in argininosuccinic acid synthase deficiency (citrullinemia type 1), with a milder elevation in argininosuccinic acid lyase deficiency in the presence of argininosuccinic acid (argininosuccinic aciduria). Arginine levels are usually normal to low in these conditions and become markedly elevated only in patients with arginase deficiency. In addition to urea cycle defects, hyperammonemia can also be caused by liver disease from any cause and several organic acidemias and fatty acid oxidation defects (the latter two excluded by the analysis of urine organic acids and plasma acylcarnitine profile).

TREATMENT

TREATMENT

Urea Cycle Defects

Therapy is aimed at stopping catabolism and ammonia production by providing adequate calories (as IV glucose and lipids in the comatose patient) and, if needed, insulin. Excess nitrogen is removed by IV phenylacetate and benzoate (0.25 g/kg for the priming dose and subsequently as an infusion over 24 h) that conjugate with glutamine and glycine, respectively, to form phenylacetylglutamine and hippuric acid, water-soluble molecules efficiently excreted in urine. Arginine (200 mg/kg per day) becomes an essential amino acid (except in arginase deficiency) and should be provided intravenously to resume protein synthesis. If these measures fail to reduce ammonia, hemodialysis should be initiated promptly. Chronic therapy consists of a protein-restricted diet, phenylbutyrate, glycerol phenylbutyrate (a liquid drug better tolerated by most patients), arginine, or citrulline supplements, depending on the specific diagnosis. Oral carglumic acid can restore a functional urea cycle in patients with *N*-acetylglutamate synthase deficiency and can render other therapies unnecessary. Liver transplantation should be considered in patients with severe urea cycle defects that are difficult to control medically.

Hyperammonemia due to a functional deficiency of glutamine synthase can occur in patients receiving chemotherapy for different malignancies or undergoing solid organ transplants. It can also be seen with hepatic cirrhosis. Several of these patients have been successfully rescued from hyperammonemia using the protocol described above for urea cycle defects.

FURTHER READING

Camp KM et al: Phenylketonuria Scientific Review Conference: State of the science and future research needs. Mol Genet Metab 112:87, 2014. [PubMed: 24667081]

Ranganath LR et al: Suitability Of Nitisinone In Alkaptonuria 1 (SONIA 1): An international, multicentre, randomised, open-label, no-treatment controlled, parallel-group, dose-response study to investigate the effect of once daily nitisinone on 24-h urinary homogentisic acid excretion in patients with alkaptonuria after 4 weeks of treatment. Ann Rheum Dis 75:362, 2016. [PubMed: 25475116]

Smith AD, Refsum H: Homocysteine, B vitamins, and cognitive impairment. Annu Rev Nutr 36:211, 2016. [PubMed: 27431367]

Chapter 414: Inherited Defects of Membrane Transport

Nicola Longo

INTRODUCTION

Specific membrane transporters mediate the passage of a wide variety of substances across cellular membranes. Classes of substrates include amino acids, sugars, cations, anions, vitamins, and water. The number of inherited disorders of membrane transport continues to increase with the identification of new transporters on the plasma membrane or intracellular organelles and the clarification of the molecular basis of diseases with previously unknown pathophysiology. The first transport disorders identified affected the gut or the kidney, but transport processes are now proving essential for the normal function of every organ. Mutations in transporter molecules cause disorders of the heart, muscle, brain, and endocrine and sensory organs (Table 414-1). Inherited defects impairing the transport of selected amino acids that can present in adults are discussed here as examples of the abnormalities encountered; others are considered elsewhere in this text.

TABLE 414-1

Genetic Disorders of Membrane Transport (Selected Examples)

Class of Substance and Disorder	Individual Substrates	Tissues Manifesting Transport Defect	Molecular Defect	Major Clinical Manifestations	Inheritance
Amino Acids					
Cystinuria	Cystine, lysine, arginine, ornithine	Proximal renal tubule, jejunal mucosa	Shared dibasic-cystine transporter *SLC3A1, SLC7A9*	Cystine nephrolithiasis	AR
Lysinuric protein intolerance	Lysine, arginine, ornithine	Proximal renal tubule, jejunal mucosa	Dibasic transporter *SLC7A7*	Protein intolerance, hyperammonemia, intellectual disability	AR
Hartnup disease	Neutral amino acids	Proximal renal tubule, jejunal mucosa	Neutral amino acid transporter *SLC6A19*	Constant neutral aminoaciduria, intermittent symptoms of pellagra	AR
Brain branched-chain amino acid deficiency	Leucine, Isoleucine, Valine	Plasma membrane of blood brain barrier	Branched-chain amino acid transporter *SLC7A5*	Microcephaly, intellectual disability, seizures	AR
Citrullinemia type 2	Aspartate, glutamate, malate	Inner mitochondrial membrane	Mitochondrial aspartate/glutamate carrier 2 *SLC25A13*	Sudden behavioral changes with stupor, coma, hyperammonemia	AR

Class of Substance and Disorder	Individual Substrates	Tissues Manifesting Transport Defect	Molecular Defect	Major Clinical Manifestations	Inheritance
Hyperornithinemia, hyperammonemia, homocitrullinuria	Ornithine, citrulline	Inner mitochondrial membrane	Mitochondrial ornithine carrier *SLC25A15*	Vomiting, lethargy, failure to thrive, intellectual disability, episodic confusion, hyperammonemia, protein intolerance	AR
Histidinuria	Histidine	Proximal renal tubule, jejunal mucosa	Histidine transporter	Intellectual disability	AR
Iminoglycinuria	Glycine, proline, hydroxyproline	Proximal renal tubule, jejunal mucosa	Shared glycine–imino acid transporter *SLC36A2, SLC6A19, SLC6A20*	None	AR
Dicarboxylic aminoaciduria	Glutamic acid, aspartic acid	Proximal renal tubule, jejunal mucosa	Shared dicarboxylic amino acid transporter *SLC1A1*	None	AR
Cystinosis	Cystine	Lysosomal membranes	Lysosomal cystine transporter	Renal failure, hypothyroidism, blindness	AR
Hexoses					
Glucose-galactose malabsorption	D-Glucose D-Galactose	Proximal renal tubule, jejunal mucosa	Sodium-dependent glucose/galactose transporter SGLT1	Watery diarrhea on feeding glucose, lactose, sucrose, or galactose	AR

Class of Substance and Disorder	Individual Substrates	Tissues Manifesting Transport Defect	Molecular Defect	Major Clinical Manifestations	Inheritance
Glucose-transport defect	D-Glucose	Ubiquitous blood-brain barrier	Facilitative glucose transporter GLUT1	Seizures, intellectual disability	AD
Fanconi-Bickel syndrome	D-Glucose	Liver, kidney, pancreas, intestine	Facilitative glucose transporter GLUT2	Growth retardation, rickets, hepatorenal glycogenosis, hypo- and hyperglycemia	AR
Urate					
Hypouricemia	Uric acid	Proximal renal tubule	Urate transporter *SLC22A12*	Hypouricemia, uric acid urolithiasis	AR
Vitamins					
Thiamine-responsive megaloblastic anemia	Thiamine	Ubiquitous	Thiamine transporter *SLC19A2*	Megaloblastic anemia, deafness, diabetes mellitus	AR
Biotin-thiamine-responsive basal ganglia disease	Biotin, thiamine	Ubiquitous	Biotin-thiamine transporter *SLC19A3*	Dystonia, seizures, psychomotor delay, Wernicke-like encephalopathy	AR
Riboflavin transporter deficiencies	Riboflavin	Blood brain barrier	Riboflavin transporters *SLC52A2, SLC52A3*	Ataxia, weakness, neuropathy, hearing loss,	AR
Other					

Class of Substance and Disorder	Individual Substrates	Tissues Manifesting Transport Defect	Molecular Defect	Major Clinical Manifestations	Inheritance
Carnitine deficiency	Carnitine	Kidney, muscle, heart	Carnitine transporter OCTN2	Hypoketotic hypoglycemia, cardiomyopathy, sudden death	AR
Creatine deficiency	Creatine	Brain	Creatine transporter *SLC6A8*	Intellectual disability, seizures, hypotonia	XL

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; XL, X-linked recessive.

CYSTINURIA

Cystinuria (frequency of 1 in 10,000 to 1 in 15,000) is an autosomal recessive disorder caused by defective transporters in the apical brush border of proximal renal tubule and small intestinal cells. It is characterized by impaired reabsorption and excessive urinary excretion of the dibasic amino acids lysine, arginine, ornithine, and cystine. Because cystine is poorly soluble, its excess excretion predisposes to the formation of renal, ureteral, and bladder stones. Such stones are responsible for the signs and symptoms of the disorder.

There are two variants of cystinuria. Homozygotes for both variants have high urinary excretion of cystine, lysine, arginine, and ornithine. Type I heterozygotes usually have normal urinary amino acid excretion, whereas most non–type I (formerly type II and type III) heterozygotes have moderately increased urinary excretion of each of the four amino acids. The gene for type I cystinuria (*SLC3A1*, chromosome 2p16.3) encodes a membrane glycoprotein. Non–type I cystinuria is caused by mutations in *SLC7A9* (chromosome 19q13) that encodes the $b^{0,+}$ amino acid transporter. The glycoprotein encoded by *SLC3A1* favors the correct processing of the $b^{0,+}$ membrane transporter and explains why mutations in two different genes cause a similar disease.

Cystine stones account for 1–2% of all urinary tract calculi but are the most common cause of stones in children. Cystinuria homozygotes regularly excrete 2400–7200 µmol (600–1800 mg) of cystine daily. Since the maximum solubility of cystine in the physiologic urinary pH range of 4.5–7.0 is about 1200 µmol/L (300 mg/L), cystine needs to be diluted to 2.5–7 L of water to prevent crystalluria. Stone formation usually manifests in the second or third decade but may occur in the first year of life. Symptoms and signs are those typical of urolithiasis: hematuria, flank pain, renal colic, obstructive uropathy, and infection (Chap. 312). Recurrent urolithiasis may lead to progressive renal insufficiency.

Cystinuria is suspected after observing typical hexagonal crystals in the sediment of acidified, concentrated, chilled urine, or after performing a urinary nitroprusside test. Quantitative urine amino acid analysis confirms the diagnosis of cystinuria by showing selective overexcretion of cystine, lysine, arginine, and ornithine. Quantitative measurements are important for differentiating heterozygotes from homozygotes and for following free cystine excretion during therapy.

Management is aimed at preventing cystine crystal formation by increasing urinary volume and by maintaining an alkaline urine pH. Fluid ingestion in excess of 4 L/d is essential, and 5–7 L/d is optimal. Urinary cystine concentration should be <1000 µmol/L (250 mg/L). The daily fluid ingestion necessary to maintain this dilution of excreted cystine should be spaced over 24 h, with one-third of the total volume ingested between bedtime and 3 A.M. Cystine solubility rises sharply above pH 7.5, and urinary alkalinization (with bicarbonate or potassium citrate) can be therapeutic. Penicillamine (1–3 g/d) and tiopronin (α-mercaptopropionylglycine, 800–1200 mg/d in four divided doses) undergo sulfhydryl-disulfide exchange with cystine to form mixed disulfides. Because these disulfides are much more soluble than cystine, pharmacologic therapy can prevent and promote dissolution of calculi. Penicillamine can have significant side effects and should be reserved for patients who fail to respond to hydration alone or who are in a high-risk category (e.g., one remaining kidney, renal insufficiency). When medical management fails, shock waves lithotripsy, ureteroscopy, and percutaneous nephrolithotomy are effective for most stones. Open urologic surgery is considered only for complex staghorn stones or when the patient has concomitant renal or ureteral abnormalities. Occasional patients progress to renal failure and require kidney transplantation.

LYSINURIC PROTEIN INTOLERANCE

This disorder is characterized by a defect in renal tubular reabsorption of the three dibasic amino acids lysine, arginine, and ornithine but *not* cystine (*lysinuric protein intolerance*). Homozygotes show defective intestinal transport of dibasic amino acids as well as exaggerated renal losses. Lysinuric protein intolerance is most common in Finland (1 in 60,000), southern Italy, and Japan, but is rare elsewhere. The transport defect affects basolateral rather than luminal membrane transport and is associated with impairment of the urea cycle. The defective gene (*SLC7A7*, chromosome 14q11.2) encodes the y^+LAT membrane transporter, which associates with the cell-surface glycoprotein 4F2 heavy chain to form the complete sodium-independent transporter y^+L.

Manifestations are related to impairment of the urea cycle and to immune dysfunction potentially attributable to nitric oxide overproduction secondary to arginine intracellular trapping within macrophages. Affected patients present in childhood with hepatosplenomegaly, protein intolerance, and episodic ammonia intoxication. Older patients may present with severe osteoporosis, impairment of kidney function, pulmonary alveolar proteinosis, various autoimmune disorders, and an incompletely characterized immune deficiency. Plasma concentrations of lysine, arginine, and ornithine are reduced, whereas urinary excretion of lysine and orotic acid are increased. Hyperammonemia may develop after the ingestion of protein loads or with infections, probably because of insufficient amounts of arginine and ornithine to maintain proper function of the urea cycle. Therapy consists of dietary protein restriction and supplementation of citrulline (2–8 g/d), a neutral amino acid that fuels the urea cycle when metabolized to arginine and ornithine. Pulmonary disease

responds to glucocorticoids or recombinant human GM-CSF in some patients. Women with lysinuric protein intolerance who become pregnant have an increased risk of anemia, toxemia, and bleeding complications during delivery. These can be minimized by aggressive nutritional therapy and control of blood pressure. Their infants can have intrauterine growth restriction but have normal neurologic function.

CITRULLINEMIA TYPE 2 (CITRIN DEFICIENCY)

Citrullinemia type 2 is a recessive condition caused by deficiency of the mitochondrial aspartate-glutamate carrier AGC2 (citrin). A defect in this transporter reduces the availability of cytoplasmic aspartate to combine with citrulline to form argininosuccinate **(see Fig. 413-1)**, impairing the urea cycle and decreasing the transfer of reducing equivalents from the cytosol to the mitochondria through the malate–aspartate NADH shuttle. Mutations in the *SLC25A13* gene on chromosome 7q21.3 that encodes for this transporter are rare in Caucasians, but affect about 1:20,000 people with ancestry from Japan, China, and Southeast Asia with variable penetrance.

The disease can present in children with neonatal intrahepatic cholestasis, failure to thrive, and dyslipidemia, but usually presents with sudden onset between 20 and 50 years of age with recurring episodes of hyperammonemia with associated neuropsychiatric symptoms such as altered mental status, irritability, seizures, or coma-resembling hepatic encephalopathy. Some patients might come to medical attention for hypertriglyceridemia, pancreatitis, hepatoma, or fatty liver histologically similar to nonalcoholic steatohepatitis. Without therapy, most patients die with cerebral edema within a few years of onset. Episodes are usually triggered by medications (such as acetaminophen), surgery, alcohol consumption or high sugar intake, the latter conditions causing excess NADH production. NADH is not generated by the metabolism of proteins or fats, and many individuals with citrullinemia type 2 spontaneously prefer foods such as meat, eggs, and fish, and avoid carbohydrates.

Laboratory studies during an acute attack include elevated ammonia, citrulline, and arginine with low or normal levels of glutamine (the latter is usually increased in classic urea cycle defects). Levels of galactose-1-phosphate in red blood cells are also increased, reflecting defective transfer of reducing equivalents from the cytosol to mitochondria. The diagnosis is confirmed by demonstrating mutations in the *SLC25A13* gene. Liver transplantation prevents progression of the disease and normalizes biochemical parameters. A diet high in fats and proteins and low in carbohydrates with supplements of medium chain triglycerides, arginine, and pyruvate is also effective in preventing further episodes, at least in the short term.

HARTNUP DISEASE

Hartnup disease (frequency 1 in 24,000) is an autosomal recessive disorder characterized by pellagra-like skin lesions, variable neurologic manifestations, and neutral and aromatic aminoaciduria. Alanine, serine, threonine, valine, leucine, isoleucine, phenylalanine, tyrosine, tryptophan, glutamine, asparagine, and histidine are excreted in urine in quantities 5–10 times greater than normal, and intestinal transport of these same amino acids is defective. The defective neutral amino acid transporter, B°AT1 encoded by the *SLC6A19*

gene on chromosome 5p15, requires either collectrin or angiotensin-converting enzyme 2 for surface expression in the kidney and intestine, respectively.

The clinical manifestations result from nutritional deficiency of the essential amino acid tryptophan, caused by its intestinal and renal malabsorption, and of niacin, which derives in part from tryptophan metabolism. Only a small fraction of patients with the chemical findings of this disorder develop a pellagra-like syndrome, implying that manifestations depend on other factors in addition to the transport defect. The diagnosis of Hartnup disease should be suspected in any patient with clinical features of pellagra who does not have a history of dietary niacin deficiency **(Chap. 326)**. The neurologic and psychiatric manifestations range from attacks of cerebellar ataxia to mild emotional lability to frank delirium, and they are usually accompanied by exacerbations of the erythematous, eczematoid skin rash. Fever, sunlight, stress, and sulfonamide therapy provoke clinical relapses. Diagnosis is made by detection of the neutral aminoaciduria, which does not occur in dietary niacin deficiency. Treatment is directed at niacin repletion and includes a high-protein diet and daily nicotinamide supplementation (50–250 mg).

CYSTINOSIS

Cystinosis (frequency 1:100,000–1:200,000) is an autosomal recessive disorder caused by mutations in the *CTNS* gene encoding the lysosomal cystine/proton transporter (cystinosin). In this condition, cystine derived from protein degradation accumulates inside lysosomes and forms crystals due to its poor solubility. Depending on the degree of impairment of transporter function, three clinical forms are recognized. The most severe form, classic nephropathic cystinosis, causes renal Fanconi syndrome during the first year of life and, without treatment, evolves to renal failure usually by ten years of age. Intermediate nephropathic cystinosis leads to kidney failure between 15 and 25 years of age while photophobia, caused by deposition of cystine crystals in the cornea, is the only manifestation of ocular non-nephropathic cystinosis. Cystinosis is suspected by the identification of cystine crystals in the cornea by slit lamp examination and diagnosed by measuring cystine content in white blood cells. DNA testing (including deletion analysis) of the *CTNS* gene can further confirm the diagnosis. Therapy consists in the administration of cysteamine that enters lysosomes, forms a mixed disulfide with cysteine, and is exported from the lysosome using a cationic amino acid transporter. Oral cysteamine therapy (60–90 mg/kg per day up to 2 g per day in adults, 0.2 to 0.3 grams/m^2 per day divided into two doses given every 12 hours for the extended release formulation) can delay renal failure and is more effective if started early in the course of the disease. Therapy with cysteamine reduces intracellular cystine accumulation in white blood cells, but compliance with therapy is difficult due to the unpleasant odor of the drug and the need for frequent administration. Cysteamine eye drops can relieve photophobia. Renal replacement therapy with salts, alkali, and activated vitamin D is necessary for renal Fanconi syndrome. Cystine accumulation occurs in virtually all organs and tissues, causing additional complications such as hypothyroidism, hypohydrosis, diabetes, delayed puberty in both males and females with primary hypogonadism in males. Growth hormone replacement, L-thyroxine for hypothyroidism, insulin for diabetes mellitus, and testosterone for hypogonadism in males may be necessary. Despite therapy, many patients with cystinosis progress to end-stage renal failure and require kidney transplantation. Late-onset complications include hepatomegaly and splenomegaly that occur in approximately one-third of subjects and a vacuolar

myopathy causing weakness (initially involving the distal extremities), swallowing difficulties, gastrointestinal dysmotility and pulmonary insufficiency. Before the availability of cystine-depleting therapy and renal transplantation, the life span in nephropathic cystinosis was less than ten years. With current therapies, affected individuals can survive into the late forties with satisfactory quality of life.

FURTHER READING

Emma F et al: Nephropathic cystinosis: An international consensus document. Nephrol Dial Transplant. 29 Suppl 4:87, 2014.

Tanner LM et al: Inhaled sargramostim induces resolution of pulmonary alveolar proteinosis in lysinuric protein intolerance. JIMD Rep 34:97, 2017.

Tărlungeanu DC et al: Impaired amino acid transport at the blood brain barrier is a cause of autism spectrum disorder. Cell 167:1481, 2016. [PubMed: 27912058]

Chapter 415: Approach to the Patient with Neurologic Disease

Daniel H. Lowenstein; Joseph B. Martin; Stephen L. Hauser

INTRODUCTION

Neurologic diseases are common and costly. According to estimates by the World Health Organization, neurologic disorders affect over 1 billion people worldwide, constitute 12% of the global burden of disease, and cause 14% of global deaths (Table 415-1). These numbers are only expected to increase as the world's population ages. Because therapies now exist for many neurologic disorders, a skillful approach to diagnosis is essential. Errors commonly result from an overreliance on costly neuroimaging procedures and laboratory tests, which, while useful, do not substitute for an adequate history and examination. The proper approach begins with the patient and focuses the clinical problem first in anatomic and then in pathophysiologic terms; only then should a specific neurologic diagnosis be entertained. This method ensures that technology is judiciously applied, a correct diagnosis is established in an efficient manner, and treatment is promptly initiated.

TABLE 415-1

Global Disability-Adjusted Life-Years (DALYs) and Number of Annual Deaths for Selected Neurologic Disorders in 2015

Disorder	DALYs	Deaths
Low back and neck pain	94,941,000	—
Cerebrovascular diseases	118,627,000	6,326,000
Meningitis and encephalitis	33,848,000	529,000
Migraine	32,898,000	—
Epilepsy	12,418,000	125,000
Dementia	23,779,000	1,908,000
Parkinson's disease	2,059,000	117,000
% of total DALYs or deaths for all causes that are neurologic	9.7%	16.3%
% change of DALYs for neurologic disorders between 2005 and 2015	17.0%	35.2%

Source: GBD 2015 DALYs and HALE Collaborators: Global, Regional, and National Disability-Adjusted Life-Years (Dalys) for 315 Diseases and Injuries and Healthy Life Expectancy (HALE), 1990–2015: A Systematic Analysis for the Global Burden of Disease Study 2015. Lancet 388:1603, 2016; and GBD 2015 Mortality and Causes of Death Collaborators: Global, Regional, and National Life Expectancy, All-Cause Mortality, and Cause-Specific Mortality for 249 Causes of Death, 1980–2015: A Systematic Analysis for the Global Burden of Disease Study 2015.Lancet 388:1459, 2016.

THE NEUROLOGIC METHOD

DEFINE THE ANATOMY

The first priority is to identify the region of the nervous system that is likely to be responsible for the symptoms. Can the disorder be mapped to one specific location, is it multifocal, or is a diffuse process present? Are the symptoms restricted to the nervous system, or do they arise in the context of a systemic illness? Is the problem in the central nervous system (CNS), the peripheral nervous system (PNS), or both? In the CNS, is the cerebral cortex, basal ganglia, brainstem, cerebellum, or spinal cord responsible? Are the pain-sensitive meninges involved? In the PNS, could the disorder be located in peripheral nerves and, if so,

are motor or sensory nerves primarily affected, or is a lesion in the neuromuscular junction or muscle more likely?

The first clues to defining the anatomic area of involvement appear in the history, and the examination is then directed to confirm or rule out these impressions and to clarify uncertainties. A more detailed examination of a particular region of the CNS or PNS is often indicated. For example, the examination of a patient who presents with a history of ascending paresthesias and weakness should be directed toward deciding, among other things, if the lesion is in the spinal cord or peripheral nerves. Focal back pain, a spinal cord sensory level, and incontinence suggest a spinal cord origin, whereas a stocking-glove pattern of sensory loss suggests peripheral nerve disease; areflexia usually indicates peripheral neuropathy but may also be present with spinal shock in acute spinal cord disorders.

Deciding “where the lesion is” accomplishes the task of limiting the possible etiologies to a manageable, finite number. In addition, this strategy safeguards against making serious errors. Symptoms of recurrent vertigo, diplopia, and nystagmus should not trigger “multiple sclerosis” as an answer (etiology) but “brainstem” or “pons” (location); then a diagnosis of brainstem arteriovenous malformation will not be missed for lack of consideration. Similarly, the combination of optic neuritis and spastic ataxic paraparesis suggests optic nerve and spinal cord disease; multiple sclerosis (MS), CNS syphilis, and vitamin B_{12} deficiency are treatable disorders that can produce this syndrome. Once the question, “Where is the lesion?” is answered, then the question “What is the lesion?” can be addressed.

IDENTIFY THE PATHOPHYSIOLOGY

Clues to the pathophysiology of the disease process may also be present in the history. Primary neuronal (gray matter) disorders often present as early cognitive disturbances, movement disorders, or seizures, whereas white matter involvement produces “long tract” disorders of motor, sensory, visual, and cerebellar pathways. Progressive and symmetric symptoms often have a metabolic or degenerative origin; in such cases lesions are usually not sharply circumscribed. Thus, a patient with paraparesis and a clear spinal cord sensory level is unlikely to have vitamin B_{12} deficiency as the explanation. A Lhermitte symptom (electric shock–like sensations evoked by neck flexion) is due to ectopic impulse generation in white matter pathways and occurs with demyelination in the cervical spinal cord; among many possible causes, this symptom may indicate MS in a young adult or compressive cervical spondylosis in an older person. Symptoms that worsen after exposure to heat or exercise may indicate conduction block in demyelinated axons, as occurs in MS. A patient with recurrent episodes of diplopia and dysarthria associated with exercise or fatigue may have a disorder of neuromuscular transmission such as myasthenia gravis. Slowly advancing visual scotoma with luminous edges, termed *fortification spectra*, indicates spreading cortical depression, typically with migraine.

THE NEUROLOGIC HISTORY

Attention to the description of the symptoms experienced by the patient and substantiated by family members and others often permits an accurate localization and determination of the probable cause, even before the neurologic examination is performed. The history also helps focus the neurologic examination that follows. Each complaint should be pursued as far as possible to identify the location of the lesion, the likely underlying pathophysiology, and potential etiologies. For example, a patient complains of weakness of the right arm. What are the associated features? Does the patient have difficulty with brushing hair or reaching upward (proximal) or buttoning buttons or opening a twist-top bottle (distal)? Negative associations may also be crucial. A patient with a right hemiparesis without a language deficit likely has a lesion (internal capsule, brainstem, or spinal cord) different from that of a patient with a right hemiparesis and aphasia (left hemisphere). Other pertinent features of the history include the following:

1. *Temporal course of the illness*. It is important to determine the precise time of appearance and rate of progression of the symptoms experienced by the patient. The rapid onset of a neurologic complaint, occurring within seconds or minutes, usually indicates a vascular event, a seizure, or migraine. The onset of sensory symptoms located in one extremity that spread over a few seconds to adjacent portions of that extremity and then to the other regions of the body suggests a seizure. A similar but slower temporal march of symptoms accompanied by headache, nausea, or visual disturbance suggests migraine. Less well-localized symptoms that are either sudden or more gradual in onset point to the possibility of a transient ischemic attack (TIA). The presence of “positive” sensory symptoms (e.g., tingling or sensations that are difficult to describe) or involuntary motor movements suggests a seizure; in contrast, transient loss of function (negative symptoms) suggests a TIA. A stuttering onset where symptoms appear, stabilize, and then progress over hours or days also suggests cerebrovascular disease; an additional history of transient remission or regression indicates that the process is more likely due to ischemia rather than hemorrhage. A gradual evolution of symptoms over hours or days suggests a toxic, metabolic, infectious, or inflammatory process. Progressing symptoms associated with the systemic manifestations of fever, stiff neck, and altered level of consciousness imply an infectious process. Relapsing and remitting symptoms involving different levels of the nervous system suggest MS or other inflammatory processes. Slowly progressive symptoms without remissions are characteristic of neurodegenerative disorders, chronic infections, gradual intoxications, and neoplasms.

2. *Patients’ descriptions of the complaint*. The same words often mean different things to different patients. “Dizziness” may imply impending syncope, a sense of disequilibrium, or true spinning vertigo. “Numbness” may mean a complete loss of feeling, a positive sensation such as tingling, or even weakness. “Blurred vision” may be used to describe unilateral visual loss, as in transient monocular blindness, or diplopia. The interpretation of the true meaning of the words used by patients to describe symptoms obviously becomes even more complex when there are differences in primary languages and cultures.

3. *Corroboration of the history by others*. It is almost always helpful to obtain additional information from family, friends, or other observers to corroborate or expand the patient’s description. Memory loss, aphasia, loss of insight, intoxication, and other factors may impair the patient’s capacity to communicate

normally with the examiner or prevent openness about factors that have contributed to the illness. Episodes of loss of consciousness necessitate that details be sought from observers to ascertain precisely what has happened during the event.

4. *Family history*. Many neurologic disorders have an underlying genetic component. The presence of a Mendelian disorder, such as Huntington's disease or Charcot-Marie-Tooth neuropathy, is often obvious if family data are available. More detailed questions about family history are often necessary in polygenic disorders such as MS, migraine, and many types of epilepsy. It is important to elicit family history about all illnesses, in addition to neurologic and psychiatric disorders. A familial propensity to hypertension or heart disease is relevant in a patient who presents with a stroke. There are numerous inherited neurologic diseases that are associated with multisystem manifestations that may provide clues to the correct diagnosis (e.g., neurofibromatosis, Wilson's disease, mitochondrial disorders).

5. *Medical illnesses*. Many neurologic diseases occur in the context of systemic disorders. Diabetes mellitus, hypertension, and abnormalities of blood lipids predispose to cerebrovascular disease. A solitary mass lesion in the brain may be an abscess in a patient with valvular heart disease, a primary hemorrhage in a patient with a coagulopathy, a lymphoma or toxoplasmosis in a patient with AIDS, or a metastasis in a patient with underlying cancer. Patients with malignancy may also present with a neurologic paraneoplastic syndrome **(Chap. 90)** or complications from chemotherapy or radiotherapy. Marfan's syndrome and related collagen disorders predispose to dissection of the cranial arteries and aneurysmal subarachnoid hemorrhage; the latter may also occur with polycystic kidney disease. Various neurologic disorders occur with dysthyroid states or other endocrinopathies. It is especially important to look for the presence of systemic diseases in patients with peripheral neuropathy. Most patients with coma in a hospital setting have a metabolic, toxic, or infectious cause.

6. *Drug use and abuse and toxin exposure*. It is essential to inquire about the history of drug use, both prescribed and illicit. Sedatives, antidepressants, and other psychoactive medications are frequently associated with acute confusional states, especially in the elderly. Aminoglycoside antibiotics may exacerbate symptoms of weakness in patients with disorders of neuromuscular transmission, such as myasthenia gravis, and may cause dizziness secondary to ototoxicity. Vincristine and other antineoplastic drugs can cause peripheral neuropathy, and immunosuppressive agents such as cyclosporine can produce encephalopathy. Excessive vitamin ingestion can lead to disease; examples include vitamin A and pseudotumor cerebri or pyridoxine and peripheral neuropathy. Many patients are unaware that over-the-counter sleeping pills, cold preparations, and diet pills are actually drugs. Alcohol, the most prevalent neurotoxin, is often not recognized as such by patients, and other drugs of abuse such as cocaine and heroin can cause a wide range of neurologic abnormalities. A history of environmental or industrial exposure to neurotoxins may provide an essential clue; consultation with the patient's coworkers or employer may be required.

7. *Formulating an impression of the patient*. Use the opportunity while taking the history to form an impression of the patient. Is the information forthcoming, or does it take a circuitous course? Is there evidence of anxiety, depression, or hypochondriasis? Are there any clues to problems with language,

memory, insight, comportment, or behavior? The neurologic assessment begins as soon as the patient comes into the room and the first introduction is made.

THE NEUROLOGIC EXAMINATION

The neurologic examination is challenging and complex; it has many components and includes a number of skills that can be mastered only through repeated use of the same techniques on a large number of individuals with and without neurologic disease. Mastery of the complete neurologic examination is usually important only for physicians in neurology and associated specialties. However, knowledge of the basics of the examination, especially those components that are effective in screening for neurologic dysfunction, is essential for all clinicians, especially generalists.

There is no single, universally accepted sequence of the examination that must be followed, but most clinicians begin with assessment of mental status followed by the cranial nerves (CN), motor system, reflexes, sensory system, coordination, and gait. Whether the examination is basic or comprehensive, it is essential that it is performed in an orderly and systematic fashion to avoid errors and serious omissions. Thus, the best way to learn and gain expertise in the examination is to choose one's own approach and practice it frequently and do it in the same exact sequence each time.

The detailed description that follows describes the more commonly used parts of the neurologic examination, with a particular emphasis on the components that are considered most helpful for the assessment of common neurologic problems. Each section also includes a brief description of the minimal examination necessary to adequately screen for abnormalities in a patient who has no symptoms suggesting neurologic dysfunction. A screening examination done in this way can be completed in 3–5 min. *Video demonstrations of the neurologic screening examination (V6) and the detailed neurologic examination (V7) can be found in the Harrison's Video Collection included in this textbook*.

Several additional points about the examination are worth noting. First, in recording observations, it is important to describe what is found rather than to apply a poorly defined medical term (e.g., "patient groans to sternal rub" rather than "obtunded"). Second, subtle CNS abnormalities are best detected by carefully comparing a patient's performance on tasks that require simultaneous activation of both cerebral hemispheres (e.g., eliciting a pronator drift of an outstretched arm with the eyes closed; extinction on one side of bilaterally applied light touch, also with eyes closed; or decreased arm swing or a slight asymmetry when walking). Third, if the patient's complaint is brought on by some activity, reproduce the activity in the office. If the complaint is of dizziness when the head is turned in one direction, have the patient do this and also look for associated signs on examination (e.g., nystagmus or dysmetria). If pain occurs after walking two blocks, have the patient leave the office and walk this distance and immediately return, and repeat the relevant parts of the examination. Finally, the use of tests that are individually tailored to the patient's problem can be of value in assessing changes over time. Tests of walking a 7.5-m (25-ft) distance (normal, 5–6 s; note assistance, if any), repetitive finger or toe tapping (normal, 20–25 taps in 5 s), or handwriting are examples.

MENTAL STATUS EXAMINATION

The bare minimum: During the interview, look for difficulties with communication and determine whether the patient has recall and insight into recent and past events.

The mental status examination is under way as soon as the physician begins observing and speaking with the patient. If the history raises any concern for abnormalities of higher cortical function or if cognitive problems are observed during the interview, then detailed testing of the mental status is indicated. The patient's ability to understand the language used for the examination, cultural background, educational experience, sensory or motor problems, or comorbid conditions needs to be factored into the applicability of the tests and interpretation of results.

The Mini-Mental State Examination (MMSE) is a standardized screening examination of cognitive function that is extremely easy to administer and takes <10 min to complete **(Chap. 25)**. Using age-adjusted values for defining normal performance, the test is ~85% sensitive and 85% specific for making the diagnosis of dementia that is moderate or severe, especially in educated patients. When there is sufficient time available, the MMSE is one of the best methods for documenting the current mental status of the patient, and this is especially useful as a baseline assessment to which future scores of the MMSE can be compared.

Individual elements of the mental status examination can be subdivided into level of consciousness, orientation, speech and language, memory, fund of information, insight and judgment, abstract thought, and calculations.

Level of consciousness is the patient's relative state of awareness of the self and the environment, and ranges from fully awake to comatose. When the patient is not fully awake, the examiner should describe the responses to the minimum stimulus necessary to elicit a reaction, ranging from verbal commands to a brief, painful stimulus such as a squeeze of the trapezius muscle. Responses that are directed toward the stimulus and signify some degree of intact cerebral function (e.g., opening the eyes and looking at the examiner or reaching to push away a painful stimulus) must be distinguished from reflex responses of a spinal origin (e.g., triple flexion response—flexion at the ankle, knee, and hip in response to a painful stimulus to the foot).

Orientation is tested by asking the person to state his or her name, location, and time (day of the week and date); time is usually the first to be affected in a variety of conditions.

Speech is assessed by observing articulation, rate, rhythm, and prosody (i.e., the changes in pitch and accentuation of syllables and words).

Language is assessed by observing the content of the patient's verbal and written output, response to spoken commands, and ability to read. A typical testing sequence is to ask the patient to name successively more detailed components of clothing, a watch, or a pen; repeat the phrase "No ifs, ands, or buts"; follow a three-step, verbal command; write a sentence; and read and respond to a written command.

Memory should be analyzed according to three main time scales: (1) immediate memory is assessed by saying a list of three items and having the patient repeat the list immediately; (2) short-term memory is tested by asking the patient to recall the same three items 5 and 15 min later; and (3) long-term memory is evaluated by determining how well the patient is able to provide a coherent chronologic history of his or her illness or personal events.

Fund of information is assessed by asking questions about major historic or current events, with special attention to educational level and life experiences.

Abnormalities of *insight and judgment* are usually detected during the patient interview; a more detailed assessment can be elicited by asking the patient to describe how he or she would respond to situations having a variety of potential outcomes (e.g., “What would you do if you found a wallet on the sidewalk?”).

Abstract thought can be tested by asking the patient to describe similarities between various objects or concepts (e.g., apple and orange, desk and chair, poetry and sculpture) or to list items having the same attributes (e.g., a list of four-legged animals).

Calculation ability is assessed by having the patient carry out a computation that is appropriate to the patient’s age and education (e.g., serial subtraction of 7 from 100 or 3 from 20; or word problems involving simple arithmetic).

CRANIAL NERVE EXAMINATION

The bare minimum: Check the fundi, visual fields, pupil size and reactivity, extraocular movements, and facial movements.

The CN are best examined in numerical order, except for grouping together CN III, IV, and VI because of their similar function.

CN I (Olfactory)

Testing is often omitted unless there is suspicion for inferior frontal lobe disease (e.g., meningioma). With eyes closed, ask the patient to sniff a mild stimulus such as toothpaste or coffee and identify the odorant.

CN II (Optic)

Check visual acuity (with eyeglasses or contact lens correction) using a Snellen chart or similar tool. Test the visual fields by confrontation, i.e., by comparing the patient’s visual fields to your own. As a screening test, it is usually sufficient to examine the visual fields of both eyes simultaneously; individual eye fields should be tested if there is any reason to suspect a problem of vision by the history or other elements of the examination, or if the screening test reveals an abnormality. Face the patient at a distance of ~0.6–1.0 m (2–3 ft) and place your hands at the periphery of your visual fields in the plane that is equidistant between you and the patient. Instruct the patient to look directly at the center of your face and to indicate when and

where he or she sees one of your fingers moving. Beginning with the two inferior quadrants and then the two superior quadrants, move your index finger of the right hand, left hand, or both hands simultaneously and observe whether the patient detects the movements. A single small-amplitude movement of the finger is sufficient for a normal response. Focal perimetry and tangent screen examinations should be used to map out visual field defects fully or to search for subtle abnormalities. Optic fundi should be examined with an ophthalmoscope, and the color, size, and degree of swelling or elevation of the optic disc noted, as well as the color and texture of the retina. The retinal vessels should be checked for size, regularity, arteriovenous nicking at crossing points, hemorrhage, exudates, etc.

CN III, IV, VI (Oculomotor, Trochlear, Abducens)

Describe the size and shape of pupils and reaction to light and accommodation (i.e., as the eyes converge while following your finger as it moves toward the bridge of the nose). To check extraocular movements, ask the patient to keep his or her head still while tracking the movement of the tip of your finger. Move the target slowly in the horizontal and vertical planes; observe any paresis, nystagmus, or abnormalities of smooth pursuit (saccades, oculomotor ataxia, etc.). If necessary, the relative position of the two eyes, both in primary and multidirectional gaze, can be assessed by comparing the reflections of a bright light off both pupils. However, in practice it is typically more useful to determine whether the patient describes diplopia in any direction of gaze; true diplopia should almost always resolve with one eye closed. Horizontal nystagmus is best assessed at 45° and not at extreme lateral gaze (which is uncomfortable for the patient); the target must often be held at the lateral position for at least a few seconds to detect an abnormality.

CN V (Trigeminal)

Examine sensation within the three territories of the branches of the trigeminal nerve (ophthalmic, maxillary, and mandibular) on each side of the face. As with other parts of the sensory examination, testing of two sensory modalities derived from different anatomic pathways (e.g., light touch and temperature) is sufficient for a screening examination. Testing of other modalities, the corneal reflex, and the motor component of CN V (jaw clench—masseter muscle) is indicated when suggested by the history.

CN VII (Facial)

Look for facial asymmetry at rest and with spontaneous movements. Test eyebrow elevation, forehead wrinkling, eye closure, smiling, and cheek puff. Look in particular for differences in the lower versus upper facial muscles; weakness of the lower two-thirds of the face with preservation of the upper third suggests an upper motor neuron lesion, whereas weakness of an entire side suggests a lower motor neuron lesion.

CN VIII (Vestibulocochlear)

Check the patient's ability to hear a finger rub or whispered voice with each ear. Further testing for air versus mastoid bone conduction (Rinne) and lateralization of a 512-Hz tuning fork placed at the center of the forehead (Weber) should be done if an abnormality is detected by history or examination. Any suspected

problem should be followed up with formal audiometry. **For further discussion of assessing vestibular nerve function in the setting of dizziness, hearing loss, or coma, see Chaps. 19, 30, and 300, respectively.**

CN IX, X (Glossopharyngeal, Vagus)

Observe the position and symmetry of the palate and uvula at rest and with phonation ("aah"). The pharyngeal ("gag") reflex is evaluated by stimulating the posterior pharyngeal wall on each side with a sterile, blunt object (e.g., tongue blade), but the reflex is often absent in normal individuals.

CN XI (Spinal Accessory)

Check shoulder shrug (trapezius muscle) and head rotation to each side (sternocleidomastoid) against resistance.

CN XII (Hypoglossal)

Inspect the tongue for atrophy or fasciculations, position with protrusion, and strength when extended against the inner surface of the cheeks on each side.

MOTOR EXAMINATION

The bare minimum: Look for muscle atrophy and check extremity tone. Assess upper extremity strength by checking for pronator drift and strength of wrist or finger extensors. Assess lower extremity strength by checking strength of the toe extensors and having the patient walk normally and on heels and toes.

The motor examination includes observations of muscle appearance, tone, and strength. Although gait is in part a test of motor function, it is usually evaluated separately at the end of the examination.

Appearance

Inspect and palpate muscle groups under good light and with the patient in a comfortable and symmetric position. Check for muscle fasciculations, tenderness, and atrophy or hypertrophy. Involuntary movements may be present at rest (e.g., tics, myoclonus, choreoathetosis), during maintained posture (pill-rolling tremor of Parkinson's disease), or with voluntary movements (intention tremor of cerebellar disease or familial tremor).

Tone

Muscle tone is tested by measuring the resistance to passive movement of a relaxed limb. Patients often have difficulty relaxing during this procedure, so it is useful to distract the patient to minimize active movements. In the upper limbs, tone is assessed by rapid pronation and supination of the forearm and flexion and extension at the wrist. In the lower limbs, while the patient is supine the examiner's hands are placed behind the knees and rapidly raised; with normal tone, the ankles drag along the table surface for a variable distance

before rising, whereas increased tone results in an immediate lift of the heel off the surface. Decreased tone is most commonly due to lower motor neuron or peripheral nerve disorders. Increased tone may be evident as spasticity (resistance determined by the angle and velocity of motion; corticospinal tract disease), rigidity (similar resistance in all angles of motion; extrapyramidal disease), or paratonia (fluctuating changes in resistance; frontal lobe pathways or normal difficulty in relaxing). Cogwheel rigidity, in which passive motion elicits jerky interruptions in resistance, is seen in parkinsonism.

Strength

Testing for pronator drift is an extremely useful method for screening upper limb weakness. The patient is asked to hold both arms fully extended and parallel to the ground with eyes closed. This position should be maintained for ~10 s; any flexion at the elbow or fingers or pronation of the forearm, especially if asymmetric, is a sign of potential weakness. Muscle strength is further assessed by having the patient exert maximal effort for the particular muscle or muscle group being tested. It is important to isolate the muscles as much as possible, i.e., hold the limb so that only the muscles of interest are active. It is also helpful to palpate accessible muscles as they contract. Grading muscle strength and evaluating the patient's effort is an art that takes time and practice. Muscle strength is traditionally graded using the following scale:

0 = no movement

1 = flicker or trace of contraction but no associated movement at a joint

2 = movement with gravity eliminated

3 = movement against gravity but not against resistance

4– = movement against a mild degree of resistance

4 = movement against moderate resistance

4+ = movement against strong resistance

5 = full power

However, in many cases, it is more practical to use the following terms:

Paralysis = no movement
Severe weakness = movement with gravity eliminated
Moderate weakness = movement against gravity but not against mild resistance
Mild weakness = movement against moderate resistance
Full strength

Noting the pattern of weakness is as important as assessing the magnitude of weakness. Unilateral or bilateral weakness of the upper limb extensors and lower limb flexors ("pyramidal weakness") suggests a

lesion of the pyramidal tract, bilateral proximal weakness suggests myopathy, and bilateral distal weakness suggests peripheral neuropathy.

REFLEX EXAMINATION

The bare minimum: Check the biceps, patellar, and Achilles reflexes.

Muscle Stretch Reflexes

Those that are typically assessed include the biceps (**C5**, C6), brachioradialis (C5, **C6**), triceps (C6, **C7**), and sometimes finger flexor (**C8**, T1) reflexes in the upper limbs and the patellar or quadriceps (**L3**, L4) and Achilles (**S1**, S2) reflexes in the lower limbs. The patient should be relaxed and the muscle positioned midway between full contraction and extension. Reflexes may be enhanced by asking the patient to voluntarily contract other, distant muscle groups (Jendrassik maneuver). For example, upper limb reflexes may be reinforced by voluntary teeth-clenching, and the Achilles reflex by hooking the flexed fingers of the two hands together and attempting to pull them apart. For each reflex tested, the two sides should be tested sequentially, and it is important to determine the smallest stimulus required to elicit a reflex rather than the maximum response. Reflexes are graded according to the following scale:

0 = absent

1 = present but diminished

2 = normoactive

3 = exaggerated

4 = clonus

Cutaneous Reflexes

The plantar reflex is elicited by stroking, with a noxious stimulus such as a tongue blade, the lateral surface of the sole of the foot beginning near the heel and moving across the ball of the foot to the great toe. The normal reflex consists of plantar flexion of the toes. With upper motor neuron lesions above the S1 level of the spinal cord, a paradoxical extension of the toe is observed, associated with fanning and extension of the other toes (termed an *extensor plantar response*, or *Babinski sign*). However, despite its popularity, the reliability and validity of the Babinski sign for identifying upper motor neuron weakness is limited—it is far more useful to rely on tests of tone, strength, stretch reflexes, and coordination. Superficial abdominal reflexes are elicited by gently stroking the abdominal surface near the umbilicus in a diagonal fashion with a sharp object (e.g., the wooden end of a cotton-tipped swab) and observing the movement of the umbilicus. Normally, the umbilicus will pull toward the stimulated quadrant. With upper motor neuron lesions, these reflexes are absent. They are most helpful when there is preservation of the upper (spinal cord level T9) but not lower (T12) abdominal reflexes, indicating a spinal lesion between T9 and T12, or when the response is

asymmetric. Other useful cutaneous reflexes include the cremasteric (ipsilateral elevation of the testicle following stroking of the medial thigh; mediated by L1 and L2) and anal (contraction of the anal sphincter when the perianal skin is scratched; mediated by S2, S3, S4) reflexes. It is particularly important to test for these reflexes in any patient with suspected injury to the spinal cord or lumbosacral roots.

Primitive Reflexes

With disease of the frontal lobe pathways, several primitive reflexes not normally present in the adult may appear. The suck response is elicited by lightly touching with a tongue blade the center of the lips, and the root response the corner of the lips; the patient will move the lips to suck or root in the direction of the stimulus. The grasp reflex is elicited by touching the palm between the thumb and index finger with the examiner's fingers; a positive response is a forced grasp of the examiner's hand. In many instances, stroking the back of the hand will lead to its release. The palmomental response is contraction of the mentalis muscle (chin) ipsilateral to a scratch stimulus diagonally applied to the palm.

SENSORY EXAMINATION

The bare minimum: Ask whether the patient can feel light touch and the temperature of a cool object in each distal extremity. Check double simultaneous stimulation using light touch on the hands. Perform the Romberg maneuver.

Evaluating sensation is usually the most unreliable part of the examination because it is subjective and is difficult to quantify. In the compliant and discerning patient, the sensory examination can be extremely helpful for the precise localization of a lesion. With patients who are uncooperative or lack an understanding of the tests, it may be useless. The examination should be focused on the suspected lesion. For example, in spinal cord, spinal root, or peripheral nerve abnormalities, all major sensory modalities should be tested while looking for a pattern consistent with a spinal level and dermatomal or nerve distribution. In patients with lesions at or above the brainstem, screening the primary sensory modalities in the distal extremities along with tests of "cortical" sensation is usually sufficient.

The five primary sensory modalities—light touch, pain, temperature, vibration, and joint position—are tested in each limb. Light touch is assessed by stimulating the skin with single, very gentle touches of the examiner's finger or a wisp of cotton. Pain is tested using a new pin, and temperature is assessed using a metal object (e.g., tuning fork) that has been immersed in cold and warm water. Vibration is tested using a 128-Hz tuning fork applied to the distal phalanx of the great toe or index finger just below the nail bed. By placing a finger on the opposite side of the joint being tested, the examiner compares the patient's threshold of vibration perception with his or her own. For joint position testing, the examiner grasps the digit or limb laterally and distal to the joint being assessed; small 1- to 2-mm excursions can usually be sensed. The Romberg maneuver is primarily a test of proprioception. The patient is asked to stand with the feet as close together as necessary to maintain balance while the eyes are open, and the eyes are then closed. A loss of balance with the eyes closed is an abnormal response.

"Cortical" sensation is mediated by the parietal lobes and represents an integration of the primary sensory modalities; testing cortical sensation is only meaningful when primary sensation is intact. Double simultaneous stimulation is especially useful as a screening test for cortical function; with the patient's eyes closed, the examiner lightly touches one or both hands and asks the patient to identify the stimuli. With a parietal lobe lesion, the patient may be unable to identify the stimulus on the contralateral side when both hands are touched. Other modalities relying on the parietal cortex include the discrimination of two closely placed stimuli as separate (two-point discrimination), identification of an object by touch and manipulation alone (stereognosis), and the identification of numbers or letters written on the skin surface (graphesthesia).

COORDINATION EXAMINATION

The bare minimum: Observe the patient at rest and during spontaneous movements. Test rapid alternating movements of the hands and feet and finger to nose.

Coordination refers to the orchestration and fluidity of movements. Even simple acts require cooperation of agonist and antagonist muscles, maintenance of posture, and complex servomechanisms to control the rate and range of movements. Part of this integration relies on normal function of the cerebellar and basal ganglia systems. However, coordination also requires intact muscle strength and kinesthetic and proprioceptive information. Thus, if the examination has disclosed abnormalities of the motor or sensory systems, the patient's coordination should be assessed with these limitations in mind.

Rapid alternating movements in the upper limbs are tested separately on each side by having the patient make a fist, partially extend the index finger, and then tap the index finger on the distal thumb as quickly as possible. In the lower limb, the patient rapidly taps the foot against the floor or the examiner's hand. Finger-to-nose testing is primarily a test of cerebellar function; the patient is asked to touch his or her index finger repetitively to the nose and then to the examiner's outstretched finger, which moves with each repetition. A similar test in the lower extremity is to have the patient raise the leg and touch the examiner's finger with the great toe. Another cerebellar test in the lower limbs is the heel-knee-shin maneuver; in the supine position the patient is asked to slide the heel of each foot from the knee down the shin of the other leg. For all these movements, the accuracy, speed, and rhythm are noted.

GAIT EXAMINATION

The bare minimum: Observe the patient while walking normally, on the heels and toes, and along a straight line.

Watching the patient walk is the most important part of the neurologic examination. Normal gait requires that multiple systems—including strength, sensation, and coordination—function in a highly integrated fashion. Unexpected abnormalities may be detected that prompt the examiner to return in more detail to other aspects of the examination. The patient should be observed while walking and turning normally, walking on the heels, walking on the toes, and walking heel-to-toe along a straight line. The examination may reveal decreased arm swing on one side (corticospinal tract disease), a stooped posture and short-

stepped gait (parkinsonism), a broad-based unstable gait (ataxia), scissoring (spasticity), or a high-stepped, slapping gait (posterior column or peripheral nerve disease), or the patient may appear to be stuck in place (apraxia with frontal lobe disease).

NEUROLOGIC DIAGNOSIS

The clinical data obtained from the history and examination are interpreted to arrive at an anatomic localization that best explains the clinical findings **(Table 415-2)**, to narrow the list of diagnostic possibilities, and to select the laboratory tests most likely to be informative. The laboratory assessment may include (1) serum electrolytes; complete blood count; and renal, hepatic, endocrine, and immune studies; (2) cerebrospinal fluid examination; (3) focused neuroimaging studies **(Chap. 416)**; or (4) electrophysiologic studies. The anatomic localization, mode of onset and course of illness, other medical data, and laboratory findings are then integrated to establish an etiologic diagnosis.

TABLE 415-2

Findings Helpful for Localizations within the Nervous System

	Signs
Cerebrum	Abnormal mental status or cognitive impairment Seizures Unilateral weakness[a] and sensory abnormalities including head and limbs Visual field abnormalities Movement abnormalities (e.g., diffuse incoordination, tremor, chorea)
Brainstem	Isolated cranial nerve abnormalities (single or multiple) "Crossed" weakness[a] and sensory abnormalities of head and limbs, e.g., weakness of right face and left arm and leg
Spinal cord	Back pain or tenderness Weakness[a] and sensory abnormalities sparing the head Mixed upper and lower motor neuron findings Sensory level Sphincter dysfunction
Spinal roots	Radiating limb pain Weakness[b] or sensory abnormalities following root distribution **(see Figs. 22-2 and 22-3)** Loss of reflexes
Peripheral nerve	Mid or distal limb pain Weakness[b] or sensory abnormalities following nerve distribution **(see Figs. 22-2 and 22-3)** "Stocking or glove" distribution of sensory loss Loss of reflexes
Neuromuscular junction	Bilateral weakness including face (ptosis, diplopia, dysphagia) and proximal limbs Increasing weakness with exertion Sparing of sensation
Muscle	Bilateral proximal or distal weakness Sparing of sensation

[a]Weakness along with other abnormalities having an “upper motor neuron” pattern, i.e., spasticity, weakness of extensors > flexors in the upper extremity and flexors > extensors in the lower extremity, and hyperreflexia. [b]Weakness along with other abnormalities having a “lower motor neuron” pattern, i.e., flaccidity and hyporeflexia.

The neurologic examination may be normal even in patients with a serious neurologic disease, such as seizures, chronic meningitis, or a TIA. A comatose patient may arrive with no available history, and in such cases, the approach is as described in **Chap. 300**. In other patients, an inadequate history may be overcome by a succession of examinations from which the course of the illness can be inferred. In perplexing cases it is useful to remember that uncommon presentations of common diseases are more likely than rare etiologies. Thus, even in tertiary care settings, multiple strokes are usually due to emboli and not vasculitis, and dementia with myoclonus is usually Alzheimer’s disease and not a prion disorder or a paraneoplastic illness. Finally, the most important task of a primary care physician faced with a patient who has a new neurologic complaint is to assess the urgency of referral to a specialist. Here, the imperative is to rapidly identify patients likely to have nervous system infections, acute strokes, and spinal cord compression or other treatable mass lesions and arrange for immediate care.

FURTHER READING

Campbell WW: *DeJong’s Neurological Examination*, 7th ed. Philadelphia, Wolters Kluwer/Lippincott Williams & Wilkins, 2013.

O’Brien M: *Aids to the Examination of the Peripheral Nervous System*, 5th ed. Edinburgh, WB Saunders, 2010.

Ropper AH et al: *Adams and Victor’s Principles of Neurology*, 10th ed. New York, McGraw-Hill, 2014.

Chapter 416: Neuroimaging in Neurologic Disorders

William P. Dillon

INTRODUTION

Numerous noninvasive imaging options are available to clinicians evaluating patients with neurologic disorders. These include computed tomography (CT) and variations CT angiography (CTA), perfusion CT (pCT), and dual energy CT and magnetic resonance (MR) imaging (MRI) and variations MR angiography (MRA), MR vessel wall imaging, functional MRI (fMRI), MR spectroscopy (MRS), MR neurography (MRN), diffusion and diffusion tensor MR imaging, susceptibility-weighted MR imaging (SWI), arterial spin label imaging (ASL) and perfusion MRI (pMRI). Furthermore, a number of interventional neuroradiologic techniques have matured including catheter embolization, stent retrieval thrombolysis, aneurysm coiling and stenting, as well as numerous techniques for spine disorders, including CT myelography, fluoroscopy and CT-guided transforaminal and translaminar epidural and nerve root injections, radiofrequency ablation and blood patches. Multidetector CTA (MDCTA) and gadolinium-enhanced MRA techniques have reduced the need for catheter-based angiography, which is now reserved for patients in whom small-vessel detail is essential for diagnosis or for whom concurrent interventional therapy is planned (Table 416-1).

TABLE 416-1

Guidelines for the Use of CT, Ultrasound, and MRI

Condition	Recommended Technique
Hemorrhage	
Acute parenchymal	CT, MR
Subacute/chronic	MRI
Subarachnoid hemorrhage	CT, CTA, lumbar puncture → angiography
Aneurysm	Angiography > CTA, MRA
Ischemic infarction	
Hemorrhagic infarction	CT or MRI
Bland infarction	MRI with diffusion > CT, CTA, angiography
Carotid or vertebral dissection	MRI/MRA
Vertebral basilar insufficiency	CTA, MRI/MRA
Carotid stenosis	CTA, MRA > US
Suspected mass lesion	
Neoplasm, primary or metastatic	MRI + contrast
Infection/abscess	MRI + contrast
Immunosuppressed with focal findings	MRI + contrast
Vascular malformation	MRI ± angiography
White matter disorders	MRI
Demyelinating disease	MRI ± contrast
Dementia	MRI > CT

Condition	Recommended Technique
Trauma	
Acute trauma	CT
Shear injury/chronic hemorrhage	MRI + susceptibility-weighted imaging
Headache/migraine	CT/MRI
Seizure	
First time, no focal neurologic deficits	MRI > CT
Partial complex/refractory	MRI
Cranial neuropathy	MRI with contrast
Meningeal disease	MRI with contrast
Spine	
Low back pain	
No neurologic deficits	MRI or CT after >6 weeks
With focal deficits	MRI > CT
Spinal stenosis	MRI or CT
Cervical spondylosis	MRI, CT, CT myelography
Infection	MRI + contrast, CT
Myelopathy	MRI + contrast
Arteriovenous malformation	MRI + contrast, angiography

Abbreviations: CT, computed tomography; CTA, CT angiography; MRA, magnetic resonance angiography; MRI, magnetic resonance imaging.

In general, MRI is more sensitive than CT for the detection of lesions affecting the peripheral and central nervous system (CNS), particularly those of the spinal cord, cranial nerves, and posterior fossa structures. Diffusion MR, a sequence sensitive to the microscopic motion of water, is the most sensitive technique for detecting acute ischemic stroke of the brain or spinal cord, and it is also useful in the detection and characterization of encephalitis, abscess, Creutzfeldt-Jacob disease, cerebral tumors and acute demyelinating lesions. CT, however, is acquired quickly, making it a pragmatic choice for patients with acute changes in mental status, suspected hemorrhage, and acute intracranial or spinal trauma. CT is also more sensitive than MRI for visualizing fine osseous detail and is indicated in the initial imaging evaluation of conductive hearing loss as well as lesions affecting the skull base and calvarium. MR may, however, add important diagnostic information regarding bone marrow infiltrative processes that are difficult to detect on CT.

COMPUTED TOMOGRAPHY

TECHNIQUE

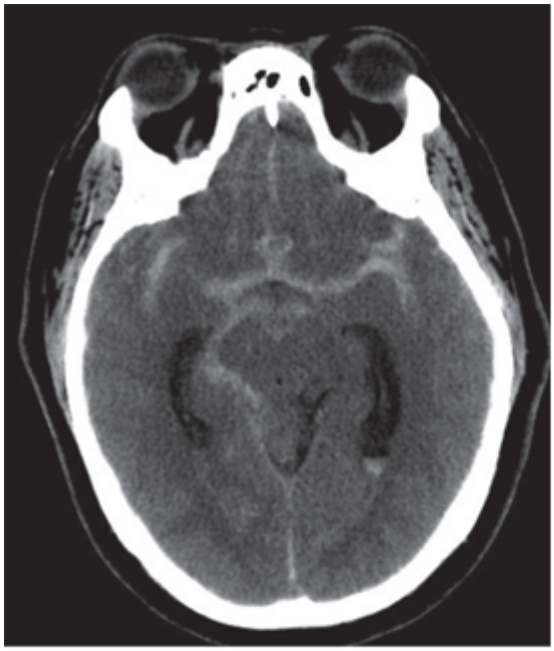
The CT image is a cross-sectional representation of anatomy created by a computer-generated analysis of the attenuation of x-ray beams passed through a section of the body. As the x-ray beam, collimated to the desired slice width, rotates around the patient, it passes through selected regions in the body. X-rays that are not attenuated by body structures are detected by sensitive x-ray detectors aligned 180° from the x-ray tube. A computer calculates a “back projection” image from the 360° x-ray attenuation profile. Greater x-ray attenuation (e.g., as caused by bone), results in areas of high “density” (whiter) on the scan, whereas soft tissue structures that have poor attenuation of x-rays, such as organs and air-filled cavities, are lower (blacker) in density. The resolution of an image depends on the radiation dose, the detector size, collimation (slice thickness), the field of view, and the matrix size of the display. A modern CT scanner is capable of obtaining sections as thin as 0.5–1 mm with 0.4-mm in-plane resolution at a speed of 0.3 s per rotation; complete studies of the brain can be completed in 1–10 s.

Multidetector CT (MDCT) is now standard. Single or multiple (from 4 to 320) solid-state detectors positioned opposite to the x-ray source result in multiple slices per revolution of the beam around the patient. In helical mode, the table moves continuously through the rotating x-ray beam, generating a continuous “helix” of information that can be reformatted into various slice thicknesses and planes. Advantages of MDCT include shorter scan times and thus reduced patient and organ motion, and the ability to acquire images dynamically during the infusion of intravenous contrast, the basis of CTA and CT perfusion **(Figs. 416-1B and C)**. CTA is displayed in three dimensions to yield angiogram-like images **(Figs. 416-1C, 416-2E and *F*, and see Fig. 420-3)**.

FIGURE 416-1

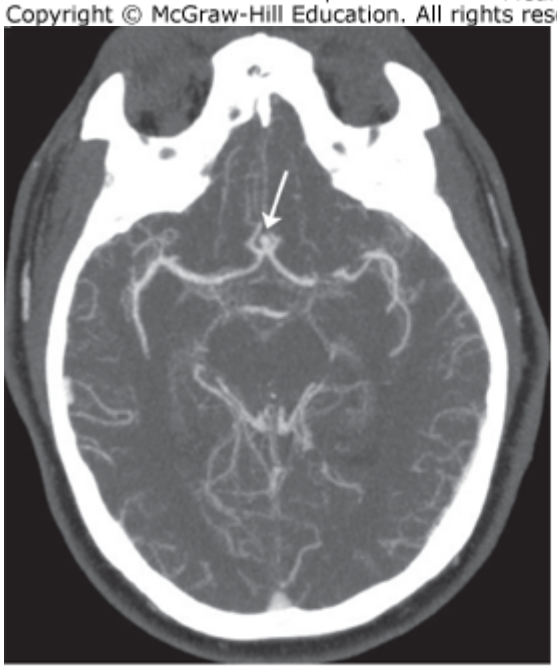
Computed tomography (CT) angiography (CTA) of ruptured anterior cerebral artery aneurysm in a patient presenting with acute headache. ***A.*** Noncontrast CT demonstrates subarachnoid hemorrhage and mild obstructive hydrocephalus. ***B.*** Axial maximum-intensity projection from CTA demonstrates enlargement of the anterior cerebral artery (*arrow*). ***C.*** Three-dimensional surface reconstruction using a workstation

confirms the anterior cerebral aneurysm and demonstrates its orientation and relationship to nearby vessels (*arrow*). CTA image is produced by 0.5- to 1-mm helical CT scans performed during a rapid bolus infusion of intravenous contrast medium.



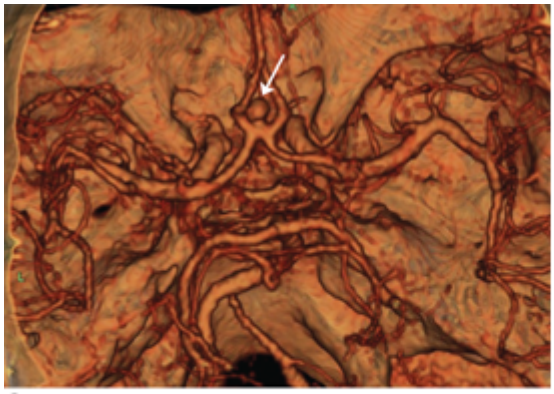
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

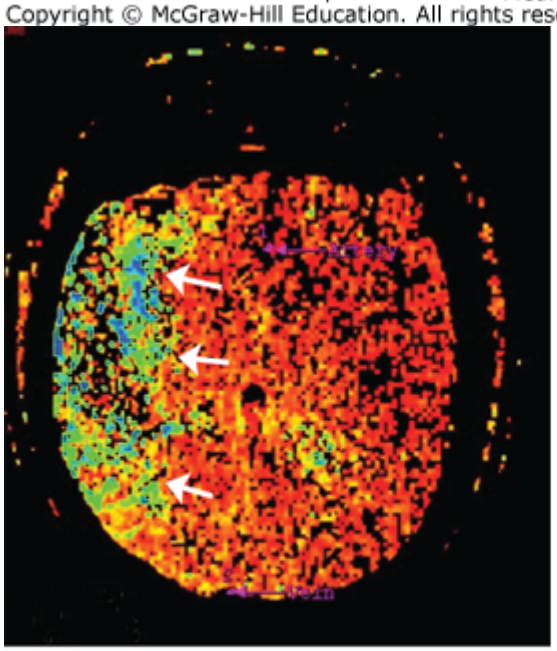
FIGURE 416-2

Acute left hemiparesis due to middle cerebral artery occlusion. ***A.*** Axial noncontrast computed tomography (CT) scan demonstrates high density within the right middle cerebral artery (*arrow*) associated with subtle low density involving the right putamen (*arrowheads*). ***B.*** Mean transit time CT perfusion parametric map indicating prolonged mean transit time involving the right middle cerebral territory (*arrows*). ***C.*** Cerebral blood volume (CBV) map shows reduced CBV involving an area within the defect shown in ***B***, indicating a high likelihood of infarction (*arrows*). ***D.*** Axial maximum-intensity projection from a CT angiography (CTA) study through the circle of Willis demonstrates an abrupt occlusion of the proximal right middle cerebral artery (*arrow*). ***E.*** Sagittal reformation through the right internal carotid artery demonstrates a low-density lipid-laden plaque (*arrowheads*) narrowing the lumen (*black arrow*). ***F.*** Three-dimensional surface-rendered CTA image demonstrates calcification and narrowing of the right internal carotid artery (*arrow*), consistent with atherosclerotic disease. ***G.*** Coronal maximum-intensity projection from magnetic resonance angiography shows right middle cerebral artery (MCA) occlusion (*arrow*). ***H.*** and ***I.*** Axial diffusion-weighted image (***H***) and apparent diffusion coefficient image (***I***) documents the presence of a right middle cerebral artery infarction.



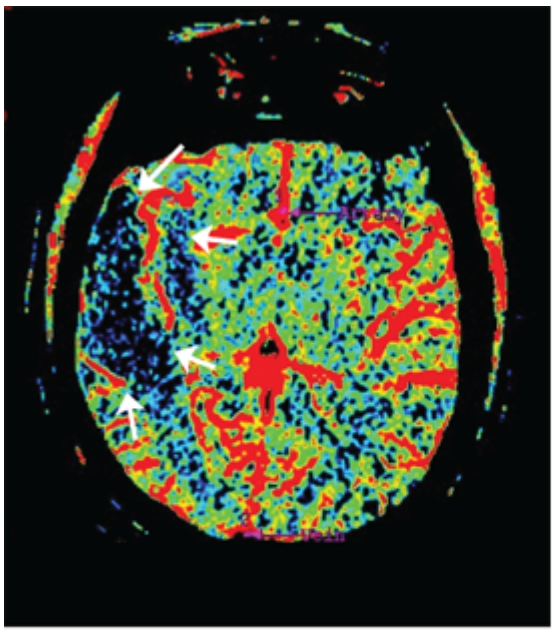
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



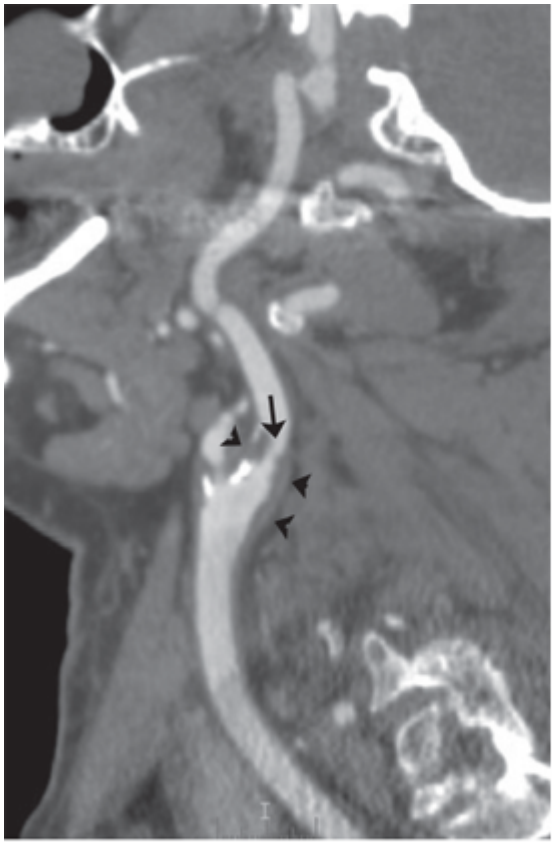
C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



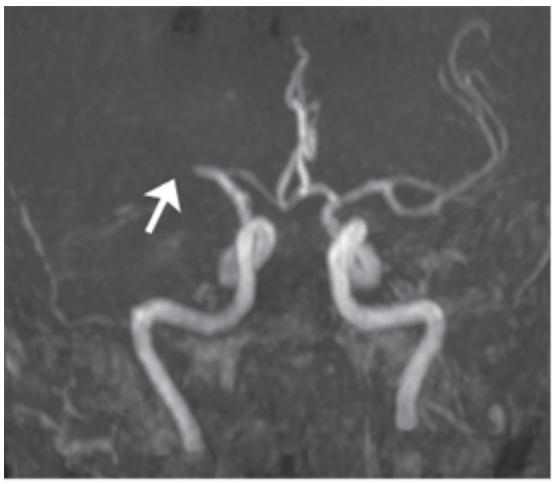
E

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



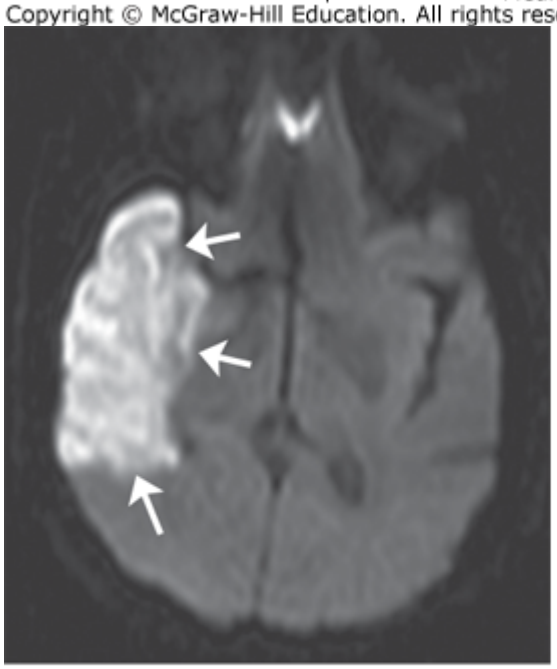
F

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



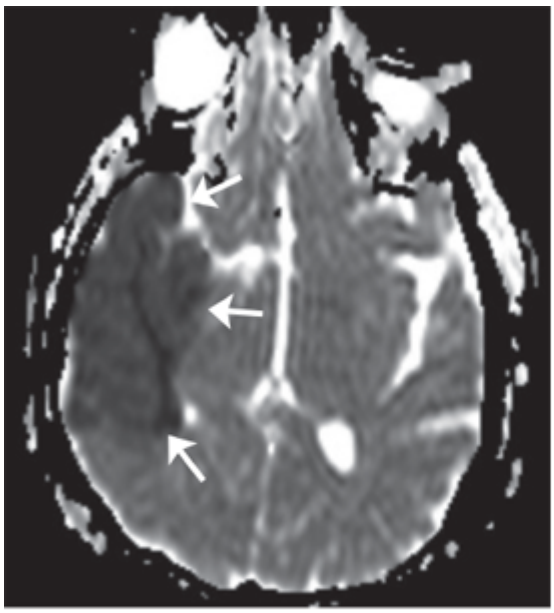
G

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



H

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



I

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Intravenous iodinated contrast is used to identify vascular structures and to detect defects in the blood-brain barrier (BBB) that are caused by tumors, infarcts, and infections. In the normal CNS, only vessels and structures lacking a BBB (e.g., the pituitary gland, choroid plexus, and dura) enhance after contrast administration. While helpful in characterizing mass lesions as well as essential for the acquisition of CTA studies, the decision to use contrast material should always be considered carefully as it carries a small risk of allergic reaction and adds additional expense.

INDICATIONS

CT is the primary study of choice in the evaluation of an acute change in mental status, focal neurologic findings, acute trauma to the brain and spine, suspected subarachnoid hemorrhage, and conductive hearing loss (Table 416-1). CT often is complementary to MR in the evaluation of the skull base, orbit, and osseous structures of the spine. In the spine, CT is useful in evaluating patients with osseous spinal stenosis and spondylosis, but MRI is often preferred in those with neurologic deficits. CT is often acquired following intrathecal contrast injection to evaluate for spinal and intracranial cerebrospinal fluid (CSF) fistula, as well as the spinal subarachnoid space (*CT myelography*) in failed back surgery syndromes.

COMPLICATIONS

CT is safe, fast, and reliable. Radiation exposure depends on the dose used but is normally between 2 and 5 mSv (millisievert) for a routine brain CT study. Care must be taken to reduce exposure when imaging children, who are typically best studied with MR. With the advent of MDCT, CTA, and CT perfusion, the benefit must be weighed against the increased radiation doses associated with these techniques. Advances in postprocessing software now permit acceptable diagnostic CT scans at 30–40% lower radiation doses.

The most frequent complications are those associated with use of intravenous contrast agents. While two broad categories of contrast media, ionic and nonionic, are in use, ionic agents have been largely replaced by safer nonionic compounds.

Contrast nephropathy is rare. It may result from hemodynamic changes, renal tubular obstruction and cell damage, or immunologic reactions to contrast agents. A rise in serum creatinine of at least 44 µmol/L (0.5 mg/dL) within 48 h of contrast administration is often used as a definition of contrast nephropathy, although there is no accepted definition and other causes of acute renal failure must be excluded. The prognosis is usually favorable, with serum creatinine levels returning to baseline within 1–2 weeks. Risk factors for contrast nephropathy include age (>80 years), preexisting renal disease (serum creatinine exceeding 2 mg/dL), solitary kidney, diabetes mellitus, dehydration, paraproteinemia, concurrent use of nephrotoxic medication or chemotherapeutic agents, and high contrast dose. Patients with diabetes and those with mild renal failure should be well hydrated prior to the administration of contrast agents; careful consideration should be given to alternative imaging techniques such as MRI, noncontrast CT, or ultrasound (US). Nonionic, low-osmolar media produce fewer abnormalities in renal blood flow and less endothelial cell damage but should still be used carefully in patients at risk for allergic reaction. Estimated glomerular filtration rate (eGFR) is a more reliable indicator of renal function compared to creatinine alone because it takes into account age, race, and sex. In one study, 15% of outpatients with a normal serum creatinine had an estimated creatinine clearance of ≤50 mL/min/1.73 m^2 (normal is ≥90 mL/min/1.73 m^2). The exact eGFR threshold, below which withholding intravenous contrast should be considered, is controversial. The risk of contrast nephropathy is minimal in patients with eGFR >30 mL/min/1.73 m^2; however, the majority of these patients will only have a temporary rise in creatinine. The risk of dialysis after receiving contrast significantly increases in patients with eGFR <30 mL/min/1.73 m^2. A creatinine of 1.6 in a 70-year-old, non-African-American male corresponds to an eGFR of ~45 mL/min/1.73 m^2. The American College of Radiology suggests using an eGFR of 30 mL/min/1.73 m^2 as a threshold below which iodinated contrast should not be given without serious consideration of the potential for contrast nephropathy. If contrast must be administered to a patient with an eGFR <30 mL/min/1.73 m^2, the patient should be well hydrated, and a reduction in the dose of contrast should be considered. Use of other agents such as bicarbonate and acetylcysteine may reduce the incidence of contrast nephropathy.

Suggested guidelines for creatinine testing prior to contrast administration:

If serum creatinine is not available, it should be performed IF the patient has ANY of the following risk factors:

Age >60

History of “kidney disease” as an adult, including tumor and transplant

Family history of kidney failure

Diabetes mellitus treated with insulin or other prescribed medications

Hypertension

Paraproteinemia syndromes or diseases (e.g., myeloma)

Collagen vascular disease (e.g., SLE, scleroderma, rheumatoid arthritis)

Solid organ transplant recipient

If creatinine testing is required, a creatinine level within the prior 6 weeks is sufficient in most clinical settings.

Allergy

Immediate reactions following intravenous contrast media occur through several mechanisms. The most severe reactions are related to allergic hypersensitivity (anaphylaxis) and range from mild hives to bronchospasm and death. The pathogenesis of allergic hypersensitivity reactions is thought to include the release of mediators such as histamine, antibody-antigen reactions, and complement activation. Severe allergic reactions occur in ~0.04% of patients receiving nonionic media, sixfold lower than with ionic media. Risk factors include a history of prior contrast reaction (fivefold increased likelihood), food and or drug allergies, and atopy (asthma and hay fever). The predictive value of specific allergies, such as those to shellfish, once thought important, actually is now recognized to be unreliable. Nonetheless, in patients with a history worrisome for potential allergic reaction, a noncontrast CT or MRI procedure should be considered as an alternative to contrast administration. If iodinated contrast is absolutely required, a nonionic agent should be used in conjunction with pretreatment with glucocorticoids and antihistamines **(Table 416-2)**; however, pretreatment does not guarantee safety. Patients with allergic reactions to iodinated contrast material do not usually react to gadolinium-based MR contrast material, although such reactions can occur. It would be wise to pretreat patients with a prior allergic history to MR contrast administration in a similar fashion. Subacute (>1 h after injection) reactions are frequent and probably related to T cell–mediated immune reactions. These are typically urticarial but can occasionally be more severe. Drug provocation and skin testing may be required to determine the culprit agent involved as well as determine a safe alternative.

TABLE 416-2

Guidelines for Premedication of Patients with Prior Contrast Allergy

12 h prior to examination:

Prednisone, 50 mg PO *or* methylprednisolone, 32 mg PO

2 h prior to examination:

Prednisone, 50 mg PO *or* methylprednisolone, 32 mg PO *and* cimetidine, 300 mg PO *or* ranitidine, 150 mg PO

Immediately prior to examination:

Benadryl, 50 mg IV (alternatively, can be given PO 2 h prior to exam)

Other side effects of CT scanning are rare but include a sensation of warmth throughout the body and a metallic taste during intravenous administration of iodinated contrast media. Extravasation of contrast media, although rare, can be painful and lead to compartment syndrome. When this occurs, consultation with plastic surgery is indicated. Patients with significant cardiac disease may be at increased risk for contrast reactions, and in these patients, limits to the volume and osmolality of the contrast media should be considered. Patients who may undergo systemic radioactive iodine therapy for thyroid disease or cancer should not receive iodinated contrast media if possible, because this will decrease the uptake of the radioisotope into the tumor or thyroid (see the *American College of Radiology Manual on Contrast Media*, Version 10.3, 2017; *https://www.acr.org/-/media/ACR/Files/Clinical-Resources/ContrastMedia.pdf.)*

MAGNETIC RESONANCE IMAGING

TECHNIQUE

MRI is a complex interaction between hydrogen protons in biologic tissues, a static magnetic field (the magnet), and energy in the form of radiofrequency (Rf) waves of a specific frequency introduced by coils placed next to the body part of interest. Images are made by computerized processing of resonance information received from protons in the body. Field strength of the magnet is directly related to signal-to-noise ratio. While 1.5 Tesla (T) and 3-T magnets are now widely available and have distinct advantages in the brain and musculoskeletal systems, even igher field magnets (7-T) and positron emission tomography (PET) MR machines promise increased resolution and anatomic-functional information on a variety of disorders. Spatial localization is achieved by magnetic gradients surrounding the main magnet, which impart slight changes in magnetic field throughout the imaging volume. Rf pulses transiently excite the energy state of the hydrogen protons in the body. Rf is administered at a frequency specific for the field strength of the magnet. The subsequent return to equilibrium energy state (*relaxation*) of the hydrogen protons results in a release of Rf energy (the *echo*), which is detected by the coils that delivered the Rf pulses. Fourier analysis is used to transform the echo into the information used to form an MR image. The MR image thus consists of a map of the distribution of hydrogen protons, with signal intensity imparted by both density of hydrogen protons and

differences in the relaxation times (see below) of hydrogen protons on different molecules. Although clinical MRI currently makes use of the ubiquitous hydrogen proton, research into sodium and carbon imaging and spectroscopy appears promising.

T1 and T2 Relaxation Times

The rate of return to equilibrium of perturbed protons is called the *relaxation rate*. The relaxation rate varies among normal and pathologic tissues. The relaxation rate of a hydrogen proton in a tissue is influenced by local interactions with surrounding molecules and atomic neighbors. Two relaxation rates, T1 and T2, influence the signal intensity of the image. The T1 relaxation time is the time, measured in milliseconds, for 63% of the hydrogen protons to return to their normal equilibrium state, whereas the T2 relaxation is the time for 63% of the protons to become dephased owing to interactions among nearby protons. The intensity and image contrast of the signal within various tissues can be modulated by altering acquisition parameters such as the interval between Rf pulses (TR) and the time between the Rf pulse and the signal reception (TE). T1-weighted (T1W) images are produced by keeping the TR and TE relatively short, whereas using longer TR and TE times produces T2-weighted (T2W) images. Fat and subacute hemorrhage have relatively shorter T1 relaxation rates and thus higher signal intensity than brain on T1W images. Structures containing more water, such as CSF and edema, have long T1 and T2 relaxation rates, resulting in relatively lower signal intensity on T1W images and higher signal intensity on T2W images **(Table 416-3)**. Gray matter contains 10–15% more water than white matter, which accounts for much of the intrinsic contrast between the two on MRI **(Fig. 416-4A)**. T2W images are more sensitive than T1W images to edema, demyelination, infarction, and chronic hemorrhage, whereas T1W imaging is more sensitive to subacute hemorrhage and fat-containing structures.

TABLE 416-3

Some Common Intensities on T1- and T2-Weighted MRI Sequences

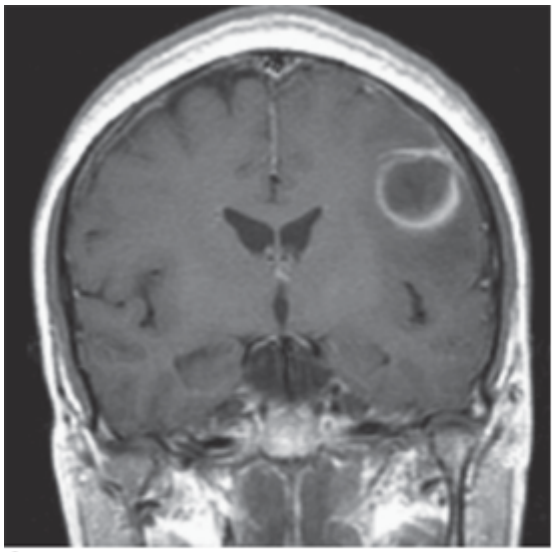
			Signal Intensity			
Image	**TR**	**TE**	**CSF**	**Fat**	**Brain**	**Edema**
T1W	Short	Short	Low	High	Low	Low
T2W	Long	Long	High	Low	High	High
FLAIR (T2)	Long	Long	Low	Medium	High	High

Abbreviations: CSF, cerebrospinal fluid; FLAIR, fluid-attenuated inversion recovery; TE, interval between radiofrequency pulse and signal reception; TR, interval between radiofrequency pulses; T1W and T2W, T1- and T2-weighted.

Many different MR pulse sequences exist, and each can be obtained in various planes **(Figs. 416-2, 416-3, and 416-4)**. The selection of a proper protocol that will best answer a clinical question depends on an accurate clinical history and indication for the examination. Fluid-attenuated inversion recovery (FLAIR) is a very useful pulse sequence that produces T2W images in which the normally high signal intensity of CSF is suppressed (Fig. 416-4B). FLAIR images are more sensitive than standard spin echo images for water-containing lesions or edema, especially those close to CSF filled cisterns and sulci. Diffusion weighted imaging is also routinely obtained in most brain protocols. This sequence interrogates the microscopic motion of water, which is restricted in areas of infarction, abscess, and some tumors. SWI is very sensitive to alterations in local magnetic field generated by blood, calcium, and air. SWI is routinely obtained and helps detects microhemorrhages, such as is typical of amyloid, hemorrhagic metastases, traumatic brain injury, and thrombotic states **(Fig. 416-5C)**. MR images can be generated in any plane without changing the patient's position. Each sequence, however, must be obtained separately and takes 1–10 min on average to complete. Three-dimensional volumetric imaging is also possible with MRI, resulting in a three-dimensional volume of data that can be reformatted in any orientation to highlight certain disease processes.

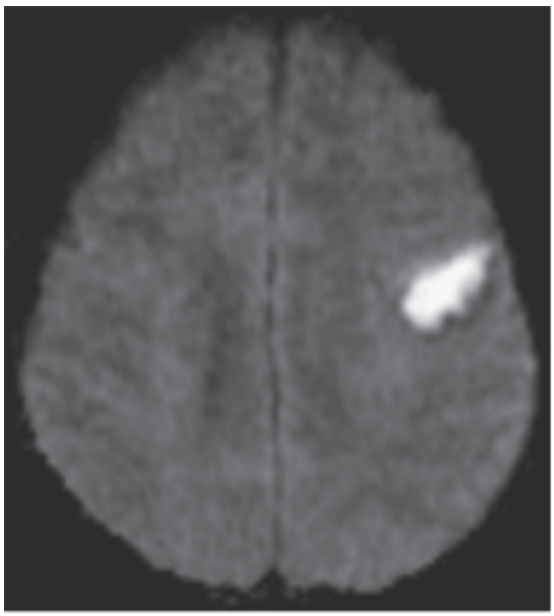
FIGURE 416-3

Cerebral abscess in a patient with fever and a right hemiparesis. ***A.*** Coronal postcontrast T1-weighted image demonstrates a ring-enhancing mass in the left frontal lobe. ***B.*** Axial diffusion-weighted image demonstrates restricted diffusion (high signal intensity) within the lesion, which in this setting is highly suggestive of cerebral abscess.



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

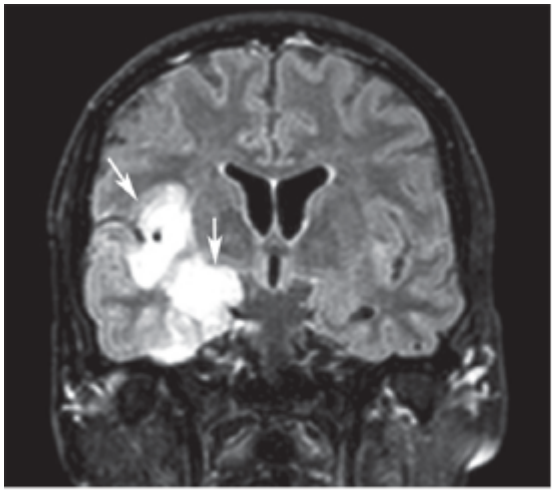


B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

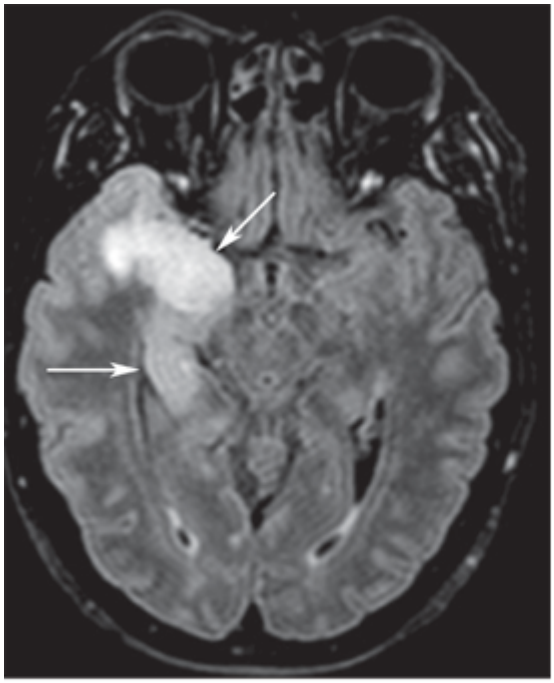
FIGURE 416-4

Herpes simplex encephalitis in a patient presenting with altered mental status and fever. ***A.*** and ***B.*** Coronal (***A***) and axial (***B***) T2-weighted fluid-attenuated inversion recovery images demonstrate expansion and high signal intensity involving the right medial temporal lobe and insular cortex (*arrows*). ***C.*** Coronal diffusion-weighted image demonstrates high signal intensity indicating restricted diffusion involving the right medial temporal lobe and hippocampus (*arrows*) as well as subtle involvement of the left inferior temporal lobe (*arrowhead*). This is most consistent with neuronal death and can be seen in acute infarction as well as encephalitis and other inflammatory conditions. The suspected diagnosis of herpes simplex encephalitis was confirmed by cerebrospinal fluid polymerase chain reaction analysis.



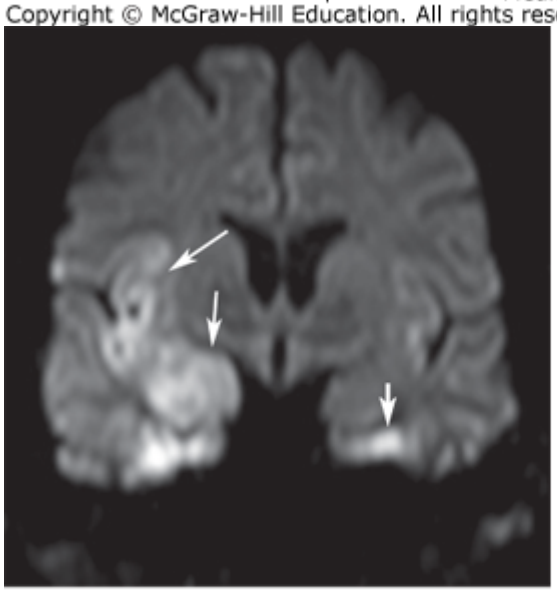
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

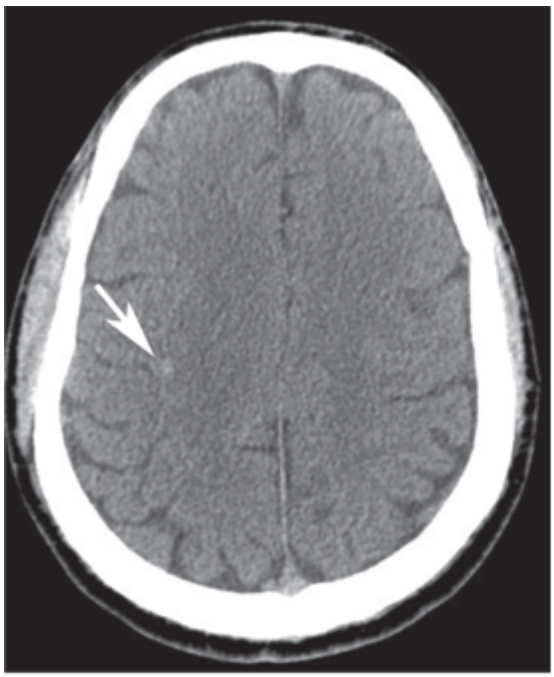


C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

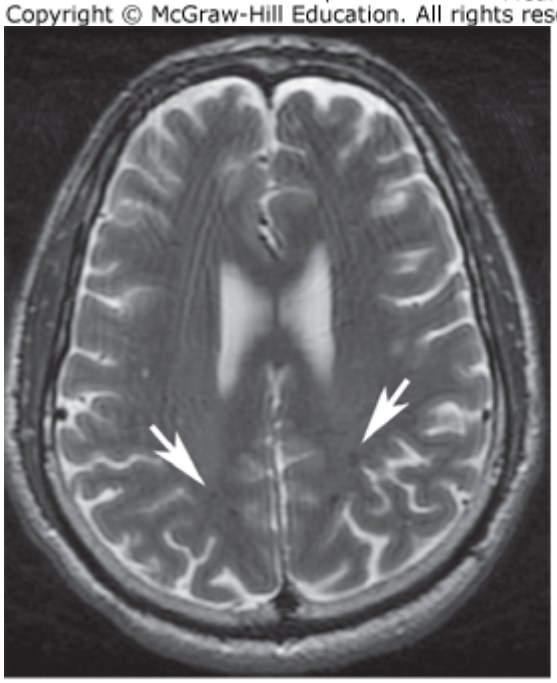
FIGURE 416-5

Susceptibility-weighted imaging in a patient with familial cavernous malformations. ***A.*** Noncontrast computed tomography scan shows one hyperdense lesion in the right hemisphere (*arrow*). ***B.*** T2-weighted fast spin echo image shows subtle low-intensity lesions (*arrows*). ***C.*** Susceptibility-weighted image shows numerous low-intensity lesions consistent with hemosiderin-laden cavernous malformations (*arrow*).



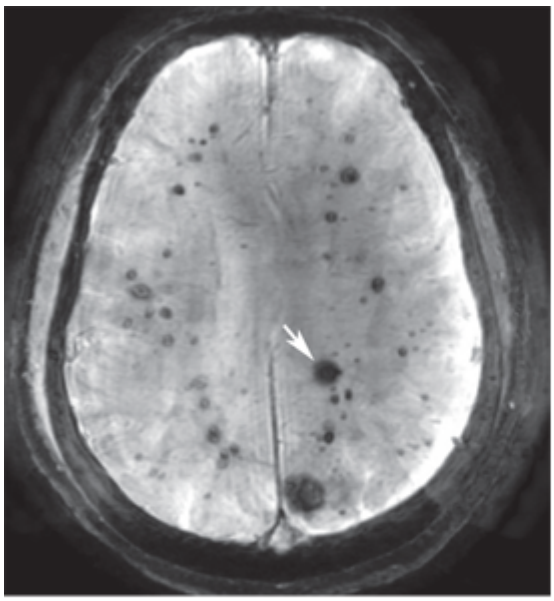
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MR Contrast Material

The heavy-metal element gadolinium forms the basis of all currently approved intravenous MR contrast agents. Gadolinium is a paramagnetic substance that reduces the T1 and T2 relaxation times of nearby water protons, resulting in a high signal on T1W images and a low signal on T2W images (the latter requires a sufficient local concentration, usually in the form of an intravenous bolus). Unlike iodinated contrast agents, the effect of MR contrast agents depends on the presence of local hydrogen protons on which it must act to achieve the desired effect. There are nine different gadolinium agents approved in the United States for use with MRI. These differ according to the attached chelated moiety, which also affects the strength of chelation of the otherwise toxic gadolinium element. The chelating carrier molecule for gadolinium can be classified by whether it is macrocyclic or has linear geometry and whether it is ionic or nonionic. Macrocyclic ligands (Group 2 agents) are considered more stable as the gadolinium ion is “caged” in the cavity of the ligand, and thus the rate of dissociation of gadolinium is slower compared to linear ligands (Group 1 agents). Most agents are excreted by the renal system.

BRAIN ACCUMULATION OF GADOLINIUM

It recently has become evident that gadolinium accumulates in the dentate nuclei and globus pallidus of the brain after serial administration of some linear Group 1 gadolinium agents. This has not been demonstrated for Group 2 macrocyclic agents. To date there is no clinical effect of this deposition that has been detected.

ALLERGIC HYPERSENSITIVITY

Gadolinium-DTPA (diethylenetriaminepentaacetic acid) does not normally cross the intact BBB immediately but will enhance lesions lacking a BBB (Fig. 416-3A) as well as areas of the brain that normally are devoid of the BBB (pituitary, dura, choroid plexus). However, gadolinium contrast slowly crosses an intact BBB over time and especially in the setting of reduced renal clearance or inflamed meninges. The agents are generally well tolerated; overall adverse events after injection range from 0.07 to 2.4%. True allergic reactions are rare

(0.004–0.7%) but have been reported. Severe life-threatening reactions are exceedingly rare; in one report, only 55 reactions out of 20 million doses occurred. However, the adverse reaction rate in patients with a prior history of reaction to gadolinium is eight times higher than normal. Other risk factors include atopy or asthma (3.7%). There is no cross reactivity between different classes of contrast media; a prior reaction to gadolinium-based contrast does not predict a future reaction to iodinated contrast medium, or vice versa, more than any other unrelated allergy. Gadolinium contrast material can be administered safely to children as well as adults, although these agents are generally avoided in those aged <6 months.

NEPHROTOXICITY

Contrast-induced renal failure does not occur with gadolinium agents. A rare complication, nephrogenic systemic fibrosis (NSF), has occurred in patients with severe renal insufficiency who have been exposed to linear (Group 1 and 3) gadolinium contrast agents. The onset of NSF has been reported between 5 and 75 days following exposure; histologic features include thickened collagen bundles with surrounding clefts, mucin deposition, and increased numbers of fibrocytes and elastic fibers in skin. In addition to dermatologic symptoms, other manifestations include widespread fibrosis of the skeletal muscle, bone, lungs, pleura, pericardium, myocardium, kidney, muscle, bone, testes, and dura. The American College of Radiology recommends that a glomerular filtration rate (GFR) assessment be obtained within 6 weeks prior to elective gadolinium-based MR contrast agent administration in patients with:

1. A history of renal disease (including solitary kidney, renal transplant, renal tumor)
2. Age >60 years
3. History of hypertension
4. History of diabetes
5. History of severe hepatic disease, liver transplant, or pending liver transplant; for these patients, it is recommended that the patient's GFR assessment be nearly contemporaneous with the MR examination.

The incidence of NSF in patients with severe renal dysfunction (GFR <30) varies from 0.19 to 4%. Other risk factors for NSF include acute kidney injury, the use of non-macrocyclic agents, and repeated or high-dose exposure to gadolinium. The American College of Radiology Committee on Drugs and Contrast Media considers the risk of NSF among patients exposed to standard or lower doses of Group 2 gadolinium agents (macrocyclic agents) is sufficiently low or possibly non-existent that the assessment of renal function is optimal prior to administration. However, patients receiving any Group 1 (linear) or 3 gadolinium-containing agent should be considered at risk of NSF if they are on dialysis (of any form); have severe or end-stage chronic renal disease (eGFR <30 mL/min/1.73 m^2) without dialysis; eGFR of 30–40 mL/min/1.73 m^2 without dialysis (as the GFR may fluctuate); or have acute renal insufficiency. The use of gadolinium in young children and infants is discouraged due to the unknown risks and their immature renal systems.

COMPLICATIONS AND CONTRAINDICATIONS

From the patient's perspective, an MRI examination can be intimidating, and a higher level of cooperation is required than with CT. The patient lies on a table that is moved into a long, narrow gap within the magnet. Approximately 5% of the population experiences severe claustrophobia in the MR environment. This can be reduced by mild sedation but remains a problem for some. Because it takes between 3 and 10 min per sequence, movement of the patient during an MR examination distorts all of the images; therefore, uncooperative patients should either be sedated for the MR study or scanned with CT. Generally, children aged <8 years usually require conscious sedation in order to complete the MR examination without motion degradation.

MRI is considered safe for patients, even at very high field strengths. Serious injuries have been caused, however, by attraction of ferromagnetic objects into the magnet, which act as missiles if brought too close to the magnet. Likewise, ferromagnetic implants, such as aneurysm clips, may torque within the magnet, causing damage to vessels and even death. Metallic foreign bodies in the eye have moved and caused intraocular hemorrhage; screening for ocular metallic fragments is indicated in those with a history of metal work or ocular metallic foreign bodies. Implanted cardiac pacemakers are generally a contraindication to MRI owing to the risk of induced arrhythmias; however, some newer pacemakers have been shown to be safe. All health care personnel and patients must be screened and educated thoroughly to prevent such disasters because the magnet is always "on." **Table 416-4** lists common contraindications for MRI.

TABLE 416-4

Common Contraindications to Magnetic Resonance Imaging

Cardiac pacemaker or permanent pacemaker leads
Internal defibrillatory device
Cochlear prostheses
Bone growth stimulators
Spinal cord stimulators
Electronic infusion devices
Intracranial aneurysm clips (some but not all)
Ocular implants (some) or ocular metallic foreign body
McGee stapedectomy piston prosthesis
Duraphase penile implant
Swan-Ganz catheter
Magnetic stoma plugs
Magnetic dental implants
Magnetic sphincters
Ferromagnetic inferior vena cava filters, coils, stents—safe 6 weeks after implantation
Tattooed eyeliner (contains ferromagnetic material and may irritate eyes)

Note: See also *http://www.mrisafety.com*.

MAGNETIC RESONANCE ANGIOGRAPHY

MR angiography is a general term describing several MR techniques that result in vascular-weighted images. Non contrast MRA provides a flow map rather than the anatomic map shown by conventional angiography. On routine spin echo MR sequences, moving protons (e.g., flowing blood, CSF) exhibit complex MR signals that range from high- to low-signal intensity relative to background stationary tissue. Fast-flowing blood returns no signal (flow void) on routine T1W or T2W spin echo MR images. Slower-flowing blood, as occurs in veins or distal to arterial stenosis, may appear high in signal. However, using special pulse sequences called *gradient echo sequences*, it is possible to increase the signal intensity of moving protons in contrast to the low signal background intensity of stationary tissue. This creates angiography-like images, which can be manipulated in three dimensions to highlight vascular anatomy and relationships.

So-called time-of-flight (TOF) MRA relies on the suppression of nonmoving tissue to provide a low-intensity background for the high signal intensity of flowing blood entering the section; arterial or venous structures may be highlighted. A typical TOF MRA sequence results in a series of contiguous, thin MR sections (0.6–0.9 mm thick), which can be viewed as a stack and manipulated to create an angiographic image data set that

can be reformatted and viewed in various planes and angles, much like that seen with conventional angiography **(Fig. 416-2G)**.

Phase-contrast MRA has a longer acquisition time than TOF MRA, but in addition to providing anatomic information similar to that of TOF imaging, it can be used to reveal the velocity and direction of blood flow in a given vessel.

MRA is often acquired during infusion of contrast material. Advantages include faster imaging times (1–2 min vs 10 min), fewer flow-related artifacts, and 4D temporal imaging resulting in arterial and venous phases. Recently, contrast-enhanced MRA has become the standard for extracranial vascular MRA. This technique entails rapid imaging using coronal three-dimensional TOF sequences during a bolus infusion of gadolinium contrast agent.

MRA has lower spatial resolution compared with conventional film-based angiography, and therefore the detection of small-vessel abnormalities, such as vasculitis and distal vasospasm, is problematic. MRA is also less sensitive to slowly flowing blood and thus may not reliably differentiate complete from near-complete occlusions. Motion, either by the patient or by anatomic structures, may distort the MRA images, creating artifacts. These limitations notwithstanding, MRA has proved useful in evaluation of the extracranial carotid and vertebral circulation as well as of larger-caliber intracranial arteries and dural sinuses. It has also proved useful in the noninvasive detection of intracranial aneurysms and vascular malformations.

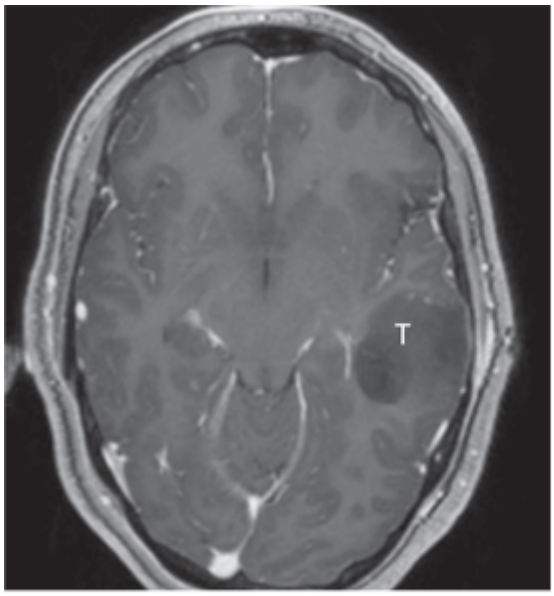
ECHO-PLANAR MRI

Recent improvements in gradients, software, and high-speed computer processors now permit extremely rapid MRI of the brain. With echo-planar MRI (EPI), fast gradients are switched on and off at high speeds to create the information used to form an image. With EPI, all of the information required for processing an image is accumulated in milliseconds, and the information for the entire brain can be obtained in <1–2 min, depending on the degree of resolution required or desired. Fast MRI reduces patient and organ motion and is the basis of perfusion imaging during contrast infusion and kinematic motion studies. EPI is also the sequence used to obtain diffusion imaging and tractography, as well as fMRI and arterial spin-labeled studies (**Figs. 416-2H**, 416-3, **416-4C,** and 416-6; and **Fig. 419-13**).

FIGURE 416-6

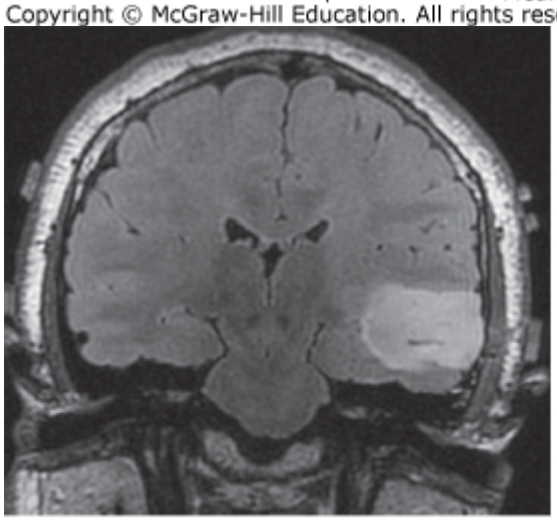
Diffusion tractography in cerebral glioma. Associative and descending pathways in a healthy subject (***A***) and in a patient with parietal lobe glioblastoma (***B***) presenting with a language deficit: the mass causes a disruption of the arcuate-SLF complex, in particular of its anterior portion (SLF III). Also shown are bilateral optic tract and left optic radiation pathways in a healthy subject (***C***) and in a patient with left occipital grade II oligoastrocytoma (***D***): the mass causes a disruption of the left optic radiation. *Shown in neurologic orientation, i.e., the left brain appears on the left side of the image.* AF, long segment of the arcuate fascicle; CST, corticospinal tract; IFOF: inferior fronto-occipital fascicle; ILF, inferior longitudinal fascicle; SLF III, superior longitudinal fascicle III or anterior segment of the arcuate fascicle; SLF-tp, temporo-parietal portion

of the superior longitudinal fascicle or posterior segment of the arcuate fascicle; T, tumor; UF, uncinated fascicle. *(Part D courtesy of Eduardo Caverzasi and Roland Henry.)*



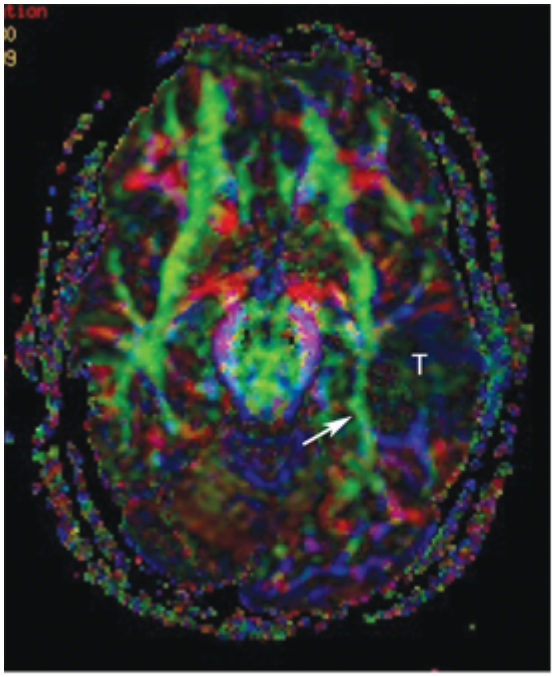
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



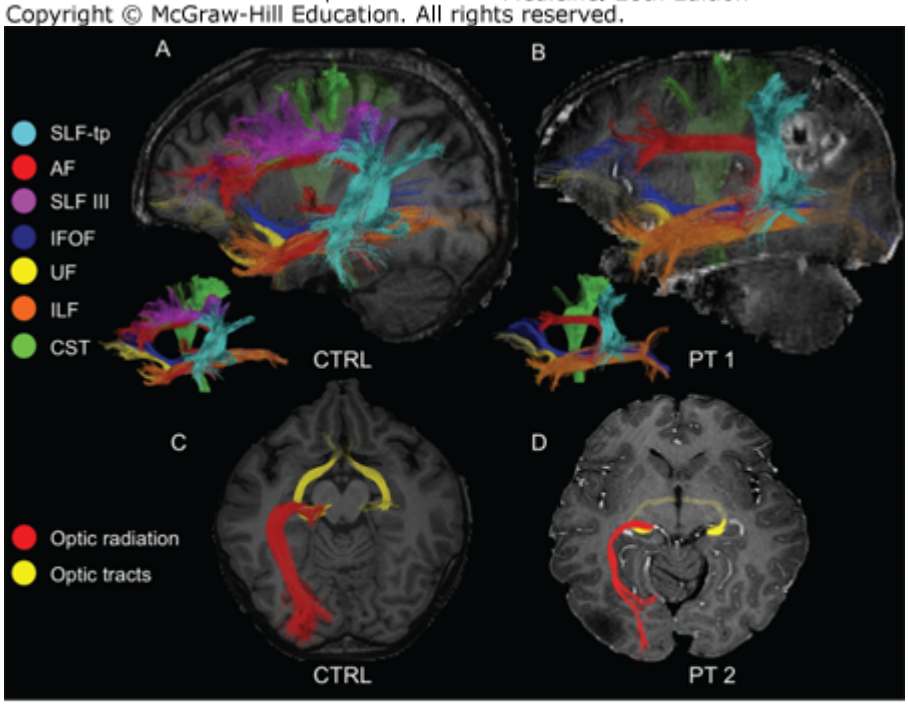
B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Perfusion and diffusion imaging are EPI techniques that are useful in early detection of ischemic injury of the brain and may be useful together to demonstrate infarcted tissue as well as ischemic but potentially viable tissue at risk of infarction (e.g., the ischemic penumbra). Diffusion-weighted imaging (DWI) assesses microscopic motion of water; abnormal restriction of motion appears as relative high-signal intensity on diffusion-weighted images. Infarcted tissue reduces the water motion within cells and in the interstitial tissues, resulting in high signal on DWI. DWI is the most sensitive technique for detection of acute cerebral

infarction of <7 days in duration (Fig. 416-2H). It is also quite sensitive for detecting dying or dead brain tissue secondary to encephalitis, as well as abscess formation (Fig. 416-3B).

Perfusion MRI can be performed by the acquisition of fast echo planar gradient images during a rapid intravenous bolus of gadolinium contrast material or by arterial spin labeling (ASL) without the use of contrast material. With contrast perfusion imaging, relative cerebral blood volume, mean transit time, and cerebral blood flow maps are derived. Delay in mean transit time and reduction in cerebral blood volume and cerebral blood flow are typical of infarction. In the setting of reduced blood flow, a prolonged mean transit time of contrast but normal or elevated cerebral blood volume may indicate tissue supplied by collateral flow that is at risk of infarction. Perfusion MRI imaging can also be used in the assessment of brain tumors to differentiate intraaxial primary tumors, whose BBB is relatively intact, from extraaxial tumors or metastases, which demonstrate a relatively more permeable BBB.

Diffusion tensor imaging is derived from diffusion MRI imaging sequences. This technique assesses the direction and integrity of white matter architecture. It has proven valuable in preoperative assessment of subcortical white matter tract anatomy prior to brain tumor surgery, as well as determining normal and abnormal white matter architecture in congenital and acquired pathologies (Fig. 416-6).

fMRI of the brain is an EPI technique that localizes regions of activity in the brain following task activation or at rest (so-called resting state fMRI). Neuronal activity elicits a slight increase in the delivery of oxygenated blood flow to a specific region of activated brain. This results in an alteration in the balance of oxyhemoglobin and deoxyhemoglobin, which yields a 2–3% increase in signal intensity within veins and local capillaries. Currently, preoperative somatosensory and auditory cortex localization is possible, and methods to assess motor and language function are in development. This technique has proved useful to neuroscientists interested in interrogating the localization of certain brain functions.

ARTERIAL SPIN LABELING

ASL is a quantitative noninvasive MR technique that measures cerebral blood flow. Blood traversing in the neck is labeled by an MR pulse and then imaged in the brain after a short (2 second) delay. The signal in the brain is reflective of blood flow. ASL is an especially important technique for patients with kidney failure and for pediatric patients in whom the use of radioactive tracers or exogenous contrast agents is contraindicated. Increased cerebral flow is more easily identified than slow flow, which can be sometimes difficult to quantify. This technique has also been useful in detecting arterial venous shunting in arteriovenous malformations and arteriovenous fistulas, as well as increased blood flow in brain tumors, and patients post TIA, post-ictal, or post migrane.

MAGNETIC RESONANCE NEUROGRAPHY

MRN is a T2W MR technique that shows promise in detecting increased signal in irritated, inflamed, or infiltrated peripheral nerves. Images are obtained with fat-suppressed fast spin echo imaging or short

inversion recovery sequences. Irritated or infiltrated nerves will demonstrate high signal on T2W imaging. This is indicated in patients with radiculopathy whose conventional MR studies of the spine are normal, or in those suspected of peripheral nerve entrapment or trauma.

POSITRON EMISSION TOMOGRAPHY

PET relies on the detection of positrons emitted during the decay of a radionuclide that has been injected into a patient. The most frequently used moiety is 2-[^{18}F]fluoro-2-deoxy-D-glucose (FDG), which is an analogue of glucose and is taken up by cells competitively with 2-deoxyglucose. Multiple images of glucose uptake activity are formed after 45–60 min. Images reveal differences in regional glucose activity among normal and pathologic brain structures. FDG-PET is used primarily for the detection of extracranial metastatic disease; however, a lower activity of FDG in the parietal lobes is associated with Alzheimer's disease, a finding that may simply reflect atrophy that occurs in the later stages of the disease. Combination PET-CT scanners, in which both CT and PET are obtained at one sitting, have largely replaced PET scans alone for most clinical indications. MR-PET scanners have also been developed and may prove useful for imaging the brain and other organs without the radiation exposure of CT. More recent PET ligand developments include beta-amyloid tracers, such as Pittsburgh compound B (PIB) and 18-F AV-45 (florbetapir), and tau PET tracers, such as 18F-T807 and T808. Studies have shown an increased percentage of amyloid deposition in patients with Alzheimer's disease compared with mild cognitive impairment and healthy controls; however, up to 25% of cognitively "normal" patients show abnormalities on amyloid PET imaging. This may either reflect subclinical disease processes or variation of normal. Tau imaging may be more specific for Alzheimer's disease, and clinical studies are under way.

MYELOGRAPHY

TECHNIQUE

Myelography involves the intrathecal instillation of specially formulated water-soluble iodinated contrast medium into the lumbar or cervical subarachnoid space. CT scanning is typically performed after myelography (*CT myelography*) to better demonstrate the spinal cord and roots, which appear as filling defects in the opacified subarachnoid space. *Low-dose CT myelography*, in which CT is performed after the subarachnoid injection of a small amount of relatively dilute contrast material, has replaced conventional myelography for many indications, thereby reducing exposure to radiation and contrast media. CT is obtained at a slice thickness ~ 2.5 mm and reconstructed at 0.625-mm thick slices which can quickly be reformated in sagittal and coronal planes, equivalent to traditional myelography projections.

INDICATIONS

CT myelography and MRI have largely replaced conventional myelography for the diagnosis of diseases of the spinal canal and cord (Table 416-1). Remaining indications for conventional plain-film myelography include the evaluation of suspected meningeal or arachnoid cysts and the localization of CSF fistulas. Conventional

myelography and CT myelography provide the most precise information in patients with failed back syndrome following spinal fusion procedures.

CONTRAINDICATIONS

Myelography is relatively safe; however, it should be performed with caution in any patient with elevated intracranial pressure, evidence of a spinal block, or a history of allergic reaction to intrathecal contrast media. In patients with a suspected spinal block, MR is the preferred imaging technique. If myelography is necessary, only a small amount of contrast medium should be instilled below the block in order to minimize the risk of neurologic deterioration. Lumbar puncture is to be avoided in patients with bleeding disorders, including patients receiving anticoagulant therapy, as well as in those with infections of the overlying soft tissues.

COMPLICATIONS

Headache is the most frequent complication of myelography and is reported to occur in 5–30% of patients. Nausea and vomiting may also occur rarely. Postural headache (post–lumbar puncture headache) is generally due to continued leakage of CSF from the dural puncture site. A higher incidence is noted among younger women and with the use of larger gauge cutting-type spinal needles. If significant headache persists for >48 h, placement of an epidural blood patch should be considered. **Management of lumbar puncture headache is discussed in Chap. 13.** Vasovagal syncope may occur during lumbar puncture; it is accentuated by the upright position used during lumbar myelography. Adequate hydration before and after myelography will reduce the incidence of this complication.

Hearing loss is a rare complication of myelography. It may result from a direct toxic effect of the contrast medium or from an alteration of the pressure equilibrium between CSF and perilymph in the inner ear. Puncture of the spinal cord is a rare but serious complication of cervical (C1–2) or high lumbar puncture. The risk of cord puncture is greatest in patients with spinal stenosis, Chiari malformations, or conditions that reduce CSF volume. In these settings, a low-dose lumbar injection followed by thin-section CT or MRI is a safer alternative to cervical puncture. Intrathecal contrast reactions are rare, but aseptic meningitis and encephalopathy are reported complications. The latter is usually dose related and associated with contrast entering the intracranial subarachnoid space. Seizures rarely occur following myelography, historically reported in 0.1–0.3% of patients. Risk factors include a preexisting seizure disorder and the use of a total iodine dose of >4500 mg. Other reported complications include hyperthermia, hallucinations, depression, and anxiety states. These side effects have been reduced by the development of nonionic, water-soluble contrast agents as well as by head elevation and generous hydration following myelography.

SPINE INTERVENTIONS

DISKOGRAPHY

The evaluation of back pain and radiculopathy may require diagnostic procedures that attempt either to reproduce the patient's pain or relieve it, indicating its correct source prior to lumbar fusion. Diskography is performed by fluoroscopic placement of a 22- to 25-gauge needle into the intervertebral disk and subsequent injection of 1–3 mL of contrast media. The intradiskal pressure is recorded, as is an assessment of the patient's response to the injection of contrast material. Typically little or no pain is felt during injection of a normal disk, which does not accept much more than 1 mL of contrast material, even at pressures as high as 415–690 kPa (60–100 lb/in^2). CT and plain films are obtained following the procedure. Concerns have been raised that diskography may contribute to an accelerated rate of disk degeneration; furthermore, patients who suffer from depression or anxiety are more likely to find diskography painful and in some cases the procedure-associated pain became persistent, lasting a year or longer.

SELECTIVE NERVE ROOT AND EPIDURAL SPINAL INJECTIONS

Percutaneous selective nerve root and epidural blocks with glucocorticoid and anesthetic mixtures may be both therapeutic and diagnostic, especially if a patient's pain is relieved. Typically, 1–2 mL of an equal mixture of a long-acting glucocorticoid such as betamethasone and a long-acting anesthetic such as bupivacaine 0.75% is instilled under CT or fluoroscopic guidance in the intraspinal epidural space or adjacent to an existing nerve root. This can also be performed in the facet joints, or around the medial nerve branches that supply innervation to the facet joints.

ANGIOGRAPHY

Catheter angiography is indicated for evaluating intracranial small-vessel pathology (such as vasculitis), for assessing vascular malformations and aneurysms, and in endovascular therapeutic procedures (Table 416-1). As noted above, angiography has been replaced for many indications by CT/CTA or MRI/MRA.

Angiography carries the greatest risk of morbidity of all diagnostic imaging procedures, owing to the necessity of inserting a catheter into a blood vessel, directing the catheter to the required location, injecting contrast material to visualize the vessel, and removing the catheter while maintaining hemostasis. Therapeutic transcatheter procedures (see below) have become important options for the treatment of some cerebrovascular diseases. The decision to undertake a diagnostic or therapeutic angiographic procedure requires careful assessment of the goals of the investigation and its attendant risks.

To improve tolerance to contrast agents, patients undergoing angiography should be well hydrated before and after the procedure. Because the femoral route is used most commonly, the femoral artery must be compressed after the procedure to prevent a hematoma from developing. The puncture site and distal pulses should be evaluated carefully after the procedure; complications can include thigh hematoma or lower extremity emboli.

COMPLICATIONS

A common femoral arterial puncture provides retrograde access via the aorta to the aortic arch and great vessels. The most feared complication of cerebral angiography is stroke. Thrombus can form on or inside the tip of the catheter, and atherosclerotic thrombus or plaque can be dislodged by the catheter or guide wire or by the force of injection and can embolize distally in the cerebral circulation. Risk factors for ischemic complications include limited experience on the part of the angiographer, atherosclerosis, vasospasm, low cardiac output, decreased oxygen-carrying capacity, advanced age, and prior history of migraine. The risk of a neurologic complication varies but is ~4% for transient ischemic attack and stroke, 1% for permanent deficit, and <0.1% for death.

Ionic contrast material injected into the cerebral vasculature can be neurotoxic if the BBB is breached, either by an underlying disease or by the injection of hyperosmolar contrast agent. Ionic contrast media are less well tolerated than nonionic media, probably because they can induce changes in cell membrane electrical potentials. Patients with dolichoectasia of the basilar artery can suffer reversible brainstem dysfunction and acute short-term memory loss during angiography, owing to the slow percolation of the contrast material and the consequent prolonged exposure of the brain. Rarely, an intracranial aneurysm ruptures during an angiographic contrast injection, causing subarachnoid hemorrhage, perhaps as a result of injection under high pressure.

SPINAL ANGIOGRAPHY

Spinal angiography may be indicated to evaluate vascular malformations and tumors and to identify the artery of Adamkiewicz (Chap. 434) prior to aortic aneurysm repair. The procedure is lengthy and requires the use of relatively large volumes of contrast; the incidence of serious complications, including paraparesis, subjective visual blurring, and altered speech, is ~2%. Gadolinium-enhanced MRA has been used successfully in this setting, as has iodinated contrast CTA, which has promise for replacing diagnostic spinal angiography for some indications.

INTERVENTIONAL NEURORADIOLOGY

This rapidly developing field is providing new therapeutic options for patients with challenging neurovascular problems. Available procedures include detachable coil therapy for aneurysms, particulate or liquid adhesive embolization of arteriovenous malformations, stent retrieval systems for embolectomy, balloon angioplasty and stenting of arterial stenosis or vasospasm, transarterial or transvenous embolization of dural arteriovenous fistulas, balloon occlusion of carotid-cavernous and vertebral fistulas, endovascular treatment of vein-of-Galen malformations, preoperative embolization of tumors, and thrombolysis of acute arterial or venous thrombosis. Many of these disorders place the patient at high risk of cerebral hemorrhage, stroke, or death.

The highest complication rates are found with the therapies designed to treat the highest risk diseases. The advent of electrolytically detachable coils ushered in a new era in the treatment of cerebral aneurysms. Two randomized trials found reductions of morbidity and mortality at 1 year among those treated for aneurysm

with detachable coils compared with neurosurgical clipping. It remains to be determined what the role of coils will be relative to surgical options, but in many centers, coiling has become standard therapy for many aneurysms.

Finally, recent studies of stent retrieval systems used to withdraw emboli have shown improved clinical outcomes in patients presenting with large vessel occlusions and signs of stroke **(Chap. 420)**.

FURTHER READING

Haller S et al: Arterial spin labeling perfusion of the brain: Emerging clinical applications. Radiology 281:337, 2016. [PubMed: 27755938]

Kim SJ, Kim KA: Safety issues and updates under MR environments. Eur J Radiol 89:7, 2017. [PubMed: 28267552]

Mandell DM et al: Intracranial vessel wall MRI: Principles and expert consensus recommendations of the American Society of Neuroradiology. AJNR Am J Neuroradiol 38:218, 2017. [PubMed: 27469212]

Nguyen HS et al: Progressing bevacizumab-induced diffusion restriction is associated with coagulative necrosis surrounded by viable tumor and decreased overall survival in patients with recurrent glioblastoma. AJNR Am J Neuroradiol 37:2201, 2016. [PubMed: 27492073]

Ramalho J et al: Gadolinium-based contrast agent accumulation and toxicity: An update. AJNR Am J Neuroradiol 37:1192, 2016. [PubMed: 26659341]

Chapter 417: Pathobiology of Neurologic Diseases

Stephen L. Hauser; Arnold R. Kriegstein; Stanley B. Prusiner

INTRODUCTION

The human nervous system is the organ of consciousness, cognition, ethics, and behavior; as such, it is the most intricate structure known to exist. More than one-third of the 23,000 genes encoded in the human genome are expressed in the nervous system. Each mature brain is composed of 100 billion neurons, several million miles of axons and dendrites, and $>10^{15}$ synapses. Neurons exist within a dense parenchyma of multifunctional glial cells that synthesize myelin, preserve homeostasis, and regulate immune responses. Measured against this background of complexity, the achievements of molecular neuroscience have been extraordinary. Advances have occurred in parallel with the development of new enabling technologies—in bioengineering and computational sciences, imaging, and cell, molecular and chemical biology—and moving forward it is likely that the pace of new discoveries will only increase. This chapter reviews a number of the most dynamic areas in neuroscience, specifically highlighting advances in immunology and inflammation, neurodegeneration, and stem cell biology. In each of these areas, recent discoveries are providing context for an understanding of the triggers and mechanisms of disease, and offering new hope for prevention, treatment, and repair of nervous system injuries. Discussions of the neurogenetics of behavior, advances in addiction science, and diseases caused by network dysfunction can be found in **Chap. 443** (Biology of Psychiatric Disorders); and new approaches to rehabilitation via harnessing of neuroplasticity, neurostimulation, and computer-brain interfaces are presented in **Chap. 477** (Emerging Neurotherapeutic Technologies).

NEUROIMMUNOLOGY AND NEUROINFLAMMATION

OLIGODENDROCYTES AND MYELIN

Myelin is the multilayered insulating substance that surrounds axons and speeds impulse conduction by permitting action potentials to jump between naked regions of axons (nodes of Ranvier) and across myelinated segments. Molecular interactions between the myelin membrane and axon are required to maintain the stability, function, and normal life span of both structures. A single oligodendrocyte usually ensheaths multiple axons in the central nervous system (CNS), whereas in the peripheral nervous system (PNS), each Schwann cell typically myelinates a single axon. Myelin is a lipid-rich material formed by a spiraling process of the membrane of the myelinating cell around the axon, creating multiple membrane bilayers that are tightly apposed (compact myelin) by charged protein interactions. Several inhibitors of axon growth are expressed on the innermost (periaxonal) lamellae of the myelin membrane (see below). A number of clinically important neurologic disorders are caused

by inherited mutations in myelin proteins of the CNS or PNS **(Chap. 438)**, and constituents of myelin also have a propensity to be targeted as autoantigens in autoimmune demyelinating disorders **(Chap. 436)**.

Premyelinating oligodendrocyte precursor cells (OPCs) are highly motile cells that migrate extensively during development and in the adult brain following injuries to the myelin sheath. OPCs migrate along the inner (or abluminal) surface of endothelial cells, a process regulated by *Wnt* pathway signaling and upregulation of the chemokine receptor Cxcr4 that drives their attachment and retention to the vasculature. Initial specification to OPCs is transcriptionally regulated by the *Olig 2* and *Yin Yang 1* genes, whereas the later stage of myelination mediated by postmitotic oligodendrocytes depends on a different transcription factor, *myelin gene regulatory factor (MRF)*. In the normal adult brain, large numbers of OPCs (expressing PDGFR-α and NG2) are widely distributed but do not myelinate axons, even in demyelinating environments such as in lesions of multiple sclerosis (MS). In addition to *Wnt*, several families of molecules have been identified that regulate oligodendrocyte differentiation and myelination, including LINGO-1, PSA-NCAM, hyaluronan, Nogo-A, the Wnt pathway, notch signaling (and its receptor Jagged), and the M1 muscarinic receptor Chrm1, all of which are inhibitory, and the retinoic acid receptor RXRγ, which is excitatory. All are also potential targets for myelin repair therapies. In an experimental allergic encephalomyelitis (EAE) model, oligodendrocyte-specific knockout of Chrm1 improved remyelination, protected axons and restored function, directly demonstrating that remyelination can be neuroprotective following injury. A recently reported pivotal trial of a monoclonal antibody against LINGO-1 failed to promote remyelination, a disappointing result given that the antibody appeared to have promising clinical effects in an earlier phase 2 trial.

A series of observations has called into question the traditional concept that axon-derived cues are always required for myelination to occur. Fixed (i.e., dead) axons could be efficiently myelinated by oligodendrocytes in vitro, as could artificial polystyrene nanowires of a similar diameter. This led to development of new high-throughput screening assays based on myelination of polystyrene nanowires to identify compounds that could promote myelination and in a preliminary human trial a molecule that emerged from this assay, the antihistamine clemastine, had clear efficacy as a remyelinating agent in patients with chronic optic neuropathy due to MS. Remarkably, the drug appears to work via binding to the Chrm1 muscarinic receptor.

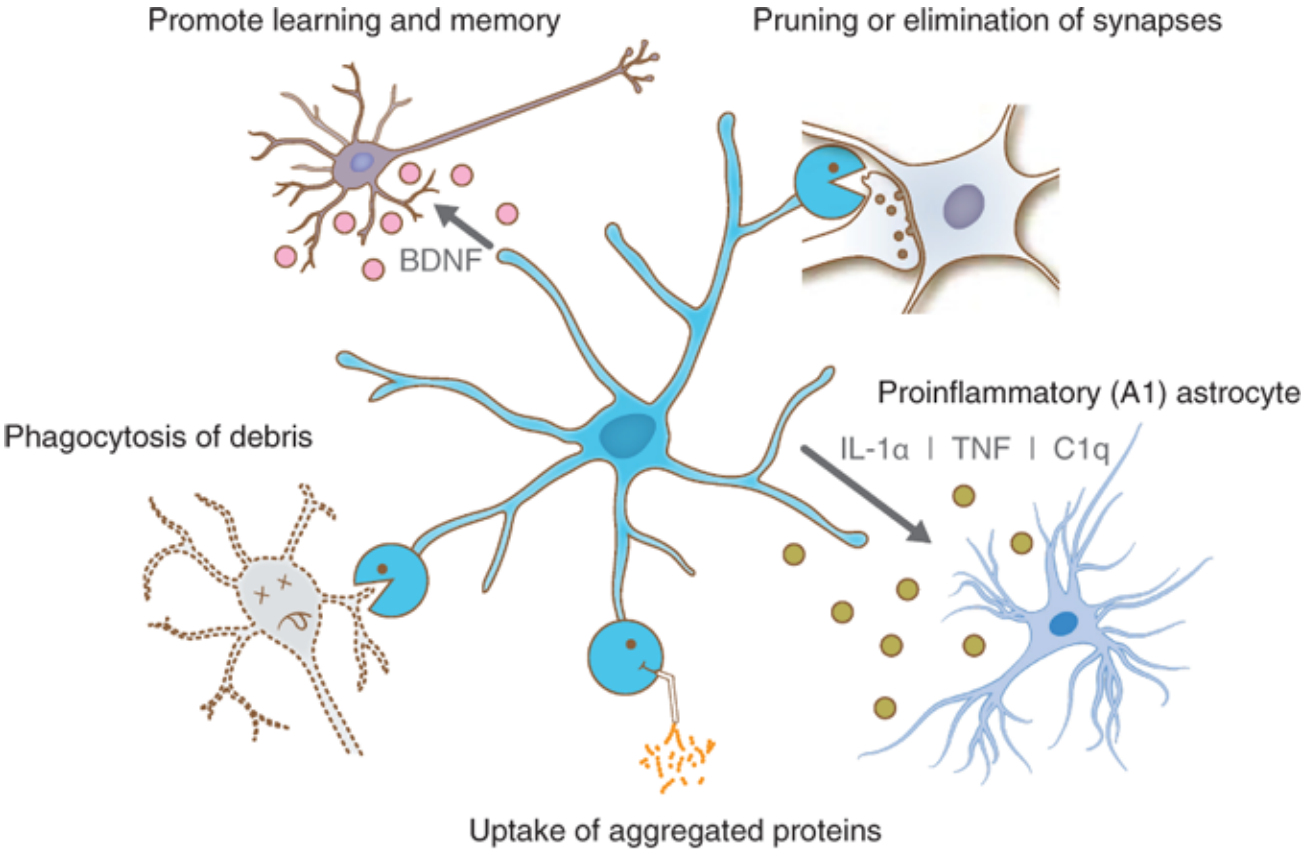
MACROPHAGES AND MICROGLIA

These represent the major cell types in the nervous system responsible for antigen presentation and innate immunity. Brain microglia migrate from the yolk sac early in embryogenesis before the blood-brain barrier is formed, and are believed to maintain their cell numbers through cell division within the nervous system and not via repopulation from the circulation. Depletion of microglia in adult mice by administration of a selective inhibitor of colony-stimulating factor receptor 1 (CSFR1) was followed by their rapid repopulation, suggesting that a pool of resident microglial precursor cells exist throughout the CNS. Additional roles for brain microglia are known to exist in neurogenesis, through secretion of brain derived neurotrophic factor (BDNF) and other molecules, as well as in the development and regulation of neural circuits through pruning of excitatory synapses and control of dendritic spine densities **(Fig. 417-1)**. Mice depleted of microglia during development exhibit a variety of cognitive, learning and behavioral deficits, including abnormal social behaviors; these processes are dependent on the classical complement pathway molecules and the chemokine receptor CX3CR1. A challenge to the field has been that tools to definitively separate brain microglia from perivascular macrophages do not

currently exist. A recent advance that could provide a possible solution utilizes adult skin-derived pluripotent stem cells (iPSCs), and the development of methods to generate microglia-like cells from iPSCs using media containing IL-34 and colony-stimulating factor 1.

FIGURE 417-1

The multifunctional microglial cell. Microglia have diverse functions that can support healthy development and maintain homeostasis, or contribute to tissue damage in pathologic conditions. Homeostatic functions include promotion of learning and memory through secretion of soluble proteins such as brain derived neurotrophic factor (BDNF); participation in normal synaptic pruning; and clearing cellular debris and protein aggregates via phagocytosis. However, in pathologic states activated microglia also contribute to tissue damage, by targeting normal healthy neurons and synapses; by promoting formation of β-amyloid or other misfolded proteins deposited in neurodegenerative diseases; and secreting cytokines (such as IL-1α, TNF, and the complement component C1q) incriminated in induction of neurotoxic A1 astrocytes. In addition, microglia have diverse functions in adaptive immunity, including roles in antigen presentation and immune regulation (Fig. 417-2). *(Figure adapted from J Herz et al: Myeloid cells in the central nervous system. Immunity 46:943, 2017, Fig. 2.)*



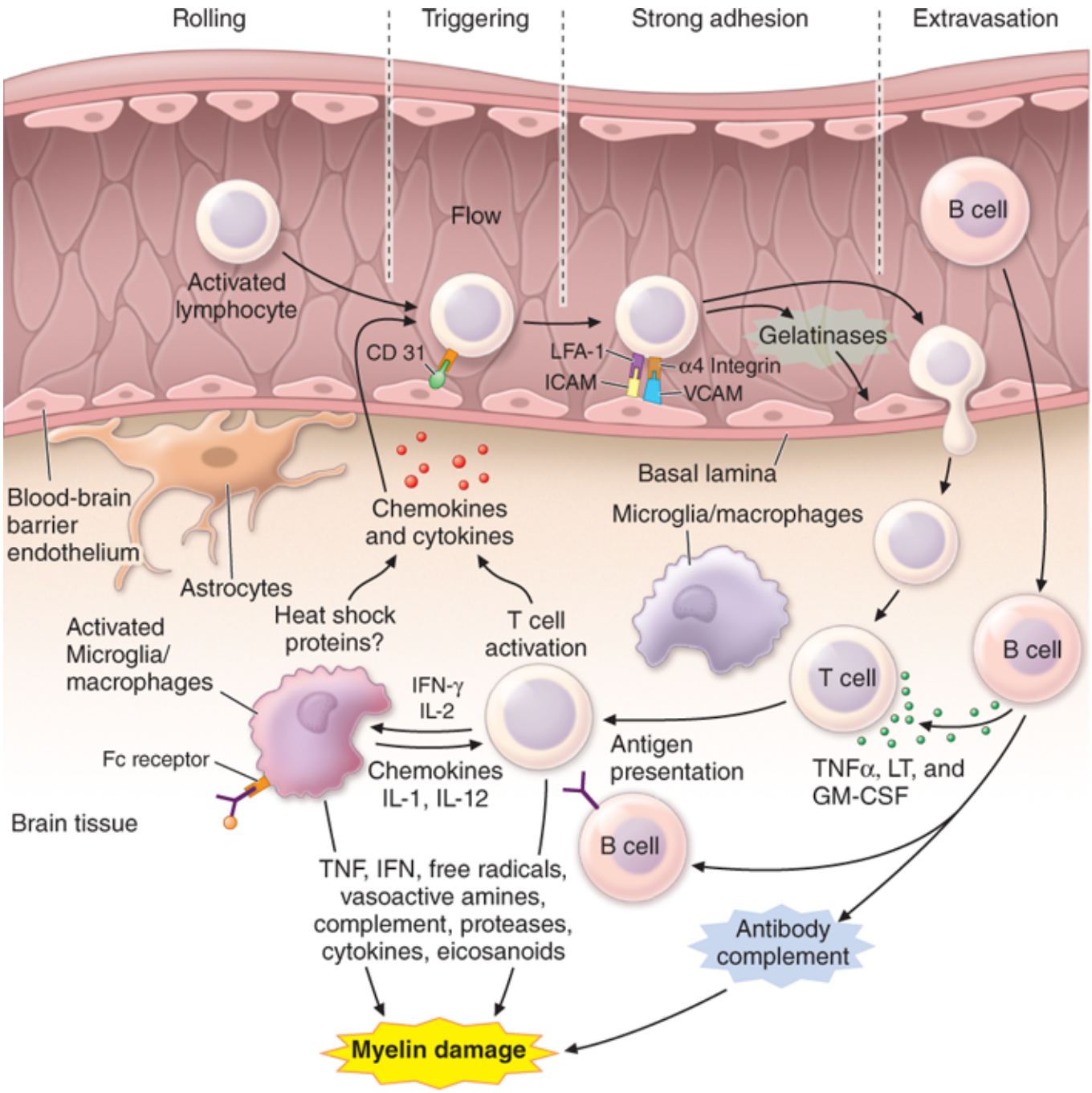
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Microglia are located throughout the brain parenchyma, whereas brain macrophages occur primarily in perivascular regions, including the meninges and choroid plexus. In contrast to microglia, brain macrophages are derived from monocytes that enter the nervous system at low levels from the bloodstream on a continuing basis and in higher numbers during pathologic states. A possible exception to this rule are the meningeal macrophages, located primarily in the subdural space, that appear to enter the brain at an early developmental stage and remain throughout the life of the individual. In a murine model of autoimmune demyelination, EAE **(Fig. 417-2)**,

macrophages derived from bone marrow monocytes, but not microglia, were the critical population that initiated inflammatory demyelination at paraxonal regions near nodes of Ranvier. Brain macrophages have been found to have multiple pro-inflammatory functions, including promoting adhesion, attraction and activation of B and T lymphocytes; providing antigen-specific activation of T cells via antigen presentation of specific immunogenic peptides, including autoantigens, complexed to surface class II major histocompatibility complex (MHC II) molecules; and contributing to cell injury through generation of oxidative stress and cytotoxicity. By contrast, microglia have been traditionally thought to downregulate inflammatory responses and promote tissue repair. This model of M1 (pro-inflammatory) and M2 (regulatory/repair) macrophage/microglial functions, derived primarily from experimental models of autoimmunity, is certainly an oversimplification, and more nuanced functions of these cell types can be revealed depending on the specific context and environmental cues.

FIGURE 417-2

A model for experimental allergic encephalomyelitis (EAE). Crucial steps for disease initiation and progression include peripheral activation of preexisting autoreactive T cells; homing to the central nervous system (CNS) and extravasation across the blood-brain barrier; reactivation of T cells by exposed autoantigens; secretion of cytokines; activation of microglia and astrocytes and recruitment of a secondary inflammatory wave; and immune-mediated myelin destruction. ICAM, intercellular adhesion molecule; IFN, interferon; IL, interleukin; LFA-1, leukocyte function-associated antigen-1; TNF, tumor necrosis factor; VCAM, vascular cell adhesion molecule.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Evidence also supports a primary role for brain macrophages and microglia in neurodegenerative diseases, in contrast to earlier views in which their role was seen as largely secondary and involving phagocytosis of cell debris. In different situations, macrophages and microglia can be either protective or pathogenic. In mice, macrophages contribute to spatial memory when activated in the presence of the cytokine interleukin (IL)-4 produced by invading lymphocytes, and microglia through secretion of BDNF support learning and memory through promoting synaptic plasticity. Experimentally, microglia and brain macrophages also participate in clearance of pathogenic β-amyloid aggregates in Alzheimer's disease (AD) mice, and disruption of brain macrophages by knockout of CCR2, a chemokine required for entry of bloodstream monocytes into the CNS, exacerbated AD pathology. On the other hand, data indicate that disease exacerbating effects of microglia and macrophages may predominate in other situations. A direct role for microglia in human AD was suggested by

genetic evidence implicating the phagocytosis-associated gene TREM2, and other genes belonging to the complement system, in AD susceptibility. Activation of the classical complement cascade is also assuming an increased role in concepts of pathogenesis, as follows; synapses targeted for elimination express the complement proteins C1q and C3, the levels of which increase in the presence of excess β-amyloid; C3-bearing synapses are then targeted for elimination by microglia that express the complement 3 receptor (CR3); and knockout of C3 can rescue the clinical and pathologic abnormalities associated with neurodegeneration in AD-prone mice. In familial frontotemporal degeneration (FTD) due to mutations of progranulin (pgrn), a prominent immune pathology has also been identified, including the presence of activated microglia expressing high levels of pro-inflammatory cytokines. In pgrn-/- mice, an age-dependent microglial activation is associated with upregulation of genes associated with innate immunity including complement proteins, and with enhanced pruning of inhibitory synapses in key regions of the CNS, leading to behavioral disorders reminiscent of human FTD. Moreover, inhibition of complement activation rescued all of these deficits. Taken together, these data indicate a primary role for microglial activation in pgrn associated FTD, likely occurring through enhanced lysosomal trafficking and increased production of cleavage products of the C3 complement component, and leading to enhanced and deleterious synaptic pruning in regions of the brain affected in FTD. Although it is likely that the specific mechanisms of complement dependent neurodegeneration will differ in distinct neurodegenerative conditions, these data provide hope that complement pathway interventions could represent a possible approach to control of neurodegenerative pathologies mediated at least in part through the innate immune system.

ASTROCYTES

Astrocytes represent half or more of all cells in the CNS. Traditionally thought to function as simple interstitial supporting cells that provide scaffolds for neuronal migration and contribute to homeostasis, emerging data indicate far more pleiotropic functions for this cell type. Astrocytes exert profound roles in the life of synapses by secreting factors (such as apolipoprotein E, thrombospondins, and glypicans) that regulate development, maintenance, and pruning of presynaptic and postsynaptic structures. Influenced by local neuronal activity, astrocytes actively phagocytose synapses. Astrocytes also participate in dynamic regulation of vascular tone, in part through astrocyte-astrocyte communication mediated through gap junctions and calcium waves modulated by neuronal activity; support blood brain barrier and glymphatic (see below) integrity through extension of foot process to the vascular structures and expression of aquaporin-4 water channels; and carry out additional metabolic functions essential for the maintenance of neuronal health.

One characteristic of the response to many types of brain injury is reactive astrocytosis, or the formation of a glial scar. Recent work has identified two fundamentally different types of reactive astrocytes that appear to have countervailing functions; the terms A1 and A2 astrocytes have been proposed, by analogy to brain macrophage/microglia M1 and M2 designations, described above. A2 astrocytes are induced by ischemia and may serve beneficial functions, including a contribution to tissue repair after injury. By contrast, A1 astrocytes are induced in diverse inflammatory and degenerative states, and appear to actively participate in the injury process. Interestingly, secreted products of activated microglia, specifically IL-1a, TNF, and C1q, induce astrocytes to transform to the A1 type. Functionally, A1 astrocytes lose the capacity to phagocytose synapses and myelin debris, and are strikingly toxic in vitro to various populations of neurons and to mature oligodendrocytes, potentially at least in part via complement mediated damage. Interestingly, OPCs, abundant in active lesions of

multiple sclerosis (MS; Chap. 436) despite the inflammatory milieu, are resistant to A1 mediated death. The nature of the toxic factor is unknown. There is speculation that products of A1 astrocytes could promote damage in conditions as varied as MS, AD (Chap. 423), Parkinson's disease (PD) (Chap. 427), and amyotrophic lateral sclerosis (ALS) (Chap. 429), despite their distinct etiologies and pathologies.

GLYMPHATICS

Two newly identified lymphatic structures of the CNS are the glymphatic and deep dural lymphoid systems, responsible for clearance of debris in the CNS, and likely also serving a role in immune surveillance. The brain has traditionally been considered to lack a classical lymphatic system, and immune responses against antigens are less effectively generated in the CNS than in other organ systems, a concept termed "immune privilege." However, there is abundant evidence that the immune privilege status of the brain is only relative and not absolute. Furthermore, given the high metabolic demands of the brain some mechanism for efficient removal of solute and debris must be present. One well-established pathway involves the passive flow of solutes from the brain parenchyma into the cerebrospinal fluid (CSF), and their exit via the arachnoid granulations, as well as along cranial and spinal nerve roots to a series of lymphoid structures located in the cribriform plate and nasal mucosa and elsewhere.

The glymphatic system derives its name from a distinctive architecture involving lymphoid-like structures and astroglial cells. CSF synthesized in the arachnoid villi circulates through the ventricles and subarachnoid space surrounding the convexities of the brain and spinal cord, and exits through conduits surrounding arterioles penetrating into the brain parenchyma. These spaces are lined by endothelial cells internally, and by astrocyte foot processes that form the external walls. Aided by arterial propulsion, CSF moves out of these specialized conduits and into astrocytes via foot processes rich in aquaporin-4 water channels, and then in the interstitium of brain parenchyma picks up solutes and particulate debris that are then carried to perivenous spaces where they passage to exit the brain and drain into the lymphatic system. In mice, knockout of aquaporin-4 markedly reduced the flow of interstitial fluids in the brain, underscoring the critical role of astrocyte uptake of CSF in this process. Interstitial flow in the CNS is also impaired with aging, possibly related to changes in astrocytic aquaporin-4 expression. Another fascinating aspect of the glymphatic system is that the transport of fluids and solutes accelerates with sleep, arguing for a critical role for sleep in promoting clearance of debris needed to meet the high metabolic demands of the nervous system. Furthermore, in disease models, aggregated proteins associated with neurodegenerative disease, such as β-amyloid associated with AD (Chap. 423), were also more efficiently cleared during sleep. Indeed, in mice genetically engineered to produce excess β-amyloid and develop Alzheimer's-like cognitive decline, sleep deprivation increased accumulation of amyloid plaques. Glymphatic pathways are also likely to represent an important egress pathway for lymphocytes in the CNS and a route for lymphocyte encounter with CNS antigens in cervical lymph nodes. In this regard, recent data indicate that deep cervical lymph nodes might be a site for antigen-specific stimulation of B-cells in MS (Chap. 436).

A second recently identified pathway consists of a plexus of small lymphatic-like vessels located on the external surface of meningeal arteries and deep dural sinuses (including the sagittal and transverse sinuses), structures that exit the brain along the surface of veins and arteries and drain to the deep cervical lymph nodes. These conduits are comprised of cells that express a transcriptome indicating that they are components of a lymphoid drainage system distinct from vascular endothelium. These sinus-associated lymphoid structures may be most

important in clearing solutes from the CSF, in contrast to the glymphatic system that likely functions to remove waste products from the brain interstitium; however, the exact functions of these two systems and their interrelationships are only beginning to be understood.

MICROBIOTA AND NEUROLOGIC DISEASE

The human microbiome **(Chap. 459)** represents the collective set of genes from the 10^{14} organisms living in our gut, skin, mucosa, and other sites. Different microbial communities are associated with different ethnicities, diets, and environments. In any individual, the predominant gut microbiota can be remarkably stable over decades, but also can be altered by exposure to certain microbial species, for example by ingestion of probiotics.

There is compelling evidence that gut microbes can shape immune responses through the interaction of their metabolism with that of humans. These gut-brain interactions are likely to be important in understanding the pathogenesis of many autoimmune neurologic diseases. For example, mice treated with broad-spectrum antibiotics are resistant to EAE, an effect associated with decreases in production of proinflammatory cytokines, and conversely more production of the immunosuppressive cytokines IL-10 and IL-13 and an increase in regulatory T and B lymphocytes. Oral administration of polysaccharide A (PSA) from *Bacillus fragilis* also protects mice from EAE, via increases in IL-10. Intestinal microbiota from patients with MS were found to promote EAE when transferred to germ free mice, possibly due to imbalances between bacterial species that promote inflammation (such as *Akkermansia muciniphila* and *Acinetobacter calcoaceticus*) and those that induce regulatory immune responses (such as *Parabacteroides distasonis*).

In addition to nonspecific effects on immune homeostasis mediated by cytokines and regulatory cells, some microbial proteins can trigger, in susceptible individuals, a cross-reactive immune response against a homologous protein in the nervous system, a mechanism termed *molecular mimicry*. Examples include cross-reactivity between the astrocyte water channel aquaporin-4 and an ABC transporter permease from *Clostridia perfringens* in neuromyelitis optica **(Chap. 437)**; the neural ganglioside Gm1 and similar sialic acid–containing structures from *Campylobacter jejuni* in Guillain-Barré syndrome **(Chap. 439)**; and the sleep-promoting protein hypocretin and hemagglutinin from H1N1 influenza virus in narcolepsy **(Chap. 27)**.

Recently, a number of tantalizing observations have incriminated the microbial environment in the pathogenesis of a much wider spectrum of neurologic conditions and behaviors, extending well beyond the traditional boundaries of immune-mediated pathologies. It has long been known that gut bacteria can influence brain function, based mostly on classic studies demonstrating that products of gut microbes can worsen hepatic encephalopathy, forming the basis of treatment with antibiotics for this condition.

Mice that developed in a completely germ-free environment displayed less anxiety, lower responses to stressful situations, more exploratory locomotive behaviors, and impaired memory formation compared with non-germ-free counterparts. These behaviors were related to changes in gene expression in pathways related to neural signaling, synaptic function, and modulation of neurotransmitters. Moreover, this behavior could be reversed when the germ-free mice were co-housed with non-germ-free mice. In other experiments, intestinal microbiota were also found to be required for the normal development and function of brain microglia, potentially linking these behavioral effects to specific cellular targets in the CNS.

The enteric autonomic nervous system in humans provides a bidirectional neural connection between the brain and gut. The vagus nerve, which innervates the upper gut and proximal colon, has been implicated in anxiety- and depression-like behaviors in mice. Ingestion of *Lactobacillus rhamnosus* induced changes in expression of the inhibitory neurotransmitter GABA1b in neurons of the limbic cortex, hippocampus, and amygdala, associated with reduced levels of corticosteroids and reduced anxiety- and depression-like behaviors. Remarkably, these changes could be blocked by vagotomy.

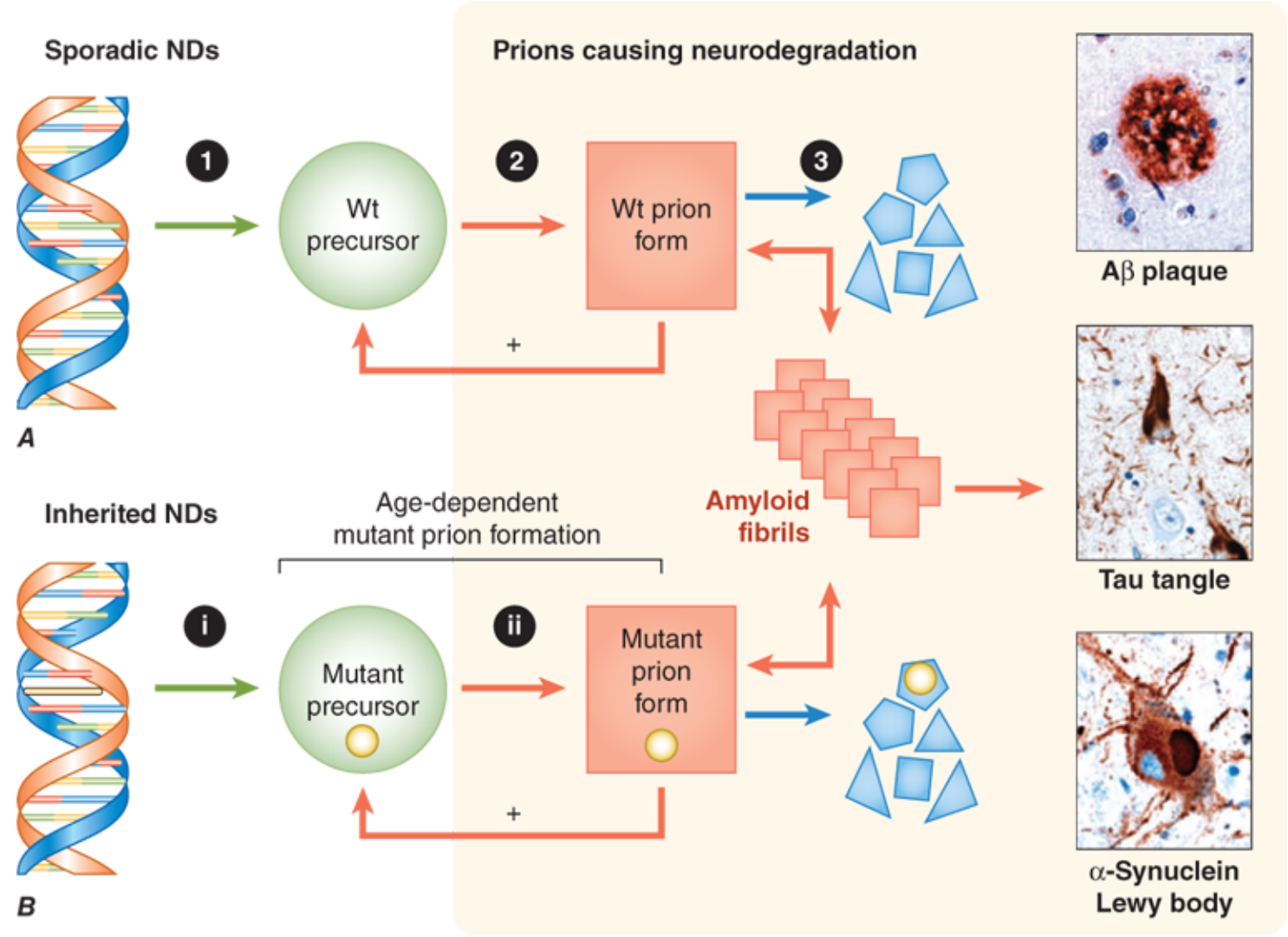
Another area of emerging interest is in a possible contribution of the gut microbiome to autism and related disorders. Children with autistic spectrum disorders have long been known to have gastrointestinal disturbances, and it has been claimed that the severity of dysbiosis correlates with the severity of autism. A murine model of autism was recently induced in offspring after injecting the pregnant mother with the viral RNA mimic polyinosinic:polycytidylic acid (poly I:C). Remarkably, oral treatment of offspring with *B. fragilis* corrected a range of autistic behaviors in these mice and also improved gut permeability.

PATHOLOGIC PROTEINS, PRIONS, AND NEURODEGENERATION

The term “protein aggregation” has become widely used to describe easily recognizable hallmarks of neurodegeneration **(Fig. 417-3)**. While such neuropathologic hallmarks including plaques, neurofibrillary tangles, and inclusion bodies are often thought to cause neurologic dysfunction, numerous new discoveries over the past several decades have rendered this view increasingly unlikely. Instead, protein aggregates represent accumulations of toxic proteins that become less harmful when they are sequestered into plaques, tangles, and inclusion bodies.

FIGURE 417-3

Neurodegeneration caused by prions. ***A.*** In sporadic neurodegenerative diseases (NDs), wild-type (Wt) prions multiply through self-propagating cycles of posttranslational modification, during which the precursor protein (*green circle*) is converted into the prion form (*red square*), which generally is high in β-sheet content. Pathogenic prions are most toxic as oligomers and less toxic after polymerization into amyloid fibrils. The small polygons (*blue*) represent proteolytic cleavage products of the prion. Depending on the protein, the fibrils coalesce into Aβ amyloid plaques in AD, neurofibrillary tangles in AD and Pick’s disease, or Lewy bodies in PD and Lewy body dementia. Drug targets for the development of therapeutics include: (1) lowering the precursor protein, (2) inhibiting prion formation, and (3) enhancing prion clearance. ***B.*** Late-onset heritable neurodegeneration argues for two discrete events: The (*i*) first event is the synthesis of mutant precursor protein (*green circle*), and the (*ii*) second event is the age-dependent formation of mutant prions (*red square*). The *highlighted yellow bar* in the DNA structure represents mutation of a base pair within an exon, and the *small yellow circles* signify the corresponding mutant amino acid substitution. *Green arrows* represent a normal process; *red arrows*, a pathogenic process; and *blue arrows*, a process that is known to occur but unknown whether it is normal or pathogenic. *(Micrographs prepared by Stephen J. DeArmond. Reprinted with permission from SB Prusiner: Biology and genetics of prions causing neurodegeneration. Annu Rev Genet 47:601, 2013.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Deposition of β-amyloid is strongly implicated in the pathogenesis of AD. Genetic mutations in familial AD cause increased production of β-amyloid with 42 amino acids, which has an increased propensity to aggregate, as compared to β-amyloid with 40 amino acids. Furthermore, mutations in the amyloid precursor protein (APP) which reduce the production of β-amyloid protect against the development of AD and are associated with preserved cognition in the elderly. Mutations in genes encoding MAPT lead to altered splicing of tau and the production of neurofibrillary tangles in frontotemporal dementia and progressive supranuclear palsy. Familial PD is associated with mutations in *leucine-rich repeat kinase 2 (LRRK2), α-synuclein, parkin, PINK1,* and *DJ-1*. PINK1 is a mitochondrial kinase (see below), and DJ-1 is a protein involved in protection from oxidative stress. Parkin, which causes autosomal recessive early-onset PD, is a ubiquitin ligase. The characteristic histopathologic feature of PD is the Lewy body, an eosinophilic cytoplasmic inclusion that contains both neurofilaments and α-synuclein. Huntington's disease (HD) and cerebellar degenerations are associated with expansions of polyglutamine repeats in proteins, which aggregate to produce neuronal intranuclear inclusions. Familial ALS is associated with superoxide dismutase mutations and cytoplasmic inclusions containing superoxide dismutase. An important finding was the discovery that ubiquinated inclusions observed in most cases of ALS and the most common form of frontotemporal dementia are composed of TAR DNA binding protein 43 (TDP-43). Subsequently, mutations in the TDP-43 gene, and in the fused in sarcoma gene (FUS), were found in familial ALS. Both of these proteins are

involved in transcription regulation as well as RNA metabolism. In autosomal dominant neurohypophyseal diabetes insipidus, mutations in vasopressin result in abnormal protein processing, accumulation in the endoplasmic reticulum, and cell death.

Another key mechanism linked to cell death is mitochondrial dynamics, which refers to the processes involved in movement of mitochondria, as well as in mitochondrial fission and fusion, which play a critical role in mitochondrial turnover and in replenishment of damaged mitochondria. Mitochondrial dysfunction is strongly linked to the pathogenesis of a number of neurodegenerative diseases such as Friedreich's ataxia, which is caused by mutations in an iron-binding protein that plays an important role in transferring iron to iron-sulfur clusters in aconitase and complex I and II of the electron transport chain. Mitochondrial fission is dependent on the dynamin-related proteins (Drp1), which bind to its receptor Fis, whereas mitofuscins 1 and 2 (MF 1/2) and optic atrophy protein 1 (OPA1) are responsible for fusion of the outer and inner mitochondrial membrane, respectively. Mutations in MFN2 cause Charcot-Marie-Tooth neuropathy type 2A, and mutations in OPA1 cause autosomal dominant optic atrophy. Both β-amyloid and mutant huntingtin protein induce mitochondrial fragmentation and neuronal cell death associated with increased activity of Drp1. In addition, mutations in genes causing autosomal recessive PD, *parkin* and *PINK1*, cause abnormal mitochondrial morphology and result in impairment of the ability of the cell to remove damaged mitochondria by autophagy.

One major scientific question is whether protein aggregates directly contribute to neuronal death or whether they are merely secondary bystanders. A current focus in all the neurodegenerative diseases is on small protein aggregates termed *oligomers*. These may be the toxic species of β-amyloid, α-synuclein, and proteins with expanded polyglutamines such as are associated with HD. Protein aggregates are usually ubiquinated, which targets them for degradation by the 26S component of the proteasome. An inability to degrade protein aggregates could lead to cellular dysfunction, impaired axonal transport, and cell death by apoptotic mechanisms.

Autophagy is the degradation of cystolic components in lysosomes. There is increasing evidence that autophagy plays an important role in degradation of protein aggregates in the neurodegenerative diseases, and it is impaired in AD, PD, and HD. Autophagy is particularly important to the health of neurons, and failure of autophagy contributes to cell death. In HD, a failure of cargo recognition occurs, contributing to protein aggregates and cell death. Rapamycin, which induces autophagy, exerts beneficial therapeutic effects in transgenic mouse models of AD, PD, and HD.

There is other evidence for lysosomal dysfunction and impaired autophagy in PD. Mutations in glucocerebrosidase are associated with 5% of all PD cases as well as 8–9% of patients with dementia with Lewy bodies. Therefore, this is the most important genetic cause of both disorders thus far identified. There appear to be reciprocal interactions between glucocerebrosidase and α-synuclein. It has been shown that glucocerebrosidase concentrations and enzymatic activity are reduced in the substantia nigra of sporadic PD patients. Furthermore, α-synuclein is degraded by chaperone-mediated and macro autophagy. The degradation of α-synuclein has been shown to be impaired in transgenic mice deficient in glucocerebrosidase as well as in mice in which the enzyme has been inhibited. Finally, it is known that α-synuclein inhibits the activity of glucocerebrosidase. Therefore, there is bidirectional feedback between α-synuclein and glucocerebrosidase. An attractive therapeutic intervention could be to use protein chaperones to increase the activity and duration of

action of glucocerebrosidase. This would also reduce α-synuclein levels and block the degeneration of dopaminergic neurons.

The retromer complex is a conserved membrane-associated protein complex that functions in the endosome-to-Golgi complex. The retromer complex contains a cargo selective complex consisting of VPS35, VPS26, and VPS29, along with a sorting nexin dimer. Recently, mutations in VPS35 were shown to be a cause of late-onset autosomal dominant PD. The retromer also traffics APP away from endosomes, where it is cleaved to generate β-amyloid. Deficiencies of VPS35 and VPS26 were also identified in hippocampal brain tissue from AD. A new therapeutic approach to these diseases might therefore be to use chaperones to stabilize the retromer and reduce the generation of β-amyloid and α-synuclein.

The LRRK2 mutations were shown to have effects on clearance of Golgi-derived vesicles through the autophagy-lysosome system both in vitro and in vivo. LRRK2 mutations also are linked to elevated protein synthesis mediated by ribosomal protein s15 phosphorylation. Blocking this phosphorylation reduces LRRK2-mediated neurite loss and cell death in human dopamine and cortical neurons.

Interestingly, in experimental models of HD and cerebellar degeneration, protein aggregates are not well correlated with neuronal death and may be protective. A substantial body of evidence suggests that the mutant proteins with polyglutamine expansions in these diseases bind to transcription factors and that this contributes to disease pathogenesis. In HD, there is dysfunction of the transcriptional co-regulator, PGC-1α, a key regulator of mitochondrial biogenesis. There is evidence that impaired function of PGC-1α is also important in both PD and AD, making it an attractive target for treatments. Agents that upregulate gene transcription are neuroprotective in animal models of these diseases. A number of compounds have been developed to block β-amyloid production and/or aggregation, and these agents are being studied in early clinical trials in humans. Another approach under investigation is immunotherapy with antibodies that bind β-amyloid, tau, or α-synuclein. These studies have shown efficacy in preventing the spread of amyloid, tau, and α-synuclein in animal studies, raising hopes that this could lead to effective therapies by blocking neuron-to-neuron propagation. Two large clinical trials of β-amyloid immunotherapy, however, did not show efficacy, although this therapeutic strategy is still being studied.

PRIONS AND NEURODEGENERATIVE DISEASES

As we have learned more about the etiology and pathogenesis of the neurodegenerative diseases, it has become clear that the histologic abnormalities that were once curiosities, in fact, are likely to reflect the etiologies. For example, the amyloid plaques in kuru and Creutzfeldt-Jakob disease (CJD) are filled with the PrP^{Sc} prions that have assembled into fibrils. The past three decades have witnessed an explosion of new knowledge about prions. For many years, kuru, CJD, and scrapie of sheep were thought to be caused by slow-acting viruses, but a large body of experimental evidence argues that the infectious pathogens causing these diseases are devoid of nucleic acid. Such pathogens are called prions, which are composed of host-encoded proteins that adopt alternative conformations that undergo self-propagation (Chap. 430). Prions impose their conformations on the normal, precursor proteins, which in turn become self-templating resulting in faithful copies; most prions are enriched for β-sheet and can assemble into amyloid fibrils.

Similar to the plaques in kuru and CJD that are composed of PrP prions, the amyloid plaques in AD are filled with Aβ prions that have polymerized into fibrils. This relationship between the neuropathologic findings and the etiologic prion was strengthened by the genetic linkage between familial CJD and mutations in the PrP gene, as well as (as noted above) between familial AD and mutations in the *APP* gene. Moreover, a mutation in the *APP* gene that prevents Aβ peptide formation was correlated with a decreased incidence of AD in Iceland.

The heritable neurodegenerative diseases offer an important insight into the pathogenesis of the more common, sporadic ones. Although the mutant proteins that cause these disorders are expressed in the brains of people early in life, the diseases do not occur for many decades. Many explanations for the late onset of familial neurodegenerative diseases have been offered, but none are supported by substantial experimental evidence. The late onset might be due to a second event in which a mutant protein, after its conversion into a prion, begins to accumulate at some rather advanced age. Such a formulation is also consistent with data showing that the protein quality-control mechanisms diminish in efficiency with age. Thus, the prion forms of both wild-type and mutant proteins are likely to be efficiently degraded in younger people but are less well handled in older individuals. This explanation is consistent with the view that neurodegenerative diseases are disorders of the aging nervous system.

A new classification for neurodegenerative diseases can be proposed based on not only the traditional phenotypic presentation and neuropathology, but also the prion etiology **(Table 417-1)**. Over the past decade, an expanding body of experimental data has accumulated implicating prions in each of these illnesses. In addition to kuru and CJD, Gerstmann-Sträussler-Scheinker disease (GSS) and fatal insomnia in humans are caused by PrP^{Sc} prions. In animals, PrP^{Sc} prions cause scrapie of sheep and goats, bovine spongiform encephalopathy (BSE), chronic wasting disease (CWD) of deer and elk, feline spongiform encephalopathy, and transmissible mink encephalopathy (TME). Similar to PrP, Aβ, tau, α-synuclein, superoxide dismutase 1 (SOD1), and possibly huntingtin all adopt alternative conformations that become self-propagating, and thus, each protein can become a prion and be transferred to synaptically connected neurons. Moreover, each of these prions causes a distinct constellation of neurodegenerative diseases.

TABLE 417-1

A Prion-Based Classification of Neurodegenerative Diseases

Neurodegenerative Diseases	Causative Prion Proteins
Creutzfeldt-Jakob disease (CJD) Kuru Gerstmann-Sträussler-Scheinker (GSS) Fatal insomnia Bovine spongiform encephalopathy (BSE) Scrapie Chronic wasting disease (CWD) Feline spongiform encephalopathy Transmissible mink encephalopathy	PrP^{Sc}
Alzheimer's disease (AD)	Aβ → tau
Parkinson's disease Multiple system atrophy	α-Synuclein
Frontotemporal dementias (FTDs) Posttraumatic FTD, called chronic traumatic encephalopathy	Tau, TDP43, FUS (C9orf72, progranulin)
Amyotrophic lateral sclerosis	SOD1, TDP43, FUS (C9orf72)
Huntington's disease	Huntingtin

Evidence for a prion etiology of AD comes from a series of transmission experiments initially performed in marmosets and more recently in transgenic (Tg) mice inoculated with a synthetic Aβ peptide folded into a prion. Studies with the tau protein have shown that it not only features in the pathogenesis of AD, but also causes such illnesses as the frontotemporal dementias including chronic traumatic encephalopathy, which has been reported in both contact sport athletes and military personnel who have suffered traumatic brain injuries. A series of incisive studies using cultured cells and Tg mice has demonstrated that tau can become a prion and multiply in the brain. In contrast to the Aβ and tau prions, a strain of α-synuclein prions found in the brains of patients who died of multiple system atrophy (MSA) killed the Tg mouse host ~90 days after intracerebral inoculation, whereas mutant α-synuclein (A53T) prions formed spontaneously in Tg mouse brains killed recipient mice in ~200 days.

For many years, the most frequently cited argument against prions was the existence of strains that produced distinct clinical presentations and different patterns of neuropathologic lesions. Some investigators argued that the biologic information carried in different prion strains could only be encoded within a nucleic acid. Subsequently, many studies demonstrated that strain-specified variation is enciphered in the conformation of PrP^{Sc}, but the molecular mechanisms responsible for the storage of biologic information remains enigmatic. The

neuroanatomical patterns of prion deposition have been shown to be dependent on the particular strain of prion. Convincing evidence in support of this proposition has been accumulated for PrP, Aβ, tau, and α-synuclein prions.

Although the number of prions identified in mammals and in fungi continues to expand, the existence of prions in other phylogeny remains undetermined. Some mammalian prions perform vital functions and do not cause disease; such nonpathogenic prions include the cytoplasmic polyadenylation element binding (CPEB) protein, the mitochondrial antiviral-signaling (MAVS) protein, and T cell–restricted intracellular antigen 1 (TIA-1).

All mammalian prion proteins adopt a β-sheet-rich conformation and appear to readily oligomerize as this process becomes self-propagating. Control of the self-propagating state of benign mammalian prions is not well understood but is critical for the well-being of the host. In contrast, pathogenic mammalian prions appear to multiply exponentially, but the mechanisms by which they cause disease are poorly defined. We do not know if prions multiply as monomers or as oligomers; notably, the ionizing radiation target size of PrP^{Sc} prions seems to suggest it is a trimer. The oligomeric states of pathogenic mammalian prions are thought to be the toxic forms, and assembly into larger polymers, such as amyloid fibrils, seems to be a mechanism for minimizing toxicity.

To date, there is no medication that halts or even slows a human neurodegenerative disease. The development of drugs designed to inhibit the conversion of the normal precursor proteins into prions or to enhance the degradation of prions focuses on the initial step in prion accumulation. Although a dozen drugs that cross the blood-brain barrier have been identified that prolong the lives of mice infected with scrapie prions, none have been identified that extend the lives of Tg mice that replicate human CJD prions. Despite doubling or tripling the length of incubation times in mice inoculated with scrapie prions, all of the mice eventually succumb to illness. Because all of the treated mice develop neurologic dysfunction at the same time, the mutation rate as judged by drug resistance is likely to approach 100%, which is much higher than mutation rates recorded for bacteria and viruses. Mutations in prions seem likely to represent conformational variants that are selected for in mammals where survival becomes limited by the fastest-replicating prions. The results of these studies make it likely that cocktails of drugs that attack a variety of prion conformers will be required for the development of effective therapeutics.

NEURAL STEM CELL BIOLOGY

Normal and genetically modified (“transgenic”) mice are the most widely used model systems to study features of human nervous system diseases. However, modeling genetic diseases in rodents is limited to the relatively small number of monogenic human diseases where the specific gene mutations are known, and is further limited by species differences. The latter can be particularly important in brain regions such as the cerebral cortex that have undergone significant evolutionary expansion in humans. These shortcomings, that likely contribute to the low probability that therapeutic efficacy translates from animal models to humans, can potentially be overcome through stem cell models that enable the use of human cells and tissues to model human diseases. The advent of new stem cell technologies is transforming our understanding of the pathobiology of human neurologic diseases. Stem cell platforms are being used to screen for therapeutic agents, to uncover adverse drug effects, and to discover novel therapeutic targets.

Among the most exciting recent advances in stem cell technology is the ability to convert somatic cells, either skin fibroblasts or blood cells, into pluripotent stem cells known as induced pluripotent stem cells (iPSCs). This

technology has introduced an entirely new and powerful approach to study the pathobiology of heritable diseases. Pluripotent stem cells can be easily obtained through minimally invasive procedures such as a skin biopsy or blood sample, and converted to pluripotency through application of a cocktail of reprogramming factors to create iPSCs. Initially, a set of four programming factors, Oct3/4, Klf-4, Sox2, and c-Myc, were delivered to cells using lentiviruses which stably integrated the reprogramming factor genes into the iPSC genome, potentially altering disease phenotypes and also abrogating expression of native genes at the DNA sites where the factors integrated. Newer techniques have been developed that use non-integrating approaches such as through the use of Sendai virus, messenger RNA (mRNA), or episomal vectors that circumvent these problems. Once created, iPSC lines can be expanded indefinitely to produce a limitless supply of stem cells. These cells are the starting material for the derivation of specific cell types based on protocols that use small molecules, proteins, or direct gene induction to recapitulate developmental programs. Most current protocols derive neuronal progenitors through “dual-SMAD inhibition,” a step that involves the use of small molecule inhibitors to block endoderm and mesodermal cell fates, thereby creating neural cells by default. Multiple protocols have been developed over the last decade for creating large numbers of human neuron progenitor cell types and directing them toward specific nervous system cell fates, including neuron subtypes from multiple regions of brain and spinal cord as well as retinal cells, glial cells including astrocytes and oligodendrocytes, immune cells, and PNS cells.

The primary medical benefit of iPSC technology is that it enables the creation of patient-specific cells or tissues that are genetically matched to individual patients. This approach not only enables the study of monogenetic disorders, but also sporadic forms of disease, and complex polygenic disorders including those with unidentified risk loci. Furthermore, by deriving iPSC cell lines from multiple patients it would be possible to explore how disease phenotypes may vary according to genetic background. Another approach that has been used to generate specific neuron and glial cell types from somatic cells such as fibroblasts is through direct reprogramming. This approach relies on a cocktail of specific transcription factors to directly convert somatic cells into the alternate desired cell type. This approach bypasses the epigenetic reset that accompanies cells as they are reprogrammed to a pluripotent state. The advantage of this approach is that age-related epigenetic signatures are not erased, so that derived neurons may more readily reflect diseases that manifest in older cells.

Despite the advantages of using *in vitro* models of nervous system diseases derived from patient-specific iPSCs, several potential roadblocks remain. There are no standard reprogramming or derivation protocols, and the different methods can result in considerable variability in the disease phenotypes reported by different laboratories. Confidence in the specificity of a particular phenotype is therefore increased if it has been validated across multiple laboratories. There is also the problem of inherent variability between patient lines that may result from their different genetic backgrounds. One solution, available only in the case of monogenic disorders, is to use isogenic controls generated using gene editing, such as with CRISPR-Cas9 technology, to create disease and control lines on an identical genetic background. However, because differences in genetic background can influence the penetrance of a particular trait, it will still be necessary to compare disease lines from multiple patients to discern a true disease phenotype. For polygenic disorders where the causative mutations are unknown it will not be possible to create isogenic controls, and in these situations the best strategy for improving reliability and sensitivity is to compare lines from multiple patients.

Organoids

Most nervous system disorders, including autism spectrum disorder, schizophrenia, PD, AD, and ALS are complex disorders, resulting from an unknown combination of gene mutations and manifest not only in specific cell types, but also in alterations of the local tissue environment. These disorders are difficult to model in animals, but they are approachable using three-dimensional human iPSC stem cell models, often referred to as “organoids.” Organoids are derived from pluripotent stem cells that are directed along a tissue-specific lineage through the timed application of growth factors, genes, or small molecule activators or inhibitors, and allowed to aggregate into three-dimensional structures. With time, cell intrinsic programs are spontaneously engaged and the cellular aggregates begin to self-organize and develop into structures that recapitulate the complex topographical and cellular diversity of normal organ development. In this way it has been possible to create, at least in part, in vitro brain-like organoids that resemble the human forebrain at early stages of development. These structures, when allowed to develop from an anterior neural tube stage, can become heterogeneous containing regions with forebrain, midbrain, and/or hindbrain identity and can often include retina-like structures. The high degree of variability in such “cerebral organoids” can be a liability for controlled studies, and can be reduced by the use of more directed protocols that restrict outcomes to more defined brain regions, such as forebrain, cortex, or ganglionic eminence. A variety of protocols have now been developed to generate organoids with specific regional identity, and fusing organoids of different regional identity with each other has been used to reproduce cellular interactions such as neuronal migration across regions. Many protocols are focused on modeling cortical development, and they can reproduce developmental features including a diversity of progenitor and neuronal cell types topographically distributed within ventricular and subventricular progenitor regions and rudimentary cortical layers. However, the organoids follow a human developmental timetable and still remain at stages roughly comparable to late fetal development after 6–9 months. Moreover they lack key cell types such as endothelial cells, pericytes, microglia, and have few if any astrocytes or oligodendrocytes. Nonetheless, while still only reflecting rudimentary organizational and compositional features, organoids have become attractive models to study human brain development and the pathophysiology of human nervous system diseases in the context of an organized brain-like structure.

Brain Development and Developmental Disorders: Microcephaly and Lissencephaly

Transcriptional analysis has suggested that the neurons produced by most stem cell protocols resemble early to mid-gestational stages of human brain development. The immaturity of stem cell-derived human neurons may limit their utility for modeling adult diseases, but makes them ideally suited for the study of brain development and the pathophysiology of neurodevelopmental disorders.

Primary autosomal recessive microcephaly (MCPH) is a rare neurodevelopmental disorder producing severe microcephaly with simplified cortical gyration and intellectual disability. MCPH was one of the first disorders to be studied using cerebral organoids. Mutations in genes encoding microtubule spindle components and spindle-associated proteins are the most frequent causes of congenital microcephaly. Among them is cyclin-dependent kinase 5 related activator protein 2 (CDK5RAP2). Skin fibroblasts derived from a single microcephalic patient carrying a mutation in CDK5RAP2 were used to generate four iPSC lines. Cerebral organoids grown from these cell lines contained fewer proliferating progenitor cells and showed premature neural differentiation compared to wild type controls. Introducing functional CDK5RAP2 by electroporation partially rescued the disease phenotype, supporting the notion that failure of the founder population of neural progenitors to properly expand underlies the smaller brain. This study demonstrated that brain organoids derived from patients with microcephaly can be

used to reproduce features of the disease, but did not reveal new insights or disease features of CDK5RAP2 microcephaly that had not already been described in mouse models.

In a study using cortical organoids to model Miller-Dieker syndrome (MDS), a severe congenital form of lissencephaly or “smooth-brain,” features of the human disease were observed that had not been noted in murine models. Classical lissencephaly is a genetic neurological disorder associated with mental retardation and intractable epilepsy, and MDS is a severe form of the disorder. Cortical folding in humans begins toward the end of the second trimester, a stage of development that has not yet been modeled in organoids, but gyrencephaly depends upon earlier events such as neural progenitor cell proliferation and neuronal migration that can be modeled in organoids. The human organoid model of MDS exhibited several neural progenitor cell phenotypes that had already been reported in mouse models, including altered mitotic spindle orientation and neuronal migration defects. But the organoids also displayed a mitotic defect in a specific neural stem cell subtype, the outer radial glia cell (oRG) that had not been observed in mice. oRG cells are enriched in the outer subventricular zone, a proliferative region that is large in primates and not present in rodents. These cells are particularly numerous in the developing human cortex and are thought to underlie the developmental and evolutionary expansion of the human cortex. oRG cells from MDS patients behaved abnormally and had arrested or delayed mitoses. MDS organoids also identified non-cell autonomous defects in WNT signaling as an underlying mechanism. These insights into mechanistic and cell type specific features of human disease highlight how organoid technology can provide new and valuable perspectives on the pathophysiology of disorders of in utero development.

Acquired Neurodevelopmental Disorders: Zika

The recent outbreak of Zika virus (ZIKV) and associated microcephaly cases in the Americas provided a test case for the utility of brain organoids to model acquired human microcephaly. Despite a correlation between Zika infection rates and the incidence of congenital microcephaly, compelling evidence that ZIKV caused microcephaly was lacking in the early phases of the epidemic. The causal link between ZIKV and congenital microcephaly was buttressed by two studies in 2016 that used human iPSC-derived neural progenitor cells and organoids to demonstrate ZIKV tropism for human neural progenitor cells. Neural progenitor cells (radial glia) were readily infected in vitro with subsequent progenitor cell death and involution of organoid size. Forebrain organoids were further used to highlight the role of the flavivirus entry factor, AXL, in determining viral tropism, and were also used to explore the disease mechanism by demonstrating upregulation of the innate immune receptor toll-like receptor 3 (TLR) in response to ZIKV infection. Stem cell-derived models of human brain development have also demonstrated centrosomal abnormalities in radial glia and alteration in the cleavage plane of mitotic radial glia associated with premature neural differentiation. Mouse models are also being used to study the pathophysiology of congenital ZIKV syndrome, but the availability of unlimited numbers of human neural cells produced using stem cell technology has enabled high-throughput screening assays to test libraries of clinically approved compounds for potential therapeutic agents. This strategy has already highlighted several compounds that could potentially help protect against ZIKV microcephaly.

Neurodevelopmental Disorders: Autism and Schizophrenia

Autism spectrum disorders (ASD) are complex and heterogeneous neurodevelopmental disorders usually manifesting in childhood with difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors. The cellular and molecular mechanisms underlying ASD are thought to arise at stages of fetal brain development, making them well-suited for exploration using human iPSC-derived disease models. The pathophysiology of disorders associated with ASD that are caused by monogenic mutations have been studied using iPSC-derived neurons; these include Fragile X, Rett, and Timothy syndromes.

Fragile X is the most common heritable cause of intellectual disability, affecting 1 in 4000 males and 1 in 8000 females, and is a leading genetic cause of ASD. Patients also have speech delay, growth and motor abnormalities, hyperactivity, and anxiety. The causative mutation lies in the *FMR1* gene and produces a CGG triplet repeat expansion from a normal number of 5–20 to >200, leading to epigenetic silencing of the *FMR1* gene and loss of the Fragile X mental retardation protein. The epigenetic mechanism means that unlike a simple gene deletion that would lead to ubiquitous loss of expression, the FMR1 locus becomes hypermethylated and epigenetically silenced during differentiation, thus FMR1 protein is expressed by the early embryo and becomes absent only around the beginning of the second trimester. Interestingly, this expression pattern is recapitulated during cellular differentiation in stem cell models. Pluripotent Fragile X stem cell lines have been derived from embryos identified through pre-implantation genetic diagnosis and by reprogramming skin fibroblasts from Fragile X patients to create iPSC lines. In both cases, *FMR1* was expressed by the pluripotent stem cells, but underwent transcriptional silencing following differentiation. Fragile X stem cell lines can therefore be used to study the mechanism of FMR1 silencing, an effort that is ongoing. Neurons generated from Fragile X iPSC cells reproduce features observed in neurons from transgenic *FMR1* mouse models and patients, including stunted neurites with decreased branching, increasing confidence in the iPSC model. In addition to providing a model that can be used to study disease pathogenesis, Fragile X iPSC-derived neurons could be used to screen for potential therapeutic agents or gene editing strategies that could be able to remove the repressive epigenetic marks induced by the mutation and rescue the phenotype.

Rett syndrome is an X-linked neurodevelopmental disorder with dominant inheritance caused by a mutation in the MECP2 gene. Because males carrying one copy of the defect gene usually die in infancy, most patients are girls. Random inactivation of the X chromosome in girls results in mosaic cellular expression of the mutation that circumvents fatality and produces a variable phenotype. The symptoms are present in early childhood and include microcephaly associated with developmental delay, autistic-like behaviors and cognitive dysfunction, seizures, and repetitive motor actions; these then progress to include difficulties with gait, swallowing, and breathing before usually stabilizing with patients surviving to adulthood. The pathophysiology of RETT syndrome is presumed to involve abnormal epigenetic regulation leading to decreased transcriptional repression of genes whose overexpression produces the disease phenotype, although this concept has been contested. In one of the first studies to use iPSC modeling to study RETT syndrome, it was discovered that when fibroblasts from patients were reprogrammed to pluripotent stem cells, X inactivation was erased. In apparent recapitulation of endogenous events, X chromosome inactivation re-occurred during neuronal differentiation, producing a mosaic of cells carrying the mutant gene intermingled with normal cells. RETT neurons had fewer dendritic spines and synapses, smaller cell bodies, and reduced network activity. Another iPSC model of RETT syndrome highlighted the potential role of altered inhibitory function. RETT neurons were found to have a deficit of a potassium/chloride cotransporter (KCC2) that is developmentally regulated and normally leads to a switch in GABA signaling from excitatory at embryonic ages to inhibitory by birth. In RETT neurons KCC2 expression level

was low, and the functional switch in GABA effects was delayed, contributing to some of the disease features and possibly accounting for the developmental onset of the disease. One curious feature of some iPSC RETT lines was that despite the mosaic expression of the mutation, disease phenotypes were observed in all cells. Possibly, this could reflect a non-cell autonomous effect, but as in all iPSC disease models, confidence in disease-specific features will be increased when similar phenotypes are seen across multiple independent studies.

Timothy syndrome, another severe neurodevelopmental disease associated with ASD has been modeled using iPSC-derived organoids. Timothy syndrome is caused by a mutation in the *CACNA1C* gene coding for a voltage-gated calcium channel, and neuron defects in Timothy syndrome organoids were rescued by selectively altering calcium channel activity. In one study two separate organoids were produced with different regional identity, one represented neocortex and one a more ventral structure known as the medial ganglionic eminence, which is the source of most cortical interneurons. The two organoids were then fused together to allow the interneurons to migrate into the cortex, mimicking their endogenous behavior. The ability to model interneuron migration led to the discovery of a cell-autonomous migration defect in the disease-carrying neurons.

The majority of nervous system diseases, including ASD, are multigenic and cannot be modeled in animals but can be modeled using patient-derived iPSCs. For example, a subset of patients with ASD have large head size, and a cohort of patients with this phenotype were used to generate iPSCs which were converted to neural progenitor cells and forebrain neurons. The progenitors had an accelerated cell cycle and produced an excess of inhibitory interneurons and had exuberant cellular overgrowth of neurites and synapses. This last feature is in contrast to the decrease in spines and synapses observed in other iPSC models of ASD such as Fragile X and Rett syndrome and underscores the need for replication and validation of purported disease phenotypes given the high variability based on differences between stem cell lines, protocols, patient genetic background, and other factors. Moreover, the clinical features of most neuropsychiatric diseases reflect disorders in processes such as circuit formation and refinement that occur after birth and may be difficult to capture at the fetal stage of development reflected in stem cell models.

Patient stem cells have also been used by multiple groups to study the pathophysiology of schizophrenia, producing a variety of diverse and sometimes contradictory results. Reports claim obvious phenotypes such as disruptions in the adherens junctions of forebrain radial glia or aberrant neuronal migration, although such gross abnormalities observed at the equivalent of in utero stages of development seem very unlikely to underlie a disease that usually manifests at adolescence or young adulthood. Other studies report abnormalities related to abnormal microRNA expression, disordered cyclic AMP and Wnt signaling, abnormal stress responses, diminished neuronal connectivity, fewer neuronal processes, problems with neuronal differentiation, and mitochondrial abnormalities among others. While the pathophysiology of as complex a neurodevelopmental disorder as schizophrenia may be multidimensional, it is unclear which, if any, of the reported findings in iPSC models reflect the true pathology of schizophrenia. Progress will likely depend on the adoption of more standard and reproducible protocols, more rigorous identification of cell types, markers of regional identity, and indicators of maturity.

Alzheimer's Disease

As noted above, the leading concept of AD pathogenesis, the amyloid hypothesis, suggests that an imbalance between production and clearance of β-amyloid leads to excessive accumulation of β-amyloid peptide and the

formation of neurofibrillary tangles within neurons, composed of aggregated hyperphosphorylated tau proteins. Additionally, aggregates of amyloid fibrils are deposited outside neurons in the form of neuritic plaques. Recent failures of anti-β-amyloid therapies, which were highly effective in mouse models, have led to a search for alternative models that might be more predictive of therapeutic effectiveness in humans. Among the causes of familial AD are mutations in genes involved in β-amyloid production, including APP and presenilin 1 and 2. Shortly after the introduction of iPSC technology, human stem cell-derived neurons were generated from patients carrying mutations in AD causative genes as well as from sporadic AD cases. The disease neurons developed hallmarks of AD including intracellular accumulation of β-amyloid and phosphorylated tau, as well as secretion of APP cleavage products, features that could be reduced by adding β- or γ-secretase inhibitors or β-amyloid specific antibodies. The neurons also demonstrated other disease features observed in postmortem AD tissues. However, extracellular β-amyloid aggregation and neurofibrillary tangles were not robustly modeled in these two-dimensional systems, presumably because secreted factors were able to readily diffuse away. The use of three-dimensional organoids to model AD overcame this limitation, presumably by recreating a more faithful extracellular matrix. Organoid models promoted the aggregation of β-amyloid, and more readily recapitulated the pathologic features of AD, including the formation of neurofibrillary tangles and neuritic plaques.

It is hoped that the new stem cell models, particularly organoid models, will accelerate our understanding of AD by enabling the study of human disease-carrying cells in a quasi in situ setting. These new models may lead to discovery of novel druggable targets and new diagnostic and prognostic biomarkers. One concern is that the pathogenic features of AD usually appear in the sixth or seventh decade of life and progress slowly over years, while most protocols for the derivation of human cortical neurons generate cells over weeks or months and most remain comparable to immature neurons at fetal stages of development. Nonetheless, these young cells have been used to model neurodegenerative diseases such as AD and HD that strike patients in mid to late adulthood. Possibly the onset of disease phenotype is accelerated in stem cell models due to increased cellular stress, or disease features may actually have a subtle onset at earlier stages than generally suspected. Indeed, 3-year-old children at genetic risk of developing early-onset AD appear to have smaller hippocampal size and lower scores on memory tests than children in a non-risk group. The phenotypes of adult neurodegenerative diseases that are visible at fetal stages may or may not correspond to those manifest at later, adult stages, but they may offer the possibility of devising preventative strategies effective at very early stages of the disease.

Cell Type Disorders: ALS and HD

In diseases such as ALS, PD, and HD, that mostly target specific neuron subtypes, stem cells provide an ideal means to study the vulnerable human cell populations. By enabling the production of unlimited numbers of normal and diseased human midbrain dopaminergic neurons for the study of PD, medium spiny striatal neurons for HD, and spinal and cortical motor neurons for ALS, iPSC approaches have the potential to transform our understanding and management of these diseases. Stem cell-derived neurons serve as platforms to explore mechanisms of cell vulnerability, to screen drugs for neural protection, and potentially to derive neurons for replacement therapy.

Amyotrophic Lateral Sclerosis

One of the first protocols for producing neurons of a specific subtype from embryonic stem cells recapitulated normal developmental programs to generate mouse spinal motor neurons. Pluripotent mouse stem cells underwent neural induction and adopted a caudal identity through the application of retinoic acid, and subsequently adopted motor neuron fate through the action of sonic hedgehog, a ventralizing factor. Generating human motor neurons proved more complex, requiring additional steps, such as early exposure to the growth factor, FGF2. The first application of stem cell-derived motor neurons to study ALS involved the use of mouse motor neurons generated from transgenic mice expressing a mutation in the SOD1 gene, the most common mutation responsible for familial ALS. Only 5–10% of ALS cases are familial, but the known mutations provide a useful entry point to tease apart the causative pathophysiology. Mutations in SOD1 produce ALS through a toxic gain of function for which the mechanism remains unclear, despite the use of multiple transgenic animal and iPSC models. The use of mouse ESC-derived motor neurons, however, demonstrated that toxic factors secreted by SOD1 astrocytes contribute to the death of motor neurons. Interestingly, stem cell-derived interneurons were spared, indicating a specific vulnerability of motor neurons. These findings helped establish the notion that a non-cell autonomous toxic mechanism contributes to ALS pathogenesis and may ultimately lead to novel treatment strategies. These findings also highlight that modeling the full pathophysiology of ALS may require the reproduction of a complex environment including motor neurons, astrocytes, and possibly additional cell types such as microglia. A variety of approaches including co-culture of specific cell types, three-dimensional spinal cord organoids, and microfluidic organ-on-chip models are being explored to achieve a more complete facscimile of spinal cord organization. Similar to other neurologic disorders where a clearly defined phenotype has been observed in human stem cell-derived models, there is hope that drug screening using human disease-expressing cells will identify a potential therapeutic compound.

Huntington's Disease

HD is caused by an expansion in CAG triplet repeats in the huntingtin gene which leads to an expanded polyglutamine tract in the huntingtin protein. HD is dominantly inherited, with symptoms of cognitive decline and uncontrollable gait and limb motions beginning in the third to fifth decade of life with progression to dementia and death ~20 years later. Mutant huntingtin causes a toxic gain-of-function, with the degree of effect related to the CAG repeat length. For example, a CAG length of 40–60 repeats produces adult onset HD, while repeats of ≥60 produce juvenile onset disease. Although it has been 25 years since the discovery of this causative mutation, the disease mechanism remains poorly understood. Excess huntingtin protein and protein fragments accumulate in specific subtypes of neurons where they misfold and form aggregates that are visible as cellular inclusions. Affected cells eventually die, possibly as a result of metabolic toxicity. The medium spiny neurons of the striatum are the most vulnerable neurons, spurring ongoing attempts to produce replacement cells derived from stem cells, but neuron loss is widespread including in the cortex, complicating a cell replacement approach for this disease. HD iPSCs have been generated from patients with various CAG repeat lengths, but those from juvenile onset disease with the longest repeat lengths have been favored as being most likely to express robust disease phenotypes at an early stage. This is particularly important given the immature stage of maturation of stem cell-derived human neurons. This approach has been able to produce disease phenotypes observed in patients including huntingtin protein aggregation, decreased metabolic capacity, increased oxidative stress with mitochondrial fragmentation, and apoptosis enhanced by withdrawal of growth factor support. However, many of these phenotypes were observed in pluripotent cells prior to neural differentiation and in neural progenitors and a broad array of CNS neurons in contrast to the cell type specific features of the disease. Nonetheless, neurons

that assumed striatal fate appear to be more vulnerable to stress and apoptosis than other cell types. As with other iPSC models of nervous system diseases, there have so far been few efforts to validate results in multiple iPSC lines having different genetic backgrounds but with similar CAG repeat lengths. An HD consortium has been formed to address this problem by generating a series of iPSC lines from multiple patients. An alternative strategy to validate disease phenotypes has been to use gene editing to create isogenic iPSC lines that are corrected to produce wild type control and HD iPSC lines against the same genetic background.

Future Perspectives

Despite early successes, it may prove difficult to reconstitute neurodegenerative disease conditions in human cells in vitro over a short time-course because the pathogenic changes of degenerative diseases progress slowly and commence in the later stages of life. The differentiation and maturation of human neurons from stem cell lines occurs over a span of months, which may not be long enough to establish the aged brain conditions under which patients develop robust neurodegenerative pathology. Possible manipulation through gene editing or by application of aging-associated stresses, such as DNA damaging agents or proteasome inhibitors, may accelerate the expression of degenerative phenotypes in human iPSC-derived cellular models. Stem cell-derived organoid models are also ideal platforms to apply methods for cellular level visualization such as clarity and multi-electrode recording techniques to better evaluate three-dimensional organoid structures and explore early-forming circuits. These applications are only just beginning.

Two-dimensional cell cultures are ideal for production and evaluation of large numbers of specific cells of a particular identity, but may not provide the complex extracellular environment necessary to model certain disease processes, such as extracellular protein aggregation. These features can be best modeled using three-dimensional organoids, but current methods do not reproduce all the relevant features of brain tissue. Optimization will be needed to better reproduce the cellular composition of brain, including endothelial cells, astrocytes, microglia, and oligodendrocytes. It may also be necessary to combine different brain regions generated separately, possibly by fusion of tissues such as dorsal cortex, subpallium, thalamus, retina, and others. However, currently there is a limited ability to recreate tissues or neurons with regional brain identity, such as hippocampus, thalamus, or cerebellum. More faithful organoid models could also emerge through the application of bioengineered scaffolds, matrices, or perfusion systems that might allow the growth of larger structures. Of course, not all aspects of mature brain architecture and function will be modeled by these tissue structures, particularly as they represent fetal stages of development, but perhaps the most precocious events in disease etiology can be captured and investigated and these may share mechanistic pathways with disease features that manifest at later stages.

The current excitement surrounding human stem cells has more to do with their promise to improve on animal models of disease than their potential as a source for cell based therapies. Even without new insights into disease pathogenesis, there is promise that iPSC models such as brain organoids will act as drug screening platforms for discovery of novel therapeutics and for detection of off-target and toxic effects. The failure of many neurotherapeutic approaches to translate from animal models to clinical practice underscores the need for better predictive models, and stem cell models and brain organoids based on human cells may be ideally suited to bridge this divide.

A Current Perspective on Neural Stem Cells in the Clinic

The prospect of stem cell therapies to treat diseases or injuries of the nervous system has captured the attention of researchers, clinicians, and the public. The pace of research is usually slow and deliberate, but in the stem cell arena there has been enormous pressure to accelerate the pace of progress in order to bring cell-based therapies to the clinic. Expectations have been raised, and clinics have already begun offering unproven or dangerous treatments to a public that is ill-informed and vulnerable to exploitation. Nonetheless, there is cautious optimism that stem cells will eventually realize the promise of regenerative therapy for at least some currently untreatable or incurable nervous system diseases.

Pursuit of a cell-based therapy for PD has been ongoing for many decades. Following anecdotal success in a handful of patients who appeared to improve following striatal grafts of fetal midbrain dopaminergic cells, two NIH-funded double-blind control studies were launched in the 1990s. However, only a small number of younger patients showed some benefit, and several patients developed spontaneous dyskinetic movements related to the therapy. These efforts constituted a failed trial as the treated patients who did not experience side effects failed to improve significantly. The dyskinesias that curtailed the trials were eventually ascribed to an abundance of serotonergic neurons that were inadvertently included in some of the cell grafts. Protocols for deriving dopaminergic neurons from stem cells could potentially avoid this complication by providing a more purified cell population, and several groups in the United States and Europe have been aggressively pursuing a stem cell-based approach and are nearing clinical trials. Meanwhile, techniques to extract dopaminergic cells from fetal tissue have been improved, and on the basis of encouraging results in individual transplanted patients, some of whom have managed to go off their Parkinson's medication, a new trial of fetal cell transplantation for PD has started in Europe. This is a very consequential trial, as a poor clinical outcome could dampen enthusiasm for the planned follow-on stem cell trials in PD and possibly in other disorders as well.

One of the first cell-based clinical trials for a neurological disease targeted patients suffering from an untreatable childhood disorder, Batten disease. Batten disease is an autosomal recessive metabolic disorder resulting from an inability to synthesize a lysosomal enzyme critical to brain function. The Phase 1 trial involved six patients with infantile and late infantile forms of the disease who received neural stem cells rather than any specific postmitotic cell type. Neural stem cells derived from donated fetal tissue were expanded in vitro prior to surgical grafting into the brain. This approach was not without risk, as the neural stem cells were proliferating and could potentially form an abnormal growth. The rationale was that the cells would be capable of synthesizing and secreting the missing lysosomal enzyme and would therefore serve as a delivery device. Animal studies using a transgenic mouse model of Batten disease demonstrated rescue, and this promising result led to a small Phase 1 trial. The Phase 1 study was considered a success as no adverse events were reported and the cells appeared to be safe, though there was no clinical improvement and no clear evidence of whether the cells had dispersed, transformed into neurons or glia, or indeed survived at all. Despite clearing the Phase 1 trial, the company did not pursue further trials for Batten disease, but instead initiated clinical trials using the same cell product for several other indications, including an inherited fatal dysmyelination syndrome known as Pelizaeus-Merzbacher disease (PMD). The human neural stem cells have both neurogenic and gliogenic potential, and when delivered to white matter regions in experimental animals most persisting cells had become oligodendrocytes. This supported use of the cells to promote myelin formation in conditions such as PMD. The company also initiated trials in spinal cord injury. However, the spinal cord trial failed to achieve sufficient benefit in Phase II and the company ceased its work on stem cell therapies.

Spinal cord injury is an attractive target for novel therapies since there are no effective treatment options currently. A series of stem cell trials designed to treat subacute spinal cord injury are underway in the United States and Europe. The first to enter clinical trials in the United States was based on a protocol designed to generate oligodendrocytes from pluripotent embryonic stem cells. Evidence of efficacy was obtained in animal models following surgical grafting of cells to sites of spinal cord injury. However, evidence of myelination of host axons was minimal, and other mechanisms were invoked for improvement in gait, including trophic support and immune modulation. Regulatory permission for a Phase 1 trial for subacute mid-thoracic injury was initially stalled by concern over abnormal growths at sites of cell deposit in some animals, but this was satisfactorily addressed and patient trials commenced. However, following a change in leadership, the stem cell program was terminated. The program was acquired by another company that has resumed the spinal cord injury trial and received regulatory approval to advance to include cervical level injuries.

The possibility of treating ALS by replacing dying motor neurons with stem cell–derived substitutes has excited interest but this prospect seems very remote. Even if new neurons are able to integrate into spinal cord circuits and become properly innervated, they would have to grow long axons that would take many months to years to project to appropriate targets and attract myelinating Schwann cells. Furthermore, cells would need to be grafted at multiple spinal cord and brainstem levels, and the upper motor neuron deficit would need to be treated by replacing projecting neurons in the motor cortex. An additional complication is the recent finding that spinal motor neurons have unique segmental identity, and replacement cells might need to be generated with a range of molecular identities in order to integrate at multiple spinal levels. This would still leave unaddressed the toxic effects recently shown to be produced in ALS by diseased astrocytes and microglia that could attack the replacement cells. A more tractable near-term solution would be to graft support cells that could rescue or protect endogenous motor neurons from damage. This approach was tried in a mouse model of ALS. Human stem cell–derived neural progenitor cells engineered to express GDNF, a growth factor known to provide trophic support for neurons, were grafted to the spinal cord of young ALS mice. The cells dispersed and were able to rescue motor neurons, a very promising result, but disappointingly, the animals became weak and died at the same rate as untreated control animals. However, ALS is a deadly disease with no known treatment. In the hope that patients will respond differently than mice, a clinical trial based on this approach has been approved by the U.S. Food and Drug Administration (FDA) and will begin soon.

Following Shinya Yamanaka's discovery of iPSCs, the Japanese government has invested in bringing iPSC-derived cell therapy to the clinic. Banks of iPSC lines selected to capture the diversity of HLA haplotypes found in the Japanese population are being produced in the hope that these will allow cell therapies to be matched to individual patient haplotypes in order to avoid immune rejection. While these stem cell banks were still being produced, the first Japanese study to use stem cells was approved in August, 2013, and involved patients who were to receive customized therapy using cells derived from their own skin fibroblasts. The targeted disease was age-related macular degeneration, a common cause of blindness in the elderly that results from loss of retinal pigment epithelial (RPE) cells. RPE cells are relatively easy to generate from pluripotent stem cells, making replacement therapy an attractive target in this condition. A challenge is to coax the replacement cells to recreate an epithelium in the subretinal space. The Japanese approach involves surgical insertion of a biofilm seeded with RPE cells into the retina. One patient was treated with his/her own stem cell-derived RPE cells, but prior to treating a second patient, the genome of the RPE cell line was sequenced, and a mutation was discovered in a known oncogene. The trial was halted and a decision made to discontinue the effort for customized cell therapy in

favor of using RPE cells derived from the national repository of banked iPSC lines which undergo extensive gene sequencing and quality controls. This outcome serves as a caution for the challenges involved in bringing a customized cell therapy to the clinic.

By far the largest number of human trials have been performed using mesenchymal stem cells (MSCs) sourced from a variety of sites including bone marrow, peripheral blood, adipose tissue, umbilical cord, etc. Interest in the potential utility of MSCs for regenerative therapy began with the optimistic report that bone marrow stem cells were pluripotent and capable of generating nerve and heart muscle as well as blood cells. The possibility that easily obtainable MSCs could be used to regenerate injured or diseased cells or organs to treat diseases ranging from stroke, neurodegenerative disease, myocardial infarct, and even diabetes, generated enormous enthusiasm. The enthusiasm proved irresistible to many, and even after the initial reports were discredited—MSCs turned out not to be pluripotent stem cells as initially thought—a veritable flood of papers began to appear claiming disease-modifying activity of MSCs in mouse models of almost every degenerative disease and injury model. But when it became clear that the MSCs were not transforming into or generating new neurons or cardiac myocytes, alternative mechanisms of action were invoked, including the release of trophic factors, cytokines, or inflammatory modulators that were credited with producing their remarkable restorative effects. The relative ease with which blood or adipose tissue can be harvested from patients or donors and MSCs extracted has led to a rapidly expanding number of clinical trials for conditions ranging from stroke and MS to AD and PD. Furthermore, a loophole in the regulatory framework of the FDA allows autologous cell therapy to escape regulation provided that the cells have not been significantly processed. This lax regulation has spawned a veritable industry of stem cell clinics making unsubstantiated claims of success in treating nervous system diseases. Patients have died from treatments in unregulated clinics operating in countries around the world and three patients became blind in a well-publicized incident following stem cell treatments delivered by a Florida clinic. The “stem cells” were derived from the patients’ own fat tissue and blood. These activities represent the dark side of the stem cell revolution perpetrated by practitioners who exploit the desperation of patients and their families. Legitimate and effective stem cell therapies will emerge over time, but given the prevalence and abundance of misleading information available on the internet and elsewhere, a trusted and well-informed physician can play a key role in helping patients navigate the current cell therapy minefield.

FURTHER READING

Chung WS et al: Do glia drive synaptic and cognitive impairment in disease? Nat Neurosci 18:1539, 2015. [PubMed: 26505565]

Di Lullo E, Kriegstein AR: The use of brain organoids to investigate neural development and disease. Nat Rev Neurosci 18:573, 2017. [PubMed: 28878372]

Fung TC et al: Interactions between the microbiota, immune and nervous systems in health and disease. Nat Neurosci 20:145, 2017. [PubMed: 28092661]

Herz J et al: Myeloid cells in the central nervous system. Immunity 46:943, 2017. [PubMed: 28636961]

Hong S et al: Complement and microglia mediate early synapse loss in Alzheimer mouse models. Science 352:712, 2016. [PubMed: 27033548]

Liddelow SA et al: Neurotoxic reactive astrocytes are induced by activated microglia. Nature 481:541, 2017.

Prusiner SB et al: Evidence for α-synuclein prions causing multiple system atrophy in humans with parkinsonism. Proc Natl Acad Sci USA 112:E5308, 2015. [PubMed: 26324905]

Turner L, Knoepfler P: Selling stem cells in the USA: Assessing the direct-to-consumer industry. Cell Stem Cell 19:154, 2016. [PubMed: 27374789]

Yoo BB, Masmanian SK: The Enteric Network: Interactions between the immune and nervous systems of the gut. Immunity 46:910, 2016.

Chapter 418: Seizures and Epilepsy

Daniel H. Lowenstein

INTRODUCTION

A *seizure* (from the Latin *sacire*, "to take possession of") is a transient occurrence of signs or symptoms due to abnormal excessive or synchronous neuronal activity in the brain. Depending on the distribution of discharges, this abnormal brain activity can have various manifestations, ranging from dramatic convulsive activity to experiential phenomena not readily discernible by an observer. Although a variety of factors influence the incidence and prevalence of seizures, ~5–10% of the population will have at least one seizure, with the highest incidence occurring in early childhood and late adulthood.

The meaning of the term *seizure* needs to be carefully distinguished from that of epilepsy. *Epilepsy* describes a condition in which a person has a risk of *recurrent* seizures due to a chronic, underlying process. This definition implies that a person with a single seizure, or recurrent seizures due to correctable or avoidable circumstances, does not necessarily have epilepsy (although a single seizure associated with particular clinical or electroencephalographic features may establish the diagnosis of epilepsy). Epilepsy refers to a clinical phenomenon rather than a single disease entity, because there are many forms and causes of epilepsy. However, among the many causes of epilepsy there are various *epilepsy syndromes* in which the clinical and pathologic characteristics are distinctive and suggest a specific underlying etiology.

Using the definition of epilepsy as two or more unprovoked seizures, the incidence of epilepsy is ~0.3–0.5% in different populations throughout the world, and the prevalence of epilepsy has been estimated at 5–30 persons per 1000.

CLASSIFICATION OF SEIZURES

Determining the type of seizure that has occurred is essential for focusing the diagnostic approach on particular etiologies, selecting the appropriate therapy, and providing potentially vital information regarding prognosis. The International League Against Epilepsy (ILAE) Commission on Classification and Terminology provided an updated approach to classification of seizures in 2017 **(Table 418-1)**. This system is based on the clinical features of seizures and associated electroencephalographic findings. Other potentially distinctive features such as etiology or cellular substrate are not considered in this classification system, although this will undoubtedly change in the future as more is learned about the pathophysiologic mechanisms that underlie specific seizure types.

TABLE 418-1

Classification of Seizures*

1. **Focal Onset**
   (Can be further described as having intact or impaired awareness, motor or nonmotor onset, or evolve from focal to bilateral tonic clonic)
2. **Generalized Onset**
   a. Motor
      Tonic-clonic
      Other motor (e.g., atonic, myoclonic)
   b. Nonmotor (absence)
3. **Unknown Onset**
   (Can be further described as motor or nonmotor, or unclassified)

*Based on the new 2017 International League Against Epilepsy classification of seizure types (RS Fisher et al: Epilepsia 58: 522, 2017).

A fundamental principle is that seizures may be either focal or generalized. *Focal seizures* originate within networks limited to one brain region (note that the term *partial seizures* is no longer used). *Generalized seizures* arise within and rapidly engage networks distributed across both cerebral hemispheres. Focal seizures are usually associated with structural abnormalities of the brain. In contrast, generalized seizures may result from cellular, biochemical, or structural abnormalities that have a more widespread distribution. There are clear exceptions in both cases, however.

FOCAL ONSET SEIZURES

Focal seizures arise from a neuronal network either discretely localized within one brain region or more broadly distributed but still within a cerebral hemisphere. With the new classification system, the subcategories of “simple focal seizures” and “complex focal seizures” have been eliminated. Instead, the classification emphasizes the effect on awareness (intact or impaired) and nature of the onset (motor or nonmotor). Focal seizures can also evolve into generalized seizures. In the past this was referred to as *focal seizures with secondary generalization*, but the new system relies on descriptions of the type of generalized seizures that evolve from the focal seizure.

The routine interictal (i.e., between seizures) electroencephalogram (EEG) in patients with focal seizures is often normal or may show brief discharges termed *epileptiform spikes*, or *sharp waves*. Because focal seizures can arise from the medial temporal lobe or inferior frontal lobe (i.e., regions distant from the scalp), the EEG recorded during the seizure may be nonlocalizing. However, the region of seizure onset may be detected using surgically placed intracranial electrodes.

Focal Seizures with Intact Awareness

Focal seizures can have motor manifestations (such as tonic, clonic, or myoclonic movements) or nonmotor manifestations (such as sensory, autonomic, or emotional symptoms) without impairment of awareness. For example, a patient having a focal motor seizure arising from the right primary motor cortex near the area controlling hand movement will note the onset of involuntary movements of the contralateral, left hand. Since the cortical region controlling hand movement is immediately adjacent to the region for facial expression, the seizure may also cause abnormal movements of the face synchronous with the movements of the hand. The EEG recorded with scalp electrodes during the seizure (i.e., an ictal EEG) may show abnormal discharges in a very limited region over the appropriate area of cerebral cortex if the seizure focus involves the cerebral convexity.

Three additional features of focal motor seizures are worth noting. First, in some patients, the abnormal motor movements may begin in a very restricted region such as the fingers and gradually progress (over seconds to minutes) to include a larger portion of the extremity. This phenomenon, described by Hughlings Jackson and known as a “Jacksonian march,” represents the spread of seizure activity over a progressively larger region of motor cortex. Second, patients may experience a localized paresis (Todd’s paralysis) for minutes to many hours in the involved region following the seizure. Third, in rare instances, the seizure may continue for hours or days. This condition, termed *epilepsia partialis continua*, is often refractory to medical therapy.

Focal seizures may also manifest as changes in somatic sensation (e.g., paresthesias), vision (flashing lights or formed hallucinations), equilibrium (sensation of falling or vertigo), or autonomic function (flushing, sweating, piloerection). Focal seizures arising from the temporal or frontal cortex may also cause alterations in hearing, olfaction, or emotional state. This includes the sensation of unusual, intense odors (e.g., burning rubber or kerosene) or sounds (crude or highly complex sounds), or an epigastric sensation that rises from the stomach or chest to the head. Some patients describe odd, internal feelings such as fear, a sense of impending change, detachment, depersonalization, déjá vu, or illusions that objects are growing smaller (micropsia) or larger (macropsia). These subjective, “internal” events that are not directly observable by someone else are referred to as *auras*.

Focal Seizures with Impaired Awareness

Focal seizures may also be accompanied by a transient impairment of the patient’s ability to maintain normal contact with the environment. The patient is unable to respond appropriately to visual or verbal commands during the seizure and has impaired recollection or awareness of the ictal phase. The seizures frequently begin with an aura (i.e., a focal seizure without cognitive disturbance) that is stereotypic for the patient. The start of the ictal phase is often a motionless stare, which marks the onset of the period of impaired awareness. The impaired awareness is usually accompanied by *automatisms*, which are involuntary, automatic behaviors that have a wide range of manifestations. Automatisms may consist of very basic behaviors such as chewing, lip smacking, swallowing, or “picking” movements of the hands, or more elaborate behaviors such as a display of emotion or running. The patient is typically confused following the seizure, and the transition to full recovery of consciousness may range from seconds up to an hour or longer. Examination immediately following the seizure may show an anterograde amnesia or transient neurological deficits (such as aphasia, hemi-neglect, or visual loss) caused by postictal inhibition of the cortical regions most involved in the seizure itself.

The range of potential clinical behaviors linked to focal seizures is so broad that extreme caution is advised before concluding that stereotypic episodes of bizarre or atypical behavior are not due to seizure activity. In such cases additional, detailed EEG studies may be helpful.

EVOLUTION OF FOCAL SEIZURES TO GENERALIZED SEIZURES

Focal seizures can spread to involve both cerebral hemispheres and produce a generalized seizure, usually of the tonic-clonic variety (discussed below). This evolution is observed frequently following focal seizures arising from a region in the frontal lobe, but may also be associated with focal seizures occurring elsewhere in the brain. A focal seizure that evolves into a generalized seizure is often difficult to distinguish from a primary generalized onset tonic-clonic seizure, because bystanders tend to emphasize the more dramatic, generalized convulsive phase of the seizure and overlook the more subtle, focal symptoms present at onset. In some cases, the focal onset of the seizure becomes apparent only when a careful history

identifies a preceding aura. Often, however, the focal onset is not clinically evident and may be established only through careful EEG analysis. Nonetheless, distinguishing between these two entities is extremely important, because there may be substantial differences in the evaluation and treatment of epilepsies characterized by focal versus generalized onset seizures.

GENERALIZED ONSET SEIZURES

Generalized seizures arise at some point in the brain but immediately and rapidly engage neuronal networks in both cerebral hemispheres. Several types of generalized seizures have features that place them in distinctive categories and facilitate clinical diagnosis.

Typical Absence Seizures

Typical absence seizures are characterized by sudden, brief lapses of consciousness without loss of postural control. The seizure usually lasts for only seconds, consciousness returns as suddenly as it was lost, and there is no postictal confusion. Although the brief loss of consciousness may be clinically inapparent or the sole manifestation of the seizure discharge, absence seizures are usually accompanied by subtle, bilateral motor signs such as rapid blinking of the eyelids, chewing movements, or small-amplitude, clonic movements of the hands.

Typical absence seizures are associated with a group of genetically determined epilepsies with onset usually in childhood (ages 4–10 years) or early adolescence and are the main seizure type in 15–20% of children with epilepsy. The seizures can occur hundreds of times per day, but the child may be unaware of or unable to convey their existence. Because the clinical signs of the seizures are subtle, especially to parents who may not have had previous experience with seizures, it is not surprising that the first clue to absence epilepsy is often unexplained “daydreaming” and a decline in school performance recognized by a teacher.

The electrophysiologic hallmark of typical absence seizures is a generalized, symmetric, 3-Hz spike-and-slow-wave discharges that begins and ends suddenly, superimposed on a normal EEG background. Periods of spike-and-slow-wave discharges lasting more than a few seconds usually correlate with clinical signs, but the EEG often shows many more brief bursts of abnormal cortical activity than were suspected clinically. Hyperventilation tends to provoke these electrographic discharges and even the seizures themselves and is routinely used when recording the EEG.

Atypical Absence Seizures

Atypical absence seizures have features that deviate both clinically and electrophysiologically from typical absence seizures. For example, the lapse of consciousness is usually of longer duration and less abrupt in onset and cessation, and the seizure is accompanied by more obvious motor signs that may include focal or lateralizing features. The EEG shows a generalized, slow spike-and-slow-wave pattern with a frequency of ≤2.5 per second, as well as other abnormal activity. Atypical absence seizures are usually associated with diffuse or multifocal structural abnormalities of the brain and therefore may accompany other signs of neurologic dysfunction such as mental retardation. Furthermore, the seizures are less responsive to anticonvulsants compared to typical absence seizures.

Generalized, Tonic-Clonic Seizures

Generalized onset tonic-clonic seizures are the main seizure type in ~10% of all persons with epilepsy. They are also the most common seizure type resulting from metabolic derangements and are therefore frequently encountered in many different clinical settings. The seizure usually begins abruptly without warning, although some patients describe vague premonitory symptoms in the hours leading up to the seizure. This prodrome is distinct from the stereotypic auras associated with focal seizures that generalize. The initial phase of the seizure is usually tonic contraction of muscles throughout the body, accounting for a number of the classic features of the event. Tonic contraction of the muscles of expiration and the larynx at the onset will produce a loud moan or “ictal cry.” Respirations are impaired, secretions pool in the oropharynx, and cyanosis develops. Contraction of the jaw muscles may cause biting of the tongue. A marked enhancement of sympathetic tone leads to increases in heart rate, blood pressure, and pupillary size. After 10–20 s, the tonic phase of the seizure typically evolves into the clonic phase, produced by the superimposition of periods of muscle relaxation on the tonic muscle contraction. The periods of relaxation progressively increase until the end of the ictal phase, which usually lasts no more than 1 min. The postictal phase is characterized by unresponsiveness, muscular flaccidity, and excessive salivation that can cause stridorous breathing and partial airway obstruction. Bladder or bowel incontinence may occur at this point. Patients gradually regain consciousness over minutes to hours, and during this transition, there is typically a period of postictal confusion. Patients subsequently complain of headache, fatigue, and muscle ache that can last for many hours. The duration of impaired consciousness in the postictal phase can be extremely long (i.e., many hours) in patients with prolonged seizures or underlying central nervous system (CNS) diseases such as alcoholic cerebral atrophy.

The EEG during the tonic phase of the seizure shows a progressive increase in generalized low-voltage fast activity, followed by generalized high-amplitude, polyspike discharges. In the clonic phase, the high-amplitude activity is typically interrupted by slow waves to create a spike-and-slow-wave pattern. The postictal EEG shows diffuse suppression of all cerebral activity, then slowing that gradually recovers as the patient awakens.

There are a number of variants of generalized motor seizures, including pure tonic and pure clonic seizures. Brief tonic seizures lasting only a few seconds are especially noteworthy since they are usually associated with specific epilepsy syndromes having mixed seizure phenotypes, such as the

Lennox-Gastaut syndrome (discussed below).

Atonic Seizures

Atonic seizures are characterized by sudden loss of postural muscle tone lasting 1–2 s. Consciousness is briefly impaired, but there is usually no postictal confusion. A very brief seizure may cause only a quick head drop or nodding movement, whereas a longer seizure will cause the patient to collapse. This can be extremely dangerous, because there is a substantial risk of direct head injury with the fall. The EEG shows brief, generalized spike-and-wave discharges followed immediately by diffuse slow waves that correlate with the loss of muscle tone. Similar to pure tonic seizures, atonic seizures are usually seen in association with known epilepsy syndromes.

Myoclonic Seizures

Myoclonus is a sudden and brief muscle contraction that may involve one part of the body or the entire body. A normal, common physiologic form of myoclonus is the sudden jerking movement observed while falling asleep. Pathologic myoclonus is most commonly seen in association with metabolic disorders, degenerative CNS diseases, or anoxic brain injury **(Chap. 301)**. Although the distinction from other forms of myoclonus is imprecise, myoclonic seizures are considered to be true epileptic events because they are caused by cortical (versus subcortical or spinal) dysfunction. The EEG shows bilaterally synchronous spike-and-slow-wave discharges immediately prior to the movement and muscle artifact associated with the myoclonus. Myoclonic seizures usually coexist with other forms of generalized seizures but are the predominant feature of juvenile myoclonic epilepsy (JME) (discussed below).

Epileptic Spasms

Epileptic spasms are characterized by a briefly sustained flexion or extension of predominantly proximal muscles, including truncal muscles. The EEG usually shows hypsarrhythmia, which consist of diffuse, giant slow waves with a chaotic background of irregular, multifocal spikes and sharp waves. During the clinical spasm, there is a marked suppression of the EEG background (the “electrodecremental response”). The electromyogram (EMG) also reveals a characteristic rhomboid pattern that may help distinguish spasms from brief tonic and myoclonic seizures. Epileptic spasms occur predominantly in infants and likely result from differences in neuronal function and connectivity in the immature versus mature CNS.

EPILEPSY SYNDROMES

Epilepsy syndromes are disorders in which epilepsy is a predominant feature, and there is sufficient evidence (e.g., through clinical, EEG, radiologic, or genetic observations) to suggest a common underlying mechanism. Three important epilepsy syndromes are listed below; additional examples with a known genetic basis are shown in **Table 418-2**.

TABLE 418-2

Examples of Genes Associated with Epilepsy Syndromes[a]

Gene (Locus)	Function of Gene	Clinical Syndrome	Comments
CHRNA4 (20q13.2)	Nicotinic acetylcholine receptor subunit; mutations cause alterations in Ca^{2+} flux through the receptor; this may reduce the amount of GABA release in presynaptic terminals	Autosomal dominant nocturnal frontal lobe epilepsy (ADNFLE); childhood onset; brief, nighttime seizures with prominent motor movements; often misdiagnosed as primary sleep disorder	Rare; first identified in a large Australian family; other families found to have mutations in *CHRNA2* or *CHRNB2*, and some families appear to have mutations at other loci
KCNQ2 (20q13.3)	Voltage-gated potassium channel subunits; mutation in pore regions may cause a 20–40% reduction of potassium currents, which will lead to impaired repolarization	Benign familial neonatal seizures (BFNSs); autosomal dominant inheritance; onset in first week of life in infants who are otherwise normal; remission usually within weeks to months; long-term epilepsy in 10–15%	Rare; other families found to have mutations in *KCNQ3* or an inversion in chromosomal 5; sequence and functional homology to *KCNQ1*, mutations of which cause long QT syndrome and a cardiac-auditory syndrome
SCN1A (2q24.3)	α-Subunit of a voltage-gated sodium channel; numerous mutations affecting sodium currents that cause either gain or loss of function; network effects appear related to expression in excitatory or inhibitory cells	Generalized epilepsy with febrile seizures plus (GEFS+); autosomal dominant inheritance; presents with febrile seizures at median 1 year, which may persist >6 years, then variable seizure types not associated with fever; numerous other syndromes, including almost 80% of patients with Dravet's syndrome (severe myoclonic epilepsy of infancy) and some cases of Lennox-Gastaut syndrome	Incidence uncertain; GEFS+ identified in other families with mutations in other sodium channel subunits (*SCN2B* and *SCN2A*) and $GABA_A$ receptor subunit (*GABRG2* and *GABRA1*); significant phenotypic heterogeneity within same family, including members with febrile seizures only
LGI1 (10q24)	Leucine-rich glioma-inactivated 1 gene; previous evidence for role in glial tumor progression; recent studies suggest an influence in the postnatal development of glutamatergic circuits in the hippocampus	Autosomal dominant partial epilepsy with auditory features (ADPEAF); a form of idiopathic lateral temporal lobe epilepsy with auditory symptoms or aphasia as a major focal seizure manifestation; age of onset usually between 10 and 25 years	Mutations found in up to 50% of families containing two or more subjects with idiopathic localization-related epilepsy with ictal auditory symptoms, suggesting that at least one other gene may underlie this syndrome
DEPDC5 (22q12.2)	Disheveled, Egl-10 and pleckstrin domain containing protein 5; exerts an inhibitory effect on mammalian target of rapamycin (mTOR)-mediated processes, such as cell growth and proliferation	Autosomal dominant familial focal epilepsy with variable foci (FFEVF); family members have seizures originating from different cortical regions; neuroimaging usually normal but may harbor subtle malformations; recent studies also suggest association with benign epilepsy with centrotemporal spikes	Study of families with the limited number of affected members revealed mutations in ~12% of families; thus may be a relatively common cause of lesion-negative focal epilepsies with suspected genetic basis
CSTB (21q22.3)	Cystatin B, a noncaspase cysteine protease inhibitor; normal protein may block neuronal apoptosis by inhibiting caspases directly or indirectly (via cathepsins), or controlling proteolysis	Progressive myoclonus epilepsy (PME) (Unverricht-Lundborg disease); autosomal recessive inheritance; age of onset between 6 and 15 years, myoclonic seizures, ataxia, and progressive cognitive decline; brain shows neuronal degeneration	Overall rare, but relatively common in Finland and Western Mediterranean (>1 in 20,000); precise role of cystatin B in human disease unknown, although mice with null mutations of cystatin B have similar syndrome

Gene (Locus)	Function of Gene	Clinical Syndrome	Comments
EPM2A (6q24)	Laforin, a protein tyrosine phosphatase (PTP); involved in glycogen metabolism and may have antiapoptotic activity	Progressive myoclonus epilepsy (Lafora's disease); autosomal recessive inheritance; age of onset 6–19 years, death within 10 years; brain degeneration associated with polyglucosan intracellular inclusion bodies in numerous organs	Most common PME in Southern Europe, Middle East, Northern Africa, and Indian subcontinent; genetic heterogeneity; unknown whether seizure phenotype due to degeneration or direct effects of abnormal laforin expression
Doublecortin (Xq21-24)	Doublecortin, expressed primarily in frontal lobes; directly regulates microtubule polymerization and bundling	Classic lissencephaly associated with severe mental retardation and seizures in males; subcortical band heterotopia with more subtle findings in females (presumably due to random X-inactivation); X-linked dominant	Relatively rare but of uncertain incidence; recent increased ascertainment due to improved imaging techniques; relationship between migration defect and seizure phenotype unknown

[a]The first five syndromes listed in the table (ADNFLE, BFNC, GEFS+, ADPEAF, and FFEVF) are examples of idiopathic epilepsies associated with identified gene mutations. The last three syndromes are examples of the numerous Mendelian disorders in which seizures are one part of the phenotype.

Abbreviations: GABA, γ-aminobutyric acid; PME, progressive myoclonus epilepsy.

JUVENILE MYOCLONIC EPILEPSY

JME is a generalized seizure disorder of unknown cause that appears in early adolescence and is usually characterized by bilateral myoclonic jerks that may be single or repetitive. The myoclonic seizures are most frequent in the morning after awakening and can be provoked by sleep deprivation. Consciousness is preserved unless the myoclonus is especially severe. Many patients also experience generalized tonic-clonic seizures, and up to one-third have absence seizures. Although complete remission is relatively uncommon, the seizures usually respond well to appropriate anticonvulsant medication. There is often a family history of epilepsy, and genetic linkage studies suggest a polygenic cause.

LENNOX-GASTAUT SYNDROME

Lennox-Gastaut syndrome occurs in children and is defined by the following triad: (1) multiple seizure types (usually including generalized tonic-clonic, atonic, and atypical absence seizures); (2) an EEG showing slow (<3 Hz) spike-and-wave discharges and a variety of other abnormalities; and (3) impaired cognitive function in most but not all cases. Lennox-Gastaut syndrome is associated with CNS disease or dysfunction from a variety of causes, including *de novo* mutations, developmental abnormalities, perinatal hypoxia/ischemia, trauma, infection, and other acquired lesions. The multifactorial nature of this syndrome suggests that it is a nonspecific response of the brain to diffuse neuronal dysfunction. Unfortunately, many patients have a poor prognosis due to the underlying CNS disease and the physical and psychosocial consequences of severe, poorly controlled epilepsy.

MESIAL TEMPORAL LOBE EPILEPSY SYNDROME

Mesial temporal lobe epilepsy (MTLE) is the most common syndrome associated with focal seizures with impairment of consciousness and is an example of an epilepsy syndrome with distinctive clinical, electroencephalographic, and pathologic features **(Table 418-3)**. High-resolution magnetic resonance imaging (MRI) can detect the characteristic hippocampal sclerosis that appears to be essential in the pathophysiology of MTLE for many patients **(Fig. 418-1)**. Recognition of this syndrome is especially important because it tends to be refractory to treatment with anticonvulsants but responds well to surgical intervention. Advances in the understanding of basic mechanisms of epilepsy have come through studies of experimental models of MTLE, discussed below.

TABLE 418-3

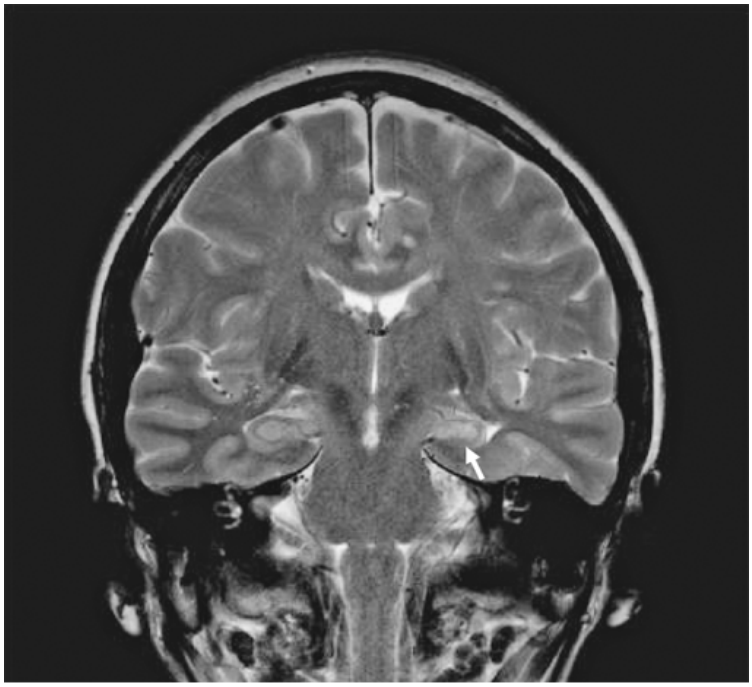
Characteristics of the Mesial Temporal Lobe Epilepsy Syndrome

History	
History of febrile seizures	Rare generalized seizures
Family history of epilepsy	Seizures may remit and reappear
Early onset	Seizures often intractable
Clinical Observations	
Aura common	Postictal disorientation
Behavioral arrest/stare	Memory loss
Complex automatisms	Dysphasia (with focus in dominant hemisphere)
Unilateral posturing	
Laboratory Studies Unilateral or bilateral anterior temporal spikes on EEG Hypometabolism on interictal PET Hypoperfusion on interictal SPECT Material-specific memory deficits on intracranial amobarbital (Wada) test	
MRI Findings Small hippocampus with increased signal on T2-weighted sequences Small temporal lobe Enlarged temporal horn	
Pathologic Findings	
Highly selective loss of specific cell populations within hippocampus in most cases	

Abbreviations: EEG, electroencephalogram; MRI, magnetic resonance imaging; PET, positron emission tomography; SPECT, single-photon emission computed tomography.

FIGURE 418-1

Mesial temporal lobe epilepsy. The electroencephalogram and seizure semiology were consistent with a left temporal lobe focus. This coronal high-resolution T2-weighted fast spin echo magnetic resonance image obtained at 3 tesla is at the level of the hippocampal bodies, and shows abnormal high signal intensity, blurring of internal laminar architecture, and reduced size of the left hippocampus (*arrow*) relative to the right. This triad of imaging findings is consistent with hippocampal sclerosis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

THE CAUSES OF SEIZURES AND EPILEPSY

Seizures are a result of a shift in the normal balance of excitation and inhibition within the CNS. Given the numerous properties that control neuronal excitability, it is not surprising that there are many different ways to perturb this normal balance, and therefore many different causes of both seizures and epilepsy. Three clinical observations emphasize how a variety of factors determine why certain conditions may cause seizures or epilepsy in a given patient.

1. *The normal brain is capable of having a seizure under the appropriate circumstances, and there are differences between individuals in the susceptibility or threshold for seizures*. For example, seizures may be induced by high fevers in children who are otherwise normal and who never develop other neurologic problems, including epilepsy. However, febrile seizures occur only in a relatively small proportion of children. This implies there are various underlying *endogenous factors* that influence the threshold for having a seizure. Some of these factors are genetic, as a family history of epilepsy has a clear influence on the likelihood of seizures occurring in otherwise normal individuals. Normal development also plays an important role, because the brain appears to have different seizure thresholds at different maturational stages.

2. *There are a variety of conditions that have an extremely high likelihood of resulting in a chronic seizure disorder*. One of the best examples of this is severe, penetrating head trauma, which is associated with up to a 45% risk of subsequent epilepsy. The high propensity for severe traumatic brain injury to lead to epilepsy suggests that the injury results in a long-lasting pathologic change in the CNS that transforms a presumably normal neural network into one that is abnormally hyperexcitable. This process is known as *epileptogenesis*, and the specific changes that result in a lowered seizure threshold can be considered *epileptogenic factors*. Other processes associated with epileptogenesis include stroke, infections, and abnormalities of CNS development. Likewise, the genetic abnormalities associated with epilepsy likely involve processes that trigger the appearance of specific sets of epileptogenic factors.

3. *Seizures are episodic*. Patients with epilepsy have seizures intermittently and, depending on the underlying cause, many patients are completely normal for months or even years between seizures. This implies there are important provocative or *precipitating factors* that induce seizures in patients with epilepsy. Similarly, precipitating factors are responsible for causing the single seizure in someone without epilepsy. Precipitants include those due to intrinsic physiologic processes such as psychological or physical stress, sleep deprivation, or hormonal changes. They also include exogenous factors such as exposure to toxic substances and certain medications.

These observations emphasize the concept that the many causes of seizures and epilepsy result from a dynamic interplay between endogenous factors, epileptogenic factors, and precipitating factors. The potential role of each needs to be carefully considered when determining the appropriate management of a patient with seizures. For example, the identification of predisposing factors (e.g., family history of epilepsy) in a patient with febrile seizures may increase the necessity for closer follow-up and a more aggressive diagnostic evaluation. Finding an epileptogenic lesion may help in the estimation of seizure recurrence and duration of therapy. Finally, removal or modification of a precipitating factor may be an effective and safer method for preventing further seizures than the prophylactic use of anticonvulsant drugs.

CAUSES ACCORDING TO AGE

In practice, it is useful to consider the etiologies of seizures based on the age of the patient, because age is one of the most important factors determining both the incidence and the likely causes of seizures or epilepsy **(Table 418-4)**. During the *neonatal period and early infancy*, potential causes include hypoxic-ischemic encephalopathy, trauma, CNS infection, congenital CNS abnormalities, and metabolic disorders. Babies born to mothers using neurotoxic drugs such as cocaine, heroin, or ethanol are susceptible to drug-withdrawal seizures in the first few days after delivery. Hypoglycemia and hypocalcemia, which can occur as secondary complications of perinatal injury, are also causes of seizures early after delivery. Seizures due to inborn errors of metabolism usually present once regular feeding begins, typically 2–3 days after birth. Pyridoxine (vitamin B_6) deficiency, an important cause of neonatal seizures, can be effectively treated with pyridoxine replacement. The idiopathic or inherited forms of benign neonatal seizures are also seen during this time period.

TABLE 418-4

Causes of Seizures

Neonates (<1 month)	Perinatal hypoxia and ischemia Intracranial hemorrhage and trauma CNS infection Metabolic disturbances (hypoglycemia, hypocalcemia, hypomagnesemia, pyridoxine deficiency) Drug withdrawal Developmental disorders Genetic disorders
Infants and children (>1 month and <12 years)	Febrile seizures Genetic disorders (metabolic, degenerative, primary epilepsy syndromes) CNS infection Developmental disorders Trauma
Adolescents (12–18 years)	Trauma Genetic disorders Infection Illicit drug use Brain tumor
Young adults (18–35 years)	Trauma Alcohol withdrawal Illicit drug use Brain tumor Autoantibodies
Older adults (>35 years)	Cerebrovascular disease Brain tumor Alcohol withdrawal Metabolic disorders (uremia, hepatic failure, electrolyte abnormalities, hypoglycemia, hyperglycemia) Alzheimer's disease and other degenerative CNS diseases Autoantibodies

Abbreviation: CNS, central nervous system.

The most common seizures arising in *late infancy and early childhood* are febrile seizures, which are seizures associated with fevers but without evidence of CNS infection or other defined causes. The overall prevalence is 3–5% and even higher in some parts of the world such as Asia. Patients often have a family history of febrile seizures or epilepsy. Febrile seizures usually occur between 3 months and 5 years of age and have a peak incidence between 18 and 24 months. The typical scenario is a child who has a generalized, tonic-clonic seizure during a febrile illness in the setting of a common childhood infection such as otitis media, respiratory infection, or gastroenteritis. The seizure is likely to occur during the rising phase of the temperature curve (i.e., during the first day) rather than well into the course of the illness. A *simple* febrile seizure is a single, isolated event, brief, and

symmetric in appearance. *Complex* febrile seizures are characterized by repeated seizure activity, duration >15 minutes, or by focal features. Approximately one-third of patients with febrile seizures will have a recurrence, but <10% have three or more episodes. Recurrences are much more likely when the febrile seizure occurs in the first year of life. Simple febrile seizures are not associated with an increase in the risk of developing epilepsy, while complex febrile seizures have a risk of 2–5%; other risk factors include the presence of preexisting neurologic deficits and a family history of nonfebrile seizures.

Childhood marks the age at which many of the well-defined epilepsy syndromes present. Some children who are otherwise normal develop idiopathic, generalized tonic-clonic seizures without other features that fit into specific syndromes. Temporal lobe epilepsy usually presents in childhood and may be related to mesial temporal lobe sclerosis (as part of the MTLE syndrome) or other focal abnormalities such as cortical dysgenesis. Other types of focal seizures, including those that evolve into generalized seizures, may be the relatively late manifestation of a developmental disorder, an acquired lesion such as head trauma, CNS infection (especially viral encephalitis), or very rarely a CNS tumor.

The period of *adolescence and early adulthood* is one of transition during which the idiopathic or genetically based epilepsy syndromes, including JME and juvenile absence epilepsy, become less common, while epilepsies secondary to acquired CNS lesions begin to predominate. Seizures that arise in patients in this age range may be associated with head trauma, CNS infections (including parasitic infections such as cysticercosis), brain tumors, congenital CNS abnormalities, illicit drug use, or alcohol withdrawal. Autoantibodies directed against CNS antigens such as potassium channels or glutamate receptors are a newly recognized cause of epilepsy that also begins to appear in this age group (although cases of autoimmunity are being increasingly described in the pediatric population), including patients without an identifiable cancer. This etiology should be suspected when a previously normal individual presents with a particularly aggressive seizure pattern developing over weeks to months and characterized by increasingly frequent and prolonged seizures, especially when combined with psychiatric symptoms and changes in cognitive function **(Chap. 90)**.

Head trauma is a common cause of epilepsy in adolescents and adults. The head injury can be caused by a variety of mechanisms, and the likelihood of developing epilepsy is strongly correlated with the severity of the injury. A patient with a penetrating head wound, depressed skull fracture, intracranial hemorrhage, or prolonged posttraumatic coma or amnesia has a 30–50% risk of developing epilepsy, whereas a patient with a closed head injury and cerebral contusion has a 5–25% risk. Recurrent seizures usually develop within 1 year after head trauma, although intervals of >10 years are well known. In controlled studies, mild head injury, defined as a concussion with amnesia or loss of consciousness of <30 min, was found to be associated with only a slightly increased likelihood of epilepsy. Nonetheless, most epileptologists know of patients who have focal seizures within hours or days of a mild head injury and subsequently develop chronic seizures of the same type; such cases may represent rare examples of chronic epilepsy resulting from mild head injury.

The causes of seizures in *older adults* include cerebrovascular disease, trauma (including subdural hematoma), CNS tumors, and degenerative diseases. Cerebrovascular disease may account for ~50% of new cases of epilepsy in patients >65 years. Acute seizures (i.e., occurring at the time of the stroke) are seen more often with embolic rather than hemorrhagic or thrombotic stroke. Chronic seizures typically appear months to years after the initial event and are associated with all forms of stroke.

Metabolic disturbances such as electrolyte imbalance, hypo- or hyperglycemia, renal failure, and hepatic failure may cause seizures at any age. Similarly, endocrine disorders, hematologic disorders, vasculitides, and many other systemic diseases may cause seizures over a broad age range. A wide variety of medications and abused substances are known to precipitate seizures as well **(Table 418-5)**.

TABLE 418-5

Drugs and Other Substances That Can Cause Seizures

Alkylating agents (e.g., busulfan, chlorambucil)
Antimalarials (chloroquine, mefloquine)
Antimicrobials/antivirals
β-lactam and related compounds
Quinolones
Acyclovir
Isoniazid
Ganciclovir
Anesthetics and analgesics
Meperidine
Fentanyl
Tramadol
Local anesthetics
Dietary supplements
Ephedra (ma huang)
Gingko
Immunomodulatory drugs
Cyclosporine
OKT3 (monoclonal antibodies to T cells)
Tacrolimus
Interferons
Psychotropics
Antidepressants (e.g., bupropion)
Antipsychotics (e.g., clozapine)
Lithium
Radiographic contrast agents
Drug withdrawal
Alcohol

Baclofen
Barbiturates (short-acting)
Benzodiazepines (short-acting)
Zolpidem
Drugs of abuse
Amphetamine
Cocaine
Phencyclidine
Methylphenidate
Flumazenil[a]

[a]In benzodiazepine-dependent patients.

BASIC MECHANISMS

MECHANISMS OF SEIZURE INITIATION AND PROPAGATION

Focal seizure activity can begin in a very discrete region of cortex and then slowly invade the surrounding regions. The hallmark of an established seizure is typically an electrographic "spike" due to intense near-simultaneous firing of a large number of local excitatory neurons, resulting in an apparent hypersynchronization of the excitatory bursts across a relatively large cortical region. The bursting activity in individual neurons (the "paroxysmal depolarization shift") is caused by a relatively long-lasting depolarization of the neuronal membrane due to influx of extracellular calcium (Ca^{2+}), which leads to the opening of voltage-dependent sodium (Na^{+}) channels, influx of Na^{+}, and generation of repetitive action potentials. This is followed by a hyperpolarizing afterpotential mediated by γ-aminobutyric acid (GABA) receptors or potassium (K^{+}) channels, depending on the cell type. The synchronized bursts from a sufficient number of neurons result in a so-called spike discharge on the EEG.

The spreading seizure wavefront is thought to slow and ultimately halt by intact hyperpolarization and a "surround" inhibition created by feedforward activation of inhibitory neurons. With sufficient activation, there is a recruitment of surrounding neurons via a number of synaptic and nonsynaptic mechanisms, including: (1) an increase in extracellular K^{+}, which blunts hyperpolarization and depolarizes neighboring neurons; (2) accumulation of Ca^{2+} in presynaptic terminals, leading to enhanced neurotransmitter release; (3) depolarization-induced activation of the *N*-methyl-D-aspartate (NMDA) subtype of the excitatory amino acid receptor, which causes additional Ca^{2+} influx and neuronal activation; and (4) ephaptic interactions related to changes in tissue osmolarity and cell swelling. The recruitment of a sufficient number of neurons leads to the propagation of excitatory currents into contiguous areas via local cortical connections and to more distant areas via long commissural pathways such as the corpus callosum.

Many factors control neuronal excitability, and thus there are many potential mechanisms for altering a neuron's propensity to have bursting activity. Mechanisms *intrinsic* to the neuron include changes in the conductance of ion channels, response characteristics of membrane receptors, cytoplasmic buffering, second-messenger systems, and protein expression as determined by gene transcription, translation, and posttranslational modification. Mechanisms *extrinsic* to the neuron include changes in the amount or type of neurotransmitters present at the synapse, modulation of receptors by extracellular ions and other molecules, and temporal and spatial properties of synaptic and nonsynaptic input. Nonneural cells, such as astrocytes and oligodendrocytes, have an important role in many of these mechanisms as well.

Certain recognized causes of seizures are explained by these mechanisms. For example, accidental ingestion of domoic acid, which is an analogue of glutamate (the principal excitatory neurotransmitter in the brain), causes profound seizures via direct activation of excitatory amino acid receptors throughout the CNS. Penicillin, which can lower the seizure threshold in humans and is a potent convulsant in experimental models, reduces inhibition by antagonizing the effects of GABA at its receptor. The basic mechanisms of other precipitating factors of seizures, such as sleep deprivation, fever, alcohol withdrawal, hypoxia, and infection, are not as well understood but presumably involve analogous perturbations in neuronal excitability. Similarly, the endogenous factors that determine an individual's seizure threshold may relate to these properties as well.

Knowledge of the mechanisms responsible for initiation and propagation of most generalized seizures (including tonic-clonic, myoclonic, and atonic types) remains rudimentary and reflects the limited understanding of the connectivity of the brain at a systems level. Much more is understood about the origin of generalized spike-and-wave discharges in absence seizures. These appear to be related to oscillatory rhythms normally generated during sleep by circuits connecting the thalamus and cortex. This oscillatory behavior involves an interaction between $GABA_B$ receptors, T-type Ca^{2+} channels, and K^+ channels located within the thalamus. Pharmacologic studies indicate that modulation of these receptors and channels can induce absence seizures, and there is good evidence that the genetic forms of absence epilepsy may be associated with mutations of components of this system.

MECHANISMS OF EPILEPTOGENESIS

Epileptogenesis refers to the transformation of a normal neuronal network into one that is chronically hyperexcitable. There is often a delay of months to years between an initial CNS injury such as trauma, stroke, or infection and the first clinically evident seizure. The injury appears to initiate a process that gradually lowers the seizure threshold in the affected region until a spontaneous seizure occurs. In many genetic and idiopathic forms of epilepsy, epileptogenesis is presumably determined by developmentally regulated events.

Pathologic studies of the hippocampus from patients with temporal lobe epilepsy have led to the suggestion that some forms of epileptogenesis are related to *structural changes in neuronal networks*. For example, many patients with MTLE have a highly selective loss of neurons that normally contribute to inhibition of the main excitatory neurons within the dentate gyrus. There is also evidence that, in response to the loss of neurons, there is reorganization of surviving neurons in a way that affects the excitability of the network. Some of these changes can be seen in experimental models of prolonged electrical seizures or traumatic brain injury. Thus, an initial injury such as head injury may lead to a very focal, confined region of structural change that causes local hyperexcitability. The local hyperexcitability leads to further structural changes that evolve over time until the focal lesion produces clinically evident seizures. Similar models have provided strong evidence for long-term alterations in *intrinsic, biochemical properties of cells* within the network such as chronic changes in glutamate or GABA receptor function. Induction of inflammatory cascades may be a critical factor in these processes as well.

GENETIC CAUSES OF EPILEPSY



The most important recent progress in epilepsy research has been the identification of genetic mutations associated with a variety of epilepsy syndromes (Table 418-2). Although most of the mutations identified to date cause rare forms of epilepsy, their discovery has led to extremely important conceptual advances. For example, it appears that many of the inherited epilepsies are due to mutations affecting ion channel function. These syndromes are therefore part of the larger group of channelopathies causing paroxysmal disorders such as cardiac arrhythmias, episodic ataxia, periodic weakness, and familial hemiplegic migraine. Other gene mutations are proving to be associated with pathways influencing CNS development or neuronal homeostasis. *De novo* mutations may explain a significant proportion of these syndromes, especially those with onset in early childhood. A current challenge is to identify the multiple susceptibility genes that underlie the more common forms of idiopathic epilepsies. Recent studies suggest that ion channel mutations and copy number variants may contribute to causation in a subset of these patients.

MECHANISMS OF ACTION OF ANTIEPILEPTIC DRUGS

Antiepileptic drugs appear to act primarily by blocking the initiation or spread of seizures. This occurs through a variety of mechanisms that modify the activity of ion channels or neurotransmitters, and in most cases, the drugs have pleiotropic effects. The mechanisms include inhibition of Na^+-dependent action potentials in a frequency-dependent manner (e.g., phenytoin, carbamazepine, lamotrigine, topiramate, zonisamide, lacosamide, rufinamide), inhibition of voltage-gated Ca^{2+} channels (phenytoin, gabapentin, pregabalin), facilitating the opening of potassium channels (ezogabine), attenuation of glutamate activity (lamotrigine, topiramate, felbamate), potentiation of GABA receptor function (benzodiazepines and barbiturates), increase in the availability of GABA (valproic acid, gabapentin, tiagabine), and modulation of release of synaptic vesicles (levetiracetam, brivaracetam). Two of the effective drugs for absence seizures, ethosuximide and valproic acid, probably act by inhibiting T-type Ca^{2+} channels in thalamic neurons.

In contrast to the relatively large number of antiepileptic drugs that can attenuate seizure activity, there are currently no drugs known to prevent the formation of a seizure focus following CNS injury. The eventual development of such “antiepileptogenic” drugs will provide an important means of preventing the emergence of epilepsy following injuries such as head trauma, stroke, and CNS infection.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Seizure

When a patient presents shortly after a seizure, the first priorities are attention to vital signs, respiratory and cardiovascular support, and treatment of seizures if they resume (see "Treatment: Seizures and Epilepsy"). Life-threatening conditions such as CNS infection, metabolic derangement, or drug toxicity must be recognized and managed appropriately.

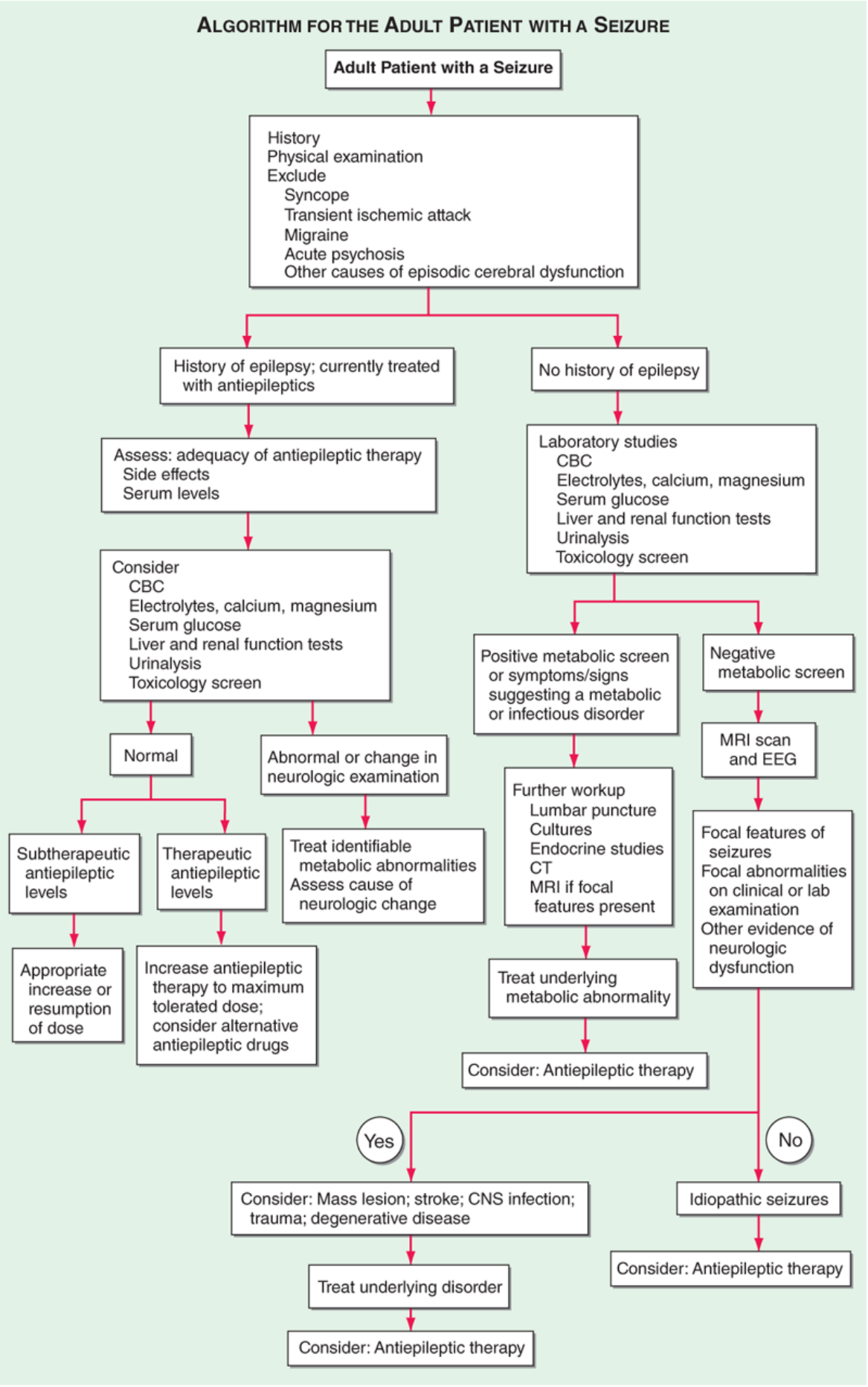
When the patient is not acutely ill, the evaluation will initially focus on whether there is a history of earlier seizures **(Fig. 418-2)**. If this is the first seizure, then the emphasis will be to: (1) establish whether the reported episode was a seizure rather than another paroxysmal event, (2) determine the cause of the seizure by identifying risk factors and precipitating events, and (3) decide whether anticonvulsant therapy is required in addition to treatment for any underlying illness.

In the patient with prior seizures or a known history of epilepsy, the evaluation is directed toward: (1) identification of the underlying cause and precipitating factors, and (2) determination of the adequacy of the patient's current therapy.

FIGURE 418-2

Evaluation of the adult patient with a seizure. CBC, complete blood count; CNS, central nervous system; CT, computed tomography; EEG, electroencephalogram; MRI, magnetic resonance imaging.

ALGORITHM FOR THE ADULT PATIENT WITH A SEIZURE



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HISTORY AND EXAMINATION

The first goal is to determine whether the event was truly a seizure. An in-depth history is essential, because *in many cases the diagnosis of a seizure is based solely on clinical grounds—the examination and laboratory studies are often normal.* Questions should focus on the symptoms before, during, and after the episode in order to differentiate a seizure from other paroxysmal events (see "Differential Diagnosis of Seizures" below). Seizures frequently occur out-of-hospital, and the patient may be unaware of the ictal and immediate postictal phases; thus, witnesses to the event should be interviewed carefully.

The history should also focus on risk factors and predisposing events. Clues for a predisposition to seizures include a history of febrile seizures, a family history of seizures, and, of particular importance, earlier auras or brief seizures not recognized as such. Epileptogenic factors such as prior head trauma, stroke, tumor, or CNS infection should be identified. In children, a careful assessment of developmental milestones may provide evidence for underlying CNS disease. Precipitating factors such as sleep deprivation, systemic diseases, electrolyte or metabolic derangements, acute infection, drugs that lower the seizure threshold (Table 418-5), or alcohol or illicit drug use should also be identified.

The general physical examination includes a search for signs of infection or systemic illness. Careful examination of the skin may reveal signs of neurocutaneous disorders such as tuberous sclerosis or neurofibromatosis, or chronic liver or renal disease. A finding of organomegaly may indicate a metabolic storage disease, and limb asymmetry may provide a clue to brain injury early in development. Signs of head trauma and use of alcohol or illicit drugs should be sought. Auscultation of the heart and carotid arteries may identify an abnormality that predisposes to cerebrovascular disease.

All patients require a complete neurologic examination, with particular emphasis on eliciting signs of cerebral hemispheric disease **(Chap. 415)**. Careful assessment of mental status (including memory, language function, and abstract thinking) may suggest lesions in the anterior frontal, parietal, or temporal lobes. Testing of visual fields will help screen for lesions in the optic pathways and occipital lobes. Screening tests of motor function such as pronator drift, deep tendon reflexes, gait, and coordination may suggest lesions in motor (frontal) cortex, and cortical sensory testing (e.g., double simultaneous stimulation) may detect lesions in the parietal cortex.

LABORATORY STUDIES

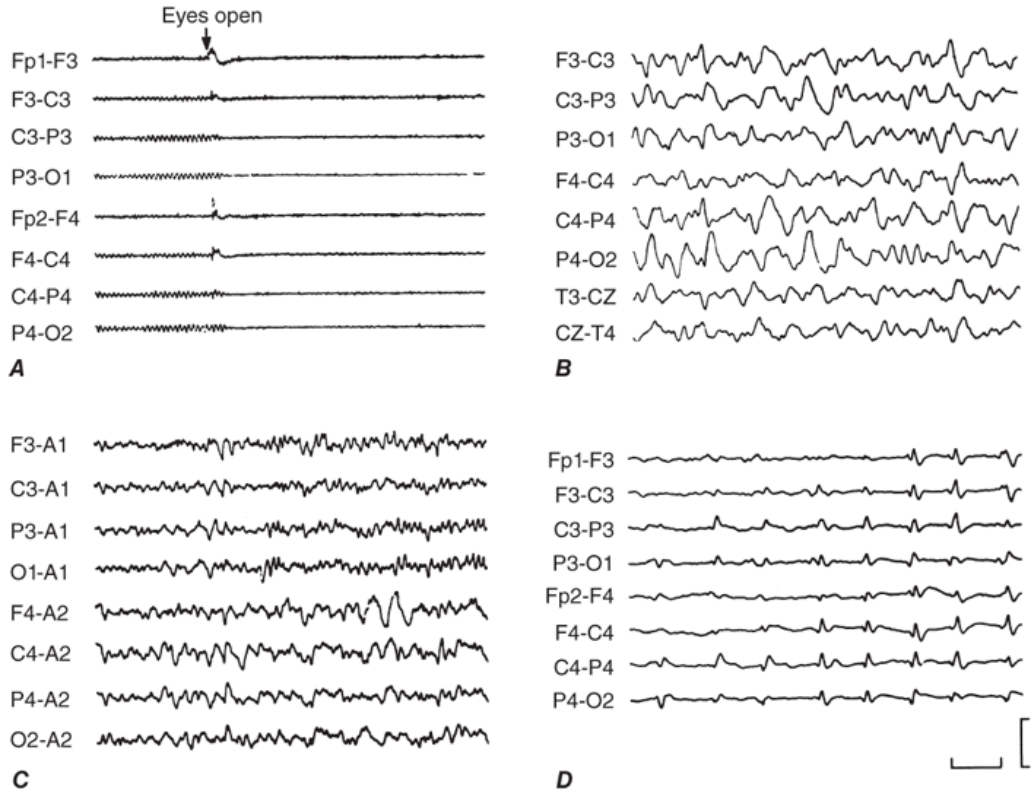
Routine blood studies are indicated to identify the more common metabolic causes of seizures such as abnormalities in electrolytes, glucose, calcium, or magnesium, and hepatic or renal disease. A screen for toxins in blood and urine should also be obtained from all patients in appropriate risk groups, especially when no clear precipitating factor has been identified. A lumbar puncture is indicated if there is any suspicion of meningitis or encephalitis, and it is mandatory in all patients infected with HIV, even in the absence of symptoms or signs suggesting infection. Testing for autoantibodies in the serum and cerebrospinal fluid (CSF) should be considered in patients presenting with a seemingly aggressive form of epilepsy associated with other abnormalities such as psychiatric symptoms or cognitive disturbances.

ELECTROPHYSIOLOGIC STUDIES

The electrical activity of the brain (the EEG) is easily recorded from electrodes placed on the scalp. The potential difference between pairs of electrodes on the scalp (bipolar derivation) or between individual scalp electrodes and a relatively inactive common reference point (referential derivation) is amplified and displayed on a computer monitor, oscilloscope, or paper. Digital systems allow the EEG to be reconstructed and displayed with any desired format and to be manipulated for more detailed analysis and also permit computerized techniques to be used to detect certain abnormalities. The characteristics of the normal EEG depend on the patient's age and level of arousal. The rhythmic activity normally recorded represents the postsynaptic potentials of vertically oriented pyramidal cells of the cerebral cortex and is characterized by its frequency. In normal awake adults lying quietly with the eyes closed, an 8- to 13-Hz alpha rhythm is seen posteriorly in the EEG, intermixed with a variable amount of generalized faster (beta) activity (>13 Hz); the alpha rhythm is attenuated when the eyes are opened **(Fig. 418-3)**. During drowsiness, the alpha rhythm is also attenuated; with light sleep, slower activity in the theta (4–7 Hz) and delta (<4 Hz) ranges becomes more conspicuous.

FIGURE 418-3

A. Normal electroencephalogram (EEG) showing a posteriorly situated 9-Hz alpha rhythm that attenuates with eye opening. ***B.*** Abnormal EEG showing irregular diffuse slow activity in an obtunded patient with encephalitis. ***C.*** Irregular slow activity in the right central region, on a diffusely slowed background, in a patient with a right parietal glioma. *D.* Periodic complexes occurring once every second in a patient with Creutzfeldt-Jakob disease. Horizontal calibration: 1 s; vertical calibration: 200 μV in A, 300 μV in other panels. In this and the following figure, electrode placements are indicated at the left of each panel and accord with the international 10:20 system. A, earlobe; C, central; F, frontal; Fp, frontal polar; P, parietal; T, temporal; O, occipital. Right-sided placements are indicated by even numbers, left-sided placements by odd numbers, and midline placements by Z. (*From MJ Aminoff [ed]: Aminoff's Electrodiagnosis in Clinical Neurology, 6th ed. Oxford, Elsevier Saunders, 2012.*)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

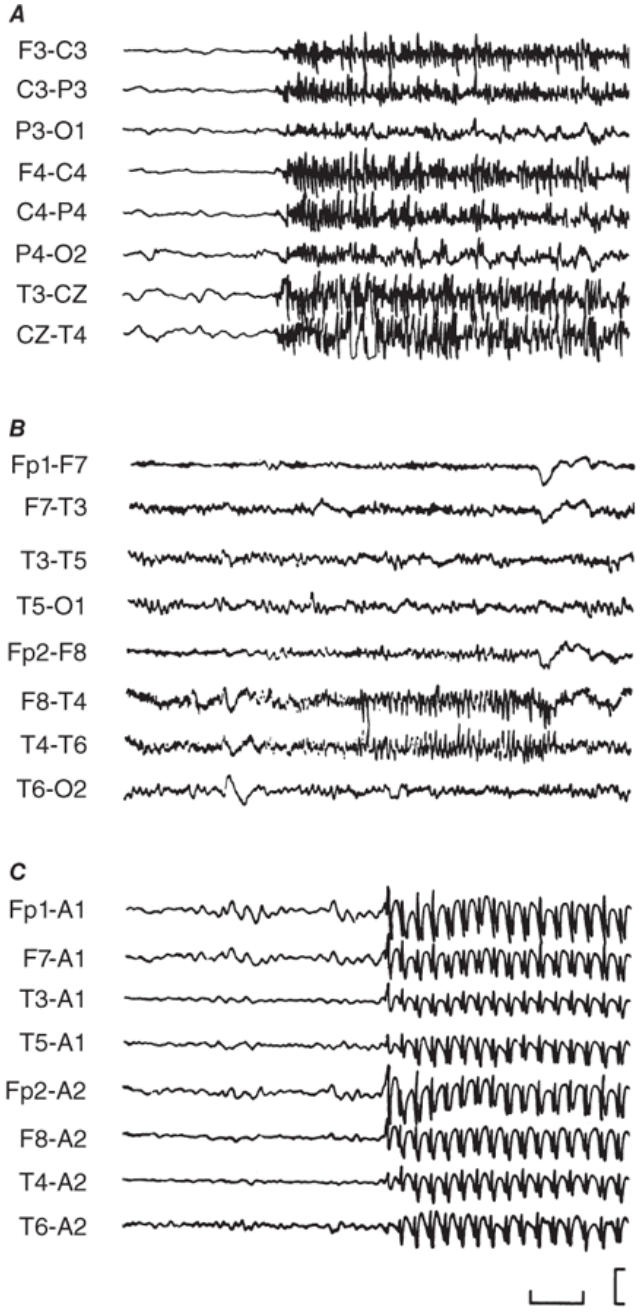
All patients who have a possible seizure disorder should be evaluated with an EEG as soon as possible. In the evaluation of a patient with suspected epilepsy, the presence of *electrographic seizure activity* during the clinically evident event (i.e., abnormal, repetitive, rhythmic activity having a discrete onset and termination) clearly establishes the diagnosis. The absence of electrographic seizure activity does not exclude a seizure disorder, however, because focal seizures may originate from a region of the cortex that cannot be detected by standard scalp electrodes. The EEG is always abnormal during generalized tonic-clonic seizures. Because seizures are typically infrequent and unpredictable, it is often not possible to obtain the EEG during a clinical event. In such situations, activating procedures are generally undertaken while the EEG is recorded in an attempt to provoke abnormalities. These procedures commonly include hyperventilation (for 3 or 4 min), photic stimulation, sleep, and sleep deprivation on the night prior to the recording. Continuous monitoring for prolonged periods in video-EEG telemetry units for hospitalized patients or the use of portable equipment to record the EEG continuously for ≥24 h in ambulatory patients has made it easier to capture the electrophysiologic accompaniments of clinical events. In particular, video-EEG telemetry is now a routine approach for the accurate diagnosis of epilepsy in patients with poorly characterized events or seizures that are difficult to control.

The EEG may also be helpful in the interictal period by showing certain abnormalities that are highly supportive of the diagnosis of epilepsy. Such *epileptiform activity* consists of bursts of abnormal discharges containing spikes or sharp waves. The presence of epileptiform activity is not entirely specific for epilepsy, but it has a much greater prevalence in patients with epilepsy than in normal individuals. However, even in an individual who is known to have epilepsy, the initial routine interictal EEG may be normal up to 60% of the time. Thus, the EEG cannot establish the diagnosis of epilepsy in many cases.

The EEG is also used for classifying seizure disorders and aiding in the selection of anticonvulsant medications (Fig. 418-4). For example, episodic generalized spike-wave activity is usually seen in patients with typical absence epilepsy and may be seen with other generalized epilepsy syndromes. Focal interictal epileptiform discharges would support the diagnosis of a focal seizure disorder such as temporal lobe epilepsy or frontal lobe seizures, depending on the location of the discharges.

FIGURE 418-4

Electrographic seizures. ***A.*** Onset of a tonic seizure showing generalized repetitive sharp activity with synchronous onset over both hemispheres. ***B.*** Burst of repetitive spikes occurring with sudden onset in the right temporal region during a clinical spell characterized by transient impairment of external awareness. ***C.*** Generalized 3-Hz spike-wave activity occurring synchronously over both hemispheres during an absence (petit mal) attack. Horizontal calibration: 1 s; vertical calibration: 400 μV in *A*, 200 μV in *B*, and 750 μV in *C. (From MJ Aminoff [ed]: Aminoff's Electrodiagnosis in Clinical Neurology, 6th ed. Oxford, Elsevier Saunders, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The routine scalp-recorded EEG may also be used to assess the prognosis of seizure disorders; in general, a normal EEG implies a better prognosis, whereas an abnormal background or profuse epileptiform activity suggests a worse outcome. Unfortunately, the EEG has not proved to be useful in predicting which patients with predisposing conditions such as head injury or brain tumor will go on to develop epilepsy, because in such circumstances epileptiform activity is commonly encountered regardless of whether seizures occur.

Magnetoencephalography (MEG) provides another way of looking noninvasively at cortical activity. Instead of measuring electrical activity of the brain, it measures the small magnetic fields that are generated by this activity. The source of epileptiform activity seen on MEG can be analyzed, and its source in the brain can be estimated using a variety of mathematical techniques. These source estimates can then be plotted on an anatomic image of the brain such as an MRI (discussed below) to generate a magnetic source image (MSI). MSI can be useful to localize potential seizure foci.

BRAIN IMAGING

Almost all patients with new-onset seizures should have a brain imaging study to determine whether there is an underlying structural abnormality that is responsible. The only potential exception to this rule is children who have an unambiguous history and examination suggestive of a benign, generalized seizure disorder such as absence epilepsy. MRI has been shown to be superior to computed tomography (CT) for the detection of cerebral lesions associated with epilepsy. In some cases, MRI will identify lesions such as tumors, vascular malformations, or other pathologies that need

urgent therapy. The availability of newer MRI methods such as 3-tesla scanners, parallel imaging with multichannel head coils, three-dimensional structural imaging at submillimeter resolution, and widespread use of pulse sequences such as fluid-attenuated inversion recovery (FLAIR), has increased the sensitivity for detection of abnormalities of cortical architecture, including hippocampal atrophy associated with mesial temporal sclerosis, as well as abnormalities of cortical neuronal migration. In such cases, the findings may not lead to immediate therapy, but they do provide an explanation for the patient's seizures and point to the need for chronic antiepileptic drug therapy or possible surgical resection.

In the patient with a suspected CNS infection or mass lesion, CT scanning should be performed emergently when MRI is not immediately available. Otherwise, it is usually appropriate to obtain an MRI study within a few days of the initial evaluation. Functional imaging procedures such as positron emission tomography (PET) and single-photon emission computed tomography (SPECT) are also used to evaluate certain patients with medically refractory seizures (discussed below).

GENETIC TESTING

With the increasing recognition of specific gene mutations causing epilepsy, genetic testing is beginning to emerge as part of the diagnostic evaluation of patients with epilepsy. In addition to providing a definitive diagnosis (which may be of great benefit to the patient and family members, and curtail the pursuit of additional, unrevealing laboratory testing), genetic testing may offer a guide for therapeutic options (see section "Selection of Antiepileptic Drugs" below). Presently, genetic testing is being done mainly in infants and children with epilepsy syndromes thought to have a genetic cause. However, genetic testing should also be considered in older patients with a history suggesting an undiagnosed genetic epilepsy syndrome that began early in life.

DIFFERENTIAL DIAGNOSIS OF SEIZURES

Disorders that may mimic seizures are listed in **Table 418-6**. In most cases, seizures can be distinguished from other conditions by meticulous attention to the history and relevant laboratory studies. On occasion, additional studies such as video-EEG monitoring, sleep studies, tilt-table analysis, or cardiac electrophysiology may be required to reach a correct diagnosis. Two of the more common nonepileptic syndromes in the differential diagnosis are detailed below.

TABLE 418-6

Differential Diagnosis of Seizures

Syncope	**Transient ischemic attack (TIA)**
Vasovagal syncope	Basilar artery TIA
Cardiac arrhythmia	**Sleep disorders**
Valvular heart disease	Narcolepsy/cataplexy
Cardiac failure	Benign sleep myoclonus
Orthostatic hypotension	**Movement disorders**
Psychological disorders	Tics
Psychogenic seizure	Nonepileptic myoclonus
Hyperventilation	Paroxysmal choreoathetosis
Panic attack	**Special considerations in children**
Metabolic disturbances	Breath-holding spells
Alcoholic blackouts	Migraine with recurrent abdominal pain and cyclic vomiting
Delirium tremens	Benign paroxysmal vertigo
Hypoglycemia	Apnea
Hypoxia	Night terrors
Psychoactive drugs (e.g., hallucinogens)	Sleepwalking
Migraine	
Confusional migraine	
Basilar migraine	

SYNCOPE

(See also Chap. 18) The diagnostic dilemma encountered most frequently is the distinction between a generalized seizure and syncope. Observations by the patient and bystanders that can help differentiate between the two are listed in **Table 418-7**. Characteristics of a seizure include the presence of an aura, cyanosis, unconsciousness, motor manifestations lasting >15 s, postictal disorientation, muscle soreness, and sleepiness. In contrast, a

syncopal episode is more likely if the event was provoked by acute pain or emotional stress or occurred immediately after arising from the lying or sitting position. Patients with syncope often describe a stereotyped transition from consciousness to unconsciousness that includes tiredness, sweating, nausea, and tunneling of vision, and they experience a relatively brief loss of consciousness. Headache or incontinence usually suggests a seizure but may on occasion also occur with syncope. A brief period (i.e., 1–10 s) of convulsive motor activity is frequently seen immediately at the onset of a syncopal episode, especially if the patient remains in an upright posture after fainting (e.g., in a dentist's chair) and therefore has a sustained decrease in cerebral perfusion. Rarely, a syncopal episode can induce a full tonic-clonic seizure. In such cases, the evaluation must focus on both the cause of the syncopal event as well as the possibility that the patient has a propensity for recurrent seizures.

TABLE 418-7

Features That Distinguish Generalized Tonic-Clonic Seizure from Syncope

Features	Seizure	Syncope
Immediate precipitating factors	Usually none	Emotional stress, Valsalva, orthostatic hypotension, cardiac etiologies
Premonitory symptoms	None or aura (e.g., odd odor)	Tiredness, nausea, diaphoresis, tunneling of vision
Posture at onset	Variable	Usually erect
Transition to unconsciousness	Often immediate	Gradual over seconds[a]
Duration of unconsciousness	Minutes	Seconds
Duration of tonic or clonic movements	30–60 s	Never >15 s
Facial appearance during event	Cyanosis, frothing at mouth	Pallor
Disorientation and sleepiness after event	Many minutes to hours	<5 min
Aching of muscles after event	Often	Sometimes
Biting of tongue	Sometimes	Rarely
Incontinence	Sometimes	Sometimes
Headache	Sometimes	Rarely

[a]May be sudden with certain cardiac arrhythmias.

PSYCHOGENIC SEIZURES

Psychogenic seizures are nonepileptic behaviors that resemble seizures. They are often part of a conversion reaction precipitated by underlying psychological distress. Certain behaviors such as side-to-side turning of the head, asymmetric and large-amplitude shaking movements of the limbs, twitching of all four extremities without loss of consciousness, and pelvic thrusting are more commonly associated with psychogenic rather than epileptic seizures. Psychogenic seizures often last longer than epileptic seizures and may wax and wane over minutes to hours. However, the distinction is sometimes difficult on clinical grounds alone, and there are many examples of diagnostic errors made by experienced epileptologists. This is especially true for psychogenic seizures that resemble focal seizures, because the behavioral manifestations of focal seizures (especially of frontal lobe origin) can be extremely unusual, and in both cases, the routine surface EEG may be normal. Video-EEG monitoring is very useful when historic features are nondiagnostic. Generalized tonic-clonic seizures always produce marked EEG abnormalities during and after the seizure. For suspected focal seizures of temporal lobe origin, the use of additional electrodes may help to localize a seizure focus. Measurement of serum prolactin levels may also help to distinguish between epileptic and psychogenic seizures, because most generalized seizures and some focal seizures are accompanied by rises in serum prolactin (during the immediate 30-min postictal period), whereas psychogenic seizures are not. The diagnosis of psychogenic seizures does not exclude a concurrent diagnosis of epilepsy, because the two may coexist.

TREATMENT

TREATMENT

Seizures and Epilepsy

Therapy for a patient with a seizure disorder is almost always multimodal and includes treatment of underlying conditions that cause or contribute to the seizures, avoidance of precipitating factors, suppression of recurrent seizures by prophylactic therapy with antiepileptic medications or surgery, and addressing a variety of psychological and social issues. Treatment plans must be individualized, given the many different types and causes of seizures as well as the differences in efficacy and toxicity of antiepileptic medications for each patient. In almost all cases, a neurologist with experience in the treatment of epilepsy should design and oversee implementation of the treatment strategy. Furthermore, patients with refractory epilepsy or those who require polypharmacy with antiepileptic drugs should remain under the regular care of a neurologist.

TREATMENT OF UNDERLYING CONDITIONS

If the sole cause of a seizure is a metabolic disturbance such as an abnormality of serum electrolytes or glucose, then treatment is aimed at reversing the metabolic problem and preventing its recurrence. Therapy with antiepileptic drugs is usually unnecessary unless the metabolic disorder cannot be corrected promptly and the patient is at risk of having further seizures. If the apparent cause of a seizure was a medication (e.g., theophylline) or illicit drug use (e.g., cocaine), then appropriate therapy is avoidance of the drug; there is usually no need for antiepileptic medications unless subsequent seizures occur in the absence of these precipitants.

Seizures caused by a structural CNS lesion such as a brain tumor, vascular malformation, or brain abscess may not recur after appropriate treatment of the underlying lesion. However, despite removal of the structural lesion, there is a risk that the seizure focus will remain in the surrounding tissue or develop *de novo* as a result of gliosis and other processes induced by surgery, radiation, or other therapies. Most patients are therefore maintained on an antiepileptic medication for at least 1 year, and an attempt is made to withdraw medications only if the patient has been completely seizure free. If seizures are refractory to medication, the patient may benefit from surgical removal of the epileptic brain region (see below).

AVOIDANCE OF PRECIPITATING FACTORS

Unfortunately, little is known about the specific factors that determine precisely when a seizure will occur in a patient with epilepsy. An almost universal precipitating factor for seizures is sleep deprivation, so patients should do everything possible to optimize their sleep quality. Many patients can identify other particular situations that appear to lower their seizure threshold; these situations should be avoided. For example, patients may note an association between alcohol intake and seizures, and they should be encouraged to modify their drinking habits accordingly. There are also relatively rare cases of patients with seizures that are induced by highly specific stimuli such as a video game monitor, music, or an individual's voice ("reflex epilepsy"). Because there is often an association between stress and seizures, stress reduction techniques such as physical exercise, meditation, or counseling may be helpful.

ANTIEPILEPTIC DRUG THERAPY

Antiepileptic drug therapy is the mainstay of treatment for most patients with epilepsy. The overall goal is to completely prevent seizures without causing any untoward side effects, preferably with a single medication and a dosing schedule that is easy for the patient to follow. Seizure classification is an important element in designing the treatment plan, because some antiepileptic drugs have different activities against various seizure types. However, there is considerable overlap between many antiepileptic drugs such that the choice of therapy is often determined more by the patient's specific needs, especially his or her assessment of side effects.

When to Initiate Antiepileptic Drug Therapy

Antiepileptic drug therapy should be started in any patient with recurrent seizures of unknown etiology or a known cause that cannot be reversed. Whether to initiate therapy in a patient with a single seizure is controversial. Patients with a single seizure due to an identified lesion such as a CNS tumor, infection, or trauma, in which there is strong evidence that the lesion is epileptogenic, should be treated. The risk of seizure recurrence in a patient with an apparently unprovoked or idiopathic seizure is uncertain, with estimates ranging from 31 to 71% in the first 12 months after the initial seizure. This uncertainty arises from differences in the underlying seizure types and etiologies in various published epidemiologic studies. Generally accepted risk factors associated with recurrent seizures include the following: (1) an abnormal neurologic examination, (2) seizures presenting as status epilepticus, (3) postictal Todd's paralysis, (4) a strong family history of seizures, or (5) an abnormal EEG. Most patients with one or more of these risk factors should be treated. Issues such as employment or driving may influence the decision whether to start medications as well. For example, a patient with a single, idiopathic seizure whose job depends on driving may prefer taking antiepileptic drugs rather than risk a seizure recurrence and the potential loss of driving privileges.

Selection of Antiepileptic Drugs

Antiepileptic drugs available in the United States are shown in **Table 418-8**, and the main pharmacologic characteristics of commonly used drugs are listed in **Table 418-9**. Worldwide, older medications such as phenytoin, valproic acid, carbamazepine, phenobarbital, and ethosuximide are generally used as first-line therapy for most seizure disorders because, overall, they are as effective as recently marketed drugs and significantly less expensive overall. Most of the new drugs that have become available in the past decade are used as add-on or alternative therapy, although many are now being used as first-line monotherapy.

In addition to efficacy, factors influencing the choice of an initial medication include the convenience of dosing (e.g., once daily versus three or four times daily) and potential side effects. In this regard, a number of the newer drugs have the advantage of reduced drug–drug interactions and easier dosing. Almost all of the commonly used antiepileptic drugs can cause similar, dose-related side effects such as sedation, ataxia, and diplopia. Long-term use of some agents in adults, especially the elderly, can lead to osteoporosis. Close follow-up is required to ensure these side effects are promptly recognized and reversed. Most of the older drugs and some of the newer ones can also cause idiosyncratic toxicity such as rash, bone marrow suppression, or hepatotoxicity. Although rare, these side effects should be considered during drug selection, and patients must be instructed about symptoms or signs that should signal the need to alert their health care provider. For some drugs, laboratory tests (e.g., complete blood count and liver function tests) are recommended prior to the institution of therapy (to establish baseline values) and during initial dosing and titration of the agent.

An important recent advance in the care of patients with epilepsy has been the application of genetic testing to help guide the choice of therapy (as well as establishing the underlying cause of a patient's syndrome). For example, the identification of a mutation in the *SLC2A1* gene, which encodes the glucose type 1 transporter (GLUT-1) and is a cause of GLUT-1 deficiency, should immediately prompt treatment with the ketogenic diet. Mutations of the *ALDH7A1* gene, which encodes antiquitin, can cause alterations in pyridoxine metabolism that are reversed by treatment with pyridoxine. There is also mounting evidence that certain gene mutations may indicate better or worse response to specific antiepileptic drugs. For example, patients with mutations in the sodium channel subunit *SCN1A* should generally avoid taking phenytoin or lamotrigine, whereas patients with mutations in the *SCN2A* or *SCN8A* sodium channel subunits appear to respond favorably to high-dose phenytoin. Genetic testing may also help predict antiepileptic drug toxicity. Studies have shown that Asian individuals carrying the human leukocyte antigen allele, HLA-B*1502, are at particularly high risk of developing serious skin reactions from carbamazepine, phenytoin, oxcarbazepine, and lamotrigine. HLA-A*31:01 has also been found to be associated with carbamazepine-induced hypersensitivity reactions in patients of European or Japanese ancestry. As a result, racial background and genotype are additional factors to consider in drug selection.

Antiepileptic Drug Selection for Focal Seizures

Carbamazepine (or a related drug, oxcarbazepine), lamotrigine, phenytoin, and levetiracetam are currently the drugs of choice approved for the initial treatment of focal seizures, including those that evolve into generalized seizures. Overall they have very similar efficacy, but differences in pharmacokinetics and toxicity are the main determinants for use in a given patient. For example, an advantage of carbamazepine (which is also available in an extended-release form) is that its metabolism follows first-order pharmacokinetics, which allows for a linear relationship between drug dose, serum levels, and toxicity. Carbamazepine can cause leukopenia, aplastic anemia, or hepatotoxicity and would therefore be contraindicated in patients with predispositions to these problems. Oxcarbazepine has the advantage of being metabolized in a way that avoids an intermediate metabolite associated with some of the side effects of carbamazepine. Oxcarbazepine also has fewer drug interactions than carbamazepine. Lamotrigine tends to be well tolerated in terms of side effects. However, patients need to be particularly vigilant about the possibility of a skin rash during the initiation of therapy. This can be extremely severe and lead to Stevens-Johnson syndrome if unrecognized and if the medication is not discontinued immediately. This risk can be reduced by the use of low initial doses and slow titration. Lamotrigine must be started at lower initial doses when used as add-on therapy with valproic acid, because valproic acid inhibits lamotrigine metabolism and results in a substantially prolonged half-life. Phenytoin has a relatively long half-life and offers the advantage of once or twice daily dosing compared to two or three times daily dosing for many of the other drugs. However, phenytoin shows properties of nonlinear kinetics, such that small increases in phenytoin doses above a standard maintenance dose can precipitate marked side effects. This is one of the main causes of acute phenytoin toxicity. Long-term use of phenytoin is associated with untoward cosmetic effects (e.g., hirsutism, coarsening of facial features, gingival hypertrophy) and effects on bone metabolism. Due to these side effects, phenytoin is often avoided in young patients who are likely to require the drug for many years. Levetiracetam has the advantage of having no known drug–drug interactions, making it especially useful in the elderly and patients on other medications. However, a significant number of patients taking levetiracetam complain of irritability, anxiety, and other psychiatric symptoms. Topiramate can be used for both focal and generalized seizures. Similar to some of the other antiepileptic drugs, topiramate can cause significant psychomotor slowing and other cognitive problems. Additionally, it should not be used in patients at risk for the development of glaucoma or renal stones.

Valproic acid is an effective alternative for some patients with focal seizures, especially when the seizures generalize. Gastrointestinal side effects are fewer when using the delayed-release formulation (Depakote). Laboratory testing is required to monitor toxicity because valproic acid can rarely cause reversible bone marrow suppression and hepatotoxicity. This drug should generally be avoided in patients with preexisting bone marrow or liver disease. Valproic acid also has relatively high risks of unacceptable adverse effects for women of childbearing age, including hyperandrogenism that may affect fertility and teratogenesis (e.g., neural tube defects) in offspring. Irreversible, fatal hepatic failure appearing as an idiosyncratic rather than dose-related side effect is a relatively rare complication; its risk is highest in children <2 years old, especially those taking other antiepileptic drugs or with inborn errors of metabolism.

Zonisamide, brivaracetam, tiagabine, gabapentin, and lacosamide are additional drugs currently used for the treatment of focal seizures with or without evolution into generalized seizures. Phenobarbital and other barbiturate compounds were commonly used in the past as first-line therapy for many forms of epilepsy. However, the barbiturates frequently cause sedation in adults, hyperactivity in children, and other more subtle cognitive changes; thus, their use should be limited to situations in which no other suitable treatment alternatives exist.

Antiepileptic Drug Selection for Generalized Seizures

Lamotrigine, valproic acid and levetiracetam are currently considered the best initial choice for the treatment of primary generalized, tonic-clonic seizures. Topiramate, zonisamide, phenytoin, carbamazepine, and oxcarbazepine are suitable alternatives, although carbamazepine, oxcarbazepine, and phenytoin can worsen certain types of generalized seizures. Valproic acid is particularly effective in absence, myoclonic, and atonic seizures. It is therefore commonly used in patients with generalized epilepsy syndromes having mixed seizure types. However, levetiracetam, rather than valproic acid, is increasingly considered the initial drug of choice for women with epilepsies having mixed seizure types given the adverse effects of valproic acid for women of childbearing age. Lamotrigine is also an alternative to valproate, especially for absence epilepsies. Ethosuximide is a particularly effective drug for the treatment of uncomplicated absence seizures, but it is not useful for tonic-clonic or focal seizures. Periodic monitoring of blood cell counts is required since ethosuximide rarely causes bone marrow suppression.

INITIATION AND MONITORING OF THERAPY

Because the response to any antiepileptic drug is unpredictable, patients should be carefully educated about the approach to therapy. The goal is to prevent seizures and minimize the side effects of treatment; determination of the optimal dose is often a matter of trial and error. This process may take months or longer if the baseline seizure frequency is low. Most antiepileptic drugs need to be introduced relatively slow to minimize side effects. Patients should expect that minor side effects such as mild sedation, slight changes in cognition, or imbalance will typically resolve within a few days. Starting doses are usually the lowest value listed under the dosage column in **Table 418-9**. Subsequent increases should be made only after achieving a steady state with the previous dose (i.e., after an interval of five or more half-lives).

Monitoring of serum antiepileptic drug levels can be very useful for establishing the initial dosing schedule. However, the published therapeutic ranges of serum drug concentrations are only an approximate guide for determining the proper dose for a given patient. The key determinants are the clinical measures of seizure frequency and presence of side effects, not the laboratory values. Conventional assays of serum drug levels measure the total drug (i.e., both free and protein bound). However, it is the concentration of free drug that reflects extracellular levels in the brain and correlates best with efficacy. Thus, patients with decreased levels of serum proteins (e.g., decreased serum albumin due to impaired liver or renal function) may have an increased ratio of free to bound drug, yet the concentration of free drug may be adequate for seizure control. These patients may have a "subtherapeutic" drug level, but the dose should be changed only if seizures remain uncontrolled, not just to achieve a "therapeutic" level. It is also useful to monitor free drug levels in such patients. In practice, other than during the initiation or modification of therapy, monitoring of antiepileptic drug levels is most useful for documenting adherence or assessing clinical suspicion of toxicity.

If seizures continue despite gradual increases to the maximum tolerated dose and documented compliance, then it becomes necessary to switch to another antiepileptic drug. This is usually done by maintaining the patient on the first drug while a second drug is added. The dose of the second drug should be adjusted to decrease seizure frequency without causing toxicity. Once this is achieved, the first drug can be gradually withdrawn (usually over weeks unless there is significant toxicity). The dose of the second drug is then further optimized based on seizure response and side effects. Monotherapy should be the goal whenever possible.

WHEN TO DISCONTINUE THERAPY

Overall, about 50–60% of patients who have their seizures completely controlled with antiepileptic drugs can eventually discontinue therapy. The following patient profile yields the greatest chance of remaining seizure free after drug withdrawal: (1) complete medical control of seizures for 1–5 years; (2) single seizure type, with generalized seizures having a better prognosis than focal seizures; (3) normal neurologic examination, including intelligence; (4) no family history of epilepsy; and (5) normal EEG. The appropriate seizure-free interval is unknown and undoubtedly varies for different forms of epilepsy. However, it seems reasonable to attempt withdrawal of therapy after 2 years in a patient who meets all of the above criteria, is motivated to discontinue the medication, and clearly understands the potential risks and benefits. In most cases, it is preferable to reduce the dose of the drug gradually over 2–3 months. Most recurrences occur in the first 3 months after discontinuing therapy, and patients should be advised to avoid potentially dangerous situations such as driving or swimming during this period.

TREATMENT OF REFRACTORY EPILEPSY

Approximately one-third of patients with epilepsy do not respond to treatment with a single antiepileptic drug, and it becomes necessary to try a combination of drugs to control seizures. Patients who have focal epilepsy related to an underlying structural lesion or those with multiple seizure types and developmental delay are particularly likely to require multiple drugs. There are currently no clear guidelines for rational polypharmacy, although in theory a combination of drugs with different mechanisms of action may be most useful. In most cases, the initial combination therapy combines first-line drugs (i.e., carbamazepine, oxcarbazepine, lamotrigine, valproic acid, levetiracetam, and phenytoin). If these drugs are unsuccessful, then the addition of other drugs such as zonisamide, brivaracetam, topiramate, lacosamide, or tiagabine is indicated. Patients with myoclonic seizures resistant to valproic acid may benefit from the addition of clonazepam or clobazam, and those with absence seizures may respond to a combination of valproic acid and ethosuximide. The same principles concerning the monitoring of therapeutic response, toxicity, and serum levels for monotherapy apply to polypharmacy, and potential drug interactions need to be recognized. If there is no improvement, a third drug can be added while the first two are maintained. If there is a response, the less effective or less well tolerated of the first two drugs should be gradually withdrawn.

SURGICAL TREATMENT OF REFRACTORY EPILEPSY

Approximately 20–30% of patients with epilepsy continue to have seizures despite efforts to find an effective combination of antiepileptic drugs. For some, surgery can be extremely effective in substantially reducing seizure frequency and even providing complete seizure control. Understanding the potential value of surgery is especially important when a patient's seizures are not controlled with initial treatment, as such patients often do not respond to subsequent medication trials. Rather than submitting the patient to years of unsuccessful medical therapy and the psychosocial trauma and increased mortality associated with ongoing seizures, the patient should have an efficient but relatively brief attempt at medical therapy and then be referred for surgical evaluation.

The most common surgical procedure for patients with temporal lobe epilepsy involves resection of the anteromedial temporal lobe (temporal lobectomy) or a more limited removal of the underlying hippocampus and amygdala (amygdalohippocampectomy). Focal seizures arising from extratemporal regions may be abolished by a focal neocortical resection with precise removal of an identified lesion (lesionectomy). Localized neocortical resection without a clear lesion identified on MRI is also possible when other tests (e.g., MEG, PET, SPECT) implicate a focal cortical region as a seizure onset zone. When the cortical region cannot be removed, multiple subpial transection, which disrupts intracortical connections, is sometimes used to prevent seizure spread. Hemispherectomy or multilobar resection is useful for some patients with severe seizures due to hemispheric abnormalities such as hemimegalencephaly or other dysplastic abnormalities, and corpus callosotomy has been shown to be effective for disabling tonic or atonic seizures, usually when they are part of a mixed-seizure syndrome (e.g., Lennox-Gastaut syndrome).

Presurgical evaluation is designed to identify the functional and structural basis of the patient's seizure disorder. Inpatient video-EEG monitoring is used to define the anatomic location of the seizure focus and to correlate the abnormal electrophysiologic activity with behavioral manifestations of the seizure. Routine scalp or scalp-sphenoidal recordings and a high-resolution MRI scan are usually sufficient for localization of the epileptogenic focus, especially when the findings are concordant. Functional imaging studies such as SPECT, PET, and MEG are adjunctive tests that may help to reveal or verify the localization of an apparent epileptogenic region. Once the presumed location of the seizure onset is identified, additional studies, including neuropsychological testing, the intracarotid amobarbital test (Wada test), and functional MRI may be used to assess language and memory localization and to determine the possible functional consequences of surgical removal of the epileptogenic region. In some cases, standard noninvasive evaluation is not sufficient to localize the seizure onset zone, and invasive electrophysiologic monitoring, such as implanted depth or subdural electrodes, is required for more definitive localization. The exact extent of the resection to be undertaken can also be determined by performing cortical mapping at the time of the surgical procedure, allowing for a tailored resection. This involves electrocorticographic recordings made with electrodes on the surface of the brain to identify the extent of epileptiform disturbances. If the region to be resected is within or near brain regions suspected of having sensorimotor or language function, electrical cortical stimulation mapping is performed on the awake patient to determine the function of cortical regions in question in order to avoid resection of so-called eloquent cortex and thereby minimize postsurgical deficits.

Advances in presurgical evaluation and microsurgical techniques have led to a steady increase in the success of epilepsy surgery. Clinically significant complications of surgery are <5%, and the use of functional mapping procedures has markedly reduced the neurologic sequelae due to removal or sectioning of brain tissue. For example, about 70% of patients treated with temporal lobectomy will become seizure free, and another 15–25% will have at least a 90% reduction in seizure frequency. Marked improvement is also usually seen in patients treated with hemispherectomy for catastrophic seizure disorders due to large hemispheric abnormalities. Postoperatively, patients generally need to remain on antiepileptic drug therapy, but the marked reduction of seizures following resective surgery can have a very beneficial effect on quality of life.

Not all medically refractory patients are suitable candidates for resective surgery. For example, some patients have seizures arising from more than one location, making the risk of ongoing seizures or potential harm from the surgery unacceptably high. Vagus nerve stimulation (VNS) has been used in some of these cases, although the results are limited and it is difficult to predict who will benefit. An implantable device that can detect the onset of a seizure (in some instances before the seizure becomes clinically apparent) and deliver an electrical stimulation to abort the seizure (Responsive NeuroStimulation) has proved to be of benefit in selected patients. Studies are currently evaluating the efficacy of stereotactic radiosurgery, laser thermoablation, and deep brain stimulation (DBS) as other options for surgical treatment of refractory epilepsy.

TABLE 418-8

Selection of Antiepileptic Drugs

Generalized-Onset Tonic-Clonic	Focal	Typical Absence	Atypical Absence, Myoclonic, Atonic
First-Line			
Lamotrigine Valproic acid	Lamotrigine Carbamazepine Oxcarbazepine Phenytoin Levetiracetam	Valproic acid Ethosuximide Lamotrigine	Valproic acid Lamotrigine Topiramate
Alternatives			
Zonisamide[a] Phenytoin Levetiracetam Carbamazepine Oxcarbazepine Topiramate Phenobarbital Primidone Felbamate	Zonisamide[a] Brivaracetam Topiramate Valproic acid Tiagabine[a] Gabapentin[a] Lacosamide[a] Phenobarbital Primidone Felbamate	Lamotrigine Clonazepam	Clonazepam Felbamate Clobazam Rufinamide

[a]As adjunctive therapy.

TABLE 418-9

Dosage and Adverse Effects of Commonly Used Antiepileptic Drugs

Generic Name	Trade Name	Principal Uses	Typical Dose; Dose Interval	Half-Life	Therapeutic Range	Adverse Effects		Drug Interactions[a]
						Neurologic	Systemic	
Brivaracetam	Briviact	Focal-onset	100–200 mg/d; bid	7–10 h	Not established	Fatigue Dizziness Weakness Ataxia Mood changes	Gastrointestinal irritation	May increase carbamazepine-epoxide causing decreased tolerability May increase phenytoin
Carbamazepine	Tegretol[c]	Tonic-clonic Focal-onset	600–1800 mg/d (15–35 mg/kg, child); bid (capsules or tablets), tid-qid (oral suspension)	10–17 h (variable due to autoinduction: complete 3–5 wk after initiation)	4–12 µg/mL	Ataxia Dizziness Diplopia Vertigo	Aplastic anemia Leukopenia Gastrointestinal irritation Hepatotoxicity Hyponatremia	Level decreased by enzyme-inducing drugs[b] Level increased by erythromycin, propoxyphene, isoniazid, cimetidine, fluoxetine
Clobazam	Onfi	Lennox-Gastaut syndrome	10–40 mg/d (5–20 mg/d for patients <30 kg body weight); bid	36–42 h (71–82 h for less active metabolite)	Not established	Fatigue Sedation Ataxia Aggression Insomnia	Constipation Anorexia Skin rash	Level increased by CYP2C19 inhibitors
Clonazepam	Klonopin	Absence Atypical absence Myoclonic	1–12 mg/d; qd-tid	24–48 h	10–70 ng/mL	Ataxia Sedation Lethargy	Anorexia	Level decreased by enzyme-inducing drugs[b]
Ethosuximide	Zarontin	Absence	750–1250 mg/d (20–40 mg/kg); qd-bid	60 h, adult 30 h, child	40–100 µg/mL	Ataxia Lethargy Headache	Gastrointestinal irritation Skin rash Bone marrow suppression	Level decreased by enzyme-inducing drugs[b] Level increased by valproic acid
Felbamate	Felbatol	Focal-onset Lennox-Gastaut syndrome Tonic-clonic	2400–3600 mg/d, tid-qid	16–22 h	30–60 µg/mL	Insomnia Dizziness Sedation Headache	Aplastic anemia Hepatic failure Weight loss Gastrointestinal irritation	Increases phenytoin, valproic acid, active carbamazepine metabolite

Generic Name	Trade Name	Principal Uses	Typical Dose; Dose Interval	Half-Life	Therapeutic Range	Adverse Effects		Drug Interactions[a]
						Neurologic	Systemic	
Gabapentin	Neurontin	Focal-onset	900–2400 mg/d; tid-qid	5–9 h	2–20 µg/mL	Sedation Dizziness Ataxia Fatigue	Gastrointestinal irritation Weight gain Edema	No known significant interactions
Lacosamide	Vimpat	Focal-onset	200–400 mg/d; bid	13 h	Not established	Dizziness Ataxia Diplopia Vertigo	Gastrointestinal irritation Cardiac conduction (PR interval prolongation)	Level decreased by enzyme-inducing drugs[b]
Lamotrigine	Lamictal[c]	Focal-onset Tonic-clonic Atypical absence Myoclonic Lennox-Gastaut syndrome	150–500 mg/d; bid (immediate release), daily (extended release) (lower daily dose for regimens with valproic acid; higher daily dose for regimens with an enzyme inducer)	25 h 14 h (with enzyme-inducers), 59 h (with valproic acid)	2.5–20 µg/mL	Dizziness Diplopia Sedation Ataxia Headache	Skin rash Stevens-Johnson syndrome	Level decreased by enzyme-inducing drugs[b] and oral contraceptives Level increased by valproic acid
Levetiracetam	Keppra[c]	Focal-onset	1000–3000 mg/d; bid (immediate release), daily (extended release)	6–8 h	5–45 µg/mL	Sedation Fatigue Incoordination Mood changes	Anemia Leukopenia	No known significant interactions
Oxcarbazepine[c]	Trileptal	Focal-onset Tonic-clonic	900–2400 mg/d (30–45 mg/kg, child); bid	10–17 h (for active metabolite)	10–35 µg/mL	Fatigue Ataxia Dizziness Diplopia Vertigo Headache	See carbamazepine	Level decreased by enzyme-inducing drugs[b] May increase phenytoin

Generic Name	Trade Name	Principal Uses	Typical Dose; Dose Interval	Half-Life	Therapeutic Range	Adverse Effects		Drug Interactions[a]
						Neurologic	Systemic	
Phenobarbital	Luminal	Tonic-clonic Focal-onset	60–180 mg/d; qd-tid	90 h	10–40 µg/mL	Sedation Ataxia Confusion Dizziness Decreased libido Depression	Skin rash	Level increased by valproic acid, phenytoin
Phenytoin (diphenylhydantoin)	Dilantin	Tonic-clonic Focal-onset	300–400 mg/d (3–6 mg/kg, adult; 4–8 mg/kg, child); qd-tid	24 h (wide variation, dose-dependent)	10–20 µg/mL	Dizziness Diplopia Ataxia Incoordination Confusion	Gingival hyperplasia Lymphadenopathy Hirsutism Osteomalacia Facial coarsening Skin rash	Level increased by isoniazid, sulfonamides, fluoxetine Level decreased by enzyme-inducing drugs[b] Altered folate metabolism
Primidone	Mysoline	Tonic-clonic Focal-onset	750–1000 mg/d; bid-tid	Primidone, 8–15 h Phenobarbital, 90 h	Primidone, 4–12 µg/mL Phenobarbital, 10–40 µg/mL	Same as phenobarbital		Level increased by valproic acid Level decreased by phenytoin (increased conversion to phenobarbital)
Rufinamide	Banzel	Lennox-Gastaut syndrome	3200 mg/d (45 mg/kg, child); bid	6–10 h	Not established	Sedation Fatigue Dizziness Ataxia Headache Diplopia	Gastrointestinal irritation Leukopenia Cardiac conduction (QT interval shortening)	Level decreased by enzyme-inducing drugs[b] Level increased by valproic acid May increase phenytoin
Tiagabine	Gabitril	Focal-onset	32–56 mg/d; bid-qid (as adjunct to enzyme-inducing antiepileptic drug regimen)	2–5 h (with enzyme inducer), 7–9 h (without enzyme inducer)	Not established	Confusion Sedation Depression Dizziness Speech or language problems Paresthesias Psychosis	Gastrointestinal irritation	Level decreased by enzyme-inducing drugs[b]

Generic Name	Trade Name	Principal Uses	Typical Dose; Dose Interval	Half-Life	Therapeutic Range	Adverse Effects		Drug Interactions[a]
						Neurologic	Systemic	
Topiramate	Topamax	Focal-onset Tonic-clonic Lennox-Gastaut syndrome	200–400 mg/d; bid (immediate release), daily (extended release)	20 h (immediate release), 30 h (extended release)	2–20 µg/mL	Psychomotor slowing Sedation Speech or language problems Fatigue Paresthesias	Renal stones (avoid use with other carbonic anhydrase inhibitors) Glaucoma Weight loss Hypohidrosis	Level decreased by enzyme-inducing drugs[b]
Valproic acid (valproate sodium, divalproex sodium)	Depakene Depakote[c]	Tonic-clonic Absence Atypical absence Myoclonic Focal-onset Atonic	750–2000 mg/d (20–60 mg/kg); bid-qid (immediate and delayed release), daily (extended release)	15 h	50–125 µg/mL	Ataxia Sedation Tremor	Hepatotoxicity Thrombocytopenia Gastrointestinal irritation Weight gain Transient alopecia Hyperammonemia	Level decreased by enzyme-inducing drugs[b]
Zonisamide	Zonegran	Focal-onset Tonic-clonic	200–400 mg/d; qd-bid	50–68 h	10–40 µg/mL	Sedation Dizziness Confusion Headache Psychosis	Anorexia Renal stones Hypohidrosis	Level decreased by enzyme-inducing drugs[b]

[a]Examples only; please refer to other sources for comprehensive listings of all potential drug–drug interactions. [b]Phenytoin, carbamazepine, phenobarbital.

[c]Extended-release product available.

STATUS EPILEPTICUS

Status epilepticus refers to continuous seizures or repetitive, discrete seizures with impaired consciousness in the interictal period. Status epilepticus has numerous subtypes, including generalized convulsive status epilepticus (GCSE) (e.g., persistent, generalized electrographic seizures, coma, and tonic-clonic movements) and nonconvulsive status epilepticus (e.g., persistent absence seizures or focal seizures with confusion or partially impaired consciousness, and minimal motor abnormalities). The duration of seizure activity sufficient to meet the definition of status epilepticus has traditionally been specified as 15–30 min. However, a more practical definition is to consider status epilepticus as a situation in which the duration of seizures prompts the acute use of anticonvulsant therapy. For GCSE, this is typically when seizures last beyond 5 min.

GCSE is an emergency and must be treated immediately, because cardiorespiratory dysfunction, hyperthermia, and metabolic derangements can develop as a consequence of prolonged seizures, and these can lead to irreversible neuronal injury. Furthermore, CNS injury can occur even when the patient is paralyzed with neuromuscular blockade but continues to have electrographic seizures. The most common causes of GCSE are anticonvulsant withdrawal or noncompliance, metabolic disturbances, drug toxicity, CNS infection, CNS tumors, refractory epilepsy, and head trauma.

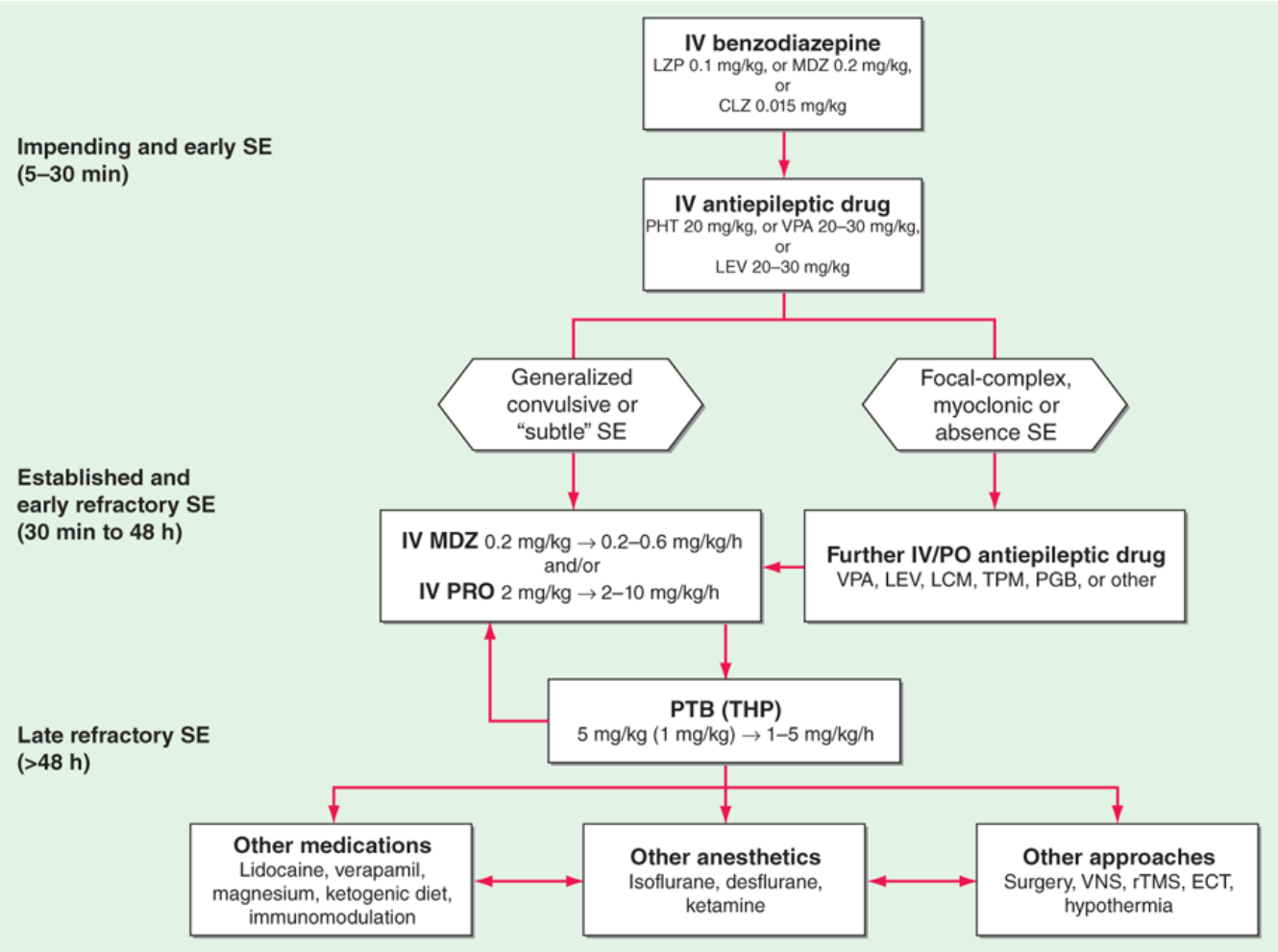
GCSE is obvious when the patient is having overt seizures. However, after 30–45 min of uninterrupted seizures, the signs may become increasingly subtle. Patients may have mild clonic movements of only the fingers or fine, rapid movements of the eyes. There may be paroxysmal episodes of tachycardia, hypertension, and pupillary dilation. In such cases, the EEG may be the only method of establishing the diagnosis. Thus, if the patient stops having overt seizures, yet remains comatose, an EEG should be performed to rule out ongoing status epilepticus. This is obviously also essential when a patient with GCSE has been paralyzed with neuromuscular blockade in the process of protecting the airway.

The first steps in the management of a patient in GCSE are to attend to any acute cardiorespiratory problems or hyperthermia, perform a brief medical and neurologic examination, establish venous access, and send samples for laboratory studies to identify metabolic abnormalities. Anticonvulsant

therapy should then begin without delay; a treatment approach is shown in Fig. 418-5.

FIGURE 418-5

Pharmacologic treatment of generalized tonic-clonic status epilepticus (SE) in adults. CLZ, clonazepam; ECT, electroconvulsive therapy; LCM, lacosamide; LEV, levetiracetam; LZP, lorazepam; MDZ, midazolam; PGB, pregabalin; PHT, phenytoin or fos-phenytoin; PRO, propofol; PTB, pentobarbital; rTMS, repetitive transcranial magnetic stimulation; THP, thiopental; TPM, topiramate; VNS, vagus nerve stimulation; VPA, valproic acid. *(From AO Rossetti, DH Lowenstein: Lancet Neurol 10:922, 2011.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The treatment of nonconvulsive status epilepticus is thought to be less urgent than GCSE, because the ongoing seizures are not accompanied by the severe metabolic disturbances seen with GCSE. However, evidence suggests that nonconvulsive status epilepticus, especially that caused by ongoing, focal seizure activity, is associated with cellular injury in the region of the seizure focus; therefore, this condition should be treated as promptly as possible using the general approach described for GCSE.

BEYOND SEIZURES: OTHER MANAGEMENT ISSUES

EPILEPSY COMORBIDITIES

The adverse effects of epilepsy often go beyond clinical seizures. Many epilepsy patients are completely normal between seizures and live highly successful and productive lives. However, a significant proportion of patients suffer from varying degrees of cognitive dysfunction, including psychiatric disease, and it has become increasingly clear that the network dysfunction underlying epilepsy can have effects well beyond the occurrence of seizures. For example, patients with seizures secondary to developmental abnormalities or acquired brain injury may have impaired cognitive function and other neurologic deficits due to abnormal brain structure. Frequent interictal EEG abnormalities are associated with subtle dysfunction of memory and attention. Patients with many seizures, especially those emanating from the temporal lobe, often note an impairment of short-term memory that may progress over time.

The psychiatric problems associated with epilepsy include depression, anxiety, and psychosis. This risk varies considerably depending on many factors, including the etiology, frequency, and severity of seizures and the patient's age and previous personal or family history of psychiatric disorder. Depression occurs in ~20% of patients, and the incidence of suicide is higher in patients with epilepsy than in the general population. Depression should be treated through counseling or medication. The selective serotonin reuptake inhibitors (SSRIs) typically have minimal effect on seizures, whereas tricyclic antidepressants may lower the seizure threshold. Anxiety can be a seizure symptom, and anxious or psychotic behavior can occur during a postictal delirium. Postictal psychosis is a rare phenomenon that typically occurs after a period of increased seizure frequency. There is usually a brief lucid interval lasting up to a week, followed by days to weeks of agitated, psychotic behavior. The psychosis usually resolves spontaneously but frequently will require short-term treatment with antipsychotic or anxiolytic medications.

MORTALITY OF EPILEPSY

Patients with epilepsy have a risk of death that is roughly two to three times greater than expected in a matched population without epilepsy. Most of the increased mortality is due to the underlying etiology of epilepsy (e.g., tumors or strokes in older adults). However, a significant number of patients die from accidents, status epilepticus, and a syndrome known as *sudden unexpected death in epilepsy* (SUDEP), which usually affects young people with convulsive seizures and tends to occur at night. The cause of SUDEP is unknown; it may result from brainstem-mediated effects of seizures on pulmonary, cardiac, and arousal functions. Recent studies suggest that, in some cases, a genetic mutation may be the cause of both epilepsy and a cardiac conduction defect that gives rise to sudden death.

PSYCHOSOCIAL ISSUES

There continues to be a cultural stigma about epilepsy, although it is slowly declining in societies with effective health education programs. Many patients with epilepsy harbor fears such as the fear of becoming mentally retarded or dying during a seizure. These issues need to be carefully addressed by educating the patient about epilepsy and by ensuring that family members, teachers, fellow employees, and other associates are equally well informed. A useful source of educational material is the website *www.epilepsy.com*.

EMPLOYMENT, DRIVING, AND OTHER ACTIVITIES

Many patients with epilepsy face difficulty in obtaining or maintaining employment, even when their seizures are well controlled. Federal and state legislation is designed to prevent employers from discriminating against patients with epilepsy, and patients should be encouraged to understand and claim their legal rights. Patients in these circumstances also benefit greatly from the assistance of health providers who act as strong patient advocates.

Loss of driving privileges is one of the most disruptive social consequences of epilepsy. Physicians should be very clear about local regulations concerning driving and epilepsy, because the laws vary considerably among states and countries. In all cases, it is the physician's responsibility to warn patients of the danger imposed on themselves and others while driving if their seizures are uncontrolled (unless the seizures are not associated with impairment of consciousness or motor control). In general, most states allow patients to drive after a seizure-free interval (on or off medications) of between 3 months and 2 years.

Patients with incompletely controlled seizures must also contend with the risk of being in other situations where an impairment of consciousness or loss of motor control could lead to major injury or death. Thus, depending on the type and frequency of seizures, many patients need to be instructed to avoid working at heights or with machinery or to have someone close by for activities such as bathing and swimming.

SPECIAL ISSUES RELATED TO WOMEN AND EPILEPSY

CATAMENIAL EPILEPSY

Some women experience a marked increase in seizure frequency around the time of menses. This is believed to be mediated by either the effects of estrogen and progesterone on neuronal excitability or changes in antiepileptic drug levels due to altered protein binding or metabolism. Some patients may benefit from increases in antiepileptic drug dosages during menses. Natural progestins or intramuscular medroxyprogesterone may be of benefit to a subset of women.

PREGNANCY

Most women with epilepsy who become pregnant will have an uncomplicated gestation and deliver a normal baby. However, epilepsy poses some important risks to a pregnancy. Seizure frequency during pregnancy will remain unchanged in ~50% of women, increase in ~30%, and decrease in ~20%. Changes in seizure frequency are attributed to endocrine effects on the CNS, variations in antiepileptic drug pharmacokinetics (such as acceleration of hepatic drug metabolism or effects on plasma protein binding), and changes in medication compliance. It is useful to see patients at

frequent intervals during pregnancy and monitor serum antiepileptic drug levels. Measurement of the unbound drug concentrations may be useful if there is an increase in seizure frequency or worsening of side effects of antiepileptic drugs.

The overall incidence of fetal abnormalities in children born to mothers with epilepsy is 5–6%, compared to 2–3% in healthy women. Part of the higher incidence is due to teratogenic effects of antiepileptic drugs, and the risk increases with the number of medications used (e.g., 10–20% risk of malformations with three drugs) and possibly with higher doses. A meta-analysis of published pregnancy registries and cohorts found that the most common malformations were defects in the cardiovascular and musculoskeletal system (1.4–1.8%). Valproic acid is strongly associated with an increased risk of adverse fetal outcomes (7–20%). Findings from a large pregnancy registry suggest that, other than topiramate, the newer antiepileptic drugs are far safer than valproic acid.

Because the potential harm of uncontrolled convulsive seizures on the mother and fetus is considered greater than the teratogenic effects of antiepileptic drugs, it is currently recommended that pregnant women be maintained on effective drug therapy. When possible, it seems prudent to have the patient on monotherapy at the lowest effective dose, especially during the first trimester. For some women, however, the type and frequency of their seizures may allow for them to safely wean off antiepileptic drugs prior to conception. Patients should also take folate (1–4 mg/d), because the antifolate effects of anticonvulsants are thought to play a role in the development of neural tube defects, although the benefits of this treatment remain unproved in this setting.

Enzyme-inducing drugs such as phenytoin, carbamazepine, oxcarbazepine, topiramate, phenobarbital, and primidone cause a transient and reversible deficiency of vitamin K–dependent clotting factors in ~50% of newborn infants. Although neonatal hemorrhage is uncommon, the mother should be treated with oral vitamin K (20 mg/d, phylloquinone) in the last 2 weeks of pregnancy, and the infant should receive intramuscular vitamin K (1 mg) at birth.

CONTRACEPTION

Special care should be taken when prescribing antiepileptic medications for women who are taking oral contraceptive agents. Drugs such as carbamazepine, phenytoin, phenobarbital, and topiramate can significantly decrease the efficacy of oral contraceptives via enzyme induction and other mechanisms. Patients should be advised to consider alternative forms of contraception, or their contraceptive medications should be modified to offset the effects of the antiepileptic medications.

BREAST-FEEDING

Antiepileptic medications are excreted into breast milk to a variable degree. The ratio of drug concentration in breast milk relative to serum ranges from ~5% (valproic acid) to 300% (levetiracetam). Given the overall benefits of breast-feeding and the lack of evidence for long-term harm to the infant by being exposed to antiepileptic drugs, mothers with epilepsy can be encouraged to breast-feed. This should be reconsidered, however, if there is any evidence of drug effects on the infant such as lethargy or poor feeding.

ACKNOWLEDGMENT

Dr. Michael J. Aminoff contributed to the section on EEG interpretation in earlier editions.

FURTHER READING

Chen DK et al: Psychogenic non-epileptic seizures. Curr Neurol Neurosci Rep 17:71, 2017. [PubMed: 28770498]

Crepeau AZ, Sirven JI: Management of adult onset seizures. Mayo Clin Proc 92:306, 2017. [PubMed: 28160877]

Dalkilic EB: Neurostimulation devices used in treatment of epilepsy. Curr Treat Options Neurol 19:7, 2017. [PubMed: 28251589]

Epi PM Consortium: A roadmap for precision medicine in the epilepsies. Lancet Neurol 14:1219, 2015. [PubMed: 26416172]

Gavvala JR, Schuele SU: New-onset seizure in adults and adolescents: A review. JAMA 316:2657, 2016. [PubMed: 28027373]

Golyala A, Kwan P: Drug development for refractory epilepsy: The past 25 years and beyond. Seizure 44:147, 2017. [PubMed: 28017578]

Jetté N et al: Surgical treatment for epilepsy: The potential gap between evidence and practice. Lancet Neurol 15:982, 2016. [PubMed: 27478955]

Keezer MR et al: Comorbidities of epilepsy: Current concepts and future perspectives. Lancet Neurol 15:106, 2016. [PubMed: 26549780]

Lamberink HJ et al: Individualised prediction model of seizure recurrence and long-term outcomes after withdrawal of antiepileptic drugs in seizure-free patients: A systematic review and individual participant data meta-analysis. Lancet Neurol 16:523, 2017. [PubMed: 28483337]

McGovern RA et al: New techniques and progress in epilepsy surgery. Curr Neurol Neurosci Rep 16:65, 2016. [PubMed: 27181271]

Pitkänen A et al: Advances in the development of biomarkers for epilepsy. Lancet Neurol 15:843, 2016. [PubMed: 27302363]

Chapter 419: Cerebrovascular Diseases

Wade S. Smith; S. Claiborne Johnston; J. Claude Hemphill, III

INTRODUCTION

Cerebrovascular diseases include some of the most common and devastating disorders: ischemic stroke and hemorrhagic stroke. Stroke is the second leading cause of death worldwide, with 6.2 million dying from stroke in 2015, an increase of 830,000 since the year 2000. While stroke has grown in incidence worldwide, it is declining among the affluent and rising among those with less access to medical care. In the United States, the incidence of stroke has declined steadily since at least 1958, and stroke is currently the fifth leading cause of death with 133,000 dying in 2014. Despite this progress, however, stroke remains the most common disabling disease in the United States and in many forms is preventable.

A stroke, or cerebrovascular accident, is defined as an abrupt onset of a neurologic deficit that is attributable to a focal vascular cause. Thus, the definition of stroke is clinical, and laboratory studies including brain imaging are used to support the diagnosis. The clinical manifestations of stroke are highly variable because of the complex anatomy of the brain and its vasculature. *Cerebral ischemia* is caused by a reduction in blood flow that lasts longer than several seconds. Neurologic symptoms are manifest within seconds because neurons lack glycogen, so energy failure is rapid. If the cessation of flow lasts for more than a few minutes, *infarction* or death of brain tissue results. When blood flow is quickly restored, brain tissue can recover fully and the patient's symptoms are only transient: this is called a *transient ischemic attack* (TIA). The definition of TIA requires that all neurologic signs and symptoms resolve within 24 h without evidence of brain infarction on brain imaging. Stroke has occurred if the neurologic signs and symptoms last for >24 h or brain infarction is demonstrated. A generalized reduction in cerebral blood flow due to systemic hypotension (e.g., cardiac arrhythmia, myocardial infarction, or hemorrhagic shock) usually produces syncope (Chap. 18). If low cerebral blood flow persists for a longer duration, then infarction in the border zones between the major cerebral artery distributions may develop. In more severe instances, *global hypoxia-ischemia* causes widespread brain injury; the constellation of cognitive sequelae that ensues is called *hypoxic-ischemic encephalopathy* (Chap. 301). *Focal ischemia* or infarction, conversely, is usually caused by thrombosis of the cerebral vessels themselves or by emboli from a proximal arterial source or the heart (Chap. 420). *Intracranial hemorrhage* is caused by bleeding directly into or around the brain; it produces neurologic symptoms by producing a mass effect on neural structures, from the toxic effects of blood itself, or by increasing intracranial pressure (Chap. 421).

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Cerebrovascular Disease

Rapid evaluation is essential for use of acute treatments such as thrombolysis or thrombectomy. However, patients with acute stroke often do not seek medical assistance on their own because they may lose the appreciation that something is wrong (anosognosia) or lack the knowledge that acute treatment is beneficial; it is often a family member or a bystander who calls for help. Therefore, patients and their family members should be counseled to call emergency medical services immediately if they experience or witness the sudden onset of any of the following: loss of sensory and/or motor function on one side of the body (nearly 85% of ischemic stroke patients have hemiparesis); change in vision, gait, or ability to speak or understand; or a sudden, severe headache. The acronym FAST (Facial weakness, Arm weakness, Speech abnormality and Time) is simple and helpful to

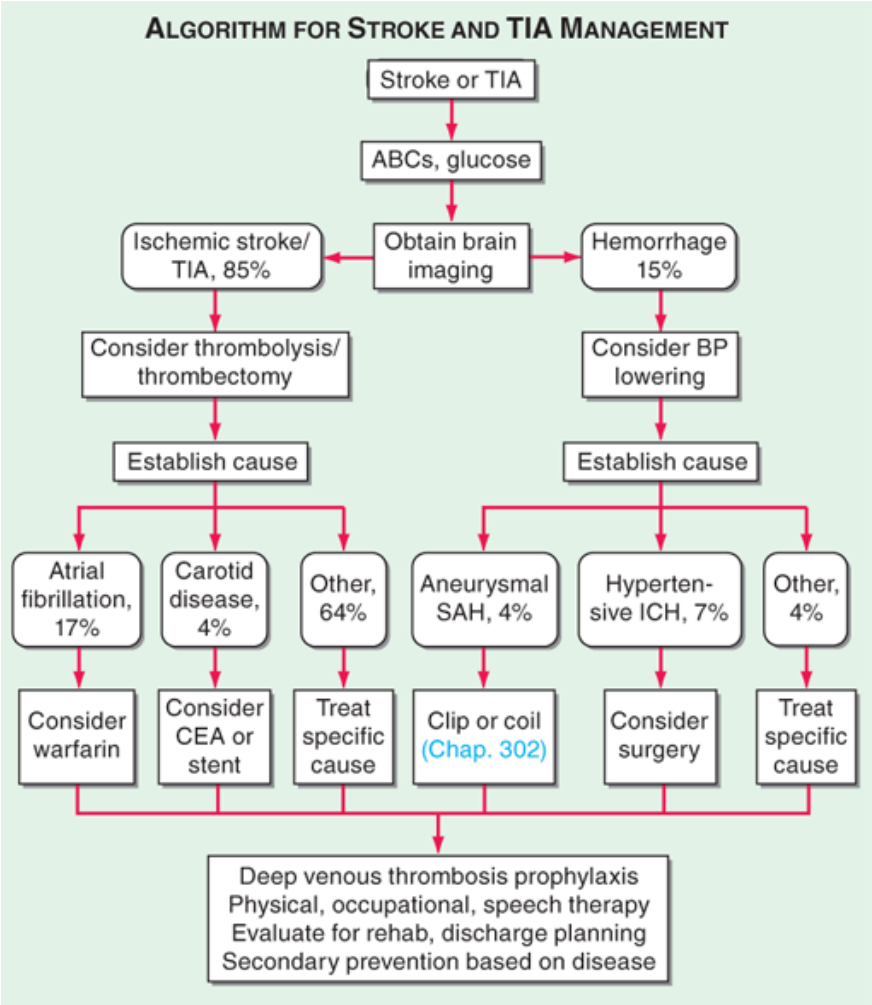
teach to the lay public about the common physical symptoms of stroke and to underscore that treatments are highly time-sensitive.

Other causes of sudden-onset neurologic symptoms that may mimic stroke include seizure, intracranial tumor, migraine, and metabolic encephalopathy. An adequate history from an observer that no convulsive activity occurred at the onset usually excludes seizure, although ongoing complex partial seizures without tonic-clonic activity can on occasion mimic stroke. Tumors may present with acute neurologic symptoms due to hemorrhage, seizure, or hydrocephalus. Surprisingly, migraine **(Chap. 422)** can mimic stroke, even in patients without a significant migraine history. When migraine develops without head pain (*acephalgic migraine*), the diagnosis can be especially difficult. Patients without any prior history of migraine may develop acephalgic migraine even after age 65. A sensory disturbance is often prominent, and the sensory deficit, as well as any motor deficits, tends to migrate slowly across a limb, over minutes rather than seconds as with stroke. The diagnosis of migraine becomes more secure as the cortical disturbance begins to cross vascular boundaries or if classic visual symptoms are present such as scintillating scotomata. At times, it may be impossible to make the diagnosis of migraine until there have been multiple episodes with no residual symptoms or signs and no changes on brain magnetic resonance imaging (MRI). Metabolic encephalopathies typically produce fluctuating mental status changes without focal neurologic findings. However, in the setting of prior stroke or brain injury, a patient with fever or sepsis may manifest a recurrent hemiparesis, which clears rapidly when the infection is treated. The metabolic process serves to “unmask” a prior deficit.

Once the diagnosis of stroke is made, a brain imaging study is necessary to determine if the cause of stroke is ischemia or hemorrhage **(Fig. 419-1)**. Computed tomography (CT) imaging of the brain is the standard imaging modality to detect the presence or absence of intracranial hemorrhage (see “Imaging Studies,” below). If the stroke is ischemic, administration of recombinant tissue plasminogen activator (rtPA) or endovascular mechanical thrombectomy may be beneficial in restoring cerebral perfusion **(Chap. 420)**. Medical management to reduce the risk of complications becomes the next priority, followed by plans for secondary prevention. For ischemic stroke, several strategies can reduce the risk of subsequent stroke in all patients, while other strategies are effective for patients with specific causes of stroke such as cardiac embolus and carotid atherosclerosis. For hemorrhagic stroke, aneurysmal subarachnoid hemorrhage (SAH) and hypertensive intracerebral hemorrhage are two important causes. **The treatment and prevention of hypertensive intracerebral hemorrhage are discussed in Chap. 421. SAH is discussed in Chap. 302**.

FIGURE 419-1

Medical management of stroke and TIA. *Rounded boxes* are diagnoses; *rectangles* are interventions. Numbers are percentages of stroke overall. ABCs, airway, breathing, circulation; BP, blood pressure; CEA, carotid endarterectomy; ICH, intracerebral hemorrhage; SAH, subarachnoid hemorrhage; TIA, transient ischemic attack.



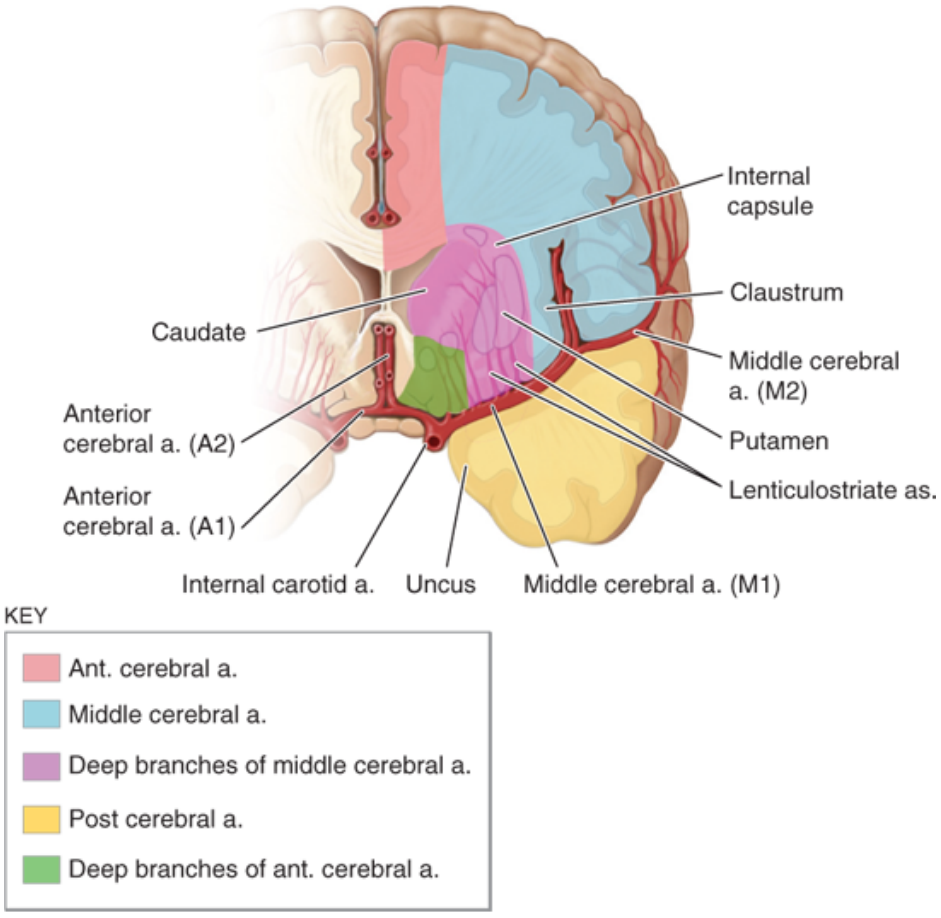
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

STROKE SYNDROMES

A careful history and neurologic examination can often localize the region of brain dysfunction; if this region corresponds to an arterial distribution, the possible causes responsible for the syndrome can be narrowed. This is of particular importance when the patient presents with a TIA and a normal examination. For example, if a patient develops language loss and a right homonymous hemianopia, a search for causes of left middle cerebral emboli should be performed. A finding of an isolated stenosis of the right internal carotid artery in that patient, for example, suggests an asymptomatic carotid stenosis, and the search for other causes of stroke should continue. The following sections describe the clinical findings of cerebral ischemia associated with cerebral vascular territories depicted in Figs. 419-2, 419-3, 419-4, 419-5, 419-6, 419-7, 419-8, 419-9, 419-10, 419-11. Stroke syndromes are divided into: (1) large-vessel stroke within the anterior circulation, (2) large-vessel stroke within the posterior circulation, and (3) small-vessel disease of either vascular bed.

FIGURE 419-2

Diagram of a cerebral hemisphere in coronal section showing the territories of the major cerebral vessels that branch from the internal carotid arteries.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 419-3

Diagram of a cerebral hemisphere, lateral aspect, showing the branches and distribution of the middle cerebral artery (MCA) and the principal regions of cerebral localization. Note the bifurcation of the MCA into a superior and inferior division.

Signs and symptoms: *Structures involved*

Paralysis of the contralateral face, arm, and leg; sensory impairment over the same area (pinprick, cotton touch, vibration, position, two-point discrimination, stereognosis, tactile localization, barognosis, cutaneographia): *Somatic motor area for face and arm and the fibers descending from the leg area to enter the corona radiata and corresponding somatic sensory system*

Motor aphasia: *Motor speech area of the dominant hemisphere*

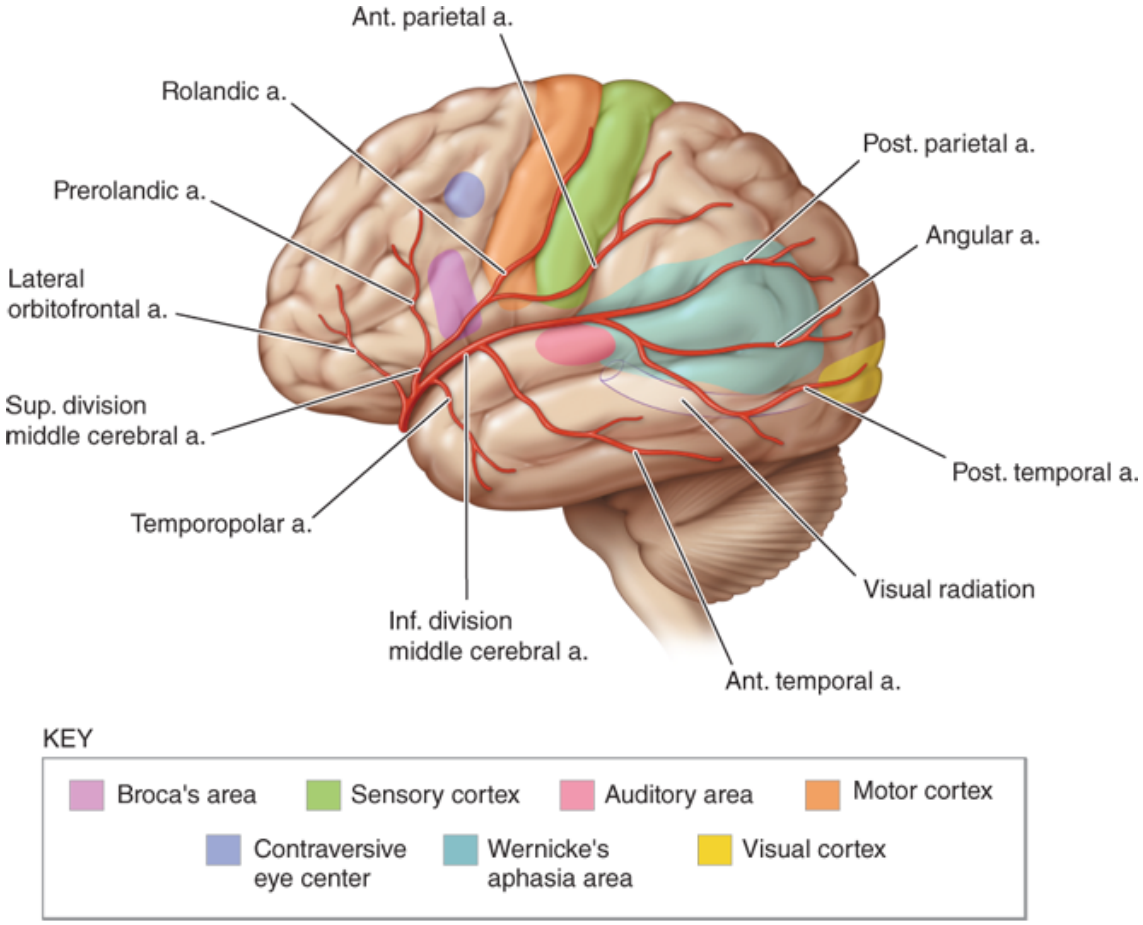
Central aphasia, word deafness, anomia, jargon speech, sensory agraphia, acalculia, alexia, finger agnosia, right-left confusion (the last four comprise the Gerstmann syndrome): *Central, suprasylvian speech area and parietooccipital cortex of the dominant hemisphere*

Conduction aphasia: *Central speech area (parietal operculum)*

Apractagnosia of the nondominant hemisphere, anosognosia, hemiasomatognosia, unilateral neglect, agnosia for the left half of external space, dressing "apraxia," constructional "apraxia," distortion of visual coordinates, inaccurate localization in the half field, impaired ability to judge distance, upside-down reading, visual illusions (e.g., it may appear that another person walks through a table): *Nondominant parietal lobe (area corresponding to speech area in dominant hemisphere); loss of topographic memory is usually due to a nondominant lesion, occasionally to a dominant one*

Homonymous hemianopia (often homonymous inferior quadrantanopia): *Optic radiation deep to second temporal convolution*

Paralysis of conjugate gaze to the opposite side: *Frontal contraversive eye field or projecting fibers*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 419-4

Diagram of a cerebral hemisphere, medial aspect, showing the branches and distribution of the anterior cerebral artery and the principal regions of cerebral localization.

Signs and symptoms: *Structures involved*

Paralysis of opposite foot and leg: *Motor leg area*

A lesser degree of paresis of opposite arm: *Arm area of cortex or fibers descending to corona radiata*

Cortical sensory loss over toes, foot, and leg: *Sensory area for foot and leg*

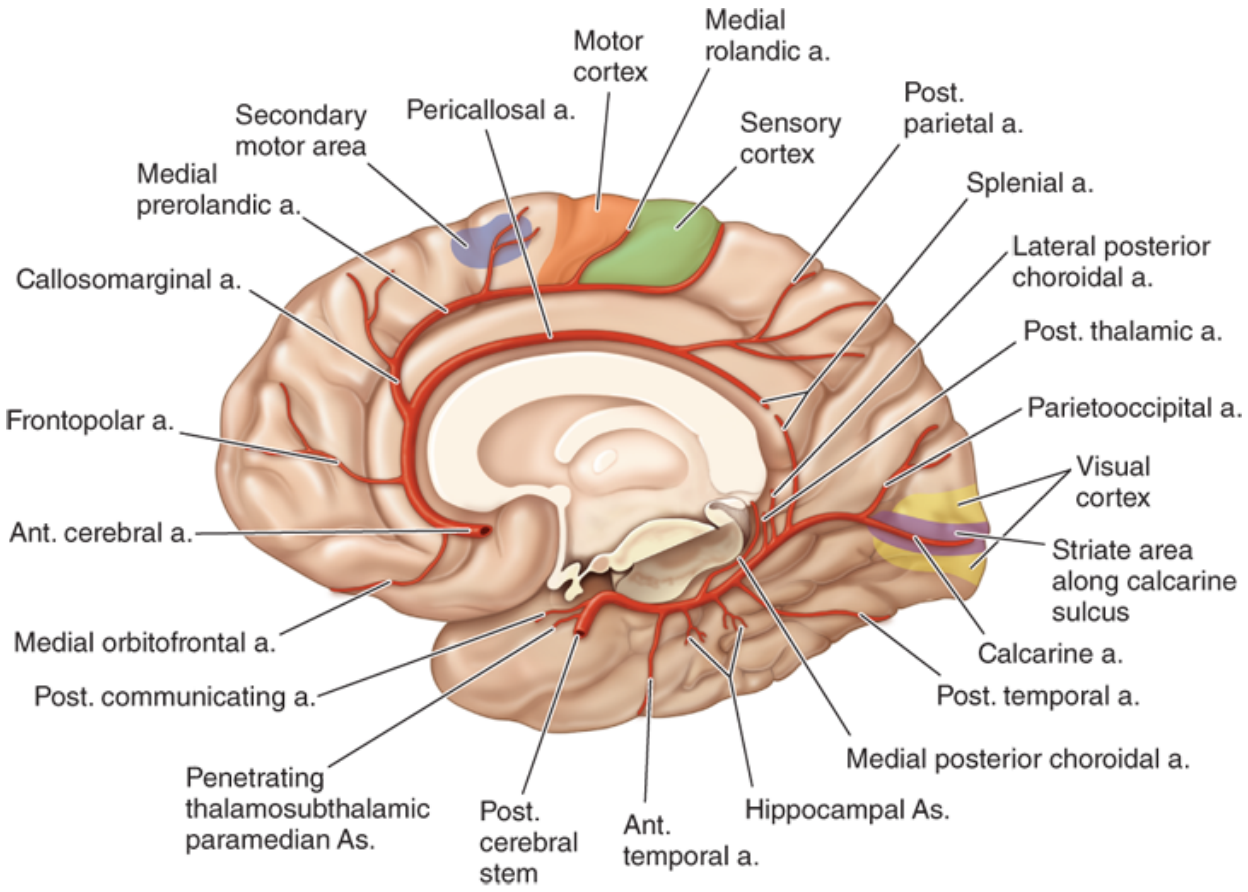
Urinary incontinence: *Sensorimotor area in paracentral lobule*

Contralateral grasp reflex, sucking reflex, gegenhalten (paratonic rigidity): *Medial surface of the posterior frontal lobe; likely supplemental motor area*

Abulia (akinetic mutism), slowness, delay, intermittent interruption, lack of spontaneity, whispering, reflex distraction to sights and sounds: *Uncertain localization—probably cingulate gyrus and medial inferior portion of frontal, parietal, and temporal lobes*

Impairment of gait and stance (gait apraxia): *Frontal cortex near leg motor area*

Dyspraxia of left limbs, tactile aphasia in left limbs: Corpus callosum



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 419-5

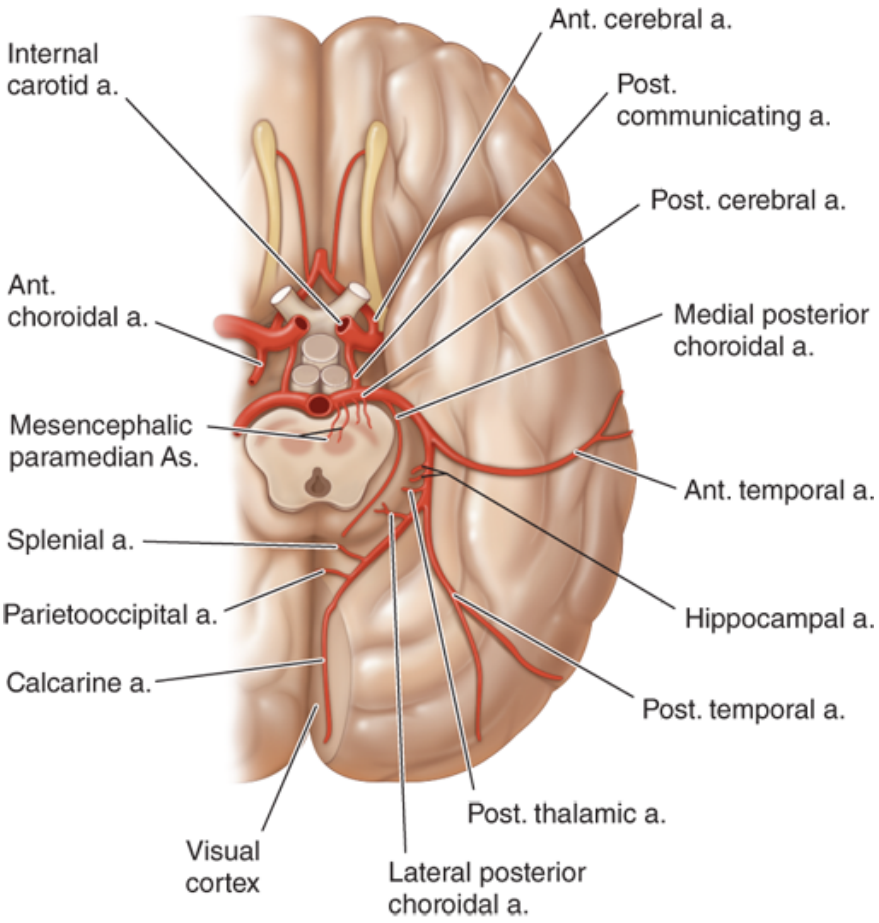
Inferior aspect of the brain with the branches and distribution of the posterior cerebral artery and the principal anatomic structures shown.

Signs and symptoms: *Structures involved*

Peripheral territory (see also Fig. 419-9). Homonymous hemianopia (often upper quadrantic): *Calcarine cortex or optic radiation nearby*. Bilateral homonymous hemianopia, cortical blindness, awareness or denial of blindness; tactile naming, achromatopia (color blindness), failure to see to-and-fro movements, inability to perceive objects not centrally located, apraxia of ocular movements, inability to count or enumerate objects, tendency to run into things that the patient sees and tries to avoid: *Bilateral occipital lobe with possibly the parietal lobe involved*. Verbal dyslexia without agraphia, color anomia: *Dominant calcarine lesion and posterior part of corpus callosum*. Memory defect: *Hippocampal lesion bilaterally or on the dominant side only*. Topographic disorientation and prosopagnosia: *Usually with lesions of nondominant, calcarine, and lingual gyrus*. Simultanagnosia, hemivisual neglect: *Dominant visual cortex, contralateral hemisphere*. Unformed visual hallucinations, peduncular hallucinosis, metamorphopsia, teleopsia, illusory visual spread, palinopsia, distortion of outlines, central photophobia: *Calcarine cortex*. Complex hallucinations: *Usually nondominant hemisphere*.

Central territory. Thalamic syndrome: sensory loss (all modalities), spontaneous pain and dysesthesias, choreoathetosis, intention tremor, spasms of hand, mild hemiparesis: *Posteroventral nucleus of thalamus; involvement of the adjacent subthalamus body or its afferent tracts*. Thalamoperforate syndrome: crossed cerebellar ataxia with ipsilateral third nerve palsy (Claude's syndrome): *Dentatothalamic tract and issuing third nerve*. Weber's syndrome: third nerve palsy and contralateral hemiplegia: *Third nerve and cerebral peduncle*. Contralateral hemiplegia: *Cerebral peduncle*. Paralysis or paresis of vertical eye movement, skew deviation, sluggish pupillary responses to light, slight miosis and ptosis (retraction nystagmus and "tucking" of the eyelids may be associated): *Supranuclear fibers to third nerve, interstitial nucleus of Cajal, nucleus of Darkschewitsch, and*

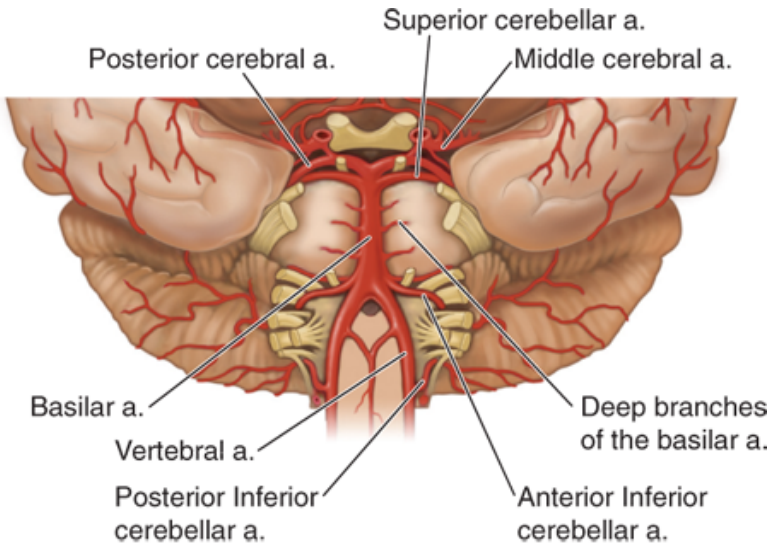
posterior commissure. Contralateral rhythmic, ataxic action tremor; rhythmic postural or “holding” tremor (rubral tremor): *Dentatothalamic tract*.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 419-6

Diagram of the posterior circulation, showing the intracranial vertebral arteries forming the basilar artery that gives off the anterior inferior cerebellar, superior cerebellar, and posterior cerebral arteries. The posterior inferior cerebellar artery arises from each of the vertebral segments. The majority of brainstem blood flow arises from numerous deep branches of the basilar artery that penetrate directly into the brainstem.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 419-7

Axial section at the level of the medulla, depicted schematically on the left, with a corresponding magnetic resonance image on the right. Note that in Figs. 419-7, 419-8, 419-9, 419-10, 419-11, all drawings are oriented with the dorsal surface at the bottom, matching the orientation of the brainstem that is commonly seen in all modern neuroimaging studies. Approximate regions involved in medial and lateral medullary stroke syndromes are shown.

Signs and symptoms: *Structures involved*

1. Medial medullary syndrome (occlusion of vertebral artery or of branch of vertebral or lower basilar artery)

   On side of lesion

   Paralysis with atrophy of one-half half the tongue: *Ipsilateral twelfth nerve*

   On side opposite lesion

   Paralysis of arm and leg, sparing face; impaired tactile and proprioceptive sense over one-half the body: *Contralateral pyramidal tract and medial lemniscus*

2. Lateral medullary syndrome (occlusion of any of five vessels may be responsible—vertebral, posterior inferior cerebellar, superior, middle, or inferior lateral medullary arteries)

   On side of lesion

   Pain, numbness, impaired sensation over one-half the face: *Descending tract and nucleus fifth nerve*

   Ataxia of limbs, falling to side of lesion: *Uncertain—restiform body, cerebellar hemisphere, cerebellar fibers, spinocerebellar tract (?)*

   Nystagmus, diplopia, oscillopsia, vertigo, nausea, vomiting: *Vestibular nucleus*

   Horner's syndrome (miosis, ptosis, decreased sweating): *Descending sympathetic tract*

   Dysphagia, hoarseness, paralysis of palate, paralysis of vocal cord, diminished gag reflex: *Issuing fibers ninth and tenth nerves*

   Loss of taste: *Nucleus and tractus solitarius*

Numbness of ipsilateral arm, trunk, or leg: *Cuneate and gracile nuclei*

Weakness of lower face: *Genuflected upper motor neuron fibers to ipsilateral facial nucleus*

On side opposite lesion

Impaired pain and thermal sense over half the body, sometimes face: *Spinothalamic tract*

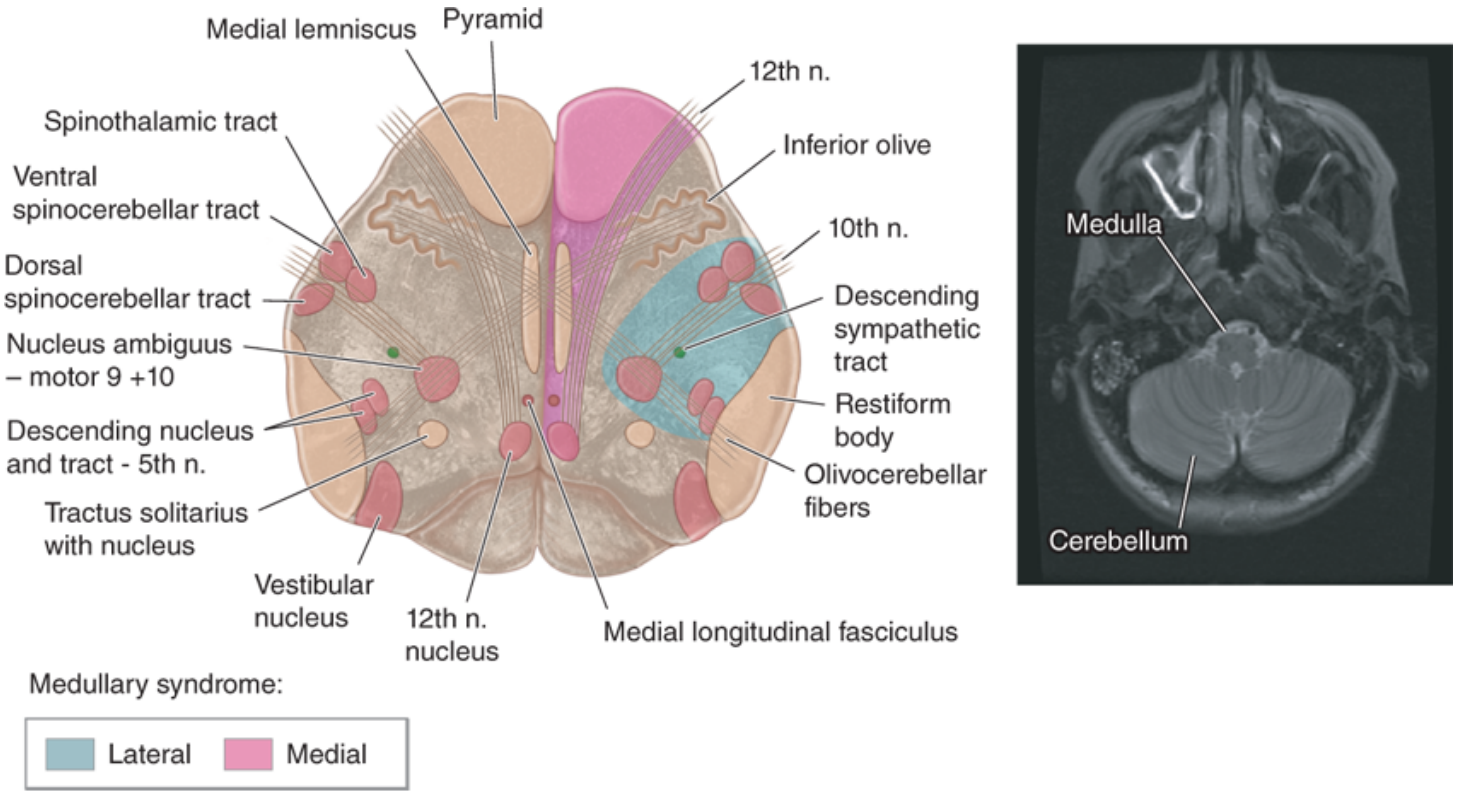
3. Total unilateral medullary syndrome (occlusion of vertebral artery): Combination of medial and lateral syndromes

4. Lateral pontomedullary syndrome (occlusion of vertebral artery): Combination of lateral medullary and lateral inferior pontine syndrome

5. Basilar artery syndrome (the syndrome of the lone vertebral artery is equivalent): A combination of the various brainstem syndromes plus those arising in the posterior cerebral artery distribution.

   Bilateral long tract signs (sensory and motor; cerebellar and peripheral cranial nerve abnormalities): *Bilateral long tract; cerebellar and peripheral cranial nerves*

   Paralysis or weakness of all extremities, plus all bulbar musculature: *Corticobulbar and corticospinal tracts bilaterally*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 419-8

Axial section at the level of the inferior pons, depicted schematically on the left, with a corresponding magnetic resonance image on the right. Approximate regions involved in medial and lateral inferior pontine stroke syndromes are shown.

Signs and symptoms: *Structures involved*

1. Medial inferior pontine syndrome (occlusion of paramedian branch of basilar artery)

   On side of lesion

Paralysis of conjugate gaze to side of lesion (preservation of convergence): *Center for conjugate lateral gaze*

Nystagmus: *Vestibular nucleus*

Ataxia of limbs and gait: Likely *middle cerebellar peduncle*

Diplopia on lateral gaze: *Abducens nerve*

On side opposite lesion

Paralysis of face, arm, and leg: *Corticobulbar and corticospinal tract in lower pons*

Impaired tactile and proprioceptive sense over one-half of the body: *Medial lemniscus*

2. Lateral inferior pontine syndrome (occlusion of anterior inferior cerebellar artery)

On side of lesion

Horizontal and vertical nystagmus, vertigo, nausea, vomiting, oscillopsia: *Vestibular nerve or nucleus*

Facial paralysis: *Seventh nerve*

Paralysis of conjugate gaze to side of lesion: *Center for conjugate lateral gaze*

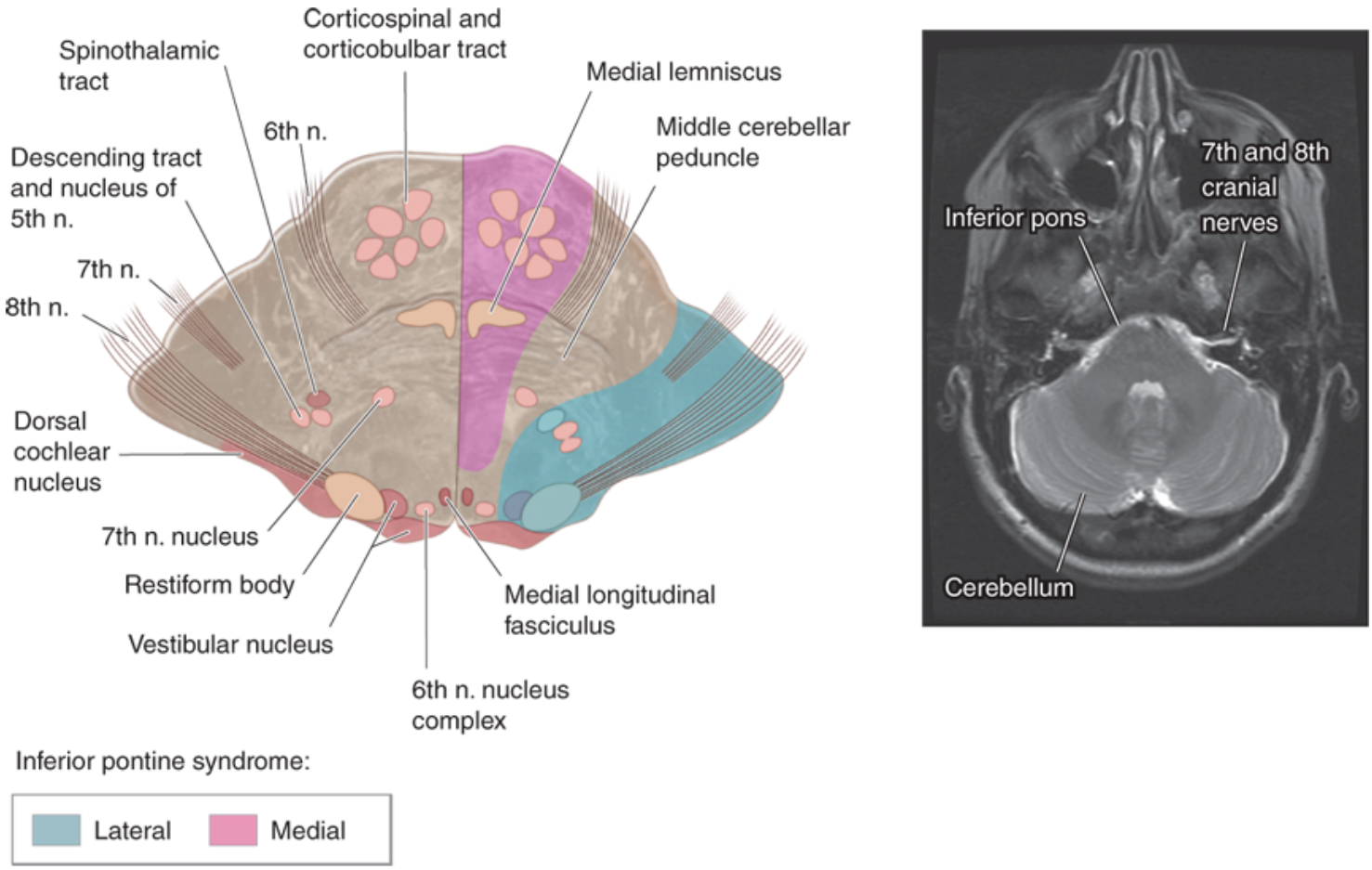
Deafness, tinnitus: *Auditory nerve or cochlear nucleus*

Ataxia: *Middle cerebellar peduncle and cerebellar hemisphere*

Impaired sensation over face: *Descending tract and nucleus fifth nerve*

On side opposite lesion

Impaired pain and thermal sense over one-half the body (may include face): *Spinothalamic tract*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 419-9

Axial section at the level of the midpons, depicted schematically on the left, with a corresponding magnetic resonance image on the right. Approximate regions involved in medial and lateral midpontine stroke syndromes are shown.

Signs and symptoms: *Structures involved*

1. Medial midpontine syndrome (paramedian branch of midbasilar artery)

   On side of lesion

   Ataxia of limbs and gait (more prominent in bilateral involvement): *Pontine nuclei*

   On side opposite lesion

   Paralysis of face, arm, and leg: *Corticobulbar and corticospinal tract*

   Variable impaired touch and proprioception when lesion extends posteriorly: *Medial lemniscus*

2. Lateral midpontine syndrome (short circumferential artery)

   On side of lesion

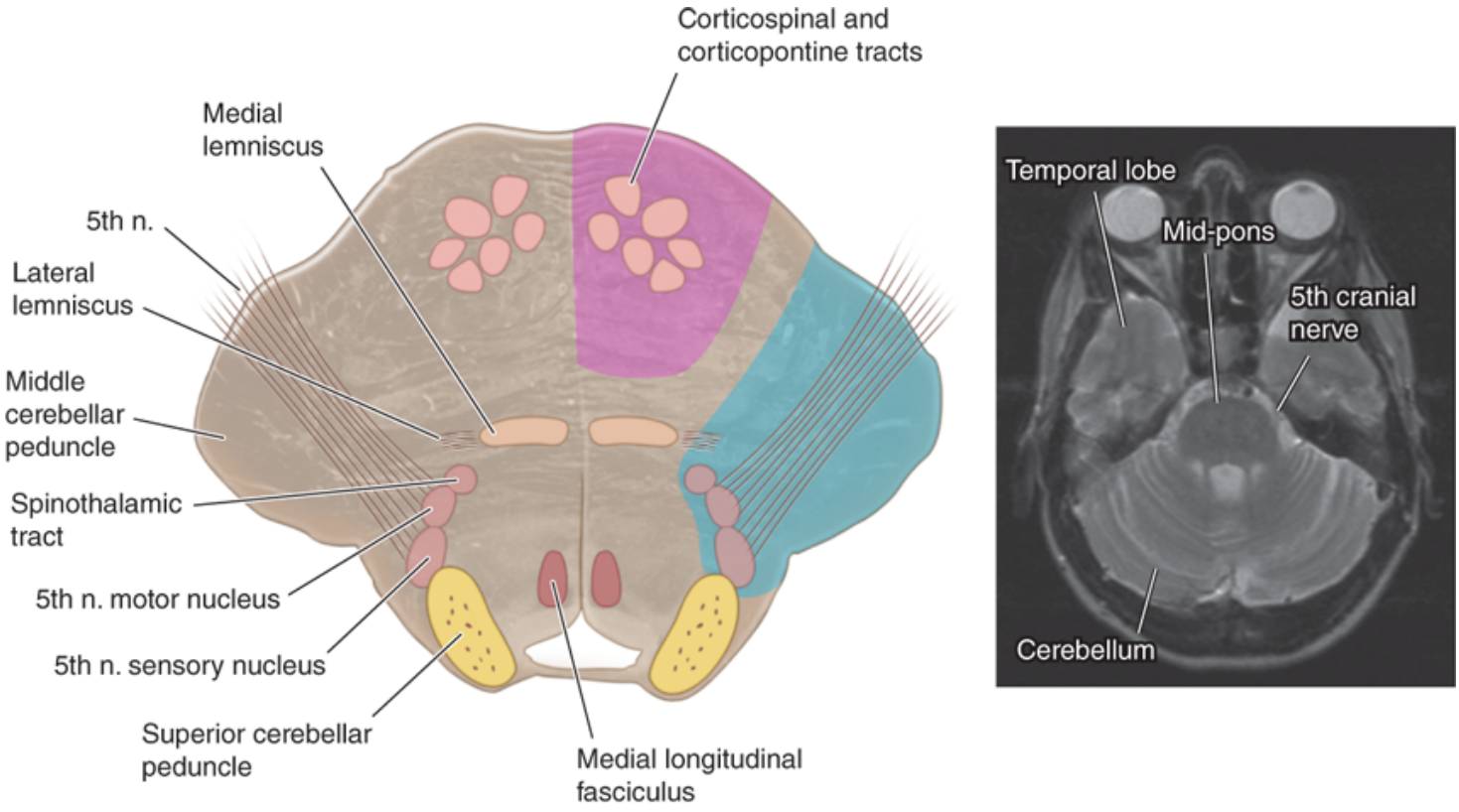
   Ataxia of limbs: *Middle cerebellar peduncle*

   Paralysis of muscles of mastication: *Motor fibers or nucleus of fifth nerve*

   Impaired sensation over side of face: *Sensory fibers or nucleus of fifth nerve*

On side opposite lesion

Impaired pain and thermal sense on limbs and trunk: *Spinothalamic tract*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 419-10

Axial section at the level of the superior pons, depicted schematically on the left, with a corresponding magnetic resonance image on the right. Approximate regions involved in medial and lateral superior pontine stroke syndromes are shown.

Signs and symptoms: *Structures involved*

1. Medial superior pontine syndrome (paramedian branches of upper basilar artery)

   On side of lesion

   Cerebellar ataxia (probably): *Superior and/or middle cerebellar peduncle*

   Internuclear ophthalmoplegia: *Medial longitudinal fasciculus*

   Myoclonic syndrome, palate, pharynx, vocal cords, respiratory apparatus, face, oculomotor apparatus, etc.: *Localization uncertain—central tegmental bundle, dentate projection, inferior olivary nucleus*

   On side opposite lesion

   Paralysis of face, arm, and leg: *Corticobulbar and corticospinal tract*

   Rarely touch, vibration, and position are affected: *Medial lemniscus*

2. Lateral superior pontine syndrome (syndrome of superior cerebellar artery)

   On side of lesion

   Ataxia of limbs and gait, falling to side of lesion: *Middle and superior cerebellar peduncles, superior surface of cerebellum, dentate nucleus*

   Dizziness, nausea, vomiting; horizontal nystagmus: *Vestibular nucleus*

   Paresis of conjugate gaze (ipsilateral): *Pontine contralateral gaze*

   Skew deviation: *Uncertain*

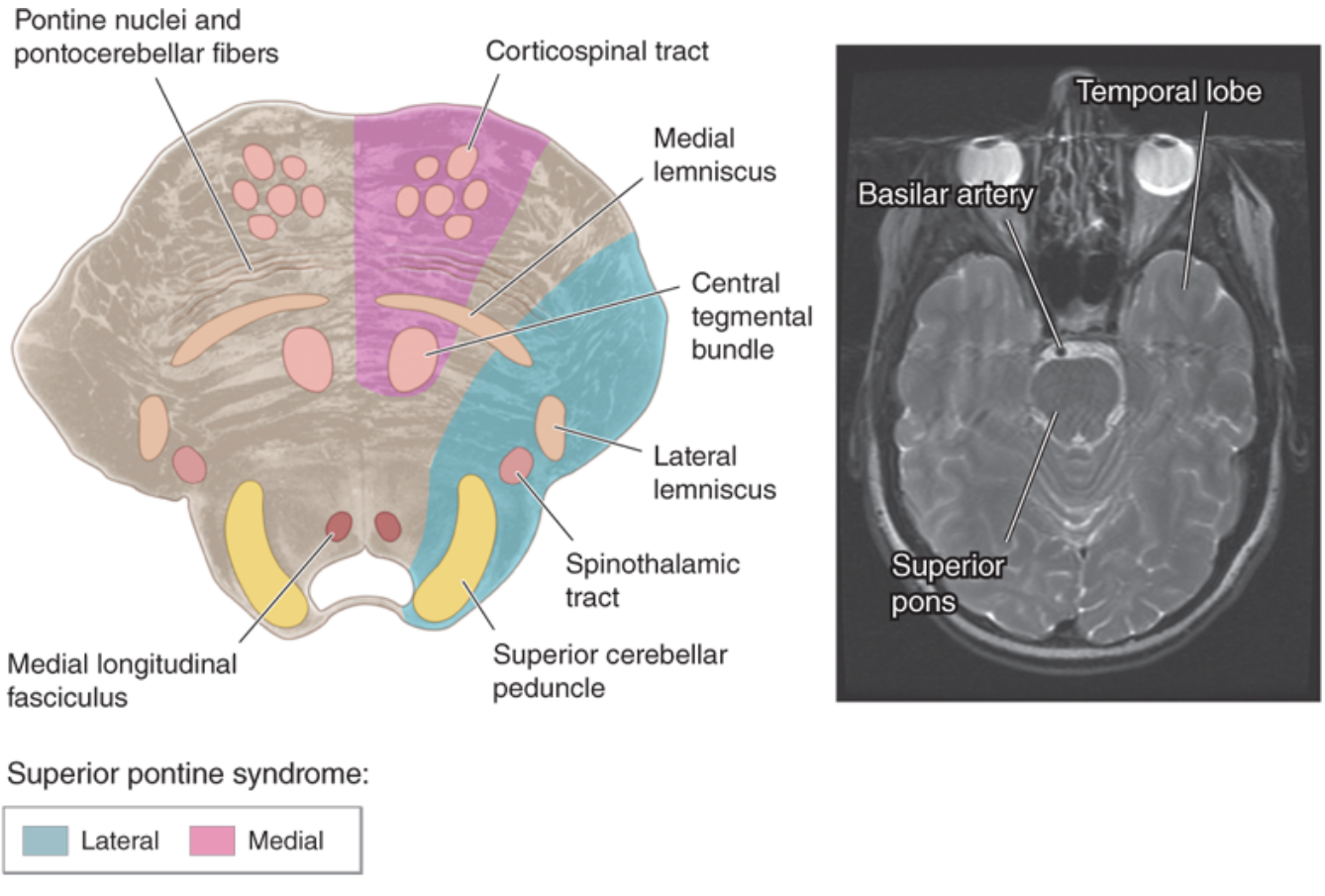
   Miosis, ptosis, decreased sweating over face (Horner's syndrome): *Descending sympathetic fibers*

   Tremor: Localization unclear—*Dentate nucleus, superior cerebellar peduncle*

   On side opposite lesion

   Impaired pain and thermal sense on face, limbs, and trunk: *Spinothalamic tract*

   Impaired touch, vibration, and position sense, more in leg than arm (there is a tendency to incongruity of pain and touch deficits): *Medial lemniscus (lateral portion)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 419-11

Axial section at the level of the midbrain, depicted schematically on the left, with a corresponding magnetic resonance image on the right. Approximate regions involved in medial and lateral midbrain stroke syndromes are shown.

Signs and symptoms: *Structures involved*

1. Medial midbrain syndrome (paramedian branches of upper basilar and proximal posterior cerebral arteries)

   On side of lesion

   Eye “down and out” secondary to unopposed action of fourth and sixth cranial nerves, with dilated and unresponsive pupil: *Third nerve fibers*

   On side opposite lesion

   Paralysis of face, arm, and leg: *Corticobulbar and corticospinal tract descending in crus cerebri*

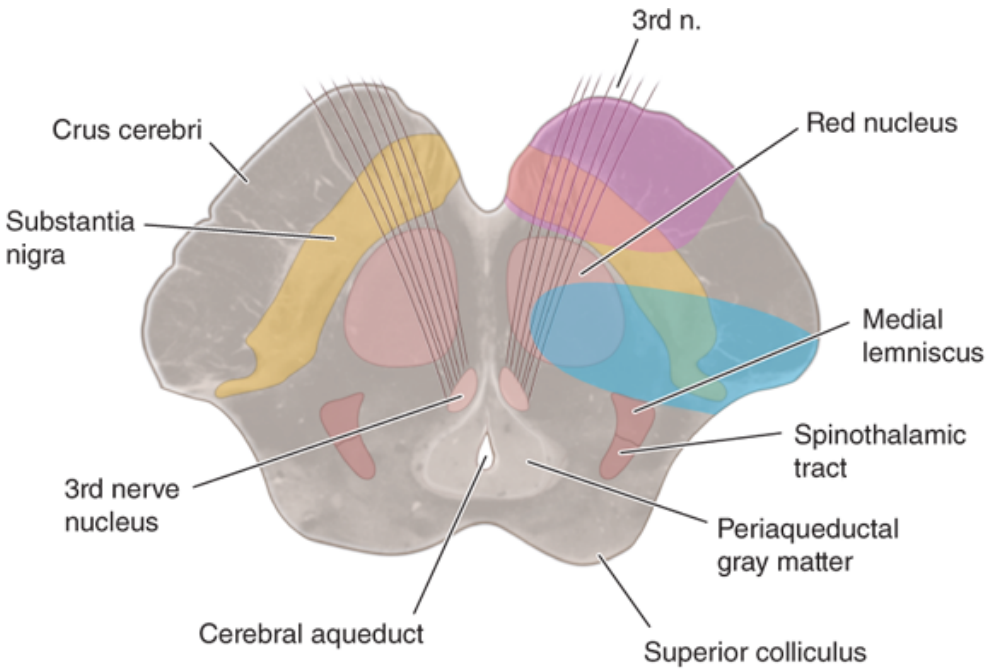
2. Lateral midbrain syndrome (syndrome of small penetrating arteries arising from posterior cerebral artery)

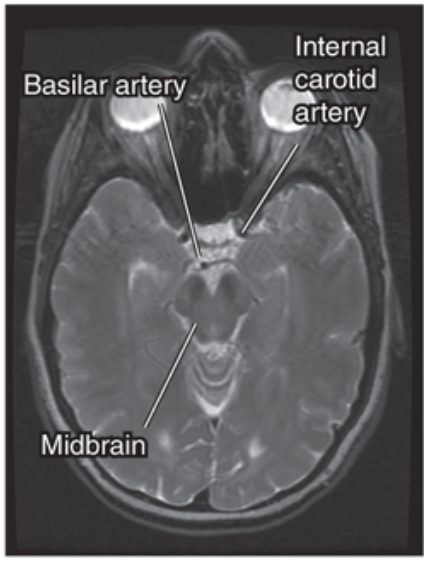
   On side of lesion

   Eye “down and out” secondary to unopposed action of fourth and sixth cranial nerves, with dilated and unresponsive pupil: *Third nerve fibers and/or third nerve nucleus*

   On side opposite lesion

   Hemiataxia, hyperkinesias, tremor: *Red nucleus, dentatorubrothalamic pathway*





Midbrain syndrome:

Lateral Medial

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Stroke within the Anterior Circulation

The internal carotid artery and its branches comprise the anterior circulation of the brain. These vessels can be occluded by intrinsic disease of the vessel (e.g., atherosclerosis or dissection) or by embolic occlusion from a proximal source as discussed

above. Occlusion of each major intracranial vessel has distinct clinical manifestations.

MIDDLE CEREBRAL ARTERY

Occlusion of the proximal middle cerebral artery (MCA) or one of its major branches is most often due to an embolus (artery-to-artery, cardiac, or of unknown source) rather than intracranial atherothrombosis. Atherosclerosis of the proximal MCA may cause distal emboli to the middle cerebral territory or, less commonly, may produce low-flow TIAs. Collateral formation via leptomeningeal vessels often prevents MCA stenosis from becoming symptomatic.

The cortical branches of the MCA supply the lateral surface of the hemisphere except for (1) the frontal pole and a strip along the superomedial border of the frontal and parietal lobes supplied by the anterior cerebral artery (ACA) and (2) the lower temporal and occipital pole convolutions supplied by the posterior cerebral artery (PCA) (Figs. 419-2, Figs. 419-3, Figs. 419-4, 419-5).

The proximal MCA (M1 segment) gives rise to penetrating branches (termed *lenticulostriate arteries*) that supply the putamen, outer globus pallidus, posterior limb of the internal capsule, adjacent corona radiata, and most of the caudate nucleus (Fig. 419-2). In the sylvian fissure, the MCA in most patients divides into *superior* and *inferior* divisions (M2 branches). Branches of the inferior division supply the inferior parietal and temporal cortex, and those from the superior division supply the frontal and superior parietal cortex (Fig. 419-3).

If the entire MCA is occluded at its origin (blocking both its penetrating and cortical branches) and the distal collaterals are limited, the clinical findings are contralateral hemiplegia, hemianesthesia, homonymous hemianopia, and a day or two of gaze preference to the ipsilateral side. Dysarthria is common because of facial weakness. When the dominant hemisphere is involved, global aphasia is present also, and when the nondominant hemisphere is affected, anosognosia, constructional apraxia, and neglect are found **(Chap. 26)**.

Complete MCA syndromes occur most often when an embolus occludes the stem of the artery. Cortical collateral blood flow and differing arterial configurations are probably responsible for the development of many partial syndromes. Partial syndromes may also be due to emboli that enter the proximal MCA without complete occlusion, occlude distal MCA branches, or fragment and move distally.

Partial syndromes due to embolic occlusion of a single branch include hand, or arm and hand, weakness alone (brachial syndrome) or facial weakness with nonfluent (Broca) aphasia **(Chap. 26)**, with or without arm weakness (frontal opercular syndrome). A combination of sensory disturbance, motor weakness, and nonfluent aphasia suggests that an embolus has occluded the proximal superior division and infarcted large portions of the frontal and parietal cortices (Fig. 419-3). If a fluent (Wernicke's) aphasia occurs without weakness, the inferior division of the MCA supplying the posterior part (temporal cortex) of the dominant hemisphere is probably involved. Jargon speech and an inability to comprehend written and spoken language are prominent features, often accompanied by a contralateral, homonymous superior quadrantanopia. Hemineglect or spatial agnosia without weakness indicates that the inferior division of the MCA in the nondominant hemisphere is involved.

Occlusion of a lenticulostriate vessel produces small-vessel (lacunar) stroke within the internal capsule (Fig. 419-2). This produces pure motor stroke or sensory-motor stroke contralateral to the lesion. Ischemia within the genu of the internal capsule causes primarily facial weakness followed by arm and then leg weakness as the ischemia moves posterior within the capsule. Alternatively, the contralateral hand may become ataxic, and dysarthria will be prominent (clumsy hand, dysarthria lacunar syndrome). Lacunar infarction affecting the globus pallidus and putamen often has few clinical signs, but parkinsonism and hemiballismus have been reported.

ANTERIOR CEREBRAL ARTERY

The ACA is divided into two segments: the precommunal (A1) circle of Willis, or stem, which connects the internal carotid artery to the anterior communicating artery, and the postcommunal (A2) segment distal to the anterior communicating artery (Figs.

419-2 and 419-4). The A1 segment gives rise to several deep penetrating branches that supply the anterior limb of the internal capsule, the anterior perforate substance, amygdala, anterior hypothalamus, and the inferior part of the head of the caudate nucleus.

Occlusion of the proximal ACA is usually well tolerated because of collateral flow through the anterior communicating artery and collaterals through the MCA and PCA. Occlusion of a single A2 segment results in the contralateral symptoms noted in Fig. 419-4. If both A2 segments arise from a single anterior cerebral stem (contralateral A1 segment atresia), the occlusion may affect both hemispheres. Profound abulia (a delay in verbal and motor response) and bilateral pyramidal signs with paraparesis or quadriparesis and urinary incontinence result.

ANTERIOR CHOROIDAL ARTERY

This artery arises from the internal carotid artery and supplies the posterior limb of the internal capsule and the white matter posterolateral to it, through which pass some of the geniculocalcarine fibers (Fig. 419-5). The complete syndrome of anterior choroidal artery occlusion consists of contralateral hemiplegia, hemianesthesia (hypesthesia), and homonymous hemianopia. However, because this territory is also supplied by penetrating vessels of the proximal MCA and the posterior communicating and posterior choroidal arteries, minimal deficits may occur, and patients frequently recover substantially. Anterior choroidal strokes are usually the result of in situ thrombosis of the vessel, and the vessel is particularly vulnerable to iatrogenic occlusion during surgical clipping of aneurysms arising from the internal carotid artery.

INTERNAL CAROTID ARTERY

The clinical picture of internal carotid occlusion varies depending on whether the cause of ischemia is propagated thrombus, embolism, or low flow. The cortex supplied by the MCA territory is affected most often. With a competent circle of Willis, occlusion may go unnoticed. If the thrombus propagates up the internal carotid artery into the MCA or embolizes it, symptoms are identical to proximal MCA occlusion (see above). Sometimes there is massive infarction of the entire deep white matter and cortical surface. When the origins of both the ACA and MCA are occluded at the top of the carotid artery, abulia or stupor occurs with hemiplegia, hemianesthesia, and aphasia or anosognosia. When the PCA arises from the internal carotid artery (a configuration called a *fetal PCA*), it may also become occluded and give rise to symptoms referable to its peripheral territory (Figs. 419-4 and 419-5).

In addition to supplying the ipsilateral brain, the internal carotid artery perfuses the optic nerve and retina via the ophthalmic artery. In ~25% of symptomatic internal carotid disease, recurrent transient monocular blindness (amaurosis fugax) warns of the lesion. Patients typically describe a horizontal shade that sweeps down or up across the field of vision. They may also complain that their vision was blurred in that eye or that the upper or lower half of vision disappeared. In most cases, these symptoms last only a few minutes. Rarely, ischemia or infarction of the ophthalmic artery or central retinal arteries occurs at the time of cerebral TIA or infarction.

A high-pitched prolonged carotid bruit fading into diastole is often associated with tightly stenotic lesions. As the stenosis grows tighter and flow distal to the stenosis becomes reduced, the bruit becomes fainter and may disappear when occlusion is imminent.

COMMON CAROTID ARTERY

All symptoms and signs of internal carotid occlusion may also be present with occlusion of the common carotid artery. Jaw claudication may result from low flow in the external carotid branches. Bilateral common carotid artery occlusions at their origin may occur in Takayasu's arteritis **(Chap. 356)**.

Stroke within the Posterior Circulation

The posterior circulation is composed of the paired vertebral arteries, the basilar artery, and the paired PCAs. The vertebral arteries join to form the basilar artery at the pontomedullary junction. The basilar artery divides into two PCAs in the interpeduncular fossa (Figs. 419-4, 419-5, 419-6). These major arteries give rise to long and short circumferential branches and to smaller deep penetrating branches that supply the cerebellum, medulla, pons, midbrain, subthalamus, thalamus, hippocampus, and medial temporal and occipital lobes. Occlusion of each vessel produces its own distinctive syndrome.

POSTERIOR CEREBRAL ARTERY

In 75% of cases, both PCAs arise from the bifurcation of the basilar artery; in 20%, one has its origin from the ipsilateral internal carotid artery via the posterior communicating artery; in 5%, both originate from the respective ipsilateral internal carotid arteries (Figs. 419-4, 419-5, 419-6). The precommunal, or P1, segment of the true PCA is atretic in such cases.

PCA syndromes usually result from atheroma formation or emboli that lodge at the top of the basilar artery; posterior circulation disease may also be caused by dissection of either vertebral artery or fibromuscular dysplasia.

Two clinical syndromes are commonly observed with occlusion of the PCA: (1) *P1 syndrome*: midbrain, subthalamic, and thalamic signs, which are due to disease of the proximal P1 segment of the PCA or its penetrating branches (thalamogeniculate, Percheron, and posterior choroidal arteries); and (2) *P2 syndrome*: cortical temporal and occipital lobe signs, due to occlusion of the P2 segment distal to the junction of the PCA with the posterior communicating artery.

P1 SYNDROMES

Infarction usually occurs in the ipsilateral subthalamus and medial thalamus and in the ipsilateral cerebral peduncle and midbrain (Figs. 419-5 and 419-11). A third nerve palsy with contralateral ataxia (Claude's syndrome) or with contralateral hemiplegia (Weber's syndrome) may result. The ataxia indicates involvement of the red nucleus or dentatorubrothalamic tract; the hemiplegia is localized to the cerebral peduncle (Fig. 419-11). If the subthalamic nucleus is involved, contralateral hemiballismus may occur. Occlusion of the artery of Percheron produces paresis of upward gaze and drowsiness and often abulia. Extensive infarction in the midbrain and subthalamus occurring with bilateral proximal PCA occlusion presents as coma, unreactive pupils, bilateral pyramidal signs, and decerebrate rigidity.

Occlusion of the penetrating branches of thalamic and thalamogeniculate arteries produces less extensive thalamic and thalamocapsular lacunar syndromes. The *thalamic Déjérine-Roussy syndrome* consists of contralateral hemisensory loss followed later by an agonizing, searing, or burning pain in the affected areas. It is persistent and responds poorly to analgesics. Anticonvulsants (carbamazepine or gabapentin) or tricyclic antidepressants may be beneficial.

P2 SYNDROMES

(Figs. 419-4 and 419-5) Occlusion of the distal PCA causes infarction of the medial temporal and occipital lobes. Contralateral homonymous hemianopia without macula sparing is the usual manifestation. (MCA strokes often produce hemianopia but typically spare the macula as calcarine cortex is perfused by the P2 segment). Occasionally, only the upper quadrant of visual field is involved or the macula vision is spared. If the visual association areas are spared and only the calcarine cortex is involved, the patient may be aware of visual defects. Medial temporal lobe and hippocampal involvement may cause an acute disturbance in memory, particularly if it occurs in the dominant hemisphere. The defect usually clears because memory has bilateral representation. If the dominant hemisphere is affected and the infarct extends to involve the splenium of the corpus callosum, the patient may demonstrate alexia without agraphia. Visual agnosia for faces, objects, mathematical symbols, and colors and anomia with paraphasic errors (amnestic aphasia) may also occur, even without callosal involvement. Occlusion of the PCA can produce *peduncular hallucinosis* (visual hallucinations of brightly colored scenes and objects).

Bilateral infarction in the distal PCAs produces cortical blindness (blindness with preserved pupillary light reaction). The patient is often unaware of the blindness or may even deny it (*Anton's syndrome*). Tiny islands of vision may persist, and the patient

may report that vision fluctuates as images are captured in the preserved portions. Rarely, only peripheral vision is lost and central vision is spared, resulting in “gun-barrel” vision. Bilateral visual association area lesions may result in *Balint’s syndrome*, a disorder of the orderly visual scanning of the environment **(Chap. 26)**, usually resulting from infarctions secondary to low flow in the “watershed” between the distal PCA and MCA territories, as occurs after cardiac arrest. Patients may experience persistence of a visual image for several minutes despite gazing at another scene (*palinopsia*) or an inability to synthesize the whole of an image (*asimultanagnosia*). Embolic occlusion of the top of the basilar artery can produce any or all the central or peripheral territory symptoms. The hallmark is the sudden onset of bilateral signs, including ptosis, pupillary asymmetry or lack of reaction to light, and somnolence. Patients will often have posturing and myoclonic jerking that simulates seizure. Interrogation of the noncontrast CT scan for a hyperdense basilar artery sign (indicating thrombus in the basilar artery), or CT angiography (CTA) establishes this diagnosis. Physicians should be suspicious of this rare, but potentially treatable stroke syndrome in the setting of presumed new onset seizure and cranial nerve deficits.

VERTEBRAL AND POSTERIOR INFERIOR CEREBELLAR ARTERIES

The vertebral artery, which arises from the innominate artery on the right and the subclavian artery on the left, consists of four segments. The first (V1) extends from its origin to its entrance into the sixth or fifth transverse vertebral foramen. The second segment (V2) traverses the vertebral foramina from C6 to C2. The third (V3) passes through the transverse foramen and circles around the arch of the atlas to pierce the dura at the foramen magnum. The fourth (V4) segment courses upward to join the other vertebral artery to form the basilar artery (Fig. 419-6); only the fourth segment gives rise to branches that supply the brainstem and cerebellum. The posterior inferior cerebellar artery (PICA) in its proximal segment supplies the lateral medulla and, in its distal branches, the inferior surface of the cerebellum.

Atherothrombotic lesions have a predilection for V1 and V4 segments of the vertebral artery. The first segment may become diseased at the origin of the vessel and may produce posterior circulation emboli; collateral flow from the contralateral vertebral artery or the ascending cervical, thyrocervical, or occipital arteries is usually sufficient to prevent low-flow TIAs or stroke. When one vertebral artery is atretic and an atherothrombotic lesion threatens the origin of the other, the collateral circulation, which may also include retrograde flow down the basilar artery, is often insufficient (Figs. 419-5 and 419-6). In this setting, low-flow TIAs may occur, consisting of syncope, vertigo, and alternating hemiplegia; this state also sets the stage for thrombosis. Disease of the distal fourth segment of the vertebral artery can promote thrombus formation manifest as embolism or with propagation as basilar artery thrombosis. Stenosis proximal to the origin of the PICA can threaten the lateral medulla and posterior inferior surface of the cerebellum.

If the subclavian artery is occluded proximal to the origin of the vertebral artery, there is a reversal in the direction of blood flow in the ipsilateral vertebral artery. Exercise of the ipsilateral arm may increase demand on vertebral flow, producing posterior circulation TIAs, or “subclavian steal.”

Although atheromatous disease rarely narrows the second and third segments of the vertebral artery, this region is subject to dissection, fibromuscular dysplasia, and, rarely, encroachment by osteophytic spurs within the vertebral foramina.

Embolic occlusion or thrombosis of a V4 segment causes ischemia of the lateral medulla. The constellation of vertigo, numbness of the ipsilateral face and contralateral limbs, diplopia, hoarseness, dysarthria, dysphagia, and ipsilateral Horner’s syndrome is called the *lateral medullary (or Wallenberg’s) syndrome* (Fig. 419-7). Ipsilateral upper motor neuron facial weakness can also occur. Most cases result from ipsilateral vertebral artery occlusion; in the remainder, PICA occlusion is responsible. Occlusion of the medullary penetrating branches of the vertebral artery or PICA results in partial syndromes. *Hemiparesis is not a typical feature of vertebral artery occlusion; however, quadriparesis may result from occlusion of the anterior spinal artery.*

Rarely, a *medial medullary syndrome* occurs with infarction of the pyramid and contralateral hemiparesis of the arm and leg, sparing the face. If the medial lemniscus and emerging hypoglossal nerve fibers are involved, contralateral loss of joint position sense and ipsilateral tongue weakness occur.

Cerebellar infarction can lead to *respiratory arrest* due to raised intracranial pressure from cerebellar swelling, closure of the aqueduct of Silvius or fourth ventricle, followed by hydrocephalus and central herniation. Displacement of the brainstem from cerebellar edema will also cause respiratory and hemodynamic instability. Drowsiness, Babinski signs, dysarthria, and bifacial weakness may be absent, or present only briefly, before respiratory arrest ensues. Gait unsteadiness, headache, dizziness, nausea, and vomiting may be the only early symptoms and signs and should arouse suspicion of this impending complication, which may require neurosurgical decompression, often with an excellent outcome. Separating these symptoms from those of viral labyrinthitis can be a challenge, but headache, neck stiffness, and unilateral dysmetria favor stroke.

BASILAR ARTERY

Branches of the basilar artery (Fig. 419-6) supply the base of the pons and superior cerebellum and fall into three groups: (1) paramedian, 7–10 in number, which supply a wedge of pons on either side of the midline; (2) short circumferential, 5–7 in number, that supply the lateral two-thirds of the pons and middle and superior cerebellar peduncles; and (3) bilateral long circumferential (superior cerebellar and anterior inferior cerebellar arteries), which course around the pons to supply the cerebellar hemispheres.

Atheromatous lesions can occur anywhere along the basilar trunk but are most frequent in the proximal basilar and distal vertebral segments. Typically, lesions occlude either the proximal basilar and one or both vertebral arteries. The clinical picture varies depending on the availability of retrograde collateral flow from the posterior communicating arteries. Rarely, dissection of a vertebral artery may involve the basilar artery and, depending on the location of true and false lumen, may produce multiple penetrating artery strokes.

Although atherothrombosis occasionally occludes the distal portion of the basilar artery, emboli from the heart or proximal vertebral or basilar segments are more commonly responsible for “top of the basilar” syndromes.

Because the brainstem contains many structures in close apposition, a diversity of clinical syndromes may emerge with ischemia, reflecting involvement of the corticospinal and corticobulbar tracts, ascending sensory tracts, and cranial nerve nuclei (Figs. 419-7, 419-8, 419-9, 419-10, 419-11).

The symptoms of transient ischemia or infarction in the territory of the basilar artery often do not indicate whether the basilar artery itself or one of its branches is diseased, yet this distinction has important implications for therapy. *The picture of complete basilar occlusion, however, is easy to recognize as a constellation of bilateral long tract signs (sensory and motor) with signs of cranial nerve and cerebellar dysfunction*. Patients may have spontaneous posturing movements that are myoclonic in nature and simulate seizure activity. A “locked-in” state of preserved consciousness with quadriplegia and cranial nerve signs suggests complete pontine and lower midbrain infarction. The therapeutic goal is to identify *impending* basilar occlusion before devastating infarction occurs. A series of TIAs and a slowly progressive, fluctuating stroke are extremely significant, because they often herald an atherothrombotic occlusion of the distal vertebral or proximal basilar artery.

TIAs in the proximal basilar distribution may produce vertigo (often described by patients as “swimming,” “swaying,” “moving,” “unsteadiness,” or “light-headedness”). Other symptoms that warn of basilar thrombosis include diplopia, dysarthria, facial or circumoral numbness, and hemisensory symptoms. In general, symptoms of basilar branch TIAs affect one side of the brainstem, whereas symptoms of basilar artery TIAs usually affect both sides, although a “herald” hemiparesis has been emphasized as an initial symptom of basilar occlusion. Most often, TIAs, whether due to impending occlusion of the basilar artery or a basilar branch, are short lived (5–30 min) and repetitive, occurring several times a day. The pattern suggests intermittent reduction of flow. Although treatment with intravenous heparin or various combinations of antiplatelet agents have been used to prevent clot propagation there is no specific evidence to support any one approach, and endovascular intervention is also an option.

Atherothrombotic occlusion of the basilar artery with infarction usually causes *bilateral* brainstem signs. A gaze paresis or internuclear ophthalmoplegia associated with ipsilateral hemiparesis may be the only manifestation of bilateral brainstem

ischemia. More often, unequivocal signs of bilateral pontine disease are present. Complete basilar thrombosis carries a high mortality.

Occlusion of a branch of the basilar artery usually causes *unilateral* symptoms and signs involving motor, sensory, and cranial nerves. If symptoms remain unilateral, concern over pending basilar occlusion should be reduced.

Occlusion of the superior cerebellar artery results in severe ipsilateral cerebellar ataxia, nausea and vomiting, dysarthria, and contralateral loss of pain and temperature sensation over the extremities, body, and face (spino- and trigeminothalamic tract). Partial deafness, ataxic tremor of the ipsilateral upper extremity, Horner's syndrome, and palatal myoclonus may occur rarely. Partial syndromes occur frequently (Fig. 419-10). With large strokes, swelling and mass effects may compress the midbrain or produce hydrocephalus; these symptoms may evolve rapidly. Neurosurgical intervention may be lifesaving in such cases.

Occlusion of the anterior inferior cerebellar artery produces variable degrees of infarction because the size of this artery and the territory it supplies vary inversely with those of the PICA. The principal symptoms include: (1) ipsilateral deafness, facial weakness, vertigo, nausea and vomiting, nystagmus, tinnitus, cerebellar ataxia, Horner's syndrome, and paresis of conjugate lateral gaze; and (2) contralateral loss of pain and temperature sensation. An occlusion close to the origin of the artery may cause corticospinal tract signs (Fig. 419-8).

Occlusion of one of the short circumferential branches of the basilar artery affects the lateral two-thirds of the pons and middle or superior cerebellar peduncle, whereas occlusion of one of the paramedian branches affects a wedge-shaped area on either side of the medial pons (Figs. 419-8, 419-9, 419-10).

IMAGING STUDIES

See also Chap. 416.

CT Scans

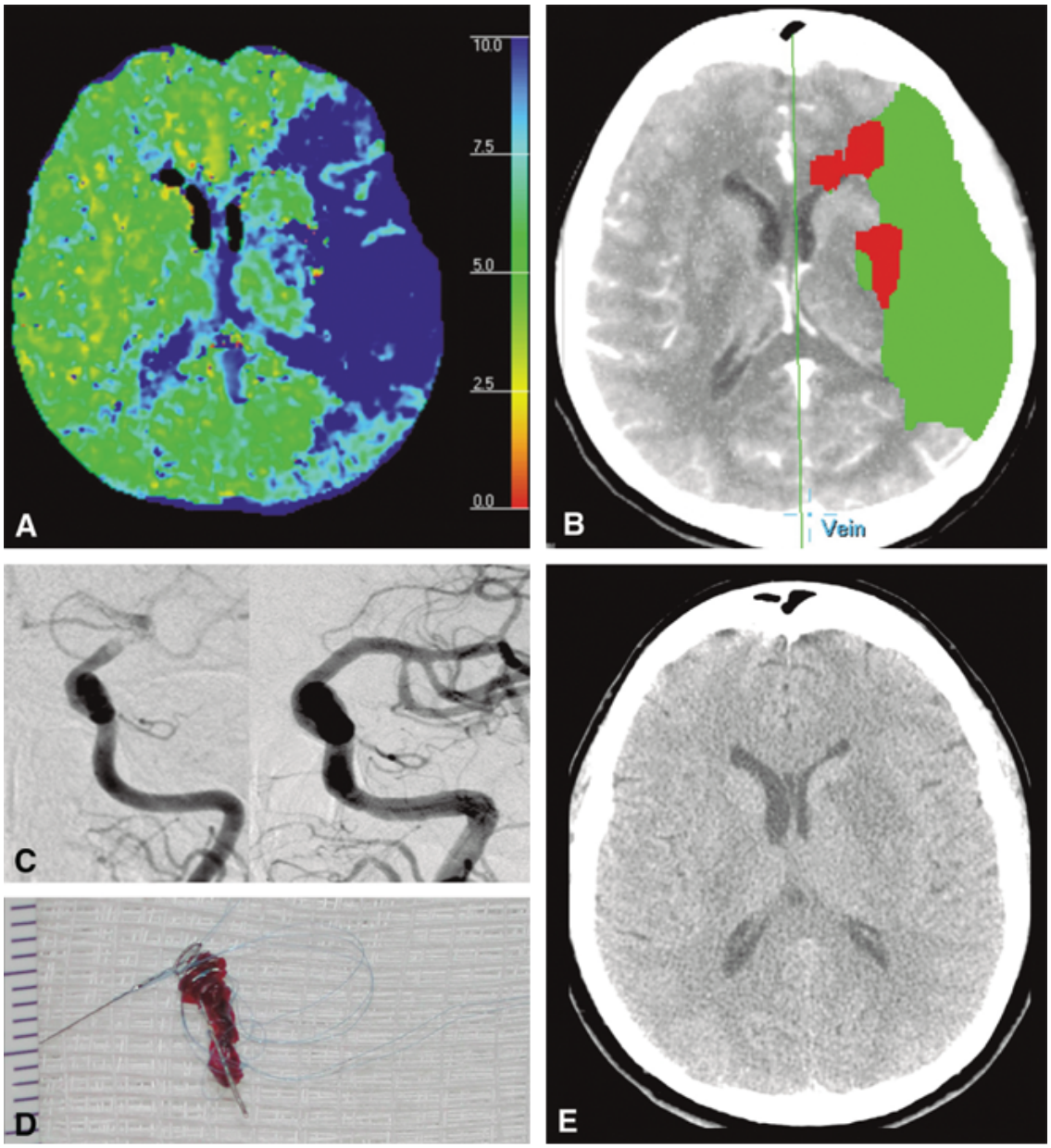
CT radiographic images identify or exclude hemorrhage as the cause of stroke, and they identify extraparenchymal hemorrhages, neoplasms, abscesses, and other conditions masquerading as stroke. Brain CT scans obtained in the first several hours after an infarction generally show no abnormality, and the infarct may not be seen reliably for 24–48 h. CT may fail to show small ischemic strokes in the posterior fossa because of bone artifact; small infarcts on the cortical surface may also be missed.

Contrast-enhanced CT scans add specificity by showing contrast enhancement of subacute infarcts and allow visualization of venous structures. Coupled with multidetector scanners, CT angiography can be performed with administration of IV iodinated contrast allowing visualization of the cervical and intracranial arteries, intracranial veins, aortic arch, and even the coronary arteries in one imaging session. Carotid disease and intracranial vascular occlusions are readily identified with this method (**see Fig. 420-2**). After an IV bolus of contrast, deficits in brain perfusion produced by vascular occlusion can also be demonstrated **(Fig. 419-12)** and used to predict the region of infarcted brain and the brain at risk of further infarction (i.e., the ischemic penumbra, see "Pathophysiology of Ischemic Stroke" in Chap. 420). CT imaging is also sensitive for detecting SAH (although by itself does not rule it out), and CTA can readily identify intracranial aneurysms **(Chap. 301)**. Because of its speed and wide availability, noncontrast head CT is the imaging modality of choice in patients with acute stroke (Fig. 419-1), and CTA and CT perfusion imaging may also be useful and convenient adjuncts.

FIGURE 419-12

Acute left middle cerebral artery (MCA) stroke with right hemiplegia but preserved language. ***A.*** Computed tomography (CT) perfusion mean-transit time map showing delayed perfusion of the left MCA distribution (*blue*). ***B.*** Predicted region of infarct (*red*) and penumbra (*green*) based on CT perfusion data. ***C.*** Conventional angiogram showing occlusion of the left internal carotid–MCA bifurcation (*left panel*), and revascularization of the vessels following successful thrombectomy 8 h after stroke symptom onset (*right panel*). ***D.*** The clot removed with a thrombectomy device (L5, Concentric Medical, Inc.). ***E.*** CT scan of the

brain 2 days later; note infarction in the region predicted in ***B*** but preservation of the penumbral region by successful revascularization.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

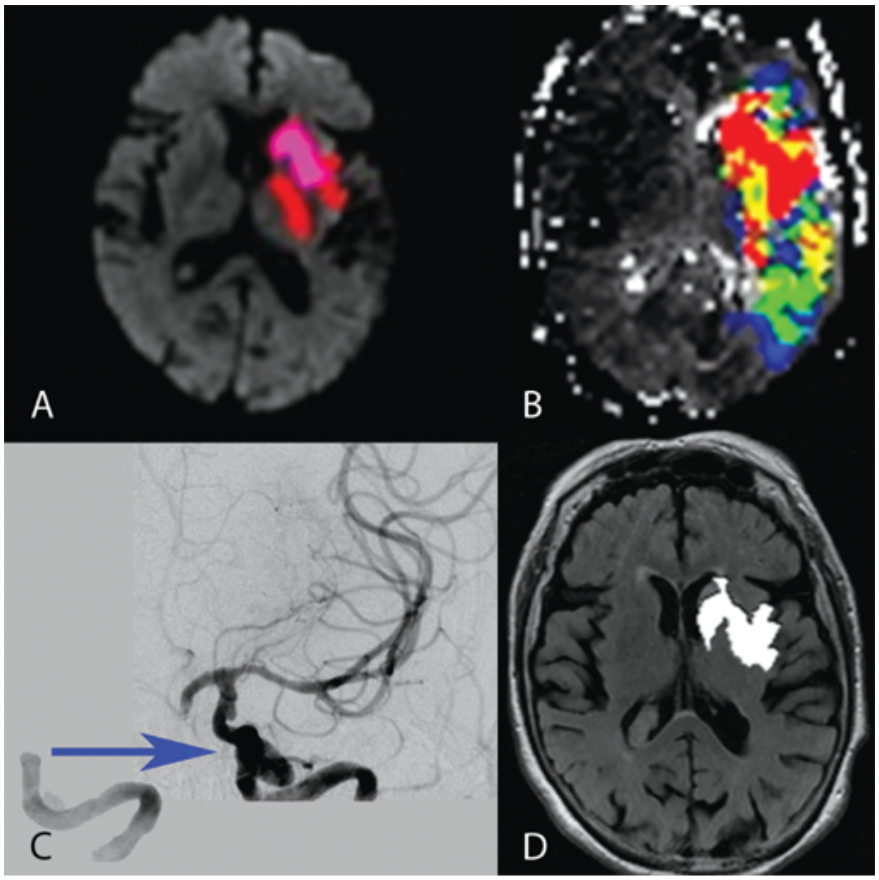
MRI

MRI reliably documents the extent and location of infarction in all areas of the brain, including the posterior fossa and cortical surface. It also identifies intracranial hemorrhage and other abnormalities and, using special sequences, can be as sensitive as CT for detecting acute intracerebral hemorrhage. MRI scanners with magnets of higher field strength produce more reliable and precise images. Diffusion-weighted imaging is more sensitive for early brain infarction than standard MR sequences or CT **(Fig. 419-13)**, as is fluid-attenuated inversion recovery (FLAIR) imaging **(Chap. 416)**. Using IV administration of gadolinium contrast, MR perfusion studies can be performed. Brain regions showing poor perfusion but no abnormality on diffusion provide, compared to CT, an equivalent measure of the ischemic penumbra. MR angiography is highly sensitive for stenosis of extracranial internal carotid arteries and of large intracranial vessels. With higher degrees of stenosis, MR angiography tends to overestimate the degree of stenosis when compared to conventional x-ray angiography. MRI with fat saturation is an imaging

sequence used to visualize extra or intracranial arterial dissection. This sensitive technique images clotted blood within the dissected vessel wall. Iron-sensitive imaging (ISI) is helpful to detect cerebral microbleeds that may be present in cerebral amyloid angiopathy and other hemorrhagic disorders.

FIGURE 419-13

Magnetic resonance imaging (MRI) of acute stroke. ***A.*** MRI diffusion-weighted image (DWI) of an 82-year-old woman 2.5 h after onset of right-sided weakness and aphasia reveals restricted diffusion within the left basal ganglia and internal capsule (*colored regions*). ***B.*** Perfusion defect within the left hemisphere (*colored signal*) imaged after administration of an IV bolus of gadolinium contrast. The discrepancy between the region of poor perfusion shown in ***B*** and the diffusion deficit shown in ***A*** is called *diffusion-perfusion mismatch* and provides an estimate of the ischemic penumbra. Without specific therapy, the region of infarction will expand into much or all the perfusion deficit. ***C.*** Cerebral angiogram of the left internal carotid artery in this patient before (*left*) and after (*right*) successful endovascular embolectomy. The occlusion is within the carotid terminus. ***D.*** Fluid-attenuated inversion recovery image obtained 3 days later showing a region of infarction (*coded as white*) that corresponds to the initial DWI image in ***A***, but not the entire area at risk shown in ***B***, suggesting that successful embolectomy saved a large region of brain tissue from infarction. *(Courtesy of Gregory Albers, MD, Stanford University; with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MRI is more expensive and time consuming than CT and less readily available. Claustrophobia and the logistics of imaging acutely critically ill patients also limit its application. Most acute stroke protocols use CT because of these limitations. However, MRI is useful outside the acute period by more clearly defining the extent of tissue injury and discriminating new from old regions of brain infarction. MRI may have utility in patients with TIA, because it is also more likely to identify new infarction, which is a strong predictor of subsequent stroke.

Cerebral Angiography

Conventional x-ray cerebral angiography is the gold standard for identifying and quantifying atherosclerotic stenoses of the cerebral arteries and for identifying and characterizing other pathologies, including aneurysms, vasospasm, intraluminal thrombi, fibromuscular dysplasia, arteriovenous fistulae, vasculitis, and collateral channels of blood flow. Conventional angiography carries risks of arterial damage, groin hemorrhage, embolic stroke, and renal failure from contrast nephropathy, so it should be reserved for situations where less invasive means are inadequate. Acute stroke treatment with endovascular thrombectomy has proven effective in ischemic strokes caused by internal carotid terminus or MCA occlusions and has now part of routine clinical practice at centers that have this capability (see Chap. 420).

Ultrasound Techniques

Stenosis at the origin of the internal carotid artery can be identified and quantified reliably by ultrasonography that combines a B-mode ultrasound image with a Doppler ultrasound assessment of flow velocity ("duplex" ultrasound). Transcranial Doppler (TCD) assessment of MCA, ACA, and PCA flow and of vertebrobasilar flow is also useful. This latter technique can detect stenotic lesions in the large intracranial arteries because such lesions increase systolic flow velocity. TCD can also detect microemboli from otherwise asymptomatic carotid plaques. In many cases, MR angiography combined with carotid and transcranial ultrasound studies eliminates the need for conventional x-ray angiography in evaluating vascular stenosis. Alternatively, CTA of the entire head and neck can be performed during the initial imaging of acute stroke. Because this images the entire arterial system relevant to stroke, with the exception of the heart, much of the clinician's stroke workup can be completed with this single imaging study.

Perfusion Techniques

Both xenon techniques (principally xenon-CT) and positron emission tomography (PET) can quantify cerebral blood flow. These tools are generally used for research **(Chap. 416)** but can be useful for determining the significance of arterial stenosis and planning for revascularization surgery. Single-photon emission computed tomography (SPECT) and MR perfusion techniques report relative cerebral blood flow. As noted above, CT imaging is used as the initial imaging modality for acute stroke, and some centers combine both CTA and CT perfusion imaging together with the noncontrast CT scan. CT perfusion imaging increases the sensitivity for detecting ischemia and can measure the ischemic penumbra (Fig. 419-12). Alternatively, MR perfusion can be combined with MR diffusion imaging to identify the ischemic penumbra as the mismatch between these two imaging sequences (Fig. 419-13).

FURTHER READING

Caplan LR: *Caplan's Stroke: A Clinical Approach*, 5th ed. Cambridge, UK, Cambridge University Press, 2016.

Tamutzer AA et al: ED misdiagnosis of cerebrovascular events in the era of modern neuroimaging: A meta-analysis. Neurology 88:1468, 2017. [PubMed: 28356464]

Chapter 420: Ischemic Stroke

Wade S. Smith; S. Claiborne Johnston; J. Claude Hemphill, III

INTRODUCTION

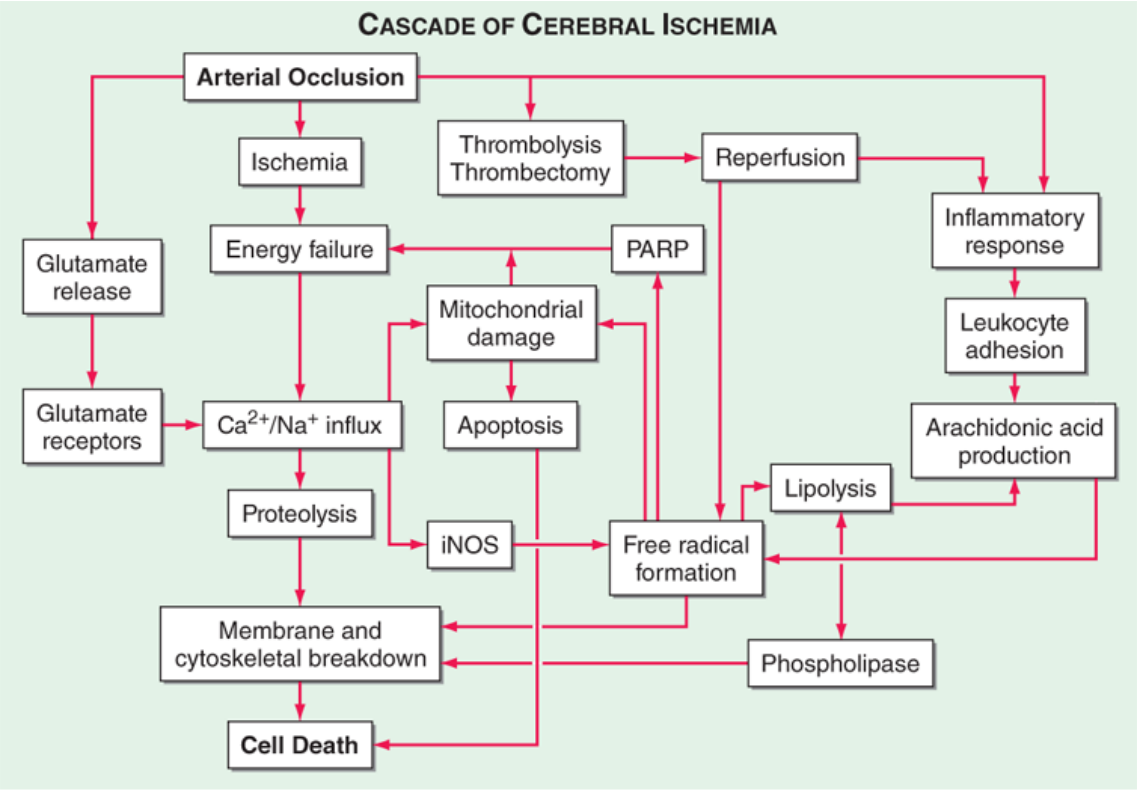
The clinical diagnosis of stroke is discussed in **Chap. 419**. Once this diagnosis is made, and either a non-contrast CT scan or MRI has been performed, rapid reversal of ischemia is paramount. This chapter will focus on the stroke treatment timeline and subsequent secondary stroke prevention.

PATHOPHYSIOLOGY OF ISCHEMIC STROKE

Acute occlusion of an intracranial vessel causes reduction in blood flow to the brain region it supplies. The magnitude of flow reduction is a function of collateral blood flow, and this depends on individual vascular anatomy (which may be altered by disease), the site of occlusion, and systemic blood pressure. A decrease in cerebral blood flow to zero causes death of brain tissue within 4–10 min; values <16–18 mL/100 g tissue per minute cause infarction within an hour; and values <20 mL/100 g tissue per minute cause ischemia without infarction unless prolonged for several hours or days. If blood flow is restored to ischemic tissue before significant infarction develops, the patient may experience only transient symptoms, and the clinical syndrome is called a transient ischemic attack (TIA). Another important concept is the *ischemic penumbra*, defined as the ischemic but reversibly dysfunctional tissue surrounding a core area of infarction. The penumbra can be imaged by perfusion imaging using MRI or CT (see below and **Figs. 419-12** and **419-13**). The ischemic penumbra will eventually progress to infarction if no change in flow occurs, and hence saving the ischemic penumbra is the goal of revascularization therapies.

Focal cerebral infarction occurs via two distinct pathways **(Fig. 420-1)**: (1) a necrotic pathway in which cellular cytoskeletal breakdown is rapid, due principally to energy failure of the cell; and (2) an apoptotic pathway in which cells become programmed to die. Ischemia produces necrosis by starving neurons of glucose and oxygen, which in turn results in failure of mitochondria to produce ATP. Without ATP, membrane ion pumps stop functioning and neurons depolarize, allowing intracellular calcium to rise. Cellular depolarization also causes glutamate release from synaptic terminals; excess extracellular glutamate produces neurotoxicity by activating postsynaptic glutamate receptors that increase neuronal calcium influx. Free radicals are produced by degradation of membrane lipids and mitochondrial dysfunction. Free radicals cause catalytic destruction of membranes and likely damage other vital functions of cells. Lesser degrees of ischemia, as are seen within the ischemic penumbra, favor apoptotic cellular death causing cells to die days to weeks later. Fever dramatically worsens brain injury during ischemia, as does hyperglycemia (glucose >11.1 mmol/L [200 mg/dL]), so it is reasonable to suppress fever and prevent hyperglycemia as much as possible. The value of induced mild hypothermia to improve stroke outcomes is the subject of continuing clinical research.

FIGURE 420-1
Major steps in the cascade of cerebral ischemia. See text for details. iNOS, inducible nitric oxide synthase; PARP, poly-A ribose polymerase.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Acute Ischemic Stroke

After the clinical diagnosis of stroke is made, an orderly process of evaluation and treatment should follow. The first goal is to prevent or reverse brain injury. Attend to the patient's airway, breathing, and circulation (ABCs), and treat hypoglycemia or hyperglycemia if identified by finger stick testing. Perform an emergency noncontrast head CT scan to differentiate between ischemic stroke and hemorrhagic stroke; there are no reliable clinical findings that conclusively separate ischemia from hemorrhage, although a more depressed level of consciousness, higher initial blood pressure, or worsening of symptoms after onset favor hemorrhage, and a deficit that is maximal at onset, or remits, suggests ischemia. Treatments designed to reverse or lessen the amount of tissue infarction and improve clinical outcome fall within six categories: (1) medical support, (2) IV thrombolysis, (3) endovascular revascularization, (4) antithrombotic treatment, (5) neuroprotection, and (6) stroke centers and rehabilitation.

MEDICAL SUPPORT

When ischemic stroke occurs, the immediate goal is to optimize cerebral perfusion in the surrounding ischemic penumbra. Attention is also directed toward preventing the common complications of bedridden patients—infections (pneumonia, urinary, and skin) and deep-venous thrombosis (DVT) with pulmonary embolism. Subcutaneous heparin (unfractionated and low-molecular-weight) is safe and can be used concomitantly. Use of pneumatic compression stockings is of proven benefit in reducing risk of DVT and is a safe alternative to heparin.

Because collateral blood flow within the ischemic brain may be blood pressure dependent, there is controversy about whether blood pressure should be lowered acutely. Blood pressure should be reduced if it exceeds 220/120 mmHg, if there is malignant hypertension (Chap. 271), concomitant myocardial ischemia, or if blood pressure is >185/110 mmHg and thrombolytic therapy is anticipated. When faced with the competing demands of myocardium and brain, lowering the heart rate with a β_1-adrenergic blocker (such as esmolol) can be a first step to decrease cardiac work and maintain blood pressure. Routine lowering of blood pressure below the limits listed above has the potential to worsen outcomes. Fever is detrimental and should be treated with antipyretics and surface cooling. Serum glucose should be monitored and kept <10.0 mmol/L (180 mg/dL) using an insulin infusion if necessary, and above at least 3.3 mmol/L (60 mg/dL).

Between 5 and 10% of patients develop enough cerebral edema to cause obtundation or brain herniation. Edema peaks on the second or third day but can cause mass effect for ~10 days. The larger the infarct, the greater the likelihood that clinically significant edema will develop. Water restriction and IV mannitol may be used to raise the serum osmolarity, but hypovolemia should be avoided because this may contribute to hypotension and worsening infarction. Combined analysis of three randomized European trials of hemicraniectomy (craniotomy and temporary removal of part of the skull) shows that hemicraniectomy reduces mortality by 50%, and the clinical outcomes of survivors are significantly improved. Older patients (age >60 years) benefit less, but still significantly. The size of the diffusion-weighted imaging volume of brain infarction during the acute stroke is a predictor of deterioration requiring hemicraniectomy.

Special vigilance is warranted for patients with cerebellar infarction. These strokes may mimic labyrinthitis because of prominent vertigo and vomiting; the presence of head or neck pain should alert the physician to consider cerebellar stroke due to vertebral artery dissection. Even small amounts of cerebellar edema can acutely increase intracranial pressure (ICP) by obstructing cerebrospinal fluid (CSF) flow leading to hydrocephalus or by directly compressing the brainstem. The resulting brainstem compression can manifest as coma and respiratory arrest and require emergency surgical decompression. Suboccipital decompression is recommended in patients with cerebellar infarcts who demonstrate neurological deterioration and should be performed before significant brainstem compression occurs.

INTRAVENOUS THROMBOLYSIS

The National Institute of Neurological Disorders and Stroke (NINDS) rtPA Stroke Study showed a clear benefit for IV rtPA in selected patients with acute stroke. The NINDS study used IV rtPA (0.9 mg/kg to a 90-mg maximum; 10% as a bolus, then the remainder over 60 min) versus placebo in ischemic stroke within 3 h of onset. One-half of the patients were treated within 90 min. Symptomatic intracranial hemorrhage occurred in 6.4% of patients on rtPA and 0.6% on placebo. In the rTPA group, there was a significant 12% absolute increase in the number of patients with only minimal disability (32% on placebo and 44% on rtPA) and a nonsignificant 4% reduction in mortality (21% on placebo and 17% on rtPA). Thus, despite an increased incidence of symptomatic intracranial hemorrhage, treatment with IV rtPA within 3 h of the onset of ischemic stroke improved clinical outcome.

Three subsequent trials of IV rtPA did not confirm this benefit, perhaps because of the dose of rtPA used, the timing of its delivery, and small sample size. When data from all randomized IV rtPA trails were combined, however, efficacy was confirmed in the <3-h time window, and efficacy likely extended to 4.5 h and possibly to 6 h. Based on these combined results, the European Cooperative Acute Stroke Study (ECASS) III explored the safety and efficacy of rtPA in the 3- to 4.5-h time window. Unlike the NINDS study, patients aged >80 years and diabetic patients with a previous stroke were excluded. In this 821-patient randomized study, efficacy was again confirmed, although the treatment effect was less robust than in the 0- to 3-h time window. In the rtPA group, 52.4% of patients achieved a good outcome at 90 days, compared to 45.2% of the placebo group (odds ratio [OR] 1.34, $p = .04$). The symptomatic intracranial hemorrhage rate was 2.4% in the rtPA group and 0.2% in the placebo group ($p = .008$).

Based on these data, rtPA is approved in the 3- to 4.5-h window in Europe and Canada, but is still only approved for 0–3 h in the United States. A dose of 0.6 mg/kg is typically used in Japan and other Asian countries based on observation of >600 patients given this lower dose, and observing similar outcomes to historical controls and a lower rate of intracranial hemorrhage. This dose also mitigates concerns that patients of Asian descent have a higher propensity to bleed from most antithrombotic and thrombolytic medications.. Use of IV tPA is now considered a central component of primary stroke centers (see below). It represents the first treatment proven to improve clinical outcomes in ischemic stroke and is cost-effective and cost-saving. Advanced neuroimaging techniques **(see Chap. 419)** may help to select patients beyond the 4.5-h window who will benefit from thrombolysis. The time of stroke onset is defined as the time the patient's symptoms were witnessed to begin or the time the patient was last seen as normal. Patients who awaken with stroke have the onset defined as when they went to bed. **Table 420-1** summarizes eligibility criteria and instructions for administration of IV rtPA.

ENDOVASCULAR REVASCULARIZATION

Ischemic stroke from large-vessel intracranial occlusion results in high rates of mortality and morbidity. Occlusions in such large vessels (middle cerebral artery [MCA], intracranial internal carotid artery, and the basilar artery) generally involve a large clot volume and often fail to open with IV rtPA alone. As proof of concept, thrombolytics were tested via an intraarterial route to increase the concentration of drug at the clot and minimize systemic bleeding complications. The Prolyse in Acute Cerebral Thromboembolism (PROACT) II trial found benefit for intraarterial prourokinase in acute MCA occlusions up to the sixth hour following onset of stroke. Intraarterial treatment of basilar artery occlusions may also be beneficial for selected patients but has not been tested in a randomized trial. Intraarterial administration of a thrombolytic agent for acute ischemic stroke (AIS) is not approved by the U.S. Food and Drug Administration (FDA); however, based on these data many stroke centers consider this treatment if more advanced techniques of mechanical thrombectomy fail.

Endovascular mechanical thrombectomy has been studied as an alternative or adjunctive treatment of acute stroke in patients who are ineligible for, or have contraindications to, thrombolytics or in those who failed to achieve vascular recanalization with IV thrombolytics (**see Fig. 419-12**). First generation thrombectomy devices produced promising results with recanalization in observational studies, leading to FDA approval. Three

randomized stroke trials published in 2013 concluded that endovascular therapy did not improve outcomes, but results may have been influenced by methodologic issues: angiography was not required for study entry and less effective mechanical devices were employed. In 2015, the results of six randomized trials were published, all demonstrating that endovascular therapy improved clinical outcomes for internal carotid and MCA occlusions proven by CTA, under 6 h from stroke onset, with or without pretreatment with IV t-PA. One study concluded that patients were home nearly 2 months earlier if they received endovascular therapy. A combined meta-analysis of all 1287 patients in these trials confirmed a large benefit with endovascular therapy (OR 2.49, 95% CI 1.76–3.53; p<0.001). The percentage of patients who achieved modified Rankin scores of 0–2 (normal or symptomatic but independent) was 46% in the endovascular group and 26.5% in the medical arm. Mortality was unchanged. As with IV t-PA treatment, clinical outcome is dependent on time to effective therapy. The odds of a good outcome exceed 3 if groin puncture occurs within 2 h of symptom onset, but is only 2 if 8 h elapse. Over 80% of patients who had vessel opening within 1 h of arrival to the Emergency Department had a good outcome, while only one-third had a good outcome if 6 h elapsed.

Extending the time window beyond 6 h appears to be effective if the patient has specific imaging findings demonstrating good vascular collaterals (CT perfusion or MR perfusion techniques, see Chap. 419) and can be treated within 24 h; the Clinical Mismatch in the Triage of Wake Up and Late Presenting Strokes Undergoing Neurointervention With Trevo (DAWN) trial reported good outcomes more frequently with endovascular therapy than with medical care alone (47 vs 13%, p<0.0001). The Endovascular Therapy Following Imaging Evaluation for Ischemic Stroke 3 (DEFUSE-3) trial confirmed these results (45 vs 17%, p<0.001) if treated up to 16 hours from stroke onset.

Now that endovascular stroke therapy is proven to be effective, the creation of comprehensive stroke centers designed to rapidly identify and treat patients with large vessel cerebral ischemia are a major focus internationally. Creating geographical systems of care whereby stroke patients are first evaluated at primary stroke centers (which can administer IV t-PA) then transferred to comprehensive centers if needed, or directly triaged to comprehensive centers based on field assessment, appears to be an effective strategy to improve patient outcomes.

ANTITHROMBOTIC TREATMENT

Platelet Inhibition

Aspirin is the only antiplatelet agent that has been proven to be effective for the acute treatment of ischemic stroke; there are several antiplatelet agents proven for the secondary prevention of stroke (see below). Two large trials, the International Stroke Trial (IST) and the Chinese Acute Stroke Trial (CAST), found that the use of aspirin within 48 h of stroke onset reduced both stroke recurrence risk and mortality minimally. Among 19,435 patients in IST, those allocated to aspirin, 300 mg/d, had slightly fewer deaths within 14 days (9.0 vs 9.4%), significantly fewer recurrent ischemic strokes (2.8 vs 3.9%), no excess of hemorrhagic strokes (0.9 vs 0.8%), and a trend toward a reduction in death or dependence at 6 months (61.2 vs 63.5%). In CAST, 21,106 patients with ischemic stroke received 160 mg/d of aspirin or a placebo for up to 4 weeks. There were very small reductions in the aspirin group in early mortality (3.3 vs 3.9%), recurrent ischemic strokes (1.6 vs 2.1%), and dependency at discharge or death (30.5 vs 31.6%). These trials demonstrate that the use of aspirin in the treatment of AIS is safe and produces a small net benefit. For every 1000 acute strokes treated with aspirin, about 9 deaths or nonfatal stroke recurrences will be prevented in the first few weeks and ~13 fewer patients will be dead or dependent at 6 months.

Anticoagulation

Numerous clinical trials have failed to demonstrate any benefit of routine anticoagulation in the primary treatment of atherothrombotic cerebral ischemia, and have also shown an increase in the risk of brain and systemic hemorrhage. Therefore the routine use of heparin or other anticoagulants for patients with atherothrombotic stroke is not warranted. Heparin and oral anticoagulation are likely no more effective than aspirin for stroke associated with arterial dissection. However, there may be benefit of anticoagulation for halting progression of dural sinus thrombosis.

NEUROPROTECTION

Neuroprotection is the concept of providing a treatment that prolongs the brain's tolerance to ischemia. Drugs that block the excitatory amino acid pathways have been shown to protect neurons and glia in animals, but despite multiple human trials, they have not yet been proven to be beneficial. Hypothermia is a powerful neuroprotective treatment in patients with cardiac arrest (Chap. 301) and is neuroprotective in animal models of stroke, but it has not been adequately studied in patients with ischemic stroke and is associated with an increase in pneumonia rates that could adversely impact stroke outcomes.

STROKE CENTERS AND REHABILITATION

Patient care in stroke units followed by rehabilitation services improves neurologic outcomes and reduces mortality. Use of clinical pathways and staff dedicated to the stroke patient can improve care. This includes use of standardized stroke order sets. Stroke teams that provide emergency 24-h evaluation of acute stroke patients for acute medical management and consideration of thrombolysis or endovascular treatments are essential components of primary and comprehensive stroke centers, respectively.

Proper rehabilitation of the stroke patient includes early physical, occupational, and speech therapy. It is directed toward educating the patient and family about the patient's neurologic deficit, preventing the complications of immobility (e.g., pneumonia, DVT and pulmonary embolism, pressure sores of the skin, and muscle contractures), and providing encouragement and instruction in overcoming the deficit. Use of pneumatic compression stockings is of proven benefit in reducing risk of DVT and is a safe alternative to heparin. The goal of rehabilitation is to return the patient home and to maximize recovery by providing a safe, progressive regimen suited to the individual patient. Additionally, the use of constrained movement therapy (immobilizing the unaffected side) has been shown to improve hemiparesis following stroke, even years after the stroke, suggesting that physical therapy can recruit unused neural pathways. Newer robotic therapies appear promising as well. The human nervous system is more adaptable than previously thought, and developing physical and pharmacologic strategies to enhance long-term neural recovery is an active area of research.

TABLE 420-1

Administration of Intravenous Recombinant Tissue Plasminogen Activator (rtPA) for Acute Ischemic Stroke (AIS)[a]

Indication	Contraindication
Clinical diagnosis of stroke Onset of symptoms to time of drug administration ≤4.5 h[b] CT scan showing no hemorrhage or edema of >1/3 of the MCA territory Age 18 ≥ years	Sustained BP >185/110 mmHg despite treatment Bleeding diathesis Recent head injury or intracerebral hemorrhage Major surgery in preceding 14 days Gastrointestinal bleeding in preceding 21 days Recent myocardial infarction
Administration of rtPA	
IV access with two peripheral IV lines (avoid arterial or central line placement) Review eligibility for rtPA Administer 0.9 mg/kg IV (maximum 90 mg) IV as 10% of total dose by bolus, followed by remainder of total dose over 1 h[c] Frequent cuff blood pressure monitoring No other antithrombotic treatment for 24 h For decline in neurologic status or uncontrolled blood pressure, stop infusion, give cryoprecipitate, and reimage brain emergently Avoid urethral catheterization for ≥2 h	

[a]See Activase (tissue plasminogen activator) package insert for complete list of contraindications and dosing. [b]Depending on the country, IV rtPA may be approved for up to 4.5 h with additional restrictions. [c]A dose of 0.6 mg/kg is commonly used in Asia (Japan and China) based on randomized data indicating less hemorrhage and similar efficacy using this lower-dose.

Abbreviations: BP, blood pressure; CT, computed tomography; HCT, hematocrit; INR, international normalized ratio; MCA, middle cerebral artery; PTT, partial thromboplastin time.

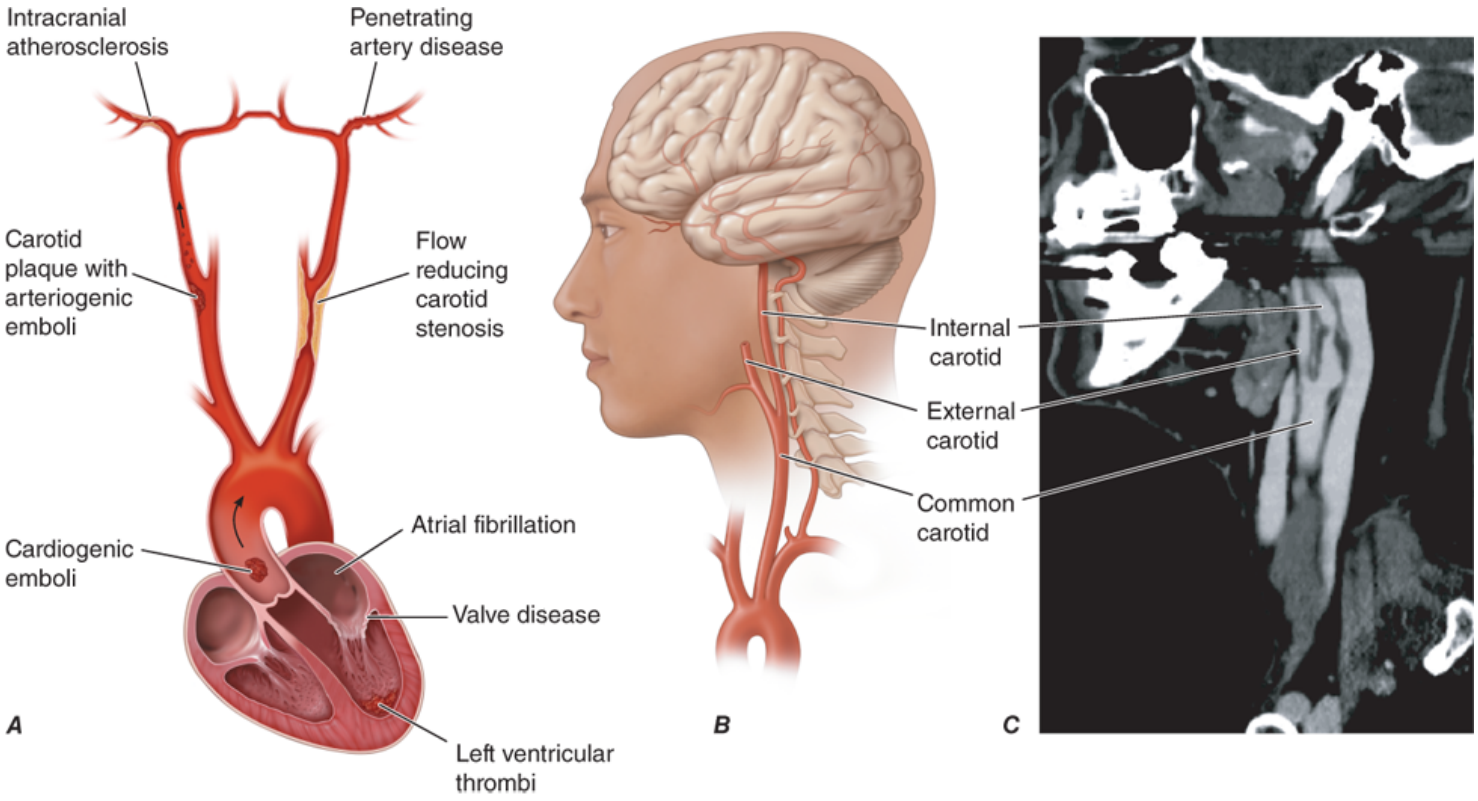
ETIOLOGY OF ISCHEMIC STROKE

(Fig. 420-2 and Table 420-2) Although the initial management of AIS often does not depend on the etiology, establishing a cause is essential to reduce the risk of recurrence. Focus should be on atrial fibrillation and carotid atherosclerosis, because these etiologies have proven secondary prevention strategies. The clinical presentation and examination findings often establish the cause of stroke or narrow the possibilities to a few. Judicious use of laboratory testing and imaging studies completes the initial evaluation. Nevertheless, nearly 30% of strokes remain unexplained despite extensive evaluation.

FIGURE 420-2

Pathophysiology of ischemic stroke. ***A.*** Diagram illustrating the three major mechanisms that underlie ischemic stroke: (1) occlusion of an intracranial vessel by an embolus that arises at a distant site (e.g., cardiogenic sources such as atrial fibrillation or artery-to-artery emboli from carotid atherosclerotic plaque), often affecting the large intracranial vessels; (2) in situ thrombosis of an intracranial vessel, typically affecting the small penetrating arteries that arise from the major intracranial arteries; (3) hypoperfusion caused by flow-limiting stenosis of a major extracranial (e.g., internal carotid) or intracranial vessel, often producing "watershed" ischemia. ***B.*** and ***C.*** Diagram and reformatted computed

tomography angiogram of the common, internal, and external carotid arteries. High-grade stenosis of the internal carotid artery, which may be associated with either cerebral emboli or flow-limiting ischemia, was identified in this patient.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 420-2

Causes of Ischemic Stroke

Common Causes	Uncommon Causes
Thrombosis Lacunar stroke (small vessel) Large-vessel thrombosis Dehydration Embolic occlusion Artery-to-artery Carotid bifurcation Aortic arch Arterial dissection Cardioembolic Atrial fibrillation Mural thrombus Myocardial infarction Dilated cardiomyopathy Valvular lesions Mitral stenosis Mechanical valve Bacterial endocarditis Paradoxical embolus Atrial septal defect Patent foramen ovale Atrial septal aneurysm Spontaneous echo contrast Stimulant drugs: cocaine, amphetamine	Hypercoagulable disorders Protein C deficiency[a] Protein S deficiency[a] Antithrombin III deficiency[a] Antiphospholipid syndrome Factor V Leiden mutation[a] Prothrombin G20210 mutation[a] Systemic malignancy Sickle cell anemia β Thalassemia Polycythemia vera Systemic lupus erythematosus Homocysteinemia Thrombotic thrombocytopenic purpura Disseminated intravascular coagulation Dysproteinemias[a] Nephrotic syndrome[a] Inflammatory bowel disease[a] Oral contraceptives Venous sinus thrombosis[b] Fibromuscular dysplasia Vasculitis Systemic vasculitis (PAN, granulomatosis with polyangiitis [Wegener's], Takayasu's, giant cell arteritis) Primary CNS vasculitis Meningitis (syphilis, tuberculosis, fungal, bacterial, zoster) Noninflammatory vasculopathy Reversible vasoconstriction syndrome Fabry's disease Angiocentric lymphoma Cardiogenic Mitral valve calcification Atrial myxoma Intracardiac tumor Marantic endocarditis Libman-Sacks endocarditis Subarachnoid hemorrhage vasospasm Moyamoya disease Eclampsia

[a]Chiefly cause venous sinus thrombosis. [b]May be associated with any hypercoagulable disorder.

Abbreviations: CNS, central nervous system; PAN, polyarteritis nodosa.

Clinical examination should focus on the peripheral and cervical vascular system (carotid auscultation for bruits and blood pressure), the heart (dysrhythmia, murmurs), extremities (peripheral emboli), and retina (effects of hypertension and cholesterol emboli [Hollenhorst plaques]). A complete neurologic examination is performed to localize the anatomic site of stroke. An imaging study of the brain is nearly always indicated and is required for patients being considered for thrombolysis; it may be combined with CT- or MRI-based angiography to visualize the

vasculature of the neck and intracranial vessels **(see "Imaging Studies," Chap. 419)**. A chest x-ray, electrocardiogram (ECG), urinalysis, complete blood count, erythrocyte sedimentation rate (ESR), serum electrolytes, blood urea nitrogen (BUN), creatinine, blood glucose, serum lipid profile, prothrombin time (PT), and partial thromboplastin time (PTT) are often useful and should be considered in all patients. An ECG, and subsequent cardiac telemetry, may demonstrate arrhythmias or reveal evidence of recent myocardial infarction (MI). Of all these studies, only brain imaging and capillary blood glucose are necessary prior to IV rtPA; the results of other studies should not delay the rapid administration of IV rtPA if the patient is eligible.

Cardioembolic Stroke

Cardioembolism is responsible for ~20% of all ischemic strokes. Stroke caused by heart disease is primarily due to embolism of thrombotic material forming on the atrial or ventricular wall or the left heart valves. These thrombi then detach and embolize into the arterial circulation. The thrombus may fragment or lyse quickly, producing only a TIA. Alternatively, the arterial occlusion may last longer, producing stroke. Embolic strokes tend to occur suddenly with maximum neurologic deficit present at onset. With reperfusion following more prolonged ischemia, petechial hemorrhages can occur within the ischemic territory. These are usually of no clinical significance and should be distinguished from frank intracranial hemorrhage into a region of ischemic stroke where the mass effect from the hemorrhage can cause a significant decline in neurologic function.

Emboli from the heart most often lodge in the intracranial internal carotid artery, the MCA, the posterior cerebral artery (PCA), or one of their branches; infrequently, the anterior cerebral artery (ACA) is involved. Emboli large enough to occlude the stem of the MCA (3–4 mm) or internal carotid terminus lead to large infarcts that involve both deep gray and white matter and some portions of the cortical surface and its underlying white matter. A smaller embolus may occlude a small cortical or penetrating arterial branch. The location and size of an infarct within a vascular territory depend on the extent of the collateral circulation.

The most significant cause of cardioembolic stroke in most of the world is nonrheumatic (often called nonvalvular) atrial fibrillation. MI, prosthetic valves, rheumatic heart disease, and ischemic cardiomyopathy are other considerations (Table 420-2). Cardiac disorders causing brain embolism are discussed in the chapters on heart diseases, but a few pertinent aspects are highlighted here.

Nonrheumatic atrial fibrillation is the most common cause of cerebral embolism overall. The presumed stroke mechanism is thrombus formation in the fibrillating atrium or atrial appendage, with subsequent embolization. Patients with atrial fibrillation have an average annual risk of stroke of ~5%. The risk of stroke can be estimated by calculating the CHA_2DS_2-VASc score **(Table 420-3)**. Left atrial enlargement is an additional risk factor for formation of atrial thrombi. Rheumatic heart disease usually causes ischemic stroke when there is prominent mitral stenosis or atrial fibrillation. Recent MI may be a source of emboli, especially when transmural and involving the anteroapical ventricular wall, and prophylactic anticoagulation following MI has been shown to reduce stroke risk. Mitral valve prolapse is not usually a source of emboli unless the prolapse is severe.

TABLE 420-3

Recommendations on Chronic Use of Antithrombotics for Various Cardiac Conditions

Condition	Recommendation
Nonvalvular atrial fibrillation	Calculate CHA_2DS_2-VASc score[a]
• CHA_2DS_2-VASc score of 0	Aspirin or no antithrombotic
• CHA_2DS_2-VASc score of 1	Aspirin or OAC
• CHA_2DS_2-VASc score of 2 or greater	OAC
Rheumatic mitral valve disease	
• With atrial fibrillation, previous embolization, or atrial appendage thrombus, or left atrial diameter >55 mm	OAC
• Embolization or appendage clot despite OAC	OAC plus aspirin
Mitral valve prolapse	
• Asymptomatic	No therapy
• With otherwise cryptogenic stroke or TIA	Aspirin
• Atrial fibrillation	OAC
Mitral annular calcification	
• Without atrial fibrillation but systemic embolization, or otherwise cryptogenic stroke or TIA	Aspirin
• Recurrent embolization despite aspirin	OAC
• With atrial fibrillation	OAC
Aortic valve calcification	
• Asymptomatic	No therapy
• Otherwise cryptogenic stroke or TIA	Aspirin
Aortic arch mobile atheroma	
• Otherwise cryptogenic stroke or TIA	Aspirin or OAC

Condition	Recommendation
Patent foramen ovale	
• Otherwise cryptogenic ischemic stroke or TIA	Aspirin or closure with device
• Indication for OAC (deep-venous thrombosis or hypercoagulable state)	OAC
Mechanical heart value	
• Aortic position, bileaflet or Medtronic Hall tilting disk with normal left atrial size and sinus rhythm	VKA INR 2.5, range 2–3
• Mitral position tilting disk or bileaflet valve	VKA INR 3.0, range 2.5–3.5
• Mitral or aortic position, anterior-apical myocardial infarct or left atrial enlargement	VKA INR 3.0, range 2.5–3.5
• Mitral or aortic position, with atrial fibrillation, or hypercoagulable state, or low ejection fraction, or atherosclerotic vascular disease	Aspirin plus VKA INR 3.0, range 2.5–3.5
• Systemic embolization despite target INR	Add aspirin and/or increase INR: prior target was 2.5 increase to 3.0, range 2.5–3.5; prior target was 3.0 increase to 3.5, range 3–4
Bioprosthetic valve	
• No other indication for VKA therapy	Aspirin
Infective endocarditis	Avoid antithrombotic agents
Nonbacterial thrombotic endocarditis	
• With systemic embolization	Full-dose unfractionated heparin or SC LMWH

[a] CHA_2DS_2-VASc score is calculated as follows: 1 point for Congestive heart failure, 1 point for Hypertension, 2 points for Age ≥ 75 y, 1 point for Diabetes mellitus, 2 points for Stroke or TIA, 1 point for Vascular disease (prior MI, peripheral vascular disease or aortic plaque), 1 point for Age 65–74 y, 1 point for female Sex category; sum of point is the total [b] CHA_2DS_2-VASc score.

Note: Dose of aspirin is 50–325 mg/d; target INR for OAC is between 2 and 3 unless otherwise specified.

Abbreviations: INR, international normalized ratio; LMWH, low-molecular-weight heparin; OAC, oral anticoagulant (VKA, thrombin inhibitor, or oral factor Xa inhibitors); TIA, transient ischemic attack; VKA, vitamin K antagonist.

Sources: Modified from DE Singer et al: Chest 133:546S, 2008; DN Salem et al: Chest 133:593S, 2008; CT January et al: JACC 64:2246, 2014.

Paradoxical embolization occurs when venous thrombi migrate to the arterial circulation, usually via a patent foramen ovale (PFO) or atrial septal defect. Bubble-contrast echocardiography (IV injection of agitated saline coupled with either transthoracic or transesophageal echocardiography) can demonstrate a right-to-left cardiac shunt, revealing the conduit for paradoxical embolization. Alternatively, a right-to-left shunt is implied if immediately following IV injection of agitated saline, the ultrasound signature of bubbles is observed during transcranial Doppler insonation of the MCA; pulmonary arteriovenous malformations should be considered if this test is positive yet an echocardiogram fails to reveal an intracardiac shunt. Both techniques are highly sensitive for detection of right-to-left shunts. Besides venous clot, fat and tumor emboli, bacterial

endocarditis, IV air, and amniotic fluid emboli at childbirth may occasionally be responsible for paradoxical embolization. The importance of a PFO as a cause of stroke is debated, particularly because they are present in ~15% of the general population. Some studies have suggested that the risk is only elevated in the presence of a coexisting atrial septal aneurysm. The presence of a venous source of embolus, most commonly a deep-venous thrombus, may provide confirmation of the importance of a PFO with an accompanying right-to-left shunt in a particular case. Two recent trials found about a 1% per year absolute reduction in stroke risk using percutaneous occlusion devices in patients with no other explanation for their stroke.

Bacterial endocarditis can be a source of valvular vegetations that give rise to septic emboli. The appearance of multifocal symptoms and signs in a patient with stroke makes bacterial endocarditis more likely. Infarcts of microscopic size occur, and large septic infarcts may evolve into brain abscesses or cause hemorrhage into the infarct, which generally precludes use of anticoagulation or thrombolytics. Mycotic aneurysms caused by septic emboli may also present as subarachnoid hemorrhage (SAH) or intracerebral hemorrhage.

Artery-to-Artery Embolic Stroke

Thrombus formation on atherosclerotic plaques may embolize to intracranial arteries producing an artery-to-artery embolic stroke. Less commonly, a diseased vessel may acutely thrombose. Unlike the myocardial vessels, artery-to-artery embolism, rather than local thrombosis, appears to be the dominant vascular mechanism causing large-vessel brain ischemia. Any diseased vessel may be an embolic source, including the aortic arch, common carotid, internal carotid, vertebral, and basilar arteries.

CAROTID ATHEROSCLEROSIS

Atherosclerosis within the carotid artery occurs most frequently within the common carotid bifurcation and proximal internal carotid artery; the carotid siphon (portion within the cavernous sinus) is also vulnerable to atherosclerosis. Male gender, older age, smoking, hypertension, diabetes, and hypercholesterolemia are risk factors for carotid disease, as they are for stroke in general **(Table 420-4)**. Carotid atherosclerosis produces an estimated 10% of ischemic stroke. **For further discussion of the pathogenesis of atherosclerosis, see Chap. 232**.

TABLE 420-4

Risk Factors for Stroke

Risk Factor	Relative Risk	Relative Risk Reduction with Treatment	Number Needed to Treat[a]	
			Primary Prevention	Secondary Prevention
Hypertension	2–5	38%	100–300	50–100
Atrial fibrillation	1.8–2.9	68% warfarin, 21% aspirin	20–83	13
Diabetes	1.8–6	No proven effect		
Smoking	1.8	50% at 1 year, baseline risk at 5 years postcessation		
Hyperlipidemia	1.8–2.6	16–30%	560	230
Asymptomatic carotid stenosis	2.0	53%	85	N/A
Symptomatic carotid stenosis (70–99%)		65% at 2 years	N/A	12
Symptomatic carotid stenosis (50–69%)		29% at 5 years	N/A	77

[a]Number needed to treat to prevent one stroke annually. Prevention of other cardiovascular outcomes is not considered here.

Abbreviation: N/A, not applicable.

Carotid disease can be classified by whether the stenosis is symptomatic or asymptomatic and by the degree of stenosis (percent narrowing of the narrowest segment compared to a nondiseased segment). Symptomatic carotid disease implies that the patient has experienced a stroke or TIA within the vascular distribution of the artery, and it is associated with a greater risk of subsequent stroke than asymptomatic stenosis, in which the patient is symptom free and the stenosis is detected through screening. Greater degrees of arterial narrowing are generally associated with a higher risk of stroke, except that those with near occlusions are at lower risk of stroke.

OTHER CAUSES OF ARTERY-TO-ARTERY EMBOLIC STROKE

Intracranial atherosclerosis produces stroke either by an embolic mechanism or by in situ thrombosis of a diseased vessel. It is more common in patients of Asian and African-American descent. Recurrent stroke risk is ~15% per year, similar to untreated symptomatic carotid atherosclerosis.

Dissection of the internal carotid or vertebral arteries or even vessels beyond the circle of Willis is a common source of embolic stroke in young (age <60 years) patients. The dissection is usually painful and precedes the stroke by several hours or days. Extracranial dissections do not cause hemorrhage, presumably because of the tough adventitia of these vessels. Intracranial dissections, conversely, may produce SAH because the adventitia of intracranial vessels is thin and pseudoaneurysms may form, requiring urgent treatment to prevent rerupture. Treating asymptomatic pseudoaneurysms following dissection is likely not necessary. The cause of dissection is usually unknown, and recurrence is rare. Ehlers-Danlos type IV, Marfan's disease, cystic medial necrosis, and fibromuscular dysplasia are associated with dissections. Trauma (usually a motor vehicle accident or a sports injury) can cause carotid and vertebral artery dissections. Spinal manipulative therapy is associated with vertebral artery dissection and stroke. Most dissections heal spontaneously, and stroke or TIA is uncommon beyond 2 weeks. A recent trial showed no difference in stroke prevention with aspirin compared to anticoagulation, with a low recurrent stroke rate of 2%.

SMALL-VESSEL STROKE

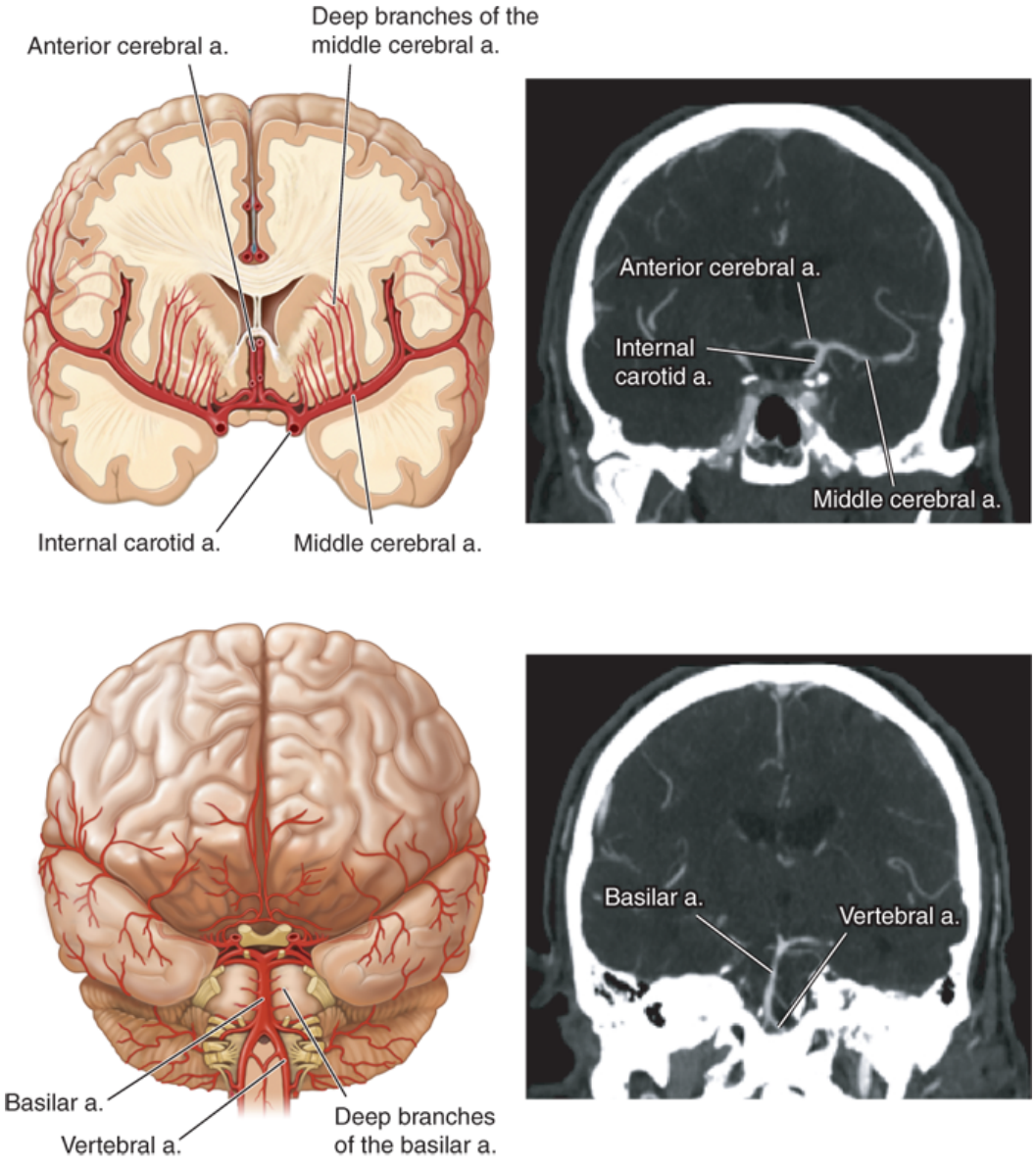
The term *lacunar infarction* refers to infarction following atherothrombotic or lipohyalinotic occlusion of a small artery in the brain. The term *small-vessel stroke* denotes occlusion of such a small penetrating artery and is now the preferred term. Small-vessel strokes account for ~20% of all strokes.

Pathophysiology

The MCA stem, the arteries comprising the circle of Willis (A1 segment, anterior and posterior communicating arteries, and P1 segment), and the basilar and vertebral arteries all give rise to 30- to 300-µm branches that penetrate the deep gray and white matter of the cerebrum or brainstem (Fig. 420-3). Each of these small branches can occlude either by atherothrombotic disease at its origin or by the development of lipohyalinotic thickening. Thrombosis of these vessels causes small infarcts that are referred to as *lacunes* (Latin for "lake" of fluid noted at autopsy). These infarcts range in size from 3 mm to 2 cm in diameter. Hypertension and age are the principal risk factors.

FIGURE 420-3

Diagrams and reformatted computed tomography (CT) angiograms in the coronal section illustrating the deep penetrating arteries involved in small-vessel strokes. In the anterior circulation, small penetrating arteries called *lenticulostriates* arise from the proximal portion of the anterior and middle cerebral arteries and supply deep subcortical structures (***upper panels***). In the posterior circulation, similar arteries arise directly from the vertebral and basilar arteries to supply the brainstem (***lower panels***). Occlusion of a single penetrating artery gives rise to a discrete area of infarct (pathologically termed a "lacune," or lake). Note that these vessels are too small to be visualized on CT angiography.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Clinical Manifestations

The most common small-vessel stroke syndromes are the following: (1) *Pure motor hemiparesis* from an infarct in the posterior limb of the internal capsule or the pons; the face, arm, and leg are almost always involved; (2) *pure sensory stroke* from an infarct in the ventral thalamus; (3) *ataxic hemiparesis* from an infarct in the ventral pons or internal capsule; (4) and *dysarthria and a clumsy hand* or arm due to infarction in the ventral pons or in the genu of the internal capsule.

Transient symptoms (small-vessel TIAs) may herald a small-vessel infarct; they may occur several times a day and last only a few minutes. Recovery from small-vessel strokes tends to be more rapid and complete than recovery from large-vessel strokes; in some cases, however, there is severe permanent disability.

A large-vessel source (either thrombosis or embolism) may manifest initially as a small-vessel infarction. Therefore, the search for embolic sources (carotid and heart) should not be completely abandoned in the evaluation of these patients. Secondary prevention of small-vessel stroke involves risk factor modification, specifically reduction in blood pressure (see "Treatment: Primary and Secondary Prevention of Stroke and TIA," below).

LESS COMMON CAUSES OF STROKE

(Table 420-2) *Hypercoagulable disorders* **(Chap. 61)** primarily increase the risk of cortical vein or cerebral venous sinus thrombosis. Systemic lupus erythematosus with Libman-Sacks endocarditis can be a cause of embolic stroke. These conditions overlap with the antiphospholipid syndrome, which probably requires long-term anticoagulation to prevent further stroke. Homocysteinemia may cause arterial thromboses as well; this disorder is caused by various mutations in the homocysteine pathways and responds to different forms of cobalamin depending on the mutation.

Venous sinus thrombosis of the lateral or sagittal sinus or of small cortical veins (cortical vein thrombosis) occurs as a complication of oral contraceptive use, pregnancy and the postpartum period, inflammatory bowel disease, intracranial infections (meningitis), and dehydration. It is also seen in patients with laboratory-confirmed thrombophilia including antiphospholipid syndrome, polycythemia, sickle cell anemia, deficiencies of proteins C and S, factor V Leiden mutation (resistance to activated protein C), antithrombin III deficiency, homocysteinemia, and the prothrombin G20210 mutation. Women who take oral contraceptives and have the prothrombin G20210 mutation may be at particularly high risk for sinus thrombosis. Patients present with headache and may also have focal neurologic signs (especially paraparesis) and seizures. Often, CT imaging is normal unless an intracranial venous hemorrhage has occurred, but the venous sinus occlusion is readily visualized using magnetic resonance (MR) or CT venography or conventional x-ray angiography. With greater degrees of sinus thrombosis, the patient may develop signs of increased ICP and coma. Intravenous heparin, regardless of the presence of intracranial hemorrhage, reduces morbidity and mortality, and the long-term outcome is generally good. Heparin prevents further thrombosis and reduces venous hypertension and ischemia. If an underlying hypercoagulable state is not found, many physicians treat with vitamin K antagonists (VKAs) for 3–6 months and then convert to aspirin, depending on the degree of resolution of the venous sinus thrombus. Anticoagulation is often continued indefinitely if thrombophilia is diagnosed.

Sickle cell anemia (SS disease) is a common cause of stroke in children. A subset of homozygous carriers of this hemoglobin mutation develop stroke in childhood, and this may be predicted by documenting high-velocity blood flow within the MCAs using transcranial Doppler ultrasonography. In children who are identified to have high velocities, treatment with aggressive exchange transfusion dramatically reduces risk of stroke, and if exchange transfusion is ceased, their stroke rate increases again along with MCA velocities.

Fibromuscular dysplasia affects the cervical arteries and occurs mainly in women. The carotid or vertebral arteries show multiple rings of segmental narrowing alternating with dilatation. Vascular occlusion is usually incomplete. The process is often asymptomatic but occasionally is associated with an audible bruit, TIAs, or stroke. Involvement of the renal arteries is common and may cause hypertension. The cause and natural history of fibromuscular dysplasia are unknown **(Chap. 275)**. TIA or stroke generally occurs only when the artery is severely narrowed or dissects. Anticoagulation or antiplatelet therapy may be helpful.

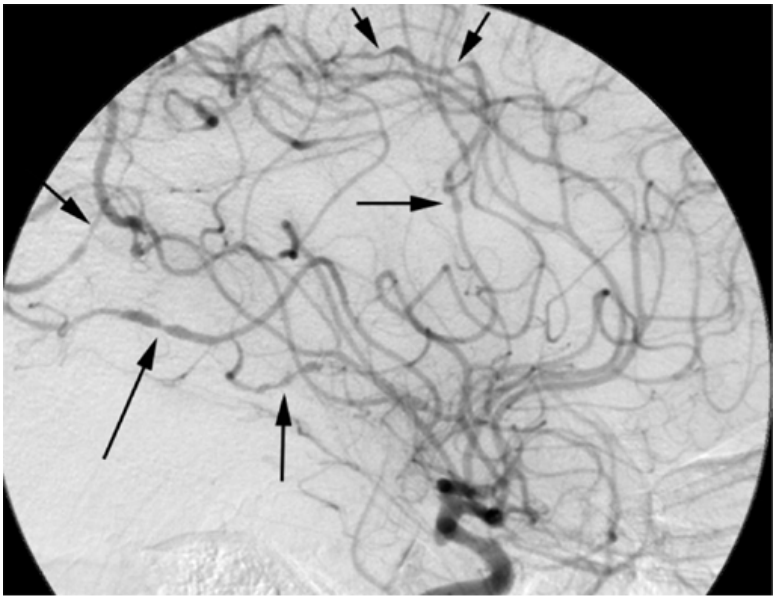
Temporal (giant cell) arteritis **(Chap. 356)** is a relatively common affliction of elderly individuals in which the external carotid system, particularly the temporal arteries, undergo subacute granulomatous inflammation with giant cells. Occlusion of posterior ciliary arteries derived from the ophthalmic artery results in blindness in one or both eyes and can be prevented with glucocorticoids. It rarely causes stroke because the internal carotid artery is usually not inflamed. Idiopathic giant cell arteritis involving the great vessels arising from the aortic arch (*Takayasu's arteritis*) may cause carotid or vertebral thrombosis; it is rare in the Western Hemisphere.

Necrotizing (or granulomatous) arteritis, occurring alone or in association with generalized polyarteritis nodosa or granulomatosis with polyangiitis (Wegener's), involves the distal small branches (<2 mm diameter) of the main intracranial arteries and produces small ischemic infarcts in the brain, optic nerve, and spinal cord. The CSF often shows pleocytosis, and the protein level is elevated. *Primary central nervous system vasculitis* is rare; small or medium-sized vessels are usually affected, without apparent systemic vasculitis. The differential diagnosis includes other inflammatory vasculopathies including infection (tuberculous, fungal), sarcoidosis, angiocentric lymphoma, carcinomatous meningitis, and noninflammatory causes such as atherosclerosis, emboli, connective tissue disease, vasospasm, migraine-associated vasculopathy, and drug-associated causes. Some cases develop in the postpartum period and are self-limited.

Patients with any form of vasculopathy may present with insidious progression of combined white and gray matter infarctions, prominent headache, and cognitive decline. Brain biopsy or high-resolution conventional x-ray angiography is usually required to make the diagnosis **(Fig. 420-4)**. An inflammatory profile (elevated WBCs, elevated IgG index, bands on electrophoresis) found on lumbar puncture favors an inflammatory cause. In cases where inflammation is confirmed, aggressive immunosuppression with glucocorticoids, and often cyclophosphamide, is usually necessary to prevent progression; a diligent investigation for infectious causes such as tuberculosis is essential prior to immunosuppression. With prompt recognition and treatment, many patients can make an excellent recovery.

FIGURE 420-4

Cerebral angiogram from a 32-year-old male with central nervous system vasculopathy. Dramatic beading (*arrows*) typical of vasculopathy is seen.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Drugs, in particular amphetamines and perhaps cocaine, may cause stroke on the basis of acute hypertension or drug-induced vasculopathy. This vasculopathy is commonly due to vasospasm or atherosclerosis but cases of inflammatory vasculitis have also been reported. No data exist on the value of any treatment, but cessation of stimulants is prudent. Phenylpropanolamine has been linked with intracranial hemorrhage, as has cocaine and methamphetamine, perhaps related to a drug-induced vasculopathy. *Moyamoya disease* is a poorly understood occlusive disease involving large intracranial arteries, especially the distal internal carotid artery and the stem of the MCA and ACA. Vascular inflammation is absent. The lenticulostriate arteries develop a rich collateral circulation around the occlusive lesion, which gives the impression of a "puff of smoke" (*moyamoya* in Japanese) on conventional x-ray angiography. Other collaterals include transdural anastomoses between the cortical surface branches of the meningeal and scalp arteries. The disease occurs mainly in Asian children or young adults, but the appearance may be identical in adults who have atherosclerosis, particularly in association with diabetes. Intracranial hemorrhage may result from rupture of the moyamoya collaterals; thus, anticoagulation is risky. Progressive occlusion of large surface arteries can occur, producing large-artery distribution strokes. Surgical bypass of extracranial carotid arteries to the dura or MCAs may prevent stroke and hemorrhage.

Posterior reversible encephalopathy syndrome (PRES) can occur with head injury, seizure, migraine, sympathomimetic drug use, eclampsia, and in the postpartum period. The pathophysiology is uncertain but likely involves a hyperperfusion state where blood pressure exceeds the upper limit of cerebral autoregulation resulting in cerebral edema **(Chap. 301)**. Patients complain of headache and manifest fluctuating neurologic symptoms and signs, especially visual symptoms. Sometimes cerebral infarction ensues, but typically the clinical and imaging findings reverse completely. MRI findings are characteristic with the edema present within the occipital lobes but can be generalized and do not respect any single vascular territory. A closely related *reversible cerebral vasoconstriction syndrome* (RCVS) typically presents with sudden, severe headache closely

mimicking SAH. Patients may experience ischemic infarction and intracerebral hemorrhage and typically have new-onset, severe hypertension. Conventional x-ray angiography reveals changes in the vascular caliber throughout the hemispheres resembling vasculitis, but the process is noninflammatory. Oral calcium channel blockers may be effective in producing remission, and recurrence is rare.

Leukoaraiosis, or *periventricular white matter disease*, is the result of multiple small-vessel infarcts within the subcortical white matter. It is readily seen on CT or MRI scans as areas of white matter injury surrounding the ventricles and within the corona radiata. The pathophysiologic basis of the disease is lipohyalinosis of small penetrating arteries within the white matter, likely produced by chronic hypertension. Patients with periventricular white matter disease may develop a subcortical dementia syndrome, and it is likely that this common form of dementia may be delayed or prevented with antihypertensive medications **(Chap. 425)**.

CADASIL (cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy) is an inherited disorder that presents as small-vessel strokes, progressive dementia, and extensive symmetric white matter changes often including the anterior temporal lobes visualized by MRI. Approximately 40% of patients have migraine with aura, often manifest as transient motor or sensory deficits. Onset is usually in the fourth or fifth decade of life. This autosomal dominant condition is caused by one of several mutations in *Notch-3*, a member of a highly conserved gene family characterized by epidermal growth factor repeats in its extracellular domain. Other monogenic ischemic stroke syndromes include cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy (CARASIL) and hereditary endotheliopathy, retinopathy, nephropathy, and stroke (HERNS). Fabry's disease also produces both a large-vessel arteriopathy and small-vessel infarctions. The COL4A1 mutation is associated with multiple small vessel strokes with hemorrhagic transformation.

TRANSIENT ISCHEMIC ATTACKS

TIAs are episodes of stroke symptoms that last only briefly; the standard definition of duration is <24 h, but most TIAs last <1 h. If a relevant brain infarction is identified on brain imaging, the clinical entity is now classified as stroke regardless of the duration of symptoms. A normal brain imaging study following a TIA does not rule-out TIA; rather, the clinical syndrome is diagnostic. The causes of TIA are similar to the causes of ischemic stroke, but because TIAs may herald stroke, they are an important risk factor that should be considered separately and urgently. TIAs may arise from emboli to the brain or from in situ thrombosis of an intracranial vessel. With a TIA, the occluded blood vessel reopens and neurologic function is restored.

The risk of stroke after a TIA is ~10–15% in the first 3 months, with most events occurring in the first 2 days. This risk can be directly estimated using the well-validated $ABCD^2$ score **(Table 420-5)**. Therefore, urgent evaluation and treatment are justified. Because etiologies for stroke and TIA are identical, evaluation for TIA should parallel that of stroke (Fig. 420-2). The improvement characteristic of TIA is a contraindication to thrombolysis. However, because the risk of subsequent stroke in the first few days after a TIA is high, the opportunity to give rtPA rapidly if a subsequent stroke occurs may justify hospital admission for most patients. The combination of aspirin and clopidogrel was found to prevent stroke following TIA better than aspirin alone in a large Chinese randomized trial and is undergoing similar evaluation in an ongoing National Institutes of Health (NIH)-sponsored trial (POINT study). Failure to respond to the combination of aspirin and clopidogrel is linked to carriage of a common CYP2C19 polymorphism that leads to poor metabolism of clopidogrel into its active form. This mutation is common, particularly in Asians.

TABLE 420-5

Risk of Stroke Following Transient Ischemic Attack: The ABCD^2 Score

Clinical Factor	Score
A: Age ≥60 years	1
B: SBP >140 mmHg or DBP >90 mmHg	1
C: Clinical symptoms	
Unilateral weakness	2
Speech disturbance without weakness	1
D: Duration	
>60 min	2
10–59 min	1
D: Diabetes (oral medications or insulin)	1
Total Score	**Sum Each Category**
ABCD^2 Score Total	3-Month Rate of Stroke (%)[a]
0	0
1	2
2	3
3	3
4	8
5	12
6	17
7	22

[a]Data ranges are from five cohorts.

Abbreviations: DBP, diastolic blood pressure; SBP, systolic blood pressure.

Source: SC Johnston et al: Validation and refinement of score to predict very early stroke risk after transient ischaemic attack. Lancet 369:283, 2007.

TREATMENT

TREATMENT

Primary and Secondary Prevention of Stroke and TIA

GENERAL PRINCIPLES

Many medical and surgical interventions, as well as lifestyle modifications, are available for preventing stroke. Some of these can be widely applied because of their low cost and minimal risk; others are expensive and carry substantial risk but may be valuable for selected high-risk patients. Identification and control of modifiable risk factors, and especially hypertension, is the best strategy to reduce the burden of stroke, and the total number of strokes could be reduced substantially by these means (Table 420-4).

ATHEROSCLEROSIS RISK FACTORS

The relationship of various factors to the risk of atherosclerosis is described in **Chaps. 232** and **233**. Older age, diabetes mellitus, hypertension, tobacco smoking, abnormal blood cholesterol (particularly, low high-density lipoprotein [HDL] and/or elevated low-density lipoprotein [LDL]), and other factors are either proven or probable risk factors for ischemic stroke, largely by their link to atherosclerosis. Risk of stroke is much greater in those with prior stroke or TIA. Many cardiac conditions predispose to stroke, including atrial fibrillation and recent MI. Oral contraceptives and hormone replacement therapy increase stroke risk, and although rare, certain inherited and acquired hypercoagulable states predispose to stroke.

Hypertension is the most significant of the risk factors; in general, all hypertension should be treated to a target of <130/80 mmHg. Recent data (the Systolic Blood Pressure Intervention Trial—SPRINT) suggest that lowering systolic blood pressure <120 mmHg reduces stroke and heart attack 43% compared to systolic blood pressure <140 mmHg, without an increased risk of syncope or falls. The presence of known cerebrovascular disease is not a contraindication to treatment aimed at achieving normotension. Data are particularly strong in support of thiazide diuretics and angiotensin-converting enzyme inhibitors.

Several trials have confirmed that statin drugs reduce the risk of stroke even in patients without elevated LDL or low HDL. The Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) trial showed benefit in secondary stroke reduction for patients with recent stroke or TIA who were prescribed atorvastatin, 80 mg/d. The primary prevention trial, Justification for the Use of Statins in Prevention: An Intervention Trial Evaluating Rosuvastatin (JUPITER), found that patients with low LDL (<130 mg/dL) caused by elevated C-reactive protein benefitted by daily use of this statin. Primary stroke occurrence was reduced by 51% (hazard ratio 0.49, $p = .004$), and there was no increase in the rates of intracranial hemorrhage. Meta-analysis has also supported a primary treatment effect for statins given acutely for ischemic stroke. Therefore, a statin should be considered in all patients with prior ischemic stroke. Tobacco smoking should be discouraged in all patients **(Chap. 448)**. The use of pioglitazone (an agonist of peroxisome proliferator-activated receptor gamma) in patients with type 2 diabetes and previous stroke does not lower stroke, MI, or vascular death rates, but is effective in lowering vascular events patients with stroke and insulin resistance alone. Diabetes prevention is likely the most effective strategy for primary and secondary stroke prevention.

ANTIPLATELET AGENTS FOR STROKE PREVENTION

Platelet antiaggregation agents can prevent atherothrombotic events, including TIA and stroke, by inhibiting the formation of intraarterial platelet aggregates. These can form on diseased arteries, induce thrombus formation, and occlude or embolize into the distal circulation. Aspirin, clopidogrel, and the combination of aspirin plus extended-release dipyridamole are the antiplatelet agents most commonly used for this purpose. Ticlopidine has been largely abandoned because of its adverse effects but may be used as an alternative to clopidogrel. Ticagrelor has not been found to be better than aspirin for stroke prevention.

Aspirin is the most widely studied antiplatelet agent. Aspirin acetylates platelet cyclooxygenase, which irreversibly inhibits the formation in platelets of thromboxane A_2, a platelet aggregating and vasoconstricting prostaglandin. This effect is permanent and lasts for the usual 8-day life of the platelet. Paradoxically, aspirin also inhibits the formation in endothelial cells of prostacyclin, an antiaggregating and vasodilating prostaglandin. This effect is transient. As soon as aspirin is cleared from the blood, the nucleated endothelial cells again produce prostacyclin. Aspirin in low doses given once daily inhibits the production of thromboxane A_2 in platelets without substantially inhibiting prostacyclin formation. Higher doses of aspirin have not been proven to be more effective than lower doses.

Ticlopidine and clopidogrel block the adenosine diphosphate (ADP) receptor on platelets and thus prevent the cascade resulting in activation of the glycoprotein IIb/IIIa receptor that leads to fibrinogen binding to the platelet and consequent platelet aggregation. Ticlopidine is more effective than aspirin; however, it has the disadvantage of causing diarrhea, skin rash, and, in rare instances, neutropenia and thrombotic thrombocytopenic purpura (TTP). Clopidogrel rarely causes TTP but does not cause neutropenia. The Clopidogrel versus Aspirin in Patients at Risk of Ischemic Events (CAPRIE) trial, which led to FDA approval, found that it was only marginally more effective than aspirin in reducing risk of stroke. The Management of Atherothrombosis with Clopidogrel in High-Risk Patients (MATCH) trial was a large multicenter, randomized, double-blind study that compared clopidogrel in combination with aspirin to clopidogrel alone in the secondary prevention of TIA or stroke. The MATCH trial found no difference in TIA or stroke prevention with this combination, but did show a small but significant increase in major bleeding complications (3 vs 1%). In the Clopidogrel for High Atherothrombotic Risk and Ischemic Stabilization, Management, and Avoidance (CHARISMA) trial, which included a subgroup of patients with prior stroke or TIA along with other groups at high risk of cardiovascular events, there was no benefit of clopidogrel combined with aspirin compared to aspirin alone. Lastly, the SPS3 trial looked at the long-term combination of clopidogrel

and aspirin versus clopidogrel alone in small-vessel stroke and found no improvement in stroke prevention and a significant increase in both hemorrhage and death. Thus, the long-term use of clopidogrel in combination with aspirin is not recommended for stroke prevention.

The short-term combination of clopidogrel with aspirin may be effective in preventing second stroke, however. A trial of 5170 Chinese patients enrolled within 24 h of TIA or minor ischemic stroke found that a clopidogrel-aspirin regimen (clopidogrel 300 mg load then 75 mg/d with aspirin 75 mg for the first 21 days) was superior to aspirin (75 mg/d) alone, with 90-day stroke risk decreased from 11.7 to 8.2% (p <.001) and no increase in major hemorrhage. This benefit was limited to those not carrying the CYP2C19 polymorphism associated with clopidogrel hypometabolism. An international NIH-sponsored trial of similar design is ongoing.

Dipyridamole is an antiplatelet agent that inhibits the uptake of adenosine by a variety of cells, including those of the vascular endothelium. The accumulated adenosine is an inhibitor of aggregation. At least in part through its effects on platelet and vessel wall phosphodiesterases, dipyridamole also potentiates the antiaggregatory effects of prostacyclin and nitric oxide produced by the endothelium and acts by inhibiting platelet phosphodiesterase, which is responsible for the breakdown of cyclic AMP. The resulting elevation in cyclic AMP inhibits aggregation of platelets. Dipyridamole is erratically absorbed depending on stomach pH, but a newer formulation combines timed-release dipyridamole, 200 mg, with aspirin, 25 mg, and has better oral bioavailability. This combination drug was studied in three trials. The European Stroke Prevention Study (ESPS) II showed efficacy of both 50 mg/d of aspirin and extended-release dipyridamole in preventing stroke, and a significantly better risk reduction when the two agents were combined. The open-label ESPRIT (European/Australasian Stroke Prevention in Reversible Ischaemia Trial) trial confirmed the ESPS-II results. After 3.5 years of follow-up, 13% of patients on aspirin and dipyridamole and 16% on aspirin alone (hazard ratio 0.80, 95% confidence index [CI] 0.66–0.98) met the primary outcome of death from all vascular causes. In the Prevention Regimen for Effectively Avoiding Second Strokes (PRoFESS) trial, the combination of extended-release dipyridamole and aspirin was compared directly with clopidogrel with and without the angiotensin receptor blocker telmisartan; there were no differences in the rates of second stroke (9% each) or degree of disability in patients with median follow-up of 2.4 years. Telmisartan also had no effect on these outcomes. This suggests that these antiplatelet regimens are similar and raises questions about default prescription of agents to block the angiotensin pathway in all stroke patients. The principal side effect of dipyridamole is headache. The combination capsule of extended-release dipyridamole and aspirin is approved for prevention of stroke.

Many large clinical trials have demonstrated clearly that most antiplatelet agents reduce the risk of all important vascular atherothrombotic events (i.e., ischemic stroke, MI, and death due to all vascular causes) in patients at risk for these events. The overall *relative* reduction in risk of nonfatal stroke is about 25–30% and of all vascular events is about 25%. The *absolute* reduction varies considerably, depending on the patient's risk. Individuals at very low risk for stroke seem to experience the same relative reduction, but their risks may be so low that the "benefit" is meaningless. Conversely, individuals with a 10–15% risk of vascular events per year experience a reduction to about 7.5–11%.

Aspirin is inexpensive, can be given in low doses, and could be recommended for all adults to prevent both stroke and MI. However, it causes epigastric discomfort, gastric ulceration, and gastrointestinal hemorrhage, which may be asymptomatic or life threatening. Consequently, not every 40- or 50-year-old should be advised to take aspirin regularly because the risk of atherothrombotic stroke is extremely low and is outweighed by the risk of adverse side effects. Conversely, every patient who has experienced an atherothrombotic stroke or TIA and has no contraindication should be taking an antiplatelet agent regularly because the average annual risk of another stroke is 8–10%; another few percent will experience an MI or vascular death. Clearly, the likelihood of benefit far outweighs the risks of treatment.

The choice of antiplatelet agent and dose must balance the risk of stroke, the expected benefit, and the risk and cost of treatment. However, there are no definitive data, and opinions vary. Many authorities believe low-dose (30–75 mg/d) and high-dose (650–1300 mg/d) aspirin are about equally effective. Some advocate very low doses to avoid adverse effects, and still others advocate very high doses to be sure the benefit is maximal. Most physicians in North America recommend 81–325 mg/d, whereas most Europeans recommend 50–100 mg. Clopidogrel and extended-release dipyridamole plus aspirin are being increasingly recommended as first-line drugs for secondary prevention. Similarly, the choice of aspirin, clopidogrel, or dipyridamole plus aspirin must balance the fact that the latter are more effective than aspirin but the cost is higher, and this is likely to affect long-term patient adherence. The use of platelet aggregation studies in individual patients taking aspirin is controversial because of limited data.

ANTICOAGULATION THERAPY AND EMBOLIC STROKE PREVENTION

Several trials have shown that anticoagulation (INR range, 2–3) in patients with chronic nonvalvular (nonrheumatic) atrial fibrillation (NVAF) prevents cerebral embolism and stroke and is safe. For primary prevention and for patients who have experienced stroke or TIA, anticoagulation with a VKA reduces the risk by about 67%, which clearly outweighs the 1–3% risk per year of a major bleeding complication. VKAs are difficult to dose, their effects vary with dietary intake of vitamin K, and they require frequent blood monitoring of the PTT/INR. Several newer oral anticoagulants (OACs) have recently been shown to be more convenient and efficacious for stroke prevention in NVAF. A randomized trial compared the oral thrombin inhibitor dabigatran to VKAs in a noninferiority trial to prevent stroke or systemic embolization in NVAF. Two doses of dabigatran were used: 110 mg/d and 150 mg/d. Both dose tiers of dabigatran were noninferior to VKAs in preventing second stroke and systemic

embolization, and the higher dose tier was superior (relative risk 0.66; 95% CI 0.53–0.82; p <.001) and the rate of major bleeding was lower in the lower dose tier of dabigatran compared to VKAs. Dabigatran requires no blood monitoring to titrate the dose, and its effect is independent of oral intake of vitamin K. Newer oral factor Xa inhibitors have also been found to be equivalent or safer and more effective than VKAs in NVAF stroke prevention. In the Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation (ARISTOTLE) trial, patients were randomized between apixaban, 5 mg twice daily, and dose-adjusted warfarin (INR 2–3). The combined endpoint of ischemic or hemorrhagic stroke or system embolism occurred in 1.27% of patients in the apixaban group and in 1.6% in the warfarin group (p <.001 for noninferiority and p <.01 for superiority). Major bleeding was 1% less, favoring apixaban (p <.001). Similar results were obtained in the Rivaroxaban Once Daily Oral Direct Factor Xa Inhibition Compared with Vitamin K Antagonism for Prevention of Stroke and Embolism Trial in Atrial Fibrillation (ROCKET-AF). Here, patients with NVAF were randomized to rivaroxaban versus warfarin: 1.7% of the factor Xa group and 2.2% of the warfarin group reached the endpoint of stroke and systemic embolism (p <.001 for noninferiority); intracranial hemorrhage was also lower with rivaroxaban. Finally, the factor Xa inhibitor edoxaban was also found to be noninferior to warfarin. Thus, oral factor Xa inhibitors are at least a suitable alternative to VKAs, and likely are superior both in efficacy and perhaps compliance.

For patients who cannot take anticoagulant medications, clopidogrel plus aspirin was compared to aspirin alone in the Atrial Fibrillation Clopidogrel Trial with Irbesartan for Prevention of Vascular Events (ACTIVE-A). Clopidogrel combined with aspirin was more effective than aspirin alone in preventing vascular events, principally stroke, but increased the risk of major bleeding (relative risk 1.57, p <.001).

The decision to use anticoagulation for primary prevention is based primarily on risk factors (Table 420-3). The history of a TIA or stroke tips the balance in favor of anticoagulation regardless of other risk factors. Intermittent atrial fibrillation carries the same risk of stroke as chronic atrial fibrillation, and several ambulatory studies of seemingly “cryptogenic” stroke have found evidence of intermittent atrial fibrillation in nearly 20% of patients monitored for a few weeks. Interrogation of implanted pacemakers also confirms an association between subclinical atrial fibrillation and stroke risk. Therefore, for patients with otherwise cryptogenic embolic stroke (no evidence of any other cause for stroke), ambulatory monitoring for 3–4 weeks is a reasonable strategy to determine the best prophylactic therapy.

Because of the high annual stroke risk in untreated rheumatic heart disease with atrial fibrillation, primary prophylaxis against stroke has not been studied in a double-blind fashion. These patients generally should receive long-term anticoagulation. Dabigatran and the oral Xa inhibitors have not been studied in this population.

Anticoagulation also reduces the risk of embolism in acute MI. Most clinicians recommend a 3-month course of anticoagulation when there is anterior Q-wave infarction, substantial left ventricular dysfunction, congestive heart failure, mural thrombosis, or atrial fibrillation. OACs are recommended long-term if atrial fibrillation persists.

Stroke secondary to thromboembolism is one of the most serious complications of prosthetic heart valve implantation. The intensity of anticoagulation and/or antiplatelet therapy is dictated by the type of prosthetic valve and its location. Dabigatran may be less effective than warfarin, and the oral Xa inhibitors have not been studied in this population.

If the embolic source cannot be eliminated, anticoagulation should in most cases be continued indefinitely. Many neurologists recommend combining antiplatelet agents with anticoagulants for patients who “fail” anticoagulation (i.e., have another stroke or TIA), but the evidence basis for this is lacking.

ANTICOAGULATION THERAPY AND NONCARDIOGENIC STROKE

Data do not support the use of long-term VKAs for preventing atherothrombotic stroke for either intracranial or extracranial cerebrovascular disease. The Warfarin-Aspirin Recurrent Stroke Study (WARSS) found no benefit of warfarin sodium (INR 1.4–2.8) over aspirin, 325 mg, for secondary prevention of stroke but did find a slightly higher bleeding rate in the warfarin group; a European study confirmed this finding. The Warfarin and Aspirin for Symptomatic Intracranial Disease (WASID) study (see below) demonstrated no benefit of warfarin (INR 2–3) over aspirin in patients with symptomatic intracranial atherosclerosis and found a higher rate of bleeding complications. Trials are ongoing testing Factor Xa medications for prevention of embolic stroke of unknown source.

TREATMENT

TREATMENT

Carotid Atherosclerosis

Carotid atherosclerosis can be removed surgically (endarterectomy) or mitigated with endovascular stenting with or without balloon angioplasty. Anticoagulation has not been directly compared with antiplatelet therapy for carotid disease.

SURGICAL THERAPY

Symptomatic carotid stenosis was studied in the North American Symptomatic Carotid Endarterectomy Trial (NASCET) and the European Carotid Surgery Trial (ECST). Both showed a substantial benefit for surgery in patients with stenosis of ≥70%. In NASCET, the average cumulative ipsilateral stroke risk at 2 years was 26% for patients treated medically and 9% for those receiving the same medical treatment plus a carotid endarterectomy. This 17% *absolute* reduction in the surgical group is a 65% *relative* risk reduction favoring surgery (Table 420-4). NASCET also showed a significant, although less robust, benefit for patients with 50–70% stenosis. ECST found harm for patients with stenosis <30% treated surgically.

A patient's risk of stroke and possible benefit from surgery are related to the presence of retinal versus hemispheric symptoms, degree of arterial stenosis, extent of associated medical conditions (of note, NASCET and ECST excluded "high-risk" patients with significant cardiac, pulmonary, or renal disease), institutional surgical morbidity and mortality, timing of surgery relative to symptoms, and other factors. A recent meta-analysis of the NASCET and ECST trials demonstrated that endarterectomy is most beneficial when performed within 2 weeks of symptom onset. In addition, benefit is more pronounced in patients >75 years, and men appear to benefit more than women.

In summary, a patient with recent symptomatic hemispheric ischemia, high-grade stenosis in the appropriate internal carotid artery, and an institutional perioperative morbidity and mortality rate of ≤6% generally should undergo carotid endarterectomy. If the perioperative stroke rate is >6% for any particular surgeon, however, the benefits of carotid endarterectomy are questionable.

The indications for surgical treatment of *asymptomatic carotid disease* have been clarified by the results of the Asymptomatic Carotid Atherosclerosis Study (ACAS) and the Asymptomatic Carotid Surgery Trial (ACST). ACAS randomized asymptomatic patients with ≥60% stenosis to medical treatment with aspirin or the same medical treatment plus carotid endarterectomy. The surgical group had a risk over 5 years for ipsilateral stroke (and any perioperative stroke or death) of 5.1%, compared to a risk in the medical group of 11%. Although this demonstrates a 53% *relative* risk reduction, the *absolute* risk reduction is only 5.9% over 5 years, or 1.2% annually (Table 420-4). Nearly one-half of the strokes in the surgery group were caused by preoperative angiograms. ACST randomized asymptomatic patients with >60% carotid stenosis to endarterectomy or medical therapy. The 5-year risk of stroke in the surgical group (including perioperative stroke or death) was 6.4%, compared to 11.8% in the medically treated group (46% relative risk reduction and 5.4% absolute risk reduction).

In both ACAS and ACST, the perioperative complication rate was higher in women, perhaps negating any benefit in the reduction of stroke risk within 5 years. It is possible that with longer follow-up, a clear benefit in women will emerge. At present, carotid endarterectomy in asymptomatic women remains particularly controversial.

In summary, the natural history of asymptomatic stenosis is an ~2% per year stroke rate, whereas symptomatic patients experience a 13% per year risk of stroke. Whether to recommend carotid revascularization for an asymptomatic patient is somewhat controversial and depends on many factors, including patient preference, degree of stenosis, age, gender, and comorbidities. Medical therapy for reduction of atherosclerosis risk factors, including cholesterol-lowering agents and antiplatelet medications, is generally recommended for patients with asymptomatic carotid stenosis. As with atrial fibrillation, it is imperative to counsel the patient about TIAs so that therapy can be revised if symptoms develop.

ENDOVASCULAR THERAPY

Balloon angioplasty coupled with stenting is being used with increasing frequency to open stenotic carotid arteries and maintain their patency. These techniques can treat carotid stenosis not only at the bifurcation but also near the skull base and in the intracranial segments. The Stenting and Angioplasty with Protection in Patients at High Risk for Endarterectomy (SAPPHIRE) trial randomized high-risk patients (defined as patients with clinically significant coronary or pulmonary disease, contralateral carotid occlusion, restenosis after endarterectomy, contralateral laryngeal-nerve palsy, prior radical neck surgery or radiation, or age >80) with symptomatic carotid stenosis >50% or asymptomatic stenosis >80% to either stenting combined with a distal emboli-protection device or endarterectomy. The risk of death, stroke, or MI within 30 days and ipsilateral stroke or death within 1 year was 12.2% in the stenting group and 20.1% in the endarterectomy group ($p = .055$), suggesting that stenting is at the very least comparable to endarterectomy as a treatment option for this patient group at high risk of surgery. However, the outcomes with both interventions may not have been better than leaving the carotid stenoses untreated, particularly for the asymptomatic patients, and much of the benefit seen in the stenting group was due to a reduction in periprocedure MI. Two randomized trials comparing stents to endarterectomy in lower-risk patients have been published. The Carotid Revascularization Endarterectomy versus Stenting Trial (CREST) enrolled patients with either asymptomatic or symptomatic stenosis. The 30-day risk of stroke was 4.1% in the stent group and 2.3% in the surgical group, but the 30-day risk of MI was 1.1% in the stent group and 2.3% in the surgery group, suggesting relative equivalence of risk between the procedures. At median follow-up of 2.5 years, the combined endpoint of stroke, MI, and death was the same (7.2% stent vs 6.8% surgery) and remained so at 10-year follow up. The rate of restenosis at 2 years was also similar in both groups. The International Carotid Stenting Study (ICSS) randomized symptomatic patients to stents versus endarterectomy and found a different result: At 120 days, the incidence of stroke, MI, or death was 8.5% in the stenting group versus 5.2% in the endarterectomy group ($p = .006$). At median follow-up of 5 years these differences were no longer significant except a small increase in non-disabling stroke in the stenting group but no change in the average disability. In meta-analysis, carotid endarterectomy (CEA) is less morbid in older patients (aged ≥70) then is stenting. Investigation is on-going in asymptomatic

patients to compare medical therapy to stenting and CEA. This will likely answer how well medical patients do with more modern medical therapy (statins, close blood pressure control, and life-style modification).

BYPASS SURGERY

Extracranial-to-intracranial (EC-IC) bypass surgery has been proven ineffective for atherosclerotic stenoses that are inaccessible to conventional carotid endarterectomy. In patients with recent stroke, an associated carotid occlusion, and evidence of inadequate perfusion of the brain as measured with positron emission tomography, no benefit from EC-IC bypass was found in a trial stopped for futility.

INTRACRANIAL ATHEROSCLEROSIS

The WASID trial randomized patients with symptomatic stenosis (50–99%) of a major intracranial vessel to either high-dose aspirin (1300 mg/d) or warfarin (target INR, 2.0–3.0), with a combined primary endpoint of ischemic stroke, brain hemorrhage, or death from vascular cause other than stroke. The trial was terminated early because of an increased risk of adverse events related to warfarin anticoagulation. With a mean follow-up of 1.8 years, the primary endpoint was seen in 22.1% of patients in the aspirin group and 21.8% of the warfarin group. Death from any cause was seen in 4.3% of the aspirin group and 9.7% of the warfarin group; 3.2% of patients on aspirin experienced major hemorrhage, compared to 8.3% of patients taking warfarin.

Intracranial stenting of intracranial atherosclerosis was found to be dramatically harmful compared to aspirin in the Stenting and Aggressive Medical Management for Preventing Recurrent Stroke in Intracranial Stenosis (SAMMPRIS) trial. This trial enrolled newly symptomatic TIA or minor stroke patients with associated 70–99% intracranial stenosis to primary stenting with a self-expanding stent or to medical management. Both groups received clopidogrel, aspirin, statin, and aggressive control of blood pressure. The endpoint of stroke or death occurred in 14.7% of the stented group and 5.8% of the medically treated groups ($p = .002$). This low rate of second stroke was significantly lower than in the WASID trial and suggests that aggressive medical management had a marked influence on secondary stroke risk. A concomitant study of balloon-expandable stenting was halted early at 125 patients because of the negative SAMMPRIS results and due to harm. Therefore, routine use of intracranial stenting is harmful, and medical therapy is superior for intracranial atherosclerosis.

Dural Sinus Thrombosis

Limited evidence exists to support short-term use of anticoagulants, regardless of the presence of intracranial hemorrhage, for venous infarction following sinus thrombosis. The long-term outcome for most patients, even those with intracerebral hemorrhage, is excellent.

FURTHER READING

Goyal M et al: Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials. Lancet 387:1723, 2016. [PubMed: 26898852]

Group SR et al: A randomized trial of intensive versus standard blood-pressure control. N Engl J Med 373:2103, 2015. [PubMed: 26551272]

January CT et al: 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. J Am Coll Cardiol 64:e1, 2014. [PubMed: 24685669]

Larsson SC et al: Prognosis of carotid dissecting aneurysms: Results from CADISS and a systematic review. Neurology 88:646, 2017. [PubMed: 28087823]

Powers WJ et al: 2018 Guidelines for the Early Management of Patients With Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke 49:e46, 2018.

Saver JL et al: Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: A meta-analysis. JAMA 316:1279, 2016. [PubMed: 27673305]

Torbey MT et al: Evidence-based guidelines for the management of large hemispheric infarction: A statement for health care professionals from the Neurocritical Care Society and the German Society for Neuro-intensive Care and Emergency Medicine. Neurocrit Care 22:146, 2015. [PubMed: 25605626]

Chapter 421: Intracranial Hemorrhage

Wade S. Smith; J. Claude Hemphill, III; S. Claiborne Johnston

FIGURE 421-1

INTRODUCTION

Intracranial hemorrhage is a form of stroke **(see Chap. 419)**. Compared to ischemic stroke, patients with intracranial hemorrhage are more likely to present with headache; however, brain imaging is required to distinguish these entities. CT imaging of the head is highly sensitive and specific for intracranial hemorrhage and determines the location(s) of bleeding. Hemorrhages are classified by their location and the underlying vascular pathology. Hemorrhage directly into the brain parenchyma, also known as intracerebral hemorrhage (ICH), and arteriovenous malformations (AVMs) of the brain will be considered here. Other categories of hemorrhage include bleeding into subdural and epidural spaces, usually caused by trauma **(Chap 435)**, and subarachnoid hemorrhage due to trauma or the rupture of an intracranial aneurysm **(Chap. 302)**.

DIAGNOSIS

Intracranial hemorrhage is often identified on noncontrast CT imaging of the brain during the acute evaluation of stroke. Because CT is more widely available and may be logistically easier to perform than MRI, CT imaging is generally the preferred method for acute stroke evaluation **(Fig. 421-1)**. The location of the hemorrhage narrows the differential diagnosis to a few entities. **Table 421-1** lists the causes and anatomic spaces involved in hemorrhages.

FIGURE 421-1

Hypertensive hemorrhage. Transaxial noncontrast computed tomography scan through the region of the basal ganglia reveals a hematoma involving the left putamen in a patient with rapidly progressive onset of right hemiparesis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 421-1

Causes of Intracranial Hemorrhage

Cause	Location	Comments
Head trauma	Intraparenchymal: frontal lobes, anterior temporal lobes; subarachnoid; extra-axial (subdural, epidural)	Coup and contrecoup injury during brain deceleration
Hypertensive hemorrhage	Putamen, globus pallidus, thalamus, cerebellar hemisphere, pons	Chronic hypertension produces hemorrhage from small (~30–100 μm) vessels in these regions
Transformation of prior ischemic infarction	Basal ganglion, subcortical regions, lobar	Occurs in 1–6% of ischemic strokes with predilection for large hemispheric infarctions
Metastatic brain tumor	Lobar	Lung, choriocarcinoma, melanoma, renal cell carcinoma, thyroid, atrial myxoma
Coagulopathy	Any	Risk for ongoing hematoma expansion
Drug	Any, lobar, subarachnoid	Cocaine, amphetamine
Arteriovenous malformation	Lobar, intraventricular, subarachnoid	Risk is ~2–3% per year for bleeding if previously unruptured
Aneurysm	Subarachnoid, intraparenchymal, rarely subdural	Mycotic and nonmycotic forms of aneurysms
Amyloid angiopathy	Lobar	Degenerative disease of intracranial vessels; associated with dementia, rare in patients <60 years
Cavernous angioma	Intraparenchymal	Multiple cavernous angiomas linked to mutations in *KRIT1*, *CCM2*, and *PDCD10* genes

Cause	Location	Comments
Dural arteriovenous fistula	Lobar, subarachnoid	Produces bleeding by venous hypertension
Capillary telangiectasias	Usually brainstem	Rare cause of hemorrhage

EMERGENCY MANAGEMENT

Close attention should be paid to airway management because a reduction in the level of consciousness is common and often progressive. The initial blood pressure should be maintained until the results of the CT scan are reviewed and demonstrate ICH. In theory, a higher blood pressure should promote hematoma expansion, but it remains unclear if lowering of blood pressure reduces hematoma growth. Recent clinical trials have shown that systolic blood pressure (SBP) can be safely lowered acutely and rapidly to <140 mmHg in patients with spontaneous ICH whose initial SBP was 150–220 mmHg. The INTERACT2 trial was a large phase 3 clinical trial to address the effect of acute blood pressure lowering on ICH functional outcome. INTERACT2 randomized patients with spontaneous ICH within 6 h of onset and a baseline SBP of 150–220 mmHg to two different SBP targets (<140 and <180 mmHg). In those with the target SBP <140 mmHg, 52% had an outcome of death or major disability at 90 days compared with 55.6% of those with a target SBP <180 mmHg ($p = .06$). There was a significant shift to improved outcomes in the lower blood pressure arm, whereas both groups had a similar mortality. ATACH2 was a similarly designed clinical trial that assessed the same blood pressure targets but demonstrated no difference in outcome between groups. Current U.S. and European guidelines emphasize that blood pressure lowering to a target SBP is likely safe and possibly beneficial. However, these guidelines were completed prior to publication of the ATACH2 results, thus the specific optimal target remains a point of debate. In patients who have higher SBP on presentation or who are deeply comatose with possible elevated intracranial pressure (ICP), it is unclear whether these clinical trial results apply. In patients who have ICP monitors in place, current recommendations are that maintaining the cerebral perfusion pressure (mean arterial pressure [MAP] minus ICP) at 50–70 mmHg is reasonable, depending on the individual patient's cerebral autoregulation status **(Chap. 301)**. Blood pressure should be lowered with nonvasodilating IV drugs such as nicardipine, labetalol, or esmolol. Patients with cerebellar hemorrhages with depressed mental status or radiographic evidence of hydrocephalus should undergo urgent neurosurgical evaluation; these patients require close monitoring because they can deteriorate rapidly. Based on the clinical examination and CT findings, further imaging studies may be necessary, including MRI or conventional x-ray angiography. Stuporous or comatose patients with clinical and imaging signs of herniation are generally treated presumptively for elevated ICP with tracheal intubation and sedation, administration of osmotic diuretics such as mannitol or hypertonic saline, and elevation of the

head of the bed while surgical consultation is obtained **(Chap. 301)**. Reversal of coagulopathy and consideration of surgical evacuation of the hematoma (detailed below) are two other principal aspects of initial emergency management.

INTRACEREBRAL HEMORRHAGE

ICH accounts for ~10% of all strokes, and about 35–45% of patients die within the first month. Incidence rates are particularly high in Asians and blacks. Hypertension, coagulopathy, sympathomimetic drugs (cocaine, methamphetamine), and cerebral amyloid angiopathy (CAA) cause most of these hemorrhages. Advanced age and heavy alcohol consumption increase the risk, and cocaine and methamphetamine use is one of the most important causes in the young.

Hypertensive ICH

PATHOPHYSIOLOGY

Hypertensive ICH usually results from spontaneous rupture of a small penetrating artery deep in the brain. The most common sites are the basal ganglia (especially the putamen), thalamus, cerebellum, and pons. The small arteries in these areas seem most prone to hypertension-induced vascular injury. When hemorrhages occur in other brain areas or in nonhypertensive patients, greater consideration should be given to other causes such as hemorrhagic disorders, neoplasms, vascular malformations, vasculitis, and CAA. The hemorrhage may be small, or a large clot may form and compress adjacent tissue, causing herniation and death. Blood may also dissect into the ventricular space, which substantially increases morbidity and may cause hydrocephalus.

Most hypertensive ICHs initially develop over 30–90 min, whereas those associated with anticoagulant therapy may evolve for as long as 24–48 h. However, it is now recognized that about a third of patients even with no coagulopathy may have significant hematoma expansion with the first day. Within 48 h, macrophages begin to phagocytize the hemorrhage at its outer surface. After 1–6 months, the hemorrhage is generally resolved to a slitlike cavity lined with a glial scar and hemosiderin-laden macrophages.

CLINICAL MANIFESTATIONS

ICH generally presents as the abrupt onset of a focal neurologic deficit. Seizures are uncommon. Although clinical symptoms may be maximal at onset, commonly the focal deficit worsens over 30–90 min and is associated with a diminishing level of consciousness and signs of increased ICP such as headache and vomiting.

The putamen is the most common site for hypertensive hemorrhage, and the adjacent internal capsule is usually damaged (Fig. 421-1). Contralateral hemiparesis is therefore the sentinel sign. When mild, the face sags on one side over 5–30 min, speech becomes slurred, the arm and leg gradually weaken, and the eyes deviate away from the side of the hemiparesis. The paralysis may worsen until the affected limbs become

flaccid or extend rigidly. When hemorrhages are large, drowsiness gives way to stupor as signs of upper brainstem compression appear. Coma ensues, accompanied by deep, irregular, or intermittent respiration, a dilated and fixed ipsilateral pupil, and decerebrate rigidity. In milder cases, edema in adjacent brain tissue may cause progressive deterioration over 12–72 h.

Thalamic hemorrhages also produce a contralateral hemiplegia or hemiparesis from pressure on, or dissection into, the adjacent internal capsule. A prominent sensory deficit involving all modalities is usually present. Aphasia, often with preserved verbal repetition, may occur after hemorrhage into the dominant thalamus, and constructional apraxia or mutism occurs in some cases of nondominant hemorrhage. There may also be a homonymous visual field defect. Thalamic hemorrhages cause several typical ocular disturbances by extension inferiorly into the upper midbrain. These include deviation of the eyes downward and inward so that they appear to be looking at the nose, unequal pupils with absence of light reaction, skew deviation with the eye opposite the hemorrhage displaced downward and medially, ipsilateral Horner's syndrome, absence of convergence, paralysis of vertical gaze, and retraction nystagmus. Patients may later develop a chronic, contralateral pain syndrome (Déjérine-Roussy syndrome).

In pontine hemorrhages, deep coma with quadriplegia often occurs over a few minutes. Typically, there is prominent decerebrate rigidity and "pinpoint" (1 mm) pupils that react to light. There is impairment of reflex horizontal eye movements evoked by head turning (doll's-head or oculocephalic maneuver) or by irrigation of the ears with ice water **(Chap. 300)**. Hyperpnea, severe hypertension, and hyperhidrosis are common. Most patients with deep coma from pontine hemorrhage ultimately die, or develop a locked-in state, but small hemorrhages are compatible with survival and significant recovery.

Cerebellar hemorrhages usually develop over several hours and are characterized by occipital headache, repeated vomiting, and ataxia of gait. In mild cases, there may be no other neurologic signs except for gait ataxia. Dizziness or vertigo may be prominent. There is often paresis of conjugate lateral gaze toward the side of the hemorrhage, forced deviation of the eyes to the opposite side, or an ipsilateral sixth nerve palsy. Less frequent ocular signs include blepharospasm, involuntary closure of one eye, ocular bobbing, and skew deviation. Dysarthria and dysphagia may occur. As the hours pass, the patient often becomes stuporous and then comatose from brainstem compression or obstructive hydrocephalus; immediate surgical evacuation before severe brainstem compression occurs may be lifesaving. Hydrocephalus from fourth ventricle compression can be relieved by external ventricular drainage; however, in this situation definitive hematoma evacuation is recommended rather than treatment with ventricular drainage alone. If the deep cerebellar nuclei are spared, full recovery is common.

Lobar Hemorrhage

The major neurologic deficit with an occipital hemorrhage is hemianopsia; with a left temporal hemorrhage, aphasia and delirium; with a parietal hemorrhage, hemisensory loss; and with frontal hemorrhage, arm weakness. Large hemorrhages may be associated with stupor or coma if they compress the thalamus or midbrain. Most patients with lobar hemorrhages have focal headaches, and more than one-half vomit or are drowsy. Stiff neck and seizures are uncommon.

Other Causes of ICH

CAA is a disease of the elderly in which arteriolar degeneration occurs and amyloid is deposited in the walls of the cerebral arteries. Amyloid angiopathy causes both single and recurrent lobar hemorrhages and is probably the most common cause of lobar hemorrhage in the elderly. It accounts for some intracranial hemorrhages associated with IV thrombolysis given for myocardial infarction. This disorder can be suspected in patients who present with multiple hemorrhages (and infarcts) over several months or years or in patients with “microbleeds” in the cortex, seen on brain MRI sequences sensitive for hemosiderin (iron-sensitive imaging), but it is definitively diagnosed by pathologic demonstration of Congo red staining of amyloid in cerebral vessels. The ε2 and ε4 allelic variations of the apolipoprotein E gene are associated with increased risk of recurrent lobar hemorrhage and may therefore be markers of amyloid angiopathy. Positron emission tomography imaging can image amyloid-beta deposits in CAA using specific antibody labels and may be helpful in diagnosing CAA noninvasively. Although cerebral biopsy is the most definitive method of diagnosis, evidence of inflammation on lumbar puncture should prompt consideration of CAA-associated vasculitis as an underlying cause and oral glucocorticoids may be beneficial. Non-inflammatory CAA has no specific treatment. Oral anticoagulants are typically avoided.

Cocaine and *methamphetamine* are frequent causes of stroke in young (age <45 years) patients. ICH, ischemic stroke, and subarachnoid hemorrhage (SAH) are all associated with stimulant use. Angiographic findings vary from completely normal arteries to large-vessel occlusion or stenosis, vasospasm, or changes consistent with vasculopathy. The mechanism of sympathomimetic-related stroke is not known, but cocaine enhances sympathetic activity causing acute, sometimes severe, hypertension, and this may lead to hemorrhage. Slightly more than one-half of stimulant-related intracranial hemorrhages are intracerebral and the rest are subarachnoid. In cases of SAH, a saccular aneurysm is usually identified. Presumably, acute hypertension causes aneurysmal rupture.

Head injury often causes intracranial bleeding. The common sites are intraparenchymal (especially temporal and inferior frontal lobes) and into the subarachnoid, subdural, and epidural spaces. Trauma must be considered in any patient with an unexplained acute neurologic deficit (hemiparesis, stupor, or confusion), particularly if the deficit occurred in the context of a fall (Chap. 435).

Intracranial hemorrhages associated with *anticoagulant therapy* can occur at any location; they are often lobar or subdural. Anticoagulant-related ICHs may continue to evolve over 24–48 h, especially if coagulopathy is insufficiently reversed. Coagulopathy and thrombocytopenia should be reversed rapidly, as discussed below. ICH associated with *hematologic disorders* (leukemia, aplastic anemia, thrombocytopenic purpura) can occur at any site and may present as multiple ICHs. Skin and mucous membrane bleeding may be evident and offers a diagnostic clue.

Hemorrhage into a *brain tumor* may be the first manifestation of neoplasm. Choriocarcinoma, malignant melanoma, renal cell carcinoma, and bronchogenic carcinoma are among the most common metastatic tumors associated with ICH. Glioblastoma multiforme in adults and medulloblastoma in children may also have areas of ICH.

Hypertensive encephalopathy is a complication of malignant hypertension. In this acute syndrome, severe hypertension is associated with headache, nausea, vomiting, convulsions, confusion, stupor, and coma. Focal or lateralizing neurologic signs, either transitory or permanent, may occur but are infrequent and therefore suggest some other vascular disease (hemorrhage, embolism, or atherosclerotic thrombosis). There are retinal hemorrhages, exudates, papilledema (hypertensive retinopathy), and evidence of renal and cardiac disease. In most cases, ICP and CSF protein levels are elevated. MRI brain imaging shows a pattern of typically posterior (occipital > frontal) brain edema that is reversible and termed *reversible posterior leukoencephalopathy*. The hypertension may be essential or due to chronic renal disease, acute glomerulonephritis, acute toxemia of pregnancy, pheochromocytoma, or other causes. Lowering the blood pressure reverses the process, but stroke can occur, especially if blood pressure is lowered too rapidly. Neuropathologic examination reveals multifocal to diffuse cerebral edema and hemorrhages of various sizes from petechial to massive. Microscopically, there is necrosis of arterioles, minute cerebral infarcts, and hemorrhages. The term *hypertensive encephalopathy* should be reserved for this syndrome and not for chronic recurrent headaches, dizziness, recurrent transient ischemic attacks, or small strokes that often occur in association with high blood pressure. Distinguishing hypertensive encephalopathy with ICH from hypertensive ICH is important since aggressive lowering of SBP to 140–180 mmHg acutely is usually considered in hypertensive ICH but less aggressive measures should be used in hypertensive encephalopathy. Having no alteration in mental status or other prodrome prior to the ICH favors hypertensive ICH as the disease.

Primary intraventricular hemorrhage is rare and should prompt investigation for an underlying vascular anomaly. Sometimes bleeding begins within the periventricular substance of the brain and dissects into the ventricular system without leaving signs of intraparenchymal hemorrhage. Alternatively, bleeding can arise from periependymal veins. Vasculitis, usually polyarteritis nodosa or lupus erythematosus, can produce hemorrhage in any region of the central nervous system; most hemorrhages are associated with hypertension, but the arteritis itself may cause bleeding by disrupting the vessel wall. Nearly one-half of patients with primary intraventricular hemorrhage have identifiable bleeding sources seen using conventional angiography.

Sepsis can cause small petechial hemorrhages throughout the cerebral white matter. *Moyamoya disease* (Chap. 421), mainly an occlusive arterial disease that causes ischemic symptoms, may on occasion produce ICH, particularly in the young. Hemorrhages into the spinal cord are usually the result of an AVM, cavernous malformation, or metastatic tumor. *Epidural spinal hemorrhage* produces a rapidly evolving syndrome of spinal cord or nerve root compression (Chap. 434). Spinal hemorrhages usually present with sudden back pain and some manifestation of myelopathy.

Laboratory and Imaging Evaluation

Patients should have routine blood chemistries and hematologic studies. Specific attention to the platelet count and PT/PTT/INR is important to identify coagulopathy. CT imaging reliably detects acute focal hemorrhages in the supratentorial space. Rarely very small pontine or medullary hemorrhages may not be

well delineated because of motion and bone-induced artifact that obscure structures in the posterior fossa. After the first 2 weeks, x-ray attenuation values of clotted blood diminish until they become isodense with surrounding brain. Mass effect and edema may remain. In some cases, a surrounding rim of contrast enhancement appears after 2–4 weeks and may persist for months. MRI, although more sensitive for delineating posterior fossa lesions, is generally not necessary for primary diagnosis in most instances. Images of flowing blood on MRI scan may identify AVMs as the cause of the hemorrhage. MRI, CT angiography (CTA), and conventional x-ray angiography are used when the cause of intracranial hemorrhage is uncertain, particularly if the patient is young or not hypertensive and the hematoma is not in one of the usual sites for hypertensive hemorrhage. CTA or postcontrast CT imaging may reveal one or more small areas of enhancement within a hematoma; this “spot sign” is thought to represent ongoing bleeding. The presence of a spot sign is associated with an increased risk of hematoma expansion, increased mortality, and lower likelihood of favorable functional outcome. Because patients typically have focal neurologic signs and obtundation and often show signs of increased ICP, a lumbar puncture is generally unnecessary and should usually be avoided because it may induce cerebral herniation.

TREATMENT

TREATMENT

Intracerebral Hemorrhage

ACUTE MANAGEMENT

After immediate attention to blood pressure and airway protection (see above), focus can switch to medical and surgical management. Approximately 40% of patients with a hypertensive ICH die, but survivors can have a good to complete recovery. The ICH Score **(Table 421-2)** is a validated clinical grading scale that is useful for stratification of mortality risk and clinical outcome. However, a specific ICH clinical grading scale should not be used to precisely prognosticate outcome because of the concern of creating a self-fulfilling prophecy of poor outcome if early aggressive care is withheld. Any identified coagulopathy should be corrected as soon as possible. For patients taking vitamin K antagonists (VKAs), rapid correction of coagulopathy can be achieved by infusing prothrombin complex concentrates (PCCs), which can be administered quickly, with vitamin K administered concurrently. Fresh frozen plasma (FFP) is an alternative but since it requires larger fluid volumes and longer time to achieve adequate reversal than PCC, it is not recommended if PCC is available. Idarucizumab is a monoclonal antibody to dabigatran and the administration of two doses reverses the anticoagulation effect of dabigatran quickly. PCC may partially reverse the effects of oral factor Xa inhibitors and are reasonable to administer if available; targeted drugs to reverse Xa inhibitors are under clinical investigation. When ICH is associated with thrombocytopenia (platelet count <50,000/μL), transfusion of fresh platelets is indicated. A recent clinical trial of platelet transfusions in patients with ICH and without thrombocytopenia who are taking antiplatelet drugs suggested no benefit and possible harm.

Hematomas may expand for several hours following the initial hemorrhage, even in patients without coagulopathy. However, the precise mechanism is unclear. A phase 3 trial of treatment with recombinant

factor VIIa reduced hematoma expansion; however, clinical outcomes were not improved, so use of this drug is not recommended. Blood pressure lowering has been considered due to the theoretical risk of acutely elevated blood pressure on hematoma expansion, although recent clinical trials did not find a difference in hematoma expansion between the SBP targets of 140–180 mmHg.

Evacuation of supratentorial hematomas does not appear to improve outcome for most patients. The International Surgical Trial in Intracerebral Haemorrhage (STICH) randomized patients with supratentorial ICH to either early surgical evacuation or initial medical management. No benefit was found in the early surgery arm, although analysis was complicated by the fact that 26% of patients in the initial medical management group ultimately had surgery for neurologic deterioration. The follow-up study STICH-II found that surgery within 24 h of lobar, supratentorial hemorrhage did not improve overall outcome, but might have a role in select severely affected patients. Therefore, existing data do not support routine surgical evacuation of supratentorial hemorrhages in stable patients. However, many centers still consider surgery for patients deemed salvageable and who are experiencing progressive neurologic deterioration due to herniation. Surgical techniques continue to evolve, and minimally invasive endoscopic hematoma evacuation is currently being investigated in clinical trials.

For cerebellar hemorrhages, a neurosurgeon should be consulted immediately to assist with the evaluation; most cerebellar hematomas >3 cm in diameter will require surgical evacuation. If the patient is alert without focal brainstem signs and if the hematoma is <1 cm in diameter, surgical removal is usually unnecessary. Patients with hematomas between 1 and 3 cm require careful observation for signs of impaired consciousness, progressive hydrocephalus, and precipitous respiratory failure. Hydrocephalus due to cerebellar hematoma requires surgical evacuation and should not be treated solely with ventricular drainage.

Tissue surrounding hematomas is displaced and compressed but not necessarily infarcted. Hence, in survivors, major improvement commonly occurs as the hematoma is reabsorbed and the adjacent tissue regains its function. Thus, careful management of the patient during the acute phase of the hemorrhage can lead to considerable recovery.

Surprisingly, ICP is often normal even with large ICHs. However, if the hematoma causes marked midline shift of structures with consequent obtundation, coma, or hydrocephalus, osmotic agents can be instituted in preparation for placement of a ventriculostomy or parenchymal ICP monitor **(Chap. 301)**. Once ICP is recorded, CSF drainage (if available), osmotic therapy, and blood pressure management can be tailored to the individual patient to keep cerebral perfusion pressure (MAP minus ICP) at least 50–70 mmHg. For example, if ICP is found to be high, CSF can be drained from the ventricular space and osmotic therapy continued; persistent or progressive elevation in ICP may prompt surgical evacuation of the clot. Alternately, if ICP is normal or only mildly elevated, interventions such as osmotic therapy may be tapered. Because hyperventilation may actually produce ischemia by cerebral vasoconstriction, induced hyperventilation should be limited to acute resuscitation of the patient with presumptive high ICP and eliminated once other treatments (osmotic therapy or surgical treatments) have been instituted. Glucocorticoids are not helpful for the edema from intracerebral hematoma.

PREVENTION

Hypertension is the leading cause of primary ICH. Prevention is aimed at reducing chronic hypertension, eliminating excessive alcohol use, and discontinuing use of illicit drugs such as cocaine and amphetamines. Current guidelines recommend that patients with CAA should generally avoid oral anticoagulant medications, but antiplatelet agents may be administered if there is an indication based on atherothrombotic vascular disease.

TABLE 421-2

The ICH Score

Clinical or Imaging Factor	Point Score
Age	
<80 years	0
≥80 years	1
Hematoma Volume	
<30 cc	0
≥30 cc	1
Intraventricular Hemorrhage Present	
No	0
Yes	1
Infratentorial Origin of Hemorrhage	
No	0
Yes	1
Glasgow Coma Scale Score	
13–15	0
5–12	1
3–4	2
Total Score	0–6 Sum of each category above

Source: JC Hemphill et al: Stroke 32:891, 2001.

VASCULAR ANOMALIES

Vascular anomalies can be divided into congenital vascular malformations and acquired vascular lesions.

CONGENITAL VASCULAR MALFORMATIONS

True AVMs, venous anomalies, and capillary telangiectasias are lesions that usually remain clinically silent through life. AVMs are probably congenital, but cases of acquired lesions have been reported.

True AVMs are congenital shunts between the arterial and venous systems that may present with headache, seizures, and intracranial hemorrhage. AVMs consist of a tangle of abnormal vessels across the cortical surface or deep within the brain substance. AVMs vary in size from a small blemish a few millimeters in diameter to a large mass of tortuous channels composing an arteriovenous shunt of sufficient magnitude to raise cardiac output and precipitate heart failure. Blood vessels forming the tangle interposed between arteries and veins are usually abnormally thin and histologically resemble both arteries and veins. AVMs occur in all parts of the cerebral hemispheres, brainstem, and spinal cord, but the largest ones are most frequently located in the posterior half of the hemispheres, commonly forming a wedge-shaped lesion extending from the cortex to the ventricle.

Bleeding, headache, and seizures are most common between the ages of 10 and 30, occasionally as late as the fifties. AVMs are more frequent in men, and rare familial cases have been described. Familial AVM may be a part of the autosomal dominant syndrome of hereditary hemorrhagic telangiectasia (Osler-Rendu-Weber) syndrome due to mutations in either endoglin or activin receptor-like kinase 1, both involved in transforming growth factor (TGF) signaling and angiogenesis.

Headache (without bleeding) may be hemicranial and throbbing, like migraine, or diffuse. Focal seizures, with or without generalization, occur in ~30% of cases. One-half of AVMs become evident as ICHs. In most, the hemorrhage is mainly intraparenchymal with extension into the subarachnoid space in some cases. Unlike primary subarachnoid hemorrhages (Chap. 302), blood from a ruptured AVM is usually not deposited in the basal cisterns, and symptomatic cerebral vasospasm is rare. The risk of AVM rupture is strongly influenced by a history of prior rupture. Although unruptured AVMs have a hemorrhage rate of ~2–4% per year, previously ruptured AVMs may have a rate as high as 17% a year, at least for the first year. Hemorrhages may be massive, leading to death, or may be as small as 1 cm in diameter, leading to minor focal symptoms or no deficit. The AVM may be large enough to steal blood away from adjacent normal brain tissue or to increase venous pressure significantly to produce venous ischemia locally and in remote areas of the brain. This is seen most often with large AVMs in the territory of the middle cerebral artery.

Large AVMs of the anterior circulation may be associated with a systolic and diastolic bruit (sometimes self-audible) over the eye, forehead, or neck and a bounding carotid pulse. Headache at the onset of AVM rupture is generally not as explosive as with aneurysmal rupture. MRI is better than CT for diagnosis, although noncontrast CT scanning sometimes detects calcification of the AVM and contrast may demonstrate the

abnormal blood vessels. Once identified, conventional x-ray angiography is the gold standard for evaluating the precise anatomy of the AVM.

Surgical treatment of AVMs presenting with hemorrhage often done in conjunction with preoperative embolization to reduce operative bleeding is usually indicated for accessible lesions. Stereotactic radiosurgery, an alternative to conventional surgery, can produce a slow sclerosis of the AVM over 2–3 years.

Several angiographic features can be used to help predict future bleeding risk. Paradoxically, smaller lesions seem to have a higher hemorrhage rate. The presence of deep venous drainage, venous outflow stenosis, and intranidal aneurysms may increase rupture risk. Because of the relatively low annual rate of hemorrhage and the risk of complications due to surgical or endovascular treatment, the indication for surgery in asymptomatic AVMs is debated. The ARUBA (A Randomized Trial of Unruptured Brain Arteriovenous Malformations) trial randomized patients to medical management versus intervention (surgery, endovascular embolization, combination embolization and surgery, or gamma-knife). The trial was stopped prematurely for harm, with the medical arm achieving the combined endpoint of death or symptomatic stroke in 10.1% of patients compared to 30.7% in the intervention group at an average follow-up time of 33 months. This highly significant finding argues against routine intervention for patients presenting without hemorrhage, although debate ensues regarding the generalizability of these results.

Venous anomalies are the result of development of anomalous cerebral, cerebellar, or brainstem venous drainage. These structures, unlike AVMs, are functional venous channels. They are of little clinical significance and should be ignored if found incidentally on brain imaging studies. Surgical resection of these anomalies may result in venous infarction and hemorrhage. Venous anomalies may be associated with cavernous malformations (see below), which do carry some bleeding risk.

Capillary telangiectasias are true capillary malformations that often form extensive vascular networks through an otherwise normal brain structure. The pons and deep cerebral white matter are typical locations, and these capillary malformations can be seen in patients with hereditary hemorrhagic telangiectasia (Osler-Rendu-Weber) syndrome. If bleeding does occur, it rarely produces mass effect or significant symptoms. No treatment options exist.

ACQUIRED VASCULAR LESIONS

Cavernous angiomas are tufts of capillary sinusoids that form within the deep hemispheric white matter and brainstem with no normal intervening neural structures. The pathogenesis is unclear. Familial cavernous angiomas have been mapped to several different genes: *KRIT1, CCM2*, and *PDCD10*. Both *KRIT1* and *CCM2* have roles in blood vessel formation, whereas *PDCD10* is an apoptotic gene. Cavernous angiomas are typically <1 cm in diameter and are often associated with a venous anomaly. Bleeding is usually of small volume, causing slight mass effect only. The bleeding risk for single cavernous malformations is 0.7–1.5% per year and may be higher for patients with prior clinical hemorrhage or multiple malformations. Seizures may occur if the malformation is located near the cerebral cortex. Surgical resection eliminates bleeding risk and

may reduce seizure risk, but it is usually reserved for those malformations that form near the brain surface. Radiation treatment has not been shown to be of benefit.

Dural arteriovenous fistulas are acquired connections usually from a dural artery to a dural sinus. Patients may complain of a pulse-synchronous cephalic bruit ("pulsatile tinnitus") and headache. Depending on the magnitude of the shunt, venous pressures may rise high enough to cause cortical ischemia or venous hypertension and hemorrhage, particularly SAH. Surgical and endovascular techniques are usually curative. These fistulas may form because of trauma, but most are idiopathic. There is an association between fistulas and dural sinus thrombosis. Fistulas have been observed to appear months to years following venous sinus thrombosis, suggesting that angiogenesis factors elaborated from the thrombotic process may cause these anomalous connections to form. Alternatively, dural arteriovenous fistulas can produce venous sinus occlusion over time, perhaps from the high pressure and high flow through a venous structure.

FURTHER READING

Anderson CS et al: Rapid blood-pressure lowering in patients with acute intracerebral hemorrhage. N Engl J Med 368:2355, 2013. [PubMed: 23713578]

Hemphill JC et al: Guidelines for the management of spontaneous intracerebral hemorrhage: A guideline for healthcare professionals from the American heart association/American stroke association. Stroke 46:2032, 2015. [PubMed: 26022637]

Mohr JP et al: Medical management with or without interventional therapy for unruptured brain arteriovenous malformations (ARUBA): A multicentre, non-blinded, randomised trial. Lancet 383:614, 2014. [PubMed: 24268105]

Qureshi AI et al: Intracerebral haemorrhage. Lancet 373:1632, 2009. [PubMed: 19427958]

Steiner T et al: European Stroke Organisation (ESO) guidelines for the management of spontaneous intracerebral hemorrhage. Int J Stroke 9:840, 2014. [PubMed: 25156220]

Chapter 422: Migraine and Other Primary Headache Disorders

Peter J. Goadsby

INTRODUCTION

The general approach to headache as a cardinal symptom are covered elsewhere (Chap. 13); here, disorders in which headache and associated features occur in the absence of any exogenous cause are discussed. The most common are migraine, tension-type headache (TTH), and the trigeminal autonomic cephalalgias (TACs), notably cluster headache; the complete list is summarized in Table 422-1.

TABLE 422-1

Primary Headache Disorders, Modified from International Classification of Headache Disorders-III-Beta (Headache Classification Committee of the International Headache Society, 2018)

1. Migraine	1.1 Migraine without aura 1.2 Migraine with aura 1.2.1 Migraine with typical aura 1.2.1.1 Typical aura with headache 1.2.1.2 Typical aura without headache 1.2.2 Migraine with brainstem aura 1.2.3 Hemiplegic migraine 1.2.3.1 Familial hemiplegic migraine (FHM) 1.2.3.1.1 Familial hemiplegic migraine type 1 1.2.3.1.2 Familial hemiplegic migraine type 2 1.2.3.1.3 Familial hemiplegic migraine type 3 1.2.3.1.4 Familial hemiplegic migraine, other loci 1.2.3.2 Sporadic hemiplegic migraine 1.2.4 Retinal migraine 1.3 Chronic migraine 1.4 Complications of migraine 1.4.1 Status migrainosus 1.4.2 Persistent aura without infarction 1.4.3 Migrainous infarction 1.4.4 Migraine aura-triggered seizure 1.5 Probable migraine 1.5.1 Probable migraine without aura 1.5.2 Probable migraine with aura 1.6 Episodic syndromes that may be associated with migraine 1.6.1 Recurrent gastrointestinal disturbance 1.6.1.1 Cyclical vomiting syndrome 1.6.1.2 Abdominal migraine 1.6.2 Benign paroxysmal vertigo 1.6.3 Benign paroxysmal torticollis
2. Tension-type headache	2.1 Infrequent episodic tension-type headache 2.2 Frequent episodic tension-type headache 2.3 Chronic tension-type headache 2.4 Probable tension-type headache
3. Trigeminal autonomic cephalalgias	3.1 Cluster headache 3.1.1 Episodic cluster headache 3.1.2 Chronic cluster headache 3.2 Paroxysmal hemicrania 3.2.1 Episodic paroxysmal hemicrania 3.2.2 Chronic paroxysmal hemicrania 3.3 Short-lasting unilateral neuralgiform headache attacks 3.3.1 Short-lasting unilateral neuralgiform headache attacks with conjunctival injection and tearing (SUNCT)

	3.3.1.1 Episodic SUNCT 3.3.1.2 Chronic SUNCT 3.3.2 Short-lasting unilateral neuralgiform headache attacks with cranial autonomic symptoms (SUNA) 3.3.2.1 Episodic SUNA 3.3.2.2 Chronic SUNA 3.4 Hemicrania continua 3.5 Probable trigeminal autonomic cephalalgia
4. Other primary headache disorders	4.1 Primary cough headache 4.2 Primary exercise headache 4.3 Primary headache associated with sexual activity 4.4 Primary thunderclap headache 4.5 Cold-stimulus headache 4.5.1 Headache attributed to external application of a cold stimulus 4.5.2 Headache attributed to ingestion or inhalation of a cold stimulus 4.6 External-pressure headache 4.6.1 External-compression headache 4.6.2 External-traction headache 4.7 Primary stabbing headache 4.8 Nummular headache 4.9 Hypnic headache 4.10 New daily persistent headache (NDPH)

MIGRAINE

Migraine, the second most common cause of headache, and the most common headache-related, and indeed neurologic, cause of disability in the world, afflicts ~15% of women and 6% of men over a 1-year period. It is usually an episodic headache associated with certain features such as sensitivity to light, sound, or movement; nausea and vomiting often accompany the headache. A useful description of migraine is a recurring syndrome of headache associated with other symptoms of neurologic dysfunction in varying admixtures **(Table 422-2)**. A migraine attack has three phases: premonitory (prodrome), headache phase, and postdrome; each has distinct and sometimes disabling symptoms. About 20–25% of migraine patients have a fourth, aura, phase. Migraine can often be recognized by its activators, referred to as *triggers*.

TABLE 422-2

Symptoms Accompanying Severe Migraine Attacks in 500 Patients

Symptom	Patients Affected, %
Nausea	87
Photophobia	82
Lightheadedness	72
Scalp tenderness	65
Vomiting	56
Visual disturbances	36
Paresthesias	33
Vertigo	33
Photopsia	26
Alteration of consciousness	18
Diarrhea	16
Fortification spectra	10
Syncope	10
Seizure	4
Confusional state	4

Source: From NH Raskin: *Headache*, 2nd ed. New York, Churchill Livingston, 1988; with permission.

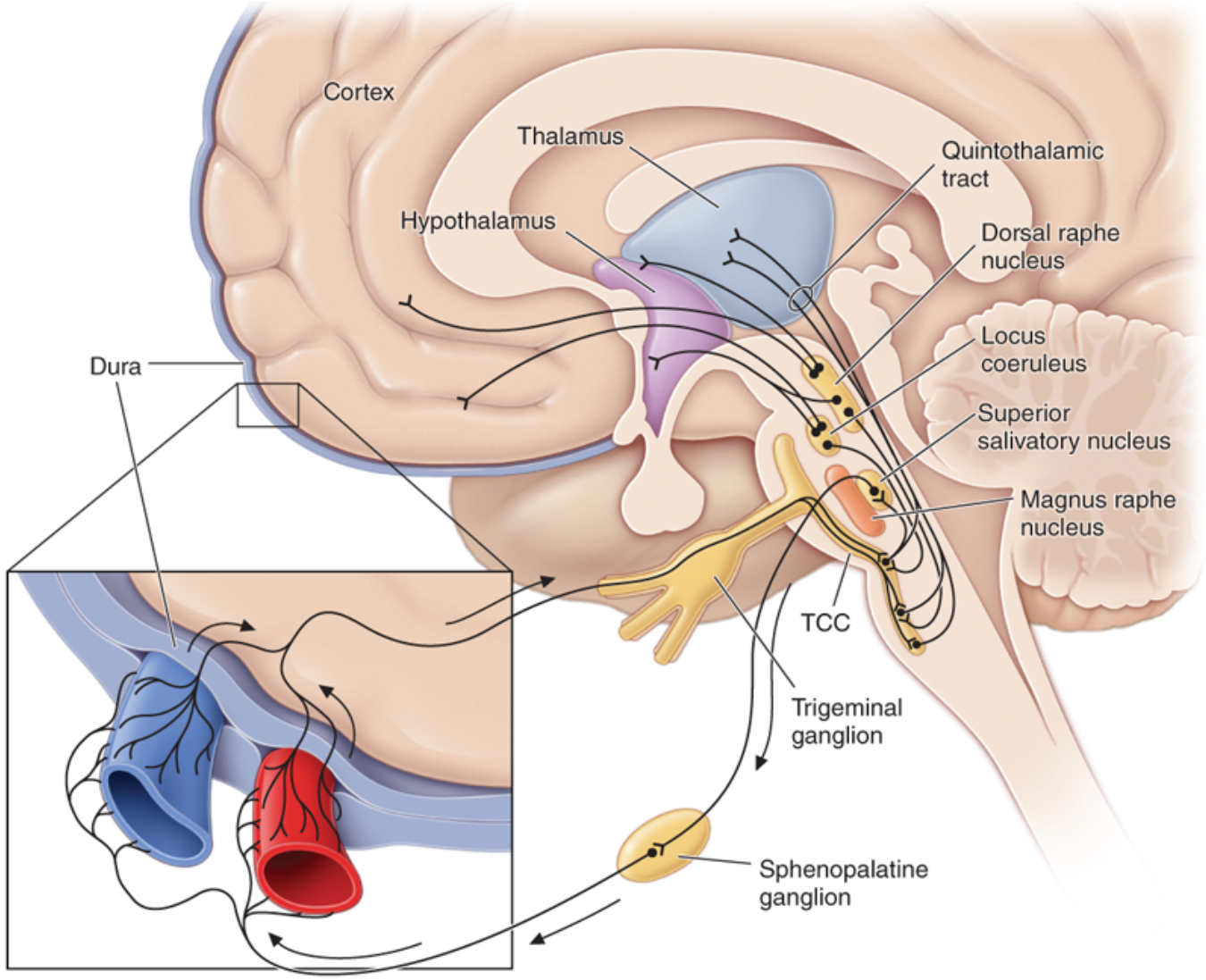
Migraineurs are particularly sensitive to environmental and sensory stimuli; migraine-prone patients do not habituate easily to sensory stimuli. This sensitivity is amplified in females during the menstrual cycle. Headache can be initiated or amplified by various triggers, including glare, bright lights, sounds, or other types of afferent stimulation; hunger; let-down from stress; physical exertion; stormy weather or barometric pressure changes; hormonal fluctuations during menses; lack of or excess sleep; and alcohol or other chemical stimulation, such as with nitrates. Knowledge of a patient's susceptibility to specific triggers can be useful in management strategies involving lifestyle adjustments, although it is becoming recognized that some apparent triggers may in fact be part of the initial phase of the attack; i.e., the premonitory phase or prodrome.

Pathogenesis

The sensory sensitivity that is characteristic of migraine is probably due to dysfunction of monoaminergic sensory control systems located in the brainstem and hypothalamus **(Fig. 422-1)**.

FIGURE 422-1

Brainstem pathways that modulate sensory input. The key pathway for pain in migraine is the trigeminovascular input from the meningeal vessels, which passes through the trigeminal ganglion and synapses on second-order neurons in the trigeminocervical complex (TCC). These neurons in turn project in the quintothalamic tract and, after decussating in the brainstem, synapse on neurons in the thalamus. Important modulation of the trigeminovascular nociceptive input comes from the dorsal raphe nucleus, locus coeruleus, and nucleus raphe magnus.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Activation of cells in the trigeminal nucleus results in the release of vasoactive neuropeptides, particularly calcitonin gene–related peptide (CGRP), at vascular terminals of the trigeminal nerve and within the trigeminal nucleus. Six CGRP receptor antagonists, *gepants*, have now been shown to be effective in the acute treatment of migraine, and four monoclonal antibodies to CGRP or its receptor have been shown to be effective in migraine prevention. Centrally, the second-order trigeminal neurons cross the midline and project to ventrobasal and posterior nuclei of the thalamus for further processing. Additionally, there are projections to the periaqueductal gray and hypothalamus, from which reciprocal descending systems have established antinociceptive effects. Other brainstem regions likely to be involved in descending modulation of trigeminal pain include the nucleus locus coeruleus in the pons and the rostroventromedial medulla.

Pharmacologic and other data point to the involvement of the neurotransmitter 5-hydroxytryptamine (5-HT; also known as serotonin) in migraine. In the late 1950s methysergide was found to antagonize certain peripheral actions of 5-HT and was introduced, based on its anti-inflammation properties, as the first drug capable of preventing migraine attacks. The *triptans* were

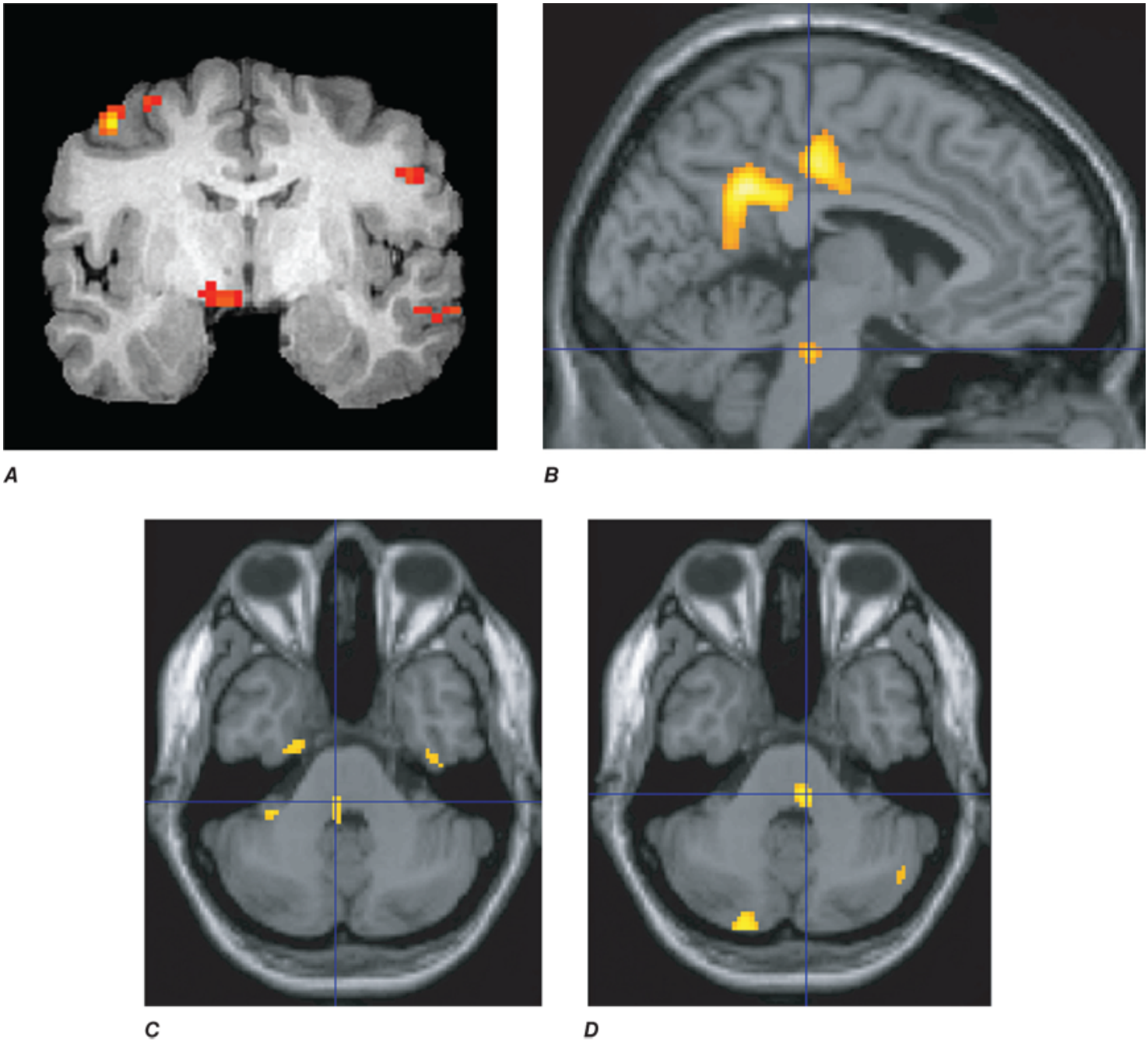
designed to stimulate selectively subpopulations of 5-HT receptors; at least 14 different 5-HT receptors exist in humans. The triptans are potent agonists of 5-HT_{1B} and 5-HT_{1D} receptors, and some are active at the 5-HT_{1F} receptor; the latter's exclusive agonists are called *ditans*. Triptans arrest nerve signaling in the nociceptive pathways of the trigeminovascular system, at least in the trigeminal nucleus caudalis and trigeminal sensory thalamus, in addition to promoting cranial vasoconstriction, while *ditans*, now shown conclusively to be effective in acute migraine, act only at neural and not vascular targets. A range of neural targets are currently under investigation for the acute and preventive management of migraine.

Data also support a role for dopamine in the pathophysiology of migraine. Most migraine symptoms can be induced by dopaminergic stimulation. Moreover, there is dopamine receptor hypersensitivity in migraineurs, as demonstrated by the induction of yawning, nausea, vomiting, hypotension, and other symptoms of a migraine attack by dopaminergic agonists at doses that do not affect nonmigraineurs. Dopamine receptor antagonists are effective therapeutic agents in migraine, especially when given parenterally or concurrently with other antimigraine agents. Moreover, hypothalamic activation, anterior to that seen in cluster headache, has now been shown in the premonitory (prodromal) phase of migraine using functional imaging, and this may hold a key to understanding some part of the role of dopamine in the disorder.

Migraine genes identified by studying families with familial hemiplegic migraine (FHM) reveal involvement of ion channels, suggesting that alterations in membrane excitability can predispose to migraine. Mutations involving the $Ca_V2.1$ (P/Q)–type voltage-gated calcium channel *CACNA1A* gene are now known to cause FHM 1; this mutation is responsible for about 50% of FHM cases. Mutations in the Na^+-K^+ATPase *ATP1A2* gene, designated FHM 2, are responsible for about 20% of FHMs. Mutations in the neuronal voltage-gated sodium channel *SCN1A* cause FHM 3. Functional neuroimaging has suggested that brainstem regions in migraine **(Fig. 422-2)** and the posterior hypothalamic gray matter region close to the human circadian pacemaker cells of the suprachiasmatic nucleus in cluster headache **(Fig. 422-3)** are good candidates for specific involvement in these primary headaches.

FIGURE 422-2

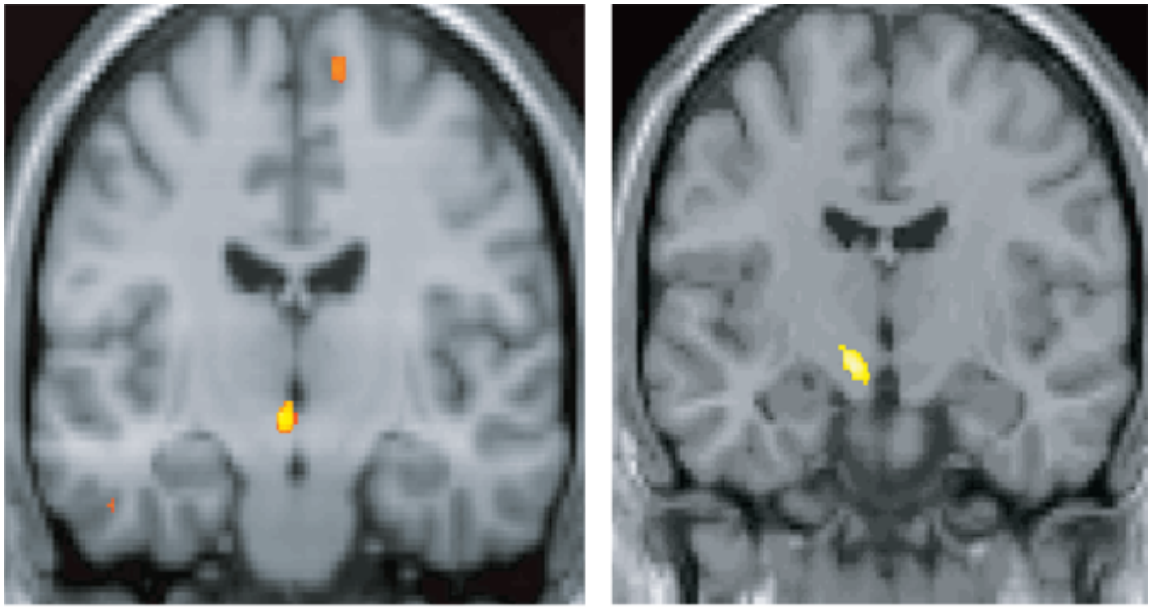
Positron emission tomography (PET) activation in migraine. Hypothalamic, dorsal midbrain, and dorsolateral pontine activation is seen in triggered attacks in the premonitory phase before pain, whereas in migraine attacks, dorsolateral pontine activation persists, as it does in chronic migraine (not shown). The dorsolateral pontine area, which includes the noradrenergic locus coeruleus, is fundamental to the expression of migraine. Moreover, lateralization of changes in this region of the brainstem correlates with lateralization of the head pain in hemicranial migraine; the scans shown in panels ***C*** and ***D*** are of patients with acute migraine headache on the right and left side, respectively. *(Panel A from FH Maniyar et al: Brain 137:232, 2014; panel B from SK Afridi et al: Arch Neurol 62:1270, 2005; Panels C and D from SK Afridi et al: Brain 128:932, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 422-3

A. Posterior hypothalamic gray matter region activation by positron emission tomography in a patient with acute cluster headache. *(From A May et al: Lancet 352:275, 1998.)* ***B.*** High-resolution T1-weighted magnetic resonance image obtained using voxel-based morphometry demonstrates increased gray matter activity, lateralized to the side of pain in a patient with cluster headache. *(From A May et al: Nat Med 5:836, 1999.)*



A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Diagnosis and Clinical Features

Diagnostic criteria for migraine headache are listed in **Table 422-3**. A high index of suspicion is required to diagnose migraine: the migraine aura, consisting of visual disturbances with flashing lights or zigzag lines moving across the visual field or of other neurologic symptoms, is reported in only 20–25% of patients. It should be distinguished from the pan-field television static-like disturbance now recognized as the *visual snow syndrome*. The first phase of a migraine attack for most patients is the premonitory (prodromal) phase consisting of some or all of the following: yawning, tiredness, cognitive dysfunction, mood change, neck discomfort, polyuria, and food cravings; this can last from a few hours to days. Typically, the headache phase follows with its associated features, such as nausea, photophobia, and phonophobia as well as allodynia. When questioned, these typical migraine symptoms also emerge in the premonitory phase, and typical premonitory symptoms also continue into the headache phase. As the headache lessens many patients enter a postdrome, most commonly feeling tired/weary, having problems concentrating, and experiencing mild neck discomfort that can last for hours and sometimes up to a day. A headache diary can often be helpful in making the diagnosis; this is also helpful in assessing disability and the frequency of treatment for acute attacks. Patients with episodes of migraine on eight or more days per month and with at least 15 total days of headache per month are considered to have chronic migraine **(see "Chronic Daily Headache" in Chap. 13)**. Migraine must be differentiated from TTH (discussed below), which is reported to be the most common primary headache syndrome. Migraine has several forms that have been defined (Table 422-1): migraine with and without aura and chronic migraine are the most important. *Migraine at its most basic level is headache with associated features, and tension-type headache is headache that is featureless. Most patients with disabling headache probably have migraine.*

TABLE 422-3

Simplified Diagnostic Criteria for Migraine

Repeated attacks of headache lasting 4–72 h in patients with a normal physical examination, no other reasonable cause for the headache, and:	
At Least 2 of the Following Features:	**Plus at Least 1 of the Following Features:**
Unilateral pain Throbbing pain Aggravation by movement Moderate or severe intensity	Nausea/vomiting Photophobia and phonophobia

Source: Adapted from the International Headache Society Classification (Headache Classification Committee of the International Headache Society, Cephalalgia 38:1-211, 2018).

Patients with acephalgic migraine (typical aura without headache, 1.2.1.2 in Table 422-1) experience recurrent neurologic symptoms, often with nausea or vomiting, but with little or no headache. Vertigo can be prominent; it has been estimated that one-third of patients referred for vertigo or dizziness have a primary diagnosis of migraine. Migraine aura can have prominent brainstem symptoms, and the terms *basilar artery* and *basilar-type migraine* have now been replaced by *migraine with brainstem aura* (Table 422-1).

TREATMENT

TREATMENT

Migraine Headache

Once a diagnosis of migraine has been established, it is important to assess the extent of a patient's disease and disability. The Migraine Disability Assessment Score (MIDAS) is a well-validated, easy-to-use tool **(Fig. 422-4)**.

Patient education is an important aspect of migraine management. Information for patients is available at websites such as the American Migraine Foundation (www.americanmigrainefoundation.org) and the Migraine Trust (www.migrainetrust.org). It is helpful for patients to understand that migraine is an inherited tendency to headache; that migraine can be modified and controlled by lifestyle adjustments and medications, but it cannot be eradicated; and that, except on some occasions in women on oral estrogens or contraceptives, migraine is not associated with serious or life-threatening illnesses.

NONPHARMACOLOGIC MANAGEMENT

Migraine can often be managed to some degree by a variety of nonpharmacologic approaches. When patients can identify reliable triggers, their avoidance can be useful. A regulated lifestyle is helpful, including a healthy diet, regular exercise, regular sleep patterns, avoidance of excess caffeine and alcohol, and avoidance of acute changes in stress levels, being particularly wary of the let-down effect.

The measures that benefit a given individual should be used routinely because they provide a simple, cost-effective approach to migraine management. Patients with migraine do not encounter more stress than headache-free individuals; over-responsiveness to changes in stress appears to be the issue. Because the stresses of everyday living cannot be eliminated, lessening one's response to stress by various techniques is helpful for many patients. These may include yoga, transcendental meditation, hypnosis, and conditioning techniques such as biofeedback. For most patients seen in clinical practice, this approach

is, at best, an adjunct to pharmacotherapy. Nonpharmacologic measures are unlikely to prevent all migraine attacks. If these measures fail to prevent an attack, pharmacologic approaches are then needed.

ACUTE ATTACK THERAPIES FOR MIGRAINE

The mainstay of pharmacologic therapy is the judicious use of one or more of the many medicines that are effective in migraine **(Table 422-4)**. The selection of the optimal regimen for a given patient depends on a number of factors, the most important of which is the severity of the attack. Mild migraine attacks can usually be managed by oral agents; the average efficacy rate is 50–70%. Severe migraine attacks may require parenteral therapy. Most drugs effective in the treatment of migraine are members of one of three major pharmacologic classes: nonsteroidal anti-inflammatory drugs, 5-$HT_{1B/1D}$ receptor agonists, and dopamine receptor antagonists. Two new classes of therapeutic agents, CGRP receptor antagonists, such as rimegepant and ubrogepant, and 5-HT_{1F} receptor agonists, such as lasmiditan, should soon be available.

In general, an adequate dose of whichever agent is chosen should be used as soon as possible after the onset of an attack. If additional medication is required within 60 min because symptoms return or have not abated, the initial dose should be increased for subsequent attacks or a different class of drug tried as first-line treatment. Migraine therapy must be individualized; a standard approach for all patients is not possible. A therapeutic regimen may need to be constantly refined until one is identified that provides the patient with rapid, complete, and consistent relief with minimal side effects **(Table 422-5)**.

Nonsteroidal Anti-Inflammatory Drugs (NSAIDs)

Both the severity and duration of a migraine attack can be reduced significantly by NSAIDs (Table 422-4). Indeed, many undiagnosed migraineurs self-treat with nonprescription NSAIDs. A general consensus is that NSAIDs are most effective when taken early in the migraine attack. However, the effectiveness of these agents in migraine is usually less than optimal in moderate or severe migraine attacks. The combination of acetaminophen (paracetamol), aspirin, and caffeine has been approved for use by the U.S. Food and Drug Administration (FDA) for the treatment of mild to moderate migraine. The combination of aspirin and metoclopramide has been shown to be comparable to a single dose of oral sumatriptan. Important side effects of NSAIDs include dyspepsia and gastrointestinal irritation.

5-$HT_{1B/1D}$ RECEPTOR AGONISTS

Oral

Stimulation of 5-$HT_{1B/1D}$ receptors can stop an acute migraine attack. Ergotamine and dihydroergotamine are nonselective receptor agonists, whereas the triptans are selective 5-$HT_{1B/1D}$ receptor agonists. A variety of triptans—sumatriptan, almotriptan, eletriptan, frovatriptan, naratriptan, rizatriptan, and zolmitriptan—are available for the treatment of migraine.

Each drug in the triptan class has similar pharmacologic properties but varies slightly in terms of clinical efficacy. Rizatriptan and eletriptan are, on a population basis, the most efficacious of the triptans currently available in the United States. Sumatriptan and zolmitriptan have similar rates of efficacy as well as time to onset, with an advantage of having multiple formulations, whereas almotriptan has a similar rate of efficacy to sumatriptan and is better tolerated, and frovatriptan and naratriptan are somewhat slower in onset and are also well tolerated. Clinical efficacy appears to be related more to the t_{max} (time to peak plasma level) than to the potency, half-life, or bioavailability. This observation is consistent with a large body of data indicating that faster-acting analgesics are more effective than slower-acting agents.

Unfortunately, monotherapy with a selective oral 5-$HT_{1B/1D}$ receptor agonist does not result in rapid, consistent, and complete relief of migraine in all patients. Triptans are generally not effective in migraine with aura unless given after the aura is completed and the headache initiated. Side effects are common, although often mild and transient. Moreover, 5-$HT_{1B/1D}$ receptor agonists are contraindicated in individuals with a history of cardiovascular and cerebrovascular disease. Recurrence of headache, within usual time course of an attack, is another important limitation of triptan use and occurs at least occasionally in most patients. Evidence from randomized controlled trials shows that coadministration of a longer-acting NSAID, naproxen 500 mg, with sumatriptan will augment the initial effect of sumatriptan and, importantly, reduce rates of headache recurrence.

Ergotamine preparations offer a nonselective means of stimulating 5-HT_1 receptors. A nonnauseating dose of ergotamine should be sought because a dose that provokes nausea is too high and may intensify head pain. Oral (excluding sublingual) formulations of ergotamine also contain 100 mg caffeine (theoretically to enhance ergotamine absorption and possibly to add additional analgesic activity). The average oral ergotamine dose for a migraine attack is 2 mg. Because the clinical studies demonstrating the efficacy of ergotamine in migraine predated the clinical trial methodologies used with the triptans, it is difficult to assess the comparative efficacy of ergotamine versus the triptans. In general, with use of ergotamine there appears to be a much higher incidence of nausea than with triptans but less headache recurrence.

Nasal

Nasal formulations of dihydroergotamine, zolmitriptan, or sumatriptan can be useful in patients requiring a nonoral route of administration. The nasal sprays result in substantial blood levels within 30–60 min. Although in theory nasal sprays might provide faster and more effective relief of a migraine attack than oral formulations, their reported efficacy is only ~50–60%. Studies with a new inhalational formulation of dihydroergotamine indicate that its absorption problems can be overcome to produce rapid onset of action with good tolerability.

Parenteral

Administration of drugs by injection, such as dihydroergotamine and sumatriptan, is approved by the FDA for the rapid relief of a migraine attack. Peak plasma levels of dihydroergotamine are achieved 3 min after IV dosing, 30 min after IM dosing, and 45 min after SC dosing. If an attack has not already peaked, SC or IM administration of 1 mg of dihydroergotamine is adequate for about 80–90% of patients. Sumatriptan, 4–6 mg SC, is effective in ~50–80% of patients, and can now be administered by a needle-free device.

DOPAMINE RECEPTOR ANTAGONISTS

Oral

Oral dopamine receptor antagonists can be considered as adjunctive therapy in migraine. Drug absorption is impaired during migraine because of reduced gastrointestinal motility. Delayed absorption occurs even in the absence of nausea and is related to the severity of the attack and not its duration. Therefore, when oral NSAIDs and/or triptan agents fail, the addition of a dopamine receptor antagonist, such as metoclopramide 10 mg or domperidone 10 mg (not available in the United States), should be considered to enhance gastric absorption. In addition, dopamine receptor antagonists decrease nausea/vomiting and restore normal gastric motility.

Parenteral

Dopamine receptor antagonists (e.g., chlorpromazine, prochlorperazine, metoclopramide) by injection can also provide significant acute relief of migraine; they can be used in combination with parenteral 5-$HT_{1B/1D}$ receptor agonists. A common IV protocol used for the treatment of severe migraine is the administration over 2 min of a mixture of 5 mg of prochlorperazine and 0.5 mg of dihydroergotamine.

OTHER OPTIONS FOR ACUTE MIGRAINE

Oral

The combination of acetaminophen, dichloralphenazone, and isometheptene, one to two capsules, has been classified by the FDA as “possibly” effective in the treatment of migraine. Because the clinical studies demonstrating the efficacy of this combination analgesic in migraine predated the clinical trial methodologies used with the triptans, it is difficult to compare the efficacy of this sympathomimetic compound to other agents.

Parenteral

Opioids are modestly effective in the acute treatment of migraine. For example, IV meperidine (50–100 mg) is given frequently in the emergency room. This regimen “works” in the sense that the pain of migraine is eliminated. Importantly, it is clear from a recent randomized controlled trial that prochlorperazine is superior to hydromorphone in the emergency room setting. However, opioids are clearly suboptimal for patients with recurrent headache. Opioids do not treat the underlying headache mechanism;

rather, they act to alter the pain sensation, and there is evidence their use may decrease the likelihood of a response to triptans in the future. Moreover, in patients taking oral opioids, such as oxycodone or hydrocodone, habituation or addiction can greatly confuse the treatment of migraine. Opioid craving and/or withdrawal can aggravate and accentuate migraine. Therefore, it is recommended that opioid use in migraine be limited to patients with severe, but infrequent, headaches that are unresponsive to other pharmacologic approaches or who have contraindications to other therapies.

Neuromodulation

Single pulse transcranial magnetic stimulation (sTMS) is FDA-approved for the acute treatment of migraine. Two pulses can be applied at the onset of an attack and this can be repeated. The use of sTMS is safe where there is no cranial metal implant, and offers an option to patients seeking non-pharmaceutical approaches to treatment. Similarly, a noninvasive vagus nerve stimulator (nVNS) is FDA-approved for the treatment of migraine attacks in adults. One to two 120-second doses may be applied for attack treatment.

MEDICATION-OVERUSE HEADACHE

Acute attack medications, particularly opioid or barbiturate-containing compound analgesics, have a propensity to aggravate headache frequency and induce a state of refractory daily or near-daily headache called *medication-overuse headache*. This condition is likely not a separate headache entity but a reaction of the migraine patient's biology to a particular medicine. Migraine patients who have two or more headache days a week should be cautioned about frequent analgesic use **(see "Chronic Daily Headache" in Chap. 13)**.

PREVENTIVE TREATMENTS FOR MIGRAINE

Patients with an increasing frequency of migraine attacks or with attacks that are either unresponsive or poorly responsive to abortive treatments are good candidates for preventive agents. In general, a preventive medication should be considered in patients with four or more attacks a month. Significant side effects are associated with the use of many of these agents; furthermore, determination of dose can be difficult because the recommended doses have been derived for conditions other than migraine. The mechanism of action of these drugs is unclear; it seems likely that the brain sensitivity that underlies migraine is modified. Patients are usually started on a low dose of a chosen treatment; the dose is then gradually increased, up to a reasonable maximum, to achieve clinical benefit.

Treatments that have the capacity to stabilize migraine are listed in **Table 422-6**. Most treatments must be taken daily, and there is usually a lag of between 2 and 12 weeks before an effect is seen. The drugs that have been approved by the FDA for the preventive treatment of migraine include propranolol, timolol, sodium valproate, and topiramate. In addition, a number of other drugs appear to display preventive efficacy. This group includes amitriptyline, nortriptyline, flunarizine, phenelzine, and cyproheptadine. Placebo-controlled trials of onabotulinum toxin type A in episodic migraine were negative, whereas, overall, placebo-controlled trials in chronic migraine were positive. The FDA has approved sTMS for the preventive treatment of migraine. It offers a well-tolerated, effective option for patients. Phenelzine is a monoamine oxidase inhibitor (MAOI); therefore, tyramine-containing foods, decongestants, and meperidine are contraindicated, and it is reserved for only very recalcitrant cases. Methysergide is now of historical interest only, since it is no longer manufactured. Melatonin has been reported to be useful, with controlled trial evidence but is not approved in the U.S. Monoclonal antibodies to the CGRP receptor (erenumab) or to the peptide (eptinezumab, fremanezumab, and galcanezumab) have all proven effective and well tolerated in migraine and should be available soon as preventive agents.

The probability of success with any one of the antimigraine drugs is 50%. Many patients are managed adequately with well-tolerated doses of candesartan, propranolol, amitriptyline, topiramate, or valproate. If these agents fail or produce unacceptable side effects, neuromodulation approaches, such as sTMS, or related agents from the above classes, can be used (Table 422-6). Once effective stabilization is achieved, the drug is continued for ~6 months and then slowly tapered to assess the continued need. Many patients are able to discontinue medication and experience fewer and milder attacks for long periods, suggesting that these drugs may alter the natural history of migraine.

FIGURE 422-4

The Migraine Disability Assessment Score (MIDAS) Questionnaire.

***MIDAS Questionnaire**

INSTRUCTIONS: Please answer the following questions about ALL headaches you have had over the last 3 months. Write zero if you did not do the activity in the last 3 months.

1. On how many days in the last 3 months did you miss work or school because of your headaches? ..... ____ days
2. How many days in the last 3 months was your productivity at work or school reduced by half or more because of your headaches *(do not include days you counted in question 1 where you missed work or school)*? ..... ____ days
3. On how many days in the last 3 months did you **not** do household work because of your headaches? ..... ____ days
4. How many days in the last 3 months was your productivity in household work reduced by half or more because of your headaches *(do not include days you counted in question 3 where you did not do household work)*? ..... ____ days
5. On how many days in the last 3 months did you miss family, social, or leisure activities because of your headaches? ..... ____ days

A. On how many days in the last 3 months did you have a headache? *(If a headache lasted more than one day, count each day.)* ..... ____ days

B. On a scale of 0–10, on average how painful were these headaches? *(Where 0 = no pain at all, and 10 = pain as bad as it can be.)* ..... ____

*Migraine Disability Assessment Score
(Questions 1–5 are used to calculate the MIDAS score.)
Grade I—Minimal or Infrequent Disability: 0–5
Grade II—Mild or Infrequent Disability: 6–10
Grade III—Moderate Disability: 11–20
Grade IV—Severe Disability: > 20

© Innovative Medical Research 1997

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 422-4

Treatment of Acute Migraine

Drug	Trade Name	Dosage
Simple Analgesics		
Acetaminophen, aspirin, caffeine	Excedrin Migraine	Two tablets or caplets q6h (max 8 per day)
NSAIDs		
Naproxen	Aleve, Anaprox, generic	220–550 mg PO bid
Ibuprofen	Advil, Motrin, Nuprin, generic	400 mg PO q3–4h
Tolfenamic acid	Clotam Rapid	200 mg PO; may repeat ×1 after 1–2 h
Diclofenac K	Cambia	50 mg PO with water
5-HT$_{1B/1D}$ Receptor Agonists		
Oral		
Ergotamine 1 mg, caffeine 100 mg	Cafergot	One or two tablets at onset, then one tablet q½h (max 6 per day, 10 per week)
Naratriptan	Amerge	2.5-mg tablet at onset; may repeat once after 4 h
Rizatriptan	Maxalt	5–10-mg tablet at onset; may repeat after 2 h (max 30 mg/d)
	Maxalt-MLT	
Sumatriptan	Imitrex	50–100-mg tablet at onset; may repeat after 2 h (max 200 mg/d)
Frovatriptan	Frova	2.5-mg tablet at onset, may repeat after 2 h (max 5 mg/d)
Almotriptan	Axert	12.5-mg tablet at onset, may repeat after 2 h (max 25 mg/d)
Eletriptan	Relpax	40 or 80 mg
Zolmitriptan	Zomig	2.5-mg tablet at onset; may repeat after 2 h (max 10 mg/d)
	Zomig Rapimelt	

Drug	Trade Name	Dosage
Nasal		
Dihydroergotamine	Migranal Nasal Spray	Prior to nasal spray, the pump must be primed 4 times; 1 spray (0.5 mg) is administered, followed in 15 min by a second spray
Sumatriptan	Imitrex Nasal Spray	5–20 mg intranasal spray as 4 sprays of 5 mg or a single 20 mg spray (may repeat once after 2 h, not to exceed a dose of 40 mg/d)
Zolmitriptan	Zomig	5 mg intranasal spray as one spray (may repeat once after 2 h, not to exceed a dose of 10 mg/d)
Parenteral		
Dihydroergotamine	DHE-45	1 mg IV, IM, or SC at onset and q1h (max 3 mg/d, 6 mg per week)
Sumatriptan	Imitrex Injection Alsuma Sumavel DosePro	6 mg SC at onset (may repeat once after 1 h for max of 2 doses in 24 h)
Dopamine Receptor Antagonists		
Oral		
Metoclopramide	Reglan,[a] generic[a]	5–10 mg/d
Prochlorperazine	Compazine,[a] generic[a]	1–25 mg/d
Parenteral		
Chlorpromazine	Generic[a]	0.1 mg/kg IV at 2 mg/min; max 35 mg/d
Metoclopramide	Reglan,[a] generic	10 mg IV
Prochlorperazine	Compazine,[a] generic[a]	10 mg IV
Other		
Oral		

Drug	Trade Name	Dosage
Acetaminophen, 325 mg, *plus* dichloralphenazone, 100 mg, *plus* isometheptene, 65 mg	Midrin, generic	Two capsules at onset followed by 1 capsule q1h (max 5 capsules)
Parenteral		
Opioids	Generic[a]	Multiple preparations and dosages; **see Table 10-1**
Other		
Neuromodulation	SpringTMS	Two pulses at onset followed by two further pulses
Noninvasive Vagus Nerve Stimulation (nVNS)	gammaCore	Two doses each of 120 seconds
Single pulse transcranial magnetic stimulation (sTMS)		

[a]Not all drugs are specifically indicated by the FDA for migraine. Local regulations and guidelines should be consulted.

Note: Antiemetics (e.g., domperidone 10 mg or ondansetron 4 or 8 mg) or prokinetics (e.g., metoclopramide 10 mg) are sometimes useful adjuncts.

Abbreviations: 5-HT, 5-hydroxytryptamine; NSAIDs, nonsteroidal anti-inflammatory drugs.

TABLE 422-5

Clinical Stratification of Acute Specific Migraine Treatments

Clinical Situation	Treatment Options
Failed NSAIDs/analgesics	**First tier** Sumatriptan 50 mg or 100 mg PO Almotriptan 12.5 mg PO Rizatriptan 10 mg PO Eletriptan 40 mg PO Zolmitriptan 2.5 mg PO **Slower effect/better tolerability** Naratriptan 2.5 mg PO Frovatriptan 2.5 mg PO **Infrequent headache** Ergotamine/caffeine 1–2/100 mg PO Dihydroergotamine nasal spray 2 mg
Early nausea or difficulties taking tablets	Zolmitriptan 5 mg nasal spray Sumatriptan 20 mg nasal spray Rizatriptan 10 mg MLT wafer
Headache recurrence	Ergotamine 2 mg (most effective PR/usually with caffeine) Naratriptan 2.5 mg PO Almotriptan 12.5 mg PO Eletriptan 40 mg
Tolerating acute treatments poorly	Naratriptan 2.5 mg Almotriptan 12.5 mg Single pulse transcranial magnetic stimulation Noninvasive vagus nerve stimulation
Early vomiting	Zolmitriptan 5 mg nasal spray Sumatriptan 25 mg PR Sumatriptan 6 mg SC
Menses-related headache	**Prevention** Ergotamine PO at night Estrogen patches **Treatment** Triptans Dihydroergotamine nasal spray
Very rapidly developing symptoms	Zolmitriptan 5 mg nasal spray Sumatriptan 6 mg SC Dihydroergotamine 1 mg IM

Abbreviation: NSAIDs, nonsteroidal anti-inflammatory drugs.

TABLE 422-6

Preventive Treatments in Migraine[a]

Drug	Dose	Selected Side Effects
Beta blocker		
Propranolol	40–120 mg bid	Reduced energy
Metoprolol	25–100 mg bid	Tiredness
		Postural symptoms
		Contraindicated in asthma
Antidepressants		
Amitriptyline	10–75 mg at night	Drowsiness
Dosulepin	25–75 mg at night	
Nortriptyline	25–75 mg at night	***Note:*** Some patients may only need a total dose of 10 mg, although generally 1–1.5 mg/kg body weight is required
Venlafaxine	75–150 mg/d	
Anticonvulsants		
Topiramate	25–200 mg/d	Paresthesias
		Cognitive symptoms
		Weight loss
		Glaucoma
		Caution with nephrolithiasis
Valproate	400–600 mg bid	Drowsiness
		Weight gain

Drug	Dose	Selected Side Effects
		Tremor
		Hair loss
		Fetal abnormalities
		Hematologic or liver abnormalities
Serotonergic drugs		
Pizotifen[b]	0.5–2 mg qd	Weight gain
		Drowsiness
Other classes		
Flunarizine[b]	5–15 mg qd	Drowsiness
		Weight gain
		Depression
		Parkinsonism
Candesartan	4–24 mg daily	Dizziness
Neuromodulation		
Single pulse transcranial magnetic stimulation (sTMS)	4–24 pulses per day	Lightheadedness
		Tingling
		Tinnitus
Chronic migraine		
Onabotulinum toxin type A	155 U	Loss of brow furrow
No convincing evidence from controlled trials		
Verapamil		
Controlled trials demonstrate *no effect*		
Nimodipine		
Clonidine		

Drug	Dose	Selected Side Effects
Selective serotonin reuptake inhibitors: fluoxetine		

[a]Commonly used preventives are listed with typical doses and common side effects. Not all listed medicines are approved by the U.S. Food and Drug Administration; local regulations and guidelines should be consulted. [b]Not available in the United States. [c]Not currently available worldwide.

TENSION-TYPE HEADACHE

Clinical Features

The term *tension-type headache* is commonly used to describe a chronic head-pain syndrome characterized by bilateral tight, band-like discomfort. The pain typically builds slowly, fluctuates in severity, and may persist more or less continuously for many days. The headache may be episodic or chronic (present >15 days per month).

A useful clinical approach is to diagnose TTH in patients whose headaches are completely without accompanying features such as nausea, vomiting, photophobia, phonophobia, osmophobia, throbbing, and aggravation with movement. Such an approach neatly separates migraine, which has one or more of these features and is the main differential diagnosis, from TTH. The International Headache Society's main definition of TTH allows an admixture of nausea, photophobia, or phonophobia in various combinations, although the appendix definition does not; this illustrates the difficulty in distinguishing these two clinical entities. In clinical practice using the appendix definition to dichotomize patients on the basis of the presence of associated features (migraine) and the absence of associated features (TTH) is highly recommended. Indeed patients whose headaches fit the TTH phenotype and who have migraine at other times, along with a family history of migraine, migrainous illnesses of childhood, or typical migraine triggers to their migraine attacks, may be biologically different from those who have TTH headache with none of the features. TTH may be infrequent (episodic) or occur on 15 days or more a month (chronic).

Pathophysiology

The pathophysiology of TTH is incompletely understood. It seems likely that TTH is due to a primary disorder of central nervous system pain modulation alone, unlike migraine, which involves a more generalized disturbance of sensory modulation. Data suggest a genetic contribution to TTH, but this may not be a valid finding: given the current diagnostic criteria, the studies undoubtedly included many migraine patients. The name *tension-type headache* implies that pain is a product of *nervous tension*, but there is no clear evidence for tension as an etiology. Muscle contraction has been considered to be a feature that distinguishes TTH from migraine, but there appear to be no differences in contraction between the two headache types.

TREATMENT

TREATMENT

Tension-Type Headache

The pain of TTH can generally be managed with simple analgesics such as acetaminophen, aspirin, or NSAIDs. Behavioral approaches including relaxation can also be effective. Clinical studies have demonstrated that triptans in pure TTH are not helpful, although triptans are effective in TTH when the patient also has migraine. For chronic TTH, amitriptyline is the only proven treatment (Table 422-6); other tricyclics, selective serotonin reuptake inhibitors, and the benzodiazepines have not been

shown to be effective. There is no evidence for the efficacy of acupuncture. Placebo-controlled trials of onabotulinum toxin type A in chronic TTH were negative.

TRIGEMINAL AUTONOMIC CEPHALALGIAS, INCLUDING CLUSTER HEADACHE

The TACs describe a grouping of primary headaches including cluster headache, paroxysmal hemicrania (PH), SUNCT (*s*hort-lasting *u*nilateral *n*euralgiform headache attacks with *c*onjunctival injection and *t*earing)/SUNA (*s*hort-lasting *u*nilateral *n*euralgiform headache attacks with cranial *a*utonomic symptoms), and hemicrania continua (Table 422-1). TACs are characterized by relatively short-lasting attacks of head pain associated with cranial autonomic symptoms, such as lacrimation, conjunctival injection, aural fullness, or nasal congestion **(Table 422-7)**. Pain is usually severe and may occur more than once a day. Because of the associated nasal congestion or rhinorrhea, patients are often misdiagnosed with “sinus headache” and treated with decongestants, which are ineffective.

TABLE 422-7

Clinical Features of the Trigeminal Autonomic Cephalalgias

	Cluster Headache	Paroxysmal Hemicrania	SUNCT/SUNA
Gender	M > F	F = M	F ~ M
Pain			
Type	Stabbing, boring	Throbbing, boring, stabbing	Burning, stabbing, sharp
Severity	Excruciating	Excruciating	Severe to excruciating
Site	Orbit, temple	Orbit, temple	Periorbital
Attack frequency	1/alternate day–8/d	1–20/d (>5/d for more than half the time)	3–200/d
Duration of attack	15–180 min	2–30 min	5–240 s
Autonomic features	Yes	Yes	Yes (prominent conjunctival injection and lacrimation)[a]
Migrainous features[b]	Yes	Yes	Yes
Alcohol trigger	Yes	No	No
Cutaneous triggers	No	No	Yes
Indomethacin effect	—	Yes[c]	—
Abortive treatment	Sumatriptan injection or nasal spray	No effective treatment	Lidocaine (IV)
	Oxygen		
	nVNS[c]		
Prophylactic treatment	Verapamil	Indomethacin[d]	Lamotrigine
	Topiramate		Topiramate
	Melatoinin		

	Cluster Headache	Paroxysmal Hemicrania	SUNCT/SUNA
	Lithium		Gabapentin

[a]If conjunctival injection and tearing are not present, consider SUNA. [b]Nausea, photophobia, or phonophobia; photophobia and phonophobia are typically unilateral on the side of the pain. [c]Non-invasive vagus nerve stimulation is FDA approved in episodic cluster headache [d]Indicates complete response to indomethacin.

Abbreviations: SUNA, *s*hort-lasting *u*nilateral *n*euralgiform headache attacks with cranial *a*utonomic features; SUNCT, *s*hort-lasting *u*nilateral *n*euralgiform headache attacks with *c*onjunctival injection and *t*earing.

TACs must be differentiated from short-lasting headaches that do not have prominent cranial autonomic syndromes, notably trigeminal neuralgia (TN), primary stabbing headache, and hypnic headache. The cycling pattern and length, frequency, and timing of attacks are useful in classifying patients. Patients with TACs should undergo pituitary imaging and pituitary function tests because there is an excess of TAC presentations in patients with pituitary tumor–related headache, particularly prolactin and growth hormone secreting tumors.

Cluster Headache

Cluster headache is a relatively rare form of primary headache, although nonetheless a common condition, with a population frequency of ~0.1%. The pain is deep, usually retroorbital, often excruciating in intensity, nonfluctuating, and explosive in quality. A core feature of cluster headache is periodicity. At least one of the daily attacks of pain recurs at about the same hour each day for the duration of a cluster bout. The typical cluster headache patient has daily bouts of one to two attacks of relatively short-duration unilateral pain for 8–10 weeks a year; this is usually followed by a pain-free interval that averages a little less than 1 year. Cluster headache is characterized as chronic when there is <3 months of sustained remission without treatment. Patients are generally perfectly well between episodes. Onset of attacks is nocturnal in about 50% of patients, and men are affected three times more often than women. Patients with cluster headache tend to move about during attacks, pacing, rocking, or rubbing their head for relief; some may even become aggressive during attacks. This is in sharp contrast to patients with migraine, who prefer to remain motionless during attacks.

Cluster headache is associated with ipsilateral symptoms of cranial parasympathetic autonomic activation: conjunctival injection or lacrimation, aural fullness, rhinorrhea or nasal congestion, or cranial sympathetic dysfunction such as ptosis. The sympathetic deficit is peripheral and likely to be due to parasympathetic activation with injury to ascending sympathetic fibers surrounding a dilated carotid artery as it passes into the cranial cavity. When present, photophobia and phonophobia are far more likely to be unilateral and on the same side of the pain, rather than bilateral, as is seen in migraine. This phenomenon of unilateral photophobia/phonophobia is characteristic of TACs. Cluster headache is likely to be a disorder involving central pacemaker neurons and neurons in the posterior hypothalamic region (Fig. 422-3).

TREATMENT

TREATMENT

Cluster Headache

The most satisfactory treatment is the administration of drugs to prevent cluster attacks until the bout is over. However, treatment of acute attacks is required for all cluster headache patients at some time.

ACUTE ATTACK TREATMENT

Cluster headache attacks peak rapidly, and thus a treatment with rapid onset is required. Many patients with acute cluster headache respond very well to oxygen inhalation. This should be given as 100% oxygen at 10–12 L/min for 15–20 min. It appears that high flow and high oxygen content are important. Sumatriptan 6 mg SC is rapid in onset and will usually shorten an attack to 10–15 min; there is no evidence of tachyphylaxis. Sumatriptan (20 mg) and zolmitriptan (5 mg) nasal sprays are both effective in acute cluster headache, offering a useful option for patients who may not wish to self-inject daily. Noninvasive vagus nerve stimulation (nVNS) is FDA approved for the acute treatment of attacks in episodic cluster headache using three 2-min stimulation cycles applied consecutively at the onset of headache on the side of pain; this may be repeated after nine minutes. Oral sumatriptan is not effective for prevention or for acute treatment of cluster headache.

PREVENTIVE TREATMENTS

The choice of a preventive treatment in cluster headache depends in part on the length of the bout **(TABLE 422-8)**. Patients with long bouts or those with chronic cluster headache require medicines that are safe when taken for long periods. For patients with relatively short bouts, limited courses of oral glucocorticoids can be very useful. A 10-day course of prednisone, beginning at 60 mg daily for 7 days and followed by a rapid taper, may interrupt the pain bout for many patients. Greater occipital nerve injection with lidocaine and glucocorticoids has been shown to be effective in randomized controlled trials, with a benefit that lasts up to 6–8 weeks.

Most experts favor verapamil as the first-line preventive treatment for patients with chronic cluster headache or with prolonged bouts. While verapamil compares favorably with lithium in practice, some patients require verapamil doses far in excess of those administered for cardiac disorders. The initial dose range is 40–80 mg twice daily; effective doses may be as high as 960 mg/d. Side effects such as constipation, leg swelling, or gingival hyperplasia can be problematic. Of paramount concern, however, is the cardiovascular safety of verapamil, particularly at high doses. Verapamil can cause heart block by slowing conduction in the atrioventricular node, a condition that can be monitored by following the PR interval on a standard electrocardiogram (ECG). Approximately 20% of patients treated with verapamil develop ECG abnormalities, which can be observed with doses as low as 240 mg/d; these abnormalities can worsen over time in patients on stable doses. A baseline ECG is recommended for all patients. The ECG is repeated 10 days after a dose change in patients whose dose is being increased above 240 mg daily. Dose increases are usually made in 80-mg increments. For patients on long-term verapamil, ECG monitoring every 6 months is advised.

NEUROMODULATION THERAPY

When medical therapies fail in chronic cluster headache, neuromodulation strategies can be used. Sphenopalatine ganglion (SPG) stimulation with an implanted battery-free stimulator has been shown in randomized controlled trials to be effective in aborting attacks and reducing their frequency over time. nVNS compares favorably to standard-of-care in open label experience. Similarly, occipital nerve stimulation has been used open label and appears to be beneficial. Deep-brain stimulation of the region of the posterior hypothalamic gray matter is successful in about 50% of patients treated, although its risk-benefit ratio makes it inappropriate before all other less invasive options have been explored.

TABLE 422-8

Preventive Management of Cluster Headache

Short-Term Prevention	Long-Term Prevention
Episodic Cluster Headache	**Episodic Cluster Headache and Prolonged Chronic Cluster Headache**
Prednisone 1 mg/kg up to 60 mg qd, tapering over 21 days Verapamil 160–960 mg/d Greater occipital nerve injection	Verapamil 160–960 mg/d Topiramate[a] 100–400 mg/d nVNS[b] 6 to 24 stimulations/d Lithium 400–800 mg/d Melatonin[a] 9–12 mg/d Gabapentin[a] 1200–3600 mg/d

[a]Unproven but of potential benefit. [b]Non-invasive vagus nerve stimulation.

PAROXYSMAL HEMICRANIA

PH is characterized by frequent unilateral, severe, short-lasting episodes of headache. Like cluster headache, the pain tends to be retroorbital but may be experienced all over the head and is associated with autonomic phenomena such as lacrimation and nasal congestion. Patients with remissions are said to have episodic PH, whereas those with the nonremitting form are said to have chronic PH. The essential features of PH are unilateral; very severe pain; short-lasting attacks (2–45 min); very frequent attacks (usually >5 a day); marked autonomic features ipsilateral to the pain; rapid course (<72 h); and excellent response to indomethacin. In contrast to cluster headache, which predominantly affects males, the male-to-female ratio in PH is close to 1:1.

Indomethacin (25–75 mg tid), which can completely suppress attacks of PH, is the treatment of choice. Although therapy may be complicated by indomethacin-induced gastrointestinal side effects, currently there are no consistently effective alternatives. Topiramate is helpful in some cases. Piroxicam has been used, although it is not as effective as indomethacin. Verapamil, an effective treatment for cluster headache, does not appear to be useful for PH. nVNS can be useful in these patients and can be very effective. In occasional patients, PH can coexist with TN (PH-tic syndrome); similar to cluster-tic syndrome, each component may require separate treatment.

Secondary PH has been reported with lesions in the region of the sella turcica, including arteriovenous malformation, cavernous sinus meningioma, pituitary pathology, and epidermoid tumors. Secondary PH is more likely if the patient requires high doses (>200 mg/d) of indomethacin. In patients with apparent bilateral PH, raised cerebrospinal fluid (CSF) pressure should be suspected. It is important to note that indomethacin reduces CSF pressure. When a diagnosis of PH is considered, magnetic resonance imaging (MRI) is indicated to exclude a pituitary lesion.

SUNCT/SUNA

SUNCT is a rare primary headache syndrome characterized by severe, unilateral orbital or temporal pain that is stabbing or throbbing in quality. Diagnosis requires at least 20 attacks, lasting for 5–240 s; ipsilateral conjunctival injection and lacrimation should be present. In some patients, conjunctival injection or lacrimation is missing, and the diagnosis of SUNA can be made.

DIAGNOSIS

The pain of SUNCT/SUNA is unilateral and may be located anywhere in the head. Three basic patterns can be seen: single stabs, which are usually short-lived; groups of stabs; or a longer attack comprising many stabs between which the pain does not completely resolve, thus giving a “saw-tooth” phenomenon with attacks lasting many minutes. Each pattern may be seen in the context of an underlying continuous head pain. Characteristics that lead to a suspected diagnosis of SUNCT are the cutaneous (or other) triggers of attacks, a lack of refractory period to triggering between attacks, and the lack of a response to indomethacin. Apart from trigeminal sensory disturbance, the neurologic examination is normal in primary SUNCT/SUNA.

The diagnosis of SUNCT/SUNA is often confused with TN particularly in first-division TN **(Chap. 433)**. Minimal or no cranial autonomic symptoms and a clear refractory period to triggering indicate a diagnosis of TN.

SECONDARY (SYMPTOMATIC) SUNCT

SUNCT can be seen with posterior fossa or pituitary lesions. All patients with SUNCT/SUNA should be evaluated with pituitary function tests and a brain MRI with pituitary views.

TREATMENT

TREATMENT

SUNCT/SUNA

ABORTIVE THERAPY

Therapy of acute attacks is not a useful concept in SUNCT/SUNA because the attacks are of such short duration. However, IV lidocaine, which arrests the symptoms, can be used in hospitalized patients.

PREVENTIVE THERAPY

Long-term prevention to minimize disability and hospitalization is the goal of treatment. The most effective treatment for prevention is lamotrigine, 200–400 mg/d. Topiramate and gabapentin may also be effective. Carbamazepine, 400–500 mg/d, has been reported by patients to offer modest benefit.

Surgical approaches such as microvascular decompression or destructive trigeminal procedures are seldom useful and often produce long-term complications. Greater occipital nerve injection has produced limited benefit in some patients. Occipital nerve stimulation is probably helpful in a subgroup of these patients. For intractable cases, short-term prevention with IV lidocaine can be effective.

Hemicrania Continua

The essential features of hemicrania continua are moderate and continuous unilateral pain associated with fluctuations of severe pain; complete resolution of pain with indomethacin; and exacerbations that may be associated with autonomic features, including conjunctival injection, lacrimation, and photophobia on the affected side. The age of onset ranges from 10 to 70 years; women are affected twice as often as men. The cause is unknown.

TREATMENT

TREATMENT

Hemicrania Continua

Treatment consists of indomethacin; other NSAIDs appear to be of little or no benefit. The IM injection of 100 mg of indomethacin has been proposed as a diagnostic tool, and administration with a placebo injection in a blinded fashion can be very useful diagnostically. Alternatively, a trial of oral indomethacin, starting with 25 mg tid, then 50 mg tid, and then 75 mg tid, can be given. Up to 2 weeks at the maximal dose may be necessary to assess whether a dose has a useful effect. Topiramate can be helpful in

some patients. nVNS can be useful in these patients. Occipital nerve stimulation probably has a role in patients with hemicrania continua who are unable to tolerate indomethacin.

OTHER PRIMARY HEADACHES

Primary Cough Headache

Primary cough headache is a generalized headache that begins suddenly, lasts for seconds or several minutes, sometimes up to a few hours, and is precipitated by coughing; it is preventable by avoiding coughing or other precipitating events, which can include sneezing, straining, laughing, or stooping. In all patients with this syndrome, serious etiologies must be excluded before a diagnosis of "benign" primary cough headache can be established. A Chiari malformation or any lesion causing obstruction of CSF pathways or displacing cerebral structures can be the cause of the head pain. Other conditions that can present with cough or exertional headache as the initial symptom include cerebral aneurysm, carotid stenosis, and vertebrobasilar disease. Benign cough headache can resemble benign exertional headache (below), but patients with the former condition are typically older.

TREATMENT

TREATMENT

Primary Cough Headache

Indomethacin 25–50 mg two to three times daily is the treatment of choice. Some patients with cough headache obtain complete cessation of their attacks with lumbar puncture; this is a simple option when compared to prolonged use of indomethacin, and it is effective in about one-third of patients. The mechanism of this response is unclear.

Primary Exercise Headache

Primary exercise headache has features resembling both cough headache and migraine. It may be precipitated by any form of exercise; it often has the pulsatile quality of migraine. The pain lasts <48 h, is bilateral and often throbbing at onset; migrainous features may develop in patients susceptible to migraine. The duration tends to be shorter in adolescents than in older adults. Primary exercise headache can be prevented by avoiding excessive exertion, particularly in hot weather or at high altitude.

The mechanism of primary exercise headache is unclear. Acute venous distension likely explains one syndrome—the acute onset of headache with straining and breath holding, as in weightlifter's headache. Because exercise can result in headache in a number of serious underlying conditions (Chap. 13), these must be considered in patients with exercise headache. Pain from angina may be referred to the head, probably by central connections of vagal afferents, and may present as exercise headache (cardiac cephalgia). The link to exercise is the main clinical clue that headache is of cardiac origin. Pheochromocytoma may occasionally cause exercise headache. Intracranial lesions and stenosis of the carotid arteries are other possible etiologies.

TREATMENT

TREATMENT

Primary Exercise Headache

Exercise regimens should begin modestly and progress gradually to higher levels of intensity. Indomethacin at daily doses from 25 to 150 mg is generally effective in benign exertional headache. Indomethacin (50 mg), ergotamine (1 mg orally), and dihydroergotamine (2 mg by nasal spray) are useful prophylactic measures.

Primary Headache Associated with Sexual Activity

Three types of sex headache are reported: a dull bilateral ache in the head and neck that intensifies as sexual excitement increases; a sudden, severe, explosive headache occurring at orgasm; and a postural headache developing after coitus. The last arises from vigorous sexual activity and is a form of low CSF pressure headache and thus not a primary headache disorder **(Chap. 13)**. Headaches developing at the time of orgasm are not always benign; 5–12% of cases of subarachnoid hemorrhage are precipitated by sexual intercourse. Sex headache is reported by men more often than women and may occur at any time during the years of sexual activity. It may appear on several occasions in succession and then not trouble the patient again, even without an obvious change in sexual activity. In patients who stop sexual activity when headache is first noticed, the pain may subside within a period of 5 min to 2 h. In about half of patients, sex headache will subside within 6 months. Most patients with sex headache do not have exercise or cough headache; this clinical paradox is generally a marker of primary sex headache. Migraine is probably more common in patients with sex headache.

TREATMENT

TREATMENT

Primary Sex Headache

Benign sex headaches recur irregularly and infrequently. Management can often be limited to reassurance and advice about ceasing sexual activity if a mild, warning headache develops. Propranolol can be used to prevent headache that recurs regularly or frequently, but the dosage required varies from 40 to 200 mg/d. An alternative is the calcium channel–blocking agent diltiazem, 60 mg tid. Indomethacin (25–50 mg) or frovatriptan (2.5 mg) taken 30–45 min prior to sexual activity can also be helpful.

Primary Thunderclap Headache

Sudden onset of severe headache may occur in the absence of any known provocation. The differential diagnosis includes the sentinel bleed of an intracranial aneurysm, cervicocephalic arterial dissection, and cerebral venous thrombosis. Headaches of explosive onset may also be caused by the ingestion of sympathomimetic drugs or of tyramine-containing foods in a patient who is taking MAOIs, or they may be a symptom of pheochromocytoma. Whether thunderclap headache can be the presentation of an unruptured cerebral aneurysm is uncertain. When neuroimaging studies and lumbar puncture exclude subarachnoid hemorrhage, patients with thunderclap headache usually do very well over the long term. In one study of patients whose computed tomography (CT) scans and CSF findings were negative, ~15% had recurrent episodes of thunderclap headache, and nearly half subsequently developed migraine or TTH.

The first presentation of any sudden-onset severe headache should be diligently investigated with neuroimaging (CT or, when possible, MRI with MR angiography) and CSF examination. Reversible segmental cerebral vasoconstriction may be seen in primary thunderclap headache without an intracranial aneurysm, and it is thought that this may be an under-diagnosed condition. In the presence of posterior leukoencephalopathy, the differential diagnosis includes cerebral angiitis, drug toxicity (cyclosporine, intrathecal methotrexate/cytarabine, pseudoephedrine, or cocaine), posttransfusion effects, and postpartum angiopathy. Treatment with nimodipine may be helpful, although the vasoconstriction of primary thunderclap headache resolves spontaneously.

Cold-Stimulus Headache

This refers to head pain triggered by application or ingestion/inhalation of something cold. It is bought on quickly and typically resolves within 10–30 min of the stimulus being removed. It is best recognized as “brain-freeze” headache or ice-cream headache when due to ingestion. Although cold may be uncomfortable at some level for many people, it is the reliable, severe, and somewhat prolonged nature of these pains that set them apart. The transient receptor potential cation subfamily M member 8 (TRPM8) channel, a known cold temperature sensor, may be a mediator of this syndrome.

External Pressure Headache

External pressure from compression or traction on the head can produce a pain that may have some generalized component, although the pain is largely focused around the site of the pressure. It typically resolves within an hour of the stimulus being removed. Examples of stimuli include helmets, swimming goggles, or very long ponytails. Treatment is to recognize the problem and remove the stimulus.

Primary Stabbing Headache

The essential features of primary stabbing headache are stabbing pain confined to the head or, rarely, the face, lasting from 1 to many seconds and occurring as a single stab or a series of stabs; absence of associated cranial autonomic features; absence of cutaneous triggering of attacks; and a pattern of recurrence at irregular intervals (hours to days). The pains have been variously described as “ice-pick pains” or “jabs and jolts.” They are more common in patients with other primary headaches, such as migraine, the TACs, and hemicrania continua.

TREATMENT

TREATMENT

Primary Stabbing Headache

The response of primary stabbing headache to indomethacin (25–50 mg two to three times daily) is usually excellent. As a general rule, the symptoms wax and wane, and after a period of control on indomethacin, it is appropriate to withdraw treatment and observe the outcome.

Nummular Headache

Nummular headache is felt as a round or elliptical discomfort that is fixed in place, ranges in size from 1 to 6 cm, and may be continuous or intermittent. Uncommonly it may be multifocal. It may be episodic but is more often continuous during exacerbations. Accompanying the pain there may be a local sensory disturbance, such as allodynia or hypesthesia. Local dermatologic or bony lesions need to be excluded by examination and investigation. This condition can be difficult to treat when present in isolation; tricyclics, such as amitriptyline, or anticonvulsants, such as topiramate or valproate, are most often tried. This phenotype can be seen in combination with migraine and the TACs, in which cases treatment of the associated condition is often effective for the nummular headache as well.

Hypnic Headache

This headache syndrome typically begins a few hours after sleep onset. The headaches last from 15 to 30 min and are typically moderately severe and generalized, although they may be unilateral and can be throbbing. Patients may report falling back to sleep only to be awakened by a further attack a few hours later; up to three repetitions of this pattern occur through the night. Daytime naps can also precipitate head pain. Most patients are female, and the onset is usually after age 60 years. Headaches are bilateral in most, but may be unilateral. Photophobia, phonophobia, and nausea are usually absent. The major secondary consideration in this headache type is poorly controlled hypertension; 24-h blood pressure monitoring is recommended to detect this treatable condition.

TREATMENT

TREATMENT

Hypnic Headache

Patients with hypnic headache generally respond to a bedtime dose of lithium carbonate (200–600 mg). For those intolerant of lithium, verapamil (160 mg) is an alternative strategy. One to two cups of coffee or caffeine, 60 mg orally, at bedtime may be

effective in approximately one-third of patients. Case reports also suggest that flunarizine, 5 mg nightly, or indomethacin, 25–75 mg nightly, can be effective.

New Daily Persistent Headache

Primary new daily persistent headache (NDPH) occurs in both males and females. It can be of the migrainous type, with features of migraine, or it can be featureless, appearing as new-onset TTH. Those with migrainous features are the most common form, and include unilateral headache and throbbing pain; each feature is present in about one-third of patients. Nausea, photophobia, and/or phonophobia occur in about half of patients. Some patients have a previous history of migraine; however, the proportion of NDPH sufferers with preexisting migraine is no greater than the frequency of migraine in the general population. NDPH may be more common in adolesents. Treatment of migrainous-type primary NDPH consists of using the preventive therapies effective in migraine (see above). Featureless NDPH is one of the primary headache forms most refractory to treatment. Standard preventive therapies can be offered but are often ineffective. The secondary NDPHs are discussed elsewhere **(Chap. 13)**.

ACKNOWLEDGMENT

The Editors acknowledge the contributions of Neil H. Raskin to earlier editions of this chapter.

FURTHER READING

Charles A: Migraine. N Engl J Med 377:553, 2017. [PubMed: 28792865]

Goadsby PJ: Bench to bedside advances in the 21st century for primary headache disorders: Migraine treatments for migraine patients. Brain 139:2571, 2016. [PubMed: 27671024]

Goadsby PJ et al: Pathophysiology of migraine: A disorder of sensory processing. Physiol Rev 97:553, 2017. [PubMed: 28179394]

Lipton RB et al: Migraine prevalence, disease burden, and the need for preventive therapy. Neurology 68:343, 2007. [PubMed: 17261680]

May A: Cluster headache: Pathogenesis, diagnosis, and management. Lancet 366:843, 2005. [PubMed: 16139660]

Schankin CJ et al: “Visual snow”—a disorder distinct from persistent migraine aura. Brain 137:1419, 2014. [PubMed: 24645145]

Tolner EA et al: From migraine genes to mechanisms. Pain 156 Suppl 1:S64, 2015. [PubMed: 25789438]

Chapter 423: Alzheimer's Disease

William W. Seeley; Bruce L. Miller

ALZHEIMER'S DISEASE

Approximately 10% of all persons aged >70 years have significant memory loss, and in more than half the cause is Alzheimer's disease (AD). It is estimated that the median annual total cost of caring for a single patient with advanced AD is >$50,000, while the emotional toll for family members and caregivers is immeasurable. AD can manifest as early as the third decade of life, but it is the most common cause of dementia in the elderly. Patients most often present with an insidious loss of episodic memory followed by a slowly progressive dementia. In typical amnestic AD, brain atrophy begins in the medial temporal lobes before spreading to lateral and medial parietal and temporal lobes and lateral frontal cortex. Microscopically, there are widespread neuritic plaques containing amyloid beta (Aβ), neurofibrillary tangles (NFTs) composed of hyperphosphorylated tau filaments, and Aβ accumulation in blood vessel walls in cortex and leptomeninges (see "Pathology," below). The identification of causative mutations and susceptibility genes for AD has provided a foundation for rapid progress in understanding the biological basis of the disorder. The major genetic risk for AD is the ε4 allele of the apolipoprotein E (ApoE) gene. Carrying one ε4 allele increases the risk for AD by two- to threefold whereas two alleles increase the risk sixteenfold in both sexes.

CLINICAL MANIFESTATIONS

The cognitive changes of AD tend to follow a characteristic pattern, beginning with memory impairment and progressing to language and visuospatial deficits, followed by executive dysfunction. Yet, ~20% of patients with AD present with non-memory complaints such as word-finding, organizational, or navigational difficulty. In other patients, visual processing dysfunction (referred to as posterior cortical atrophy syndrome) or a progressive "logopenic" aphasia characterized by difficulties with naming and repetition are the primary manifestations of AD for years before progressing to involve memory and other cognitive domains. Still other patients may present with an asymmetric akinetic-rigid-dystonic ("corticobasal") syndrome or a dysexecutive/behavioral, i.e., "frontal" variant of AD.

In the early stages of typical amnestic AD, the memory loss may go unrecognized or be ascribed to benign forgetfulness of aging. Once the memory loss becomes noticeable to the patient and spouse and falls 1.5 standard deviations below normal on standardized memory tests, the term mild cognitive impairment (MCI) is often used. This construct provides useful prognostic information, because ~50% of patients with MCI (roughly 12% per year) will progress to AD over 4 years. Increasingly, the MCI construct is being replaced by the notion of "early symptomatic AD" to signify that AD is considered the underlying disease (based on clinical or biomarker evidence) in a patient who remains functionally compensated. Even earlier in the course, "prodromal AD" refers to a person with biomarker evidence of AD (amyloid imaging positive with positron emission tomography (PET) or low cerebrospinal $A\beta_{42}$ and mildly elevated tau) in the absence of symptoms. These refinements have been developed in anticipation of early-stage treatment and prevention trials that are well underway in humans. New evidence suggests that partial and sometimes generalized seizures herald AD and can occur even prior to dementia onset, especially in younger patients.

Eventually, with AD, the cognitive problems begin to interfere with daily activities, such as keeping track of finances, following instructions on the job, driving, shopping, and housekeeping. Some patients are unaware of these difficulties (*anosognosia*), but most remain acutely attuned to their deficits. Changes in environment (travel, relocation, hospitalization) tend to destabilize the patient. Over time, patients become lost on walks or while driving. Social graces, routine behavior, and superficial conversation may be surprisingly intact, even into the later stages of the illness.

In the middle stages of AD, the patient is unable to work, is easily lost and confused, and requires daily supervision. Language becomes impaired—first naming, then comprehension, and finally fluency. Word-finding difficulties and circumlocution can be evident in the early stages, even when formal testing demonstrates intact naming and fluency. *Apraxia* emerges, manifesting as trouble performing learned sequential motor tasks such as using utensils or appliances. Visuospatial deficits begin to interfere with dressing, eating, or even walking, and patients fail to solve simple puzzles or copy geometric figures. Simple calculations and clock reading become difficult in parallel.

In the late stages, some persons remain ambulatory, wandering aimlessly. Loss of judgment and reasoning is inevitable. Delusions are prevalent and usually simple, with common themes of theft, infidelity, or misidentification. Disinhibition and uncharacteristic belligerence may occur and alternate with passivity and withdrawal. Sleep-wake patterns are disrupted, and nighttime wandering becomes disturbing to the household. Some patients develop a shuffling gait with generalized muscle rigidity associated with slowness and awkwardness of movement. Patients often look parkinsonian (Chap. 427) but rarely have a high-amplitude, low-frequency tremor at rest. There is a strong overlap between dementia with Lewy bodies (DLB) (Chap. 426) and AD, and some AD patients develop more classical parkinsonian features.

In the end stages, patients with AD become rigid, mute, incontinent, and bedridden, and help is needed with eating, dressing, and toileting. Hyperactive tendon reflexes and myoclonic jerks (sudden brief contractions of various muscles or the whole body) may occur spontaneously or in response to physical or auditory stimulation. Often death results from malnutrition, secondary infections, pulmonary emboli, heart disease, or, most commonly, aspiration. The typical duration of symptomatic AD is 8–10 years, but the course ranges from 1 to 25 years. For unknown reasons, some patients with AD show a steady decline in function while others have prolonged plateaus without major deterioration.

DIFFERENTIAL DIAGNOSIS

A detailed discussion of the diagnosis of dementia is presented in Chap. 25. Early in the disease course, other etiologies of dementia should be excluded **(see Tables 25-1, 25-3, and 25-4)**. Neuroimaging studies (computed tomography [CT] and magnetic resonance imaging [MRI]) do not show a single specific pattern with AD and may be normal early in the disease. As AD progresses, more distributed but usually posterior-predominant cortical atrophy becomes apparent, along with atrophy of the medial temporal memory structures **(see Fig. 25-1)**. The main purpose of imaging is to exclude other disorders, such as primary and secondary neoplasms, vascular dementia, diffuse white matter disease, and normal-pressure hydrocephalus (NPH). Imaging also helps to distinguish AD from other degenerative disorders, such as frontotemporal dementia (FTD) (Chap. 424) or the prion disorder Creutzfeldt-Jakob disease (CJD) (Chap. 430), which feature distinctive imaging patterns. Functional imaging studies, such as PET, reveal hypometabolism in the posterior temporal-parietal cortex in AD **(see Fig. 25-1)**. PET can also be used to detect the presence of fibrillar amyloid in the brain **(see Fig. 25-4)**, and amyloid PET positivity is becoming a criterion for entry into AD treatment trials. Use of amyloid PET in routine clinical evaluation may be limited to specific clinical scenarios, however. Although amyloid PET binding is detected in AD, many asymptomatic healthy older individuals also show amyloid uptake, and the likelihood that these individuals will convert to clinical AD is still under study. Similarly, dementia due to a non-AD disorder can be the underlying etiology in a patient who tests positively on amyloid PET. Electroencephalogram (EEG) is normal or shows nonspecific slowing; prolonged EEG can be used to seek out intermittent nonconvulsive seizures. Routine spinal fluid examination is also normal. Cerebrospinal fluid (CSF) $A\beta_{42}$ level is reduced, whereas phosphorylated tau protein is elevated, but the test characteristics of these assays can still make interpretation challenging in individual patients. *Slowly progressive decline in memory and orientation, normal results on laboratory tests, and an MRI or CT scan showing only distributed or posteriorly predominant cortical and hippocampal atrophy are highly suggestive of AD.* A clinical diagnosis of AD reached after careful evaluation is confirmed at autopsy about 90% of the time, with misdiagnosed cases usually resulting from pathological frontotemporal lobar degeneration (FTLD), DLB, hippocampal sclerosis of the elderly, or a mixture of mild AD changes with vascular or DLB pathology.

Simple clinical clues are useful in the differential diagnosis. Early prominent gait disturbance with only mild memory loss suggests vascular dementia or, rarely, NPH (see below). Resting tremor with stooped posture, bradykinesia, and masked facies suggest PD (Chap. 427) or DLB (Chap. 426). When dementia occurs after a well-established diagnosis of PD, PD dementia (PDD) is usually the correct diagnosis, but many patients with this diagnosis will show a mixture of AD and Lewy body disease at autopsy. The early appearance of parkinsonian features in association with fluctuating alertness, visual hallucinations, or delusional misidentification suggests DLB. Chronic alcoholism should prompt the search for vitamin deficiency. Loss of joint position and vibration sensibility accompanied by Babinski signs suggests vitamin B_{12} deficiency, especially in a patient with a history of autoimmune disease, small bowel resection or irradiation, or veganism (Chap. 95). Early onset of a focal seizure suggests a metastatic or primary brain neoplasm (Chap. 86). Previous or ongoing depression raises suspicion for depression-related cognitive impairment, although AD can feature a depressive prodrome. A history of treatment for insomnia, anxiety, psychiatric disturbance, or epilepsy suggests chronic drug intoxication. Rapid progression over a few weeks or months associated with rigidity and myoclonus suggests CJD (Chap. 430). Prominent behavioral changes with intact navigation and focal anterior-predominant atrophy on brain imaging are typical of FTD. A positive family history of dementia suggests either one of the familial forms of AD or one of the other genetic disorders associated with dementia, such as FTD (Chap. 424), Huntington's disease (HD) (Chap. 428), prion disease (Chap. 430), or rare hereditary ataxias (Chap. 431).

EPIDEMIOLOGY

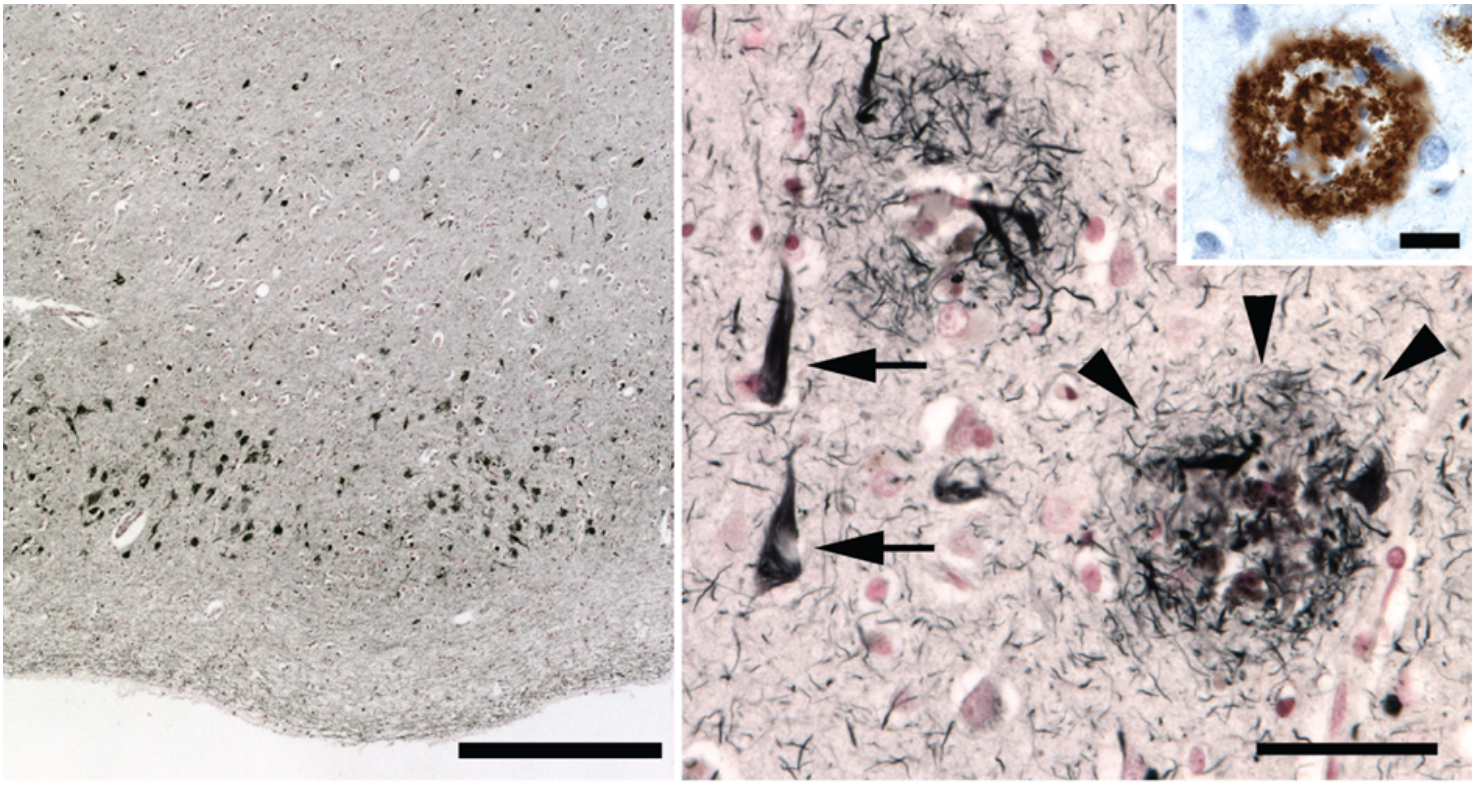
The most important risk factors for AD are age >70 years and a positive family history. The prevalence of AD increases with each decade of adult life, reaching 20–40% of the population aged >85. A positive family history of dementia suggests a genetic contribution to AD, although autosomal dominant inheritance occurs in only 2% of patients. Female sex is a risk factor independent of the greater longevity of women, and women who carry a single Apo ε4 allele are more susceptible than are male ε4 carriers. A history of head trauma with concussion increases the risk for AD. AD is more common in groups with low educational attainment, but education influences test-taking ability, and it is clear that AD can affect persons of all intellectual levels. One study found that the capacity to express complex written language in early adulthood correlated with a decreased risk for AD. Numerous environmental factors, including aluminum, mercury, and viruses, have been proposed as causes of AD, but rigorous studies have failed to demonstrate a significant role for any of these exposures. Similarly, several studies suggest that the use of nonsteroidal anti-inflammatory agents is associated with a decreased risk of AD, but this risk has not been confirmed in large prospective studies. Vascular disease, and stroke in particular, seems to lower the threshold for the clinical expression of AD. Also, in many patients with AD, amyloid angiopathy can lead to microhemorrhages, large lobar hemorrhages, ischemic infarctions most often in the subcortical white matter, or in rare cases an inflammatory leukoencephalopathy. Diabetes increases the risk of AD threefold. Elevated homocysteine and cholesterol levels; hypertension; diminished serum levels of folic acid; low dietary intake of fruits, vegetables, and red wine; and low levels of exercise are all being explored as potential risk factors for AD.

PATHOLOGY

At autopsy, the earliest and most severe degeneration is usually found in the medial temporal lobe (entorhinal/perirhinal cortex and hippocampus), inferolateral temporal cortex, and nucleus basalis of Meynert. The characteristic microscopic findings are neuritic plaques and NFTs **(Fig. 423-1)**. These lesions accumulate in small numbers during normal brain aging but dominate the picture in AD. Increasing evidence suggests that soluble amyloid species called *oligomers* may cause cellular dysfunction and represent the early toxic molecule in AD. Eventually, further amyloid polymerization and fibril formation lead to neuritic plaques, which contain a central core of amyloid, proteoglycans, ApoE, α-antichymotrypsin, and other proteins. Aβ is a protein of 39–42 amino acids that is derived proteolytically from a larger transmembrane protein, *amyloid precursor protein* (APP), when APP is cleaved by β and γ secretases **(Fig. 423-2)**. The normal function of the Aβ peptides remains uncertain. APP has neurotrophic and neuroprotective properties. The plaque core is surrounded by a halo, which contains dystrophic, tau-immunoreactive neurites and activated microglia. The accumulation of Aβ in cerebral arterioles is termed *amyloid angiopathy*. NFTs are composed of silver-staining neuronal cytoplasmic fibrils composed of abnormally phosphorylated tau protein; they appear as paired helical filaments by electron microscopy. Tau binds to and stabilizes microtubules, supporting axonal transport of organelles, glycoproteins, neurotransmitters, and other important cargoes throughout the neuron. Once hyperphosphorylated, tau can no longer bind properly to microtubules and redistributes from the axon to throughout the neuronal cytoplasm and distal dendrites, compromising function. Other theories emphasize that abnormal conformations of tau induce misfolding of native (unfolded) tau into pathological conformations and that this prion-like templating process is responsible for tau spreading **(Chap. 417)**. Finally, patients with AD often show comorbid DLB or vascular pathology. Most prevailing rodent models of AD involve expression of mutant transgenes that leads to $A\beta_{42}$ accumulation in the absence of tauopathy. Even in these models, diminishing neuronal tau ameliorates cognitive deficits and nonconvulsive seizures while $A\beta_{42}$ continues to accumulate, raising hope for tau-lowering therapies in humans. Biochemically, AD is associated with a decrease in the cortical levels of several proteins and neurotransmitters, especially acetylcholine, its synthetic enzyme choline acetyltransferase, and nicotinic cholinergic receptors. Reduction of acetylcholine reflects degeneration of cholinergic neurons in the nucleus basalis of Meynert, located just below the thalamus and adjacent to the third ventricle, that project throughout the cortex. There is also noradrenergic and serotonergic depletion due to degeneration of upper brainstem nuclei such as the locus coeruleus (norepinephrine) and dorsal raphe (serotonin), where tau-immunoreactive neuronal cytoplasmic inclusions can be identified in early adult life, even in individuals lacking entorhinal cortex NFTs.

FIGURE 423-1

Neuropathology of Alzheimer's disease. ***A.*** Early neurofibrillary degeneration, consisting of neurofibrillary tangles and neuropil threads, preferentially affects the medial temporal lobes, especially the stellate pyramidal neurons that compose the layer 2 islands of entorhinal cortex, as shown using Gallyas silver staining. ***B.*** Higher magnification view reveals the fibrillar nature of tangles (*arrows*) and the complex structure of neuritic plaques (*arrowheads*), whose major component is Aβ (*inset* shows immunohistochemistry for Aβ). Scale bars are 500 μM in ***A***, 50 μM in ***B***, and 20 μM in ***B*** inset.

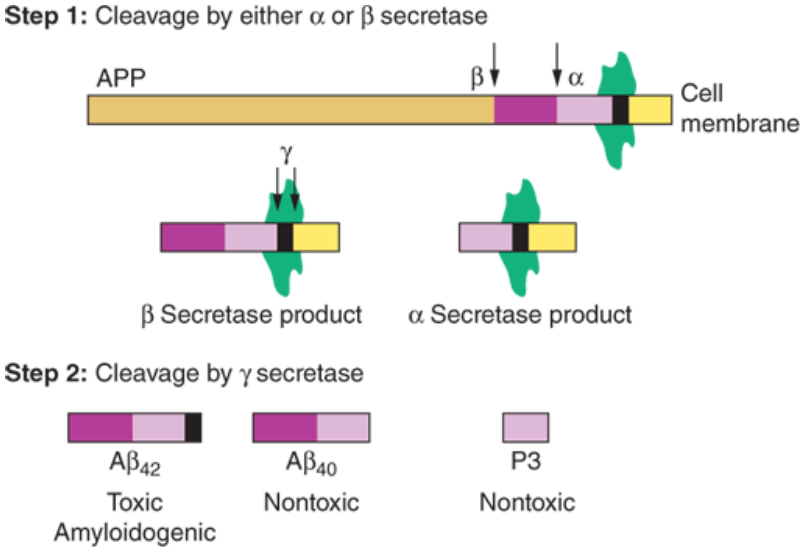


A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 423-2

Amyloid precursor protein (APP) is catabolized by α, β, and γ secretases. A key initial step is the digestion by either β secretase (BASE) or α secretase (ADAM10 or ADAM17 [TACE]), producing smaller nontoxic products. Cleavage of the β secretase product by γ secretase (Step 2) results in either the toxic $A\beta_{42}$ or the nontoxic $A\beta_{40}$ peptide; cleavage of the α secretase product by γ secretase produces the nontoxic P3 peptide. Excess production of $A\beta_{42}$ is a key initiator of cellular damage in Alzheimer's disease (AD). Therapeutics for AD have focused on attempts to reduce accumulation of $A\beta_{42}$ by antagonizing β or γ secretases, promoting α secretase, or clearing $A\beta_{42}$ that has already formed by use of specific antibodies.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GENETIC CONSIDERATIONS

 Several genes play an important role in the pathogenesis of AD. One is the *APP* gene on chromosome 21. Adults with trisomy 21 (Down's syndrome) consistently develop the typical neuropathologic hallmarks of AD if they survive beyond age 40 years, and many develop a progressive dementia superimposed on their baseline mental retardation. The extra dose of the *APP* gene on chromosome 21 is the initiating cause of AD in adult Down's syndrome and results in excess cerebral amyloid production. Supporting this hypothesis, some families with early age-of-onset familial AD (FAD) have point mutations in *APP*. Although very rare, these families were the first examples of single-gene autosomal dominant transmission of AD.

Investigation of large families with multigenerational FAD led to the discovery of two additional AD-causing genes, the *presenilins*. Presenilin-1 (*PSEN-1*) is on chromosome 14 and encodes presenilin-1 protein (also known as S182). Mutations in this gene cause an early age-of-onset AD, with onset typically before age 60 and often before age 50, transmitted in an autosomal dominant, highly penetrant fashion. More than 100 different mutations have been found in the *PSEN-1* gene in families from a wide range of ethnic backgrounds. Presenilin-2 (*PSEN-2*) is on chromosome 1 and encodes the presenilin-2 protein (also known as STM2). A mutation in the *PSEN-2* gene was first found in a group of American families with Volga German ethnic background. Mutations in *PSEN-1* are much more common than those in *PSEN-2*. The presenilins are highly homologous and encode similar proteins that at first appeared to have seven transmembrane domains (hence the designation *STM*), but subsequent studies have suggested eight such domains, with a ninth submembrane region. Both presenilins are cytoplasmic neuronal proteins that are widely expressed throughout the nervous system. They are homologous to a cell-trafficking protein, sel 12, found in the nematode *Caenorhabditis elegans*. Patients with mutations in the presenilin genes have elevated plasma levels of $A\beta_{42}$, and *PSEN-1* mutations produce increased $A\beta_{42}$ in the media in cell culture. *PSEN-1* is involved in the cleavage of APP at the γ secretase site and mutations in either gene (*PSEN-1* or *APP*) may disturb γ secretase cleavage. Mutations in *PSEN-1* are the most common cause of early-age-of-onset FAD, representing perhaps 40–70% of all cases. Mutations in *PSEN-1* tend to produce AD with an earlier age of onset (mean onset 45 years) and a shorter, more rapidly progressive course (mean duration 6–7 years) than the disease caused by mutations in *PSEN-2* (mean onset 53 years; duration 11 years). Although some carriers of *PSEN-2* mutations have had onset of dementia after the age of 70, mutations in the presenilins rarely lead to late-age-of-onset AD. Clinical genetic testing for these uncommon mutations is available but likely to be revealing only in early-age-of-onset FAD and should be performed in association with formal genetic counseling.

The *Apo* ε gene on chromosome 19 is involved in the pathogenesis of AD. The protein product, apolipoprotein E, participates in cholesterol transport (Chap. 400), and the gene has three alleles: ε2, ε3, and ε4. The Apo ε4 allele confers increased risk of AD in the general population, including sporadic and late age-of-onset familial forms. Approximately 24–30% of the non-demented white population has at least one ε4 allele (12–15% allele frequency), and about 2% are ε4/ε4 homozygotes. Among patients with AD, 40–65% have at least one ε4 allele, a highly significant elevation compared with controls. The increased risk associated with a single ε4 allele is especially prominent in women. Conversely, many patients with AD have no ε4 allele, and ε4 carriers may never develop AD. Therefore, ε4 is neither necessary nor sufficient to cause AD. Nevertheless, the Apo ε4 allele represents the most important genetic risk factor for sporadic AD and acts as a dose-dependent disease modifier, with the earliest age of onset associated with the ε4 homozygosity. Precise mechanisms through which Apo ε4 confers AD risk or hastens onset remain unclear, but ε4 leads to less efficient amyloid clearance and to the production of toxic fragments from cleavage of the molecule. Apo ε can be identified in neuritic plaques and may also be involved in neurofibrillary tangle formation, because it binds to tau protein. Apo ε4 decreases neurite outgrowth in dorsal root ganglion neuronal cultures, perhaps indicating a deleterious role in the brain's response to injury. Some evidence suggests that the ε2 allele may reduce AD risk. Use of Apo ε testing in AD diagnosis remains controversial because its predictive value remains unclear and many individuals with the ε4 allele will never develop dementia. Cognitively normal ε4 heterozygotes and homozygotes may show decreased posterior cerebral cortical metabolic function by PET imaging, suggesting presymptomatic abnormalities due to AD or an inherited vulnerability of the AD-targeted network. In demented persons who meet clinical criteria for AD, finding an ε4 allele increases the reliability of diagnosis; however, the absence of an ε4 allele cannot be considered evidence against AD. Nevertheless, Apo ε4 remains the single most important biologic marker associated with AD risk, and studies of ε4's functional role and diagnostic utility are progressing rapidly. The ε4 allele is not associated with risk for FTD, DLB, or CJD, although some evidence suggests that ε4 may worsen the expression of non-AD degenerative disorders, head trauma, and other brain injuries. Additional genes are also likely to be involved in AD, especially as minor risk alleles for sporadic forms of the disease. Genome-wide association studies have implicated the clusterin (*CLU*), phosphatidylinositol-binding clathrin assembly protein (*PICALM*), and complement component (3b/4b) receptor 1 (*CR1*) genes, among others. *CLU* may play a role in synapse turnover, *PICALM* participates in clathrin-mediated endocytosis, and *CR1* may be involved in amyloid clearance or synapse loss through the complement pathway. *TREM2* is a gene involved with inflammation that increases the likelihood of dementia. Homozygous mutation carriers develop a frontal dementia with bone cysts (Nasu-Hakola disease), whereas heterozygotes are predisposed to the development of AD.

TREATMENT

TREATMENT

Alzheimer's Disease

The management of AD is challenging and gratifying despite the absence of a cure or a robust pharmacologic treatment. The primary focus is on long-term amelioration of associated behavioral and neurologic problems, as well as providing caregiver support.

Building rapport with the patient, family members, and other caregivers is essential. In the early stages of AD, memory aids such as notebooks and posted daily reminders can be helpful. Family members should emphasize activities that are pleasant while curtailing those that increase stress on the patient. Kitchens, bathrooms, stairways, and bedrooms need to be made safe, and eventually patients will need to stop driving. Loss of independence and change of environment may worsen confusion, agitation, and anger. Communication and repeated calm reassurance are necessary. Caregiver "burnout" is common, often resulting in nursing home placement of the patient or new health problems for the caregiver. Respite breaks for the caregiver help to maintain a successful long-term therapeutic milieu. Use of adult day care centers can be helpful. Local and national support groups, such as the Alzheimer's Association and the Family Caregiver Alliance, are valuable resources. Internet access to these resources has become available to clinicians and families in recent years.

Donepezil (target dose, 10 mg daily), rivastigmine (target dose, 6 mg twice daily or 9.5-mg patch daily), galantamine (target dose 24 mg daily, extended-release), and memantine (target dose, 10 mg twice daily) are approved by the U.S. Food and Drug Administration (FDA) for the treatment of AD. Due to hepatotoxicity, tacrine is no longer used. Dose escalations for each of these medications must be carried out over 4–6 weeks to minimize side effects. The pharmacologic action of donepezil, rivastigmine, and galantamine is inhibition of the cholinesterases, primarily acetylcholinesterase, with a resulting increase in cerebral acetylcholine levels. Memantine appears to act by blocking overexcited *N*-methyl-D-aspartate (NMDA) glutamate receptors. Double-blind, placebo-controlled, crossover studies with cholinesterase inhibitors and memantine in moderate to severe AD have shown them to be associated with improved caregiver ratings of patients' functioning and with an apparent decreased rate of decline in cognitive test scores over periods of up to 3 years. The average patient on an anticholinesterase inhibitor maintains his or her mini-mental state examination (MMSE) score for close to a year, whereas a placebo-treated patient declines 2–3 points over the same time period. Memantine, used in conjunction with cholinesterase inhibitors or by itself, slows cognitive deterioration and decreases caregiver burden for patients with moderate to severe AD, but is not approved for mild AD. Each of these compounds has only modest efficacy for AD. Cholinesterase inhibitors are relatively easy to administer, and their major side effects are gastrointestinal symptoms (nausea, diarrhea, cramps), altered sleep with unpleasant or vivid dreams, bradycardia (usually benign), and muscle cramps.

In a prospective observational study, the use of estrogen replacement therapy appeared to protect—by about 50%—against development of AD in women. This study seemed to confirm the results of two earlier case-controlled studies. Sadly, a prospective placebo-controlled study of a combined estrogen-progesterone therapy for asymptomatic postmenopausal women increased, rather than decreased, the prevalence of dementia. This study markedly dampened enthusiasm for hormonal treatments to prevent dementia. Additionally, no benefit has been found in the treatment of AD with estrogen alone.

A controlled trial of an extract of *Ginkgo biloba* found modest improvement in cognitive function in subjects with AD and vascular dementia. Unfortunately, a comprehensive 6-year multicenter prevention study using ginkgo found no slowing of progression to dementia in the treated group.

Vaccination against $A\beta_{42}$ has proved highly efficacious in mouse models of AD, helping clear brain amyloid and preventing further amyloid accumulation. In human trials, this approach led to life-threatening complications, including meningoencephalitis, in a minority of patients. Another experimental approach to AD treatment has been the use of β and γ secretase inhibitors that diminish the production of $A\beta_{42}$, but the first two placebo-controlled trials of γ secretase inhibitors, tarenflurbil and semagacestat, were negative, and semagacestat may have accelerated cognitive decline compared to placebo. Passive immunization with monoclonal antibodies against $A\beta_{42}$ has been tried in mild to moderate AD. These studies were negative, leading some to suggest that the patients treated were too advanced to respond to amyloid-lowering therapies. Therefore, new trials have started in asymptomatic individuals with mild AD, in asymptomatic autosomal dominant forms of AD, and in cognitively normal elderly who are amyloid positive with PET. Medications that modify tau phosphorylation and aggregation, including tau antibodies, are beginning to be studied as possible treatments for both AD and non-AD tau-related disorders including FTD and progressive supranuclear palsy (PSP) (Chap. 424).

Several retrospective studies suggest that nonsteroidal anti-inflammatory agents and 3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA) reductase inhibitors (statins) may have a protective effect on dementia if used prior to the onset of disease but do not influence clinically symptomatic AD. Finally, there is now a strong interest in the relationship between diabetes and AD, and insulin-regulating studies are being conducted.

Mild to moderate depression is common in the early stages of AD and may respond to antidepressants or cholinesterase inhibitors. Selective serotonin reuptake inhibitors (SSRIs) are commonly used due to their low anticholinergic side effects (for example, escitalopram, target dose 5–10 mg daily). Seizures can be treated with levetiracetam unless the patient had a different regimen that was effective prior to the onset of AD. Agitation, insomnia, hallucinations, and belligerence are especially troublesome characteristics of some AD patients, and these behaviors can lead to nursing home placement. The newer generation of atypical antipsychotics, such as risperidone, quetiapine, and olanzapine, are being used in low doses to treat these neuropsychiatric symptoms. The few controlled studies comparing drugs against behavioral intervention in the treatment of agitation suggest mild efficacy with significant side effects related to sleep, gait, and cardiovascular complications, including an increased risk of death. All antipsychotics carry a black box FDA warning for use in elderly patients with dementia and thus should be prescribed only with caution; however, careful, daily, nonpharmacologic behavior management is often not available, rendering medications necessary for some patients. Finally, medications with strong anticholinergic effects should be vigilantly avoided, including prescription and over-the-counter sleep aids (e.g., diphenhydramine) or incontinence therapies (e.g., oxybutynin).

A general approach to the symptomatic management of dementia is presented in Chap. 25.

OTHER CAUSES OF DEMENTIA

FTD **(Chap. 424)**, vascular dementia **(Chap. 425)**, and DLB **(Chap. 426)**, are covered in dedicated chapters. Additional important causes of dementia are described here.

Prion diseases such as CJD are rare neurodegenerative conditions (prevalence ~1 per million) that produce dementia. CJD is a rapidly progressive disorder associated with dementia, focal cortical signs, rigidity, and myoclonus, causing death <1 year after first symptoms appear. The rapidity of progression seen with CJD is uncommon in AD so that the distinction between the two disorders is usually straightforward. Cortical basal degeneration (CBD) **(Chap. 424)** and DLB **(Chap. 426)**, more rapid degenerative dementias with prominent movement abnormalities, are more likely to be mistaken for CJD. The differential diagnosis for CJD includes other rapidly progressive dementing conditions such as viral or bacterial encephalitides, Hashimoto's encephalopathy, central nervous system (CNS) vasculitis, lymphoma, or paraneoplastic/autoimmune syndromes. The markedly abnormal periodic complexes on EEG and cortical ribboning and basal ganglia hyperintensities on fluid-attenuate inversion recovery MRI are diagnostic features of CJD, although rarely, prolonged focal or generalized seizures can produce a similar imaging appearance. **Prion diseases are discussed in detail in Chap. 430.**

Huntington's disease (HD) **(Chap. 428)** is an autosomal dominant degenerative brain disorder. HD clinical hallmarks include chorea, behavioral disturbance, and executive impairment. Symptoms typically begin in the fourth or fifth decade, but there is a wide range, from childhood to >70 years. Memory is frequently not impaired until late in the disease, but attention, judgment, self-awareness, and executive functions are often deficient at an early stage. Depression, apathy, social withdrawal, irritability, and intermittent disinhibition are common. Delusions and obsessive-compulsive behavior may occur. Disease duration is variable but typically lasts ~15 years.

Normal-pressure hydrocephalus is a relatively uncommon but treatable syndrome. The clinical, physiologic, and neuroimaging characteristics of NPH must be carefully distinguished from those of other dementias associated with gait impairment. Historically, many patients treated for NPH have suffered from other dementias, particularly AD, vascular dementia, DLB, and PSP. For NPH, the clinical triad includes an abnormal gait (ataxic or apractic), dementia (usually mild to moderate, with an emphasis on executive impairment), and urinary urgency or incontinence. Neuroimaging reveals enlarged lateral ventricles (hydrocephalus) with little or no cortical atrophy, although the sylvian fissures may appear propped open (so-called "boxcarring"), which can be mistaken for perisylvian atrophy. Crowding of dorsal frontal-parietal gyri helps distinguish NPH from other movement disorders, such as PSP and CBD, in which dorsal atrophy with sulcal widening is common. NPH is a communicating hydrocephalus with a patent aqueduct of Sylvius **(see Fig. 25-3)**, in contrast to aqueductal stenosis, in which the aqueduct is small. Lumbar puncture opening pressure falls in the high-normal range, and the CSF protein, glucose, and cell counts are normal. NPH may be caused by obstruction to normal CSF flow over the cerebral convexities and delayed resorption into the venous system. The indolent nature of the process results in enlarged lateral ventricles with relatively little increase in CSF pressure. Presumed edema, stretching, and distortion of subfrontal white matter tracts may lead to clinical symptoms, but the precise underlying pathophysiology remains unclear. Some patients provide a history of conditions that produce meningeal scarring (blocking CSF resorption)

such as previous meningitis, subarachnoid hemorrhage, or head trauma. Others with long-standing but asymptomatic congenital hydrocephalus may have adult-onset deterioration in gait or memory that is confused with NPH. In contrast to AD, the patient with NPH complains of an early and prominent gait disturbance without cortical atrophy on CT or MRI.

Numerous attempts to improve NPH diagnosis with various special studies and predict the success of ventricular shunting have been undertaken. These tests include radionuclide cisternography (showing a delay in CSF absorption over the convexity) and various efforts to monitor and alter CSF flow dynamics, including a constant-pressure infusion test. None has proven to be specific or consistently useful. A transient improvement in gait or cognition may follow lumbar puncture (or serial punctures) with removal of 30–50 mL of CSF, but this finding has also not proved to be consistently predictive of postshunt improvement. Perhaps the most reliable strategy is a period of close inpatient evaluation before, during, and after lumbar CSF drainage. Occasionally, when a patient with AD presents with gait impairment (at times due to comorbid subfrontal vascular injury) and absent or only mild cortical atrophy on CT or MRI, distinguishing NPH from AD can be challenging. Hippocampal atrophy on MRI favors AD, whereas a characteristic "magnetic" gait with external hip rotation, low foot clearance, and short strides, along with prominent truncal sway or instability, favors NPH. The diagnosis of NPH should be avoided when hydrocephalus is not detected on imaging studies, even if the symptoms otherwise fit. Thirty to fifty percent of patients identified by careful diagnosis as having NPH will improve with ventricular shunting. Gait may improve more than cognition, but many reported failures to improve cognitively may have resulted from comorbid AD. Short-lasting improvement is common. Patients should be carefully selected for shunting, because subdural hematoma, infection, and shunt failure are known complications and can be a cause for early nursing home placement in an elderly patient with previously mild dementia.

Intracranial hypotension, sometimes called sagging brain syndrome, is a disorder caused by low CSF pressure, leading to downward pressure on the subcortical structures and disruption of cerebral function. It presents in a variable manner with headache, often exacerbated by coughing or a Valsalva maneuver or by moving from lying to standing. Other common symptoms include dizziness, vomiting, disruption of sleep-wake cycles, and sometimes a progressive behavioral variant FTD-like syndrome **(Chap. 424)**. Although sometimes idiopathic, this syndrome can be caused by CSF leaks secondary to lumbar puncture, head trauma, or spinal cord arachnoid cysts. Treatment consists of finding and patching CSF leaks.

Dementia can accompany *chronic alcoholism* **(Chap. 445)** and may result from associated malnutrition, especially of B vitamins, particularly thiamine. Other poorly defined aspects of chronic alcoholism may, however, also produce cerebral damage. A rare idiopathic syndrome of dementia and seizures with degeneration of the corpus callosum has been reported primarily in male Italian red wine drinkers (Marchiafava-Bignami disease).

Thiamine (vitamin B_1) deficiency causes Wernicke's encephalopathy **(Chap. 301)**. The clinical presentation is usually a malnourished patient (frequently but not necessarily alcoholic) with confusion, ataxia, and diplopia resulting from inflammation and necrosis of periventricular midline structures, including dorsomedial thalamus, mammillary bodies, midline cerebellum, periaqueductal gray matter, and trochlear and abducens nuclei. Damage to the dorsomedial thalamus correlates most closely with the memory loss. Prompt administration of parenteral thiamine (100 mg intravenously for 3 days followed by daily oral dosage) may reverse the disease if given within the first days of symptom onset. Prolonged untreated thiamine deficiency can result in an irreversible and profound amnestic syndrome (Korsakoff's syndrome) or even death.

In *Korsakoff's syndrome*, the patient is unable to recall new information despite normal immediate memory, attention span, and level of consciousness. Memory for new events is seriously impaired, whereas knowledge acquired prior to the illness remains relatively intact. Patients are easily confused, disoriented, and cannot store information for more than a few minutes. Superficially, they may be conversant, engaging, and able to perform simple tasks and follow immediate commands. Confabulation is common, although not always present. There is no specific treatment because the previous thiamine deficiency has produced irreversible damage to the medial thalamic nuclei and mammillary bodies. Mammillary body atrophy may be visible on MRI in the chronic phase **(see Fig. 301-6)**.

Vitamin B_{12} deficiency, as can occur in pernicious anemia, causes a megaloblastic anemia and may also damage the nervous system **(Chaps. 95 and 434)**. Neurologically, it most commonly produces a spinal cord syndrome (myelopathy) affecting the posterior columns (loss of vibration and position sense) and corticospinal tracts (hyperactive tendon reflexes with Babinski signs); it also damages peripheral nerves (neuropathy), resulting in sensory loss with depressed tendon reflexes. Damage to myelinated axons may also cause dementia. The mechanism of neurologic damage is unclear but may be related to a deficiency of *S*-adenosyl methionine (required for methylation of myelin phospholipids) due to reduced methionine synthase activity or accumulation of methylmalonate, homocysteine, and propionate, providing abnormal substrates for fatty acid synthesis in myelin. Use of histamine blockers or metformin, vegan diets, autoimmunity against gastric

parietal cells, and various causes of malabsorption are the typical causes for vitamin B_{12} deficiency. The neurologic sequelae of vitamin B_{12} deficiency may occur in the absence of hematologic manifestations, making it critical to avoid using the complete blood count (CBC) and blood smear as a substitute for measuring B_{12} blood levels. Treatment with parenteral vitamin B_{12} (1000 µg intramuscularly daily for a week, weekly for a month, and monthly for life for pernicious anemia) stops progression of the disease if instituted promptly, but complete reversal of advanced nervous system damage will not occur.

Deficiency of nicotinic acid (*pellagra*) is associated with skin rash over sun-exposed areas, glossitis, and angular stomatitis **(Chap. 326)**. Severe dietary deficiency of nicotinic acid along with other B vitamins such as pyridoxine may result in spastic paraparesis, peripheral neuropathy, fatigue, irritability, and dementia. This syndrome has been seen in prisoners of war and in concentration camps but should be considered in any malnourished individual. Low serum folate levels appear to be a rough index of malnutrition, but isolated folate deficiency has not been proved as a specific cause of dementia.

CNS infections usually cause delirium and other acute neurologic syndromes. However, some chronic CNS infections, particularly those associated with chronic meningitis **(Chap. 134)**, may produce a dementing illness. The possibility of chronic infectious meningitis should be suspected in patients presenting with a dementia or behavioral syndrome, who also have headache, meningismus, cranial neuropathy, and/or radiculopathy. Between 20 and 30% of patients in the advanced stages of HIV infection become demented **(Chap. 197)**. Cardinal features include psychomotor retardation, apathy, and impaired memory. This syndrome may result from secondary opportunistic infections but can also be caused by direct infection of CNS neurons with HIV. Neurosyphilis **(Chap. 177)** was a common cause of dementia in the preantibiotic era; it is now uncommon but can still be encountered in patients with multiple sex partners, particularly among patients with HIV. Characteristic CSF changes consist of pleocytosis, increased protein, and a positive Venereal Disease Research Laboratory (VDRL) test.

Primary and metastatic *neoplasms of the CNS* **(Chap. 86)** usually produce focal neurologic findings and seizures rather than dementia, but if tumor growth begins in the frontal or temporal lobes, the initial manifestations may be memory loss or behavioral changes. A paraneoplastic syndrome of dementia associated with occult carcinoma (often small-cell lung cancer) is termed *limbic encephalitis*. In this syndrome, confusion, agitation, seizures, poor memory, emotional changes, and frank dementia may occur. Paraneoplastic *encephalitis associated with NMDA receptor antibodies* presents as a progressive psychiatric disorder with memory loss and seizures; affected patients are often young women with ovarian teratoma **(Chap. 90)**.

A *nonconvulsive seizure disorder* **(Chap. 418)** may underlie a syndrome of confusion, clouding of consciousness, and garbled speech. Often, psychiatric disease is suspected, but an EEG demonstrates the epileptic nature of the illness. If recurrent or persistent, the condition may be termed *complex partial status epilepticus*. The cognitive disturbance often responds to anticonvulsant therapy. The etiology may be previous small strokes or head trauma; some cases are idiopathic. Nonconvulsive temporal lobe seizures can also emerge early in the course of AD.

It is important to recognize *systemic diseases* that indirectly affect the brain and produce chronic confusion or dementia. Such conditions include hypothyroidism; vasculitis; and hepatic, renal, or pulmonary disease. Hepatic encephalopathy may begin with irritability and confusion and slowly progress to agitation, lethargy, and coma.

Isolated vasculitis of the CNS (CNS granulomatous angiitis) **(Chaps. 356 and 419)** occasionally causes a chronic encephalopathy associated with confusion, disorientation, and clouding of consciousness. Headache is common, and strokes and cranial neuropathies may occur. Brain imaging studies may be normal or nonspecifically abnormal. CSF analysis reveals a mild pleocytosis or protein elevation. Cerebral angiography can show multifocal stenoses involving medium-caliber vessels, but some patients have only small-vessel disease that is not revealed on angiography. The angiographic appearance is not specific and may be mimicked by atherosclerosis, infection, or other causes of vascular disease. Brain or meningeal biopsy demonstrates endothelial cell proliferation and mononuclear infiltrates within blood vessel walls. The prognosis is often poor, although the disorder may remit spontaneously. Some patients respond to glucocorticoids or chemotherapy.

Chronic metal exposure represents a rare cause of dementia. The key to diagnosis is to elicit a history of exposure at work or home. Chronic lead poisoning from inadequately fire-glazed pottery has been reported. Fatigue, depression, and confusion may be associated with episodic abdominal pain and peripheral neuropathy. Gray lead lines appear in the gums, usually accompanied by an anemia with basophilic stippling of red blood cells. The clinical presentation can resemble that of acute intermittent porphyria, including elevated levels of urine porphyrins as a result of the inhibition of δ-aminolevulinic acid dehydrase. The treatment is chelation therapy with agents such as ethylenediamine tetraacetic acid (EDTA). Chronic mercury poisoning produces dementia, peripheral neuropathy, ataxia, and tremulousness that may progress to a cerebellar intention tremor or choreoathetosis. The confusion and memory loss of chronic arsenic intoxication is also associated with nausea, weight loss, peripheral neuropathy, pigmentation and scaling of the skin, and transverse white lines of the fingernails (Mees' lines).

Treatment is chelation therapy with dimercaprol (BAL). Aluminum poisoning is rare but was documented with the dialysis dementia syndrome, in which water used during renal dialysis was contaminated with excessive amounts of aluminum. This poisoning resulted in a progressive encephalopathy associated with confusion, nonfluent aphasia, memory loss, agitation, and, later, lethargy and stupor. Speech arrest and myoclonic jerks were common and associated with severe and generalized EEG changes. The condition has been eliminated by the use of deionized water for dialysis.

Recurrent head trauma in professional athletes may lead to a dementia previously referred to as "punch-drunk" syndrome or *dementia pugilistica* but now known as chronic traumatic encephalopathy (CTE) to signify its relevance to contact sport athletes other than boxers **(Chap. 435)**. The symptoms can be progressive, beginning late in an athlete's career or, more often, after retirement. Early in the course, a personality change associated with social instability and sometimes paranoia and delusions occurs. Later, memory loss progresses to full-blown dementia, often associated with parkinsonian signs and ataxia or intention tremor. At autopsy, the cerebral cortex shows tau-immunoreactive NFTs that are more prominent than amyloid plaques (which are usually diffuse or absent rather than neuritic). NFTs and tau-positive reactive astrocytes are often clustered in the depths of cortical sulci and in a perivascular distribution. TDP-43 inclusions **(Chap. 424)** have also been reported, highlighting the overlap with the FTD spectrum. Loss of neurons in the substantia nigra is a variable feature, and some with TDP-43 inclusions also develop motor neuron disease (MND) **(Chap. 429)**.

Chronic subdural hematoma **(Chap. 435)** is also occasionally associated with dementia, often in the context of underlying cortical atrophy from conditions such as AD or HD.

Transient global amnesia (TGA) is characterized by the sudden onset of a severe episodic memory deficit, usually occurring in persons aged >50 years. Often the amnesia occurs in the setting of an emotional stimulus or physical exertion. During the attack, the individual is alert and communicative, general cognition seems intact, and there are no other neurologic signs or symptoms. The patient may seem confused and repeatedly ask about his or her location in place and time. The ability to form new memories returns after a period of hours, and the individual returns to normal with no recall for the period of the attack. Frequently no cause is determined, but cerebrovascular disease, epilepsy (7% in one study), migraine, or cardiac arrhythmias have all been implicated. Approximately one-quarter of patients experience recurrent attacks. Rare instances of permanent memory loss have been reported in patients with TGA-like spells, usually representing ischemic infarction of the hippocampus or dorsomedial thalamic nucleus bilaterally. Seizure activity due to AD should always be suspected in this syndrome.

The *ALS/parkinsonian/dementia complex of Guam* is a rare degenerative disease that has occurred in the Chamorro natives on the island of Guam. Individuals may have any combination of parkinsonian features, dementia, and MND. The most characteristic pathologic features are the presence of NFTs in degenerating neurons of the cortex and substantia nigra and loss of motor neurons in the spinal cord, although recent reanalysis has shown that some patients with this illness also show coexisting TDP-43 pathology. Epidemiologic evidence supports a possible environmental cause, such as exposure to a neurotoxin or an infectious agent with a long latency period. One interesting but unproven candidate neurotoxin is the seed of the false palm tree, which Guamanians traditionally used to make flour. The amyotrophic lateral sclerosis (ALS) syndrome is no longer present in Guam, but a dementing illness with rigidity continues to be seen.

Rarely, adult-onset leukodystrophies, lysosomal storage diseases, and other genetic disorders can present as a dementia in middle to late life. Metachromatic leukodystrophy (MLD) causes a progressive psychiatric or dementia syndrome associated with an extensive, confluent frontal white matter abnormality. MLD is diagnosed by measuring reduced arylsulfatase A enzyme activity in peripheral white blood cells. Adult-onset presentations of adrenoleukodystrophy have been reported in female carriers, and these patients often feature spinal cord and posterior white matter involvement. Adrenoleukodystrophy is diagnosed by demonstrating increased levels of plasma very-long-chain fatty acids. CADASIL is another genetic syndrome associated with white matter disease, often frontally and temporally predominant. Diagnosis is made with skin biopsy, which shows osmophilic granules in arterioles, or, increasingly, through genetic testing for mutations in Notch 3. The neuronal ceroid lipofuscinoses are a genetically heterogeneous group of disorders associated with myoclonus, seizures, vision loss, and progressive dementia. Diagnosis is made by finding eosinophilic curvilinear inclusions within white blood cells or neuronal tissue.

Psychogenic amnesia for personally important memories can be seen. Whether this results from deliberate avoidance of unpleasant memories, outright malingering, or unconscious repression remains unknown and probably depends on the patient. Event-specific amnesia is more likely to occur after violent crimes such as homicide of a close relative or friend or sexual abuse. It may develop in association with severe drug or alcohol intoxication and sometimes with schizophrenia. More prolonged psychogenic amnesia occurs in fugue states that also commonly follow severe emotional stress. The patient with a fugue state suffers from a sudden loss of personal identity and may be found wandering far from home. *In contrast to neurologic amnesia, fugue states are associated with amnesia for personal identity and events closely associated with the personal past.* At the same time, memory for other recent events and the ability to learn and use new information

are preserved. The episodes usually last hours or days and occasionally weeks or months while the patient takes on a new identity. On recovery, there is a residual amnesia gap for the period of the fugue. Very rarely does selective loss of autobiographic information reflect a focal injury to the brain areas involved with these functions.

Psychiatric diseases may mimic dementia. Severely depressed or anxious individuals may appear demented, a phenomenon sometimes called *pseudodementia*. Memory and language are usually intact when carefully tested, and a significant memory disturbance usually suggests an underlying dementia, even if the patient is depressed. Patients in this condition may feel confused and unable to accomplish routine tasks. Vegetative symptoms, such as insomnia, lack of energy, poor appetite, and concern with bowel function, are common. Onset is often more abrupt, and the psychosocial milieu may suggest prominent reasons for depression. Such patients respond to treatment of the underlying psychiatric illness. Schizophrenia is usually not difficult to distinguish from dementia, but occasionally the distinction can be problematic. Schizophrenia generally has a much earlier age of onset (second and third decades) than most dementing illnesses and is associated with intact memory. The delusions and hallucinations of schizophrenia are usually more complex, bizarre, and threatening than those of dementia. Some chronic schizophrenics develop an unexplained progressive dementia late in life that is not related to AD. Conversely, FTD, HD, vascular dementia, DLB, AD, or leukoencephalopathy can begin with schizophrenia-like features, leading to the misdiagnosis of a psychiatric condition. Later age of onset, significant deficits on cognitive testing, or the presence of abnormal neuroimaging suggest a degenerative condition. Memory loss may also be part of a *conversion disorder*. In this situation, patients commonly complain bitterly of memory loss, but careful cognitive testing either does not confirm the deficits or demonstrates inconsistent or unusual patterns of cognitive problems. The patient's behavior and "wrong" answers to questions often indicate that he or she understands the question and knows the correct answer.

Clouding of cognition by *chronic drug or medication use*, often prescribed by physicians, is an important cause of dementia. Sedatives, tranquilizers, and analgesics used to treat insomnia, pain, anxiety, or agitation may cause confusion, memory loss, and lethargy, especially in the elderly. Discontinuation of the offending medication often improves mentation.

FURTHER READING

Braak H, Del Tredici K: Where, when, and in what form does sporadic Alzheimer's disease begin? Curr Opin Neurol 25:708, 2012. [PubMed: 23160422]

Cohen AD, Klunk WE: Early detection of Alzheimer's disease using PiB and FDG PET. Neurobiol Dis 72 Pt A:117, 2014. [PubMed: 24825318]

Haas C: Strategies, development, and pitfalls of therapeutic options for Alzheimer's disease. J Alzheimers Dis 28:241, 2012. [PubMed: 21987594]

Musiek ES, Holtzman DM: Origins of Alzheimer's disease: Reconciling cerebrospinal fluid biomarker and neuropathology data regarding the temporal sequence of amyloid-beta and tau involvement. Curr Opin Neurol 25:715, 2012. [PubMed: 23041958]

Selkoe DJ, Hardy J: The amyloid hypothesis of Alzheimer's disease at 25 years. EMBO Mol Med 8:595, 2016. [PubMed: 27025652]

Chapter 424: Frontotemporal Dementia

William W. Seeley; Bruce L. Miller

INTRODUCTION

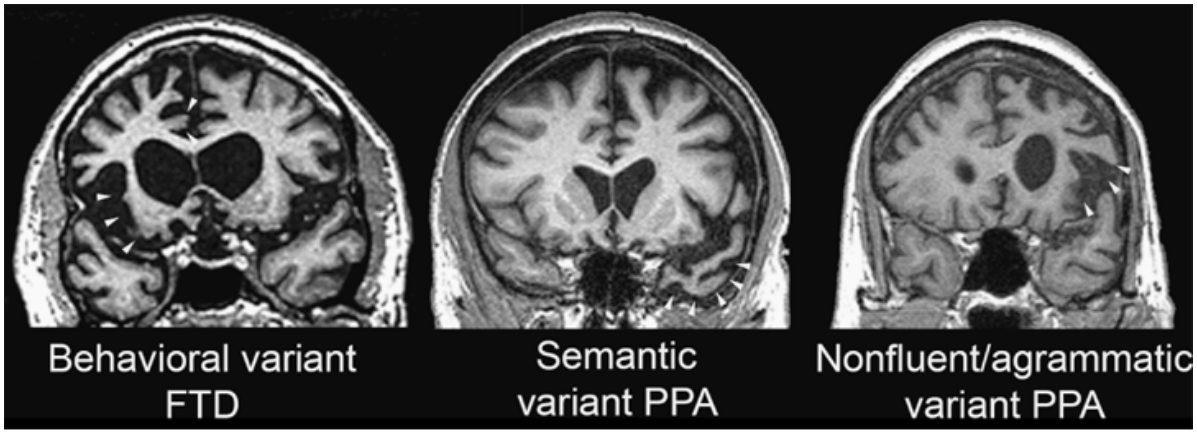
Frontotemporal dementia (FTD) refers to a group of clinical syndromes united by their links to underlying frontotemporal lobar degeneration (FTLD) pathology. FTD most often begins in the fifth to seventh decades and is nearly as prevalent as AD in this age group. Early studies suggested that FTD may be more common in men than women, however more recent reports cast doubt on this finding. Although a family history of dementia is common, autosomal dominant inheritance is seen in only 10–20% of all FTD cases.

CLINICAL MANIFESTATIONS

The clinical heterogeneity seen in both familial and sporadic forms of FTD is remarkable. Three core clinical syndromes have been described (Fig. 424-1). In the behavioral variant (bvFTD), the most common FTD syndrome, social and emotional systems dysfunction manifests as apathy, disinhibition, compulsivity, loss of empathy, and overeating, often but not always accompanied by deficits in executive control. Two forms of primary progressive aphasia (PPA), the semantic and nonfluent/agrammatic variants, are commonly due to FTLD and included under the FTD umbrella. In the semantic variant, patients slowly lose the ability to decode word, object, person-specific, and emotion meaning, whereas patients with the nonfluent/agrammatic variant develop profound inability to produce words, often with prominent motor speech impairment. Any of these three clinical syndromes, but most often bvFTD, may be accompanied by motor neuron disease (MND) (Chap. 429), in which case the term FTD-MND is applied. In addition, the corticobasal syndrome (CBS) and progressive supranuclear palsy syndrome (PSP-S) can be considered part of the FTLD clinical spectrum. Furthermore, patients may evolve from any of the major syndromes described above to have prominent features of another syndrome.

FIGURE 424-1

Three major frontotemporal dementia (FTD) clinical syndromes. Coronal magnetic resonance imaging sections from representative patients with behavioral variant FTD (*left*) and the semantic (*center*) and nonfluent/agrammatic (*right*) variants of primary progressive aphasia (PPA). Areas of early and severe atrophy in each syndrome are highlighted (*white arrowheads*). The behavioral variant features anterior cingulate and frontoinsular atrophy, spreading to orbital and dorsolateral prefrontal cortex. Semantic variant PPA shows prominent temporopolar atrophy, more often on the left. Nonfluent/agrammatic variant PPA is associated with dominant frontal opercular and dorsal insula degeneration.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Findings at the bedside are dictated by the anatomic localization of the disorder. Medial and orbital frontal and anterior insula degeneration predicts bvFTD. Patients with nonfluent/agrammatic PPA show dominant hemisphere lateral frontal and precentral gyrus atrophy. Anterior temporal degeneration presents with semantic variant PPA. Parietal functions such as visuospatial processing and arithmetic may remain normal late into any FTD syndrome. Many patients with nonfluent aphasia or bvFTD later develop aspects of PSP-S, as disease spreads into subcortical or brainstem structures, or CBS-like features, as disease moves into peri-rolandic cortices.

GENETIC CONSIDERATIONS



The most common autosomal dominantly inherited mutations causing FTD involve the *C9ORF72* (chromosome 9), *GRN* (chromosome 17), and *MAPT* (chromosome 17) genes. Hexanucleotide (GGGGCC) expansions in a noncoding exon of *C9ORF72* are the most recently identified and represent the most common genetic cause of familial or sporadic FTD (usually presenting as bvFTD with or without MND) and amyotrophic lateral sclerosis (ALS). The expansion is associated with *C9ORF72* haploinsuffiency, nuclear mRNA foci containing transcribed portions of the expansion and other mRNAs, neuronal cytoplasmic inclusions containing dipeptide repeat proteins translated from the repeat mRNA, and transactive response DNA-binding protein of 43 kDa (TDP-43) neuronal cytoplasmic and glial inclusions. The pathogenic significance of these various features is a topic of vigorous investigation. *MAPT* mutations lead to a change in the alternate splicing of tau or cause loss of function in the tau molecule, thereby altering microtubule binding. With *GRN*, mutations in the coding sequence of the gene encoding progranulin protein result in mRNA degradation due to nonsense-mediated decay, leading to a ~50% reduction in circulating progranulin protein levels. Intriguingly, homozygous *GRN* mutations were recently reported to cause neuronal ceroid lipofuscinosis, focusing investigators on the lysosome as a site of molecular dysfunction in *GRN*-related FTD. Progranulin is a growth factor that binds to tumor necrosis factor (TNF) receptors and participates in tissue repair and tumor growth. How progranulin mutations lead to FTD remains unknown, but the most likely mechanisms include lysosomal dysfunction and enhanced neuroinflammation. Both *MAPT* and *GRN* mutations can be associated with parkinsonian features, whereas ALS is rare. Infrequently, mutations in the valosin-containing protein (*VCP*, chromosome 9), TANK binding kinase 1 (TBK-1), T cell-restricted intracellular antigen-1 (TIA1), and charged multivesicular body protein 2b (*CHMP2b,* chromosome 3) genes also lead to autosomal dominant familial FTD. Mutations in the *TARDBP* (encoding TDP-43) and *FUS* (encoding fused in sarcoma [FUS]) genes (see below) cause familial ALS, sometimes in association with an FTD syndrome, although a few patients presenting with FTD alone have been reported.

NEUROPATHOLOGY

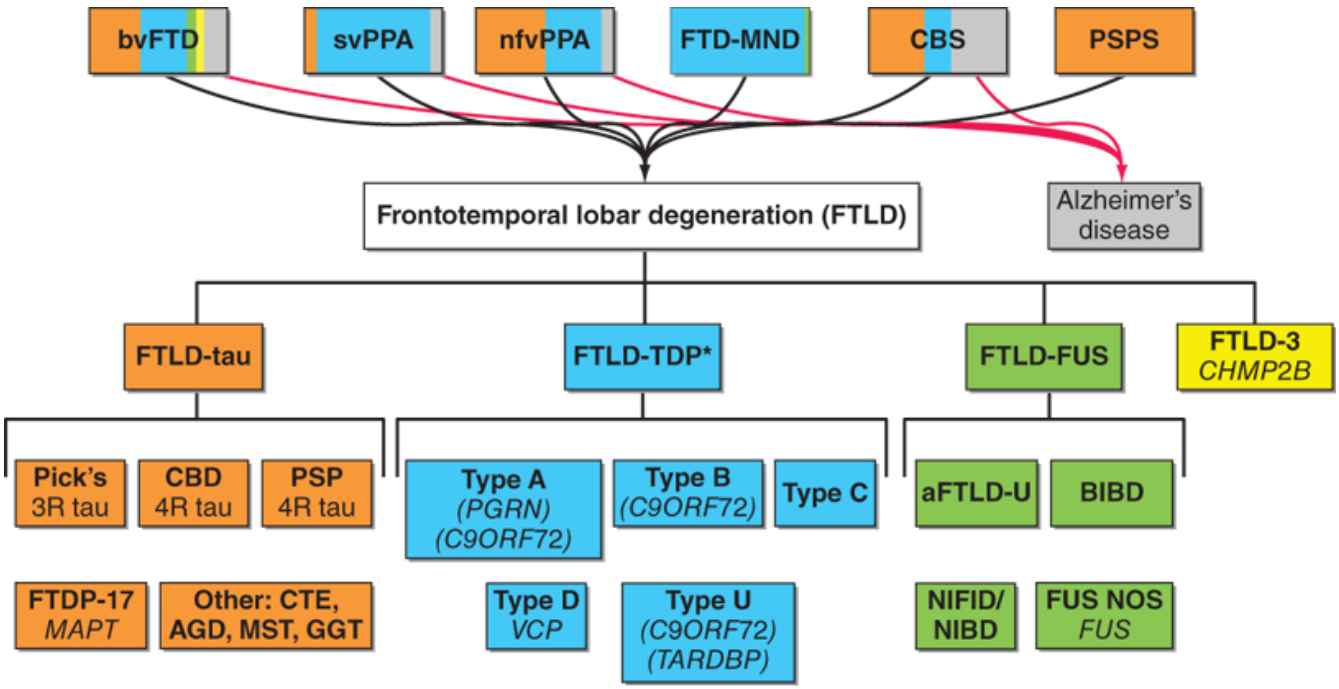
The gross pathologic hallmark of FTLD is a focal atrophy of frontal, insular, and/or temporal cortex, which can be visualized with neuroimaging studies (Fig. 424-1) and is often profound at autopsy. Despite the appearance of advanced disease, however, imaging studies suggest that atrophy often begins focally in one hemisphere before spreading to anatomically interconnected cortical and

subcortical regions. Loss of cortical serotonergic innervation is seen in many patients. In contrast to AD, the cholinergic system is relatively spared in FTD, which accounts for the poor efficacy of acetylcholinesterase inhibitors in this group.

Although early studies suggested that 15–30% of patients with FTD showed underlying AD at autopsy, progressive refinement in clinical diagnosis has improved pathologic prediction accuracy, and most patients diagnosed with FTD at a dementia clinic with expertise in FTD will show underlying FTLD pathology. Microscopic findings seen across all patients with FTLD include gliosis, microvacuolation, and neuronal loss, but the disease is subtyped according to the protein composition of neuronal and glial inclusions, which contain either tau or TDP-43 in ~90% of patients, with the remaining ~10% showing inclusions containing FUS (Fig. 424-2).

FIGURE 424-2

Frontotemporal dementia syndromes are united by underlying frontotemporal lobar degeneration pathology, which can be divided according to the presence of tau, TDP-43, or FUS-containing inclusions in neurons and glia. Correlations between clinical syndromes and major molecular classes are shown with colored shading. Despite improvements in clinical syndromic diagnosis, a small percentage of patients with some frontotemporal dementia syndromes will show Alzheimer's disease neuropathology at autopsy (*gray shading*). aFTLD-U, atypical frontotemporal lobar degeneration with ubiquitin-positive inclusions; AGD, argyrophilic grain disease; BIBD, basophilic inclusion body disease; bvFTD, behavioral variant frontotemporal dementia; CBD, corticobasal degeneration; CBS, corticobasal syndrome; CTE, chronic traumatic encephalopathy; FTD-MND, frontotemporal dementia with motor neuron disease; FTDP-17, frontotemporal dementia with parkinsonism linked to chromosome 17; FUS, fused in sarcoma; GGT, globular glial tauopathy; MST, multisystem tauopathy; nfvPPA, nonfluent/agrammatic variant primary progressive aphasia; NIBD, neurofilament inclusion body disease; NIFID, neuronal intermediate filament inclusion disease; PSP, progressive supranuclear palsy; PSP-S, progressive supranuclear palsy syndrome; svPPA, semantic variant primary progressive aphasia; Type U, unclassifiable type.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PATHOGENESIS

The toxicity and spreading capacity of misfolded tau underlies the pathogenesis of many familial cases and is emerging as a key factor in sporadic tauopathies, although loss of tau microtubule stabilizing function may also play a role. TDP-43 and FUS, in contrast, are RNA/DNA binding proteins whose roles in neuronal function are still being actively investigated, but one key role may be the chaperoning of mRNAs to the distal neuron for activity-dependent translation within dendritic spines. Because these proteins also

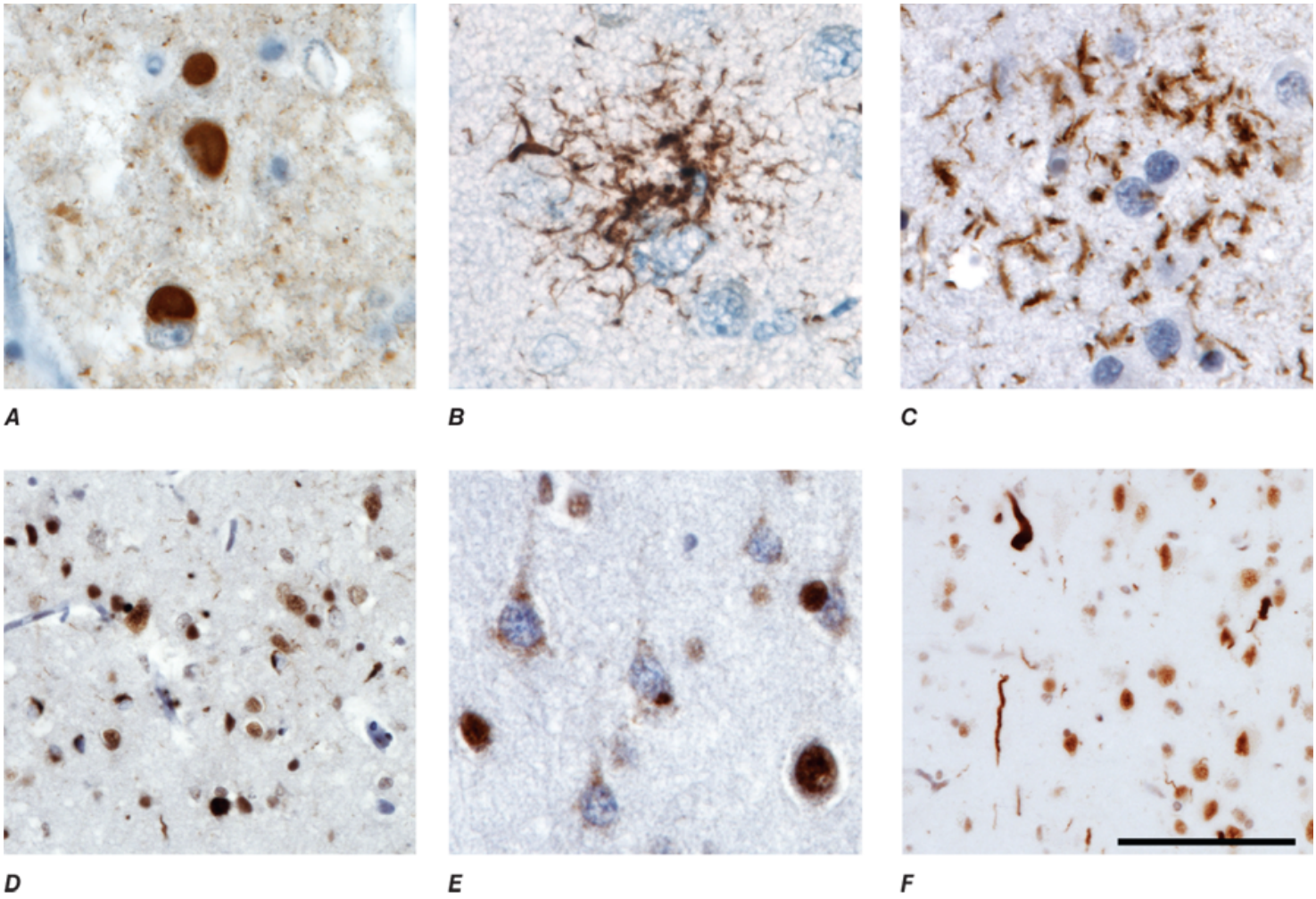
form intracellular aggregates and produce similar anatomic progression, protein toxicity and spreading may also factor heavily in the pathogenesis of FTLD-TDP and FTLD-FUS.

Increasingly, misfolded proteins in neurodegenerative disease are being recognized as having "prion-like" properties in that they can template the misfolding of their natively folded (or unfolded) protein counterparts, a process that creates exponential amplification of protein misfolding within a cell and may promote transcellular and even transsynaptic protein propagation between cells. This hypothesis could provide a unifying explanation for the stereotypical patterns of disease spread observed in each syndrome **(Chap. 417)**.

Although the term *Pick's disease* was once used to describe a progressive degenerative disorder characterized by selective involvement of the anterior frontal and temporal neocortex and pathologically by intraneuronal cytoplasmic inclusions (*Pick bodies*), it is now used only in reference to a specific FTLD-tau histopathologic entity. Classical Pick bodies are argyrophilic, staining positively with the Bielschowsky silver method (but not with the Gallyas method) and also with immunostaining for hyperphosphorylated tau. Recognition of the three FTLD major molecular classes has allowed delineation of distinct FTLD subtypes within each class. These subtypes, based on the morphology and distribution of the neuronal and glial inclusions **(Fig. 424-3)**, account for the vast majority of patients, and some subtypes show strong clinical or genetic associations (Fig. 424-2). Despite this progress, clinical features do not allow reliable prediction of the underlying FTLD subtype, or even the major molecular class, for all clinical syndromes. Molecular PET imaging with ligands chosen to bind misfolded tau protein shows great promise and is already being applied to the study of patients with AD and FTD. Because FTLD-tau and FTLD-TDP account for 90% of FTLD patients, the ability to detect pathological tau protein deposition in vivo would greatly improve prediction accuracy, especially when amyloid PET imaging is negative.

FIGURE 424-3

Neuropathology in frontotemporal lobar degeneration (FTLD). FTLD-tau (***A–C***) and FTLD-TDP (***D–F***) account for >90% of patients with FTLD, and immunohistochemistry reveals characteristic lesions in each of the major histopathologic subtypes within each class: ***A.*** Pick bodies in Pick's disease; ***B.*** a tufted astrocyte in progressive supranuclear palsy; ***C.*** an astrocytic plaque in corticobasal degeneration; ***D.*** small compact or crescentic neuronal cytoplasmic inclusions and short, thin neuropil threads in FTLD-TDP, type A; ***E.*** diffuse/granular neuronal cytoplasmic inclusions (with a relative paucity of neuropil threads) in FTLD-TDP, type B; and ***F.*** long, tortuous dystrophic neurites in FTLD-TDP, type C. TDP can be seen within the nucleus in neurons lacking inclusions but mislocalizes to the cytoplasm and forms inclusions in FTLD-TDP. Immunostains are 3-repeat tau (***A***), phospho-tau (***B*** and ***C***), and TDP-43 (***D–F***). Sections are counterstained with hematoxylin. Scale bar applies to all panels and represents 50 µm in ***A, B, C***, and ***E*** and 100 µm in ***D*** and ***F***.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

The burden on caregivers of patients with FTD is extremely high, especially when the illness disrupts core emotional and personality functions of the loved one. Treatment is symptomatic, and there are currently no therapies known to slow progression or improve symptoms. Many of the behaviors that may accompany FTD, such as depression, hyperorality, compulsions, and irritability, can be ameliorated with antidepressants, especially SSRIs. Because FTD is often accompanied by parkinsonism, antipsychotics, which can exacerbate this problem, must be used with caution. **A general approach to the symptomatic management of dementia is presented in Chap. 25.**

PROGRESSIVE SUPRANUCLEAR PALSY SYNDROME

PSP-S (also known as Richardson syndrome) is a degenerative disorder that involves the brainstem, basal ganglia, diencephalon, and selected areas of cortex. Clinically, PSP-S begins with falls and executive or subtle personality changes (such as mental rigidity, impulsivity, or apathy). Shortly thereafter, a progressive oculomotor syndrome ensues that begins with square wave jerks, followed by slowed saccades (vertical worse than horizontal) before resulting in progressive supranuclear ophthalmoparesis. Dysarthria, dysphagia, and symmetric axial rigidity can be prominent features that emerge at any point in the illness. A stiff, unstable posture with hyperextension of the neck and a slow, jerky, toppling gait are characteristic. Frequent unexplained and sometimes spectacular falls are common secondary to a combination of axial rigidity, inability to look down, and poor judgment. Even once patients have severely limited voluntary eye movements, they retain oculocephalic reflexes (demonstrated using a vertical doll's head maneuver); thus, the oculomotor disorder is supranuclear. The dementia overlaps with bvFTD, featuring apathy, frontal-executive dysfunction,

poor judgment, slowed thought processes, impaired verbal fluency, and difficulty with sequential actions and with shifting from one task to another. These features are common at presentation and often precede the motor syndrome. Some patients with a pathologic diagnosis of PSP begin with a nonfluent aphasia or motor speech disorder and progress to classical PSP-S. Response to L-dopa is limited or absent; no other treatments exist. Death occurs within 5–10 years of onset. Like Pick's disease, increasingly the term PSP is used to refer to a specific histopathologic entity within the FTLD-tau class. In PSP, accumulation of hyperphosphorylated 4-repeat tau is seen within neurons and glia. Tau neuronal inclusions often appear tangle-like and may be large, spherical ("globose") and coarse in subcortical structures. The most prominent involvement is in the subthalamic nucleus, globus pallidus, substantia nigra, locus coeruleus, periaqueductal gray, tectum, oculomotor nuclei, and dentate nucleus of cerebellum. Neocortical tangle-like inclusions, like those in AD, often take on a more flame-shaped morphology, but on electron microscopy, PSP tangles can be shown to consist of straight tubules rather than the paired helical filaments found in AD. Furthermore, PSP is associated with prominent tau-positive glial inclusions, such as tufted astrocytes (Fig. 424-3), coiled oligodendroglial inclusions ("coiled bodies"), or, least often, thorny astrocytes. Most patients with PSP-S show PSP at autopsy, although small numbers will show another tauopathy (corticobasal degeneration [CBD] or Pick's disease; Fig. 424-2).

In addition to its overlap with FTD and CBS (see below), PSP is often confused with idiopathic *Parkinson's disease* (PD). Although elderly patients with PD may have restricted upgaze, they do not develop downgaze paresis or other abnormalities of voluntary eye movements typical of PSP. Dementia occurs in ~20% of patients with PD, often due to the emergence of a full-blown DLB-like syndrome. Furthermore, the behavioral syndromes seen with DLB differ from PSP (see below). Dementia in PD becomes more likely with increasing age, increasing severity of extrapyramidal signs, long disease duration, and the presence of depression. Patients with PD who develop dementia also show cortical atrophy on brain imaging. Neuropathologically, there may be AD-related changes in the cortex or LBD-related α-synuclein inclusions in both the limbic system and cerebral cortex. **PD is discussed in detail in Chap. 427.**

CORTICOBASAL SYNDROME

CBS is a slowly progressive dementia-movement disorder associated with severe degeneration in perirolandic cortex and basal ganglia (substantia nigra and striatopallidum). Patients typically present with asymmetric onset of rigidity, dystonia, myoclonus, and apraxia of one limb, at times associated with *alien limb* phenomena in which the limb exhibits unintended motor actions such as grasping, groping, drifting, or undoing. Eventually CBS becomes bilateral and leads to dysarthria, slow gait, action tremor, and a frontal-predominant dementia. Whereas CBS refers to the clinical syndrome, CBD refers to a specific histopathological FTLD-tau entity (Fig. 424-2). Although CBS was once thought to be pathognomonic for CBD, increasingly it has been recognized that CBS can be due to CBD, PSP, FTLD-TDP, or even AD, which accounts for up to 30% of CBS in some series. In CBD, the microscopic features include ballooned, achromatic, tau-positive neurons; astrocytic plaques (Fig. 424-3); and other dystrophic glial tau pathomorphologies that overlap with those seen in PSP. Most specifically, CBD features a severe tauopathy burden in the subcortical white matter, consisting of axonal threads and oligodendroglial coiled bodies. As shown in Fig. 424-2, patients with bvFTD, nonfluent/agrammatic PPA, and PSP-S may also show CBD at autopsy, emphasizing the importance of distinguishing clinical and pathologic constructs and terminology. Treatment of CBS remains symptomatic; no disease-modifying therapies are available.

FURTHER READING

Irwin DJ et al: Frontotemporal lobar degeneration: Defining phenotypic diversity through personalized medicine. Acta Neuropathol 129:469, 2015. [PubMed: 25549971]

Mackenzie I et al: Nomenclature and nosology for neuropathologic subtypes of frontotemporal lobar degeneration: An update. Acta Neuropathol 119:1, 2010. [PubMed: 19924424]

Olney NT et al: Frontotemporal dementia. Neurol Clin 35:339, 2017. [PubMed: 28410663]

Onyike CU, Diehl-Schmid J: The epidemiology of frontotemporal dementia. Int Rev Psychiatry 25:130, 2013. [PubMed: 23611343]

Roberson ED: Mouse models of frontotemporal dementia. Ann Neurol 72:837, 2012. [PubMed: 23280835]

Chapter 425: Vascular Dementia

Joel Kramer; William W. Seeley; Bruce L. Miller

INTRODUCTION

Vascular cognitive impairment and vascular dementia (VCI-VaD) denote deficits in cognition and behavior, along a spectrum of severity, that are associated with cerebrovascular disease (CVD). A dementia syndrome results when CVD is severe enough to cause significant deficits in occupational, social, or functional abilities. VaD is among the most common causes of dementia in the elderly, although its prevalence is disputed. Vascular disease can disrupt structural cognitive networks with lesions such as microinfarcts, microbleeds, macroinfarcts, large hemorrhages, and chronic progressive white matter degeneration, as well as altered cerebral hemodynamics, such as hypoperfusion, disrupted cerebrovascular autoregulation **(Chap. 301)** neurovascular decoupling (loss of normal hemodynamic responses to neural activity), and blood brain barrier dysfunction. The pathophysiological underpinnings of VCI-VaD remain an active area of research.

Age remains the strongest risk factor for CVD and stroke. By the age of 70, 70% of the population has white matter disease and lesions on neuroimaging, with small infarcts (lacunar infarcts) found in 11–24% of the population. In addition to genetic predisposition, risk factors that directly contribute to CVD include chronic hypertension, hyperlipidemia, diabetes, and smoking. Cardiac disease, such as atrial fibrillation or heart failure, can also cause cognitive impairment via embolic infarcts and hypoxemia due to inadequate cerebral blood flow.

A review of data from across the globe indicates good evidence for variability in CVD and stroke risk. Intracranial atherosclerosis, for example, is higher in Asians, Hispanics, and American blacks than it is in European and American whites, while whites may have more extracranial disease. The causes of these disparities remain under investigation, but likely include genetics, lifestyle, and access to health care.

VaD is strongly associated with hemorrhagic and ischemic strokes, with an estimated one-third of stroke survivors affected by post-stroke dementia or cognitive impairment. Hemorrhages, including subdural, intracerebral and subarachnoid bleeds, account for roughly 20% of all strokes. The disruption of cerebral networks caused by hemorrhage depends to a certain extent on size and location. Subarachnoid hemorrhage (SAH) has a more intricate relation with cognitive deficits. For instance, a history of SAH can triple the lifetime risk of developing a dementia syndrome; the molecular underpinnings of this observation are being actively studied. Of note, hemorrhagic strokes may occur as a result of vessel wall damage and inflammation associated with cerebral amyloid angiopathy (CAA). The build-up of β-amyloid protein in cerebral blood

vessels that increases their susceptibility to rupture. Although CAA is frequently present in patients with Alzheimer's disease (AD) **(Chap. 423)**, it can also be found in the absence of neocortical amyloid plaques or in individuals with specific genetic predispositions and lead to cognitive impairment in the absence of AD.

Ischemic strokes compose 80% of all strokes. Large vessel and small vessel disease (SVD) can lead to dementia, although the mechanisms and clinical presentation vary. In a cross-sectional study of 706 VaD cases, large vessel disease, often referred to as multi-infarct dementia, made up roughly 18% of all cases. Neurobehavioral symptoms vary as a function of cerebral lesion size and location, and can include aphasia, apraxia, agnosia, and inattention syndromes. Some ambiguity between lesion location and cognitive symptoms continues to persist. This may be a result of the interconnected nature of cognitive, behavioral functional, and structural networks of the brain, as well as the remote consequences of cerebral inflammation and blood-brain barrier disruption caused by strokes.

Of all CVD subtypes, chronic cerebral SVD shows the strongest association with cognitive impairment. SVD accounts for 36–67% of all VaDs. Nonetheless, the relationship between SVD and dementia remains complex and the neuropsychology of dementia due to SVD remains controversial. SVD typically causes occlusion of the deep-penetrating arterioles and disease of draining venules, causing damage to subcortical structures such as the basal ganglia, thalami, and white matter tracts. In addition, cerebral microinfarcts, first observed microscopically in autopsy series, then visualized on 7T MRI, have also showed strong associations with cognitive impairment. Larger lesions, referred to as lacunar infarcts, can result in either a stepwise or gradual decline in cognition sometimes referred to as etat lacunaire. Lacunes average 2 mm in volume, but can range from 0.2 to 15 mm.

Although most VCI-VaD cases caused by SVD are sporadic, there are also several genetic SVD-related VaD syndromes. The most prevalent is cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), a genetic disorder linked to a mutation in the *NOTCH3* gene on chromosome 19. CADASIL presents as small vessel strokes, progressive dementia and extensive white matter disease often beginning in mid-adult life **(Chap. 420)**. Through altered extracellular molecular signaling pathways and protein elimination failure and accumulation, pericytes and endothelia of small vessels are involved. CADASIL may offer a unique opportunity to study "pure" VaD, as individuals diagnosed with CADASIL tend to display cognitive decline at an early age, when the likelihood of comorbid neurodegenerative diseases is lower.

In addition to infarcts and hemorrhages, SVD is also associated with blood-brain-barrier compromise, acculmulating white matter disease, and a state of chronic cerebral hypoperfusion, hippocampal atrophy and sclerosis. Neuroimaging markers of SVD, such as white matter hyperintensities (WMH) of presumed vascular origin, microbleeds, cortical microinfarcts, and enlarged Virchow-Robin spaces (eVRS) increase the risk of dementia, with conventional vascular risk factors explaining little of the variance in the absence of these changes. Otto Binswanger described the drastic effect of white matter injury when he observed eight patients with gradual cognitive deterioration and notable white matter changes. Binswanger's disease, commonly referred to as subcortical arteriosclerotic encephalopathy, is considered a prototypical clinical syndrome of VCI and a pathologically homogenous subgroup. On neuroimaging, a progressive confluent

subcortical and periventricular white matter disease is seen (**see Fig. 25-2**), with hypoperfusion and hypometabolism. Novel neuroimaging techniques for assessment of blood brain barrier integrity, such as dynamic contrast-enhanced MRI (DCE-MRI), in combination with biofluid markers, have shown a characteristic unrelenting progressive course of hypoxic injury with inflammatory disruption of blood brain barrier. Individuals with Binswanger's disease typically have hypertension or systemic vascular disease, and the clinical course may include gradual accumulation of focal neurological deficits. Neuropathological features include extensive demyelination and destruction of white matter with relative sparing of the subcortical U fibers. The resulting dementia syndrome is largely dysexecutive. From a cognitive and behavioral standpoint, affected individuals typically display apathy, mood alterations with mild depression, executive dysfunction, and slowed information processing. They can also have marked motor slowing and symmetric parkinsonism.

A strategically placed infarct, usually in the thalamus, basal ganglia or angular gyrus can also result in marked cognitive dysfunction and dementia. For example, a single paramedian artery (artery of Percheron) supplies both anteromedial thalamic regions, and occlusion of this artery can lead to bilateral infarction of the dorsomedial nucleus and the mammillothalamic tracts, resulting in altered mental status, vertical gaze palsy, and memory impairment. Similarly, an infarct in the inferior genu of the internal capsule may strategically disrupt the inferior and medial thalamic peduncles carrying thalamo-cortical fibers important for cognition and memory.

A consensus on frequency, causal factors, underlying neuropathology, clinical symptoms, characteristic neuropsychological presentations, and developmental course of VaD has yet to emerge. Issues related to comorbidity are particularly challenging. A large number of dementia patients with significant CVD show multiple pathologies, most frequently a mix of AD and/or Lewy body disease. Despite variability in reported prevalence, the consistent decrease of pure VaD with age and the related increase in mixed pathologies is well established and reflect the complex interactions between CVD, brain aging, and accumulation of abnormal proteins in neurodegeneration. Unsurprisingly, post-stroke dementia risk factors extensively overlap with those identified for AD, including older age, low education, hypertension, diabetes, smoking, and cerebral atrophy and hippocampal sclerosis. Additionally, concurrent CVD lowers the threshold for dementia due to AD and Lewy body disease **(Chap. 426)**.

Treatment of VaD must be focused on accurate diagnosis of VaD so that new ischemic injury can be prevented by stabilizing or removing the underlying causes. Effective control of modifiable risk factors **(Chap. 2)**, including hypertension **(Chap. 271)**, smoking **(Chap. 448)**, alcohol intake **(Chap. 445)**, sodium consumption, diabetes mellitus **(Chap. 397)**, obesity **(Chap. 395)**, and the metabolic syndrome **(Chap. 401)**, are key to slowing the rising global prevalence of this condition. There is a great need for sensitive and specific in vivo biomarkers of VCI-VaD for early diagnosis and ultimately a surrogate marker of therapeutic efficacy in clinical trials.

FURTHER READING

Gardener H et al: Brain health and shared risk factors for dementia and stroke. Nat Rev Neurol 11:651, 2015. [PubMed: 26481296]

Gorelick PB et al: Vascular contributions to cognitive impairment and dementia. Stroke 42:2672, 2011. [PubMed: 21778438]

McAleese KE et al: Post-mortem assessment in vascular dementia: Advances and aspirations. BMC Medicine 14:129, 2016. [PubMed: 27600683]

Rizzi L et al: Global epidemiology of dementia: Alzheimer's and vascular types. Biomed Res Int 2014:908915, 2014. [PubMed: 25089278]

Rosenberg GA et al: Consensus statement for diagnosis of subcortical small vessel disease. J Cereb Blood Flow Metab 36:6, 2016. [PubMed: 26198175]

Sachdev P et al: Diagnostic criteria for vascular cognitive disorders: A VASCOG statement. Alzheimer Dis Assoc Disord 28:206, 2014. [PubMed: 24632990]

Chapter 426: Dementia with Lewy Bodies

William W. Seeley; Caroline M. Tanner; Bruce L. Miller

INTRODUCTION

Although Lewy body disease (LBD) was once conceptualized as a disease of the substantia nigra, modern human postmortem studies with alpha-synuclein immunohistochemistry revealed that Lewy body and Lewy neurite pathology most often begins in the enteric and autonomic nervous systems before ascending through the brainstem to the substantia nigra, limbic system, and ultimately the cerebral cortex. In other individuals, disease may begin in the olfactory bulb and spread inward through olfactory system connections. These sites of onset, positioned as they are at neural interfaces with the environment, have suggested to some that a toxic environmental exposure may trigger the disease. Individual patients vary in their adherence to these general patterns, and the tempo and topology of progression dictate the clinical syndrome. Some patients with long-standing Parkinson's disease (PD) (Chap. 427) without cognitive impairment slowly develop a dementia that is associated with visual hallucinations and fluctuating alertness. In this scenario, the term *Parkinson's disease dementia* (PDD) is often used. In other patients, dementia and a neuropsychiatric syndrome precede or co-emerge with the parkinsonism, and the patient is diagnosed with *dementia with Lewy bodies* (DLB).

CLINICAL MANIFESTATIONS

The DLB syndrome is characterized by visual hallucinations, parkinsonism, fluctuating alertness, neuroleptic sensitivity, rapid eye movement (REM) sleep behavior disorder (RBD), and often hyposmia and excessive daytime sleepiness. Delusions related to persecution, invasion, and person or place identity (reduplicative paramnesia) are common. Both PDD and DLB may be accompanied or preceded by symptoms referable to brainstem pathology below the substantia nigra including constipation, orthostatic lightheadedness, depression/anxiety, and RBD, and most researchers now conceptualize DLB and PDD as points on a spectrum of LBD pathology. When orthostatic hypotension is present, DLB must be distinguished from multiple system atrophy with parkinsonism (MSA-P) (Chap. 432). Recurrent, disabling syncope early in the course, accompanied by laryngeal spasms and anterocollis, suggest MSA-P. In DLB, orthostasis can appear early but rarely becomes disabling until well into the course, when neuropsychiatric symptoms and cognitive dysfunction are well-established.

Patients with DLB and PDD are highly sensitive to metabolic perturbations, and in some patients the first manifestation of illness is a delirium, often precipitated by an infection, new medicine, or other systemic

disturbance. A hallucinatory delirium induced by L-dopa, prescribed for parkinsonian symptoms attributed to PD, may likewise provide the initial clue to a DLB or PDD diagnosis. Conversely, patients with mild cognitive deficits and hallucinations may receive typical or atypical antipsychotic medications, which induce profound parkinsonism at low doses due to a subclinical LBD-related nigral dopaminergic neuron loss. Minor day-to-day variation in cognitive functioning is common across dementias, but in DLB fluctuations can be marked, with bouts of confusion, lethargy, or even stupor that may rapidly resolve. Cognitively, DLB features relative preservation of episodic memory but often more severe visuospatial and executive deficits than seen in patients with early stages of Alzheimer's disease. Iodine-123-meta-iodobenzylguanidine (MIBG) cardiac scintigraphy to illustrate cardiac postganglionic sympathetic denervation and dopamine transporter imaging using SPECT or PET can both be useful in distinguishing dementia with parkinsonism due to LBD from that due to AD alone.

NEUROPATHOLOGY

The key neuropathological feature in LBD is the presence of Lewy bodies and Lewy neurites throughout specific brainstem nuclei, substantia nigra, amygdala, cingulate gyrus, and, ultimately, the neocortex. Formal staging criteria identify three stages of ascension: (1) brainstem predominant, (2) transitional limbic, and (3) diffuse neocortical, although healthy older individuals may also show isolated scattered Lewy body pathology in the substantia nigra, amygdala, or olfactory bulb. Lewy bodies are intraneuronal cytoplasmic inclusions that stain with periodic acid–Schiff (PAS) but are now identified with antibodies to the presynaptic protein, α-synuclein. Lewy bodies are composed of straight neurofilaments 7–20 nm long with surrounding amorphous material and contain epitopes recognized by antibodies against phosphorylated and nonphosphorylated neurofilament proteins, ubiquitin, and α-synuclein. Generally neuronal and synaptic loss, rather than Lewy pathology per se, best predict the clinical deficits. A profound cholinergic deficit, owing to basal forebrain and pedunculopontine nucleus involvement, is present in many patients with DLB and may be a factor responsible for the fluctuations, inattention, and visual hallucinations. Adrenergic deficits from locus ceruleus involvement further undermine arousal and alerting.

TREATMENT

TREATMENT

Dementia with Lewy Bodies

Due to the frequent comorbidity with AD and the cholinergic deficit in DLB, cholinesterase inhibitors often provide significant benefit, reducing hallucinosis, stabilizing delusional symptoms, and even helping with RBD in some patients. Exercise programs maximize motor function and protect against fall-related injury. Antidepressants are often necessary. Atypical antipsychotics may be required for psychosis but can worsen extrapyramidal syndromes, even at low doses, and should be used cautiously, given the side effects including an increased risk of death. Patients with DLB are extremely sensitive to dopaminergic medications **(Chap.**

427), which must be carefully titrated; tolerability may be improved by concomitant use of a cholinesterase inhibitor. **A general approach to the symptomatic management of dementia is presented in Chap. 25**.

FURTHER READING

Goedert M et al: 100 years of Lewy pathology. Nat Rev Neurol 9:13, 2013. [PubMed: 23183883]

McKeith IG et al: Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. Neurology 89:88, 2017. [PubMed: 28592453]

Morra LF, Donovick PJ: Clinical presentation and differential diagnosis of dementia with Lewy bodies: A review. Int J Geriatr Psychiatry 29:569, 2014. [PubMed: 24150834]

Sonni I et al: Clinical validity of presynaptic dopaminergic imaging with 123I-ioflupane and noradrenergic imaging with 123I-MIBG in the differential diagnosis between Alzheimer's disease and dementia with Lewy bodies in the context of a structured 5-phase development framework. Neurobiol Aging 52:228, 2017. [PubMed: 28317651]

Velayudhan L et al: New therapeutic strategies for Lewy body dementias. Curr Neurol Neurosci Rep 17:68, 2017. [PubMed: 28741230]

Chapter 427: Parkinson's Disease

C. Warren Olanow; Christine Klein; Anthony H.V. Schapira

PARKINSON'S DISEASE AND RELATED DISORDERS

Parkinson's disease (PD) is the second most common age-related neurodegenerative disease, exceeded only by Alzheimer's disease (AD). Its cardinal clinical features were first described by the English physician James Parkinson in 1817. It is noteworthy that James Parkinson was a general physician who captured the essence of this condition based on a visual inspection of a mere handful of patients, several of whom he only observed and did not formally examine. It is estimated that the number of people with PD in the most populous nations worldwide was ~4 million persons in 2005, and this number is expected to more than double to ~9 million by the year 2030 based on the aging of the population. The mean age of onset of PD is about 60 years, and the lifetime risk is ~2% for men and 1.3% for women. The frequency of PD increases with aging, but cases can be seen in individuals in their twenties and even younger, particularly in association with a gene mutation.

Clinically, PD is characterized by rest tremor, rigidity (stiffness), bradykinesia (slowing), and gait dysfunction with postural instability. These are known as the "cardinal features" of the disease. Additional clinical features can include freezing of gait, speech difficulty, swallowing impairment, autonomic disturbances, and a series of nonmotor features that include sensory alterations, mood disorders, sleep dysfunction, cognitive impairment, and dementia **(see Table 427-1 and discussion below)**.

TABLE 427-1

Clinical Features of Parkinson's Disease

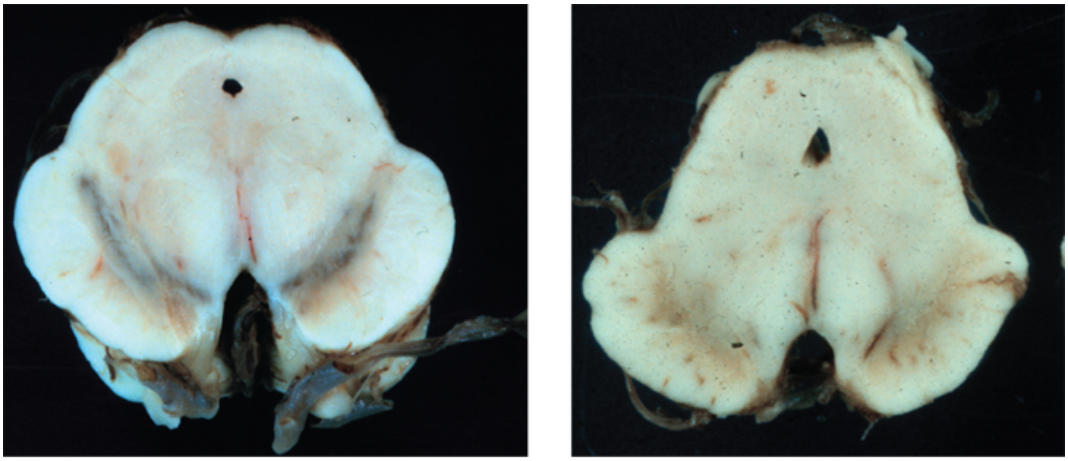
Cardinal Motor Features	Other Motor Features	Nonmotor Features
Bradykinesia	Micrographia	Anosmia
Rest tremor	Masked facies (hypomimia)	Sensory disturbances (e.g., pain)
Rigidity	Reduced eye blinking	Mood disorders (e.g., depression)
Postural instability	Drooling	Sleep disturbances (e.g., RBD)
	Soft voice (hypophonia)	Autonomic disturbances
	Dysphagia	Orthostatic hypotension
	Freezing	Gastrointestinal disturbances
		Genitourinal disturbances
		Sexual dysfunction
		Cognitive impairment/Dementia

Abbreviations: RBD, rapid eye movement behavior disorder.

Pathologically, the hallmark features of PD are degeneration of dopaminergic neurons in the substantia nigra pars compacta (SNc), reduced striatal dopamine, and intraneuronal proteinaceous inclusions known as Lewy bodies and Lewy neurites that primarily contain the protein α-synuclein **(Fig. 427-1)**. While interest has primarily focused on the dopamine system, neuronal degeneration with inclusion body formation can also affect cholinergic neurons of the nucleus basalis of Meynert (NBM), norepinephrine neurons of the locus coeruleus (LC), serotonin neurons in the raphe nuclei of the brainstem, and neurons of the olfactory system, cerebral hemispheres, spinal cord, and peripheral autonomic nervous system. This "nondopaminergic" pathology is likely responsible for the development of the nondopaminergic clinical features listed in Table 427-1. There is some evidence that Lewy body pathology can begin in the peripheral autonomic nervous system, olfactory system, and dorsal motor nucleus of the vagus nerve in the lower brainstem, and then spread in a predictable and sequential manner to affect the upper brainstem (SNc) and cerebral hemispheres (Braak staging). These studies suggest that the classic degeneration of SNc dopamine neurons and the cardinal motor features of PD develop at a mid-stage of the disease. Indeed, epidemiologic studies suggest that clinical symptoms reflecting early involvement of nondopaminergic neurons such as constipation, anosmia, rapid eye movement (REM) behavior sleep disorder, and cardiac denervation can precede the onset of the classic motor features of PD by several years if not decades. Based on these findings, efforts are underway to accurately define a premotor stage of PD.

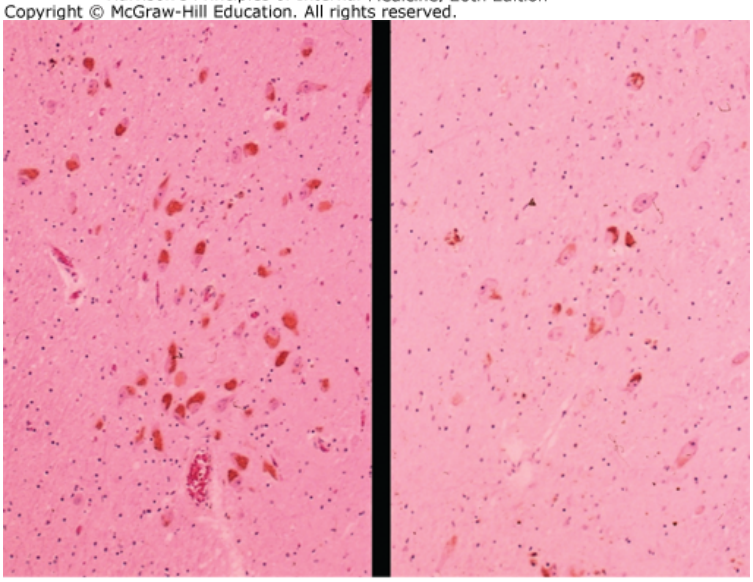
FIGURE 427-1

Pathologic specimens from a patient with Parkinson's disease (PD) compared to a normal control demonstrating (***A***) reduction of pigment in SNc in PD (*right*) versus control (*left*), (***B***) reduced numbers of cells in SNc in PD (*right*) compared to control (*left*), and (***C***) Lewy bodies (*arrows*) within melanized dopamine neurons in PD. SNc, substantia nigra pars compacta.



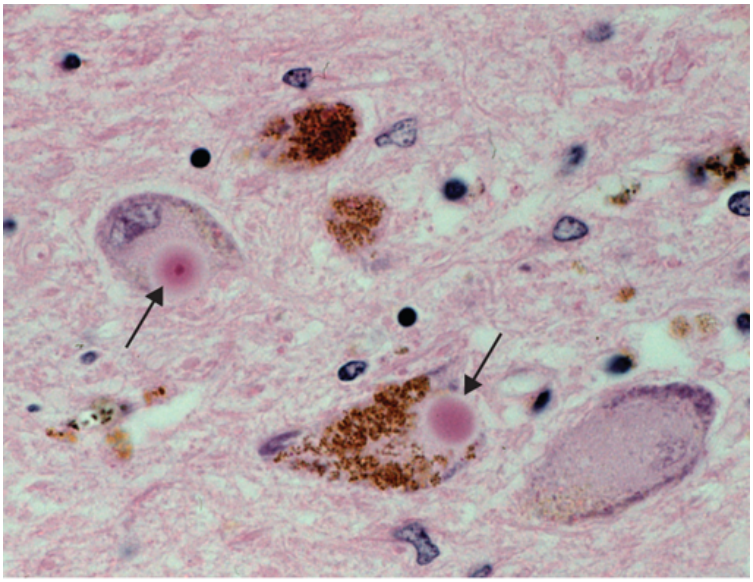
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DIAGNOSIS AND DIFFERENTIAL DIAGNOSIS

Parkinsonism is a generic term that is used to define a syndrome manifest by bradykinesia with rigidity and/or tremor. It has a differential diagnosis (Table 427-2) that reflects differences in the site of damage and pathology in the various components of the basal ganglia. The basal ganglia comprise a group of subcortical nuclei that include the striatum (putamen and caudate nucleus), subthalamic nucleus (STN), globus pallidus pars externa (GPe), globus pallidus pars interna (GPi), and the SNc (Fig. 427-2). Among the different forms of parkinsonism, PD is the most common (~75% of cases). Historically, PD was diagnosed based on the presence of two of three parkinsonian features (tremor, rigidity, bradykinesia). However, postmortem studies found a 24% error rate when diagnosis was based solely on these criteria. Clinicopathologic correlation studies subsequently determined that parkinsonism associated with rest tremor, asymmetry of motor impairment, and a good response to levodopa was more likely to predict the correct pathologic diagnosis. With these revised criteria (known as the U.K. Brain Bank Criteria), a clinical diagnosis of PD could be confirmed pathologically in as many as 99% of cases. The International Parkinson's Disease and Movement Disorder Society (MDS) has recently suggested revised clinical criteria for PD (known as the MDS Clinical Diagnostic Criteria for Parkinson's disease) that are currently undergoing international validation. While motor parkinsonism has been retained as the core feature of the disease, the diagnosis of PD as the cause of parkinsonism relies on three additional categories of diagnostic features: supportive criteria (features that increase confidence in the diagnosis of PD), absolute exclusion criteria, and red flags (which must be counterbalanced by supportive criteria to permit a diagnosis of PD). Utilizing these criteria, two levels of certainty have been delineated; clinically established PD, and probable PD. The key diagnostic criteria for PD based on MDS criteria are illustrated in Table 427-2.

TABLE 427-2

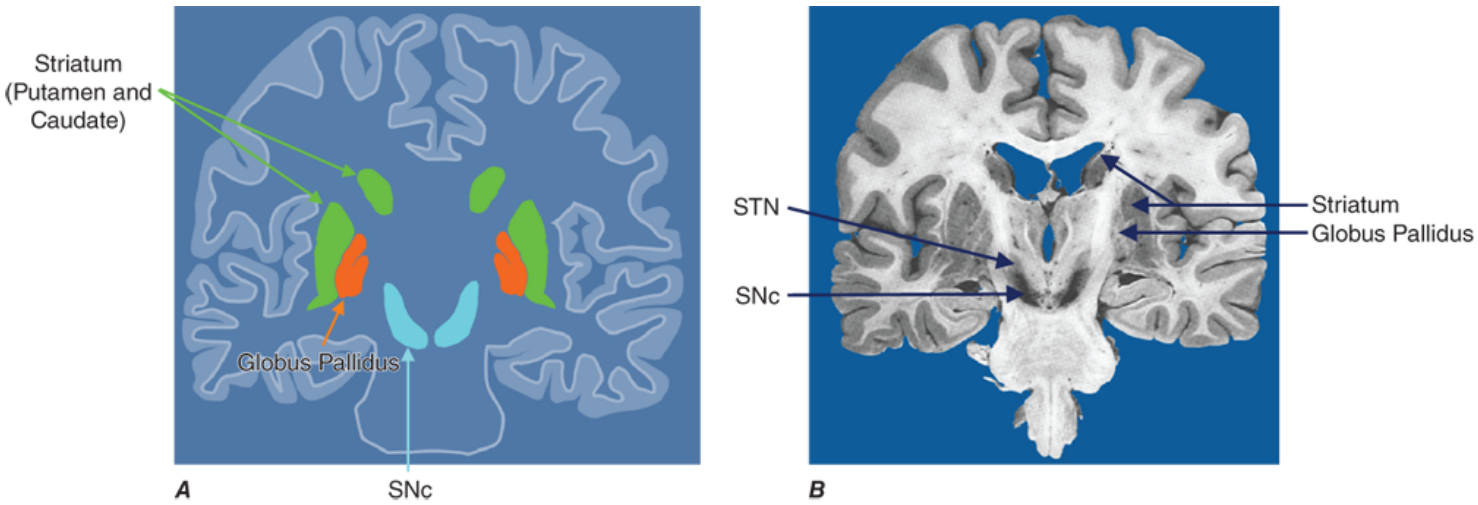
Differential Diagnosis of Parkinsonism

Parkinson's Disease Sporadic Genetic Dementia with Lewy bodies	Atypical Parkinsonism Multiple-system atrophy (MSA) Cerebellar type (MSA-c) Parkinson type (MSA-p) Progressive supranuclear palsy Parkinsonism Richardson variant Corticobasal Syndrome Frontotemporal dementia	Secondary Parkinsonism Drug-induced Tumor Infection Vascular Normal-pressure hydrocephalus Trauma Liver failure Toxins (e.g., carbon monoxide, manganese, MPTP, cyanide, hexane, methanol, carbon disulfide)	Neurodegenerative Disorders and other forms of parkinsonism Wilson's disease Huntington's disease Neurodegeneration with brain iron accumulation SCA 3 (spinocerebellar ataxia) Fragile X–associated ataxia-tremor-parkinsonism Prion disease X-linked Dystonia-parkinsonism Alzheimer's disease with parkinsonism Dopa-Responsive Dystonia

Abbreviations: MPTP, 1-methyl-4-phenyl-1,2,5,6-tetrahydropyridine.

FIGURE 427-2

Basal ganglia nuclei. Schematic (***A***) and postmortem (***B***) coronal sections illustrating the various components of the basal ganglia. SNc, substantia nigra pars compacta; STN, subthalamic nucleus.



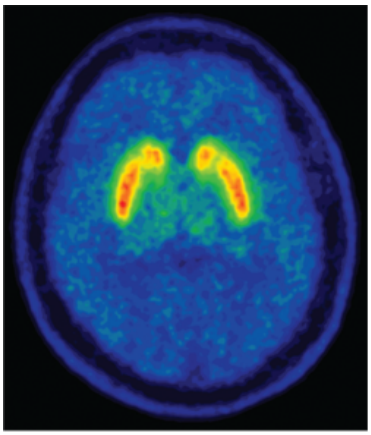
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Imaging of the brain dopamine system in patients with PD can be performed using positron emission tomography (PET) or single-photon emission computed tomography (SPECT). These studies typically show reduced and asymmetric uptake of striatal dopaminergic biomarkers, particularly in the posterior putamen with relative sparing of the caudate nucleus **(Fig. 427-3)**. These findings reflect the degeneration of nigrostriatal dopaminergic neurons and the loss of striatal terminals. Imaging can be useful in patients where there is diagnostic uncertainty (e.g., essential tremor, dystonic tremor, psychogenic tremor) or in research studies, but is rarely necessary in routine practice because the diagnosis can usually be established on clinical criteria alone. This may change in the future when there is a disease-modifying therapy and it is critically important to make a correct diagnosis as early as possible. Genetic testing can be helpful for establishing a diagnosis, but is not routinely employed as monogenic forms are rare and likely account for no more than 5% of cases (see discussion below). A genetic form of PD should be considered in patients with a positive family history, early age of onset (<40 years), a specific clinical picture or a particular ethnic background, and in research studies. Mutations of the *LRRK2* gene have attracted particular interest because they are the most common known cause of familial PD and are responsible for ~1% of typical sporadic cases of the disease. Mutations in *LRRK2* are a particularly frequent cause (~25%) of PD in Ashkenazi Jews and North African Berber Arabs; however,

there is considerable variability in penetrance and many carriers never develop clinical features of PD. Genetic testing is of particular interest to identify at-risk individuals in a research setting. There is also some evidence that diagnosis of PD, and even pre-PD, may possible based on the presence of increased iron accumulation in the SNc using transcranial sonography or special MRI protocols.

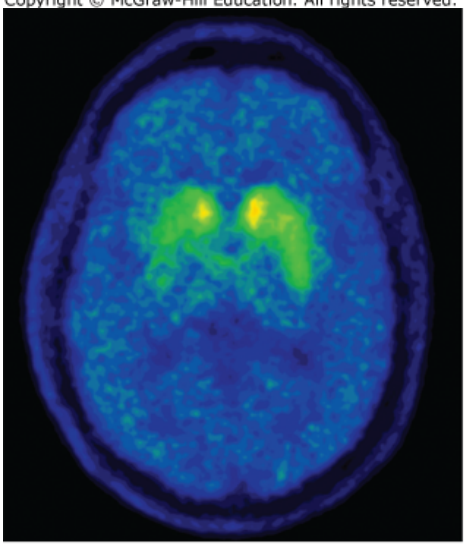
FIGURE 427-3

[^{11}C]Dihydrotetrabenazine positron emission tomography (a marker of VMAT2) in healthy control (*A*) and Parkinson's disease (*B*) patient. Note the reduced striatal uptake of tracer, which is most pronounced in the posterior putamen and tends to be asymmetric. *(Courtesy of Dr. Jon Stoessl.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Atypical, Secondary and Other Forms of Parkinsonism

Atypical parkinsonism refers to a group of neurodegenerative conditions that usually are associated with more widespread pathology than found in PD (potentially with degeneration of striatum, globus pallidus, cerebellum and brainstem as well as the SNc). These include Multiple System Atrophy (MSA), Progressive Supranuclear Palsy (PSP), and Corticobasal syndrome (CBS). As a group, they present with parkinsonism (rigidity and bradykinesia) but manifest clinical differences from PD reflecting the differences in their underlying pathology. In comparison to PD, the atypical parkinsonisms are characterized clinically by early involvement of speech and gait, absence of rest tremor, lack of motor asymmetry, poor or no response to levodopa, and a more aggressive clinical course. In the early stages, they may show a modest benefit from levodopa and can be difficult to distinguish from PD, but the diagnosis becomes clearer with disease evolution. Pathologically, neurodegeneration involves the SNc (typically without Lewy bodies) and has more extensive neurodegeneration than occurs in PD (see below for individual conditions). Neuroimaging of the dopamine system is usually not helpful, as striatal dopamine depletion can be seen in both PD and atypical parkinsonism. By contrast, metabolic imaging of the basal ganglia/thalamus network (using 2-F-deoxyglucose) may be helpful, showing a pattern of decreased activity in the GPi with increased activity in the thalamus, the reverse of what is seen in PD.

MSA manifests as a combination of parkinsonian, cerebellar, and autonomic features and can be divided into a predominant parkinsonian (MSA-p) or cerebellar (MSA-c) form. Clinically, MSA is suspected when a patient presents with features of atypical parkinsonism as described above in conjunction with cerebellar signs and/or prominent autonomic dysfunction, usually orthostatic hypotension **(Chap. 432)**. Pathologically, MSA is characterized by degeneration of the SNc, striatum, cerebellum, and inferior olivary nuclei coupled with characteristic glial cytoplasmic inclusions (GCIs) that stain positively for α-synuclein. Magnetic resonance imaging (MRI) can show pathologic iron accumulation in the striatum on T2-weighted scans, high signal change in the region of the external surface of the putamen (putaminal rim) in MSA-p, or cerebellar and brainstem atrophy (the pontine "hot cross bun" sign **[Fig. 432-2]**) in MSA-c. There is currently no established evidence for any gene mutation/genetic risk factor for MSA. Recent studies suggest the possibility that MSA may be a prion disorder (see discussion below).

PSP is a form of atypical parkinsonism that is characterized by slow ocular saccades, eyelid apraxia, and restricted vertical eye movements with particular impairment of downward gaze. Patients frequently experience hyperextension of the neck with early gait disturbance and falls. In later stages, speech and swallowing difficulty and cognitive impairment may become evident. There is usually little or no response to levodopa. Two clinical forms of PSP have been identified; a "Parkinson" form that can closely resemble PD in the early stages including a positive response to levodopa, and the more classic "Richardson" form that is characterized by the features described above. MRI may reveal a characteristic atrophy of the midbrain with relative preservation of the pons on midsagittal images (the so-called "hummingbird sign"). Pathologically, PSP is characterized by degeneration of the SNc, striatum, STN, midline thalamic nuclei, and pallidum, coupled with neurofibrillary tangles and inclusions that stain for the tau protein. Mutations in the MAPT gene which encodes for the tau protein have been detected in some familial cases.

CBS is the least common of the three atypical parkinsonisms and usually presents with asymmetric dystonic contractions and clumsiness of one hand coupled with cortical sensory disturbances manifest as apraxia, agnosia, focal limb myoclonus, or alien limb phenomenon (where the limb assumes a position in space without the patient being aware of the position or recognizing that the limb belongs to him/her). Dementia may occur at any stage of the disease. Both cortical and basal ganglia features are required to make this diagnosis. MRI frequently shows asymmetric cortical atrophy but this must be carefully sought. Pathologic findings include achromatic neuronal degeneration with tau deposits. Considerable overlap may occur both clinically and pathologically between CBS and PSP, and they may be difficult to distinguish without pathologic confirmation.

Secondary parkinsonisms occur as a result of a variety of primary conditions including drugs, stroke, tumor, infection, or exposure to toxins such as carbon monoxide or manganese that can cause damage to specific regions of the basal ganglia. Clinical features reflect the region of the basal ganglia that has been damaged. For example, strokes or tumors that affect the SNc may have a clinical picture identical to PD, whereas toxins such as carbon monoxide or manganese that damage the globus pallidus more closely resemble atypical parkinsonism. Dopamine-blocking agents such as neuroleptics are the most common cause of secondary parkinsonism. These drugs are most widely used in psychiatry, *but medical physicians should be aware that drugs such as metoclopramide which are primarily used to treat gastrointestinal problems are also neuroleptic agents* and may induce secondary parkinsonism. These drugs can also cause acute and tardive dyskinesias **(see Chap. 428)**. Other drugs that can cause secondary parkinsonism include tetrabenazine, calcium channel blockers (flunarizine, cinnarizine), amiodarone, and lithium.

Parkinsonism can also be seen in Dopa-Responsive Dystonia, a condition that results from a mutation in the *GTP-Cyclohydrolase 1 gene* which can lead to a defect in a cofactor for tyrosine hydroxylase and the impaired manufacture of dopa and dopamine. While it typically presents as dystonia **(Chap. 428)**, it can present as a biochemically based form of parkinsonism (due to reduced synthesis of dopamine) which closely resembles PD and responds to levodopa, but is not associated with abnormalities on fluoro-dopa positron emission tomography (FD-PET) nor neurodegeneration. This diagnosis should be considered in individuals aged <20 years who present with a clinical picture resembling PD.

Finally, parkinsonism can be seen as a feature of a variety of other degenerative disorders such as Wilson's disease, Huntington's disease (especially the juvenile form known as the Westphal variant), certain forms of spinocerebellar ataxias, and neurodegenerative disorders with brain iron accumulation such as pantothenate kinase (PANK)–associated neurodegeneration (formerly known as Hallervorden-Spatz disease).

Some features that suggest that parkinsonism might be due to a condition other than PD are shown in **Table 427-3**.

TABLE 427-3

Features Suggesting an Atypical or Secondary Cause of Parkinsonism

Symptoms/Signs	Alternative Diagnosis to Consider
History	
Early speech and gait impairment (Lack of tremor, lack of motor asymmetry)	Atypical parkinsonism
Exposure to neuroleptics	Drug-induced parkinsonism
Onset prior to age 40	Genetic form of PD
Liver disease	Wilson's disease, non-Wilsonian hepatolenticular degeneration
Early hallucinations and dementia with later development of PD features	Dementia with Lewy bodies
Diplopia, impaired down gaze	PSP
Poor or no response to an adequate trial of levodopa	Atypical or secondary parkinsonism
Physical Examination	
Dementia as first or early feature	Dementia with Lewy bodies
Prominent orthostatic hypotension	MSA-p
Prominent cerebellar signs	MSA-c
Slow saccades with impaired down gaze	PSP
High-frequency (6–10 Hz) symmetric postural tremor with a prominent kinetic component	Essential tremor

Abbreviations: MSA-c, multiple-system atrophy–cerebellar type; MSA-p, multiple-system atrophy–Parkinson's type; PD, Parkinson's disease; PSP, progressive supranuclear palsy.

ETIOLOGY AND PATHOGENESIS

Most PD cases occur sporadically (~85–90%) and are of unknown cause. Gene mutations (see below) are the only known causes of PD. Twin studies performed several decades ago suggested that environmental factors might play an important role in patients with an age of onset ≥50 years, with genetic factors being more important in younger-onset patients. However, the demonstration of later onset genetic variants (e.g., *LRRK2* and *GBA*) argues against the emphasis on environmental factors, even in individuals >50 years of age. The environmental hypothesis received some support in the 1980s with the demonstration that MPTP (1-methyl-4-phenyl-1,2,5,6-tetrahydropyridine), a by-product of the illicit manufacture of a heroin-like drug, caused a PD-like syndrome in addicts in northern California. MPTP is transported into the central nervous system, where it is oxidized to form MPP^+, a mitochondrial toxin that is selectively taken up by, and damages, dopamine neurons. However, MPTP or MPTP-like compounds have not been linked to sporadic PD. Epidemiologic studies have reported an increased risk of developing PD in association with exposure to pesticides, rural living, farming, and drinking well water. Dozens of other associations have also been reported in individual studies but results have been inconsistent, and no environmental factor has yet been proven to be a cause or to contribute to the cause of PD. Some possible protective factors have also been identified in epidemiologic studies including caffeine, smoking, intake of nonsteroidal anti-inflammatory drugs, and calcium channel blockers. The validity of these findings and the responsible mechanism also remain to be established.

About 5–15% of cases are familial in origin, and mutations in several PD-linked genes have been identified (Table 427-4). While monogenic mutations have been shown to be causative of PD, genetic risk factors that increase the risk of developing PD have also been identified. Large-size genome-wide association studies (GWASs) have identified 26 independent gene variants (single nucleotide polymorphisms) as PD risk factors including variants in the *SNCA, LRRK2, MAPT*, and *GBA* genes as well as in the HLA region on chromosome 6. It has been proposed that many cases of PD may be due to a

“double hit” involving an interaction between (a) one or more genetic risk factors that induce susceptibility coupled with (b) exposure to a toxic environmental factor that may induce epigenetic or somatic DNA alterations or has the potential to directly damage the dopaminergic system. In this scenario, both factors are required for PD to ensue, while the presence of either one alone is not sufficient to cause the disease. Notably, however, even if a genetic or environmental risk factor doubles the risk to develop PD, this only results in a lifetime risk of 4% or lower, and thus cannot presently be used for individual patient counseling.

TABLE 427-4

Confirmed Genetic Causes of Parkinson's Disease*

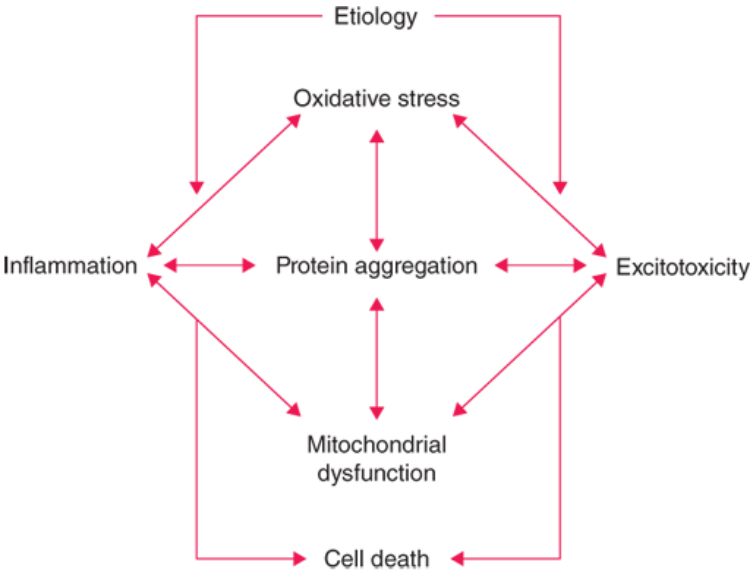
Designation* and Reference	GeneReviews and OMIM Reference	Clinical Clues	Inheritance	Previous Locus Symbol
1. Classical PD				
PARK-*SNCA*	GeneReviews *http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/books/NBK1223/* OMIM 168601	Missense mutations cause classical parkinsonism. Duplication or triplication mutations in this gene cause early onset parkinsonism with prominent dementia.	AD	PARK1
PARK-*LRRK2*	GeneReviews *http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/books/NBK1208/* OMIM 607060	Clinically typical PD	AD	PARK8
PARK-*VPS35*	GeneReviews *http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/books/NBK1223/* OMIM 614203	Clinically typical PD	AD	PARK17
2. Early-onset PD				
PARK-*Parkin*	GeneReviews *http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/books/NBK1155/* OMIM 600116	Often presents with dystonia, typically in a leg	AR	PARK2
PARK-*PINK1*	GeneReviews *http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/books/NBK1223/* OMIM 605909	Often presents with psychiatric features	AR	PARK6
PARK-*DJ1*	GeneReviews *http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/books/NBK1223/* OMIM 606324		AR	PARK7
3. Parkinsonism				
PARK-*ATP13A2*	GeneReviews *http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/books/NBK1223/* OMIM 606693	Kufor-Rakeb syndrome with parkinsonism and dystonia; additional features: Supranuclear gaze palsy, spasticity/pyramidal signs, dementia, facial-faucial-finger mini-myoclonus, dysphagia, dysarthria, olfactory dysfunction	AR	PARK9
PARK-*FBXO7*	GeneReviews *http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/books/NBK1223/* OMIM: 260300	Early onset parkinsonism with pyramidal signs	AR	PARK15
PARK-*DNAJC6*	GeneReviews: n/a OMIM 615528	May present with mental retardation and seizures	AR	PARK19
PARK-*SYNJ1*	GeneReviews: n/a OMIM 615530	May have seizures, cognitive decline, abnormal eye movements, and dystonia	AR	PARK20

*According to the recommendations of the International Parkinson's and Movement Disorder Society (C Marras: Mov Disord 31:436, 2016).

Several factors have been implicated in the pathogenesis of cell death in PD, including oxidative stress, inflammation, excitotoxicity, mitochondrial dysfunction, and the accumulation of misfolded proteins with consequent proteolytic stress. Recent studies have demonstrated that with aging, dopamine neurons switch from sodium to calcium pacing through calcium channels, potentially making these high-energy neurons vulnerable to calcium-mediated neurotoxicity. Whatever the pathogenic mechanism, cell death appears to occur, at least in part, by way of a signal-mediated apoptotic or "suicidal" process. Each of these mechanisms offers a potential target for putative neuroprotective drugs. However, it is not clear which of these factors is primary, if they are the same in all cases or specific to individual (genetic) patient subgroups, if they act by way of a network such that multiple insults are required for neurodegeneration to ensue, or if the findings to date merely represent an epiphenomenon unrelated to the true cause of cell death that still remains undiscovered **(Fig. 427-4)**.

FIGURE 427-4

Schematic representation of how pathogenetic factors implicated in Parkinson's disease interact in a network manner, ultimately leading to cell death. This figure illustrates how interference with any one of these factors may not necessarily stop the cell death cascade. *(Adapted from CW Olanow: Movement Disorders 22:S-335, 2007.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although gene mutations cause only a minority of cases of PD, they may be helpful in pointing to specific pathogenic pathways and molecular mechanisms that are central to a neurodegenerative process that might be relevant to all forms of the disease. To date, most interest has focused on pathways implicated by mutations in *α-synuclein (SNCA)*, *GBA*, *LRRK2*, and *PINK1/Parkin*.

Although mutations in *SNCA* are an extremely rare cause of PD, *SNCA* was the first PD-linked and most intensely investigated PD gene, with respect to causative mutations but also risk variants, function of the gene and of the encoded protein. Shared clinical features of patients with *SNCA* mutations include earlier age of disease onset than in nongenetic PD, a faster progression of motor signs that are mostly levodopa-responsive, early occurrence of motor fluctuations, and presence of prominent nonmotor features. Intriguingly, SNCA constitutes the major component of Lewy bodies in patients with both monogenic and sporadic forms of PD (Fig. 427-1). Duplication or triplication of the wild-type *SNCA* gene also causes PD with triplication carriers being more severely affected than carriers of duplications. These findings indicate that increased production of the normal protein alone can cause the disease in a dose-dependent fashion. More recently, Lewy pathology was discovered to have developed in healthy embryonic dopamine neurons that had been implanted into the striatum of PD patients, suggesting that the abnormal protein had transferred from affected cells to healthy unaffected dopamine neurons. Based on these findings, it has been proposed that the SNCA protein may be a prion, and PD a prion disorder **(Chaps. 417 and 430)**. Like the prion protein PrP^C, SNCA can misfold to form β-rich sheets, join to form toxic oligomers and aggregates, polymerize to form amyloid plaques (i.e., Lewy bodies), and cause neurodegeneration with spread to involve unaffected neurons. Indeed, injection of SNCA fibrils into the striatum of both transgenic and wild-type rodents leads to the development of Lewy pathology in host neurons, neurodegeneration, behavioral abnormalities, with spread of SNCA pathology to anatomically connected sites. Further support for this hypothesis comes from the demonstration that inoculation of SNCA derived from human Lewy bodies induces dopamine cell degeneration and widespread Lewy pathology in mice and primates. Collectively, this evidence supports the possibility that neuroprotective therapies for PD might be developed based on inhibiting the accumulation or accelerating the removal of SNCA aggregates, knocking down levels of host SNCA, or blocking the templating phenomenon whereby misfolded SNCA promotes misfolding of the native protein in a prion-like chain reaction.

Mutations in the glucocerebrosidase (*GBA*) gene represent the most important risk factor in terms of effect size for the development of PD, and experimentally there is a direct pathophysiological link between increased levels of SNCA and reduced levels of GBA. GBA encodes the enzyme

glucocerebrosidase (GCase) which promotes lysosomal function and enhances the clearance of misfolded proteins. The identification of *GBA* as a risk gene for PD resulted from the clinical observation that patients with Gaucher's disease (GD) and their relatives commonly show signs of parkinsonism. This clinical observation of a link between GD and PD led to the discovery that several mutations in *GBA*, which cause Gaucher's disease in an autosomal recessive manner, confer risk for the development of PD, also in a heterozygous state. Further, reduced GCase activity due to GBA mutations impairs lysosomal function which results in the accumulation of SNCA. Accumulation of SNCA can also lead to inhibition of lysosomal function and a further reduction in levels of wild-type GBA by interfering with endoplasmic reticulum-to-Golgi trafficking. This in turn, leads to decreased GBA activity and a further increase in the accumulation of SNCA. In this regard, it is noteworthy that lysosomal function is impaired and levels of GCase are reduced in patients with sporadic PD. These findings suggest that this molecular pathway may apply not only to patients with GD or with a *GBA* mutation, but also to patients with sporadic PD or other synucleinopathies who have two wild-type *GBA* alleles. These bidirectional effects of SNCA and GBA form a positive feedback loop that, after surpassing a theoretical threshold, could lead to self-propagating disease. Studies of drugs that enhance GCase activity are currently underway.

Seven different *LRRK2* mutations have now been clearly linked to PD, with p.G2019S being the most common due to a founder effect in the Ashkenazi Jewish and North African Arab populations. Mutations in *LRRK2* account for 3–41% of familial PD cases (depending on specific population) and are also found in apparently sporadic cases, albeit at a lower rate. The phenotype of LRRK2 p.G2019S mutations is indistinguishable from that of sporadic PD, although tremor appears to be more common, and leg tremor may be a useful diagnostic clue. The mechanism responsible for cell death with this mutation is not conclusively known but is thought to involve changes in kinase activity with altered phosphorylation of target proteins (including autophosphorylation) with possible impairment of lysosomal function. Kinase inhibitors can block toxicity associated with *LRRK2* mutations in laboratory models, and there has been much interest in developing drugs directed at this target. However, kinase inhibitors are potentially toxic, and the majority of PD patients do not carry a *LRRK2* mutation.

Mutations in *Parkin* and *PINK1* have also been identified as a cause of PD. *Parkin* mutations are the more common, and the major cause of autosomal recessive and early-onset PD, accounting for up to 77% of cases of juvenile PD with an age of onset <20 years, and for 10–20% of early-onset PD patients in general. The disease is slowly progressive, responds well to antiparkinsonian treatment, and is commonly complicated by dystonia, but very rarely by dementia. At pathology, neurodegeneration tends to be restricted to the SNc and LC in patients with *Parkin* mutations, and Lewy bodies are typically absent. The reason for these differences from classic PD are not known, but may related to impaired ubiquitination of damaged proteins (parkin is a ubiquitin ligase). The clinical phenotypes of *Parkin*- and *PINK1*-linked PD are similar. Recent studies suggest a role for Parkin and PINK1 proteins in the turnover and clearance of damaged mitochondria (mitophagy), and mutations in *Parkin* and *PINK1* cause mitochondrial dysfunction in transgenic animals that can be corrected with overexpression of Parkin or with drugs acting on the mitochondrial electron transfer chain, such as Vitamin K2. Improving mitochondrial function is a particularly attractive potential therapeutic target because postmortem studies in PD patients show a defect in complex I of the respiratory chain in SNc neurons.

Thus, evidence is accumulating that genetics plays an important role in both familial and "sporadic" forms of PD. It is anticipated that better understanding of the pathways responsible for cell death caused by these mutations will permit the development of more relevant animal models of PD and targets for the development of gene-specific neuroprotective drugs.

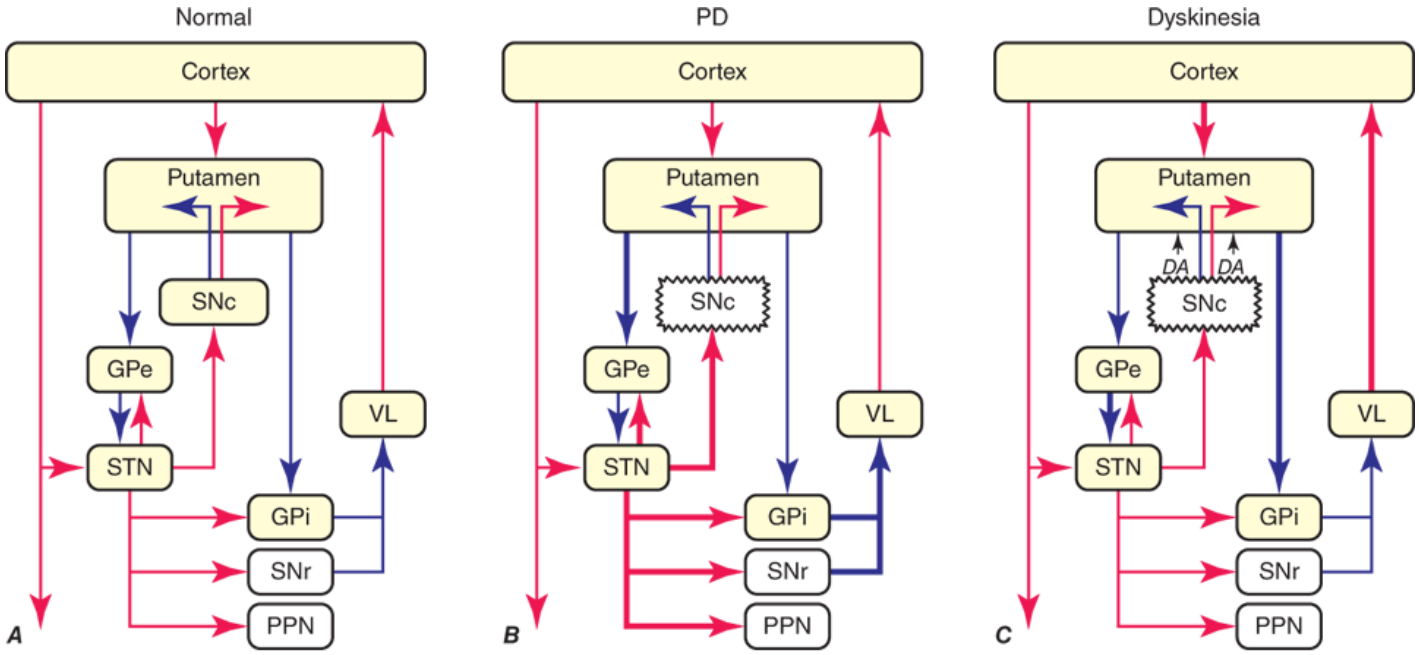
PATHOPHYSIOLOGY OF PD

The classic model of the organization of the basal ganglia in the normal and PD states is provided in **Fig. 427-5**. With respect to motor function, a series of neuronal circuits or loops link the basal ganglia nuclei with corresponding cortical motor regions in a somatotopic manner. The striatum is the major input region of the basal ganglia, while the GPi and SNr are the major output regions. The input and output regions are connected via direct and indirect pathways that have reciprocal effects on the activity of the basal ganglia output pathway. The output of the basal ganglia provides inhibitory (GABAergic) tone to thalamic and brainstem neurons that in turn connect to motor systems in the cerebral cortex and spinal cord that control motor function. An increase in neuronal activity in the output regions of the basal ganglia (GPi/SNr) is associated with poverty of movement or parkinsonism, while decreased output results in movement facilitation and involuntary movements. Dopaminergic projections from SNc neurons serve to modulate neuronal firing and to stabilize the basal ganglia network. Normal dopamine innervation thus serves to facilitate the selection of the desired movement and reject unwanted movements. Cortical loops integrating the cortex and the basal ganglia are now thought to also play an important role in regulating behavioral, emotional, and cognitive functions.

FIGURE 427-5

Basal ganglia organization. Classic model of the organization of the basal ganglia in the normal (***A***), Parkinson's disease (PD) (***B***), and levodopa-induced dyskinesia (***C***) state. Inhibitory connections are shown as *blue arrows* and excitatory connections as *red arrows*. The striatum is the major input region and receives its major input from the cortex. The GPi and SNr are the major output regions, and they project to the thalamocortical and brainstem motor regions. The striatum and GPi/SNr are connected by direct and indirect pathways. This model predicts that parkinsonism results from increased neuronal firing in the STN and GPi and that lesions or DBS of these targets might provide benefit. This concept led to the rationale for surgical therapies for PD. The model also predicts that dyskinesia results from decreased firing of the output regions, resulting in excessive cortical activation by the thalamus. This component of the model is not completely correct because lesions of the GPi ameliorate rather than increase

dyskinesia in PD, suggesting that firing frequency is just one of the components that lead to the development of dyskinesia. DBS, deep brain stimulation; GPe, external segment of the globus pallidus; GPi, internal segment of the globus pallidus; PPN, pedunculopontine nucleus; SNc, substantia nigra, pars compacta; SNr, substantia nigra, pars reticulata; STN, subthalamic nucleus; VL, ventrolateral thalamus. *(Derived from JA Obeso et al: Trends Neurosci 23:S8, 2000.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In PD, dopamine denervation with loss of dopaminergic tone leads to increased firing of neurons in the STN and GPi, excessive inhibition of the thalamus, reduced activation of cortical motor systems, and the development of parkinsonian features (Fig. 427-5). The current role of surgery in the treatment of PD is based on this model, which predicted that lesions or high-frequency stimulation of the STN or GPi might reduce this neuronal overactivity and improve PD features.

TREATMENT

TREATMENT

Parkinson's Disease

LEVODOPA

Since its introduction in the late 1960s, levodopa has been the mainstay of therapy for PD. Experiments in the late 1950s by Carlsson and colleagues demonstrated that blocking dopamine uptake with reserpine caused rabbits to become parkinsonian; this could be reversed with the dopamine precursor, levodopa. Subsequently, Hornykiewicz demonstrated a dopamine deficiency in the striatum of PD patients, and suggested the potential benefit of dopamine replacement therapy. Dopamine does not cross the blood-brain barrier (BBB), so clinical trials were initiated with levodopa, the precursor of dopamine. Studies over the course of the next decade confirmed the value of levodopa and revolutionized the treatment of PD.

Levodopa is routinely administered in combination with a peripheral decarboxylase inhibitor to prevent its peripheral metabolism to dopamine and the development of nausea, vomiting, and orthostatic hypotension due to activation of dopamine receptors in the area postrema that are not protected by the BBB. In the United States, levodopa is combined with the decarboxylase inhibitor carbidopa (Sinemet®), whereas in many other countries it is combined with benserazide (Madopar®). Levodopa plus a decarboxylase inhibitor is also available in a methylated formulation, a controlled-release formulation (Sinemet CR® or Madopar HP®) and in combination with a catechol-*O*-methyltransferase (COMT) inhibitor (Stalevo®). A long-acting formulation of levodopa (Rytary®) has also recently been approved. An inhaled form of levodopa that is rapidly and reliably absorbed is currently in late stage investigation as a rescue therapy for the treatment of individual "off" episodes (see below).

Levodopa remains the most effective symptomatic treatment for PD and the gold standard against which new therapies are compared. No current medical or surgical treatment provides antiparkinsonian benefits superior to what can be achieved with levodopa. Levodopa benefits the classic motor features of PD, prolongs independence and employability, improves quality of life, and increases life span. Almost all PD patients experience improvement, and failure to respond to an adequate trial of levodopa should cause the diagnosis to be questioned.

There are, however, important limitations of levodopa therapy. Acute dopaminergic side effects include nausea, vomiting, and orthostatic hypotension as indicated above. These are usually transient and can generally be avoided by starting with low doses and gradual titration. If they persist, they can be treated with additional doses of a peripheral decarboxylase inhibitor (e.g., carbidopa), administering with food, or adding a peripheral dopamine-blocking agent such as domperidone (not available in the United States). More important are motor complications (see below) that develop in the majority of patients treated long-term with levodopa. In addition, the disease continues to progress, and features such as neuropsychiatric problems, falling, freezing, autonomic dysfunction, sleep disorders, and dementia may emerge that are not adequately controlled by levodopa. Indeed, these nondopaminergic features (especially falling and dementia) are the primary source of disability and the main reason in the present era for nursing home placement for patients with advanced PD.

Levodopa-induced motor complications consist of fluctuations in motor response ("on" episodes when the drug is working and "off" episodes when parkinsonian features return) and involuntary movements known as dyskinesias which typically complicate "on" periods **(Fig. 427-6)**. When patients initially take levodopa, benefits are long-lasting (many hours) even though the drug has a relatively short half-life (60–90 min). With continued treatment, however, the duration of benefit following an individual dose becomes progressively shorter until it approaches the half-life of the drug. This loss of benefit is known as the *wearing-off effect*. In more severe cases, the response to a given dose may be variable with patients potentially experiencing a delay in turning on (delayed-on) or no response at all (no-on). Peak-dose dyskinesias occur at the time of levodopa peak plasma concentration and maximal clinical benefit. They are usually choreiform, but can manifest as dystonic movements, myoclonus, or other movement disorders. They are not troublesome when mild, but can be disabling when severe, and can limit the ability to use higher doses of levodopa to better control PD motor features. In more advanced states, patients may cycle between "on" periods complicated by disabling dyskinesias and "off" periods in which they suffer from severe parkinsonism and painful dystonic postures. Patients may also experience "diphasic dyskinesias," which occur as the levodopa dose begins to take effect and again as it wears off. These dyskinesias typically consist of transient, stereotypic, rhythmic movements that predominantly involve the lower extremities and are frequently associated with parkinsonism in other body regions. They can be relieved by increasing the dose of levodopa, although higher doses may induce more severe peak-dose dyskinesia. Long-term double blind studies show that motor complications are dose related, and can be minimized by using the lowest dose of levodopa that provides satisfactory benefit and through the use of polypharmacy to avoid raising the dose of levodopa.

The cause of levodopa-induced motor complications is not precisely known. They are more likely to occur in females, younger individuals with more severe disease, and with the use of higher doses of levodopa. The classic model of the basal ganglia has been useful for understanding the origin of motor features in PD, but has proved less valuable for understanding levodopa-induced dyskinesias (Fig. 427-5). The model predicts that dopamine replacement might excessively inhibit the pallidal output system, thereby leading to increased thalamocortical activity, enhanced stimulation of cortical motor regions, and the development of dyskinesia. However, lesions of the pallidum are associated with amelioration rather than induction of dyskinesia as would be suggested by the classic model. It is now thought that dyskinesia results from alterations in the GPi/SNr neuronal firing pattern (pauses, bursts, synchrony, etc.) and not simply the firing frequency alone. This in turn leads to the transmission of "misinformation" from pallidum to thalamus/cortex, resulting in dyskinesia. Surgical lesions or high-frequency stimulation targeted at the GPi or STN can ameliorate dyskinesia by interfering with (blocking or masking) this abnormal neuronal activity and preventing the transfer of misinformation to motor systems. There has also been recent interest in the use of ultrasound to lesion these target regions in a relatively noninvasive manner.

Current information suggests that altered neuronal firing patterns and motor complications develop in response to nonphysiologic levodopa replacement. Striatal dopamine levels are normally maintained at a relatively constant level. In PD, dopamine neurons degenerate and striatal dopamine is dependent on the peripheral availability of levodopa. Intermittent oral doses of levodopa result in fluctuating plasma levels because of variability in the transit of the drug from the stomach to the duodenum where it is absorbed and the short half-life of the drug. This variability is translated to the brain and results in exposure of striatal dopamine receptors to alternating high and low concentrations of dopamine. It has been hypothesized that more continuous delivery of levodopa might prevent the development of motor complications. Indeed, a recent double-blind, double-dummy, double titration study demonstrated that continuous intraintestinal infusion of levodopa/carbidopa is associated with significant improvement in "off" time and in "on" time without dyskinesia in advanced PD patients compared with optimized standard oral levodopa. These benefits are superior to what has been observed in double blind controlled studies with other dopaminergic agents, and this therapy is now approved in the United States and Europe (Duodopa®, Duopa®). The treatment is, however, complicated by potentially serious adverse events related to the surgical procedure and the tubing, and the inconvenience of the infusion system. New approaches are currently being tested in which levodopa is continuously administered by subcutaneous infusion or by long-acting oral levodopa formulations in an effort to avoid the need for a surgical procedure. An inhaled formulation of levodopa is in late stage development as an acute rescue therapy for individual off episodes.

Behavioral complications can also be encountered in levodopa-treated patients. A dopamine dysregulation syndrome has been described where patients have a craving for levodopa and take frequent and unnecessary doses of the drug in an addictive manner. PD patients taking high doses of levodopa can also develop purposeless, stereotyped behaviors such as the assembly and disassembly or collection and sorting of objects. This is known as punding, a term taken from the Swedish description of the meaningless behaviors seen in chronic amphetamine users. Hypersexuality and other impulse-control disorders are occasionally encountered with levodopa, although these are more commonly seen with dopamine agonists.

DOPAMINE AGONISTS

Dopamine agonists are a diverse group of drugs that act directly on dopamine receptors. Unlike levodopa, they do not require metabolic conversion to an active product and do not undergo oxidative metabolism. Initial dopamine agonists were ergot derivatives (e.g., bromocriptine, pergolide, cabergoline) and were associated with potentially serious ergot-related side effects such as cardiac valvular damage and pulmonary fibrosis. They have largely been replaced by a second generation of nonergot dopamine agonists (e.g., pramipexole, ropinirole, rotigotine). In general, dopamine agonists do not have comparable efficacy to levodopa. They were initially introduced as adjuncts to levodopa to enhance motor function and reduce "off" time in fluctuating patients. Subsequently, it was shown that dopamine agonists are less prone than levodopa to induce dyskinesia, possibly because they are relatively long-acting. For this reason, many physicians initiate therapy with a dopamine agonist particularly in younger patients, although supplemental levodopa is eventually required in virtually all patients. This view has been tempered by the recognition that dopamine agonists are associated with potentially serious adverse effects such as unwanted sleep episodes and impulse control disorders (see below). Both ropinirole and pramipexole are available as orally administered immediate (tid) and extended-release (qd) formulations. Rotigotine is administered as a once-daily transdermal patch, and may be useful in managing surgical patients who are NPO. Apomorphine is a dopamine agonist with efficacy comparable to levodopa, but it must be administered parenterally as it is rapidly and extensively metabolized if taken orally. It has a short half-life and duration of activity (45 min). It can be administered by subcutaneous injection as a rescue agent for the treatment of severe "off" episodes, but can also be administered by continuous subcutaneous infusion where it has been demonstrated to reduce both "off" time and dyskinesia in advanced patients. This latter approach has not been approved in the United States. A sublingual bilayer formulation of apomorphine is in late stage development as a rapid and reliable therapy for individual "off" periods that avoids the need for a subcutaneous (SC) injection.

Dopamine agonist use is associated with a variety of side effects. Acute side effects are primarily dopaminergic and include nausea, vomiting, and orthostatic hypotension. As with levodopa, these can usually be avoided by starting with low doses and using slow titration. Side effects associated with chronic use include hallucinations and cognitive impairment. Sedation with sudden unintended episodes of falling asleep that can occur in dangerous situations such as while driving a motor vehicle have been reported. Patients should be informed about this potential problem and should not drive when tired. Dopamine agonists can also be associated with impulse-control disorders, including pathologic gambling, hypersexuality, and compulsive eating and shopping. Patients should also be advised of these risks and specifically questioned for their occurrence at follow-up examinations. The precise cause of these problems, and why they appear to occur more frequently with dopamine agonists than levodopa, remains to be resolved, but reward systems associated with dopamine and alterations in the ventral striatum and orbitofrontal regions have been implicated. In general, chronic side effects are dose-related and can be avoided or minimized with lower doses. Injections of apomorphine can be complicated by skin lesions at sites of administration, but this has not been a problem with the sublingual bilayer formulation.

MAO-B INHIBITORS

Inhibitors of monoamine oxidase type B (MAO-B) block central dopamine metabolism and increase synaptic concentrations of the neurotransmitter. Selegiline and rasagiline are relatively selective suicide inhibitors of the MAO-B isoform of the enzyme. Clinically, these agents provide antiparkinsonian benefits when used as monotherapy in early disease stages and reduced "off" time when used as an adjunct to levodopa in patients with motor fluctuations. MAO-B inhibitors are generally safe and well tolerated. They may increase dyskinesia in levodopa-treated patients, but this can usually be controlled by down-titrating the dose of levodopa. Inhibition of the MAO-A isoform prevents metabolism of tyramine in the gut, leading to a potentially fatal hypertensive reaction known as a "cheese effect" because it can be precipitated by foods rich in tyramine such as some cheeses, aged meats, and red wine. Selegiline and rasagiline do not functionally inhibit MAO-A and are not associated with a cheese effect with doses used in clinical practice. There are theoretical risks of a serotonin reaction in patients receiving concomitant selective serotonin reuptake inhibitor (SSRI) antidepressants, but these are rarely encountered. Safinamide (Xadago®) is a reversible MAO-B inhibitor that has recently been approved as an adjunct to levodopa in advanced PD patients with motor fluctuations. The drug also acts to block activated sodium channels and inhibit glutamate release, and is currently being studied as a possible antidyskinetic agent.

Interest in MAO-B inhibitors has also focused on their potential to have disease-modifying effects. MPTP toxicity can be prevented experimentally by coadministration of an MAO-B inhibitor that blocks its conversion to the toxic pyridinium ion MPP^+ that selectively damages dopamine neurons. MAO-B inhibitors also have the potential to block the oxidative metabolism of dopamine and prevent oxidative stress. In addition, both selegiline and rasagiline incorporate a propargyl ring within their molecular structure that provides antiapoptotic effects in laboratory models. The DATATOP study showed that selegiline significantly delayed the time until the emergence of disability necessitating the introduction of levodopa in untreated PD patients. However, it could not be definitively determined whether this was due to a neuroprotective effect that slowed disease progression or a symptomatic effect that masked ongoing neurodegeneration. More recently, the ADAGIO study used a two-period delayed-start design and demonstrated that early treatment with rasagiline 1 mg/d, but not 2 mg/d, provided benefits that could not be achieved when treatment with the same drug was initiated at a later time point. This benefit is consistent with a disease-modifying effect; however, the long-term significance of these findings is uncertain.

COMT INHIBITORS

When levodopa is administered with a decarboxylase inhibitor, it is primarily metabolized in the periphery by the catechol-O-methyl transferase (COMT) enzyme. Inhibitors of COMT increase the elimination half-life of levodopa and enhance its brain availability. Combining levodopa with a COMT inhibitor reduces "off" time and prolongs "on" time in fluctuating patients while enhancing motor scores. Two COMT inhibitors, tolcapone and

entacapone, have been approved for use. More recently, opicapone (a long-acting, once daily COMT inhibitor) has been approved in Europe. There is also a combination tablet of levodopa, carbidopa, and entacapone (Stalevo®).

Side effects of COMT inhibitors are primarily dopaminergic (nausea, vomiting, increased dyskinesia) and can usually be controlled by down-titrating the dose of levodopa by 20–30%. Severe diarrhea has been described with tolcapone, and to a lesser degree with entacapone, and necessitates stopping the medication in 5–10% of individuals. Cases of fatal hepatic toxicity have been reported with tolcapone. It is still used because it is the most effective of the COMT inhibitors, but periodic monitoring of liver function is required. This problem has not been encountered with entacapone. Discoloration of urine can be seen with COMT inhibitors due to accumulation of a metabolite, but it is of no clinical concern.

It has been proposed that initiating levodopa in combination with a COMT inhibitor to enhance its elimination half-life could provide more continuous levodopa delivery and reduce the risk of motor complications if administered at frequent intervals. While this result has been demonstrated in a preclinical MPTP model, and continuous infusion reduces both “off” time and dyskinesia in advanced PD patients, no benefit of initiating levodopa with a COMT inhibitor compared to levodopa alone was detected in early PD patients in the STRIDE-PD study. This may have been because the combination was not administered at frequent enough intervals to provide continuous levodopa availability. For now, the main value of COMT inhibitors continues to be in patients who experience motor fluctuations.

OTHER MEDICAL THERAPIES

Centrally acting anticholinergic drugs such as trihexyphenidyl and benztropine were used historically for the treatment of PD, but they lost favor with the introduction of dopaminergic agents. Their major clinical effect is on tremor, although it is not certain that this benefit is superior to what can be obtained with agents such as levodopa and dopamine agonists. Still, they can be helpful in individual patients with severe tremor. Their use is limited particularly in the elderly, due to their propensity to induce a variety of side effects including urinary dysfunction, glaucoma, and particularly cognitive impairment.

Amantadine was originally introduced as an antiviral agent, but was appreciated to also have antiparkinsonian effects that are thought to be due to *N*-methyl-D-aspartate (NMDA) receptor antagonism. While some physicians use amantadine in patients with early disease for its mild symptomatic effects, it is most widely used as an antidyskinesia agent in patients with advanced PD. Indeed, it is the only oral agent that has been demonstrated in controlled studies to reduce dyskinesia without worsening parkinsonian features, although benefits may be relatively transient. Cognitive impairment is a major concern. Other side effects include livedo reticularis and weight gain. Amantadine should always be discontinued gradually because patients can experience withdrawal-like symptoms. An extended release formulation of amantadine has recently been approved in the United States.

The anticonvulsant zonisamide has also been shown to have antiparkinsonian effects and is approved for use in Japan. Its mechanism of action is unknown.

Several new classes of drugs are currently being investigated in an attempt to enhance antiparkinsonian effects, reduce off time, and treat or prevent dyskinesia. These include adenosine A_{2A} antagonists, nicotinic agonists, glutamate antagonists, and 5-HT_{1A} agonists. The A_{2A} antagonist Istradefylline is approved in Japan.

A list of the major drugs and available dosage strengths currently available to treat PD is provided in **Table 427-5**.

NEUROPROTECTION

Despite the many therapeutic agents available for the treatment of PD, patients continue to progress and to develop intolerable disability. A neuroprotective therapy that slows or stops disease progression remains the major unmet therapeutic need. Numerous trials have shown positive results (e.g., selegiline, rasagiline, pramipexole, ropinirole) consistent with a disease-modifying effect. However, it has not been possible to determine with certainty if the positive results were due to neuroprotection with slowing of disease progression or confounding symptomatic or pharmacologic effects that mask disease progression. There is a flurry of clinical activity testing interventions targeting etiopathogenic factors; these include calcium channel blockers, urate, and agents that enhance glucocerebrocidase (GCase) or interfere with SNCA or LRRK2 in the hope that they might provide disease-modifying effects. A major limitation is the uncertainty as to a specific clinical development plan and trial design that will prove acceptable to both clinicians and regulatory authorities.

SURGICAL TREATMENT

Surgical treatments for PD have been used for more than a century. Lesions were initially placed in the motor cortex and improved tremor but were associated with motor deficits, and this approach was abandoned. Subsequently, it was appreciated that lesions placed into the ventral intermediate (VIM) nucleus of the thalamus reduced contralateral tremor without inducing hemiparesis, but these lesions did not meaningfully help other more disabling features of PD. In the 1990s, it was shown that lesions placed in the posteroventral portion of the GPi (motor territory) improved rigidity and bradykinesia as well as tremor. Importantly, pallidotomy was also associated with marked improvement in contralateral dyskinesia. This procedure gained favor with greater understanding of the pathophysiology of PD (see above). However, this procedure is not optimal, because bilateral lesions are associated with side effects such as dysphagia, dysarthria, and impaired cognition. Lesions of the STN are associated with antiparkinsonian benefit and reduced levodopa requirement, but there is a concern about the risk of hemiballismus, and this procedure is not commonly performed.

Most surgical procedures for PD performed today use deep brain stimulation (DBS). Here, an electrode is placed into the target area and connected to a stimulator inserted subcutaneously over the chest wall. DBS simulates the effects of a lesion without necessitating making a brain lesion. The precise mechanism whereby DBS works is not fully resolved but may act by disrupting the abnormal neurophysiological signals associated with PD and motor complications. The stimulation variables can be adjusted with respect to electrode configuration, voltage, frequency, and pulse duration in order to maximize benefit and minimize adverse side effects. The procedure does not require making a lesion in the brain and is thus suitable for performing bilateral procedures with relative safety. In cases with intolerable side effects, stimulation can be stopped and the system removed.

DBS for PD primarily targets the STN or the GPi. It provides dramatic results, particularly with respect to tremor and reducing both “off” time and dyskinesias, but does not provide superior clinical benefits or improve features that do not respond to levodopa such as freezing, falling, and dementia. The procedure is thus primarily indicated for patients who suffer disability resulting from severe tremor, or levodopa-induced motor complications that cannot be satisfactorily controlled with drug manipulation. In such patients, DBS has been shown to provide benefits in comparison to best medical therapy. Side effects can be seen with respect to the surgical procedure (hemorrhage, infarction, infection), the DBS system (infection, lead break, lead displacement, skin ulceration), or the stimulation (ocular and speech abnormalities, muscle twitches, paresthesias, depression, and rarely suicide). Recent studies indicate that benefits following DBS of the STN and GPi are comparable, but that GPi stimulation may be associated with a reduced frequency of depression. Although not all PD patients are candidates, the procedure can be profoundly beneficial for many. Long-term studies demonstrate continued benefits with respect to the classic motor features of PD, but DBS does not prevent the development of nondopaminergic features, which continue to evolve and to be a source of disability. Studies continue to evaluate the optimal way to use DBS (low- vs high-frequency stimulation, closed loop systems, etc.). Studies of DBS in early PD patients show benefits in comparison to medical therapy, but this must be weighed against the cost of the procedure and the risk of side effects in patients who might otherwise be well controlled with medical therapies. Controlled studies comparing DBS to other therapies aimed at improving motor function without causing dyskinesia, such as Duodopa® and apomorphine infusions, remain to be performed. The utility of DBS may also be reduced in future years if new medical therapies are developed that provide the benefits of levodopa without motor complications. New targets for DBS that might benefit gait dysfunction, depression, and cognitive impairment are being actively explored (Chap. 477).

MRI-guided ultrasound is also now being used as a means of damaging critical target regions such as the GPi in PD patients with motor complications in a noninvasive manner that avoids the needs for a surgical procedure. Preliminary results suggest good target localization and safety.

EXPERIMENTAL THERAPIES FOR PD

There has been considerable scientific and public interest in a number of novel interventions that are being investigated as possible treatments for PD. These include cell-based therapies (such as transplantation of fetal nigral dopamine cells or dopamine neurons derived from stem cells), gene therapies, trophic factors, and therapies directed against gene-specific targets. Transplant strategies are based on the concept of implanting dopaminergic cells into the striatum to replace degenerating SNc dopamine neurons. Fetal nigral mesencephalic cells have been demonstrated to survive implantation, re-innervate the striatum in an organotypic manner, and restore motor function in PD models. However, two double-blind studies failed to show significant benefit of fetal nigral transplantation in comparison to a sham operation with respect to their primary endpoints. Additionally, grafting of fetal nigral cells is associated with a previously unrecognized form of dyskinesia (graft-induced dyskinesia) that persists after lowering or even stopping levodopa. This has been postulated to be related to suboptimal release of dopamine from grafted cells leading to a sustained form of diphasic dyskinesia. In addition, there is evidence that after many years, transplanted healthy embryonic dopamine neurons from unrelated donors develop PD pathology and become dysfunctional, suggesting transfer of α-synuclein from affected to unaffected neurons in a prion-like manner (see discussion above). Perhaps most importantly, it is not clear how replacing dopamine cells alone will improve nondopaminergic features such as falling and dementia, which are the major sources of disability for patients with advanced disease. While stem cells, and specifically induced pluripotent stem cells derived from the recipient, may overcome problems related to immunity, type and number of cells, and physiologic integration, many of these same concerns still apply. To date, stem cells have not yet been properly tested in PD patients and bear the additional concern of tumors and other unanticipated side effects. While there remains a need for scientifically based studies attempting to evaluate the potential role of cell-based therapies in PD, there is no scientific basis to warrant routine treatment of PD patients with stem cells as is being marketed in some countries.

Trophic factors are a series of proteins that enhance neuronal growth and restore function to damaged neurons. There are several different trophic factors that have been demonstrated to have beneficial effects on dopamine neurons in laboratory studies. Glial-derived neurotrophic factor (GDNF) and neurturin have attracted particular attention as possible therapies for PD. However, double-blind trials of intraventricular and intraputaminal infusions of GDNF failed to show benefits compared to placebo in PD patients, possibly because of inadequate delivery of the trophic molecule to the target region.

Gene therapy offers the potential of providing long-term expression of a therapeutic protein with a single procedure. Gene therapy involves placing the DNA of a therapeutic protein into a viral vector that can then be taken up and incorporated into the genome of host cells and then synthesized and released on a continual basis. The AAV2 virus has been most often used as the vector because it does not promote an inflammatory response, is not incorporated into the host genome, and is associated with long-lasting transgene expression. Clinical trials of AAV2 delivery of the trophic factor neurturin showed promising results in open label trials but failed in double-blind trials, even when injected into both the putamen and the SNc. This

likely reflects α-synuclein-mediated downregulation of Nurr1 and RET receptors, thereby limiting the potential of the trophic factor to interact with its receptor and induce upregulation of repair genes. Gene delivery is also being explored as a means of delivering aromatic amino acid decarboxylase with or without tyrosine hydroxylase to the striatum to facilitate the conversion of orally administered levodopa to dopamine. Animal studies suggest that this approach can provide antiparkinsonian benefits with reduced motor complications, and clinical trials in PD patients are underway. Although gene delivery technology has great potential and will likely be used to deliver novel therapies in the future (e.g. Parkin), current approaches carry the risk of unanticipated side effects and do not address the nondopaminergic features of the illness.

MANAGEMENT OF THE NONMOTOR AND NONDOPAMINERGIC FEATURES OF PD

Although PD management has primarily focused on dopaminergic features, management of the nondopaminergic features should not be ignored. Some nonmotor features, although not thought to reflect dopaminergic pathology, nonetheless benefit from dopaminergic drugs. For example, problems such as anxiety, panic attacks, depression, pain, sweating, sensory problems, freezing, and constipation all tend to be worse during "off" periods and have been reported to improve with better dopaminergic control. Approximately 50% of PD patients suffer depression during the course of the disease, and depression is frequently underdiagnosed and undertreated. Antidepressants should not be withheld, particularly for patients with major depression, although dopaminergic agents such as pramipexole may prove helpful for both depression and PD motor features. Serotonin syndromes have been a theoretical concern with the combined use of SSRIs and MAO-B inhibitors but these problems are rarely encountered. Anxiety is also a common problem, and if not adequately controlled with better antiparkinsonian drugs can be treated with short-acting benzodiazepines.

Psychosis can be a problem for some PD patients, and is often a harbinger of developing dementia. In contrast to AD, hallucinations are typically visual, formed, and nonthreatening. Importantly, they can limit the use of dopaminergic agents necessary to obtain satisfactory motor control. They can be associated with dopaminergic drugs, and the first approach is typically to withdraw agents that are less effective than levodopa such as anticholinergics, amantadine, and dopamine agonists followed by lowering the dose of levodopa if possible. Psychosis in PD often responds to low doses of atypical neuroleptics and may permit higher doses of levodopa to be tolerated. Clozapine is an effective drug, but it can be associated with agranulocytosis, and regular monitoring is required. Quetiapine avoids these problems but it has not been established to be effective in placebo-controlled trials. Pimavanserin (Nuplazid®) differs from other atypical neuroleptics in that it is also an inverse agonist of the serotonin 5-HT_{2A} receptor. It has been shown to be effective in double-blind trials with a relatively good safety profile. It was recently approved for use in the United States.

Dementia in PD (PDD) is common, ultimately affecting as many as 80% of patients. Its frequency increases with aging and, in contrast to AD, primarily affects executive functions and attention, with relative sparing of language, memory, and calculation domains. When dementia precedes or develops within 1 year after the onset of motor dysfunction, it is by convention referred to as dementia with Lewy bodies (DLB; **Chap. 426**). These patients are particularly prone to have hallucinations and diurnal fluctuations. Pathologically, DLB is characterized by Lewy bodies distributed throughout the cerebral cortex (especially the hippocampus and amygdala) and is often also associated with AD pathology. It is likely that DLB and PD with dementia represent a spectrum of PD rather than separate disease entities. Mild cognitive impairment (MCI) frequently precedes the onset of dementia and is a more reliable index of impending PDD than in the general population. Dopaminergic drugs can worsen cognitive function in demented patients and should be stopped or reduced to try and provide a compromise between antiparkinsonian benefit and preserved cognitive function. Drugs are usually discontinued in the following sequence: anticholinergics, amantadine, dopamine agonists, COMT inhibitors, and MAO-B inhibitors. Eventually, patients with cognitive impairment should be managed with the lowest dose of standard levodopa that provides meaningful antiparkinsonian effects and does not worsen mental function. Anticholinesterase agents such as memantine, rivastigmine, and donepezil reduce the rate of deterioration of measures of cognitive function and can improve attention in PD, but do not typically improve cognitive function in any meaningful way. More effective therapies that treat or prevent dementia are a critical unmet need in the therapy of PD.

Autonomic disturbances are common and frequently require attention. Orthostatic hypotension can be problematic and contribute to falling. Initial treatment should include adding salt to the diet and elevating the head of the bed to prevent overnight sodium natriuresis. Low doses of fludrocortisone (Florinef) or midodrine provide control for most cases. The norepinephrine precursor 3-0-methylDOPS (Droxidopa®) has been shown to provide mild and transient benefits for patients with orthostatic hypotension, and was recently approved by the U.S. Food and Drug Administration. Vasopressin and erythropoietin can be used in more severe or refractory cases. If orthostatic hypotension is prominent in early parkinsonian cases, a diagnosis of MSA should be considered **(Chap. 432)**. Sexual dysfunction can be helped with sildenafil or tadalafil. Urinary problems, especially in males, should be treated in consultation with a urologist to exclude prostate problems. Anticholinergic agents, such as oxybutynin (Ditropan), may be helpful. Constipation can be a very important problem for PD patients. Mild laxatives or enemas can be useful, but physicians should first ensure that patients are drinking adequate amounts of fluid and consuming a diet rich in bulk with green leafy vegetables and bran. Agents that promote gastrointestinal (GI) motility can also be helpful.

Sleep disturbances are common in PD patients, with many experiencing fragmented sleep with excess daytime sleepiness. Restless leg syndrome, sleep apnea, and other sleep disorders should be treated as appropriate. REM behavior disorder (RBD) is a syndrome composed of violent movements and vocalizations during REM sleep, possibly representing acting out of dreams due to a failure of motor inhibition that typically accompanies REM sleep **(Chap. 27)**. Low doses of clonazepam (0.5–1 mg at bedtime) are usually effective in controlling this problem. Consultation with a sleep specialist and polysomnography may be necessary to identify and optimally treat sleep problems. Many PD patients have a history of RBD preceding the onset of the classic motor features of PD, and most cases of RBD go on to develop an α-synucleinopathy (PD or MSA).

NONPHARMACOLOGIC THERAPY

Gait dysfunction with falling is an important cause of disability in PD. Dopaminergic therapies can help patients whose gait is worse in “off” time, but there are currently no specific therapies for gait dysfunction. Canes and walkers may become necessary to increase stability and reduce the risk of falling. An effective therapy for gait impairment is an important unmet need in PD.

Freezing, where patients suddenly become stuck in place for seconds to minutes as if their feet were glued to the ground, is a major cause of falling. Freezing may occur during “on” or “off” periods. Freezing during “off” periods may respond to dopaminergic therapies, but there are no specific treatments for “on” period freezing. Some patients will respond to sensory cues such as marching in place, singing a song, or stepping over an imaginary line.

Speech impairment is another source of disability for many advanced PD patients. Speech therapy programs may be helpful, but benefits are generally transient.

Exercise has been shown to maintain and even improve function for PD patients, and active and passive exercises with full range of motion reduce the risk of arthritis and frozen joints. Some laboratory studies suggest the possibility that exercise might also have neuroprotective effects, but this has not been confirmed in PD. Exercise is generally recommended for all PD patients. It is less clear that physical therapy or specific exercise programs such as tai chi or dance offer any specific advantage. It is important for patients to maintain social and intellectual activities to the extent possible. Education, assistance with financial planning, social services, and attention to home safety are important elements of the overall care plan. Information is available through numerous PD foundations and on the web, but should be reviewed with physicians to ensure accuracy. The needs of the caregiver should not be neglected. Caring for a person with PD involves a substantial work effort and there is an increased incidence of depression among caregivers. Support groups for patients and caregivers may be useful.

CURRENT MANAGEMENT OF PD

The management of PD should be tailored to the needs of the individual patient, and there is no single treatment approach that is universally accepted and applicable to all individuals. Clearly, if an agent could be demonstrated to have disease-modifying effects, it should be initiated at the time of diagnosis. Indeed, recent studies suggest that dopamine terminal degeneration may be complete within 4 years of diagnosis. Epidemiologic and pathologic studies suggest that constipation, RBD, and anosmia may represent premotor features of PD and could permit diagnosis and the initiation of a disease-modifying therapy even prior to the onset of the classical motor features of the disease. However, no therapy has yet been conclusively proven to be a disease-modifying agent. For now, physicians must use their judgment in deciding whether or not to introduce a drug such as rasagiline (see above) for its possible disease-modifying effects based on available preclinical and clinical information.

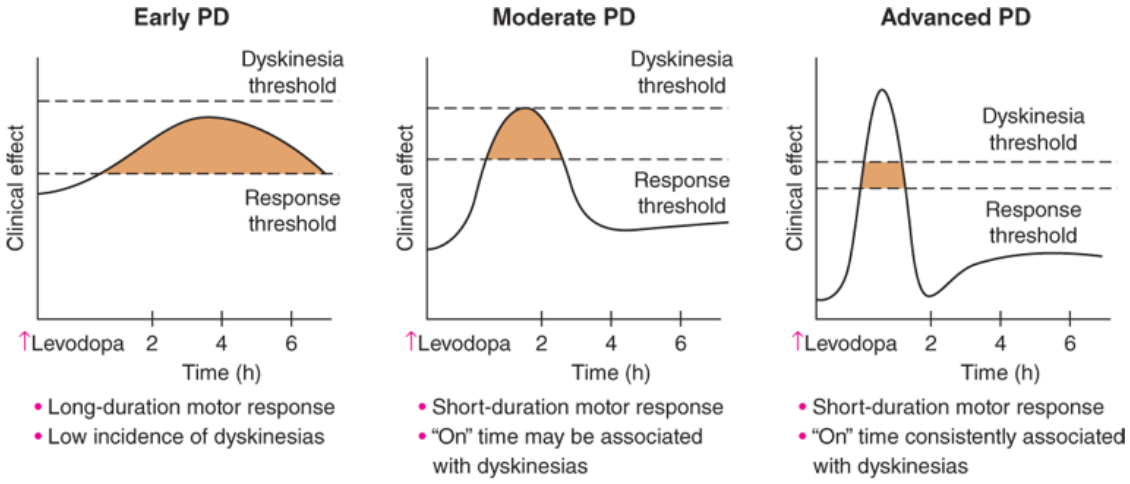
The next important issue to address is when to initiate symptomatic therapy. Several studies suggest that it may be best to start therapy at the time of diagnosis in order to preserve beneficial compensatory mechanisms and possibly provide functional benefits even in the early stage of the disease. Levodopa remains the most effective symptomatic therapy for PD, and some recommend starting it immediately using low doses (≤400 mg/d), as motor complications have now clearly been shown to be dose-related. Others, however, prefer to delay levodopa treatment, particularly in younger patients, in order to reduce the risk of inducing motor complications entirely. An alternate approach is to begin with a MAO-B inhibitor and/or a dopamine agonist, and reserve levodopa for later stages when these drugs no longer provide satisfactory control. In making this decision, the age, degree of disability, and side effect profile of the drug must all be considered. In patients with more severe disability, the elderly, those with cognitive impairment, those with significant comorbidities, or those in whom the diagnosis is uncertain, most physicians would initiate therapy with levodopa. Regardless of initial choice, most patients ultimately require polypharmacy (combination of levodopa, an MAO-B inhibitor, and a dopamine agonist). While it is important to use low doses of each agent in order to reduce the risk of side effects, it is important not to deny patients levodopa when they cannot be adequately controlled with alternative medications.

If motor complications develop, patients can initially be treated by manipulating the frequency and dose of levodopa or by combining lower doses of levodopa with a dopamine agonist, a COMT inhibitor, or a MAO-B inhibitor. Amantadine is the only drug that has been demonstrated to treat dyskinesia without worsening parkinsonism, but benefits may be short-lasting, and there are important side effects related to cognitive function. In advanced cases, it may be necessary to consider a surgical therapy such as DBS or Duodopa® if the patient is a suitable candidate, but as described above, these procedures have their own set of complications. The use of DBS in early PD patients has been advocated by some, but there is considerable skepticism about this approach considering the costs and potential side effects, when inexpensive, well tolerated, and effective medical alternatives are available. Continuous intraintestinal infusion of levodopa/carbidopa intestinal gel (Duodopa) appears to offer similar benefits to DBS, but also requires a surgical intervention with potentially serious complications. Continuous infusion of apomorphine is a treatment option that does not require surgery but is associated with potentially troublesome skin nodules. Comparative studies of these approaches in more advanced patients are awaited. There are ongoing efforts aimed at developing a long-acting oral or subcutaneous formulation of levodopa that mirrors the pharmacokinetic properties of a levodopa infusion. Such a formulation might provide all of the benefits of levodopa without motor complications and avoid the need for polypharmacy and surgical intervention.

A decision tree that considers the various treatment options and decision points for the management of PD is provided in **Fig. 427-7**.

FIGURE 427-6

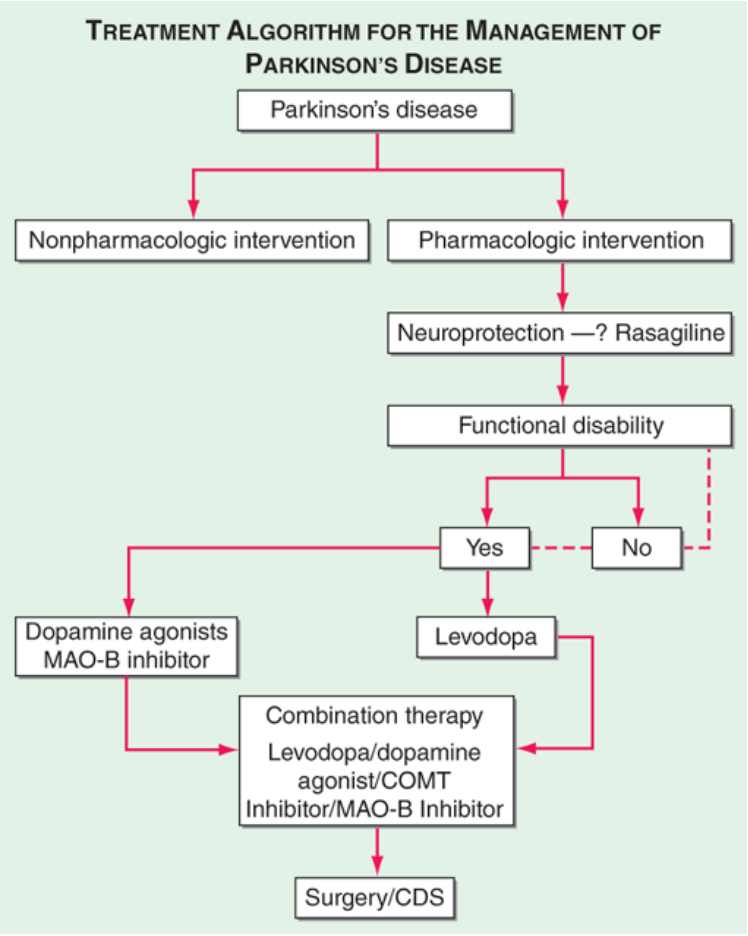
Changes in motor response associated with chronic levodopa treatment. Levodopa-induced motor complications. Schematic illustration of the gradual shortening of the duration of a beneficial motor response to levodopa (wearing off) and the appearance of dyskinesias complicating "on" time. PD, Parkinson's disease.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 427-7

Treatment options for the management of Parkinson's disease (PD). Decision points include: (1) Introduction of a neuroprotective therapy: no drug has been established to have or is currently approved for neuroprotection or disease modification, but there are several agents that have this potential based on laboratory and preliminary clinical studies (e.g., rasagiline 1 mg/d, coenzyme Q10 1200 mg/d, the dopamine agonists ropinirole, and pramipexole). (2) When to initiate symptomatic therapy: There is a trend toward initiating therapy at the time of diagnosis or early in the course of the disease because patients may have some disability even at an early stage, and there is the possibility that early treatment may preserve beneficial compensatory mechanisms; however, some experts recommend waiting until there is functional disability before initiating therapy. (3) What therapy to initiate: many experts favor starting with a monoamine oxidase type B (MAO-B) inhibitor in mildly affected patients because of the good safety profile of the drug and the potential for a disease-modifying effect; dopamine agonists for younger patients with functionally significant disability to reduce the risk of motor complications; and levodopa for patients with more advanced disease, the elderly, or those with cognitive impairment. Recent studies suggest the early employment of polypharmacy using low doses of multiple drugs to avoid side effects associated with high doses of any one agent. (4) Management of motor complications: motor complications are typically approached with combination therapy to try and reduce dyskinesia and enhance the "on" time. When medical therapies cannot provide satisfactory control, surgical therapies such as DBS or continuous infusion of levodopa/carbidopa intestinal gel can be considered. (5) Nonpharmacologic approaches: interventions such as exercise, education, and support should be considered throughout the course of the disease. CDS, continuous dopaminergic stimulation; COMT, catechol-*O*-methyltransferase. *(Adapted from CW Olanow et al: Neurology 72:S1, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 427-5

Drugs Commonly Used for Treatment of Parkinson's Disease[a]

Agent	Available Dosages	Typical Dosing
Levodopa[a]		
Carbidopa/levodopa	10/100, 25/100, 25/250 mg	200–1000 mg levodopa/day
Benserazide/levodopa	25/100, 50/200 mg	
Carbidopa/levodopa CR	25/100, 50/200 mg	
Benserazide/levodopa MDS	25/200, 25/250 mg	
Parcopa	10/100, 25/100, 25/250	
Rytary (carbidopa/levodopa)	23.75/95, 36.25/145, 48.75/195, 61.25/245	See conversion tables
Carbidopa/levodopa/entacapone	12.5/50/200, 18.75/75/200, 25/100/200, 31.25/125/200, 37.5/150/200, 50/200/200 mg	
Dopamine agonists		
Pramipexole	0.125, 0.25, 0.5, 1.0, 1.5 mg	0.25–1.0 mg tid
Pramipexole ER	0.375, 0.75, 1.5. 3.0, 4.5 mg	1–3 mg/d
Ropinirole	0.25, 0.5, 1.0, 3.0 mg	6–24 mg/d
Ropinirole XL	2, 4, 6, 8 mg	6–24 mg/d
Rotigotine patch	2-, 4-, 6-, 8-mg patches	4–24 mg/d
Apomorphine SC	2-8 mg	2–8 mg
COMT inhibitors		
Entacapone	200 mg	200 mg with each levodopa dose
Tolcapone	100, 200 mg	100–200 mg tid
Opicapone	50 mg	50 mg HS
MAO-B inhibitors		
Selegiline	5 mg	5 mg bid
Rasagiline	0.5, 1.0 mg	mg QAM
Safinamide	100 mg	100 mg QAM

[a]Treatment should be individualized. Generally, drugs should be started in low doses and titrated to optimal dose.

Note: Drugs should not be withdrawn abruptly but should be gradually lowered or removed as appropriate.

Abbreviations: COMT, catechol-*O*-methyltransferase; MAO-B, monoamine oxidase type B; QAM, every morning.

FURTHER READING

Ascherio A, Schwarzschild MA: The epidemiology of Parkinson's disease: Risk factors and prevention. Lancet Neurol 15:1257, 2016. [PubMed: 27751556]

Berg D et al: MDS research criteria for prodromal Parkinson's disease. Mov Disord 12:1600, 2015.

Dorsey ER et al: Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. Neurology 68:384, 2007. [PubMed: 17082464]

Hernandez DG et al: Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. J Neurochem 139 Suppl 1:59, 2016. [PubMed: 27090875]

Höglinger GU et al: Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria. Movement Disorder Society-endorsed PSP Study Group. Mov Disord 32:853, 2017.

Marras C et al: Nomenclature of genetic movement disorders: Recommendations of the International Parkinson and Movement Disorder Society task force. Mov Disord 32:724, 2017. [PubMed: 28513081]

Obeso JA et al: Past, present and future of Parkinson's disease: A special essay on the 200th Anniversary of the Shaking Palsy. Mov Disord 32:1264, 2017. [PubMed: 28887905]

Olanow CW et al: Scientific and clinical basis for the treatment of PD—2009. Neurology 72:S1, 2009. [PubMed: 19470958]

Postuma RB et al: MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 12:1591, 2015.

Schapira AH et al: Slowing of neurodegeneration in Parkinson's disease and Huntington's disease: Future therapeutic perspectives. Lancet 384:545, 2014. [PubMed: 24954676]

Schapira AHV et al: Non-motor features of Parkinson disease. Nat Rev Neurosci 18:435, 2017. [PubMed: 28592904]

Chapter 428: Tremor, Chorea, and Other Movement Disorders

C. Warren Olanow; Christine Klein; Jose A. Obeso

FIGURE 428-1

HYPERKINETIC MOVEMENT DISORDERS

Hyperkinetic movement disorders are characterized by involuntary movements unaccompanied by weakness (Table 428-1). This term is somewhat arbitrary and potentially misleading as hypokinetic disorders such as Parkinson's disease (PD) are often accompanied by tremor which is a hyperkinetic feature, and hyperkinetic disorders such as dystonia may manifest slow movements because of the severe muscle contractions. Nonetheless, the terms continue to be used because of convention. The major hyperkinetic movement disorders and the diseases with which they are associated are considered in this section.

TABLE 428-1

Hyperkinetic Movement Disorders

Tremor	Rhythmic oscillation of a body part due to intermittent muscle contractions
Dystonia	Involuntary, patterned, sustained, or repeated muscle contractions often associated with twisting movements and abnormal posture
Athetosis	Slow, distal, writhing, involuntary movements with a propensity to affect the arms and hands (this represents a form of dystonia with increased mobility)
Chorea	Rapid, semi-purposeful, graceful, dance-like nonpatterned involuntary movements involving distal or proximal muscle groups. When the movements are of large amplitude and predominant proximal distribution, the term *ballism* is used.
Myoclonus	Sudden, brief (<100 ms), jerk-like, arrhythmic muscle twitches
Tic	Brief, repeated, stereotyped muscle contractions that can often be suppressed for a short time. These can be simple and involve a single muscle group or complex and affect a range of motor activities.

TREMOR

CLINICAL FEATURES

Tremor consists of alternating contractions of agonist and antagonist muscles in an oscillating, rhythmic manner. It can be most prominent at rest (rest tremor), on assuming a posture (postural tremor), on actively reaching for a target (kinetic tremor), or on carrying out a movement (action tremor). Tremor may also be characterized based on distribution, frequency, amplitude, and related neurologic dysfunction.

PD (Chap. 427) is characterized by a resting tremor, essential tremor (ET) by a tremor that typically occurs while trying to sustain a posture coupled with an action tremor, and cerebellar dysfunction by a kinetic tremor and is usually associated with hypotonia and past pointing. Normal individuals can have a physiologic tremor that typically manifests as a mild, high-frequency (10–12 Hz), postural or action tremor typically affecting the upper extremities. This tremor is usually of no clinical consequence and often is only appreciated with an accelerometer or under stress. An enhanced physiologic tremor (EPT) can be seen in up to 10% of the population, and tends to occur in association with anxiety, fatigue, a metabolic disturbance (e.g., hyperthyroidism, electrolyte abnormalities), drugs (e.g., valproate, lithium), or toxins (e.g., caffeine,

smoking, alcohol). Treatment is initially directed at control of any underlying disorder and, if necessary, can often be improved with a beta blocker.

ESSENTIAL TREMOR

ET is the most common movement disorder, affecting ~5% of the population (an estimated 5–10 million persons in the United States or Western Europe). It can present in childhood but dramatically increases in prevalence in those aged >70 years. ET is characterized by a high-frequency tremor (6–10 Hz) that predominantly affects the upper extremities. The tremor is most often manifest as a postural or action tremor and, in severe cases, can interfere with functions such as eating and drinking. It is typically bilateral and symmetric but may begin on one side and remain asymmetric. Patients with severe ET can have an intention tremor with overshoot and slowness of movement suggesting the possibility of a cerebellar origin. Tremor involves the head in ~30% of cases, voice in ~20%, tongue in ~20%, face/jaw in ~10%, and lower limbs in ~10%. Multiple body parts are involved in at least 50% of cases. The tremor is characteristically improved by alcohol and worsened by stress. Subtle impairment of coordination or tandem walking may be present, and disturbances of hearing, cognition, personality, mood, and olfaction have been described, but usually the neurologic examination is normal aside from tremor. The differential diagnosis includes dystonic tremor (see below) or PD. PD can usually be differentiated from ET because the former stops at the onset of a voluntary action and is typically associated with bradykinesia with progressive slowing of sequential movements (sequence effect), rigidity, gait and postural instability, and other parkinsonian features. However, the examiner should be aware that PD patients may have a postural tremor and ET patients may develop a rest tremor, and that these typically begin after a latency of a few seconds (emergent tremor). In contrast to the micrographia of PD, ET patients have relatively large handwriting with evidence of the effect of tremor. The examiner must also differentiate the effect of tremor when assessing tone in ET to distinguish this from the cogwheel rigidity found in PD.

ETIOLOGY AND PATHOPHYSIOLOGY

The etiology and pathophysiology of ET are not known. Approximately 50% of cases have a positive family history with an autosomal dominant pattern of inheritance. Linkage studies have detected possibly linked loci in large ET families, but no independently confirmed causative genes have been identified to date. It is likely, however, that there are undiscovered genes for ET that have escaped detection to date because of the heterogeneity of the syndrome and the high population frequency of ET likely resulting in a large number of phenocopies, (i.e., family members with a similar clinical syndrome, but not carrying the causative mutation). The cerebellum and inferior olives have been implicated as possible sites of a “tremor pacemaker” based on the presence of cerebellar signs in about 10% of ET patients, and increased metabolic activity and blood flow in these regions in some patients. Some pathologic studies have described cerebellar pathology with a loss of Purkinje cells and axonal torpedoes, but these findings are controversial, and the precise pathologic correlate of ET remains to be defined. It is likely that multiple causes of ET will ultimately be identified.

TREATMENT

Many cases are mild and require no treatment other than reassurance. Occasionally, tremor can be severe and interfere with eating, writing, and activities of daily living. This is more likely to occur as the patient ages and is often associated with a reduction in tremor frequency. Beta blockers and primidone are the standard drug therapies for ET and help in about 50% of cases. Propranolol (20–120 mg daily, given in divided doses) is usually effective at relatively low doses, but higher doses may be needed in some patients. The drug is contraindicated in patients with bradycardia or asthma. Hand tremor tends to be most improved, while head tremor is often refractory. Primidone can be helpful but should be started at low doses (12.5 mg) and gradually increased (125–250 mg tid) to avoid sedation, nausea, and dizziness. Benefits have also been reported with gabapentin and topiramate, but these drugs have not been widely employed. Botulinum toxin injections may be helpful for limb or voice tremor, but treatment can be associated with muscle weakness. Surgical therapies targeting the ventro-intermediate (VIM) nucleus of the thalamus can be very effective for severe and drug-resistant cases. Recently, focal ultrasound (which is a procedure that does not require surgery) has also been shown to be an effective therapy against tremor in ET.

DYSTONIA

CLINICAL FEATURES

Dystonia is a movement disorder characterized by sustained or intermittent muscle contractions of antagonist muscles causing abnormal often repetitive movements and postures. Dystonic movements are typically patterned and twisting and may be associated with a “dystonic tremor”. Dystonia is often initiated or worsened by voluntary action and associated with overflow muscle activation. Dystonia can range from focal minor contractions affecting only an individual muscle group to severe and disabling involvement of multiple muscle groups. The frequency is estimated to be 16 per 100,000 (~50,000 cases in the United States) but is likely to be much higher because many cases are not recognized. Dystonia is often brought out by voluntary movements (action dystonia) and can extend to involve other muscle groups and body regions not required for a given action (overflow). It can be aggravated by stress and fatigue and attenuated by relaxation and sensory tricks such as touching the affected body part (geste antagoniste).

Historically, dystonia has been described as primary or secondary. However, because of a confusing and not always congruent combination of phenotypic and etiologic features, the older terms are no longer recommended. A Movement Disorder Society Task Force charged with redefining dystonia recommends classifying dystonia along two main axes: clinical and etiologic. On clinical grounds, dystonia can be categorized by age of onset (infancy, childhood, adolescence, early and late adulthood), body distribution (focal, segmental, multifocal, and generalized), temporal pattern (static or progressive, action-specific [diurnal and paroxysmal]), and association with additional features. Clinical description along these lines enables formulating specific dystonia syndromes (e.g., early-onset generalized isolated dystonia).

Etiology of dystonia primarily reflects genetic abnormalities, although occasionally there may be other causes such as trauma and stroke. Genetic features used for classification include mode of inheritance or

identification of a specific genetic defect. In the past three decades, more than 200 genes have been linked to different, mainly childhood-onset and generalized forms of dystonia. These include forms in which dystonia is the only disease manifestation with the exception of tremor ("isolated dystonia"), forms in which dystonia co-occurs with another movement disorder such as parkinsonism or myoclonus ("combined dystonia") and disorders in which dystonia is only one of several clinical manifestations and may be a less prominent or even inconsistent feature ("complex dystonia"). Most of the genetic forms belong to the latter phenotypic group, which also represents the most heterogeneous class in terms of clinical expression.

ISOLATED DYSTONIAS

Focal (Multifocal, Segmental) Dystonia

Adult-onset, focal dystonia is by far the most frequent form of isolated dystonia, with women being affected about twice as often as men. Focal dystonia typically presents in the fourth to sixth decade and can be focal, multifocal, or segmental. The major clinical phenotypes are as follows: (1) *Cervical dystonia*—dystonic contractions of neck muscles causing the head to deviate to one side (*laterocollis*), twist (*torticollis*), move in a forward direction (*anterocollis*), or move in a backward direction (*retrocollis*). Muscle contractions can be painful and occasionally can be complicated with a secondary cervical radiculopathy. (2) *Blepharospasm*—dystonic contractions of the eyelids with increased blinking that can interfere with reading, watching television, working on a computer, and driving. This can sometimes be so severe as to cause functional blindness. (3) *Oromandibular dystonia* (OMD)—contractions of muscles of the lower face, lips, tongue, and jaw (opening or closing). Meige's syndrome is a combination of OMD and blepharospasm that predominantly affects women aged >60 years. (4) *Spasmodic dysphonia*—dystonic contractions of the vocal cords during phonation, causing impaired speech. Most cases affect the adductor muscles and cause speech to have a choking or strained quality. Less commonly, the abductors are affected, leading to speech with a breathy or whispering quality. (5) *Limb dystonias*—these can be present in either arms or legs and are often brought out by task-specific activities such as handwriting (writer's cramp), playing a musical instrument (musician's cramp), or putting in golf (the yips). The vast majority of patients have dystonia of the neck (cervical dystonia; ~50%) or the eye lid (blepharospasm; ~20%). Focal hand or leg dystonia (~5%), spasmodic dysphonia (~2%), musician's dystonia (~3%), or OMD (~1%) are much less common. Focal dystonias can extend to involve other body regions (about 30% of cases) and are frequently misdiagnosed as psychiatric or orthopedic in origin. Their cause is usually not known, but genetic factors, autoimmunity, and trauma have been suggested. Focal dystonias are often associated with a high-frequency tremor that can resemble ET. Dystonic tremor can usually be distinguished from ET because it tends to occur in conjunction with the dystonic contraction and disappears when the dystonia is relieved (e.g., turning the head in the opposite direction of the dystonia).

GENERALIZED DYSTONIA

Generalized dystonia is often hereditary in nature and, unlike focal dystonia, generally has an age of onset in childhood or adolescence. There are currently at least four well established genes for *isolated* dystonia; *TOR1A, THAP1, ANO3*, and *GNAL*. According to the recommendations of the International Parkinson's Disease

and Movement Disorder Society, confirmed monogenic forms are classified according to absence or presence of accompanying clinical features and preceded by a "DYT" prefix, e.g., DYT-TOR1A. These genetic forms are all inherited in an autosomal dominant fashion and found in <5% of dystonia patients. Not all mutation carriers develop generalized dystonia; about 35% remain unaffected despite harboring a pathogenic mutation (reduced penetrance), and rarely they present with dystonia that remains focal or segmental in nature.

Mutations in the *TOR1A* gene (torsin family 1 member A—*formerly known as the DYT1 gene*) are the most common cause of early-onset generalized dystonia. The first, and currently the only clearly established mutation, is a 3-base pair deletion in the *TOR1A* gene. The mutation is frequently found among Ashkenazi Jewish patients due to a founder effect. Mutation carriers usually present with dystonia in an extremity in childhood that later progress to other body parts, but typically spare the face and neck.

THAP1 gene (*THAP domain containing, apoptosis associated protein 1*) mutations have been linked to adolescent-onset dystonia with mixed phenotype. About 100 different mutations have been reported in *THAP1*. Mutations typically manifest with dysphonia or writer's cramp beginning in late childhood or adolescence. Over the course of the disease, dystonia spreads to other body parts with prominent craniocervical involvement.

Mutations in the *ANO3* gene (*anoctamin 3*) were first reported in patients with predominantly craniocervical dystonia with a broad range of ages of onset. While a large number of missense variants can be found in healthy individuals, a pathogenic role of *ANO3* mutations has recently been supported by the description of additional families with dystonia and myoclonic jerks.

Mutations in the *GNAL* gene (*guanine nucleotide-binding protein subunit alpha L*) are a rare cause of cervical or cranial dystonia with a mean age of onset in the thirties. About 30 different *GNAL* mutations have been reported in dystonia patients.

In addition to the above, missense mutations in *KMT2B* (*lysine methyltransferase 2B*) have recently been identified, and confirmed to be a cause of an early-onset generalized dystonia which may be accompanied by other syndromic features including intellectual disability, microcephaly, psychiatric features, dysmorphia, or skin lesions. The majority of the mutations occurred de novo. *KMT2B* mutations may account for up to 10% of early-onset generalized dystonia but further validation is warranted and placement into the group of isolated vs complex dystonias is currently under debate.

Combined Dystonia

A number of other well-established genes have been described for combined forms of dystonia in which dystonia occurs in conjunction with a different movement disorder, such as parkinsonism or myoclonus.

Dopa-responsive dystonia (DRD; also known as Segawa syndrome) is caused by mutations in the *GCH1* gene (*GTP cyclohydrolase-1*) that encodes for the rate-limiting enzyme in the biosynthesis of dopamine via the biopterin pathway. It is manifest as a childhood-onset form of dystonia with diurnal fluctuations and is

important to recognize as the condition dramatically responds to low doses of levodopa. Parkinsonism can be a major, or even the only finding, and there may be a presynaptic dopaminergic deficit as evidenced by SPECT. To date, more than 100 different mutations have been reported with a penetrance of around 50% which is considerably higher in women compared to men. Recessively inherited (biallelic) mutations in *GCH1* result in a much more severe clinical phenotype with developmental delay and infantile onset. Due to the enzymatic defect in the levodopa biosynthesis, there is a lifelong and dramatic response to levodopa therapy. Indeed, all young onset forms of dystonia should be tested with levodopa to exclude the possibility of DRD.

X-linked dystonia-parkinsonism (Lubag) is a combined form of dystonia and parkinsonism that is found exclusively in patients of Filipino origin due to a founder effect and seems to be fully penetrant. Patients usually develop focal (cranial) dystonia first that rapidly generalizes and, after 5–10 years, is gradually replaced by a form of L-dopa-unresponsive parkinsonism. The exact mutation causing X-linked dystonia-parkinsonism (Lubag) is not yet known but several variants in a disease haplotype segregate with the disease and a retrotransposon insertion in the *TAF1* (*TATA-Box Binding Protein Associated Factor 1*) gene has been suggested as the most likely disease cause.

Biallelic mutations in the *PRKRA* (*protein activator of interferon-induced protein kinase EIF2AK2*) gene are linked to a dystonia-parkinsonism syndrome and mostly due to the same missense mutation that seems to result from a shared founder. The phenotype includes early-onset generalized dystonia, often with laryngeal dystonia, tongue protrusion, prominent oromandibular involvement, dysphagia, and retrocollis. Parkinsonian features are mild (or even absent) and do not respond to levodopa therapy.

Mutations in the *ATP1A3* (*ATPase Na^+/K^+ transporting subunit alpha 3*) gene present with a characteristic, sudden onset usually in adolescence or young adulthood, often triggered by high fever, physical exertion, or emotional stress. Dystonic symptoms frequently show a rostrocaudal gradient with a strong involvement of the bulbar region and are often accompanied by bradykinesia as a parkinsonian feature. In addition, mutations in *ATP1A3* have also been linked to a variety of clinical syndromes (pleiotropy) including epileptic or hemiplegic attacks, ataxia, cognitive decline, and other neurological disorders, often with a more severe course and an earlier age at onset.

Myoclonic-dystonia is characterized by action-induced, alcohol-responsive myoclonic jerks predominantly involving the upper body half. Onset is usually in childhood or adolescence. Additionally, many individuals develop psychiatric features such as depression, anxiety-related disorders, and alcohol dependence. The disorder is primarily related to mutations in the *SGCE* gene (*sarcoglycan epsilon*) which codes for the ε member of the sarcoglycan family. About 80 different mutations have been reported in *SGCE* including deletions of the entire gene. The latter type of mutation often also involves loss of adjacent genes leading to additional clinical features such as joint problems. *SGCE* mutations are incompletely penetrant and only manifest when inherited from the father due to the epigenetic effect of maternal imprinting of *SGCE*.

A number of additional monogenic causes have been suggested for isolated and combined forms of dystonia but still await independent confirmation. **Table 428-2** provides a list of the confirmed Monogenic Forms of

Isolated and Combined Dystonias.

TABLE 428-2

Confirmed Monogenic Forms of Isolated and Combined Dystonia[a]

Form of Dystonia		Gene	Locus Name	Designation and Phenotypic Subgroup[a]	Additional Distinguishing Features	MOI
Isolated		*TOR1A*	DYT1	DYT-TOR1A	Childhood or adolescent-onset, generalized	AD
		THAP1	DYT6	DYT-THAP1	Adolescent-onset, cranial or generalized	AD
		ANO3	DYT24	DYT-ANO3	Adult-onset, focal or segmental	AD
		GNAL	DYT25	DYT-GNAL	Mostly adult-onset, focal or segmental	AD
		KMT2B[b]	DYT28	DYT-KMT2B	Early-onset, generalized, mild syndromic features	AD
Combined	**Dystonia plus Parkinsonism**	*GCH1*	DYT5a	DYT-GCH1	Dopa-responsive	AD
		TAF1	DYT3	DYT-TAF1	Neurodegeneration	XL
		PRKRA	DYT16	DYT-PRKRA	Dystonia with mild parkinsonism	AR
		ATP1A3	DYT12	DYT-ATP1A3	Rapid-onset	AD
	Dystonia plus Myoclonus	*SGCE*	DYT11	DYT-SGCE	Psychiatric disease	AD

[a]According to C Marras et al: Mov Disord 31:436, 2016. [b]Several, but not all, patients show syndromic features; DYT-KMT2B may thus be better placed with the complex dystonias.

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; MOI, mode of inheritance; XL, X-linked.

COMPLEX DYSTONIAS

In the complex dystonias, dystonia is one part of a syndrome with multiple different disease manifestations. Most frequently, they are hereditary such as Wilson's disease (WD), Huntington's disease (HD), Lesh Nyhan syndrome, corticobasal ganglionic disorders, and a variety of other neurologic, neurometabolic, and mitochondrial disorders. Complex dystonias may also develop as a consequence of drugs or toxins (previously referred to as secondary dystonias). Drug-induced dystonia may be acute or chronic, and is most commonly seen with neuroleptic drugs or after chronic levodopa treatment in PD patients. Dystonia can also be observed following discrete lesions in the striatum, and occasionally in the pallidum, thalamus, cortex, and brainstem due to infarction, hemorrhage anoxia, trauma, tumor, infection, or toxins such as manganese or carbon monoxide. In these cases, dystonia often assumes a segmental distribution, but may be generalized when lesions are bilateral or widespread. More rarely, dystonia can develop following peripheral nerve injury and be associated with features of complex regional pain syndrome **(Chap. 432)**. A psychogenic origin is responsible for some cases of dystonia; these typically present with fixed, immobile dystonic postures (see below).

PATHOPHYSIOLOGY OF DYSTONIA

The pathophysiologic basis of dystonia is not completely known. The phenomenon is characterized by co-contracting synchronous bursts of agonist and antagonist muscle groups with recruitment of muscle groups that are not required for a given movement (overflow). Dystonia is characterized by derangement of the basic physiological principle of action-selection, leading to abnormal recruitment of inappropriate muscles for a given action with inadequate inhibition of this undesired motor activity. Physiologically, loss of surround inhibition is observed at multiple levels of the motor system (e.g., cortex, brainstem, spinal cord) accompanied by increased cortical excitability and reorganization. Attention has focused on the basal ganglia as the site of origin of at least some types of dystonia because there are alterations in blood flow and metabolism in these structures. Further, lesions of the basal ganglia (particularly the putamen) can induce dystonia, and surgical ablation or deep brain stimulation (DBS) of specific regions of the globus pallidus may ameliorate dystonia. The dopamine system has also been implicated, because dopaminergic therapies can both induce and treat some forms of dystonia in different circumstances. Interestingly, no specific pathology has been consistently identified in dystonia.

TREATMENT

TREATMENT

Dystonia

Treatment of dystonia is for the most part symptomatic except in rare cases where correction of a primary underlying condition is possible. Wilson's disease should be ruled out in young patients with dystonia. Levodopa should be tried in all cases of childhood-onset dystonia to test for DRD. High-dose anticholinergics (e.g., trihexyphenidyl 20–120 mg/d) may be beneficial in children, but adults can rarely tolerate high doses

because of side effects related to cognitive impairment and hallucinations. Oral baclofen (20–120 mg) may also be helpful, but benefits, if present, are usually modest, and side effects of sedation, weakness, and memory loss can be problematic. Intrathecal infusion of baclofen is more likely to be useful, particularly for leg and trunk dystonia, but benefits are frequently not sustained, and complications can be serious and include infection, seizures, and coma. Tetrabenazine is another consideration (the usual starting dose is 12.5 mg/d and the average treating dose is 25–75 mg/d), but its use may be limited by sedation and the development of parkinsonism. Neuroleptics can improve as well as induce dystonia, but they are typically not recommended because of their potential to induce parkinsonism and other movement disorders, including tardive dystonia. Clonazepam and diazepam are rarely effective.

Botulinum toxin has become the preferred treatment for patients with focal dystonia, particularly where involvement is limited to small muscle groups such as in blepharospasm, torticollis, and spasmodic dysphonia. Botulinum toxin acts by blocking the release of acetylcholine at the neuromuscular junction, leading to reduced dystonic muscle contractions. However, treatment with botulinum toxin can be complicated by excessive weakness that can be troublesome, particularly if it involves neck and swallowing muscles. Two serotypes of botulinum toxin are currently available (A and B). Both are effective, and it is not clear that there are advantages of one over the other. No systemic side effects are encountered with the doses typically used, but benefits are transient, and repeat injections are required at 2–5 month intervals. Some patients fail to respond after having experienced an initial benefit. This has been attributed to antibody formation, but improper muscle selection, injection technique, and inadequate dose should be excluded.

Surgical therapy is an alternative for patients with severe dystonia who are not responsive to other treatments. Peripheral procedures such as rhizotomy and myotomy were used in the past to treat cervical dystonia, but are now rarely employed. DBS of the pallidum can provide dramatic benefits for some patients with various forms of hereditary and nonhereditary generalized dystonia. This represents a major therapeutic advance because previously there was no consistently effective therapy, especially for patients with severe disability. Benefits tend to be obtained with a lower frequency of stimulation and often occur after a relatively longer latency (weeks to months) than in PD. Better results are typically obtained in younger patients with shorter disease duration. Recent studies suggest that DBS may also be valuable for patients with focal and secondary dystonias, although results are less consistent. Supportive treatments such as physical therapy and education should be a part of the treatment regimen.

Physicians should be aware of dystonic storm, a rare but potentially fatal condition that can occur in response to a stress situation such as surgery or a systemic infection in patients with preexisting dystonia. It consists of the acute onset of generalized and persistent dystonic contractions that can involve the vocal cords or laryngeal muscles, leading to airway obstruction. Patients may experience rhabdomyolysis with renal failure and should be managed in an intensive care unit with airway protection if required. Treatment can be instituted with one or a combination of anticholinergics, diphenhydramine, baclofen, benzodiazepines, and dopaminergic agents. Spasms may be difficult to control, and anesthesia with muscle paralysis may be required.

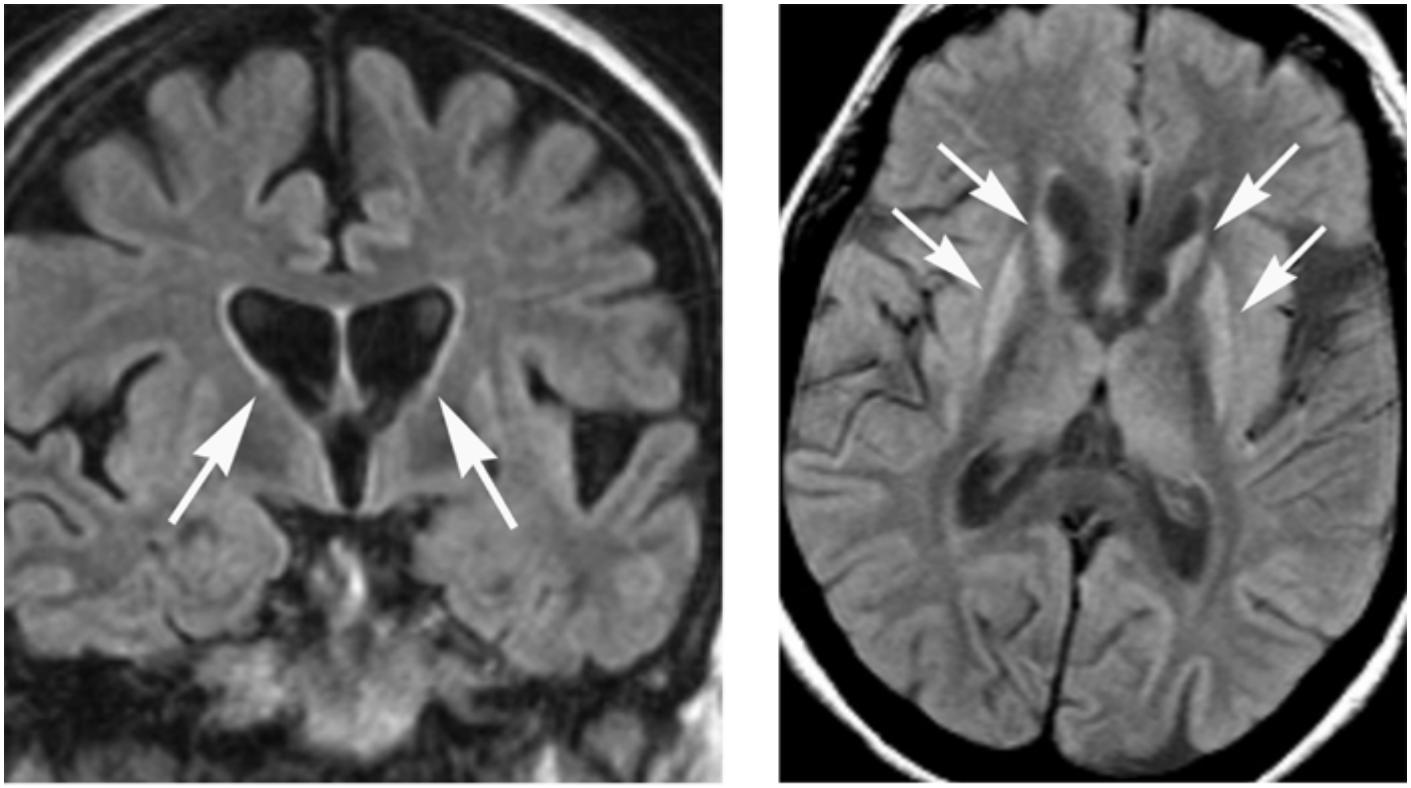
CHOREAS

HUNTINGTON'S DISEASE

HD is a progressive, fatal, highly penetrant autosomal dominant disorder characterized by motor, behavioral, oculomotor, and cognitive dysfunction. The disease is named for George Huntington, a family physician who described cases on Long Island, New York, in the nineteenth century. Onset is typically between the ages of 25 and 45 years (range, 3–70 years) with a prevalence of 2–8 cases per 100,000 and an average age at death of 60 years. It is prevalent in Europe, North America, South America, and Australia but is rare in African blacks and Asians. HD is characterized by rapid, nonpatterned, semi-purposeful, involuntary choreiform movements, and for this reason was formerly referred to as Huntington's chorea. However, dysarthria, gait disturbance, oculomotor abnormalities, behavioral disturbance, and cognitive impairment with dementia are also common features, thus the condition is currently referred to as HD. In the early stages, chorea tends to be focal or segmental, but progresses over time to involve multiple body regions. With advancing disease, there tends to be a reduction in chorea and the emergence of dystonia, rigidity, bradykinesia, and myoclonus. Functional decline is often predicted by progressive weight loss despite adequate calorie intake. In younger patients (~10% of cases), HD can present as an akinetic-rigid or parkinsonian syndrome (Westphal variant). HD patients eventually develop behavioral and cognitive disturbances, and the majority progress to dementia. Depression with suicidal tendencies, aggressive behavior, and psychosis can be prominent features. HD patients may also develop noninsulin-dependent diabetes mellitus and neuroendocrine abnormalities (e.g., hypothalamic dysfunction). A clinical diagnosis of HD can be strongly suspected in cases of chorea with a positive family history, but genetic testing provides the ultimate confirmation of the diagnosis. The disease predominantly affects the striatum but progresses to involve the cerebral cortex and other brain regions. Progressive atrophy of the head of the caudate nucleus, which form the lateral margin of the lateral ventricle, can be visualized by MRI **(Fig. 428-1)**, but the putamen can be equally or even more severely affected. More diffuse cortical atrophy can be seen in the middle and late stages of the disease. Supportive studies include reduced metabolic activity in the caudate nucleus and putamen, and reduced brain metabolites on MR spectroscopy. Genetic testing can be used to confirm the diagnosis and to detect at-risk individuals in the family, but must be performed with caution and in conjunction with trained counselors, because positive results can worsen depression and generate suicidal reactions. The neuropathology of HD consists of prominent neuronal loss and gliosis in the caudate nucleus and putamen; similar changes are also widespread in the cerebral cortex. Intraneuronal inclusions containing aggregates of ubiquitin and the mutant protein huntingtin are found in the nuclei of affected neurons.

FIGURE 428-1

Huntington's disease. ***A.*** Coronal fluid attenuated inversion recovery (FLAIR) magnetic resonance imaging shows enlargement of the lateral ventricles reflecting typical atrophy (*arrows*). ***B.*** Axial FLAIR image demonstrates abnormal high signal in the caudate and putamen (*arrows*).



A B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In anticipation of developing neuroprotective therapies, there has been an intensive effort to define the premanifest stage of HD. Subtle motor impairment, cognitive alterations, and imaging changes can be detected in at-risk individuals who later go on to develop the manifest form of the disease. Defining the rate of progression of these features is paramount for future studies of putative disease-modifying therapies designed to slow the rate of disease progression and the development of cumulative disability.

ETIOLOGY

HD is caused by an increase in the number of polyglutamine (CAG) repeats (>40) in the coding sequence of the *huntingtin* gene located on the short arm of chromosome 4. The larger the number of repeats, the earlier the disease is manifest. Intermediate forms of the disease with 36–39 repeats are described in some patients, typically with less severe clinical involvement. Acceleration of the process tends to occur, particularly in males, with subsequent generations having larger numbers of repeats and earlier age of disease onset, a phenomenon referred to as anticipation. The gene encodes the highly conserved cytoplasmic protein huntingtin, which is widely distributed in neurons throughout the central nervous system (CNS), but whose function is largely unknown. Mitochondrial dysfunction has been demonstrated in the striatum and skeletal muscle of symptomatic and presymptomatic individuals. Fragments of the mutant huntingtin protein can be toxic, possibly by translocating into the nucleus and interfering with transcriptional regulation of proteins. Neuronal inclusions found in affected regions in HD may represent a protective mechanism aimed at segregating and facilitating the clearance of these toxic proteins. There is also interest in the possibility that protein accumulation and aggregation in HD, like Alzheimer's disease **(Chap. 423)** and PD **(Chap. 427)**, may be

critical to the disease process and reflect a prion-like disorder (Chap. 430). Models of HD with striatal pathology can be induced in transgenic animals that express the mutant gene and by excitotoxic agents such as kainic acid and 3-nitropropionic acid which promote calcium entry into the cell and cytotoxicity.

TREATMENT

TREATMENT

Huntington's Disease

Although the gene for HD was identified 25 years ago, there is still no disease-modifying therapy for this disorder and symptomatic treatment is limited. Current treatment involves a multidisciplinary approach, with medical, neuropsychiatric, social, and genetic counseling for patients and their families. Dopamine-blocking agents may control the choreatic movements. Tetrabenazine (a presynaptic dopamine depleting agent) has been approved for the treatment of chorea, but can cause secondary parkinsonism. More recently, deuterated tetrabenazine (Austedo™) has been approved as a treatment for chorea in HD. Deuteration interferes with the metabolism of tetrabenazine and avoids a high Cmax. In clinical trials, it has been shown to have fewer dose-related side effects than tetrabenazine, and therefore can be administered in higher doses with potentially superior clinical benefits. Neuroleptics are generally not recommended because of their potential to induce other troubling movement disorders and because HD chorea tends to be self-limited and is usually not disabling. These drugs may be used however in patients with severe and disabling chorea. Unfortunately, no medications have been developed as yet that interfere with the nonchoreic aspects of motor dysfunction in HD, although many promising agents are currently in clinical trials. Depression and anxiety can be major problems, and patients should be treated with appropriate antidepressant and antianxiety drugs and monitored for mania and suicidal ideations. Psychosis can be treated with atypical antipsychotics such as clozapine (50–600 mg/d), quetiapine (50–600 mg/d), and risperidone (2–8 mg/d). There is no adequate treatment for the cognitive or motor decline. A neuroprotective therapy that slows or stops disease progression is the major unmet medical need in HD. Drugs that enhance mitochondrial function and increase the clearance of defective mitochondria are being tested as possible disease-modifying therapies. Other investigative approaches include antiglutamate agents, dopamine stabilizers, caspase inhibitors, neurotrophic factors, anti-inflammatory agents, transplantation of fetal striatal cells or stem cells, and DBS of the globus pallidus pars interna (GPi), but none has as yet been demonstrated to have a beneficial effect in HD. The potential to block/edit the mutant huntingtin gene with small interfering RNAs (siRNAs) or CRISPR/cas9 technology is an exciting area of research that is currently being investigated as a possible future therapy.

HUNTINGTON'S DISEASE-LIKE DISORDERS

A group of rare inherited conditions that can mimic HD, designated HD-like (HDL) disorders, have also been identified. HDL-1, 2, and 4 are autosomal dominant conditions that typically present in adulthood. HDL-1 is due to expansion of an octapeptide repeat in *PRNP*, the gene encoding the prion protein (Chap. 430). Thus HDL-1 is properly considered a prion disease. Patients exhibit onset of personality change in the third or

fourth decade, followed by chorea, rigidity, myoclonus, ataxia, and epilepsy. HDL-2 manifests in the third or fourth decade with a variety of movement disorders, including chorea, dystonia, or parkinsonism and dementia. Most patients are of African descent. Acanthocytosis can sometimes be seen in these patients, and this condition must be distinguished from neuroacanthocytosis (below). HDL-2 is caused by an abnormally expanded CTG/CAG trinucleotide repeat expansion in the *junctophilin-3* (*JPH3*) gene. The pathology of HDL-2 consists of intranuclear inclusions immunoreactive for ubiquitin and expanded polyglutamine repeats. HDL-4, the most common condition in this group, is caused by expansion of trinucleotide repeats in *TBP*, the gene that encodes the TATA box-binding protein involved in regulating transcription; this condition is identical to spinocerebellar ataxia (SCA) 17 **(Chap. S10)**, and most patients present primarily with ataxia rather than chorea. Mutations of the *C9Orf72* gene associated with amyotrophic lateral sclerosis **(Chap. 429)** have also been reported in some individuals with an HDL phenotype.

OTHER CHOREAS

Chorea can be seen in a number of additional disorders related to genetic mutations or other disease states.

Among the hereditary forms of childhood-onset chorea, mutations in the *ADCY5* (adenylate cyclase 5) gene are an increasingly recognized and probably relatively common cause of childhood-onset chorea, often in combination with dystonia and developmental delay in some cases. Characteristic perioral movements are a hallmark of the disorder.

Chorea-acanthocytosis (neuroacanthocytosis) is a progressive and typically fatal autosomal recessive disorder that is characterized by chorea coupled with red cell abnormalities on peripheral blood smear (acanthocytes). The chorea can be severe and associated with self-mutilating behavior, dystonia, tics, seizures, and a polyneuropathy. Mutations in the *VPS13A* gene encoding chorein have been described. A phenotypically similar X-linked form of the disorder has been described in older individuals who have reactivity with Kell blood group antigens (McLeod syndrome). A benign hereditary chorea of childhood (BHC1) due to mutations in the gene for thyroid transcription factor 1 and a late-onset benign senile chorea (BHC2) have also been described. It is important to ensure that patients with these types of choreas do not have HD.

Chorea may also occur in association with a variety of infections and degenerative disorders as well as vascular diseases and hypo- and hyperglycemia. Sydenham's chorea (originally called St. Vitus's dance) is more common in females and is typically seen in childhood (5–15 years). It often develops in association with prior exposure to group A streptococcal infection **(Chap. 143)** and is thought to be autoimmune in nature. It is characterized by the acute onset of choreiform movements and behavioral disturbances. With the reduction in the incidence of rheumatic fever, the incidence of Sydenham's chorea has fallen, but it can still be seen in developing countries. The chorea generally responds to dopamine-blocking agents, valproic acid, and carbamazepine, but is self-limited, and treatment is generally restricted to those with severe chorea. Chorea may recur in later life, particularly in association with pregnancy (chorea gravidarum) or treatment with sex hormones. Several reports have documented cases of chorea associated with N-methyl-D-aspartate (NMDA) receptor antibody–positive encephalitis **(Chap. 90)** following herpes simplex virus encephalitis. Systemic

lupus erythematosus (Chap. 349) is the most common systemic disorder that is associated with chorea. The chorea can last for days to years. Chorea can also be seen with hyperthyroidism, autoimmune disorders including Sjögren's syndrome, infectious disorders including HIV disease, metabolic alterations, and polycythemia rubra vera. Chorea has also been described following open-heart surgery in the pediatric population and in association with many medications (especially anticonvulsants, cocaine, CNS stimulants, estrogens, and lithium). Chorea is commonly seen in association with chronic levodopa treatment (Parkinson's Disease, Chap. 427). Chorea may also be encountered in paraneoplastic syndromes associated with anti-CRMP-5 or anti-Hu antibodies (Chap. 90).

HEMIBALLISMUS

Ballism is a violent form of choreiform movement composed of wild, flinging, large-amplitude movements most frequently affecting proximal limb muscles on one side of the body (hemiballism). The movements may only affect one limb (monoballism) or, more exceptionally, both upper or lower limbs (paraballism). The movements may be so severe as to cause exhaustion, dehydration, local injury, and, in extreme cases, death. Fortunately, dopamine-blocking drugs can be very helpful, and importantly, hemiballismus is usually self-limiting and tends to resolve spontaneously after weeks or months. The most common cause is a partial lesion (infarct or hemorrhage) in the subthalamic nucleus (STN), but in 30–40% of cases the lesion is found in the putamen, thalamus, or parietal cortex. Hemiballismus is also a common feature of the paroxysmal dyskinesias (see below). In extreme cases, pallidotomy or DBS of the GPi can be effective and abolish the involuntary movements. Interestingly, surgically induced lesions and DBS of the STN in PD patients are usually not associated with hemiballismus.

TICS

A *tic* is a brief, rapid, recurrent, and seemingly purposeless stereotyped motor contraction. Motor tics can be simple, with movement only affecting an individual muscle group (e.g., blinking, twitching of the nose, jerking of the neck), or complex, with coordinated involvement of multiple muscle groups (e.g., jumping, sniffing, head banging, and echopraxia [mimicking movements]). Phonic (or vocal) tics can also be simple (e.g., grunting) or complex (e.g., echolalia [repeating other people's words], palilalia [repeating one's own words], and coprolalia [expression of obscene words]). Patients may also experience sensory tics, composed of unpleasant focal sensations in the face, head, or neck. These can be mild and of little clinical consequence or severe and disabling to the patient. Tics may present in adulthood and can be seen in association with a variety of disorders, including PD, HD, trauma, dystonia, drugs (e.g., levodopa, neuroleptics), and toxins.

TOURETTE'S SYNDROME (TS)

TS is a neurobehavioral disorder named after the French neurologist Georges Gilles de la Tourette. It predominantly affects males, and the prevalence is estimated to be 0.03–1.6%, but it is likely that many mild cases do not come to medical attention. TS is characterized by multiple motor tics often accompanied by vocalizations (phonic tics). Patients characteristically can voluntarily suppress tics for short periods of time,

but then experience an irresistible urge to express them. Tics vary in intensity and may be absent for days or weeks only to recur, occasionally in a different pattern. Tics tend to present between ages 2 and 15 years (mean 7 years) and often lessen or even disappear in adulthood, particularly in males. Associated behavioral disturbances include anxiety, depression, attention deficit hyperactivity disorder, and obsessive-compulsive disorder. Patients may experience personality disorders, self-destructive behaviors, difficulties in school, and impaired interpersonal relationships.

Etiology and Pathophysiology

TS is thought to be a genetic disorder, but no specific monogenic cause has yet been identified. Current evidence supports a complex inheritance pattern with an important contribution of de-novo, likely gene-disrupting variants. Four likely risk genes with multiple de novo damaging variants in unrelated probands include *WWC1, CELSR3, NIPBL*, and *FN1*. The risk of a family with one affected child having a second is about 25%. The pathophysiology of TS is not known, but alterations in dopamine neurotransmission, opioids, and second-messenger systems have been proposed. Some cases of TS may be the consequence of an autoimmune response to β-hemolytic streptococcal infection (pediatric autoimmune neuropsychiatric disorder associated with streptococcal infection [PANDAS]); however, this entity remains controversial.

TREATMENT

TREATMENT

Tics

Patients with mild disease often only require education and counseling (for themselves and family members). In a high proportion of patients the severity of tics wane in adult life becoming less of a medical problem, thus arguing for a conservative management when possible during the first decades of life. Drug treatment is indicated when the tics are disabling and interfere with quality of life. Therapy is individualized, and there is no singular treatment regimen that has been properly evaluated in double-blind trials. Some physicians use the α-agonist clonidine, starting at low doses and gradually increasing the dose and frequency until satisfactory control is achieved. Guanfacine (0.5–2 mg/d) is an α-agonist that is preferred by some because it only requires once-a-day dosing. Other physicians prefer to use neuroleptics. Atypical neuroleptics are usually used initially (risperidone, olanzapine, ziprasidone) because they are thought to be associated with a reduced risk of tardive dyskinesia. If they are not effective, low doses of classical neuroleptics such as haloperidol, fluphenazine, pimozide, or tiapride can be tried because the risk of tardive dyskinesia in young people is relatively low. Tetrabenazine and deuterated tetrabenazine are also currently being evaluated. Botulinum toxin injections can be effective in controlling focal tics that involve small muscle groups. Behavioral features, and particularly anxiety and compulsions, can be a disabling feature of TS and should be treated. The potential value of DBS targeting the anterior portion of the internal capsule, the GPi, or the thalamus is currently being explored.

MYOCLONUS

Myoclonus is a brief, rapid (<100 ms), shock-like, jerky movement consisting of single or repetitive muscle discharges. Myoclonic jerks can be focal, multifocal, segmental, or generalized and can occur spontaneously, in association with voluntary movement (action myoclonus) or in response to an external stimulus (reflex myoclonus). Negative myoclonus consists of a brief loss of muscle activity (e.g., asterixis in hepatic failure). Myoclonic jerks can be severe and interfere with normal movement or benign and of no clinical consequence as is commonly observed in normal people when waking up or falling asleep (hypnagogic jerks).

Myoclonic jerks differ from tics in that they are not typically repetitive, can severely interfere with normal voluntary movement, and are not suppressible. They can arise in association with abnormal neuronal discharges in cortical, subcortical, brainstem, or spinal cord regions, particularly in association with hypoxemia (especially following cardiac arrest), encephalopathy, and neurodegeneration. Reversible myoclonus can be seen with metabolic disturbances (renal failure, electrolyte imbalance, hypocalcemia), toxins, and many medications. The combination of action myoclonus (cortical origin) with ataxia and generalized epilepsy is associated with several recognized causes. The most common is myoclonic epilepsy or Unverricht-Lundborg disease (EPM-1) which can have a variable but often progressive course. This is an autosomal recessive disease caused by mutations in the *CSBT* gene. Other causes are Lafora body epilepsy or progressive myoclonic epilepsy (PME-2) caused by mutations in in the *EPM2A* gene or the *NHLRC1* gene and ceroid lipofuscinosis. In patients with less severe or absent epilepsy, mitochondrial disorders and neurodegenerative disorders affecting the cerebellum (i.e., SCAs) should be considered. Essential myoclonus is a relatively benign familial condition characterized by multifocal, very brief, lightning-like movements that are frequently alcohol-sensitive. Mutations in the *epsilon-sarcoglycan* gene have been associated with myoclonus seen in association with dystonia (myoclonic-dystonia).

TREATMENT

TREATMENT

Myoclonus

Treatment primarily consists of managing the underlying condition or removing an offending agent. Pharmacologic therapy involves one or a combination of GABAergic agents such as valproic acid (800–3000 mg/d), piracetam (8–20 g/d), clonazepam (2–15 mg/d), levetiracetam (1000–3000 mg/d), or primidone (500–1000 mg/d) and may be associated with striking clinical improvement in chronic cases (e.g., postanoxic myoclonus, progressive myoclonic epilepsy) in which a cortical origin for the myoclonic discharges has been identified. The serotonin precursor 5-hydroxytryptophan (plus carbidopa) may be useful in some cases of postanoxic myoclonus.

DRUG-INDUCED MOVEMENT DISORDERS

This important group of movement disorders is primarily associated with drugs that block dopamine receptors (neuroleptics) or central dopaminergic transmission. These drugs are widely used in psychiatry, but it is important to appreciate that drugs used in the treatment of nausea or vomiting (e.g.,

prochlorperazine [Compazine]) or gastroesophageal disorders (e.g., metoclopramide) are neuroleptic agents and can also cause these disorders. Hyperkinetic movement disorders secondary to neuroleptic drugs can be divided into those that present acutely, subacutely, or after prolonged exposure (tardive syndromes). Dopamine-blocking drugs can also be associated with a reversible parkinsonian syndrome for which anticholinergics are often concomitantly prescribed, but there is concern that this may increase the risk of developing a tardive syndrome and an underlying subclinical PD syndrome should be considered.

ACUTE

Dystonia is the most common acute hyperkinetic drug reaction. It is typically generalized in children and focal in adults (e.g., blepharospasm, torticollis, or OMD). The reaction can develop within minutes of exposure and can be successfully treated in most cases with parenteral administration of anticholinergics (benztropine or diphenhydramine), benzodiazepines (lorazepam, clonazepam, or diazepam), or dopamine agonists. The abrupt onset of severe spasms may occasionally be confused with a seizure; however, there is no loss of consciousness, automatisms, or postictal features typical of epilepsy. The acute onset of chorea, stereotypic behavior, and tics may also be seen, particularly following exposure to CNS stimulants such as methylphenidate, cocaine, or amphetamines.

SUBACUTE

Akathisia is the most common reaction in this category. It consists of motor restlessness with a need to move that is alleviated by movement. Therapy consists of removing the offending agent. When this is not possible, symptoms may be ameliorated with benzodiazepines, anticholinergics, beta blockers, or dopamine agonists.

TARDIVE SYNDROMES

These disorders develop months to years after initiation of the neuroleptic agent. Tardive dyskinesias (TD) are most common, and typically present with choreiform and/or dystonic movements involving the mouth, lips, and tongue. In severe cases, the trunk, limbs, and respiratory muscles may also be affected. In approximately one-third of patients, TD remit within 3 months of stopping the drug, and most patients gradually improve over the course of several years. However, abnormal movements may also develop or worsen after stopping the offending agent. The movements are often mild and more upsetting to the family than to the patient, but they can be severe and disabling, particularly in the context of an underlying psychiatric disorder. Atypical antipsychotics (e.g., clozapine, risperidone, olanzapine, quetiapine, ziprasidone, and aripiprazole) are associated with a lower risk of causing TD in comparison to traditional antipsychotics. Younger patients have a lower risk of developing neuroleptic-induced TD, whereas elderly, females, and those with underlying organic cerebral dysfunction have been reported to be at greater risk. Chronic use is associated with increased risk, and specifically, the U.S. Food and Drug Administration has warned that use of metoclopramide for more than 12 weeks increases the risk of TD. Because TD can be permanent and resistant to treatment, antipsychotics should be used judiciously, atypical neuroleptics should be the preferred agent when possible, and the need for continued use should be regularly monitored.

Treatment primarily consists of stopping the offending agent. If the patient is receiving a traditional antipsychotic, and withdrawal is not possible, replacement with an atypical antipsychotic should be tried. Abrupt cessation of a neuroleptic should be avoided because acute withdrawal can induce worsening. TD can persist after withdrawal of antipsychotics and can be difficult to treat. Valbenazine (Ingrezza™) is an ester of tetrabenazine that has recently been approved for the treatment of tardive dyskinesia based on results of efficacy in double blind trials, but it is associated with sleepiness and QT prolongation. It acts as a vesicular monoamine transporter type 2 (VMAT-2) inhibitor and blocks storage of dopamine. Deuterated tetrabenazine is also being studied for this indication. Benefits in open label studies have been reported with valproic acid (750–3000 mg/d), anticholinergics, or botulinum toxin injections. Other approaches that have been tried include baclofen (40–80 mg/d) or clonazepam (1–8 mg/d). In some cases, the abnormal movement is refractory to therapy.

Chronic neuroleptic exposure can also be associated with tardive dystonia, with preferential involvement of axial muscles and characteristic rocking movements of the trunk and pelvis. Tardive dystonia can be more troublesome than tardive dyskinesia and frequently persists despite stopping medication. Valproic acid, anticholinergics, and botulinum toxin may occasionally be beneficial, but patients are frequently refractory to medical therapy. Tardive akathisia, tardive TS, and tardive tremor syndromes are rare but may also occur after chronic neuroleptic exposure.

Neuroleptic medications can also be associated with a neuroleptic malignant syndrome (NMS). NMS is characterized by the acute or subacute onset of muscle rigidity, elevated temperature, altered mental status, hyperthermia, tachycardia, labile blood pressure, renal failure, and markedly elevated creatine kinase levels. Symptoms typically evolve within days or weeks after initiating the drug. NMS can also be precipitated by the abrupt withdrawal of dopaminergic medications in PD patients. Treatment involves immediate cessation of the offending antipsychotic drug and the introduction of a dopaminergic agent (e.g., a dopamine agonist or levodopa), dantrolene, or a benzodiazepine. In very severe cases, when oral intake is not possible, a patch (delivering rotigotine subcutaneously) or an infusion pump (delivering apomorphine subcutaneously) may be the best approach to provide dopaminergic treatment. Treatment may need to be undertaken in an intensive care setting and include supportive measures such as control of body temperature (antipyretics and cooling blankets), hydration, electrolyte replacement, and control of renal function and blood pressure.

Drugs that have serotonin-like activity (tryptophan, MDMA or “ecstasy,” meperidine) or that block serotonin reuptake can induce a rare, but potentially fatal, serotonin syndrome that is characterized by confusion, hyperthermia, tachycardia, and coma as well as rigidity, ataxia, and tremor. Myoclonus is often a prominent feature, in contrast to NMS, which it resembles in other respects. Patients can be managed with propranolol, diazepam, diphenhydramine, chlorpromazine, or cyproheptadine as well as supportive measures.

A variety of drugs can also be associated with parkinsonism and other hyperkinetic movement disorders. Some examples include phenytoin (chorea, dystonia, tremor, myoclonus), carbamazepine (tics and dystonia), tricyclic antidepressants (dyskinesias, tremor, myoclonus), fluoxetine (myoclonus, chorea, dystonia), oral contraceptives (dyskinesia), β-adrenergics (tremor), buspirone (akathisia, dyskinesias, myoclonus), and digoxin, cimetidine, diazoxide, lithium, methadone, and fentanyl (dyskinesias).

PAROXYSMAL DYSKINESIAS

Paroxysmal dyskinesias are a group of rare disorders characterized by episodic, brief involuntary movements that can manifest as various types of hyperkinetic movements, including chorea, dystonia, tremor, myoclonus, and ballism. There are three main types: (1) *paroxysmal kinesigenic dyskinesia (PKD)*, where the involuntary movements are triggered by sudden movement, (2) *paroxysmal nonkinesigenic dyskinesias (PNKD)*, where the attacks are not induced by movement, and (3) rare cases of *exertion-induced dyskinesia (PED)*, where attacks are induced by prolonged exercise.

PKD are characterized by brief, self-limited attacks induced by movement onset such as running but also occasionally by unexpected sound or photic stimulation. Attacks may affect one side of the body, last seconds to minutes at a time, and recur several times a day. They usually manifest as a mixed hyperkinetic movement disorder with dystonic posturing of a limb, ballismus, and chorea, which may also become generalized. PKD is most commonly familial with an autosomal dominant pattern of inheritance and mutations in the *proline-rich transmembrane protein 2* (*PRRT2*) gene, but may also occur secondary to various brain disorders such as multiple sclerosis or hyperglycemia. PKD is more frequent in males (4:1), and the onset is typically in the first or second decade of life. About 70% report sensory symptoms such as tingling or numbness of the affected limb preceding the attack by a few milliseconds. The evolution is relatively benign, and there is a trend toward resolution of the attacks over time. Treatment with low-dose anticonvulsant therapy such as carbamazepine or phenytoin is advised when the attacks are frequent and interfere with daily life activities, and is effective in about 80% of patients. Some clinical features of PKD (abrupt and short-lasting attacks preceded by an “aura”), the association with true seizure episodes, and its favorable response to anticonvulsant drugs have led to speculation that it is epileptic in origin, but this has not been established.

PNKD involve attacks of generalized dyskinesias precipitated by alcohol, caffeine, stress, or fatigue. In comparison to PKD, the episodes have a relatively longer duration (minutes to hours) and are less frequent (one to three per day). PNKD is inherited as an autosomal dominant condition with high (~80%) but incomplete penetrance. A missense mutation in the *myofibrillogenesis regulator* (*PNKD*) gene has been identified in several families. Recognition of the condition and elimination of the underlying precipitating factors, where possible, are the first priorities. Tetrabenazine, neuroleptics, dopamine-blocking agents, propranolol, clonazepam, and baclofen may be helpful. Treatment may not be required if the condition is mild and self-limited. Most patients with PNKD do not benefit from anticonvulsant drugs, but some may respond to clonazepam or other benzodiazepines.

The *SLC2A1* (solute carrier family 2 member 1) gene, previously linked to GLUT1 (glucose transporter of the blood brain barrier) deficiency syndrome, has been identified to also cause paroxysmal PED. The attacks in this disorder are characterized by a combination of chorea, athetosis, and dystonia in excessively exercised body regions with the legs being most frequently affected. A single attack lasts from a few minutes to an hour and occurs after prolonged physical exercise. In addition to the movement disorder, several patients have

other disease manifestations such as epilepsy, hemolytic anemia, and migraine. A ketogenic diet is an effective therapeutic option.

RESTLESS LEGS SYNDROME (RLS)

RLS is a neurologic disorder that affects ~10% of the adult population (it is rare in Asians) and can cause significant morbidity in some individuals. It was first described in the seventeenth century by the English physician Thomas Willis, but has only recently been recognized as being a bona fide movement disorder. The four core symptoms required for diagnosis are as follows: an urge to move the legs usually caused or accompanied by an unpleasant sensation in the legs; symptoms that begin or worsen with rest; partial or complete relief by movement; and worsening during the evening or night.

Symptoms most commonly begin in the legs, but can spread to or even begin in the upper limbs. The unpleasant sensation is often described as a creepy-crawly feeling, paresthesia, or burning. In about 80% of patients, RLS is associated with periodic leg movements (PLMs) during sleep and occasionally while awake. These involuntary movements are usually brief, lasting no more than a few seconds, and recur every 5–90 s. The restlessness and PLMs are a major cause of sleep disturbance in patients, leading to poor-quality sleep and daytime sleepiness.

Primary RLS has a strong genetic component, and several loci have been associated with an autosomal dominant pattern of inheritance, although penetrance may be variable and no specifically causative gene has been identified to date. The mean age of onset in familial forms is 27 years, although pediatric cases are recognized. The severity of symptoms is variable. Secondary RLS may be associated with pregnancy or a range of underlying disorders, including anemia, ferritin deficiency, renal failure, and peripheral neuropathy. The pathogenesis probably involves disordered dopamine function, which may be peripheral or central, possibly in association with an abnormality of iron metabolism. Diagnosis is made on clinical grounds but can be supported by polysomnography and the demonstration of PLMs. The neurologic examination is normal. Secondary causes of RLS should be excluded, and ferritin levels, glucose, and renal function should be measured.

Most RLS sufferers have mild symptoms that do not require specific treatment. General measures to improve sleep hygiene and quality should be attempted first. If symptoms remain intrusive, low doses of dopamine agonists, e.g., pramipexole (0.25–0.5 mg), ropinirole (1–2 mg), or patch rotigotine (2–3 mg), taken 1–2 h before bedtime are generally effective. Levodopa may also be effective but is more likely to be associated with augmentation (spread and worsening of restlessness and its appearance earlier in the day) or rebound (reappearance sometimes with worsening of symptoms at a time related to the drug's short half-life). Augmentation can also be seen with dopamine agonists, particularly if higher doses are employed. Other drugs that can be effective include anticonvulsants, analgesics, and opiates. Management of secondary RLS should be directed to correcting the underlying disorder; for example, iron replacement for anemia.

OTHER DISORDERS THAT MAY PRESENT WITH A COMBINATION OF PARKINSONISM AND HYPERKINETIC MOVEMENTS

WILSON'S DISEASE

WD is an autosomal recessive inherited disorder of copper metabolism that manifests with neurologic, psychiatric, and liver disorders, alone or in combination. It is caused by mutations in the *ATP7B* gene encoding a P-type ATPase. The disease was first described by the English neurologist Kinnier Wilson at the beginning of the twentieth century, although at around the same time the German physicians Kayser and Fleischer separately noted the characteristic association of corneal pigmentation with hepatic and neurologic features. WD has a worldwide prevalence of ~1 in 30,000, with a mutation carrier frequency of 1 in 90. About half of WD patients (especially younger patients) manifest with liver abnormalities. The remainder present with neurologic disease (with or without underlying liver abnormalities), and a small proportion have hematologic or psychiatric problems at disease onset.

Neurologic onset usually manifests in the second decade with tremor, rigidity, and dystonia. The tremor is usually in the upper limbs, bilateral, and asymmetric. Tremor can be on intention or occasionally at rest and, in advanced disease, can take on a wing-beating characteristic (a flapping movement when the arms are held outstretched with the fingers opposed). Other features can include parkinsonism with bradykinesia, dystonia (particularly facial grimacing), dysarthria, and dysphagia. More than half of those with neurologic features have a history of psychiatric disturbances, including depression, mood swings, and overt psychosis. Kayser-Fleischer (KF) rings are seen virtually in all patients with neurologic features and 80% of those with hepatic presentations. KF rings represent the deposition of copper in Descemet's membrane around the cornea. They consist of a characteristic grayish rim or circle at the limbus of the cornea and are best detected by slit-lamp examination. Neuropathologic examination is characterized by neurodegeneration and astrogliosis in the basal ganglia, particularly in the striatum.

WD should always be considered in the differential diagnosis of a movement disorder in the first decades of life. Low levels of blood copper and ceruloplasmin and high levels of urinary copper may be present, but normal levels do not exclude the diagnosis. Brain imaging usually reveals generalized brain atrophy in established cases, and ~50% have signal hypointensity in the caudate head, putamen, globus pallidus, substantia nigra, and red nucleus on T2-weighted MRI scans. However, correlation of imaging changes with clinical features is not good. Liver biopsy with demonstration of high copper levels and genetic testing remain the gold standard for the diagnosis.

In the absence of treatment, the course is progressive and leads to severe neurologic dysfunction and early death in the majority of patients, although a small proportion experience a relatively benign course. Treatment is directed at reducing tissue copper levels and maintenance therapy to prevent reaccumulation. There is no clear consensus on optimal treatment, and patients should be managed in a unit with expertise in WD. Penicillamine is frequently used to increase copper excretion, but may lead to a worsening of symptoms in the initial stages of therapy. Side effects are common and can to some degree be attenuated by

coadministration of pyridoxine. Tetrathiomolybdate blocks the absorption of copper and can be used instead of penicillamine. Trientine and zinc are useful drugs for maintenance therapy. Effective treatment can reverse the neurologic features in most patients, particularly when started early. However, some patients may still progress, especially those with hepatocerebral disease. KF rings tend to decrease after 3–6 months and disappear by 2 years. Adherence to maintenance therapy is a major challenge in long-term care. Patients with advanced hepatic disease may require a liver transplant, and research is looking into the potential role of organ-specific chelators.

NEURODEGENERATION WITH BRAIN IRON ACCUMULATION (NBIA)

NBIA represents a group of inherited disorders characterized by iron accumulation in the basal ganglia. Clinically, they can manifest as a progressive neurologic disorder with a variety of clinical features including parkinsonism, dystonia, neuropsychiatric abnormalities, and retinal degeneration. Cognitive disorders and cerebellar dysfunction may also be seen. Presentation is usually in childhood, but adult cases have been described. Multiple genes have been identified to date. Pantothenate kinase–associated neurodegeneration (PKAN) formerly known as Hallervorden-Spatz disease is caused by a mutation in the *PANK2* gene, and is the most common form of NBIA accounting for about 50% of cases. Onset is usually in early childhood and is manifest as a combination of dystonia, parkinsonism, and spasticity. MRI shows a characteristic low signal abnormality in the center of the globus pallidus on T2-weighted scans caused by iron accumulation known as the “eye of the tiger” sign. Numerous other gene mutations have been described associated with iron accumulation including mutations in *PLA2G6, C19orf12, FA2H, ATP13A2, WDR45, FTL, CP*, and *DCAF17*. One must be cautious, however, not to assume that all cases with iron accumulation in the basal ganglia represent an NBIA, because iron accumulation in specific basal ganglia regions is normal, and excess iron accumulation may occur in the basal ganglia region as a consequence of neurodegeneration associated with multiple causes unrelated to a defect in iron metabolism.

PSYCHOGENIC (FUNCTIONAL) DISORDERS

Virtually all movement disorders including tremor, tics, dystonia, myoclonus, chorea, ballism, and parkinsonism can be psychogenic in origin. Tremor affecting the upper limbs is the most common psychogenic movement disorder. Psychogenic movements can result from a somatoform or conversion disorder, malingering (e.g., seeking financial gain), or a factitious disorder (e.g., seeking psychological gain). Psychogenic movement disorders are relatively common (estimated to be 2–3% of patients seen in a movement disorder clinic), more frequent in women, disabling for the patient and family, and expensive for society. Clinical features suggesting a psychogenic movement disorder include an acute onset with a pattern of abnormal movement that is inconsistent with a known movement disorder. Diagnosis is based on the nonorganic quality of the movement, the absence of findings of an organic disease process, and positive features that specifically point to a psychogenic illness such as variability and distractibility. For example, the magnitude of a psychogenic tremor is increased with attention and diminishes or even disappears when the patient is distracted by being asked to perform a different task or is unaware that he or she is being observed. This is the opposite of an organic tremor where the magnitude is increased with distraction and tends to be

reduced when observed. Other positive features suggesting a psychogenic problem include a tremor frequency that is variable or that entrains with the frequency of a designated movement in the contralateral limb, or a response to placebo interventions. Associated features can include nonanatomic sensory findings, give-way weakness, astasia-abasia (an odd, gyrating gait or posture); **(Chap. 23)**, and multiple somatic complaints with no underlying pathology (somatoform disorder). Comorbid psychiatric problems such as anxiety, depression, and emotional trauma may be present but are not necessary for the diagnosis of a psychogenic movement disorder to be made. Psychogenic movement disorders can occur as an isolated entity or in association with an underlying organic problem. The diagnosis can often be made based on clinical features alone, and unnecessary tests or medications can be avoided. Underlying psychiatric problems may be present and should be identified and treated, but many patients with psychogenic movement disorders have no obvious psychiatric pathology. Psychotherapy and hypnosis may be of value for patients with conversion reaction, and cognitive behavioral therapy may be helpful for patients with somatoform disorders. Patients with hypochondriasis, factitious disorders, and malingering have a poor prognosis.

FURTHER READING

Albanese A et al: Therapeutic advances in dystonia. Mov Disord 30:1547, 2015. [PubMed: 26301801]

Albanese A et al: Phenomenology and classification of dystonia: A consensus update. Mov Disord 28:863, 2013. [PubMed: 23649720]

Balint B, Bhatia KP: Dystonia: An update on phenomenology, classification, pathogenesis and treatment. Curr Opin Neurol 27:468, 2014. [PubMed: 24978640]

Elble RJ: The essential tremor syndromes. Curr Opin Neurol 29:507, 2016. [PubMed: 27257943]

Espay AJ et al: Essential pitfalls in “essential” tremor. Mov Disord 32:325, 2017. [PubMed: 28116753]

Kieburtz K et al: Huntington’s disease: Current and future therapeutic prospects. Mov Disord, 2018.

Krack P et al: Current applications and limitations of surgical treatments for movement disorders. Mov Disord 32:36, 2017. [PubMed: 28124435]

Marras C et al: Nomenclature of genetic movement disorders: Recommendations of the International Parkinson and Movement Disorder Society task force. Mov Disord 32:724, 2017. [PubMed: 28513081]

Schapira AH et al: Slowing of neurodegeneration in Parkinson’s disease and Huntington’s disease: Future therapeutic perspectives. Lancet 384:545, 2014. [PubMed: 24954676]

Chapter 429: Amyotrophic Lateral Sclerosis and Other Motor Neuron Diseases

Robert H. Brown, Jr.

FIGURE 429-1

AMYOTROPHIC LATERAL SCLEROSIS (ALS)

ALS is the most common progressive motor neuron disease. It is a prime example of a neurodegenerative disease and is arguably the most devastating of the neurodegenerative disorders.

PATHOLOGY

The pathologic hallmark of motor neuron degenerative disorders is death of lower motor neurons (consisting of anterior horn cells in the spinal cord and their brainstem homologues innervating bulbar muscles) and upper, or corticospinal, motor neurons (originating in layer five of the motor cortex and descending via the pyramidal tract to synapse with lower motor neurons, either directly or indirectly via interneurons) **(Chap. 21)**. Although at its onset ALS may involve selective loss of function of only upper or lower motor neurons, it ultimately causes progressive loss of both categories of motor neurons. Indeed, in the absence of clear involvement of both motor neuron types, the diagnosis of ALS is questionable. In a subset of cases, ALS arises concurrently with frontotemporal dementia **(Chap. 424)**; in these instances, there is degeneration of frontotemporal cortical neurons and corresponding cortical atrophy.

Other motor neuron diseases involve only particular subsets of motor neurons **(Tables 429-1 and 429-2)**. Thus, in bulbar palsy and spinal muscular atrophy (SMA; also called *progressive muscular atrophy*), the lower motor neurons of brainstem and spinal cord, respectively, are most severely involved. By contrast, pseudobulbar palsy, primary lateral sclerosis (PLS), and hereditary spastic paraplegia (HSP) affect only upper motor neurons innervating the brainstem and spinal cord.

TABLE 429-1

Etiology of Motor Neuron Disorders

Diagnostic Category	Investigation
Structural lesions Parasagittal or foramen magnum tumors Cervical spondylosis Chiari malformation of syrinx Spinal cord arteriovenous malformation	MRI scan of head (including foramen magnum and cervical spine)
Infections Bacterial—tetanus, Lyme Viral—poliomyelitis, herpes zoster Retroviral—myelopathy	CSF exam, culture Lyme titer Anti-viral antibody HTLV-1 titers
Intoxications, physical agents Toxins—lead, aluminum, others Drugs—strychnine, phenytoin Electric short, x-irradiation	24-h urine for heavy metals Serum lead level
Immunologic mechanisms Plasma cell dyscrasias Autoimmune polyradiculopathy Motor neuropathy with conduction block Paraneoplastic Paracarcinomatous	Complete blood count[a] Sedimentation rate[a] Total protein[a] Anti-GM1 antibodies[a] Anti-Hu antibody MRI scan, bone marrow biopsy
Metabolic Hypoglycemia Hyperparathyroidism Hyperthyroidism Deficiency of folate, vitamin B_{12}, vitamin E Malabsorption Deficiency of copper, zinc Mitochondrial dysfunction	Fasting blood sugar[a] Routine chemistries including calcium[a] PTH Thyroid function[a] Vitamin B_{12}, vitamin E, folate[a] Serum zinc, copper[a] 24-h stool fat, carotene, prothrombin time Fasting lactate, pyruvate, ammonia Consider mtDNA
Hyperlipidemia	Lipid electrophoresis
Hyperglycinuria	Urine and serum amino acids CSF amino acids
Hereditary disorders C9orf72 Superoxide dismutase TDP43 FUS/TLS Androgen receptor defect (Kennedy's disease)	WBC DNA for mutational analysis

[a]Should be obtained in all cases.

Abbreviations: CSF, cerebrospinal fluid; FUS/TLS, fused in sarcoma/translocated in liposarcoma; HTLV-1, human T-cell lymphotropic virus; MRI, magnetic resonance imaging; PTH, parathyroid; WBC, white blood cell.

TABLE 429-2

Sporadic Motor Neuron Diseases

Chronic	**Entity**
Upper and lower motor neuron	Amyotrophic lateral sclerosis
Predominantly upper motor neuron	Primary lateral sclerosis
Predominantly lower motor neuron	Multifocal motor neuropathy with conduction block
	Motor neuropathy with paraproteinemia or cancer
	Motor predominant peripheral neuropathies
Other	
Associated with other neurodegenerative disorders	
Secondary motor neuron disorders (see Table 429-1)	
Acute	
Poliomyelitis	
Herpes zoster	
Coxsackie virus	
West Nile virus	

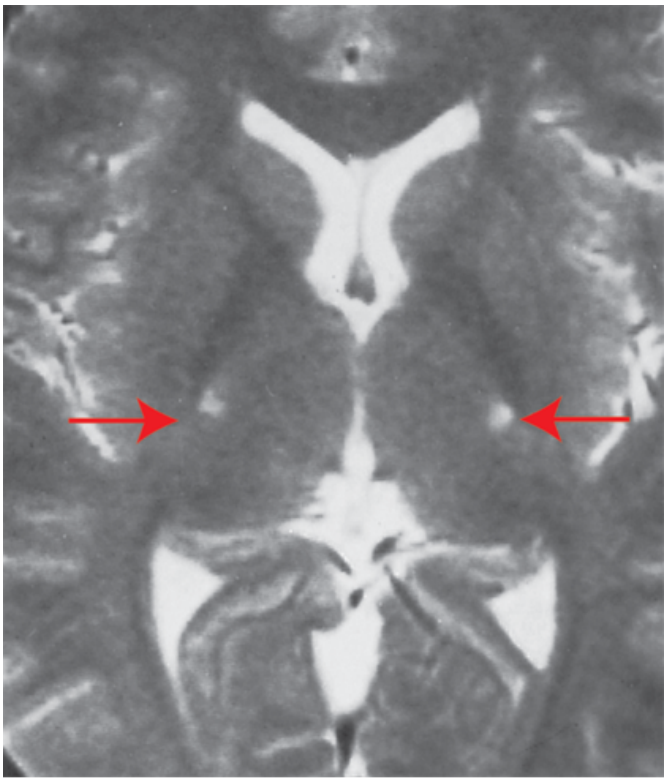
In each of these diseases, the affected motor neurons undergo shrinkage, often with accumulation of the pigmented lipid (lipofuscin) that normally develops in these cells with advancing age. In ALS, the motor neuron cytoskeleton is typically affected early in the illness. Focal enlargements are frequent in proximal motor axons; ultrastructurally, these “spheroids” are composed of accumulations of neurofilaments and other proteins. Commonly in both sporadic and familial ALS, the affected neurons demonstrate ubiquitin-positive aggregates, typically associated with the protein TDP43 (see below). Also seen is proliferation of astroglia and microglia, the inevitable accompaniment of all degenerative processes in the central nervous system (CNS).

The death of the peripheral motor neurons in the brainstem and spinal cord leads to denervation and atrophy of the corresponding muscle fibers. Histochemical and electrophysiologic evidence indicates that in the early phases of the illness denervated muscle can be reinnervated by sprouting of nearby distal motor nerve terminals, although reinnervation in this disease is considerably less extensive than in most other disorders affecting motor neurons (e.g., poliomyelitis, peripheral neuropathy). As denervation progresses, muscle atrophy is readily recognized in muscle biopsies and on clinical examination. This is the basis for the term *amyotrophy*. The loss of cortical motor neurons results in thinning of the corticospinal tracts that travel via the internal capsule **(Fig. 429-1)** and pyramidal tracts in the brainstem to the lateral and anterior white matter columns of the spinal cord. The loss of fibers in the lateral columns and resulting fibrillary gliosis impart a particular firmness (*lateral sclerosis*). A remarkable feature of the disease is the selectivity of neuronal cell death. By light microscopy, the entire sensory apparatus, the regulatory mechanisms for the control and coordination of movement, remains intact.

Except in cases of frontotemporal dementia, the components of the brain required for cognitive processing are also preserved. However, immunostaining indicates that neurons bearing ubiquitin, a marker for degeneration, are also detected in nonmotor systems. Moreover, studies of glucose metabolism in the illness also indicate that there is neuronal dysfunction outside of the motor system. Pathological studies reveal proliferation of microglial cells and astrocytes in affected regions; in some cases, this phenomenon, designated neuroinflamation, can be visualized using positron emission tomography (PET) scanning for ligands that are recognized by activated microglia. Within the motor system, there is some selectivity of involvement. Thus, motor neurons required for ocular motility remain unaffected, as do the parasympathetic neurons in the sacral spinal cord (the nucleus of Onufrowicz, or Onuf) that innervate the sphincters of the bowel and bladder.

FIGURE 429-1

Amyotrophic lateral sclerosis. Axial T2-weighted magnetic resonance imaging (MRI) scan through the lateral ventricles of the brain reveals abnormal high signal intensity within the corticospinal tracts (*arrows*). This MRI feature represents an increase in water content in myelin tracts undergoing Wallerian degeneration secondary to cortical motor neuronal loss. This finding is commonly present in ALS, but can also be seen in AIDS-related encephalopathy, infarction, or other disease processes that produce corticospinal neuronal loss in a symmetric fashion.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CLINICAL MANIFESTATIONS

The manifestations of ALS are somewhat variable depending on whether corticospinal neurons or lower motor neurons in the brainstem and spinal cord are more prominently involved. With lower motor neuron dysfunction and early denervation, typically the first evidence of the disease is insidiously developing asymmetric weakness, usually first evident distally in one of the limbs. A detailed history often discloses recent development of cramping with volitional movements, typically in the early hours of the morning (e.g., while stretching in bed). Weakness caused by denervation is associated with progressive wasting and atrophy of muscles and, particularly early in the illness, spontaneous twitching of motor units, or fasciculations. In the hands, a preponderance of extensor over flexor weakness is common. When the initial denervation involves bulbar rather than limb muscles, the problem at onset is difficulty with chewing, swallowing, and movements of the face and tongue. Rarely, early involvement of the muscles of respiration may lead to death before the disease is far advanced elsewhere. With prominent corticospinal involvement, there is hyperactivity of the muscle-stretch reflexes (tendon jerks) and, often, spastic resistance to passive movements of the affected limbs. Patients with significant reflex hyperactivity complain of muscle

stiffness often out of proportion to weakness. Degeneration of the corticobulbar projections innervating the brainstem results in dysarthria and exaggeration of the motor expressions of emotion. The latter leads to involuntary excess in weeping or laughing (pseudobulbar affect).

Virtually any muscle group may be the first to show signs of disease, but, as time passes, more and more muscles become involved until ultimately the disorder takes on a symmetric distribution in all regions. It is characteristic of ALS that, regardless of whether the initial disease involves upper or lower motor neurons, both will eventually be implicated. Even in the late stages of the illness, sensory, bowel and bladder, and cognitive functions are preserved. Even when there is severe brainstem disease, ocular motility is spared until the very late stages of the illness. As noted, in some cases (particularly those that are familial), ALS develops concurrently with frontotemporal dementia, characterized by early behavioral abnormalities with prominent behavioral features indicative of frontal lobe dysfunction.

A committee of the World Federation of Neurology has established diagnostic guidelines for ALS. Essential for the diagnosis is simultaneous upper and lower motor neuron involvement with progressive weakness and the exclusion of all alternative diagnoses. The disorder is ranked as “definite” ALS when three or four of the following are involved: bulbar, cervical, thoracic, and lumbosacral motor neurons. When two sites are involved, the diagnosis is “probable,” and when only one site is implicated, the diagnosis is “possible.” An exception is made for those who have progressive upper and lower motor neuron signs at only one site and a mutation in the gene encoding superoxide dismutase (SOD1; see below).

EPIDEMIOLOGY

The illness is relentlessly progressive, leading to death from respiratory paralysis; the median survival is from 3 to 5 years. There are very rare reports of stabilization or even regression of ALS. In most societies, there is an incidence of 1–3 per 100,000 and a prevalence of 3–5 per 100,000. It is striking that at least 1 in 1000 deaths in North America and Western Europe (and probably elsewhere) are due to ALS; this finding predicts that more than 300,000 individuals now alive in the United States will die of ALS. Several endemic foci of higher prevalence exist in the western Pacific (e.g., in specific regions of Guam or Papua New Guinea). In the United States and Europe, males are somewhat more frequently affected than females. Epidemiologic studies have incriminated risk factors for this disease including exposure to pesticides and insecticides, smoking, and possibly service in the military. Although ALS is overwhelmingly a sporadic disorder, some 10% of cases are inherited as an autosomal dominant trait.

FAMILIAL ALS

Several forms of selective motor neuron disease are inheritable **(Table 429-3)**. Familial ALS (FALS) involves both corticospinal and lower motor neurons. Apart from its inheritance as an autosomal dominant trait, it is clinically indistinguishable from sporadic ALS. Genetic studies have identified mutations in multiple genes, including those encoding the protein C9orf72 (open reading frame 72 on chromosome 9), cytosolic enzyme SOD1 (superoxide dismutase), the RNA binding proteins TDP43 (encoded by the TAR DNA binding protein gene), and fused in sarcoma/translocated in liposarcoma (FUS/TLS), as the most common causes of FALS. Mutations in C9orf72 account for ~45–50% of FALS and perhaps 5% of sporadic ALS cases. Mutations in SOD1 explain another 20% of cases of FALS, whereas TDP43 and FUS/TLS each represent about 5% of familial cases. Mutations in several other genes (such as optineurin, TBK1 and profilin-1) each cause about ~1% of cases.

TABLE 429-3

Genetic Motor Neuron Diseases

Disease	Gene Symbol	Gene Name	Inheritance	Frequency (in the United States)	Usual Onset	Protein Function	Unusual Features
I. Upper and Lower Motor Neurons (Familial ALS)							
ALS1	SOD1	Cu/Zn superoxide dismutase 1	AD	20% FALS	Adult	Protein antioxidant	
ALS2	ALS2	Alsin	AR	<1% FALS	Juvenile	GEF signaling	Severe corticobulbar, corticospinal featuresk may mimic PLS
ALS4	SETX	Senataxin	AD	~1% FALS	Late juvenile	DNA helicase	Late childhood onset
ALS6	FUS/TLS	Fused in Sarcoma/Translocated in liposarcoma	AD	5% FALS	Adult	DNA, RNA binding	
ALS8 /SMA	VAPB	Vesicle associated protein B	AD	<1%	Adult	Vesicular trafficking	
ALS9	ANG	Angiogenin	AD	<1%	Adult	RNAse, angiogenesis	
ALS10	TARDBP	TAR DNA binding protein	AD	5% FALS	Adult	DNA, RNA binding	
ALS12	OPTN	Optineurin	AD/AR	~1% FALS	Adult	Attenuates NF-κB	
ALS13	ATXN2	Ataxin 2	AD	<1%	Adult	Cytotoxic expanded CAG repeat	
ALS14	VCP	Valosin-containing protein	AD	~ 1% FALS	Adult	ATPase	
ALS18	PFN1	Profilin 1	AD	~1% FALS	Adult	Involved in actin polymerization	

Disease	Gene Symbol	Gene Name	Inheritance	Frequency (in the United States)	Usual Onset	Protein Function	Unusual Features
ALS19	ERB4	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 4	AD		Adult	Signaling molecule	
ALS20	HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1	AD	<1%	Adult	Heteronuclear RNA binding protein	
ALS	DCTN1	Dynactin	AD	<1%	Adult	Axonal transport	May cause vocal cord paralysis or PLSe
ALS-FTD	TBK1	TankBinding Protein 1	AD		Adult	NF-κB signalling	also mimics PLS
ALS-FTD	UBQLN2	Ubiquilin 2	X-LD	<1%	Adult or Juvenile	Protein degradation	
ALS-FTD	CHMP2B	Chromatin modifying protein 2B	AD	<1% FALS	Adult	Chromatin binding protein	
ALS-FTD	C9ORF72	Chromosome 9 Open Reading Frame 72	AD	40-50% FALS	Adult	Regulates vessicle trafficking	May also be associated with Parkinosonism, PLS
ALS-FTD	MAPT	Microtubule Associated Protein Tau	AD		Adult	cytoskeletal protein	Usually causes only FTD
ALS	COX	Cytochrome c oxidase	Maternally inherited		Adult	Mitochondrial: ATP generation	
ALS	tRNA-isoleucine		Maternally inherited		Adult	Mitochondrial: ATP generation	
II. Lower Motor Neurons							
Spinal muscular atrophies	SMN	Survival motor neuron	AR	1/10,000 live births	Infancy	RNA metabolism	
GM2-gangliosidosis							
1. Sandhoff's disease	HEXB	Hexosaminidase B	AR		Childhood	Ganglioside recycling	

Disease	Gene Symbol	Gene Name	Inheritance	Frequency (in the United States)	Usual Onset	Protein Function	Unusual Features
2. AB variant	GM2A	GM2-activator protein	AR		Childhood	Ganglioside recycling	
3. Adult Tay-Sachs disease	HEXA	Hexosaminidase A	AR		Childhood	Ganglioside recycling	
X-linked spinobulbar muscular atrophy	AR	Androgen receptor	XR		Adult	Nuclear signaling	
III. Upper Motor Neuron (Selected HSPs)							
SPG3A	ATL1	Atlastin	AD	10% AD FSP	Childhood	GTPase—vesicle recycling	
SPG4	SPAST	Spastin	AD	50-60% AD FSP	Early adulthood	ATPase family—microtubule associate	Some sensory loss
SPG6	NIPA1	Non imprinted in Prader-Willi/Angelman syndrome 1	AD		Early adulthood	Membrane transporter or receptor	Deleted in Prader-Willi, Angelman's
SPG8	WASHC5	Strumpellin	AD		Early adulthood	Ubiquitous, spectrin-like	
SPG10	KIF5A	Kinesin heavy chain isoform 5A	AD	10% AD FSP	Second–third decade	Motor-associated protein	± Peripheral neuropathy, retardation
SPG12	RTN2	Reticulon 2	AD		Childhood	ER protein, interacts with spastin	
SPG13	HSP60	Heat shock protein 60	AD		Early adulthood	Chaperone protein	
SPG17/ variants of charcot Marie Tooth type 2/ Silver syndrome	BSCL2	Seipin lipid droplet biogenesis associated	AD		Variable	Membrane protein in ER	Amyotrophy hands, feet
SPG31	REEP1	Receptor Expression Enhancing Protein 1	AD	10% AD FSP	Early	Mitochondrial protein	Rarely, amyotrophy
SPG33	ZFYVE27	Zinc Finger FYVE-Type Containing 27	AD		Adult	Interacts with spastin	Pes equinus

Disease	Gene Symbol	Gene Name	Inheritance	Frequency (in the United States)	Usual Onset	Protein Function	Unusual Features
SPG42	SLC33A1	Acetyl-CoA-transporter	AD		Variable	Solute carrier	
SPG72	REEP2	Receptor Expression Enhancing Protein 2	AD		Childhood	ER protein	
SPG5	CYP7B1	Cytochrome P450	AR	5-10% AR FSP	Variable	Degrades endogenous substances	Sensory loss
SPG7	SPG7	Paraplegin	AR	5-10% AR FSP	Variable	Mitochondrial protein	Rarely, optic atrophy, ataxia, rarely PLS
SPG11	SPG11	Spatacsin	AR	20-70% AR FSP depends on ethnicity	Predominantly childhood	Cytosolic, ? membrane-associated	Some sensory loss, thin corpus callosum; may mimic ALS (ALS5)
SPG15	ZFYVE26	Spastizin	AR		Childhood	Zinc finger protein	Some amyotrophy, some CNS features
SPG20	SPG20	Spartin	AR		Childhood	Endosomal trafficking protein	
SPG21	SPG21	Maspardin	AR		Childhood	Endosomal trafficking protein	
SPG35	Fatty acid 2 hydrolase		AR		Childhood	Membrane protein	Multiple CNS features
SPG39	PNPLA6	patatin-like phospholipase domain-containing protein 6 / Neuropathy target esterase	AR		Early childhood	Esterase	May have PLS=like phenotype
SPG44	GJC2	Gap junction protein gamma 2/ Connexin 47	AR		Childhood	Gap junction protein	Possible mild CNS features

Disease	Gene Symbol	Gene Name	Inheritance	Frequency (in the United States)	Usual Onset	Protein Function	Unusual Features
SPG46	GBA2	β-Glucosidase 2	AR		Childhood	Glycoside hydrolase	Thin corpus callosum, mental retardation
SPG2	PLP	Proteolipid protein	XR		Early childhood	Myelin protein	Sometimes multiple CNS features
SPG1	L1-CAM	Neural cell adhesion molecule L1 precursor	XR		Infancy	Cell adhesion molecule	
SPG22	SLC16A2	Solute Carrier Family 16 Member 2	XR		Infancy	Monocarboxylic acid transporter	
Adrenoleukodystrophy	ALDP	Adrenoleukodystrophy protein	XR		Early adulthood	ATP binding transporter protein	Possible adrenal insufficiency, CNS inflammation

Abbreviations: AD, autosomal dominant; ALS, amyotrophic lateral sclerosis; AR, autosomal recessive; CNS, central nervous system; BSCL2, Bernadelli-Seip congenital lipodystrophy 2B; FUS/TLS, fused in sarcoma/translocated in liposarcoma; GEF, Guanidine nucleotide exchange factor; HSP, hereditary spastic paraplegia; TDP43, Tar DNA binding protein 43 kd; XR, X-linked recessive.

Rare mutations in other genes are also clearly implicated in ALS-like diseases. Thus, a familial, dominantly inherited motor disorder that in some individuals closely mimics the ALS phenotype arises from mutations in a gene that encodes a vesicle-binding protein. Mutations in senataxin, a helicase, cause an early adult-onset, slowly evolving ALS variant. Kennedy's syndrome is an X-linked, adult-onset disorder that may mimic ALS, as described below. Tau gene mutations usually underlie frontotemporal dementia, but in some instances may be associated with prominent motor neuron findings.

Genetic analyses are also beginning to illuminate the pathogenesis of some childhood-onset motor neuron diseases. For example, a slowly disabling degenerative, predominantly upper motor neuron disease that starts in the first decade is caused by mutations in a gene that expresses a novel signaling molecule with properties of a guanine-exchange factor, termed *alsin*.

DIFFERENTIAL DIAGNOSIS

Because ALS is currently untreatable, it is imperative that potentially remediable causes of motor neuron dysfunction be excluded (Table 429-1). This is particularly true in cases that are atypical by virtue of (1) restriction to either upper or lower motor neurons, (2) involvement of neurons other than motor neurons, and (3) evidence of motor neuronal conduction block on electrophysiologic testing. Compression of the cervical spinal cord or cervicomedullary junction from tumors in the cervical regions or at the foramen magnum or from cervical spondylosis with osteophytes projecting into the vertebral canal can produce weakness, wasting, and fasciculations in the upper limbs and spasticity in the legs, closely resembling ALS. The absence of cranial nerve involvement may be helpful in differentiation, although some foramen magnum lesions may compress the twelfth cranial (hypoglossal) nerve, with resulting paralysis of the tongue. Absence of pain or of sensory changes, normal bowel and bladder function, normal radiologic studies of the spine, and normal cerebrospinal fluid (CSF) all

favor ALS. Where doubt exists, magnetic resonance imaging (MRI) scans and possibly contrast myelography should be performed to visualize the cervical spinal cord.

Another important entity in the differential diagnosis of ALS is *multifocal motor neuropathy with conduction block* (MMCB), discussed below. A diffuse, lower motor axonal neuropathy mimicking ALS sometimes evolves in association with hematopoietic disorders such as lymphoma or multiple myeloma. In this clinical setting, the presence of an M-component in serum should prompt consideration of a bone marrow biopsy. Lyme disease **(Chap. 181)** may also cause an axonal, lower motor neuropathy, although typically with intense proximal limb pain and a CSF pleocytosis.

Other treatable disorders that occasionally mimic ALS are chronic lead poisoning and thyrotoxicosis. These disorders may be suggested by the patient's social or occupational history or by unusual clinical features. When the family history is positive, disorders involving the genes encoding C9orf72, cytosolic SOD1, TDP43, FUS/TLS, and adult hexosaminidase A or α-glucosidase deficiency must be excluded **(Chap. 411)**. These are readily identified by appropriate laboratory tests. Benign fasciculations are occasionally a source of concern because on inspection they resemble the fascicular twitchings that accompany motor neuron degeneration. The absence of weakness, atrophy, or denervation phenomena on electrophysiologic examination usually excludes ALS or other serious neurologic disease. Patients who have recovered from poliomyelitis may experience a delayed deterioration of motor neurons that presents clinically with progressive weakness, atrophy, and fasciculations. Its cause is unknown, but it is thought to reflect sublethal prior injury to motor neurons by poliovirus **(Chap. 199)**.

Rarely, ALS develops concurrently with features indicative of more widespread neurodegeneration. Thus, one infrequently encounters otherwise typical ALS patients with a parkinsonian movement disorder or frontotemporal dementia, particularly in instances of C9orf72 mutations, which strongly suggests that the simultaneous occurrence of two disorders is a direct consequence of the gene mutation. As another example, prominent amyotrophy has been described as a dominantly inherited disorder in individuals with bizarre behavior and a movement disorder suggestive of parkinsonism; many such cases have now been ascribed to mutations that alter the expression of tau protein in brain **(Chap. 424)**. In other cases, ALS develops simultaneously with a striking frontotemporal dementia. An ALS-like disorder has also been described in some individuals with chronic traumatic encephalopathy, associated with deposition of TDP43 and neurofibrillary tangles in motor neurons.

PATHOGENESIS

The cause of sporadic ALS is not well defined. Several mechanisms that impair motor neuron viability have been elucidated in rodents induced to develop motor neuron disease by SOD1 or profilin-1 transgenes with ALS-associated mutations. One may loosely group the genetic causes of ALS into three categories. In one group, the primary problem is inherent instability of the mutant proteins, with subsequent perturbations in protein degradation (SOD1, ubiquilin-1 and 2, p62). In the second category, the causative mutant genes perturb RNA processing, transport, and metabolism (C9orf73, TDP43, FUS). In the case of C9orf72, the molecular pathology is an expansion of an intronic hexanucleotide repeat (-GGGGCC-) beyond an upper normal of 30 repeats to hundreds or even thousands of repeats. As observed in other intronic repeat disorders such as myotonic dystrophy **(Chap. 441)** and spinocerebellar atrophy type 8 **(Chap. 431)**, data suggest that the expanded intronic repeats generate expanded RNA repeats that form intranuclear foci and may confer toxicity by sequestering transcription factors or by undergoing noncanonical protein translation across all possible reading frames of the expanded RNA tracts. Importantly, the latter process generates lengthy dipeptides that are detected in the spinal fluid and are a unique biomarker for C9orf72 ALS. TDP43 and FUS are multifunctional proteins that bind RNA and DNA and shuttle between the nucleus and the cytoplasm, playing multiple roles in the control of cell proliferation, DNA repair and transcription, and gene translation, both in the cytoplasm and locally in dendritic spines in response to electrical activity. How mutations in FUS/TLS provoke motor neuron cell death is not clear, although this may represent loss of function of FUS/TLS in the nucleus or an acquired, toxic function of the mutant proteins in the cytosol. In the third group of ALS genes, the primary problem is defective axonal cytoskeleton and transport (dynactin, profilin-1). It is striking that variants in other genes (e.g., EphA4) influence survival in ALS but not ALS susceptibility. Data indicate that intermediate-length polyglutamine-coding expansions (-CAG-) in the gene *ataxin-2* confer increased ALS susceptibility; suppression of ataxin-2 expression extends survival in transgenic ALS mice. Beyond the upstream, primary defects, it is also evident that the ultimate neuronal cell death process is complex, involving multiple cellular processes acting in diverse components of the motor neuron (dendrites, cell body, axons, neuromuscular junction) to accelerate cell death. These include but are not limited to excitotoxicity, defective autophagy, impairment of axonal transport, oxidative stress, activation of endoplasmic reticulum stress and the unfolded protein response, and mitochondrial dysfunction. As well, the hexanucleotide expansions that cause C9orf72 ALS disrupt nucleocytoplasmic transport; the importance of this

observation is underscored by the finding that mutations in the gene encoding GLE1, a protein that mediates mRNA export, cause an aggressive, infantile motor neuron disease.

Multiple recent studies have convincingly demonstrated that proliferating, activated nonneuronal cells such as microglia and astrocytes importantly influence the disease course, at least in ALS transgenic mice. A striking additional finding in ALS and most neurodegenerative disorders is that miscreant proteins arising from gene defects in familial forms of these diseases are often implicated in sporadic forms of the same disorder. For example, some reports propose that nonheritable, posttranslational modifications in SOD1 are pathogenic in sporadic ALS; indeed, SOD1 aggregates are sometimes observed in spinal cord in sporadic ALS without SOD1 mutations. Germline mutations in the genes encoding β-amyloid and α-synuclein cause familial forms of Alzheimer's and Parkinson's diseases, and posttranslational, noninherited abnormalities in these proteins are also central to sporadic Alzheimer's and Parkinson's diseases.

TREATMENT

TREATMENT

Amyotrophic Lateral Sclerosis

No treatment arrests the underlying pathologic process in ALS. The drug riluzole (100 mg/d) was approved for ALS because it produces a modest lengthening of survival. In one trial, the survival rate at 18 months with riluzole was similar to placebo at 15 months. The mechanism of this effect is not known with certainty; riluzole may reduce excitotoxicity by diminishing glutamate release. Riluzole is generally well tolerated; nausea, dizziness, weight loss, and elevated liver enzymes occur occasionally. A second drug, edaravone, has also been approved by the U.S. Food and Drug Administration based on a single 6-month study in a highly selected ALS population that demonstrated a modest reduction in the trajectory of worsening on an ALS disability scale; survival was not included as an endpoint. This drug, which is believed to act as an antioxidant, is administered via recurring monthly 10-day series of daily intravenous infusions. Pathophysiologic studies of mutant SOD1–related ALS in mice have disclosed diverse targets for therapy; consequently, multiple therapies are presently in clinical trials for ALS including experimental trials of small molecules, mesenchymal stem cells, and immunosuppression. Interventions such as antisense oligonucleotides (ASO) that diminish expression of mutant SOD1 protein prolong survival in transgenic ALS mice and rats are also now in clinical trials for SOD1-mediated ALS.

In the absence of a primary therapy for ALS, a variety of rehabilitative aids may substantially assist ALS patients. Foot-drop splints facilitate ambulation by obviating the need for excessive hip flexion and by preventing tripping on a floppy foot. Finger extension splints can potentiate grip. Respiratory support may be life-sustaining. For patients electing against long-term ventilation by tracheostomy, positive-pressure ventilation by mouth or nose provides transient (weeks to months) relief from hypercarbia and hypoxia. Also extremely beneficial for some patients is a respiratory device (Cough Assist Device) that produces an artificial cough. This is highly effective in clearing airways and preventing aspiration pneumonia. When bulbar disease prevents normal chewing and swallowing, gastrostomy is uniformly helpful, restoring normal nutrition and hydration. Fortunately, an increasing variety of speech synthesizers are now available to augment speech when there is advanced bulbar palsy. These facilitate oral communication and may be effective for telephone use.

In contrast to ALS, several of the disorders (Tables 429-1 and 429-3) that bear some clinical resemblance to ALS are treatable. For this reason, a careful search for causes of secondary motor neuron disease is warranted.

OTHER MOTOR NEURON DISEASES

SELECTED LOWER MOTOR NEURON DISORDERS

In these motor neuron diseases, the peripheral motor neurons are affected without evidence of involvement of the corticospinal motor system (Tables 429-1, Tables 429-2, 429-3).

X-Linked Spinobulbar Muscular Atrophy (Kennedy's Disease)

This is an X-linked lower motor neuron disorder in which progressive weakness and wasting of limb and bulbar muscles begins in males in mid-adult life and is conjoined with androgen insensitivity manifested by gynecomastia and reduced fertility **(Chap. 384)**. In addition to gynecomastia, which may be subtle, two findings distinguishing this disorder from ALS are the absence of signs of pyramidal tract disease (spasticity) and the presence of a subtle sensory neuropathy in some patients. The underlying molecular defect is an expanded

trinucleotide repeat (-CAG-) in the first exon of the androgen receptor gene on the X chromosome. An inverse correlation appears to exist between the number of CAG- repeats and the age of onset of the disease.

Adult Tay-Sachs Disease

Several reports have described adult-onset, predominantly lower motor neuropathies arising from deficiency of the enzymeβ-hexosaminidase (hex A). These tend to be distinguishable from ALS because they are very slowly progressive and in some cases may have been symptomatic for years; dysarthria and radiographically evident cerebellar atrophy may be prominent. In rare cases, spasticity may also be present, although it is generally absent **(Chap. 411)**.

Spinal Muscular Atrophy

The SMAs are a family of selective lower motor neuron diseases of early onset. Despite some phenotypic variability (largely in age of onset), the defect in the majority of families with SMA is loss of a protein (SMN, for survival motor neuron) that is important in the formation and trafficking of RNA complexes across the nuclear membrane. Neuropathologically these disorders are characterized by extensive loss of large motor neurons; muscle biopsy reveals evidence of denervation atrophy. Several clinical forms exist.

Infantile SMA (SMA I, Werdnig-Hoffmann disease) has the earliest onset and most rapidly fatal course. In some instances, it is apparent even before birth, as indicated by decreased fetal movements late in the third trimester. Though alert, afflicted infants are weak and floppy (hypotonic) and lack muscle stretch reflexes. Death generally ensues within the first year of life. *Chronic childhood SMA* (SMA II) begins later in childhood and evolves with a more slowly progressive course. *Juvenile SMA* (SMA III, Kugelberg-Welander disease) manifests during late childhood and runs a slow, indolent course. Unlike most denervating diseases, in this chronic disorder weakness is greatest in the proximal muscles; indeed, the pattern of clinical weakness can suggest a primary myopathy such as limb-girdle dystrophy. Electrophysiologic and muscle biopsy evidence of denervation distinguish SMA III from the myopathic syndromes. Remarkably, two treatments have shown dramatic benefit in infantile SMA. One, nusinersen, now an approved therapy, entails administering small oligonucleotides that alter mRNA splicing of one of the SMN genes, generating sufficient normal SMN protein to provide clinical benefit (including prolonged survival). The other uses systemically administered adeno-associated virus (AAV) to deliver the missing SMN gene to motor neurons and other cells.

Multifocal Motor Neuropathy with Conduction Block

In this disorder, lower motor neuron function is regionally and chronically disrupted by focal blocks in conduction. Many cases have elevated serum titers of mono- and polyclonal antibodies to ganglioside GM1; it is hypothesized that the antibodies produce selective, focal, paranodal demyelination of motor neurons. MMCB is not typically associated with corticospinal signs. In contrast with ALS, MMCB may respond dramatically to therapy such as IV immunoglobulin or chemotherapy; thus, it is imperative that MMCB be excluded when considering a diagnosis of ALS.

Other Forms of Lower Motor Neuron Disease

In individual families, other syndromes characterized by selective lower motor neuron dysfunction in an SMA-like pattern have been described. There are rare X-linked and autosomal dominant forms of apparent SMA. There is an ALS variant of juvenile onset, the Fazio-Londe syndrome, that involves mainly the musculature innervated by the brainstem. A component of lower motor neuron dysfunction is also found in degenerative disorders such as Machado-Joseph disease and the related olivopontocerebellar degenerations **(Chap. 431)**.

SELECTED DISORDERS OF THE UPPER MOTOR NEURON

Primary Lateral Sclerosis

This rare disorder arises sporadically in adults in mid to late life. Clinically, PLS is characterized by progressive spastic weakness of the limbs, preceded or followed by spastic dysarthria and dysphagia, indicating combined involvement of the corticospinal and corticobulbar tracts. Fasciculations, amyotrophy, and sensory changes are absent; neither electromyography nor muscle biopsy shows denervation. On neuropathologic examination, there is selective loss of the large pyramidal cells in the precentral gyrus and degeneration of the corticospinal and corticobulbar projections. The peripheral motor neurons and other neuronal systems are spared. The course of PLS is variable; although long-term survival is documented, the course may be as aggressive as in ALS, with ~3-year survival from onset to death. Early in its course, PLS raises the question of multiple sclerosis or other demyelinating diseases as diagnostic considerations **(Chap. 436)**. A

myelopathy suggestive of PLS is infrequently seen with infection with the retrovirus human T cell lymphotropic virus 1 (HTLV-1) **(Chap. 434)**. The clinical course and laboratory testing will distinguish these possibilities.

Hereditary Spastic Paraplegia

In its pure form, HSP is usually transmitted as an autosomal trait; most adult-onset cases are dominantly inherited. There are more than 80 genetic types of HSP for which causative mutations in more than 60 genes have been identified. Table 429-3 lists more commonly identified genetic types of HSP. Symptoms usually begin in the third or fourth decade, presenting as progressive spastic weakness beginning in the lower extremities; however, there are variants with onset so early that the differential diagnosis includes cerebral palsy. HSP typically has a long survival, presumably because respiratory function is spared. Late in the illness, there may be urinary urgency and incontinence and sometimes fecal incontinence; sexual function tends to be preserved.

In pure forms of HSP, the spastic leg weakness is often accompanied by posterior column (vibration and position) abnormalities and disturbance of bowel and bladder function. Some family members may have spasticity without clinical symptoms.

By contrast, particularly when recessively inherited, HSP may have complex or complicated forms in which altered corticospinal and dorsal column function is accompanied by significant involvement of other regions of the nervous system, including amyotrophy, mental retardation, optic atrophy, and sensory neuropathy.

Neuropathologically, in HSP there is degeneration of the corticospinal tracts, which appear nearly normal in the brainstem but show increasing atrophy at more caudal levels in the spinal cord; in effect, this pathologic picture is of a dying-back or distal axonopathy of long neuronal fibers within the CNS.

Defects at numerous loci underlie both dominantly and recessively inherited forms of HSP (Table 429-3). The gene most commonly implicated in dominantly inherited HSP is *spastin*, which encodes a microtubule interacting protein. The most common childhood-onset dominant form arises from mutations in the *atlastin* gene.

An infantile-onset form of X-linked, recessive HSP arises from mutations in the gene for myelin proteolipid protein. This is an example of rather striking allelic variation, as most other mutations in the same gene cause not HSP but Pelizaeus-Merzbacher disease, a widespread disorder of CNS myelin. Another recessive variant is caused by defects in the *paraplegin* gene. Paraplegin has homology to metalloproteases that are important in mitochondrial function in yeast.

FURTHER READING

Brown RH, Al-Chalabi A: Review article: Amyotrophic lateral sclerosis. N Engl J Med 377:162, 2017. [PubMed: 28700839]

Finkel RS et al: Treatment of infantile-onset spinal muscular atrophy with nusinersin: A phase 2, open-label, dose-escalation study. Lancet 388:3017, 2016. [PubMed: 27939059]

Gendron TF et al: Poly(GP) proteins are a useful pharmacodynamic marker for C9ORF72-associated amyotrophic lateral sclerosis. Sci Transl Med 9:pii:eaai7866, 2017.

Miller TM et al: An antisense oligonucleotide against SOD1 delivered intrathecally for patients with SOD1 familial amyotrophic lateral sclerosis: A phase 1, randomised, first-in-man study. Lancet Neurol 12:435, 2013. [PubMed: 23541756]

Robberecht W, Philips T: The changing scene of amyotrophic lateral sclerosis. Nat Rev Neurosci 14:248, 2013. [PubMed: 23463272]

Schüle R et al: Hereditary spastic paraplegia: Clinicogenetic lessons from 608 patients. Ann Neurol 79:646, 2016. [PubMed: 26856398]

Taylor JP et al: Decoding ALS: From genes to mechanism. Nature 539:197, 2016. [PubMed: 27830784]

The Writing Group on Behalf of the Edaravone (MCI-186) ALS 19 Study Group: Safety and efficacy of edaravone in well defined patients with amyotrophic lateral sclerosis: A randomised, double-blind, placebo controlled trial. Lancet Neurol 16:505, 2017. [PubMed: 28522181]

WEBSITES

Several websites provide valuable information on ALS including those offered by the Muscular Dystrophy Association (www.mdausa.org), the Amyotrophic Lateral Sclerosis Association (www.alsa.org), and the World Federation of Neurology and the Neuromuscular Unit at Washington University in St. Louis (www.neuro.wustl.edu).

Chapter 430: Prion Diseases

Stanley B. Prusiner; Bruce L. Miller

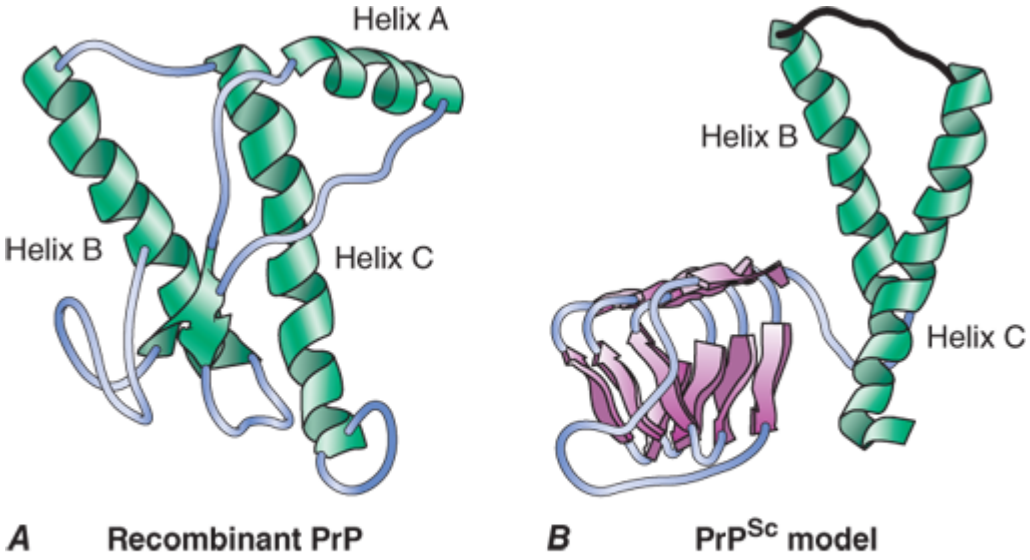
INTRODUCTION

Prions are proteins that adopt an alternative conformation, which becomes self-propagating. Some prions cause degeneration of the central nervous system (CNS). Once relegated to causing a group of rare disorders of the CNS such as Creutzfeldt-Jakob disease (CJD), prions also appear to play a role in more common illnesses such as Alzheimer's disease (AD) and Parkinson's disease (PD). While CJD is caused by the accumulation of PrP^{Sc} prions, recent investigations demonstrate unequivocally that α-synuclein prions cause multiple system atrophy (MSA). Infectious MSA prions have been recovered from human brain samples stored in formalin for up to 20 years. Similar resistance to formalin was demonstrated for brain samples from sheep with scrapie. Increasing data argue that Aβ prions contribute to AD, α-synuclein prions to PD, and tau prions to some types of frontotemporal dementia (FTD). In this chapter, we confine our discussion to CJD, which typically presents with a rapidly progressive dementia as well as motor abnormalities. The illness is relentlessly progressive and generally causes death within 9 months of onset. Most CJD patients are between 50 and 75 years of age; however, patients as young as 17 and as old as 83 have been recorded. The role of prions in the pathogenesis of neurodegenerative diseases is reviewed in **Chap. 417**.

CJD is one malady in a group of disorders caused by prions composed of the prion protein (PrP). PrP prions reproduce by binding to the normal, *c*ellular isoform of the *pri*on *p*rotein (PrP^{C}) and stimulating conversion of PrP^{C} into the di*s*ease-*c*ausing isoform PrP^{Sc}. PrP^{C} is rich in α-helix and has little β-structure, whereas PrP^{Sc} has less α-helix and a high amount of β-structure **(Fig. 430-1)**. This α-to-β structural transition in PrP is the fundamental event underlying this group of prion diseases **(Table 430-1)**.

FIGURE 430-1

Structures of PrP prion proteins. ***A.*** Nuclear magnetic resonance structure of Syrian hamster recombinant (rec) PrP(90–231). Presumably, the structure of the α-helical form of recPrP(90–231) resembles that of PrP^{C}. recPrP(90–231) is viewed from the interface where PrP^{Sc} is thought to bind to PrP^{C}. Shown are α-helices A (residues 144–157), B (172–193), and C (200–227). Flat ribbons depict β-strands S1 (129–131) and S2 (161–163). ***B.*** Structural model of PrP^{Sc}. The 90–160 region has been modeled onto a β-helical architecture while the COOH terminal helices B and C are preserved as in PrP^{C}.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 430-1

Glossary of PrP Prion Terminology

Prion	*Pro*teinaceous *in*fectious particle that lacks nucleic acid. Prions are composed entirely of alternatively folded proteins that undergo self-propagation. Distinct strains of prions exhibit different biologic properties, which are epigenetically heritable. PrP prions cause scrapie in sheep and goats, mad cow disease, and related neurodegenerative diseases of humans such as Creutzfeldt-Jakob disease (CJD).
PrP^{Sc}	Di*s*ease-*c*ausing isoform of the prion protein. This protein is the only identifiable macromolecule in purified preparations of scrapie prions.
PrP^{C}	*C*ellular isoform of the prion protein. PrP^{C} is the precursor of PrP^{Sc}.
PrP 27-30	A fragment of PrP^{Sc}, generated by truncation of the NH_2-terminus by limited digestion with proteinase K. PrP 27-30 retains prion infectivity and polymerizes into amyloid.
PRNP	PrP gene located on human chromosome 20.
Prion rod	An aggregate of prions composed largely of PrP 27-30 molecules. Created by detergent extraction and limited proteolysis of PrP^{Sc}. Morphologically and histochemically indistinguishable from many amyloids.
PrP amyloid	Amyloid containing PrP in the brains of animals or humans with prion disease; often accumulates as plaques.

Four new concepts have emerged from studies of PrP prions: (1) Prions are the only known transmissible pathogens that are devoid of nucleic acid; all other infectious agents possess genomes composed of either RNA or DNA that direct the synthesis of their progeny. (2) Prion diseases may manifest as infectious, genetic, or sporadic disorders; no other group of illnesses with a single etiology presents with such a wide spectrum of clinical manifestations. (3) Prion diseases result from the accumulation of PrP^{Sc}, the conformation of which differs substantially from that of its precursor, PrP^{C}. (4) Distinct strains of prions exhibit different biologic properties, which are epigenetically inherited. In other words, PrP^{Sc} can exist in a variety of different conformations, many of which seem to specify particular disease phenotypes.

How a specific conformation of a PrP^{Sc} molecule is imparted to PrP^{C} during prion replication to produce nascent PrP^{Sc} with the same conformation is unknown. Additionally, it is unclear what factors determine where in the CNS a particular PrP^{Sc} molecule will be deposited.

SPECTRUM OF PrP PRION DISEASES

The sporadic form of CJD is the most common prion disorder in humans. Sporadic CJD (sCJD) accounts for ~85% of all cases of human PrP prion disease, whereas inherited prion diseases account for 10–15% of all cases **(Table 430-2)**. Familial CJD (fCJD), Gerstmann-Sträussler-Scheinker (GSS) disease, and fatal familial insomnia (FFI) are all dominantly inherited prion diseases that are caused by mutations in the PrP gene.

TABLE 430-2

The PrP Prion Diseases

Disease	Host	Mechanism of Pathogenesis
Human		
Kuru	Fore people	Infection through ritualistic cannibalism
iCJD	Humans	Infection from prion-contaminated hGH, dura mater grafts, etc.
vCJD	Humans	Infection from bovine prions
fCJD	Humans	Germline mutations in *PRNP*
GSS	Humans	Germline mutations in *PRNP*
FFI	Humans	Germline mutation in *PRNP* (D178N, M129)
sCJD	Humans	Somatic mutation or spontaneous conversion of PrP^{C} into PrP^{Sc}?
sFI	Humans	Somatic mutation or spontaneous conversion of PrP^{C} into PrP^{Sc}?
Animal		
Scrapie	Sheep, goats	Infection in genetically susceptible sheep
BSE	Cattle	Infection with prion-contaminated MBM
TME	Mink	Infection with prions from sheep or cattle
CWD	Mule deer, elk	Unknown
FSE	Cats	Infection with prion-contaminated beef
Exotic ungulate encephalopathy	Greater kudu, nyala, or oryx	Infection with prion-contaminated MBM

Abbreviations: BSE, bovine spongiform encephalopathy; CJD, Creutzfeldt-Jakob disease; CWD, chronic wasting disease; fCJD, familial Creutzfeldt-Jakob disease; FFI, fatal familial insomnia; FSE, feline spongiform encephalopathy; GSS, Gerstmann-Sträussler-Scheinker disease; hGH, human growth hormone; iCJD, iatrogenic Creutzfeldt-Jakob disease; MBM, meat and bone meal; sCJD, sporadic Creutzfeldt-Jakob disease; sFI, sporadic fatal insomnia; TME, transmissible mink encephalopathy; vCJD, variant Creutzfeldt-Jakob disease.

Although infectious PrP prion diseases account for <1% of all cases and infection does not seem to play an important role in the natural history of these illnesses, the transmissibility of PrP prions is an important biologic feature. Kuru of the Fore people of New Guinea is thought to have resulted from the consumption of brains from dead relatives during ritualistic cannibalism. After the cessation of ritualistic cannibalism in the late 1950s, kuru nearly disappeared, with the exception of a few recent patients exhibiting incubation periods of >40 years. Iatrogenic CJD (iCJD) seems to be the result of the accidental inoculation of patients with prions. Variant CJD (vCJD) in teenagers and young adults in Europe is the result of exposure to tainted beef from cattle with bovine spongiform encephalopathy (BSE). Although occasional cases of iatrogenic CJD still occur, this form of CJD is currently on the decline due to public health measures aimed at preventing the spread of PrP prions.

Six diseases of animals are caused by prions (Table 430-2). Scrapie of sheep and goats is the prototypic PrP prion disease. Mink encephalopathy, BSE, feline spongiform encephalopathy, and exotic ungulate encephalopathy are all thought to occur after the consumption of prion-infected foodstuffs. The BSE epidemic emerged in Britain in the late 1980s and was shown to be due to industrial cannibalism. Whether BSE began as a sporadic case of BSE in a cow or started with scrapie in sheep is unknown. The origin of chronic wasting disease (CWD), a prion disease endemic in deer and elk in regions of North America, is uncertain. In contrast to other prion diseases, CWD is highly communicable. Feces from asymptomatic, infected cervids contain prions that are likely to be responsible for the spread of CWD.

EPIDEMIOLOGY

CJD is found throughout the world. The incidence of sCJD is ~1 case per million population, and accounts for ~1 in every 10,000 deaths. Because sCJD is an age-dependent neurodegenerative disease, its incidence is expected to increase steadily as older segments of populations in developed and developing countries continue to expand. Although many geographic clusters of CJD have been reported, each has been shown to segregate with a PrP gene mutation. Attempts to identify common exposure to some etiologic agent have been unsuccessful for both the sporadic and familial cases. Ingestion of scrapie-infected sheep or goat as a cause of CJD in humans has not been demonstrated by epidemiologic studies, although speculation about this potential route of infection continues. Of particular interest are deer hunters who develop CJD, because up to 90% of culled deer in some game herds have been shown to harbor CWD prions. Whether prion disease in deer or elk has passed to cows, sheep, or directly to humans remains unknown. Studies with rodents demonstrate that oral infection with prions can occur, but the process is inefficient compared to intracerebral inoculation.

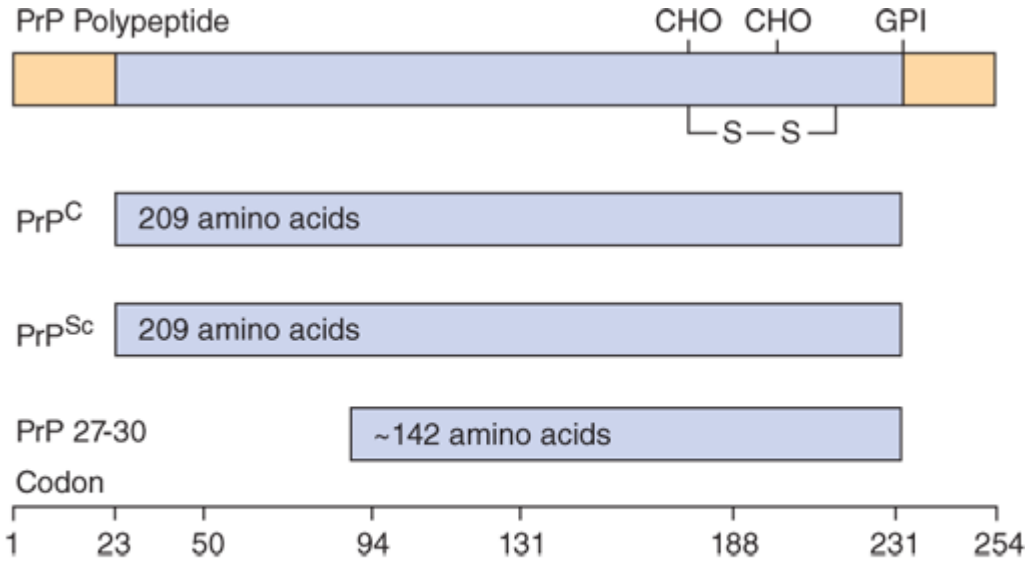
PATHOGENESIS

The human PrP prion diseases were initially classified as neurodegenerative disorders of unknown etiology on the basis of pathologic changes being confined to the CNS. With the transmission of kuru and CJD to apes, investigators began to view these diseases as infectious CNS illnesses caused by slow viruses. Even though the familial nature of a subset of CJD cases was well described, the significance of this observation became more obscure with the transmission of CJD to animals. Eventually, the meaning of heritable CJD became clear with the discovery of mutations in the *PRNP* gene of these patients. The prion concept explains how a disease can manifest as a heritable as well as an infectious illness. Moreover, the hallmark of all PrP prion diseases, whether sporadic, dominantly inherited, or acquired by infection, is that they involve the aberrant metabolism of PrP.

A major feature that distinguishes PrP prions from viruses is the finding that both PrP isoforms are encoded by a chromosomal gene. In humans, the PrP gene is designated *PRNP* and is located on the short arm of chromosome 20. Limited proteolysis of PrP^{Sc} produces a smaller, protease-resistant molecule of ~142 amino acids designated PrP 27-30; PrP^{C} is completely hydrolyzed under the same conditions **(Fig. 430-2)**. In the presence of detergent, PrP 27-30 polymerizes into amyloid. Prion rods formed by limited proteolysis and detergent extraction are indistinguishable from the filaments that aggregate to form PrP amyloid plaques in the CNS. Both the rods and the PrP amyloid filaments found in brain tissue exhibit similar ultrastructural morphology and green-gold birefringence after staining with Congo red dye.

FIGURE 430-2

PrP prion protein isoforms. Bar diagram of Syrian hamster PrP, which consists of 254 amino acids. After processing of the NH_2 and COOH termini, both PrP^{C} and PrP^{Sc} consist of 209 residues. After limited proteolysis, the NH_2 terminus of PrP^{Sc} is truncated to form PrP 27–30 composed of ~142 amino acids. GPI, glycosylphosphatidylinositol anchor attachment site; S—S, disulfide bond; CHO, N-linked sugars.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prion Strains

Distinct strains of PrP prions exhibit different biologic properties, which are epigenetically heritable. The existence of prion strains raised the question of how heritable biologic information can be enciphered in a molecule other than nucleic acid. Various strains of PrP prions have been defined by incubation times, distribution of neuronal vacuolation, and stabilities of PrP^{Sc} to denaturation. Subsequently, the patterns of PrP^{Sc} deposition were found to correlate with vacuolation profiles, and these patterns were also used to characterize prion strains.

Persuasive evidence that strain-specific information is enciphered in the tertiary structure of PrP^{Sc} comes from transmission of two different inherited human prion diseases to mice expressing a chimeric human-mouse PrP transgene. In FFI, the protease-resistant fragment of PrP^{Sc} after deglycosylation has a molecular mass of 19 kDa, whereas in fCJD and most sporadic prion diseases, it is 21 kDa **(Table 430-3)**. This difference in molecular mass was shown to be due to different sites of proteolytic cleavage at the NH_2 termini of the two human PrP^{Sc} molecules, reflecting different tertiary structures. These distinct conformations were not unexpected because the amino acid sequences of the PrP fragments differ. Extracts from the brains of patients with FFI transmitted disease to the mice expressing the chimeric human-mouse PrP transgene and induced formation of 19-kDa PrP^{Sc}, whereas brain extracts from fCJD and sCJD patients produced 21-kDa PrP^{Sc} in mice expressing the same transgene. On second passage, these differences were maintained, demonstrating that chimeric PrP^{Sc} can exist in two different conformations based on the sizes of the protease-resistant fragments, even though the amino acid sequence of PrP^{Sc} is invariant.

TABLE 430-3

Distinct Prion Strains Generated in Humans with Inherited Prion Diseases and Transmitted to Transgenic Mice[a]

Inoculum	Host Species	Host PrP Genotype	Incubation Time [days ± SEM] (n/n_0)	PrP^{Sc} (kDa)
None	Human	FFI(D178N, M129)		19
FFI	Mouse	Tg(MHu2M)	206 ± 7 (7/7)	19
FFI → Tg(MHu2M)	Mouse	Tg(MHu2M)	136 ± 1 (6/6)	19
None	Human	fCJD(E200K)		21
fCJD	Mouse	Tg(MHu2M)	170 ± 2 (10/10)	21
fCJD → Tg(MHu2M)	Mouse	Tg(MHu2M)	167 ± 3 (15/15)	21

[a]Tg(MHu2M) mice express a chimeric mouse-human PrP gene.

Notes: Clinicopathologic phenotype is determined by the conformation of PrP^{Sc} in accord with the results of the transmission of human prions from patients with FFI to transgenic mice.

Abbreviations: fCJD, familial Creutzfeldt-Jakob disease; FFI, fatal familial insomnia; SEM, standard error of the mean.

This analysis was extended when patients with sporadic fatal insomnia (sFI) were identified. Although they did not carry a *PRNP* gene mutation, the patients demonstrated a clinical and pathologic phenotype that was indistinguishable from that of patients with FFI. Furthermore, 19-kDa PrP^{Sc} was found in their brains, and on passage of prion disease to mice expressing the chimeric human-mouse PrP transgene, 19-kDa PrP^{Sc} was also found. These findings indicate that the disease phenotype is dictated by the conformation of PrP^{Sc} and not the amino acid sequence. PrP^{Sc} acts as a template for the conversion of PrP^{C} into nascent PrP^{Sc}. On the passage of prions into mice expressing a chimeric hamster-mouse PrP transgene, a change in the conformation of PrP^{Sc} was accompanied by the emergence of a new strain of prions.

Many new strains of prions were generated using recombinant (rec) PrP produced in bacteria; recPrP was polymerized into amyloid fibrils and inoculated into transgenic mice expressing high levels of wild-type mouse PrP^{C}. Approximately 500 days later, the mice died of prion disease. The incubation times of the "synthetic prions" in mice were dependent on the conditions used for polymerization of the amyloid fibrils. Highly stable amyloids gave rise to stable prions with long incubation times; low-stability amyloids led to

prions with short incubation times. Amyloids of intermediate stability gave rise to prions with intermediate stabilities and intermediate incubation times. Such findings are consistent with earlier studies showing that the incubation times of synthetic and naturally occurring prions are directly proportional to the stability of the prion.

Species Barrier

Studies on the role of the primary and tertiary structures of PrP in the transmission of prion disease have provided new insights into the pathogenesis of these maladies. The amino acid sequence of PrP encodes the species of the prion, and the prion derives its PrP^{Sc} sequence from the last mammal in which it was passaged. While the primary structure of PrP is likely to be the most important or even the sole determinant of the tertiary structure of PrP^{C}, PrP^{Sc} seems to function as a template in determining the tertiary structure of nascent PrP^{Sc} molecules as they are formed from PrP^{C}. In turn, prion diversity appears to be enciphered in the conformation of PrP^{Sc}, and thus prion strains seem to represent different conformers of PrP^{Sc}.

In general, transmission of PrP prion disease from one species to another is inefficient, in that not all intracerebrally inoculated animals develop disease, and those that fall ill do so only after long incubation times that can approach the natural life span of the animal. This “species barrier” to transmission is correlated with the degree of similarity between the amino acid sequences of PrP^{C} in the inoculated host and of PrP^{Sc} in the inoculum. The importance of sequence similarity between the host and donor PrP argues that PrP^{C} directly interacts with PrP^{Sc} in the prion conversion process.

SPORADIC AND INHERITED PrP PRION DISEASES

Several different scenarios might explain the initiation of sporadic prion disease: (1) A somatic mutation may be the cause and thus follow a path similar to that for germline mutations in inherited disease. In this situation, the mutant PrP^{Sc} must be capable of targeting wild-type PrP^{C}, a process known to be possible for some mutations but less likely for others. (2) The activation energy barrier separating wild-type PrP^{C} from PrP^{Sc} could be crossed on rare occasions when viewed in the context of a population. Most individuals would be spared, while presentations in the elderly with an incidence of ~1 per million would be seen. (3) PrP^{Sc} may be present at low levels in some normal cells, where it performs an important, but as yet unknown, function. The level of PrP^{Sc} in such cells is hypothesized to be sufficiently low as not to be detected by routine bioassay. In some altered metabolic states, the cellular mechanisms for clearing PrP^{Sc} might become compromised, and the rate of PrP^{Sc} formation would then begin to exceed the capacity of the cell to clear it. The third possible mechanism is attractive because it suggests that PrP^{Sc} is not simply a misfolded protein, as proposed for the first and second mechanisms, but that it is an alternatively folded molecule with a function. Moreover, the multitude of conformational states that PrP^{Sc} can adopt, as described above, raises

the possibility that PrP^{Sc} or another prion-like protein might function in a process like short-term memory where information storage occurs in the absence of new protein synthesis.

More than 40 different mutations resulting in nonconservative substitutions in the human *PRNP* gene have been found to segregate with inherited human prion diseases. Missense mutations and expansions in the octapeptide repeat region of the gene are responsible for familial forms of prion disease. Five different mutations of the *PRNP* gene have been linked genetically to heritable prion disease.

Although phenotypes may vary dramatically within families, specific phenotypes tend to be observed with certain mutations. A clinical phenotype indistinguishable from typical sCJD is usually seen with substitutions at codons 180, 183, 200, 208, 210, and 232. Substitutions at codons 102, 105, 117, 198, and 217 are associated with the GSS variant of prion disease with prominent Parkinsonian and cerebellar features. The normal human PrP sequence contains five repeats of an eight-amino-acid sequence. Insertions from two to nine extra octarepeats frequently cause variable phenotypes ranging from a condition indistinguishable from sCJD to a slowly progressive dementing illness of many years in duration to an early-age-of-onset disorder that is similar to AD. A mutation at codon 178 that results in substitution of asparagine for aspartic acid produces FFI if a methionine is encoded at the polymorphic residue 129 on the same allele. Typical CJD is seen if the D178N mutation occurs with a valine encoded at position 129 of the same allele.

HUMAN *PRNP* GENE POLYMORPHISMS

Polymorphisms influence the susceptibility to sporadic, inherited, and infectious forms of PrP prion disease. The methionine/valine polymorphism at position 129 not only modulates the age of onset of some inherited prion diseases but also can determine the clinical phenotype. The finding that homozygosity at codon 129 predisposes an individual to sCJD supports a model of prion production that favors PrP interactions between homologous proteins.

Substitution of the basic residue lysine at position 218 in mouse PrP produced dominant-negative inhibition of prion replication in transgenic mice. This same lysine at position 219 in human PrP has been found in 12% of the Japanese population, a group that appears to be resistant to prion disease. Dominant-negative inhibition of prion replication was also found with substitution of the basic residue arginine at position 171; sheep with arginine were resistant to scrapie prions but were susceptible to BSE prions that were inoculated intracerebrally.

INFECTIOUS PrP PRION DISEASES

IATROGENIC CJD

Accidental transmission of CJD to humans appears to have occurred with corneal transplantation, contaminated electroencephalogram (EEG) electrode implantation, and surgical procedures. Corneas from donors with unsuspected CJD have been transplanted to apparently healthy recipients who developed CJD after variable incubation periods. The same improperly decontaminated EEG electrodes that caused CJD in

two young patients with intractable epilepsy caused CJD in a chimpanzee 18 months after their experimental implantation.

Surgical procedures may have resulted in accidental inoculation of patients with prions, presumably because some instrument or apparatus in the operating theater became contaminated when a CJD patient underwent surgery. Although the epidemiology of these studies is highly suggestive, no proof for such episodes exists.

Dura Mater Grafts

More than 160 cases of CJD after implantation of dura mater grafts have been recorded. All of the grafts appear to have been acquired from a single manufacturer whose preparative procedures were inadequate to inactivate human prions. One case of CJD occurred after repair of an eardrum perforation with a pericardium graft.

Human Growth Hormone and Pituitary Gonadotropin Therapy

The transmission of CJD prions from contaminated human growth hormone (hGH) preparations derived from human pituitaries has been responsible for fatal cerebellar disorders with dementia in >180 patients ranging in age from 10 to 41 years. These patients received injections of hGH every 2–4 days for 4–12 years. If it is thought that these patients developed CJD from injections of prion-contaminated hGH preparations, the possible incubation periods range from 4 to 30 years. Only recombinant hGH is now used therapeutically so that possible contamination with prions is no longer an issue. Four cases of CJD have occurred in women receiving human pituitary gonadotropin.

VARIANT CJD

The restricted geographic occurrence and chronology of vCJD raised the possibility that BSE prions had been transmitted to humans through the consumption of tainted beef. More than 190 cases of vCJD have occurred, with >90% of these in Britain. Variant CJD has also been reported in people either living in or originating from France, Ireland, Italy, the Netherlands, Portugal, Spain, Saudi Arabia, the United States, Canada, and Japan.

The steady decline in the number of vCJD cases over the past decade argues that there will not be a prion disease epidemic in Europe, similar to those seen for BSE and kuru. What is certain is that prion-tainted meat should be prevented from entering the human food supply.

The most compelling evidence that vCJD is caused by BSE prions was obtained from experiments in mice expressing the bovine PrP transgene. Both BSE and vCJD prions were efficiently transmitted to these transgenic mice and with similar incubation periods. In contrast to sCJD prions, vCJD prions did not transmit disease efficiently to mice expressing a chimeric human-mouse PrP transgene. Earlier studies with nontransgenic mice suggested that vCJD and BSE might be derived from the same source because both inocula transmitted disease with similar but very long incubation periods.

Attempts to determine the origin of BSE and vCJD prions have relied on passaging studies in mice, some of which are described above, as well as studies of the conformation and glycosylation of PrP^{Sc}. One scenario suggests that a particular conformation of bovine PrP^{Sc} was selected for heat resistance during the rendering process and was then reselected multiple times as cattle infected by ingesting prion-contaminated meat and bone meal (MBM) were slaughtered and their offal rendered into more MBM. Variant CJD cases have virtually disappeared with protection of the beef supply in Europe.

NEUROPATHOLOGY

Frequently, the brains of patients with CJD have no recognizable abnormalities on gross examination. Patients who survive for several years have variable degrees of cerebral atrophy.

On light microscopy, the pathologic hallmarks of CJD are spongiform degeneration and astrocytic gliosis. The lack of an inflammatory response in CJD and other prion diseases is an important pathologic feature of these degenerative disorders. Spongiform degeneration is characterized by many 1- to 5-µm vacuoles in the neuropil between nerve cell bodies. Generally, the spongiform changes occur in the cerebral cortex, putamen, caudate nucleus, thalamus, and molecular layer of the cerebellum. Astrocytic gliosis is a constant but nonspecific feature of PrP prion diseases. Widespread proliferation of fibrous astrocytes is found throughout the gray matter of brains infected with CJD prions. Astrocytic processes filled with glial filaments form extensive networks.

Amyloid plaques have been found in ~10% of CJD cases. Purified CJD prions from humans and animals exhibit the ultrastructural and histochemical characteristics of amyloid when treated with detergents during limited proteolysis. On first passage of samples from some human Japanese CJD cases into mice, amyloid plaques were found. These plaques stain with antibodies raised against PrP.

The amyloid plaques of GSS disease are morphologically distinct from those seen in kuru or scrapie. GSS plaques consist of a central dense core of amyloid surrounded by smaller globules of amyloid. Ultrastructurally, they consist of a radiating fibrillar network of amyloid fibrils, with scant or no neuritic degeneration. The plaques can be distributed throughout the brain but are most frequently found in the cerebellum. They are often located adjacent to blood vessels. Congophilic angiopathy has been noted in some cases of GSS disease.

In vCJD, a characteristic feature is the presence of “florid plaques.” These are composed of a central core of PrP amyloid, surrounded by vacuoles in a pattern suggesting petals on a flower.

CLINICAL FEATURES

Nonspecific prodromal symptoms occur in approximately a third of patients with CJD and may include fatigue, sleep disturbance, weight loss, headache, anxiety, vertigo, malaise, and ill-defined pain. Most patients with CJD present with deficits in higher cortical function. Similarly, psychiatric symptoms, such as depression, psychosis, and visual hallucinations, are often the defining features of the illness. These deficits

almost always progress over weeks or months to a state of profound dementia characterized by memory loss, impaired judgment, and a decline in virtually all aspects of intellectual function. A few patients present with either visual impairment or cerebellar gait and coordination deficits. Frequently, the cerebellar deficits are rapidly followed by progressive dementia. Visual problems often begin with blurred vision and diminished acuity, rapidly followed by dementia.

Other symptoms and signs include extrapyramidal dysfunction manifested as rigidity, masklike facies, or (less commonly) choreoathetoid movements; pyramidal signs (usually mild); seizures (usually major motor) and, less commonly, hypoesthesia; supranuclear gaze palsy; optic atrophy; and vegetative signs such as changes in weight, temperature, sweating, or menstruation.

Myoclonus

Most patients (~90%) with CJD exhibit myoclonus that appears at various times throughout the illness. Unlike other involuntary movements, myoclonus persists during sleep. Startle myoclonus elicited by loud sounds or bright lights is frequent. It is important to stress that myoclonus is neither specific nor confined to CJD and tends to occur later in the course of CJD. Dementia with myoclonus can also be due to AD **(Chap. 423)**, dementia with Lewy bodies **(Chap. 426)**, corticobasal degeneration **(Chap. 424)**, cryptococcal encephalitis **(Chap. 210)**, or the myoclonic epilepsy disorder Unverricht-Lundborg disease **(Chap. 418)**.

Clinical Course

In documented cases of accidental transmission of CJD to humans, an incubation period of 1.5–2 years preceded the development of clinical disease. In other cases, incubation periods of up to 40 years have been suggested. Most patients with CJD live 6–12 months after the onset of clinical signs and symptoms, whereas some live for up to 5 years.

DIAGNOSIS

The constellation of dementia, myoclonus, and periodic electrical bursts in an afebrile 60-year-old patient generally indicates CJD. Clinical abnormalities in CJD are confined to the CNS. Fever, elevated sedimentation rate, leukocytosis in blood, or a pleocytosis in cerebrospinal fluid (CSF) should alert the physician to another etiology to explain the patient's CNS dysfunction, although there are rare cases of CJD in which mild CSF pleocytosis is observed.

Variations in the typical course appear in inherited and transmitted forms of the disease. Familial CJD has an earlier mean age of onset than sCJD. In GSS disease, ataxia is usually a prominent and presenting feature, with dementia occurring late in the disease course. GSS disease presents earlier than CJD (mean age 43 years) and is typically more slowly progressive than CJD; death usually occurs within 5 years of onset. FFI is characterized by insomnia and dysautonomia; dementia occurs only in the terminal phase of the illness. Rare sporadic cases have been identified. Variant CJD has an unusual clinical course, with a prominent psychiatric

prodrome that may include visual hallucinations and early ataxia, whereas frank dementia is usually a late sign of vCJD.

DIFFERENTIAL DIAGNOSIS

Many conditions mimic CJD. Dementia with Lewy bodies **(Chap. 426)** is the most common disorder to be mistaken for CJD. It can present in a subacute fashion with delirium, myoclonus, and extrapyramidal features. Other neurodegenerative disorders to consider include AD, FTD, corticobasal degeneration, progressive supranuclear palsy, ceroid lipofuscinosis, and myoclonic epilepsy with Lafora bodies. The absence of abnormalities on diffusion-weighted and fluid-attenuated inversion recovery (FLAIR) magnetic resonance imaging (MRI) will almost always distinguish these dementing conditions from CJD.

Hashimoto's encephalopathy, which presents as a subacute progressive encephalopathy with myoclonus and periodic triphasic complexes on the EEG, should be excluded in every case of suspected CJD. It is diagnosed by the finding of high titers of antithyroglobulin or antithyroid peroxidase (antimicrosomal) antibodies in the blood and improves with glucocorticoid therapy. Unlike CJD, fluctuations in severity typically occur in Hashimoto's encephalopathy.

Intracranial vasculitides **(Chap. 356)** may produce nearly all of the symptoms and signs associated with CJD, sometimes without systemic abnormalities. Myoclonus is exceptional with cerebral vasculitis, but focal seizures may confuse the picture. Prominent headache, absence of myoclonus, stepwise change in deficits, abnormal CSF, and focal white matter change on MRI or angiographic abnormalities all favor vasculitis.

Paraneoplastic conditions **(Chap. 90)**, particularly limbic encephalitis and cortical encephalitis, can also mimic CJD. In many of these patients, dementia appears prior to the diagnosis of a tumor, and in some, no tumor is ever found. Detection of the paraneoplastic antibodies is often the only way to distinguish these cases from CJD.

Other diseases that can simulate CJD include neurosyphilis **(Chap. 177)**, AIDS dementia complex **(Chap. 197)**, progressive multifocal leukoencephalopathy **(Chap. 132)**, subacute sclerosing panencephalitis, progressive rubella panencephalitis, herpes simplex encephalitis **(Chap. 132)**, diffuse intracranial tumor (gliomatosis cerebri; **Chap. 86**), anoxic encephalopathy, dialysis dementia, uremia, hepatic encephalopathy, voltage-gated potassium channel (VGkC) autoimmune encephalopathy, and lithium or bismuth intoxication.

LABORATORY TESTS

The only specific diagnostic tests for CJD and other human PrP prion diseases measure PrP^{Sc}. The most widely used method involves limited proteolysis that generates PrP 27-30, which is detected by immunoassay after denaturation. The conformation-dependent immunoassay (CDI) is based on immunoreactive epitopes that are exposed in PrP^{C} but buried in PrP^{Sc}. In humans, the diagnosis of CJD can be established by brain biopsy if PrP^{Sc} is detected although biopsy is rarely indicated. If no attempt is made to measure PrP^{Sc}, but the constellation of pathologic changes frequently found in CJD is seen in a brain

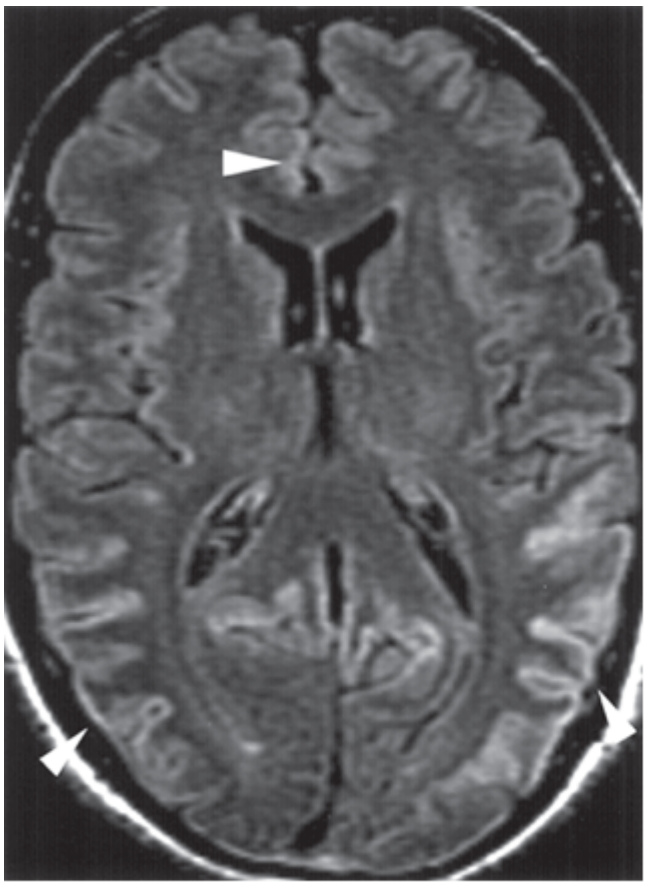
biopsy, then the diagnosis is reasonably secure (see "Neuropathology," above). The high sensitivity and specificity of cortical ribboning and basal ganglia hyperintensity on FLAIR and diffusion-weighted MRI for the diagnosis of CJD have greatly diminished the need for brain biopsy in patients with suspected CJD. Because PrP^{Sc} is not uniformly distributed throughout the CNS, the apparent absence of PrP^{Sc} in a limited sample such as a biopsy does not rule out prion disease. At autopsy, sufficient brain samples should be taken for both PrP^{Sc} immunoassay, preferably by CDI, and immunohistochemistry of tissue sections.

To establish the diagnosis of either sCJD or familial prion disease, sequencing the *PRNP* gene must be performed. Finding the wild-type *PRNP* gene sequence permits the diagnosis of sCJD if there is no history to suggest infection from an exogenous source of prions. The identification of a mutation in the *PRNP* gene sequence that encodes a nonconservative amino acid substitution argues for familial prion disease.

CT may be normal or show cortical atrophy. MRI is valuable for distinguishing sCJD from most other conditions. On FLAIR sequences and diffusion-weighted imaging, ~90% of patients show increased intensity in the basal ganglia and cortical ribboning **(Fig. 430-3)**. This pattern is not seen with other neurodegenerative disorders but has been seen infrequently with viral encephalitis, paraneoplastic syndromes, or seizures. When the typical MRI pattern is present, in the proper clinical setting, diagnosis is facilitated. However, some cases of sCJD do not show this typical pattern, and other early diagnostic approaches are still needed.

FIGURE 430-3

T2-weighted FLAIR MRI showing hyperintensity in the cortex in a patient with sCJD. This so-called "cortical ribboning" along with increased intensity in the basal ganglia on T2- or diffusion-weighted imaging can aid in the diagnosis of CJD.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CSF is nearly always normal but may show protein elevation and, rarely, mild pleocytosis. Although the stress protein 14-3-3 is elevated in the CSF of some patients with CJD, similar elevations of 14-3-3 are found in patients with other disorders; thus this elevation is not specific. Similarly, elevations of CSF neuron-specific enolase and tau occur in CJD but lack specificity for diagnosis.

The EEG is often useful in the diagnosis of CJD, although only ~60% of individuals show the typical pattern, which appears quite late in the clinical course. During the early phase of CJD, the EEG is usually normal or shows only scattered theta activity. In most advanced cases, repetitive, high-voltage, triphasic, and polyphasic sharp discharges are seen, but in many cases their presence is transient. The presence of these stereotyped periodic bursts of <200 ms in duration, occurring every 1–2 s, makes the diagnosis of CJD very likely. These discharges are frequently but not always symmetric; there may be a one-sided predominance in amplitude. As CJD progresses, normal background rhythms become fragmentary and slower.

CARE OF CJD PATIENTS

Although CJD should not be considered either contagious or communicable, it is transmissible. The risk of accidental inoculation by aerosols is very small; nonetheless, procedures producing aerosols should be performed in certified biosafety cabinets. Biosafety level 2 practices, containment equipment, and facilities are recommended by the Centers for Disease Control and Prevention and the National Institutes of Health. The primary worry in caring for patients with CJD is the inadvertent infection of health care workers by

needle and stab wounds, although with the possible exception of vCJD even blood transfusions appear to carry little risk for transmission. Electroencephalographic and electromyographic needles should not be reused after studies on patients with CJD have been performed.

There is no reason for pathologists or other morgue employees to resist performing autopsies on patients whose clinical diagnosis was CJD. Standard microbiologic practices outlined here, along with specific recommendations for decontamination, seem to be adequate precautions for the care of patients with CJD and the handling of infected specimens.

DECONTAMINATION OF CJD PRIONS

Prions are extremely resistant to common inactivation procedures, and there is some disagreement about the optimal conditions for sterilization. Some investigators recommend treating CJD-contaminated materials once with 1 *N* NaOH at room temperature, but we believe this procedure may be inadequate for sterilization. Autoclaving at 134°C for 5 h or treatment with 2 *N* NaOH for several hours is recommended for sterilization of prions. The term *sterilization* implies complete destruction of prions; any residual infectivity can be hazardous. Transgenic mouse studies show that sCJD prions bound to stainless steel surfaces are resistant to inactivation by autoclaving at 134°C for 2 h; exposure of bound prions to an acidic detergent solution prior to autoclaving rendered prions susceptible to inactivation. Recent studies show that α-synuclein prions in brain homogenates prepared from MSA patients bind to stainless steel wires and that the bound prions can be transmitted to transgenic mice expressing mutant human α-synuclein.

PREVENTION AND THERAPEUTICS

There is no known effective therapy for preventing or treating CJD. The finding that phenothiazines and acridines inhibit PrP^{Sc} formation in cultured cells led to clinical studies of quinacrine in CJD patients. Unfortunately, quinacrine failed to slow the rate of cognitive decline in CJD, possibly because therapeutic concentrations in the brain were not achieved. Although inhibition of the P-glycoprotein (Pgp) transport system resulted in substantially increased quinacrine levels in the brains of mice, the prion incubation times were not extended by treatment with the drug. Whether such an approach can be used to treat CJD remains to be established.

Like the acridines, anti-PrP antibodies have been shown to eliminate PrP^{Sc} from cultured cells. Additionally, such antibodies in mice, either administered by injection or produced from a transgene, have been shown to prevent prion disease when prions are introduced by a peripheral route, such as intraperitoneal inoculation. Unfortunately, the antibodies were ineffective in mice inoculated intracerebrally with prions. Several drugs, including pentosan polysulfate as well as porphyrin and phenylhydrazine derivatives, delay the onset of disease in animals inoculated intracerebrally with prions if the drugs are given intracerebrally beginning soon after inoculation.

DIFFERENT PRIONS CAUSING OTHER NEURODEGENERATIVE DISEASES

There is a rapidly expanding body of literature demonstrating that besides PrP, other proteins including amyloid beta (Aβ), tau, α-synuclein, and huntingtin can all become prions **(Chap. 417)**. Experimental studies have shown that transgenic mice expressing mutant amyloid precursor protein (APP) develop amyloid plaques containing fibrils composed of the Aβ peptide ~6 months after inoculation with synthetic Aβ peptides polymerized into amyloid fibrils or extracts prepared from the brains of patients with AD. Mutant tau aggregates in transgenic mice and cultured cells can trigger the aggregation of tau into fibrils that resemble those found in neurofibrillary tangles and Pick bodies. Such tangles have been found in AD, FTDs, Pick's disease, and some cases of posttraumatic brain injury called chronic traumatic encephalopathy, all of which are thought to be caused by the prion isoforms of Aβ and/or tau.

In patients with advanced PD who received grafts of fetal substantia nigral neurons, Lewy bodies containing β-sheet-rich α-synuclein were identified in grafted cells ~10 years after transplantation, arguing for the axonal transport of misfolded α-synuclein crossing into grafted neurons, where it initiated aggregation of nascent α-synuclein into fibrils that coalesced into Lewy bodies. These findings combined with MSA studies argue that the synucleinopathies are caused by prions. Brain homogenates from MSA patients injected into transgenic mice transmitted lethal neurodegeneration in ~3 months; moreover, recombinant synuclein injected into wild-type mice initiated the deposition of synuclein fibrils.

In summary, a wealth of evidence continues to accumulate arguing that proteins causing AD, PD, FTDs, amyotrophic lateral sclerosis (ALS), and even Huntington's disease (HD) acquire alternative conformations that become self-propagating. Each of these neurodegenerative diseases is thought to be caused by the abnormal aggregation of a different protein that undergoes a self-replicating conformational change to become a prion. Prions explain many of the features that the neurodegenerative diseases have in common: (1) incidence increases with age, (2) steady progression over years, (3) spread from one region of the CNS to another, (4) protein deposits often but not always consisting of amyloid fibrils, and (5) late onset of inherited forms. Notably, amyloid plaques containing PrP^{Sc} are a nonobligatory feature of PrP prion disease in humans and animals. Furthermore, amyloid plaques in AD do not correlate with the level of dementia; however, the level of soluble (oligomeric) Aβ peptide does correlate with memory loss and other intellectual deficits.

FURTHER READING

Collinge J: Mammalian prions and their wider relevance in neurodegenerative diseases. Nature 539:217, 2016. [PubMed: 27830781]

Prusiner SB et al: Evidence for α-synuclein prions causing multiple system atrophy in humans with parkinsonism. Proc Natl Acad Sci U S A 112:E5308, 2015. [PubMed: 26324905]

Prusiner SB (ed): *Prion Biology*. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press, 2017.

Prusiner SB (ed): *Prion Diseases*. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press, 2017.

Zweckstetter M, Requena JR, Wille H: Elucidating the structure of an infectious protein. PLoS Pathog 13:e1006229, 2017. [PubMed: 28406976]

Chapter 431: Ataxic Disorders

Roger N. Rosenberg

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Ataxic Disorders

Symptoms and signs of ataxia consist of gait impairment, unclear (“scanning”) speech, visual blurring due to nystagmus, hand incoordination, and tremor with movement. These result from the involvement of the cerebellum and its afferent and efferent pathways, including the spinocerebellar pathways, and the frontopontocerebellar pathway originating in the rostral frontal lobe. True cerebellar ataxia must be distinguished from ataxia associated with vestibular nerve or labyrinthine disease, as the latter results in a disorder of gait associated with a significant degree of dizziness, light-headedness, or the perception of movement **(Chap. 19)**. True cerebellar ataxia is devoid of these vertiginous complaints and is clearly an unsteady gait due to imbalance. Sensory disturbances can also on occasion simulate the imbalance of cerebellar disease; with sensory ataxia, imbalance dramatically worsens when visual input is removed (Romberg sign). Rarely, weakness of proximal leg muscles mimics cerebellar disease. In the patient who presents with ataxia, the rate and pattern of the development of cerebellar symptoms help to narrow the diagnostic possibilities **(Table 431-1)**. A gradual and progressive increase in symptoms with bilateral and symmetric involvement suggests a genetic, metabolic, immune, or toxic etiology. Conversely, focal, unilateral symptoms with headache and impaired level of consciousness accompanied by ipsilateral cranial nerve palsies and contralateral weakness imply a space-occupying cerebellar lesion.

SYMMETRIC ATAXIA

Progressive and symmetric ataxia can be classified with respect to onset as acute (over hours or days), subacute (weeks or months), or chronic (months to years). Acute and reversible ataxias include those caused by intoxication with alcohol, phenytoin, lithium, barbiturates, and other drugs. Intoxication caused by toluene exposure, gasoline sniffing, glue sniffing, spray painting, or exposure to methyl mercury or bismuth are additional causes of acute or subacute ataxia, as is treatment with cytotoxic chemotherapeutic drugs such as fluorouracil and paclitaxel. Patients with a postinfectious syndrome (especially after varicella) may develop gait ataxia and mild dysarthria, both of which are reversible **(Chap. 436)**. Rare infectious causes of acquired ataxia include poliovirus, coxsackievirus, echovirus, Epstein-Barr virus, toxoplasmosis, *Legionella*, and Lyme disease.

The subacute development of ataxia of gait over weeks to months (degeneration of the cerebellar vermis) may be due to the combined effects of alcoholism and malnutrition, particularly with deficiencies of vitamins B_1 and B_{12}. Hyponatremia has also been associated with ataxia. Paraneoplastic cerebellar ataxia is associated with a number of different tumors (and autoantibodies) such as breast and ovarian cancers (anti-Yo), small-cell lung cancer (anti-PQ-type voltage-gated calcium channel), and Hodgkin's disease (anti-Tr) **(Chap. 90)**. Another paraneoplastic syndrome associated with myoclonus and opsoclonus occurs with breast (anti-Ri) and lung cancers and neuroblastoma. Elevated serum anti-glutamic acid decarboxylase (GAD) antibodies have been associated with a progressive ataxic syndrome affecting speech and gait. For all of these paraneoplastic ataxias, the neurologic syndrome may be the presenting symptom of the cancer. Another immune-mediated progressive ataxia is associated with antigliadin (and antiendomysium) antibodies and the human leukocyte antigen (HLA) DQB1*0201 haplotype; in some affected patients, biopsy of the small intestine reveals villus atrophy consistent with gluten-sensitive enteropathy **(Chap. 318)**. Finally, subacute progressive ataxia may be caused by a prion disorder, especially when an infectious etiology, such as transmission from contaminated human growth hormone, is responsible **(Chap. 430)**.

Chronic symmetric gait ataxia suggests an inherited ataxia (discussed below), a metabolic disorder, or a chronic infection. Hypothyroidism must always be considered as a readily treatable and reversible form of gait ataxia. Infectious diseases that can present with ataxia are meningovascular syphilis and tabes dorsalis due to degeneration of the posterior columns and spinocerebellar pathways in the spinal cord.

FOCAL ATAXIA

Acute focal ataxia commonly results from cerebrovascular disease, usually ischemic infarction or cerebellar hemorrhage. These lesions typically produce cerebellar symptoms ipsilateral to the injured cerebellum and may be associated with an impaired level of consciousness due to brainstem compression and increased intracranial pressure; ipsilateral pontine signs, including sixth and seventh nerve palsies, may be present. Focal and worsening signs of acute ataxia should also prompt consideration of a posterior fossa subdural hematoma, bacterial abscess, or primary or metastatic cerebellar tumor. Computed tomography (CT) or magnetic resonance imaging (MRI) studies will reveal clinically significant processes of this type. Many of these lesions represent true neurologic emergencies, as sudden herniation, either rostrally through the tentorium or caudal herniation of cerebellar tonsils through the foramen magnum, can occur and is usually devastating. Acute surgical decompression may be required **(Chap. 301)**. Lymphoma or progressive multifocal leukoencephalopathy (PML) in a patient with AIDS may present with an acute or subacute focal cerebellar syndrome. Chronic etiologies of progressive ataxia include multiple sclerosis **(Chap. 436)** and congenital lesions such as a Chiari malformation **(Chap. 434)** or a congenital cyst of the posterior fossa (Dandy-Walker syndrome).

TABLE 431-1

Etiology of Cerebellar Ataxia

Symmetric and Progressive Signs			Focal and Ipsilateral Cerebellar Signs		
Acute (Hours to Days)	**Subacute (Days to Weeks)**	**Chronic (Months to Years)**	**Acute (Hours to Days)**	**Subacute (Days to Weeks)**	**Chronic (Months to Years)**
Intoxication: alcohol, lithium, phenytoin, barbiturates (positive history and toxicology screen) Acute viral cerebellitis (CSF supportive of acute viral infection) Postinfection syndrome	Intoxication: mercury, solvents, gasoline, glue; cytotoxic chemotherapeutic, hemotherapeutic drugs Alcoholic-nutritional (vitamin B_1 and B_{12} deficiency) Lyme disease	Paraneoplastic syndrome Antigliadin antibody syndrome Hypothyroidism Inherited diseases Tabes dorsalis (tertiary syphilis) Phenytoin toxicity Amiodarone	Vascular: cerebellar infarction, hemorrhage, or subdural hematoma Infectious: cerebellar abscess (mass lesion on MRI/CT, history in support of lesion)	Neoplastic: cerebellar glioma or metastatic tumor (positive for neoplasm on MRI/CT) Demyelinating: multiple sclerosis (history, CSF, and MRI are consistent) AIDS-related multifocal leukoencephalopathy (positive HIV test and CD4+ cell count for AIDS)	Stable gliosis secondary to vascular lesion or demyelinating plaque (stable lesion on MRI/CT older than several months) Congenital lesion: Chiari or Dandy-Walker malformations (malformation noted on MRI/CT)

Abbreviations: CSF, cerebrospinal fluid; CT, computed tomography; MRI, magnetic resonance imaging.

THE INHERITED ATAXIAS

These may show autosomal dominant, autosomal recessive, or maternal (mitochondrial) modes of inheritance. A genomic classification **(Chap. S10)** has now largely superseded previous ones based on clinical expression alone.

Although the clinical manifestations and neuropathologic findings of cerebellar disease dominate the clinical picture, there may also be characteristic changes in the basal ganglia, brainstem, spinal cord, optic nerves, retina, and peripheral nerves. In large families with dominantly inherited ataxias, many gradations are

observed from purely cerebellar manifestations to mixed cerebellar and brainstem disorders, cerebellar and basal ganglia syndromes, and spinal cord or peripheral nerve disease. Rarely, dementia is present as well. The clinical picture may be homogeneous within a family with dominantly inherited ataxia, but sometimes most affected family members show one characteristic syndrome, while one or several members have an entirely different phenotype.

AUTOSOMAL DOMINANT ATAXIAS

The autosomal spinocerebellar ataxias (SCAs) include SCA types 1 through 40, dentatorubropallidoluysian atrophy (DRPLA), and episodic ataxia (EA) types 1 to 7 **(Chap. S10)**. SCA1, SCA2, SCA3 (Machado-Joseph disease [MJD]), SCA6, SCA7, and SCA17 are caused by CAG triplet repeat expansions in different genes. SCA8 is due to an untranslated CTG repeat expansion, SCA12 is linked to an untranslated CAG repeat, and SCA10 is caused by an untranslated pentanucleotide repeat. The clinical phenotypes of these SCAs overlap. The genotype has become the gold standard for diagnosis and classification. CAG encodes glutamine, and these expanded CAG triplet repeat expansions result in expanded polyglutamine proteins, termed *ataxins*, that produce a toxic gain of function with autosomal dominant inheritance. Although the phenotype is variable for any given disease gene, a pattern of neuronal loss with gliosis is produced that is relatively unique for each ataxia. Immunohistochemical and biochemical studies have shown cytoplasmic (SCA2), neuronal (SCA1, MJD, SCA7), and nucleolar (SCA7) accumulation of the specific mutant polyglutamine-containing ataxin proteins. Expanded polyglutamine ataxins with more than ~40 glutamines are potentially toxic to neurons for a variety of reasons including: high levels of gene expression for the mutant polyglutamine ataxin in affected neurons; conformational change of the aggregated protein to a β-pleated structure; abnormal transport of the ataxin into the nucleus (SCA1, MJD, SCA7); binding to other polyglutamine proteins, including the TATA-binding transcription protein and the CREB-binding protein, impairing their functions; altering the efficiency of the ubiquitin-proteasome system of protein turnover; and inducing neuronal apoptosis. An earlier age of onset (anticipation) and more aggressive disease in subsequent generations are due to further expansion of the CAG triplet repeat and increased polyglutamine number in the mutant ataxin. The most common disorders are discussed below.

SCA1

SCA1 was previously referred to as *olivopontocerebellar atrophy*, but genomic data have shown that that entity represents several different genotypes with overlapping clinical features.

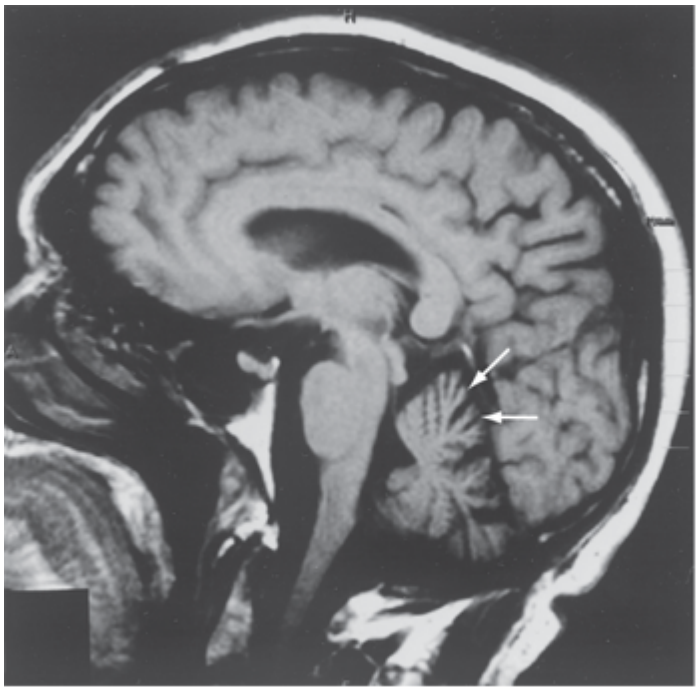
Symptoms and Signs

SCA1 is characterized by the development in early or middle adult life of progressive cerebellar ataxia of the trunk and limbs, impairment of equilibrium and gait, slowness of voluntary movements, scanning speech, nystagmoid eye movements, and oscillatory tremor of the head and trunk. Dysarthria, dysphagia, and oculomotor and facial palsies may also occur. Extrapyramidal symptoms include rigidity, an immobile face, and parkinsonian tremor. The reflexes are usually normal, but knee and ankle jerks may be lost, and extensor plantar responses may occur. Dementia may be noted but is usually mild. Impairment of sphincter function is

common, with urinary and sometimes fecal incontinence. Cerebellar and brainstem atrophy are evident on MRI **(Fig. 431-1)**.

FIGURE 431-1

Sagittal magnetic resonance imaging (MRI) of the brain of a 60-year-old man with gait ataxia and dysarthria due to spinocerebellar ataxia type 1 (SCA1), illustrating cerebellar atrophy (*arrows*). *(Reproduced with permission from RN Rosenberg, P Khemani, in RN Rosenberg, JM Pascual [eds]: Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease, 5th ed. London, Elsevier, 2015.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Marked shrinkage of the ventral half of the pons, disappearance of the olivary eminence on the ventral surface of the medulla, and atrophy of the cerebellum are evident on gross postmortem inspection of the brain. Variable loss of Purkinje cells, reduced numbers of cells in the molecular and granular layer, demyelination of the middle cerebellar peduncle and the cerebellar hemispheres, and severe loss of cells in the pontine nuclei and olives are found on histologic examination. Degenerative changes in the striatum, especially the putamen, and loss of the pigmented cells of the substantia nigra may be found in cases with extrapyramidal features. More widespread degeneration in the central nervous system (CNS), including involvement of the posterior columns and the spinocerebellar fibers, is often present.

GENETIC CONSIDERATIONS



SCA1 encodes a gene product, called *ataxin-1*, which is a novel protein of unknown function. The mutant allele has 40 CAG repeats located within the coding region, whereas alleles from unaffected individuals have ≤36 repeats. A few patients with 38–40 CAG repeats have been described. There is a direct correlation between a larger number of repeats and a younger age of onset for SCA1. Juvenile patients have higher

numbers of repeats, and anticipation is present in subsequent generations. Transgenic mice carrying SCA1 developed ataxia and Purkinje cell pathology. Leucine-rich acidic nuclear protein localization, but not aggregation, of ataxin-1 appears to be required for cell death initiated by the mutant protein.

SCA2

Symptoms and Signs

Another clinical phenotype, SCA2, has been described in patients from Cuba and India. Cuban patients probably are descendants of a common ancestor, and the population may be the largest homogeneous group of patients with ataxia yet described. The age of onset ranges from 2 to 65 years, and there is considerable clinical variability within families. Although neuropathologic and clinical findings are compatible with a diagnosis of SCA1, including slow saccadic eye movements, ataxia, dysarthria, parkinsonian rigidity, optic disc pallor, mild spasticity, and retinal degeneration, SCA2 is a unique form of cerebellar degenerative disease.

GENETIC CONSIDERATIONS


The gene in SCA2 families also contains CAG repeat expansions coding for a polyglutamine-containing protein, ataxin-2. Normal alleles contain 15–32 repeats; mutant alleles have 35–77 repeats.

MACHADO-JOSEPH DISEASE/SCA3

MJD was first described among the Portuguese and their descendants in New England and California. Subsequently, MJD has been found in families from Portugal, Australia, Brazil, Canada, China, England, France, India, Israel, Italy, Japan, Spain, Taiwan, and the United States. In most populations, it is the most common autosomal dominant ataxia.

Symptoms and Signs

MJD has been classified into three clinical types. In type I MJD (amyotrophic lateral sclerosis-parkinsonism-dystonia type), neurologic deficits appear in the first two decades and involve weakness and spasticity of extremities, especially the legs, often with dystonia of the face, neck, trunk, and extremities. Patellar and ankle clonus are common, as are extensor plantar responses. The gait is slow and stiff, with a slightly broadened base and lurching from side to side; this gait results from spasticity, not true ataxia. There is no truncal titubation. Pharyngeal weakness and spasticity cause difficulty with speech and swallowing. Of note is the prominence of horizontal and vertical nystagmus, loss of fast saccadic eye movements, hypermetric and hypometric saccades, and impairment of upward vertical gaze. Facial fasciculations, facial myokymia, lingual fasciculations without atrophy, ophthalmoparesis, and ocular prominence are common early manifestations.

In type II MJD (ataxic type), true cerebellar deficits of dysarthria and gait and extremity ataxia begin in the second to fourth decades along with corticospinal and extrapyramidal deficits of spasticity, rigidity, and dystonia. Type II is the most common form of MJD. Ophthalmoparesis, upward vertical gaze deficits, and facial and lingual fasciculations are also present. Type II MJD can be distinguished from the clinically similar disorders SCA1 and SCA2.

Type III MJD (ataxic-amyotrophic type) presents in the fifth to the seventh decades with a pancerebellar disorder that includes dysarthria and gait and extremity ataxia. Distal sensory loss involving pain, touch, vibration, and position senses and distal atrophy are prominent, indicating the presence of peripheral neuropathy. The deep tendon reflexes are depressed to absent, and there are no corticospinal or extrapyramidal findings.

The mean age of onset of symptoms in MJD is 25 years. Neurologic deficits invariably progress and lead to death from debilitation within 15 years of onset, especially in patients with types I and II disease. Usually, patients retain full intellectual function.

The major pathologic findings are variable loss of neurons and glial replacement in the corpus striatum and severe loss of neurons in the pars compacta of the substantia nigra. A moderate loss of neurons occurs in the dentate nucleus of the cerebellum and in the red nucleus. Purkinje cell loss and granule cell loss occur in the cerebellar cortex. Cell loss also occurs in the dentate nucleus and in the cranial nerve motor nuclei. Sparing of the inferior olives distinguishes MJD from other dominantly inherited ataxias.

GENETIC CONSIDERATIONS


The gene for MJD maps to 14q24.3-q32. Unstable CAG repeat expansions are present in the MJD gene coding for a polyglutamine-containing protein named ataxin-3, or MJD-ataxin. An earlier age of onset is associated with longer repeats. Alleles from normal individuals have between 12 and 37 CAG repeats, whereas MJD alleles have 60–84 CAG repeats. Polyglutamine-containing aggregates of ataxin-3 (MJD-ataxin) have been described in neuronal nuclei undergoing degeneration. MJD-ataxin codes for a ubiquitin protease, which is inactive due to expanded polyglutamines. Proteosome function is impaired, resulting in altered clearance of proteins and cerebellar neuronal loss.

SCA6

Genomic screening for CAG repeats in other families with autosomal dominant ataxia and vibratory and proprioceptive sensory loss have yielded another locus. Of interest is that different mutations in the same gene for the α_{1A} voltage-dependent calcium channel subunit (CACNLIA4; also referred to as the *CACNA1A* gene) at 19p13 result in different clinical disorders. CAG repeat expansions (21–27 in patients; 4–16 triplets in normal individuals) result in late-onset progressive ataxia with cerebellar degeneration. Missense mutations in this gene result in familial hemiplegic migraine. Nonsense mutations resulting in termination of protein

synthesis of the gene product yield hereditary paroxysmal cerebellar ataxia or EA. Some patients with familial hemiplegic migraine develop progressive ataxia and also have cerebellar atrophy.

SCA7

This disorder is distinguished from all other SCAs by the presence of retinal pigmentary degeneration. The visual abnormalities first appear as blue-yellow color blindness and proceed to frank visual loss with macular degeneration. In almost all other respects, SCA7 resembles several other SCAs in which ataxia is accompanied by various noncerebellar findings, including ophthalmoparesis and extensor plantar responses. The genetic defect is an expanded CAG repeat in the SCA7 gene at 3p14-p21.1. The expanded repeat size in SCA7 is highly variable. Consistent with this, the severity of clinical findings varies from essentially asymptomatic to mild late-onset symptoms to severe, aggressive disease in childhood with rapid progression. Marked anticipation has been recorded, especially with paternal transmission. The disease protein, ataxin-7, forms aggregates in nuclei of affected neurons, as has also been described for SCA1 and SCA3/MJD. Ataxin 7 is a subunit of GCN5, a histone acetyltransferase-containing complex.

SCA8

This form of ataxia is caused by a CTG repeat expansion in an untranslated region of a gene on chromosome 13q21. There is marked maternal bias in transmission, perhaps reflecting contractions of the repeat during spermatogenesis. The mutation is not fully penetrant. Symptoms include slowly progressive dysarthria and gait ataxia beginning at ~40 years of age with a range between 20 and 65 years. Other features include nystagmus, leg spasticity, and reduced vibratory sensation. Severely affected individuals are nonambulatory by the fourth to sixth decades. MRI shows cerebellar atrophy. The mechanism of disease may involve a dominant “toxic” effect occurring at the RNA level, as occurs in myotonic dystrophy.

DENTATORUBROPALLIDOLUYSIAN ATROPHY

DRPLA has a variable presentation that may include progressive ataxia, choreoathetosis, dystonia, seizures, myoclonus, and dementia. DRPLA is due to unstable CAG triplet repeats in the open reading frame of a gene named *atrophin* located on chromosome 12p12-ter. Larger expansions are found in patients with earlier onset. The number of repeats is 49 in patients with DRPLA and ≤26 in normal individuals. Anticipation occurs in successive generations, with earlier onset of disease in association with an increasing CAG repeat number in children who inherit the disease from their father. One well-characterized family in North Carolina has a phenotypic variant known as the *Haw River syndrome*, now recognized to be due to the DRPLA mutation.

EPISODIC ATAXIA

EA types 1 and 2 are two rare dominantly inherited disorders that have been mapped to chromosomes 12p (a potassium channel gene, KCNA1, Phe249Leu mutation) for type 1 and 19p for type 2. Patients with EA-1 have brief episodes of ataxia with myokymia and nystagmus that last only minutes. Startle, sudden change in posture, and exercise can induce episodes. Acetazolamide or anticonvulsants may be therapeutic. Patients

with EA-2 have episodes of ataxia with nystagmus that can last for hours or days. Stress, exercise, or excessive fatigue may be precipitants. Acetazolamide may be therapeutic and can reverse the relative intracellular alkalosis detected by magnetic resonance spectroscopy. Stop codon, nonsense mutations causing EA-2 have been found in the *CACNA1A* gene, encoding the α_{1A} voltage-dependent calcium channel subunit (see "SCA6," above).

AUTOSOMAL RECESSIVE ATAXIAS

Friedreich's Ataxia

This is the most common form of inherited ataxia, comprising one-half of all hereditary ataxias. It can occur in a classic form or in association with a genetically determined vitamin E deficiency syndrome; the two forms are clinically indistinguishable.

SYMPTOMS AND SIGNS

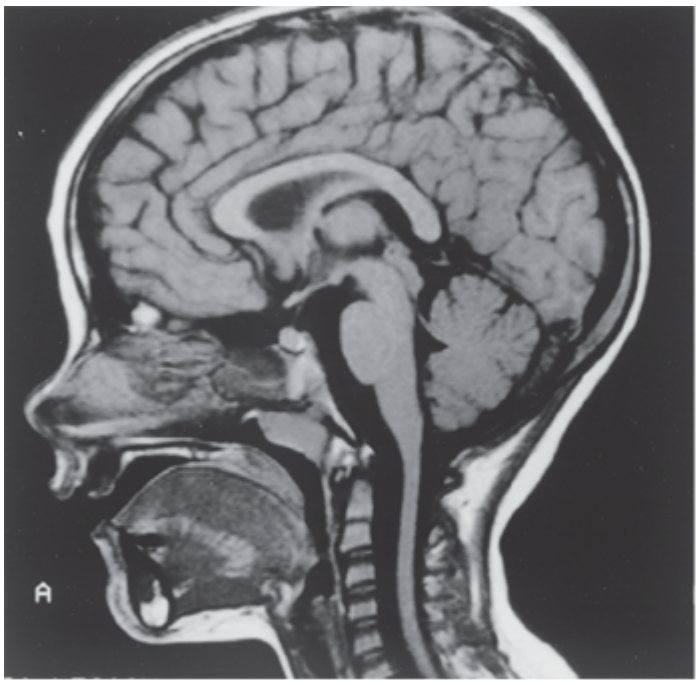
Friedreich's ataxia presents before 25 years of age with progressive staggering gait, frequent falling, and titubation. The lower extremities are more severely involved than the upper ones. Dysarthria occasionally is the presenting symptom; rarely, progressive scoliosis, foot deformity, nystagmus, or cardiopathy is the initial sign.

The neurologic examination reveals nystagmus, loss of fast saccadic eye movements, truncal titubation, dysarthria, dysmetria, and ataxia of trunk and limb movements. Extensor plantar responses (with normal tone in trunk and extremities), absence of deep tendon reflexes, and weakness (greater distally than proximally) are usually found. Loss of vibratory and proprioceptive sensation occurs. The median age of death is 35 years. Women have a significantly better prognosis than men.

Cardiac involvement occurs in 90% of patients. Cardiomegaly, symmetric hypertrophy, murmurs, and conduction defects are reported. Moderate mental retardation or psychiatric syndromes are present in a small percentage of patients. A high incidence of diabetes mellitus (20%) is found and is associated with insulin resistance and pancreatic β-cell dysfunction. Musculoskeletal deformities are common and include pes cavus, pes equinovarus, and scoliosis. MRI of the spinal cord shows atrophy **(Fig. 431-2)**.

FIGURE 431-2

Sagittal magnetic resonance imaging (MRI) of the brain and spinal cord of a patient with Friedreich's ataxia, demonstrating spinal cord atrophy. *(Reproduced with permission from RN Rosenberg, P Khemani, in RN Rosenberg, JM Pascual [eds]: Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease, 5th ed. London, Elsevier, 2015.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The primary sites of pathology are the spinal cord, dorsal root ganglion cells, and the peripheral nerves. Slight atrophy of the cerebellum and cerebral gyri may occur. Sclerosis and degeneration occur predominantly in the spinocerebellar tracts, lateral corticospinal tracts, and posterior columns. Degeneration of the glossopharyngeal, vagus, hypoglossal, and deep cerebellar nuclei is described. The cerebral cortex is histologically normal except for loss of Betz cells in the precentral gyri. The peripheral nerves are extensively involved, with a loss of large myelinated fibers. Cardiac pathology consists of myocytic hypertrophy and fibrosis, focal vascular fibromuscular dysplasia with subintimal or medial deposition of periodic acid-Schiff (PAS)-positive material, myocytopathy with unusual pleomorphic nuclei, and focal degeneration of nerves and cardiac ganglia.

GENETIC CONSIDERATIONS



The classic form of Friedreich's ataxia has been mapped to 9q13-q21.1, and the mutant gene, *frataxin*, contains expanded GAA triplet repeats in the first intron. There is homozygosity for expanded GAA repeats in >95% of patients. Normal persons have 7–22 GAA repeats, and patients have 200–900 GAA repeats. A more varied clinical syndrome has been described in compound heterozygotes who have one copy of the GAA expansion and the other copy a point mutation in the *frataxin* gene. When the point mutation is located in the region of the gene that encodes the amino-terminal half of frataxin, the phenotype is milder, often consisting of a spastic gait, retained or exaggerated reflexes, no dysarthria, and mild or absent ataxia.

Patients with Friedreich's ataxia have undetectable or extremely low levels of *frataxin* mRNA, as compared with carriers and unrelated individuals; thus, disease appears to be caused by a loss of expression of the frataxin protein. Frataxin is a mitochondrial protein involved in iron homeostasis. Mitochondrial iron accumulation due to loss of the iron transporter coded by the mutant *frataxin* gene results in oxidized

intramitochondrial iron. Excess oxidized iron results in turn in the oxidation of cellular components and irreversible cell injury.

Two forms of hereditary ataxia associated with abnormalities in the interactions of vitamin E (α-tocopherol) with very-low-density lipoprotein (VLDL) have been delineated. These are abetalipoproteinemia (Bassen-Kornzweig syndrome) and ataxia with vitamin E deficiency (AVED). Abetalipoproteinemia is caused by mutations in the gene coding for the larger subunit of the microsomal triglyceride transfer protein (MTP). Defects in MTP result in impairment of formation and secretion of VLDL in liver. This defect results in a deficiency of delivery of vitamin E to tissues, including the central and peripheral nervous system, as VLDL is the transport molecule for vitamin E and other fat-soluble substitutes. AVED is due to mutations in the gene for α-tocopherol transfer protein (α-TTP). These patients have an impaired ability to bind vitamin E into the VLDL produced and secreted by the liver, resulting in a deficiency of vitamin E in peripheral tissues. Hence, either absence of VLDL (abetalipoproteinemia) or impaired binding of vitamin E to VLDL (AVED) causes an ataxic syndrome. Once again, a genotype classification has proved to be essential in sorting out the various forms of the Friedreich's disease syndrome, which may be clinically indistinguishable.

Ataxia Telangiectasia

SYMPTOMS AND SIGNS

Patients with ataxia telangiectasia (AT) present in the first decade of life with progressive telangiectatic lesions associated with deficits in cerebellar function and nystagmus. The neurologic manifestations correspond to those in Friedreich's disease, which should be included in the differential diagnosis. Truncal and limb ataxia, dysarthria, extensor plantar responses, myoclonic jerks, areflexia, and distal sensory deficits may develop. There is a high incidence of recurrent pulmonary infections and neoplasms of the lymphatic and reticuloendothelial system in patients with AT. Thymic hypoplasia with cellular and humoral (IgA and IgG2) immunodeficiencies, premature aging, and endocrine disorders such as type 1 diabetes mellitus are described. There is an increased incidence of lymphomas, Hodgkin's disease, acute T cell leukemias, and breast cancer.

The most striking neuropathologic changes include loss of Purkinje, granule, and basket cells in the cerebellar cortex as well as of neurons in the deep cerebellar nuclei. The inferior olives of the medulla may also have neuronal loss. There is a loss of anterior horn neurons in the spinal cord and of dorsal root ganglion cells associated with posterior column spinal cord demyelination. A poorly developed or absent thymus gland is the most consistent defect of the lymphoid system.

GENETIC CONSIDERATIONS



The gene for AT (the *ATM* gene) at 11q22-23 encodes a protein that is similar to several yeast and mammalian phosphatidylinositol-3′ kinases involved in mitogenic signal transduction, meiotic recombination, and cell cycle control. Defective DNA repair in AT fibroblasts exposed to ultraviolet light has been demonstrated. The discovery of *ATM* permits early diagnosis and identification of heterozygotes who

are at risk for cancer (e.g., breast cancer). Elevated serum alpha-fetoprotein and immunoglobulin deficiency are noted.

MITOCHONDRIAL ATAXIAS

Spinocerebellar syndromes have been identified with mutations in mitochondrial DNA (mtDNA). Thirty pathogenic mtDNA point mutations and 60 different types of mtDNA deletions are known, several of which cause or are associated with ataxia **(Chap. 441)**.

TREATMENT

TREATMENT

Ataxic Disorders

The most important goal in management of patients with ataxia is to identify treatable disease entities. Mass lesions must be recognized promptly and treated appropriately. Paraneoplastic disorders can often be identified by the clinical patterns of disease that they produce, measurement of specific autoantibodies, and uncovering the primary cancer; these disorders are often refractory to therapy, but some patients improve following removal of the tumor or immunotherapy **(Chap. 90)**. Ataxia with antigliadin antibodies and gluten-sensitive enteropathy may improve with a gluten-free diet. Malabsorption syndromes leading to vitamin E deficiency may lead to ataxia. The vitamin E deficiency form of Friedreich's ataxia must be considered, and serum vitamin E levels measured. Vitamin E therapy is indicated for these rare patients. Vitamin B_1 and B_{12} levels in serum should be measured, and the vitamins administered to patients having deficient levels. Hypothyroidism is easily treated. The cerebrospinal fluid should be tested for a syphilitic infection in patients with progressive ataxia and other features of tabes dorsalis. Similarly, antibody titers for Lyme disease and *Legionella* should be measured and appropriate antibiotic therapy should be instituted in antibody-positive patients. Aminoacidopathies, leukodystrophies, urea-cycle abnormalities, and mitochondrial encephalomyopathies may produce ataxia, and some dietary or metabolic therapies are available for these disorders. The deleterious effects of phenytoin and alcohol on the cerebellum are well known, and these exposures should be avoided in patients with ataxia of any cause.

There is no proven therapy for any of the autosomal dominant ataxias (SCA1 to SCA40). There is preliminary evidence that idebenone, a free-radical scavenger, can improve myocardial hypertrophy in patients with classic Friedreich's ataxia; there is no current evidence, however, that it improves neurologic function. A small preliminary study in a mixed population of patients with different inherited ataxias raised the possibility that the glutamate antagonist riluzole may offer modest benefit. Iron chelators and antioxidant drugs are potentially harmful in Friedreich's patients because they may increase heart muscle injury. Acetazolamide can reduce the duration of symptoms of EA. At present, identification of an at-risk person's genotype, together with appropriate family and genetic counseling, can reduce the incidence of these cerebellar syndromes in future generations **(Chap. 457)**.

GENETIC DIAGNOSTIC LABORATORIES

1. Baylor College of Medicine; Houston, Texas, 1-713-798-6522

   http://www.bcm.edu/genetics/index.cfm?pmid=21387

2. GeneDx

   http://www.genedx.com

3. Transgenomic, 1-877-274-9432

   http://www.transgenomic.com/labs/neurology

GLOBAL FEATURES

 Ataxias with autosomal dominant, autosomal recessive, X-linked, or mitochondrial forms of inheritance are present on a worldwide basis. Machado-Joseph disease (SCA3) (autosomal dominant) and Friedreich's ataxia (autosomal recessive) are the most common types in most populations. Genetic markers are now commercially available to precisely identify the genetic mutation for correct diagnosis and also for family planning. Early detection of asymptomatic preclinical disease can reduce or eliminate the inherited form of ataxia in some families on a global, worldwide basis.

FURTHER READING

Anheim M, Tranchant C, Koenig M: The autosomal recessive cerebellar ataxias. N Engl J Med 16:636, 2012.

Jacobi H et al: Long-term disease progression in spinocerebellar ataxia types 1, 2, 3, and 6: A longitudinal cohort study. Lancet Neurol 14:1101, 2015. [PubMed: 26377379]

Romano S et al: Riluzole in patients with hereditary cerebellar ataxia: A randomised, double-blind, placebo-controlled trial. Lancet Neurol 14:985, 2015. [PubMed: 26321318]

Chapter 432: Disorders of the Autonomic Nervous System

Christopher H. Gibbons; John W. Engstrom

INTRODUCTION

The autonomic nervous system (ANS) innervates the entire neuraxis and influences all organ systems. It regulates blood pressure (BP), heart rate, sleep, glandular, pupillary, bladder and bowel function. It maintains organ homeostasis and operates automatically; its full importance becomes recognized only when ANS function is compromised, resulting in dysautonomia. **Hypothalamic disorders that cause disturbances in homeostasis are discussed in Chaps. 15 and 371.**

ANATOMIC ORGANIZATION

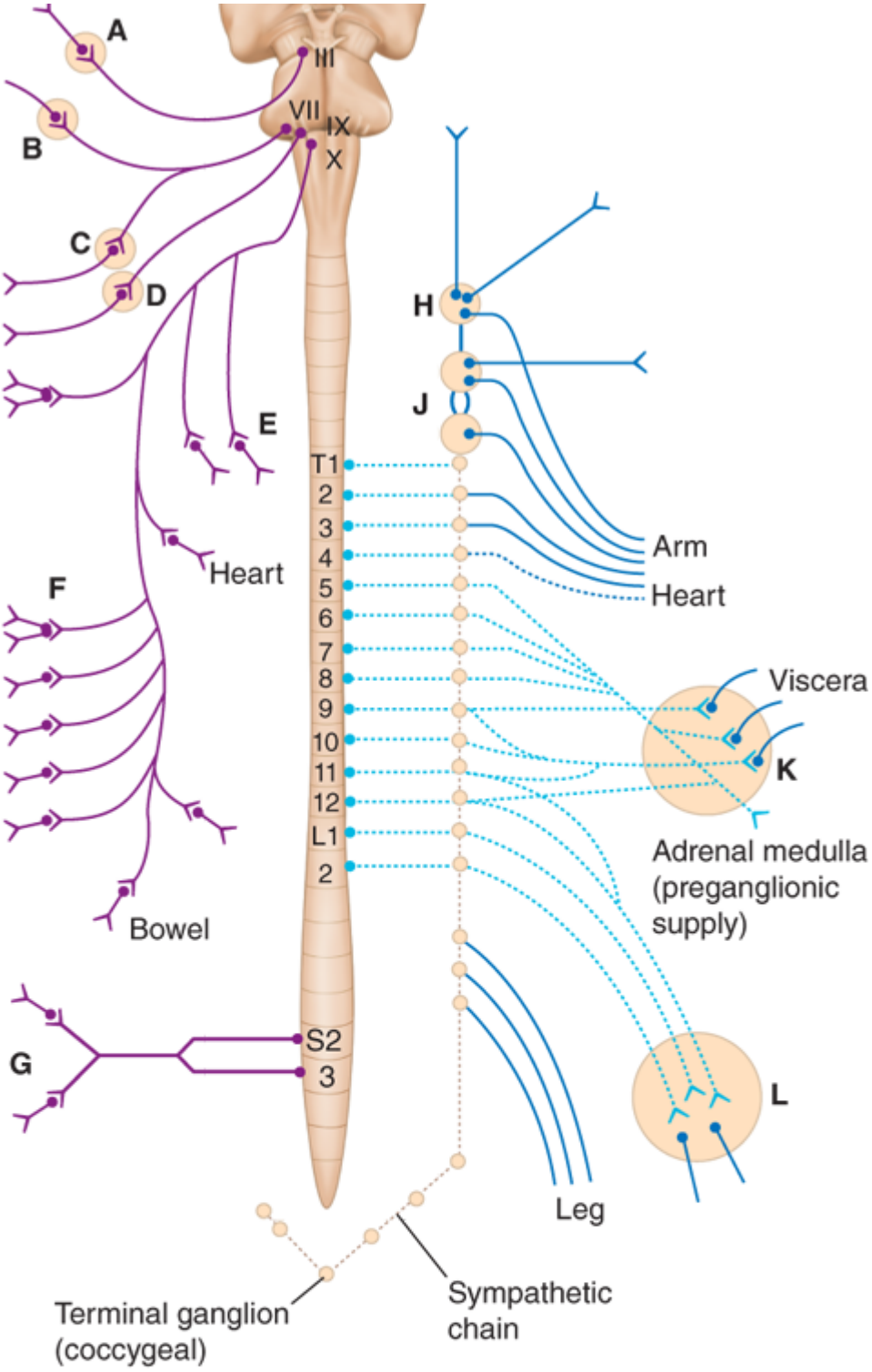
The activity of the ANS is regulated by central neurons responsive to diverse afferent inputs. After central integration of afferent information, autonomic outflow is adjusted to permit the functioning of the major organ systems in accordance with the needs of the whole organism. Connections between the cerebral cortex and the autonomic centers in the brainstem coordinate autonomic outflow with higher mental functions.

The preganglionic neurons of the parasympathetic nervous system leave the central nervous system (CNS) in the third, seventh, ninth, and tenth cranial nerves as well as the second and third sacral nerves, while the preganglionic neurons of the sympathetic nervous system exit the spinal cord between the first thoracic and the second lumbar segments **(Fig. 432-1)**. The autonomic preganglionic fibers are thinly myelinated. The postganglionic neurons, located in ganglia outside the CNS, give rise to the postganglionic unmyelinated autonomic nerves that innervate organs and tissues throughout the body. Responses to sympathetic and parasympathetic stimulation are frequently antagonistic **(Table 432-1)**, reflecting highly coordinated interactions within the CNS; the resultant changes in parasympathetic and sympathetic activity provide more precise control of autonomic responses than could be achieved by the modulation of a single system.

FIGURE 432-1

Schematic representation of the autonomic nervous system. *(From M Moskowitz: Clin Endocrinol Metab 6:77, 1977.)*





Parasympathetic system
from cranial nerves III, VII, IX, X and from sacral nerves 2 and 3

Sympathetic system
from T1-L2
Preganglionic fibers
Postganglionic fibers

A Ciliary ganglion
B Sphenopalatine (pterygopalatine) ganglion
C Submandibular ganglion
D Otic ganglion

H Superior cervical ganglion
J Middle cervical ganglion and inferior cervical (stellate) ganglion including T1 ganglion

E Vagal ganglion cells in the heart wall
F Vagal ganglion cells in bowel wall
G Pelvic ganglia
K Coeliac and other abdominal ganglia
L Lower abdominal sympathetic ganglia

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 432-1

Functional Consequences of Normal ANS Activation

	Sympathetic	Parasympathetic
Heart rate	Increased	Decreased
Blood pressure	Increased	Mildly decreased
Bladder	Increased sphincter tone	Voiding (decreased tone)
Bowel motility	Decreased motility	Increased
Lung	Bronchodilation	Bronchoconstriction
Sweat glands	Sweating	—
Pupils	Dilation	Constriction
Adrenal glands	Catecholamine release	—
Sexual function	Ejaculation, orgasm	Erection
Lacrimal glands	—	Tearing
Parotid glands	—	Salivation

Acetylcholine (ACh) is the preganglionic neurotransmitter for both the sympathetic and parasympathetic divisions of the ANS as well as the postganglionic neurotransmitter of the parasympathetic neurons; the preganglionic receptors are nicotinic, and the postganglionic are muscarinic in type. Norepinephrine (NE) is the neurotransmitter of the postganglionic sympathetic neurons, except for cholinergic neurons innervating the eccrine sweat glands.

The gastrointestinal (GI) tract has long been described as part of the sympathetic and parasympathetic nervous systems. However, it is has many unique characteristics such that it is now considered separately as the enteric, or intrinsic, nervous system. Parasympathetic control of the GI system is through the craniospinal nerves (vagus and S2-S4 nerves) while sympathetic control is through the thoracolumbar region. The enteric nervous system itself is made up of a series of ganglia that form a network of plexuses with several hundred million cells (the equivalent of the number of cells in the spinal cord). Meissner's (submucosal) plexus, Auerbach's (myenteric), Cajal's (deep muscular), mucosal and submucosal plexuses comprise the majority of nerves within the enteric nervous system. Numerous neurotransmitters have now been identified within the enteric nervous system, with many neurons containing both primary and co-transmitter neurotransmitters.

CLINICAL EVALUATION

CLASSIFICATION

Disorders of the ANS may result from pathology of either the CNS or the peripheral nervous system (PNS) (Table 432-2). Signs and symptoms may result from interruption of the afferent limb, CNS processing centers, or efferent limb of reflex arcs controlling autonomic responses. For example, a lesion of the medulla produced by a posterior fossa tumor can impair BP responses to postural changes and result in orthostatic hypotension (OH). OH can also be caused by lesions of the afferent limb of the baroreflex arc (e.g., radiation or congenital disease), spinal cord or peripheral vasomotor nerve fibers (e.g., diabetic autonomic neuropathy). Lesions of the efferent limb cause the most consistent and severe OH. The site of reflex interruption is usually established by the clinical context in which the dysautonomia arises, combined with judicious use of ANS testing and neuroimaging studies. The presence or absence of CNS signs, association with sensory or motor polyneuropathy, medical illnesses, medication use, and family history are often important considerations. Some syndromes do not fit easily into any classification scheme.

TABLE 432-2

Classification of Clinical Autonomic Disorders

I. Autonomic Disorders with Brain Involvement

A. Associated with multisystem degeneration
1. Multisystem degeneration: autonomic failure clinically prominent
   a. Multiple system atrophy (MSA)
   b. Parkinson's disease with autonomic failure
   c. Diffuse Lewy body disease with autonomic failure
2. Multisystem degeneration: autonomic failure clinically not usually prominent
   a. Parkinson's disease without autonomic failure
   b. Other extrapyramidal disorders (inherited spinocerebellar atrophies, progressive supranuclear palsy, corticobasal degeneration, Machado-Joseph disease, fragile X syndrome [FXTAS])

B. Unassociated with multisystem degeneration (focal CNS disorders)
1. Disorders mainly due to cerebral cortex involvement
   a. Frontal cortex lesions causing urinary/bowel incontinence
   b. Focal seizures (temporal lobe or anterior cingulate)
   c. Cerebral infarction of the insula
2. Disorders of the limbic and paralimbic circuits
   a. Shapiro's syndrome (agenesis of corpus callosum, hyperhidrosis, hypothermia)
   b. Autonomic seizures
   c. Limbic encephalitis
3. Disorders of the hypothalamus
   a. Thiamine deficiency (Wernicke-Korsakoff syndrome)
   b. Diencephalic syndrome
   c. Neuroleptic malignant syndrome
   d. Serotonin syndrome
   e. Fatal familial insomnia
   f. Antidiuretic hormone (ADH) syndromes (diabetes insipidus, inappropriate ADH secretion)
   g. Disturbances of temperature regulation (hyperthermia, hypothermia)
   h. Disturbances of sexual function
   i. Disturbances of appetite
   j. Disturbances of BP/HR and gastric function
   k. Horner's syndrome
4. Disorders of the brainstem and cerebellum
   a. Posterior fossa tumors
   b. Syringobulbia and Arnold-Chiari malformation
   c. Disorders of BP control (hypertension, hypotension)
   d. Cardiac arrhythmias

e. Central sleep apnea
f. Baroreflex failure
g. Horner's syndrome
h. Vertebrobasilar and lateral medullary (Wallenberg's) syndromes
i. Brainstem encephalitis

II. Autonomic Disorders with Spinal Cord Involvement

A. Traumatic quadriplegia
B. Syringomyelia
C. Subacute combined degeneration
D. Multiple sclerosis and neuromyelitis optica
E. Amyotrophic lateral sclerosis
F. Tetanus
G. Stiff-person syndrome
H. Spinal cord tumors

III. Autonomic Neuropathies

A. Acute/subacute autonomic neuropathies
   a. Subacute autoimmune autonomic ganglionopathy (AAG)
   b. Subacute paraneoplastic autonomic neuropathy
   c. Guillain-Barré syndrome
   d. Botulism
   e. Porphyria
   f. Drug induced autonomic neuropathies-stimulants, drug withdrawal, vasoconstrictor, vasodilators, beta-receptor antagonists, beta-agonists
   g. Toxin-induced autonomic neuropathies
   h. Subacute cholinergic neuropathy
B. Chronic peripheral autonomic neuropathies
   1. Distal small fiber neuropathy
   2. Combined sympathetic and parasympathetic failure
      a. Amyloid
      b. Diabetic autonomic neuropathy
      c. AAG (paraneoplastic and idiopathic)
      d. Sensory neuronopathy with autonomic failure
      e. Familial dysautonomia (Riley-Day syndrome)
      f. Diabetic, uremic, or nutritional deficiency
      g. Geriatric dysautonomia (age >80 years)

3. Disorders of orthostatic intolerance: reflex syncope; POTS; prolonged bed rest; space flight; chronic fatigue

Abbreviations: BP, blood pressure; CNS, central nervous system; HR, heart rate; POTS, postural orthostatic tachycardia syndrome.

SYMPTOMS OF AUTONOMIC DYSFUNCTION

Clinical manifestations can result from loss of function, overactivity, or dysregulation of autonomic circuits. Disorders of autonomic function should be considered in patients with unexplained OH, syncope, sleep dysfunction, altered sweating (hyperhidrosis or hypohidrosis), impotence, constipation or other GI symptoms (bloating, nausea, vomiting of old food, diarrhea), or bladder disorders (urinary frequency, hesitancy, or incontinence). Symptoms may be widespread or regional in distribution. An autonomic history focuses on systemic functions (orthostatic symptoms, BP, heart rate, sleep, fever, sweating) and involvement of individual organ systems (pupils, bowel, bladder, sexual function). The autonomic symptom profile is a self-report questionnaire that can be used for formal assessment. A recent consensus guideline has provided some useful screening questions for detection of OH, and there is a recently developed OH questionnaire for more detailed symptom assessment. It is important to recognize the modulating effects of age and duration of disease. For example, OH typically produces lightheadedness when of acute onset, but may present with subtle cognitive manifestations in chronic disease. Specific symptoms of orthostatic intolerance are diverse (Table 432-3). Autonomic symptoms may vary dramatically, reflecting the dynamic nature of autonomic control over homeostatic function. For example, OH might be manifest only in the early morning, following a meal, with exercise, or with raised ambient temperature, depending on the regional vascular bed affected by the dysautonomia.

TABLE 432-3

Symptoms of Orthostatic Intolerance

Lightheadedness (dizziness)	88%
Weakness or tiredness	72%
Cognitive difficulty (thinking/concentrating)	47%
Blurred vision	47%
Tremulousness	38%
Vertigo	37%
Pallor	31%
Anxiety	29%
Palpitations	26%
Clammy feeling	19%
Nausea	18%

Source: PA Low et al: Mayo Clin Proc 70:617, 1995.

Early autonomic symptoms may be overlooked. Impotence, although not specific for autonomic failure, often heralds autonomic failure in men and may precede other symptoms by years (Chap. 390). A decrease in the frequency of spontaneous early morning erections may occur months before loss of nocturnal penile tumescence and development of total impotence. Bladder dysfunction may appear early in men and women, particularly in those with a CNS etiology. Cold feet may indicate increased peripheral vasomotor constriction, although this symptom is a very common complaint among healthy individuals as well. Brain and spinal cord disease above the level of the lumbar spine results first in urinary frequency and small bladder volumes and eventually in incontinence (upper motor neuron or spastic bladder). By contrast, PNS disease of autonomic nerve fibers results in large bladder volumes, urinary frequency, and overflow incontinence (lower motor neuron flaccid bladder). Measurements of bladder volume (postvoid residual), or urodynamic studies, are useful tests for distinguishing between upper and lower motor neuron bladder dysfunction in the early stages of dysautonomia. GI autonomic dysfunction typically presents as progressively severe constipation. Diarrhea may develop (typically in diabetes mellitus) due to many reasons including rapid transit of contents, uncoordinated small-bowel motor activity, an osmotic basis from bacterial overgrowth associated with

small-bowel stasis, and anorectal dysfunction with diminished sphincter control and increased intestinal secretion. Impaired glandular secretory function may cause difficulty with food intake due to decreased salivation or eye irritation due to decreased lacrimation. Loss of sweat function (anhidrosis), a critical element of thermoregulation, may result in hyperthermia. Patients with a length-dependent neuropathy may present with distal anhidrosis but the primary complaint is proximal hyperhidrosis that occurs to maintain thermoregulation **(Chap. 15)**. Lack of sweating after a hot bath, during exercise, or on a hot day can suggest sudomotor failure.

OH (also called *orthostatic or postural hypotension*) is perhaps the most disabling feature of autonomic dysfunction. There are numerous causes of OH (e.g., medications, anemia, dehydration or volume depletion), but when the OH is specifically due to dysfunction of the ANS it is referred to as neurogenic OH. The prevalence of OH is relatively high, especially when OH associated with aging and diabetes mellitus is included **(Table 432-4)**. OH can cause a variety of symptoms, including dimming or loss of vision, lightheadedness, diaphoresis, diminished hearing, pallor, weakness, and shortness of breath. Syncope results when the drop in BP impairs cerebral perfusion. Other manifestations of impaired baroreflexes are supine hypertension, a heart rate that is fixed regardless of posture, postprandial hypotension, and an excessively high nocturnal BP. Many patients with OH have a preexisting diagnosis of hypertension or have concomitant supine hypertension, reflecting the great importance of baroreflexes in maintaining postural and supine normotension. The appearance of OH in patients receiving antihypertensive treatment may indicate overtreatment or the onset of an autonomic disorder. The most common causes of OH are not neurologic in origin **(Table 432-5)**; these must be distinguished from the neurogenic causes. The mortality rates of nonneurogenic OH are similar to that of the general population while neurogenic OH carries a three- to sevenfold higher mortality rate. **Neurocardiogenic and cardiac causes of syncope are considered in Chap. 18**.

TABLE 432-4

Prevalence of Orthostatic Hypotension in Different Situations

Disorder	Prevalence
Aging	14–20%
Diabetic neuropathy	10%
Other autonomic neuropathies	>60%
Multiple system atrophy	>90%
Pure autonomic failure	>95%

TABLE 432-5

Nonneurogenic Causes of Orthostatic Hypotension

Cardiac Pump Failure
Myocardial infarction Myocarditis Constrictive pericarditis Aortic stenosis Tachyarrhythmias Bradyarrhythmias Salt-losing nephropathy Adrenal insufficiency Diabetes insipidus Venous obstruction
Reduced Intravascular Volume
Straining or heavy lifting, urination, defecation Dehydration Diarrhea, emesis Hemorrhage Burns
Metabolic
Adrenocortical insufficiency Hypoaldosteronism Pheochromocytoma Severe potassium depletion
Venous Pooling
Alcohol Postprandial dilation of splanchnic vessel beds Vigorous exercise with dilation of skeletal vessel beds Heat: hot environment, hot showers and baths, fever Prolonged recumbency or standing Sepsis
Medications

Antihypertensives
Diuretics
Vasodilators: nitrates, hydralazine
Alpha- and beta-blocking agents
Central nervous system sedatives: barbiturates, opiates
Tricyclic antidepressants
Phenothiazines

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Orthostatic Hypotension and Other ANS Disorders

The first step in the evaluation of symptomatic OH is the exclusion of treatable causes. The history should include a review of medications that may affect the ANS **(Table 432-6)**. The main classes of drugs that may cause OH are diuretics, antihypertensive agents (preload reducers, vasodilators, negative inotropic or chronotropic agents), antidepressants (tricyclic antidepressants and SSRIs), ethanol, opioids, insulin, dopamine agonists, and barbiturates. However, the precipitation of OH by medications may also be the first sign of an underlying autonomic disorder. The history may reveal an underlying cause for symptoms (e.g., diabetes, Parkinson's disease) or specific underlying mechanisms (e.g., cardiac pump failure, reduced intravascular volume). The relationship of symptoms to meals (splanchnic pooling), standing on awakening in the morning (intravascular volume depletion), ambient warming (vasodilatation), or exercise (muscle arteriolar vasodilatation) should be sought. Standing time to first symptom and to presyncope **(Chap. 18)** should be followed for management.

Physical examination includes measurement of supine and standing pulse and BP. OH is defined as a sustained drop in systolic (≥20 mmHg) or diastolic (≥10 mmHg) BP after 3 min of standing. In non-neurogenic causes of OH (such as hypovolemia), the BP drop is accompanied by a compensatory increase in heart rate of >15 beats/min. A clue that the patient has neurogenic OH is the aggravation or precipitation of OH by autonomic stressors (a meal, hot bath, or exercise). Neurologic examination should include mental status (neurodegenerative disorders such as Lewy body dementia can be accompanied by significant dysautonomia), cranial nerves (abnormal pupils with Horner's or Adie's syndrome), motor tone (parkinsonian syndromes), motor strength and sensation (polyneuropathies). In patients without a clear diagnosis initially, follow-up evaluations every few months or whenever symptoms worsen may reveal the underlying cause.

Disorders of autonomic function should be considered in patients with symptoms of altered sweating (hyperhidrosis or hypohidrosis), gastroparesis (bloating, nausea, vomiting of old food), impotence, constipation, or bladder disturbances (urinary frequency, hesitancy, or incontinence).

AUTONOMIC TESTING

Autonomic function tests are helpful to document and localize abnormalities when findings on history and examination are inconclusive; to detect subclinical involvement; or to follow the course of an autonomic disorder.

Heart Rate Variation With Deep Breathing

This tests the parasympathetic component of cardiovascular reflexes via the vagus nerve. Results are influenced by multiple factors including the subject's position (recumbent, sitting, or standing), rate and depth of respiration (6 breaths per minute and a forced vital capacity [FVC] >1.5 L are optimal), age, medications, weight, and degree of hypocapnia. Interpretation of results requires comparison of test data with results from age-matched controls collected under identical test conditions. For example, the lower limit of normal heart rate variation with deep breathing in persons <20 years is >15–20 beats/min, but for persons aged >60 it is 5–8 beats/min. Heart rate variation with deep breathing (respiratory sinus arrhythmia) is abolished by the muscarinic ACh receptor antagonist atropine but is unaffected by sympathetic postganglionic blockade (e.g., propranolol).

Valsalva Response

This response **(Table 432-7)** assesses the integrity of the baroreflex control of heart rate (parasympathetic) and BP (sympathetic adrenergic). Under normal conditions, increases in BP at the carotid bulb trigger a reduction in heart rate (increased vagal tone), and decreases in BP trigger an increase in heart rate (reduced vagal tone). The Valsalva response is tested in the supine position. The subject exhales against a closed glottis (or into a manometer maintaining a constant expiratory pressure of 40 mmHg) for 15 s while measuring changes in heart rate and beat-to-beat BP. Without directly measuring expiratory pressure, heart rate and beat-to-beat blood pressure the Valsalva maneuver cannot be interpreted correctly. There are four phases of the BP and heart rate response to the Valsalva maneuver. Phases I and III are mechanical and related to changes in intrathoracic and intraabdominal pressure. In early phase II, reduced venous return results in a fall in stroke volume and BP, counteracted by a combination of reflex tachycardia and increased total peripheral resistance. Increased total peripheral resistance arrests the BP drop ~5–8 s after the onset of the maneuver. Late phase II begins with a progressive rise in BP toward or above baseline. Venous return and cardiac output return to normal in phase IV. Persistent peripheral arteriolar vasoconstriction and increased cardiac adrenergic tone result in a temporary BP overshoot and phase IV bradycardia (mediated by the baroreceptor reflex). Abnormalities in blood pressure during phase II recovery or phase IV overshoot suggest sympathetic adrenergic dysfunction.

Autonomic parasympathetic function during the Valsalva maneuver is measured using heart rate changes. The *Valsalva ratio* is defined as the maximum phase II tachycardia divided by the minimum phase IV bradycardia **(Table 432-8)** and is predominantly a measure of parasympathetic function.

Sudomotor Function

Sweating is induced by release of ACh from sympathetic postganglionic fibers. The quantitative sudomotor axon reflex test (QSART) is a measure of regional autonomic function mediated by ACh-induced sweating. A

reduced or absent response indicates a lesion of the postganglionic sudomotor axon. For example, sweating may be reduced in the feet as a result of distal polyneuropathy (e.g., diabetes). The thermoregulatory sweat test (TST) is a qualitative measure of global sweat production in response to an elevation of body temperature under controlled conditions. An indicator powder placed on the anterior surface of the body changes color with sweat production during temperature elevation. The pattern of color change measures the integrity of both the preganglionic and postganglionic sudomotor function. A postganglionic lesion is present if both QSART and TST show absent sweating. In a preganglionic lesion, the QSART is normal but TST shows anhidrosis.

Orthostatic Bp Recordings

Beat-to-beat BP measurements determined in supine, 70° tilt, and tilt-back positions are useful to quantitate orthostatic failure of BP control. Allow a 20-min period of rest in the supine position before assessing changes in BP during tilting. The BP change combined with heart rate monitoring is useful for the evaluation of patients with suspected OH or unexplained syncope.

Tilt Table Testing For Syncope

The great majority of patients with syncope do not have autonomic failure. Tilt table testing can be used to make the diagnosis of vasovagal syncope with sensitivity, specificity, and reproducibility. A standardized protocol is used that specifies the tilt apparatus, tilt angle, and duration of tilt. A passive phase for 30–40 min with a tilt angle at 60–70 degrees can identify reflex syncope, psychogenic syncope, or be nondiagnostic. Pharmacologic provocation of syncope (with intravenous, sublingual, or spray nitroglycerin) is controversial because it increases sensitivity at the cost of specificity. Recommendations for the performance of tilt studies for syncope have been incorporated in consensus guidelines.

TABLE 432-6

Some Drugs That Affect Autonomic Function

Symptom	Drug Class	Specific Examples
Impotence	Opioids	Tylenol #3
	Anabolic steroids	—
	Some antiarrhythmics	Prazosin
	Some antihypertensives	Clonidine
	Some diuretics	Benazepril
	Some SSRIs	Venlafaxine
Urinary retention	Opioids	Fentanyl
	Decongestants	Brompheniramine
		Diphenhydramine
Diaphoresis	Some antihypertensives	Amlodipine
	Some SSRIs	Citalopram
	Opioids	Morphine

Abbreviations: CCBs, calcium channel blockers; HCTZ, hydrochlorothiazide; SSRIs, selective serotonin reuptake inhibitors.

TABLE 432-7

Normal Blood Pressure and Heart Rate Changes During the Valsalva Maneuver

Phase	Maneuver	Blood Pressure	Heart Rate	Comments
I	Forced expiration against a partially closed glottis	Rises; aortic compression from raised intrathoracic pressure	Decreases	Mechanical
II *early*	Continued expiration	Falls; decreased venous return to the heart	Increases (reflex tachycardia)	Reduced vagal tone
II *late*	Continued expiration	Rises; reflex increase in peripheral vascular resistance	Increases at slower rate	Requires intact efferent sympathetic response
III	End of expiration	Falls; increased capacitance of pulmonary bed	Increases further	Mechanical
IV	Recovery	Rises; persistent vasoconstriction and increased cardiac output	Compensatory bradycardia	Requires intact efferent sympathetic response

TABLE 432-8

Neural Pathways Underlying Some Standardized Autonomic Tests

Test Evaluated	Procedure	Autonomic Function
HRDB	6 deep breaths/min	Cardiovagal (parasympathetic) function
Valsalva ratio	Expiratory pressure, 40 mmHg for 10–15 s	Cardiovagal (parasympathetic) function
QSART	Axon-reflex test 4 limb sites	Postganglionic (sympathetic cholinergic) sudomotor function
BP_{BB} to VM	BP_{BB} response to VM	Sympathetic adrenergic function: baroreflex adrenergic control of vagal and vasomotor function
HUT	BP_{BB} and heart rate response to HUT	Sympathetic adrenergic and cardiovagal (parasympathetic) responses to HUT

Abbreviations: BP_{BB}, beat-to-beat blood pressure; HRDB, heart rate response to deep breathing; HUT, head-up tilt; QSART, quantitative sudomotor axon reflex test; VM, Valsalva maneuver.

SPECIFIC SYNDROMES OF ANS DYSFUNCTION

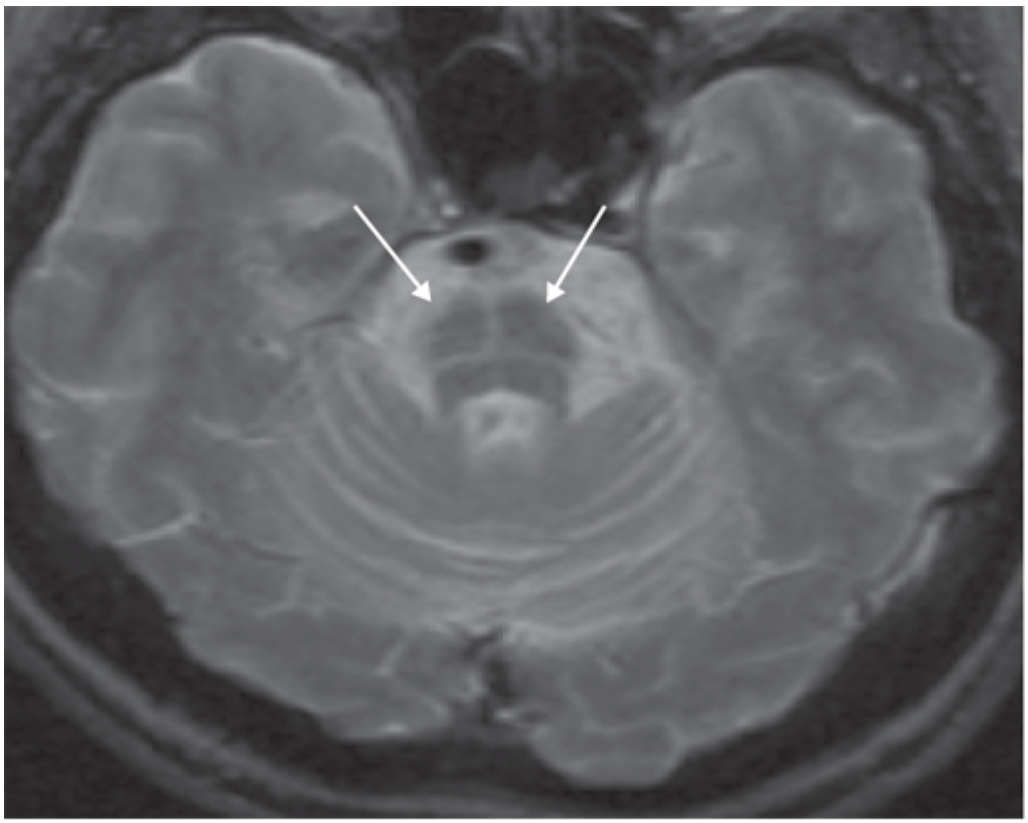
MULTIPLE SYSTEM ATROPHY

Multiple system atrophy (MSA) is an entity that comprises autonomic failure (OH or a neurogenic bladder) *and* either parkinsonism (MSA-p) or a cerebellar syndrome (MSA-c). MSA-p is the more common form; the parkinsonism is atypical in that there is more symmetric motor involvement than in Parkinson's disease (PD; **Chap. 427**), tremor is not as prominent, and there is a poor or only transient response to levodopa. Symptomatic OH within 1 year of onset of parkinsonism is suggestive of MSA-p. There is a very high frequency of impotence in men. Although autonomic abnormalities are common in advanced PD, the severity and distribution of autonomic failure are more severe and generalized in MSA. Brain magnetic resonance imaging (MRI) is a useful diagnostic adjunct: in MSA-p, iron deposition in the striatum may be evident as T2 hypointensity, and in MSA-c, cerebellar atrophy is present with a characteristic T2 hyperintense signal ("hot cross buns sign") in the pons **(Fig. 432-2)**. However, these MRI findings are typically present only with advanced disease. Cardiac postganglionic adrenergic innervation, measured by uptake of fluorodopamine on positron emission tomography, is markedly impaired in the dysautonomia of PD but is usually normal in MSA. Neuropathologic changes include neuronal loss and gliosis in many CNS regions,

including the brainstem, cerebellum, striatum, and intermediolateral cell column of the thoracolumbar spinal cord.

FIGURE 432-2

Multiple system atrophy, cerebellar type (MSA-c). Axial T2-weighted magnetic resonance image at the level of the pons shows a characteristic hyperintense signal, the “hot cross buns” sign (*arrows*). This appearance can also be seen in some spinocerebellar atrophies, as well as other neurodegenerative conditions affecting the brainstem.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MSA is uncommon, with a prevalence estimated at 2–5 per 100,000 individuals. Onset is typically in the mid-fifties, men are slightly more often affected than women, and most cases are sporadic. The diagnosis should be considered in adults aged >30 years who present with OH or urinary incontinence and either parkinsonism that is poorly responsive to dopamine replacement or a cerebellar syndrome. MSA generally progresses relentlessly to death 7–10 years after onset, but survival beyond 15 years has been reported. MSA-p is more prevalent in Western countries, while MSA-c is more common in Japan. Factors that predict a worse prognosis include early autonomic dysfunction, rapid progression of disability, bladder dysfunction, female gender, the MSA-p subtype, and an older age at onset. Attempts to slow the progression of MSA have thus far been unsuccessful, including trials of lithium, growth hormone, riluzole, rasagiline, minocycline, and rifampicin.

Management is symptomatic for neurogenic OH (see below), sleep disorders including laryngeal stridor, GI, and urinary dysfunction. GI management includes frequent small meals, soft diet, stool softeners, and bulk

agents. Gastroparesis is difficult to treat; metoclopramide stimulates gastric emptying but worsens parkinsonism by blocking central dopamine receptors. The peripheral dopamine (D_2 and D_3) receptor antagonist domperidone has been used patients with various GI conditions in many countries, and although not available in the United States, it can be obtained through the U.S. Food and Drug Administration's (FDA) Expanded Access to Investigational Drugs program.

Autonomic dysfunction is also a common feature in dementia with Lewy bodies **(Chap. 426)**; with the severity usually intermediate between that found in MSA and PD. In multiple sclerosis (MS; **Chap. 436**), autonomic complications reflect the CNS location of MS involvement and generally worsen with disease duration and disability, but are generally a secondary complaint and not of the severity seen in the synucleinopathies.

SPINAL CORD

Spinal cord lesions from any cause can result in focal autonomic deficits or autonomic hyperreflexia (e.g., spinal cord transection or hemisection) affecting bowel, bladder, sexual, temperature-regulation, or cardiovascular functions. Quadriparetic patients exhibit both supine hypertension and OH after upward tilting. *Autonomic dysreflexia* describes a dramatic increase in BP in patients with traumatic spinal cord lesions above the T6 level, often in response to irritation of the bladder, skin, or muscles. The triggers may be clinically silent because perception of painful sensations arising from structures innervated below the level of a spinal cord lesion is often blunted or absent. A distended bladder, often from an obstructed Foley catheter or a urinary infection, are common triggers of dysreflexia. Associated symptoms can include facial flushing, headache, hypertension, or piloerection. Potential complications include intracranial vasospasm or hemorrhage, cardiac arrhythmia, and death. Awareness of the syndrome, identifying the trigger, and careful monitoring of BP during procedures in patients with acute or chronic spinal cord injury are essential. In patients with supine hypertension, BP can be lowered by tilting the head upward or sitting the patient up. Vasodilator drugs may be used to treat acute elevations in BP. Clonidine can be used prophylactically to reduce the hypertension resulting from bladder stimulation. Dangerous increases or decreases in body temperature may result from an inability to sense heat or cold exposure or control peripheral vasoconstriction or sweating below the level of the spinal cord injury.

PERIPHERAL NERVE AND NEUROMUSCULAR JUNCTION DISORDERS

Peripheral neuropathies **(Chap. 438)** are the most common cause of chronic autonomic insufficiency. Polyneuropathies that affect small myelinated and unmyelinated fibers of the sympathetic and parasympathetic nerves commonly occur in diabetes mellitus, amyloidosis, chronic alcoholism, porphyria, and Guillain-Barré syndrome. Neuromuscular junction disorders with autonomic involvement include botulism and Lambert-Eaton syndrome **(Chap. 440)**.

Diabetes Mellitus

The presence of autonomic neuropathy in patients with diabetes increases the mortality rate 1.5- to 3-fold, even after adjusting for other cardiovascular risk factors. Estimates of 5-year mortality risk among these

patients range from 15 to 53%. Although many deaths are due to secondary vascular disease, there are patients who specifically suffer cardiac arrest due to autonomic neuropathy. The autonomic involvement is also predictive of other complications including renal disease, stroke, and sleep apnea. Tight glycemic control with insulin significantly reduces the long-term risk of autonomic cardiovascular neuropathy. **Diabetes mellitus is discussed in Chaps. 396, 397, 398.**

Amyloidosis

Autonomic neuropathy occurs in both sporadic and familial forms of amyloidosis. The AL (immunoglobulin light chain) type is associated with primary amyloidosis or amyloidosis secondary to multiple myeloma. The amyloid transthyretin (ATTR) type, with transthyretin as the primary protein component, is responsible for the most common form of inherited amyloidosis. Although patients usually present with a distal sensorimotor polyneuropathy accompanied by autonomic insufficiency that can precede the development of the polyneuropathy or occur in isolation. The diagnosis can be made by protein electrophoresis of blood and urine, tissue biopsy (abdominal fat pad, rectal mucosa, or sural nerve) to search for amyloid deposits, and genetic testing for transthyretin mutations in familial cases. Death is usually due to cardiac or renal involvement. Postmortem studies reveal amyloid deposition in many organs, including two sites that contribute to autonomic failure: intraneural blood vessels and autonomic ganglia. Pathologic examination reveals a loss of both unmyelinated and myelinated nerve fibers. **Clinical manifestations and treatment of the various forms of amyloidosis are discussed in detail in Chap. 108.**

Alcoholic Neuropathy

Abnormalities in parasympathetic vagal and efferent sympathetic function are usually mild in alcoholic polyneuropathy. OH is usually due to brainstem involvement, rather than injury to the PNS. Impotence is a major problem, but concurrent gonadal hormone abnormalities may play a role in this symptom. Clinical symptoms of autonomic failure generally appear only when the stocking-glove polyneuropathy is severe, and there is usually coexisting Wernicke's encephalopathy **(Chap. 301)**. Autonomic involvement may contribute to the high mortality rates associated with alcoholism **(Chap. 445)**.

Porphyria

Autonomic dysfunction is most extensively documented in acute intermittent porphyria but can also occur with variegate porphyria and hereditary coproporphyria **(Chap. 409)**. Autonomic symptoms include tachycardia, sweating, urinary retention, abdominal pain, nausea and vomiting, insomnia, hypertension, and (less commonly) hypotension. Another prominent symptom is anxiety. Abnormal autonomic function can occur both during acute attacks and during remissions. Elevated catecholamine levels during acute attacks correlate with the degree of tachycardia and hypertension that is present.

Guillain-Barré Syndrome

BP fluctuations and arrhythmias from autonomic instability can be severe (Chap. 439). It is estimated that between 2 and 10% of patients with severe Guillain-Barré syndrome suffer fatal cardiovascular collapse. GI autonomic involvement, sphincter disturbances, abnormal sweating, and pupillary dysfunction can also occur. Demyelination has been described in the vagus and glossopharyngeal nerves, the sympathetic chain, and the white rami communicantes. Interestingly, the degree of autonomic involvement appears to be independent of the severity of motor or sensory neuropathy. Acute autonomic and sensory neuropathy is a variant that spares the motor system and presents with neurogenic OH and varying degrees of sensory loss. It is treated similarly to Guillain-Barré syndrome, but prognosis is less favorable, with persistent severe sensory deficits and variable degrees of OH in many patients.

Autoimmune Autonomic Ganglionopathy (AAG)

This disorder presents with the subacute development of autonomic disturbances including OH, enteric neuropathy (gastroparesis, ileus, constipation/diarrhea), flaccid bladder, and cholinergic failure (e.g., loss of sweating, sicca complex, and a tonic pupil). A chronic form of AAG resembles pure autonomic failure (PAF) (see below). Autoantibodies against the α3 subunit of the ganglionic Ach receptor, present in approximately half of patients, are considered diagnostic of AAG. Pathology shows preferential involvement of small unmyelinated nerve fibers, with sparing of larger myelinated ones. Onset of the neuropathy follows a viral infection in approximately half of cases. Up to one-third of untreated patients experience significant functional improvement over time. Immunotherapies that have been reported to be helpful include plasmapheresis, intravenous immune globulin, glucocorticoids, azathioprine, rituximab, and mycophenolate mofetil. OH, gastroparesis, and sicca symptoms can be managed symptomatically.

AAG can also occur on a paraneoplastic basis, with adenocarcinoma or small-cell carcinoma of the lung, lymphoma, or thymoma being the most common (Chap. 90). Cerebellar involvement or dementia may coexist (see Tables 90-1, 90-2, 90-3), and the neoplasm can be occult.

Botulism

Botulinum toxin binds presynaptically to cholinergic nerve terminals and, after uptake into the cytosol, blocks ACh release. This acute cholinergic neuropathy presents as motor paralysis and autonomic disturbances that include blurred vision, dry mouth, nausea, unreactive or sluggishly reactive pupils, constipation, and urinary retention (Chap. 148).

PURE AUTONOMIC FAILURE (PAF)

This sporadic syndrome consists of postural hypotension, impotence, bladder dysfunction, and impaired sweating. The disorder begins in midlife and occurs in women more often than men. The symptoms can be disabling, but life span is unaffected. The clinical and pharmacologic characteristics suggest primary involvement of postganglionic autonomic neurons. A severe reduction in the density of neurons within sympathetic ganglia results in low supine plasma NE levels and noradrenergic supersensitivity. Some patients who are initially labeled with this diagnosis subsequently go on to develop AAG, but more often a

neurodegenerative disease supervenes, typically Lewy body dementia, PD, or MSA. In one recent series, more than one-third of patients initially diagnosed with PAF developed a CNS synucleinopathy within 4 years, and the presence of rapid eye movement sleep behavior disorder (RBD; Chap. 27) was predictive of subsequent CNS disease. Skin biopsies and autopsy studies demonstrate phosphorylated α-synuclein inclusions in postganglionic sympathetic adrenergic and cholinergic nerve fibers, distinguishing PAF from AAG and indicating that PAF is a synucleinopathy; notably, patients with PD also have α-synuclein inclusions in sympathetic nerve biopsies.

POSTURAL ORTHOSTATIC TACHYCARDIA SYNDROME (POTS)

This syndrome is characterized by symptomatic orthostatic intolerance without OH, accompanied by either an increase in heart rate to >120 beats/min or an increase of 30 beats/min with standing that subsides on sitting or lying down. Women are affected approximately five times more often than men, and most develop the syndrome between the ages of 15 and 50. Presyncopal symptoms (lightheadedness, weakness, blurred vision) combined with symptoms of autonomic overactivity (palpitations, tremulousness, nausea) are common. The pathogenesis is typically multifactorial which frequently confounds the clinical picture. A number of potential causes have been reported, including sympathetic denervation distally in the legs with preserved cardiovascular function or reduced cardiac function due to deconditioning. Hypovolemia, venous pooling, impaired brainstem baroreceptor regulation, or increased sympathetic activity may also play a role. No standardized approach to diagnosis has been established, and therapy typically has included symptomatic relief with a focus on cardiovascular rehabilitation, including a sustained exercise program. Expansion of fluid volume with water, salt, and fludrocortisone can be helpful as an initial intervention. In some patients, low-dose propranolol (20 mg) provides a modest improvement in heart rate control and exercise capacity. If these approaches are inadequate, then midodrine, pyridostigmine, or clonidine can be considered.

INHERITED DISORDERS

Five hereditary sensory and autonomic neuropathies (HSANs) exist, designated HSAN I–V. The most important autonomic variants are HSAN I and HSAN III. HSAN I is dominantly inherited and often presents as a distal small-fiber neuropathy (burning feet syndrome) associated with sensory loss and foot ulcers. The most common responsible gene, on chromosome 9q, is *SPTLC1*. SPTLC is a key enzyme in the regulation of ceramide. Cells from HSAN I patients with the mutation produce higher-than-normal levels of glucosyl ceramide, perhaps triggering apoptosis. HSAN III (Riley-Day syndrome; familial dysautonomia) is an autosomal recessive disorder of Ashkenazi Jewish children and adults and is much less prevalent than HSAN I. Decreased tearing, hyperhidrosis, reduced sensitivity to pain, areflexia, absent fungiform papillae on the tongue, and labile BP may be present. Individuals with HSAN III have afferent, but not efferent, baroreflex failure that causes the classic episodic abdominal crises and blood pressure surges in response to emotional stimuli. Pathologic examination of nerves reveals a loss of sympathetic, parasympathetic, and sensory neurons. The defective gene, *IKBKAP*, prevents normal transcription of important molecules in neural development.

PRIMARY HYPERHIDROSIS

This syndrome presents with excess sweating of the palms of the hands and soles of the feet beginning in childhood or early adulthood. The condition tends to improve with age. The disorder affects 0.6–1.0% of the population. The etiology is unclear, but there may be a genetic component because 25% of patients have a positive family history. The condition can be socially embarrassing (e.g., shaking hands) or even disabling (e.g., inability to write without soiling the paper). Topical antiperspirants are occasionally helpful. More useful are potent anticholinergic drugs such as glycopyrrolate 1–2 mg PO tid or oxybutynin 5 mg po bid. T2 ganglionectomy or sympathectomy is successful in >90% of patients with palmar hyperhidrosis. The advent of endoscopic transaxillary T2 sympathectomy has lowered the complication rate of the procedure. The most common postoperative complication is compensatory hyperhidrosis, which improves spontaneously over months. Other potential complications include recurrent hyperhidrosis (16%), Horner's syndrome (<2%), gustatory sweating, wound infection, hemothorax, and intercostal neuralgia. Local injection of botulinum toxin has also been used to block cholinergic, postganglionic sympathetic fibers to sweat glands. This approach is effective but limited by the need for repetitive injections (the effect usually lasts 4 months before waning).

ACUTE SYMPATHETIC OVERACTIVITY SYNDROMES

An *autonomic storm* is an acute state of sustained sympathetic surge that results in variable combinations of alterations in BP and heart rate, body temperature, respiration, and sweating. Causes of autonomic storm include brain and spinal cord injury, toxins and drugs, autonomic neuropathy, and chemodectomas (e.g., pheochromocytoma). Brain injury is the most common cause of autonomic storm and typically follows severe head trauma and postresuscitation anoxic-ischemic brain injury. Autonomic storm can also occur with other acute intracranial lesions such as hemorrhage, cerebral infarction, rapidly expanding tumors, subarachnoid hemorrhage, hydrocephalus, or (less commonly) an acute spinal cord lesion. The most consistent setting is that of an acute intracranial catastrophe of sufficient size and rapidity to produce a massive catecholaminergic surge. The surge can cause seizures, neurogenic pulmonary edema, and myocardial injury. Manifestations include fever, tachycardia, hypertension, tachypnea, hyperhidrosis, pupillary dilatation, and flushing. Lesions of the afferent limb of the baroreflex can result in milder recurrent autonomic storms; these can be associated with tumors or follow neck irradiation or surgery that damages the vagus and glossopharyngeal nerves.

Drugs and toxins may also be responsible, including sympathomimetics such as phenylpropanolamine, cocaine, amphetamines, and tricyclic antidepressants; tetanus; and, less often, botulinum toxin. Cocaine, including "crack," can cause a hypertensive state with CNS hyperstimulation. An overdose of tricyclic antidepressants, such as amitriptyline, can cause flushing, hypertension, tachycardia, fever, mydriasis, anhidrosis, and a toxic psychosis. The hyperadrenergic state associated with Guillain-Barré syndrome can produce a moderate autonomic storm. Pheochromocytoma presents with a paroxysmal or sustained hyperadrenergic state, headache, hyperhidrosis, palpitations, anxiety, tremulousness, and hypertension.

Neuroleptic malignant syndrome refers to a syndrome of muscle rigidity, hyperthermia, and hypertension in patients treated with neuroleptic agents (including lower potency and atypical antipsychotic agents, and even antiemetic drugs such as metoclopramide, promethazine) **(Chap. 429)**. Management of autonomic storm includes ruling out other causes of autonomic instability, including malignant hyperthermia, porphyria, and seizures. Sepsis and encephalitis need to be excluded with appropriate studies. An electroencephalogram (EEG) should be done to search for seizure activity; MRI of the brain and spine is often necessary. The patient should be managed in an intensive care unit. Management with morphine sulphate (10 mg every 4 h) and labetalol (100–200 mg twice daily) may be helpful. Supportive treatment may need to be maintained for several weeks. For chronic and milder autonomic storm, propranolol and/or clonidine can be effective.

MISCELLANEOUS

Other conditions associated with autonomic failure include infections, malignancy, and poisoning (organophosphates). Disorders of the hypothalamus can affect autonomic function and produce abnormalities in temperature control, satiety, sexual function, and circadian rhythms **(Chap. 373)**.

COMPLEX REGIONAL PAIN SYNDROMES (CRPS)

The failure to identify a primary role of the ANS in the pathogenesis of these disorders has resulted in a change of nomenclature. The terms CRPS types I and II are now used in place of reflex sympathetic dystrophy (RSD) and causalgia.

CRPS type I is a regional pain syndrome that often develops after tissue injury and most commonly affects one limb. Examples of associated injury include minor shoulder or limb trauma, fractures, myocardial infarction, or stroke. *Allodynia* (the perception of a nonpainful stimulus as painful), *hyperpathia* (an exaggerated pain response to a painful stimulus), and spontaneous pain occur. The symptoms are unrelated to the severity of the initial trauma and are not confined to the distribution of a single peripheral nerve. CRPS type II is a regional pain syndrome that develops after injury to a specific peripheral nerve, often a major nerve trunk. Spontaneous pain initially develops within the territory of the affected nerve but eventually may spread outside the nerve distribution. Although CRPS type I (RSD) has been classically divided into three clinical phases, there is little evidence that CRPS “progresses” from one stage to another. Currently, the Budapest consensus criteria for clinical diagnosis of CRPS delete staging and require at least three symptoms and two signs in the following four categories: (1) sensory, (2) vasomotor, (3) sudomotor/edema, and (4) motor/trophic. Pain (usually burning or electrical in quality) is the primary clinical feature of CRPS. Limb pain syndromes that do not meet these criteria are best classified as “limb pain—not otherwise specified.” In CRPS, localized sweating (increased resting sweat output) and changes in blood flow may produce temperature differences between affected and unaffected limbs.

The natural history of typical CRPS may be more benign and more variable than previously recognized. A variety of surgical and medical treatments have been developed, with conflicting reports of efficacy. Clinical trials suggest that early mobilization with physical therapy or a brief course of glucocorticoids may be helpful

for early CRPS type I or II. Chronic glucocorticoid treatment is not recommended. Current treatment paradigms are multidisciplinary with a focus on early mobilization, physical therapy, pain management, patient education, and psychological support.

TREATMENT

TREATMENT

Autonomic Failure

Management of autonomic failure is aimed at specific treatment of the cause and alleviation of symptoms. Of particular importance is the removal of drugs or amelioration of underlying conditions that cause or aggravate the autonomic symptoms, especially in the elderly. For example, OH can be caused or aggravated by antihypertensive agents, antidepressants, levodopa or dopaminergic agonists, ethanol, opioids, insulin, and barbiturates. A summary of drugs that can cause impotence, urinary retention, or diaphoresis by class and putative mechanism is shown in Table 432-6.

PATIENT EDUCATION

Only a minority of patients with OH require drug treatment. All patients should be taught the mechanisms of postural normotension (volume status, resistance and capacitance bed, autoregulation) and the nature of orthostatic stressors (time of day and the influence of meals, heat, standing, and exercise). Patients should learn to recognize orthostatic symptoms early (especially subtle cognitive symptoms, weakness, and fatigue) and to modify or avoid activities that provoke episodes. Other helpful measures may include keeping a BP log and dietary education (salt/fluids). Learning physical counter-maneuvers that reduce standing OH and practicing postural and resistance training and cardiovascular reconditioning are helpful measures.

SYMPTOMATIC TREATMENT

Nonpharmacologic approaches are summarized in **Table 432-9**. Adequate intake of salt and fluids to produce a voiding volume between 1.5 and 2.5 L of urine (containing >170 meq/L of Na^+) each 24 h is essential. Sleeping with the head of the bed elevated will minimize the effects of supine nocturnal hypertension. Prolonged recumbency should be avoided when possible. Patients are advised to sit with legs dangling over the edge of the bed for several minutes before attempting to stand in the morning; other postural stressors should be similarly approached in a gradual manner. One maneuver that can reduce OH is leg-crossing with maintained contraction of leg muscles for 30 s; this compresses leg veins and increases systemic resistance. Compressive garments, such as compression stockings or abdominal binders, are helpful on occasion but are uncomfortable for many patients. For transient worsening of OH, drinking two 250-mL (8-oz) glasses of water within 5 min can raise standing BP 20–30 mmHg for about 2 h, beginning ~5 min after the fluid load. The patient can increase intake of salt and fluids (bouillon treatment), increase use of physical counter-maneuvers (elevate the legs when supine), or temporarily resort to a full-body stocking (compression pressure 30–40 mmHg).

Anemia can be corrected with erythropoietin, administered subcutaneously at doses of 25–75 U/kg three times per week. The hematocrit increases after 2–6 weeks. A weekly maintenance dose is usually necessary. However, the increased intravascular volume that accompanies the rise in hematocrit can exacerbate supine hypertension and requires monitoring.

If these measures are not sufficient, additional pharmacologic treatment may be necessary. Midodrine, a directly acting α_1-agonist that does not cross the blood-brain barrier, is effective. It has a duration of action of 2–4 h. The usual dose is 5–10 mg orally tid, but some patients respond best to a decremental dose (e.g., 15 mg on awakening, 10 mg at noon, and 5 mg in the afternoon). Midodrine should not be taken after 6:00 P.M. Side effects include pruritus, uncomfortable piloerection, and supine hypertension, especially at higher doses. Droxidopa (Northera) for treatment of neurogenic OH associated with PAF, PD, or MSA is effective in decreasing symptoms of OH. Pyridostigmine appears to improve OH without aggravating supine hypertension by enhancing ganglionic transmission (maximal when orthostatic, minimal when supine), but with only modest clinical effects on BP. Fludrocortisone will reduce OH but aggravates supine hypertension. At doses between 0.1 mg/d and 0.3 mg bid orally, it enhances renal sodium conservation and increases the sensitivity of arterioles to NE. Susceptible patients may develop fluid overload, congestive heart failure, supine hypertension, or hypokalemia. Potassium supplements are often necessary with chronic administration of fludrocortisone. Sustained elevations of supine BP >180/110 mmHg should be avoided. Supine hypertension (>180/110 mmHg) can be self-treated by avoiding the supine position (e.g., sleeping in a recumbent chair or elevating the head of the bed) and reducing fludrocortisone. If these simple measures are not adequate, drugs to be considered include oral hydralazine (25 mg qhs), oral nifedipine (Procardia; 10 mg qhs), or a nitroglycerin patch.

A promising drug combination (atomoxetine and yohimbine) has been studied for use in human subjects with severe OH not responsive to other agents, as can occur is some patients with diabetes and severe autonomic neuropathy not responsive to other medications. The atomoxetine blocks the NE reuptake transporter, and yohimbine blocks α_2 receptors that mediate the sympathetic feedback loop for downregulation of BP in response to atomoxetine. The result is a dramatic increase in BP and standing tolerance. Yohimbine is no longer produced commercially and must be obtained from a compounding pharmacy. This combination is not FDA approved for this purpose.

Postprandial OH may respond to several measures. Frequent, small, low-carbohydrate meals may diminish splanchnic shunting of blood after meals and reduce postprandial OH. Prostaglandin inhibitors (ibuprofen or indomethacin) taken with meals or midodrine (10 mg with the meal) can be helpful. The somatostatin analogue octreotide can be useful in the treatment of postprandial syncope by inhibiting the release of GI peptides that have vasodilator and hypotensive effects. The subcutaneous dose ranges from 25 μg bid to 200 μg tid.

TABLE 432-9

Initial Treatment of Orthostatic Hypotension (OH)

Patient education: mechanisms and stressors of OH
High-salt diet (10–20 g/d)
High-fluid intake (2 L/d)
Elevate head of bed 10 cm (4 in.) to minimize supine hypertension
Maintain postural stimuli
Learn physical counter-maneuvers
Compression garments
Correct anemia

ACKNOWLEDGMENT

The authors want to thank Phillip A. Low for his contributions to previous editions of this chapter.

FURTHER READING

Gibbons CH et al: The recommendations of a consensus panel for the screening, diagnosis, and treatment of neurogenic orthostatic hypotension and associated supine hypertension. J Neurol 264:1587, 2017.

Kaufmann H et al: Natural history of pure autonomic failure: A United States prospective cohort. Ann Neurol 81:287, 2017. [PubMed: 28093795]

McKeon A, Benarroch EE: Autoimmune autonomic disorders. Handb Clin Neurol 133:405, 2016. [PubMed: 27112689]

Singer W et al: Pure autonomic failure: Predictors of conversion to clinical CNS involvement. Neurology 88:1129, 2017. [PubMed: 28202694]

Vinik AI, Erbas T: Diabetic autonomic neuropathy. Handb Clin Neurol 117:279, 2013. [PubMed: 24095132]

Chapter 433: Trigeminal Neuralgia, Bell's Palsy, and Other Cranial Nerve Disorders

M. Flint Beal; Stephen L. Hauser

INTRODUCTION

Symptoms and signs of cranial nerve pathology are common in internal medicine. They often develop in the context of a widespread neurologic disturbance, and in such situations, cranial nerve involvement may represent the initial manifestation of the illness. In other disorders, involvement is largely restricted to one or several cranial nerves; these distinctive disorders are reviewed in this chapter. **Disorders of ocular movement are discussed in Chap. 28, disorders of hearing in Chap. 30, and vertigo and disorders of vestibular function in Chap. 19**.

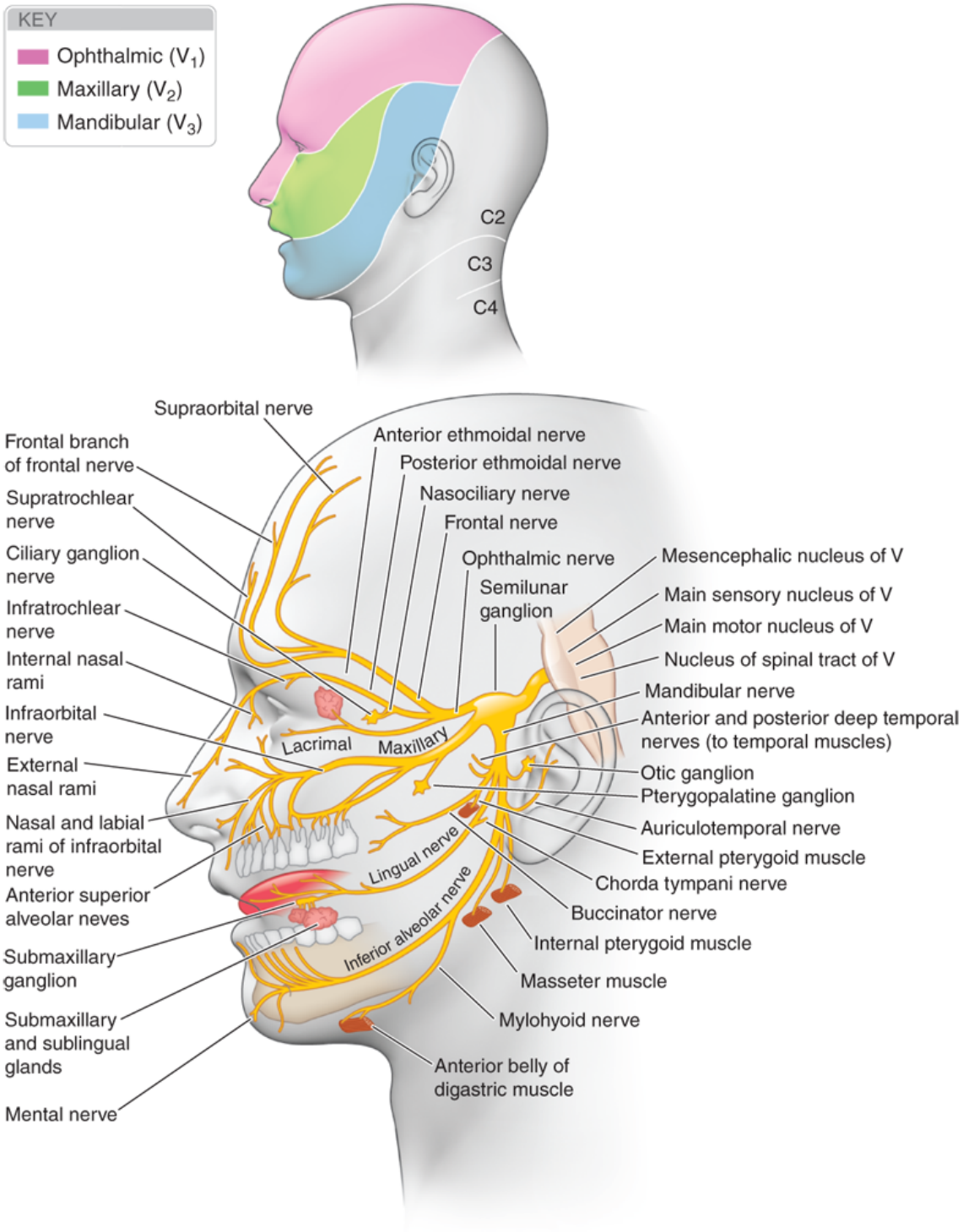
FACIAL PAIN OR NUMBNESS

ANATOMIC CONSIDERATIONS

The trigeminal (fifth cranial) nerve supplies sensation to the skin of the face and anterior half of the head **(Fig. 433-1)**. The motor part innervates the muscles involved in chewing (including masseters and pterygoids) as well as the tensor tympani of the middle ear (hearing especially for high-pitched tones). It is the largest of the cranial nerves. It exits in the lateral midpons and traverses the middle cranial fossa to the semilunar (gasserian, trigeminal) ganglion in Meckel's cave, where the nerve divides into three divisions (ophthalmic [V1], maxillary [V2], and mandibular [V3]). V1 and V2 traverse the cavernous sinus to exit in the superior orbital fissure and foramen rotundum, located above and below the eye socket respectively; V3 exits through the foramen ovale. The trigeminal nerve is predominantly sensory, and motor innervation is exclusively carried in V3. The cornea is primarily innervated by V1, although an inferior crescent may be V2. Upon entering the pons, pain and temperature fibers descend ipsilaterally to the upper cervical spinal cord as the spinal tract of V, before synapsing with the spinal nucleus of V; this accounts for the facial numbness that can occur with spinal cord lesions above C2. In the brainstem, the spinal tract of V is also located adjacent to crossed ascending fibers of the spinothalamic tract, producing a "crossed" sensory loss for pain and temperature (ipsilateral face, contralateral arm/trunk/leg) with lesions of the lateral lower brainstem. CN V is also ensheathed by oligodendrocyte-derived, rather than Schwann cell–derived, myelin for up to 7 mm after it leaves the brainstem, unlike just a few millimeters for other cranial and spinal nerves; this may explain the high frequency of trigeminal neuralgia in multiple sclerosis (MS) **(Chap. 436)**, a disorder of oligodendrocyte myelin.

FIGURE 433-1

The trigeminal nerve and its branches and sensory distribution on the face. The three major sensory divisions of the trigeminal nerve consist of the ophthalmic, maxillary, and mandibular nerves. *(Adapted from Waxman SG: Clinical Neuroanatomy, 26th ed. http://www.accessmedicine.com.ezproxy1.library.usyd.edu.au. Copyright © The McGraw-Hill Companies, Inc. All rights reserved.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TRIGEMINAL NEURALGIA (TIC DOULOUREUX)

Clinical Manifestations

Trigeminal neuralgia is characterized by excruciating paroxysms of pain in the lips, gums, cheek, or chin and, very rarely, in the distribution of the ophthalmic division of the fifth nerve. The pain seldom lasts more than a few seconds or a minute or two but may be so intense that the patient winces, hence the term tic. The paroxysms, experienced as single jabs or clusters, tend to recur frequently, both day and night, for several weeks at a time. They may occur spontaneously or with movements of affected areas evoked by speaking, chewing, or smiling. Another characteristic feature is the presence of trigger zones, typically on the face, lips, or tongue, that provoke attacks; patients may report that tactile stimuli—e.g., washing the face, brushing the teeth, or exposure to a draft of air—generate excruciating pain. An essential feature of trigeminal neuralgia is that objective signs of sensory loss cannot be demonstrated on examination.

Trigeminal neuralgia is relatively common, with an estimated annual incidence of 4–8 per 100,000 individuals. Middle-aged and elderly persons are affected primarily, and ~60% of cases occur in women. Onset is typically sudden, and bouts tend to persist for weeks or months before remitting spontaneously. Remissions may be long-lasting, but in most patients, the disorder ultimately recurs.

Pathophysiology

Symptoms result from ectopic generation of action potentials in pain-sensitive afferent fibers of the fifth cranial nerve root just before it enters the lateral surface of the pons. Compression or other pathology in the nerve leads to demyelination of large myelinated fibers that do not themselves carry pain sensation but become hyperexcitable and electrically coupled with smaller unmyelinated or poorly myelinated pain fibers in close proximity; this may explain why tactile stimuli, conveyed via the large myelinated fibers, can stimulate paroxysms of pain. Compression of the trigeminal nerve root by a blood vessel, most often the superior cerebellar artery or on occasion a tortuous vein, is now believed to be the source of trigeminal neuralgia in most patients. In cases of vascular compression, age-related brain sagging and increased vascular thickness and tortuosity may explain the prevalence of trigeminal neuralgia in later life.

Differential Diagnosis

Trigeminal neuralgia must be distinguished from other causes of face and head pain **(Chap. 13)** and from pain arising from diseases of the jaw, teeth, or sinuses. Pain from migraine or cluster headache tends to be deep-seated and steady, unlike the superficial stabbing quality of trigeminal neuralgia; rarely, cluster headache is associated with trigeminal neuralgia, a syndrome known as *cluster-tic*. Other rare headaches include short-lasting unilateral headache attacks with conjunctival injection and tearing (SUNCT; **Chap. 422**). In temporal arteritis, superficial facial pain is present but is not typically shock-like, the patient frequently complains of myalgias and other systemic symptoms, and an elevated erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP) is usually present **(Chap. 356)**. When trigeminal neuralgia develops in a young adult or is bilateral, MS is a key consideration, and in such cases, the cause is a demyelinating plaque near the root entry zone of the fifth nerve in the pons; often, evidence of facial sensory loss can be found on careful examination. Cases that are secondary to mass lesions—such as aneurysms, neurofibromas, acoustic schwannomas, or meningiomas—usually produce objective signs of sensory loss in the trigeminal nerve distribution (trigeminal neuropathy, see below).

Laboratory Evaluation

An ESR or CRP is indicated if temporal arteritis is suspected. In typical cases of trigeminal neuralgia, neuroimaging studies are usually unnecessary but may be valuable if MS is a consideration or in assessing overlying vascular lesions in order to plan for decompression surgery.

TREATMENT

TREATMENT

Trigeminal Neuralgia

Drug therapy with carbamazepine is effective in ~50–75% of patients. Carbamazepine should be started as a single daily dose of 100 mg taken with food and increased gradually (by 100 mg daily in divided doses every 1–2 days) until substantial (>50%) pain relief is achieved. Most patients require a maintenance dose of 200 mg qid. Doses >1200 mg daily provide no additional benefit. Dizziness, imbalance, sedation, and rare cases of agranulocytosis are the most important side effects of carbamazepine. If treatment is effective, it is usually continued for 1 month and then tapered as tolerated. Oxcarbazepine (300–1200 mg bid) is an alternative to carbamazepine that has less bone marrow toxicity and probably is equally efficacious. If these agents are not well tolerated or are ineffective, lamotrigine, 400 mg daily, and phenytoin, 300–400 mg daily, are other options. Baclofen may also be tried, either alone or in combination with an anticonvulsant. The initial dose is 5–10 mg tid, gradually increasing as needed to 20 mg qid.

If drug treatment fails, surgical therapy should be offered. The most widely used method is currently microvascular decompression to relieve pressure on the trigeminal nerve as it exits the pons. This procedure requires a suboccipital craniotomy. This procedure appears to have a >70% efficacy rate and a low rate of pain recurrence in responders; the response is better for classic tic-like symptoms than for nonlancinating facial pains. In a small number of cases, there is perioperative damage to the eighth or seventh cranial nerves or to the cerebellum or a postoperative cerebrospinal fluid leak syndrome. High-resolution magnetic resonance angiography is useful preoperatively to visualize the relationships between the fifth cranial nerve root and nearby blood vessels.

Gamma knife radiosurgery of the trigeminal nerve root is also used for treatment and results in complete pain relief, sometimes delayed in onset, in more than two-thirds of patients and a low risk of persistent facial numbness; the response is sometimes long-lasting, but recurrent pain develops over 2–3 years in half of patients. Compared with surgical decompression, gamma knife surgery appears to be somewhat less effective but has few serious complications.

Another procedure, radiofrequency thermal rhizotomy, creates a heat lesion of the trigeminal ganglion or nerve. It is used less often now than in the past. Short-term relief is experienced by >95% of patients; however, long-term studies indicate that pain recurs in up to one-third of treated patients.

Postoperatively, partial numbness of the face is common, masseter (jaw) weakness may occur especially following bilateral procedures, and corneal denervation with secondary keratitis can follow rhizotomy for first-division trigeminal neuralgia.

TRIGEMINAL NEUROPATHY

A variety of diseases can affect the trigeminal nerve (Table 433-1). Most present with sensory loss on the face or with weakness of the jaw muscles. Deviation of the jaw on opening indicates weakness of the pterygoids on the side to which the jaw deviates. Some cases are due to Sjögren's syndrome or a collagen-vascular disease such as systemic lupus erythematosus, scleroderma, or mixed connective tissue disease. Among infectious causes, herpes zoster (acute or postherpetic) and leprosy should be considered. Tumors of the middle cranial fossa (meningiomas), of the trigeminal nerve (schwannomas), or of the base of the skull (metastatic tumors) may cause a combination of motor and sensory signs. Lesions in the cavernous sinus can affect the first and second divisions of the trigeminal nerve, and lesions of the superior orbital fissure can affect the first (ophthalmic) division; the accompanying corneal anesthesia increases the risk of ulceration (neurokeratitis).

TABLE 433-1

Trigeminal Nerve Disorders

Nuclear (Brainstem) Lesions
Multiple sclerosis Stroke Syringobulbia Glioma Lymphoma
Preganglionic Lesions
Acoustic neuroma Meningioma Metastasis Chronic meningitis Cavernous carotid aneurysm
Gasserian Ganglion Lesions
Trigeminal neuroma Herpes zoster Infection (spread from otitis media or mastoiditis)
Peripheral Nerve Lesions
Nasopharyngeal carcinoma Trauma Guillain-Barré syndrome Sjögren's syndrome Collagen-vascular diseases Sarcoidosis Leprosy Drugs (stilbamidine, trichloroethylene) Idiopathic trigeminal neuropathy

Isolated sensory loss over the chin (mental neuropathy) can be the only manifestation of systemic malignancy. Rarely, an idiopathic form of trigeminal neuropathy is observed. It is characterized by numbness and paresthesias, sometimes bilaterally, with loss of sensation in the territory of the trigeminal nerve but without weakness of the jaw. Gradual recovery is the rule. Tonic spasm of the masticatory muscles, known as trismus, is symptomatic of tetanus (Chap. 147) or may occur in patients treated with phenothiazines.

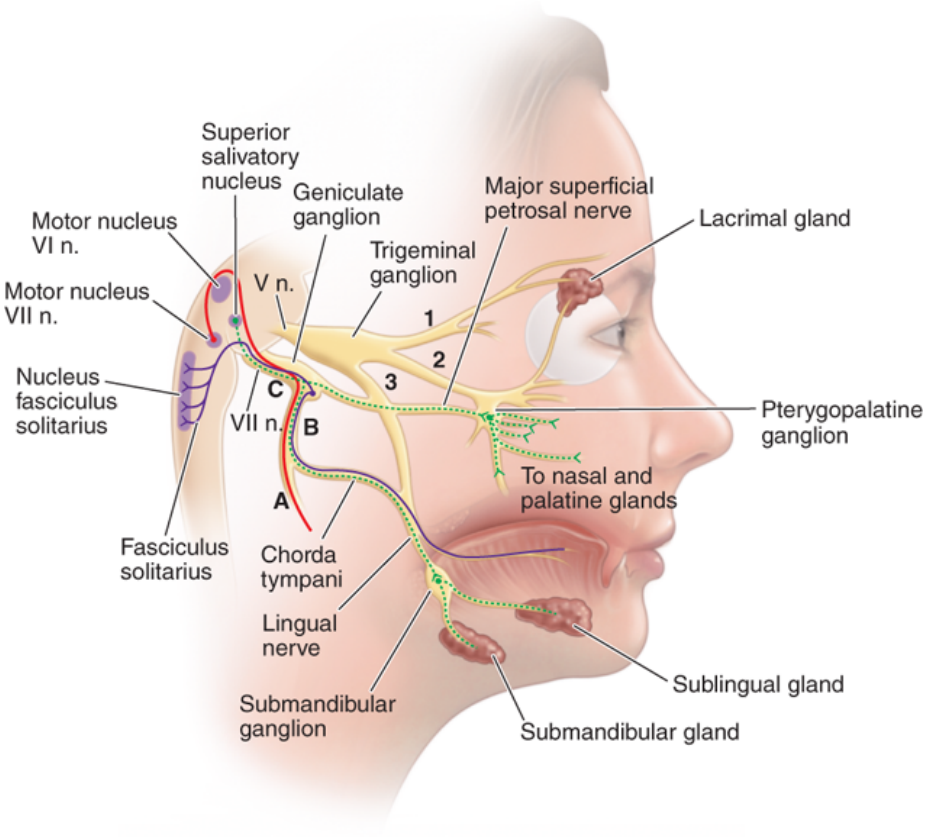
FACIAL WEAKNESS

ANATOMIC CONSIDERATIONS

(Fig. 433-2) The seventh cranial nerve supplies all the muscles concerned with facial expression. The sensory component is small (the nervus intermedius); it conveys taste sensation from the anterior two-thirds of the tongue and probably cutaneous impulses from the anterior wall of the external auditory canal. The motor nucleus of the seventh nerve lies anterior and lateral to the abducens nucleus. After leaving the pons, the seventh nerve enters the internal auditory meatus with the acoustic nerve. The nerve continues its course in its own bony channel, the facial canal, and exits from the skull via the stylomastoid foramen. It then passes through the parotid gland and subdivides to supply the facial muscles.

FIGURE 433-2

The facial nerve. A, B, and C denote lesions of the facial nerve at the stylomastoid foramen, distal and proximal to the geniculate ganglion, respectively. *Green lines* indicate the parasympathetic fibers, *red line* indicates motor fibers, and *purple lines* indicate visceral afferent fibers (taste). *(Adapted from MB Carpenter: Core Text of Neuroanatomy, 2nd ed. Williams & Wilkins, 1978.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

A complete interruption of the facial nerve at the stylomastoid foramen paralyzes all muscles of facial expression. The corner of the mouth droops, the creases and skinfolds are effaced, the forehead is unfurrowed, and the eyelids will not close. Upon attempted closure of the lids, the eye on the paralyzed side rolls upward (Bell's phenomenon). The lower lid sags and falls away from the conjunctiva, permitting tears to spill over the cheek. Food collects between the teeth and lips, and saliva may dribble from the corner of the mouth. The patient complains of a heaviness or numbness in the face, but sensory loss is rarely demonstrable and taste is intact.

If the lesion is in the middle-ear portion, taste is lost over the anterior two-thirds of the tongue on the same side. If the nerve to the stapedius is interrupted, there is hyperacusis (sensitivity to loud sounds). Lesions in the internal auditory meatus may affect the adjacent auditory and vestibular nerves, causing deafness, tinnitus, or dizziness. Intrapontine lesions that paralyze the face usually affect the abducens nucleus as well, and often the corticospinal and sensory tracts.

If the peripheral facial paralysis has existed for some time and recovery of motor function is incomplete, a continuous diffuse contraction of facial muscles may appear. The palpebral fissure becomes narrowed, and the nasolabial fold deepens. Attempts to move one group of facial muscles may result in contraction of all (associated movements, or synkinesis). Facial spasms, initiated by movements of the face, may develop (hemifacial spasm). Anomalous regeneration of seventh nerve fibers may result in other troublesome phenomena. If fibers originally connected with the orbicularis oculi come to innervate the orbicularis oris, closure of the lids may cause a retraction of the mouth, or if fibers originally connected with muscles of the face later innervate the lacrimal gland, anomalous tearing ("crocodile tears") may occur with any activity of the facial muscles, such as eating. Another facial synkinesia is triggered by jaw opening, causing closure of the eyelids on the side of the facial palsy (jaw-winking).

BELL'S PALSY

The most common form of facial paralysis is Bell's palsy. The annual incidence of this idiopathic disorder is ~25 per 100,000 annually, or about 1 in 60 persons in a lifetime. Risk factors include pregnancy and diabetes mellitus.

Clinical Manifestations

The onset of Bell's palsy is fairly abrupt, with maximal weakness being attained by 48 h as a general rule. Pain behind the ear may precede the paralysis for a day or two. Taste sensation may be lost unilaterally, and hyperacusis may be present. In some cases, there is mild cerebrospinal fluid lymphocytosis. Magnetic resonance imaging (MRI) may reveal swelling and uniform enhancement of the geniculate ganglion and facial nerve and, in some cases, entrapment of the swollen nerve in the temporal bone. Approximately 80% of patients recover within a few weeks or months. Electromyography may be of some prognostic value; evidence of denervation after 10 days indicates there has been axonal degeneration, that there will be a long delay (3 months as a rule) before regeneration occurs, and that it may be incomplete. The presence of incomplete paralysis in the first week is the most favorable prognostic sign. Recurrences are reported in ~7% of cases.

Pathophysiology

In acute Bell's palsy, there is inflammation of the facial nerve with mononuclear cells, consistent with an infectious or immune cause. Herpes simplex virus (HSV) type 1 DNA was frequently detected in endoneurial fluid and posterior auricular muscle, suggesting that a reactivation of this virus in the geniculate ganglion may be responsible for most cases. Reactivation of varicella-zoster virus is associated with Bell's palsy in up to one-third of cases and may represent the second most frequent cause. A variety of other viruses have also been implicated less commonly. An increased incidence of Bell's palsy was also reported among recipients of inactivated intranasal influenza vaccine, and it was hypothesized that this could have resulted from the *Escherichia coli* enterotoxin used as adjuvant or reactivation of latent virus.

Differential Diagnosis

There are many other causes of acute facial palsy that must be considered in the differential diagnosis of Bell's palsy. Lyme disease can cause unilateral or bilateral facial palsies; in endemic areas, ≥10% of cases of facial palsy are likely due to infection with *Borrelia burgdorferi* **(Chap. 181)**. The Ramsay Hunt syndrome, caused by reactivation of herpes zoster in the geniculate ganglion, consists of a severe facial palsy associated with a vesicular eruption in the external auditory canal and sometimes in the pharynx and other parts of the cranial integument; often the eighth cranial nerve is affected as well. Facial palsy that is often bilateral occurs in sarcoidosis **(Chap. 360)** and in Guillain-Barré syndrome **(Chap. 439)**. Leprosy frequently involves the facial nerve, and facial neuropathy may also occur in diabetes mellitus, connective tissue diseases including Sjögren's syndrome, and amyloidosis. The rare Melkersson-Rosenthal syndrome consists of recurrent facial paralysis; recurrent—and eventually permanent—facial (particularly labial) edema; and, less constantly, plication of the tongue. Its cause is unknown. Acoustic neuromas frequently involve the facial nerve by local compression. Infarcts, demyelinating lesions of MS, and tumors are the common pontine lesions that interrupt the facial nerve fibers; other signs of brainstem involvement are usually present. Tumors that invade the temporal bone (carotid body, cholesteatoma, dermoid) may produce a facial palsy, but the onset is insidious and the course progressive.

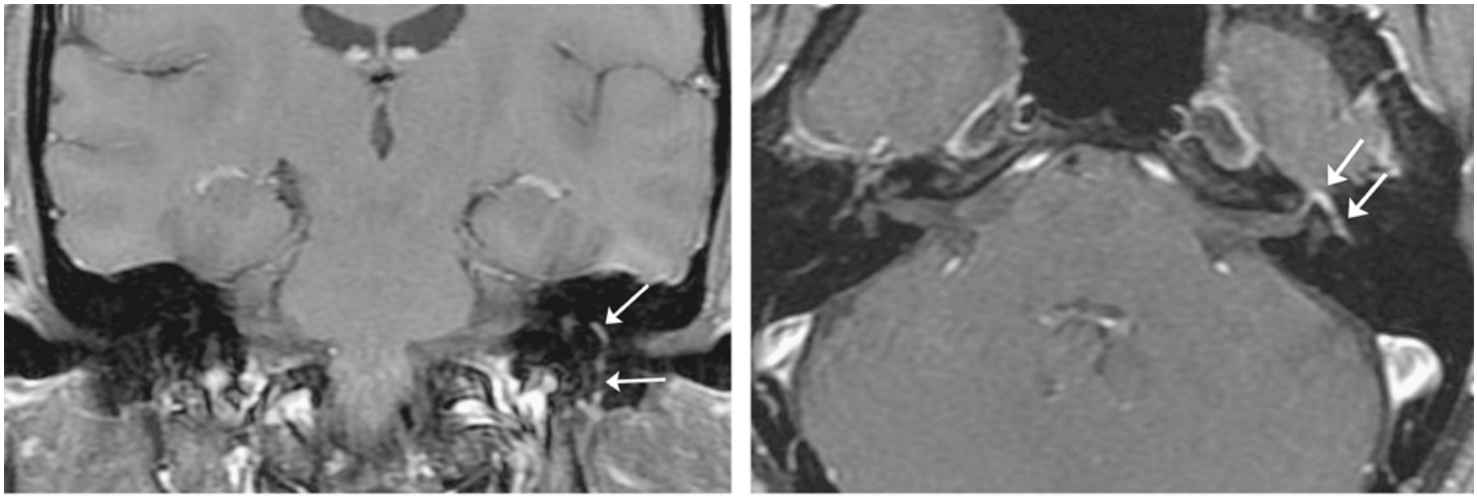
All these forms of nuclear or peripheral facial palsy must be distinguished from the supranuclear type. In the latter, the frontalis and orbicularis oculi muscles of the forehead are involved less than those of the lower part of the face, since the upper facial muscles are innervated by corticobulbar pathways from both motor cortices, whereas the lower facial muscles are innervated only by the opposite hemisphere. In supranuclear lesions, there may be a dissociation of emotional and voluntary facial movements, and often some degree of paralysis of the arm and leg or an aphasia (in dominant hemisphere lesions) is present.

Laboratory Evaluation

The diagnosis of Bell's palsy can usually be made clinically in patients with (1) a typical presentation, (2) no risk factors or preexisting symptoms for other causes of facial paralysis, (3) absence of cutaneous lesions of herpes zoster in the external ear canal, and (4) a normal neurologic examination with the exception of the facial nerve. Particular attention to the eighth cranial nerve, which courses near to the facial nerve in the pontomedullary junction and in the temporal bone, and to other cranial nerves is essential. In atypical or uncertain cases, an ESR or CRP, testing for diabetes mellitus, a Lyme titer, angiotensin-converting enzyme and chest imaging studies for possible sarcoidosis, a lumbar puncture for possible Guillain-Barré syndrome, or MRI scanning may be indicated. MRI often shows swelling and enhancement of the facial nerve in idiopathic Bell's palsy **(Fig. 433-3)**.

FIGURE 433-3

Axial and coronal T1-weighted images after gadolinium with fat suppression demonstrate diffuse smooth linear enhancement of the left facial nerve, involving the genu, tympanic, and mastoid segments within the temporal bone (*arrows*), without evidence of mass lesion. Although highly suggestive of Bell's palsy, similar findings may be seen with other etiologies such as Lyme disease, sarcoidosis, and perineural malignant spread.



A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Bell's Palsy

Symptomatic measures include (1) the use of paper tape to depress the upper eyelid during sleep and prevent corneal drying, (2) artificial tears; and (3) massage of the weakened muscles. A course of glucocorticoids, given as prednisone 60–80 mg daily during the first 5 days and then tapered over the next 5 days, modestly shortens the recovery period and improves the functional outcome. Although large and well-controlled randomized trials found no added benefit of the antiviral agents valacyclovir (1000 mg daily for 5–7 days) or acyclovir (400 mg five times daily for 10 days) compared to glucocorticoids alone, some earlier data suggested that combination therapy with prednisone plus valacyclovir might be marginally better than prednisone alone, especially in patients with severe clinical presentations. For patients with permanent paralysis from Bell's palsy, a number of cosmetic surgical procedures have been used to restore a relatively symmetric appearance to the face.

OTHER MOTOR DISORDERS OF THE FACE

Hemifacial spasm consists of painless irregular involuntary contractions on one side of the face. Most cases appear related to vascular compression of the exiting facial nerve in the pons. Other cases develop as a sequela to Bell's palsy or are secondary to compression and/or demyelination of the nerve by tumor, infection, or MS. Mild cases can be treated with carbamazepine, gabapentin, or, if these drugs fail, baclofen. Local injections of botulinum toxin into affected muscles can relieve spasms for 3–4 months, and the injections can be repeated. Refractory cases due to vascular compression usually respond to surgical decompression of the facial nerve. Blepharospasm is an involuntary recurrent spasm of both eyelids that usually occurs in elderly persons as an isolated phenomenon or with varying degrees of spasm of other facial muscles. Severe, persistent cases of blepharospasm can be treated by local injection of botulinum toxin into the orbicularis oculi. Facial myokymia refers to a fine rippling activity of the facial muscles; it may be caused by MS or follow Guillain-Barré syndrome (Chap. 439).

Facial hemiatrophy occurs mainly in women and is characterized by a disappearance of fat in the dermal and subcutaneous tissues on one side of the face. It usually begins in adolescence or the early adult years and is slowly progressive. In its advanced form, the affected side of the face is gaunt, and the skin is thin, wrinkled, and brown. The facial hair may turn white and fall out, and the sebaceous glands become atrophic. Bilateral involvement may occur. A limited form of systemic sclerosis (scleroderma) may be the cause of some cases. Treatment is cosmetic, consisting of transplantation of skin and subcutaneous fat.

OTHER CRANIAL NERVE DISORDERS

GLOSSOPHARYNGEAL NEURALGIA

This form of neuralgia involves the ninth (glossopharyngeal) and sometimes portions of the tenth (vagus) cranial nerves. It resembles trigeminal neuralgia in many respects but is much less common. The pain is intense and paroxysmal; it originates on one side of the throat, approximately in the tonsillar fossa. In some cases, the pain is localized in the ear or may radiate from the throat to the ear because of involvement of the tympanic branch of the glossopharyngeal nerve. Spasms of pain may be initiated by swallowing or coughing. There is no demonstrable motor or sensory deficit; the glossopharyngeal nerve supplies taste sensation to the posterior third of the tongue and, together with the vagus nerve, sensation to the posterior

pharynx. Cardiac symptoms—bradycardia or asystole, hypotension, and fainting—have been reported. Glossopharyngeal neuralgia can result from vascular compression, MS, or tumors, but many cases are idiopathic. Medical therapy is similar to that for trigeminal neuralgia, and carbamazepine is generally the first choice. If drug therapy is unsuccessful, surgical procedures—including microvascular decompression if vascular compression is evident—or rhizotomy of glossopharyngeal and vagal fibers in the jugular bulb is frequently successful.

Glossopharyngeal neuropathy in conjunction with vagus and accessory nerve palsies may occur with herpes zoster infection or with a tumor or aneurysm in the posterior fossa or in the jugular foramen. Hoarseness due to vocal cord paralysis, some difficulty in swallowing, deviation of the soft palate to the intact side, anesthesia of the posterior wall of the pharynx, and weakness of the upper part of the trapezius and sternocleidomastoid muscles make up the jugular foramen syndrome **(Table 433-2)**.

TABLE 433-2

Cranial Nerve Syndromes

Site	Cranial Nerves	Usual Cause
Sphenoid fissure (superior orbital)	III, IV, first division V, VI	Invasive tumors of sphenoid bone; aneurysms
Lateral wall of cavernous sinus	III, IV, first division V, VI, often with proptosis	Infection, thrombosis, aneurysm, or fistula of cavernous sinus; invasive tumors from sinuses and sella turcica; benign granuloma responsive to glucocorticoids
Retrosphenoid space	II, III, IV, V, VI	Large tumors of middle cranial fossa
Apex of petrous bone	V, VI	Petrositis; tumors of petrous bone
Internal auditory meatus	VII, VIII	Tumors of petrous bone (dermoids, etc.); infectious processes; acoustic neuroma
Pontocerebellar angle	V, VII, VIII, and sometimes IX	Acoustic neuroma; meningioma
Jugular foramen	IX, X, XI	Tumors and aneurysms
Posterior laterocondylar space	IX, X, XI, XII	Tumors of parotid gland and carotid body and metastatic tumors
Posterior retroparotid space	IX, X, XI, XII, and Horner's syndrome	Tumors of parotid gland, carotid body, lymph nodes; metastatic tumor; tuberculous adenitis

DYSPHAGIA AND DYSPHONIA

When the intracranial portion of one vagus (tenth cranial) nerve is interrupted, the soft palate droops ipsilaterally and does not rise in phonation. There is loss of the gag reflex on the affected side, as well as of the "curtain movement" of the lateral wall of the pharynx, whereby the faucial pillars move medially as the palate rises in saying "ah." The voice is hoarse and slightly nasal, and the vocal cord lies immobile midway between abduction and adduction. Loss of sensation at the external auditory meatus and the posterior pinna may also be present.

The pharyngeal branches of both vagal nerves may be affected in diphtheria; the voice has a nasal quality, and regurgitation of liquids through the nose occurs during swallowing.

Injury to the vagus nerve in the carotid sheath can also occur with carotid dissection or following endarterectomy. The vagus nerve may be involved at the meningeal level by neoplastic and infectious processes and within the medulla by tumors, vascular lesions (e.g., the lateral medullary syndrome), and motor neuron disease. This nerve may be involved by infection with varicella zoster virus. Polymyositis and dermatomyositis, which cause hoarseness and dysphagia by direct involvement of laryngeal and pharyngeal muscles, may be confused with diseases of the vagus nerves. Dysphagia is also a symptom in some patients with myotonic dystrophy. **Nonneurologic causes of dysphagia are discussed in Chap. 40**.

The recurrent laryngeal nerves, especially the left, are most often damaged as a result of intrathoracic disease. Aneurysm of the aortic arch, an enlarged left atrium, and tumors of the mediastinum and bronchi are much more frequent causes of an isolated vocal cord palsy than are intracranial disorders. However, a substantial number of cases of recurrent laryngeal palsy remain idiopathic.

When confronted with a case of laryngeal palsy, the physician must attempt to determine the site of the lesion. If it is intramedullary, there are usually other signs, such as ipsilateral cerebellar dysfunction, loss of pain and temperature sensation over the ipsilateral face and contralateral arm and leg, and an ipsilateral Horner's syndrome. If the lesion is extramedullary, the glossopharyngeal and spinal accessory nerves are frequently involved (jugular foramen syndrome). If it is extracranial in the posterior laterocondylar or retroparotid space, there may be a combination of ninth, tenth, eleventh, and twelfth cranial nerve palsies and a Horner's syndrome (Table 433-2). If there is no sensory loss over the palate and pharynx and no palatal weakness or dysphagia, the lesion is below the origin of the pharyngeal branches, which leave the vagus nerve high in the cervical region; the usual site of disease is then the mediastinum.

NECK WEAKNESS

Isolated involvement of the accessory (eleventh cranial) nerve can occur anywhere along its route, resulting in partial or complete paralysis of the sternocleidomastoid and trapezius muscles. More commonly, involvement occurs in combination with deficits of the ninth and tenth cranial nerves in the jugular foramen or after exit from the skull (Table 433-2). An idiopathic form of accessory neuropathy, akin to Bell's palsy, has been described, and it may be recurrent in some cases. Most but not all patients recover.

TONGUE PARALYSIS

The hypoglossal (twelfth cranial) nerve supplies the ipsilateral muscles of the tongue. The nucleus of the nerve or its fibers of exit may be involved by intramedullary lesions such as tumor, poliomyelitis, or most often motor neuron disease. Lesions of the basal meninges and the occipital bones (platybasia, invagination of occipital condyles, Paget's disease) may compress the nerve in its extramedullary course or in the hypoglossal canal. Isolated lesions of unknown cause can occur. Atrophy and fasciculation of the tongue develop weeks to months after interruption of the nerve.

MULTIPLE CRANIAL NERVE PALSIES

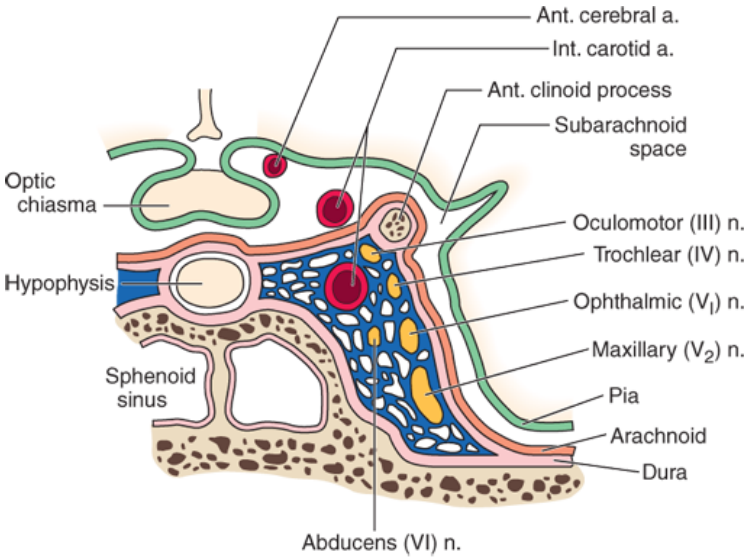
Several cranial nerves may be affected by the same disease process. In this situation, the main clinical problem is to determine whether the lesion lies within the brainstem or outside it. Lesions that lie on the surface of the brainstem are characterized by involvement of adjacent cranial nerves (often occurring in succession) and late and rather slight involvement of the long sensory and motor pathways and segmental structures lying within the brainstem. The opposite is true of primary lesions within the brainstem. The extramedullary lesion is more likely to cause bone erosion or enlargement of the foramens of exit of cranial nerves. The intramedullary lesion involving cranial nerves often produces a crossed sensory or motor paralysis (cranial nerve signs on one side of the body and tract signs on the opposite side).

Involvement of multiple cranial nerves outside the brainstem is frequently the result of trauma, localized infections including varicella-zoster virus, infectious and noninfectious (especially carcinomatous) causes of meningitis **(Chaps. 133 and 134)**, granulomatous diseases such as granulomatosis with polyangiitis **(Chap. 356)**, Behçet's disease, vascular disorders including those associated with diabetes, enlarging aneurysms, or locally infiltrating tumors. Among the tumors, nasopharyngeal cancers, lymphomas, neurofibromas, meningiomas, chordomas, cholesteatomas, carcinomas, and sarcomas have all been observed to involve a succession of lower cranial nerves. Owing to their anatomic relationships, the multiple cranial nerve palsies form a number of distinctive syndromes, listed in Table 433-2. Sarcoidosis is the cause of some cases of multiple cranial neuropathy; tuberculosis, the Chiari malformation, platybasia, and basilar invagination of the skull are additional causes. A purely motor disorder without atrophy always raises the question of myasthenia gravis **(Chap. 440)**. As noted above, Guillain-Barré syndrome commonly affects the facial nerves bilaterally. In the Fisher variant of the Guillain-Barré syndrome, oculomotor paresis occurs with ataxia and areflexia in the limbs **(Chap. 439)**. Wernicke's encephalopathy can cause a severe ophthalmoplegia combined with other brainstem signs **(Chap. 301)**.

The cavernous sinus syndrome **(Fig. 433-4)** is a distinctive and frequently life-threatening disorder. It often presents as orbital or facial pain; orbital swelling and chemosis due to occlusion of the ophthalmic veins; fever; oculomotor neuropathy affecting the third, fourth, and sixth cranial nerves; and trigeminal neuropathy affecting the ophthalmic (V1) and occasionally the maxillary (V2) divisions of the trigeminal nerve. Cavernous sinus thrombosis, often secondary to infection from orbital cellulitis (frequently *Staphylococcus aureus*), a cutaneous source on the face, or sinusitis (especially with mucormycosis in diabetic patients), is the most frequent cause; other etiologies include aneurysm of the carotid artery, a carotid-cavernous fistula (orbital bruit may be present), meningioma, nasopharyngeal carcinoma, other tumors, or an idiopathic granulomatous disorder (Tolosa-Hunt syndrome). The two cavernous sinuses directly communicate via intercavernous channels; thus, involvement on one side may extend to become bilateral. Early diagnosis is essential, especially when due to infection, and treatment depends on the underlying etiology.

FIGURE 433-4

Anatomy of the cavernous sinus in coronal section, illustrating the location of the cranial nerves in relation to the vascular sinus, internal carotid artery (which loops anteriorly to the section), and surrounding structures.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In infectious cases, prompt administration of broad-spectrum antibiotics, drainage of any abscess cavities, and identification of the offending organism are essential. Anticoagulant therapy may benefit cases of primary thrombosis. Repair or occlusion of the carotid artery may be required for treatment of fistulas or aneurysms. The Tolosa-Hunt syndrome generally responds to glucocorticoids. A dramatic improvement in pain is usually evident within a few days; oral prednisone (60 mg daily) is usually continued for 2 weeks and then gradually tapered over a month, or longer if pain recurs. Occasionally an immunosuppressive medication, such as azathioprine or methotrexate, needs to be added to maintain an initial response to glucocorticoids.

An idiopathic form of multiple cranial nerve involvement on one or both sides of the face is occasionally seen. The syndrome consists of a subacute onset of boring facial pain, followed by paralysis of motor cranial nerves. The clinical features overlap those of the Tolosa-Hunt syndrome and appear to be due to idiopathic inflammation of the dura mater, which may be visualized by MRI. The syndrome is usually responsive to glucocorticoids.

FURTHER READING

Cruccu G et al: Trigeminal neuralgia: New classification and diagnostic grading for practice and research. Neurology 87:220, 2016. [PubMed: 27306631]

Gagyor I et al: Antiviral treatment for Bell's palsy (idiopathic facial paralysis). Cochrane Database Syst Rev (11):CD001869, 2015.

Kelly HR, Curtin HD: Imaging of skull base lesions. Handb Clin Neurol 135:637, 2016. [PubMed: 27432686]

Madhok VB et al: Corticosteroids for Bell's palsy (idiopathic facial paralysis). Cochrane Database Syst Rev 7:CD001942, 2016. [PubMed: 27428352]

Chapter 434: Diseases of the Spinal Cord

Stephen L. Hauser

INTRODUCTION

Diseases of the spinal cord are frequently devastating. They produce quadriplegia, paraplegia, and sensory deficits far beyond the damage they would inflict elsewhere in the nervous system because the spinal cord contains, in a small cross-sectional area, almost the entire motor output and sensory input of the trunk and limbs. Many spinal cord diseases are reversible if recognized and treated at an early stage (Table 434-1); thus, they are among the most critical of neurologic emergencies. The efficient use of diagnostic procedures, guided by knowledge of the anatomy and the clinical features of spinal cord diseases, is required to maximize the likelihood of a successful outcome.

TABLE 434-1

Treatable Spinal Cord Disorders

Compressive
Epidural, intradural, or intramedullary neoplasm
Epidural abscess
Epidural hemorrhage
Cervical spondylosis
Herniated disk
Posttraumatic compression by fractured or displaced vertebra or hemorrhage
Vascular
Arteriovenous malformation and dural fistula
Antiphospholipid syndrome and other hypercoagulable states
Inflammatory
Multiple sclerosis
Neuromyelitis optica
Transverse myelitis
Sarcoidosis
Sjögren-related myelopathy
Systemic lupus erythematosus-related myelopathy
Vasculitis
Infectious
Viral: VZV, HSV-1 and 2, CMV, HIV, HTLV-1, others

Bacterial and mycobacterial: *Borrelia*, *Listeria*, syphilis, others
Mycoplasma pneumoniae
Parasitic: schistosomiasis, toxoplasmosis, cystercercosis
Developmental
Syringomyelia
Meningomyelocele
Tethered cord syndrome
Metabolic
Vitamin B_{12} deficiency (subacute combined degeneration)
Copper deficiency

Abbreviations: CMV, cytomegalovirus; HSV, herpes simplex virus; HTLV, human T cell lymphotropic virus; VZV, varicella-zoster virus.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Spinal Cord Disease

SPINAL CORD ANATOMY RELEVANT TO CLINICAL SIGNS

The spinal cord is a thin, tubular extension of the central nervous system contained within the bony spinal canal. It originates at the medulla and continues caudally to the conus medullaris at the lumbar level; its fibrous extension, the filum terminale, terminates at the coccyx. The adult spinal cord is ~46 cm (18 in.) long, oval in shape, and enlarged in the cervical and lumbar regions, where neurons that innervate the upper and lower extremities, respectively, are located. The white matter tracts containing ascending sensory and descending motor pathways are located peripherally, whereas nerve cell bodies are clustered in an inner region of gray matter shaped like a four-leaf clover that surrounds the central canal (anatomically an extension of the fourth ventricle). The membranes that cover the spinal cord—the pia, arachnoid, and dura—are continuous with those of the brain, and the cerebrospinal fluid is contained within the subarachnoid space between the pia and arachnoid.

The spinal cord has 31 segments, each defined by an exiting ventral motor root and entering dorsal sensory root. During embryologic development, growth of the cord lags behind that of the vertebral column, and the mature spinal cord ends at approximately the first lumbar vertebral body. The lower spinal nerves take an increasingly downward course to exit via intervertebral foramina. The first seven pairs of cervical spinal nerves exit above the same-numbered vertebral bodies, whereas all the subsequent nerves exit below the same-numbered vertebral bodies because of the presence of eight cervical spinal cord segments but only seven cervical vertebrae. The relationship between spinal cord segments and the corresponding vertebral bodies is shown in Table 434-2. These relationships assume particular importance for localization of lesions that cause spinal cord compression. Sensory loss below the circumferential level of the umbilicus, for example, corresponds to the T10 cord segment but indicates involvement of the cord adjacent to the seventh or eighth thoracic vertebral body **(see Figs. 22-2 and 22-3)**. In addition, at every level, the main ascending and descending tracts are somatotopically organized with a laminated distribution that reflects the origin or destination of nerve fibers.

Determining the Level of the Lesion

The presence of a horizontally defined level below which sensory, motor, and autonomic function is impaired is a hallmark of a lesion of the spinal cord. This *sensory level* is sought by asking the patient to identify a pinprick or cold stimulus applied to the proximal legs and lower trunk and successively moved up toward the neck on each side. Sensory loss below this level is the result of damage to the spinothalamic tract on the opposite side, one to two segments higher in the case of a unilateral spinal cord lesion, and at the level of a bilateral lesion. The discrepancy in the level of a unilateral lesion is the result of the course of the second-order sensory fibers, which originate in the dorsal horn, and ascend for one or two levels as they cross anterior to the central canal to join the opposite spinothalamic tract. Lesions that transect the descending corticospinal and other motor tracts cause paraplegia or quadriplegia with heightened deep tendon reflexes, Babinski signs, and eventual spasticity (the upper motor neuron syndrome). Transverse damage to the cord also produces autonomic disturbances consisting of absent sweating below the implicated cord level and bladder, bowel, and sexual dysfunction.

The uppermost level of a spinal cord lesion can also be localized by attention to the *segmental signs* corresponding to disturbed motor or sensory innervation by an individual cord segment. A band of altered sensation (hyperalgesia or hyperpathia) at the upper end of the sensory disturbance, fasciculations or atrophy in muscles innervated by one or several segments, or a muted or absent deep tendon reflex may be noted at this level. These signs also can occur with focal root or peripheral nerve disorders; thus, they are most useful when they occur together with signs of long tract damage. With severe and acute transverse lesions, the limbs initially may be flaccid rather than spastic. This state of “spinal shock” lasts for several days, rarely for weeks, and may be mistaken for extensive damage to the anterior horn cells over many segments of the cord or for an acute polyneuropathy.

The main features of transverse damage at each level of the spinal cord are summarized below.

Cervical Cord

Upper cervical cord lesions produce quadriplegia and weakness of the diaphragm. The uppermost level of weakness and reflex loss with lesions at C5-C6 is in the biceps; at C7, in finger and wrist extensors and triceps; and at C8, finger, and wrist flexion. Horner's syndrome (miosis, ptosis, and facial hypohidrosis) may accompany a cervical cord lesion at any level.

Thoracic Cord

Lesions here are localized by the sensory level on the trunk and, if present, by the site of midline back pain. Useful markers of the sensory level on the trunk are the nipples (T4) and umbilicus (T10). Leg weakness and disturbances of bladder and bowel function accompany the paralysis. Lesions at T9-T10 paralyze the lower—but not the upper—abdominal muscles, resulting in upward movement of the umbilicus when the abdominal wall contracts (*Beevor's sign*).

Lumbar Cord

Lesions at the L2-L4 spinal cord levels paralyze flexion and adduction of the thigh, weaken leg extension at the knee, and abolish the patellar reflex. Lesions at L5-S1 paralyze only movements of the foot and ankle, flexion at the knee, and extension of the thigh, and abolish the ankle jerks (S1).

Sacral Cord/Conus Medullaris

The conus medullaris is the tapered caudal termination of the spinal cord, comprising the sacral and single coccygeal segments. The distinctive conus syndrome consists of bilateral saddle anesthesia (S3-S5), prominent bladder and bowel dysfunction (urinary retention and incontinence with lax anal tone), and impotence. The bulbocavernosus (S2-S4) and anal (S4-S5) reflexes are absent **(Chap. 415)**. Muscle strength is largely preserved. By contrast, lesions of the cauda equina, the nerve roots derived from the lower cord, are characterized by low back and radicular pain, asymmetric leg weakness and sensory loss, variable areflexia in the lower extremities, and relative sparing of bowel and bladder function. Mass lesions in the lower spinal canal often produce a mixed clinical picture with elements of both cauda equina and conus medullaris syndromes. **Cauda equina syndromes are also discussed in Chap. 14.**

Special Patterns of Spinal Cord Disease

The location of the major ascending and descending pathways of the spinal cord are shown in **Fig. 434-1**. Most fiber tracts—including the posterior columns and the spinocerebellar and pyramidal tracts—are situated on the side of the body they innervate. However, afferent fibers mediating pain and temperature sensation ascend in the spinothalamic tract contralateral to the side they supply. The anatomic configurations of these tracts produce characteristic syndromes that provide clues to the underlying disease process.

Brown-Sequard Hemicord Syndrome

This consists of ipsilateral weakness (corticospinal tract) and loss of joint position and vibratory sense (posterior column), with contralateral loss of pain and temperature sense (spinothalamic tract) one or two levels below the lesion. Segmental signs, such as radicular pain, muscle atrophy, or loss of a deep tendon reflex, are unilateral. Partial forms are more common than the fully developed syndrome.

Central Cord Syndrome

This syndrome results from selective damage to the gray matter nerve cells and crossing spinothalamic tracts surrounding the central canal. In the cervical cord, the central cord syndrome produces arm weakness out of proportion to leg weakness and a “dissociated” sensory loss, meaning loss of pain and temperature sensations over the shoulders, lower neck, and upper trunk (cape distribution), in contrast to preservation of light touch, joint position, and vibration sense in these regions. Spinal trauma, syringomyelia, and intrinsic cord tumors are the main causes.

Anterior Spinal Artery Syndrome

Infarction of the cord is generally the result of occlusion or diminished flow in this artery. The result is bilateral tissue destruction at several contiguous levels that spares the posterior columns. All spinal cord functions—motor, sensory, and autonomic—are lost below the level of the lesion, with the striking exception of retained vibration and position sensation.

Foramen Magnum Syndrome

Lesions in this area interrupt decussating pyramidal tract fibers destined for the legs, which cross caudal to those of the arms, resulting in weakness of the legs (*crural paresis*). Compressive lesions near the foramen magnum may produce weakness of the ipsilateral shoulder and arm followed by weakness of the ipsilateral leg, then the contralateral leg, and finally the contralateral arm, an “around the clock” pattern that may begin in any of the four limbs. There is typically suboccipital pain spreading to the neck and shoulders.

Intramedullary and Extramedullary Syndromes

It is useful to differentiate *intramedullary* processes, arising within the substance of the cord, from *extramedullary* ones that lie outside the cord and compress the spinal cord or its vascular supply. The differentiating features are only relative and serve as clinical guides. With extramedullary lesions, radicular pain is often prominent, and there is early sacral sensory loss and spastic weakness in the legs with incontinence due to the superficial location of the corresponding sensory and motor fibers in the spinothalamic and corticospinal tracts (Fig. 434-1). Intramedullary lesions tend to produce poorly localized burning pain rather than radicular pain and to spare sensation in the perineal and sacral areas (“sacral sparing”), reflecting the laminated configuration of the spinothalamic tract with sacral fibers outermost; corticospinal tract signs appear later. Regarding extramedullary lesions, a further distinction is made between extradural and intradural masses, as the former are generally malignant and the latter benign (neurofibroma being a common cause). Consequently, a long duration of symptoms favors an intradural origin.

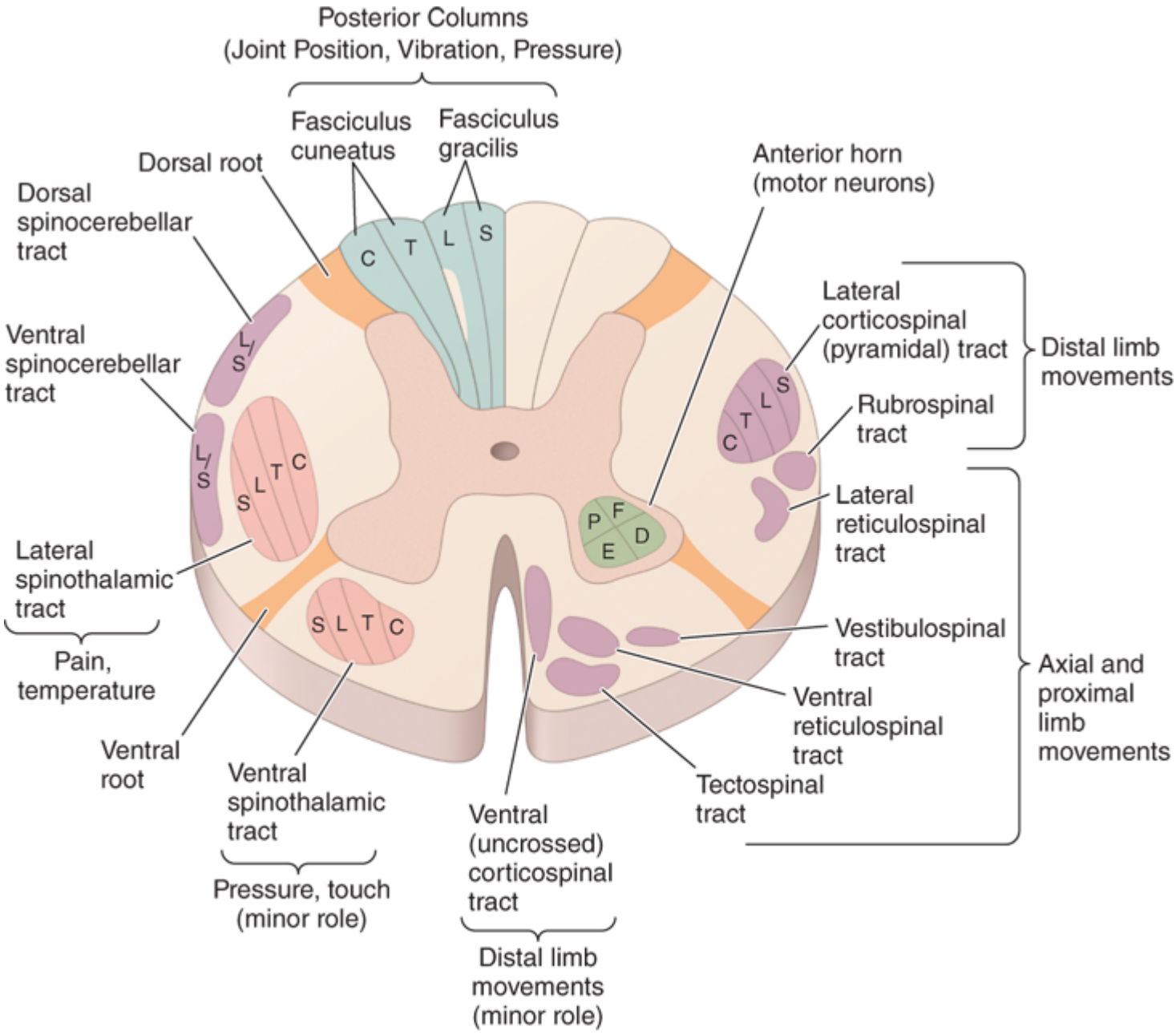
TABLE 434-2

Spinal Cord Levels Relative to the Vertebral Bodies

Spinal Cord Level	Corresponding Vertebral Body
Upper cervical	Same as cord level
Lower cervical	1 level higher
Upper thoracic	2 levels higher
Lower thoracic	2–3 levels higher
Lumbar	T10-T12
Sacral	T12-L1

FIGURE 434-1

Transverse section through the spinal cord, composite representation, illustrating the principal ascending (*left*) and descending (*right*) pathways. The lateral and ventral spinothalamic tracts ascend contralateral to the side of the body that is innervated. C, cervical; D, distal; E, extensors; F, flexors; L, lumbar; P, proximal; S, sacral; T, thoracic.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ACUTE AND SUBACUTE SPINAL CORD DISEASES

Symptoms of the cord diseases that evolve over days or weeks are focal neck or back pain, followed by various combinations of paresthesias, sensory loss, motor weakness, and sphincter disturbance. There may be mild sensory symptoms only or a devastating functional transection of the cord. When paresthesias begin in the feet and then ascend a polyneuropathy is often considered, and in such cases the presence of bladder disturbances and a sharply demarcated spinal cord level provide important clues to the spinal cord origin of the disease.

In severe and abrupt cases, areflexia reflecting spinal shock may be present, but hyperreflexia supervenes over days or weeks; persistent areflexic paralysis with a sensory level usually indicates necrosis over multiple

segments of the spinal cord.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Compressive and Noncompressive Myelopathy

DISTINGUISHING COMPRESSIVE FROM NONCOMPRESSIVE MYELOPATHY

The first priority is to exclude treatable compression of the cord by a mass lesion. The common causes are tumor, epidural abscess or hematoma, herniated disk, and spondylitic vertebral pathology. Epidural compression due to malignancy or abscess often causes warning signs of neck or back pain, bladder disturbances, and sensory symptoms that precede the development of paralysis. Spinal subluxation, hemorrhage, and noncompressive etiologies such as infarction are more likely to produce myelopathy without antecedent symptoms. Magnetic resonance imaging (MRI) with gadolinium, centered on the clinically suspected level, is the initial diagnostic procedure if it is available; it is often appropriate to image the entire spine (cervical through sacral regions) to search for additional clinically silent lesions. Once compressive lesions have been excluded, noncompressive causes of acute myelopathy that are intrinsic to the cord are considered, primarily vascular, inflammatory, and infectious etiologies.

COMPRESSIVE MYELOPATHIES

Neoplastic Spinal Cord Compression

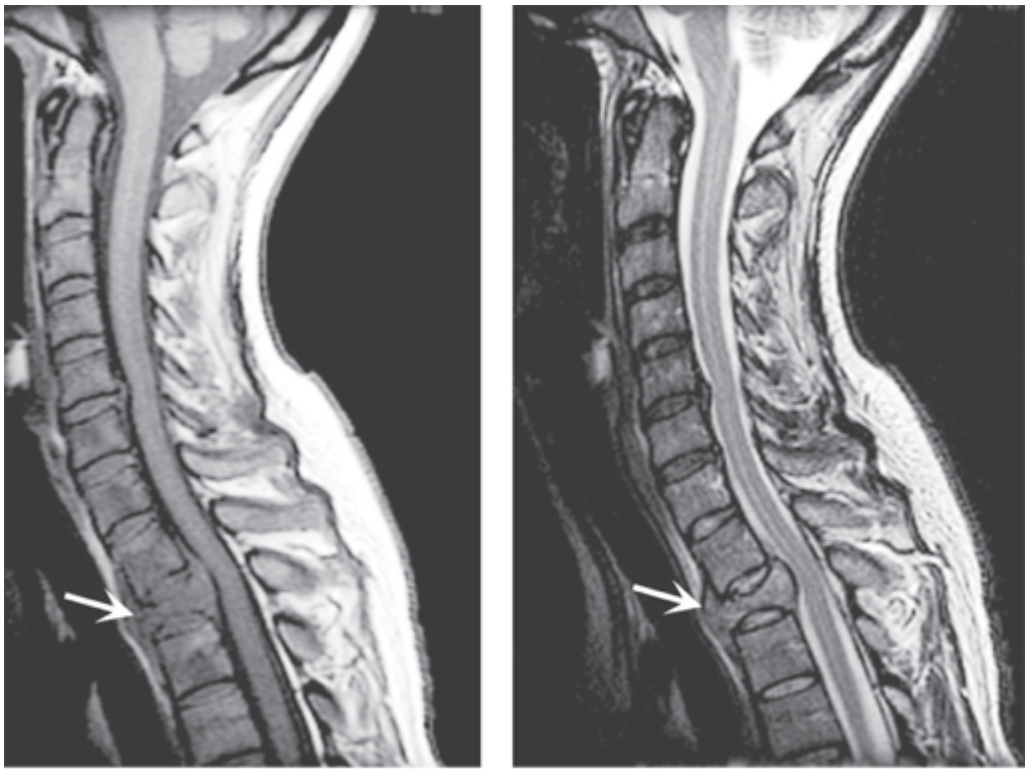
In adults, most neoplasms are epidural in origin, resulting from metastases to the adjacent vertebral column. The propensity of solid tumors to metastasize to the vertebral column probably reflects the high proportion of bone marrow located in the axial skeleton. Almost any malignant tumor can metastasize to the spinal column, with breast, lung, prostate, kidney, lymphoma, and myeloma being particularly frequent. The thoracic spinal column is most commonly involved; exceptions are metastases from prostate and ovarian cancer, which occur disproportionately in the sacral and lumbar vertebrae, probably from spread through Batson's plexus, a network of veins along the anterior epidural space. Retroperitoneal neoplasms (especially lymphomas or sarcomas) enter the spinal canal laterally through the intervertebral foramina and produce radicular pain with signs of weakness that corresponds to the level of involved nerve roots.

Pain is usually the initial symptom of spinal metastasis; it may be aching and localized or sharp and radiating in quality and typically worsens with movement, coughing, or sneezing and characteristically awakens patients at night. A recent onset of persistent back pain, particularly if in the thoracic spine (which is uncommonly involved by spondylosis), should prompt consideration of vertebral metastasis. Rarely, pain is mild or absent. Plain radiographs of the spine and radionuclide bone scans have a limited role in diagnosis because they do not identify 15–20% of metastatic vertebral lesions and fail to detect paravertebral masses that reach the epidural space through the intervertebral foramina. MRI provides excellent anatomic resolution of the extent of spinal tumors **(Fig. 434-2)** and is able to distinguish between malignant lesions and

other masses—epidural abscess, tuberculoma, lipoma, or epidural hemorrhage, among others—that present in a similar fashion. Vertebral metastases are usually hypointense relative to a normal bone marrow signal on T1-weighted MRI; after the administration of gadolinium, contrast enhancement may deceptively "normalize" the appearance of the tumor by increasing its intensity to that of normal bone marrow. Infections of the spinal column (osteomyelitis and related disorders) are distinctive in that, unlike tumor, they often cross the disk space to involve the adjacent vertebral body.

FIGURE 434-2

Epidural spinal cord compression due to breast carcinoma. Sagittal T1-weighted (***A***) and T2-weighted (***B***) magnetic resonance imaging scans through the cervicothoracic junction reveal an infiltrated and collapsed second thoracic vertebral body with posterior displacement and compression of the upper thoracic spinal cord. The low-intensity bone marrow signal in ***A*** signifies replacement by tumor.



A ***B***

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

If spinal cord compression is suspected, imaging should be obtained promptly. If there are radicular symptoms but no evidence of myelopathy, it may be safe to defer imaging for 24–48 h. Up to 40% of patients who present with cord compression at one level are found to have asymptomatic epidural metastases elsewhere; thus, imaging of the entire length of the spine is important to define the extent of disease.

TREATMENT

TREATMENT

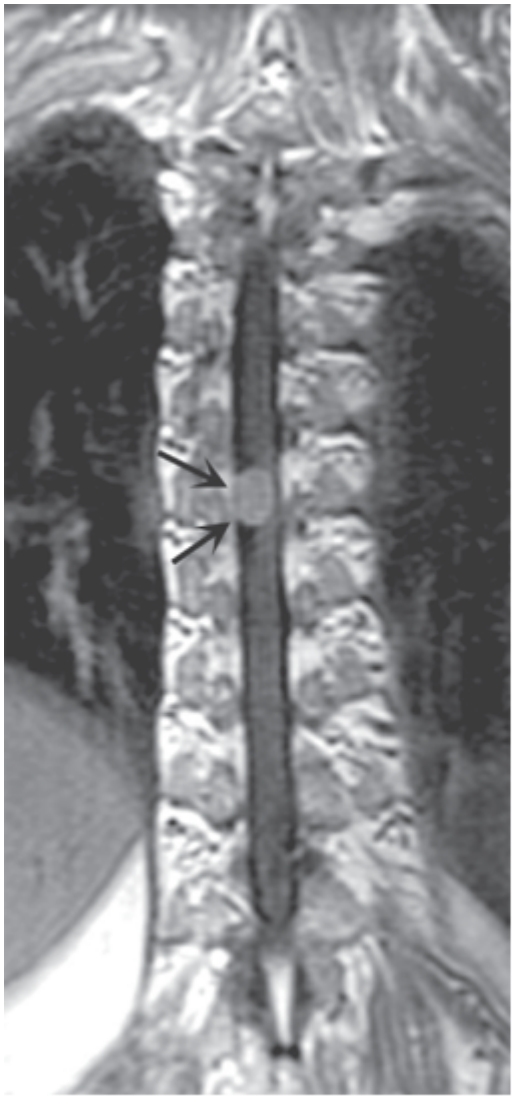
Neoplastic Spinal Cord Compression

Management of cord compression includes glucocorticoids to reduce cord edema, local radiotherapy (initiated as early as possible) to the symptomatic lesion, and specific therapy for the underlying tumor type. Glucocorticoids (typically dexamethasone, 10 mg intravenously) can be administered before an imaging study if there is clinical suspicion of cord compression and continued at a lower dose (4 mg every 6 h orally) until definitive treatment with radiotherapy (generally 30–40 Gy administered in 8–10 fractions) and/or surgical decompression is completed. In one trial, initial management with surgery followed by radiotherapy was more effective than radiotherapy alone for patients with a single area of spinal cord compression by extradural tumor; however, patients with recurrent cord compression, brain metastases, radiosensitive tumors, or severe motor symptoms of >48 h in duration were excluded from this study. Radiotherapy alone may be effective even for some typically radioresistant metastases. A good response to therapy can be expected in individuals who are ambulatory at presentation. Treatment usually prevents new weakness, and some recovery of motor function occurs in up to one-third of patients. Motor deficits (paraplegia or quadriplegia), once established for >12 h, do not usually improve, and beyond 48 h the prognosis for substantial motor recovery is poor. Although most patients do not experience recurrences in the months following radiotherapy, with survival beyond 2 years recurrence becomes increasingly likely and can be managed with additional radiotherapy. Newer techniques such as stereotactic radiosurgery can deliver high doses of focused radiation with similar rates of response compared to traditional radiotherapy, and these are increasingly being used, particularly for patients with traditionally radioresistant tumors or those requiring re-irradiation. Biopsy of the epidural mass is unnecessary in patients with known primary cancer, but it is indicated if a history of underlying cancer is lacking. Surgery, either decompression by laminectomy or vertebral body resection, is also indicated when signs of cord compression worsen despite radiotherapy, when the maximum-tolerated dose of radiotherapy has been delivered previously to the site, or when a vertebral compression fracture or spinal instability contributes to cord compression.

In contrast to tumors of the epidural space, most intradural mass lesions are slow-growing and benign. Meningiomas and neurofibromas account for most of these, with occasional cases caused by chordoma, lipoma, dermoid, or sarcoma. Meningiomas **(Fig. 434-3)** are often located posterior to the thoracic cord or near the foramen magnum, although they can arise from the meninges anywhere along the spinal canal. Neurofibromas are benign tumors of the nerve sheath that typically arise from the posterior root; when multiple, neurofibromatosis is the likely etiology. Symptoms usually begin with radicular sensory symptoms followed by an asymmetric, progressive spinal cord syndrome. Therapy is surgical resection.

FIGURE 434-3

Magnetic resonance imaging of a thoracic meningioma. Coronal T1-weighted postcontrast image through the thoracic spinal cord demonstrates intense and uniform enhancement of a well-circumscribed extramedullary mass (*arrows*) that displaces the spinal cord to the left.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Primary intramedullary tumors of the spinal cord are uncommon. They present as central cord or hemicord syndromes, often in the cervical region. There may be poorly localized burning pain in the extremities and sparing of sacral sensation. In adults, these lesions are ependymomas, hemangioblastomas, or low-grade astrocytomas (Fig. 434-4). Complete resection of an intramedullary ependymoma is often possible with microsurgical techniques. Debulking of an intramedullary astrocytoma can also be helpful, as these are often slowly growing lesions; the value of adjunctive radiotherapy and chemotherapy is uncertain. Secondary (metastatic) intramedullary tumors also occur, especially in patients with advanced metastatic disease (Chap. 86), although these are not nearly as frequent as brain metastases.

FIGURE 434-4

Magnetic resonance imaging of an intramedullary astrocytoma. Sagittal T1-weighted postcontrast image through the cervical spine demonstrates expansion of the upper cervical spine by a mass lesion emanating from within the spinal cord at the cervicomedullary junction. Irregular peripheral enhancement occurs within the mass (*arrows*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Spinal Epidural Abscess

Spinal epidural abscess presents with midline back or neck pain, fever, and progressive limb weakness. Prompt recognition of this distinctive process may prevent permanent sequelae. Aching pain is almost always present, either over the spine or in a radicular pattern. The duration of pain prior to presentation is generally ≤2 weeks but may on occasion be several months or longer. Fever is typically but not invariably present, accompanied by elevated white blood cell count, sedimentation rate, and C-reactive protein. As the abscess expands, further spinal cord damage results from venous congestion and thrombosis. Once weakness and other signs of myelopathy appear, progression may be rapid and irreversible. A more chronic sterile granulomatous form of abscess is also known, usually after treatment of an acute epidural infection.

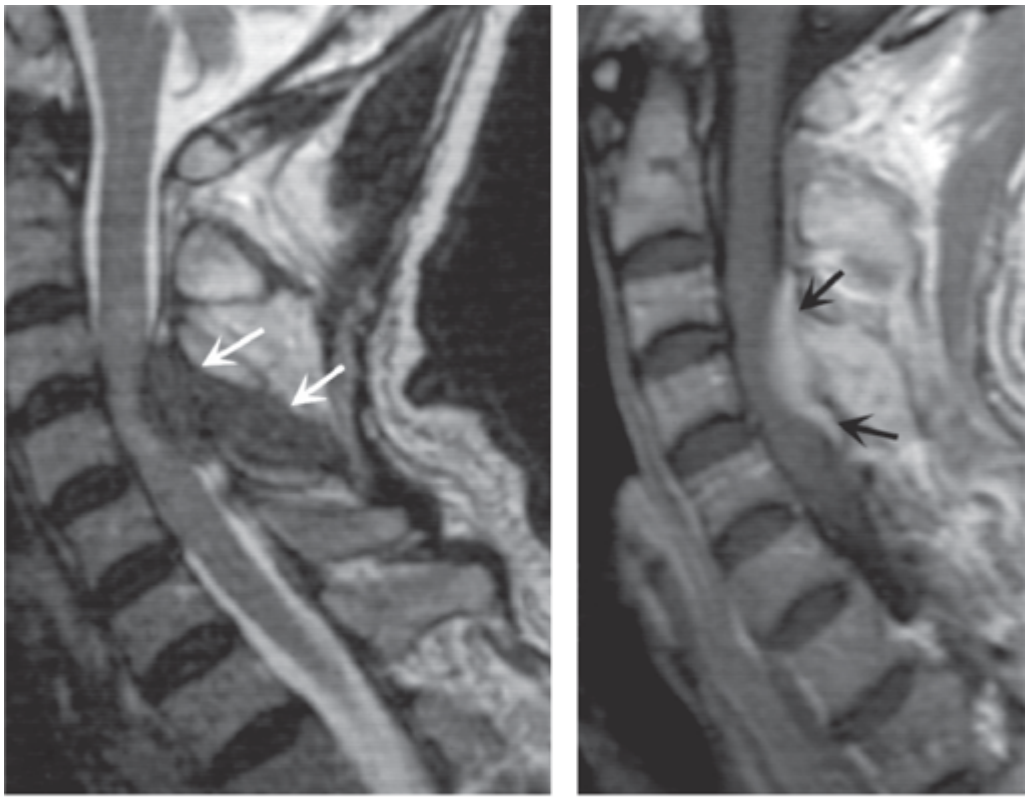
Risk factors include an impaired immune status (HIV, diabetes mellitus, renal failure, alcoholism, malignancy), intravenous drug abuse, and infections of the skin or other tissues. Two-thirds of epidural infections result from hematogenous spread of bacteria from the skin (furunculosis), soft tissue (pharyngeal or dental abscesses; sinusitis), or deep viscera (bacterial endocarditis). The remainder arises from direct extension of a local infection to the subdural space; examples of local predisposing conditions are vertebral osteomyelitis, decubitus ulcers, lumbar puncture, epidural anesthesia, or spinal surgery. Most cases are due to *Staphylococcus aureus*; gram-negative bacilli, *Streptococcus*, anaerobes, and fungi can also cause epidural abscesses. Methicillin resistant *Staphylococcus aureus* (MRSA) is an important consideration, and

therapy should be tailored to this possibility. Tuberculosis from an adjacent vertebral source (Pott's disease) remains an important cause in the developing world.

MRI (Fig. 434-5) localizes the abscess and excludes other causes of myelopathy. Blood cultures are positive in more than half of cases, but direct aspiration of the abscess at surgery is often required for a microbiologic diagnosis. Lumbar puncture is only required if encephalopathy or other clinical signs raise the question of associated meningitis, a feature that is found in <25% of cases. The level of the puncture should be planned to minimize the risk of meningitis due to passage of the needle through infected tissue. A high cervical tap is sometimes the safest approach. Cerebrospinal fluid (CSF) abnormalities in epidural and subdural abscess consist of pleocytosis with a preponderance of polymorphonuclear cells, an elevated protein level, and a reduced glucose level, but the responsible organism is not cultured unless there is associated meningitis.

FIGURE 434-5

Magnetic resonance (MR) imaging of a spinal epidural abscess due to tuberculosis. ***A.*** Sagittal T2-weighted free spin-echo MR sequence. A hypointense mass replaces the posterior elements of C3 and extends epidurally to compress the spinal cord (*arrows*). ***B.*** Sagittal T1-weighted image after contrast administration reveals a diffuse enhancement of the epidural process (*arrows*) with extension into the epidural space.



A *B*

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Spinal Epidural Abscess

Treatment is by decompressive laminectomy with debridement combined with long-term antibiotic treatment. Surgical evacuation prevents development of paralysis and may improve or reverse paralysis in evolution, but it is unlikely to improve deficits of more than several days in duration. Broad-spectrum antibiotics (typically vancomycin 15–20 mg/kg q12h (staphylococcus including MRSA, streptococcus), ceftriaxone 2 gm q24h (gram-negative bacilli), and when indicated metronidazole 30 mg/kg/d divided into q6h intervals (anaerobes) should be started empirically before surgery and then modified on the basis of culture results; medication is generally continued for 6–8 weeks. If surgery is contraindicated or if there is a fixed paraplegia or quadriplegia that is unlikely to improve following surgery, long-term administration of systemic and oral antibiotics can be used; in such cases, the choice of antibiotics may be guided by results of blood cultures. Surgical management remains the treatment of choice unless the abscess is limited in size and causes few or no neurologic signs.

With prompt diagnosis and treatment of spinal epidural abscess, up to two-thirds of patients experience significant recovery.

Spinal Epidural Hematoma

Hemorrhage into the epidural (or subdural) space causes acute focal or radicular pain followed by variable signs of a spinal cord or conus medullaris disorder. Therapeutic anticoagulation, trauma, tumor, or blood dyscrasias are predisposing conditions. Rare cases complicate lumbar puncture or epidural anesthesia. MRI and computed tomography (CT) confirm the clinical suspicion and can delineate the extent of the bleeding. Treatment consists of prompt reversal of any underlying clotting disorder and surgical decompression. Surgery may be followed by substantial recovery, especially in patients with some preservation of motor function preoperatively. Because of the risk of hemorrhage, lumbar puncture should be avoided whenever possible in patients with severe thrombocytopenia or other coagulopathies.

Hematomyelia

Hemorrhage into the substance of the spinal cord is a rare result of trauma, intraparenchymal vascular malformation (see below), vasculitis due to polyarteritis nodosa or systemic lupus erythematosus (SLE), bleeding disorders, or a spinal cord neoplasm. Hematomyelia presents as an acute painful transverse myelopathy. With large lesions, extension into the subarachnoid space results in subarachnoid hemorrhage (Chap. 302). Diagnosis is by MRI or CT. Therapy is supportive, and surgical intervention is generally not useful. An exception is hematomyelia due to an underlying vascular malformation, for which spinal angiography and endovascular occlusion may be indicated, or surgery to evacuate the clot and remove the underlying vascular lesion.

NONCOMPRESSIVE MYELOPATHIES

The most frequent causes of noncompressive acute transverse myelopathy are spinal cord infarction; systemic inflammatory disorders, including SLE and sarcoidosis; demyelinating diseases, including multiple sclerosis (MS); neuromyelitis optica (NMO); postinfectious or idiopathic transverse myelitis, which is presumed to be an immune condition related to acute disseminated encephalomyelitis **(Chap. 436)**; and infectious (primarily viral) causes. After spinal cord compression is excluded, the evaluation generally requires a lumbar puncture and a search for underlying systemic disease **(Table 434-3)**.

TABLE 434-3

Evaluation of Acute Transverse Myelopathy

1. MRI of spinal cord with and without contrast (exclude compressive causes).
2. CSF studies: Cell count, protein, glucose, IgG index/synthesis rate, oligoclonal bands, VDRL; Gram's stain, acid-fast bacilli, and India ink stains; PCR for VZV, HSV-2, HSV-1, EBV, CMV, HHV-6, enteroviruses, HIV; antibody for HTLV-1, *Borrelia burgdorferi*, *Mycoplasma pneumoniae*, and *Chlamydia pneumoniae*; viral, bacterial, mycobacterial, and fungal cultures.
3. Blood studies for infection: HIV; RPR; IgG and IgM enterovirus antibody; IgM mumps, measles, rubella, group B arbovirus, *Brucella melitensis*, *Chlamydia psittaci*, *Bartonella henselae*, schistosomal antibody; cultures for *B. melitensis*. Also consider nasal/pharyngeal/anal cultures for enteroviruses; stool O&P for *Schistosoma* ova.
4. Systemic immune-mediated disorders: ESR; ANA; ENA; dsDNA; rheumatoid factor; anti-SSA; anti-SSB, complement levels; antiphospholipid and anticardiolipin antibodies; p-ANCA; antimicrosomal and antithyroglobulin antibodies; if Sjögren's syndrome suspected, Schirmer test, salivary gland scintigraphy, and salivary/lacrimal gland biopsy.
5. Neuromyelitis Optica: Serum anti-aquaporin-4 antibody, anti-MOG antibody.
6. Demyelinating disease: Brain MRI scan; evoked potentials.
7. Sarcoidosis: Serum angiotensin-converting enzyme; serum Ca; 24-h urine Ca; chest x-ray; chest CT; total-body gallium scan; lymph node biopsy.
8. Vascular causes: MRI, CT myelogram; spinal angiogram.

Abbreviations: ANA, antinuclear antibodies; CMV, cytomegalovirus; CSF, cerebrospinal fluid; CT, computed tomography; EBV, Epstein-Barr virus; ENA, epithelial neutrophil-activating peptide; ESR, erythrocyte sedimentation rate; HHV, human herpes virus; HSV, herpes simplex virus; HTLV, human T cell leukemia/lymphoma virus; MOG, myelin oligodendrocyte glycoprotein; MRI, magnetic resonance imaging; O&P, ova and parasites; p-ANCA, perinuclear antineutrophilic cytoplasmic antibodies; PCR, polymerase chain reaction; RPR, rapid plasma reagin (test); VDRL, Venereal Disease Research Laboratory; VZV, varicella-zoster virus.

Spinal Cord Infarction

The cord is supplied by three arteries that course vertically over its surface: a single anterior spinal artery and paired posterior spinal arteries. The anterior spinal artery originates in paired branches of the vertebral arteries at the cranciocervical junction and is fed by additional radicular vessels that arise at C6, at an upper thoracic level, and, most consistently, at T11-L2 (artery of Adamkiewicz). At each spinal cord segment, paired penetrating vessels branch from the anterior spinal artery to supply the anterior two-thirds of the cord; the posterior spinal arteries, which often become less distinct below the midthoracic level, supply the posterior columns.

Spinal cord ischemia can occur at any level; however, the presence of the artery of Adamkiewicz below, and the anterior spinal artery circulation above, creates a region of marginal blood flow in the upper thoracic segments. With hypotension or cross-clamping of the aorta, cord infarction typically occurs at the level of T3-T4, and also at boundary zones between the anterior and posterior spinal artery territories. The latter may result in a rapidly progressive syndrome over hours of weakness and spasticity with little sensory change.

Acute infarction in the territory of the *anterior spinal artery* produces paraplegia or quadriplegia, dissociated sensory loss affecting pain and temperature sense but sparing vibration and position sense, and loss of sphincter control ("anterior cord syndrome"). Onset may be sudden but more typically is progressive over minutes or a few hours, unlike stroke in the cerebral hemispheres. Sharp midline or radiating back pain localized to the area of ischemia is frequent. Areflexia due to spinal shock is often present initially; with time, hyperreflexia and spasticity appear. Less common is infarction in the territory of the *posterior spinal arteries*, resulting in loss of posterior column function either on one side or bilaterally.

Causes of spinal cord infarction include aortic atherosclerosis, dissecting aortic aneurysm, vertebral artery occlusion or dissection in the neck, aortic surgery, or profound hypotension from any cause. A *surfer's myelopathy* usually in the thoracic region, has been associated with prolonged back extension due to lifting the upper body off the board while waiting for waves; it, typically manifests as back pain followed by an anterior cord syndrome with progressive paralysis and loss of sphincter control, and is likely vascular in origin. Cardiogenic emboli, vasculitis **(Chap. 356)**, and collagen vascular disease (particularly SLE **[Chap. 349]**, Sjögren's syndrome **[Chap. 354]**, and the antiphospholipid antibody syndrome **[Chap. 350]**) are other etiologies. Occasional cases develop from *embolism of nucleus pulposus* material into spinal vessels, usually from local spine trauma. In a substantial number of cases, no cause can be found, and thromboembolism in arterial feeders is suspected. MRI may fail to demonstrate infarctions of the cord, especially in the first day, but often the imaging becomes abnormal at the affected level.

In cord infarction due to presumed thromboembolism, acute anticoagulation is not indicated, with the possible exception of the unusual transient ischemic attack or incomplete infarction with a stuttering or progressive course. The antiphospholipid antibody syndrome is treated with anticoagulation **(Chap. 350)**. Increasing systemic blood pressure to a mean arterial pressure of >90 mmHg, or lumbar drainage of spinal fluid, was reportedly helpful in a few published cases of cord infarction, but neither of these approaches has been studied systematically. Prognosis following spinal cord infarction is influenced by the severity of the deficits at presentation; patients with severe motor weakness and those with persistent areflexia usually do

poorly, but in one recent large series some improvement over time occurred in many patients, with more than half ultimately regaining some ambulation.

Inflammatory and Immune Myelopathies (Myelitis)

This broad category includes the demyelinating conditions MS, NMO, and postinfectious myelitis, as well as sarcoidosis and systemic autoimmune disease. In approximately one-quarter of cases of myelitis, no underlying cause can be identified. Some will later manifest additional symptoms of an immune-mediated disease. *Recurrent episodes of myelitis* are usually due to one of the immune-mediated diseases or to infection with herpes simplex virus (HSV) type 2 (below).

MULTIPLE SCLEROSIS

MS may present with acute myelitis, particularly in individuals of Asian or African ancestry. In Caucasians, MS attacks rarely cause a transverse myelopathy (i.e., attacks of bilateral sensory disturbances, unilateral or bilateral weakness, and bladder or bowel symptoms), but it is among the most common causes of a partial cord syndrome. MRI findings in MS-associated myelitis typically consist of mild swelling of the cord and diffuse or multifocal "shoddy" areas of abnormal signal on T2-weighted sequences. Contrast enhancement, indicating disruption in the blood-brain barrier associated with inflammation, is present in many acute cases. In one study 68% of patients presenting with partial myelitis developed MS after a mean follow-up of 4 years; risk factors for conversion to MS included age <40 years; inflammatory CSF, and >3 periventricular lesions on brain MRI.

Treatment of acute episodes of MS-associated myelitis consists of intravenous methylprednisolone (500 mg qd for 3 days) followed by oral prednisone (1 mg/kg/d for several weeks, then gradual taper). A course of plasma exchange may be indicated for severe cases if glucocorticoids are ineffective. **MS is discussed in Chap. 436.**

NEUROMYELITIS OPTICA

NMO is an immune-mediated demyelinating disorder consisting of a severe myelopathy that is typically longitudinally extensive, meaning that the lesion spans three or more vertebral segments. NMO is associated with optic neuritis that is often bilateral and may precede or follow myelitis by weeks or months, and also by brainstem and, in some cases, hypothalamic or focal cerebral white matter involvement. Recurrent myelitis without optic nerve or other involvement can also occur in NMO. CSF studies reveal a variable mononuclear pleocytosis of up to several hundred cells per microliter; unlike MS, oligoclonal bands are generally absent. Diagnostic serum autoantibodies against the water channel protein aquaporin-4 are present in 60–70% of patients with NMO, and less commonly autoantibodies against the CNS myelin protein myelin oligodendrocyte glycoprotein (MOG) are found. NMO has also been associated with SLE (see below) as well as with other systemic autoimmune diseases; rare cases are paraneoplastic in origin. There have been no definitive trials of therapy for NMO. Recommended treatment of acute relapses is with glucocorticoids and, for refractory cases, plasma exchange. Prophylactic treatment with azathioprine, mycophenylate, or rituximab may protect against subsequent relapses; treatment for 5 years or longer is generally recommended. **NMO is discussed in Chap. 437.**

SYSTEMIC IMMUNE-MEDIATED DISORDERS

Myelitis occurs in a small number of patients with SLE, many cases of which are associated with antibodies to aquaporin-4 and satisfy diagnostic criteria for NMO-spectrum disorder (Chap. 437). These patients are at high risk of developing future episodes of myelitis and/or optic neuritis. In others the etiology of SLE-associated myelitis is uncertain; anti-phospholipid antibodies have been suggested to play a role, however the presence of these antibodies appears to be no more frequent in SLE patients with and without myelitis. The CSF in NMO-associated myelitis typically shows a pleocytosis with polymorphonuclear leukocytes and no oligoclonal bands; in cases not due to NMO a mild lymphocytic pleocytosis and oligoclonal bands are variable findings. Although there are no systematic trials of therapy for SLE myelitis, based on limited data, high-dose glucocorticoids followed by cyclophosphamide have been recommended. Acute severe episodes of transverse myelitis that do not initially respond to glucocorticoids are often treated with a course of plasma exchange. Sjögren's syndrome (Chap. 354) can also be associated with NMO-spectrum disorder and also with cases of chronic progressive myelopathy. Other immune-mediated myelitides include antiphospholipid antibody syndrome (Chap. 350), mixed connective tissue disease (Chap. 353), Behçet's syndrome (Chap. 357), and vasculitis related to polyarteritis nodosa, perinuclear antineutrophilic cytoplasmic (p-ANCA) antibodies, or primary central nervous system vasculitis (Chap. 356).

Another important consideration in this group is sarcoid myelopathy that may present as a slowly progressive or relapsing disorder. MRI reveals an edematous swelling of the spinal cord that may mimic tumor; there is almost always gadolinium enhancement of active lesions and in some cases nodular enhancement of the adjacent surface of the cord; lesions may be single or multiple, and on axial images, enhancement of the central cord is often present. The typical CSF profile consists of a mild lymphocytic pleocytosis and mildly elevated protein level; in a minority of cases, reduced glucose and oligoclonal bands are found. The diagnosis is particularly difficult when systemic manifestations of sarcoid are minor or absent (nearly 50% of cases) or when other typical neurologic manifestations of the disease—such as cranial neuropathy, hypothalamic involvement, or meningeal enhancement visualized by MRI—are lacking. A slit-lamp examination of the eye to search for uveitis, chest x-ray and CT to assess pulmonary involvement and mediastinal lymphadenopathy, serum or CSF angiotensin-converting enzyme (ACE; CSF values elevated in only a minority of cases), serum calcium, and a gallium scan may assist in the diagnosis. The usefulness of spinal fluid ACE is uncertain. Initial treatment is with oral glucocorticoids; immunosuppressant drugs, including the tumor necrosis factor α inhibitor infliximab, have been used for resistant cases. **Sarcoidosis is discussed in Chap. 360.**

POSTINFECTIOUS MYELITIS

Many cases of myelitis, termed *postinfectious* or *postvaccinal*, follow an infection or vaccination. Numerous organisms have been implicated, including Epstein-Barr virus (EBV), cytomegalovirus (CMV), mycoplasma, influenza, measles, varicella, mumps, and yellow fever. As in the related disorder acute disseminated encephalomyelitis (Chap. 436), postinfectious myelitis often begins as the patient appears to be recovering from an acute febrile infection, or in the subsequent days or weeks, but an infectious agent cannot be isolated from the nervous system or CSF. The presumption is that the myelitis represents an autoimmune disorder triggered by infection and is not due to direct infection of the spinal cord. No randomized controlled trials of therapy exist; treatment is usually with glucocorticoids or, in fulminant cases, plasma exchange.

ACUTE INFECTIOUS MYELITIS

Many viruses have been associated with an acute myelitis that is infectious in nature rather than postinfectious. Nonetheless, the two processes are often difficult to distinguish. Herpes zoster is the best characterized viral myelitis, but HSV types 1 and 2, EBV, CMV, and rabies virus are other well-described causes and Zika virus has also been recognized as a cause of infectious myelitis. HSV-2 (and less commonly HSV-1) produces a distinctive syndrome of recurrent sacral cauda equina neuritis in association with outbreaks of genital herpes (Elsberg's syndrome). Poliomyelitis is the prototypic viral myelitis, but it is more or less restricted to the anterior gray matter of the cord containing the spinal motoneurons. A polio-like syndrome can also be caused by a large number of enteroviruses (including enterovirus 71 and coxsackie), and with Japanese encephalitis and other flaviviruses such as West Nile virus. Recently, cases of paralysis in children and adolescents were associated with enterovirus D-68 infection but a causal role for this virus has not been established. Chronic viral myelitic infections, such as those due to HIV or human T cell lymphotropic virus type 1 (HTLV-1), are discussed below.

Bacterial and mycobacterial myelitis (most are essentially abscesses) are less common than viral causes and much less frequent than cerebral bacterial abscess. Almost any pathogenic species may be responsible, including *Borrelia burgdorferi* (Lyme disease), *Listeria monocytogenes, Mycobacterium tuberculosis,* and *Treponema pallidum* (syphilis). *Mycoplasma pneumoniae* may be a cause of myelitis, but its status is uncertain because many cases are more properly classified as postinfectious.

Schistosomiasis **(Chap. 229)** is an important cause of parasitic myelitis in endemic areas. The process is intensely inflammatory and granulomatous, caused by a local response to tissue-digesting enzymes from the ova of the parasite, typically *Schistosoma hematobium* or *Schistosoma mansoni*. Toxoplasmosis **(Chap. 223)** can occasionally cause a focal myelopathy, and this diagnosis should especially be considered in patients with AIDS **(Chap. 197)**. *Cysticercosis* **(Chap. 230)** is another consideration, although myelitis from this helminth is far less common than parenchymal brain or meningeal involvement.

In cases of suspected viral myelitis, it may be appropriate to begin specific therapy pending laboratory confirmation. Herpes zoster, HSV, and EBV myelitis are treated with intravenous acyclovir (10 mg/kg q8h) or oral valacyclovir (2 g tid) for 10–14 days; CMV is treated with ganciclovir (5 mg/kg IV bid) plus foscarnet (60 mg/kg IV tid) or cidofovir (5 mg/kg per week for 2 weeks).

High-Voltage Electrical Injury

Spinal cord injuries are prominent following electrocution from lightning strikes or other accidental electrical exposures. The syndrome consists of transient weakness acutely (often with an altered sensorium and focal cerebral disturbances), sometimes followed several days or even weeks later by a myelopathy that can be severe and permanent. This is a rare injury type, and limited data incriminate a vascular pathology involving the anterior spinal artery and its branches in some cases. Therapy is supportive.

CHRONIC MYELOPATHIES

SPONDYLOTIC MYELOPATHY

Spondylotic myelopathy is one of the most common causes of chronic cord compression and of gait difficulty in the elderly. Neck and shoulder pain with stiffness are early symptoms; impingement of bone and soft tissue overgrowth on nerve roots results in radicular arm pain, most often in a C5 or C6 distribution. Compression of the cervical cord, which occurs in fewer than one-third of cases, produces a slowly progressive spastic paraparesis, at times asymmetric and often accompanied by paresthesias in the feet and hands. Vibratory sense is diminished in the legs, there is a Romberg sign, and occasionally there is a sensory level for vibration or pinprick on the upper thorax. In some cases, coughing or straining produces leg weakness or radiating arm or shoulder pain. Dermatomal sensory loss in the arms, atrophy of intrinsic hand muscles, increased deep-tendon reflexes in the legs, and extensor plantar responses are common. Urinary urgency or incontinence occurs in advanced cases, but there are many alternative causes of these problems in older individuals. A tendon reflex in the arms is often diminished at some level; most often at the biceps (C5-C6). In individual cases, radicular, myelopathic, or combined signs may predominate. The diagnosis should be considered in appropriate cases of progressive cervical myelopathy, paresthesias of the feet and hands, or wasting of the hands.

Diagnosis is usually made by MRI and may be suspected from CT images; plain x-rays are less helpful. Extrinsic cord compression and deformation are appreciated on axial MRI views, and T2-weighted sequences may reveal areas of high signal intensity within the cord adjacent to the site of compression. A cervical collar may be helpful in milder cases, but the likelihood of progression of medically-treated myelopathy is high, estimated at 8% over 1 year. Definitive therapy consists of surgical decompression, either posterior laminectomy or an anterior approach with resection of the protruded disk and bony material. **Cervical spondylosis and related degenerative diseases of the spine are discussed in Chap. 14.**

VASCULAR MALFORMATIONS OF THE CORD AND DURA

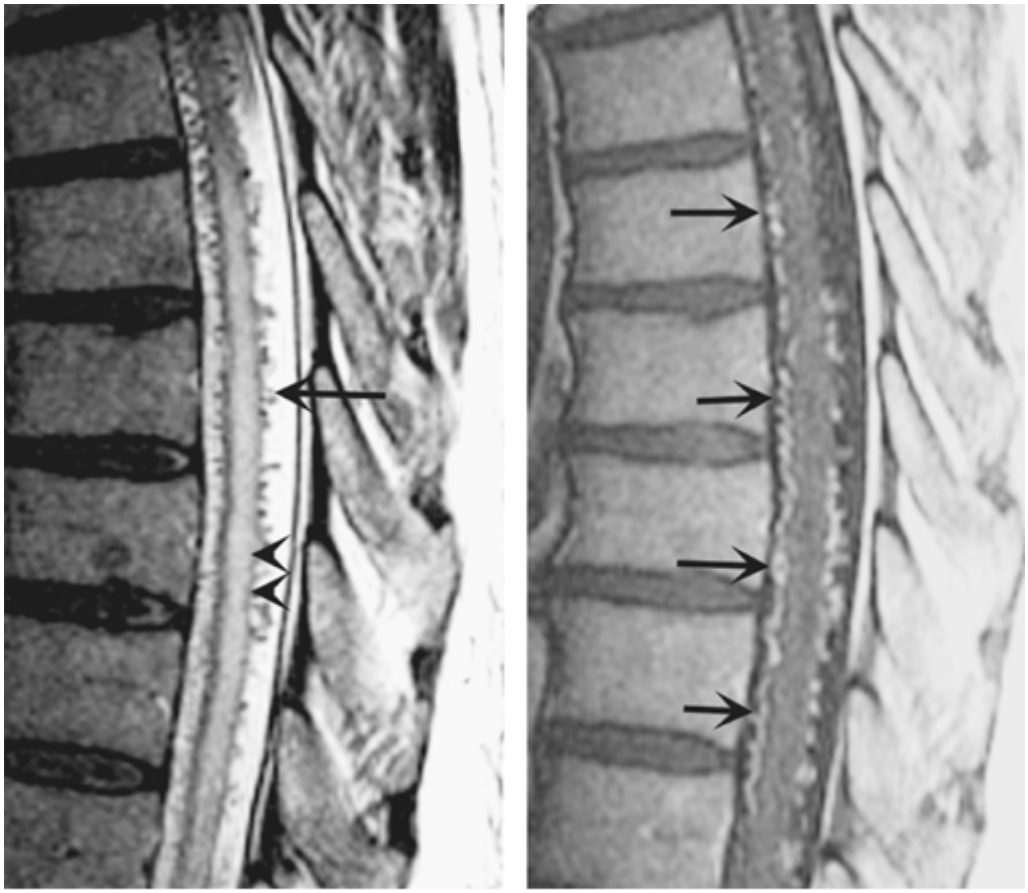
Vascular malformations, comprising ~4% of all mass lesions of the cord and overlying dura, are treatable causes of progressive myelopathy. Most common are fistulas located within the dura or posteriorly along the surface of the cord. Most dural arteriovenous (AV) fistulas are located at or below the midthoracic level, usually consisting of a direct connection between a radicular feeding artery in the nerve root sleeve with dural veins. The typical presentation is a middle-aged man with a progressive myelopathy that worsens slowly or intermittently and may have periods of remission, sometimes mimicking MS. Acute deterioration due to hemorrhage into the spinal cord (hematomyelia) or subarachnoid space may also occur but is rare. In many cases, progression results from local ischemia and edema due to venous congestion. Most patients have incomplete sensory, motor, and bladder disturbances. The motor disorder may predominate and produce a mixture of upper and restricted lower motor neuron signs, simulating amyotrophic lateral sclerosis (ALS). Pain over the dorsal spine, dysesthesias, or radicular pain may be present. Other symptoms suggestive of AV malformation (AVM) or dural fistula include intermittent claudication; symptoms that change with posture, exertion, Valsalva maneuver, or menses; and fever.

Less commonly, AVM disorders are intramedullary rather than dural. One unusual disorder is a progressive thoracic myelopathy with paraparesis developing over weeks or months, characterized pathologically by abnormally thick, hyalinized vessels within the cord (subacute necrotic myelopathy, or Foix-Alajouanine syndrome).

Spinal bruits are infrequent but may be sought at rest and after exercise in suspected cases. A vascular nevus on the overlying skin may indicate an underlying vascular malformation as occurs with Klippel-Trenaunay-Weber syndrome. MR angiography and CT angiography can detect the draining vessels of many AVMs **(Fig. 434-6)**. Definitive diagnosis requires selective spinal angiography, which defines the feeding vessels and the extent of the malformation. Treatment is tailored to the anatomy and location of the lesion, and generally consists of microsurgical resection, endovascular embolization of the major feeding vessels, or a combination of the two approaches.

FIGURE 434-6

Arteriovenous malformation. Sagittal magnetic resonance scans of the thoracic spinal cord: T2 fast spin-echo technique (*left*) and T1 postcontrast image (*right*). On the T2-weighted image (*left*), abnormally high signal intensity is noted in the central aspect of the spinal cord (*arrowheads*). Numerous punctate flow voids indent the dorsal and ventral spinal cord (*arrow*). These represent the abnormally dilated venous plexus supplied by a dural arteriovenous fistula. After contrast administration (*right*), multiple, serpentine, enhancing veins (*arrows*) on the ventral and dorsal aspect of the thoracic spinal cord are visualized, diagnostic of arteriovenous malformation. This patient was a 54-year-old man with a 4-year history of progressive paraparesis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

RETROVIRUS-ASSOCIATED MYELOPATHIES

The myelopathy associated with HTLV-1, formerly called tropical spastic paraparesis, is a slowly progressive spastic syndrome with variable sensory and bladder disturbance. Approximately half of patients have mild back or leg pain. The neurologic signs may be asymmetric, often lacking a well-defined sensory level; the only sign in the arms may be hyperreflexia after several years of illness. The onset is usually insidious, and the tempo of progression of the illness occurs at a variable rate; in one study, median time for progression to cane, walker, or wheelchair dependent state was 6, 13, and 21 years, respectively. Progression appears to be more rapid in older patients and those with higher viral loads. Diagnosis is made by demonstration of HTLV-1-specific antibody in serum by enzyme-linked immunosorbent assay (ELISA), confirmed by radioimmunoprecipitation or Western blot analysis. Especially in endemic areas, a finding of HTLV-1 seropositivity in a patient with myelopathy does not necessarily prove that HTLV-1 is causative. The CSF/serum antibody index may provide support by establishing intrathecal synthesis of antibodies, including oligoclonal antibodies, favoring HTVL-1 myelopathy over asymptomatic carriage. Measuring proviral DNA by polymerase chain reaction (PCR) in serum and CSF cells can be useful as an ancillary part of diagnosis. The pathogenesis of the myelopathy is uncertain. It could result from an immune response directed against HTLV-1 antigens in the nervous system, or alternatively to secondary autoimmunity triggered by the viral infection. There is no proven effective treatment. Based on limited evidence, the use of chronic low dose oral

glucocorticoids can be tried; interferon is of uncertain value, and antiviral treatment is ineffective. Symptomatic therapy for spasticity and bladder symptoms may be helpful.

A progressive myelopathy can also result from HIV infection **(Chap. 197)**. It is characterized by vacuolar degeneration of the posterior and lateral tracts, resembling subacute combined degeneration (see below).

SYRINGOMYELIA

Syringomyelia is a developmental cavity of the cervical cord that may enlarge and produce progressive myelopathy or may remain asymptomatic. Symptoms begin insidiously in adolescence or early adulthood, progress irregularly, and may undergo spontaneous arrest for several years. Many young patients acquire a cervical-thoracic scoliosis. More than half of all cases are associated with Chiari type 1 malformations in which the cerebellar tonsils protrude through the foramen magnum and into the cervical spinal canal. The pathophysiology of syrinx expansion is controversial, but some interference with the normal flow of CSF seems likely, perhaps by the Chiari malformation. Acquired cavitations of the cord in areas of necrosis are also termed *syrinx cavities*; these follow trauma, myelitis, necrotic spinal cord tumors, and chronic arachnoiditis due to tuberculosis and other etiologies.

The presentation is a central cord syndrome consisting of a regional dissociated sensory loss (loss of pain and temperature sensation with sparing of touch and vibration) and areflexic weakness in the upper limbs. The sensory deficit has a distribution that is “suspended” over the nape of the neck, shoulders, and upper arms (cape distribution) or in the hands. Most cases begin asymmetrically with unilateral sensory loss in the hands that leads to injuries and burns that are not appreciated by the patient. Muscle wasting in the lower neck, shoulders, arms, and hands with asymmetric or absent reflexes in the arms reflects expansion of the cavity in the gray matter of the cord. As the cavity enlarges and compresses the long tracts, spasticity and weakness of the legs, bladder and bowel dysfunction, and a Horner’s syndrome appear. Some patients develop facial numbness and sensory loss from damage to the descending tract of the trigeminal nerve (C2 level or above). In cases with Chiari malformations, cough-induced headache and neck, arm, or facial pain may be reported. Extension of the syrinx into the medulla, syringobulbia, causes palatal or vocal cord paralysis, dysarthria, horizontal or vertical nystagmus, episodic dizziness or vertigo, and tongue weakness with atrophy.

MRI accurately identifies developmental and acquired syrinx cavities and their associated spinal cord enlargement **(Fig. 434-7)**. Images of the brain and the entire spinal cord should be obtained to delineate the full longitudinal extent of the syrinx, assess posterior fossa structures for the Chiari malformation, and determine whether hydrocephalus is present.

FIGURE 434-7

Magnetic resonance imaging of syringomyelia associated with a Chiari malformation. Sagittal T1-weighted image through the cervical and upper thoracic spine demonstrates descent of the cerebellar tonsils below the level of the foramen magnum (*black arrows*). Within the substance of the cervical and thoracic spinal cord, a cerebrospinal fluid collection dilates the central canal (*white arrows*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Syringomyelia

Treatment of syringomyelia is generally unsatisfactory. The Chiari tonsillar herniation may be decompressed, generally by suboccipital craniectomy, upper cervical laminectomy, and placement of a dural graft. Fourth ventricular outflow is reestablished by this procedure. If the syrinx cavity is large, some surgeons recommend direct decompression or drainage, but the added benefit of this procedure is uncertain, and complications are common. With Chiari malformations, shunting of hydrocephalus generally precedes any attempt to correct the syrinx. Surgery may stabilize the neurologic deficit, and some patients improve. Patients with few symptoms and signs from the syrinx do not require surgery and are followed by serial clinical and imaging examinations.

Syrinx cavities secondary to trauma or infection, if symptomatic, are treated with a decompression and drainage procedure in which a small shunt is inserted between the cavity and subarachnoid space; alternatively, the cavity can be fenestrated. Cases due to intramedullary spinal cord tumor are generally managed by resection of the tumor.

CHRONIC MYELOPATHY OF MULTIPLE SCLEROSIS

A chronic progressive myelopathy is the most frequent cause of disability in both primary progressive and secondary progressive forms of MS. Involvement is typically bilateral but asymmetric and produces motor, sensory, and bladder/bowel disturbances. Fixed motor disability appears to result from extensive loss of axons in the corticospinal tracts. Diagnosis is facilitated by identification of earlier attacks such as optic neuritis. MRI, CSF, and evoked response testing are confirmatory. Treatment with ocrelizumab, an anti-CD20 B-cell monoclonal antibody, is effective in patients with primary progressive MS, and disease modifying therapy is also indicated in patients with secondary progressive MS who have coexisting MS relapses. **MS is discussed in Chap. 436.**

SUBACUTE COMBINED DEGENERATION (VITAMIN B_{12} DEFICIENCY)

This treatable myelopathy presents with subacute paresthesias in the hands and feet, loss of vibration and position sensation, and a progressive spastic and ataxic weakness. Loss of reflexes due to an associated peripheral neuropathy in a patient who also has Babinski signs is an important diagnostic clue. Optic atrophy and irritability or other cognitive changes may be prominent in advanced cases and are occasionally the presenting symptoms. The myelopathy of subacute combined degeneration tends to be diffuse rather than focal; signs are generally symmetric and reflect predominant involvement of the posterior and lateral tracts, including Romberg's sign. Causes include dietary deficiency, especially in vegans, and gastric malabsorption syndromes including pernicious anemia (Chap. 95). The diagnosis is confirmed by the finding of macrocytic red blood cells, a low serum B_{12} concentration, and elevated serum levels of homocysteine and methylmalonic acid. Treatment is by replacement therapy, beginning with 1000 μg of intramuscular vitamin B_{12} repeated at regular intervals or by subsequent oral treatment.

HYPOCUPRIC MYELOPATHY

This myelopathy is similar to subacute combined degeneration (described above), except there is no neuropathy, and explains cases with normal serum levels of B_{12}. Low levels of serum copper are found, and often there is also a low level of serum ceruloplasmin. Some cases follow gastrointestinal procedures, particularly bariatric surgery, that result in impaired copper absorption; others have been associated with excess zinc from health food supplements or in the past zinc-containing denture creams, all of which impair copper absorption via induction of metallothionein, a copper-binding protein. Many cases are idiopathic. Improvement or at least stabilization may be expected with reconstitution of copper stores by oral supplementation. There is microcytic or macrocytic anemia. The pathophysiology and pathology of the idiopathic form are not known.

TABES DORSALIS

The classic syphilitic syndromes of tabes dorsalis and meningovascular inflammation of the spinal cord are now less frequent than in the past but must be considered in the differential diagnosis of spinal cord

disorders. The characteristic symptoms of tabes are fleeting and repetitive lancinating pains, primarily in the legs or less often in the back, thorax, abdomen, arms, and face. Ataxia of the legs and gait due to loss of position sense occurs in half of patients. Paresthesias, bladder disturbances, and acute abdominal pain with vomiting (visceral crisis) occur in 15–30% of patients. The cardinal signs of tabes are loss of reflexes in the legs; impaired position and vibratory sense; Romberg sign; and, in almost all cases, bilateral Argyll Robertson pupils, which fail to constrict to light but accommodate. Diabetic polyradiculopathy may simulate this condition. Treatment of tabes dorsalis and other forms of neurosyphilis consists of penicillin G administered intravenously, or intramuscularly in combination with oral probenecid **(Chap. 177)**.

HEREDITARY SPASTIC PARAPLEGIA

Many cases of slowly progressive myelopathy are genetic in origin **(Chap. 429)**. More than 60 different causative loci have been identified, including autosomal dominant, autosomal recessive, and X-linked forms. Especially for the recessive and X-linked forms, a family history of myelopathy may be lacking. Most patients present with almost imperceptibly progressive spasticity and weakness in the legs, usually but not always symmetrical. Sensory symptoms and signs are absent or mild, but sphincter disturbances may be present. In some families, additional neurologic signs are prominent, including nystagmus, ataxia, or optic atrophy. The onset may be as early as the first year of life or as late as middle adulthood. Only symptomatic therapies are available.

ADRENOMYELONEUROPATHY

This X-linked disorder is a variant of adrenoleukodystrophy (ALD). Most affected males have a history of adrenal insufficiency and then develop a progressive spastic (or ataxic) paraparesis beginning in early or sometimes middle adulthood; some patients also have a mild peripheral neuropathy. Female heterozygotes may develop a slower, insidiously progressive spastic myelopathy beginning later in adulthood and without adrenal insufficiency. Diagnosis is usually made by demonstration of elevated levels of very-long-chain fatty acids in plasma and in cultured fibroblasts. The responsible gene encodes the adrenoleukodystrophy protein (ALDP), a peroxisomal membrane transporter involved in carrying long-chain fatty acids to peroxisomes for degradation. Corticosteroid replacement is indicated if hypoadrenalism is present. Allogeneic bone marrow transplantation has been successful in slowing progression of cognitive decline in ALD, but appears to be ineffective for the myelopathy of ALD. Nutritional supplements (Lorenzo’s oil) have also been attempted for this condition without evidence of efficacy.

OTHER CHRONIC MYELOPATHIES

Primary lateral sclerosis **(Chap. 429)** is a mid to late life onset degenerative disorder characterized by progressive spasticity with weakness, eventually accompanied by dysarthria and dysphonia; bladder symptoms occur in approximately half of patients. Sensory function is spared. The disorder resembles ALS and is considered a variant of the motor neuron degenerations, but without the characteristic lower motor neuron disturbance and with typically a slower progression. Some cases may represent late-onset cases of

familial spastic paraplegia, particularly autosomal recessive or X-linked varieties in which a family history may be absent.

Tethered cord syndrome is a developmental disorder of the lower spinal cord and nerve roots that rarely presents in adulthood as low back pain accompanied by a progressive lower spinal cord and/or nerve root syndrome. Some patients have a small leg or foot deformity indicating a long-standing process, and in others, a dimple, patch of hair, or sinus tract on the skin overlying the lower back is the clue to a congenital lesion. Diagnosis is made by MRI, which demonstrates a low-lying conus medullaris and thickened filum terminale. The MRI may also reveal diastematomyelia (division of the lower spinal cord into two halves), lipomas, cysts, or other congenital abnormalities of the lower spine coexisting with the tethered cord. Treatment is with surgical release.

There are a number of rare toxic causes of spastic myelopathy, including lathyrism due to ingestion of chickpeas containing the excitotoxin β-*N*-oxalylamino-L-alanine (BOAA), seen primarily in the developing world, and nitrous oxide inhalation producing a myelopathy identical to subacute combined degeneration. SLE, Sjögren's syndrome, and sarcoidosis may each cause a myelopathy without overt evidence of systemic disease. Cancer-related causes of chronic myelopathy, besides the common neoplastic compressive myelopathy discussed earlier, include radiation injury **(Chap. 86)** and rare paraneoplastic myelopathies. The last of these are most often associated with lung cancer and anti-Hu or anti-CV2/CRMP5 antibodies **(Chap. 90)** or with lymphoma that causes a syndrome of destruction of anterior horn cells; NMO with aquaporin-4 antibodies **(Chap. 437)** can also rarely be paraneoplastic in origin. Metastases to the cord are probably more common than either of these in patients with cancer. Often, a cause of intrinsic myelopathy can be identified only through periodic reassessment.

REHABILITATION OF SPINAL CORD DISORDERS

The prospects for recovery from an acute destructive spinal cord lesion fade after ~6 months. There are currently no effective means to promote repair of injured spinal cord tissue; promising but entirely experimental approaches include the use of factors that influence reinnervation by axons of the corticospinal tract, nerve and neural sheath graft bridges, forms of electrical stimulation at the site of injury, and the local introduction of stem cells. The disability associated with irreversible spinal cord damage is determined primarily by the level of the lesion and by whether the disturbance in function is complete or incomplete **(Table 434-4)**. Even a complete high cervical cord lesion may be compatible with a productive life. The primary goals are development of a rehabilitation plan framed by realistic expectations and attention to the neurologic, medical, and psychological complications that commonly arise.

TABLE 434-4

Expected Neurologic Function Following Complete Cord Lesions

Level	Self-Care	Transfers	Maximum Mobility
High quadriplegia (C1-C4)	Dependent on others; requires respiratory support	Dependent on others	Motorized wheelchair
Low quadriplegia (C5-C8)	Partially independent with adaptive equipment	May be dependent or independent	May use manual wheelchair, drive an automobile with adaptive equipment
Paraplegia (below T1)	Independent	Independent	Ambulates short distances with aids

Source: Adapted from JF Ditunno, CS Formal: Chronic spinal cord injury. N Engl J Med 330:550, 1994; with permission.

Many of the usual symptoms associated with medical illnesses, especially somatic and visceral pain, may be lacking because of the destruction of afferent pain pathways. Unexplained fever, worsening of spasticity, or deterioration in neurologic function should prompt a search for infection, thrombophlebitis, or an intraabdominal pathology. The loss of normal thermoregulation and inability to maintain normal body temperature can produce recurrent fever (*quadriplegic fever*), although most episodes of fever are due to infection of the urinary tract, lung, skin, or bone.

Bladder dysfunction generally results from loss of supraspinal innervation of the detrusor muscle of the bladder wall and the sphincter musculature. Detrusor spasticity is treated with anticholinergic drugs (oxybutynin, 2.5–5 mg qid) or tricyclic antidepressants with anticholinergic properties (imipramine, 25–200 mg/d). Failure of the sphincter muscle to relax during bladder emptying (urinary dyssynergia) may be managed with the α-adrenergic blocking agent terazosin hydrochloride (1–2 mg tid or qid), with intermittent catheterization, or, if that is not feasible, by use of a condom catheter in men or a permanent indwelling catheter. Surgical options include the creation of an artificial bladder by isolating a segment of intestine that can be catheterized intermittently (enterocystoplasty) or can drain continuously to an external appliance (urinary conduit). Bladder areflexia due to acute spinal shock or conus lesions is best treated by catheterization. Bowel regimens and disimpaction are necessary in most patients to ensure at least biweekly evacuation and avoid colonic distention or obstruction.

Patients with acute cord injury are at risk for venous thrombosis and pulmonary embolism. Use of calf-compression devices and anticoagulation with low-molecular-weight heparin is recommended. In cases of persistent paralysis, anticoagulation should probably be continued for 3 months.

Prophylaxis against decubitus ulcers should involve frequent changes in position in a chair or bed, the use of special mattresses, and cushioning of areas where pressure sores often develop, such as the sacral prominence and heels. Early treatment of ulcers with careful cleansing, surgical or enzyme debridement of necrotic tissue, and appropriate dressing and drainage may prevent infection of adjacent soft tissue or bone.

Spasticity is aided by stretching exercises to maintain mobility of joints. Drug treatment is effective but may result in reduced function, as some patients depend on spasticity as an aid to stand, transfer, or walk. Baclofen (up to 240 mg/d in divided doses) is effective; it acts by facilitating γ-aminobutyric acid–mediated inhibition of motor reflex arcs. Diazepam acts by a similar mechanism and is useful for leg spasms that interrupt sleep (2–4 mg at bedtime). Tizanidine (2–8 mg tid), an α_2 adrenergic agonist that increases presynaptic inhibition of motor neurons, is another option. For nonambulatory patients, the direct muscle inhibitor dantrolene (25–100 mg qid) may be used, but it is potentially hepatotoxic. In refractory cases, intrathecal baclofen administered via an implanted pump, botulinum toxin injections, or dorsal rhizotomy may be required to control spasticity.

Despite the loss of sensory function, many patients with spinal cord injury experience chronic pain sufficient to diminish their quality of life. Randomized controlled studies indicate that gabapentin or pregabalin is useful in this setting. Epidural electrical stimulation and intrathecal infusion of pain medications have been tried with some success. **Management of chronic pain is discussed in Chap. 10.**

A paroxysmal autonomic hyperreflexia may occur following lesions above the major splanchnic sympathetic outflow at T6. Headache, flushing, and diaphoresis above the level of the lesion, as well as hypertension with bradycardia or tachycardia, are the major symptoms. The trigger is typically a noxious stimulus—for example, bladder or bowel distention, a urinary tract infection, or a decubitus ulcer—below the level of the cord lesion. Treatment consists of removal of offending stimuli; ganglionic blocking agents (mecamylamine, 2.5–5 mg) or other short-acting antihypertensive drugs are useful in some patients.

Attention to these details allows longevity and a productive life for patients with complete transverse myelopathies.

FURTHER READING

Kato S et al: Comparison of anterior and posterior surgery for degenerative cervical myelopathy: An MRI-based propensity-score-matched analysis using data from the prospective multicenter AOSpine CSM North America and international studies. J Bone Joint Surg Am 99:1013, 2017. [PubMed: 28632590]

Kühl JS et al: Long-term outcomes of allogeneic haematopoietic stem cell transplantation for adult cerebral X-linked adrenoleukodystrophy. Brain 140:953, 2017. [PubMed: 28375456]

Loblaw DA et al: A 2011 updated systematic review and clinical practice guideline for the management of malignant extradural spinal cord compression. Int J Radiat Oncol Biol Phys 84:312, 2012. [PubMed:

22420969]

Ozpinar A et al: Epidemiology, clinical presentation, diagnostic evaluation, and prognosis of spinal arteriovenous malformations. Handb Clin Neurol 143:145, 2017. [PubMed: 28552136]

Patchell RA et al: Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: A randomised trial. Lancet 366:643, 2005. [PubMed: 16112300]

Robertson CE et al: Recovery after spinal cord infarcts: Long-term outcome in 115 patients. Neurology 78:114, 2012. [PubMed: 22205760]

Ropper AE, Ropper AH: Acute spinal cord compression. N Engl J Med 376:1358, 2017. [PubMed: 28379788]

Ruet A et al: Predictive factors for multiple sclerosis in patients with clinically isolated spinal cord syndrome. Mult Scler 17:312, 2011. [PubMed: 21071465]

Yáñez ML et al: Diagnosis and treatment of epidural metastases. Cancer 123:1106, 2017. [PubMed: 28026861]

Chapter 435: Concussion and Other Traumatic Brain Injuries

Geoffrey T. Manley; Stephen L. Hauser; Michael McCrea

INTRODUCTION

Traumatic brain injury (TBI) represents a significant global public health problem facing the United States and other countries around the world. In the United States, estimates of the frequency of TBI range between 2.5 and 4 million cases per year, depending on the study and methods used to define and include cases. Age-specific rates show a bimodal distribution, with highest risk in younger individuals and older adults. The most common mechanism of injury in the young is motor vehicle accidents and is more common in men, while in older adults falls are the major cause of injury and are more likely to occur in women.

TBI imposes substantial demands on health care systems. Worldwide, at least 10 million TBIs are serious enough to result in death or hospitalization. In the United States, the estimated annual cost is >$76 billion. Due to advances in medical care and other factors, more people are surviving traumatic brain injury than ever before. Brain injury accounts for more lost productivity at work among Americans than any other form of injury. An estimated 5.3 million Americans are living with significant disabilities resulting from TBI that complicate their return to a full and productive life. Increased media attention to military and sports-related TBI has highlighted the growing concern that injuries that were previously dismissed can have life-long consequences for some individuals.

Head injuries are so common that almost all physicians will be called upon to provide some aspect of immediate care or to see patients who are suffering from various sequelae. Patients initially need education regarding the natural history of TBI along with treatment of acute symptoms such as headache. Follow-up of TBI patients is important to make sure that the sequelae that some patients experience—such as postconcussive disorder (PCD), depression, or sleep disorders—can be identified and treated by a coordinated multidisciplinary team.

DEFINITION AND CLASSIFICATION

TBI is commonly defined as *an alteration in brain function, or other evidence of brain pathology, caused by an external force, and characterized by the following: (1) any period of loss or decreased level of consciousness (LOC), (2) any loss of memory for events immediately before (retrograde) or after*

(posttraumatic) the injury, (3) any neurological deficits, and/or (4) any alteration in mental state at the time of injury.

Evidence of TBI can include visual, neuroradiologic, or laboratory confirmation of damage to the brain, but TBI is more often diagnosed on the basis of acute clinical criteria. In addition to standard computed tomography (CT) imaging, modern structural magnetic resonance imaging (MRI), and functional imaging (resting state functional MRI) techniques show increasing sensitivity, and it is likely that sensitive blood-based biomarkers will be developed in the near future.

Mechanisms of TBI: Common mechanisms of TBI include the head being struck by an object, the head striking an object, the brain undergoing an acceleration/deceleration movement, a foreign body penetrating the brain, or forces generated from events such as a blast or explosion. Motor vehicle crashes have historically been cited as the most common cause of TBI. All forms of transportation, however, are common causes of TBI, including motorcycle crashes, bicycle accidents, skateboard and pedestrian injuries. The other leading causes of TBI are falls, assaults, and sports, with varied frequency across the lifespan. Certainly, there has been an increased focus on the high frequency of mild TBI (mTBI), often referred to as concussion, encountered by athletes participating in contact and collision sports at all competitive levels, as well as the potential short-term effects and long-term risks associated with sport-related concussion.

Classification of TBI Severity: Numerous systems have been developed over the years to define and classify TBI severity along a continuum from mild to moderate to severe. These systems are usually most applicable to closed head injuries. In nearly all classification systems, traumatic brain injury severity is graded based on acute injury characteristics rather than postacute injury status, as other factors can intervene to influence functional outcome. Historically, the presence and duration of unconsciousness and amnesia have been the main points of distinction along the gradient of TBI severity. Current TBI classification systems are symptom-based and do not incorporate patho-anatomical or molecular features.

The Glasgow Coma Scale (GCS) is the most recognized and widely used method for grading TBI severity. The GCS provides a practical indicator of gross neurologic status by assessing motor function, verbal responses, and the patient's ability to open his or her eyes voluntarily or in response to external commands and stimuli. The grading is applied to the best response that can be elicited from the patient at the time of assessment, preferably before any paralyzing or sedating medication is administered or the patient is intubated, as these interventions confound interpretation of the score. The GCS assessment produces scores ranging from 3 to 15 **(Table 435-1)**.

TABLE 435-1

Glasgow Coma Scale

Eye Opening (E)		**Verbal Response (V)**	
Spontaneous	4	Oriented	5
To speech	3	Confused	4
To pressure	2	Words	3
None	1	Sounds	2
		None	1
Best Motor Response (M)			
Obeying commands	6		
Localizing	5		
Normal flexion	4		
Abnormal flexion	3		
Extension	2		
None	1		

Note: Revised GCS (2014).

Source: From G Teasdale et al: The Glasgow Coma Scale at 40 years: Standing the test of time. Lancet Neurol 13:844, 2014.

Upon the 40th anniversary of the GCS, the wording for responses was revised, and recommendations were made to improve its utility. Importantly, individual patients are best described by the three components of the coma scale (eye, verbal, motor, e.g., E3V4M6); the derived total coma score (e.g., 13) is less informative and should only be used to characterize groups of patients.

Several injury classification systems have been developed to go beyond GCS score or acute injury characteristics and incorporate chief signs and symptoms in defining mTBI. The use of multiple severity indicators is intended to improve sensitivity in the detection of mTBI (GCS 13–15), while also taking into

consideration traditional acute injury characteristics that have been presumed to predict outcome following mild and moderate brain injury. Loss of consciousness (LOC) and posttraumatic amnesia (PTA) remain the most common injury characteristics referenced in these classification systems. In the case of moderate (GCS 9–12) and severe (GCS 3–8) TBI, GCS score and the duration of LOC and PTA can be robust predictors of long-term outcome and morbidity. In cases of mTBI, however, while PTA and LOC are important indicators of acute injury severity, they are less predictive of eventual recovery time and outcome.

TBI TYPES AND PATHOLOGIES

Mild TBI (Concussion): It is estimated that between 70 and 90% of all treated traumatic brain injuries are mild in severity based on traditional case definitions and acute injury characteristics, with most reported estimates in the order of 85%. The published figures likely under-represent the true incidence of mTBI because of variable case definitions and heterogeneous methods. Moreover, because a subgroup of individuals with milder brain injuries do not seek medical attention, epidemiological studies that depend on hospital-based data also underestimate the true incidence.

The term concussion, while popular, is vague and is not based on widely accepted objective criteria, resulting in multiple definitions from various groups. There has been debate as to whether concussion is part of the TBI spectrum or a separate entity. In 2017, the Concussion in Sports Group issued a consensus statement that “concussion is a traumatic brain injury” (McCrory et al, 2017). By firmly placing concussion in the spectrum of TBI, the underlying pathophysiological processes common to all TBI presentations can now be considered together.

SKULL FRACTURE, EXTRA-AXIAL HEMATOMA, CONTUSION, AND AXONAL INJURY

Skull Fracture

A blow to the skull that exceeds the elastic tolerance of the bone causes a fracture. Intracranial lesions accompany roughly two-thirds of skull fractures, and the presence of a fracture increases many-fold the chances of an underlying subdural or epidural hematoma. Consequently, fractures are primarily markers of the site and severity of injury. If the underlying arachnoid membrane has been torn, fractures also provide potential pathways for entry of bacteria to the cerebrospinal fluid (CSF) with a risk of meningitis and for leakage of CSF outward through the dura. If there is leakage of CSF, severe orthostatic headache results from lowered pressure in the spinal fluid compartment.

Most fractures are linear and extend from the point of impact toward the base of the skull. Basilar skull fractures are often extensions of adjacent linear fractures over the convexity of the skull but may occur independently owing to stresses on the floor of the middle cranial fossa or occiput. Basilar fractures are usually parallel to the petrous bone or along the sphenoid bone and directed toward the sella turcica and ethmoidal groove. Although most basilar fractures are uncomplicated, they can cause CSF leakage,

pneumocephalus, and delayed cavernous-carotid fistulas. Hemotympanum (blood behind the tympanic membrane), ecchymosis over the mastoid process (Battle sign), and periorbital ecchymosis (“raccoon sign”) are associated with basilar fractures and should be suspected if these clinical signs are present.

EPIDURAL AND SUBDURAL HEMATOMAS

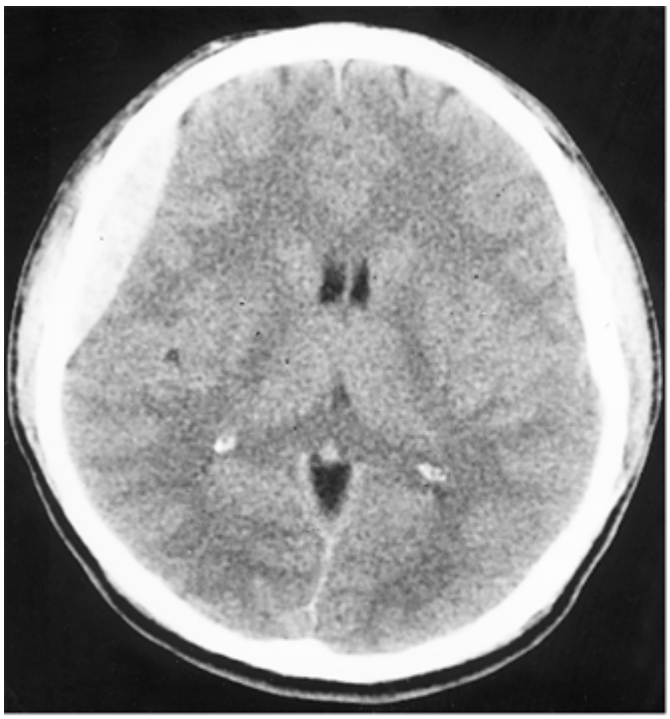
Hemorrhages between the dura and skull (epidural) or beneath the dura (subdural) have characteristic clinical and imaging features. They are sometimes associated with underlying brain contusions and other injuries, often making it difficult to determine the relative contribution of each component to the clinical state. The mass effect of raised intracranial pressure (ICP) caused by these hematomas can be life threatening, making it imperative to identify them rapidly by CT or MRI scan and to surgically remove them when appropriate.

Epidural Hematoma

These highly dangerous lesions usually arise from an injury to a meningeal arterial vessel and evolve rapidly (Fig. 435-1). They are often accompanied by a “lucid interval” of several minutes to hours prior to neurological deterioration. They occur in up to 10% of cases of severe head injury, but are less often associated with underlying cortical damage compared to subdural hematomas. Rapid surgical evacuation and ligation or cautery of the damaged vessel, usually the middle meningeal artery that has been lacerated by an overlying skull fracture, is indicated. If recognized and treated rapidly, patients often have a favorable outcome.

FIGURE 435-1

Acute epidural hematoma. The tightly attached dura is stripped from the inner table of the skull, producing a characteristic lenticular-shaped hemorrhage on noncontrast computed tomography scan. Epidural hematomas are usually caused by tearing of the middle meningeal artery following fracture of the temporal bone.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Acute Subdural Hematoma

Direct cranial trauma may be minor and is not always required for acute subdural hemorrhage to occur, especially in the elderly and those taking anticoagulant medications **(Fig. 435-2)**. Acceleration forces alone, as from whiplash, are sometimes sufficient to produce subdural hematoma. Up to one-third of patients have a lucid interval lasting minutes to hours before coma supervenes, but most are drowsy or comatose from the moment of injury. A unilateral headache and slightly enlarged pupil on the side of the hematoma are frequently, but not invariably, present. Small subdural hematomas may be asymptomatic and usually do not require surgical evacuation if they do not enlarge. Stupor or coma, hemiparesis, and unilateral pupillary enlargement are signs of larger hematomas. The bleeding that causes larger subdural hematomas is primarily venous in origin, although arterial bleeding sites are sometimes found at operation, and a few large hematomas have a purely arterial origin. In an acutely deteriorating patient, an emergency craniotomy is required. In contrast to epidural hematomas, there is significant morbidity and mortality associated with acute subdural hematomas that require surgery.

FIGURE 435-2

Acute subdural hematoma. Noncontrast computed tomography scan reveals a hyperdense clot that has an irregular border with the brain and causes more horizontal displacement (mass effect) than might be expected from its thickness. The disproportionate mass effect is the result of the large rostral-caudal extent of these hematomas. Compare to Fig. 435-1.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Chronic Subdural Hematoma

A subacutely evolving syndrome due to subdural hematoma occurs days or weeks after injury with drowsiness, headache, confusion, or mild hemiparesis, usually in the elderly with age-related atrophy and often after only minor or unnoticed trauma. On imaging studies, chronic subdural hematomas appear as crescentic clots over the convexity of one or both hemispheres, most commonly in the frontotemporal region (Fig. 435-3). A history of trauma may or may not be elicited in relation to chronic subdural hematoma; the injury may have been trivial and forgotten, particularly in the elderly and those with clotting disorders. Headache is common but not invariable. Additional features that may appear weeks later include slowed thinking, vague change in personality, seizure, or a mild hemiparesis. The headache typically fluctuates in severity, sometimes with changes in head position. Drowsiness, inattentiveness, and incoherence of thought are generally more prominent than focal signs such as hemiparesis. Rarely, chronic hematomas cause brief episodes of hemiparesis or aphasia that are indistinguishable from transient ischemic attacks.

FIGURE 435-3

Computed tomography scan of chronic bilateral subdural hematomas of different ages. The collections began as acute hematomas and have become hypodense in comparison to the adjacent brain after a period during which they were isodense and difficult to appreciate. Some areas of resolving blood are contained on the more recently formed collection on the left (*arrows*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CT without contrast initially shows a low-density mass over the convexity of the hemisphere. Between 2 and 6 weeks after the initial bleeding, the clot becomes isodense compared to adjacent brain and may be inapparent. Many subdural hematomas that are several weeks in age contain areas of blood and intermixed serous fluid. Infusion of contrast material demonstrates enhancement of the vascular fibrous capsule surrounding the collection. MRI reliably identifies both subacute and chronic hematomas.

Clinical observation coupled with serial imaging is a reasonable approach to patients with few symptoms and small chronic subdural collections that do not cause mass effect. Treatment with surgical evacuation through burr holes is usually successful, if a cranial drain is used postoperatively. The fibrous membranes that grow from the dura and encapsulate the collection may require removal with a craniotomy to prevent recurrent fluid accumulation.

TRAUMATIC SUBARACHNOID HEMORRHAGE

Subarachnoid hemorrhage (SAH) is common in TBI. Rupture of small cortical arteries or veins can cause bleeding into the subarachnoid space. Traumatic SAH is often seen in the sulci and is frequently the only radiographic finding on CT following mild TBI. SAH occurs diffusely after severe TBI and confers an increase in mortality. In mild TBI, SAH provides an objective imaging biomarker for TBI, and in some patients is associated with unfavorable outcomes.

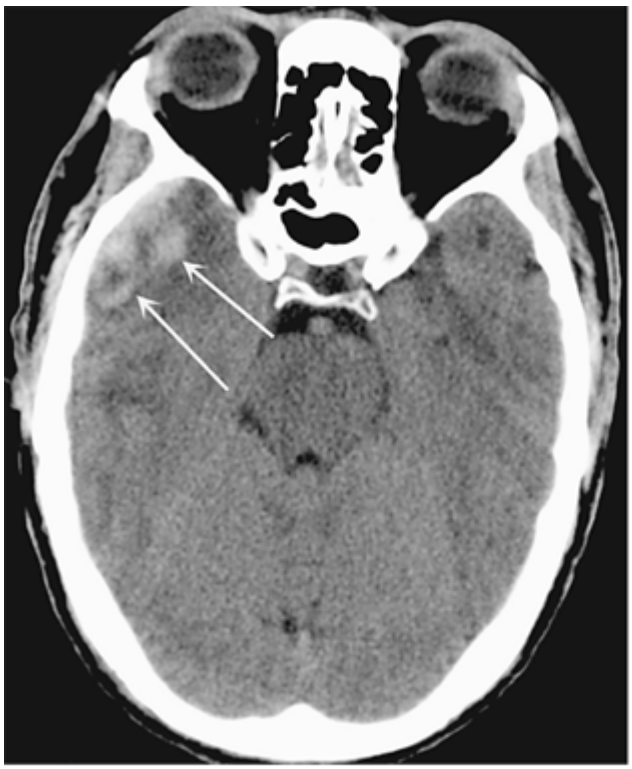
Contusion

A surface bruise of the brain, or contusion, consists of varying degrees of petechial hemorrhage, edema, and tissue destruction **(Fig. 435-4)**. Contusions and deeper hemorrhages result from mechanical forces that

displace and compress the hemispheres forcefully and by deceleration of the brain against the inner skull, either under a point of impact (coup lesion) or, as the brain swings back, in the antipolar area (contrecoup lesion). Trauma sufficient to cause prolonged unconsciousness usually produces some degree of contusion. Blunt deceleration impact, as occurs against an automobile dashboard or from falling forward onto a hard surface, causes contusions on the orbital surfaces of the frontal lobes and the anterior and basal portions of the temporal lobes. With lateral forces, as from impact on an automobile door frame, contusions are situated on the lateral convexity of the hemisphere. The clinical signs of contusion are determined by the location and size of the lesion; often, there are no focal abnormalities with a routine neurological exam, but these injured regions are later the sites of gliotic scars that may produce seizures. A hemiparesis or gaze preference is fairly typical of moderately sized contusions. Large bilateral contusions produce stupor with extensor posturing, while those limited to the frontal lobes cause a taciturn state. Contusions in the temporal lobe may cause delirium or an aggressive, combative syndrome. Torsional or shearing forces within the brain can cause hemorrhages of the basal ganglia and other deep regions. Large contusions and hemorrhages after minor trauma should raise concerns for coagulopathy due to an underlying disease or more commonly anticoagulant therapy.

FIGURE 435-4

Traumatic cerebral contusion. Noncontrast computed tomography scan demonstrating a hyperdense hemorrhagic region in the anterior temporal lobe.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Acute contusions are easily visible on CT and MRI scans, appearing as inhomogeneous hyperdensities on CT and as hyperintensities on T2 and fluid-attenuated inversion recovery (FLAIR) MRI sequences; there is usually surrounding localized brain edema and some subarachnoid bleeding. Blood in the CSF due to trauma may

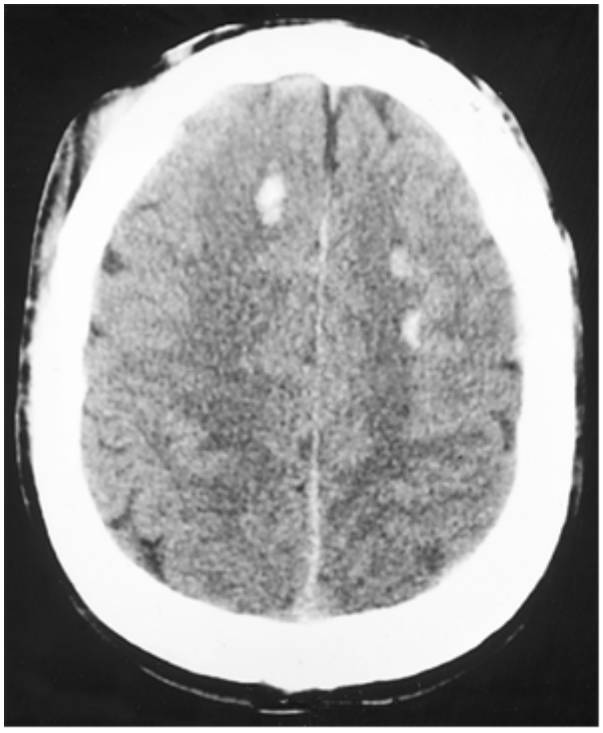
provoke a mild inflammatory reaction. Over a few days, contusions acquire a surrounding contrast enhancement and edema that may be mistaken for tumor or abscess.

Axonal Injury

Traumatic axonal injury (TAI) is one of the most common injuries after TBI **(Fig. 435-5)**. There is disruption, or shearing, of axons at the time of impact and this is associated with microhemorrhages. It occurs following high-speed deceleration injuries, such as motor vehicle collisions (Johnson et al, 2013). The presence of ≥4 areas of TAI is called diffuse axonal injury (DAI), and when widespread, has been proposed to explain persistent coma and the vegetative state after TBI **(Chap. 300)**. Only severe TAI lesions that contain substantial blood are visualized by CT, usually in the corpus callosum and centrum semiovale. More commonly, the CT will be negative for TAI, but subsequent MRI, particularly gradient-echo or susceptibility weighted imaging, will show hemosiderin deposits reflective of microhemorrhages in addition to the axonal damage on diffusion sequences.

FIGURE 435-5

Multiple small areas of hemorrhage and tissue disruption in the white matter of the frontal lobes on noncontrast computed tomography scan. These appear to reflect an extreme type of the diffuse axonal shearing lesions that occur with closed head injury.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CRANIAL NERVE INJURIES

The cranial nerves most often injured with TBI are the olfactory, optic, oculomotor, and trochlear; the first and second branches of the trigeminal nerve; and the facial and auditory nerves. Anosmia and an apparent loss of taste (actually a loss of perception of aromatic flavors, with retained elementary taste perception) occur in ~10% of persons with serious head injuries, particularly from falls on the back of the head. This is the result of displacement of the brain and shearing of the fine olfactory nerve filaments that course through the cribriform bone. At least partial recovery of olfactory and gustatory function is expected, but if bilateral anosmia persists for several months, the prognosis is poor. Partial optic nerve injuries from closed trauma result in blurring of vision, central or paracentral scotomas, or sector defects. Direct orbital injury may cause short-lived blurred vision for close objects due to reversible iridoplegia. Diplopia limited to downward gaze and corrected when the head is tilted away from the side of the affected eye indicates trochlear (fourth nerve) nerve damage. It occurs frequently as an isolated problem after minor head injury or may develop for unknown reasons after a delay of several days. Facial nerve injury caused by a basilar fracture is present immediately in up to 3% of severe injuries; it may also be delayed for 5–7 days. Fractures through the petrous bone, particularly the less common transverse type, are liable to produce facial palsy. Delayed facial palsy occurring up to a week after injury, the mechanism of which is unknown, has a good prognosis. Injury to the eighth cranial nerve from a fracture of the petrous bone causes loss of hearing, vertigo, and nystagmus immediately after injury. Deafness from eighth nerve injury is rare and must be distinguished from blood in the middle ear or disruption of the middle ear ossicles. Dizziness, tinnitus, and high-tone hearing loss occur from cochlear concussion.

SEIZURES

Convulsions are surprisingly uncommon immediately after TBI, but a brief period of tonic extensor posturing or a few clonic movements of the limbs just after the moment of impact can occur. However, the cortical scars that evolve from contusions are highly epileptogenic and may later manifest as seizures, even after many months or years (Chap. 418). The severity of injury roughly determines the risk of future seizures. It has been estimated that 17% of individuals with brain contusion, subdural hematoma, or prolonged LOC will develop a seizure disorder and that this risk extends for an indefinite period of time, whereas the risk is ≤2% after mild injury. The majority of convulsions in the latter group occur within 5 years of injury but may be delayed for decades. Penetrating injuries have a much higher rate of subsequent epilepsy.

CLINICAL SYNDROMES AND TREATMENT OF HEAD INJURY

CONCUSSION/MILD TBI

The patient who has briefly lost consciousness or been stunned after a minor head injury usually becomes fully alert and attentive within minutes but may complain of headache, dizziness, faintness, nausea, a single episode of emesis, difficulty with concentration, a brief amnestic period, or slight blurring of vision. This typical concussion syndrome has a good prognosis with little risk of subsequent deterioration. Children are particularly prone to drowsiness, vomiting, and irritability, symptoms that are sometimes delayed for several hours after apparently minor injuries. Vasovagal syncope that follows injury may cause undue concern.

Generalized or frontal headache is common in the following days. It may be migrainous (throbbing and hemicranial) in nature or aching and bilateral. After several hours of observation, patients with minor injury may be accompanied home and observed for a day by a family member or friend, with written instructions to return if symptoms worsen.

Persistent severe headache and repeated vomiting in the context of normal alertness and no focal neurologic signs is usually benign, but CT should be obtained and a longer period of observation is appropriate. The decision to perform imaging tests also depends on clinical signs that indicate that the impact was severe (e.g., persistent confusion, repeated vomiting, palpable skull fracture); the presence of other serious bodily injuries, an underlying coagulopathy, or age >65 years; and on the degree of surveillance that can be anticipated after discharge. Guidelines have also indicated that older age (>65), two or more episodes of vomiting, >30 min of retrograde or persistent anterograde amnesia, seizure, and concurrent drug or alcohol intoxication are sensitive (but not specific) indicators of intracranial hemorrhage that justify CT scanning.

SPORT-RELATED CONCUSSION

Based on its reported prevalence and acute effects, and fears over potential long-term neurological consequences, sport-related concussion has become the focus of increasing concern from clinicians, researchers, sporting organizations, and athletes themselves. Concussion is a frequent injury in contact and collision sports (e.g., football, hockey, wrestling) at all levels of participation, including youth sports. One study indicated that from 1997 to 2007 emergency department visits for 8- to 13-year-old children affected by concussion in organized team sports doubled, and increased by >200% in the 14- to 19-year-old group; these increases could represent improvements in identification in addition to actual changes in incidence rates.

Research over the last decade has advanced our understanding of the true natural history of clinical recovery following sport-related concussion. In general, the findings on acute recovery are favorable. A 2003 report was the first to chart the continuous time course of acute recovery within several days after concussion, indicating that >90% of athletes reported symptom recovery within 1 week. Several other prospective studies have since demonstrated that the overwhelming majority of athletes achieve a complete recovery in symptoms, cognitive functioning, postural stability, and other functional impairments over a period of 1–3 weeks following concussion.

There are frequent anecdotal reports, however, of athletes who remain symptomatic or impaired on functional testing well beyond the window of recovery commonly reported in group studies. The greatest challenge arguably still facing sport medicine clinicians and public health experts is how to most effectively manage and reduce risk in this subset of athletes who do not follow the “typical” course of recovery. The precise frequency of athletes who do not follow the typical course of rapid, spontaneous recovery and instead exhibit prolonged postconcussive symptoms or other functional impairments after concussion remains unclear. There is little empirical evidence regarding which risk factors may be associated with prolonged recovery time or poor outcome in athletes and how these risks can be modified in a clinical setting.

In the current absence of adequate data, a common sense approach to athletic concussion has been to remove the individual from play immediately and avoid contact sports for at least several days after a mild injury, and for a longer period if there are more severe injuries or if there are protracted neurologic symptoms such as headache and difficulty concentrating. No individual should return to play unless all symptoms have resolved and an assessment has been made by a health care professional who has experience with treatment of concussion. Once cleared, the individual can then begin a graduated program of increasing activity. Younger athletes are particularly likely to experience protracted concussive symptoms, and a slower return to play in this age group may be reasonable. These guidelines are designed in part to avoid a perpetuation of symptoms but also to prevent the rare *second impact syndrome*, in which diffuse and fatal cerebral swelling follows a second minor head injury.

In the past, mental decline in boxers late in their careers had been called *dementia pugilistica*. There is some evidence that repeated concussions from other sports, and especially in professional American football players, are associated with a similar delayed and progressive cognitive disorder, sometimes with prominent behavioral symptoms that can include depression, insomnia, violent behaviors, and suicidality. The brains of these patients display a characteristic deposition of tau protein in neurons located in the superficial cortical layers and perivascular regions, and particularly in the depths of sulci, a pattern named *chronic traumatic encephalopathy* (CTE). CTE is an intensively studied and provocative entity. A recent neuropathologic study of athletes who had donated their brains for research reported that changes of CTE were extremely common findings. However, the majority of former football players do not complain of cognitive symptoms, and at this time the true prevalence of CTE is unknown. Its contribution, if any, to late-life dementia and parkinsonism in former athletes, soldiers, or others who have sustained repeated concussive injuries is unknown. **CTE is also discussed in Chap. 417**.

POSTCONCUSSIVE STATES

The *postconcussion syndrome* (PCS) refers to a state following minor TBI consisting of combinations of fatigue, dizziness, headache, and difficulty in concentration. Management is difficult and generally requires the identification and management of the specific problem or problems that are most troubling to the individual. A clear explanation of the symptoms that may follow concussion has been shown to reduce subsequent complaints. Care is taken to avoid prolonged use of drugs that produce dependence. Headache may initially be treated with acetaminophen and small doses of amitriptyline. Vestibular exercises **(Chap. 19)** and small doses of vestibular suppressants such as promethazine (Phenergan) may be helpful when dizziness is the main problem. Patients who after minor or moderate injury have difficulty with memory or with complex cognitive tasks at work may be reassured that these problems usually improve over several months, and workload may be reduced in the interim.

For the vast majority of individuals with mTBI, the symptoms of PCS subside and resolve within a few weeks of injury. For a subset of individuals with mTBI, complaints of postconcussion symptoms persist beyond the expectation derived from TBI severity markers. The term PCD has been proposed for diagnostic use when symptoms following mTBI such as neurologic, cognitive, behavioral or somatic complaints persist beyond

the acute and subacute periods and become chronic, often operationalized as persisting beyond 3 months. Although the overall risk of developing PCD following mTBI is low, the frequency of mTBI patients who meet criteria for a diagnosis of PCD and present in a clinical setting is believed to be higher.

mTBI patients with PCD frequently present to the outpatient clinics of primary care physicians, physiatrists or neurologists seeking relief for lingering PCD-related symptoms. While some patients will have already received an initial medical work-up to rule out a more serious brain injury during the acute phase, many patients will have had no prior contact with health care specialists. A medical work-up ordered in the outpatient setting for PCD-related complaints is typically unremarkable for any identifiable neurologic cause to account for the persisting symptoms reported by the patient. The development of uniform decision trees or “standard of care” treatment regimens for PCD-related symptoms has been limited by the diversity of symptoms that patients experience, even within mTBI subgroups that have sustained very similar injury patterns. While some patients experience somatic symptoms, others complain of subjective cognitive or behavioral changes.

PCD is not a unidimensional condition but rather an outcome influenced by diverse cognitive, emotional, medical, psychosocial, and motivational factors. Because of this complexity, treatments targeting persistent and refractory PCD-related symptoms should be tailored to the needs and expectations of the individual patient, with referrals to specialists as needed for assistance with management of headache, neck and back pain, dizziness and vertigo, and other symptoms reported within the context of PCD. In addition, patients are frequently referred to behavioral health providers such as neuropsychologists, rehabilitation psychologists, health psychologists, and/or psychiatrists for a variety of reasons, but particularly when they are experiencing cognitive, emotional, or behavioral changes that accompany PCD. Patients with mood disorders (e.g., depression), anxiety disorders (e.g., posttraumatic stress disorder), or adjustment reactions may benefit from psychiatric consultation for appropriate medication trials or from time-limited psychotherapy such as cognitive behavioral therapy.

INJURY OF INTERMEDIATE SEVERITY

Patients who are not fully alert or have persistent confusion, behavioral changes, extreme dizziness, or focal neurologic signs such as hemiparesis should be admitted to the hospital and undergo a cerebral imaging study. A cerebral contusion or hematoma will usually be found. Common syndromes include: (1) delirium with a disinclination to be examined or moved, expletive speech, and resistance if disturbed (anterior temporal lobe contusions); (2) a quiet, disinterested, slowed mental state (abulia) alternating with irascibility (inferior frontal and frontopolar contusions); (3) a focal deficit such as aphasia or mild hemiparesis (due to subdural hematoma or convexity contusion or, less often, carotid artery dissection); (4) confusion and inattention, poor performance on simple mental tasks, and fluctuating orientation (associated with several types of injuries, including those described above, and with medial frontal contusions and interhemispheric subdural hematoma); (5) repetitive vomiting, nystagmus, drowsiness, and unsteadiness (labyrinthine concussion, but occasionally due to a posterior fossa subdural hematoma or vertebral artery dissection); and (6) diabetes insipidus (damage to the median eminence or pituitary stalk). Injuries of this degree are often

complicated by drug or alcohol intoxication, and clinically inapparent cervical spine injury may be present. Blast injuries are often accompanied by rupture of the tympanic membranes.

After surgical removal of hematomas, patients in this category improve over weeks to months. During the first week, the state of alertness, memory, and other cognitive functions often fluctuate, and agitation and somnolence are common. Behavioral changes tend to be worse at night, as with many other encephalopathies, and may be treated with small doses of antipsychotic medications. Subtle abnormalities of attention, intellect, spontaneity, and memory return toward normal weeks or months after the injury, sometimes abruptly. However, the full extent of recovery may not be realized for several years. Persistent cognitive problems are discussed below.

SEVERE INJURY

Patients who are comatose from the moment of injury require immediate neurologic attention and resuscitation. After intubation, with care taken to immobilize the cervical spine, the depth of coma, pupillary size and reactivity, limb movements, and Babinski responses are assessed. As soon as vital functions permit and cervical spine x-rays and a CT scan have been obtained, the patient should be transported to a critical care unit. Hypoxia should be reversed, and normal saline used as the resuscitation fluid in preference to albumin. The finding of an epidural or subdural hematoma or large intracerebral hemorrhage is usually an indication for prompt surgery and intracranial decompression in an otherwise salvageable patient. Measurement of ICP with a ventricular catheter or fiberoptic device in order to guide treatment has been favored by many units but has not improved outcome. Hyperosmolar intravenous solutions are used in various regimens to limit intracranial pressure. Prophylactic antiepileptic medications are recommended for 7 days and should be discontinued unless there are multiple seizures postinjury. **Management of raised ICP, a frequent feature of severe head injury, is discussed in Chap. 301**.

Despite the improvement in mortality for severe TBI over the past few decades, a great deal of therapeutic nihilism persists in TBI. The common use of a 6-month outcome for TBI clinical studies reinforces this misconception. The recovery from severe TBI can take years. Furthermore, the ability to predict long-term outcome is limited and frequently incorrect. Recent best practice guidelines recommend, in the absence of brain death, that aggressive therapy be instituted for at least 72 h in the acute injury period.

ACKNOWLEDGMENT

The authors wish to acknowledge the contributions of Dr. Allan Ropper to earlier editions of this chapter.

FURTHER READING

Jagoda AS et al: American College of Emergency Physicians and Centers for Disease Control and Prevention. Clinical policy: Neuroimaging and decisionmaking in adult mild traumatic brain injury in the acute setting. Ann Emerg Med 52:714, 2008. [PubMed: 19027497]

Johnson VE et al: Axonal pathology in traumatic brain injury. Exp Neurol 246:35, 2013. [PubMed: 22285252]

McCrory P et al: Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in Berlin, October 2016. Br J Sports Med 51:838, 2017. [PubMed: 28446457]

Mez J et al: Clinicopathological evaluation of chronic traumatic encephalopathy in players of American football. JAMA 25;318:360, 2017.

Taylor CA et al: Traumatic brain injury-related emergency department visits, hospitalizations, and deaths—United States, 2007 and 2013. MMWR Surveill Summ 66:1, 2017. [PubMed: 28301451]

Chapter 436: Multiple Sclerosis

Bruce A. C. Cree; Stephen L. Hauser

MULTIPLE SCLEROSIS

Multiple sclerosis (MS) is an autoimmune disease of the central nervous system (CNS) characterized by chronic inflammation, demyelination, gliosis (plaques or scarring), and neuronal loss; the course can be relapsing or progressive. MS plaques typically develop at different times and in different CNS locations (i.e., MS is said to be disseminated in time and space). Approximately more than 900,000 individuals in the United States and millions of individuals worldwide are affected. The clinical course is extremely variable, ranging from a relatively benign condition to a rapidly evolving and incapacitating disease requiring profound lifestyle adjustments.

CLINICAL MANIFESTATIONS

The onset of MS may be abrupt or insidious. Symptoms may be severe or seem so trivial that a patient may not seek medical attention for months or years. Indeed, at autopsy, ~0.1% of individuals who were asymptomatic during life will be found, unexpectedly, to have pathologic evidence of MS. Similarly, an magnetic resonance imaging (MRI) scan obtained for an unrelated reason may show evidence of asymptomatic MS. Symptoms of MS are extremely varied and depend on the location and severity of lesions within the CNS (Table 436-1). Examination often reveals evidence of neurologic dysfunction, often in asymptomatic locations. For example, a patient may present with symptoms in one leg but signs in both.

TABLE 436-1

Initial Symptoms of Multiple Sclerosis (MS)

Symptom	Percentage of Cases	Symptom	Percentage of Cases
Sensory loss	37	Lhermitte	3
Optic neuritis	36	Pain	3
Weakness	35	Dementia	2
Paresthesias	24	Visual loss	2
Diplopia	15	Facial palsy	1
Ataxia	11	Impotence	1
Vertigo	6	Myokymia	1
Paroxysmal attacks	4	Epilepsy	1
Bladder	4	Falling	1

Source: After WB Matthews et al: *McAlpine's Multiple Sclerosis.* New York, Churchill Livingstone, 1991.

Sensory symptoms are varied and include both paresthesias (e.g., tingling, prickling sensations, formications, "pins and needles," or painful burning) and hypesthesia (e.g., reduced sensation, numbness, or a "dead" feeling). Unpleasant sensations (e.g., feelings that body parts are swollen, wet, raw, or tightly wrapped) are also common. Sensory impairment of the trunk and legs below a horizontal line on the torso (a sensory level) indicates that the spinal cord is the origin of the sensory disturbance. It is often accompanied by a bandlike sensation of tightness around the torso. Pain is a common symptom of MS, experienced by >50% of patients. Pain can occur anywhere on the body and can change locations over time.

Optic neuritis (ON) presents as diminished visual acuity, dimness, or decreased color perception (desaturation) in the central field of vision. These symptoms can be mild or may progress to severe visual loss. Rarely, there is complete loss of light perception. Visual symptoms are generally monocular

but may be bilateral. Periorbital pain (aggravated by eye movement) often precedes or accompanies the visual loss. An afferent pupillary defect (Chap. 28) is usually present. Funduscopic examination may be normal or reveal optic disc swelling (papillitis). Pallor of the optic disc (optic atrophy) commonly follows ON. Uveitis is uncommon and should raise the possibility of alternative diagnoses such as sarcoid or lymphoma.

Weakness of the limbs may manifest as loss of strength, speed, or dexterity, as fatigue, or as a disturbance of gait. Exercise-induced weakness is a characteristic symptom of MS. The weakness is of the upper motor neuron type (Chap. 21) and is usually accompanied by other pyramidal signs such as spasticity, hyperreflexia, and Babinski signs. Occasionally, a tendon reflex may be lost (simulating a lower motor neuron lesion) if an MS lesion disrupts the afferent reflex fibers in the spinal cord (see Fig. 21-2).

Facial weakness due to a lesion in the pons may resemble idiopathic Bell's palsy (Chap. 433). Unlike Bell's palsy, facial weakness in MS is usually not associated with ipsilateral loss of taste sensation or retroauricular pain.

Spasticity (Chap. 21) is commonly associated with spontaneous and movement-induced muscle spasms. More than 30% of MS patients have moderate to severe spasticity, especially in the legs. This is often accompanied by painful spasms interfering with ambulation, work, or self-care. Occasionally, spasticity provides support for the body weight during ambulation, and in these cases, treatment of spasticity may actually do more harm than good.

Visual blurring in MS may result from ON or diplopia (double vision); if the symptom resolves when either eye is covered, the cause is diplopia. *Diplopia* may result from internuclear ophthalmoplegia (INO) or from palsy of the sixth cranial nerve (rarely the third or fourth). An INO consists of impaired adduction of one eye due to a lesion in the ipsilateral medial longitudinal fasciculus (Chaps. 28 and V3). Prominent nystagmus is often observed in the abducting eye, along with a small skew deviation. A bilateral INO is particularly suggestive of MS. Other common gaze disturbances in MS include (1) a horizontal gaze palsy, (2) a "one and a half" syndrome (horizontal gaze palsy plus an INO), and (3) acquired pendular nystagmus.

Ataxia usually manifests as cerebellar tremors (Chap. 431). Ataxia may also involve the head and trunk or the voice, producing a characteristic cerebellar dysarthria (scanning speech).

Vertigo may appear suddenly from a brainstem lesion, superficially resembling acute labyrinthitis (Chap. 19). *Hearing loss* (Chap. 30) may also occur in MS but is uncommon.

Ancillary Symptoms

Paroxysmal symptoms are distinguished by their brief duration (10 s to 2 min), high frequency (5–40 episodes per day), lack of any alteration of consciousness or change in background electroencephalogram during episodes, and a self-limited course (generally lasting weeks to months). They may be precipitated by hyperventilation or movement. These syndromes may include Lhermitte's symptom; tonic contractions of a limb, face, or trunk (tonic seizures); paroxysmal dysarthria and ataxia; paroxysmal sensory disturbances; and several other less well-characterized syndromes. Paroxysmal symptoms probably result from spontaneous discharges, arising at the edges of demyelinated plaques and spreading to adjacent white matter tracts.

Lhermitte's symptom is an electric shock–like sensation (typically induced by flexion or other movements of the neck) that radiates down the back into the legs. Rarely, it radiates into the arms. It is generally self-limited but may persist for years. Lhermitte's symptom can also occur with other disorders of the cervical spinal cord (e.g., cervical spondylosis).

Trigeminal neuralgia, hemifacial spasm, and glossopharyngeal neuralgia (Chap. 433) can occur when the demyelinating lesion involves the root entry (or exit) zone of the fifth, seventh, and ninth cranial nerve, respectively. Trigeminal neuralgia (tic douloureux) is a very brief lancinating facial pain often triggered by an afferent input from the face or teeth. Most cases of trigeminal neuralgia are not MS related; however, atypical features such as onset before age 50 years, bilateral symptoms, objective sensory loss, or nonparoxysmal pain should raise the possibility that MS could be responsible.

Facial myokymia consists of either persistent rapid flickering contractions of the facial musculature (especially the lower portion of the orbicularis oculus) or a contraction that slowly spreads across the face. It results from lesions of the corticobulbar tracts or brainstem course of the facial nerve.

Heat sensitivity refers to neurologic symptoms produced by an elevation of the body's core temperature. For example, unilateral visual blurring may occur during a hot shower or with physical exercise (*Uhthoff's symptom*). It is also common for MS symptoms to worsen transiently, sometimes dramatically, during febrile illnesses (see "Acute Attacks or Initial Demyelinating Episodes," below). Such heat-related symptoms probably result from transient conduction block (see above).

Bladder dysfunction is present in >90% of MS patients, and in a third of patients, dysfunction results in weekly or more frequent episodes of incontinence. During normal reflex voiding, relaxation of the bladder sphincter (α-adrenergic innervation) is coordinated with contraction of the detrusor muscle in the bladder wall (muscarinic cholinergic innervation). *Detrusor hyperreflexia*, due to impairment of suprasegmental inhibition, causes urinary frequency, urgency, nocturia, and uncontrolled bladder emptying. *Detrusor sphincter dyssynergia*, due to loss of synchronization between detrusor and sphincter muscles, causes difficulty in initiating and/or stopping the urinary stream, producing hesitancy, urinary retention, overflow incontinence, and recurrent infection.

Constipation occurs in >30% of patients. Fecal urgency or *bowel incontinence* is less common (<15%) but can be socially debilitating.

Sexual dysfunction may manifest as decreased libido, impaired genital sensation, impotence in men, and diminished vaginal lubrication or adductor spasms in women.

Cognitive dysfunction can include memory loss; impaired attention; difficulties in executive functioning, memory, and problem solving; slowed information processing; and problems shifting between cognitive tasks. Euphoria (elevated mood) was once thought to be characteristic of MS but is actually uncommon, occurring in <20% of patients. Cognitive dysfunction sufficient to impair activities of daily living is rare.

Depression, experienced by approximately half of patients, can be reactive, endogenous, or part of the illness itself and can contribute to fatigue.

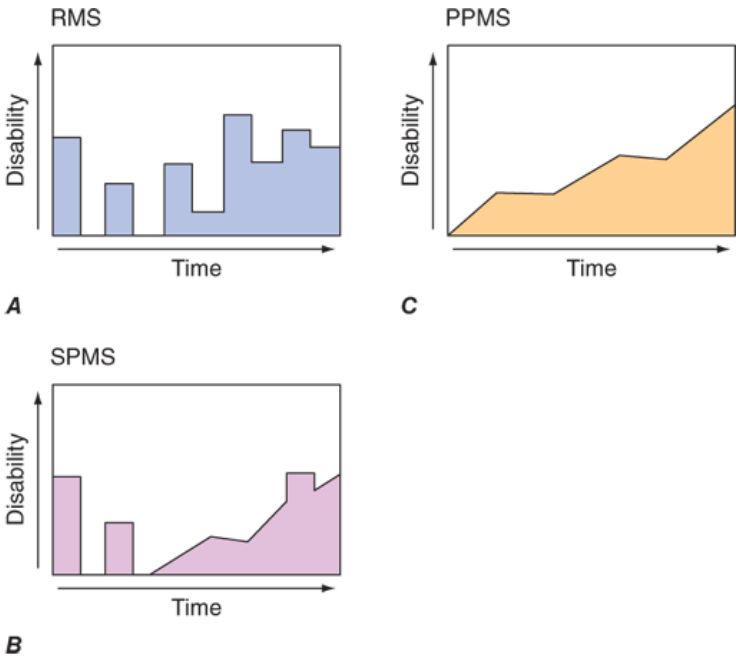
Fatigue **(Chap. 20)** is experienced by 90% of patients; this symptom is the most common reason for work-related disability in MS. Fatigue can be exacerbated by elevated temperatures, depression, expending exceptional effort to accomplish basic activities of daily living, or sleep disturbances (e.g., from frequent nocturnal awakenings to urinate).

DISEASE COURSE

Three clinical types of MS exist **(Fig. 436-1)**:

FIGURE 436-1

Clinical course of multiple sclerosis (MS). ***A.*** Relapsing MS (RMS). ***B.*** Secondary progressive MS (SPMS). ***C.*** Primary progressive MS (PPMS).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

1. *Relapsing or bout onset MS* (RMS) accounts for 90% of MS cases and is characterized by discrete attacks of neurological dysfunction that generally evolve over days to weeks (rarely over hours). With initial attacks, there is often substantial or complete recovery over the ensuing weeks to months. However, as attacks continue recovery may be less evident **(Fig. 436-1A)**. Between attacks, patients are neurologically stable.

2. *Secondary progressive MS* (SPMS) always begins as RMS **(Fig. 436-1B)**. At some point, however, the clinical course changes so that the patient experiences deterioration in function unassociated with acute attacks. SPMS produces a greater amount of fixed neurologic disability than RMS. For a patient with RMS, the risk of developing SPMS is ~2% each year, meaning that the great majority of RMS ultimately evolves into SPMS. As such, SPMS appears to represent a late stage of the same underlying illness as RMS.

3. *Primary progressive MS* (PPMS) accounts for ~10% of cases. These patients do not experience attacks but rather steadily decline in function from disease onset **(Fig. 436-1C)**. Compared to RMS, the sex distribution is more even, the disease begins later in life (mean age ~40 years), and disability develops faster (relative to the onset of the first clinical symptom). Despite these differences, PPMS appears to represent the same underlying illness as RMS.

Progressive MS and disease activity. Patients with SPMS or even PPMS will occasionally experience relapses, albeit far less often than in RMS. Progressive MS patients experiencing relapses or who are found to have acute new lesions on MRI are considered to have “active” MS. In contrast, the term “progression” is reserved to describe neurological worsening that accumulates independently from disease activity.

Epidemiology

MS is approximately threefold more common in women than men. The age of onset is typically between 20 and 40 years (slightly later in men than in women), but the disease can present across the lifespan. Approximately 10% of cases begin before the age of 18 years, and a small percentage of cases begin before the age of 10 years.

Geographical gradients are observed in MS, with the highest known prevalence for MS (250 per 100,000) in the Orkney Islands, located north of Scotland. In other temperate zone areas (e.g., northern North America, northern Europe, southern Australia, and southern New Zealand), the prevalence of MS is 0.1–0.2%. By contrast, in the tropics (e.g., Asia, equatorial Africa, and the Middle East), the prevalence is often tenfold to twentyfold less.

The prevalence of MS has increased steadily (and dramatically) in several regions around the world over the past half-century, presumably reflecting the impact of some environmental shift. Moreover, the fact that this increase has occurred primarily (or exclusively) in women indicates that women are more responsive to this environmental change.

Well-established risk factors for MS include a genetic predisposition, vitamin D deficiency, Epstein-Barr virus (EBV) exposure after early childhood, and cigarette smoking.

Vitamin D deficiency is associated with an increase in MS risk, and data suggest that ongoing deficiency also increases disease activity after MS begins. Immunoregulatory effects of vitamin D could explain these apparent relationships. Exposure of the skin to ultraviolet-B (UVB) radiation from the sun is essential for the biosynthesis of vitamin D, and this endogenous production is the most important source of vitamin D in most individuals. A diet rich in fatty fish represents another source of vitamin D. At high latitudes, the amount of UVB radiation reaching the earth's surface is often insufficient, particularly during winter months, and consequently, low serum levels of vitamin D are common in temperate zones. The common practice to avoid direct sun exposure and the widespread use of sun block would be expected to exacerbate any population-wide vitamin D deficiency (sun protection factor [SPF] 15 blocks 94% of incoming UVB radiation).

Evidence of a remote EBV infection playing some role in MS is supported by numerous epidemiologic and laboratory studies. A higher risk of infectious mononucleosis (associated with relatively late EBV infection) and higher antibody titers to latency-associated EBV nuclear antigen have been repeatedly associated with MS risk, although a causal role for EBV is not established.

A history of cigarette smoking also is associated with MS risk. Interestingly, in an animal model of MS, the lung was identified as a critical site for activation of pathogenic T lymphocytes responsible for autoimmune demyelination.

GENETIC CONSIDERATIONS



Whites are inherently at higher risk for MS than Africans or Asians, even when residing in a similar environment. MS also aggregates within some families, and adoption, half-sibling, twin, and spousal studies indicate that familial aggregation is due to genetic, and not environmental, factors **(Table 436-2)**.

TABLE 436-2

Risk of Developing Multiple Sclerosis (MS)

1 in 3	If an identical twin has MS
1 in 15	If a fraternal twin has MS
1 in 25	If a sibling has MS
1 in 50	If a parent or half-sibling has MS
1 in 100	If a first cousin has MS
1 in 1000	If a spouse has MS
1 in 1000	If no one in the family has MS

Susceptibility to MS is polygenic, with each gene contributing a relatively small amount to the overall risk. The strongest susceptibility signal genome-wide maps to the HLA-DRB1 gene in the class II region of the major histocompatibility complex (MHC), and this association accounts for ~10% of the disease risk. This HLA association, first described in the early 1970s, suggests that MS, at its core, is an autoimmune disease. Whole-genome association studies have now identified ~200 other MS susceptibility variants, each of which individually has only a very small effect on MS risk. Many of these MS-associated genes have known roles in the adaptive immune system, for example the genes for the interleukin (IL) 7 receptor (CD127), IL-2 receptor (CD25),

and T cell costimulatory molecule LFA-3 (CD58); some variants also influence susceptibility to other autoimmune diseases in addition to MS. The variants identified so far all lack specificity and sensitivity for MS; thus, at present, they are not useful for diagnosis or prediction of the future disease course.

PATHOGENESIS

Pathology

New MS lesions begin with perivenular cuffing by inflammatory mononuclear cells, predominantly T cells and macrophages, which also infiltrate the surrounding white matter. At sites of inflammation, the blood-brain barrier (BBB) is disrupted, but unlike vasculitis, the vessel wall is preserved. Involvement of the humoral immune system is also evident; small numbers of B lymphocytes infiltrate the nervous system, myelin-specific autoantibodies are present on degenerating myelin sheaths, and complement is activated. Demyelination is the pathological hallmark and evidence of myelin degeneration is found at the earliest time points of tissue injury. Although relative sparing of axons is typical of MS, partial or total axonal destruction can also occur, especially within highly inflammatory lesions. In some lesions, surviving oligodendrocytes or those that differentiate from precursor cells partially remyelinate the surviving axons, producing so-called *shadow plaques*. However, in many lesions, although oligodendrocyte precursor cells are present, they fail to differentiate into mature myelin-producing cells. As lesions evolve, there is prominent astrocytic proliferation (gliosis) and the term *sclerosis* refers to these gliotic plaques that have a rubbery or hardened texture at autopsy.

MS is not solely a disease of myelin, and neuronal pathology is increasingly recognized as a major contributor to irreversible neurologic disability. Inflammation, demyelination, and plaque formation are also present in the cerebral cortex, and significant axon loss indicating death of neurons is widespread, especially in advanced cases (see "Neurodegeneration," below). Cortical plaques may extend upward from demyelinated white matter, or may be restricted to the cortex itself, or located underneath the pia. A recently recognized feature of MS pathology is the presence of ectopic clusters of lymphocytes, termed lymphoid follicles, consisting of aggregates of T, B, and plasma cells resembling secondary lymphoid tissue located in the meninges, especially overlying deep cortical sulci; they are also present in perivascular spaces and less commonly within brain parenchyma. These structures appear to be more prevalent in progressive MS and are located in proximity to cortical plaques suggesting that perhaps diffused factors from these ectopic follicles contribute to subpial cortical demyelination and neurodegeneration.

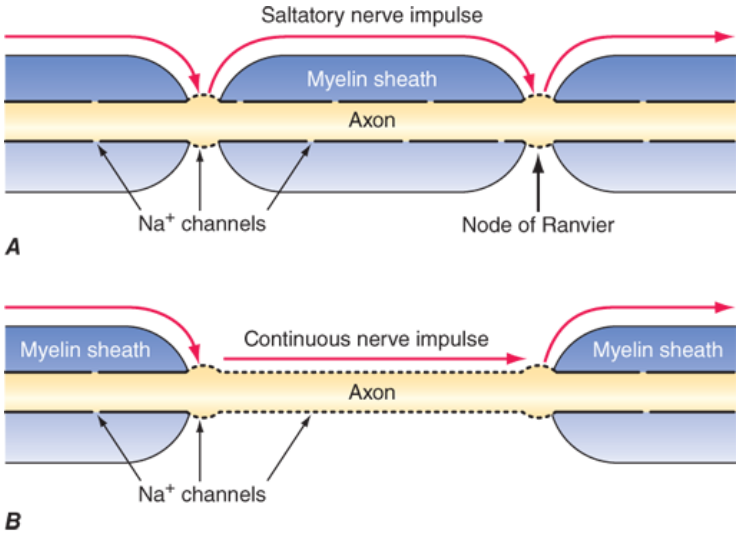
Inflammation is always present when active demyelination or axonal injury occurs, and the presence of T-cell and B-cell infiltration is related to the extent of demyelination and axonal injury. However, the nature of the inflammatory response appears to be somewhat different between early and later stages of MS. In relapsing MS, inflammation is associated with focal perivenular parenchymal infiltration of lymphocytes and monocytes associated with BBB disruption and active demyelination. In contrast, inflammation in progressive MS is more diffuse and is characterized by widespread microglial activation. Acute perivascular infiltrates are fewer in number, and lymphocytes and monocytes in chronic MS plaques aggregate at the lesion border suggesting ongoing inflammatory injury at the lesion edge. In addition, a diffuse low-grade inflammation with microglial proliferation is observed across large areas of white matter, associated with reduced myelin staining and axonal injury ("dirty white matter"). Activated astrocytes induced by microglia may also contribute to tissue damage **(Chap. 417)**. These observations imply that ongoing inflammation occurs behind a partially repaired BBB in many patients with progressive MS, and this feature could explain the failure of immunotherapies not capable of crossing the BBB to benefit patients with progressive MS.

Physiology

Nerve conduction in myelinated axons occurs in a saltatory manner, with the nerve impulse jumping from one node of Ranvier to the next without depolarization of the axonal membrane underlying the myelin sheath between nodes **(Fig. 436-2)**. This produces considerably faster conduction velocities (~70 m/s) than the slow velocities (~1 m/s) produced by continuous propagation in unmyelinated nerves. Conduction block occurs when the nerve impulse is unable to traverse the demyelinated segment. This can happen when the resting axon membrane becomes hyperpolarized due to the exposure of voltage-dependent potassium channels that are normally buried underneath the myelin sheath. A temporary conduction block often follows a demyelinating event before sodium channels (originally concentrated at the nodes) redistribute along the naked axon (Fig. 436-2). This redistribution ultimately allows continuous propagation of nerve action potentials through the demyelinated segment. Conduction block may be incomplete, affecting high- but not low-frequency volleys of impulses. Variable conduction block can occur with raised body temperature or metabolic alterations and may explain clinical fluctuations that vary from hour to hour or appear with fever or exercise. Conduction slowing occurs when the demyelinated segments of the axonal membrane are reorganized to support continuous (slow) nerve impulse propagation.

FIGURE 436-2

Nerve conduction in myelinated and demyelinated axons. ***A.*** Saltatory nerve conduction in myelinated axons occurs with the nerve impulse jumping from one node of Ranvier to the next. Sodium channels (shown as breaks in the solid black line) are concentrated at the nodes where axonal depolarization occurs. ***B.*** Following demyelination, additional sodium channels are redistributed along the axon itself, thereby allowing continuous propagation of the nerve action potential despite the absence of myelin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IMMUNOLOGY

A proinflammatory autoimmune response directed against a component of CNS myelin, and perhaps other neural elements as well, remains the cornerstone of current concepts of MS pathogenesis.

AUTOREACTIVE T LYMPHOCYTES

Myelin basic protein (MBP), an intracellular protein involved in myelin compaction, is an important T cell antigen in experimental allergic encephalomyelitis (EAE), a laboratory model, and possibly also in human MS. Activated MBP-reactive T cells have been identified in the blood, in cerebrospinal fluid (CSF), and within MS lesions. Moreover, *DRB1* 15:01* may influence the autoimmune response because it binds with high affinity to a fragment of MBP (spanning amino acids 89–96), stimulating T cell responses to this self-protein. Two different populations of proinflammatory T cells are likely to mediate autoimmunity in MS. T-helper type 1 (T_H1) cells producing interferon γ (IFN-γ) are one key effector population, and a role for highly proinflammatory T_H17 T cells has also been established, at least in some patients. T_H17 cells are induced by transforming growth factor β (TGF-β) and IL-6 and are amplified by IL-21 and IL-23. T_H17 cells and levels of their corresponding cytokine IL-17 are increased in MS lesions. T_H1 cytokines, including IL-2, tumor necrosis factor (TNF)-α, and IFN-γ, also play key roles in activating and maintaining autoimmune responses, and TNF-α and IFN-γ may directly injure oligodendrocytes or the myelin membrane.

HUMORAL AUTOIMMUNITY

B cell activation and antibody responses are also involved in the development of demyelinating lesions. Clonally restricted populations of activated, antigen-experienced, memory B cells and plasma cells are present in MS lesions, in meningeal lymphoid follicle-like structures overlying the cerebral cortex, and in the CSF. Similar populations are found in each compartment, indicating that a highly focused B cell response occurs locally within the CNS. Myelin-specific autoantibodies, some directed against an extracellular myelin protein, myelin oligodendrocyte glycoprotein (MOG), have been detected bound to vesiculated myelin debris in MS plaques. In the CSF, elevated levels of locally synthesized immunoglobulins and oligoclonal antibodies, derived from clonally restricted CNS B cells and plasma cells, are also characteristic of MS. The pattern of oligoclonal banding is unique to each individual, and attempts to identify the targets of these antibodies have been largely unsuccessful. Moreover, when proteins recognized by CSF restricted oligoclonal bands (OCBs) have been found, they appear to recognize a variety of antigens including intracellular ubiquitous proteins. Therefore, although intrathecal OCBs and elevated intrathecal synthesis of immunoglobulins are characteristic of MS, their role in disease pathogenesis remains uncertain.

NEURODEGENERATION

Axonal damage occurs in every newly formed MS lesion, and cumulative axonal and neuronal loss is considered to be the most important contributor to irreversible neurologic disability. As many as 70% of axons are lost from the lateral corticospinal (e.g., motor) tracts in patients with advanced paraparesis from MS, and longitudinal MRI studies suggest that there is progressive axonal loss over time within established lesions. Demyelination can result in reduced trophic support for axons, redistribution of ion channels, and destabilization of action potential membrane potentials. Axons can adapt initially to these injuries, but over time distal and retrograde degeneration (“dying–back” axonopathy) occurs. Therefore, promoting remyelination to protect axons remains an important therapeutic goal.

In addition to white matter plaques and axonopathy, as noted above (see Pathology), recent studies in progressive MS have highlighted an important role for a primary injury to the cerebral cortex, perhaps related to overlying meningeal inflammation.

Data also support a role for one, or more likely several, of the following mechanisms in progressive MS. Axonal and neuronal death may result from glutamate-mediated excitotoxicity, oxidative injury, iron accumulation, and/or mitochondrial failure either occurring as a consequence of free-radical damage or due to accumulation of deletions in mitochondrial DNA.

DIAGNOSIS

There is no single diagnostic test for MS. Diagnostic criteria for clinically definite MS require documentation of two or more episodes of symptoms and two or more signs that reflect pathology in anatomically noncontiguous white matter tracts of the CNS (Table 436-3). Symptoms must last for >24 h and occur as distinct episodes that are separated by a month or more. In patients who have only one of the two required signs on neurologic examination, the second may be documented by abnormal tests such as MRI or evoked potentials (EPs). Similarly, in the most recent diagnostic scheme, the second clinical event (in time) may be supported solely by MRI findings, consisting of either the development of new focal white matter lesions on MRI or the simultaneous presence of both an enhancing lesion and a nonenhancing lesion in an asymptomatic location. In patients whose course is progressive from onset for ≥6 months without superimposed relapses, documentation of intrathecal IgG synthesis may be used to support a diagnosis of PPMS.

TABLE 436-3

Diagnostic Criteria for Multiple Sclerosis (MS)

Clinical Presentation	Additional Data Needed for MS Diagnosis
2 or more attacks; objective clinical evidence of 2 or more lesions or objective clinical evidence of 1 lesion with reasonable historical evidence of a prior attack	None
2 or more attacks; objective clinical evidence of 1 lesion	Dissemination in space, demonstrated by • ≥1 T2 lesion on MRI in at least 2 out of 4 MS-typical regions of the CNS (periventricular, juxtacortical, infratentorial, or spinal cord) OR • Await a further clinical attack implicating a different CNS site
1 attack; objective clinical evidence of 2 or more lesions	Dissemination in time, demonstrated by • Simultaneous presence of asymptomatic gadolinium-enhancing and nonenhancing lesions at any time OR • A new T2 and/or gadolinium-enhancing lesion(s) on follow-up MRI, irrespective of its timing with reference to a baseline scan OR • Await a second clinical attack
1 attack; objective clinical evidence of 1 lesion (clinically isolated syndrome)	Dissemination in space and time, demonstrated by: For dissemination in space • ≥1 T2 lesion in at least 2 out of 4 MS-typical regions of the CNS (periventricular, juxtacortical, infratentorial, or spinal cord) OR • Await a second clinical attack implicating a different CNS site AND For dissemination in time • Simultaneous presence of asymptomatic gadolinium-enhancing and nonenhancing lesions at any time OR • A new T2 and/or gadolinium-enhancing lesion(s) on follow-up MRI, irrespective of its timing with reference to a baseline scan OR • Await a second clinical attack
Insidious neurologic progression suggestive of MS (PPMS)	1 year of disease progression (retrospectively or prospectively determined) PLUS 2 out of the 3 following criteria Evidence for dissemination in space in the brain based on ≥1 T2+ lesions in the MS-characteristic periventricular, juxtacortical, or infratentorial regions Evidence for dissemination in space in the spinal cord based on ≥2 T2+ lesions in the cord Positive CSF (isoelectric focusing evidence of oligoclonal bands and/or elevated IgG index)

Abbreviations: CNS, central nervous system; CSF, cerebrospinal fluid; MRI, magnetic resonance imaging; PPMS, primary progressive multiple sclerosis.

Source: From CH Polman et al: Diagnostic criteria for multiple sclerosis: 2010 Revisions to the "McDonald Criteria." Ann Neurol 69:292, 2011.

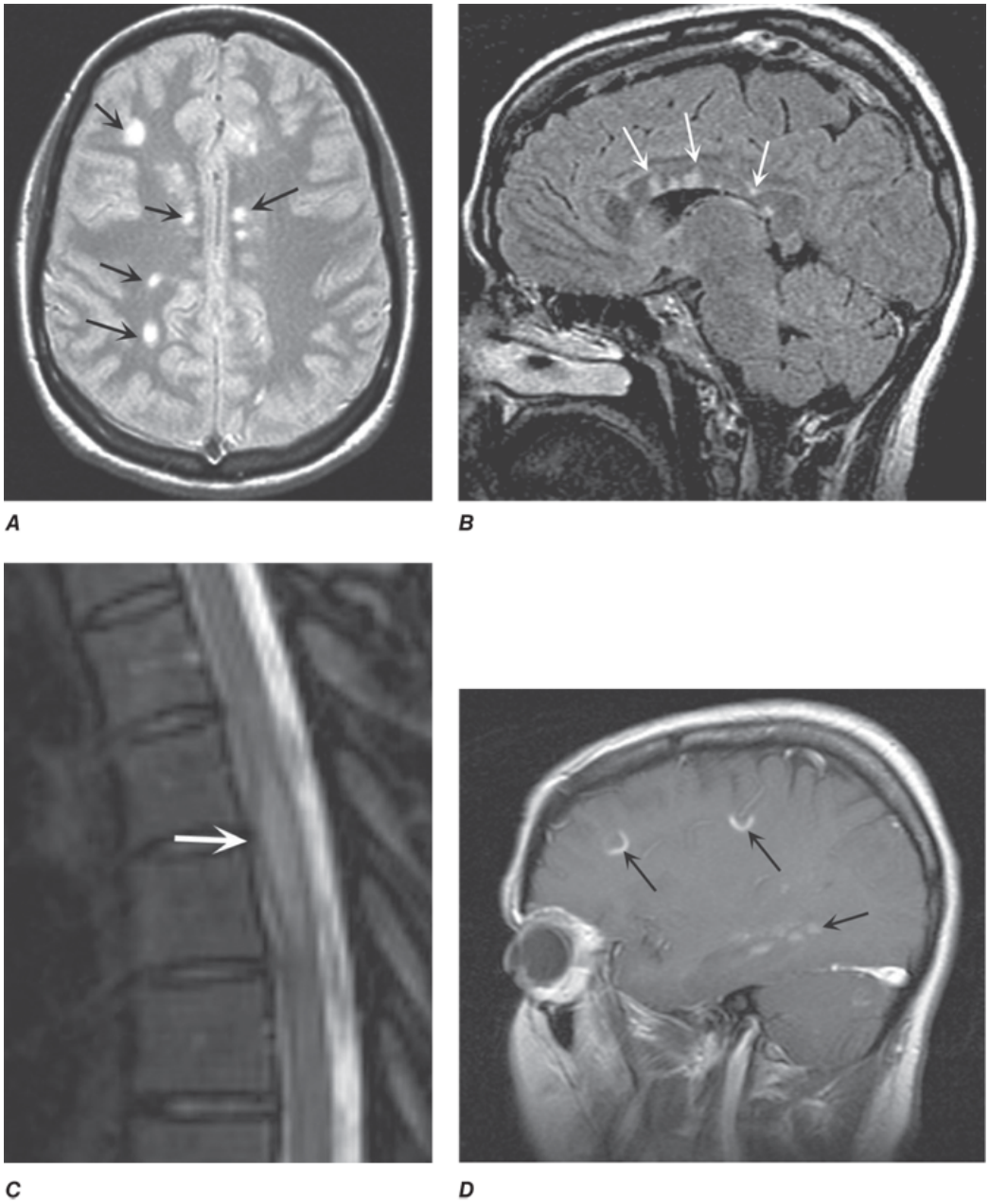
DIAGNOSTIC TESTS

Magnetic Resonance Imaging

MRI has revolutionized the diagnosis and management of MS **(Fig. 436-3)**; characteristic abnormalities are found in >95% of patients, although >90% of the lesions visualized by MRI are asymptomatic. An increase in vascular permeability from a breakdown of the BBB is detected by leakage of intravenous gadolinium (Gd) into the parenchyma. Such leakage occurs early in the development of an MS lesion and serves as a useful marker of inflammation. Gd enhancement typically persists for <1 month, and the residual MS plaque remains visible indefinitely as a focal area of hyperintensity (a lesion) on T2-weighted images. Lesions are frequently oriented perpendicular to the ventricular surface, corresponding to the pathologic pattern of perivenous demyelination (Dawson's fingers). Lesions are multifocal within the brain, brainstem, and spinal cord. Lesions >6 mm located in the corpus callosum, periventricular white matter, brainstem, cerebellum, or spinal cord are particularly helpful diagnostically. Current criteria for the use of MRI in the diagnosis of MS are shown in Table 436-3.

FIGURE 436-3

Magnetic resonance imaging findings in multiple sclerosis (MS). ***A.*** Axial first-echo image from T2-weighted sequence demonstrates multiple bright signal abnormalities in white matter, typical for MS. ***B.*** Sagittal T2-weighted fluid-attenuated inversion recovery (FLAIR) image in which the high signal of cerebrospinal fluid (CSF) has been suppressed. CSF appears dark, whereas areas of brain edema or demyelination appear high in signal, as shown here in the corpus callosum (*arrows*). Lesions in the anterior corpus callosum are frequent in MS and rare in vascular disease. ***C.*** Sagittal T2-weighted fast spin echo image of the thoracic spine demonstrates a fusiform high-signal-intensity lesion in the midthoracic spinal cord. ***D.*** Sagittal T1-weighted image obtained after the intravenous administration of gadolinium DTPA reveals focal areas of blood-brain barrier disruption, identified as high-signal-intensity regions (*arrows*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Serial MRI studies in early relapsing-remitting MS reveal that bursts of focal inflammatory disease activity occur far more frequently than would have been predicted by the frequency of relapses. Thus, early in MS, most disease activity is clinically silent.

The total volume of T2-weighted signal abnormality (the “burden of disease”) shows a significant (albeit weak) correlation with clinical disability. Quantitative measures of brain and spinal cord atrophy are evidence of diffuse tissue injury and correlate more strongly with measures of disability or progressive MS. Serial MRI studies also indicate that progressive whole brain atrophy occurs even in very early MS and continues throughout the disease course. Approximately one-third of T2-weighted lesions appear as hypointense lesions (black holes) on T1-weighted imaging. Black holes may be a marker of irreversible demyelination and axonal loss, although even this measure depends on the timing of the image acquisition (e.g., most acute Gd-enhancing T2 lesions are T1 dark).

Evoked Potentials

EP testing assesses function in afferent (visual, auditory, and somatosensory) or efferent (motor) CNS pathways. EPs use computer averaging to measure CNS electric potentials evoked by repetitive stimulation of selected peripheral nerves or of the brain. These tests provide the most information when the pathways studied are clinically uninvolved. For example, in a patient with a relapsing spinal cord syndrome with sensory deficits in the legs, an abnormal somatosensory EP following posterior tibial nerve stimulation provides little new information. By contrast, an abnormal visual EP in this circumstance would permit a diagnosis of clinically definite MS (Table 436-3). Abnormalities on one or more EP modalities occur in 80–90% of MS patients. EP

abnormalities are not specific to MS, although a marked delay in the latency of a specific EP component (as opposed to a reduced amplitude or distorted wave-shape) is suggestive of demyelination.

Cerebrospinal Fluid

CSF abnormalities found in MS include a mononuclear cell pleocytosis and an increased level of intrathecally synthesized IgG. The total CSF protein is usually normal or mildly elevated. Various formulas distinguish intrathecally synthesized IgG from IgG that entered the CNS passively from the serum. One formula, the CSF IgG index, expresses the ratio of IgG to albumin in the CSF divided by the same ratio in the serum. The IgG synthesis rate uses serum and CSF IgG and albumin measurements to calculate the rate of CNS IgG synthesis. The measurement of OCBs by agarose gel electrophoresis in the CSF also assesses intrathecal production of IgG. Two or more discrete OCBs, not present in a paired serum sample, are found in >75% of patients with MS. OCBs may be absent at the onset of MS, and in individual patients, the number of bands may increase with time.

A mild CSF pleocytosis (>5 cells/μL) is present in ~25% of cases, usually in young patients with RMS. A pleocytosis of >75 cells/μL, the presence of polymorphonuclear leukocytes, or a protein concentration >1 g/L (>100 mg/dL) in CSF should raise concern that the patient may not have MS.

DIFFERENTIAL DIAGNOSIS

The possibility of an alternative diagnosis should always be considered **(Table 436-4)**, particularly when (1) symptoms are localized exclusively to the posterior fossa, craniocervical junction, or spinal cord; (2) the patient is <15 or >60 years of age; (3) the clinical course is progressive from onset; (4) the patient has never experienced visual, sensory, or bladder symptoms; or (5) laboratory findings (e.g., MRI, CSF, or EPs) are atypical. Similarly, uncommon or rare symptoms in MS (e.g., aphasia, parkinsonism, chorea, isolated dementia, severe muscular atrophy, peripheral neuropathy, episodic loss of consciousness, fever, headache, seizures, or coma) should increase concern about an alternative diagnosis. Diagnosis is also difficult in patients with a rapid or explosive (stroke-like) onset or with mild symptoms and a normal neurologic examination. Rarely, intense inflammation and swelling may produce a mass lesion that mimics a primary or metastatic tumor. Disorders possibly mistaken for MS include: neuromyelitis optica **(Chap. 437)**, sarcoidosis, vascular disorders (antiphospholipid syndrome and vasculitis), rarely CNS lymphoma and still more rarely infections such as syphilis or Lyme disease. The specific tests required to exclude alternative diagnoses will vary with each clinical situation; however, an erythrocyte sedimentation rate, serum B_{12} level, anti-nuclear antibodies, and treponemal antibody should probably be obtained in all patients with suspected MS.

TABLE 436-4

Disorders That Can Mimic Multiple Sclerosis (MS)

Acute disseminated encephalomyelitis (ADEM)
Antiphospholipid antibody syndrome
Behçet's disease
Cerebral autosomal dominant arteriopathy, subcortical infarcts, and leukoencephalopathy (CADASIL)
Congenital leukodystrophies (e.g., adrenoleukodystrophy, metachromatic leukodystrophy)
Human immunodeficiency virus (HIV) infection
Ischemic optic neuropathy (arteritic and nonarteritic)
Lyme disease
Mitochondrial encephalopathy with lactic acidosis and stroke (MELAS)
Neoplasms (e.g., lymphoma, glioma, meningioma)
Sarcoid
Sjögren's syndrome
Stroke and ischemic cerebrovascular disease
Syphilis
Systemic lupus erythematosus and related collagen vascular disorders
Tropical spastic paraparesis (HTLV-1/2 infection)
Vascular malformations (especially spinal dural AV fistulas)
Vasculitis (primary CNS or other)
Vitamin B_{12} deficiency

Abbreviations: AV, arteriovenous; CNS, central nervous system; HTLV, human T cell lymphotropic virus.

PROGNOSIS

Most patients with clinically evident MS ultimately experience progressive neurologic disability. In older studies conducted before disease-modifying therapies for MS were available, 15 years after onset, only 20% of patients had no functional limitation, and between one-third and one-half of RMS patients progressed to SPMS and required assistance with ambulation; furthermore, 25 years after onset, ~80% of MS patients reached this level of disability. The long-term prognosis for MS has improved substantially in recent years, and transition from RMS to SPMS now occurs at approximately a 1% annual rate compared with 2–3% in the pretreatment era. This improvement is almost certainly due, at least in part, to widespread use of disease modifying therapies for RMS. Although the prognosis in an individual is difficult to establish, certain clinical features suggest a more favorable prognosis. These include ON or sensory symptoms at onset; fewer than two relapses in the first year of illness; and minimal impairment after 5 years. By contrast, patients with truncal ataxia, action tremor, pyramidal symptoms, or a progressive disease course are more likely to become disabled. Patients with a long-term favorable course are likely to have developed fewer MRI lesions and have less brain atrophy during the early years of disease, and vice versa. Importantly, some MS patients have a benign variant of MS and never develop neurologic disability. The likelihood of having benign MS is thought to be <10%. Patients with benign MS 15 years after onset who have entirely normal neurologic examinations are likely to maintain their benign course.

In patients with their first demyelinating event (i.e., a clinically isolated syndrome), the brain MRI provides prognostic information. With three or more typical T2-weighted lesions, the risk of developing MS after 20 years is ~80%. Conversely, with a normal brain MRI, the likelihood of developing MS is <20%. Similarly, the presence of two or more Gd-enhancing lesions at baseline is highly predictive of future MS, as is the appearance of either new T2-weighted lesions or new Gd enhancement ≥3 months after the initial episode.

Effect of Pregnancy

Pregnant MS patients experience fewer attacks than expected during gestation (especially in the last trimester), but more attacks than expected in the first 3 months postpartum. When considering the pregnancy year as a whole (i.e., 9 months of pregnancy plus 3 months postpartum), the overall disease course is unaffected. Decisions about childbearing should thus be made based on (1) the mother's physical state, (2) her ability to care for the child, and (3) the availability of social support. Disease-modifying therapy is generally discontinued during pregnancy, although the actual risk from the interferons and glatiramer acetate (see below) appears to be low.

TREATMENT

TREATMENT

Multiple Sclerosis

Therapy for MS can be divided into several categories: (1) treatment of acute attacks, (2) treatment with disease-modifying agents that reduce the biologic activity of MS, and (3) symptomatic therapy. Treatments that promote remyelination or neural repair do not currently exist, but several promising approaches are being actively investigated.

The Expanded Disability Status Score (EDSS) is a widely used measure of neurologic impairment in MS **(Table 436-5)**. Most patients with EDSS scores <3.5 walk normally, and are generally not disabled; by contrast, patients with EDSS scores >4.0 have progressive MS (SPMS or PPMS), are gait-impaired, and often are occupationally disabled.

ACUTE ATTACKS OR INITIAL DEMYELINATING EPISODES

When patients experience acute deterioration, it is important to consider whether this change reflects new disease activity or a "pseudoexacerbation" resulting from an increase in ambient temperature, fever, or an infection. When the clinical change is thought to reflect a pseudoexacerbation, glucocorticoid treatment is inappropriate. Glucocorticoids are used to manage either first attacks or acute exacerbations. They provide short-term clinical benefit by reducing the severity and shortening the duration of attacks. Whether treatment provides any long-term benefit on the course of the illness is less clear. Therefore, mild attacks are often not treated. Physical and occupational therapy can help with mobility and manual dexterity.

Glucocorticoid treatment is usually administered as intravenous methylprednisolone, 500–1000 mg/d for 3–5 days, either without a taper or followed by a course of oral prednisone beginning at a dose of 60–80 mg/d and gradually tapered over 2 weeks. Orally administered methylprednisolone, prednisone, or dexamethasone (in equivalent dosages) can be substituted for the intravenous portion of the therapy. Outpatient treatment is almost always possible.

Side effects of short-term glucocorticoid therapy include fluid retention, potassium loss, weight gain, gastric disturbances, acne, and emotional lability. Concurrent use of a low-salt, potassium-rich diet and avoidance of potassium-wasting diuretics are advisable. Lithium carbonate (300 mg orally bid) may help manage emotional lability and insomnia associated with glucocorticoid therapy. Patients with a history of peptic ulcer disease may require cimetidine (400 mg bid) or ranitidine (150 mg bid). Proton pump inhibitors such as pantoprazole (40 mg orally bid) may reduce the likelihood of gastritis, especially when large doses are administered orally. Plasma exchange (five to seven exchanges: 40–60 mL/kg per exchange, every other day for 14 days) may benefit patients with fulminant attacks of demyelination that are unresponsive to glucocorticoids. However, the cost is high, and conclusive evidence of efficacy is lacking.

DISEASE-MODIFYING THERAPIES FOR RELAPSING FORMS OF MS (RMS, SPMS WITH EXACERBATIONS)

More than a dozen immunomodulatory and immunosuppressive agents are approved by regulatory bodies for treatment of RMS. In phase 3 clinical trials, each was shown to reduce the frequency of clinical relapses and evolution of new brain MRI lesions in relapsing forms of MS **(Table 436-6)**. Each can also be used in SPMS patients who continue to experience attacks, both because SPMS can be difficult to distinguish from relapsing MS and because the available clinical trials, although not definitive, suggest that such patients may sometimes derive therapeutic benefit. When considering the data in Table 436-6, however, it is important to note that the relative efficacy of the different agents has not been directly tested in head-to head studies and that cross-trial comparisons are inaccurate. However, given the increasingly complex landscape of therapeutics for MS, for convenience the discussion of these agents has been divided into those used more and less frequently; and also by an estimate of their relative (modest, moderate or high) perceived level of efficacy. These are meant to serve as only very rough guides, and considerable variance exists in practice patterns, as well as availability of these agents, in different parts of the world.

FREQUENTLY USED AGENTS FOR RMS

Interferon β (Modestly Effective)

Interferon β (IFN-β) is a class I interferon originally identified by its antiviral properties. Efficacy in MS probably results from immunomodulatory properties including (1) downregulating expression of MHC molecules on antigen-presenting cells, (2) reducing proinflammatory and increasing regulatory cytokine levels, (3) inhibiting T-cell proliferation, and (4) limiting the trafficking of inflammatory cells in the CNS. IFN-β reduces the attack rate and slows accumulation of disability and MRI-documented disease burden. IFN-β should be considered in patients with either relapsing forms of MS (either RMS or SPMS with superimposed relapses). Head-to-head trials suggest that dosing IFN-β more frequently and at higher doses has better efficacy but is also more likely to induce neutralizing antibodies (see below). IFN-β-1a (Avonex), 30 µg, is administered by intramuscular injection once every week. IFN-β-1a (Rebif), 44 µg, is administered by subcutaneous injection three times per week. IFN-β-1b (Betaseron or Extavia), 250 µg, is administered by subcutaneous injection every other day. Pegylated IFN-β-1a (Plegridy), 125 µg, is administered by subcutaneous injection once every 14 days. Pegylated IFN-β-1a is an interferon to which a single, linear 20,000 dalton methoxy poly(ethyleneglycol)-O-2-methylpopionaldehyde molecule is covalently attached; the pegylated molecule contributes to reduced in vivo clearance allowing less frequent administration.

Common side effects of IFN-β therapy include flulike symptoms (e.g., fevers, chills, and myalgias) and mild abnormalities on routine laboratory evaluation (e.g., elevated liver function tests or lymphopenia). Rarely, more severe hepatotoxicity may occur. Subcutaneous IFN-β also causes reactions at the injection site (e.g., pain, redness, induration, or, rarely, skin necrosis). Side effects can usually be managed with concomitant nonsteroidal anti-inflammatory medications. Depression, increased spasticity, and cognitive changes have been reported, although these symptoms can also be due to the underlying disease. Side effects due to IFN-β therapy usually subside over time.

Approximately 2–10% of IFN-β-1a (Avonex) recipients, 15–25% of IFN-β-1a (Rebif) recipients, and 30–40% of IFN-β-1b (Betaseron/Extavia) recipients develop neutralizing antibodies to IFN-β, which may disappear over time. Less than 1% of patients treated with pegylated IFN-β-1a develop neutralizing antibodies. For a patient doing well on therapy, the presence of antibodies should not affect treatment. Conversely, for a patient doing poorly on therapy, alternative treatment should be considered, even if there are no detectable antibodies.

Glatiramer Acetate (Modestly Effective)

Glatiramer acetate is a synthetic, random polypeptide composed of four amino acids (L-glutamic acid, L-lysine, L-alanine, and L-tyrosine). Its mechanism of action may include (1) induction of antigen-specific suppressor T cells; (2) binding to MHC molecules, thereby displacing bound MBP; or (3) altering the balance between proinflammatory and regulatory cytokines. Glatiramer acetate reduces the attack rate (whether measured clinically or by MRI) in RMS. Glatiramer acetate also benefits disease severity measures, although, for clinical disability, this is less well established than for IFN-β. Nevertheless, two head-to-head trials demonstrated that the impact of glatiramer acetate on clinical relapse rates and disability was comparable to high-dose, high-frequency IFN-β. Therefore, glatiramer acetate should be considered as an equally effective alternative to IFN-β in RMS patients. Its usefulness in progressive disease is unknown. Glatiramer acetate is administered by subcutaneous injection of either 20 mg every day or 40 mg thrice weekly. Injection-site reactions also occur with glatiramer acetate. In addition, ~15% of patients experience one or more episodes of flushing, chest tightness, dyspnea, palpitations, and anxiety after injection. This systemic reaction is unpredictable, brief (duration <1 h), and tends not to recur. Finally, some patients experience lipoatrophy, which, on occasion, can be disfiguring and require cessation of treatment. Recently, glatiramer acetate was U.S. Food and Drug Administration (FDA) approved as a biosimilar medication (Glatopa) and is dosed at 20 mg every day. Although clinical trials were not performed with biosimilar glatiramer acetate, the efficacy and safety are presumed to be similar to the branded product.

Fingolimod (Moderately Effective)

Fingolimod is a sphingosine-1-phosphate (S1P) inhibitor that prevents the egress of lymphocytes from secondary lymphoid organs such as the lymph nodes and spleen. Its mechanism of action is probably due to sequestration of lymphocytes in the periphery, thereby inhibiting their trafficking to the CNS. Fingolimod reduces the attack rate and significantly improves all measures of disease severity in MS. It is well tolerated, and the daily oral dosing schedule makes it convenient for patients. A head-to-head phase 3 randomized study demonstrated the superiority of fingolimod over low-dose (weekly) IFN-β-1a.

Fingolimod, 0.5 mg, is administered orally each day. Treatment with fingolimod is well tolerated. Mild abnormalities on routine laboratory evaluation (e.g., elevated liver function tests or lymphopenia) are more common than in controls, sometimes requiring discontinuation of the medication. First- and second-degree heart block and bradycardia can also occur when fingolimod therapy is initiated. A 6-h period of observation (including electrocardiogram monitoring) is recommended for all patients receiving their first dose. Other side effects include macular edema and, rarely, disseminated varicella-zoster virus (VZV) and cryptococcal infections; prior to initiating therapy with fingolimod, an ophthalmic examination and VZV vaccination for seronegative individuals are indicated. Fingolimod can also cause QT prolongation with the potential for drug-drug interactions with other medications that also prolong the QT interval.

Dimethyl Fumarate (DMF) (Moderately Effective)

DMF is a small molecule and is a Krebs cycle metabolite with anti-inflammatory effects in psoriasis. Although the precise mechanisms of action of DMF are not fully understood, it seems to modulate the expression of proinflammatory and anti-inflammatory cytokines. Also, DMF inhibits the ubiquitylation and degradation of nuclear factor E2-related factor 2 (Nrf2)—a transcription factor that binds to the antioxidant response elements (AREs) located on the DNA and thereby induces the transcription of several antioxidant proteins. DMF reduces the attack rate and significantly improves all measures of disease severity in MS patients. However, its twice-daily oral dosing schedule makes it somewhat less convenient for patients than daily oral therapies. In

addition, compliance is likely to be less with a twice-daily dosing regimen—a factor that could be of concern given the observation (in a small clinical trial) that once-daily DMF lacks efficacy. A head-to-head trial provided evidence that DMF was superior to glatiramer acetate on some outcome measures.

DMF, 240 mg, is administered orally twice each day. Gastrointestinal side effects (abdominal discomfort, nausea, vomiting, flushing, and diarrhea) are common at the start of therapy but generally subside with continued administration. Other adverse events include flushing mild decreases in neutrophil and lymphocyte counts and elevations in liver enzymes. Nevertheless, in general, treatment with DMF is well tolerated after an initial period of adjustment. Following the release of DMF, several cases of progressive multifocal leukoencephalopathy (PML) were reported in patients receiving products that contained DMF. Most of these patients were lymphopenic and monitoring for lymphopenia every 6 months is recommended. Patients who are persistently lympopenic (lymphocyte count <500 cells/mL) are recommended to consider alternate treatments due to the PML risk. Clinically significant liver injury has been reported with DMF treatment. Liver function tests should be assessed before treatment and when clinically indicated. Elevations in liver function tests resolve the following treatment discontinuation.

Natalizumab (Highly Effective)

Natalizumab is a humanized monoclonal antibody directed against the α_4 subunit of $\alpha_4\beta_1$ integrin, a cellular adhesion molecule expressed on the surface of lymphocytes. It prevents lymphocytes from binding to endothelial cells, thereby preventing lymphocytes from penetrating the BBB and entering the CNS. Natalizumab is highly effective in reducing the attack rate and significantly improves all measures of disease severity in MS (both clinical and MRI). Moreover, it is well-tolerated, and the dosing schedule of monthly intravenous infusions makes it very convenient for patients. Natalizumab, 300 mg, is administered by IV infusion each month. Treatment with natalizumab is, in general, well tolerated. A small percentage (<10%) of patients experience hypersensitivity reactions (including anaphylaxis), and ~6% develop neutralizing antibodies to the molecule (only half of which persist).

The major concern with long-term treatment is the risk of PML, a life-threatening condition resulting from infection by the John Cunningham (JC) virus. PML has occurred in ~0.4% of patients treated with natalizumab. The incidence of PML is very low in the first year of treatment but then rises in subsequent years of treatment to reach a level of about 2 cases per 1000 patients per year. Nevertheless, the measurement of antibodies against the JC virus in the serum can be used to stratify this risk. Approximately half of the adult population is JC antibody positive, indicating that they experienced an asymptomatic infection with the JC virus at some time in the past. Thus, in patients who do not have these antibodies, the risk of PML is minimal (<1:10,000 as long as they remain JC antibody free). Conversely, in patients who have these antibodies (especially those who have them in high titer), the risk may be as high as ≥1.1%. Up to 2% of seronegative MS patients undergoing treatment with natalizumab seroconvert annually; thus, it is recommended that JC antibody status be assessed at 6-month intervals in all patients receiving natalizumab treatment. In antibody-positive patients, a change to another disease-modifying therapy should be strongly considered. The risk of PML is also high in patients who previously received immunosuppressive therapy. Natalizumab is generally recommended only for JC antibody–negative patients, unless they have failed alternative therapies or if they have a particularly aggressive disease course.

Ocrelizumab (Highly Effective)

Ocrelizumab is a humanized monoclonal antibody directed against the CD20 molecule present on the surface of mature B cells. CD20 is not expressed on early B-cell precursors or on antibody-producing plasma cells, thus treatment with ocrelizumab selectively depletes mature B cells while preserving preexisting humoral immunity and the capacity for B-cell reconstitution by lymphoid stem cells. Ocrelizumab rapidly depletes circulating B cells through antibody-dependent cellular toxicity and complement-dependent cytotoxicity. The beneficial effects of B-cell depletion in MS are not fully understood but may involve interruption in trafficking of B cells from the periphery to the CNS and through reduction in antigen presentation and/or modulation of cytokine secretion by B cells. Ocrelizumab targets the same molecule as rituximab, a monoclonal antibody indicated for non-Hodgkin's lymphoma and rheumatoid arthritis, and ofatumumab, indicated for treatment of chronic lymphocytic leukemia. In two phase 3 trials, ocrelizumab demonstrated a high degree of efficacy against RMS, reducing annualized relapse rates by 47%, reducing new MRI lesions by 95%, and improving other measures of inflammatory and degenerative disease activity, compared with three times per week interferon β-1a (Rebif). Ocrelizumab 600 mg is administered by intravenous infusion every 24 weeks (administered as two 300-mg infusions spaced 2 weeks apart for the first dose, and as a single 600-mg infusion thereafter); intravenous methylprednisolone 100 mg is given prior to each infusion and optional prophylaxis with analgesics/antipyretics and antihistamines are recommended, along with adjustment of the infusion rate to manage infusion-related reactions.

Ocrelizumab is generally well tolerated with infusion-related reactions occurring in a minority of patients; these are most often observed with the first infusion and are usually mild in degree. Rituximab is associated with a very small risk of PML (estimated at <1:25,000/year), thus it is possible that ocrelizumab will also carry a nonzero risk. Ocrelizumab may also carry some risk of increased malignancies including breast cancer, although rituximab is not associated with an increased risk of malignancy. Additional study of ocrelizumab in the postmarketing setting is needed to determine whether there is in fact a fractional increased malignancy risk associated with ocrelizumab.

LESS COMMONLY USED AGENTS FOR RMS

Teriflunomide (Modestly Effective)

Teriflunomide inhibits the mitochondrial enzyme dihydro-orotate dehydrogenase, which is a key part of the pathway for de novo pyrimidine biosynthesis from carbamoyl phosphate and aspartate. It is the active metabolite of the drug leflunomide (FDA-approved for rheumatoid arthritis), and it exerts its anti-inflammatory effects by limiting the proliferation of rapidly dividing T and B cells. This enzyme is not involved in the so-called "salvage pathway," by which existing pyrimidine pools are recycled for DNA and RNA synthesis in resting and homeostatically proliferating cells. Consequently, teriflunomide is

considered to be cytostatic rather than cytotoxic. Teriflunomide reduces the attack rate and significantly improves all measures of disease severity in MS patients. It is well tolerated, and its daily oral dosing schedule makes it very convenient for patients. A head-to-head trial suggested the equivalence, but not superiority, of teriflunomide and thrice-weekly IFN-β-1a. Teriflunomide, either 7 or 14 mg, is administered orally each day. In the pivotal clinical trials, mild hair thinning and gastrointestinal symptoms (nausea and diarrhea) were more common than in controls, but in general, treatment with teriflunomide was well tolerated. Teriflunomide rarely causes toxic epidermal necrolysis or Stevens-Johnson syndrome. A major limitation, especially in women of childbearing age, is its possible teratogenicity (pregnancy category X); teriflunomide can remain in the bloodstream for 2 years due to hepatobiliary reabsorption. Therefore, it is recommended that exposed men and women who wish to conceive receive cholestyramine or activated charcoal to eliminate residual drug.

Alemtuzumab (Highly Effective)

Alemtuzumab is a humanized monoclonal antibody directed against the CD52 antigen that is expressed on both monocytes and lymphocytes. It causes lymphocyte depletion (of both B and T cells) and a change in the composition of lymphocyte subsets. Both of these changes, particularly the impact on lymphocyte subsets, are long lasting. In two phase 3 trials, which used the active comparator of thrice-weekly, high-dose IFN-β-1a, alemtuzumab markedly reduced the attack rate and significantly improved measures of disease severity in MS patients although its impact on clinical disability was found in only one of the two trials. The European and Canadian drug agencies were the first to approve this agent for use in RMS; the FDA has also approved alemtuzumab, but only after an appeal following initial disapproval. The reasons for the initial disapproval were based on a perceived lack of a convincing disability effect and concerns over potential toxicity. The toxicities of concern were the occurrence of (1) autoimmune diseases including thyroiditis, Graves' disease, thrombocytopenia, hemolytic anemia, pancytopenia, antiglomerular basement membrane disease, and membranous glomerulonephritis; (2) malignancies including thyroid cancer, melanoma, breast cancer, human papillomavirus (HPV)–related cancers, and lymphoproliferative disorders including lymphoma; (3) serious infections; and (4) infusion reactions. Because of its toxicity profile, the FDA indicated alemtuzumab only in patients who have tried and failed at least two other DMTs.

Mitoxantrone Hydrochloride (Highly Effective)

Mitoxantrone, an anthracenedione, exerts its antineoplastic action by (1) intercalating into DNA and producing both strand breaks and interstrand cross-links, (2) interfering with RNA synthesis, and (3) inhibiting topoisomerase II (involved in DNA repair). The FDA approved mitoxantrone on the basis of a single phase 3 clinical trial in Europe, in addition to an even smaller phase 2 studies. Mitoxantrone is indicated for use in patients with rapidly worsening MS (defined as patients whose neurologic status remains significantly abnormal between MS attacks). Despite this broad indication, however, data supporting its efficacy are less robust compared to other approved therapies.

Mitoxantrone is cardiotoxic (e.g., cardiomyopathy, reduced left ventricular ejection fraction, and irreversible congestive heart failure). As a result, a cumulative dose >140 mg/m^2 is not recommended. At currently approved doses (12 mg/m^2 every 3 months), the maximum duration of therapy can be only 2–3 years. Furthermore, >40% of women will experience amenorrhea, which may be permanent. Finally, there is risk of acute leukemia from mitoxantrone, estimated as at least a 1.4% lifetime risk. Because of these risks, and the availability of alternative therapies, mitoxantrone is now rarely used for MS.

Daclizumab, a monoclonal antibody against CD25, the α subunit of the interleukin 2 receptor, was removed from the market in 2018 because of reports of brain inflammation; it was previously approved for patients who had failed at least two other therapies.

Initiating and Changing Treatment

Previously, most patients with relapsing forms of MS received injectable agents (IFN-β or glatiramer acetate) as first-line therapy. However, with the introduction of effective oral agents that include dimethyl fumarate, fingolimod, and teriflunomide, this has begun to change. In addition, the monthly infusion therapy natalizumab, which is highly effective, well tolerated, and apparently safe in JC antibody–negative patients, provides an attractive option for many patients. Ocrelizumab can also be used first-line; the combination of high efficacy, infrequently administered infusions, and a favorable safety profile make its use an attractive option. With the exception of the first-generation injectable agents, mitoxantrone and natalizumab long-term (>10 year) safety data are not available for the newer therapies, and in many cases comparative data are lacking. The value of combination therapy is also largely unknown, although a clinical trial demonstrated no added benefit to the combination of glatiramer acetate with once-weekly IFN-β-1a.

Despite these unknowns, clinicians need to make decisions based on the best available evidence, coupled with practical considerations. One reasonable approach stratifies initial decision-making based on two levels of disease aggressiveness **(Fig. 436-4)**.

Mild Initial Course

In the case of recent onset, normal examination or minimal impairment (EDSS ≤2.5 or less), or low disease activity, either an injectable (IFN-β or glatiramer acetate) or an oral (DMF, fingolimod, or teriflunomide) agent is reasonable. Although head-to-head comparisons are not available, natalizumab is thought to be more effective than these other agents, and, therefore, this therapy can be considered even in minimally affected, JCV antibody–seronegative patients. Ocrelizumab is more effective than IFN-β-1a TIW and appears to be safe although long-term data are not available. Therefore, this therapy can also be considered in recent onset MS patients regardless of JCV serology status. The injectable agents (IFN-β and glatiramer acetate) have a superb track record for safety but have a high nuisance factor due to the need for frequent injections, as well as bothersome side effects that contribute to noncompliance. Some of the oral agents (DMF and fingolimod) are probably more effective than the injectables, but long-term risks are still unknown;

DMF produces flushing and bothersome gastrointestinal symptoms in many patients at least initially (can be mitigated by beginning at one-quarter strength and gradually advancing to full dose), and fingolimod can lead to bradycardia and other cardiac conduction disturbances of unclear clinical significance. Teriflunomide may be less effective than the other oral agents, and there are concerns about its possible long-lasting pregnancy risks as well as its association with toxic epidermal necrolysis and Stevens-Johnson syndrome. Nevertheless, its long-term safety is likely known because teriflunomide is the active metabolite of leflunomide—a drug long approved by the FDA for treatment of rheumatoid arthritis.

Moderate or Severe Initial Course

In highly active disease or moderate impairment (EDSS >2.5), either a highly effective oral agent (DMF or fingolimod) or ocrelizumab or, if the patient is JC virus antibody seronegative, infusion therapy with natalizumab is recommended.

Regardless of which agent is chosen first, treatment should probably be changed in patients who continue to have relapses, progressive neurologic impairment or, arguably, ongoing evidence of subclinical MRI activity (Fig. 436-4).

The long-term impact of these treatments on the disease course remains controversial, although several recent observational studies showed that these agents improve the long-term outcome of MS including a prolongation of the time to reach certain disability outcomes (e.g., SPMS and requiring assistance to ambulate) and reduction in MS-related mortality. These benefits seem most conspicuous when treatment begins early in the relapsing stage of the illness. Unfortunately, however, already established progressive symptoms do not respond well to treatment with these disease-modifying therapies. Because progressive symptoms are likely to result from accumulated axonal and neuronal loss, many experts now believe that very early treatment with a disease-modifying drug is appropriate for most MS patients. It may also be reasonable to delay initiating treatment in patients with (1) normal neurologic examinations, (2) a single attack or a low attack frequency, and (3) a low burden of disease as assessed by brain MRI. Untreated patients, however, should be followed closely with periodic brain MRI scans; the need for therapy is reassessed if scans reveal evidence of ongoing, subclinical disease. Finally, vitamin D deficiency should be corrected in all patients with MS, and generally this requires oral supplementation with vitamin D_3, 4000–5000 IU daily. Several clinical trials showed that supplementation with vitamin D in relapsing MS patients reduces MRI measures of disease activity and may also reduce the relapse frequency in patients actively treated with either interferon or glatiramer acetate.

DISEASE-MODIFYING THERAPIES FOR PROGRESSIVE MS

SPMS

High-dose IFN-β probably has a beneficial effect in patients with SPMS with active disease (see above). IFN-β is probably ineffective in patients with SPMS who do not have active disease. All of the other agents have not yet been studied in this patient population. Although mitoxantrone was approved for patients with progressive MS, this is not the population studied in the pivotal trial. Therefore, no evidence-based recommendation can be made with regard to its use in this setting.

PPMS

Ocrelizumab (see above), a humanized monoclonal antibody that targets CD20 B-cells, was shown in a single phase 3 trial to reduce progression of clinical disability in PPMS by 24%, and also to improve other clinical and MRI markers of inflammatory and degenerative disease activity, compared with placebo treatment. Ocrelizumab represents the first agent to convincingly modify the course of PPMS. The dosing of ocrelizumab for PPMS is identical as for RMS (above).

OFF-LABEL TREATMENT OPTIONS FOR RMS AND SPMS

Azathioprine (2–3 mg/kg per day) has been used primarily in relapsing MS. Meta-analysis of published trials suggests that azathioprine is marginally effective at lowering relapse rates, although a benefit on disability progression has not been demonstrated.

Methotrexate (7.5–20 mg/week) was shown in one study to slow the progression of upper extremity dysfunction in SPMS. Because of the possibility of developing irreversible liver damage, some experts recommend a blind liver biopsy after 2 years of therapy.

Cyclophosphamide (700 mg/m^2, every other month) may be helpful for treatment-refractory patients who are (1) otherwise in good health, (2) ambulatory, and (3) <40 years of age. Because cyclophosphamide can be used for periods in excess of 3 years, it may be preferable to mitoxantrone in these circumstances.

Intravenous immunoglobulin (IVIg), administered in monthly pulses (up to 1 g/kg) for up to 2 years, appears to reduce annual exacerbation rates. However, its use is limited because of its high cost, questions about optimal dose, and uncertainty about its having any impact on long-term disability.

Methylprednisolone in one study, administered as monthly high-dose intravenous pulses, reduced disability progression (see above).

Hematopoietic stem cell transplantation appears to be highly effective in reducing the occurrence of relapses and may improve disability in relapsing MS. However, this procedure carries a significant mortality risk. Randomized trials with appropriate comparators are needed in order to position this procedure with respect to available pharmacological interventions.

OTHER THERAPEUTIC CLAIMS

Many purported treatments for MS have never been subjected to scientific scrutiny. These include dietary therapies (e.g., the Swank diet, the Paleo Diet, the Wahls diet), megadose vitamins, calcium orotate, bee stings, cow colostrum, hyperbaric oxygen, procarin (a combination of histamine and caffeine), chelation, acupuncture, acupressure, various Chinese herbal remedies, and removal of mercury-amalgam tooth fillings, among many others. Patients should avoid costly or potentially hazardous unproven treatments. Many such treatments lack biologic plausibility. For example, no reliable case of mercury poisoning resembling typical MS has ever been described, therefore challenging the notion that removal of mercury-amalgam tooth fillings would be beneficial.

Although potential roles for EBV, human herpesvirus (HHV) 6, or chlamydia have been suggested for MS, these reports are unconfirmed, and treatment with antiviral agents or antibiotics is not recommended.

Recently, chronic cerebrospinal insufficiency (CCSVI) was proposed as a cause of MS with vascular-surgical intervention recommended. However, multiple independent studies have subsequently failed to even approximate the initial claims, and patients should be strongly advised to avoid diagnostic procedures and potentially dangerous surgery for this condition.

SYMPTOMATIC THERAPY

For all patients, it is useful to encourage attention to a healthy lifestyle, including maintaining an optimistic outlook, a healthy diet, and regular exercise as tolerated (swimming is often well-tolerated because of the cooling effect of cold water). It is reasonable also to correct vitamin D deficiency with oral vitamin D.

Ataxia/tremor is often intractable. Clonazepam, 1.5–20 mg/d; primidone, 50–250 mg/d; propranolol, 40–200 mg/d; or ondansetron, 8–16 mg/d, may help. Wrist weights occasionally reduce tremor in the arm or hand. Thalamotomy and deep-brain stimulation have been tried with mixed success.

Spasticity and *spasms* may improve with physical therapy, regular exercise, and stretching. Avoidance of triggers (e.g., infections, fecal impactions, bed sores) is extremely important. Effective medications include baclofen (20–120 mg/d), diazepam (2–40 mg/d), tizanidine (8–32 mg/d), dantrolene (25–400 mg/d), and cyclobenzaprine hydrochloride (10–60 mg/d). For severe spasticity, a baclofen pump (delivering medication directly into the CSF) can provide substantial relief.

Weakness can sometimes be improved with the use of potassium channel blockers such as 4-aminopyridine (20 mg/d) and 3,4-di-aminopyridine (40–80 mg/d), particularly in the setting where lower extremity weakness interferes with the patient's ability to ambulate. The FDA approved extended release 4-aminopyridine (at 10 mg twice daily), and this can be obtained either as dalfampridine (Ampyra) or through a compounding pharmacy. The principal concern with the use of these agents is the possibility of inducing seizures at high doses.

Pain is treated with anticonvulsants (carbamazepine, 100–1000 mg/d; phenytoin, 300–600 mg/d; gabapentin, 300–3600 mg/d; or pregabalin, 50–300 mg/d), antidepressants (amitriptyline, 25–150 mg/d; nortriptyline, 25–150 mg/d; desipramine, 100–300 mg/d; or venlafaxine, 75–225 mg/d), or antiarrhythmics (mexiletine, 300–900 mg/d). If these approaches fail, patients should be referred to a comprehensive pain management program.

Bladder dysfunction management is best guided by urodynamic testing. Evening fluid restriction or frequent voluntary voiding may help *detrusor hyperreflexia*. If these methods fail, propantheline bromide (10–15 mg/d), oxybutynin (5–15 mg/d), hyoscyamine sulfate (0.5–0.75 mg/d), tolterodine tartrate (2–4 mg/d), or solifenacin (5–10 mg/d) may help. Coadministration of pseudoephedrine (30–60 mg) is sometimes beneficial.

Detrusor/sphincter dyssynergia may respond to phenoxybenzamine (10–20 mg/d) or terazosin hydrochloride (1–20 mg/d). Loss of reflex bladder wall contraction may respond to bethanechol (30–150 mg/d). However, both conditions often require catheterization.

Urinary tract infections should be treated promptly. Patients with postvoid residual urine volumes >200 mL are predisposed to infections. Prevention by urine acidification (with cranberry juice or vitamin C) inhibits some bacteria. Prophylactic administration of antibiotics is sometimes necessary but may lead to colonization by resistant organisms. Intermittent catheterization may help to prevent recurrent infections and reduce overflow incontinence.

Treatment of *constipation* includes high-fiber diets and fluids. Natural or other laxatives may help. *Fecal incontinence* may respond to a reduction in dietary fiber.

Depression should be treated. Useful drugs include the selective serotonin reuptake inhibitors (fluoxetine, 20–80 mg/d, or sertraline, 50–200 mg/d), the tricyclic antidepressants (amitriptyline, 25–150 mg/d; nortriptyline, 25–150 mg/d; or desipramine, 100–300 mg/d), and the nontricyclic antidepressants (venlafaxine, 75–225 mg/d).

Fatigue may improve with assistive devices, help in the home, or successful management of spasticity. Patients with frequent nocturia may benefit from anticholinergic medication at bedtime. Excessive daytime somnolence caused by MS may respond to amantadine (200 mg/d), methylphenidate (5–25 mg/d), modafinil (100–400 mg/d), or armodafinil (150–250 mg/d).

Cognitive problems may respond marginally to lisdexamfetamine (40 mg/d).

Paroxysmal symptoms respond dramatically to low-dose anticonvulsants (acetazolamide, 200–600 mg/d; carbamazepine, 50–400 mg/d; phenytoin, 50–300 mg/d; or gabapentin, 600–1800 mg/d).

Heat sensitivity may respond to heat avoidance, air-conditioning, or cooling garments.

Sexual dysfunction may be helped by lubricants to aid in genital stimulation and sexual arousal. Management of pain, spasticity, fatigue, and bladder/bowel dysfunction may also help. Sildenafil (50–100 mg), tadalafil (5–20 mg), or vardenafil (5–20 mg), taken 1–2 h before sex, are standard treatments for erectile dysfunction.

PROMISING EXPERIMENTAL THERAPIES

Numerous clinical trials are currently under way. These include studies on (1) selective oral sphingosine-1-phosphate receptor antagonists to sequester lymphocytes in secondary lymphoid organs; (2) high dose biotin to improve disability in progressive forms of MS; (3) molecules to promote remyelination; and (4) bone marrow transplantation.

TABLE 436-5

Scoring Systems for Multiple Sclerosis (MS)

Kurtzke Expanded Disability Status Score (EDSS)

0.0 = Normal neurologic examination (all grade 0 in functional status [FS])
1.0 = No disability, minimal signs in one FS (i.e., grade 1)
1.5 = No disability, minimal signs in more than one FS (more than one grade 1)
2.0 = Minimal disability in one FS (one FS grade 2, others 0 or 1)
2.5 = Minimal disability in two FS (two FS grade 2, others 0 or 1)
3.0 = Moderate disability in one FS (one FS grade 3, others 0 or 1) or mild disability in three or four FS (three/four FS grade 2, others 0 or 1) although fully ambulatory
3.5 = Fully ambulatory but with moderate disability in one FS (one grade 3) and one or two FS grade 2; or two FS grade 3; or five FS grade 2 (others 0 or 1)
4.0 = Ambulatory without aid or rest for ~500 m
4.5 = Ambulatory without aid or rest for ~300 m
5.0 = Ambulatory without aid or rest for ~200 m
5.5 = Ambulatory without aid or rest for ~100 m
6.0 = Unilateral assistance required to walk about 100 m with or without resting
6.5 = Constant bilateral assistance required to walk about 20 m without resting
7.0 = Unable to walk beyond about 5 m even with aid; essentially restricted to wheelchair; wheels self and transfers alone
7.5 = Unable to take more than a few steps; restricted to wheelchair; may need aid to transfer
8.0 = Essentially restricted to bed or chair or perambulated in wheelchair, but out of bed most of day; retains many self-care functions; generally has effective use of arms
8.5 = Essentially restricted to bed much of the day; has some effective use of arm(s); retains some self-care functions
9.0 = Helpless bed patient; can communicate and eat
9.5 = Totally helpless bed patient; unable to communicate or eat
10.0 = Death due to MS

Functional Status (FS) Score

A. Pyramidal functions
0 = Normal
1 = Abnormal signs without disability
2 = Minimal disability
3 = Mild or moderate paraparesis or hemiparesis, or severe monoparesis
4 = Marked paraparesis or hemiparesis, moderate quadriparesis, or monoplegia
5 = Paraplegia, hemiplegia, or marked quadriparesis
6 = Quadriplegia

B. Cerebellar functions
0 = Normal
1 = Abnormal signs without disability
2 = Mild ataxia
3 = Moderate truncal or limb ataxia
4 = Severe ataxia all limbs
5 = Unable to perform coordinated movements due to ataxia

C. Brainstem functions
0 = Normal
1 = Signs only
2 = Moderate nystagmus or other mild disability
3 = Severe nystagmus, marked extraocular weakness, or moderate disability of other cranial nerves
4 = Marked dysarthria or other marked disability
5 = Inability to swallow or speak

D. Sensory functions
0 = Normal
1 = Vibration or figure-writing decrease only, in 1 or 2 limbs
2 = Mild decrease in touch or pain or position sense, and/or moderate decrease in vibration in 1 or 2 limbs, or vibratory decrease alone in 3 or 4 limbs
3 = Moderate decrease in touch or pain or position sense, and/or essentially lost vibration in 1 or 2 limbs, or mild decrease in touch or pain, and/or moderate decrease in all proprioceptive tests in 3 or 4 limbs
5 = Loss (essentially) of sensation in 1 or 2 limbs or moderate decrease in touch or pain and/or loss of proprioception for most of the body below the head
6 = Sensation essentially lost below the head

E. Bowel and bladder functions
0 = Normal
1 = Mild urinary hesitancy, urgency, or retention
2 = Moderate hesitancy, urgency, retention of bowel or bladder, or rare urinary incontinence
3 = Frequent urinary incontinence
4 = In need of almost constant catheterization
5 = Loss of bladder function
6 = Loss of bowel and bladder function

F. Visual (or optic) functions
0 = Normal
1 = Scotoma with visual acuity (corrected) better than 20/30
2 = Worse eye with scotoma with maximal visual acuity (corrected) of 20/30 to 20/59
3 = Worse eye with large scotoma, or moderate decrease in fields, but with maximal visual acuity (corrected) of 20/60 to 20/99
4 = Worse eye with marked decrease of fields and maximal acuity (corrected) of 20/100 to 20/200; grade 3 plus maximal acuity of better eye of 20/60 or less
5 = Worse eye with maximal visual acuity (corrected) <20/200; grade 4 plus maximal acuity of better eye of ≤20/60
6 = Grade 5 plus maximal visual acuity of better eye of ≤20/60

G. Cerebral (or mental) functions
0 = Normal
1 = Mood alteration only (does not affect EDSS score)
2 = Mild decrease in mentation
3 = Moderate decrease in mentation

4 = Marked decrease in touch or pain or loss of proprioception, alone or combined, in 1 or 2 limbs or moderate decrease in touch or pain and/or severe proprioceptive decrease in >2 limbs	4 = Marked decrease in mentation 5 = Chronic brain syndrome—severe or incompetent

Source: Adapted from JF Kurtzke: Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). Neurology 33:1444, 1983.

TABLE 436-6

Outcomes for FDA-Approved Therapies for Multiple Sclerosis (MS)[a]

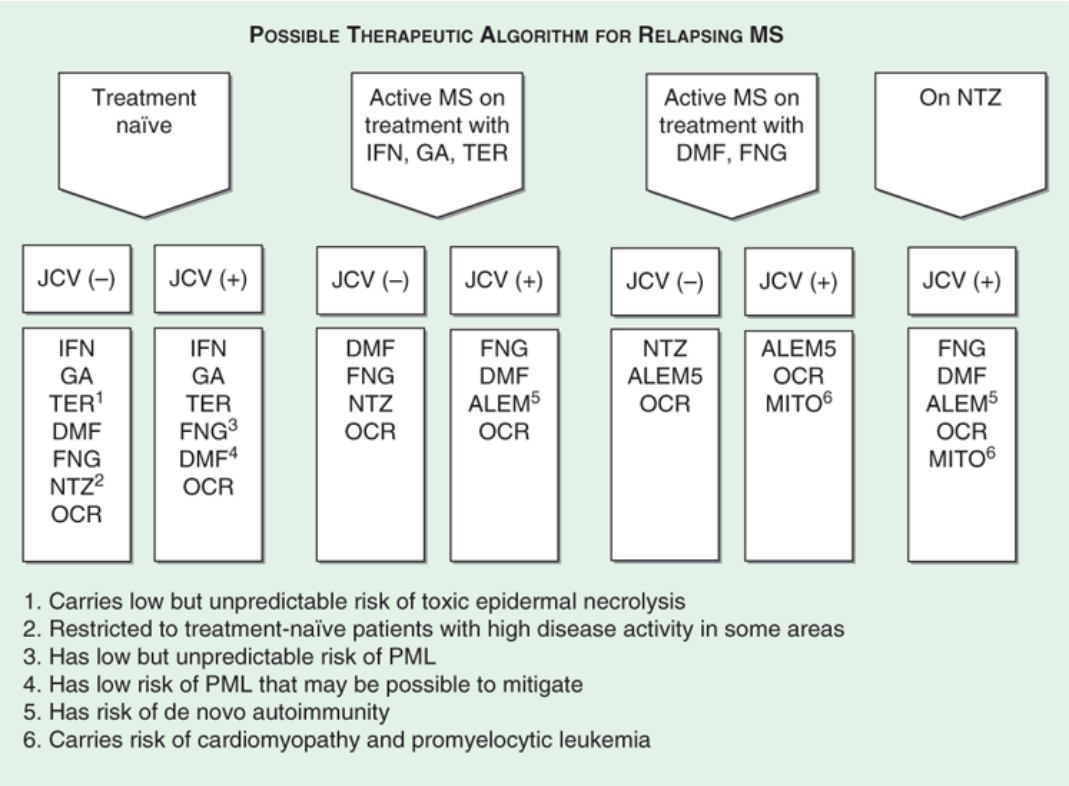
Relapsing MS						
			Clinical Outcomes[b]		**MRI Outcomes[c]**	
Dose, Route, and Schedule	**Study Duration Weeks**	**Comparator**	**Attack Rate, Mean**	**Change in Disease Severity**	**New T2 Lesions[d]**	**Total Burden of Disease**
IFN-β-1b, 250 µg SC qod	96	PBO	–34%[e]	–29% (NS)	–83%[f]	–17%[e]
IFN-β-1a, 30 µg IM qw	96	PBO	–18%[g]	–37%[g]	–36%[f]	NS
IFN-β-1a, 44 µg SC tiw	96	PBO	–32%[e]	–30%[g]	–78%[e]	–15%[e]
Peg- IFN-β-1a, 125 µg SC q2w	48	PBO	–36%[e]	–38%[g]	–67%[e]	–2%[e]
GA, 20 mg SC qd	96	PBO	–29%[f]	–12% (NS)	–38%[f]	–8%[f]
MTX, 12 mg/m^2 IV q3mo	96	PBO	–66%[e]	–75%[g]	–79%[g]	NR
NTZ, 300 mg IV qmo	96	PBO	–68%[e]	–42%[e]	–83%[e]	–18%[e]
FNG, 0.5 mg PO qd	96	PBO	–55%[e]	–34%[f]	–74%[e]	–23%[e]
DMF, 240 mg PO bid	96	PBO	–52%[e]	–40%[f]	–71%[e]	NR
TF, 14 mg PO qd	96	PBO	–31%[e]	–26%[g]	–70%[e]	–20%[g]
FNG, 0.5 mg PO qd	48	IFN-β-1a, 30 µg IM qw	–52%[e]	NS	–35%[e]	NS
ALEM, 12 mg/m^2 IV/5 d	104	IFN-β-1a, 44 µg SC tiw	–49%[e]	–42%[f]	–32%[e]	NS
OCR, 600 mg IV, Q6 mo	96	IFN-β-1a, 44 µg SC tiw	–46%[e,h]	–33%[e,h]	–80%[e,h]	NR
Primary Progressive MS						
OCR, 600 mg IV, Q6 mo	96	PBO	–NR	–24%	–92%	–11%

[a]Percentage reductions (or increases) have been calculated by dividing the reported rates in the treated group by the comparable rates in the placebo group, except for magnetic resonance imaging (MRI) disease burden, which was calculated as the difference in the median percent change between the treated and placebo groups. [b]Severity = 1 point Expanded Disability Status Score progression, sustained for 3 months (in the IFN-β-1a 30 µg qw trial, this change was sustained for 6 months; in the IFN-β-1b trial, this was over 3 years). [c]Different studies measured these MRI measures differently, making comparisons difficult (numbers for new T2 represent the best case scenario for each trial). [d]New lesions seen on T2-weighted MRI. [e]$p = .001$. [f]$p = .01$. [g]$p = .05$. [h]Pooled analysis from OPERA 1 and 2 studies.

Abbreviations: DMF, dimethyl fumarate; FDA, U.S. Food and Drug Administration; FNG, fingolimod; GA, glatiramer acetate; IFN-β, interferon β; IM, intramuscular; IV, intravenous; MTX, mitoxantrone; NR, not reported; NS, not significant; NTZ, natalizumab; PO, oral; q3mo, once every 3 months; qd, daily; qmo, once per month; qod, every other day; qw, once per week; qyr, once per year; SC, subcutaneous; TF, teriflunomide; tiw, three times per week.

FIGURE 436-4

Therapeutic decision-making for relapsing MS. Options are shown for different clinical scenarios and based on JCV status. Active MS is defined by clinical relapses or the development of new focal MRI white matter lesions. Treatment options can also include trials of different preparations of interferon β (IFN-β), particularly advancing from once-weekly (Avonex) to a more frequent (e.g., Rebif, Betaseron/Extavia) dosing regimen, as well as use of natalizumab in JC virus–positive patients.



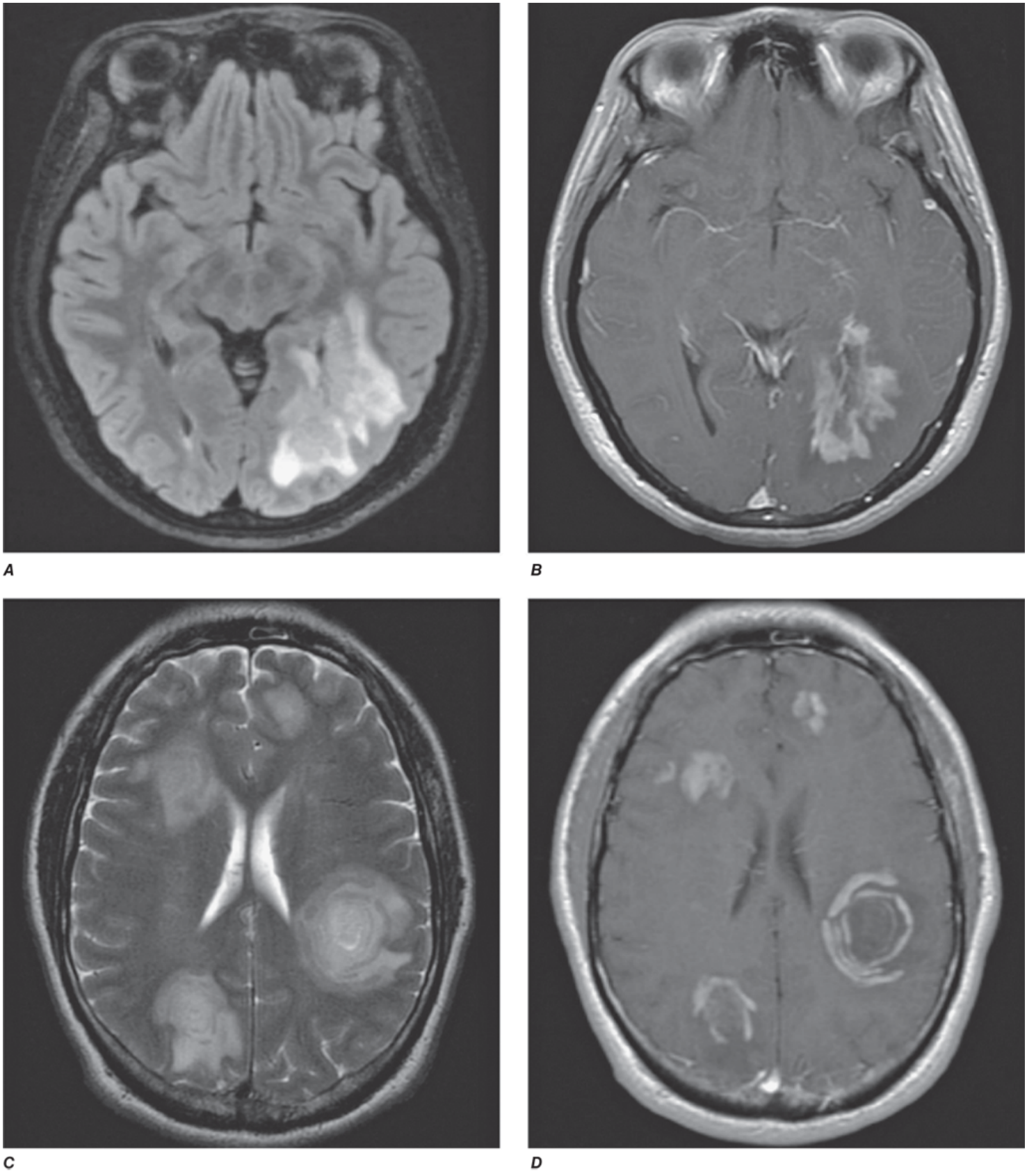
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

CLINICAL VARIANTS OF MS

Acute MS (*Marburg's variant*) is a fulminant demyelinating process that in some cases progresses inexorably to death within 1–2 years. Typically, there are no remissions. When acute MS presents as a solitary, usually cavitary, lesion, a brain tumor is often suspected (Fig. 436-5). In such cases, a brain biopsy is usually required to establish the diagnosis. Marburg's variant does not seem to follow infection or vaccination, and it is unclear whether this syndrome represents an extreme form of MS or another disease altogether.

FIGURE 436-5

Magnetic resonance imaging findings in variants of MS. ***A*** and ***B.*** Acute tumefactive MS. In ***A***, a sagittal T2-weighted fluid-attenuated inversion recovery (FLAIR) image of a large solitary right parieto-occipital white matter lesion is shown, with effacement of overlying cortical sulci consistent with mass effect. In ***B***, T1-weighted image obtained after the intravenous administration of gadolinium DTPA reveals a large serpiginous area of blood-brain barrier disruption consistent with acute inflammation. ***C*** and ***D***. Balo's concentric sclerosis. In ***C***, an axial T2-weighted sequence shows multiple areas of abnormal ovoid bright signal in the supratentorial white matter bilaterally; some lesions reveal concentric layers, typical of Balo's concentric sclerosis. In ***D***, T1-weighted MR images postgadolinium demonstrate abnormal enhancement of all lesions with some lesions demonstrating concentric ring enhancement.



A B C D

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Balo's concentric sclerosis is another fulminant demyelinating syndrome characterized by concentric brain or spinal cord lesions with alternating spheres of demyelination and remyelination (Fig. 436-5). For these fulminant demyelinating states, no controlled trials of therapy exist; high-dose glucocorticoids, plasma exchange, and cyclophosphamide have been tried, with uncertain benefit.

ACUTE DISSEMINATED ENCEPHALOMYELITIS (ADEM)

ADEM has a monophasic course and is most frequently associated with an antecedent infection (postinfectious encephalomyelitis); ~5% of ADEM cases follow immunization (postvaccinal encephalomyelitis). ADEM is far more common in children than adults, and many adult cases initially thought to represent ADEM subsequently experience late relapses qualifying as either MS or another chronic inflammatory disorder such as vasculitis, sarcoid, or lymphoma. The hallmark of ADEM is the presence of widely scattered foci of perivenular inflammation and demyelination that can involve both white matter and grey matter structures, in contrast to larger confluent white matter lesions typical of MS. In the most explosive form of ADEM, acute hemorrhagic leukoencephalitis, the lesions are vasculitic and hemorrhagic, and the clinical course is devastating.

Postinfectious encephalomyelitis is most frequently associated with the viral exanthems of childhood. Infection with measles virus is the most common antecedent (1 in 1000 cases). Worldwide, measles encephalomyelitis is still common, although use of the live measles vaccine has dramatically reduced its incidence in developed countries. An ADEM-like illness rarely follows vaccination with live measles vaccine (1–2 in 10^6 immunizations). ADEM is now most frequently associated with varicella (chickenpox) infections (1 in 4000–10,000 cases). It may also follow infection with rubella, mumps, influenza, parainfluenza, EBV, HHV-6, HIV, dengue, Zika, other viruses, and *Mycoplasma pneumoniae*. Some patients may have a nonspecific upper respiratory infection or no known antecedent illness. In addition to measles, postvaccinal encephalomyelitis may also follow the administration of vaccines for smallpox (5 cases per million), the Semple rabies, and Japanese encephalitis. Modern vaccines that do not require viral culture in CNS tissue have reduced the ADEM risk.

All forms of ADEM presumably result from a cross-reactive immune response to the infectious agent or vaccine that then triggers an inflammatory demyelinating response. Autoantibodies to MBP and to other myelin antigens have been detected in the CSF from some patients with ADEM. Attempts to demonstrate direct viral invasion of the CNS have been unsuccessful.

CLINICAL MANIFESTATIONS

In severe cases, onset is abrupt and progression rapid (hours to days). In postinfectious ADEM, the neurologic syndrome generally begins late in the course of the viral illness as the exanthem is fading. Fever reappears, and headache, meningismus, and lethargy progressing to coma may develop. Seizures are common. Signs of disseminated neurologic disease are consistently present (e.g., hemiparesis or quadriparesis, extensor plantar responses, lost or hyperactive tendon reflexes, sensory loss, and brainstem involvement). In ADEM due to chickenpox, cerebellar involvement is often conspicuous. CSF protein is modestly elevated (0.5–1.5 g/L [50–150 mg/dL]). Lymphocytic pleocytosis, generally ≥200 cells/μL, occurs in 80% of patients. Occasional patients have higher counts or a mixed polymorphonuclear-lymphocytic pattern during the initial days of the illness. Transient CSF oligoclonal banding has been reported. MRI usually reveals extensive changes in the brain and spinal cord, consisting of white matter hyperintensities on T2 and fluid-attenuated inversion recovery (FLAIR) sequences with Gd enhancement on T1-weighted sequences.

DIAGNOSIS

The diagnosis is most reliably established when there is a history of recent vaccination or viral exanthematous illness. In severe cases with predominantly cerebral involvement, acute encephalitis due to infection with herpes simplex or other viruses including HIV may be difficult to exclude **(Chap. 132)**; other considerations include hypercoagulable states including the antiphospholipid antibody syndrome, vasculitis, neurosarcoid, primary CNS lymphoma, or metastatic cancer. An explosive presentation of MS can mimic ADEM, and, especially in adults, it may not be possible to distinguish these conditions at onset. The simultaneous onset of disseminated symptoms and signs is common in ADEM and rare in MS. Similarly, meningismus, drowsiness, coma, and seizures suggest ADEM rather than MS. Unlike MS, in ADEM, optic nerve involvement is generally bilateral and transverse myelopathy complete. MRI findings that favor ADEM include extensive and relatively symmetric white matter abnormalities, basal ganglia or cortical gray matter lesions, and Gd enhancement of all abnormal areas. By contrast, OCBs in the CSF are more common in MS. In one study of adult patients initially thought to have ADEM, 30% experienced additional relapses over a follow-up period of 3 years, and they were reclassified as having MS. Other patients initially classified as ADEM are subsequently found to have neuromyelitis optica spectrum disorder **(Chap. 437)**. Occasional patients with “recurrent ADEM” have also been reported, especially children; however, it is not possible to distinguish this entity from atypical MS. Because of the clinical overlap at presentation between ADEM and MS, it is crucial that routine surveillance imaging be performed following recovery from ADEM so that subclinical disease activity due to MS can be recognized and treatment for MS initiated.

TREATMENT

TREATMENT

Acute Disseminated Encephalomyelitis

Initial treatment is with high-dose glucocorticoids; depending on the response, treatment may need to be continued for 8 weeks. Patients who fail to respond within a few days may benefit from a course of plasma exchange or intravenous immunoglobulin. The prognosis reflects the severity of the underlying acute illness. In recent case series of presumptive ADEM in adults, mortality rates of 5–20% are reported, and many survivors have permanent neurologic sequelae.

ACKNOWLEDGMENT

The authors want to thank Douglas E. Goodin for his contributions to previous editions of this chapter.

FURTHER READING

Arnold DL et al: Peginterferon beta-1a improves MRI measures and increases the proportion of patients with no evidence of disease activity in relapsing-remitting multiple sclerosis: 2-Year results from the advance randomized controlled trial. BMC Neurol 17:29, 2017. [PubMed: 28183276]

Correale J et al: Progressive multiple sclerosis: From pathogenic mechanisms to treatment. Brain 140:527, 2017. [PubMed: 27794524]

Giovannoni G et al: Alemtuzumab improves preexisting disability in active relapsing-remitting MS patients. Neurology 87:1985, 2016. [PubMed: 27733571]

Hauser SL et al: Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis. N Engl J Med 376:221, 2017. [PubMed: 28002679]

Mahad DH et al: Pathological mechanisms in progressive multiple sclerosis. Lancet Neurol 14:183, 2015. [PubMed: 25772897]

Montalban X et al: Ocrelizumab versus placebo in primary progressive multiple sclerosis. N Engl J Med 376:209, 2017. [PubMed: 28002688]

Pohl D et al: Acute disseminated encephalomyelitis: Updates on an inflammatory CNS syndrome. Neurology 87(9 Suppl 2):S38, 2016. [PubMed: 27572859]

Chapter 437: Neuromyelitis Optica

Bruce A. C. Cree; Stephen L. Hauser

FIGURE 437-1

INTRODUCTION

Neuromyelitis optica (NMO; Devic's disease) is an aggressive inflammatory disorder characterized by recurrent attacks of ON and myelitis; the more inclusive term *NMO Spectrum Disorder* (NMOSD) has been proposed to incorporate individuals with partial forms, and also those with involvement of additional structures in the central nervous system **(Table 437-1)**. NMO is more frequent in women than men (>3:1), and typically begins in adulthood but can arise at any age. An important consideration, especially early in its presentation, is distinguishing between NMO and multiple sclerosis (MS; **Chap. 436**). In patients with NMO, attacks of ON can be bilateral and produce severe visual loss (uncommon in MS); myelitis can be severe and transverse (rare in MS) and is typically longitudinally extensive **(Fig. 437-1)** involving three or more contiguous vertebral segments. Also in contrast to MS, progressive symptoms typically do not occur in NMO. The brain MRI was earlier thought to be normal in NMO, but it is now recognized that in many cases brain lesions are present, including areas of nonspecific signal change as well as lesions associated with specific syndromes such as the hypothalamus causing an endocrinopathy; the area postrema in the lower medulla presenting as intractable hiccoughs or vomiting; or the cerebral hemispheres producing focal symptoms, encephalopathy, or seizures. Large MRI lesions in the cerebral hemispheres can be asymptomatic, sometimes have a "cloud-like" appearance and, unlike MS lesions, are often not destructive, and can resolve completely. Spinal cord MRI lesions typically consist of focal enhancing areas of swelling and tissue destruction, extending over three or more spinal cord segments, and on axial sequences, these are centered on the gray matter of the cord. Cerebrospinal fluid (CSF) findings include pleocytosis greater than that observed in MS, with neutrophils and eosinophils present in many acute cases; OCBs are uncommon, occurring in <20% of NMO patients. The pathology of NMO is a distinctive astrocytopathy with inflammation, loss of astrocytes, and an absence of staining of the water channel protein AQP4 by immunohistochemistry, plus thickened blood vessel walls, demyelination, and deposition of antibody and complement.

TABLE 437-1

Diagnostic Criteria for Neuromyelitis Optica Spectrum Disorder

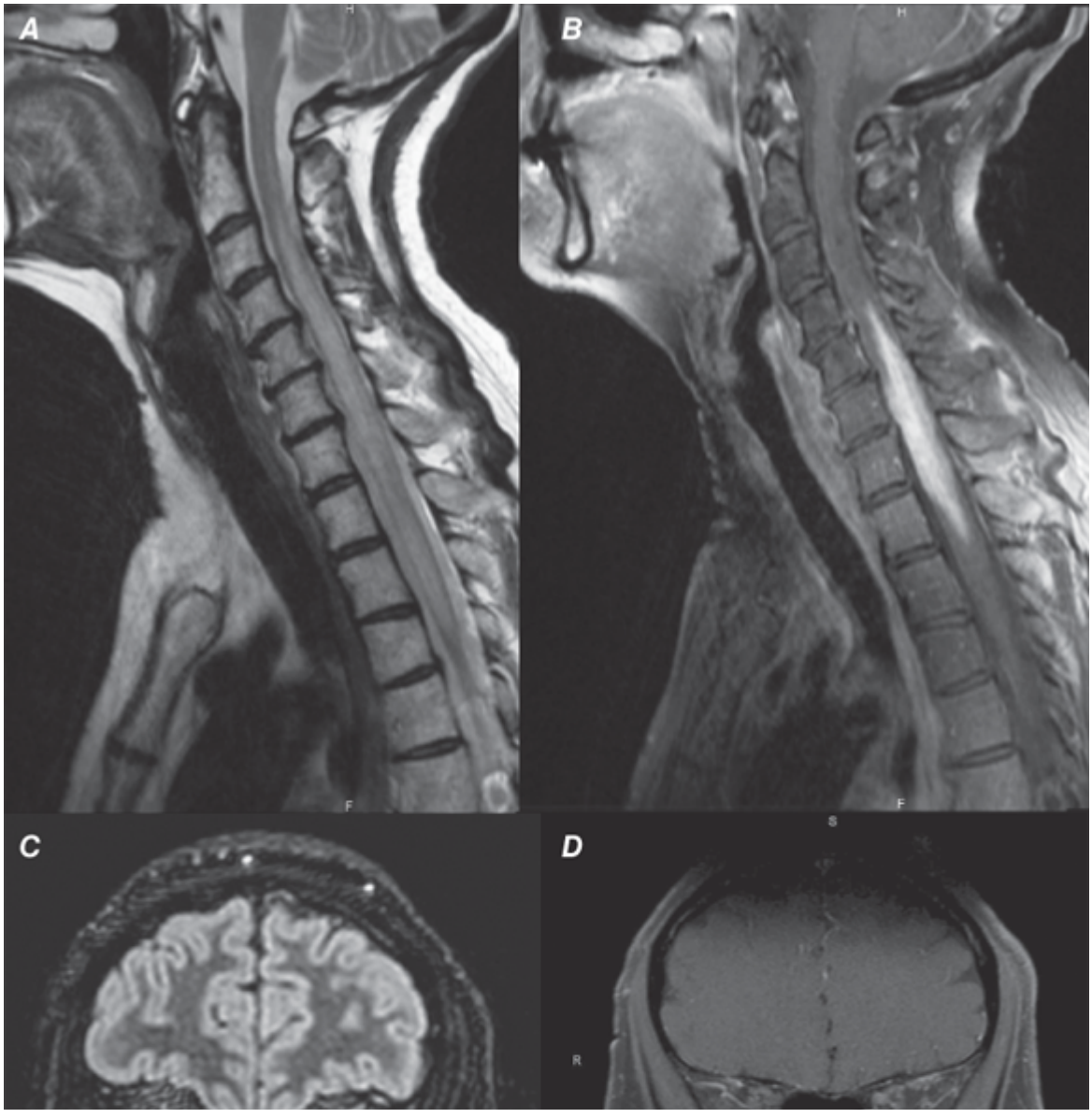
Diagnostic Criteria for NMOSA with AQP4-IgG
1. At least 1 core clinical characteristic 2. Positive test for AQP4-IgG using best available detection method (cell-based assay strongly recommended) 3. Exclusion of alternative diagnoses
Diagnostic Criteria for NMOSA Without AQP4-IgG or NMOSD with Unknown AQP4-IgG Status
1. At least 2 core clinical characteristics occurring as a result of one or more clinical attacks and meeting all of the following requirements: a. At least 1 core clinical characteristic must be optic neuritis, acute myelitis with LETM, or area postrema syndrome b. Dissemination in space (2 or more different clinical characteristics) c. Fulfillment of additional MRI requirements, as applicable 2. Negative test for AQP4-IgG using best available detection method or testing unavailable 3. Exclusion of alternative diagnoses
Core Clinical Characteristics
1. Optic neuritis 2. Acute myelitis 3. Area postrema syndrome: episode of otherwise unexplained hiccups or nausea or vomiting 4. Acute brainstem syndrome 5. Symptomatic narcolepsy or acute diencephalic clinical syndrome with NMOSD-typical diencephalic MRI lesions 6. Symptomatic cerebral syndrome with NMOSD-typical brain lesions
Additional MRI Requirements for NMOSD Without AQP4-IgG and NMOSD with Unknown AQP4-IgG Status
1. Acute optic neuritis: requires brain MRI showing (a) normal findings or only nonspecific white matter lesions, OR (b) optic nerve MRI with T2-hyperintense lesion of T1-weighted gadolinium-enhancing lesion extending over >1/2 optic nerve length or involving optic chiasm 2. Acute myelitis: requires associated intramedullary NMRI lesion extending ≥3 contiguous segments (LETM) OR ≥3 contiguous segments of focal spinal cord atrophy in patients with history compatible with acute myelitis

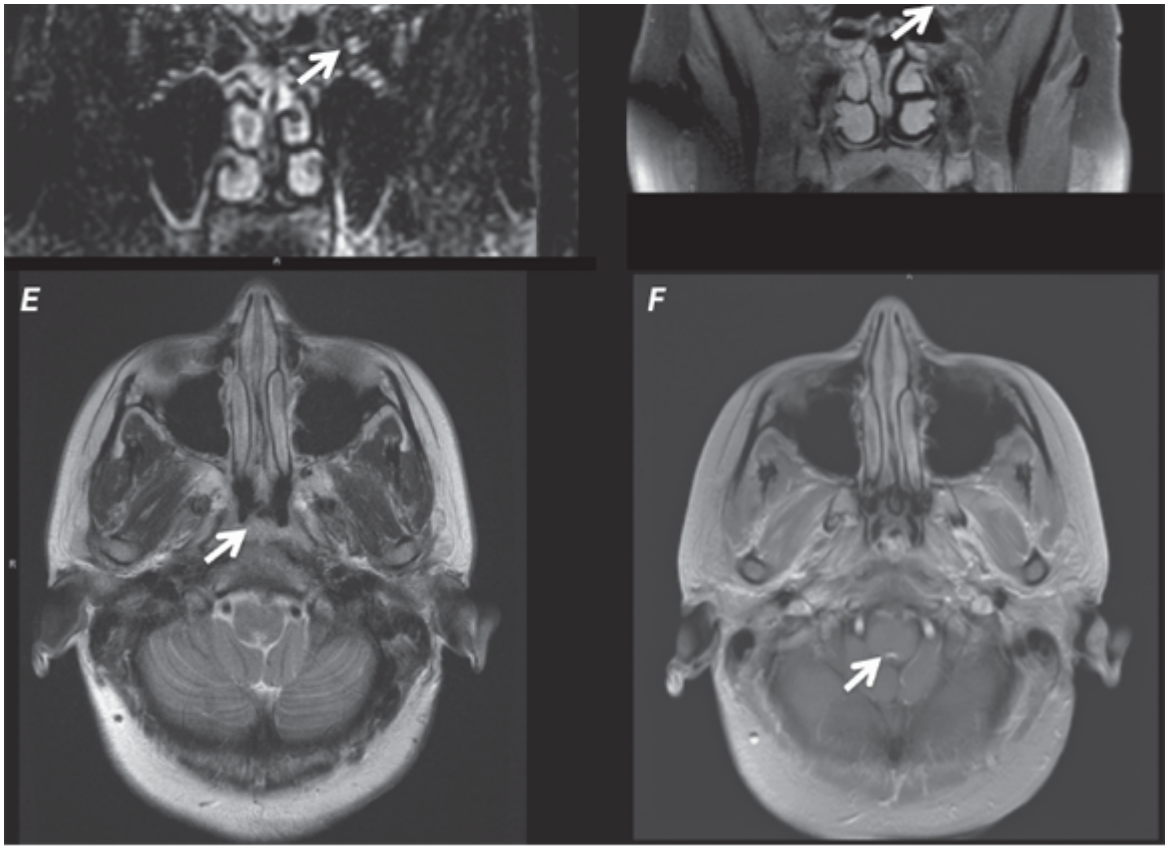
3. Area postrema syndrome requires associated dorsal medulla/area postrema lesions
4. Acute brainstem syndrome requires periependymal brainstem lesions

Source: Adapted from DM Wingerchuk et al: Neurology 85:177–189, 2015.

FIGURE 437-1

Imaging findings in neuromyelitis optica: longitudinally extensive transverse myelitis, optic neuritis, and brainstem involvement. ***A.*** Sagittal fluid attenuation inversion recovery (FLAIR) cervical-spine MRI showing an area of increased signal change on T2-weighted imaging spanning >3 vertebral segments in length. ***B.*** Sagittal T1-weighted cervical-spine MRI following gadolinium-diethylene triamine pentaacetic acid (DPTA) infusion showing enhancement. ***C.*** Coronal brain MRI shows hyperintense signal on FLAIR imaging within the left optic nerve. ***D.*** Coronal T1-weighted brain MRI following gadolinium-DPTA infusion shows enhancement of the left optic nerve. ***E.*** Axial brain MRI shows an area of hyperintense signal on T2-weighted imaging within the area postrema (arrow). ***F.*** Axial T1-weighted brain MRI following gadolinium-DPTA infusion shows punctate enhancement of the area postrema (arrow).





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IMMUNOLOGY

NMO is an autoimmune disease associated with a highly specific autoantibody directed against aquaporin-4 (AQP4) that is present in the sera of ~70% of patients with a clinical diagnosis of NMO. AQP4 is localized to the foot processes of astrocytes in close apposition to endothelial surfaces, as well as at paranodal regions near nodes of Ranvier. It is likely that AQP4 antibodies are pathogenic because passive transfer of AQP4 antibodies into laboratory animals can reproduce histologic features of the disease; complement fixation is thought to mediate astrocyte injury. During acute attacks of myelitis, CSF levels of interleukin-6 (IL-6; a proinflammatory cytokine) and glial fibrillary acidic protein (GFAP) levels are markedly elevated, consistent with active inflammation and astrocyte injury. Proinflammatory T-lymphocytes of the Th17 type recognize an immunodominant epitope of AQP4 and may also contribute to pathogenesis. Because of the high specificity of the antibody, its presence is considered to be diagnostic when found in conjunction with a typical clinical presentation. Anti-AQP4 seropositive patients have a high risk for future relapses; more than half will relapse within 1 year if untreated.

CLINICAL COURSE

NMO is typically a recurrent disease; the course is monophasic in fewer than 10% of patients. Individuals who test negative for AQP-4 antibodies are somewhat more likely to have a monophasic course. Untreated NMO is

usually quite disabling over time; in one series, respiratory failure from cervical myelitis was present in one-third of patients, and 8 years after onset, 60% of patients were blind and more than half had permanent paralysis of one or more limbs. There is limited data indicating that the long-term course of NMO has been substantially improved with the development of therapies to treat acute attacks and prevent relapses.

GLOBAL CONSIDERATIONS

The incidence and prevalence of NMO shows considerable variation between populations and geographic regions, with prevalence estimates that range from <1 to > 4 per 100,000. Although NMO can occur in people of any ethnic background, individuals of Asian and African origin are disproportionately affected. The highest reported prevalence is from Martinique. Among white populations, MS **(Chap. 436)** is far more common than NMO.

Interestingly, when MS affects individuals of African or Asian ancestry, there is a propensity for demyelinating lesions to involve predominantly the optic nerve and spinal cord, an MS subtype termed *opticospinal MS*. Some individuals with opticospinal MS are seropositive for AQP-4 antibodies, indicating that such cases represent NMOSD.

ASSOCIATED CONDITIONS

Up to 40% of NMO patients have a systemic autoimmune disorder, such as systemic lupus erythematosus, Sjögren's syndrome, perinuclear antineutrophil cytoplasmic antibody (p-ANCA)–associated vasculitis, myasthenia gravis, Hashimoto's thyroiditis, or mixed connective tissue disease. In others, onset may be associated with acute infection with varicella zoster virus, Epstein-Barr virus, HIV, or tuberculosis. Rare cases appear to be paraneoplastic and associated with breast, lung, or other cancers.

TREATMENT

TREATMENT

Neuromyelitis Optica

Disease-modifying therapies have not been rigorously studied in NMO. Acute attacks are usually treated with high-dose glucocorticoids (e.g., methylprednisolone 1 g/d for 5–10 days followed by a prednisone taper). Plasma exchange (typically 5–7 exchanges of 1.5 plasma volumes/exchange) is used empirically for acute episodes that do not respond to glucocorticoids. Given the unfavorable natural history of untreated NMO, prophylaxis against relapses is recommended for most patients using one of the following regimens: mycophenolate mofetil (1000 mg bid); rituximab a B cell depleting anti-CD20 monoclonal antibody (2 g IV Q 6 months); or a combination of glucocorticoids (500 mg IV methylprednisolone daily for 5 days; then oral prednisone 1 mg/kg per day for 2 months, followed by slow taper) plus azathioprine (2 mg/kg per day started on week 3). Some therapies with proven efficacy in MS do not appear to be useful for NMO. Available evidence suggests that interferon beta is ineffective and paradoxically may increase the risk of NMO relapses,

and based on limited data glatiramer acetate, fingolimod, natalizumab, and alemtuzumab also appear to be ineffective. That therapies not commonly used in MS are empirically used in NMOSD highlight the need for efficient diagnosis of this disorder. Clinical trials with the B-cell depleting anti-CD19 monoclonal antibody (inebilizumab), the terminal complement inhibitor (eculizumab), and an IL-6 receptor blocking antibody (SA-237) are ongoing.

DEMYELINATION ASSOCIATED WITH ANTI-MOG ANTIBODIES

Although long considered to be a likely target for antibody-mediated demyelination, anti-MOG antibodies detected by a cell-based assay that enables recognition of myelin oligodendrocyte glycoprotein (MOG) epitopes in a lipid bilayer were only recently found to be associated with cases of acute disseminated encephalomyelitis (ADEM) **(Chap. 436)** in children, and then with cases of AQP4 seronegative NMO. Further studies showed that patients who are seropositive for anti-MOG antibodies are at risk for bilateral, synchronous optic neuritis and myelitis. A clinical feature that can help distinguish ON associated with anti-MOG antibodies from NMO or MS is the presence of papillitis seen by funduscopy or orbital MRI. ON associated with anti-MOG antibodies is typically longitudinally extensive on MRI, and brain MRI can be normal or show fluffy areas of increased signal change in white or grey matter structures, similar to NMO. MRI lesions that are typical for MS, including finger-like lesions oriented perpendicular to the ventricular surface (Dawson fingers) and T1 hypointense lesions **(Chap. 436)**, are uncommon. Spinal cord lesions can be longitudinally extensive or short. Demyelination associated with anti-MOG antibodies is sometimes monophasic, as in ADEM, but can also be recurrent. Acute episodes are managed with high dose glucocorticoids followed by a prednisone taper and sometimes by plasmapheresis, as with NMO. Brain lesions associated with anti-MOG antibodies often respond rapidly to treatment with glucocorticoids and may resolve entirely. Some patients experience disease recurrence following discontinuation of prednisone and can become glucocorticoid dependent. Clinical trials have not been undertaken and there is limited data on use of other immune-suppressing medications typically used in NMO.

FURTHER READING

Cree BA: Placebo controlled trials in neuromyelitis optica are needed and ethical. Mult Scler Relat Disord 4:536, 2015. [PubMed: 26590660]

Hinson SR et al: Autoimmune AQP4 channelopathies and neuromyelitis optica spectrum disorders. Handb Clin Neurol 133:377, 2016. [PubMed: 27112688]

Peschl P et al: Myelin oligodendrocyte glycoprotein: Deciphering a target in inflammatory demyelinating diseases. Front Immunol 8:529, 2017. [PubMed: 28533781]

Wingerchuk DM et al: International consensus diagnostic criteria for neuromyelitis optica spectrum disorders. Neurology 85:177, 2015. [PubMed: 26092914]

Chapter 438: Peripheral Neuropathy

Anthony A. Amato; Richard J. Barohn

INTRODUCTION

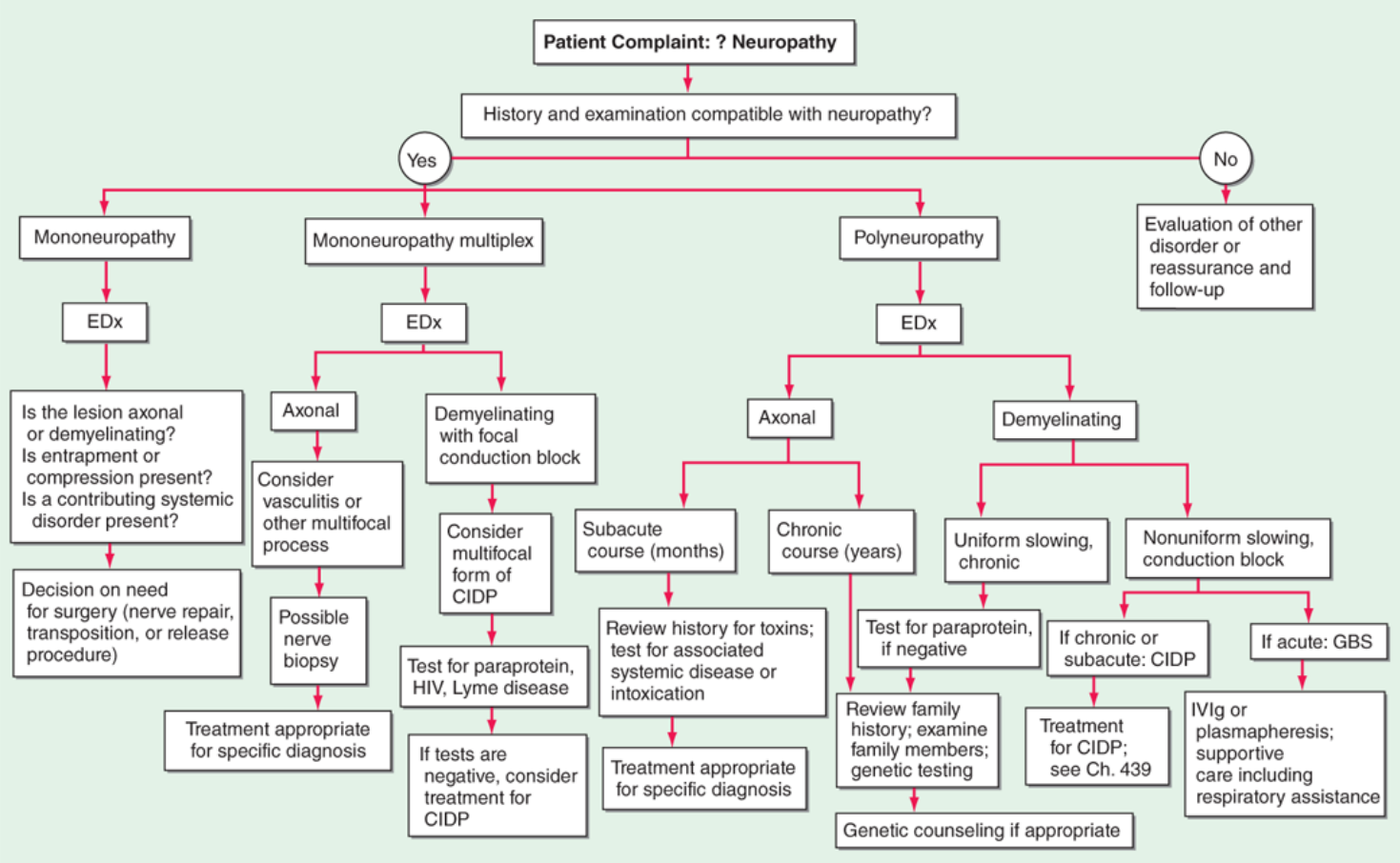
Peripheral nerves are composed of sensory, motor, and autonomic elements. Diseases can affect the cell body of a neuron or its peripheral processes, namely the axons or the encasing myelin sheaths. Most peripheral nerves are mixed and contain sensory and motor as well as autonomic fibers. Nerves can be subdivided into three major classes: large myelinated, small myelinated, and small unmyelinated. Motor axons are usually large myelinated fibers that conduct rapidly (~50 m/s). Sensory fibers may be any of the three types. Large-diameter sensory fibers conduct proprioception and vibratory sensation to the brain, while the smaller-diameter myelinated and unmyelinated fibers transmit pain and temperature sensation. Autonomic nerves are also small in diameter. Thus, peripheral neuropathies can impair sensory, motor, or autonomic function, either singly or in combination. Peripheral neuropathies are further classified into those that primarily affect the cell body (e.g., neuronopathy or ganglionopathy), myelin (myelinopathy), and the axon (axonopathy). These different classes of peripheral neuropathies have distinct clinical and electrophysiologic features. This chapter discusses the clinical approach to a patient suspected of having a peripheral neuropathy, as well as specific neuropathies, including hereditary and acquired neuropathies. **The inflammatory neuropathies are discussed in Chap. 439.**

GENERAL APPROACH

In approaching a patient with a neuropathy, the clinician has three main goals: (1) identify where the lesion is, (2) identify the cause, and (3) determine the proper treatment. The first goal is accomplished by obtaining a thorough history, neurologic examination, and electrodiagnostic and other laboratory studies **(Fig. 438-1)**. While gathering this information, seven key questions are asked **(Table 438-1)**, the answers to which help identify the pattern of involvement and the cause of the neuropathy **(Table 438-2)**. Despite an extensive evaluation, in approximately half of patients no etiology is ever found; these patients typically have a predominately sensory polyneuropathy and have been labeled as having idiopathic or cryptogenic sensory polyneuropathy (CSPN).

FIGURE 438-1

Approach to the evaluation of peripheral neuropathies. CIDP, chronic inflammatory demyelinating polyradiculoneuropathy; EDx, electrodiagnostic; GBS, Guillain-Barré syndrome; IVIg, intravenous immunoglobulin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 438-1

Approach to Neuropathic Disorders: Seven Key Questions

1. **What systems are involved?**
   - Motor, sensory, autonomic, or combinations
2. **What is the distribution of weakness?**
   - Only distal versus proximal and distal
   - Focal/asymmetric versus symmetric
3. **What is the nature of the sensory involvement?**
   - Temperature loss or burning or stabbing pain (e.g., small fiber)
   - Vibratory or proprioceptive loss (e.g., large fiber)
4. **Is there evidence of upper motor neuron involvement?**
   - Without sensory loss
   - With sensory loss
5. **What is the temporal evolution?**
   - Acute (days to 4 weeks)
   - Subacute (4–8 weeks)
   - Chronic (>8 weeks)
   - Monophasic, progressive, or relapsing-remitting
6. **Is there evidence for a hereditary neuropathy?**
   - Family history of neuropathy
   - Lack of sensory symptoms despite sensory signs
7. **Are there any associated medical conditions?**
   - Cancer, diabetes mellitus, connective tissue disease or other autoimmune diseases, infection (e.g., HIV, Lyme disease, leprosy)
   - Medications including over-the-counter drugs that may cause a toxic neuropathy
   - Preceding events, drugs, toxins

TABLE 438-2

Patterns of Neuropathic Disorders

Pattern 1: Symmetric proximal and distal weakness with sensory loss
Consider: inflammatory demyelinating polyneuropathy (GBS and CIDP)
Pattern 2: Symmetric distal sensory loss with or without distal weakness
Consider: cryptogenic or idiopathic sensory polyneuropathy (CSPN), diabetes mellitus and other metabolic disorders, drugs, toxins, familial (HSAN), CMT, amyloidosis, and others
Pattern 3: Asymmetric distal weakness with sensory loss
With involvement of multiple nerves
Consider: multifocal CIDP, vasculitis, cryoglobulinemia, amyloidosis, sarcoid, infectious (leprosy, Lyme, hepatitis B, C, or E, HIV, CMV), HNPP, tumor infiltration
With involvement of single nerves/regions
Consider: may be any of the above but also could be compressive mononeuropathy, plexopathy, or radiculopathy
Pattern 4: Asymmetric proximal and distal weakness with sensory loss
Consider: polyradiculopathy or plexopathy due to diabetes mellitus, meningeal carcinomatosis or lymphomatosis, sarcoid, amyloid, hereditary plexopathy (HNPP, HNA), idiopathic
Pattern 5: Asymmetric distal weakness without sensory loss
With upper motor neuron findings
Consider: motor neuron disease
Without upper motor neuron findings
Consider: progressive muscular atrophy, juvenile monomelic amyotrophy (Hirayama's disease), multifocal motor neuropathy, multifocal acquired motor axonopathy
Pattern 6: Symmetric sensory loss and distal areflexia with upper motor neuron findings
Consider: Vitamin B_{12}, vitamin E, and copper deficiency with combined system degeneration with peripheral neuropathy, chronic liver disease, hereditary leukodystrophies (e.g., adrenomyeloneuropathy) HSP-plus
Pattern 7: Symmetric weakness without sensory loss
With proximal and distal weakness
Consider: SMA
With distal weakness
Consider: hereditary motor neuropathy ("distal" SMA) or atypical CMT
Pattern 8: Focal midline proximal symmetric weakness
Neck extensor weakness
Consider: ALS
Bulbar weakness

Consider: ALS/PLS, isolated bulbar ALS (IBALS), Kennedy's syndrome (X-linked, bulbospinal SMA), bulbar presentation GBS
Diaphragm weakness (SOB)
Consider: ALS
Pattern 9: Asymmetric proprioceptive sensory loss without weakness
Consider causes of a sensory neuronopathy (ganglionopathy):
Cancer (paraneoplastic)
Sjögren's syndrome
Idiopathic sensory neuronopathy (possible GBS variant)
Cisplatin and other chemotherapeutic agents
Vitamin B_6 toxicity
HIV-related sensory neuronopathy
Pattern 10: Autonomic symptoms and signs
Consider neuropathies associated with prominent autonomic dysfunction:
Hereditary sensory and autonomic neuropathy
Amyloidosis (familial and acquired)
Diabetes mellitus
Idiopathic pandysautonomia (may be a variant of Guillain-Barré syndrome)
Porphyria
HIV-related autonomic neuropathy
Vincristine and other chemotherapeutic agents

Abbreviations: CIDP, chronic inflammatory demyelinating polyneuropathy; CMT, Charcot-Marie-Tooth disease; CMV, cytomegalovirus; GBS, Guillain-Barré syndrome; HIV, human immunodeficiency virus; HNA, hereditary neuralgic amyotrophy; HNPP, hereditary neuropathy with liability to pressure palsies; HSAN, hereditary sensory and autonomic neuropathy; HSP-plus, hereditary spastic paraplegia plus neuropathy; SMA, spinal muscular atrophy; SOB, shortness of breath.

INFORMATION FROM THE HISTORY AND PHYSICAL EXAMINATION: SEVEN KEY QUESTIONS

(Table 438-1)

1. What Systems Are Involved?

It is important to determine if the patient's symptoms and signs are motor, sensory, autonomic, or a combination of these. If the patient has only weakness without any evidence of sensory or autonomic dysfunction, a motor neuropathy, neuromuscular junction abnormality, or myopathy should be considered. Some peripheral neuropathies are associated with significant autonomic nervous system dysfunction. Symptoms of autonomic involvement include fainting spells or orthostatic lightheadedness; heat intolerance; or any bowel, bladder, or sexual dysfunction **(Chap. 432)**. There will typically be an orthostatic fall in blood pressure without an appropriate increase in heart rate. Autonomic dysfunction in the absence of diabetes should alert the clinician to the possibility of amyloid polyneuropathy. Rarely, a pandysautonomic syndrome can be the only manifestation of a peripheral neuropathy without other motor or sensory findings. The majority of neuropathies are predominantly sensory in nature.

2. What Is the Distribution of Weakness?

Delineating the pattern of weakness, if present, is essential for diagnosis, and in this regard two additional questions should be answered: (1) Does the weakness only involve the distal extremity, or is it both proximal and distal? and (2) Is the weakness focal and asymmetric, or is it symmetric? Symmetric proximal and distal weakness is the hallmark of acquired immune demyelinating polyneuropathies, both the acute form (Guillain-Barré syndrome [GBS]) and the chronic form (chronic inflammatory demyelinating polyneuropathy [CIDP]) **(Chap. 439)**. The importance of finding symmetric proximal and distal weakness in a patient who presents with both motor and sensory symptoms cannot be overemphasized because this identifies the important subset of patients who may have a treatable acquired demyelinating neuropathic disorder (i.e., GBS or CIDP).

Findings of an asymmetric or multifocal pattern of weakness narrow the differential diagnosis. Some neuropathic disorders may present with unilateral extremity weakness. In the absence of sensory symptoms and signs, such weakness evolving over weeks or months would be worrisome for motor neuron disease (e.g., amyotrophic lateral sclerosis [ALS]), but it would be important to exclude multifocal motor neuropathy that may be treatable **(Chap. 439)**. In a patient presenting with asymmetric subacute or acute sensory and motor symptoms and signs, radiculopathies, plexopathies, compressive mononeuropathies, or multiple mononeuropathies (e.g., mononeuropathy multiplex) must be considered.

ALS can produce prominent neck extensor weakness (head drop), tongue and pharyngeal weakness (dysarthia and dysphagia), or shortness of breath. These focal symmetric weakness patterns can also be seen in neuromuscular junction disorders (myasthenia gravis, Lambert-Eaton myasthenic syndrome [LEMS] **[Chap. 440]**) and some myopathies, particularly isolated neck extensor myopathy **(Chap. 441)**.

3. What Is the Nature of the Sensory Involvement?

The patient may have loss of sensation (numbness), altered sensation to touch (hyperpathia or allodynia), or uncomfortable spontaneous sensations (tingling, burning, or aching) **(Chap. 22)**. Neuropathic pain can be burning, dull, and poorly localized (protopathic pain), presumably transmitted by polymodal C nociceptor fibers, or sharp and lancinating (epicritic pain), relayed by A-delta fibers. If pain and temperature perception are lost, while vibratory and position sense are preserved along with muscle strength, deep tendon reflexes, and normal nerve conduction studies (NCS), a small-fiber neuropathy is likely. This is important, because the most likely cause of small-fiber neuropathies, when one is identified, is diabetes mellitus (DM) or glucose intolerance. Amyloid neuropathy should be considered as well in such cases, but most of these small-fiber neuropathies remain idiopathic despite extensive evaluation.

Severe proprioceptive loss also narrows the differential diagnosis. Affected patients will note imbalance, especially in the dark. A neurologic examination revealing a dramatic loss of proprioception with vibration loss and normal strength should alert the clinician to consider a sensory neuronopathy/ganglionopathy (Table 438-2, Pattern 9). In particular, if this loss is asymmetric or affects the arms more than the legs, this pattern suggests a non-length-dependent process as seen in sensory neuronopathies.

4. Is There Evidence of Upper Motor Neuron Involvement?

If the patient presents with symmetric distal sensory symptoms and signs suggestive of a distal sensory neuropathy, but there is additional evidence of symmetric upper motor neuron involvement **(Chap. 21)**, the physician should consider a combined system degeneration with neuropathy. The most common cause for this pattern is vitamin B_{12} deficiency, but other etiologies should also be considered (e.g., copper deficiency, human immunodeficiency virus [HIV] infection, severe hepatic disease, adrenomyeloneuropathy [AMN]), and hereditary spastic paraplegia plus a neuropathy.

5. What Is the Temporal Evolution?

It is important to determine the onset, duration, and evolution of symptoms and signs. Does the disease have an acute (days to 4 weeks), subacute (4–8 weeks), or chronic (>8 weeks) course? Is the course monophasic, progressive, or relapsing? Most neuropathies are insidious and slowly progressive in nature. Neuropathies with acute and subacute presentations include GBS, vasculitis, and radiculopathies related to diabetes or Lyme disease. A relapsing course can be present in CIDP and porphyria.

6. Is There Evidence for a Hereditary Neuropathy?

In patients with slowly progressive distal weakness over many years with few sensory symptoms yet significant sensory deficits on clinical examination, the clinician should consider a hereditary neuropathy (e.g., Charcot-Marie-Tooth disease [CMT]). On examination, the feet may show high or flat arches or hammer toes, and scoliosis may be present. In suspected cases, it may be necessary to perform neurologic and electrophysiologic studies on family members in addition to the patient.

7. Does the Patient Have Any Other Medical Conditions?

It is important to inquire about associated medical conditions (e.g., DM, systemic lupus erythematosus [SLE]); preceding or concurrent infections (e.g. diarrheal illness preceding GBS); surgeries (e.g., gastric bypass and nutritional neuropathies); medications (toxic neuropathy), including over-the-counter vitamin preparations (B_6); alcohol; dietary habits; and use of dentures (e.g., fixatives contain zinc that can lead to copper deficiency).

PATTERN RECOGNITION APPROACH TO NEUROPATHIC DISORDERS

Based on the answers to the seven key questions, neuropathic disorders can be classified into several patterns based on the distribution or pattern of sensory, motor, and autonomic involvement (Table 438-2). Each pattern has a limited differential diagnosis, and information from laboratory studies usually permits a final diagnosis to be established.

ELECTRODIAGNOSTIC STUDIES

The electrodiagnostic (EDx) evaluation of patients with a suspected peripheral neuropathy consists of NCS and needle electromyography (EMG). In addition, studies of autonomic function can be valuable. The electrophysiologic data can confirm whether the neuropathic disorder is a mononeuropathy, multiple mononeuropathy (mononeuropathy multiplex), radiculopathy, plexopathy, or generalized polyneuropathy. Similarly, EDx evaluation can ascertain whether the process involves only sensory fibers, motor fibers, autonomic fibers, or a combination of these. Finally, the electrophysiologic data can help distinguish axonopathies from myelinopathies as well as axonal degeneration secondary to ganglionopathies from the more common length-dependent axonopathies.

NCS are most helpful in classifying a neuropathy as due to axonal degeneration or segmental demyelination **(Table 438-3)**. In general, low-amplitude potentials with relatively preserved distal latencies, conduction velocities, and late potentials, along with fibrillations on needle EMG, suggest an axonal neuropathy. On the other hand, slow conduction velocities, prolonged distal latencies and late potentials, relatively preserved amplitudes, and the absence of fibrillations on needle EMG imply a primary demyelinating neuropathy. The presence of nonuniform slowing of conduction velocity, conduction block, or temporal dispersion further suggests an acquired demyelinating neuropathy (e.g., GBS or CIDP) as opposed to a hereditary demyelinating neuropathy (e.g., CMT type 1).

TABLE 438-3

Electrophysiologic Features: Axonal Degeneration versus Segmental Demyelination

	Axonal Degeneration	Segmental Demyelination
Motor Nerve Conduction Studies		
CMAP amplitude	Decreased	Normal (except with CB or distal dispersion)
Distal latency	Normal	Prolonged
Conduction velocity	Normal	Slow
Conduction block	Absent	Present
Temporal dispersion	Absent	Present
F wave	Normal or absent	Prolonged or absent
H reflex	Normal or absent	Prolonged or absent
Sensory Nerve Conduction Studies		
SNAP amplitude	Decreased	Normal or decreased
Distal latency	Normal	Prolonged
Conduction velocity	Normal	Slow
Needle EMG		
Spontaneous activity		
Fibrillations	Present	Absent
Fasciculations	Present	Absent
Motor unit potentials		
Recruitment	Decreased	Decreased
Morphology	Long duration, large amplitude, polyphasic (if there is reinnervation)	Normal

Abbreviations: CB, conduction block; CMAP, compound motor action potential; EMG, electromyography; SNAP, sensory nerve action potential.

Autonomic studies are used to assess small myelinated (A-delta) or unmyelinated (C) nerve fiber involvement. Such testing includes heart rate response to deep breathing, heart rate, and blood pressure response to both the Valsalva maneuver and tilt-table testing and quantitative sudomotor axon reflex testing (Chap. 432). These studies are particularly useful in patients who have pure small-fiber neuropathy or autonomic neuropathy in which routine NCS are normal.

OTHER IMPORTANT LABORATORY INFORMATION

In patients with generalized symmetric peripheral neuropathy, a standard laboratory evaluation should include a complete blood count, basic chemistries including serum electrolytes and tests of renal and hepatic function, fasting blood glucose (FBS), hemaglobin (Hb) A_{1c}, urinalysis, thyroid function tests, B_{12}, folate, erythrocyte sedimentation rate (ESR), rheumatoid factor, antinuclear antibodies (ANA), serum protein electrophoresis (SPEP) and immunoelectrophoresis or immunofixation, and urine for Bence Jones protein. Quantification of the concentration of serum-free light chains and the kappa/lambda ratio is more sensitive than SPEP, immunoelectrophoresis, or immunofixation to detect a monoclonal gammopathy and therefore should be done if amyloidosis is suspected. A skeletal survey should be performed in patients with acquired demyelinating neuropathies and M-spikes to look for osteosclerotic or lytic lesions. Patients with monoclonal gammopathy should also be referred to a hematologist for consideration of a bone marrow biopsy.

An oral glucose tolerance test is indicated in patients with painful sensory neuropathies even if FBS and HbA_{1c} are normal, as the test is abnormal in about one-third of such patients. In addition to the above tests, patients with a mononeuropathy multiplex pattern of involvement should have a vasculitis workup, including antineutrophil cytoplasmic antibodies (ANCA), cryoglobulins, hepatitis serology, Western blot for Lyme disease, HIV, and occasionally a cytomegalovirus (CMV) titer.

There are many autoantibody panels (various antiganglioside antibodies) marketed for screening routine neuropathy patients for a treatable condition. These autoantibodies have no proven clinical utility or added benefit beyond the information obtained from a complete clinical examination and detailed EDx. A heavy metal screen is also not necessary as a screening procedure, unless there is a history of possible exposure or suggestive features on examination (e.g., severe painful sensorimotor and autonomic neuropathy and alopecia—thallium; severe painful sensorimotor neuropathy with or without gastrointestinal [GI] disturbance and Mee's lines—arsenic; wrist or finger extensor weakness and anemia with basophilic stippling of red blood cells—lead).

In patients with suspected GBS or CIDP, a lumbar puncture is indicated to look for an elevated cerebrospinal fluid (CSF) protein. In idiopathic cases of GBS and CIDP, CSF pleocytosis is usually absent. If cells are present, one should consider HIV infection, Lyme disease, sarcoidosis, or lymphomatous or leukemic infiltration of nerve roots. Some patients with GBS and CIDP have abnormal liver function tests. In these cases, it is important to also check for hepatitis B and C, HIV, CMV, and Epstein-Barr virus (EBV) infection. In patients with an axonal GBS (by EMG/NCS) or those with a suspicious coinciding history (e.g., unexplained abdominal pain, psychiatric illness, significant autonomic dysfunction), it is reasonable to screen for porphyria.

In patients with a severe sensory ataxia, a sensory ganglionopathy or neuronopathy should be considered. The most common causes of sensory ganglionopathies are Sjögren's syndrome **(Chap. 354)** and a paraneoplastic neuropathy **(Chap. 90)**. Neuropathy can be the initial manifestation of Sjögren's syndrome. Thus, one should always inquire about dry eyes and mouth in patients with sensory signs and symptoms. Further, some patients can manifest sicca complex without other manifestations of Sjögren's syndrome. Thus, patients with sensory ataxia should be tested for antibodies to SS-A/Ro and SS-B/La, in addition to the routine ANA. To evaluate a possible paraneoplastic sensory ganglionopathy, antineuronal nuclear antibodies (e.g., anti-Hu antibodies) should be obtained. These antibodies are most commonly seen in patients with small-cell carcinoma of the lung but are also present with breast, ovarian, lymphoma, and other cancers. Importantly, the paraneoplastic neuropathy can precede the detection of the cancer, and detection of these autoantibodies should lead to a search for malignancy.

NERVE BIOPSIES

Nerve biopsies are now rarely performed in the evaluation of neuropathies. The primary indication for nerve biopsy is suspicion for amyloid neuropathy or vasculitis. In most instances, the abnormalities present on biopsies do not help distinguish one form of peripheral neuropathy from another (beyond what is already apparent by clinical examination and the NCS). Nerve biopsies should only be performed when the NCS are abnormal. The sural nerve is most commonly biopsied because it is a pure sensory nerve and biopsy will not result in loss of motor function. In suspected vasculitis, a combination biopsy of a superficial peroneal nerve (pure sensory) and the underlying peroneus brevis muscle obtained from a single small incision increases the diagnostic yield. Tissue can be analyzed to assess for evidence of inflammation, vasculitis, or amyloid deposition. Semithin plastic sections, teased fiber preparations, and electron microscopy are used to assess the morphology of the nerve fibers and to distinguish axonopathies from myelinopathies.

SKIN BIOPSIES

Skin biopsies are sometimes used to diagnose a small-fiber neuropathy. Following a punch biopsy of the skin in the distal lower extremity, immunologic staining can be used to measure the density of small unmyelinated fibers. The density of these nerve fibers is reduced in patients with small-fiber neuropathies in whom NCS and routine nerve biopsies are often normal. This technique may allow for an objective measurement in patients with mainly subjective symptoms. However, it often adds little to what one already knows from the clinical examination and EDx.

SPECIFIC DISORDERS

HEREDITARY NEUROPATHIES

CMT disease is the most common type of hereditary neuropathy (Pattern 2, Table 438-2). Rather than one disease, CMT is a syndrome of many genetically distinct disorders **(Table 438-4)**. The various subtypes of CMT are classified according to the nerve conduction velocities (NCVs) and predominant pathology (e.g., demyelination or axonal degeneration), inheritance pattern (autosomal dominant, recessive, or X-linked), and the specific mutated genes. Type 1 CMT (or CMT1) refers to inherited demyelinating sensorimotor neuropathies, whereas the axonal sensory neuropathies are classified as CMT2. By definition, motor conduction velocities in the arms are slowed to <38 m/s in CMT1 and are >38 m/s in CMT2. However, most cases of CMT1 actually have motor NCVs between 20 and 25 m/s. CMT1 and CMT2 usually begin in childhood or early adult life; however, onset later in life can occur, particularly in CMT2. Both are inherited in an autosomal dominant fashion, with a few exceptions. CMT3 is an autosomal dominant neuropathy that appears in infancy and is associated with severe demyelination or hypomyelination. CMT4 is an autosomal recessive neuropathy that typically begins in childhood or early adult life. There are no medical therapies for any of the CMTs, but physical and occupational therapy can be beneficial, as can bracing (e.g., ankle-foot orthotics for footdrop) and other orthotic devices.

TABLE 438-4

Classification of Charcot-Marie-Tooth Disease and Related Neuropathies

Name	Inheritance	Gene Location	Gene Product
CMT1			
CMT1A	AD	17p11.2	PMP-22 (usually duplication of gene)
CMT1B	AD	1q21-23	MPZ
CMT1C	AD	16p13.1-p12.3	LITAF
CMT1D	AD	10q21.1-22.1	ERG2
CMT1E (with deafness)	AD	17p11.2	Point mutations in PMP 22 gene
CMT1F	AD	8p13-21	Neurofilament light chain
CMT1G	AD	14q32.33	INF2
CMT1X	X-linked dominant	Xq13	Connexin-32
HNPP	AD	17p11.2	PMP-22
		1q21-23	MPZ
CMT dominant-intermediate (CMTD1)			
CMTD1A	AD	10q24.1-25.1	?
CMTD1B	AD	19.p12-13.2	Dynamin 2
CMTD1C	AD	1p35	YARS
CMTD1D	AD	1q22	MPZ
CMT2			
CMT2A2 (allelic to HMSN VI with optic atrophy)	AD	1p36.2	MFN2
CMT2B	AD	3q13-q22	RAB7
CMT2B1 (allelic to LGMD 1B)	AR	1q21.2	Lamin A/C
CMT2B2	AR and AD	19q13	MED25 for AR; Unknown for AD
CMT2C (with vocal cord and diaphragm paralysis)	AD	12q23-24	TRPV4
CMT2D (allelic to distal SMA5)	AD	7p14	Glycine tRNA synthetase
CMT2E (allelic to CMT 1F)	AD	8p21	Neurofilament light chain
CMT2F	AD	7q11-q21	Heat-shock 27-kDa protein-1
CMT2G	AD	9q31.3-34.2	LRSAM1
CMT2I (allelic to CMT1B)	AD	1q22	MPZ
CMT2J	AD	1q22	MPZ

Name	Inheritance	Gene Location	Gene Product
CMT2H, CMT2K (allelic to CMT4A)	AD	8q13-q21	GDAP1
CMT2L (allelic to distal hereditary motor neuropathy type 2)	AD	12q24	Heat-shock protein 8
CMT2M	AD	16q22	Dynamin-2
CMT2N	AD	16q22.1	AARS
CMT2O	AD	14q32.31	DYNC1H1
CMT2P	AD	9q31.3-34.2	LRSAM1
CMT2P-Okinawa (HSMN2P)	AD	3q13-q14	TFG
CMT2X	X-linked	Xq22-24	PRPS1
CMT3	AD	17p11.2	PMP-22
(Dejerine-Sottas disease, congenital hypomyelinating neuropathy)	AD	1q21-23	MPZ
	AR	10q21.1-22.1	ERG2
	AR	19q13	Periaxon
CMT4			
CMT4A	AR	8q13-21.1	GDAP1
CMT4B1	AR	11q23	MTMR2
CMT4B2	AR	11p15	MTMR13
CMT4C	AR	5q23-33	SH3TC2
CMT4D (HMSN-Lom)	AR	8q24	NDRG1
CMT4E (congenital hypomyelinating neuropathy)	AR	Multiple	Includes PMP22, MPZ, and ERG-2
CMT4F	AR	19q13.1-13.3	Periaxin
CMT4G	AR	10q23.2	HK1
CMT4H	AR	12q12-q13	Frabin
CMT4J	AR	6q21	FIG4
HNA	AD	17q24	SEPT9
HSAN1A	AD	9q22	SPTLC1
HSAN1B	AD	3q21	RAB7
HSAN1C	AD	14q24.3	SPTLC2
HSAN1D	AD	14q21.3	ATL1
HSAN1E	AD	19p13.2	DNMT1

Name	Inheritance	Gene Location	Gene Product
HSAN2A	AR	12p13.33	PRKWNK1
HSAN2B	AR	5p15.1	FAM134B
HSAN2C	AR	12q13.13	KIF1A
HSAN2D	AR	2q24.3	SCN9A
HSAN3A	AR	9q21	IKAP
HSAN3B	AR	6p12.1	Dystonin
HSAN4	AR	3q	trkA/NGF receptor
HSAN5	AD or AR	1p11.2-p13.2	NGFb
		2q24.3	SCN9A
		3p22.2	SCN11A
HSAN6	AR	6p12.1	Dystonin

Abbreviations: AARS, alanyl-tRNA synthetase; AD, autosomal dominant; AR, autosomal recessive; ATL, atlastin; CMT, Charcot-Marie-Tooth; DNMT1, DNA methyltransferase 1; DYNC1HI, cytoplasmic dynein 1 heavy chain 1; ERG2, early growth response-2 protein; FAM134B, family with sequence similarity 134, member B; FIG4, FDG1-related F actin-binding protein; GDAP1, ganglioside-induced differentiation-associated protein-1; HK1, hexokinase 1; HMSN-P, hereditary motor and sensory neuropathyproximal; HNA, hereditary neuralgic amyotrophy; HNPP, hereditary neuropathy with liability to pressure palsies; HSAN; hereditary sensory and autonomic neuropathy; IFN2, inverted formin-2; IKAP, $_{\kappa}$B kinase complex-associated protein; LGMD, limb girdle muscular dystrophy; LITAF, lipopolysaccharide-induced tumor necrosis factor α factor; LRSAM1, E3 ubiquitin-protein ligase; MED25, mediator 25; MFN2, mitochondrial fusion protein mitofusin 2 gene; MPZ, myelin protein zero protein; MTMR2, myotubularin-related protein-2; NDRG1, N-myc downstream regulated 1; PMP-22, peripheral myelin protein-22; PRKWNK1, protein kinase, lysine deficient 1; PRPS1, phosphoribosylpyrophosphate synthetase 1; RAB7, Ras-related protein 7; SEPT9, Septin 9; SH3TC2, SH3 domain and tetratricopeptide repeats 2; SMA, spinal muscular atrophy; SPTLC, serine palmitoyltransferase long-chain base; TFG, TRK-fused gene; TrkA/NGF, tyrosine kinase A/nerve growth factor; tRNA, transfer ribonucleic acid; TRPV4, transient receptor potential cation channel, subfamily V, member 4; WNK1, WNK lysine deficient; YARS, tyrosyl-tRNA synthetase.

Source: Modified from AA Amato, J Russell: *Neuromuscular Disorders,* 2nd ed. New York, McGraw-Hill, 2016, Table 11-1, pp 265–266.

CMT1

CMT1 is the most common form of hereditary neuropathy. Affected individuals usually present in the first to third decade of life with distal leg weakness (e.g., foot drop), although patients may remain asymptomatic even late in life. People with CMT generally do not complain of numbness or tingling, which can be helpful in distinguishing CMT from acquired forms of neuropathy in which sensory symptoms usually predominate. Although usually asymptomatic, reduced sensation to all modalities is apparent on examination. Muscle stretch reflexes are unobtainable or reduced throughout. There is often atrophy of the muscles below the knee (particularly the anterior compartment), leading to so-called inverted champagne bottle legs.

Motor NCVs are generally in the 20–25 m/s range. Nerve biopsies usually are not performed on patients suspected of having CMT1, because the diagnosis usually can be made by less invasive testing (e.g., NCS and genetic studies). However, when done, the biopsies reveal reduced numbers of myelinated nerve fibers with a predilection for loss of large-diameter fibers and Schwann cell proliferation around thinly or demyelinated fibers, forming so-called onion bulbs.

CMT1A is the most common subtype of CMT1, representing 70% of cases, and is caused by a 1.5-megabase (Mb) duplication within chromosome 17p11.2-12 encoding the gene for peripheral myelin protein-22 (PMP-22). This results in patients having three copies of the PMP-22 gene rather than two. This protein accounts for 2–5% of myelin protein and is expressed in compact regions of the peripheral myelin sheath. Approximately 20% of patients with CMT1 have CMT1B, caused by mutations in the myelin protein zero (MPZ). CMT1B is for the most part clinically, electrophysiologically, and histologically indistinguishable from CMT1A. MPZ is an integral myelin protein and accounts for more than half of the myelin protein in peripheral nerves. Other forms of CMT1 are much less common and also indistinguishable from one another clinically and electrophysiologically.

CMT2

CMT2 occurs approximately half as frequently as CMT1 and CMT2 tends to present later in life. Affected individuals usually become symptomatic in the second decade; some cases present earlier in childhood, whereas others remain asymptomatic into late adult life. Clinically, CMT2 is for the most part indistinguishable from CMT1. NCS are helpful in this regard; in contrast to CMT1, the velocities are normal or only slightly slowed. The most common cause of CMT2 is a mutation in the gene for mitofusin 2 (MFN2), which accounts for approximately 20–30% of CMT2 cases overall. MFN2 localizes to the outer mitochondrial membrane, where it regulates the mitochondrial network architecture by participating in mitochondrial fusion. The other genes associated with CMT2 are much less common.

CMTDI

In dominant-intermediate CMTs (CMTDIs), the NCVs are faster than usually seen in CMT1 (e.g., >38 m/s) but slower than in CMT2.

CMT3

CMT3 was originally described by Dejerine and Sottas as a hereditary demyelinating sensorimotor polyneuropathy presenting in infancy or early childhood. Affected children are severely weak. Motor NCVs are markedly slowed, typically ≤5–10 m/s. Most cases of CMT3 are caused by point mutations in the genes for PMP-22, MPZ, or ERG-2, which are also the genes responsible for CMT1.

CMT4

CMT4 is extremely rare and is characterized by a severe, childhood-onset sensorimotor polyneuropathy that is usually inherited in an autosomal recessive fashion. Electrophysiologic and histologic evaluations can show demyelinating or axonal features. CMT4 is genetically heterogeneous (Table 438-4).

CMT1X

CMT1X is an X-linked dominant disorder with clinical features similar to CMT1 and CMT2, except that the neuropathy is much more severe in males than in females. CMT1X accounts for ~10–15% of CMT overall. Males usually present in the first two decades of life with atrophy and weakness of the distal arms and legs, areflexia, pes cavus, and hammer toes. Obligate female carriers are frequently asymptomatic, but can develop signs and symptoms of CMT. Onset in females is usually after the second decade of life, and the neuropathy is milder in severity.

NCS reveal features of both demyelination and axonal degeneration. In males, motor NCVs in the arms and legs are moderately slowed (in the low to mid 30-m/s range). About 50% of males with CMT1X have motor NCVs between 15 and 35 m/s with about 80% of these falling between 25 and 35 m/s (intermediate slowing). In contrast, about 80% of females with CMT1X have NCVs in the normal range and 20% have NCVs in the intermediate range. CMT1X is caused by mutations in the connexin 32 gene. Connexins are gap junction structural proteins that are important in cell-to-cell communication.

Hereditary Neuropathy with Liability to Pressure Palsies (HNPP)

HNPP is an autosomal dominant disorder related to CMT1A. While CMT1A is usually associated with a 1.5-Mb duplication in chromosome 17p11.2 that results in an extra copy of PMP-22 gene, HNPP is caused by inheritance of the chromosome with the corresponding 1.5-Mb deletion of this segment, and thus affected individuals have only one copy of the PMP-22 gene. Patients usually manifest in the second or third decade of life with painless numbness and weakness in the distribution of single peripheral nerves, although multiple mononeuropathies can occur (Pattern 3, Table 438-2). Symptomatic mononeuropathy or multiple mononeuropathies are often precipitated by trivial compression of nerve(s) as can occur with wearing a backpack, leaning on the elbows, or crossing one's legs for even a short period of time. These pressure-related mononeuropathies may take weeks or months to resolve. In addition, some affected individuals manifest with a progressive or relapsing, generalized and symmetric, sensorimotor peripheral neuropathy that resembles CMT.

Hereditary Neuralgic Amyotrophy (HNA)

HNA is an autosomal dominant disorder characterized by recurrent attacks of pain, weakness, and sensory loss in the distribution of the brachial plexus often beginning in childhood (Pattern 4, Table 438-2). These attacks are similar to those seen with idiopathic brachial plexitis (see below). Attacks may occur in the postpartum period, following surgery, or at other times of stress. Most patients recover over several weeks or months. Slightly dysmorphic features, including hypotelorism, epicanthal folds, cleft palate, syndactyly, micrognathia, and facial asymmetry, are evident in some individuals. EDx demonstrate an axonal process. HNA is genetically heterogeneous but can be caused by mutations in septin 9 (*SEPT9*). Septins may be important in formation of the neuronal cytoskeleton and have a role in cell division, but it is not known how mutations in SEPT9 lead to HNA.

Hereditary Sensory and Autonomic Neuropathy (HSAN)

The HSANs are a very rare group of hereditary neuropathies in which sensory and autonomic dysfunction predominates over muscle weakness, unlike CMT, in which motor findings are most prominent (Pattern 2, Table 438-2; Table 438-4). Nevertheless, affected individuals can develop motor weakness and there can be overlap with CMT. There are no medical therapies available to treat these neuropathies, other than prevention and treatment of mutilating skin and bone lesions.

Of the HSANs, only HSAN1 typically presents in adults. HSAN1 is the most common of the HSANs and is inherited in an autosomal dominant fashion. Affected individuals usually manifest in the second through fourth decades of life. HSAN1 is associated with the degeneration of small myelinated and unmyelinated nerve fibers leading to severe loss of pain and temperature sensation, deep dermal ulcerations, recurrent osteomyelitis, Charcot joints, bone loss, gross foot and hand deformities, and amputated digits. Although most people with HSAN1 do not complain of numbness, they often describe burning, aching, or lancinating pains. Autonomic neuropathy is not a prominent feature, but bladder dysfunction and reduced sweating in the feet may occur. HSAN1A, which is most common, is caused by mutations in the serine palmitoyltransferase long-chain base 1 (*SPTLC1*) gene.

OTHER HEREDITARY NEUROPATHIES

(Table 438-5)

TABLE 438-5

Rare Hereditary Neuropathies

Hereditary Disorders of Lipid Metabolism
Metachromatic leukodystrophy
Krabbe's disease (globoid cell leukodystrophy)
Fabry's disease
Adrenoleukodystrophy/adrenomyeloneuropathy
Refsum's disease
Tangier disease
Cerebrotendinous xanthomatosis
Hereditary Ataxias with Neuropathy
Friedreich's ataxia
Vitamin E deficiency
Spinocerebellar ataxia
Abetalipoproteinemia (Bassen-Kornzweig disease)
Disorders of Defective DNA Repair
Ataxia-telangiectasia
Cockayne's syndrome
Giant Axonal Neuropathy
Porphyria
Acute intermittent porphyria (AIP)
Hereditary coproporphyria (HCP)
Variegate porphyria (VP)
Familial Amyloid Polyneuropathy (FAP)
Transthyretin-related
Gelsolin-related
Apolipoprotein A1-related

Source: Modified from AA Amato, J Russell: *Neuromuscular Disorders,* 2nd ed. New York, McGraw-Hill, 2016, Table 12-1, p. 299.

FABRY'S DISEASE

Fabry's disease (angiokeratoma corporis diffusum) is an X-linked dominant disorder. Although men are more commonly and severely affected, women can also manifest symptoms and signs of the disease. Angiokeratomas are reddish-purple maculopapular lesions that are usually found around the umbilicus,

scrotum, inguinal region, and perineum. Burning or lancinating pain in the hands and feet often develops in males in late childhood or early adult life (Pattern 2, Table 438-2). However, the neuropathy is usually overshadowed by complications arising from an associated premature atherosclerosis (e.g., hypertension, renal failure, cardiac disease, and stroke) that often lead to death by the fifth decade of life. Some patients also manifest primarily with a dilated cardiomyopathy.

Fabry's disease is caused by mutations in the a-galactosidase gene that leads to the accumulation of ceramide trihexoside in nerves and blood vessels. A decrease in a-galactosidase activity is evident in leukocytes and cultured fibroblasts. Glycolipid granules may be appreciated in ganglion cells of the peripheral and sympathetic nervous systems and in perineurial cells. Enzyme replacement therapy with a-galactosidase B can improve the neuropathy if patients are treated early, before irreversible nerve fiber loss develops.

ADRENOLEUKODYSTROPHY/ADRENOMYELONEUROPATHY

Adrenoleukodystrophy (ALD) and AMN are allelic X-linked dominant disorders caused by mutations in the peroxisomal transmembrane adenosine triphosphate-binding cassette (ABC) transporter gene. Patients with ALD manifest with central nervous system (CNS) abnormalities. However, ~30% of patients with mutations in this gene present with the AMN phenotype that typically manifests in the third to fifth decade of life as mild to moderate peripheral neuropathy combined with progressive spastic paraplegia (Pattern 6, Table 438-2) **(Chap. 434)**. Rare patients present with an adult-onset spinocerebellar ataxia or only with adrenal insufficiency.

EDx is suggestive of a primary axonopathy with secondary demyelination. Nerve biopsies demonstrate a loss of myelinated and unmyelinated nerve fibers with lamellar inclusions in the cytoplasm of Schwann cells. Very long chain fatty acid (VLCFA) levels (C24, C25, and C26) are increased in the urine. Laboratory evidence of adrenal insufficiency is evident in approximately two-thirds of patients. The diagnosis can be confirmed by genetic testing.

Adrenal insufficiency is managed by replacement therapy; however, there is no proven effective therapy for the neurologic manifestations of ALD/AMN. Diets low in VLCFAs and supplemented with Lorenzo's oil (erucic and oleic acids) reduce the levels of VLCFAs and increase the levels of C22 in serum, fibroblasts, and liver; however, several large, open-label trials of Lorenzo's oil failed to demonstrate efficacy.

REFSUM'S DISEASE

Refsum's disease can manifest in infancy to early adulthood with the classic tetrad of (1) peripheral neuropathy, (2) retinitis pigmentosa, (3) cerebellar ataxia, and (4) elevated CSF protein concentration. Most affected individuals develop progressive distal sensory loss and weakness in the legs leading to foot drop by their 20 s (Pattern 2, Table 438-2). Subsequently, the proximal leg and arm muscles may become weak. Patients may also develop sensorineural hearing loss, cardiac conduction abnormalities, ichthyosis, and anosmia.

Serum phytanic acid levels are elevated. Sensory and motor NCS reveal reduced amplitudes, prolonged latencies, and slowed conduction velocities. Nerve biopsy demonstrates a loss of myelinated nerve fibers, with remaining axons often thinly myelinated and associated with onion bulb formation.

Refsum's disease is genetically heterogeneous but autosomal recessive in nature. Classical Refsum's disease with childhood or early adult onset is caused by mutations in the gene that encodes for phytanoyl-CoA α-hydroxylase (*PAHX*). Less commonly, mutations in the gene encoding peroxin 7 receptor protein (PRX7) are responsible. These mutations lead to the accumulation of phytanic acid in the central and peripheral nervous systems. Treatment is removal of phytanic precursors (phytols: fish oils, dairy products, and ruminant fats) from the diet.

TANGIER DISEASE

Tangier disease is a rare autosomal recessive disorder that can present as (1) asymmetric multiple mononeuropathies, (2) a slowly progressive symmetric polyneuropathy predominantly in the legs, or (3) a pseudo-syringomyelia pattern with dissociated sensory loss (i.e., abnormal pain/temperature perception but preserved position/vibration in the arms **[Chap. 434]**). The tonsils may appear swollen and yellowish-orange in color, and there may also be splenomegaly and lymphadenopathy.

Tangier disease is caused by mutations in the ATP-binding cassette transporter 1 (ABC1) gene, which leads to markedly reduced levels of high-density lipoprotein (HDL) cholesterol levels, whereas triacylglycerol levels are increased. Nerve biopsies reveal axonal degeneration with demyelination and remyelination. Electron microscopy demonstrates abnormal accumulation of lipid in Schwann cells, particularly those encompassing umyelinated and small myelinated nerves. There is no specific treatment.

PORPHYRIA

Porphyria is a group of inherited disorders caused by defects in heme biosynthesis **(Chap. 409)**. Three forms of porphyria are associated with peripheral neuropathy: acute intermittent porphyria (AIP), hereditary coproporphyria (HCP), and variegate porphyria (VP). The acute neurologic manifestations are similar in each, with the exception that a photosensitive rash is seen with HCP and VP but not in AIP. Attacks of porphyria can be precipitated by certain drugs (usually those metabolized by the P450 system), hormonal changes (e.g., pregnancy, menstrual cycle), and dietary restrictions.

An acute attack of porphyria may begin with sharp abdominal pain. Subsequently, patients may develop agitation, hallucinations, or seizures. Several days later, back and extremity pain followed by weakness ensues, mimicking GBS (Pattern 1, Table 438-2). Weakness can involve the arms or the legs and can be asymmetric, proximal, or distal in distribution, as well as affecting the face and bulbar musculature. Dysautonomia and signs of sympathetic overactivity are common (e.g., pupillary dilation, tachycardia, and hypertension). Constipation, urinary retention, and incontinence can also be seen.

The CSF protein is typically normal or mildly elevated. Liver function tests and hematologic parameters are usually normal. Some patients are hyponatremic due to inappropriate secretion of antidiuretic hormone **(Chap. 371)**. The urine may appear brownish in color secondary to the high concentration of porphyrin metabolites. Accumulation of intermediary precursors of heme (i.e., d-aminolevulinic acid, porphobilinogen, uroporphobilinogen, coproporphyrinogen, and protoporphyrinogen) is found in urine. Specific enzyme activities can also be measured in erythrocytes and leukocytes. The primary abnormalities on EDx are marked reductions in compound motor action potential (CMAP) amplitudes and signs of active axonal degeneration on needle EMG.

The porphyrias are inherited in an autosomal dominant fashion. AIP is associated with porphobilinogen deaminase deficiency, HCP is caused by defects in coproporphyrin oxidase, and VP is associated with protoporphyrinogen oxidase deficiency. The pathogenesis of the neuropathy is not completely understood. Treatment with glucose and hematin may reduce the accumulation of heme precursors. Intravenous glucose is started at a rate of 10–20 g/h. If there is no improvement within 24 h, intravenous hematin 2–5 mg/kg per day for 3–14 days should be administered.

FAMILIAL AMYLOID POLYNEUROPATHY

Familial amyloid polyneuropathy (FAP) is phenotypically and genetically heterogeneous and is caused by mutations in the genes for transthyretin (TTR), apolipoprotein A1, or gelsolin **(Chap. 108)**. The majority of patients with FAP have mutations in the TTR gene. Amyloid deposition may be evident in abdominal fat pad, rectal, or nerve biopsies. The clinical features, histopathology, and EDx reveal abnormalities consistent with a generalized or multifocal, predominantly axonal but occasionally demyelinating, polyneuropathy.

Patients with TTR-related FAP usually develop insidious onset of numbness and painful paresthesias in the distal lower limbs in the third to fourth decade of life, although some patients develop the disorder later in life (Pattern 2, Table 438-2). Carpal tunnel syndrome (CTS) is common. Autonomic involvement can be severe, leading to postural hypotension, constipation or persistent diarrhea, erectile dysfunction, and impaired sweating (Pattern 10, Table 438-2). Amyloid deposition also occurs in the heart, kidneys, liver, and corneas. Patients usually die 10–15 years after the onset of symptoms from cardiac failure or complications from malnutrition. Because the liver produces much of the body's TTR, liver transplantation has been used to treat FAP related to TTR mutations. Serum TTR levels decrease after transplantation, and improvement in clinical and EDx features has been reported. Both tafamidis meglumine (20 mg daily) and diflunisal (250 mg twice daily), which prevent misfolding and deposition of mutated TTR, appear to slow the rate of deterioration patients with TTR-related FAP. Recent studies have shown very promising results with gene therapy.

Patients with apolipoprotein A1-related FAP (Van Allen type) usually present in the fourth decade with numbness and painful dysesthesias in the distal limbs. Gradually, the symptoms progress, leading to proximal and distal weakness and atrophy. Although autonomic neuropathy is not severe, some patients develop diarrhea, constipation, or gastroparesis. Most patients die from systemic complications of amyloidosis (e.g., renal failure) 12–15 years after the onset of the neuropathy.

Gelsolin-related amyloidosis (Finnish type) is characterized by the combination of lattice corneal dystrophy and multiple cranial neuropathies that usually begin in the third decade of life. Over time, a mild generalized sensorimotor polyneuropathy develops. Autonomic dysfunction does not occur.

ACQUIRED NEUROPATHIES

PRIMARY OR AL AMYLOIDOSIS

Besides FAP, amyloidosis can also be acquired **(See Chap. 108)**. In primary or AL amyloidosis, the abnormal protein deposition is composed of immunoglobulin light chains. AL amyloidosis occurs in the setting of multiple myeloma (MM), Waldenström's macroglobulinemia, lymphoma, other plasmacytomas, or lymphoproliferative disorders, or without any other identifiable disease.

Approximately 30% of patients with AL primary amyloidosis present with a polyneuropathy, most typically painful dysesthesias and burning sensations in the feet (Pattern 2, Table 438-2). However, the trunk can be involved, and some patients manifest with a mononeuropathy multiplex pattern. CTS occurs in 25% of patients and may be the initial manifestation. The neuropathy is slowly progressive, and eventually weakness develops along with large-fiber sensory loss. Most patients develop autonomic involvement with postural hypertension, syncope, bowel and bladder incontinence, constipation, impotence, and impaired sweating (Pattern 10, Table 438-2). Patients generally die from their systemic illness (renal failure, cardiac disease).

The monoclonal protein may be composed of IgG, IgA, IgM, or only free light chain. Lambda (λ) is more common than κ light chain (>2:1) in AL amyloidosis. The CSF protein is often increased (with normal cell count), and thus the neuropathy may be mistaken for CIDP **(Chap. 439)**. Nerve biopsies reveal axonal degeneration and amyloid deposition in either a globular or diffuse pattern infiltrating the perineurial, epineurial, and endoneurial connected tissue and in blood vessel walls.

The median survival of patients with primary amyloidosis is <2 years, with death usually from progressive congestive heart failure or renal failure. Chemotherapy with melphalan, prednisone, and colchicine, to reduce the concentration of monoclonal proteins, and autologous stem cell transplantation may prolong survival, but whether the neuropathy improves is controversial.

DIABETIC NEUROPATHY

DM is the most common cause of peripheral neuropathy in developed countries. DM is associated with several types of polyneuropathy: distal symmetric sensory or sensorimotor polyneuropathy, autonomic neuropathy, diabetic neuropathic cachexia, polyradiculoneuropathies, cranial neuropathies, and other mononeuropathies. Risk factors for the development of neuropathy include long-standing, poorly controlled DM and the presence of retinopathy and nephropathy.

Diabetic Distal Symmetric Sensory and Sensorimotor Polyneuropathy (DSPN)

DSPN is the most common form of diabetic neuropathy and manifests as sensory loss beginning in the toes that gradually progresses over time up the legs and into the fingers and arms (Pattern 2, Table 438-2). When severe, a patient may develop sensory loss in the trunk (chest and abdomen), initially in the midline anteriorly and later extending laterally. Tingling, burning, deep aching pains may also be apparent. NCS usually show reduced amplitudes and mild to moderate slowing of conduction velocities. Nerve biopsy reveals axonal degeneration, endothelial hyperplasia, and, occasionally, perivascular inflammation. Tight control of glucose can reduce the risk of developing neuropathy or improve the underlying neuropathy. A variety of medications have been used with variable success to treat painful symptoms associated with DSPN, including antiepileptic medications, antidepressants, sodium channel blockers, and other analgesics **(Table 438-6)**.

TABLE 438-6

Treatment of Painful Sensory Neuropathies

Therapy	Route	Dose	Side Effects
First-Line			
Lidoderm 5% patch	Apply to painful area	Up to 3 patches qd	Skin irritation
Tricyclic antidepressants (e.g., amitriptyline, nortriptyline)	PO	10–100 mg qhs	Cognitive changes, sedation, dry eyes and mouth, urinary retention, constipation
Gabapentin	PO	300–1200 mg tid	Cognitive changes, sedation, peripheral edema
Pregabalin	PO	50–100 mg tid	Cognitive changes, sedation, peripheral edema
Duloxetine	PO	30–60 mg qd	Cognitive changes, sedation, dry eyes, diaphoresis, nausea, diarrhea, constipation
Second-Line			
Carbamazepine	PO	200–400 mg q 6–8 h	Cognitive changes, dizziness, leukopenia, liver dysfunction
Phenytoin	PO	200–400 mg qhs	Cognitive changes, dizziness, liver dysfunction
Venlafaxine	PO	37.5–150 mg/d	Asthenia, sweating, nausea, constipation, anorexia, vomiting, somnolence, dry mouth, dizziness, nervousness, anxiety, tremor, and blurred vision as well as abnormal ejaculation/orgasm and impotence
Tramadol	PO	50 mg qid	Cognitive changes, gastrointestinal upset
Third-Line			
Mexiletine	PO	200–300 mg tid	Arrhythmias
Other Agents			
EMLA cream	Apply cutaneously	qid	Local erythema
2.5% lidocaine			
2.5% prilocaine			
Capsaicin 0.025–0.075% cream	Apply cutaneously	qid	Painful burning skin

Source: Modified from AA Amato, J Russell: *Neuromuscular Disorders,* 2nd ed. New York, McGraw-Hill, 2016, Table 22-3, p 485.

Diabetic Autonomic Neuropathy

Autonomic neuropathy is typically seen in combination with DSPN. The autonomic neuropathy can manifest as abnormal sweating, dysfunctional thermoregulation, dry eyes and mouth, pupillary abnormalities, cardiac arrhythmias, postural hypotension, GI abnormalities (e.g., gastroparesis, postprandial bloating, chronic diarrhea, or constipation), and genitourinary dysfunction (e.g., impotence, retrograde ejaculation, incontinence) (Pattern 10, Table 438-2). Tests of autonomic function are generally abnormal, including sympathetic skin responses and quantitative sudomotor axon reflex testing. Sensory and motor NCS generally demonstrate features described above with DSPN.

Diabetic Radiculoplexus Neuropathy (Diabetic Amyotrophy or Bruns-Garland Syndrome)

Diabetic radiculoplexus neuropathy is the presenting manifestation of DM in approximately one-third of patients. Typically, patients present with severe pain in the low back, hip, and thigh in one leg. Rarely, the diabetic polyradiculoneuropathy begins in both legs at the same time (Pattern 4, Table 438-2). Atrophy and weakness of proximal and distal muscles in the affected leg become apparent within a few days or weeks. The neuropathy is often accompanied or heralded by severe weight loss. Weakness usually progresses over several weeks or months, but can continue to progress for 18 months or more. Subsequently, there is slow recovery but many are left with residual weakness, sensory loss, and pain. In contrast to the more typical lumbosacral radiculoplexus neuropathy, some patients develop thoracic radiculopathy or, even less commonly, a cervical polyradiculoneuropathy. CSF protein is usually elevated, while the cell count is normal. ESR is often increased. EDx reveals evidence of active denervation in affected proximal and distal muscles in the affected limbs and in paraspinal muscles. Nerve biopsies may demonstrate axonal degeneration along with perivascular inflammation. Patients with severe pain are sometimes treated in the acute period with glucocorticoids, although a randomized controlled trial has yet to be performed, and the natural history of this neuropathy is gradual improvement.

Diabetic Mononeuropathies or Multiple Mononeuropathies

The most common mononeuropathies are median neuropathy at the wrist and ulnar neuropathy at the elbow, but peroneal neuropathy at the fibular head, and sciatic, lateral femoral, cutaneous, or cranial neuropathies also occur (Pattern 3, Table 438-2). In regard to cranial mononeuropathies, seventh nerve palsies are relatively common but may have other, nondiabetic etiologies. In diabetics, a third nerve palsy is most common, followed by sixth nerve, and, less frequently, fourth nerve palsies. Diabetic third nerve palsies are characteristically pupil-sparing **(Chap. 28)**.

HYPOTHYROIDISM

Hypothyroidism is more commonly associated with a proximal myopathy, but some patients develop a neuropathy, most typically CTS. Rarely, a generalized sensory polyneuropathy characterized by painful paresthesias and numbness in both the legs and hands can occur. Treatment is correction of the hypothyroidism.

SJÖGREN'S SYNDROME

Sjögren's syndrome, characterized by the sicca complex of xerophthalmia, xerostomia, and dryness of other mucous membranes, can be complicated by neuropathy **(Chap. 354)**. Most common is a length-dependent axonal sensorimotor neuropathy characterized mainly by sensory loss in the distal extremities (Pattern 2, Table 438-2). A pure small-fiber neuropathy or a cranial neuropathy, particularly involving the trigeminal nerve, can also be seen. Sjögren's syndrome is also associated with sensory neuronopathy/ganglionopathy. Patients with sensory ganglionopathies develop progressive numbness and tingling of the limbs, trunk, and face in a non-length-dependent manner such that symptoms can involve the face or arms more than the legs. The onset can be acute or insidious. Sensory examination demonstrates severe vibratory and proprioceptive loss leading to sensory ataxia.

Patients with neuropathy due to Sjögren's syndrome may have ANAs, SS-A/Ro, and SS-B/La antibodies in the serum, but most do not. NCS demonstrate reduced amplitudes of sensory studies in the affected limbs. Nerve biopsy demonstrates axonal degeneration. Nonspecific perivascular inflammation may be present, but only rarely is there necrotizing vasculitis. There is no specific treatment for neuropathies related to Sjögren's syndrome. When vasculitis is suspected, immunosuppressive agents may be beneficial. Occasionally, the sensory neuronopathy/ganglionopathy stabilizes or improves with immunotherapy, such as IVIg.

RHEUMATOID ARTHRITIS

Peripheral neuropathy occurs in at least 50% of patients with rheumatoid arthritis (RA) and may be vasculitic in nature **(Chap. 351)**. Vasculitic neuropathy can present with a mononeuropathy multiplex (Pattern 3, Table 438-2), a generalized symmetric pattern of involvement (Pattern 2, Table 438-2), or a combination of these patterns **(Chap. 356)**. Neuropathies may also result from drugs used to treat RA (e.g., tumor necrosis blockers, leflunomide). Nerve biopsy often reveals thickening of the epineurial and endoneurial blood vessels as well as perivascular inflammation or vasculitis, with transmural inflammatory cell infiltration and fibrinoid necrosis of vessel walls. The neuropathy is usually responsive to immunomodulating therapies.

SYSTEMIC LUPUS ERYTHEMATOSUS

Between 2 and 27% of individuals with SLE develop a peripheral neuropathy **(Chap. 349)**. Affected patients typically present with a slowly progressive sensory loss beginning in the feet. Some patients develop burning pain and paresthesias with normal reflexes, and NCS suggest a pure small-fiber neuropathy (Pattern 2, Table 438-2). Less common are multiple mononeuropathies presumably secondary to necrotizing vasculitis (Pattern 3, Table 438-2). Rarely, a generalized sensorimotor polyneuropathy meeting clinical, laboratory, electrophysiologic, and histologic criteria for either GBS or CIDP may occur. Immunosuppressive therapy may be beneficial in SLE patients with neuropathy due to vasculitis. Immunosuppressive agents are less likely to be effective in patients with a generalized sensory or sensorimotor polyneuropathy without evidence of vasculitis. Patients with a GBS or CIDP-like neuropathy should be treated accordingly **(Chap. 439)**.

SYSTEMIC SCLEROSIS (SCLERODERMA)

A distal symmetric, mainly sensory, polyneuropathy complicates 5–67% of scleroderma cases (Pattern 2, Table 438-2) **(Chap. 353)**. Cranial mononeuropathies can also develop, most commonly of the trigeminal nerve, producing numbness and dysesthesias in the face. Multiple mononeuropathies also occur (Pattern 3, Table 438-2). The EDx and histologic features of nerve biopsy are those of an axonal sensory greater than motor polyneuropathy.

MIXED CONNECTIVE TISSUE DISEASE (MCTD)

A mild distal axonal sensorimotor polyneuropathy occurs in ~10% of patients with MCTD.

SARCOIDOSIS

The peripheral or CNS is involved in about 5% of patients with sarcoidosis **(Chap. 360)**. The most common cranial nerve involved is the seventh nerve, which can be affected bilaterally. Some patients develop radiculopathy or polyradiculopathy (Pattern 4, Table 438-2). With a generalized root involvement, the clinical presentation can mimic GBS or CIDP. Patients can also present with multiple mononeuropathies (Pattern 3, Table 438-2) or a generalized, slowly progressive, sensory greater than motor polyneuropathy (Pattern 2, Table 438-2). Some have features of a pure small-fiber neuropathy. EDx reveals an axonal neuropathy. Nerve biopsy can reveal noncaseating granulomas infiltrating the endoneurium, perineurium, or epineurium along with lymphocytic necrotizing angiitis. Neurosarcoidosis may respond to treatment with glucocorticoids or other immunosuppressive agents.

HYPEREOSINOPHILIC SYNDROME

Hypereosinophilic syndrome is characterized by eosinophilia associated with various skin, cardiac, hematologic, and neurologic abnormalities. A generalized peripheral neuropathy or a mononeuropathy multiplex occurs in 6–14% of patients (Pattern 2, Table 438-2).

CELIAC DISEASE (GLUTEN-INDUCED ENTEROPATHY OR NONTROPICAL SPRUE)

Neurologic complications, particularly ataxia and peripheral neuropathy, are estimated to occur in 10% of patients with celiac disease **(Chap. 318)**. A generalized sensorimotor polyneuropathy, pure motor neuropathy, multiple mononeuropathies, autonomic neuropathy, small-fiber neuropathy, and neuromyotonia have all been reported in association with celiac disease or antigliadin/antiendomysial antibodies (Patterns 2, 3, and 9; Table 438-2). Nerve biopsy may reveal a loss of large myelinated fibers. The neuropathy may be secondary to malabsorption of vitamins B_{12} and E. However, some patients have no appreciable vitamin deficiencies. The pathogenic basis for the neuropathy in these patients is unclear but may be autoimmune in etiology. The neuropathy does not appear to respond to a gluten-free diet. In patients with vitamin B_{12} or vitamin E deficiency, replacement therapy may improve or stabilize the neuropathy.

INFLAMMATORY BOWEL DISEASE

Ulcerative colitis and Crohn's disease may be complicated by GBS, CIDP, generalized axonal sensory or sensorimotor polyneuropathy, small-fiber neuropathy, or mononeuropathy (Patterns 2 and 3, Table 438-2) **(Chap. 319)**. These neuropathies may be autoimmune, nutritional (e.g., vitamin B_{12} deficiency), treatment related (e.g., metronidazole), or idiopathic in nature. An acute neuropathy with demyelination resembling GBS, CIDP, or multifocal motor neuropathy may occur in patients treated with tumor necrosis factor α blockers.

UREMIC NEUROPATHY

Approximately 60% of patients with renal failure develop a polyneuropathy characterized by length-dependent numbness, tingling, allodynia, and mild distal weakness (Pattern 2, Table 438-2). Rarely, a rapidly progressive weakness and sensory loss very similar to GBS can occur that improves with an increase in the intensity of renal dialysis or with transplantation (Pattern 1, Table 438-2). Mononeuropathies can also occur, the most common of which is CTS. Ischemic monomelic neuropathy (see below) can complicate arteriovenous shunts created in the arm for dialysis (Pattern 3, Table 438-2). EDx in uremic patients reveals features of a length-dependent, primarily axonal, sensorimotor polyneuropathy. Sural nerve biopsies demonstrate a loss of nerve fibers (particularly large myelinated nerve fibers), active axonal degeneration, and segmental and paranodal demyelination. The sensorimotor polyneuropathy can be stabilized by hemodialysis and improved with successful renal transplantation.

CHRONIC LIVER DISEASE

A generalized sensorimotor neuropathy characterized by numbness, tingling, and minor weakness in the distal aspects of primarily the lower limbs commonly occurs in patients with chronic liver failure. EDx studies are consistent with a sensory greater than motor axonopathy. Occasionally patients with severe liver disease develop a combined neuropathy and myopathy. Sural nerve biopsy reveals both segmental demyelination and axonal loss. It is not known if hepatic failure in isolation can cause peripheral neuropathy, as the majority of patients have liver disease secondary to other disorders, such as alcoholism or viral hepatitis, which can also cause neuropathy.

CRITICAL ILLNESS POLYNEUROPATHY

The most common causes of acute generalized weakness leading to admission to a medical intensive care unit (ICU) are GBS and myasthenia gravis (Pattern 1, Table 438-2) **(Chaps. 439 and 440)**. However, weakness developing in critically ill patients while in the ICU is usually caused by critical illness polyneuropathy (CIP) or critical illness myopathy (CIM) or, much less commonly, by prolonged neuromuscular blockade. From a clinical and EDx standpoint, it can be quite difficult to distinguish these disorders. Most specialists believe that CIM is more common. Both CIM and CIP develop as a complication of sepsis and multiple organ failure. They usually present as an inability to wean a patient from a ventilator. A coexisting encephalopathy may limit the neurologic examination, in particular the sensory examination. Muscle stretch reflexes are absent or reduced.

Serum creatine kinase (CK) is usually normal; an elevated serum CK would point to CIM as opposed to CIP. NCS reveal absent or markedly reduced amplitudes of motor and sensory studies in CIP, whereas sensory studies are relatively preserved in CIM. Needle EMG usually reveals profuse positive sharp waves and fibrillation potentials, and it is not unusual in patients with severe weakness to be unable to recruit motor unit action potentials. The pathogenic basis of CIP is not known. Perhaps circulating toxins and metabolic abnormalities associated with sepsis and multiorgan failure impair axonal transport or mitochondrial function, leading to axonal degeneration.

LEPROSY (HANSEN'S DISEASE)


Leprosy, caused by the acid-fast bacteria *Mycobacterium leprae*, is the most common cause of peripheral neuropathy in Southeast Asia, Africa, and South America **(Chap. 174)**. Clinical manifestations range from tuberculoid leprosy at one end of the spectrum to lepromatous leprosy at the other end, with borderline leprosy in between. Neuropathies are most common in patients with borderline leprosy. Superficial cutaneous nerves of the ears and distal limbs are commonly affected. Mononeuropathies, multiple mononeuropathies, or a slowly progressive symmetric sensorimotor polyneuropathy may develop (Patterns 2 and 3, Table 438-2). Sensory NCS are usually absent in the lower limb and are reduced in amplitude in the arms. Motor NCS may demonstrate reduced amplitudes in affected nerves but occasionally can reveal demyelinating features. Leprosy is usually diagnosed by skin lesion biopsy. Nerve biopsy can also be diagnostic, particularly when there are no apparent skin lesions. The tuberculoid form is characterized by granulomas, and bacilli are not seen. In contrast, with lepromatous leprosy, large numbers of infiltrating bacilli, T_H2 lymphocytes, and organism-laden, foamy macrophages with minimal granulomatous infiltration are evident. The bacilli are best appreciated using the Fite stain, where they can be seen as red-staining rods often in clusters free in the endoneurium, within macrophages, or within Schwann cells.

Patients are generally treated with multiple drugs: dapsone, rifampin, and clofazimine. Other medications that are used include thalidomide, pefloxacin, ofloxacin, sparfloxacin, minocycline, and clarithromycin. Patients are generally treated for 2 years. Treatment is sometimes complicated by the so-called reversal reaction, particularly in borderline leprosy. The reversal reaction can occur at any time during treatment and develops because of a shift to the tuberculoid end of the spectrum, with an increase in cellular immunity during treatment. The cellular response is upregulated as evidenced by an increased release of tumor necrosis factor α, interferon γ, and interleukin 2, with new granuloma formation. This can result in an exacerbation of the rash and the neuropathy as well as in appearance of new lesions. High-dose glucocorticoids blunt this adverse reaction and may be used prophylactically at treatment onset in high-risk patients. Erythema nodosum leprosum (ENL) is also treated with glucocorticoids or thalidomide.

LYME DISEASE

Lyme disease is caused by infection with *Borrelia burgdorferi*, a spirochete usually transmitted by the deer tick *Ixodes dammini* **(Chap. 181)**. Neurologic complications may develop during the second and third stages of infection. Facial neuropathy is most common and is bilateral in about half of cases, which is rare for idiopathic Bell's palsy. Involvement of nerves is frequently asymmetric. Some patients present with a polyradiculoneuropathy or multiple mononeuropathies (Pattern 3 or 4, Table 438-2). EDx is suggestive of a primary axonopathy. Nerve biopsies can reveal axonal degeneration with perivascular inflammation. Treatment is with antibiotics **(Chap. 181)**.

DIPHTHERITIC NEUROPATHY

Diphtheria is caused by the bacteria *Corynebacterium diphtheriae* **(Chap. 145)**. Infected individuals present with flulike symptoms of generalized myalgias, headache, fatigue, low-grade fever, and irritability within a week to 10 days of the exposure. About 20–70% of patients develop a peripheral neuropathy caused by a toxin released by the bacteria. Three to four weeks after infection, patients may note decreased sensation in their throat and begin to develop dysphagia, dysarthria, hoarseness, and blurred vision due to impaired accommodation. A generalized polyneuropathy may manifest 2 or 3 months following the initial infection, characterized by numbness, paresthesias, and weakness of the arms and legs and occasionally ventilatory failure (Pattern 1,

Table 438-2). CSF protein can be elevated with or without lymphocytic pleocytosis. EDx suggests a diffuse axonal sensorimotor polyneuropathy. Antitoxin and antibiotics should be given within 48 h of symptom onset. Although early treatment reduces the incidence and severity of some complications (i.e., cardiomyopathy), it does not appear to alter the natural history of the associated peripheral neuropathy. The neuropathy usually resolves after several months.

HUMAN IMMUNODEFICIENCY VIRUS

HIV infection can result in a variety of neurologic complications, including peripheral neuropathies **(Chap. 197)**. Approximately 20% of HIV-infected individuals develop a neuropathy either as a direct result of the virus itself, other associated viral infections (e.g., CMV), or neurotoxicity secondary to antiviral medications (see below). The major presentations of peripheral neuropathy associated with HIV infection include (1) distal symmetric polyneuropathy (DSP), (2) inflammatory demyelinating polyneuropathy (including both GBS and CIDP), (3) multiple mononeuropathies (e.g., vasculitis, CMV-related), (4) polyradiculopathy (usually CMV-related), (5) autonomic neuropathy, and (6) sensory ganglionitis.

HIV-Related Distal Symmetric Polyneuropathy

DSP is the most common form of peripheral neuropathy associated with HIV infection and usually is seen in patients with AIDS. It is characterized by numbness and painful paresthesias involving the distal extremities (Pattern 2, Table 438-2). The pathogenic basis for DSP is unknown but is not due to actual infection of the peripheral nerves. The neuropathy may be immune mediated, perhaps caused by the release of cytokines from surrounding inflammatory cells. Vitamin B_{12} deficiency may contribute in some instances but is not a major cause of most cases of DSP. Older antiretroviral agents (e.g., dideoxycytidine, dideoxyinosine, stavudine) are also neurotoxic and can cause a painful sensory neuropathy.

HIV-Related Inflammatory Demyelinating Polyradiculoneuropathy

Both acute inflammatory demyelinating polyneuropathy (AIDP) and CIDP can occur as a complication of HIV infection (Pattern 1, Table 438-2). AIDP usually develops at the time of seroconversion, whereas CIDP can occur any time in the course of the infection. Clinical and EDx features are indistinguishable from idiopathic AIDP or CIDP **(Chap. 439)**. In addition to elevated protein levels, lymphocytic pleocytosis is evident in the CSF, a finding that helps distinguish this HIV-associated polyradiculoneuropathy from idiopathic AIDP/CIDP.

HIV-Related Progressive Polyradiculopathy

An acute, progressive lumbosacral polyradiculoneuropathy usually secondary to CMV infection can develop in patients with AIDS (Pattern 4, Table 438-2). Patients present with severe radicular pain, numbness, and weakness in the legs, which is usually asymmetric. CSF is abnormal, demonstrating a high protein level, along with a reduced glucose concentration and notably a neutrophilic pleocytosis. EDx studies reveal features of active axonal degeneration. The polyradiculoneuropathy may improve with antiviral therapy.

HIV-Related Multiple Mononeuropathies

Multiple mononeuropathies can also develop in patients with HIV infection, usually in the context of AIDS. Weakness, numbness, paresthesias, and pain occur in the distribution of affected nerves (Pattern 3, Table 438-2). Nerve biopsies can reveal axonal degeneration with necrotizing vasculitis or perivascular inflammation. Glucocorticoid treatment is indicated for vasculitis directly due to HIV infection.

HIV-Related Sensory Neuronopathy/Ganglionopathy

Dorsal root ganglionitis is a very rare complication of HIV infection, and neuronopathy can be the presenting manifestation. Patients develop sensory ataxia similar to idiopathic sensory neuronopathy/ganglionopathy (Pattern 9, Table 438-2). NCS reveal reduced amplitudes or absence of sensory nerve action potentials (SNAPs).

HERPES VARICELLA-ZOSTER VIRUS

Peripheral neuropathy from herpes varicella-zoster (HVZ) infection results from reactivation of latent virus or from a primary infection **(Chap. 188)**. Two-thirds of infections in adults are characterized by dermal zoster in which severe pain and paresthesias develop in a dermatomal region followed within a week or two by a vesicular rash in the same distribution (Pattern 3, Table 438-2). Weakness in muscles innervated by roots corresponding to the dermatomal distribution of skin lesions occurs in 5–30% of patients. Approximately 25% of affected patients have continued pain (postherpetic neuralgia [PHN]). A large clinical trial demonstrated that vaccination against zoster reduces the incidence of HVZ among vaccine recipients by 51% and reduces the incidence of PHN by 67%. Treatment of PHN is symptomatic (Table 438-6).

CYTOMEGALOVIRUS

CMV can cause an acute lumbosacral polyradiculopathy and multiple mononeuropathies in patients with HIV infection and in other immune deficiency conditions (Pattern 4, Table 438-2) **(Chap. 190)**.

EPSTEIN-BARR VIRUS

EBV infection has been associated with GBS, cranial neuropathies, mononeuropathy multiplex, brachial plexopathy, lumbosacral radiculoplexopathy, and sensory neuronopathies (Patterns 1, 3, 4, and 9, Table 438-2) **(Chap. 189)**.

HEPATITIS VIRUSES

Hepatitis B and C can cause multiple mononeuropathies related to vasculitis, AIDP, or CIDP (Patterns 1 and 3, Table 438-2) **(Chap. 334)**.

NEUROPATHIES ASSOCIATED WITH MALIGNANCY

Patients with malignancy can develop neuropathies due to (1) a direct effect of the cancer by invasion or compression of the nerves, (2) remote or paraneoplastic effect, (3) a toxic effect of treatment, or (4) as a consequence of immune compromise caused by immunosuppressive medications. The most common associated malignancy is lung cancer, but neuropathies also complicate carcinoma of the breast, ovaries, stomach, colon, rectum, and other organs, including the lymphoproliferative system.

PARANEOPLASTIC SENSORY NEURONOPATHY/GANGLIONOPATHY

Paraneoplastic encephalomyelitis/sensory neuronopathy (PEM/SN) usually complicates small-cell lung carcinoma **(Chap. 90)**. Patients usually present with numbness and paresthesias in the distal extremities that are often asymmetric. The onset can be acute or insidiously progressive. Prominent loss of proprioception leads to sensory ataxia (Pattern 9; Table 438-2). Weakness can be present, usually secondary to an associated myelitis, motor neuronopathy, or concurrent LEMS. Many patients also develop confusion, memory loss, depression, hallucinations or seizures, or cerebellar ataxia. Polyclonal antineuronal antibodies (IgG) directed against a 35- to 40-kDa protein or complex of proteins, the so-called Hu antigen, are found in the sera or CSF in the majority of patients with paraneoplastic PEM/SN. CSF may be normal or may demonstrate mild lymphocytic pleocytosis and elevated protein. PEM/SN is probably the result of antigenic similarity between proteins expressed in the tumor cells and neuronal cells, leading to an immune response directed against both cell types. Treatment of the underlying cancer generally does not affect the course of PEM/SN. However, occasional patients may improve following treatment of the tumor. Unfortunately, plasmapheresis, intravenous immunoglobulin, and immunosuppressive agents have not shown benefit.

NEUROPATHY SECONDARY TO TUMOR INFILTRATION

Malignant cells, in particular leukemia and lymphoma, can infiltrate cranial and peripheral nerves, leading to mononeuropathy, mononeuropathy multiplex, polyradiculopathy, plexopathy, or even a generalized symmetric distal or proximal and distal polyneuropathy (Patterns 1, 2, 3, and 4; Table 438-2). Neuropathy related to tumor infiltration is often painful; it can be the presenting manifestation of the cancer or the heralding symptom of a relapse. The neuropathy may improve with treatment of the underlying leukemia or lymphoma or with glucocorticoids.

NEUROPATHY AS A COMPLICATION OF BONE MARROW TRANSPLANTATION

Neuropathies may develop in patients who undergo bone marrow transplantation (BMT) because of the toxic effects of chemotherapy, radiation, infection, or an autoimmune response directed against the peripheral nerves. Peripheral neuropathy in BMT is often associated with graft-versus-host disease (GVHD). Chronic GVHD shares many features with a variety of autoimmune disorders, and it is possible that an immune-mediated response directed against peripheral nerves is responsible. Patients with chronic GVHD may develop cranial neuropathies, sensorimotor polyneuropathies, multiple mononeuropathies, and severe generalized peripheral neuropathies resembling AIDP or CIDP (Patterns 1, 2, and 3; Table 438-2). The neuropathy may improve by increasing the intensity of immunosuppressive or immunomodulating therapy and resolution of the GVHD.

LYMPHOMA

Lymphomas may cause neuropathy by infiltration or direct compression of nerves or by a paraneoplastic process. The neuropathy can be purely sensory or motor, but most commonly is sensorimotor. The pattern of involvement may be symmetric, asymmetric, or multifocal, and the course may be acute, gradually progressive, or relapsing and remitting (Patterns 1, 2, and 3; Table 438-2). EDx can be compatible with either an axonal or demyelinating process. CSF may reveal lymphocytic pleocytosis and an elevated protein. Nerve biopsy may demonstrate endoneurial inflammatory cells in both the infiltrative and the paraneoplastic etiologies. A monoclonal population of cells favors lymphomatous invasion. The neuropathy may respond to treatment of the underlying lymphoma or immunomodulating therapies.

MULTIPLE MYELOMA

MM usually presents in the fifth to seventh decade of life with fatigue, bone pain, anemia, and hypercalcemia **(Chap. 107)**. Clinical and EDx features of neuropathy occur in as many as 40% of patients. The most common pattern is that of a distal, axonal, sensory, or sensorimotor polyneuropathy (Pattern 2; Table 438-2). Less frequently, a chronic demyelinating polyradiculoneuropathy may develop (Pattern 1; Table 438-2) (see POEMS, **Chap. 439**). MM can be complicated by amyloid polyneuropathy and should be considered in patients with painful paresthesias, loss of pinprick and temperature discrimination, and autonomic dysfunction (suggestive of a small-fiber neuropathy) and CTS. Expanding plasmacytomas can compress cranial nerves and spinal roots as

well. A monoclonal protein, usually composed of γ or μ heavy chains or κ light chains, may be identified in the serum or urine. EDx usually shows reduced amplitudes with normal or only mildly abnormal distal latencies and conduction velocities. A superimposed median neuropathy at the wrist is common. Abdominal fat pad, rectal, or sural nerve biopsy can be performed to look for amyloid deposition. Unfortunately, the treatment of the underlying MM does not usually affect the course of the neuropathy.

NEUROPATHIES ASSOCIATED WITH MONOCLONAL GAMMOPATHY OF UNCERTAIN SIGNIFICANCE

(See Chap. 439)

Toxic Neuropathies Secondary to Chemotherapy

Many of the commonly used chemotherapy agents can cause a toxic neuropathy (Table 438-7). The mechanisms by which these agents cause toxic neuropathies vary, as does the specific type of neuropathy produced. The risk of developing a toxic neuropathy or more severe neuropathy appears to be greater in patients with a preexisting neuropathy (e.g., CMT disease, diabetic neuropathy) and those who also take other potentially neurotoxic drugs (e.g., nitrofurantoin, isoniazid, disulfiram, pyridoxine). Chemotherapeutic agents usually cause a sensory greater than motor length-dependent axonal neuropathy or neuronopathy/ganglionopathy (Patterns 2 and 9; Table 438-2).

TABLE 438-7

Toxic Neuropathies Secondary to Chemotherapy

Drug	Mechanism of Neurotoxicity	Clinical Features	Nerve Histopathology	EMG/NCS
Vinca alkaloids (vincristine, vinblastine, vindesine, vinorelbine)	Interfere with axonal microtubule assembly; impairs axonal transport	Symmetric, S-M, large-/small-fiber PN; autonomic symptoms common; infrequent cranial neuropathies	Axonal degeneration of myelinated and unmyelinated fibers; regenerating clusters, minimal segmental demyelination	Axonal sensorimotor PN; distal denervation on EMG; abnormal QST, particularly vibratory perception
Cisplatin	Preferential damage to dorsal root ganglia: ? binds to and cross-links DNA ? inhibits protein synthesis ? impairs axonal transport	Predominant large-fiber sensory neuronopathy; sensory ataxia	Loss of large > small myelinated and unmyelinated fibers; axonal degeneration with small clusters of regenerating fibers; secondary segmental demyelination	Low-amplitude or unobtainable SNAPs with normal CMAPs and EMG; abnormal QST, particularly vibratory perception
Taxanes (paclitaxel, docetaxel)	Promotes axonal microtubule assembly; interferes with axonal transport	Symmetric, predominantly sensory PN; large-fiber modalities affected more than small-fiber	Loss of large > small myelinated and unmyelinated fibers; axonal degeneration with small clusters of regenerating fibers; secondary segmental demyelination	Axonal sensorimotor PN; distal denervation on EMG; abnormal QST, particularly vibratory perception
Suramin				
Axonal PN	Unknown; ? inhibition of neurotrophic growth factor binding; ? neuronal lysosomal storage	Symmetric, length-dependent, sensory-predominant PN	None described	Abnormalities consistent with an axonal S-M PN
Demyelinating PN	Unknown; ? immunomodulating effects	Subacute, S-M PN with diffuse proximal and distal weakness; areflexia; increased CSF protein	Loss of large and small myelinated fibers with primary demyelination and secondary axonal degeneration; occasional epi- and endoneurial inflammatory cell infiltrates	Features suggestive of an acquired demyelinating sensorimotor PN (e.g., slow CVs, prolonged distal latencies and F-wave latencies, conduction block, temporal dispersion)
Cytarabine (ARA-C)	Unknown; ? selective Schwann cell toxicity; ? immunomodulating effects	GBS-like syndrome; pure sensory neuropathy; brachial plexopathy	Loss of myelinated nerve fibers; axonal degeneration; segmental demyelination; no inflammation	Axonal, demyelinating, or mixed S-M PN; denervation on EMG
Etoposide (VP-16)	Unknown; ? selective dorsal root ganglia toxicity	Length-dependent, sensory-predominant PN; autonomic neuropathy	None described	Abnormalities consistent with an axonal S-M PN
Bortezomib (Velcade)	Unknown	Length-dependent, sensory, predominantly small-fiber PN	Not reported	Abnormalities consistent with an axonal sensory neuropathy with early small-fiber involvement (abnormal autonomic studies)

Abbreviations: CMAP, compound motor action potential; CSF, cerebrospinal fluid; CVs, conduction velocities; EMG, electromyography; GBS, Guillain-Barré syndrome; NCS, nerve conduction studies; PN, polyneuropathy; QST, quantitative sensory testing; S-M, sensorimotor; SNAP, sensory nerve action potential.

Source: Modified from AA Amato, J Russell: *Neuromuscular Disorders,* 2nd ed. New York, McGraw-Hill, 2016, Table 19-3, p. 439.

OTHER TOXIC NEUROPATHIES

Neuropathies can develop as complications of toxic effects of various drugs and other environmental exposures **(Table 438-8)**. The more common neuropathies associated with these agents are discussed here.

TABLE 438-8

Toxic Neuropathies

Drug	Mechanism of Neurotoxicity	Clinical Features	Nerve Histopathology	EMG/NCS
Misonidazole	Unknown	Painful paresthesias and loss of large- and small-fiber sensory modalities and sometimes distal weakness in length-dependent pattern	Axonal degeneration of large myelinated fibers; axonal swellings; segmental demyelination	Low-amplitude or unobtainable SNAPs with normal or only slightly reduced CMAPs amplitudes
Metronidazole	Unknown	Painful paresthesias and loss of large- and small-fiber sensory modalities and sometimes distal weakness in length-dependent pattern	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal CMAPs
Chloroquine and hydroxychloroquine	Amphiphilic properties may lead to drug-lipid complexes that are indigestible and result in accumulation of autophagic vacuoles	Loss of large- and small-fiber sensory modalities and distal weakness in length-dependent pattern; superimposed myopathy may lead to proximal weakness	Axonal degeneration with autophagic vacuoles in nerves as well as muscle fibers	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes; distal denervation on EMG; irritability and myopathic-appearing MUAPs proximally in patients with superimposed toxic myopathy
Amiodarone	Amphiphilic properties may lead to drug-lipid complexes that are indigestible and result in accumulation of autophagic vacuoles	Paresthesias and pain with loss of large- and small-fiber sensory modalities and distal weakness in length-dependent pattern; superimposed myopathy may lead to proximal weakness	Axonal degeneration and segmental demyelination with myeloid inclusions in nerves and muscle fibers	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes; can also have prominent slowing of CVs; distal denervation on EMG; irritability and myopathic-appearing MUAPs proximally in patients with superimposed toxic myopathy
Colchicine	Inhibits polymerization of tubulin in microtubules and impairs axoplasmic flow	Numbness and paresthesias with loss of large-fiber modalities in a length-dependent fashion; superimposed myopathy may lead to proximal in addition to distal weakness	Nerve biopsy demonstrates axonal degeneration; muscle biopsy reveals fibers with vacuoles	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes; irritability and myopathic-appearing MUAPs proximally in patients with superimposed toxic myopathy
Podophyllin	Binds to microtubules and impairs axoplasmic flow	Sensory loss, tingling, muscle weakness, and diminished muscle stretch reflexes in length-dependent pattern; autonomic neuropathy	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Thalidomide	Unknown	Numbness, tingling, and burning pain and weakness in a length-dependent pattern	Axonal degeneration; autopsy studies reveal degeneration of dorsal root ganglia	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes

Drug	Mechanism of Neurotoxicity	Clinical Features	Nerve Histopathology	EMG/NCS
Disulfiram	Accumulation of neurofilaments and impaired axoplasmic flow	Numbness, tingling, and burning pain in a length-dependent pattern	Axonal degeneration with accumulation of neurofilaments in the axons	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Dapsone	Unknown	Distal weakness that may progress to proximal muscles; sensory loss	Axonal degeneration and segmental demyelination	Low-amplitude or unobtainable CMAPs with normal or reduced SNAP amplitudes
Leflunomide	Unknown	Paresthesias and numbness in a length-dependent pattern	Unknown	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Nitrofurantoin	Unknown	Numbness, painful paresthesias, and severe weakness that may resemble GBS	Axonal degeneration; autopsy studies reveal degeneration of dorsal root ganglia and anterior horn cells	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Pyridoxine (vitamin B_6)	Unknown	Dysesthesias and sensory ataxia; impaired large-fiber sensory modalities on examination	Marked loss of sensory axons and cell bodies in dorsal root ganglia	Reduced amplitudes or absent SNAPs
Isoniazid	Inhibits pyridoxal phosphokinase leading to pyridoxine deficiency	Dysesthesias and sensory ataxia; impaired large-fiber sensory modalities on examination	Marked loss of sensory axons and cell bodies in dorsal root ganglia and degeneration of the dorsal columns	Reduced amplitudes or absent SNAPs and, to a lesser extent, CMAPs
Ethambutol	Unknown	Numbness with loss of large-fiber modalities on examination	Axonal degeneration	Reduced amplitudes or absent SNAPs
Antinucleosides	Unknown	Dysesthesia and sensory ataxia; impaired large-fiber sensory modalities on examination	Axonal degeneration	Reduced amplitudes or absent SNAPs
Phenytoin	Unknown	Numbness with loss of large-fiber modalities on examination	Axonal degeneration and segmental demyelination	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Lithium	Unknown	Numbness with loss of large-fiber modalities on examination	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Acrylamide	Unknown; may be caused by impaired axonal transport	Numbness with loss of large-fiber modalities on examination; sensory ataxia; mild distal weakness	Degeneration of sensory axons in peripheral nerves and posterior columns, spinocerebellar tracts, mammillary bodies, optic tracts, and corticospinal tracts in the CNS	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Carbon disulfide	Unknown	Length-dependent numbness and tingling with mild distal weakness	Axonal swellings with accumulation of neurofilaments	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes

Drug	Mechanism of Neurotoxicity	Clinical Features	Nerve Histopathology	EMG/NCS
Ethylene oxide	Unknown; may act as alkylating agent and bind DNA	Length-dependent numbness and tingling; may have mild distal weakness	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Organophosphates	Bind and inhibit neuropathy target esterase	Early features are those of neuromuscular blockade with generalized weakness; later axonal sensorimotor PN ensues	Axonal degeneration along with degeneration of gracile fasciculus and corticospinal tracts	Early: repetitive firing of CMAPs and decrement with repetitive nerve stimulation; late: axonal sensorimotor PN
Hexacarbons	Unknown; may lead to covalent cross-linking between neurofilaments	Acute, severe sensorimotor PN that may resemble GBS	Axonal degeneration and giant axons swollen with neurofilaments	Features of a mixed axonal and/or demyelinating sensorimotor axonal PN—reduced amplitudes, prolonged distal latencies, conduction block, and slowing of CVs
Lead	Unknown; may interfere with mitochondria	Encephalopathy; motor neuropathy (often resembles radial neuropathy with wrist and finger drop); autonomic neuropathy; bluish-black discoloration of gums	Axonal degeneration of motor axons	Reduction of CMAP amplitudes with active denervation on EMG
Mercury	Unknown; may combine with sulfhydryl groups	Abdominal pain and nephrotic syndrome; encephalopathy; ataxia; paresthesias	Axonal degeneration; degeneration of dorsal root ganglia, calcarine, and cerebellar cortex	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Thallium	Unknown	Encephalopathy; painful sensory symptoms; mild loss of vibration; distal or generalized weakness may also develop; autonomic neuropathy; alopecia	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Arsenic	Unknown; may combine with sulfhydryl groups	Abdominal discomfort, burning pain, and paresthesias; generalized weakness; autonomic insufficiency; can resemble GBS	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes; may have demyelinating features: prolonged distal latencies and slowing of CVs
Gold	Unknown	Distal paresthesias and reduction of all sensory modalities	Axonal degeneration	Low-amplitude or unobtainable SNAPs

Abbreviations: CMAP, compound motor action potential; CVs, conduction velocities; EMG, electromyography; GBS, Guillain-Barré syndrome; MUAP, muscle action potential; NCS, nerve conduction studies; PN, polyneuropathy; S-M, sensorimotor; SNAP, sensory nerve action potential.

Source: Modified from AA Amato, J Russell: *Neuromuscular Disorders,* 2nd ed. New York, McGraw-Hill, 2016, Table 20-1, pp. 449–451.

CHLOROQUINE AND HYDROXYCHLOROQUINE

Chloroquine and hydroxychloroquine can cause a toxic myopathy characterized by slowly progressive, painless, proximal weakness and atrophy, which is worse in the legs than the arms. In addition, neuropathy can also develop with or without the myopathy leading to sensory loss and distal weakness. The "neuromyopathy" usually appears in patients taking 500 mg daily for a year or more but has been reported with doses as low as 200 mg/d. Serum CK levels

are usually elevated due to the superimposed myopathy. NCS reveal mild slowing of motor and sensory NCVs with a mild to moderate reduction in the amplitudes, although NCS may be normal in patients with only the myopathy. EMG demonstrates myopathic muscle action potentials (MUAPs), increased insertional activity in the form of positive sharp waves, fibrillation potentials, and occasionally myotonic potentials, particularly in the proximal muscles. Neurogenic MUAPs and reduced recruitment are found in more distal muscles. Nerve biopsy demonstrates autophagic vacuoles within Schwann cells. Vacuoles may also be evident in muscle biopsies. The pathogenic basis of the neuropathy is not known but may be related to the amphiphilic properties of the drug. These agents contain both hydrophobic and hydrophilic regions that allow them to interact with the anionic phospholipids of cell membranes and organelles. The drug-lipid complexes may be resistant to digestion by lysosomal enzymes, leading to the formation of autophagic vacuoles filled with myeloid debris that may in turn cause degeneration of nerves and muscle fibers. The signs and symptoms of the neuropathy and myopathy are usually reversible following discontinuation of medication.

AMIODARONE

Amiodarone can cause a neuromyopathy similar to chloroquine and hydroxychloroquine. The neuromyopathy typically appears after patients have taken the medication for 2–3 years. Nerve biopsy demonstrates a combination of segmental demyelination and axonal loss. Electron microscopy reveals lamellar or dense inclusions in Schwann cells, pericytes, and endothelial cells. The inclusions in muscle and nerve biopsies have persisted as long as 2 years following discontinuation of the medication.

COLCHICINE

Colchicine can also cause a neuromyopathy. Patients usually present with proximal weakness and numbness and tingling in the distal extremities. EDx reveals features of an axonal polyneuropathy. Muscle biopsy reveals a vacuolar myopathy, whereas sensory nerves demonstrate axonal degeneration. Colchicine inhibits the polymerization of tubulin into microtubules. The disruption of the microtubules probably leads to defective intracellular movement of important proteins, nutrients, and waste products in muscle and nerves.

THALIDOMIDE

Thalidomide is an immunomodulating agent used to treat MM, GVHD, leprosy, and other autoimmune disorders. Thalidomide is associated with severe teratogenic effects as well as peripheral neuropathy that can be dose-limiting. Patients develop numbness, painful tingling, and burning discomfort in the feet and hands and less commonly muscle weakness and atrophy. Even after stopping the drug for 4–6 years, as many as 50% patients continue to have significant symptoms. NCS demonstrate reduced amplitudes or complete absence of SNAPs, with preserved conduction velocities when obtainable. Motor NCS are usually normal. Nerve biopsy reveals a loss of large-diameter myelinated fibers and axonal degeneration. Degeneration of dorsal root ganglion cells has been reported at autopsy.

PYRIDOXINE (VITAMIN B_6) TOXICITY

Pyridoxine is an essential vitamin that serves as a coenzyme for transamination and decarboxylation. However, at high doses (116 mg/d), patients can develop a severe sensory neuropathy with dysesthesias and sensory ataxia. NCS reveal absent or markedly reduced SNAP amplitudes with relatively preserved CMAPs. Nerve biopsy reveals axonal loss of fiber at all diameters. Loss of dorsal root ganglion cells with subsequent degeneration of both the peripheral and central sensory tracts have been reported in animal models.

ISONIAZID

One of the most common side effects of isoniazid (INH) is peripheral neuropathy. Standard doses of INH (3–5 mg/kg per day) are associated with a 2% incidence of neuropathy, whereas neuropathy develops in at least 17% of patients taking in excess of 6 mg/kg per d. The elderly, malnourished, and “slow acetylators” are at increased risk for developing the neuropathy. INH inhibits pyridoxal phosphokinase, resulting in pyridoxine deficiency and the neuropathy. Prophylactic administration of pyridoxine 100 mg/d can prevent the neuropathy from developing.

ANTIRETROVIRAL AGENTS

The nucleoside analogues zalcitabine (dideoxycytidine or ddC), didanosine (dideoxyinosine or ddI), stavudine (d4T), lamivudine (3TC), and antiretroviral nucleoside reverse transcriptase inhibitor (NRTI) are used to treat HIV infection. One of the major dose-limiting side effects of these medications is a predominantly sensory, length-dependent, symmetrically painful neuropathy (Pattern 2; Table 438-2). Zalcitabine (ddC) is the most extensively studied of the nucleoside analogues, and at doses >0.18 mg/kg per d, it is associated with a subacute onset of severe burning and lancinating pains in the feet and hands. NCS reveal decreased amplitudes of the SNAPs with normal motor studies. The nucleoside analogues inhibit mitochondrial DNA polymerase, which is the suspected pathogenic basis for the neuropathy. Because of a “coasting effect,” patients can continue to worsen even 2–3 weeks after stopping the medication. Following dose reduction, improvement in the neuropathy is seen in most patients after several months (mean time about 10 weeks).

HEXACARBONS (*n*-HEXANE, METHYL *n*-BUTYL KETONE)/GLUE SNIFFER’S NEUROPATHY

n-Hexane and methyl *n*-butyl ketone are water-insoluble industrial organic solvents that are also present in some glues. Exposure through inhalation, accidentally or intentionally (glue sniffing), or through skin absorption can lead to a profound subacute sensory and motor polyneuropathy (Pattern 2; Table 438-2). NCS demonstrate decreased amplitudes of the SNAPs and CMAPs with slightly slow CVs. Nerve biopsy reveals a loss of myelinated fibers and giant axons that are filled with 10-nm neurofilaments. Hexacarbon exposure leads to covalent cross-linking between axonal neurofilaments that result in their aggregation, impaired axonal transport, swelling of the axons, and eventual axonal degeneration.

LEAD

Lead neuropathy is uncommon, but it can be seen in children who accidentally ingest lead-based paints in older buildings and in industrial workers exposed to lead-containing products. The most common presentation of lead poisoning is an encephalopathy; however, symptoms and signs of a primarily motor neuropathy can also occur. The neuropathy is characterized by an insidious and progressive onset of weakness usually beginning in the arms, in particular involving the wrist and finger extensors, resembling a radial neuropathy. Sensation is generally preserved; however, the autonomic nervous system can be affected (Patterns 2, 3, and 10; Table 438-2). Laboratory investigation can reveal a microcytic hypochromic anemia with basophilic stippling of erythrocytes, an elevated serum lead level, and an elevated serum coproporphyrin level. A 24-h urine collection demonstrates elevated levels of lead excretion. The NCS may reveal reduced CMAP amplitudes, while the SNAPs are typically normal. The pathogenic basis may be related to abnormal porphyrin metabolism. The most important principle of management is to remove the source of the exposure. Chelation therapy with calcium disodium ethylene-diaminetetraacetic acid (EDTA), British anti-Lewisite (BAL), and penicillamine also demonstrates variable efficacy.

MERCURY

Mercury toxicity may occur as a result of exposure to either organic or inorganic mercurials. Mercury poisoning presents with paresthesias in hands and feet that progress proximally and may involve the face and tongue. Motor weakness can also develop. CNS symptoms often overshadow the neuropathy. EDx shows features of a primarily axonal sensorimotor polyneuropathy. The primary site of neuromuscular pathology appears to be the dorsal root ganglia. The mainstay of treatment is removing the source of exposure.

THALLIUM

Thallium can exist in a monovalent or trivalent form and is primarily used as a rodenticide. The toxic neuropathy usually manifests as burning paresthesias of the feet, abdominal pain, and vomiting. Increased thirst, sleep disturbances, and psychotic behavior may be noted. Within the first week, patients develop pigmentation of the hair, an acne-like rash in the malar area of the face, and hyperreflexia. By the second and third weeks, autonomic instability with labile heart rate and blood pressure may be seen. Hyporeflexia and alopecia also occur but may not be evident until the third or fourth week following exposure. With severe intoxication, proximal weakness and involvement of the cranial nerves can occur. Some patients require mechanical ventilation due to respiratory muscle involvement. The lethal dose of thallium is variable, ranging from 8 to 15 mg/kg body weight. Death can result in <48 h following a particularly large dose. NCS demonstrate features of a primarily axonal sensorimotor polyneuropathy. With acute intoxication, potassium ferric ferrocyanide II may be effective in preventing absorption of thallium from the gut. However, there may be no benefit once thallium has been absorbed. Unfortunately, chelating agents are not very efficacious. Adequate diuresis is essential to help eliminate thallium from the body without increasing tissue availability from the serum.

ARSENIC

Arsenic is another heavy metal that can cause a toxic sensorimotor polyneuropathy. The neuropathy manifests 5–10 days after ingestion of arsenic and progresses for several weeks, sometimes mimicking GBS. The presenting symptoms are typically an abrupt onset of abdominal discomfort, nausea, vomiting, pain, and diarrhea followed within several days by burning pain in the feet and hands. Examination of the skin can be helpful in the diagnosis as the loss of the superficial epidermal layer results in patchy regions of increased or decreased pigmentation on the skin several weeks after an acute exposure or with chronic low levels of ingestion. Mee's lines, which are transverse lines at the base of the fingernails and toenails, do not become evident until 1 or 2 months after the exposure. Multiple Mee's lines may be seen in patients with long fingernails who have had chronic exposure to arsenic. Mee's lines are not specific for arsenic toxicity as they can also be seen following thallium poisoning. Because arsenic is cleared from blood rapidly, the serum concentration of arsenic is not diagnostically helpful. However, arsenic levels are increased in the urine, hair, and fingernails of patients exposed to arsenic. Anemia with stippling of erythrocytes is common, and occasionally pancytopenia and aplastic anemia can develop. Increased CSF protein levels without pleocytosis can be seen; this can lead to misdiagnosis as GBS. NCS are usually suggestive of an axonal sensorimotor polyneuropathy; however, demyelinating features can be present. Chelation therapy with BAL has yielded inconsistent results; therefore, it is not generally recommended.

NUTRITIONAL NEUROPATHIES

COBALAMIN (VITAMIN B_{12})

Pernicious anemia is the most common cause of cobalamin deficiency. Other causes include dietary avoidance (vegetarians), gastrectomy, gastric bypass surgery, inflammatory bowel disease, pancreatic insufficiency, bacterial overgrowth, and possibly histamine-2 blockers and proton pump inhibitors. An underappreciated cause of cobalamin deficiency is food-cobalamin malabsorption. This typically occurs in older individuals and results from an inability to

adequately absorb cobalamin in food protein. No apparent cause of deficiency is identified in a significant number of patients with cobalamin deficiency. The use of nitrous oxide as an anesthetic agent or as a recreational drug can produce acute cobalamin deficiency neuropathy and subacute combined degeneration.

Complaints of numb hands typically appear before lower extremity paresthesias are noted. A preferential large-fiber sensory loss affecting proprioception and vibration with sparing of small-fiber modalities is present; an unsteady gait reflects sensory ataxia. These features, coupled with diffuse hyperreflexia and absent Achilles reflexes, should always focus attention on the possibility of cobalamin deficiency (Patterns 2 and 6; Table 438-2). Optic atrophy and, in severe cases, behavioral changes ranging from mild irritability and forgetfulness to severe dementia and frank psychosis may appear. The full clinical picture of subacute combined degeneration is uncommon. CNS manifestations, especially pyramidal tract signs, may be missing, and in fact some patients may only exhibit symptoms of peripheral neuropathy.

EDx shows an axonal sensorimotor neuropathy. CNS involvement produces abnormal somatosensory and visual evoked potential latencies. The diagnosis is confirmed by finding reduced serum cobalamin levels. In up to 40% of patients, anemia and macrocytosis are lacking. Serum methylmalonic acid and homocysteine, the metabolites that accumulate when cobalamin-dependent reactions are blocked, are elevated. Antibodies to intrinsic factor are present in ~60%, and antiparietal cell antibodies in about 90%, of individuals with pernicious anemia.

Cobalamin deficiency can be treated with various regimens of cobalamin. One typical regimen consists of 1000 µg cyanocobalamin IM weekly for 1 month and monthly thereafter. Patients with food cobalamin malabsorption can absorb free cobalamin and therefore can be treated with oral cobalamin supplementation. An oral cobalamin dose of 1000 µg per day should be sufficient. Treatment for cobalamin deficiency usually does not completely reverse the clinical manifestations, and at least 50% of patients exhibit some permanent neurologic deficit.

THIAMINE DEFICIENCY

Thiamine (vitamin B_1) deficiency is an uncommon cause of peripheral neuropathy in developed countries. It is now most often seen as a consequence of chronic alcohol abuse, recurrent vomiting, total parenteral nutrition, and bariatric surgery. Thiamine deficiency polyneuropathy can occur in normal, healthy young adults who do not abuse alcohol but who engage in inappropriately restrictive diets. Thiamine is water-soluble. It is present in most animal and plant tissues, but the greatest sources are unrefined cereal grains, wheat germ, yeast, soybean flour, and pork. Beriberi means “I can’t, I can’t” in Singhalese, the language of natives of what was once part of the Dutch East Indies (now Sri Lanka). *Dry beriberi* refers to neuropathic symptoms. The term *wet beriberi* is used when cardiac manifestations predominate (in reference to edema). Beriberi was relatively uncommon until the late 1800s when it became widespread among people for whom rice was a dietary mainstay. This epidemic was due to a new technique of processing rice that removed the germ from the rice shaft, rendering the so-called polished rice deficient in thiamine and other essential nutrients.

Symptoms of neuropathy follow prolonged deficiency. These begin with mild sensory loss and/or burning dysesthesias in the toes and feet and aching and cramping in the lower legs. Pain may be the predominant symptom. With progression, patients develop features of a nonspecific generalized polyneuropathy, with distal sensory loss in the feet and hands.

Blood and urine assays for thiamine are not reliable for diagnosis of deficiency. Erythrocyte transketolase activity and the percentage increase in activity (in vitro) following the addition of thiamine pyrophosphate (TPP) may be more accurate and reliable. EDx shows nonspecific findings of an axonal sensorimotor polyneuropathy. When a diagnosis of thiamine deficiency is made or suspected, thiamine replacement should be provided until proper nutrition is restored. Thiamine is usually given intravenously or intramuscularly at a dose of 100 mg/d. Although cardiac manifestations show a striking response to thiamine replacement, neurologic improvement is usually more variable and less dramatic.

VITAMIN E DEFICIENCY

The term *vitamin E* is usually used for a-tocopherol, the most active of the four main types of vitamin E. Because vitamin E is present in animal fat, vegetable oils, and various grains, deficiency is usually due to factors other than insufficient intake. Vitamin E deficiency usually occurs secondary to lipid malabsorption or in uncommon disorders of vitamin E transport. One hereditary disorder is abetalipoproteinemia, a rare autosomal dominant disorder characterized by steatorrhea, pigmentary retinopathy, acanthocytosis, and progressive ataxia. Patients with cystic fibrosis may also have vitamin E deficiency secondary to steatorrhea. There are genetic forms of isolated vitamin E deficiency not associated with lipid malabsorption. Vitamin E deficiency may also occur as a consequence of various cholestatic and hepatobiliary disorders as well as short-bowel syndromes resulting from the surgical treatment of intestinal disorders.

Clinical features may not appear until many years after the onset of deficiency. The onset of symptoms tends to be insidious, and progression is slow. The main clinical features are spinocerebellar ataxia and polyneuropathy, thus resembling Friedreich’s ataxia or other spinocerebellar ataxias. Patients manifest progressive ataxia and signs of posterior column dysfunction, such as impaired joint position and vibratory sensation. Because of the polyneuropathy, there is hyporeflexia, but plantar responses may be extensor as a result of the spinal cord involvement (Patterns 2 and 6; Table 438-2). Other neurologic manifestations may include ophthalmoplegia, pigmented retinopathy, night blindness, dysarthria, pseudoathetosis, dystonia, and tremor. Vitamin E deficiency may present as an isolated polyneuropathy, but this is very rare. The yield of checking serum vitamin E levels in patients with isolated polyneuropathy is extremely low, and this test should not be part of routine practice.

Diagnosis is made by measuring a-tocopherol levels in the serum. EDx shows features of an axonal neuropathy. Treatment is replacement with oral vitamin E, but high doses are not needed. For patients with isolated vitamin E deficiency, treatment consists of 1500–6000 IU/d in divided doses.

VITAMIN B_6 DEFICIENCY

Vitamin B_6, or pyridoxine, can produce neuropathic manifestations from both deficiency and toxicity. Vitamin B_6 toxicity was discussed above. Vitamin B_6 deficiency is most commonly seen in patients treated with isoniazid or hydralazine. The polyneuropathy of vitamin B_6 is nonspecific, manifesting as a generalized axonal sensorimotor polyneuropathy. Vitamin B_6 deficiency can be detected by direct assay. Vitamin B_6 supplementation with 50–100 mg/d is suggested for patients being treated with isoniazid or hydralazine. This same dose is appropriate for replacement in cases of nutritional deficiency.

PELLAGRA (NIACIN DEFICIENCY)

Pellagra is produced by deficiency of niacin. Although pellagra may be seen in alcoholics, this disorder has essentially been eradicated in most Western countries by means of enriching bread with niacin. Nevertheless, pellagra continues to be a problem in a number of underdeveloped regions, particularly in Asia and Africa, where corn is the main source of carbohydrate. Neurologic manifestations are variable; abnormalities can develop in the brain and spinal cord as well as peripheral nerves. When peripheral nerves are involved, the neuropathy is usually mild and resembles beriberi. Treatment is with niacin 40–250 mg/d.

COPPER DEFICIENCY

A syndrome that has only recently been described is myeloneuropathy secondary to copper deficiency. Most patients present with lower limb paresthesias, weakness, spasticity, and gait difficulties (Pattern 6; Table 438-2). Large-fiber sensory function is impaired, reflexes are brisk, and plantar responses are extensor. In some cases, light touch and pinprick sensation are affected, and NCS indicate sensorimotor axonal polyneuropathy in addition to myelopathy.

Hematologic abnormalities are a known complication of copper deficiency; these can include microcytic anemia, neutropenia, and occasionally pancytopenia. Because copper is absorbed in the stomach and proximal jejunum, many cases of copper deficiency occur in the setting of prior gastric surgery. Excess zinc is an established cause of copper deficiency. Zinc upregulates enterocyte production of metallothionine, which results in decreased absorption of copper. Excessive dietary zinc supplements or denture cream containing zinc can produce this clinical picture. Other potential causes of copper deficiency include malnutrition, prematurity, total parenteral nutrition, and ingestion of copper-chelating agents.

Following oral or IV copper replacement, some patients show neurologic improvement, but this may take many months or not occur at all. Replacement consists of oral copper sulfate or gluconate 2 mg one to three times a day. If oral copper replacement is not effective, elemental copper in the copper sulfate or copper chloride forms can be given as 2 mg IV daily for 3–5 days, then weekly for 1–2 months until copper levels normalize. Thereafter, oral daily copper therapy can be resumed. In contrast to the neurologic manifestations, most of the hematologic indices normalize in response to copper replacement therapy.

NEUROPATHY ASSOCIATED WITH GASTRIC SURGERY

Polyneuropathy may occur following gastric surgery for ulcer, cancer, or weight reduction. This usually occurs in the context of rapid, significant weight loss and recurrent, protracted vomiting. The clinical picture is one of acute or subacute sensory loss and weakness. Neuropathy following weight loss surgery usually occurs in the first several months after surgery. Weight reduction surgical procedures include gastrojejunostomy, gastric stapling, vertical banded gastroplasty, and gastrectomy with Roux-en-Y anastomosis. The initial manifestations are usually numbness and paresthesias in the feet (Pattern 2; Table 438-2). In many cases, no specific nutritional deficiency factor is identified.

Management consists of parenteral vitamin supplementation, especially including thiamine. Improvement has been observed following supplementation, parenteral nutritional support, and reversal of the surgical bypass. The duration and severity of deficits before identification and treatment of neuropathy are important predictors of final outcome.

CRYPTOGENIC (IDIOPATHIC) SENSORY AND SENSORIMOTOR POLYNEUROPATHY

Cryptogenic (idiopathic) sensory and sensorimotor polyneuropathy (CSPN) is a diagnosis of exclusion, established after a careful medical, family, and social history; neurologic examination; and directed laboratory testing. Despite extensive evaluation, the cause of polyneuropathy in as many as 50% of all patients is idiopathic. CSPN should be considered a distinct diagnostic subset of peripheral neuropathy. The onset of CSPN is predominantly in the sixth and seventh decades. Patients complain of distal numbness, tingling, and often burning pain that invariably begins in the feet and may eventually involve the fingers and hands. Patients exhibit a distal sensory loss to pinprick, touch, and vibration in the toes and feet, and occasionally in the fingers (Pattern 2; Table 438-2). It is uncommon to see significant proprioception deficits, even though patients may complain of gait unsteadiness. However, tandem gait may be abnormal in a minority of cases. Neither subjective nor objective evidence of weakness is a prominent feature. Most patients have evidence of both large- and small-fiber loss on neurologic examination and EDx. Approximately 10% of patients have only evidence of small-fiber involvement. The ankle muscle stretch reflex is frequently absent, but in cases with predominantly small-fiber loss, this may be preserved. The EDx findings range from isolated

SNAP abnormalities (usually with loss of amplitude), to evidence for an axonal sensorimotor neuropathy, to a completely normal study (if primarily small fibers are involved). Therapy primarily involves the control of neuropathic pain (Table 438-6) if present. These drugs should not be used if the patient has only numbness and tingling but no pain.

Although no treatment is available that can reverse an idiopathic distal peripheral neuropathy, the prognosis is good. Progression often does not occur or is minimal, with sensory symptoms and signs progressing proximally up to the knees and elbows. The disorder does not lead to significant motor disability over time. The relatively benign course of this disorder should be explained to patients.

MONONEUROPATHIES/PLEXOPATHIES/RADICULOPATHIES

(Pattern 3; Table 438-2)

MEDIAN NEUROPATHY

CTS is a compression of the median nerve in the carpal tunnel at the wrist. The median nerve enters the hand through the carpal tunnel by coursing under the transverse carpal ligament. The symptoms of CTS consist of numbness and paresthesias variably in the thumb, index, middle, and half of the ring finger. At times, the paresthesias can include the entire hand and extend into the forearm or upper arm or can be isolated to one or two fingers. Pain is another common symptom and can be located in the hand and forearm and, at times, in the proximal arm. CTS is common and often misdiagnosed as thoracic outlet syndrome. The signs of CTS are decreased sensation in the median nerve distribution; reproduction of the sensation of tingling when a percussion hammer is tapped over the wrist (Tinel's sign) or the wrist is flexed for 30–60 s (Phalen's sign); and weakness of thumb opposition and abduction. EDx is extremely sensitive and shows slowing of sensory and, to a lesser extent, motor median potentials across the wrist. Ultrasound can show focal swelling of the median nerve at the wrist. Treatment options consist of avoidance of precipitating activities; control of underlying systemic-associated conditions if present; nonsteroidal anti-inflammatory medications; neutral (volar) position wrist splints, especially for night use; glucocorticoid/anesthetic injection into the carpal tunnel; and surgical decompression by dividing the transverse carpal ligament. The surgical option should be considered if there is a poor response to nonsurgical treatments; if there is thenar muscle atrophy and/or weakness; and if there are significant denervation potentials on EMG.

Other proximal median neuropathies are very uncommon and include the pronator teres syndrome and anterior interosseous neuropathy. These often occur as a partial form of brachial plexitis.

ULNAR NEUROPATHY AT THE ELBOW—"CUBITAL TUNNEL SYNDROME"

The ulnar nerve passes through the condylar groove between the medial epicondyle and the olecranon. Symptoms consist of paresthesias, tingling, and numbness in the medial hand and half of the fourth and the entire fifth fingers, pain at the elbow or forearm, and weakness. Signs consist of decreased sensation in an ulnar distribution, Tinel's sign at the elbow, and weakness and atrophy of ulnar-innervated hand muscles. The Froment sign indicates thumb adductor weakness and consists of flexion of the thumb at the interphalangeal joint when attempting to oppose the thumb against the lateral border of the second digit. EDx may show slowing of ulnar motor NCV across the elbow with prolonged ulnar sensory latencies. Ultrasound can show swelling of the ulnar nerve around the elbow as well. Treatment consists of avoiding aggravating factors, using elbow pads, and surgery to decompress the nerve in the cubital tunnel. Ulnar neuropathies can also rarely occur at the wrist in the ulnar (Guyon) canal or in the hand, usually after trauma.

RADIAL NEUROPATHY

The radial nerve winds around the proximal humerus in the spiral groove and proceeds down the lateral arm and enters the forearm, dividing into the posterior interosseous nerve and superficial nerve. The symptoms and signs consist of wrist drop; finger extension weakness; thumb abduction weakness; and sensory loss in the dorsal web between the thumb and index finger. Triceps and brachioradialis strength is often normal, and triceps reflex is often intact. Most cases of radial neuropathy are transient compressive (neuropraxic) injuries that recover spontaneously in 6–8 weeks. If there has been prolonged compression and severe axonal damage, it may take several months to recover. Treatment consists of cock-up wrist and finger splints, avoiding further compression, and physical therapy to avoid flexion contracture. If there is no improvement in 2–3 weeks, an EDx study is recommended to confirm the clinical diagnosis and determine the degree of severity.

LATERAL FEMORAL CUTANEOUS NEUROPATHY (MERALGIA PARESTHETICA)

The lateral femoral cutaneous nerve arises from the upper lumbar plexus (spinal levels L2/3), crosses through the inguinal ligament near its attachment to the iliac bone, and supplies sensation to the anterior lateral thigh. The neuropathy affecting this nerve is also known as meralgia paresthetica. Symptoms and signs consist of paresthesias, numbness, and occasionally pain in the lateral thigh. Symptoms are increased by standing or walking and are relieved by sitting. There is normal strength, and knee reflexes are intact. The diagnosis is clinical, and further tests usually are not performed. EDx is only needed to rule out lumbar plexopathy, radiculopathy, or femoral neuropathy. If the symptoms and signs are classic, EMG is not necessary. Symptoms often resolve spontaneously over weeks or months, but the patient may be left with permanent numbness. Treatment consists of weight loss and avoiding tight belts. Analgesics in the form of a lidocaine patch, nonsteroidal agents, and occasionally medications for neuropathic pain can be used (Table 438-6). Rarely, locally injecting the nerve with an anesthetic can be tried. There is no role for surgery.

FEMORAL NEUROPATHY

Femoral neuropathies can arise as complications of retroperitoneal hematoma, lithotomy positioning, hip arthroplasty or dislocation, iliac artery occlusion, femoral arterial procedures, infiltration by hematogenous malignancy, penetrating groin trauma, pelvic surgery including hysterectomy and renal transplantation, and diabetes (a partial form of lumbosacral diabetic plexopathy); some cases are idiopathic. Patients with femoral neuropathy have difficulty extending their knee and flexing the hip. Sensory symptoms occurring either on the anterior thigh and/or medial leg occur in only half of reported cases. A prominent painful component is the exception rather than the rule, may be delayed, and is often self-limited in nature. The quadriceps (patellar) reflex is diminished.

SCIATIC NEUROPATHY

Sciatic neuropathies commonly complicate hip arthroplasty, pelvic procedures in which patients are placed in a prolonged lithotomy position, trauma, hematomas, tumor infiltration, and vasculitis. In addition, many sciatic neuropathies are idiopathic. Weakness may involve all motions of the ankles and toes as well as flexion of the leg at the knee; abduction and extension of the thigh at the hip are spared. Sensory loss occurs in the entire foot and the distal lateral leg. The ankle jerk and on occasion the internal hamstring reflex are diminished or more typically absent on the affected side. The peroneal subdivision of the sciatic nerve is typically involved disproportionately to the tibial counterpart. Thus, patients may have only ankle dorsiflexion and eversion weakness with sparing of knee flexion, ankle inversion, and plantar flexion; these features can lead to misdiagnosis of a common peroneal neuropathy.

PERONEAL NEUROPATHY

The sciatic nerve divides at the distal femur into the tibial and peroneal nerve. The common peroneal nerve passes posterior and laterally around the fibular head, under the fibular tunnel. It then divides into the superficial peroneal nerve, which supplies the ankle evertor muscles and sensation over the anterolateral distal leg and dorsum of the foot, and the deep peroneal nerve, which supplies ankle dorsiflexors and toe extensor muscles and a small area of sensation dorsally in the area of the first and second toes.

Symptoms and signs consist of foot drop (ankle dorsiflexion, toe extension, and ankle eversion weakness) and variable sensory loss, which may involve the superficial and deep peroneal pattern. There is usually no pain. Onset may be on awakening in the morning. Peroneal neuropathy needs to be distinguished from L5 radiculopathy. In L5 radiculopathy, ankle invertors and evertors are weak and needle EMG reveals denervation. EDx can help localize the lesion. Peroneal motor conduction velocity shows slowing and amplitude drop across the fibular head. Management consists of rapid weight loss and avoiding leg crossing. Footdrop is treated with an ankle brace. A knee pad can be worn over the lateral knee to avoid further compression. Most cases spontaneously resolve over weeks or months.

RADICULOPATHIES

Radiculopathies are most often due to compression from degenerative joint disease and herniated disks, but there are a number of unusual etiologies (Table 438-9). Degenerative spine disease affects a number of different structures, which narrow the diameter of the neural foramen or canal of the spinal column and compromise nerve root integrity; **these are discussed in detail in Chap. 14**.

TABLE 438-9

Causes of Radiculopathy

- Herniated nucleus pulposus
- Degenerative joint disease
- Rheumatoid arthritis
- Trauma
- Vertebral body compression fracture
- Pott's disease
- Compression by extradural mass (e.g., meningioma, metastatic tumor, hematoma, abscess)
- Primary nerve tumor (e.g., neurofibroma, schwannoma, neurinoma)
- Carcinomatous meningitis
- Perineurial spread of tumor (e.g., prostate cancer)
- Acute inflammatory demyelinating polyradiculopathy
- Chronic inflammatory demyelinating polyradiculopathy
- Sarcoidosis
- Amyloidoma
- Diabetic radiculopathy
- Infection (Lyme disease, herpes zoster, HIV, cytomegalovirus, syphilis, schistosomiasis, strongyloides)
- Arachnoiditis (e.g., postsurgical)
- Radiation

PLEXOPATHIES

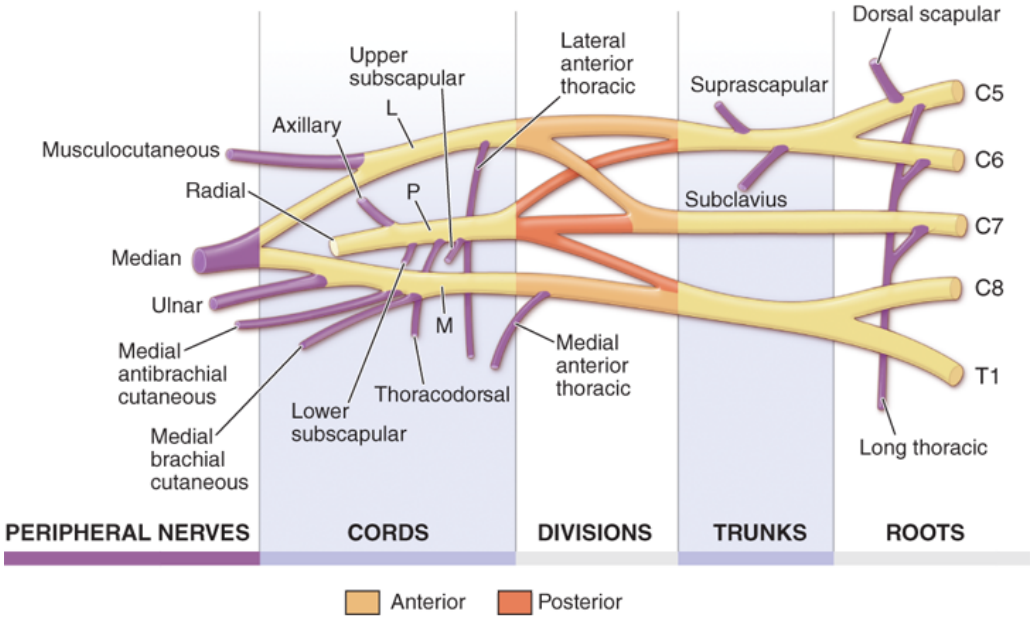
(Pattern 4; Table 438-2)

BRACHIAL PLEXUS

The brachial plexus is composed of three trunks (upper, middle, and lower), with two divisions (anterior and posterior) per trunk **(Fig. 438-2)**. Subsequently, the trunks divide into three cords (medial, lateral, and posterior), and from these arise the multiple terminal nerves innervating the arm. The anterior primary rami of C5 and C6 fuse to form the upper trunk; the anterior primary ramus of C7 continues as the middle trunk, while the anterior rami of C8 and T1 join to form the lower trunk. There are several disorders commonly associated with brachial plexopathy.

FIGURE 438-2

Brachial plexus anatomy. L, lateral; M, medial; P, posterior. *(From J Goodgold: Anatomical Correlates of Clinical Electromyography. Baltimore, Williams and Wilkins, 1974, p. 56, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Immune-Mediated Brachial Plexus Neuropathy

Immune-mediated brachial plexus neuropathy (IBPN) goes by various terms, including *acute brachial plexitis, neuralgic amyotrophy*, and *Parsonage-Turner syndrome*. IBPN usually presents with an acute onset of severe pain in the shoulder region. The intense pain usually lasts several days to a few weeks, but a dull ache can persist. Individuals who are affected may not appreciate weakness of the arm early in the course because the pain limits movement. However, as the pain dissipates, weakness and often sensory loss are appreciated. Attacks can occasionally recur.

Clinical findings are dependent on the distribution of involvement (e.g., specific trunk, divisions, cords, or terminal nerves). The most common pattern of IBPN involves the upper trunk or a single or multiple mononeuropathies primarily involving the suprascapular, long thoracic, or axillary nerves. Additionally, the phrenic and anterior interosseous nerves may be concomitantly affected. Any of these nerves may also be affected in isolation. EDx is useful to confirm and localize the site(s) of involvement. Empirical treatment of severe pain with glucocorticoids is often used in the acute period.

Brachial Plexopathies Associated with Neoplasms

Neoplasms involving the brachial plexus may be primary nerve tumors, local cancers expanding into the plexus (e.g., Pancoast lung tumor or lymphoma), and metastatic tumors. Primary brachial plexus tumors are less common than the secondary tumors and include schwannomas, neurinomas, and neurofibromas. Secondary tumors affecting the brachial plexus are more common and are always malignant. These may arise from local tumors, expanding into the plexus. For example, a Pancoast tumor of the upper lobe of the lung may invade or compress the lower trunk, whereas a primary lymphoma arising from the cervical or axillary lymph nodes may also infiltrate the plexus. Pancoast tumors typically present as an insidious onset of pain in the upper arm, sensory disturbance in the medial aspect of the forearm and hand, and weakness and atrophy of the intrinsic hand muscles along with an ipsilateral Horner's syndrome. Chest computed tomography (CT) scans or magnetic resonance imaging (MRI) can demonstrate extension of the tumor into the plexus. Metastatic involvement of the brachial plexus may occur with spread of breast cancer into the axillary lymph nodes and local spread into the nearby nerves.

Perioperative Plexopathies (Median Sternotomy)

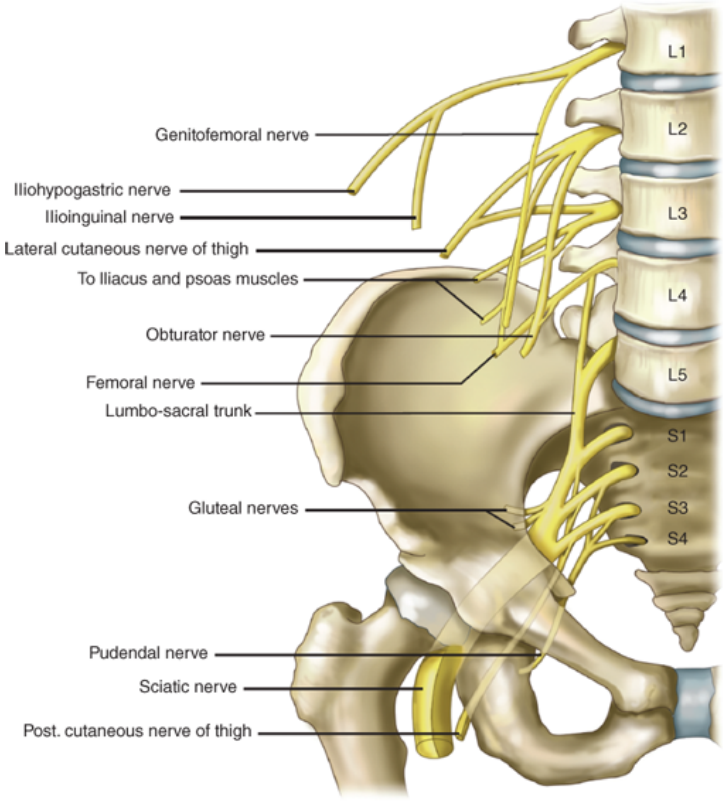
The most common surgical procedures associated with brachial plexopathy as a complication are those that involve median sternotomies (e.g., open-heart surgeries and thoracotomies). Brachial plexopathies occur in as many as 5% of patients following a median sternotomy and typically affect the lower trunk. Thus, individuals manifest with sensory disturbance affecting the medial aspect of forearm and hand along with weakness of the intrinsic hand muscles. The mechanism is related to the stretch of the lower trunk, so most individuals who are affected recover within a few months.

Lumbosacral Plexus

The lumbar plexus arises from the ventral primary rami of the first to the fourth lumbar spinal nerves (Fig. 438-3). These nerves pass downward and laterally from the vertebral column within the psoas major muscle. The femoral nerve derives from the dorsal branches of the second to the fourth lumbar ventral rami. The obturator nerve arises from the ventral branches of the same lumbar rami. The lumbar plexus communicates with the sacral plexus by the lumbosacral trunk, which contains some fibers from the fourth and all of the fibers from the fifth lumbar ventral rami (Fig. 438-4).

FIGURE 438-3

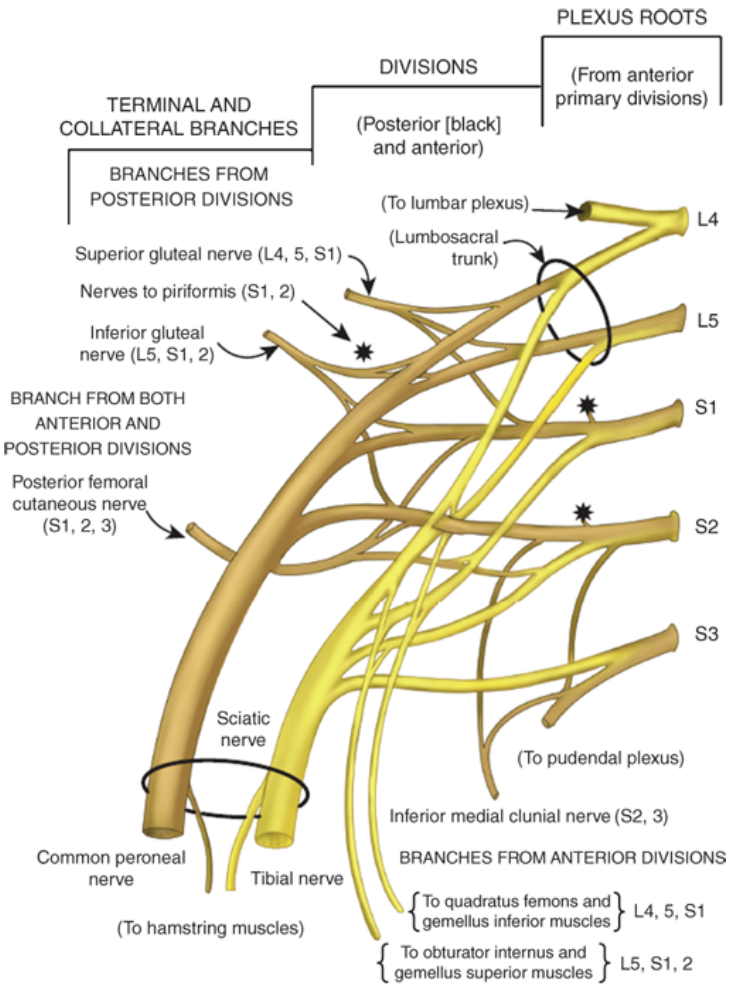
Lumbosacral plexus. *(From AA Amato, JA Russell (eds): Neuromuscular Disorders, 2nd ed. McGraw-Hill Education, 2016, Table 24-3, p. 541, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 438-4

Lumbosacral trunk sacral plexus, and sciatic nerve. *(From AA Amato, JA Russell (eds): Neuromuscular Disorders, 2nd ed. McGraw-Hill Education, 2016, Table 24-4, p. 542, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The sacral plexus is the part of the lumbosacral plexus that is formed by the union of the lumbosacral trunk with the ventral rami of the first to fourth sacral nerves. The plexus lies on the posterior and posterolateral wall of the pelvis with its components converging toward the sciatic notch. The lateral trunk of the sciatic nerve (which forms the common peroneal nerve) arises from the union of the dorsal branches of the lumbosacral trunk (L4, L5) and the dorsal branches of the S1 and S2 spinal nerve ventral rami. The medial trunk of the sciatic nerve (which forms the tibial nerve) derives from the ventral branches of the same ventral rami (L4-S2).

LUMBOSACRAL PLEXOPATHIES

Plexopathies are typically recognized when motor, sensory, and if applicable, reflex deficits occur in multiple nerve and segmental distributions confined to one extremity. If localization within the lumbosacral plexus can be accomplished, designation as a lumbar plexopathy, a sacral plexopathy, a lumbosacral trunk lesion, or a pan-plexopathy is the best localization that can be expected. Although lumbar plexopathies may be bilateral, usually occurring in a stepwise and chronologically dissociated manner, sacral plexopathies are more likely to behave in this manner due to their closer anatomic proximity. The differential diagnosis of plexopathy includes disorders of the conus medullaris and cauda equina (polyradiculopathy). If there is a paucity of pain and sensory involvement, motor neuron disease should be considered as well.

The causes of lumbosacral plexopathies are listed in Table 438-10. Diabetic radiculopathy (discussed above) is a fairly common cause of painful leg weakness. Lumbosacral plexopathies are a well-recognized complication of retroperitoneal hemorrhage. Various primary and metastatic malignancies can affect the lumbosacral plexus as well; these include carcinoma of the cervix, endometrium, and ovary; osteosarcoma; testicular cancer; MM; lymphoma; acute myelogenous leukemia; colon cancer; squamous cell carcinoma of the rectum; adenocarcinoma of unknown origin; and intraneural spread of prostate cancer.

TABLE 438-10

Lumbosacral Plexopathies: Etiologies

- Retroperitoneal hematoma
- Psoas abscess
- Malignant neoplasm
- Benign neoplasm
- Radiation
- Amyloid
- Diabetic radiculoplexus neuropathy
- Idiopathic radiculoplexus neuropathy
- Sarcoidosis
- Aortic occlusion/surgery
- Lithotomy positioning
- Hip arthroplasty
- Pelvic fracture
- Obstetric injury

RECURRENT NEOPLASTIC DISEASE OR RADIATION-INDUCED PLEXOPATHY

The treatment for various malignancies is often radiation therapy, the field of which may include parts of the brachial plexus. It can be difficult in such situations to determine if a new brachial or lumbosacral plexopathy is related to tumor within the plexus or from radiation-induced nerve damage. Radiation can be associated with microvascular abnormalities and fibrosis of surrounding tissues, which can damage the axons and the Schwann cells. Radiation-induced plexopathy can develop months or years following therapy and is dose dependent.

Tumor invasion is usually painful and more commonly affects the lower trunk, whereas radiation injury is often painless and affects the upper trunk. Imaging studies such as MRI and CT scans are useful but can be misleading, especially when there is small microscopic invasion of the plexus. EMG can be informative if myokymic discharges are appreciated, as this finding strongly suggests radiation-induced damage.

EVALUATION AND TREATMENT OF PLEXOPATHIES

Most patients with plexopathies will undergo both imaging with MRI and EDx evaluations. Severe pain from acute idiopathic lumbosacral plexopathy may respond to a short course of glucocorticoids.

FURTHER READING

Amato AA, Russell J: *Neuromuscular Disorders,* 2nd ed. New York, McGraw-Hill, 2016.

Arnold WD, Elsheikh B: Entrapment neuropathies. Neurol Clin 31:405, 2013. [PubMed: 23642716]

Barohn RJ, Amato AA. Pattern-recognition approach to neuropathy and neuronopathy. Neurol Clin 31:343, 2013. [PubMed: 23642713]

England JD et al: Evaluation of distal symmetric polyneuropathy: The role of autonomic testing, nerve biopsy, and skin biopsy (an evidence-based review). Muscle Nerve 39:106, 2009. [PubMed: 19086069]

England JD et al: Evaluation of distal symmetric polyneuropathy: The role of laboratory and genetic testing (an evidence-based review). Muscle Nerve 39:116, 2009. [PubMed: 19086068]

Klein CJ et al: Inherited neuropathies: Clinical overview and update. Muscle Nerve 48:604, 2013. [PubMed: 23801417]

Pasnoor M et al: Diabetic neuropathy part 1: Overview and symmetric phenotypes. Neurol Clin 31:425, 2013. [PubMed: 23642717]

Pasnoor M et al: Diabetic neuropathy part 2: Proximal and asymmetric phenotypes. Neurol Clin 31:447, 2013. [PubMed: 23642718]

Saporta AS et al: Charcot-Marie-Tooth disease subtypes and genetic testing strategies. Ann Neurol 69:22, 2011. [PubMed: 21280073]

Waldfogel JM et al: Pharmacotherapy for diabetic peripheral neuropathy pain and quality of life: A systematic review. Neurology 87:978, 2016. [PubMed: 27511182]

Chapter 439: Guillain-Barré Syndrome and Other Immune-Mediated Neuropathies

Stephen L. Hauser; Anthony A. Amato

GUILLAIN-BARRÉ SYNDROME

Guillain-Barré syndrome (GBS) is an acute, frequently severe, and fulminant polyradiculoneuropathy that is autoimmune in nature. It occurs year-round at a rate of between 1 and 4 cases per 100,000 annually; in the United States, ~5000–6000 cases occur per year. Males are at slightly higher risk for GBS than females, and in Western countries, adults are more frequently affected than children.

Clinical Manifestations

GBS manifests as a rapidly evolving areflexic motor paralysis with or without sensory disturbance. The usual pattern is an ascending paralysis that may be first noticed as rubbery legs. Weakness typically evolves over hours to a few days and is frequently accompanied by tingling dysesthesias in the extremities. The legs are usually more affected than the arms, and facial diparesis is present in 50% of affected individuals. The lower cranial nerves are also frequently involved, causing bulbar weakness with difficulty handling secretions and maintaining an airway; the diagnosis in these patients may initially be mistaken for brainstem ischemia. Pain in the neck, shoulder, back, or diffusely over the spine is also common in the early stages of GBS, occurring in ~50% of patients. Most patients require hospitalization, and in different series, up to 30% require ventilatory assistance at some time during the illness. The need for mechanical ventilation is associated with more severe weakness on admission, a rapid tempo of progression, and the presence of facial and/or bulbar weakness during the first week of symptoms. Fever and constitutional symptoms are absent at the onset and, if present, cast doubt on the diagnosis. Deep tendon reflexes attenuate or disappear within the first few days of onset. Cutaneous sensory deficits (e.g., loss of pain and temperature sensation) are usually relatively mild, but functions subserved by large sensory fibers, such as deep tendon reflexes and proprioception, are more severely affected. Bladder dysfunction may occur in severe cases but is usually transient. If bladder dysfunction is a prominent feature and comes early in the course or there is a sensory level on examination, diagnostic possibilities other than GBS should be considered, particularly spinal cord disease **(Chap. 434)**. Once clinical worsening stops and the patient reaches a plateau (almost always within 4 weeks of onset), further progression is unlikely.

Autonomic involvement is common and may occur even in patients whose GBS is otherwise mild. The usual manifestations are loss of vasomotor control with wide fluctuations in blood pressure, postural hypotension, and cardiac dysrhythmias. These features require close monitoring, and management and can be fatal. Pain is another common feature of GBS; in addition to the acute pain described above, a deep aching pain may be present in weakened muscles that patients liken to having overexercised the previous day. Other pains in GBS include dysesthetic pain in the extremities as a manifestation of sensory nerve fiber involvement. These pains are self-limited and often respond to standard analgesics **(Chap. 10)**.

Several subtypes of GBS are recognized, as determined primarily by electrodiagnostic (Edx) and pathologic distinctions **(Table 439-1)**. The most common variant is acute inflammatory demyelinating polyneuropathy (AIDP). Additionally, there are two axonal variants, which are often clinically severe—the acute motor axonal neuropathy (AMAN) and acute motor sensory axonal neuropathy (AMSAN) subtypes. In addition, a range of limited or regional GBS syndromes are also encountered. Notable among these is the Miller Fisher syndrome (MFS), which presents as rapidly evolving ataxia and areflexia of limbs without weakness, and ophthalmoplegia, often with pupillary paralysis. The MFS variant accounts for ~5% of all cases and is strongly associated with antibodies to the ganglioside GQ1b (see "Immunopathogenesis," below). Other regional variants of GBS include (1) pure sensory forms; (2) ophthalmoplegia with anti-GQ1b antibodies as part of severe motor-sensory GBS; (3) GBS with severe bulbar and facial paralysis, sometimes associated with antecedent cytomegalovirus (CMV) infection and anti-GM2 antibodies; and (4) acute pandysautonomia **(Chap. 432)**.

TABLE 439-1

Subtypes of Guillain-Barré Syndrome (GBS)

Subtype	Features	Electrodiagnosis	Pathology
Acute inflammatory demyelinating polyneuropathy (AIDP)	Adults affected more than children; 90% of cases in Western world; recovery rapid; anti-GM1 antibodies (<50%)	Demyelinating	First attack on Schwann cell surface; widespread myelin damage, macrophage activation, and lymphocytic infiltration; variable secondary axonal damage
Acute motor axonal neuropathy (AMAN)	Children and young adults; prevalent in China and Mexico; may be seasonal; recovery rapid; anti-GD1a antibodies	Axonal	First attack at motor nodes of Ranvier; macrophage activation, few lymphocytes, frequent periaxonal macrophages; extent of axonal damage highly variable
Acute motor sensory axonal neuropathy (AMSAN)	Mostly adults; uncommon; recovery slow, often incomplete; closely related to AMAN	Axonal	Same as AMAN, but also affects sensory nerves and roots; axonal damage usually severe
Miller Fisher syndrome (MFS)	Adults and children; ophthalmoplegia, ataxia, and areflexia; anti-GQ1b antibodies (90%)	Axonal or demyelinating	Few cases examined; resembles AIDP

Antecedent Events

Approximately 70% of cases of GBS occur 1–3 weeks after an acute infectious process, usually respiratory or gastrointestinal. Culture and seroepidemiologic techniques show that 20–30% of all cases occurring in North America, Europe, and Australia are preceded by infection or reinfection with *Campylobacter jejuni*. A similar proportion is preceded by a human herpes virus infection, often CMV or Epstein-Barr virus. Other viruses (e.g., HIV, hepatitis E, Zika) and also *Mycoplasma pneumoniae* have been identified as agents involved in antecedent infections, as have recent immunizations. The swine influenza vaccine, administered widely in the United States in 1976, is the most notable example. Influenza vaccines in use from 1992 to 1994, however, resulted in only one additional case of GBS per million persons vaccinated, and the more recent seasonal influenza vaccines appear to confer a GBS risk of <1 per million. Epidemiologic studies looking at H1N1 vaccination demonstrated at most only a slight increased risk of GBS. Meningococcal vaccinations (Menactra) does not appear to carry an increased risk. Older-type rabies vaccine, prepared in nervous system tissue, is implicated as a trigger of GBS in developing countries where it is still used; the mechanism is presumably immunization against neural antigens. GBS also occurs more frequently than can be attributed to chance alone in patients with lymphoma (including Hodgkin's disease), in HIV-seropositive individuals, and in patients with systemic lupus erythematosus (SLE).

C. jejuni has also been implicated in summer outbreaks of AMAN among children and young adults exposed to chickens in rural China. Infection by Zika virus recently has been implicated in the increased incidence of GBS in Brazil and other endemic regions.

Immunopathogenesis

Several lines of evidence support an autoimmune basis for acute inflammatory demyelinating polyneuropathy (AIDP), the most common and best-studied type of GBS; the concept extends to all of the subtypes of GBS (Table 439-1).

It is likely that both cellular and humoral immune mechanisms contribute to tissue damage in AIDP. T cell activation is suggested by the finding that elevated levels of cytokines and cytokine receptors are present in serum (interleukin [IL] 2, soluble IL-2 receptor) and in cerebrospinal fluid (CSF) (IL-6, tumor necrosis factor α, interferon γ). AIDP is also closely analogous to an experimental T cell–mediated immunopathy designated *experimental allergic neuritis* (EAN). EAN is induced in laboratory animals by immune sensitization against protein fragments derived from peripheral nerve proteins, and in particular against the P2 protein. Based on analogy to EAN, it was initially thought that AIDP was likely to be primarily a T cell–mediated disorder; however, abundant data now suggest that autoantibodies directed against T-cell independent nonprotein determinants may be central to many cases.

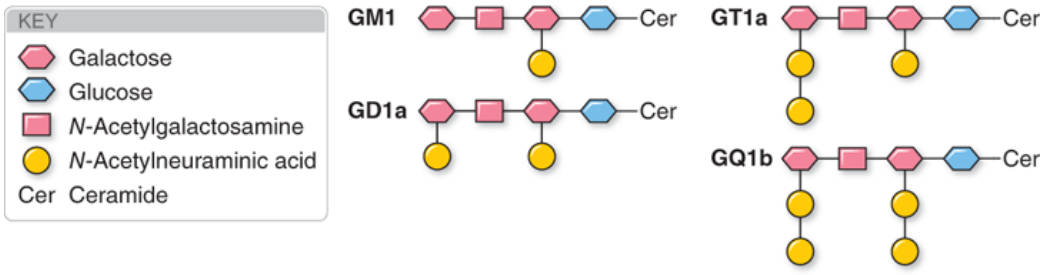
Circumstantial evidence suggests that all GBS results from immune responses to nonself antigens (infectious agents, vaccines) that misdirect to host nerve tissue through a resemblance-of-epitope (molecular mimicry) mechanism **(Fig. 439-1)**. The neural targets are likely to be glycoconjugates, specifically gangliosides **(Table 439-2; Fig. 439-2)**. Gangliosides are complex glycosphingolipids that contain one or more sialic acid residues; various gangliosides participate in cell-cell interactions (including those between axons and glia), modulation of receptors, and regulation of growth. They are typically exposed on the plasma membrane of cells, rendering them susceptible to an antibody-mediated attack. Gangliosides and other glycoconjugates are present in large quantity in human nervous tissues and in key sites, such as nodes of Ranvier. Antiganglioside antibodies, most frequently to GM1, are common in GBS (20–50% of cases), particularly in AMAN and AMSAN and in those cases preceded by *C. jejuni* infection. Some AIDP autoantibodies may

recognize glycolipid heterocomplexes, rather than single species, present on cell membranes. Furthermore, isolates of *C. jejuni* from stool cultures of patients with GBS have surface glycolipid structures that antigenically cross react with gangliosides, including GM1, concentrated in human nerves. Sialic acid residues from pathogenic *C. jejuni* strains can also trigger activation of dendritic cells via signaling through toll-like receptor 4 (TLR4), promoting B cell differentiation and further amplifying humoral autoimmunity. Another line of evidence implicating humoral autoimmunity is derived from cases of GBS that followed intravenous administration of bovine brain gangliosides for treatment of various neuropathies; 5–15 days after injection, some recipients developed AMAN with high titers of anti-GM1 antibodies that recognized epitopes at nodes of Ranvier and motor endplates. Experimentally, anti-GM1 antibodies can trigger complement-mediated injury at paranodal axon-glial junctions, disrupting the clustering of sodium channels and likely contributing to conduction block (see "Pathophysiology," below).

FIGURE 439-1

Spectrum of disorders in Guillain–Barré Syndrome and associated antiganglioside antibodies. IgG autoantibodies against GM1 or GD1a are strongly associated with AMAN, as well as the more extensive AMSAN, and the less extensive acute motor-conduction-block neuropathy. IgG anti-GQ1b antibodies, which cross-react with GT1a, are strongly associated with Miller Fisher syndrome, its incomplete forms (acute ophthalmoparesis [without ataxia] and acute ataxic neuropathy [without ophthalmoplegia]), and its more extensive form, Bickerstaff's brain-stem encephalitis. Pharyngeal–cervical–brachial weakness is categorized as a localized form of acute motor axonal neuropathy or an extensive form of Miller Fisher syndrome. Half of patients with pharyngeal–cervical–brachial weakness have IgG anti-GT1a antibodies, which often cross-react with GQ1b. IgG anti-GD1a antibodies have also been detected in a small percentage of patients. The anti-GQ1b antibody syndrome includes Miller Fisher syndrome, acute ophthalmoparesis, acute ataxic neuropathy, Bickerstaff's brain-stem encephalitis, and pharyngeal–cervical–brachial weakness. The presence of clinical overlap also indicates that Miller Fisher syndrome is part of a continuous spectrum with these conditions. Patients who have had Guillain–Barré syndrome overlapped with Miller Fisher syndrome or with its related conditions have IgG antibodies against GM1 or GD1a as well as against GQ1b or GT1a, supporting a link between AMAN and anti-GQ1b syndrome. AMAN, acute motor axonal neuropathy; AMSAN, acute motor–sensory axonal neuropathy. *(Adapted from N Yuki, H-P Hartung: Guillain-Barré syndrome. N Engl J Med 366:2294, 2012, Figure 1.)*

Subtypes and variants	IgG autoantibodies to
Guillain-Barré syndrome	
Acute inflammatory demyelinating polyneuropathy	None
Facial variant: Facial diplegia and paresthesia	None
Acute motor axonal neuropathy	GM1, GD1a
More and less extensive forms	
Acute motor-sensory axonal neuropathy	GM1, GD1a
Acute motor-conduction-block neuropathy	GM1, GD1a
Pharyngeal-cervical-brachial weakness	GT1a>GQ1b>>GD1a
Miller Fisher syndrome	GQ1b, GT1a
Incomplete forms	
Acute ophthalmoparesis (without ataxia)	GQ1b, GT1a
Acute ataxic neyropathy (without ophthalmoplegia)	GQ1b, GT1a
CNS variant: Bickerstaff's brainstem encephalitis	GQ1b, GT1a

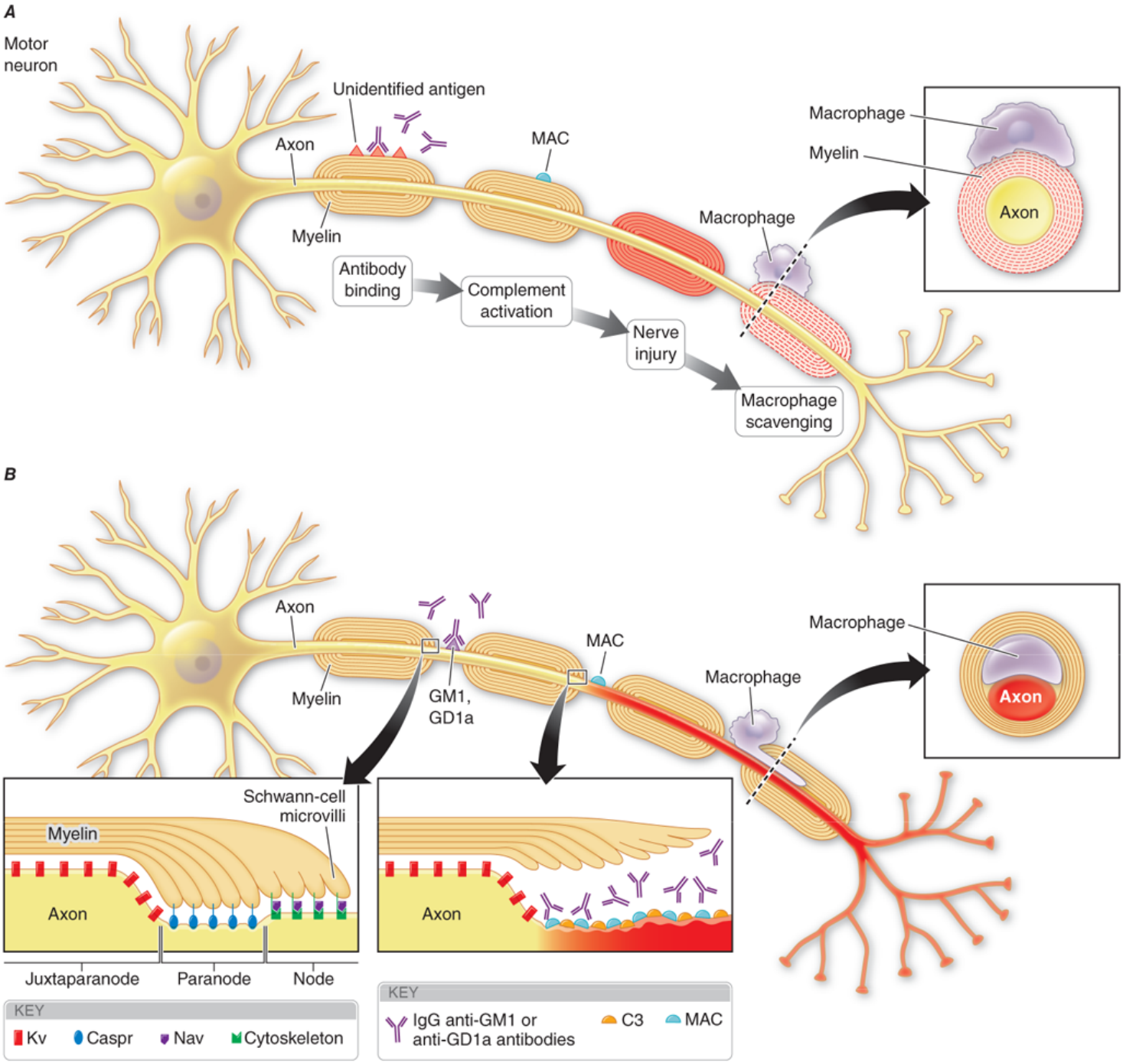


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 439-2

Possible immune mechanisms in GBS. Panel ***A*** shows the immunopathogenesis of AIDP. Although autoantigens have yet to be unequivocally identified, autoantibodies may bind to myelin antigens and activate complement. This is followed by the formation of membrane-attack complex (MAC) on the outer surface of Schwann cells and the initiation of vesicular degeneration. Macrophages subsequently invade myelin and act as scavengers to remove myelin debris. Panel ***B*** shows the immunopathogenesis of acute axonal forms of GBS (AMAN and AMSAN). Myelinated axons are divided into four functional regions: the nodes of Ranvier, paranodes, juxtaparanodes, and internodes. Gangliosides GM1 and GD1a are strongly expressed at the nodes of Ranvier, where the voltage-gated sodium (Nav) channels are localized. Contactin-associated protein (Caspr) and voltage-gated potassium (Kv) channels are respectively present at the paranodes and juxtaparanodes. IgG anti-GM1 or anti-GD1a autoantibodies bind to the nodal axolemma, leading to MAC formation. This results in the disappearance of Nav clusters and the detachment of paranodal myelin, which can lead to nerve-conduction failure and

muscle weakness. Axonal degeneration may follow at a later stage. Macrophages subsequently invade from the nodes into the periaxonal space, scavaenging the injured axons. *(Adapted from N Yuki, H-P Hartung: Guillain-Barré syndrome. N Engl J Med 366:2294, 2012, Figure 2.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 439-2

Principal Antiglycolipid Antibodies Implicated in Immune Neuropathies

Clinical Presentation	Antibody Target	Usual Isotype
Acute Immune Neuropathies (Guillain-Barré Syndrome)		
Acute inflammatory demyelinating polyneuropathy (AIDP)	No clear patterns	IgG (polyclonal)
	GM1 most common	
Acute motor axonal neuropathy (AMAN)	GD1a, GM1, GM1b, GalNAc–GD1a (<50% for any)	IgG (polyclonal)
Miller Fisher syndrome (MFS)	GQ1b (>90%)	IgG (polyclonal)
Acute pharyngeal cervicobrachial neuropathy (APCBN)	GT1a (? most)	IgG (polyclonal)
Chronic Immune Neuropathies		
Chronic inflammatory demyelinating polyneuropathy (CIDP) (75%)	<10% to CNTN1 or NF155 and even more rare to P0, myelin P2 protein, orPMP22,	IgG4 with CNTN1 and NF155
CIDP-M (MGUS associated) (25%)	Neural binding sites	IgG, IgA (monoclonal)
Chronic sensory > motor neuropathy	SPGP, SGLPG (on MAG) (50%)	IgM (monoclonal)
	Uncertain (50%)	IgM (monoclonal)
Multifocal motor neuropathy (MMN)	GM1, GalNAc–GD1a, others (25–50%)	IgM (polyclonal, monoclonal)
Chronic sensory ataxic neuropathy	GD1b, GQ1b, and other b-series gangliosides	IgM (monoclonal)

Abbreviations: CIDP-M, CIDP with a monoclonal gammopathy; CNTN1, contactin-1; MAG, myelin-associated glycoprotein; MGUS, monoclonal gammopathy of undetermined significance; NF155, neurofascin-155.

Source: Modified from HJ Willison, N Yuki: Brain 125:2591, 2002.

Anti-GQ1b IgG antibodies are found in >90% of patients with MFS (Table 439-2; Fig. 439-2), and titers of IgG are highest early in the course. Anti-GQ1b antibodies are not found in other forms of GBS unless there is extraocular motor nerve involvement. A possible explanation for this association is that extraocular motor nerves are enriched in GQ1b gangliosides in comparison to limb nerves. In addition, a monoclonal anti-GQ1b antibody raised against *C. jejuni* isolated from a patient with MFS blocked neuromuscular transmission experimentally.

Taken together, these observations provide strong but still inconclusive evidence that autoantibodies play an important pathogenic role in GBS. Although antiganglioside antibodies have been studied most intensively, other antigenic targets may also be important. Proof that these antibodies are pathogenic requires that they be capable of mediating disease following direct passive transfer to naïve hosts; this has not yet been demonstrated, although one case of possible maternal-fetal transplacental transfer of GBS has been described.

In AIDP, an early step in the induction of tissue damage appears to be complement deposition along the outer surface of the Schwann cell. Activation of complement initiates a characteristic vesicular disintegration of the myelin sheath and also leads to recruitment of activated macrophages, which participate in damage to myelin and axons. In AMAN, the pattern is different in that complement is deposited along with IgG at the nodes of Ranvier along large motor axons. Interestingly, in cases of AMAN, antibodies against GD1a appear to have a fine specificity that favors binding to motor rather than sensory nerve roots, even though this ganglioside is expressed on both fiber types.

Pathophysiology

In the demyelinating forms of GBS, the basis for flaccid paralysis and sensory disturbance is conduction block. This finding, demonstrable electrophysiologically, implies that the axonal connections remain intact. Hence, recovery can take place rapidly as remyelination occurs. In severe cases of demyelinating GBS, secondary axonal degeneration usually occurs; its extent can be estimated electrophysiologically. More secondary axonal

degeneration correlates with a slower rate of recovery and a greater degree of residual disability. When a severe primary axonal pattern is encountered electrophysiologically, the implication is that axons have degenerated and become disconnected from their targets, specifically the neuromuscular junctions, and must therefore regenerate for recovery to take place. In motor axonal cases in which recovery is rapid, the lesion is thought to be localized to preterminal motor branches, allowing regeneration and reinnervation to take place quickly. Alternatively, in mild cases, collateral sprouting and reinnervation from surviving motor axons near the neuromuscular junction may begin to reestablish physiologic continuity with muscle cells over a period of several months.

Laboratory Features

CSF findings are distinctive, consisting of an elevated CSF protein level (1–10 g/L [100–1000 mg/dL]) without accompanying pleocytosis. The CSF is often normal when symptoms have been present for ≤48 h; by the end of the first week, the level of protein is usually elevated. A transient increase in the CSF white cell count (10–100/μL) occurs on occasion in otherwise typical GBS; however, a sustained CSF pleocytosis suggests an alternative diagnosis (viral myelitis) or a concurrent diagnosis such as unrecognized HIV infection, leukemia or lymphoma with infiltration of nerves, or neurosarcoidosis. Edx features are mild or absent in the early stages of GBS and lag behind the clinical evolution. In AIDP, the earliest features are prolonged F-wave latencies, prolonged distal latencies, and reduced amplitudes of compound muscle action potentials (CMAPs), probably owing to the predilection for involvement of nerve roots and distal motor nerve terminals early in the course. Later, slowing of conduction velocity, conduction block, and temporal dispersion may be appreciated (Table 439-1). Occasionally, sensory nerve action potentials (SNAPs) may be normal in the feet (e.g., sural nerve) when abnormal in the arms. This is also a sign that the patient does not have one of the more typical “length-dependent” polyneuropathies. In cases with primary axonal pathology, the principal Edx finding is reduced amplitude of CMAPs (and also SNAPS with AMSAN) without conduction slowing or prolongation of distal latencies.

Diagnosis

GBS is a descriptive entity. The diagnosis of AIDP is made by recognizing the pattern of rapidly evolving paralysis with areflexia, absence of fever or other systemic symptoms, and characteristic antecedent events. In 2011, the Brighton Collaboration developed a new set of case definitions for GBS in response to needs of epidemiologic studies of vaccination and assessing risks of GBS **(Table 439-3)**. These criteria have subsequently been validated. Other disorders that may enter into the differential diagnosis include acute myelopathies (especially with prolonged back pain and sphincter disturbances); diphtheria (early oropharyngeal disturbances); Lyme polyradiculitis and other tick-borne paralyses; porphyria (abdominal pain, seizures, psychosis); vasculitic neuropathy (check erythrocyte sedimentation rate, described below); poliomyelitis (fever and meningismus common); West Nile virus; CMV polyradiculitis (in immunocompromised patients); critical illness neuropathy or myopathy; neuromuscular junction disorders such as myasthenia gravis and botulism (pupillary reactivity lost early); poisonings with organophosphates, thallium, or arsenic; paralytic shellfish poisoning; or severe hypophosphatemia (rare). Laboratory tests are helpful primarily to exclude mimics of GBS. Edx features may be minimal, and the CSF protein level may not rise until the end of the first week. If the diagnosis is strongly suspected, treatment should be initiated without waiting for evolution of the characteristic Edx and CSF findings to occur. GBS patients with risk factors for HIV or with CSF pleocytosis should have a serologic test for HIV.

TABLE 439-3

Brighton Criteria for Diagnosis of Guillain-Barré Syndrome (GBS) and Miller Fisher Syndrome

Clinical case definitions for diagnosis of GBS

Level 1 of diagnostic certainty

Bilateral AND flaccid weakness of the limbs

AND

Decreased or absent deep tendon reflexes in weak limbs

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

Electrophysiologic findings consistent with GBS

AND

Cytoalbuminologic dissociation (i.e., elevation of CSF protein level above laboratory normal value AND CSF total white cell count <50 cells/μL)

AND

Absence of an identified alternative diagnosis for weakness

Level 2 of diagnostic certainty

Bilateral AND flaccid weakness of the limbs

AND

Decreased or absent deep tendon reflexes in weak limbs

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

CSF total white cell count <50 cells/μL (with or without CSF protein elevation above laboratory normal value)

OR

If CSF not collected or results not available, electrophysiologic studies consistent with GBS

AND

Absence of identified alternative diagnosis for weakness

Level 3 of diagnostic certainty

Bilateral and flaccid weakness of the limbs

AND

Decreased or absent deep tendon reflexes in weak limbs

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

Absence of identified alternative diagnosis for weakness

Clinical case definitions for diagnosis of Miller Fisher syndrome

Level 1 of diagnostic certainty

Bilateral ophthalmoparesis and bilateral reduced or absent tendon reflexes, and ataxia

AND

Absence of limb weakness

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

Cytoalbuminologic dissociation (i.e., elevation of cerebrospinal protein above the laboratory normal and total CSF white cell count <50 cells/μL)

AND

Nerve conduction studies are normal, OR indicate involvement of sensory nerves only

AND

No alterations in consciousness or corticospinal tract signs

AND

Absence of identified alternative diagnosis

Level 2 of diagnostic certainty

Bilateral ophthalmoparesis and bilateral reduced or absent tendon reflexes and ataxia

AND

Absence of limb weakness

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

CSF with a total white cell count <50 cells/μL) (with or without CSF protein elevation above laboratory normal value)

OR

Nerve conduction studies are normal, OR indicate involvement of sensory nerves only

AND

No alterations in consciousness or corticospinal tract signs

AND

Absence of identified alternative diagnosis

Level 3 of diagnostic certainty

Bilateral ophthalmoparesis and bilateral reduced or absent tendon reflexes and ataxia

AND

Absence of limb weakness

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

No alterations in consciousness or corticospinal tract signs

AND

Absence of identified alternative diagnosis

Abbreviation: CSF, cerebrospinal fluid.

Source: From JJ Sejvar et al: Guillain-Barré syndrome and Fisher syndrome: Case definitions and guidelines for collection, analysis, and presentation of immunization safety data. Vaccine 29:599, 2011. Validation study published by C Fokke et al: Diagnosis of Guillain-Barré syndrome and validation of Brighton criteria. Brain 137:33, 2014.

TREATMENT

TREATMENT

Guillain-Barré Syndrome

In the vast majority of patients with GBS, treatment should be initiated as soon after diagnosis as possible. Each day counts; ~2 weeks after the first motor symptoms, it is not known whether immunotherapy is still effective. If the patient has already reached the plateau stage, then treatment probably is no longer indicated, unless the patient has severe motor weakness and one cannot exclude the possibility that an immunologic attack is still ongoing. Either high-dose intravenous immune globulin (IVIg) or plasmapheresis can be initiated, as they are equally effective for typical GBS. A combination of the two therapies is not significantly better than either alone. IVIg is often the initial therapy chosen because of its ease of administration and good safety record. Anecdotal data have also suggested that IVIg may be preferable to plasma exchange (PE) for the AMAN and MFS variants of GBS. IVIg is administered as five daily infusions for a total dose of 2 g/kg body weight. There is some evidence that GBS autoantibodies are neutralized by anti-idiotypic antibodies present in IVIg preparations, perhaps accounting for the therapeutic effect. A course of plasmapheresis usually consists of ~40–50 mL/kg PE 4–5 times over 7–10 days. Meta-analysis of randomized clinical trials indicates that treatment reduces the need for mechanical ventilation by nearly half (from 27 to 14% with PE) and increases the likelihood of full recovery at 1 year (from 55 to 68%). Functionally significant improvement may occur toward the end of the first week of treatment or may be delayed for several weeks. The lack of noticeable improvement following a course of IVIg or PE is not an indication to treat with the alternate treatment. However, there are occasional patients who are treated early in the course of GBS and improve, who then relapse within a month. Brief retreatment with the original therapy is usually effective in such cases. Glucocorticoids have not been found to be effective in GBS. Occasional patients with very mild forms of GBS, especially those who appear to have already reached a plateau when initially seen, may be managed conservatively without IVIg or PE.

In the worsening phase of GBS, most patients require monitoring in a critical care setting, with particular attention to vital capacity, heart rhythm, blood pressure, nutrition, deep-vein thrombosis prophylaxis, cardiovascular status, early consideration (after 2 weeks of intubation) of tracheotomy, and chest physiotherapy. As noted, ~30% of patients with GBS require ventilatory assistance, sometimes for prolonged periods of time (several weeks or longer). Frequent turning and assiduous skin care are important, as are daily range-of-motion exercises to avoid joint contractures and daily reassurance as to the generally good outlook for recovery.

Prognosis and Recovery

Approximately 85% of patients with GBS achieve a full functional recovery within several months to a year, although minor findings on examination (such as areflexia) may persist and patients often complain of continued symptoms, including fatigue. The mortality rate is <5% in optimal settings; death usually results from secondary pulmonary complications. The outlook is worst in patients with severe proximal motor and sensory axonal damage. Such axonal damage may be either primary or secondary in nature (see "Pathophysiology," above), but in either case successful regeneration cannot occur. Other factors that worsen the outlook for recovery are advanced age, a fulminant or severe attack, and a delay in the onset of treatment. Between 5 and 10% of patients with typical GBS have one or more late relapses; many of these cases are then classified as chronic inflammatory demyelinating polyneuropathy (CIDP).

CHRONIC INFLAMMATORY DEMYELINATING POLYNEUROPATHY

CIDP is distinguished from GBS by its chronic course. In other respects, this neuropathy shares many features with the common demyelinating form of GBS, including elevated CSF protein levels and the Edx findings of acquired demyelination. Most cases occur in adults, and males are affected slightly more often than females. The incidence of CIDP is lower than that of GBS, but due to the protracted course, the prevalence is greater.

Clinical Manifestations

Onset is usually gradual over a few months or longer, but in a few cases, the initial attack is indistinguishable from that of GBS. An acute-onset form of CIDP may mimic GBS but should be considered if it deteriorates >9 weeks after onset or relapses at least three times. Symptoms are both motor and sensory in most cases. Weakness of the limbs is usually symmetric but can be strikingly asymmetric in multifocal acquired demyelinating sensory and motor (MADSAM) neuropathy variant (Lewis-Sumner syndrome) in which discrete peripheral nerves are involved. There is considerable variability from case to case. Some patients experience a chronic progressive course, whereas others, usually younger patients, have a relapsing and remitting course. Some have only motor findings, and a small proportion present with a relatively pure syndrome of sensory ataxia. Tremor occurs in ~10% and may become more prominent during periods of subacute worsening or improvement. A small proportion have cranial nerve findings, including external ophthalmoplegia. CIDP tends to ameliorate over time with treatment; the result is that many years after onset, nearly 75% of patients have reasonable functional status. Death from CIDP is uncommon.

Diagnosis

The diagnosis rests on characteristic clinical, CSF, and electrophysiologic findings. The CSF is usually acellular with an elevated protein level, sometimes several times normal. As with GBS, a CSF pleocytosis should lead to the consideration of HIV infection, leukemia or lymphoma, and neurosarcoidosis. Edx findings reveal variable degrees of conduction slowing, prolonged distal latencies, distal and temporal dispersion of CMAPs, and conduction block as the principal features. In particular, the presence of conduction block is a certain sign of an acquired demyelinating process. Evidence of axonal loss, presumably secondary to demyelination, is present in >50% of patients. Serum protein electrophoresis with immunofixation is indicated to search for

monoclonal gammopathy and associated conditions (see "Monoclonal Gammopathy of Undetermined Significance," below). In all patients with presumptive CIDP, it is also reasonable to exclude vasculitis, collagen vascular disease (especially SLE), chronic hepatitis, HIV infection, amyloidosis, and diabetes mellitus. Other associated conditions include inflammatory bowel disease and lymphoma.

Pathogenesis

Biopsy typically reveals little inflammation and onion-bulb changes (imbricated layers of attenuated Schwann cell processes surrounding an axon) that result from recurrent demyelination and remyelination (Fig. 439-1). The response to therapy suggests that CIDP is immune-mediated; CIDP responds to glucocorticoids, whereas GBS does not. Passive transfer of demyelination into experimental animals has been accomplished using IgG purified from the serum of some patients with CIDP, lending support for a humoral autoimmune pathogenesis. A minority of patients have serum antibodies against P0, myelin P2 protein, or PMP22 (proteins whose genes are mutated in certain forms of hereditary Charcot-Marie-Tooth neuropathy). More recently antibodies of IgG4 isotype directed against contactin-1 (CNTN1) or neurofascin-155 (NF155) have been associated with early axonal damage, severe distal motor involvement or sensory ataxia with tremor, and a poor response to IVIg. CNTN1 and its partner contactin-associated protein-1 (CASPR1) interact with NF155 at paranodal axoglial junctions. Passive transfer of IgG4 CNTN1 antibodies produces paranodal damage and ataxia in rodents. It is also of interest that a CIDP-like illness developed spontaneously in the nonobese diabetic (NOD) mouse when the immune co-stimulatory molecule B7-2 (CD86) was genetically deleted; this suggests that CIDP can result from altered triggering of T cells by antigen-presenting cells.

As many as 25% of patients with clinical features of CIDP also have a monoclonal gammopathy of undetermined significance (MGUS). Cases associated with monoclonal IgA or IgG kappa usually respond to treatment as favorably as cases without a monoclonal gammopathy. Patients with IgM-kappa monoclonal gammopathy and antibodies directed against myelin-associated glycoprotein (MAG) have a distinct demyelinating polyneuropathy with more sensory findings, usually only distal weakness, and a poor response to immunotherapy.

TREATMENT

TREATMENT

Chronic Inflammatory Demyelinating Polyneuropathy

Most authorities initiate treatment for CIDP when progression is rapid or walking is compromised. If the disorder is mild, management can be expectant, awaiting spontaneous remission. Controlled studies have shown that high-dose IVIg, PE, and glucocorticoids are all more effective than placebo. Initial therapy is usually with IVIg, administered as 2.0 g/kg body weight given in divided doses over 2–5 days; three monthly courses are generally recommended before concluding a patient is a treatment failure. If the patient responds, the infusion intervals can be gradually increased or the dosage decreased (e.g., starting at 1 g/kg every 3–4 weeks). PE, which appears to be as effective as IVIg, is initiated at 2-3 treatments per week for 6 weeks; periodic re-treatment may also be required. Treatment with glucocorticoids is another option (60–80 mg prednisone PO daily for 1–2 months, followed by a gradual dose reduction of 10 mg per month as tolerated), but long-term adverse effects including bone demineralization, gastrointestinal bleeding, and cushingoid changes are problematic. As many as one-third of patients with CIDP fail to respond adequately to the initial therapy chosen; a different treatment should then be tried. Patients who fail therapy with IVIg, PE, and glucocorticoids may benefit from treatment with immunosuppressive agents such as azathioprine, methotrexate, cyclosporine, and cyclophosphamide, either alone or as adjunctive therapy. CIDP associated with anti-CNTN1 and NF155 antibodies is typically refractory to IVIg, but a few studies suggest a response to rituximab. Use of these therapies requires periodic reassessment of their risks and benefits. In patients with a CIDP-like neuropathy who fail to respond to treatment, it is important to evaluate for POEMS syndrome (*p*olyneuropathy, *o*rganomegaly, *e*ndocrinopathy, *m*onoclonal gammopathy, *s*kin changes; see below).

MULTIFOCAL MOTOR NEUROPATHY

Multifocal motor neuropathy (MMN) is a distinctive but uncommon neuropathy that presents as slowly progressive motor weakness and atrophy evolving over years in the distribution of selected nerve trunks, associated with sites of persistent focal motor conduction block in the same nerve trunks. Sensory fibers are relatively spared. The arms are affected more frequently than the legs, and >75% of all patients are male. Some cases have been confused with lower motor neuron forms of amyotrophic lateral sclerosis **(Chap. 429)**. Less than 50% of patients present with high titers of polyclonal IgM antibody to the ganglioside GM1. It is uncertain how this finding relates to the discrete foci of persistent motor conduction block, but high concentrations of GM1 gangliosides are normal constituents of nodes of Ranvier in peripheral nerve fibers. Pathology reveals demyelination and mild inflammatory changes at the sites of conduction block.

Most patients with MMN respond to high-dose IVIg (dosages as for CIDP, above); periodic re-treatment is required (usually at least monthly) to maintain the benefit. Some refractory patients have responded to rituximab or cyclophosphamide. Glucocorticoids and PE are not effective.

NEUROPATHIES WITH MONOCLONAL GAMMOPATHY

MULTIPLE MYELOMA

Clinically overt polyneuropathy occurs in ~5% of patients with the commonly encountered type of multiple myeloma, which exhibits either lytic or diffuse osteoporotic bone lesions. These neuropathies are sensorimotor, are usually mild and slowly progressive but may be severe, and generally do not reverse with successful suppression of the myeloma. In most cases, Edx and pathologic features are consistent with a process of axonal degeneration.

In contrast, myeloma with osteosclerotic features, although representing only 3% of all myelomas, is associated with polyneuropathy in one-half of cases. These neuropathies, which may also occur with solitary plasmacytoma, are distinct because they (1) are usually demyelinating in nature and resemble CIDP; (2) often respond to radiation therapy or removal of the primary lesion; (3) are associated with different monoclonal proteins and light chains (almost always lambda as opposed to primarily kappa in the lytic type of multiple myeloma); (4) are typically refractory to standard treatments of CIDP; and (5) may occur in association with other systemic findings including thickening of the skin, hyperpigmentation, hypertrichosis, organomegaly, endocrinopathy, anasarca, and clubbing of fingers. These are features of POEMS syndrome. Levels of vascular endothelial growth factor (VEGF) are increased in the serum, and this factor is thought to somehow play a pathogenic role in this syndrome. Treatment of the neuropathy is best directed at the osteosclerotic myeloma using surgery, radiotherapy, chemotherapy, or autologous peripheral blood stem cell transplantation.

Neuropathies are also encountered in other systemic conditions with gammopathy, including Waldenström's macroglobulinemia, primary systemic amyloidosis, and cryoglobulinemic states (mixed essential cryoglobulinemia, some cases of hepatitis C).

MONOCLONAL GAMMOPATHY OF UNDETERMINED SIGNIFICANCE

Chronic polyneuropathies occurring in association with MGUS are usually associated with the immunoglobulin isotypes IgG, IgA, and IgM. Most patients present with isolated sensory symptoms in their distal extremities and have Edx features of an axonal sensory or sensorimotor polyneuropathy. These patients otherwise resemble idiopathic sensory polyneuropathy, and the MGUS might just be coincidental. They usually do not respond to immunotherapies designed to reduce the concentration of the monoclonal protein. Some patients, however, present with generalized weakness and sensory loss and Edx studies indistinguishable from CIDP without monoclonal gammopathy (see "Chronic Inflammatory Demyelinating Polyneuropathy," above), and their response to immunosuppressive agents is also similar. An exception is the syndrome of IgM kappa monoclonal gammopathy associated with an indolent, long-standing, sometimes static sensory neuropathy, frequently with tremor and sensory ataxia. Most patients are male and aged >50 years. In the majority, the monoclonal IgM immunoglobulin binds to a normal peripheral nerve constituent, MAG, found in the paranodal regions of Schwann cells. Binding appears to be specific for a polysaccharide epitope that is also found in other normal peripheral nerve myelin glycoproteins, P0 and PMP22, and also in other normal nerve-related glycosphingolipids (Fig. 439-1). In the MAG-positive cases, IgM paraprotein is incorporated into the myelin sheaths of affected patients and widens the spacing of the myelin lamellae, thus producing a distinctive ultrastructural pattern. Demyelination and remyelination are the hallmarks of the lesions, but axonal loss develops over time. These anti-MAG polyneuropathies are typical refractory to immunotherapy. In a small proportion of patients (30% at 10 years), MGUS will in time evolve into frankly malignant conditions such as multiple myeloma or lymphoma.

VASCULITIC NEUROPATHY

Peripheral nerve involvement is common in polyarteritis nodosa (PAN), appearing in half of all cases clinically and in 100% of cases at postmortem studies (Chap. 356). The most common pattern is multifocal (asymmetric) motor-sensory neuropathy (mononeuropathy multiplex) due to ischemic lesions of nerve trunks and roots; however, some cases of vasculitic neuropathy present as a distal, symmetric sensorimotor polyneuropathy. Symptoms of neuropathy are a common presenting complaint in patients with PAN. The Edx findings are those of an axonal process. Small- to medium-sized arteries of the vasa nervorum, particularly the epineural vessels, are affected in PAN, resulting in a widespread ischemic neuropathy. A high frequency of neuropathy occurs in eosinophilic granulomatosis with polyangiitis (Churg-Strauss syndrome [CSS]).

Systemic vasculitis should always be considered when a subacute or chronically evolving mononeuropathy multiplex occurs in conjunction with constitutional symptoms (fever, anorexia, weight loss, loss of energy, malaise, and nonspecific pains). Diagnosis of suspected vasculitic neuropathy is made by a combined nerve and muscle biopsy, with serial section or skip-serial techniques.

Approximately one-third of biopsy-proven cases of vasculitic neuropathy are "nonsystemic" in that the vasculitis appears to affect only peripheral nerves. Constitutional symptoms are absent, and the course is more indolent than that of PAN. The erythrocyte sedimentation rate may be elevated, but other tests for systemic disease are negative. Nevertheless, clinically silent involvement of other organs is likely, and vasculitis is frequently found in muscle biopsied at the same time as nerve.

Vasculitic neuropathy may also be seen as part of the vasculitis syndrome occurring in the course of other connective tissue disorders (Chap. 356). The most frequent is rheumatoid arthritis, but ischemic neuropathy due to involvement of vasa nervorum may also occur in mixed cryoglobulinemia, Sjögren's syndrome, granulomatosis with polyangiitis (formerly known as Wegener's), hypersensitivity angiitis, SLE, and progressive systemic sclerosis.

Some vasculitides are associated with antineutrophil cytoplasmic antibodies (ANCAs), which in turn, are subclassified as cytoplasmic (cANCA) or perinuclear (pANCA). cANCAs are directed against proteinase 3 (PR3), whereas pANCAs target myeloperoxidase (MPO). PR3/cANCAs are associated with eosinophilic granulomatosis with polyangiitis, whereas MPO/pANCAs are typically associated with microscopic polyangiitis, CSS, and less commonly PAN. Of note, MPO/pANCA has also been seen in minocycline-induced vasculitis.

Management of these neuropathies, including the “nonsystemic” vasculitic neuropathy, consists of treatment of the underlying condition as well as the aggressive use of glucocorticoids and cyclophosphamide. Use of these immunosuppressive agents has resulted in dramatic improvements in outcome, with 5-year survival rates now >80%. Clinical trials found that the combination of rituximab and glucocorticoids is not inferior to cyclophosphamide and glucocorticoids. Thus, combination therapy with glucocorticoids and rituximab is recommended as the standard initial treatment, particularly for ANCA-associated vasculitis. A recent study demonstrated that mepolizumab, an anti-interleukin-5 monoclonal antibody, when added to standard care is effective for treatment of eosinophilic granulomatosis with polyangiitis.

ANTI-Hu PARANEOPLASTIC NEUROPATHY

This uncommon immune-mediated disorder manifests as a sensory neuronopathy (i.e., selective damage to sensory nerve bodies in dorsal root ganglia) (Chap. 90). The onset is often asymmetric with dysesthesias and sensory loss in the limbs that soon progress to affect all limbs, the torso, and the face. Marked sensory ataxia, pseudoathetosis, and inability to walk, stand, or even sit unsupported are frequent features and are secondary to the extensive deafferentation. Subacute sensory neuronopathy may be idiopathic, but more than half of cases are paraneoplastic, primarily related to lung cancer, and most of those are small-cell lung cancer (SCLC). Diagnosis of the underlying SCLC requires awareness of the association, testing for the paraneoplastic antibody, and often positron emission tomography (PET) scanning for the tumor. The target antigens are a family of RNA-binding proteins (HuD, HuC, and Hel-N1) that in normal tissues are only expressed by neurons. The same proteins are usually expressed by SCLC, triggering in some patients an immune response characterized by antibodies and cytotoxic T cells that cross-react with the Hu proteins of the dorsal root ganglion neurons, resulting in immune-mediated neuronal destruction. An encephalomyelitis may accompany the sensory neuronopathy and presumably has the same pathogenesis. Neurologic symptoms usually precede, by ≤6 months, the identification of SCLC. The sensory neuronopathy runs its course in a few weeks or months and stabilizes, leaving the patient disabled. Most cases are unresponsive to treatment with glucocorticoids, IVIg, PE, or immunosuppressant drugs.

FURTHER READING

Amato AA, Russell JA (eds): *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill, 2016, pp 320–383.

Collins MP: The vasculitic neuropathies: an update. Curr Opin Neurol 25:573, 2012. [PubMed: 22918487]

Devaux JJ et al: Neurofascin-155 IgG4 in chronic inflammatory demyelinating polyneuropathy. Neurology 86:800, 2016. [PubMed: 26843559]

Fokke C et al: Diagnosis of Guillain-Barré syndrome and validation of Brighton criteria. Brain 137:33, 2014. [PubMed: 24163275]

Manso C et al: Contactin-1 IgG4 antibodies cause paranode dismantling and conduction defects. Brain 139:1700, 2016. [PubMed: 27017186]

Querol L, Illa I: Paranodal and other autoantibodies in chronic inflammatory neuropathies. Curr Opin Neurol 28:474, 2015. [PubMed: 26263474]

Willison HJ et al: Guillain-Barré syndrome. Lancet 388: 717, 2016. [PubMed: 26948435]

Yuki N, Hartung H-P: Guillain-Barré syndrome. N Engl J Med 366:2294, 2012. [PubMed: 22694000]

Chapter 440: Myasthenia Gravis and Other Diseases of the Neuromuscular Junction

Anthony A. Amato

INTRODUCTION

Myasthenia gravis (MG) is a neuromuscular junction (NMJ) disorder characterized by weakness and fatigability of skeletal muscles. The underlying defect is a decrease in the number of available acetylcholine receptors (AChRs) at NMJs due to an antibody-mediated autoimmune attack. Treatment now available for MG is highly effective, although a specific cure has remained elusive.

PATHOPHYSIOLOGY

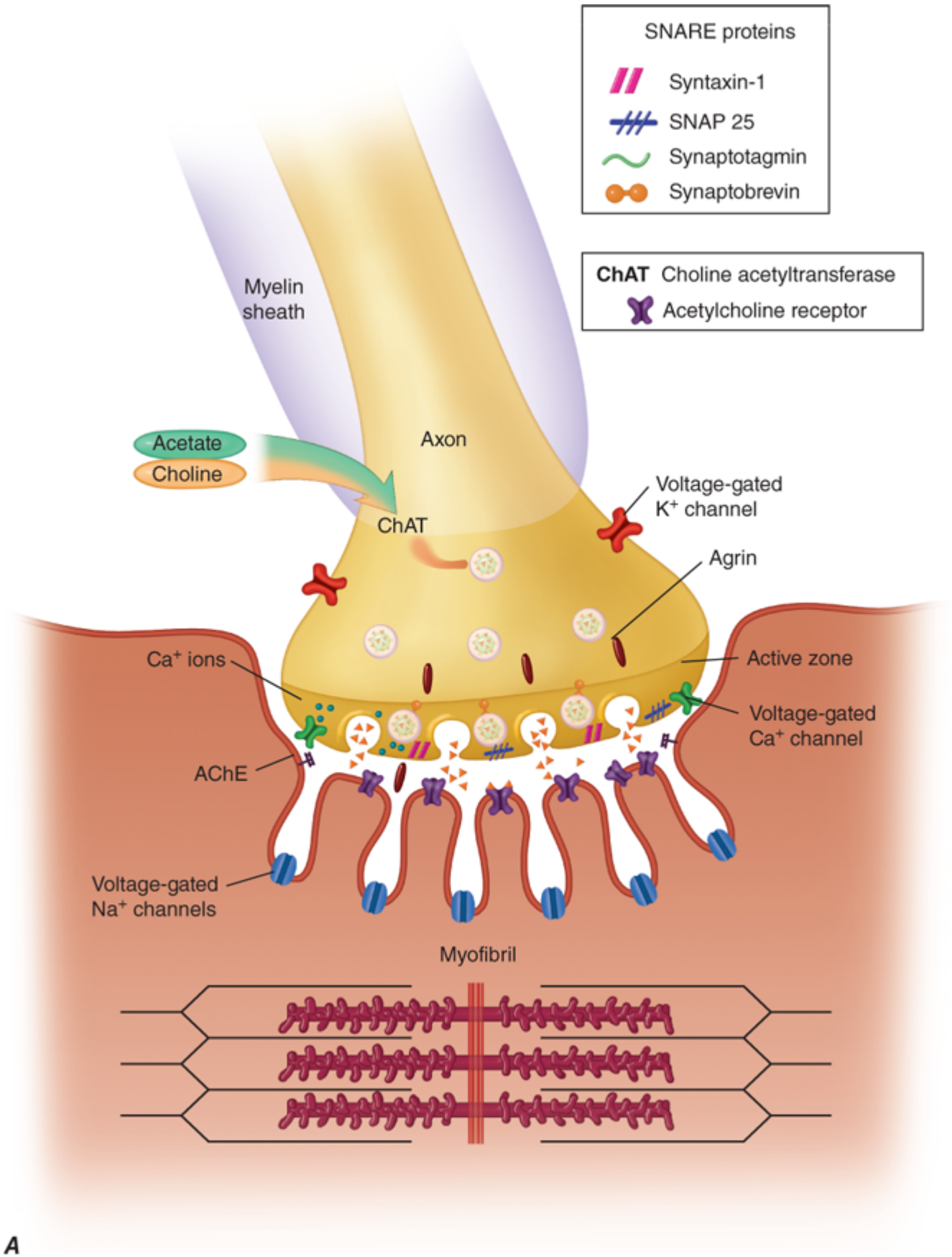
At the NMJ **(Fig. 440-1, Video 440-1)**, acetylcholine (ACh) is synthesized in the motor nerve terminal and stored in vesicles (quanta). When an action potential travels down a motor nerve and reaches the nerve terminal, ACh from 150 to 200 vesicles is released and combines with AChRs that are densely packed at the peaks of postsynaptic folds. The AChR consists of five subunits (2α, 1β, 1δ, 1γ, or ε) arranged around a central pore. When ACh combines with the binding sites on the α subunits of the AChR, the channel in the AChR opens, permitting the rapid entry of cations, chiefly sodium, which produces depolarization at the end-plate region of the muscle fiber. If the depolarization is sufficiently large, it initiates an action potential that is propagated along the muscle fiber, triggering muscle contraction. This process is rapidly terminated by hydrolysis of ACh by acetylcholinesterase (AChE), which is present within the synaptic folds, and by diffusion of ACh away from the receptor.

VIDEO 440-1: Myasthenia Gravis and Other Diseases of the Neuromuscular Junction
Anthony A. Amato

Image not available.
Play Video

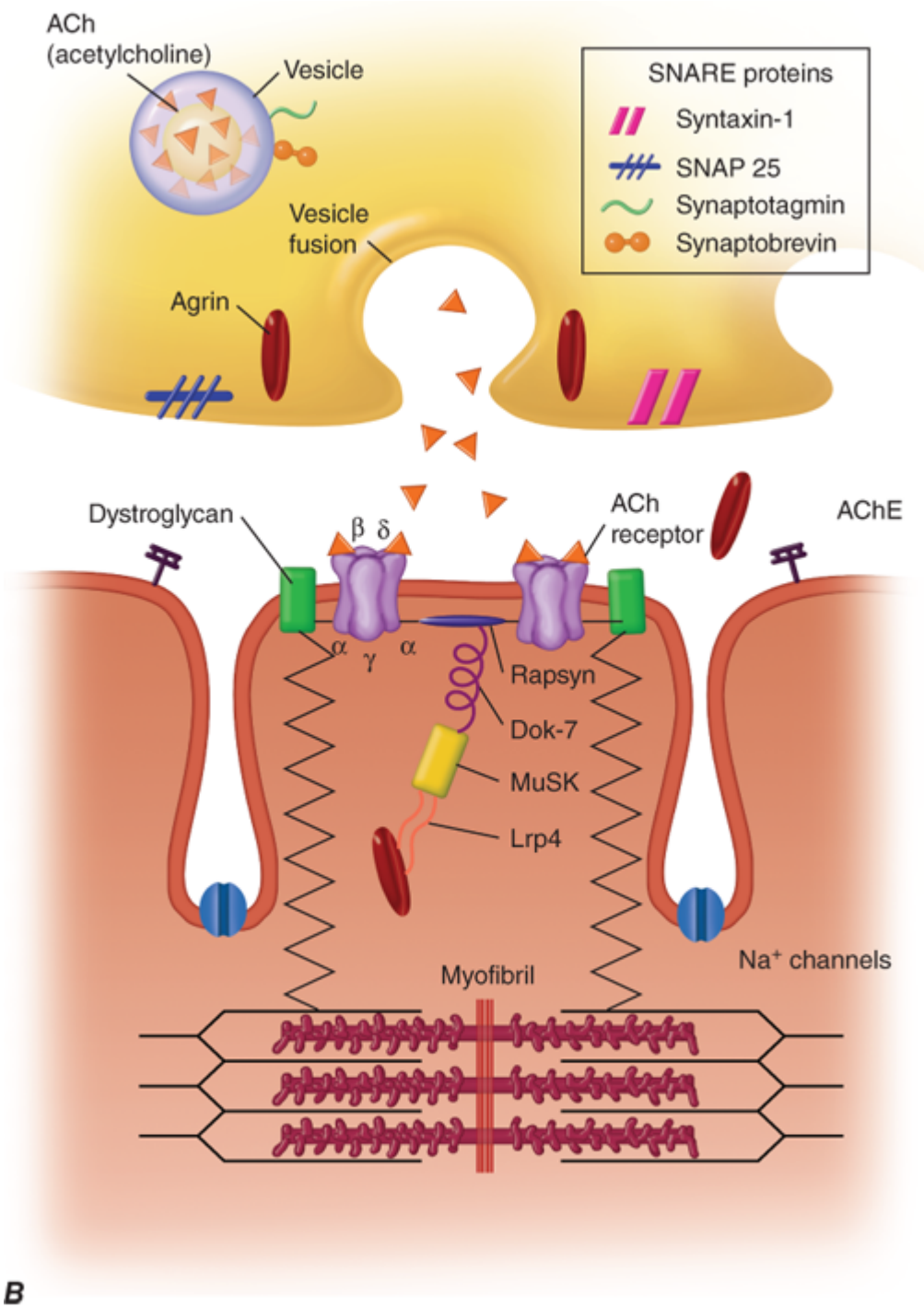
FIGURE 440-1

Illustrations of (*A*) a normal presynaptic neuromuscular junction, (*B*) a normal postsynapatic terminal, and (*C*) a myasthenic neuromuscular junction. AChE, acetylcholinesterase. See text for description of normal neuromuscular transmission. The myasthenia gravis (MG) junction demonstrates a reduced number of acetylcholine receptors (AChRs); flattened, simplified postsynaptic folds; and a widened synaptic space. See Video 440-1 also. *(From AA Amato, J Russell: Neuromuscular Disorders, 2nd ed. New York, McGraw-Hill, 2016, Figures 25-3 [p 588], 25-4 [p 589], and 25-5 [p 590]; with permission.*)



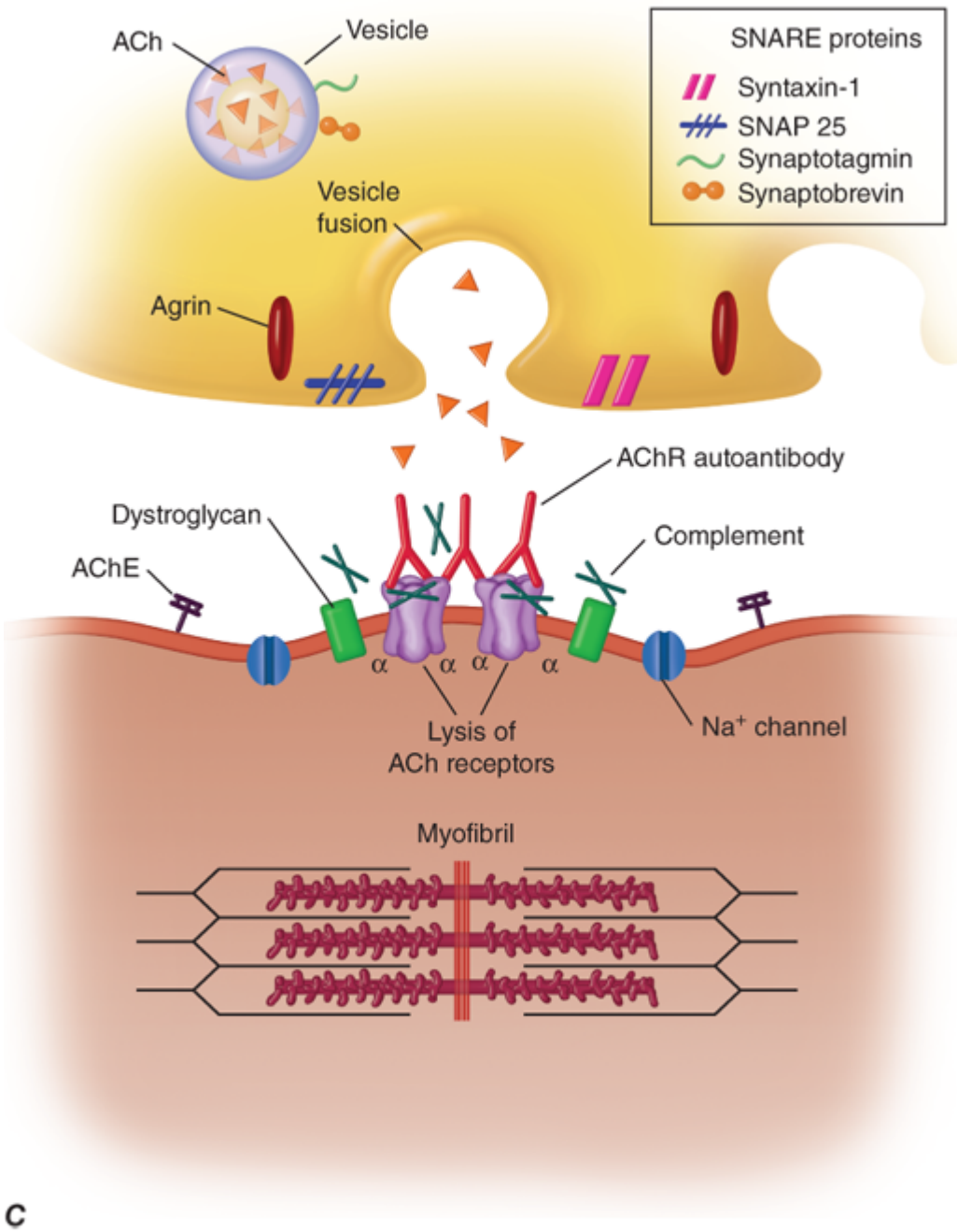
A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In MG, the fundamental defect is a decrease in the number of available AChRs at the postsynaptic muscle membrane. In addition, the postsynaptic folds are flattened, or “simplified.” These changes result in decreased efficiency of neuromuscular transmission. Therefore, although ACh is released normally, it produces small end-plate potentials that may fail to trigger muscle action potentials. Failure of transmission results in weakness of muscle contraction.

The amount of ACh released per impulse normally declines on repeated activity (termed *presynaptic rundown*). In the myasthenic patient, the decreased efficiency of neuromuscular transmission combined with the normal rundown results in the activation of fewer and fewer muscle fibers by successive nerve impulses and hence increasing weakness, or *myasthenic fatigue*. This mechanism also accounts for the decremental response to repetitive nerve stimulation seen on electrodiagnostic testing.

MG is an autoimmune disorder most commonly caused by anti-AChR antibodies. The anti-AChR antibodies reduce the number of available AChRs at NMJs by three distinct mechanisms: (1) accelerated turnover of AChRs by a mechanism involving cross-linking and rapid endocytosis of the receptors; (2) damage to the postsynaptic muscle membrane by the antibody in collaboration with complement; and (3) blockade of the active site of the AChR (i.e., the site that normally binds Ach). An immune response to muscle-specific kinase (MuSK), a protein involved in AChR clustering at the NMJ, can also result in MG, with reduction of AChRs demonstrated experimentally. Anti-MuSK antibody occurs in about 10% of patients (about 40% of AChR antibody negative patients), while 1–3% have antibodies to another protein at the NMJ—low-density lipoprotein receptor-related protein 4 (LRP4)—that is also important for clustering of AChRs. The pathogenic antibodies are IgG and are T cell dependent. Thus, immunotherapeutic strategies directed against either the antibody-producing B cells or helper T cells are effective in this antibody-mediated disease.

How the autoimmune response is initiated and maintained in MG is not completely understood, but the thymus appears to play a role in this process. The thymus is abnormal in ~75% of patients with AChR antibody–positive MG; in ~65% the thymus is “hyperplastic,” with the presence of active germinal centers detected histologically, although the hyperplastic thymus is not necessarily enlarged. An additional 10% of patients have thymic tumors (thymomas). Muscle-like cells within the thymus (myoid cells), which express AChRs on their surface, may serve as a source of autoantigen and trigger the autoimmune reaction within the thymus gland.

CLINICAL FEATURES

MG has a prevalence as high as 200 in 100,000. It affects individuals in all age groups, but peak incidences occur in women in their twenties and thirties and in men in their fifties and sixties. Overall, women are affected more frequently than men, in a ratio of ~3:2. The cardinal features are *weakness* and *fatigability* of muscles. The weakness increases during repeated use (fatigue) or late in the day and may improve following rest or sleep. The course of MG is often variable. Exacerbations and remissions may occur, particularly during the first few years after the onset of the disease. Unrelated infections or systemic disorders can lead to increased myasthenic weakness and may precipitate “crisis” (see below).

The distribution of muscle weakness often has a characteristic pattern. The cranial muscles, particularly the lids and extraocular muscles (EOMs), are typically involved early in the course of MG; diplopia and ptosis are common initial complaints. Facial weakness produces a “snarling” expression when the patient attempts to smile. Weakness in chewing is most noticeable after prolonged effort, as in chewing meat. Speech may have a nasal timbre caused by weakness of the palate or a dysarthric “mushy” quality due to tongue weakness. Difficulty in swallowing may occur as a result of weakness of the palate, tongue, or pharynx, giving rise to nasal regurgitation or aspiration of liquids or food. Bulbar weakness is especially prominent in MuSK antibody–positive MG. In ~85% of patients, the weakness becomes generalized, affecting the limb muscles as well. If weakness remains restricted to the EOMs for 3 years, it is likely that it will not become generalized, and these patients are said to have *ocular MG*. The limb weakness in MG is often proximal and may be

asymmetric. Despite the muscle weakness, deep tendon reflexes are preserved. If ventilatory weakness becomes requires respiratory assistance, the patient is said to be in *crisis*.

DIAGNOSIS AND EVALUATION

The diagnosis is suspected on the basis of weakness and fatigability in the typical distribution described above, without loss of reflexes or impairment of sensation or other neurologic function **(Table 440-1)**. The suspected diagnosis should always be confirmed definitively before treatment is undertaken; this is essential because (1) other treatable conditions may closely resemble MG and (2) the treatment of MG may involve surgery and the prolonged use of drugs with potentially adverse side effects.

TABLE 440-1

Diagnosis of Myasthenia Gravis (MG)

History
Diplopia, ptosis, dysarthria, dysphagia, dyspnea
Weakness in characteristic distribution: proximal limbs, neck extensors, generalized
Fluctuation and fatigue: worse with repeated activity, improved by rest
Effects of previous treatments
Physical examination
Evaluation for ptosis at rest and following one minute of exercise, extraocular muscles and subjective diplopia, orbicularis oculi and oris strength, jaw opening and closure
Assessment of muscle strength in neck and extremities
Weakness following repeated shoulder abduction
Vital capacity measurement
Absence of other neurologic signs
Laboratory testing
Anti-AChR radioimmunoassay: ~85% positive in generalized MG; 50% in ocular MG; definite diagnosis if positive; negative result does not exclude MG; ~40% of AChR antibody–negative patients with generalized MG have anti-MuSK antibodies
Repetitive nerve stimulation: decrement of >10% at 3 Hz: highly probable
Single-fiber electromyography: blocking and jitter, with normal fiber density; confirmatory, but not specific
Edrophonium chloride (Enlon®) 2 mg + 8 mg IV; highly probable diagnosis if unequivocally positive
For ocular or cranial MG: exclude intracranial lesions by CT or MRI

Abbreviations: AChR, acetylcholine receptor; CT, computed tomography; MRI, magnetic resonance imaging; MuSK, muscle-specific tyrosine kinase.

Ice-pack Test

If a patient has ptosis, application of a pack of ice over a ptotic eye often results in improvement if the ptosis is due to an NMJ defect. This is hypothesized to be due to less depletion of quanta of AChR in the cold and reduced activity of AChE at the NMJ. It is a quick and easy test to do in the clinic or at the bedside of a hospitalized patient.

Autoantibodies Associated with MG

As previously mentioned, anti-AChR antibodies are detectable in the serum of ~85% of all myasthenic patients but in only about 50% of patients with weakness confined to the ocular muscles. The presence of anti-AChR antibodies is virtually diagnostic of MG, but a negative test does not exclude the disease. The measured level of anti-AChR antibody does not correspond well with the severity of MG in different patients. Antibodies to MuSK are present in ~40% of AChR antibody–negative patients with generalized MG. MuSK antibodies are rarely present in AChR antibody–positive patients or in patients with MG limited to ocular muscles. These antibodies may interfere with clustering of AChRs at NMJs. A small proportion of MG patients without antibodies to AChR or MuSK have antibodies to LRP4. Interestingly, antibodies against agrin have recently been found in some patients with MG. Agrin is a protein derived from motor nerves that normally binds to LRP4 and important for normal clustering of AChRs at NMJ. Additionally, anti-striated muscle antibodies directed against titin and other skeletal muscle components are found in ~30% of myasthenic without thymoma, 24% of thymoma patients without myasthenia, and 70–80% of patients with both myasthenia and thymoma. Furthermore, antibodies directed against Netrin-1 receptors and Caspr2 (contactin-associated protein-like 2) often coexist and are associated in patients with thymoma who have MG and neuromyotonia or Morvan syndrome.

Electrodiagnostic Testing

Repetitive nerve stimulation may provide helpful diagnostic evidence of MG. Anti-AChE medication should be stopped 6–12 h before testing. It is best to test weak muscles or proximal muscle groups. Electrical stimulation is delivered at a rate of two or three per second to the appropriate nerves, and action potentials are recorded from the muscles. In normal individuals, the amplitude of the evoked muscle action potentials does not change by >10% at these rates of stimulation. However, in myasthenic patients, there is a rapid reduction of >10% in the amplitude of the evoked responses.

Anticholinesterase Test

Drugs that inhibit the enzyme AChE allow ACh to interact repeatedly with the limited number of AChRs in MG, producing improvement in muscle strength. Edrophonium is used most commonly for diagnostic testing

because of the rapid onset (30 s) and short duration (~5 min) of its effect. An objective end point must be selected to evaluate the effect of edrophonium, such as weakness of EOMs, impairment of speech, or the length of time that the patient can maintain the arms in forward. An initial IV dose of 2 mg of edrophonium is given. If definite improvement occurs, the test is considered positive and is terminated. If there is no change, the patient is given an additional 8 mg IV. The dose is administered in two parts because some patients react to edrophonium with side effects such as nausea, diarrhea, salivation, fasciculations, and rarely with severe symptoms of syncope or bradycardia. Atropine (0.6 mg) should be drawn up in a syringe and ready for IV administration if these symptoms become troublesome. The edrophonium test is now reserved for patients with clinical findings that are suggestive of MG but who have negative antibody, electrodiagnostic testing, or ice-pack test. False-positive tests occur in occasional patients with other neurologic disorders, such as amyotrophic lateral sclerosis **(Chap. 429)**, and in placebo-reactors. False-negative or equivocal tests may also occur.

Pulmonary Function Tests

Measurements of ventilatory function are valuable because of the frequency and seriousness of respiratory impairment in myasthenic patients **(Chap. 278)**.

Differential Diagnosis

Other conditions that cause weakness of the cranial and/or somatic musculature include the nonautoimmune congenital myasthenia, drug-induced myasthenia, Lambert-Eaton myasthenic syndrome (LEMS), neurasthenia, hyperthyroidism (Graves' disease), botulism, intracranial mass lesions, oculopharyngeal dystrophy, and mitochondrial myopathy (Kearns-Sayre syndrome, progressive external ophthalmoplegia). Treatment with penicillamine (used for scleroderma or rheumatoid arthritis) and check point inhibitors for cancer may also result in autoimmune MG. Aminoglycoside antibiotics or procainamide can cause exacerbation of weakness in myasthenic patients; very large doses can cause neuromuscular weakness in normal individuals.

The *congenital myasthenic syndromes (CMS) comprise* a rare heterogeneous group of disorders of the NMJ that are not autoimmune but rather are due to genetic mutations in which virtually any component of the NMJ may be affected. Alterations in function of the presynaptic nerve terminal, in the various subunits of the AChR, AChE, or the other molecules involved in end-plate development or maintenance, have been identified in the different forms of CMS. These disorders share many of the clinical features of autoimmune MG, including weakness and fatigability of proximal or distal extremity muscles, and often involving EOMs and the eyelids similar to the distribution in autoimmune MG. CMS should be suspected when symptoms of myasthenia have begun in infancy or childhood, but they can present in early adulthood. As in acquired autoimmune MG, repetitive nerve stimulation is associated with a decremental response. Some forms (e.g., AChE deficiency, prolonged open channel syndrome) have a feature of after-discharges which are not seen in MG. An additional clue is the absence of AChR and MuSK antibodies though these are absent in ~10% of generalized MG patients (so-called double seronegative MG) antibodies.

The prevalence of CMS is estimated at ~3.8 per 100,000. The most common genetic defects occur in the ε subunit of the AChR, accounting for ~50% of CMS cases, with mutations in the genes encoding for rapsin, COLQ, DOK7, agrin, and GFPT together accounting for ~40%. In most of the recessively inherited forms of CMS, the mutations are heteroallelic; that is, different mutations affecting each of the two alleles are present. Features of the most common forms of CMS are summarized in **Table 440-2**. Molecular analysis is required for precise elucidation of the defect; this may lead to helpful treatment as well as genetic counseling. Some forms of CMS improve with AChE inhibitors, while others (e.g., slow channel syndrome, AChE deficiency, DOK-7 related CMS) actually worsen. Fluoxetine and quinidine can be useful for slow channel syndrome, and albuterol for mutations affecting AChE, DOK-7, rapsyn, and agrin. Additionally, ephedrine and 3,4 diaminopyridine (3,4 DAP) may be of benefit in some forms of CMS.

TABLE 440-2

Congenital Myasthenic Syndromes

CMS Subtype	Gene	Clinical Features	Electrophysiological Features	Response to AChE Inhibitors	Treatment
Presynaptic Disorders					
CMS with paucity of ACh release	CHAT CHT	AR; early onset, respiratory failure at birth, episodic apnea, improvement with age	Decremental response to RNS	Improve	AChE inhibitors; 3,4 DAP
Synaptic Disorders					
AChE deficiency	COLQ	AR; early onset; variable severity; axial weakness with scoliosis; apnea; +/- EOM involvement, slow or absent pupillary responses	After discharges on nerve stimulation and decrement on RNS	Worsen	Albuterol; ephedrine; 3,4 DAP; avoid AChE inhibitors
Postsynaptic Disorders Involving AChR Deficiency or Kinetics					
Primary AChR deficiency	AChR subunit genes	AR; early onset; variable severity; fatigue; typical MG features	Decremental response to RNS	Improve	AChE inhibitors; 3,4 DAP
AChR kinetic disorder: slow channel syndrome	AChR subunit genes	AD; onset childhood to early adult; weak forearm extensors and neck; respiratory weakness; variable severity	After discharges on nerve stimulation and decrement on RNS	Worsen	Fluoxetine and quinidine; avoid AChE inhibitors

CMS Subtype	Gene	Clinical Features	Electrophysiological Features	Response to AChE Inhibitors	Treatment
AChR kinetic disorder: fast channel syndrome	AChR subunit genes	AR; early onset; mild to severe; ptosis, EOM involvement; weakness and fatigue	Decremental response to RNS	Improve	AChE inhibitors; caution with 3,4 DAP
Postsynaptic Disorders Involving Abnormal Clustering/Function of AChR					
	DOK 7	AR; limb girdle weakness with ptosis but no EOM involvement	Decremental response to RNS	Variable	Albuterol; ephedrine; may worsen with AChE inhibitors
	Rapsyn	AR; early onset with hypotonia, respiratory failure, and arthrogryposis at birth to early adult onset resembling MG	Decremental response to RNS	Variable	Albuterol
	Agrin	AR; limb girdle or distal weakness, apnea	Decremental response to RNS	Variable	Albuterol; may worsen with AChE inhibitors

CMS Subtype	Gene	Clinical Features	Electrophysiological Features	Response to AChE Inhibitors	Treatment
	MuSK	AR; congenital or childhood onset of ptosis, EOM and progressive limb girdle weakness	Decremental response to RNS	Variable	Variable response to AChE inhibitors and 3,4, DAP Positive response to albuterol
	LPR4	AR; congenital onset with hypotonia; ventilatory failure, mild ptosis, and EOM weakness	Decremental response to RNS	Worsen	Worsen with AChE inhibitors
Other Postsynaptic Disorders					
Limb-girdle CMS with tubular aggregates	GFPT1; DPAGT1; ALG2; ALG14; DPAGT1	AR; limb girdle weakness usually without ptosis or EOM weakness; onset in infancy or early adult	Decremental response to RNS	Variable	Albuterol; ephedrine; variable response to AChE inhibitors and 3,4, DAP; albuterol
Congenital muscular dystrophy with myasthenia	Plectin	AR; infantile or childhood onset of generalized weakness including ptosis and EOM; epidermolysis bullosa simplex; elevated CK	Decremental response to RNS	Variable	No response to AChE and 3,4 DAP

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; CHAT, choline acetyl transferase; CHT, sodium-dependent high-affinity choline transport 1; CMA, congenital myasthenic syndrome; COLQ, collaganic tail of endplate

acetylcholinesterase; 3,4 DAP, 3,4-diaminopyridine; Dok7, downstream of tyrosine kinase 7; DPAGT1, UDP-N-acetylglucosamine-dolichyl-phosphate N-acetylglucosamine phosphotransferase; GFPT1, glutamine-fructose-6-phosphate aminotransferase 1; LRP4, lipoprotein receptor-related protein 4; MuSK, muscle specific kinase; RNS, repetitive nerve stimulation

Source: From AA Amato, J Russell: *Neuromuscular Disorders*, 2nd ed. New York: McGraw-Hill, 2016, p 627, Figure 26-2; with permission.

LEMS is a presynaptic disorder of the NMJ that can cause weakness similar to that of MG. The proximal muscles of the lower limbs are most commonly affected, but other muscles may be involved as well. Cranial nerve findings, including ptosis of the eyelids and diplopia, occur in up to 70% of patients and resemble features of MG. However, the two conditions are usually readily distinguished because patients with LEMS have depressed or absent reflexes and experience autonomic changes such as dry mouth and impotence. Nerve stimulation produces an initial low-amplitude compound muscle action potential and, at low rates of repetitive stimulation (2–3 Hz), a decremental responses as seen in MG; however, at high rates (20–50 Hz), or following brief exercise, incremental responses occur. LEMS is caused by autoantibodies directed against P/Q-type calcium channels at the motor nerve terminals detected in ~85% of LEMS patients. These autoantibodies impair the release of ACh from nerve terminals. In young adults, particularly women, LEMS is not associated with an underlying cancer. However, in older adults, most LEMS is associated with malignancy, most commonly small-cell lung cancer (SCLC) The tumor cells may express calcium channels that stimulate the autoimmune response. Treatment of LEMS involves plasmapheresis and immunotherapy, as for MG. 3,4-Diaminopyridine (3,4-DAP) and pyridostigmine can also help with symptoms. 3,4-DAP acts by blocking potassium channels, which results in prolonged depolarization of the motor nerve terminals and thus enhances ACh release. Pyridostigmine prolongs the action of ACh, allowing repeated interactions with AChRs.

Botulism (Chap. 148) is due to potent bacterial toxins produced by any of eight different strains of *Clostridium botulinum*. The toxins enzymatically cleave specific proteins essential for the release of ACh from the motor nerve terminal, thereby interfering with neuromuscular transmission. Most commonly, botulism is caused by ingestion of improperly prepared food containing toxin. Rarely, the nearly ubiquitous spores of *C. botulinum* may germinate in wounds. In infants, the spores may germinate in the gastrointestinal (GI) tract and release toxin, causing muscle weakness. Patients present with myasthenia-like bulbar weakness (e.g., diplopia, dysarthria, dysphagia) and lack sensory symptoms and signs. Weakness may generalize to the limbs and may result in respiratory failure. Reflexes are present early, but they may be diminished as the disease progresses. Mentation is normal. Autonomic findings include paralytic ileus, constipation, urinary retention, dilated or poorly reactive pupils, and dry mouth. The demonstration of toxin in serum by bioassay is definitive, but the results usually take a relatively long time to be completed and may be negative. Nerve stimulation studies reveal reduced compound muscle action potential (CMAP) amplitudes that increase following high-frequency repetitive stimulation. Treatment includes ventilatory support and aggressive inpatient supportive care (e.g., nutrition, deep vein thrombosis prophylaxis) as needed. Antitoxin should be given as early as possible to be effective and can be obtained through the Centers for Disease Control and Prevention. A preventive vaccine is available for laboratory workers or other highly exposed individuals.

Neurasthenia is the historic term for a myasthenia-like fatigue syndrome without an organic basis. These patients may present with subjective symptoms of weakness and fatigue, but muscle testing usually reveals the "give-away weakness" characteristic of nonorganic disorders; the complaint of fatigue in these patients means tiredness or apathy rather than decreasing muscle power on repeated effort. *Hyperthyroidism* is readily diagnosed or excluded by tests of thyroid function, which should be carried out routinely in patients with suspected MG. Abnormalities of thyroid function (hyper- or hypothyroidism) may increase myasthenic weakness. Diplopia resembling that in MG may occasionally be due to an intracranial mass lesion that compresses nerves to the EOMs (e.g., sphenoid ridge meningioma), but magnetic resonance imaging (MRI) of the head and orbits usually reveals the lesion.

Progressive external ophthalmoplegia is a rare condition resulting in weakness of the EOMs, which may be accompanied by weakness of the proximal muscles of the limbs and other systemic features. Most patients with this condition have mitochondrial disorders that can be detected on muscle biopsy **(Chap. 441)**.

Search for Associated Conditions

Myasthenic patients have an increased incidence of several associated disorders **(Table 440-3)**. Thymic abnormalities occur in ~75% of AChR antibody–positive patients, as noted above. Neoplastic change (thymoma) may produce enlargement of the thymus, which is detected by chest computed tomography (CT). A thymic shadow on CT scan may normally be present through young adulthood, but enlargement of the thymus in a patient age >40 years is highly suspicious of thymoma. Hyperthyroidism occurs in 3–8% of patients and may aggravate the myasthenic weakness. Thyroid function tests should be obtained in all patients with suspected MG. Other autoimmune disorders, most commonly systemic lupus erythematosus and rheumatoid arthritis, can coexist with MG; associations also occur with neuromyelitis optica, neuromyotonia, Morvan's syndrome (encephalitis, insomnia, confusion, hallucinations, autonomic dysfunction, and neuromyotonia), rippling muscle disease, granulomatous myositis/myocarditis, and chronic inflammatory demyelinating polyneuropathy.

TABLE 440-3

Disorders Associated with Myasthenia Gravis and Recommended Laboratory Tests

Associated disorders
Disorders of the thymus: thymoma, hyperplasia
Other autoimmune neurological disorders: chronic inflammatory demyelinating polyneuropathy, neuromyelitis optica
Other autoimmune disorders: Hashimoto's thyroiditis, Graves' disease, rheumatoid arthritis, systemic lupus erythematosus, skin disorders, family history of autoimmune disorder
Disorders or circumstances that may exacerbate myasthenia gravis: hyperthyroidism or hypothyroidism, occult infection, medical treatment for other conditions (see Table 440-4)
Disorders that may interfere with therapy: tuberculosis, diabetes, peptic ulcer, gastrointestinal bleeding, renal disease, hypertension, asthma, osteoporosis, obesity
Recommended laboratory tests or procedures
CT or MRI of chest
Tests for antinuclear antibodies, rheumatoid factor
Thyroid function tests
Testing for tuberculosis
Fasting blood glucose, hemoglobin A1c
Pulmonary function tests
Bone densitometry

Abbreviations: CT, computed tomography; MRI, magnetic resonance imaging.

An infection of any kind can exacerbate typical MG, and should be sought carefully in patients with relapses. Because of the side effects of glucocorticoids and other immunotherapies used in the treatment of MG, a thorough medical investigation should be undertaken, searching specifically for evidence of chronic or latent infection (such as tuberculosis or hepatitis), hypertension, diabetes, renal disease, and glaucoma.

TREATMENT

TREATMENT

Myasthenia Gravis

The prognosis has improved strikingly as a result of advances in treatment. Nearly all myasthenic patients can be returned to full productive lives with proper therapy. The most useful treatments for MG include anticholinesterase medications, immunosuppressive agents, thymectomy, plasmapheresis, and intravenous immunoglobulin (IVIg) **(Fig. 440-2)**.

ANTICHOLINESTERASE MEDICATIONS

Anticholinesterase medication produces at least partial improvement in most myasthenic patients, although improvement is complete in only a few. Patients with anti-MuSK MG generally obtain less benefit from anticholinesterase agents than those with AChR antibodies and may actually worsen. Pyridostigmine is the most widely used anticholinesterase drug and is initiated at a dosage of 30–60 mg three to four times daily. The beneficial action of oral pyridostigmine begins within 15–30 min and lasts for 3–4 h, but individual responses vary. The frequency and amount of the dose should be tailored to the patient's individual requirements throughout the day. For example, patients with weakness in chewing and swallowing may benefit by taking the medication before meals so that peak strength coincides with mealtimes. Long-acting pyridostigmine may occasionally be useful to get the patient through the night but should not be used for daytime medication because of variable absorption. The maximum useful dose of pyridostigmine rarely exceeds 300 mg daily. Overdosage with anticholinesterase medication may cause increased weakness and other side effects. In some patients, muscarinic side effects of the anticholinesterase medication (diarrhea, abdominal cramps, salivation, nausea) may limit the dose tolerated. Atropine/diphenoxylate or loperamide is useful for the treatment of GI symptoms.

THYMECTOMY

Two separate issues should be distinguished: (1) surgical removal of thymoma, and (2) thymectomy as a treatment for MG. Surgical removal of a thymoma is necessary because of the possibility of local tumor spread, although most thymomas are histologically benign. Until recently there was a debate regarding the role of thymectomy in non-thymotous MG, but a recent large international trial of extended transternal thymectomy in non-thymomatous AChR antibody positive, generalized MG demonstrated that participants who underwent thymectomy had improved strength and function, required less prednisone and additions of second line agents (e.g., azathioprine), and fewer hospitalizations for exacerbations. Whether or not less invasive thymectomy may be beneficial is unknown. Also, patients with ocular myasthenia, MuSK-positive, and seronegative MG were excluded from the study; retrospective and anecdotal evidence suggest that these patients may not benefit from thymectomy. Thymectomy should never be carried out as an emergency procedure, but only when the patient is adequately prepared. If necessary, treatment with IVIg or plasmapheresis may be used before surgery to maximize strength in weak patients.

IMMUNOSUPPRESSION

Immunosuppression using one or more of the available agents is effective in nearly all patients with MG. The choice of drugs or other immunomodulatory treatments should be guided by the relative benefits and risks for the individual patient and the urgency of treatment. It is helpful to develop a treatment plan based on short-term, intermediate-term, and long-term objectives. For example, if immediate improvement is essential either because of the severity of weakness or because of the patient's need to return to activity as soon as possible, IVIg should be administered or plasmapheresis should be undertaken. For the intermediate term, glucocorticoids and cyclosporine or tacrolimus generally produce clinical improvement within a period of 1–3 months. The beneficial effects of azathioprine and mycophenolate mofetil usually begin after many months (as long as a year), but these drugs have advantages for the long-term treatment of patients with MG. There is a growing body of evidence that rituximab is effective in many MG patients, especially those with MuSK antibody.

Glucocorticoid Therapy

Glucocorticoids, when used properly, produce improvement in myasthenic weakness in the great majority of patients. To minimize adverse side effects, prednisone should be given in a single dose rather than in divided doses throughout the day. In patients with only mild or moderate weakness, the initial dose should be relatively low (15–25 mg/d) to avoid the early weakening that occurs in perhaps 10–15% of patients treated initially with a high-dose regimen. The dose is increased stepwise, as tolerated by the patient (usually by 5 mg/d at 2–3 day intervals), until there is marked clinical improvement or a dose of 50–60 mg/d is reached. In patients with more severe weakness and those already in the hospital, starting at a high dose is reasonable. Patients are maintained on the dose that seems to control their symptoms for about a month, and then the dosage is slowly tapered (no faster than 10 mg a month until on 20 mg daily and then by 2.5–5 mg a month) to determine the minimum effective dose, and close monitoring is required. Some patients are able to be managed without the addition of other immunotherapies. Patients on long-term glucocorticoid therapy must be followed carefully to prevent or treat adverse side effects. The most common errors in glucocorticoid treatment of myasthenic patients include (1) insufficient persistence—improvement may be delayed and gradual; (2) tapering the dosage too early, too rapidly, or excessively; and (3) lack of attention to prevention and treatment of side effects.

The management of patients treated with glucocorticoids is discussed in Chap. 379.

Other Immunotherapies

Mycophenolate mofetil, azathioprine, cyclosporine, tacrolimus, rituximab, and occasionally cyclophosphamide are effective in many patients, either alone or in various combinations.

Mycophenolate mofetil is widely used because of its presumed effectiveness and relative lack of side effects. A dose of 1–1.5 g bid is recommended. Its mechanism of action involves inhibition of purine synthesis by the de novo pathway. Since lymphocytes have only the de novo pathway, but lack the alternative salvage pathway that is present in all other cells, mycophenolate inhibits proliferation of lymphocytes but not proliferation of other cells. It does not kill or eliminate preexisting autoreactive lymphocytes, and therefore clinical improvement may be delayed for many months to a year, until the preexisting autoreactive lymphocytes die spontaneously. The advantage of mycophenolate lies in its relative lack of adverse side

effects, with only occasional production of GI symptoms, rare development of leukopenia, and very small risks of malignancy or progressive multifocal leukoencephalopathy inherent in nearly all immunosuppressive treatments. Although two published studies did not show positive outcomes, most experts attribute the negative results to flaws in the trial designs, and mycophenolate is widely used for long-term treatment of myasthenic patients.

Azathioprine has long been used for MG and a randomized, clinical trial demonstrated that it was effective in reducing the dosage of prednisone necessary to control symptoms. However, the beneficial effect of azathioprine can take a year or more to become evident. Approximately 10–15% of patients are unable to tolerate azathioprine because of idiosyncratic reactions consisting of flulike symptoms of fever and malaise, bone marrow suppression, or abnormalities of liver function. An initial dose of 50 mg/d is given for about a week to test for these side effects. If this dose is tolerated, it is increased gradually to about 2–3 mg/kg of total body weight, or until the white blood count falls to 3000–4000/μL. Allopurinol should never be used in combination with azathioprine because the two drugs share a common degradation pathway; the result may be severe bone marrow suppression due to increased effects of the azathioprine.

The calcineurin inhibitors cyclosporine and tacrolimus seem to be effective in MG and appear to work more rapidly than azathioprine and mycophenolate. However, they are associated with more frequent severe side effects including hypertension and nephrotoxicity. The usual dose of cyclosporine is 4–5 mg/kg per d, and the average dose of tacrolimus is 0.07–0.1 mg/kg per d, given in two equally divided doses. “Trough” blood levels are measured 12 h after the evening dose. The therapeutic range for the trough level of cyclosporine is 150–200 ng/L, and for tacrolimus, it is 5–15 ng/L.

Rituximab (Rituxan) is a monoclonal antibody that binds to the CD20 molecule on B lymphocytes. It has been widely used for the treatment of B cell lymphomas and has also proven successful in the treatment of several autoimmune diseases including rheumatoid arthritis, pemphigus, and some IgM-related neuropathies. There is an increasing literature on the benefit of rituximab in MG. It appears particularly effective in MuSK antibody–positive MG, although some patients with AChR antibody MG also respond. A large NIH sponsored trial is underway in AChR-positive MG. The usual dose is 1 g IV on two occasions 2 weeks apart. Periodically, a repeat course needs to be administered; some MuSK patients go 2–3 years between infusions.

Eculizumab is a monoclonal antibody that binds membrane attack complex and was beneficial in a small pilot study of MG patients. The results of a large phase 3 clinical trial were recently published and largely positive leading to recent FDA approval. The drug is administered intravenously every 2 weeks.

For the rare refractory MG patient, a course of high-dose cyclophosphamide may induce long-lasting benefit by “rebooting” the immune system. At high doses, cyclophosphamide eliminates mature lymphocytes but spares hematopoietic precursors (stem cells), because they express the enzyme aldehyde dehydrogenase, which hydrolyzes cyclophosphamide. This procedure is reserved for refractory patients and should be administered only in a facility fully familiar with this approach. Maintenance immunotherapy after rebooting is usually required to sustain the beneficial effect.

PLASMAPHERESIS AND INTRAVENOUS IMMUNOGLOBULIN

Plasmapheresis has been used therapeutically in MG. Plasma, which contains the pathogenic antibodies, is mechanically separated from the blood cells, which are returned to the patient. A course of five exchanges (3–4 L per exchange) is generally administered over a 10- to 14-day period. Plasmapheresis produces a short-term reduction in anti-AChR antibodies, with clinical improvement in many patients. It is useful as a temporary expedient in seriously affected patients or to improve the patient's condition prior to surgery (e.g., thymectomy).

The indications for the use of IVIg are the same as those for plasma exchange: to produce rapid improvement to help the patient through a difficult period of myasthenic weakness or prior to surgery. This treatment has the advantages of not requiring special equipment or large-bore venous access. The usual dose is 2 g/kg, which is typically administered >2–5 days. Improvement occurs in ~70% of patients, beginning during treatment or within a week, and continuing for weeks to months. The mechanism of action of IVIg is not known; the treatment has no consistent effect on the measurable amount of circulating AChR antibody. Adverse reactions are generally not serious but may include headache, fluid overload, and rarely aseptic meningitis or renal failure. IVIg or plasma exchange is occasionally used in combination with other immunosuppressive therapy for maintenance treatment of difficult MG.

MANAGEMENT OF MYASTHENIC CRISIS

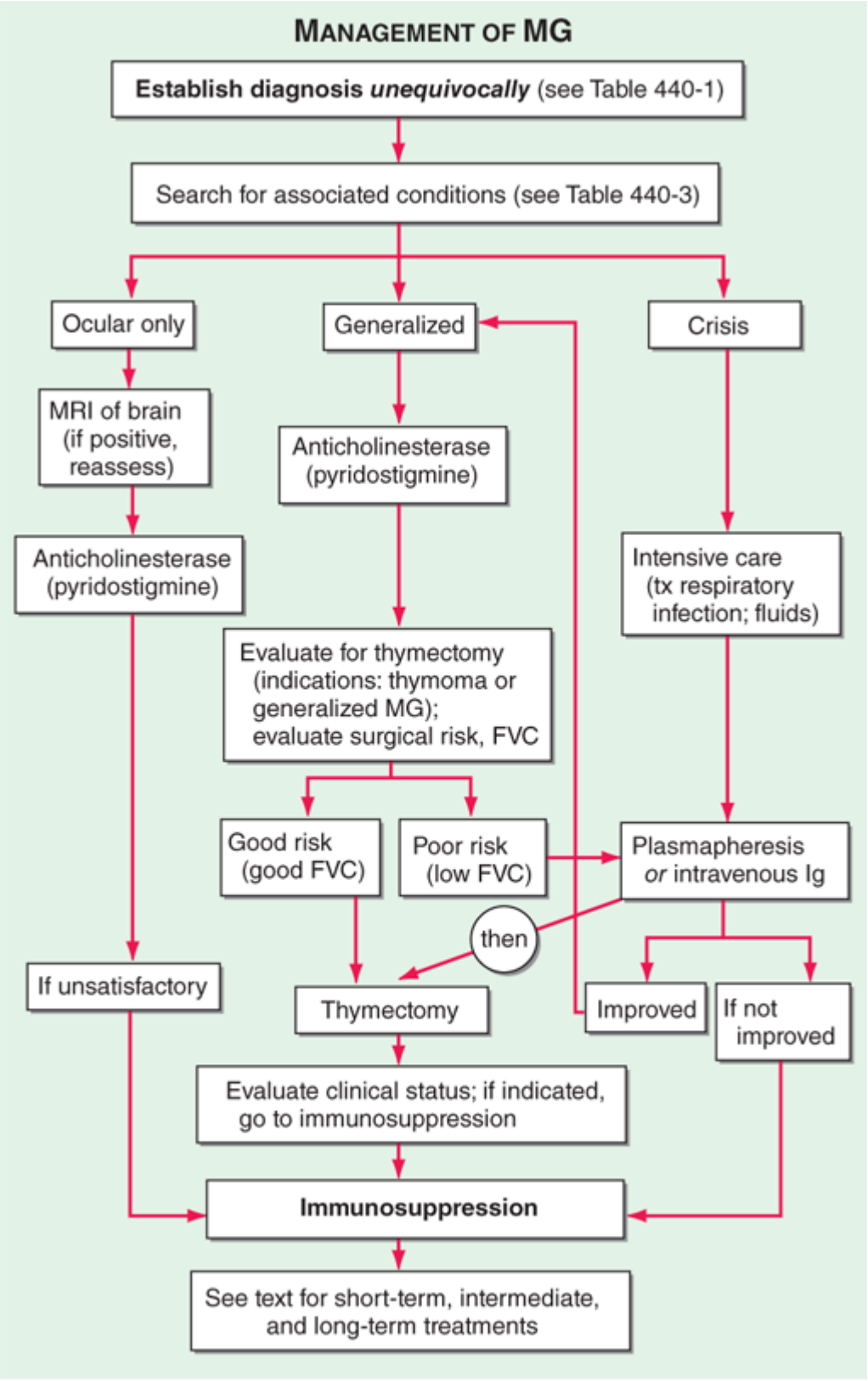
Myasthenic crisis is defined as an exacerbation of weakness sufficient to endanger life; it usually includes ventilatory failure caused by diaphragmatic and intercostal muscle weakness. Treatment should be carried out in intensive care units staffed with teams experienced in the management of MG. The possibility that deterioration could be due to excessive anticholinesterase medication ("cholinergic crisis") is best excluded by temporarily stopping anticholinesterase drugs. The most common cause of crisis is intercurrent infection. This should be treated immediately because the mechanical and immunologic defenses of the patient can be assumed to be compromised. The myasthenic patient with fever and early infection should be treated like other immunocompromised patients. Early and effective antibiotic therapy, ventilatory assistance, and pulmonary physiotherapy are essentials of the treatment program. As discussed above, plasmapheresis or IVIg is frequently helpful in hastening recovery.

DRUGS TO AVOID IN MYASTHENIC PATIENTS

Many drugs can potentially exacerbate weakness in patients with MG **(Table 440-4)**. As a rule, the listed drugs should be avoided whenever possible.

FIGURE 440-2

Algorithm for the management of myasthenia gravis. FVC, forced vital capacity; MRI, magnetic resonance imaging.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 440-4

Drugs with Interactions in Myasthenia Gravis (MG)

Drugs That May Exacerbate MG
Antibiotics
Aminoglycosides: e.g., streptomycin, tobramycin, kanamycin
Quinolones: e.g., ciprofloxacin, levofloxacin, ofloxacin, gatifloxacin
Macrolides: e.g., erythromycin, azithromycin
Nondepolarizing muscle relaxants for surgery
D-Tubocurarine (curare), pancuronium, vecuronium, atracurium
Beta-blocking agents
Propranolol, atenolol, metoprolol
Local anesthetics and related agents
Procaine, Xylocaine in large amounts
Procainamide (for arrhythmias)
Botulinum toxin
Botox exacerbates weakness
Quinine derivatives
Quinine, quinidine, chloroquine, mefloquine (Lariam)
Magnesium
Decreases acetylcholine release
Penicillamine
May cause MG

Check point inhibitors
May cause MG and other autoimmune neuromuscular disorders (e.g., myositis, inflammatory neuropathy)
Drugs with Important Interactions in MG
Cyclosporine and Tacrolimus
Broad range of drug interactions, which may raise or lower levels.
Azathioprine
Avoid allopurinol—combination may result in myelosuppression.

PATIENT ASSESSMENT

To evaluate the effectiveness of treatment as well as drug-induced side effects, it is important to assess the patient's clinical status systematically at baseline and on repeated interval examinations. Spirometry with determination of forced vital capacity and mean inspiratory and expiratory pressures are important to follow.

PROGNOSIS

Approximately 20% of patients with MG can be tapered off all immunotherapies and achieve a sustained remission. There does not appear to be a correlation with maximal disease severity and chance for remission. Thymectomy may increase the likelihood of achieving remission in anti-AChR MG, but the large randomized trial was too short in duration to examine this endpoint; rather, the results revealed only that thymectomy was efficacious and led to less use of glucocorticoids and second line agents. Mortality from MG diminished greatly during the twentieth century, changing from a "grave" illness with mortality of nearly 70% a century ago, to 2–30% by the 1950s, with contemporary estimates in the 1–2% range. Anti-MuSK patients typically do not experience myasthenic crises, but are generally more difficult to treat than anti-AChR MG. However, as noted above, recent series suggest that rituximab is effective in this subgroup. Non-paraneoplastic LEMS is usually responsive to immunotherapy and symptomatic treatment with pyridostigmine and 3,4 DAP. In older adults, LEMS is most often paraneoplastic, and screening for an underlying tumor is indicated. Recent studies suggest that survival in patients with LEMS has improved, for uncertain reasons and likely not due to earlier diagnosis and treatment of the tumor. There is wide variability in age of onset, severity, and prognosis of the many types of CMS.

GLOBAL ISSUES

The incidence of MG and its subtpes vary in different populations, for example occurring in ~2–10/10^6 individuals in the United States and the Netherlands, and up to 20/10^6 in Spain. Estimates of prevalence in different parts of the world range widely from 2–200/10^6. The age of onset may also be influenced by geographical and/or ethnic differences. Juvenile onset MG is uncommon in Western populations but may represent more than half of cases in Asians. MuSK MG appearts to be more common in the Mediterranean area of Europe than in northern Europe and is also more common in the northern regions of East Asia than in the southern regions.

FURTHER READING

Amato AA, Russell JA: *Neuromuscular Disorders*, 2nd ed. New York: McGraw-Hill; 2016, pp. 581–655.

Beeson D: Congenital myasthenic syndromes: Recent advances. Curr Opin Neurol 29:565, 2016. [PubMed: 27472506]

Garg N et al: Late presentations of congenital myasthenic syndromes: How many do we miss? Muscle Nerve 54:721, 2016. [PubMed: 26910802]

Gilhus NE: Myasthenia Gravis. N Engl J Med 375:2570, 2016. [PubMed: 28029925]

Howard JF Jr et al: Safety and efficacy of eculizumab in anti-acetylcholine receptor antibody-positive refractory generalised myasthenia gravis (REGAIN): A phase 3, randomised, double-blind, placebo-controlled, multicentre study. Lancet Neurol 16:976, 2017. [PubMed: 29066163]

Nicolle MW: Myasthenia gravis and Lambert-Eaton myasthenic syndrome. Continuum (Minneap Minn) 22(6, Muscle and neuromuscular junction disorders):1978, 2016. [PubMed: 27922503]

Tandan R et al: Rituximab treatment of myasthenia gravis: A systematic review. Muscle Nerve 56:185, 2017. [PubMed: 28164324]

Titulaer MJ et al: Lambert-Eaton myasthenic syndrome: From clinical characteristics to therapeutic strategies. Lancet Neurol 10:1098, 2011. [PubMed: 22094130]

Wolfe GI et al: Randomized trial of thymectomy in myasthenia gravis. N Engl J Med 375:511, 2016. [PubMed: 27509100]

Chapter 441: Muscular Dystrophies and Other Muscle Diseases

Anthony A. Amato; Robert H. Brown, Jr.

INTRODUCTION

Myopathies are disorders with structural changes or functional impairment of muscle and can be differentiated from other diseases of the motor unit (e.g., lower motor neuron or neuromuscular junction pathologies) by characteristic clinical and laboratory findings. **Myasthenia gravis and related disorders are discussed in Chap. 440; inflammatory myopathies are discussed in Chap. 358.**

CLINICAL FEATURES

The most important aspect of assessing individuals with neuromuscular disorders is taking a thorough history of the patient's symptoms, disease progression, past medical and family history as well as performing a detailed neurologic examination. Based on this and additional laboratory workup (e.g., serum creatine kinase [CK], electromyography [EMG]) one can usually localize the site of the lesion to muscle (as opposed to motor neurons, peripheral nerves, or neuromuscular junction) and the pattern of muscle involvement. It is this pattern of muscle involvement which is most useful in narrowing the differential diagnosis **(Table 441-1)**. Most myopathies present with proximal, symmetric limb weakness with preserved reflexes and sensation. However, asymmetric and predominantly distal weakness can be seen in some myopathies. An associated sensory loss suggests a peripheral neuropathy or a central nervous system (CNS) abnormality (e.g., myelopathy) rather than a myopathy. On occasion, disorders affecting the motor nerve cell bodies in the spinal cord (anterior horn cell disease), the neuromuscular junction, or peripheral nerves can mimic findings of myopathy.

TABLE 441-1

Myopathies by Pattern of Weakness/Muscle Involvement

Proximal (Limb-Girdle) Weakness
Most dystrophies (e.g., dystrophinophies, limb-girdle, myotonic dystrophy type 2, rare FSHD) Congenital myopathies (e.g., central core, multiminicore, centronuclear, nemaline rod) Metabolic myopathies (e.g., glycogen and lipid storage diseases) Mitochondrial myopathies Inflammatory myopathies (DM, PM, IMNM) Toxic myopathies (see Table 441-6) Endocrine myopathies Neuromuscular junction disorders (myasthenia gravis, LEMS, congenital myasthenia, botulism, **see Chap. 440**)
Distal Weakness
Distal muscular dystrophies/myofibrillar myopathy (see Table 441-5) Congenital myopathies (e.g., late onset centronuclear and nemaline rod myopathies) Metabolic Glygogen storage disease (e.g., brancher and debrancher deficiency, rarely McArdle disease) Lipid storage disease (e.g., neutral lipid storage myopathy, multiacyldehydrogenase deficiency) NMJ disorders (e.g., rare myasthenia gravis and congenital myasthenia)
Proximal Arm/Distal Leg Weakness (Scapuloperoneal or Humeroperonal) Weakness
Facioscapulohumeral muscular dystrophy (FSHD) Scapuloperoneal myopathy and neuropathy Myofibrillar myopathies Emery-Dreifuss muscular dystrophy (EDMD) Bethlem myopathy
Distal Arm/Proximal Leg Weakness
Inclusion body myositis (usually wrist and finger flexors in arms, hip flexors and knee extensors in legs, and asymmetric) Myotonic dystrophy (uncommon presentation)
Axial Muscle Weakness
Inflammatory (cervico-brachial myositis) sIBM and hIBM Myotonic dystrophy 2 Isolated neck extensor myopathy/bent spine syndrome FSHD
Late onset central core (RYR1 mutations) Sporadic lat onset nemaline rod with or without a monoclonal gammopathy Metabolic (late onset Pompe, McArdle disease, lipid storage, mitochondrial) Hyperparathyroidism/ostemalacia/vitamin D deficiency Myasthenia gravis
Eye Muscle Weakness (Ptosis/Ophthalmoparesis)
Ptosis without ophthalmoparesis Myotonic dystrophy Congenital myopathies Neuromuscular junction disorders Ptosis with ophthalmoparesis Oculopharyngeal dystrophy

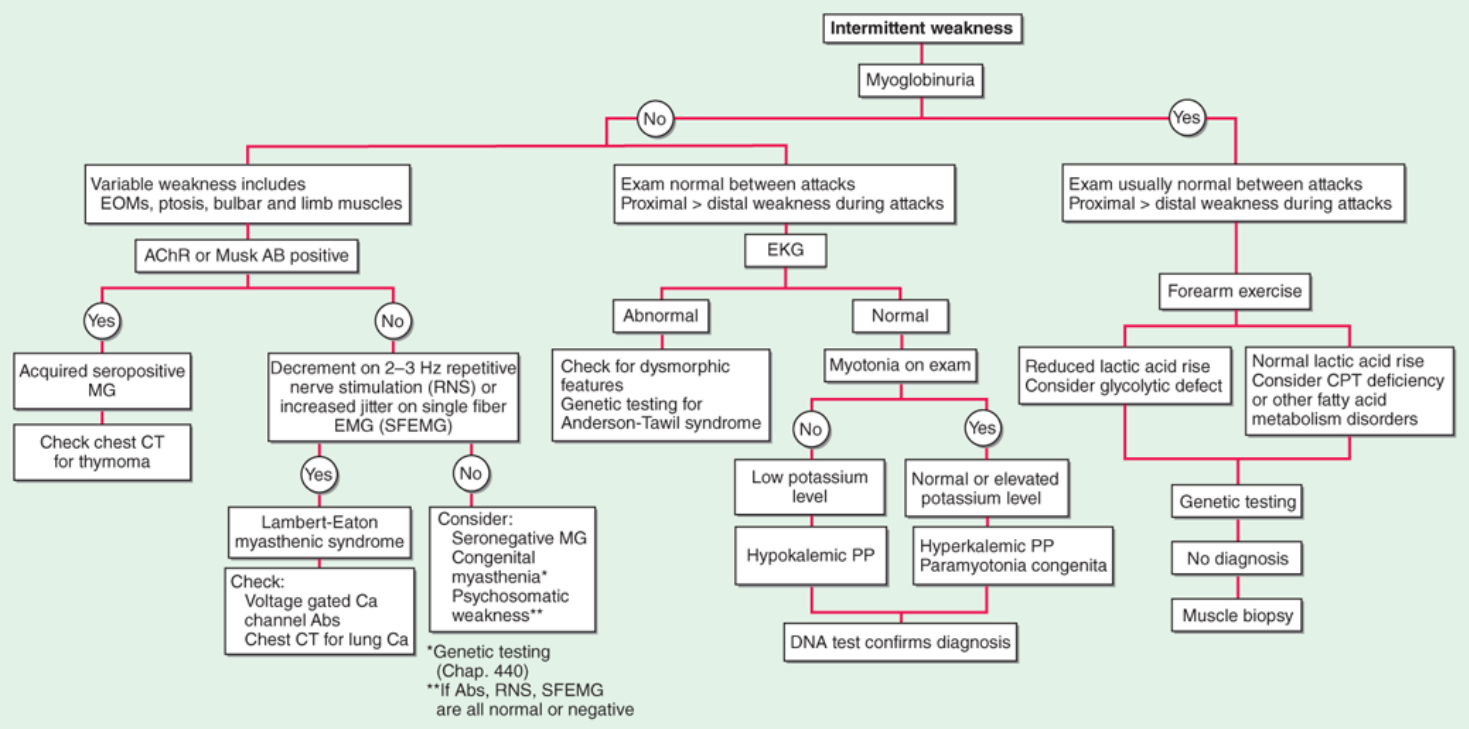
Mitochondrial myopathy hIBM type 3 Neuromuscular junction disorders
Episodic Weakness or Myoglobinuria
Related to exercise Glycogenoses (e.g., McArdle disease, etc.) Lipid disorders (e.g., CPT2 deficiency) Mitochondrial myopathies (e.g., cytrochrome B deficiency) Not Related to Exercise Malignant hyperthermia Drugs/toxins (e.g., statins) Prolonged/intensive eccentric exercise Inflammatory (e.g., PM/DM—rare, viral/bacterial infections) Delayed or unrelated to exercise Periodic paralysis (e.g., hereditary hyper- or hyokalemic, thyrotoxic, associated renal tubular acidosis, acquired electrolyte imbalance) NMJ disorders
Muscle Stiffness/Decreased Ability To Relax
Myotonic dystrophy 1 and 2 Myotonia congenita Paramyotonia congenita Hyperkalemic periodic paralysis with myotonia Potassium aggravated myotonia Schwartz-Jampel syndrome Other: rippling muscle disease (acquired and hereditary), acquired neuromyotonia (Isaacs syndrome), stiff-person syndrome, Brody disease

Muscle Weakness

Symptoms of muscle weakness can be either intermittent or persistent. Disorders causing *intermittent weakness* (Table 441-1 and **Fig. 441-1**) include myasthenia gravis, periodic paralyses (hypokalemic or hyperkalemic), and metabolic energy deficiencies of glycolysis (especially myophosphorylase deficiency), fatty acid utilization (carnitine palmitoyltransferase [CPT] deficiency), and some mitochondrial myopathies. The states of energy deficiency cause activity-related muscle breakdown accompanied by myoglobinuria.

FIGURE 441-1

Diagnostic evaluation of intermittent weakness. AChR AB, acetylcholine receptor antibody; CPT, carnitine palmitoyltransferase; EOMs, extraocular muscles; MG, myasthenia gravis; PP, periodic paralysis.

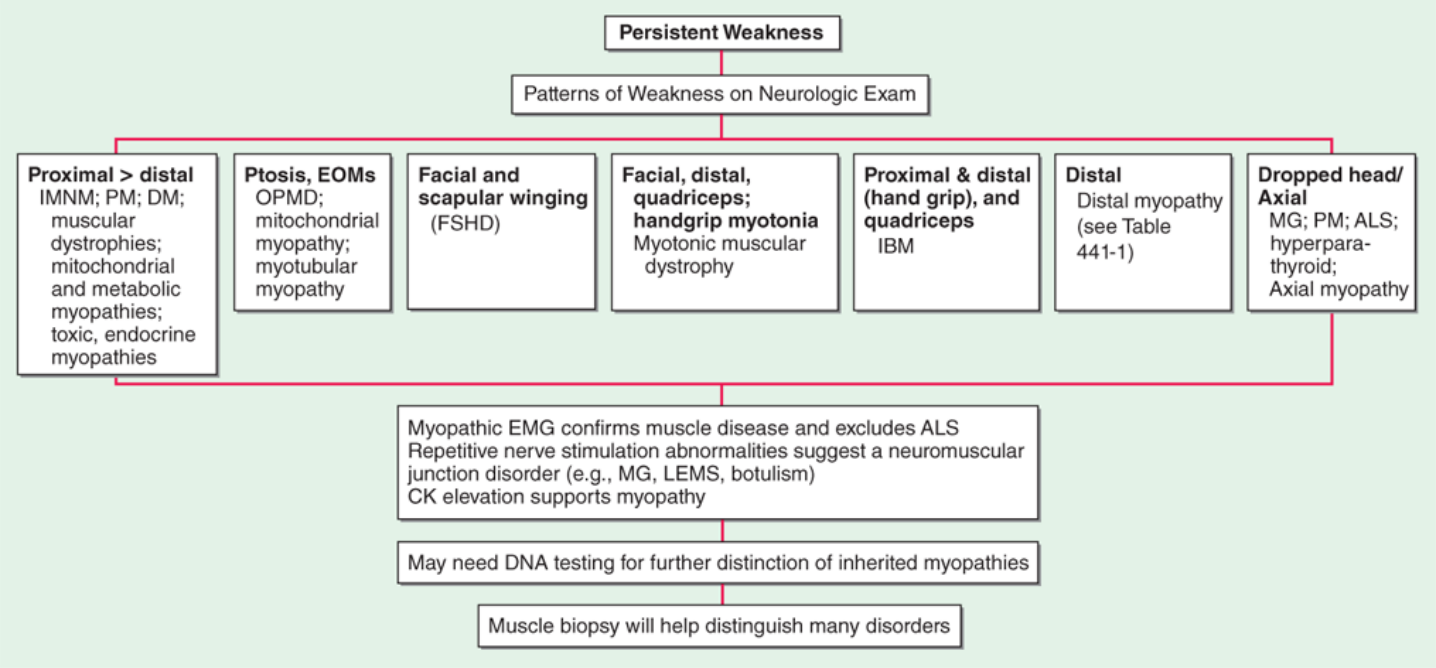


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Most muscle disorders cause *persistent weakness* (Table 441-1 and **Fig. 441-2**). In the majority of these, including most types of muscular dystrophy, polymyositis, and dermatomyositis, the proximal muscles are weaker than the distal and are symmetrically affected, and the facial muscles are spared, a pattern referred to as *limb-girdle weakness*. The differential diagnosis is more restricted for other patterns of weakness. Facial weakness (difficulty with eye closure and impaired smile) and scapular winging **(Fig. 441-3)** are characteristic of facioscapulohumeral dystrophy (FSHD). Facial and distal limb weakness associated with hand grip myotonia is virtually diagnostic of myotonic dystrophy type 1. When other cranial nerve muscles are weak, causing ptosis or extraocular muscle weakness, the most important disorders to consider include neuromuscular junction disorders, oculopharyngeal muscular dystrophy, mitochondrial myopathies, or some of the congenital myopathies (Table 441-1). A pathognomonic pattern characteristic of inclusion body myositis is atrophy and weakness of the flexor forearm (e.g., wrist and finger flexors) and quadriceps muscles that is often asymmetric. Less frequently, but important diagnostically, is the presence of a dropped head syndrome indicative of selective neck extensor muscle weakness. The most important neuromuscular diseases associated with this pattern of weakness include myasthenia gravis, amyotrophic lateral sclerosis, late-onset core or nemaline rod myopathy, hyperparathyroidism, focal myositis, and some forms of inclusion body myopathy. A final pattern, recognized because of preferential distal extremity weakness, is seen in the distal myopathies.

FIGURE 441-2

Diagnostic evaluation of persistent weakness. Examination reveals one of seven patterns of weakness. The pattern of weakness in combination with the laboratory evaluation leads to a diagnosis. ALS, amyotrophic lateral sclerosis; CK, creatine kinase; DM, dermatomyositis; EMG, electromyography; EOMs, extraocular muscles; FSHD, facioscapulohumeral dystrophy; IBM, inclusion body myositis; IMNM, immune-mediated necrotizing myopathy; MG, myasthenia gravis; OPMD, oculopharyngeal muscular dystrophy; PM, polymyositis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 441-3

Facioscapulohumeral dystrophy with prominent scapular winging.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

It is important to examine functional capabilities to help disclose certain patterns of weakness (Table 441-1 and **Table 441-2**). The Gower sign **(Fig. 441-4)** is particularly useful. Observing the gait of an individual may disclose a lordotic posture caused by combined trunk and hip weakness, frequently exaggerated by toe walking **(Fig. 441-5)**. A waddling gait is caused by the inability of weak hip muscles to prevent hip drop or hip dip. Hyperextension of the knee (genu recurvatum or back-kneeing) is characteristic of quadriceps muscle weakness; and a steppage gait, due to footdrop, accompanies distal weakness.

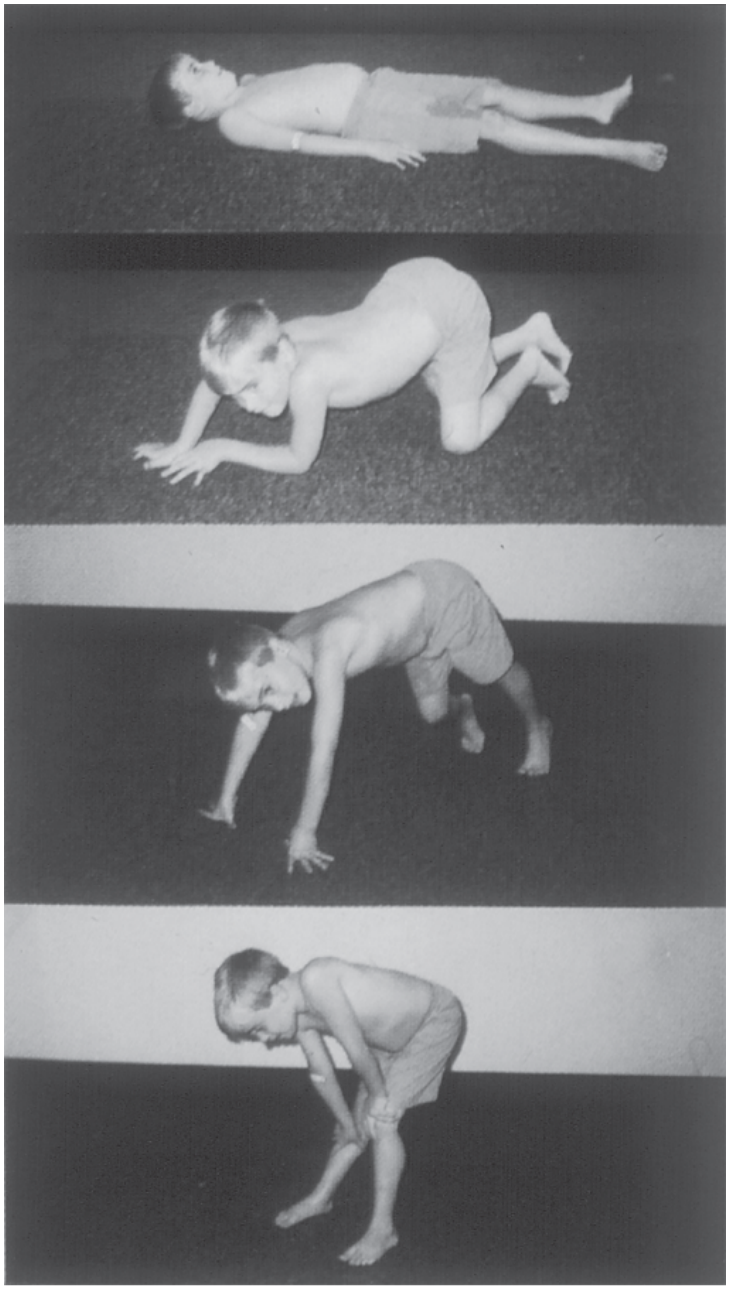
TABLE 441-2

Observations on Examination That Disclose Muscle Weakness

Functional Impairment	Muscle Weakness
Inability to forcibly close eyes	Upper facial muscles
Impaired pucker	Lower facial muscles
Inability to raise head from prone position	Neck extensor muscles
Inability to raise head from supine position	Neck flexor muscles
Inability to raise arms above head	Proximal arm muscles (may be only scapular stabilizing muscles)
Inability to walk without hyperextending knee (back-kneeing or genu recurvatum)	Knee extensor muscles
Inability to walk with heels touching the floor (toe walking)	Shortening of the Achilles tendon
Inability to lift foot while walking (steppage gait or footdrop)	Anterior compartment of leg
Inability to walk without a waddling gait	Hip muscles
Inability to get up from the floor without climbing up the extremities (Gowers' sign)	Hip, thigh, and trunk muscles
Inability to get up from a chair without using arms	Hip muscles

FIGURE 441-4

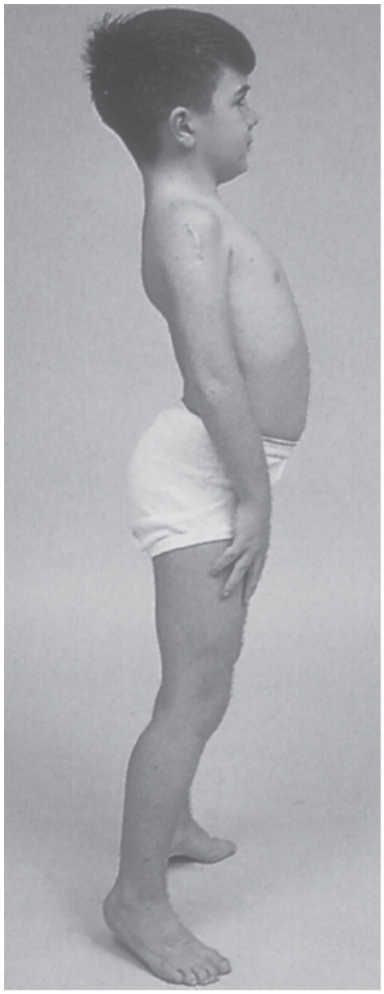
Gowers sign showing a patient using his arms to climb up the legs in attempting to get up from the floor.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 441-5

Lordotic posture, exaggerated by standing on toes, associated with trunk and hip weakness.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Any disorder causing muscle weakness may be accompanied by *fatigue*, referring to an inability to maintain or sustain a force (pathologic fatigability). This condition must be differentiated from asthenia, a type of fatigue caused by excess tiredness or lack of energy. Associated symptoms may help differentiate asthenia and pathologic fatigability. Asthenia is often accompanied by a tendency to avoid physical activities, complaints of daytime sleepiness, necessity for frequent naps, and difficulty concentrating on activities such as reading. There may be feelings of overwhelming stress and depression. In contrast, pathologic fatigability occurs in disorders of neuromuscular transmission and in disorders altering energy production, including defects in glycolysis, lipid metabolism, or mitochondrial energy production. Pathologic fatigability also occurs in chronic myopathies because of difficulty accomplishing a task with less muscle. Pathologic fatigability is accompanied by abnormal clinical or laboratory findings. Fatigue without those supportive features almost never indicates a primary muscle disease.

Muscle Pain (Myalgias), Cramps, and Stiffness

Some myopathies can be associated with muscle pain, cramps, contractures, stiff or rigid muscles, or inability to relax the muscles (e.g., myotonia) (Table 441-1). *Muscle cramps* are abrupt in onset, short in duration, triggered by voluntary muscle contraction, and may cause abnormal posturing of the joint. Muscle cramps often occur in neurogenic disorders, especially motor neuron disease **(Chap. 429)**, radiculopathies and polyneuropathies **(Chap. 438)**, but are not a feature of most primary muscle diseases.

A *muscle contracture* is different from a muscle cramp. In both conditions, the muscle becomes hard, but a contracture is associated with energy failure in glycolytic disorders. The muscle is unable to relax after an active muscle contraction. The EMG shows electrical silence. Confusion is created because contracture also refers to a muscle that cannot be passively stretched to its proper length (fixed contracture) because of fibrosis. In some muscle disorders, especially in Emery-Dreifuss muscular dystrophy and Bethlem myopathy, fixed contractures occur early and represent distinctive features of the disease.

Myotonia is a condition of prolonged muscle contraction followed by slow muscle relaxation. It always follows muscle activation (action myotonia), usually voluntary, but may be elicited by mechanical stimulation (percussion myotonia) of the muscle. Myotonia typically causes difficulty in releasing objects after a firm grasp. In myotonic muscular dystrophy type 1 (DM1), distal weakness usually accompanies myotonia, whereas in DM2, proximal muscles are more affected. Myotonia also occurs with *myotonia congenita* (a chloride channel disorder), but in this condition muscle weakness is not prominent. Myotonia may also be seen in individuals with sodium channel mutations (*hyperkalemic periodic paralysis* or *potassium-sensitive myotonia*). Another sodium channelopathy, *paramyotonia congenita* (PC), also is associated with muscle stiffness. In contrast to other disorders associated with myotonia in which the

myotonia is eased by repetitive activity, PC is named for a paradoxical phenomenon whereby the myotonia worsens with repetitive activity. Potassium-aggravated myotonia is an allelic disorder in which myotonia is brought on by consumption of too much potassium-containing foods.

Muscle stiffness can refer to different phenomena. Some patients with inflammation of joints and periarticular surfaces feel stiff. This condition is different from the disorders of hyperexcitable motor nerves causing stiff or rigid muscles. In *stiff-person syndrome*, spontaneous discharges of the motor neurons of the spinal cord cause involuntary muscle contractions mainly involving the axial (trunk) and proximal lower extremity muscles. The gait becomes stiff and labored, with hyperlordosis of the lumbar spine. Superimposed episodic muscle spasms are precipitated by sudden movements, unexpected noises, and emotional upset. The muscles relax during sleep. Serum antibodies against glutamic acid decarboxylase are present in approximately two-thirds of cases. In *acquired neuromyotonia* (*Isaacs syndrome*), there is hyperexcitability of the peripheral nerves manifesting as continuous muscle fiber activity in the form of widespread fasciculations and myokymia with impaired muscle relaxation. Muscles of the leg are stiff, and the constant contractions of the muscle cause increased sweating of the extremities. This peripheral nerve hyperexcitability is mediated by antibodies that target voltage-gated potassium channels.

There are two painful muscle conditions of particular importance, neither of which is associated with muscle weakness. *Fibromyalgia* is a common, yet poorly understood myofascial pain syndrome in which patients complain of severe muscle pain and tenderness, severe fatigue, and often poor sleep. Serum CK, erythrocyte sedimentation rate (ESR), EMG, and muscle biopsy are normal **(Chap. 366)**. *Polymyalgia rheumatica* occurs mainly in patients aged >50 years and is characterized by stiffness and pain in the shoulders, lower back, hips, and thighs **(Chap. 356)**. The ESR and CRP are elevated, while serum CK, EMG, and muscle biopsy are normal.

Muscle Enlargement and Atrophy

In most myopathies muscle tissue is replaced by fat and connective tissue, but the size of the muscle is usually not affected. However, in many limb-girdle muscular dystrophies, enlarged calf muscles are typical. The enlargement represents true muscle hypertrophy; thus the term *pseudohypertrophy* should be avoided when referring to these patients. The calf muscles remain very strong even late in the course of these disorders. Muscle enlargement can also result from infiltration by sarcoid granulomas, amyloid deposits, bacterial and parasitic infections, and focal myositis. In contrast, muscle atrophy is characteristic of other myopathies. In dysferlinopathies (LGMD2B) and anoctaminopathies (LGMD2L), there is a predilection for early atrophy of the gastrocnemius muscles, particularly the medial aspect. Atrophy of the humeral muscles is characteristic of FSHD and EDMD.

LABORATORY EVALUATION

Various tests can be used to evaluate a suspected myopathy, including CK levels, endocrine studies (e.g, thyroid function tests, parathyroid hormone and vitamin D levels), autoantibodies (associated with myositis and systemic disorders), forearm exercise test, muscle biopsy, and genetic testing. Electrodiagnostic studies can be useful to differentiate muscle disorders from other motor unit diseases but in most instances do not help distinguish the specific type of myopathy.

Serum Enzymes

CK is the most sensitive measure of muscle damage. The MM isoenzyme predominates in skeletal muscle, whereas creatine kinase-myocardial bound (CK-MB) is the marker for cardiac muscle. Serum CK can be elevated in normal individuals without provocation, presumably on a genetic basis or after strenuous activity, trauma, a prolonged muscle cramp, or a generalized seizure. Aspartate aminotransferase (AST), alanine aminotransferase (ALT), aldolase, and lactic dehydrogenase (LDH) are enzymes sharing an origin in both muscle and liver. Problems arise when the levels of these enzymes are found to be elevated in a routine screening battery, leading to the erroneous assumption that liver disease is present when in fact muscle could be the cause. An elevated γ-glutamyl transferase (GGT) helps to establish a liver origin because this enzyme is not found in muscle.

Electrodiagnostic Studies

EMG, repetitive nerve stimulation, and nerve conduction studies (NCS) **(Chap. 438)** are helpful in differentiating myopathies from neuropathies and neuromuscular junction diseases. Routine NCS are typically normal in myopathies but reduced amplitudes of compound muscle action potentials may be seen in atrophied muscles. The needle EMG may reveal irritability on needle insertion and spontaneously that is suggestive of a necrotizing myopathy (inflammatory myopathies, dystrophies, toxic myopathies, myotonic myopathies), whereas a lack of irritability is characteristic of long-standing myopathic disorders (muscular dystrophies with severe fibro-fatty replacement, endocrine myopathies, disuse atrophy, and many of the metabolic myopathies). In addition, the EMG may demonstrate myotonic discharges that will narrow the differential diagnosis (Table 441-1). Another important EMG finding is the presence of short-duration, small-amplitude, polyphasic motor unit action potentials (MUAPs). In myopathies, the MUAPs fire early but at a normal rate to compensate for the loss of individual muscle fibers, whereas in neurogenic disorders the MUAPs fire faster. An EMG is usually normal in steroid or disuse myopathy, both of which are associated with type 2 fiber atrophy; this is because the EMG preferentially assesses the physiologic function of type 1 fibers. The EMG can supplement the clinical examination in choosing an appropriately affected muscle to biopsy.

Genetic Testing

This is increasingly available and is the gold standard for diagnosing patients with hereditary myopathies.

Forearm Exercise Test

With exercise-induced muscle pain and myoglobinuria, there may be a defect in glycolysis. For safety, the test should not be performed under ischemic conditions to avoid an unnecessary insult to the muscle, causing rhabdomyolysis. The test is performed by placing a small indwelling catheter into an antecubital vein. A baseline blood sample is obtained for lactic acid and ammonia. The forearm muscles are exercised by asking the patient to vigorously open and close the hand for 1 min. Blood is then obtained at intervals of 1, 2, 4, 6, and 10 min for comparison with the baseline sample. A three- to fourfold rise of lactic acid is typical. The simultaneous measurement of ammonia serves as a control because it should also rise with exercise. In patients with myophosphorylase deficiency and certain other glycolytic defects, the lactic acid rise will be absent or below normal, while the rise in ammonia will reach control values. If there is lack of effort, neither lactic acid nor ammonia will rise. Patients with selective failure to increase ammonia may have myoadenylate deaminase deficiency. This condition has been reported to be a cause of myoglobinuria, but deficiency of this enzyme in asymptomatic individuals makes interpretation controversial.

Muscle Biopsy

Muscle biopsy is extremely helpful in evaluation of acquired myopathies but is performed less frequently in suspected hereditary myopathies as genetic testing has become more widely available. Almost any superficial muscle can be biopsied, but it is important to biopsy one that is affected clinically but not too severely (for example grade 4 out of 5 strength or movement against moderate resistance by manual muscle testing **[Chap. 415]**). A specific diagnosis can be established in many disorders.

HEREDITARY MYOPATHIES

Muscular dystrophy refers to a group of hereditary progressive diseases each with unique phenotypic and genetic features **(Tables 441-3, 441-4, 441-5, and Fig. 441-6)**. The prognosis of dystrophies is slow progressive weakness, though the severity and course is variable between and even within subtypes. Some are associated with cardiac and ventilatory muscle involvement, which are the leading cause of mortality. Unfortunately, there are no specific medical therapies for most of the muscular dystrophies and treatment is aimed at maintaining function with physical and occupational therapy. Non-invasive ventilation and tracheostomy may be warranted. Those with cardiomyopathy may require afterload reduction, antiarrhythmic agents, pacemakers or intracardiac defibrillators, and occasionally cardiac transplantation. We will focus primarily on those that manifest in adulthood.

TABLE 441-3

Autosomal Dominant Limb-Girdle Muscular Dystrophies (LGMDs)

Disease	Clinical Features	Laboratory Features	Abnormal Protein
LGMD1A	Onset second to eighth decade	Serum CK 2× normal	Myotilin
	Muscle weakness affects proximal and distal limb muscles, vocal cords, and pharyngeal muscles	EMG myopathic and may have pseudomotonic discharges	
		Muscle biopsy: features of MFM	
LGMD1B	Onset first or second decade	Serum CK 3–5× normal	Lamin A/C
	Proximal lower limb weakness and cardiomyopathy with conduction defects	EMG myopathic	
	Some cases indistinguishable from Emery-Dreifuss muscular dystrophy with joint contractures		
LGMD1C	Onset in early childhood	Serum CK 4–25× normal	Caveolin-3
	Proximal weakness	EMG myopathic	
	Gower sign, calf hypertrophy, rippling muscles		
	Exercise-related muscle cramps		
LGMD1D	Onset second to sixth decade	Serum CK 2–3× normal	DNAJB6
	Proximal and distal muscle weakness	EMG myopathic	
		Muscle biopsy: features of MFM	
LGMD1E	Onset first to sixth decade	Serum CK 2–4× normal	Desmin
	Proximal or distal muscle weakness	EMG myopathic and may have pseudomotonic discharges	
	Cardiomyopathy and arrhythmias	Muscle biopsy: features of MFM	
LGMD1F	Onset infancy to sixth decade	Serum CK normal to 20× normal	TNPO3
	Proximal or distal weakness	EMG myopathic	
	May have early contractures resembling Emery-Dreifuss syndrome	Muscle biopsy may show enlarged nuclei with central pallor, rimmed vacuoles, and filamentous inclusions	
LGMD1G	Onset teens to sixth decade	Muscle biopsies show rimmed vaculoes	HNRNPDL
	Proximal weakness contractures of fingers and toes	CK normal to 9× normal	
LGMD1H	Onset teens to sixth decade. Proximal weakness	Muscle biopsy with many COX negative/SDH positive fibers	
LGMD1I	Onset in teens to eighties, proximal weakness, scapular winging	CK normal to 50× normal	Calpain-3
	Allelic to LGMD2A but milder phenotype	EMG myopathic	

Abbreviations: CK, creatine kinase; EMG, electromyography; MFM, myofibrillar myopathy; NCS, nerve conduction studies; HNRNPDL, heterogeneous nuclear ribonucleoprotein D-like protein.

TABLE 441-4

Autosomal Recessive Limb-Girdle Muscular Dystrophies (LGMDs)

Disease	Clinical Features	Laboratory Features	Abnormal Protein
LGMD2A	Onset first or second decade	Serum CK 3–15× normal	Calpain-3
	Scapular winging; no calf hypertrophy; no cardiac or respiratory muscle weakness	EMG myopathic	
	Proximal and distal weakness; may have contractures at elbows, wrists, and fingers	Muscle biopsy may show lobulated muscle fibers	
LGMD2B	Onset second or third decade	Serum CK 3–100× normal	Dysferlin
	Proximal muscle weakness at onset, later distal (calf) muscles affected	EMG myopathic	
	Miyoshi's myopathy is variant of LGMD2B with calf muscles affected at onset	Inflammation on muscle biopsy may simulate polymyositis	
LGMD2C–F	Onset in childhood to teenage years	Serum CK 5–100× normal	γ, α, β, δ sarcoglycans
	Clinical condition similar to Duchenne and Becker muscular dystrophies	EMG myopathic	
	Cognitive function normal		
LGMD2G	Onset age 10–15	Serum CK 3–17× normal	Telethonin
	Proximal and distal muscle weakness	EMG myopathic	
		Muscle biopsy may show rimmed vacuoles	
LGMD2H	Onset first to third decade	Serum CK 2–25× normal	*TRIM32* gene
	Allelic to sarcotubular congenital myopathy	EMG myopathic	
	Proximal muscle weakness	Muscle biopsy reveals dilated T-tubules	
LGMD2I	Onset first to third decade	Serum CK 10–30× normal	Fukutin-related protein
	Clinical condition similar to Duchenne or Becker dystrophies	EMG myopathic	
	Cardiomyopathy and respiratory failure may occur early before significant weakness		
	Cognitive function normal		
LGMD2J[a]	Onset first to third decade	Serum CK 1.5–2× normal	Titin
	Proximal lower limb weakness	EMG myopathic	
	Mild distal weakness	Muscle biopsy reveals rimmed vacuoles	
	Progressive weakness causes loss of ambulation		
LGMD2K	Usually presents in infancy as Walker-Warburg syndrome but can present in early adult life with proximal weakness and only minor CNS abnormalities	CK 10–20× normal EMG myopathic	POMT1

Disease	Clinical Features	Laboratory Features	Abnormal Protein
LGMD2L	Presents in childhood or adult life	CK 8–20× normal	Anoctamin 5
	May manifest with quadriceps atrophy and myalgia	EMG myopathic	
	Some present with early involvement of the calves in the second decade of life, resembling Miyoshi myopathy type 1 (dysferlinopathy)		
LGMD2M	Usually presents in infancy as Fukuyama congenital muscular dystrophy but can present in early adult life with proximal weakness and only minor CNS abnormalities	CK 10–50× normal EMG myopathic	Fukutin
LGMD2N	Usually presents in infancy as muscle-eye-brain disease but can present in early adult life with proximal weakness and only minor CNS abnormalities	CK 5–20× normal EMG myopathic	POMGnT1
LGMD2O	Usually presents in infancy as Walker-Warburg syndrome but can present in early adult life with proximal weakness and only minor CNS abnormalities	CK 5–20× normal EMG myopathic	POMT2
LGMD2P	One case reported presenting in early childhood	CK >10× normal	α-Dystroglycan
LGMD2Q	Onset in infancy to fourth decade; proximal weakness; may have ptosis and extraocular weakness; epidermolysis bullosa (also considered a congenital myasthenic syndrome)	CK variable, but usually only mildly elevated EMG myopathic Repetitive nerve stimulation may show decrement	Plectin 1
LGMD2R	See LGMD1E (Table 441-6)	See LGMD1E	Desmin
LGMD2S	Onset in infancy to sixth decade	CK 1.5–20× normal	TRAPC11
	Proximal weakness		
	Eye abnormalities common; truncal ataxia and chorea		
	Mild to moderate intellectual disability		
	Hutterite descent		
LGMD2T	Onset in early childhood to fourth decade	CK3-> 10× normal	GMPPB
	Proximal weakness	EMG myopathic	
	CNS abnormalities, cataracts, cardiomyopathy, and neuromuscular junction dysfunction		

[a]Udd type distal myopathy is a form of titin deficiency with only distal muscle weakness (see Table 441-9).

Abbreviations: CK, creatine kinase; EMG, electromyography; GMPPB; guanosine diphosphate (GDP)-mannose pyrophosphorylase B; NCS, nerve conduction studies; POMGNT1, O-linked mannose beta 1,2-N-acetylglucosaminyltransferase; POMT1, protein-O-mannosyltransferase 1; POMT2, protein-O-mannosyltransferase 2; TNPO3, transportin 3; TRAPC11, transport (trafficking) protein particle complex, subunit 11.

TABLE 441-5

Distal Myopathies

Disease	Clinical Features	Laboratory Features	Abnormal Protein
Welander distal myopathy	Onset in fifth decade Weakness begins in hands Slow progression with spread to distal lower extremities Lifespan normal	Serum CK 2–3× normal EMG myopathic NCS normal Muscle biopsy shows dystrophic features and rimmed vacuoles	AD TIA1
Tibial muscular dystrophy (Udd)	Onset fourth to eighth decade Distal lower extremity weakness (tibial distribution) Upper extremities usually normal Lifespan normal	Serum CK 2–4× normal EMG myopathic NCS normal Muscle biopsy shows dystrophic features and rimmed vacuoles Titin absent in M-line of muscle	AD Titin AR (associated with more proximal weakness—LGMD2J)
Markesbery-Griggs distal myopathy	Onset fourth to eighth decade Distal lower extremity weakness (tibial distribution) with progression to distal arms and proximal muscles	Serum CK is usually mildly elevated EMG reveals irritative myopathy Muscle biopsies demonstrate rimmed vacuoles and features of MFM	AD Z-band alternatively spliced PDX motif-containing protein (ZASP)
Laing distal myopathy	Onset childhood to third decade Distal lower extremity weakness (anterior tibial distribution) and neck flexors affected early May have cardiomyopathy	Serum CK is normal or slightly elevated Muscle biopsies do not typically show rimmed vacuoles, but may show hyaline bodies with accumulation of myosin Large deposits of myosin heavy chain are seen in type 1 muscle fibers	AD Myosin heavy chain 7
GNE myopathy (Nonaka distal myopathy and autosomal recessive hereditary inclusion body myopathy)	Onset: second to third decade Distal lower extremity weakness (anterior tibial distribution) Mild distal upper limb weakness may be present early Progression to other muscles sparing quadriceps Ambulation may be lost in 10–15 y	Serum CK 3–10× normal EMG myopathic NCS normal Dystrophic features on muscle biopsy plus rimmed vacuoles and 15- to 19-nm filaments within vacuoles	AR *GNE* gene: UDP-*N*-acetylglucosamine 2-epimerase/*N*-acetylmannosamine kinase Allelic to hereditary inclusion body myopathy
Miyoshi myopathy[a]	Onset: second to third decade Lower extremity weakness in posterior compartment muscles Progression leads to weakness in other muscle groups Ambulation lost after 10–15 y in about one-third of cases	Serum CK 20–100× normal EMG myopathic NCS normal Muscle biopsy shows nonspecific dystrophic features often with prominent inflammatory cell infiltration; no rimmed vacuoles	AR Dysferlin (allelic to LGMD2B) ANO-5 (allelic to LGMD2L)
Williams myopathy	Distal lower extremity weakness (anterior tibial distribution)	Muscle biopsy may show rimmed vacuoles and features of MFM	X-linked Filamin-C

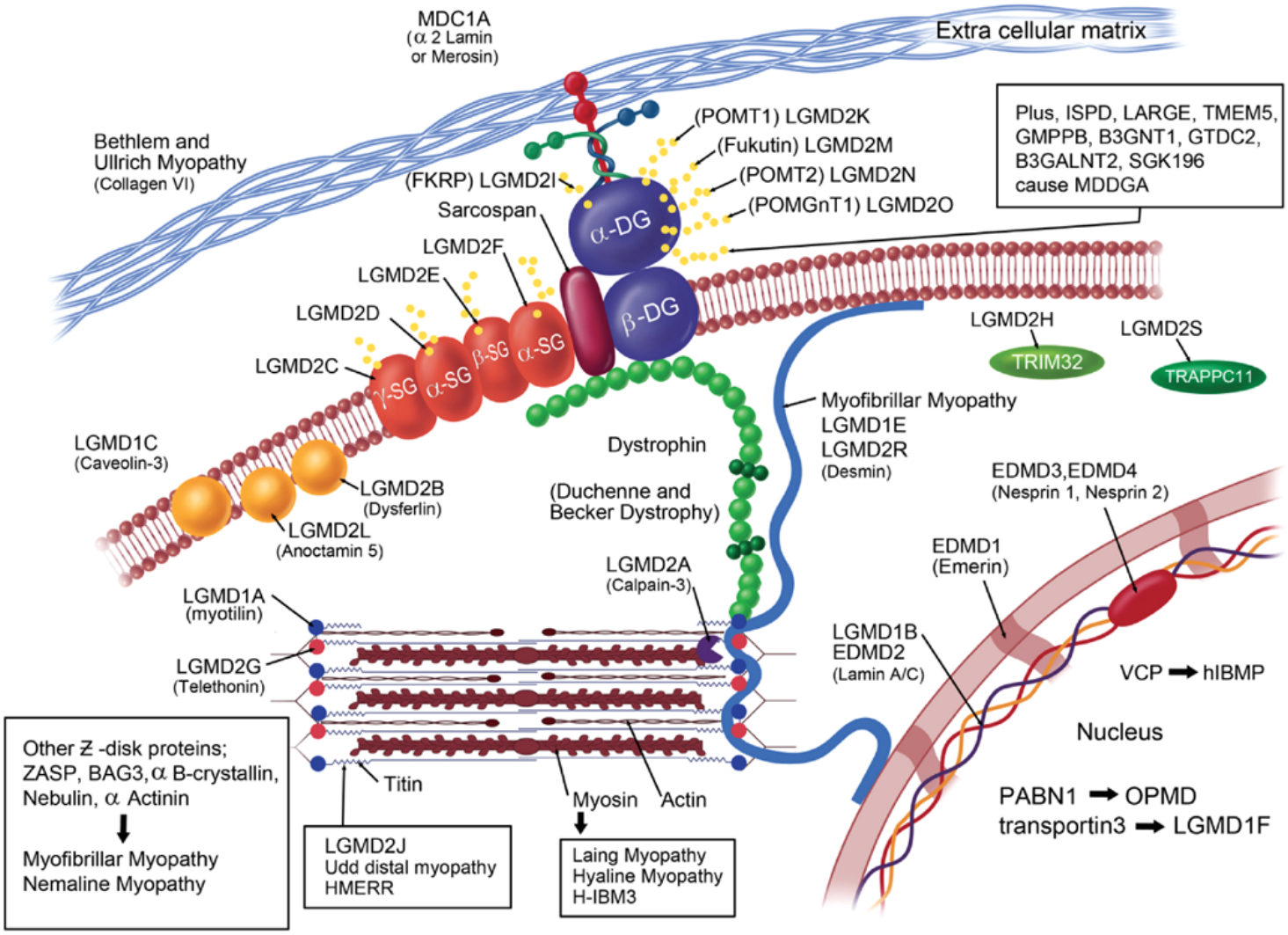
Disease	Clinical Features	Laboratory Features	Abnormal Protein
Myofibrillar myopathies	Onset from early childhood to late adult life Weakness may be distal, proximal, or generalized Cardiomyopathy and respiratory involvement is not uncommon	Serum CKs can be normal or moderately elevated EMG is myopathic and often associated with myotonic discharges Muscle biopsy demonstrates abnormal accumulation of desmin and other proteins, rimmed vacuoles, and myofibrillar degeneration	Genetically heterogeneous AD Myotilin (also known as LGMD 1A) ZASP (see Markesbery-Griggs distal myopathy) Filamin-C Desmin Alpha B crystallin Bag3 Titin DNAJB6 TNPO3 AR Desmin X-linked FHL1

[a]Miyoshi myopathy phenotype may also be seen with mutations in ANO-5 that encodes for anoctamin 5 (allelic to LGMD2L).

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; CK, creatine kinase; EMG, electromyography; MFM, myofibrillar myopathy; NCS, nerve conduction studies.

FIGURE 441-6

Proteins involved in the muscular dystrophies. This schematic shows the location of various sarcolemmal, sarcomeric, nuclear, and enzymatic proteins associated with muscular dystrophies. The diseases associated with mutations in the genes responsible for encoding these proteins are shown in boxes. Dystrophin, via its interaction with the dystroglycan complex, connects the actin cytoskeleton to the extracellular matrix. Extracellularly, the sarcoglycan complex interacts with biglycan, which connects this complex to the dystroglycan complex and the extracellular matrix collagen. Various enzymes are important in the glycosylation of the α-dystroglycan and mediate its binding to the extracellular matrix and usually cause a congenital muscular dystrophy with severe brain and eye abnormalities but may cause milder LGMD phenotype. Mutations in genes that encode for sarcomeric and Z-disk proteins cause forms of LGMD and distal myopathies (including myofibrillar myopathy, forms of hereditary inclusion body myopathy) as well as nemaline rod myopathy and other "congenital" myopathies. Mutations affecting nuclear membrane proteins are responsible for most forms of EDMD. Mutations in other nuclear genes cause other forms of dystrophy. *(Used with permission from AA Amato, J Russell (eds): Neuromuscular Disorders, 2nd ed. New York, McGraw-Hill, 2016, Figure 27-1, p 657.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

DUCHENNE AND BECKER MUSCULAR DYSTROPHY (DMD AND BMD)

DMD and BMD are X-linked recessive muscular dystrophies caused by mutations in the *dystrophin* gene. Affecting 1/3,000 male births, DMD is the most common mutational disease affecting boys. The incidence of BMD is ~5 per 100,000.

Clinical Features

Proximal muscles, especially of the lower extremities, are prominently involved in both disorders. This becomes evident in DMD very early; boys with DMD have difficulty climbing stairs and never run well. As the disease progresses, weakness becomes more generalized. Hypertrophy of muscles, particularly in the calves, is an early and prominent finding. Most patients with BMD first experience difficulties between ages 5 and 15 years, although onset in the third or fourth decade or even later can occur. Life expectancy for DMD and BMD is reduced, but most BMD cases survive into the fourth or fifth decade. Mental retardation may occur in both disorders, but is less common in BMD. Cardiac involvement is common in both DMD and BMD and may result in heart failure; some BMD patients manifest with only heart failure. Other less common presentations of dystrophinopathy are asymptomatic hyper-CK-emia, myalgias without weakness, and myoglobinuria.

Laboratory Features

Serum CK levels are usually elevated. Muscle biopsies demonstrate dystrophic features. Western blot analysis of muscle biopsy samples demonstrate absent dystrophin in DMD, or reduction in levels or size of dystrophin in BMD. In both disorders, mutations can be established using DNA from peripheral blood leukocytes. In most cases, muscle biopsies are no longer performed when DMD or BMD is suspected, as genetic testing is less invasive, less costly, and routinely available. Deletions within or duplications of the *dystrophin* gene are common in both DMD and BMD; in ~95% of cases, the mutation does not alter the translational reading frame of messenger RNA. These “in-frame” mutations allow for production of some dystrophin, which accounts for the presence of altered rather than absent dystrophin on Western blot analysis and a milder clinical phenotype.

TREATMENT

TREATMENT

Duchenne and Becker Muscular Dystrophy

Glucocorticoids slow progression in DMD but their use has not been adequately studied in Becker dystrophy. Physical and occupational therapy are important in helping maintain function. As patients often die from the associated cardiomyopathy, it is important to follow patients with a cardiologist and treat appropriately. Recent studies suggest that there is clinical benefit in selected cases of DMD from short oligonucleotides that permit skipping of mutant exons, leading to expression of a short but nonetheless functional dystrophin protein. In parallel, other studies suggest that small molecules may permit read-through of protein-truncating mutations in some Duchenne cases, again with clinical benefit.

LIMB-GIRDLE MUSCULAR DYSTROPHY

The LGMDs are a genetically heterogenous group of dystrophies in which males and females are affected equally, with onset ranging from late in the first decade to the fourth decade. The LGMDs typically manifest with progressive weakness of pelvic and shoulder girdle musculature and are often clinically indistinguishable from DMD and BMD. Respiratory insufficiency from weakness of the diaphragm may occur, as may cardiomyopathy. Serum CKs are elevated and the EMG is myopathic. Muscle biopsy reveal dystrophic features, but the findings are not specific to differentiate subtypes from one another unless immunohistochemistry is employed (e.g., immunostaining for various sarcoglycans, dysferlin, alpha-dystroglycan, merosin) or there are features to suggest one of the myofibrillar myopathies. Nonetheless, definitive diagnosis requires genetic testing.

A systematic classification of LGMD is based on autosomal dominant (LGMD1) and autosomal recessive (LGMD2) inheritance. Superimposed on the backbone of LGMD1 and LGMD2, the classification uses a sequential alphabetical lettering system (LGMD1A, LGMD2A, etc.) based on genotype. This results in an ever-expanding list of conditions summarized in Tables 441-3 and 441-4. The prevalence of LGMD ranges from 80 to 700 per 100,000 while estimated prevalence of individual specific subtypes LGMDs vary. The most common types of adult-onset LGMD are calpainopathy (LGMD2A), Fukutin-related protein (FKRP) deficiency (LGMD2I), and anoctaminopathy (LGMD2L). LGMD2A, the most common cause of LGMD in those with ancestry from Spain, France, Italy, and Great Britain, is associated with marked scapular winging, lack of calf muscle hypertrophy, and lack of cardiac and lung involvement. LGMD2I is more common in those from northern European ancestry, is associated with calf muscle hypertrophy, and can have cardiac and lung involvement out of proportion to extremity weakness. LGMD2L is becoming increasingly recognized as a common form of LGMD; as seen in dysferlinopathies (LGMD2B and Miyoshi myopathy type 1), anoctaminopathy has an early predilection for medial calf atrophy and weakness.

EMERY-DREIFUSS MUSCULAR DYSTROPHY (EDMD)

There are at least five genetically distinct forms of EDMD. Emerin mutations are the most common cause of X-linked EDMD, although mutations in *FHL1* may also be associated with a similar phenotype, which is X-linked as well. Mutations involving the gene for lamin A/C are the most common cause of autosomal dominant EDMD (also known as LGMD1B) and are also a common cause of hereditary cardiomyopathy. Less commonly, autosomal dominant EDMD has been reported with mutations in *nesprin-1, nesprin-2*, and *TMEM43* genes.

Clinical Features

Prominent contractures can be recognized in early childhood and teenage years, often preceding muscle weakness. The contractures persist throughout the course of the disease and are present at the elbows, ankles, and neck. Muscle weakness affects humeral and peroneal muscles at first and later spreads to a limb-girdle distribution (Table 441-1). The cardiomyopathy is potentially life threatening and may result in sudden death. A spectrum of atrial rhythm and conduction defects includes atrial fibrillation and paralysis and atrioventricular heart block. Some patients have a dilated cardiomyopathy. Female carriers of the X-linked variant may manifest with a cardiomyopathy.

Laboratory Features

Serum CK is usually slightly elevated and the EMG is myopathic. Muscle biopsy usually shows nonspecific dystrophic features, although cases associated with *FHL1* mutations have features of myofibrillar myopathy. Immunohistochemistry reveals absent emerin staining of myonuclei in X-linked EDMD due to *emerin* mutations. ECGs demonstrate atrial and atrioventricular rhythm disturbances.

X-linked EDMD usually arises from defects in the *emerin* gene encoding a nuclear envelope protein. *FHL1* mutations are also a cause of X-linked scapuloperoneal dystrophy, but can also present with an X-linked form of EDMD. The autosomal dominant disease can be caused by mutations in the *LMNA* gene encoding lamin A and C; in the synaptic nuclear envelope protein 1 (*SYNE1*) or 2 (*SYNE2*) encoding nesprin-1 and nesprin-2, respectively; and in *TMEM43* encoding LUMA. These proteins are essential components of the filamentous network underlying the inner nuclear membrane. Loss of structural integrity of the nuclear envelope from defects in emerin, lamin A/C, nesprin-1, nesprin-2, and LUMA accounts for overlapping phenotypes.

TREATMENT

TREATMENT

Emery-Dreifuss Muscular Dystrophy

Supportive care should be offered for neuromuscular disability, including ambulatory aids, if necessary. Stretching of contractures is difficult. Management of cardiomyopathy and arrhythmias (e.g., early use of a defibrillator or cardiac pacemaker) may be lifesaving.

MYOTONIC DYSTROPHY

There are two distinct forms of myotonic dystrophy (*dystrophia myotonica* or DM), namely myotonic dystrophy type 1 (DM1), and myotonic dystrophy type 2 (DM2) also called *proximal myotonic myopathy* (PROMM).

Clinical Features

The clinical expression of DM1 varies widely and involves many systems other than muscle. Affected patients may have a “hatchet-faced” appearance due to temporalis, masseter, and facial muscle atrophy and weakness. Frontal baldness is frequent. Weakness of wrist and fingers occurs early as does footdrop. Proximal muscles are less affected. Palatal, pharyngeal, and tongue involvement can lead to dysarthria and dysphagia. Some patients have diaphragm and intercostal muscle weakness, resulting in ventilatory insufficiency. Myotonia is usually apparent by the age of 5 years and is best demonstrable by percussion of the thenar eminence or asking patients to close their fingers very tightly and then relax.

ECG abnormalities include first-degree heart block and more extensive conduction system involvement. Complete heart block and sudden death can occur. Congestive heart failure occurs infrequently but may result from cor pulmonale secondary to respiratory failure. Other associated features include intellectual impairment, hypersomnia, posterior subcapsular cataracts, gonadal atrophy, insulin resistance, and decreased esophageal and colonic motility.

Congenital myotonic dystrophy is a more severe form of DM1 and occurs in ~25% of infants of affected mothers. It is characterized by severe facial and bulbar weakness, transient neonatal respiratory insufficiency, and mental retardation.

DM2 or PROMM involves mainly proximal muscles. Other features of the disease overlap with DM1, including cataracts, testicular atrophy, insulin resistance, constipation, hypersomnia, and cognitive defects. Cardiac conduction defects occur but are less common. The hatchet face and frontal baldness are also less consistent features. A very striking difference is the failure to clearly identify a congenital form of DM2.

Laboratory Features

The diagnosis of myotonic dystrophy can usually be made on the basis of clinical findings. Serum CK levels may be normal or mildly elevated. EMG evidence of myotonia is present in most cases of DM1 but is more patchy in DM2. Muscle biopsy is not typically performed for diagnosis but is sometimes done when the clinical features and electrophysiological features are not recognized. The major histopathological feature in both DM1 and DM2 are numerous internalized nuclei can be seen in individual muscle fibers combined with many atrophic fibers with pyknotic nuclear clumps.

DM1 and DM2 are autosomal dominant disorders. DM1 is transmitted by an intronic mutation consisting of an unstable expansion of a CTG trinucleotide repeat in a serine-threonine protein kinase gene (named *DMPK*). An increase in the severity of the disease phenotype in successive generations (genetic anticipation) is accompanied by an increase in the number of trinucleotide repeats. The unstable triplet repeat in myotonic dystrophy can be used for prenatal diagnosis. Congenital disease occurs almost exclusively in infants born to affected mothers.

DM2 is caused by a DNA expansion mutation consisting of a CCTG repeat in intron 1 of the *CNBP* gene encoding the CCHC-type zinc finger nucleic acid binding protein. The DNA expansions in DM1 and DM2 impair muscle function by a toxic gain of function of the mutant mRNA. In both DM1 and DM2, the mutant RNA appears to form intranuclear inclusions composed of aberrant RNA. These RNA inclusions sequester RNA-binding proteins essential for proper splicing of a variety of other mRNAs. This leads to abnormal transcription of multiple proteins in a variety of tissues/organ systems, in turn causing the systemic manifestations of DM1 and DM2.

TREATMENT

TREATMENT

Myotonic Dystrophy

The myotonia in DM1 and DM2 is usually not so bothersome to warrant treatment, but when it is mexiletine may be helpful. A cardiac pacemaker or implantable cardioverter defibrillator should be considered for patients with significant arrhythmia. Molded ankle-foot orthoses help stabilize gait in patients with foot drop. Excessive daytime somnolence with or without sleep apnea is not uncommon. Sleep studies, noninvasive respiratory support (biphasic positive airway pressure [BiPAP]), and treatment with modafinil may be beneficial.

FACIOSCAPULOHUMERAL (FSH) MUSCULAR DYSTROPHY

There are two forms of FSHD that have similar pathogenesis. Most patients have FSHD type 1 (95%), whereas ~5% have FSHD2. Both forms are clinically and histopathologically identical. The prevalence FSHD is ~5 per 100,000 individuals.

Clinical Features

FSHD typically presents in childhood or young adulthood. In most cases, facial weakness is the initial manifestation, appearing as an inability to smile, whistle, or fully close the eyes. Loss of scapular stabilizer muscles makes arm elevation difficult. Scapular winging (Fig. 441-3) becomes apparent with attempts at abduction and forward movement of the arms. Biceps and triceps muscles may be severely affected, with relative sparing of the deltoid muscles. Weakness is invariably worse for wrist extension than for wrist flexion, and weakness of the anterior compartment muscles of the legs may lead to footdrop. In 20% of patients, weakness progresses to involve the pelvic muscles, and severe functional impairment and possible wheelchair dependency result. The heart is not involved but there can be ventilatory muscle weakness in 5% of affected individuals. There is an increased incidence of nerve deafness. *Coats' disease*, a disorder consisting of telangiectasia, exudation, and retinal detachment, also occurs.

Laboratory Features

The serum CK level may be normal or mildly elevated. EMG and muscle biopsy show nonspecific abnormalities but on occasion can reveal a prominent inflammatory infiltrate leading to an incorrect diagnosis of myositis **(Chap. 358)**.

FSHD1 is associated with deletions of tandem 3.3-kb repeats at 4q35. The deletion reduces the number of repeats to a fragment of <35 kb in most patients. Within these repeats lies the *DUX4* gene, which usually is not expressed after early muscle development. In patients with FSHD1 these deletions in the setting of a specific polymorphism leads to hypomethylation of the region and toxic expression of the *DUX4* gene. In patients with FSHD2, there is no deletion, but a mutation in *SMCHD1*, leading to hypomethylation of the DUX4 region and the permissive expression of the *DUX4* gene. In both FSHD1 and FSHD2, there is overexpression of the *DUX4* transcript.

TREATMENT

TREATMENT

Facioscapulohumeral Muscular Dystrophy

No specific treatment is available; ankle-foot orthoses are helpful for footdrop. Scapular stabilization procedures improve scapular winging but may not improve function.

OCULOPHARYNGEAL DYSTROPHY (OPMD)

OPMD represents one of several disorders characterized by progressive external ophthalmoplegia, which consists of slowly progressive ptosis and limitation of eye movements with sparing of pupillary reactions for light and accommodation. Patients usually do not complain of diplopia, in contrast to patients having conditions with a more acute onset of ocular muscle weakness (e.g., myasthenia gravis).

Clinical Features

OPMD has a late onset; it usually presents in the fourth to sixth decade with ptosis or dysphagia. The extraocular muscle impairment is less prominent in the early phase but may become severe over time. The swallowing problem may lead to aspiration. Weakness of the neck and proximal extremities can develop but is usually mild in degree.

Laboratory Features

The serum CK level may be two to three times normal. EMG can identify myopathic changes in weak muscles. Muscle biopsies are no longer necessary for diagnosis in most cases, but when performed demonstrate muscle fibers with rimmed vacuoles. On electron microscopy, a distinctive feature of OPMD is the presence of 8.5 nm tubular filaments in some muscle cell nuclei.

OPMD is an autosomal dominant disorder that has a high incidence in certain populations (e.g., French-Canadians, individuals of Spanish ancestry, and Ashkenazi Jews). The molecular defect in OPMD is an expansion of a polyalanine repeat tract in a poly-RNA-binding protein (*PABP2*) gene.

TREATMENT

TREATMENT

Oculopharyngeal Dystrophy

Dysphagia can lead to significant undernourishment and aspiration. Cricopharyngeal myotomy may improve swallowing. Eyelid crutches can improve vision when ptosis obstructs vision; candidates for ptosis surgery must be carefully selected—those with severe facial weakness are not suitable.

DISTAL MYOPATHIES / DYSTROPHIES

The distal myopathies are notable for their preferential distal distribution of muscle weakness in contrast to most muscle conditions associated with proximal weakness. The major distal myopathies are summarized in Tables 441-1 and 441-5.

Clinical Features

Welander, Udd, and *Markesbery-Griggs type distal myopathies* are all late-onset, dominantly inherited disorders of distal limb muscles, usually beginning after age 40 years. Welander distal myopathy preferentially involves the wrist and finger extensors, whereas the others are associated with anterior tibial weakness leading to progressive footdrop. *Laing distal myopathy* is also a dominantly inherited disorder heralded by tibial weakness; however, it is distinguished by onset in childhood or early adult life. GNE myopathy (also known as *Nonaka distal myopathy* and autosomal recessive hereditary inclusion body myopathy) and *Miyoshi myopathy* are distinguished by autosomal recessive inheritance and onset in the late teens or twenties. GNE and Williams myopathy produce prominent anterior tibial weakness, whereas Miyoshi myopathy is unique in that gastrocnemius muscles are preferentially affected at onset. Finally, the *myofibrillar myopathies* (MFMs) are a clinically and genetically heterogeneous group of muscular dystrophies that can be associated with prominent distal weakness; they can be inherited in an autosomal dominant or recessive pattern. Of note, Markesbery-Griggs myopathy (caused by mutations in ZASP) and LGMD1B (caused by mutations in myotilin) are subtypes of myofibrillar myopathy (MFM).

Laboratory Features

Serum CK levels are markedly elevated in Miyoshi myopathy, but in the other conditions, serum CK is only slightly increased. EMGs are myopathic and can be irritable with myotonic discharges in MFM. Muscle biopsy shows nonspecific dystrophic features and, with the exception of Laing and Miyoshi myopathies, often shows rimmed vacuoles. MFM is associated with the accumulation of dense inclusions and amorphous material best seen on Gomori trichrome staining along with myofibrillar disruption on electron microscopy. Immune staining sometimes demonstrates accumulation of desmin and other proteins in MFM, large deposits of myosin heavy chain in the subsarcolemmal region of type 1 muscle fibers in Laing myopathy, and reduced or absent dysferlin in Miyoshi myopathy type 1.

TREATMENT

TREATMENT

Distal Myopathies

Occupational therapy is offered for loss of hand function; ankle-foot orthoses can support distal lower limb muscles. The MFMs can be associated with cardiomyopathy (congestive heart failure or arrhythmias) and respiratory failure that may require medical management. Laing-type distal myopathy can also be associated with a cardiomyopathy.

DISORDERS OF MUSCLE ENERGY METABOLISM

There are two principal sources of energy for skeletal muscle—fatty acids and glucose. Abnormalities in either glucose or lipid utilization can be associated with distinct clinical presentations that can range from an acute, painful syndrome with rhabdomyolysis and myoglobinuria to a chronic, progressive muscle weakness simulating muscular dystrophy (Table 441-1). As with the muscular dystrophies there are no specific medical treatments available.

GLYCOGEN STORAGE AND GLYCOLYTIC DEFECTS

Disorders of Glycogen Storage Causing Progressive Weakness

ALPHA-GLUCOSIDASE, OR ACID MALTASE, DEFICIENCY (POMPE DISEASE)

Three clinical forms of α-glucosidase, or acid maltase, deficiency (*type II glycogenosis*) can be distinguished. The infantile form is the most common, with onset of symptoms in the first 3 months of life. Infants develop severe muscle weakness, cardiomegaly, hepatomegaly, and respiratory insufficiency. Glycogen accumulation in motor neurons of the spinal cord and brainstem contributes to muscle weakness. Death usually occurs by 1.5 years of age. In the childhood form, the picture resembles Duchenne muscular dystrophy with delayed motor milestones resulting from proximal limb muscle weakness and involvement of respiratory muscles. The heart may be involved, but the liver and brain are unaffected. The adult form usually begins in the third or fourth decade but can present as late as the seventh decade. Ventilatory weakness can be the initial and only manifestation in 20–30% of late-onset cases.

The serum CK level is 2–10 times normal in infantile or childhood-onset Pompe's disease but can be normal in adult-onset cases. EMG can demonstrate muscle membrane irritability, particularly in the paraspinal muscles. The muscle biopsy in infants typically reveals vacuoles containing glycogen and the lysosomal enzyme acid phosphatase. Electron microscopy reveals membrane-bound and free tissue glycogen. However, muscle biopsies in late-onset Pompe disease may demonstrate only nonspecific abnormalities. Enzyme analysis of dried blood spots is a sensitive technique to screen for Pompe's disease. A definitive diagnosis is established by genetic testing.

Pompe disease is inherited as an autosomal recessive disorder caused by mutations of the *α-glucosidase* gene. Enzyme replacement therapy (ERT) with IV recombinant human α-glucosidase is beneficial in infantile-onset Pompe disease. In late-onset cases, ERT has a more modest benefit.

OTHER GLYCOGEN STORAGE DISEASES WITH PROGRESSIVE WEAKNESS

In *de-branching enzyme deficiency* (*type III glycogenosis*), a slowly progressive form of muscle weakness can develop after puberty. Rarely, myoglobinuria may be seen. Patients are usually diagnosed in infancy, however, because of hypotonia and delayed motor milestones; hepatomegaly, growth retardation, and hypoglycemia are other manifestations. *Branching enzyme deficiency* (*type IV glycogenosis*) is a rare and fatal glycogen storage disease characterized by failure to thrive and hepatomegaly. Hypotonia and muscle wasting may be present, but the skeletal muscle manifestations are minor compared to liver failure.

Disorders of Glycolysis Causing Exercise Intolerance

Several glycolytic defects are associated with recurrent myoglobinuria. The most common is *McArdle disease* caused by mutations in the *PYGM* gene leading to *myophosphorylase deficiency*. Symptoms of muscle pain and stiffness usually begin in adolescence. With severe episodes myoglobinuria can occur.

Certain features help distinguish some enzyme defects. In McArdle disease, exercise tolerance can be enhanced by a slow induction phase (warm-up) or brief periods of rest, allowing for the start of the “second-wind” phenomenon (switching to utilization of fatty acids). Varying degrees of hemolytic anemia accompany deficiencies of both phosphofructokinase (mild) and phosphoglycerate kinase (severe). In phosphoglycerate kinase deficiency, the usual clinical presentation is a seizure disorder associated with mental retardation; exercise intolerance is an infrequent manifestation.

In all of these conditions, the serum CK levels fluctuate widely and may be elevated even during symptom-free periods. CK levels >100 times normal are expected accompanying myoglobinuria. A forearm exercise test reveals a blunted rise in venous lactate with a normal rise in ammonia. A definitive diagnosis of glycolytic disease can be made by muscle biopsy with appropriate staining and enzyme assays, but genetic testing is now done in lieu of biopsy in most cases.

Training may enhance exercise tolerance, perhaps by increasing perfusion to muscle. Dietary intake of free glucose or fructose prior to activity may improve function but care must be taken to avoid obesity from ingesting too many calories.

LIPID AS AN ENERGY SOURCE AND ASSOCIATED DEFECTS

Lipid is an important muscle energy source during rest and during prolonged, submaximal exercise. Oxidation of fatty acids occurs in the mitochondria. To enter the mitochondria, a fatty acid must first be converted to an “activated fatty acid,” acyl-CoA. The acyl-CoA must be linked with carnitine by the enzyme CPT for transport into the mitochondria.

Carnitine Palmitoyltransferase 2 Deficiency

CPT2 deficiency is the most common recognizable cause of recurrent myoglobinuria. Onset is usually in the teenage years or early twenties. Muscle pain and myoglobinuria typically occur after prolonged exercise but can also be precipitated by fasting or infections; up to 20% of patients do not exhibit myoglobinuria, however. Strength is normal between attacks. In contrast to disorders caused by defects in glycolysis, in which muscle cramps follow short, intense bursts of exercise, the muscle pain in CPT2 deficiency does not occur until the limits of utilization have been exceeded and muscle breakdown has already begun.

Serum CK levels and EMG findings are both usually normal between episodes. A normal rise of venous lactate during forearm exercise distinguishes this condition from glycolytic defects. Muscle biopsy does not show lipid accumulation and is usually normal between attacks. The diagnosis requires direct measurement of muscle CPT or genetic testing. Attempts to improve exercise tolerance with frequent meals and a low-fat, high-carbohydrate diet, or by substituting medium-chain triglycerides in the diet, have not proven to be beneficial.

MITOCHONDRIAL MYOPATHIES

Mitochondria play a key role in energy production. Oxidation of the major nutrients derived from carbohydrate, fat, and protein leads to the generation of reducing equivalents. The latter are transported through the respiratory chain in the process known as *oxidative phosphorylation*. The energy generated by the oxidation-reduction reactions of the respiratory chain is stored in an electrochemical gradient coupled to ATP synthesis.

A novel feature of mitochondria is their genetic composition. Each mitochondrion possesses a DNA genome that is distinct from that of the nuclear DNA. Human mitochondrial DNA (mtDNA) consists of a double-strand, circular molecule comprising 16,569 base pairs. It codes for 22 transfer RNAs, 2 ribosomal RNAs, and 13 polypeptides of the respiratory chain enzymes. The genetics of mitochondrial diseases differ from the genetics of chromosomal disorders. The DNA of mitochondria is directly inherited from the cytoplasm of the gametes, mainly from the oocyte. The sperm contributes very little of its mitochondria to the offspring at the time of fertilization. Thus, mitochondrial genes are derived almost exclusively from the mother, accounting for maternal inheritance of some mitochondrial disorders.

Patients with mitochondrial myopathies have clinical manifestations that usually fall into three groups: chronic progressive external ophthalmoplegia (CPEO), skeletal muscle–CNS syndromes, and pure myopathy simulating muscular dystrophy or metabolic myopathy. Unfortunately, no specific medical therapies are clearly beneficial, although Co-enzyme Q10 supplements are often prescribed.

Kearn-Sayre Syndrome (KSS)

This is a widespread multi-organ system disorder with a defined triad of clinical findings: onset before age 20, CPEO, and pigmentary retinopathy, plus one or more of the following features: complete heart block, cerebrospinal fluid (CSF) protein >1 g/L (100 mg/dL), or cerebellar ataxia. The cardiac disease includes syncopal attacks and cardiac arrest related to the abnormalities in the cardiac conduction system: prolonged intraventricular conduction time, bundle branch block, and complete atrioventricular block. Death attributed to heart block occurs in ~20% of the patients. Varying degrees of progressive limb muscle weakness and easy fatigability affect activities of daily living. Many affected individuals have intellectual disabilities. Endocrine abnormalities are also common, including gonadal dysfunction in both sexes with delayed puberty, short stature, and infertility. Diabetes mellitus occurs in ~13% of KSS patients. Other less common endocrine disorders include thyroid disease, hyperaldosteronism, Addison disease, and hypoparathyroidism.

Serum CK and lactate levels are normal or slightly elevated. EMG is myopathic. NCS may be abnormal related to an associated neuropathy. Muscle biopsies reveal ragged red fibers and cytochrome oxidase (COX) negative fibers. By electron microscopy, there are increased numbers of mitochondria that often appear enlarged and contain paracrystalline inclusions.

KSS is a sporadic disorder caused by single mtDNA deletions that are presumed to arise spontaneously in the ovum or zygote. The most common deletion, occurring in about one-third of patients, removes 4977 bp of contiguous mtDNA. Monitoring for cardiac conduction defects is critical. Prophylactic pacemaker implantation is indicated when ECGs demonstrate a bifascicular block.

Progressive External Ophthalmoplegia (PEO)

PEO can be caused by nuclear DNA mutations affecting mtDNA and thus inherited in a Mendelian fashion or by mutations in mtDNA. Onset is usually after puberty. Fatigue, exercise intolerance, dysphagia, and complaints of muscle weakness are typical. The neurologic examination confirms the ptosis and ophthalmoplegia, usually asymmetric in distribution. Patients do not complain of diplopia. Mild facial, neck flexor, and proximal weakness are typical. Rarely, respiratory muscles may be progressively affected and may be the direct cause of death.

Serum CK and lactate can be normal or mildly elevated. The EMG can myopathic. Ragged red and COX-negative fibers are prominently displayed in the muscle biopsy.

This autosomal dominant form of CPEO is most commonly caused by mutations in the genes encoding adenine nucleotide translocator 1 (*ANT1*), twinkle gene (*C10orf2*), and mtDNA polymerase 1(*POLG1*). Autosomal recessive PEO can also be caused by mutations in *POLG1*. Point mutations have been identified within various mitochondrial tRNA (Leu, Ile, Asn, Trp) genes in families with maternal inheritance of PEO.

There is no specific medical treatment available; exercise may improve function but will depend on the patient's ability to participate.

Myoclonic Epilepsy with Ragged Red Fibers (MERRF)

The onset of MERRF is variable, ranging from late childhood to middle adult life. Characteristic features include myoclonic epilepsy, cerebellar ataxia, and progressive proximal muscle weakness. The seizure disorder is an integral part of the disease and may be the initial symptom. Cerebellar ataxia precedes or accompanies epilepsy. Other more variable features include dementia, peripheral neuropathy, optic atrophy, hearing loss, and diabetes mellitus.

Serum CK levels and lactate may be normal or elevated. EMG is myopathic, and in some patients NCS show a neuropathy. The electroencephalogram is abnormal, corroborating clinical findings of epilepsy. Typical ragged red fibers are seen on muscle biopsy. MERRF is caused by maternally inherited point mutations of mitochondrial tRNA genes. The most common mutation found in 80% of MERRF patients is an A to G substitution at nucleotide 8344 of tRNA lysine (A8344G $tRNA^{lys}$). Only supportive treatment is possible, with special attention to epilepsy.

Mitochondrial Myopathy, Encephalopathy, Lactic Acidosis, and Stroke-like Episodes (MELAS)

MELAS is the most common mitochondrial encephalomyopathy. The term *stroke-like* is appropriate because the cerebral lesions do not conform to a strictly vascular distribution. The onset in the majority of patients is before age 20. Seizures, usually partial motor or generalized, are common and may represent the first clearly recognizable sign of disease. The cerebral insults that resemble strokes cause hemiparesis, hemianopia, and cortical blindness. A presumptive stroke occurring before age 40 should place this mitochondrial encephalomyopathy high in the differential diagnosis. Associated conditions include hearing loss, diabetes mellitus, hypothalamic pituitary dysfunction causing growth hormone deficiency, hypothyroidism, and absence of secondary sexual characteristics. In its full expression, MELAS leads to dementia, a bedridden state, and a fatal outcome. Serum lactic acid is typically elevated.

The CSF protein is also increased but is usually ≤1 g/L (100 mg/dL). Muscle biopsies show ragged red fibers. Neuroimaging demonstrates basal ganglia calcification in a high percentage of cases. Focal lesions that mimic infarction are present predominantly in the occipital and parietal lobes. Strict vascular territories are not respected, and cerebral angiography fails to demonstrate lesions of the major cerebral blood vessels.

MELAS is usually caused by maternally inherited point mutations of mitochondrial tRNA genes. The A3243G point mutation in $tRNA^{Leu(UUR)}$ is the most common, occurring in ~80% of MELAS cases. No specific treatment is available. Supportive treatment is essential for the strokelike episodes, seizures, and

endocrinopathies.

Mitochondrial DNA Depletion Syndromes

Mitochondrial DNA depletion syndrome (MDS) is a heterogeneous group of disorders that are inherited in an autosomal recessive fashion and can present in infancy or adults. MDS can be caused by mutations in several genes (*TK2*, *DGUOK*, *RRM2B*, *TYMP*, *SUCLA1*, and *SUCLA2*) that lead to depletion of mitochondrial deoxyribonucleotides (dNTP) necessary for mtDNA replication. The other major cause of MDS is a set of mutations in genes essential for mtDNA replication (e.g., *POLG1* and *C10orf2*). The clinical phenotypes associated with MDS vary. Patients may develop a severe encephalopathy (e.g., Leigh's syndrome), PEO, an isolated myopathy, myo-neuro-gastrointestinal-encephalopathy (MNGIE), and a sensory neuropathy with ataxia.

DISORDERS OF MUSCLE MEMBRANE EXCITABILITY

Muscle membrane excitability is affected in a group of disorders referred to as *channelopathies*. These disorders usually present with episodic muscle weakness (periodic paralysis) and sometimes myotonia or paramyotonia (Table 441-1).

CALCIUM CHANNEL DISORDERS OF MUSCLE

Hypokalemic Periodic Paralysis (HypoKPP)

This is an autosomal dominant disorder with onset in adolescence. Males are more often affected because of decreased penetrance in females. Episodic weakness with onset after age 25 is almost never due to periodic paralyses, with the exception of thyrotoxic periodic paralysis. Attacks are often provoked by meals high in carbohydrates or sodium and may accompany rest following prolonged exercise. Weakness usually affects proximal limb muscles more than distal. Ocular and bulbar muscles are less likely to be affected. Respiratory muscles are usually spared, but when they are involved, the condition may prove fatal. Weakness may take as long as 24 h to resolve. Life-threatening cardiac arrhythmias related to hypokalemia may occur during attacks. As a late complication, patients commonly develop severe, disabling proximal lower extremity weakness.

Attacks of thyrotoxic periodic paralysis resemble those of primary HypoKPP. Despite a higher incidence of thyrotoxicosis in women, men, particularly those of Asian descent, are more likely to manifest this complication. Attacks abate with treatment of the underlying thyroid condition.

A low serum potassium level during an attack, excluding secondary causes, establishes the diagnosis. In the midst of an attack of weakness, motor conduction studies may demonstrate reduced amplitudes, whereas EMG may show electrical silence in severely weak muscles. In between attacks, the EMG and routine NCS are normal. However, a long exercise NCS test may demonstrate decrementing amplitudes.

HypoKPP type 1 is the most common form, and is caused by mutations in the voltage-sensitive, skeletal muscle calcium channel gene, *CALCL1A3*. Approximately 10% of cases are HypoKPP type 2, arising from mutations in the voltage-sensitive sodium channel gene (*SCN4A*). In both forms the mutations lead to an abnormal gating pore current that predisposes the muscle cell to depolarize when potassium levels are low.

TREATMENT

TREATMENT

Hypokalemic Periodic Paralysis

Mild attacks usually do not require medical treatment. However, severe attacks of weakness can be improved by the administration of potassium. Oral KCl (0.2–0.4 mmol/kg) can be given every 30 min. Only rarely is IV therapy necessary (e.g., when swallowing problems or vomiting is present). The long-term goal of therapy is to avoid attacks. Patients should be made aware of the importance of a low-carbohydrate, low-sodium diet, and consequences of intense exercise. Prophylactic administration of acetazolamide or dichlorphenamide can reduce attacks of periodic weakness. However, in patients with HypoKPP type 2, attacks of weakness can be exacerbated with these medications.

SODIUM CHANNEL DISORDERS OF MUSCLE

Hyperkalemic Periodic Paralysis (HyperKPP)

The term *hyperkalemic* is misleading because patients are often normokalemic during attacks. That attacks are precipitated by potassium administration best defines the disease. The onset is usually in the first decade; males and females are affected equally. Attacks are brief and mild, usually lasting 30 min to several hours. Weakness affects proximal muscles, sparing bulbar muscles. Attacks are precipitated by rest following exercise and fasting.

Potassium may be slightly elevated or normal during an attack. As in HypoKPP, NCS in HyperKPP muscle may demonstrate reduced motor amplitudes and the EMG may be silent in very weak muscles. A long exercise NCS test can reveal diminished amplitudes as well. The EMG may demonstrate myotonic discharges. HyperKPP is caused by mutations of the voltage-gated sodium channel *SCN4A* gene. Acetazolamide or dichlorphenamide can reduce the frequency and severity of attacks. Mexiletine to be helpful in patients with significant clinical myotonia.

Paramyotonia Congenita

In PC, the attacks of weakness are cold-induced or occur spontaneously and are mild. Myotonia is a prominent feature but worsens with muscle activity (paradoxical myotonia). This is in contrast to classic myotonia in which exercise alleviates the condition. Attacks of weakness are seldom severe enough to require emergency room treatment. Over time patients develop interattack weakness as they do in other forms of periodic paralysis.

Serum CK is usually mildly elevated. Routine NCS are normal. Short exercise NCS test may be abnormal however, and cooling of the muscle often dramatically reduces the amplitude of the compound muscle action potentials. EMG reveals diffuse myotonic potentials in PC. Upon local cooling of the muscle, the myotonic discharges disappear as the patient becomes unable to activate MUAPs.

PC is inherited as an autosomal dominant condition; voltage-gated sodium channel mutations are responsible, and thus this disorder is allelic with HyperKPP. Mexiletine is reported to be helpful in reducing the myotonia.

POTASSIUM CHANNEL DISORDERS

Andersen-Tawil Syndrome

This rare disease is characterized by episodic weakness, cardiac arrhythmias, and dysmorphic features (short stature, scoliosis, clinodactyly, hypertelorism, small or prominent low-set ears, micrognathia, and broad forehead). The cardiac arrhythmias are potentially serious and life threatening. They include long QT, ventricular ectopy, bidirectional ventricular arrhythmias, and tachycardia. The disease is most commonly caused by mutations of the inwardly rectifying potassium channel (*Kir 2.1*) gene that heighten muscle cell excitability. The episodes of weakness may differ between patients because of potassium variability. Acetazolamide may decrease the attack frequency and severity.

CHLORIDE CHANNEL DISORDERS

Two forms of this disorder, autosomal dominant (*Thomsen disease*) and autosomal recessive (*Becker disease*) are both caused by mutations in the chloride channel 1 gene (*CLCN1*). Symptoms are noted in infancy and early childhood. The severity lessens in the third to fourth decade. Myotonia is worsened by cold and improved by activity. The gait may appear slow and labored at first but improves with walking. In Thomsen disease, muscle strength is normal, but in Becker disease, which is usually more severe, there may be muscle weakness. Muscle hypertrophy is usually present. Myotonic discharges are prominently displayed by EMG recordings. Serum CK is normal or mildly elevated. Mexiletine is helpful in relieving the myotonia.

ENDOCRINE AND METABOLIC MYOPATHIES

Endocrinopathies can cause weakness, but fatigue is more common than true weakness. The serum CK level is often normal (except in hypothyroidism) and the muscle histology is characterized by atrophy rather than destruction of muscle fibers. Nearly all endocrine myopathies respond to treatment.

THYROID DISORDERS

Hypothyroidism

Patients with hypothyroidism have frequent muscle complaints, and about one-third have proximal muscle weakness **(Chap. 376)**. Muscle cramps, pain, and stiffness are common. Some patients have enlarged muscles. Features of slow muscle contraction and relaxation occur in 25% of patients; the relaxation phase of muscle stretch reflexes is characteristically prolonged and best observed at the ankle or biceps brachii reflexes. The serum CK level is often elevated (up to 10 times normal). EMG is typically normal. Muscle biopsy shows no distinctive morphologic abnormalities.

Hyperthyroidism

Patients who are thyrotoxic commonly have proximal muscle weakness, but they rarely complain of myopathic symptoms **(Chap. 377)**. Activity of deep tendon reflexes may be enhanced. Fasciculations may be apparent and, when coupled with increased muscle stretch reflexes, may lead to an erroneous diagnosis of amyotrophic lateral sclerosis. A form of hypokalemic periodic paralysis can occur in patients who are thyrotoxic. Mutations in the *KCNJ18* gene that encodes for the inwardly rectifying potassium channel, Kir 2.6, have been discovered in up to a third of cases.

PARATHYROID DISORDERS

(See also Chap. 403)

Hyperparathyroidism

Proximal muscle weakness, muscle wasting, and brisk muscle stretch reflexes are the main features of this endocrinopathy. Some patients develop neck extensor weakness (part of the dropped head syndrome). Serum CK levels are usually normal or slightly elevated. Serum parathyroid hormone levels are elevated, while vitamin D and calcium levels are usually reduced. Muscle biopsies show only mild type 2 fiber atrophy.

Hypoparathyroidism

An overt myopathy due to hypocalcemia rarely occurs. Neuromuscular symptoms are usually related to localized or generalized tetany. Serum CK levels may be increased secondary to muscle damage from sustained tetany. Hyporeflexia or areflexia is usually present and contrasts with the hyperreflexia in hyperparathyroidism.

ADRENAL DISORDERS

Conditions associated with glucocorticoid excess cause a myopathy; steroid myopathy is the most commonly diagnosed endocrine muscle disease **(See also Chap. 379)**. Proximal muscle weakness combined with a cushingoid appearance are the key clinical features. Serum CK and EMG are normal. Muscle biopsy, not typically done for diagnostic purposes, reveals type 2b muscle fiber atrophy. In primary hyperaldosteronism (*Conn's syndrome*), neuromuscular complications are due to potassium depletion. The clinical picture is one of persistent muscle weakness. Long-standing hyperaldosteronism may lead to proximal limb weakness and wasting. Serum CK levels may be elevated, and a muscle biopsy may demonstrate necrotic fibers. These changes relate to hypokalemia and are not a direct effect of aldosterone on skeletal muscle.

PITUITARY DISORDERS

Patients with acromegaly usually have mild proximal weakness **(See also Chap. 373)**. Muscles often appear enlarged but exhibit decreased force generation. The duration of acromegaly, rather than the serum growth hormone levels, correlates with the degree of myopathy.

DIABETES MELLITUS

Neuromuscular complications of diabetes mellitus are most often related to neuropathy **(See also Chap. 398)**. The only notable myopathy is ischemic infarction of leg muscles, usually involving one of the thigh muscles but on occasion affecting the distal leg. This condition occurs in patients with poorly controlled diabetes and presents with the abrupt onset of pain, tenderness, and edema of a thigh or calf. The area of muscle infarction is hard and indurated. The muscles most often affected include the vastus lateralis, thigh adductors, and biceps femoris. Computed tomography (CT) or MRI can demonstrate focal abnormalities in the affected muscle. Diagnosis by imaging is preferable to muscle biopsy, if possible, as hemorrhage into the biopsy site can occur.

MYOPATHIES OF SYSTEMIC ILLNESS

Systemic illnesses such as chronic respiratory, cardiac, or hepatic failure are frequently associated with severe muscle wasting and complaints of weakness. Fatigue is usually a more significant problem than weakness, which is typically mild.

DRUG-INDUCED OR TOXIC MYOPATHIES

The most common toxic myopathies are caused by the cholesterol-lowering agents and glucocorticoids. Others impact practice to a lesser degree but are important to consider in specific situations. **Table 441-6** provides a comprehensive list of drug-induced myopathies with their distinguishing features.

TABLE 441-6

Drug-Induced Myopathies

Drugs	Major Toxic Reaction
Lipid-lowering agents HMG-CoA reductase inhibitors Fibric acid derivatives Niacin (nicotinic acid)	Drugs belonging to all three of the major classes of lipid-lowering agents can produce a spectrum of toxicity: asymptomatic serum creatine kinase elevation, myalgias, exercise-induced pain, rhabdomyolysis, and myoglobinuria.
Glucocorticoids	Acute, high-dose glucocorticoid treatment can cause acute quadriplegic myopathy. These high doses of steroids are often combined with nondepolarizing neuromuscular blocking agents but the weakness can occur without their use. Chronic steroid administration produces predominantly proximal weakness.
Nondepolarizing neuromuscular blocking agents	Acute quadriplegic myopathy can occur with or without concomitant glucocorticoids.
Zidovudine	Mitochondrial myopathy with ragged red fibers
Drugs of abuse Alcohol Amphetamines Cocaine Heroin Phencyclidine Meperidine	All drugs in this group can lead to widespread muscle breakdown, rhabdomyolysis, and myoglobinuria. Local injections cause muscle necrosis, skin induration, and limb contractures.
Autoimmune myopathy Statins Check point inhibitors D-Penicillamine	Use of statins may cause an immune mediated necrotizing myopathy associated with HMG-CoA reductase antibodies. Check point inhibitors can be complicated by myositis, myasthenia gravis, and immune-mediated neuropathies. Myasthenia gravis has also been reported with penicillamine.
Amphophilic cationic drugs Amiodarone Chloroquine Hydroxychloroquine	All amphophilic drugs have the potential to produce painless, proximal weakness associated with necrosis and autophagic vacuoles in the muscle biopsy.
Antimicrotubular drugs Colchicine	This drug produces painless, proximal weakness especially in the setting of renal failure. Muscle biopsy shows necrosis and fibers with autophagic vacuoles.

MYOPATHY FROM LIPID-LOWERING AGENTS

All classes of lipid-lowering agents have been implicated in muscle toxicity, including HMG-CoA reductase inhibitors (statins), and to a much lesser extent, fibrates, niacin, and ezetimibe. Myalgia and elevated CKs are the most common manifestations. Rarely patients exhibit proximal weakness or myoglobinuria. Concomitant use of statins with fibrates and cyclosporine increase the risk of severe myotoxicity. EMG demonstrates irritability and myopathic units and muscle biopsies reveal necrotic muscle fibers in weak muscles. Severe myalgia, weakness, marked elevations in serum CK (>3–5 times

baseline), and myoglobinuria are indications for stopping the drug. Patients usually improve with drug cessation, although this may take several weeks. Rare cases continue to progress after the offending agent is discontinued. It is possible that in such cases the statin may have triggered an immune-mediated necrotizing myopathy, as these individuals require aggressive immunotherapy (e.g., prednisone and sometimes other agents) to improve and often relapse when these therapies are discontinued **(Chap. 358)**. Interestingly, antibodies directed against HMG-CoA reductase have been identified in many of these cases.

GLUCOCORTICOID-RELATED MYOPATHIES

Glucocorticoid myopathy occurs with chronic treatment or as "acute quadriplegic" myopathy secondary to high-dose IV glucocorticoid use. Chronic administration produces proximal weakness accompanied by cushingoid manifestations, which can be quite debilitating; the chronic use of prednisone at a daily dose of ≥30 mg/d is most often associated with toxicity. Patients taking fluorinated glucocorticoids (triamcinolone, betamethasone, dexamethasone) appear to be at especially high risk for myopathy. In chronic steroid myopathy, the serum CK is usually normal. Serum potassium may be low. The muscle biopsy in chronic cases shows preferential type 2 muscle fiber atrophy; this is not reflected in the EMG, which is usually normal.

Patients receiving high-dose IV glucocorticoids for status asthmaticus, chronic obstructive pulmonary disease, organ transplantation, or other indications may develop severe generalized weakness (critical illness myopathy). This myopathy, also known as acute quadriplegic myopathy, can also occur in the setting of sepsis. Involvement of the diaphragm and intercostal muscles causes ventilatory muscle weakness and is usually appreciated when patients are unable to be weaned off a ventilatory in the ICU. NCS demonstrate reduced compound muscle action potentials in the setting of relatively preserved sensory potentials. EMG can demonstrate abnormal insertional and spontaneous activity and early recruitment of myopathic appearing units in those muscles that can be activated. Muscle biopsy can show a distinctive loss of thick filaments (myosin) by electron microscopy. Treatment is withdrawal of glucocorticoids and physical therapy but the recovery is slow. Patients require supportive care and rehabilitation.

OTHER DRUG-INDUCED MYOPATHIES

Certain drugs produce painless, largely proximal, muscle weakness. These drugs include the amphophilic cationic drugs (amiodarone, chloroquine, hydroxychloroquine) and antimicrotubular drugs (colchicine) (Table 441-6). Muscle biopsy can be useful in the identification of toxicity because autophagic vacuoles are prominent pathologic features of these toxins.

GLOBAL ISSUES

As previously discussed, certain dystrophies have an increased prevalence in different parts of the world. LGMD2A is the most common LGMD in individuals from Spain, France, Italy, and Great Britain; LGMD2I is more common in those with northern European ancestry. GNE myopathy is the most common form of distal myopathy in Japan but is also prevalent in the Ashkenazi population. OPMD is most comment in those with ancestry from Spain and French-Canada as well as among Ashkenazi. Epidemiological studies are lacking regarding other forms of myopathy and their prevalence in different areas of the world.

FURTHER READING

Amato AA, Russell JA (eds): *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill Education, 2016.

Finsterer J et al: Secondary myopathy due to systemic diseases. Acta Neurol Scand 134:388, 2016. [PubMed: 26915593]

Iyadurai SJ, Kissel JT: The limb-girdle muscular dystrophies and the dystrophinopathies. Continuum (Minneap Minn) 22:1954, 2016. [PubMed: 27922502]

Katzberg HD, Kassardjian CD: Toxic and endocrine myopathies. Continuum (Minneap Minn) 22:1815, 2016. [PubMed: 27922495]

Mah JK et al: A systematic review and meta-analysis on the epidemiology of the muscular dystrophies. Can J Neurol Sci 43:163, 2016. [PubMed: 26786644]

Narayanaswami P et al: Summary of evidence-based guideline: Diagnosis and treatment limb-girdle and distal muscular dystrophies. Neurology 83:1453, 2014. [PubMed: 25313375]

Rosow LK, Amato AA: The role of electrodiagnostic testing, imaging, and muscle biopsy in the investigation of muscle disease. Continuum (Minneap Minn) 22:1787, 2016. [PubMed: 27922493]

Sansone VA: The dystrophic and nondystrophic myotonias. Continuum (Minneap Minn) 22:1889, 2016. [PubMed: 27922499]

Sansone VA et al: Randomized, placebo-controlled trials of dichlorphenamide in periodic paralysis. Neurology 86:1408, 2016. [PubMed: 26865514]

Statland JM, Tawil R: Facioscapulohumeral muscular dystrophy. Continuum (Minneap Minn) 22:1916, 2016. [PubMed: 27922500]

Tawil R et al: Evidence-based guideline summary: Evaluation, diagnosis, and management of facioscapulohumeral muscular dystrophy: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology and the Practice Issues Review Panel of the American Association of Neuromuscular & Electrodiagnostic Medicine. Neurology 85:357, 2015. [PubMed: 26215877]

Chapter 442: Chronic Fatigue Syndrome

Gijs Bleijenberg; Jos W. M. van der Meer

FIGURE 442-1

DEFINITION

Chronic fatigue syndrome (CFS) is a disorder characterized by persistent and unexplained fatigue resulting in severe impairment in daily functioning. Besides intense fatigue, most patients with CFS report concomitant symptoms such as pain, cognitive dysfunction, and unrefreshing sleep. Additional symptoms can include headache, sore throat, tender lymph nodes, muscle aches, joint aches, post-exertional malaise, feverishness, difficulty sleeping, psychiatric problems, allergies, and abdominal cramps. Criteria for the diagnosis of CFS have been developed by the U.S. Centers for Disease Control and Prevention (Table 442-1). The Institute of Medicine (IOM; now the National Academy of Medicine) has recently changed the diagnostic criteria and proposed the name systemic exercise intolerance disease (SEID). To date, however, no studies have been reported to demonstrate the usefulness of this change.

TABLE 442-1

Diagnostic Criteria for Chronic Fatigue Syndrome

Characteristic Persistent or Relapsing Unexplained Chronic Fatigue
Fatigue lasts for at least 6 months.
Fatigue is of new or definite onset.
Fatigue is not the result of an organic disease or of continuing exertion.
Fatigue is not alleviated by rest.
Fatigue results in a substantial reduction in previous occupational, educational, social, and personal activities.
Four or more of the following symptoms are concurrently present for 6 months: impaired memory or concentration, sore throat, tender cervical or axillary lymph nodes, muscle pain, pain in several joints, new headaches, unrefreshing sleep, or malaise after exertion.
Exclusion Criteria
Medical condition explaining fatigue
Major depressive disorder (psychotic features) or bipolar disorder
Schizophrenia, dementia, or delusional disorder
Anorexia nervosa, bulimia nervosa
Alcohol or substance abuse
Severe obesity (body mass index >40)

EPIDEMIOLOGY

CFS is seen worldwide, with adult prevalence rates varying between 0.2 and 0.4%. In the United States, the prevalence is higher among women (~75% of cases), members of minority groups (African and Native Americans), and individuals with lower levels of education and occupational status. CFS has been reported to be associated with an increase in mortality from suicide. The mean age of onset is between 29 and 35 years. Many patients probably go undiagnosed and/or do not seek help.

ETIOLOGY

There are numerous hypotheses about the etiology of CFS; there is no definitively identified cause. Distinguishing between predisposing, precipitating, and perpetuating factors in CFS helps to provide a framework for understanding this complex condition **(Table 442-2)**.

TABLE 442-2

Predisposing, Precipitating, and Perpetuating Factors in Chronic Fatigue Syndrome

Predisposing Factors
Childhood trauma (sexual, physical, emotional abuse; emotional and physical neglect)
Physical inactivity during childhood
Premorbid psychiatric illness or psychopathology
Premorbid hyperactivity
↓
Precipitating Factors
Somatic events: infection (e.g., mononucleosis, Q fever, Lyme disease), surgery, pregnancy
Psychosocial stress, life events
↓
Perpetuating Factors
Non-acknowledgment by physician
Negative self-efficacy
Strong physical attributions
Strong focus on bodily symptoms
Fear of fatigue
Lack of social support
Low physical activity pattern

Predisposing Factors

Physical inactivity and trauma in childhood tend to increase the risk of CFS in adults. Neuroendocrine dysfunction may be associated with childhood trauma, reflecting a biological correlate of vulnerability. Psychiatric illness and physical hyperactivity in adulthood raise the risk of CFS in later life. Twin studies suggest a familial predisposition to CFS, but research into causative genes has yielded variable results. In a recent systematic review, a number of single nucleotide polymorphisms encoding for TNFalpha, IL-1beta, IL-6 and IL-4 were found to be associated with increased fatigue in CFS, cancer-related fatigue, and other disease-related fatigue states.

Precipitating Factors

Physical or psychological stress may elicit the onset of CFS. A substantial number of patients report an infection (often a flulike illness or infectious mononucleosis) as the trigger of their fatigue. Relatively high percentages of CFS cases follow Q fever and Lyme disease. However, no differences in Epstein-Barr virus load and immunologic reactivity were found between individuals who developed CFS and those who did not. While antecedent infections are associated with CFS, a direct microbial causality is unproven and unlikely. Patients also often report other precipitating somatic events such as serious injury, surgery, pregnancy, or childbirth. Serious life events, such as the loss of a loved one or a job, military combat, and other stressful situations, may also precipitate CFS. One-third of all patients cannot recall a trigger.

Perpetuating Factors.

Once CFS has developed, numerous factors may impede recovery. Physicians may contribute to chronicity by ordering unnecessary diagnostic procedures, by persistently suggesting psychological causes, and by not acknowledging CFS as a diagnosis.

A patient’s focus on symptoms and avoidance of activities may perpetuate symptoms. A firm belief in a physical cause, a strong focus on bodily sensations, and a poor sense of control over symptoms may also prolong or exacerbate the fatigue and functional impairment. In most patients, inactivity is caused by negative illness perceptions rather than by poor physical fitness. Solicitous behavior of others may reinforce a patient’s illness-related perceptions and behavior. A lack of social support is another known perpetuating factor.

PATHOPHYSIOLOGY

The pathophysiology of CFS is unclear. Neuroimaging studies have suggested that CFS is associated with reduced gray matter volume, but recent studies seem to indicate that pain rather than fatigue is associated with these changes. Functional MRI data have suggested that abnormal patterns of activation correlate with self-reported problems with information processing. Neurophysiologic studies have shown altered CNS activation patterns during muscle contraction.

Evidence for immunologic dysfunction is inconsistent. Modest elevations in titers of antinuclear antibodies, reductions in immunoglobulin subclasses, deficiencies in mitogen-driven lymphocyte proliferation, reductions in natural killer cell activity, disturbances in cytokine production, and shifts in lymphocyte subsets have been described. None of these immune findings has been firmly established and none of these changes appear in most patients, nor does any correlate with the severity of CFS. In theory, symptoms of CFS could result from excessive

production of a cytokine, such as interleukin 1, that induces asthenia and other flulike symptoms; however, compelling data in support of this hypothesis are lacking. There is some evidence that CFS patients have mild hypocortisolism, the degree of which is associated with a poorer response to cognitive-behavioral therapy (CBT). Discrepancies in perceived and actual cognitive performance are consistent findings in patients with CFS.

DIAGNOSIS

In addition to a thorough history, a systematic physical examination is warranted to exclude disorders causing fatigue (e.g., endocrine disorders, neoplasms, heart failure). The heart rate of CFS patients is often slightly above normal, but postural hypotension, which has been included in the diagnostic criteria put forward by the IOM (see above), is not more common in CFS patients than in controls. Laboratory tests serve primarily to exclude other diagnoses; no test can diagnose CFS. The following laboratory screen usually suffices: complete blood count; erythrocyte sedimentation rate; C-reactive protein; serum creatinine, electrolytes, calcium, and iron; blood glucose; creatine kinase; liver function tests; thyroid-stimulating hormone; anti-gliadin antibodies; and urinalysis. Serology for viral or bacterial infections usually is not helpful. No diagnostic abnormalities have been identified on MRI or CT scans. Extensive, unfocused, and expensive testing in a search for the “hidden” cause of the fatigue is not productive. CFS is a constellation of symptoms with no pathognomonic features and remains a diagnosis of exclusion.

Bipolar disorders, schizophrenia, and substance abuse exclude a diagnosis of CFS, as do eating disorders, unless these health problems have been resolved ≥5 years before symptom onset. In addition, CFS is excluded if the chronic fatigue developed immediately after a depressive episode. Depression developing in the course of the fatigue, however, does not preclude CFS. Concurrent psychiatric disorders, especially anxiety and mood disorders, are present in 30–60% of cases.

INITIAL MANAGEMENT

In cases of suspected CFS, the clinician should acknowledge the impact of the patient’s symptoms on daily functioning. Disbelief or denial can provoke an exacerbation of genuine symptoms, which in turn strengthens the clinician’s disbelief, leading to an unfortunate cycle of miscommunication. The possibility of CFS should be considered if a patient fulfills all criteria (Table 442-1) and if other diagnoses have been excluded.

The patient should be asked to describe the symptoms (fatigue and accompanying symptoms) and their duration as well as their consequences (reduction in daily activities). To assess symptom severity and the extent of daily-life impairment, the patient should describe a typical day, from waking to retiring, and, for comparison, an average day prior to symptom onset. Next, potential fatigue-precipitating factors are sought. The severity of fatigue is commonly difficult to assess quantitatively; a brief questionnaire is often helpful **(Fig. 442-1)**.

FIGURE 442-1

Shortened fatigue questionnaire.

How have you felt during the last two weeks?

Please rate all four statements and per statement check the box that reflects your situation best.

1.	I feel tired	Yes, that is true							No, that is not true
2.	I tire easily	Yes, that is true							No, that is not true
3.	I feel fit	Yes, that is true							No, that is not true
4.	Physically I feel exhausted	Yes, that is true							No, that is not true

Scoring:

1, 2 and 4:	Yes, that is true	7	6	5	4	3	2	1	No, that is not true

3: Reversed

Sum scores >18 indicate severe fatigue

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The patient should be informed of the current understanding of precipitating and perpetuating factors and effective treatments and provided general advice about disease management. If CBT for CFS is not available as an initial option (see below) and depression and anxiety are present, these symptoms should be treated. For patients with headache, diffuse pain, and feverishness, nonsteroidal anti-inflammatory drugs may be helpful. Even modest improvements in symptoms can make an important difference in the patient's degree of self-sufficiency and ability to appreciate life's pleasures.

Controlled therapeutic trials have established that acyclovir, fludrocortisone, galantamine, modafinil, and IV immunoglobulin, among other agents, offer no significant benefit in CFS. Countless anecdotes circulate regarding other traditional and nontraditional therapies. It is important to guide patients away from those therapeutic modalities that are toxic, expensive, or unreasonable.

The patient should be encouraged to maintain regular sleep patterns, to remain as active as possible, and to gradually return to previous levels of exercise and other activity (work).

TREATMENT

TREATMENT

Chronic Fatigue Syndrome

CBT and graded exercise therapy (GET) have been found to be the only beneficial interventions in CFS. Some patient groups argue against these approaches because of the implication that CFS is a purely mental disorder. CBT is a psychotherapeutic approach directed at changing unhealthy disease-perpetuating patterns of thoughts and behaviors. It includes educating the patient about the etiologic model, setting goals, restoring fixed bedtimes and wake-up times, challenging and changing fatigue- and activity-related concerns, reducing a focus on symptoms, spreading activities evenly throughout the day, gradually increasing physical activity, planning a return to work, and resuming other activities. The intervention, typically consisting of 12–14 sessions over 6 months performed by an experienced cognitive behavior therapist, helps CFS patients gain control over their symptoms.

GET targets deconditioning and exercise intolerance and usually involves a home exercise program that continues for 3–5 months. Walking or cycling is systematically increased, with set goals for maximal heart rates. Evidence that deconditioning is the basis for symptoms in CFS is lacking, however. CBT and GET appear to improve fatigue primarily by changing the patient's perception of the fatigue and also by reducing the focus on symptoms.

CBT and GET seem equally effective, however not all patients benefit from these interventions. Predictors of poor outcome are insufficient motivation for the treatment, medical (including psychiatric) comorbidities, current disability claims, and severe pain. CBT offered in an early stage of the illness reduces the burden of CFS for the patient as well as for society in terms of decreased medical and disability-related costs.

PROGNOSIS

Full recovery from untreated CFS is rare: the median annual recovery rate is 5% (range, 0–31%), and the median improvement rate is 39% (range, 8–63%). Patients with an underlying psychiatric disorder and those who continue to attribute their symptoms to an undiagnosed medical condition have poorer outcomes.

FURTHER READING

Blundell S et al: Chronic fatigue syndrome and circulating cytokines: A systematic review. Brain Behav Immun 50:186, 2015. [PubMed: 26148446]

Castell BD et al: Cognitive behavioral therapy and graded exercise for chronic fatigue syndrome: A meta-analysis. Clin Psychol Sci Prac 18:311, 2011.

Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome; Board on the Health of Select Populations; Institute of Medicine. Beyond myalgic encephalomyelitis/chronic fatigue syndrome: Redefining an illness. National Academies Press, Washington, DC, 2015.

Larun L et al: Exercise therapy for chronic fatigue syndrome. Cochrane Database Syst Rev 12:CD003200, 2016. [PubMed: 27995604]

Chapter 443: Biology of Psychiatric Disorders

Robert O. Messing; Eric J. Nestler; Matthew W. State

INTRODUCTION

Psychiatric disorders are central nervous system diseases characterized by disturbances in emotion, cognition, motivation, and socialization. They are highly heritable, with genetic risk comprising 20–90% of disease vulnerability. As a result of their prevalence, early onset, and persistence, they contribute substantially to the burden of illness worldwide. All psychiatric disorders are broad heterogeneous syndromes that currently lack well-defined neuropathology and bona fide biologic markers. Therefore, diagnoses continue to be made solely from clinical observations using criteria in the *Diagnostic and Statistical Manual of Mental Disorders*, *Fifth Edition (DSM-5)*, of the American Psychiatric Association.

There is increasing agreement that the classification of psychiatric illnesses in DSM does not accurately reflect the underlying biology of these disorders. Uncertainties in diagnosis complicate efforts to study the genetic basis and attendant neurobiological mechanisms underlying mental illness, though recent advances in genomic and neuroscience technologies along with the consolidation of very large patient cohorts have, for multiple disorders, led to major progress in these realms. In addition, there have been recent efforts to address the limitations of a categorical nosology directly through the development of an alternative diagnostic scheme, termed Research Domain Criteria (RDoC). This system classifies mental illness on the basis of core behavioral abnormalities shared across several syndromes—such as psychosis (loss of reality) or anhedonia (decreased ability to experience pleasure)—and the associated brain circuitry that controls these behavioral domains. It is anticipated that that such classifications will assist in defining the biologic basis of key symptoms. Other factors that have impeded progress in understanding mental illness include the lack of access to pathologic brain tissue except upon death and the inherent limitations of animal models for disorders defined largely by behavioral abnormalities (e.g., hallucinations, delusions, guilt, suicidality) that are inaccessible in animals.

Despite these limitations, the past decade has been marked by real progress. Neuroimaging methods are beginning to provide evidence of brain pathology; genome-wide association studies and high-throughput sequencing are reliably identifying genes and genomic loci that confer risk for severe forms of mental illness; and investigations of better validated animal models, leveraging a host of new methods to study molecular, cellular and circuit level processes, are offering new insight into disease pathogenesis. There is also excitement in the utility of neurons and brain organoids induced in vitro from patient-derived pluripotent stem cells, providing novel ways to study disease pathophysiology and screen for new treatments. There is consequently justified optimism that the field of psychiatry will better integrate behaviorally defined syndromes with an understanding of biological substrates in a way that will drive the development of improved treatments and eventually cures and preventive measures. This chapter describes several examples of recent discoveries in basic neuroscience and genetics that have informed our current understanding of disease mechanisms in psychiatry.

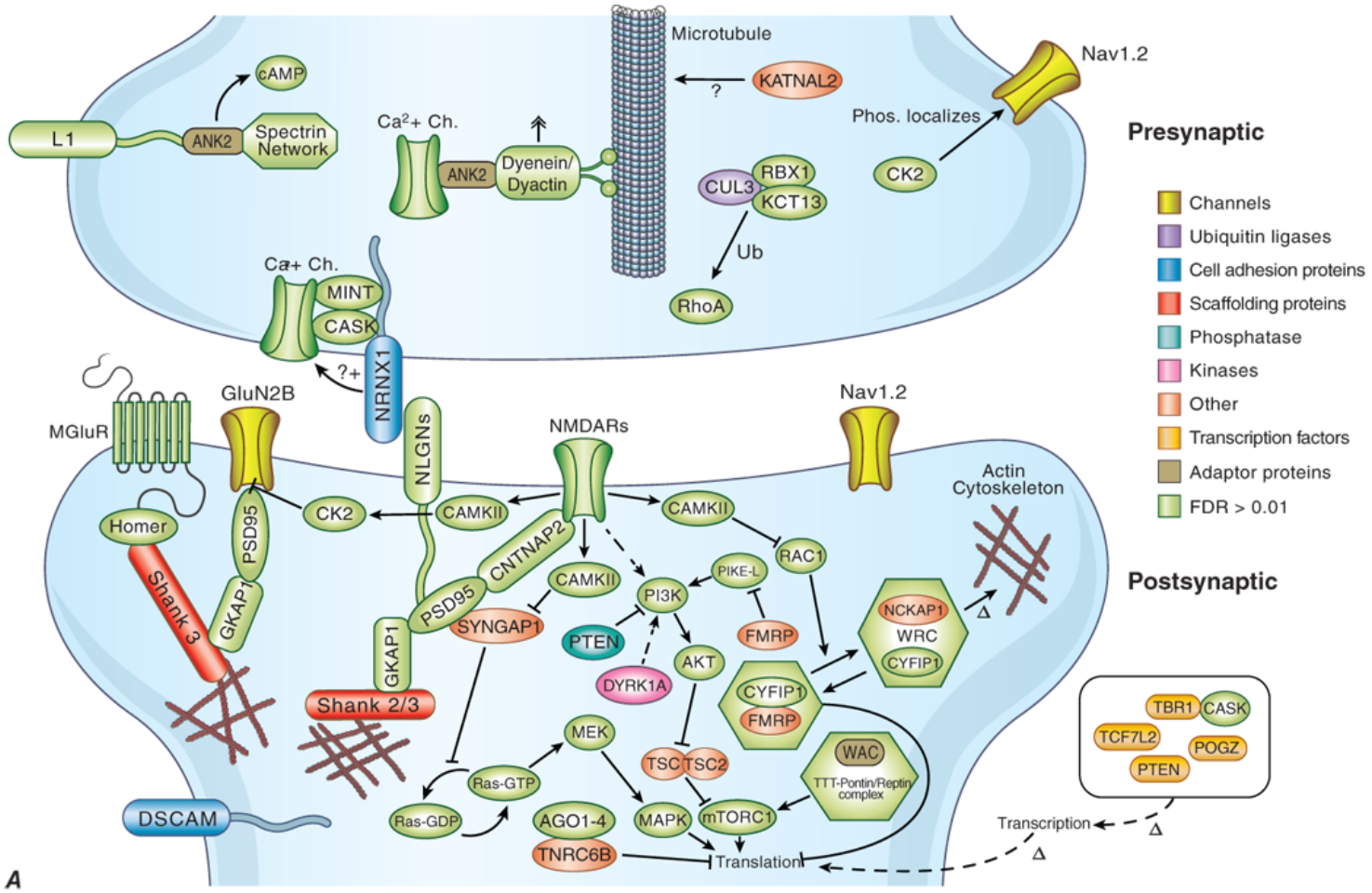
NEUROGENETICS

Because the human brain can only be examined indirectly during life, genome analyses have been extremely important for obtaining molecular clues about the pathogenesis of psychiatric disorders. Moreover, the identification of germ-line risk alleles and mutations provides potential traction on the question of cause versus effect. In other types of cross-sectional studies, it may be impossible to determine whether a phenotype or biomarker observed in affected humans or model systems reflects an etiological factor or a compensatory response. In contrast, germ-line genetic risk is present before the brain develops—at least theoretically allowing for experiments to address temporal sequencing.

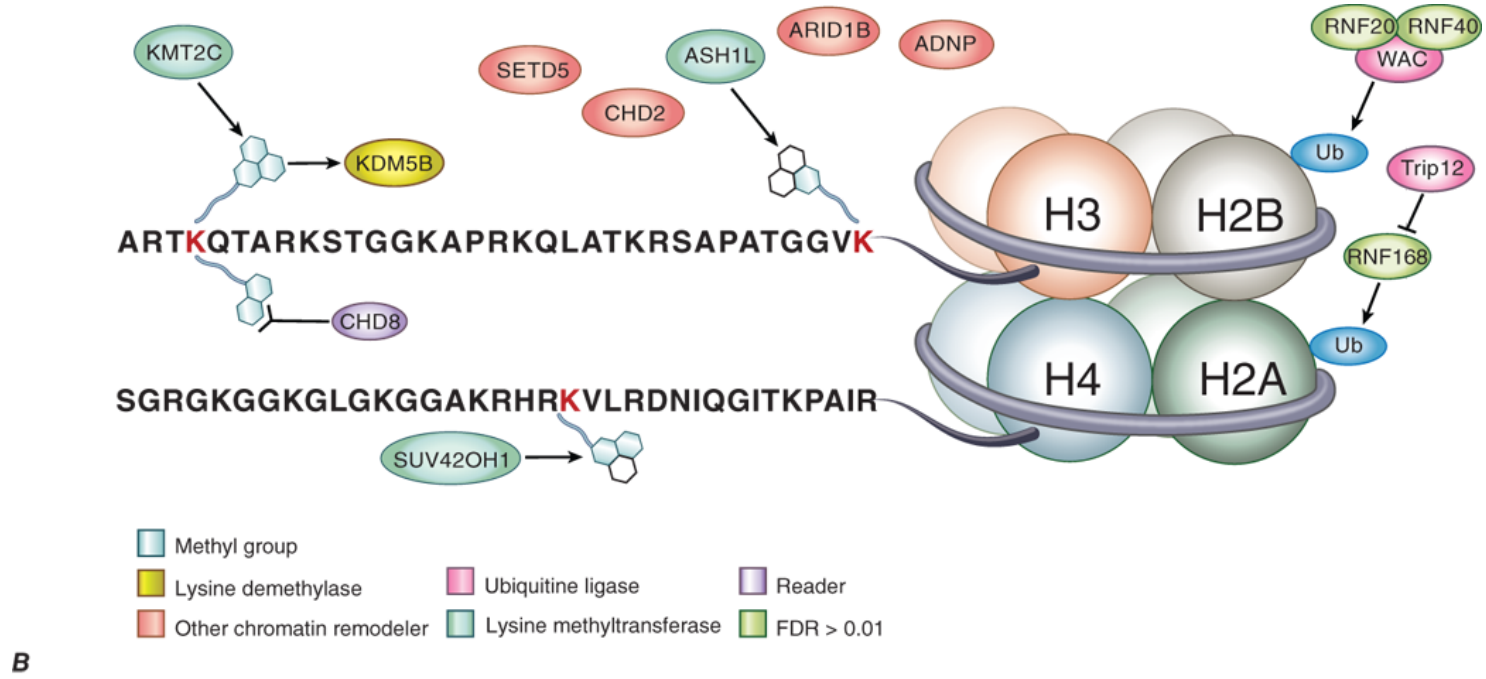
A wealth of new information has been made possible by recent technological developments that have permitted affordable, large-scale genome-wide association studies and high-throughput sequencing. As an example of the latter, significant progress has been made in the genetics of autism spectrum disorders (ASDs), which are a heterogeneous group of neurodevelopmental diseases that share clinical features of impaired social communication and restricted, repetitive patterns of behavior. ASDs are highly heritable; concordance rates in monozygotic twins (~60–90%) are five- to tenfold higher than in dizygotic twins and siblings, and first-degree relatives show approximately tenfold increased risk compared with the general population. ASDs are also genetically heterogeneous. At present, ~70 individual risk genes, along with dozens of submicroscopic deletions and duplications often containing multiple genes, have been identified, almost exclusively through the study of rare, large-effect new (de novo) mutations **(Fig. 443-1)**. All told, genes and genomic regions vulnerable to these types of mutations account for about 20–30% of formerly idiopathic cases that present in the clinic, although none individually accounts for >1%. In addition, ~10% of individuals with ASD have well-described intellectual disability syndromes including *fragile X, Rett syndrome and tuberous sclerosis* **(Chap. 86)**. However, it appears that most of the risk for ASD in the population involves true polygenic inheritance. There is considerable evidence, for example, that >50% of the genetic liability is carried in common alleles of very small individual effect. To date, however, studies of many thousands of cases have yet to identify a reproducible association of a specific nucleotide polymorphism (SNP) using gold standard genome-wide association methods—although with continually increasing cohort sizes, and thus power, this is certain to change in the near future.

FIGURE 443-1

Functional characteristics and developmental convergence of autism spectrum disorders (ASDs) associated genes: genes associated with risk for ASD based on recurrent de novo mutations are shown in ***A*** and ***B*** (Sanders et al: *Neuron* 2015). Those genes encoding proteins meeting criteria for the highest confidence statistical association (false discovery rate [FDR] < 0.01) are highlighted with respect to their putative functions. Additional interacting and functionally related molecules that do not meet this threshold are shown in green. As a group, genes with FDR <0.01 carry large effects, conveying approximately a twentyfold increase in risk. Multiple gene ontology analyses of ASD genes have highlighted both pre- and postsynaptic molecules (***A***) and chromatin modifiers (***B***) as points of enrichment. In ***C***, an alternative strategy for grouping ASD risk genes is highlighted, based on their spatiotemporal expression patterns as opposed to putative functions. One analytic strategy, illustrated in ***C***, leveraged only high confidence ASD genes and examined their developmental expression patterns using the BrainSpan dataset. Convergence for ASD risk was identified in deep layer (V and VI) excitatory neurons in mid-fetal human cortex. Multiple analyses have similarly found glutamatergic neurons in mid-fetal prefrontal cortex as one point of convergence, with somewhat less agreement on layer-specificity and potential additional spatiotemporal points of convergence. *(Figure drawn by Montana Morris and Sarah Pyle.)*

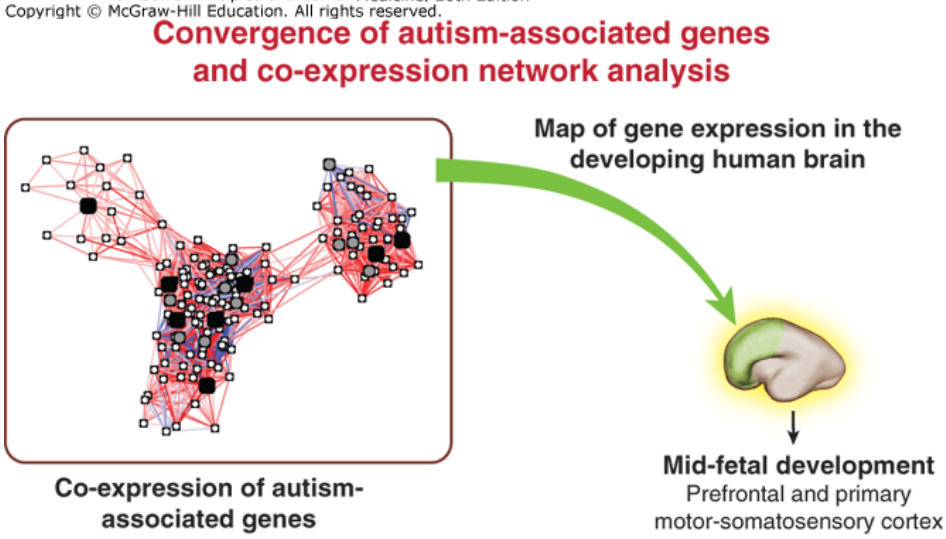


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



C

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Amidst the genetic heterogeneity that has so far been identified, common themes have emerged that inform pathogenesis of ASDs. For instance, many identified rare mutations are in genes that encode proteins involved in synaptic function and early transcriptional regulation (Fig. 443-1) and have a clear relationship to activity-dependent neural responses that can affect the development of neural systems underlying cognition and social behaviors. One particularly intriguing hypothesis is that these genes may lead to ASD risk by changing the balance of excitatory versus inhibitory synaptic signaling in local and extended circuits and by altering mechanisms that control brain growth. Some mutations affect genes (e.g., *PTEN, TSC1*, and *TSC2*) that negatively regulate signaling from several types of extracellular stimuli, including those transduced by receptor tyrosine kinases. Their dysregulation can alter neuronal growth as well as synaptic development and function. Finally, several recent studies have focused on the question of when and where multiple functionally diverse risk genes converge with respect to human brain development. Interestingly, these studies have thus far tended to overlap with expression patterns of glutamatergic neurons in mid-fetal cortex (Fig. 443-1). Given the pleiotropic biological effects of the ASD genes identified to date, an understanding of the developmental or "spatiotemporal" dimensions of risk is likely to serve as a useful complement to studies of the function of individual genes. In short, it may turn out that when and where genetic variation has its impact in the developing brain may be as important as the key processes that are identified.

With further understanding of pathogenesis and the definition of specific ASD subtypes, there is reason to believe that effective therapies will be identified. Work in mouse models has already demonstrated that some autism-like behaviors can be reversed, even in fully developed adult animals, by modifying the underlying pathology; these results encourage hope for many affected individuals. Treatments that target excitation-inhibition imbalance or altered mRNA translation appear to offer early promise. For example, the genes *TSC1, TSC2*, and *PTEN* are negative regulators of signaling through the target of rapamycin

complex 1 (TORC1), which regulates protein synthesis. Rapamycin, a selective inhibitor of TORC1, can reverse several behavioral and synaptic defects in mice carrying null mutations in these genes. Another example is fragile X syndrome, which is the leading cause of inherited autism and mental disability and is due to mutations in *FMR1* that result in loss of the encoded fragile X mental retardation protein (FMRP). FMRP is a polyribosome-associated mRNA-binding protein that represses the translation of a subset (~5%) of all mRNAs, several of which encode proteins that comprise the postsynaptic density, including the metabotropic glutamate receptor 5 (mGluR5). Treatment of *Fmr1* knockout mice with mGluR5 antagonists reduces several behavioral and morphologic abnormalities in these mice; these promising preclinical results have led to ongoing trials of mGluR5 antagonists in humans with fragile X and other ASD syndromes. While early clinical data have been disappointing, this work nonetheless illustrates a potential path forward in therapeutics development.

The ability to catalog common genetic variants and assay them on array-based platforms and, more recently, to carry out whole-exome sequencing has allowed investigators to leverage very large patient cohorts to detect genetic risk loci for schizophrenia and bipolar disorder with genome-wide significance. In contrast to ASD, where the lion's share of early success in gene identification has resulted from the study of rare large effect de novo mutations, much of gene discovery to date for these syndromes has resulted from genome-wide association studies of common inherited polymorphisms. It is noteworthy that there is also striking overlap among the submicroscopic deletions and duplications, called copy number variants (CNVs), that have been found to carry large risks for ASD, schizophrenia, and bipolar disorders, as well as epilepsy and intellectual disability.

To date more than a hundred distinct genomic regions, marked by associated SNPs, have been identified in schizophrenia, some of which show risk as well for bipolar disorder. Several of the identified genes are parts of molecular complexes, such as voltage-gated calcium channels (in particular, *CACNA1C* and *CACNB2*) and the postsynaptic density of excitatory synapses. Genes that promote risk for addiction and depression have also begun to emerge from large studies. The best-established susceptibility locus for addiction is the *CHRNA5-A3-B4* nicotinic acetylcholine receptor gene cluster on chromosome 15 associated with nicotine and alcohol addiction. Recent genome-wide association studies of depression have required hundreds of thousands of cases and controls to identify the first statistically significant loci using state-of-the-art approaches. These findings collectively point to the tremendous heterogeneity of depressive disorders as well as the very small biological effects conferred by any individual common allele.

A recurrent theme that has emerged from genetic studies of psychiatric disorders is phenotypic pleiotropy, namely, that many genes are associated with multiple psychiatric syndromes. For example, mutations in *MECP2, FMR1*, and *TSC1* and *TSC2* (see **Table 443-1** for abbreviations) can cause mental retardation without ASD, others in *MECP2* can cause obsessive-compulsive and attention-deficit hyperactivity disorders, some alleles of *NRXN1* are associated with symptoms of both ASD and schizophrenia, and common polymorphisms in *CACNA1C* are strongly associated with both schizophrenia and bipolar disorder. Likewise, duplication of chromosome 16p is associated with both schizophrenia and autism, whereas deletions in the DiGeorge's (velocardiofacial) syndrome region are associated with schizophrenia, autism, and bipolar disorder. The association of genes and genomic regions with multiple syndromes attests to the complexity of psychiatric disorders, the very large gap between molecular mechanisms and the current categorical diagnostic schemes, and the influence of additional factors that combine to specify the ultimate phenotype. The latter might include polygenic "background," variations in regulatory regions of the genome that determine cell-type specificity and timing of gene expression, protective variants, stochastic events, and epigenetic effects.

TABLE 443-1

Initial Actions of Drugs of Abuse

Drug	Neurotransmitter Affected	Drug Target (Action)
Opiates	Endorphins, enkephalins	μ- and δ-opioid receptors (agonist)
Psychostimulants (cocaine, amphetamine, methamphetamine)	Dopamine	Dopamine transporter (antagonist—cocaine; reverse transport—amphetamine, methamphetamine)
Nicotine	Acetylcholine	Nicotinic cholinergic receptors (agonist)
Ethanol	GABA	$GABA_A$ receptors (positive allosteric modulator)
	Glutamate	NMDA glutamate receptors (antagonist)
	Acetylcholine	Nicotinic cholinergic receptors (allosteric modulator)
	Serotonin	5HT-3 receptor (positive allosteric modulator)
	Others	Calcium-activated K^+ channel (activator)
Marijuana	Endocannabinoids (anandamide, 2-arachidonoylglycerol)	CB_1 receptor (agonist)
Phencyclidine	Glutamate	NMDA glutamate receptor (antagonist)

Abbreviations: GABA, γ-aminobutyric acid; NMDA, *N*-methyl-D-aspartate

SIGNAL TRANSDUCTION

Studies of signal transduction have revealed numerous intracellular signaling pathways that are perturbed in psychiatric disorders, and such research has provided insight into development of new therapeutic agents. For example, lithium is a highly effective drug for bipolar disorder and competes with magnesium to inhibit numerous magnesium-dependent enzymes, including the enzyme GSK3β and several enzymes involved in phosphoinositide signaling that lead to activation of protein kinase C. These findings have led to discovery programs focused on developing GSK3β or protein kinase C inhibitors as potential novel treatments for mood disorders, although none have demonstrated clinical efficacy to date.

The observations that tricyclic antidepressants (e.g., imipramine) inhibit serotonin and/or norepinephrine reuptake and that monoamine oxidase inhibitors (e.g., tranylcypromine) are effective antidepressants initially led to the view that depression is caused by a deficiency of these monoamines. However, this hypothesis has not been substantiated. A cardinal feature of these drugs is that long-term (weeks to months) administration is needed for their antidepressant effects. This means that their short-term actions, namely promotion of serotonin or norepinephrine function, are not per se antidepressant but rather induce a cascade of adaptations in the brain that underlie their slowly developing clinical effects. The nature of these therapeutic drug-induced adaptations has not been identified with certainty. One theory holds that, in a subset of depressed patients who display upregulation of the hypothalamic-pituitary-adrenal (HPA) axis characterized by increased secretion of corticotropin-releasing factor (CRF) and glucocorticoids, excessive glucocorticoids cause atrophy of hippocampal neurons, which is associated with reduced hippocampal volumes seen clinically. Chronic antidepressant administration might reverse this atrophy by increasing brain-derived neurotrophic factor (BDNF) or a host of other neurotrophic factors in hippocampus. A role for stress-induced decreases in the generation of newly born hippocampal granule cell neurons, and its reversal by antidepressants through BDNF or other growth factors, has also been suggested.

A major advance in recent years has been the identification of several rapidly acting antidepressants with non-monoamine-based mechanisms of action. The best established is ketamine, a noncompetitive antagonist of *N*-methyl-D-aspartate (NMDA) glutamate receptors, which exerts rapid (hours) and robust antidepressant effects in severely depressed patients who have not responded to other treatments. Ketamine, which at higher doses is psychotomimetic and anesthetic, exerts these antidepressant effects at lower doses with minimal side effects. However, the response to ketamine is transient, which has led to several approaches to maintain treatment response, such as repeated ketamine delivery. The mechanism underlying ketamine's antidepressant action is not known, and its action as an NMDA receptor antagonist has recently been called into question. Nevertheless, ketamine's striking clinical efficacy has stimulated animal research on the role of glutamate neurotransmission and synaptic plasticity in key limbic regions. Recent evidence supports a role for

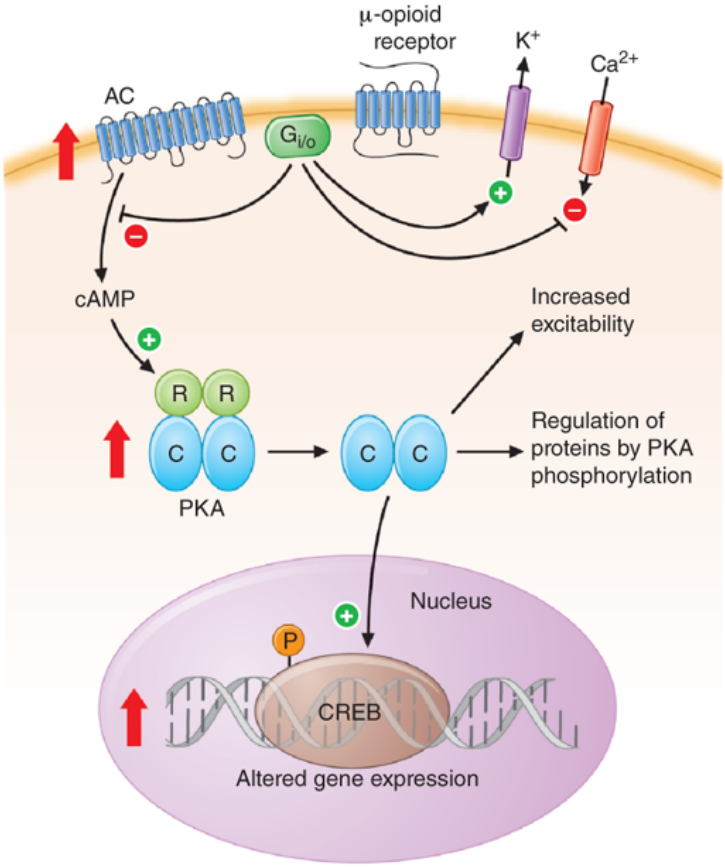
TORC1 or BDNF activation, as blockade of either blocks the antidepressant-like effects of ketamine in animal models. Mechanisms by which ketamine activates these signaling cascades are currently an active area of investigation.

A major goal in the field of drug abuse has been to identify neuroadaptive mechanisms that lead from recreational use to addiction. Such research has determined that repeated intake of abused drugs induces specific changes in cellular signal transduction, leading to changes in synaptic strength (long-term potentiation or depression) and neuronal structure (altered dendritic branching or cell soma size) within the brain's reward circuitry. These drug-induced modifications are mediated in part by changes in gene expression, achieved by regulation of transcription factors (e.g., CREB [cAMP response element-binding protein] and ΔFosB [a Fos family protein]) and their target genes. Such alterations in gene expression are associated with lasting alterations in epigenetic modifications, including histone acetylation and methylation and DNA methylation. These adaptations provide opportunities for developing treatments targeted to drug-addicted individuals. The fact that the spectrum of these adaptations differs in part depending on the particular addictive substance used raises hope that treatments could be developed that are specific for different classes of addictive drugs and less likely to disturb basic mechanisms that govern normal motivation and reward.

Increasingly, causal relationships are being established between individual molecular and cellular adaptations and specific behavioral abnormalities that characterize the addicted state. For example, acute activation of μ-opioid receptors by morphine or other opiates activates $G_{i/o}$ proteins, leading to inhibition of adenylyl cyclase (AC), resulting in reduced cyclic AMP (cAMP) production, protein kinase A (PKA) activation, and activation of the transcription factor CREB. Repeated administration of these drugs **(Fig. 443-2)** evokes a homeostatic response involving upregulation of ACs and PKA and increased activation of CREB. Such upregulation of cAMP-CREB signaling has been identified in the locus coeruleus (LC), periaqueductal gray, ventral tegmental area (VTA), nucleus accumbens (NAc), and several other CNS regions, and contributes to opiate craving and signs of opiate withdrawal. The fact that endogenous opioid peptides do not produce tolerance and dependence, while morphine and heroin do, may relate to the observation that, unlike endogenous opioids, morphine and heroin are weak inducers of μ-opioid receptor desensitization and endocytosis. Therefore, these drugs cause prolonged receptor activation and inhibition of ACs, which provides a powerful stimulus for the upregulation of cAMP-CREB signaling that characterizes the opiate-dependent state.

FIGURE 443-2

Opiate action in the locus coeruleus (LC). Binding of opiate agonists to μ-opioid receptors catalyzes nucleotide exchange on G_i and G_o proteins, leading to inhibition of adenylyl cyclase (AC), neuronal hyperpolarization via activation of K^+ channels, and inhibition of neurotransmitter release via inhibition of Ca^{2+} channels. Inhibition of AC reduces protein kinase A (PKA) activity and phosphorylation of several PKA substrate proteins, thereby altering their function. For example, opiates reduce phosphorylation of the cAMP response element-binding protein (CREB), which initiates longer term changes in neuronal function. Chronic administration of opiates increases levels of AC isoforms, PKA catalytic (C) and regulatory (R) subunits, and the phosphorylation of several proteins, including CREB (indicated by *red arrows*). These changes contribute to the altered phenotype of the drug-addicted state. For example, the excitability of LC neurons is increased by enhanced cAMP signaling. Activation of CREB causes upregulation of AC isoforms and tyrosine hydroxylase, the rate-limiting enzyme in catecholamine biosynthesis.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SYSTEMS NEUROSCIENCE

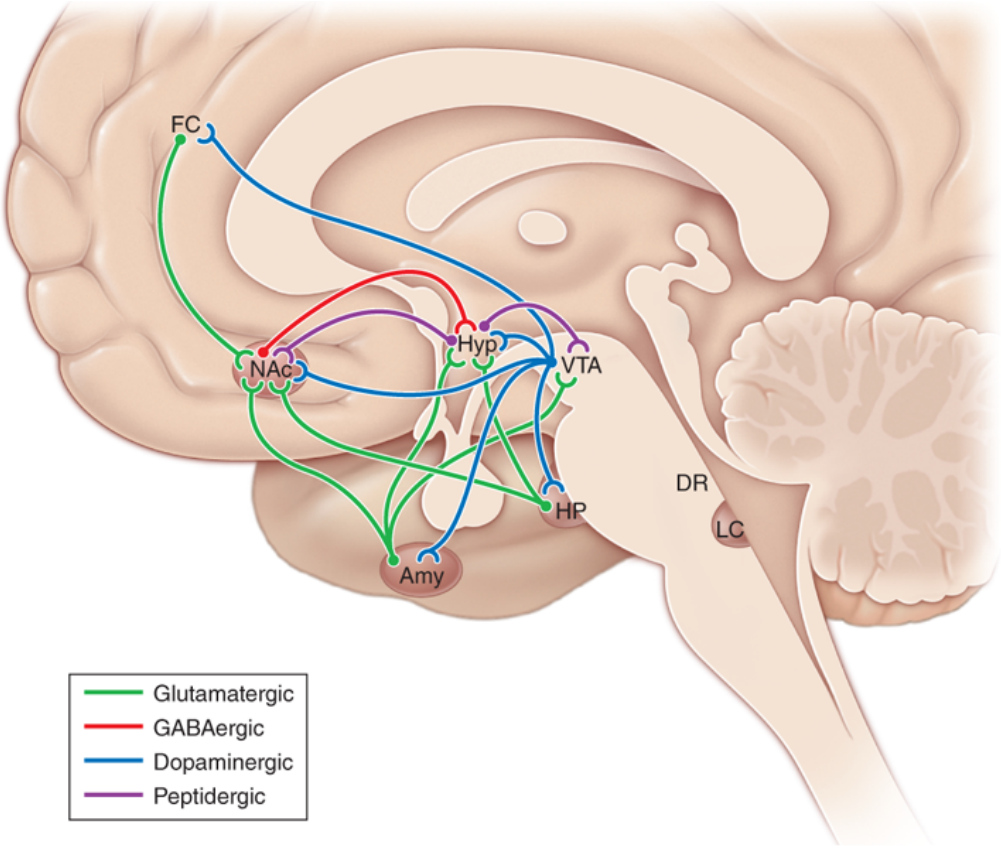
The study of interconnected brain circuits that drive behavior has been greatly advanced through newer methods in brain imaging that have documented abnormalities in neural function and connectivity in psychiatric disorders. Electroceutical devices, which use electrical or magnetic stimulation to control neuronal activity, have had some success in depression, obsessive compulsive disorder, pain, and addiction. The past decade has also witnessed the development of revolutionary new techniques—optogenetics, designer receptors and ligands—that provide unprecedented temporal and spatial control of neural circuits. The development of genetically encoded calcium detectors and electrode arrays has allowed in vivo monitoring of thousands of neurons in multiple brain regions simultaneously. Advances in histology and microscopy now permit three-dimensional imaging of specific proteins in the intact brain, while advances in endoscopic microscopy allow imaging of hundreds of neurons within deep brain structures in awake, freely moving animals. These new methods promise to revolutionize our ability to understand the circuit basis of brain function.

Positron emission tomography (PET), diffusion tensor imaging (DTI), and functional magnetic resonance imaging (fMRI) have identified neural circuits that contribute to psychiatric disorders, for example, defining the neural circuitry of mood within the brain's limbic system **(Fig. 443-3)**. Integral to this system are the NAc (important also for brain reward—see below), amygdala, hippocampus, and regions of prefrontal cortex. Recent optogenetic research in animals, where the activity of specific types of neurons in defined circuits can be controlled with light, has confirmed the importance of this limbic circuitry in controlling depression-related behavioral abnormalities. Given that many symptoms of depression (so-called neurovegetative symptoms) involve physiologic functions, a key role for the hypothalamus is presumed as well. A subset of depressed individuals shows a small reduction in hippocampal size, as noted above. In addition, brain imaging investigations have revealed increased activation of the amygdala by negative stimuli and reduced activation of the NAc by rewarding stimuli. There is also evidence for altered activity in prefrontal cortex, such as hyperactivity of subgenual area 25 in anterior cingulate cortex. Such findings have led to trials of deep brain stimulation (DBS) of either the NAc or subgenual area 25, which appears to be therapeutic in some severely depressed individuals.

FIGURE 443-3

Neural circuitry of depression and addiction. The figure shows a simplified summary of a series of limbic circuits in brain that regulate mood and motivation and are implicated in depression and addiction. Shown in the figure are the hippocampus (HP) and amygdala (Amy) in the temporal lobe, regions of prefrontal cortex, NAc, and hypothalamus (Hyp). Only a subset of the known interconnections among these brain regions is shown. Also shown is the innervation of several of these brain regions by monoaminergic neurons. The ventral tegmental area (VTA) provides dopaminergic input to each of the limbic structures. Norepinephrine (from the locus coeruleus [LC]) and serotonin (from the dorsal raphe [DR] and other raphe nuclei) innervate all of the

regions shown. In addition, there are strong connections between the hypothalamus and the VTA–NAc pathway. Important peptidergic projections from the hypothalamus include those from the arcuate nucleus that release β-endorphin and melanocortin and from the lateral hypothalamus that release orexin.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In schizophrenia, structural and functional imaging studies have confirmed earlier pathologic studies that show enlargement of the ventricular system and reduction of cortical and subcortical gray matter in frontal and temporal lobes and in the limbic system. Functional imaging studies show reduced metabolic (presumably neural) activity in the dorsolateral prefrontal cortex at rest and when performing tests of executive function, including working memory. There is also evidence for impaired structural and task-related functional connectivity, mainly in frontal and temporal lobes. The reduction in cortical thickness seen in schizophrenia is associated with increased cell packing density and reduced neuropil (defined as axons, dendrites, and glial cell processes) without an apparent change in neuronal cell number. Specific classes of interneurons in prefrontal cortex consistently show reduced expression of the gene encoding the enzyme glutamic acid decarboxylase 1 (*GAD1*), which synthesizes γ-aminobutyric acid (GABA), the principal inhibitory neurotransmitter in the brain. Recently, results from well-powered genome-wide association studies point to synaptic pruning as a potential contributing mechanism. In the region of the genome most strongly statistically associated with schizophrenia risk, variations in the relative expression of two isotypes of complement component 4, C4A and C4B, have been found to account for a significant proportion of this genetic signal. Studies of loss of C4 in mice show deficient synaptic pruning, leading to the hypothesis that increased expression of C4A in humans may result in excessive synaptic pruning. Such results point to the potential for a gene-driven understanding of pathophysiology; however, the findings also leave some important questions unanswered. The strongest effect haplotype in humans still only accounts for a very small increase in risk, with an odds ratio of less than 1.3. In contrast, having a sibling with schizophrenia increases risk approximately tenfold. In short, whether this allele reflects a driving pathophysiological mechanism remains to be determined. Moreover, humans have diverged at the C4 locus compared with rodents such that only a single C4 isotype is present in the mouse, preventing any analysis of the putative effects of changing the ratio of C4A to C4B—the phenomenon associated with disease risk in humans. Nonetheless, all the aforementioned findings support the notion that schizophrenia is a developmental neurodegenerative disorder with some evidence pointing to loss of cortical interneurons in frontal and temporal lobes.

Work in rodent and nonhuman primate models of addiction has established the brain's reward regions as key neural substrates for the acute actions of drugs of abuse and for addiction induced in vulnerable individuals by repeated drug administration (Fig. 443-3). Midbrain dopamine neurons in the VTA function normally as rheostats of reward: they are activated by natural rewards (food, sex, social interaction) or even by the expectation of such rewards, and many are suppressed by the absence of an expected reward or by aversive stimuli. These neurons thereby transmit crucial survival signals to the rest of the limbic brain to promote reward-related behavior, including motor responses to seek and obtain the rewards (NAc), memories of reward-related cues (amygdala, hippocampus), and executive control of obtaining rewards (prefrontal cortex).

Drugs of abuse alter neurotransmission through initial actions at different classes of ion channels, neurotransmitter receptors, or neurotransmitter transporters (Table 443-1). Studies in animal models have demonstrated that although the initial targets differ, the actions of these drugs converge on the

brain's reward circuitry by promoting dopamine neurotransmission in the NAc and other limbic targets of the VTA. In addition, some drugs promote activation of opioid and cannabinoid receptors, which modulate this reward circuitry. By these mechanisms, drugs of abuse produce powerful rewarding signals, which, after repeated drug administration, corrupt a vulnerable brain's reward circuitry in ways that promote addiction. Three major pathologic adaptations have been described. First, drugs produce tolerance and dependence in reward circuits, which promote escalating drug intake and a negative emotional state during drug withdrawal that promotes relapse. Second, sensitization to the rewarding effects of the drugs and associated cues is seen during prolonged abstinence and also triggers relapse. Third, executive function is impaired in such a way as to increase impulsivity and compulsivity, both of which promote relapse.

Imaging studies in humans confirm that addictive drugs, as well as craving for them, activate the brain's reward circuitry. In addition, patients who abuse alcohol or psychostimulants show reduced gray matter in the prefrontal cortex as well as reduced activity in anterior cingulate and orbitofrontal cortex during tasks of attention and inhibitory control. It is thought that damage to these cortical areas contributes to addiction by impairing decision-making and increasing impulsivity.

NEUROINFLAMMATION

There is increasing evidence for the involvement of inflammatory mechanisms in a wide range of psychiatric syndromes. For example, a subset of depressed patients display elevated blood levels of interleukin 6 (IL-6), tumor necrosis factor α (TNF-α), and other cytokines. Moreover, rodents exposed to chronic stress exhibit similar increases in peripheral cytokines, and peripheral or central delivery of those cytokines to normal rodents increases their susceptibility to chronic stress. These findings have led to the novel idea of using peripheral cytokines as biomarkers of a subtype of depression and the potential utility of developing new antidepressants that oppose cytokine action.

Recent evidence has also linked proinflammatory signaling in the brain to addiction, particularly to alcohol. Human alcoholism is associated with impaired innate immunity, increases in circulating proinflammatory cytokines, and increases in brain expression of several immune-related genes. Many of these genes are expressed by astrocytes and microglia, and by neurons under certain pathologic conditions, where they play important roles in modifying neuronal function and plasticity. For example, cytokine monocyte chemotactic protein-1 (MCP-1) modulates the release of certain neurotransmitters and, when administered into the VTA, increases neuronal excitability, promotes dopamine release, and increases locomotor activity. Recent gene expression studies of alcohol drinking in mice have identified a network of regulated neuroimmune proteins in brain, and a role in regulation of alcohol consumption has been recently validated for several, including chemokines MCP-1 and chemokine (C-C motif) ligand 3 (CCL3), beta-2 microglobulin, CD14, IL-1 receptor antagonist, and cathepsins S and F. This work has led to discovery of anti-inflammatory medications that reduce alcohol intake in animals, such as agonists of peroxisome proliferator-activated receptors (PPARs), which are transcription factors that repress key inflammatory signaling molecules such as nuclear factor κB (NFκB) and nuclear factor of activated T cells (NFAT). A major focus of current research is to define the sites and mechanisms by which proinflammatory cytokines impair brain function to elicit a depressive episode or promote drug abuse.

CONCLUSIONS

This brief narrative illustrates the substantial progress that is being made in understanding the genetic and neurobiological basis of mental illness. It is anticipated that biologic measures will be used increasingly to diagnose and subtype psychiatric disorders and that targeted therapeutics will become available for these complex conditions.

FURTHER READING

Gandal MJ et al: The road to precision psychiatry: Translating genetics into disease mechanisms. Nat Neurosci 19:1397, 2016. [PubMed: 27786179]

Koob GF, Volkow ND: Neurobiology of addiction: A neurocircuitry analysis. Lancet Psychiatry 3:760, 2016. [PubMed: 27475769]

Rajasethupathy P et al: Targeting neural circuits. Cell 165:524, 2016. [PubMed: 27104976]

Ron D, Barak S: Molecular mechanisms underlying alcohol-drinking behaviours. Nat Rev Neurosci 17:576, 2016. [PubMed: 27444358]

Sanders SJ et al: Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron 87:1215, 2015. [PubMed: 26402605]

Schizophrenia Working Group of the Psychiatric Genomics Consortium: Biological insights from 108 schizophrenia-associated genetic loci. Nature 511:421, 2014. [PubMed: 25056061]

Wohleb ES et al: Integrating neuroimmune systems in the neurobiology of depression. Nat Rev Neurosci 17:497, 2016. [PubMed: 27277867]

Chapter 444: Psychiatric Disorders

Victor I. Reus

FIGURE 444-1

INTRODUCTION

Psychiatric disorders are common in medical practice and may present either as a primary disorder or as a comorbid condition. The prevalence of mental or substance use disorders in the United States is ~30%, but only one-third of affected individuals are currently receiving treatment. Global burden of disease statistics indicates that 4 of the 10 most important causes of morbidity and attendant health care costs worldwide are psychiatric in origin.

Changes in health care delivery underscore the need for primary care physicians to assume responsibility for the initial diagnosis and treatment of the most common mental disorders. Prompt diagnosis is essential to ensure that patients have access to appropriate medical services and to maximize the clinical outcome. Validated patient-based questionnaires have been developed that systematically probe for signs and symptoms associated with the most prevalent psychiatric diagnoses and guide the clinician into targeted assessment. The Primary Care Evaluation of Mental Disorders (PRIME-MD; and a self-report form, the Patient Health Questionnaire) and the Symptom-Driven Diagnostic System for Primary Care (SDDS-PC) are inventories that require only 10 min to complete and link patient responses to the formal diagnostic criteria of anxiety, mood, somatoform, and eating disorders and to alcohol abuse or dependence.

A physician who refers patients to a psychiatrist should know not only when doing so is appropriate but also how to refer because societal misconceptions and the stigma of mental illness impede the process. Primary care physicians should base referrals to a psychiatrist on the presence of signs and symptoms of a mental disorder and not simply on the absence of a physical explanation for a patient's complaint. The physician should discuss with the patient the reasons for requesting the referral or consultation and provide reassurance that he or she will continue to provide medical care and work collaboratively with the mental health professional. Consultation with a psychiatrist or transfer of care is appropriate when physicians encounter evidence of psychotic symptoms, mania, severe depression, or anxiety; symptoms of posttraumatic stress disorder (PTSD); suicidal or homicidal preoccupation; or a failure to respond to first-order treatment. This chapter reviews the clinical assessment and treatment of some of the most common mental disorders presenting in primary care and is based on the *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition (DSM-5), the framework for categorizing psychiatric illness used in the United States.

Eating disorders are discussed later in this chapter, and the biology of psychiatric and addictive disorders is discussed in Chap. 443.

GLOBAL CONSIDERATIONS


The DSM-5 and the tenth revision of the International Classification of Diseases (ICD-10), which is used more commonly worldwide, have taken somewhat differing approaches to the diagnosis of mental illness, but considerable effort has been expended to provide an operational translation between the two nosologies. Both systems are in essence purely descriptive and emphasize clinical pragmatism, in distinction to the Research Domain Criteria (RDOC) proposed by National Institute of Mental Health, which aspires to provide a causal framework for classification of behavioral disturbance. None of these diagnostic systems has as yet achieved adequate validation. The Global Burden of Disease Study 2010, using available epidemiologic data, nevertheless has reinforced the conclusion that, regardless of nosologic differences, mental and substance abuse disorders are the major cause of life-years lost to disability among all medical illnesses. There is general agreement that high-income countries will need to build capacity in professional training in low- and middle-income countries in order to provide an adequate balanced care model for the delivery of evidence-based therapies for mental disorders. Recent surveys that indicate a dramatic increase in mental disorder prevalence in rapidly developing countries, such as China, may reflect both an increased recognition of the issue, but also the consequence of social turmoil, stigma, and historically inadequate resources. The need for improved prevention strategies and for more definitive and effective interventional treatments remains a global concern.

ANXIETY DISORDERS

Anxiety disorders, the most prevalent psychiatric illnesses in the general community, are present in 15–20% of medical clinic patients. Anxiety, defined as a subjective sense of unease, dread, or foreboding, can indicate a primary psychiatric condition or can be a component of, or reaction to, a primary medical disease. The primary anxiety disorders are classified according to their duration and course and the existence and nature of precipitants.

When evaluating the anxious patient, the clinician must first determine whether the anxiety antedates or postdates a medical illness or is due to a medication side effect. Approximately one-third of patients presenting with anxiety have a medical etiology for their psychiatric symptoms, but an anxiety disorder can also present with somatic symptoms in the absence of a diagnosable medical condition.

PANIC DISORDER

Clinical Manifestations

Panic disorder is defined by the presence of recurrent and unpredictable panic attacks, which are distinct episodes of intense fear and discomfort associated with a variety of physical symptoms, including

palpitations, sweating, trembling, shortness of breath, chest pain, dizziness, and a fear of impending doom or death. Paresthesias, gastrointestinal distress, and feelings of unreality are also common. Diagnostic criteria require at least 1 month of concern or worry about the attacks or a change in behavior related to them. The lifetime prevalence of panic disorder is 2–3%. Panic attacks have a sudden onset, developing within 10 min and usually resolving over the course of an hour, and they occur in an unexpected fashion. Some may occur when waking from sleep. The frequency and severity of panic attacks vary, ranging from once a week to clusters of attacks separated by months of well-being. The first attack is usually outside the home, and onset is typically in late adolescence to early adulthood. In some individuals, anticipatory anxiety develops over time and results in a generalized fear and a progressive avoidance of places or situations in which a panic attack might recur. *Agoraphobia*, which occurs commonly in patients with panic disorder, is an acquired irrational fear of being in places where one might feel trapped or unable to escape. It may, however, be diagnosed even if panic disorder is not present. Typically, it leads the patient into a progressive restriction in lifestyle and, in a literal sense, in geography. Frequently, patients are embarrassed that they are housebound and dependent on the company of others to go out into the world and do not volunteer this information; thus, physicians will fail to recognize the syndrome if direct questioning is not pursued.

Differential Diagnosis

A diagnosis of panic disorder is made after a medical etiology for the panic attacks has been ruled out. A variety of cardiovascular, respiratory, endocrine, and neurologic conditions can present with anxiety as the chief complaint. Patients with true panic disorder will often focus on one specific feature to the exclusion of others. For example, 20% of patients who present with syncope as a primary medical complaint have a primary diagnosis of a mood, anxiety, or substance abuse disorder, the most common being panic disorder. The differential diagnosis of panic disorder is complicated by a high rate of comorbidity with other psychiatric conditions, especially alcohol and benzodiazepine abuse, which patients initially use in an attempt at self-medication. Some 75% of panic disorder patients will also satisfy criteria for major depression at some point in their illness.

When the history is nonspecific, physical examination and focused laboratory testing must be used to rule out anxiety states resulting from medical disorders such as pheochromocytoma, thyrotoxicosis, or hypoglycemia. Electrocardiogram (ECG) and echocardiogram may detect some cardiovascular conditions associated with panic such as paroxysmal atrial tachycardia and mitral valve prolapse. In two studies, panic disorder was the primary diagnosis in 43% of patients with chest pain who had normal coronary angiograms and was present in 9% of all outpatients referred for cardiac evaluation. Panic disorder has also been diagnosed in many patients referred for pulmonary function testing or with symptoms of irritable bowel syndrome.

Etiology and Pathophysiology

The etiology of panic disorder is unknown but appears to involve a genetic predisposition, altered autonomic responsivity, and social learning. Panic disorder shows familial aggregation; the disorder is concordant in 30–45% of monozygotic twins, and genome-wide screens have identified suggestive risk loci. Acute panic attacks

appear to be associated with increased noradrenergic discharges in the locus coeruleus. Intravenous infusion of sodium lactate evokes an attack in two-thirds of panic disorder patients, as do the α_2-adrenergic antagonist yohimbine, cholecystokinin tetrapeptide (CCK-4), and carbon dioxide inhalation. It is hypothesized that each of these stimuli activates a pathway involving noradrenergic neurons in the locus coeruleus and serotonergic neurons in the dorsal raphe. Resting state fMRI has identified abnormalities in the default mode network involving the medial temporal lobe, with greater activation in the sensorimotor cortex in panic disorder and in amygdala-frontal connectivity in social anxiety disorder. Agents that block serotonin reuptake can prevent attacks. Patients with panic disorder have a heightened sensitivity to somatic symptoms, which triggers increasing arousal, setting off the panic attack; accordingly, therapeutic intervention involves altering the patient's cognitive interpretation of anxiety-producing experiences as well as preventing the attack itself.

TREATMENT

TREATMENT

Panic Disorder

Achievable goals of treatment are to decrease the frequency of panic attacks and to reduce their intensity. The cornerstone of drug therapy is antidepressant medication **(Tables 444-1, Tables 444-2, 444-3)**. Selective serotonin reuptake inhibitors (SSRIs) benefit the majority of panic disorder patients and do not have the adverse effects of tricyclic antidepressants (TCAs). Fluoxetine, paroxetine, sertraline, and the selective serotonin-norepinephrine reuptake inhibitor (SNRI) venlafaxine have received approval from the U.S. Food and Drug Administration (FDA) for this indication. These drugs should be started at one-third to one-half of their usual antidepressant dose (e.g., 5–10 mg fluoxetine, 25–50 mg sertraline, 10 mg paroxetine, venlafaxine 37.5 mg). Monoamine oxidase inhibitors (MAOIs) are also effective and may specifically benefit patients who have comorbid features of atypical depression (i.e., hypersomnia and weight gain). Insomnia, orthostatic hypotension, and the need to maintain a low-tyramine diet (avoidance of cheese and wine) have limited their use, however. Antidepressants typically take 2–6 weeks to become effective, and doses may need to be adjusted based on the clinical response.

Because of anticipatory anxiety and the need for immediate relief of panic symptoms, benzodiazepines are useful early in the course of treatment and sporadically thereafter **(Table 444-4)**. FDA-approved agents include alprazolam and clonazepam. A recent Cochrane review found no difference between antidepressants and benzodiazepines in response rate, although benzodiazepines were somewhat better tolerated by patients. In treatment resistant cases, short-term augmentation with aripiprazole, divalproex sodium, or pindolol has some evidence for efficacy. There also is no clear difference in short-term efficacy between psychological therapies and antidepressant or benzodiazepine treatment, alone or in combination,

Early psychotherapeutic intervention and education aimed at symptom control enhance the effectiveness of drug treatment. Patients can be taught breathing techniques, be educated about physiologic changes that occur with panic, and learn to expose themselves voluntarily to precipitating events in a treatment program

spanning 12–15 sessions. Homework assignments and monitored compliance are important components of successful treatment. Once patients have achieved a satisfactory response, drug treatment should be maintained for 1–2 years to prevent relapse. Controlled trials indicate a success rate of 75–85%, although the likelihood of complete remission is somewhat lower.

TABLE 444-1

Antidepressants

Name	Usual Daily Dose, mg	Side Effects	Comments
SSRIs			
Fluoxetine (Prozac)	10–80	Headache; nausea and other GI effects; jitteriness; insomnia; sexual dysfunction; can affect plasma levels of other medicines (except sertraline); akathisia rare	Once-daily dosing, usually in the morning; fluoxetine has very long half-life; must not be combined with MAOIs
Sertraline (Zoloft)	50–200		
Paroxetine (Paxil)	20–60		
Fluvoxamine (Luvox)	100–300		
Citalopram (Celexa)	20–60		
Escitalopram (Lexapro)	10–30		
TCAs and Tetracyclics			
Amitriptyline (Elavil)	150–300	Anticholinergic (dry mouth, tachycardia, constipation, urinary retention, blurred vision); sweating; tremor; postural hypotension; cardiac conduction delay; sedation; weight gain	Once-daily dosing, usually qhs; blood levels of most TCAs available; can be lethal in overdose (lethal dose = 2 g); nortriptyline best tolerated, especially by elderly
Nortriptyline (Pamelor)	50–200		
Imipramine (Tofranil)	150–300		
Desipramine (Norpramin)	150–300		FDA approved for OCD

Name	Usual Daily Dose, mg	Side Effects	Comments
Doxepin (Sinequan)	150–300		
Clomipramine (Anafranil)	150–300		
Maprotiline (Ludiomil)	25–150		
Mixed Norepinephrine/Serotonin Reuptake Inhibitors (SNRI) and Receptor Blockers			
Venlafaxine (Effexor)	75–375	Nausea; dizziness; dry mouth; headaches; increased blood pressure; anxiety and insomnia	Bid-tid dosing (extended release available); lower potential for drug interactions than SSRIs; contraindicated with MAOIs
Desvenlafaxine (Pristiq)	50–400	Nausea, dizziness, insomnia	Primary metabolite of venlafaxine; no increased efficacy with higher dosing
Duloxetine (Cymbalta)	40–60	Nausea, dizziness, headache, insomnia, constipation	May have utility in treatment of neuropathic pain and stress incontinence
Mirtazapine (Remeron)	15–45	Somnolence, weight gain; neutropenia rare	Once-a-day dosing
Vilazodone (Viibryd)	40	Nausea, diarrhea, headache; dosage adjustment if given with CYP3A4 inhibitor/stimulator	Also 5-HT_{1a} receptor partial agonist
Vortioxetine (Brintellix)	5–20	Nausea, diarrhea, sweating, headache; low incidence of sedation or weight gain	No specific p450 effects; 5-HT_{3a} and 5-HT_7 receptor antagonist, 5-HT_{1b} partial agonist, and 5-HT_{1a} agonist

Name	Usual Daily Dose, mg	Side Effects	Comments
Levomilnacipran (Fetzima)	40–120	Nausea, constipation, sweating; rare increase in blood pressure/pulse	Most noradrenergic of SNRIs
Mixed-Action Drugs			
Bupropion (Wellbutrin)	250–450	Jitteriness; flushing; seizures in at-risk patients; anorexia; tachycardia; psychosis	Tid dosing, but sustained release also available; fewer sexual side effects than SSRIs or TCAs; may be useful for adult ADD
Trazodone (Desyrel)	200–600	Sedation; dry mouth; ventricular irritability; postural hypotension; priapism rare	Useful in low doses for sleep because of sedating effects with no anticholinergic side effects
Trazodone extended release (Oleptro)	150–375	Daytime somnolence, dizziness, nausea	
Amoxapine (Asendin)	200–600	Sexual dysfunction	Lethality in overdose; EPS possible
MAOIs			
Phenelzine (Nardil)	45–90	Insomnia; hypotension; edema; anorgasmia; weight gain; neuropathy; hypertensive crisis; toxic reactions with SSRIs; narcotics	May be more effective in patients with atypical features or treatment-refractory depression
Tranylcypromine (Parnate)	20–50		
Isocarboxazid (Marplan)	20–60		Less weight gain and hypotension than phenelzine

Name	Usual Daily Dose, mg	Side Effects	Comments
Transdermal selegiline (Emsam)	6–12	Local skin reaction hypertension	No dietary restrictions with 6-mg dose

Abbreviations: ADD, attention deficit disorder; EPS, extrapyramidal symptoms; FDA, U.S. Food and Drug Administration; GI, gastrointestinal; MAOIs, monoamine oxidase inhibitors; OCD, obsessive-compulsive disorder; SSRIs, selective serotonin reuptake inhibitors; TCAs, tricyclic antidepressants.

TABLE 444-2

Management of Antidepressant Side Effects

Symptoms	Comments and Management Strategies
Gastrointestinal	
Nausea, loss of appetite	Usually short-lived and dose-related; consider temporary dose reduction or administration with food and antacids
Diarrhea	Famotidine, 20–40 mg/d
Constipation	Wait for tolerance; try diet change, stool softener, exercise; avoid laxatives
Sexual dysfunction	Consider dose reduction; drug holiday
Anorgasmia/impotence; impaired ejaculation	Bethanechol, 10–20 mg, 2 h before activity, or cyproheptadine, 4–8 mg 2 h before activity, or bupropion, 100 mg bid, or amantadine, 100 mg bid/tid
Orthostasis	Tolerance unlikely; increase fluid intake, use calf exercises/support hose; fludrocortisone, 0.025 mg/d
Anticholinergic	Wait for tolerance
Dry mouth, eyes	Maintain good oral hygiene; use artificial tears, sugar-free gum
Tremor/jitteriness	Antiparkinsonian drugs not effective; use dose reduction/slow increase; lorazepam, 0.5 mg bid, or propranolol, 10–20 mg bid
Insomnia	Schedule all doses for the morning; trazodone, 50–100 mg qhs
Sedation	Caffeine; schedule all dosing for bedtime; bupropion, 75–100 mg in afternoon
Headache	Evaluate diet, stress, other drugs; try dose reduction; amitriptyline, 50 mg/d
Weight gain	Decrease carbohydrates; exercise; consider fluoxetine
Loss of therapeutic benefit over time	Related to tolerance? Increase dose or drug holiday; add amantadine, 100 mg bid, buspirone, 10 mg tid, or pindolol, 2.5 mg bid

TABLE 444-3

Possible Drug Interactions with Selective Serotonin Reuptake Inhibitors

Agent	Effect
Monoamine oxidase inhibitors	Serotonin syndrome—absolute contraindication
Serotonergic agonists, e.g., tryptophan, fenfluramine, tryptans	Potential serotonin syndrome
Drugs that are metabolized by P450 isoenzymes: tricyclics, other SSRIs, antipsychotics, beta blockers, codeine, triazolobenzodiazepines, calcium channel blockers	Delayed metabolism resulting in increased blood levels and potential toxicity
Drugs that are bound tightly to plasma proteins, e.g., warfarin	Increased bleeding secondary to displacement
Drugs that inhibit the metabolism of SSRIs by P450 isoenzymes, e.g., quinidine	Increased SSRI side effects

Abbreviation: SSRIs, selective serotonin reuptake inhibitors.

TABLE 444-4

Anxiolytics

Name	Equivalent PO Dose, mg	Onset of Action	Half-Life, h	Comments
Benzodiazepines				
Diazepam (Valium)	5	Fast	20–70	Active metabolites; quite sedating
Flurazepam (Dalmane)	15	Fast	30–100	Flurazepam is a prodrug; metabolites are active; quite sedating
Triazolam (Halcion)	0.25	Intermediate	1.5–5	No active metabolites; can induce confusion and delirium, especially in elderly
Lorazepam (Ativan)	1	Intermediate	10–20	No active metabolites; direct hepatic glucuronide conjugation; quite sedating; FDA approved for anxiety with depression
Alprazolam (Xanax)	0.5	Intermediate	12–15	Active metabolites; not too sedating; FDA approved for panic disorder and anxiety with depression; tolerance and dependence develop easily; difficult to withdraw
Chlordiazepoxide (Librium)	10	Intermediate	5–30	Active metabolites; moderately sedating
Oxazepam (Serax)	15	Slow	5–15	No active metabolites; direct glucuronide conjugation; not too sedating
Temazepam (Restoril)	15	Slow	9–12	No active metabolites; moderately sedating
Clonazepam (Klonopin)	0.5	Slow	18–50	No active metabolites; moderately sedating; FDA approved for panic disorder

Name	Equivalent PO Dose, mg	Onset of Action	Half-Life, h	Comments
Clorazepate (Tranxene)	15	Fast	40–200	Low sedation; unreliable absorption
Nonbenzodiazepines				
Buspirone (BuSpar)	7.5	2 weeks	2–3	Active metabolites; tid dosing—usual daily dose 10–20 mg tid; nonsedating; no additive effects with alcohol; useful for controlling agitation in demented or brain-injured patients

Abbreviation: FDA, U.S. Food and Drug Administration.

GENERALIZED ANXIETY DISORDER

Clinical Manifestations

Patients with generalized anxiety disorder (GAD) have persistent, excessive, and/or unrealistic worry associated with muscle tension, impaired concentration, autonomic arousal, feeling “on edge” or restless, and insomnia **(Table 444-5)**. Onset is usually before age 20 years, and a history of childhood fears and social inhibition may be present. The lifetime prevalence of GAD is 5–6%; the risk is higher in first-degree relatives of patients with the diagnosis. Interestingly, family studies indicate that GAD and panic disorder segregate independently. More than 80% of patients with GAD also suffer from major depression, dysthymia, or social phobia. Comorbid substance abuse is common in these patients, particularly alcohol and/or sedative/hypnotic abuse. Patients with GAD worry excessively over minor matters, with life-disrupting effects; unlike in panic disorder, complaints of shortness of breath, palpitations, and tachycardia are relatively rare.

TABLE 444-5

Diagnostic Criteria for Generalized Anxiety Disorder

A. Excessive anxiety and worry (apprehensive expectation), occurring more days than not for at least 6 months, about a number of events or activities (such as work or school performance).
B. The individual finds it difficult to control the worry.
C. The anxiety and worry are associated with three (or more) of the following six symptoms (with at least some symptoms present for more days than not for the past 6 months): (1) restlessness or feeling keyed up or on edge; (2) being easily fatigued; (3) difficulty concentrating or mind going blank; (4) irritability; (5) muscle tension; (6) sleep disturbance (difficulty falling or staying asleep, or restless, unsatisfying sleep).
D. The anxiety, worry, or physical symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.
E. The disturbance is not attributable to the physiologic effects of a substance (e.g., a drug of abuse, a medication) or another medical condition (e.g., hyperthyroidism).
F. The disturbance is not better explained by another mental disorder (e.g., anxiety or worry about having panic attacks in panic disorder, negative evaluation in social anxiety disorder [social phobia], contamination or other obsessions in obsessive-compulsive disorder, separation from attachment figures in separation anxiety disorder, reminders of traumatic events in posttraumatic stress disorder, gaining weight in anorexia nervosa, physical complaints in somatic symptom disorder, perceived appearance flaws in body dysmorphic disorder, having a serious illness in illness anxiety disorder, or the content of delusional beliefs in schizophrenia or delusional disorder).

Source: Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC, American Psychiatric Association, 2013.

Etiology and Pathophysiology

All anxiogenic agents act on the γ-aminobutyric acid $(GABA)_A$ receptor/chloride ion channel complex, implicating this neurotransmitter system in the pathogenesis of anxiety and panic attacks. Benzodiazepines are thought to bind two separate $GABA_A$ receptor sites: type I, which has a broad neuroanatomic distribution, and type II, which is concentrated in the hippocampus, striatum, and neocortex. The antianxiety effects of the various benzodiazepines are influenced by their relative binding to alpha 2 and 3 subunits of the $GABA_A$ receptor, and sedation and memory impairment to the alpha 1 subunit. Serotonin (5-hydroxytryptamine [5-HT]), and 3α-reduced neuroactive steroids (allosteric modulators of $GABA_A$) also appear to have a role in anxiety, and buspirone, a partial $5\text{-}HT_{1A}$ receptor agonist, and certain $5\text{-}HT_{2A}$ and $5\text{-}HT_{2C}$ receptor antagonists (e.g., nefazodone) may have beneficial effects.

TREATMENT

TREATMENT

Generalized Anxiety Disorder

A combination of pharmacologic and psychotherapeutic interventions is most effective in GAD, but complete symptomatic relief is rare. A short course of a benzodiazepine is usually indicated, preferably lorazepam, oxazepam, or alprazolam. (The first two of these agents are metabolized via conjugation rather than oxidation and thus do not accumulate if hepatic function is impaired; the latter also has limited active metabolites.) Treatment should be initiated at the lowest dose possible and prescribed on an as-needed basis as symptoms warrant. Benzodiazepines differ in their milligram per kilogram potency, half-life, lipid solubility, metabolic pathways, and presence of active metabolites. Agents that are absorbed rapidly and are lipid soluble, such as diazepam, have a rapid onset of action and a higher abuse potential. Benzodiazepines should generally not be prescribed for >4–6 weeks because of the development of tolerance and the risk of abuse and dependence. Withdrawal must be closely monitored as relapses can occur. It is important to warn patients that concomitant use of alcohol or other sedating drugs may exacerbate side effects and impair their ability to function. An optimistic approach that encourages the patient to clarify environmental precipitants, anticipate his or her reactions, and plan effective response strategies is an essential element of therapy.

Adverse effects of benzodiazepines generally parallel their relative half-lives. Longer-acting agents, such as diazepam, chlordiazepoxide, flurazepam, and clonazepam, tend to accumulate active metabolites, with resultant sedation, impairment of cognition, and poor psychomotor performance. Shorter-acting compounds, such as alprazolam, lorazepam, and oxazepam, can produce daytime anxiety, early morning insomnia, and, with discontinuation, rebound anxiety and insomnia. Although patients develop tolerance to the sedative effects of benzodiazepines, they are less likely to habituate to the adverse psychomotor effects. Withdrawal from the longer half-life benzodiazepines can be accomplished through gradual, stepwise dose reduction (by 10% every 1–2 weeks) over 6–12 weeks. It is usually more difficult to taper patients off shorter-acting benzodiazepines. Physicians may need to switch the patient to a benzodiazepine with a longer half-life or use an adjunctive medication such as a beta blocker or carbamazepine, before attempting to discontinue the benzodiazepine. Withdrawal reactions vary in severity and duration; they can include depression, anxiety, lethargy, diaphoresis, autonomic arousal, and, rarely, seizures.

Buspirone is a nonbenzodiazepine anxiolytic agent. It is nonsedating, does not produce tolerance or dependence, does not interact with benzodiazepine receptors or alcohol, and has no abuse or disinhibition potential. However, it requires several weeks to take effect and requires thrice-daily dosing. Patients who were previously responsive to a benzodiazepine are unlikely to rate buspirone as equally effective, but patients with head injury or dementia who have symptoms of anxiety and/or agitation may do well with this agent. Escitalopram, paroxetine, duloxetine, and venlafaxine are FDA approved for the treatment of GAD, usually at doses that are comparable to their efficacy in major depression, and may be preferable to usage of benzodiazepines in the treatment of chronic anxiety. Benzodiazepines are contraindicated during pregnancy and breast-feeding.

Anticonvulsants with GABAergic properties may also be effective against anxiety. Gabapentin, oxcarbazepine, tiagabine, pregabalin, and divalproex have all shown some degree of benefit in a variety of anxiety-related syndromes in off-label usage.

PHOBIC DISORDERS

Clinical Manifestations

The cardinal feature of phobic disorders is a marked and persistent fear of objects or situations, exposure to which results in an immediate anxiety reaction. The patient avoids the phobic stimulus, and this avoidance usually impairs occupational or social functioning. Panic attacks may be triggered by the phobic stimulus or may occur spontaneously. Unlike patients with other anxiety disorders, individuals with phobias usually experience anxiety only in specific situations. Common phobias include fear of closed spaces (claustrophobia), fear of blood, and fear of flying. Social phobia is distinguished by a specific fear of social or performance situations in which the individual is exposed to unfamiliar individuals or to possible examination and evaluation by others. Examples include having to converse at a party, use public restrooms, and meet strangers. In each case, the affected individual is aware that the experienced fear is excessive and unreasonable given the circumstance. The specific content of a phobia may vary across gender, ethnic, and cultural boundaries.

Phobic disorders are common, affecting ~7–9% of the population. Twice as many females are affected than males. Full criteria for diagnosis are usually satisfied first in early adulthood, but behavioral avoidance of unfamiliar people, situations, or objects dating from early childhood is common.

In one study of female twins, concordance rates for agoraphobia, social phobia, and animal phobia were found to be 23% for monozygotic twins and 15% for dizygotic twins. A twin study of fear conditioning, a model for the acquisition of phobias, demonstrated a heritability of 35–45%. Animal studies of fear conditioning have indicated that processing of the fear stimulus occurs through the lateral nucleus of the amygdala, extending through the central nucleus and projecting to the periaqueductal gray region, lateral hypothalamus, and paraventricular hypothalamus.

TREATMENT

TREATMENT

Phobic Disorders

Beta blockers (e.g., propranolol, 20–40 mg orally 2 h before the event) are particularly effective in the treatment of “performance anxiety” (but not general social phobia) and appear to work by blocking the peripheral manifestations of anxiety such as perspiration, tachycardia, palpitations, and tremor. MAOIs alleviate social phobia independently of their antidepressant activity, and paroxetine, sertraline, and venlafaxine have received FDA approval for treatment of social anxiety. Benzodiazepines can be helpful in reducing fearful avoidance, but the chronic nature of phobic disorders limits their usefulness.

Behaviorally focused psychotherapy is an important component of treatment because relapse rates are high when medication is used as the sole treatment. Cognitive-behavioral strategies are based on the finding that distorted perceptions and interpretations of fear-producing stimuli play a major role in perpetuation of phobias. Individual and group therapy sessions teach the patient to identify specific negative thoughts associated with the anxiety-producing situation and help to reduce the patient's fear of loss of control. In desensitization therapy, hierarchies of feared situations are constructed, and the patient is encouraged to pursue and master gradual exposure to the anxiety-producing stimuli.

Patients with social phobia, in particular, have a high rate of comorbid alcohol abuse, as well as of other psychiatric conditions (e.g., eating disorders), necessitating the need for parallel management of each disorder if anxiety reduction is to be achieved.

STRESS DISORDERS

Clinical Manifestations

Patients may develop anxiety after exposure to extreme traumatic events such as the threat of personal death or injury or the death of a loved one. The reaction may occur shortly after the trauma (*acute stress disorder*) or be delayed and subject to recurrence (PTSD) **(Table 444-6)**. In both syndromes, individuals experience associated symptoms of detachment and loss of emotional responsivity. The patient may feel depersonalized and unable to recall specific aspects of the trauma, although typically it is reexperienced through intrusions in thought, dreams, or flashbacks, particularly when cues of the original event are present. Patients often actively avoid stimuli that precipitate recollections of the trauma and demonstrate a resulting increase in vigilance, arousal, and startle response. Patients with stress disorders are at risk for the development of other disorders related to anxiety, mood, and substance abuse (especially alcohol). Between 5 and 10% of Americans will at some time in their life satisfy criteria for PTSD, with women more likely to be affected than men. A validated 4-item screen for PTSD (PC-PTSD) is available.

TABLE 444-6

Diagnostic Criteria for Posttraumatic Stress Disorder

A. Exposure to actual or threatened death, serious injury, or sexual violence in one (or more) of the following ways:
   1. Directly experiencing the traumatic event(s).
   2. Witnessing, in person, the event(s) as it occurred to others.
   3. Learning that the traumatic event(s) occurred to a close family member or close friend. In cases of actual or threatened death of a family member or friend, the event(s) must have been violent or accidental.
   4. Experiencing repeated or extreme exposure to aversive details of the traumatic event(s) (e.g., first responders collecting human remains; police officers repeatedly exposed to details of child abuse).

B. Presence of one (or more) of the following intrusion symptoms associated with the traumatic event(s), beginning after the traumatic event(s) occurred:
   1. Recurrent, involuntary, and intrusive distressing memories of the traumatic event(s).
   2. Recurrent distressing dreams in which the content and/or affect of the dream are related to the traumatic event(s).
   3. Dissociative reactions (e.g., flashbacks) in which the individual feels or acts as if the traumatic event(s) were recurring. (Such reactions may occur on a continuum, with the most extreme expression being a complete loss of awareness of present surroundings.)
   4. Intense or prolonged psychological distress at exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event(s).
   5. Marked physiologic reactions to internal or external cues that symbolize or resemble an aspect of the traumatic event(s).

C. Persistent avoidance of stimuli associated with the traumatic event(s), beginning after the traumatic event(s) occurred, as evidenced by one or both of the following:
   1. Avoidance of or efforts to avoid distressing memories, thoughts, or feelings about or closely associated with the traumatic event(s).
   2. Avoidance of or efforts to avoid external reminders (people, places, conversations, activities, objects, situations) that arouse distressing memories, thoughts, or feelings about or closely associated with the traumatic event(s).

D. Negative alterations in cognitions and mood associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred as evidenced by two (or more) of the following:
   1. Inability to remember an important aspect of the traumatic event(s) (typically due to dissociative amnesia and not to other factors such as head injury, alcohol, or drugs).
   2. Persistent and exaggerated negative beliefs or expectations about oneself, others, or the world (e.g., “I am bad,” “No one can be trusted,” “The world is completely dangerous,” “My whole nervous system is permanently ruined”).
   3. Persistent, distorted cognitions about the cause or consequences of the traumatic event(s) that lead the individual to blame himself/herself or others.

4. Persistent negative emotional state (e.g., fear, horror, anger, guilt, or shame).
5. Markedly diminished interest or participation in significant activities.
6. Feelings of detachment or estrangement from others.
7. Persistent inability to experience positive emotions (e.g., inability to experience happiness, satisfaction, or loving feelings).

E. Marked alterations in arousal and reactivity associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred, as evidenced by two (or more) of the following:
1. Irritable behavior and angry outbursts (with little or no provocation) typically expressed as verbal or physical aggression toward people or objects.
2. Reckless or self-destructive behavior.
3. Hypervigilance.
4. Exaggerated startle response.
5. Problems with concentration.
6. Sleep disturbance (e.g., difficulty falling or staying asleep or restless sleep).

F. Duration of the disturbance (criteria B, C, D, and E) is >1 month.

G. The disturbance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.

H. The disturbance is not attributable to the physiologic effects of a substance (e.g., medication, alcohol) or another medical condition.

Source: Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC, American Psychiatric Association, 2013.

Risk factors for the development of PTSD include a past psychiatric history and personality characteristics of high neuroticism and extroversion. Twin studies show a substantial genetic influence on all symptoms associated with PTSD, with less evidence for an environmental effect.

Etiology and Pathophysiology

It is hypothesized that in PTSD there is excessive release of norepinephrine from the locus coeruleus in response to stress and increased noradrenergic activity at projection sites in the hippocampus and amygdala. These changes theoretically facilitate the encoding of fear-based memories. Greater sympathetic responses to cues associated with the traumatic event occur in PTSD, although pituitary adrenal responses are blunted. In addition to fear learning, changes in threat detection (insula overactivity), executive function, emotional regulation and contextual learning have been documented.

TREATMENT

TREATMENT

Stress Disorders

Acute stress reactions are usually self-limited, and treatment typically involves the short-term use of benzodiazepines and supportive/expressive psychotherapy. The chronic and recurrent nature of PTSD, however, requires a more complex approach using drug and behavioral treatments. PTSD is highly correlated with peritraumatic dissociative symptoms and the development of an acute stress disorder at the time of the trauma. The SSRIs (paroxetine and sertraline are FDA approved for PTSD), venlafaxine, nefazadone, and topiramate can all reduce anxiety, symptoms of intrusion, and avoidance behaviors. Hydrocortisone, intranasal oxytocin, and opiates such as morphine, given shortly after the acute stress, may have beneficial effects in preventing the development of PTSD, and adjunctive naltrexone can be effective when comorbid alcoholism is present. Low dose trazodone and mirtazepine, sedating antidepressants, are frequently used at night to help with insomnia. Benzodiazepines and SSRIs, however, should not be given in the early aftermath of trauma. Psychotherapeutic strategies for PTSD help the patient overcome avoidance behaviors and demoralization and master fear of recurrence of the trauma; therapies that encourage the patient to dismantle avoidance behaviors through stepwise focusing on the experience of the traumatic event, such as trauma-focused cognitive-behavioral therapy, exposure therapy, and eye movement desensitization and reprocessing, are the most effective. Debriefing after the traumatic event does not prevent PTSD and may exacerbate symptoms.

OBSESSIVE-COMPULSIVE DISORDER

Clinical Manifestations

Obsessive-compulsive disorder (OCD) is characterized by obsessive thoughts and compulsive behaviors that impair everyday functioning. Fears of contamination and germs are common, as are handwashing, counting behaviors, and having to check and recheck such actions as whether a door is locked. The degree to which the disorder is disruptive for the individual varies, but in all cases, obsessive-compulsive activities take up >1 h per day and are undertaken to relieve the anxiety triggered by the core fear. Patients often conceal their symptoms, usually because they are embarrassed by the content of their thoughts or the nature of their actions. Physicians must ask specific questions regarding recurrent thoughts and behaviors, particularly if physical clues such as chafed and reddened hands or patchy hair loss (from repetitive hair pulling, or trichotillomania) are present. Comorbid conditions are common, the most frequent being depression, other anxiety disorders, eating disorders, and tics. OCD has a lifetime prevalence of 2–3% worldwide. Onset is usually gradual, beginning in early adulthood, but childhood onset is not rare. The disorder usually has a waxing and waning course, but some cases may show a steady deterioration in psychosocial functioning.

Etiology and Pathophysiology

A genetic contribution to OCD is suggested by twin studies, but no susceptibility gene for OCD has been identified to date. Family studies show an aggregation of OCD with Tourette's disorder, and both are more common in males and in first-born children.

The anatomy of obsessive-compulsive behavior is thought to include the orbital frontal cortex, caudate nucleus, and globus pallidus. The caudate nucleus appears to be involved in the acquisition and

maintenance of habit and skill learning, and interventions that are successful in reducing obsessive-compulsive behaviors also decrease metabolic activity measured in the caudate.

TREATMENT

TREATMENT

Obsessive-Compulsive Disorder

Clomipramine, fluoxetine, fluvoxamine, and sertraline are approved for the treatment of OCD in adults (fluvoxamine is also approved for children). Clomipramine is a TCA that is often tolerated poorly owing to anticholinergic and sedative side effects at the doses required to treat the illness (25–250 mg/d); its efficacy in OCD is unrelated to its antidepressant activity. Fluoxetine (5–60 mg/d), fluvoxamine (25–300 mg/d), and sertraline (50–150 mg/d) are as effective as clomipramine and have a more benign side effect profile. Only 50–60% of patients with OCD show adequate improvement with pharmacotherapy alone. In treatment-resistant cases, augmentation with other serotonergic agents such as buspirone, or with a neuroleptic or benzodiazepine, may be beneficial, and in severe cases, deep brain stimulation has been found to be effective. When a therapeutic response is achieved, long-duration maintenance therapy is usually indicated.

For many individuals, particularly those with time-consuming compulsions, behavior therapy, and exposure response prevention will result in as much improvement as that afforded by medication. Effective techniques include the gradual increase in exposure to stressful situations, maintenance of a diary to clarify stressors, and homework assignments that substitute new activities for compulsive behaviors.

MOOD DISORDERS

Mood disorders are characterized by a disturbance in the regulation of mood, behavior, and affect. Mood disorders are subdivided into (1) depressive disorders, (2) bipolar disorders, and (3) depression in association with medical illness or alcohol and substance abuse **(Chaps. 445, Chaps. 446, Chaps. 447, 448)**. Major depressive disorder (MDD) is differentiated from bipolar disorder by the absence of a manic or hypomanic episode. The relationship between pure depressive syndromes and bipolar disorders is not well understood; MDD is more frequent in families of bipolar individuals, but the reverse is not true. In the Global Burden of Disease Study conducted by the World Health Organization, unipolar major depression ranked fourth among all diseases in terms of disability-adjusted life-years and was projected to rank second by the year 2020. In the United States, lost productivity directly related to mood disorders has been estimated at $55.1 billion per year.

DEPRESSION IN ASSOCIATION WITH MEDICAL ILLNESS

Depression occurring in the context of medical illness is difficult to evaluate. Depressive symptomatology may reflect the psychological stress of coping with the disease, may be caused by the disease process itself or by the medications used to treat it, or may simply coexist in time with the medical diagnosis.

Virtually every class of *medication* includes some agent that can induce depression. Antihypertensive drugs, anticholesterolemic agents, and antiarrhythmic agents are common triggers of depressive symptoms. Iatrogenic depression should also be considered in patients receiving glucocorticoids, antimicrobials, systemic analgesics, antiparkinsonian medications, and anticonvulsants. To decide whether a causal relationship exists between pharmacologic therapy and a patient's change in mood, it may sometimes be necessary to undertake an empirical trial of an alternative medication.

Between 20 and 30% of *cardiac* patients manifest a depressive disorder; an even higher percentage experience depressive symptomatology when self-reporting scales are used. Depressive symptoms following unstable angina, myocardial infarction, cardiac bypass surgery, or heart transplant impair rehabilitation and are associated with higher rates of mortality and medical morbidity. Depressed patients often show decreased variability in heart rate (an index of reduced parasympathetic nervous system activity), which may predispose individuals to ventricular arrhythmia and increased morbidity. Depression also appears to increase the risk of coronary heart disease, possibly through increased platelet aggregation. TCAs are contraindicated in patients with bundle branch block, and TCA-induced tachycardia is an additional concern in patients with congestive heart failure. SSRIs appear not to induce ECG changes or adverse cardiac events and thus are reasonable first-line drugs for patients at risk for TCA-related complications. SSRIs may interfere with hepatic metabolism of anticoagulants, however, causing increased anticoagulation.

In patients with *cancer*, the mean prevalence of depression is 25%, but depression occurs in 40–50% of patients with cancers of the pancreas or oropharynx. This association is not due to the effect of cachexia alone, as the higher prevalence of depression in patients with pancreatic cancer persists when compared to those with advanced gastric cancer. Initiation of antidepressant medication in cancer patients has been shown to improve quality of life as well as mood. Psychotherapeutic approaches, particularly group therapy, may have some effect on short-term depression, anxiety, and pain symptoms.

Depression occurs frequently in patients with *neurologic disorders*, particularly cerebrovascular disorders, Parkinson's disease, dementia, multiple sclerosis, and traumatic brain injury. One in five patients with left-hemisphere stroke involving the dorsolateral frontal cortex experiences major depression. Late-onset depression in otherwise cognitively normal individuals increases the risk of a subsequent diagnosis of Alzheimer's disease. All classes of antidepressant agents are effective against these depressions, as are, in some cases, stimulant compounds.

The reported prevalence of depression in patients with *diabetes mellitus* varies from 8 to 27%, with the severity of the mood state correlating with the level of hyperglycemia and the presence of diabetic complications. Treatment of depression may be complicated by effects of antidepressive agents on glycemic control. MAOIs can induce hypoglycemia and weight gain, whereas TCAs can produce hyperglycemia and carbohydrate craving. SSRIs and SNRIs, like MAOIs, may reduce fasting plasma glucose, but they are easier to use and may also improve dietary and medication compliance.

Hypothyroidism is frequently associated with features of depression, most commonly depressed mood and memory impairment. Hyperthyroid states may also present in a similar fashion, usually in geriatric

populations. Improvement in mood usually follows normalization of thyroid function, but adjunctive antidepressant medication is sometimes required. Patients with subclinical hypothyroidism can also experience symptoms of depression and cognitive difficulty that respond to thyroid replacement.

The lifetime prevalence of depression in *HIV-positive* individuals has been estimated at 22–45%. The relationship between depression and disease progression is multifactorial and likely to involve psychological and social factors, alterations in immune function, and central nervous system (CNS) disease. Chronic hepatitis C infection is also associated with depression, which may worsen with interferon-α treatment.

Some chronic disorders of uncertain etiology, such as chronic fatigue syndrome **(Chap. 442)** and fibromyalgia **(Chap. 366)**, are strongly associated with depression and anxiety; patients may benefit from antidepressant treatment or anticonvulsant agents such as pregabalin.

DEPRESSIVE DISORDERS

Clinical Manifestations

Major depression is defined as depressed mood on a daily basis for a minimum duration of 2 weeks **(Table 444-7)**. An episode may be characterized by sadness, indifference, apathy, or irritability and is usually associated with changes in sleep patterns, appetite, and weight; motor agitation or retardation; fatigue; impaired concentration and decision-making; feelings of shame or guilt; and thoughts of death or dying. Patients with depression have a profound loss of pleasure in all enjoyable activities, exhibit early morning awakening, feel that the dysphoric mood state is qualitatively different from sadness, and often notice a diurnal variation in mood (worse in morning hours). Patients experiencing bereavement or grief may exhibit many of the same signs and symptoms of major depression, although the emphasis is usually on feelings of emptiness and loss, rather than anhedonia and loss of self-esteem, and the duration is usually limited. In certain cases, however, the diagnosis of major depression may be warranted even in the context of a significant loss.

TABLE 444-7

Criteria for a Major Depressive Episode

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure. (*Note:* Do not include symptoms that are clearly attributable to another medical condition.)
   1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad, empty, hopeless) or observation made by others (e.g., appears tearful).
   2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).
   3. Significant weight loss when not dieting or weight gain (e.g., a change of >5% of body weight in a month), or decrease or increase in appetite nearly every day.
   4. Insomnia or hypersomnia nearly every day.
   5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
   6. Fatigue or loss of energy nearly every day.
   7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
   8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
   9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

B. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning

C. The episode is not attributable to the physiologic effects of a substance or to another medical condition.

D. The occurrence of the major depressive episode is not better explained by seasonal affective disorder, schizophrenia, schizophreniform disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and other psychotic disorders.

E. There has never been a manic episode or a hypomanic episode.

Source: Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC, American Psychiatric Association, 2013.

Approximately 15% of the population experiences a major depressive episode at some point in life, and 6–8% of all outpatients in primary care settings satisfy diagnostic criteria for the disorder. Depression is often undiagnosed, and even more frequently, it is treated inadequately. If a physician suspects the presence of a major depressive episode, the initial task is to determine whether it represents unipolar or bipolar depression or is one of the 10–15% of cases that are secondary to general medical illness or substance abuse.

Physicians should also assess the risk of suicide by direct questioning, as patients are often reluctant to verbalize such thoughts without prompting. If specific plans are uncovered or if significant risk factors exist (e.g., a past history of suicide attempts, profound hopelessness, concurrent medical illness, substance abuse, or social isolation), the patient must be referred to a mental health specialist for immediate care. The physician should specifically probe each of these areas in an empathic and hopeful manner, being sensitive to denial and possible minimization of distress. The presence of anxiety, panic, or agitation significantly increases near-term suicidal risk. Approximately 4–5% of all depressed patients will commit suicide; most will have sought help from physicians within 1 month of their deaths.

In some depressed patients, the mood disorder does not appear to be episodic and is not clearly associated with either psychosocial dysfunction or change from the individual's usual experience in life. Persistent depressive disorder (*dysthymic disorder*) consists of a pattern of chronic (at least 2 years), ongoing depressive symptoms that are usually less severe and/or less numerous than those found in major depression, but the functional consequences may be equivalent to or even greater; the two conditions are sometimes difficult to separate and can occur together ("double depression"). Many patients who exhibit a profile of pessimism, disinterest, and low self-esteem respond to antidepressant treatment. Persistent and chronic depressive disorders occur in ~2% of the general population.

Depression is approximately twice as common in women as in men, and the incidence increases with age in both sexes. Twin studies indicate that the liability to major depression of early onset (before age 25) is largely genetic in origin. Negative life events can precipitate and contribute to depression, but genetic factors influence the sensitivity of individuals to these stressful events. In most cases, both biologic and psychosocial factors are involved in the precipitation and unfolding of depressive episodes. The most potent stressors appear to involve death of a relative, assault, or severe marital or relationship problems.

Unipolar depressive disorders usually begin in early adulthood and recur episodically over the course of a lifetime. The best predictor of future risk is the number of past episodes; 50–60% of patients who have a first episode have at least one or two recurrences. Some patients experience multiple episodes that become more severe and frequent over time. The duration of an untreated episode varies greatly, ranging from a few months to ≥1 year. The pattern of recurrence and clinical progression in a developing episode are also variable. Within an individual, the nature of episodes (e.g., specific presenting symptoms, frequency, and duration) may be similar over time. In a minority of patients, a severe depressive episode may progress to a psychotic state; in elderly patients, depressive symptoms may be associated with cognitive deficits mimicking dementia ("pseudodementia"). A seasonal pattern of depression, called *seasonal affective disorder*, may manifest with onset and remission of episodes at predictable times of the year. This disorder is more common in women, whose symptoms are anergy, fatigue, weight gain, hypersomnia, and episodic carbohydrate craving. The prevalence increases with distance from the equator, and improvement may occur by altering light exposure.

Etiology and Pathophysiology

Although evidence for genetic transmission of unipolar depression is not as strong as in bipolar disorder, monozygotic twins have a higher concordance rate (46%) than dizygotic siblings (20%), with little support for any effect of a shared family environment. A recent association study of over 5000 Chinese women identified genome wide significant loci near the gene for sirtuin 1(*SIRT1*) and in an intron of the phospholysine phosphohistidine inorganic pyrophosphate phosphatase (*LHPP*) gene.

Neuroendocrine abnormalities that reflect the neurovegetative signs and symptoms of depression include increased cortisol and corticotropin-releasing hormone (CRH) secretion, a decreased inhibitory response of glucocorticoids to dexamethasone, and a blunted response of thyroid-stimulating hormone (TSH) level to infusion of thyroid-releasing hormone (TRH). Antidepressant treatment leads to normalization of these abnormalities. Major depression is also associated with changes in levels of proinflammatory cytokines and neurotrophins, an increase in measures of oxidative stress and cellular aging, telomere shortening, and mitochondrial dysfunction.

Diurnal variations in symptom severity and alterations in circadian rhythmicity of a number of neurochemical and neurohumoral factors suggest that biologic differences may be secondary to a primary defect in regulation of biologic rhythms. Patients with major depression show consistent findings of a decrease in rapid eye movement (REM) sleep onset (REM latency), an increase in REM density, and, in some subjects, a decrease in stage IV delta slow-wave sleep.

Although antidepressant drugs inhibit neurotransmitter uptake within hours, their therapeutic effects typically emerge over several weeks, implicating adaptive changes in second messenger systems and transcription factors as possible mechanisms of action.

TREATMENT

TREATMENT

Depressive Disorders

Treatment planning requires coordination of short-term strategies to induce remission combined with longer term maintenance designed to prevent recurrence. The most effective intervention for achieving remission and preventing relapse is medication, but combined treatment, incorporating psychotherapy to help the patient cope with decreased self-esteem and demoralization, improves outcome **(Fig. 444-1)**. Approximately 40% of primary care patients with depression drop out of treatment and discontinue medication if symptomatic improvement is not noted within a month, unless additional support is provided. Outcome improves with (1) increased intensity and frequency of visits during the first 4–6 weeks of treatment, (2) supplemental educational materials, and (3) psychiatric consultation as indicated. Despite the widespread use of SSRIs and other second-generation antidepressant drugs, there is no convincing evidence that these classes of antidepressants are more efficacious than TCAs. Between 60 and 70% of all depressed patients respond to any drug chosen, if it is given in a sufficient dose for 6–8 weeks.

A rational approach to selecting which antidepressant to use involves matching the patient's preference and medical history with the metabolic and side effect profile of the drug (Tables 444-4 and 444-5). A previous response, or a family history of a positive response, to a specific antidepressant often suggests that that drug be tried first. Before initiating antidepressant therapy, the physician should evaluate the possible contribution of comorbid illnesses and consider their specific treatment. In individuals with suicidal ideation, particular attention should be paid to choosing a drug with low toxicity if taken in overdose. Newer antidepressant drugs are distinctly safer in this regard; nevertheless, the advantages of TCAs have not been completely superseded. The existence of generic equivalents makes TCAs relatively cheap, and for secondary tricyclics, particularly nortriptyline and desipramine, well-defined relationships among dose, plasma level, and therapeutic response exist. The steady-state plasma level achieved for a given drug dose can vary more than tenfold between individuals, and plasma levels may help in interpreting apparent resistance to treatment and/or unexpected drug toxicity. The principal side effects of TCAs are antihistamine (sedation) and anticholinergic (constipation, dry mouth, urinary hesitancy, blurred vision). TCAs are contraindicated in patients with serious cardiovascular risk factors, and overdoses of tricyclic agents can be lethal, with desipramine carrying the greatest risk. It is judicious to prescribe only a 10-day supply when suicide is a risk. Most patients require a daily dose of 150–200 mg of imipramine or amitriptyline or its equivalent to achieve a therapeutic blood level of 150–300 ng/mL and a satisfactory remission; some patients show a partial effect at lower doses. Geriatric patients may require a low starting dose and slow escalation. Ethnic differences in drug metabolism are significant, with Hispanic, Asian, and African-American patients generally requiring lower doses to achieve a comparable blood level.

Second-generation antidepressants are similar to tricyclics in their effect on neurotransmitter reuptake, although some also have specific actions on catecholamine and indolamine receptors as well. Amoxapine is a dibenzoxazepine derivative that blocks norepinephrine and serotonin reuptake and has a metabolite that shows a degree of dopamine blockade. Long-term use of this drug carries a risk of tardive dyskinesia. Maprotiline is a potent noradrenergic reuptake blocker that has little anticholinergic effect but may produce seizures. Bupropion is a novel antidepressant whose mechanism of action is thought to involve enhancement of noradrenergic function. It has no anticholinergic, sedating, or orthostatic side effects and has a low incidence of sexual side effects. It may, however, be associated with stimulant-like side effects, may lower seizure threshold, and has an exceptionally short half-life, requiring frequent dosing. An extended-release preparation is available.

SSRIs such as fluoxetine, sertraline, paroxetine, citalopram, and escitalopram cause a lower frequency of anticholinergic, sedating, and cardiovascular side effects but a possibly greater incidence of gastrointestinal complaints, sleep impairment, and sexual dysfunction than do TCAs. Akathisia, involving an inner sense of restlessness and anxiety in addition to increased motor activity, may also be more common, particularly during the first week of treatment. One concern is the risk of "serotonin syndrome," which is thought to result from hyperstimulation of brainstem 5-HT_{1A} receptors and characterized by myoclonus, agitation, abdominal cramping, hyperpyrexia, hypertension, and potentially death. Serotonergic agonists taken in combination should be monitored closely for this reason. Considerations such as half-life, compliance, toxicity, and drug-drug interactions may guide the choice of a particular SSRI. Fluoxetine and its principal active metabolite,

norfluoxetine, for example, have a combined half-life of almost 7 days, resulting in a delay of 5 weeks before steady-state levels are achieved and a similar delay for complete drug excretion once its use is discontinued; paroxetine appears to incur a greater risk of withdrawal symptoms with abrupt discontinuation. All the SSRIs may impair sexual function, resulting in diminished libido, impotence, or difficulty in achieving orgasm. Sexual dysfunction frequently results in noncompliance and should be asked about specifically. Sexual dysfunction can sometimes be ameliorated by lowering the dose, by instituting weekend drug holidays (two or three times a month), or by treatment with amantadine (100 mg tid), bethanechol (25 mg tid), buspirone (10 mg tid), or bupropion (100–150 mg/d). Paroxetine appears to be more anticholinergic than either fluoxetine or sertraline, and sertraline carries a lower risk of producing an adverse drug interaction than the other two. Rare side effects of SSRIs include angina due to vasospasm and prolongation of the prothrombin time. Escitalopram is the most specific of currently available SSRIs and appears to have no specific inhibitory effects on the P450 system.

Venlafaxine, desvenlafaxine, duloxetine, vilazodone, vortioxetine, and levomilnacipran block the reuptake of both norepinephrine and serotonin but produce relatively little in the way of traditional tricyclic side effects. Unlike the SSRIs, venlafaxine and vortioxetine have relatively linear dose-response curves. Patients on immediate release venlafaxine should be monitored for a possible increase in diastolic blood pressure, and multiple daily dosing is required because of the drug's short half-life. An extended-release form is available and has a somewhat lower incidence of gastrointestinal side effects. Mirtazapine is a TCA that has a unique spectrum of activity. It increases noradrenergic and serotonergic neurotransmission through a blockade of central α_2-adrenergic receptors and postsynaptic 5-HT_2 and 5-HT_3 receptors. It is also strongly antihistaminic and, as such, may produce sedation. Levomilnacipran is the most noradrenergic of the SNRIs and theoretically may be appropriate for patients with more severe fatigue and anergia.

With the exception of citalopram and escitalopram, each of the SSRIs may inhibit one or more cytochrome P450 enzymes. Depending on the specific isoenzyme involved, the metabolism of a number of concomitantly administered medications can be dramatically affected. Fluoxetine and paroxetine, for example, by inhibiting 2D6, can cause dramatic increases in the blood level of type 1C antiarrhythmics, whereas sertraline, by acting on 3A4, may alter blood levels of carbamazepine or digoxin. Depending on drug specificity for a particular CYP enzyme for its own metabolism, concomitant medications or dietary factors, such as grapefruit juice, may in turn affect the efficacy or toxicity of the SSRI.

The MAOIs are highly effective, particularly in atypical depression, but the risk of hypertensive crisis following intake of tyramine-containing food or sympathomimetic drugs makes them inappropriate as first-line agents. Transdermal selegiline may avert this risk at low dose. Common side effects include orthostatic hypotension, weight gain, insomnia, and sexual dysfunction. MAOIs should not be used concomitantly with SSRIs, because of the risk of serotonin syndrome, or with TCAs, because of possible hyperadrenergic effects.

Electroconvulsive therapy is at least as effective as medication, but its use is reserved for treatment-resistant cases and delusional depressions. Repetitive transcranial magnetic stimulation (rTMS) is approved for treatment-resistant depression and has been shown to have efficacy in several controlled trials. Vagus nerve

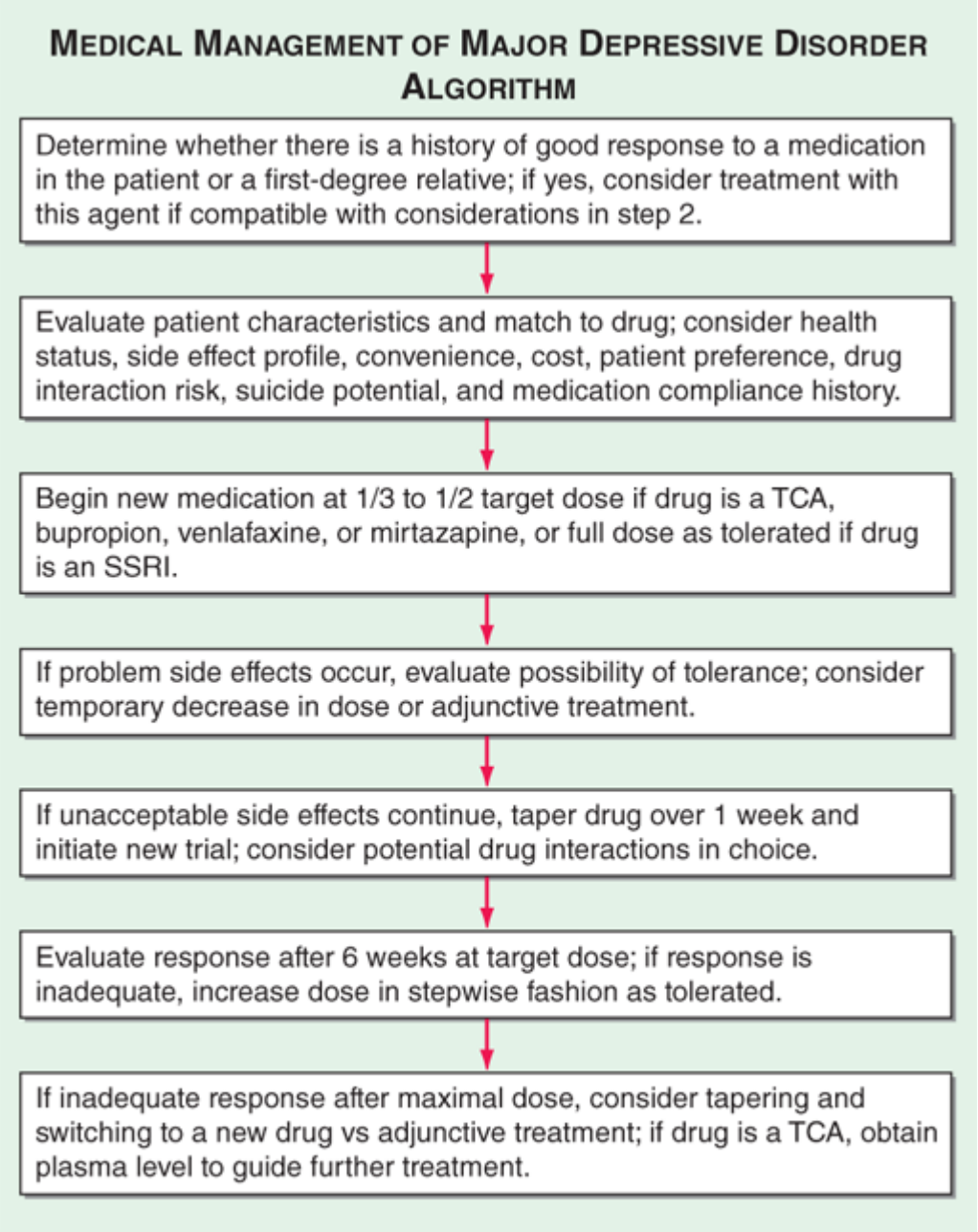
stimulation (VNS) has also recently been approved for treatment-resistant depression, but its degree of efficacy is controversial. Some metanalyses of low intensity transcranial current stimulation (tCS) have shown a positive benefit over sham treatment, but whether this is comparable to or synergistic with antidepressant treatment is unclear. In off-label usage, intravenous ketamine or esketamine and intranasal esketamine have been shown to have short-term antidepressant efficacy, often after a single administration, suggesting a possible utility in addressing suicidality. Questions remain, however, about the risk/benefit ratio over the longer term. Lastly, deep brain stimulation of the ventral anterior limb of the internal capsule and of the subcallosal cingulate region have demonstrable efficacy in randomized experimental trials of treatment-resistant depression.

Regardless of the treatment undertaken, the response should be evaluated after ~2 months. Three-quarters of patients show improvement by this time, but if remission is inadequate, the patient should be questioned about compliance, and an increase in medication dose should be considered if side effects are not troublesome. If this approach is unsuccessful, referral to a mental health specialist is advised. Strategies for treatment then include selection of an alternative drug, combinations of antidepressants, and/or adjunctive treatment with other classes of drugs, including lithium, thyroid hormone, l-methylfolate or s-adenylmethionine, atypical antipsychotic agents, and dopamine agonists. In switching to a different monotherapy, other drugs from the same class appear to be as likely to be efficacious as choosing a drug from a different class. A large randomized trial (STAR-D) was unable to show preferential efficacy, but the addition of certain atypical antipsychotic drugs (quetiapine extended-release; aripiprazole; brexpiprazole) has received FDA approval, as has usage of a combined medication, olanzapine and fluoxetine (Symbyax). Patients whose response to an SSRI wanes over time may benefit from the addition of buspirone (10 mg tid) or pindolol (2–5 mg tid) or small amounts of a TCA such as nortriptyline (25 mg bid or tid). Most patients will show some degree of response, but aggressive treatment should be pursued until remission is achieved, and drug treatment should be continued for at least 6–9 months to prevent relapse. In patients who have had two or more episodes of depression, indefinite maintenance treatment should be considered. Pharmacogenomic testing focusing on cytochrome p450 allelic variation may sometimes be helpful in identifying individuals who are poor or rapid metabolizers, but assessing pharmacodynamic gene variants has not been shown to be cost-effective or affect clinical outcomes.

It is essential to educate patients both about depression and the benefits and side effects of medications they are receiving. Advice about stress reduction and cautions that alcohol may exacerbate depressive symptoms and impair drug response are helpful. Patients should be given time to describe their experience, their outlook, and the impact of the depression on them and their families. Occasional empathic silence may be as helpful for the treatment alliance as verbal reassurance. Controlled trials have shown that cognitive-behavioral and interpersonal therapies are effective in improving psychological and social adjustment and that a combined treatment approach is more successful than medication alone for many patients.

FIGURE 444-1

A guideline for the medical management of major depressive disorder. SSRI, selective serotonin reuptake inhibitor; TCA, tricyclic antidepressant.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

BIPOLAR DISORDER

Clinical Manifestations

Bipolar disorder is characterized by unpredictable swings in mood from mania (or hypomania) to depression. Some patients suffer only from recurrent attacks of *mania*, which in its pure form is associated with increased psychomotor activity; excessive social extroversion; decreased need for sleep; impulsivity and impaired judgment; and expansive, grandiose, and sometimes irritable mood (Table 444-8). In severe mania, patients may experience delusions and paranoid thinking indistinguishable from schizophrenia. One-half of patients with bipolar disorder present with a mixture of psychomotor agitation and activation with dysphoria, anxiety, and irritability. It may be difficult to distinguish *mixed mania* from *agitated depression*. In some

bipolar patients (*bipolar II disorder*), the full criteria for mania are lacking, and the requisite recurrent depressions are separated by periods of mild activation and increased energy (hypomania). In *cyclothymic disorder*, there are numerous hypomanic periods, usually of relatively short duration, alternating with clusters of depressive symptoms that fail, either in severity or duration, to meet the criteria of major depression. The mood fluctuations are chronic and should be present for at least 2 years before the diagnosis is made.

TABLE 444-8

Criteria for a Manic Episode

A. A distinct period of abnormally and persistently elevated, expansive, or irritable mood and abnormally and persistently increased goal-directed activity or energy, lasting at least 1 week and present most of the day, nearly every day (or any duration if hospitalization is necessary).

B. During the period of the mood disturbance and increased energy or activity, three (or more) of the following symptoms (four if the mood is only irritable) are present to a significant degree and represent a noticeable change from usual behavior:
   1. Inflated self-esteem or grandiosity.
   2. Decreased need for sleep (e.g., feels rested after only 3 h of sleep).
   3. More talkative than usual or pressure to keep talking.
   4. Flight of ideas or subjective experience that thoughts are racing.
   5. Distractibility (i.e., attention too easily drawn to unimportant or irrelevant external stimuli), as reported or observed.
   6. Increase in goal-directed activity (either socially, at work or school, or sexually) or psychomotor agitation (i.e., purposeless non-goal-directed activity).
   7. Excessive involvement in activities that have a high potential for painful consequences (e.g., engaging in unrestrained buying sprees, sexual indiscretions, or foolish business investments).

C. The mood disturbance is sufficiently severe to cause marked impairment in social or occupational functioning or to necessitate hospitalization to prevent harm to self or others, or there are psychotic features.

D. The episode is not attributable to the physiologic effects of a substance (e.g., a drug of abuse, a medication, or other treatment) or another medical condition.

Source: Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC, American Psychiatric Association, 2013.

Manic episodes typically emerge over a period of days to weeks, but onset within hours is possible, usually in the early morning hours. An untreated episode of either depression or mania can be as short as several weeks or last as long as 8–12 months, and rare patients have an unremitting chronic course. The term *rapid cycling* is used for patients who have four or more episodes of either depression or mania in a given year.

This pattern occurs in 15% of all patients, most of whom are women. In some cases, rapid cycling is linked to an underlying thyroid dysfunction, and in others, it is iatrogenically triggered by prolonged antidepressant treatment. Approximately one-half of patients have sustained difficulties in work performance and psychosocial functioning, with depressive phases being more responsible for impairment than mania.

Bipolar disorder is common, affecting ~1.5% of the population in the United States. Onset is typically between 20 and 30 years of age, but many individuals report premorbid symptoms in late childhood or early adolescence. The prevalence is similar for men and women; women are likely to have more depressive and men more manic episodes over a lifetime.

Differential Diagnosis

The differential diagnosis of mania includes secondary mania induced by stimulant or sympathomimetic drugs, hyperthyroidism, AIDS, and neurologic disorders such as Huntington's or Wilson's disease and cerebrovascular accidents. Comorbidity with alcohol and substance abuse is common, either because of poor judgment and increased impulsivity or because of an attempt to self-treat the underlying mood symptoms and sleep disturbances.

Etiology and Pathophysiology

Genetic predisposition to bipolar disorder is evident from family studies; the concordance rate for monozygotic twins approaches 80%. A number of risk genes that have been identified to date overlap with those conveying risk for other psychiatric disorders, such as schizophrenia and autism, implying some degree of shared pathophysiology. Replicated loci include the alpha subunit of the L-type calcium channel (*CACNA1C*), teneurin transmembrane protein 4 (*ODZ4*), ankryn 3 (*ANK3*), neurocan (NCAN), and tetratricopeptide repeat and ankyrin repeat containing 1 (*TRANK1*). No clear biomarkers have been identified, but there is evidence for circadian rhythm dysregulation and oxidative stress, mitochondrial, and endoplasmic reticulum abnormalities.

TREATMENT

TREATMENT

Bipolar Disorder

(Table 444-9) Lithium carbonate is the mainstay of treatment in bipolar disorder, although sodium valproate and carbamazepine, as well as a number of second-generation antipsychotic agents (aripiprazole, asenapine, olanzapine, quetiapine, risperidone, ziprasidone), also have FDA approval for the treatment of acute mania. Oxcarbazepine is not FDA approved, but appears to enjoy carbamazepine's spectrum of efficacy. The response rate to lithium carbonate is 70–80% in acute mania, with beneficial effects appearing in 1–2 weeks. Lithium also has a prophylactic effect in prevention of recurrent mania and, to a lesser extent, in the prevention of recurrent depression, which is more difficult to treat than unipolar depression. A simple cation,

lithium is rapidly absorbed from the gastrointestinal tract and remains unbound to plasma or tissue proteins. Some 95% of a given dose is excreted unchanged through the kidneys within 24 h.

Serious side effects from lithium are rare, but minor complaints such as gastrointestinal discomfort, nausea, diarrhea, polyuria, weight gain, skin eruptions, alopecia, and edema are common. Over time, urine-concentrating ability may be decreased, but significant nephrotoxicity is relatively rare. Lithium exerts an antithyroid effect by interfering with the synthesis and release of thyroid hormones. More serious side effects include tremor, poor concentration and memory, ataxia, dysarthria, and incoordination. There is suggestive, but not conclusive, evidence that lithium is teratogenic, inducing cardiac malformations in the first trimester.

In the treatment of acute mania, lithium is initiated at 300 mg bid or tid, and the dose is then increased by 300 mg every 2–3 days to achieve blood levels of 0.8–1.2 meq/L. Because the therapeutic effect of lithium may not appear until after 7–10 days of treatment, adjunctive usage of lorazepam (1–2 mg every 4 h) or clonazepam (0.5–1 mg every 4 h) may be beneficial to control agitation. Antipsychotics are indicated in patients with severe agitation who respond only partially to benzodiazepines. Patients using lithium should be monitored closely, since the blood levels required to achieve a therapeutic benefit are close to those associated with toxicity.

Valproic acid may be more effective than lithium for patients who experience rapid cycling (i.e., more than four episodes a year) or who present with a mixed or dysphoric mania. Tremor and weight gain are the most common side effects; hepatotoxicity and pancreatitis are rare toxicities.

The recurrent nature of bipolar mood disorder necessitates maintenance treatment. A sustained blood lithium level of at least 0.8 meq/L is important for optimal prophylaxis and has been shown to reduce the risk of suicide, a finding not yet apparent for other mood stabilizers. Combinations of mood stabilizers together or with atypical antipsychotic drugs are sometimes required to maintain mood stability. Quetiapine extended release, olanzapine, risperidone, and lamotrigine have been approved for maintenance treatment as sole agents, in combination with lithium and with aripiprazole and ziprasidone as adjunctive drugs. Lurasidone, olanzapine/fluoxetine, and quetiapine are also approved to treat acute depressive episodes in bipolar disorder. Compliance is frequently an issue and often requires enlistment and education of concerned family members. Efforts to identify and modify psychosocial factors that may trigger episodes are important, as is an emphasis on lifestyle regularity (social rhythm therapy). Mobile apps for smartphones that alert the individual and clinician to changes in activity and speech are proving useful in early detection of behavioral change and in delivering clinical interventions and education. Antidepressant medications are sometimes required for the treatment of severe breakthrough depressions, but their use should generally be avoided during maintenance treatment because of the risk of precipitating mania or accelerating the cycle frequency. Alternative off-label agents for bipolar depression include pramipexole, modafenil, omega-3 fatty acids, and *N*-acetyl cysteine; interventions such as ECT, light therapy and rTMS may also be effective. Loss of efficacy over time may be observed with any of the mood-stabilizing agents. In such situations, an alternative agent or combination therapy is usually helpful.

TABLE 444-9

Clinical Pharmacology of Mood Stabilizers

Agent and Dosing	Side Effects and Other Effects
Lithium	**Common Side Effects**
Starting dose: 300 mg bid or tid Therapeutic blood level: 0.8–1.2 meq/L	Nausea/anorexia/diarrhea, fine tremor, thirst, polyuria, fatigue, weight gain, acne, folliculitis, neutrophilia, hypothyroidism Blood level is increased by thiazides, tetracyclines, and NSAIDs
	Blood level is decreased by bronchodilators, verapamil, and carbonic anhydrase inhibitors
	Rare side effects: Neurotoxicity, renal toxicity, hypercalcemia, ECG changes
Valproic Acid	**Common Side Effects**
Starting dose: 250 mg tid Therapeutic blood level: 50–125 µg/mL	Nausea/anorexia, weight gain, sedation, tremor, rash, alopecia Inhibits hepatic metabolism of other medications
	Rare side effects: Pancreatitis, hepatotoxicity, Stevens-Johnson syndrome
Carbamazepine/Oxcarbazepine	**Common Side Effects**
Starting dose: 200 mg bid for carbamazepine, 150 mg bid for oxcarbazepine	Nausea/anorexia, sedation, rash, dizziness/ataxia
Therapeutic blood level: 4–12 µg/mL for carbamazepine	Carbamazepine, but not oxcarbazepine, induces hepatic metabolism of other medications
	Rare side effects: Hyponatremia, agranulocytosis, Stevens-Johnson syndrome
Lamotrigine	**Common Side Effects**
Starting dose: 25 mg/d	Rash, dizziness, headache, tremor, sedation, nausea

Agent and Dosing	Side Effects and Other Effects
	Rare side effect: Stevens-Johnson syndrome

Abbreviations: ECG, electrocardiogram; NSAIDs, nonsteroidal anti-inflammatory drugs.

SOMATIC SYMPTOM DISORDER

Many patients presenting in general medical practice, perhaps as many as 5–7%, will experience a somatic symptom(s) as particularly distressing and preoccupying, to the point that it comes to dominate their thoughts, feelings, and beliefs and interferes to a varying degree with everyday functioning. Although the absence of a medical explanation for these complaints was historically emphasized as a diagnostic element, it has been recognized that the patient's interpretation and elaboration of the experience is the critical defining factor and that patients with well-established medical causation may qualify for the diagnosis. Multiple complaints are typical, but severe single symptoms can occur as well. Comorbidity with depressive and anxiety disorders is common and may affect the severity of the experience and its functional consequences. Personality factors may be a significant risk factor, as may a low level of educational or socioeconomic status or a history of recent stressful life events. Cultural factors are relevant as well and should be incorporated into the evaluation. Individuals who have persistent preoccupations about having or acquiring a serious illness, but who do not have a specific somatic complaint, may qualify for a related diagnosis—illness anxiety disorder. The diagnosis of conversion disorder (functional neurologic symptom disorder) is used to specifically identify those individuals whose somatic complaints involve one or more symptoms of altered voluntary motor or sensory function that cannot be medically explained and that causes significant distress or impairment or requires medical evaluation.

In *factitious illnesses*, the patient consciously and voluntarily produces physical symptoms of illness. The term *Munchausen's syndrome* is reserved for individuals with particularly dramatic, chronic, or severe factitious illness. In true factitious illness, the sick role itself is gratifying. A variety of signs, symptoms, and diseases have been either simulated or caused by factitious behavior, the most common including chronic diarrhea, fever of unknown origin, intestinal bleeding or hematuria, seizures, and hypoglycemia. Factitious disorder is usually not diagnosed until 5–10 years after its onset, and it can produce significant social and medical costs. In *malingering*, the fabrication derives from a desire for some external reward such as a narcotic medication or disability reimbursement.

TREATMENT

TREATMENT

Somatic Symptom Disorder and Related Disorders

Patients with somatic symptom disorder are frequently subjected to many diagnostic tests and exploratory surgeries in an attempt to find their “real” illness. Such an approach is doomed to failure and does not address the core issue. Successful treatment is best achieved through behavior modification, in which access to the physician is tightly regulated and adjusted to provide a sustained and predictable level of support that is less clearly contingent on the patient’s level of presenting distress. Visits can be brief and should not be associated with a need for a diagnostic or treatment action. Although the literature is limited, some patients may benefit from antidepressant treatment.

Any attempt to confront the patient usually creates a sense of humiliation and causes the patient to abandon treatment from that caregiver. A better strategy is to introduce psychological causation as one of a number of possible explanations in the differential diagnoses that are discussed. Without directly linking psychotherapeutic intervention to the diagnosis, the patient can be offered a face-saving means by which the pathologic relationship with the health care system can be examined and alternative approaches to life stressors developed. Specific medical treatments also may be indicated and effective in treating some of the functional consequences of conversion disorder.

FEEDING AND EATING DISORDERS

CLINICAL MANIFESTATIONS

Feeding and eating disorders constitute a group of conditions in which there is a persistent disturbance of eating or associated behaviors that significantly impair an individual’s physical health or psychosocial functioning. In DSM-5 the described categories (with the exception of pica) are defined to be mutually exclusive in a given episode, based on the understanding that although they are phenotypically similar in some ways, they differ in course, prognosis, and effective treatment interventions. Compared with DSM-IV-TR, three disorders (i.e., avoidant/restrictive food intake disorder, rumination disorder, pica) that were previously classified as disorders of infancy or childhood have been grouped together with the disorders of anorexia and bulimia nervosa. Binge-eating disorder is also now included as a formal diagnosis; the intent of each of these modifications is to encourage clinicians to be more specific in their codification of eating and feeding pathology.

PICA

Pica is diagnosed when the individual, aged >2, eats one or more nonnutritive, nonfood substances for a month or more and requires medical attention as a result. There is usually no specific aversion to food in general but a preferential choice to ingest substances such as clay, starch, soap, paper, or ash. The diagnosis requires the exclusion of specific culturally approved practices and has not been commonly found to be caused by a specific nutritional deficiency. Onset is most common in childhood but the disorder can occur in association with other major psychiatric conditions in adults. An association with pregnancy has been observed, but the condition is only diagnosed when medical risks are increased by the behavior.

RUMINATION DISORDER

In this condition, individuals who have no demonstrable associated gastrointestinal or other medical condition repeatedly regurgitate their food after eating and then either rechew or swallow it or spit it out. The behavior typically occurs on a daily basis and must persist for at least 1 month. Weight loss and malnutrition are common sequelae, and individuals may attempt to conceal their behavior, either by covering their mouth or through social avoidance while eating. In infancy, the onset is typically between 3 and 12 months, and the behavior may remit spontaneously, although in some it appears to be recurrent.

AVOIDANT/RESTRICTIVE FOOD INTAKE DISORDER

The cardinal feature of this disorder is avoidance or restriction of food intake, usually stemming from a lack of interest in or distaste of food and associated with weight loss, nutritional deficiency, dependency on nutritional supplementation, or marked impairment in psychosocial functioning, either alone or in combination. Culturally approved practices, such as fasting or a lack of available food, must be excluded as possible causes. The disorder is distinguished from anorexia nervosa by the presence of emotional factors, such as a fear of gaining weight and distortion of body image in the latter condition. Onset is usually in infancy or early childhood, but avoidant behaviors may persist into adulthood. The disorder is equally prevalent in males and females and is frequently comorbid with anxiety and cognitive and attention-deficit disorders and situations of familial stress. Developmental delay and functional deficits may be significant if the disorder is long-standing and unrecognized.

ANOREXIA NERVOSA

Individuals are diagnosed with anorexia nervosa if they restrict their caloric intake to a degree that their body weight deviates significantly from age, gender, health, and developmental norms and if they also exhibit a fear of gaining weight and an associated disturbance in body image. The condition is further characterized by differentiating those who achieve their weight loss predominantly through restricting intake or by excessive exercise (restricting type) from those who engage in recurrent binge eating and/or subsequent purging, self-induced vomiting, and usage of enemas, laxatives, or diuretics (binge-eating/purging type). Such subtyping is more state than trait specific, as individuals may transition from one profile to the other over time. Determination of whether an individual satisfies the primary criterion of significant low weight is complex and must be individualized, using all available historical information and comparison of body habitus to international body mass norms and guidelines.

Individuals with anorexia nervosa frequently lack insight into their condition and are in denial about possible medical consequences; they often are not comforted by their achieved weight loss and persist in their behaviors despite having met previously self-designated weight goals. Recent research has identified alterations in the circuitry of reward sensitivity and executive function in anorexia and implicated disturbances in frontal cortex and anterior insula regulation of interoceptive awareness of satiety and hunger. Neurochemical findings, including the role of ghrelin, remain controversial.

Onset is most common in adolescence, although onset in later life can occur. Many more females than males are affected, with a lifetime prevalence in women of up to 4%. The disorder appears most prevalent in

postindustrialized and urbanized countries and is frequently comorbid with preexisting anxiety disorders. The medical consequences of prolonged anorexia nervosa are multisystemic and can be life-threatening in severe presentations. Changes in blood chemistry include leukopenia with lymphocytosis, elevations in blood urea nitrogen, and metabolic alkalosis and hypokalemia when purging is present. History and physical examination may reveal amenorrhea in females, skin abnormalities (petechiae, lanugo hair, dryness), and signs of hypometabolic function, including hypotension, hypothermia, and sinus bradycardia. Endocrine effects include hypogonadism, growth hormone resistance, and hypercortisolemia. Osteoporosis is a longer-term concern.

The course of the disorder is variable, with some individuals recovering after a single episode, while others exhibit recurrent episodes or a chronic course. Untreated anorexia has a mortality of 5.1/1000, the highest among psychiatric conditions. Maudsley Anorexia Nervosa Treatment for Adults (MANTRA) and eating disorder focused cognitive behavior therapy have proven to be effective therapies, with strict behavioral contingencies used when weight loss becomes critical. No pharmacologic intervention has proven to be specifically beneficial, but comorbid depression and anxiety should be treated. Weight gain should be undertaken gradually with a goal of 0.5–1 pound per week to prevent refeeding syndrome. Most individuals are able to achieve remission within 5 years of the original diagnosis.

BULIMIA NERVOSA

Bulimia nervosa describes individuals who engage in recurrent and frequent (at least once a week for 3 months) periods of binge eating and who then resort to compensatory behaviors, such as self-induced purging, enemas, use of laxatives, or excessive exercise to avoid weight gain. Binge eating itself is defined as excessive food intake in a prescribed period of time, usually <2 h. As in anorexia nervosa, disturbances in body image occur and promote the behavior, but unlike in anorexia, individuals are of normal weight or even somewhat overweight. Subjects typically describe a loss of control and express shame about their actions, and often relate that their episodes are triggered by feelings of negative self-esteem or social stresses. The lifetime prevalence in women is ~2%, with a 10:1 female-to-male ratio. The disorder typically begins in adolescence and may be persistent over a number of years. Transition to anorexia occurs in only 10–15% of cases. Many of the medical risks associated with bulimia nervosa parallel those of anorexia nervosa and are a direct consequence of purging, including fluid and electrolyte disturbances and conduction abnormalities. Physical examination often results in no specific findings, but dental erosion and parotid gland enlargement may be present. Effective treatment approaches include SSRI antidepressants, usually in combination with cognitive-behavioral, emotion regulation, or interpersonal-based psychotherapies.

BINGE-EATING DISORDER

Binge-eating disorder is distinguished from bulimia nervosa by the absence of compensatory behaviors to prevent weight gain after an episode and by a lack of effort to restrict weight gain between episodes. Other features are similar, including distress over the behavior and the experience of loss of control, resulting in eating more rapidly or in greater amounts than intended or eating when not hungry. The 12-month

prevalence in females is 1.6%, with a much lower female-to-male ratio than bulimia nervosa. Little is known about the course of the disorder, given its recent categorization, but its prognosis is markedly better than for other eating disorders, both in terms of its natural course and response to treatment. Transition to other eating disorder conditions is thought to be rare.

PERSONALITY DISORDERS

CLINICAL MANIFESTATIONS

Personality disorders are characteristic patterns of thinking, feeling, and interpersonal behavior that are relatively inflexible and cause significant functional impairment or subjective distress for the individual. The observed behaviors are not secondary to another mental disorder, nor are they precipitated by substance abuse or a general medical condition. This distinction is often difficult to make in clinical practice, because personality change may be the first sign of serious neurologic, endocrine, or other medical illness. Patients with frontal lobe tumors, for example, can present with changes in motivation and personality while the results of the neurologic examination remain within normal limits. Individuals with personality disorders are often regarded as “difficult patients” in clinical medical practice because they are seen as excessively demanding and/or unwilling to follow recommended treatment plans. Although DSM-5 portrays personality disorders as qualitatively distinct categories, there is an alternative and emerging perspective that personality characteristics vary as a continuum between normal functioning and formal mental disorder, the essential features being moderate or greater impairment in self/interpersonal functioning and one or more pathological personality traits.

Personality disorders have been grouped into three overlapping clusters. *Cluster A* includes paranoid, schizoid, and schizotypal personality disorders. It includes individuals who are odd and eccentric and who maintain an emotional distance from others. Individuals have a restricted emotional range and remain socially isolated. Patients with schizotypal personality disorder frequently have unusual perceptual experiences and express magical beliefs about the external world. The essential feature of paranoid personality disorder is a pervasive mistrust and suspiciousness of others to an extent that is unjustified by available evidence. *Cluster B* disorders include antisocial, borderline, histrionic, and narcissistic types and describe individuals whose behavior is impulsive, excessively emotional, and erratic. *Cluster C* incorporates avoidant, dependent, and obsessive-compulsive personality types; enduring traits are anxiety and fear. The boundaries between cluster types are to some extent artificial, and many patients who meet criteria for one personality disorder also meet criteria for aspects of another. The risk of a comorbid major mental disorder is increased in patients who qualify for a diagnosis of personality disorder.

ETIOLOGY AND PATHOPHYSIOLOGY

Genetic studies have increasingly suggested a genetic contribution to the development of personality disorders. One study of 106,000 subjects identified 9 loci significantly linked to aspects of neuroticism.

TREATMENT

TREATMENT

Personality Disorders

Dialectical behavior therapy (DBT) is a cognitive-behavioral approach that focuses on behavioral change while providing acceptance, compassion, and validation of the patient. Several randomized trials have demonstrated the efficacy of DBT in the treatment of personality disorders. Antidepressant medications and low-dose antipsychotic drugs have some efficacy in cluster A personality disorders, whereas anticonvulsant mood-stabilizing agents and MAOIs may be considered for patients with cluster B diagnoses who show marked mood reactivity, behavioral dyscontrol, and/or rejection hypersensitivity. Anxious or fearful cluster C patients often respond to medications used for axis I anxiety disorders (see above). It is important that the physician and the patient have reasonable expectations vis-à-vis the possible benefit of any medication used and its side effects. Improvement may be subtle and observable only over time.

SCHIZOPHRENIA

CLINICAL MANIFESTATIONS

Schizophrenia is a heterogeneous syndrome characterized by perturbations of language, perception, thinking, social activity, affect, and volition. There are no pathognomonic features. The syndrome commonly begins in late adolescence, has an insidious (and less commonly acute) onset, and, often, a poor outcome, progressing from social withdrawal and perceptual distortions to recurrent delusions and hallucinations. Patients may present with positive symptoms (such as conceptual disorganization, delusions, or hallucinations) or negative symptoms (loss of function, anhedonia, decreased emotional expression, impaired concentration, and diminished social engagement) and must have at least two of these for a 1-month period and continuous signs for at least 6 months to meet formal diagnostic criteria. Disorganized thinking or speech and grossly disorganized motor behavior, including catatonia, may also be present. As individuals age, positive psychotic symptoms tend to attenuate, and some measure of social and occupational function may be regained. “Negative” symptoms predominate in one-third of the schizophrenic population and are associated with a poor long-term outcome and a poor response to drug treatment. However, marked variability in the course and individual character of symptoms is typical.

The term *schizophreniform disorder* describes patients who meet the symptom requirements but not the duration requirements for schizophrenia, and *schizoaffective disorder* is used for those who manifest symptoms of schizophrenia and independent periods of mood disturbance. The terms *schizotypal* and *schizoid* refer to specific personality disorders and are discussed in that section. The diagnosis of delusional disorder is used for individuals who have delusions of various content for at least 1 month but who otherwise do not meet criteria for schizophrenia. Patients who experience a sudden onset of a brief (<1 month) alteration in thought processing, characterized by delusions, hallucinations, disorganized speech, or gross motor behavior, are most appropriately designated as having a brief psychotic disorder. Catatonia is recognized as a nonspecific syndrome that can occur as a consequence of other severe psychiatric/medical disorders and is diagnosed by the documentation of three or more of a cluster of motor and behavioral

symptoms, including stupor, cataplexy, mutism, waxy flexibility, and stereotypy, among others. Prognosis depends not on symptom severity but on the response to antipsychotic medication. A permanent remission without recurrence does occasionally occur. About 10% of schizophrenic patients commit suicide.

Schizophrenia is present in 0.85% of individuals worldwide, with a lifetime prevalence of ~1–1.5%. An estimated 300,000 episodes of acute schizophrenia occur annually in the United States, resulting in direct and indirect costs of $62.7 billion.

DIFFERENTIAL DIAGNOSIS

The diagnosis is principally one of exclusion, requiring the absence of significant associated mood symptoms, any relevant medical condition, and substance abuse. Drug reactions that cause hallucinations, paranoia, confusion, or bizarre behavior may be dose-related or idiosyncratic; parkinsonian medications, clonidine, quinacrine, and procaine derivatives are the most common prescription medications associated with these symptoms. Drug causes should be ruled out in any case of newly emergent psychosis. The general neurologic examination in patients with schizophrenia is usually normal, but motor rigidity, tremor, and dyskinesias are noted in one-quarter of untreated patients.

EPIDEMIOLOGY AND PATHOPHYSIOLOGY

Epidemiologic surveys identify several risk factors for schizophrenia, including genetic susceptibility, early developmental insults, winter birth, and increasing parental age. Genetic factors are involved in at least a subset of individuals who develop schizophrenia. Schizophrenia is observed in ~6.6% of all first-degree relatives of an affected proband. If both parents are affected, the risk for offspring is 40%. The concordance rate for monozygotic twins is 50%, compared to 10% for dizygotic twins. Schizophrenia-prone families are also at risk for other psychiatric disorders, including schizoaffective disorder and *schizotypal* and *schizoid personality disorders*, the latter terms designating individuals who show a lifetime pattern of social and interpersonal deficits characterized by an inability to form close interpersonal relationships, eccentric behavior, and mild perceptual distortions. Large scale genomewide association studies have identified >100 small effect risk loci and a few larger effect copy number variants, along with epigenetic effects. Pathways identified include ones involved in immunity, inflammation, and cell signaling.

TREATMENT

TREATMENT

Schizophrenia

Antipsychotic agents **(Table 444-10)** are the cornerstone of acute and maintenance treatment of schizophrenia and are effective in the treatment of hallucinations, delusions, and thought disorders, regardless of etiology. The mechanism of action involves, at least in part, binding to dopamine D_2/D_3 receptors in the ventral striatum; the clinical potencies of traditional antipsychotic drugs parallel their

affinities for the D_2 receptor, and even the newer "atypical" agents exert some degree of D_2 receptor blockade. All neuroleptics induce expression of the immediate-early gene c-*fos* in the nucleus accumbens, a dopaminergic site connecting prefrontal and limbic cortices. The clinical efficacy of newer atypical neuroleptics, however, may involve *N*-methyl-D-aspartate (NMDA) receptor blockade, α_1- and α_2-noradrenergic activity, altering the relationship between 5-HT_2 and D_2 receptor activity, and faster dissociation of D_2 binding and effects on neuroplasticity.

Conventional neuroleptics differ in their potency and side effect profile. Older agents, such as chlorpromazine and thioridazine, are more sedating and anticholinergic and more likely to cause orthostatic hypotension, whereas higher potency antipsychotics, such as haloperidol, perphenazine, and thiothixene, are more likely to induce extrapyramidal side effects. The model "atypical" antipsychotic agent is clozapine, a dibenzodiazepine that has a greater potency in blocking the 5-HT_2 than the D_2 receptor and a much higher affinity for the D_4 than the D_2 receptor. Its principal disadvantage is a risk of blood dyscrasias. Paliperidone is a recently approved agent that is a metabolite of risperidone and shares many of its properties. Unlike other antipsychotics, clozapine does not cause a rise in prolactin level. Approximately 30% of patients who do not benefit from conventional antipsychotic agents will have a better response to this drug, which also has a demonstrated superiority to other antipsychotic agents in preventing suicide; however, its side effect profile makes it most appropriate for treatment-resistant cases. Risperidone, a benzisoxazole derivative, is more potent at 5-HT_2 than D_2 receptor sites, like clozapine, but it also exerts significant α_2 antagonism, a property that may contribute to its perceived ability to improve mood and increase motor activity. Risperidone is not as effective as clozapine in treatment-resistant cases but does not carry a risk of blood dyscrasias. Olanzapine is similar neurochemically to clozapine but has a significant risk of inducing weight gain. Quetiapine is distinct in having a weak D_2 effect but potent α_1 and histamine blockade. Ziprasidone causes minimal weight gain and is unlikely to increase prolactin but may increase QT prolongation. Aripiprazole also has little risk of weight gain or prolactin increase but may increase anxiety, nausea, and insomnia as a result of its partial agonist properties. Asenapine is associated with minimal weight gain and anticholinergic effect but may have a higher than expected risk of extrapyramidal symptoms.

Antipsychotic agents are effective in 70% of patients presenting with a first episode. Improvement may be observed within hours or days, but full remission usually requires 6–8 weeks. The choice of agent depends principally on the side effect profile and cost of treatment or on a past personal or family history of a favorable response to the drug in question. Atypical agents appear to be more effective in treating negative symptoms and improving cognitive function. An equivalent treatment response can usually be achieved with relatively low doses of any drug selected (i.e., 4–6 mg/d of haloperidol, 10–15 mg of olanzapine, or 4–6 mg/d of risperidone). Doses in this range result in >80% D_2 receptor blockade, and there is little evidence that higher doses increase either the rapidity or degree of response. Maintenance treatment requires careful attention to the possibility of relapse and monitoring for the development of a movement disorder. Intermittent drug treatment is less effective than regular dosing, but gradual dose reduction is likely to improve social functioning in many schizophrenic patients who have been maintained at high doses. If

medications are completely discontinued, however, the relapse rate is 60% within 6 months. Long-acting injectable preparations (risperidone, paliperidone, olanzapine, aripiprazole) are considered when noncompliance with oral therapy leads to relapses but should not be considered interchangeable, because the agents differ in their indications, injection intervals and sites/volumes, and possible adverse reactions, among other factors. In treatment-resistant patients, a transition to clozapine usually results in rapid improvement, but a prolonged delay in response in some cases necessitates a 6- to 9-month trial for maximal benefit to occur.

Antipsychotic medications can cause a broad range of side effects, including lethargy, weight gain, postural hypotension, constipation, and dry mouth. Extrapyramidal symptoms such as dystonia, akathisia, and akinesia are also frequent with first-generation agents and may contribute to poor adherence if not specifically addressed. Anticholinergic and parkinsonian symptoms respond well to trihexyphenidyl, 2 mg bid, or benztropine mesylate, 1–2 mg bid. Akathisia may respond to beta blockers. In rare cases, more serious and occasionally life-threatening side effects may emerge, including hyperprolactinemia, ventricular arrhythmias, gastrointestinal obstruction, retinal pigmentation, obstructive jaundice, and neuroleptic malignant syndrome (characterized by hyperthermia, autonomic dysfunction, muscular rigidity, and elevated creatine phosphokinase levels). The most serious adverse effects of clozapine are agranulocytosis, which has an incidence of 1%, and induction of seizures, which has an incidence of 10%. Weekly white blood cell counts are required, particularly during the first 3 months of treatment.

The risk of type 2 diabetes mellitus appears to be increased in schizophrenia, and second-generation agents as a group produce greater adverse effects on glucose regulation, independent of effects on obesity, than traditional agents. Clozapine, olanzapine, and quetiapine seem more likely to cause hyperglycemia, weight gain, and hypertriglyceridemia than other atypical antipsychotic drugs. Close monitoring of plasma glucose and lipid levels are indicated with the use of these agents.

A serious side effect of long-term use of first-generation and, to a lesser extent, second-generation antipsychotic agents is tardive dyskinesia, characterized by repetitive, involuntary, and potentially irreversible movements of the tongue and lips (bucco-linguo-masticatory triad) and, in approximately half of cases, choreoathetosis. Tardive dyskinesia has an incidence of 2–4% per year of exposure and a prevalence of 20% in chronically treated patients. The prevalence increases with age, total dose, and duration of drug administration and may involve formation of free radicals and perhaps mitochondrial energy failure. Valbenazine, a vesicular monoamine transporter 2 inhibitor that depletes presynaptic dopamine, has recently received FDA approval for treatment of tardive dyskinesia.

The CATIE study, a large-scale investigation of the effectiveness of antipsychotic agents in “real-world” patients, revealed a high rate of discontinuation of treatment >18 months. Olanzapine showed greater effectiveness than quetiapine, risperidone, perphenazine, or ziprasidone but also a higher discontinuation rate due to weight gain and metabolic effects. Surprisingly, perphenazine, a first-generation agent, showed little evidence of inferiority to newer drugs.

Drug treatment of schizophrenia is by itself insufficient. Educational efforts directed toward families and relevant community resources have proved to be necessary to maintain stability and optimize outcome. A treatment model using social cognition interventions and involving a multidisciplinary case-management team that seeks out and closely follows the patient in the community has proved particularly effective. Attempts to prevent schizophrenia through early identification and treatment (both psychosocial and psychopharmacologic) of high-risk children and adolescents are currently being evaluated.

TABLE 444-10

Antipsychotic Agents

Name	Usual PO Daily Dose, mg	Side Effects	Sedation	Comments
First-Generation Antipsychotics				
Low potency				
Chlorpromazine (Thorazine) Thioridazine (Mellaril)	100–1000 100–600	Anticholinergic effects; orthostasis; photosensitivity; cholestasis; QT prolongation	+++	EPSEs usually not prominent; can cause anticholinergic delirium in elderly patients
Midpotency				
Trifluoperazine (Stelazine)	2–50	Fewer anticholinergic side effects	++	Well tolerated by most patients
Perphenazine (Trilafon)	4–64	Fewer EPSEs than with higher potency agents	++	
Loxapine (Loxitane)	30–100	Frequent EPSEs	++	
Molindone (Moban)	30–100	Frequent EPSEs	0	Little weight gain
High potency				
Haloperidol (Haldol)	5–20	No anticholinergic side effects; EPSEs often prominent	0/+	Often prescribed in doses that are too high; long-acting injectable forms of haloperidol and fluphenazine available

Name	Usual PO Daily Dose, mg	Side Effects	Sedation	Comments
Fluphenazine (Prolixin)	1–20	Frequent EPSEs	0/+	
Thiothixene (Navane)	2–50	Frequent EPSEs	0/+	
Second-Generation Antipsychotics				
Clozapine (Clozaril)	150–600	Agranulocytosis (1%); weight gain; seizures; drooling; hyperthermia	+ +	Requires weekly WBC count for first 6 months, then biweekly if stable
Risperidone (Risperdal)	2–8	Orthostasis	+	Requires slow titration; EPSEs observed with doses >6 mg qd
Olanzapine (Zyprexa)	10–30	Weight gain	++	Mild prolactin elevation
Quetiapine (Seroquel)	350–800	Sedation; weight gain; anxiety	+++	Bid dosing
Ziprasidone (Geodon)	120–200	Orthostatic hypotension	+/++	Minimal weight gain; increases QT interval
Aripiprazole (Abilify)	10–30	Nausea, anxiety, insomnia	0/+	Mixed agonist/antagonist; ER available
Paliperidone (Invega)	3–12	Restlessness, EPSEs, increased prolactin, headache	+	Active metabolite of risperidone
Iloperidone (Fanapt)	12–24	Dizziness, hypotension	0/+	Requires dose titration; long acting injectable available

Name	Usual PO Daily Dose, mg	Side Effects	Sedation	Comments
Asenapine (Saphris)	10–20	Dizziness, anxiety, EPSEs, minimal weight gain	++	Sublingual tablets; bid dosing
Lurasidone (Latuda)	40–80	Nausea, EPSEs	++	Uses CYP3A4
Brexpiprazole (Rexulti)	1–4	Anxiety, dizziness, fatigue	++	CYP3A4 and 2D6 interactions
Pimavanserin (Nuplazid)	34	Edema, confusion, sedation	++	Approved for Parkinson's disease psychosis
Cariprazine (Vraylar)	1.5–6	EPSEs, vomiting	++	Preferential D3 receptor affinity

Abbreviations: EPSEs, extrapyramidal side effects; WBC, white blood cell.

ASSESSMENT AND EVALUATION OF VIOLENCE

Primary care physicians may encounter situations in which family, domestic, or societal violence is discovered or suspected. Such an awareness can carry legal and moral obligations; many state laws mandate reporting of child, spousal, and elder abuse. Physicians are frequently the first point of contact for both victim and abuser. Approximately 2 million older Americans and 1.5 million U.S. children are thought to experience some form of physical maltreatment each year. Spousal abuse is thought to be even more prevalent. An interview study of 24,000 women in 10 countries found a lifetime prevalence of physical or sexual violence that ranged from 15 to 71%; these individuals are more likely to suffer from depression, anxiety, and substance abuse and to have attempted suicide. In addition, abused individuals frequently express low self-esteem, vague somatic symptomatology, social isolation, and a passive feeling of loss of control. Although it is essential to treat these elements in the victim, the first obligation is to ensure that the perpetrator has taken responsibility for preventing any further violence. Substance abuse and/or dependence and serious mental illness in the abuser may contribute to the risk of harm and require direct intervention. Depending on the situation, law enforcement agencies, community resources such as support groups and shelters, and individual and family counseling can be appropriate components of a treatment

plan. A safety plan should be formulated with the victim, in addition to providing information about abuse, its likelihood of recurrence, and its tendency to increase in severity and frequency. Antianxiety and antidepressant medications may sometimes be useful in treating the acute symptoms, but only if independent evidence for an appropriate psychiatric diagnosis exists.

FURTHER READING

Combs G, Oshman L: Pearls for working with people who have personality disorder diagnoses. Prim Care 43:263, 2017.

Craske MG, Stein MB: Anxiety. Lancet 388:3048, 2016. [PubMed: 27349358]

Gelauff J, Stone J: Prognosis of functional neurologic disorders. Handb Clin Neurol 139:523, 2016. [PubMed: 27719869]

Gitlin M, Miklowitz DJ: The difficult lives of individuals with bipolar disorder: A review of functional outcomes and their implications for treatment. J Affec Disord 209:147, 2017.

Kok RM, Reynolds CF: Management of depression in older adults: A review. JAMA 317:2114, 2017. [PubMed: 28535241]

Maselko J: Social epidemiology and global mental health: Expanding the evidence from high-income to low- and middle-income countries. Curr Epidemiol Rep 3:166, 2017

Morin L, Franck N: Rehabilitation interventions to promote recovery from schizophrenia: A systematic review. Front Psychiatry 8:100, 2017. [PubMed: 28659832]

National Guideline Alliance (UK): Eating Disorders: Recognition and treatment. National Institute for Health and Care Excellence (NICE) May 2017.

Ortiz-Orendain J et al: Antipsychotic combinations for schizophrenia. Cochrane Database Syst Rev 28:6, 2017.

Shalev A et al: Post-traumatic stress disorder. N Engl J Med 376:2459, 2017. [PubMed: 28636846]

Chapter 445: Alcohol and Alcohol Use Disorders

Marc A. Schuckit

FIGURE 445-1

INTRODUCTION

Alcohol (beverage ethanol) has diverse and widespread effects on the body, and impacts directly or indirectly on almost every neurochemical system in the brain. A large majority of patients in most clinical settings consume alcohol, with the highest proportions of drinkers of at least modest levels of alcohol seen in more educated and affluent patient groups. At even relatively low doses, this drug can exacerbate most medical problems and affect medications metabolized in the liver, and at higher doses can temporarily mimic many medical (e.g., diabetes) and psychiatric (e.g., depression) conditions. The lifetime risk for repetitive serious alcohol problems (e.g., alcohol use disorders as described below) in patients is about 20% for men and 10% for women, regardless of a person's education or income. Although low doses of alcohol might have healthful benefits, greater than three standard drinks per day enhances the risk for cancer and vascular disease, and alcohol use disorders decrease the life span by about 10 years. Unfortunately, most clinicians have had only limited training regarding identifying and treating alcohol-related disorders. This chapter presents a brief overview of clinically useful information about alcohol use and associated problems.

PHARMACOLOGY AND NUTRITIONAL IMPACT OF ETHANOL

Ethanol blood levels are expressed as milligrams or grams of ethanol per deciliter (e.g., 100 mg/dL = 0.10 g/dL), with values of ~0.02 g/dL resulting from the ingestion of one typical drink. In round figures, a standard drink is 10–12 g, as seen in 340 mL (12 oz) of beer, 115 mL (4 oz) of nonfortified wine, and 43 mL (1.5 oz) (a shot) of 80-proof beverage (e.g., whisky); 0.5 L (1 pint) of 80-proof beverage contains ~160 g of ethanol (about 16 standard drinks), and 750 mL of wine contains ~60 g of ethanol. These beverages also have additional components (*congeners*) that affect the drink's taste and might contribute to adverse effects on the body. Congeners include methanol, butanol, acetaldehyde, histamine, tannins, iron, and lead. Alcohol acutely decreases neuronal activity and has similar behavioral effects and cross-tolerance with other depressants, including benzodiazepines and barbiturates.

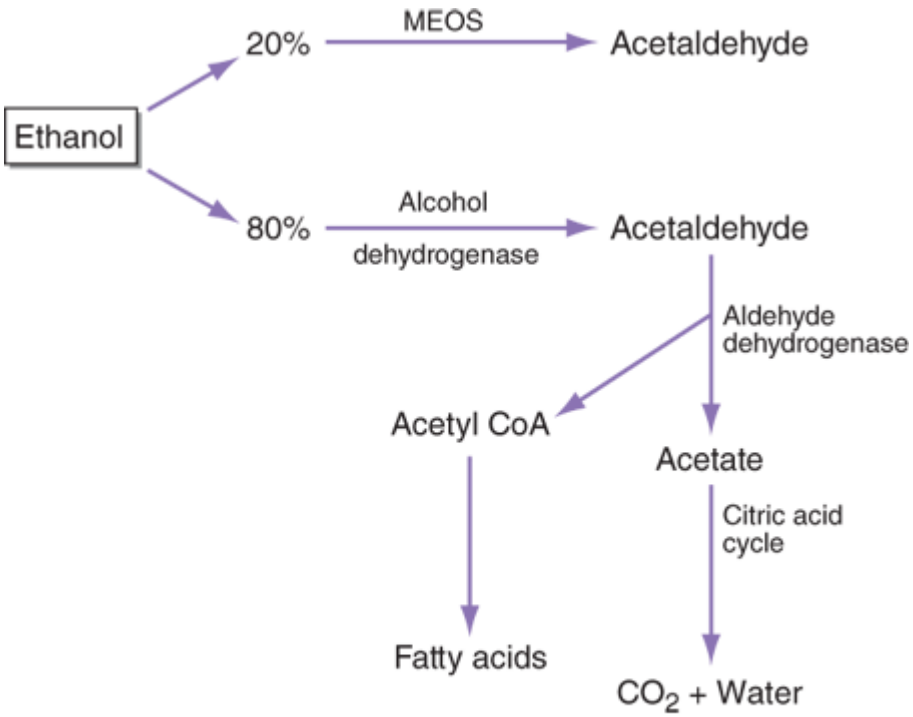
Alcohol is absorbed from mucous membranes of the mouth and esophagus (in small amounts), from the stomach and large bowel (in modest amounts), and from the proximal portion of the small intestine (the major site). The rate of absorption is increased by rapid gastric emptying (as seen with carbonation); by the

absence of proteins, fats, or carbohydrates (which interfere with absorption); and by dilution to a modest percentage of ethanol (maximum at ~20% by volume).

Between 2% (at low blood alcohol concentrations) and 10% (at high blood alcohol concentrations) of ethanol is excreted directly through the lungs, urine, or sweat, but most is metabolized to acetaldehyde, primarily in the liver. The most important pathway occurs in the cell cytosol where alcohol dehydrogenase (ADH) produces acetaldehyde, which is then rapidly destroyed by aldehyde dehydrogenase (ALDH) in the cytosol and mitochondria **(Fig. 445-1)**. A second pathway occurs in the microsomes of the smooth endoplasmic reticulum (the microsomal ethanol-oxidizing system or MEOS) that is responsible for ≥10% of ethanol oxidation at high blood alcohol concentrations.

FIGURE 445-1

The metabolism of alcohol. CoA, coenzyme A; MEOS, microsomal ethanol oxidizing system.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Although a drink contains ~300 kJ, or 70–100 kcal, these are devoid of minerals, proteins, and vitamins. In addition, alcohol interferes with absorption of vitamins in the small intestine and decreases their storage in the liver with modest effects on folate (folacin or folic acid), pyridoxine (B_6), thiamine (B_1), nicotinic acid (niacin, B_3), and vitamin A.

Heavy drinking in a fasting, healthy individual can produce transient hypoglycemia within 6–36 h, secondary to the acute actions of ethanol that decrease gluconeogenesis. This can result in temporary abnormal glucose tolerance tests (with a resulting erroneous diagnosis of diabetes mellitus) until the heavy drinker has abstained for 2–4 weeks. Alcohol ketoacidosis, probably reflecting a decrease in fatty acid oxidation coupled with poor diet or persistent vomiting, can be misdiagnosed as diabetic ketosis. With alcohol-related

ketoacidosis, patients show an increase in serum ketones along with a mild increase in glucose but a large anion gap, a mild to moderate increase in serum lactate, and a β-hydroxybutyrate/lactate ratio of between 2:1 and 9:1 (with normal being 1:1).

In the brain, alcohol affects almost all neurotransmitter systems, with acute effects that are often the opposite of those seen following desistance after a period of heavy drinking. The most prominent acute actions relate to boosting γ-aminobutyric acid (GABA) activity, especially at $GABA_A$ receptors. Enhancement of this complex chloride channel system contributes to anticonvulsant, sleep-inducing, antianxiety, and muscle relaxation effects of all GABA-boosting drugs. Acutely administered alcohol produces a release of GABA, and continued use increases density of $GABA_A$ receptors, whereas alcohol withdrawal states are characterized by decreases in GABA-related activity. Equally important is the ability of acute alcohol to inhibit postsynaptic *N*-methyl-D-aspartate (NMDA) excitatory glutamate receptors, whereas chronic drinking and desistance are associated with an upregulation of these excitatory receptor subunits. The relationships between greater GABA and diminished NMDA receptor activity during acute intoxication and diminished GABA with enhanced NMDA actions during alcohol withdrawal explain much of intoxication and withdrawal phenomena.

As with all pleasurable activities, alcohol acutely increases dopamine levels in the ventral tegmentum and related brain regions, and this effect plays an important role in continued alcohol use, craving, and relapse. The changes in dopamine pathways are also linked to increases in "stress hormones," including cortisol and adrenocorticotropic hormone (ACTH) during intoxication and in the context of the stresses of withdrawal. Such alterations are likely to contribute to both feelings of reward during intoxication and depression during falling blood alcohol concentrations. Also closely linked to alterations in dopamine (especially in the nucleus accumbens) are alcohol-induced changes in opioid receptors, with acute alcohol causing release of β-endorphins.

Additional neurochemical changes include increases in synaptic levels of serotonin during acute intoxication and subsequent upregulation of serotonin receptors. Acute increases in nicotinic acetylcholine systems contribute to the impact of alcohol in the ventral tegmental region, which occurs in concert with enhanced dopamine activity. In the same regions, alcohol impacts on cannabinol receptors, with resulting release of dopamine, GABA, and glutamate as well as subsequent effects on brain reward circuits.

BEHAVIORAL EFFECTS, TOLERANCE, AND WITHDRAWAL

The acute effects of a drug depend on the dose, the rate of increase in plasma, the concomitant presence of other drugs, and past experience with the agent. "Legal intoxication" with alcohol in most states is based on a blood alcohol concentration of 0.08 g/dL, some states are considering lowering acceptable levels to <.05 g/dL, and levels of 0.04 are cited for pilots in the United States and automobile drivers in some other countries. However, behavioral, psychomotor, and cognitive changes are seen at 0.02–0.04 g/dL (i.e., after one to two drinks) **(Table 445-1)**. Deep but disturbed sleep can be seen at .15 g/dL in individuals who have

not developed tolerance, and death can occur with levels between 0.30 and 0.40 g/dL. Beverage alcohol is probably responsible for more overdose deaths than any other drug.

TABLE 445-1

Effects of Blood Alcohol Levels in the Absence of Tolerance

Blood Level, g/dL	Usual Effect
0.02	Decreased inhibitions, a slight feeling of intoxication
0.08	Decrease in complex cognitive functions and motor performance
0.20	Obvious slurred speech, motor incoordination, irritability, and poor judgment
0.30	Light coma and depressed vital signs
0.40	Death

Repeated use of alcohol contributes to the need for a greater number of standard drinks to produce effects originally observed with fewer drinks (acquired tolerance), a phenomenon involving at least three compensatory mechanisms. (1) After 1–2 weeks of daily drinking, *metabolic or pharmacokinetic tolerance* can be seen, with up to 30% increases in the rate of hepatic ethanol metabolism. This alteration disappears almost as rapidly as it develops. (2) *Cellular or pharmacodynamic tolerance* develops through neurochemical changes that maintain relatively normal physiologic functioning despite the presence of alcohol. Subsequent decreases in blood levels contribute to symptoms of withdrawal. (3) Individuals learn to adapt their behavior so that they can function better than expected under the influence of the drug (*learned or behavioral tolerance*).

The cellular changes caused by chronic ethanol exposure may not resolve for several weeks or longer following cessation of drinking. Rapid decreases in blood alcohol levels before that time can produce a withdrawal syndrome, which is most intense during the first 5 days, but with some symptoms (e.g., disturbed sleep and anxiety) lasting up to 4–6 months as part of a “protracted withdrawal” syndrome.

THE EFFECTS OF ETHANOL ON ORGAN SYSTEMS

Relatively low doses of alcohol (one or two drinks per day) may have potential beneficial effects of increasing high-density lipoprotein cholesterol and decreasing aggregation of platelets, with a resulting possible decrease in risk for occlusive coronary disease and embolic strokes. Red wine has additional potential health-promoting qualities at relatively low doses due to flavinols and related substances. Such modest drinking might also decrease the risk for vascular dementia and, possibly, Alzheimer’s disease. However, any

potential healthful effects disappear with the regular consumption of three or more drinks per day, and knowledge about the deleterious effects of alcohol can both help the physician to identify patients with alcohol use disorders and to supply them with information that might help motivate changes in behavior.

NERVOUS SYSTEM

Approximately 35% of drinkers overall, including as many as 50% of drinking college students and a much higher proportion of individuals with alcohol use disorders, ever experience a *blackout*. This is an episode of temporary anterograde amnesia, in which the person was awake but forgot all (en bloc blackouts at blood alcohol levels >.20 mg/dL) or part (fragmentary blackouts at >.12 mg/dL) of what occurred during a drinking period.

Another common problem, one seen after as few as one or two drinks shortly before bedtime, is disturbed sleep. Although alcohol might initially help a person fall asleep, it disrupts sleep throughout the rest of the night. The stages of sleep are altered, and times spent in rapid eye movement (REM) and deep sleep early in the night are reduced. Alcohol relaxes muscles in the pharynx, which can cause snoring and exacerbate sleep apnea; symptoms of the latter occur in 75% of men with alcohol use disorders aged ≥60 years. Patients may also experience prominent and sometimes disturbing dreams later in the night. All these sleep impairments can contribute to relapses to drinking in persons with alcohol use disorders.

Another common consequence of alcohol use even at relatively low alcohol levels is impaired judgment and coordination, which increases the risk of injuries. In the United States, ~40% of drinkers have at some time driven while intoxicated. Heavy drinking can also be associated with headache, thirst, nausea, vomiting, and fatigue the following day, a hangover syndrome that is responsible for much missed work and school time and temporary cognitive deficits.

Chronic high alcohol doses cause *peripheral neuropathy* in ~10% of individuals with alcohol use disorders: similar to diabetes, patients experience bilateral limb numbness, tingling, and paresthesias, all of which are more pronounced distally. Approximately 1% of those with alcohol use disorders develop *cerebellar degeneration or atrophy*, producing a syndrome of progressive unsteady stance and gait often accompanied by mild nystagmus; neuroimaging studies reveal atrophy of the cerebellar vermis. Perhaps as few as 1 in 500 individuals with alcohol use disorders develop full *Wernicke's* (ophthalmoparesis, ataxia, and encephalopathy) and *Korsakoff's* (severe retrograde and anterograde amnesia) *syndromes*, although a higher proportion has one or more neuropathologic findings related to these conditions. These result from low levels of thiamine, especially in predisposed individuals with transketolase deficiencies. Repeated heavy drinking can contribute to *cognitive problems* and temporary memory impairment lasting for weeks to months after abstinence. Brain atrophy, evident as ventricular enlargement and widened cortical sulci on magnetic resonance imaging (MRI) and computed tomography (CT) scans, occurs in ~50% of individuals with long-term alcohol use disorders; these changes are usually reversible if abstinence is maintained. Adolescents may be especially vulnerable to alcohol-related brain changes. There is no single "alcoholic dementia" syndrome; rather, this label describes patients who have irreversible cognitive changes (possibly from diverse causes) in the context of chronic alcohol use disorders.

Psychiatric Comorbidity

As many as two-thirds of individuals with alcohol use disorders meet criteria for another independent or temporary substance-induced psychiatric syndrome as defined in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5) of the American Psychiatric Association **(Chap. 444)**. A substantial proportion of those with independent psychiatric conditions (i.e., not just temporary symptoms only seen during intoxication or withdrawal) relate to a preexisting antisocial personality disorder (ASPD) manifesting as severe impulsivity and disinhibition that contribute to both alcohol and drug use disorders. The lifetime ASPD risk is 3% in males, and ≥80% of such individuals demonstrate alcohol- and/or drug-related conditions. Another common psychiatric comorbidity occurs with problems regarding other substances of abuse. The remainder of individuals with alcohol use disorders who have an independent psychiatric syndrome relate to preexisting conditions such as schizophrenia or manic-depressive disease or anxiety syndromes such as panic disorder. The comorbidities of alcohol use disorders with independent psychiatric disorders might represent an overlap in genetic vulnerabilities, impaired judgment regarding the use of alcohol from the independent psychiatric condition, or an attempt to use alcohol to alleviate symptoms of the disorder or side effects of medications.

Many alcohol-related psychiatric syndromes can be seen *temporarily* during heavy drinking and subsequent withdrawal. These alcohol-induced conditions include an intense *sadness* lasting for days to weeks in the midst of heavy drinking seen in 40% of individuals with alcohol use disorders, which tends to disappear over several weeks of abstinence (alcohol-induced mood disorder); 10–30% have temporary severe *anxiety*, often beginning during alcohol withdrawal, which can persist for a month or more after cessation of drinking (alcohol-induced anxiety disorder); and 3–5% have auditory *hallucinations* and/or paranoid delusions while they are otherwise alert and oriented (*alcohol-induced psychotic disorder*).

Treatment of all forms of alcohol-induced psychopathology includes helping patients achieve abstinence and offering supportive care, as well as reassurance and “talk therapy” such as cognitive-behavioral approaches. However, with the exception of short-term antipsychotic medications for substance-induced psychoses, substance-induced psychiatric conditions only rarely require medications. Recovery is likely within several days to 4 weeks of abstinence. Conversely, because alcohol-induced conditions are temporary and do not indicate a need for long-term pharmacotherapy, a history of heavy alcohol intake is an important part of the workup for any patient who presents with any of these psychiatric symptoms.

THE GASTROINTESTINAL SYSTEM

Esophagus and Stomach

Alcohol can cause inflammation of the esophagus and stomach causing epigastric distress and gastrointestinal bleeding, making alcohol one of the most common causes of hemorrhagic gastritis. Violent vomiting can produce severe bleeding through a Mallory-Weiss lesion, a longitudinal tear in the mucosa at the gastroesophageal junction.

The incidence of acute pancreatitis (~25 per 1000 per year) is almost threefold higher in individuals with alcohol use disorders than in the general population, accounting for an estimated 10% or more of the total cases. Alcohol impairs gluconeogenesis in the liver, resulting in a fall in the amount of glucose produced from glycogen, increased lactate production, and decreased oxidation of fatty acids. These contribute to an increase in fat accumulation in liver cells. In healthy individuals, these changes are reversible, but with repeated exposure to ethanol, especially daily heavy drinking, more severe changes in the liver occur, including alcohol-induced hepatitis, perivenular sclerosis, and cirrhosis, with the latter observed in an estimated 15% of individuals with alcohol use disorders **(Chap. 335)**. Perhaps through an enhanced vulnerability to infections, individuals with alcohol use disorders have an elevated rate of hepatitis C, and drinking in the context of that disease is associated with more severe liver deterioration.

CANCER

As few as 1.5 drinks per day increases a woman's risk of breast cancer 1.4-fold. For both sexes, four drinks per day increases the risk for oral and esophageal cancers approximately threefold and rectal cancers by a factor of 1.5; seven to eight or more drinks per day produces an approximately fivefold increased risk for many other cancers. These consequences may result directly from cancer-promoting effects of alcohol and acetaldehyde or indirectly by interfering with immune homeostasis.

HEMATOPOIETIC SYSTEM

Ethanol causes an increase in red blood cell size (mean corpuscular volume [MCV]), which reflects its effects on stem cells. If heavy drinking is accompanied by folic acid deficiency, there can also be hypersegmented neutrophils, reticulocytopenia, and a hyperplastic bone marrow; if malnutrition is present, sideroblastic changes can be observed. Chronic heavy drinking can decrease production of white blood cells, decrease granulocyte mobility and adherence, and impair delayed-hypersensitivity responses to novel antigens (with a possible false-negative tuberculin skin test). Associated immune deficiencies can contribute to vulnerability toward infections, including hepatitis and HIV, and interfere with their treatment. Finally, many individuals with alcohol use disorders have mild thrombocytopenia, which usually resolves within a week of abstinence unless there is hepatic cirrhosis or congestive splenomegaly.

CARDIOVASCULAR SYSTEM

Acutely, ethanol decreases myocardial contractility and causes peripheral vasodilation, with a resulting mild decrease in blood pressure and a compensatory increase in cardiac output. Exercise-induced increases in cardiac oxygen consumption are higher after alcohol intake. These acute effects have little clinical significance for the average healthy drinker but can be problematic when persisting cardiac disease is present.

The consumption of three or more drinks per day results in a dose-dependent increase in blood pressure, which returns to normal within weeks of abstinence. Thus, heavy drinking is an important factor in mild to moderate hypertension. Chronic heavy drinkers also have a sixfold increased risk for coronary artery disease, related, in part, to increased low-density lipoprotein cholesterol, and carry an increased risk for cardiomyopathy through direct effects of alcohol on heart muscle. Symptoms of the latter include unexplained arrhythmias in the presence of left ventricular impairment, heart failure, hypocontractility of heart muscle, and dilation of all four heart chambers with associated potential mural thrombi and mitral valve regurgitation. Atrial or ventricular arrhythmias, especially paroxysmal tachycardia, can also occur temporarily after heavy drinking in individuals showing no other evidence of heart disease—a syndrome known as the "holiday heart."

GENITOURINARY SYSTEM CHANGES, SEXUAL FUNCTIONING, AND FETAL DEVELOPMENT

Heavy drinking in adolescence can affect normal sexual development and reproductive onset. At any age, modest ethanol doses (e.g., blood alcohol concentrations of 0.06 g/dL) can increase sexual drive but also decrease erectile capacity in men. Even in the absence of liver impairment, a significant minority of chronic heavy drinking men show irreversible testicular atrophy with shrinkage of the seminiferous tubules, decreases in ejaculate volume, and a lower sperm count **(Chap. 384)**.

The repeated ingestion of high doses of ethanol by women can result in amenorrhea, a decrease in ovarian size, absence of corpora lutea with associated infertility, and an increased risk of spontaneous abortion. Drinking during pregnancy results in the rapid placental transfer of both ethanol and acetaldehyde, which may contribute to a range of consequences known as fetal alcohol spectrum disorder (FASD). One severe result is the *fetal alcohol syndrome* (FAS), seen in ~5% of children born to heavy-drinking mothers, which can include any of the following: facial changes with epicanthal eye folds; poorly formed ear concha; small teeth with faulty enamel; cardiac atrial or ventricular septal defects; an aberrant palmar crease and limitation in joint movement; and microcephaly with intellectual impairment. Less pervasive FASD conditions include combinations of low birth weight, a lower intelligence quotient (IQ), hyperactive behavior, and some modest cognitive deficits. The amount of ethanol required and the time of vulnerability during pregnancy have not been defined, making it advisable for pregnant women to abstain from alcohol completely.

OTHER EFFECTS

Between one-half and two-thirds of individuals with alcohol use disorders have skeletal muscle weakness caused by acute *alcoholic myopathy*, a condition that improves but which might not fully remit with abstinence. Effects of repeated heavy drinking on the *skeletal system* include changes in calcium metabolism, lower bone density, and decreased growth in the epiphyses, leading to an increased risk for fractures and osteonecrosis of the femoral head. *Hormonal changes* include an increase in cortisol levels, which can remain elevated during heavy drinking; inhibition of vasopressin secretion at rising blood alcohol concentrations and enhanced secretion at falling blood alcohol concentrations (with the final result that most individuals with alcohol use disorders are likely to be slightly overhydrated); a modest and reversible

decrease in serum thyroxine (T_4); and a more marked decrease in serum triiodothyronine (T_3). Hormone irregularities may disappear after a month or more of abstinence.

ALCOHOL USE DISORDERS

Because many drinkers occasionally imbibe to excess, temporary alcohol-related problems are common, especially in the late teens to the late twenties. However, repeated problems in multiple life areas can indicate an alcohol use disorder as defined in DSM-5.

DEFINITIONS AND EPIDEMIOLOGY

An *alcohol use disorder* (aka alcoholism or alcohol dependence in prior diagnostic manuals) is defined as repeated alcohol-related difficulties in at least 2 of 11 life areas that cluster together in the same 12-month period **(Table 445-2)**. Ten of the 11 items in DSM-5 (published in 2013) were taken directly from the 7 dependence and 4 abuse criteria in DSM-IV (most recently revised in 2000), after deleting legal problems and adding craving. Severity of an alcohol use disorder is based on the number of items endorsed: mild is two or three items; moderate is four or five; and severe is six or more of the criterion items. The 2013 diagnostic approach is similar enough to DSM-IV that the following descriptions of associated phenomena are still accurate.

TABLE 445-2

***Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition, Classification of Alcohol Use Disorder (AUD)**

Criteria
Two or more of the following items occurring in the same 12-month period must be endorsed for the diagnosis of an alcohol use disorder[a]: Drinking resulting in recurrent failure to fulfill role obligations Recurrent drinking in hazardous situations Continued drinking despite alcohol-related social or interpersonal problems Tolerance Withdrawal, or substance use for relief/avoidance of withdrawal Drinking in larger amounts or for longer than intended Persistent desire/unsuccessful attempts to stop or reduce drinking Great deal of time spent obtaining, using, or recovering from alcohol Important activities given up/reduced because of drinking Continued drinking despite knowledge of physical or psychological problems caused by alcohol Alcohol craving

[a] *Mild AUD*: 2–3 criteria required; *Moderate AUD*: 4–5 items endorsed; *severe AUD:* 6 or more items endorsed.

The lifetime risk for an alcohol use disorder in most Western countries is about 10–15% for men and 5–8% for women; higher rates are seen in individuals who seek help from health-care deliverers. Between 2001 and 2013, the proportion of the United States population with a current (i.e., past 12-month) alcohol use disorder increased by 49% with increases of almost 100% in women, African Americans, and individuals aged ≥45. Rates are similar in the United States, Canada, Germany, Australia, and the United Kingdom tend to be lower in most Mediterranean countries, such as Italy, Greece, and Israel, and may be higher in Ireland, France, Eastern Europe (e.g., Russia), and Scandinavia. An even higher lifetime prevalence has been reported for most native cultures, including Native Americans, Eskimos, Maori groups, and aboriginal tribes of Australia. These differences in prevalence reflect both cultural and genetic influences, as described below. In Western countries, the typical person with an alcohol use disorder is more often a blue- or white-collar worker or homemaker. The lifetime risk for this disorder among physicians is similar to that of the general population.

GENETICS

Approximately 60% of the risk for alcohol use disorders is attributed to genes, as indicated by the fourfold higher risk in children with an alcohol use disorder parent (even if adopted early in life and raised by nonalcoholics) and a higher risk in identical twins compared to fraternal twins of affected individuals. The genetic variations operate primarily through intermediate characteristics that subsequently combine with environmental influences to alter the risk for heavy drinking and alcohol problems. These include genes

relating to a high risk for all substance use disorders that operate through impulsivity, schizophrenia, and bipolar disorder. Another characteristic, an intense skin flushing response when drinking, decreases the risk for only alcohol use disorders through gene variations for several alcohol-metabolizing enzymes, especially ALDH (a mutation only seen in Asians), and to a lesser extent, variations in ADH.

An additional genetically influenced characteristic that increases the risk for heavy drinking, a low level of response to alcohol, can be seen very early in the drinking career and before acquired tolerance or alcohol used disorders develop. The low response per drink operates, in part, through variations in genes relating to calcium and potassium channels, GABA, nicotinic, dopamine, and serotonin systems. Follow-up studies have demonstrated that this need for higher doses of alcohol to achieve effects predicts future heavy drinking, alcohol problems, and alcohol use disorders. The impact of a low response to alcohol on adverse drinking outcomes is partially mediated by a range of environmental and attitudinal influences, including the selection of heavier-drinking friends, more positive expectations of the effects of high doses of alcohol, and using alcohol to cope with stress.

NATURAL HISTORY

Although the average age of the first drink (~15 years) is similar in individuals who do and do not go on to develop alcohol use disorders, an earlier onset of regular drinking and drunkenness, especially in the context of conduct problems, is associated with a higher risk for later alcohol-related diagnoses. By the mid-twenties, most nonalcoholic men and women begin to moderate their drinking (perhaps learning from negative consequences), whereas those with alcohol use disorders are likely to escalate their drinking despite difficulties. The first major life problem from alcohol often appears in the late teens to early twenties, and a pattern of multiple alcohol difficulties by the midtwenties. Once established, the course is likely to include exacerbations and remissions, with little difficulty in temporarily stopping or controlling alcohol use when problems develop, but without help, desistance usually gives way to escalations in alcohol intake and subsequent problems. Following treatment, between half and two-thirds of those with alcohol use disorders maintain abstinence for at least a year, and often permanently. Even without formal treatment or self-help groups, there is at least a 20% chance of spontaneous remission with long-term abstinence. However, should the individual continue to drink heavily, the life span is shortened by ~10 years on average, with the leading causes of early death being enhanced rates of heart disease, cancer, accidents, and suicide.

TREATMENT

The approach to treating alcohol-related conditions is relatively straightforward: (1) recognize that at least 20% of patients have an alcohol use disorder; (2) learn how to identify and treat acute alcohol-related conditions (e.g., severe intoxication); (3) know how to help patients begin to address their alcohol problems; and (4) know how to treat alcohol withdrawal symptoms and to appropriately refer patients for additional help.

IDENTIFICATION OF PATIENTS WITH ALCOHOL USE DISORDERS

Even in affluent locales, the ~20% of patients who have an alcohol use disorder can be identified by asking questions about *alcohol problems* and noting laboratory test results that can reflect regular consumption of six to eight or more drinks per day. The two blood tests with ≥60% sensitivity and specificity for heavy alcohol consumption are γ-glutamyl transferase (GGT) (>35 U) and carbohydrate-deficient transferrin (CDT) (>20 U/L or >2.6%); the combination of the two tests is likely to be more accurate than either alone. The values for these serologic markers are likely to return toward normal within several weeks of abstinence. Other useful blood tests include high-normal MCVs (≥91 μm^3) and serum uric acid (>416 mol/L, or 7 mg/dL).

The diagnosis of an alcohol use disorder ultimately rests on the documentation of a pattern of repeated difficulties associated with alcohol (Table 445-2). Thus, in screening, it is important to probe for marital or job problems, legal difficulties, histories of accidents, medical problems, evidence of tolerance, and so on, and then attempt to relate these issues to use of alcohol. Some standardized questionnaires can be helpful, including the 10-item Alcohol Use Disorders Identification Test (AUDIT) **(Table 445-3)**, but these are only screening tools, and a face-to-face interview is still required for a meaningful diagnosis.

TABLE 445-3

The Alcohol Use Disorders Identification Test (AUDIT)[a]

Item	5-Point Scale (Least to Most)
1. How often do you have a drink containing alcohol?	Never (0) to 4+ per week (4)
2. How many drinks containing alcohol do you have on a typical day?	1 or 2 (0) to 10+ (4)
3. How often do you have six or more drinks on one occasion?	Never (0) to daily or almost daily (4)
4. How often during the last year have you found that you were not able to stop drinking once you had started?	Never (0) to daily or almost daily (4)
5. How often during the last year have you failed to do what was normally expected from you because of drinking?	Never (0) to daily or almost daily (4)
6. How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session?	Never (0) to daily or almost daily (4)
7. How often during the last year have you had a feeling of guilt or remorse after drinking?	Never (0) to daily or almost daily (4)
8. How often during the last year have you been unable to remember what happened the night before because you had been drinking?	Never (0) to daily or almost daily (4)
9. Have you or someone else been injured as a result of your drinking?	No (0) to yes, during the last year (4)
10. Has a relative, friend, doctor, or other health worker been concerned about your drinking or suggested that you should cut down?	No (0) to yes, during the last year (4)

[a]The AUDIT is scored by simply summing the values associated with the endorsed response. A score ≥8 may indicate harmful alcohol use.

TREATMENT

TREATMENT

Alcohol-Related Conditions

ACUTE INTOXICATION

The first priority in treating severe intoxication is to assess vital signs and manage respiratory depression, cardiac arrhythmias, and blood pressure instability, if present. The possibility of intoxication with other drugs should be considered by obtaining, if needed, toxicology screens for other central nervous system (CNS) depressants such as benzodiazepines and for opioids. Aggressive behavior should be handled by offering reassurance but also by considering a possible show of force with an intervention team. If the aggressive behavior continues, relatively low doses of a short-acting benzodiazepine such as lorazepam (e.g., 1–2 mg PO or IV) may be used and can be repeated as needed, but care must be taken not to destabilize vital signs or worsen confusion. An alternative approach is to use an antipsychotic medication (e.g., 0.5–5 mg of haloperidol PO or IM every 4–8 h as needed, or olanzapine 2.5–10 mg IM repeated at 2 and 6 h, if needed).

INTERVENTION

There are two main elements to highlighting the need for compliance with treatment in a person with an alcohol use disorder: motivational interviewing and brief interventions. During motivational interviewing, the clinician helps the patient to think through the assets (e.g., comfort in social situations) and liabilities (e.g., health- and interpersonal-related problems) of the current pattern of drinking. The patient's responses are key, and the clinician should listen empathetically, helping patients to weigh options and encouraging them to take responsibility for needed changes. Patients should be reminded that only they can decide to avoid the consequences that will occur without changes in drinking. The process of the similar approach, brief intervention, has been summarized by the acronym FRAMES: *F*eedback to the patient; *R*esponsibility to be taken by the patient; *A*dvice, rather than orders, on what needs to be done; *M*enus of options that might be considered; *E*mpathy for understanding the patient's thoughts and feelings; and *S*elf-efficacy, i.e., offering support for the capacity of the patient to make changes.

Once the patient begins to consider change, the discussions can focus more on the consequences of high alcohol consumption, suggested approaches to stopping drinking, and help in recognizing and avoiding situations likely to lead to heavy drinking. Both motivational interviewing and brief interventions can be carried out in 15-min sessions, but because patients often do not change behavior immediately, multiple meetings are often required to explore the problem and possible options, discuss optimal treatments, and explain the benefits of abstinence.

ALCOHOL WITHDRAWAL

If the patient agrees to stop drinking, sudden decreases in alcohol intake can produce withdrawal symptoms, most of which are the opposite of those produced by intoxication. Features include tremor of the hands (shakes); agitation and anxiety; autonomic nervous system overactivity including an increase in pulse, respiratory rate, sweating, and body temperature; and insomnia. These symptoms usually begin within 5–10 h of decreasing ethanol intake, peak on day 2 or 3, and improve by day 4 or 5, although mild levels of these problems may persist for 4–6 months as a protracted abstinence syndrome.

About 2% of individuals with alcohol use disorders experience a withdrawal seizure, with the risk increasing in the context of older age, concomitant medical problems, misuse of additional drugs, and higher alcohol quantities. The same risk factors also contribute to the even lower rate of withdrawal delirium, also known as *delirium tremens* (DTs), where the withdrawal includes delirium (mental confusion, agitation, and fluctuating levels of consciousness) associated with a tremor and autonomic overactivity (e.g., marked increases in pulse, blood pressure, and respirations). The risks for seizures and DTs can be diminished by identifying and treating any underlying medical conditions early in the course of withdrawal.

Thus, the first step in dealing with possible withdrawal phenomena is a thorough physical examination in all very heavy drinkers who are considering abstinence. This includes a search for evidence of liver failure, gastrointestinal bleeding, cardiac arrhythmias, infection, and glucose or electrolyte imbalances. It is also important to offer adequate nutrition and oral multiple B vitamins, including 50–100 mg of oral thiamine daily for a week or more. Because most patients with alcohol use disorders who enter withdrawal are either normally hydrated or mildly overhydrated, IV fluids should be avoided unless there is a relevant medical problem or significant recent bleeding, vomiting, or diarrhea.

The next step is to recognize that because withdrawal symptoms reflect the rapid removal of a CNS depressant, alcohol, the symptoms can be controlled by administering any depressant in doses that decrease symptoms (e.g., a rapid pulse and tremor) and then tapering the dose over 3–5 days. Although most depressants are effective, benzodiazepines **(Chap. 444)** have the most supportive data for use in this situation, highest margin of safety and lowest cost and are, therefore, the preferred class of drugs. Short-half-life benzodiazepines can be considered for patients with serious liver impairment or evidence of significant brain damage, but they must be given every 4 h to avoid abrupt blood-level fluctuations that may increase the risk for seizures. Therefore, most clinicians use drugs with longer half-lives (e.g., chlordiazepoxide), adjusting the dose if signs of withdrawal escalate, and withholding the drug if the patient is sleeping or has orthostatic hypotension. The average patient requires 25–50 mg of chlordiazepoxide or 10 mg of diazepam given PO every 4–6 h on the first day, with doses then decreased to zero over the next 5 days. Although alcohol withdrawal can be treated in a hospital, patients in good physical condition who demonstrate mild signs of withdrawal despite low blood alcohol concentrations and who have no prior history of DTs or withdrawal seizures can be considered for outpatient detoxification. For the next 4 or 5 days, these patients should receive only 1 or 2 days of medications at a time, and return daily for evaluation of vital signs. They can be hospitalized if signs and symptoms of withdrawal markedly escalate.

Treatment of patients with DTs can be challenging, and the condition is likely to run a course of 3–5 days regardless of the therapy used. However, conditions that meet the criteria for DTs outlined above represent medical emergencies that carry an estimated mortality as high as 5%, and treatment is best carried out in an intensive care unit by well-trained clinicians who closely monitor vital signs. Medications can include high-dose benzodiazepines (e.g., as much as 800 mg/d of chlordiazepoxide has been reported), or, for those who do not respond to that regimen, closely monitored doses of propofol or dexmedetomidine. The focus of care is to identify and correct medical problems and to control behavior and prevent injuries. We do not recommend the use of antipsychotic medications in the treatment of alcohol withdrawal symptoms;

although antipsychotics are less likely than benzodiazepines to exacerbate confusion, they may increase the risk of seizures.

Generalized withdrawal seizures rarely require more than the administration of an adequate dose of benzodiazepines. There is little evidence that anticonvulsants such as phenytoin or gabapentin are more effective than benzodiazepines for alcohol-withdrawal seizures, and the risk of seizures has usually passed by the time effective drug levels are reached. The rare patient with status epilepticus must be treated aggressively **(Chap. 418)**.

HELPING INDIVIDUALS WITH ALCOHOL USE DISORDERS TO STOP DRINKING: THE REHABILITATION PHASE

An Overview

After completing alcoholic rehabilitation, ≥60% of individuals with alcohol use disorders, especially highly functioning patients, maintain abstinence for at least a year; many also achieve long-term sobriety. The core components of the rehabilitation phase of treatment include cognitive-behavioral approaches to help patients recognize the need to change, while working with them to alter their behaviors to enhance compliance. A key step is to optimize motivation toward abstinence through education of patients and their significant others about alcohol use disorders and their likely course over time. After years of heavy drinking, many patients also require vocational or avocational counseling to help to structure their days, and all patients should try self-help groups such as Alcoholics Anonymous (AA) to assist them in developing a sober peer group and to learn how to deal with life's stresses while remaining sober. *Relapse prevention education* helps patients identify situations in which a return to drinking is likely (e.g., spending time with heavily drinking friends or stopping in a bar to meet friends but planning to only have a nonalcoholic beverage), formulate ways to avoid the risky situation and if not possible to mitigate the risks to which they are exposed. It is also important to develop coping strategies that increase the chances of a return to abstinence quickly after an episode of drinking.

Although many patients can be treated as outpatients, more intense interventions are more effective, and some individuals with alcohol use disorders do not respond to AA or outpatient groups. Whatever the setting, ongoing contact with outpatient treatment staff should be maintained for at least 6 months and preferably for a year after abstinence. Counseling focuses on areas of improved functioning in the absence of alcohol (i.e., why it is a good idea to continue abstinence), helping patients to manage free time without alcohol, encouraging them to develop a nondrinking peer group, and discussions of ways to handle stress without drinking.

The physician serves an important role in identifying the alcohol problem, diagnosing and treating associated medical and independent or substance-induced psychiatric syndromes, overseeing detoxification, referring the patient to outpatient or inpatient rehabilitation programs, providing counseling, and, if appropriate, selecting which (if any) medication might be needed. For insomnia, patients should be reassured that troubled sleep is temporary after alcohol withdrawal and will begin to improve over subsequent weeks. They should be taught the elements of "sleep hygiene" including maintaining consistent schedules for bedtime and awakening, avoiding exercising or eating large meals before bedtime, and keeping the bedroom cool, dark, and quiet at night **(Chap. 27)**. Depressant sleep medications are not the

optimal approach for this type of insomnia that often continues for several weeks or months. Patients are likely to develop rebound insomnia when the depressant dose is decreased or stopped. The rebound increases the chance they will increase the dose and potentially develop problems controlling the prescribed depressant drug. Sedating antidepressants (e.g., trazodone) should not be used because they interfere with cognitive functioning the next morning and disturb the normal sleep architecture, but occasional use of over-the-counter sleeping medications (sedating antihistamines) can be considered. An additional problem, anxiety symptoms, can be addressed by increasing patients' insights into the temporary nature of the symptoms and helping them develop strategies to achieve relaxation by using forms of cognitive therapy.

Medications for the Alcohol Rehabilitation Treatment Phase

Several medications have modest benefits when used in the first 6–12 months of recovery. The opioid antagonist, naltrexone, may shorten subsequent relapses, whether used in the oral form (50–150 mg/d) or as a once-per-month 380-mg injection. By blocking opioid receptors, naltrexone decreases activity in the dopamine-rich ventral tegmental reward system and decreases the feeling of pleasure if alcohol is imbibed. A second medication, acamprosate (Campral) (~2 g/d divided into three oral doses), has similar modest effects. Acamprosate inhibits NMDA receptors, decreasing mild symptoms of protracted withdrawal. Several trials of combined naltrexone and acamprosate have reported that the combination is well tolerated and the efficacy might be superior to either drug alone, although not all studies agree.

It is more difficult to establish the asset-to-liability ratio of a third drug, disulfiram, an ALDH inhibitor, used clinically at doses of 250 mg/d. This drug produces vomiting and autonomic nervous system instability in the presence of alcohol as a result of rapidly rising blood levels of acetaldehyde. This reaction can be dangerous, especially for patients with heart disease, stroke, diabetes mellitus, or hypertension. The drug itself carries potential risks of temporary depressive or psychotic symptoms, peripheral neuropathy, and liver damage. Disulfiram is best given under supervision by someone (such as a spouse), especially during high-risk drinking situations (such as the Christmas holidays). Additional drugs under investigation include another opioid antagonist nalmefene, the nicotinic receptor agonist varenicline, the serotonin antagonist ondansetron, the α-adrenergic agonist prazosin, the $GABA_B$ receptor agonist baclofen, the anticonvulsant topiramate, and cannabinol receptor antagonists. At present, there are insufficient data to determine the asset-to-liability ratio for these medications in treating alcohol use disorders and, therefore, few data yet offer solid support for their routine use in clinical settings.

GLOBAL CONSIDERATIONS



As described above, rates of alcohol use disorders differ across sex, age, ethnicity, and country. There are also differences across countries regarding the definition of a standard drink (e.g., 10–12 g of ethanol in the United States and 8 g in the United Kingdom) and the definition of being legally drunk. The preferred alcoholic beverage also varies across groups, even within countries. That said, regardless of sex, ethnicity, or country, the actual drug in the drink is still ethanol, and the risks for problems, course of alcohol use disorders, and approaches to treatment are similar across the world.

FURTHER READING

Grant BF et al: Prevalence of 12-month alcohol use, high-risk drinking, and DSM-IV alcohol use disorder in the United States, 2001-2002 to 2012-2013: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. JAMA Psychiatry 74:911, 2017. [PubMed: 28793133]

Heikkinen N et al: Alcohol consumption during adolescence is associated with reduced grey matter volumes. Addiction 112:604, 2017. [PubMed: 27865039]

Litten RZ et al: Potential medications for the treatment of alcohol use disorder: An evaluation of clinical efficacy and safety. Subst Abuse 37:286, 2016.

Reilly MT et al: Genetic studies of alcohol dependence in the context of the addiction cycle. Neuropsychopharmacology 122:3, 2017.

Schuckit MA: Recognition and management of withdrawal delirium (delirium tremens). N Engl J Med 371:2109, 2014. [PubMed: 25427113]

Wetherill RR, Fromme K: Alcohol-induced blackouts: A review of recent clinical research with practical implications and recommendations for future studies. Alcohol Clin Exp Res 40:922, 2016. [PubMed: 27060868]

Chapter 446: Opioid-Related Disorders

Thomas R. Kosten; Colin N. Haile

FIGURE 446-1

INTRODUCTION

Opioid analgesics have been used since at least 300 B.C. Nepenthe (Greek “free from sorrow”) helped the hero of the *Odyssey*, but widespread opium smoking in China and the Near East has caused harm for centuries. Since the first chemical isolation of opium and codeine 200 years ago, a wide range of synthetic opioids have been developed, and opioid receptors were cloned in the 1990s. Two of the most important adverse effects of all these agents are the development of opioid use disorder and overdose. Prescription opioids are primarily used for pain management, but due to ease of availability individuals procure and misuse these drugs with dire consequences. In 2015, for example, 3.8 million individuals in the United States were current misusers of pain relievers. More concerning, during 2015 >20,000 overdose deaths involved opioids with an additional 12,990 overdose deaths related to heroin alone. These numbers continue to increase and have accelerated due to mixing high potency fentanyl derivatives with heroin. The accelerating death rates are partially because reversal of fentanyl overdoses can require several-fold larger doses of naloxone than the doses in the intranasal devices used for nonmedical street resuscitations. Indeed, according to the most recent World Drug Report, opioid misuse causes the greatest global burden of morbidity and mortality; disease transmission; increased health care, crime, and law enforcement costs; and less tangible costs of family distress and lost productivity.

The terms “dependence” and “addiction” are no longer used to describe substance use disorders. Opioid-related disorders encompass opioid use disorder, opioid intoxication, and opioid withdrawal. The diagnosis of opioid use disorder as defined in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5) requires the repeated use of the opiate while producing problems in two or more areas in a 12-month period. The areas include tolerance, withdrawal, use of greater amounts of opioids than intended, craving, and use despite adverse consequences. This new definition of opioid use disorder, reducing the criteria for diagnosis from three problem areas to two, is not expected to change the rates of these disorders because most individuals using these substances meet more than three criteria.

A striking recent aspect of illicit opioid use has been its marked increase as the gateway to illicit drugs in the United States. Since 2007, prescription opiates have surpassed marijuana as the most common illicit drug that adolescents initially use, although overall rates of opioid use are far lower than marijuana. The most commonly used opioids are diverted prescriptions for oxycodone and hydrocodone, followed by heroin and morphine, and—among health professionals—meperidine and fentanyl. Heroin is metabolized into 6-monoacetylmorphine and morphine thus acting as a prodrug that more readily penetrates the brain and is converted rapidly to morphine in the body. Two opioid maintenance treatment agents—methadone and buprenorphine—are also misused, but at substantially lower rates, and the partial opioid agonists such as butorphanol, tramadol, and pentazocine are misused even less frequently. Because the chemistry and general pharmacology of these agents are covered in major pharmacology texts, this chapter focuses on the neurobiology and pharmacology relevant to opioid use disorder and its treatments. Although the neurobiology of misuse involves all four of the known opioid receptors—mu, kappa, delta, and nociceptin/orphanin—this discussion focuses on the mu receptor targeted by most of the clinically used opioids.

NEUROBIOLOGY

The neurobiology of opioids and their effects not only include opioid receptors, but also downstream intracellular messenger systems and ion channels that the receptors regulate. The different functional activities of opioid receptors are summarized in **Table 446-1**. Abuse liability of opioids is primarily associated with the mu receptor. All opioid receptors are G protein–linked and coupled to the cyclic adenosine monophosphate (cAMP) second messenger system and to G protein–coupled, inwardly rectifying potassium channels (GIRKs). Opioids activate GIRKs, increasing permeability to potassium ions to cause hyperpolarization, which inhibits the production of action potentials. Thus, opioids inhibit the activity of diverse and widely distributed neuronal types. The major effects of opioids, such as analgesia, sedation, and drug reinforcement, are produced through this inhibition of neurons that belong to specific brain pathways.

TABLE 446-1

Actions of Opioid Receptors

Receptor Type	Actions
Mu (μ) (e.g., morphine, buprenorphine)	Analgesia, reinforcement euphoria, cough and appetite suppression, decreased respirations, decreased GI motility, sedation, hormone changes, dopamine and acetylcholine release
Kappa (κ) (e.g., butorphanol)	Dysphoria, decreased GI motility, decreased appetite, decreased respiration, psychotic symptoms, sedation, diuresis, and analgesia
Delta (δ) (e.g., etorphine)	Analgesia, euphoria, physical dependence, hormone changes, appetite suppression, and dopamine release
Nociceptin/orphanin (e.g., buprenorphine)	Analgesia, appetite, anxiety, tolerance to opioids, hypotension, decreased GI motility, and 5-HT and NE release

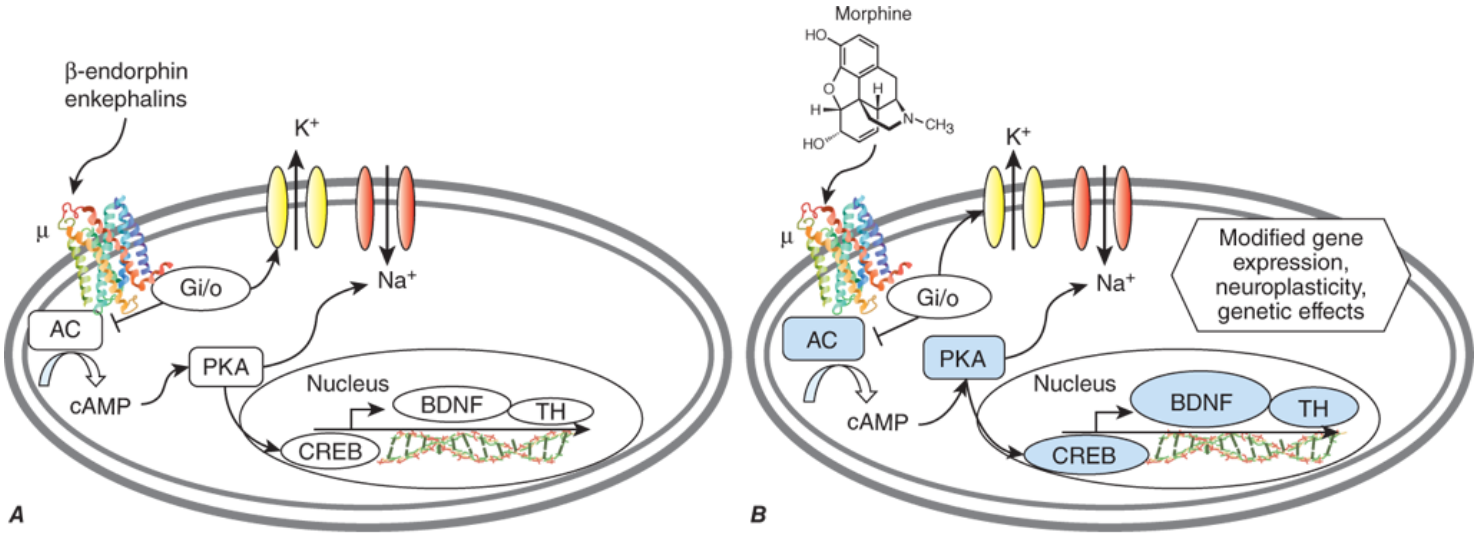
Abbreviations: 5-HT, serotonin; GI, gastrointestinal; NE, norepinephrine.

Many opioid actions are related to the specific neuroanatomic locations of mu receptors. Reinforcing and euphoric effects of opioids relate primarily to activation of the mesolimbic dopaminergic pathway from the ventral tegmental area (VTA) to the nucleus accumbens (NAc), where opioids increase synaptic levels of dopamine. This increase is due to inhibition of GABAergic neurons that inhibit both the activity of neurons within the VTA and the NAc. The positive subjective effects of opioid drugs also include mu receptor desensitization and internalization, potentially related to stimulation of beta-arrestin signaling pathways. However, the "high" only occurs when the *rate of change* in dopamine is fast. Large, rapidly administered doses of opioids block γ-aminobutyric acid (GABA) inhibition and produce a burst of VTA dopamine neuron activity that is associated with a "high" in commonly misused substances. Therefore, routes of administration that slowly increase opioid blood and brain levels, such as oral and transdermal routes, are effective for analgesia and sedation but do not produce an opioid "high" that follows smoking and intravenous routes. Other acute effects such as analgesia and respiratory depression involve opioid receptors located in other brain areas such as the locus coeruleus (LC).

Opioid tolerance and withdrawal are chronic effects related to the cAMP-protein kinase A (PKA)-cAMP response-element binding protein (CREB) intracellular cascade **(Fig. 446-1)**. These effects are also reflective of genetic risk factors for developing opioid use disorder, with estimates of up to 50% of the risk due to polygenic inheritance. Specific functional polymorphisms in the mu opiate receptor gene appear to be associated with this risk for opioid misuse, including one producing a threefold increase in this receptor's affinity for opiates and the endogenous ligand β-endorphin. Epigenetic methylation changes also occur on DNA in the region of the mu receptor gene in individuals with opioid use disorder, inhibiting gene transcription. This molecular cascade links acute intoxication and sedation to opioid tolerance and withdrawal mediated by the LC. Noradrenergic (NE) neurons in the LC mediate activation of the cortical hemispheres. When large opioid doses saturate and activate all of its mu receptors, action potentials cease. When this direct inhibitory effect is sustained over weeks and months of opioid use, a secondary set of adaptive changes occur that lead to tolerance and withdrawal symptoms (Fig. 446-1). Withdrawal symptoms reflect, in part, overactivity of NE neurons in the LC. This molecular model of NE neuronal activation during withdrawal has had important treatment implications, such as the use of the alpha-2 agonist clonidine to treat opioid withdrawal. Other contributors to withdrawal include deficits within the dopamine reward system.

FIGURE 446-1

Normal mu-receptor activation by endogenous opioids inhibits the cyclic adenosine monophosphate (cAMP)-protein kinase A (PKA)-cAMP response-element binding protein (CREB) cascade in noradrenergic (NE) neurons within the locus coeruleus (***A***) through inhibitory Gi/o protein influence on adenylyl cyclase (AC). Similarly, acute exposure to opioids (e.g., morphine) inhibits this system, whereas chronic exposure to opiates (***B***) leads to upregulation of the cAMP pathway in an attempt to oppose opioid-induced inhibitory influence. Upregulation of this system is involved in opioid tolerance, and when the opioid is removed, unopposed NE neurotransmission is involved in opioid withdrawal. Upregulated PKA phosphorylates CREB, initiating the expression of various genes such as tyrosine hydroxylase (*TH*) and brain-derived neurotrophic factor (*BDNF*). *BDNF* is implicated in long-term neuroplastic changes in response to chronic opioids.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

PHARMACOLOGY

Tolerance and withdrawal commonly occur with chronic daily use, developing as quickly as 6–8 weeks depending on dose concentration and dosing frequency. Tolerance appears to be primarily a pharmacodynamic rather than pharmacokinetic effect, with relatively limited induction of cytochrome P450 or other liver enzymes. The metabolism of opioids occurs in the liver, primarily through the cytochrome P450 systems of 2D6 and 3A4. They then are conjugated to glucuronic acid and excreted in small amounts in feces. The plasma half-lives generally range from 2.5 to 3 h for morphine and >22 h for methadone. The shortest half-lives of several minutes are for fentanyl-related opioids and the longest are for buprenorphine and its active metabolites, which can block opioid withdrawal for up to 3 days after a single dose. Tolerance to opioids leads to the need for increasing amounts of drugs to sustain the desired euphoric effects—as well as to avoid the discomfort of withdrawal. This combination has the expected consequence of strongly reinforcing misuse once it has started. Methadone taken chronically at maintenance doses is stored in the liver, which may reduce the occurrence of withdrawal between daily doses. The role of endogenous opioid peptides in tolerance and withdrawal is uncertain.

The clinical features of opioid misuse are tied to route of administration and rapidity of the drug reaching the brain. Intravenous and smoked administration rapidly produces high drug concentrations in the brain. This produces a “rush,” followed by euphoria, a feeling of tranquility, and sleepiness (“the nod”). Heroin produces effects that last 3–5 h, and several doses a day are required to forestall manifestations of withdrawal in chronic users. Symptoms of opioid withdrawal begin 8–10 h after the last dose; lacrimation, rhinorrhea, yawning, and sweating appear first. Restless sleep followed by weakness, chills, gooseflesh (“cold turkey”), nausea and vomiting, muscle aches, and involuntary movements (“kicking the habit”), hyperpnea, hyperthermia, and hypertension occur in later stages of the withdrawal syndrome. The acute course of withdrawal may last 7–10 days. A secondary phase of protracted abstinence lasts for 26–30 weeks and is characterized by hypotension, bradycardia, hypothermia, mydriasis, and decreased responsiveness of the respiratory center to carbon dioxide.

Besides the brain effects of opioids on sedation and euphoria and the combined brain and peripheral nervous system effects on analgesia, a wide range of other organs can be affected. The release of several pituitary hormones is inhibited, including corticotropin-releasing factor (CRF) and luteinizing hormone, which reduces levels of cortisol and sex hormones and can lead to impaired stress responses and reduced libido. An increase in prolactin also contributes to the reduced sex drive in males. Two other hormones affected are thyrotropin, which is reduced, and growth hormone, which is increased. Respiratory depression results from opioid-induced insensitivity of brainstem neurons to increases in carbon dioxide, and in patients with pulmonary disease, this can result in clinically significant complications. In overdoses, aspiration pneumonia is commonly due to loss of the gag reflex. Opioids reduce gut motility, which is helpful for treating diarrhea, but can lead to nausea, constipation, and anorexia with weight loss. Deaths occurred in early methadone maintenance programs due to severe constipation and toxic megacolon. Opioids such as methadone may prolong QT intervals and lead to sudden death in some patients. Orthostatic hypotension may occur due to histamine release and peripheral blood vessel dilation, which is an opioid effect usefully applied to managing acute myocardial infarction. During opioid maintenance, interactions with other medications are of concern; these include inducers of the cytochrome P450 system (usually CYP3A4) such as rifampin and carbamazepine.

Heroin users in particular tend to use opioids intravenously and are likely to be polydrug users, also using alcohol, sedatives, cannabinoids, and stimulants. None of these other drugs are substitutes for opioids, but they have desired additive effects. Therefore, one needs to be sure that the person undergoing a withdrawal reaction is not also withdrawing from alcohol or sedatives, which might be more dangerous and more difficult to manage.

Intravenous opioid use carries with it the risk of serious complications. The common sharing of hypodermic syringes can lead to infections with hepatitis B and HIV/AIDS, among others. Bacterial infections can lead to septic complications such as meningitis, osteomyelitis, and abscesses in various organs.

Off-target effects of opioids synthesized in illicit drug labs can lead to serious toxicity. For example, attempts to illicitly manufacture meperidine in the 1980s resulted in the production of a highly specific neurotoxin, MPTP, which produced parkinsonism in users **(Chap. 427)**.

Lethal overdose is a relatively common complication of opioid use disorder. Rapid recognition and treatment with naloxone, a highly specific reversal agent that is relatively free of complications, is essential. The diagnosis is based on recognition of characteristic signs and symptoms, including shallow and slow respirations, pupillary miosis (mydriasis does not occur until significant brain anoxia supervenes), bradycardia, hypothermia, and stupor or coma. Blood or urine toxicology studies can confirm a suspected diagnosis, but immediate management must be based on clinical criteria. If naloxone is not administered, progression to respiratory and cardiovascular collapse leading to death occurs. At autopsy, cerebral edema and sometimes frothy pulmonary edema are generally found. Opioids generally do not produce seizures except for unusual cases of polydrug use with the opioid meperidine, with high doses of tramadol, or in the newborn.

TREATMENT

TREATMENT

Opioid Overdose

Beyond the acute treatment of opioid overdose with naloxone, clinicians have two general treatment options: opioid maintenance or detoxification. Opioid agonist and partial agonist medications are commonly used for both maintenance and detoxification purposes. Alpha-2-adrenergic agonists are primarily used for detoxification. Antagonists are used to accelerate detoxification and then continued after detoxification to prevent relapse. Only the residential medication-free programs have had success that comes close to matching that of the medication-based programs. Success of the various treatment approaches is assessed as retention in treatment and reduced opioid and other drug use; secondary outcomes, such as reduced HIV risk behaviors, crime, psychiatric symptoms, and medical comorbidity, also indicate successful treatment.

Stopping opioid use is much easier than preventing relapse. Long-term relapse prevention for individuals with opioid use disorder requires combined pharmacologic and psychosocial approaches. Chronic users tend to prefer pharmacologic approaches; those with shorter histories of drug use are more amenable to detoxification and psychosocial interventions.

OPIOID OVERDOSE

Managing overdose requires naloxone and support of vital functions, including intubation if needed **(Table 446-2)**. If the overdose is due to buprenorphine, then naloxone might be required at total doses of ≥10 mg, but primary buprenorphine overdose is nearly impossible because this agent is a partial opioid agonist, meaning that as the dose of buprenorphine is increased it has greater opioid antagonist than agonist activity. Thus, a 0.2-mg buprenorphine dose leads to analgesia and sedation, while a hundred times greater 20-mg dose produces profound opioid antagonism, precipitating opioid withdrawal in a person who had opioid use disorder on morphine or methadone. It is important to recognize that the goal is to reverse the respiratory depression and not to administer so much naloxone that it precipitates opiate withdrawal. Because naloxone only lasts a few hours and most opioids last considerably longer, an IV naloxone drip with close monitoring is frequently employed to provide a continuous level of antagonism for 24–72 h depending on the opioid used in the overdose (e.g., morphine vs methadone). Whenever naloxone has only a limited effect, other sedative drugs that produce significant overdoses must be considered. The most common are benzodiazepines, which have produced overdoses and deaths in combination with buprenorphine. A specific antagonist for benzodiazepines—flumazenil at 0.2 mg/min—can be given to a maximum of 3 g/h, but it may precipitate seizures and increase intracranial pressure. Like naloxone, administration for a prolonged period is usually required because most benzodiazepines remain active for considerably longer than flumazenil. Support of vital functions may include oxygen and positive-pressure breathing, IV fluids, pressor agents for hypotension, and cardiac monitoring to detect QT prolongation, which might require specific treatment. Activated charcoal and gastric lavage may be helpful for oral ingestions, but intubation will be needed if the patient is stuporous.

OPIOID WITHDRAWAL

The principles of detoxification are the same for all drugs: to substitute a longer-acting, orally active, pharmacologically equivalent medication for the substance being used, stabilize the patient on that medication, and then gradually withdraw the substituted medication. Methadone and buprenorphine are the two medications used to treat opioid use disorder. Clonidine, a centrally acting sympatholytic agent, has also been used for detoxification in the United States. By reducing central sympathetic outflow, clonidine mitigates many of the signs of sympathetic overactivity but typically requires augmentation with other agents. Clonidine has no narcotic action and is not addictive. Lofexidine, a clonidine analogue with less hypotensive effect, is not yet approved in the United States.

Methadone for Detoxification

Dose-tapering regimens for detoxification using methadone range from 2 to 3 weeks to as long as 180 days, but this approach is controversial given the relative effectiveness of methadone maintenance and the low success rates of detoxification. Unfortunately, the vast majority of patients tend to relapse to heroin or other opioids during or after the detoxification period, indicative of the chronic and relapsing nature of opioid use disorder.

Buprenorphine for Detoxification

Buprenorphine does not appear to lead to better outcomes than methadone but is superior to clonidine in reducing symptoms of withdrawal, retaining patients in a withdrawal protocol, and in completing treatment.

Alpha-2-Adrenergic Agonists for Detoxification

Several alpha-2-adrenergic agonists have relieved opioid withdrawal by suppressing brain NE hyperactivity. Clonidine relieves some signs and symptoms of opioid withdrawal such as lacrimation, rhinorrhea, muscle pain, joint pain, restlessness, and gastrointestinal symptoms. Related agents are lofexidine, guanfacine, and guanabenz acetate. Lofexidine can be dosed up to ~2 mg/d and appears to be associated with fewer adverse effects. Clonidine or lofexidine is typically administered orally, in three or four doses per day, with dizziness, sedation, lethargy, and dry mouth as the primary adverse side effects. Outpatient-managed withdrawal will require close follow-up often with naltrexone maintenance to prevent relapse.

Rapid and Ultrarapid Opioid Detoxification

The opioid antagonist naltrexone typically combined with an alpha-2-adrenergic agonist has been purported to shorten the duration of withdrawal without significantly increasing patient discomfort. Completion rates using naltrexone and clonidine range from 75 to 81% compared to 40 to 65% for methadone or clonidine alone. Ultrarapid opioid detoxification is an extension of this approach using anesthetics but is highly controversial due to the medical risks and mortality associated with it.

TABLE 446-2

Management of Opioid Overdose

Establish airway. Intubation and mechanical ventilation may be necessary.
Naloxone 0.4–2.0 mg (IV, IM, or endotracheal tube). Onset of action with IV is ~1–2 min.
Repeat doses of naloxone if needed to restore adequate respiration or a continuous infusion of naloxone can be used.
One-half to two-thirds of the initial naloxone dose that reversed the respiratory depression is administered on an hourly basis (note: naloxone dosing is not necessary if the patient has been intubated).

Opioid Agonist Medications for Maintenance

Methadone maintenance substitutes a once-daily oral opioid dose for three- to four-times daily heroin. Methadone saturates the opioid receptors and, by inducing a high level of opioid tolerance, blocks the euphoria from additional opioids. Buprenorphine, a partial opioid agonist, also can be given once daily at sublingual doses of 4–32 mg daily, and in contrast to methadone, it can be given in an office-based primary care setting.

METHADONE MAINTENANCE

Methadone's slow onset of action when taken orally, long elimination half-life (24–36 h), and production of cross-tolerance at doses from 80 to 150 mg are the basis for its efficacy in treatment retention and reductions in IV drug use, criminal activity, and HIV risk behaviors and mortality. Methadone can prolong the QT interval at rates as high as 16% above the rates in nonmethadone-maintained, drug-injecting patients, but it has been used safely in the treatment of opioid use disorder for 40 years.

BUPRENORPHINE MAINTENANCE

While France and Australia have had sublingual buprenorphine maintenance since 1996, it was first approved by the U.S. Food and Drug Administration (FDA) in 2002 as a Schedule III drug for managing opioid use disorder. Unlike the full agonist methadone, buprenorphine is a partial agonist of mu-opioid receptors with a slow onset and long duration of action. Its partial agonism reduces the risk of unintentional overdose but limits its efficacy to patients who need the equivalent of only 60–70 mg of methadone, and many patients in methadone maintenance require higher doses up to 150 mg daily. Buprenorphine is combined with naloxone at a 4:1 ratio in order to reduce its abuse liability. Because of pediatric exposures and diversion of buprenorphine to illicit use, a new formulation, using mucosal films rather than sublingual pills that were crushed and snorted, is now marketed. A subcutaneous buprenorphine implant that lasts up to 6 months has recently been approved by the FDA as a formulation improvement to prevent pediatric exposures and illicit diversion and to enhance compliance.

In the United States, the ability of primary care physicians to prescribe buprenorphine for opioid use disorder represents an important opportunity to improve access and quality of treatment as well as reduce social harm. Europe, Asia, and Australia have found reduced opioid-related deaths and drug-injection-related medical morbidity with buprenorphine available in primary care. Retention in office-based buprenorphine treatment has been high as 70% at 6-month follow-ups.

Opioid Antagonist Medications

The rationale for using narcotic antagonist therapy is that blocking the action of self-administered opioids should eventually extinguish the habit, but this therapy is poorly accepted by patients. Naltrexone, a long-acting orally active pure opioid antagonist, can be given three times a week at doses of 100–150 mg. Because it is an antagonist, the patient must first be detoxified from opioids before starting naltrexone. It is safe even when taken chronically for years, is associated with few side effects (headache, nausea, and abdominal pain), and can be given to patients infected with hepatitis B or C without producing hepatotoxicity. However, most providers refrain from prescribing naltrexone if liver function tests are three times above normal levels. Naltrexone maintenance combined with psychosocial therapy is effective in reducing heroin use, but medication adherence is low. Depot injection formulations lasting up to 4 weeks markedly improve adherence, retention, and drug use. Subcutaneous naltrexone implants in Russia, China, and Australia have doubled treatment retention and reduced relapse to half that of oral naltrexone. In the United States, a depot naltrexone formulation is available for monthly use and maintains blood levels equivalent to 25 mg of daily oral use.

Medication-Free Treatment

Most opioid users enter medication-free treatments in inpatient, residential, or outpatient settings, but 1- to 5-year outcomes are very poor compared to pharmacotherapy except for residential settings lasting 6–18 months. The residential programs require full immersion in a regimented system with progressively increasing levels of independence and responsibility within a controlled community of fellow drug users. These medication-free programs, as well as the pharmacotherapy programs, also include counseling and behavioral treatments designed to teach interpersonal and cognitive skills for coping with stress and for avoiding situations leading to easy access to drugs or to craving. Relapse is prevented by having the individual very gradually reintroduced to greater responsibilities and to the working environment outside of the protected therapeutic community.

PREVENTION

Preventing the development of opioid use disorder represents a critically important challenge for physicians. Opioid prescriptions are the most common source of drugs accessed by adolescents who begin a pattern of illicit drug use. The major sources of these drugs are family members, not drug dealers or the Internet. Pain management involves providing sufficient opioids to relieve the pain over as short a period of time as the pain warrants **(Chap. 10)**. The patient then needs to dispose of any remaining opioids, not save them in the medicine cabinet, because this behavior leads to diversion by adolescents. Finally, physicians should never prescribe opioids for themselves.

FURTHER READING

Brady KT et al: Prescription opioid misuse, abuse, and treatment in the United States: An update. Am J Psychiatry 173:18, 2016. [PubMed: 26337039]

D'onofrio G et al: Emergency department-initiated buprenorphine/naloxone treatment for opioid dependence: A randomized clinical trial. JAMA 313:1636, 2015. [PubMed: 25919527]

Dugosh K et al: A systematic review on the use of psychosocial interventions in conjunction with medications for the treatment of opioid addiction. J Addict Med 10:93, 2016. [PubMed: 26808307]

Trang T et al: Pain and poppies: The good, the bad, and the ugly of opioid analgesics. J Neurosci 35:13879, 2015. [PubMed: 26468188]

Chapter 447: Cocaine and Other Commonly Used Drugs

Karran A. Phillips; Antonello Bonci

INTRODUCTION

The use of cocaine and other psychostimulants reflects a complex interaction between the pharmacology of the drug, the personality and expectations of the user, and the environmental context in which the drug is used. Polydrug use involving the concurrent use of several drugs with different pharmacologic effects is increasingly common. Sometimes one drug is used to enhance the effects of another, as with the combined use of cocaine and nicotine, or cocaine and heroin in methadone-maintained patients. Some forms of polydrug use, such as the combined use of IV heroin and cocaine, are especially dangerous and account for many hospital emergency room visits. Chronic cocaine and psychostimulant use may cause a number of adverse health consequences and may exacerbate preexisting disorders such as hypertension and cardiac disease. In addition, the combined use of two or more drugs may accentuate medical complications associated with use of one drug. Chronic drug use is often associated with immune system dysfunction and increased vulnerability to infections, including risk for HIV infection. The concurrent use of cocaine and opiates (“speedball”) is frequently associated with needle sharing by people using drugs intravenously. People who use IV drugs represent the largest single group of individuals with HIV infection in several major metropolitan areas in the United States as well as in many parts of Europe and Asia.

Stimulants and hallucinogens have been used to induce euphoria and alter consciousness for centuries. Cocaine and marijuana are two of the most commonly used drugs today. Synthetic variations of marijuana and a variety of hallucinogens have become popular recently, and new drugs are continually being developed. This chapter describes the subjective and adverse medical effects of cocaine, other psychostimulants including methamphetamine, 3,4-methylenedioxymethamphetamine (MDMA), and cathinones; as well as hallucinogens such as phencyclidine (PCP), D-lysergic acid diethylamide (LSD), salvia divinorum; and marijuana and the synthetic cannabinoids. Some options for medical management of severe adverse effects are also described.

PSYCHOSTIMULANTS

PHARMACOKINETICS/DYNAMICS, NEUROBIOLOGY, AND EPIDEMIOLOGY

Cocaine

Cocaine is a powerful stimulant drug made from the cocoa plant. It has local anesthetic, vasoconstrictor, and stimulant properties. Cocaine is a Schedule II drug, which means that it has high potential for abuse but can be administered by a physician for legitimate medical uses, such as local anesthesia for some eye, ear, and throat surgeries.

Cocaine comes in a variety of forms, the most commonly used being the hydrochloride salt, sulfate, and a base. The salt is an acidic, water-soluble powder with a high melting point, used by snorting or sniffing intranasally or by dissolving it in water and injecting it intravenously. When used intranasally the bioavailability of cocaine is about 60 per percent. Cocaine sulfate ("paste") has a melting point of almost 200°C, so it has limited use, but is sometimes smoked with tobacco. The base form can be freebase or crystallized as crack. Cocaine freebase is made by adding a strong base to an aqueous solution of cocaine and extracting the alkaline freebase precipitate. It has a melting point of 98°C and can be vaporized and inhaled. Freebase cocaine can also be crystallized and sold as crack or rock, which is also smoked or inhaled. Street dealers often dilute (or "cut") cocaine with nonpsychoactive substances such as cornstarch, talcum powder, flour, or baking soda, or adulterate it with other substances with similar effects (like procaine or amphetamine) to increase their profits.

Given the extensive pulmonary vasculature, smoked or inhaled cocaine reaches the brain very quickly and produces a rapid and intense (yet transient) high, which enhances its addictive potential. Cocaine binds to the dopamine (DA) transporter and blocks DA reuptake, which increases synaptic levels of the monoamine neurotransmitters DA, norepinephrine (NE), and serotonin, in both the central nervous system (CNS) and the peripheral nervous system (PNS). Use of cocaine, like other drugs of abuse, induces long-term changes in the brain. Animal studies have shown adaptations in neurons that release the excitatory neurotransmitter glutamate after cocaine exposure.

According to the National Survey on Drug Use and Health (NSDUH), in 2015 about 1.9 million people (~0.7% of the population) were current users of cocaine, including about 394,000 current users of crack (0.1% of the population in the United States). There were 53,000 adolescents aged 12–17 (0.2% of adolescents) who were current users of cocaine in 2015. About 896,000 people aged ≥12 (0.3% of the population) in 2015 had a cocaine use disorder in the past year. The Drug Abuse Warning Network (DAWN) reported that in 2011 there were 505,224 cocaine-related emergency department (ED) visits, or about 162 ED visits per 100,000 of the U.S. population.

Methamphetamine

Methamphetamine is a stimulant drug usually used as a white, bitter-tasting powder or a pill. Crystal methamphetamine is a form of the drug that looks like glass fragments or shiny, bluish-white rocks. It can be inhaled/smoked, swallowed (pill), snorted, or injected after being dissolved in water or alcohol. When smoked, methamphetamine exhibits 90.3% bioavailability, compared to 67.2% for oral ingestion. Methamphetamine exists in two stereoisomers, the L- and D-forms. D-Methamphetamine, or the dextrorotatory enantiomer, is a more powerful psychostimulant, with 3–5 times the CNS activity as compared to L-methamphetamine. Methamphetamine is a cationic lipophilic molecule which stimulates the

release, and partially blocks the reuptake, of newly synthesized catecholamines in the CNS. Methamphetamine has a similar structure to the DA, NE, serotonin, and vesicular monoamine transporters and reverses their endogenous function, resulting in release of monoamines from storage vesicles into the synapse. Methamphetamine also attenuates the metabolism of monoamines by inhibiting monoamine oxidase.

Methamphetamine is more potent and more efficacious than amphetamine, resulting in much higher concentrations of synaptic DA and more toxic effects on nerve terminals. Outside the medical context, methamphetamine's pharmacokinetics and low cost often result in a chronic and continuous, high dose self-administered use pattern.

According to the NSDUH, ~897,000 people (0.3% of the population) aged ≥12 were current users of methamphetamine in 2015. Meanwhile, about 13,000 adolescents (0.1%) aged 12–17 were current methamphetamine users in 2015. There were also 51.3 ED visits, per 100,000 of the population, related to illicit stimulants (predominately amphetamines and methamphetamine) in 2011.

MDMA and Cathinones

MDMA is an illegal drug that has stimulant and psychedelic effects. With MDMA use, individuals experience increased physical and mental energy, distortions in time and perception, emotional warmth, empathy toward others, a general sense of well-being, decreased anxiety, and an enhanced enjoyment of tactile experience. MDMA is usually taken orally in a tablet, capsule, or liquid form, and its effects last ~3–6 h. MDMA alters brain chemistry by binding to serotonin transporters and increasing the release of serotonin, NE, and DA. Research in animals has shown that MDMA in moderate to high doses can cause loss of serotonin-containing nerve endings and permanent damage. MDMA is a Schedule I drug, along with other substances with no proven therapeutic value. MDMA is currently in clinical trials as a possible treatment for posttraumatic stress disorder and anxiety in terminally ill patients, and for social anxiety in autistic adults.

Adulteration of MDMA tablets with methamphetamine, ketamine, caffeine, the over-the-counter cough suppressant dextromethorphan (DXM), the diet drug ephedrine, and cocaine is common. MDMA is rarely used alone and is often mixed with other substances, such as alcohol and marijuana, making the scope of its use difficult to ascertain. The Monitoring the Future study estimated that, in 2016, the lifetime prevalence of MDMA use among eighth graders was 1.7%, tenth graders was 2.8%, and twelfth graders was 4.9%, with the most use among 18–25 year olds.

Cathinone is an alkaloid psychostimulant found in the khât (Catha edulis) plant, which grows at high altitudes in East Africa and the Middle East. The actions and effects of khât are like those of the amphetamines, and misusers are at increased risk for acute myocardial infarction and stroke, due to inotropic and chronotropic effects on the heart, vasospasm of coronary arteries, and catecholamine-induced platelet aggregation.

Prescribed Psychostimulants

Methylphenidate, amphetamine, and methamphetamine are psychostimulants approved in the United States for treatment of attention-deficit hyperactivity disorder (ADHD), weight control, and narcolepsy. Phenylpropanolamine, a psychostimulant used primarily for weight control, was found to be related to hemorrhagic stroke in women and removed from the market in 2005. These drugs deserve mention here, as there has been increased use of nonprescribed amphetamines or methylphenidate as a study aid among college students, and an energy and productivity booster for so-called “supermoms.” According to the NSDUH, of the 7.7 million people, aged ≥12, who had a past year stimulant use disorder (SUD) related to their use of illicit drugs, 0.4 million misused prescription stimulants.

CLINICAL MANIFESTATIONS

Psychostimulants produce the same acute CNS effects: euphoria, increased energy/decreased fatigue, reduced need for sleep, decreased appetite, decreased distractibility, increased self-confidence and alertness, increased libido, and prolonged orgasm, independent of the specific psychostimulant or route of administration. Peripheral effects may include tremor, diaphoresis, hypertonia, tachypnea, hyperreflexia, and hyperthermia. Many of the effects are biphasic; for example, low doses improve psychomotor performance, while higher doses may cause tremors or convulsions. α-adrenergically mediated cardiovascular effects are also biphasic, with low doses resulting in increased vagal tone and decreased heart rate, and high doses causing increased heart rate and blood pressure. Psychostimulant use can result in restlessness, irritability, and insomnia and, at higher doses, suspiciousness, repetitive stereotyped behaviors, and bruxism. Endocrine effects may include impotence, gynecomastia, menstrual function disruptions, and persistent hyperprolactinemia **(Table 447-1)**.

TABLE 447-1

Complications of Cocaine Use

Cardiovascular	Acute • Arterial vasoconstriction • Thrombosis • Tachycardia • Hypertension • Increased myocardial oxygen demand • Increased vascular shearing forces • Coronary vasoconstriction • Cardiac ischemia • Left ventricular dysfunction/heart failure (high blood concentrations) • Supraventricular and ventricular dysrhythmias • Aortic dissection/rupture Chronic • Accelerated atherogenesis • Left ventricular hypertrophy • Dilated cardiomyopathy
Central and Peripheral Nervous	• Hyperthermia • Psychomotor agitation • Tremor • Hyperreflexia • Hypertonia • Headache • Seizures • Coma • Intracranial hemorrhage • Focal neurologic symptoms
Pulmonary	• Angioedema (inhaled) • Pharyngeal burns (inhaled) • Pneumothorax • Pneumomediastinum • Pneumopericardium • Reversible airway disease exacerbations • Bronchospasm

	• Shortness of breath (“crack lung”) • Tachypnea • Pulmonary infarction
Gastrointestinal	• Perforated ulcers • Ischemic colitis • Bowel infarction • Impaction (body packing) • Hepatic enzyme elevation
Renal	• Metabolic acidosis • Renal infarction • Rhabdomyolysis
Endocrine	• Impotence • Gynecomastia • Menstrual function disruptions • Hyperprolactinemia
Other	• Diaphoresis • Irritability • Insomnia • Bruxism • Stereotypy • Splenic infarction • Acute angle-closure glaucoma • Vasospasm of the retinal vessels (unilateral or bilateral vision loss) • Mydriasis • Madarosis • Abruptio placentae

Overdose presents as sympathetic nervous system overactivity with psychomotor agitation, hypertension, tachycardia, headache, and mydriasis, and can lead to convulsions, cerebral hemorrhage or infarction, cardiac arrhythmias or ischemia, respiratory failure, or rhabdomyolysis. It is a medical emergency; treatment is largely symptomatic and should occur in an intensive care or telemetry unit. Inhalation of crack cocaine that is vaporized at high temperatures can cause airway burns, bronchospasm and other symptoms of

pulmonary disease. MDMA has also been shown to raise body temperature and can occasionally result in liver, kidney, or heart failure, or even death.

Psychostimulants are often used with other drugs, including opioids and alcohol, whose CNS-depressant effects tend to attenuate psychostimulant-induced CNS stimulation. These combinations often have additive deleterious effects, increasing the risk of morbidity and mortality. An example of this risk is the use of cocaine with alcohol, which results in the metabolite, cocaethylene. Cocaethylene's effects on the cardiovascular system are additive to that of cocaine's effects, resulting in intensified pathophysiologic consequences.

Adulteration of psychostimulants, particularly cocaine, with other drugs is common and can have additional potential health consequences. Levamisole, an anthelminthic and immunomodulator used primarily in veterinary medicine, has been found in cocaine and can cause agranulocytosis, leukoencephalopathy, and cutaneous vasculitis, which has resulted in cutaneous necrosis. Clenbuterol, a sympathomimetic amine used clinically as a bronchodilator, has also been found in cocaine and can result in tachycardia, hyperglycemia, palpitations, and hypokalemia. Studies in Europe have found that in addition to levamisole some of the most common adulterants in cocaine include: phenacetin, lidocaine, caffeine, diltiazem, hydroxyzine, procaine, tetracaine, paracetamol, creatine, and benzocaine.

Withdrawal from psychostimulants often includes hypersomnia, increased appetite, and depressed mood. Acute withdrawal typically lasts 7–10 days, but residual symptoms, possibly associated with neurotoxicity, may persist for several months. Psychostimulant withdrawal is not thought to be a driver of ongoing use. Debate remains as to whether, in psychostimulant withdrawal, symptoms decline monotonically or occur in discrete phases, getting worse before they get better. Most current theories of psychostimulant addiction emphasize the primary role of conditioned craving, which can persist long after physiological withdrawal has abated.

Injection of psychostimulants places people at increased risk of contracting infectious diseases from exposure to blood or other bodily fluids, such as HIV and hepatitis B and C. Psychostimulant use can also increase risk for infection by causing altered judgment and decision-making, leading to risky behaviors, such as unprotected sex. There is some evidence that psychostimulant use may worsen the progression of HIV/AIDS via increased injury to nerve cells exacerbating cognitive problems.

SCREENING AND DIAGNOSIS

The *Diagnostic and Statistical Manual of Psychiatric Disorders*, 5th edition (*DSM*-5) defines a SUD as a pattern of use of amphetamine-type substances, cocaine, or other stimulants leading to clinically significant impairment or distress, as manifested by at least two of the following 11 problems within a 12-month period: taking larger amounts, or over a longer period of time, than intended; persistent desire or unsuccessful efforts to cut down or control; a great deal of time spent in activities necessary to obtain, use, or recover; craving; use resulting in failure to fulfill major role obligations; continued use, despite recurrent social or interpersonal problems; giving up social, occupational, or recreational activities; recurrent use in physically

hazardous situations; continued use despite persistent or recurrent physical or psychological problems; tolerance; and withdrawal symptoms, or avoidance of withdrawal symptoms, by continued use.

TREATMENT

TREATMENT

Psychostimulants

COCAINE ACUTE INTOXICATION

As with all emergency situations the first task is to ensure a patent airway, breathing, and circulation. With cocaine use, succinylcholine is relatively contraindicated in rapid sequence intubation; consider rocuronium (1 mg/kg IV) or another nondepolarizing agent as an alternative. If psychomotor agitation occurs, rule out hypoglycemia and hypoxemia first, and then administer benzodiazepines (e.g., diazepam 10 mg IV and then 5–10 mg IV every 3–5 min until agitation controlled). Benzodiazepines are usually sufficient to address cardiovascular side effects. Severe or symptomatic hypertension can be treated with phentolamine, nitroglycerin, or nitroprusside. Hyperthermic patients should be cooled within ≤30 min with the goal to achieve a core body temperature of <39°C (102°F). Evaluation of chest pain in someone using cocaine should include an electrocardiogram, chest radiograph, and biomarkers to exclude myocardial infarction. The treatment approach is similar to noncocaine-induced chest pain, however, it is recommended that whenever possible beta blockers not be used in people who use cocaine. The concern arises from the potential unopposed alpha-adrenergic stimulation that results from beta blockade possibly causing coronary arterial vasoconstriction, ischemia, and infarction and also limited data supporting the benefit of beta blockers in cocaine-related cardiovascular complications. If beta blockers are to be given, it is suggested that mixed alpha/beta blockers, e.g., labetalol and carvedilol, be used rather than nonselective beta blockers, and only in situations where the benefits outweigh the risks. Because many instances of cocaine-related mortality have been associated with concurrent use of other illicit drugs (particularly heroin), the physician must be prepared to institute effective emergency treatment for multiple drug toxicities.

COCAINE USE DISORDERS

Treatment of cocaine use disorders requires the combined efforts of primary care physicians, psychiatrists, and psychosocial care providers. Early abstinence from cocaine use is often complicated by symptoms of depression and guilt, insomnia, and anorexia, which may be as severe as those observed in major affective disorders and can last for months and even years after use has stopped.

Behavioral therapies, including cognitive-behavioral therapy (CBT), the community reinforcement approach (CRA), contingency management (CM; providing rewards to patients who remain substance free), motivational enhancement therapy (MET), combinations of these, and others remain the mainstay of treatment for stimulant use disorders and show modest benefit. These behavioral therapies are designed to help modify the patient’s thinking, expectancies, and behaviors, and to increase life-coping skills, with behavioral interventions to support long-term, drug-free recovery.

There are no U.S. Food and Drug Administration (FDA)-approved medications for psychostimulant addiction. Current research includes several neurotransmitter-based strategies, including DA agonist-, serotonin-, γ-aminobutyric acid (GABA)-, and glutamate-based approaches. Other therapies being studied for the treatment of psychostimulant use disorder include: acamprosate (possibly via a role in Ca^{2+} supply), galantamine (reversible acetylcholine esterase inhibitor, which may strengthen impulse control, as well as cognitive and social abilities depleted by long-term psychostimulant use), naltrexone (opiate receptor antagonist), doxazosin (alpha-adrenergic antagonist), and varenicline (partial agonist at the α4β2 nicotinic acetylcholine receptors and DA neurotransmission enhancer). Vaccines for cocaine and methamphetamine use disorders are also being developed. Finally, recent preliminary studies have brought attention to the use of brain stimulation techniques such as transcranial magnetic stimulation (TMS), theta-burst stimulation (TBS), and transcranial direct current stimulation (tDCS) to treat psychostimulant use disorders, although further studies are warranted.

HALLUCINOGENS

Hallucinogens are a diverse group of drugs causing alteration of thoughts, feelings, sensations, and perceptions. Their use in religious and spiritual rituals goes back centuries. Hallucinogens can be found naturally in plants and mushrooms, or can be human-made. They include: ayahuasca (a tea made from Amazonian plants containing *dimethyltryptamine* (DMT), the primary mind-altering ingredient); DMT (aka Dimitri, can also be synthesized in a lab); *LSD* (clear or white odorless material made from lysergic acid found in rye and other grain fungus); peyote (mescaline, derived from a small, spineless cactus or made synthetically); and *4-phosphoryloxy-N,N-dimethyltryptamine* (psilocybin, comes from certain South and North American mushrooms).

A subgroup of hallucinogens produces the added sensation of feeling out of control or disconnected from one's body or surroundings, these include: DXM (an over-the-counter cough suppressant, when used in high doses); ketamine (a human and veterinary anesthetic); PCP (cyclohexylamine derivative and dissociative anesthetic); and *Salvia divinorum* (salvia, a Mexican, Central, and South American plant).

Hallucinogens are used in a wide variety of ways, including smoking, snorting, and transmucosally. Except for salvia, whose effects last 30 min, the onset of action of hallucinogens is within 20–90 min and the duration of action can be as long as 6–12 h. Hallucinogens disrupt brain chemistry, specifically the neurotransmitters serotonin and glutamate. Effects on the serotonin system can disturb mood, sensory perception, sleep, appetite, body temperature, sexual behavior, and muscle control. Glutamate system effects include perturbations in pain perception, responses to the environment, emotion, and learning and memory.

According to the NSDUH, in 2015, an estimated 1.2 million (0.5%) of people aged ≥12 reported current hallucinogen use. The highest rates were among young adults aged 18–25, with 1.8% (636,000) young adults reporting current hallucinogen use.

Clinical manifestations of hallucinogen use include: hallucinations, intensified feelings, heightened sensory experiences, and time perturbations. Additional physiologic responses include: nausea, increased heart rate,

blood pressure, respiratory rate, or body temperature, loss of appetite, xerostomia, sleep problems, synesthesia, impaired coordination, and hyperhidrosis. “Bad trips” (negative experiences with hallucinogen use) can include panic, paranoia, and psychosis, and may persist for up to 24 h. Such experiences are best treated with supportive reassurance. There is some evidence that chronic effects of hallucinogen use can occur, including persistent psychosis, memory loss, anxiety, depression, and flashbacks.

There are currently no FDA-approved medications for the treatment of hallucinogen addiction. Research on behavioral treatments for hallucinogen addiction is underway.

MARIJUANA

Marijuana policies in several states in the United States have legalized marijuana for medical and/or recreational use. Marijuana refers to the dried leaves, flowers, stems, and seeds from the hemp plant, *Cannabis sativa*. There are >480 natural components found within the *Cannabis sativa* plant, of which 66 have been classified as “cannabinoids”; chemicals unique to the plant. The degree of psychological activity allows for the differentiation of the cannabinoids. Three classes of cannabinoids, the cannabigerols (CBGs), cannabichromenes (CBCs), and cannabidiols (CBDs) are not known to have psychological effects. The psychologically active cannabinoids include: tetrahydrocannabinols (THC), cannabinol (CBN), and cannabinodiol (CBDL), among other cannabinoids. Delta-9-tetrahydrocannabinol (THC) is the main psychoactive chemical, responsible for most of the intoxicating effects. Stronger forms of marijuana include sinsemilla (from specially tended female plants) and concentrated resins, including honey-like *hash oil*, waxy *budder*, and hard amber-like *shatter*.

When smoked, marijuana is quickly absorbed from the lungs into the blood and then sequestered in tissues and metabolized by the liver. Marijuana can also be baked into foods (edibles) and eaten with a resulting slower onset of action of 30–60 min. Cannabinoid receptors (CB_1 and CB_2) have been identified in the CNS (cerebral cortex, basal ganglia, and hippocampus) and PNS, as well as on T and B lymphocytes. Endogenous cannabinoids (such as anandamide) as well as exogenous cannabinoids (THC) bind to the CB_1 receptors.

Cannabinoid effects occur in the limbic system, affecting memory, cognition and psychomotor performance, and the mesolimbic pathway, impacting the reward pathway and areas of pain perception. Effects include: altered senses, altered sense of time, laughter, changes in mood, psychomotor retardation, difficulty with thinking and problem-solving, and impaired memory.

Marijuana is the most commonly used illicit drug in the United States, with 22.2 million (8.3%) current marijuana users aged ≥12 (i.e., users in the past 30 days). In 2015, >11 million young adults, ages 18–25, used marijuana in the past year, and 19.8% used in the past month. Of the 7.7 million people aged ≥12 who had a past year SUD related to their use of illicit drugs, 4.0 million had a past year disorder related to their use of marijuana. In 2015, 2.6% of adolescents aged 12–17, 5.1% of young adults aged 18–25, and 0.8% of adults aged ≥26 had a marijuana use disorder in the past year. Emergency room visits involving marijuana have increased, which may be due to increased THC levels in marijuana over the past few decades resulting in a greater chance of a harmful reaction.

Acute intoxication brings with it a perceived sense of relaxation and mild euphoria, accompanied by some degree of impairment in memory, concentration, judgment, and perceptual and psychomotor function, as well as anxiety, paranoia, and rarely, psychosis. As with all psychoactive compounds, the experience changes depending on the individual's environment and state of mind at the time of use. Physical signs of marijuana use include conjunctival injection and tachycardia. Adverse physical effects of marijuana include: respiratory problems due to inhaled pulmonary irritants and lower birth weights in pregnancy.

Chronic marijuana use may also have adverse psychological effects, which may not be permanent, such as impaired concentration and learning, insomnia, and worsening symptoms in schizophrenia. Upon cessation, or cutting back, there is evidence of a withdrawal syndrome consisting of irritability, insomnia, anorexia, anxiety, and craving. Individuals who begin marijuana use before age 17, while the brain is still developing, may be more prone to cognitive deficits, and may be at higher risk for polydrug addiction in the future.

There are no current medications to treat marijuana use disorder. Behavioral therapies (CBT, CM, MET) and symptomatic treatment of withdrawal, for example selective serotonin reuptake inhibitors (SSRIs) to treat related anxiety, may be effective. Preliminary studies and small clinical trials with zolpidem (sleep aid), buspirone (antianxiety/antistress medication), and gabapentin (antiepileptic) have been promising. Other agents being studied include *N*-acetylcysteine; fatty acid amid hydrolase (FAAH) inhibitors, which may reduce withdrawal by inhibiting the breakdown of endocannabinoids; and allosteric modulators that interact with cannabinoid receptors to inhibit THC's rewarding effects.

Therapeutic use of marijuana includes as an antiemetic in chemotherapy, appetite promoter in AIDS, intraocular pressure reducer in glaucoma, and spasticity reducer in multiple sclerosis and other neurologic disorders.

EMERGING DRUGS

With the aid of the Internet, and some basic over-the-counter (and other) ingredients, the rise of the "kitchen chemist" is upon us. The production of new psychoactive substances (NPSs), such as synthetic cathinones (bath salts) and synthetic cannabinoids (spice), is on the rise and has resulted in the use of unregulated psychoactive substances that are intended to copy the effects of more expensive illegal drugs, such as methamphetamine and cocaine.

Synthetic cathinones (bath salts) are human-made drugs that are chemically like khât, and are often stronger and more dangerous than the natural product. They usually take the form of a white or brown crystal-like powder, packaged in small plastic or foil bundles labeled "not for human consumption," or as "plant food," "jewelry cleaner," or "phone screen cleaner," and sold online and in drug paraphernalia stores. The popular nickname Molly (slang for "molecular") often refers to the purported "pure" crystalline powder form of MDMA, usually sold in capsules. However, people who purchase powder or capsules sold as Molly often actually get other drugs, such as synthetic cathinones. The uncertainty of what is actual in these synthetic products, whose components might change from batch to batch, makes them even more dangerous as anyone using them is unaware of what they contain and how their bodies will react.

The three most common synthetic cathinones are mephedrone, methylone, and MDPV (*3,4-methylenedioxypyrovalerone)*. With oral ingestion, these drugs have an onset of action from 15 to 45 min, and a duration of action that varies from 2 to 7 h. A recent study found that MDPV affects the brain in a manner similar to cocaine, but is at least 10 times more powerful. MDPV is the most common synthetic cathinone found in the blood and urine of patients admitted to EDs after taking “bath salts.” High doses, or chronic use, of synthetic cathinones can lead to dangerous medical consequences, including psychosis, violent behaviors, tachycardia, hyperthermia, and even death.

Synthetic cannabinoids refer to a growing number of human-made psychoactive chemicals that are either sprayed on dried, shredded plant material so they can be smoked (herbal incense), or sold as liquids to be vaporized and inhaled in e-cigarettes and other devices (liquid incense). Synthetic cannabinoids act on the same brain cell receptors as THC, the psychoactive ingredient in marijuana, and with use people report elevated mood, relaxation, altered perception, and symptoms of psychosis, including extreme anxiety, confusion, paranoia, and hallucinations.

Overdose with synthetic cannabinoids can result in tachycardia, vomiting, violent behavior, and suicidal thoughts. Elevations in blood pressure due to vasoconstriction can impair blood flow to the heart, brain, kidney, liver, and other vital organs. Withdrawal symptoms include: headaches, anxiety, depression, and irritability. Behavioral and pharmacologic therapies for treatment of synthetic cannabinoid addiction have not yet been tested.

The ability to synthesize addictive and dangerous drugs relatively simply and rapidly, changing just a few molecules, yet retaining the effects, has allowed many of these emerging drugs to outpace the attempt to regulate them, resulting in a developing global public health concern.

GLOBAL CONSIDERATIONS

 Cannabis remains the most commonly used drug globally, with an estimated 183 million people having used the drug in 2014, while amphetamines are the second most commonly used drug. Overall global trends show the use of cannabis has remained stable over the past 3 years; however, in subregions of North America and Western and Central Europe, cannabis use has increased. Cocaine use had remained stable until 2010 when it also began to rise, driven by an increase in cocaine use in South America. The use of amphetamines appears to be stable; however, drug use data may be an underestimate particularly in subregions in East and South-East Asia, where information is sparse. Globally, men are three times more likely than women to use cannabis, cocaine, or amphetamines, whereas women are more likely than men to participate in the nonmedical use of opioids and tranquilizers. Opioids, cocaine, amphetamines, and cannabis together accounted for almost 12 million life years lost due to premature death or disability in 2013 according to the United Nation’s World Drug Report 2016. Stigma and marginalization makes treatment of drug use disorders difficult and hinders sustainable development incorporating gender equality and the empowerment of women and girls. Drug use further corrodes environmental and economic well-being as well as the ability to develop and sustain safe communities.

FUTURE DIRECTIONS

Despite their prevalence and public health impact, psychostimulant, hallucinogen, and marijuana use disorders have no FDA-approved treatment medications. While behavioral therapies such as CBT, CM, and MET have been shown effective in psychostimulant use disorders, further research needs to be done regarding their utility for hallucinogen and marijuana use disorders. Furthermore, based upon experience with opioid and alcohol use disorders, it is likely that the most efficacious treatments will employ a combination of behavioral and pharmacological therapy.

Additionally, new approaches that utilize emerging technologies have considerable potential for future treatment of psychostimulant use disorders. These include neurostimulation/neuromodulation (TMS, TBS, tDCS), optogenetic techniques (use of light to control neurons that have been genetically modified to express light-sensitive ion channels), wearable biosensors, and mobile technology, including ecological and geographical momentary assessment (EMA/GMA) as well as real-time interventions delivered via smartphone or other mobile devices.

FURTHER READING

Koob GF, Volkow ND: The neurobiology of addiction: A neurocircuitry analysis. Lancet Psychiatry 3:760, 2016. [PubMed: 27475769]

Volkow ND et al: Neurobiologic advances from the brain disease model of addiction. N Engl J Med 374:363, 2016. [PubMed: 26816013]

WEBSITES

American Society of Addiction Medicine: *https://www.asam.org/public-resources*

National Institute on Drug Abuse: *https://www.drugabuse.gov/drugs-abuse*

World Health Organization: *http://www.who.int/substance_abuse/en/*

Chapter 448: Nicotine Addiction

David M. Burns

INTRODUCTION

The use of tobacco leaf to create and satisfy nicotine addiction was introduced to Columbus by Native Americans and spread rapidly to Europe. Use of tobacco as cigarettes, however, only became popular in the twentieth century and so is a modern phenomenon, as is the epidemic of disease caused by this form of tobacco use.

Nicotine is the principal constituent of tobacco responsible for its addictive character, but other smoke constituents and behavioral associations contribute to the strength of the addiction. Addicted smokers regulate their nicotine intake by adjusting the frequency and intensity of their tobacco use both to obtain the desired psychoactive effects and avoid withdrawal.

Unburned cured tobacco used orally contains nicotine, carcinogens, and other toxicants capable of causing gum disease, oral and pancreatic cancers, and an increase in the risk of heart disease. When tobacco is burned, the resultant smoke contains, in addition to nicotine, >7000 other compounds that result from volatilization, pyrolysis, and pyrosynthesis of tobacco leaf and various chemical additives used in making different tobacco products. Tobacco smoke is composed of a fine aerosol and a vapor phase; the aerosol is of a size range that results in deposition in the airways and alveolar surfaces of the lungs. The aggregate of particulate matter, after subtracting nicotine and moisture, is referred to as tar.

The alkaline pH of smoke from blends of tobacco used for pipes and cigars allows sufficient absorption of nicotine across the oral mucosa to satisfy the smoker’s need for this drug. Therefore, those who smoke pipes and cigars exclusively tend not to inhale the smoke into the lung, confining the toxic and carcinogenic exposure (and the increased rates of disease) largely to the upper airway. The acidic pH of smoke generated by the tobacco used in cigarettes dramatically reduces absorption of nicotine in the mouth, necessitating inhalation of the smoke into the larger surface of the lungs in order to absorb quantities of nicotine sufficient to satisfy the smoker’s addiction. The shift to using tobacco as cigarettes, with resultant increased deposition and absorption of smoke in the lung, has created the epidemic of heart disease, lung disease, and lung cancer that dominates the current disease manifestations of tobacco use.

Several genes have been associated with nicotine addiction. Some reduce the clearance of nicotine, and others have been associated with an increased likelihood of becoming dependent on tobacco and other drugs as well as a higher incidence of depression. It is likely that genetic susceptibility can influence the

probability that adolescent experimentation with tobacco will lead to addiction as an adult. Rates of smoking cessation have increased, and rates of nicotine addiction have decreased dramatically, since the mid-1950s, suggesting that factors other than genetics are more important influences for tobacco use.

Adult cigarette smoking prevalence has declined to about 15% in the United States, with only 11.4% of the population smoking every day. However, the rapidly rising smoking rate observed in the developing world is of concern, and the World Health Organization Framework Convention on Tobacco Control is encouraging effective tobacco control approaches in these countries with the hope of preventing a future epidemic of tobacco-related illness.

DISEASE MANIFESTATIONS OF CIGARETTE SMOKING

More than 480,000 individuals die prematurely each year in the United States from cigarette use; this represents almost one of every five deaths in the United States. Approximately 40% of cigarette smokers will die prematurely due to cigarette smoking unless they are able to quit.

The major diseases caused by cigarette smoking are listed in **Table 448-1**. The ratio of smoking-related disease rates in smokers compared to never smokers (relative risk) increases with advancing age for most cancers and for chronic obstructive pulmonary disease (COPD). However, relative risk declines with advancing age for cardiovascular diseases due to the increasing contribution of other risk factors to cardiovascular disease occurrence as age advances. Nevertheless, even for cardiovascular disease, the absolute difference in mortality rate between smokers and never smokers, called excess death rate, continues to increase with advancing age, as one would expect from a process of cumulative injury.

TABLE 448-1

Relative Risks for Current Smokers of Cigarettes

Age	35–44	45–64	65–74	≥75
Males				
Lung cancer	14.33	19.03	28.29	22.51
Coronary heart disease	3.88	2.99	2.76	1.98
Cerebrovascular disease	2.17	1.48	1.23	1.12
Other vascular diseases			7.25	4.93
Chronic obstructive pulmonary disease (COPD)			29.69	23.01
All causes	2.55	2.97	3.02	2.40
Females				
Lung cancer	13.30	18.95	23.65	23.08
Other tobacco-related cancers	1.28	2.08	2.06	1.93
Coronary heart disease	4.98	3.25	3.29	2.25
Cerebrovascular disease	2.27	1.70	1.24	1.10
Other vascular diseases			6.81	5.77
COPD			38.89	20.96
All causes	1.79	2.63	2.87	2.47
Relative Risks for Selected Other Cancers				
Other cancers	Male		Female	
Larynx	14.6		13	
Lip, oral cavity, pharynx	10.9		5.1	

Age	35–44	45–64	65–74	≥75
Esophagus	6.8		7.8	
Bladder	3.3		2.2	
Kidney	2.7		1.3	
Pancreas	2.3		2.3	
Stomach	2		1.4	
Liver	1.7		1.7	
Colorectal	1.2		1.2	
Cervix			1.6	
Acute myeloid leukemia	1.4		1.4	

CARDIOVASCULAR DISEASES

Cigarette smokers are more likely than nonsmokers to develop both large-vessel atherosclerosis and small-vessel disease. Approximately 90% of peripheral vascular disease in the nondiabetic population can be attributed to cigarette smoking, as can ~50% of aortic aneurysms. In contrast, 24% of coronary artery disease and ~11% of ischemic and hemorrhagic strokes are caused by cigarette smoking. There is a multiplicative interaction between cigarette smoking and other cardiac risk factors such that the increment in risk produced by smoking among individuals with hypertension or elevated serum lipids is substantially greater than the increment in risk produced by smoking for individuals without these risk factors.

In addition to its role in promoting atherosclerosis, cigarette smoking also increases the likelihood of myocardial infarction and sudden cardiac death by promoting platelet aggregation and vascular occlusion. Reversal of these effects on coagulation may explain the rapid benefit of smoking cessation for a new coronary event demonstrable among those who have survived a first myocardial infarction. This effect may also explain the substantially higher rates of graft occlusion among continuing smokers following vascular bypass surgery for cardiac or peripheral vascular disease.

Cessation of cigarette smoking reduces the risk of a second coronary event within 6–12 months. Rates of first myocardial infarction and death from coronary heart disease decline within 2–4 years following cessation among those with no prior cardiovascular history. After 15 years of abstinence, the risk of a new myocardial

infarction or death from coronary heart disease in former smokers is similar to that for those who have never smoked.

CANCER

Tobacco smoking causes cancer of the lung; lip; oral cavity; naso-, oro-, and hypopharynx; nasal cavity and paranasal sinuses; larynx; esophagus; stomach; pancreas; liver (hepatocellular); colon and rectum; kidney (body and pelvis); ureter; urinary bladder; uterine cervix; and acute myeloid leukemia. There is evidence suggesting that cigarette smoking may play a role in increasing the risk of breast cancer, particularly for premenopausal women. There does not appear to be a causal link between cigarette smoking and cancer of the endometrium, and there is a lower risk of uterine cancer among postmenopausal women who smoke. The risks of cancer increase with the increasing number of cigarettes smoked per day and with increasing duration of smoking. Additionally, there are synergistic interactions between cigarette smoking and alcohol use for cancer of the oral cavity and esophagus. Several occupational exposures synergistically increase lung cancer risk among cigarette smokers, most notably occupational asbestos and radon exposure.

Cessation of cigarette smoking reduces the risk of developing cancer relative to continuing smoking after about 4 years of abstinence, but even 20 years after cessation there is a modest persistent increased risk of developing lung cancer.

RESPIRATORY DISEASE

Cigarette smoking is responsible for 80% of COPD. Within 1–2 years of beginning to smoke regularly, many young smokers will develop inflammatory changes in their small airways, although lung function measures of these changes do not predict development of chronic airflow obstruction. Pathophysiologic changes in the lungs manifest and progress over longer durations of smoking proportional to smoking intensity and duration. Chronic mucous hyperplasia of the larger airways results in a chronic productive cough in as many as 80% of smokers >60 years of age. Chronic inflammation and narrowing of the small airways, and/or enzymatic digestion of alveolar walls resulting in pulmonary emphysema, can reduce expiratory airflow sufficiently to produce clinical symptoms of respiratory limitation in ~15–25% of smokers.

Changes in the small airways of young smokers will reverse after 1–2 years of cessation. There is also a small increase in measures of expiratory airflow following cessation among many individuals who have developed chronic airflow obstruction, but the major change following cessation is a slowing of the rate of decline in lung function with advancing age rather than a return of lung function toward normal.

PREGNANCY

Cigarette smoking is associated with several maternal complications of pregnancy: premature rupture of membranes, abruptio placentae, and placenta previa; there is also a small increase in the risk of spontaneous abortion among smokers. Infants of smoking mothers are more likely to experience preterm delivery, have a higher perinatal mortality rate, be small for their gestational age, and have higher rates of

infant respiratory distress syndrome. They are more likely to die of sudden infant death syndrome and appear to have a developmental lag for at least the first several years of life.

OTHER CONDITIONS

Smoking delays healing of peptic ulcers; increases the risk of developing periodontal disease, diabetes, active tuberculosis, rheumatoid arthritis, osteoporosis, senile cataracts, and neovascular and atrophic forms of macular degeneration; and results in premature menopause, wrinkling of the skin, gallstones and cholecystitis in women, and male impotence. Patients who continue to smoke during treatment for cancer with chemotherapy or radiation have poorer outcomes and reduced survival.

ENVIRONMENTAL TOBACCO SMOKE

Long-term exposure to environmental tobacco smoke increases the risk of lung cancer and coronary artery disease among nonsmokers. It also increases the incidence of respiratory infections, chronic otitis media, and asthma in children and causes exacerbation of asthma in children. Some evidence suggests that environmental tobacco smoke exposure may increase the risk of premenopausal breast cancer.

PHARMACOLOGIC INTERACTIONS

Cigarette smoking may interact with a variety of other drugs **(Table 448-2)**. Cigarette smoking induces the cytochrome P450 system, which may alter the metabolic clearance of drugs such as warfarin. This may result in inadequate serum levels in smokers as outpatients when the dosage is established in the hospital under nonsmoking conditions. Correspondingly, serum levels may rise when smokers are hospitalized and not allowed to smoke. Smokers may also have higher first-pass clearance for drugs such as lidocaine, and the stimulant effects of nicotine may reduce the effect of benzodiazepines or beta blockers.

TABLE 448-2

Interactions of Smoking and Prescription Drugs

Drug	Interaction
Cardiovascular and Pulmonary Drugs	
β blockers	Reduced lowering of heart rate and blood pressure
Flecainide	Increased first-pass clearance
Heparin	Faster clearance
Lidocaine	Increased first-pass clearance
Mexiletine	Increased first-pass clearance
Propranolol	Increased first-pass clearance
Theophylline	Faster metabolic clearance
Verapamil	Increased clearance
Warfarin	Increased metabolism lowers serum levels
Neuropsychiatric Drugs	
Amitriptyline	Increased clearance
Benzodiazepines	Less sedation
Clomipramine, Imipramine	Decreased serum concentrations
Chlorpromazine	Decreased serum concentrations
Clozapine	Decreased serum concentrations
Duloxetine	Decreased serum concentrations
Fluphenazine	Decreased serum concentrations

Drug	Interaction
Fluvoxamine	Decreased serum concentrations
Haloperidol	Decreased serum concentrations
Naratriptan	Increased clearance
Olanzapine	Faster clearance
Trazodone	Decreased serum concentrations
Anticancer Drugs	
Erlotinib	Increased clearance, higher response rate, and improved survival in non-smokers
Gefitinib	Higher response rate and improved survival in non-smokers
Irinotecan	Increased clearance
Other Drugs	
Dextropropoxyphene	Less analgesia
Estrogens (oral)	Increased hepatic clearance
Fentanyl	Increased clearance
Insulin	Delayed absorption due to skin vasoconstriction
Rivastigmine	Increased clearance

OTHER FORMS OF TOBACCO USE

Other major forms of tobacco use are moist snuff deposited between the cheek and gum, chewing tobacco, pipes and cigars, and recently bidi (tobacco wrapped in tendu or temburni leaf; commonly used in India), clove cigarettes, and water pipes. Oral tobacco use leads to gum disease and can result in oral and pancreatic cancer as well as heart disease. There are dramatic differences in the risks evident for products used in Africa or Asia as compared to those in the United States and Europe.

The risk of upper airway cancers is similar among cigarette, pipe, and cigar smokers, whereas those who have smoked only pipes and cigars have a much lower risk of lung cancer, heart disease, and COPD. Cigarette smokers who switch to pipes or cigars do tend to inhale the smoke, increasing their risk.

A resurgence of cigar, bidi, and water pipe use among adolescents of both genders has raised concerns that these older forms of tobacco use are once again causing a public health problem.

ELECTRONIC CIGARETTES

Use of electronic nicotine delivery systems, often called e-cigarettes, is a new behavior where both the products and patterns of use are rapidly evolving. These devices electronically heat a solution, which may or may not contain nicotine, to produce a vapor that can be inhaled. Smaller devices resembling cigarettes often deliver less nicotine than combustible cigarettes, whereas the larger devices with substantial tanks for the nicotine solution can deliver amounts of nicotine similar to combustible cigarettes. Absent the identification of new toxicants or new disease risks, the existing evidence on toxicant exposure establishes that smokers who use e-cigarettes exclusively will receive less toxicant exposure than combustible cigarette smokers and, with sustained exclusive use, would be expected to experience less disease risk.

Addicted cigarette smokers can derive sufficient nicotine from e-cigarettes to satisfy their addiction and some will persist in e-cigarette use for multiple months. Given the newness of this behavior, it is not clear whether smokers who switch to exclusive e-cigarette use will persist in that behavior, quit nicotine use entirely, or relapse back to smoking over the multiyear period needed to alter disease risks.

Patterns of e-cigarette use indicate that only small percentages of adults use e-cigarettes. The highest prevalence of e-cigarette use is among adolescents and young adults where the prevalence of e-cigarette use substantially exceeds that of cigarettes. It is currently uncertain whether the high prevalence of e-cigarette use among youth will translate into nicotine addiction and combustible cigarette use as they age. Nevertheless, the prevalence of combustible cigarette use among adolescents and young adults has continued to decline over recent years, even with widespread uptake of e-cigarettes.

LOWER TAR AND NICOTINE CIGARETTES

Filtered cigarettes with lower machine-measured yields of tar and nicotine commonly use ventilation holes in the filters and other engineering designs to artificially lower the machine measurements. Smokers compensate for the lowered nicotine delivery resulting from these design changes by changing the manner in which they puff on the cigarette or the number of cigarettes smoked per day, and actual tar and nicotine deliveries are not reduced with use of these products negating any reduction in disease risks with their use. In addition, the amount of carcinogenic tobacco-specific nitrosamines in the tobacco used in cigarettes has increased over time.

Cigarette design changes that reduce machine-measured tar and nicotine also lead to deeper inhalation of the smoke. Presentation of more carcinogenic smoke to the alveolar portions of the lung increases the risk of

adenocarcinoma of the lung. The increased adenocarcinoma risk produces a substantively greater overall risk for lung cancer among current smokers compared with smokers of cigarettes manufactured prior to the 1960s. This increased risk may also be present for COPD. It is the changes in cigarette design and composition of cigarettes over the past six decades that caused the dramatic rise in rates of adenocarcinoma of the lung observed over the past half century. There has been no increase in risk of lung cancer or adenocarcinoma of the lung over the same period among never smokers.

CESSATION

The process of stopping smoking is commonly a cyclical one, with the smoker sometimes making multiple attempts to quit and failing before finally being successful. Approximately 70–80% of smokers would like to quit smoking. More than one-half of current smokers attempted to quit in the last year, but only 6% quit for 6 months, and only 3% remain abstinent for 2 years. Clinician-based smoking interventions should repeatedly encourage smokers to try to quit, and to use different forms of cessation assistance with each new cessation attempt, rather than focusing exclusively on immediate cessation at the time of the first visit.

Advice from a physician to quit smoking, particularly at the time of an acute illness, is a powerful trigger for cessation attempts, with up to half of patients who are advised to quit making a cessation effort. Other triggers that may be enhanced by timely physician advice to quit include increases in the tax on cigarettes, media campaigns, and changes in rules to restrict smoking in the workplace.

PHYSICIAN INTERVENTION

All patients should be asked whether they smoke, how much they smoke, how long they have smoked, their past experience with quitting, and whether they are currently interested in quitting **(Table 448-3)**. The number of cigarettes smoked per day and smoking within 30 min of waking are useful measures of the intensity of nicotine addiction. Even those who are not interested in quitting should be encouraged and motivated to quit; provided a clear, strong, and personalized message by the clinician that smoking is an important health concern; and offered assistance if they become interested in quitting in the future. Many of those not currently expressing an interest in quitting may nevertheless make an attempt to quit in the subsequent year. For those interested in quitting, a quit date should be negotiated, usually not the day of the visit but within the next few weeks, and a follow-up contact by office staff around the time of the quit date should be provided. There is a relationship between the amount of assistance a patient is willing to accept and the success of the cessation attempt.

TABLE 448-3

Clinical Practice Guidelines

Physician Actions
Ask: Systematically identify all tobacco users at every visit
Advise: Strongly urge all smokers to quit
Identify smokers willing to quit
Assist the patient in quitting
Arrange follow-up contact
Effective Pharmacologic Interventions[a]
First-line therapies
Nicotine gum (1.5)
Nicotine patch (1.9)
Nicotine nasal inhaler (2.3)
Nicotine oral inhaler (2.1)
Nicotine lozenge (2 mg 2.0, 4 mg 2.8)
Bupropion (2.0)
Varenicline (3.1)
Second-line therapies
Clonidine (2.1)
Nortriptyline (1.8)
Other Effective Interventions[a]
Physician or other medical personnel counseling (10 min) (1.84)

Intensive group smoking cessation programs (at least 4–7 sessions of 20- to 30-min duration lasting at least 2 and preferably 8 weeks) (1.3)
Intensive individual counseling (1.7)
Clinic-based smoking status identification system (3.1)
Telephone counseling (1.6)

[a]Numerical value following the intervention is the multiple for cessation success compared to no intervention.

There are a variety of nicotine-replacement products, including over-the-counter nicotine patches, gum, and lozenges, as well as nicotine nasal and oral inhalers available by prescription. These products can be used for up to 3–6 months, and some products are formulated to allow a gradual step-down in dosage with increasing duration of abstinence. Antidepressants such as bupropion (300 mg in divided doses for up to 6 months) have also been shown to be effective, as has varenicline, a partial agonist for the nicotinic acetylcholine receptor (initial dose 0.5 mg daily increasing to 1 mg twice daily at day 8; treatment duration up to 6 months). Combined use of nicotine-replacement therapy (NRT) and antidepressants as well as the use of gum or lozenges for acute cravings in patients using patches can increase cessation outcomes. Pretreatment with antidepressants or varenicline is recommended for 1–2 weeks prior to the quit date. Pretreatment with nicotine-replacement products is also useful prior to a cessation date. Longer duration of nicotine replacement as a maintenance therapy for those who are unsuccessful in quitting with a shorter duration of use is a useful strategy. NRT is provided in different dosages, with higher doses being recommended for more intense smokers. Clonidine or the tricyclic antidepressant nortriptyline should be reserved for patients who have failed on first-line pharmacologic treatment or who are unable to use other therapies. Antidepressants are more effective among smokers with a history of depression symptoms.

Current recommendations are to offer pharmacologic treatment, usually with nicotine patches or varenicline, to all who will accept it and to provide counseling and other support as a part of the cessation attempt. Cessation advice alone by a physician or his or her staff is likely to increase success compared with no intervention; a more comprehensive approach with advice, pharmacologic assistance, and counseling can increase cessation success nearly threefold.

For adult addicted smokers, switching to exclusive use of e-cigarettes, but not dual use with combusted cigarettes, may have a role in promoting cessation, particularly for those unlikely to try to quit with other proven cessation modalities. However, it is not yet clear whether e-cigarette use results in higher cessation outcomes than other pharmacological approaches in the context of physician-based smoking interventions.

Incorporation of cessation assistance into a practice requires a change of the care delivery infrastructure. Simple changes include (1) adding questions about smoking and interest in cessation on patient-intake questionnaires, (2) asking patients whether they smoke as part of the initial vital sign measurements made

by office staff, (3) listing smoking as a problem in the medical record, and (4) automating follow-up contact with the patient on the quit date. These changes are essential to institutionalizing smoking intervention within the practice setting; without this institutionalization, the best intentions of physicians to intervene with their patients who smoke are often lost in the time crush of a busy practice.

PREVENTION

Approximately 85% of individuals who become cigarette smokers initiate the behavior during adolescence. Factors that promote adolescent initiation are parental or older-sibling cigarette smoking, tobacco advertising and promotional activities, the availability of cigarettes, and the social acceptability of smoking. The need for an enhanced self-image and to imitate adult behavior is greatest for those adolescents who have the least external validation of their self-worth, which may explain in part the enormous differences in adolescent smoking prevalence by socioeconomic and school performance strata.

Prevention of smoking initiation must begin early, preferably in the elementary school years. Physicians who treat adolescents should be sensitive to the prevalence of this problem even in the preteen population. Physicians should ask all adolescents whether they have experimented with tobacco or currently use tobacco, reinforce the fact that most adolescents and adults do not smoke, and explain that all forms of tobacco are both addictive and harmful.

FURTHER READING

El Dib R et al: Electronic nicotine delivery systems and/or electronic non-nicotine delivery systems for tobacco smoking cessation or reduction: A systematic review and meta-analysis. BMJ Open 23;7:e012680, 2017.

Sohn HS et al: Evidence supporting the need for considering the effects of smoking on drug disposition and effectiveness in medication practices: A systematic narrative review. Int J Clin Pharmacol Ther 53:621, 2015. [PubMed: 26104035]

U.S. Department of Health and Human Services: *The Health Consequences of Smoking: 50 Years of Progress. A Report of the Surgeon General*. Treatment for Tobacco Use and Dependence; Appendix 14.5 Smoking Cessation Medications. Atlanta, GA, 2014. Available from https://www.surgeongeneral.gov/library/reports/50-years-of-progress/50-years-of-progress-by-section.html only available electronically. Accessed March 23, 2017.

U.S. Department of Health and Human Services: *The Health Consequences of Smoking: 50 Years of Progress. A Report of the Surgeon General*. Atlanta, GA, 2014. Available from https://www.surgeongeneral.gov/library/reports/50-years-of-progress/index.html. Accessed March 23, 2017.

Chapter 449: Heavy Metal Poisoning

Howard Hu

INTRODUCTION

Metals pose a significant threat to health through low-level environmental as well as occupational exposures. One indication of their importance relative to other potential hazards is their ranking by the U.S. Agency for Toxic Substances and Disease Registry, which maintains an updated list of all hazards present in toxic waste sites according to their prevalence and the severity of their toxicity. The first, second, third, and seventh hazards on the list are heavy metals: lead, mercury, arsenic, and cadmium, respectively (*http://www.atsdr.cdc.gov/spl/*). Specific information pertaining to each of these metals, including sources and metabolism, toxic effects produced, diagnosis, and the appropriate treatment for poisoning, is summarized in **Table 449-1**.

TABLE 449-1

Heavy Metals

MAIN SOURCES	METABOLISM	TOXICITY	DIAGNOSIS	TREATMENT
Arsenic				
Smelting and microelectronics industries; wood preservatives, pesticides, herbicides, fungicides; contaminant of deep-water wells; folk remedies; and coal; incineration of these products.	Organic arsenic (arsenobetaine, arsenocholine) is ingested in seafood and fish, but is nontoxic; inorganic arsenic is readily absorbed (lung and GI); sequesters in liver, spleen, kidneys, lungs, and GI tract; residues persist in skin, hair, and nails; biomethylation results in detoxification, but this process saturates.	Acute arsenic poisoning results in necrosis of intestinal mucosa with hemorrhagic gastroenteritis, fluid loss, hypotension, delayed cardiomyopathy, acute tubular necrosis, and hemolysis. Chronic arsenic exposure causes diabetes, vasospasm, peripheral vascular insufficiency and gangrene, peripheral neuropathy, and cancer of skin, lung, liver (angiosarcoma), bladder, and kidney. Lethal dose: 120–200 mg (adults); 2 mg/kg (children).	Nausea, vomiting, diarrhea, abdominal pain, delirium, coma, seizures; garlicky odor on breath; hyperkeratosis, hyperpigmentation, exfoliative dermatitis, and Mees' lines (transverse white striae of the fingernails); sensory and motor polyneuritis, distal weakness. Radiopaque sign on abdominal x-ray; ECG–QRS broadening, QT prolongation, ST depression, T-wave flattening; 24-h urinary arsenic >67 µmol/d or 50 µg/d; (no seafood × 24 h); if recent exposure, serum arsenic >0.9 µmol/L (7 µg/dL). High arsenic in hair or nails.	If acute ingestion, ipecac to induce vomiting, gastric lavage, activated charcoal with a cathartic. Supportive care in ICU. Dimercaprol 3–5 mg/kg IM q4h × 2 days; q6h × 1 day, then q12h × 10 days; alternative: oral succimer.

MAIN SOURCES	METABOLISM	TOXICITY	DIAGNOSIS	TREATMENT
Cadmium				
Metal-plating, pigment, smelting, battery, and plastics industries; tobacco; incineration of these products; ingestion of food that concentrates cadmium (grains, cereals).	Absorbed through ingestion or inhalation; bound by metallothionein, filtered at the glomerulus, but reabsorbed by proximal tubules (thus, poorly excreted). Biologic half-life: 10–30 y. Binds cellular sulfhydryl groups, competes with zinc, calcium for binding sites. Concentrates in liver and kidneys.	Acute cadmium inhalation causes pneumonitis after 4–24 h; acute ingestion causes gastroenteritis. Chronic exposure causes anosmia, yellowing of teeth, emphysema, minor LFT elevations, microcytic hypochromic anemia unresponsive to iron therapy, proteinuria, increased urinary β_2-microglobulin, calciuria, leading to chronic renal failure, osteomalacia, and fractures. Possible risks of cardiovascular disease and cancer.	With inhalation: pleuritic chest pain, dyspnea, cyanosis, fever, tachycardia, nausea, noncardiogenic pulmonary edema. With ingestion: nausea, vomiting, cramps, diarrhea. Bone pain, fractures with osteomalacia. If recent exposure, serum cadmium >500 nmol/L (5 µg/dL). Urinary cadmium >100 nmol/L (10 µg/g creatinine) and/or urinary β_2-microglobulin >750 µg/g creatinine (but urinary β_2-microglobulin also increased in other renal diseases such as pyelonephritis).	There is no effective treatment for cadmium poisoning (chelation not useful; dimercaprol can exacerbate nephrotoxicity). Avoidance of further exposure, supportive therapy, vitamin D for osteomalacia.
Lead				
Manufacturing of auto batteries, lead crystal,	Absorbed through ingestion or inhalation;	Acute exposure with blood lead levels (BPb) of >60–80 µg/dL can	Abdominal pain, irritability, lethargy, anorexia, anemia, Fanconi's syndrome,	Identification and correction of exposure sources is critical. In some U.S. states, screening and

MAIN SOURCES	METABOLISM	TOXICITY	DIAGNOSIS	TREATMENT
ceramics, fishing weights, etc.; demolition or sanding of lead-painted houses, bridges; stained glass–making, plumbing, soldering; environmental exposure to paint chips, house dust (in homes built <1975), firing ranges (from bullet dust), food or water from improperly glazed ceramics, lead pipes; contaminated herbal remedies, candies; exposure to the combustion of leaded fuels.	organic lead (e.g., tetraethyl lead) absorbed dermally. In blood, 95–99% sequestered in RBCs—thus, must measure lead in whole blood (not serum). Distributed widely in soft tissue, with half-life ~30 days; 15% of dose sequestered in bone with half-life of >20 years. Excreted mostly in urine, but also appears in other fluids including breast milk. Interferes with mitochondrial oxidative phosphorylation, ATPases, calcium-dependent messengers; enhances oxidation and cell apoptosis.	cause impaired neurotransmission and neuronal cell death (with central and peripheral nervous system effects); impaired hematopoiesis and renal tubular dysfunction. At higher levels of exposure (e.g., BPb >80–120 µg/dL), acute encephalopathy with convulsions, coma, and death may occur. Subclinical exposures in children (BPb 25–60 µg/dL) are associated with anemia; mental retardation; and deficits in language, motor function, balance, hearing, behavior, and school performance. Impairment of IQ appears to occur at even lower levels of exposure with no measurable threshold above	pyuria, azotemia in children with blood lead level (BPb) >80 µg/dL; may also see epiphyseal plate “lead lines” on long bone x-rays. Convulsions, coma at BPb >120 µg/dL. Noticeable neurodevelopmental delays at BPb of 40–80 µg/dL; may also see symptoms associated with higher BPb levels. Screening of all U.S. children when they begin to crawl (~6 months) is recommended by the CDC; source identification and intervention is begun if the BPb >10 µg/dL. In adults, acute exposure causes similar symptoms as in children as well as headaches, arthralgias, myalgias, depression, impaired short-term memory, loss of libido. Physical examination may	reporting to local health boards of children with BPb >10 µg/dL and workers with BPb >40 µg/dL is required. In the highly exposed individual with symptoms, chelation is recommended with oral DMSA (succimer); if acutely toxic, hospitalization and IV or IM chelation with ethylenediaminetetraacetic acid calcium disodium (CaEDTA) may be required, with the addition of dimercaprol to prevent worsening of encephalopathy. It is uncertain whether children with asymptomatic lead exposure (e.g., BPb 20–40 µg/dL) benefit from chelation; a recent randomized trial showed no benefit. Correction of dietary deficiencies in iron, calcium, magnesium, and zinc will lower lead absorption and may also improve toxicity. Vitamin C is a weak but natural chelating agent. Calcium supplements (1200 mg at bedtime) have been shown to lower blood lead levels in pregnant women.

MAIN SOURCES	METABOLISM	TOXICITY	DIAGNOSIS	TREATMENT
		the limit of detection in most assays of 1 µg/dL. In adults, chronic subclinical exposures (BPb >40 µg/dL) are associated with an increased risk of anemia, demyelinating peripheral neuropathy (mainly motor), impairments of reaction time and hearing, accelerated declines in cognition, hypertension, ECG conduction delays, higher risk of cardiovascular disease and death, interstitial nephritis and chronic renal failure, diminished sperm counts, and spontaneous abortions.	reveal a “lead line” at the gingiva-tooth border, pallor, wrist drop, and cognitive dysfunction (e.g., declines on the mini-mental state exam); lab tests may reveal a normocytic, normochromic anemia, basophilic stippling, an elevated blood protoporphyrin level (free erythrocyte or zinc), and motor delays on nerve conduction. U.S. OSHA requires regular testing of lead-exposed workers with removal if BPb >40 µg/dL. New guidelines have been proposed recommending that BPb be maintained at <10 µg/dL, removal of workers if BPb >20 µg/dL, and monitoring of cumulative exposure parameters.	
Mercury				
Metallic, mercurous, and	Elemental mercury (Hg) is	Acute inhalation of Hg vapor causes	Chronic exposure to metallic mercury	Treat acute ingestion of mercuric salts with induced

MAIN SOURCES	METABOLISM	TOXICITY	DIAGNOSIS	TREATMENT
mercuric mercury (Hg, Hg^{+}, Hg^{2+}) exposures occur in some chemical, metal-processing, electrical-equipment, automotive industries; they are also in thermometers, dental amalgams, batteries. Mercury is dispersed by waste incineration. Environmental bacteria convert inorganic to organic mercury, which then bioconcentrates up the aquatic food chain to contaminate tuna, swordfish, and other pelagic fish.	not well absorbed; however, it will volatilize into highly absorbable vapor. Inorganic mercury is absorbed through the gut and skin. Organic mercury is well absorbed through inhalation and ingestion. Elemental and organic mercury cross the blood-brain barrier and placenta. Mercury is excreted in urine and feces and has a half-life in blood of ~60 days; however, deposits will remain in the kidney and brain for years. Exposure to mercury stimulates the kidney to produce metallothionein, which provides	pneumonitis and noncardiogenic pulmonary edema leading to death, CNS symptoms, and polyneuropathy. Chronic high exposure causes CNS toxicity (mercurial *erethism*; see Diagnosis); lower exposures impair renal function, motor speed, memory, coordination. Acute ingestion of inorganic mercury causes gastroenteritis, the nephritic syndrome, or acute renal failure, hypertension, tachycardia, and cardiovascular collapse, with death at a dose of 10–42 mg/kg. Ingestion of organic mercury causes gastroenteritis, arrhythmias, and lesions in the basal ganglia, gray	vapor produces a characteristic intention tremor and mercurial *erethism*: excitability, memory loss, insomnia, timidity, and delirium (“mad as a hatter”). On neurobehavioral tests: decreased motor speed, visual scanning, verbal and visual memory, visuomotor coordination. Children exposed to mercury in any form may develop *acrodynia* (“pink disease”): flushing, itching, swelling, tachycardia, hypertension, excessive salivation or perspiration, irritability, weakness, morbilliform rashes, desquamation of palms and soles. Toxicity from elemental or inorganic mercury exposure begins when blood levels >180 nmol/L (3.6 µg/dL) and urine	emesis or gastric lavage and polythiol resins (to bind mercury in the GI tract). Chelate with dimercaprol (up to 24 mg/kg per day IM in divided doses), DMSA (succimer), or penicillamine, with 5-day courses separated by several days of rest. If renal failure occurs, treat with peritoneal dialysis, hemodialysis, or extracorporeal regional complexing hemodialysis and succimer. Chronic inorganic mercury poisoning is best treated with *N*-acetyl penicillamine.

MAIN SOURCES	METABOLISM	TOXICITY	DIAGNOSIS	TREATMENT
	some detoxification benefit. Mercury binds sulfhydryl groups and interferes with a wide variety of critical enzymatic processes.	matter, and cerebellum at doses >1.7 mg/kg. High exposure during pregnancy causes derangement of fetal neuronal migration resulting in severe mental retardation. Mild exposures during pregnancy (from fish consumption) are associated with declines in neurobehavioral performance in offspring. Dimethylmercury, a compound only found in research labs, is “supertoxic”—a few drops of exposure via skin absorption or inhaled vapor can cause severe cerebellar degeneration and death.	levels >0.7 µmol/L (15 µg/dL). Exposures that ended years ago may result in a >20-µg increase in 24-h urine after a 2-g dose of succimer. Organic mercury exposure is best measured by levels in blood (if recent) or hair (if chronic); CNS toxicity in children may derive from fetal exposures associated with maternal hair Hg >30 nmol/g (6 µg/g).	

Abbreviations: ATPase, adenosine triphosphatase; BPb, blood lead; CDC, Centers for Disease Control and Prevention; CNS, central nervous system; DMSA, dimercaptosuccinic acid; ECG, electrocardiogram; GI, gastrointestinal; ICU,

intensive care unit; IQ, intelligence quotient; LFT, liver function tests; OSHA, Occupational Safety and Health Administration; RBC, red blood cell.

Metals are inhaled primarily as dusts and fumes (the latter defined as tiny particles generated by combustion). Metal poisoning can also result from exposure to vapors (e.g., mercury vapor in creating dental amalgams). When metals are ingested in contaminated food or drink or by hand-to-mouth activity (implicated especially in children), their gastrointestinal absorption varies greatly with the specific chemical form of the metal and the nutritional status of the host. Once a metal is absorbed, blood is the main medium for its transport, with the precise kinetics dependent on diffusibility, protein binding, rates of biotransformation, availability of intracellular ligands, and other factors. Some organs (e.g., bone, liver, and kidney) sequester metals in relatively high concentrations for years. Most metals are excreted through renal clearance and gastrointestinal excretion; some proportion is also excreted through salivation, perspiration, exhalation, lactation, skin exfoliation, and loss of hair and nails. The intrinsic stability of metals facilitates tracing and measurement in biologic material, although the clinical significance of the levels measured is not always clear.

Some metals, such as copper and selenium, are essential to normal metabolic function as trace elements (Chap. 326) but are toxic at high levels of exposure. Others, such as lead and mercury, are xenobiotic and theoretically are capable of exerting toxic effects at any level of exposure. Indeed, much research is currently focused on the contribution of low-level xenobiotic metal exposure to chronic diseases and to subtle changes in health that may have significant public health consequences. Genetic factors, such as polymorphisms that encode for variant enzymes with altered properties in terms of metal binding, transport, and effects, also may modify the impact of metals on health and thereby account, at least in part, for individual susceptibility to metal effects.

The most important component of treatment for metal toxicity is the termination of exposure. *Chelating agents* are used to bind metals into stable cyclic compounds with relatively low toxicity and to enhance their excretion. The principal chelating agents are dimercaprol (British anti-Lewisite [BAL]), ethylenediamine tetraacetic acid (EDTA), succimer (dimercaptosuccinic acid [DMSA]), and penicillamine; their specific use depends on the metal involved and the clinical circumstances. Activated charcoal does not bind metals and thus is of limited usefulness in cases of acute metal ingestion.

In addition to the information provided in Table 449-1, several other aspects of exposure, toxicity, or management are worthy of discussion with respect to the four most hazardous toxicants (arsenic, cadmium, lead, and mercury).

Arsenic, even at moderate levels of exposure, has been clearly linked with increased risks for cancer of the skin, bladder, renal pelvis, ureter, kidney, liver, and lung. These risks appear to be modified by smoking, folate and selenium status, genetic traits (such as ability to methylate arsenic), and other factors. Recent studies in community-based populations have generated strong evidence that arsenic exposure is a risk factor for increased coronary heart disease and stroke, lung function impairment, acute respiratory tract

infections, respiratory symptoms, and non-malignant lung disease mortality. Evidence is also emerging that low-level arsenic may cause neurodevelopmental delays in children and diabetes.

Serious *cadmium* poisoning from the contamination of food and water by mining effluents in Japan contributed to the 1946 outbreak of “itai-itai” (“ouch-ouch”) disease, so named because of cadmium-induced bone toxicity that led to painful bone fractures. Modest exposures from environmental contamination have been associated in some studies with a lower bone density, a higher incidence of fractures, and a faster decline in height in both men and women, effects that may be related to cadmium’s calciuric effect on the kidney. There is some evidence for synergy between the adverse impacts of cadmium and lead on kidney function. Environmental exposures have also been linked to lower lung function (even after adjusting for smoking cigarettes, which contain cadmium) as well as increased risk of cardiovascular disease and mortality, stroke, and heart failure. Several studies have also raised concerns that cadmium may be carcinogenic and contribute to elevated risks of prostate, breast, and pancreatic cancer. Overall, this growing body of research indicates that cadmium exposure may be contributing significantly to morbidity and mortality rates in the general population.

Advances in our understanding of *lead* toxicity have recently benefited by the development of K x-ray fluorescence (KXRF) instruments for making safe in vivo measurements of lead levels in bone (which, in turn, reflect cumulative exposure over many years, as opposed to blood lead levels, which mostly reflect recent exposure). Higher bone lead levels measured by KXRF have been linked to increased risk of hypertension and accelerated declines in cognition in both men and women from an urban population. Upon reviewing these studies in conjunction with other epidemiologic and toxicologic studies, a recent federal expert panel concluded that the impact of lead exposure on hypertension and cognition in adults was causal. Prospective studies have also demonstrated that higher bone lead levels are a major risk factor for increased cardiovascular morbidity and mortality rates in both community-based and occupational-exposed populations. Lead exposure at community levels has also been recently associated with increased risks of hearing loss, Parkinson’s disease, and amyotrophic lateral sclerosis. With respect to pregnancy-associated risks, high maternal bone lead levels were found to predict lower birth weight, head circumference, birth length, and neurodevelopmental performance in offspring by age 2 years. Offspring have also been shown to have higher blood pressures at age 7–14 years, an age range at which higher blood pressures are known to predict an elevated risk of developing hypertension. In a randomized trial, calcium supplementation (1200 mg daily) was found to significantly reduce the mobilization of lead from maternal bone into blood during pregnancy.

The toxicity of low-level organic *mercury* exposure (as manifested by neurobehavioral performance) is of increasing concern based on studies of the offspring of mothers who ingested mercury-contaminated fish. With respect to whether the consumption of fish by women during pregnancy is good or bad for offspring neurodevelopment, balancing the trade-offs of the beneficial effects of the omega-3-fatty acids (FAs) in fish versus the adverse effects of mercury contamination in fish has led to some confusion and inconsistency in public health recommendations. Overall, it would appear that it would be best for pregnant women to either limit fish consumption to those species known to be low in mercury contamination but high in omega-3-FAs

(such as sardines or mackerel) or to avoid fish and obtain omega-3-FAs through supplements or other dietary sources. Accumulated evidence has not supported the contention that ethyl mercury, used as a preservative in multiuse vaccines administered in early childhood, has played a significant role in causing neurodevelopmental problems such as autism. With regard to adults, there is conflicting evidence as to whether mercury exposure is associated with increased risk of hypertension and cardiovascular disease. At this point, conclusions cannot be drawn. Mercury exposure may also be associated with perturbations in markers of autoimmunity. The clinical significance of these findings remains unclear.

 Heavy metals pose risks to health that are especially burdensome in selected parts of the world. For example, *arsenic* exposure from natural contamination of shallow tube wells inserted for drinking water is a major environmental problem for millions of residents in parts of Bangladesh and Western India. Contamination was formerly considered only a problem with deep wells; however, the geology of this region allows most residents only a few alternatives for potable drinking water. The combustion of leaded gasoline with resulting contamination of air and soil with *lead oxide* remains a problem in some countries of Central Asia, Southeast Asia, Africa, and the Middle East. Populations living in the Arctic have been shown to have particularly high exposures to *mercury* due to long-range transport patterns that concentrate mercury in the polar regions, as well as the traditional dependence of Arctic peoples on the consumption of fish and other wildlife that bioconcentrate methylmercury.

A few additional metals deserve brief mention but are not covered in Table 449-1 because of the relative rarity of their being clinically encountered or the uncertainty regarding their potential toxicities. *Aluminum* contributes to the encephalopathy in patients with severe renal disease, who are undergoing dialysis **(Chap. 403)**. High levels of aluminum are found in the neurofibrillary tangles in the cerebral cortex and hippocampus of patients with Alzheimer's disease, as well as in the drinking water and soil of areas with an unusually high incidence of Alzheimer's. The experimental and epidemiologic evidence for the aluminum–Alzheimer's disease link remains relatively weak, however, and it cannot be concluded that aluminum is a causal agent or a contributing factor in neurodegenerative disease. Hexavalent chromium is corrosive and sensitizing. Workers in the chromate and chrome pigment production industries have consistently had a greater risk of lung cancer. The introduction of *cobalt* chloride as a fortifier in beer led to outbreaks of fatal cardiomyopathy among heavy consumers. Occupational exposure (e.g., of miners, dry-battery manufacturers, and arc welders) to *manganese* (Mn) can cause a parkinsonian syndrome within 1–2 years, including gait disorders; postural instability; a masked, expressionless face; tremor; and psychiatric symptoms. With the introduction of methylcyclopentadienyl manganese tricarbonyl (MMT) as a gasoline additive, there is concern for the toxic potential of environmental manganese exposure. For example, a recent study found a high prevalence of parkinsonian disorders in a community with risks proportionate to estimated manganese exposures emitted by local ferroalloy industries. Epidemiologic studies have also suggested that manganese may interfere with early childhood neurodevelopment in ways similar to that of lead. Manganese toxicity is clearly associated with dopaminergic dysfunction and its toxicity is likely influenced by age, gender, ethnicity, genetics, and preexisting medical conditions. *Nickel* exposure induces an allergic response, and inhalation of nickel compounds with low aqueous solubility (e.g., nickel subsulfide and nickel oxide) in occupational settings is associated with an increased risk of lung cancer. Overexposure to selenium may cause local irritation of the

respiratory system and eyes, gastrointestinal irritation, liver inflammation, loss of hair, depigmentation, and peripheral nerve damage. Workers exposed to certain organic forms of *tin* (particularly trimethyl and triethyl derivatives) have developed psychomotor disturbances, including tremor, convulsions, hallucinations, and psychotic behavior.

Thallium, which is a component of some insecticides, metal alloys, and fireworks, is absorbed through the skin as well as by ingestion and inhalation. Severe poisoning follows a single ingested dose of >1 g or >8 mg/kg. Nausea and vomiting, abdominal pain, and hematemesis precede confusion, psychosis, organic brain syndrome, and coma. Thallium is radiopaque. Induced emesis or gastric lavage is indicated within 4–6 h of acute ingestion; Prussian blue prevents absorption and is given orally at 250 mg/kg in divided doses. Unlike other types of metal poisoning, thallium poisoning may be less severe when activated charcoal is used to interrupt its enterohepatic circulation. Other measures include forced diuresis, treatment with potassium chloride (which promotes renal excretion of thallium), and peritoneal dialysis.

Chelation Therapy The Trial to Assess Chelation Therapy (TACT), a multi-center double-blind, placebo-controlled, prospective randomized trial funded by NIH of 1708 patients aged ≥50 years who had experienced a myocardial infarction (MI), recently found that a protocol of repeated intravenous chelation with disodium EDTA, compared with placebo, modestly but significantly reduced the risk of adverse cardiovascular outcomes, many of which were revascularization procedures. The effect was particularly pronounced among those with concurrent diabetes. However, the trial did not include rigorous measures of lead exposure or any selection criteria based on lead exposure; thus, even though chelation reduces metal burdens, which have been associated with adverse cardiovascular effects (especially lead), it remains unclear whether the beneficial effects result from a reduction in metal burden. In view of the risks of side effects associated with chelation, by themselves, the results are not sufficient to support the routine use of chelation therapy for treatment of patients either who have had an MI or who have had low-level lead exposure. A follow-up trial with rigorous measures of metals exposure is planned.

FURTHER READING

Alamolhodaei NS et al: Arsenic cardiotoxicity: An overview. Environ Toxicol Pharmacol 40:1005, 2015. [PubMed: 26606645]

Aneni EC et al: Chronic toxic metal exposure and cardiovascular disease: Mechanisms of risk and emerging role of chelation therapy. Curr Atheroscler Rep 18:81, 2016. [PubMed: 27822681]

Cheng X et al: Cadmium exposure and risk of any fracture: A PRISMA-compliant systematic review and meta-analysis. Medicine (Baltimore) 95:e2932, 2016. [PubMed: 26962791]

Gidlow DA: Lead toxicity. Occup Med (Lond) 65:348, 2015. [PubMed: 26187801]

Kim KH et al: A review on the distribution of Hg in the environment and its human health impacts. J Hazard Mater 306:376, 2016. [PubMed: 26826963]

Lucchini RG, Hashim D: Tremor secondary to neurotoxic exposure: Mercury, lead, solvents, pesticides. Handb Clin Neurol 131:241, 2015. [PubMed: 26563793]

O'Neal SL, Zheng W: Manganese toxicity upon overexposure: A decade in review. Curr Environ Health Rep 2:315, 2015. [PubMed: 26231508]

Chapter 450: Poisoning and Drug Overdose

Mark B. Mycyk

INTRODUCTION

Poisoning refers to the development of dose-related adverse effects following exposure to chemicals, drugs, or other xenobiotics. To paraphrase Paracelsus, the dose makes the poison. Although most poisons have predictable dose-related effects, individual responses to a given dose may vary because of genetic polymorphism, enzymatic induction or inhibition in the presence of other xenobiotics, or acquired tolerance. Poisoning may be local (e.g., skin, eyes, or lungs) or systemic depending on the route of exposure, the chemical and physical properties of the poison, and its mechanism of action. The severity and reversibility of poisoning also depend on the functional reserve of the individual or target organ, which is influenced by age and preexisting disease.

EPIDEMIOLOGY

More than 5 million poison exposures occur in the United States each year. Most are acute, are accidental (unintentional), involve a single agent, occur in the home, result in minor or no toxicity, and involve children <6 years of age. Pharmaceuticals are involved in 47% of exposures and in 84% of serious or fatal poisonings. In the last decade, the rate of injury-related deaths from poisoning has overtaken the rate of deaths related to motor-vehicle crashes in the United States. According to the Centers for Disease Control (CDC), twice as many Americans died from drug overdoses in 2014 compared to 2000. Although prescription opioids have appropriately received attention as a major reason for the increased number of poisoning deaths, the availability of other pharmaceuticals and rapid proliferation of novel drugs of abuse also contribute to the increasing death rate. In many parts of the United States, where these issues are particularly prevalent, there are efforts to develop better prescription drug databases and enhanced training for health care professionals in pain management and the use of opiates. Unintentional exposures can result from the improper use of chemicals at work or play; label misreading; product mislabeling; mistaken identification of unlabeled chemicals; uninformed self-medication; and dosing errors by nurses, pharmacists, physicians, parents, and the elderly. Excluding the recreational use of ethanol, attempted suicide (deliberate self-harm) is the most common reported reason for intentional poisoning. Recreational use of prescribed and over-the-counter drugs for psychotropic or euphoric effects (*abuse*) or excessive self-dosing (*misuse*) is increasingly common and may also result in unintentional self-poisoning.

About 20–25% of exposures require bedside health-professional evaluation, and 5% of all exposures require hospitalization. Poisonings account for 5–10% of all ambulance transports, emergency department visits, and

intensive care unit admissions. Hospital admissions related to poisoning are also associated with longer lengths of stay and increase the utilization of resources such as radiography and other laboratory services. Up to 30% of psychiatric admissions are prompted by attempted suicide via overdosage. Overall, the mortality rate is low: <1% of all poisoning exposures. It is significantly higher (1–2%) among hospitalized patients with intentional (suicidal) overdose or complications from drugs of abuse, who account for the majority of serious poisonings. Acetaminophen is the pharmaceutical agent most often implicated in fatal poisoning. Overall, carbon monoxide is the leading cause of death from poisoning, but this prominence is not reflected in hospital or poison center statistics because patients with such poisoning are typically dead when discovered and are referred directly to medical examiners.

DIAGNOSIS

Although poisoning can mimic other illnesses, the correct diagnosis can usually be established by the history, physical examination, routine and toxicologic laboratory evaluations, and characteristic clinical course.

HISTORY

The *history* should include the time, route, duration, and circumstances (location, surrounding events, and intent) of exposure; the name and amount of each drug, chemical, or ingredient involved; the time of onset, nature, and severity of symptoms; the time and type of first-aid measures provided; and the medical and psychiatric history.

In many cases the patient is confused, comatose, unaware of an exposure, or unable or unwilling to admit to one. Suspicious circumstances include unexplained sudden illness in a previously healthy person or a group of healthy people; a history of psychiatric problems (particularly depression); recent changes in health, economic status, or social relationships; and onset of illness during work with chemicals or after ingestion of food, drink (especially ethanol), or medications. When patients become ill soon after arriving from a foreign country or being arrested for criminal activity, “body packing” or “body stuffing” (ingesting or concealing illicit drugs in a body cavity) should be suspected. Relevant information may be available from family, friends, paramedics, police, pharmacists, physicians, and employers, who should be questioned regarding the patient’s habits, hobbies, behavioral changes, available medications, and antecedent events. Patients need to be asked explicitly about their prescribed medications and recreational drug use. Drugs previously considered “illicit” such as cannabinoids are now legal in many places and prescribed for therapeutic purposes. A search of clothes, belongings, and place of discovery may reveal a suicide note or a container of drugs or chemicals. Without a clear history in a patient clinically suspected to be poisoned, all medications available anywhere in the patient’s home or belongings should be considered as possible agents, including medications for pets. The imprint code on pills and the label on chemical products may be used to identify the ingredients and potential toxicity of a suspected poison by consulting a reference text, a computerized database, the manufacturer, or a regional poison information center (800-222-1222). Occupational exposures require review of any available safety data sheet (SDS) from the worksite. Because of increasing globalization from travel and internet consumerism, unfamiliar poisonings may result in local emergency department evaluation. Pharmaceuticals, industrial

chemicals, or drugs of abuse from foreign countries may be identified with the assistance of a regional poison center or via the World Wide Web.

PHYSICAL EXAMINATION AND CLINICAL COURSE

The *physical examination* should focus initially on vital signs, the cardiopulmonary system, and neurologic status. The neurologic examination should include documentation of neuromuscular abnormalities such as dyskinesia, dystonia, fasciculations, myoclonus, rigidity, and tremors. The patient should also be examined for evidence of trauma and underlying illnesses. Focal neurologic findings are uncommon in poisoning, and their presence should prompt evaluation for a structural central nervous system (CNS) lesion. Examination of the eyes (for nystagmus and pupil size and reactivity), abdomen (for bowel activity and bladder size), and skin (for burns, bullae, color, warmth, moisture, pressure sores, and puncture marks) may reveal findings of diagnostic value. When the history is unclear, all orifices should be examined for the presence of chemical burns and drug packets. The odor of breath or vomitus and the color of nails, skin, or urine may provide important diagnostic clues.

The diagnosis of poisoning in cases of unknown etiology primarily relies on pattern recognition. The first step is to assess the pulse, blood pressure, respiratory rate, temperature, and neurologic status and to characterize the overall physiologic state as stimulated, depressed, discordant, or normal **(Table 450-1)**. Obtaining a complete set of vital signs and reassessing them frequently are critical. Measuring core temperature is especially important, even in difficult or combative patients, since temperature elevation is the most reliable prognosticator of poor outcome in poisoning from stimulants (e.g., cocaine) or drug withdrawal (e.g., alcohol or GHB). The next step is to consider the underlying causes of the physiologic state and to attempt to identify a pathophysiologic pattern or toxic syndrome (*toxidrome*) based on the observed findings. Assessing the severity of physiologic derangements **(Table 450-2)** is useful in this regard and also for monitoring the clinical course and response to treatment. The final step is to attempt to identify the particular agent involved by looking for unique or relatively poison-specific physical or ancillary test abnormalities. Distinguishing among toxidromes on the basis of the physiologic state is summarized next.

TABLE 450-1

Differential Diagnosis of Poisoning Based on Physiologic State

STIMULATED	DEPRESSED	DISCORDANT	NORMAL
Sympathetics Sympathomimetics Ergot alkaloids Methylxanthines Monoamine oxidase inhibitors Thyroid hormones Anticholinergics Antihistamines Antiparkinsonian agents Antipsychotics Antispasmodics Belladonna alkaloids Cyclic antidepressants Mushrooms and plants Hallucinogens Cannabinoids (marijuana) LSD and analogues Mescaline and analogues Mushrooms Phencyclidine and analogues Withdrawal syndromes Barbiturates Benzodiazepines	Sympatholytics α_1-Adrenergic antagonists α_2-Adrenergic agonists ACE inhibitors Angiotensin receptor blockers Antipsychotics β-Adrenergic blockers Calcium channel blockers Cardiac glycosides Cyclic antidepressants Cholinergics Acetylcholinesterase inhibitors Muscarinic agonists Nicotinic agonists Opioids Analgesics GI antispasmodics Heroin Sedative-hypnotics Alcohols Anticonvulsants Barbiturates Benzodiazepines GABA precursors Muscle relaxants	Asphyxiants Cytochrome oxidase inhibitors Inert gases Irritant gases Methemoglobin inducers Oxidative phosphorylation inhibitors AGMA inducers Alcohol (ketoacidosis) Ethylene glycol Iron Methanol Other alcohols Salicylate Toluene CNS syndromes Extrapyramidal reactions Hydrocarbon inhalation Isoniazid Lithium Neuroleptic malignant syndrome Serotonin syndrome Strychnine Membrane-active agents Amantadine Antiarrhythmics Antihistamines	Nontoxic exposure Psychogenic illness “Toxic time-bombs” Slow absorption Anticholinergics Carbamazepine Concretion formers Extended-release phenytoin sodium capsules (Dilantin Kapseals) Drug packets Enteric-coated pills Diphenoxylate-atropine (Lomotil) Opioids Salicylates Sustained-release pills Valproate Slow distribution Cardiac glycosides Lithium Metals Salicylate Valproate Toxic metabolite Acetaminophen Carbon tetrachloride Cyanogenic glycosides Ethylene glycol Methanol Methemoglobin inducers Mushroom toxins Organophosphate insecticides Paraquat Metabolism disruptors Antineoplastic agents Antiviral agents Colchicine

STIMULATED	DEPRESSED	DISCORDANT	NORMAL
Ethanol GHB products Opioids Sedative-hypnotics Sympatholytics	Other agents GHB products	Antipsychotics Carbamazepine Cyclic antidepressants Local anesthetics Opioids (some) Quinoline antimalarials	Hypoglycemic agents Immunosuppressive agents MAO inhibitors Metals Other oral anticoagulants Salicylate Warfarin

Abbreviations: ACE, angiotensin-converting enzyme; AGMA, anion-gap metabolic acidosis; CNS, central nervous system; GABA, γ-aminobutyric acid; GHB, γ-hydroxybutyrate; GI, gastrointestinal; LSD, lysergic acid diethylamide; MAO, monoamine oxidase.

TABLE 450-2

Severity of Physiologic Stimulation and Depression in Poisoning and Drug Withdrawal

Physiologic Stimulation	
Grade 1	Anxious, irritable, tremulous; vital signs normal; diaphoresis, flushing or pallor, mydriasis, and hyperreflexia sometimes present
Grade 2	Agitated; may have confusion or hallucinations but can converse and follow commands; vital signs mildly to moderately increased
Grade 3	Delirious; unintelligible speech, uncontrollable motor hyperactivity; moderately to markedly increased vital signs; tachyarrhythmias possible
Grade 4	Coma, seizures, cardiovascular collapse
Physiologic Depression	
Grade 1	Awake, lethargic, or sleeping but arousable by voice or tactile stimulation; able to converse and follow commands; may be confused
Grade 2	Responds to pain but not voice; can vocalize but not converse; spontaneous motor activity present; brainstem reflexes intact
Grade 3	Unresponsive to pain; spontaneous motor activity absent; brainstem reflexes depressed; motor tone, respirations, and temperature decreased
Grade 4	Unresponsive to pain; flaccid paralysis; brainstem reflexes and respirations absent; cardiovascular vital signs decreased

The Stimulated Physiologic State

Increased pulse, blood pressure, respiratory rate, temperature, and neuromuscular activity characterize the *stimulated* physiologic state, which can reflect sympathetic, anticholinergic, or hallucinogen poisoning or drug withdrawal (Table 450-1). Other features are noted in (Table 450-2). Mydriasis, a characteristic feature of all stimulants, is most marked in anticholinergic poisoning since pupillary reactivity relies on muscarinic control. In sympathetic poisoning (e.g., due to cocaine), pupils are also enlarged, but some reactivity to light remains. The anticholinergic toxidrome is also distinguished by hot, dry, flushed skin; decreased bowel sounds; and urinary retention. Other stimulant syndromes increase sympathetic activity and cause diaphoresis, pallor, and increased bowel activity with varying degrees of nausea, vomiting, abnormal distress, and occasionally diarrhea. The

absolute and relative degree of vital-sign changes and neuromuscular hyperactivity can help distinguish among stimulant toxidromes. Since sympathetics stimulate the peripheral nervous system more directly than do hallucinogens or drug withdrawal, markedly increased vital signs and organ ischemia suggest sympathetic poisoning. Findings helpful in suggesting the particular drug or class causing physiologic stimulation include reflex bradycardia from selective α-adrenergic stimulants (e.g., decongestants), hypotension from selective β-adrenergic stimulants (e.g., asthma therapeutics), limb ischemia from ergot alkaloids, rotatory nystagmus from phencyclidine and ketamine (the only physiologic stimulants that cause this finding), and delayed cardiac conduction from high doses of cocaine and some anticholinergic agents (e.g., antihistamines, cyclic antidepressants, and antipsychotics). Seizures suggest a sympathetic etiology, an anticholinergic agent with membrane-active properties (e.g., cyclic antidepressants, phenothiazines), or a withdrawal syndrome. Close attention to core temperature is critical in patients with grade 4 physiologic stimulation (Table 450-2).

The Depressed Physiologic State

Decreased pulse, blood pressure, respiratory rate, temperature, and neuromuscular activity are indicative of the *depressed* physiologic state caused by "functional" sympatholytics (agents that decrease cardiac function and vascular tone as well as sympathetic activity), cholinergic (muscarinic and nicotinic) agents, opioids, and sedative-hypnotic γ-aminobutyric acid (GABA)-ergic agents (Tables 450-1 and 450-2). Miosis is also common and is most pronounced in opioid and cholinergic poisoning. Miosis is distinguished from other depressant syndromes by muscarinic and nicotinic signs and symptoms (Table 450-1). Pronounced cardiovascular depression in the absence of significant CNS depression suggests a direct or peripherally acting sympatholytic. In contrast, in opioid and sedative-hypnotic poisoning, vital-sign changes are secondary to depression of CNS cardiovascular and respiratory centers (or consequent hypoxemia), and significant abnormalities in these parameters do not occur until there is a marked decrease in the level of consciousness (grade 3 or 4 physiologic depression; [Table 450-2]). Other clues that suggest the cause of physiologic depression include cardiac arrhythmias and conduction disturbances (due to antiarrhythmics, β-adrenergic antagonists, calcium channel blockers, digitalis glycosides, propoxyphene, and cyclic antidepressants), mydriasis (due to tricyclic antidepressants, some antiarrhythmics, meperidine, and diphenoxylate-atropine [Lomotil]), nystagmus (due to sedative-hypnotics), and seizures (due to cholinergic agents, propoxyphene, and cyclic antidepressants).

The Discordant Physiologic State

The *discordant* physiologic state is characterized by mixed vital-sign and neuromuscular abnormalities, as observed in poisoning by asphyxiants, CNS syndromes, membrane-active agents, and anion-gap metabolic acidosis (AGMA) inducers (Table 450-1). In these conditions, manifestations of physiologic stimulation and physiologic depression occur together or at different times during the clinical course. For example, membrane-active agents can cause simultaneous coma, seizures, hypotension, and tachyarrhythmias. Alternatively, vital signs may be normal while the patient has an altered mental status or is obviously sick or clearly symptomatic. Early, pronounced vital-sign and mental-status changes suggest asphyxiant or membrane-active agent poisoning; the lack of such abnormalities suggests an AGMA inducer; and marked neuromuscular dysfunction without significant vital-sign abnormalities suggests a CNS syndrome. The *discordant* physiologic state may also be evident in patients poisoned with multiple agents.

The Normal Physiologic State

A *normal* physiologic status and physical examination may be due to a nontoxic exposure, psychogenic illness, or poisoning by "toxic time-bombs": agents that are slowly absorbed, are slowly distributed to their sites of action, require metabolic activation, or disrupt metabolic processes (Table 450-1). Because so many medications have now been reformulated into a once-a-day preparations for the patient's convenience and adherence, toxic time-bombs are increasingly common. Diagnosing a nontoxic exposure requires that the identity of the exposure agent be known or that a toxic time-bomb exposure be excluded and the time since exposure exceed the longest known or predicted interval between exposure and peak toxicity. Psychogenic illness (fear of being poisoned, mass hysteria) may also follow a nontoxic exposure and should be considered when symptoms are inconsistent with exposure history. Anxiety reactions resulting from a nontoxic exposure can cause mild physiologic stimulation (Table 450-2) and be indistinguishable from toxicologic causes without ancillary testing or a suitable period of observation.

LABORATORY ASSESSMENT

Laboratory assessment may be helpful in the differential diagnosis. Increased anion gap metabolic acidosis (AGMA) is most common in advanced methanol, ethylene glycol, and salicylate intoxication but can occur with any poisoning that results in hepatic, renal, or respiratory failure; seizures; or shock. The serum lactate concentration is more commonly low (less than the anion gap) in the former and high (nearly equal to the anion gap) in the latter. An abnormally low anion gap can be due to elevated blood levels of bromide, calcium, iodine, lithium, or magnesium. An increased osmolal gap—a difference of >10 mmol/L between serum osmolality (measured by freezing-point depression) and osmolality calculated from serum sodium, glucose, and blood urea nitrogen levels—suggests the presence of a low-molecular-weight solute such as acetone; an alcohol (benzyl, ethanol, isopropanol, methanol); a glycol (diethylene, ethylene, propylene); ether (ethyl, glycol); or an "unmeasured" cation (calcium, magnesium) or sugar (glycerol, mannitol, sorbitol). Ketosis suggests acetone, isopropyl alcohol, salicylate poisoning, or alcoholic ketoacidosis. Hypoglycemia may be due to poisoning with β-adrenergic blockers, ethanol, insulin, oral hypoglycemic agents, quinine, and salicylates, whereas hyperglycemia can occur in poisoning with acetone, β-adrenergic agonists, caffeine, calcium channel blockers, iron, theophylline, or *N*-3-pyridylmethyl-*N*′-*p*-nitrophenylurea (PNU [Vacor]). Hypokalemia can be caused by barium, β-adrenergic agonists, caffeine, diuretics, theophylline, or toluene; hyperkalemia suggests poisoning with an α-adrenergic agonist, a β-adrenergic blocker, cardiac glycosides, or fluoride. Hypocalcemia may be seen in ethylene glycol, fluoride, and oxalate poisoning. PT and INR are useful for risk stratification in cases of warfarin or rodenticide poisoning, but are not to be relied on when evaluating overdose or complications from new anticoagulant pharmaceuticals (e.g., dabigatran).

The *electrocardiogram* (ECG) can be useful for rapid diagnostic purposes. Bradycardia and atrioventricular block may occur in patients poisoned by α-adrenergic agonists, antiarrhythmic agents, beta blockers, calcium channel blockers, cholinergic agents (carbamate and organophosphate insecticides), cardiac glycosides, lithium, or tricyclic antidepressants. QRS- and QT-interval prolongation may be caused by hyperkalemia, various antidepressants, and other membrane-active drugs (Table 450-1). Ventricular tachyarrhythmias may be seen in poisoning with cardiac glycosides, fluorides, membrane-active drugs, methylxanthines, sympathomimetics,

antidepressants, and agents that cause hyperkalemia or potentiate the effects of endogenous catecholamines (e.g., chloral hydrate, aliphatic and halogenated hydrocarbons).

Radiologic studies may occasionally be useful. Pulmonary edema (adult respiratory distress syndrome [ARDS]) can be caused by poisoning with carbon monoxide, cyanide, an opioid, paraquat, phencyclidine, a sedative-hypnotic, or salicylate; by inhalation of irritant gases, fumes, or vapors (acids and alkali, ammonia, aldehydes, chlorine, hydrogen sulfide, isocyanates, metal oxides, mercury, phosgene, polymers); or by prolonged anoxia, hyperthermia, or shock. Aspiration pneumonia is common in patients with coma, seizures, and petroleum distillate aspiration. Chest x-ray is useful for identifying complications from metal fume fever or elemental mercury. The presence of radiopaque densities on abdominal x-rays or abdominal CT scan suggests the ingestion of calcium salts, chloral hydrate, chlorinated hydrocarbons, heavy metals, illicit drug packets, iodinated compounds, potassium salts, enteric-coated tablets, or salicylates.

Toxicologic analysis of urine and blood (and occasionally of gastric contents and chemical samples) can sometimes confirm or rule out suspected poisoning. Interpretation of laboratory data requires knowledge of the qualitative and quantitative tests used for screening and confirmation (enzyme-multiplied, fluorescence polarization, and radio-immunoassays; colorimetric and fluorometric assays; thin-layer, gas-liquid, or high-performance liquid chromatography; gas chromatography; mass spectrometry), their sensitivity (limit of detection) and specificity, the preferred biologic specimen for analysis, and the optimal time of specimen sampling. Personal communication with the hospital laboratory is essential to an understanding of institutional testing capabilities and limitations.

Rapid *qualitative* hospital-based urine tests for drugs of abuse are only screening tests that cannot confirm the exact identity of the detected substance and should not be considered diagnostic or used for forensic purposes: False-positive and false-negative results are common. A positive screen may result from other pharmaceuticals that interfere with laboratory analysis (e.g., fluoroquinolones commonly cause “false-positive” opiate screens). Confirmatory testing with gas chromatography/mass spectrometry can be requested, but it often takes weeks to obtain a reported result. A negative screening result may mean that the responsible substance is not detectable by the test used or that its concentration is too low for detection at the time of sampling. For instance, recent new drugs of abuse that often result in emergency department evaluation for unexpected complications, such as synthetic cannabinoids (spice), cathinones (bath salts), and opiate substitutes (kratom), are not detectable by hospital-based tests. In cases where a drug concentration is too low to be detected early during clinical evaluation, repeating the test at a later time may yield a positive result. Patients symptomatic from drugs of abuse often require immediate management based on the history, physical examination, and observed toxidrome without laboratory confirmation (e.g., apnea from opioid intoxication). When the patient is asymptomatic or when the clinical picture is consistent with the reported history, qualitative screening is neither clinically useful nor cost-effective. Thus, qualitative drug screens are of greatest value for the evaluation of patients with severe or unexplained toxicities, such as coma, seizures, cardiovascular instability, metabolic or respiratory acidosis, and nonsinus cardiac rhythms. In contrast to qualitative drug screens, *quantitative* serum tests are useful for evaluation of patients poisoned with acetaminophen **(Chap. 333)**, alcohols (including ethylene glycol and methanol), anticonvulsants, barbiturates, digoxin, heavy metals, iron, lithium, salicylate, and theophylline as well as for the presence of carboxyhemoglobin and methemoglobin. The serum concentration in these cases guides clinical management, and results are often available within an hour.

The *response to antidotes* is sometimes useful for diagnostic purposes. Resolution of altered mental status and abnormal vital signs within minutes of IV administration of dextrose, naloxone, or flumazenil is virtually diagnostic of hypoglycemia, opioid poisoning, and benzodiazepine intoxication, respectively. The prompt reversal of dystonic (extrapyramidal) signs and symptoms following an IV dose of benztropine or diphenhydramine confirms a drug etiology. Although complete reversal of both central and peripheral manifestations of anticholinergic poisoning by physostigmine is diagnostic of this condition, physostigmine may cause some arousal in patients with CNS depression of any etiology.

TREATMENT

TREATMENT

Poisoning and Drug Overdose

GENERAL PRINCIPLES

Treatment goals include support of vital signs, prevention of further poison absorption (decontamination), enhancement of poison elimination, administration of specific antidotes, and prevention of reexposure **(Table 450-3)**. Specific treatment depends on the identity of the poison, the route and amount of exposure, the time of presentation relative to the time of exposure, and the severity of poisoning. Knowledge of the offending agents' pharmacokinetics and pharmacodynamics is essential.

During the *pretoxic phase*, prior to the onset of poisoning, decontamination is the highest priority, and treatment is based solely on the history. The maximal potential toxicity based on the greatest possible exposure should be assumed. Since decontamination is more effective when accomplished soon after exposure and when the patient is asymptomatic, the initial history and physical examination should be focused and brief. It is also advisable to establish IV access and initiate cardiac monitoring, particularly in patients with potentially serious ingestions or unclear histories.

When an accurate history is not obtainable and a poison causing delayed toxicity (i.e., a toxic time-bomb) or irreversible damage is suspected, blood and urine should be sent for appropriate toxicologic screening and quantitative analysis. During poison absorption and distribution, blood levels may be greater than those in tissue and may not correlate with toxicity. However, high blood levels of agents whose metabolites are more toxic than the parent compound (acetaminophen, ethylene glycol, or methanol) may indicate the need for additional interventions (antidotes, dialysis). Most patients who remain asymptomatic or who become asymptomatic 6 h after ingestion are unlikely to develop subsequent toxicity and can be discharged safely. Longer observation will be necessary for patients who have ingested toxic time-bombs.

During the *toxic phase*—the interval between the onset of poisoning and its peak effects—management is based primarily on clinical and laboratory findings. *Effects after an overdose usually begin sooner, peak later, and last longer than they do after a therapeutic dose*. A drug's published pharmacokinetic profile in standard references such as the *Physician's Desk Reference* (PDR) is usually different from its toxicokinetic profile in overdose. Resuscitation and stabilization are the first priority. Symptomatic patients should have an IV line placed and should undergo oxygen saturation determination, cardiac monitoring, and continuous observation. Baseline laboratory, ECG, and x-ray evaluation may also be appropriate. Intravenous glucose (unless the serum level is

documented to be normal), naloxone, and thiamine should be considered in patients with altered mental status, particularly those with coma or seizures. Decontamination should also be considered, but it is less likely to be effective during this phase than during the pretoxic phase.

Measures that enhance poison elimination may shorten the duration and severity of the toxic phase. However, they are not without risk, which must be weighed against the potential benefit. Diagnostic certainty (usually via laboratory confirmation) is generally a prerequisite. Intestinal (gut) dialysis with repetitive doses of activated charcoal (see "Multiple-Dose Activated Charcoal," later) can enhance the elimination of selected poisons such as theophylline or carbamazepine. Urinary alkalinization may enhance the elimination of salicylates and a few other poisons. Chelation therapy can enhance the elimination of selected metals. Extracorporeal elimination methods are effective for many poisons, but their expense and risk make their use reasonable only in patients who would otherwise have an unfavorable outcome.

During the *resolution phase* of poisoning, supportive care and monitoring should continue until clinical, laboratory, and ECG abnormalities have resolved. Since chemicals are eliminated sooner from the blood than from tissues, blood levels are usually lower than tissue levels during this phase and again may not correlate with toxicity. This discrepancy applies particularly when extracorporeal elimination procedures are used. Redistribution from tissues may cause a rebound increase in the blood level after termination of these procedures (e.g., lithium). When a metabolite is responsible for toxic effects, continued treatment may be necessary in the absence of clinical toxicity or abnormal laboratory studies.

SUPPORTIVE CARE

The goal of supportive therapy is to maintain physiologic homeostasis until detoxification is accomplished and to prevent and treat secondary complications such as aspiration, bedsores, cerebral and pulmonary edema, pneumonia, rhabdomyolysis, renal failure, sepsis, thromboembolic disease, coagulopathy, and generalized organ dysfunction due to hypoxemia or shock.

Admission to an intensive care unit is indicated for the following: patients with severe poisoning (coma, respiratory depression, hypotension, cardiac conduction abnormalities, cardiac arrhythmias, hypothermia or hyperthermia, seizures); those needing close monitoring, antidotes, or enhanced elimination therapy; those showing progressive clinical deterioration; and those with significant underlying medical problems. Patients with mild to moderate toxicity can be managed on a general medical service, on an intermediate care unit, or in an emergency department observation area, depending on the anticipated duration and level of monitoring needed (intermittent clinical observation versus continuous clinical, cardiac, and respiratory monitoring). Patients who have attempted suicide require continuous observation and measures to prevent self-injury until they are no longer suicidal.

Respiratory Care

Endotracheal intubation for protection against the aspiration of gastrointestinal contents is of paramount importance in patients with CNS depression or seizures as this complication can increase morbidity and mortality rates. Mechanical ventilation may be necessary for patients with respiratory depression or hypoxemia and for facilitation of therapeutic sedation or paralysis of patients in order to prevent or treat hyperthermia, acidosis, and rhabdomyolysis associated with neuromuscular hyperactivity. Since clinical assessment of

respiratory function can be inaccurate, the need for oxygenation and ventilation is best determined by continuous pulse oximetry or arterial blood-gas analysis. The gag reflex is not a reliable indicator of the need for intubation. A patient with CNS depression may maintain airway patency while being stimulated but not if left alone. Drug-induced pulmonary edema is usually noncardiac rather than cardiac in origin, although profound CNS depression and cardiac conduction abnormalities suggest the latter. Measurement of pulmonary artery pressure may be necessary to establish the cause and direct appropriate therapy. Extracorporeal measures (membrane oxygenation, ECMO, venoarterial perfusion, cardiopulmonary bypass) and partial liquid (perfluorocarbon) ventilation may be appropriate for severe but reversible respiratory failure.

Cardiovascular Therapy

Maintenance of normal tissue perfusion is critical for complete recovery to occur once the offending agent has been eliminated. Focused bedside echocardiography or measurement of CVP may help prioritize therapeutic strategies. If hypotension is unresponsive to volume expansion and appropriate goal-directed antidotal therapy, treatment with norepinephrine, epinephrine, or high-dose dopamine may be necessary. Intraaortic balloon pump counterpulsation and venoarterial or cardiopulmonary perfusion techniques should be considered for severe but reversible cardiac failure. For patients with a return of spontaneous circulation after resuscitative treatment for cardiopulmonary arrest secondary to poisoning, therapeutic hypothermia should be used according to protocol. Bradyarrhythmias associated with hypotension generally should be treated as described in **Chaps. 239 and 240**. Glucagon, calcium, and high-dose insulin with dextrose may be effective in beta blocker and calcium channel blocker poisoning. Antibody therapy may be indicated for cardiac glycoside poisoning.

Supraventricular tachycardia associated with hypertension and CNS excitation is almost always due to agents that cause generalized physiologic excitation (Table 450–1). Most cases are mild or moderate in severity and require only observation or nonspecific sedation with a benzodiazepine. In severe cases or those associated with hemodynamic instability, chest pain, or ECG evidence of ischemia, specific therapy is indicated. When the etiology is sympathetic hyperactivity, treatment with a benzodiazepine should be prioritized. Further treatment with a combined alpha and beta blocker (labetalol), a calcium channel blocker (verapamil or diltiazem), or a combination of a beta blocker and a vasodilator (esmolol and nitroprusside) may be considered for cases refractory to high doses of benzodiazepines only when adequate sedation has been achieved but cardiac conduction or blood pressure abnormalities persist. Treatment with an α-adrenergic antagonist (phentolamine) alone may sometimes be appropriate. If the cause is anticholinergic poisoning, physostigmine alone can be effective. Supraventricular tachycardia without hypertension is generally secondary to vasodilation or hypovolemia and responds to fluid administration.

For ventricular tachyarrhythmias due to tricyclic antidepressants and other membrane-active agents (Table 450-1), sodium bicarbonate is indicated, whereas class IA, IC, and III antiarrhythmic agents are contraindicated because of similar electrophysiologic effects. Although lidocaine and phenytoin are historically safe for ventricular tachyarrhythmias of any etiology, sodium bicarbonate should be considered first for any ventricular arrhythmia suspected to have a toxicologic etiology. Intravenous lipid emulsion therapy has shown benefit for treatment of arrhythmias and hemodynamic instability from various membrane-active agents. Beta blockers can be hazardous if the arrhythmia is due to sympathetic hyperactivity. Magnesium sulfate and overdrive pacing (by isoproterenol or a pacemaker) may be useful in patients with torsades des pointes and prolonged QT intervals. Magnesium and anti-digoxin antibodies should be considered in patients with severe cardiac glycoside

poisoning. Invasive (esophageal or intracardiac) ECG recording may be necessary to determine the origin (ventricular or supraventricular) of wide-complex tachycardias **(Chap. 241)**. If the patient is hemodynamically stable, however, it is reasonable to simply observe him or her rather than to administer another potentially proarrhythmic agent. Arrhythmias may be resistant to drug therapy until underlying acid-base, electrolyte, oxygenation, and temperature derangements are corrected.

Central Nervous System Therapies

Neuromuscular hyperactivity and seizures can lead to hyperthermia, lactic acidosis, and rhabdomyolysis and should be treated aggressively. Seizures caused by excessive stimulation of catecholamine receptors (sympathomimetic or hallucinogen poisoning and drug withdrawal) or decreased activity of GABA (isoniazid poisoning) or glycine (strychnine poisoning) receptors are best treated with agents that enhance GABA activity, such as benzodiazepine or barbiturates. Since benzodiazepines and barbiturates act by slightly different mechanisms (the former increases the frequency via allosteric modulation at the receptor and the latter directly increases the duration of chloride channel opening in response to GABA), therapy with both may be effective when neither is effective alone. Seizures caused by isoniazid, which inhibits the synthesis of GABA at several steps by interfering with the cofactor pyridoxine (vitamin B_6), may require high doses of supplemental pyridoxine. Seizures resulting from membrane destabilization (beta blocker or cyclic antidepressant poisoning) require GABA enhancers (benzodiazepines first, barbiturates second). Phenytoin is contraindicated in toxicologic seizures: Animal and human data demonstrate worse outcomes after phenytoin loading, especially in theophylline overdose. For poisons with central dopaminergic effects (methamphetamine, phencyclidine) manifested by psychotic behavior, a dopamine receptor antagonist, such as haloperidol or ziprasidone, may be useful. In anticholinergic and cyanide poisoning, specific antidotal therapy may be necessary. The treatment of seizures secondary to cerebral ischemia or edema or to metabolic abnormalities should include correction of the underlying cause. Neuromuscular paralysis is indicated in refractory cases. Electroencephalographic monitoring and continuing treatment of seizures are necessary to prevent permanent neurologic damage. Serotonergic receptor overstimulation in serotonin syndrome may be treated with cyproheptadine.

Other Measures

Temperature extremes, metabolic abnormalities, hepatic and renal dysfunction, and secondary complications should be treated by standard therapies.

PREVENTION OF POISON ABSORPTION

Gastrointestinal Decontamination

Whether or not to perform gastrointestinal decontamination and which procedure to use depends on the time since ingestion; the existing and predicted toxicity of the ingestant; the availability, efficacy, and contraindications of the procedure; and the nature, severity, and risk of complications. The efficacy of all decontamination procedures decreases with time, and data are insufficient to support or exclude a beneficial effect when they are used >1 h after ingestion. The average time from ingestion to presentation for treatment is >1 h for children and >3 h for adults. Most patients will recover from poisoning uneventfully with good supportive care alone, but complications of gastrointestinal decontamination, particularly aspiration, can prolong this process. Hence, gastrointestinal decontamination should be performed selectively, not routinely, in

the management of overdose patients. It is clearly unnecessary when predicted toxicity is minimal or the time of expected maximal toxicity has passed without significant effect.

Activated charcoal has comparable or greater efficacy; has fewer contraindications and complications; and is less aversive and invasive than ipecac or gastric lavage. Thus it is the preferred method of gastrointestinal decontamination in most situations. Activated charcoal suspension (in water) is given orally via a cup, straw, or small-bore nasogastric tube. The generally recommended dose is 1 g/kg body weight because of its dosing convenience, although in vitro and in vivo studies have demonstrated that charcoal adsorbs ≥90% of most substances when given in an amount equal to 10 times the weight of the substance. Palatability may be increased by adding a sweetener (sorbitol) or a flavoring agent (cherry, chocolate, or cola syrup) to the suspension. Charcoal adsorbs ingested poisons within the gut lumen, allowing the charcoal-toxin complex to be evacuated with stool. Charged (ionized) chemicals such as mineral acids, alkalis, and highly dissociated salts of cyanide, fluoride, iron, lithium, and other inorganic compounds are not well adsorbed by charcoal. In studies with animals and human volunteers, charcoal decreases the absorption of ingestants by an average of 73% when given within 5 min of ingestant administration, 51% when given at 30 min, and 36% when given at 60 min. For this reason, charcoal given before hospital arrival increases the potential clinical benefit. Side effects of charcoal include nausea, vomiting, and diarrhea or constipation. Charcoal may also prevent the absorption of orally administered therapeutic agents. Complications include mechanical obstruction of the airway, aspiration, vomiting, and bowel obstruction and infarction caused by inspissated charcoal. Charcoal is not recommended for patients who have ingested corrosives because it obscures endoscopy.

Gastric lavage should be considered for life-threatening poisons that cannot be treated effectively with other decontamination, elimination, or antidotal therapies (e.g., colchicine). Gastric lavage is performed by sequentially administering and aspirating ~5 mL of fluid per kilogram of body weight through a no. 40 French orogastric tube (no. 28 French tube for children). Except in infants, for whom normal saline is recommended, tap water is acceptable. The patient should be placed in Trendelenburg and left lateral decubitus positions to prevent aspiration (even if an endotracheal tube is in place). Lavage decreases ingestant absorption by an average of 52% if performed within 5 min of ingestion administration, 26% if performed at 30 min, and 16% if performed at 60 min. Significant amounts of ingested drug are recovered from <10% of patients. Aspiration is a common complication (occurring in up to 10% of patients), especially when lavage is performed improperly. Serious complications (esophageal and gastric perforation, tube misplacement in the trachea) occur in ~1% of patients. For this reason, the physician should personally insert the lavage tube and confirm its placement, and the patient must be cooperative during the procedure. Gastric lavage is contraindicated in corrosive or petroleum distillate ingestions because of the respective risks of gastroesophageal perforation and aspiration pneumonitis. It is also contraindicated in patients with a compromised unprotected airway and those at risk for hemorrhage or perforation due to esophageal or gastric pathology or recent surgery. Finally, gastric lavage is absolutely contraindicated in combative patients or those who refuse, as most published complications involve patient resistance to the procedure.

Syrup of ipecac, an emetogenic agent that was once the substance most commonly used for decontamination, no longer has a role in poisoning management. Even the American Academy of Pediatrics—traditionally the strongest proponent of ipecac—issued a policy statement in 2003 recommending that ipecac should no longer

be used in poisoning treatment. Chronic ipecac use (by patients with anorexia nervosa or bulimia) has been reported to cause electrolyte and fluid abnormalities, cardiac toxicity, and myopathy.

Whole-bowel irrigation is performed by administering a bowel-cleansing solution containing electrolytes and polyethylene glycol (Golytely, Colyte) orally or by gastric tube at a rate of 2 L/h (0.5 L/h in children) until rectal effluent is clear. The patient must be in a sitting position. Although data are limited, whole-bowel irrigation appears to be as effective as other decontamination procedures in volunteer studies. It is most appropriate for those who have ingested foreign bodies, packets of illicit drugs, and agents that are poorly adsorbed by charcoal (e.g., heavy metals). This procedure is contraindicated in patients with bowel obstruction, ileus, hemodynamic instability, and compromised unprotected airways.

Cathartics are salts (disodium phosphate, magnesium citrate and sulfate, sodium sulfate) or saccharides (mannitol, sorbitol) that historically have been given with activated charcoal to promote the rectal evacuation of gastrointestinal contents. However, no animal, volunteer, or clinical data have ever demonstrated any decontamination benefit from cathartics. Abdominal cramps, nausea, and occasional vomiting are side effects. Complications of repeated dosing include severe electrolyte disturbances and excessive diarrhea. Cathartics are contraindicated in patients who have ingested corrosives and in those with preexisting diarrhea. Magnesium-containing cathartics should not be used in patients with renal failure.

Dilution (i.e., drinking water, another clear liquid, or milk at a volume of 5 mL/kg of body weight) is recommended only after the ingestion of corrosives (acids, alkali). It may increase the dissolution rate (and hence absorption) of capsules, tablets, and other solid ingestants and should *not* be used in these circumstances.

Endoscopic or surgical removal of poisons may be useful in rare situations, such as ingestion of a potentially toxic foreign body that fails to transit the gastrointestinal tract, a potentially lethal amount of a heavy metal (arsenic, iron, mercury, thallium), or agents that have coalesced into gastric concretions or bezoars (heavy metals, lithium, salicylates, sustained-release preparations). Patients who become toxic from cocaine due to its leakage from ingested drug packets require immediate surgical intervention.

Decontamination of Other Sites

Immediate, copious flushing with water, saline, or another available clear, drinkable liquid is the initial treatment for topical exposures (exceptions include alkali metals, calcium oxide, phosphorus). Saline is preferred for eye irrigation. A triple wash (water, soap, water) may be best for dermal decontamination. Inhalational exposures should be treated initially with fresh air or supplemental oxygen. The removal of liquids from body cavities such as the vagina or rectum is best accomplished by irrigation. Solids (drug packets, pills) should be removed manually, preferably under direct visualization.

ENHANCEMENT OF POISON ELIMINATION

Although the elimination of most poisons can be accelerated by therapeutic interventions, the pharmacokinetic efficacy (removal of drug at a rate greater than that accomplished by intrinsic elimination) and clinical benefit (shortened duration of toxicity or improved outcome) of such interventions are often more theoretical than

proven. Accordingly, the decision to use such measures should be based on the actual or predicted toxicity and the potential efficacy, cost, and risks of therapy.

Multiple-Dose Activated Charcoal

Repetitive oral dosing with charcoal can enhance the elimination of previously absorbed substances by binding them within the gut as they are excreted in the bile, are secreted by gastrointestinal cells, or passively diffuse into the gut lumen (*reverse absorption* or *enterocapillary exsorption*). Doses of 0.5–1 g/kg of body weight every 2–4 h, adjusted downward to avoid regurgitation in patients with decreased gastrointestinal motility, are generally recommended. Pharmacokinetic efficacy approaches that of hemodialysis for some agents (e.g., phenobarbital, theophylline). Multiple-dose therapy should be considered only for selected agents (theophylline, phenobarbital, carbamazepine, dapsone, quinine). Complications include intestinal obstruction, pseudo-obstruction, and nonocclusive intestinal infarction in patients with decreased gut motility. Because of electrolyte and fluid shifts, sorbitol and other cathartics are absolutely contraindicated when multiple doses of activated charcoal are administered.

Urinary Alkalinization

Ion trapping via alteration of urine pH may prevent the renal reabsorption of poisons that undergo excretion by glomerular filtration and active tubular secretion. Since membranes are more permeable to non-ionized molecules than to their ionized counterparts, acidic (low-pK_a) poisons are ionized and trapped in alkaline urine, whereas basic ones become ionized and trapped in acid urine. Urinary alkalinization (producing a urine pH ≥7.5 and a urine output of 3–6 mL/kg of body weight per hour by the addition of sodium bicarbonate to an IV solution) enhances the excretion of chlorophenoxyacetic acid herbicides, chlorpropamide, diflunisal, fluoride, methotrexate, phenobarbital, sulfonamides, and salicylates. Contraindications include congestive heart failure, renal failure, and cerebral edema. Acid-base, fluid, and electrolyte parameters should be monitored carefully. Although acid diuresis may make theoretical sense for some overdoses (amphetamines), it is *never* indicated and is potentially harmful.

Extracorporeal Removal

Hemodialysis, charcoal or resin hemoperfusion, hemofiltration, plasmapheresis, and exchange transfusion are capable of removing any toxin from the bloodstream. Agents most amenable to enhanced elimination by dialysis have low molecular mass (<500 Da), high water solubility, low protein binding, small volumes of distribution (<1 L/kg of body weight), prolonged elimination (long half-life), and high dialysis clearance relative to total-body clearance. Molecular weight, water solubility, and protein binding do not limit the efficacy of the other forms of extracorporeal removal.

Dialysis should be considered in cases of severe poisoning due to carbamazepine, ethylene glycol, isopropyl alcohol, lithium, methanol, theophylline, salicylates, and valproate. Although hemoperfusion may be more effective in removing some of these poisons, it does not correct associated acid-base and electrolyte abnormalities, and most hospitals no longer have hemoperfusion cartridges readily available. Fortunately, recent advances in hemodialysis technology make it as effective as hemoperfusion for removing poisons such as caffeine, carbamazepine, and theophylline. Both techniques require central venous access and systemic anticoagulation and may result in transient hypotension. Hemoperfusion may also cause hemolysis,

hypocalcemia, and thrombocytopenia. Peritoneal dialysis and exchange transfusion are less effective but may be used when other procedures are unavailable, contraindicated, or technically difficult (e.g., in infants). Exchange transfusion may be indicated in the treatment of severe arsine- or sodium chlorate–induced hemolysis, methemoglobinemia, and sulfhemoglobinemia. Although hemofiltration can enhance elimination of aminoglycosides, vancomycin, and metal-chelate complexes, the roles of hemofiltration and plasmapheresis in the treatment of poisoning are not yet defined.

Candidates for extracorporeal removal therapies include patients with severe toxicity whose condition deteriorates despite aggressive supportive therapy; those with potentially prolonged, irreversible, or fatal toxicity; those with dangerous blood levels of toxins; those who lack the capacity for self-detoxification because of liver or renal failure; and those with a serious underlying illness or complication that will adversely affect recovery.

Other Techniques

The elimination of heavy metals can be enhanced by chelation, and the removal of carbon monoxide can be accelerated by hyperbaric oxygenation.

ADMINISTRATION OF ANTIDOTES

Antidotes counteract the effects of poisons by neutralizing them (e.g., antibody-antigen reactions, chelation, chemical binding) or by antagonizing their physiologic effects (e.g., activation of opposing nervous system activity, provision of a competitive metabolic or receptor substrate). Poisons or conditions with specific antidotes include acetaminophen, anticholinergic agents, anticoagulants, benzodiazepines, beta blockers, calcium channel blockers, carbon monoxide, cardiac glycosides, cholinergic agents, cyanide, drug-induced dystonic reactions, ethylene glycol, fluoride, heavy metals, hypoglycemic agents, isoniazid, membrane-active agents, methemoglobinemia, opioids, sympathomimetics, and a variety of envenomations. Intravenous lipid emulsion has been shown to be a successful antidote for poisoning from various anesthetics and membrane-active agents (e.g., cyclic antidepressants), but the exact mechanism of benefit is still under investigation. Antidotes can significantly reduce morbidity and mortality rates but are potentially toxic if used for inappropriate reasons. Since their safe use requires correct identification of a specific poisoning or syndrome, details of antidotal therapy are discussed with the conditions for which they are indicated **(Table 450-4)**.

PREVENTION OF REEXPOSURE

Poisoning is a preventable illness. Unfortunately, some adults and children are poison-prone, and recurrences are common. Unintentional polypharmacy poisoning has become especially common among adults with developmental delays, among the growing population of geriatric patients who are prescribed a large number of medications, and among adolescents and young adults experimenting with pharmaceuticals for recreational euphoria. Adults with unintentional exposures should be instructed regarding the safe use of medications and chemicals (according to labeling instructions). Confused patients may need assistance with the administration of their medications. Errors in dosing by health care providers may require educational efforts. Patients should be advised to avoid circumstances that result in chemical exposure or poisoning. Appropriate agencies and health departments (e.g., Occupational Health and Safety Administration [OSHA]) should be notified in cases of environmental or workplace exposure. The best approach to young children and patients with intentional overdose (deliberate self-harm or attempted suicide) is to limit their access to poisons. In households where

children live or visit, alcoholic beverages, medications, household products (automotive, cleaning, fuel, pet-care, and toiletry products), inedible plants, and vitamins should be kept out of reach or in locked or child-proof cabinets. Depressed or psychotic patients should undergo psychiatric assessment, disposition, and follow-up. They should be given prescriptions for a limited supply of drugs with a limited number of refills and should be monitored for compliance and response to therapy.

TABLE 450-3

Fundamentals of Poisoning Management

Supportive Care	
Airway protection	Treatment of seizures
Oxygenation/ventilation	Correction of temperature abnormalities
Treatment of arrhythmias	Correction of metabolic derangements
Hemodynamic support	Prevention of secondary complications
Prevention of Further Poison Absorption	
Gastrointestinal decontamination	Decontamination of other sites
Gastric lavage	Eye decontamination
Activated charcoal	Skin decontamination
Whole-bowel irrigation	Body cavity evacuation
Dilution	
Endoscopic/surgical removal	
Enhancement of Poison Elimination	
Multiple-dose activated charcoal administration Alteration of urinary pH Chelation	Extracorporeal removal Hemodialysis Hemoperfusion Hemofiltration Plasmapheresis Exchange transfusion Hyperbaric oxygenation
Administration of Antidotes	
Neutralization by antibodies	Metabolic antagonism
Neutralization by chemical binding	Physiologic antagonism

Prevention of Reexposure	
Adult education	Notification of regulatory agencies
Child-proofing	Psychiatric referral

TABLE 450-4

Pathophysiologic Features and Treatment of Specific Toxic Syndromes and Poisonings

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Stimulated				
Sympathetics[a]				

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Sympathomimetics	α_1-Adrenergic agonists (decongestants): phenylephrine, phenylpropanolamine β_2-Adrenergic agonists (bronchodilators): albuterol, terbutaline Nonspecific adrenergic agonists: amphetamines, cocaine, ephedrine	Stimulation of central and peripheral sympathetic receptors directly or indirectly (by promoting release or inhibiting reuptake of norepinephrine and sometimes dopamine)	Physiologic stimulation (Table 450-2). Reflex bradycardia can occur with selective α_1 agonists; β agonists can cause hypotension and hypokalemia.	Phentolamine, a nonselective α_1-adrenergic receptor antagonist, for severe hypertension due to α_1-adrenergic agonists; propranolol, a nonselective β blocker, for hypotension and tachycardia due to β_2 agonists; *either* labetalol, a β blocker with α-blocking activity, *or* phentolamine with esmolol, metoprolol, or another cardioselective β blocker for hypertension with tachycardia due to nonselective agents (β blockers, if used alone, can exacerbate hypertension and vasospasm due to unopposed α stimulation.); benzodiazepines; propofol

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Ergot alkaloids	Ergotamine, methysergide, bromocriptine, pergolide	Stimulation and inhibition of serotonergic and α-adrenergic receptors; stimulation of dopamine receptors	Physiologic stimulation (Table 450-2); formication; vasospasm with limb (isolated or generalized), myocardial, and cerebral ischemia progressing to gangrene or infarction. Hypotension, bradycardia, and involuntary movements can also occur.	Nitroprusside or nitroglycerine for severe vasospasm; prazosin (an α_1 blocker), captopril, nifedipine, and cyproheptadine (a serotonin receptor antagonist) for mild-to-moderate limb ischemia; dopamine receptor antagonists (antipsychotics) for hallucinations and movement disorders

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Methylxanthines	Caffeine, theophylline	Inhibition of adenosine synthesis and adenosine receptor antagonism; stimulation of epinephrine and norepinephrine release; inhibition of phosphodiesterase resulting in increased intracellular cyclic adenosine and guanosine monophosphate	Physiologic stimulation (Table 450-2); pronounced gastrointestinal symptoms and β agonist effects (see above). Toxicity occurs at lower drug levels in chronic poisoning than in acute poisoning.	Propranolol, a nonselective β blocker, for tachycardia with hypotension; any β blocker for supraventricular or ventricular tachycardia without hypotension; elimination enhanced by multiple-dose charcoal, hemoperfusion, and hemodialysis. Indications for hemoperfusion or hemodialysis include unstable vital signs, seizures, and a theophylline level of 80–100 µg/mL after an acute overdose and 40–60 µg/mL with chronic exposure.

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Monoamine oxidase inhibitors	Phenelzine, tranylcypromine, selegiline	Inhibition of monoamine oxidase resulting in impaired metabolism of endogenous catecholamines and exogenous sympathomimetic agents	Delayed or slowly progressive physiologic stimulation (Table 450-2); terminal hypotension and bradycardia in severe cases	Short-acting agents (e.g., nitroprusside, esmolol) for severe hypertension and tachycardia; direct-acting sympathomimetics (e.g., norepinephrine, epinephrine) for hypotension and bradycardia
Anticholinergics				
Antihistamines	Diphenhydramine, doxylamine, pyrilamine	Inhibition of central and postganglionic parasympathetic muscarinic cholinergic receptors. At high doses, amantadine, diphenhydramine, orphenadrine, phenothiazines, and tricyclic antidepressants have additional nonanticholinergic activity (see below).	Physiologic stimulation (Table 450-2); dry skin and mucous membranes, decreased bowel sounds, flushing, and urinary retention; myoclonus and picking activity. Central effects may occur without significant autonomic dysfunction.	Physostigmine, an acetylcholinesterase inhibitor (see below), for delirium, hallucinations, and neuromuscular hyperactivity. Contraindications include asthma and non-anticholinergic cardiovascular toxicity (e.g., cardiac conduction abnormalities, hypotension, and ventricular arrhythmias).

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Antipsychotics	Chlorpromazine, olanzapine, quetiapine, thioridazine	Inhibition of α-adrenergic, dopaminergic, histaminergic, muscarinic, and serotonergic receptors. Some agents also inhibit sodium, potassium, and calcium channels.	Physiologic depression (Table 450-2), miosis, anticholinergic effects (see above), extrapyramidal reactions (see below), tachycardia	Sodium bicarbonate for ventricular tachydysrhythmias associated with QRS prolongation; magnesium, isoproterenol, and overdrive pacing for torsades des pointes. Avoid class IA, IC, and III antiarrhythmics.
Belladonna alkaloids	Atropine, hyoscyamine, scopolamine	Inhibition of central and postganglionic parasympathetic muscarinic cholinergic receptors	Physiologic stimulation (Table 450-2); dry skin and mucous membranes, decreased bowel sounds, flushing, and urinary retention; myoclonus and picking activity. Central effects may occur without significant autonomic dysfunction.	Physostigmine, an acetylcholinesterase inhibitor (see below), for delirium, hallucinations, and neuromuscular hyperactivity. Contraindications include asthma and non-anticholinergic cardiovascular toxicity (e.g., cardiac conduction abnormalities, hypotension, and ventricular arrhythmias).

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Cyclic antidepressants	Amitriptyline, doxepin, imipramine	Inhibition of α-adrenergic, dopaminergic, GABA-ergic, histaminergic, muscarinic, and serotonergic receptors; inhibition of sodium channels (see membrane-active agents); inhibition of norepinephrine and serotonin reuptake	Physiologic depression (Table 450-2), seizures, tachycardia, cardiac conduction delays (increased PR, QRS, JT, and QT intervals; terminal QRS right-axis deviation) with aberrancy and ventricular tachydysrhythmias; anticholinergic toxidrome (see above)	Hypertonic sodium bicarbonate (or hypertonic saline) for ventricular tachydysrhythmias associated with QRS prolongation. Use of phenytoin is controversial. Avoid class IA, IC, and III antiarrhythmics. IV emulsion therapy may be beneficial in some cases.
Mushrooms and plants	*Amanita muscaria* and *A. pantherina*, henbane, jimson weed, nightshade	Inhibition of central and postganglionic parasympathetic muscarinic cholinergic receptors	Physiologic stimulation (Table 450-2); dry skin and mucous membranes, decreased bowel sounds, flushing, and urinary retention; myoclonus and picking activity. Central effects may occur without significant autonomic dysfunction.	Physostigmine, an acetylcholinesterase inhibitor (see below), for delirium, hallucinations, and neuromuscular hyperactivity. Contraindications include asthma and nonanticholinergic cardiovascular toxicity (e.g., cardiac conduction abnormalities, hypotension, and ventricular arrhythmias).
Depressed				
Sympatholytics				

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
α_2-Adrenergic agonists	Clonidine, guanabenz, tetrahydrozoline and other imidazoline decongestants, tizanidine and other imidazoline muscle relaxants	Stimulation of α_2-adrenergic receptors leading to inhibition of CNS sympathetic outflow. Activity at nonadrenergic imidazoline binding sites also contributes to CNS effects.	Physiologic depression (Table 450-2), miosis. Transient initial hypertension may be seen.	Dopamine and norepinephrine for hypotension; atropine for symptomatic bradycardia; naloxone for CNS depression (inconsistently effective)
Antipsychotics	Chlorpromazine, clozapine, haloperidol, risperidone, thioridazine	Inhibition of α-adrenergic, dopaminergic, histaminergic, muscarinic, and serotonergic receptors. Some agents also inhibit sodium, potassium, and calcium channels.	Physiologic depression (Table 450-2), miosis, anticholinergic effects (see above), extrapyramidal reactions (see below), tachycardia. Cardiac conduction delays (increased PR, QRS, JT, and QT intervals) with ventricular tachydysrhythmias, including torsades des pointes, can sometimes develop.	Sodium bicarbonate for ventricular tachydysrhythmias associated with QRS prolongation; magnesium, isoproterenol, and overdrive pacing for torsades des pointes. Avoid class IA, IC, and III antiarrhythmics.

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
β-Adrenergic blockers	Cardioselective (β_1) blockers: atenolol, esmolol, metoprolol Nonselective (β_1 and β_2) blockers: nadolol, propranolol, timolol Partial β agonists: acebutolol, pindolol α_1 Antagonists: carvedilol, labetalol Membrane-active agents: acebutolol, propranolol, sotalol	Inhibition of β-adrenergic receptors (class II antiarrhythmic effect). Some agents have activity at additional receptors or have membrane effects (see below).	Physiologic depression (Table 450-2), atrioventricular block, hypoglycemia, hyperkalemia, seizures. Partial agonists can cause hypertension and tachycardia. Sotalol can cause increased QT interval and ventricular tachydysrhythmias. Onset may be delayed after sotalol and sustained-release formulation overdose.	Glucagon for hypotension and symptomatic bradycardia. Atropine, isoproterenol, dopamine, dobutamine, epinephrine, and norepinephrine may sometimes be effective. High-dose insulin (with glucose and potassium to maintain euglycemia and normokalemia), electrical pacing, and mechanical cardiovascular support for refractory cases.

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Calcium channel blockers	Diltiazem, nifedipine and other dihydropyridine derivatives, verapamil	Inhibition of slow (type L) cardiovascular calcium channels (class IV antiarrhythmic effect)	Physiologic depression (Table 450-2), atrioventricular block, organ ischemia and infarction, hyperglycemia, seizures. Hypotension is usually due to decreased vascular resistance rather than to decreased cardiac output. Onset may be delayed for ≥12 h after overdose of sustained-release formulations.	Calcium and glucagon for hypotension and symptomatic bradycardia. Dopamine, epinephrine, norepinephrine, atropine, and isoproterenol are less often effective but can be used adjunctively. High-dose insulin (with glucose and potassium to maintain euglycemia and normokalemia), IV lipid emulsion therapy, electrical pacing, and mechanical cardiovascular support for refractory cases.

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Cardiac glycosides	Digoxin, endogenous cardioactive steroids, foxglove and other plants, toad skin secretions (Bufonidae spp.)	Inhibition of cardiac Na^+, K^+-ATPase membrane pump	Physiologic depression (Table 450-2); gastrointestinal, psychiatric, and visual symptoms; atrioventricular block with or without concomitant supraventricular tachyarrhythmia; ventricular tachyarrhythmias; hyperkalemia in acute poisoning. Toxicity occurs at lower drug levels in chronic poisoning than in acute poisoning.	Digoxin-specific antibody fragments for hemodynamically compromising dysrhythmias, Mobitz II or third-degree atrioventricular block, hyperkalemia (>5.5 meq/L; in acute poisoning only). Temporizing measures include atropine, dopamine, epinephrine, and external cardiac pacing for bradydysrhythmias and magnesium, lidocaine, or phenytoin, for ventricular tachydysrhythmias. Internal cardiac pacing and cardioversion can increase ventricular irritability and should be reserved for refractory cases.

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Cyclic antidepressants	Amitriptyline, doxepin, imipramine	Inhibition of α-adrenergic, dopaminergic, GABA-ergic, histaminergic, muscarinic, and serotonergic receptors; inhibition of sodium channels (see membrane-active agents); inhibition of norepinephrine and serotonin reuptake	Physiologic depression (Table 450-2), seizures, tachycardia, cardiac conduction delays (increased PR, QRS, JT, and QT intervals; terminal QRS right-axis deviation) with aberrancy and ventricular tachydysrhythmias; anticholinergic toxidrome (see above)	Hypertonic sodium bicarbonate (or hypertonic saline) for ventricular tachydysrhythmias associated with QRS prolongation. Use of phenytoin is controversial. Avoid class IA, IC, and III antiarrhythmics. IV emulsion therapy may be beneficial in some cases.
Cholinergics				
Acetylcholinesterase inhibitors	Carbamate insecticides (aldicarb, carbaryl, propoxur) and medicinals (neostigmine, physostigmine, tacrine); nerve gases (sarin, soman, tabun, VX); organophosphate insecticides (diazinon, chlorpyrifos-ethyl, malathion)	Inhibition of acetylcholinesterase leading to increased synaptic acetylcholine at muscarinic and nicotinic cholinergic receptor sites	Physiologic depression (Table 450-2). Muscarinic signs and symptoms: seizures, excessive secretions (lacrimation, salivation, bronchorrhea and wheezing, diaphoresis), and increased bowel and bladder activity with nausea, vomiting, diarrhea, abdominal cramps, and incontinence of feces and urine. Nicotinic signs and symptoms: hypertension,	Atropine for muscarinic signs and symptoms; 2-PAM, a cholinesterase reactivator, for nicotinic signs and symptoms due to organophosphates, nerve gases, or an unknown anticholinesterase
Muscarinic agonists	Bethanechol, mushrooms (*Boletus, Clitocybe, Inocybe* spp.), pilocarpine	Stimulation of CNS and postganglionic parasympathetic cholinergic (muscarinic) receptors		

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Nicotinic agonists	Lobeline, nicotine (tobacco)	Stimulation of preganglionic sympathetic and parasympathetic and striated muscle (neuromuscular junction) cholinergic (nicotine) receptors	tachycardia, muscle cramps, fasciculations, weakness, and paralysis. Death is usually due to respiratory failure. Cholinesterase activity in plasma and red cells is <50% of normal in acetylcholinesterase inhibitor poisoning.	
Sedative-hypnotics[b]				
Anticonvulsants	Carbamazepine, ethosuximide, felbamate, gabapentin, lamotrigine, levetiracetam, oxcarbazepine, phenytoin, tiagabine, topiramate, valproate, zonisamide	Potentiation of the inhibitory effects of GABA by binding to the neuronal GABA–A chloride channel receptor complex and increasing the frequency or duration of chloride channel opening in response to GABA stimulation. Baclofen and, to some extent, GHB act at the GABA–B receptor complex. Meprobamate, its metabolite carisoprodol, felbamate, and orphenadrine antagonize NDMA excitatory receptors. Ethosuximide,	Physiologic depression (Table 450-2), nystagmus. Delayed absorption can occur with carbamazepine, phenytoin, and valproate. Myoclonus, seizures, hypertension, and tachyarrhythmias can occur with baclofen, carbamazepine, and orphenadrine.	Benzodiazepines, barbiturates, or propofol for seizures.
Barbiturates	Short-acting: butabarbital, pentobarbital, secobarbital Long-acting: phenobarbital, primidone			Hemodialysis and hemoperfusion may be indicated for severe poisoning by some agents (see "Extracorporeal Removal," in text).

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Benzodiazepines	Ultrashort-acting: estazolam, midazolam, temazepam, triazolam Short-acting: alprazolam, flunitrazepam, lorazepam, oxazepam Long-acting: chlordiazepoxide, clonazepam, diazepam, flurazepam Pharmacologically related agents: zaleplon, zolpidem	valproate, and zonisamide decrease conduction through T-type calcium channels. Valproate decreases GABA degradation, and tiagabine blocks GABA reuptake. Carbamazepine, lamotrigine, oxcarbazepine, phenytoin, topiramate, valproate, and zonisamide slow the rate of recovery of inactivated sodium channels. Some agents also have α_2 agonist, anticholinergic, and sodium channel–blocking activity (see above and below).	Tachyarrhythmias can also occur with chloral hydrate. AGMA, hypernatremia, hyperosmolality, hyperammonemia, chemical hepatitis, and hypoglycemia can be seen in valproate poisoning. Carbamazepine and oxcarbazepine may produce hyponatremia from SIADH.	See above and below for treatment of anticholinergic and sodium channel (membrane)–blocking effects.
GABA precursors	γ-Hydroxybutyrate (sodium oxybate; GHB), γ-butyrolactone (GBL), 1,4-butanediol		Some agents can cause anticholinergic and sodium channel (membrane) blocking effects (see above and below).	
Muscle relaxants	Baclofen, carisoprodol, cyclobenzaprine, etomidate, metaxalone, methocarbamol, orphenadrine, propofol, tizanidine and other imidazoline muscle relaxants	Baclofen acts at GABA-B receptor complex; Stimulation of α2-adrenergic receptors inhibits CNS sympathetic outflow. Activity at nonadrenergic imidazoline binding sites also contributes to CNS effects. The others have centrally-acting and	Physiologic depression (Table 450-2)	Goal-directed supportive care; benzodiazepines and barbiturates for seizures

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	various other MECHANISM OF unknown ACTION mechanisms of	CLINICAL FEATURES	SPECIFIC TREATMENTS
Other agents	Chloral hydrate, ethchlorvynol, glutethimide, meprobamate, methaqualone, methyprylon	action		
Discordant				
Asphyxiants				

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Cytochrome oxidase inhibitors	Cyanide, hydrogen sulfide	Inhibition of mitochondrial cytochrome oxidase, with consequent blockage of electron transport and oxidative metabolism. Carbon monoxide also binds to hemoglobin and myoglobin and prevents oxygen binding, transport, and tissue uptake. (Binding to hemoglobin shifts the oxygen dissociation curve to the left.)	Signs and symptoms of hypoxemia with initial physiologic stimulation and subsequent depression (Table 450-2); lactic acidosis; normal Po_2 and calculated oxygen saturation but decreased oxygen saturation by co-oximetry. (That measured by pulse oximetry is falsely elevated but is less than normal and less than the calculated value.) Headache and nausea are common with carbon monoxide. Sudden collapse may occur with cyanide and hydrogen sulfide exposure. A bitter almond breath odor may be noted with cyanide ingestion, and hydrogen sulfide smells like rotten eggs.	High-dose oxygen; IV hydroxocobalamin or IV sodium nitrite and sodium thiosulfate (Lilly cyanide antidote kit) for coma, metabolic acidosis, and cardiovascular dysfunction in cyanide poisoning or victims from a fire

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Methemoglobin inducers	Aniline derivatives, dapsone, local anesthetics, nitrates, nitrites, nitrogen oxides, nitro- and nitrosohydrocarbons, phenazopyridine, primaquine-type antimalarials, sulfonamides	Oxidation of hemoglobin iron from ferrous (Fe^{2+}) to ferric (Fe^{3+}) state prevents oxygen binding, transport, and tissue uptake. (Methemoglobinemia shifts oxygen dissociation curve to the left.) Oxidation of hemoglobin protein causes hemoglobin precipitation and hemolytic anemia (manifesting as Heinz bodies and “bite cells” on peripheral-blood smear).	Signs and symptoms of hypoxemia with initial physiologic stimulation and subsequent depression (Table 450-2), gray-brown cyanosis unresponsive to oxygen at methemoglobin fractions >15–20%, headache, lactic acidosis (at methemoglobin fractions >45%), normal Po_2 and calculated oxygen saturation but decreased oxygen saturation and increased methemoglobin fraction by co-oximetry (Oxygen saturation by pulse oximetry may be falsely increased or decreased but is less than normal and less than the calculated value.)	High-dose oxygen; IV methylene blue for methemoglobin fraction >30%, symptomatic hypoxemia, or ischemia (contraindicated in G6PD deficiency); exchange transfusion and hyperbaric oxygen for severe or refractory cases

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
AGMA inducers	Ethylene glycol	Ethylene glycol causes CNS depression and increased serum osmolality. Metabolites (primarily glycolic acid) cause AGMA, CNS depression, and renal failure. Precipitation of oxalic acid metabolite as calcium salt in tissues and urine results in hypocalcemia, tissue edema, and crystalluria.	Initial ethanol-like intoxication, nausea, vomiting, increased osmolar gap, calcium oxylate crystalluria; delayed AGMA, back pain, renal failure; coma, seizures, hypotension, ARDS in severe cases	Sodium bicarbonate to correct acidemia; thiamine, folinic acid, magnesium, and high-dose pyridoxine to facilitate metabolism; ethanol or fomepizole for AGMA, crystalluria or renal dysfunction, ethylene glycol level >3 mmol/L (20 mg/dL), and ethanol-like intoxication or increased osmolal gap if level not readily obtainable; hemodialysis for persistent AGMA, lack of clinical improvement, and renal dysfunction; hemodialysis also useful for enhancing ethylene glycol elimination and shortening duration of treatment when ethylene glycol level is >8 mmol/L (50 mg/dL).

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
AGMA inducers	Iron	Hydration of ferric (Fe^{3+}) ion generates H^+. Non-transferrin-bound iron catalyzes formation of free radicals that cause mitochondrial injury, lipid peroxidation, increased capillary permeability, vasodilation, and organ toxicity.	Initial nausea, vomiting, abdominal pain, diarrhea; AGMA, cardiovascular and CNS depression, hepatitis, coagulopathy, and seizures in severe cases. Radiopaque iron tablets may be seen on abdominal x-ray.	Whole-bowel irrigation for large ingestions; endoscopy and gastrostomy if clinical toxicity and large number of tablets are still visible on x-ray; IV hydration; sodium bicarbonate for acidemia; IV deferoxamine for systemic toxicity, iron level >90 µmol/L (500 µg/dL)

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
	Methanol	Methanol causes ethanol-like CNS depression and increased serum osmolality. Formic acid metabolite causes AGMA and retinal toxicity.	Initial ethanol-like intoxication, nausea, vomiting, increased osmolar gap; delayed AGMA, visual (clouding, spots, blindness) and retinal (edema, hyperemia) abnormalities; coma, seizures, cardiovascular depression in severe cases; possible pancreatitis	Gastric aspiration for recent ingestion; sodium bicarbonate to correct acidemia; high-dose folinic acid or folate to facilitate metabolism; ethanol or fomepizole for AGMA, visual symptoms, methanol level >6 mmol/L (20 mg/dL), and ethanol-like intoxication or increased osmolal gap if level not readily obtainable; hemodialysis for persistent AGMA, lack of clinical improvement, and renal dysfunction; hemodialysis also useful for enhancing methanol elimination and shortening duration of treatment when methanol level is >15 mmol/L (50 mg/dL)

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
	Salicylate	Increased sensitivity of CNS respiratory center to changes in and stimulates respiration. Uncoupling of oxidative phosphorylation, inhibition of Krebs cycle enzymes, and stimulation of carbohydrate and lipid metabolism generate unmeasured endogenous anions and cause AGMA.	Initial nausea, vomiting, hyperventilation, alkalemia, alkaluria; subsequent alkalemia with both respiratory alkalosis and AGMA and paradoxical aciduria; late acidemia with CNS and respiratory depression; cerebral and pulmonary edema in severe cases. Hypoglycemia, hypocalcemia, hypokalemia, and seizures can occur.	IV hydration and supplemental glucose; sodium bicarbonate to correct acidemia; urinary alkalinization for systemic toxicity; hemodialysis for coma, cerebral edema, seizures, pulmonary edema, renal failure, progressive acid-base disturbances or clinical toxicity, salicylate level >7 mmol/L (100 mg/dL) following acute overdose
CNS syndromes				
Extrapyramidal reactions	Antipsychotics (see above), some cyclic antidepressants and antihistamines	Decreased CNS dopaminergic activity with relative excess of cholinergic activity	Akathisia, dystonia, parkinsonism	Oral or parenteral anticholinergic agent such as benztropine or diphenhydramine

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Isoniazid		Interference with activation and supply of pyridoxal-5-phosphate, a cofactor for glutamic acid decarboxylase, which converts glutamic acid to GABA, results in decreased levels of this inhibitory CNS neurotransmitter; complexation with and depletion of pyridoxine itself; inhibition of nicotine adenine dinucleotide–dependent lactate and hydroxybutyrate dehydrogenases, resulting in substrate accumulation	Nausea, vomiting, agitation, confusion; coma, respiratory depression, seizures, lactic and ketoacidosis in severe cases	High-dose IV pyridoxine (vitamin B_6) for agitation, confusion, coma, and seizures; diazepam or barbiturates for seizures

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Lithium		Interference with cell membrane ion transport, adenylate cyclase and Na^+, K^+-ATPase activity, and neurotransmitter release	Nausea, vomiting, diarrhea, ataxia, choreoathetosis, encephalopathy, hyperreflexia, myoclonus, nystagmus, nephrogenic diabetes insipidus, falsely elevated serum chloride with low anion gap, tachycardia; coma, seizures, arrhythmias, hyperthermia, and prolonged or permanent encephalopathy and movement disorders in severe cases; delayed onset after acute overdose, particularly with delayed-release formulations. Toxicity occurs at lower drug levels in chronic poisoning than in acute poisoning.	Whole-bowel irrigation for large ingestions; IV hydration; hemodialysis for coma, seizures, encephalopathy or neuromuscular dysfunction (severe, progressive, or persistent), peak lithium level >4 meq/L following acute overdose

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Serotonin syndrome	Amphetamines, cocaine, dextromethorphan, meperidine, MAO inhibitors, selective serotonin (5-HT) reuptake inhibitors, tricyclic antidepressants, tramadol, triptans, tryptophan	Promotion of serotonin release, inhibition of serotonin reuptake, or direct stimulation of CNS and peripheral serotonin receptors (primarily 5-HT-1a and 5-HT-2), alone or in combination	Altered mental status (agitation, confusion, mutism, coma, seizures), neuromuscular hyperactivity (hyperreflexia, myoclonus, rigidity, tremors), and autonomic dysfunction (abdominal pain, diarrhea, diaphoresis, fever, flushing, labile hypertension, mydriasis, tearing, salivation, tachycardia). Complications include hyperthermia, lactic acidosis, rhabdomyolysis, and multisystem organ failure.	Discontinue the offending agent(s); the serotonin receptor antagonist cyproheptadine may be helpful in severe cases.
Membrane-active agent	Amantadine, antiarrhythmics (class I and III agents; some β blockers), antipsychotics (see above), antihistamines (particularly diphenhydramine), carbamazepine, local anesthetics (including	Blockade of fast sodium membrane channels prolongs phase 0 (depolarization) of the cardiac action potential, which prolongs QRS duration and promotes reentrant (monomorphic)	QRS and JT prolongation (or both) with hypotension, ventricular tachyarrhythmias, CNS depression, seizures; anticholinergic effects with amantadine,	Hypertonic sodium bicarbonate (or hypertonic saline) for cardiac conduction delays and monomorphic ventricular tachycardia; lidocaine for monomorphic ventricular

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
	cocaine), opioids (meperidine, propoxyphene), orphenadrine, quinoline antimalarials (chloroquine, hydroxychloroquine, quinine), cyclic antidepressants (see above)	ventricular tachycardia. Class Ia, Ic, and III antiarrhythmics also block potassium channels during phases 2 and 3 (repolarization) of the action potential, prolonging the JT interval and promoting early after-depolarizations and polymorphic (torsades des pointes) ventricular tachycardia. Similar effects on neuronal membrane channels cause CNS dysfunction. Some agents also block α-adrenergic and cholinergic receptors or have opioid effects (see above and **Chap. 446**).	antihistamines, carbamazepine, disopyramide, antipsychotics, and cyclic antidepressants (see above); opioid effects with meperidine and propoxyphene **(see Chap. 446)**; cinchonism (hearing loss, tinnitus, nausea, vomiting, vertigo, ataxia, headache, flushing, diaphoresis), and blindness with quinoline antimalarials	tachycardia (except when due to class Ib antiarrhythmics); magnesium, isoproterenol, and overdrive pacing for polymorphic ventricular tachycardia; physostigmine for anticholinergic effects (see above); naloxone for opioid effects **(see Chap. 446)**; extracorporeal removal for some agents (see text).

[a]

See above and **Chap. 447**. [b]See above and **Chap. 446**.

Abbreviations: AGMA, anion-gap metabolic acidosis; ARDS, adult respiratory distress syndrome; CNS, central nervous system; GABA, γ-aminobutyric acid; GBL, γ-butyrolactone; GHB, γ-hydroxybutyrate; G6PD, glucose-6-phosphate dehydrogenase; MAO, monoamine oxidase; NDMA, *N*-methyl-D-aspartate; 2-PAM, pralidoxime; SIADH, syndrome of inappropirate antidiuretic hormone secretion.

SPECIFIC TOXIC SYNDROMES AND POISONINGS

Table 450-4 summarizes the pathophysiology, clinical features, and treatment of toxidromes and poisonings that are common, produce life-threatening toxicity, or require unique therapeutic interventions. In all cases, treatment should include attention to the general principles discussed above and, in particular, supportive care. Poisonings not covered in this chapter are discussed in specialized texts.

Alcohol, cocaine, hallucinogen, and opioid poisoning and alcohol and opioid withdrawal are discussed in Chaps. 445–447; nicotine addiction is discussed in Chap. 448; acetaminophen poisoning is discussed in Chap. 333; the neuroleptic malignant syndrome is discussed in Chap. 427; and heavy metal poisoning is discussed in Chap. 449.

GLOBAL CONSIDERATIONS


Risks of poisoning in the United States and throughout the world are in transition. Patterns of travel, immigration, and internet consumerism should always be considered in patients suspected of poisoning or overdose without a clear etiology. Immigrants into various countries may have underlying poisoning from various metals from work or the environment where they previously lived; herbal remedies, food products, and cosmetics imported from overseas may be contaminated with metals, toxic plants, or other pharmaceutical contaminants; and new drugs of abuse that originate in one part of the world quickly circulate due to the ease afforded by the internet. Expanding the history at the time of evaluation, recruiting the assistance of global health specialists, and ordering expanded laboratory panels may be indicated.

FURTHER READING

Dart RC et al: Expert consensus guidelines for stocking of antidotes in hospitals that provide emergency care. Ann Emerg Med 54:386, 2009. [PubMed: 19406507]

Hoffman RS et al (eds): *Goldfrank's Toxicologic Emergencies*, 10th ed. New York, McGraw-Hill, 2015.

Mazer-Amirshahi M et al: Trends in Emergency Department Resource Utilization for Poisoning-Related Visits, 2003–2011. J Med Toxicol 12:248, 2016. [PubMed: 27342464]

Mowry JB et al: 2015 Annual report of the American Association of Poison Control Centers' National Poison Data System (NPDS): 33rd Annual Report. Clin Toxicol 54:924, 2016.

Thompson TM et al: The general approach to the poisoned patient. Dis Mon 60:509, 2014. [PubMed: 25454682]

Welker K, Mycyk MB: Pharmacology in the geriatric patient. Emerg Med Clin North Am 34:469, 2016. [PubMed: 27475010]

Chapter 451: Disorders Caused by Venomous Snakebites and Marine Animal Exposures

Charles Lei; Kirsten B. Hornbeak; Paul S. Auerbach; Robert L. Norris

INTRODUCTION

This chapter outlines general principles for the evaluation and management of victims of envenomation and poisoning by venomous snakes and marine animals. Because the incidence of serious bites and stings is relatively low in developed nations, there is a paucity of relevant clinical research; as a result, therapeutic decision-making often is based on anecdotal information.

VENOMOUS SNAKEBITE

EPIDEMIOLOGY

The venomous snakes of the world belong to the families Viperidae (subfamily Viperinae: Old World vipers; subfamily Crotalinae: New World and Asian pit vipers), Elapidae (including cobras, coral snakes, sea snakes, kraits, and all Australian venomous snakes), Lamprophiidae (subfamily Atractaspidinae: burrowing asps), and Colubridae (a large family in which most species are nonvenomous and only a few are dangerously toxic to humans). Most snakebites occur in developing countries with temperate and tropical climates in which populations subsist on agriculture and fishing. Recent estimates indicate that somewhere between 1.2 million and 5.5 million snakebites occur worldwide each year, with 421,000–1,841,000 envenomations and 20,000–94,000 deaths. Such wide-ranging estimates reflect the challenges of collecting accurate data in the regions most affected by venomous snakes; many victims do not seek medical attention, and reporting and record keeping are generally poor.

SNAKE ANATOMY/IDENTIFICATION

The typical snake venom delivery apparatus consists of bilateral venom glands situated behind the eyes and connected by ducts to hollow anterior maxillary fangs. In viperids, these fangs are long and highly mobile; they are retracted against the roof of the mouth when the snake is at rest and brought to an upright position for a strike. In elapids, the fangs are shorter and relatively fixed in an erect position. Approximately 20% of pit viper bites and higher percentages of other snakebites (up to 75% for sea snakes) are “dry” bites; i.e., no venom is released. Significant envenomation probably occurs in ~50% of all venomous snakebites.

Differentiation of venomous from nonvenomous snake species can be difficult. Viperids are characterized by somewhat triangular heads (a feature shared with many harmless snakes), elliptical pupils (also seen in some nonvenomous snakes, such as boas and pythons), enlarged maxillary fangs, and, in pit vipers, heat-sensing organs (pits) on each side of the head that assist with locating prey and aiming strikes. Rattlesnakes possess a series of interlocking hollow keratin plates (the rattle) on the tip of the tail that emits a buzzing sound when the snake rapidly vibrates its tail; this sound serves as a warning signal to perceived threats. Identifying venomous snakes by color pattern is notoriously misleading, as many harmless snakes have color patterns that closely mimic those of venomous snakes found in the same region.

VENOMS AND CLINICAL MANIFESTATIONS

Snake venoms are highly variable and complex mixtures of enzymes, polypeptides, glycoproteins, and other constituents. Among the deleterious components are proteolytic enzymes that cause local tissue necrosis, affect the coagulation pathway at various steps, and impair organ function. Hemorrhagins cause vascular leakage, resulting in fluid shifts and spontaneous local and systemic bleeding. Hyaluronidases promote the spread of venom through connective tissue. Myocardial depressant factors reduce cardiac output, and bradykinins cause vasodilation and hypotension. Neurotoxins act at various sites of the neuromuscular junction to block transmission and cause muscle paralysis. Most snake venoms have multisystem effects on their victims.

After a venomous snakebite, the time to symptom onset and clinical presentation can be quite variable and depends on the species involved, the anatomic location of the bite, and the amount of venom injected. Envenomations by most viperids and some elapids with necrotizing venoms cause progressive local pain, soft-tissue swelling, and ecchymosis **(Fig. 451-1)**. Hemorrhagic or serum-filled vesicles and bullae may develop at the bite site over a period of hours to days. In serious bites, tissue loss can be significant **(Figs. 451-2 and 451-3)**. Systemic findings are extremely variable and can include generalized fatigue, nausea, changes in taste, mouth numbness, tachycardia or bradycardia, hypotension, muscle fasciculations, pulmonary edema, renal dysfunction, and spontaneous hemorrhage from essentially any anatomic site. Envenomations by neurotoxic elapids, such as kraits (*Bungarus* species), many Australian elapids (e.g., death adders [*Acanthophis* species] and tiger snakes [*Notechis* species]), some cobras (*Naja* species), some viperids (e.g., the South American rattlesnake [*Crotalus durissus*], and certain Indian Russell's vipers [*Daboia russelii*]), cause neurologic dysfunction. Early findings may consist of nausea and vomiting, headache, paresthesias or numbness, and altered mental status. Victims may develop cranial nerve abnormalities (e.g., ptosis, difficulty swallowing), followed by peripheral motor weakness. Severe envenomation may result in muscle paralysis, including the muscles of respiration, and lead to death from respiratory failure and aspiration. Sea snake envenomation results in local pain (variable), generalized myalgias, trismus, rhabdomyolysis, and progressive flaccid paralysis; these manifestations can be delayed for several hours.

FIGURE 451-1

Northern Pacific rattlesnake (*Crotalus oreganus oreganus*) envenomations. ***A.*** Moderately severe envenomation. Note edema and early ecchymosis 2 h after a bite to the finger. ***B.*** Severe envenomation. Note extensive ecchymosis 5 days after a bite to the ankle. *(Courtesy of Robert Norris, with permission.)*



A

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



B

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 451-2

Early stages of severe, full-thickness necrosis 5 days after a Russell's viper (*Daboia russelii*) bite in southwestern India. *(Courtesy of Robert Norris, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 451-3

Severe necrosis 10 days after a pit viper bite in a young child in Colombia. *(Courtesy of Jay Stanka, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TREATMENT

TREATMENT

Venomous Snakebite

FIELD MANAGEMENT

The most important aspect of prehospital care of a person bitten by a venomous snake is rapid transport to a medical facility equipped to provide supportive care (airway, breathing, and circulation) and antivenom therapy. Any jewelry or tight-fitting clothing near the bite should be removed to avoid constriction from anticipated soft-tissue swelling. Without a delay in transport, the wound should be cleaned with soap and running water and covered with a sterile dressing. It is reasonable to apply a splint to the bitten extremity to limit movement and lessen bleeding. If possible, the extremity should be maintained in a neutral position of comfort at approximately heart level. Attempting to capture and transport the offending snake, alive or dead, is not advised; instead, digital photographs taken from a safe distance may assist with snake identification and treatment decisions.

Most of the first-aid measures recommended in the past are of little benefit, and some actually worsen outcome. Incising and/or applying suction to the bite site should be avoided, as these measures exacerbate local tissue damage, increase the risk of infection, and have not been shown to be effective. Similarly ineffective and potentially harmful are the application of poultices, ice, and electric shock. Venom sequestration devices (e.g., lympho-occlusive bandages or tourniquets) are not advised, as they may

intensify local tissue damage by restricting the spread of potentially necrotizing venom. Tourniquet use can result in loss of function, ischemia, and limb amputation, even in the absence of envenomation. In developing countries, indigenous people should be encouraged to seek immediate treatment at a medical facility equipped with antivenom instead of consulting traditional healers and thus incurring significant delays in reaching appropriate care.

Elapid venoms that are primarily neurotoxic and have no significant effects on local tissue may be localized by pressure-immobilization, a technique in which the entire bitten limb is immediately wrapped with a bandage (e.g., crepe or elastic) and then immobilized. For this technique to be effective, the wrap pressure must be precise (40–70 mmHg in upper-extremity application and 55–70 mmHg in lower-extremity application) and the victim must be carried out of the field because walking generates muscle-pumping activity that—regardless of the anatomic site of the bite—will disperse venom into the systemic circulation. Pressure-immobilization should be used only in cases in which the offending snake is reliably identified and known to be primarily neurotoxic, the rescuer is skilled in pressure-wrap application, the necessary supplies are readily available, and the victim can be fully immobilized and carried to medical care—a rare combination of conditions, particularly in the regions of the world where such bites are most common.

HOSPITAL MANAGEMENT

Initial hospital management should focus on the victim's airway, breathing, and circulation. Patients with bites to the face or neck may require early endotracheal intubation to prevent loss of airway patency caused by rapid soft-tissue swelling. Vital signs, cardiac rhythm, oxygen saturation, and urine output should be closely monitored. Two large-bore IV lines should be established in unaffected extremities. Because of the potential for coagulopathy, venipuncture attempts should be minimized and noncompressible sites (e.g., a subclavian vein) avoided. Early hypotension is due to pooling of blood in the pulmonary and splanchnic vascular beds. Later, systemic bleeding, hemolysis, and loss of intravascular volume into the soft tissues may play important roles. Fluid resuscitation with isotonic saline (20–40 mL/kg IV) should be initiated if there is any evidence of hemodynamic instability, and a trial of 5% albumin (10–20 mL/kg IV) may be undertaken if the response to saline infusion is inadequate. Vasopressors (e.g., norepinephrine, dopamine) should be added only if venom-induced shock persists after aggressive volume resuscitation and antivenom administration (see below). Invasive hemodynamic monitoring (central venous and/or continuous arterial pressures) can be helpful in such cases, although gaining central vascular access is risky if coagulopathy has developed.

A thorough history (including the time of the bite and any symptoms of envenomation) should be obtained and a complete physical examination performed. Bandages or wraps applied in the field should be removed once IV access has been obtained, with cognizance that the release of such ligatures may result in hypotension or dysrhythmias when stagnant acidotic blood containing venom is released into the systemic circulation. To objectively evaluate the progression of local envenomation, the leading edge of swelling, ecchymosis, and tenderness should be marked and limb circumference should be measured every 15 min until the local tissue effects have stabilized. During this period of observation, the bitten extremity should be positioned at approximately heart level. Victims of neurotoxic envenomation should be monitored closely for

evidence of cranial nerve dysfunction (e.g., ptosis) that may precede more overt signs of impending airway compromise (e.g., difficulty swallowing, respiratory insufficiency) necessitating endotracheal intubation and mechanical ventilation.

Blood should be drawn for laboratory evaluation as soon as possible. Important studies include a complete blood count to determine the degree of hemorrhage or hemolysis and to detect thrombocytopenia; blood type and cross-matching; assessment of renal and hepatic function; coagulation studies to diagnose consumptive coagulopathy; measurement of creatine kinase for suspected rhabdomyolysis; and testing of urine for blood or myoglobin. In developing regions, the 20-min whole-blood clotting test can be used to reliably diagnose coagulopathy. A few milliliters of fresh blood are placed in a new, clean, plain glass receptacle (e.g., a test tube) and left undisturbed for 20 min. The tube is then tipped once to 45°. If the blood is still liquid and a clot has not formed, coagulopathy is present. Arterial blood gas studies, electrocardiography, and chest radiography may be helpful in severe envenomations or when there is significant comorbidity. Any arterial puncture in the setting of coagulopathy requires great caution and must be performed at an anatomic site amenable to direct-pressure tamponade. After antivenom therapy (see below), laboratory values should be rechecked every 6 h until clinical stability is achieved. If initial laboratory values are normal, the complete blood count and coagulation studies should be repeated every hour until it is clear that no systemic envenomation has occurred.

The mainstay of treatment of a venomous snakebite resulting in significant envenomation is prompt administration of specific antivenom. Antivenoms are produced by injecting animals (generally horses or sheep) with venoms from medically important snakes. Once the stock animals develop antibodies to the venoms, their serum is harvested and the antibodies are isolated for antivenom preparation. The goal of antivenom administration is to allow antibodies (or antibody fragments) to bind and deactivate circulating venom components before they can attach to target tissues and cause deleterious effects. Antivenoms may be monospecific (directed against a particular snake species) or polyspecific (covering several species in a geographic region) but rarely offer cross-protection against snake species other than those used in their production unless the species are known to have homologous venoms. Thus, antivenom selection must be specific for the offending snake; if the antivenom chosen does not contain antibodies to that snake's venom components, it will provide no benefit and may lead to unnecessary complications (see below). In the United States, assistance in finding appropriate antivenom can be obtained from a regional poison-control center, which can be reached by telephone 24 h a day at (800) 222-1222.

For victims of bites by viperids or cytotoxic elapids, indications for antivenom administration include significant progressive local findings (e.g., soft-tissue swelling that crosses a joint, involves more than half the bitten limb, or is rapidly spreading; extensive blistering or bruising; severe pain) and any evidence of systemic envenomation (systemic symptoms or signs, laboratory abnormalities). Caution must be used in determining the significance of isolated pain or soft-tissue swelling after the bite of an unidentified snake because the saliva of some relatively harmless species can cause mild discomfort or edema at the bite site; in such bites, antivenoms are useless and potentially harmful. Antivenoms have limited efficacy in preventing local tissue damage caused by necrotizing venoms, as venom components bind to local tissues very quickly,

before antivenom administration can be initiated. Nevertheless, antivenom should be administered as soon as the need for it is identified to limit further tissue damage and systemic effects. Antivenom administration after bites by neurotoxic elapids is indicated at the first sign of any evidence of neurotoxicity (e.g., cranial nerve dysfunction, peripheral neuropathy). In general, antivenom is effective only in reversing active venom toxicity; it is of no benefit in reversing effects that have already been established (e.g., renal failure, established paralysis) and that will improve only with time and other supportive therapies.

Specific comments related to the management of venomous snakebites in the United States and Canada appear in **Table 451-1**. The package insert for the selected antivenom should be consulted regarding species covered, method of administration, starting dose, and need (if any) for re-dosing. The mobile app SnakeBite911 ER, developed by the manufacturers of the Crotalidae Polyvalent Immune Fab (CroFab) (Ovine) antivenom, offers a treatment algorithm for pit viper envenomations, has photos and detailed descriptions of North American pit viper species, and can help locate nearby hospitals. Whenever possible, it is advisable for health care providers to seek advice from experts in snakebite management regarding indications for and dosing of antivenom.

Antivenom should be administered only by the IV route, and the infusion should be started slowly, with the treating clinician at the bedside ready to immediately intervene at the first signs of an acute adverse reaction. In the absence of an adverse reaction, the rate of infusion can be increased gradually until the full starting dose has been administered (over a total period of ~1 h). Further antivenom may be necessary if the patient's acute clinical condition worsens or fails to stabilize or if venom effects that were initially controlled recur. The decision to administer further antivenom to a stabilized patient should be based on clinical evidence of the persistent circulation of unbound venom components. For viperid bites, antivenom administration generally should be continued until the victim shows definite improvement (e.g., reduced pain, stabilized vital signs, restored coagulation). Neurotoxicity from elapid bites may be more difficult to reverse with antivenom. Once neurotoxicity is established and endotracheal intubation is required, further doses of antivenom are unlikely to be beneficial. In such cases, the victim must be maintained on mechanical ventilation until recovery, which may take days or weeks.

Adverse reactions to antivenom administration include early (anaphylaxis) and late (serum sickness) hypersensitivity reactions. Clinical manifestations of early hypersensitivity range from tachycardia, rigors, vomiting, and urticaria to more serious dyspnea, laryngeal edema, bronchospasm, and hypotension. Skin testing for potential early hypersensitivity, although recommended by some antivenom manufacturers, is neither sensitive nor specific and is of no benefit. Worldwide, the quality of antivenoms is highly variable; rates of acute anaphylactic reactions to some of these products exceed 50%. There is some evidence supporting routine pretreatment with low-dose SC epinephrine (0.25 mg of 1:1000 aqueous solution) to prevent acute anaphylactic reactions after antivenom infusion; although widely practiced, prophylactic use of antihistamines and glucocorticoids has not proved beneficial. Modest expansion of the patient's intravascular volume with crystalloids may blunt episodes of acute hypotension during antivenom infusion. Epinephrine and airway equipment should always be immediately available. If the patient develops an anaphylactic reaction to antivenom, the infusion should be stopped temporarily and the reaction treated

immediately with IM epinephrine (0.01 mg/kg up to 0.5 mg), an IV antihistamine (e.g., diphenhydramine, 1 mg/kg up to 50 mg), and a glucocorticoid (e.g., hydrocortisone, 2 mg/kg up to 100 mg). Once the reaction has been controlled, if the severity of the envenomation warrants additional antivenom, the dose should be diluted further in isotonic saline and restarted as soon as possible. In rare cases of refractory hypotension, a concomitant IV infusion of epinephrine may be initiated and titrated to clinical effect while antivenom is administered. The patient must be monitored very closely during such therapy, preferably in an intensive care setting. Serum sickness typically develops 1–2 weeks after antivenom administration and may present as myalgias, arthralgias, fever, chills, urticaria, lymphadenopathy, and renal or neurologic dysfunction. Treatment for serum sickness consists of systemic glucocorticoids (e.g., oral prednisone, 1–2 mg/kg daily) until all symptoms have resolved, with a subsequent taper over 1–2 weeks. Oral antihistamines and analgesics may provide additional relief of symptoms.

Blood products are rarely necessary in the management of an envenomated patient. The venoms of many snake species can deplete coagulation factors and cause a decrease in platelet count or hematocrit. Nevertheless, these components usually rebound within hours after administration of adequate antivenom. If the need for blood products is thought to be great (e.g., a dangerously low platelet count in a hemorrhaging patient), these products should be given only after adequate antivenom administration to avoid fueling ongoing consumptive coagulopathy.

Rhabdomyolysis should be managed in standard fashion. Victims who develop acute renal failure should be evaluated by a nephrologist and referred for hemodialysis or peritoneal dialysis as needed. Such renal failure, which usually is due to acute tubular necrosis, is frequently reversible. If bilateral cortical necrosis occurs, however, the prognosis for renal recovery is less favorable, and long-term dialysis with possible renal transplantation may be necessary.

Most snake envenomations involve subcutaneous deposition of venom. On occasion, however, venom can be injected more deeply into muscle compartments, particularly if the offending snake was large and the bite occurred on the lower leg, forearm, or hand. Intramuscular swelling of the affected extremity may be accompanied by severe pain, decreased strength, altered sensation, cyanosis, and apparent pulselessness—signs suggesting a muscle compartment syndrome. If there is clinical concern that subfascial muscle edema may be impeding tissue perfusion, intracompartmental pressures should be measured by a minimally invasive technique (e.g., with a wick catheter or digital readout device). If the intracompartmental pressure is high (>30–40 mmHg), the extremity should be kept elevated while antivenom is administered. A dose of IV mannitol (1 g/kg) may be given in an effort to reduce muscle edema if the patient is hemodynamically stable. If the intracompartmental pressure remains elevated after 1 h of such therapy, a surgical consultation should be obtained for possible fasciotomy. Although evidence from animal studies suggests that fasciotomy may actually worsen myonecrosis, compartmental decompression may still be necessary to preserve neurologic function. Fortunately, the incidence of compartment syndrome is very low after snakebite, with fasciotomies required in <1% of cases. Nevertheless, vigilance is essential.

Acetylcholinesterase inhibitors (e.g., edrophonium and neostigmine) may promote neurologic improvement in patients bitten by snakes with postsynaptic neurotoxins. Snakebite victims with objective evidence of

neurologic dysfunction should receive a test dose of acetylcholinesterase inhibitors, as outlined in Table 451-2. If they exhibit improvement, additional doses of long-acting neostigmine can be administered as needed. Close monitoring is required to prevent aspiration if repetitive dosing of neostigmine is used in an attempt to obviate endotracheal intubation. Acetylcholinesterase inhibitors are not a substitute for administration of appropriate antivenom when available.

Care of the bite wound includes simple cleansing with soap and water; application of a dry, sterile dressing; and splinting of the affected extremity with padding between the digits. Once antivenom therapy has been initiated, the extremity should be elevated above heart level to reduce swelling. Patients should receive tetanus immunization as appropriate. Prophylactic antibiotics are generally unnecessary after bites by North American snakes because the incidence of secondary infection is low. In some regions, secondary bacterial infection is more common and the consequences are serious; in these regions, prophylactic treatment with broad-spectrum antibiotics may be appropriate. Antibiotics may also be considered if misguided first-aid efforts have included incision or mouth suction of the bite site. Pain control should be achieved with acetaminophen or opioid analgesics. Salicylates and nonsteroidal anti-inflammatory agents should be avoided because of their effects on blood clotting.

Wound care in the days after the bite should include careful aseptic debridement of clearly necrotic tissue once coagulopathy has been fully reversed. Intact serum-filled vesicles or hemorrhagic blebs should be left undisturbed. If ruptured, they should be debrided with sterile technique. Any debridement of damaged muscle should be conservative because there is evidence that such muscle may recover to a significant degree after antivenom therapy.

Physical therapy should be started as soon as possible to assist the patient in returning to a functional state. The incidence of long-term loss of function (e.g., reduced range of motion, impaired sensory function) is unclear but is probably quite high (>30%), particularly after viperid bites.

Any patient with signs of envenomation should be observed in the hospital for at least 24 h. In North America, a patient with an apparently “dry” viperid bite should be closely monitored for at least 8 h before discharge, as significant toxicity occasionally develops after a delay of several hours. The onset of systemic symptoms commonly is delayed for a number of hours after bites by certain elapids (including coral snakes, *Micrurus* species), some non–North American viperids (e.g., the hump-nosed pit viper [*Hypnale hypnale*]), and sea snakes. Patients bitten by these snakes should be observed in the hospital for at least 24 h. Admission to an intensive care setting is advised for patients with progressive clinical findings despite initial antivenom administration; those bitten in the head, neck, or other high-risk sites; and those who develop an acute hypersensitivity reaction to antivenom.

At hospital discharge, victims of venomous snakebites should be warned about symptoms and signs of wound infection, antivenom-related serum sickness, and potential long-term sequelae, such as pituitary insufficiency from Russell’s viper (*D. russelii*) bites. If coagulopathy developed in the acute stages of envenomation, it can recur during the first 2–3 weeks after the bite; in such cases, victims should be warned

to avoid elective surgery or activities posing a high risk of trauma during this period. Outpatient analgesic treatment, wound management, and physical therapy should be provided.

TABLE 451-1

Management of Venomous Snakebite in the United States and Canada[a]

Pit Viper Bites: Rattlesnakes (*Crotalus* and *Sistrurus* spp.), Cottonmouth Water Moccasins (*Agkistrodon piscivorus*), and Copperheads (*Agkistrodon contortrix*)

- Stabilize airway, breathing, and circulation.
- Institute monitoring (vital signs, cardiac rhythm, and oxygen saturation).
- Establish two large-bore IV lines.
  - If the patient is hypotensive, administer a normal saline bolus (20–40 mL/kg IV).
  - If hypotension persists, consider 5% albumin (10–20 mL/kg IV).
- Take thorough history and perform complete physical examination.
- Identify offending snake if possible.
- Measure and record circumference of bitten extremity q15min until swelling has stabilized.
- Order laboratory studies (CBC, blood type and cross-matching, metabolic panel, PT/INR/PTT, fibrinogen level, FDP, CK, urinalysis).
  - If normal, repeat CBC and coagulation studies every hour until it is clear that no systemic envenomation has occurred.
  - If abnormal, repeat 6 h after antivenom administration (see below).
- Determine severity of envenomation.
  - None: fang marks only (“dry” bite)
  - Mild: local findings only (e.g., pain, ecchymosis, nonprogressive swelling)
  - Moderate: swelling that is clearly progressing, systemic symptoms or signs, and/or laboratory abnormalities
  - Severe: neurologic dysfunction, respiratory distress, and/or cardiovascular instability/shock
- Contact regional poison-control center.
- Locate and administer antivenom as indicated: Crotalidae Polyvalent Immune Fab (CroFab) (Ovine) (BTG International Inc., West Conshohocken, PA).
  - Starting dose
    - Based on severity of envenomation
    - None or mild: none
    - Moderate: 4–6 vials
    - Severe: 6 vials
    - Dilute reconstituted vials in 250 mL of normal saline.
    - Infuse IV over 1 h (with medical provider in close attendance).
    - Start at rate of 20–50 mL/h for first 10 min.
    - If there is no allergic reaction, increase rate to 250 mL/h.
  - If there is an acute reaction to antivenom:
    - Stop infusion.
    - Treat with standard doses of epinephrine (IM or IV; latter route only in setting of severe hypotension), antihistamines (IV), and glucocorticoids (IV).

- When reaction is controlled, restart antivenom as soon as possible (may further dilute in larger volume of normal saline).
  - Monitor clinical status over 1 h.
  - Stabilized or improved: Admit to hospital.
  - Progressing or unimproved: Repeat starting dose. Continue this pattern until patient's condition is stabilized or improved.
  - Blood products are rarely needed; if required, they should be given only after antivenom administration.
  - Provide tetanus immunization as needed.
  - Prophylactic antibiotics are unnecessary unless prehospital care included incision or mouth suction.
  - Pain management: Administer acetaminophen and/or opioids as needed; avoid salicylates and nonsteroidal anti-inflammatory agents.
  - Admit victim to hospital. (If there is no evidence of envenomation, monitor for 8 h before discharge.)
  - Give additional CroFab (2 vials q6h for 3 additional doses, with close monitoring).
  - Monitor for evidence of rising intracompartmental pressures (see text).
  - Provide wound care (see text).
  - Start physical therapy (see text).
  - At discharge, warn patient of possible recurrent coagulopathy and symptoms/signs of serum sickness.

Coral *Snakebites* (*Micrurus* spp. and *Micruroides euryxanthus*)

- Stabilize airway, breathing, and circulation.
- Institute monitoring (vital signs, cardiac rhythm, and oxygen saturation).
- Establish one large-bore IV line and initiate normal saline infusion.
- Take thorough history and perform complete physical examination.
- Identify offending snake if possible.
- Laboratory studies are unlikely to be helpful.
- Contact regional poison-control center.
- Locate and administer antivenom as indicated: Antivenin (*Micrurus fulvius*) (Equine) (commonly referred to as North American Coral Snake Antivenin; Wyeth Pharmaceuticals, New York, NY).[b]
  - Refer to antivenom package insert.
  - Dilute 3–5 reconstituted vials in 250 mL of normal saline.
  - Infuse IV over 1 h (with medical provider in close attendance).
  - If signs of envenomation progress despite initial antivenom, repeat the starting dose; up to 10 vials total may be required.
- If there is an acute adverse reaction to antivenom:
  - Stop infusion.
  - Treat with standard doses of epinephrine (IM or IV; latter route only in setting of severe hypotension), antihistamines (IV), and glucocorticoids (IV).
  - When reaction is controlled, restart antivenom as soon as possible (may further dilute in larger volume of normal saline).

- If there is any evidence of neurologic dysfunction (e.g., any cranial nerve abnormalities such as ptosis):
  - Administer trial of acetylcholinesterase inhibitors (see Table 451-2).
  - With any evidence of difficulty swallowing or breathing, proceed with endotracheal intubation and ventilatory support (may be required for days or weeks).
- Provide tetanus immunization as needed.
- Prophylactic antibiotics are unnecessary unless prehospital care included incision or mouth suction.
- Admit victim to hospital (intensive care unit) even if there is no evidence of envenomation; monitor for at least 24 h.

[a]These recommendations are specific to the care of victims of venomous snakebites in the United States and Canada and should not be applied to bites in other regions of the world. [b]At the time of this writing, a single lot of antivenom remained, with an extended expiration date of January 31, 2018.

Abbreviations: CBC, complete blood count; CK, creatine kinase; FDP, fibrin degradation products; PT/INR/PTT, prothrombin time/international normalized ratio/partial thromboplastin time.

TABLE 451-2

Use of Acetylcholinesterase Inhibitors in Envenomations by Neurotoxic Snakes and Cone Snails

1. Patients with clear, objective evidence of neurotoxicity (e.g., ptosis or inability to maintain upward gaze) should receive a test dose of edrophonium (if available) or neostigmine.
   a. Pretreat with atropine: 0.6 mg IV (children, 0.02 mg/kg with a minimum of 0.1 mg)
   b. Treat with:
      Edrophonium: 10 mg IV (children, 0.25 mg/kg)
      or
      Neostigmine: 0.02 mg/kg IV or IM (children, 0.04 mg/kg)
2. If objective improvement is evident after 30 min, treat with:
   a. Neostigmine: 0.5 mg IV, IM, or SC (children, 0.01 mg/kg) every hour as needed
   b. Atropine: 0.6 mg as IV continuous infusion over 8 h (children, 0.02 mg/kg over 8 h)
3. Closely monitor the airway and perform endotracheal intubation as needed.

MORBIDITY AND MORTALITY

The overall mortality rates for victims of venomous snakebites are low in regions with rapid access to medical care and appropriate antivenoms. In the United States, for example, the mortality rate is <1% for victims who

receive antivenom. Eastern and western diamondback rattlesnakes (*Crotalus adamanteus* and *Crotalus atrox*, respectively) are responsible for the majority of snakebite deaths in the United States. Snakes responsible for large numbers of deaths in other countries include cobras (*Naja* species), carpet and saw-scaled vipers (*Echis* species), Russell's vipers (*D. russelii*), large African vipers (*Bitis* species), lancehead pit vipers (*Bothrops* species), and tropical rattlesnakes (*C. durissus*).

The incidence of morbidity—defined as permanent functional loss in a bitten extremity—is difficult to estimate but is substantial. Morbidity may be due to muscle, nerve, or vascular injury or to scar contracture. Such morbidity can have devastating consequences for victims in the developing world when they lose the ability to work and provide for their families. In the United States, functional loss tends to be more common and severe after rattlesnake bites than after bites by copperheads (*Agkistrodon contortrix*) or water moccasins (*Agkistrodon piscivorus*).

GLOBAL CONSIDERATIONS

 In many developing countries where snakebites are common, limited access to medical care and antivenoms contributes to high rates of morbidity and mortality. Often, the available antivenoms are inappropriate and ineffective against the venoms of medically important indigenous snakes. In those regions, further research is necessary to determine the actual impact of venomous snakebites and the specific antivenoms needed in terms of both quantity and spectrum of coverage. Without accurate statistics, it is difficult to persuade antivenom manufacturers to begin and sustain production of appropriate antisera in developing nations. There is evidence that antivenoms can be produced by much more cost-effective methods than those currently being employed. Just as important as getting the correct antivenoms into underserved regions is the need to educate populations about snakebite prevention and train medical care providers in proper management approaches. Local protocols written with significant input from experienced providers in the region of concern should be developed and distributed. Appropriate antivenoms must be available at the most likely first point of medical contact for patients (e.g., primary health centers) in order to minimize the common practice of referring victims to more distant, higher levels of care for the initiation of antivenom therapy. Those who care for snakebite victims in these often-remote clinics must have the skills and confidence required to begin antivenom treatment (and to treat possible reactions) as soon as possible when needed.

MARINE ENVENOMATIONS

Much of the management of envenomation by marine animals is supportive in nature. A subset of envenomations can threaten life or limb. Specific antivenom can be used when appropriate.

INVERTEBRATES

Cnidarians

Cnidarians, such as hydroids, fire coral, jellyfish, Portuguese men-of-war, and sea anemones, produce specialized living stinging organelles called *cnidocysts* (a term that encompasses nematocysts, ptychocysts, and spirocysts). In the stinging process, cnidocysts are released and discharged upon mechanosensory stimulation. The venom from these organisms contains bioactive substances, such as tetramine, 5-hydroxytryptamine, histamine, serotonin, and high-molecular-weight toxins, all of which can, among other effects, change the permeability of cells to ions. Victims usually report immediate prickling or burning, pruritus, paresthesias, and painful throbbing with radiation. The skin becomes reddened, darkened, edematous, and blistered and may show signs of superficial necrosis. A legion of neurologic, cardiovascular, respiratory, rheumatologic, gastrointestinal, renal, and ocular symptoms have been described, especially following stings from anemones, *Physalia* species, and scyphozoans. Anaphylaxis is possible. Hundreds of deaths have been reported, many of them caused by *Chironex fleckeri, Stomolophus nomurai, Physalia physalis*, and *Chiropsalmus quadrumanus*. Irukandji syndrome, associated with the Australian jellyfish *Carukia barnesi* and other species, is a potentially fatal condition that most commonly is characterized by hypertension; severe back, chest, and abdominal pain; nausea and vomiting; headache; sweating; and, in the most serious cases, myocardial troponin leak, pulmonary edema, and ultimately hypotension. This syndrome is thought to be mediated, at least in part, by the release of endogenous catecholamines followed by cytokines and nitric oxide.

Envenomations by different cnidarians (typified by jellyfish) may respond differently to similar topical therapies; thus, the recommendations in this chapter must be tailored to local species and clinical practices. During stabilization, the skin should be decontaminated immediately with a generous application of saline. Vinegar (5% acetic acid) is an all-purpose agent that appears to be useful for relieving pain caused by a large number of species. If unavailable, rubbing alcohol (40–70% isopropyl alcohol) can sometimes be used, although some data show it may worsen nematocyst discharge in particular species.

For the sting of the venomous box jellyfish (*C. fleckeri*), vinegar should be used, followed by local application of heat (up to 45°C/113°F) by immersion in hot water. Hot-water application is first-line treatment for mild to moderate *Physalia utriculus* (bluebottle jellyfish) stings. If hot water is unavailable, commercial (chemical) cold packs or ice packs applied over a thin dry cloth or plastic membrane are effective in alleviating bluebottle jellyfish stings. In general, rubbing leads to further stinging by adherent cnidocysts and should be avoided.

After decontamination, topical application of an anesthetic ointment (e.g., lidocaine, benzocaine), an antihistamine (e.g., diphenhydramine), or a glucocorticoid (e.g., hydrocortisone) may be helpful for symptom control. Persistent severe pain after decontamination may be treated with IV or oral opioid analgesics. Muscle spasms may respond to IV diazepam (2–5 mg, titrated upward as necessary). An antivenom is available from Commonwealth Serum Laboratories for stings from the box jellyfish found in Australian and Indo-Pacific waters. However, despite its reported clinical efficacy, some in vitro studies suggest that this antivenom cannot bind venom rapidly enough to account for its effects. Until further notice, current recommendations for its use apply (see "Sources of Antivenoms and Other Assistance," below). Treatment of Irukandji syndrome may require administration of opioid analgesics and aggressive treatment of hypertension (e.g.,

phentolamine, 5 mg IV). All victims with systemic reactions should be observed for at least 6–8 h for rebound from any therapy, and all elderly patients should be assessed for cardiac arrhythmias. Patients may suffer post-inflammatory hyperpigmentation and persistent cutaneous hypersensitivity in areas of skin contact.

Safe Sea, a "jellyfish-safe" sunblock (*www.nidaria.com*) applied to the skin before an individual enters the water, inactivates the recognition and discharge mechanisms of nematocysts; it has been tested successfully against a number of marine stingers and may prevent or diminish the effects of coelenterate stings. Whenever possible, a dive skin or wet suit should be worn when entering ocean waters.

Sea Sponges

Many sponges produce crinotoxins. As a result, touching a sea sponge may result in allergic contact dermatitis. Irritant dermatitis may result if the sponge's small spicules of silica or calcium carbonate penetrate the skin. It is impossible to distinguish between the allergic and spicule reactions, so the treatment is the same for both. Afflicted skin should be gently dried and adhesive tape, a commercial facial peel, or a thin layer of rubber cement used to remove embedded spicules. Vinegar should then be applied immediately, with repeated application for 10–30 min three or four times a day thereafter. Rubbing alcohol may be used if vinegar is unavailable. After spicule removal and skin decontamination, glucocorticoid or antihistamine cream may be applied to the skin. Severe vesiculation should be treated with a 2-week tapering course of systemic glucocorticoids. Mild reactions subside in 3–7 days, while involvement of large areas of the skin may result in systemic symptoms of fever, dizziness, nausea, muscle cramps, and formication.

Annelid Worms

Annelid worms (bristleworms) possess rows of soft, cactus-like spines capable of inflicting painful stings. Contact results in symptoms similar to those of nematocyst envenomation. Without treatment, pain usually subsides over several hours, but inflammation may persist for up to a week **(Fig. 451-4)**. Victims should resist the urge to scratch because scratching may fracture retrievable spines. Visible bristles should be removed with forceps and adhesive tape or a commercial facial peel; alternatively, a thin layer of rubber cement can be used to entrap and then peel away the spines. Use of vinegar or rubbing alcohol or a brief application of lidocaine may provide additional relief. Local inflammation should be treated with topical or systemic glucocorticoids.

FIGURE 451-4

Rash on the hand of a diver from the spines of a bristleworm. *(Courtesy of Paul Auerbach, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Sea Urchins

Venomous sea urchins possess either hollow, venom-filled calcified spines or triple-jawed, globiferous pedicellariae with venom glands. Venom may also be found within the integumentary sheath on the external spine surface of certain species. The venom contains toxic components, including steroid glycosides, hemolysins, proteases, serotonin, and cholinergic substances. Contact with either venom apparatus produces immediate and intensely painful stings. Spines that enter a joint can cause synovitis that may progress to arthritis if the spines remain in or near the joint. If multiple spines penetrate the skin, the patient may develop systemic symptoms, including nausea, vomiting, numbness, muscular paralysis, and respiratory distress. A delayed hypersensitivity reaction 7–10 days after resolution of primary symptoms has been described.

The affected part should be immersed immediately in hot water to tolerance (up to 45°C/113°F). Pedicellariae should be removed by shaving so that envenomation cannot continue. Accessible embedded spines should be removed with care as they may fracture and leave remnants lodged in the victim. Residual dye from the surface of a spine remaining after the spine's removal may mimic a retained spine but is otherwise of no consequence. Soft-tissue radiography, ultrasonography, or MRI can confirm the presence of retained spines, which may warrant referral for surgical removal if the spines are near vital structures (e.g., joints, neurovascular bundles). Retained spines can cause the formation of granulomas that are amenable to excision or to intralesional injection with triamcinolone hexacetonide (5 mg/mL). Chronic granulomatous arthritis of the proximal interphalangeal joints has been treated with synovectomy and removal of granulation tissue. Eosinophilic pneumonia and local and diffuse neuropathies have been observed separately after penetration by multiple spines of the black sea urchin (presumed *Diadema* species). The pathophysiologies of these phenomena have not been determined.

Starfish

The crown-of-thorns *Acanthaster planci* produces venom in glandular tissue underneath the epidermis, which is released via its spiny surfaces **(Fig. 451-5)**. Skin puncture causes pain, bleeding, and local edema, usually with remission over 30–180 min. Multiple punctures may result in reactions such as local muscle paralysis; retained fragments may cause granulomatous lesions and synovitis. There has also been a case report of elevated liver enzymes after *A. planci* envenomation. Envenomated persons benefit from acute immersion therapy in hot water, local anesthesia, wound cleansing, imaging, and possible exploration to remove spines and foreign material.

FIGURE 451-5

Spines on the crown-of-thorns sea star (*Acanthaster planci*). *(Courtesy of Paul Auerbach, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Sea Cucumbers

Sea cucumbers produce holothurin (a cantharidin-like liquid toxin) in their body walls. This toxin is concentrated in the tentacular organs that are projected when the animal is threatened. Underwater, holothurin induces minimal contact dermatitis in the skin but can cause significant corneal and conjunctival irritation should ocular contact occur. A severe reaction can lead to blindness. Skin should be detoxified with vinegar or isopropyl alcohol. The eye should be anesthetized with 1–2 drops of 0.5% proparacaine and irrigated copiously with normal saline, with subsequent slit-lamp examination to identify corneal defects.

Cone Snails

Cone snails use a detachable dart-like tooth to inject conotoxins into victims, inducing paralysis. Punctures result in small, painful wounds followed by local ischemia, cyanosis, and numbness. Syncope, dysphagia,

dysarthria, ptosis, blurred vision, and pruritus also have been documented. Some envenomations induce paralysis leading to respiratory failure, coma, and death. There is no antivenom for treatment. Pressure immobilization (see "Octopuses," below), hot-water soaks, and local anesthetics have been used empirically with success. The wound should be inspected for a foreign body. Edrophonium has been recommended as therapy for paralysis if an edrophonium (Tensilon) test is positive (see Table 451-2).

Octopuses

Serious envenomations and deaths have followed bites of Australian blue-ringed octopuses (*Octopus maculosa* and *Octopus lunulata*). Although these animals rarely exceed 20 cm in length, their salivary venom contains a potent neurotoxin (maculotoxin) that inhibits peripheral-nerve transmission by blocking sodium conductance. Oral and facial numbness develop within several minutes of a serious envenomation and rapidly progress to total flaccid paralysis, including failure of respiratory muscles. Immediately after envenomation, a circumferential pressure-immobilization dressing 15 cm wide should be applied over a gauze pad (~7 × 7 × 2 cm) that has been placed directly over the sting. The dressing should be applied at venous-lymphatic pressure, with the preservation of distal arterial pulses. The limb should then be splinted. Once the victim has been transported to the nearest medical facility, the bandage can be released. There is no antidote and treatment is supportive. Patients with respiratory failure may need to be mechanically ventilated. The victim may remain awake although completely paralyzed, so analgesia and sedation should be provided as needed. Even with serious envenomations, significant recovery often takes place within 4–10 h, although complete recovery may require 2–4 days. Sequelae are uncommon unless related to hypoxia.

VERTEBRATES

As for all penetrating injuries, first-aid care should be undertaken. In addition, consideration must be given to local wound infection by marine *Vibrio* species and freshwater *Aeromonas hydrophila* as well as other "aquatic bacteria," particularly if spines and needles remain embedded.

Stingrays

A stingray injury is both an envenomation and a traumatic wound. Thoracic and cardiac penetration, major vessel laceration, and compartment syndrome have all been observed. The venom, which contains serotonin, 5′-nucleotidase, and phosphodiesterase, causes immediate, intense pain that peaks at 30–60 min and may last up to 48 h. The wound often becomes ischemic in appearance and heals poorly, with adjacent soft-tissue swelling and prolonged disability. Systemic effects include weakness, diaphoresis, nausea, vomiting, diarrhea, dysrhythmias, syncope, hypotension, muscle cramps, fasciculations, paralysis, and (in rare cases) death. Because of differences in the toxins present on the tissues covering the stingers, freshwater stingrays may cause more severe injuries than marine stingrays.

Scorpionfish

The designation *scorpionfish* encompasses members of the family Scorpaenidae and includes not only scorpionfish but also lionfish and stonefish. A complex venom with neuromuscular toxicity is delivered through 12 or 13 dorsal, 2 pelvic, and 3 anal spines. In general, the sting of a stonefish is regarded as the most serious (severe to life-threatening); that of the scorpionfish is of intermediate seriousness; and that of the lionfish is the least serious. Like that of a stingray, the sting of a scorpionfish is immediately and intensely painful. Pain from a stonefish envenomation may last for days and can be sufficiently intense to cause delirium. Systemic manifestations of scorpionfish stings are similar to those of stingray envenomations but may be more pronounced, particularly in the case of a stonefish sting, which may cause severe local tissue necrosis in addition to vital organ failure. The rare deaths that follow stonefish envenomation usually occur within 6–8 h. A commercially available stonefish antivenom from Commonwealth Serum Laboratories (see "Sources of Antivenoms and Other Assistance," below) can be used in cases of severe envenomation.

Other Fish

Three species of marine catfish—*Plotosus lineatus* (oriental catfish), *Bagre marinus* (sail catfish), and *Galeichthys felis* (common sea catfish)—as well as several species of freshwater catfish are capable of stinging humans. Venom is delivered through a single dorsal spine and two pectoral spines. Clinically, a catfish sting is comparable to that of a stingray, although marine catfish envenomations are generally more severe than those of their freshwater counterparts. Surgeonfish (doctorfish, tang), weeverfish, ratfish, and horned venomous sharks also have the capacity to envenomate humans.

Platypuses

The platypus is a venomous mammal. The male has a keratinous spur on each hind limb; the spur is connected to a venom gland within the upper thigh. Skin puncture causes soft-tissue edema and pain that may last for days or weeks. Care is supportive, and hot-water therapy does not appear to benefit the victim.

TREATMENT

TREATMENT

Marine Vertebrate Stings

The stings of all marine vertebrates are treated in a similar fashion. Except for stonefish and serious scorpionfish envenomations, no antivenom is available. The affected part should be immersed immediately in non-scalding hot water (45°C/113°F) for 30–90 min or until there is significant pain relief. Recurrent pain may respond to repeated hot-water treatment. Cryotherapy is contraindicated, and no data support the use of antihistamines or steroids. Systemic opioids as well as local wound infiltration or regional nerve block with lidocaine or bupivacaine can help alleviate pain. After soaking, the wound must be explored and debrided. Radiography (in particular, ultrasound or MRI) may be helpful in identification of foreign bodies. After exploration and debridement, the wound should be irrigated vigorously with warm sterile water, saline, or 1% povidone-iodine in solution. Bleeding usually can be controlled by sustained local pressure. In general, wounds should be left open to heal by secondary intention or treated by delayed primary closure. Tetanus

immunization should be provided as appropriate. Antibiotic treatment should be considered for serious wounds and for envenomations in immunocompromised hosts. The initial antibiotics should cover *Staphylococcus* and *Streptococcus* species. If the victim is immunocompromised, if a wound is primarily repaired, or if an infection develops, antibiotic coverage should be broadened to include *Vibrio* species. Infection with *Aeromonas* species is of similar concern for wounds sustained in freshwater.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Marine Envenomations

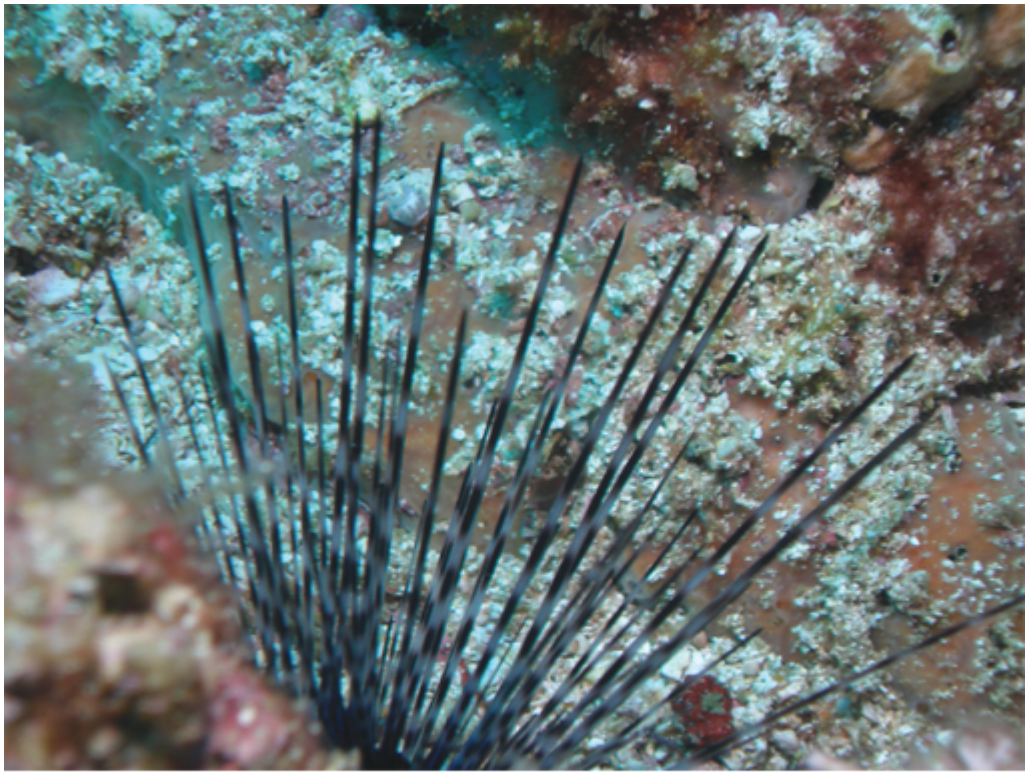
It is useful to be familiar with the local marine fauna and to recognize patterns of injury.

A large puncture wound or jagged laceration (particularly on the lower extremity) that is more painful than one would expect from the size and configuration of the wound is likely to be a stingray envenomation. Smaller punctures, sometimes associated with purple or dark discoloration, represent the activity of a sea urchin **(Fig. 451-6)** or starfish. Stony corals cause rough abrasions and, in rare instances, lacerations or puncture wounds.

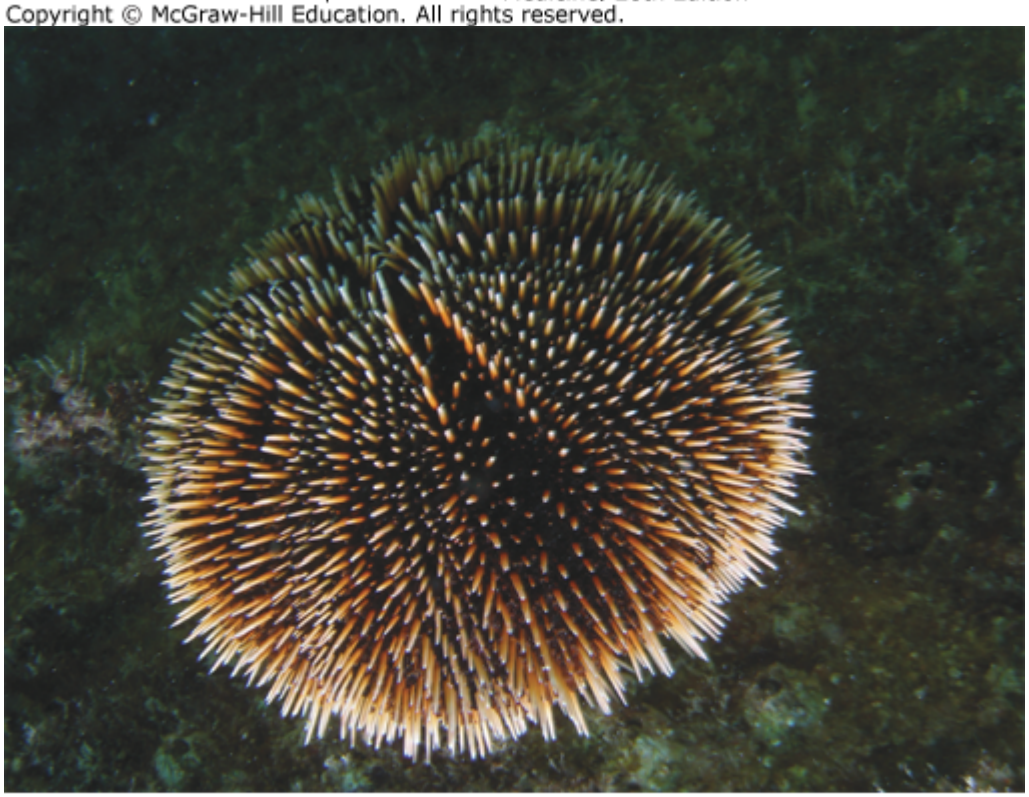
Coelenterate (marine invertebrate) stings sometimes create diagnostic skin patterns. A diffuse urticarial rash on exposed skin is often indicative of exposure to fragmented hydroids or larval anemones. A linear, whip-like print pattern appears where a jellyfish tentacle has contacted the skin. In the case of the box jellyfish, a cross-hatched appearance, followed by development of dark purple coloration within a few hours of the sting, heralds skin necrosis. An encounter with fire coral causes immediate pain and swollen red skin irritation in the pattern of contact, similar to but more severe than the imprint left by exposure to an intact feather hydroid. Seabather's eruption, caused by thimble jellyfish and larval anemones, is a diffuse, intensely pruritic rash consisting of clusters of erythematous macules or raised papules that follow the pattern of bathing attire **(Fig. 451-7)**. Toxic sponges create a burning and painful red rash on exposed skin, which may blister and later desquamate. Virtually all marine stingers invoke the sequelae of inflammation; thus local erythema, swelling, and adenopathy are fairly nonspecific.

FIGURE 451-6

Spiny sea urchins. *(Courtesy of Paul Auerbach, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 451-7

Erythematous, papular rash typical of seabather's eruption caused by thimble jellyfish and larval anemones. *(Courtesy of Paul Auerbach, with permission.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SOURCES OF ANTIVENOMS AND OTHER ASSISTANCE

In the United States, assistance in locating a specific antivenom can be obtained from a regional poison-control center. Divers Alert Network, a nonprofit organization designed to assist in the care of injured divers, also may help with the treatment of marine injuries. The network can be reached on the Internet at *www.diversalertnetwork.org* or by telephone 24 h a day at (919) 684-9111. Antivenom for box jellyfish (*C. fleckeri*) and stonefish (and severe scorpionfish) envenomations is made in Australia by Commonwealth Serum Laboratories (CSL; 45 Poplar Road, Parkville, Victoria, Australia 3052; *www.csl.com.au*; 61-3-9389-1911). When administering the box jellyfish antivenom, time is of the essence. For cardiac or respiratory decompensation, a minimum of 1 ampoule and up to 6 ampoules consecutively should be given IV, preferably in a 1:10 dilution with normal saline. For stonefish (or severe scorpionfish) envenomation, 1 ampoule of specific antivenom should be administered IM for every one or two punctures, to a maximum of 3 ampoules.

MARINE POISONINGS

CIGUATERA

Epidemiology and Pathogenesis

Ciguatera poisoning is the most common nonbacterial food poisoning associated with fish in the United States; most U.S. cases occur in Florida and Hawaii, although, with transportation of imported fish worldwide, all clinicians need to be aware of ciguatera. The poisoning almost exclusively involves tropical coral reef fish common in the Indian Ocean, the South Pacific, and the Caribbean Sea. Global estimates predict that 20,000–50,000 people may be affected by this poisoning each year, although it is suspected that 90% of cases go unreported. More than 400 different fish have been associated with ciguatera toxicity, but 75% of poisonings involve the reef-dwelling barracuda, snapper, jack, or grouper. Ciguatera toxin is created by warm-water marine dinoflagellates, primarily of the genus *Gambierdiscus toxicus*, whose consumption by grazing fish allows the toxin to bioaccumulate in the food chain. Ciguatoxins act on neuron voltage-gated sodium channels, most notably producing cardiac, gastrointestinal, and neurologic symptoms. These toxins are unaffected by freeze-drying, heat, cold, or gastric acid, and the fact that none of the toxins affects the odor, color, or taste of fish makes identification difficult. Toxins are found in the highest concentrations in the fish's skin, head, and viscera; therefore, consumption of these portions should be avoided.

Clinical Manifestations

Symptoms may develop within 15–30 min of ingestion but typically do so within 2–6 h and increase in severity over the ensuing 4–6 h. Most victims develop symptoms within 12 h of ingestion, and virtually all are afflicted within 24 h. The more than 150 symptoms and signs reported include those shown in **Table 451-3**. Diarrhea, vomiting, and abdominal pain usually occur first and develop 3–6 h after ingestion. Bradycardia, hypotension, dysesthesias, and paresthesias may occur. Symptoms may persist for 48 h and then generally resolve (even without treatment). A pathognomonic symptom is the reversal of hot and cold tactile perception, which develops in some persons after 3–5 days and may last for months. More severe reactions tend to occur on repeat exposure. Persons who have ingested parrotfish (scaritoxin) may develop classic ciguatera poisoning as well as a "second-phase" syndrome (after a delay of 5–10 days) of disequilibrium with locomotor ataxia, dysmetria, and resting or kinetic tremor. This syndrome may persist for 2–6 weeks.

TABLE 451-3

Representative Symptoms and Signs of Ciguatera Poisoning

System	Symptoms/Signs
Gastrointestinal	Abdominal pain, nausea, vomiting, diarrhea
Neurologic	Paresthesias, pruritus, tongue and throat numbness or burning, sensation of "carbonation" during swallowing, odontalgia or dental dysesthesias, dysphagia, tremor, fasciculations, athetosis, meningismus, aphonia, ataxia, vertigo, pain and weakness in the lower extremities, visual blurring, transient blindness, hyporeflexia, seizures, coma
Dermatologic	Conjunctivitis, maculopapular rash, skin vesiculations, dermographism
Cardiovascular	Bradycardia, heart block, hypotension, central respiratory failure[a]
Other	Chills, dysuria, dyspnea, dyspareunia, weakness, fatigue, nasal congestion and dryness, insomnia, hypersalivation, diaphoresis, headache, arthralgias, myalgias

[a]Tachycardia and hypertension may occur after potentially severe transient bradycardia and hypotension. Death is rare.

Diagnosis

The differential diagnosis of ciguatera includes paralytic shellfish poisoning, eosinophilic meningitis, type E botulism, organophosphate insecticide poisoning, tetrodotoxin poisoning, and psychogenic hyperventilation. At present, the diagnosis of ciguatera poisoning is made on clinical grounds. Liquid chromatography–mass spectrometry is available for ciguatoxins, and a ciguatoxin enzyme immunoassay or radioimmunoassay may be used to test suspected fish; however, these tests are of limited clinical value because most health care institutions do not have the equipment needed for their performance.

TREATMENT

TREATMENT

Ciguatera Poisoning

Therapy is supportive and based on symptoms. Hypotension should be treated with IV crystalloid and, in rare cases, a vasopressor. Bradyarrhythmias that lead to cardiac insufficiency and hypotension generally respond well to atropine (0.5 mg IV, up to 2 mg). Nausea and vomiting may be controlled with an antiemetic such as ondansetron (4–8 mg IV). Cool showers or the administration of hydroxyzine (25 mg orally every 6–8 h) may

relieve pruritus. Amitriptyline (25 mg orally twice a day) reportedly alleviates pruritus and dysesthesias and may decrease rates of subsequent development of chronic nerve symptoms. Gabapentin has shown some efficacy in the treatment of long-term sequelae of nerve pain. IV infusion of mannitol may be beneficial in moderate or severe cases in fluid-repleted patients, particularly for the relief of distressing neurologic or cardiovascular symptoms; however, the efficacy of this therapy has been challenged and has not been definitively proven. An initial IV dose of mannitol at 1 g/kg may be given over 45–60 min. If symptoms are alleviated, a second dose may be given within 3–4 h and a third dose the next day. Care must be taken to avoid dehydration. The mechanism of the drug's benefit against ciguatera intoxication is perhaps hyperosmotic water-drawing action, which reverses ciguatoxin-induced Schwann cell edema. Mannitol may also act in some fashion as a "hydroxyl scavenger" or may competitively inhibit ciguatoxin at the cell membrane. Activated charcoal is not recommended for ciguatera poisoning.

During recovery from ciguatera poisoning, the victim should exclude the following from the diet for 6 months: fish (fresh or preserved), fish sauces, shellfish, shellfish sauces, alcoholic beverages, nuts, and nut oils. Consumption of fish in ciguatera-endemic regions should be avoided. All oversized fish of any predacious reef species should be suspected of harboring ciguatoxin. Neither moray eels nor the viscera, head, or skin of tropical marine fish should be eaten.

DIARRHETIC SHELLFISH POISONING

Diarrhetic shellfish poisoning occurs with consumption of shellfish producing diarrheal illness. The causative agents are the lipophilic compound okadaic acid and the dinophysistoxins, which inhibit serine and threonine protein phosphatases, with consequent protein accumulation and continued secretion of fluid by intestinal cells leading to diarrhea. Shellfish acquire these toxins by feeding on dinoflagellates, particularly of the genera *Dinophysis* and *Prorocentrum*.

Symptoms include diarrhea, nausea, vomiting, abdominal pain, and chills. Onset occurs within 30 min to 12 h. The illness is usually self-limited; most patients recover in 3–4 days and only a few require hospitalization. Treatment is supportive and focused on hydration. Toxins can be detected in food samples by a mouse bioassay, an immunoassay, and fluorometric high-performance liquid chromatography (HPLC).

PARALYTIC SHELLFISH POISONING

Paralytic shellfish poisoning is induced by ingestion of any of a variety of feral or aquacultured filter-feeding organisms, including clams, oysters, scallops, mussels, chitons, limpets, starfish, and sand crabs. The source of their toxicity is the chemical toxin they accumulate and concentrate by feeding on various planktonic dinoflagellates (e.g., *Alexandrium*, *Pyrodinium*, *Gymnodinium*). These unicellular phytoplanktonic organisms form the foundation of the food chain, and in warm summer months these organisms "bloom" in nutrient-rich coastal temperate and semitropical waters. In the United States, paralytic shellfish poisoning is acquired primarily from seafood harvested in the Northeast, the Pacific Northwest, and Alaska. These planktonic species can release massive amounts of toxic metabolites into the water and cause mortality in bird and marine populations. The paralytic shellfish toxins are water soluble as well as heat and acid stable; they

cannot be destroyed by ordinary cooking or freezing. Contaminated seafood looks, smells, and tastes normal. The best-characterized, most potent, and most frequently identified paralytic shellfish toxin is saxitoxin, which appears to block sodium conductance, inhibiting neuromuscular transmission at the axonal and muscle membrane levels. A toxin concentration of >75 µg/100 g of foodstuff is considered hazardous to humans. During a “red tide” algal bloom, the concentration of saxitoxin in shellfish can exceed 9000 µg/100 g. A direct human serum assay to identify the toxin responsible for paralytic shellfish poisoning is not yet clinically available; a mouse bioassay that identifies saxitoxin in suspected shellfish is currently in use.

Intraoral and perioral paresthesias (notably of the lips, tongue, and gums) can occur within minutes to a few hours after ingestion of contaminated shellfish, and these paresthesias can progress rapidly to involve the neck and distal extremities. Other symptoms develop rapidly and can include lightheadedness, disequilibrium, incoordination, weakness, hyperreflexia, incoherence, dysarthria, hypersalivation, dysphagia, thirst, diarrhea, abdominal pain, nausea, vomiting, nystagmus, dysmetria, headache, diaphoresis, loss of vision, chest pain, and tachycardia. Flaccid paralysis and respiratory insufficiency may follow 2–12 h after ingestion. In the absence of hypoxia, the victim often remains alert but paralyzed. Up to 12% of patients may die.

TREATMENT

TREATMENT

Paralytic Shellfish Poisoning

Treatment is supportive and based on symptoms. If the victim comes to medical attention within the first few hours after poison ingestion, the stomach should be emptied by gastric lavage and then irrigated with 2 L (in 200-mL aliquots) of a solution of 2% sodium bicarbonate; this intervention has not been proven to be of benefit but is based on the notion that gastric acidity may enhance the potency of saxitoxin. Because respiratory insufficiency can be rapid in onset, induction of emesis is not advised. The administration of activated charcoal (50–100 g) and a cathartic (e.g., sorbitol, 20–50 g) makes empirical sense because these shellfish toxins are believed to bind well to charcoal. Some authors advise against administration of magnesium-based solutions (e.g., certain cathartics), cautioning that hypermagnesemia may contribute to suppression of nerve conduction.

The most serious concern is respiratory paralysis. The victim should be closely observed for respiratory distress for at least 24 h in a hospital. With prompt recognition of ventilatory failure and establishment of endotracheal intubation and assisted ventilation, anoxic myocardial and brain injury may be prevented. If the patient survives for 18 h, the prognosis is good for a complete recovery.

DOMOIC ACID POISONING (AMNESIC SHELLFISH POISONING)

Domoic acid poisoning occurs when humans consume shellfish containing the marine toxin domoic acid. The toxin is produced by marine diatoms of the *Pseudonitzchia* species and during algal blooms can bioaccumulate in filter-feeding shellfish. Clams, mussels, oysters, anchovies, and Dungeness crabs have all

been found to cause domoic acid poisoning. A water-soluble, heat-stable neuroexcitatory amino acid with biochemical analogues of kainic acid and glutamic acid, domoic acid binds to the kainate type of glutamate receptor with 3 times the affinity and 20 times the toxicity of kainic acid. The toxin is heat stable and is not affected by cooking or freezing. Shellfish can be tested for domoic acid by mouse bioassay and HPLC. The regulatory limit for domoic acid in shellfish is 20 parts per million.

The abnormalities noted within 24 h of ingesting contaminated shellfish include arousal, confusion, disorientation, and memory loss. The median time of onset is 5.5 h. Other prominent symptoms and signs include severe headache, nausea, vomiting, diarrhea, abdominal cramps, hiccups, arrhythmias, hypotension, seizures, ophthalmoplegia, pupillary dilation, piloerection, hemiparesis, mutism, grimacing, agitation, emotional lability, coma, copious bronchial secretions, and pulmonary edema. Histologic study of brain tissue taken at autopsy has shown neuronal necrosis or cell loss and astrocytosis, most prominently in the hippocampus and amygdaloid nucleus—findings similar to those in animals poisoned with kainic acid. Several months after the primary intoxication, victims may still display chronic residual memory deficits and motor neuropathy or axonopathy. Nonneurologic sequelae generally do not persist.

TREATMENT

TREATMENT

Domoic Acid Intoxication

Therapy is supportive and based on symptoms. IV fluids and antiemetics may be used for severe nausea, vomiting, and diarrhea. Because kainic acid neuropathology seems to be nearly entirely seizure mediated, the emphasis should be on anticonvulsive therapy; benzodiazepines, propofol, and barbiturates may be used.

HISTAMINE (SCOMBROID) FISH POISONING

Histamine fish poisoning, most often referred to as scombroid or pseudoallergenic fish poisoning, may be the most common type of seafood poisoning worldwide. It follows consumption of both scombroid (mackerel-like) fish (including albacore, bluefin, and yellowfin tuna; mackerel; saury; needlefish; wahoo; skipjack; and bonito) and nonscombroid fish (including dolphinfish, kahawai, sardine, black marlin, pilchard, anchovy, herring, amberjack, Australian ocean salmon, and bluefish).

Under conditions of inadequate preservation or refrigeration, the musculature of these dark- or red-fleshed fish undergoes decomposition by *Morganella morganii*, *Escherichia coli*, *Proteus* species, and *Klebsiella* species, with consequent decarboxylation of the amino acid L-histidine to histamine, histamine phosphate, and histamine hydrochloride. Histamine levels of 20–50 mg/100 g are noted in toxic fish, with levels >400 mg/100 g on occasion. Toxic levels can be reached with as few as 12 h of inadequate refrigeration. A second compound is thought to play a causative role in this intoxication because large doses of oral histamine do not reproduce the affliction. It is proposed that this unknown agent (possibly saurine, cadaverine, or putrescine) works by inhibiting the metabolism of histamine, promoting degranulation of mast cells to release

endogenous histamine, or acting as a histamine receptor agonist. The toxin or toxins involved are heat stable and are not destroyed by cooking or freezing. Affected fish may have a sharply metallic or peppery taste, although more often they are normal in appearance, color, and flavor. Not all persons who eat a contaminated fish necessarily become ill, perhaps because of uneven distribution of decay within the fish.

Symptoms develop within 15–90 min of ingestion. Most cases are mild, with tingling of lips and mouth, mild abdominal discomfort, and nausea. The more severe and commonly described presentation includes flushing (sharply demarcated; exacerbated by ultraviolet exposure; particularly pronounced on the face, neck, and upper trunk), a sensation of warmth, conjunctival hyperemia, pruritus, urticaria, and angioedema. This syndrome may progress to bronchospasm, nausea, vomiting, diarrhea, epigastric pain, abdominal cramps, dysphagia, headache, palpitations, tachycardia, dizziness, hypotension and cardiogenic shock. Without treatment, the symptoms generally resolve within 8–12 h. Because of blockade of gastrointestinal tract histaminase, the reaction may be more severe in a person who is concurrently ingesting isoniazid.

TREATMENT

TREATMENT

Scombroid Poisoning

Therapy is directed at reversing the histamine effect with either oral or IV antihistamines. If bronchospasm is severe, an inhaled bronchodilator—or in rare, severe circumstances, injected epinephrine—may be used. The use of activated charcoal is not recommended. Protracted nausea and vomiting may be controlled with a specific antiemetic, such as ondansetron or prochlorperazine. Hypotension should be treated with IV fluids. The persistent headache of scombroid poisoning may respond to cimetidine or a similar antihistamine if standard analgesics are not effective. It is important to inform the patient that the symptoms are related to eating improperly refrigerated fish and are not due to a fish allergy.

FURTHER READING

Auerbach PS et al: Envenomation by aquatic invertebrates, in *Auerbach's Wilderness Medicine*, 7th ed. PS Auerbach et al (eds). Philadelphia, Elsevier, 2017, pp 1679–1720.

Auerbach PS et al: Envenomations by aquatic vertebrates, in *Auerbach's Wilderness Medicine*, 7th ed. PS Auerbach et al (eds). Philadelphia, Elsevier, 2017, pp 1721–1742.

Currie BJ: Marine antivenoms. J Toxicol Clin Toxicol 41:301, 2003. [PubMed: 12807313]

Fenner P: Marine envenomation: An update—a presentation on the current status of marine envenomation first aid and medical treatments. Emerg Med Australas 12:295, 2000.

Kanaan NC et al: Wilderness Medical Society practice guidelines for the treatment of pitviper envenomations in the United States and Canada. Wilderness Environ Med 26:472, 2015. [PubMed: 26433731]

Minns AB et al: Seafood toxidromes, in *Auerbach's Wilderness Medicine*, 7th ed. PS Auerbach et al (eds). Philadelphia, Elsevier, 2017, pp 1766–1793.

Norris RL et al: Bites by venomous reptiles in Canada, the United States, and Mexico, in *Auerbach's Wilderness Medicine*, 7th ed. PS Auerbach et al (eds). Philadelphia, Elsevier, 2017, pp 729–759.

Warrell DA: Venomous bites, stings, and poisoning. Infect Dis Clin North Am 26:207, 2012. [PubMed: 22632635]

Warrell DA: Bites by venomous and nonvenomous reptiles worldwide, in *Auerbach's Wilderness Medicine*, 7th ed. PS Auerbach et al (eds.). Philadelphia, Elsevier, 2017, pp 760–827.

Chapter 452: Ectoparasite Infestations and Arthropod Injuries

Richard J. Pollack; Scott A. Norton

INTRODUCTION

Ectoparasites include arthropods and creatures from other phyla that infest the skin or hair of animals; the host animals provide them with sustenance and shelter. The ectoparasites may penetrate within or beneath the surface of the host or may attach by mouthparts and specialized claws. These organisms may inflict direct mechanical injury, consume blood or nutrients, induce hypersensitivity reactions, inoculate toxins, transmit pathogens, create openings in the skin for secondary bacterial infection, and incite fear or disgust. Human beings are the sole or obligate hosts for only a few kinds of ectoparasites but serve as facultative, dead-end, or paratenic (accidental) hosts for many others.

Arthropods that are ectoparasitic or otherwise cause injury include insects (such as lice, fleas, bedbugs, wasps, ants, bees, and flies), arachnids (spiders, scorpions, mites, and ticks), millipedes, and centipedes. Certain nematodes (helminths), such as the hookworms **(Chap. 226)**, are ectoparasitic in that they penetrate and migrate through the skin. Infrequently encountered ectoparasites in other phyla include the pentastomes (tongue worms) and leeches.

Arthropods may cause injury when they attempt to take a blood meal or as they defend themselves by biting, stinging, or exuding venoms. Papular urticaria and other lesions caused by arthropod bites and stings are so diverse and variable (depending upon the host's health status and prior exposure to the arthropod's saliva, venom, or other exudates) that it is difficult to identify the precise causative organism without a bona fide specimen and taxonomic expertise.

SCABIES

The human itch mite, *Sarcoptes scabiei* var. *hominis*, is an obligate human ectoparasite and a common cause of itchy dermatosis, infesting ~300 million persons worldwide. Gravid female mites (~0.3 mm in length) burrow superficially within the stratum corneum, depositing several eggs per day. Six-legged larvae mature to eight-legged nymphs and then to adults. Gravid adult females emerge to the surface of the skin about 8 days later and then (re)invade the skin of the same or another host. Newly fertilized female mites are transferred from person to person mainly by direct skin-to-skin contact; transfer is facilitated by crowding,

poor hygiene, and sex with multiple partners. Generally, scabies mites die within a day or so in the absence of a suitable host. Transmission via sharing of contaminated bedding or clothing occurs less frequently than is often thought. In the United States, scabies may account for up to 5% of visits to dermatologists. Outbreaks are known to occur in preschools, hospitals, nursing homes, and other institutional residences.

The itching and rash associated with scabies derive from a sensitization reaction to the mites and their secretions/excretions. A person's initial infestation remains asymptomatic for up to 6 weeks before the onset of intense pruritus, but a re-infestation produces a hypersensitivity reaction without delay. Burrows become surrounded by inflammatory infiltrates composed of eosinophils, lymphocytes, and histiocytes, and a generalized hypersensitivity rash later develops in remote sites. Immunity and associated scratching limit most infestations to <15 mites per person. Hyperinfestation with thousands of mites, a condition known as *crusted scabies* (formerly termed *Norwegian scabies*), may result from glucocorticoid use, immunodeficiency, and neurologic or psychiatric illnesses that limit the itch and/or the scratch response.

Pruritus typically intensifies at night and after hot showers. Classic burrows are often difficult to find because they are few in number and may be obscured by excoriations. Burrows appear as dark wavy lines in the upper epidermis and are 3–15 mm long. Scabetic lesions are most common on the volar wrists and along the digital web spaces. In males, the penis and scrotum become involved. Small papules and vesicles, often accompanied by eczematous plaques, pustules, or nodules, appear symmetrically at those sites and within intertriginous areas, around the navel and belt line, in the axillae, and on the buttocks and upper thighs. Except in infants, the face, scalp, neck, palms, and soles are usually spared. Crusted scabies often resembles psoriasis: both are characterized by widespread thick keratotic crusts, scaly plaques, and dystrophic nails. Characteristic burrows are not seen in crusted scabies, and patients usually do not itch, although their infestations are highly contagious and have been responsible for outbreaks of classic scabies in hospitals.

Scabies should be considered in patients with pruritus and symmetric superficial, excoriated, papulovesicular skin lesions in characteristic locations, particularly if there is a history of direct and prolonged contact with an infested person. Burrows should be sought and unroofed with a sterile needle or scalpel blade, and the scrapings should be examined microscopically for mites, eggs, and fecal pellets. Examination of skin biopsies (including superficial cyanoacrylate biopsy) or scrapings, dermatoscopic imaging of papulovesicular lesions, and microscopic inspection of clear cellophane tape lifted from lesions also may be diagnostic. In the absence of identifiable mites or eggs, the diagnosis is based on a history of pruritus, a clinical examination, and an epidemiologic link. Diverse kinds of dermatitis from other causes frequently are misdiagnosed as scabies, particularly in presumed "outbreak" situations. Scabies mites of other animals may cause transient irritation, but they do not reside or reproduce in human hosts. In some Aboriginal communities, household dogs may serve as reservoirs for human scabies mites.

TREATMENT

TREATMENT

Scabies

Permethrin cream (5%) is less toxic than 1% lindane preparations and is effective against lindane-tolerant infestations. Scabicides are applied thinly but thoroughly from the jawline down after bathing—with careful application to interdigital spaces and the umbilicus and under the fingernails—and are removed 8–14 h later with soap and water. Successful treatment of crusted scabies requires pre-application of a keratolytic agent such as 6% salicylic acid and then of scabicides to the scalp, face, and ears. Repeated treatments or the sequential use of several agents may be necessary. Ivermectin has not been approved by the U.S. Food and Drug Administration (FDA) for treatment of any form of scabies, but a single oral dose (200 µg/kg) is effective in otherwise healthy persons; patients with crusted scabies may require two doses separated by an interval of 1–2 weeks. All FDA-approved scabicides are available solely by prescription.

Within 1 day of effective treatment, scabies infestations become non-communicable, but the pruritic hypersensitivity dermatitis induced by the dead mites and their remnant products frequently persists for weeks. Unnecessary re-treatment with topical agents may provoke contact dermatitis. Antihistamines, salicylates, and calamine lotion relieve itching during treatment, and topical glucocorticoids are useful for pruritus that lingers after effective treatment. To prevent reinfestations, bedding and clothing should be washed and dried on high heat or heat-pressed. Close contacts of confirmed cases, even if asymptomatic, should be treated simultaneously.

CHIGGERS AND OTHER BITING MITES

Chiggers are the larvae of trombiculid (harvest) mites that normally feed on mice in grassy or brush-covered sites in tropical, subtropical, and (less frequently) temperate areas during warm months. They reside on low vegetation and attach themselves to passing vertebrate hosts. While feeding, larvae secrete saliva with proteolytic enzymes to create a tube-like invagination in the host's skin; this *stylostome* allows the mite to imbibe tissue fluids. The stylostomal saliva is highly antigenic and causes exceptionally pruritic papular, papulovesicular, or papulourticarial lesions (≤2 cm in diameter). In persons previously sensitized to salivary antigens, the papules develop within hours of attachment. While attached, mites appear as tiny red vesicles on the skin. Generally, lesions vesicate and develop a hemorrhagic base. Scratching invariably destroys the body of a mite, but itching and burning often persist for weeks. The rash is common on the ankles and areas where clothing obstructs the further wanderings of the mites. Repellents are useful for preventing chigger bites. Chiggers serve as vectors for *Orientia tsutsugamushi*, the agent of scrub typhus in Palearctic, Indomalayan, and Australasian regions.

Many kinds of mites associated with peridomestic birds and rodents are particularly bothersome when they invade homes and bite people. In North America, the northern fowl mite, chicken mite, tropical rat mite, and house mouse mite normally feed on poultry and diverse other birds as well as small mammals; these mites are abundant in and around their hosts' nests. After their natural hosts die or leave the nest, the mites disperse and may invade homes. Although the mites are rarely seen because of their small size, their bites can be painful and pruritic. House mouse mites (*Liponyssoides sanguineus*) serve as vectors for the agent of rickettsialpox, *Rickettsia akari*.

Once confirmed as the cause of irritation, rodent- and bird-associated mites are best eliminated by excluding their hosts, removing the nests, and cleaning and treating the nesting area with appropriate acaricides. *Pyemotes* and other mites that infest grain, straw, cheese, hay, oak leaf galls, or other products occasionally produce similar episodes of rash and discomfort and may produce a unique dermatologic “comet sign” lesion—a paisley-shaped urticarial plaque.

Diagnosis of mite-induced dermatitides (including those caused by chiggers) relies on confirmation of the mite’s identity or elicitation of a history of exposure to the mite’s source. Treatment of the patient with acaricides is not necessary, but oral antihistamines or topical steroids may suppress mite-induced pruritus temporarily.

TICK BITES AND TICK PARALYSIS

Ticks attach superficially to skin and feed painlessly; blood is their only food. Their salivary secretions are biologically active and can produce local reactions, induce fevers, and cause paralysis in addition to transmitting diverse pathogens. The two main families of ticks are the soft (argasid) and hard (ixodid) ticks. Generally, soft ticks attach for <1 h, leaving red macules after they drop off. Some species in Africa, the western United States, and Mexico produce painful hemorrhagic lesions. Hard ticks are much more common and transmit most of the tick-borne infections that are familiar to physicians and patients. Hard ticks attach to the host and feed for several days or sometimes for >1 week (depending upon the tick’s species and stage of development). At the site of hard-tick bites, small areas of induration, often purpuric, develop and may be surrounded by an erythematous rim. A necrotic eschar, called a *tâche noire* (“black spot”), occasionally develops. Chronic nodules (persistent tick-bite granulomas) can be several centimeters in diameter and may linger for months after the feeding tick has been removed. These granulomas can be treated with injected intralesional glucocorticoids or by surgical excision. Tick-induced fever, unassociated with transmission of any pathogen, is often accompanied by headache, nausea, and malaise, but usually resolves ≤36 h after the tick is removed. Salivary antigens of the lone star tick, *Amblyomma americanum*, may induce antibodies to galactose-α-1,3-galactose (alpha-gal) that result in mammalian meat allergy–alpha-gal syndrome.

Tick paralysis, an acute ascending flaccid paralysis that resembles Guillain-Barré syndrome, is believed to be caused by one or more toxins in tick saliva that block neuromuscular transmission and decrease nerve conduction. This rare complication has followed the bites of more than 60 kinds of ticks, although in the United States dog and wood ticks (*Dermacentor* species) are most commonly involved. Weakness begins symmetrically in the lower extremities ≤6 days after the tick’s attachment, ascends symmetrically during several days, and may culminate in complete paralysis of the extremities and cranial nerves. Deep tendon reflexes are diminished or absent, but sensory examination and findings on lumbar puncture are typically normal. Removal of the tick generally leads to rapid improvement within a few hours and complete recovery after several days, although the patient’s condition may continue to deteriorate for a full day. Failure to remove the tick may lead to dysarthria, dysphagia, and ultimately death from aspiration or respiratory paralysis. Diagnosis depends on finding the tick, which is often hidden beneath scalp hair. An antiserum to

the saliva of *Ixodes holocyclus*, the usual cause of tick paralysis in Australia, effectively reverses paralysis caused by these ticks.

Removal of hard ticks during the first 36 h of attachment generally prevents transmission of the agents of Lyme disease, babesiosis, anaplasmosis, and ehrlichiosis, although tick-borne viruses may be transmitted more quickly. Ticks should be removed by traction with fine-tipped forceps placed firmly around the tick's mouthparts. Careful handling (to avoid rupture of ticks) and use of gloves may avert accidental contamination with pathogens contained in tick fluids. Use of occlusive dressings, heat, or other substances (in an attempt to induce the tick to detach) merely delays tick removal. Afterward, the site of attachment should be disinfected. Tick mouthparts sometimes remain in the skin but generally are shed spontaneously within days without the need for surgical removal. Although somewhat controversial, current guidelines from the Centers for Disease Control and Prevention suggest that, rather than awaiting the onset of erythema migrans, the results of tick testing, or seroconversion to antigens diagnostic for Lyme disease, administration of prophylaxis with a single oral dose of doxycycline (200 mg) within 72 h of tick removal is appropriate in adult patients with bites thought to be associated with *Ixodes scapularis* (deer ticks) **(Fig. 452-1)** in Lyme disease–endemic areas.

FIGURE 452-1

Deer ticks (*Ixodes scapularis*, black-legged ticks) on a U.S. penny: larva (*below ear*), nymph (*right*), adult male (*above*), and adult female (*left*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

LOUSE INFESTATION (PEDICULIASIS AND PTHIRIASIS)

Nymphs and adults of all three kinds of human lice feed at least once a day, ingesting human blood exclusively. Head lice (*Pediculus capitis*) infest mainly the hair of the scalp, body lice (*Pediculus humanus*) the clothing, and crab or pubic lice (*Pthirus pubis*) mainly the hair of the pubis. The saliva of lice produces a pruritic morbilliform or urticarial rash in some sensitized persons. Female head and pubic lice cement their eggs (nits) firmly to hair, whereas female body lice cement their eggs to clothing, particularly to threads along clothing seams. After ~10 days of development within the egg, a nymph hatches. Empty eggs may remain affixed for months or years thereafter.

In North America, the prevalence of head lice is ~1% among children 6–10 years old and considerably lower among persons of other ages. Infestations can be far more prevalent elsewhere. Head lice are transmitted mainly by direct head-to-head contact rather than by fomites such as shared headgear, bed linens, and grooming implements. Chronic infestations by head lice tend to be asymptomatic. Pruritus, due mainly to hypersensitivity to the louse's saliva, generally is transient and mild and is most evident around the posterior hairline. Head lice removed from a person succumb to desiccation and starvation within ~1 day. Head lice are not known to serve as a natural vector for any pathogens.

Body lice remain on clothing except when feeding and generally succumb in ≤2 days if separated from their host. In most Western countries, body lice are generally found on a small proportion of indigent persons but may become increasingly prevalent after societal upheaval from natural or human-caused disasters, when displaced persons are in close contact with infested individuals with whom they share accommodations and lack the wherewithal to wash or change their clothes. Body lice are acquired by direct contact or by sharing of infested clothing and bedding. These lice are vectors for the agents of louse-borne (epidemic) typhus (Chap. 182), louse-borne relapsing fever (Chap. 180), and trench fever (Chap. 167). Chronic infestations result in a postinflammatory hyperpigmentation and thickening of the skin known as *vagabond's disease*.

The crab or pubic louse is transmitted mainly by sexual contact. These lice occur predominantly on pubic hair and less frequently on axillary or facial hair, including the eyelashes. Children and adults may acquire pubic lice by sexual or close nonsexual contact. Intensely pruritic, bluish macules ~3 mm in diameter (*maculae ceruleae*) develop at the site of bites. Blepharitis commonly accompanies infestations of the eyelashes.

Pediculiasis is often suspected upon the detection of nits firmly cemented to hairs or in clothing. Many bona fide nits, however, are dead or hatched relics of prior infestation, and pseudo-nits are frequently misconstrued to be signs of a louse infestation. Confirmation of a louse infestation, therefore, best relies on the discovery of a live louse.

TREATMENT

TREATMENT

Louse Infestation

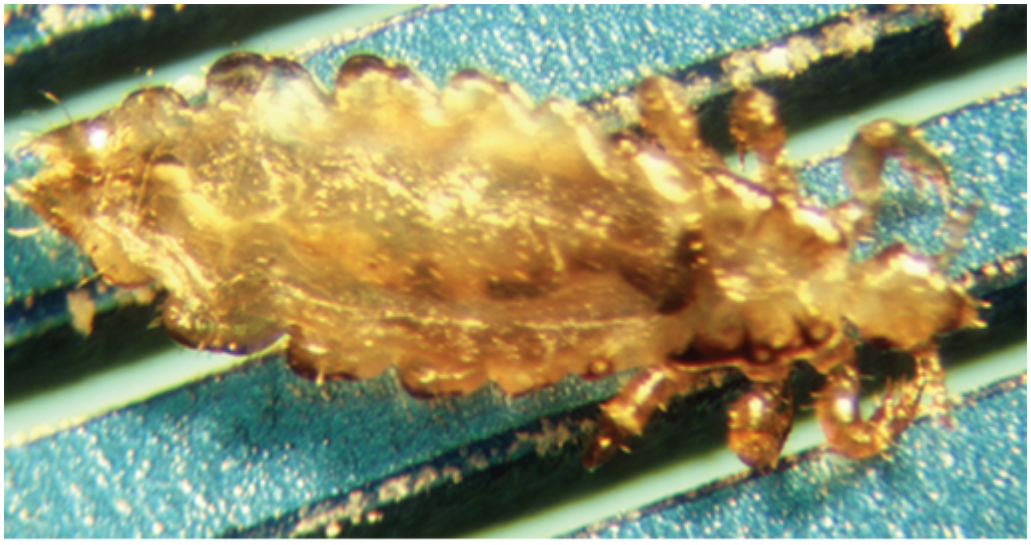
Generally, treatment is justified only if live lice are discovered. The presence of nits alone is evidence of a former—not necessarily current—infestation. Mechanical removal of head lice and their eggs with a fine-toothed louse or nit comb **(Fig. 452-2)** often fails to eliminate infestations. Treatment of newly identified active infestations traditionally relies on a 10-min topical application of ~1% permethrin or pyrethrins, with a second application ~10 days later. Lice persisting after this treatment may be resistant to pyrethroids. Chronic infestations may be treated for ≤12 h with 0.5% malathion. Lindane is applied for just 4 min but seems less effective and may pose a greater risk of adverse reactions, particularly when misused. Resistance of head lice to permethrin, malathion, and lindane has been reported. Newer FDA-approved topical pediculicides contain benzyl alcohol, dimethicone, spinosad, and ivermectin. Although children infested by head lice—or those who simply have remnant nits from a prior infestation—are frequently isolated or excluded from school, this practice increasingly is considered to be unjustified, ineffective, and counterproductive.

Body lice usually are eliminated by bathing and by changing to laundered clothes. Application of topical pediculicides from head to foot may be necessary for hirsute patients. Clothes and bedding are effectively deloused by heating in a clothes dryer at ≥55°C (≥131°F) for 30 min or by heat-pressing. Emergency mass delousing of persons and clothing may be warranted during periods of civil strife and after natural disasters to reduce the risk of pathogen transmission by body lice.

Pubic louse infestations are treated with topical pediculicides, except for eyelid infestations (*pthiriasis palpebrum*), which generally respond to a coating of petrolatum applied for 3–4 days.

FIGURE 452-2

Adult female human head louse (*Pediculus capitis*) on a nit (louse-egg) comb.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

MYIASIS (FLY INFESTATION)

Myiasis refers to infestations by fly larvae (maggots) that invade living or necrotic tissues or body cavities and produce different clinical syndromes, depending on the species of fly.

In forested parts of Central and South America, larvae of the human botfly (*Dermatobia hominis*) produce furuncular (boil-like) papules or subcutaneous nodules ≤3 cm in diameter. A gravid adult female botfly captures a mosquito or another bloodsucking insect and deposits her eggs on its abdomen. When the carrier insect attacks a human or bovine host several days later, the warmth and moisture of the host's skin stimulate the eggs to hatch. The emerging larvae promptly penetrate intact skin. After 6–12 weeks of development, mature larvae emerge from the skin and drop to the ground to pupate and then become adults.

The African tumbu fly (*Cordylobia anthropophaga*) deposits its eggs on damp sand or leaf litter or on drying laundry, particularly items contaminated by urine or sweat. Larvae hatch from eggs upon contact with a host's body and penetrate the skin, producing boil-like lesions from which mature larvae emerge ~9 days later. Furuncular myiasis is suggested by uncomfortable lesions with a central breathing pore that emit bubbles when submerged in water. A sensation of movement under the patient's skin may cause severe emotional distress.

Larvae that cause furuncular myiasis may be induced to emerge if the air pore is coated with petrolatum or another occlusive substance. Removal may be facilitated by injection of a local anesthetic into the surrounding tissue, but surgical excision is sometimes necessary because upward-pointing spines of some species hold the larvae firmly in place.

Other fly larvae cause nonfuruncular myiasis. Larvae of the horse botfly (*Gasterophilus intestinalis*) emerge from eggs deposited on the horse's flanks and may come into contact with and infest human beings. After penetrating human skin, these larvae rarely mature but instead may migrate for weeks in the dermis. The resulting pruritic and serpiginous eruption resembles cutaneous larva migrans caused by canine or feline hookworms (Chap. 226). Larvae of rabbit and rodent botflies (*Cuterebra* species) occasionally cause dermal or tracheopulmonary myiasis.

Certain flies are attracted to blood and pus, laying their eggs on open or draining sores. Newly hatched larvae enter wounds or diseased skin. Larvae of several types of green bottle flies (*Lucilia*/*Phaenicia* species) usually remain superficial and confined to necrotic tissue. Specially raised, sterile "surgical maggots" are sometimes used intentionally for wound debridement. Larvae of screwworm flies (*Cochliomyia*) and the flesh fly invade viable tissues more deeply and produce large suppurating lesions. Larvae that infest wounds also may enter body cavities such as the mouth, nose, ears, sinuses, anus, vagina, and lower urinary tract, particularly in unconscious or otherwise debilitated patients. The consequences range from harmless colonization to destruction of the nose, meningitis, and deafness. Treatment involves removal of maggots and debridement of tissue.

Larvae of the sheep botfly, *Oestrus ovis*, and others responsible for furuncular and wound myiasis also may cause ophthalmomyiasis. Sequelae include nodules in the eyelid, retinal detachment, and destruction of the

globe. Most instances in which maggots are found in human feces result from deposition of eggs or larvae by flies on recently passed stools, not from an intestinal maggot infestation.

PENTASTOMIASIS

Pentastomids (tongue worms) inhabit the respiratory passages of reptiles and carnivorous mammals. Human infestation by *Linguatula serrata* is common in the Middle East and results from the consumption of encysted larval stages in raw liver or lymph nodes of sheep and goats, which are true intermediate hosts for the tongue worms. Larvae migrate to the nasopharynx and produce an acute self-limiting syndrome—known as *halzoun* or *marrara*—characterized by pain and itching of the throat and ears, coughing, hoarseness, dysphagia, and dyspnea. Severe edema may cause obstruction that requires tracheostomy. In addition, ocular invasion has been described. Diagnostic larvae measuring ≤10 mm in length appear in copious nasal discharge or vomitus. Individuals become infected with another type of tongue worm, *Armillifer armillatus*, by consuming its eggs in contaminated food or drink or after handling the definitive host, the African python. Larvae encyst in various organs but rarely cause symptoms. Cysts may require surgical removal as they enlarge during molting, but they usually are encountered as an incidental finding at autopsy. Parasite-induced lesions may be misinterpreted as a malignancy, with the correct diagnosis confirmed histopathologically. Cutaneous larva migrans–type syndromes of other pentastomes have been reported from Southeast Asia and Central America.

LEECH INFESTATIONS

Medically important leeches are annelid worms that attach to their hosts with chitinous cutting jaws and draw blood through muscular suckers. The medicinal leech (*Hirudo medicinalis*) is still used occasionally for medical purposes to reduce venous congestion in surgical flaps or replanted body parts. This practice has been complicated by intractable bleeding, wound infections, myonecrosis, and sepsis due to *Aeromonas hydrophila*, which colonizes the gullets of commercially available leeches.

Ubiquitous aquatic leeches that parasitize fish, frogs, and turtles readily attach to the skin of human beings and avidly suck blood. More notorious are arboreal land leeches that live among moist vegetation of tropical rain forests. Attachment is usually painless, and the leeches will detach themselves when satiated with a blood meal. Hirudin, a powerful anticoagulant secreted by the leech, causes continued bleeding after the leech has detached. Healing of a leech-bite wound is slow, and bacterial infections are not uncommon. Several kinds of aquatic leeches in Africa, Asia, and southern Europe can enter the mouth, nose, and genitourinary tract and attach to mucosal surfaces at sites as deep as the esophagus and trachea. Externally attached leeches generally drop off after they have engorged, but removal is hastened by gentle scraping aside of the anterior and posterior suckers the leech uses for attachment and feeding. Some authorities dispute the wisdom of removing leeches with alcohol, salt, vinegar, insect repellent, a flame or heated instrument, or applications of other noxious substances. Internally attached leeches may detach on exposure to gargled saline or may be removed by forceps.

SPIDER BITES

Of the more than 30,000 recognized species of spiders, only ~100 defend themselves aggressively and have fangs sufficiently long to penetrate human skin. The venom that some spiders use to immobilize and digest their prey can cause necrosis of skin and systemic toxicity. Whereas the bites of most spiders are painful but not harmful, envenomations by recluse or fiddleback spiders (*Loxosceles* species) and widow spiders (*Latrodectus* species) may be life-threatening. Identification of the offending spider is important because specific treatments exist for bites of widow spiders and because injuries attributed to spiders are frequently due to other causes. Except in cases where the patient actually observes a spider immediately associated with the bite or fleeing from the site, lesions reported as spider-bite reactions are most often due to other injuries or to infections with bacteria such as methicillin-resistant *Staphylococcus aureus* (MRSA) or, infrequently, *Franciscella tularensis.*

Recluse Spider Bites and Necrotic Arachnidism

Brown recluse spiders (*Loxosceles reclusa*) live mainly in the south-central United States and have close relatives in Central and South America, Africa, and the Middle East. Bites by brown recluse spiders usually cause only minor injuries, with edema and erythema. Envenomation, however, occasionally causes severe necrosis of skin and subcutaneous tissue and more rarely causes systemic hemolysis. These spiders are not aggressive toward humans and bite only if threatened or pressed against the skin. They hide under rocks and logs or in caves and animal burrows. They invade homes and seek dark and undisturbed hiding spots in closets, in folds of clothing, or under furniture and rubbish in storage rooms, garages, and attics. Despite their impressive abundance in some homes, these spiders rarely bite humans. Bites tend to occur while the victim is dressing and are sustained primarily on the hands, arms, neck, and lower abdomen.

Initially, the bite is painless or may produce a stinging sensation. Within the next few hours, the site becomes painful and pruritic, with central induration surrounded by a pale ischemic zone that itself is encircled by a zone of erythema. In most cases, the lesion resolves without treatment in just a few days. In severe cases, the erythema spreads, and the center of the lesion becomes hemorrhagic or necrotic with an overlying bulla. A black eschar forms and sloughs several weeks later, leaving an ulcer that eventually may create a depressed scar. Healing usually takes place in ≤3 months. Local complications include injury to nerves and secondary bacterial infection. Fever, chills, weakness, headache, nausea, vomiting, myalgia, arthralgia, maculopapular rash, and leukocytosis may develop ≤72 h after the bite. Reports of deaths attributed to bites of North American brown recluse spiders have not been verified.

The Mediterranean recluse spider (*Loxosceles rufescens*) is a widely invasive species in urban areas of both the Old and New Worlds. The dorsal surface of *L. rufescens* and *L. reclusa* is adorned with a fiddle-shaped pattern. *L. rufescens* is warier than *L. reclusa*, is less likely to bite, and rarely causes necrosis. Misidentification of this spider may create spurious reports of *L. reclusa* activity outside the known range of that species.

TREATMENT

TREATMENT

Recluse Spider Bites

Initial management includes rest, ice, compression, and elevation (RICE). Analgesics, antihistamines, antibiotics, and tetanus prophylaxis should be administered if indicated. Early debridement or surgical excision of the wound without closure delays healing. Routine use of antibiotics or dapsone lacks utility. Patients should be monitored closely for signs of hemolysis, renal failure, and other systemic complications.

Widow Spider Bites

The black widow spider, common in the southeastern United States, measures ≤1 cm in body length and 5 cm in leg span and is shiny black with a red hourglass marking on the ventral abdomen. Other dangerous *Latrodectus* species occur elsewhere in temperate and subtropical parts of the world. The bites of the female widow spiders are notorious for their potent neurotoxins.

Widow spiders spin their webs under stones, logs, plants, or rock piles and in dark spaces in barns, garages, and outhouses. Bites are most common in the summer and early autumn and occur when a web is disturbed or a spider is trapped or provoked. The initial bite is perceived as a sharp pinprick or may go unnoticed. Fang-puncture marks are uncommon. The venom that is injected does not produce local necrosis, and some persons experience no other symptoms. α-Latrotoxin, the most active component of the venom, binds irreversibly to presynaptic nerve terminals and causes release and eventual depletion of acetylcholine, norepinephrine, and other neurotransmitters from those terminals. Painful cramps may spread within 60 min from the bite site to large muscles of the extremities and trunk. Extreme rigidity of the abdominal muscles and excruciating pain may suggest peritonitis, but the abdomen is not tender on palpation and surgery is not warranted. The pain begins to subside during the first 12 h but may recur during several days or weeks before resolving spontaneously. A wide range of other sequelae may include salivation, diaphoresis, vomiting, hypertension, tachycardia, labored breathing, anxiety, headache, weakness, fasciculations, paresthesia, hyperreflexia, urinary retention, uterine contractions, and premature labor. Rhabdomyolysis and renal failure have been reported, and respiratory arrest, cerebral hemorrhage, or cardiac failure may end fatally, especially in very young, elderly, or debilitated persons.

TREATMENT

TREATMENT

Widow Spider Bites

Treatment consists of RICE and tetanus prophylaxis. Hypertension that does not respond to analgesics and antispasmodics (e.g., benzodiazepines or methocarbamol) requires specific antihypertensive medication. The efficacy and safety of antivenoms for black widow and redback spiders are controversial because of concerns about potential anaphylaxis or serum sickness.

Tarantulas and Other Spiders

Tarantulas are hairy spiders of which 30 species are found in the United States, mainly in the Southwest. The tarantulas that have become popular household pets are usually imported from Central or South America. Tarantulas bite only when threatened and usually cause no more harm than a bee sting, but on occasion the venom causes deep pain and swelling. Several species of tarantulas are covered with urticating hairs that are brushed off in the thousands when a threatened spider rubs its hind legs across its dorsal abdomen. These hairs can penetrate human skin and produce pruritic papules that may persist for weeks. Failure to wear gloves or to wash the hands after handling the Chilean Rose tarantula, a popular pet spider, has resulted in transfer of hairs to the eye with subsequent devastating ocular inflammation. Treatment of bites includes local washing and elevation of the bitten area, tetanus prophylaxis, and analgesic administration. Antihistamines and topical or systemic glucocorticoids are given for exposure to urticating hairs.

Atrax robustus, a funnel-web spider of Australia, and *Phoneutria* species, the South American banana spiders, are among the most dangerous spiders in the world because of their aggressive behavior and potent neurotoxins. Envenomation by *A. robustus* causes a rapidly progressive neuromotor syndrome that can be fatal within 2 h. The bite of a banana spider causes severe local pain followed by profound systemic symptoms and respiratory paralysis that can lead to death within 2–6 h. Specific antivenoms for use after bites by each of these spiders are available. Yellow sac spiders (*Cheiracanthium* species) are common in homes worldwide. Their bites, though painful, generally lead to only minor erythema, edema, and pruritus.

SCORPION STINGS

Scorpions are arachnids that feed on ground-dwelling arthropods and small lizards. They paralyze their prey and defend themselves by injecting venom from a stinger on the tip of the tail. Painful but relatively harmless scorpion stings need to be distinguished from the potentially lethal envenomations that are produced by ~30 of the ~1000 known species and that cause more than 5000 deaths worldwide each year. Scorpions are nocturnal and remain hidden during the day in crevices or burrows or under wood, loose bark, or rocks. They occasionally enter houses and tents and may hide in shoes, clothing, or bedding. Scorpions sting humans only when threatened.

Of the 40 or so scorpion species in the United States, only bark scorpions (*Centruroides sculpturatus*/*C. exilicauda*) in the Southwest produce venom that is potentially lethal to humans. This venom contains neurotoxins that cause sodium channels to remain open. Such envenomations usually are associated with little swelling, but prominent pain, paresthesia, and hyperesthesia can be accentuated by tapping on the affected area (the *tap test*). These symptoms soon spread to other locations; dysfunction of cranial nerves and hyperexcitability of skeletal muscles develop within hours. Patients present with restlessness, blurred vision, abnormal eye movements, profuse salivation, lacrimation, rhinorrhea, slurred speech, difficulty in handling secretions, diaphoresis, nausea, and vomiting. Muscle twitching, jerking, and shaking may be mistaken for a seizure. Complications include tachycardia, arrhythmias, hypertension, hyperthermia, rhabdomyolysis, and acidosis. Symptoms progress to maximal severity in ~5 h and subside within a day or two, although pain and paresthesia can last for weeks. Fatal respiratory arrest is most common among young children and the elderly.

Envenomations by *Leiurus quinquestriatus* in the Middle East and North Africa, by *Mesobuthus tamulus* in India, by *Androctonus* species along the Mediterranean littoral and in North Africa and the Middle East, and by *Tityus serrulatus* in Brazil cause massive release of endogenous catecholamines with hypertensive crises, arrhythmias, pulmonary edema, and myocardial damage. Acute pancreatitis occurs with stings of *Tityus trinitatis* in Trinidad, and central nervous toxicity complicates stings of *Parabuthus* and *Buthotus* scorpions of South Africa. Tissue necrosis and hemolysis may follow stings of the Iranian *Hemiscorpius lepturus*.

Stings of most other species cause immediate sharp local pain followed by edema, ecchymosis, and a burning sensation. Symptoms typically resolve within a few hours, and skin does not slough. Allergic reactions to the venom sometimes develop.

TREATMENT

TREATMENT

Scorpion Stings

Identification of the offending scorpion helps to determine the course of treatment. Stings of nonlethal species require at most ice packs, analgesics, or antihistamines. Because most victims experience only local discomfort, they can be managed at home with instructions to return to the emergency department if signs of cranial-nerve or neuromuscular dysfunction develop. Aggressive supportive care and judicious use of antivenom can reduce or eliminate deaths from more severe envenomations. Keeping the patient calm and applying pressure dressings and cold packs to the sting site are measures that decrease the absorption of venom. A continuous IV infusion of midazolam controls the agitation, flailing, and involuntary muscle movements produced by scorpion stings. Close monitoring during treatment with this drug and other sedatives or narcotics is necessary for persons with neuromuscular symptoms because of the risk of respiratory arrest. Hypertension and pulmonary edema respond to nifedipine, nitroprusside, hydralazine, or prazosin. Dangerous bradydysrhythmia can be controlled with atropine.

Commercially prepared antivenoms are available in several countries for some of the most dangerous species. An FDA-approved *C. sculpturatus* antivenom in horse serum is now available. IV administration of antivenom rapidly reverses cranial-nerve dysfunction and muscular symptoms. Although effective, cost analyses suggest that antivenoms should be reserved for only the most severe envenomations.

HYMENOPTERA STINGS

Bees, wasps, hornets, yellow jackets, and ants (all of the insect order Hymenoptera) sting in defense or to subdue their prey. Their venoms contain a wide array of amines, peptides, and enzymes that cause local and systemic reactions. Although the toxic effect of multiple stings can be fatal to a human, nearly all of the ≥100 deaths due to hymenopteran stings in the United States each year result from allergic reactions.

Bee and Wasp Stings

The stinger of the honeybee (*Apis mellifera*) is unique in being barbed. The stinging apparatus and attached venom sac tear loose from the honeybee's body, and muscular contraction of the venom sac continues to inject venom into the skin. Other kinds of bees, ants, and wasps have smooth stinging mechanisms and can sting numerous times in succession. Generally, a person sustains just one sting from a bee or social wasp unless a nest was disturbed. Africanized honeybees (now present in South and Central America and the southern and western United States) respond to minimal intrusions more aggressively. The sting of an Africanized bee contains less venom than that of its non-Africanized relatives, but victims tend to sustain far more stings and receive a far greater overall volume of venom. Most patients who report having sustained a "bee sting" are more likely to have encountered stinging wasps instead.

The venoms of different kinds of hymenopterans are biochemically and immunologically distinct. Direct toxic effects are mediated by mixtures of low-molecular-weight compounds such as serotonin, histamine, acetylcholine, and several kinins. Polypeptide toxins in honeybee venom include mellitin, which damages cell membranes; mast cell–degranulating protein, which causes histamine release; the neurotoxin apamin; and the anti-inflammatory compound adolapin. Enzymes in venom include hyaluronidase and phospholipases. There appears to be little cross-sensitization between the venoms of honeybees and wasps.

Uncomplicated hymenopteran stings cause immediate pain, a wheal-and-flare reaction, and local edema, all which usually subside in a few hours. Multiple stings can lead to vomiting, diarrhea, generalized edema, dyspnea, hypotension, and non-anaphylactic circulatory collapse. Rhabdomyolysis and intravascular hemolysis may cause renal failure. Death from the direct (nonallergic) effects of venom has followed stings of several hundred honeybees. Stings to the tongue or mouth may induce life-threatening edema of the upper airways.

Large local reactions accompanied by erythema, edema, warmth, and tenderness that spread ≥10 cm around the sting site over 1–2 days are not uncommon. These reactions may resemble bacterial cellulitis but are caused by hypersensitivity rather than by secondary infection. Such reactions tend to recur on subsequent exposure but are seldom accompanied by anaphylaxis and are not prevented by venom immunotherapy.

An estimated 0.4–4.0% of the U.S. population exhibits clinical immediate-type hypersensitivity to hymenopteran stings, and 15% may have asymptomatic sensitization manifested by positive skin tests. Persons who experience severe allergic reactions are likely to have similar or more severe reactions after subsequent stings by the same or closely related species. Mild anaphylactic reactions from insect stings, as from other causes, consist of nausea, abdominal cramping, generalized urticaria or angioedema, and flushing. Serious reactions, including upper airway edema, bronchospasm, hypotension, and shock, may be rapidly fatal. Severe reactions usually begin within 10 min of the sting and only rarely develop after 5 h.

TREATMENT

TREATMENT

Bee and Wasp Stings

Honeybee stingers embedded in the skin should be removed as soon as possible to limit the quantity of venom delivered. The stinger and venom sac may be scraped off with a blade, a fingernail, or the edge of a credit card or may be removed with forceps. The site should be cleansed and disinfected and ice packs applied to slow the spread of venom. Elevation of the affected site and administration of oral analgesics, oral antihistamines, and topical calamine lotion help relieve symptoms.

Anaphylactic reactions to bee or wasp venom can be a life-threatening emergency that requires prompt life-saving actions. If the individual carries a bee-sting kit, then a subcutaneous injection of epinephrine hydrochloride (0.3 mL of a 1:1000 dilution) should be considered, with treatment repeated every 20–30 min as necessary. A tourniquet may slow the spread of venom. The patient should be transferred to a hospital emergency room where treatment for profound shock, if required, can be administered safely. Such treatment may entail the use of intravenous epinephrine and other vasopressors, intubation or supplemental oxygen, fluid resuscitation, bronchodilators, and parenteral antihistamines. Patients should be observed for 24 h for recurrent anaphylaxis, renal failure, or coagulopathy.

Persons with a history of allergy to insect stings should carry an anaphylaxis kit with a preloaded syringe containing epinephrine for self-administration. These patients should seek medical attention immediately after using the kit.

Prophylactic immunotherapy may greatly reduce the risk of life-threatening reactions to bee and wasp stings. Repeated injections of purified venom produce a blocking IgG antibody response to venom and reduce the incidence of recurrent anaphylaxis. Honeybee, wasp, and yellow jacket venoms are commercially available for desensitization and for skin testing. Results of skin tests and venom-specific radioallergosorbent tests (RASTs) aid in the selection of patients for immunotherapy and guide the design of such treatment.

STINGING ANTS

Stinging ants are an important medical problem in the United States. Imported fire ants (*Solenopsis* species) infest southern states from Texas to North Carolina, with colonies now established in California, New Mexico, Arizona, and Virginia. Slight disturbances of their mound nests have provoked massive outpourings of ants and as many as 10,000 stings on a single person. Elderly and immobile persons are at high risk for attacks when fire ants invade dwellings.

Fire ants attach to skin with powerful mandibles and rotate their bodies while repeatedly injecting venom with posteriorly situated stingers. The alkaloid venom consists of cytotoxic and hemolytic piperidines and several proteins with enzymatic activity. The initial wheal-and-flare reaction, burning, and itching resolve in ~30 min, and a sterile pustule develops within 24 h. The pustule ulcerates over the next 48 h and then heals in ≥1 week. Large areas of erythema and edema lasting several days are not uncommon and in extreme cases may compress nerves and blood vessels. Anaphylaxis occurs in fewer than 2% of victims; seizures and mononeuritis have been reported. Stings are treated with ice packs, topical glucocorticoids, and oral antihistamines. Pustules should be cleansed and then covered with bandages and antibiotic ointment to prevent bacterial infection. Epinephrine administration and supportive measures are indicated for

anaphylactic reactions. Fire ant whole-body extracts are available for skin testing and immunotherapy, which appear to lower the rate of anaphylactic reactions.

European fire (red) ants (*Myrmica rubra*) have recently become public health pests in the northeastern United States and southern Canada. The western United States is home to harvester ants (*Pogonomyrmex* species). The painful local reaction that follows harvester ant stings often extends to lymph nodes and may be accompanied by anaphylaxis. The bullet or conga ant (*Paraponera clavata*) of South America is known locally as *hormiga veinticuatro* ("24-hour ant"), a designation that refers to the 24 h of throbbing, excruciating pain following a sting that delivers the potent paralyzing neurotoxin poneratoxin.

DIPTERAN (FLY AND MOSQUITO) BITES

In the process of feeding on vertebrate blood and tissue fluids, adults of certain flies inflict painful bites, produce local allergic reactions, and may transmit pathogenic agents. Bites of mosquitoes, tiny "no-see-um" (ceratopogonid) midges, and phlebotomine sand flies typically produce a wheal and a pruritic papule. Small humpbacked black flies (simuliids) lacerate skin, resulting in a lesion with serosanguineous discharge that is often painful and pruritic. Regional lymphadenopathy, fever, or anaphylaxis occasionally ensues. The widely distributed deerflies and horseflies as well as the tsetse flies of Africa are stout flies measuring ≤25 mm in length that attack during the day and produce large and painful bleeding punctures. Houseflies (*Musca domestica*) do not consume blood but use rasping mouthparts to scarify skin and feed upon tissue fluids and salt. Beyond direct injury from bites of any kind of fly, risks include transmission of diverse pathogens and secondary infection of the lesion.

TREATMENT

TREATMENT

Fly and Mosquito Bites

Treatment of fly bites is symptom based. Topical application of antipruritic agents, glucocorticoids, or antiseptic lotions may relieve itching and pain. Allergic reactions may require oral antihistamines. Antibiotics may be necessary for the treatment of large bite wounds that become secondarily infected.

FLEA BITES

Common human-biting fleas include the dog and cat fleas (*Ctenocephalides* species) and the rat flea (*Xenopsylla cheopis*), which infest their respective hosts and their nests and resting sites. Sensitized persons develop erythematous pruritic papules (*papular urticaria*) and occasionally vesicles and bacterial superinfection at the site of the bite. Symptom-based treatment consists of antihistamines, topical glucocorticoids, and topical antipruritic agents.

Flea infestations are eliminated by removal and treatment of animal nests, frequent cleaning of pet bedding, and application of contact and systemic insecticides to pets and the dwelling. Flea infestations in the home may be abated or prevented if pets are regularly treated with veterinary antiparasitic agents, insect growth regulators, or chitin inhibitors.

Tunga penetrans, like other fleas, is a wingless, laterally flattened insect that feeds on blood. Also known as the chigoe flea, sand flea, or jigger (not to be confused with the chigger), it occurs in tropical regions of Africa and the Americas. Adult female chigoes live in sandy soil and burrow under the skin, usually between toes, under nails, or on the soles of bare feet. Gravid chigoes engorge on the host's blood and grow from pinpoint to pea size during a 2-week interval. They produce lesions that resemble a white pustule with a central black depression and that may be pruritic or painful. Occasional complications include tetanus, bacterial infections, and autoamputation of toes (*ainhum*). Tungiasis is treated by removal of the intact flea with a sterile needle or scalpel, tetanus vaccination, and topical application of antibiotics.

HEMIPTERAN / HETEROPTERAN (TRUE BUG) BITES

Most true bugs feed on plants, but some are predaceous or feed on blood. They may inflict bites in order to feed or in defense that produce allergic reactions and are sometimes painful. Bites of the cone-nose or "kissing bugs" (family Reduviidae) tend to occur at night and are painless. Reactions to such bites depend on prior sensitization and include tender and pruritic papules, vesicular or bullous lesions, extensive urticaria, fever, lymphadenopathy, and (rarely) anaphylaxis. Bug bites are treated with topical antipruritics or oral antihistamines. Persons with anaphylactic reactions to reduviid bites should keep an epinephrine kit available. Some reduviids transmit *Trypanosoma cruzi*, the agent of New World trypanosomiasis (Chagas disease) **(Chap. 222)**.

The cosmopolitan and tropical bedbugs (*Cimex lectularius and C. hemipterus*) hide in crevices of mattresses, bed frames and other furniture, walls, and picture frames and under loose wallpaper, actively seeking blood meals at night. Bedbug populations have resurged, recently attaining levels and spreading to an extent not encountered since the mid-twentieth century. These bugs are now a common pest in homes, dormitories, and hotels; on cruise ships; and even in medical facilities. Their bite is painless, but minutes to days later, sensitized persons develop erythema, itching, and wheals around a central hemorrhagic punctum. Bedbugs are not known to transmit pathogens in nature.

CENTIPEDE BITES AND MILLIPEDE DERMATITIS

The fangs of centipedes of the genus *Scolopendra* can penetrate human skin and deliver a venom that produces intense burning pain, swelling, erythema, and sterile lymphangitis. Dizziness, nausea, and anxiety are described occasionally, and rhabdomyolysis and renal failure have been reported. Treatment includes washing of the site, application of cold dressings, oral analgesic administration or local lidocaine infiltration, and tetanus prophylaxis.

Millipedes are docile and do not bite, but some secrete defensive fluids that may burn and discolor human skin. Affected skin turns brown overnight and may blister and exfoliate. Secretions in the eye cause intense pain and inflammation that can result in corneal ulcers and even blindness. Management includes irrigation with copious amounts of water or saline, use of analgesics, and local care of denuded skin.

CATERPILLAR STINGS AND DERMATITIS

Caterpillars of several moth species are covered with hairs or spines that produce mechanical irritation and may contain or be coated with venom. Contact with these caterpillars or their hairs may lead to erucism (a pruritic urticarial or papular rash) or caterpillar envenomation. The response typically consists of an immediate burning sensation followed by local swelling and erythema and occasionally by regional lymphadenopathy, nausea, vomiting, and headache. A rare reaction to a South American caterpillar, *Lonomia obliqua*, can cause disseminated coagulopathy and fatal hemorrhagic shock. In the United States, dermatitis is most often associated with caterpillars of the io, puss, saddleback, and brown-tail moths. Even contact with detached hairs of other caterpillars, such as gypsy moth larvae, can later produce erucism. Spines may be deposited on tree trunks or drying laundry or may be airborne and cause irritation of the eyes and upper airways. Treatment of caterpillar stings consists of repeated application of adhesive or cellophane tape to remove the hairs, which can then be identified microscopically. Local ice packs, topical glucocorticoids, and oral antihistamines relieve symptoms.

BEETLE VESICATION AND DERMITITIS

Several families of beetles have independently developed the ability to produce chemically unrelated vesicating toxins. When disturbed, blister beetles (family Meloidae) exude cantharidin, a low-molecular-weight toxin that produces thin-walled blisters (≤5 cm in diameter) 2–5 h after contact. The blisters are not painful or pruritic unless broken and resolve without treatment in ≤10 days. Nephritis may follow unusually heavy cantharidin exposure. The hemolymph of certain rove beetles (Staphylinidae) contains pederin, a potent vesicant. When these beetles are crushed or brushed against the skin, the released fluid causes painful, red, flaccid bullae. These beetles occur worldwide but are most numerous and problematic in parts of Africa (where they are called “Nairobi fly”) and southwestern Asia. Ocular lesions may develop after impact with flying beetles at night or unintentional transfer of the vesicant on the fingers. Treatment is rarely necessary, although ruptured blisters should be kept clean and bandaged.

Larvae of common carpet beetles are adorned with dense arrays of ornate hairs called *hastisetae*. Contact with these larvae or their setae results in delayed dermal reactions in sensitized individuals. The lesions are commonly mistaken as bites of bedbugs.

DELUSIONAL INFESTATIONS

The groundless conviction that one is infested with arthropods or other parasites (Ekbom’s syndrome, delusory parasitosis, delusions of parasitosis, and perhaps Morgellons syndrome) is extremely difficult to

treat and, unfortunately, is not uncommon. Patients describe uncomfortable sensations of something moving in or on their skin. Excoriations and self-induced ulcerations typically accompany the pruritus, dysesthesias, and imaginary insect bites. Patients often believe that some invisible or as yet undescribed creatures are infesting their skin, clothing, homes, or environment in general. Frequently, patients submit as evidence of infestation specimens that consist of plant-feeding and nonbiting peridomestic arthropods, pieces of skin, vegetable matter, lint, and other inanimate detritus. When evaluating a patient with possible delusional parasitosis, it is imperative to rule out true infestations and bites by arthropods, endocrinopathies, sensory disorders due to neuropathies, opiate and other drug use, environmental irritants (e.g., fiberglass threads), and other causes of tingling or prickling sensations. Frequently, such patients repeatedly seek medical consultations, resist alternative explanations for their symptoms, and exacerbate their discomfort by self-treatment. Long-term pharmacotherapy with pimozide or other psychotropic agents has been more helpful than psychotherapy in treating this disorder. Patients with delusory parasitosis often develop the unshakeable conviction that they are infested by a previously unknown pathogen, while their personal lives, family support, and employment collapse around them.

FURTHER READING

Arlia LG, Morgan MS: A review of *Sarcoptes scabiei*: Past, present and future. Parasit Vectors 10:297, 2017. [PubMed: 28633664]

Goddard J: *Physician's Guide to Arthropods of Medical Importance*, 5th ed. Boca Raton, FL, CRC Press, 2007.

Hinkle N: Ekbom syndrome: The challenge of "invisible bug" infestations. Annu Rev Entomol 55:77, 2010. [PubMed: 19961324]

McGraw TA, Turiansky MC: Cutaneous myiasis. J Am Acad Dermatol 58:907, 2008. [PubMed: 18485982]

Mullen G, Durden L: *Medical and Veterinary Entomology*, 2nd ed. Amsterdam, Academic Press, 2009.

Pollack RJ, Marcus L: A travel medicine guide to arthropods of medical importance. Infect Dis Clin North Am 19:169, 2005. [PubMed: 15701553]

Richards SL et al: Do tick attachment times vary between different tick–pathogen systems? Environments 4:37, 2017.

Saucier JR: Arachnid envenomation. Emerg Med Clin North Am 22:405, 2004. [PubMed: 15163574]

Steen CJ et al: Insect sting reactions to bees, wasps, and ants. Int J Dermatol 44:91, 2005. [PubMed: 15689203]

Vetter RS, Isbister GK: Medical aspects of spider bites. Annu Rev Entomol 53:409, 2008. [PubMed: 17877450]

Chapter 453: Altitude Illness

Buddha Basnyat; Geoffrey Tabin

EPIDEMIOLOGY

Mountains cover one-fifth of the earth's surface; 140 million people live permanently at altitudes ≥2500 m, and 100 million people travel to high-altitude locations each year. Skiers in the Alps or Aspen; tourists to La Paz, Ladakh, or Lahsa; religious pilgrims to Kailash-Manasarovar or Gosainkunda; trekkers and climbers to Kilimanjaro, Aconcagua, or Everest; miners working in high-altitude sites in South America; and military personnel deployed to high-altitude locations are all at risk of developing acute mountain sickness (AMS), high-altitude cerebral edema (HACE), high-altitude pulmonary edema (HAPE), and other altitude-related problems. AMS is the benign form of altitude illness, whereas HACE and HAPE are life-threatening. Altitude illness is likely to occur above 2500 m but has been documented even at 1500–2500 m. In the Mount Everest region of Nepal, ~50% of trekkers who walk to altitudes >4000 m over ≥5 days develop AMS, as do 84% of people who fly directly to 3860 m. The incidences of HACE and HAPE are much lower than that of AMS, with estimates in the range of 0.1–4%. Finally, reentry HAPE, which in the past was generally limited to highlanders (long-term residents of altitudes >2500 m) in the Americas, is now being seen in Himalayan and Tibetan highlanders—and often misdiagnosed as a viral illness—as a result of recent rapid air, train, and motorable-road access to high-altitude settlements.

PHYSIOLOGY

Ascent to a high altitude subjects the body to a decrease in barometric pressure that results in a decreased partial pressure of oxygen in the inspired gas in the lungs. This change leads in turn to less pressure driving oxygen diffusion from the alveoli and throughout the oxygen cascade. A normal initial "struggle response" to such an ascent includes increased ventilation—the cornerstone of acclimatization—mediated by the carotid bodies. Hyperventilation may cause respiratory alkalosis and dehydration. Respiratory alkalosis may be extreme, with an arterial blood pH of >7.7 (e.g., at the summit of Everest). Alkalosis may depress the ventilatory drive during sleep, with consequent periodic breathing and hypoxemia. During early acclimatization, renal suppression of carbonic anhydrase and excretion of dilute alkaline urine combat alkalosis and tend to bring the pH of the blood to normal. Other physiologic changes during normal acclimatization include increased sympathetic tone; increased erythropoietin levels, leading to increased hemoglobin levels and red blood cell mass; increased tissue capillary density and mitochondrial numbers; and higher levels of 2,3-bisphosphoglycerate, enhancing oxygen utilization. Even with normal

acclimatization, however, ascent to a high altitude decreases maximal exercise capacity (by ~1% for every 100 m gained above 1500 m) and increases susceptibility to cold injury due to peripheral vasoconstriction. If the ascent is made faster than the body can adapt to the stress of hypobaric hypoxemia, altitude-related disease states can result.

GENETICS



Hypoxia-inducible factor, which acts as a master switch in high-altitude adaptation, controls transcriptional responses to hypoxia throughout the body and is involved in the release of vascular endothelial growth factor (VEGF) in the brain, erythropoiesis, and other pulmonary and cardiac functions at high altitudes. In particular, the gene *EPAS1*, which codes for transcriptional regulator hypoxia-inducible factor 2α, appears to play an important role in the adaptation of Tibetans living at high altitude, resulting in lower hemoglobin concentrations than are found in Han Chinese or South American highlanders. Other genes implicated include *EGLN1* and *PPARA*, which are also associated with hemoglobin concentration. Some evidence indicates that these genetic changes occurred within the past 3000 years, which is very fast in evolutionary terms. An intriguing question is whether the Sherpas' well-known mountain-climbing ability is partially attributable to their Tibetan ancestry, with overrepresentation of variants of *EPAS*. A striking recent finding is that some of these genetic characteristics may stem from those of Denisovan hominids who were contemporaries of the Neanderthals.

For acute altitude illness, a single gene variant is unlikely to be found, but differences in the susceptibility of individuals and populations, familial clustering of cases, and a positive association of some genetic variants all clearly support a role for genetics. Approximately 58 candidate genes have been tested, and at least one variant from 17 of these genes is associated with altitude illness.

ACUTE MOUNTAIN SICKNESS AND HIGH-ALTITUDE CEREBRAL EDEMA

AMS is a neurologic syndrome characterized by nonspecific symptoms (headache, nausea, fatigue, and dizziness), with a paucity of physical findings, developing 6–12 h after ascent to a high altitude. AMS is a clinical diagnosis. For uniformity in research studies, the Lake Louise Scoring System, created at the 1991 International Hypoxia Symposium, is generally used. AMS must be distinguished from exhaustion, dehydration, hypothermia, alcoholic hangover, and hyponatremia. AMS and HACE are thought to represent opposite ends of a continuum of altitude-related neurologic disorders. HACE (but not AMS) is an encephalopathy whose hallmarks are ataxia and altered consciousness with diffuse cerebral involvement but generally without focal neurologic deficits. Progression to these signal manifestations can be rapid. Papilledema and, more commonly, retinal hemorrhages may develop. In fact, retinal hemorrhages occur frequently at ≥5000 m, even in individuals without clinical symptoms of AMS or HACE. It is unclear whether retinal hemorrhage and cerebral hemorrhage at high altitude are caused by the same mechanism.

Risk Factors

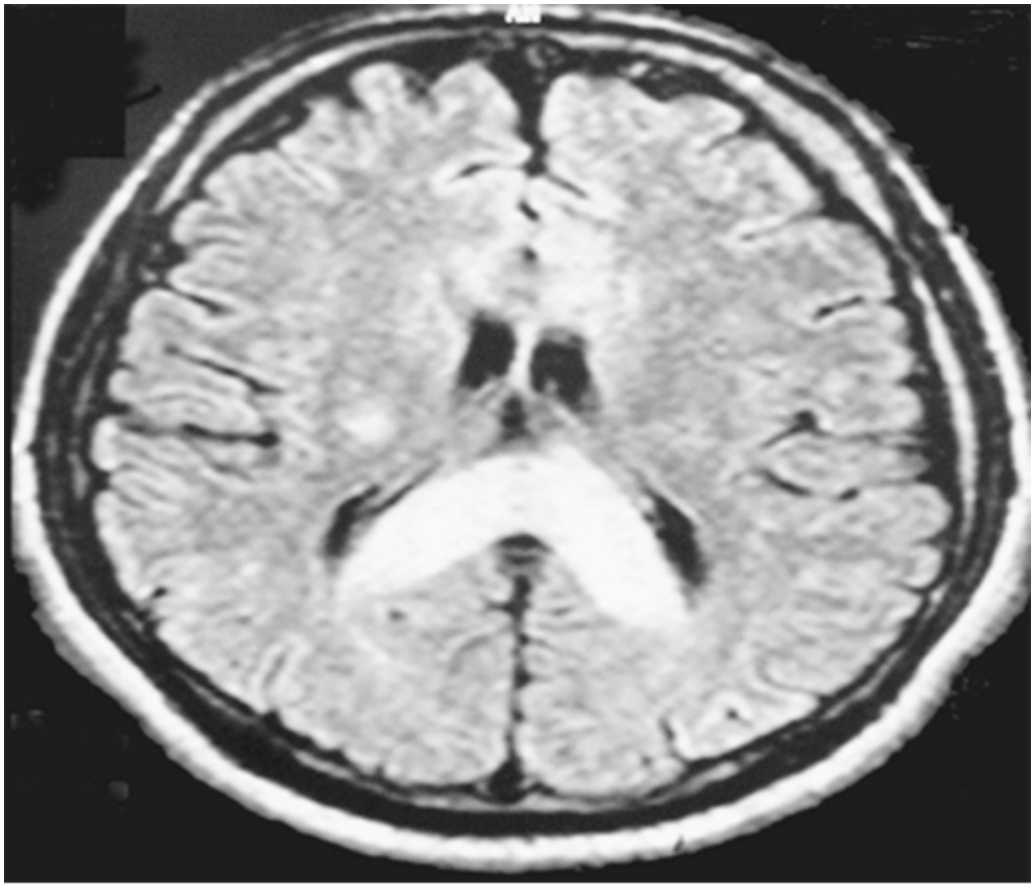
The most important risk factors for the development of altitude illness are the rate of ascent and a prior history of high-altitude illness. Exertion is a risk factor, but lack of physical fitness is not. An attractive but still speculative hypothesis proposes that AMS develops in people who have inadequate cerebrospinal capacity to buffer the brain swelling that occurs at high altitude. Children and adults seem to be equally affected, but people >50 years of age may be less likely to develop AMS than younger people. Most studies reveal no gender difference in AMS incidence. One study showed that, in women, adaptive responses to hypoxia with aging are blunted by menopause but can be maintained with endurance training. Sleep desaturation—a common phenomenon at high altitude—is associated with AMS. Debilitating fatigue consistent with severe AMS on descent from a summit is an important risk factor for death in mountaineers. A prospective study involving trekkers and climbers who ascended to altitudes between 4000 m and 8848 m showed that high oxygen desaturation and low ventilatory response to hypoxia during exercise are independent predictors of severe altitude illness. However, because there may be a large overlap between groups of susceptible and nonsusceptible individuals, accurate cutoff values are hard to define. Prediction is made more difficult because the pretest probabilities of HAPE and HACE are low. Neck irradiation or surgery damaging the carotid bodies, respiratory tract infections, and dehydration appear to be other potential risk factors for altitude illness. Unless guided by clinical signs and symptoms, pulse oximeter readings alone on a trek should not be used to predict AMS.

Pathophysiology

Hypobaric hypoxia is the main trigger for altitude illness. In established AMS, raised intracranial pressure, increased sympathetic activity, relative hypoventilation, fluid retention and redistribution, and impaired gas exchange have all been well noted; these factors may play an important role in the pathophysiology of AMS. Severe hypoxemia can lead to a greater than normal increase in cerebral blood flow. However, the exact mechanisms underlying AMS and HACE are unknown. Evidence points to a central nervous system process. MRI studies have suggested that vasogenic (interstitial) cerebral edema is a component of the pathophysiology of HACE. In the setting of high-altitude illness, the MRI findings shown in **Fig. 453-1** are confirmatory of HACE, with increased signal in the white matter and particularly in the splenium of the corpus callosum. In addition, hemosiderin deposits in the corpus callosum have been characterized as long-lasting footprints of HACE. Quantitative analysis in a 3-tesla MRI study revealed that hypoxia is associated with mild vasogenic cerebral edema irrespective of AMS. This finding is in keeping with case reports of suddenly symptomatic brain tumors and of cranial nerve palsies without AMS at high altitudes. Vasogenic edema may become cytotoxic (intracellular) in severe HACE.

FIGURE 453-1

T2 magnetic resonance image of the brain of a patient with high-altitude cerebral edema (HACE) shows marked swelling and a hyperintense posterior body and splenium of the corpus callosum (area with dense opacity). The patient, a climber, went on to climb Mount Everest about 9 months after this episode of HACE. *(With permission from Wilderness Environ Med 15:53–55, 2004.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Impaired cerebral autoregulation in the presence of hypoxic cerebral vasodilation and altered permeability of the blood–brain barrier due to hypoxia-induced chemical mediators like histamine, arachidonic acid, and VEGF may all contribute to brain edema. In 1995, VEGF was first proposed as a potent promoter of capillary leakage in the brain at high altitude, and studies in mice have borne out this role. Although studies of VEGF in climbers have yielded inconsistent results regarding its association with altitude illness, indirect evidence of a role for this growth factor in AMS and HACE comes from the observation that dexamethasone, when used in the prevention and treatment of these conditions, blocks hypoxic upregulation of VEGF. Other factors in the development of cerebral edema may be the release of calcium-mediated nitric oxide and neuronally mediated adenosine, which may promote cerebral vasodilation. Venous outflow obstruction resulting in increased brain capillary pressure is also thought to play an important role in the development of HACE. Lesions in the globus pallidum (which is sensitive to hypoxia) leading to Parkinson's disease have been reported to be complications of HACE.

The pathophysiology of the most common and prominent symptom of AMS—headache—remains unclear because the brain itself is an insensate organ; only the meninges contain trigeminal sensory nerve fibers. The cause of high-altitude headache is multifactorial. Various chemicals and mechanical factors activate a final common pathway, the trigeminovascular system. In the genesis of high-altitude headache, the response to nonsteroidal anti-inflammatory drugs and glucocorticoids provides indirect evidence for involvement of the arachidonic acid pathway and inflammation. Although high altitude may be a trigger for migraine, it is unclear whether high-altitude headache and migraine share the same pathophysiology.

Prevention and Treatment

Gradual ascent, with adequate time for acclimatization, is the best method for the prevention of altitude illness (Table 453-1). Even though there may be individual variation in the rate of acclimatization, a conservative approach would be a graded ascent of ≤300 m from the previous day's sleeping altitude above 3000 m, and taking every third day of gain in sleeping altitude as an extra day for acclimatization is helpful. Spending one night at an intermediate altitude before proceeding to a higher altitude may enhance acclimatization and attenuate the risk of AMS. Another protective factor in AMS is high-altitude exposure during the preceding 2 months; for example, the incidence and severity of AMS at 4300 m are reduced by 50% with an ascent after 1 week at an altitude ≥2000 m rather than with an ascent from sea level. Studies have examined whether exposure to a normobaric hypoxic environment (in a room or a tent) before an ascent can provide protection against AMS. In double-blind placebo-controlled trials, repeated intermittent exposure (60–90 min) to normobaric hypoxia (up to 4500 m) or continuous exposure to 3000 m during 8 h of sleep for 7 consecutive days failed to reduce the incidence of AMS at altitudes of 4300–4559 m. However, a recent study with 14 consecutive nights in a similar setting demonstrated reduced symptoms and incidence of AMS. More uniform studies need to be done to clarify the utility, if any, of this approach to preacclimatization.

TABLE 453-1

Management of Altitude Illness

Condition	Management
Acute mountain sickness (AMS), mild[a]	Discontinuation of ascent Treatment with acetazolamide (250 mg q12h) Descent[b]
AMS, moderate[a]	Immediate descent for worsening symptoms Use of low-flow oxygen if available Treatment with acetazolamide (250 mg q12h) and/or dexamethasone (4 mg q6h)[c] Hyperbaric therapy[d]
High-altitude cerebral edema (HACE)	Immediate descent or evacuation Administration of oxygen (2–4 L/min) Treatment with dexamethasone (8 mg PO/IM/IV; then 4 mg q6h) Hyperbaric therapy if descent is not possible
High-altitude pulmonary edema (HAPE)	Immediate descent or evacuation Minimization of exertion while patient is kept warm Administration of oxygen (4–6 L/min) to bring O_2 saturation to >90% Adjunctive therapy with nifedipine[e] (30 mg, extended-release, q12h) Hyperbaric therapy if descent is not possible

[a] Categorization of cases as mild or moderate is a subjective judgment based on the severity of headache and the presence and severity of other manifestations (nausea, fatigue, dizziness, insomnia). [b] No fixed altitude is specified; the patient should descend to a point below that at which symptoms developed. [c] Acetazolamide treats and dexamethasone masks symptoms. For prevention (as opposed to treatment) of AMS, 125–250 mg of acetazolamide q12h or (when acetazolamide is contraindicated—e.g., in people with sulfa allergy) 4 mg of dexamethasone q12h may be used. [d] In hyperbaric therapy (Fig. 453-2), the patient is placed in a portable altitude chamber or bag to simulate descent. [e] Nifedipine at this dose is also effective for the prevention of HAPE, as is salmeterol (125 mg inhaled twice daily), tadalafil (10 mg twice daily), or dexamethasone (8 mg twice daily).

Clearly, a flexible itinerary that permits additional rest days will be helpful. Sojourners to high-altitude locations must be aware of the symptoms of altitude illness and should be encouraged not to ascend further if these symptoms develop. Any hint of HAPE (see below) or HACE mandates descent. Proper hydration (but

not overhydration) in high-altitude trekking and climbing, aimed at countering fluid loss due to hyperventilation and sweating, may play a role in avoiding AMS. Pharmacologic prophylaxis at the time of travel to high altitudes is warranted for people with a history of AMS or when a graded ascent and acclimatization are not possible—e.g., when rapid ascent is necessary for rescue purposes or when flight to a high-altitude location is required. Acetazolamide is the drug of choice for AMS prevention. It inhibits renal carbonic anhydrase, causing prompt bicarbonate diuresis that leads to metabolic acidosis and hyperventilation. Acetazolamide (125–250 mg twice a day), administered for 1 day before ascent and continued for 2 or 3 days, is effective. Treatment can be restarted if symptoms return after discontinuation of the drug. Higher doses are not required. A meta-analysis limited to randomized controlled trials revealed that 125 mg of acetazolamide twice daily was effective in the prevention of AMS, with a relative-risk reduction of ~48% from values obtained with placebo. Paresthesia and a tingling sensation are common side effects of acetazolamide. This drug is a nonantibiotic sulfonamide that has low-level cross-reactivity with sulfa antibiotics; as a result, severe reactions are rare. Dexamethasone (8 mg/d in divided doses) is also effective. A large-scale, randomized, double-blind, placebo-controlled trial in partially acclimatized trekkers clearly showed that *Ginkgo biloba* is ineffective in the prevention of AMS. In randomized studies, ibuprofen (600 mg three times daily) has been shown to be beneficial in the prevention of AMS. Recently, acetaminophen (1 g three times daily) was as effective as ibuprofen at the above dosage in a randomized, double-blind study. However, more definitive studies and (for ibuprofen) a proper gastrointestinal-bleeding risk assessment need to be conducted before these drugs can be routinely recommended for AMS prevention. Many drugs, including spironolactone, medroxyprogesterone, magnesium, calcium channel blockers, and antacids, confer no benefit in the prevention of AMS. Starkly conflicting results from a number of trials of inhaled budesonide for the prevention of AMS have recently been published, but, in all likelihood, the drug is ineffective. Similarly, no efficacy studies are available for coca leaves (a weak form of cocaine), which are offered to high-altitude travelers in the Andes, or for *soroche* pills, which contain aspirin, caffeine, and acetaminophen and are sold over the counter in Bolivia and Peru. Finally, a word of caution applies in the pharmacologic prevention of altitude illness. A fast-growing population of climbers in pursuit of a summit are using prophylactic drugs such as glucocorticoids in an attempt to improve their performance; the outcome can be tragic because of potentially severe side effects of these drugs.

For the treatment of mild AMS, rest alone with analgesic use may be adequate. Descent and the use of acetazolamide and (if available) oxygen are sufficient to treat most cases of moderate AMS. Even a minor descent (400–500 m) may be adequate for symptom relief. For moderate AMS or early HACE, dexamethasone (4 mg orally or parenterally) is highly effective. For HACE, immediate descent is mandatory. When descent is not possible because of poor weather conditions or darkness, a simulation of descent in a portable hyperbaric chamber **(Fig. 453-2)** can be very effective. Pressurization in the bag for 1–2 h often leads to spectacular improvement and, like dexamethasone administration, “buys time.” Thus, in certain high-altitude locations (e.g., remote pilgrimage sites), the decision to bring along the lightweight hyperbaric chamber may prove lifesaving. Like nifedipine, phosphodiesterase-5 inhibitors have no role in the treatment of AMS or HACE.

FIGURE 453-2

A hyperbaric bag. The cylindrical, portable (<7 kg) nylon bag has a one-way valve to prevent carbon dioxide buildup. A patient with severe acute mountain sickness (AMS), high-altitude cerebral edema (HACE), or high-altitude pulmonary edema (HAPE) is zipped inside the bag, which is continuously inflated with a foot pedal. The increased barometric pressure (2 psi) inside the bag simulates descent; for example, at 4250 m, the equivalent "elevation" inside the bag is ~2100 m. No supplemental oxygen is required.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

HIGH-ALTITUDE PULMONARY EDEMA

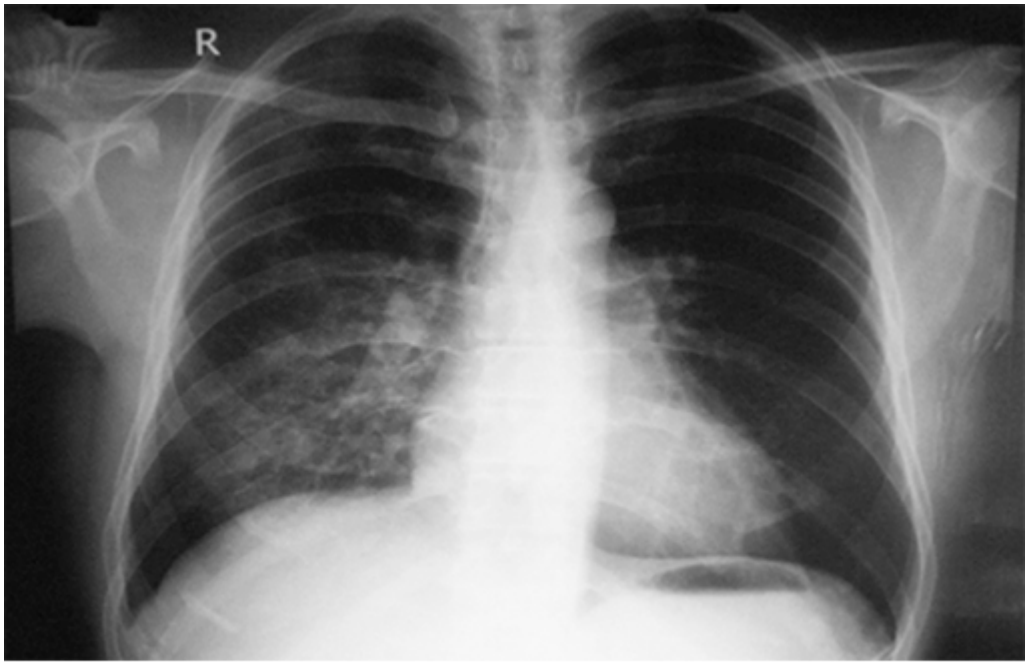
Risk Factors and Manifestations

Unlike HACE (a neurologic disorder), HAPE is primarily a pulmonary problem and therefore is not necessarily preceded by AMS. HAPE develops within 2–4 days after arrival at high altitude; it rarely occurs after more than 4 or 5 days at the same altitude, probably because of remodeling and adaptation that render the pulmonary vasculature less susceptible to the effects of hypoxia. A rapid rate of ascent, a history of HAPE, respiratory tract infections, and cold environmental temperatures are risk factors. Men are more susceptible than women. People with abnormalities of the cardiopulmonary circulation leading to pulmonary hypertension—e.g., a large patent foramen ovale, mitral stenosis, primary pulmonary hypertension, and unilateral absence of the pulmonary artery—may be at increased risk of HAPE, even at moderate altitudes. For example, patent foramen ovale is four times more common among HAPE-susceptible individuals than in the general population. Echocardiography is recommended when HAPE develops at relatively low altitudes (<3000 m) and whenever cardiopulmonary abnormalities predisposing to HAPE are suspected.

The initial manifestation of HAPE may be a reduction in exercise tolerance greater than that expected at the given altitude. Although a dry, persistent cough may presage HAPE and may be followed by the production of blood-tinged sputum, cough in the mountains is almost universal and the mechanism is poorly understood. Tachypnea and tachycardia, even at rest, are important markers as illness progresses. Crackles may be heard on auscultation but are not diagnostic. HAPE may be accompanied by signs of HACE. Patchy or localized opacities **(Fig. 453-3)** or streaky interstitial edema may be noted on chest radiography. In the past, HAPE was mistaken for pneumonia due to the cold or for heart failure due to hypoxia and exertion. Kerley B lines or a bat-wing appearance are not seen on radiography. Electrocardiography may reveal right ventricular strain or even hypertrophy. Hypoxemia and respiratory alkalosis are consistently present unless the patient is taking acetazolamide, in which case metabolic acidosis may supervene. Assessment of arterial blood gases is not necessary in the evaluation of HAPE; an oxygen saturation reading with a pulse oximeter is generally adequate. The existence of a subclinical form of HAPE has been suggested by an increased alveolar–arterial oxygen gradient in Everest climbers near the summit, but hard evidence correlating this abnormality with the development of clinically relevant HAPE is lacking. Comet-tail scoring—an ultrasound technique—has been used for evaluation of extravascular lung water at high altitude; the results have been encouraging, and this technique may potentially prove useful in detecting HAPE (clinical or subclinical) and even in ascertaining whether the presence of extravascular lung water is a harbinger of HAPE in patients with AMS.

FIGURE 453-3

Chest radiograph of a patient with high-altitude pulmonary edema shows opacity in the right middle and lower zones simulating pneumonic consolidation. The opacity cleared almost completely in 2 days with descent and supplemental oxygen.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Pathophysiology

HAPE is a noncardiogenic pulmonary edema with normal pulmonary artery wedge pressure. It is characterized by patchy pulmonary hypoxic vasoconstriction that leads to overperfusion in some areas. This abnormality leads in turn to increased pulmonary capillary pressure (>18 mmHg) and capillary "stress" failure. The exact mechanism for this hypoxic vasoconstriction is unknown. Endothelial dysfunction due to hypoxia may play a role by impairing the release of nitric oxide, an endothelium-derived vasodilator. At high altitude, HAPE-prone persons have reduced levels of exhaled nitric oxide. The effectiveness of phosphodiesterase-5 inhibitors in alleviating altitude-induced pulmonary hypertension, decreased exercise tolerance, and hypoxemia supports the role of nitric oxide in the pathogenesis of HAPE. One study demonstrated that prophylactic use of tadalafil, a phosphodiesterase-5 inhibitor, decreases the risk of HAPE by 65%. In contrast, the endothelium also synthesizes endothelin-1, a potent vasoconstrictor whose concentrations are higher than average in HAPE-prone mountaineers.

Exercise and cold lead to increased pulmonary intravascular pressure and may predispose to HAPE. In addition, hypoxia-triggered increases in sympathetic drive may lead to pulmonary venoconstriction and extravasation into the alveoli from the pulmonary capillaries. Consistent with this concept, phentolamine, which elicits α-adrenergic blockade, improves hemodynamics and oxygenation in HAPE more than do other vasodilators. The study of tadalafil cited above also investigated dexamethasone in the prevention of HAPE. Surprisingly, dexamethasone reduced the incidence of HAPE by 78%—a greater decrease than with tadalafil. Besides possibly increasing the availability of endothelial nitric oxide, dexamethasone may have altered the excessive sympathetic activity associated with HAPE: the heart rate of participants in the dexamethasone arm of the study was significantly lowered. Finally, people susceptible to HAPE also display enhanced sympathetic activity during short-term hypoxic breathing at low altitudes.

Because many patients with HAPE have fever, peripheral leukocytosis, and an increased erythrocyte sedimentation rate, inflammation has been considered an etiologic factor in HAPE. However, strong evidence suggests that inflammation in HAPE is an epiphenomenon rather than the primary cause. Nevertheless, inflammatory processes (e.g., those elicited by viral respiratory tract infections) do predispose persons to HAPE—even those who are constitutionally resistant to its development.

Another proposed mechanism for HAPE is impaired transepithelial clearance of sodium and water from the alveoli. β-Adrenergic agonists upregulate the clearance of alveolar fluid in animal models. In a double-blind, randomized, placebo-controlled study of HAPE-susceptible mountaineers, prophylactic inhalation of the β-adrenergic agonist salmeterol reduced the incidence of HAPE by 50%. Other effects of β agonists may also contribute to the prevention of HAPE, but these findings are in keeping with the concept that alveolar fluid clearance may play a pathogenic role in this illness.

Prevention and Treatment

Allowing sufficient time for acclimatization by ascending gradually (as discussed above for AMS and HACE) is the best way to prevent HAPE (Table 453-1). Sustained-release nifedipine (30 mg), given once or twice daily, prevents HAPE in people who must ascend rapidly or who have a history of HAPE. Other drugs for the prevention of HAPE are listed in Table 453-1 (footnote e). Although dexamethasone is listed for prevention, its

adverse-effect profile requires close monitoring. Acetazolamide has been shown to blunt hypoxic pulmonary vasoconstriction in animal models, and this observation warrants further study in HAPE prevention. However, one large study failed to show a decrease in pulmonary vasoconstriction in partially acclimatized individuals given acetazolamide.

Early recognition is paramount in the treatment of HAPE, especially when it is not preceded by the AMS symptoms of headache and nausea. Fatigue and dyspnea at rest may be the only initial manifestations. Descent and the use of supplementary oxygen (aimed at bringing oxygen saturation to >90%) are the most effective therapeutic interventions. Exertion should be kept to a minimum, and the patient should be kept warm. Hyperbaric therapy (Fig. 453-2) in a portable altitude chamber may be lifesaving, especially if descent is not possible and oxygen is not available. Oral sustained-release nifedipine (30 mg once or twice daily) can be used as adjunctive therapy. Inhaled β agonists, which are safe and convenient to carry, are useful in the prevention of HAPE and may be effective in its treatment, although no trials have yet been carried out. Inhaled nitric oxide and expiratory positive airway pressure may also be useful therapeutic measures but may not be available in high-altitude settings. No studies have investigated phosphodiesterase-5 inhibitors in the treatment of HAPE, but reports have described their use in clinical practice. The mainstays of treatment remain descent and (if available) oxygen.

In AMS, if symptoms abate (with or without acetazolamide), the patient may reascend gradually to a higher altitude. Unlike that in acute respiratory distress syndrome (another noncardiogenic pulmonary edema), the architecture of the lung in HAPE is usually well-preserved, with rapid reversibility of abnormalities (Fig. 453-3). This fact has allowed some people with HAPE to reascend slowly after a few days of descent and rest. In HACE, reascent after a few days may not be advisable during the same trip.

OTHER HIGH-ALTITUDE PROBLEMS

Sleep Impairment

The mechanisms underlying sleep problems, which are among the most common adverse reactions to high altitude, include increased periodic breathing; changes in sleep architecture, with increased time in lighter sleep stages; and changes in rapid eye movement sleep. Sojourners should be reassured that sleep quality improves with acclimatization. In cases where drugs do need to be used, acetazolamide (125 mg before bedtime) is especially useful because this agent decreases hypoxemic episodes and alleviates sleeping disruptions caused by excessive periodic breathing. Whether combining acetazolamide with temazepam or zolpidem is more effective than administering acetazolamide alone is unknown. In combinations, the doses of temazepam and zolpidem should not be increased by >10 mg at high altitudes. Limited evidence suggests that diazepam causes hypoventilation at high altitudes and therefore is contraindicated. For trekkers with obstructive sleep apnea who are using a continuous positive airway pressure (CPAP) machine, the addition of acetazolamide, which will decrease centrally mediated sleep apnea, may be helpful. There is evidence to show that obstructive sleep apnea at high altitude may decrease and “convert” to central sleep apnea.

Gastrointestinal Issues

High-altitude exposure may be associated with increased gastric and duodenal bleeding, but further studies are required to determine whether there is a causal effect. Because of decreased atmospheric pressure and consequent intestinal gas expansion at high altitudes, many sojourners experience abdominal bloating and distension as well as excessive flatus expulsion. In the absence of diarrhea, these phenomena are normal, if sometimes uncomfortable. Accompanying diarrhea, however, may indicate the involvement of bacteria or *Giardia* parasites, which are common at many high-altitude locations in the developing world. Prompt treatment with fluids and empirical antibiotics may be required to combat dehydration in the mountains. Hemorrhoids are common on high-altitude treks; treatment includes hot soaks, application of hydrocortisone ointment, and measures to avoid constipation.

High-Altitude Cough

High-altitude cough can be debilitating and is sometimes severe enough to cause rib fracture, especially at >5000 m. The etiology of this common problem is probably multifactorial. Although high-altitude cough has been attributed to inspiration of cold dry air, this explanation appears not to be sufficient by itself; in long-duration studies in hypobaric chambers, cough has occurred despite controlled temperature and humidity. The implication is that hypoxia also plays a role. Exercise can precipitate cough at high altitudes, possibly because of water loss from the respiratory tract. Long-acting β agonists and glucocorticoids prevent bronchoconstriction that otherwise may be brought on by cold and exercise. In general, infection does not seem to be a common etiology. Anecdotal reports have described the efficacy of an inhaled combination of fluticasone and salmeterol in the treatment of high-altitude cough. Many trekkers find it useful to wear a balaclava to trap some moisture and heat. In most situations, cough resolves upon descent.

High-Altitude Neurologic Events Unrelated to “Altitude Illness”

Transient ischemic attacks (TIAs) and strokes have been well described in high-altitude sojourners outside the setting of altitude sickness. However, these descriptions are not based on cause (hypoxia) and effect. In general, symptoms of AMS present gradually, whereas many of these neurologic events happen suddenly. The population that suffers strokes and TIAs at sea level is generally an older age group with other risk factors, whereas those so afflicted at high altitudes are generally younger and probably have fewer risk factors for atherosclerotic vascular disease. Other mechanisms (e.g., migraine, vasospasm, focal edema, hypocapneic vasoconstriction, hypoxia in the watershed zones of minimal cerebral blood flow, or cardiac right-to-left shunt) may be operative in TIAs and strokes at high altitude.

Subarachnoid hemorrhage, transient global amnesia, delirium, and cranial nerve palsies (e.g., lateral rectus palsy) occurring at high altitudes but outside the setting of altitude sickness have been well described. Syncope is common at moderately high altitudes, generally occurs shortly after ascent, usually resolves without descent, and appears to be a vasovagal event related to hypoxemia. Seizures occur rarely with HACE, but hypoxemia and hypocapnia, which are prevalent at high altitudes, are well-known triggers that may contribute to new or breakthrough seizures in predisposed individuals. Nevertheless, the consensus among experts is that sojourners with well-controlled seizure disorders can ascend to high altitudes.

Finally, persons with hypercoagulable conditions (e.g., antiphospholipid syndrome, protein C deficiency) who are asymptomatic at sea level may experience cerebral venous thrombosis (possibly due to enhanced blood viscosity triggered by polycythemia and dehydration) at high altitudes. Proper history taking, examination, and prompt investigations where possible will help define these conditions as entities separate from altitude sickness. Administration of oxygen (where available) and prompt descent are the cornerstones of treatment of most of these neurologic conditions.

Ocular Problems

Ocular issues are common in sojourners to high altitudes. Hypoxemia induced by altitude leads to increased retinal blood flow, which can be visible as engorged retinal veins on ophthalmoscopic examination. Both high flow and hypoxemic vascular damage causing permeability have been implicated in a breakdown of the blood–retina barrier and the formation of retinal hemorrhages. Blot, dot, flame, and white-centered hemorrhages can be observed. These hemorrhages usually resolve spontaneously with descent, with only mild symptoms and no lasting visual damage in most healthy eyes. The exception is hemorrhage in the macular area. Macular hemorrhages can cause devastating initial visual loss, particularly if bilateral, and have been reported to cause permanently decreased vision in a few cases.

Stroke syndromes such as retinal vein occlusion, retinal artery occlusion, ischemic optic neuropathy, and cortical visual loss have all been reported. With unilateral vision loss, it is always important to check for a relative afferent pupillary defect. Increased hematocrit combined with dehydration may contribute to these maladies. Glaucomatous optic nerve damage may progress with hypoxemia of altitude. Acetazolamide is helpful both in combating the respiratory alkalosis that comes with increased ventilation at high altitude and in lowering the interocular pressure; its use should be considered in patients with stable controlled glaucoma. Macular degeneration and diabetic eye disease are not directly exacerbated by ascent to high altitude. Dry eye and solar damage to the cornea, known as “snow blindness,” are common. Wearing of high-quality UV blocking sunglasses, even on cloudy days, and attention to protecting and supplementing the tear film with artificial tear drops can greatly improve comfort and vision. Although modern refractive surgeries, such as photorefractive keratectomy (PRK) and laser in situ keratomileusis (LASIK), are stable at high altitude, patients who have undergone radial keratotomy should be cautioned that hypoxemia to the cornea can lead to swelling that shifts the refraction during ascent.

Psychological/Psychiatric Problems

Delirium characterized by a sudden change in mental status, a short attention span, disorganized thinking, and an agitated state during the period of confusion has been well described in mountain climbers and trekkers without a prior history. In addition, anxiety attacks, often triggered at night by excessive periodic breathing, are well documented. The contribution of hypoxia to these conditions is unknown. Expedition medical kits need to include antipsychotic injectable drugs to control psychosis in patients in remote high-altitude locations.

PREEXISTING MEDICAL ISSUES

Because travel to high altitudes is increasingly popular, common conditions such as hypertension, coronary artery disease, and diabetes are more frequently encountered among high-altitude sojourners. This situation is of particular concern for the millions of elderly pilgrims with medical problems who visit high-altitude sacred areas (e.g., in the Himalayas) each year. In recent years, high-altitude travel has attracted intrepid trekkers who are taking immunosuppressive medications (e.g., kidney transplant recipients or patients undergoing chemotherapy). Recommended vaccinations and other precautions (e.g., hand washing) may be especially important for this group. Although most of these medical conditions do not appear to influence susceptibility to altitude illness, they may be exacerbated by ascent to altitude, exertion in cold conditions, and hypoxemia. Advice regarding the advisability of high-altitude travel and the impact of high-altitude hypoxia on these preexisting conditions is becoming increasingly relevant, but there are no evidence-based guidelines. In addition, recommendations made for relatively low altitudes (~3000 m) may not hold true for higher altitudes (>4000 m), where hypoxic stress is greater. Personal risks and benefits must be clearly thought through before ascent.

Hypertension

At high altitudes, enhanced sympathetic activity may lead to a transient rise in blood pressure. Occasionally, nonhypertensive, healthy, asymptomatic trekkers have pathologically high blood pressure at high altitude that rapidly normalizes without medicines on descent. Sojourners should continue to take their antihypertensive medications at high altitudes. Hypertensive patients are not more likely than others to develop altitude illness. Because the probable mechanism of high-altitude hypertension is α-adrenergic activity, anti-α-adrenergic drugs like prazosin have been suggested for symptomatic patients and those with labile hypertension. It is best to start taking the drug several weeks before the trip and to carry a sphygmomanometer if a trekker has labile hypertension. Sustained-release nifedipine may also be useful. A recent observational cohort study of 672 hypertensive and nonhypertensive trekkers in the Himalayas showed that most travelers, including those with well-controlled hypertension, can be reassured that their blood pressure will remain relatively stable at high altitude. Although blood pressure may be extremely elevated at high altitude in normotensive and hypertensive people, it is unlikely to cause symptoms.

Coronary Artery Disease

Myocardial oxygen demand and maximal heart rate are reduced at high altitudes because the VO_2 max (maximal oxygen consumption) decreases with increasing altitude. This effect may explain why signs of cardiac ischemia or dysfunction usually are not seen in healthy persons at high altitudes. Asymptomatic, fit individuals with no risk factors need not undergo any tests for coronary artery disease before ascent. For persons with ischemic heart disease, previous myocardial infarction, angioplasty, and/or bypass surgery, an exercise treadmill test is indicated. A strongly positive treadmill test is a contraindication for high-altitude trips. Patients with poorly controlled arrhythmias should avoid high-altitude travel, but patients with arrhythmias that are well controlled with antiarrhythmic medications do not seem to be at increased risk.

Sudden cardiac deaths are not noted with a greater frequency in the Alps than at lower altitudes; although sudden cardiac deaths are encountered every trekking season in the higher Himalayan range, accurate documentation is lacking.

Asthma

Although cold air and exercise may provoke acute bronchoconstriction, asthmatic patients usually have fewer problems at high than at low altitudes, possibly because of decreased allergen levels and increased circulating catecholamine levels. Nevertheless, asthmatic individuals should carry all their medications, including oral glucocorticoids, with proper instructions for use in case of an exacerbation. Severely asthmatic persons should be cautioned against ascending to high altitudes.

Pregnancy

In general, low-risk pregnant women ascending to 3000 m are not at special risk except for the relative unavailability of medical care in many high-altitude locations, especially in developing countries. Despite the lack of firm data on this point, venturing higher than 3000 m to altitudes at which oxygen saturation drops steeply seems unadvisable for pregnant women.

Obesity

Although living at a high altitude has been suggested as a means of controlling obesity, obesity has also been reported to be a risk factor for AMS, probably because nocturnal hypoxemia is more pronounced in obese individuals. Hypoxemia may also lead to greater pulmonary hypertension, thus possibly predisposing the trekker to HAPE.

Sickle Cell Disease

High altitude is one of the rare environmental exposures that occasionally provokes a crisis in persons with the sickle cell trait. Even when traversing mountain passes as low as 2500 m, people with sickle cell disease have been known to have a vaso-occlusive crisis. Sickle cell disease needs to be considered when persons traveling to high altitudes become unwell and develop left-upper-quadrant pain. Patients with known sickle cell disease who need to travel to high altitudes should use supplemental oxygen and travel with caution. Thalassemia has not been known to cause problems at high altitude.

Diabetes Mellitus

Trekking at high altitudes may enhance sugar uptake. Thus, high-altitude travel may not pose problems for persons with diabetes that is well controlled with oral hypoglycemic agents. An eye examination before travel may be useful. Patients taking insulin may require lower doses on trekking/climbing days than on rest days. Because of these variations, diabetic patients need to carry a reliable glucometer and use fast-acting

insulin. Ready access to sweets is also essential. It is important for companions of diabetic trekkers to be fully aware of potential problems like hypoglycemia.

Chronic Lung Disease

Depending on disease severity and access to medical care, preexisting lung disease may not always preclude high-altitude travel. A proper pretravel evaluation must be conducted. Supplemental oxygen may be required if the predicted PaO_2 for the altitude is <50–55 mmHg. Preexisting pulmonary hypertension may also need to be assessed in these patients. If the result is positive, patients should be discouraged from ascending to high altitudes; if such travel is necessary, treatment with sustained-release nifedipine (20 mg twice a day) should be considered. Small-scale studies have revealed that when patients with bullous disease reach ~5000 m, bullous expansion and pneumothorax are not noted. Compared with information on chronic obstructive pulmonary disease, fewer data exist about the safety of travel to high altitude for people with pulmonary fibrosis, but acute exacerbation of pulmonary fibrosis has been seen at high altitude. A handheld pulse oximeter can be useful to check for oxygen saturation.

Chronic Kidney Disease

Patients with chronic kidney disease can tolerate short-term stays at high altitudes, but theoretical concern persists about progression to end-stage renal disease. Acetazolamide, the drug most commonly used for altitude sickness, should be avoided by anyone with preexisting metabolic acidosis, which can be exacerbated by this drug. In addition, the acetazolamide dosage should be adjusted when the glomerular filtration rate falls to <50 mL/min, and the drug should not be used at all if this value falls to <10 mL/min.

Cirrhosis

Of patients with cirrhosis, 16% may have portopulmonary arterial hypertension and 32% may have hepatopulmonary syndrome; these conditions may be detrimental at high altitude as they may cause exaggerated hypoxemia. Thus, screening for these problems is important in cirrhotic patients planning a high-altitude trip. In addition, acetazolamide may be inadvisable in these patients as the drug may increase the risk of hepatic encephalopathy.

CHRONIC MOUNTAIN SICKNESS AND HIGH-ALTITUDE PULMONARY HYPERTENSION IN HIGHLANDERS

The largest populations of highlanders live in the South American Andes, the Tibetan Plateau, and parts of Ethiopia. Chronic mountain sickness (*Monge's disease*) is a disease in highlanders that is characterized by excessive erythrocytosis with moderate to severe pulmonary hypertension leading to cor pulmonale. This condition was originally described in South America and has also been documented in Colorado and in the Han Chinese population in Tibet; it is much less common in Tibetans or in Ethiopian highlanders. Migration

to a low altitude results in the resolution of chronic mountain illness. Venesection and acetazolamide are helpful.

High-altitude pulmonary hypertension is also a subacute disease of long-term high-altitude residents. Unlike Monge's disease, this syndrome is characterized primarily by pulmonary hypertension (not erythrocytosis) leading to heart failure. Indian soldiers living at extreme altitudes for prolonged periods and Han Chinese infants born in Tibet have presented with the adult and infantile forms, respectively. High-altitude pulmonary hypertension bears a striking pathophysiologic resemblance to brisket disease in cattle. Descent to a lower altitude is curative.

FURTHER READING

Basnyat B: High altitude pilgrimage medicine. High Alt Med Biol 15:434, 2014. [PubMed: 25330393]

Basnyat B, Murdoch D: High altitude illness. Lancet 361:1967, 2003. [PubMed: 12801752]

Basnyat B et al: Acute medical problems in the Himalayas outside the setting of altitude sickness. High Alt Med Biol 1:167, 2000. [PubMed: 11254226]

Kanaan NC et al: Prophylactic acetaminophen or ibuprofen results in equivalent acute mountain sickness incidence at high altitude: A prospective randomized trial. Wilderness Environ Med 28:72, 2017. [PubMed: 28479001]

Roach RC et al: Mountain medicine, in *Wilderness Medicine*, 7th ed. Auerbach PS et al (eds). Philadelphia, Elsevier, 2016, pp 2–39.

Chapter 454: Hypothermia and Peripheral Cold Injuries

Daniel F. Danzl

FIGURE 454-1

HYPOTHERMIA

Accidental hypothermia occurs when there is an unintentional drop in the body's core temperature below 35°C (95°F). At this temperature, many of the compensatory physiologic mechanisms that conserve heat begin to fail. *Primary accidental hypothermia* is a result of the direct exposure of a previously healthy individual to the cold. The mortality rate is much higher for patients who develop *secondary hypothermia* as a complication of a serious systemic disorder or injury.

CAUSES

Primary accidental hypothermia is geographically and seasonally pervasive. Although most cases occur in the winter months and in colder climates, this condition is surprisingly common in warmer regions as well. Multiple variables render individuals at the extremes of age—both the elderly and neonates—particularly vulnerable to hypothermia (Table 454-1). The elderly have diminished thermal perception and are more susceptible to immobility, malnutrition, and systemic illnesses that interfere with heat generation or conservation. Dementia, psychiatric illness, and socioeconomic factors often compound these problems by impeding adequate measures to prevent hypothermia. Neonates have high rates of heat loss because of their increased surface-to-mass ratio and their lack of effective shivering and adaptive behavioral responses. At all ages, malnutrition can contribute to heat loss because of diminished subcutaneous fat and as a result of depleted energy stores used for thermogenesis.

TABLE 454-1

Risk Factors for Hypothermia

Age extremes	Endocrine-related
Elderly	Diabetes mellitus
Neonates	Hypoglycemia
Environmental exposure	Hypothyroidism
Occupational	Adrenal insufficiency
Sports-related	Hypopituitarism
Inadequate clothing	Neurologic
Immersion	Cerebrovascular accident
Toxicologic and pharmacologic	Hypothalamic disorders
Ethanol	Parkinson's disease
Phenothiazines	Spinal cord injury
Barbiturates	Multisystemic
Anesthetics	Trauma
Neuromuscular blockers	Sepsis
Antidepressants	Shock
Insufficient fuel	Hepatic or renal failure
Malnutrition	Carcinomatosis
Marasmus	Burns and exfoliative dermatologic disorders
Kwashiorkor	Immobility or debilitation

Individuals whose occupations or hobbies entail extensive exposure to cold weather are at increased risk for hypothermia. Military history is replete with hypothermic tragedies. Hunters, sailors, skiers, and climbers also are at great risk of exposure, whether it involves injury, changes in weather, or lack of preparedness.

Ethanol causes vasodilation (which increases heat loss), reduces thermogenesis and gluconeogenesis, and may impair judgment or lead to obtundation. Phenothiazines, barbiturates, benzodiazepines, heterocyclic antidepressants, and other medications reduce centrally mediated vasoconstriction. Many hypothermic patients are admitted to intensive care because of drug overdose. Anesthetics can block shivering responses; their effects are compounded when patients are not insulated adequately in the operating or recovery units.

Several types of endocrine dysfunction can lead to hypothermia. Hypothyroidism—particularly when extreme, as in myxedema coma—reduces the metabolic rate and impairs thermogenesis and behavioral responses. Adrenal insufficiency and hypopituitarism also increase susceptibility to hypothermia. Hypoglycemia, most commonly caused by insulin or oral hypoglycemic drugs, is associated with hypothermia, in part because of neuroglycopenic effects on hypothalamic function. Increased osmolality and metabolic derangements associated with uremia, diabetic ketoacidosis, and lactic acidosis can lead to altered hypothalamic thermoregulation.

Neurologic injury from trauma, cerebrovascular accident, subarachnoid hemorrhage, and hypothalamic lesion increases susceptibility to hypothermia. Agenesis of the corpus callosum (*Shapiro's syndrome*) is one cause of episodic hypothermia; in this syndrome, profuse perspiration is followed by a rapid fall in temperature. Acute spinal cord injury disrupts the autonomic pathways that lead to shivering and prevents cold-induced reflex vasoconstrictive responses.

Hypothermia associated with sepsis is a poor prognostic sign. Hepatic failure causes decreased glycogen storage and gluconeogenesis as well as a diminished shivering response. In acute myocardial infarction associated with low cardiac output, hypothermia may be reversed after adequate resuscitation. With extensive burns, psoriasis, erythrodermas, and other skin diseases, increased peripheral-blood flow leads to excessive heat loss.

THERMOREGULATION

Heat loss occurs through five mechanisms: radiation (55–65% of heat loss), conduction (10–15% of heat loss, much increased in cold water), convection (increased in the wind), respiration, and evaporation; both of the latter two mechanisms are affected by the ambient temperature and the relative humidity.

The preoptic anterior hypothalamus normally orchestrates thermoregulation **(Chap. 15)**. The immediate defense of thermoneutrality is via the autonomic nervous system, whereas delayed control is mediated by the endocrine system. Autonomic nervous system responses include the release of norepinephrine, increased muscle tone, and shivering, leading to thermogenesis and an increase in the basal metabolic rate. Cutaneous cold thermoreception causes direct reflex vasoconstriction to conserve heat. Prolonged exposure to cold also stimulates the thyroid axis, leading to an increased metabolic rate.

CLINICAL PRESENTATION

In most cases of hypothermia, the history of exposure to environmental factors (e.g., prolonged exposure to the outdoors without adequate clothing) makes the diagnosis straightforward. In urban settings, however, the presentation is often more subtle, and other disease processes, toxin exposures, or psychiatric diagnoses should be considered. Predicting the core temperature based on the clinical presentation is very difficult.

After initial stimulation by hypothermia, there is progressive depression of all organ systems. The timing of the appearance of these clinical manifestations varies widely **(Table 454-2)**. Without knowing the core temperature, it can be difficult to interpret other vital signs. For example, tachycardia disproportionate to the core temperature suggests

secondary hypothermia resulting from hypoglycemia, hypovolemia, or a toxin overdose. Because carbon dioxide production declines progressively, the respiratory rate should be low; persistent hyperventilation suggests a central nervous system (CNS) lesion or one of the organic acidoses. A markedly depressed level of consciousness in a patient with mild hypothermia raises suspicion of an overdose or CNS dysfunction due to infection or trauma.

TABLE 454-2

Physiologic Changes Associated with Accidental Hypothermia

Severity	Body Temperature	Central Nervous System	Cardiovascular	Respiratory	Renal and Endocrine	Neuromuscular
Mild	35°C (95°F)–32.2°C (90°F)	Linear depression of cerebral metabolism; amnesia; apathy; dysarthria; impaired judgment; maladaptive behavior	Tachycardia, then progressive bradycardia; cardiac cycle prolongation; vasoconstriction; increase in cardiac output and blood pressure	Tachypnea, then progressive decrease in respiratory minute volume; declining oxygen consumption; bronchorrhea; bronchospasm	Diuresis; increase in catecholamines, adrenal steroids, triiodothyronine, and thyroxine; increase in metabolism with shivering	Increased preshivering muscle tone, then fatiguing
Moderate	<32.2°C (90°F)–28°C (82.4°F)	EEG abnormalities; progressive depression of level of consciousness; pupillary dilation; paradoxical undressing; hallucinations	Progressive decrease in pulse and cardiac output; increased atrial and ventricular arrhythmias; suggestive (J-wave) ECG changes	Hypoventilation; 50% decrease in carbon dioxide production per 8°C (17.6°F) drop in temperature; absence of protective airway reflexes	50% increase in renal blood flow; renal autoregulation intact; impaired insulin action	Hyporeflexia; diminishing shivering-induced thermogenesis; rigidity
Severe	<28°C (<82.4°F)	Loss of cerebrovascular autoregulation; decline in cerebral blood flow; coma; loss of ocular reflexes; progressive decrease in EEG abnormalities	Progressive decrease in blood pressure, heart rate, and cardiac output; reentrant dysrhythmias; maximal risk of ventricular fibrillation; asystole	Pulmonic congestion and edema; 75% decrease in oxygen consumption; apnea	Decrease in renal blood flow that parallels decrease in cardiac output; extreme oliguria; poikilothermia; 80% decrease in basal metabolism	No motion; decreased nerve-conduction velocity; peripheral areflexia; no corneal or oculocephalic reflexes

Abbreviations: ECG, electrocardiogram; EEG, electroencephalogram.

Source: Modified from DF Danzl, RS Pozos: N Engl J Med 331:1756, 1994.

Physical examination findings can also be altered by hypothermia. For instance, the assumption that areflexia is solely attributable to hypothermia can obscure and delay the diagnosis of a spinal cord lesion. Patients with hypothermia may be confused or combative; these symptoms abate more rapidly with rewarming than with chemical or physical restraint. A classic example of maladaptive behavior in patients with hypothermia is paradoxical undressing, which involves the inappropriate removal of clothing in response to a cold stress. The cold-induced ileus and abdominal rectus spasm can mimic or mask the presentation of an acute abdomen (Chap. 12).

When a patient in hypothermic cardiac arrest is first discovered, cardiopulmonary resuscitation (CPR) is indicated unless (1) a do-not-resuscitate status is verified, (2) obviously lethal injuries are identified, or (3) the depression of a frozen chest wall is not possible. Continuous CPR is normally recommended and interruptions should be avoided. In the field, when the core temperature is below 28°C, intermittent CPR may be effective.

As the resuscitation proceeds, the prognosis is grave if there is evidence of widespread cell lysis, as reflected by potassium levels >10–12 mmol/L (10–12 meq/L). Other findings that may preclude continuing resuscitation include a core temperature <10–12°C (<50–54°F), a pH <6.5, and evidence of intravascular thrombosis with a fibrinogen value <0.5 g/L (<50 mg/dL). The decision to terminate resuscitation before rewarming the patient past 33°C (91°F) should be predicated on the type and severity of the precipitants of hypothermia. Survival has occurred with a cardiac arrest time over 7 h. There are no validated prognostic indicators for recovery from hypothermia. A history of asphyxia with secondary cooling is the most important negative predictor of survival.

DIAGNOSIS AND STABILIZATION

Hypothermia is confirmed by measurement of the core temperature, preferably at two sites. Rectal probes should be placed to a depth of 15 cm and not adjacent to cold feces. A simultaneous esophageal probe should be placed 24 cm below the larynx; it may read falsely high during heated inhalation therapy. Relying solely on infrared tympanic thermography is not advisable.

After a diagnosis of hypothermia is established, cardiac monitoring should be instituted, along with attempts to limit further heat loss. If the patient is in ventricular fibrillation, it is unclear at what core temperature ventricular defibrillation (2 J/kg) should first be attempted. One attempt below 30°C is warranted. Further defibrillation attempts should be deferred until some rewarming (1°–2°C) is achieved and ventricular fibrillation is coarser. Although cardiac pacing for hypothermic bradydysrhythmias is rarely indicated, the transthoracic technique is preferable.

Supplemental oxygenation is always warranted, since tissue oxygenation is affected adversely by the leftward shift of the oxyhemoglobin dissociation curve. Pulse oximetry may be unreliable in patients with vasoconstriction. If protective airway reflexes are absent, gentle endotracheal intubation should be performed. Adequate preoxygenation will prevent ventricular arrhythmias.

Insertion of a gastric tube prevents dilation secondary to decreased bowel motility. Indwelling bladder catheters facilitate monitoring of cold-induced diuresis. Dehydration is encountered commonly with chronic hypothermia, and most patients benefit from an intravenous or intraosseous bolus of crystalloid. Normal saline is preferable to lactated Ringer’s solution, as the liver in hypothermic patients inefficiently metabolizes lactate. The placement of a pulmonary

artery catheter can cause perforation of the less compliant pulmonary artery. Insertion of a central venous catheter deeply into the cold right atrium should be avoided since this procedure, similar to transvenous pacing, can precipitate arrhythmias.

Arterial blood gases should not be corrected for temperature **(Chap. 51)**. An uncorrected pH of 7.42 and a P_{CO_2} of 40 mmHg reflect appropriate alveolar ventilation and acid-base balance at any core temperature. Acid-base imbalances should be corrected gradually, since the bicarbonate buffering system is inefficient. A common error is overzealous hyperventilation in the setting of depressed CO_2 production. When the P_{CO_2} decreases by 10 mmHg at 28°C (82°F), it doubles the pH increase of 0.08 that occurs at 37°C (99°F).

The severity of anemia may be underestimated because the hematocrit increases 2% for each 1°C drop in temperature. White blood cell sequestration and bone marrow suppression are common, potentially masking an infection. Although hypokalemia is more common in chronic hypothermia, hyperkalemia also occurs; the expected electrocardiographic changes can be obscured by hypothermia. Patients with renal insufficiency, metabolic acidoses, or rhabdomyolysis are at greatest risk for electrolyte disturbances.

Coagulopathies are common because cold inhibits the enzymatic reactions required for activation of the intrinsic cascade. In addition, thromboxane B_2 production by platelets is temperature dependent, and platelet function is impaired. The administration of platelets and fresh-frozen plasma is therefore not effective. The prothrombin or partial thromboplastin times or the international normalized ratio can be deceptively normal and contrast with the observed in vivo coagulopathy. This contradiction occurs because all coagulation tests are routinely performed at 37°C (99°F), and the enzymes are thus rewarmed.

REWARMING STRATEGIES

The key initial decision is whether to rewarm the patient passively or actively. *Passive external rewarming* simply involves covering and insulating the patient in a warm environment. With the head also covered, the rate of rewarming is usually 0.5°–2°C (1.10°–4.4°F) per hour. This technique is ideal for previously healthy patients who develop acute, mild primary accidental hypothermia. The patient must have sufficient glycogen to support endogenous thermogenesis.

The application of heat directly to the extremities of patients with chronic severe hypothermia should be avoided because it can induce peripheral vasodilation and precipitate core temperature “afterdrop,” a response characterized by a continual decline in the core temperature after removal of the patient from the cold. Truncal heat application reduces the risk of afterdrop.

Active rewarming is necessary under the following circumstances: core temperature <32°C (<90°F) (*poikilothermia*), cardiovascular instability, age extremes, CNS dysfunction, hormone insufficiency, and suspicion of secondary hypothermia. *Active external rewarming* is best accomplished with forced-air heating blankets. Other options include devices that circulate water through external heat exchange pads, radiant heat sources, and hot packs. Monitoring a patient with hypothermia in a heated tub is extremely difficult. Electric blankets should be avoided because vasoconstricted skin is easily burned.

There are numerous widely available options for *active core rewarming*. Airway rewarming with heated humidified oxygen (40°–45°C [104°–113°F]) via mask or endotracheal tube is a convenient option. Although airway rewarming

provides less heat than do some other forms of active core rewarming, it eliminates respiratory heat loss and adds 1°–2°C (2.2°–4.4°F) to the overall rewarming rate. Crystalloids should be heated to 40°–42°C (104°–108°F), but the quantity of heat provided is significant only during massive volume resuscitation. The most efficient method for heating and delivering fluid or blood is with a countercurrent in-line heat exchanger. Heated irrigation of the gastrointestinal tract or bladder transfers minimal heat because of the limited available surface area. These methods should be reserved for patients in cardiac arrest and then used in combination with all available active rewarming techniques.

Closed thoracic lavage is far more efficient in severely hypothermic patients with cardiac arrest. The hemithoraxes are irrigated through two inserted large-bore thoracostomy tubes. Thoracostomy tubes should not be placed in the left chest of a spontaneously perfusing patient for purposes of rewarming. Peritoneal lavage with the dialysate at 40°–45°C (104°–113°F) efficiently transfers heat when delivered through two catheters with outflow suction. Like peritoneal dialysis, standard hemodialysis is especially useful for patients with electrolyte abnormalities, rhabdomyolysis, or toxin ingestion. Another option involves the central venous insertion of a rapid endovascular warming device.

Extracorporeal blood rewarming options **(Table 454-3)** should be considered in severely hypothermic patients, especially those with *primary accidental hypothermia*. Cardiopulmonary bypass should be considered in nonperfusing patients without documented contraindications to resuscitation. Circulatory support may be the only effective option in patients with completely frozen extremities or those with significant tissue destruction coupled with rhabdomyolysis. There is no evidence that extremely rapid rewarming improves survival in perfusing patients. The best strategy is usually a combination of passive, truncal active, and active core rewarming techniques.

TABLE 454-3

Options for Extracorporeal Blood Rewarming

Extracorporeal Rewarming Technique	Considerations
Continuous venovenous (CVV) rewarming	Circuit: CV catheter to CV, dual-lumen CV, or peripheral catheter No oxygenator/circulatory support Flow rates 150–400 mL/min ROR 2°–3°C (4.4°–6.6°F)/h
Hemodialysis	Circuit: single- or dual-vessel cannulation
	Stabilizes electrolyte or toxicologic abnormalities
	Exchange cycle volumes 200–500 mL/min
	ROR 2°–3°C (4.4°– 6.6°F)/h
Continuous arteriovenous rewarming (CAVR)	Circuit: percutaneous 8.5-Fr femoral catheters Requires systolic blood pressure of 60 mmHg No perfusionist/pump/anticoagulation Flow rates 225–375 mL/min ROR 3°–4°C (6.6°–8.8°F)/h
Cardiopulmonary bypass (CPB)	Circuit: full circulatory support with pump and oxygenator
	Perfusate-temperature gradient (5°–10°C [41°–50°F])
	Flow rates 2–7 L/min (average 3–4 L/min)
	ROR up to 9.5°C (20.9°F)/h
Extracorporeal membrane oxygenation (ECMO)	Decreased risk of post-rewarming cardiorespiratory failure

Abbreviations: CV, central venous; ROR, rate of rewarming.

TREATMENT

TREATMENT

Hypothermia

When a patient is hypothermic, target organs and the cardiovascular system respond minimally to most medications. Generally, IV medications are withheld below 30°C (86°F). In contrast to antiarrhythmics, low-dose vasopressor

medications may improve the intra-arrest rates of return of spontaneous circulation. Because of increased binding of drugs to proteins as well as impaired metabolism and excretion, either a lower dose or a longer interval between doses should be used to avoid toxicity. As an example, the administration of repeated doses of digoxin or insulin would be ineffective while the patient is hypothermic, and the residual drugs would be potentially toxic during rewarming.

Achieving a mean arterial pressure of at least 60 mmHg should be an early objective. If the hypotension does not respond to crystalloid/colloid infusion and rewarming, low-dose dopamine support (2–5 μg/kg per min) should be considered. Perfusion of the vasoconstricted cardiovascular system also may be improved with low-dose IV nitroglycerin.

Atrial arrhythmias should be monitored initially without intervention, as the ventricular response should be slow and, unless preexistent, most will convert spontaneously during rewarming. The role of prophylaxis and treatment of ventricular arrhythmias is problematic. Preexisting ventricular ectopy may be suppressed by hypothermia and reappear during rewarming. None of the class I agents has proved to be safe and efficacious. Initiating empirical therapy for adrenal insufficiency usually is not warranted unless the history suggests steroid dependence or hypoadrenalism or efforts to rewarm with standard therapy fail. The administration of parenteral levothyroxine to euthyroid patients with hypothermia, however, is potentially hazardous. Because laboratory results can be delayed and confounded by the presence of the sick euthyroid syndrome **(Chap. 375)**, historic clues or physical findings suggestive of hypothyroidism should be sought. When myxedema is the cause of hypothermia, the relaxation phase of the Achilles reflex is prolonged more than is the contraction phase.

Hypothermia obscures most of the symptoms and signs of infection, notably fever and leukocytosis. Shaking rigors from infection may be mistaken for shivering. Except in mild cases, extensive cultures and repeated physical examinations are essential. Unless an infectious source is identified, empirical antibiotic prophylaxis is most warranted in the elderly, neonates, and immunocompromised patients.

Preventive measures should be discussed with high-risk individuals, such as the elderly and people whose work frequently exposes them to extreme cold. The importance of layered clothing and headgear, adequate shelter, increased caloric intake, and the avoidance of ethanol should be emphasized, along with access to rescue services.

FROSTBITE

Peripheral cold injuries include both freezing and nonfreezing injuries to tissue. Tissue freezes quickly when in contact with thermal conductors such as metal and volatile solutions. Other predisposing factors include constrictive clothing or boots, immobility, and vasoconstrictive medications. Frostbite occurs when the tissue temperature drops below 0°C (32°F). Ice-crystal formation subsequently distorts and destroys the cellular architecture. Once the vascular endothelium is damaged, stasis progresses rapidly to microvascular thrombosis. After the tissue thaws, there is progressive dermal ischemia. The microvasculature begins to collapse, arteriovenous shunting increases tissue pressures, and edema forms. Finally, thrombosis, ischemia, and superficial necrosis appear. The development of mummification and demarcation may take weeks to months.

CLINICAL PRESENTATION

The initial presentation of frostbite can be deceptively benign. The symptoms always include a sensory deficiency affecting light touch, pain, or temperature perception. The acral areas and distal extremities are the most common

insensate areas. Some patients describe a clumsy or “chunk of wood” sensation in the extremity.

Deep frostbitten tissue can appear waxy, mottled, yellow, or violaceous-white. Favorable presenting signs include some warmth or sensation with normal color. The injury is often superficial if the subcutaneous tissue is pliable or if the dermis can be rolled over bony prominences.

Frostnip may precede frostbite. Frostnip is a nonfreezing cold injury resulting from intense vasoconstriction of exposed acral skin.

Clinically, frostbite is superficial or deep. Superficial frostbite does not entail tissue loss but rather causes only anesthesia and erythema. The appearance of vesiculation surrounded by edema and erythema implies deeper involvement **(Fig. 454-1)**. Hemorrhagic vesicles reflect a serious injury to the microvasculature and indicate severe frostbite. Damages in subcuticular, muscular, or osseous tissues may result in amputation. An alternative classification establishes grades based on the location of presenting cyanosis; that is Grade 1, absence of cyanosis; Grade 2, cyanosis on the distal phalanx; Grade 3, cyanosis up to the MP joint; and Grade 4 cyanosis proximal to the MP joint.

FIGURE 454-1

Frostbite with vesiculation, surrounded by edema and erythema.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The two most common nonfreezing peripheral cold injuries are *chilblain* (*pernio*) and *immersion* (*trench*) *foot*. Chilblain results from neuronal and endothelial damage induced by repetitive exposure to damp cold above the freezing point. Young females, particularly those with a history of Raynaud's phenomenon, are at greatest risk. Persistent vasospasticity and vasculitis can cause erythema, mild edema, and pruritus. Eventually plaques, blue nodules, and ulcerations develop. These lesions typically involve the dorsa of the hands and feet. In contrast, immersion foot results from repetitive exposure to wet cold above the freezing point. The feet initially appear cyanotic, cold, and edematous. The subsequent development of bullae is often indistinguishable from frostbite. This vesiculation rapidly progresses to ulceration and liquefaction gangrene. Patients with milder cases report hyperhidrosis, cold sensitivity, and painful ambulation for many years.

TREATMENT

TREATMENT

Peripheral Cold Injuries

When frostbite accompanies hypothermia, hydration may improve vascular stasis. Frozen tissue should be thawed rapidly and completely by immersion in circulating water at 37°–40°C (99°–104°F) for 30–60 min. Rapid rewarming often produces an initial hyperemia. The early formation of large clear distal blebs is more favorable than that of smaller proximal dark hemorrhagic blebs. A common error is the premature termination of thawing, since the reestablishment of perfusion is intensely painful. Parenteral narcotics will be necessary with deep frostbite. If cyanosis persists after rewarming, the tissue compartment pressures should be monitored carefully.

Many antithrombotic and vasodilatory primary and adjunctive treatment regimens have been evaluated. The prostacyclin analogue iloprost given within 48 h after rewarming may prove useful. There is no conclusive evidence that sympathectomy, steroids, calcium channel blockers, or hyperbaric oxygen salvages tissue.

Patients who have deep frostbite injuries with the potential for significant morbidity should be considered for intravenous or intraarterial thrombolytic therapy. Angiography or pyrophosphate scanning may help evaluate the injury and monitor the progress of tissue plasminogen activator therapy (rt-PA). Heparin is recommended as

adjunctive therapy. Intraarterial thrombolysis may reduce the need for digital and more proximal amputations when administered within 24 h of severe injuries. A treatment protocol for frostbite is summarized in Table 454-4.

Unless infection develops, any decision regarding debridement or amputation should generally be deferred. Angiography or technetium-99 bone scan may assist in the determination of surgical margins. Magnetic resonance angiography may also demonstrate the line of demarcation earlier than does clinical demarcation.

The most common symptomatic sequelae reflect neuronal injury and persistently abnormal sympathetic tone, including paresthesias, thermal misperception, and hyperhidrosis. Delayed findings include nail deformities, cutaneous carcinomas, and epiphyseal damage in children.

Management of the chilblain syndrome is usually supportive. With refractory perniosis, alternatives include nifedipine, steroids, and limaprost, a prostaglandin E_1 analogue.

TABLE 454-4

Treatment for Frostbite

Before Thawing	During Thawing	After Thawing
Remove from environment.	Consider parenteral analgesia and ketorolac.	Gently dry and protect part; elevate; place pledgets between toes, if macerated.
Prevent partial thawing and refreezing.	Administer ibuprofen (400 mg PO).	If clear vesicles are intact, aspirate sterilely; if broken, debride and dress with antibiotic or sterile aloe vera ointment.
Stabilize core temperature and treat hypothermia.	Immerse part in 37°–40°C (99°–104°F) (thermometer-monitored) circulating water containing an antiseptic soap until distal flush (10–45 min).	Leave hemorrhagic vesicles intact to prevent desiccation and infection.
Protect frozen part—no friction or massage.	Encourage patient to gently move part.	Continue ibuprofen (400– 600 mg PO [12 mg/kg per day] q8 to 12h).
Address medical or surgical conditions.	If pain is refractory, reduce water temperature to 35°–37°C (95°–99°F) and administer parenteral narcotics.	Consider tetanus and streptococcal prophylaxis; elevate part. Administer hydrotherapy at 37°C (99°F). Consider dextran or phenoxybenzamine or, in severe cases, thrombolysis rt-PA (IV or intraarterial)

FURTHER READING

Brandstrom H et al: Accidental cold-related injury leading to hospitalization in northern Sweden: An eight-year retrospective analysis. Scan J Trauma Resusc Emerg Med 22:6, 2014.

Cauchy E et al: A controlled trial of a prostacyclin and rt-PA in the treatment of severe frostbite. N Engl J Med 364:189, 2011. [PubMed: 21226604]

McIntosh SE et al: Wilderness Medical Society practice guidelines for the prevention and treatment of frostbite. Wilderness Environ Med 25(4 suppl):S43, 2014. [PubMed: 25498262]

Petrone P et al: Management of accidental hypothermia and cold injury. [Review]. Curr Prob Surg 51:417, 2014.

Zafren K et al: Wilderness Medical Society practice guidelines for the out-of-hospital evaluation and treatment of accidental hypothermia. Wilderness Environ Med 25:425, 2014. [PubMed: 25443771]

Chapter 455: Heat-Related Illnesses

Daniel F. Danzl

INTRODUCTION

Heat-related illnesses include a spectrum of disorders ranging from heat syncope, muscle cramps, and heat exhaustion to medical emergencies such as heatstroke. The core body temperature is normally maintained within a very narrow range. Although significant levels of hypothermia are tolerated **(Chap. 454)**, multiorgan dysfunction occurs rapidly at temperatures >41°–43°C. In contrast to heatstroke, the far more common sign of fever reflects intact thermoregulation.

THERMOREGULATION

Humans are capable of significant heat generation. Strenuous exercise can increase heat generation twentyfold. The heat load from metabolic heat production and environmental heat absorption is balanced by a variety of heat dissipation mechanisms. These central integrative dissipation pathways are orchestrated by the central thermostat, which is located in the preoptic nucleus of the anterior hypothalamus. Efferent signals sent via the autonomic nervous system trigger cutaneous vasodilation and diaphoresis to facilitate heat loss.

Normally, the body dissipates heat into the environment via four mechanisms. The *evaporation* of skin moisture is the single most efficient mechanism of heat loss but becomes progressively ineffective as the relative humidity rises to >70%. The *radiation* of infrared electromagnetic energy directly into the surrounding environment occurs continuously. (Conversely, radiation is a major source of heat gain in hot climates.) *Conduction*—the direct transfer of heat to a cooler object—and *convection*—the loss of heat to air currents—become ineffective when the environmental temperature exceeds the skin temperature.

Factors that interfere with the evaporation of diaphoresis significantly increase the risk of heat illness. Examples include dripping of sweat off the skin, constrictive or occlusive clothing, dehydration, and excessive humidity. While air is an effective insulator, the thermal conductivity of water is 25 times greater than that of air at the same temperature. The *wet-bulb globe temperature* is a commonly used index to assess the environmental heat load. This calculation considers the ambient air temperature, the relative humidity, and the degree of radiant heat.

The regulation of this heat load is complex and involves the central nervous system (CNS), thermosensors, and thermoregulatory effectors. The central thermostat activates the effectors that produce peripheral vasodilation and sweating. The skin surface is in effect the radiator and the principal location of heat loss, since skin blood flow can increase 25–30 times over the basal rate. This dramatic increase in skin blood flow, coupled with the maintenance of peripheral vasodilation, efficiently radiates heat. At the same time, there is a compensatory vasoconstriction of the splanchnic and renal beds.

Acclimatization to heat reflects a constellation of physiologic adaptations that permit the body to lose heat more efficiently. This process often requires one to several weeks of exposure and work in a hot environment. During acclimatization, the thermoregulatory set point is altered, and this alteration affects the onset, volume, and content of diaphoresis. The threshold for the initiation of sweating is lowered, and the amount of sweat increases, with a lowered salt concentration. Sweating rates can be 1–2 L/h in acclimated individuals during heat stress. Plasma volume expansion also occurs and improves cutaneous vascular flow. The heart rate lowers, with a higher stroke volume. After the individual leaves the hot environment, improved tolerance to heat stress dissipates rapidly, the plasma volume decreases, and de-acclimatization occurs within weeks.

PREDISPOSING FACTORS AND DIFFERENTIAL DIAGNOSIS

When there is an excessive heat load, unacclimated individuals can develop a variety of heat-related illnesses. Heat waves exacerbate the mortality rate, particularly among the elderly and among persons lacking adequate nutrition and access to air-conditioned environments. Secondary vascular events, including cerebrovascular accidents and myocardial infarctions, occur at least 10 times more often in conditions of extreme heat.

Exertional heat illness continues to occur when laborers, military personnel, or athletes exercise strenuously in the heat. A variety of common factors predispose to heat illness. In addition to the very young and very old, preadolescents and teenagers are at risk since they may use poor judgment when vigorously exercising in high humidity and heat. Other risk factors include obesity, poor conditioning and lack of acclimatization, and mild dehydration.

Cardiovascular inefficiency is a common feature of heat illness. Any physiologic or pharmacologic impediment to cutaneous perfusion will impair heat loss. Many patients are unaware of the heat risk associated with their medications. Anticholinergic agents impair sweating and blunt the normal cardiovascular response to heat. Phenothiazines and heterocyclic antidepressants also have anticholinergic properties that interfere with the function of the preoptic nucleus of the anterior hypothalamus due to central depletion of dopamine.

Calcium channel blockers, beta blockers, and various stimulants also inhibit sweating by reducing peripheral blood flow. To maintain the mean arterial blood pressure, increased cardiac output must be capable of compensating for progressive dehydration. A variety of stimulants and substances of abuse also increase muscle activity and heat production. Careful consideration of the differential diagnosis is important in the evaluation of a patient for a potential heat-related illness. The clinical setting may suggest other etiologies,

such as malignant hyperthermia after general anesthesia. Neuroleptic malignant syndrome can be triggered by certain antipsychotic medications, including selective serotonin reuptake inhibitors. A variety of infectious and endocrine disorders as well as conditions with toxicologic or CNS etiologies may initially mimic heatstroke **(Table 455-1)**.

TABLE 455-1

Heat-Related Illness: Predisposing Factors and Differential Diagnosis

Illness	Predisposing Factors
Cardiovascular inefficiency	Age extremes Beta/calcium channel blockade Congestive heart failure Dehydration Diuresis Obesity Poor physical fitness
Central nervous system illness	Cerebral hemorrhage Hypothalamic cerebrovascular accident Psychiatric disorders Status epilepticus
Impaired heat loss	Antihistamines Heterocyclic antidepressants Occlusive clothing Skin abnormalities
Endocrine-related illness	Diabetic ketoacidosis Pheochromocytoma Thyroid storm
Excessive heat load	Environmental conditions Exertion Fever Hypermetabolic state Lack of acclimatization
Infectious illness	Cerebral abscess Encephalitis Malaria Meningitis Sepsis syndrome Tetanus Typhoid

Illness	Predisposing Factors
Toxicologic illness	Amphetamines Anticholinergic toxidrome Cocaine Dietary supplements Hallucinogens Malignant hyperthermia Neuroleptic malignant syndrome Salicylates Serotonin syndrome Strychnine Sympathomimetics Withdrawal syndromes (ethanol, hypnotics)

MINOR HEAT-EMERGENCY SYNDROMES

Heat edema is characterized by mild swelling of the hands, feet, and ankles during the first few days of significant heat exposure. The principal mechanism involves cutaneous vasodilation and pooling of interstitial fluid in response to heat stress. Heat also increases the secretion of antidiuretic hormone and aldosterone. Systemic causes of edema, including cirrhosis, nephrotic syndrome, and congestive heart failure, can usually be excluded by the history and physical examination. Heat edema generally resolves without treatment in several days. Simple leg elevation or thigh-high support hose will usually suffice. Diuretics are *not* effective and, in fact, predispose to volume depletion and the development of more serious heat-related illnesses.

Prickly heat (*miliaria rubra*, *lichen tropicus*) is a maculopapular, pruritic, erythematous rash that commonly occurs in clothed areas. Blockage of the sweat pores by debris from macerated stratum corneum causes inflammation in the sweat ducts. As the ducts dilate, they rupture and produce superficial vesicles. The predominant symptom is pruritus. In addition to antihistamines, chlorhexidine in a light cream or lotion provides some relief. In adults, localized areas may benefit from 1% salicylic acid TID, with caution taken to avoid salicylate intoxication. Clothing with breathable fabric should be clean and loose fitting, and activities or environments that induce diaphoresis should be avoided.

Heat syncope (exercise-associated collapse) can follow endurance exercise or occur in the elderly. Other common clinical scenarios include prolonged standing while stationary in the heat and sudden standing after prolonged exposure to heat. Heat stress routinely causes relative volume depletion, decreased vasomotor tone, and peripheral vasodilation. The cumulative effect of this decrease in venous return is postural hypotension, especially in nonacclimated elderly individuals. Many of those affected also have

comorbidities. Therefore, other cardiovascular, neurologic, and metabolic causes of syncope should be considered. After removal from the heat source, most patients will recover promptly with cooling and rehydration.

Hyperventilation tetany occurs in some individuals when exposure to heat stimulates hyperventilation, producing respiratory alkalosis, paresthesias, and carpopedal spasm. Unlike heat cramps, heat tetany causes very little muscle-compartment pain. Treatment includes providing reassurance, moving the patient out of the heat, and addressing the hyperventilation.

HEAT CRAMPS

Heat cramps (exercise-associated muscle cramps) are intermittent, painful, and involuntary spasmodic contractions of skeletal muscles. They typically occur in an unacclimated individual who is at rest after vigorous exertion in a humid, hot environment. In contrast, cramps that occur in athletes during exercise last longer, are relieved by stretching and massage, and resolve spontaneously.

Of note, not all muscle cramps are related to exercise, and the differential diagnosis includes many other disorders. A variety of medications, myopathies, endocrine disorders, and sickle cell trait are other possible causes.

The typical patient with heat cramps is usually profusely diaphoretic and has been replacing fluid losses with copious water or other hypotonic fluids. Roofers, firefighters, military personnel, athletes, steel workers, and field workers are commonly affected. Other predisposing factors include insufficient sodium intake before intense activity in the heat and lack of heat acclimatization, resulting in sweat with a high salt concentration.

The precise pathogenesis of heat cramps appears to involve a relative deficiency of sodium, potassium, and fluid at the intracellular level. Coupled with copious hypotonic fluid ingestion, large amounts of sodium in the diaphoresis cause hyponatremia and hypochloremia, resulting in muscle cramps due to calcium-dependent muscle relaxation. Total-body depletion of potassium may be observed during the period of heat acclimatization. Rhabdomyolysis is very rare with routine exercise-associated muscle cramps.

Heat cramps that are not accompanied by significant dehydration can be treated with commercially available electrolyte solutions. Although the flavored electrolyte solutions are far more palatable, two 650-mg salt tablets dissolved in 1 quart of water produce a 0.1% saline solution. Individuals should avoid the ingestion of undissolved salt tablets, which are a gastric irritant and may induce vomiting.

HEAT EXHAUSTION

The physiologic hallmarks of heat exhaustion—in contrast to heatstroke—are the maintenance of thermoregulatory control and CNS function. The core temperature is usually elevated but is generally <40.5°C (<105°F). The two physiologic precipitants are water depletion and sodium depletion, which often occur in combination. Laborers, athletes, and elderly individuals exerting themselves in hot environments, without

adequate fluid intake, tend to develop *water-depletion heat exhaustion*. Persons working in the heat frequently consume only two-thirds of their net water loss and are voluntarily dehydrated. In contrast, *salt-depletion heat exhaustion* occurs more slowly in unacclimated persons who have been consuming large quantities of hypotonic solutions.

Heat exhaustion is usually a diagnosis of exclusion because of the multitude of nonspecific symptoms. If any signs of heatstroke are present, rapid cooling and crystalloid resuscitation should be initiated immediately during stabilization and evaluation. Mild neurologic and gastrointestinal influenza-like symptoms are common. These symptoms may include headache, vertigo, ataxia, impaired judgment, malaise, dizziness, nausea, and muscle cramps. Orthostatic hypotension and sinus tachycardia develop frequently. More significant CNS impairment suggests heatstroke or other infectious, neurologic, or toxicologic diagnoses.

Hemoconcentration does not always develop, and rapid infusion of isotonic IV fluids should be guided by frequent electrolyte determinations and perfusion requirements. Most cases of heat exhaustion reflect mixed sodium and water depletion. Sodium-depletion heat exhaustion is characterized by hyponatremia and hypochloremia. Hepatic aminotransferases are mildly elevated in both types of heat exhaustion. Urinary sodium and chloride concentrations are usually low.

Some patients with heat exhaustion develop heatstroke after removal from the heat-stress environment. Aggressive cooling of nonresponders is indicated until their core temperature is 39°C (102.2°F). Except in mild cases, free water deficits should be replaced slowly over 24–48 h to avoid a decrease of serum osmolality by >2 mOsm/h.

The disposition of younger, previously healthy heat-exhaustion patients who have no major laboratory abnormalities may include hospital observation and discharge after IV rehydration. Older patients with comorbidities (including cardiovascular disease) or predisposing factors often require inpatient fluid and electrolyte replacement, monitoring, and reassessment.

HEATSTROKE

The clinical manifestations of heatstroke reflect a total loss of thermoregulatory function. Typical vital-sign abnormalities include tachypnea, various tachycardias, hypotension, and a widened pulse pressure. Although there is no single specific diagnostic test, the historical and physical triad of exposure to a heat stress, CNS dysfunction, and a core temperature >40.5°C helps establish the preliminary diagnosis. Some patients with impending heat stroke will initially appear lucid. The definitive diagnosis should be reserved until the other potential causes of hyperthermia are excluded. Many of the usual laboratory abnormalities seen with heatstroke overlap with other conditions. If the patient’s mental status does not improve with cooling, toxicologic screening may be indicated, and cranial CT and spinal fluid analysis can be considered.

The premonitory clinical characteristics may be nonspecific and include weakness, dizziness, disorientation, ataxia, and gastrointestinal or psychiatric symptoms. These prodromal symptoms often resemble heat exhaustion. The sudden onset of heatstroke occurs when the maintenance of adequate perfusion requires

peripheral vasoconstriction to stabilize the mean arterial blood pressure. As a result, the cutaneous radiation of heat ceases. At this juncture, the core temperature rises dramatically. Since many patients with heatstroke also meet the criteria for systemic inflammatory response syndrome and have a broad differential diagnosis, rapid cooling is essential during the extensive diagnostic evaluation (Table 455-1).

There are two forms of heatstroke with significantly different manifestations **(Table 455-2)**. Classic (epidemic) heatstroke (CHS) usually occurs during long periods of high ambient temperature and humidity, as during summer heat waves. Patients with CHS commonly have chronic diseases that predispose to heat-related illness, and they may have limited access to oral fluids. Heat dissipation mechanisms are overwhelmed by both endogenous heat production and exogenous heat stress. Patients with CHS are often compliant with prescribed medications that can impair tolerance to a heat stress. In many of these dehydrated CHS patients, sweating has ceased and the skin is hot and dry.

TABLE 455-2

Typical Manifestations of Heatstroke

Classic	Exertional
Older patient	Younger patient
Predisposing health factors/medications	Healthy condition
Epidemiology (heat waves)	Sporadic cases
Sedentary	Exercising
Anhidrosis (possible)	Diaphoresis (common)
Central nervous system dysfunction	Myocardial/hepatic injury
Oliguria	Acute renal failure
Coagulopathy (mild)	Disseminated intravascular coagulation
Mild lactic acidosis	Marked lactic acidosis
Mild creatine kinase elevation	Rhabdomyolysis
Normoglycemia/calcemia	Hypoglycemia/calcemia
Normokalemia	Hyperkalemia
Normonatremia	Hyponatremia

If cooling is delayed, severe hepatic dysfunction, renal failure, disseminated intravascular coagulation, and fulminant multisystem organ failure may occur. Hepatocytes are very heat sensitive. On presentation, the serum level of aspartate aminotransferase (AST) is routinely elevated. Eventually, levels of both AST and alanine aminotransferase (ALT) often increase to >100 times the normal values. Coagulation studies commonly demonstrate decreased platelets, fibrinogen, and prothrombin. Most patients with CHS require cautious crystalloid resuscitation, electrolyte monitoring, and—in certain refractory cases—consideration of central venous pressure (CVP) measurements. Hypernatremia is secondary to dehydration in CHS. Many patients exhibit significant stress leukocytosis, even in the absence of infection.

Patients with exertional heatstroke (EHS), in contrast to those with CHS, are often young and previously healthy, and their diagnosis is usually more obvious from the history. Athletes, laborers, and military recruits are common victims. Unlike those with CHS, many EHS patients present profusely diaphoretic despite significant dehydration. As a result of muscular exertion, rhabdomyolysis and acute renal failure are more common in EHS. Studies to detect rhabdomyolysis and its complications, including hypocalcemia and hyperphosphatemia, should be considered. Hyponatremia, hypoglycemia, and coagulopathies are frequent findings. Elevated creatine kinase and lactate dehydrogenase levels also suggest EHS. Oliguria is a common finding. Renal failure can result from direct thermal injury, untreated rhabdomyolysis, or volume depletion. Common urinalysis findings include microscopic hematuria, myoglobinuria, and granular or red cell casts.

With both CHS and EHS, heat-related reversible increases in cardiac biomarker levels are often present. Heatstroke often causes thermal cardiomyopathy. As a result, the CVP may be elevated despite significant dehydration. In addition, the patient often presents with potentially deceptive noncardiogenic pulmonary edema and basilar rales despite being significantly hypovolemic. The electrocardiogram commonly displays a variety of tachyarrhythmias, nonspecific ST-T wave changes, and heat-related ischemia or infarction. Rapid cooling—not the administration of antiarrhythmic medications—is essential.

Above 42°C (107.6°F), heat can rapidly produce direct cellular injury. Thermosensitive enzymes become nonfunctional, and eventually there is irreversible uncoupling of oxidative phosphorylation. The production of heat-shock proteins increases, and cytokines mediate a systemic inflammatory response. The vascular endothelium is also damaged, and this injury activates the coagulation cascade. Significant shunting away from the splanchnic circulation produces gastrointestinal ischemia. Endotoxins further impair normal thermoregulation. As a result, if cooling is delayed, severe hepatic dysfunction, permanent renal failure, disseminated intravascular coagulation, and fulminant multisystem organ failure may occur.

COOLING STRATEGIES

Before cooling is initiated, endotracheal intubation and continuous core-temperature monitoring should be considered. Peripheral methods to measure temperature are *not* reliable. Hypoglycemia is a frequent finding and can be addressed by glucose infusion. Since peripheral vasoconstriction delays heat dissipation, repeated administration of discrete boluses of isotonic crystalloid for hypotension is preferable to the administration of α-adrenergic agonists.

Evaporative cooling is frequently the most practical and effective technique. Rapid cooling is essential in both CHS and EHS, and an immediate improvement in vital signs and mental status may prove valuable for diagnostic purposes. Cool water (15°C [60°F]) is sprayed on the exposed skin while fans direct continuous airflow over the moistened skin. Cold packs applied to the neck, axillae, and groin are useful cooling adjuncts. If cardiac electrodes will not adhere, they can be applied to the patient's back.

Immersion cooling in ice-cold water is an alternative option in EHS but can induce peripheral vasoconstriction and shivering. The initial increase in temperature from peripheral vasoconstriction will rapidly be overcome by the large conductive thermal transfer into cold water. This technique presents

significant monitoring and resuscitation challenges in many clinical settings. The safety of immersion cooling is best established for young, previously healthy patients with EHS (but not for those with CHS). To avoid hypothermic afterdrop (continued cooling after immersion), active cooling should be terminated ~ 38°–39°C (100.4°F–102.2°F).

Cooling with commercially available cooling blankets should not be the sole technique used, since the rate of cooling is far too slow. Other methods are less efficacious and rarely indicated, such as IV infusion of cold fluids and cold irrigation of the bladder or gastrointestinal tract. Cold thoracic and peritoneal lavage are efficient maneuvers but are invasive and rarely necessary. Cardiopulmonary bypass provides effective cooling but is labor intensive and is rarely necessary.

RESUSCITATION

Aspiration commonly occurs in heatstroke, and endotracheal intubation is usually necessary. The metabolic demands are high, and supplemental oxygenation is essential due to hypoxemia induced by thermal stress and pulmonary dysfunction. The oxyhemoglobin dissociation curve is shifted to the right. Pneumonitis, pulmonary infarction, hemorrhage, edema, and acute respiratory distress syndrome occur frequently in heatstroke patients. Seizures are common, and can occur during therapeutic cooling. Cold induced tonic-clonic muscular rigidity mimics seizure activity.

The circulatory fluid requirements, particularly in CHS, may be deceptively modest. Aggressive cooling and modest volume repletion usually elevate the CVP to 12–14 mmHg. The reading, however, may be deceptive. Many patients present with a thermally induced hyperdynamic circulation accompanied by a high cardiac index, low peripheral vascular resistance, and an elevated CVP caused by right-sided heart failure. In contrast, most patients with EHS require far more zealous isotonic crystalloid resuscitation.

The hypotension that is initially common among patients with heatstroke results from both dehydration and high-output cardiac failure caused by peripheral vasodilation. Inotropes causing α-adrenergic stimulation (e.g., norepinephrine) can impede cooling by causing significant vasoconstriction. Vasoactive catecholamines such as dopamine or dobutamine may be necessary if the cardiac output remains depressed despite an elevated CVP, particularly in patients with a hyperdynamic circulation.

A wide variety of tachyarrhythmias are routinely observed on presentation and usually resolve spontaneously during cooling. The administration of atrial or ventricular antiarrhythmic medications is rarely indicated during cooling. Anticholinergic medications (including atropine) inhibit sweating and should be avoided. With a cardiac rhythm that sustains perfusion, electrical cardioversion of the hyperthermic myocardium should be deferred until the myocardium is cooled. Significant shivering, discomfort, or extreme agitation is preferably mitigated with short-acting benzodiazepines, which are ideal due to their renal clearance. On the other hand, chlorpromazine may lower the seizure threshold, has anticholinergic properties, and can exacerbate the hypotension or cause neuroleptic malignant syndrome. With hepatic dysfunction, barbiturates should be avoided and seizures treated with benzodiazepines.

Coagulopathies more commonly occur after the first day of illness. After cooling, the patient should be monitored for disseminated intravascular coagulation, and replacement therapy with fresh-frozen plasma and platelets should be considered.

There is no therapeutic role for antipyretics in the control of environmentally induced hyperthermia; these drugs block the actions of pyrogens at hypothalamic receptor sites. Salicylates can further uncouple oxidative phosphorylation in heatstroke and exacerbate coagulopathies. Acetaminophen may further stress hepatic function. The safety and efficacy of dantrolene is not established. Although aminocaproic acid impedes fibrinolysis, it may cause rhabdomyolysis and is not recommended in heatstroke.

DISPOSITION

Most patients with minor heat-emergency syndromes (including heat edema, heat syncope, and heat cramps) require only stabilization and treatment with outpatient follow-up. Although there are no decision rules to guide disposition choices in heat exhaustion, many of these patients have multiple predisposing factors and comorbidities that will require prolonged observation or hospital admission.

Essentially all patients with actual heatstroke require admission to a monitored setting, and most require intensive care. Many of these patients also require prolonged tracheal intubation, invasive hemodynamic monitoring, and support for various degrees of multiorgan dysfunction syndrome. The prognosis worsens if the initial core temperature exceeds 42°C (107.6°F) or if there was a prolonged period during which the core temperature exceeded this level. Other features of a negative prognosis include acute renal failure, massively elevated liver enzymes, and significant hyperkalemia. As expected, the number of dysfunctional organ systems also correlates directly with mortality risk.

FURTHER READING

Atha WF: Heat-related illness. Emerg Med Clin North Am 31:1097, 2013.

Casa DJ et al: National Athletic Trainers' Association Position Statement: Exertional Heat Illnesses. J Athl Train 50:986, 2015.

Lipman GS et al: Wilderness Medical Society practice guidelines for the prevention and treatment of heat-related illness. Wilderness Environ Med 24:351, 2013. [PubMed: 24140191]

Maron BJ et al: Sudden deaths in young competitive athletes: Analysis of 1866 deaths in the United States, 1980–2006. Circulation 119:1085, 2009. [PubMed: 19221222]

O'Connor FG et al: American College of Sports Medicine Roundtable on exertional heat stroke—Return to duty/return to play: Conference proceedings. Curr Sports Med Rep 9:314, 2010. [PubMed: 20827100]

Platt M et al: Heat illnesses, in *Rosen's Emergency Medicine: Concepts and Clinical Practices*, 9th ed, Marx JA et al (eds). Philadelphia, Elsevier, 2018, pp 1755–1764.

Racinais S et al: Consensus recommendations on training and competing in the heat. Br J Sports Med 49:1164, 2015. [PubMed: 26069301]

Zeller L et al: Exertional heatstroke: Clinical characteristics, diagnostic and therapeutic considerations. Eur J Intern Med 22:296, 2011. [PubMed: 21570651]

Chapter 456: Principles of Human Genetics

J. Larry Jameson; Peter Kopp

IMPACT OF GENETICS AND GENOMICS ON MEDICAL PRACTICE

Human genetics refers to the study of individual genes, their role and function in disease, and their mode of inheritance. Genomics refers to an organism's entire genetic information, the *genome*, and the function and interaction of DNA within the genome, as well as with environmental or nongenetic factors, such as a person's lifestyle. With the characterization of the human genome, genomics not only complements traditional genetics in our efforts to elucidate the etiology and pathogenesis of disease, but it plays an increasingly prominent role in diagnostics, prevention, and therapy **(Chap. 457)**. These transformative developments, emerging from the Human Genome Project, have been variably designated *genomic medicine*, *personalized medicine*, or *precision medicine*. Precision medicine aims at customizing medical decisions to an individual patient. For example, a patient's genetic characteristics (genotype) can be used to optimize drug therapy and predict efficacy, adverse events, and drug dosing of selected medications (*pharmacogenomics*) **(Chap. 64)**. The characterization of the mutational profile of a malignancy allows to identify driver mutations or overexpressed signaling molecules, which then facilitates the selection of targeted therapies. Genomic risk prediction models for common diseases are also beginning to emerge.

Genetics has traditionally been viewed through the window of relatively rare single-gene diseases. These disorders account for ~10% of pediatric admissions and childhood mortality. Historically, genetics has focused predominantly on chromosomal and metabolic disorders, reflecting the long-standing availability of techniques to diagnose these conditions. For example, conditions such as trisomy 21 (Down's syndrome) or monosomy X (Turner's syndrome) can be diagnosed using cytogenetics. Likewise, many metabolic disorders (e.g., phenylketonuria, familial hypercholesterolemia) are diagnosed using biochemical analyses. The advances in DNA diagnostics have extended the field of genetics to include virtually all medical specialties and have led to the elucidation of the pathogenesis of numerous monogenic disorders. In addition, it is apparent that virtually every medical condition has a genetic component. As is often evident from a patient's family history, many common disorders such as hypertension, heart disease, asthma, diabetes mellitus, and mental illnesses are significantly influenced by the genetic background. These polygenic or multifactorial (complex) disorders involve the contributions of many different genes, as well as environmental factors that can modify disease risk. Genome-wide association studies (GWAS) have elucidated numerous disease-associated loci and are providing novel insights into the allelic architecture of complex traits. These studies have been facilitated by the availability of comprehensive catalogues of human single-nucleotide polymorphism (SNP) haplotypes (HapMap, International Genome Sample Resource/1000 genomes project). Next-generation DNA sequencing (NGS) technologies have evolved rapidly and the cost of sequencing whole exomes (the exons within the genome, WES) or genomes (WGS) has plummeted. Comprehensive unbiased sequence analyses are now frequently used to characterize individuals with complex undiagnosed conditions or to determine the mutational profile of advanced malignancies in order to select better targeted therapies.

Cancer has a genetic basis because it results from acquired somatic mutations in genes controlling growth, apoptosis, and cellular differentiation **(Chap. 67)**. In addition, the development of many cancers is associated with a hereditary predisposition. Characterization of the genome (and epigenome) in various malignancies has led to fundamental new insights into cancer biology and reveals that the genomic profile of mutations is in many cases more important in determining the appropriate therapy than the organ in which the tumor originates. The Cancer Genome Atlas (TCGA) initiative of the National Cancer Institute and the National Human Genome Research Institute has already characterized the genomic landscape of >30 malignancies and several others will be completed in the near future. TCGA consists of comprehensive analyses of genomic and proteomic alterations and is providing fundamental new insights into the molecular pathogenesis of cancer. This knowledge has direct clinical ramifications as it impacts cancer taxonomy and the development of targeted therapies.

Genetic and genomic approaches have proven invaluable for the detection of infectious pathogens and are used clinically to identify agents that are difficult to culture such as mycobacteria, viruses, and parasites, or to track infectious agents locally or globally. In many cases, molecular genetics has improved the feasibility and accuracy of diagnostic testing and is beginning to open new avenues for therapy, including gene and cellular therapies **(Chap. 458)**. Molecular genetics has also provided the opportunity to characterize the *microbiome*, a new field that characterizes the population dynamics of bacteria, viruses, and parasites that coexist with humans and other animals **(Chap. 459)**. Emerging data indicate that the microbiome has significant effects on normal physiology as well as various disease states and the field is now focusing on defining the mechanisms underlying these interactions.

Molecular biology has significantly changed the treatment of human disease. Peptide hormones, growth factors, cytokines, and vaccines can now be produced in large amounts using recombinant DNA technology. Targeted modifications of these peptides provide the practitioner with improved therapeutic tools, as illustrated by genetically modified insulin analogues with more favorable kinetics. Lastly, there is reason to believe that a better understanding of the genetic basis of human disease will also have an increasing impact on disease prevention.

The astounding rate at which new genetic and genomic information is being generated creates a major challenge for physicians, health care providers, and basic investigators. Although many functional aspects of the genome remain unknown, there are many clinical situations where sufficient evidence exists

for the use of genetic and genomic information to optimize patient care and treatment. Much genetic information resides in databases that provide easy access to the expanding information about the human genome, genetic disease, and genetic testing **(Table 456-1)**. For example, several thousand monogenic disorders are summarized in a large, continuously evolving compendium, referred to as the *Online Mendelian Inheritance in Man* (OMIM) catalogue (Table 456-1). The constant refinement of bioinformatics and new developments in *big data analytics*, together with the widespread adoption of electronic health records (EHRs), is simplifying the access, analysis and integration of this daunting amount of new information. Importantly, genomic data can be integrated readily into EHRs, and will accelerate the impact on clinical practice.

TABLE 456-1

Selected Databases Relevant for Genomics and Genetic Disorders

Site	URL	Comment
National Center for Biotechnology Information (NCBI)	*http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/*	Broad access to biomedical and genomic information, literature (PubMed), sequence databases, software for analyses of nucleotides and proteins Extensive links to other databases, genome resources, and tutorials
National Human Genome Research Institute	*http://www.genome.gov/*	An institute of the National Institutes of Health focused on genomic and genetic research; links providing information about the human genome sequence, genomes of other organisms, and genomic research
Catalog of Published Genome-Wide Association Studies	*http://www.genome.gov/GWAStudies/*	Published high-resolution genome-wide association studies (GWAS)
Ensembl Genome browser	*http://www.ensembl.org*	Maps and sequence information of eukaryotic genomes
Online Mendelian Inheritance in Man	*http://www-ncbi-nlm-nih-gov.ezproxy1.library.usyd.edu.au/omim*	Online compendium of Mendelian disorders and human genes causing genetic disorders
Office of Biotechnology Activities, National Institutes of Health	*http://oba.od.nih.gov.ezproxy1.library.usyd.edu.au/oba*	Information about recombinant DNA and gene transfer; medical, ethical, legal, and social issues raised by genetic testing; medical, ethical, legal, and social issues raised by xenotransplantation
American College of Medical Genetics and Genomics	*http://www.acmg.net/*	Extensive links to other databases relevant for the diagnosis, treatment, and prevention of genetic disease
American Society of Human Genetics	*http://www.ashg.org*	Information about advances in genetic research, professional and public education, social and scientific policies
The Cancer Genome Atlas	*https://cancergenome-nih-gov.ezproxy1.library.usyd.edu.au/*	Comprehensive, multi-dimensional characterization of the genomic and proteomic landscape of malignancies with high public health impact.
Cancer Genome Anatomy Project (CGAP)	*http://cgap.nci.nih.gov.ezproxy1.library.usyd.edu.au/*	Information about gene expression profiles of normal, precancer, and cancer cells
GeneTests	*http://www.genetests.org/*	International directory of genetic testing laboratories and prenatal diagnosis clinics; reviews and educational materials
Genomes Online Database (GOLD)	*http://www.genomesonline.org/*	Information on published and unpublished genomes
HUGO Gene Nomenclature	*http://www.genenames.org/*	Gene names and symbols
GENECODE	*https://www.gencodegenes.org/*	High quality reference gene annotation and experimental validation for human and mouse genomes
MITOMAP, a human mitochondrial genome database	*http://www.mitomap.org/*	A compendium of polymorphisms and mutations of the human mitochondrial DNA

Site	URL	Comment
The International Genome Sample Resource (IGSR)	*http://www.internationalgenome.org*	Public catalogue of human variation and genotype data from numerous ethnic groups
ENCODE	*http://www.genome.gov/10005107*	*Enc*yclopedia *o*f *D*NA *E*lements; catalogue of all functional elements in the human genome
Dolan DNA Learning Center, Cold Spring Harbor Laboratories	*http://www.dnalc.org/*	Educational material about selected genetic disorders, DNA, eugenics, and genetic origin
The Online Metabolic and Molecular Bases of Inherited Disease (OMMBID)	*http://ommbid.mhmedical.com.ezproxy1.library.usyd.edu.au*	Online version of the comprehensive text on the metabolic and molecular bases of inherited disease
Online Mendelian Inheritance in Animals (OMIA)	*http://omia.angis.org.au/*	Online compendium of Mendelian disorders in animals
The Jackson Laboratory	*http://www.jax.org/*	Information about murine models and the mouse genome
Mouse genome informatics	*http://www.informatics.jax.org*	Mouse genome informatics

Note: Databases are evolving constantly. Pertinent information may be found by using links listed in the few selected databases.

THE HUMAN GENOME

Structure of the Human Genome

The Human Genome Project was initiated in the mid-1980s as an ambitious effort to characterize the entire human genome and culminated in the completion of the DNA sequence for the last of the human chromosomes in 2006. The scope of a whole genome sequence analysis can be illustrated by the following analogy. Human DNA consists of ~3 billion base pairs (bp) of DNA per haploid genome, which is nearly 1000-fold greater than that of the *Escherichia coli* genome. If the human DNA sequence were printed out, it would correspond to about 120 volumes of *Harrison's Principles of Internal Medicine*.

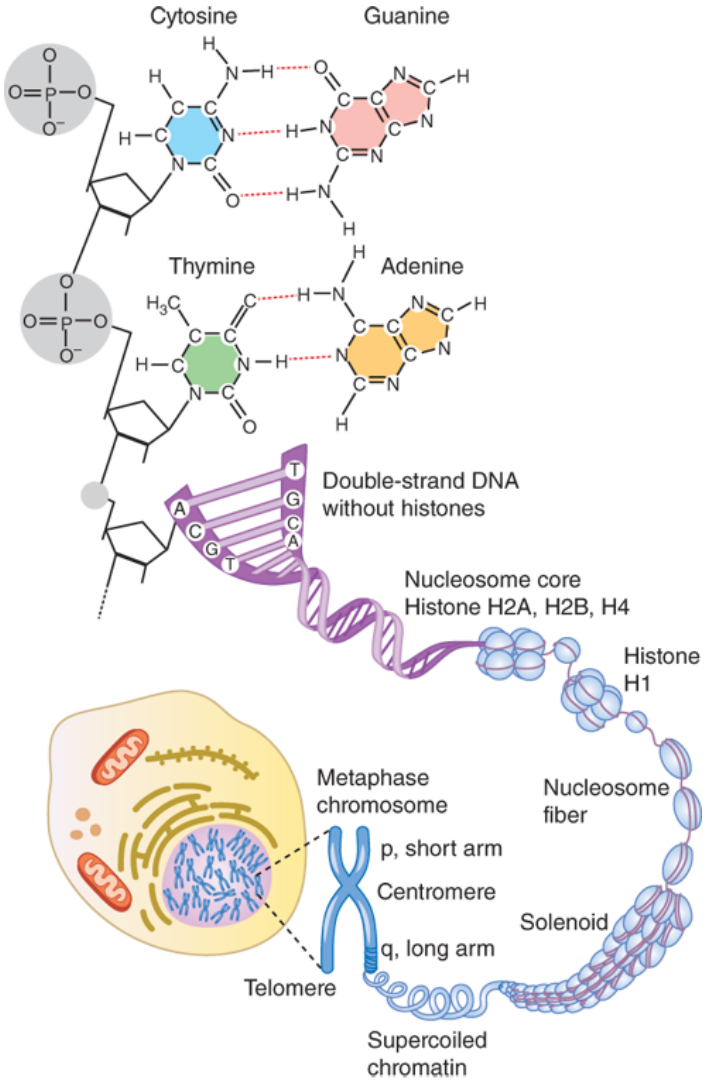
In addition to the human genome, the genomes of numerous organisms have been sequenced completely (~4000) or partially (~10,000) (Genomes Online Database [GOLD]; Table 456-1). They include, among others, eukaryotes such as the mouse (*Mus musculus), Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster*; bacteria (e.g., *E. coli*); and archaea, viruses, organelles (mitochondria, chloroplasts), and plants (e.g., *Arabidopsis thaliana*). Genomic information of infectious agents has significant impact for the characterization of infectious outbreaks and epidemics. Other ramifications arising from the availability of genomic data include, among others, (1) the comparison of entire genomes (*comparative genomics*), (2) the study of large-scale expression of RNAs (*functional genomics*) and proteins (*proteomics*) to detect differences between various tissues in health and disease, (3) the characterization of the variation among individuals by establishing catalogues of sequence variations and SNPs, and (4) the identification of genes that play critical roles in the development of polygenic and multifactorial disorders.

CHROMOSOMES

The human genome is divided into 23 different chromosomes, including 22 autosomes (numbered 1–22) and the X and Y sex chromosomes **(Fig. 456-1)**. Adult cells are *diploid*, meaning they contain two homologous sets of 22 autosomes and a pair of sex chromosomes. Females have two X chromosomes (XX), whereas males have one X and one Y chromosome (XY). As a consequence of meiosis, germ cells (sperm or oocytes) are haploid and contain one set of 22 autosomes and one of the sex chromosomes. At the time of fertilization, the diploid genome is reconstituted by pairing of the homologous chromosomes from the mother and father. With each cell division (mitosis), chromosomes are replicated, paired, segregated, and divided into two daughter cells.

FIGURE 456-1

Structure of chromatin and chromosomes. Chromatin is composed of double-strand DNA that is wrapped around histone and nonhistone proteins forming nucleosomes. The nucleosomes are further organized into solenoid structures. Chromosomes assume their characteristic structure, with short (p) and long (q) arms at the metaphase stage of the cell cycle.



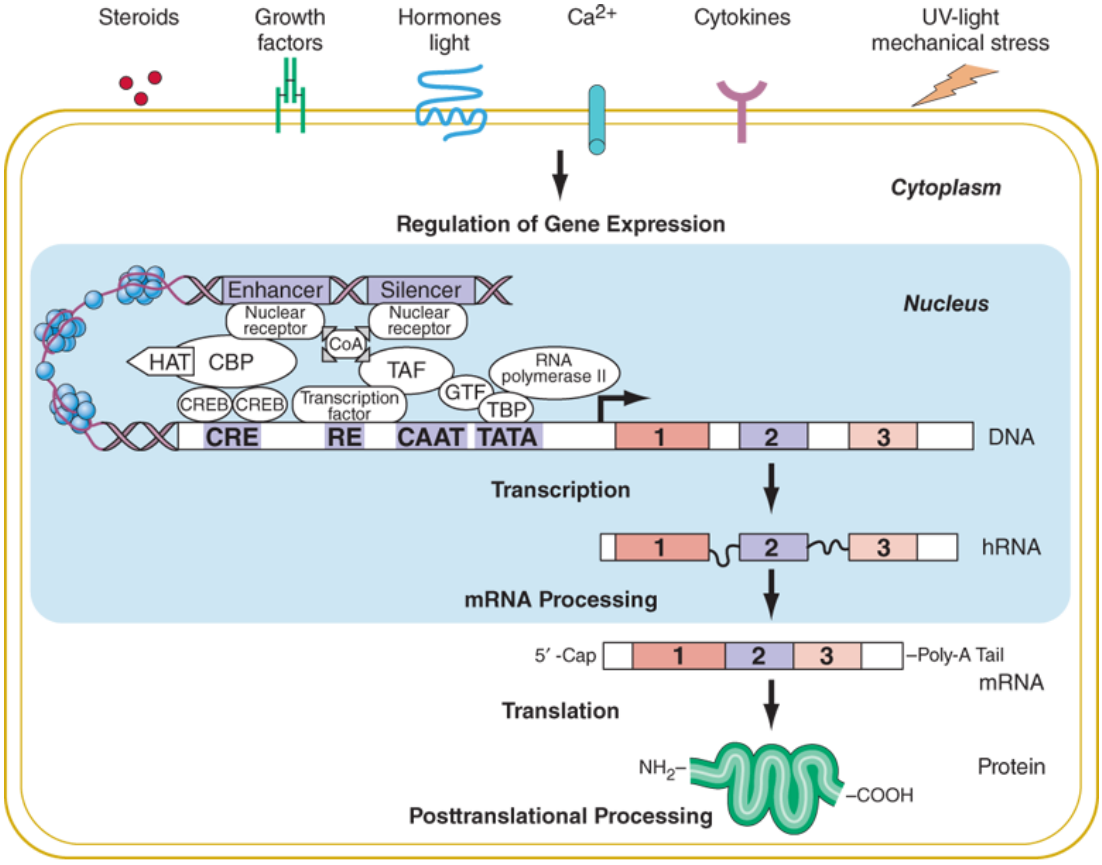
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

STRUCTURE OF DNA

DNA is a double-stranded helix composed of four different bases: adenine (A), thymidine (T), guanine (G), and cytosine (C). Adenine is paired to thymidine, and guanine is paired to cytosine, by hydrogen bond interactions that span the double helix (Fig. 456-1). DNA has several remarkable features that make it ideal for the transmission of genetic information. It is relatively stable, and the double-stranded nature of DNA and its feature of strict base-pair complementarity permit faithful replication during cell division. Complementarity also allows the transmission of genetic information from DNA → RNA → protein **(Fig. 456-2)**. mRNA is encoded by the so-called sense or coding strand of the DNA double helix and is translated into proteins by ribosomes.

FIGURE 456-2

Flow of genetic information. Multiple extracellular signals activate intracellular signal cascades that result in altered regulation of gene expression through the interaction of transcription factors with regulatory regions of genes. RNA polymerase transcribes DNA into RNA that is processed to mRNA by excision of intronic sequences. The mRNA is translated into a polypeptide chain to form the mature protein after undergoing posttranslational processing. CBP, CREB-binding protein; CoA, co-activator; COOH, carboxyterminus; CRE, cyclic AMP responsive element; CREB, cyclic AMP response element–binding protein; GTF, general transcription factors; HAT, histone acetyl transferase; NH2, aminoterminus; RE, response element; TAF, TBP-associated factors; TATA, TATA box; TBP, TATA-binding protein.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The presence of four different bases provides surprising genetic diversity. In the protein-coding regions of genes, the DNA bases are arranged into codons, a triplet of bases that specifies a particular amino acid. It is possible to arrange the four bases into 64 different triplet codons (4^3). Each codon specifies 1 of the 20 different amino acids, or a regulatory signal such as initiation and stop of translation. Because there are more codons than amino acids, the genetic code is degenerate; that is, most amino acids can be specified by several different codons. By arranging the codons in different combinations and in various lengths, it is possible to generate the tremendous diversity of primary protein structure.

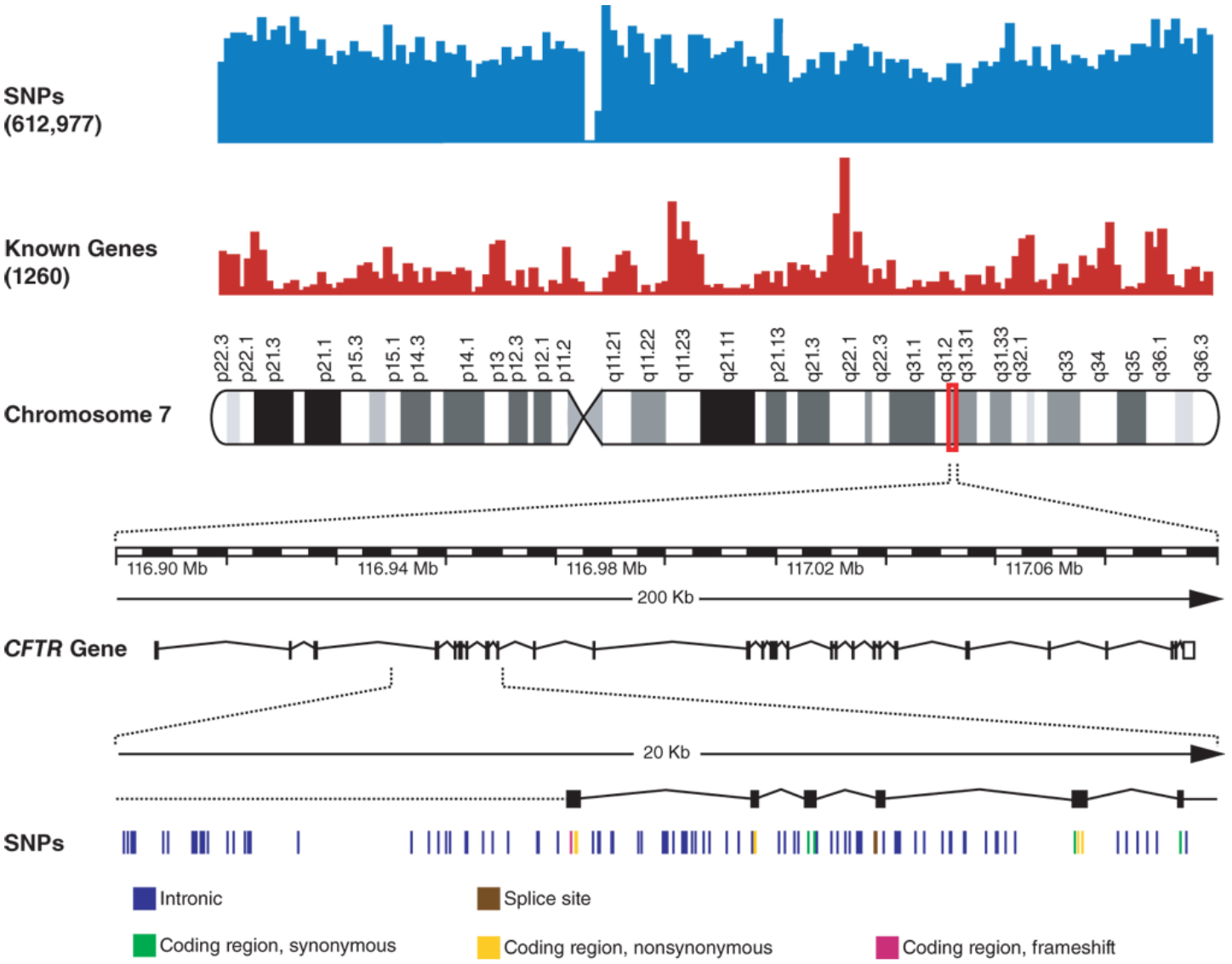
DNA length is normally measured in units of 1000 bp (kilobases, kb) or 1,000,000 bp (megabases, Mb). Not all DNA encodes genes. In fact, genes account for only ~10–15% of DNA. Much of the remaining DNA consists of sequences, often of highly repetitive nature, the function of which remains incompletely understood. These repetitive DNA regions, along with nonrepetitive sequences that do not encode genes, serve, in part, a structural role in the packaging of DNA into chromatin (i.e., DNA bound to histone proteins, and chromosomes) and exert regulatory functions (Fig. 456-1).

GENES

A *gene* is a functional unit that is regulated by transcription (see below) and encodes an RNA product, which is most commonly, but not always, translated into a protein that exerts activity within or outside the cell **(Fig. 456-3)**. Historically, genes were identified because they conferred specific traits that are transmitted from one generation to the next. Increasingly, they are characterized based on expression in various tissues (*transcriptome*). The size of genes is quite broad; some genes are only a few hundred base pairs, whereas others are extraordinarily large (2 Mb). The number of genes greatly underestimates the complexity of genetic expression, because single genes can generate multiple spliced messenger RNA (mRNA) products (*isoforms*), which are translated into proteins that are subject to complex posttranslational modification such as phosphorylation. *Exons* refer to the portion of genes that are eventually spliced together to form mRNA. *Introns* refer to the spacing regions between the exons that are spliced out of precursor RNAs during RNA processing. The gene locus also includes regions that are necessary to control its expression (Fig. 456-2). Current estimates predict roughly 20,000 protein-coding genes in the human genome with an average of about four different coding transcripts per gene. Remarkably, the exome only constitutes 1.14% of the genome. Of note, the number of transcripts is close to 200,000 and includes thousands of noncoding transcripts (RNAs of various length such as microRNAs [miRNA] and long noncoding RNAs [lncRNA]). These non-coding RNAs are involved in numerous cellular processes such as transcriptional and posttranscriptional regulation of gene expression, chromatin remodeling, and protein trafficking, among others. Not surprisingly, aberrant expression and/or mutations in these RNAs play a pathogenic role in numerous diseases.

FIGURE 456-3

Chromosome 7 is shown with the density of single-nucleotide polymorphisms (SNPs) and genes above. A 200-kb region in 7q31.2 containing the *CFTR* gene is shown below. The *CFTR* gene contains 27 exons. Close to 2000 mutations in this gene have been found in patients with cystic fibrosis. A 20-kb region encompassing exons 4–9 is shown further amplified to illustrate the SNPs in this region.



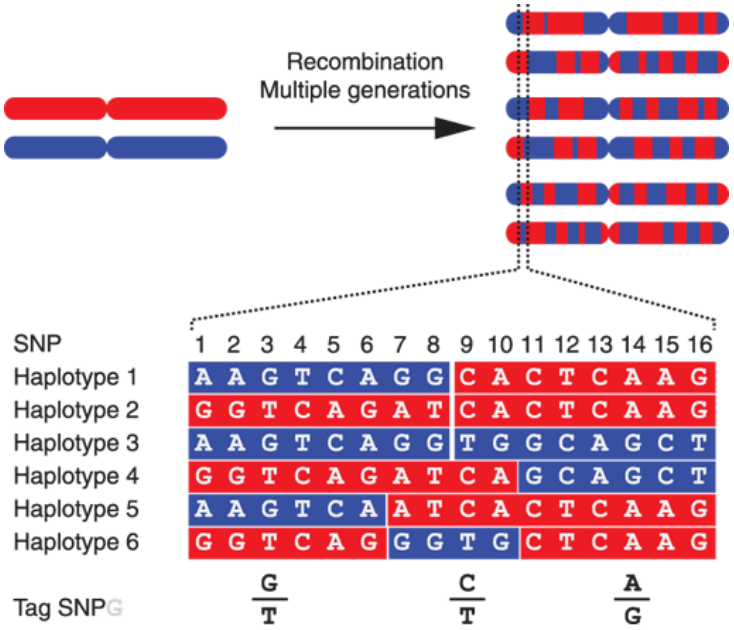
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

SINGLE-NUCLEOTIDE POLYMORPHISMS

Each individual has roughly 5 million sequence variants that differentiate one person from another. Some of these variants have no impact on health, whereas others may increase or lower the risk for developing a specific disease. Remarkably, however, the primary DNA sequence of humans has ~99.9% similarity compared to that of any other human. A SNP is a variation of a single base pair in the DNA. The identification of the ~10 million SNPs estimated to occur in the human genome has generated a catalogue of common genetic variants that occur in human beings from distinct ethnic backgrounds (Fig. 456-3). SNPs are the most common type of sequence variation and account for ~90% of all sequence variation. They occur on average every 100–300 bases and are the major source of genetic heterogeneity. SNPs that are in close proximity are inherited together (e.g., they are linked) and are referred to as *haplotypes* (Fig. 456-4). Haplotype maps describe the nature and location of these SNP haplotypes and how they are distributed among individuals within and among populations, information that is facilitating GWAS designed to elucidate the complex interactions among multiple genes and lifestyle factors in multifactorial disorders (see below). Moreover, haplotype analyses are useful to assess variations in responses to medications (*pharmacogenomics*) and environmental factors, as well as the prediction of disease predisposition.

FIGURE 456-4

The origin of haplotypes is due to repeated recombination events occurring in multiple generations. Over time, this leads to distinct haplotypes. These haplotype blocks can often be characterized by genotyping selected Tag single-nucleotide polymorphisms (SNPs), an approach that facilitates performing genome-wide association studies (GWAS).



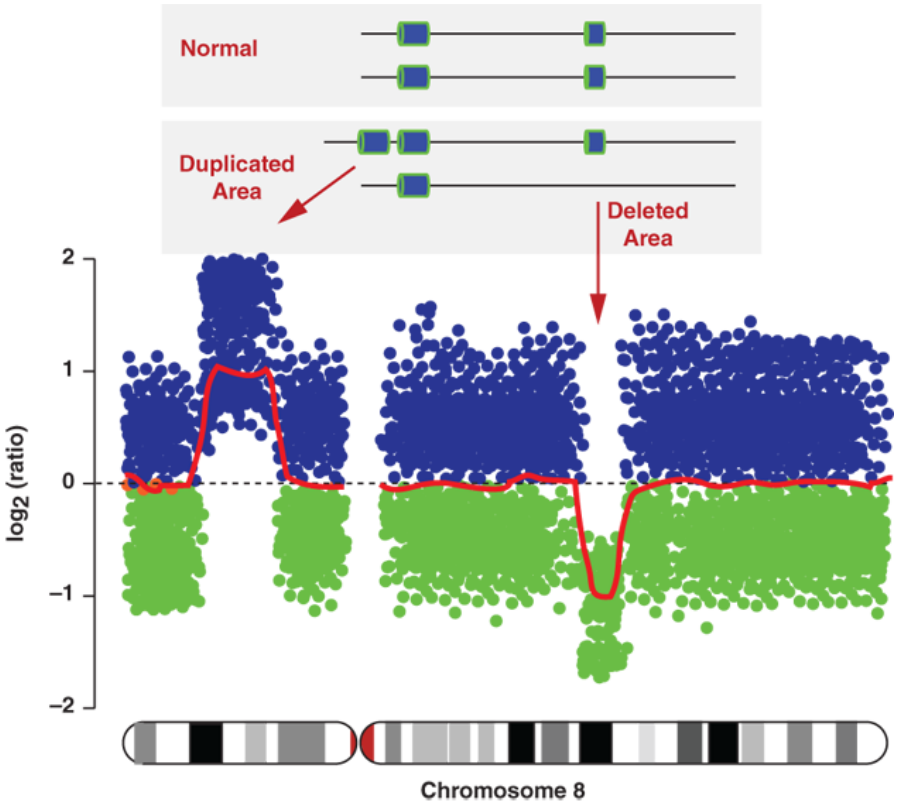
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

COPY NUMBER VARIATIONS

Copy number variations (CNVs) are relatively large genomic regions (1 kb to several Mb) that have been duplicated or deleted on certain chromosomes and hence alter the diploid status of the DNA (Fig. 456-5). It has been estimated that 5–10% of the genome can display CNVs. When comparing the genomes of two individuals, ~0.4–0.8% of their genomes differ in terms of CNVs scattered throughout the genome. Of note, de novo CNVs have been observed between monozygotic twins, who otherwise have identical genomes. Some CNVs have no functional consequences, whereas others have been associated with susceptibility or resistance to disease, and CNVs also occur in cancer cells.

FIGURE 456-5

Copy number variations (CNV) encompass relatively large regions of the genome that have been duplicated or deleted. Chromosome 8 is shown with CNV detected by genomic hybridization. An increase in the signal strength indicates a duplication, a decrease reflects a deletion of the covered chromosomal regions.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Replication of DNA and Mitosis

Genetic information in DNA is transmitted to daughter cells under two different circumstances: (1) somatic cells divide by *mitosis*, allowing the diploid ($2n$) genome to replicate itself completely in conjunction with cell division; and (2) germ cells (sperm and ova) undergo *meiosis*, a process that enables the reduction of the diploid ($2n$) set of chromosomes to the haploid state ($1n$).

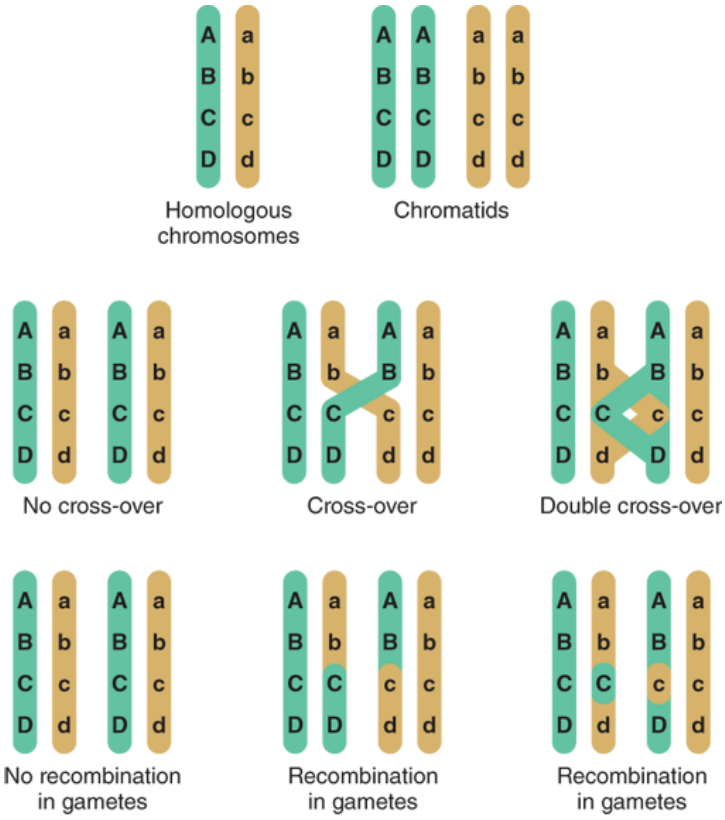
Prior to mitosis, cells exit the resting, or G_0 state, and enter the cell cycle. After traversing a critical checkpoint in G_1, cells undergo DNA synthesis (S phase), during which the DNA in each chromosome is replicated, yielding two pairs of sister chromatids ($2n \rightarrow 4n$). The process of DNA synthesis requires stringent fidelity in order to avoid transmitting errors to subsequent generations of cells. Genetic abnormalities of DNA mismatch/repair include xeroderma pigmentosum, Bloom's syndrome, ataxia telangiectasia, and hereditary nonpolyposis colon cancer (HNPCC), among others. Many of these disorders strongly predispose to neoplasia because of the rapid acquisition of additional mutations **(Chap. 67)**. After completion of DNA synthesis, cells enter G_2 and progress through a second checkpoint before entering mitosis. At this stage, the chromosomes condense and are aligned along the equatorial plate at metaphase. The two identical sister chromatids, held together at the centromere, divide and migrate to opposite poles of the cell. After formation of a nuclear membrane around the two separated sets of chromatids, the cell divides and two daughter cells are formed, thus restoring the diploid ($2n$) state.

Assortment and Segregation of Genes During Meiosis

Meiosis occurs only in germ cells of the gonads. It shares certain features with mitosis but involves two distinct steps of cell division that reduce the chromosome number to the haploid state. In addition, there is active recombination that generates genetic diversity. During the first cell division, two sister chromatids ($2n \rightarrow 4n$) are formed for each chromosome pair and there is an exchange of DNA between homologous paternal and maternal chromosomes. This process involves the formation of *chiasmata*, structures that correspond to the DNA segments that cross over between the maternal and paternal homologues **(Fig. 456-6)**. Usually there is at least one crossover on each chromosomal arm; recombination occurs more frequently in female meiosis than in male meiosis. Subsequently, the chromosomes segregate randomly. Because there are 23 chromosomes, there exist 2^{23} (>8 million) possible combinations of chromosomes. Together with the genetic exchanges that occur during recombination, chromosomal segregation generates tremendous diversity, and each gamete is genetically unique. The process of recombination and the independent segregation of chromosomes provide the foundation for performing linkage analyses, whereby one attempts to correlate the inheritance of certain chromosomal regions (or linked genes) with the presence of a disease or genetic trait (see below).

FIGURE 456-6
Crossing-over and genetic recombination. During chiasma formation, either of the two sister chromatids on one chromosome pairs with one of the chromatids of the homologous chromosome. Genetic recombination occurs through crossing-over and results in recombinant and nonrecombinant

chromosome segments in the gametes. Together with the random segregation of the maternal and paternal chromosomes, recombination contributes to genetic diversity and forms the basis of the concept of linkage.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

After the first meiotic division, which results in two daughter cells (2*n*), the two chromatids of each chromosome separate during a second meiotic division to yield four gametes with a haploid state (1*n*). When the egg is fertilized by sperm, the two haploid sets are combined, thereby restoring the diploid state (2*n*) in the zygote.

REGULATION OF GENE EXPRESSION

Regulation by Transcription Factors

The expression of genes is regulated by DNA-binding proteins that activate or repress transcription. The number of DNA sequences and transcription factors that regulate transcription is much greater than originally anticipated. Most genes contain at least 15–20 discrete regulatory elements within 300 bp of the transcription start site. This densely packed promoter region often contains binding sites for ubiquitous transcription factors. However, factors involved in cell-specific expression may also bind to these sequences. Key regulatory elements may also reside at a large distance from the proximal promoter. The globin and the immunoglobulin genes, for example, contain *locus control regions* that are several kilobases away from the structural sequences of the gene. Specific groups of transcription factors that bind to these promoter and enhancer sequences provide a combinatorial code for regulating transcription. In this manner, relatively ubiquitous factors interact with more restricted factors to allow each gene to be expressed and regulated in a unique manner that is dependent on developmental state, cell type, and numerous extracellular stimuli. Regulatory factors also bind within the gene itself, particularly in the intronic regions. The transcription factors that bind to DNA actually represent only the first level of regulatory control. Other proteins—*co-activators* and *co-repressors*—interact with the DNA-binding transcription factors to generate large regulatory complexes. These complexes are subject to control by numerous cell-signaling pathways and enzymes, leading to phosphorylation, acetylation, sumoylation, and ubiquitination. Ultimately, the recruited transcription factors interact with, and stabilize, components of the basal transcription complex that assembles at the site of the TATA box and initiator region. This basal transcription factor complex consists of >30 different proteins. Gene transcription occurs when RNA polymerase begins to synthesize RNA from the DNA template. A large number of identified genetic diseases involve transcription factors (Table 456-2).

TABLE 456-2

Selected Examples of Diseases Caused by Mutations and Rearrangements in Transcription Factors

Transcription Factor Class	Example	Associated Disorder
Nuclear receptors	Androgen receptor	Complete or partial androgen insensitivity (recessive missense mutations)
		Spinobulbar muscular atrophy (CAG repeat expansion)
Zinc finger proteins	WT1	WAGR syndrome: *W*ilms' tumor, *a*niridia, *g*enitourinary malformations, mental *r*etardation
Basic helix-loop-helix	MITF	Waardenburg's syndrome type 2A
Homeobox	IPF1	Maturity onset of diabetes mellitus type 4 (heterozygous mutation/haploinsufficiency)
		Pancreatic agenesis (homozygous mutation)
Leucine zipper	Retina leucine zipper (NRL)	Autosomal dominant retinitis pigmentosa
High mobility group (HMG) proteins	SRY	Sex reversal
Forkhead	HNF4α, HNF1α, HNF1β	Maturity onset of diabetes mellitus types 1, 3, 5
Paired box	PAX3	Waardenburg's syndrome types 1 and 3
T-box	TBX5	Holt-Oram syndrome (thumb anomalies, atrial or ventricular septum defects, phocomelia)
Cell cycle control proteins	P53	Li-Fraumeni syndrome, other cancers
Co-activators	CREB binding protein (CBP)	Rubinstein-Taybi syndrome
General transcription factors	TATA-binding protein (TBP)	Spinocerebellar ataxia 17 (CAG expansion)
Transcription elongation factor	VHL	von Hippel–Lindau syndrome (renal cell carcinoma, pheochromocytoma, pancreatic tumors, hemangioblastomas) Autosomal dominant inheritance, somatic inactivation of second allele (Knudson two-hit model)
Runt	CBFA2	Familial thrombocytopenia with propensity to acute myelogenous leukemia
Chimeric proteins due to translocations	PML-RAR	Acute promyelocytic leukemia t(15;17)(q22;q11.2-q12) translocation

Abbreviations: CREB, cAMP responsive element–binding protein; HNF, hepatocyte nuclear factor; PML, promyelocytic leukemia; RAR, retinoic acid receptor; SRY, sex-determining region Y; VHL, von Hippel–Lindau.

The field of *functional genomics* is based on the concept that understanding alterations of gene expression under various physiologic and pathologic conditions provides insight into the underlying functional role of the gene. The ENCODE (ENCyclopedia Of DNA Elements) project aims to compile and annotate all functional sequences in the human genome. By revealing specific gene expression profiles, this knowledge may be of diagnostic and

therapeutic relevance. The large-scale study of expression profiles, which takes advantage of micro and bead array technologies, is also referred to as *transcriptomics* because the complement of mRNAs transcribed by the cellular genome is called the *transcriptome*.

Most studies of gene expression have focused on the regulatory DNA elements of genes that control transcription. However, it should be emphasized that gene expression requires a series of steps, including mRNA processing, protein translation, and posttranslational modifications, all of which are actively regulated (Fig. 456-2).

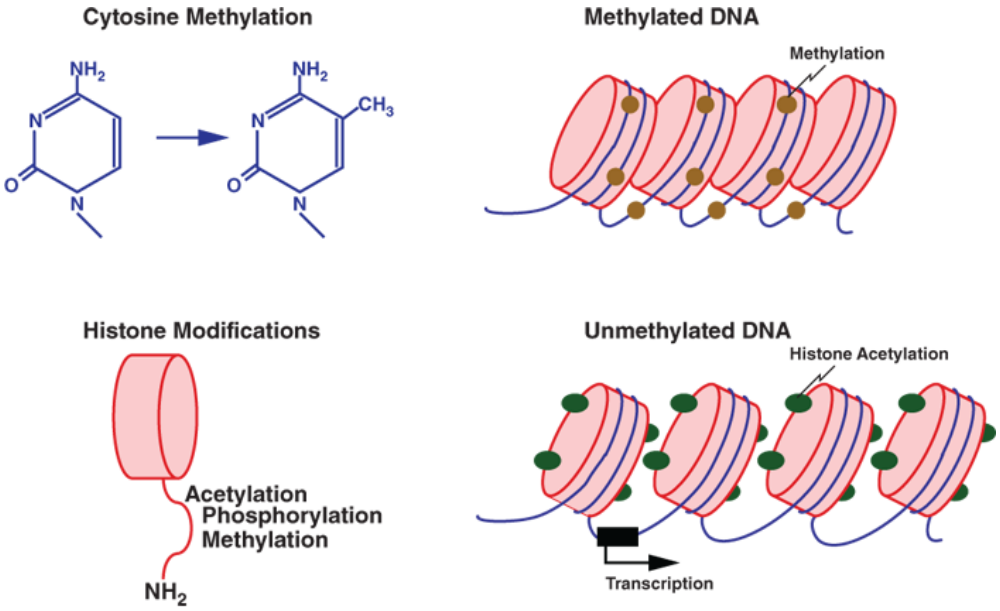
Epigenetic Regulation of Gene Expression

Epigenetics describes mechanisms and phenotypic changes that are not a result of variation in the primary DNA nucleotide sequence, but are caused by secondary modifications of DNA or histones **(see Chap. 471)**. These modifications include heritable changes such as X-inactivation and imprinting, but they can also result from dynamic posttranslational protein modifications in response to environmental influences such as diet, age, or drugs. The epigenetic modifications result in altered expression of individual genes or chromosomal loci encompassing multiple genes. The term *epigenome* describes the constellation of covalent modifications of DNA and histones that impact chromatin structure, as well as noncoding transcripts that modulate the transcriptional activity of DNA. Although the primary DNA sequence is usually identical in all cells of an organism, tissue-specific changes in the epigenome contribute to determining the transcriptional signature of a cell (transcriptome) and hence the protein expression profile (proteome).

Mechanistically, DNA and histone modifications can result in the activation or silencing of gene expression **(Fig. 456-7)**. DNA methylation involves the addition of a methyl group to cytosine residues. This is usually restricted to cytosines of CpG dinucleotides, which are abundant throughout the genome. Methylation of these dinucleotides is thought to represent a defense mechanism that minimizes the expression of sequences that have been incorporated into the genome such as retroviral sequences. CpG dinucleotides also exist in so-called *CpG islands*, stretches of DNA characterized by a high CG content, which are found in the majority of human gene promoters. CpG islands in promoter regions are typically unmethylated, and the lack of methylation facilitates transcription.

FIGURE 456-7

Epigenetic modifications of DNA and histones. Methylation of cytosine residues is associated with gene silencing. Methylation of certain genomic regions is inherited (imprinting), and it is involved in the silencing of one of the two X chromosomes in females (X-inactivation). Alterations in methylation can also be acquired, e.g., in cancer cells. Covalent posttranslational modifications of histones play an important role in altering DNA accessibility and chromatin structure and hence in regulating transcription. Histones can be reversibly modified in their amino-terminal tails, which protrude from the nucleosome core particle, by acetylation of lysine, phosphorylation of serine, methylation of lysine and arginine residues, and sumoylation. Acetylation of histones by histone acetylases (HATs), e.g., leads to unwinding of chromatin and accessibility to transcription factors. Conversely, deacetylation by histone deacetylases (HDACs) results in a compact chromatin structure and silencing of transcription.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

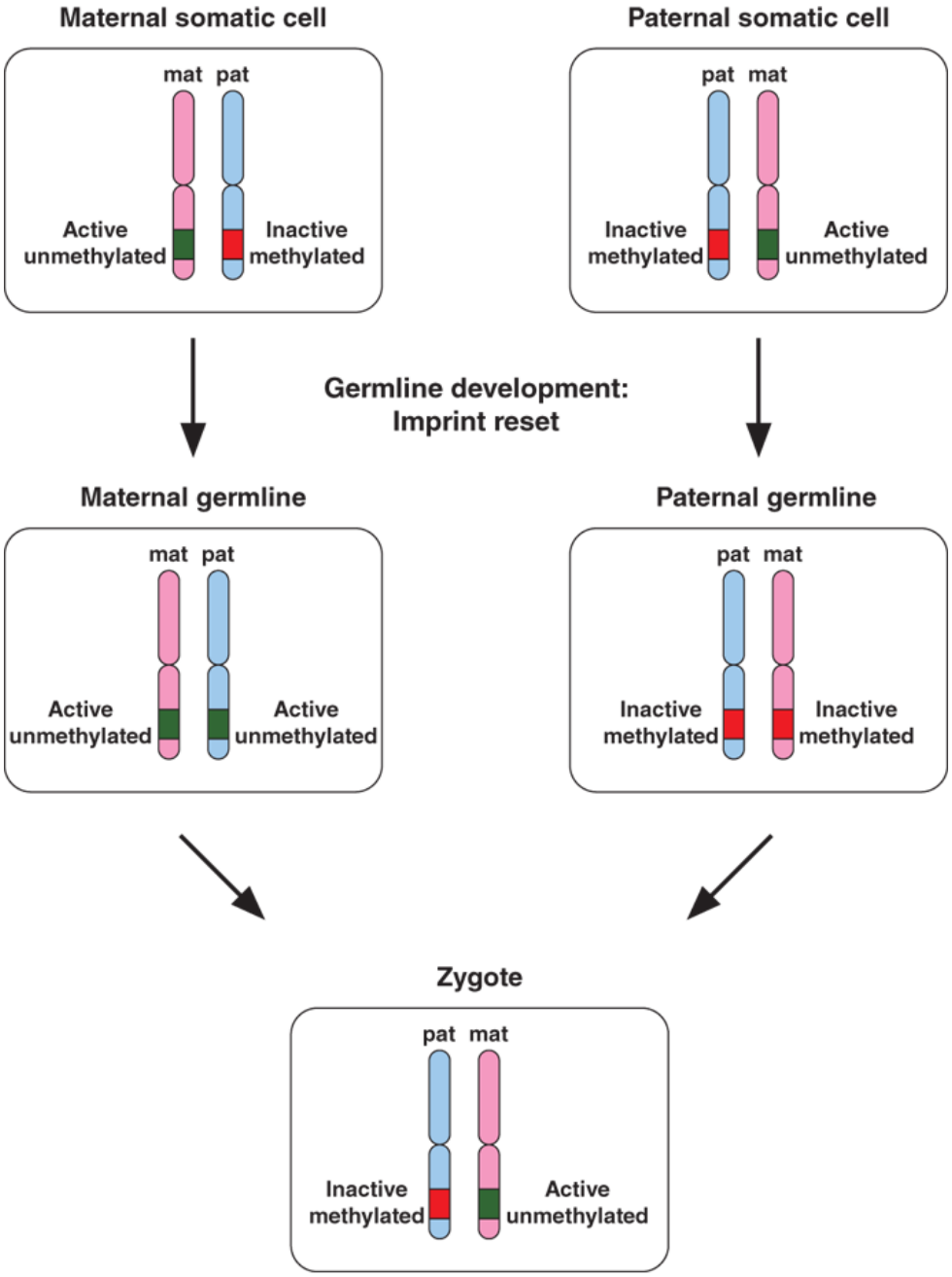
Histone methylation involves the addition of a methyl group to lysine residues in histone proteins (Fig. 456-7). Depending on the specific lysine residue being methylated, this alters chromatin configuration, either making it more open or tightly packed. Acetylation of histone proteins is another well-characterized mechanism that results in an open chromatin configuration, which favors active transcription. Acetylation is generally more dynamic than methylation, and many transcriptional activation complexes have histone acetylase activity, whereas repressor complexes often contain deacetylases and remove acetyl groups from histones. Other histone modifications, whose effects are incompletely characterized, include phosphorylation and sumoylation. Lastly, noncoding RNAs that bind to DNA can have a significant impact on transcriptional activity.

Physiologically, epigenetic mechanisms play an important role in several instances. For example, X-inactivation refers to the relative silencing of one of the two X chromosome copies present in females. The inactivation process is a form of dosage compensation such that females (XX) do not generally express twice as many X-chromosomal gene products as males (XY). In a given cell, the choice of which chromosome is inactivated occurs randomly in humans. But once the maternal or paternal X chromosome is inactivated, it will remain inactive, and this information is transmitted with each cell division. The *X-inactive specific transcript* (*Xist*) gene encodes a large noncoding RNA that mediates the silencing of the X chromosome from which it is transcribed by coating it with Xist RNA. The inactive X chromosome is highly methylated and has low levels of histone acetylation. While the majority of X-chromosomal genes are silenced by X-inactivation, about 15% escape inactivation and are expressed.

Epigenetic gene inactivation also occurs on selected chromosomal regions of autosomes, a phenomenon referred to as *genomic imprinting*. Through this mechanism, a small subset of genes is only expressed in a monoallelic fashion. Imprinting is heritable and leads to the preferential expression of one of the parental alleles, which deviates from the usual biallelic expression seen for the majority of genes. Remarkably, imprinting can be limited to a subset of tissues. Imprinting is mediated through DNA methylation of one of the alleles. The epigenetic marks on imprinted genes are maintained throughout life, but during zygote formation, they are activated or inactivated in a sex-specific manner (imprint reset) **(Fig. 456-8)**, which allows a differential expression pattern in the fertilized egg and the subsequent mitotic divisions. Appropriate expression of imprinted genes is important for normal development and cellular functions. Imprinting defects and uniparental disomy, which is the inheritance of two chromosomes or chromosomal regions from the same parent, are the cause of several developmental disorders such as Beckwith-Wiedemann syndrome, Silver-Russell syndrome, Angelman's syndrome, and Prader-Willi syndrome (see below). Monoallelic loss-of-function mutations in the *GNAS1* gene lead to Albright's hereditary osteodystrophy (AHO). Paternal transmission of *GNAS1* mutations leads to an isolated AHO phenotype (pseudopseudohypoparathyroidism), whereas maternal transmission leads to AHO in combination with hormone resistance to parathyroid hormone, thyrotropin, and gonadotropins (pseudohypoparathyroidism type IA). These phenotypic differences are explained by tissue-specific imprinting of the *GNAS1* gene, which is expressed primarily from the maternal allele in the thyroid, gonadotropes, and the proximal renal tubule. In most other tissues, the *GNAS1* gene is expressed biallelically. In patients with isolated renal resistance to parathyroid hormone (pseudohypoparathyroidism type IB), defective imprinting of the *GNAS1* gene results in decreased Gsα expression in the proximal renal tubules. Rett's syndrome is an X-linked dominant disorder resulting in developmental regression and stereotypic hand movements in affected girls. It is caused by mutations in the *MECP2* gene, which encodes a methyl-binding protein. The ensuing aberrant methylation results in abnormal gene expression in neurons, which are otherwise normally developed.

FIGURE 456-8

A few genomic regions are imprinted in a parent-specific fashion. The unmethylated chromosomal regions are actively expressed, whereas the methylated regions are silenced. In the germline, the imprint is reset in a parent-specific fashion: both chromosomes are unmethylated in the maternal (mat) germline and methylated in the paternal (pat) germline. In the zygote, the resulting imprinting pattern is identical with the pattern in the somatic cells of the parents.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Remarkably, epigenetic differences also occur among monozygotic twins. Although twins are epigenetically indistinguishable during the early years of life, older monozygotic twins exhibit differences in the overall content and genomic distribution of DNA methylation and histone acetylation, which would be expected to alter gene expression in various tissues.

In cancer, the epigenome is characterized by simultaneous losses and gains of DNA methylation in different genomic regions, as well as repressive histone modifications. Hyper- and hypomethylation are associated with mutations in genes that control DNA methylation. Hypomethylation is thought to remove normal control mechanisms that prevent expression of repressed DNA regions. It is also associated with genomic instability. Hypermethylation, in contrast, results in the silencing of CpG islands in promoter regions of genes, including tumor-suppressor genes. Epigenetic alterations are considered to be more easily reversible compared to genetic changes and modification of the epigenome with demethylating agents and histone deacetylases is being used in the treatment of various malignancies.

TRANSMISSION OF GENETIC DISEASE

Origins and Types of Mutations

The term *mutation* is used to designate the process of generating genetic variations as well as the outcome of these alterations. A *mutation* can be defined as any change in the primary nucleotide sequence of DNA regardless of its functional consequences, although it often has a negative connotation. The more

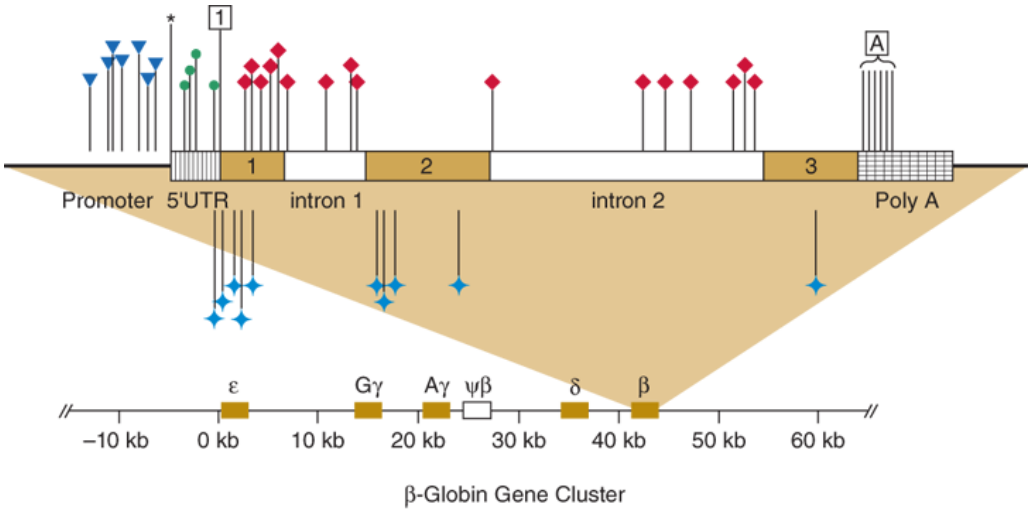
neutral term *variation* is now increasingly used to describe sequence changes and is recommended by several professional organizations and guidelines instead of *mutation*. Some variations may be lethal, others are less deleterious, and some may confer an evolutionary advantage. Variations can occur in the germline (sperm or oocytes); these can be transmitted to progeny. Alternatively, variations can occur during embryogenesis or in somatic tissues. Variations that occur during development lead to *mosaicism*, a situation in which tissues are composed of cells with different genetic constitutions. If the germline is mosaic, a mutation can be transmitted to some progeny but not others, which sometimes leads to confusion in assessing the pattern of inheritance. Somatic mutations that do not affect cell survival can sometimes be detected because of variable phenotypic effects in tissues (e.g., pigmented lesions in McCune-Albright syndrome). Other somatic mutations are associated with neoplasia because they confer a growth advantage to cells. Epigenetic events may also influence gene expression or facilitate genetic damage. With the exception of triplet nucleotide repeats, which can expand (see below), variations are usually stable.

Mutations are structurally diverse—they can involve the entire genome, as in triploidy (one extra set of chromosomes), or gross numerical or structural alterations in chromosomes or individual genes. Large deletions may affect a portion of a gene or an entire gene, or, if several genes are involved, they may lead to a *contiguous gene syndrome*. Unequal crossing-over between homologous genes can result in fusion gene mutations, as illustrated by color blindness. Variations involving single nucleotides are referred to as *point mutations*. Substitutions are called *transitions* if a purine is replaced by another purine base (A ↔ G) or if a pyrimidine is replaced by another pyrimidine (C ↔ T). Changes from a purine to a pyrimidine, or vice versa, are referred to as *transversions*. If the DNA sequence change occurs in a coding region and alters an amino acid, it is called a *missense mutation*. Depending on the functional consequences of such a missense mutation, amino acid substitutions in different regions of the protein can lead to distinct phenotypes.

Variations can occur in all domains of a gene **(Fig. 456-9)**. A point mutation occurring within the coding region leads to an amino acid substitution if the codon is altered **(Fig. 456-10)**. Point mutations that introduce a premature stop codon result in a truncated or missing protein. Large deletions may affect a portion of a gene or an entire gene, whereas small deletions and insertions alter the reading frame if they do not represent a multiple of three bases. These "frameshift" mutations, now also designated as *amphigoric* amino acid changes, lead to an entirely altered carboxy terminus. Mutations in intronic sequences or in exon junctions may destroy or create splice donor or splice acceptor sites. Variations may also be found in the regulatory sequences of genes, resulting in reduced or enhanced gene transcription.

FIGURE 456-9

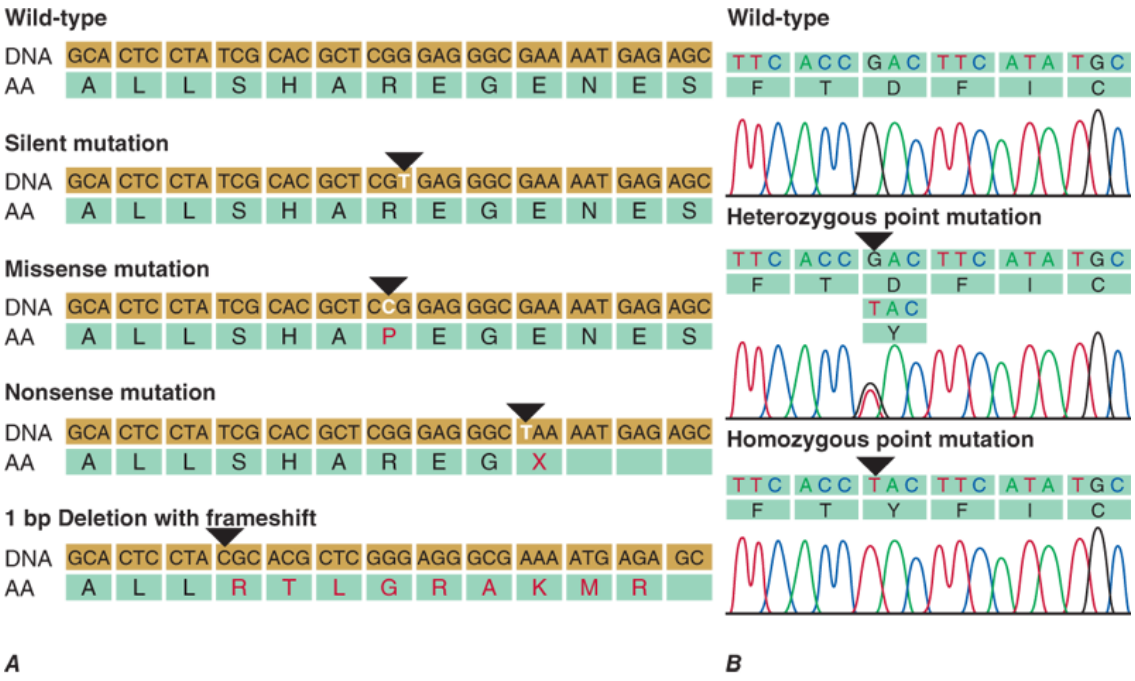
Point mutations causing β thalassemia as example of allelic heterogeneity. The β-globin gene is located in the globin gene cluster. Point mutations can be located in the promoter, the CAP site, the 5'-untranslated region, the initiation codon, each of the three exons, the introns, or the polyadenylation signal. Many mutations introduce missense or nonsense mutations, whereas others cause defective RNA splicing. Not shown here are deletion mutations of the β-globin gene or larger deletions of the globin locus that can also result in thalassemia. ▼, promoter mutations; *, CAP site; ●, 5'UTR; ◫, initiation codon; ◆, defective RNA processing; ✦, missense and nonsense mutations; A, Poly A signal.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 456-10

A. Examples of mutations (now commonly referred to as *variations*). The coding strand is shown with the encoded amino acid sequence. ***B.*** Chromatograms of sequence analyses after amplification of genomic DNA by polymerase chain reaction.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Certain DNA sequences are particularly susceptible to mutagenesis. Successive pyrimidine residues (e.g., T-T or C-C) are subject to the formation of ultraviolet light–induced photoadducts. If these pyrimidine dimers are not repaired by the nucleotide excision repair pathway, mutations will be introduced after DNA synthesis. The dinucleotide C-G, or CpG, is also a hot spot for a specific type of mutation. In this case, methylation of the cytosine is associated with an enhanced rate of deamination to uracil, which is then replaced with thymine. This C → T transition (or G → A on the opposite strand) accounts for at least one-third of point mutations associated with polymorphisms and mutations. In addition to the fact that certain types of mutations (C → T or G → A) are relatively common, the nature of the genetic code also results in overrepresentation of certain amino acid substitutions.

Polymorphisms are sequence variations that have a frequency of at least 1%. Usually, they do not result in a perceptible phenotype but because allele frequency and functional consequences are often not known, the term variation is now increasingly recommended for the description of these sequence changes. Often they consist of single base-pair substitutions that do not alter the protein coding sequence because of the degenerate nature of the genetic code (synonymous polymorphism), although it is possible that some might alter mRNA stability, translation, or the amino acid sequence (nonsynonymous polymorphism) (Fig. 456-10). The detection of sequence variants poses a practical problem because it is often unclear whether it creates a change with functional consequences or a benign variation. In this situation, the sequence alteration is also described as *variant of unknown significance (VUS)*.

MUTATION RATES

Mutations represent an important cause of genetic diversity as well as disease. Mutation rates are difficult to determine in humans because many mutations are silent and because testing is often not adequate to detect the phenotypic consequences. Mutation rates vary in different genes but are estimated to occur at a rate of ~10^{-10}/bp per cell division. Germline mutation rates (as opposed to somatic mutations) are relevant in the transmission of genetic disease. Because the population of oocytes is established very early in development, only ~20 cell divisions are required for completed oogenesis, whereas spermatogenesis involves ~30 divisions by the time of puberty and 20 cell divisions each year thereafter. Consequently, the probability of acquiring new point mutations is much greater in the male germline than the female germline, in which rates of aneuploidy are increased. Thus, the incidence of new point mutations in spermatogonia increases with paternal age (e.g., achondrodysplasia, Marfan's syndrome, neurofibromatosis). It is estimated that about 1 in 10 sperm carries a new deleterious mutation. The rates for new mutations are calculated most readily for autosomal dominant and X-linked disorders and are ~10^{-5}–10^{-6}/locus per generation. Because most monogenic diseases are relatively rare, new mutations account for a significant fraction of cases. This is important in the context of genetic counseling, because a new mutation can be transmitted to the affected individual but does not necessarily imply that the parents are at risk to transmit the disease to other children. An exception to this is when the new mutation occurs early in germline development, leading to *gonadal mosaicism*.

UNEQUAL CROSSING-OVER

Normally, DNA recombination in germ cells occurs with remarkable fidelity to maintain the precise junction sites for the exchanged DNA sequences (Fig. 456-6). However, mispairing of homologous sequences leads to unequal crossover, with gene duplication on one of the chromosomes and gene deletion on the other chromosome. A significant fraction of growth hormone (*GH*) gene deletions, for example, involve unequal crossing-over **(Chap. 372)**. The *GH* gene is a member of a large gene cluster that includes a GH variant gene as well as several structurally related chorionic somatomammotropin genes and pseudogenes (highly homologous but functionally inactive relatives of a normal gene). Because such gene clusters contain multiple homologous DNA

sequences arranged in tandem, they are particularly prone to undergo recombination and, consequently, gene duplication or deletion. On the other hand, duplication of the *PMP22* gene because of unequal crossing-over results in increased gene dosage and type IA Charcot-Marie-Tooth disease. Unequal crossing-over resulting in deletion of *PMP22* causes a distinct neuropathy called *hereditary liability to pressure palsy* (Chap. 438).

Glucocorticoid-remediable aldosteronism (GRA) is caused by a gene fusion or rearrangement involving the genes that encode aldosterone synthase (*CYP11B2*) and steroid 11β-hydroxylase (*CYP11B1*), normally arranged in tandem on chromosome 8q. These two genes are 95% identical, predisposing to gene duplication and deletion by unequal crossing-over. The rearranged gene product contains the regulatory regions of 11β-hydroxylase fused to the coding sequence of aldosterone synthetase. Consequently, the latter enzyme is expressed in the adrenocorticotropic hormone (ACTH)–dependent zona fasciculata of the adrenal gland, resulting in overproduction of mineralocorticoids and hypertension (Chap. 379).

Gene conversion refers to a nonreciprocal exchange of homologous genetic information. It has been used to explain how an internal portion of a gene is replaced by a homologous segment copied from another allele or locus; these genetic alterations may range from a few nucleotides to a few thousand nucleotides. As a result of gene conversion, it is possible for short DNA segments of two chromosomes to be identical, even though these sequences are distinct in the parents. A practical consequence of this phenomenon is that nucleotide substitutions can occur during gene conversion between related genes, often altering the function of the gene. In disease states, gene conversion often involves intergenic exchange of DNA between a gene and a related pseudogene. For example, the 21-hydroxylase gene (*CYP21A2*) is adjacent to a nonfunctional pseudogene (*CYP21A1P*). Many of the nucleotide substitutions that are found in the *CYP21A2* gene in patients with congenital adrenal hyperplasia correspond to sequences that are present in the *CYP21A1P* pseudogene, suggesting gene conversion as one cause of mutagenesis. In addition, mitotic gene conversion has been suggested as a mechanism to explain revertant mosaicism in which an inherited mutation is “corrected” in certain cells. For example, patients with autosomal recessive generalized atrophic benign epidermolysis bullosa have acquired reverse mutations in one of the two mutated *COL17A1* alleles, leading to clinically unaffected patches of skin.

INSERTIONS AND DELETIONS

Although many instances of insertions and deletions occur as a consequence of unequal crossing-over, there is also evidence for internal duplication, inversion, or deletion of DNA sequences. The fact that certain deletions or insertions appear to occur repeatedly as independent events indicates that specific regions within the DNA sequence predispose to these errors. For example, certain regions of the *DMD* gene, which encodes dystrophin, appear to be hot spots for deletions and result in muscular dystrophy (Chap. 441). Some regions within the human genome are rearrangement hot spots and lead to CNVs.

ERRORS IN DNA REPAIR

Because mutations caused by defects in DNA repair accumulate as somatic cells divide, these types of mutations are particularly important in the context of neoplastic disorders. Several genetic disorders involving DNA repair enzymes underscore their importance. Patients with xeroderma pigmentosum have defects in DNA damage recognition or in the nucleotide excision and repair pathway (Chap. 72). Exposed skin is dry and pigmented and is extraordinarily sensitive to the mutagenic effects of ultraviolet irradiation. More than 10 different genes have been shown to cause the different forms of xeroderma pigmentosum. This finding is consistent with the earlier classification of this disease into different complementation groups in which normal function is rescued by the fusion of cells derived from two different forms of xeroderma pigmentosum.

Ataxia telangiectasia causes large telangiectatic lesions of the face, cerebellar ataxia, immunologic defects, and hypersensitivity to ionizing radiation (Chap. 431). The discovery of the ataxia telangiectasia mutated (*ATM*) gene reveals that it is homologous to genes involved in DNA repair and control of cell cycle checkpoints. Mutations in the *ATM* gene give rise to defects in meiosis as well as increasing susceptibility to damage from ionizing radiation. Fanconi’s anemia is also associated with an increased risk of multiple acquired genetic abnormalities. It is characterized by diverse congenital anomalies and a strong predisposition to develop aplastic anemia and acute myelogenous leukemia (Chap. 100). Cells from these patients are susceptible to chromosomal breaks caused by a defect in genetic recombination. Currently, at least 16 different complementation groups have been identified, and the genes associated with Fanconi’s anemia have been cloned. HNPCC (Lynch’s syndrome) is characterized by autosomal dominant transmission of colon cancer, young age (<50 years) of presentation, predisposition to lesions in the proximal large bowel, and associated malignancies such as uterine cancer and ovarian cancer. HNPCC is predominantly caused by mutations in one of several different mismatch repair (MMR) genes including MutS homologue 2 (*MSH2*), MutL homologue 1 and 6 (*MLH1, MLH6*), *MSH6*, *PMS1*, and *PMS2* (Chap. 77). These proteins are involved in the detection of nucleotide mismatches and in the recognition of slipped-strand trinucleotide repeats. Germline mutations in these genes lead to microsatellite instability and a high mutation rate in colon cancer. Genetic screening tests for this disorder are now being used for families considered to be at risk. Recognition of HNPCC allows early screening with colonoscopy and the implementation of prevention strategies using nonsteroidal anti-inflammatory drugs.

UNSTABLE DNA SEQUENCES

Trinucleotide repeats may be unstable and expand beyond a critical number. Mechanistically, the expansion is thought to be caused by unequal recombination and slipped mispairing. A premutation represents a small increase in trinucleotide copy number. In subsequent generations, the expanded repeat may increase further in length and result in an increasingly severe phenotype, a process called *dynamic mutation* (see below for discussion of anticipation). Trinucleotide expansion was first recognized as a cause of the fragile X syndrome, one of the most common causes of intellectual disability. Other disorders arising from a similar mechanism include Huntington’s disease, X-linked spinobulbar muscular atrophy, and myotonic dystrophy. Malignant cells are also characterized by genetic instability, indicating a breakdown in mechanisms that regulate DNA repair and the cell cycle.

Functional Consequences of Mutations

Functionally, mutations can be broadly classified as gain-of-function and loss-of-function mutations. Gain-of-function mutations are typically dominant (e.g., they result in phenotypic alterations when a single allele is affected). Inactivating mutations are usually recessive, and an affected individual is homozygous or compound heterozygous (e.g., carrying two different mutant alleles of the same gene) for the disease-causing mutations. Alternatively, mutation in a single allele can result in *haploinsufficiency*, a situation in which one normal allele is not sufficient to maintain a normal phenotype. Haploinsufficiency is a commonly observed mechanism in diseases associated with mutations in transcription factors (Table 456-2). Remarkably, the clinical features among patients with an identical mutation often vary significantly. One mechanism underlying this variability consists in the influence of modifying genes. Haploinsufficiency can also affect the expression of rate-limiting enzymes. For example, haploinsufficiency in enzymes involved in heme synthesis can cause porphyrias **(Chap. 409)**.

An increase in dosage of a gene product may also result in disease, as illustrated by the duplication of the *DAX1* gene in dosage-sensitive sex reversal **(Chap. 383)**. Mutation in a single allele can also result in loss of function due to a dominant-negative effect. In this case, the mutated allele interferes with the function of the normal gene product by one of several different mechanisms: (1) a mutant protein may interfere with the function of a multimeric protein complex, as illustrated by mutations in type 1 collagen (*COL1A1*, *COL1A2*) genes in osteogenesis imperfecta **(Chap. 406)**; (2) a mutant protein may occupy binding sites on proteins or promoter response elements, as illustrated by thyroid hormone resistance β, a disorder in which inactivated thyroid hormone receptor β binds to target genes and functions as an antagonist of normal receptors **(Chap. 375)**; or (3) a mutant protein can be cytotoxic as in α_1 antitrypsin deficiency **(Chap. 286)** or autosomal dominant neurohypophyseal diabetes insipidus **(Chap. 374)**, in which the abnormally folded proteins are trapped within the endoplasmic reticulum and ultimately cause cellular damage.

Genotype and Phenotype

ALLELES, GENOTYPES, AND HAPLOTYPES

An observed trait is referred to as a *phenotype*; the genetic information defining the phenotype is called the *genotype*. Alternative forms of a gene or a genetic marker are referred to as *alleles*. Alleles may be polymorphic variants of nucleic acids that have no apparent effect on gene expression or function. In other instances, these variants may have subtle effects on gene expression, thereby conferring adaptive advantages associated with genetic diversity. On the other hand, allelic variants may reflect mutations that clearly alter the function of a gene product. The common Glu6Val (E6V) sickle cell mutation in the *β-globin* gene and the ΔF508 deletion of phenylalanine (F) in the *CFTR* gene are examples of allelic variants of these genes that result in disease. Because each individual has two copies of each chromosome (one inherited from the mother and one inherited from the father), an individual can have only two alleles at a given locus. However, there can be many different alleles in the population. The normal or common allele is usually referred to as *wild type*. When alleles at a given locus are identical, the individual is *homozygous*. Inheriting identical copies of a mutant allele occurs in many autosomal recessive disorders, particularly in circumstances of consanguinity or isolated populations. If the alleles are different on the maternal and the paternal copy of the gene, the individual is *heterozygous* at this locus (Fig. 456-10). If two different mutant alleles are inherited at a given locus, the individual is said to be a *compound heterozygote*. *Hemizygous* is used to describe males with a mutation in an X chromosomal gene or a female with a loss of one X chromosomal locus.

Genotypes describe the specific alleles at a particular locus. For example, there are three common alleles (E2, E3, E4) of the apolipoprotein E (*APOE*) gene. The genotype of an individual can therefore be described as *APOE3/4* or *APOE4/4* or any other variant. These designations indicate which alleles are present on the two chromosomes in the *APOE* gene at locus 19q13.2. In other cases, the genotype might be assigned arbitrary numbers (e.g., 1/2) or letters (e.g., B/b) to distinguish different alleles.

A *haplotype* refers to a group of alleles that are closely linked together at a genomic locus (Fig. 456-4). Haplotypes are useful for tracking the transmission of genomic segments within families and for detecting evidence of genetic recombination, if the crossover event occurs between the alleles (Fig. 456-6). As an example, various alleles at the histocompatibility locus antigen (HLA) on chromosome 6p are used to establish haplotypes associated with certain disease states. For example, 21-hydroxylase deficiency, complement deficiency, and hemochromatosis are each associated with specific HLA haplotypes. It is now recognized that these genes lie in close proximity to the HLA locus, which explains why HLA associations were identified even before the disease genes were cloned and localized. In other cases, specific HLA associations with diseases such as ankylosing spondylitis (HLA-B27) or type 1 diabetes mellitus (HLA-DR4) reflect the role of specific HLA allelic variants in susceptibility to these autoimmune diseases. The characterization of common SNP haplotypes in numerous populations from different parts of the world has provided the necessary tools for association studies designed to detect genes involved in the pathogenesis of complex disorders (Table 456-1). The presence or absence of certain haplotypes can also be relevant for the customized choice of medical therapies (pharmacogenomics) or may have value for preventive strategies.

Genotype-phenotype correlation describes the association of a specific mutation and the resulting phenotype. The phenotype may differ depending on the location or type of the mutation in some genes. For example, in von Hippel–Lindau disease, an autosomal dominant multisystem disease that can include renal cell carcinoma, hemangioblastomas, and pheochromocytomas, among others, the phenotype varies greatly and the identification of the specific mutation can be clinically useful in order to predict the phenotypic spectrum.

ALLELIC HETEROGENEITY

Allelic heterogeneity refers to the fact that different mutations in the same genetic locus can cause an identical or similar phenotype. For example, many different mutations of the β-globin locus can cause β thalassemia **(Table 456-3)** (Fig. 456-9). In essence, allelic heterogeneity reflects the fact that many different mutations are capable of altering protein structure and function. For this reason, maps of inactivating mutations in genes usually show a near-random distribution. Exceptions include (1) a founder effect, in which a particular mutation that does not affect reproductive capacity can be traced to a single individual; (2) “hot spots” for mutations, in which the nature of the DNA sequence predisposes to a recurring mutation; and (3) localization of mutations to certain domains that are particularly critical for protein function. Allelic heterogeneity creates a practical problem for genetic testing because one must often examine the entire genetic locus for mutations, because these can differ in each patient. For example, about 2000 variants have been identified in the *CFTR* gene to date, although some of them are very rare and some may not be disease-causing (Fig. 456-3). Mutational analysis may initially focus on a panel of mutations that are particularly frequent (often taking the ethnic background of the patient into account), but a negative result does not exclude the presence of a mutation elsewhere in the gene. One should also be aware that mutational analyses tend to focus on the coding region of a gene without considering regulatory and intronic regions. Because disease-causing mutations may be located outside the coding regions, negative results need to be interpreted with caution. The advent of more comprehensive sequencing technologies now greatly facilitates concomitant mutational analyses of several genes after targeted enrichment, or even mutational analysis of the whole exome or genome. However, comprehensive sequencing can result in significant diagnostic challenges because the detection of a sequence alteration alone is not always sufficient to establish that it has a causal role (variants of unknown significance, VUS).

TABLE 456-3

Selected Examples of Phenotypic Heterogeneity and Locus Heterogeneity

Phenotypic Heterogeneity			
Gene, Protein	**Phenotype**	**Inheritance**	**OMIM**
LMNA, Lamin A/C	Emery-Dreifuss muscular dystrophy (AD)	AD	181350
	Familial partial lipodystrophy Dunnigan	AD	151660
	Hutchinson-Gilford progeria	AD	176670
	Atypical Werner's syndrome	AD	150330
	Dilated cardiomyopathy	AD	115200
	Early-onset atrial fibrillation	AD	607554
	Emery-Dreifuss muscular dystrophy (AR)	AR	604929
	Limb-girdle muscular dystrophy type 1B	AR	159001
	Charcot-Marie-Tooth type 2B1	AR	605588
KRAS	Noonan's syndrome	AD	163950
	Cardio-facio-cutaneous syndrome	AD	115150
Locus Heterogeneity			
Phenotype	**Gene**	**Chromosomal Location**	**Protein**
Familial hypertrophic cardiomyopathy	*MYH7*	14q12	Myosin heavy chain beta
Genes encoding sarcomeric proteins	*TNNT2*	1q2	Troponin-T2
	TPM1	15q22.1	Tropomyosin alpha
			Myosin-binding
	MYBPC3	11p11q	protein C
	TNNI3	19q13.4	Troponin 1
	MYL2	12q23-24.3	Myosin light chain 2
	MYL3	3p	Myosin light chain 3
	TTN	2q24.3	Cardiac titin
	ACTC	15q11	Cardiac alpha actin
	MYH6	14q1	Myosin heavy chain alpha
	MYLK2	20q13.3	Myosin light-peptide kinase

	CAV3	3p25	Caveolin 3
Genes encoding nonsarcomeric proteins	*MTT1*	Mitochondrial	tRNA isoleucine
	MTTG	Mitochondrial	tRNA glycine
	PRKAG2	7q35-q36	AMP-activated protein kinase γ2 subunit
	DMPK	19q13.2-13.3	Myotonin protein kinase (myotonic dystrophy)
	FRDA	9q13	Frataxin (Friedreich's ataxia)
Polycystic kidney disease	*PKD1*	16p13.3-13.12	Polycystin 1 (AD)
	PKD2	4q21.-23	Polycystin 2 (AD)
	PKHD1	6p21.1-p12	Fibrocystin (AR)
Noonan's syndrome	*PTPN11*	12q24.1	Protein-tyrosine phosphatase 2c
	KRAS	12p12.1	KRAS

PHENOTYPIC HETEROGENEITY

Phenotypic heterogeneity occurs when more than one phenotype is caused by allelic mutations (e.g., different mutations in the same gene) (Table 456-3). For example, laminopathies are monogenic multisystem disorders that result from mutations in the *LMNA* gene, which encodes the nuclear lamins A and C. Twelve autosomal dominant and five autosomal recessive disorders are caused by mutations in the *LMNA* gene. They include several forms of lipodystrophies, Emery-Dreifuss muscular dystrophy, progeria syndromes, a form of neuronal Charcot-Marie-Tooth disease (type 2B1), and a group of overlapping syndromes. Remarkably, hierarchical cluster analysis has revealed that the phenotypes vary depending on the position of the mutation (*genotype-phenotype correlation*). Similarly, identical mutations in the *FGFR2* gene can result in very distinct phenotypes: Crouzon's syndrome (craniofacial synostosis) or Pfeiffer's syndrome (acrocephalopolysyndactyly).

LOCUS OR NONALLELIC HETEROGENEITY AND PHENOCOPIES

Nonallelic or locus heterogeneity refers to the situation in which a similar disease phenotype results from mutations at different genetic loci (Table 456-3). This often occurs when more than one gene product produces different subunits of an interacting complex or when different genes are involved in the same genetic cascade or physiologic pathway. For example, osteogenesis imperfecta can arise from mutations in two different procollagen genes (*COL1A1* or *COL1A2*) that are located on different chromosomes, and can involve multiple other genes **(Chap. 406)**. The effects of inactivating mutations in these two genes are similar because the protein products comprise different subunits of the helical collagen fiber. Similarly, muscular dystrophy syndromes can be caused by mutations in various genes, consistent with the fact that it can be transmitted in an X-linked (Duchenne or Becker), autosomal dominant (limb-girdle muscular dystrophy type 1), or autosomal recessive (limb-girdle muscular dystrophy type 2) manner **(Chap. 441)**. Mutations in the X-linked *DMD* gene, which encodes dystrophin, are the most common cause of muscular dystrophy. This feature reflects the large size of the gene as well as the fact that the phenotype is expressed in hemizygous males because they have only a single copy of the X chromosome. Dystrophin is associated with a large protein complex linked to the membrane-associated cytoskeleton in muscle. Mutations in several different components of this protein complex can also cause muscular dystrophy syndromes. Although the phenotypic features of some of these disorders are distinct, the phenotypic spectrum caused by mutations in different genes overlaps, thereby leading to nonallelic heterogeneity. It should be noted that mutations in dystrophin are also associated with allelic heterogeneity. For example, mutations in the *DMD* gene can cause either Duchenne's or the less severe Becker's muscular dystrophy, depending on the severity of the protein defect.

Recognition of nonallelic heterogeneity is important for several reasons: (1) the ability to identify disease loci in linkage studies is reduced by including patients with similar phenotypes but different genetic disorders; (2) genetic testing is more complex because several different genes need to be considered along with the possibility of different mutations in each of the candidate genes; and (3) novel information is gained about how genes or proteins interact, providing unique insights into molecular physiology.

Phenocopies refer to circumstances in which nongenetic conditions mimic a genetic disorder. For example, features of toxin- or drug-induced neurologic syndromes can resemble those seen in Huntington's disease, and vascular causes of dementia share phenotypic features with familial forms of Alzheimer's

dementia (Chap. 423). As in nonallelic heterogeneity, the presence of phenocopies has the potential to confound linkage studies and genetic testing. Patient history and subtle differences in phenotype can often provide clues that distinguish these disorders from related genetic conditions.

VARIABLE EXPRESSIVITY AND INCOMPLETE PENETRANCE

The same genetic mutation may be associated with a phenotypic spectrum in different affected individuals, thereby illustrating the phenomenon of *variable expressivity*. This may include different manifestations of a disorder variably involving different organs (e.g., multiple endocrine neoplasia [MEN]), the severity of the disorder (e.g., cystic fibrosis), or the age of disease onset (e.g., Alzheimer's dementia). MEN 1 illustrates several of these features. In this autosomal dominant tumor syndrome, affected individuals carry an inactivating germline mutation that is inherited in an autosomal dominant fashion. After somatic inactivation of the alternate allele (loss of heterozygosity; Knudson two-hit model), they can develop tumors of the parathyroid gland, endocrine pancreas, and the pituitary gland (Chap. 381). However, the pattern of tumors in the different glands, the age at which tumors develop, and the types of hormones produced vary among affected individuals, even within a given family. In this example, the phenotypic variability arises, in part, because of the requirement for a second somatic mutation in the normal copy of the *MEN1* gene, as well as the large array of different cell types that are susceptible to the effects of *MEN1* gene mutations. In part, variable expression reflects the influence of modifier genes, or genetic background, on the effects of a particular mutation. Even in identical twins, in whom the genetic constitution is essentially the same, one can occasionally see variable expression of a genetic disease.

Interactions with the environment can also influence the course of a disease. For example, the manifestations and severity of hemochromatosis can be influenced by iron intake (Chap. 407), and the course of phenylketonuria is affected by exposure to phenylalanine in the diet (Chap. 413). Other metabolic disorders, such as hyperlipidemias and porphyria, also fall into this category. Many mechanisms, including genetic effects and environmental influences, can therefore lead to variable expressivity. In genetic counseling, it is particularly important to recognize this variability, because one cannot always predict the course of disease, even when the mutation is known.

Penetrance refers to the proportion of individuals with a mutant genotype that express the phenotype. If all carriers of a mutant express the phenotype, penetrance is complete, whereas it is said to be *incomplete* or *reduced* if some individuals do not exhibit features of the phenotype. Dominant conditions with incomplete penetrance are characterized by skipping of generations with unaffected carriers transmitting the mutant gene. For example, hypertrophic obstructive cardiomyopathy (HCM) caused by mutations in the *myosin-binding protein C* gene is a dominant disorder with clinical features in only a subset of patients who carry the mutation (Chap. 254). Patients who have the mutation but no evidence of the disease can still transmit the disorder to subsequent generations. In many conditions with postnatal onset, the proportion of gene carriers who are affected varies with age. Thus, when describing penetrance, one has to specify age. For example, for disorders such as Huntington's disease or familial amyotrophic lateral sclerosis, which present later in life, the rate of penetrance is influenced by the age at which the clinical assessment is performed. *Imprinting* can also modify the penetrance of a disease. For example, in patients with AHO, mutations in the Gsα subunit (*GNAS1* gene) are expressed clinically only in individuals who inherit the mutation from their mother (Chap. 403).

SEX-INFLUENCED PHENOTYPES

Certain mutations affect males and females quite differently. In some instances, this is because the gene resides on the X or Y sex chromosomes (X-linked disorders and Y-linked disorders). As a result, the phenotype of mutated X-linked genes will be expressed fully in males but variably in heterozygous females, depending on the degree of X-inactivation and the function of the gene. For example, most heterozygous female carriers of factor VIII deficiency (hemophilia A) are asymptomatic because sufficient factor VIII is produced to prevent a defect in coagulation (Chap. 112). On the other hand, some females heterozygous for the X-linked lipid storage defect caused by α-galactosidase A deficiency (Fabry's disease) experience mild manifestations of painful neuropathy, as well as other features of the disease (Chap. 411). Because only males have a Y chromosome, mutations in genes such as *SRY*, which causes male-to-female sex reversal, or *DAZ* (deleted in azoospermia), which causes abnormalities of spermatogenesis, are unique to males (Chap. 383).

Other diseases are expressed in a sex-limited manner because of the differential function of the gene product in males and females. Activating mutations in the luteinizing hormone receptor cause dominant male-limited precocious puberty in boys (Chap. 384). The phenotype is unique to males because activation of the receptor induces testosterone production in the testis, whereas it is functionally silent in the immature ovary. Biallelic inactivating mutations of the follicle-stimulating hormone (FSH) receptor cause primary ovarian failure in females because the follicles do not develop in the absence of FSH action. In contrast, affected males have a more subtle phenotype, because testosterone production is preserved (allowing sexual maturation) and spermatogenesis is only partially impaired (Chap. 384). In congenital adrenal hyperplasia, most commonly caused by 21-hydroxylase deficiency, cortisol production is impaired and ACTH stimulation of the adrenal gland leads to increased production of androgenic precursors (Chap. 379). In females, the increased androgen level causes ambiguous genitalia, which can be recognized at the time of birth. In males, the diagnosis may be made on the basis of adrenal insufficiency at birth, because the increased adrenal androgen level does not alter sexual differentiation, or later in childhood, because of the development of precocious puberty. Hemochromatosis is more common in males than in females, presumably because of differences in dietary iron intake and losses associated with menstruation and pregnancy in females (Chap. 407).

Chromosomal Disorders

Chromosomal disorders and the techniques used for their characterization have been discussed in detail in a chapter in previous editions of this textbook. Chromosomal or cytogenetic disorders are caused by numerical (aneuploidy) or structural aberrations (deletions, duplications, translocations, inversions,

dicentric and ring chromosomes, Robertsonian translocations) in chromosomes. They occur in about 1% of the general population, in 8% of stillbirths, and in close to 50% of spontaneously aborted fetuses. Indications for cytogenetic and cytogenomic chromosome analyses are summarized in **Table 456-4**. *Contiguous gene syndromes* (e.g., large deletions affecting several genes) have been useful for identifying the location of new disease-causing genes. Because of the variable size of gene deletions in different patients, a systematic comparison of phenotypes and locations of deletion breakpoints allows positions of particular genes to be mapped within the critical genomic region.

TABLE 456-4

Indications for Cytogenetic and Cytogenomic Analysis across the Lifespan

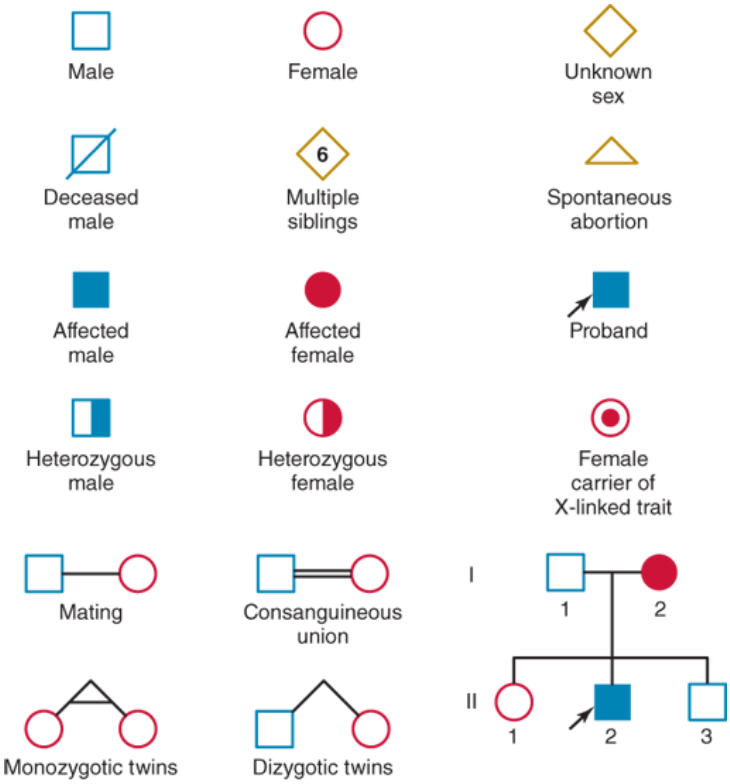
Timing of Testing	Indications for Testing
Prenatal	Advanced maternal age Abnormalities on ultrasound Increased risk for genetic disorder on maternal serum screen
Neonatal and Childhood	Multiple congenital anomalies Intellectual disability Autism Developmental delay Failure to thrive Short stature Disorders of sexual development History of familial chromosomal alteration Cancer
Adult	Infertility Recurrent miscarriage Familial cancer

Monogenic Mendelian Disorders

Monogenic human diseases are frequently referred to as *Mendelian disorders* because they obey the principles of genetic transmission originally set forth in Gregor Mendel's classic work. The continuously updated OMIM catalogue lists several thousand of these disorders and provides information about the clinical phenotype, molecular basis, allelic variants, and pertinent animal models (Table 456-1). The mode of inheritance for a given phenotypic trait or disease is determined by pedigree analysis. All affected and unaffected individuals in the family are recorded in a pedigree using standard symbols **(Fig. 456-11)**. The principles of allelic segregation, and the transmission of alleles from parents to children, are illustrated in **Fig. 456-12**. One dominant (A) allele and one recessive (a) allele can display three Mendelian modes of inheritance: autosomal dominant, autosomal recessive, and X-linked. About 65% of human monogenic disorders are autosomal dominant, 25% are autosomal recessive, and 5% are X-linked. Genetic testing is now available for many of these disorders and plays an important role in clinical medicine **(Chap. 457)**.

FIGURE 456-11

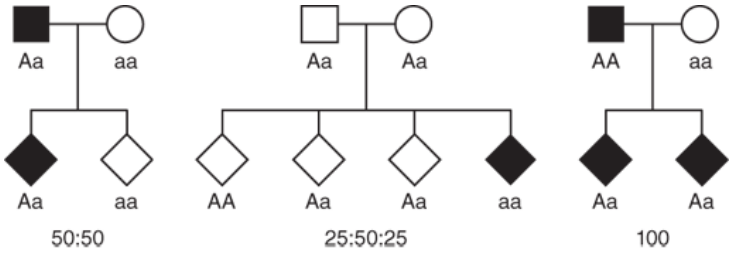
Standard pedigree symbols.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 456-12

Segregation of alleles. Segregation of genotypes in the offspring of parents with one dominant (A) and one recessive (a) allele. The distribution of the parental alleles to their offspring depends on the combination present in the parents. Filled symbols = affected individuals.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

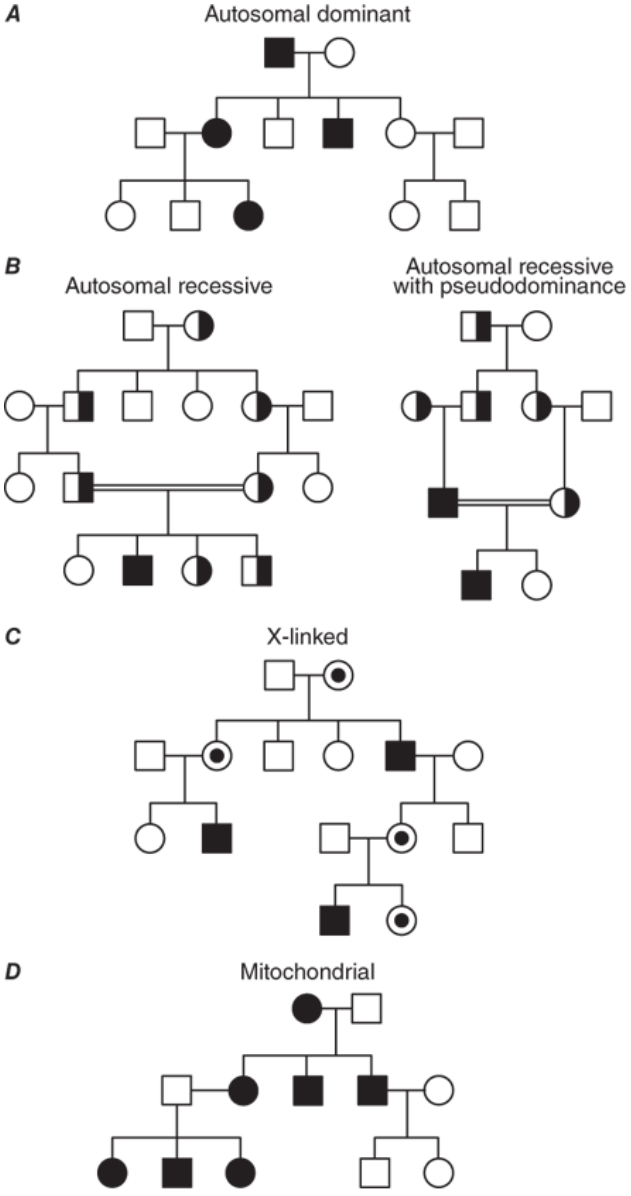
AUTOSOMAL DOMINANT DISORDERS

In autosomal dominant disorders, mutations in a single allele are sufficient to cause the disease. In contrast to recessive disorders, in which disease pathogenesis is relatively straightforward because there is a biallelic loss of gene function, dominant disorders can be caused by various disease mechanisms, many of which are unique to the function of the genetic pathway involved. Mechanistically, the mutation may confer constitutive activation (gain-of-function), exert a dominant negative effect, or result in loss-of-function and haploinsufficiency.

In autosomal dominant disorders, individuals are affected in successive generations; the disease does not occur in the offspring of unaffected individuals. Males and females are affected with equal frequency because the defective gene resides on one of the 22 autosomes (Fig. 456-13A). Autosomal dominant mutations alter one of the two alleles at a given locus. Because the alleles segregate randomly at meiosis, the probability that an offspring will be affected is 50%. Unless there is a new germline mutation, an affected individual has an affected parent. Children with a normal genotype do not transmit the disorder. Due to differences in penetrance or expressivity (see above), the clinical manifestations of autosomal dominant disorders may be variable. Because of these variations, it is sometimes challenging to determine the pattern of inheritance.

FIGURE 456-13

A. Dominant, ***B.*** recessive, ***C.*** X-linked, and ***D.*** mitochondrial (matrilinear) inheritance.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

It should be recognized, however, that some individuals acquire a mutated gene from an unaffected parent. De novo germline mutations occur more frequently during later cell divisions in gametogenesis, which explains why siblings are rarely affected. As noted before, new germline mutations occur more frequently in fathers of advanced age. For example, the average age of fathers with new germline mutations that cause Marfan's syndrome is ~37 years, whereas fathers who transmit the disease by inheritance have an average age of ~30 years.

AUTOSOMAL RECESSIVE DISORDERS

In recessive disorders, the mutated alleles result in a complete or partial loss of function. They frequently involve enzymes in metabolic pathways, receptors, or proteins in signaling cascades. In an autosomal recessive disease, the affected individual, who can be of either sex, is a homozygote or compound heterozygote for a single-gene defect. With a few important exceptions, autosomal recessive diseases are rare and often occur in the context of parental consanguinity. The relatively high frequency of certain recessive disorders such as sickle cell anemia, cystic fibrosis, and thalassemia, is partially explained by a selective biologic advantage for the heterozygous state (see below). Although heterozygous carriers of a defective allele are usually clinically normal, they may display subtle differences in phenotype that only become apparent with more precise testing or in the context of certain environmental influences. In sickle cell anemia, for example, heterozygotes are normally asymptomatic. However, in situations of dehydration or diminished oxygen pressure, sickle cell crises can also occur in heterozygotes **(Chap. 94)**.

In most instances, an affected individual is the offspring of heterozygous parents. In this situation, there is a 25% chance that the offspring will have a normal genotype, a 50% probability of a heterozygous state, and a 25% risk of homozygosity for the recessive alleles **(Figs. 456-10, 456-13B)**. In the case of one unaffected heterozygous and one affected homozygous parent, the probability of disease increases to 50% for each child. In this instance, the pedigree analysis mimics an autosomal dominant mode of inheritance (*pseudodominance*). In contrast to autosomal dominant disorders, new mutations in recessive alleles are rarely manifest because they usually result in an asymptomatic carrier state.

X-LINKED DISORDERS

Males have only one X chromosome; consequently, a daughter always inherits her father's X chromosome in addition to one of her mother's two X chromosomes. A son inherits the Y chromosome from his father and one maternal X chromosome. Thus, the characteristic features of X-linked inheritance are (1) the absence of father-to-son transmission, and (2) the fact that all daughters of an affected male are obligate carriers of the mutant allele **(Fig. 456-13C)**. The risk of developing disease due to a mutant X-chromosomal gene differs in the two sexes. Because males have only one X chromosome, they are hemizygous for the mutant allele; thus, they are more likely to develop the mutant phenotype, regardless of whether the mutation is dominant or recessive. A female may be either heterozygous or homozygous for the mutant allele, which may be dominant or recessive. The terms *X-linked dominant* or *X-linked recessive* are therefore only applicable to expression of the mutant phenotype in women. In addition, the expression of X-chromosomal genes is influenced by X chromosome inactivation.

Y-LINKED DISORDERS

The Y chromosome has a relatively small number of genes. One such gene, the sex-region determining Y factor (*SRY*), which encodes the testis-determining factor (*TDF*), is crucial for normal male development. Normally there is infrequent exchange of sequences on the Y chromosome with the X chromosome. The *SRY* region is adjacent to the pseudoautosomal region, a chromosomal segment on the X and Y chromosomes with a high degree of homology. A crossing-over event occasionally involves the *SRY* region with the distal tip of the X chromosome during meiosis in the male. Translocations can result in XY females with the Y chromosome lacking the *SRY* gene or XX males harboring the *SRY* gene on one of the X chromosomes **(Chap. 383)**. Point mutations in the *SRY* gene may also result in individuals with an XY genotype and an incomplete female phenotype. Most of these mutations occur de novo. Men with oligospermia/azoospermia frequently have microdeletions on the long arm of the Y chromosome that involve one or more of the azoospermia factor (*AZF*) genes.

Exceptions to Simple Mendelian Inheritance Patterns

MITOCHONDRIAL DISORDERS

Mendelian inheritance refers to the transmission of genes encoded by DNA contained in the nuclear chromosomes. In addition, each mitochondrion contains several copies of a small circular chromosome **(Chap. 472)**. The mitochondrial DNA (mtDNA) is ~16.5 kb and encodes transfer and ribosomal RNAs and 13 core proteins that are components of the respiratory chain involved in oxidative phosphorylation and ATP generation. The mitochondrial genome does not recombine and is inherited through the maternal line because sperm does not contribute significant cytoplasmic components to the zygote. A noncoding region of the mitochondrial chromosome, referred to as D-loop, is highly polymorphic. This property, together with the absence of mtDNA recombination, makes it a valuable tool for studies tracing human migration and evolution, and it is also used for specific forensic applications.

Inherited mitochondrial disorders are transmitted in a matrilineal fashion; all children from an affected mother will inherit the disease, but it will not be transmitted from an affected father to his children **(Fig. 456-13D)**. Alterations in the mtDNA that involves enzymes required for oxidative phosphorylation lead to reduction of ATP supply, generation of free radicals, and induction of apoptosis. Several syndromic disorders arising from mutations in the mitochondrial genome are known in humans and they affect both protein-coding and tRNA genes. The broad clinical spectrum often involves (cardio) myopathies and encephalopathies because of the high dependence of these tissues on oxidative phosphorylation. The age of onset and the clinical course are highly variable because of the unusual mechanisms of mtDNA transmission, which replicates independently from nuclear DNA. During cell replication, the proportion of wild-type and mutant mitochondria can drift among different cells and tissues. The resulting heterogeneity in the proportion of mitochondria with and without a mutation is referred to as *heteroplasmia* and underlies the phenotypic variability that is characteristic of mitochondrial diseases.

Acquired somatic mutations in mitochondria are thought to be involved in several age-dependent degenerative disorders affecting predominantly muscle and the peripheral and central nervous system (e.g., Alzheimer's and Parkinson's diseases). Establishing that an mtDNA alteration is causal for a clinical phenotype is challenging because of the high degree of polymorphism in mtDNA and the phenotypic variability characteristic of these disorders. Certain pharmacologic treatments may have an impact on mitochondria and/or their function. For example, treatment with the antiretroviral compound azidothymidine (AZT) causes an acquired mitochondrial myopathy through depletion of muscular mtDNA.

MOSAICISM

Mosaicism refers to the presence of two or more genetically distinct cell lines in the tissues of an individual. It results from a mutation that occurs during embryonic, fetal, or extrauterine development. The developmental stage at which the mutation arises will determine whether germ cells and/or somatic cells are involved. Chromosomal mosaicism results from nondisjunction at an early embryonic mitotic division, leading to the persistence of more than one cell line, as exemplified by some patients with Turner's syndrome **(Chap. 383)**. Somatic mosaicism is characterized by a patchy distribution of genetically altered somatic cells. The McCune-Albright syndrome, for example, is caused by activating mutations in the stimulatory G protein α (Gsα) that occur early in development **(Chap. 403)**. The clinical phenotype varies depending on the tissue distribution of the mutation; manifestations include ovarian cysts that secrete sex steroids and cause precocious puberty, polyostotic fibrous dysplasia, café-au-lait skin pigmentation, GH–secreting pituitary adenomas, and hypersecreting autonomous thyroid nodules.

X-INACTIVATION, IMPRINTING, AND UNIPARENTAL DISOMY

According to traditional Mendelian principles, the parental origin of a mutant gene is irrelevant for the expression of the phenotype. There are, however, important exceptions to this rule. *X-inactivation* prevents the expression of most genes on one of the two X chromosomes in every cell of a female. Gene inactivation through genomic imprinting occurs on selected chromosomal regions of autosomes and leads to inheritable preferential expression of one of the parental alleles. It is of pathophysiologic importance in disorders where the transmission of disease is dependent on the sex of the transmitting parent and, thus, plays an important role in the expression of certain genetic disorders. Two classic examples are the Prader-Willi syndrome and Angelman's syndrome. Prader-Willi syndrome is characterized by diminished fetal activity, obesity, hypotonia, mental retardation, short stature, and hypogonadotropic hypogonadism. Deletions of the paternal copy of the Prader-Willi locus located on the short arm of chromosome 15 result in a contiguous gene syndrome involving missing paternal copies of the *necdin* and *SNRPN* genes, among others. In contrast, patients with Angelman's syndrome, characterized by mental retardation, seizures, ataxia, and hypotonia, have deletions involving the maternal copy of this region on chromosome 15. These two syndromes may also result from *uniparental disomy*. In this case, the syndromes are not caused by deletions on chromosome 15 but by the inheritance of either two maternal chromosomes (Prader-Willi syndrome) or two paternal chromosomes (Angelman's syndrome). Lastly, the two distinct phenotypes can also be caused by an imprinting defect that impairs the resetting of the imprint during zygote development (defect in the father leads to Prader-Willi syndrome; defect in the mother leads to Angelman's syndrome).

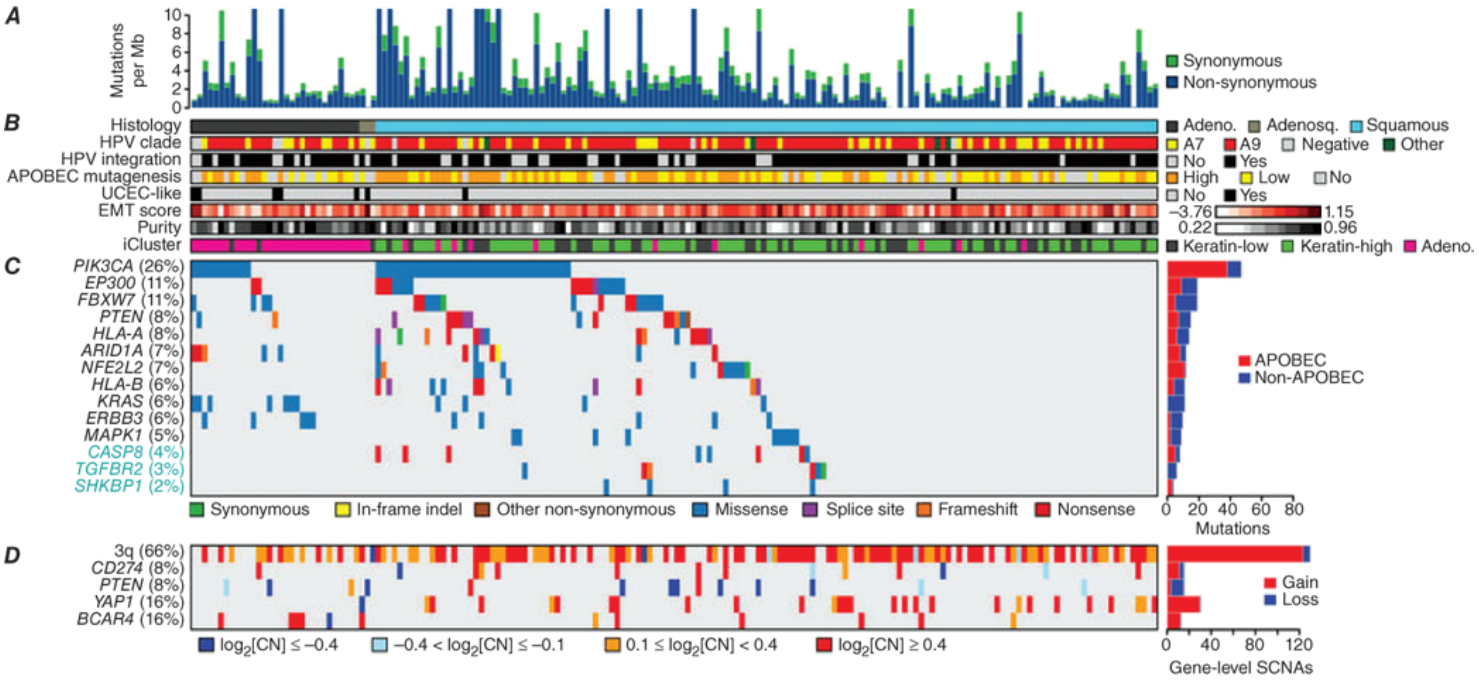
Imprinting and the related phenomenon of allelic exclusion may be more common than currently documented, because it is difficult to examine levels of mRNA expression from the maternal and paternal alleles in specific tissues or in individual cells. Genomic imprinting, or uniparental disomy, is involved in the pathogenesis of several other disorders and malignancies. For example, hydatidiform moles contain a normal number of diploid chromosomes, but they are all of paternal origin. The opposite situation occurs in ovarian teratomata, with 46 chromosomes of maternal origin. Expression of the imprinted gene for insulin-like growth factor II (IGF-II) is involved in the pathogenesis of the cancer-predisposing Beckwith-Wiedemann syndrome (BWS). These children show somatic overgrowth with organomegalies and hemihypertrophy, and they have an increased risk of embryonal malignancies such as Wilms' tumor. Normally, only the paternally derived copy of the *IGF-II* gene is active and the maternal copy is inactive. Imprinting of the *IGF-II* gene is regulated by *H19*, which encodes an RNA transcript that is not translated into protein. Disruption or lack of *H19* methylation leads to a relaxation of *IGF-II* imprinting and expression of both alleles. Alterations of the epigenome through gain and loss of DNA methylation, as well as altered histone modifications, play an important role in the pathogenesis of malignancies.

SOMATIC MUTATIONS

Cancer can be considered a genetic disease at the cellular level **(Chap. 67)**. Cancers are monoclonal in origin, indicating that they have arisen from a single precursor cell with one or several mutations in genes controlling growth (proliferation or apoptosis) and/or differentiation. These acquired somatic mutations are restricted to the tumor and its metastases and are not found in the surrounding normal tissue. The molecular alterations include dominant gain-of-function mutations in oncogenes, recessive loss-of-function mutations in tumor-suppressor genes and DNA repair genes, gene amplification, and chromosome rearrangements. Rarely, a single mutation in certain genes may be sufficient to transform a normal cell into a malignant cell. In most cancers, however, the development of a malignant phenotype requires several genetic alterations for the gradual progression from a normal cell to a cancerous cell, a phenomenon termed *multistep carcinogenesis*. Genome-wide analyses of cancers using deep sequencing often reveal somatic rearrangements resulting in fusion genes and mutations in multiple genes **(Fig. 456-14)**. Comprehensive sequence analyses provide further insight into genetic heterogeneity within malignancies; these include intratumoral heterogeneity among the cells of the primary tumor, intermetastatic and intrametastatic heterogeneity, and interpatient differences. These analyses further support the notion of cancer as an ongoing process of clonal evolution, in which successive rounds of clonal selection within the primary tumor and metastatic lesions result in diverse genetic and epigenetic alterations that require targeted (personalized) therapies (precision medicine). The heterogeneity of mutations within a tumor can also lead to resistance to target therapies because cells with mutations that are resistant to the therapy, even if they are a minor part of the tumor population, will be selected as the more sensitive cells are killed. Most human tumors express telomerase, an enzyme formed of a protein and an RNA component, which adds telomere repeats at the ends of chromosomes during replication. This mechanism impedes shortening of the telomeres, which is associated with senescence in normal cells and is associated with enhanced replicative capacity in cancer cells. Telomerase inhibitors provide a strategy for treating advanced human cancers.

FIGURE 456-14

Somatic alterations in cervical cancer. ***A.*** Cervical carcinoma samples ordered by histology and mutation frequency; ***B.*** clinical and molecular platform features; ***C.*** Significantly Mutated Genes (SMGs); and ***D.*** select somatic copy number alterations. SMGs are ordered by the overall mutation frequency and color-coded by mutation type. Adeno, adenocarcinomas; Adenosq, adenosquamous cancers; CN, copy number; SCNAs, somatic copy number alterations; Squamous, squamous cell carcinomas. *(From The Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. Nature 543:378–384, 2017. Permission for reproduction granted through a Creative Commons CC-BY (CC BY 4.0) license.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In many cancer syndromes, there is frequently an inherited *predisposition* to tumor formation. In these instances, a germline mutation is inherited in an autosomal dominant fashion inactivating one allele of an autosomal tumor-suppressor gene. If the second allele is inactivated by a somatic mutation or by epigenetic silencing in a given cell, this will lead to neoplastic growth (Knudson two-hit model). Thus, the defective allele in the germline is transmitted in a dominant mode, although tumorigenesis results from a biallelic loss of the tumor-suppressor gene in an affected tissue. The classic example to illustrate this phenomenon is retinoblastoma, which can occur as a sporadic or hereditary tumor. In sporadic retinoblastoma, both copies of the retinoblastoma (*RB*) gene are inactivated through two somatic events. In hereditary retinoblastoma, one mutated or deleted *RB* allele is inherited in an autosomal dominant manner and the second allele is inactivated by a subsequent somatic mutation. This two-hit model applies to other inherited cancer syndromes such as MEN 1 **(Chap. 381)** and neurofibromatosis type 2 **(Chap. 86)**. In contrast, in the autosomal dominant MEN2 syndrome, the predisposition for tumor formation in various organs is caused by a gain-of-function mutation in a single allele of the *RET* gene (Chap. 381).

NUCLEOTIDE REPEAT EXPANSION DISORDERS

Several diseases are associated with an increase in the number of nucleotide repeats above a certain threshold **(Table 456-5)**. The repeats are sometimes located within the coding region of the genes, as in Huntington's disease or the X-linked form of spinal and bulbar muscular atrophy (SBMA; Kennedy's syndrome). In other instances, the repeats probably alter gene regulatory sequences. If an expansion is present, the DNA fragment is unstable and tends to expand further during cell division. The length of the nucleotide repeat often correlates with the severity of the disease. When repeat length increases from one generation to the next, disease manifestations may worsen or be observed at an earlier age; this phenomenon is referred to as *anticipation*. In Huntington's disease, for example, there is a correlation between age of onset and length of the triplet codon expansion **(Chap. 417)**. Anticipation has also been documented in other diseases caused by dynamic mutations in trinucleotide repeats (Table 456-5). The repeat number may also vary in a tissue-specific manner. In myotonic dystrophy, the CTG repeat may be tenfold greater in muscle tissue than in lymphocytes **(Chap. 441)**.

TABLE 456-5

Selected Trinucleotide Repeat Disorders

Disease	Locus	Repeat	Triplet Length (Normal/Disease)	Inheritance	Gene Product
X-chromosomal spinobulbar muscular atrophy (SBMA)	Xq11-q12	CAG	11–34/40–62	XR	Androgen receptor
Fragile X syndrome (FRAXA)	Xq27.3	CGG	6–50/200–300	XR	FMR-1 protein
Fragile X syndrome (FRAXE)	Xq28	GCC	6–25/>200	XR	FMR-2 protein
Dystrophia myotonica (DM)	19q13.2-q13.3	CTG	5–30/200–1000	AD, variable penetrance	Myotonin protein kinase
Huntington's disease (HD)	4p16.3	CAG	6–34/37–180	AD	Huntingtin
Spinocerebellar ataxia type 1 (SCA1)	6p21.3-21.2	CAG	6–39/40–88	AD	Ataxin 1
Spinocerebellar ataxia type 2 (SCA2)	12q24.1	CAG	15–31/34–400	AD	Ataxin 2
Spinocerebellar ataxia type 3 (SCA3); Machado-Joseph disease (MD)	14q21	CAG	13–36/55–86	AD	Ataxin 3
Spinocerebellar ataxia type 6 (SCA6, CACNAIA)	19p13.1-13.2	CAG	4–16/20–33	AD	Alpha 1A voltage-dependent L-type calcium channel
Spinocerebellar ataxia type 7 (SCA7)	3p21.1-p12	CAG	4–19/37 to >300	AD	Ataxin 7
Spinocerebellar ataxia type 12 (SCA12)	5q31	CAG	6–26/66–78	AD	Protein phosphatase 2A
Dentatorubral pallidoluysian atrophy (DRPLA)	12p	CAG	7–23/49–75	AD	Atrophin 1
Friedreich's ataxia (FRDA1)	9q13-21	GAA	7–22/200–900	AR	Frataxin

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; XR, X-linked recessive.

Complex Genetic Disorders

The expression of many common diseases such as cardiovascular disease, hypertension, diabetes, asthma, psychiatric disorders, and certain cancers is determined by a combination of genetic background, environmental factors, and lifestyle. A trait is called *polygenic* if multiple genes contribute to the phenotype or *multifactorial* if multiple genes are assumed to interact with environmental factors. Genetic models for these complex traits need to account for genetic heterogeneity and interactions with other genes and the environment. Complex genetic traits may be influenced by modifier genes that are not linked to the main gene involved in the pathogenesis of the trait. This type of gene-gene interaction, or *epistasis*, plays an important role in polygenic traits that require the simultaneous presence of variations in multiple genes to result in a pathologic phenotype.

Type 2 diabetes mellitus provides a paradigm for considering a multifactorial disorder, because genetic, nutritional, and lifestyle factors are intimately interrelated in disease pathogenesis **(Table 456-6) (Chap. 396)**. The identification of genetic variations and environmental factors that either predispose to or protect against disease is essential for predicting disease risk, designing preventive strategies, and developing novel therapeutic approaches. The study of rare monogenic diseases may provide insight into some of the genetic and molecular mechanisms important in the pathogenesis of complex diseases. For example, the identification of the genes causing monogenic forms of permanent neonatal diabetes mellitus or maturity-onset diabetes defined them as *candidate genes* in the pathogenesis of diabetes mellitus type 2 (Tables 456-2 and 456-6) **(Fig. 456-15)**. Genome scans have identified numerous genes and loci that may be associated with susceptibility to development of diabetes mellitus in certain populations **(Fig. 456-16)**. Efforts to identify susceptibility genes require very large sample sizes, and positive results may depend on ethnicity, ascertainment criteria, and statistical analysis. Association studies analyzing the potential influence of (biologically functional) SNPs and SNP haplotypes on a particular phenotype have revealed new insights into the genes

involved in the pathogenesis of these common disorders. Large variants ([micro]deletions, duplications, and inversions) present in the human population also contribute to the pathogenesis of complex disorders, but their contributions remain poorly understood.

TABLE 456-6

Genes and Loci Involved in Mono- and Polygenic Forms of Diabetes

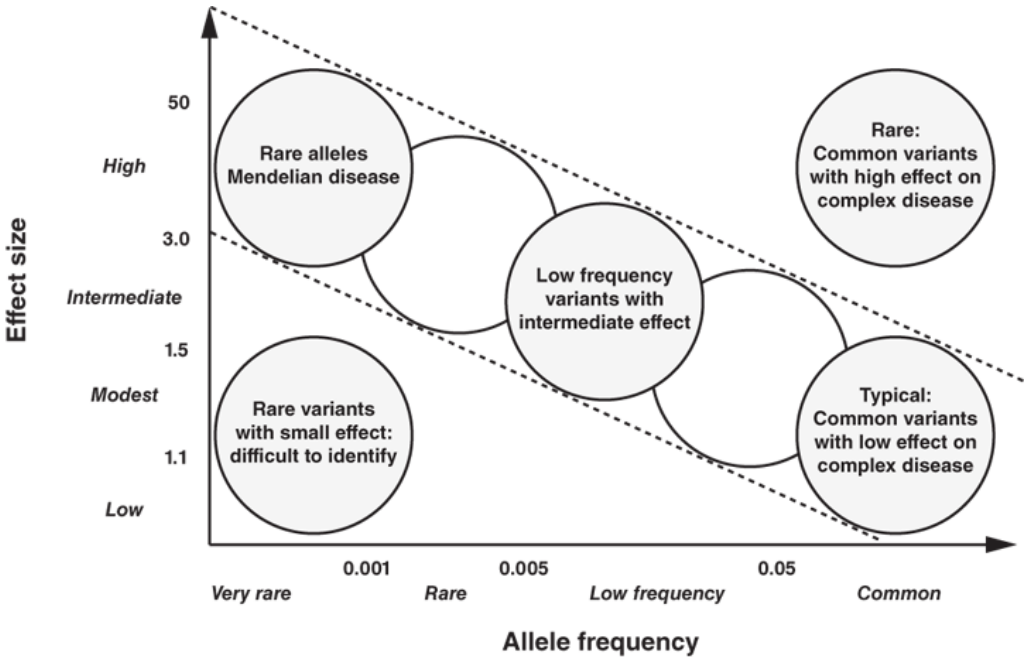
Disorder	Genes or Susceptibility Locus	Chromosomal Location	Other Factors
Monogenic permanent neonatal diabetes mellitus	*KCNJ11* (inwardly rectifying potassium channelKir6.2)	11p15.1	AD
	GCK (glucokinase)	7p15-p13	AR
	INS (insulin)	11p15.5	AR, hyperproinsulinemia
	ABCC8 (ATP-binding cassette, subfamily c, member 8; sulfonylurea receptor)	11p15.1	AD or AR
	GLIS3 (GLIS family zinc finger protein 3)	9p24.2	AR, diabetes, congenital hypothyroidism
Maturity-onset diabetes of the young (MODY): Monogenic forms of diabetes mellitus			
MODY 1	*HNF4α* (hepatocyte nuclear factor 4α)	20q12-q13.1	AD inheritance
MODY 2	*GCK* (glucokinase)	7p15-p13	
MODY 3	*HNF1α* (hepatocyte nuclear factor 1α)	12q24.2	
MODY 4	*IPF1* (insulin receptor substrate)	13q12.1	
MODY 5 (renal cysts, diabetes)	*HNF1β* (hepatocyte nuclear factor 1β)	17cen-q21.3	
MODY 6	*NeuroD1* (neurogenic differentiation factor 1)	2q32	
MODY 7	*KLF1* (Kruppel-like factor 1)	19p13.13-p13.12	
MODY 8	*CEL* (carboxyl ester lipase)	9q34.3	
MODY 9	*PAX4* (paired box transcription factor 4)	7q32	
MODY 10	*INS* (insulin)	11p15.5	
MODY 11	*BLK* (B-lymphocyte-specific tyrosine kinase)	8p23-p22	

Disorder	Genes or Susceptibility Locus	Chromosomal Location	Other Factors
Diabetes mellitus type 2; loci and genes linked and/or associated with susceptibility for diabetes mellitus type 2	Genes and loci identified by linkage/association studies		Heavily influenced by diet, energy expenditure, obesity
	PPARG, KCNJ11/ABCC8, TCF7L2, HNF1B, WFS1, SLC30A8, FTO, HHEX, IGF2BP2, CDKN2A/B, CDKAL1, TSPAN8, ADAMTs9, CDC123/CAMK1D, JAZF1, NOTCH2, THADA, KCNQ1, DUSP8, MTNR1B, IRS1, SPRY2, SRR, ZFAND6, GCK, KLF14, TP53INP1, PROX1, PRC1, BCL11A, ZBED3, RBMS1, HNF1A, DGKB/TMEM195, CCND2, C2CD4A/C2CD4B, PTPRD, ARAP1/CENTD2, HMGA2, TLE4/CHCHD9, ADCY5, UBE2E2, DUSP9, GCKR, COBLL1/GRB14, HMG20A, VPS26A, ST6GAL1, AP3S2, HNF4A, BCL2, LAMA1, GIPR, MC4R, TLE1, KCNK16, ANK1, KLHDC5, ZMIZ1, PSMD6 , FITM2/R3HDML/HNF4A, CILP2, ANKRD55, GLIS3, PEPD, GCC1/PAX4, ZFAND3, MAEA, BCAR1, RBM43/RND3 , MACF1, RASGRP1, GRK5, TMEM163, SGCG, LPP, FAF1, TMEM154, MPHOSPH9, ARL15, POU5F1/TCF19, SSR1/RREB1, HLA-B, INS-IGF2, GPSM1, LEP, SLC16A13, PAM/PPIP5K2, SLC16A11, CCDC63, C12orf51, CCND2, HNF1A, TBC1D4, CCDC85A, INAFM2, ASB3, FAM60A, ATP8B2, MIR4686, MTMR3, DMRTA1, SLC35D3, GLP2R, GIP, MAP3K11, PLEKHA1, HSD17B12, NRXN3, CMIP, ZZEF1, MNX1, ABO, ACSL1, HLA-DQA1		

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; MODY, maturity onset diabetes of the young.

FIGURE 456-15

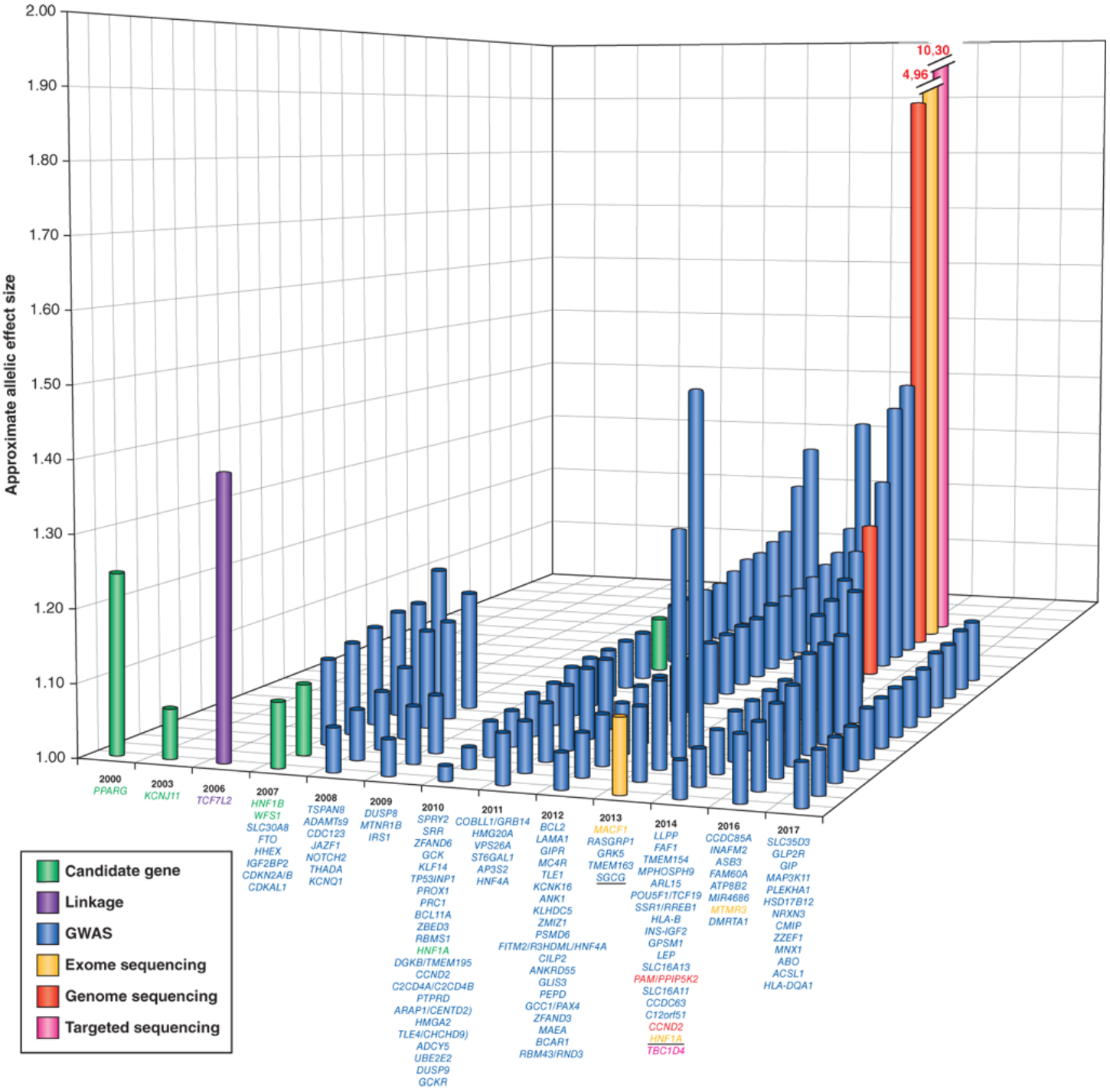
Relationship between allele frequency and effect size in monogenic and polygenic disorders. In classic Mendelian disorders, the allele frequency is typically low but has a high impact (single gene disorder). This contrasts with polygenic disorders that require the combination of multiple low impact alleles that are frequently quite common in the general population.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 456-16

Loci and genes associated with diabetes mellitus type 2. Loci are listed by year of identification, and the color indicates discovery method. Gene names indicate the locus and does not necessarily imply that the gene itself is causally involved. Approximate allelic effect sizes were either derived from the discovery cohort or, if available, from the DIAGRAM (Diabetes Genetics Replication and Meta-analysis consortium) European ancestry meta-analysis and the Asian ancestry meta-analysis. Gene names that are underlined denote identification in population isolates. *(The data have been graciously provided by Dr. Miriam Udler and Dr. Jose Florez, Harvard Medical School, Boston.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Linkage and Association Studies

There are two primary strategies for mapping genes that cause or increase susceptibility to human disease: (1) classic linkage can be performed based on a known genetic model or, when the model is unknown, by studying pairs of affected relatives; or (2) disease genes can be mapped using allelic association studies (Table 456-7).

TABLE 456-7

Genetic Approaches for Identifying Disease Genes

Method	Indications and Advantages	Limitations
Linkage Studies		
Classical linkage analysis (parametric methods)	Analysis of monogenic traits	Difficult to collect large informative pedigrees
	Suitable for genome scan	Difficult to obtain sufficient statistical power for complex traits
	Control population not required	
	Useful for multifactorial disorders in isolated populations	
Allele-sharing methods (nonparametric methods)	Suitable for identification of susceptibility genes in polygenic and multifactorial disorders	Difficult to collect sufficient number of subjects
Affected sib and relative pair analyses	Suitable for genome scan	Difficult to obtain sufficient statistical power for complex traits
Sib pair analysis	Control population not required if allele frequencies are known	Reduced power compared to classical linkage, but not sensitive to specification of genetic mode
	Statistical power can be increased by including parents and relatives	
Association Studies		
Case-control studies	Suitable for identification of susceptibility genes in polygenic and multifactorial disorders	Requires large sample size and matched control population
Linkage disequilibrium	Suitable for testing specific allelic variants of known candidate loci	False-positive results in the absence of suitable control population
Transmission disequilibrium test (TDT)	Facilitated by comprehensive catalogs of genotypes and variants	Candidate gene approach does not permit detection of novel genes and pathways
Whole-genome association studies	Does not necessarily need relatives	Susceptibility genes can vary among different populations
Next-Generations Sequencing Technologies		
Whole exome or genome sequencing	Unbiased approach, analysis can be performed without reference sequences from parents or siblings	Requires appropriate bioinformatics, may have low sensitivity if CNV analysis is not included, detects numerous VUS, can lead to the detection of unrelated deleterious alleles
Targeted sequencing of gene panels	Captures multiple candidate genes and loci with hybridization techniques followed by deep sequencing	Permits analyses of multiple candidate genes in parallel. Facilitates molecular characterization of disorders with locus heterogeneity.

Abbreviations: CNV, copy number variation; GWAS, genome-wide association study; VUS, variants of unknown significance.

GENETIC LINKAGE

Genetic linkage refers to the fact that genes are physically connected, or linked, to one another along the chromosomes. Two fundamental principles are essential for understanding the concept of linkage: (1) when two genes are close together on a chromosome, they are usually transmitted together, unless a recombination event separates them (Figs. 456-6); and (2) the odds of a crossover, or recombination event, between two linked genes is proportional to the distance that separates them. Thus, genes that are farther apart are more likely to undergo a recombination event than genes that are very close together. The detection of chromosomal loci that segregate with a disease by linkage can be used to identify the gene responsible for the disease (*positional cloning*) and to predict the odds of disease gene transmission in genetic counseling.

Polymorphic variants are essential for linkage studies because they provide a means to distinguish the maternal and paternal chromosomes in an individual. On average, 1 out of every 1000 bp varies from one person to the next. Although this degree of variation seems low (99.9% identical), it means that >3 million sequence differences exist between any two unrelated individuals and the probability that the sequence at such loci will differ on the two homologous chromosomes is high (often >70–90%). These sequence variations include variable number of tandem repeats (VNTRs), short tandem repeats (STRs), and SNPs. Most STRs, also called *polymorphic microsatellite markers*, consist of di-, tri-, or tetranucleotide repeats that can be characterized readily using the polymerase chain reaction (PCR). Characterization of SNPs, using DNA chips or beads, permits comprehensive analyses of genetic variation, linkage, and association studies. Although these sequence variations often have no apparent functional consequences, they provide much of the basis for variation in genetic traits.

In order to identify a chromosomal locus that segregates with a disease, it is necessary to characterize polymorphic DNA markers from affected and unaffected individuals of one or several pedigrees. One can then assess whether certain marker alleles cosegregate with the disease. Markers that are closest to the disease gene are less likely to undergo recombination events and therefore receive a higher linkage score. Linkage is expressed as a lod (logarithm of odds) score—the ratio of the probability that the disease and marker loci are linked rather than unlinked. Lod scores of +3 (1000:1) are generally accepted as supporting linkage, whereas a score of –2 is consistent with the absence of linkage.

ALLELIC ASSOCIATION, LINKAGE DISEQUILIBRIUM, AND HAPLOTYPES

Allelic association refers to a situation in which the frequency of an allele is significantly increased or decreased in individuals affected by a particular disease in comparison to controls. Linkage and association differ in several aspects. Genetic linkage is demonstrable in families or sibships. Association studies, on the other hand, compare a population of affected individuals with a control population. Association studies can be performed as case-control studies that include unrelated affected individuals and matched controls or as family-based studies that compare the frequencies of alleles transmitted or not transmitted to affected children.

Allelic association studies are particularly useful for identifying susceptibility genes in complex diseases. When alleles at two loci occur more frequently *in combination* than would be predicted (based on known allele frequencies and recombination fractions), they are said to be in *linkage disequilibrium*. Evidence for linkage disequilibrium can be helpful in mapping disease genes because it suggests that the two loci are tightly linked.

Detecting the genetic factors contributing to the pathogenesis of common complex disorders is challenging. In many instances, these are low-penetrance alleles (e.g., variations that individually have a subtle effect on disease development, and they can only be identified by unbiased GWAS) (Catalog of published Genome-Wide Association Studies; Table 456-1) (Fig. 456-16). Most variants occur in noncoding or regulatory sequences but do not alter protein structure. The analysis of complex disorders is further complicated by ethnic differences in disease prevalence, differences in allele frequencies in known susceptibility genes among different populations, locus and allelic heterogeneity, gene-gene and gene-environment interactions, and the possibility of phenocopies. Catalogues of human variation and genotype data (HapMap, International Genome Sample Resource) are greatly facilitating GWAS for the characterization of complex disorders. Adjacent SNPs are inherited together as blocks, and these blocks can be identified by genotyping selected marker SNPs, so-called *Tag SNPs*, thereby reducing cost and workload (Fig. 456-4). The availability of this information permits the characterization of a limited number of SNPs to identify the set of haplotypes present in an individual (e.g., in cases and controls). This, in turn, permits performing GWAS by searching for associations of certain haplotypes with a disease phenotype of interest, an essential step for unraveling the genetic factors contributing to complex disorders.

POPULATION GENETICS

In population genetics, the focus changes from alterations in an individual's genome to the distribution pattern of different genotypes in the population. In a case where there are only two alleles, A and a, the frequency of the genotypes will be $p^2 + 2pq + q^2 = 1$, with p^2 corresponding to the frequency of AA, $2pq$ to the frequency of Aa, and q^2 to aa. When the frequency of an allele is known, the frequency of the genotype can be calculated. Alternatively, one can determine an allele frequency if the genotype frequency has been determined.

Allele frequencies vary among ethnic groups and geographic regions. For example, heterozygous mutations in the *CFTR* gene are relatively common in populations of European origin but are rare in the African population. Allele frequencies may vary because certain allelic variants confer a selective advantage. For example, heterozygotes for the sickle cell mutation, which is particularly common in West Africa, are more resistant to malarial infection because the erythrocytes of heterozygotes provide a less favorable environment for *Plasmodium* parasites. Although homozygosity for the sickle cell mutation is associated with severe anemia and sickle crises, heterozygotes have a higher probability of survival because of the reduced morbidity and

mortality from malaria; this phenomenon has led to an increased frequency of the mutant allele. Recessive conditions are more prevalent in geographically isolated populations because of the more restricted gene pool.

APPROACH TO THE PATIENT

APPROACH TO THE PATIENT

Inherited Disorders

For the practicing clinician, the family history remains an essential step in recognizing the possibility of a hereditary predisposition to disease. When taking the history, it is useful to draw a detailed pedigree of the first-degree relatives (e.g., parents, siblings, and children), because they share 50% of genes with the patient. Standard symbols for pedigrees are depicted in Fig. 456-11. The family history should include information about ethnic background, age, health status, and deaths, including infants. Next, the physician should explore whether there is a family history of the same or related illnesses to the current problem. An inquiry focused on commonly occurring disorders such as cancers, heart disease, and diabetes mellitus should follow. Because of the possibility of age-dependent expressivity and penetrance, the family history will need intermittent updating. If the findings suggest a genetic disorder, the clinician should assess whether some of the patient's relatives may be at risk of carrying or transmitting the disease. In this circumstance, it is useful to confirm and extend the pedigree based on input from several family members. This information may form the basis for genetic counseling, carrier detection, early intervention, and disease prevention in relatives of the index patient **(Chap. 457)**.

In instances where a diagnosis at the molecular level may be relevant, it is important to identify an appropriate laboratory that can perform the appropriate test. Genetic testing is available for a large number of monogenic disorders through commercial laboratories. For uncommon disorders, the test may only be performed in a specialized research laboratory. Approved laboratories offering testing for inherited disorders can be identified in continuously updated online resources (e.g., GeneTests; Table 456-1). If genetic testing is considered, the patient and the family should be counseled about the potential implications of positive results, including psychological distress and the possibility of discrimination. The patient or caretakers should be informed about the meaning of a negative result, technical limitations, and the possibility of false-negative and inconclusive results. For these reasons, genetic testing should only be performed after obtaining *informed consent*. Published ethical guidelines address the specific aspects that should be considered when testing children and adolescents. Genetic testing should usually be limited to situations in which the results may have an impact on medical management.

IDENTIFYING THE DISEASE-CAUSING GENE

Precision medicine aims to enhance the quality of medical care through the use of genotypic analysis (DNA testing) to identify genetic predisposition to disease, to select more specific pharmacotherapy, and to design individualized medical care based on genotype. Genotype can be deduced by analysis of protein (e.g., hemoglobin, apoprotein E), mRNA, or DNA. However, technologic advances have made DNA analysis particularly useful because it can be readily applied.

DNA testing is performed by mutational analysis or linkage studies in individuals at risk for a genetic disorder known to be present in a family. Mass screening programs require tests of high sensitivity and specificity to be cost-effective. Prerequisites for the success of genetic screening programs include the following: that the disorder is potentially serious; that it can be influenced at a presymptomatic stage by changes in behavior, diet, and/or pharmaceutical manipulations; and that the screening does not result in any harm or discrimination. Screening in Jewish populations for the autosomal recessive neurodegenerative storage disease Tay-Sachs has reduced the number of affected individuals. In contrast, screening for sickle cell trait/disease in African Americans has led to unanticipated problems of discrimination by health insurers and employers. Mass screening programs harbor additional potential problems. For example, screening for the most common genetic alteration in cystic fibrosis, the ΔF508 mutation with a frequency of ~70% in northern Europe, is feasible and seems to be effective. One has to keep in mind, however, that there is pronounced allelic heterogeneity and that the disease can be caused by about 2000 other mutations. The search for these less common mutations would substantially increase costs, but not the effectiveness of the screening program as a whole. Next-generation genome sequencing permits comprehensive and cost-effective mutational analyses after selective enrichment of candidate genes. For example, tests that sequence all the common genes causing hereditary deafness or hereditary pheochromocytomas are commercially available. Occupational screening programs aim to detect individuals with increased risk for certain professional activities (e.g., α_1 antitrypsin deficiency and smoke or dust exposure). Integrating genomic data into electronic medical records is evolving and can provide significant decision support at the point of care, for example, by providing the clinician with genomic data and decision algorithms for the prescription of drugs that are subject to pharmacogenetic influences.

Mutational Analyses

DNA sequence analysis is now widely used as a diagnostic tool and has significantly enhanced diagnostic accuracy. It is used for determining carrier status and for prenatal testing in monogenic disorders. Numerous techniques, discussed in previous versions of this chapter, are available for the detection of mutations. In a very broad sense, one can distinguish between techniques that allow for screening of known mutations (screening mode) or techniques that definitively characterize mutations. Analyses of large alterations in the genome are possible using classic methods such as karyotype analysis, cytogenetics, fluorescent in situ hybridization (FISH), as well as more sensitive array- or bead-based techniques that search for multiple single exon deletions or duplications.

More discrete sequence alterations rely heavily on the use of PCR, which allows rapid gene amplification and analysis. Moreover, PCR makes it possible to perform genetic testing and mutational analysis with small amounts of DNA extracted from leukocytes or even from single cells, buccal cells, or hair roots.

DNA sequencing can be performed directly on PCR products or on fragments cloned into plasmid vectors amplified in bacterial host cells. Sequencing of the whole genome, exome, selected chromosomes, or sequencing of numerous candidate genes in a single run, is now possible with next-generation sequencing platforms and has entered the clinical realm. Genomic tests are also widely used for the detection of pathogens and for the identification of viral or bacterial sequence variations. The integration of genomic tests into clinical medicine is, however, associated with a number of ongoing challenges related to costs, variable sensitivities of the tests, bioinformatics analyses, storage and sharing of data, and the difficulty of interpreting all genetic variants identified with comprehensive testing. The discovery of incidental (or secondary) findings that are unrelated to the indication for the sequencing analysis but indicators of other disorders of potential relevance for patient care can pose a difficult ethical dilemma. It can lead to the detection of undiagnosed medically actionable genetic conditions, but can also reveal deleterious mutations that cannot be influenced, as numerous sequence variants are of unknown significance.

A general algorithm for the approach to mutational analysis is outlined in **Fig. 456-17**. The importance of a detailed clinical phenotype cannot be overemphasized. This is the step where one should also consider the possibility of genetic heterogeneity and phenocopies. If obvious candidate genes are suggested by the phenotype, they can be analyzed directly. After identification of a mutation, it is essential to demonstrate that it segregates with the phenotype. The functional characterization of novel mutations is labor intensive and may require analyses in vitro or in transgenic models in order to document the relevance of the genetic alteration.

Prenatal diagnosis of numerous genetic diseases in instances with a high risk for certain disorders is now possible by direct DNA analysis. *Amniocentesis* involves the removal of a small amount of amniotic fluid, usually at 16 weeks of gestation. Cells can be collected and submitted for karyotype analyses, FISH, and mutational analysis of selected genes Table 456-4). The main indications for amniocentesis include advanced maternal age (>35 years), presence of an abnormality of the fetus on ultrasound examination, an abnormal serum “quad” test (α-fetoprotein, β human chorionic gonadotropin, inhibin-A, and unconjugated estriol), a family history of chromosomal abnormalities, or a Mendelian disorder amenable to genetic testing. Prenatal diagnosis can also be performed by *chorionic villus sampling* (CVS), in which a small amount of the chorion is removed by a transcervical or transabdominal biopsy. Chromosomes and DNA obtained from these cells can be submitted for cytogenetic and mutational analyses. CVS can be performed earlier in gestation (weeks 9–12) than amniocentesis, an aspect that may be of relevance when termination of pregnancy is a consideration. Later in pregnancy, beginning at about 18 weeks of gestation, percutaneous umbilical blood sampling (PUBS) permits collection of fetal blood for lymphocyte culture and analysis. These approaches enable screening for clinically relevant and deleterious alleles inherited from the parents, as well as for de novo germline mutations, and they may have the potential to change the diagnosis of genetic disorders in the prenatal setting. Although genomic sequencing of fetal DNA circulating in the bloodstream of the mother has been achieved, it is not commonly used for non-invasive prenatal testing.

In combination with in vitro fertilization (IVF) techniques, it is possible to perform genetic diagnoses in a single cell removed from the four- to eight-cell embryo or to analyze the first polar body from an oocyte. Preconceptual diagnosis thereby avoids therapeutic abortions but is costly and labor intensive. It should be emphasized that excluding a specific disorder by any of these approaches is never equivalent to the assurance of having a normal child. Postnatal indications for cytogenetic analyses in infants or children include multiple congenital anomalies, suspicion of a known cytogenetic syndrome, developmental delay, dysmorphic features, autism, short stature and disorders of sexual development, among others (Table 456-4).

Mutations in certain cancer susceptibility genes such as *BRCA1* and *BRCA2* may identify individuals with an increased risk for the development of malignancies and result in risk-reducing interventions. The detection of cytogenetic alterations and mutations is an important diagnostic and prognostic tool in leukemias and somatic mutational analysis is transforming oncology by providing diagnostic and prognostic information, and it informs the choice of targeted therapies. However, the cancer may recur due to treatment resistance of subclones due to the continuously evolving genomic landscape. The demonstration of the presence or absence of mutations and polymorphisms is also relevant for the rapidly evolving field of pharmacogenomics, including the identification of differences in drug treatment response or metabolism as a function of genetic background. For example, the thiopurine drugs 6-mercaptopurine and azathioprine are commonly used cytotoxic and immunosuppressive agents. They are metabolized by thiopurine methyltransferase (TPMT), an enzyme with variable activity associated with genetic polymorphisms in 10% of whites and complete deficiency in about 1 in 300 individuals. Patients with intermediate or deficient TPMT activity are at risk for excessive toxicity, including fatal myelosuppression. Characterization of these polymorphisms allows mercaptopurine doses to be modified based on TPMT genotype. Pharmacogenomics may increasingly permit individualized drug therapy, improve drug effectiveness, reduce adverse side effects, and provide cost-effective pharmaceutical care **(Chap. 64)**.

ETHICAL ISSUES

Determination of the association of genetic defects with disease, comprehensive data of an individual’s genome, and studies of genetic variation raise many ethical and legal issues. Genetic information is generally regarded as sensitive information that should not be readily accessible without explicit consent (*genetic privacy*). The disclosure of genetic information may risk possible discrimination by insurers or employers. The scientific components of the Human Genome Project have been paralleled by efforts to examine ethical, social, and legal implications. An important milestone emerging from these endeavors consists in the Genetic Information Nondiscrimination Act (GINA), signed into law in 2008, which aims to protect asymptomatic individuals against the misuse of genetic information for health insurance and employment. It does not, however, protect the symptomatic individual. Provisions of the U.S. Patient Protection and Affordable Care Act, effective in 2014, have, in part, filled this gap and prohibit exclusion from, or termination of, health insurance based on personal health status. Potential threats to the maintenance of genetic privacy consist in the emerging integration of genomic data into electronic medical records, compelled disclosures of health records, and direct-to-consumer genetic testing.

It is widely accepted that identifying disease-causing genes can lead to improvements in diagnosis, treatment, and prevention. However, the information gleaned from genotypic results can have quite different impacts, depending on the availability of strategies to modify the course of disease. For example, the identification of mutations that cause MEN 2 or hemochromatosis allows specific interventions for affected family members. On the other hand, at present, the identification of an Alzheimer's or Huntington's disease gene does not currently alter therapy and outcomes. Most genetic disorders are likely to fall into an intermediate category where the opportunity for prevention or treatment is significant but limited. However, the progress in this area is unpredictable, as underscored by the finding that angiotensin II receptor blockers may slow disease progression in Marfan's syndrome. Genetic test results can generate anxiety in affected individuals and family members. Comprehensive sequence analyses are particularly challenging because most individuals can be expected to harbor several serious recessive gene mutations. Moreover, the sensitivity of comprehensive sequence analyses are not always greater, for example, if CNV analysis is not integrated, but can be associated with higher costs.

The impact of genetic testing on health care costs remain unclear. It does vary among disorders and depends on the availability of effective therapeutic modalities. A significant problem arises from the marketing of genetic testing directly to consumers by commercial companies. The validity of these tests is, in part, not well defined, and there are persisting concerns about the lack of appropriate regulatory oversight, the accuracy and confidentiality of genetic information, the availability of counseling, and the handling of these results.

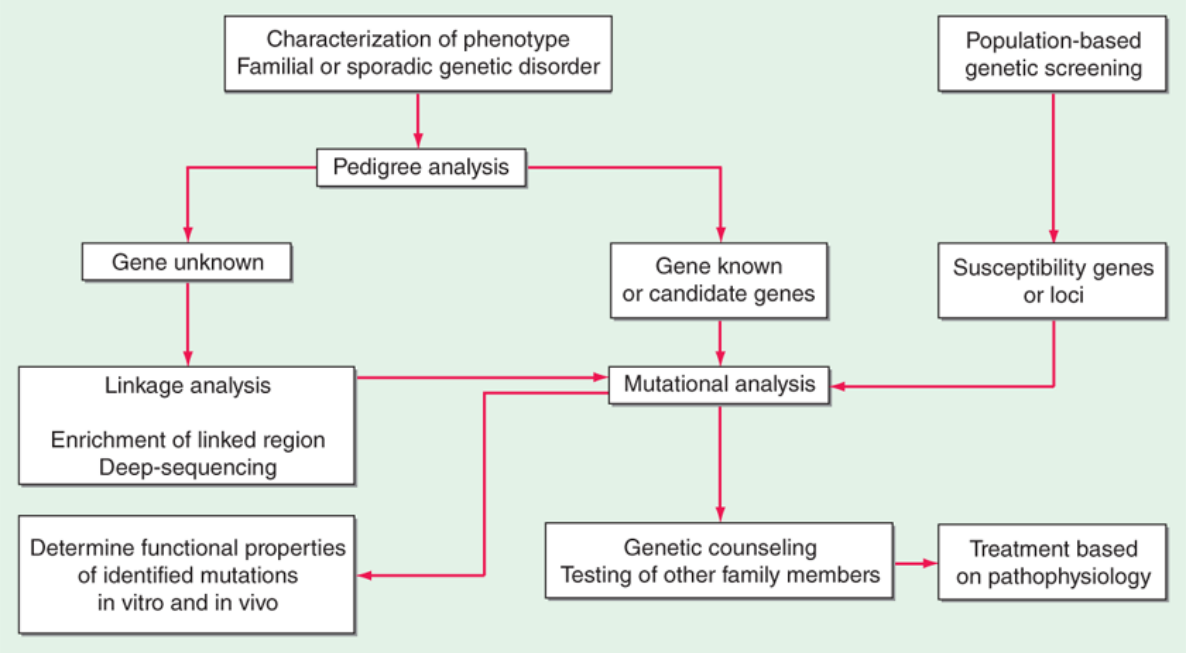
Many issues raised by the genome project are familiar, in principle, to medical practitioners. For example, an asymptomatic patient with increased low-density lipoprotein (LDL) cholesterol, high blood pressure, or a strong family history of early myocardial infarction is known to be at increased risk of coronary heart disease. In such cases, it is clear that the identification of risk factors and an appropriate intervention are beneficial. Likewise, patients with phenylketonuria, cystic fibrosis, or sickle cell anemia are often identified as having a genetic disease early in life. These precedents can be helpful for adapting policies that relate to genetic information. We can anticipate similar efforts, whether based on genotypes or other markers of genetic predisposition, to be applied to many disorders. One confounding aspect of the rapid expansion of information is that our ability to make clinical decisions often lags behind initial insights into genetic mechanisms of disease. For example, when genes that predispose to breast cancer such as *BRCA1* are described, they generate tremendous public interest in the potential to predict disease, but many years of clinical research are still required to rigorously establish genotype and phenotype correlations.

Genomics may contribute to improvements in global health by providing a better understanding of pathogens and diagnostics, and through contributions to drug development. There is, however, ongoing concern about the development of a "genomics divide" because of the costs associated with these developments and uncertainty as to whether these advances will be accessible to the populations of developing countries faced with pressing health needs associated with poverty, infectious diseases, and the relative lack of essential infrastructure. The World Health Organization has summarized these issues and inequities surrounding genomic medicine in a detailed report titled "Genomics and World Health."

Whether related to informed consent, participation in research, or the management of a genetic disorder that affects an individual or his or her family, there is a great need for more information about fundamental principles of genetics. The pervasive nature of the role of genetics in medicine makes it important for physicians and other health care professionals to become more informed about genetics and to provide advice and counseling in conjunction with trained genetic counselors (Chap. 457). The application of screening and prevention strategies does therefore require continuing patient and physician education, changes in health care financing, and legislation to protect patient's rights.

FIGURE 456-17

Approach to genetic disease.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Acknowledgment

Selected sections and Table 456-4 have been integrated from the chapter on Chromosome Disorders by Dr. Nancy B. Spinner and Dr. Laura K. Conlin, published in the 19th edition of Harrison's Principles in Internal Medicine. The data and concept for Figure 456-16 have been graciously provided by Dr. Miriam Udler and Dr. Jose Florez, Massachusetts General Hospital and Harvard Medical School, Boston.

FURTHER READING

Ashley EA: Towards precision medicine. Nat Rev Genet 17:507, 2016. [PubMed: 27528417]

Jameson JL, Longo DL: Precision medicine—personalized, problematic, and promising. N Engl J Med 372:2229, 2015. [PubMed: 26014593]

Jarvik GP, Evans JP: Mastering genomic terminology. Genet Med 19: 491, 2017. [PubMed: 27657676]

Joly Y et al: Comparative approaches to genetic discrimination: Chasing shadows? Trends Genet 33: 299, 2017. [PubMed: 28365141]

Manolio TA: Bringing genome-wide association findings into clinical use. Nat Rev Genet 14:549, 2013. [PubMed: 23835440]

Manolio TA et al: Bedside back to bench: Building bridges between basic and clinical genomic research. Cell 169:6, 2017. [PubMed: 28340351]

Rehm HL: Evolving health care through personal genomics. Nat Rev Genet 18:259, 2017. [PubMed: 28138143]

Richards S et al: ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405, 2015. [PubMed: 25741868]

Sidow A, Spies N: Concepts in solid tumor evolution. Trends Genet 31:208, 2015. [PubMed: 25733351]

Wauters A, Van Hoyweghen I: Global trends on fears and concerns of genetic discrimination: A systematic literature review. J Hum Genet 61:275, 2016. [PubMed: 26740237]

Wolf SM et al: Mapping the ethics of translational genomics: Situating return of results and navigating the research-clinical divide. J Law Med Ethics 43:486, 2015. [PubMed: 26479558]

World Health Organization: Genomics and World Health: Report of the Advisory Committee on Health Research. 1-254, 2002.

Zheng S et al: Comprehensive pan-genomic characterization of adrenocortical carcinoma. Cancer Cell 29:723, 2016. [PubMed: 27165744]

Chapter 457: The Practice of Genetics in Clinical Medicine

Susan M. Domchek; J. Larry Jameson; Susan Miesfeldt

APPLICATIONS OF MOLECULAR GENETICS IN CLINICAL MEDICINE

Genetic testing for inherited abnormalities associated with disease risk is increasingly used in the practice of clinical medicine. Germline alterations include chromosomal abnormalities specific gene mutations with autosomal dominant or recessive patterns of transmission (Chap. 456), and single nucleotide polymorphisms (SNPs) with small relative risks associated with disease. Germline alterations are responsible for disorders beyond classic Mendelian conditions with genetic susceptibility to common adult-onset diseases such as asthma, hypertension, diabetes mellitus, macular degeneration, and a number of types of cancer. For many of these diseases, there is a complex interplay of genes (often multiple) and environmental factors that effect lifetime risk, age of onset, disease severity, and treatment options.

The expansion of human genetic knowledge is changing our understanding of pathophysiology and influencing our classification of diseases. Awareness of genetic etiology can have an impact on clinical management, including prevention and screening for or treatment of a range of diseases. Primary care physicians are relied upon to help patients navigate testing and treatment options. Consequently, they must understand the genetic basis for a large number of genetically influenced diseases, incorporate personal and family history to determine the risk for a specific mutation, and be positioned to provide counseling. Even if patients are seen by genetic specialists who assess genetic risk and coordinate testing, primary care providers should offer information to their patients regarding the indications, limitations, risks, and benefits of genetic counseling and testing. They must also be prepared to offer risk-based management following genetic risk assessment. Given the pace of genetics, this is an increasingly difficult task. The field of clinical genetics has rapidly transitioned from single gene testing to multigene panel testing, with techniques such as whole-exome and genome sequencing on the horizon, increasing the complexity of test selection and interpretation, as well as patient education and medical decision-making.

COMMON ADULT-ONSET GENETIC DISORDERS

INHERITANCE PATTERNS

Adult-onset hereditary diseases follow multiple patterns of inheritance. Some are autosomal dominant conditions. These include many common cancer susceptibility syndromes such as hereditary breast and ovarian cancer (due to germline *BRCA1* and *BRCA2* mutations) and Lynch syndrome (caused by germline mutations in the mismatch repair genes *MLH1*, *MSH2*, *MSH6*, and *PMS2*). In both of these examples, inherited mutations are associated with a high penetrance (lifetime risk) of cancer, although penetrance is incomplete (risk is not 100%). In other conditions, although there is autosomal dominant transmission, penetrance is lower, thereby making the disorders more difficult to recognize. For example, germline mutations in *CHEK2* increase the risk of breast cancer, but with a moderate lifetime risk in the range of 20–30%, as opposed to 50–70% for mutations in *BRCA1* or *BRCA2*. Other adult-onset hereditary diseases are transmitted in an autosomal recessive fashion where two mutant alleles are necessary to cause full expression of disease. Examples include hemochromatosis and *MUTYH*-associated polyposis. There are more pediatric-onset autosomal recessive disorders, such as lysosomal storage diseases and cystic fibrosis.

The genetic risk for many adult-onset disorders is multifactorial. Risk can be conferred by genetic factors at a number of loci (polygenic), which individually have very small effects (usually with relative risks of <1.5). These risk loci (generally SNPs) combine with other genes and environmental factors in ways that are not well understood. Despite our incomplete understanding of gene-environment interactions, recent data suggest that a healthy lifestyle can mitigate risk associated with elevated polygenic risk for cardiovascular disease. SNP panels are available to assess risk of disease, but the optimal way of using this information in the clinical setting to improve patient outcomes remains uncertain.

Many diseases have multiple patterns of inheritance, adding to the complexity of evaluating patients and families for these conditions. For example, colon cancer can be associated with a single germline mutation in a mismatch repair gene (Lynch syndrome; autosomal dominant), biallelic mutations in *MUTYH* (autosomal recessive), or multiple SNPs (polygenic). Many more individuals will have SNP risk alleles than germline mutations in high-penetrance genes, but cumulative lifetime risk of colon cancer related to the former is modest, whereas the risk related to the latter is significant. Personal and family histories provide important insights into the possible mode of inheritance.

FAMILY HISTORY

When two or more first-degree relatives are affected with asthma, cardiovascular disease, type 2 diabetes, breast cancer, colon cancer, or melanoma, the relative risk for disease among close relatives ranges from two- to fivefold, underscoring the importance of family history for these prevalent disorders. In most situations, the key to assessing the inherited risk for common adult-onset diseases is the collection and interpretation of a detailed personal and family medical history in conjunction with a directed physical examination.

Family history should be recorded in the form of a pedigree, conveying health-related data on first- and second-degree relatives. When such pedigrees suggest inherited disease, they should be expanded to include additional family members. The determination of risk for an asymptomatic individual will vary depending on the size of the family, the number of unaffected relatives, the types of diagnoses, and the ages of disease

onset. For example, a woman with two first-degree relatives with breast cancer is at greater risk for a specific Mendelian disorder if she has a total of 3 female first-degree relatives (with only 1 unaffected) than if she has a total of 10 female first-degree relatives (with 8 unaffected). Factors such as adoption and limited family structure (few women in a family or multiple early deaths unrelated to the target disease) should to be taken into consideration in the interpretation of a pedigree. Additional considerations include young age of disease onset (e.g., a 30-year nonsmoking woman with a myocardial infarction), unusual diseases (e.g., male breast cancer or medullary thyroid cancer), and the finding of multiple potentially related diseases in an individual (e.g., a woman with a history of both colon and endometrial cancer). Some adult-onset diseases are more prevalent in certain ethnic groups. For instance, 2.5% of individuals of Ashkenazi Jewish ancestry carry one of three founder mutation in *BRCA1* and *BRCA2*. Factor V Leiden mutations are much more common in Caucasians than in Africans or Asians.

Additional variables that should be documented are nonhereditary risk factors among those with disease (such as cigarette smoking and myocardial infarction; asbestos exposure and lung disease; and mantle radiation and breast cancer). Significant associated environmental exposures or lifestyle factors decrease the likelihood of a specific genetic disorder. In contrast, the absence of nonhereditary risk factors typically associated with a disease raises concern about a genetic association. A personal or family history of deep-vein thrombosis in the absence of known environmental or medical risk factors suggests a hereditary thrombotic disorder. The physical examination may also provide important clues about the risk for a specific inherited disorder. A patient presenting with xanthomas at a young age should prompt consideration of familial hypercholesterolemia. The presence of trichilemmomas in a woman with breast cancer raises concern for Cowden syndrome, associated with *PTEN* mutations.

Recall of family history is often inaccurate. This is especially so when the history is remote and families lose contact or separate geographically. It can be helpful to ask patients to fill out family history forms before or after their visits, because this provides them with an opportunity to contact relatives. Ideally, this information should be embedded in electronic health records and updated intermittently. Attempts should be made to confirm the illnesses reported in the family history before making important and, in certain circumstances, irreversible management decisions. This process is often labor intensive and ideally involves interviews of additional family members or review of medical records (including pathology reports), and death certificates.

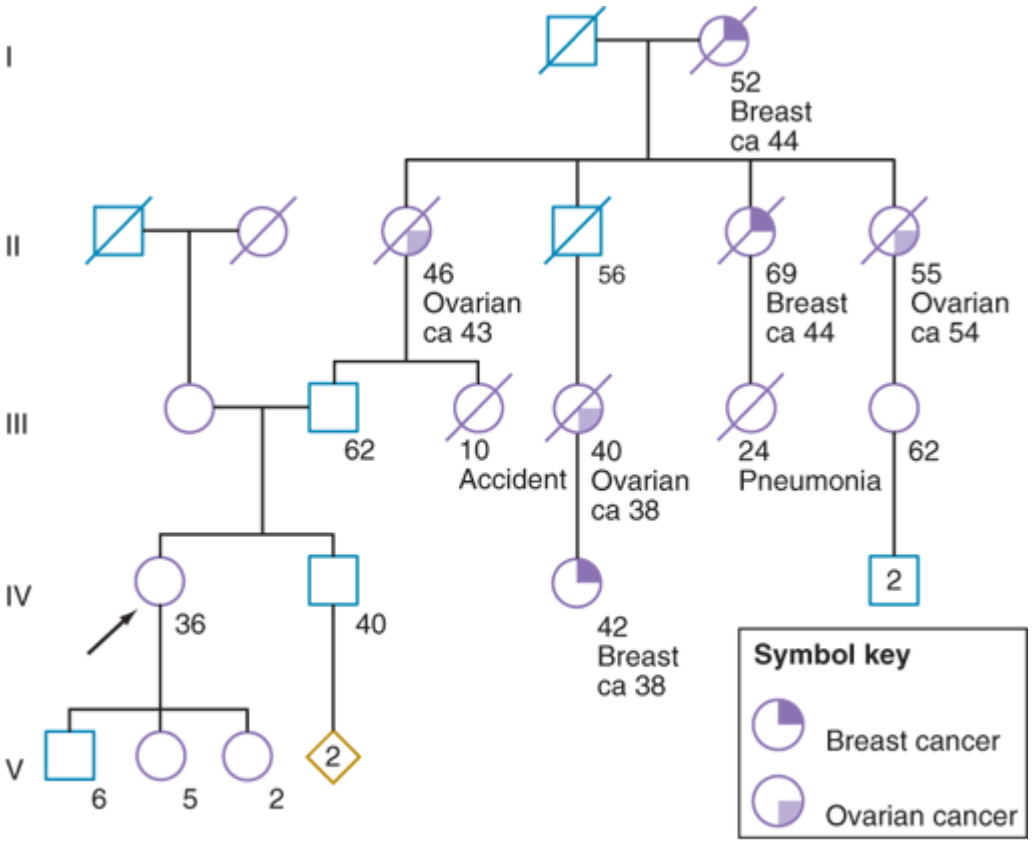
Although many inherited disorders will be suggested by the clustering of relatives with the same or related conditions, it is important to note that disease penetrance is incomplete for most genetic disorders. As a result, the pedigree obtained in such families may not exhibit a clear Mendelian inheritance pattern because not all family members carrying the disease-associated alleles will manifest clinical evidence of the condition. Furthermore, genes associated with some of these disorders often exhibit variable disease expression. For example, the breast cancer–associated gene *BRCA2* can predispose to several different malignancies in the same family, including cancers of the breast, ovary, pancreas, skin (melanoma), and prostate. For common diseases such as breast cancer, some family members without the susceptibility allele

(or genotype) may develop breast cancer (or phenotype) sporadically. Such *phenocopies* represent another confounding variable in the pedigree analysis.

Some of the aforementioned features of the family history are illustrated in **Fig. 457-1**. In this example, the proband (the individual serving as the starting point for genetic assessment in a family), a 36-year-old woman (IV-1), has a strong history of breast and ovarian cancer on the paternal side of her family. The early age of onset and the co-occurrence of breast and ovarian cancer in this family suggest the possibility of an inherited mutation in *BRCA1* or *BRCA2*, associated with the hereditary breast and ovarian cancer syndrome. It is unclear however, without genetic testing, whether her father harbors such a mutation and transmitted it to her. After appropriate genetic counseling of the proband and her family, the most informative and cost-effective approach to DNA analysis in this family is to test the cancer-affected 42-year-old living cousin for the presence of a *BRCA1* or *BRCA2* mutation. If a mutation is found, then it is possible to test for this particular alteration in other family members, if they so desire. In the example shown, if the proband's cousin has a *BRCA1* mutation, testing for this alteration (single site testing) would be offered to the proband's father. If he tests positive, there is a 50:50 probability that the mutation was transmitted to her, and genetic testing can be used to establish the absence or presence of this alteration. In contrast, if he tests negative for the known familial *BRCA1* mutation (a true negative result), the proband and her brother are not at risk for having inherited this variant.

FIGURE 457-1

36-year-old woman (*arrow*) seeks consultation because of her family history of cancer. The patient expresses concern that the multiple cancers in her relatives imply an inherited predisposition to develop cancer. The family history is recorded, and records of the patient's relatives confirm the reported diagnoses.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

GENETIC TESTING FOR ADULT-ONSET DISORDERS

A critical first step before initiating genetic testing is to ensure that the correct clinical diagnosis has been made, whether it is based on family history, characteristic physical findings, pathology, or biochemical testing. Such careful clinical assessment can define the phenotype. In the traditional model of genetic testing, testing is directed initially toward the most probable genes (determined by the phenotype), which prevents unnecessary testing. Many disorders exhibit the feature of locus heterogeneity, which refers to the fact that mutations in different genes can cause phenotypically similar disorders. For example, osteogenesis imperfecta **(Chap. 406)**, long QT syndrome **(Chap. 247)**, muscular dystrophy **(Chap. 441)**, and hereditary predisposition to breast **(Chap. 75)** or colon **(Chap. 77)** cancer can each be caused by mutations in a number of distinct genes. The patterns of disease transmission, disease risk, clinical course, and treatment may differ significantly depending on the specific gene affected. Historically, the choice of which gene to test has been determined by unique clinical and family history features and the relative prevalence of candidate genetic disorders. However, rapid changes in genetic testing techniques, as discussed below, are impacting this paradigm. It is now technically and financially feasible to sequence many genes (or even the whole exome) at one time. The incorporation of expanded testing for germline mutations is rapidly evolving both within the clinic as well as through direct-to-consumer marketing of genetic and genomic tests.

METHODOLOGIC APPROACHES TO GENETIC TESTING

Genetic testing is regulated and performed in much the same way as other specialized laboratory tests. In the United States, genetic testing laboratories are Clinical Laboratory Improvement Amendments (CLIA) approved to ensure that they meet quality and proficiency standards. A useful information source for various genetic tests is *www.genetests.org*. It should be noted that many tests need to be ordered through specialized laboratories.

Genetic testing is performed largely by DNA sequence analysis for mutations, although genotype can also be deduced through the study of RNA or protein (e.g., apolipoprotein E, hemoglobin S, and immunohistochemistry). For example, universal Lynch syndrome screening of colorectal and uterine cancers via immunohistochemical analysis for absence of expression of mismatch repair proteins is recommended by the National Comprehensive Cancer Center Network. The determination of DNA sequence alterations relies heavily on the use of polymerase chain reaction (PCR), which allows rapid amplification and analysis of the gene of interest. In addition, PCR enables genetic testing on minimal amounts of DNA extracted from a wide range of tissue sources including leukocytes, mucosal epithelial cells (obtained via saliva or buccal swabs), and archival tissues. Amplified DNA can be analyzed directly by DNA sequencing, or it can be hybridized to DNA chips or blots to detect the presence of normal and altered DNA sequences. Direct DNA sequencing is frequently used for determination of hereditary disease susceptibility and prenatal diagnosis. Analyses of large alterations (e.g., deletions, duplications, rearrangements, translocations) of the genome are possible using cytogenetics, fluorescent in situ hybridization (FISH), Southern blotting, or multiplex ligation-dependent probe amplification (MLPA).

Massively parallel sequencing (also called *next-generation sequencing*) has significantly altered the approach to genetic testing for adult-onset hereditary susceptibility disorder. This technology encompasses high-throughput approaches to DNA analysis which can reliably examine many genes at one time. Technically, this involves sequencing of millions of small fragments of DNA in parallel. Through bioinformatics, these fragments are pieced together by mapping the individual sequence reads to the human reference genome, a very different process than traditional Sanger sequencing which is time-consuming and expensive.

Multiplex panels for inherited susceptibility are commercially available and include testing of a number of genes that have been associated with the condition of interest. For example, panels are available for Brugada syndrome, hypertrophic cardiomyopathy, and Charcot-Marie-Tooth neuropathy. For many syndromes, this type of panel testing may make sense. However, in other situations, the clinical utility of panel testing is evolving and may be dependent on the particular composition of the panel. Currently available breast cancer susceptibility panels contain close to 30 genes with larger multi-cancer panels available. Some of the genes included in the larger multi-cancer panels have no known association with breast cancer or have only a modest associated risk and the clinical utility is uncertain. An additional problem of sequencing many genes, rather than focusing on leading candidate genes, is the identification of one or more variants of uncertain significance, discussed below, or an unexpected, yet clinically relevant result.

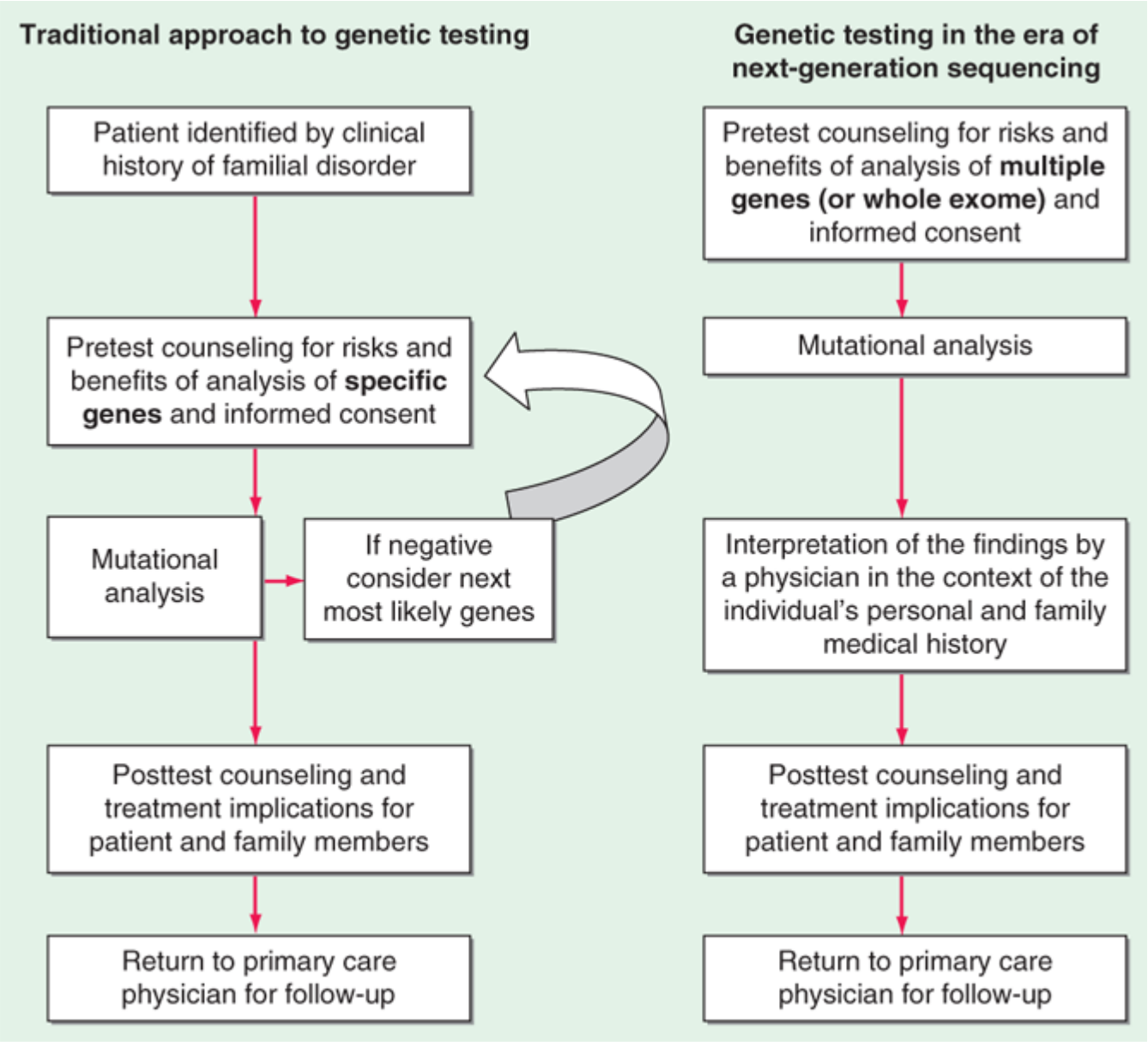
Whole-exome sequencing (WES) is also now commercially available, although largely used in individuals with syndromes unexplained by traditional genetic testing. As cost declines, WES may be more widely used. Whole-genome sequencing is also commercially available. Although it may be quite feasible to sequence the

entire genome, there are many issues in doing so, including the daunting task of analyzing the vast amount of data generated. Other issues include: (1) the optimal way in which to obtain informed consent, (2) interpretation of frequent sequence variation of uncertain significance, (3) interpretation of alterations in genes with unclear relevance to specific human pathology, and (4) management of unexpected but clinically significant genetic findings.

Testing strategies are evolving as a result of these new genetic testing platforms. As the costs of multiplex gene panels and WES continue to fall, and as interpretation and understanding of the clinical relevance of such test results improve, there has been a shift to more extensive panel-based genetic testing in the clinic. For example, in the past, a 30-year-old woman with breast cancer but no family history of cancer and no syndromic features would undergo *BRCA1/2* testing and would be offered *TP53* testing in light of her early onset disease (notably, a reasonable number of individuals offered *TP53* testing for Li-Fraumeni syndrome in the past declined because mutations are associated with extremely high cancer risks—including childhood cancers—in multiple organs and means to mitigate risk are uncertain).Without features consistent with other high-risk, breast cancer-related conditions like Cowden syndrome, the woman would not have been routinely offered *PTEN* analysis (associated with Cowden syndrome) or testing for other breast cancer-associated genes including *CHEK2* and *ATM*. It is now possible to synchronously analyze all of the aforementioned genes, along with genes such as BRIP and RAD51D (which are associated with moderate ovarian cancer risk but unclear risk for other cancers, including breast cancer) for a nominally higher cost than *BRCA1/2* testing alone. Concerns about such panels include appropriate consent strategies related to unclear findings, including one or more variants of uncertain significance, unanticipated results, and the uncertain clinical utility of some of the genes included on the panel **(Fig. 457-2)**.

FIGURE 457-2

Approach to genetic testing.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Limitations to the accuracy and interpretation of genetic testing exist. In addition to technical errors, genetics tests are sometimes designed to detect only the most common mutations. In addition, genetic testing has evolved over time. For example, it was not possible to obtain commercially available comprehensive large genomic rearrangement testing for *BRCA1* and *BRCA2* until 2006. Therefore, a negative result must be qualified by the possibility that the individual may have a variant that was not detectable in the test. In addition, the individual may have a mutation in another untested cancer-associated gene or in a gene not yet reported to be associated with elevated disease risk. As such, unless there is known baseline mutation in the family, a negative result in an individual with a suggestive personal or family history is typically classified as an uninformative negative. In this circumstance, medical management decisions should be based on personal and family history. For example, a woman with a strong family history of breast cancer who receives an uninformative negative panel-based genetic test result may still be eligible for high-risk care, including consideration of MRI-based breast cancer screening in addition to close clinical surveillance and mammograms.

The finding of a VUS is another limitation to genetic testing. A VUS (also termed *unclassified variant*) is a sequence variation in a gene where the effect of the alteration on the function of the protein is not known. Many of these variants are single nucleotide substitutions (also called missense mutations) that result in a single amino acid change. Although many VUSs will ultimately be reclassified as benign variants, some will prove to be functionally important. As more genes are sequenced (for example, in a multiplex panel or through WES), the percentage of individuals found to have one or more VUSs increases significantly. The finding of a VUS is difficult for patients and providers alike and complicates decisions regarding medical management. In this setting, until there is further reclassification of the variant, ongoing screening, surveillance and care is typically determined based on personal and family history.

Clinical utility is an important consideration because genetic testing for susceptibility to chronic diseases is increasingly integrated into the practice of medicine. In some situations, there is proven clinical utility to genetic testing with significant evidence-based changes in medical management options and recommendations based on results. For example, there is clear evidence that risk-reducing bilateral salpingo-oophorectomy benefits women with a documented *BRCA1*/*2* mutation, relative to both ovarian and breast cancer-related risk. However, in many cases, the discovery of disease-associated genes has outpaced studies that assess how such information should be used in the clinical management of the patient and family. This is particularly true for moderate- and low-penetrance gene mutations. Therefore, predictive genetic testing should be approached with caution and offered to individuals who have been adequately counseled and have provided informed consent.

Predictive genetic testing falls into two distinct categories. Presymptomatic testing applies to diseases where a specific genetic alteration is associated with a near 100% likelihood of developing disease. In contrast, predisposition testing predicts a risk for disease that is <100%. For example, presymptomatic testing is available for those at risk for Huntington's disease; whereas, predisposition testing is considered for those at risk for hereditary colon cancer. It is important to note that for the majority of adult-onset disorders, testing is only predictive. Test results cannot reveal with confidence whether, when, or how the disease will manifest itself. For example, not everyone with the apolipoprotein ε4 allele will develop Alzheimer's disease, and individuals without this genetic marker can still develop the disorder.

The optimal testing strategy for a family is to initiate testing in a [target] disease-affected family member first. Identification of a mutation can direct the testing of other at-risk family members (whether symptomatic or not). In the absence of additional familial or environmental risk factors, individuals who test negative for the mutation found in the affected family member can be informed that they are at general population risk for that particular disease. Furthermore, they can be reassured that they are not at risk for passing the mutation on to their children. On the other hand, asymptomatic family members who test positive for the known mutation must be informed that they are at increased risk for disease development and for transmitting the alteration to their children.

Pretest counseling and education are important, as is an assessment of the patient's ability to understand and cope with test results. Genetic testing has implications for entire families, and thus individuals interested in pursuing genetic testing must consider how test results might impact their relationships with relatives,

partners, spouses, and children. In families with a known genetic mutation, those who test positive must consider the impact of their carrier status on their present and future lifestyles; those who test negative may manifest survivor guilt. Parents who are found to have a disease-associated mutation often express considerable anxiety and despair as they address the issue of risk to their children. In addition, some individuals consider options such as preimplantation genetic diagnosis (PGD) in their reproductive decision-making.

When a condition does not manifest until adulthood, clinicians and parents are faced with the question of whether at-risk children should be offered genetic testing and, if so, at what age. Although the matter is debated, several professional organizations have cautioned that genetic testing for adult-onset disorders should not be offered to children. Many of these conditions have no known interventions in childhood to prevent disease; consequently, such information can pose significant psychosocial risk to the child. In addition, there is concern that testing during childhood violates a child's right to make an informed decision regarding testing upon reaching adulthood. On the other hand, testing should be offered in childhood for disorders that may manifest early in life, especially when management options are available. For example, children with multiple endocrine neoplasia 2 (MEN 2) may develop medullary thyroid cancer early in life and should be considered for prophylactic thyroidectomy **(Chap. 381)**. Similarly, children with familial adenomatous polyposis (FAP) due to a mutation in *APC* may develop polyps in their teens with progression to invasive cancer in the twenties, and therefore, colonoscopy screening is started between the ages of 10 and 15 years **(Chap. 77)**.

INFORMED CONSENT

Informed consent for genetic testing begins with education and counseling. The patient should understand the risks, benefits, and limitations of genetic testing, as well as the potential implications of test results. Informed consent should include a written document, drafted clearly and concisely in a language and format that is understandable to the patient. Because molecular genetic testing of an asymptomatic individual often allows prediction of future risk, the patient should understand all potential long-term medical, psychological, and social implications of testing. There have long been concerns about the potential for genetic discrimination. The Genetic Information Nondiscrimination Act (GINA) was passed in 2008 and provides some protections related to job and health insurance discrimination. It is important to explore with patients the potential impact of genetic test results on future health as well as disability and life insurance coverage. Patients should understand that alternatives remain available if they decide not to pursue genetic testing, including the option of delaying testing to a later date. The option of DNA banking should be presented so that samples are readily available for future use by family members, if needed.

FOLLOW-UP CARE AFTER TESTING

Depending on the nature of the genetic disorder, posttest interventions may include: (1) cautious surveillance and awareness; (2) specific medical interventions such as enhanced screening, chemoprevention, or risk-reducing surgery; (3) risk avoidance; and (4) referral to support services. For

example, patients with known deleterious mutations in *BRCA1* or *BRCA2* are strongly encouraged to undergo risk-reducing salpingo-oophorectomy and are offered intensive breast cancer screening as well as the option of risk-reducing mastectomy. In addition, such women may wish to take chemoprevention with tamoxifen, raloxifene, or exemestane. Those with more limited medical management and prevention options, such as patients with Huntington's disease, should be offered continued follow-up and supportive services, including physical and occupational therapy and social services or support groups as indicated. Specific interventions will change as research continues to enhance our understanding of the medical management of these genetic conditions and more is learned about the functions of the gene products involved.

Individuals who test negative for a mutation in a disease-associated gene identified in an affected family member must be reminded that they may still be at risk for the disease. This is of particular importance for common diseases such as diabetes mellitus, cancer, and coronary artery disease. For example, a woman who finds that she does not carry the disease-associated mutation in *BRCA1* previously discovered in the family should be reminded that she still requires the same breast cancer screening recommended for the general population.

GENETIC COUNSELING AND EDUCATION

Genetic counseling is the process of helping people understand and adapt to the medical, psychological and familial implications of genetic contributions to disease. This process integrates the following: interpretation of family and medical histories to assess the chance of disease occurrence or recurrence; education about the natural history of the condition, inheritance pattern, testing, management, prevention, support resources and research; counseling to promote informed choices in view of risk assessment, family goals, ethical and religious values. Genetic counseling should be distinguished from genetic testing and risk-based medical screening and care, though genetic counselors are involved in these later issues.

Genetic risk assessment is complex and often involves elements of uncertainty. Genetic counseling can be useful in a wide range of situations **(Table 457-1)**. The role of the genetic counselor includes the following:

TABLE 457-1

Indications for Genetic Counseling

Advanced maternal age (>35 years)
Consanguinity
Previous history of a child with birth defects or a genetic disorder
Personal or family history suggestive of a genetic disorder
High-risk ethnic groups
Documented genetic alteration in a family member
Ultrasound or prenatal testing suggesting a genetic disorder

1. Gather and document a detailed family history.
2. Educate patients about general genetic principles related to disease risk, both for themselves and for others in the family.
3. Assess and enhance the patient's ability to cope with the genetic information offered.
4. Discuss how nongenetic factors may relate to the ultimate expression of disease.
5. Address medical management issues.
6. Assist in determining the role of genetic testing for the individual and the family.
7. Ensure the patient is aware of the indications, process, risks, benefits, and limitations of the various genetic testing options.
8. Assist the patient, family, and referring physician in the interpretation of the test results.
9. Ensure that the patient has the resources necessary to alert relatives to their risk, particularly in the face of a positive genetic test result.
10. Address the reproductive implications of a positive genetic test result, including the risk for a recessive disorder as well as discussion of reproductive options, including gamete donation or PGD.
11. Refer the patient and other at-risk family members for additional medical and support services, if necessary.

The principles of voluntary and informed decision-making, and protection of the individual's privacy and confidentiality are core principles in the practice of genetic counseling. Genetic counseling is generally offered in a *nondirective, noncoercive* manner, wherein patients learn to understand how their values factor into a particular medical decision. Nondirective counseling is particularly appropriate when there are no data demonstrating a clear benefit associated with a particular intervention or when an intervention is considered experimental. For example, nondirective genetic counseling is used when a person is deciding whether to undergo genetic testing for Huntington's disease. At this time, there is no clear benefit (in terms of medical outcome) to an at-risk individual undergoing genetic testing for this disease because its course cannot be altered by therapeutic interventions. However, testing can have an important impact on the individual's perception of advanced care planning and his or her interpersonal relationships and plans for childbearing. Therefore, the decision to pursue testing rests on the individual's belief system and values. On the other hand, a more directive approach is appropriate when a condition can be treated. In a family with FAP, colon cancer screening and prophylactic colectomy should be recommended for known *APC* mutation carriers. The counselor and clinician following this family must ensure that the at-risk family members have access to the resources necessary to adhere to these recommendations.

Genetic education is central to an individual's ability to make an informed decision regarding testing options and treatment. An adequate knowledge of patterns of inheritance will allow patients to understand the probability of disease risk for themselves and other family members. It is also important to impart the concepts of disease penetrance and expression. For most complex adult-onset genetic disorders, asymptomatic patients should be advised that a positive test result does not always translate into future disease development. In addition, the role of nongenetic factors, such as environmental exposures and lifestyle, must be discussed in the context of multifactorial disease risk and disease prevention. Finally, patients should understand the natural history of the disease as well as the potential options for intervention, including screening, prevention, and in certain circumstances, pharmacologic treatment or prophylactic surgery.

THERAPEUTIC INTERVENTIONS BASED ON GENETIC RISK FOR DISEASE

Specific treatments are available for a number of genetic disorders. Strategies for the development of therapeutic interventions have a long history in childhood metabolic diseases; however, these principles have been applied in the diagnosis and management of adult-onset diseases as well **(Table 457-2)**. Hereditary hemochromatosis is usually caused by mutations in *HFE* (although other genes have been less commonly associated) and manifests as a syndrome of iron overload, which can lead to liver disease, skin pigmentation, diabetes mellitus, arthropathy, impotence in males, and cardiac issues **(Chap. 407)**. When identified early, the disorder can be managed effectively with therapeutic phlebotomy. Therefore, when the diagnosis of hemochromatosis has been made in a proband, it is important to counsel other family members in order to minimize the impact of the disorder.

TABLE 457-2

Examples of Genetic Testing and Possible Interventions

Genetic Disorder	Inheritance	Genes	Interventions
Oncologic			
Lynch syndrome (HNPCC)	AD	*MLH1, MSH2, MSH6, PMS2*	Early endoscopic screening; risk-reducing surgery
Familial adenomatous polyposis	AD	*APC*	Early and frequent endoscopy; prophylactic colectomy
Hereditary breast and ovarian cancer	AD	*BRCA1, BRCA2*	Risk reducing salpingo-oophorectomy; intensified breast surveillance including breast MRI; risk-reducing mastectomy
Hereditary diffuse gastric cancer	AD	*CDH1*	Prophylactic gastrectomy; enhanced breast cancer surveillance
Hematologic			
Factor V Leiden	AD	*F5*	Avoidance of thrombogenic risk factors
Hemophilia A	XL	*F8*	Factor VIII replacement
Hemophilia B	XL	*F9*	Factor IX replacement
Glucose 6-phosphate dehydrogenase deficiency	XL	*G6PD*	Avoidance of oxidant drugs and certain foods
Cardiovascular			
Hypertrophic cardiomyopathy	AD	>10 genes including *MYBPC3, MYH7, TNNT2, TPM1*	Echocardiographic screening; pharmacologic intervention; myomectomy

Genetic Disorder	Inheritance	Genes	Interventions
Long QT syndrome	AD, AR	>10 genes including *KCNQ1, SCN5A, KCNE1, KCNE2*	Electrocardiographic screening; pharmacologic intervention; implantable cardiac defibrillator devices
Marfan's syndrome	AD	*FBN1*	Echocardiographic screening; prophylactic beta blockers; aortic valve replacement as indicated
Gastrointestinal			
Familial Mediterranean fever	AR	*MEFV*	Colchicine
Hemochromatosis	AR	*HFE*	Phlebotomy
Pulmonary			
α_1 Antitrypsin deficiency	AR	*SERPINA1*	Avoidance of smoking and occupational and environmental toxins
Cystic fibrosis	AR	*CFTR*	Chest physiotherapy; agents to promote airway secretion clearance; *CFTR* modulators (G551D mutations); lung transplantation
Endocrine			
Neurohypophyseal diabetes insipidus	AD	*AVP*	Replace vasopressin
Familial hypocalciuric hypercalcemia	AD	*CASR*	Avoidance of parathyroidectomy
Multiple endocrine neoplasia type 2	AD	*RET*	Prophylactic thyroidectomy; screening for pheochromocytoma and hyperparathyroidism
Renal			

Genetic Disorder	Inheritance	Genes	Interventions
Polycystic kidney disease	AD, AR	*PKD1, PKD2, PKHD1*	Prevention of hypertension; prevention of urinary tract infections; kidney transplantation
Nephrogenic diabetes insipidus	XL, AR	*AVPR2, AQP2*	Fluid replacement; thiazides with or without amiloride
Neurologic			
Malignant hyperthermia	AD	*RYR1, CACNA1S*	Avoidance of precipitating anesthetics
Hyperkalemic periodic paralysis	AD	*SCN4A*	Diet rich in calcium and low in potassium; thiazides or acetazolamide
Duchenne's and Becker's muscular dystrophy	XL	*DMD*	Corticosteroids; physical therapy
Wilson's disease	AR	*ATP7B*	Zinc, trientene

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; HNPCC, hereditary nonpolyposis colorectal cancer; MRI, magnetic resonance imaging; XL, X-linked.

Preventative measures and therapeutic interventions are not restricted to metabolic disorders. Identification of familial forms of long QT syndrome, associated with ventricular arrhythmias, allows early electrocardiographic testing and the use of prophylactic antiarrhythmic therapy, overdrive pacemakers, or defibrillators. Individuals with familial hypertrophic cardiomyopathy can be screened by ultrasound, treated with beta blockers or other drugs, and counseled about the importance of avoiding strenuous exercise and dehydration. Those with Marfan's syndrome can be treated with beta blockers or angiotensin II receptor blockers and monitored for the development of aortic aneurysms.

The field of pharmacogenetics identifies genes that alter drug metabolism or confer susceptibility to toxic drug reactions. Pharmacogenetics seeks to individualize drug therapy in an attempt to improve treatment outcomes and reduce toxicity. Examples include thiopurine methyltransferase (TPMT) deficiency, dihydropyrimidine dehydrogenase deficiency, malignant hyperthermia, and glucose-6-phosphate deficiency. Despite successes in this area, it is not always clear how to incorporate pharmacogenetics into clinical care. For example, although there is an association with *CYP2C6* and *VKORC1* genotypes and warfarin dosing, there is no evidence that incorporating genotyping into clinical practice improves patient outcomes compared with clinical algorithms.

The identification of germline abnormalities that increase the risk of specific types of cancer is rapidly changing clinical management. Identifying family members with mutations that predispose to FAP or Lynch syndrome leads to recommendations of early cancer screening and prophylactic surgery, as well as consideration of chemoprevention and attention to healthy lifestyle habits. Similar principles apply to familial forms of melanoma as well as cancers of the breast, ovary, and thyroid. In addition to increased screening and prophylactic surgery, the identification of germline mutations associated with cancer may also lead to the development of targeted therapeutics, for example, the U.S. Food and Drug Administration (FDA) approval of the poly–ADP ribose polymerase (PARP) inhibitors olaparib and rucaparib for BRCA1/2-associated recurrent ovarian cancer.

Although the role of genetic testing in the clinical setting continues to evolve, such testing holds the promise of allowing early and more targeted interventions that can reduce morbidity and mortality. Rapid technologic advances are changing the ways in which genetic testing is performed. As genetic testing becomes less expensive and technically easier to perform, it is anticipated that there will be an expansion of its use. This will present challenges, but also opportunities. It is critical that physicians and other health care professionals keep current with advances in genetic medicine in order to facilitate appropriate referral for genetic counseling and judicious use of genetic testing, as well as to provide state-of-the-art, evidence-based care for affected or at-risk patients and their relatives.

FURTHER READING

Anya ER et al: Genetic information, non-discrimination, and privacy protections in genetic counseling practice. J Genet Couns 23:891, 2014. [PubMed: 25063358]

Clayton EW: Ethical, legal, and social implications of genomic medicine. N Engl J Med 349:562, 2003. [PubMed: 12904522]

Hampel H et al: A practice guideline from the American College of Medical Genetics and Genomics and the National Society of Genetic Counselors: Referral indications for cancer predisposition assessment. Genet Med 17:70, 2015. [PubMed: 25394175]

Resta R et al: A new definition of genetic counseling: National Society of Genetic Counselors' Task Force report. J Genet Couns 15:77, 2006. [PubMed: 16761103]

Roberts JS, Ostergren J: Direct-to-consumer genetic testing and personal genomics services: A review of recent empirical studies. Curr Genet Med Rep 1:182, 2013. [PubMed: 24058877]

Robson ME et al: American Society of Clinical Oncology policy statement update: Genetic and genomic testing for cancer susceptibility. J Clin Oncol 33:3660, 2015. [PubMed: 26324357]

Chapter 458: Gene and Cell Based Therapy in Clinical Medicine

Katherine A. High; Malcolm K. Brenner

FIGURE 458-1

INTRODUCTION

Gene transfer is a novel area of therapeutics in which the active agent is a nucleic acid sequence rather than a protein or small molecule. Because delivery of naked DNA or RNA to a cell is an inefficient process, most gene transfer is carried out using a vector, or gene delivery vehicle. These vehicles have generally been engineered from viruses by deleting some or all of the viral genome and replacing it with the therapeutic gene of interest under the control of a suitable promoter (Table 458-1). Gene transfer strategies can thus be described in terms of three essential elements: (1) a vector; (2) a gene to be delivered, sometimes called the transgene; and (3) a physiologically relevant target cell to which the DNA or RNA is delivered. The series of steps in which the vector and donated DNA enter the target cell and express the transgene is referred to as *transduction*. Gene delivery can take place in vivo, in which the vector is directly injected into the patient, or, in the case of hematopoietic and some other target cells, ex vivo, with removal of the target cells from the patient, followed by return of the gene-modified autologous cells to the patient after manipulation in the laboratory. The latter approach effectively combines gene transfer techniques with cellular therapies (Chap. 473).

TABLE 458-1

Characteristics of Gene Delivery Vehicles

Viral Vectors							
Features	**Retroviral**	**Lentiviral**	**Adenoviral**	**AAV**	**Human Foamy Virus**	**HSV-1**	**Alpha Viruses**
Viral genome	RNA	RNA	DNA	DNA	RNA	DNA	RNA
Cell division requirement	Yes	G_1 phase	No	No	No	No	No
Packaging limitation	8 kb	8 kb	8–30 kb	5 kb	8.5 kb	40–150 kb	5 kb
Immune responses to vector	Few	Few	Extensive	Few	Few	Few in recombinant virus	Few
Genome integration	Yes	Yes	Poor	Poor	Yes	No	No
Long-term expression	Yes	Yes	No	Yes	Yes	No	No
Main advantages	Persistent gene transfer in dividing cells	Persistent gene transfer in transduced tissues	Highly effective in transducing various tissues	Elicits few inflammatory responses, nonpathogenic	Persistent gene expression in both dividing and nondividing cells	Large packaging capacity with persistent gene transfer	Limited immune responses against the vector
Main disadvantages	Theoretical risk of insertional mutagenesis (occurred in multiple cases)	Might induce oncogenesis in some cases (not yet observed)	Viral capsid elicits strong immune responses	Limited packaging capacity	In need of a stable packaging system	Residual cytotoxicity with neuron specificity	Transduced gene expression is transient

Abbreviations: AAV, adeno-associated virus; HSV, herpes simplex virus.

Gene transfer is one of the most powerful concepts in modern molecular medicine and has the potential to address a host of diseases for which there are currently no available treatments. Clinical trials of gene therapy have been under way since 1990; the first gene therapy product to be licensed in the United States or Europe was licensed in 2012 (see below). Given that vector-mediated gene therapy is arguably one of the most complex therapeutics yet developed, consisting of both a nucleic acid and a protein component, this time course from first clinical trial to licensed product is noteworthy for being similar to those seen with other novel classes of therapeutics, i.e., monoclonal antibodies. Over 5000 subjects have been enrolled in gene transfer studies, and serious adverse events have been rare. Some of the initial trials were characterized by an overabundance of optimism and a failure to be appropriately critical of preclinical studies in animals; in addition, it was in some contexts not fully appreciated that animal studies are only a partial guide to safety profiles of products in humans (e.g., insertional mutagenesis, and human immune responses to the vector). Initial exuberance was driven by many factors, including an intense desire to develop therapies for hitherto untreatable diseases, lack of understanding of risks, and, in some cases, undisclosed financial conflicts of interest. After a teenager died of complications related to vector infusion, the field underwent a retrenchment; continued efforts led to a more nuanced understanding of the risks and benefits of these new therapies and more sophisticated selection of disease targets. Currently, gene therapies are being developed for a variety of disease entities (Table 458-2).

TABLE 458-2

Most Common Indications in Gene Therapy Trials

Indication	Number
Cancer	1554
Monogenic diseases	248
Infectious diseases	180

Source: Adapted from J Gene Med *http://www.abedia.com/wiley/indications.php, 2016*.

GENE TRANSFER FOR GENETIC DISEASE

Gene transfer strategies for genetic disease generally involve gene addition therapy, an approach characterized by transfer of the missing gene to a physiologically relevant target cell. However, other strategies are possible, including supplying a gene that achieves a similar biologic effect through an alternative pathway (e.g., factor VIIa for hemophilia A); supplying an antisense oligonucleotide to splice out a mutant exon if the sequence is not critical to the function of the protein (as has been done with the dystrophin gene in Duchenne's muscular dystrophy); or downregulating a harmful effect through a small interfering RNA (siRNA). Two distinct strategies are used to achieve long-term gene expression: one is to transduce stem cells with an integrating vector, so that all progeny cells will carry the donated gene; and the other is to transduce long-lived cells, such as skeletal muscle or neurons. In the case of long-lived cells, integration into the target cell genome is unnecessary. Instead, because the cells are nondividing, the donated DNA, if stabilized in an episomal form, will give rise to expression for the life of the cell. This approach thus avoids problems related to integration and insertional mutagenesis.

IMMUNODEFICIENCY DISORDERS: PROOF OF PRINCIPLE

Early attempts to effect gene replacement into hematopoietic stem cells (HSCs) were stymied by the relatively low transduction efficiency of retroviral vectors, which require dividing target cells for integration. Because HSCs are normally quiescent, they are a formidable transduction target. However, identification of cytokines that induced cell division without promoting differentiation of stem cells, along with technical improvements in the isolation and transduction of HSCs, led to modest but real gains in transduction efficiency.

The first convincing therapeutic effect from gene transfer occurred with X-linked severe combined immunodeficiency disease (SCID), which results from mutations in the gene (*IL2RG*) encoding the γc subunit of cytokine receptors required for normal development of T and natural killer (NK) cells **(Chap. 344)**. Affected infants present in the first few months of life with overwhelming infections and/or failure to thrive. In this disorder, it was recognized that the transduced cells, even if few in number, would have a proliferative advantage compared to the nontransduced cells, which lack receptors for the cytokines required for lymphocyte development and maturation. Complete reconstitution of the immune system, including documented responses to standard childhood vaccinations, clearing of infections, and remarkable gains in growth occurred in most of the treated children. However, among 20 children treated in the initial trials, five eventually developed a syndrome similar to T cell acute lymphocytic leukemia, with splenomegaly, rising white counts, and the emergence of a single clone of T cells. Molecular studies revealed that, in most of these children, the retroviral vector had integrated within a gene, *LMO-2* (LIM only-2), which encodes a component of a transcription factor complex involved in hematopoietic development. The retroviral long terminal repeat increases the expression of *LMO-2*, resulting in T cell leukemia.

The X-linked SCID studies were a watershed event in the evolution of gene therapy. They demonstrated conclusively that gene therapy could cure disease; of the 20 children treated in these initial trials, 18 achieved correction of the immunodeficiency disorder. Unfortunately, 5 of the 20 patients later developed a leukemia-like disorder, and one died of this complication; the rest are alive and free of complications at time periods ranging up to 17 years after initial treatment. These studies demonstrated that insertional mutagenesis leading to cancer was more than a theoretical possibility **(Table 458-3)**. As a result of the experience in these trials, all protocols using integrating vectors in hematopoietic cells must include a plan for monitoring sites of insertion and clonal proliferation. Strategies to overcome this complication have included using a “suicide” gene cassette in the vector, so that errant clones can be quickly ablated, or using “insulator” elements in the cassette, which can limit the activation of genes surrounding the insertion site. Lentiviral vectors, which can efficiently transduce nondividing target cells, are also likely to be safer than retroviral vectors, based on patterns of integration; the field is thus gradually moving toward these to replace retroviral vectors.

TABLE 458-3

Potential Complications of Gene Therapy

Gene silencing—repression of promoter
Genotoxicity—complications arising from insertional mutagenesis
Phenotoxicity—complications arising from overexpression or ectopic expression of the transgene
Immunotoxicity—harmful immune response to either the vector or transgene; or a harmful immune response of the vector (e.g. CAR T cells)
Risks of horizontal transmission—shedding of infectious vector into environment
Risks of vertical transmission—germline transmission of donated DNA

More clear-cut success has been achieved in a gene therapy trial for another form of SCID, adenosine deaminase (ADA) deficiency **(Chap. 344)**. ADA-SCID is clinically similar to X-linked SCID, although it can be treated by enzyme replacement therapy with a pegylated form of the enzyme (PEG-ADA), which leads to immune reconstitution but not always to normal T cell counts. Enzyme replacement therapy is expensive (annual costs: $200,000–$300,000). The initial trials of gene therapy for ADA-SCID were unsuccessful, but modifications of this protocol to include the use of HSCs rather than T cells as the target for transduction; discontinuation of PEG-ADA at the time of vector infusion, so that the transduced cells have a proliferative advantage over the nontransduced; and the use of a mild conditioning regimen to facilitate engraftment of the transduced autologous cells have led to success without the complications seen in the X-linked SCID trials. There have been no complications in the 10 children treated on the Milan protocol, with a median follow-up of >11 years. This therapy was approved in 2016 by the European Medicines Agency. ADA-SCID, then, is an example where gene therapy has changed therapeutic options for patients. For those with a human leukocyte antigen (HLA)-identical sibling, bone marrow transplantation is still the best treatment option, but this includes only a minority of those affected. For those without an HLA-identical match, gene therapy has comparable efficacy to PEG-ADA, does not require repetitive injections, and does not present the risk of development of neutralizing antibodies to the bovine enzyme.

NEURODEGENERATIVE DISEASES: EXTENSION OF PRINCIPLE

The SCID trials gave support to the hypothesis that gene transfer into HSCs could be used to treat any disease for which allogeneic bone marrow transplantation was therapeutic. Moreover, the use of genetically modified autologous cells carried several advantages including no risk of graft-versus-host disease, guaranteed availability of a “donor” (unless the disease itself damages the stem cell population of the patient), and low likelihood of failure of engraftment. Cartier and Aubourg capitalized on this realization to conduct the first trial of lentiviral vector transduction of HSCs for a neurodegenerative disorder, X-linked adrenoleukodystrophy (ALD). X-linked ALD is a fatal demyelinating disease of the central nervous system caused by mutations in the gene encoding an adenosine triphosphate–binding cassette transporter. Deficiency of this protein leads to accumulation of very-long-chain fatty acids in oligodendrocytes and microglia, disrupting myelin maintenance by these cells. Affected boys present with clinical and neuroradiographic evidence of disease at age 6–8 and usually die before adolescence. Following lentiviral transduction of autologous HSCs in young boys with the disease, dramatic stabilization of disease occurred, demonstrating that stem cell transduction

could work for neurodegenerative as well as immunologic disorders. Investigators in Milan carried this observation one step further to develop a treatment for another neurodegenerative disorder that has previously responded poorly to bone marrow transplantation. Metachromatic leukodystrophy is a lysosomal storage disorder caused by mutations in the gene encoding arylsulfatase A (ARSA). The late infantile form of the disease is characterized by progressive motor and cognitive impairment, and death within a few years of onset, due to accumulation of the ARSA substrate sulfatide in oligodendrocytes, microglia, and some neurons. Recognizing that endogenous levels of production of ARSA were too low to provide cross-correction by allogeneic transplant, Naldini and colleagues engineered a lentiviral vector that directed supraphysiologic levels of ARSA expression in transduced cells. Transduction of autologous HSCs from children born with the disease, at a point when they were still presymptomatic, led to preservation and continued acquisition of motor and cognitive milestones at time periods as long as 32 months after affected siblings had begun to lose milestones. These results illustrate that the ability to engineer levels of expression can allow gene therapy approaches to succeed where allogeneic bone marrow transplantation cannot. It is likely that a similar approach will be used in other neurodegenerative conditions.

Transduction of HSCs to treat the hemoglobinopathies is an obvious extension of studies already conducted but represents a higher hurdle in terms of the extent of transduction required to achieve a therapeutic effect. Trials are now under way for thalassemia and sickle cell disease, and for a number of other hematologic disorders, including Wiskott-Aldrich syndrome, and chronic granulomatous disease.

LONG-TERM EXPRESSION IN GENETIC DISEASE: IN VIVO GENE TRANSFER WITH RECOMBINANT ADENO-ASSOCIATED VIRAL VECTORS

Recombinant adeno-associated viral (AAV) vectors have emerged as attractive gene delivery vehicles for genetic disease. Engineered from a small replication-defective DNA virus, they are devoid of viral coding sequences and trigger very little immune response in experimental animals. They are capable of transducing nondividing target cells, and the donated DNA is stabilized primarily in an episomal form, thus minimizing risks arising from insertional mutagenesis. Because the vector has a tropism for certain long-lived cell types, such as skeletal muscle, the central nervous system (CNS), and hepatocytes, long-term expression can be achieved even in the absence of integration.

FIRST LICENSED PRODUCT

These features of AAV were used to develop the first licensed gene therapy product in Europe, an AAV vector for treatment of the autosomal recessive disorder lipoprotein lipase (LPL) deficiency. This rare disorder (1–2/million) is due to loss-of-function mutations in the gene encoding LPL, an enzyme normally produced in skeletal muscle and required for the catabolism of triglyceride-rich lipoproteins and chylomicrons. Affected individuals have lipemic serum and may have eruptive xanthomas, hepatosplenomegaly, and in some cases, recurrent bouts of acute pancreatitis. Clinical trials demonstrated the safety of intramuscular injection of AAV-LPL and its efficacy in reducing frequency of pancreatitis episodes in affected individuals, leading to drug approval in Europe. Additional clinical trials currently under way that use AAV vectors in the setting of genetic disease include those for muscular dystrophies, spinal muscular atrophy, Parkinson’s disease, Batten’s disease, ornithine transcarbamylase deficiency, hemophilia B and A, several forms of congenital blindness, and a variety of other inherited conditions.

HEMOPHILIA

Hemophilia (Chap. 61) has long been considered a promising disease model for gene transfer, because the gene product does not require precise regulation of expression and biologically active clotting factors can be synthesized in a variety of tissue types, permitting latitude in the choice of target tissue. Moreover, raising circulating factor levels from <1% (levels seen in those severely affected) into the range of 5% greatly improves the phenotype of the disease. Preclinical studies with recombinant AAV vectors infused into skeletal muscle or liver have resulted in long-term (>5 years) expression of factor VIII or factor IX in the hemophilic dog model. Administration to skeletal muscle of an AAV vector expressing factor IX in patients with hemophilia B was safe and resulted in long-term expression as measured on muscle biopsy, but circulating levels never rose to >1% for sustained periods, and a large number of IM injections (>80–100) was required to access a large muscle mass. Intravascular vector delivery has been used to access large areas of skeletal muscle in animal models of hemophilia and will likely be tested as a route of administration for muscular dystrophy disorders in upcoming trials.

The first trial of an AAV vector expressing factor IX delivered to the liver in humans with hemophilia B resulted in therapeutic circulating levels at the highest dose tested, but expression at these levels (>5%) lasted for only 6–10 weeks before declining to baseline (<1%). A memory T cell response to the viral capsid, present in humans but not in other animal species (which are not natural hosts for the virus), likely led to the loss of expression (Table 458-3). In response to these findings, a second trial included a short course of prednisolone, to be administered if factor IX levels began to decline. This approach resulted in long-term expression (>7 years, with observation ongoing) of factor IX, in the range of 2–7%, in men with severe hemophilia B. By using as the transgene a high specific activity variant of FIX, it has been possible to lower the vector dose required, reducing the risk of the immune response, and increasing the plateau levels of FIX into the range of 15–45%. Current efforts are focused on expanding these trials, and extending the approach to hemophilia A.

RETINAL GENE THERAPY

A logical conclusion from the early experience with AAV in liver in the hemophilia trial was that avoidance of immune responses was key to long-term expression. Thus immunoprivileged sites such as the retina began to attract substantial interest as therapeutic targets. This inference has been elegantly confirmed in the setting of the retinal degenerative disease Leber congenital amaurosis (LCA). Characterized by early-onset blindness, LCA was not previously treatable and is caused by mutations in several different genes; ~10% of cases of LCA are due to a mutation in a gene, *RPE65*, encoding a retinal pigment epithelial-associated 65-kDa protein. In dogs with a null mutation in *RPE65*, sight was restored after subretinal injection of an AAV vector expressing *RPE65*. Transgene expression appears to be stable, with the first animals treated >10 years ago continuing to manifest electroretinal and behavioral evidence of visual function. A Phase 3 trial, the first randomized controlled trial in human gene therapy, of an AAV vector encoding RPE65 was completed, and demonstrated improvement in multiple measures of retinal and visual function. This product has now been licensed by the U.S. Food and Drug Administration (FDA), and is the first licensed AAV gene therapy product in the United States. Trials for other inherited retinal degenerative disorders such as choroideremia are under way, as are studies for certain complex acquired disorders such as age-related macular degeneration, which affects several million people worldwide. The neovascularization that occurs in age-related macular degeneration can be inhibited by expression of vascular endothelial growth factor (VEGF) inhibitors such as angiostatin or through the use of RNA interference (RNAi)-mediated knockdown of VEGF. Early-phase trials of an AAV vector designed to achieve long-term inhibition of the biological effects of VEGF through a soluble VEGF receptor, however, failed to provide convincing evidence of efficacy, illustrating the challenges of developing genetic approaches for complex acquired disorders.

GENE THERAPY FOR CANCER

The majority of clinical gene transfer experience has been in subjects with cancer. The intent has been to increase the precision of cancer therapies and thereby make them less toxic and more effective. Most approaches have either modified the tumor directly, or altered the host's response to the malignancy to produce immune effector cells that are precisely targeted to the tumor phenotype.

MODIFYING THE CANCER

Since cancer is an (acquired) genetic disorder, initial efforts were directed at correcting the genetic deficits of the tumor or introducing lethal genes. Two major and persisting obstacles, however, are the poor biodistribution and transduction efficiency of all currently available vectors, and the heterogeneity and genetic instability of the tumor targets themselves, so that correction of single driver mutations does not preclude the evolution of a resistant population.

Tumor Correction

One widely used direct intratumoral approach was adenoviral-mediated expression of the tumor suppressor p53, which is mutated in many different cancers. Initial studies showed some complete and partial responses in squamous cell carcinoma of the head and neck, esophageal cancer, and non-small cell lung cancer, but as yet there have been no successful product licensing studies for this approach except in China.

Pro-Drug Metabolizing Genes

Efforts to overcome the above limitations have included the introduction of a prodrug or a suicide gene that would increase sensitivity of tumor cells to cytotoxic drugs. A frequently used strategy has been intratumoral injection of an adenoviral vector expressing the thymidine kinase (*TK*) gene. Cells that take up and express the *TK* gene can be killed after the administration of gancyclovir, which is phosphorylated to a toxic nucleoside by *TK*. The advantage of this approach is that the effects of transducing even a limited number of tumor cells are amplified by the spread of active drug to adjacent tumor cells. Although the approach continues to be examined in aggressive brain tumors and locally recurrent prostate, breast, and colon tumors, progress remains slow, and systemic benefits against metastatic disease have not been established.

MODIFYING THE HOST

Recruiting the Immune System

The successful use of monoclonal antibodies that produce anti-tumor activity by activating the immune response has demonstrated the feasibility of manipulating the immune system to recognize the abnormal pattern of antigen expression on tumor cells. Immune cells are capable of almost unlimited expansion and persistence and can provide long-term tumor control. They can also traffic to tumor sites irrespective of location and, in principle, have the potential to evolve with the changing pattern of tumor cell phenotype and function.

Vaccination

This strategy promotes more efficient recognition of tumor cells by the immune system, but the development of a therapeutic as opposed to the preventative vaccines required to combat infectious diseases has proved to be a considerable challenge. Approaches have included transduction of tumor cells with immune-enhancing genes encoding cytokines, chemokines, or co-stimulatory molecules, and the ex vivo manipulation of dendritic cells to enhance the

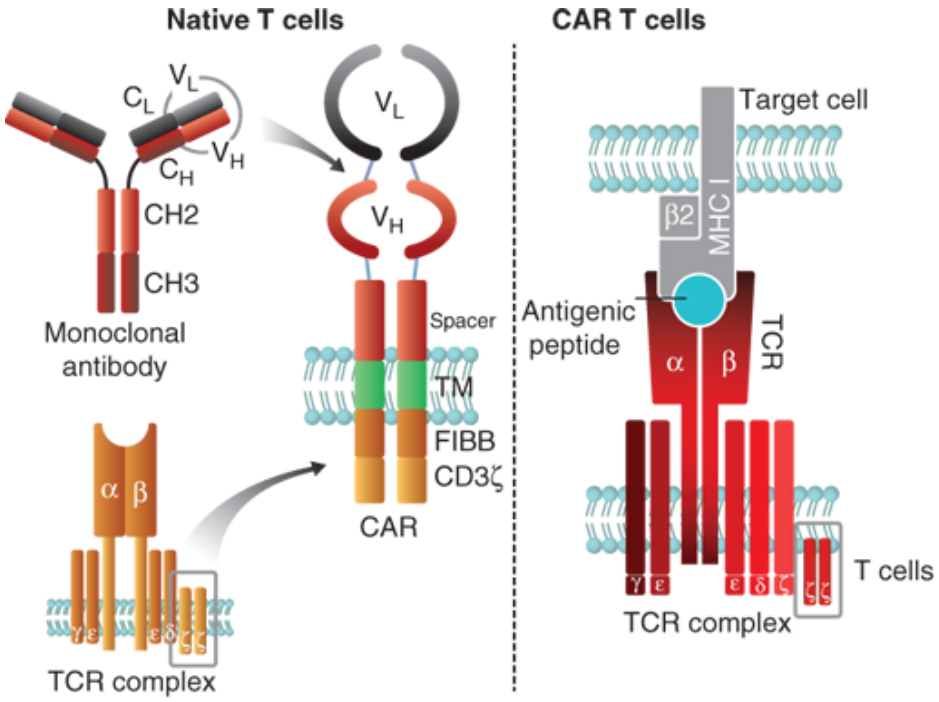
presentation of tumor antigens. A dendritic cell vaccine for treatment of recurrent prostate cancer has received approval in the United States but its limited potency and high cost constrained commercial success.

Adoptive Cell Transfer

Host immune cells such as T cells, NK cells, and others can be modified to express new transgenic receptors intended to recognize tumor cells and their microenvironment (Fig. 458-1). Retargeting may use a modification of the cells' own receptor or a synthetic chimeric antigen receptor (CAR) that is usually composed of the antigen recognition portion of an antibody and the signaling components of the cell's native antigen receptor. Both approaches have been successful, with significant responses reported with native receptors targeted to melanoma and synovial cell sarcoma and—most dramatically—with CARs targeted to CD19, an antigen expressed at high levels on normal and many malignant B cells. Infused CAR T cells can expand many thousand fold in vivo, persist long term, and have produced >90% complete response rates when targeting intractable B-acute lymphoblastic leukemia. Many responses are sustained long term and the approach has been licensed by the U.S. FDA. Broader application of adoptive T cell approaches is limited by: (1) The immune inhibitory microenvironment associated with most tumors; recent studies further modify the T cells with countermeasures to tumor inhibitory signals; (2) Acute and sometimes fatal systemic inflammatory and neurological toxicities during the phase of T-cell expansion and tumor killing; (3) The off-target/or on-target but off-tumor effects that may damage normal host tissues (such as normal B cells following CD19 CAR therapy); (4) The cost, time, and complexity of manufacture; a particular problem when antigens unique to each tumor's individual mutations are targeted (neoantigens), rather than widely shared tumor associated antigens.

FIGURE 458-1

T cell receptors. A native T cell receptor (TCR) recognizes processed peptide antigens bound to major histocompatibility (MHC) molecules through its αβ chains. Signaling then occurs through a multichain intracellular CD3 complex. A chimeric antigen receptor (CAR) usually contains an extracellular receptor component derived from the antigen binding portion (V_H and V_L) of a monoclonal antibody. This produces a receptor that can recognize either protein or non-protein antigens independent of the MHC. A transmembrane (TM) domain then connects this receptor to the ζ chain of the CD3 complex derived from the native TCR.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Non-Immunological Modifications to Host

Gene transfer can be used to protect normal cells from the toxicities of chemotherapy and thereby increase the therapeutic index of these drugs. The most extensively studied approach has been to transduce hematopoietic cells with genes encoding resistance to chemotherapeutic agents, including the multidrug resistance gene *MDRI* or the gene encoding O^6-methylguanine DNA methyltransferase (*MGMT*). Although such approaches reduce hematologic toxicity, cytotoxic dose escalation quickly reveals dose-limiting toxicities to other organ systems.

Finally, gene transfer can be used to inhibit the host angiogenesis required for tumor support, for example by constitutive expression of inhibitors such as angiostatin and endostatin, or the transfer of T cells genetically modified to recognize antigens specific to newly forming vasculature. These studies are early-phase.

COMBINATION APPROACHES—MODIFICATION OF HOST AND TUMOR BY VIROTHERAPY

Immuno Oncolytic Viruses

These viruses are genetically modified to replicate in malignant but not normal cells. The replicating vectors thus proliferate and spread within the tumor, facilitating eventual tumor clearance. However, physical limitations to viral spread, including fibrosis, intermixed normal cells, basement membranes, and necrotic areas within the tumor, may reduce clinical efficacy, and their activity against metastatic disease has proved limited. Recently, the FDA granted licensing approval to talimogene laherparepvec, an oncolytic herpes virus containing the human GM-CSF gene, for treatment of melanoma. This success has led to resurgent interest in combining the local tumor destruction and tumor antigen release mediated directly by oncolytic viruses with the recruitment of a systemic immune response mediated by immunostimulatory genes contained within the oncolytic virus. In principle such immune-oncolytic viruses should produce responses in both local and metastatic disease. Numerous novel viral agents are now entering early phase clinical testing.

OTHER APPROACHES

This chapter has focused on gene addition therapy, in which a normal gene is transferred to a target tissue to drive expression of a gene product with therapeutic effects. Another powerful technique under development is genome editing, in which a mutation is corrected in situ, generating a wild-type copy under the control of the endogenous regulatory signals. This approach makes use of novel reagents including zinc finger nucleases, TALENs and CRISPR, which introduce double-stranded breaks into the DNA near the site of the mutation and then rely on a donated repair sequence and cellular mechanisms for repair of double-strand breaks to reconstitute a functioning gene. These approaches have only recently entered the stage of clinical investigation. Another strategy recently introduced into clinical trials is the use of siRNAs or short hairpin RNAs as transgenes to knock down expression of deleterious genes (e.g., mutant huntingtin in Huntington's disease or genes of the hepatitis C genome in infected individuals).

SUMMARY

The power and versatility of gene transfer approaches are such that there are few serious disease entities for which gene transfer therapies are *not* under development. The development of new classes of therapeutics typically takes two to three decades; monoclonal antibodies and recombinant proteins are recent examples. Gene therapeutics, which entered clinical testing in the early 1990s, traversed the same time course. Examples of clinical success are now abundant, and gene therapy approaches are likely to become increasingly important as a therapeutic modality in the twenty-first century. A central question to be addressed is the long-term safety of gene transfer, and regulatory agencies have mandated a 15-year follow-up for subjects enrolled in gene therapy trials (Table 458-4). Realization of the therapeutic benefits of modern molecular medicine will depend on continued progress in gene transfer technology.

TABLE 458-4

Taking History from Subjects Enrolled in Gene Transfer Studies

Elements of History for Subjects Enrolled in Gene Transfer Trials
1. What vector was administered? Is it predominantly integrating (retroviral, lentiviral, herpesvirus [latency and reactivation]) or nonintegrating (plasmid, adenoviral, adeno-associated viral)? 2. What was the route of administration of the vector? 3. What was the target tissue? 4. What gene was transferred in? A disease-related gene? A marker? 5. Were there any adverse events noted after gene transfer?
Screening Questions for Long-Term Follow-Up in Gene Transfer Subjects[a]
1. Has a new malignancy been diagnosed? 2. Has a new neurologic/ophthalmologic disorder, or exacerbation of a preexisting disorder, been diagnosed? 3. Has a new autoimmune or rheumatologic disorder been diagnosed? 4. Has a new hematologic disorder been diagnosed?

[a]Factors influencing long-term risk include: integration of the vector into the genome, vector persistence without integration, and transgene-specific effects.

FURTHER READING

Aiuti A, Roncarolgo MG: Ten years of gene therapy for primary immune deficiencies. Hematology Am Soc Hematol Educ Program 2009:682, 2009.

Fesnak AD et al: Engineered T cells: The promise and challenges of cancer immunotherapy. Nat Rev Cancer. 16:566, 2016. [PubMed: 27550819]

Hocquemiller M et al: Adeno-associated virus-based gene therapy for CNS diseases. Hum Gene Ther 27:478, 2016. [PubMed: 27267688]

Klebanoff CA et al: Prospects for gene-engineered T cell immunotherapy for solid cancers. Nat Med 22:26, 2016. [PubMed: 26735408]

Kumar SR et al: Clinical development of gene therapy: Results and lessons from recent successes. Mol Ther Methods Clin Dev 3:16034, 2016. [PubMed: 27257611]

Rai P, Malik P: Gene Therapy for hemoglobin disorders—a mini-review. J Rare Dis Res Treat 1:25, 2016. [PubMed: 27891535]

Ramos CA et al: CAR-T cell therapy for lymphoma. Annu Rev Med 67:165, 2016. [PubMed: 26332003]

Chapter 459: The Human Microbiome

Neeraj K. Surana; Dennis L. Kasper

INTRODUCTION

"All disease begins in the gut."

–Hippocrates

Nearly two and a half millennia after Hippocrates made this statement, we are just coming to truly appreciate its profundity. Since the beginning of humankind, scholars have been investigating the underpinnings of disease with an almost singular focus on the human side of the equation. Microbes were not recognized as an important cause of disease until the inception of the "germ theory" in the late nineteenth century. During the first century of medical microbiology, research largely centered on the role of microbes as pathogens. Only recently has there been a resurgence of interest in understanding how commensal organisms—the bacteria, viruses, fungi, and Archaea that make up the *microbiota*—impact human physiology. The idea that these microorganisms are vital to the well-being of humans has challenged our traditional notions of "self." Indeed, a human being can most accurately be described as a *holobiont*: a complex assemblage of human cells and microorganisms interacting in an elaborate *pas de deux* that drives normal physiologic processes.

Aimed at a better understanding of this relationship, myriad studies during the past decade have begun to catalogue the microbiota at various body sites and in a multitude of disease conditions. Diseases in virtually every organ system have been associated with changes in the microbiota. Indeed, the microbiota has been linked to intestinal disorders, disturbances in metabolic function, autoimmune diseases, and psychiatric conditions and has been shown to influence susceptibility to infection and the efficacy of pharmaceutical therapies. Knowledge of the specific mechanism(s) underlying most of these microbe–disease associations is lacking; it remains unclear whether the disease-associated alterations in the microbiota represent mere biomarkers of disease, a causal relationship, or a combination of the two. Although cause-and-effect relationships are still being elucidated for many diseases, it is clear that humans coexist in an intricate relationship with commensal organisms. This chapter explores in detail the nature of these host–commensal interactions, focusing on how this information might be translated into clinically meaningful interventions.

HISTORICAL PERSPECTIVE

Massive undertakings, such as the Human Microbiome Project (HMP) sponsored by the National Institutes of Health and MetaHIT sponsored by the European Commission, have catalogued all the bacteria present at multiple body sites in people with and without disease. Coupled with the confluence of advances in sequencing technologies (Chap. 474), gnotobiotic animal availability, and microbial culture, significant progress has been made toward an understanding of the interplay between the microbiota and human health. However, recent findings were foreshadowed by work done centuries ago.

The human microbiota was first explored in 1683 when Antony van Leeuwenhoek described in a letter to the Royal Society of London the "very little living animalcules, very prettily a-moving" that he had observed in the plaque between his teeth. Leeuwenhoek went on to perform the first comparative "microbiota" studies by assessing how fecal and oral bacteria differ, how oral microbes change in the setting of disease (e.g., alcoholism and tobacco use), and how microbial composition changes across the age spectrum (e.g., in young children versus old men). He attempted—unsuccessfully—to eliminate these bacteria. Although Leeuwenhoek was not taken seriously when he first reported his findings, his studies laid the groundwork for what is now the field of microbiome research, and investigators are still trying to answer many of the same overarching questions that he raised more than three centuries ago.

Although Leeuwenhoek first reported the existence of bacteria and their association with humans at the end of the seventeenth century, the significance of commensal bacteria was not realized until late in the nineteenth century. In 1885, Pasteur suggested that animals could not survive if they were "artificially and completely deprived of the common microbes." Although Pasteur's preconceived ideas were proven incorrect in 1912 by the advent of germ-free animals (animals raised without exposure to any microorganisms), the underlying concept that commensal organisms are critical to health has held up. Élie Metchnikoff made another conceptual advance in this field by suggesting at the beginning of the twentieth century that clinical outcomes could be altered by the administration of specific beneficial organisms (*probiotics*). In particular, Metchnikoff believed that aging was caused by toxic bacteria in the gut and that lactic acid–producing bacteria (e.g., *Lactobacillus* species) present in sour milk and yogurt could mitigate against this process. The data behind this specific claim are still lacking, but recent discoveries offer continued hope that the microbiome can be effectively harnessed to protect against and treat a variety of diseases. Thus, although the field of microbiome research is sometimes considered to

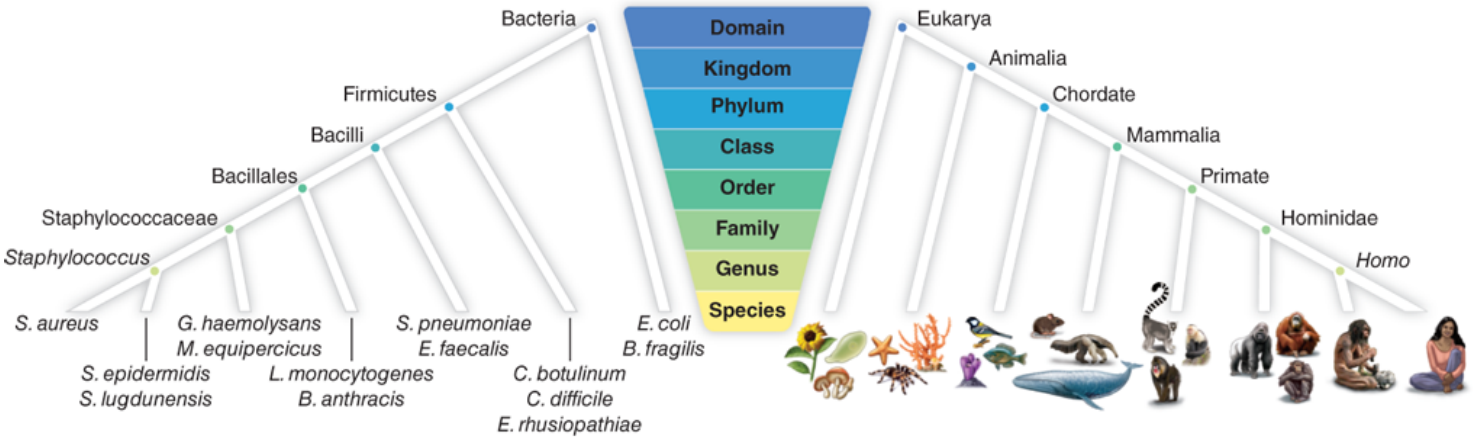
have emerged over the last one or two decades, the basic tenets—that the microbiota varies according to body site and clinical characteristics, that microbes are critical for human health, and that specific modulation of the microbiota may lead to improved clinical outcomes—are far from new.

A PRIMER ON TAXONOMY

Given that microbiome-based studies have identified and compared microbes at different levels of taxonomic resolution **(Fig. 459-1)**, some understanding of taxonomy is essential for better comprehension of the implications of these studies. Of the ~100 bacterial phyla that exist in nature, only five (Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria, and Proteobacteria) are dominant members of the human microbiome. Each of these phyla can be further categorized into multiple classes, orders, families, genera, and species. Early studies on the microbiota focused on changes in the relative abundance at the phylum level between different groups (e.g., obese versus normal-weight patients); however, these comparisons are at such a broad taxonomic level that they often provide little or no biological insight. As illustrated in Fig. 459-1, drawing comparisons between organisms in two different bacterial phyla is analogous to comparing humans to sea stars: the evolutionary distance between the two is tremendous. The limitations of current bioinformatic tools require lumping together of taxonomically related strains and thus cloud the richness of microbial ecology. Examining microbial profiles at the phylum, family, or even genus level—as is often done at present—ignores the great heterogeneity within different strains of the same bacterial species. The analytical pipelines are just now beginning to enable strain-level comparisons, and these improvements will likely facilitate our ongoing investigation of host–commensal interactions.

FIGURE 459-1

Juxtaposition of bacterial and human taxonomy highlights the evolutionary distance between different taxonomic levels. The listed species represent exemplars that are members of the taxon to which they are connected but that are not contained within the next-lower-level taxon listed. For example, *Clostridium botulinum*, *Clostridium difficile*, and *Erysipelothrix rhusiopathiae* are members of the phylum Firmicutes, but are in classes other than Bacilli. Similarly, starfish and humans are both members of the kingdom Animalia, but they are in different phyla.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

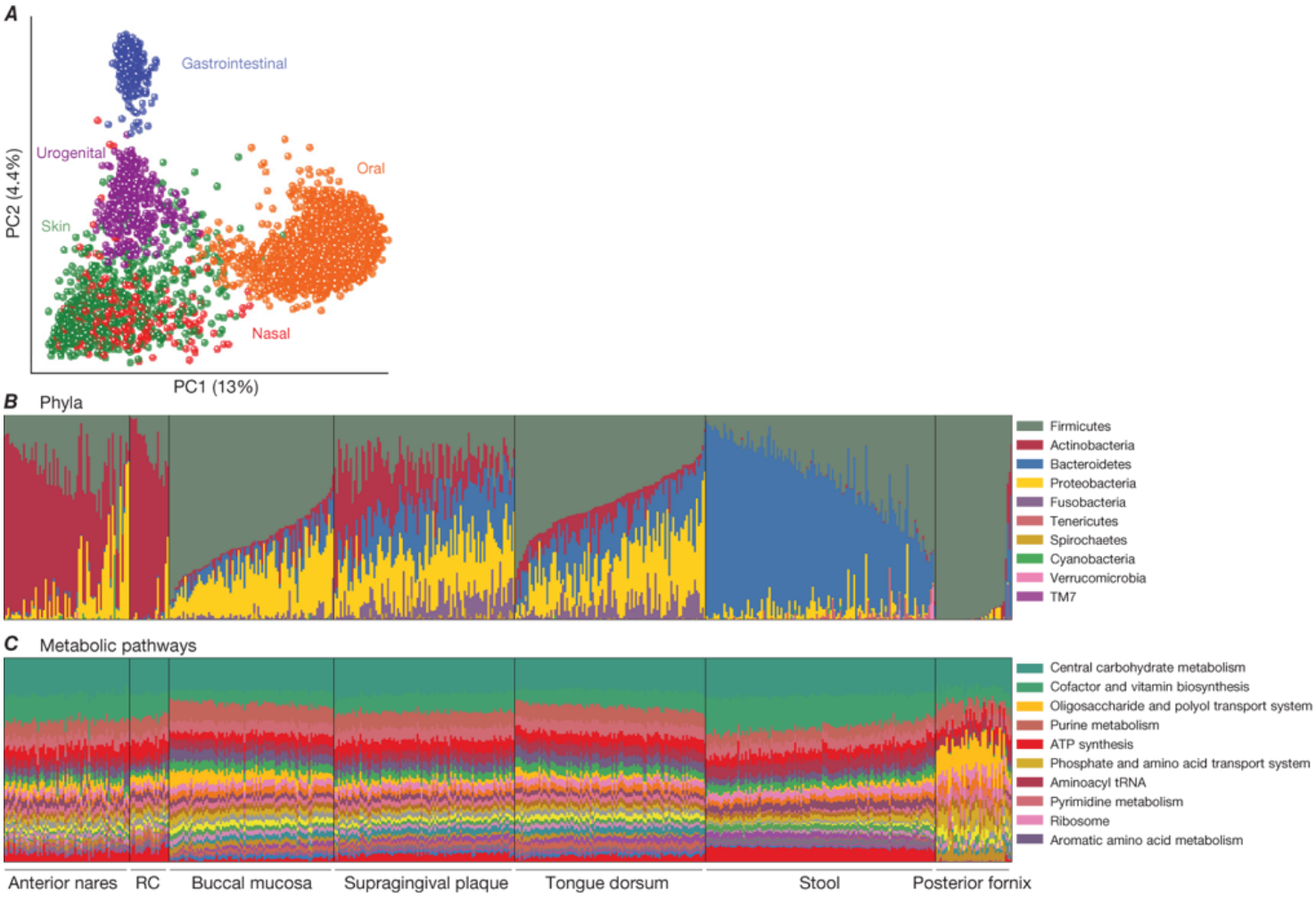
THE MICROBIOTA AND HUMAN HEALTH

OVERVIEW OF THE HUMAN MICROBIOTA

The overwhelming majority of microbiota studies have focused on stool, given that this sample type represents the most ecologically rich anatomic site, is easy to obtain, and can readily be followed longitudinally in the same individual. A landmark study by the HMP sought to define the "normal" microbiota throughout the entire body in healthy Western adults. To this end, the microbial populations at 15–18 body sites were characterized in 242 people. One striking finding was that all samples from a given body region (e.g., skin) were more similar to each other than they were to samples from a different body region (e.g., stool), even in the same individual **(Fig. 459-2A)**. In essence, the effect of the anatomic site on microbial composition is far greater than the effect of heterogeneity between individuals. That said, there was a remarkable amount of inter-individual variation at any given body site **(Fig. 459-2B)**. In stool, for example, the abundance of the phylum Bacteroidetes ranged from ~10% in some individuals to >90% in others. Remarkably, even with person-to-person variability and differences among body sites, the functional capacity of the microbiota—assessed using metagenomic data to identify gene pathways—was quite similar across different people and different body sites **(Fig. 459-2C)**. This discrepancy between the substantial differences in microbial composition and the little or no resulting change in the functional properties of the microbiota reflects an important ecologic property of the microbiota: the microbial communities at different body sites and in different people assemble in such a way that all the core metabolic functions are maintained. This finding also hints at the likely possibility of significant functional redundancy within the microbiota, with different species executing the same biological functions at different anatomic sites.

FIGURE 459-2

The human microbiome exhibits significant taxonomic variability among body sites and between individuals while maintaining core metabolic pathways. ***A.*** Principal coordinates (PC) plot showing variation among samples demonstrates that primary clustering is by body area, with the oral, gastrointestinal, skin, and urogenital habitats separate; the nares habitat bridges oral and skin habitats. Each circle represents an individual sample. ***B, C.*** Vertical bars represent microbiome samples by body habitat, with each bar within a given body site representing a different individual. Bars indicate relative abundances colored by microbial phyla (*B*) and metabolic pathways (***C***). The legend on the right indicates the most abundant phyla/pathways. RC, retroauricular crease. *(Reprinted by permission from Macmillan Publishers Ltd: Human Microbiome Project Consortium: Structure, function and diversity of the healthy human microbiome. Nature 486:207, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

While the HMP provided the first large-scale catalogue of the microbiome in multiple people and at many different body sites, the amount of data generated by what, at the time, was by far the largest microbiome study has been dwarfed by subsequent studies. These more recent studies have confirmed the HMP's major tenets: the composition of the microbiota differs by body site, there is tremendous inter-individual variation, and the microbial gene content is relatively conserved irrespective of the body site or individual. No microbial species are ubiquitous in all individuals and at all body sites, but some species are highly prevalent at a given body site: in the HMP study, *Staphylococcus epidermidis* was present in 93% of nares samples and *Escherichia coli* in 61% of stool samples. These findings highlight the remarkable personalization of the human microbiome. While the human genome is typically >99.5% identical in different people, the microbiotas of two individuals may not overlap at all. Although the "precision medicine" approach currently focuses on teasing out how differences in the human genome relate to different clinical end points, the human microbiome clearly represents a critical component for consideration.

THE MICROBIOTA BY THE NUMBERS

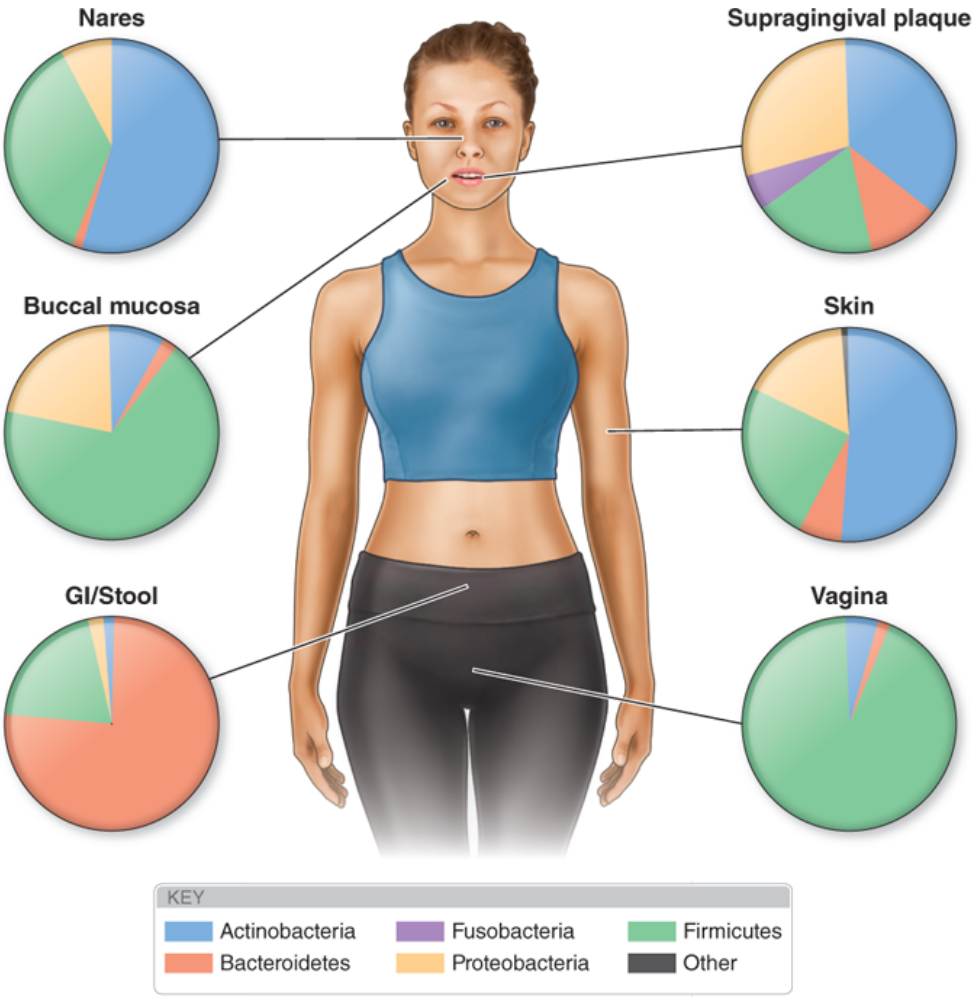
It has long been known that the human-associated microbiota is numerically dense. Leeuwenhoek estimated that there were more "animals living in the scrum on the teeth in man's mouth than there are men in a kingdom." Specific enumeration of the components of the microbiota has been challenging, in part because of its variability across time, space (body region), and clinical conditions. Moreover, the majority of human-associated microbes are not readily cultivable—a situation that raises questions about the best methodology for such quantitation. Initial back-of-the-envelope calculations performed in the 1970s suggested that there were roughly tenfold more bacteria in the body than there were human cells. This rather astounding estimate suggested that humans are really only ~10% "human" and that by far the greatest part of the holobiont is represented by

microbes. This stark numerical discrepancy has prompted some to question “who parasitizes whom.” However, a more recent estimate has suggested that there are “only” ~1.3 times more bacteria in the body than there are human cells and thus that humans are ~56% “bacterial.” Of note, this more recent study does not include the numbers of viruses (known to generally be approximately tenfold more abundant than other microbes), fungi, or Archaea. Given these additional microorganisms, the notion that microbes constitute >90% of the cells present in a human body is likely correct. These ratios are even starker when one considers the genetic potential of human cells versus that of commensal organisms. In contrast to the ~20,000 genes in the human genome, the estimated total number of genes in the microbiota (which together constitute the *microbiome*)—i.e., >2,000,000—indicates that the human genome contributes <1% to the total genetic potential of the overall holobiont. Most microbiome studies to date have focused almost exclusively on the bacterial component; much remains to be learned about the functional interplay of bacteria, viruses, fungi, and Archaea and how these other classes of microorganisms impact human health.

In terms of overall diversity, >10,000 different bacterial species are present in the human microbiota; the intestines alone contain >1000 species. At any given time, the body of any given individual harbors 500–1000 bacterial species, with 100–200 bacterial species in the gut alone. If one considers different strains of the same bacterial species, which may be functionally different from one another, the diversity of the microbiota is probably at least a magnitude greater. Although marked diversity exists at the strain and species level, only limited bacterial phyla are generally found in the human microbiota at any given body site **(Fig. 459-3)**.

FIGURE 459-3

Different anatomic sites harbor very different microbiomes. The figure indicates the relative proportion of sequences determined at the taxonomic phylum level at six anatomic sites. *(Data for stool, vagina, nares, buccal mucosa, and supragingival plaque are from the Human Microbiome Project; data for the skin is from EA Grice et al: Topographical and temporal diversity of the human skin microbiome. Science 324:1190, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

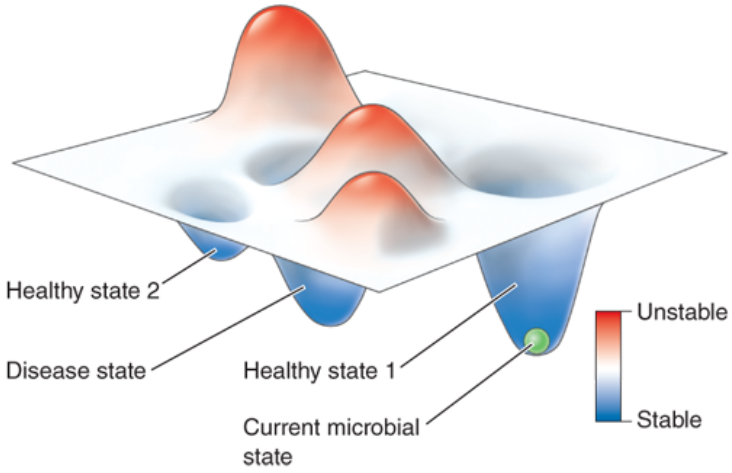
INFLUENCES ON THE MICROBIOTA

An individual’s specific microbial configuration is dynamic and is quickly altered in response to subtle changes in the microenvironments in which the bacteria reside. On a day-to-day basis, these changes usually reflect alterations in the relative abundance of the various microbes. However, some exposures have a greater effect on the microbiota and can shift the microbial population to a new equilibrium via the loss of specific species and/or the

acquisition of others; this new microbial equilibrium can be associated with either health or a disease state (Fig. 459-4). Identification of the factors that influence the microbiota's composition is critical to an understanding of what leads to and controls intra- and inter-individual variation. Moreover, an understanding of the influences on the microbiota will facilitate the design and proper interpretation of microbiota studies. While it is clear that the microbiota can be altered through these various mechanisms, it is not yet clear whether these changes are biologically significant.

FIGURE 459-4

A stability landscape of the human microbial ecosystem. A stable state, illustrated as a depression in the landscape, can be associated with either a healthy state or a disease state. The topology of an individual's landscape reflects that person's genetics, age, diet, medications, medical history, and lifestyle. The position of the green ball represents the current microbial state. Clinical changes (e.g., administration of antibiotics, development of disease) can influence both the current state and the overall topology.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Genetics

Studies of monozygotic and dizygotic twins have revealed that host genetics have a small but statistically significant effect on the microbiota's composition. Notably, some taxa, such as *Christensenella* species, are more heritable than others. Thus, some inter-individual variation may be due to underlying differences in host genetics; however, it is likely that other factors explain more of the observed variability. That said, the host's genetic contribution to the microbiota, albeit small, may be meaningful. Studies in mice have demonstrated that genetic variation in the major histocompatibility complex, a specific set of immune-related genes, leads to changes in the microbiota that alter susceptibility to an autoimmune disease. These studies offer a proof of concept for the notion that the genetic predisposition observed for certain diseases may actually be mediated by indirect alterations in the microbiota.

Age

Burgeoning evidence now indicates that microbial exposure begins in utero: bacterial DNA has been identified in otherwise healthy placentas, in amniotic fluid obtained at early stages of gestation, and in meconium of term newborns. Although some controversy persists about whether these results reflect contamination and/or the presence of nonviable bacteria, they raise the possibility that human exposure to the microbial world begins before birth. The delivery mode (vaginal versus cesarean section) and the method of feeding (breast milk versus formula, timing of solid food introduction) are major determinants of an infant's early microbiota. After birth, the infant's microbiota goes through a stereotyped succession process; with increases in bacterial diversity and functional capacity, the child's microbiota resembles that of an adult by the age of 2–3 years. Cross-sectional studies that have examined the microbiota across the entire age spectrum have revealed a general stability of the fecal microbiota after 2–3 years of age; however, the microbiota of the elderly (persons >80 years of age) demonstrates notable differences from those of their younger counterparts, with increases in *Bacteroides* and *Eubacterium* species and decreases in the bacterial family Lachnospiraceae.

Diet

Diet is a strong determinant of human health. The impact of diet is mediated, in part, by its effects on the composition of the gut microbiota. This makes intuitive sense, as the human diet provides nutrients needed not only by our own cells but also by the microbes living in the alimentary tract. In young children, this dietary influence is marked by major shifts (e.g., a decrease in *Bifidobacterium* species) in the intestinal microbiota that occur at weaning and with the introduction of solid food. In adults, long-term dietary patterns are associated with relatively stable microbial compositions. However, drastic changes in short-term macronutrient availability cause rapid (within 1 day) and reproducible fluctuations in the fecal microbiota that reflect the biological processes needed to degrade and metabolize the nutrients in the new diet. For example, vegetarian diets are associated with a microbiota

that has an increased ability to metabolize plant polysaccharides (e.g., *Roseburia* species, *Eubacterium rectale*, *Ruminococcus bromii*), while animal-based diets result in an increased abundance of bile-tolerant organisms (e.g., *Alistipes*, *Bilophila*, and *Bacteroides* species). At the completion of dietary interventions and the resumption of the individual's normal dietary pattern, the microbial communities revert back to their previous states, probably because the individual resumes his or her typical diet. Taken together, dietary studies confirm that the microbiota is highly adaptable and varies in relation to changes in the diet. Of note, virtually all of these studies have focused on how the diet influences the fecal microbiota. It will be interesting to determine whether dietary changes similarly influence the microbiota at nonintestinal sites.

Drugs

Virtually all drugs have the capacity to change the microbiota by altering the chemical landscape in which the microorganisms live (e.g., statins, bile acid sequestrants), modulating the host's ability to recognize and react to microbes (e.g., immunosuppressants), and/or directly interfering with the microbiota's constituents (e.g., antibiotics). These potential effects have made critical interpretation of microbiota studies much more difficult. A prominent study that claimed to identify a fecal microbiota signature associated with type 2 diabetes was later found actually to have identified a signature for patients taking metformin instead; the effects of this drug on the microbiota were far greater than the effects of the disease itself. These results highlight the importance of controlling for clinical variables in microbiota studies.

Antibiotics are the most obvious and best-studied class of drugs that modulate the microbiota. Multiple groups have demonstrated that antibiotics exert a considerable effect on the gut microbiota by depleting antibiotic-sensitive strains. What is more surprising is that many strains resistant to the antibiotic tested are also eliminated. This observation highlights the intricate microbe–microbe interactions that are fundamental to maintenance of the overall microbial community. For example, treatment with ciprofloxacin, which has little to no activity against clinically relevant anaerobes, leads to a loss of roughly one-third of the bacterial taxa in the gut. This broad effect is likely mediated by the depletion of certain "keystone" species that are required for the persistence of other, unrelated species. While many of the observed antibiotic effects (e.g., loss of specific taxa) are shared across many different individuals, some effects vary greatly among people. For example, studies found that microbiota recovery following antibiotic treatment differed significantly in terms of timing and degree. The microbiota of most healthy people who received ciprofloxacin for 5 days had completely recovered within 4 weeks, whereas microbiologic changes lasted up to 6 months in other individuals. Moreover, the degree of variation was compounded by repeated antibiotic administration, with fewer individuals reverting to their baseline microbiota after a second course of ciprofloxacin given 6 months after the first. These findings are consistent with those of microbial ecology experiments, which also showed that this type of repeated disturbance leads to less predictable results.

Lifestyle

Many seemingly innocuous lifestyle decisions can impact the human microbiota. For example, a person's skin and fecal microbiotas are more similar to those of their household members, regardless of genetic relatedness, than to those of residents of different households. The degree of similarity in skin microbiotas is even greater if a dog also lives in the home; in contrast, the presence of a young child does not accentuate this microbial relatedness. The presumption is that the dog serves as a more effective "vector" for transmitting microbes during its frequent direct contact with adults in the household. The type of setting in which a person lives also impacts the microbiota. Living in a rural or farm setting leads to a different fecal microbiota than living in an urban environment. Similarly, the individual's country of residence affects the microbiota. An analysis of daily fecal samples from an individual who temporarily (i.e., for a couple of months) moved from the United States to Thailand demonstrated a large shift in the fecal microbiota that coincided with arrival in Thailand and a reversion in most respects to the "American" microbial configuration upon return to the United States. These geography-driven changes probably reflect a combination of environmental and dietary differences between locations.

Circadian Rhythms

Many human biological processes follow a circadian clock; aspects of physiology are tuned by external cues, including the degree and timing of ambient light, temperature, and availability of nutrients. This endogenous biological clock enables animals to efficiently adapt to changing environmental conditions. Similarly, the microbiota maintains a circadian rhythm that is linked to the host's circadian clock. If circadian oscillations are disrupted in the host, they are similarly disrupted in the microbiota, and vice versa. These bacterial vacillations occur at the level of spatial localization within the intestine, relative species abundance, and bacterial metabolite secretion. Work in the 1960s showed that mice exhibited daily periodicity of susceptibility to infection with either *S. pneumoniae* or *E. coli* lipopolysaccharide. Although the fundamental basis for this difference was not known at the time, it is likely to be related, in part, to the microbial circadian clock. Derangements of these microbial oscillations have also been linked to the development of metabolic diseases and may underlie some of the health hazards associated with shift work and jet lag.

THE MICROBIOTA AND DISEASE

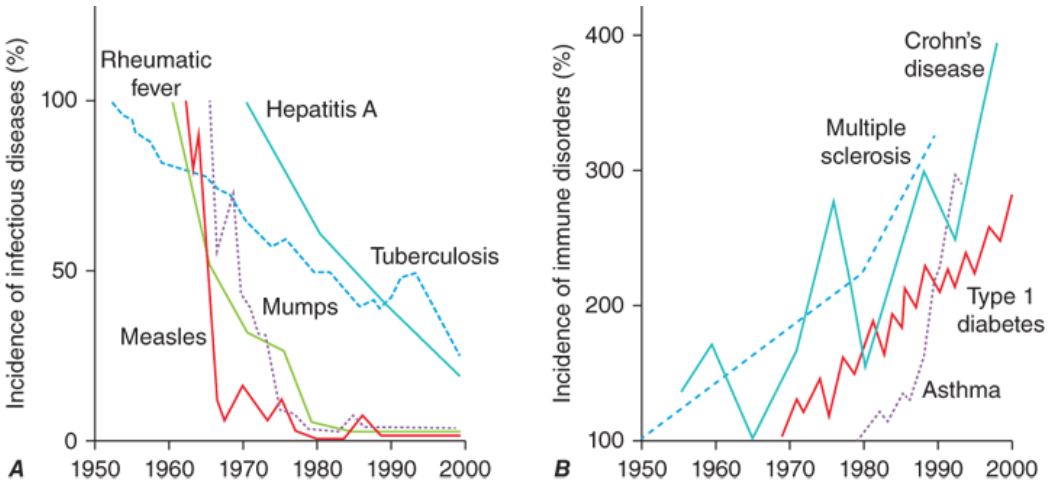
THE HYGIENE HYPOTHESIS

Over the past few decades, abundant epidemiologic data have revealed an inverse correlation between exposure to microbes and the incidence of autoimmune and/or atopic diseases **(Fig. 459-5)**. This type of epidemiologic correlation led to the proposal of the "hygiene hypothesis" in 1989. Initially,

this hypothesis focused on the development of atopic diseases in young children, with the idea that these epidemiologic observations could "be explained if allergic diseases were prevented by infection in early childhood, transmitted by unhygienic contact with older siblings, or acquired prenatally from a mother infected by contact with her older children."[1] In fact, this notion that differences in living conditions and environmental exposures contribute to susceptibility to hay fever (*summer catarrh*) dates back to at least the early nineteenth century. The hygiene hypothesis has continued to evolve over the past three decades and now posits that inadequacies in microbial exposure—in combination with genetic susceptibilities—lead to a collapse of the normally highly coordinated, homeostatic immune response. At its core, the hygiene hypothesis holds that specific early-life microbial exposures are required to prevent subsequent disease and that the "Westernization" of society has led to a decrease in such exposures. This concept is now being applied beyond atopic diseases to other inflammatory and autoimmune diseases and is thought to reflect processes that occur in later life as well.

FIGURE 459-5

There was an inverse relationship between the incidence of select infectious diseases and the incidence of autoimmune disorders during the latter half of the twentieth century. ***A***. Relative incidence of prototypical infectious diseases from 1950 to 2000. ***B***. Relative incidence of select autoimmune disorders from 1950 to 2000. *(From JF Bach: The effect of infections on susceptibility to autoimmune and allergic diseases. N Engl J Med 347:911, 2002. Copyright ©2002, Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

[1]D Strachan: BMJ 299:1259, 1989.

RELATIONSHIP BETWEEN THE MICROBIOTA AND SPECIFIC DISEASE STATES

The ideas inherent in the hygiene hypothesis—in sum, that microbial exposure can affect long-term health outcomes—laid the theoretical foundation for translational microbiome studies. While most of the studies described earlier sought to describe how the microbiota responds to specific and often transient influences (e.g., a course of antibiotics, dietary interventions, travel), a multitude of studies have characterized the microbiota in patients with various diseases in the hope that a better understanding of the nature of disease-specific microbial communities will provide insight into disease pathogenesis and potentially uncover novel treatment modalities. Remarkably, virtually all of these studies have demonstrated differences between the microbiotas of healthy controls and patients, irrespective of the specific disease process examined. Although it is difficult to generalize across all studies, a couple of general themes have emerged. First, disease states are typically associated with microbiotas that are less diverse than those of healthy individuals. This loss of diversity can be measured either as a decrease in the number of species (*alpha diversity*; often measured as the number of operational taxonomic units, which are the bioinformatic equivalent of species) or as a reduction in the microbial relatedness of the species present (*beta diversity*). Often, both alpha and beta diversity decrease in the setting of disease. Second, states of inflammation—regardless of site or underlying disease process—are often associated with a relative increase in the abundance of the bacterial family Enterobacteriaceae and a decrease in the abundance of Lachnospiraceae.

Dissecting Correlation and Causality

Given that most of these investigations have been designed as case–control studies, it is difficult to determine whether microbiologic findings are the cause or the effect of the disease. Even studies that examine treatment-naïve patients at the time of initial diagnosis are still confounded by this "chicken or egg" issue. Moreover, prospective, longitudinal clinical studies—still rare in the microbiome field—may simply yield correlations between the microbiome and subclinical disease rather than necessarily proving causality. Experiments in animals—specifically, studies using gnotobiotic mice (germ-free mice that have been colonized with specified microbial communities)—have been critical in this regard as they allow investigation of specific differences in microbial components while controlling for the host's genetics, diet, and housing conditions. Moreover, human microbes can be

transplanted into gnotobiotic mice to permit in-depth mechanistic studies of how these microbial communities affect disease pathogenesis. This marriage of human samples and animal experiments has facilitated the identification of causal roles played by some microbes in disease pathogenesis; these findings provide a critical proof of concept for the interplay of the microbiota with human health. However, the vast majority of microbiome studies are still at the level of correlation. The next several sections describe the clinical and animal data for many different disease processes. Given the voluminous and rapidly changing nature of this field, it is impossible to cover all of the disease associations known to date; rather, the following discussion represents a combination of the leading exemplars of microbiome data and nascent areas of significant clinical interest. In all cases, the hope is that further study of the role of the microbiota will provide novel diagnostics, new therapeutic modalities, and/or additional insight into disease pathogenesis.

Gastrointestinal Diseases

Given that the intestines harbor the largest number and greatest diversity of organisms in the body, much work has focused on how the microbiota impacts gastrointestinal diseases. Even though the luminal surface area of the gastrointestinal tract is 30–40 square meters (~90% of which is contained within the small intestine) and features marked anatomic and functional differences that result in many discrete macro- and micro-ecosystems, stool is often used as a surrogate for the intestinal microbiota given the relative ease of collecting samples. A few studies that have compared the microbial profile in stool with the mucosa-adherent organisms present in biopsy samples have demonstrated that stool is, in fact, a reasonable proxy for biopsy samples; however, the relative microbial “noise” present in stool can sometimes overwhelm the “signal,” making biopsy samples more informative for some scientific questions. The key issue is to ensure that the biopsy samples evaluated represent relatively similar intestinal regions, as there are significant differences between the organisms present in the crypt and the tip of the villus and between microbes found in the ascending versus the descending colon.

OBESITY

Obesity is a worsening epidemic throughout the world, and multiple studies have linked the composition of the intestinal microbiota to the development of obesity in animal models and in humans. Indeed, many of the initial translational microbiome studies performed in mice at the beginning of the twenty-first century focused on obesity. These early studies suggested that the ratio of the relative abundance of Bacteroidetes to Firmicutes was lower in obese mice than in control animals. Moreover, a causal relationship between the microbiota and obesity was established by the finding that gnotobiotic mice colonized with the microbiota from obese individuals had more rapid and more extensive weight gain than gnotobiotic mice colonized with the microbiota from lean individuals. Biologically, it is posited on the basis of metagenomic surveys that the obesity-associated microbiome has an increased capacity to harvest energy from the diet. Notably, the relationship between the Bacteroidetes/Firmicutes ratio and obesity did not hold in initial human studies; however, the finding that this ratio increased in obese patients who lost weight while on a fat- or carbohydrate-restricted diet suggested some generalizability between mice and humans. Beyond this ratio of major bacterial phyla, obesity was linked to a microbiome with a lower alpha diversity. Over the past ~15 years, numerous human studies examining the relationship between the microbiome and obesity have been completed, all with mixed results. A recent meta-analysis of 10 studies including nearly 3000 individuals revealed an apparent lack of relationship between the Bacteroidetes/Firmicutes ratio and obesity, though there is ~2% lower diversity associated with obesity that is statistically significant but of unclear biological significance. This finding highlights a problem common to microbiome studies: i.e., there is no sense as to what magnitude of change is biologically meaningful. Ultimately, although murine studies have indicated a causal link between the microbiota and obesity, the human data are less convincing, and their significance may be limited by the studies’ having primarily examined only high-level taxonomic information rather than also assessing transcriptional or metabolic differences.

The rise in obesity has elicited a plethora of ideas about the type of diet that might be most successful in leading to sustained weight loss. It has become clear that the same dietary ingredient can have highly diverse effects on blood glucose measurements in different people and that this effect is mediated largely by the microbiome. These observations suggest that the “optimal” diet needs to be individualized in the context of the person’s microbiome, which itself may continue to change over the course of the diet. An intriguing parallel question is whether the microbiota may also influence dietary preferences; such an influence would suggest important feedback loops between the microbiome and diet.

MALNUTRITION

Representing the other end of the metabolic spectrum from obesity, malnutrition is also linked to an altered microbiome. Analysis of Malawian twin pairs (≤3 years of age) who were discordant for kwashiorkor—a severe form of malnutrition—revealed that kwashiorkor is associated with a microbiologically “immature” fecal microbiota that resembles that of a chronologically younger child. Transplantation of the fecal microbiota from these discordant twins into gnotobiotic mice that were fed a diet similar in composition to a typical Malawian diet established that the kwashiorkor-associated microbiome is causally related to poor weight gain. Subsequent studies demonstrated these same general trends in malnourished Bangladeshi children. Investigators were able to identify five bacterial species (*Faecalibacterium prausnitzii*, *Ruminococcus gnavus*, *Clostridium nexile*, *Clostridium symbiosum*, and *Dorea formicigenerans*) that—when administered together as a “cocktail” to mice colonized with a kwashiorkor-associated microbiome—was able to prevent growth impairments. These results demonstrate that rationally designed modulation of the murine microbiota can lead to improved health outcomes. The clinical significance of these findings for humans remains to be clarified.

INFLAMMATORY BOWEL DISEASE

Ulcerative colitis and Crohn’s disease, the two predominant forms of inflammatory bowel disease (IBD), are chronic gastrointestinal inflammatory conditions that differ in their locations and patterns of inflammation (Chap. 319). The following observations have led to the suggestion that IBD is the

result of an immune response to a dysbiotic microbiota in a genetically susceptible individual: genes account for only ~20% of susceptibility to IBD (and many of the relevant genes are related to host–microbe interactions), antibiotic treatment reduces the clinical severity of disease, and relapses of Crohn's disease are prevented by diversion of the fecal stream. While the microbiota clearly is not the only driver of disease, it is considered to be an important element. Accordingly, numerous animal and clinical studies have been designed to tease out the nature of the relationship between the microbiota and IBD.

Most of these studies have focused on comparing the microbiome's composition in IBD patients with that in healthy controls, concentrating on microbial diversity and specific bacterial taxa that are associated with health or disease. Unfortunately, few, if any, results have been universally obtained, probably because of differences in study design, inclusion criteria, and methodology (e.g., the use of stool, rectal swabs, or biopsy samples; the choice of sequencing primers; the analysis pipeline). Even with these differences among studies, patients with IBD have been shown typically to have reduced alpha and beta diversity in their fecal microbiotas. Moreover, *Clostridium* clusters IV and XIVa, which are polyphyletic and encompass several different bacterial families, are generally reduced in patients with IBD. *F. prausnitzii* is a notable example from *Clostridium* cluster IV that is often underrepresented in the stool of patients who have Crohn's disease, with more mixed results in biopsy samples. The bacterial family Lachnospiraceae, which is largely contained in *Clostridium* cluster XIVa, and other butyrate-producing organisms are also reduced in the stool of patients with IBD. Some of these species produce butyrate by using acetate generated by other members of the microbiome, and some of these acetate-producing species are similarly reduced (e.g., *Ruminococcus albus*). These complex interactions and dependencies among bacterial species pose unique challenges to definitive ascertainment of the cause–effect relationships between microbes and disease. Even before researchers were able to assess the entire microbiome at once, they often noted that patients with Crohn's disease had a higher representation of adherent invasive *E. coli* in the ileal mucosa, an observation consistent with the increased abundance of Enterobacteriaceae seen in more recent microbiome studies. Beyond bacteria, burgeoning evidence supports a role for Caudovirales bacteriophages in IBD pathogenesis, though these findings may merely reflect the underlying dysbiosis related to the loss of bacterial diversity in IBD. Moreover, some data suggest that IBD is also associated with fungal dysbiosis; several studies have demonstrated an increased ratio of Basidiomycota to Ascomycota. It is still unclear whether any of these microbial associations reflect the cause of IBD or merely serve as biomarkers of disease.

Studies of antibiotic-treated mice and gnotobiotic mice colonized with IBD-associated microbiotas have been useful in confirming that the microbiota affects colitis severity. Several bacterial species have been identified as either promoting colitis in mice (e.g., *Klebsiella pneumoniae*, *Prevotella copri*) or protecting against it (e.g., *Bacteroides fragilis*, *Clostridium* species); however, these organisms do not always correlate with the taxa identified as differentially abundant across multiple clinical studies. In contrast, IgA-coated commensal organisms isolated from patients with IBD promote more severe colitis in mice than either IgA-uncoated bacteria from patients with IBD or IgA-coated bacteria from healthy controls. These data suggest that functional categorization of the microbiota based on immune recognition (e.g., IgA coating) may be a useful approach for identifying pathogenic organisms.

Cardiovascular Disease

Inflammation helps drive the pathogenesis of atherosclerosis, and it has long been postulated that microbes are involved in the atherosclerotic process. Early work demonstrated that patients with cardiovascular disease have higher titers of antibody to *Chlamydia pneumoniae* than control patients, that *C. pneumoniae* is present within atherosclerotic lesions, and that *C. pneumoniae* can both initiate and exacerbate atherosclerotic lesions in animal models. This type of analysis has been extended to other bacteria, such as *Porphyromonas gingivalis*, with the idea that multiple different bacteria may play some role in the pathogenesis of atherosclerosis.

More recent studies have demonstrated clinical correlations between serum levels of trimethylamine N-oxide (TMAO) and atherosclerotic heart disease. Given that red meat, eggs, and dairy products are important sources of carnitine and choline (both precursors of TMAO), it is not surprising that levels of TMAO are higher in omnivores than in vegans. Animal studies have confirmed that transfer of the gut microbiota from atherosclerosis-susceptible strains of mice to atherosclerosis-resistant animals leads to increased serum levels of TMAO and a dietary choline-dependent increase in atherosclerotic plaques; this observation confirms the role of the gut microbiota in the generation of TMAO and atherosclerosis. Moreover, treatment of atherosclerosis-susceptible strains of mice with a structural analogue of choline that inhibits the first enzymatic step in TMAO formation leads to decreased circulating TMAO levels and, more importantly, restrains macrophage foam-cell formation and atherosclerotic lesion development. In a study of more than 4000 patients, plasma TMAO levels were also predictive of incident thrombosis risk (myocardial infarction, stroke). Gnotobiotic animals were used to demonstrate that this risk was dependent on the microbiota; although eight bacterial taxa were identified as being associated with both plasma TMAO levels and thrombotic risk, organisms with choline-utilization genes that represent the first step of TMAO production were not more abundant in animals at greater risk for thrombosis. This discrepancy highlights the complexity of the microbiota and suggests that other aspects of the overall dynamics of the microbial community may be in play.

Oncology

Recent studies exploring the link between the microbiota and cancer have demonstrated that specific members of the microbiota can affect treatment efficacy in both a positive and a negative manner. For example, therapy with antibody to programmed cell death ligand 1 (anti-PD-L1) has proven highly effective for a number of different cancers (Chap. 69); however, a significant proportion of patients do not respond even when their tumors have high

PD-L1 expression levels, a prerequisite for this type of checkpoint blockade inhibition. Using a murine melanoma model, investigators showed that variations in the microbiota resulted in differences in melanoma growth, an impact that was accentuated by anti-PD-L1 therapy. Ultimately, *Bifidobacterium* species were bioinformatically associated with improved anti-tumor responses, and administration of a "cocktail" of *Bifidobacterium* species (*B. bifidum*, *B. longum*, *B. lactis*, and *B. breve*) to melanoma-susceptible mice resulted in improved tumor-specific immunity and responses to anti-PD-L1 therapy. In a separate set of studies involving both patient data and animal experiments, the efficacy of therapy with antibody to cytotoxic T lymphocyte–associated antigen 4 (anti-CTLA-4) was associated with T-cell responses specific for either *Bacteroides theta-iotaomicron* or *B. fragilis*. In particular, administration of *B. fragilis* to germ-free or antibiotic-treated mice restored the normally absent anti-cancer response to anti-CTLA-4 therapy. While both of these examples demonstrate potentiation of anti-cancer therapies by the microbiota, other therapies can be antagonized. Some cancers, such as pancreatic ductal adenocarcinoma, contain intratumor bacteria, particularly Gammaproteobacteria, that can metabolize the chemotherapeutic agent gemcitabine and thereby contribute to the drug resistance of these tumors. Overall, these examples highlight the microbiota's critical impact—both direct and indirect—on the efficacy of drugs. Many other notable examples have been described (e.g., involving cyclophosphamide, digoxin, levodopa, and sulfasalazine), and many more likely remain to be discovered.

The application of microbiome science to hematopoietic stem cell transplantation (HSCT) is an area of expanding interest, particularly given the significant morbidity and mortality related to graft-versus-host disease (GVHD). In light of studies in the 1970s showing that germ-free mice developed less frequent and less severe gut GVHD than wild-type mice, clinicians began to use antibiotics to decontaminate the gut of patients undergoing HSCT. This decontamination approach yielded mixed results, probably because of differences in the antibiotic regimens used. The natural history of patients undergoing allogeneic HSCT includes a substantial loss of diversity in the fecal microbiota, with lower levels of bacterial diversity associated with increased mortality. Moreover, a retrospective analysis of ~850 patients undergoing allogeneic HSCT revealed that receipt of imipenem-cilastatin or piperacillin-tazobactam for neutropenic fever was associated with increased GVHD-related mortality at 5 years; this observation suggested that specific bacteria may help protect against GVHD-related mortality. More detailed analyses revealed an association between the abundance of *Blautia* species and protection against GVHD and mortality, though this correlation is still being examined with regard to its causal relationship. Despite significant interest in examining these microbial relationships with HSCT, little has yet been studied in the context of solid organ transplantation, which likely represents the next frontier of transplantation-related microbiome investigation.

Autoimmune Diseases

The dramatic rise in the incidence of many autoimmune diseases over the past few decades has been far more rapid than can be explained simply by genetic factors (Fig. 459-5). It is increasingly thought that environmental triggers, including the microbiome, are partially responsible for the development of these autoimmune diseases.

TYPE 1 DIABETES

Type 1 diabetes (T1D) is an autoimmune disorder characterized by T cell–mediated destruction of insulin-producing pancreatic islets **(Chap. 396)**. There is a clear genetic predisposition for the disease: ~70% of patients with T1D have human leukocyte antigen (HLA) risk alleles. However, only 3–7% of children with these risk alleles actually develop disease, an observation that suggests a role for other environmental factors. Studying a prospective, densely sampled, longitudinal cohort of at-risk, HLA-matched children from Finland and Estonia, investigators detailed changes in the microbiota prior to development of disease. Although only 4 of the 33 children studied developed T1D within the time frame of the study, a marked decrease of ~25% in alpha diversity occurred after seroconversion but before disease diagnosis. The low number of cases in this study unfortunately precluded identification of any specific disease-associated taxa. A follow-up study compared the microbiomes of a larger cohort of these high-risk northern European children with those of low-risk Russian children who lived in geographic proximity. *Bacteroides* species were more abundant in the high-risk group than in the low-risk group, particularly at early ages. This difference was postulated to be associated with an altered structure of the bacterial lipopolysaccharide to which children were exposed at a young age. It was further suggested that *Bacteroides*-derived lipopolysaccharide was not able to provide the immunogenic stimulus necessary to prevent T1D. These two studies offer attractive—though logistically complicated—options for future clinical investigations aimed at exploring the role of the microbiome. The first approach—longitudinally following individuals who are at high risk for a given disease—may provide insight into host–microbe relationships by mapping temporal changes in the microbiome with disease onset. An important caveat with this type of study, though, is that the associations identified may reflect preclinical disease rather than specifically indicating causality for any observed changes. The second approach illustrates how careful selection of study participants may offer an opportunity to uncover more meaningful associations that can subsequently be experimentally verified.

RHEUMATOID ARTHRITIS

Similar to many other autoimmune diseases, rheumatoid arthritis (RA) is a multifactorial disease that comes to clinical attention after an environmental factor triggers symptoms in an individual with pre-existing autoantibodies. Multiple lines of evidence support the notion that RA pathogenesis is reliant on the microbiota, including the findings that germ-free mice do not develop symptoms in several RA models and that antibiotic treatment of mice mitigates against RA development. Several taxa (e.g., *Bacteroides* species, *Lactobacillus bifidus*, and segmented filamentous bacteria) have been implicated in promoting RA in murine models, and analysis of the fecal microbiota of patients with newly diagnosed RA have indicated that *P. copri* is a biomarker of disease. That this association with *P. copri* does not exist for chronic, treated RA or for psoriatic arthritis suggests some specificity for new-onset RA. A major limitation of this approach is that the identified association is shown to be a biomarker of disease (and, in this case, potentially of response to treatment) but no added insight is gained into a possible causal relationship between *P. copri* and RA. In fact, many of the patients with

new-onset RA had no *Prevotella* detected, and several of the healthy controls had significant levels of *Prevotella*. The lack of a strict concordance between the presence (or absence) of a specific taxon and a given disease state argues against a possible causal role.

MULTIPLE SCLEROSIS

Epidemiologic studies of twin pairs and at-risk individuals moving between high- and low-risk geographic areas indicate that genetics plays a minor component in multiple sclerosis (MS) susceptibility relative to environmental factors. For example, in monozygotic twin pairs in which one sibling has MS, the other sibling also develops MS in only ~30% of cases. Although MS is a disease of the central nervous system (CNS), there is growing evidence of a link between MS and the microbiota, specifically that of the gut. Germ-free animals and antibiotic-treated animals display reduced disease incidence and severity in an MS model. Similarly, some clinical studies suggest improved disease outcomes in patients with MS who have been treated with minocycline, while patients treated with long-term penicillin appear to have an increased disease risk. Although several studies have compared the fecal microbiotas of healthy controls to those of patients with MS, these studies have all been relatively small and have yielded few results (if any) that are common throughout. Although work relating the microbiome to MS is ongoing, it has opened the door to exploring this link with other neurologic diseases. Already, there are animal data demonstrating links between the microbiota and both Parkinson's disease and autism, and there are clinical data assessing fecal microbiomes in relation to a variety of neurologic conditions. It is not quite clear how the gut microbiota is communicating with the CNS—i.e., whether communication takes place via bacterial metabolites that travel in the bloodstream and cross the blood–brain barrier, via migration of whole organisms into the CNS, or via feedback through the vagus nerve. Although our understanding of this brain–gut axis is still in its infancy, research in this area has elicited tremendous excitement as a tractable approach to potential treatments for these challenging diseases.

Atopic Diseases

The incidence and prevalence of allergic diseases continue to steadily increase, as do more severe clinical presentations. Life-threatening food allergies are now such a public health issue that nut-free classrooms are the norm in many cities. The development of allergic diseases often follows a stereotyped progression that begins with atopic dermatitis (AD) and continues, in order, with food allergy, asthma, and allergic rhinitis. The microbiome has been linked to all of these conditions and has the potential to modulate effects anywhere along this spectrum.

ATOPIC DERMATITIS

The skin is the largest organ in the body, and its different anatomic sites (e.g., antecubital fossa, volar forearm, alar crease) represent distinct ecologic niches and harbor unique microbial communities. Moreover, given that the skin serves as a critical interface between the body and the external environment (e.g., microbes), it must be able to respond to unwanted microbes with an adequate immune response. AD is an inflammatory skin disorder involving immune dysfunction and a dysbiotic skin microbiota that is typically marked by greater abundances of *Staphylococcus aureus* and a lesser degree of bacterial diversity. Effective treatment of AD does not require complete elimination of *S. aureus* but is associated with restoration of the normal level of diversity. It is likely that this increase in bacterial diversity re-establishes normal immune homeostasis in the skin; specific members of the skin microbiota have been shown to induce protective skin-restricted immune responses. Coagulase-negative staphylococci (CoNS; primarily *S. epidermidis* and *S. hominis*) obtained from lesional and nonlesional skin of patients with AD were functionally screened and compared to CoNS from healthy controls; AD-lesional CoNS were much less often able to produce antimicrobial peptides (lantibiotics) directed against *S. aureus*. To demonstrate that these lantibiotic-producing CoNS were biologically relevant, they were incorporated into a lotion and applied to the arms of patients with AD. Surprisingly, a single application of the probiotic-laced lotion led to a decrease in the abundance of *S. aureus* recovered; no such decrease was observed when lantibiotic-negative strains were used. The authors of this study did not specifically comment on the clinical improvement of the AD lesions. Nevertheless, this is one of a limited number of studies that is beginning to extend microbiome-related findings into clinical trials.

ASTHMA

Asthma is characterized by the clinical triad of airflow obstruction, bronchial hyperresponsiveness, and inflammation in the lower respiratory tract. Although the long-standing dogma was that the lungs are sterile, there is now convincing evidence for a constant ebb and flow of bacteria within the lower airways. In healthy states, the mucociliary escalator continually eliminates these bacteria soon after they land in the airways; in disease states (e.g., cystic fibrosis, chronic obstructive pulmonary disease), these bacteria establish long-term colonization of the airways and influence disease pathogenesis. In asthma specifically, both fecal and airway microbes have been linked to clinical outcomes.

Early studies of the microbiome's influence on asthma used culture-based methods to assess the hypopharyngeal microbiota of asymptomatic 1-month-old infants. Intriguingly, in one study, early-life colonization with *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Moraxella catarrhalis*, or a combination of these organisms—but not *S. aureus*—was significantly associated with persistent wheeze and asthma at 5 years of age. Eosinophilia and total IgE levels at 4 years of age were also increased in children who were neonatally colonized with these organisms. Although this study examined a fairly focused set of bacteria, it laid the experimental groundwork indicating that early-life microbial exposures influence subsequent development of asthma. A later longitudinal investigation of the fecal microbiota in a general-population birth cohort of more than 300 children demonstrated that lower abundances of the genera *Lachnospira*, *Veillonella*, *Faecalibacterium*, and *Rothia* at 3 months of age were associated with an increased risk for development of asthma. The fact that these bacterial changes were no longer apparent when the children were 1 year of age is consistent with the notion that microbial exposures early in life are important to disease pathogenesis later in life. Transplantation of stool samples from 3-month-old children at risk for asthma into gnotobiotic mice resulted in significant airway inflammation in a murine model of asthma; pre- and postnatal exposure of mice to a four-species cocktail (*F. prausnitzii*, *Veillonella parvula*, *Rothia mucilaginosa*, and *Lachnospira multipara*) inhibited airway inflammation, with a marked reduction in neutrophil numbers in bronchoalveolar lavage fluid. These data suggest that early-life modulation of the microbiome may

be an effective strategy to help prevent asthma, though the specific logistics (e.g., strains, dose, timing of exposure, patient selection) remain to be clarified.

Infectious Diseases

The increased susceptibility of antibiotic-treated mice to infection with a wide range of enteric pathogens was initially observed in the 1950s and led soon thereafter to the concept of *colonization resistance*, which holds that the normal intestinal microbiota plays a critical role in preventing colonization—and therefore disease production—by invading pathogens. Seminal work in the 1970s demonstrated that this protection is largely reliant on clostridial organisms, and the subsequent half-century has been spent trying to identify the specific microbes involved. Although much of the work relating the microbiota to infection has focused on enteric pathogens, the intestinal microbiota has also been clearly linked to bacterial pneumonia in mouse models, and changes in the microbial composition of the gut have been causally related to changes in the severity of disease. Although this gut–lung axis clearly exists in animals, its relevance in humans is still unclear. Several groups are beginning to study the human lung microbiome in the context of pneumonia and tuberculosis. Moreover, the relationships between the microbiota and both systemic infections (e.g., HIV infection, sepsis) and the response to vaccination are starting to be explored.

ENTERIC INFECTIONS

Clostridium difficile infection (CDI) represents a growing worldwide epidemic and is the leading cause of antibiotic-associated diarrhea **(Chap. 129)**. Roughly 15–30% of patients who are successfully treated for CDI end up with recurrent disease. The strong association between antibiotic exposure and CDI initially raised the idea that the microbiota is inextricably linked to acquisition of disease, presumably because of the loss of colonization resistance. Consistent with the epidemiologic data, characterization of the fecal microbiota of patients with CDI revealed that it is a markedly less diverse, dysbiotic community. Fecal microbiota transplantation (FMT)—the "transplantation" of stool from a healthy individual into patients with disease—was successfully used in the 1950s to treat four patients with severe CDI and has recently been demonstrated in numerous studies to be an effective therapy for recurrent CDI, with clinical cure in 85–90% of patients (as detailed below). Thus, FMT for recurrent CDI has become the "poster child" for the idea that microbiome-based therapies may transform the management of many diseases previously considered to be refractory to medical therapy. Although FMT is agnostic as to the underlying mechanism of protection, work is ongoing to identify specific microbes and host pathways that can protect against CDI. Studying mice with differential susceptibilities to CDI due to antibiotic-induced changes in their microbiota, investigators identified a cocktail of four bacteria (*Clostridium scindens*, *Barnesiella intestihominis*, *Pseudoflavonifractor capillosus*, and *Blautia hansenii*) that conferred protection against CDI in a mouse model. Intriguingly, treatment of mice with just *C. scindens* offered significant, though not complete, protection in a bile acid–dependent manner. Clinical data from patients who underwent HSCT also associated *C. scindens* with protection from CDI, an observation that suggests the possibility of translating these findings from mice to humans. This study provides another example of the identification of relevant bacterial factors through examination of microbial differences in populations that differ in disease risk.

Microbiome-related changes associated with *Vibrio cholerae* infection include a striking loss of diversity (largely due to *V. cholerae*'s becoming the dominant member of the microbiota) and an altered composition that rapidly follows the onset of disease. These changes, which occur in a reproducible and stereotypical manner, are reversible with treatment of the disease. This recovery phase involves a microbial succession that is similar to the assembly and maturation of the microbiota of healthy infants. In addition to *V. cholerae*, streptococcal and fusobacterial species bloom during the early phases of diarrhea, and the relative abundances of *Bacteroides*, *Prevotella*, *Ruminococcus/Blautia*, and *Faecalibacterium* species increase during the resolution phase and mark the return to a healthy adult microbiota. Analysis of these microbial changes occurring in patients with cholera and in healthy children led to the selection of 14 bacteria that were transplanted into gnotobiotic mice, which were then challenged with *V. cholerae*. Bioinformatic analysis of specific taxa changing during cholera determined that *Ruminococcus obeum* restrained *V. cholerae* growth. Subsequently, this relationship was experimentally confirmed, and the *R. obeum* quorum-sensing molecule AI-2 (autoinducer 2) was found to be responsible for restricting *V. cholerae* colonization via an unclear mechanism. These studies highlight the potential for use of microbiome-based therapies to prevent and/or treat infectious diseases. Moreover, they suggest that temporal analysis of longitudinal microbiome data may be an effective strategy for identifying microbes with causal relationships to disease.

HIV INFECTION

The augmentation of HIV pathogenesis by some viral, bacterial, and parasitic co-infections suggests that a patient's underlying microbial environment can influence the severity of HIV disease. Moreover, it has been hypothesized that the intestinal immune system plays a significant role in regulating HIV-induced immune activation; this seems particularly likely since the intestines are an early site for viral replication and exhibit immune defects before peripheral CD4+ T cell counts decrease. Several studies have examined the intestinal microbiotas of HIV-infected individuals. Initial studies performed in nonhuman primates infected with simian immunodeficiency virus found no alteration in the bacterial components of the fecal microbiota; however, there were profound changes in the enteric virome. In contrast, many recent studies exploring this issue in patients have identified substantial differences in the HIV-associated fecal microbiota that correlate with systemic markers of inflammation. Curiously, these microbial changes do not necessarily normalize with antiretroviral therapy; this finding suggests that the microbiota may have some "memory" of the previously high HIV loads and/or that HIV infection helps reset the "normal" microbiota. This memory-like capacity of the microbiota has been demonstrated in animal models in the context of other infections and in response to dieting.

Given that the majority of new HIV transmission events follow heterosexual intercourse, there has been significant interest in examining the relationship between the vaginal microbiota and HIV acquisition. A longitudinal study of South African adolescent girls who underwent high-frequency testing for incident HIV infection facilitated the identification of bacteria that were associated with reduced risk of HIV acquisition (*Lactobacillus* species other than *L. iners*) or with enhanced risk (*Prevotella melaninogenica*, *Prevotella bivia*, *Veillonella montpellierensis*, *Mycoplasma*, and *Sneathia sanguinegens*). In mice inoculated intravaginally with *Lactobacillus crispatus* or *P. bivia*, the latter organism induced a greater number of activated CD4+ T cells in the female genital tract, a result suggesting that the increased risk of HIV acquisition associated with *P. bivia* may be secondary to the increased presence of target cells. In a separate study, the composition of the vaginal microbiota was shown to modulate the antiviral efficacy of a tenofovir gel microbicide. Although tenofovir reduced HIV acquisition by 61% in women who had a *Lactobacillus*-dominant vaginal microbiota, it reduced HIV acquisition by only 18% in women whose vaginal microbiota comprised primarily *Gardnerella vaginalis* and other anaerobes. This difference in efficacy was due to the ability of *G. vaginalis* to metabolize tenofovir faster than the target cells can take up the drug and convert it into its active form, tenofovir diphosphate. These findings illustrate how microbial ecology can be an important consideration in choosing effective treatment regimens.

RESPONSE TO VACCINATION

Second only to the provision of clean water, vaccination has been the most effective public health intervention in the prevention of serious infectious diseases. Its effects are mediated by antigen-specific antibodies and, in some cases, effector T-cell responses. Although vaccines are clearly effective on a population scale, the magnitude of the immune response to vaccines can vary among individuals by tenfold to a hundredfold. Although many factors (e.g., genetics, maternal antibody levels, prior antigen exposures) can affect vaccine immunogenicity, the microbiota is now recognized as another important factor. Analysis of the fecal microbiotas of ~50 Bangladeshi children identified specific taxa that exhibited positive associations (e.g., *Actinomyces*, *Rothia*, and *Bifidobacterium* species) and negative associations (e.g., *Acinetobacter*, *Prevotella*, and staphylococcal species) with responses to vaccines against polio, tuberculosis (bacille Calmette-Guérin), tetanus, and hepatitis B. A study of infants from Ghana revealed an inverse relationship between the fecal abundance of Bacteroidetes and a response to the rotavirus vaccine. Moreover, the nasal microbiota has been implicated as a factor that contributes to the IgA response to live, attenuated influenza vaccines. These correlations based on clinical data have been partially confirmed in animal studies. The best example is the demonstration that the responses to non-adjuvanted viral subunit vaccines (inactivated influenza and polio vaccines) are reliant on the microbiota, whereas the responses to live or adjuvanted vaccines (live attenuated yellow fever, Tdap/alum, an HIV envelope protein/alum vaccine) are not. An interesting note is that the antibody response to inactivated influenza vaccine is dependent on recognition of the microbiota by Toll-like receptor 5, presumably via flagellin-expressing microbes. These data suggest that the microbiota may serve as an adjuvant for certain vaccine types. Confirmation of these findings in clinical settings may suggest ways to improve vaccine efficacy in the future.

MECHANISMS OF MICROBIOME-MEDIATED EFFECTS

As highlighted in the examples above, numerous associations have been made between the microbiome and various disease states. These correlations have often been established at broad taxonomic levels, with little or no insight into causality. Given that most clinical studies of these relationships have a fairly small sample size (often <100) and are simultaneously comparing numerous variables (i.e., each of the bacterial species in the microbiota is effectively a different feature being compared), many of these studies may not be adequately powered and therefore may yield false-positive results. Testing of these correlations in animal models of disease has been critical in demonstrating a causal relationship between microbes and specific phenotypes. Because microbiome-wide association studies typically result in a long list of bacterial taxa that are correlated with a disease, it has been challenging to know which organism to test further in mechanistic studies. Moreover, even if a specific bacterial species is identified in these analyses, there is potentially enough strain-to-strain variation that the “functional” isolate may need to be recovered from the individuals studied; a publicly available representative of the species may not confer the same phenotype.

Despite all these difficulties, a handful of specific microbes have now been linked to disease effects; some examples have been mentioned above. The next layer of challenges relates to identification of the specific mechanisms that underlie these causal relationships. Successes along these lines have been more limited, but approaches are being developed to tackle the issue of defining specific bacterial factors and metabolites that are responsible for the phenotypic changes. Complicating factors are that many organisms, particularly those in the phylum Firmicutes, are not readily genetically tractable and that many of the phenotypes are not easy to assess with high-throughput screening.

BACTERIAL FACTORS

B. fragilis polysaccharide A (PSA) is perhaps the best-studied commensal-derived molecule that has been demonstrated to influence disease outcomes in mouse models. PSA—one of at least eight capsular polysaccharides expressed by *B. fragilis*—has a unique zwitterionic structure that incorporates both a positive and a negative charge within each repeating unit. Studies in which mice have been treated either with isogenic strains of *B. fragilis* that differ in PSA expression or with purified PSA have shown that PSA confers protection—prophylactically and therapeutically—against experimental colitis and MS. PSA is recognized by Toll-like receptor 2 on antigen-presenting cells, particularly plasmacytoid dendritic cells, and—in the setting of inflammation—induces interleukin 10 (IL-10)–producing regulatory T cells (Tregs) that help restrain inflammation.

B. fragilis is also the source of the only other microbiota-based bacterial factor identified thus far: an immunomodulatory glycosphingolipid that affects the numbers of invariant natural killer T (iNKT) cells. It is not clear whether these glycosphingolipids activate or inhibit iNKT cells; results have been

discordant, probably because different glycosphingolipid species have been tested. Analysis of a specific purified glycosphingolipid (Bf717) demonstrated that it inhibits endogenous iNKT cell agonists in vitro and in vivo. Treatment of neonatal mice with Bf717 leads to a decreased number of colonic iNKT cells in adulthood and to improved outcomes in a model of colitis.

BACTERIAL METABOLITES

The use of mass spectrometry to detect and profile tens of thousands of different metabolites present in different bodily fluids has offered the promise of deeper insight into microbially mediated processes that underlie disease susceptibility. However, the fact that the overwhelming majority of these metabolites are not annotated, coupled with the sheer volume of data generated, has so far limited the general utility of these untargeted approaches. Instead, interest in more targeted approaches has increased, with a current emphasis on examining the role of short-chain fatty acids (SCFAs) and bile acids.

Short-Chain Fatty Acids

Several groups have demonstrated that SCFAs, the intestinal levels of which are largely determined by bacterial metabolism, are important for the induction of Tregs, though there is not agreement on which specific SCFA (propionate, acetate, or butyrate) is most relevant. Wild-type mice colonized with bacteria known to induce colonic Tregs have elevated cecal levels of SCFAs. Colonization with any of three *Bacteroides* species (*B. caccae*, *B. massiliensis*, and *B. thetaiotaomicron*) increases levels of acetate and propionate, whereas colonization with *Parabacteroides distasonis* or a mix of 17 human-derived *Clostridium* species elevates levels of all three SCFAs. In all of these cases, though, the SCFAs inhibit histone deacetylase, with a consequent increase in Foxp3 expression. Notably, microbe-induced SCFA production has not been shown to be critical for Treg induction by any of these organisms. In contrast, there appears to be no correlation between SCFA levels and Treg numbers in mice monocolonized with various Treg-inducing bacterial species. Taken together, these data suggest important heterogeneity in the mechanisms underlying Treg development and do not rule out the possibility of other, redundant mechanisms for Treg induction.

Bile Acids

Bile acids are produced in the liver but then are metabolized by intestinal bacteria to form deconjugated and secondary bile acids. These microbially produced bile acid profiles act through complex signaling pathways to balance the metabolism of lipids and carbohydrates and to affect immune responses. Therefore, bile acids are now being investigated as microbial metabolites that are critical to maintaining human health. As mentioned above, *C. scindens* helps protect mice against CDI through a bile acid–dependent process. Alterations in bile acid profiles due to underlying microbial dysbiosis have also been associated with hepatic and colonic inflammation, hepatic cellular carcinoma, colorectal cancer, and impaired gut motility. Almost all of these relationships have been documented at the level of correlation and, at best, reflect a partial change in phenotype in the setting of bile acid sequestrants (e.g., cholestyramine). Work is ongoing to determine causal relationships between bacterial metabolism of bile acids and changes in host physiology.

Other Bacterial Metabolites

Although most work has thus far focused on SCFAs and bile acids, a few notable examples of other bacterial metabolites have been implicated in maintaining health. Taurine enhances NLRP6 inflammasome–induced colonic IL-18 secretion, while histamine, spermine, and putrescine suppress IL-18 secretion. Desaminotyrosine produced by *Clostridium orbiscindens* confers protection from influenza by inducing type I interferon activity. In both of these cases, the microbiota was initially shown to influence the phenotype, with either untargeted metabolomics or a more targeted screen indicating a potential role for the indicated metabolites. Given the thousands of different bacterial metabolites throughout the body, many more metabolites will undoubtedly be linked to health and disease.

MOVING MICROBIOME SCIENCE FROM BENCH TO BEDSIDE

The numerous microbiome–disease associations identified thus far have generated a great deal of hope that understanding the relevant microbe–host interactions will open the door to unlimited therapeutic applications. Microbiome-based therapies offer several potential benefits. Patients often view such treatment as more “natural” than conventional drug therapy and are therefore more likely to comply with it. Biologically, microbiome-based therapies are more likely to address one of the root causes of disease (microbial dysbiosis) rather than simply affecting the downstream sequelae. Finally, a given microbiome-based therapy may serve as a “polypill” that is effective against several different diseases stemming from similar microbial changes. Despite tremendous interest in therapeutically exploiting the microbiome, there have thus far been few clinical successes along these lines.

The most successful therapeutic application of microbiome science has been the use of FMT, particularly for CDI. As mentioned earlier, FMT involves “transplanting” stool from a healthy individual to a diseased patient, with the idea that the “healthy” microbiota will correct whatever derangement may exist in the ill patient and therefore will alleviate symptoms. Fundamentally, this notion is agnostic as to the specific microbial dysbiosis and holds that any healthy microbiota will be curative. The idea of FMT dates back to at least the fourth century, when traditional Chinese doctors used a “yellow soup” (fresh human fecal suspension) to successfully treat food poisoning and severe diarrhea. The continued use of FMT through the centuries for the

treatment of diarrheal illnesses in both humans and animals, along with the growing appreciation in recent years of the importance of the microbiota, laid the groundwork for using FMT to treat CDI. Since the first major prospective trial assessing FMT for recurrent CDI in 2013, most of the numerous studies of FMT for CDI have demonstrated remarkable efficacy, with an average clinical cure rate of ~85%. The donor stool can be fresh or frozen (use of the latter allows biobanking of samples from a limited number of pre-screened donors) and can be administered via nasogastric tube, nasoduodenal tube, colonoscopy, enema, or oral capsules; the cure rate is slightly higher with lower-gastrointestinal administration than with upper-gastrointestinal treatment. The optimal screening, preparation, and concentration of infused donor stool have not yet been determined. The most common adverse effects of FMT include altered gastrointestinal motility (with constipation or diarrhea), abdominal cramps, and bloating, all of which are generally transient and resolve within 48 h. At least 80 immunosuppressed patients have undergone FMT with no serious adverse events noted during 3 months of follow-up.

The successful use and the favorable short-term safety profile of FMT for CDI have led to its expanded application for other indications. At the end of 2017, 190 trials (listed at ClinicalTrials.gov) were investigating the efficacy of FMT for a range of indications, including CDI, IBD (ulcerative colitis and Crohn's disease), obesity, eradication of multidrug-resistant organisms, anxiety and depression, cirrhosis, and type 2 diabetes. The few published studies regarding indications other than CDI have generally included small sample sizes and have offered mixed results. In contrast to the successes in CDI, the results have been more varied for patients with IBD, which is perhaps the second-best-studied indication. It is not clear whether these discrepancies are due to heterogeneity in recipients (e.g., in terms of underlying disease mechanisms or endogenous microbiotas), the donor material, and/or the logistical details of FMT administration (e.g., route, frequency, dose). However, these results demonstrate that—under the right circumstances—modulation of the microbiota can be an effective therapy for IBD.

Although FMT offers an important proof of concept that microbiome-based therapies can be effective, treatment is difficult to standardize across large populations because of variability among stool donors and among the endogenous microbiotas of recipients. In addition, FMT is fraught with safety concerns, and its mechanism(s) of action are unclear. FMT likely represents the first generation of microbiome-based therapies; subsequent generations will include the use of more refined bacterial cocktails, single strains of bacteria, or bacterial metabolites as the therapeutic intervention. The field of probiotics has a complicated history: many different strains have been tested against a multitude of diseases. Several meta-analyses have combined results across bacterial strains and/or disease indications and have generally concluded that the data are not yet convincing enough to support the use of the tested regimens. It should be noted that the tested organisms have been chosen mainly on the basis of their presumed safety profile rather than in light of a plausible biological link to disease. The hope is that more focused, mechanistic microbiome studies will identify specific commensal organisms—and their underlying mechanisms of action—that are involved in disease pathogenesis and that will serve as the basis for the next wave of rationally chosen probiotics, a few of which are currently in clinical trials. The main hurdle in this endeavor has been identifying specific microbes that are causally related to protection from disease.

PERSPECTIVE

The medical view of microbes has changed radically, moving from the early-twentieth-century notion that we are engaged in a constant struggle with microbes—an "us-versus-them" mentality that focused on the necessity of eradicating bacteria—to the more recent understanding that we live in a carefully negotiated state of détente with our commensal organisms. Instead of holding a simple view of microbes as enemies to be eliminated with antibiotics, scientists are increasingly recognizing the critical role these organisms play in maintaining human health; loss of these host–microbe interactions in the increasingly sterile environment typical of Western civilization may have predisposed to the increased incidence of autoimmune and inflammatory diseases. The field of microbiome research has made great strides over the past decade in cataloguing the normal microbiota and is now on the cusp of being able to identify clinically actionable microbe–host relationships.

The recent explosion of "–omics" technologies (e.g., metagenomics, metatranscriptomics, metabolomics) has enabled the generation of vast amounts of data, but it is not yet clear how best to integrate datasets in order to gain useful insights into host–microbe relationships. The use of FMT has demonstrated that modulation of an individual's microbiota can effectively treat certain diseases; however, models with which to predict specifically how a microbiota will change after modulation—and what potentially untoward effects these changes might have—are still lacking. Implicit in this limitation is our ignorance about what microbial configuration is optimal and how a given microbiota should be rationally altered to obtain an ideal outcome.

Despite initial hyperbolic hype and a few false starts, microbiome research now stands at the precipice of an ability to treat the fundamental basis of many diseases. As the field continues to mature, it will need to move beyond correlations and address causation. The identification of causal microbes and their mechanisms of action will create a "microbial toolbox" from which relevant bioactive strains can be chosen on a per-patient basis to correct specific underlying microbial dysbioses. In the near future, our knowledge base regarding the microbiome and its relationship to health and disease will be robust enough that this information can be applied in making important treatment decisions.

FURTHER READING

Goodrich JK et al: Conducting a microbiome study. Cell 158:250, 2014. [PubMed: 25036628]

Human Microbiome Project Consortium: Structure, function and diversity of the healthy human microbiome. Nature 486:207, 2012. [PubMed: 22699609]

Lloyd-Price J et al: Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550:61, 2017. (Erratum: Nature 551:256, 2017.) [PubMed: 28953883]

Rajilić-Stojanović M, de Vos WM: The first 1000 cultured species of the human gastrointestinal microbiota. FEMS Microbiol Rev 38:996, 2014. [PubMed: 24861948]

Surana NK, Kasper DL: Deciphering the tête-à-tête between the microbiota and the immune system. J Clin Invest 124:4197, 2014. [PubMed: 25036709]

Chapter 460: Global Issues in Medicine

Joseph J. Rhatigan; Paul Farmer

WHY GLOBAL HEALTH?

Global health has emerged as an important field within medicine. Some scholars have defined global health as the field of study and practice concerned with improving the health of all people and achieving health equity worldwide, with an emphasis on addressing transnational problems. No single review can do much more than identify the leading problems in applying evidence-based medicine in settings of great poverty or across national boundaries. However, this is a moment of opportunity: only recently, persistent epidemics, improved metrics, and growing interest have been matched by an unprecedented investment in addressing the health problems of poor people in the developing world. To ensure that this opportunity is not wasted, the facts need to be laid out for specialists and laypeople alike. This chapter introduces the major international bodies that address health problems; identifies the more significant barriers to improving the health of people who to date have not, by and large, had access to modern medicine; and summarizes population-based data on the most common health problems faced by people living in poverty. Examining specific problems—notably HIV/AIDS **(Chap. 197)** but also tuberculosis (Chap. 173), malaria **(Chap. 219)**, Ebola (Chap. 205), and key "noncommunicable" chronic diseases (NCDs)—helps sharpen the discussion of barriers to prevention, diagnosis, and care as well as the means of overcoming them. This chapter closes by discussing global health equity, drawing on notions of social justice that once were central to international public health but had fallen out of favor during the last decades of the twentieth century.

A BRIEF HISTORY OF GLOBAL HEALTH INSTITUTIONS

Concern about illness across national boundaries dates back many centuries, predating the Black Plague and other pandemics. One of the first organizations founded explicitly to tackle cross-border health issues was the Pan American Sanitary Bureau, which was formed in 1902 by 11 countries in the Americas. The primary goal of what later became the Pan American Health Organization was the control of infectious diseases across the Americas. Of special concern was yellow fever, which had been running a deadly course through much of South and Central America and halted the construction of the Panama Canal. In 1948, the United Nations formed the first truly global health institution: the World Health Organization (WHO). In 1958, under the aegis of the WHO and in line with a long-standing focus on communicable diseases that cross borders, leaders in global health initiated the effort that led to what some see as the greatest success in international health: the eradication of smallpox. Naysayers were surprised when the smallpox eradication campaign, which engaged public health officials throughout the world, proved successful in 1979 despite Cold War tensions.

At the International Conference on Primary Health Care in Alma-Ata (in what is now Kazakhstan) in 1978, public health officials from around the world agreed on a commitment to "Health for All by the Year 2000," a goal to be achieved by providing universal access to primary health care worldwide. Critics argued that the attainment of this goal by the proposed date was impossible. In the ensuing years, a strategy for the provision of selective primary health care emerged. This strategy included four inexpensive interventions collectively known as GOBI: *g*rowth monitoring, *o*ral rehydration, *b*reast-feeding, and *i*mmunizations for diphtheria, whooping cough, tetanus, polio, tuberculosis, and measles. GOBI later was expanded to GOBI-FFF, which also included *f*emale education, *f*ood, and *f*amily planning. Some public health figures saw GOBI-FFF as an interim strategy to achieve "health for all," but others criticized it as a retreat from the bolder commitments of Alma-Ata.

The influence of the WHO waned during the 1980s. In the early 1990s, many observers argued that, with its vastly superior financial resources and its close—if unequal—relationships with the governments of poor countries, the World Bank had eclipsed the WHO as the most important multilateral institution working in the area of health. One of the stated goals of the World Bank was to help poor countries identify "cost-effective" interventions worthy of public funding and international support. At the same time, international financial institutions encouraged many of those nations to reduce public expenditures in health and education in order to stimulate economic growth as part of (later discredited) policies imposing restrictions as a condition for access to credit and assistance through the World Bank, the International Monetary Fund, and regional development banks. There was a resurgence of many diseases—including malaria, trypanosomiasis, and schistosomiasis—in Africa. Tuberculosis, an eminently curable disease, remained the world's leading infectious killer of adults. Half a million women per year died in childbirth during the last decade of the twentieth century, and few of the world's largest philanthropic or funding institutions focused on global health equity.

HIV/AIDS, first described in the medical literature in 1981, precipitated a change. In the United States, the advent of this newly described infectious killer marked the culmination of a series of events that dashed previous hopes of "closing the book" on infectious diseases. In Africa, which would emerge as the global epicenter of the pandemic, HIV disease strained tuberculosis control programs, and malaria continued to claim as many lives as ever: at the dawn of the twenty-first century, these three diseases alone killed nearly 6 million people each year. New research, new policies, and new funding mechanisms were called for. The past decade has seen the rise of important multilateral global health financing institutions such as the Global Fund to Fight AIDS, Tuberculosis, and Malaria; bilateral efforts such as the U.S. President's Emergency Plan for AIDS Relief (PEPFAR); and private philanthropic organizations such as the Bill & Melinda Gates Foundation. With its 193 member states and 147 country offices, the WHO remains important in matters relating to the cross-border spread of infectious diseases and other health threats. In the aftermath of the epidemic of severe acute respiratory syndrome in 2003, the WHO's International Health Regulations—which provide a legal foundation for that organization's direct investigation into a wide range of global health problems, including pandemic influenza, in any member state—were strengthened and brought into force in May 2007.

Even as attention to and resources for health problems in poor countries grow, the lack of coherence in and among global health institutions may undermine efforts to forge a more comprehensive and effective response. The WHO remains underfunded despite the ever-growing need to engage a wider and more complex range of health issues, such as the Ebola outbreak of 2014–2015 in West Africa. This may be what some have called "the golden age of global health," but leaders of major global health organizations must work together to design an effective architecture that will make the most of opportunities to link new resources for and commitments to global health equity with the emerging understanding of disease burden and the unmet need to create robust and resilient national health systems. To this end, new and old players in global health must invest heavily in *discovery* (relevant basic science), *development* of new tools (preventive, diagnostic, and therapeutic), and modes of *delivery* that will ensure the equitable provision of health products and services to all who need them.

The adoption of the Sustainable Development Goals (SDGs) in 2015 by the United Nations serves as an example of effective cooperation. The SDGs articulate 17 overarching goals across a number of domains to be achieved by 2030. Goal 3 specifically relates to global health and contains 13 distinct targets to be met, including reducing maternal and child mortality; ending the epidemics of HIV, tuberculosis, and malaria; and reducing the burden of NCDs.

Included in the SDGs is a commitment to achieve universal health coverage (UHC), providing universal access to high-quality essential health services at an affordable cost worldwide. Championed by the WHO, the World Bank, and many civil society organizations, Goal 3 will measure coverage of 16 essential health services and assess the financial burden of health spending by households in every country.

THE ECONOMICS OF GLOBAL HEALTH

Political and economic concerns have often guided global health interventions. As mentioned, early efforts to control yellow fever were tied to the completion of the Panama Canal. However, the precise nature of the link between economics and health remains a matter for debate. Some economists and demographers argue that improving the health status of populations must begin with economic development; others maintain that addressing ill health is the starting point for development in poor countries. In either case, there is increasing consensus that investments in health care delivery and the control of communicable diseases lead to increased productivity. The question is where to find the necessary resources to start the predicted "virtuous cycle."

During the past two decades, spending on health in poor countries has increased dramatically. According to a study from the Institute for Health Metrics and Evaluation (IHME) at the University of Washington, total development assistance for health worldwide grew to $36.4 billion in 2015—up from $5.6 billion in 1990. In 2015, the leading contributors included PEPFAR, the Global Fund, the GAVI Alliance, nongovernmental organizations (NGOs), the WHO, the Gates Foundation, and the World Bank, which have dramatically increased their development spending for health since 2014. It appears, however, that total development assistance for health plateaued in 2010, and it is unclear whether growth will continue in the future.

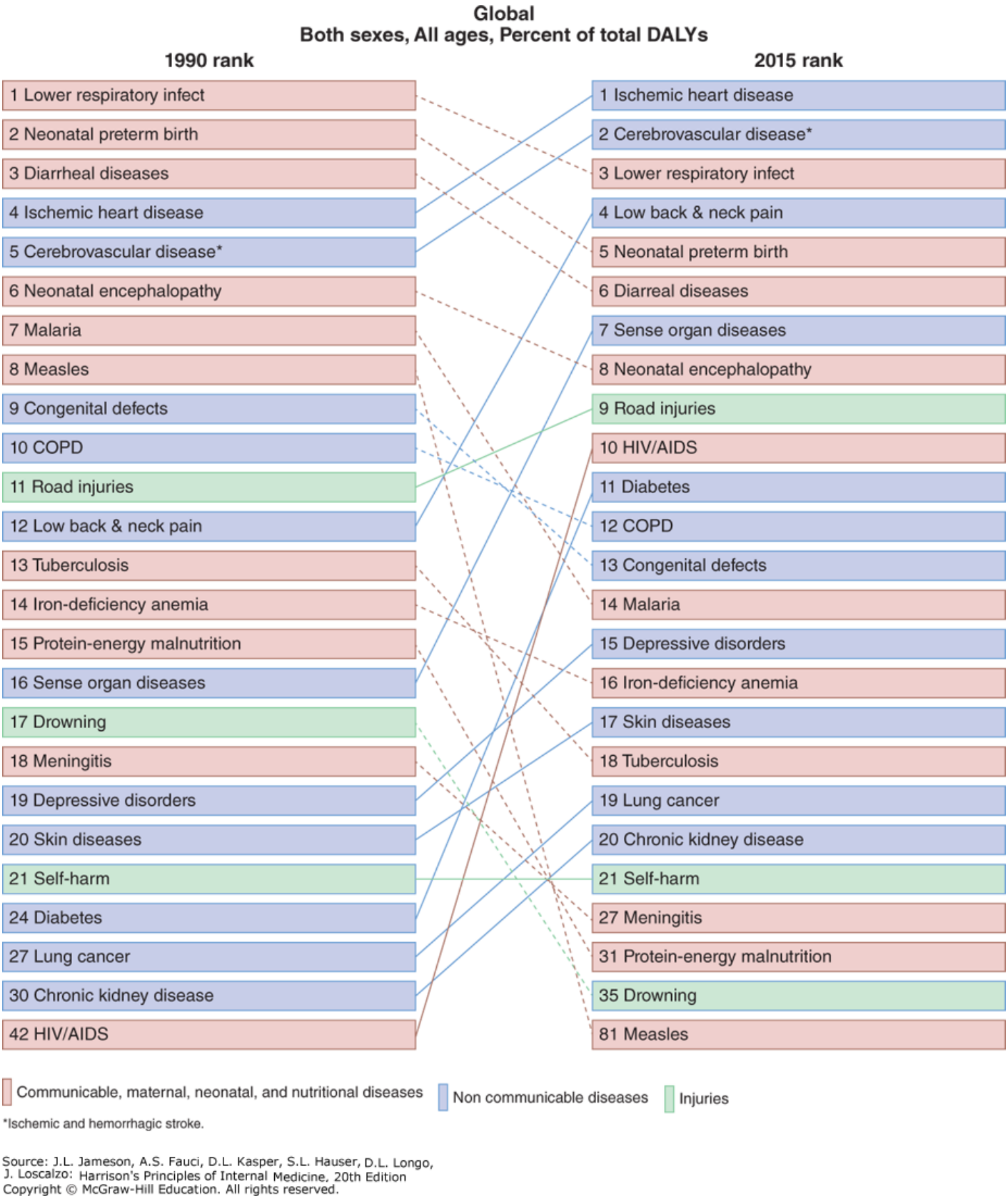
MORTALITY AND THE GLOBAL BURDEN OF DISEASE

Refining metrics is an important task for global health: only relatively recently have there been solid assessments of the global burden of disease. The first study to look seriously at this issue, conducted in 1990, laid the foundation for the first report on *Disease Control Priorities in Developing Countries* and for the World Bank's 1993 World Development Report *Investing in Health*. Those efforts represented a major advance in the understanding of health status in developing countries. *Investing in Health* has been especially influential: it familiarized a broad audience with cost-effectiveness analysis for specific health interventions and with the notion of disability-adjusted life years (DALYs). The DALY, which has become a standard measure of the impact of a specific health condition on a population, combines absolute years of life lost and

years lost due to disability for incident cases of a condition. (See Fig. 460-1 and Table 460-1 for an analysis of the global disease burden by DALYs.)

FIGURE 460-1

Global DALY (disability-adjusted life year) ranks for the top causes of disease burden in 1990 and 2015. COPD, chronic obstructive pulmonary disease. *(From the Institute for Health Metrics and Evaluation [IHME]. GBD Compare. Seattle, WA: IHME, University of Washington, 2016. Available at http://vizhub.healthdata.org/gbd-compare. Accessed July 18, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 460-1

Leading Causes of Burden of Disease (DALYs), 2015

Disease or Injury	DALYs (Millions)	Percentage of Total DALYs
World	2464.9	100
1. Ischemic heart disease	164.0	6.7
2. Lower respiratory infections	103.0	4.3
3. Preterm birth complications	74.8	3.0
4. Hemorrhagic stroke	73.4	3.0
5. Diarrheal diseases	71.6	2.9
6. Neonatal encephalopathy	67.9	2.8
7. Diabetes mellitus	64.1	2.6
8. COPD	63.9	2.6
9. Low back pain	60.1	2.4
10. Malaria	55.8	2.3
Low- and Middle-Income Countries[a]		
1. Ischemic heart disease	138.2	6.4
2. Cerebrovascular disease	106.6	5.0
3. Lower respiratory infection	97.2	4.5
4. Preterm birth complications	72.7	3.4
5. Diarrheal disease	71.2	3.3
6. Hemorrhagic stroke	67.5	3.1
7. Neonatal encephalopathy	67.2	3.1
8. HIV/AIDS	65.8	3.1
9. Road injury	60.6	2.8
10. COPD	56.0	2.6
High-Income Countries[a]		
1. Ischemic heart disease	25.6	8.2
2. Low back pain	14.0	4.4

Disease or Injury	DALYs (Millions)	Percentage of Total DALYs
3. Lung cancer	11.5	3.7
4. Diabetes mellitus	10.8	3.4
5. Alzheimer's disease	9.5	3.0
6. Neck pain	8.3	2.6
7. Major depression	8.2	2.6
8. Hearing loss	8.0	2.6
9. COPD	7.8	2.6
10. Self-harm	6.9	2.2
Sub-Saharan Africa		
1. Malaria	49.8	9.4
2. HIV/AIDS	41.1	7.8
3. Lower respiratory infections	41.0	7.8
4. Diarrheal diseases	37.0	7.00
5. Neonatal sepsis	19.9	3.8
6. Neonatal encephalopathy	19.5	3.7
7. Preterm birth complications	17.6	3.3
8. Protein-energy malnutrition	14.7	2.8
9. Iron-deficiency anemia	11.2	2.1
10. Tuberculosis	10.9	2.1

[a] The World Bank classifies high-income countries as those whose gross national income per capita is ≥$12,476. Low- and middle-income countries are categorized as low-income (GNI per capita, <$1025), lower-middle income (GNI per capita, $1026–$4035), and upper-middle income (GNI per capita, $4036–$12,475) (*http://data.worldbank.org.ezproxy1.library.usyd.edu.au/about/country-classifications*).

Abbreviations: COPD, chronic obstructive pulmonary disease; DALYs, disability-adjusted life years.

Source: Institute for Health Metrics and Evaluation, University of Washington (2017). Data available at *http://www.healthmetricsandevaluation.org/gbd/visualizations/country.* Accessed July 18, 2017.

In 2012, the IHME and partner institutions began publishing results from the Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010). GBD 2010 is the most comprehensive effort to date to produce longitudinal, globally ambitious, and comparable estimates of the burden of diseases, injuries, and risk factors. This report reflects the expansion of the available data on health in the poorest countries and of the capacity to quantify the impact of specific conditions on a population. It measures current levels and recent trends for major diseases, injuries, and risk factors worldwide. The GBD 2010 team revised and improved the health-state severity weight system, collated published data, and used household surveys to enhance the breadth and accuracy of disease burden data. Updated reports were released in 2013 and 2015. As

analytic methods and data quality improve, important trends can be identified in a comparison of global disease burden estimates from 1990 to 2015.

GLOBAL MORTALITY

Of the 55.8 million deaths worldwide in 2015, 20.2% (11.3 million) were due to communicable diseases, maternal and perinatal conditions, and nutritional deficiencies—a marked decrease compared with figures for 1990, when these conditions accounted for 34% of global mortality. Among the fraction of all deaths related to communicable diseases, maternal and perinatal conditions, and nutritional deficiencies, 74.2% occurred in sub-Saharan Africa and southern Asia. While the proportion of deaths due to these conditions has decreased significantly in the past decade, there has been a dramatic rise in the number of deaths from NCDs, which constituted the top four causes of death in 2015. The leading cause of death among adults in 2015 was ischemic heart disease, accounting for 8.9 million deaths (16% of total deaths) worldwide. In high-income countries ischemic heart disease accounted for 18.0% of total deaths, and in low- and middle-income countries it accounted for 15.5%. It is noteworthy that ischemic heart disease was responsible for just 4.7% of total deaths in sub-Saharan Africa **(Table 460-2)**. In second place—causing 11.3% of global mortality—was cerebrovascular disease (ischemic and hemorrhagic stroke), which accounted for 9.0% of deaths in high-income countries, 11.9% in low- and middle-income countries, and 5.0% in sub-Saharan Africa. Although the third leading cause of death in high-income countries was lung cancer (accounting for 5.8% of all deaths), this condition did not figure among the top 25 causes in sub-Saharan Africa. Among the 10 leading causes of death in sub-Saharan Africa, six were infectious diseases, with malaria and HIV/AIDS ranking as the dominant contributors to disease burden. In high-income countries, however, only one infectious disease—lower respiratory infection—ranked among the top 10 causes of death.

TABLE 460-2

Leading Causes of Death Worldwide, 2015

Disease or Injury	Deaths (Millions)	Percentage of Total Deaths
World	**55.8**	**100**
1. Ischemic heart disease	8.9	16.0
2. Hemorrhagic stroke	3.3	6.0
3. COPD	3.2	5.7
4. Ischemic stroke	3.0	5.3
5. Lower respiratory infections	2.7	4.9
6. Alzheimer's disease	1.9	3.4
7. Lung cancer	1.7	3.1
8. Diabetes	1.5	2.7
9. Diarrheal diseases	1.3	2.4
10. Tuberculosis	1.1	2.0
Low- and Middle-Income Countries[a]		
1. Ischemic heart disease	7.0	15.5
2. Cerebrovascular disease	5.3	11.9
3. COPD	2.7	6.0
4. Lower respiratory infections	2.2	4.9
5. Diarrheal diseases	1.3	2.9
6. Diabetes	1.3	2.8
7. Road injury	1.2	2.7
8. HIV/AIDS	1.2	2.6
9. Tuberculosis	1.1	2.4
10. Lung cancer	1.1	2.4
High-Income Countries[a]		
1. Ischemic heart disease	1.9	18.0
2. Alzheimer's disease	0.9	8.6

Disease or Injury	Deaths (Millions)	Percentage of Total Deaths
3. Lung cancer	0.6	5.8
4. Ischemic stroke	0.6	5.4
5. Lower respiratory infection	0.5	4.9
6. COPD	0.5	4.5
7. Hemorrhagic stroke	0.4	3.6
8. Colon and rectum cancers	0.4	3.3
9. Diabetes mellitus	0.3	2.3
10. Breast cancer	0.2	1.8
Sub-Saharan Africa		
1. Lower respiratory infections	0.7	9.1
2. HIV/AIDS	0.7	8.8
3. Malaria	0.6	7.1
4. Diarrheal diseases	0.6	7.1
5. Ischemic heart disease	0.4	4.7
6. Tuberculosis	0.3	3.3
7. Hemorrhagic stroke	0.2	3.0
8. Neonatal sepsis	0.2	2.9
9. Neonatal encephalopathy	0.2	2.7
10. Preterm birth complications	0.2	2.5

[a]
The World Bank classifies high-income countries as those whose gross national income per capita is ≥$12,476. Low- and middle-income countries are categorized as low-income (GNI per capita, <$1025), lower-middle income (GNI per capita, $1026–$4035), and upper-middle income (GNI per capita, $4036–$12,475) (*http://data.worldbank.org.ezproxy1.library.usyd.edu.au/about/country-classifications*).

Abbreviation: COPD, chronic obstructive pulmonary disease.

Source: Institute for Health Metrics and Evaluation, University of Washington (2017). Data available at *http://www.healthmetricsandevaluation.org/gbd/visualizations/country*. Accessed July 18, 2017.

The GBD 2015 found that the worldwide mortality figure among children <5 years of age dropped from 16.39 million in 1970 to 12.1 million in 1990 and to 5.8 million in 2015—a decrease that surpassed predictions. Of childhood deaths in 2015, 2.6 million (45.0%) occurred in the neonatal period. About one-third of deaths among children <5 years old occurred in southern Asia and almost one-half in sub-Saharan Africa; ~1% occurred in high-income countries.

The global burden of death due to HIV/AIDS and malaria was on an upward slope until 2004; significant improvements have since been documented. Global deaths from AIDS fell from 1.7 million in 2006 to 1.0 million in 2016, while malaria deaths dropped from 1.2 million to 730,000 over the same period. Despite these improvements, malaria and HIV/AIDS continue to be major burdens in particular regions, with global implications. Although it has only a minor impact on mortality outside sub-Saharan Africa and Southeast Asia, malaria is the fifth leading cause of death of children <5 years of age worldwide. HIV infection ranked 42nd in global DALYs in 1990 but was the 10th leading cause of disease burden in 2015, with sub-Saharan Africa bearing the vast majority of this burden (Fig. 460-1).

The world's population is living longer: global life expectancy has increased significantly over the past 45 years from 58.8 years in 1970 to 71.8 years in 2015. This demographic change, accompanied by the fact that the prevalence of NCDs increases with age, is dramatically shifting the burden of disease toward NCDs, which have surpassed communicable, maternal, nutritional, and neonatal causes. By 2015, 71.3% of total deaths at all ages and 59.7% of all DALYs were due to NCDs. Increasingly, the global burden of disease comprises conditions and injuries that cause disability rather than death.

Worldwide, although both life expectancy and years of life lived in good health have risen, years of life lived with disability have also increased. Despite the higher prevalence of diseases common in older populations (e.g., dementia and musculoskeletal disease) in developed and high-income countries, best estimates from 2015 reveal that disability resulting from cardiovascular diseases, chronic respiratory diseases, and the long-term impact of communicable diseases was greater in low- and middle-income countries. In most developing countries, people lived shorter lives and experienced disability and poor health for a greater proportion of their lives. Indeed, 47.7% of the global burden of disease occurred in southern Asia and sub-Saharan Africa, which together account for only 37.4% of the world's population.

HEALTH AND WEALTH

Clear disparities in burden of disease (both communicable and noncommunicable) across country income levels are strong indicators that poverty and health are inherently linked. Poverty remains one of the most important root causes of poor health worldwide, and the global burden of poverty continues to be high. Among the 7.3 billion people alive in 2015, 10.7% (767 million) lived on <$1.90 (U.S.) per day—one standard measurement of extreme poverty—and another 1 billion lived on <$3.10 per day. Approximately 385 million children—19.5% of all children in low-income countries—lived in extreme poverty in 2013. Comparison of national health indicators with gross domestic product per capita among nations shows a clear relationship between higher gross domestic product and better health, with only a few outliers. Numerous studies have also documented the link between poverty and health within nations as well as across them.

RISK FACTORS FOR DISEASE BURDEN

The GBD study found that the three leading risk factors for global disease burden in 2015 were (in order of frequency) high blood pressure, tobacco smoking (including secondhand smoke), and high fasting plasma glucose—a substantial change from 1990, when childhood undernutrition was ranked first. Although ranking fifth in 2015, childhood undernutrition remains the leading risk factor for death worldwide among children <5 years of age. In an era that has seen obesity become a major health concern in many developed countries—and the fourth leading risk factor for disease burden worldwide—the persistence of undernutrition is cause for consternation. Low body weight is still the dominant risk factor for disease burden in sub-Saharan Africa. In its rural reaches, no health care initiative, however generously funded, will be effective or comprehensive without addressing undernutrition.

In a 2016 publication that examined how specific diseases and injuries are affected by environmental risk, the WHO estimated that 23% of all deaths and 26% of deaths among children <5 years of age were due to modifiable environmental factors: some 1.7 million children die every year from causes related to unhealthy environments, including the more than 360,000 deaths stemming from a lack of access to clean water and sanitation. Many of these modifiable factors lead to child and adult deaths from infectious pathologies; others lead to deaths from malignancies. Etiology and nosology are increasingly difficult to parse with regard to environmental harm. Risk factors such as indoor air pollution due to use of solid fuels, exposure to secondhand tobacco smoke, and outdoor air pollution account for 35% of lower respiratory infections globally. Various forms of unintentional injury and malaria top the list of health problems to which environmental factors contribute.

The third edition of *Disease Control Priorities in Developing Countries* (DCP3), published as a set of serial volumes based on content area, provides evidence-based recommendations and cost-effectiveness analyses for numerous interventions, with attention to strategies for strengthening health systems. Cost-effectiveness analyses that compare relatively equivalent interventions in order to facilitate sound decisions under constraint are necessary; however, these analyses, as the DCP3 authors acknowledge, are unreliable when based on an incomplete knowledge of cost and evolving evidence of effectiveness. As both resources and objectives for global health initiatives grew, cost-effectiveness analyses (particularly those based on older evidence) sometimes steered policy makers and public health experts toward low-cost but ultimately ineffective interventions or away from higher-priced but effective ones. Thus we use the term *global health equity* to emphasize the

need to ensure equitable access to high-value health interventions. To illustrate these points, it is instructive to look to HIV/AIDS, which in the course of the last four decades has become one of the world's leading infectious causes of adult death.

HIV INFECTION/AIDS

Chapter 197 provides an overview of the global HIV epidemic today. Approximately 38.8 million people worldwide were living with HIV infection in 2015; >10.3 million of those in sub-Saharan African countries were then receiving antiretroviral therapy (ART)—a twofold increase over the corresponding figure for 2010, when ART was already the largest therapeutic rollout in the continent's history. Here the discussion will be limited to HIV/AIDS in the developing world. Lessons learned from tackling HIV/AIDS in low-resource settings are highly relevant to discussions of other chronic diseases, including NCDs, for which effective therapies have been developed. In the United States, after the mid-1990s, ART transformed HIV infection from an inescapably fatal disease into a manageable chronic illness. Across high-income countries, improved ART has dramatically prolonged life expectancy for people living with HIV infection, which now approaches that of the general population. This success rate exceeds that obtained with almost any treatment for adulthood cancer or for complications of coronary artery disease. In developing countries, treatment has been offered broadly only since 2003. Before 2003, many arguments were raised to justify not moving forward rapidly with ART programs for people living with HIV/AIDS in resource-limited settings. The standard litany included the price of therapy compared with the poverty of patients, the complexity of the intervention, the lack of infrastructure for laboratory monitoring, and the lack of trained health care providers. Narrow cost-effectiveness arguments that created false dichotomies—prevention *or* treatment rather than their synergistic integration—too often went unchallenged by policy makers, public health experts, and health economists. As a cumulative result of these delays in the face of health disparities both old and new, there were millions of premature deaths.

Disparities in access to HIV treatment did give rise to widespread moral indignation and a new type of health activism. In several middle-income countries, including Brazil, public programs have helped bridge the global access gap. Other innovative projects pioneered by international NGOs in diverse settings such as Haiti and Rwanda have established that a simple approach to ART based on intensive community engagement and social and economic support for patients and their village-based health workers can achieve remarkable results (Fig. 460-2).

FIGURE 460-2

An HIV/TB-co-infected patient in Rwanda before (***left***) and after (***right***) 6 months of treatment.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

During the past decade, the availability of ART has increased sharply in the low- and middle-income countries that have borne the greatest burden of the HIV/AIDS pandemic. In 2000 very few people living with HIV/AIDS in these nations had access to ART, whereas by 2015 46% of people living with HIV infection were receiving ART, including 51% of people living with HIV in sub-Saharan Africa and 41% in southern Asia. It

remains to be seen how many of these people are receiving ART regularly and with the requisite social support. In light of these dramatic gains, coverage targets have grown more ambitious; for example, the UNAIDS 90-90-90 Initiative aims to have 90% of people living with HIV know their status, 90% of those detected treated with ART, and 90% of those on treatment achieving viral load suppression by 2020.

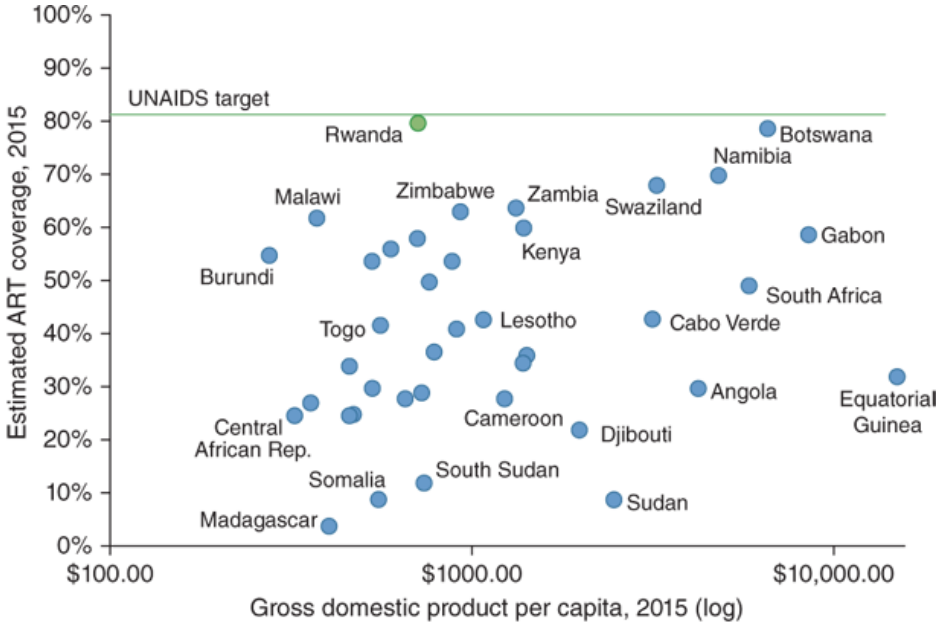
This scale-up was made possible by a number of developments: a staggering drop in the cost of generically manufactured ART, the development of a standardized approach to treatment, substantial investments by funders, and the political commitment of governments to afford ART as a public good. Civil-society AIDS activists spurred many of these efforts.

Starting in the early 2000s, a combination of factors, including work by the Clinton Foundation HIV/AIDS Initiative and Médecins Sans Frontières, led to the availability of generic ART medications. While first-line ART cost more than $10,000 per patient per year in 2000, first-line regimens in low- and middle-income countries are now available for <$100 per year. At the same time, fixed-dose combinations made multidrug regimens easier to administer. Also around this time, the WHO began advocating a public health approach to the treatment of people with AIDS in low-resource settings; this approach promised—thanks to dropping viremia—to lower transmission rates and, if universally available, to end almost all mother-to-child transmission. Derived from models of care pioneered by the NGO Partners in Health and other groups, this approach proposed the use of standard first-line treatment regimens based on a simple five-drug formulary, with a more complex (and more expensive) set of second-line options in reserve. Clinical protocols were standardized, and intensive training packages for health professionals and community health workers were developed and implemented in many countries. Early rollout efforts were supported by new funding from the Global Fund and PEPFAR. In 2003, lack of access to ART was declared a global public health emergency by the WHO and UNAIDS, and those two agencies launched the 3 by 5 Initiative, setting an ambitious target: to have 3 million people in developing countries on treatment by the end of 2005. Many countries set corresponding national targets and have worked to integrate ART into their national AIDS programs and health systems and to harness the synergies between HIV/AIDS treatment and prevention activities. External funding to fight HIV/AIDS in such settings increased dramatically during this period and beyond, rising from $300 million in 1996 to over $10.8 billion in 2015. The integration of prevention and care led to a sharp drop in transmission—a 96% decline according to one review of the impact of ART rollout in heavily burdened countries in Africa and the Caribbean.

Further lessons with implications for policy and action have come from efforts now under way among lower-income countries. Rwanda provides an example: since 2000, mortality from HIV disease has fallen by >80% as the country—despite its relatively low gross national income **(Fig. 460-3)**—has provided almost universal access to ART. The reasons for this success include strong national leadership, evidence-based policy, cross-sector collaboration, community-based care, and a deliberate focus on a health-systems approach that embeds HIV/AIDS treatment and prevention in the primary health care service delivery platform. As we will discuss later in this chapter, these principles can be applied to other conditions, including NCDs.

FIGURE 460-3

Antiretroviral therapy (ART) coverage (percentage of people living with HIV who received ART) in sub-Saharan Africa, 2015. *(Source: World Bank; data available through data.worldbank.org.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TUBERCULOSIS

Chapter 173 provides a concise overview of the pathophysiology and treatment of tuberculosis. In 2015, an estimated 1.4 million people died from *Mycobacterium tuberculosis* infection; this figure made tuberculosis the leading infectious killer of adults globally. The disease is closely linked to HIV infection in much of the world: of the 10.4 million estimated new cases of tuberculosis in 2015, 1.2 million occurred among people living with HIV. A much more substantial proportion of the resurgence of tuberculosis registered in southern Africa is attributed to HIV co-infection. Even before the advent of HIV, however, it was estimated that fewer than one-half of all cases of tuberculosis in developing countries were ever diagnosed. Primarily because of the common failure to diagnose and treat tuberculosis, international authorities devised a single strategy to reduce the burden of disease. In the early 1990s, the World Bank, the WHO, and other international bodies promoted the DOTS strategy (*d*irectly *o*bserved *t*herapy using *s*hort-course isoniazid- and rifampin-based regimens) as highly cost-effective. Passive case-finding of smear-positive patients was central to the strategy, as was an uninterrupted drug supply.

DOTS was clearly effective for most uncomplicated cases of drug-susceptible tuberculosis, but a number of shortcomings were soon identified. First, the diagnosis of tuberculosis based solely on sputum smear microscopy—a method dating from the late nineteenth century—is not sensitive. Many cases of pulmonary tuberculosis and all cases of exclusively extrapulmonary tuberculosis are missed by smear microscopy, as are most cases of active disease in children. Second, passive case-finding relies on the availability of health care services, which is uneven in the settings where tuberculosis is most prevalent. Third, patients with multidrug-resistant tuberculosis (MDR-TB) are by definition infected with strains of *M. tuberculosis* resistant to isoniazid and rifampin; thus exclusive reliance on these drugs is unwarranted in settings in which drug resistance is an established problem.

The crisis of antibiotic resistance registered in U.S. hospitals is not confined to the industrialized world or to common bacterial infections. While the great majority of patients sick with and dying from tuberculosis are afflicted with strains susceptible to all first-line drugs, a substantial minority of patients with tuberculosis in some settings are infected with strains of *M. tuberculosis* resistant to at least one first-line antituberculosis drug. Globally in 2015, an estimated 3.9% of all patients with new *M. tuberculosis* infections and 21% of all previously treated patients were infected with MDR strains; most of these cases resulted from primary transmission. It was clear that poor infection control in hospitals and clinics in the face of delays in the initiation of effective therapy led to explosive and lethal epidemics due to these strains. To improve DOTS-based responses to MDR-TB, global health authorities adopted DOTS-Plus, which adds the diagnostics and drugs necessary to manage drug-resistant disease. Even as DOTS-Plus was being piloted in resource-constrained settings, however, new strains of extensively drug-resistant (XDR) *M. tuberculosis* (resistant to isoniazid and rifampin, any fluoroquinolone, and at least one injectable second-line drug) had already threatened the success of tuberculosis control programs in beleaguered South Africa, for example, where high rates of HIV infection had led to a doubling in the incidence of tuberculosis over the preceding decade. Gene probes of cultures of infected sputum and tissues suggest that patients may be infected by more than one strain. Despite the poor capacity for detection of MDR- and XDR-TB in most resource-limited settings, an estimated 580,000 cases of MDR-TB were thought to occur in 2015. Approximately 9.5% of these cases were caused by XDR strains.

TUBERCULOSIS AND AIDS AS CHRONIC DISEASES: LESSONS LEARNED

Strategies effective against MDR-TB have implications for the management of drug-resistant HIV infection and even drug-resistant malaria, which, through repeated infections and a lack of effective therapy, has become a chronic disease in parts of Africa (see "Malaria," below). As new therapies, whether for tuberculosis or for hepatitis C infection, become available, many of the problems encountered in the past will recur. Indeed, examining AIDS and tuberculosis as chronic diseases—instead of simply communicable ones—makes it possible to draw a number of conclusions, many of them pertinent to global health equity in general.

First, the chronic infections discussed here are best treated with multidrug regimens to which the infecting strains are susceptible. This is true of chronic infections due to many bacteria, fungi, parasites, or viruses; even acute infections such as those caused by *Plasmodium* species are not reliably treated with a single drug.

Second, charging fees for AIDS prevention and care poses insurmountable problems for people living in poverty, many of whom are unable to pay even modest amounts for services or medications. Like efforts to battle airborne tuberculosis, such services might best be seen as a public good promoting public health. Initially, a subsidy approach will require sustained donor contributions, but many African countries have set targets for increased national investments in health—a pledge that could render ambitious programs sustainable in the long run, as the Rwanda experience suggests. Meanwhile, as local investments increase, the price of AIDS care continues to decrease. The use of generic medications means that ART can now cost <$0.25 per day.

Third, the effective scale-up of pilot projects requires strengthening and sometimes rebuilding of health care systems, including those charged with delivering primary care. In the past, the lack of health care infrastructure has been cited as a barrier to providing ART in the world's poorest

regions; however, AIDS resources, which are at last considerable, may be marshaled to rebuild public health systems in sub-Saharan Africa and other HIV-burdened regions—precisely the settings in which tuberculosis is resurgent. Failure to pursue such a health-systems approach after civil wars ended in Sierra Leone and Liberia accounts for much of their extreme vulnerability to Ebola a decade later.

Fourth, the lack of trained health care personnel, most notably doctors and nurses, is incorrectly invoked as a primary reason for failure to treat AIDS in poor countries and must still be addressed. The WHO recommends a minimum of 4.45 physicians, nurses, and midwives per 1000 persons, but recent reports from that organization and others confirm that many countries, especially in sub-Saharan Africa, fall far short of those target numbers. Specifically, 44% of WHO member states reportedly have <1 physician per 1000 population. In contrast, the United States and Cuba report 2.55 and 7.52 doctors per 1000 population, respectively. Similarly, about 48% of WHO member states report having fewer than three nurses and midwives per 1000 population. Further inequalities in health care staffing exist within countries. Rural–urban disparities in health care personnel mirror disparities of both wealth and health. For instance, nearly 90% of Malawi's population lives in rural areas, but more than 95% of clinical officers work at urban facilities, and 47% of nurses work in urban tertiary-care facilities. Even community health workers trained to provide first-line services to rural populations often transfer to urban districts.

In what is termed the "brain drain," many physicians and nurses emigrate from their home countries to pursue opportunities abroad, leaving behind health systems that are understaffed and ill-equipped to deal with either emergencies like Ebola or the usual burden of disease. One reason doctors and nurses leave sub-Saharan Africa and other resource-poor areas is that they lack the tools to practice there. Funding for "vertical" (disease-specific) programs can be used not only to strengthen health systems but also to recruit and train physicians and nurses to underserved regions where they, in turn, can help to train and then work with community health workers in supervising care for patients with AIDS and many other diseases within their communities. Such training should be undertaken even where physicians are abundant, since close community-based supervision represents the highest standard of care for chronic disease, whether in developing or developed countries. The United States, which has a dearth of health care providers in many of its poor and rural communities, has much to learn from Rwanda in this regard.

Fifth, the many barriers to adequate health care and patient adherence that are raised by extreme poverty can be removed only with the deployment of "wrap-around services": food supplements for the hungry, help with transportation to clinics, child care, and housing. Extreme poverty makes it difficult for many patients to comply with therapy for chronic diseases, whether communicable or not. Experience shows, however, that these many barriers can be more readily surmounted than the extreme poverty itself to which chronic disease and acute infection contribute substantially. Indeed, poverty in its many dimensions is far and away the greatest obstacle to the scale-up of treatment and prevention services.

Finally, there is a need for a renewed basic-science commitment to the discovery and development of vaccines; more reliable, less expensive diagnostic tools; and new classes of therapeutic agents. This need applies not only to the three leading infectious killers—against none of which there is an effective vaccine—but also to most other neglected diseases of poverty.

MALARIA

Chapter 219 reviews the etiology, pathogenesis, and clinical treatment of malaria, the world's fifth-ranking infectious killer. In 2015, there were ~212 million cases of malaria, and the disease is thought to have killed 429,000 people; 70% of these deaths (~303,000) occurred among children <5 years old. The poor disproportionately experience the burden of malaria: >75% of estimated malaria deaths occur in just 13 countries, and mortality rates are highest in sub-Saharan Africa. The Democratic Republic of the Congo and Nigeria account for >36% of total estimated malaria deaths globally.

Malaria's human cost has been enormous, with the highest toll among children—especially African children—living in poverty. Macroeconomic analyses estimate that malaria may reduce the per capita gross national product of a disease-endemic country by 50% relative to that of a nonmalaria-endemic country. The causes of this drag include impaired cognitive development of children, decreased schooling, decreased savings, decreased foreign investment, and restriction of worker mobility. In light of this enormous cost, it is little wonder that an important review by the economists Sachs and Malaney concludes that "where malaria prospers most, human societies have prospered least."

Microeconomic analyses focusing on direct and indirect costs estimate that malaria may consume >10% of a household's annual income. A study in rural Kenya shows that mean direct-cost burdens vary between the wet and dry seasons (7.1% and 5.9% of total household expenditure, respectively) and that this proportion is >10% in the poorest households in both seasons. A Ghanaian study that categorized the population by income group highlighted the regressive nature of this cost: responding to malaria consumed only 1% of a wealthy family's income but 34% of a poor household's income.

In part because of differences in vector distribution and climate, resource-rich countries offer few blueprints for malaria control and treatment that are applicable in tropical (and resource-poor) settings. In 2001, African heads of state endorsed the WHO Roll Back Malaria (RBM) campaign, which prescribes strategies appropriate for sub-Saharan African countries. In 2008, the RBM partnership launched the Global Malaria Action Plan (GMAP). This strategy integrates prevention and care and calls for the avoidance of single-dose regimens and an awareness of existing drug resistance; the use of insecticide-treated bed nets (ITNs); indoor residual spraying; artemisinin-based combination therapy (ACT); intermittent preventive treatment during pregnancy; prompt diagnosis; and other vector control measures such as larviciding and environmental management.

Between 2000 and 2015, global malaria deaths were reduced by an estimated 62%, a figure equating to some 6.8 million deaths averted. Again, the experience in Rwanda is instructive: from 2000 to 2015, malaria deaths dropped by >85% for the same reasons mentioned earlier in recounting that nation's successes in battling HIV. A recent resurgence there has been linked to inadequately treated ITNs and other delivery failures, rising temperatures, and reintroduction from surrounding countries (e.g., by refugees from Burundi).

Meeting the challenge of malaria control will continue to require careful study of appropriate preventive and therapeutic strategies in the context of an increasingly sophisticated molecular understanding of pathogen, vector, and host. However, an appreciation of the economic and social devastation wrought by malaria—like that inflicted by diarrhea, AIDS, and tuberculosis—on the most vulnerable populations should heighten the level of commitment to critical analysis of ways to implement proven strategies for prevention and treatment.

Funding from the Global Fund, the Gates Foundation, the World Bank's International Development Association, and the U.S. President's Malaria Initiative, along with leadership from public health authorities, is critical to sustain the benefits of prevention and treatment. Building on the growing momentum of the last decade with adequate financial support, innovative strategies, and effective tools for prevention, diagnosis, and treatment, we may yet achieve the goal of a world largely free of malaria.

EBOLA

Chapter 205 provides an overview of the epidemiology, pathogenesis, and clinical manifestations of Ebola virus and Marburg virus infections. The 2013–2016 outbreak of Ebola virus disease in West Africa was the largest documented Ebola epidemic to date, with more than 28,000 recorded cases and 11,000 recorded deaths. Such figures underestimate the true burden, however, given inaccuracies in reporting and a growing body of evidence documenting the occurrence of minimally symptomatic Ebola virus infection.

Prior to the outbreak, the health systems of the three most affected countries—Liberia, Guinea, and Sierra Leone—were among the world's weakest. Histories of extractive colonial and postcolonial commerce, the conditional aid policies of international financial institutions, recent civil conflict, and under-resourced health ministries left this part of West Africa bereft of the means to deliver modern medicine and promote public health. In 2013, Sierra Leone had the world's highest maternal mortality ratio, with 1460 deaths per 100,000 live births. According to one estimate, Liberia had just 51 physicians working in the entire country before the Ebola epidemic, or roughly one physician per 100,000 people. Clinics and hospitals were scarce across the region, especially in rural areas, and routinely lacked drugs, supplies, electricity, running water, laboratories, and personal protective equipment for the prevention of nosocomial infection. Such deficits were not surprising given these countries' meager public and private expenditures on health. In 2012, for example, the Guinean and Sierra Leonean governments allocated just 7 and 12% of their annual budgets, respectively, to health, according to WHO figures. Rwanda, in contrast, spent 22% of its national budget on health that year, the highest in the WHO African region.

The unprecedented scale of the recent West African Ebola epidemic was largely a symptom of these chronically weak health systems. As a result, clinicians, patients' families, and other caregivers—tasked with nursing the sick and interring the dead, but lacking the means to do so safely—faced disproportionately high risks of Ebola infection. Health facilities with poor infection control and unsafe burials served as amplifiers of transmission.

The quest to contain Ebola in West Africa was one of the largest global public health efforts in recent history, but it was far from ambitious clinically. As in previous Ebola outbreaks, preventing new infections was often prioritized over improving survival among those already infected, leading to substandard care for most West African patients and high case–fatality rates—by WHO estimates, ~70%. However, in settings in which quality supportive and critical care could be provided, clinical outcomes among Ebola-infected patients affirmed that Ebola virus disease is treatable, even in the absence of specific antiviral therapies and experimental drugs. Of the 27 Ebola patients treated in the United States and Europe in 2014 and 2015, more than 80% survived. This disparity of outcomes was also registered between European and African outbreaks of Marburg virus disease decades earlier.

As with efforts to combat AIDS and tuberculosis, the global response to Ebola reveals the unintended consequences of pitting preventive strategies against therapeutic ones—and the pull of debates about scarcity. Misguided (and often contradictory) public health messaging,

distrust of disease-control and social mobilization teams, punitive containment measures, and the unavailability of safe Ebola treatment units capable of delivering effective clinical care deterred individuals from presenting to health facilities, reporting symptomatic patients and their contacts, and cooperating with epidemic response activities. The resulting epidemic of mistrust facilitated the further spread of new infections by impeding surveillance, timely diagnosis, contact tracing, and patient isolation.

The aftermath of this West African epidemic provides valuable opportunities to improve global responses to acute health crises while addressing the chronic ones that fuel them. Funds pledged to support short-term, outbreak-response operations in West Africa often went undisbursed, highlighting the paucity of measures to hold donors accountable.

Moreover, the withdrawal of humanitarian organizations and of funding earmarked solely for disease containment illustrates the inadequacies of emergency responses detached from long-term efforts to strengthen national health systems. In the absence of such systems and in the face of national economies burdened by high youth unemployment, the clinical, social, and economic needs of thousands of Ebola survivors remain vast and largely neglected. The epidemic also claimed sizeable proportions of the region's health workforce, shuttered or incapacitated health centers, and collapsed the delivery of even the modest clinical and public health interventions that had been delivered prior to Ebola's explosive spread. Yet the public sectors of Liberia, Guinea, and Sierra Leone continue to lack adequate resources to rebuild their economies and health systems: estimates as of June 2017 show that, of the roughly $9.1 billion required to finance the national and regional plans presented at a 2015 United Nations conference on Ebola recovery, only $4.5 billion was pledged, of which just 26.4% has been disbursed. It remains unclear to what extent disbursed funds have been channeled through recipient governments (or other national institutions) and whether donors and implementing partners will allow these governments the flexibility to deploy the funds to best correspond to a rapidly changing burden of disease.

"NONCOMMUNICABLE" CHRONIC DISEASES

Although the burden of communicable diseases—especially HIV infection, tuberculosis, and malaria—still accounts for the majority of deaths in resource-poor regions within sub-Saharan Africa and in the poorest reaches of several first-world cities, 71.3% of all deaths worldwide in 2015 were attributed to NCDs. Although we use this term to describe cardiovascular diseases, cancers, diabetes, and chronic lung diseases, this usage masks important distinctions. For instance, two significant NCDs in low-income countries, rheumatic heart disease (RHD) and cervical cancer, represent the chronic sequelae of infections with group A *Streptococcus* and human papillomavirus, respectively, and it is in these countries that the burden of disease due to NCDs is rising most rapidly: close to three-quarters of deaths attributable to NCDs occur in low- and middle-income countries, which also account for 82% of all early NCD-related deaths—a figure representing ~16 million people and exceeding the total number of deaths due to AIDS, tuberculosis, and malaria combined. By 2020, NCDs will account for 80% of the global burden of disease and for seven of every 10 deaths in developing countries. The recent increase in resources for and attention to communicable diseases is both welcome and long overdue, but developing countries are already carrying a "double burden" of communicable and noncommunicable diseases.

Diabetes, Cardiovascular Disease, and Cancer: A Global Perspective

In contrast to tuberculosis, HIV infection, and malaria—diseases caused by single pathogens that damage multiple organs—cardiovascular diseases reflect injury to a single organ system downstream of a variety of insults, both infectious and noninfectious. Some of these insults result from rapid changes in diet and labor conditions; others are of a less recent vintage. The burden of cardiovascular disease in low-income countries represents one consequence of decades of neglect of health systems. Furthermore, cardiovascular research and investment have long focused on the ischemic conditions that are increasingly common in high- and middle-income countries.

Predictions of an imminent rise in the share of deaths and disabilities due to NCDs in developing countries have led to calls for preventive policies to improve diet, increase exercise, and restrict tobacco use, along with the prescription of multidrug regimens for persons at high-level vascular risk. Although this agenda could do much to prevent pandemic NCDs, it will do little to help persons with established heart disease stemming from nonatherogenic pathologies.

The misperception of cardiovascular diseases as a problem primarily of elderly populations in middle- and high-income countries has contributed to the neglect of these diseases by global health institutions, including regionally focused ones. Even in Eastern Europe and Central Asia, where the collapse of the Soviet Union was followed by a catastrophic surge in cardiovascular disease deaths (mortality rates from ischemic heart disease nearly doubled between 1991 and 1994 in Russia, for example), the modest flow of overseas development assistance to the health sector during these troubled years focused on the communicable causes that accounted for <1 in 20 excess deaths during that period. Even these focused investments were inadequate to halt resurgent tuberculosis (much of it caused by MDR strains), which is only now coming under control in much of the region.

DIABETES

The International Diabetes Federation reports that the number of diabetic patients in the world is expected to increase from 415 million in 2015 to 642 million by 2040—nearly one in 10 adults. Already, a significant proportion of diabetic patients live in developing countries where, because those affected are far more frequently between ages 40 and 59, the complications of micro- and macrovascular disease take a far greater toll. Globally, these complications are a major cause of disability and reduced quality of life: a high fasting plasma glucose level ranks third among risks for disability and global mortality. The GBD 2015 estimates that diabetes accounted for 1.5 million deaths in 2015; more than 80% of these deaths occurred in low- and middle-income countries.

CARDIOVASCULAR DISEASE

Because systemic investigation of the causes of stroke and heart failure in sub-Saharan Africa has begun only recently, little is known about the impact of elevated blood pressure in this portion of the continent. Modestly elevated blood pressure in the absence of tobacco use in populations with low rates of obesity may confer little risk of adverse events in the short run. In contrast, persistently elevated blood pressure above 180/110 goes largely undetected, untreated, and uncontrolled in this part of the world. In the cohort of men assessed in the Framingham Heart Study, the prevalence of blood pressures above 210/120—severe hypertension—declined from 1.8% in the 1950s to 0.1% by the 1960s with the introduction of effective antihypertensive agents. Although debate continues about appropriate screening strategies and treatment thresholds, Africa's rural health centers, run largely by nurses, must quickly gain access to antihypertensive medications.

The epidemiology of heart failure also reflects inequalities in risk factor prevalence and in access to therapy. The reported burden of this condition has remained unchanged since the 1950s, but the causes of heart failure and the age of the people affected vary across the globe. Heart failure as a consequence of pericardial, myocardial, endocardial, or valvular injury accounts for as many as 5% of all medical admissions to hospitals around the world. In high-income countries, coronary artery disease and hypertension among the elderly account for most cases of heart failure. Among the world's poorest 1 billion people, however, heart failure reflects poverty-driven exposure of children and young adults to rheumatogenic strains of streptococci and cardiotropic microorganisms (e.g., HIV, *Trypanosoma cruzi*, enteroviruses, *M. tuberculosis*), untreated high blood pressure, and nutrient deficiencies. The mechanisms underlying other causes of heart failure common in these populations—such as idiopathic dilated cardiomyopathy, peripartum cardiomyopathy, and endomyocardial fibrosis—remain unclear.

In stark contrast to the extraordinary lengths to which clinicians in wealthy countries will go to treat ischemic cardiomyopathy among elderly patients, little attention has been paid to young patients with non-ischemic cardiomyopathies in resource-poor settings. Non-ischemic cardiomyopathies, such as those due to hypertension, RHD, and chronic lung disease, account for >90% of cases of cardiac failure in sub-Saharan Africa and include poorly understood entities such as peripartum cardiomyopathy (which has an incidence in rural Haiti of one per 300 live births) and HIV-associated cardiomyopathy. Lessons learned in the scale-up of chronic care for HIV infection and tuberculosis may be illustrative as progress is made in establishing the means to deliver heart-failure medications to these patients.

Some of the lessons learned from the chronic infections discussed above are, of course, relevant to cardiovascular disease, especially those classified as NCDs but caused by infectious pathogens. Integration of prevention and care remains as important today as in 1960 when Paul Dudley White and his colleagues found little evidence of myocardial infarction in the region near the Albert Schweitzer Hospital in Lambaréné, Gabon, but reported that "the high prevalence of mitral stenosis is astonishing." They termed it a duty to integrate prevention with penicillin prophylaxis and care, including medical management and surgery, when indicated. "The same responsibility," they agreed, "exists for those with correctable congenital cardiovascular defects."

RHD affects more than 30 million people worldwide, with more than 282,000 new cases each year. Among the 2.4 million cases of pediatric RHD, an estimated 42% occur in sub-Saharan Africa. This disease, which may cause endocarditis or stroke, leads to >233,000 deaths per year—almost all occurring in developing countries. A survey of acute heart failure among adults in sub-Saharan Africa showed that ~14.3% of these cases were due to RHD. Researchers in Ethiopia have reported annual death rates as high as 12.5% in rural areas. In part because the prevention of RHD has not advanced since the disease's disappearance in wealthy countries, no part of sub-Saharan Africa has eradicated RHD despite examples of success in Costa Rica, Cuba, and some Caribbean nations.

Strategies to eliminate RHD may depend on active case-finding, with confirmation by echocardiography, among high-risk groups as well as on efforts to expand access to surgical interventions among children with advanced valvular damage. Partnerships between established surgical programs and areas with limited or nonexistent facilities may help expand the capacity to provide lifesaving interventions to patients who otherwise would die early and painfully. Such partnerships can speed the further development of regional centers of excellence equipped to provide consistent, accessible, high-quality services to those now without them.

CANCER

Low- and middle-income countries accounted for more than two-thirds of the 17.5 million cases and 8.7 million deaths due to cancer worldwide in 2015. By 2030, annual mortality from cancer is expected to increase by 4 million—with developing countries experiencing a sharper increase than developed nations. "Western" lifestyle changes may be responsible for the increased incidence of cancers of the breast, colon, and prostate among populations in low- and middle-income countries, but historic realities, sociocultural and behavioral factors, genetics, and poverty itself

already have a profound impact on cancer-related mortality and morbidity rates. In 2012, 15.4% of the more than 14 million cases of cancer globally were attributable to infectious causes, which are responsible for <10% of cancers in developed countries but account for up to 25% of all malignancies in low- and middle-income countries. Infectious causes of cancer such as human papillomavirus, hepatitis B virus, and *Helicobacter pylori* will continue to have a much larger impact in developing countries. Environmental and dietary factors, such as indoor air pollution and high-salt diets, also contribute to increased rates of certain cancers (e.g., lung and gastric cancers). Tobacco use (both smoking and chewing) is the most important source of increased mortality rates from lung and oral cancers. In contrast to decreasing tobacco use in many developed countries, the number of smokers is growing in developing countries, especially among women and young persons.

For many reasons, outcomes of malignancies are far worse in developing countries than in developed nations. As currently funded, overstretched health systems in poor countries are not capable of early detection; at the time of tissue diagnosis, the majority of patients already have incurable malignancies. Treatment of cancers is available for only a very small number of mostly wealthy citizens in the majority of poor countries, and, even when treatment is available, the range and quality of services are often substandard. Yet this need not be the future. Fifteen years ago, MDR-TB and HIV infections were widely deemed untreatable in settings of great poverty. The feasibility of creating innovative programs that reduce technical and financial barriers to the provision of care for treatable malignancies among the world's poorest populations is now clear **(Fig. 460-4)**. Several middle-income countries, including Mexico, have expanded publicly funded cancer care to reach poorer populations. This commitment of resources has dramatically improved outcomes for cancers, from childhood leukemia to cervical cancer.

FIGURE 460-4

An 11-year-old Rwandan patient with embryonal rhabdomyosarcoma before (***left***) and after (***right***) 48 weeks of chemotherapy plus surgery. Eleven years later, she is healthy with no evidence of disease.



A ***B***

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Prevention of Noncommunicable Diseases

False dichotomies, including those pitting prevention against care, persist in global health and reflect, in part, outmoded paradigms or a limited understanding of shifts in disease burden and causality as well as the dramatic variations in risk within a single nation. Moreover, such dichotomies or debates are sometimes politicized as a result of vested interests. Although globalization has had many positive effects, one negative effect has been the growth in both developed and developing countries of well-financed lobbies that have aggressively promoted unhealthy dietary changes and increased consumption of alcohol and tobacco. The WHO's 2003 Framework Convention on Tobacco Control represented a major advance, committing all of its signatories to a set of policy measures shown to reduce tobacco consumption. In 2004, the WHO released its Global Strategy on Diet, Physical Activity, and Health, which focused on population-level promotion of healthy diet and regular physical activity in an effort to reduce the growing global problem of obesity. Passing this strategy at the World Health Assembly proved difficult

because of strong opposition from the food industry and from a number of WHO member states, including the United States. The strategy fails to focus on the NCD risk factors of the bottom billion.

The WHO estimates that 80% of all cases of cardiovascular disease and type 2 diabetes as well as 40% of all cancers can be prevented through healthier diets, increased physical activity, and avoidance of tobacco. These estimates mask large local variations. Although some evidence indicates that population-based measures can have some impact on these behaviors, it is sobering to note that increasing obesity levels have not been reversed in any population. Tobacco avoidance may be the most important and most difficult behavioral modification of all. In the twentieth century, 100 million people worldwide died of tobacco-related diseases; it is projected that >1 billion people will die of these diseases in the twenty-first century, with the vast majority of those deaths in developing countries. Today, >80% of the world's 1 billion smokers live in low- and middle-income countries. If trends continue, tobacco-related deaths will increase to 8 million per year by 2030, with 80% of those deaths in low- and middle-income countries. Investment in curbing NCDs remains disproportionately low despite the WHO's 2008–2013 Action Plan for the Global Strategy for the Prevention and Control of Noncommunicable Diseases.

MENTAL HEALTH

The WHO reports that some 676 million people worldwide suffer from depression and anxiety disorders; one in four patients visiting a health service has at least one mental, neurologic, or behavioral disorder, but most of these disorders are neither diagnosed nor treated. More than 800,000 people die by suicide every year, and major depression is the leading cause of years lost to disability in the world today. Most low- and middle-income countries devote <1% of their health expenditures to mental health.

Increasingly effective therapies exist for many of the major causes of mental disorders. One of the greatest barriers to delivery of such therapies is the paucity of skilled personnel. Most sub-Saharan African countries have only a handful of psychiatrists, for example; almost all of them practice in cities and are unavailable within the public sector or to patients living in poverty. Among the few patients who are fortunate enough to see a psychiatrist or neurologist, fewer still are able to adhere to treatment regimens: several surveys of already diagnosed patients ostensibly receiving daily therapy have revealed that, among the poor, multiple barriers prevent patients from taking their medications as prescribed. In one study from Kenya, no patients being seen in an epilepsy clinic had therapeutic blood levels of antiseizure medications, even though all had been prescribed these drugs. Moreover, many patients in this study had no detectable blood levels of these agents. The same barriers that prevent the poor from having reliable access to insulin or ART prevent them from benefiting from antidepressant, antipsychotic, and antiepileptic agents. To alleviate this problem, some authorities are proposing the training of health workers to provide community-based adherence support, counseling services, and referrals for patients in need of mental health services. One such program instituted in Goa, India, used lay counselors and resulted in a significant reduction in symptoms of common mental disorders among the target population.

CONCLUSION: TOWARD A SCIENCE OF IMPLEMENTATION

There is a long way to go before evidence-based internal medicine is applied effectively among the world's poor. Public health strategies draw largely on quantitative methods—epidemiology, biostatistics, and economics. Clinical practice, including the practice of internal medicine, draws on a rapidly expanding knowledge base and remains focused on individual patient care; clinical interventions are rarely population-based. However, global health equity depends on avoiding the false dichotomies of the past: neither public health nor clinical approaches alone are adequate to address the problems of global health. The integration of prevention and care, along with adequate funding, has shown that complex infectious diseases such as HIV/AIDS and tuberculosis are not impossible to manage, even though drug resistance and lack of effective health systems have complicated such work. Beyond what is usually termed communicable disease—i.e., in the arena of chronic diseases such as cardiovascular disease and mental illness—global health is a nascent endeavor. Efforts to address any one of these problems in settings of great scarcity need to be integrated into broader efforts to strengthen failing health systems and alleviate the growing personnel crisis within these systems. Such efforts must include the building of platforms for care delivery that are robust enough to incorporate new preventive, diagnostic, and therapeutic technologies rapidly in response to changes both in the burden of disease and in the needs not met by existing paradigms and systems of care delivery.

Academic medical centers have tried to address this "know–do" gap as new technologies are introduced and assessed through clinical trials, but the reach of these institutions into settings of poverty is limited in rich and poor countries alike. When such centers link their capacities effectively to the public institutions charged with the delivery of health care to the poor, great progress can be made. For these reasons, scholarly work and practice in the field once known as "international health" and now often designated *global health equity* are changing rapidly. That work is still informed by the tension between clinical practice and population-based interventions, between analysis and action, and between prevention and care.

A number of university hospitals are developing training programs for physicians with an interest in global health. In medical schools across the United States and in other wealthy countries, interest in global health has exploded. One study has shown that more than 25% of medical students take part in at least one global health experience prior to graduation. Half a century or even a decade ago, such high levels of interest would have been unimaginable.

An estimated 400 million people lack access to essential health services; the consequence is millions of preventable deaths each year. An absolute majority of these premature deaths occur in Africa, with the poorer regions of Asia not far behind. They include deaths from vaccine-preventable illness, deaths during childbirth, deaths from infectious diseases that might be cured with access to antibiotics and other essential medicines, deaths from malaria that would have been prevented by ITNs and access to therapy, and deaths from waterborne illnesses. Other excess mortality is attributable to the inadequacy of efforts to develop new preventive, diagnostic, and therapeutic tools.

The development of tools must be followed quickly by their equitable distribution. Those funding the discovery and development of new tools typically neglect the concurrent need for strategies to make them available to the poor. Indeed, some would argue that the biggest challenge facing those who seek to address this outcome gap is the lack of practical means of delivery in the most heavily affected regions. When new preventive and therapeutic tools are developed without concurrent attention to delivery or implementation, one encounters what are sometimes termed “perverse effects”: even as new tools are developed, inequalities of outcome—lower morbidity and mortality rates among those who can afford access, with sustained high morbidity and mortality among those who cannot—grow in the absence of an equity plan to deliver the tools to those most at risk. Preventing such a future is the most important goal of global health.

FURTHER READING

Dieleman JL et al: Development assistance for health: Past trends, associations, and the future of international financial flows for health. Lancet 387:2536, 2016. [PubMed: 27086170]

Farmer P: Chronic infectious disease and the future of health care delivery. N Engl J Med 369:2424, 2013. [PubMed: 24350951]

Farmer P et al (eds): *Reimagining Global Health: An Introduction.* Berkeley, University of California Press, 2013.

GBD 2015 Mortality and Cause of Death Collaborators: Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015. Lancet 388:1459, 2016. [PubMed: 27733281]

GBD 2015 DALYs and HALE Collaborators: Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. Lancet 388:1603, 2016. [PubMed: 27733283]

Kim JY et al: Redefining global health-care delivery. Lancet 382:1060, 2013. [PubMed: 23697823]

Mills A: Health care systems in low-and middle-income countries. N Engl J Med 370:552, 2014. [PubMed: 24499213]

Chapter 461: Worldwide Changes in Patterns of Infectious Disease

George W. Rutherford; Jaime Sepulveda

THE CONCEPT OF EMERGING INFECTIOUS DISEASES

Pathogenic organisms have been a constant companion of humans, their livestock, and their cultivated plants throughout evolution. Over the centuries, new organisms emerged as ecology changed or as humans managed to cross ecologic barriers such as oceans and mountains. Throughout history, there have been severe epidemics of infectious diseases, with devastating consequences to human populations over vast geographic regions. From the Plague of Justinian in Europe in the sixth century, to the Black Death in the fourteenth century, to the five cholera pandemics of the nineteenth century, to the 1918 Spanish influenza pandemic, to the current HIV/AIDS pandemic, the death toll in human populations has been enormous. The concept of emerging infectious diseases arose in the 1970s and 1980s with the recognition of several “new” diseases, such as legionellosis, HIV infection, Lyme disease, and toxic shock syndrome and was later expanded to include reemerging infectious diseases, such as tuberculosis. In 1991, the Institute of Medicine (IOM) convened a multidisciplinary committee to elucidate emerging microbial threats to health, with particular reference to the United States. In its report, the committee defined an emerging infectious disease as a disease “of infectious origin whose incidence in humans has either increased within the past two decades or threatens to increase in the near future.” In the year following the publication of the committee’s report, large outbreaks of *Escherichia coli* O157:H7 infection, cryptosporidiosis, and hantavirus pulmonary syndrome spurred the development of a national plan to recognize and interdict emerging and reemerging infectious disease threats by the Centers for Disease Control and Prevention (CDC). Since then, the list of emerging and reemerging viral, bacterial, fungal, and parasitic diseases has grown to include multiple infections and syndromes. Examples of emerging and reemerging infectious diseases, as of 2017, are shown in Table 461-1.

TABLE 461-1

Examples of Emerging and Reemerging Infectious Diseases

Bacterial and Rickettsial	Viral and Prion	Fungal and Parasitic
Anaplasmosis Anthrax Lyme disease *Vibrio cholerae* O139 infection Diphtheria Ehrlichiosis *Escherichia coli* O157:H7 infection *Escherichia coli* O154:H4 infection *Legionella pneumophila* infection Plague Vancomycin-resistant *Staphylococcus aureus* infection Tuberculosis	Chikungunya virus infection Variant Creutzfeldt-Jakob disease Dengue Ebola virus, Marburg virus infection Enterovirus D68 infection Hantavirus (Sin Nombre, Seoul) infection Hendra virus, Nipah virus infection Hepatitis C Hepatitis E HIV-1 and -2 infection Human herpesvirus 6, 8 infection Human T-lymphotropic virus 1 and 2 infection Influenza A H1N1pdm, H5N7, H7N7, H7N9 Lassa fever Lyssavirus infection Middle East respiratory syndrome (MERS) Monkeypox Rift Valley fever virus infection Severe acute respiratory syndrome (SARS) West Nile virus infection Whitewater Arroyo virus infection Yellow fever Zika	Coccidioidomycosis *Cryptosporidium parvum* infection *Cyclospora cayetanensis* infection Drug-resistant malaria

The reasons for the emergence of previously unrecognized diseases and the reemergence of diseases that have previously been largely under control are legion. At its core, however, emergence has to do with genetic changes in disease agents or changes in ecology, which includes human behavior. The IOM committee listed six primary reasons for disease emergence or reemergence: human demographics and behavior, technology

and industry, economic development and land use, international travel and commerce, microbial adaptation and change, and breakdown of public health measures. A disease can emerge or reemerge for one or more of these reasons. For example, the worldwide spread of severe acute respiratory syndrome (SARS) began as a species cross-over, most likely involving transmission of a previously unknown coronavirus of horseshoe bats to Himalayan palm civets that were subsequently captured and transported to live-animal (i.e., “wet”) markets in Guangzhou, China, for human consumption. The SARS coronavirus was then transmitted to humans—most likely by restaurant workers—and from them to medical personnel and eventually to individuals around the world. This spread had nothing to do with migratory patterns of bats or civets but was, instead, a consequence of human travel. The cities most effected by SARS—Hong Kong, Beijing, and Toronto—became involved because of rapid human movement via international passenger aircraft.

To this list, additional factors can now be added. Either therapeutic or acquired immunosuppression (e.g., in HIV infection) can render populations susceptible to infections that have not previously been recognized, such as that with human herpesvirus 8—the cause of Kaposi’s sarcoma. Climate change, in particular, can expand the host range of disease-transmitting vectors. In addition, the weaponization of pathogenic organisms for biological terrorism or warfare can lead, at least theoretically, to prolonged chains of human-to-human transmission. One factor is clear: the preponderance of emerging infectious diseases are zoonotic in origin. The authors of a 2008 review calculated that 60.3% of all emerging infectious disease events from 1940 to 2004 were zoonotic in origin, and 71.8% of these zoonotic events originated in wildlife.

In this chapter, we review the recent changing epidemiology of four emerging infectious viral diseases that exemplify some of the IOM’s principles for emergence: infections caused by West Nile virus, dengue virus, Ebola virus, and Zika virus. This list is clearly not exhaustive but highlights a few prominent instances of the recent emergence of infectious diseases and their root causes.

EXAMPLES OF EMERGING INFECTIOUS DISEASES

WEST NILE VIRUS (WNV)

WNV is a flavivirus that was originally discovered in Uganda in 1937 and emerged as a cause of neurologic disease in humans and equines. WNV exists in nature in an enzootic cycle that involves certain birds and mosquitoes, particularly those of the genera *Culex* and *Aedes*. Humans, horses, and other vertebrates are incidental hosts and, except through blood transfusion, are unlikely to transmit WNV because levels of viremia are insufficiently high to infect mosquitoes. When originally described, WNV was believed to cause a mild febrile illness, but subsequent experience showed that it caused neuroinvasive disease in some cases. The first cases of neuroinvasive disease were described in an outbreak among elderly patients in Israel and subsequently in humans and horses in the Mediterranean basin, India, and South Africa. By the 1990s, outbreaks had been reported from Romania, Russia, and Central Asia; these outbreaks were probably a result of seasonal bird migrations from endemic Mediterranean countries, with introduction of infected mosquitoes and the establishment of infection in local bird species.

An explosive outbreak of WNV infection began in the United States in the summer of 1999 and initially involved infection of birds of the family Corvidae (e.g., the American crow and blue jay) that were susceptible to neuroinvasive disease. The first human cases appeared in New York City that same summer. Subsequently, sufficient numbers of birds more resistant to neuroinvasive disease and mosquitoes of the genus *Culex* became infected and an enzootic cycle was established in North America. Over the next 3 years, WNV spread across the continental United States, Canada, and Mexico and became an important cause of human and equine neurologic disease. The WNV clade causing the North American outbreak was the same (clade 1a) as that causing disease in the Middle East, Europe, North Africa, and parts of Asia.

In 2016, 2038 cases of WNV infection in humans, including 1140 cases of neuroinvasive disease, were reported in the United States; these figures are certainly gross underestimations of the actual number of cases. There were 94 deaths, primarily among the elderly. An additional 377 cases were reported in horses from 32 states despite the availability of a reasonably protective equine vaccine. Human cases were reported from 44 states, with only Alaska, Delaware, Hawaii, Maine, New Hampshire, and West Virginia free of the disease. Infected mosquito pools were even more widespread; Maine was the only state in the continental United States to be free of all WNV activity. Thus, from an initial introduction into New York City, WNV has successfully established itself across North America and infected an estimated 2.6–6.1 million people in the United States (1.1% of the population).

Why did this happen? First, microorganisms and larger organisms, such as plants and animals, have been exchanged between the Old and New Worlds since the initial voyages of exploration in the fifteenth and sixteenth centuries. However, it is the advent of modern high-speed transportation that allows vectors, such as mosquitoes, to move between continents in hours or days as opposed to months or years. In the most likely scenario for the introduction of WNV into North America, a single viremic mosquito was accidentally transported from an area endemic for clade 1a to New York City in the cargo hold of an airplane in 1999. The original strain associated with the 1999 outbreak (NY99) had caused outbreaks in Tunisia and Israel in 1997 and 1998, respectively; this information suggests that one of those countries was the source. The imported strain in turn infected crows, which in turn infected more competent mosquitoes, establishing an enzootic life cycle in North America with at least three *Culex* species and multiple species of birds involved. This scenario represents a successful invasion of WNV into a new ecologic niche.

The likelihood that WNV will gradually disappear is low. The virus has many avian hosts and more than one mosquito vector; it has undergone at least one successful mutation in the North American outbreak, thereby becoming infectious to *Culex piperans* and *Culex tarsalis*—mosquitoes with a broad range in the western United States. Moreover, the occurrence of outbreaks in 17 consecutive years in North America suggests that WNV has been successfully introduced onto the continent and will remain endemic for years to come.

DENGUE VIRUS

Dengue is the most important of the human arboviral infections, with almost half of the world's population at risk. Occurring in the range of *Aedes* mosquitoes, dengue virus infection imposes a heavy burden of morbidity and mortality worldwide, with as many as 50–200 million infections, 500,000 severe cases, and

20,000 deaths annually. Dengue virus is a flavivirus and exists in four serotypes (DENV1–4) that circulate independently of one another; immunity to one serotype does not confer immunity to the others.

Dengue is transmitted primarily by *Aedes aegypti* and secondarily by *Aedes albopictus*. The original life cycle of dengue virus was most likely similar to that of yellow fever, consisting of sylvatic transmission from mosquitoes to nonhuman primates and back to mosquitoes; over the past few centuries, the virus has adapted to an urban and periurban mosquito–human–mosquito cycle as well. Dengue and its more severe manifestations, dengue hemorrhagic fever and dengue shock syndrome, were first described in outbreaks in Japan in 1943 and Hawaii in 1945. However, clinically similar diseases had been reported during the previous two centuries in a geographic band extending from India south to Queensland, Australia, and east through Polynesia; in addition, there had been occasional outbreaks in areas as disparate as Greece, Panama, and southern Texas.

The ecology of dengue changed dramatically in the second half of the twentieth century, led by the successful invasion by *A. aegypti* of the global tropics after World War II, with the postwar dispersion of troops and materiel. From its ancestral roots in Southeast Asia, all four dengue serotypes spread globally in the tropics. DENV2 had been introduced into West Africa by the 1960s and established both sylvatic enzootic nonhuman primate and urban endemic human cycles. Travel and commerce facilitated importations, probably through both viremic human hosts and infected mosquitoes. In the Americas in particular, a campaign to eradicate *A. aegypti*, which is also the principal vector of yellow fever, failed in the mid-1970s, and both *A. aegypti* and dengue virus, especially DENV2, rapidly reinvaded their prior habitat; thus dengue reemerged as a major arboviral disease extending from the southern United States in the north, through northern Argentina in the south. Recent outbreaks have occurred along the U.S.–Mexico border and in the state of São Paulo in Brazil, where DENV1, DENV2, and DENV4 are co-circulating.

Dengue's emergence and spread have been intimately linked to human activity. In particular, globalization, with the movement of viremic people and mosquitoes through modern transportation of both passengers and goods, has been critical to dengue's success. One particular adaptation has also facilitated its urban spread: *Aedes* is able to breed in standing water associated with human habitation, such as cisterns, ornamental ponds, puddles, and water trapped in abandoned tires. This ability of *Aedes* has allowed dengue to be one of the only two known arboviruses (the other being Zika) that are adapted to an urban environment and can replicate entirely in a mosquito-to-human cycle. Together, these factors have led to widespread dengue transmission in a band extending across the tropics worldwide.

EBOLA AND MARBURG VIRUSES

Ebola virus is a filovirus that most likely exists in a sylvatic cycle in bats in Central and West Africa. Four strains are known to cause human disease. The first outbreak was described in Zaire in 1976. Since then, 29 outbreaks have been reported across tropical Africa, ranging in size from tens of cases to tens of thousands of cases in the West African outbreak of 2013–2016.

The life cycle of Ebola virus in the wild is not fully understood. There is evidence for sustained transmission in fruit bats, with occasional nonhuman primate spillover infections. It has been speculated that humans become infected from contact with infected bats or nonhuman primates, but, once an index case has occurred, essentially all transmission is from human-to-human contact with blood and other body fluids. Preparing bodies for burial has been an especially efficient means of transmission. In addition, health care providers who do not wear adequate personal protective equipment while caring for Ebola patients are particularly vulnerable to acquiring infection. In the West African epidemic, there was only a single zoonotic introduction and all subsequent transmission was from human to human.

The principal cause of Ebola outbreaks prior to the West African outbreak was the migration of humans into sylvatic areas, with enzootic transmission and accidental infection. In West Africa, only a single case had been recognized in Côte d'Ivoire before the 2013–2016 outbreak in the Republic of Guinea, Liberia, and Sierra Leone. It has been speculated that cultivation of palm oil attracted fruit bats, who feed on palm fruit; if so, environmental modification from dense tropical forest to palm oil plantations may have been a contributory cause. Other evidence suggests that the index patient—a 2-year-old boy—was exposed to insectivorous free-tailed bats (*Mops condylurus*). Whatever the initial event, the explosive amplification that occurred in these countries and the seven countries to which cases were exported was due to an inadequate medical and public health infrastructure. In fact, when Ebola virus was imported to countries with more functional public health systems, such as Nigeria, transmission was extinguished within three generations.

Other filovirus outbreaks have involved the transport of infected primates for medical research. The original Marburg virus outbreak, which occurred in Marburg and Frankfurt, West Germany, and Belgrade, Yugoslavia, in 1967, was likely caused by the importation of African vervet monkeys (*Cercopithecus aethiops*) from Uganda for medical research. This outbreak resulted in 31 human cases and 7 deaths. In addition, an outbreak among five crab-eating macaques (*Macaca fascicularis*) imported from the Philippines and infected with Reston Ebola virus—a strain nonpathogenic for humans—led to an epizootic in northern Virginia in 1989, eventually resulting in the culling of more than 500 primates. This outbreak, however, had no human cases associated with it, although epidemiologic investigation identified a handful of asymptomatic primate handlers who were seropositive for Reston Ebola virus. Since 1989, four additional outbreaks have been recognized in *Cynomolgus* monkeys imported from the Philippines to the United States and Italy.

A new reservoir of Ebola virus infection has now been identified: the semen of patients who have survived Ebola infection. The occurrence of several small clusters of sexually transmitted cases developing up to 284 days after symptom onset indicates prolonged carriage of Ebola virus in the testes. Moreover, the virus may remain viable over the long term in the vitreous humor.

Thus, Ebola represents a spillover event to humans and nonhuman primates from their interaction with certain species of infected and infectious bats. Contact with either the bats themselves or an infected nonhuman primate leads to infection of an index patient, which leads in turn to ongoing transmission from humans to humans. Several factors clearly contribute to the continued transmission. First, medical and public health systems are weak in severely affected countries. As experience with Ebola grows and the capacity for surveillance and response improves, numbers of secondary cases can fall; for example, in five

outbreaks in Uganda stretching from 2000 to 2012, the numbers of secondary cases and the geographic spread of the outbreaks decreased with each new introduction. Second, behavioral factors contribute, in particular funeral practices that bring mourners into close contact with infectious blood and tissues during preparation of a body for burial. Third, the areas in which the initial waves of transmission occur are often remote; thus, recognition of the outbreak can be delayed and, in the case of the West African outbreak, highly mobile populations can travel to larger cities to seek care.

ZIKA VIRUS

Zika virus is a flavivirus that is transmitted by *Aedes* mosquitoes and was originally described as an infection of nonhuman primates in Uganda in 1947. The first human cases were reported in Uganda in 1962 and 1963. Zika was thought to be an illness causing a mild rash and fever in humans in tropical Africa and southern Asia. The clinical and serologic similarity of Zika infection to dengue virus infection may have led to missed outbreaks. Since 2007, an Asian lineage of Zika virus has spread from the Western Pacific (initially, Yap Island) through Polynesia and on to Easter Island, Chile, where it was documented in 2014. From Polynesia, it also spread to Brazil, most likely through viremic travelers attending the world Va'a World Sprint Championships (Polynesian canoe racing) in Rio de Janeiro in the late summer of 2014. From there, Zika virus has spread hemisphere-wide, following the host range of *A. aegypti*. Forty-eight countries in the Americas have now reported autochthonously transmitted Zika virus infections.

In tropical Africa and Asia, Zika virus is most likely transmitted in a nonhuman primate–mosquito sylvatic cycle. Other animals may be involved in Zika's life cycle as well. A number of *Aedes* species are competent vectors, although *A. aegypti* may be the source of the majority of infections worldwide.

As Zika virus spread through Latin America and the Caribbean, a parallel epidemic of fetal microcephaly appeared; this epidemic was both temporally and geographically associated with the spread of Zika virus. More than 1.6 million cases of Zika virus infection, including 41,473 cases in pregnant women and 1950 cases of Zika-associated microcephaly, were reported from Brazil alone in 2015 and 2016. Data from a large registry of Zika-exposed pregnancies in the U.S. territories show that the overall risk of microcephaly following confirmed Zika virus infection is ~5%, ranging from 8% for infection in the first trimester to 4% for infection in the third trimester. Other fetal complications include stillbirth, neural tube defects, eye abnormalities, and sensorineural deafness. Complications in adults occur in about one of every thousand cases and include Guillain-Barré syndrome, encephalitis, leukopenia, and thrombocytopenic purpura. Moreover, it is now recognized that Zika virus can be transmitted sexually and via blood transfusion.

Thus, the introduction of Zika into the Americas represents viral invasion of a new ecosystem already widely populated by a highly competent mosquito host with an established urban habitat and an immunologically inexperienced human population. The invasion by Zika virus is in many ways similar to the original dengue invasion in the Americas in the 1950s and to the introduction of WNV into North America in 1999. Both the original importation of Zika virus and its establishment of new foci in the Americas (e.g., Florida and the Caribbean) were consequences of modern travel. Zika's spread has also been linked to climate variations, deforestation, and urban poverty.

CONTROL OF EMERGING INFECTIOUS DISEASES

Humans will continue to experience outbreaks of emerging and reemerging infectious diseases. Emerging diseases will most likely come from two sources. The first source consists of organisms that have acquired new genetic materials from other strains of the same species or from different species altogether. An example of this is influenza A virus, in which strains can acquire new genetic material through a process called *reassortment*. If the new gene is a hemagglutinin gene, the resulting reassortant virus will have a new surface hemagglutinin that is unrecognized immunologically by most human populations. An interesting case is influenza A H1N1 virus, which emerged in 2009 from the reassortment of H1N1 swine influenza virus with human seasonal H3N2 influenza virus, North American avian influenza virus, and Eurasian avian-origin swine influenza viruses. Despite a worldwide pandemic, people born before 1950 were relatively spared because they had earlier exposure to an H1N1 strain sufficiently similar to provide them with cross-immunity. Another example of this is *E. coli* O157:H7, which acquired a virulence gene from *Shigella*, probably as the result of horizontal genetic exchange. The resulting organism and several other serotypes of *E. coli* that have acquired the gene are the leading cause of hemolytic–uremic syndrome worldwide. The second source for emergence consists of existing organisms entering new ecologic niches and spreading broadly, usually through insect vectors, to immunologically naïve humans—as occurred with Zika virus. A variation on this theme is humans entering new ecosystems and becoming infected with organisms to which they have no immunity. An organism's epidemic potential will be determined by whether it is largely incapable of leaving the human host to continue onward via human-to-human transmission (e.g., *Coccidioides*) or can be efficiently transmitted from human to human (e.g., HIV and Ebola virus).

In its 1994 strategic plan to address emerging infectious disease threats, the CDC listed four goals: (1) to detect, promptly investigate, and monitor emerging pathogens, the diseases they cause, and the factors influencing their emergence; (2) to integrate laboratory science and epidemiology in order to optimize public health practice; (3) to enhance communication of public health information about emerging diseases and ensure prompt implementation of prevention strategies; and (4) to strengthen local, state, and federal public health infrastructures in order to support surveillance and implement prevention and control programs. Much of this plan has been implemented. The concept of "emerging infectious diseases" has been broadly accepted, and molecular biological methods have improved to the point that, for example, the SARS coronavirus was completely sequenced in a matter of days. In addition, there has been an increasing recognition of the "one health" concept: the nexus among human, livestock, wildlife, and plant health and the development of surveillance systems to provide early warnings of emerging and reemerging infections. New vaccines and new vector-control agents are important promising weapons in the struggle to contain existing diseases; vaccines against both dengue and Ebola have been shown to be efficacious in phase 3 trials, and a new vector-control technique involving deliberate infection of the *Aedes* population with *Wolbachia*, a bacterial genus that inhibits the transmission of arbovirus from mosquito hosts, is being evaluated.

The World Health Organization has developed new international health regulations that are designed, in part, to facilitate the recognition and reporting of infectious disease threats. However, as evidenced by the

2013–2016 Ebola virus epidemic in West Africa, additional capacity and possibly new forms of global health governance and response may be required. What is clear is that humans will continue to experience new and reemerging infections, and we will need robust, flexible, and timely responses in order to control them.

FURTHER READING

Campbell-Lendrum D et al: Climate change and vector-borne diseases: What are the implications for public health research and policy? Philos Trans R Soc Lond B Biol Sci 370:20130552, 2015. [PubMed: 25688013]

Heymann DL et al: Global health security: The wider lessons from the west African Ebola virus disease epidemic. Lancet 385:1884, 2015. [PubMed: 25987157]

Jones KE et al: Global trends in emerging infectious diseases. Nature 451:990, 2008. [PubMed: 18288193]

Lederberg J et al (eds): Committee on Emerging Microbial Threats to Health. Emerging infections. Microbial threats to health in the United States. Washington, DC, National Academy Press, 1992.

Morens DM, Fauci AS: Emerging infectious diseases in 2012: 20 years after the Institute of Medicine report. mBio 3:e00494, 2012. [PubMed: 23232716]

Chapter 462: Primary Care and Global Health

Tim Evans; Kumanan Rasanathan

INTRODUCTION

The twentieth century witnessed the rise of an unprecedented global health divide. Industrialized or high-income countries experienced rapid improvement in standards of living, nutrition, health, and health care. Meanwhile, in low- and middle-income countries with much less favorable conditions, health and health care progressed much more slowly. The scale of this divide is reflected in the current extremes of life expectancy at birth, with Japan at the high end (85 years) and Chad at the low end (50 years). This 35-year difference reflects the daunting range of health challenges faced by low- and middle-income countries. These nations must deal not only with a complex mixture of diseases (both infectious and chronic) and illness-promoting conditions but also, and more fundamentally, with the fragility of the foundations underlying good health (e.g., sufficient food, water, sanitation, and education) and of the systems necessary for universal access to good-quality health care. In the last decades of the twentieth century, the need to bridge this global health divide and establish health equity was increasingly recognized. The Declaration of Alma Ata in 1978 crystallized a vision of justice in health, regardless of income, gender, ethnicity, or education, and called for "health for all by the year 2000" through primary health care. While much progress has been made since the declaration, at the end of the first decade and half of the twenty-first century, much remains to be done to achieve global health equity.

This chapter looks first at the nature of the health challenges that underlie the health divide in low- and middle-income countries. It then outlines the values and principles of a primary health care approach, with a focus on primary care services. Next, the chapter reviews the experience of low- and middle-income countries in addressing health challenges through primary care and a primary health care approach. Finally, the chapter identifies how current challenges and global context provide an agenda and opportunities for the renewal of primary health care and primary care, allied to the movement to achieve universal health coverage.

PRIMARY CARE AND PRIMARY HEALTH CARE

The term *primary care* has been used in many different ways: to describe a level of care or the setting of the health system, a set of treatment and prevention activities carried out by specific personnel, a set of attributes for the way care is delivered, or an approach to organizing health systems that is synonymous with the term *primary health care*. In 1996, the U.S. Institute of Medicine encompassed many of these different usages, defining primary care as "the provision of integrated, accessible health care services by clinicians who are accountable for addressing a large majority of personal health care needs, developing a sustained partnership with patients, and practicing in the context of family and community."[1] We use this definition of *primary care* in this chapter. Primary care performs an essential function for health systems, providing the first point of contact when people seek health care, dealing with most problems, and referring patients onward to other services when necessary. As is increasingly evident in countries of all income levels, without strong primary care, health systems cannot function properly or address the health challenges of the communities they serve.

Primary care is only one part of a primary health care approach. The Declaration of Alma Ata, drafted in 1978 at the International Conference on Primary Health Care in Alma Ata (now Almaty in Kazakhstan), identified many features of primary care as being essential to achieving the goal of "health for all by the year 2000." However, it also identified the need to work across different sectors, address the social and economic factors that determine health, mobilize the participation of communities in health systems, and ensure the use and development of technology that was appropriate in terms of setting and cost. The declaration drew from the experiences of low- and middle-income countries in trying to improve the health of their people following independence. Commonly, these countries had built hospital-based systems similar to those in high-income countries. This effort had resulted in the development of high-technology services in urban areas while leaving the bulk of the population without access to health care unless they traveled great distances to these urban facilities. Furthermore, much of the population lacked access to basic public health measures. Primary health care efforts aimed to move care closer to where people lived, to ensure their involvement in decisions about their own health care, and to address key aspects of the physical and social environment essential to health, such as water, sanitation, and education.

After the Declaration of Alma Ata, many countries implemented reforms of their health systems based on primary health care. Most progress involved strengthening of primary care services; unexpectedly, however, much of this progress was seen in high-income countries, most of which constructed systems that made primary care available at low or no cost to their entire populations and that delivered the bulk of services in primary care settings. This endeavor also saw the reinforcement of family medicine as a specialty to provide primary care services. Even in the United States (an obvious exception to this trend), it became clear that the populations of states with more primary care physicians and services were healthier than those with fewer such resources.

Progress was also made in many low- and middle-income countries. However, the target of “health for all by the year 2000” was missed by a large margin. The reasons were complex but partly entailed a general failure to implement all aspects of the primary health care approach, particularly work across sectors to address social and economic factors that affect health and provision of sufficient human and other resources in order to make possible the access to primary care attained in high-income countries. Furthermore, despite the consensus in Alma Ata in 1978, the global health community rapidly became fractured in its commitment to the far-reaching measures called for by the declaration. Economic recession tempered enthusiasm for primary health care, and momentum shifted to programs concentrating on a few priority measures such as immunization, oral rehydration, breast-feeding, and growth monitoring for child survival. Success with these initiatives supported the continued movement of health development efforts away from the comprehensive approach of primary health care and toward programs that targeted specific public health priorities. This approach was reinforced by the need to address the HIV/AIDS epidemic. By the 1990s, primary health care had fallen out of favor in many global-health policy circles, and low- and middle-income countries were being encouraged to reduce public sector spending on health and to focus on cost-effectiveness analysis to provide a package of health care measures thought to offer the greatest health benefits.

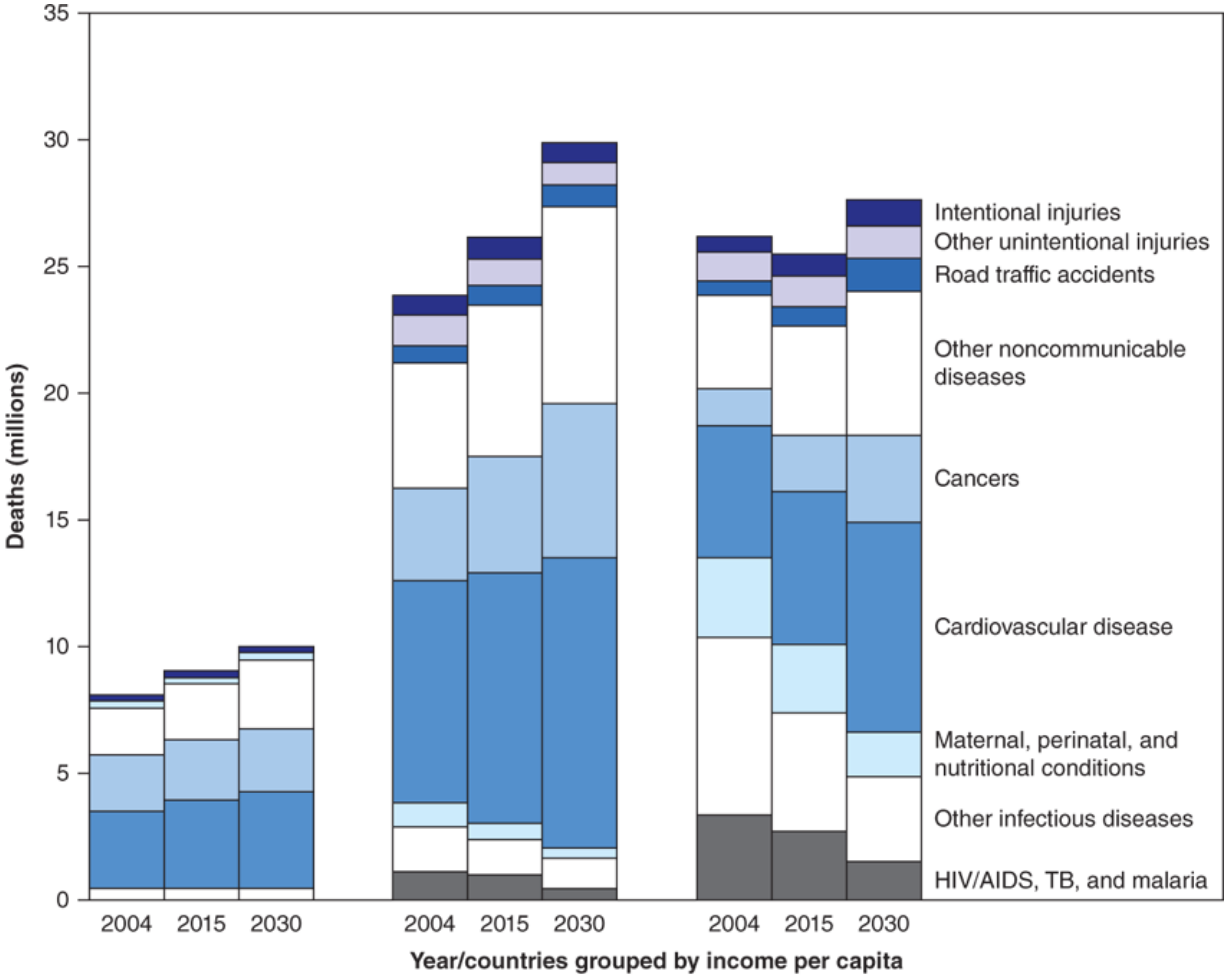
[1]Institute of Medicine. Primary Care: America’s Health in a New Era (1996).

HEALTH CHALLENGES IN LOW- AND MIDDLE-INCOME COUNTRIES

Low- and middle-income countries, defined by a per capita gross national income of <$12,476 (U.S.) per person per year, account for >80% of the world’s population. Average life expectancy in these countries lags far behind that in high-income countries: whereas the average life expectancy at birth for a girl in high-income countries is 76 years, it is only 63 years for a girl in low-income countries. This discrepancy has received growing attention over the past 40 years. Initially, the situation in poor countries was characterized primarily in terms of high fertility and high infant, child, and maternal mortality rates, with most deaths and illnesses attributable to infectious or tropical diseases among remote, largely rural populations. With growing adult (and especially elderly) populations and changing lifestyles linked to global forces of urbanization, a new set of health challenges characterized by chronic diseases, environmental overcrowding, and road traffic injuries has emerged rapidly **(Fig. 462-1)**. The majority of tobacco-related deaths globally now occur in low- and middle-income countries, and the risk of a child’s dying from a road traffic injury in Africa is more than twice that in Europe. Thus, low- and middle-income countries in the twenty-first century face a full spectrum of health challenges—infectious, chronic, and injury-related—at much higher incidences and prevalences than are documented in high-income countries and with many fewer resources to address these challenges.

FIGURE 462-1

Projections of disease burden to 2030 for high-, middle-, and low-income countries (*left*, *center*, and *right*, respectively). TB, tuberculosis. *(Source: World Health Organization: The Global Burden of Disease 2004 Update, 2008.)*

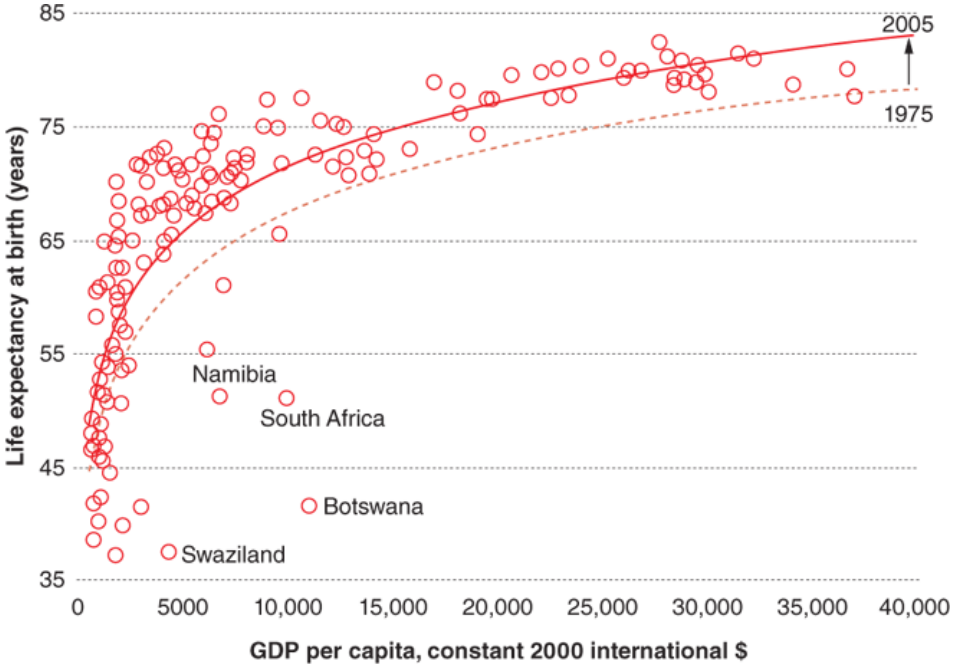


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Addressing these challenges, however, does not mean simply waiting for economic growth. Analysis of the association between wealth and health across countries reveals that, for any given level of wealth, there is a substantial variation in life expectancy at birth that has persisted despite overall global progress in life expectancy during the past 30 years (Fig. 462-2). Health status in low- and middle-income countries varies enormously. Nations such as Cuba and Costa Rica have life expectancies and childhood mortality rates similar to or even better than those in high-income countries; in contrast, countries in sub-Saharan Africa and the former Soviet bloc have at times experienced significant reverses in these health markers in recent decades, particularly in the 1990s.

FIGURE 462-2

Gross domestic product (GDP) per capita and life expectancy at birth in 169 countries, 1975 and 2005. Only outlying countries are named. *(Source: World Health Organization: Primary Health Care: Now More Than Ever. World Health Report 2008.)*

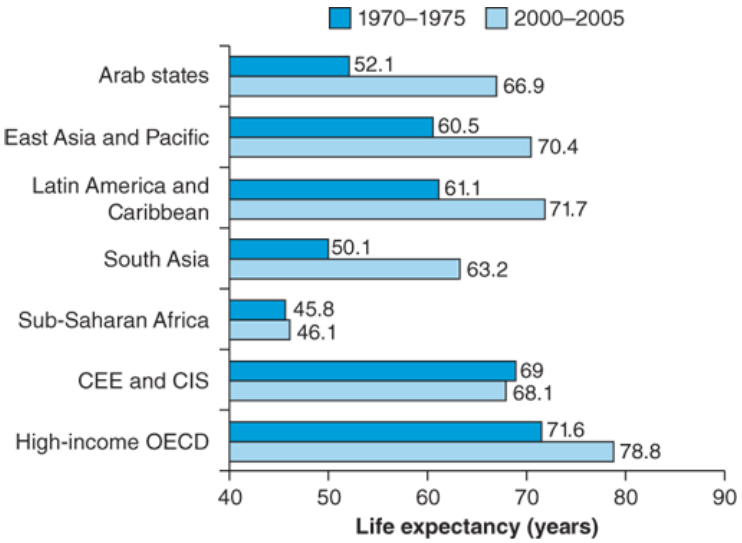


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

As Angus Deaton stated in the World Institute for Development Economics Research annual lecture on September 29, 2006, “People in poor countries are sick not primarily because they are poor but because of other social organizational failures, including health delivery, which are not automatically ameliorated by higher income.” This analysis concurs with classic studies of the array of societal factors explaining good health in poor settings such as Cuba and Kerala State in India in the 1980s. Analyses conducted over the past three decades indeed show that rapid health improvement is possible in very different contexts. That some countries continue to lag far behind can be understood through a comparison of regional differences in progress in terms of life expectancy over this period (Fig. 462-3).

FIGURE 462-3

Regional trends in life expectancy. CEE, Central and Eastern Europe; CIS, the Commonwealth of Independent States; OECD, Organization for Economic Cooperation and Development. *(Source: World Health Organization: Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health. Commission on Social Determinants of Health Final Report, 2008.)*



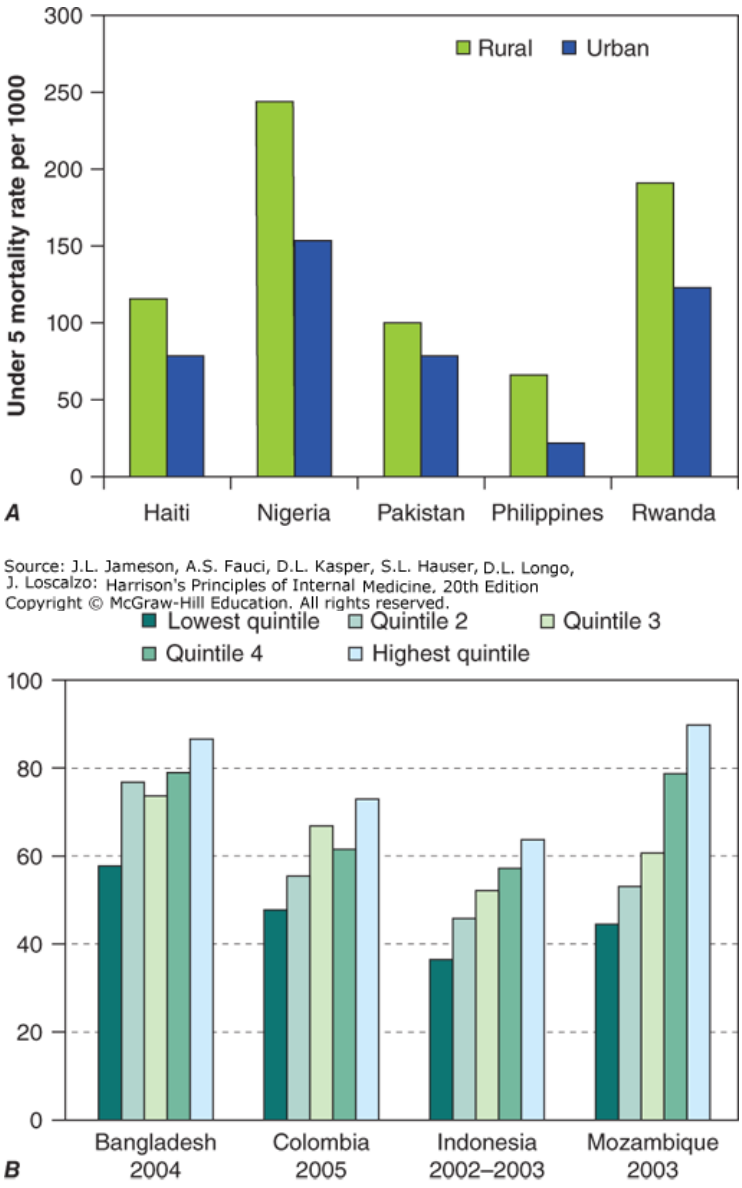
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

As average levels of health vary across regions and countries, so too do they vary within countries (Fig. 462-4). Indeed, disparities within countries are often greater than those between high-income and low-income countries. For example, if low- and middle-income countries could reduce their overall childhood mortality rate to that of the richest one-fifth of their populations, global childhood mortality could be decreased by 40%. Disparities in health are mostly a result of social and economic factors such as daily living conditions, access to resources, and ability to participate in life-affecting decisions. In most countries, the health care sector actually tends to exacerbate health inequalities (the “inverse-care law”); because of neglect and discrimination,

poor and marginalized communities are much less likely to benefit from public health services than those that are better off. Reforming health systems toward people-centered primary care provides an opportunity to reverse these negative trends.

FIGURE 462-4

A. Mortality of children under 5 years old, by place of residence, in five countries. *(Source: Data from the World Health Organization.)* ***B.*** Full basic immunization coverage (%), by income group. *(Source: Primary Health Care: Now More Than Ever. World Health Report 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Health services have failed to make their contribution to reducing these pervasive social inequalities by ensuring universal access to existing, scientifically validated, low-cost interventions such as insecticide-treated bed nets for malaria, taxes on cigarettes, short-course chemotherapy for tuberculosis, antibiotic treatment for pneumonia, dietary modification and secondary prevention measures for high blood pressure and high cholesterol levels, and water treatment and oral rehydration therapy for diarrhea. Despite decades of "essential packages" and "basic" health campaigns, the effective implementation of what is already known to work appears (deceptively) to be difficult.

Recent analyses have begun to focus on "the how" (as opposed to "the what") of health care delivery, exploring why health progress is slow and sluggish despite the abundant availability of proven interventions for health conditions in low- and middle-income countries. Three general categories of reasons are being identified: (1) shortfalls in performance of health systems; (2) stratifying social conditions; and (3) skews in science.

SHORTFALLS IN PERFORMANCE OF HEALTH SYSTEMS

Specific health problems often require the development of specific health interventions (e.g., tuberculosis requires short-course chemotherapy). However, the delivery of different interventions is often facilitated by a common set of resources or functions: money or financing, trained health

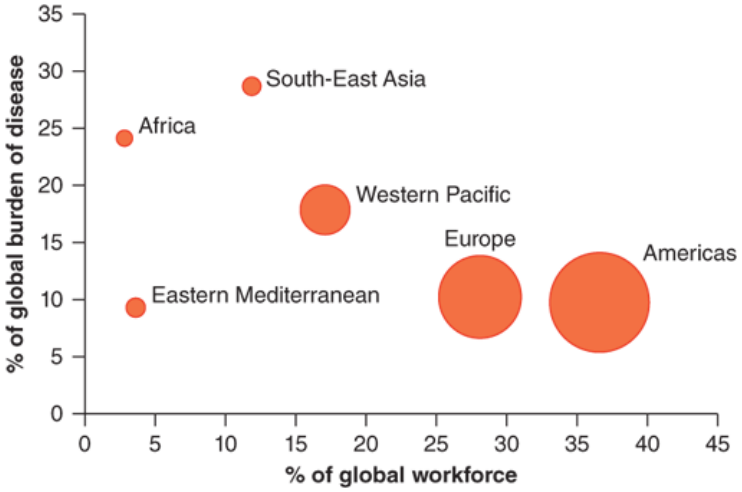
workers, and facilities with reliable supplies fit for multiple purposes. Unfortunately, health systems in most low- and middle-income countries are largely dysfunctional at present.

In the large majority of low- and middle-income countries, the level of public financing for health is woefully insufficient: whereas high-income countries spend, on average, 7% of the gross domestic product on health, middle-income countries spend <4%, and low-income countries <3%. External financing for health through various donor channels grew rapidly in the first decade of the twenty-first century but has grown more slowly in the second decade to its current level of $37 billion. While these funds for health are significant, they represent <2% of total health expenditures in low- and middle-income countries and therefore are neither a sufficient nor a long-term solution to chronic underfinancing. In Africa, 70% of health expenditures come from domestic sources. The predominant form of health care financing—charging patients at the point of service—is the least efficient and the most inequitable, tipping millions of households into poverty annually.

Health workers, who represent another critical resource, are often inadequately trained and supported in their work. Recent estimates indicate a shortage of >7 million health workers, constituting a crisis that is greatly exacerbated by the migration of health workers from low- and middle-income countries to high-income countries. Sub-Saharan Africa carries 24% of the global disease burden but has only 3% of the health workforce **(Fig. 462-5)**. The International Organization for Migration estimated in 2006 that there were more Ethiopian physicians practicing in Chicago than in Ethiopia itself.

FIGURE 462-5

Global burden of disease and health workforce. *(Source: World Health Organization: Working Together for Health, 2006.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Critical diagnostics and drugs often do not reach patients in need because of supply-chain failures. Moreover, facilities fail to provide safe care: new evidence suggests much higher rates of adverse events among hospitalized patients in low- and middle-income countries than in high-income countries. Weak government planning, regulatory, monitoring, and evaluation capacities are associated with rampant, unregulated commercialization of health services and chaotic fragmentation of these services as donors “push” their respective priority programs. With such fragile foundations, it is not surprising that low-cost, affordable, validated interventions are not reaching those who need them.

STRATIFYING SOCIAL CONDITIONS

Health care delivery systems do not exist in a vacuum but rather are embedded in a complex of social and economic forces that often stratify opportunities for health unfairly. Most worrisome are the pervasive forces of social inequality that serve to marginalize populations with disproportionately large health needs (e.g., the urban poor; illiterate mothers). Why should a poor slum dweller with no income be expected to come up with the money for a bus fare needed to travel to a clinic to learn the results of a sputum test for tuberculosis? How can a mother living in a remote rural village and caring for an infant with febrile convulsions find the means to get her child to appropriate care? Shaky or nonexistent social security systems, dangerous work environments, isolated communities with little or no infrastructure, and systematic discrimination against minorities are among the myriad forces with which efforts for more equitable health care delivery must contend.

SKEWS IN SCIENCE

While science has yielded enormous breakthroughs in health in high-income countries, with some spillover to low- and middle-income countries, many important health problems continue to affect primarily low- and middle-income countries whose research and development investments are deplorably inadequate. The past decade has seen growing efforts to right this imbalance with research and development investment for new drugs, vaccines, and diagnostics that effectively cater to the specific health needs of populations in low- and middle-income countries. For example, the Medicines for Malaria Venture has revitalized a previously “dry” pipeline for new malaria drugs. This is but one of many such efforts, but much more needs to be done.

As discussed above, the primary constraint on better health in low- and middle-income countries is related less to the availability of health technologies and more to their effective delivery. Underlying these systems and social challenges to greater equity in health is a major bias regarding what constitutes legitimate "science" to improve health equity. The lion's share of health research financing is channeled toward the development of new technologies—drugs, vaccines, and diagnostics; in contrast, virtually no resources are directed toward research on how health care delivery systems can become more reliable and overcome adverse social conditions. The complexity of systems and social context is such that this issue of delivery requires an enormous investment in terms not only of money but also of scientific rigor, with the development of new research methods and measures and the attainment of greater legitimacy in the mainstream scientific establishment.

These common challenges to low- and middle-income countries partly explain the resurgence of interest in the primary health care approach and the emergence of a global movement toward universal health coverage, now enshrined as one of the Sustainable Development Goal targets adopted in Agenda 2030 by all countries at the United Nations in September 2015. In some countries (mostly middle-income), significant progress has been made in expanding coverage by health systems based on primary care and even in improving indicators of population health. More countries are embarking on the creation of primary care services despite the challenges that exist, particularly in low-income countries. Even when these challenges are acknowledged, there are many reasons for optimism that low- and middle-income countries can accelerate progress in building primary care as a key vehicle toward achieving universal health coverage.

PRIMARY HEALTH CARE IN THE TWENTY-FIRST CENTURY

The new millennium has seen a resurgence of interest in primary health care as a means of addressing global health challenges. This interest has been driven by many of the same issues that led to the Declaration of Alma Ata: rapidly increasing disparities in health between and within countries, spiraling costs of health care at a time when many people lack quality care, dissatisfaction of communities with the care they are able to access, and failure to address changes in health threats, especially noncommunicable disease epidemics. These challenges require a comprehensive approach and strong health systems with effective primary care. Global health development agencies have recognized that sustaining gains in public health priorities such as HIV/AIDS requires not only robust health systems but also the tackling of social and economic factors related to disease incidence and progression. Weak health systems have proved a major obstacle to delivering new technologies, such as antiretroviral therapy, to all who need them. Changing disease patterns have led to a demand for health systems that can treat people as individuals whether or not they present to a health facility with the public health "priority" (e.g., HIV/AIDS or tuberculosis) to which that facility is targeted. We discuss experiences in low- and middle-income countries in relation to primary care in greater detail below. First, we consider the features of primary health care and primary care as currently understood.

REVITALIZATION OF PRIMARY HEALTH CARE

At the 2009 World Health Assembly (an annual meeting of all countries to discuss the work of the World Health Organization [WHO]), a resolution was passed reaffirming the principles of the Declaration of Alma Ata and the need for national health systems to be based on primary health care. This resolution did not suggest that nothing had changed in the intervening 30 years since the declaration, nor did it dispute that its prescription needed reframing in light of changing public health needs. The 2008 WHO World Health Report describes how a primary health care approach is necessary "now more than ever" to address global health priorities, especially in terms of disparities and new health challenges. As discussed below, this report highlights four broad areas in which reform is required **(Fig. 462-6)**. One of these areas—the need to organize health care so that it places the needs of people first—essentially relates to the necessity for strong primary care in health systems and what this requirement entails. The other three areas also relate to primary care. All four areas require action to move health systems in a direction that will reduce disparities and increase the satisfaction of those they serve. The World Health Report's recommendations present a vision of primary health care that is based on the principles of Alma Ata, but that differs from many attempts to implement primary health care in the 1970s and 1980s.

FIGURE 462-6

The four reforms of primary health care renewal. *(Source: World Health Organization: Primary Health Care: Now More Than Ever. World Health Report 2008.)*



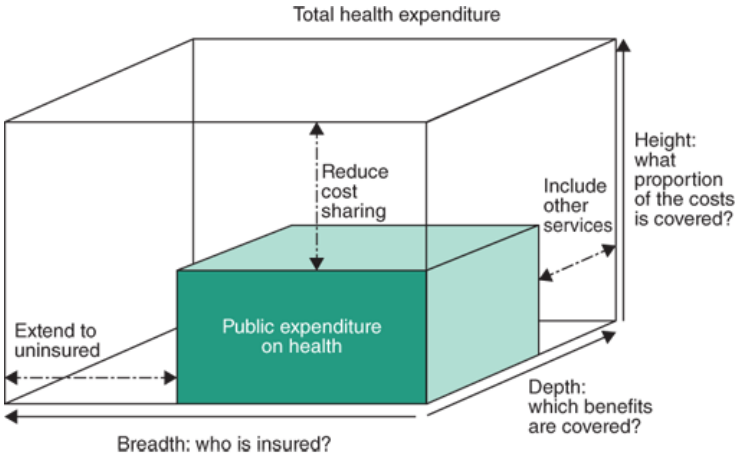
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Universal Coverage Reforms to Improve Health Equity

Despite progress in many countries, most people in the world can receive health care services only if they can pay at the point of service. Disparities in health are caused not only by a lack of access to necessary health services but also by the impact of expenditure on health. More than 100 million people are driven into poverty each year by health care costs, with countless others deterred from accessing services at all. Moving toward prepayment financing systems for universal coverage, which ensure access to a comprehensive package of services according to need without precipitating economic ruin, has therefore emerged as a major priority in low- and middle-income countries. Increasing coverage of health services can be considered in terms of three axes: the proportion of the population covered, the range of services underwritten, and the percentage of costs paid (Fig. 462-7). Moving toward universal health coverage requires ensuring the availability of health care services to all, eliminating barriers to access, and organizing pooled financing mechanisms, such as taxation or insurance, to remove user fees at the point of service. It also requires measures beyond financing, including expansion of health services in poorly served areas, improvement in the quality of services provided to marginalized communities, and increased coverage of other social services that significantly affect health (e.g., education).

FIGURE 462-7

Three ways of moving toward universal coverage.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Service Delivery Reforms to Make Health Systems People-Centered

Health systems have often been organized around the needs of those who provide health care services, such as clinicians and policy makers. The result is a centralization of services or the provision of vertical programs that target single diseases. The principles of primary health care, including the development of primary care, reorient care around the needs of the people to whom services cater. This “people-centered” approach aims to provide health care that is both more effective and appropriate.

The increase in noncommunicable diseases in low- and middle-income countries offers a further stimulus for urgent reform of service delivery to improve chronic disease care. As discussed above, large numbers of people currently fail to receive relatively low-cost interventions that have reduced the incidence of these diseases in high-income countries. Delivery of these interventions requires health systems that can address multiple problems and manage people over a long period within their own communities, yet many low- and middle-income countries are only now starting to adapt and build primary care services that can address noncommunicable diseases and communicable diseases requiring chronic care. Even some countries (e.g., Iran) that have had significant success in reducing communicable diseases and improving child survival have been slow to adapt their health systems to rapidly accelerating noncommunicable disease epidemics.

People-centered care requires a safe, comprehensive, and integrated response to the needs of those presenting to health systems, with treatment at the first point of contact or referral to appropriate services. Because no discrete boundary separates people’s needs for health promotion, curative interventions, and rehabilitation services across different diseases, primary care services must address all presenting problems in a unified way. Meeting people’s needs also involves improved communication between patients and their clinicians, who must take the time to understand the impact of the patients’ social context on the problems they present with. This enhanced understanding is made possible by improvements in the continuity of care so that responsibility transcends the limited time people spend in health care facilities. Primary care plays a vital role in navigating people through the health system; when people are referred elsewhere for services, primary care providers must monitor the resulting consultations and perform follow-up. All too often, people do not receive the benefit of complex interventions undertaken in hospitals because they lose contact with the health care system once discharged. Comprehensiveness and continuity of care are best achieved by ensuring that people have an ongoing personal relationship with a care team.

Public Policy Reforms to Promote and Protect the Health of Communities

Public policies in sectors other than health care are essential to reduce disparities in health and to make progress toward global public health targets. The 2008 final report of the WHO Commission on Social Determinants of Health provides an exhaustive review of the multisectoral policies required to address health inequities at the local, national, and global levels. Advances against major challenges such as HIV/AIDS, tuberculosis, emerging pandemics, cardiovascular disease, cancers, and injuries require effective collaboration with sectors such as transport, housing, labor, agriculture, urban planning, trade, and energy. While tobacco control provides a striking example of what is possible if different sectors work together toward health goals, the lack of implementation of many evidence-based tobacco control measures in most countries just as clearly illustrates the difficulties encountered in such multisectoral work and the unrealized potential of public policies to improve health. At the local level, primary care services can help enact health-promoting public policies in other sectors.

Leadership Reforms to Make Health Authorities More Reliable

The Declaration of Alma Ata emphasized the importance of participation by people in their own health care. In fact, participation is important at all levels of decision-making. Contemporary health challenges require new models of leadership that acknowledge the role of government in reducing disparities in health but that also recognize the many types of organizations that provide health care services. Governments need to guide and negotiate among these different groups, including nongovernmental organizations (NGOs) and the private sector, and to provide strong regulation where necessary. This difficult task requires a massive reinvestment in leadership and governance capacity, especially if action by different sectors is to be effectively implemented. Moreover, disadvantaged groups and other actors are increasingly expecting that their voices and health needs will be included in the decision-making process. The complex landscape for leadership at the national level is mirrored in many ways at the international or global level. The transnational character of health and the increasing interdependence of countries with respect to outbreak diseases, climate change, security, migration, and agriculture place a premium on more effective global health governance.

EXPERIENCES WITH PRIMARY CARE IN LOW- AND MIDDLE-INCOME COUNTRIES

Aspects of the primary health care approach described above, with an emphasis on primary care services, have been implemented to varying degrees in many low- and middle-income countries over the past half-century. As discussed above, some of these experiences inspired and informed the Declaration of Alma Ata, which itself led many more countries to attempt to implement primary health care. This section describes the experiences of a selection of low- and middle-income countries in improving primary care services that have enhanced the health of their populations.

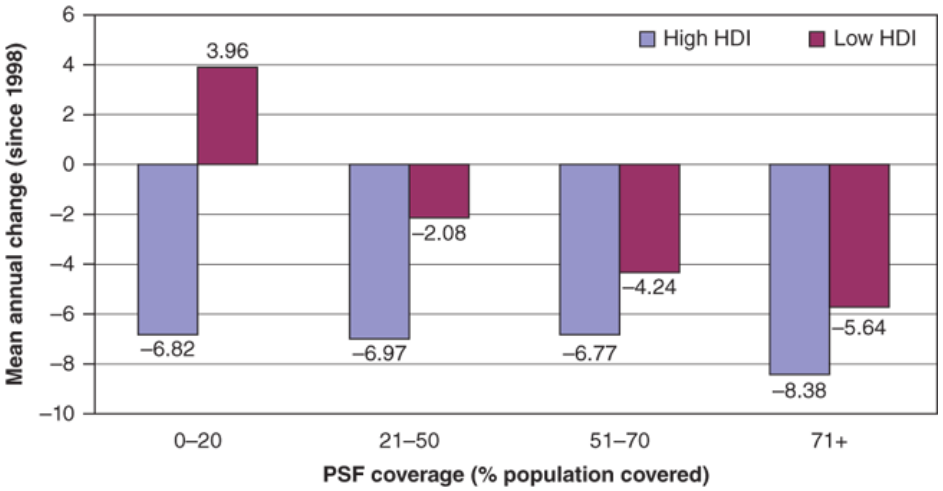
Before Alma Ata, few countries had attempted to develop primary care on a national level. Rather, most focused on expanding primary care services to specific communities (often rural villages), making use of community volunteers to compensate for the absence of facility-based care. In contrast, in the post–World War II period, China invested in primary care on a national scale, and life expectancy doubled within roughly 20 years. The Chinese expansion of primary care services included a massive investment in infrastructure for public health (e.g., water and sanitation systems) linked to innovative use of community health workers. These “barefoot doctors” lived in and expanded care to rural villages. They received a basic level of training that enabled

them to provide immunizations, maternal care, and basic medical interventions, including the use of antibiotics. Through the work of the barefoot doctors, China brought low-cost universal basic health care coverage to its entire population, most of which had previously had no access to these services.

In 1982, the Rockefeller Foundation convened a conference to review the experiences of China along with those of Costa Rica, Sri Lanka, and the state of Kerala in India. In all of these locations, good health care at low cost appeared to have been achieved. Despite lower levels of economic development and health spending, all of these jurisdictions, along with Cuba, had health indicators approaching—or in some cases exceeding—those of developed countries. Analysis of these experiences revealed a common emphasis on primary care services, with expansion of care to the entire population free of charge or at low cost, combined with community participation in decision-making about health services and coordinated work in different sectors (especially education) toward health goals. During the more than three decades since the Rockefeller meeting, some of these countries have built on this progress, while others have experienced setbacks. Recent experiences in developing primary care services show that the same combination of features is necessary for success. For example, Brazil—a large country with a dispersed population—has made major strides in increasing the availability of health care in the past quarter century. In this millennium, the Brazilian Family Health Program has expanded progressively across the country, with almost all areas now covered. This program provides communities with free access to primary care teams made up of primary care physicians, community health workers, nurses, dentists, obstetricians, and pediatricians. These teams are responsible for the provision of primary care to all people in a specified geographic area—not only those who access health clinics. Moreover, individual community health workers are responsible for a named list of people within the area covered by the primary care team. Problems with access to health care persist in Brazil, especially in isolated areas and urban slums. However, solid evidence indicates that the Family Health Program has already contributed to impressive gains in population health, particularly in terms of childhood mortality and health inequities. In fact, this program has already had an especially marked impact on childhood mortality reduction in less developed areas **(Fig. 462-8)**.

FIGURE 462-8

Improvements in childhood mortality following the Family Health Program in Brazil. HDI, Human Development Index; PSF, Program Saúde da Família (Family Health Program). *(Source: Ministry of Health, Brazil.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Chile has also built on its existing primary care services in the past decade, aiming to improve the quality of care and the extent of coverage in remote areas, above all for disadvantaged populations. This effort has been made in concert with measures aimed at reducing social inequalities and fostering development, including social welfare benefits for families and disadvantaged groups and increased access to early-childhood educational facilities. As in Brazil, these steps have improved maternal and child health and have reduced health inequities. In addition to directly enhancing primary care services, Brazil and Chile have instituted measures to increase both the accountability of health providers and the participation of communities in decision-making. In Brazil, national and regional health assemblies with high levels of public participation are integral parts of the health policy-making process. Chile has instituted a patient's charter that explicitly specifies the rights of patients in terms of the range of services to which they are entitled.

Other countries that have made recent progress with primary health care include Bangladesh, once one of the poorest countries in the world and still relatively poor. Since achieving its independence from Pakistan in 1971, Bangladesh has seen a dramatic increase in life expectancy, and childhood mortality rates are now lower than those in neighboring nations such as India and Pakistan. The expansion of access to primary care services has played a major role in these achievements. This progress has been spearheaded by a vibrant NGO community that has focused its attention on improving the lives and livelihoods of poor women and their families through innovative and integrated microcredit, education, and primary care programs.

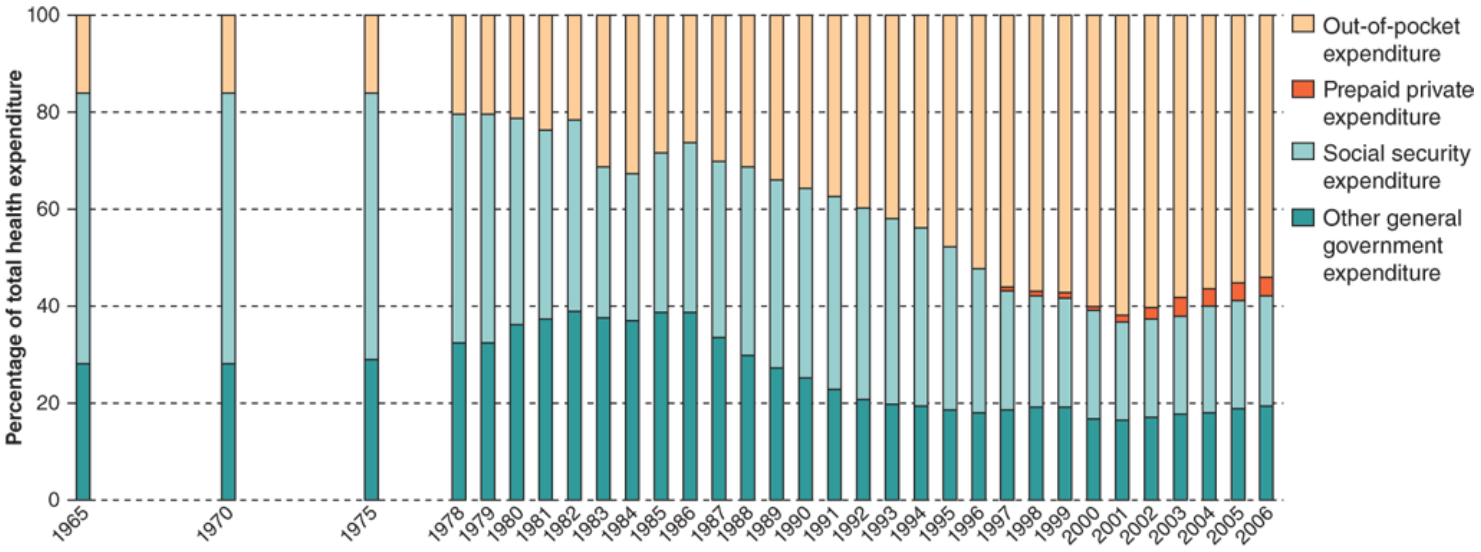
The above examples, along with others from the past 30 years in countries such as Thailand, Malaysia, Portugal, and Oman, illustrate how the implementation of a primary health care approach, with a greater emphasis on primary care, has led to better access to health care services—a trend that has not been seen in many other low- and middle-income countries. This trend, in turn, has contributed to improvements in population health and

reductions in health inequities. However, as these nations have progressed, other countries have shown how previous gains in primary care can easily be eroded. In sub-Saharan Africa, undermining of primary care services contributed to catastrophic reversals in health outcomes catalyzed by the HIV/AIDS epidemic. Countries such as Botswana and Zimbabwe implemented primary health care strategies in the 1980s, increasing access to care and making impressive gains in child health. Both countries were severely affected by HIV/AIDS, with pronounced decreases in life expectancy. However, Zimbabwe has also seen political turmoil, a decline of health and other social services, and the flight of health personnel, whereas Botswana has maintained primary care services to a greater extent and has managed to organize widespread access to antiretroviral therapy for people living with HIV/AIDS. Zimbabwe's health situation has therefore become more desperate than that in Botswana.

China provides a particularly striking example of how changes in health policy relevant to the organization of health systems **(Fig. 462-9)** can have rapid, far-reaching consequences for population health. Even as the 1982 Rockefeller conference was celebrating China's achievements in primary care, its health system was unraveling. The decision to open up the economy in the early 1980s led to rapid privatization of the health sector and the breakdown of universal health coverage. As a result, by the end of the 1980s, most people, especially the poorer segments of the population, were paying directly out of pocket for health care, and almost no Chinese had insurance—a dramatic transformation. The "barefoot doctor" schemes collapsed, and the population either turned to care paid for at hospitals or simply became unable to access care. This undermining of access to primary care services in the Chinese system and the resulting increase in impoverishment due to illness contributed to the stagnation of progress in health in China at the same time that incomes in that country increased at an unprecedented rate. Reversals in primary care have meant that China now increasingly faces health care issues similar to those faced by India, although the country has more recently implemented measures to restore universal health coverage, with significant success. In both countries, rapid economic growth has been linked to lifestyle changes and noncommunicable disease epidemics. The health care systems of the two nations share two negative features that are common when primary care is weak: a disproportionate focus on specialty services provided in hospitals and unregulated commercialization of health services. China and India have both seen expansion of private hospital services that cater to middle-class and urban populations who can afford care; at the same time, hundreds of millions of people in rural areas now struggle to access the most basic services. Even in the former groups, a lack of primary care services has been associated with late presentation with illness and with insufficient investment in primary prevention approaches. This neglect of prevention poses a risk of large-scale epidemics of cardiovascular disease, which could endanger continued economic growth. In addition, the health systems of both countries now depend for the majority of their funding on out-of-pocket payments by people when they use services. Thus substantial proportions of the population must sacrifice other essential goods as a result of health expenditure and may even be driven into poverty by this cost. The commercial nature of health services with inadequate or no regulation has also led to the proliferation of charlatan providers, inappropriate care, and pressure for people to pay for expensive and sometimes unnecessary care. Commercial providers have limited incentives to use interventions (including public health measures) that cannot be charged for or that are what the person who is paying can afford.

FIGURE 462-9

Changes in source of health expenditure in China over the past 40 years. *(Source: World Health Organization: Primary Health Care: Now More Than Ever. World Health Report 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Faced with these problems, China and India have implemented measures to strengthen primary health care. China has increased government funding of health care, has taken steps toward restoring health insurance, and has enacted a target of universal access to primary care services. India has similarly mobilized funding to greatly expand primary care services in rural areas and is now duplicating this process in urban settings. Both countries are increasingly using public resources from their growing economies to fund primary care services.

These encouraging trends are illustrative of new opportunities to implement a primary health care approach and strengthen primary care services in low- and middle-income countries. Brazil, India, China, and Chile are being joined by many other low- and middle-income countries, including Indonesia, Mexico, the Philippines, Turkey, Rwanda, Ethiopia, South Africa, and Ghana, in ambitious initiatives mobilizing new resources to move toward universal health coverage—the provision of quality health services in a timely manner at affordable cost.

OPPORTUNITIES TO BUILD PRIMARY CARE IN LOW- AND MIDDLE-INCOME COUNTRIES

Global public health targets will not be met unless health systems are significantly strengthened. More money is currently being spent on health than ever before. In 2013, global health spending totaled $7.8 trillion (U.S.)—more than double the amount spent a decade earlier. Although most expenditure occurs in high-income countries, spending in many emerging middle-income countries has rapidly accelerated, as has the allocation of monies for this purpose by both governments in and donors to low-income countries. These twin trends—greater emphasis on building health systems based on primary care and allotment of more money for health care—provide opportunities to address many of the challenges discussed above in low- and middle-income countries.

Accelerating progress requires a better understanding of how global health initiatives can more effectively facilitate the development of primary care in low-income countries. A review by the WHO Maximizing Positive Synergies Collaborative Group looked at programs funded by the Global Fund to Fight AIDS, Tuberculosis and Malaria; the Global Alliance for Vaccines and Immunisation (GAVI); the U.S. President's Emergency Plan for AIDS Relief (PEPFAR); and the World Bank (HIV/AIDS). This group found that global health initiatives had improved access to and quality of the targeted health services and had led to better information systems and more adequate financing. The review also identified the need for better alignment of global health initiatives with other national health priorities and systematic exploitation of potential synergies. If global health initiatives implement programs that work in tandem with other components of national health systems without undermining staffing and procurement of supplies, they have the potential to contribute substantially to the capacity of health systems to provide comprehensive primary care services.

Even in the aftermath of the global financial crisis, global health initiatives continued to draw significant funding. In 2009, for example, U.S. President Barack Obama announced increasing development assistance from the United States for global health, earmarking $63 billion over the period 2009–2014 for a Global Health Initiative. New funding is also promised through a range of other initiatives focusing particularly on maternal and child health in low-income countries. The general trend is to coordinate this funding in order to reduce fragmentation of national health systems and to concentrate more on strengthening these systems. Comprehensive primary care in low-income countries must inevitably deal with the rapid emergence of chronic diseases and the growing prominence of injury-related health problems; thus international health development assistance must become more responsive to these needs. More recent political currents that threaten global health funding from traditional sources only underscore the need to take a more comprehensive and integrated health systems approach to the use of these funds.

Beyond funding for health services, other opportunities exist. Increased social participation in health systems can help build primary care services. In many countries, political pressure from community advocates for more holistic and accountable care as well as entrepreneurial initiatives to scale up community-based services through NGOs have accelerated progress in primary care without major increases in funding. Participation of the population in the provision of health care services and in relevant decision-making often drives services to cater to people's needs as a whole rather than to narrow public health priorities.

Participation and innovation can help address critical issues related to the health workforce in low- and middle-income countries by establishing effective people-centered primary care services. Many primary care services do not need to be delivered by a physician or a nurse. Multidisciplinary teams can include paid community workers who have access to a physician if necessary but who can provide a range of health services on their own. In Ethiopia, more than 38,000 community health workers have been trained and deployed to improve access to primary care services, and there is increasing evidence that this measure is contributing to better health outcomes. In India, more than 600,000 community health advocates have been recruited as part of expanded rural primary care services. In Niger, the deployment of community health workers to deliver essential child health interventions (as a component of integrated community case management) has had impressive results in reducing childhood mortality and decreasing disparities. After the Declaration of Alma Ata, experiences with community health workers were mixed, with particular problems about levels of training and lack of payment. Current endeavors are not immune from these concerns. However, with access to physician support and the deployment of teams, some of these concerns may be addressed. Growing evidence from many countries indicates that shifting appropriate tasks to primary care workers who have had shorter, less expensive training than physicians will be essential to address the human resources crisis.

Finally, recent improvements in information and communication technologies, particularly mobile phone and internet systems, have created the potential for systematic implementation of e-health, telemedicine, and improved health data initiatives in low- and middle-income countries. These developments raise the tantalizing possibility that health systems in these countries, which have long lagged behind those in high-income countries but are less encumbered by legacy systems that have proved hard to modernize in many settings, could leapfrog their wealthier counterparts in exploiting these technologies. Although the challenges posed by poor or absent infrastructure and investment in many low- and middle-income countries cannot be underestimated and will need to be addressed to make this possibility a reality, the rapid rollout of mobile networks and their use for health and other social services in many low-income countries where access to fixed telephone lines was previously very limited offer great promise in building primary care services in low- and middle-income countries.

CONCLUSION

As concern continues to mount about glaring inequities in global health, there is a growing commitment to redress these egregious shortfalls, as exemplified by the central place of equity in the United Nations' Sustainable Development Goals adopted in 2015, including a specific target on the achievement of universal health coverage in all countries by 2030. This commitment begins first and foremost with a clear vision of the fundamental importance of health in all countries, regardless of income. The values of health and health equity are shared across all borders, and primary health care provides a framework for their effective translation across all contexts.

The translation of these fundamental values has its roots in four types of reform that reflect the distinct but interlinked challenges of (re)orienting a society's resources on the basis of its citizens' health needs: (1) organizing health care services around the needs of people and communities; (2) harnessing services and sectors beyond health care to promote and protect health more effectively; (3) establishing sustainable and equitable financing mechanisms for universal health coverage; and (4) investing in effective leadership of the whole of society. This common primary health care agenda highlights the striking similarity, despite enormous differences in context, in the nature and direction of the reforms that national health systems must undertake to promote greater equity in health. This shared agenda is complemented by the growing reality of global health interconnectedness due, for example, to shared microbial threats, bridging of ethnolinguistic diversity, flows in migrant health workers, and mobilization of global funds to support the neediest populations. Embracing solidarity in global health while strengthening health systems through a primary health care approach is fundamental to sustained progress in global health.

FURTHER READING

Aquino R et al: Impact of the family health program on infant mortality in Brazilian municipalities. Am J Public Health 99:87, 2009. [PubMed: 19008516]

Commission on Social Determinants of Health: *Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health: Commission on Social Determinants of Health Final Report*. Geneva, World Health Organization, 2008.

Kruk ME et al: The contribution of primary care to health and health systems in low- and middle-income countries: A critical review of major primary care initiatives. Soc Sci Med 70:904, 2010. [PubMed: 20089341]

Li X et al: The primary health-care system in China. Lancet 390:2584, 2017. [PubMed: 29231837]

Macinko J et al: The impact of primary healthcare on population health in low- and middle-income countries. J Ambul Care Manage 32:150, 2009. [PubMed: 19305227]

Patel V et al: Assuring health coverage for all in India. Lancet 386:2422, 2015. [PubMed: 26700532]

Rasanathan K et al: Primary health care and the social determinants of health: Essential and complementary approaches for reducing inequities in health. J Epidemiol Community Health 65:656, 2011. [PubMed: 19933684]

Starfield B et al: Contribution of primary care to health systems and health. Milbank Q 83:457, 2005. [PubMed: 16202000]

Tangcharoensathien V et al: Primary Health Care: Now More Than Ever. World Health Report 2008. Geneva, World Health Organization, 2008.

Tangcharoensathien V et al: Health systems development in Thailand: A solid platform for successful implementation of universal health coverage. Lancet pii:S0140-6736(18)30198-3, 2018.

Chapter 463: The Biology of Aging

Rafael de Cabo; David G. Le Couteur

THE IMPACT OF AGING ON MEDICINE

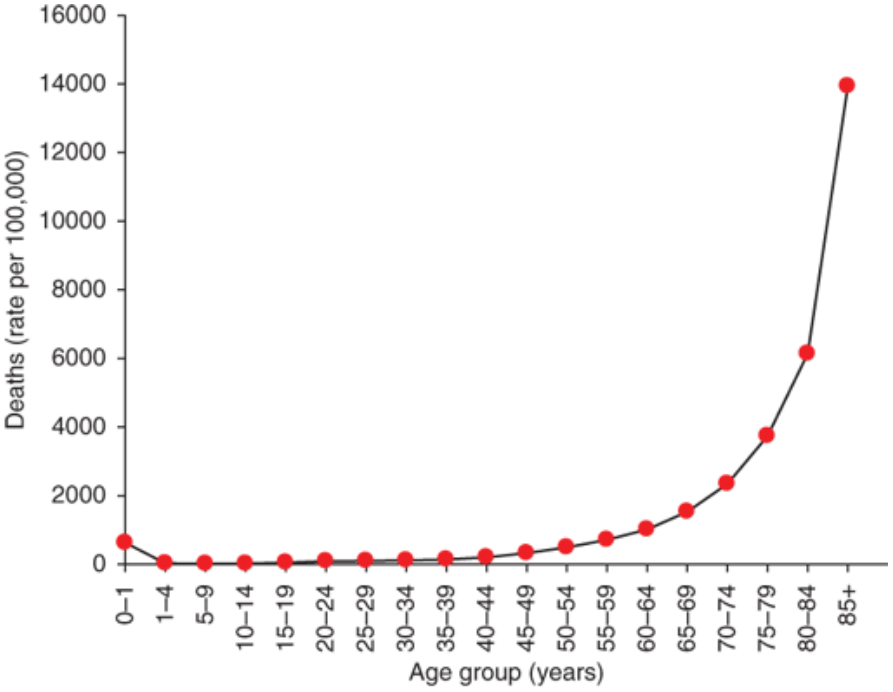
Aging and old age are among the most significant challenges facing medicine this century. The aging process is the major risk factor underlying disease and disability in developed nations, while older people respond differently to therapies developed for younger adults (usually with less effectiveness and more adverse reactions). Modern medicine and healthier lifestyles have increased the likelihood that younger adults will now achieve old age. However, this has led to rapidly increasing numbers of older people, often encumbered with age-related disorders that are predicted to overwhelm health care systems. Improved health in old age and further extension of human health span are now likely to be generated primarily from increased understanding of the biology of aging, age-related susceptibility to disease, and modifiable factors that influence the aging process.

Definitions of Aging

Aging is easy to recognize but difficult to define. Most definitions of aging indicate that it is a progressive process associated with declines in structure and function, impaired maintenance and repair systems, increased susceptibility to disease and death, and reduced reproductive capacity. There are both statistical and phenotypic components to aging. As recognized by Gompertz in the nineteenth century, aging in humans is associated with an exponential risk of mortality with time **(Fig. 463-1)**, although it is now realized that this plateaus in extreme old age because of healthy survivor bias. The phenotypic components of aging include structural and functional changes that are separated, somewhat artificially, into either primary aging changes (e.g., sarcopenia, grey hair, oxidative stress, increased peripheral vascular resistance) or age-related disease (e.g., dementia, osteoporosis, arthritis, insulin resistance, hypertension).

Figure 463-1

The rates of death in the United States (2010) showing exponential increase in mortality risk with chronological age.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

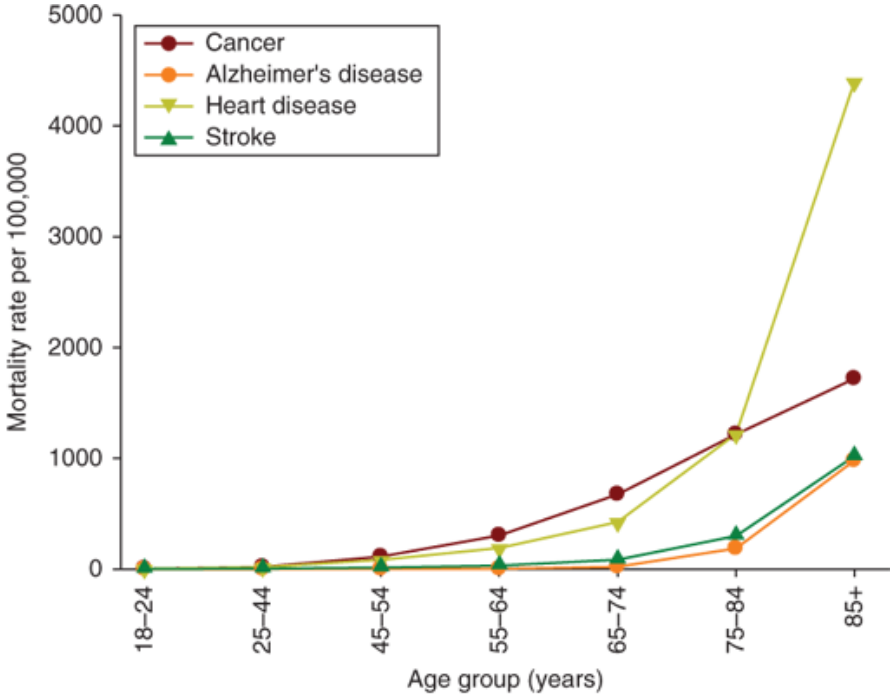
Definitions of aging rarely acknowledge the possibility that some of those biological and functional changes with aging might be adaptive or even reflect improvement and gain. Nor do they emphasize the impact of aging on responses to medical treatments. Old age is associated with increased vulnerability to many perturbations, including therapeutic interventions. This is a critical issue for clinicians—aging would be a less difficult problem if our disease-specific therapies retained their balance of risk to benefit into old age.

Aging and Disease Susceptibility

Old age is the major independent risk factor for chronic diseases (and associated mortality) that are most prevalent in developed countries such as cardiovascular disease, cancers, and neurodegenerative disorders (Fig. 463-2). Consequently, older people have multiple comorbidities, usually in the range of 5–10 illnesses per person.

Figure 463-2

The rates of most common chronic diseases and related mortality increase with old age. *(Data from USA 2008-2010 CDC.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Disease in older people is typically multifactorial with a strong component related to the underlying aging process. For example, in younger patients with dementia, Alzheimer's disease is a single disorder confirmed by examining brain tissue for plaques and tangles containing amyloid and tau proteins. However, the vast majority of people with dementia are elderly, and here the association between typical Alzheimer's neuropathology and dementia becomes less definitive. In the oldest-old, the prevalence of Alzheimer-type brain pathology is similar in people with and without clinical features of dementia. On the other hand, brains of older people with dementia usually show mixed pathology with evidence of Alzheimer's pathology along with features of other dementias such as vascular lesions, Lewy bodies, and non-Alzheimer's tauopathy. Many typical aging changes, such as microvascular dysfunction, oxidative injury, and mitochondrial impairment underlie many of the pathological changes.

The Longevity Dividend

"Compression of morbidity" refers to the concept that the burden of lifetime illness might be compressed by medical interventions into a shorter period before death without necessarily increasing longevity. However, continuing development of successful therapeutic and preventive interventions focusing on individual diseases is less effective in older people because of multiple comorbidities, complications of overtreatment, and competing causes of death. Therefore, it has been proposed that further gains in health span and life expectancy will be achieved by a single intervention that delays aging and age-related disease susceptibility, rather than multiple treatments each targeting different individual age-related illnesses. This is called the "longevity dividend" and is driving an explosion of research into aging biology and, more importantly, interventions—genetic, pharmaceutical, and nutritional—that influence the rate of aging and delay age-related disease.

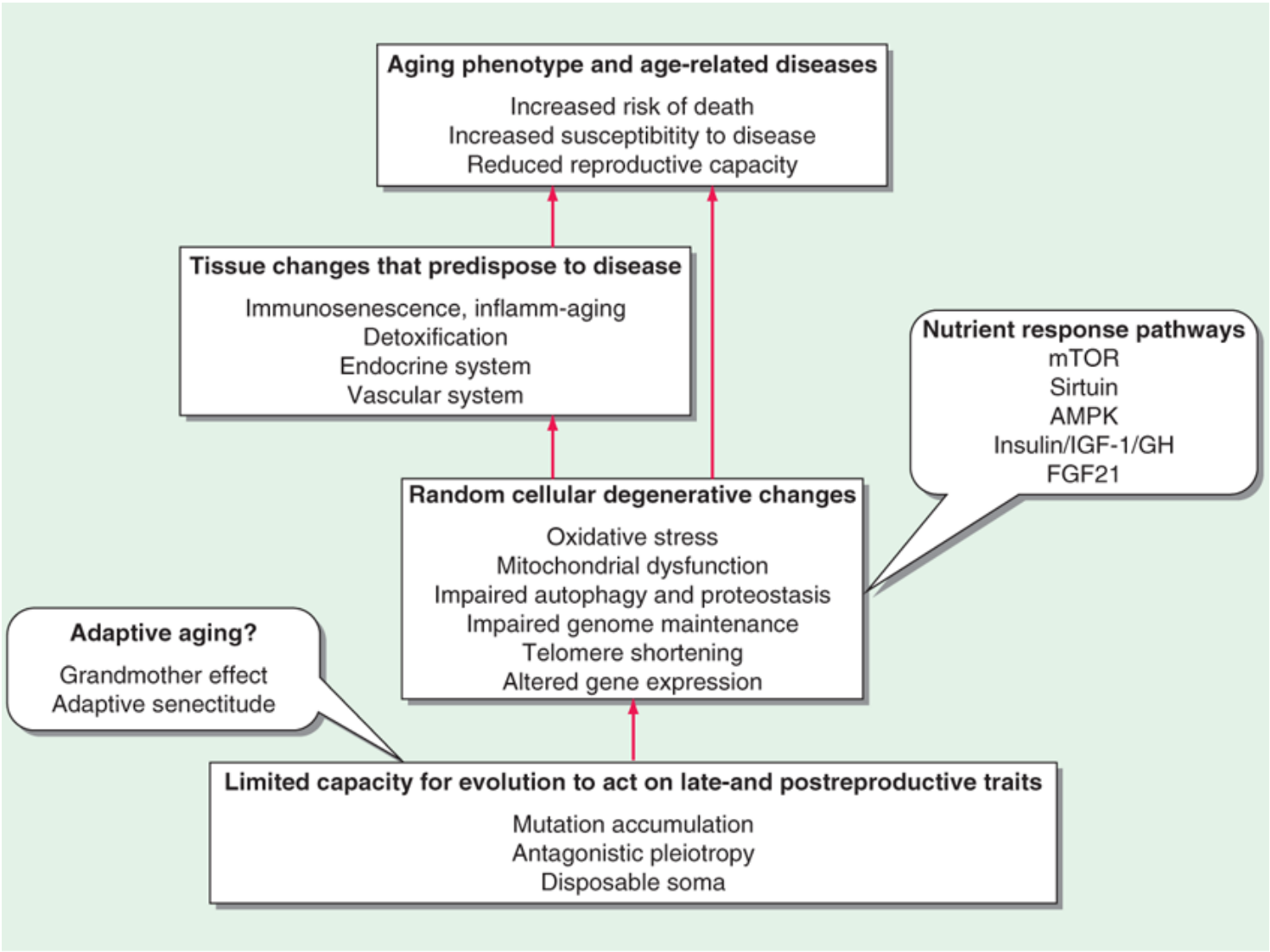
EVOLUTIONARY MECHANISMS FOR AGING

At the most basic level, living things have only two approaches to maintain their existence: immortality or reproduction. In a changing environment, reproduction combined with a finite life span has proved to be the successful strategy. Of course, finite life span is not the same as aging—although aging, by definition, contributes to a finite life span.

Many evolutionary theories related to aging are linked by their attempts to explain this interaction between reproduction and longevity (Fig. 463-3). Most mainstream aging theories stem from the fact that evolution is driven by early reproductive success, whereas there is minimal selection pressure for late-life reproduction or postreproductive survival. Aging is seen as the random degeneration resulting from the inability of evolution to prevent it, i.e., the nonadaptive consequence of evolutionary "neglect." This conclusion is supported by studies that restricted reproduction to later life in the fruit fly, *Drosophila melanogaster*, thus permitting natural selection to operate on later life traits and leading to an increase in longevity. Likewise, many of the interventions that delay aging are associated with reduced reproductive capacity.

Figure 463-3

Schema linking evolution, cellular and tissue changes with aging. The call-out blue boxes indicate factors that might delay the aging process including nutrient response pathways and, possibly, adaptive evolutionary effects.

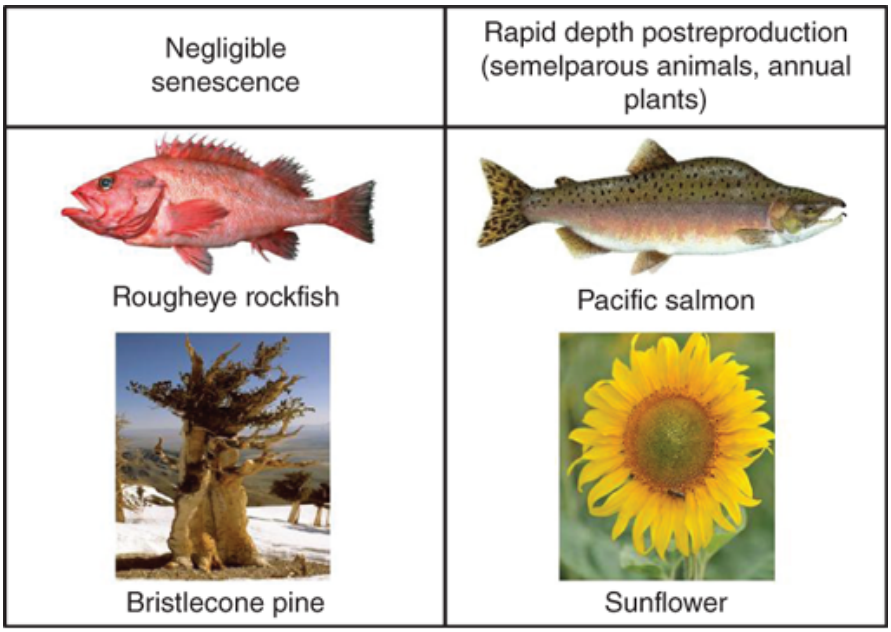


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

There are some species of plants and animals that do not appear to age, or at least they undergo an extremely slow aging process, termed "negligible senescence." The mortality rates of these species are relatively constant with time and they do not display any obvious phenotypic changes of aging. Conversely, some living things undergo programmed death immediately after reproduction such as annual plants and semelparous animals (Fig. 463-4). However, many other living things from yeast to humans undergo a gradual aging process leading to death that is surprisingly similar at the cellular and biochemical level across taxa.

Figure 463-4

The typical features of aging (aging phenotype and exponential increase in risk of death) are not universal findings in living things. Some living things (e.g., Rockfish and the Bristlecone pine, sometimes called the Methuselah tree) undergo negligible senescence while others die almost immediately after first reproductive effort is completed (e.g., semelparous animals and annual plants).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Some of the major classical evolutionary theories of aging include:

Programmed death. The first evolutionary theory of aging was proposed by Weismann in 1882. This theory states that aging and death are programmed and have evolved to remove older animals from the population so that environmental resources such as food and water are freed up for younger members of the species.

Mutation accumulation. This theory was proposed by Medawar in 1952. Natural selection is most powerful for those traits that influence reproduction in early life, and therefore, the ability of evolution to shape our biology declines with age. Germline mutations that are deleterious in later life can accumulate simply because natural selection cannot act to prevent them.

Antagonistic pleiotropy. George C. Williams extended Medawar's theory when he proposed that evolution can allow for the selection of genes that are pleiotropic, i.e., beneficial for survival and reproduction in early life, but harmful in old age. For example, genes for sex hormones are necessary for reproduction in early life but contribute to the risk of cancer in old age.

Life history theory. Evolution is influenced by the way that limited resources are allocated to all aspects of life including development, sexual maturation, reproduction, number of offspring, and senescence and death. Therefore "trade-offs" occur between these phases of life. For example, in a hostile environment, survival is highest for those species that have large numbers of offspring and short life span while in a safe and abundant environment, survival is highest for those species that invest resources in a smaller number of offspring and a longer life.

Disposable soma theory. Kirkwood and Holliday in 1979 combined many of these ideas in the disposable soma theory of aging. There are finite resources available for the maintenance and repair of both germ and soma cells so there must be a trade-off between germ cells (i.e., reproduction) versus soma cells (i.e., longevity and aging). The soma cells are disposable from an evolutionary perspective, so they accumulate damage that causes aging while resources are preferentially diverted to the maintenance and repair of the germ cells. For example, the longevity of nematode worm, *Caenorhabditis elegans* is increased when its germ cells are ablated early in life.

All of these theories assume that natural selection has negligible or negative influences on aging. Some ideas propose that aspects of aging might be adaptive and raise the possibility that evolution can act on the aging process in a positive way. These include:

Grandmother hypothesis. The grandmother hypothesis proposed by Hamilton in 1966 describes how evolution can enhance old age. In some animals, including humans, the survival of multiple, dependent offspring is beyond the capacity and resources of their mother. In this situation, the presence of a long-lived grandmother who shares in care of her grandchildren can have a major impact on their survival. These children share some of the genes of their grandmother including those that promoted their grandmother's longevity.

Mother's curse. Mitochondrial dysfunction is a key component of the aging process. Mitochondria contain their own DNA and are only passed on from mother to child, because sperm cells contain almost no mitochondria. Therefore, natural selection can only act on the evolution of mitochondrial DNA in females. The "mother's curse" of the maternal inheritance of mitochondrial DNA might explain why females live longer and age more slowly than males.

Adaptive senectitude. Many traits that are harmful in younger humans such as obesity, hypertension, oxidative stress, and declines in growth hormone and insulin-like growth factor type I (IGF-1) paradoxically appear to be associated with greater survival and function in old age. Perhaps driven by the grandmother effect, this might represent "adaptive senectitude" or "reverse antagonistic pleiotropy" whereby some traits that are harmful in young people become beneficial in older people.

CELLULAR PROCESSES THAT ACCOMPANY AGING

Many cellular processes change with aging. These are generally considered to be degenerative and stochastic or random changes that reflect some sort of time-dependent damage (Fig. 463-3) and have been called the hallmarks and pillars of aging. Whether any of these is the root cause of aging is unknown but they all contribute to the aging phenotype and disease susceptibility.

Oxidative Stress and the "Free Radical Theory of Aging"

Free radicals are chemical species that are highly reactive because they contain unpaired electrons. Oxidants are oxygen-based free radicals that include the hydroxyl free radical, superoxide, and hydrogen peroxide. Most cellular oxidants are waste products generated by mitochondria during the production of ATP from oxygen. More recently, the role of oxidants in cellular signaling and inflammatory responses has been recognized. Unchecked, oxidants can generate chain reactions leading to widespread damage to biological molecules. Cells contain numerous antioxidant defense mechanisms to prevent such oxidative stress including enzymes (superoxide dismutase, catalase, glutathione peroxidase) and chemicals (uric acid, ascorbate). In 1956, Harman proposed the "free radical theory of aging" whereby oxidants generated by metabolism or irradiation are responsible for age-related damage. It is now well established that old age in most species is associated with increased oxidative stress, for example to DNA (8-hydroxyguanosine derivatives), proteins (carbonyls), lipids (lipoperoxides, malondialdehydes), and prostaglandins (isoprostanes). Conversely, many of the cellular antioxidant defense mechanisms including the antioxidant enzymes decline in old age. The free radical theory of aging has spawned numerous studies of supplementation with antioxidants such as vitamin E to delay aging in animals and humans. However, meta-analyses of human clinical trials performed to treat and prevent various diseases with antioxidant supplements indicate that they have no effect on, or may even increase, mortality.

Mitochondrial Dysfunction

Aging is characterized by altered mitochondrial production of ATP and oxygen-derived free radicals. This leads to a vicious cycle mediated by accumulation of oxidative injury to mitochondrial proteins and DNA. With age, the number of mitochondria

in cells decreases and there is an increase in their size (megamitochondria) associated with other structural changes including vacuolization and disrupted cristae. These morphological aging changes are linked with decreased activity of mitochondrial complexes I, II, and IV and decreased ATP production. Of all of the complexes involved in ATP production, the activity of complex IV (COX) is usually reported to be most impaired in old age. Reduced energy production is linked with generation of hydrogen peroxide and superoxide radicals leading to oxidative injury to mitochondrial DNA and accumulation of carbonylated mitochondrial proteins and mitochondrial lipoperoxides. As well as being implicated in the aging process, common geriatric syndromes including sarcopenia, frailty, and cognitive impairment are associated with mitochondrial dysfunction.

Telomere Shortening and Replicative Senescence

Cells that are isolated from animal tissue and grown in culture only divide for a certain number of times before entering a senescent phase. This number of divisions is known as the Hayflick limit and tends to be less in cells isolated from older animals compared to younger animals. It has been suggested that aging in vivo might in part be secondary to some cells ceasing to divide because they have reached their Hayflick limit. Senescent cells produce a variety of cytokines, chemokines, and proteases, termed the senescence-associated secretory phenotype (SASP), that are major drivers of age-related inflammation. Eliminating senescent cells delayed aging in mice. On the other hand, cellular senescence may have a role in preventing proliferation of cells at risk of malignant transformation. One mechanism for replicative senescence relates to telomeres. Telomeres are repeat sequences of DNA at the end of linear chromosomes that shorten by around 50–200 base pairs during each cell division by mitosis. Once telomeres become too short, cell division can no longer occur. This mechanism contributes to the Hayflick limit and has been called the cellular clock. There are some studies that suggest that the length of telomeres in circulating leukocytes (leukocyte telomere length, LTL) decreases with age in humans. However, the aging process also occurs in tissues that do not undergo repeated cell division such as neurons.

Altered Gene Expression, Epigenetics, and microRNA

The expression of many genes and proteins changes during the aging process. These changes are complicated and vary between species and tissue. Such heterogeneity reflects increasing dysregulation of gene expression with age while appearing to exclude a programmed and/or uniform response. With old age, reductions in the expression of genes and proteins associated with mitochondrial function and increased expression of those involved with inflammation, genome repair, and oxidative stress are noted. Several factors control the regulation of gene and protein expression that change with aging. These include the epigenetic state of the chromosomes (e.g., DNA methylation and histone acetylation) and microRNAs (miRNAs). DNA methylation correlates with age, although the pattern of change is complex. Histone acetylation is regulated by many enzymes including Sirtuin 1 (SIRT1), a protein that has marked effects on aging and the response to dietary restriction in many species. miRNAs are a very large group of noncoding lengths of RNA (18–25 nucleotides) that inhibit translation of multiple different mRNAs through binding their 3' untranslated regions (3'UTRs). The expression of miRNAs usually decreases with aging and is altered in some age-related diseases. Specific miRNAs linked with aging pathways include miR-21 (associated with target of rapamycin pathway) and miR-1 (associated with insulin/insulin-like growth factor 1 pathway).

Impaired Autophagy and Proteostasis

Cells can remove damaged macromolecules and organelles in a number of ways, often generating cellular energy as a by-product. Intracellular degradation is undertaken by the lysosomal system and the ubiquitin proteasomal system. Both are impaired with aging, leading to the accumulation of waste products that alter cellular functions. Such waste products include lipofuscin, a brown auto-fluorescent pigment found within lysosomes of most cells in old age and often considered to be one of the most characteristic histological features of aging cells. Lysosomes are organelles that contain proteases, lipases, glycases, and nucleotidases that degrade intracellular macromolecules, membrane components, organelles, and some pathogens through a process called autophagy. The lysosomal process most impaired with aging is macroautophagy, which is

regulated by numerous autophagy-related genes. Proteostasis refers to the maintenance of protein quality through regulation of protein folding and protein degradation. Chaperones orchestrate appropriate folding of proteins while degradation involves ubiquitin tagging, proteases, and the unfolded protein response. With aging damaged, aggregated and misfolded proteins increase because of age-related changes in proteostasis. This may contribute to the aggregation of proteins such as tau, β-amyloid, α-synuclein in age-related neurodegenerative diseases such as dementia and Parkinson's disease.

AGING CHANGES IN SPECIFIC TISSUES THAT PREDISPOSE TO DISEASE

Aging changes in some tissues increase susceptibility to age-related disease as a secondary or downstream phenomenon (Fig. 463-3). In humans, this includes, but is not limited to, the immune system (leading to increased infections and autoimmunity), hepatic detoxification (leading to increased exposure to disease-inducing endobiotics and xenobiotics), the endocrine system (leading to hypogonadism and bone disease), and the vascular system (leading to segmental or global ischemic changes in many tissues).

Inflamm-aging and Immunosenescence

Old age is associated with increased background levels of inflammation including blood measurements of C reactive protein (CRP), erythrocyte sedimentation rate (ESR), and cytokines such as interleukin 6 (IL-6) and tumor necrosis factor alpha (TNFα). This has been termed "inflamm-aging" and elevated IL-6 in particular has been associated with frailty and dementia. T cells are less numerous because of age-related atrophy of the thymus, while B cells overproduce autoantibodies, leading to the age-related increase in autoimmune diseases and gammopathies. Thus, older people are generally considered to be immunucompromised and have reduced responses to infection (fever, leukocytosis) with increased mortality.

Detoxification and the Liver

Old age is associated with impaired detoxification of various disease-causing endobiotics (e.g., lipoproteins) and xenobiotics (e.g., neurotoxins and carcinogens) leading to increased systemic exposure. In humans, the liver is the major organ for the clearance of such toxins. Hepatic clearance of many substrates is reduced in old age as a consequence of reduced hepatic blood flow, impaired hepatic microcirculation and, in some cases, reduced expression of xenobiotic metabolizing enzymes. These changes in hepatic detoxification also increase the likelihood of increased blood levels of, and adverse reactions to, medications.

Endocrine System

Hormonal changes with aging have been a focus for aging research for over a century, partly because of the erroneous belief that supplementation with sex hormones will delay aging and rejuvenate older people. There are age-related reductions in sex steroids, growth hormone, IGF-1, and dehydroepiandrosterone (DHEA). These hormonal changes may contribute to some features of aging such as sarcopenia and osteoporosis but also provide protection against cancer and cardiovascular disease. Adverse effects of long-term hormonal supplementation outweigh any potential beneficial effects on life span.

Vascular Changes

There is a continuum from vascular aging through to atherosclerotic disease, present in many, but not all, older people. Vascular aging changes overlap with the early stages of hypertension and atherosclerosis, with increasing arterial stiffness and vascular resistance. This contributes to myocardial ischemia and strokes but also appears to be associated with geriatric conditions such as dementia, sarcopenia, and osteoporosis. In these conditions, impaired exchange between blood and tissues is a common pathogenic factor. For example, the risk of Alzheimer's disease and dementia is increased in patients with risk factors for vascular disease, and pathological evidence for microvascular changes is seen in postmortem studies of brains

of people with established Alzheimer's disease. Similarly, strong epidemiological links have been found between osteoporosis and standard vascular risk factors, while significant age-associated changes are in the microcirculation of osteoporotic bone. Sarcopenia might also be related to the effects of age on the muscle vasculature, which is altered in old age. The sinusoidal microcirculation of the liver becomes markedly altered during aging (pseudocapillarization), which influences hepatic uptake of lipoproteins, insulin, and other substrates. In fact, it has often been overlooked that in his original exposition of the free radical theory of aging, Harman proposed that the primary target of oxidative stress was the vasculature and that many aging changes were secondary to impaired exchange across the damaged blood vessels.

GENETIC INFLUENCES ON AGING

There is variability in aging and life span in populations of genetically identical species such as mice that are housed in the same environment. The heritability of life span in human twin studies is estimated to be only 25% (although there is stronger hereditary contribution to extreme longevity). These two observations indicate that the cause of aging is unlikely to lie only within the DNA code. On the other hand, genetic studies initially undertaken in the nematode worm *C. elegans* and, more recently, in models from yeast to mice have shown that manipulating genes can have profound effects on the rate of aging. Perhaps surprisingly, this can often be generated by variability in *single* genes, and for some genetic mechanisms, there is very strong evolutionary conservation.

Genetic Progeroid Syndromes

There are a few very rare, genetic premature aging conditions that are called progeroid syndromes. These conditions recapitulate some, but not all, age-related diseases and senescent phenotypes. They are mostly caused by impairment of genome and nuclear maintenance. These syndromes include the following:

Werner's syndrome. This is an autosomal recessive condition caused by a mutation in the *Werner's (WRN)* gene. This gene codes for a RecQ helicase, which unwinds DNA for both repair and replication. It is typically diagnosed in teen years and is associated with premature onset of atherosclerosis, osteoporosis, cancers, and diabetes with death by age of 50 years.

Hutchinson–Gilford progeria syndrome (HGPS). This usually occurs as a de novo, noninherited mutation in the lamin A gene (*LMNA*) leading to an abnormal protein called progerin. LMNA is required for the nuclear lamina, which provides structural support to the nucleus. Marked development changes are obvious in infancy with subsequent onset of atherosclerosis, kidney failure, and scleroderma-like features and death during the teen years. Lamin A-dependent nuclear defects have been reported in normal human aging.

Cockayne syndrome. This includes a number of autosomal recessive disorders with features such as impaired neurological growth, photosensitivity (xeroderma pigmentosum), and death during childhood years. These are caused by mutations in the genes for DNA excision repair proteins, *ERCC-6* and *ERCC-8*.

Gene Studies in Long-Lived Humans

The main genes that have been consistently associated with increased longevity in human candidate gene studies are *APOE* and *FOXO3A*. ApoE is an apoprotein found in chylomicrons while the ApoE4 isoform is a risk factor for Alzheimer's disease and cardiovascular disease, which might explain its association with reduced life span. FOXO3A is a transcription factor involved in the insulin/IGF-1 pathway, and its homolog in *C. elegans*, daf16, has a marked impact on aging in these nematodes. Genome-wide association studies (GWASs) of centenarians have confirmed the association of longevity with *APOE*. GWAS has been used to identify a range of other single nucleotide polymorphisms (SNPs) that might be associated with longevity including SNPs in the sirtuin genes and the progeroid syndrome genes, *LMNA* and *WRN*. Gene set analysis of GWAS studies has shown that both the insulin/IGF-1 signaling pathway and the telomere maintenance pathway are associated with longevity.

One group of particular interest are people with Laron-type dwarfism. These people have mutations in the growth hormone receptor that causes severe growth hormone resistance. In mice, similar knockout of the growth hormone receptor (GHRKO mice, Methuselah mice) is associated with extremely long life. Therefore, subjects with Laron syndrome have been carefully studied and it was found that they have very low rates of cancer and diabetes mellitus, and possibly, longer lives.

Nutrient-Sensing Pathways

Many living things have evolved to respond to periods of nutritional shortage and famine by increasing cellular resilience and delaying reproduction until food supply becomes abundant once again. This increases the chances of reproductive success and survival of offspring. Lifelong food shortage, often termed "caloric restriction (CR)" (or "dietary restriction"), increases life span and delays aging in many animals, probably as a side effect of this famine response. Many of the genes and pathways that regulate the way that cells respond to nutritional undersupply have been identified, initially in yeast and *C. elegans*. In general, manipulation of these pathways (through genetic knockout or overexpression, or pharmacological agonists and antagonists) alters the aging benefits of CR, and in some cases, the life span of animals on normal diets. These pathways are all very influential cellular "switches" that control a wide range of key functions including protein translation, autophagy, mitochondrial function and bioenergetics, and the cellular metabolism of fats, proteins, and carbohydrates. The discovery of these nutrient-sensing pathways has led to targets for pharmacological extension of life span. The main nutrient-sensing pathways that influence aging and responses to CR include:

SIRT1. The activity of SIRT1 is regulated by levels of reduced nicotinamide adenine dinucleotide (NAD^+), which are increased when cellular energy stores are depleted. Important downstream targets include PGC-1α and Nrf2, which act on mitochondrial biogenesis.

Mechanistic target of rapamycin (mTOR). mTOR is activated by branched chain amino acids, providing a link to dietary protein intake. It has two main components, mTORC1 and mTORC2, of which mTORC1 is most relevant to aging. Key downstream targets of mTOR of relevance to aging include the tuberous sclerosis protein (TSC) and 4EBP1, which influences protein production.

5' adenosine monophosphate-activated protein kinase (AMPK). AMPK is activated by increased levels of AMP, which reflect cellular energy status.

Insulin signaling and IGF-1/growth hormone (IIS). These two pathways are usually considered together because they overlap in lower animals and have diverged only in higher animals. They respond to carbohydrate and protein intake. An important downstream target for this pathway is a transcription factor called FOXO.

Fibroblast growth factor 21 (FGF21). FGF21 is produced mostly by the liver in response to CR and reduced protein intake.

Mitochondrial Genes

Mitochondrial function is influenced by genes located both in the mitochondria (mtDNA) and the nucleus. mtDNA is considered to have a prokaryotic origin and is highly conserved across taxa. It forms a circular loop of 16,569 nucleotides in humans. Aging is associated with increased frequency of mutations in mtDNA as a consequence of its high exposure to oxygen-derived free radicals and relatively inefficient DNA repair machinery. Nuclear DNA encodes ~1000–1500 genes for mitochondrial function including genes involved with oxidative phosphorylation, mitochondrial metabolic pathways, and enzymes required for biogenesis. These genes are thought to have originated in mtDNA but subsequently translocated to the nucleus and, unlike mtDNA genes, their sequence is stable with aging.

Genetic manipulation of mitochondrial genes in animals influences aging and life span. In *C. elegans*, many mutants with defective electron transfer chain function have increased life span. The mtDNA "mutator" mice which lack the mtDNA

proofreading enzyme have increased mtDNA mutations and premature aging, while overexpression of mitochondrial uncoupling proteins leads to longer life span. In humans, hereditary variability in mtDNA is associated with diseases (mitochondriopathies such as Leigh's disease) and aging. For example, in Europeans, mitochondrial DNA haplogroup J (haplogroups are combinations of genetic variants that exist in specific populations) is associated with longevity, and haplogroup D is overrepresented in Asian centenarians.

Nuclear DNA

Genomic DNA damage accumulates in cells with aging while genetic progeroid conditions such as Werner's syndrome and HGPS are associated with impaired DNA maintenance and repair. Age-related DNA changes include mutations, chromosomal aneuploidy, copy number variations, and telomere shortening. Apart from telomere attrition, these changes are random and vary between cells, and may contribute to age-related cancers.

STRATEGIES THAT INCREASE HEALTH SPAN AND DELAY AGING

Aging is an intrinsic feature of human life whose manipulation has fascinated humans ever since becoming conscious of their own existence. Several long-term experimental interventions (e.g., resveratrol, rapamycin, spermidine, and metformin) may open doors for corresponding pharmacological strategies. Surprisingly, most of the effective aging interventions proposed converge on only a few molecular pathways: nutrient signaling, mitochondrial proteostasis, and the autophagic machinery.

Life span is inevitably accompanied by functional decline, steady increase of a plethora of chronic diseases, and ultimately death. For millennia, it has been a dream of mankind to prolong both life span and health span. Developed countries have profited from the medical improvements and their transfer to public health care systems—as well as from better living conditions derived from their socioeconomic power—to achieve remarkable increases in life expectancy during the last century. In the United States, the percentage of the population aged ≥65 years is projected to increase from 13% in 2010 to 19.3% in 2030. However, old age remains the main risk factor for major life-threatening disorders, and the number of people suffering from age-related diseases is anticipated to almost double over the next two decades. The prevalence of age-related pathologies represents a major threat as well as an economic burden that urgently needs effective interventions.

Molecules, drugs, and other interventions that might decelerate aging processes continue to raise interest among both the general public and scientists of all biological and medical fields. Over the past two decades, this interest has taken root in the fact that many of the molecular mechanisms underlying aging are interconnected and linked with pathways that cause disease, including cancer, cardiovascular and neurodegenerative disorders. Unfortunately, among the many proposed aging interventions, only a few have reached a certain age themselves. Results often lack reproducibility because of a simple inherent problem: interventions in aging research take a lifetime to assess. Experiments lasting the lifetime of animal models are prone to develop artifacts, increasing the possibilities and time windows for experimental discrepancies. Some inconsistencies in the field arise from overinterpreting life span-shortening models and scenarios as being accelerated aging.

Many substances and interventions have been claimed to be antiaging throughout history and into the present. In the following sections, interventions will be restricted to those that meet the following highly selective criteria: (1) promotion of life span and/or health span, (2) validation in at least three model organisms, and (3) confirmation by at least three different laboratories. These include: (1) CR and fasting regimens, (2) some pharmacotherapies (resveratrol, rapamycin, spermidine, and metformin), and (3) exercise.

Caloric Restriction

One of the most important and robust interventions that delays aging is CR. This outcome has been recorded in rodents, dogs, worms, flies, yeasts, monkeys, and prokaryotes. CR is defined as a reduction in the total caloric intake, usually of about 30%

and without malnutrition. CR reduces the release of growth factors such as growth hormone, insulin, and IGF-1, which are activated by nutrients and have been shown to accelerate aging and enhance the probability for mortality in many organisms. Yet the effects of CR on aging were first discovered by McCay in 1935 long before the effects of such hormones and growth factors on aging were recognized. The cellular pathways that mediate this remarkable response have been explored in many experimental models. These include the nutrient sensing pathways (mTOR, AMPK, insulin/IGF-1, and sirtuins) as well as transcription factors (FOXO in *D. melanogaster* and daf-16 in *C. elegans*). The transcription factor Nrf2 appears to confer most of the anticancer properties of CR in mice, even though it is dispensable for life span extension.

The effects of CR in monkeys have been assessed in two studies with different outcomes: one study observed prolonged life while the other did not. However, both studies confirmed that CR increases health span by reducing the risk for diabetes, cardiovascular disease, and cancer. In humans, CR is associated with increased life and health span. This is most convincingly demonstrated in Okinawa, Japan, where one of the most long-lived human populations resides. In comparison to the rest of the Japanese population, Okinawan people usually combine an above-average amount of daily exercise with a below-average food intake. However, when Okinawan families move to Brazil, they adopt a Western lifestyle that affects both exercise and nutrition, causing a rise in weight and a reduction in life expectancy by nearly two decades. In the Biosphere II project, where volunteers lived together for 24 months undergoing an unforeseen severe CR, there were improvements in insulin, blood sugar, glycated hemoglobin, cholesterol levels, and blood pressure—all outcomes that would be expected to benefit life span. CR changes many aspects of human aging that might influence life span such as the transcriptome, hormonal status (especially IGF-1 and thyroid hormones), oxidative stress, inflammation, mitochondrial function, glucose homeostasis, and cardiometabolic risk factors. Epigenetic modifications are an emerging target for CR.

It must be noted that maintaining CR and avoiding malnutrition is not only arduous in humans but is also linked with substantial side effects. For instance, prolonged reduction of calorie intake may decrease fertility and libido, impair wound healing, reduce the potential to combat infections, and lead to amenorrhea and osteoporosis.

While extreme obesity (body mass index [BMI] >35) leads to a 29% increased risk of dying, people with BMI in the overweight range seem to have reduced mortality, at least in population studies of middle-aged and older subjects. People with a BMI in the overweight range seem more able to counteract and respond to disease, trauma, and infection, whereas CR impairs healing and immune responses. On the other hand, BMI is an insufficient denominator of body and body fat composition. A well-trained athlete may have an equivalent BMI compared to a fat person because of the higher muscle mass density. The waist:hip ratio is a much better indicator for body fat and an excellent and stringent predictor for the risk to die from cardiovascular disease: the lower the waist:hip ratio, the lower the risk.

PERIODIC FASTING

How can CR be translated to humans in a socially and medically feasible way? A whole series of periodic fasting regimens are asserting themselves as suitable strategies, among them the alternate-day fasting diet, the “five:two” intermittent fasting diet, and a 48-h fast once or twice each month. Periodic fasting is psychologically more viable, lacks some of the negative side effects and is only accompanied by minimal weight loss.

It is striking that many cultures implement periodic fasting rituals, for example Buddhists, Christians, Hindus, Jews, Muslims, and some African animistic religions. It could be speculated that a selective advantage of fasting versus nonfasting populations is conferred by health-promoting attributes of religious routines that periodically limit caloric intake. Indeed, several lines of evidence indicate that intermittent fasting regimens exert antiaging effects. For example, improved morbidity and longevity were observed among Spanish home nursing residents who underwent alternate-day fasting. Even rats subjected to alternate-day fasting live up to 83% longer than normally fed control animals and one 24-h fasting period every 4 days is sufficient to generate life span extension.

Repeated fasting and eating cycles may circumvent the negative side effects of sustained CR. This strategy may even yield effects despite extreme overeating during the nonfasting periods. In a spectacular experiment, mice fed a high-fat diet in a time-restricted manner, i.e., with regular fasting breaks, showed reduced inflammation markers, no fatty liver and were slim in comparison to mice with equivalent total calorie consumption but ad libitum. From an evolutionary point of view, this kind of feeding pattern may reflect mammalian adaptation to food availability: overeating in times of nutrient availability (e.g., after a hunting success) and starvation in between. This is how some indigenous peoples who have avoided Western lifestyles live today; those who have been investigated show limited signs of age-induced diseases such as cancer, neurodegeneration, diabetes, cardiovascular disease, and hypertension.

Fasting exerts beneficial effects on health span by minimizing the risk of developing age-related diseases including hypertension, neurodegeneration, cancer, and cardiovascular diseases. The most effective and rapid repercussion of fasting is reduction in hypertension. Two weeks of water-only fasting resulted in a blood pressure below 120/80 mmHg in 82% of subjects with borderline hypertension. Ten days of fasting cured all hypertensive patients who had been taking antihypertensive medication previously.

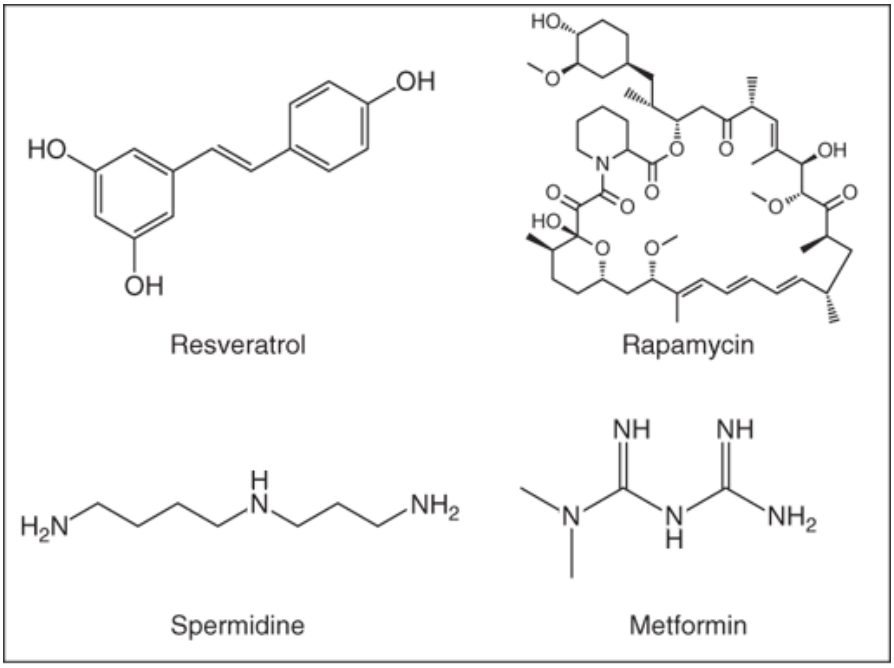
Periodic fasting dampens the consequences of many age-related neurodegenerative diseases (Alzheimer's disease, Parkinson's disease, Huntington's disease, and frontotemporal dementia but not amyotrophic lateral sclerosis in mouse models). Fasting cycles are as effective as chemotherapy against certain tumors in mice. In a combination with chemotherapy, fasting protected mice against the negative side effects of chemotherapeutic drugs, while it enhances their efficacy against tumors. Combining fasting and chemotherapy rendered 20–60% mice cancer-free when inoculated with highly aggressive tumors like glioblastoma or pancreatic tumors, which have 100% mortality even with chemotherapy.

Pharmacological Interventions to Delay Aging and Increase Life Span

Virtually all obese people know that stable weight reduction will reduce their elevated risk of cardiometabolic disease and enhance their overall survival, yet only 20% of overweight individuals are able to lose 10% weight for a period of at least 1 year. Even in the most motivated people (such as the "Cronies" who deliberately attempt long-term CR in order to extend their lives), long-term CR is extremely difficult. Thus, focus has been directed at the possibility of developing medicines that replicate the beneficial effects of CR without the need for reducing food intake ("CR-mimetics," Fig. 463-5):

Figure 463-5

Chemical structure of four agents (resveratrol, rapamycin, spermidine, and metformin) that have been shown to delay aging in experimental animal models.


HO
OH
OH
Resveratrol
HO
OH
O
O
O

In humans and animals, regular exercise reduces the risk of morbidity and mortality. Given that cardiovascular diseases are the dominant cause of aging in humans but not in mice, the effects on human health may be even stronger than those seen in mouse experiments. An increase in aerobic exercise capacity, which declines during aging, is associated with favorable effects on blood pressure, lipids, glucose tolerance, bone density, and depression in older people. Likewise, exercise training protects against aging disorders such as cardiovascular diseases, diabetes mellitus, and osteoporosis. Exercise is the only treatment that can prevent or even reverse sarcopenia (age-related muscle wasting). Even moderate or low levels of exercise (30 min walking per day) have significant protective effects in obese subjects. In older people, regular physical activity has been found to increase the duration of independent living.

While clearly promoting health and quality of life, regular exercise does not extend life span. Furthermore, the combination of exercise with CR has no additive effect on maximal life span in rodents. On the other hand, alternate-day fasting with exercise is more beneficial for the muscle mass than single treatments alone. In nonobese humans, exercise combined with CR has synergistic effects on insulin sensitivity and inflammation. From the evolutionary perspective, the responses to hunger and exercise are linked: when food is scarce, increased activity is required to hunt and gather.

Hormesis

The term hormesis describes the, at first sight paradoxical, protective effects conferred by the exposure to low doses of stressors or toxins (or as Nietzsche stated “What does not kill me makes me stronger”). Adaptive stress responses elicited by noxious agents (chemical, thermal, or radioactive) precondition an organism rendering it resistant to subsequent higher and otherwise lethal doses of the same trigger. Hormetic stressors have been found to influence aging and life span presumably by increasing cellular resilience to factors that might contribute to aging such as oxidative stress.

Yeast cells that have been exposed to low doses oxidative stress exhibit a marked antistress response that inhibits death following exposure to lethal doses of oxidants. During ischemic preconditioning in humans, short periods of ischemia protect the brain and the heart against a more severe deprivation of oxygen and subsequent reperfusion-induced oxidative stress. Similarly, the lifelong and periodic exposure to various stressors can inhibit or retard the aging process. Consistent with this concept, heat or mild doses of oxidative stress can lead to life span extension in *C. elegans*. CR can also be considered as a type of hormetic stress that results in the activation of antistress transcription factors (Rim15, Gis1, and Msn2/Msn4 in yeast and FOXO in mammals) that enhance the expression of free radical-scavenging factors and heat shock proteins.

CONCLUSIONS

Clinicians need to understand aging biology in order to better manage those people who are elderly. Moreover, there is an urgent need to develop strategies based on aging biology that delay aging, reduce the onset of age-related disorders, and increase health span for future generations. Interventions related to nutrition and those drugs that act on nutrient-sensing pathways are being developed and, in some cases, are already being tested in humans.

FURTHER READING

Campisi J: Aging, cellular senescence, and cancer. Annu Rev Physiol 75:685, 2013. [PubMed: 23140366]

de Cabo R et al: The search for antiaging interventions: From elixirs to fasting regimens. Cell 157:1515, 2014. [PubMed: 24949965]

Kaeberlein M et al: Healthy aging: The ultimate preventative medicine. Science 350:1191, 2015. [PubMed: 26785476]

Kennedy BK et al: Geroscience: Linking aging to chronic disease. Cell 159:709, 2014. [PubMed: 25417146]

Longo VD et al: Interventions to slow aging in humans: Are we ready? Aging Cell 14:497, 2015. [PubMed: 25902704]

López-Otín C et al: The hallmarks of aging. Cell 153:1194, 2013. [PubMed: 23746838]

López-Otín C et al: Metabolic control of longevity. Cell 166:802, 2016. [PubMed: 27518560]

Chapter 464: Clinical Problems Associated with the Aging Process

Joseph G. Ouslander; Bernardo Reyes

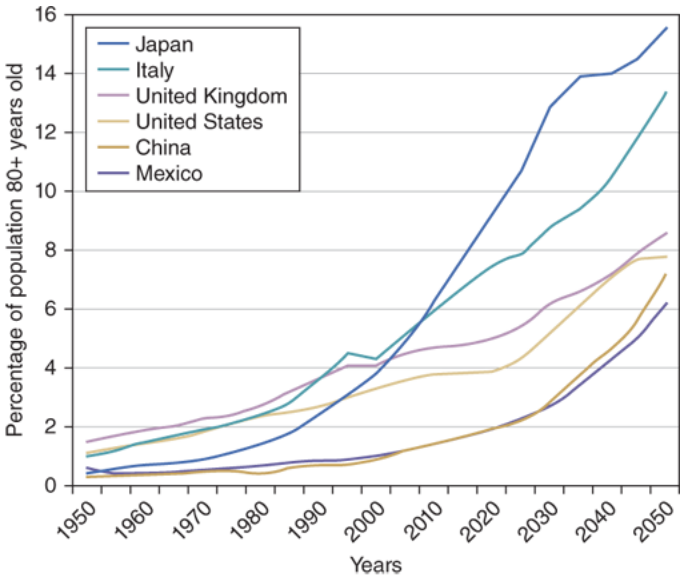
AGING AND GERIATRIC CARE

Demographics of Aging and its Implications for Geriatric Care

The United States and other countries will continue to experience a rapid increase in the number of older adults who seek health care. The most rapidly growing segment of the population in the United States and many other developed countries is those aged >80 years (Fig. 464-1). Sex composition of the aging population around the world is also expected to change. Although females outlive males, an improvement in survival of the oldest males could result in more balanced sex distribution in the geriatric population in the future.

FIGURE 464-1

Percentage of the population age >80 years from 1950 to 2050 in representative nations. *(Updated data available at https://esa.un.org/unpd/wpp/Graphs/DemographicProfiles/. Accessed December 30, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

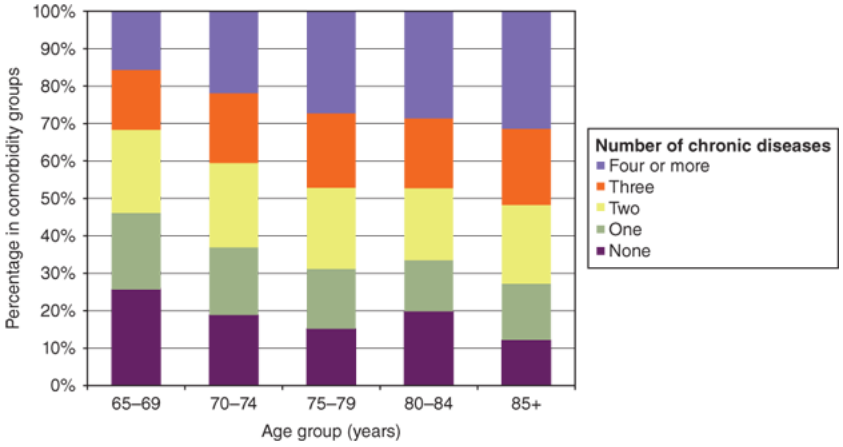
Based on the United Nations' 2015 World Population Aging Report, in high-income countries, consumption of health care resources will be most affected by the shift in the age distribution of the population over the next several decades. The World Health Organization continues to work actively to raise awareness of the changes necessary in current health care systems beyond increments in their budgets. Planning is increasingly being based on expected levels of disability and comorbidity. As lifespan increases, efforts should continue to focus on promoting healthy aging to reduce the burden of disability in health care systems all over the world.

Implications of the Aging Population for Health Care Systems and System-Based Practice

The geriatric population requires different approaches to care for several reasons. For example, acute illnesses are most often not treated in isolation, but in the context of multiple co-morbidities. Close to half of those aged >80 have three, and about one-third have four or more chronic conditions (Fig. 464-2). Functional disabilities are prevalent (Fig. 464-3), which require careful attention in the evaluation of the older patient, along with assessment of social supports available for assistance when needed for independent and safe living.

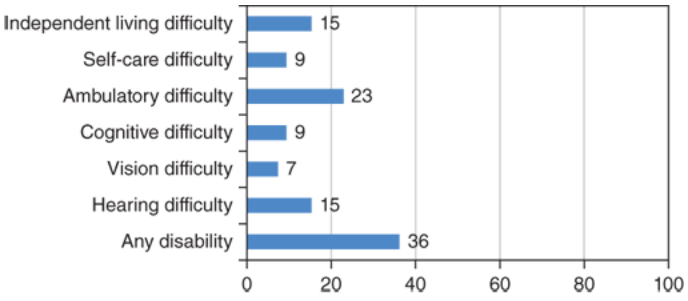
FIGURE 464-2

Prevalence of comorbidity by age group in persons ≥65 years old living in the United States and enrolled in Medicare parts A and B in 1999. (*From JL Wolff et al: Arch Intern Med 162:2269, 2002.*)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.
FIGURE 464-3

Percentage of people age 65+ with various disabilities. (*Source: US Census Bureau, American Community Survey, 2013. Available at https://aoa.acl.gov/Aging_Statistics/Profile/2014/index.aspx. Accessed December 30, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Effectively caring for the geriatric population requires consideration of several key principles:

1. Aging is not a disease; normal aging changes generally do not cause symptoms, but do increase susceptibility to many diseases and conditions due to diminished physiologic reserve (which has been termed “homeostenosis”). Aging is also associated with greater heterogeneity in virtually every measureable variable. Lab values outside the “normal” range are more common and may not reflect pathology.
2. Medical conditions are commonly multiple (“multi-morbidity”) and multifactorial in origin, requiring a comprehensive approach to evaluation and management.
3. Many potentially reversible and treatable conditions are under-diagnosed and under-evaluated in this population, such as fall risk, urinary incontinence, and elder abuse and neglect; simple screening tools can help detect them.
4. Similarly, cognitive and affective disorders (e.g. mild cognitive impairment, dementia, depression, anxiety) are common and may be undiagnosed in early stages; simple screening tools can help detect them.
5. Iatrogenic illnesses are common, especially related to adverse drug reactions and immobility and related deconditioning and other complications.
6. Functional ability and quality of life, as opposed to cure, are key goals of care.
7. Social history, social support, and patient preferences are critical to treat older people in a safe and person-centered manner.
8. Effective geriatric care requires inter-professional collaboration among many different disciplines.
9. Geriatric care is provided largely outside the hospital—at home, in skilled nursing and assisted living settings; and attention to care transitions between settings is essential for effective care.
10. Ethical issues, palliative care, and end-of-life care are critical aspects of caring for the geriatric population.

In this chapter, these key principles are highlighted. The reader is referred to textbooks of geriatric medicine for more details on each of the principles, and the management of common diseases and conditions in this population.

Models of Geriatric Care and Care Transitions

Several innovative models of care have been developed over the last three decades designed to provide high quality and effective care for the burgeoning geriatric population with multi-morbidity, functional and cognitive impairment, and challenges with social support. These include outpatient comprehensive geriatric

assessment programs, inpatient acute care for the elderly (ACE) units and consultation services, and home-based programs such as Geriatric Resources for Assessment and Care of Elders (GRACE), Home Based Primary Care (in the VA system), and Independence at Home. These models of care are assuming greater importance in the emerging era of value-based purchasing for health care services. While they may be challenging and inefficient to implement in the Medicare fee-for-service system, they may also result in improved care and lower costs overall as Medicare shifts from fee-for-service to other models of reimbursement, such as accountable care organizations, bundled payment programs, and increasing the number of older people enrolled in Medicare managed care.

Improving transitions of care between settings has become a major focus of the federal government, health systems, hospitals, post-acute (PAC) and long-term care (LTC) organizations and programs, physicians, and other health care professionals. Geriatric patients are especially vulnerable to complications at the time of discharge from an acute medical or psychiatric hospital, as well as at the time of discharge from a PAC facility (skilled nursing facility [SNF]; acute rehabilitation or long-term hospital) or home care program. With the increasing role of hospitalists and physicians who specialize in SNF care, medical care for geriatric patients has become fragmented at the time transitions, creating opportunities for communication problems and medical errors. Changes in reimbursement and financial penalties for high rates of hospital readmissions have driven the development of many care transition interventions **(Table 464-1)**. These interventions involve inter-professional collaboration and a variety of strategies targeted at making care transitions safer, and reducing unnecessary return visits to the emergency department, hospital readmissions, and related complications and costs.

TABLE 464-1

Examples Care Transitions Interventions

Intervention	Website
Re-Engineered Discharge (Project RED) (Jack et al: 2009)	*https://www.bu.edu/fammed/projectred/*
Transitional Care Model (Naylor et al: 2004; Naylor et al: 1999)	*https://www.nursing.upenn.edu/ncth/transitional-care-model/*

Intervention	Website
Care Transitions Program® (Coleman et al: 2004)	*http://www.caretransitions.org*
Better Outcomes for Older Adults through Safe Transitions (BOOST) (Hansen et al: 2013)	*http://www.hospitalmedicine.org/Web/Quality_Innovation/Implementation_Toolkits/Project_BOOST/Web/Quality___Innovation/Implementation_Toolkit/Boost/Ov hkey=09496d80-8dae-4790-af72-efed8c3e3161*
Interventions to Reduce Acute Care Transfers (INTERACT) (Ouslander et al: 2013)	*https://interact.fau.edu*

Inter-Professional Teams and Co-Managed Care

The complexity of caring for the aging population is more evident during a hospitalization due to a new acute illness or exacerbation of pre-existing chronic conditions. Inter-professional teams integrate different areas of expertise with the aim of providing patient-centered care. Physicians should understand and respect the roles of nurses, physical, occupational, and speech therapists, nutritionists, pharmacists, psychologists, social workers, clergy, and other direct care staff. The

evolution of inter-professional teams has resulted in a comprehensive approach to care by opening channels of communication between these health professionals from different disciplines.

“Huddles” are a mechanism of enhanced communication for inter-professional teams. The implementation of efficient huddles has been associated with improved safety and better utilization of resources by predicting patient needs and making appropriate changes in staffing and care plans. Huddles can also help identify potential threats to patient care, such as socioeconomic challenges that can make care plans ineffective of even harmful.

Another strategy for enhanced communication and collaboration in the care of complex geriatric patients is “Co-Managed Medicine.” In this model, internists serve as part of a multispecialty team of physicians (that often include surgeons) that provides daily assessments, addresses medical comorbidities, and facilitates transitions of care; thereby enhancing the typical consultant model. Co-managed medicine is another example of how enhanced communication between different providers improves outcomes, avoids common complications, and saves resources. In the era of person-centered care and value-based medicine, effective co-managed medicine appears to deliver consistently high quality care at a lower cost. Since the rise of hospitalist-based care, the use of co-managed care has increased significantly. Hip fracture co-management, as well as trauma co-management, and collaborations between internists and geriatricians are examples of this strategy.

FUNDAMENTALS OF GERIATRIC CARE

Person-Centered Care

Person-centered care is a critical concept in caring for older people because of the complexity of their medical, functional, and psychosocial problems, and in many instances the lack of rigorous data on the most effective strategies for caring for specific conditions in patients with multi-morbidity. Thus, decision-making on goals and approaches to care must account for patient and family preferences and goals, values, perception of risk, prognosis, and other individual factors. For almost any condition, from common disorders such as hypertension and diabetes, to geriatric syndromes such as fall risk and urinary incontinence, the answer to how best to treat medical conditions in an older patient with multi-morbidity does not only depend on evidence-based medicine—it also depends on careful weighing of the factors listed above. In everyday practice with complex older patients, a focus on improving or maintaining function and independence, quality of life, comfort, and dignity will be consistent with patient and family goals.

The American Geriatrics Society (AGS) identifies the following elements as key to person-centered care: (1) an individualized, goal-oriented care plan based on the person’s preferences; (2) ongoing review of the person’s goals and care plan; (3) continual information sharing and integrated communication; (4) education and training for providers and, when appropriate, the person and those important to the person; and (5) performance measurement and quality improvement using feedback from the person and caregivers. Several tools are available to assist in implementing person-centered care, including estimation of prognosis (e.g., “ePrognosis”), and “choosing wisely” recommendations from the AGS and AMDA—The Society for Post-Acute and Long-Term Care Medicine. Examples of these recommendations that are relevant to internal medicine practice are illustrated in **Table 464-2**.

TABLE 464-2

Examples of Choosing Wisely Recommendations Helpful in Implementing Person-Centered Care in Complex Geriatric Patients

- Don’t recommend percutaneous feeding tubes in patients with advanced dementia; instead offer oral assisted feeding.
- Don’t use antipsychotics as the first choice to treat behavioral and psychological symptoms of dementia.
- Avoid using medications other than metformin to achieve hemoglobin A1c <7.5% in most older adults; moderate control is generally better.
- Don’t use benzodiazepines or other sedative-hypnotics in older adults as first choice for insomnia, agitation or delirium.
- Don’t use antimicrobials to treat bacteriuria in older adults unless specific urinary tract symptoms are present.
- Don’t prescribe cholinesterase inhibitors for dementia without periodic assessment for perceived cognitive benefits and adverse gastrointestinal effects.
- Don’t recommend screening for breast, colorectal, prostate or lung cancer without considering life expectancy and the risks of testing, over-diagnosis and overtreatment.
- Don’t routinely prescribe lipid-lowering medications in individuals with a limited life expectancy.
- Don’t obtain a *Clostridium difficile* toxin test to confirm “cure” if symptoms have resolved.
- Don’t recommend aggressive or hospital-level care for a frail elder without a clear understanding of the individual’s goals of care and the possible benefits and burdens.

Adapted from *http://www.choosingwisely.org/societies/american-geriatrics-society/* and *http://www.choosingwisely.org/societies/amda-the-society-for-post-acute-and-long-term-care-medicine/amda-choosing-wisely-list/*. Accessed December 31, 2016.

Evaluation of the Geriatric Patient

GERIATRIC ASSESSMENT

A series of screening questions can be useful as a “geriatric review of systems” in clinical practice with older patients because of the importance and high prevalence of functional impairments and disabilities, limited social support to assist with functional limitations, cognitive and affective disorders, and geriatric conditions that may go undetected and cause patient safety issues and complications **(Table 464-3)**. These questions may be especially helpful in conducting annual Medicare “Wellness Visits.” Positive responses to one or more of the screening questions for each item should prompt consideration of further assessments, many of which can be accomplished using standard and validated tools available on the internet, such as activities of daily living scales, depression scales, sleep questionnaires, and mental status examinations.

TABLE 464-3

Examples of Screening Questions and Tools and Strategies for Further Evaluation of Social Support, Functional Status, Geriatric Syndromes, and Cognition and Affect

Geriatric Assessment Domains		Recommended Screens	Further Assessment for Positive Screen*
SOCIAL	**Social Support**	Do you live alone? Do you have a caregiver? Are you a caregiver?	• Consider referral to a social worker • Refer to Area Agency on Aging
	Elder Neglect/Abuse	Do you ever feel unsafe where you live? Has anyone ever threatened or hurt you? Has anyone been taking your money without your permission?	• Consider referral to a social worker and/or Adult protective services
	Advance Directives	Would you like information or forms for a power of attorney for health care? Would you like information on a living will?	• Discussion on advance directives • Physician Orders for Life-Sustaining Treatment (POLST) (or MOLST or POST)
FUNCTIONAL	**Functional Status**	Do you need assistance with shopping or finances? Do you need assistance with bathing or taking a shower?	• Instrumental Activities of Daily Living (ADL) Scale • Basic ADL Scale
	Driving	Do you still drive? If yes: While driving, have you had an accident in the past 6 months? Driving concerns by family member?	• Vision testing • Consider occupational therapy and/or formal driving evaluation
	Vision	Do you have trouble seeing, reading, or watching TV? (with glasses, if used)	• Vision testing • Consider referral for eye exam
	Hearing	Do you have difficulty hearing conversation in a quiet room? Unable to hear whisper test 6 inches away?	• Check for cerumen in ear canals and remove if impacted • Hearing Handicap Inventory • Consider audiology referral
GERIATRIC SYNDROMES	**Medications**	Do you take 5 or more routine medications? Do you understand the reason for each of your medications?	• Match medications with diagnoses • Consider reducing doses, stopping drugs, adherence aides, and/or consultation with a pharmacist
	Fall Risk	Have you fallen in the past year? Are you afraid of falling? Do you have trouble climbing stairs or rising from chairs?	• "Get Up and Go" test • Consider full Fall Assessment • Consider Physical Therapy Evaluation • Consider Home Safety Assessment
	Continence	Do you have any trouble with your bladder? Do you lose urine or stool when you do not want to? Do you wear pads or adult diapers?	• Consider full continence Assessment • 3 IQ Questionnaire (women) • American Urological Association (AUA) 7 symptom inventory (men)
	Weight Loss	Weight <100 pounds or Unintentional weight loss >10 pounds over 6 months?	• Assess for common risk factors for malnutrition • Consider referral to dietician for nutritional evaluation
	Sleep	Do you often feel sleepy during the day? Do you have difficulty falling asleep at night?	• Epworth Sleepiness Scale or Pittsburgh Sleep Index • Consider referral for sleep evaluation
	Pain	Are you experiencing pain or discomfort?	• Pain Assessment
	Alcohol Abuse	Do you drink > 2 drinks / day?	• AUDIT-C

Geriatric Assessment Domains		Recommended Screens	Further Assessment for Positive Screen*
COGNITION AND AFFECT	Depression	Do you often feel sad or depressed? Have you lost pleasure in doing things over the past few months?	• PHQ – 9 or Geriatric Depression Scale • Screen for suicide risk
	Cognition	Self-reported memory loss? Cognitive screen positive? (3-item recall and Clock Draw test "Mini-Cog") Confusion Assessment (CAM) for delirium	• Montreal Cognitive Assessment or Mini Mental State Examination • If diagnosis is unclear, consider neuropsychological testing

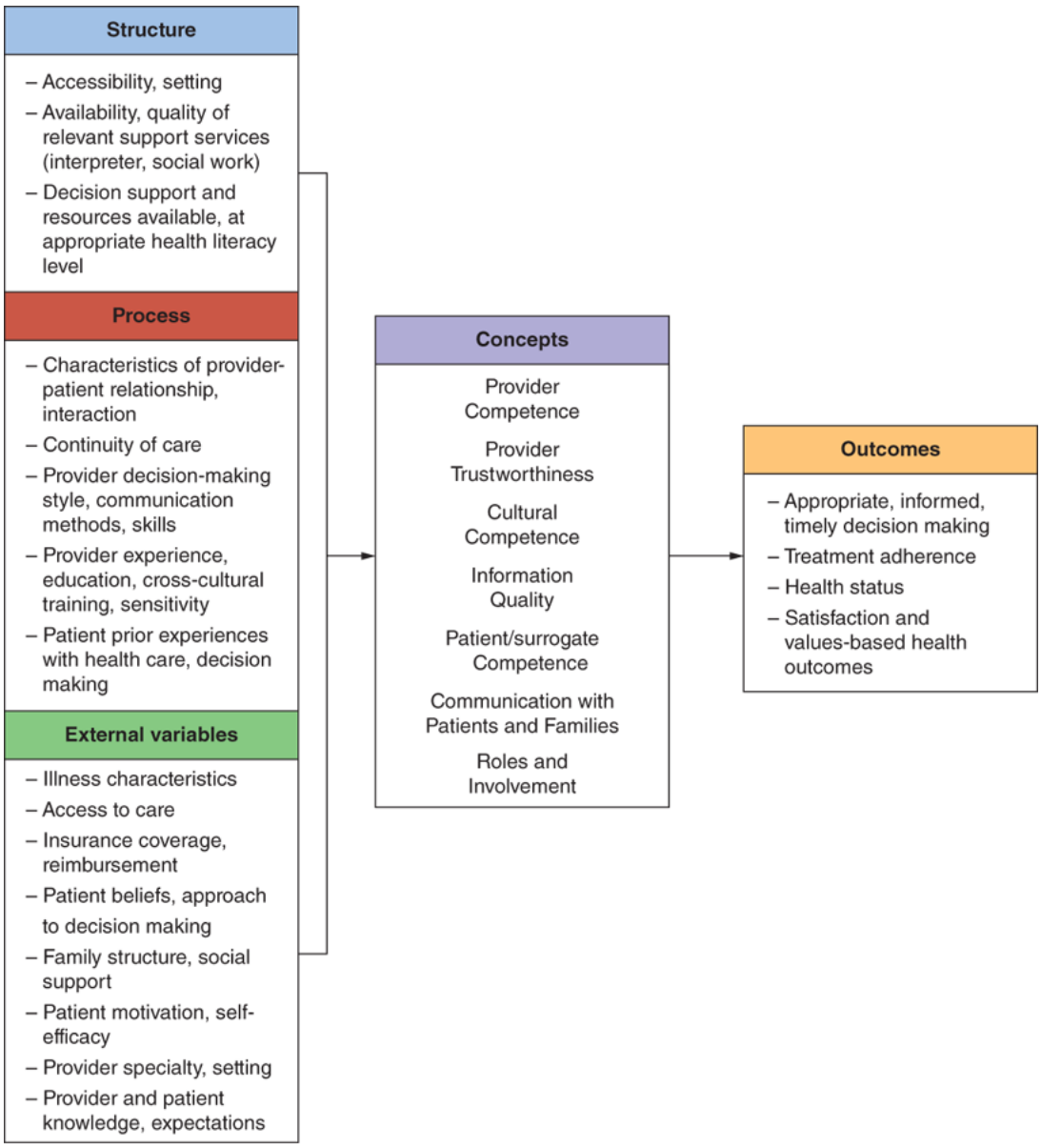
Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

EVALUATION OF MEDICAL DECISION-MAKING CAPACITY

Key aspects of decision-making in older adults are illustrated in Fig. 464-4. Including the patient in the consent process for any treatment is the foundation of patient autonomy and person-centered care. Because aging is associated with an increasing potential to develop cognitive impairment, determination of decision-making capacity is important not only to protect the patients against potential abuse, but also to preserve autonomy when possible and when it is not, that an appropriate surrogate decision-making process is followed. Assessing for capacity is usually triggered by specific circumstances (e.g., the need for invasive diagnostic testing or surgery). Determination of decision-making capacity limited to medical circumstances should be differentiated from declaring a patient "incompetent" to make all decisions. Declaring someone incompetent is a legal definition and usually is reserved for court settings. Another caveat about evaluating decision-making capacity is distinguishing lack of capacity from poorly presented information, sensory impairment, and/or low level of literacy. The clinician should corroborate that the patient has received all the necessary information, comprehends the information provided, and there are no major auditory or visual impairments. For geriatric patients, it is important to determine if the patient uses hearing aids of prescription glasses and they are available for their use.

FIGURE 464-4

Key aspects of decision-making in older adults. *(From SM Dy, SP Tanjala: Key concepts relevant to quality of complex and shared decision-making in health care: A literature review. Soc Sci Med 74:582, 2012.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Standard tests of cognitive function such as the Mini Mental State Examination correlate poorly with capacity to consent for specific interventions. Several standardized tools have been validated to determine decision-making capacity. The MacArthur Competence Assessment Tool-Treatment (MacCAT-T) is a structured tool that has been validated, but it is lengthy and can be difficult to administer in some patients. The Capacity to Consent to Treatment Instrument (CCTI) is another tool that has been validated in patients with mild to moderate Alzheimer's disease. It is structured in two different vignettes and the patient is asked to answer a series of questions. The test has high inter-rater reliability and validity.

EVALUATION OF THE OLDER DRIVER

For many older adults in the United States, driving is essential for maintaining independence and driving cessation is associated with negative outcomes including social isolation and depression. On the other hand, older adults have the highest risk of being involved in fatal crashes with up to a nine-time higher risk for those 85 years old compared to younger people. Older people should be routinely assessed for their driving status and if they have been in any car crashes, as well as for sensory, functional, and cognitive impairments that can make driving unsafe (Table 464-3). Like many geriatric conditions described below, many different types of drugs can impair various aspects of driving performance, and should be carefully considered in older people who continue to drive **(Table 464-4)**.

TABLE 464-4

Medications with Strong Potential to Affect Driving

Anticholinergics
Anticonvulsants
Antidepressants
Antiemetics
Antihistamines
Antihypertensives
Antiparkinsonian agents
Antipsychotics
Benzodiazepines and other sedatives/anxiolytics/hypnotics
Muscle relaxants
Narcotic analgesics
Stimulants
Other agents with significant anticholinergic properties

Suspected driving impairment can be a source of conflict between the patient (who wants to maintain independence), the family (who may want their relative to continue driving due to lack of other transportation; or may be concerned about their safety, or both) and the physician (who is concerned about the patient's, passengers', and other drivers' safety). There is liability involved in these decisions, since any states do not require driving re-testing for all older drivers, and many require physicians to report older people who they believe are unsafe drivers. Evaluation of driving should be inter-professional and aimed to first try to correct any reversible causes of losing driving skills, such as vision and hearing impairment. Although tests of executive function such as the Trails B have been associated with poor driving performance, no single screening test predicts unsafe driving. A combination of neuropsychological testing by a psychologist, and on-road testing by a trained occupational therapist can provide the physician with essential input in making the difficult decision on driving cessation. The AGS and the U.S. Department of Transportation's National Highway Traffic Safety Administration have updated a "Physician's Guide to Assessing and Counseling Older Drivers," which can be helpful to practicing clinicians and is available on the AGS website (see "Further Reading").

INTERPRETATION OF DIAGNOSTIC TESTS

Atypical presentations of medical conditions are a common feature of geriatric medicine. Physiologic changes associated with aging can affect the results of common diagnostic tests as well. The large variation of many physiologic measures that is associated with normal aging makes establishing what is "normal" for many tests challenging. For this reason, the results of several diagnostic tests must be interpreted with caution. Examples include creatinine clearance, pulmonary function, and sedimentation rate (which can confound the diagnosis of polymyalgia). Ambulatory cardiac monitoring may identify a variety of arrhythmias, but they have a high incidence in older people and must be linked with symptoms before considering potentially toxic or invasive treatment. Musculoskeletal imaging, such as an MRI of the spine, may reveal multiple abnormalities that may or may not be related to symptoms. For the most part, however, abnormal diagnostic tests require further evaluation in older patients, unless further evaluation would not lead to a change in the goals of care and treatment plan. Examples include low hemoglobin levels, abnormal thyroid function tests, age/sex/weight adjusted creatinine clearance, and elevated liver function tests; these examples do not result from normal aging and generally indicate a physiologic abnormality resulting from a disease or disorder that may or may not be reversible.

Prevention in Older Adults

AGE-APPROPRIATE SCREENING

Screening tests for specific diseases, as opposed to screening for geriatric conditions requires a careful person-centered approach. The focus of preventive medicine depends heavily on the ability to identify those who are at risk for specific conditions **(see Chap. 4)**. Several professional societies have provided guidance regarding specific tests in older adults **(Table 464-5)**. An important caveat about screening to prevent disease in older patients (e.g., colonoscopy for colon cancer; PAP smears; PSA testing) is that abnormal results may lead to subsequent testing and treatment among individuals who will not suffer morbidity or mortality from the disease because of limited life expectancy. Thus, geriatric patients pose a significant challenge for deciding what screening tests could offer a reasonable ratio of benefit and risk as well as being cost-effective. The *ePrognosis.com* website is a very helpful tool in these determinations.

TABLE 464-5

Recommendations for Primary Prevention Screening for Specific Diseases in Older Adults from Different Professional Societies

			Professional Society Issuing Recommendations		
Type of Screening	**Test**	**Frequency**	**USPSTF[a]**	**ACS[b]**	**ACP[c]**
Colorectal	Fecal occult blood test or fecal immunochemical test (FIT) or Sigmoidoscopy or Colonoscopy	Annual Every 5 y Every 10 y	Screen all adults age 50–75; prognosis may support screening individuals of age 76–85; not recommended for adults over age 85	Screen all adults age >50; discontinuing screening is reasonable in people with severe comorbidity that would preclude treatment	Screen all adults age 50–75 People with life expectancy <10 y should not be screened
Breast	Mammography	Every 1–2 y	Biennial screening all women age 50–74; evidence of benefits and harms is insufficient for women age >75	Annual screening starting at age 40; continue while in good health	**ECOG[d]** Annual screening starting at 40 y/o
Cervical	Pap smear HPV test	Pap only, every 3 y HPV + Pap, every 5 y	Screen women age 21–65; discontinue at age 65 if adequate prior screening	Screen women age 21–65; discontinue at age 65 if regular screening normal	Screening should stop at age 65 if evidence of negative adequate prior screening.
Lung	Low-dose CT scan	Annual	Screen age 55–80 current and former smokers with a 30+ pack-year smoking history; discontinue screening once a person has not smoked for 15 years or develops a health problem that limits their ability or willingness to have curative surgery	Screen 55–74-year-old current and former smokers in good health with a 30+ pack-year smoking history	**ACCP[e]** In settings that can deliver the comprehensive care provided to National Lung Screening Trial participants, offer screening to people age 55–74 current and former smokers with 30+ pack-year smoking history
Prostate	Prostate Specific Antigen (PSA)	1–2 y	Do not screen men for prostate cancer with PSA test	Screen men aged ≥50 with a life expectancy >10 y after discussion about the risks, benefits, and uncertainties of PSA screening. Follow-up screening should occur annually if PSA >2.5 ng/mL or biennially if PSA <2.5 ng/mL	**AUA[f]** Biennial PSA screening in men of 55–69 y with life expectancy >10–15 y, after shared decision-making discussions accounting for values and preferences
Osteoporosis	Dual-energy x-ray absorptiometry (DEXA) Measure height, preferably with a wall mounted stadiometer	Perform BMD testing 1 to 2 years after initiating medical therapy for osteoporosis and every 2 years annually thereafter	**AAFP[g]** Screening women age 65 or men age 70	**NOF[h]** Screen women age 65 and older and men age 70 and older; postmenopausal women and men age 50–69, based on risk factor profile; postmenopausal women and men age 50 and older who have had an adult age fracture	
Carotid Disease	Carotid ultrasound	Once	**Society of Vascular Surgery**		

			Professional Society Issuing Recommendations		
Type of Screening	**Test**	**Frequency**	**USPSTF[a]**	**ACS[b]**	**ACP[c]**
			Age >65, coronary artery disease, need for coronary bypass, symptomatic lower extremity arterial occlusive disease, history of tobacco use and high cholesterol would be appropriate risk factors to prompt ultrasound in patients with a bruit		
Coronary Artery Disease	Coronary Calcium Score (CCS)	Once	**SCCT[i]**	**AHA/ACC[j]**	
			Do not use CCS for patients with known CAD.	CCS of 0 may have a strong negative predictive value for coronary events in older adults.	
Abdominal Aortic Aneurysm	Abdominal Ultrasound	Once	**USPSTF[a]**	**AAFP[g]**	
			Insufficient evidence to assess the balance of benefits and harms of screening women aged 65–75 years who have ever smoked.	Recommended for men aged 65–75 years who have ever smoked.	
Diabetes	Fasting blood glucose, glucose tolerance test, or HbA1C	Annually	**USPSTF[a]**	**ADA[k]**	
			Abnormal blood glucose as part of cardiovascular risk assessment in adults aged 40–70 years who are overweight or obese	Screening people 45 years and older	

[a]United States Prevention Service Task Force; [b]American Cancer Society; [c]American College of Physicians; [d]Eastern Cooperative Oncology Group; [e]American College of Chest Physicians; [f]American Urology Association; [g]American Academy of Family Physicians; [h]National Osteoporosis Foundation; [i]Society of Computed Tomography; [j]American Heart Association/American College of Cardiology; [k]American Diabetes Association.

VACCINATIONS

The use of vaccines in older adults is aimed at creating immunity against common infections that could lead to serious complications, as well as for rebuilding previously obtained immunity. Currently, the CDC recommends routine vaccination against influenza, pneumococcus and shingles as they are prevalent in this age group. Other countries in Europe and Asia have similar trends on vaccinations with small variances.

SEXUALLY TRANSMITTED DISEASES (STDs)

Although most STDs occur in younger people **(see Chap. 131)**, a portion of older adults have high-risk sexual behavior. Most Americans remain sexually active in their sixties and seventies, and up to a quarter of individuals in their eighties consider themselves sexually active. Sexually active older adults may have a lower awareness of the need for safe sexual practices, such as the risks of multiple sexual partners and condom use. The incidence of STDs in older people is still relatively low. Individuals born in the United States between 1945 and 1965 are at higher risk of having hepatitis C due to lack of awareness of the disease and lack of implementation of universal precautions before the 1980s for blood transfusions. Other factors that could affect such risk are use of intravenous drugs and unprotected sex with multiple partners. The prevalence of tertiary syphilis is higher than newly contracted syphilis in older adults. The incidence of gonococcal infection decreases with age. Nonetheless, patients presenting with symptoms compatible with syphilis or gonococcal infection (cervicitis, urethritis, proctitis, epididymitis) should be screened for high-risk sexual behavior and educated if necessary. Clinical symptoms of herpes simplex infection and the possibility of becoming contagious also decrease with age. As ulcerative lesions are less frequent, HSV-2 specific serologic testing should be considered for patients with recurrent nonspecific genital symptoms. Therapy should not be started unless the patients are symptomatic.

Almost 3 million patients aged >50 live with HIV. Since the introduction of HAART, life expectancy of patients with HIV has increased, resulting in a significant increase in the number of older adults living with the disease. De novo infections have also contributed to the rising number of HIV cases in older adults. The low rate of condom use and lack of knowledge of the disease play a key role in the transmission rate. Age is an independent predictor of HIV progression and associated mortality. There are no specific age-specific guidelines for treating HIV. Like all other conditions, there is a higher incidence of medication-related side effects in older patients, especially those with other comorbidities and on multiple other medications, and this should be considered in treatment decisions.

Treatment of Common Diseases in the Geriatric Population

HYPERTENSION

There have been several clinical trials demonstrating the benefits of hypertension treatment on the reduction of risk of cardiovascular events in older people. Nonetheless, blood pressure targets remain controversial. The balance between the cardiovascular protective benefits versus the risk of treatment-related adverse events must be considered in individual patients based on their comorbidities. For example, hypotension and postural hypotension related to antihypertensive therapy are common causes of near-syncope and falls and related injuries in the geriatric population, especially those with multi-morbidity. On the other hand, control of systolic blood pressure, in addition to preventing cardiovascular events, may reduce the burden of white matter changes in the brain, which are associated with gait abnormalities and falls and cognitive decline. However, no studies in older patients with multi-morbidity to date have documented any beneficial effects of tight control of hypertension on the incidence of falls and cognitive decline.

Two large studies (HYVET and SPRINT) have shed some light on these issues. HYVET was a multicenter study conducted in several countries involving ~3,800 patients ≥80 years old. The study demonstrated that active treatment of hypertension with a target of ≤150 mmHg not only significantly reduced the risk of stroke and heart failure, but also the mortality risk. As with other large hypertension studies like ALLHAT, there was a linear association between blood pressure and stroke reduction. Nonetheless, in the HYVET study, this association was less prominent as age increased. SPRINT was another large randomized trial targeting lowering systolic blood pressure to targets of <140 vs 120 mmHg (measured with an automated device) with a subgroup analysis in those aged ≥75 years. There were significant reductions in the primary end-point—a composite of cardiovascular disease events (including myocardial infarction, acute coronary syndrome, heart failure, stroke, or death from cardiovascular causes). However, it is critical to recognize that patients with diabetes, history of stroke or heart failure, and systolic blood pressure <110 mmHg after 1 min of standing were excluded from the SPRINT trial.

Overall, these data strongly suggest a person-centered approach to hypertension in the heterogeneous older population. For older patients with minimal comorbidity, no postural hypotension, and low risk of falls and volume depletion, the benefit/risk ratio favors lower targets for systolic blood pressure (<130 mmHg measured by a hand sphygmomanometer). However, for those with diabetes, heart failure, history of stroke, postural hypotension, careful treatment of blood pressure with higher systolic targets (< 150 mmHg) is probably a safer approach.

DIABETES

The prevalence of diabetes in the older adult population is now over 25% and expected to increase due to adverse lifestyle changes and an increased incidence of obesity. Due to a lack of data on patients with multi-morbidity and those aged ≥80, and the high incidence of hypoglycemia in this population when treated with multiple hypoglycemic agents, the approach to management of diabetes requires a person-centered approach like that described for hypertension. Older diabetic patients are at significant risk of hypoglycemia because of potential medication errors, progressive decline in renal function, and inconsistent oral intake among other reasons. Hypoglycemic episodes are associated with progressive cognitive decline in older adults, especially those with existing cognitive impairment. On the other hand, uncontrolled diabetes is associated with an increased risk of all-cause dementia.

Data from randomized clinical trials that have largely excluded those aged ≥80 suggest that intensive glycemic control does not reduce major macrovascular events in older adults for at least 10 years or result in improved microvascular outcomes for at least 8 years, and at the same time increases the risk of severe hypoglycemia 1.5- to 3-fold. Thus, the AGS guideline on diabetes in older adults and the Choosing Wisely recommendations (Table 464-2) suggest that in most adults >65 years, the harms associated with a hemoglobin A1c (HbA1c) target <7.5% or >9% are likely to outweigh the benefits. These recommendations are consistent with the American Diabetes Association. Thus, the goals of treating diabetes in the geriatric population should be tailored to the patient's functional and medical status, social support, personal goals, perception of risk, and life expectancy. For specifics of treatment options, refer to **Table 464-6**. Regardless of the therapeutic goals for HbA1c, older diabetic patients should be regularly examined for the development of neuropathy, which can lead to the development of lesions on the feet that could become infected, as well as for retinopathy and vision loss that may require ophthalmologic intervention.

TABLE 464-6

Recommendations and Considerations for Pharmacologic Therapy of Diabetes in Older Adults[a]

Medication	Recommendations and Considerations
Metformin	• Metformin is the first-line agent for older adults with type 2 diabetes • Recent studies suggest it may be used safely in patients with estimated glomerular filtration rate ≥30 mL/min/1.73 m^2 • Contraindicated in patients with advanced renal insufficiency or significant heart failure • May be temporarily discontinued before procedures or during hospitalizations and when acute illness compromises renal or liver function
Sodium–Glucose Cotransporter 2 Inhibitors	• Offer an oral route, which may be convenient for older adults • Long-term experience is limited despite the initial efficacy and safety data
Thiazolidinediones	• If used at all, should be used very cautiously in those with, or at risk for, congestive heart failure and those at risk for falls or fractures
Secretagogues	• Associated with hypoglycemia and should be used with caution • Shorter-duration sulfonylureas such as glipizide are preferred • Glyburide is a longer-duration and contraindicated in older adults
Incretin-Based Therapies	• Few side effects and minimal hypoglycemia, but costs may be a barrier • No evidence of increase in major adverse cardiovascular events • Glucagon-like peptide 1 receptor agonists are injectable, which require visual, motor, and cognitive skills ○ Associated with nausea, vomiting, diarrhea, and weight loss, which may not be desirable in some older patients, particularly those with cachexia
Insulin Therapy	• Requires that patients or their caregivers have good visual and motor skills and cognitive ability • Insulin doses should be titrated to meet individualized glycemic targets and to avoid hypoglycemia • Once-daily basal insulin injection therapy is associated with minimal side effects and may be a reasonable option in many older patients • Multiple daily injections of insulin may be too complex for the older patient with advanced diabetes complications, life-limiting comorbid illnesses, or limited functional status

[a]Based on recommendations from the American Diabetes Association 2017.

HYPERLIPIDEMIA

While good evidence exists regarding the benefits of statins on primary cardiovascular risk prevention in patients ≤75 years old, for those aged >75 the data are very limited. The use of statins in those aged >75 or 80 for prevention of cardiovascular events and mortality is the subject of ongoing debate in the geriatric literature**.** No evidence from randomized controlled trials exists to guide statin initiation after age 80 years; treatment of hypercholesterolemia for patients at risk of atherosclerotic cardiovascular disease should start before they turn 80 years old. There are two other factors that make the use of statins in older adults controversial. First, the major benefits of statins have been demonstrated over long-term use; thus, life expectancy is a limiting factor to observe any meaningful change in outcomes. A substantial proportion of patients are maintained on statins at the end of life, but they can be safely discontinued. On the other hand, statins are well tolerated in older adults especially at moderate to low doses. Although many older adults on statins complain of muscle pain, the risk of myositis and rhabdomyolysis is increased mostly with the use of high doses; adverse effects of statins on cognitive function appear to be uncommon. Thus, some relatively healthy adults aged >75 years with life expectancy of >10 years may benefit, and the approach to hyperlipidemia should be person-centered in this population, as discussed for both hypertension and diabetes.

OSTEOARTHRITIS (OA)

The approach to the management of symptomatic OA in the geriatric population differs from the approach in younger patients **(see Chaps. 363 and 364)** because of the risks of toxicity of nonsteroidal anti-inflammatory drugs (NSAIDs) in older patients. Nonpharmacologic interventions should be the first line of treatment. While some patients aged >65 years can tolerate NSAID use with concomitant protection from gastrointestinal (GI) bleeding with a proton pump inhibitor (PPI), this regimen exposes patients to two drugs with numerous potential adverse drug effects. NSAIDs are well known to be associated with GI bleeding; they are also associated with worsening renal function based on multiple potential mechanisms, and with sodium and fluid retention and exacerbation of hypertension and congestive heart failure. In addition, a substantial number of older patients are on anticoagulants or platelet aggregation inhibitors, which could further increase the

risk of bleeding from NSAIDs. PPIs are associated with a higher incidence of pneumonia, osteoporosis, and *C. difficile*-associated diarrhea; and possibly with a higher risk of dementia.

Topical NSAIDs are better tolerated; lidocaine patches and non-prescription creams may also be effective. The AGS guideline on the management of chronic pain recommends that routine acetaminophen in doses up to 1 g four times daily should be the basis of pharmacologic treatment. Failure to respond could be followed up with careful trials of tramadol or a narcotic agent (started in a short-acting preparation) with appropriate attention to avoiding narcotic-induced constipation. Although prescription of narcotics is getting increasingly cumbersome because of high rates of abuse, this should not deter prescription of these agents to relieve pain and disability in older patients.

Many older patients respond well to a variety of non-pharmacologic interventions, including stretching, strengthening, timely and appropriate use of heat and ice, massage, swimming and whirlpool therapy, bracing, acupuncture, and therapeutic electrical stimulation. These interventions are best carried out under the supervision of physical therapists or other professionals with appropriate expertise to avoid injury. Surgical interventions, including replacement of major joints, have improved over the last several years, and even older patients with multi-morbidity may experience improved function and quality of life. Total knee replacement, for example, has been shown to be effective in generally healthy older patients, and should be considered in selected higher risk patients. “Pre-habilitation” with targeted strengthening and endurance exercises, and willingness to go through several weeks of post-operative physical therapy should be prerequisites for referring older patients for joint replacement.

CANCER

More than half of new cases of cancer and mortality associated with it occur after the age of 65. There are limited data regarding older adults with multiple comorbid conditions and their response to cancer treatment. While only ~10% of clinical trials have had age-stratification analyses, the available evidence suggests that age alone is not a predictor of harm. Nonetheless, making treatment decisions is challenging due to both shorter life expectancy in older adults and the cumulative effect of multiple comorbidities. Thus, a person-centered approach is essential.

Older adults generally experience decreases in functional status after receiving chemotherapy. Most of this negative effect appears to be related to comorbidity and baseline functional status, rather than due to age alone. For this reason, specialists in geriatric oncology have proposed using comprehensive geriatric assessment, including many of the issues addressed in Table 464-3 as a strategy to better predict which older adults will tolerate and benefit most from cancer treatment. Other considerations before making decisions about treatment plans should include socioeconomic factors. Lack of social support has been associated with poor outcomes after radiation and chemotherapy, especially in older women. Other important issues in cancer treatment planning include availability of transportation for treatments, economic and insurance status, the patient’s ability to follow treatment plans, and family and social support available during therapy, when adverse effects and functional decline may occur.

GERIATRIC SYNDROMES AND CONDITIONS

In this section selected geriatric syndromes and conditions likely to be encountered by internists in hospital, clinic, office, PAC and LTC settings are discussed.

Falls

EPIDEMIOLOGY AND IMPACT

Among all geriatric syndromes, falls are probably the most common that internists will encounter. Falls are responsible for potentially devastating consequences for function and quality of life, as well as mortality, in the geriatric population. About one in three older community-dwelling, and one in two older LTC residents fall annually; many more are at risk for falls. The impacts of falls include fear of falling with adverse effects on quality of life, painful injuries including hip and wrist fractures, subdural hematomas, and death. Falls are associated with loss of function and death within the year after a fall. For these reasons, internists should regularly screen older people for falling using questions such as: “Have you fallen in the past year?” “Are you afraid of falling?” “Do you have trouble climbing stairs or rising from chairs?” (Table 464-3).

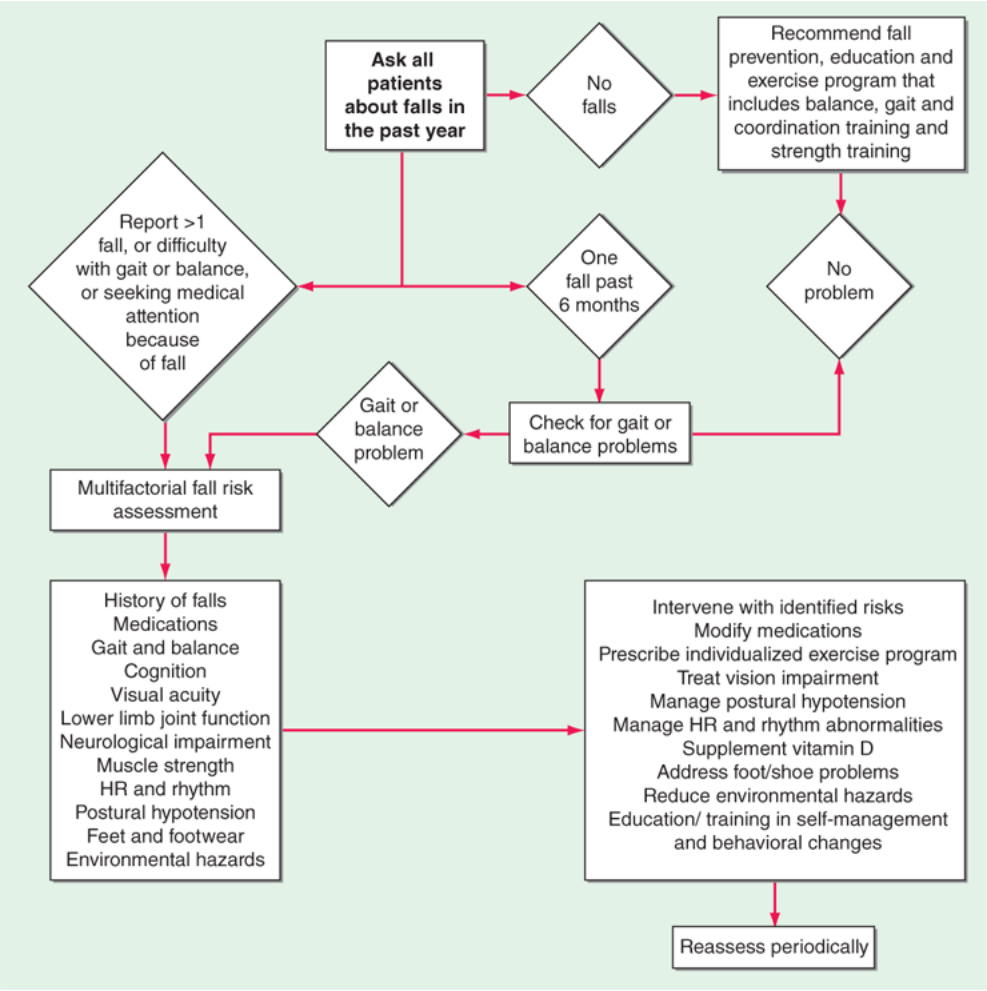
EVALUATION

The risks and causes of falls are multifactorial. Most older people at risk for a fall or who have suffered a fall have more than one potential underlying risk factor or cause. Many falls are labeled as “mechanical” and attributed to simply tripping or slipping. It is essential to recognize, however, that older people who trip or slip may have a variety of underlying reversible conditions that could have contributed to the event. Thus, a thorough evaluation of all falls is warranted. In addition to evaluating the patient who has fallen for injury, it is critical to determine, to the extent possible, whether the patient had a syncopal episode or a seizure, which dictate a very different approach to evaluation and management. As many as half of “unexplained falls” in older people with dementia (e.g., found on the floor) may be due to near-syncope or syncope related to postural hypotension.

Figure 464-5 illustrates an overview of the approach to an older person who reports a history of one or more fall in the past six months, and **Table 464-7** provides more detail on the immediate evaluation of an older person who has fallen. **Chapter 23 provides more detail on the evaluation of gait and balance disorders.**

FIGURE 464-5

Algorithm depicting assessment and management of falls in older patients. HR, heart rate. *(From American Geriatrics Society and British Geriatrics Society: Clinical Practice Guideline for the Prevention of Falls in Older Persons. New York, American Geriatric Society, 2010.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 464-7

Evaluating the Older Person Who Falls—Immediate Post-Fall Evaluation

History

- Circumstances surrounding the fall
  - Relationship to changes in posture, turning of head, after a meal or medication intake, rushing to the toilet, nocturia, straining to urinate or defecate
  - Accidental trip or slip (note that many correctable factors can contribute to a reported “mechanical” fall – see text)
  - Hazards in the living environment (loose rugs; cords; unsafe steps; slippery floors; etc.)
- Premonitory or associated symptoms
  - Dizziness (lightheadedness vs vertigo); cardiovascular (postural lightheadedness, palpitations, chest pain, shortness of breath); focal neurological symptoms suggestive of stroke or transient ischemic attack (weakness, sensory disturbance, dysarthria, ataxia, aphasia); symptoms of a seizure (witnessed clinic movements; incontinence of urine or stool; tongue biting)
  - Symptoms over the previous few days that may have led to volume depletion (poor food/fluid intake; nausea/vomiting; diarrhea; urinary frequency/polyuria)
- Exclude loss of consciousness or seizure (may be difficult without a witness)
- Medications – chronic and within the few hours before the fall
  - Diuretics and other antihypertensive drugs
  - Nitrates
  - Drugs that cause bradycardia – beta blockers; cholinesterase inhibitors
  - Psychotropic - antipsychotics; hypnotics; sedatives; antidepressants
  - Antiparkinsonian
  - Hypoglycemic drugs
- Excessive alcohol intake

Physical Examination

- Exclude physical injury
  - Head trauma; hip range of motion; pubic bone tenderness; wrist pain; other signs of trauma
  - Bruising in patients on anticoagulants/platelet inhibitors
- Exclude acute illness
  - Vital signs
  - Postural vital signs (if feasible/safe)
  - Finger-stick glucose in diabetics
  - Poor skin turgor suggesting volume depletion (over chest; other areas unreliable)
  - Signs of an acute respiratory, cardiovascular, abdominal conditions
  - Focal neurological signs suggestive of stroke
- Signs of conditions that increase risk for falls
  - Poor visual acuity; use of bifocals
  - Limited range of motion of neck (to detect possible cervical arthritis/disk disease)
  - Cardiovascular—arrhythmias; carotid bruits; aortic stenosis; mitral insufficiency; heart failure
  - Degenerative joint disease in lower extremities causing pain, limited range of motion, and/or deformity
  - Podiatric conditions (calluses; bunions; ulcerations; poorly fitted, inappropriate, or unsafe shoes)
  - Neurological signs—lower extremity muscle weakness; peripheral neuropathy; tremor, rigidity, and/or bradykinesia suggestive of undiagnosed Parkinson’s disease; cerebellar signs (abnormal heel to shin or heel tapping); abnormal reflexes that could reflect upper motor neuron disorder such as spinal cord compression or subdural hematoma; cognitive deficits that can result in poor judgement
  - Observation of gait and balance – simple get up and go test (see text) with observation for short steps, poor foot elevation, wide-based gait, multiple steps to turn 180 degrees; other abnormalities that might suggest normal pressure hydrocephalus (especially in combination with symptoms of incontinence and/or cognitive impairment)

Laboratory and/or Imaging Studies

- Should be guided by history and physical examination – common examples include:
  - Complete blood count, basic metabolic panel to exclude/verify acute illness
  - Urinalysis (only when additional symptoms of urinary tract infection present)
  - Electrocardiogram (in patients suspected of acute coronary syndrome or with significant known cardiovascular disease)
  - X-rays to exclude fractures
  - Brain imaging if signs present to exclude subdural hematoma, stroke
  - Cardiac monitoring in patients with history suggestive of syncope or near-syncope
  - Electroencephalography in patients with history suggestive of seizure

MANAGEMENT

Table 464-8 illustrates approaches to the management of falls. Immediately after a fall injuries and underlying acute illnesses should be identified and treated. It is common practice for older patients who come to an Emergency Department with a history of a fall to have a brain imaging study. While this is understandable from a potential liability standpoint, it is also reasonable to avoid such studies if there is no history or signs of head trauma, neurologic symptoms or signs, or anticoagulation, and monitor the patient carefully over the next 48–72 h for the development of specific indications for a brain imaging study.

TABLE 464-8

Examples of Management for Underlying Causes of Falls in Older Patients

Causes	Examples of treatment
Cardiovascular	
Arrhythmias	Antiarrhythmic medication, ablation, pacemaker (depending on nature of arrhythmia)
Aortic stenosis with syncope or near syncope	Valve surgery (trans-catheter procedure if appropriate)
Postural hypotension	Reduce or eliminate hypotensive drugs Hydration, support stockings Medication (proamitine, fludrocortisone, droxidopa) Adaptive behaviors (e.g., pausing and getting up slowly)
Hypertension	Manage carefully to avoid hypotension and near syncope; control may be important in patients with periventricular white matter changes in preventing further gait disturbance
Neurologic	
Autonomic dysfunction with postural hypotension	As above
Cervical spondylosis (with spinal cord compression)	Neck brace; physical therapy; consider surgery
Parkinson disease	Antiparkinsonian drugs
Visual impairment	Ophthalmological/optometric evaluation and specific treatment
Seizure disorder	Anticonvulsants
Normal-pressure hydrocephalus	Surgery (ventricular-peritoneal shunt)
Dementia	Supervised activities Hazard-free environment
Benign positional vertigo	Habituation exercises Anti-vertiginous medication
Others	
Foot disorders	Podiatric evaluation and treatment
Gait and balance disorders	Properly fitted shoes Physical therapy Exercise with balance training (including Tai Chi where available)
Muscle weakness, deconditioning	Lower extremity strength training
Drug overuse (eg, sedatives, alcohol, other psychotropic drugs, antihypertensive)	Elimination of drug(s) when feasible
Vitamin D deficiency	Vitamin D supplementation
Recurrent falls	Fall alert system for those who live alone; hip protectors in selected patients

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

Because the causes of and risk factors for falls are often multifactorial, management commonly requires multiple interventions in the same patient. Among the most common and effective interventions are physical therapy for strengthening and balance; Tai Chi has also been shown to be effective in multiple trials. Although many older people who fall are vitamin D–deficient, the role of vitamin D replacement in preventing falls, or preventing injuries from falls when combined with interventions such as strength and balance training is not clear. The risk/benefit ratio probably favors vitamin D replacement with at least 800 IU per day, but high

dose vitamin D (60,000 IU in one oral dose monthly) has been associated with an increase in risk of falls. Patients who suffer a fracture after a fall should be investigated and treated for osteoporosis. Patients at high risk for recurrent falls and injuries should be encouraged to use a fall alert system; selected patients may benefit from hip protectors.

Polypharmacy

EPIDEMIOLOGY AND IMPACT

Polypharmacy has been defined as the prescription of multiple medications using various thresholds (generally ranging from five up to nine simultaneous drugs), and has been identified as a major challenge in the geriatric population for decades. About 40% of the U.S. population aged ≥65 take 5–9 medications, and ~20% take 10 or more. Polypharmacy is an increasingly complex challenge because of the rising prevalence of multi-morbidity, a plethora of clinical practice guidelines, proliferation of medications that can effectively treat common geriatric conditions, and rising patient and family demand for medications due in part to television advertising and information available on the internet. For example, based on several condition specific clinical practice guidelines (which do not account for multi-morbidity), an 80-year-old person with multi-morbidity including diabetes, chronic obstructive lung disease, hypertension, osteoporosis, and degenerative joint disease might be prescribed an extremely complicated non-pharmacologic regimen and over a dozen medications with the potential for multiple drug-drug and drug-disease interactions.

Polypharmacy increases the risks associated with age-related changes in the pharmacology of many drugs, and the risk of adverse drug events (ADEs). ADEs cause over 100,000 hospitalizations per year; the main culprits are warfarin and other antiplatelet agents, and insulin and other hypoglycemic agents. Other categories of drugs are also involved, including cardiovascular drugs that can cause electrolyte and volume disturbances and hypotension, falls, and syncope; central nervous system drugs associated with altered mental status and falls; and antimicrobials which cause allergic reactions, diarrhea, and other ADEs.

EVALUATION

All older patients should have careful medication reconciliation at each office or clinic visit, and especially at the time of care transitions, including acute hospitalization, hospital discharge, admission to a PAC facility or home health program, and discharge from a PAC facility to home. At each transition, all medications should be considered in terms of unclear diagnosis or indication, uncertain dose or route of administration, stop date, hold parameters, lab tests needed for monitoring, dosages different than the last care setting, medication duplication, medications that should be restarted, and the potential for drug-drug and drug-disease interactions. At each clinic or office visit for community-dwelling older people, possible ADEs, effectiveness of drug therapy, and adherence should be evaluated.

MANAGEMENT

Table 464-9 lists several general recommendations for geriatric prescribing that should help make drug therapy more effective and safer in older patients, especially those with multi-morbidity. Chapter 63 also provides information on general principles of clinical pharmacology. Because these patients often see multiple specialists, the internist should serve as the “quarterback” for all prescribing to help ensure adherence and minimize the potential for ADEs. In hospital, PAC, and LTC settings clinical pharmacists can be extremely helpful in achieving these recommendations and goals.

TABLE 464-9

General Recommendations for Geriatric Prescribing

1. Evaluate geriatric patients thoroughly to identify all conditions that could (a) benefit from drug treatment; (b) be adversely affected by drug treatment; and (c) influence the efficacy of drug treatment
2. Manage medical conditions without drugs as often as possible
3. Know the pharmacology of the drug(s) being prescribed
4. Consider how the clinical status (e.g., renal function, hydration) of each patient could influence the pharmacology of the drug(s)
5. Avoid potentially serious adverse drug-drug interactions
6. For drugs or their active metabolites eliminated predominantly by the kidney, use a formula to approximate age-related changes in renal function and adjust dosages accordingly—the Cockcroft-Gault formula (below) is probably safer as it tends to underestimate creatinine clearance

$$\text{Creatinine clearance} = \frac{(140 - \text{age}) \times \text{body weight (kg)}}{72 \times \text{serum creatinine level}} \; (\times 0.85 \text{ for women})$$

7. If there is a question about drug dosage, start with smaller doses and increase gradually until the drug is effective or intolerable side effects are observed
8. Drug blood concentrations can be helpful in monitoring several potentially toxic drugs used in the geriatric population
9. Help to ensure adherence by:
   a. Making drug regimens and instructions as simple as possible
   b. Using the same dosage schedule for all drugs whenever feasible (e.g., once or twice per day)
   c. Timing the doses in conjunction with a daily routine
   d. Paying attention to impaired cognitive function, diminished hearing, and poor vision when instructing patients and labeling prescriptions
   e. Instructing relatives and caregivers on the drug regimen
   f. Enlisting other health professionals (e.g., home health aides, pharmacists) to help ensure compliance
   g. Making sure the older patient can get to a pharmacist (or vice versa), can afford the prescriptions, and can open the container
   h. Using aids (such as special pillboxes and drug calendars) whenever appropriate
   i. Performing careful medication adjudication and patient/family education at the time of every hospital discharge
   j. Keeping updated medication records and review them at each visit
   k. Reviewing knowledge of and adherence with drug regimens regularly
10. Monitor older patients frequently for adherence, drug effectiveness, and adverse effects, and adjust drug therapy accordingly

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

While there may be undertreatment of certain conditions in older people (such as osteoporosis, depression, and overactive bladder), more attention is now being paid to “de-prescribing.” De-prescribing must be done carefully, especially at the time care transitions, when indications for specific drugs and patient preferences may not be clear. The American Geriatrics Society’s updated Beers Criteria includes a comprehensive list of drugs that may be inappropriate in older people and the rationale for this rating. The STOPP criteria are also useful in identifying drugs that should be reconsidered on older people.

Several commonly prescribed drugs should be considered for “de-prescribing” efforts, including: (1) diuretics and hypotensive agents when patients have systolic hypotension or postural hypotension that can precipitate near-syncope and falls; (2) over-reliance on antianxiety and hypnotic medications, especially benzodiazepines; (3) psychotropic and other drugs with anticholinergic activity that can cause dry mouth, constipation, and increase the long-term risk of cognitive impairment; (4) PPIs with unclear indications because of numerous reported potential ADEs including increased risk of pneumonia, osteoporosis, and dementia; (5) cholinesterase inhibitors and memantine in patients with severe cognitive impairment who have been on them for years: and (6) hypoglycemic agents in patients with multi-morbidity who should not have tightly controlled blood sugar with increased risk of hypoglycemia; and (7) statins in patients with severe chronic illness who are near the end of life.

Careful de-prescribing is a critical aspect of person-centered care in the geriatric population. Several general principles, including some in Table 464-9, may assist with de-prescribing efforts, including: (1) ascertain all drugs the patient is currently taking and the reasons for each one; (2) consider overall risk of drug-induced harm in individual patients in determining the required intensity of deprescribing intervention; (3) assess each drug as to its current or future benefit potential compared with current or future harm or burden potential; (4) prioritize drugs for discontinuation that have the lowest benefit-harm ratio and lowest likelihood of adverse withdrawal reactions or disease rebound syndromes; and (5) implement a discontinuation regimen based on the pharmacology of the drug being discontinued, and monitor patients closely for improvement in outcomes or onset of adverse effects.

Cognitive Impairment—Delirium and Dementia

The reader is referred to other chapters in this text **(Chapters 423–426)** for detailed information on delirium and dementia in the overall population and among older patients specifically.

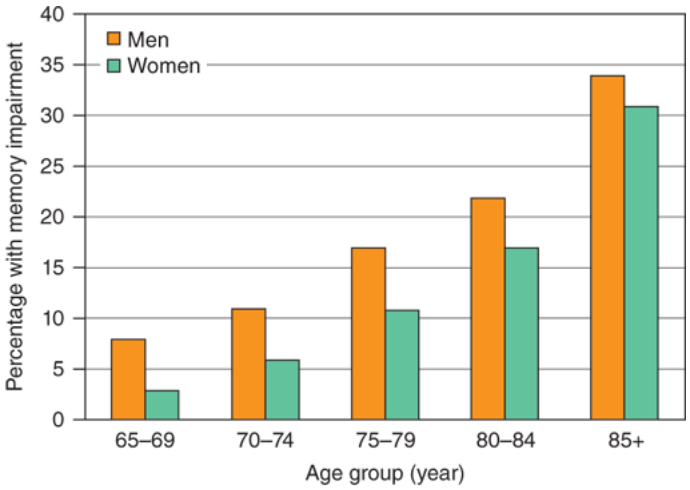
EPIDEMIOLOGY AND IMPACT

Delirium occurs in up to 40% of hospitalized older patients, and is associated with increased morbidity, need for institutional care, and mortality in this population. While most episodes of delirium clear within a few days if the underlying cause(s) are identified and treated, delirium may persist for weeks, and in a few cases for months, after an acute hospitalization.

Normal aging does not cause impairment of cognitive function of sufficient severity to render an individual dysfunctional, which is the hallmark of a dementia syndrome. Slowed thinking and reaction time, mild recent memory loss, and impaired executive function can occur with increasing age and may or may not progress to dementia. Figure 464-6 illustrates the prevalence of memory impairment with increasing age. Just over 20% of people aged >70 in the United States have cognitive impairment without dementia (generally referred to as "Mild cognitive impairment [MCI])." Up to 15 to 20% of those diagnosed with MCI will progress to dementia over the course of a year; thus, most people with MCI will progress to dementia within 5 years. Therapeutic implications of MCI are subjects of intensive research. No non-pharmacological or pharmacological intervention has been shown to prevent progression to dementia.

FIGURE 464-6

Rates of memory impairment in different age groups. The definition of "moderate or severe memory impairment" is 4 or fewer words recalled out of 20. *(Source: Health and Retirement Survey; details can be found at: http://hrsonline.isr.umich.edu.ezproxy1.library.usyd.edu.au/. Accessed December 30, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

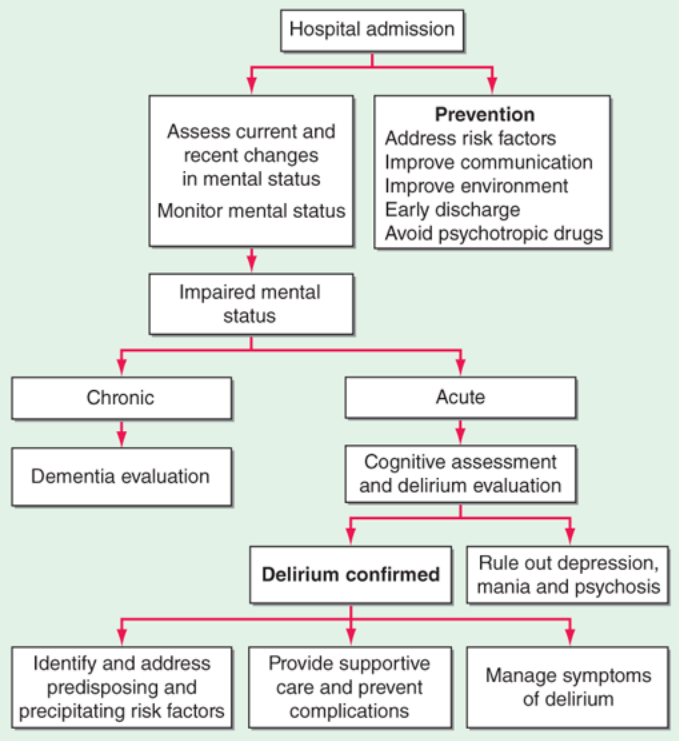
The definitions of dementia syndromes (now also termed "major neurocognitive disorder") have been updated by the American Psychiatric Association. The prevalence of dementia increases with age; by age 85 between 30 and 40% have a dementia syndrome. Alzheimer's disease and vascular dementia, which often occur together based on pathologic studies, account for most dementias in older people. Dementia with Lewy bodies accounts for up to 25% of dementia, and is characterized by Parkinsonian features early in the disease (as opposed to dementia in Parkinson's disease, which generally occurs years after the onset of Parkinson's), personality changes, alterations in alertness and attention, and visual hallucinations that can cause paranoia. Most dementia syndromes are slowly progressive over several years; however, dementia is a terminal illness among patients who do not succumb to other comorbidities, and results in devastating loss of cognition and function in the later stages.

EVALUATION

Regardless of setting, the new onset of delirium should be treated as a medical emergency because it can be the manifestation of an underlying critical illness. Figure 464-7 illustrates an overview of the assessment and management of impaired mental status and delirium in older hospitalized patients. The most validated evaluation for delirium is the Confusion Assessment Method, which requires an acute onset and fluctuating course *and* inattention *and* disorganized thinking *or* altered level of consciousness. Because the causes and risk factors for delirium are multifactorial, evaluation requires a careful history, physical examination, and selected laboratory studies based on the findings.

FIGURE 464-7

Algorithm depicting assessment and management of delirium in hospitalized older patients. *(Modified from SK Inouye: N Engl J Med 354:1157, 2006.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The benefits of screening older adults for cognitive impairment are controversial, but there are many interventions that may benefit patients and families early in the course of the disease (see below). Older patients in outpatient settings with complaints (or family reports of) early signs of cognitive impairment benefit from neuropsychological testing, which can help differentiate MCI and dementia, and identify concomitant factors such as depression and anxiety. The “Mini-Cog” is a sensitive screening tool for cognitive impairment, and consists of a 3-item recall test and clock drawing. Further evaluation of dementia includes a comprehensive history and physical examination, functional status assessment (since the diagnosis depends on impaired function), a brain imaging study, and selected laboratory tests, including a complete blood count, comprehensive metabolic panel, thyroid function tests, vitamin B12 level, and, if suspected tests for syphilis and human immunodeficiency virus antibodies.

MANAGEMENT

Table 464-10 lists non-pharmacologic management strategies for various underlying risk factors and causes of delirium. Every attempt should be made to avoid or discontinue any medication that may be worsening cognitive function in a delirious geriatric patient. This may not be possible, and in some patients, psychotropic drugs may be needed to treat delirium if the patient is a danger to themselves or others. Low dose haloperidol (0.25–0.5 mg) is generally recommended; more sedating antipsychotics and benzodiazepines should be avoided unless the goal is to put the patient to sleep for a short time. If a benzodiazepine is used, it should be short-acting and in a low dose.

TABLE 464-10

Interventions for Risk Factors for Delirium

Risk Factor	Intervention Protocol
Cognitive impairment	• Orienting communication, including orientation board • Therapeutic activities program
Immobilization	• Early mobilization (e.g., ambulation or bedside exercises) • Minimizing immobilizing equipment (e.g., restraints, bladder catheters)
Psychoactive medications	• Restricted use of PRN sleep and psychoactive medications (e.g., sedative–hypnotics, narcotics, anticholinergic drugs) • Nonpharmacological protocols for management of sleep and anxiety
Sleep deprivation	• Noise-reduction strategies • Scheduling of night time medications, procedures, and nursing activities to allow uninterrupted period of sleep.
Vision impairment	• Provision of vision aids (e.g., magnifiers, special lighting) • Provision of adaptive equipment (e.g., illuminated phone dials, large-print books)
Hearing impairment	• Provision of amplifying devices; repair hearing aids • Instruct staff in communication methods
Dehydration	• Early recognition and volume repletion

Source: Data from SK Inouye et al: A clinical trial of a multicomponent intervention to prevent delirium in hospitalized older patients. N Engl J Med 340:669, 1990.

Although the benefits of screening for cognitive impairment in older people are controversial, there are many non-pharmacologic interventions of older patients, their families, and other caregivers that may be beneficial **(Table 464-11)**. There are four basic approaches to the pharmacological treatment of dementia: (1) avoid drugs that can worsen cognitive function, mainly those with strong anticholinergic activity; (2) agents that enhance cognition and function; (3) drug treatment of coexisting depression, which is common throughout the course of dementia; and (4) pharmacological treatment of complications such as paranoia, delusions, psychosis, and behavioral symptoms such as agitation (verbal and physical). The use of antipsychotics to treat the neuropsychiatric symptoms of dementia is highly controversial. Most experts and guidelines recommend avoiding these drugs and using nonpharmacological strategies unless patients are a danger to themselves and others or if nonpharmacological interventions have failed. Patients with new or worsening behavioral symptoms associated with dementia should have a medical evaluation to identify potentially treatable precipitating conditions. Pain may be especially hard to detect, and if suspected, a therapeutic trial of acetaminophen should be considered.

TABLE 464-11

Key Principles in the Management of Dementia

Optimize the patient's physical and mental function through physical activity and mind plasticity principles and activities
Treatment underlying medical and other conditions (e.g., hypertension, Parkinson disease, depression
Avoid use of drugs with central nervous system side effects unless required for management of psychological or behavioral disturbances
Assess the environment and suggest alterations, if necessary
Encourage physical and mental activity
Avoid situations stressing intellectual capabilities; use memory aids whenever possible
Prepare the patient for changes in location
Emphasize good nutrition
Identify and manage behavioral symptoms and complications
Driving (consider a formal driving evaluation)
Wandering
Dangerous driving
Behavioral disorders
Depression
Agitation or aggressiveness
Psychosis (delusions, hallucinations)
Malnutrition
Incontinence
Provide ongoing care
Reassessment of cognitive and physical function
Treatment of medical conditions
Provide information to patient and family
Nature of the disease
Extent of impairment
Prognosis
Provide social service information to patient and family
Local Alzheimer's Association
Community health-care resources (day centers, homemakers, home health aides)
Legal and financial counseling
Use of advance directives
Provide family counseling for:
Setting realistic goals and expectations
Identification and resolution of family conflicts
Handling anger and guilt
Decisions on respite or institutional care
Legal concerns
Ethical concerns
Consideration of palliative and hospice care
Protect the caregiver from effects of caregiver stress

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

The effectiveness of cholinesterase inhibitors and memantine in improving function and quality of life in patients with various types of dementia is controversial, and the potential benefits of these drugs versus their risks and costs must be weighed carefully to provide optimal person-centered care. The best evidence for effectiveness of cholinesterase inhibitors is in delaying progression of Alzheimer's disease and increasing the time before institutional placement is needed. Gastrointestinal side effects can be problematic and include nausea, vomiting, and diarrhea; nightmares can be bothersome as well. In addition to these bothersome side effects, cholinesterase inhibitors can cause bradycardia, and have been associated with syncope, injurious falls, and pacemaker placement. Memantine can cause dizziness, headache, confusion, and constipation. In one study, vitamin E was more effective than memantine in preventing functional decline in patients with Alzheimer's disease.

Urinary Incontinence

EPIDEMIOLOGY AND IMPACT

Urinary incontinence is curable or controllable in many geriatric patients, especially those who have adequate mobility and mental functioning. Even when it is not curable, incontinence can be managed in a manner that keeps people comfortable, makes life easier for caregivers, and minimizes the costs of caring for the condition and its complications. Approximately one in three women and 15 to 20% of men aged >65 years have some degree of urinary incontinence. Between 5 and

10% of community-dwelling older adults have incontinence more often than weekly and/or use a pad for protection from urinary accidents. The prevalence is as high as 60–80% in many nursing homes, where residents often have both urinary and stool incontinence. Many older people (~40%) suffer from "overactive bladder," which may or may not include symptoms of incontinence. Symptoms of overactive bladder include urinary urgency (with or without incontinence), urinary frequency (voiding every two hours or more often), and nocturia (awakening at night to void). If nocturia alone is the predominant symptom, the patient should be asked about sleep disorders (see section that follows). The pathophysiology, evaluation, and management of overactive bladder are essentially the same as for urge urinary incontinence.

Incontinence is associated with social isolation and depression, and can be a precipitating factor in the decision to seek nursing home care when it cannot be managed in a manner that maintains hygiene and safety. In addition to predisposing to skin irritation and pressure ulcers, the most important potential complication of urinary incontinence and overactive bladder are falls and resultant injuries related to rushing to get to a toilet. Older people with gait disorders, especially those who have multiple episodes of nocturia or nocturnal incontinence, are at especially high risk for injuries. In addition to the bother of the condition to the older person or a caregiver, fall risk is a compelling reason for undertaking a diagnostic evaluation and specific treatment for incontinence and overactive bladder in the geriatric population.

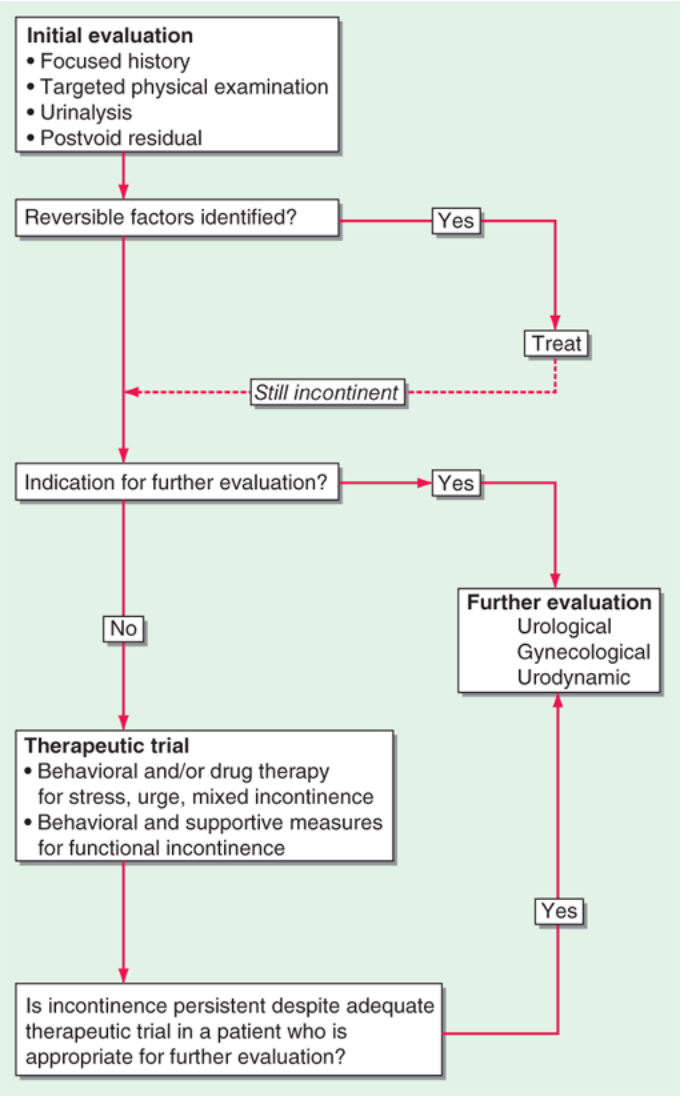
EVALUATION

Internists should ask older people about symptoms of urinary incontinence because these symptoms are often hidden out of embarrassment or fear. Simple questions can help identify incontinent patients, such as: "Do you have trouble with your bladder?" "Do you ever lose urine when you don't want to?" "Do you ever wear padding to protect yourself in case you lose urine?" (Table 464-3). A substantial number of older people will respond "no" to the first two questions, but "yes" to the third one.

Figure 464-8 illustrates an overall approach to the evaluation of the older patient with symptoms of incontinence and overactive bladder. The history and physical examination should focus on identifying potentially reversible causes and contributing factors **(Table 464-12)** and identifying the specific lower urinary tract symptoms. A simple, 3-item validated questionnaire can assist in distinguishing between the most common types of incontinence (Fig. 464-9). Among older women, the most common symptoms are a mixture of urge and stress incontinence (Fig. 464-10); urge is usually the more bothersome. Stress incontinence can often be objectively observed during a physical examination with a comfortably full bladder by having the patient cough in the standing position; leakage of urine simultaneously with coughing indicates that stress incontinence is present. Older men commonly have symptoms associated with overactive bladder and/or symptoms of voiding difficulty (hesitancy, poor or intermittent urinary stream, post-void dribbling); the overactive bladder symptoms are usually more bothersome. These symptoms overlap with those of both benign and malignant disorders of the prostate, and many internists may choose to consult a urologist for further management **(see Chap. 83)** because a urinary flow rate and post-void residual determination (PVR), and further evaluation if malignancy is suspected, are helpful in determining therapy.

FIGURE 464-8

Algorithm for the basic evaluation and management of geriatric urinary incontinence in outpatient practice. *(Adapted from RL Kane et al: Essentials of Clinical Geriatrics, 8th ed. New York, McGraw-Hill, 2018.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 464-9

The 3 Incontinence Questions (3IQ) Assessment Tool. *(From JS Brown et al: The sensitivity and specificity of a simple test to distinguish between urge and stress urinary incontinence. Ann Intern Med 144:715, 2006.)*

The 3IQ is a patient questionaire that helps your doctor distinguish urge incontinence from stress incontinence. It should take no more than a couple of minutes. Complete the quiz and bring it to your next appointment.

1. During the last 3 months, have you leaked urine (even a small amount)?

☐ **Yes** ☐ **No** (if this response is marked, the 3IQ test is complete)

2. During the last 3 months, did you leak urine (check all that apply):

☐ When you were performing some physical activity, such as coughing, sneezing, litting, or exercising?

☐ When you had the urge or the feeling that you needed to empty your bladder, but you could not get to the toilet fast enough?

☐ Without physical activity and without sense of urgency?

3. During the last 3 months, did you leak urine most often (check only one):

☐ When you were performing some physical activity, such as coughing, sneezing, litting, or exercising?

☐ When you had the urge or the feeling that you needed to empty your bladder, but you could not get to the toilet fast enough?

☐ Without physical activity and without sense of urgency?

☐ About equally as often with physical activity as with a sense of urgency?

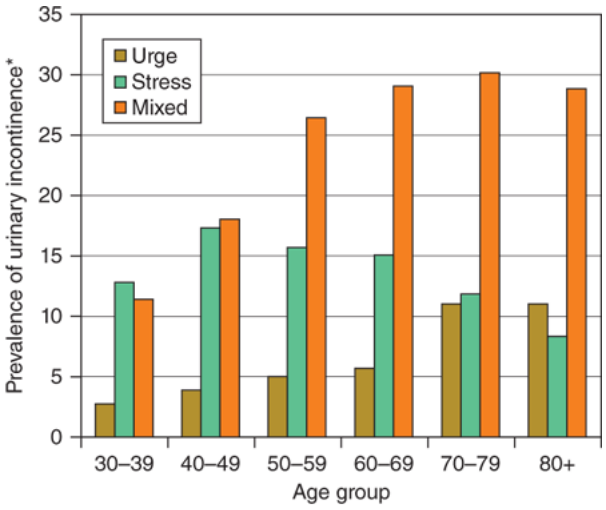
Definitions of type of urinary incontinence are based on responses to question 3.

Response to Question 3	Type of incontinence
Most often with physical activity	Stress only or stress predominant
Most often with the urge to empty the bladder	Urge only or urge predominant
Without physical activity or sense of urgency	Other cause only or other cause predominant
About equally with physical activity and sense of urgency	Mixed

Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 464-10

Rates of urge, stress, and mixed incontinence, by age group, in a sample of 3552 women. *Based on a sample of 3553 participants. *(From JL Melville et al: Arch Intern Med 165:537, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 464-12

Reversible Conditions That Cause or Contribute to Urinary Incontinence and Overactive Bladder Symptoms in Older People

Condition	Management
Lower urinary tract conditions	
Urinary tract infection (symptomatic with frequency, urgency, dysuria, etc.)	Antimicrobial therapy
Atrophic vaginitis/urethritis	Topical estrogen (not a primary treatment for incontinence but may help prevent recurrent infections and ameliorate symptoms of overactive bladder; oral estrogens can cause or worsen incontinence)
Stool impaction with irritation of bladder/urethral innervation and/or partial bladder outlet obstruction	Dis-impaction; appropriate use of stool softeners, bulk-forming agents, and laxatives if necessary; implement
Increased urine production	
Metabolic (hyperglycemia, hypercalcemia)	Better control of diabetes mellitus Therapy for hypercalcemia depends on underlying cause
Excess caffeine or fluid intake	Reduction in intake caffeinated beverages; reduction in fluid intake (most older people with incontinence or overactive bladder self-restrict fluid intake)
Volume overload with increased urine production at night	Support stockings
Venous insufficiency with edema	Leg elevation Sodium restriction Diuretic therapy (late afternoon dose may be effective)
Congestive heart failure	Medical therapy
Impaired ability or willingness to reach a toilet	
Delirium	Diagnosis and treatment of underlying cause(s)
Chronic illness, injury, or restraint that interferes with mobility	Regular toileting Use of toilet substitutes Environmental alterations (eg, bedside commode, urinal) Remove restraints if possible
Psychological (depression, anxiety)	Appropriate non-pharmacologic and/or pharmacologic treatment
Drug Side Effects	Remove offending drug(s) if feasible; modification of dose, frequency or timing may also reduce symptoms for some drugs: Diuretics (polyuria, frequency, urgency) Anticholinergics (constipation, incomplete bladder emptying) Psychotropic drugs Tricyclic antidepressants (anticholinergic effects) Antipsychotics (immobility, sedation) Sedative-hypnotics (immobility, sedation) Narcotic analgesics (constipation, incomplete bladder emptying) Alpha-Adrenergic blockers (urethral relaxation) Alpha-Adrenergic agonists (urethral contraction and potential incomplete bladder emptying) Cholinesterase inhibitors (urinary frequency, urgency) Angiotensin-converting enzyme inhibitors (cough precipitating stress incontinence) Calcium channel blockers, gabapentin, pregabalin, glitazones (edema with nocturia) Alcohol (polyuria, frequency, urgency, sedation, delirium, immobility) Caffeine (polyuria, bladder irritation)

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

Most older patients with symptoms of incontinence should have a PVR, especially men, diabetics, those with neurological disorders, and those with symptoms of voiding difficulty, because incomplete bladder emptying is common in older patients and is difficult to detect by history and physical examination alone. There is no specific cutoff for an abnormal PVR; the test must be done with a full bladder and straining during the test can alter the results. In older patients, a PVR between 0 and 100 mL is normal, between 100 and 200 mL must be interpreted based on symptoms, and a PVR higher than 200 is abnormal and usually influences treatment.

MANAGEMENT

Patients who meet certain criteria should be referred for further urological, gynecological, and/or urodynamic evaluation before initiating specific therapy. Examples include history of lower urinary tract surgery or radiation or recurrent symptomatic urinary tract infections, marked pelvic prolapse on physical examination of a woman, suspected prostate cancer, and sterile hematuria.

Potentially reversible conditions should be addressed, including the many types of medications that can affect bladder function, which should be eliminated if possible (Table 464-12). **Table 464-13** lists treatments for different types of incontinence. Many patients respond well to properly taught and adhered to behavioral interventions. Physical therapists and nurses who specialize in treating lower urinary tract symptoms can be very helpful and should be consulted if available. Pharmacologic treatment of incontinence and overactive bladder is dictated by the innervation of the lower urinary tract. Alpha-adrenergic stimulation increases tone in the smooth muscle of the urethra, thus alpha agonists have been used to treat stress incontinence in women, and alpha blocker are used to decrease urethral tone in men with overactive bladder associated with prostate enlargement. Anticholinergic/antimuscarinic agents and beta-3 stimulation inhibit bladder contraction and are used for overactive bladder and urge incontinence. Patients with severe cognitive impairment and/or immobility can generally be managed effectively by prompted voiding and/or incontinence undergarments, as long comfort, dignity, and safety are maintained.

TABLE 464-13

Primary Treatments for Different Types of Geriatric Urinary Incontinence

Type of Incontinence	Primary Treatments
Stress	Pelvic muscle (Kegel) exercises Other behavioral interventions including timed voiding and double voiding to avoid residual urine α-Adrenergic agonist (none are FDA approved for this purpose) Topical estrogen to strengthen periurethral tissue (not effective alone; oral estrogens contraindicated) Periurethral injections to provide bulking and support Surgical bladder neck suspension or sling for severe incontinence, based on patient preference
Urge and overactive bladder symptoms	Pelvic muscle (Kegel) exercises Other behavioral interventions-—timed voiding and double voiding to avoid residual urine Antimuscarinic and beta-3 adrenergic drugs
Incontinence with incomplete bladder emptying	α-Adrenergic antagonists (men) Bladder training, double voiding Intermittent catheterization Indwelling catheterization in selected patients in whom risks and discomforts of urinary retention outweigh risks of a chronic indwelling catheter
Incontinence with impaired physical and/or cognitive function	Behavioral interventions (prompted voiding, habit training) Environmental manipulation including use of urinal or bedside commode, safe lit path to bathroom Incontinence undergarments and pads

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

Sleep Disorders

Sleep disorders are discussed in more detail for the general adult population in **Chap. 27**. Because they are so common and have some unique features in older patients, they are discussed briefly here.

EPIDEMIOLOGY AND IMPACT

Aging is associated with multiple changes in sleep architecture as well as multiple diseases and disorders that can disrupt sleep. Thus, complaints of sleep difficulty are common in older adults. Consequences of sleep difficulty include lower health-related quality of life, increased medication use, more cognitive decline, and greater health care utilization. Four types of primary sleep disorders are common in the geriatric population: insomnia, sleep disordered breathing due to obstructive sleep apnea (OSA), restless leg syndrome (RLS), periodic leg movements in sleep (PLMS). Complaints of bothersome insomnia—the inability to fall asleep or stay asleep despite a conducive environment—increase with age and occur in ~30% of people aged >65. Insomnia is commonly associated with depression,

anxiety, alcohol intake, and ingestion of caffeinated beverages later in the day. OSA occurs in ~10% of older adults, but is probably under-reported and under-diagnosed. It is associated with medical comorbidities, such as obesity and congestive heart failure. RLS occurs in 5–10% of adults, and its prevalence increases in those aged >70. It is almost twice as common in women than men; family history, iron deficiency, and intake of antihistamines and most antidepressants are risk factors. PLMS can be found in up to 45% of older people, but is often of unknown clinical consequence and remains undiagnosed.

EVALUATION

Older people should be screened for sleep difficulty with questions such as: "Do you often feel sleepy during the day?" "Do you have difficulty falling asleep at night?" Further evaluation of the nature and impact of the complaints can be accomplished with standardized questionnaires (Table 464-3). Patients with significant sleep complaints should be asked about conditions that can interrupt sleep, such as nocturia, gastroesophageal reflux, chronic pain, and caffeine and alcohol intake. Specific questions characterizing the complaints should include inquiring about loud snoring (for OSA), the urge to move legs associated with uncomfortable sensations (RLS), and leg movements during sleep (PLMS; which may result in kicking a bed partner).

MANAGEMENT

Patients suspected for OSA, RLS, or PLMS should be referred for formal sleep evaluation. While hypnotics are among the most commonly prescribed drugs in the geriatric population, non-pharmacologic management of sleep should be the initial and primary approach, as many patients can benefit from properly taught and adhered to interventions (Table 464-14). Benzodiazepine hypnotics should be avoided whenever feasible because they are associated with next-day hangover effects, which may manifest as cognitive impairment and can precipitate falls and car crashes; and rebound insomnia.

TABLE 464-14

Non-Pharmacologic Management of Insomnia in Older Adults

Sleep Hygiene Rules
Check effect of medication on sleep and wakefulness Avoid caffeine, alcohol, and cigarettes after lunch Limit liquids in the evening Keep a regular bedtime–waketime schedule Avoid naps or limit to 1 nap a day, no longer than 30 min Spend time outdoors (without sunglasses), particularly in the late afternoon or early evening Exercise—but limit exercise immediately before bedtime
Instructions for Stimulus-Control Therapy
Only go to bed when tired or sleepy If unable to fall asleep within 20 min, get out of bed (and bedroom if possible); while out of bed, do something quiet and relaxing Only return to bed when sleepy If unable to fall asleep within 20 min, again get out of bed Repeat these behaviors until able to fall asleep within a few minutes Get up at the same time each morning (even if only a few hours of sleep) Avoid naps

Source: Adapted from JB Halter et al (eds): *Hazzard's Geriatric Medicine and Gerontology*, 7th ed. New York, McGraw-Hill, 2017.

Elder Abuse and Neglect

EPIDEMIOLOGY AND IMPACT

The incidence of elder abuse and neglect, and self-neglect are unknown because it is often not asked about or reported. The best data suggest that the incidence over 12 months is at least 8–10%. Abuse and neglect can result in physical injuries and related pain, worsening of chronic medical conditions, dehydration and pressure ulcers, emotional distress, and loss of income and savings and related consequences.

EVALUATION

Because abuse and neglect are so under-reported, may be unsuspected, and have such devastating consequences, older adults should be screened (in the absence of caregivers) with questions such as: "Do you ever feel unsafe where you live? "Has anyone ever threatened or hurt you?" "Has anyone been taking your money without your permission?" (Table 464-3). **Table 464-15** outlines the definitions, symptoms and signs, and key aspects of evaluating suspected abuse and neglect.

TABLE 464-15

Elder Abuse and Neglect

Category	Definition and Examples	Symptoms and Signs	Key Aspects of Evaluation
Physical Abuse	Acts of violence that may result in pain, injury, or impairment • Pushing, slapping, hitting, force-feeding • Improper positioning or use of restraints • Improper use of medications	Abrasions Lacerations Bruises Fractures Use of restraints Burns Pain Depression Delirium or onset or worsening of dementia-related behavioral symptoms	The interview should be conducted alone with the patient; it may reveal discordant histories or findings inconsistent with the history provided by the caregiver. Ankles and wrists should be examined for abrasions suggestive of the use of restraints. Findings that are discordant with the mechanism of injury reported or multiple injuries in various stages of healing should raise the suspicion of abuse. Injuries to the head, neck, and upper arms occur in victims of physical elder abuse, but must be distinguished from accidental injuries. Jaw and zygomatic fractures are more likely to be sustained from a punch than from a fall, which more typically result in fractures to orbital and nasal bones.
Psychological or Verbal Abuse	Conduct that causes mental or emotional distress • Verbal harassment or intimidation • Threats of punishment or deprivation • Isolation	Direct observation of verbal abuse Subtle signs of intimidation, such as deferring questions to a caregiver or potential abuser Evidence of isolation Depression, anxiety, or both	Assess the size and quality of the patient's social network (beyond the suspected abuser). Conduct standardized assessments of depression, anxiety, and cognition, directly or through referral. Ask specifically about verbal or psychological abuse with questions such as "Does your relative/caregiver ever yell or curse at you?"; "Have you been threatened with being put into a nursing home?"; or "Are you ever prevented from seeing friends and family members whom you wish to see?"
Financial Abuse	Misuse of the person's income or resources for the financial or personal gain of a caregiver or advisor • Stealing money or possessions • Denying a home • Coercing to sign contracts or spend money	Inability to pay for medicine, medical care, food, rent, or other necessities Failure to renew prescriptions, adhere to medication regimens or other treatments, or keep medical appointments Malnutrition, weight loss, or both, without an obvious medical cause Evidence of poor financial decision making Firing of home care or other service providers by abuser Unpaid utility bills Initiation of eviction proceedings	Ask about financial exploitation with questions such as "Has money or property been taken from you without your consent?"; "Have your credit cards or automated-teller-machine card been used without your consent?"; and "At the end of the month, do you have enough money left for food and other necessities?" Abrupt changes in financial circumstances of the caregiver in either direction may herald an increased risk of financial exploitation or exploitation already under way. Abuse of the power of attorney; if the person with power of attorney or health care proxy is suspected of not acting in the best interest of the patient, documents necessary to ensure that the assumption of fiduciary responsibilities is authorized.
Sexual Abuse	Sexual coercion or assault	Bruising, abrasions, lacerations in the genital or anal areas or abdomen Newly acquired sexually transmitted diseases, especially in nursing home Urinary tract infection	Inquire directly about sexual assault or coercion. For patients with dementia, direct queries to caregivers about hypersexual behavior as part of a larger history regarding dementia-related behaviors and assess patient's capacity for decision-making about sexual activity If indicated, refer to an emergency department for assessment for sexual assault and collection of specimens (forensic evidence should be collected by experienced professionals, such as nurses who have undergone Sexual Assault Nurse Examiners (SANE) training.

Category	Definition and Examples	Symptoms and Signs	Key Aspects of Evaluation
Neglect (by caregiver or self-neglect)	Failure to provide the materials, supplies, food and drink or services necessary for optimal functioning or to avoid harm	Malnutrition Dehydration Poor hygiene Pressure ulcers Nonadherence to medication regimen or other treatments Worsening of dementia-related behavioral symptoms	Interview primary caregiver about his or her understanding of the nature of the patient's care needs and how well care is being rendered. Neglect by a caregiver may be intentional or unintentional Assess hygiene, cleanliness, and appropriateness of dress. Examine the skin for pressure ulcers, infections, and infestations. Assess nutrition and hydration, including measuring body-mass index And blood urea nitrogen and creatinine to assess hydration.

Source: Adapted from JB Halter et al (eds): *Hazzard's Geriatric Medicine and Gerontology*, 7th ed. New York, McGraw-Hill, 2017 and MS Lachs, KA Pillemer: Elder abuse. N Engl J Med 373:1947, 2015.

MANAGEMENT

In addition to treating the physical, medical, and emotional consequences, patients suspected of elder abuse or neglect should be reported to the appropriate local or state agency to investigate and ensure the patient's safety. The reader is referred to two reviews of this topic for further information on specific aspects of management.

END-OF-LIFE AND PALLIATIVE CARE

As stated earlier, end-of-life and palliative care are critical aspects of caring for the geriatric population, and require a comprehensive, person-centered approach. End-of-life and palliative care are addressed in detail in **Chap. 9**, and pain management is addressed in **Chap. 10**. For geriatric patients, limited life expectancy is a critical factor to consider when making end-of-life care decisions. General principles of decision-making are especially relevant when considering palliative and/or end-of-life care in older patients (Fig. 464-4). Decision-making becomes complicated, however, among older patients with multi-morbidity. Without a clear terminal diagnosis, when to start palliative care/end-of-life care could be challenging. While it is sometimes clear when an older patient has a terminal condition, such as end-stage congestive heart failure or chronic obstructive pulmonary disease, many older patients with multi-morbidity have combinations of conditions of varying severity. Moreover, neurogenerative disorders, including most forms of dementia, Parkinson's disease, and patients with multiple strokes commonly have a gradually progressive course, and it can be challenging to determine when discussions about palliative and end-of-life care should be initiated. Dementia, however, should be considered a terminal illness in the advanced stages.

Internists should play a pivotal role in making the decision when to initiate these discussions, and should be proactive in encouraging patients and their families to execute advance directives before a health care crisis occurs. The survivability of cardiopulmonary resuscitation in hospitalized patients aged ≥65 is <20%; among the old-old with multi-morbidity it is much lower. The survivability of cardiopulmonary resuscitation (CPR) in nursing home residents is almost zero, making it a futile intervention for most in this setting. Enteral feeding tubes should not be placed in patients with end-stage dementia (Table 464-2). Estimation of prognosis using tools such as ePrognosis and careful attention to other factors that contribute to person-centered care will assist internists in dealing with these challenging issues in geriatric care.

FURTHER READING

American Geriatrics Society 2015 Beers Criteria Update Expert Panel. American Geriatrics Society 2015 updated Beers criteria for potentially inappropriate medication use in older adults. J Am Geriatr Soc 63:2227, 2015. [PubMed: 26446832]

American Geriatrics Society Expert Panel on the Care of Older Adults with Diabetes Mellitus: Guidelines Abstracted from the American Geriatrics Society Guidelines for Improving the Care of Older Adults with Diabetes Mellitus: 2013 Update. J Am Geriatr Soc 61:2020, 2013. [PubMed: 24219204]

American Geriatrics Society Panel on Pharmacologic Management of Persistent Pain in Older Persons: Pharmacological management of persistent pain in older persons. J Am Geriatr Soc 57:1331, 2009. [PubMed: 19573219]

American Geriatrics Society: Choosing Wisely: Ten things clinicians and patients should question. *http://www.choosingwisely.org/societies/american-geriatrics-society/*. Accessed January 14, 2017.

AMDA—The Society for Post-Acute and Long-Term Care Medicine: Ten things clinicians and patients should question. *http://www.choosingwisely.org/societies/amda-the-society-for-post-acute-and-long-term-care-medicine/*. Accessed January 14, 2017.

CDC Recommended Adult Immunization Schedule United States—2016 *https://www.cdc.gov/vaccines/schedules/downloads/adult/adult-combined-schedule.pdf*. Accessed January 12, 2017.

Clinician's Guide to Assessing and Counseling Older Drivers. *http://www.michigan.gov/documents/sos/Clinicians_Guide_To_OlderDrivers_3rd_edition_523147_7.pdf*. Accessed January 12, 2017.

Halter J et al (eds): *Hazzard's Geriatric Medicine and Gerontology*, 7th ed. New York, McGraw Hill, 2016.

Kane RL et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw Hill, 2018.

Malone ML et al (eds): *Geriatric Models of Care*. New York, Springer, 2015.

Chapter 465: Approach to Medical Consultation

Jack Ende; Jeffrey Berns

INTRODUCTION

Effective health care requires teams of generalists and specialists with complementary expertise. Many clinical conditions require the input of more than one clinical provider, either because the diagnosis and recommended treatment is uncertain or because a patient may have multiple diseases that may be best managed by involving multiple specialists.

To *consult* is to seek advice from someone with expertise in a particular area, whereas *consultation* refers to the meeting or comparable outcome arising from that request. Medical consultation takes several forms. Its most traditional forms include in-hospital consultation in which physicians provide recommendations or perform procedures for a hospitalized patient, and out-patient consultations, in which patients are seen in the office setting. More contemporary forms of consultation include e-consultations, telemedicine evaluations (see "Consultation Involving Telemedicine," below), and remote medical second opinions. In these forms, the consultant may not actually see the patient but, nonetheless, assumes the responsibility of evaluating the patient's clinical condition, assessing and analyzing pertinent clinical data, and offering a synthesis and appropriate recommendations.

While forms of medical consultation evolve, basic responsibilities associated with medical consultation endure. These responsibilities can be divided into those that fall to the requesting physician or non-physician practitioner; the consultant, who provides the consultation; and the health system, hospital, or organization that must support this important medical encounter **(Table 465-1)**.

TABLE 465-1

Stakeholder Responsibilities in the Medical Consultation Process

Referring Physician or Provider	Consultant Physician	Health System, Hospital, or Care Organization
• Ensure patient participation and engagement • Be specific regarding clinical question and desired outcome • Communicate level of urgency • Avoid consulting for nonclinical purposes	• Maintain standards of professionalism, including those pertaining to availability, communication, respect, and collegiality • Appreciate levels of urgency and respond appropriately • Assemble and develop one's own database • Be specific in synthesis and recommendations • Understand desired outcomes, including arrangements for follow-up • Communicate with referring provider in whatever manner is mutually desirable	• Maintain adequate specialty workforce to enable appropriate access • Support systems for efficient exchange of clinical information • Develop culture of collegiality and team-based care

RESPONSIBILITIES OF THE REQUESTING PRACTITIONER

Before requesting a consultation, the provider should ensure that the patient endorses the purpose of the consultation, understands the role of the consultant, and anticipates the likely outcomes of the encounter. Further responsibilities of the requesting practitioner include being specific and communicating clearly the reason for the consultation. Vague messages such as, "Please evaluate" are not as helpful as more specific inquiries such as, "What is the cause of the declining kidney function?" or, "How should this asymptomatic pulmonary nodule be evaluated?" To the extent possible, the requesting practitioner should provide the relevant clinical information, summarized as succinctly as possible. Urgency should be clearly conveyed, typically with a phone call or other direct communication.

The requesting practitioner should be explicit regarding the intended outcome of the consultation, i.e., is this for a single evaluation or ongoing co-management? Communication between the requesting and the consulting providers is paramount. Whether this communication includes direct contact is less important than that the relevant information and desired outcome be explicit and clear, regardless of communication medium. Consultations should be requested for clinical purposes and always directed to qualified consultants; they should not be driven by entrepreneurial or relationship-building purposes. Another responsibility of the referring provider is not to "over-consult." Medical care should be focused on value, not volume.

RESPONSIBILITIES OF THE CONSULTANT

Just as the referring provider should attend to clear and explicit communication, so too should the consultant follow the precepts of effective interactions between professionals, which include courtesy, availability, and clarity. Particularly on the inpatient service, where consultants may receive several requests each day, it is important that the incoming consultations are triaged and dispatched as clinically appropriate. Consultants also need to determine the requested level of involvement going forward and not assume that long-term co-management is being sought. While consultants can and should make use of available clinical data, they should also assemble independently their own database, including taking a history, performing a physical exam, and reviewing pertinent clinical studies. Absent that, they may be unable to provide an independent and actionable synthesis. Just as the referring provider needs to be clear and concise, so too should the consultant be specific and focused in the recommendations provided. “Possible malignant ascites” is less helpful than, “I will arrange for paracentesis to exclude the possibility of malignant ascites.” For the most part, recommendations to “consider” some diagnosis or test are less helpful than more specific and concrete advice. Some referring practitioners wish to be called after a patient is seen; others prefer that communication be handled as part of the medical record. How this communication is handled must also align with the complexity and urgency of the consultation and clinical circumstances.

RESPONSIBILITIES OF HEALTH SYSTEMS, HOSPITALS, AND MEDICAL ORGANIZATIONS

Health systems, hospitals, and medical organizations also have responsibilities in the consultation process. This responsibility includes ensuring that qualified consultants are accessible and available on the medical staff. Consultations within a single system are aided by common shared electronic medical records, particularly when consultations originate in the hospital, but can also involve care in the outpatient setting. Finally, health care entities should strive to foster a culture of team-based care and collegiality.

SPECIAL ISSUES IN MEDICAL CONSULTATION

Curbside Consults

Curbside consults are requests from one practitioner to another for an informal and unwritten opinion about a specific patient care matter. They are typically limited in scope, mostly regarding management or questions regarding procedures, and developed from information provided by the consulting practitioner and perhaps the medical record (such as labs and imaging studies), but without a comprehensive review of the record or any direct contact with the patient. Although often viewed as convenient, efficient, and a common aspect of clinical care, by their very nature, curbside consults have been found to often be incomplete or even flawed. It is not uncommon for the question being asked to be deemed too complex for a curbside consult, or for it not to be the actual or only issue the consultant feels needs to be addressed. As a general rule, curbside consults should be avoided. While medicolegal liability is often cited as a reason to limit curbside consults,

the risk is actually negligible as U.S. courts have ruled that curbside consults do not establish a doctor-patient relationship necessary for creating the basis for medical malpractice litigation. An important exception, however, is when a curbside consult is provided by a resident or fellow in training; in this circumstance the trainee's supervising physician, whether aware of the curbside consult or not, is responsible for the recommendations of the trainee.

Second Opinions

Physicians may find themselves providing consultations requested by patients who have already been evaluated for the same problem by another physician. Not a "consult" in the usual context of one physician referring a patient to another, the service provided by the consultant here is, nonetheless, very much aligned with a physician-referred consult. Second opinions, which often are encouraged by the patient's physician, may be sought by patients for reassurance that a diagnosis and treatment recommendation is correct, out of dissatisfaction with the initial physician, or with the hope of an entirely different opinion and recommendation. The physician providing the second opinion should strive to understand the patient's motivations for seeking the additional opinion. While a second opinion may have been initiated by the patient rather than referral from another physician, it is recommended that the consulting physician communicate with the patient's primary physician or specialist as would be done following a standard consultation unless the patient insists otherwise. In addition, professional behavior in how the consulting physician refers to the recommendations or actions of previously consulted physicians is important, even when there is disagreement. Likewise, it is important that a transfer of care from prior consultants to the one providing a second opinion be enacted only if specifically requested by the patient or the physician who encouraged the second opinion.

Consults Involving Mid-Level Providers

Increasingly, specialist physicians may find themselves being consulted by nurse practitioners and physician assistants rather than other physicians. Whether the quality of the information provided to the consultant physician by a mid-level provider is different from physician-to-physician referrals has not been studied. Consulting physicians should know whether they should respond back to the mid-level provider or to the supervising physician. As with physician-to-physician consults, it is also important for the consultant to know whether the individual calling for the consult has an ongoing role in the care of the patient or is simply covering for a limited period of time. Finally, the consultant, if responding back to the mid-level provider, should make sure that the information provided meets the needs of that provider, and that questions are answered as they would be if responding back to another physician.

Consultation Involving Telemedicine

Consultations making use of electronic health records, patient portals, and various forms of telecommunication technology, including video conferencing or cell phone communication, can improve access to care, reduce cost, and improve outcomes. This is particularly true when employed in geographic areas of health care shortage and when the clinical issues can be handled without direct contact with the

patient, e.g., radiology or dermatology. However, the absence of direct contact between patient and consultant introduces special issues related to diagnostic accuracy and physician-patient relationship. Regulatory issues, liability, security, and confidentiality issues arise as well. Consultation via telemedicine holds considerable promise, but the aforementioned concerns will need to be better understood.

FURTHER READING

Daniel H, Sulmasy LS: Policy recommendations to guide the use of telemedicine in primary care: An American College of Physicians Position Paper. Ann Intern Med 163:787, 2015. [PubMed: 26344925]

Pearson SD: Principles of generalist-specialist relationships. J Gen Intern Med 14(Suppl 1):S13, 1999. [PubMed: 9933490]

Chapter 466: Medical Disorders During Pregnancy

Robert L. Barbieri; John T. Repke

INTRODUCTION

Each year, approximately 4 million births occur in the United States, and more than 130 million births occur worldwide. A significant proportion of births are complicated by medical disorders. Advances in medical care and fertility treatment have increased the number of women with serious medical problems who attempt to become pregnant. Medical problems that interfere with the physiologic adaptations of pregnancy increase the risk for poor pregnancy outcome; conversely, in some instances, pregnancy may adversely impact an underlying medical disorder.

HYPERTENSION

(See also Chap. 271) In pregnancy, cardiac output increases by 40%, with most of the increase due to an increase in stroke volume. Heart rate increases by ~10 beats/min during the third trimester. In the second trimester, systemic vascular resistance decreases, and this decline is associated with a fall in blood pressure. During pregnancy, a blood pressure of 140/90 mmHg is considered to be abnormally elevated and is associated with an increase in perinatal morbidity and mortality. In all pregnant women, the measurement of blood pressure should be performed in the sitting position, because the lateral recumbent position may result in a lower blood pressure. The diagnosis of hypertension requires the measurement of two elevated blood pressures at least 4 h apart. Hypertension during pregnancy is usually caused by preeclampsia, chronic hypertension, gestational hypertension, or renal disease.

PREECLAMPSIA

Approximately 5–7% of all pregnant women develop *preeclampsia*, the new onset of hypertension (blood pressure >140/90 mmHg) and proteinuria (either a 24 h urinary protein >300 mg/24 h, or a protein-creatinine ratio ≥0.3) after 20 weeks of gestation. Recent revisions to the diagnostic criteria include: proteinuria is no longer an absolute requirement for making the diagnosis; the terms mild and severe preeclampsia have been replaced; and the disease is now termed preeclampsia either with or without severe features and fetal growth restriction is no longer a defining criterion for preeclampsia with severe features. Although the precise pathophysiology of preeclampsia remains unknown, recent studies show excessive placental production of antagonists to both vascular endothelial growth factor (VEGF) and transforming growth factor β (TGF-β). These antagonists to VEGF and TGF-β disrupt endothelial and renal glomerular function resulting in edema, hypertension, and proteinuria. The renal histological feature of preeclampsia is glomerular endotheliosis. Glomerular endothelial cells are swollen and encroach on the vascular lumen. Preeclampsia is associated with abnormalities of cerebral circulatory autoregulation, which increase the risk of stroke at mildly and moderately elevated blood pressures. Risk factors for the development of preeclampsia include nulliparity, diabetes mellitus, a history of renal disease or chronic hypertension, a prior history of preeclampsia, extremes of maternal age (>35 years or <15 years), obesity, antiphospholipid antibody syndrome, and multiple gestation. Low-dose aspirin (81 mg daily, initiated at the end of the first trimester) modestly reduces the risk of preeclampsia in pregnant women at high risk of developing the disease.

Preeclampsia with severe features is the presence of new-onset hypertension and proteinuria accompanied by end-organ damage. Features may include severe elevation of blood pressure (>160/110 mmHg), evidence of central nervous system (CNS) dysfunction (headaches, blurred vision, seizures, coma), renal dysfunction (oliguria or creatinine >1.5 mg/dL), pulmonary edema, hepatocellular injury (serum alanine aminotransferase level more than twofold the upper limit of normal), hematologic dysfunction (platelet count <100,000/L or disseminated intravascular coagulation [DIC]). The *HELLP syndrome* (*h*emolysis, *e*levated *l*iver enzymes, *l*ow *p*latelets) is a special subtype of severe preeclampsia and is a major cause of morbidity and mortality in this disease. Platelet dysfunction and coagulation disorders further increase the risk of stroke.

TREATMENT

TREATMENT

Preeclampsia

Preeclampsia resolves within a few weeks after delivery. For pregnant women with preeclampsia prior to 37 weeks of gestation, delivery reduces the mother's morbidity but exposes the fetus to the risk of premature birth. The management of preeclampsia is challenging because it requires the clinician to balance the health of the mother and fetus simultaneously. In general, prior to term, women with preeclampsia without severe features may be managed conservatively with limited physical activity, although bed rest is not recommended, close monitoring of blood pressure and renal function, and careful fetal surveillance. For women with preeclampsia with severe features, delivery is recommended unless the patient is eligible for expectant management in a tertiary hospital setting. Expectant management of preeclampsia with severe features remote from term affords some benefits for the fetus, but significant risks for the mother. Postponing delivery beyond 34 weeks gestation in this group of patients is not recommended. In preeclampsia without severe features delivery at 37 weeks is recommended.

The definitive treatment of preeclampsia is delivery of the fetus and placenta. For women with preeclampsia with severe features, aggressive management of blood pressures >160/105 mmHg reduces the risk of cerebrovascular accidents. IV labetalol or hydralazine is most commonly used to acutely manage severe hypertension in preeclampsia; labetalol is associated with fewer episodes of maternal hypotension. Elevated arterial pressure should be reduced slowly to avoid hypotension and a decrease in blood flow to the fetus.

Magnesium sulfate is the preferred agent for the prevention and treatment of eclamptic seizures. Large, randomized clinical trials have demonstrated the superiority of magnesium sulfate over phenytoin and diazepam in reducing the risk of seizure and, possibly, the risk of maternal death. Magnesium may prevent seizures by interacting with *N*-methyl-D-aspartate (NMDA) receptors in the CNS. The universal use of magnesium sulfate for seizure prophylaxis in preeclampsia without severe features is no longer recommended by most experts. There is consensus that magnesium sulfate should be used in all cases of preeclampsia with severe features, or in cases of eclampsia. Women who have had preeclampsia appear to be at increased risk of cardiovascular and renal disease later in life.

CHRONIC ESSENTIAL HYPERTENSION

Pregnancy complicated by chronic essential hypertension is associated with intrauterine growth restriction and increased perinatal mortality. Pregnant women with chronic hypertension are at increased risk for superimposed preeclampsia and abruptio placentae. Women with chronic hypertension should have a thorough prepregnancy evaluation, both to identify remediable causes of hypertension and to ensure that the prescribed antihypertensive agents (e.g., angiotensin-converting enzyme [ACE] inhibitors, angiotensin-receptor blockers) are not associated with an adverse outcome of pregnancy. Labetalol and nifedipine are the most commonly used medications for the treatment of chronic hypertension in pregnancy.

The target blood pressure is in the range of 130–150 mmHg systolic and 80–100 mmHg diastolic. Should hypertension worsen during pregnancy, baseline evaluation of renal function (see below) is necessary to help differentiate the effects of chronic hypertension from those of superimposed preeclampsia. There are no convincing data that the treatment of mild chronic hypertension improves perinatal outcome.

GESTATIONAL HYPERTENSION

The development of elevated blood pressure after 20 weeks of pregnancy or in the first 24 h post-partum in the absence of preexisting chronic hypertension or proteinuria is referred to as *gestational hypertension*. Mild gestational hypertension that does not progress to preeclampsia has not been associated with adverse pregnancy outcome or adverse long-term prognosis.

RENAL DISEASE

Normal pregnancy is characterized by an increase in glomerular filtration rate and creatinine clearance. This increase occurs secondary to a rise in renal plasma flow and increased glomerular filtration pressures. Patients with underlying renal disease and hypertension may expect a worsening of hypertension during pregnancy. If superimposed preeclampsia develops, the additional endothelial injury results in a capillary leak syndrome that may make management challenging. In general, patients with underlying renal disease and hypertension benefit from aggressive management of blood pressure. Preconception counseling is also essential for these patients so that accurate risk assessment and medication changes can occur prior to pregnancy. In general, a prepregnancy serum creatinine level <133 μmol/L (<1.5 mg/dL) is associated with a favorable prognosis. When renal disease worsens during pregnancy, close collaboration between the internist and the maternal-fetal medicine specialist is essential so that decisions regarding delivery can be weighed to balance the sequelae of prematurity for the neonate versus long-term sequelae for the mother with respect to future renal function.

CARDIAC DISEASE

VALVULAR HEART DISEASE

(See also Chaps. 256–263) Valvular heart disease is the most common cardiac problem complicating pregnancy.

Mitral Stenosis

This is the valvular disease most likely to cause death during pregnancy. The pregnancy-induced increase in blood volume, cardiac output, and tachycardia can increase the transmitral pressure gradient and cause pulmonary edema in women with mitral stenosis. Women with moderate to severe mitral stenosis (mitral valve area ≤1.5 cm^2) who are planning pregnancy and have either symptomatic disease or pulmonary hypertension should undergo valvuloplasty prior to conception, preferably with percutaneous balloon valvotomy (PBV). Pregnancy associated with long-standing mitral stenosis may result in pulmonary hypertension. Sudden death has been reported when hypovolemia occurs. Careful control of heart rate, especially during labor and delivery, minimizes the impact of tachycardia and reduced ventricular filling times on cardiac function. Pregnant women with mitral stenosis are at increased risk for the development of atrial fibrillation and other tachyarrhythmias. The immediate postpartum period is a time of particular concern secondary to rapid volume shifts. Careful monitoring of cardiac and fluid status should be observed.

Mitral Regurgitation and Aortic Regurgitation and Stenosis

The pregnancy-induced decrease in systemic vascular resistance reduces the risk of cardiac failure with these conditions, especially in women with chronic lesions. Acute onset of mitral or aortic regurgitation may not be well tolerated during

pregnancy. For women with severe aortic stenosis, treatment before pregnancy should be considered for a peak-to-peak valve gradient >50 mmHg. In women with aortic stenosis and a mean valve gradient <25 mmHg, pregnancy is likely to be well tolerated. For women with mitral or aortic regurgitation and left ventricular dysfunction (LVEF <30%) pregnancy should be avoided.

CONGENITAL HEART DISEASE

(See also Chap. 264) Reparative surgery has markedly increased the number of adult women with surgically repaired congenital heart disease. Maternal morbidity and mortality are greater among these women than among those without surgical cardiac repair. When pregnant, these patients should be jointly managed by a cardiologist and an obstetrician familiar with these problems. The presence of a congenital cardiac lesion in the mother increases the risk of congenital cardiac disease in the newborn. Prenatal screening of the fetus for congenital cardiac disease with ultrasound is recommended.

OTHER CARDIAC DISORDERS

Supraventricular tachycardia **(Chap. 241)** is a common cardiac complication of pregnancy. Treatment is the same as in the nonpregnant patient, and fetal tolerance of medications such as adenosine and calcium channel blockers is acceptable. When necessary, pharmacologic or electric cardioversion may be performed to improve cardiac performance and reduce symptoms. This intervention is generally well tolerated by mother and fetus.

Peripartum cardiomyopathy **(Chap. 254)** is an uncommon disorder of pregnancy and its etiology remains unknown. Approximately 10% of women with peripartum cardiomyopathy carry a truncating mutation in the gene encoding the titin sarcomere protein. Treatment is directed toward symptomatic relief and improvement of cardiac function. Many patients recover completely; others are left with progressive dilated cardiomyopathy. Recurrence in a subsequent pregnancy has been reported, and women who do not have normal baseline left-ventricular function after an episode of peripartum cardiomyopathy should be counseled to avoid pregnancy.

SPECIFIC HIGH-RISK CARDIAC LESIONS

Marfan Syndrome

(See also Chap. 406) This autosomal dominant disease is associated with an increased risk of aortic dissection and rupture. An aortic root diameter <40 mm is associated with a favorable outcome of pregnancy; conversely, an aortic root diameter >40 mm is associated with an increased risk of aortic dissection. Prophylactic therapy with beta blockers has been advocated to reduce aortic dilation and the risk of dissection. A “cardiac delivery” with reduced pushing and early intervention with operative delivery is often recommended to reduce increases in aortic wall stress caused by the Valsalva maneuver.

Ehlers-Danlos syndrome (*EDS*) may be associated with premature labor, and in type IV EDS there is increased risk of organ or vascular rupture that may cause death. For women with vascular EDS, pregnancy is relatively contraindicated because of the high risk of vascular and uterine rupture.

Pulmonary Hypertension

(See also Chap. 277) Maternal mortality in the setting of severe pulmonary hypertension is high, and primary pulmonary hypertension is a contraindication to pregnancy. Termination of pregnancy may be advisable in these circumstances to preserve the life of the mother. In the *Eisenmenger syndrome*, i.e., the combination of pulmonary hypertension with right-

to-left shunting due to congenital abnormalities (Chap. 264), maternal and fetal deaths occur frequently. Systemic hypotension may occur after blood loss, prolonged Valsalva maneuver, or regional anesthesia; sudden death secondary to hypotension is a dreaded complication. Management of these patients is challenging, and invasive hemodynamic monitoring during labor and delivery is recommended in severe cases.

In patients with pulmonary hypertension, vaginal delivery is less stressful hemodynamically than cesarean section, which should be reserved for accepted obstetric indications.

DEEP VENOUS THROMBOSIS AND PULMONARY EMBOLISM

(See also Chap. 273) Pregnancy is associated with venous stasis, endothelial injury and a hypercoagulable state. Inherited thrombophilias and the presence of antiphospholipid antibodies increase the risk of venous thromboembolism (VTE) in pregnancy. Deep venous thrombosis (DVT) or pulmonary embolism (PE) occurs in about 1 in 500 pregnancies, with DVT being three times more common than PE. VTE occurs more commonly in the 6 weeks post-partum than antepartum. In pregnant women, most unilateral DVTs occur in the left leg because the left iliac vein is compressed by the right iliac artery and the uterus compresses the inferior vena cava.

TREATMENT

TREATMENT

Deep Venous Thrombosis

Aggressive diagnosis and management of DVT and suspected pulmonary embolism optimize the outcome for mother and fetus. In general, all diagnostic and therapeutic modalities afforded that the nonpregnant patient should be utilized in pregnancy except for D-dimer measurement, in which values are elevated in normal pregnancy. Anticoagulant therapy with low-molecular-weight heparin (LMWH) or unfractionated heparin is indicated in pregnant women with DVT. LMWH may be associated with an increased risk of epidural hematoma in women receiving an epidural anesthetic in labor and must be discontinued at least 24 h before placement of an epidural catheter. Warfarin therapy is contraindicated in the first trimester due to its association with fetal chondrodysplasia punctata. In the second and third trimesters, warfarin may cause fetal optic atrophy and mental retardation. In pregnancy the use of warfarin is restricted to women with mechanical heart valves. Warfarin is not contraindicated in breast-feeding women. For women at moderate or high risk of DVT who have a cesarean delivery, mechanical and/or pharmacologic prophylaxis is warranted.

ENDOCRINE DISORDERS

DIABETES MELLITUS

(See also Chaps. 396–398) In pregnancy, the fetoplacental unit induces major metabolic changes, the purpose of which is to shunt glucose and amino acids to the fetus while the mother uses ketones and triglycerides to fuel her metabolic needs. These metabolic changes are accompanied by maternal insulin resistance caused in part by placental production of steroids, a growth hormone variant, and placental lactogen. Although pregnancy has been referred to as a state of “accelerated starvation,” it is better characterized as “accelerated ketosis.” In pregnancy, after an overnight fast, plasma glucose is lower by 0.8–1.1 mmol/L (15–20 mg/dL) than in the nonpregnant state. This difference is due to the use of glucose by the fetus. In early pregnancy, fasting may result in circulating glucose concentrations in the range of 2.2 mmol/L (40 mg/dL) and may be associated with symptoms of hypoglycemia. In contrast to the decrease in maternal glucose concentration, plasma hydroxybutyrate and acetoacetate levels rise to two to four times normal after a fast.

TREATMENT

TREATMENT

Diabetes Mellitus in Pregnancy

Pregnancy complicated by diabetes mellitus is associated with higher maternal and perinatal morbidity and mortality rates. Preconception counseling and treatment are important for the diabetic patient contemplating pregnancy and can reduce the risk of congenital malformations and improve pregnancy outcome. Folate supplementation reduces the incidence of fetal neural tube defects, which occur with greater frequency in fetuses of diabetic mothers. In addition, optimizing glucose control during key periods of organogenesis reduces other congenital anomalies, including sacral agenesis, caudal dysplasia, renal agenesis, and ventricular septal defect.

Once pregnancy is established, glucose control should be managed more aggressively than in the nonpregnant state. In addition to dietary changes, this enhanced management requires more frequent blood glucose monitoring and often involves additional injections of insulin or conversion to an insulin pump. Fasting blood glucose levels should be maintained at <5.8 mmol/L (<105 mg/dL), with avoidance of values >7.8 mmol/L (140 mg/dL). Sequential measurement of hemoglobin A1c is of minimal value for monitoring glucose control during pregnancy because of the higher rate of red blood cell turnover during pregnancy. Commencing in the third trimester, regular surveillance of maternal glucose control as well as assessment of fetal growth (obstetric sonography) and fetoplacental oxygenation (fetal heart rate monitoring or biophysical profile) optimize pregnancy outcome. Pregnant diabetic patients without vascular disease are at greater risk for delivering a macrosomic fetus, and attention to fetal growth via clinical and ultrasound examination is important. Fetal macrosomia is associated with an increased risk of maternal and fetal birth trauma, including permanent injury to the brachial plexus. Pregnant women with diabetes have an increased risk of developing preeclampsia, and those with vascular disease are at greater risk for developing intrauterine growth restriction, which is associated with an increased risk of fetal and neonatal death. Excellent pregnancy outcomes in patients with diabetic nephropathy and proliferative retinopathy have been reported with aggressive glucose control and intensive maternal and fetal surveillance.

As pregnancy progresses, glycemic control may become more difficult to achieve due to an increase in insulin resistance. In pregnant women with Type 1 diabetes, closed-loop insulin delivery with both continuous interstitial glucose monitoring and sensor-augmented insulin pump therapy is helpful in normalizing circulating glucose with few episodes of hypoglycemia. In general, efforts to control glucose and avoid preterm delivery result in the best overall outcome for both mother and newborn. Preterm delivery is generally performed only for the usual obstetric indications (e.g., preeclampsia, fetal growth restriction, non-reassuring fetal testing) or for worsening maternal renal or active proliferative retinopathy.

GESTATIONAL DIABETES (GDM)

GDM occurs in ~4% of pregnancies. Because about 90% of women have at least one risk factor for GDM, all pregnant women should be screened for GDM. A typical two-step strategy for establishing the diagnosis of GDM is performed at 24–28 weeks of gestation and involves administration of a 50-g oral glucose challenge with a single serum glucose measurement at 60 min. If the plasma glucose is <7.8 mmol/L (<130 mg/dL) the test is considered normal. Plasma glucose >7.8 mmol/L (>130 mg/dL) warrants administration of a 100-g oral glucose challenge with plasma glucose measurements obtained in the fasting state and at 1, 2, and 3 h. Normal plasma glucose concentrations at these time points are <5.3 mmol/L (<95 mg/dL), <10 mmol/L (<180 mg/dL), <8.6 mmol/L (<155 mg/dL), and <7.8 mmol/L (<140 mg/dL) as the upper norms for a 3-h glucose tolerance test. Two elevated glucose values indicate a positive test. Adverse pregnancy outcomes for mother and fetus appear to increase with glucose as a continuous variable; thus it is challenging to define the optimal threshold for establishing the diagnosis of GDM.

Pregnant women with GDM are at increased risk of stillbirth, preeclampsia, and delivery of infants who are large for their gestational age, with resulting birth lacerations, shoulder dystocia, and birth trauma including brachial plexus injury. These fetuses are at risk of hypoglycemia, hyperbilirubinemia, and polycythemia. Tight control of blood sugar during pregnancy and labor can reduce these risks.

TREATMENT

TREATMENT

Gestational Diabetes

Treatment of GDM with a two-step strategy—dietary intervention followed by insulin injections if diet alone does not adequately control blood sugar (fasting glucose <5.6 mmol/L [<100 mg/dL] and 2-h postprandial glucose <7.0 mmol/L [<126 mg/dL])—is associated with a decreased risk of birth trauma for the fetus. Oral hypoglycemic agents such as glyburide and metformin have become more commonly utilized for managing GDM refractory to nutritional management, but most experts favor insulin therapy. For women with GDM, there is a 40% risk of being diagnosed with diabetes within the 10 years after the index pregnancy. All women with GDM should have a formal glucose tolerance test (GTT) to screen for T2DM at ~6 weeks post-partum. In women with a history of GDM, exercise, weight loss, and treatment with metformin reduce the risk of developing diabetes. Lactation also reduces the risk of GDM progressing to T2DM. All women with a history of GDM should be counseled about prevention strategies and evaluated regularly for diabetes.

OBESITY

(See also Chap. 395) Pregnant women who are obese have an increased risk of stillbirth, congenital fetal malformations, GDM, preeclampsia, urinary tract infections, preterm and post-date delivery, and cesarean delivery. Women contemplating pregnancy should attempt to attain a healthy weight prior to conception. For morbidly obese women who have not been able to lose weight with lifestyle changes, bariatric surgery reduces the risks for GDM, macrosomia, and preterm delivery. Following bariatric surgery, women should delay conception for 1 year to avoid pregnancy during an interval of rapid metabolic changes. The National Academy of Medicine guidelines for weight gain during pregnancy recommend that for BMI ranges of <18.5, 18.5–24.9, 25.0–29.9, and ≥30 kg/m^2, weight gain targets should be 12.5–18 kg, 11.5–16 kg, 7–11.5 kg, and 5–9 kg, respectively.

THYROID DISEASE

(See also Chap. 375) In pregnancy, the estrogen-induced increase in thyroxine-binding globulin increases circulating levels of total T_3 and total T_4. Placental human chorionic gonadotropin (hCG) directly stimulates the thyroid causing an increase in free T_3 and T_4. Interpretation of the measurement of free T_4, free T_3, and thyroid-stimulating hormone (TSH) should use trimester-specific ranges.

TREATMENT

TREATMENT

Hyperthyroidism in Pregnancy

HYPERTHYROIDISM

Methimazole crosses the placenta to a greater degree than propylthiouracil and has been associated with fetal aplasia cutis. However, propylthiouracil can be associated with liver failure. Some experts recommend propylthiouracil in the first

trimester and methimazole thereafter. Radioiodine should not be used during pregnancy, either for scanning or for treatment, because of effects on the fetal thyroid. In emergent circumstances, additional treatment with beta blockers may be necessary. Hyperthyroidism is most difficult to control in the first trimester of pregnancy and easiest to control in the third trimester. In women with high-titer thyroid stimulating antibodies, the newborn may be born with neonatal Graves' disease.

HYPOTHYROIDISM

The goal of therapy for hypothyroidism is to maintain the serum TSH in the normal range, and thyroxine is the drug of choice. During pregnancy, the dose of thyroxine required to keep the TSH in the normal range rises. In one study, the mean replacement dose of thyroxine required to maintain the TSH in the normal range was 0.1 mg daily before pregnancy and increased to 0.15 mg daily during pregnancy. Since the increased thyroxine requirement occurs as early as the fifth week of pregnancy, one approach is to increase the thyroxine dose by 30% (two additional pills weekly) as soon as pregnancy is diagnosed and then adjust the dose by serial measurements of TSH.

HEMATOLOGIC DISORDERS

Pregnancy has been described as a state of physiologic anemia. Part of the reduction in hemoglobin concentration is dilutional, but iron and folate deficiencies are major causes of correctable anemia during pregnancy.

In populations at high risk for hemoglobinopathies (Chap. 94), hemoglobin electrophoresis should be performed as part of the prenatal screen. Hemoglobinopathies can be associated with increased maternal and fetal morbidity and mortality. Management is tailored to the specific hemoglobinopathy and is generally the same for both pregnant and nonpregnant women. Prenatal diagnosis of hemoglobinopathies in the fetus is readily available and should be discussed with prospective parents either prior to or early in pregnancy.

Thrombocytopenia occurs commonly during pregnancy. The majority of cases are benign gestational thrombocytopenias, but the differential diagnosis should include immune thrombocytopenia (Chap. 111), preeclampsia, and thrombotic thrombocytopenic purpura. Benign gestational thrombocytopenia is unlikely if the platelet count is <100,000 per µL.

NEOPLASIA

Cancer complicates ~1 in every 1000 pregnancies. Of all the cancers that occur in women, <1% complicate pregnancies. The four cancers that occur most commonly in pregnancy are cervical cancer, breast cancer, melanoma, and lymphomas (particularly Hodgkin's lymphoma); however, virtually every form of cancer has been reported in pregnant women (Table 466-1). In addition to cancers developing in other organs of the mother, gestational trophoblastic tumors can arise from the placenta.

Managing cancer in a pregnant woman is complex. One must take into account (1) the possible influence of the pregnancy on the natural history of the cancer, (2) effects on the mother and fetus of complications from the malignancy (e.g., anorexia, nausea, vomiting, malnutrition), (3) potential effects of diagnostic and staging procedures, and (4) potential effects of cancer treatments on both the mother and the developing fetus. Generally, the management that optimizes maternal physiology is also best for the fetus. The dilemma occasionally arises that what is best for the mother may be harmful to the fetus, and what is best for the fetus may compromise the ultimate prognosis for the mother. The best way to approach management of a pregnant woman with cancer is to ask, "What would one do in this clinical situation if she was not pregnant? Then, which, if any aspect of those plans need to be modified because she is pregnant?"

TABLE 466-1

Incidence of Malignant Tumors During Gestation

Tumor Type	Incidence per 10,000 Pregnancies[a]	% of Cases[b]
Breast cancer	1–3	25%
Cervical cancer	1.2–4.5	25%
Thyroid cancer	1.2	15%
Hodgkin's disease	1.6	10%
Melanoma	1–2.6	8%
Ovarian cancer	0.8	2%
All sites	10	100%

[a]These are estimates based on extrapolations from a review of more than 3 million pregnancies (LH Smith et al: Am J Obstet Gynecol 184:1504, 2001). [b]Based on accumulating case reports from the literature; the precision of these data is not high.

TREATMENT

TREATMENT

Special Therapeutic Considerations in Pregnancy

Exposure of developing fetuses to ionizing radiation may cause adverse fetal effects; awareness of this potential toxicity has resulted in a disproportionate aversion to diagnostic imaging in pregnancy. The fetus is most sensitive to teratogenesis during organogenesis in the first trimester. Imaging that uses ionizing radiation should not be done without a compelling reason and due consideration to obtaining the necessary information by alternative imaging modalities. Exposure to diagnostic and therapeutic radionuclides, especially radioactive iodine, poses unique risks, but a full discussion of these is beyond the scope of this chapter.

Generally, toxic chemotherapy should be avoided during pregnancy, if at all possible. It should virtually never be given in the first trimester. A variety of single agents and combinations have been administered in the second and third trimesters, without a high frequency of toxic effects to the pregnancy or the fetus, but data on safety are sparse. A database on the risks associated with individual chemotherapy agents is available (*http://ntp.niehs.nih.gov.ezproxy1.library.usyd.edu.au/ntp/ohat/cancer_chemo_preg/chemopregnancy_monofinal_508.pdf*). If the malignancy is slowly progressive, and if the patient is near her delivery date, and if waiting until delivery to begin treatment is not anticipated to compromise maternal prognosis, then delaying treatment until after delivery to avoid fetal exposure to chemotherapy is desirable. If there is a greater sense of urgency to begin definitive treatment to avoid compromising maternal prognosis, and the patient is beyond 24 weeks of gestation but remote from her delivery date, then treatment (surgical, medical, or both) might be initiated during pregnancy and plans made to deliver the fetus early to avoid exposure to more chemotherapy than absolutely necessary. Since neonatal prognosis is most closely linked to gestational

age at delivery, decisions regarding timing of delivery should include input from Maternal-Fetal Medicine, Neonatology, and Oncology. Finally, if the patient is in her first trimester and toxic chemotherapy must be initiated promptly to avoid a very poor maternal outcome, then it may be necessary to consider therapeutic abortion to avoid maternal disaster and fetal survival with injury resulting in long-term morbid sequelae. In general, pregnancy has relatively little or no impact on the natural history of malignancies, despite the hormonal influences. Spread of the mother's cancer to the fetus (so-called *vertical transmission*) is exceedingly rare.

NEUROLOGIC DISORDERS

For women with epilepsy planning pregnancy, consideration should be given to switching from valproate, a known teratogen, to another medication. If valproate is continued during pregnancy, folic acid supplementation should be increased to 4 mg daily.

Patients with preexisting *multiple sclerosis* (Chap. 436) experience a gradual decrease in the risk of relapses as pregnancy progresses and, conversely, an increase in attack risk during the postpartum period. Disease-modifying agents, including interferon β, should *not* be administered to pregnant multiple sclerosis patients, but moderate or severe relapses can be safely treated with pulse glucocorticoid therapy. Finally, certain tumors, particularly pituitary adenoma and meningioma (Chap. 373), may manifest during pregnancy because of accelerated growth, possibly driven by hormonal factors.

Peripheral nerve disorders associated with pregnancy include *Bell's palsy* (idiopathic facial paralysis) (Chap. 438), which is approximately threefold more likely to occur during the third trimester and immediate postpartum period than in the general population. Therapy with glucocorticoids should follow the guidelines established for nonpregnant patients. Entrapment neuropathies are common in the later stages of pregnancy, presumably as a result of fluid retention. *Carpal tunnel syndrome* (median nerve) presents first as pain and paresthesia in the hand (often worse at night) and later with weakness in the thenar muscles. Treatment is generally conservative; wrist splints may be helpful, and glucocorticoid injections or surgical section of the carpal tunnel can usually be postponed. *Meralgia paresthetica* (lateral femoral cutaneous nerve entrapment) consists of pain and numbness in the lateral aspect of the thigh without weakness. Patients are usually reassured to learn that these symptoms are benign and can be expected to remit spontaneously after the pregnancy has been completed. *Restless leg syndrome* is the most common peripheral nerve and movement disorder in pregnancy. Disordered iron metabolism is the suspected etiology. Management is expectant in most cases.

GASTROINTESTINAL AND LIVER DISEASE

Up to 90% of pregnant women experience nausea and vomiting during the first trimester of pregnancy. *Hyperemesis gravidarum* is a severe form that prevents adequate fluid and nutritional intake and may require hospitalization to prevent dehydration and malnutrition.

Crohn's disease may be associated with exacerbations in the second and third trimesters. Ulcerative colitis is associated with disease exacerbations in the first trimester and during the early postpartum period. Medical management of these diseases during pregnancy is similar to management in the nonpregnant state (Chap. 319).

Exacerbation of gallbladder disease is common during pregnancy. In part, this aggravation may be due to pregnancy-induced alteration in the metabolism of bile and fatty acids. Intrahepatic cholestasis of pregnancy is generally a third-trimester event. Profound pruritus may accompany this condition, and it may be associated with increased fetal mortality. Placental bile salt deposition may contribute to progressive uteroplacental insufficiency. Therefore, regular fetal surveillance should be undertaken once the diagnosis of intrahepatic cholestasis is made, and delivery should be planned once the fetus reaches about 37 weeks of gestation. Favorable results with ursodiol have been reported.

Acute fatty liver is a rare complication of pregnancy. Frequently confused with the HELLP syndrome (see "Preeclampsia" above) and severe preeclampsia, the diagnosis of acute fatty liver of pregnancy may be facilitated by imaging studies and laboratory evaluation. Acute fatty liver of pregnancy is generally characterized by markedly increased serum levels of bilirubin and ammonia and by hypoglycemia. Management of acute fatty liver of pregnancy is supportive; recurrence in subsequent pregnancies has been reported.

All pregnant women should be screened for hepatitis B. This information is important for pediatricians after delivery of the infant. All infants receive hepatitis B vaccine. Infants born to mothers who are carriers of hepatitis B surface antigen should also receive hepatitis B immune globulin as soon after birth as possible and preferably within the first 72 h. Screening for hepatitis C is recommended for individuals at high risk for exposure.

INFECTIONS

BACTERIAL INFECTIONS

Other than bacterial vaginosis, the most common bacterial infections during pregnancy involve the urinary tract **(Chap. 130)**. Many pregnant women have asymptomatic bacteriuria, most likely due to stasis caused by progestational effects on ureteral and bladder smooth muscle and later in pregnancy due to compression effects of the enlarging uterus. In itself, this condition is not associated with an adverse outcome of pregnancy. If asymptomatic bacteriuria is left untreated, symptomatic pyelonephritis may occur. Indeed, ~75% of pregnancy-associated pyelonephritis cases are the result of untreated asymptomatic bacteriuria. All pregnant women should be screened with a urine culture for asymptomatic bacteriuria at the first prenatal visit. Subsequent screening with nitrite/leukocyte esterase strips is indicated for high-risk women, such as those with sickle cell trait or a history of urinary tract infections. All women with positive screens should be treated. Pregnant women who develop pyelonephritis need inpatient IV antibiotic administration due to the elevated risk of urosepsis and acute respiratory distress syndrome in pregnancy. Pregnant women with recurrent urinary tract infections, or one episode of pyelonephritis, should be considered for daily antibiotic suppressive treatment throughout the remainder of their pregnancy.

All pregnant patients are screened prenatally for syphilis, gonorrhea, and chlamydial infections, and the detection of any of these should result in prompt evaluation and treatment **(Chaps. 151 and 184)**.

VIRAL INFECTIONS

Zika Virus (ZV)

ZV can be transmitted from mother to fetus throughout gestation and often results in fetal death, severe microcephaly, or other malformations of the central nervous system. Pregnant symptomatic women with relevant epidemiologic exposure within 2 weeks of symptom onset should have serum and urine tested for ZV ribonucleic acid by real-time reverse transcriptase-polymerase chain reaction (RT-PCR). Testing 2–12 weeks after symptom onset utilizes serum measurement of Zika and dengue virus IgM. Sequential obstetrical ultrasound is recommended to assess for fetal growth and anomalies. Couples considering pregnancy should avoid travel to areas with known mosquito transmission of ZV.

Influenza

(See also Chap. 195) Pregnant women with influenza are at increased risk of serious complications and death. All women who are pregnant or plan to become pregnant in the near future should receive inactivated influenza vaccine. The prompt initiation of antiviral treatment is recommended for pregnant women in whom influenza is suspected. Treatment can be

reconsidered once the results of high-sensitivity tests are available. Prompt initiation of treatment lowers the risk of admission to an intensive care unit and death.

Cytomegalovirus Infection

The most common cause of congenital viral infection in the United States is cytomegalovirus (CMV) **(Chap. 190)**. As many as 50–90% of women of childbearing age have antibodies to CMV, but only rarely does CMV reactivation result in neonatal infection. More commonly, primary CMV infection during pregnancy creates a risk of congenital CMV. No currently accepted treatment of CMV infection during pregnancy has been demonstrated to protect the fetus effectively. Moreover, it is difficult to predict which fetus will sustain a life-threatening CMV infection. Severe CMV disease in the newborn is characterized most often by petechiae, hepatosplenomegaly, and jaundice. Chorioretinitis, microcephaly, intracranial calcifications, hepatitis, hemolytic anemia, and purpura may also develop. CNS involvement, resulting in the development of psychomotor, ocular, auditory, and dental abnormalities over time, has been described. Women with a primary CMV infection should delay conception for 6 months.

Rubella

(See also Chap. 201) Rubella virus is a known teratogen; first-trimester rubella carries a high risk of fetal anomalies, though the risk significantly decreases later in pregnancy. Congenital rubella may be diagnosed by percutaneous umbilical-blood sampling with the detection of IgM antibodies in fetal blood. All pregnant women and all women of childbearing age should be tested for their immune status to rubella. All women who might become pregnant and who are not immune to rubella should be vaccinated at least 3 months before conception.

Herpesvirus Infection

(See also Chap. 187) The acquisition of genital herpes during pregnancy is associated with spontaneous abortion, prematurity, and congenital and neonatal herpes. A cohort study of pregnant women without evidence of previous herpesvirus infection demonstrated that ~2% acquired a new herpesvirus infection during the pregnancy. Approximately 60% of the newly infected women had no clinical symptoms. Infection occurred with equal frequency in all three trimesters. If herpesvirus seroconversion occurred early in pregnancy, the risk of transmission to the newborn was very low. In women who acquired genital herpes shortly before delivery, the risk of transmission was high. The risk of active genital herpes lesions at term can be reduced by prescribing acyclovir for the last 4 weeks of pregnancy to all women who had an episode of genital herpes during the pregnancy.

Herpesvirus infection in the newborn can be devastating. Disseminated neonatal herpes carries with it high mortality and morbidity rates from CNS involvement. It is recommended that pregnant women with active genital herpes lesions at the time of presentation in labor be delivered by cesarean section.

Parvovirus Infection

(See also Chap. 192) Parvovirus infection (caused by human parvovirus B19) may occur during pregnancy. It rarely causes sequelae, but susceptible women infected during pregnancy may be at risk for fetal hydrops secondary to erythroid aplasia and profound anemia.

HIV Infection

(See also Chap. 197) The predominant cause of HIV infection in children is transmission of the virus from mother to newborn during the perinatal period. All pregnant women should be screened for HIV infection. Factors that increase the risk of

mother-to-newborn transmission include high maternal viral load, low maternal CD4+ T cell count, prolonged labor, prolonged duration of membrane rupture, and the presence of other genital tract infections, such as syphilis or herpes. Prior to the widespread use of antiretroviral treatment, the perinatal transmission rate was in the range of 20%. In women with a good response to antiretroviral treatment, the transmission rate is about 1%. Measurement of maternal plasma HIV RNA copy number guides the decision for vaginal versus cesarean delivery. For women with <1000 copies of plasma HIV RNA/mL who are receiving combination antiretroviral therapy, the risk of transmission to the newborn is ~1% regardless of mode of delivery or duration of membrane rupture. These women may elect to attempt a vaginal birth following the spontaneous onset of labor. For women with a viral load of ≥1000 copies/mL prior to 38 weeks of gestation, a scheduled prelabor cesarean at 38 weeks is recommended to reduce the risk of HIV transmission to the newborn.

VACCINATIONS

(See also Chap. 118) For rubella-nonimmune individuals contemplating pregnancy, measles-mumps-rubella vaccine should be administered, ideally at least 3 months prior to conception, but otherwise in the immediate postpartum period. In addition, pregnancy is not a contraindication for vaccination against influenza, tetanus, diphtheria, and pertussis (Tdap), and these vaccines are recommended for appropriate individuals.

MATERNAL MORTALITY

Maternal death is defined as death occurring during pregnancy or within 42 days of completion of pregnancy from a cause related to or aggravated by pregnancy, but not due to accident or incidental causes. The maternal mortality ratio is the number of maternal deaths per 100,000 live births. From 1935 to 2007, the U.S. maternal mortality ratio decreased from nearly 600/100,000 births to 12.7/100,000 births. Since 2007, the U.S. maternal mortality ratio has increased to 21.5/100,000 births. There are significant health disparities in the maternal mortality ratio. In the United States, in the period from 2005 to 2014, the maternal mortality ratios (per 100,000 live births) by race were 11.3 among Hispanic women, 14.1 among non-Hispanic white women, and 40.2 among non-Hispanic black women. The most common causes of maternal death in the United States today are pulmonary embolism, obstetric hemorrhage, hypertension, sepsis, cardiovascular conditions (including peripartum cardiomyopathy and stroke), and ectopic pregnancy. Specialists in internal medicine play an important role in national efforts to reduce the maternal mortality ratio.


As stated above, the maternal mortality ratio in the United States is about 21.5/100,000 live births. In some countries in sub-Saharan Africa and southern Asia, the maternal mortality ratio is >500/100,000 live births. The most common causes of maternal death in these countries are maternal hemorrhage, hypertensive disorders, infection, obstructed labor, and complications from unsafe pregnancy termination. The health interventions that would have the greatest impact on maternal health include improving the following components of the health system: (1) access to contraceptive services in order to space births and limit total family size; (2) access to safe pregnancy termination; (3) presence of trained birth attendants at all deliveries; and (4) transportation to emergency obstetrical centers that can provide intensive medical and surgical services, including cesarean delivery. Maternal death is a global public-health tragedy that could be mitigated with the application of modest resources.

SUMMARY

With improved diagnostic and therapeutic modalities as well as advances in the treatment of infertility, more patients with serious medical complications will be seeking to become pregnant and will require complex obstetric care. Improved outcomes of pregnancy in these women will be best attained by a team of internists, maternal-fetal medicine (high-risk obstetrics) specialists, pediatricians and anesthesiologists assembled to counsel these patients about the risks of pregnancy

and to plan their treatment prior to, and following, conception. The importance of preconception counseling cannot be overstated. It is the responsibility of all physicians caring for women in the reproductive age group to assess their patients' reproductive plans as part of their overall health evaluation.

ACKNOWLDGEMENT

The authors are grateful to Michael F. Greene and Dan L. Longo for their contributions to the content on neoplasia in pregnancy based upon material from previous editions of *Harrison's*.

FURTHER READING

Brasil P et al: Zika virus infection in pregnant women in Rio de Janeiro-preliminary report. N Engl J Med 375:2321, 2016.

Esposito S et al: Chemotherapy against cancer during pregnancy: A systematic review on neonatal outcomes. Medicine (Baltimore) 95:e4899, 2016. [PubMed: 27661036]

LeFevre ML et al: Low-dose aspirin use for the prevention of morbidity and mortality from preeclampsia: U.S. Preventive Services Task Force recommendations statement. Ann Int Med 161:819, 2014. [PubMed: 25200125]

Moaddab A et al: Health care disparities and state-specific pregnancy-related mortality in the United States, 2005–2014. Obstet Gynecol 126:869, 2016.

Stewart ZA et al: Closed-loop insulin delivery during pregnancy in women with Type 1 Diabetes. N Engl J Med 375:644, 2016. [PubMed: 27532830]

Ware JS et al: Shared genetic predisposition in peripartum and dilated cardiomyopathies. N Engl J Med 374:233, 2016. [PubMed: 26735901]

Chapter 467: Medical Evaluation of the Surgical Patient

Prashant Vaishnava; Kim A. Eagle

INTRODUCTION

Cardiovascular and pulmonary complications continue to account for major morbidity and mortality in patients undergoing noncardiac surgery. Emerging evidence-based practices dictate that the internist should perform an individualized evaluation of the surgical patient to provide an accurate preoperative risk assessment and stratification that will guide optimal perioperative risk-reduction strategies. This chapter reviews cardiovascular and pulmonary preoperative risk assessment, emphasizing the goal-directed management of patients at elevated risk for adverse cardiovascular outcomes in the perioperative period. In addition, perioperative management of diabetes mellitus and prophylaxis of endocarditis and for venous thromboembolism are reviewed.

EVALUATION OF INTERMEDIATE- AND HIGH-RISK PATIENTS

Simple, standardized preoperative screening questionnaires, such as the one shown in Table 467-1, have been developed for the purpose of identifying patients at intermediate or high risk who may benefit from a more detailed clinical evaluation. Evaluation of such patients for surgery should always begin with a thorough history and physical examination and with a 12-lead resting electrocardiogram, in accordance with the American College of Cardiology/American Heart Association guidelines. The history should focus on symptoms of occult cardiac or pulmonary disease. The urgency of the surgery should be determined, as true emergency procedures are associated with unavoidably higher morbidity and mortality risk. Preoperative laboratory testing should be carried out only for specific clinical conditions, as noted during clinical examination. Thus, healthy patients of any age who are undergoing elective surgical procedures without coexisting medical conditions should not require any testing unless the degree of surgical stress may result in unusual changes from the baseline state.

TABLE 467-1

Standardized Preoperative Questionnaire[a]

1. Age, weight, height
2. Are you:
   Female and 55 years of age or older or male and 45 years of age of older?
   If yes, are you 70 years of age or older?
3. Do you take anticoagulant medications (“blood thinners”)?
4. Do you have or have you had any of the following heart-related conditions?
   Heart disease
   Heart attack within the last 6 months
   Angina (chest pain)
   Irregular heartbeat
   Heart failure
5. Do you have or have you ever had any of the following?
   Rheumatoid arthritis
   Kidney disease
   Liver disease
   Diabetes
6. Do you get short of breath when you lie flat?
7. Are you currently on oxygen treatment?
8. Do you have a chronic cough that produces any discharge or fluid?
9. Do you have lung problems or diseases?
10. Have you or any blood member of your family ever had a problem other than nausea with any anesthesia?
    If yes, describe:
11. If female, is it possible that you are pregnant?
    Pregnancy test:
    Please list date of last menstrual period:

[a]University of Michigan Health System patient information report. Patients who answer yes to any of questions 2–9 should receive a more detailed clinical evaluation.

Source: Adapted from KK Tremper, P Benedict: Anesthesiology 92:1212, 2000; with permission.

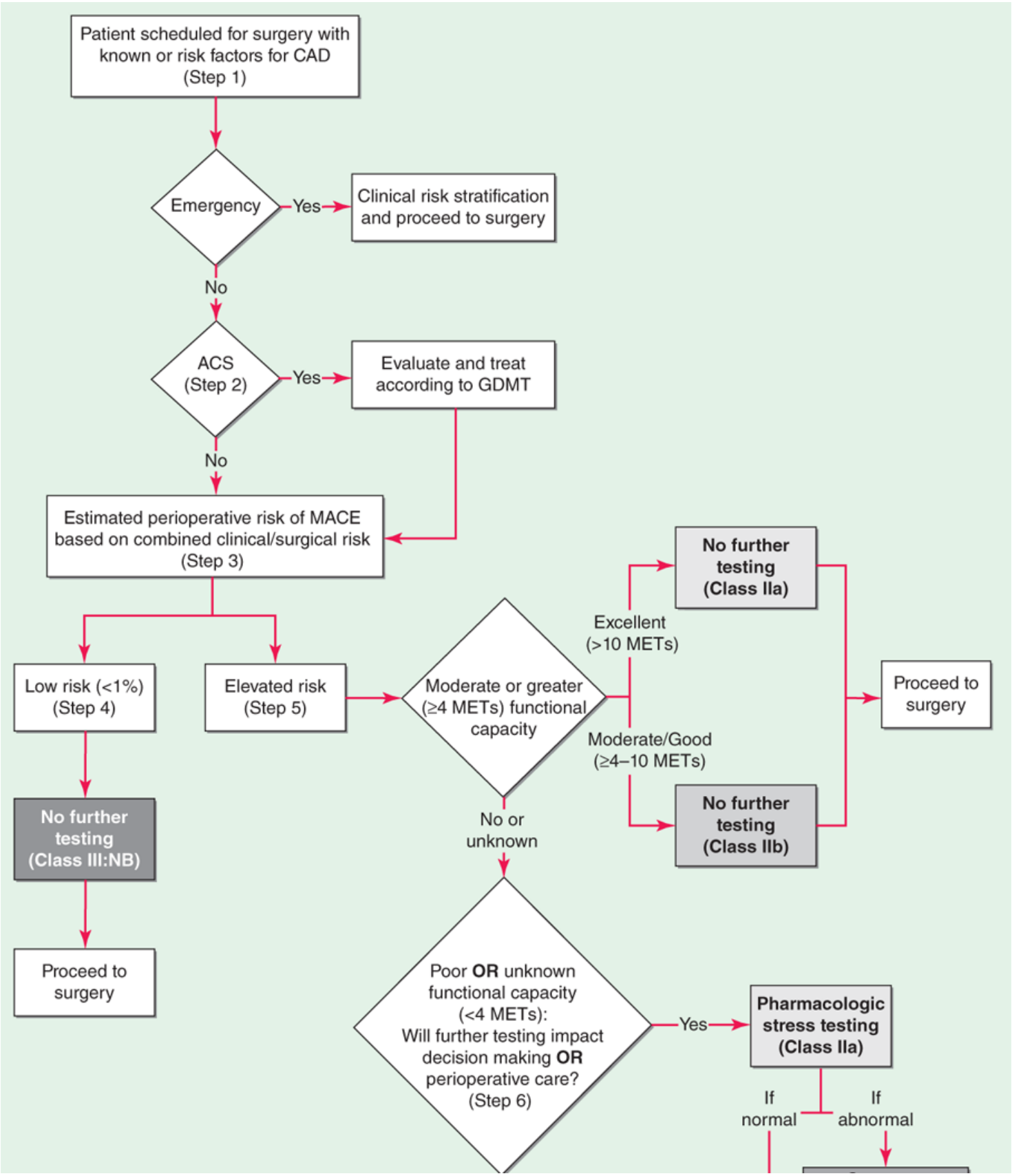
PREOPERATIVE CARDIAC RISK ASSESSMENT

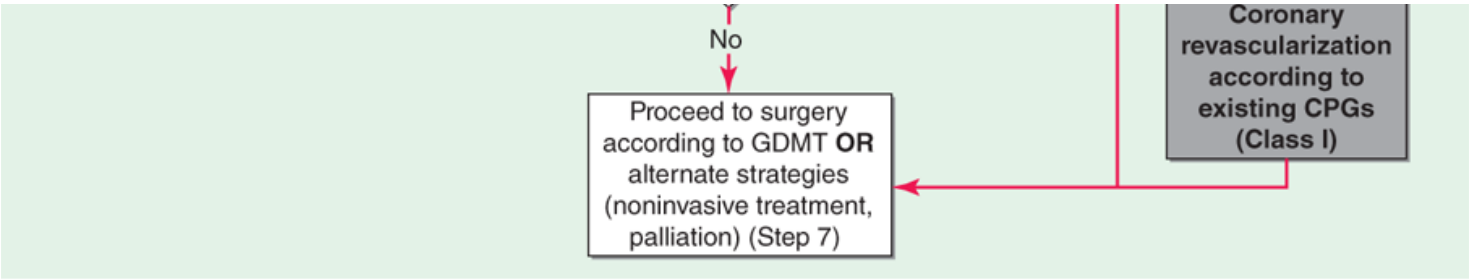
A stepwise approach to cardiac risk assessment and stratification in patients undergoing noncardiac surgery is illustrated in **Fig. 467-1**. The evaluation begins with characterization of the combined surgical and clinical risk into categories of low (<1%) and elevated risk for major adverse cardiovascular events (MACE). Select surgeries are associated with very low risk for MACE; these surgeries and procedures include select ophthalmologic surgeries (e.g., cataract surgery), select endoscopic procedures, and select superficial procedures. Patients undergoing these low-risk procedures should proceed to surgery without further testing. Clinical risk may be estimated with the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) risk calculator (*http://www.riskcalculator.facs.org*) or with calculation of the Revised Cardiac Risk Index (RCRI).

FIGURE 467-1

Composite algorithm for cardiac risk assessment and stratification in patients undergoing noncardiac surgery. Preoperative evaluation involves a stepwise clinical evaluation. Those individuals requiring emergency surgery should proceed without further risk stratification. Acute coronary syndrome (step 2) should be evaluated and treated, accordingly to goal-directed medical therapy. For

patients awaiting non-emergent surgeries and without acute coronary syndrome, perioperative risk is a combination of clinical and surgical risk. Select procedures and surgeries (e.g., select endoscopic procedures) are associated with low perioperative (<1%) risk and no further clinical testing is generally necessary. For those procedures associated with elevated risk, an assessment of functional capacity informs the decision for further testing. Those individuals with moderate or greater functional capacity do not require further testing and should proceed to surgery. Individuals with poor or unknown functional capacity may require pharmacologic stress testing if it would change decision-making or perioperative care. *(From LA Fleisher et al: Circulation 2014;130:e278-e333, with permission.)*





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Previous studies have compared several cardiac risk indices. The American College of Surgeons' National Surgical Quality Improvement Program prospective database has identified five predictors of perioperative myocardial infarction (MI) and cardiac arrest based on increasing age, American Society of Anesthesiologists class, type of surgery, dependent functional status, and abnormal serum creatinine level. However, given its accuracy and simplicity, the RCRI **(Table 467-2)** is often the favored risk index. The RCRI relies on the presence or absence of six identifiable predictive factors: high-risk surgery, ischemic heart disease, congestive heart failure, cerebrovascular disease, diabetes mellitus treated with insulin, and renal insufficiency with a creatinine >2.0 mg/dL. Each of these predictors is assigned one point. The risk of major cardiac events—defined as MI, pulmonary edema, ventricular fibrillation or primary cardiac arrest, and complete heart block—can then be predicted. Based on the presence of none, one, two, three, or more of these clinical predictors, the rate of development of one of these four major cardiac events is estimated to be 0.4, 0.9, 7, and 11%, respectively **(Fig. 467-2)**. An RCRI score of 0 signifies a 0.4–0.5% risk of cardiac events; RCRI 1, 0.9–1.3%; RCRI 2, 4–7%; and RCRI ≥3, 9–11%. The clinical utility of the RCRI is to identify patients with three or more predictors who are at very high risk (≥11%) for cardiac complications and who may benefit from further risk stratification with noninvasive cardiac testing, initiation of preoperative preventive medical management, or avoidance of surgery.

TABLE 467-2

Clinical Markers Included in the Revised Cardiac Risk Index

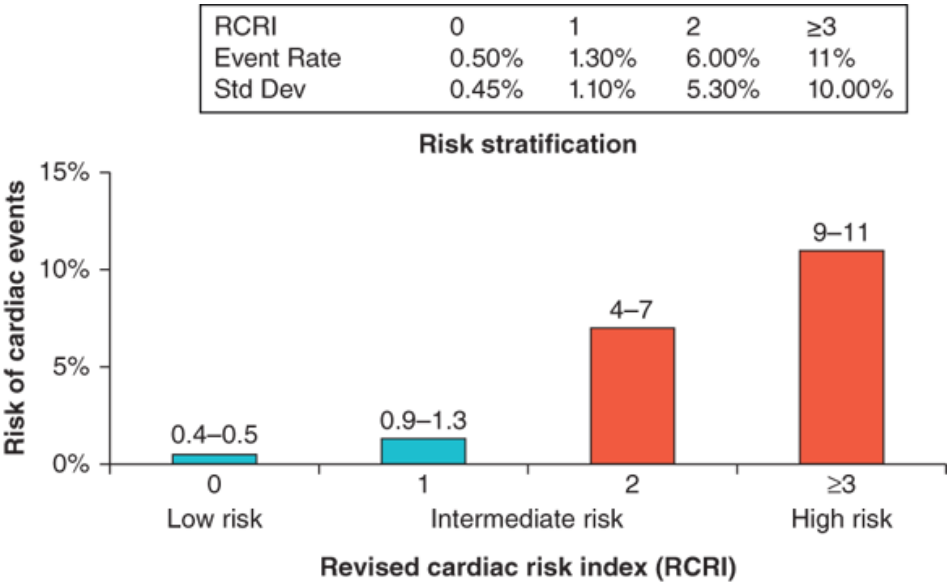
High-Risk Surgical Procedures
Vascular surgery (except carotid endarterectomy)
Major intraperitoneal or intrathoracic procedures
Ischemic Heart Disease
History of myocardial infarction
Current angina considered to be ischemic
Requirement for sublingual nitroglycerin
Positive exercise test
Pathological Q-waves on ECG
History of PCI and/or CABG with current angina considered to be ischemic
Congestive Heart Failure
Left ventricular failure by physical examination
History of paroxysmal nocturnal dyspnea
History of pulmonary edema
S_3 gallop on cardiac auscultation
Bilateral rales on pulmonary auscultation
Pulmonary edema on chest x-ray
Cerebrovascular Disease
History of transient ischemic attack
History of cerebrovascular accident
Diabetes Mellitus
Treatment with insulin
Chronic Renal Insufficiency
Serum creatinine >2 mg/dL

Abbreviations: CABG, coronary artery bypass grafting; ECG, electrocardiogram; PCI, percutaneous coronary interventions.

Source: Adapted from TH Lee et al: Circulation 100:1043, 1999.

FIGURE 467-2

Risk stratification based on the revised cardiac risk index; derivation and prospective validation of a simple index for prediction of cardiac risk in patients undergoing major noncardiac surgery. Cardiac events include myocardial infarction, pulmonary edema, ventricular fibrillation, cardiac asystole, and complete heart block. *(Adapted from TH Lee et al: Circulation 100:1043, 1999.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

For patients at elevated combined clinical and surgical risk for MACE, the stepwise perioperative cardiac assessment for coronary artery disease (CAD) proceeds with consideration of functional capacity. Participation in activities of daily living offers an expression of functional capacity, often expressed in terms of metabolic equivalents (METs). For predicting perioperative events, poor exercise tolerance has been defined as the inability to walk four blocks or climb two flights of stairs at a normal pace or to meet a MET level of 4 (e.g., carrying objects of 15–20 lb or playing golf or doubles tennis) because of the development of dyspnea, angina, or excessive fatigue (Table 467-3). Patients with moderate or greater (≥4 METs) functional capacity (e.g., climbing up a flight of stairs, walking up a hill, or walking on level ground at 4 mph) generally should not undergo further non-invasive cardiac testing prior to elective non-cardiac surgery. Those patients with poor (<4 METs) or unknown functional capacity should undergo pharmacological stress testing if the results of such testing would impact decision-making or perioperative care.

TABLE 467-3

Assessment of Cardiac Risk by Functional Status

Risk	Higher ↕ Lower	• Has difficulty with adult activities of daily living • Cannot walk four blocks or up two flights of stairs or does not meet a MET level of 4 • Is inactive but has no limitations • Is active: easily does vigorous tasks • Performs regular vigorous exercises

Source: From LA Fleisher et al: Circulation 116:1971, 2007.

PREOPERATIVE NONINVASIVE CARDIAC TESTING FOR RISK STRATIFICATION

There is little evidence to support widespread application of preoperative noninvasive cardiac testing for all patients undergoing major surgery. The current paradigm to guide the need for noninvasive cardiac testing is to perform such testing in patients with poor or unknown capacity if it would alter clinical management or modify perioperative care. Options for pharmacological stress testing include dobutamine stress echocardiography or myocardial perfusion imaging with coronary vasodilator stress (dipyridamole, adenosine, or regadenoson) with thallium-201 and/or technetium-99m. Routine screening with noninvasive stress testing is not recommended in patients at low risk for noncardiac surgery. Furthermore, coronary revascularization before noncardiac surgery is not recommended for the express purpose of reducing perioperative cardiac events. That said, revascularization before noncardiac surgery should be considered in patients if it would be indicated regardless of the surgery planned and instead according to clinical practice guidelines. In the Coronary Artery Revascular Prophylaxis trial, there were no differences in perioperative and long-term cardiac outcomes with or without preoperative coronary revascularization; of note, patients with left main disease were excluded.

RISK MODIFICATION: PREVENTIVE STRATEGIES TO REDUCE CARDIAC RISK

Perioperative Coronary Revascularization

Prophylactic coronary revascularization with either coronary artery bypass grafting (CABG) or percutaneous coronary intervention (PCI) provides no short- or mid-term survival benefit for patients *without* left main CAD or three-vessel CAD in the presence of poor left ventricular systolic function and is *not recommended* for patients with stable CAD before noncardiac surgery. Although PCI is associated with lower procedural risk than is CABG in the perioperative setting, the placement of a coronary artery stent soon before noncardiac surgery may increase the risk of bleeding during surgery if dual antiplatelet therapy (DAPT) (aspirin and thienopyridine) is administered; moreover, stent placement shortly before noncardiac surgery increases the perioperative risk of MI and cardiac death due to stent thrombosis if such therapy is withdrawn prematurely (Chap. 270). It is recommended that, if possible, elective noncardiac surgery be delayed 30 days after placement of a bare metal intracoronary stent and ideally for 6 months after deployment of a drug-eluting stent. Contemporary stent platforms allow for greater flexibility in the earlier interruption of DAPT; current clinical practice guidelines do suggest consideration of elective noncardiac surgery 6 months after drug eluting stent (DES) implantation if the risk of further delaying surgery exceeds the risk of stent thrombosis/myocardial ischemia. For patients who *must* undergo noncardiac surgery early (>14 days) after PCI, balloon angioplasty without stent placement appears to be a reasonable alternative because DAPT is not necessary in such patients.

PERIOPERATIVE PREVENTIVE MEDICAL THERAPIES

The goal of perioperative preventive medical therapies with β-adrenergic antagonists, hydroxymethylglutaryl-coenzyme A (HMG-CoA) reductase inhibitors (statins), and antiplatelet agents is to reduce perioperative adrenergic stimulation, ischemia, and inflammation, all of which are heightened during the perioperative period.

B-ADRENERGIC ANTAGONISTS

The use of perioperative beta blockade should be based on a thorough assessment of a patient's perioperative clinical and surgery-specific cardiac risk (e.g., as with the RCRI). The paradigm for beta blockade in the perioperative period has shifted in recent years owing, firstly, to the publication of the PeriOperative Ischemic Evaluation (POISE) trial demonstrating that, while perioperative beta blockade reduces the perioperative risk for MI, this is at the expense of increased death and stroke. Regarding POISE, this trial has been criticized for the use of an excessive dose of beta blocker in the perioperative period and one that may not be reflective of clinical practice, nor one that was titrated in the days or weeks preceding the procedure or surgery. Secondly, research misconduct has discredited the Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography (DECREASE) family of studies, which previously contributed to the bedrock of data supporting the use of perioperative beta blockade but have now been retracted.

Current guidelines emphasize the following key points:

1. Continuation of beta blockade in patients undergoing surgery and who have been receiving such therapy chronically.
2. Avoidance of beta-blocker withdrawal or initiation on the day of surgery.
3. Consideration of initiation of beta-blocker therapy perioperatively (ideally far enough in advance to assess safety and tolerability) in very select high-risk patients, namely, those with intermediate- or high-risk ischemia or three more RCRI risk factors.

HMG-COA REDUCTASE INHIBITORS (STATINS)

A number of prospective and retrospective studies support the perioperative prophylactic use of statins for reduction of cardiac complications in patients with established atherosclerosis. For patients undergoing noncardiac surgery and currently taking statins, statin therapy *should be continued* to reduce perioperative cardiac risk. Initiation of statin therapy is *reasonable* for patients undergoing vascular surgery independent of clinical risk. Perioperative initiation of statin therapy should be considered in patients undergoing elevated risk procedures if there is an indication for such therapy separate from the surgery and according to clinical practice guidelines.

ANGIOTENSIN-CONVERTING ENZYME (ACE) INHIBITORS

It is important to maintain continuity of therapy with ACE inhibitors (when such therapy is used for the treatment of heart failure or hypertension).

ORAL ANTIPLATELET AGENTS

The 4- to 6-week period following implantation of an intracoronary stent (bare metal or drug eluting) constitutes the period of time of greatest risk for the development of stent thrombosis. If possible, noncardiac surgery should be avoided in this vulnerable period. The duration of DAPT thereafter is dictated by the circumstances in which PCI was performed and whether the indication was stable ischemic heart disease or acute coronary syndrome. For the former among patients treated with a drug eluting stent, dual anti-platelet therapy should be given for at least 6 months. For the latter, dual anti-platelet therapy should be given for at least 12 months. However, DAPT may be interrupted to allow for noncardiac surgery 30 days after BMS and 6 months after DES, respectively. If $P2Y_{12}$ inhibitor therapy (clopidogrel, prasugrel, or ticagrelor) is interrupted or discontinued in patients who have received intracoronary stents, aspirin should be continued perioperatively (save select circumstances where the risk of bleeding may be catastrophic as in neurosurgical or spinal procedures) and the $P2Y_{12}$ receptor inhibitor should be restarted as soon as possible post-operatively. Decisions surrounding antiplatelet management in the perioperative setting among patients who have received intracoronary stents are complex and should involve multidisciplinary decision-making.

ALPHA$_2$ AGONISTS

Based on the results of POISE-2 (a large multicenter, international, blinded randomized clinical trial of aspirin and clonidine), α_2 agonists for prevention of cardiac events are not recommended in patients who are undergoing noncardiac surgery. In this trial, clonidine increased the rate of nonfatal cardiac arrest and clinically important hypotension, while reducing the rate of death or nonfatal MI.

CALCIUM CHANNEL BLOCKERS

Evidence is lacking to support the use of calcium channel blockers as a prophylactic strategy to decrease perioperative risk in major noncardiac surgery.

ANESTHETICS

Mortality risk is low with safe delivery of modern anesthesia, especially among low-risk patients undergoing low-risk surgery **(Table 467-4)**. Inhaled anesthetics have predictable circulatory and respiratory effects: all decrease arterial pressure in a dose-dependent manner by reducing sympathetic tone and causing systemic vasodilation, myocardial depression, and decreased cardiac output. Inhaled anesthetics also cause respiratory depression, with diminished responses to both hypercapnia and hypoxemia, in a dose-dependent manner; in addition, these agents have a variable effect on heart rate. Prolonged residual neuromuscular blockade also increases the risk of postoperative pulmonary complications due to reduction in functional residual lung capacity, loss of diaphragmatic and intercostal muscle function, atelectasis, and arterial hypoxemia from ventilation-perfusion mismatch.

TABLE 467-4

Gradation of Mortality Risk of Common Noncardiac Surgical Procedures

Higher	• Emergent major operations, especially in the elderly • Aortic and other noncarotid major vascular surgery (endovascular and nonendovascular) • Prolonged surgery associated with large fluid shift and/or blood loss
Intermediate	• Major thoracic surgery • Major abdominal surgery • Carotid endarterectomy surgery • Head/neck surgery • Orthopedic surgery • Prostate surgery
Lower	• Eye, skin, and superficial surgery • Endoscopic procedures

Source: From LA Fleisher et al: Circulation 116:1971, 2007, with permission.

Several meta-analyses have shown that rates of pneumonia and respiratory failure are lower among patients receiving neuroaxial anesthesia (epidural or spinal) rather than general anesthesia. However, there were no significant differences in cardiac events between the two approaches. Evidence from a meta-analysis of randomized controlled trials supports postoperative epidural analgesia for >24 h for the purpose of pain relief. However, the risk of epidural hematoma in the setting of systemic anticoagulation for venous thromboembolism prophylaxis (see below) and postoperative epidural catheterization must be considered.

PREOPERATIVE PULMONARY RISK ASSESSMENT

Perioperative pulmonary complications occur frequently and lead to significant morbidity and mortality. Clinical practice guidelines recommend the following:

1. All patients undergoing noncardiac surgery should be assessed for risk of pulmonary complications **(Table 467-5)**.
2. Patients undergoing emergency or prolonged (3–4 h) surgery; aortic aneurysm repair; vascular surgery; major abdominal, thoracic, neurologic, head, or neck surgery; and general anesthesia should be considered to be at elevated risk for postoperative pulmonary complications.
3. Patients at higher risk of pulmonary complications should undergo incentive spirometry, deep-breathing exercises, cough encouragement, postural drainage, percussion and vibration, suctioning and ambulation, intermittent positive-pressure breathing, continuous positive airway pressure, and selective use of a nasogastric tube for postoperative nausea, vomiting, or symptomatic abdominal distention to reduce postoperative risk. Multiple pulmonary risk indices are available to estimate the postoperative risk of respiratory failure, pneumonia, and other pulmonary complications; among these is the ARISCAT risk index, which accounts for the following seven risk factors: age, low preoperative oxygen saturation, respiratory infection within the preceding month, upper abdominal or thoracic surgery, surgery lasting >2 h, and emergency surgery **(Table 467-6)**.
4. Preoperative spirometry and chest radiography should not be used routinely for predicting risk of postoperative pulmonary complications but may be appropriate for patients with chronic obstructive pulmonary disease or asthma.
5. Spirometry is of value before lung resection in determining candidacy for coronary artery bypass; however, it does not provide a spirometric threshold for extrathoracic surgery below which the risks of surgery are unacceptable.

6. Pulmonary artery catheterization, administration of total parenteral nutrition (as opposed to no supplementation), or total enteral nutrition have no consistent benefit in reducing postoperative pulmonary complications.

TABLE 467-5

Predisposing Risk Factors for Pulmonary Complications

1. Upper respiratory tract infection: cough, dyspnea
2. Age >60 years
3. Chronic obstructive pulmonary disease
4. Cigarette use
5. American Society of Anesthesiologists Class ≥2
6. Functional dependence
7. Congestive heart failure
8. Serum albumin <3.5 g/dL
9. Obstructive sleep apnea
10. Impaired sensorium (confusion, delirium, or mental status changes)
11. Abnormal findings on chest examination
12. Alcohol use
13. Weight loss
14. Spirometry threshold before lung resection
    a. FEV_1 <2 L
    b. MVV <50% of predicted
    c. PEF <100 L or 50% predicted value
    d. Pco_2 ≥45 mmHg
    e. Po_2 ≤50 mmHg

Abbreviations: FEV_1, forced expiratory volume in 1 s; MVV, maximal voluntary ventilation; PEF, peak expiratory flow rate; Pco_2, partial pressure of carbon dioxide; Po_2, partial pressure of oxygen.

Source: A Qaseem et al: Ann Intern Med 144:575, 2006. Modified from GW Smetana et al: Ann Intern Med 144:581, 2006, and from DN Mohr et al: Postgrad Med 100:247, 1996.

TABLE 467-6

Risk Modification to Reduce Perioperative Pulmonary Complications

Preoperatively
• Cessation of smoking for at least 8 weeks before and until at least 10 days after surgery
• Training in proper lung expansion techniques
• Inhalation bronchodilator and/or steroid therapy, when indicated
• Control of infection and secretion, when indicated
• Weight reduction, when appropriate
Intraoperatively
• Limited duration of anesthesia
• Avoidance of long-acting neuromuscular blocking drugs, when indicated
• Prevention of aspiration and maintenance of optimal bronchodilation
Postoperatively
• Optimization of inspiratory capacity maneuvers, with attention to: • Mobilization of secretions • Early ambulation • Encouragement of coughing • Selective use of a nasogastric tube • Adequate pain control without excessive narcotics

Source: From VA Lawrence et al: Ann Intern Med 144:596, 2006, and WF Dunn, PD Scanlon: Mayo Clin Proc 68:371, 1993.

PERIOPERATIVE MANAGEMENT AND PROPHYLAXIS

DIABETES MELLITUS

(See also Chaps. 396–398) Many patients with diabetes mellitus have significant symptomatic or asymptomatic CAD and may have silent myocardial ischemia due to autonomic dysfunction. Intensive (versus lenient) glycemic control in the perioperative period is generally not associated with improved outcomes, and may increase the risk of hypoglycemia. Practice guidelines advocate a target glucose range from 100 to 180 mg/dL in the perioperative period. Oral hypoglycemic agonists should not be given on the morning of surgery. Perioperative hyperglycemia should be treated with IV infusion of short-acting insulin or SC sliding-scale insulin. Patients whose diabetes is diet controlled may proceed to surgery with close postoperative monitoring.

INFECTIVE ENDOCARDITIS

(See also Chap. 123) Prophylactic antibiotics should be administered to the following patients before dental procedures that involve manipulation of gingival tissue, manipulation of the periapical region of teeth, or perforation of the oral mucosa: those with prosthetic

cardiac valves (including transcatheter prosthetic valves); prosthetic material used in valve repair (annuloplasty ring or artificial chord); previous infective endocarditis; cardiac transplant recipients with valvular regurgitation from a structurally abnormal valve; and unrepaired cyanotic congenital heart disease or repaired congenital heart disease, with residual shunts or valvular regurgitation at the site of adjacent to the site of a prosthetic patch or prosthetic device.

VENOUS THROMBOEMBOLISM

(See also Chap. 273) Perioperative prophylaxis of venous thromboembolism should follow established guidelines of the American College of Chest Physicians. Aspirin is not supported as a single agent for thromboprophylaxis. Low-dose unfractionated heparin (≤5000 units SC bid), low-molecular weight heparin (e.g., enoxaparin, 30 mg bid or 40 mg qd), or a pentasaccharide (fondaparinux, 2.5 mg qd) is appropriate for patients at moderate risk; unfractionated heparin (5000 units SC tid) is appropriate for patients at high risk. Graduated compression stockings and pneumatic compression devices are useful supplements to anticoagulant therapy or in patients at excessive bleeding risk.

FURTHER READING

Fleisher LA et al: 2014 ACC/AHA Guideline on perioperative cardiovascular evaluation and management of patients undergoing noncardiac surgery. Circulation 130:e278, 2014. [PubMed: 25085961]

Levine GN et al: 2016 ACC/AHA guideline focused update on duration of dual antiplatelet therapy in patients with coronary artery disease. A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 68:1082, 2016. [PubMed: 27036918]

Nishimura RA et al: 2017 AHA/ACC focused update of the 2014 AHA/ACC guideline for the management of patients with valvular heart disease. A report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. Circulation 135:1, 2017. [PubMed: 28028057]

Smetana GW et al: American College of Physicians. Preoperative pulmonary risk stratification for noncardiothoracic surgery: Systematic review for the American College of Physicians. Ann Intern Med 144:581, 2006. [PubMed: 16618956]

Chapter 468: Behavioral Economics and Health

Kevin G. Volpp; George Loewenstein; David A. Asch

INTRODUCTION

Tobacco use, physical inactivity, unhealthy diet, excessive alcohol use, and other individual behaviors are estimated to underlie 40% of premature mortality in the United States. Approximately 75% of the $3 trillion currently spent on health care in the United States is attributable to cancer, heart disease, type 2 diabetes, and obesity, and each of these conditions is strongly influenced by behavior. Nearly one-half of patients prescribed medications to lower their cholesterol within 1 year following myocardial infarction stop taking these drugs—even when they are provided free of charge. Despite great advances in the science and technology of health care, a large gap separates theoretically achievable goals in health and health care from what individuals and populations actually reach. Human behavior is both a major contributor to health problems and a barrier to the successful implementation of solutions to address them.

Recognizing that there are many reasons people do not take action to improve their health, behavior change experts have focused efforts on strategies targeted at individual behavior, introducing incentives for weight loss and smoking cessation, environmental strategies such as mandated food labeling, and combination approaches. For example, Section 2705 of the Affordable Care Act (ACA) allows employers to provide incentives of up to 50% of total premiums based on outcomes such as reduced body mass index, lowered blood pressure or cholesterol, and smoking cessation; this legislation puts as much as $300 billion worth of employee health incentives in play annually.

Many of these approaches have built-in limitations, largely because they have been designed around the pervasive view that people always act to improve their self-interest. Existing policy solutions presuppose that health care decisions are rationally based economic transactions and that rational people will dispassionately assess the net present value of the costs and benefits of alternative paths and pursue the best path forward. These approaches are normatively appealing but seem better suited to support the health of people who behave as economists assume they do, and perhaps less effective when exposed to such realities of human behavior as limited attention, overconfidence, and problems of self-control. It is not just the magnitude of incentives that matters, but also other critical features such as the specific nature of rewards, feedback frequency, saliency, and framing. Public health programs, including those involving financial incentives, are more likely to achieve their goals if designed based not on how perfectly rational people *ought* to make health decisions but rather on how humans *actually* make them. The field of behavioral economics, which uses insights from psychology to identify ways that human decision-making often falls short of the ideal, and which offers a more realistic account of human behavior, provides a natural framework for such efforts, including a range of insights relevant to understanding how people respond to financial and nonfinancial (e.g., health) incentives.

CONCEPTS OF CLASSICAL ECONOMICS AND HOW THEY DIFFER FROM BEHAVIORAL ECONOMICS

Classical economics posits that individuals are rational utility maximizers, meaning that they are able to dispassionately identify alternative decisions, calculate the probabilities of and utility/disutility for each potential outcome, and then, through a process of backward induction, implement the decision that has the highest net present value. When it comes to health behaviors, classical economic theory would assume that if people are obese they must have decided that the costs of obesity are outweighed by the benefits of the behaviors that lead to it and that if people smoke they must have decided that the pleasures of doing so outweigh the costs. This assumption of rational utility maximization has two major consequences for public policy, both of which are applicable in the health domain.

The first is that conventional economic thinking radically limits the range of situations in which it makes sense to intervene at a policy level. Under the assumptions of conventional economics, regulatory interventions such as targeted taxes and subsidies are considered appropriate only in situations characterized by externalities (costs that an individual's actions impose on others such as second-hand cigarette smoke), the presence of "market failures" (e.g., monopolies), or the presence of certain types of information asymmetries. The second is that standard economics offers a relatively restricted array of policy tools, including taxes, subsidies, and mandates regarding information provision, and makes unrealistic assumptions about how each of these approaches will influence the behavior of individuals.

Behavioral economics builds on conventional economics by enriching its conception of individual behavior, offering enhancements on both of these dimensions. First, it broadens the range of situations in which policy interventions make sense by introducing the notion of "internalities." While externalities are costs (or benefits) individual behaviors impose on others, internalities are the costs individuals impose on themselves—typically their future selves. While interventions to address smoking and obesity can be justified on externality grounds (e.g., the health care costs are borne by individuals other than the immediately affected individual), they can also be justified on grounds of internalities—e.g., people often irrationally discount the delayed consequences of their behavior. In that sense people might want to be protected not just from others, but from the decisions of their prior selves. Policy intervention can be further justified by the ubiquitous exploitation of individual weaknesses by businesses. Commercial enterprises may take advantage of such vulnerabilities for individual profit rather than customer health, such as with the formulation of tempting but unhealthy processed foods or the pricing of meal "deals" that do not account for the health consequences of economically predatory offers. This behavior is found across industries: Credit card companies and automobile manufacturers lure new customers with "$0 down" and fleeting but appealing teaser rates of "0% interest," playing on the common propensity to focus on the present rather than on the future. Banks earn revenue by charging high fees for minor mistakes such as account overdrafts or breaches of minimum balance rules hiding the description of such fees in small print and complicated jargon. States market lottery tickets that return $0.45 on the dollar and promote these games in ways that ignore more realistic expectations using one-sided messages such as "you can't win if you don't play" rather than, for example, the equally accurate message, "you can't lose if you don't play."

Second, behavioral economics substantially broadens the range of potential policy interventions far beyond those offered by traditional economics. Behavioral economics has become best known for the concepts of "libertarian paternalism" and "asymmetric paternalism." In contrast to "heavy-handed" paternalism, asymmetric paternalism attempts to protect people without limiting freedom of choice. That is, it is asymmetric because it seeks to help individuals who are prone to making irrational decisions without restricting the freedom of choice of others making informed, deliberate decisions. For example, arranging the presentation of food in a cafeteria line so that the healthy foods appear first is likely to increase the amount of healthy food chosen, without depriving those who want the unhealthy foods of the opportunity to purchase them. People who believe that individuals behave optimally should not

object to asymmetric paternalism because it does not limit freedom, and those who acknowledge the limits of human rationality should endorse such measures. Frequently, these measures are called “nudges.”

A nudge has been defined as “any aspect of the choice architecture that alters people’s behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid.” The most prominent, and to date successful, application of nudges has been the use of defaults to increase enrollment in defined contribution retirement savings plans, and secondarily the use of automatic escalation to encourage higher savings rates. These ideas and research findings have had a major impact on retirement savings policies worldwide, including the Pension Protection Act of 2006 in the United States. Building on this success story in savings, and bolstered by the establishment of so-called “nudge units” worldwide, the nudge agenda has positioned behavioral economics at the center of public policy.

The applicability of behavioral economics to policy, including health-related policies, however, goes well beyond nudges. Many health-relevant insights from behavioral economics do not fit the definition of a nudge. For example, there have been incentive programs aimed at changing health behaviors for maximum cost-effectiveness, improvements in the delivery of health-related information, such as nutrition labels, and new designs for physician incentives or health insurance. A common theme of this work harnesses natural tendencies toward predictable decision errors by redirecting those tendencies to help people achieve longer-term health goals or other socially valuable purposes, much the way some martial arts redirect an adversary’s strength against him.

Many of the same messages, incentives, and choice structures used so effectively to lure people into self-destructive health behaviors can be redirected to attract them to healthier choices that improve their long-term health and well-being. Some features of human decision-making, such as our propensity to experience regret when we make a bad choice, and our aversion to putting ourselves into such situations (a phenomenon known as “regret aversion” and arguably not really an example of an error) are additionally important features of human psychology that can be exploited in the service of improving health behaviors and outcomes.

There is a common misconception that if you deploy financial incentives in order to promote behavior change, then you are engaged in behavioral economics. But that kind of activity is not behavioral economics; it is simply economics. Indeed, a large number of everyday transactions, such as being paid to go to work, or getting a fine for parking in the wrong place, reflect traditional economic incentives to encourage or discourage certain behaviors. A central lesson from the field of behavioral economics is that how incentives are delivered can matter more than their objective magnitude. There are ways of delivering large incentives that make them ineffective in changing behavior, and there are ways that can greatly magnify the effectiveness of comparatively small incentives. This observation is a source of optimism, implying that with careful design we can leverage relatively small investments to improve public health.

USING BEHAVIORAL ECONOMICS TO PROMOTE SELF-BENEFICIAL HEALTH BEHAVIORS

Behavioral economics builds on neoclassical economics, which has at its core *expected utility theory*. Expected utility theory both presumes to describe how people make decisions and offers a prescription for how such decisions should be made. While utility maximization is a powerful normative model of how we ought to behave, it turns out to be a poor descriptive model of how real people actually behave. Efforts to alter human behavior that rely on this incomplete model often fall short.

Over the past several decades, behavioral economics has described various ways in which people's decisions differ from standard economic models (Table 468-1). While economists mapped out concepts of "bounded rationality" in the 1950s and identified limitations in the dominant expected utility model of decision making under risk, the publication of Prospect Theory by Kahneman and Tversky is widely credited with being seminal in the development of behavioral economics. Prospect theory provides an overarching conceptual framework for describing observations about human behaviors that could not be explained by expected utility theory.

TABLE 468-1

Traditional vs Behavioral Economics

Traditional Economics	Behavioral Economics
Core theory: Expected utility maximization	Core theory: Prospect theory
Assumes perfect rationality	Recognizes that people make decision errors
Starting point independent	Assessment depends on your starting point
Framing doesn't matter	Framing affects assessment even when utilities are the same
Stable preferences	Time-inconsistent preferences
People discount the future at constant rates	People discount the near future to a greater degree and have time inconsistent discounting
Intervene only when my actions adversely affect others (negative externalities)	Consider interventions when people will harm their future selves (internalities)
Regulations and policies generally geared to protecting people from the actions of others	Regulations and policies often geared to protecting people from themselves

Although these deviations from expected utility theory can be seen as psychological foibles and errors, the real value of this work has come less from identifying these errors, and more from recognizing that they occur in predictable ways. It is the predictability of these errors that allows the design of strategies to overcome them (Table 468-2). Indeed, it is the predictability and reliability of a number of key features of human decision-making identified by behavioral economics, including but not limited to those connected to prospect theory, allows them to be used in order to promote self-advantageous health-related behaviors.

TABLE 468-2

Key Decision Errors and Suggestions for Addressing Them

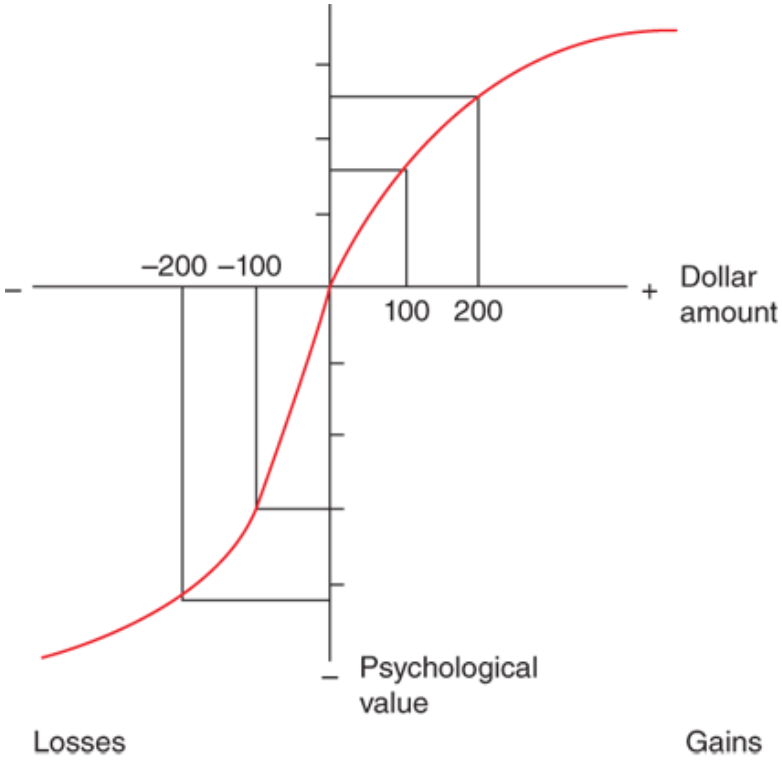
Present-biased preferences	Provide feedback and rewards quickly
Nonlinear probability weighting	Motivate people with probabilistic rewards (lotteries)
Overoptimism and loss aversion	Get people to precommit and put money at risk as an effective motivational tool
Peanuts effect	Deliver rewards in bundles, avoiding many small rewards
Narrow bracketing	Frame rewards in terms of effort per day rather than per month or year
Regret aversion	Help people anticipate the regret of poor choices
Defaults/status quo bias	Change the architecture or environment of choice to shift the path of least resistance to favor healthy decisions
Rational world bias	Move beyond the assumption that simply providing information will lead to desired behaviors

Loss Aversion and Framing Effects

Key tenets of prospect theory include (1) how people feel about a set of possible outcomes depends on their starting point; this is the notion of reference-dependence in which decision-makers evaluate outcomes as gains or losses depending on their starting point; (2) people dislike losses much more than they like equivalent-sized gains, a phenomenon known as “loss aversion”; and (3) there is diminished sensitivity to both gains and losses **(Fig. 468-1)**.

Figure 468-1

Prospect theory. (*From D Kahneman: Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux, 2011.*)



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

One example of how we can apply the insights of loss aversion can be seen in designing provider payment systems. Financial incentives to physicians to achieve quality goals might be presented—"framed" in the language of behavioral economics—as either rewards (in which physicians receive a bonus if, for example, they increase their colorectal cancer screening rates or increase the percentage of their patients who reach a glycosylated hemoglobin target) or penalties (in which physicians would fail to receive an expected payment if they do not meet these targets). Classical economists would consider equivalently sized rewards and penalties as identically motivating because each reflects a structure in which one receives $X for a certain outcome (Fig. 468-1). However, loss aversion reminds us that the disutility of losing money is much greater than the utility of gaining the same amount of money. A number of studies have shown that people have a "loss aversion ratio" in a range of 1.5 to 2.5. That means that a potential penalty of $1000 for failing to meet a quality target ought to be about as potent a motivator as a potential reward of $1500 to $2500 for meeting the same target. That multiplier is nonsensical from the standpoint of classical economics, but as an empirically verifiable descriptor of human behavior, it can be exploited in the designs of programs for clinicians or patients to improve health.

In a now classic example in public health, Tversky and Kahneman presented research participants with the following problem:

> *Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows:*
>
> > *If program A is adopted, 200 people will be saved*
> >
> > *If program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no one will be saved.*

Given these choices, a substantial majority (about 70%) choose Option A. People preferred to remove uncertainty in favor of the sure bet of saving 200 lives.

A randomly selected second group was presented with an alternative frame:

> *If program A' is adopted, 400 people will die.*
>
> *If program B' is adopted, there is a one-third probability that no one will die and a two-thirds probability that 600 people will die.*

About 70% of participants in this group choose option B'. However, in a population of 600 people, 200 people surviving (option A) is equivalent to 400 people dying (option A') and options B and B' are similarly different statements of the same outcome. When outcomes are viewed as gains, as they are in the choice between A and B, decision-makers tend to avoid gambles and are risk averse, choosing option A. When choosing between two bad outcomes, as in the choice between A' and B', decision-makers tend to take a gamble (be risk seeking), choosing option B'. In this case people would rather take a chance that everyone could be saved rather than consign themselves to a choice that involved a large number of deaths with 100% certainty.

All of us—including physicians—are highly susceptible to how information is framed. In a set of experiments patients, students, and physicians were presented cases of lung cancer that could be treated either by surgery or radiation therapy. Among all three groups, the choice of surgery was more popular when its outcomes were framed in terms of the probability of survival (e.g., a 68% chance of living for more than 1 year) rather than in terms of the probability of death (e.g., a 32% chance of dying by the end of 1 year). We can say that such sensitivity to framing is irrational since a 68% chance of survival is logically equivalent to a 32% chance of dying—but these irrational decisions fall into predictable patterns of behavior, and that predictability can be used to influence those decisions. This means that clinicians have enormous opportunities to lead patients toward particular decisions by framing the outcomes of those decisions in specific ways, even as they remain truthful. Such an understanding could lead to the view that clinicians should be careful to balance their framing (for example, by following up the statement of 68% chance of survival with a statement like “That means there is a 32% chance of death”) in order to provide information in a way that is likely to lead to a particular choice. Alternatively, it could lead to the view that clinicians should deliberately frame outcomes in certain ways in order to lead patients to particular choices—a much more paternalistic stance. Patients often rely on trusted clinicians to help them make the best decisions and, in some settings, that reliance may justify using the principles of behavioral economics strategically even if the same actions in other settings might be seen as paternalistic, anti-libertarian, or coercive.

Loss Aversion and Overoptimism

The power of loss aversion can be most effectively leveraged when combined with a well-documented decision error: overoptimism, or unrealistically high expectations about future outcomes. Overoptimism is especially pronounced in the context of people’s predictions about their own likelihood of exerting self-control, sometimes referred to as the “false hope syndrome.” Although in some contexts overoptimism seems to be beneficial, it can also result in suboptimal patterns of behavior. For example, people prefer paying a flat rate for gym memberships even though they would spend less if they were to pay on a per-visit basis, in part because they overestimate their future gym attendance.

As described earlier, loss aversion reflects the tendency to put greater weight on losses than on equivalently sized gains. It can produce a variety of undesired behaviors, from excessive risk aversion to the tendency for people to hold on to losing investments, such as houses or stocks for too long.

Loss aversion can, in theory, be deployed to advance social goals by framing reward outcomes in terms of losses by "fronting" a sum of money that gets lost if goals are not met, rather than providing equivalent gains for meeting goals (the economic but not psychological equivalent). Yet, despite the greater potency of losses, program administrators are often reluctant to use loss framing, perhaps because such programs can seem more punitive than organizations may wish to appear. However, it is possible to take advantage of loss aversion by designing programs in which people voluntarily put their own money at risk in the service of achieving health-behavior goals that they themselves desire. Many people enter into commercial weight loss programs and pay for a full year up front because they are overly optimistic about their chances of success. Once deposited, however, such optimism can become a self-fulfilling prophecy as loss aversion provides extra motivation to meet goals.

The combination of overoptimism and loss aversion has been used to help people lose weight by giving them the opportunity to participate in deposit contracts, in which they could deposit $0.01–$3.00 per day of their own money, with a 100% match. Participants reported their weight daily and received the sum of the deposit and the matching funds each day they were on track to meeting their monthly weight loss targets, but they forfeited both if they were not on track. The deposit contract leveraged participants' overly optimistic self-predictions of how much weight they would lose as well as loss aversion once deposits were made. In this 16-week study, average weight loss was 14.0 pounds in the deposit group compared with 3.9 pounds in the control group. This work was extended in a 32-week study in which weight loss was sustained for the duration of the intervention (8.7 vs 2.2 pounds in the control group). Although these results are promising, to be effective as a health strategy this approach needs to achieve high ongoing participation rates to sustain its population effect. Deposit contract approaches are powerful motivators of behavior change, but they are not always popular even among those who initially opt to try them, posing a challenge to long-term success.

Peanuts Effects

The prospect theory value function assumes diminishing marginal utility, which means that small gains and losses are disproportionately motivating relative to larger ones (e.g., two $500 rewards would be more potent than one $1000 reward). However, this may be too simple for small rewards (e.g., two $5 rewards may be less motivating than one $10 reward). This "peanuts effect" may be part of the reason why charities and retailers often describe costs in terms of "pennies a day." This observation challenges the expected efficacy of programs that emphasize efforts to repeatedly achieve small changes such as in weight loss programs. It is easy for a patient to rationalize that no single cigarette causes lung cancer or that no single trip to the gym prevents heart disease. If self-destructive patterns of behavior, such as cigarette smoking, weight gain, or cell phone use while driving, are seen as individual instances rather than parts of a composite whole, it is easier to understand how they can be seen as acceptable. The pleasure of smoking a cigarette or eating a dessert and the convenience and engagement of conducting business or socializing in otherwise "dead" travel time are immediate and tangible, but the marginal costs—increased risks of developing lung cancer, being overweight, or having a car accident—seem inappreciably small. Across a lifetime or a population, however, the cumulative costs and/or probabilities are not small at all.

The tendency to underweight small events can also be used to people's advantage, such as by inducing people to put away small sums for retirement savings in short cycle lengths or to make small periodic investments in their health via medication adherence. In each of these efforts, we need to think asymmetrically. In our incentive programs we should provide frequent (often daily) feedback on rewards because of present-biased preferences, but if we are *delivering* financial rewards, we want to aggregate them so that the rewards appear substantial enough to warrant attention. To use the peanuts effect to advantage, one might alert people to their rewards daily but then deliver them monthly to create larger aggregate payments.

PRESENT-BIASED PREFERENCES

Another central observation from behavioral economics is the concept of hyperbolic discounting, or "present-bias." It is standard in conventional economics to assume that people discount the future; for example, $1000 today is worth more than $1000 a year from now, since money can be invested and earn interest. However, people tend to discount outcomes that are close in time more steeply than outcomes that are farther off in time; the degree of time discounting is disproportionately greater for short time delays than for long ones, in contrast to the assumption of standard economic models.

The medical implications of present bias are profound. For example, most people would desperately like to avoid a stroke and many patients with hypertension have an understanding that taking their antihypertensive medications is one of the best ways to avoid a stroke in the future. While the classical economist would see daily adherence to antihypertensive treatment as "a good investment" to avoid a future stroke, the stroke that is avoided is far in the future and uncertain; moreover, the stroke that is prevented is never noticed. In contrast, even the relatively small effort required to stay on antihypertensive medication is immediate, continuous, and comes without any immediate compensatory benefit. To the extent that patients overly discount the future harms of a later stroke, they will be less motivated to invest today in their own medication adherence. To the classical economist, these patients are behaving irrationally because they are failing to make investments that, if they did the calculations, would clearly make them more likely to be better off; this reasoning is parallel to how classical economists consider people who undersave for retirement (the vast majority of Americans) to be irrational. To the behavioral economist, these errors are targets for therapy, in the way that we can see genetic mutations or defects in chemical pathways as therapeutic targets in the management of illness.

Present bias is somewhat subtler than simply steep discounting of the future. In fact, it reflects *two* behavioral tendencies: (1) the tendency to overweight immediate costs and benefits relative to those occurring in the future, as just discussed, and (2) the tendency to take a more evenhanded approach to *future* costs and benefits. People are much more willing to begin dieting *tomorrow*, because the overweighting of immediate costs deters us from the immediate deprivation of dieting, and the more balanced perspective on delayed deprivation makes us willing to impose these costs on ourselves in the future.

Although present-biased preferences typically promote unhealthy behaviors, policy makers can use them for beneficial effects. The motivational impact of benefits and costs, such as rewards for good behavior and punishments for bad behavior, can be increased substantially if they are made immediate. These consequences should coincide as closely as possible with the timing of the behaviors they are meant to encourage or deter. Funds for this could be provided by employers or insurers for whom this might be a cost-effective way to improve worker health and productivity.

Such programs have been shown to have dramatic effects in the area of drug addiction. This success is particularly striking, because many individuals with drug addiction already face major adverse consequences, such as loss of livelihood and disenfranchisement from their families; yet these costs are often insufficient to motivate abstinence. Similarly, small incentives offered on proof of abstinence have succeeded in tripling smoking-cessation rates where the far larger (but delayed) incentives in terms of improved health have failed. Small, daily, lottery-based incentives have significantly increased medication adherence and weight loss in part because they bring immediate rewards (money, excitement) to a situation in which the benefits of avoiding ill health are typically distant and uncertain.

Thus, rather than requiring individuals to make decisions based on consideration of their long-term best interests, it might be useful to change short-term incentives so that beneficial actions are easier and more attractive to choose.

Some school districts have begun to use this approach by removing various products such as soda and candy from vending machines so that the cost of obtaining them now includes a walk off campus while healthier food and beverage options are immediately and readily available. In addition, people's willingness to commit to future changes can be leveraged by giving them choices between health-benefiting and health-harming behaviors well before they actually have to act on them. An example of this is scheduling gym visits and laboratory tests to monitor cholesterol ahead of time and having patients voluntarily accept financial penalties for last-minute cancellations in advance.

Nonlinear Probability Weighting

In discussing prospect theory, we have so far discussed the concepts of reference-dependence, loss aversion, and diminishing sensitivity, collectively the properties that define prospect theory's value function (Fig. 468-1). However, prospect theory also encompasses a second important dimension: the way that people deal with—and weight—probabilities. In contrast to the standard expected utility model, which assumes that people weight outcomes according to their raw probabilities, prospect theory assumes that people overweight small probabilities but are insensitive to differences in probabilities—for example, between a 0.001 and a 0.00001 chance of winning a prize, even though the probabilities differ by several orders of magnitude—except where they provide a transition to certainty. Such overweighting of small probabilities is partly responsible for the enormous attraction of lottery tickets; yet, like present-biased preferences, this overweighting can be used to advantage in public health interventions.

Following these cognitive pathways, lottery-based reward systems have been introduced in programs aimed at motivating diverse health behaviors (described more fully below). These interventions exploit overweighting of small probabilities and also play on other psychological insights. Because people tend to be motivated by both the experience of past rewards and the prospect of future rewards, these lottery-based systems provide frequent small payoffs and infrequent large payoffs. This approach has been demonstrated effective in a variety of areas, including helping people lose weight (52.6% of people achieve 16-week weight loss goals compared with about 10.5% in a control group) and reducing medication nonadherence (from about 23% to about 3%). However, results across studies and contexts have been inconsistent, suggesting that the active ingredients of these lottery-based incentives have not been fully elucidated.

Regret Aversion

People dislike regretting decisions they have made, often voicing laments such as "If only I had …." Moreover, people are sufficiently far-sighted to anticipate possible future regret and seek to make decisions today that reduce that risk in the future. The avoidance of anticipated regret is a useful exception to present-biased preferences.

Regret aversion helps to explain the success of the Dutch postal code lottery, in which winning postal codes are selected and those living within the selected areas who purchased tickets receive prizes. Those who did not purchase tickets learn that they would have won had they done so. Individuals see their neighbors winning large prizes that they do not share, and their desire to avoid future regret drives subsequent lottery participation.

Anticipated regret has been shown to affect a variety of preventive behaviors, such as the significant increase in vaccination use among people who experienced illness after failing to get vaccinated. Lottery-based incentive programs where eligibility is conditioned on adherence (for example, you aren't eligible to play the lottery unless you had taken your medication or checked your blood pressure the previous day) can incorporate regret aversion into their design by notifying both winners and losers. Those whose number comes up but are ineligible for a reward because of nonadherence are told they would have won had they only taken their medication, checked their blood pressure, or

done whatever it was the lottery program was designed to promote. Because people hate the feeling of regret, they are more likely to engage in behaviors to avoid that feeling. Indeed, the advertising campaigns of some traditional lottery systems take advantage of regret aversion. Many people play the same favorite number when the buy lottery tickets. It easy to imagine the disappointment you would feel if you missed buying a ticket one week and that was the week your favorite number came through. Advertising slogans like “don’t let your number win without you” keep people in the game in order to escape that feeling of regret. The same techniques used to promote the marketing of lottery tickets can be deployed toward health promotion.

Defaults

Although most of the interventions playing on behavioral economics that we have discussed so far do not qualify as “nudges,” nudges remain one of the most powerful ways to influence choice. Traditional economic thinking is silent on the power of the default option—the path that is “selected” when no selection is made. However, the default bias, or status quo bias, reflects our tendency to take “the path of least resistance”—to continue doing what we have been doing, or to do what comes automatically, even when superior alternatives exist. Defaults have been blamed for a wide range of suboptimal outcomes, from the failure of employees to put aside retirement funds in companies whose default contribution rate is zero, to suboptimal allocation between investment alternatives, to excessive consumption of French fries and large sodas as part of “supersized” meals at fast-food franchises. In western European countries that have “opt-in” policies for organ donation—that is, the default is nonparticipation (as in the United States), donation rates tend to be in the range of 10%. In contrast, in countries with “opt-out” policies, in which citizens are automatically enrolled as organ donors unless they actively choose not to be, organ donation rates are often close to 99%. Defaults have been shown to increase the rate at which patients with terminal lung diseases choose comfort-oriented plans of care, and they could be used more widely to encourage the choice of beneficial health options that would vary based on the clinical context.

If used tactically, a choice architect (the person who makes decisions on how choices are presented to the end user) can utilize defaults such as changing scheduled automatic prescription refills from 30 to 90 days (or longer) for patients requiring lifelong chronic-disease therapy or changing the default option in fast-food restaurants in combination meals from large sodas to small sodas or water or french fries into carrots unless you ask for french fries to help propel people toward self-beneficial behaviors. A default toward generic prescription embedded in a health system electronic physician order system moved generic prescribing to nearly 100% in a context in which previous attempts at physician education had shown no effect at all.

Such approaches cost nothing since a default has to be set one way or another, preserve freedom of choice, and could change behavior substantially. There is little question we miss out on many opportunities to nudge people toward healthier lifestyles because we haven’t considered the default option as a tactical choice to be actively incorporated into an overall health-promoting strategy.

The Rational-World Bias

Perhaps the most consequential decision error affecting health-related behaviors is the tendency of public health officials and private-sector benefit designers to make policy decisions based on the assumption that people’s choices are deliberative and rational. This in turn leads to assumptions that information provision is all that is needed for optimal decision-making, and that when financial incentives are offered, the amount is all that really matters.

One significant manifestation of the rational-world bias is the complexity of health insurance plans. Health insurance is complicated for many reasons, including insurance companies' desire to shield themselves from costs and, possibly, to hide or "shroud" information that might be perceived as unfavorable. The result is benefit designs with layers of rules about different forms of cost sharing for different kinds of services that vary in and out of networks. A major rationale for this complexity is to include benefit-design elements that incentivize patients to engage in specific cost-minimizing and health-maximizing behaviors across a wide range of activities. For these built-in incentives to exert their desired influence, however, patients need to understand the incentives they are facing, and there is considerable evidence that they do not. This complexity of insurance benefit design, rather than creating perfectly tailored incentives, is part of insurance design's undoing—because an incentive that can't be understood can't be effective. Most consumers lack an understanding of the most basic insurance concepts, such as deductibles, copays, and coinsurance and, when given a simplified version of a traditional insurance plan, are unable to compute the costs they would incur for basic services.

Giving a nod to the benefits of simplicity, the Affordable Care Act in the United States requires insurance plans to be presented in a standardized document that describes plan features such as premiums, deductibles, and coinsurance. However, describing something that is inherently complex in simple terms raises the risk of glossing over important subtleties. A more productive approach would be to provide consumers with a truly simplified insurance product. For example, insurers might create a plan designed with copayments only, since copayments are most easily understood by consumers: copayments are analogous to paying a price for a good or service when shopping, whereas coinsurance and deductibles make it difficult for people to estimate how much their care would cost. It is difficult, if not impossible, for a consumer to estimate 10% coinsurance on a hospital stay or an emergency room visit since they have no idea of the full amount on which the 10% will be calculated. For medical markets and patient cost-sharing elements embedded in plan design to affect decision-making *ex ante*, consumers need some way of accurately gauging not only benefits and risks of treatments but also the costs. This should occur before receiving a diagnostic test or treatment as opposed to what typically happens now: a medical bill is received after the fact.

Another manifestation of the rational-world bias is the more insidious and pervasive belief that if only people knew more about how to advance their health they would do so. This error leads to making education the centerpiece of many health interventions even though public health experts have long recognized that knowledge alone rarely translates to health-enhancing behavior. For example, nearly everyone knows the health hazards of smoking. While raising and maintaining awareness of the dangers of smoking is an important goal, the financial budgets of the Centers for Disease Control and Prevention, or the time budgets of clinicians, might be more efficiently allocated toward efforts to change behavior that don't presume that the deficit is a lack of knowledge. Indeed, it has been argued that smokers tend to *overestimate* the hazards of tobacco, just as many women tend to *overestimate* their risks of developing breast cancer. In these cases, better or more accurate education might lead people to *lower* their estimates of risk and might, if they were perfectly rational, lead them to reduce health-promoting behaviors. In cases where the deficit is not in knowledge but in behavior, reliance on education as a primary avenue for reducing health risks may divert efforts and resources from other activities that might be much more effective.

APPLICATIONS

Weight Loss

Efforts to combat obesity using incentives started in the 1970s. This early work was motivated by the observation that participants who deposited money and other valuables with a therapist and signed contracts in which return of their valuables was contingent on progress towards pre-specified goals lost tremendous amounts of weight. Even though

participants received no training in weight loss or maintenance strategies, participants lost an average of 32 pounds. This small initial study lacked control groups and long-term follow-up but provided an important proof of concept.

In the first systematic study of deposit or pre-commitment contracts, participants responding to an advertisement for a weight loss program were informed that study participation required a deposit of $200 (1974 dollars), which would be fully refunded contingent on satisfactory weight loss. After an 11-session, 10-week program, those who received incentives for losing weight or for limiting calories lost significantly more weight than those who received incentives for attending sessions. Of the participants in the former two groups, 70% lost more than 15 pounds. The major limitation to this approach was that only 15% of the prospective participants who responded to the initial newspaper advertisement ended up enrolling, suggesting that deposit contracts that require participants to commit substantial funds up front are very effective for people who agree to participate, but that this requirement likely deters a substantial portion of high-risk participants from entering and its effectiveness may just as much reflect selection of those highly motivated to lose weight.

A subsequent study tested the effects on weight loss of deposit contracts of $30, $150, or $300, with the deposits returned based on either individual or group weight loss over 15 weeks. Participants in the intervention group could win $1, $5, or $10 per pound lost up to a maximum cumulative weight loss of 2 pounds per week (either individual or group average). Mean weight loss was large in all 3 groups but did not differ significantly based on contract size. However, the proportion who reached the goal of 30 pounds weight loss was significantly higher in the larger dollar groups.

Because it is increasingly difficult to shed pounds as weight is lost, the investigators also tested whether a deposit contract with increasing payments ($5, $10, $20, $40, $75) for each 5-pound increment of weight loss would be more effective than offering $30 for each 5-pound increment of weight loss. Participants in both conditions were also offered a maintenance program requiring a $100 deposit, returned in $25 increments for attendance at follow-up visits every 3 months. The increasing contract resulted in qualitatively larger weight loss during the weight loss phase, but the maintenance program did not prevent weight re-gain, likely because the magnitude of the deposit contract for maintenance was small, and feedback was infrequent (only every 3 months).

Paying participants for weight loss using direct payments was less effective than deposit contracts. In a randomized trial, cash payments up to $25 per week for making 100% of proportional progress toward goal, $12.50 for 50% of goal, and $2.50 for not gaining weight did not result in greater weight loss in the payment group than among control subjects.

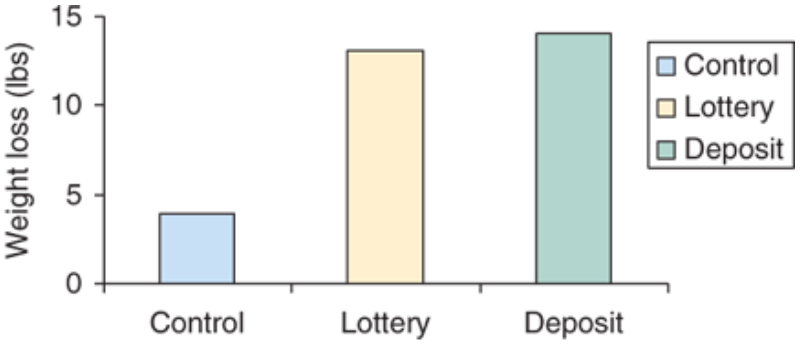
Studies that have shown no effects on either initial weight loss or maintenance typically have used incentives of small magnitude or were targeted at behaviors, like attendance at weight loss programs that, by themselves, do not ensure weight loss. In recent years, weight loss incentives have become a common feature in programs used by employers and health plans and a variety of start-ups like *stickk.com* and *dietbet.com* use deposit contracts as a way of trying to help people lose weight.

Newer approaches have provided proof of concept that daily lottery-type incentives and precommitment contract incentives promote initial weight loss over 16 weeks (lottery = 13.1 lbs; $p = .014$ for lottery vs control; deposit contract = 14.0 lbs; $p = .003$ vs control; Fig. 468-2). However, participants regained most of the weight they had lost over the following 3 months. Longer trials (8 months) found no difference in effectiveness of deposit contracts for continuous 8-month weight loss versus 6-month weight loss with 2 months of maintenance, but both were successful in achieving a mean weight loss of about 10 pounds. Subsequent tests included an employer-based study showing greater

effectiveness of team competitive versus individual incentives, although there was a confound in that participants in the teams had the potential to win more money than those in the individual arms.

Figure 468-2

Weight loss in incentives versus control. *(Figure created using data from KG Volpp et al: Financial incentive-based approaches for weight loss: A randomized trial. JAMA 300:2631, 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

While both of these studies suggest that financial incentives promote weight loss, many employers use premium adjustments (increases or decreases to health insurance payments) as their standard approach to using financial incentives for health promotion. The effectiveness of premium-based financial incentives to promote weight loss has been assessed in a workplace wellness program and with a goal of losing 5% of initial weight over the next year. Participants were randomly assigned to a control group with no other intervention or one of three financial incentive groups. Two intervention groups were offered a premium reduction of $550 if they achieved their weight loss goal by 1 year. The “delayed” group would receive the premium adjustment in the following year, spread across each pay period. The “immediate” group would have their premiums adjusted as soon as the weight loss goal was met. A third intervention group was offered a daily lottery with about a 1 in 5 chance of winning $10 and a 1 in 100 chance of winning $100. To be eligible to win each day, participants had to weigh in and be on track to lose 5% of their initial weight loss by 6 months, with maintenance for the subsequent 6 months. After 1 year in the program, none of the intervention approaches showed a significant degree of weight loss. The control group gained <0.1kg.

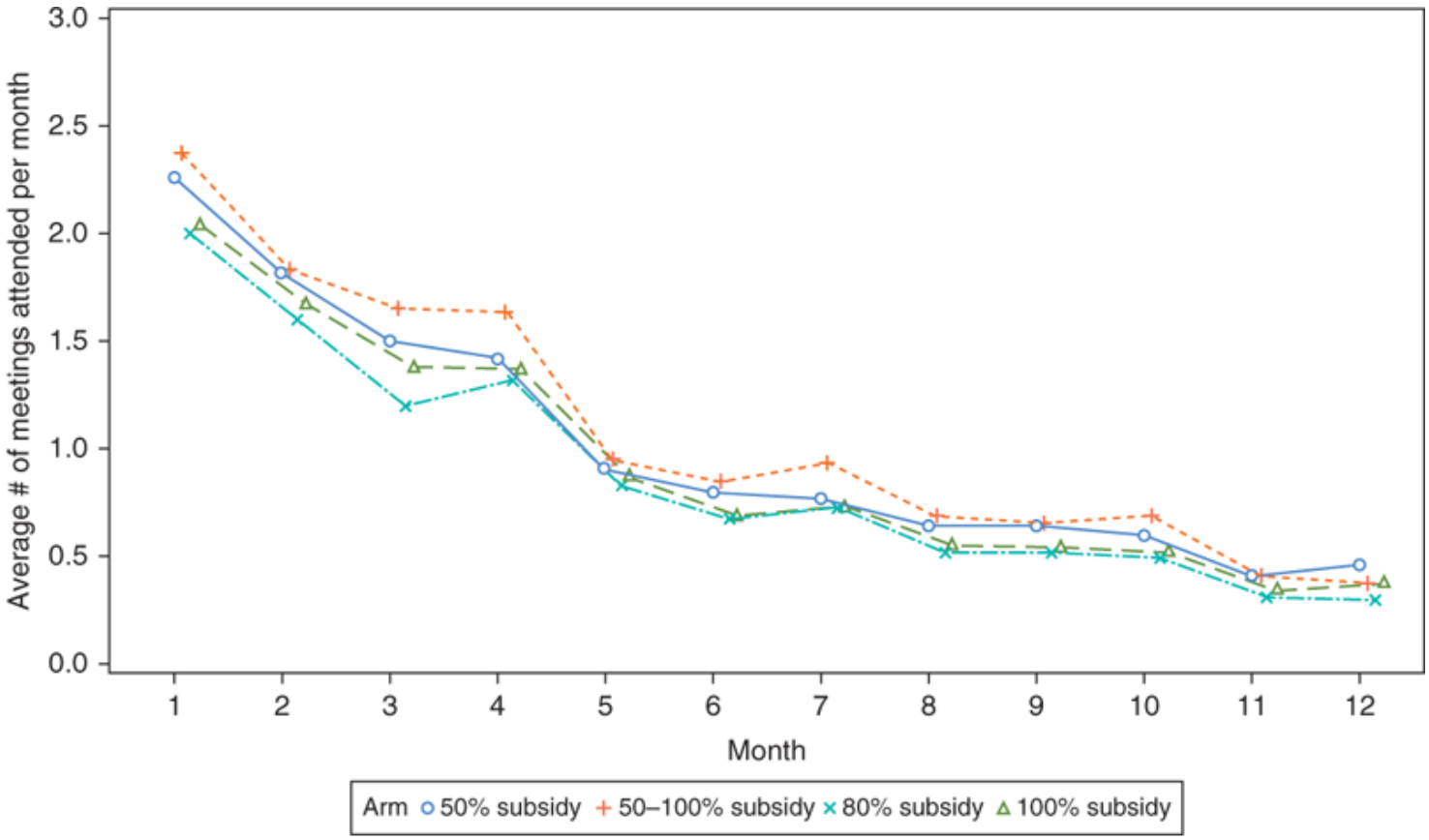
The relative ineffectiveness of premium-based financial incentives is not surprising given their design. Premium adjustments are logistically appealing because the infrastructure to do so is already in place, but the evidence for their effectiveness in on the negative side of unknown and theory argues against it. Such incentives are typically hidden in paychecks that are directly deposited in bank accounts and may go unnoticed by the individual. While a $550 incentive seems like a large amount, it is only $20 in each biweekly paycheck. Typically, these incentives are administered on an all-or-nothing basis contingent on meeting a specific threshold such as a body mass index of ≤25 kg/m^2, meaning that those who are close to that target may be motivated by a goal within reach, but those who are farther away (and have the most weight to lose and the most health to gain) may be less motivated (perhaps even demotivated) by a goal that doesn’t seem attainable. The best evidence and theory now suggests that the standard approach of using premium-based incentives is not that effective and that employers should consider alternative delivery channels for financial incentives to promote health.

A common problem employers and health plans struggle with is getting high levels of engagement among employees/health plan members. Weight Watchers, the largest commercial weight loss program in the United States, worked with a team of academic investigators to conduct a randomized controlled trial involving more than 23,000 participants testing the impact of offering employer subsidization of Weight Watchers membership fees of 50%, 80%,

100% or 50% that could turn into 100% conditional on attending at least 3 Weight Watchers classes a month. Higher subsidies led to higher program enrollment (p<.0001). Enrollment differed significantly by subsidy level (p<.0001). The 100% subsidy produced the highest enrollment (7.7%), significantly higher than each of the lower subsidies (vs 80% subsidy: 6.2%, p = .002; vs 50% subsidy: 3.9%, p<.0001; vs hybrid: 3.7%, p<.0001). Enrollment in the 80% subsidy group was significantly higher than both lower subsidy groups (vs 50% subsidy: 3.9%, p<.0001; vs hybrid: 3.7%, p<.0001). Among enrollees, there were no differences among the four groups in attendance or weight loss. In all groups overall weight loss was modest, with a mean weight loss of 2.6 pounds (95% confidence interval [CI], 5.7-lb loss to 0.3-lb gain) in the 100% subsidy arm; 1.5 pounds (95% CI, 5.6-lb loss to 1.3-lb gain) in the 80% subsidy arm; 3.8 pounds (95% CI, 7.9-lb loss to 0.4-lb gain) in the 50% subsidy arm, and 4.0 pounds (95% CI, 8.1-lb loss to 0.1-lb gain) in the hybrid subsidy arm. In all arms, attendance rates drop steadily over time, suggesting that while the hypothesized tradeoff between higher subsidies and lower ongoing engagement in the program did not exist, ongoing participation is a challenge across the board and subsidies that lower the cost of participating are insufficient to achieve high levels of ongoing engagement **(Fig. 468-3)**.

Figure 468-3

Attendance rates at Weight Watchers over time in different subsidy groups. *(From LK John et al: "The effect of cost sharing on an employee weight loss program": A randomized trial. Am J Health Promot 32:170, 2018.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

In another study, 281 overweight and obese adults were randomly assigned to a control group or one of three incentive groups for a 13-week physical activity program. All participants were given a goal of achieving 7000 steps a day, tracked automatically using their smart phones. Participants in each of the three incentive arms were offered the same magnitude incentive, $1.40 per day, and were told that accumulated earnings would be sent via a check at the end of each month. However, the incentive in each group was framed differently. In the standard gain incentive group,

participants were told that they could earn $1.40 each day they achieved the goal. In the regret lottery incentive group, participants were in a daily regret lottery in which they had an 18 in 100 chance of winning $5 and a one in 100 chance of winning $50, which averages to $1.40 per day. In the loss framing incentive group, participants were told at the beginning of each month that $42 had been placed in a virtual account and that they would lose $1.40 each day the goal was not achieved. During the 13-week intervention, participants in the control, standard gain, and regret gain arms achieved their daily step goal about 30%, 35% and 36% of the time, respectively, but those gain-framed incentive arms were not significantly different from control. However, in the loss framing group, participants achieved the goal 45% of the time, a 50% relative increase, which was significantly greater than the control arm ($p = .001$). This study demonstrated how loss framing can be used to motivate behavior change. It is also one of the first studies to create a loss frame without requiring participants to put their own money at risk using a deposit contract. This is important because fewer people are willing to engage in deposit contract-based incentives than reward-based incentives.

Experience with behavioral economics and weight loss provides several general lessons that are transferable to other health applications. While early studies largely emphasized research in the size of financial incentives, later studies have revealed that the design of the incentive strategy is at least as important. Moreover, designs can vary considerably based on the timing of incentives that can be immediate or delayed, or frequent or one time; the setting of targets that can be achievable, aspirational, or demotivating; the certainty of incentives that can be fixed or probabilistic; the channel for incentives that can either be delivered separately or bundled through payrolls; or the framing of incentives as gains or losses. Classical economists would see only the size of the incentive as an available lever for motivation, but behavioral economists face a much larger set of considerations in designing and testing effective therapies.

Medication Adherence

Numerous studies have shown that at least a third of patients fail to adhere to medication regimens. One approach to improving medication adherence is to change some of the underlying defaults by, for example, using 90-day prescriptions for chronic illness medications as opposed to 30-day prescriptions. While we are unaware of empirical evidence supporting longer prescription cycles, it seems logical that adherence rates would be higher over a year if people had to get refills three times as opposed to 11 times—the latter provides more opportunities to forget, experience delays, or fall off the wagon. Automatic refills through prescription mail order might similarly prevent some people from inadvertently falling off the wagon of medication adherence, giving them one less thing to think about, or fail to think about.

Ninety-day prescriptions or automatic refills could be set up as the default and patients or their providers could opt out if desired. Of course, opt-out defaults are not always possible. A large pharmacy benefits manager wanted to encourage automatic refills for patients on long-term medication but could not have members opt out of such a system because of the potential that those who missed the implications of the opt-out would be surprised or angry about finding credit card charges for automatic refills. In some non-health settings, marketers have used a process sometimes called “active choice” to force explicit consideration of alternatives. Often when ordering airline tickets online the sale cannot proceed unless you affirmatively accept or reject the offer to buy travel insurance. In those settings, the requirement is often seen as a nuisance, particularly if the offer is perceived as a way to cross sell an otherwise unwanted insurance product.

But the same principles can be applied in more health promoting and prosocial settings. Switching from an opt-in system to embedding a choice (yes or no) within the prescription refill process and highlighting the advantages in terms of convenience (“we can send your refills to you automatically or you can get your refills manually if you prefer”) resulted in more than twice as many patients choosing to be in the automatic-refill program.

A feature of many health insurance plans is that they require patients to pay some costs out-of-pocket, and hence discourage the use of a number of high-value elements of care, such as the treatment of hypertension or the use of statins by patients with diabetes—care that is widely seen as worth its cost. Support for the use of health insurance deductibles that require patients to have “skin in the game” for their health expenditures derives from insurance theory as well as the seminal RAND health insurance experiment, which demonstrated that these deductibles help overcome moral hazard and reduce the consumption of health care services. Copayments, deductibles, and other out-of-pocket costs make consumers more cost conscious and so aim to make them better shoppers for health care services. Indeed, the rise of high deductible health plans largely aims to increase patients' skin in the game, to make them more value-conscious shoppers.

However, while deductibles and copayments make sense as a way to reduce overutilization of some lower-value health care services, deductibles and copayments make considerably less sense when patients receive medications to manage their hypertension, diabetes, or hyperlipidemia. Given that deductibles and copayments are designed to reduce utilization, why would we ever want to apply them to antihypertensives or statins or insulin given the high health value of these drugs? Why put any barriers between patients and these drugs? Indeed, as the RAND health insurance study showed high-deductible health plans are as likely to discourage the use of high-value services as the use of low-value services. Because patients lack knowledge about what tests or services are of high or low value, and do not have information about the relationship between price and quality, such plans discourage spending on *all* tests and services, including those of high value.

Value-based insurance design—which involves discounting, or making free, services that are deemed to be high in value—is an attempt to sharpen the blunt incentives inherent in deductibles and copayments. Value-based insurance design was inspired by research that showed the use of higher copayments significantly reduced the use of services such as prescriptions but ultimately raised costs, because lower rates of medication nonadherence led to higher rates of emergency department visits and adverse outcomes. Extrapolating from these results, it was natural to conclude that lowering cost sharing for high-value activities, such as taking medications for chronic conditions, would increase adherence and potentially thereby reduce costs. The Affordable Care Act incorporates a kind of value-based insurance design in its requirement that preventive services be offered to patients at no charge.

Unfortunately, value-based insurance design has not delivered on the hope that it would both save money and improve health. From the perspective of the purchaser (for example, the employer or insurer) the economic impact of value-based insurance design depends on whether it can make enough people adherent who were previously nonadherent—and on the health and cost consequences of that improved adherence—to offset the loss of the copayments from those who were already adherent. Although some experimental tests of value-based insurance design have found that copayment reductions increase adherence, those effects have typically been small, in the range of 3–6 percentage points. Even among patients who had recent heart attacks and were given their cardiovascular medications for free, average adherence was only about 45%, just a few percentage points higher than that seen with regular copayments. One reason for these disappointing results is the “dog that didn’t bark” problem. People who are nonadherent don’t notice that their copays have been reduced because they aren’t using (and thus aren’t paying for) the service.

Indeed, one of the valuable lessons learned from efforts to introduce value-based insurance design has been a reminder of the asymmetry of the forces that surround patient engagement. Based on conventional economic thought, it might seem reasonable to assume that decreasing copayments would create effects equal and opposite to those of increasing copayments. However, behavioral economic research reveals that framing matters and that losses (in this case, higher copayments intended to reduce use) loom larger in patients' minds than gains (lowered copayments). Furthermore, people who would be deterred by higher copayments are different from people who might become adherent with lower

copayments, because the first group consists of those who take their medications while the second group consists of those who do not. Behavioral economic thinking, therefore, helps to explain what has been observed: Increases in copayments have larger effects in reducing adherence than decreases in copayments have in raising adherence. In general, copayment increases lead to far smaller decreases in medication adherence than copayment decreases lead to increases in medication adherence.

Value-based insurance design is an appealing idea. But its benefits could be increased through the application of ideas from behavioral economics, such as simple changes in reward delivery to increase salience (e.g., retaining the copay, but sending a rebate) and communications from insurers to patients so that even those who are nonadherent are aware of the benefit. Better designs might also reflect that most medication adherence happens at least daily, and so reinforcements to that behavior probably need to occur more frequently than the 30- or 90-day cycles coinciding with prescription refills.

A series of studies have used daily lottery-based financial incentives to improve medication adherence. Early work tested the impact of a lottery on medication adherence among patients on warfarin. Participants were eligible for the lottery daily if they correctly took their warfarin the day before. In the first study where the lottery had an expected value (EV) of $5 per day, the proportion of incorrect pill taking was 2.3% (97.7% adherence), compared with a historic mean of 22% incorrect pill taking in this clinic population. In the second study (EV of $3 per day), the overall mean adherence was 98.4% (1.6% days nonadherent), similar to the $5/day study. In a two-arm randomized trial of lotteries for warfarin adherence, the *a priori* subgroup with baseline international normalized ratios (INRs) below the therapeutic range showed no change relative control, but in the a priori subgroup with out-of-range INRs there was a significant reduction in out-of-range INR in the lottery arm vs the control arm (adjusted odds ratio, 0.39; 95% CI, 0.25–0.62). This study highlighted the importance of targeting nonadherent patients in terms of interventions and provided evidence from a randomized, controlled trial (RCT) that lottery incentives can be effective in achieving improved clinical outcomes.

A four-arm NHLBI-funded RCT tested the impact of daily lotteries and daily reminders in a 2 × 2 factorial design on warfarin adherence to address the question of the degree to which a daily lottery is effective due to the fact it also constitutes a daily reminder. However, this study found that while participants in the reminder group had the lowest percentage of time out of target INR range, with an adjusted odds of an out-of-range INR 36% lower than among those in the control group (95% CI, 7–55), the only group with significant improvement in incorrect adherence was the lottery group (incorrect adherence: 12.1% compared with 23.7% in the control group; difference of 7.4%; 95% CI, –14 to –0.3). There was no relationship between changes in adherence and anticoagulation control in the lottery group, highlighting that participants may appear to change their behavior without perhaps taking the medication, highlighting the importance of serologic or biometric confirmation when possible.

The 5000 Hours Problem

A major challenge is determining the optimal method to reach patients and reinforce their behavior each day if we want to significantly improve medication adherence. Even patients with chronic illnesses may spend only a few hours a year with a doctor or nurse, but they spend about 5000 waking hours a year doing just about everything else. Those 5000 hours are when they live their lives and make choices about what to eat and whether to exercise, smoke, take their medications, or visit the doctor.

Although what people do in those hours almost certainly affects their health outcomes, the hours are typically ignored by the U.S. health care system, in part because current approaches to U.S. health care financing support health care

during visits to the doctor, not between them, and because “hovering over” people during the hours between visits is personnel-intensive, often requiring nurses or other clinicians to call or visit patients or to staff telemedicine programs. Hovering also requires a fair amount of the very kind of engagement in their own health and health care that is so often missing in the patients these interventions aim to reach. As a result, many of the most promising efforts in telemedicine and home health care have been disappointing.

If some form of hovering is required to engage people who are otherwise hard to engage during the 5000 hours, it almost certainly has to become substantially more automated—both because providers must reduce the need for expensive personnel and because patients have in many cases already revealed limits to their willingness to exert themselves to improve their health. Nevertheless, there is reason for optimism based on the increasing use of cell phones and other wireless devices that make it technologically easier to embed reminders and other forms of touch into patients' lives. Indeed, one key lesson from behavioral economics is that rather than trying to change people’s behavior patterns to promote health, it is better to restructure their environment and circumstances so that their existing behavior patterns are more likely to lead to better outcomes. Those efforts require a substantial amount of hovering over patients. Cell phones and other wireless devices don’t necessarily change behavior on their own, but because they are already part of many patients' everyday lives they allow previously private behaviors to be witnessed and at times acted upon.

Indeed, an error in early approaches to technology and health behavior overgeneralized from the technologies that support the “quantified self” movement. Apps and wearables that track your diet, physical activity, and biometrics were largely designed for people passionate about measuring themselves. Such individuals don’t need a lot of encouragement to wear devices or enter data. The same approaches are far less likely to be useful for patients with difficult-to-manage chronic illness. Many of the internal and external challenges that make their chronic illness hard to manage also make such monitoring hard to manage. A patient who is nonadherent to medication is likely also to be nonadherent to using a new electronic device, but devices like cell phones that are already in use, or other devices that require much less active involvement (like wireless pill bottles) offer more conceptual appeal.

Patients with poorly controlled diseases typically exhibit multiple risk behaviors: poorly controlled diabetes, for example, can be exacerbated by poor diet, lack of exercise, obesity, and medication nonadherence. In testing daily use rates of wireless glucometers and blood pressure cuffs within a population of patients with poorly controlled blood sugars within a health system population, patients who were randomized to be asked to use their wireless glucometer or blood pressure cuff daily used these devices only 50% of the time by the end of 3 months, whereas patients randomized to receive daily lotteries conditional on device use used their devices more than 80% of the time and achieved better glycemic control.

FUTURE PERSPECTIVES

Human health derives from the interaction of basic biologic processes, environmental exposures, social structures, and behavior. The field of behavioral economics has greatly contributed to our understanding of behavior and has made significant contributions to the science of public policy. Given that understanding, we have the opportunity to replace older policies based on unrealistic normative models of rational decision-making with newer policies reflecting our most up-to-date understanding of how humans actually make decisions. Individual and population health outcomes would be very different if people were able to weigh the present and future costs of their actions carefully and dispassionately and had the necessary information and self-control to implement behavioral plans and overcome decision errors that contribute to unhealthy behaviors. Because few people can meet any of those challenges, let alone

all, we should not structure our behavior change interventions and public health policies around such models of behavior.

There is broad potential for using understanding of human motivation rooted in behavioral economics to improve private and public approaches to health behavior. One major question is whether developed economies will continue to invest the majority of their health care dollars in treatment (typically about 97% of health care dollars) rather than shifting it toward innovations that seek to keep people healthy. This will be particularly important in settings where treatment options are limited and highly effective methods of prevention exist. For example, it has been estimated that a combination of low-cost cardiovascular drugs could reduce cardiovascular events by 62–88% with perfect adherence, revealing that reducing atherosclerotic cardiovascular disease risk is largely a *behavioral* challenge, given that adherence to medications remain low despite effective pharmacologic solutions. Shifts in health care financing away from fee-for-service toward various forms of payment that require health delivery systems to take on financial risk for populations of patients may drive greater interest in addressing these behavioral and social determinants of health. Research expenditures, which in many developed countries are also roughly allocated about 97% to new treatments and about 3% to prevention, similarly could be shifted to focus more on testing of innovative approaches to improve population health.

The same errors that misdirect patients and providers also misdirect policy makers. In part because of present bias, preventive services often are covered by insurance only if they show a positive return on investment, and yet treatments of existing disease are not held to the same standard. An employer wondering whether to introduce a smoking cessation program for employees wonders about the return on that investment in terms of reduced illness and its cost. The same employer might never question the return on investment of treating lung cancer in the same employee pool, despite what would almost certainly be a negative return on investment. These asymmetries are so embedded in policy making as to be nearly invisible or at least unchallengeable. In fact, the cost of treatments is not even allowed to be considered in Medicare coverage decisions. This prohibition naturally leads to overinvestment in treatments of low value and underinvestment in prevention. The same standard for assessing the impact of health programs, with the goal of achieving the most health possible with the available resources, should be used for both preventive and therapeutic services.

Despite the promise of behavioral economics in structuring policy solutions to social goals, plenty of existing policies that have nothing to do with behavioral economics are effective. For example, raising taxes on cigarettes and other unhealthy goods where it is in the public interest to consume less is a powerful policy tool derived from classical economics. Indeed, tobacco taxes represent one of the most effective ways to curb the use of tobacco and its initiation among youth. Behavioral economics can help make such policies more effective but should not be seen as a substitute for them.

For private-sector entities, the implications of choosing defaults wisely are recognized by many organizations that aim to shift the “path of least resistance” towards healthier choices. Setting up defaults in benefit program design to favor health plans that provide better coverage of preventive services, changing the environment in workplaces to make it easier to take the stairs, and serving more healthful food in cafeterias represent approaches to gently lead people toward individual and population goals.

While medical research continues to generate new tests, interventions, and drugs successfully targeting conditions recently seen as intractable, even the most effective drugs will not work if physicians fail to prescribe them and if patients fail to take them. Although the dominant forms of investigation in medicine seek cellular or molecular therapeutic targets to modify disease, behavioral sciences have revealed cognitive pathways that operate nearly as

predictably as the genetic code. The opportunity for behavioral economics to improve health and health care delivery derives from its recognition of these behavioral pathways and the growing empirical evidence about how to best make use of them.

FURTHER READING

Asch DA et al: Automated hovering in health care—watching over the 5000 hours. N Engl J Med 367:1, 2012. [PubMed: 22716935]

Asch DA et al: Asymmetric thinking about return on investment. N Engl J Med 374:606, 2016. [PubMed: 26886519]

Connolly T, Butler DU: Regret in economic and psychological theories of choice. J Behav Decis Mak 19:148, 2006.

John LK et al: The effect of cost sharing on an employee weight loss program: A randomized trial. Am J Health Promot 32:170, 2018.

Kahneman D: *Thinking, Fast and Slow*. New York, Farrar, Straus and Giroux, 2011.

Kahneman D, Tversky A: Prospect theory: An analysis of decision under risk. Econometrica 47:263, 1979.

Loewenstein G et al: Asymmetric paternalism to improve health behaviors. JAMA 298:2415, 2007. [PubMed: 18042920]

Loewenstein G et al: Behavioral economics holds potential to deliver better results for patients, insurers, and employers. Health Aff (Millwood) 32:1244, 2013. [PubMed: 23836740]

Patel MS et al: Generic medication prescription rates after health system-wide redesign of default options within the electronic health record. JAMA Intern Med 176:847, 2016. [PubMed: 27159011]

Thaler RH, Sunstein CR: *Nudge: Improving Decisions About Health, Wealth and Happiness.* New York, Penguin Books, 2008.

Volpp KG et al: Financial incentive-based approaches for weight loss: A randomized trial. JAMA 300:2631, 2008. [PubMed: 19066383]

Volpp KG et al: Assessing value in health care programs. JAMA 307:2153, 2012. [PubMed: 22618919]

Chapter 469: Complementary, Alternative, and Integrative Health Approaches

Josephine P. Briggs

INTRODUCTION

The search for health includes many beliefs and practices that are outside conventional medicine. Physicians are important sources for information and guidance about health matters, but our patients also rely on a wide range of other sources including family and friends, cultural traditions, alternative practitioners, and increasingly the Internet, popular media, and advertising. An important step in patient-centered care is understanding what patients are doing to manage their health. This understanding is important to harness potential benefits and to help patients avoid harm.

DEFINITIONS AND SCOPE

Complementary health approaches include a broad range of practices, interventions, and natural products, which are not typically part of conventional medical care, or which may have origins outside of usual Western practice. *Complementary* approaches are defined as those used together with conventional therapies, distinguishing them from *alternative* practices, those used as a substitute for standard care. Complementary practices can roughly be divided into two major groups—mind and body practices, and natural products. Mind and body practices and disciplines are usually administered by or taught to others by a clinician, trained practitioner, or teacher, and include acupuncture, massage, meditation, and hypnosis. Natural products include a diverse group of orally or topically administered substances such as botanical products, unconventional diets, dietary supplements, herbal medicines, homeopathic remedies, probiotics, and others. Brief definitions for some of the common complementary and alternative health practices are provided in Table 469-1. Although some complementary health practices are recommended or provided by a physician or a complementary health care provider such as a chiropractor, acupuncturist, or naturopathic practitioner, many of these practices are undertaken as “self-care.” Although some are reimbursed, most are paid for out of pocket.

TABLE 469-1

Terminology of Complementary Health Approaches

Mind and Body Practices	
Acupuncture and acupressure	A family of procedures involving stimulation of defined anatomic points, a component of the major Asian medical traditions; most common application involves the insertion and manipulation of thin metallic needles
Alexander technique	A movement therapy that uses guidance and education to improve posture, movement, and efficient use of muscles for improvement of overall body functioning
Guided imagery	The use of relaxation techniques followed by the visualization of images, usually calm and peaceful in nature, to invoke specific images to alter neurologic function or physiologic states
Hypnosis	The induction of an altered state of consciousness characterized by increased responsiveness to suggestion
Massage	Manual therapies that manipulate muscle and connective tissues to promote muscle relaxation, healing, and sense of well-being
Meditation	A group of practices, largely based in Eastern spiritual traditions, intended to focus or control attention and obtain greater awareness of the present moment, or mindfulness
Reflexology	Manual stimulation of points on hands or feet that are believed to affect organ function
Rolfing/structural integration	A manual therapy that attempts to realign the body by deep tissue manipulation of fascia
Spinal manipulation	A range of manual techniques, employed by chiropractors and osteopaths, for adjustments of the spine to affect neuromuscular function and other health outcomes
Tai chi	A mind and body practice originating in China that involves slow, gentle movements and sometimes is described as “moving meditation”
Therapeutic touch	Secular version of the laying on of hands, described as “healing meditation”
Yoga	An exercise practice, originally East Indian, that combines breathing exercises, physical postures, and meditation

Traditional Medical Systems	
Ayurvedic medicine	The major East Indian traditional medicine system; treatment includes meditation, diet, exercise, herbs, and elimination regimens (using emetics and diarrheals)
Curanderismo	A spiritual healing tradition common in Latin American communities that uses ritual cleansing, herbs, and incantations
Native American medicine	Diverse traditional systems that incorporate chanting, shaman healing ceremonies, herbs, laying on of hands, and smudging (ritual cleansing with smoke from sacred plants)
Siddha medicine	An East Indian medical system (prevalent among Tamil-speaking people)
Tibetan medicine	A medical system that uses diagnosis by pulse and urine examination; therapies include herbs, diet, and massage
Traditional Chinese medicine	A medical system that uses acupuncture, herbal mixtures, massage, exercise, and diet
Unani medicine	An East Indian medical system, derived from Persian medicine, practiced primarily in the Muslim community; also called “hikmat”
“Modern” Medical Systems	
Anthroposophic medicine	A spiritually based system of medicine that incorporates herbs, homeopathy, diet, and a movement therapy called eurythmy
Chiropractic	Chiropractic care involves the adjustment of the spine and joints to alleviate pain and improve general health; primarily used to treat back problems, musculoskeletal complaints, and headaches
Homeopathy	A medical system with origins in Germany that is based on a core belief in the theory of “like cures like”—compounds that produce certain syndromes, if administered in very diluted solutions, will be curative
Naturopathy	A clinical discipline that emphasizes a holistic approach to the patient, herbal medications, diet, and exercise; practitioners have degrees as doctors of naturopathy
Osteopathy	A clinical discipline, now incorporated into mainstream medicine, that historically emphasized spinal manipulative techniques to relieve pain, restore function, and promote overall health

In the last decade or so, the terms *integrative health care* and *integrative medicin*e have entered the dialogue. The term *integrative health* care emphasizes a holistic, patient-focused approach to health care and wellness. Most integrative health care is team-based, often bringing conventional and complementary approaches together with self-care in a coordinated way. Physicians advocating this approach generally included selected complementary health practices in the care they offer patients, and many have established practice settings that include complementary health practitioners. Integration of select complementary approaches is common in Veterans Administration and Department of Defense facilities, particularly as part of management of pain and post-traumatic stress disorder. A 2007 national survey conducted by the Centers for Disease Control and Prevention's National Center for Health Statistics found that 42% of hospices had integrated complementary health practices into the care they provide. Although the integrative approach appears to be attractive to many patients, the heavy use of dietary supplements and the weaknesses in the evidence base for a number of the interventions offered in integrative practices continue to attract substantial concern and controversy.

Until a decade ago or so, *complementary and alternative medicine* could be defined as practices that are neither taught in medical schools nor reimbursed, but this definition is no longer workable, since medical students increasingly seek and receive some instruction about complementary health practices, and some practices are reimbursed by third-party payers.

By its nature, the demarcation between mainstream medicine and complementary health practices is porous, varying from culture to culture and over time. Traditional Chinese medicine and the Indian practice of Ayurvedic medicine were once the dominant health teachings in those cultures. Certain health practices that arose as challenges to the mainstream have been integrated gradually into conventional care. Examples include the teachings of Fernand Lamaze that led to the widespread use of relaxation techniques during childbirth, the promotion of lactation counseling by the La Leche League, and the teaching of Cicely Saunders and Elizabeth Kübler-Ross that established the hospice movement.

The late nineteenth century saw the development of a number of healing philosophies by care providers who were critical of the medicine of the time. Of these, naturopathy and homeopathy, which arose in Germany, and chiropractic and osteopathy, which developed in the United States, have endured. Osteopathic medicine is currently thoroughly integrated into conventional medicine, although the American Medical Association (AMA) labeled it a cult as late as 1960. The other three traditions have remained largely separate from mainstream medicine, although chiropractic care is increasingly available in some conventional care settings.

PATTERNS OF USE

The first large survey of use of these practices was performed by David Eisenberg and associates in 1993. It surprised the medical community by showing that >30% of Americans use complementary approaches. Many studies since that time have extended those conclusions. The National Health Interview Survey (NHIS),

a large, national household survey of health practices conducted by the National Center for Health Statistics, a component of the Centers for Disease Control and Prevention, has addressed the use of complementary health practices in 2002, 2007, and 2012. The NHIS survey uses methods that create a nationally representative sample and has a sample size large enough to permit valid estimates about some subgroups. Information was obtained from 31,000 adults in 2002; 23,300 adults and 9400 children in 2007; and 34,500 adults and 10,200 children in 2012. In all three surveys, approximately one-third of adults report using some form of complementary therapy or health practice. In the 2012 survey, 32.2% of adults and 11.6% of children had used one or more modalities. These surveys yield the estimate that nonvitamin, nonmineral dietary supplements are used by ~18% of adults and 5% of children. The most prevalent mind and body practices are yoga, chiropractic or osteopathic manipulation, meditation, and therapeutic massage. Americans are willing to pay for these services; the estimated out-of-pocket expenditure for complementary health practices in 2012 was $30.2 billion ($28.3 billion for adults and $1.9 billion for children), representing 1.1% of total health expenditures and 9.2% of out-of-pocket costs.

The appeal of unproven complementary health approaches continues to perplex many physicians. Many factors contribute to these choices. In NHIS surveys, the most common reason cited for use of complementary health approaches is the management of pain. In one analysis of data from the 2012 NHIS survey, >40% of people with a musculoskeletal pain disorder used a complementary health approach. This was significantly higher than use by people without a musculoskeletal pain disorder (24%). Some patients seek out complementary health practitioners because they offer optimism or greater personal attention. For others, alternative approaches reflect a “self-help” approach to health and wellness or satisfy a search for “natural” or less invasive alternatives, since dietary supplements labelled as natural are believed, often incorrectly, to be inherently healthy.

PRACTITIONER-BASED DISCIPLINES

Licensure and Accreditation

At present, six fields of complementary health practice—osteopathic manipulation, chiropractic, acupuncture and traditional Chinese medicine, therapeutic massage, naturopathy, and homeopathy—are subject to some form of educational accreditation and state licensure. Accreditation of educational programs is the responsibility of professional organizations or commissions under federal oversight by the Department of Education. Licensure, in contrast, is strictly a state matter, generally determined by state legislatures. Legal recognition establishes public access to therapies even when there is no scientific consensus about their clinical value.

Osteopathic Manipulative Therapy

Founded in 1892 by the physician Andrew Taylor Still, osteopathic medicine was originally based on the belief that manipulation of soft tissue and bone can correct a wide range of diseases of the musculoskeletal and other organ systems. Over the ensuing century, the osteopathic profession has welcomed increasing

integration with conventional medicine. Today, the postgraduate training, practice, credentialing, and licensure of osteopathic physicians are virtually indistinguishable from those of allopathic physicians. Osteopathic medical schools, however, include training in manual therapies, particularly spinal manipulation.

Chiropractic

The practice of chiropractic care, founded by David Palmer in 1895, is the most widespread practitioner-based complementary health practice in the United States. Chiropractic practice emphasizes manual therapies for treatment of musculoskeletal complaints, although the scope of practice varies widely, and in some rural areas, chiropractors may serve a primary care role, due in part to the lack of other providers. According to the NHIS, ~8% of American adults and 3% of children receive chiropractic manipulation in a given year.

Since the mid-1970s, chiropractors have been licensed in all 50 states and reimbursed by Medicare. Chiropractic educational standards mandate 2 years of undergraduate training, 4 years of training at an accredited school of chiropractic, and in most states, successful completion of a standardized board examination. Postgraduate training is not required. The U.S. Department of Labor estimates that there are 45,000 licensed chiropractors (2015 figure). There is substantial geographic variation, with greater numbers of practitioners and greater use in the midwest, particularly in rural areas, and lower use in the southeast.

Historically, the relationship between the medical and chiropractic professions has been strained. Extending through the 1970s, the AMA set forth standards prohibiting physicians consulting or entering into professional relationships with chiropractors, but in 1987, after a decade of complex litigation, the U.S. District Court found the AMA in violation of antitrust laws. An uneasy truce has followed, with continued physician skepticism, but also evidence for robust patient demand and satisfaction.

The role of both osteopathic and chiropractic spinal manipulative therapies (SMTs) in management of low-back pain has been the subject of a number of carefully performed trials and many systematic reviews. Conclusions are not consistent, but the most recent 2017 guidelines from the American College of Physicians on noninvasive treatments for acute, subacute, and chronic low-back pain conclude that spinal manipulation has a small effect on improving function and pain compared with control—either a sham manipulation or an inert treatment. Although evidence for spinal manipulation for chronic low-back pain is graded as *low-quality*, the recommendation for consideration of nonpharmacologic treatment including spinal manipulation is graded as a *strong recommendation*, reflecting increasing concern with the impact of chronic opioid use for low-back pain.

The evidence of benefit of spinal manipulation for neck pain is not as extensive, and continued concern that cervical manipulation may occasionally precipitate vascular injury clouds a contentious debate.

Naturopathy

Naturopathy is a discipline that emerged in central Europe in the nineteenth century as part of the Natural Cure movement and was introduced to the United States in the early twentieth century by Benjamin Lust. Nineteen states and the District of Columbia currently license naturopathic physicians, with considerable variation in the scope of practice. The naturopathic profession is actively seeking licensure in other states. There are estimated to be ~5000 licensed naturopathic physicians in the United States. There is also a robust naturopathy presence in Canada. Conventional and unconventional diagnostic tests and medications are prescribed, with an emphasis on relatively low doses of drugs, herbal medicines, healthy diet, and exercise. While there is some support for success of naturopathic practitioners in motivating healthy behaviors, concern exists about the heavy promotion of dietary supplements, most with little rigorous evidence.

Homeopathy

Homeopathy was widespread in the United States in the late nineteenth and early twentieth centuries and continues to be a common alternative practice in many European countries, but estimates from the NHIS suggest that <2.2% of Americans visit a homeopathic practitioner in any given year. In the United States, licensure as a homeopathic physician is only possible in three states (Arizona, Connecticut, and Nevada) where it is restricted to licensed physicians. The number of practitioners is uncertain, however, because some states include homeopathy within the scope of practice of other fields, including chiropractic and naturopathy, and some practitioners may self-identify as homeopathic practitioners. As discussed below, the regulatory framework for homeopathic remedies differs from other dietary supplements. Homeopathic remedies are widely available and commonly recommended by naturopathic physicians, chiropractors, and other licensed and unlicensed practitioners.

Therapeutic Massage

The field of therapeutic massage is growing rapidly, as use by the public is increasing. According to U.S. Department of Labor statistics, there are ~168,000 licensed massage therapists employed in the United States, and by 2024, this number is projected to grow by 22%. Forty-five states and the District of Columbia currently have laws regulating massage therapy; however, there is little consistency, and in some states, regulation is by town ordinance. States that do provide licensure for massage therapists typically require a minimum of 500 h of training at an accredited institution, as well as meeting specific continuing education requirements and carrying malpractice insurance. Massage training programs generally are approved by a state board, but some may also be accredited by an independent agency, such as the Commission on Massage Therapy Accreditation (COMTA). The development of regulatory standards for therapeutic massage has not yet caught up with the evolution of the field or the high demand. Many techniques used are also employed by physical therapists.

Acupuncture and Traditional Chinese Medicine

A venerable component of traditional Chinese medicine, with a history of use that extends at least 2000 years, acupuncture became better known in the United States in 1971, when *New York Times* reporter James Reston wrote about how doctors in China used needles to ease his pain after surgery. More than 3 million

adults in the United States use acupuncture, according to NHIS data. In a number of European countries, acupuncture is performed primarily by physicians. In the United States, the training and licensure processes for physicians and nonphysicians differ. Currently, acupuncture is licensed in 47 states and the District of Columbia, with licensure standards varying within the scope of practice of each state. Licensure for nonphysicians generally requires 3 years of accredited training and the successful completion of a standardized examination. The main accrediting organization is the Accreditation Commission for Acupuncture and Oriental Medicine. Acupuncture is included in doctor of medicine (MD) and doctor of osteopathic medicine (DO) licensure in 35 states, with 15 states requiring additional training for physicians performing acupuncture.

MIND AND BODY INTERVENTIONS

Mind and body practices and disciplines consist of physical procedures or exercises, manual therapies, or mental techniques that are administered or taught to others by a clinician, trained practitioner, or teacher. Examples include acupuncture, massage therapy, meditation, relaxation techniques, spinal manipulation, and yoga. These approaches are being used more frequently in mainstream health care facilities for both patients and health care providers. Mind and body practices such as meditation and yoga are not licensed in any state, and training in those practices is not subject to national accreditation.

Painful conditions are the most common reasons why American adults use complementary health approaches. About 40 million American adults experience severe pain in any given year, and they spend >$14 billion out-of-pocket on complementary approaches to manage such painful conditions as back pain, neck pain, and arthritis. Chronic pain management is often refractory to conventional medical approaches, and standard pharmacologic approaches have substantial drawbacks. Health care guidelines of the American Pain Society, the American College of Physicians, and other professional organizations recognize the value of certain complementary approaches as alternatives or adjuncts to pharmacologic management.

The evidence base for the effectiveness of these modalities is still relatively incomplete, but a few rigorous examples where there is promise of usefulness and safety include acupuncture and tai chi for osteoarthritis of the knee pain; massage for neck pain; tai chi for fibromyalgia pain; relaxation techniques for headaches and migraine; and acupuncture, massage, yoga, and spinal manipulation for chronic back pain. In addition, new research is shedding light on the effects of meditation and acupuncture on central mechanisms of pain processing and perception and regulation of emotion and attention. Although many unanswered questions remain about these effects, findings are pointing to scientifically plausible mechanisms by which these modalities might yield benefit.

DIETARY SUPPLEMENTS

Regulation

The Dietary Supplements Health and Education Act (DSHEA), passed in 1994, gives authority to the U.S. Food and Drug Administration (FDA) to regulate dietary supplements, but with expectations that differ in many respects from the regulation of drugs or food additives. Purveyors of dietary supplements cannot claim that they prevent or treat any disease. They can, however, claim that they maintain “normal structure and function” of body systems. For example, a product cannot claim to treat arthritis, but it can claim to maintain “normal joint health.” Homeopathic products predate FDA drug regulations and are sold with no requirement that they be proved effective. Although homeopathic products are widely believed to be safe because they are highly dilute, one product, a nasal spray called Zicam, was withdrawn from the market when it was found to produce anosmia, probably because of a significant zinc content. In January 2017, the FDA warned consumers about homeopathic teething tablets containing belladonna that pose a serious risk to infants and children.

Regulation of advertising and marketing claims is the purview of the Federal Trade Commission (FTC). The FTC does take legal action against promoters or websites that advertise or sell dietary supplements with false or deceptive statements. Misleading marketing of homeopathic products, and indeed other complementary health products and practices, contributes to the very significant risk that individuals will use them instead of effective conventional modalities.

Inherent Toxicity

Although the public may believe that “natural” equates with “safe,” it is abundantly clear that natural products can be toxic. Misidentification of medicinal mushrooms has led to liver failure. Contamination of tryptophan supplements caused the eosinophilia-myalgia syndrome. Herbal products containing particular species of *Aristolochia* were associated with genitourinary malignancies and interstitial nephritis. In 2013, dietary supplements containing 1,3-dimethylamylamine (DMAA), often touted as a “natural” stimulant, led to cardiovascular problems, including heart attacks. Among the most controversial dietary supplements is *Ephedra sinica*, or ma huang, a product used in traditional Chinese medicine for short-term treatment of asthma and bronchial congestion. The scientific basis for these indications was revealed when ephedra was shown to contain the ephedrine alkaloids, especially ephedrine and pseudoephedrine. With the promulgation of the DSHEA regulations, supplements containing ephedra and herbs rich in caffeine sold widely in the U.S. marketplace because of their claims to promote weight loss and enhance athletic performance. Reports of severe and fatal adverse events associated with use of ephedra-containing products led to an evidence-based review of the data surrounding them, and in 2004, the FDA banned their sale in the United States.

Another major current concern with dietary supplements is adulteration with pharmacologic active compounds. Multi-ingredient products marketed for weight loss, body building, “sexual health,” and athletic performance are of particular concern. Recent FDA recalls have involved contamination with steroids, diuretics, stimulants, and phosphodiesterase type 5 inhibitors.

Herb–Drug Interactions

A number of herbal products have potential impact on the metabolism of drugs. This effect was illustrated most compellingly with the demonstration in 2000 that consumption of St. John's wort interferes with the bioavailability of the HIV protease inhibitor indinavir. Later studies showed its similar interference with metabolism of topoisomerase inhibitors such as irinotecan, with cyclosporine, and with many other drugs. The breadth of interference stems from the ability of hyperforin in St. John's wort to upregulate expression of the pregnane X receptor, a promiscuous nuclear regulatory factor that promotes the expression of many hepatic oxidative, conjugative, and efflux enzymes involved in drug and food metabolism.

Because of the large number of compounds that alter drug metabolism and the large number of agents some patients are taking, identification of all potential interactions can be a daunting task. Several useful Web resources are available as information sources **(Table 469-2)**. Clearly, attention to this problem is particularly important with drugs with a narrow therapeutic index, such as anticoagulants, antiseizure medications, antibiotics, immunosuppressants, and cancer chemotherapeutic agents.

TABLE 469-2

Resources for Dietary Supplement–Drug Interactions

Medscape
http://www.medscape.com/druginfo/druginterchecker?cid=med This website is maintained by WebMD and includes a free drug interaction checker tool that provides information on interactions between two or more drugs, herbals, and/or dietary supplements.
Natural Medicines Comprehensive Database
http://naturaldatabase.therapeuticresearch.com.ezproxy1.library.usyd.edu.au This website provides an interactive natural product–drug interaction checker tool that identifies interactions between drugs and natural products, including herbals and dietary supplements. This service is available by subscription. A PDA version is available.

Abbreviation: PDA, personal digital assistant.

PATIENT AND PROVIDER RESOURCES

Physicians regularly face difficult challenges in providing patients with advice and education about complementary practices. Of particular concern to all physicians are practices of uncertain safety and practices that raise inappropriate hopes. Cancer therapies, antiaging regimens, weight-loss programs, sexual function, and athletic performance are frequently targeted for excessive claims and irresponsible marketing. A number of Internet resources provide critical tools for patient education **(Table 469-3)**. Because many

complementary health products and practices are used as self-care and because many patients research these approaches extensively on the Internet, directing patients to responsible websites can often be very helpful.

TABLE 469-3

Internet Resources on Complementary Health Approaches

The Cochrane Collaboration Complementary Medicine Reviews
This website offers rigorous systematic reviews of mainstream and complementary health interventions using standardized methods. It includes >300 reviews of complementary health practices. Complete reviews require institutional or individual subscription, but summaries are available to the public. *http://www.cochrane.org.ezproxy1.library.usyd.edu.au/evidence*
MedlinePlus All Herbs and Supplements, A–Z List
MedlinePlus Complementary and Alternative Medicine
NLM FAQ: Dietary Supplements, Complementary, or Alternative Medicines
These National Library of Medicine (NLM) Web pages provide an A–Z database of science-based information on herbal and dietary supplements; basic facts about complementary health practices; and federal government sources on information about using natural products, dietary supplements, medicinal plants, and other complementary health modalities. *http://www.nlm.nih.gov.ezproxy1.library.usyd.edu.au/medlineplus/druginfo/herb_All.html* *http://www.cochrane.org.ezproxy1.library.usyd.edu.au/evidence* *http://www.nlm.nih.gov.ezproxy1.library.usyd.edu.au/medlineplus/dietarysupplements.html*
National Institutes of Health National Center for Complementary and Integrative Health (NCCIH)
This National Institutes of Health NCCIH website contains information for consumers and health care providers on many aspects of complementary health products and practices. Downloadable information sheets include short summaries of complementary health approaches, uses and risks of herbal therapies, and advice on wise use of dietary supplements. *http://www.nccih.nih.gov.ezproxy1.library.usyd.edu.au* Resources for Health Care Providers: *http://www.nccih.nih.gov.ezproxy1.library.usyd.edu.au/health/providers* NCCIH Clinical Digest e-Newsletter: *http://www.nccih.nih.gov.ezproxy1.library.usyd.edu.au/health/providers/digest* Continuing medical education lectures: *http://www.nccih.nih.gov.ezproxy1.library.usyd.edu.au/training/videolectures*

The scientific evidence regarding complementary therapies is fragmentary and incomplete. Nonetheless, in some areas, particularly pain management, it is increasingly possible to perform the kind of rigorous systematic reviews of complementary health approaches that are the cornerstone of evidence-based

medicine. A particularly valuable resource in this respect is the Cochrane Collaboration, which has performed >300 systematic reviews of complementary health practices. Practitioners will find this a valuable source to answer patient questions. Practice guidelines, particularly for pain management, are also available from several professional organizations. Links to these resources are provided in Table 469-3.

SUMMARY

The use of complementary, alternative, and integrative health approaches reflects an active interest in improved health. An array of unproven modalities will always be used by our patients. While some of these choices need to be actively discouraged, many are in fact innocuous and can be accommodated. Some may be genuinely helpful, particularly in the management of troublesome symptoms. An important step in patient-centered care is understanding patients' beliefs and expectations and what they are doing to manage their health, including the use of complementary health approaches, and using those insights to help guide health-seeking practices in a constructive way.

ACKNOWLEDGMENT

The late Dr. Stephen Straus contributed this chapter in prior editions, and some material from his chapter has been retained here.

FURTHER READING

Barnes PM et al: Complementary and alternative medicine use among adults and children: United States, 2007. Natl Health Stat Rep 12:1, 2008.

Eisenberg DM et al: Trends in alternative medicine use in the United States, 1990–1997: Results of a follow-up national survey. JAMA 280:1569, 1998. [PubMed: 9820257]

MacPherson H et al: Closing the evidence gap in integrative medicine. BMJ 339:b3335, 2009. [PubMed: 19723708]

Manek NJ et al: What rheumatologists in the United States think of complementary and alternative medicine: Results of a national survey. BMC Complement Altern Med 10:5, 2010. [PubMed: 20109215]

Nahin RL et al: Evidence-based evaluation of complementary health approaches for pain management in the United States. Mayo Clin Proc 91:1292, 2016. [PubMed: 27594189]

Nahin RL et al: Expenditures on complementary health approaches: United States, 2012. Natl Health Stat Rep 95:1, 2016.

Paige NM et al: Association of spinal manipulative therapy with clinical benefit and harm for acute low back pain: Systematic review and meta-analysis. JAMA 317:1451, 2017. [PubMed: 28399251]

Qaseem A et al: Noninvasive treatments for acute, subacute, and chronic low back pain: A clinical practice guideline from the American College of Physicians. Ann Intern Med 166:514, 2017. [PubMed: 28192789]

Chapter 470: Telomere Disease

Rodrigo T. Calado; Neal S. Young

DEFINITION

In telomere diseases (also called *telomeropathies* or *telomere spectrum disorders*), organ dysfunction is caused by loss of the ends of chromosomes, a process termed *accelerated telomere attrition*. Inadequate repair or insufficient protection of telomeres and their resulting erosion induces cell death, deficient cell proliferation, and chromosome instability; affected tissues show defective organ regeneration, fibrosis or replacement by fat, and a proclivity for cancer. A variety of regenerative disorders affecting especially the bone marrow, lungs, liver, and skin share telomere dysfunction and loss as their common molecular mechanism. It is important to note that telomeres appear to shorten with time based on cross-sectional data of average telomere length in groups of people at different ages. However, limited data exist about telomere shortening longitudinally in individual people. Despite shortening of telomeres over time, normal aging is not associated with the development of disease from short telomeres. In normal aging, sufficient stem cell number and function are maintained to sustain vital processes. Even a patient who receives a limited number of hematopoietic cells from an adult donor is capable of maintaining normal hematopoiesis for many years, at least in part related to normal telomerase function and telomere repair. When symptoms develop as a consequence of short telomeres, a disease process is at work.

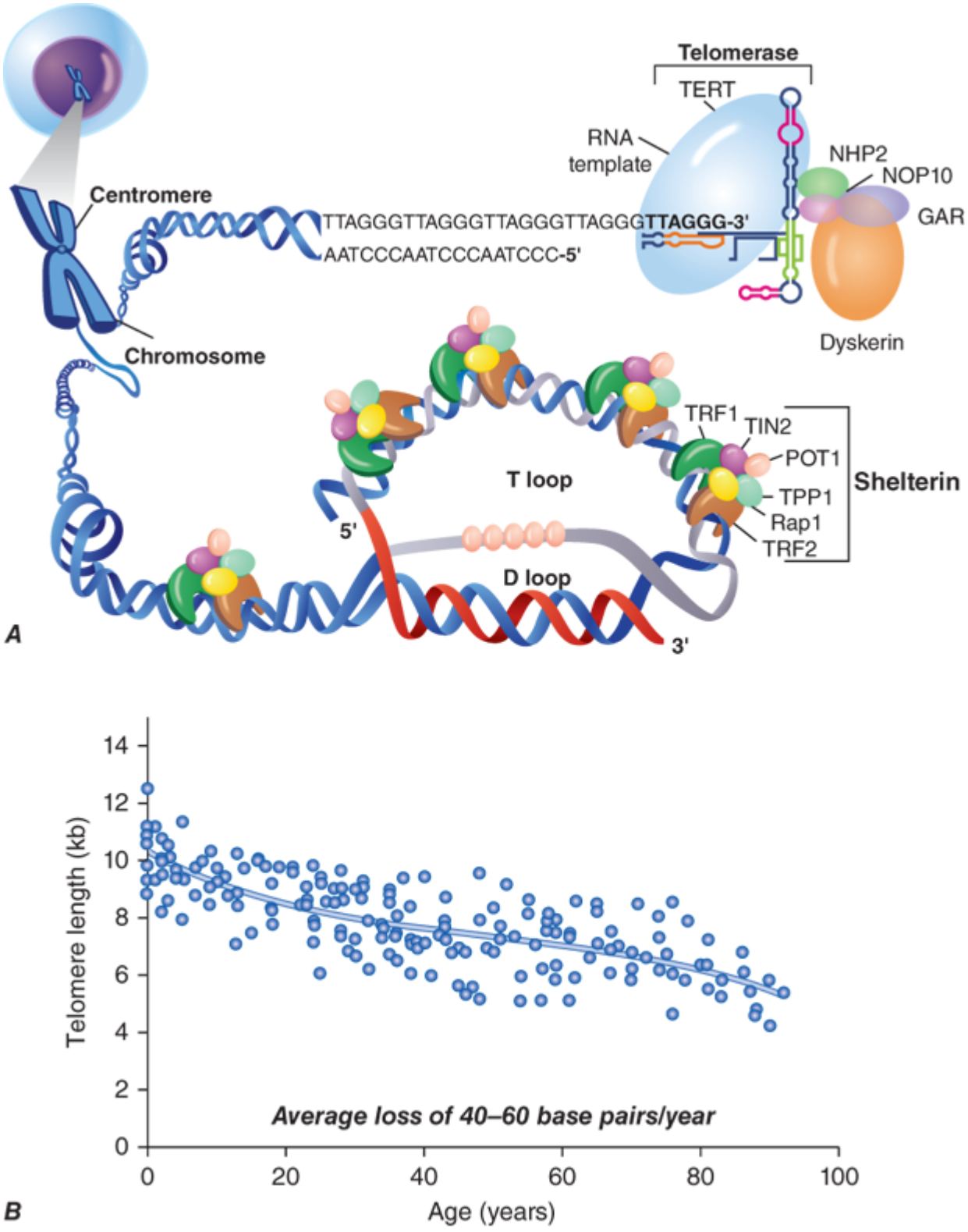
DISEASE MECHANISM

Telomeres, the physical termini of linear chromosomes, are repeated hexanucleotide sequences physically associated with specific proteins. Telomeres function to protect the chromosome ends against recognition as damaged or infectious DNA by the DNA repair machinery **(Fig. 470-1)**. During mitosis, the DNA polymerase employs an RNA oligonucleotide with a 3' hydroxyl group to prime replication. The primer dissociates as the DNA polymerase advances along the template strand, and a gap is left at the ends of linear DNA molecules: the newly synthesized DNA strand is necessarily shorter than the original template—the “end-replication problem.” Chromosome erosion is thus inevitable with mitotic cell division, but the noncoding telomeric long, repetitive structure buffers loss of genetic information. In human cells, telomeres are composed of hundreds to thousands of TTAGGG tandem repeats in the leading and CCCTAA in the lagging DNA strand. At birth, telomeres are relatively long but they inexorably shorten with chronological aging (Fig. 470-1). In an individual cell, critically short telomere length triggers the p53 pathway, usually leading to proliferative arrest, senescence, and apoptosis. Telomere loss is the molecular basis for the “Hayflick phenomenon,” the

limit to cell division and thus to cell proliferation in tissue culture. If a cell overrides proliferation arrest, extremely short telomeres may engage the DNA damage repair machinery, and chromosome end-to-end fusions, chromosome breaks, aneuploidy, and chromosome instability may occur. In addition to the telomere repeated sequences, a group of specialized proteins, collectively termed *shelterins*, directly bind to or indirectly associate with telomeres, assisting in the organization of the telomere tertiary structure and inhibiting activity of DNA damage response proteins (Fig. 470-1).

Figure 470-1

Telomeres and telomerase. ***A.*** Telomeres are ribonucleoprotein structures located at the termini of linear chromosomes inside the cell nucleus composed of hundreds of tandem hexameric DNA repeats. A group of proteins bind directly or indirectly to telomere sequences in order to provide protection to the structure and are collectively termed *shelterin* or *telosome* (TRF1, TRF2, TIN2, POT1, TPP1, and RAP1). As the 3' end of the leading strand forms a single-stranded overhang, it folds back and invades the telomeric double helix, forming a lariat termed *T loop*. The telomerase complex is composed of the enzyme telomerase reverse transcriptase (TERT), its RNA component (TERC), the protein dyskerin, and associated proteins (NHP2, NOP10, and GAR). This enzymatic complex elongates telomeres by adding GTTAGG hexameric repeats to the 3' end of the telomeric leading strand, using a sequence in TERC as the template. ***B.*** The average telomere length in human leukocytes varies: it is longer at birth (10–11 kilobases) and progressively shortens with aging (6–7 kilobases at age 90 years) at an average loss of 40–60 base pairs/year. However, there is significant variability in telomere length in each given age.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

To escape telomere attrition, cells with high proliferative demand, including embryonic and adult stem cells, lymphocytes, and the majority of cancer cells have at least two mechanisms to preserve telomere length: recombination and the capability to synthesize telomeric repeats. GTTAGG hexanucleotides are added to the 5' end of the leading DNA strand by telomerase, a reverse transcriptase enzyme (TERT), and TERC, its RNA template, (Fig. 470-1). The telomerase holoenzyme complex is composed of two copies of TERT, TERC and

dyskerin, and associated proteins. TERC binds to TERT and serves as the RNA template for its function as a reverse transcriptase. Dyskerin, encoded by *DKC1*, stabilizes the complex, and TCAB1, encoded by the *WRAP53* gene, aids telomerase trafficking to the Cajal bodies, nuclear structures for ribonucleoprotein processing where telomerase associates with telomeres for elongation. Telomerase expression is tightly regulated. MYC and sex hormones stimulate *TERT* transcription, whereas in mature cells the *TERT* gene is heavily repressed. In addition, shelterin proteins can also regulate telomerase function and processivity, modulating its catalytic activity on telomeres. Other proteins also are important for telomere length maintenance. RTEL1, an essential DNA helicase, dismantles t-loops and resolves g-quadruplexes, ensuring adequate telomere elongation.

Pathologic accelerated telomere attrition has a genetic origin. Germ-line loss-of-function mutations in genes involved in telomere maintenance and function impair telomere length repair, thus increasing the rate of telomere erosion in highly proliferative cells. Telomeres reach critically short lengths faster than with normal aging; the consequences are limited cell proliferation and impaired tissue regeneration. Some organs appear to be particularly susceptible to telomere erosion. Billions of blood cells are produced daily **(Chap. 92)**, and telomere attrition curtails cell proliferation, producing a hypocellular marrow and often low blood counts. The liver also is an organ with high proliferative capacity, and telomere dysfunction impairs hepatic regeneration after injury, with a variety of pathologic consequences. The lung alveolar epithelium is in contact with exogenous toxins that stimulate regeneration, and telomere loss may hamper these physiologic responses. However, it remains unclear why other regenerative tissues, like the intestine, are relatively unaffected by telomere dysfunction, or the mechanism by which telomere loss provokes a fibrotic response in the lungs (pulmonary fibrosis), an adipose response in the marrow (aplastic anemia), and both in the liver (hepatic steatosis and cirrhosis).

When telomeres are critically short, the DNA damage response machinery may be recruited, mistaking telomeres for damaged or infectious DNA and forcing inappropriate repair. Activation of this pathway may cause chromosome instability due to end-to-end fusion of chromosomes or translocations; these alterations generate unstable, potentially malignant clones. That telomere dysfunction increases the risk of cancer development has been demonstrated in murine models of telomerase deficiency, and patients with telomere diseases are especially prone to develop acute myeloid leukemia and head and neck squamous cell carcinomas.

GENETICS



The pattern of inheritance is variable: X-linked, autosomal recessive, and autosomal dominant. At least 13 genes are implicated in the etiology of telomeropathies, which may be grouped in three categories **(Table 470-1)**.

TABLE 470-1

Genetic Variants in 13 Genes Involved in Telomere Maintenance, Inheritance Pattern, and Phenotype

Gene	Dyskeratosis Congenita	Aplastic Anemia	Pulmonary Fibrosis	Cirrhosis	MDS/Leukemia
Telomerase					
DKC1	X-L				
TERT	AD/AR	AD/AR	AD	AD	AD/AR
TERC	AD/AR	AD	AD	AD	AD
NOP10	AR				
NHP2	AR				
WRAP53	AR				
Shelterin					
TINF2	AD	AD	AD		
TERF2		AD			
ACD	AD				
Others					
RTEL1	AR	AD/AR	AD		AD
CTC1	AR	AR			
PARN			AD		
USB1	AD				

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; X-L, X-linked.

CLINICAL MANIFESTATIONS

Presentation of telomere disease in the clinic is highly variable—in the tissues affected, in the severity of organ dysfunction, and in patterns of disease within families and between families with similar mutations. In a same family, one individual may be severely affected while close relatives carrying the same mutation are asymptomatic and have normal laboratory results. Asymptomatic carriers, however, may display subclinical organ dysfunction, which may be detected by directed or specialized testing (reduced forced vital capacity on pulmonary function test, hypocellular bone marrow at biopsy, hepatic steatosis on ultrasound). In addition, relatives sharing the same mutation and short telomeres may have different organs affected: for instance, while one individual may have aplastic anemia, another within the pedigree may suffer from pulmonary fibrosis. Environmental factors (smoking, alcohol consumption, viral infection) appear to cooperate with organ damage in these patients and contribute to disease heterogeneity among individuals.

Disease anticipation, in which clinical phenotype manifests at an earlier age in successive generations, is observed in some families with telomeropathies, due to the inheritance of short telomeres in sperm and oocytes.

The diagnosis usually is suggested by personal and family history and can be confirmed by simple measurement of leukocyte telomeres and next-generation sequencing of the most prominent genes encoding for telomere repair enzyme complex and shelterin components.

Dyskeratosis Congenita

Dyskeratosis congenita is the classic telomere disease, a mainly pediatric syndrome diagnosed in the first two decades of life. Affected children have at least two features of the mucocutaneous triad of ungual dystrophy, reticular skin pigmentation, and oral leukoplakia **(Fig. 470-2)**. In severe cases, affected newborns have cerebellar hypoplasia (Hoyeraal-Hreidarsson syndrome) or exudative retinopathy (Revesz syndrome) (Fig. 470-3). Telomeres are usually very short, below the first percentile expected for age **(Fig. 470-4)**. Most patients with dyskeratosis congenita develop bone marrow failure, often requiring transfusions and bone marrow transplant. Pulmonary fibrosis appears in as many as 20% of cases and liver disease in 10%, not infrequently after bone marrow transplant. Other tissues and organs also may be affected (Fig. 470-3). The genetic defects most commonly observed in dyskeratosis congenita patients are in *DKC1*, *TINF2*, *TERT*, and *TERC* genes (Table 470-1).

Figure 470-2

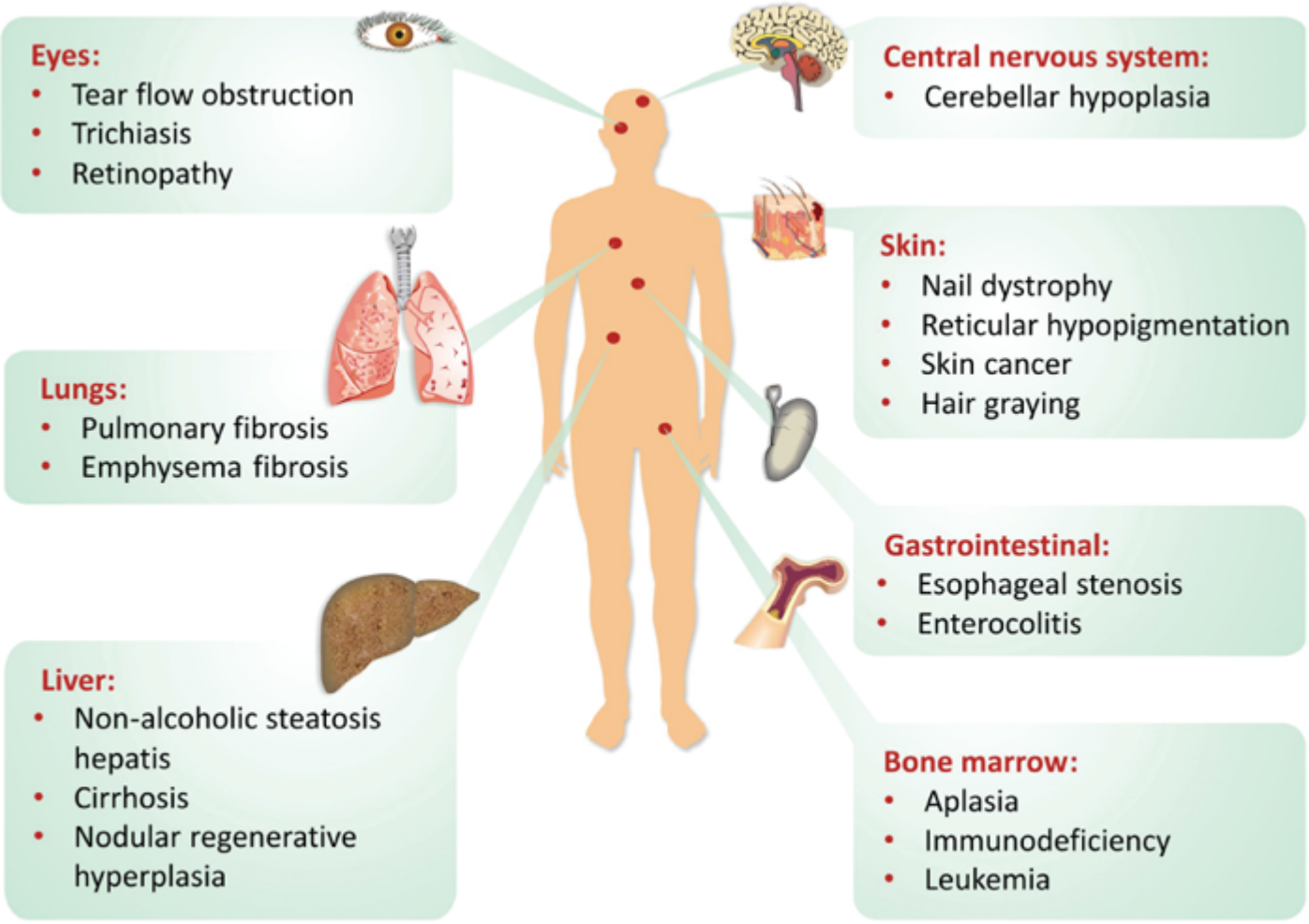
Skin manifestations of dyskeratosis congenita. The pediatric syndrome dyskeratosis congenita characterizes by the mucocutaneous triad of ***(A)*** reticular skin pigmentation, ***(B)*** oral leukoplakia, and ***(C, D)*** nail dystrophy.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Figure 470-3

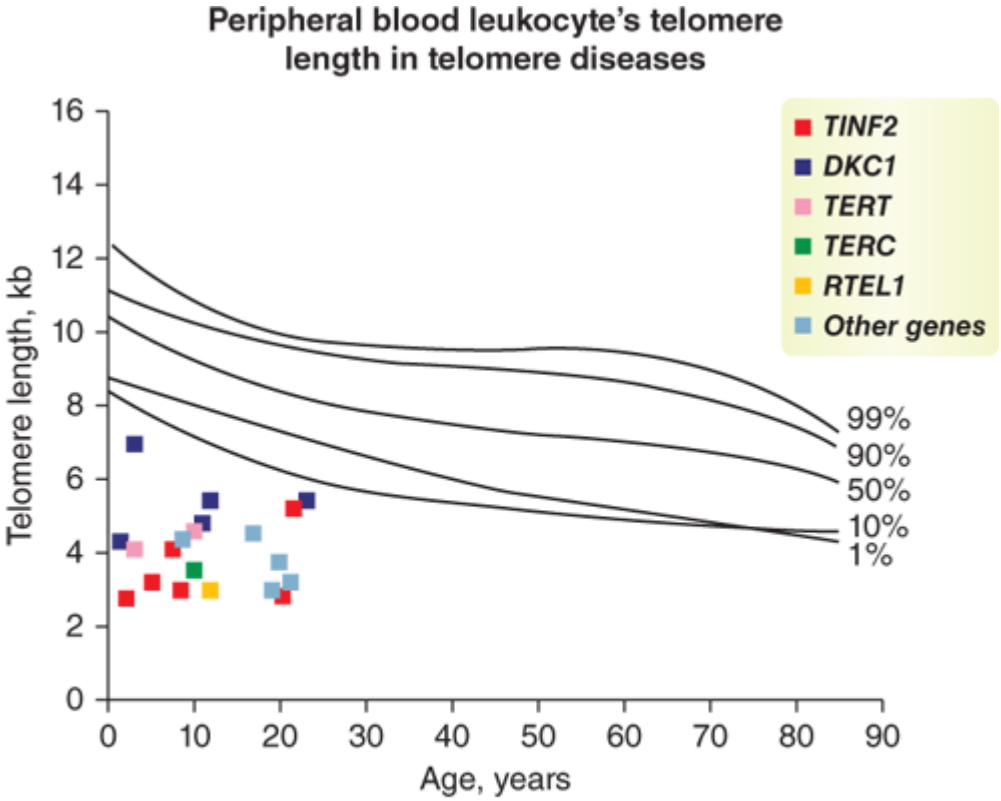
Clinical consequences of telomere diseases. Telomere dysfunction affects a variety of organs: cerebellum, eyes, lungs, liver, skin, gastrointestinal tract, and the bone marrow.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Figure 470-4

Telomere length measurement in dyskeratosis congenita. Telomeres shorten with aging, and solid curves represent the percentiles for age in healthy subjects. Telomeres are considered “short” when below the tenth percentile and very short when below the first percentile. In patients with dyskeratosis congenita, telomeres are usually below the first percentile, regardless of the gene lesion.



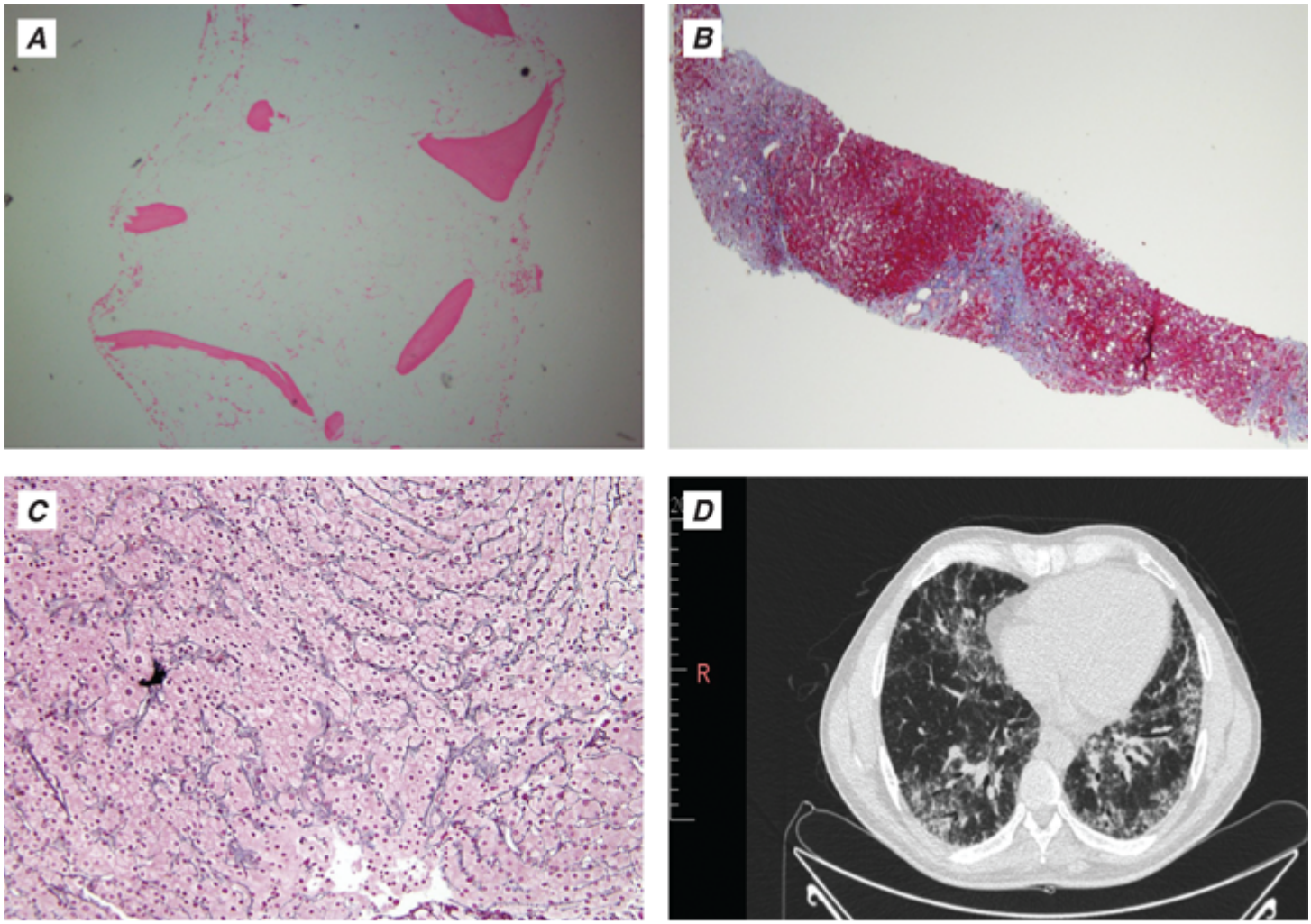
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Bone Marrow Failure

Aplastic anemia (Chap. 98) is the most common major clinical manifestation of dyskeratosis congenita. However, young or older patients carrying a telomere defect, without typical physical stigmata of dyskeratosis, also may also develop marrow failure. Mutations are monoallelic (one mutated allele and one wild-type allele), and *TERT*, *TERC*, and *RTEL1* are the genes usually affected. Telomere loss in these cases is often less intense than in classic dyskeratosis congenita. As a result of inadequate telomerase function, the stem cell pool is limited in size and in its ability to regenerate. There is insufficient production of erythrocytes, platelets, and granulocytes with anemia, thrombocytopenia, and leukopenia of peripheral blood and relatively low marrow cellularity (Fig. 470-5). Patients can present with apparently sudden onset of severe aplastic anemia, but more common is a long history of macrocytic mild to moderate anemia and/or thrombocytopenia, with preservation of leukocyte numbers. A comprehensive personal and family history is important, querying especially for blood abnormalities, and also for lung and liver disease; early hair graying, while not specific to telomeropathies, strongly suggests telomere disease in the appropriate context.

Figure 470-5

Pathologic manifestations of telomere diseases. ***A.*** In the bone marrow, telomere erosion predisposes to aplastic anemia, characterized by an empty hematopoietic marrow replaced by fat (hematoxylin and eosin). ***B.*** In the liver, telomere attrition predisposes to cirrhosis (hematoxylin and eosin). ***C.*** Telomere shortening may also result in nodular regenerative hyperplasia of the liver (reticulin stain). ***D.*** In the lungs, telomere dysfunction predisposes to pulmonary fibrosis mainly in the subpleural regions, which may be detected by high-resolution CT scan.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

When treated, immunosuppressive therapy is generally ineffective in these patients, and they may be more susceptible to pulmonary or hepatic complications after hematopoietic stem cell transplant.

Myeloid Neoplasms

Some patients diagnosed with myelodysplastic syndrome (Chap. 98) or acute myeloid leukemia (Chap. 100) have a family history of bone marrow failure or of other myeloid neoplasms. One of the genetic causes for myeloid neoplasia predisposition is a telomere defect, and these disorders are now classified together by the World Health Organization as "myeloid neoplasms associated with telomere biology disorders." Telomere length measurement may be confounded by the presence of circulating immature cells, which may have very short telomeres, precluding appropriate test interpretation.

Pulmonary Fibrosis

Pulmonary fibrosis appears in about 20% of children with dyskeratosis congenita, and approximately 10–15% of patients with idiopathic pulmonary fibrosis (Chap. 287) or familial pulmonary fibrosis have an etiologic telomerase gene mutation. Most pulmonary fibrosis patients, regardless of mutation status, have

short telomeres for their age but not as short as in dyskeratosis congenita. How telomere erosion causes pulmonary fibrosis is unknown, but it may prevent adequate proliferation and regeneration of pneumocytes type II. Idiopathic pulmonary fibrosis due to a telomere disease usually manifests after the fourth decade of life, with a restrictive pattern on pulmonary function testing associated with decreased diffusion capacity for carbon monoxide (DLCO) and diffuse interstitial lesion on high-resolution CT scan (Fig. 470-5). Histopathology of biopsied lung most commonly shows usual interstitial pneumonia. The pulmonary clinical presentation of patients with telomere disease is indistinguishable from idiopathic pulmonary fibrosis, except that those with an underlying telomere defect may have cryptic hepatic cirrhosis, macrocytosis, cytopenias, or a family history of lung, liver, or bone marrow disease. Pulmonary arteriovenous malformation leading to right-to-left shunting is observed in patients with pulmonary fibrosis due to telomere disease. Patients with idiopathic pulmonary fibrosis or familial pulmonary fibrosis should have their leukocyte telomere length measured and, if telomeres are short, undergo screening for mutations in telomere-associated genes and telomeres; however, telomere length may be normal in some cases despite the presence of pathogenic mutations. *TERT*, *TERC*, *RTEL1*, and *PARN* are the most commonly mutated genes.

Liver Disease

Genetic telomere defects may cause hepatic cirrhosis **(Chap. 337)**, nodular regenerative hyperplasia of the liver, non-alcoholic fatty liver disease **(Chap. 336)**, and hepatocellular carcinoma **(Chap. 78)**, and hepatocytes of most patients with cirrhosis have very short telomeres. Eroded telomeres limit hepatocyte proliferation, especially upon chronic injury. Additionally, hepatocytes with short telomeres display abnormal metabolic pattern and defective mitochondrial function. Abnormal liver pathology may be uncovered incidentally during the evaluation of telomeropathy patients with aplastic anemia or pulmonary fibrosis, but cirrhosis also may be the sole or most prominent clinical presentation of a telomere defect. A minority of individuals with cirrhosis associated with virus B or C infection or alcohol-associated liver disease may carry a telomere-associated gene mutation. Liver histopathology is variable, but cirrhosis is usually associated with inflammation and inflammatory cells (Fig. 470-5), increased iron deposit, positivity for CD34 in sinusoid endothelial cells, and widening of hepatocyte plates. Defective telomere maintenance may increase susceptibility of the liver to environmental challenges, such as viruses and toxins, increasing the risk for developing severe hepatic disease in mutation carriers.

TELOMERE LENGTH MEASUREMENT

Length of telomeres can be accurately measured in peripheral blood leukocytes by commercial laboratories. Of several methods available, flow-FISH and quantitative real-time polymerase chain reaction (qPCR) are the most widely utilized. Both methods have advantages and limitations and require high-quality samples, usually fresh or freshly processed, as cell death and DNA degradation impact the accuracy of testing. Results are usually expressed as leukocyte telomere length in kilobases. However, the interpretation of length must account for patient age, due to physiologic telomere loss. A normal range for telomeres is available for each year of age, longest at birth and shortening at 40–60 base pairs per year (Fig. 470-1). For each age bracket, the

percentile curves are calculated and a given patient's test result is interpreted in the context of normal age variation: telomeres below the tenth percentiles for age are defined as “short” and telomeres below the first percentile are considered “very short” (Fig. 470-4).

Telomeres may also be short in groups of patients with some chronic conditions, such as cardiovascular disease or diabetes. However, in these settings telomere erosion is not assumed to be etiologic, but rather a consequence of chronic inflammation; telomere testing is not known to have clinical utility and is not recommended. Likewise, telomere length tests have no known clinical utility in the assessment of aging and longevity or as a basis for therapeutic interventions absent a diagnosis of telomere disease.

Flow-FISH uses a fluorescent-labeled nucleotide probe specific for telomere repeats in order to estimate telomere content in an individual cell by flow cytometry. It has the advantage of determining telomere length in individual cells and in leukocyte subpopulations (neutrophils, lymphocytes, monocytes); lymphocyte telomere shortening is more specific for telomere diseases than in other cells. However, flow-FISH requires intact cells for analysis, not always available, and neutrophils are susceptible to damage during processing, freezing, and thawing.

Quantitative PCR (qPCR) utilizes telomere-binding modified primers to measure telomere content in comparison to a housekeeping gene in the whole leukocyte population and thus does not require intact cells. qPCR provides an estimate of the average telomere length of a given sample without determining telomere length in individual cells. Good DNA quality is essential for adequate qPCR testing and automation or semi-automation required for clinical purposes, as variability in conditions among batches may result in inter-assay variation.

GENETIC TESTING



When a patient with a suspected telomeropathy has short or very short telomeres, genetic screening for mutations in genes involved in telomere maintenance and biology is warranted (Table 470-1). Genetic testing should be restricted to patients with suspected telomere disorders. Sanger sequencing has been used for screening purposes but has been replaced by next-generation sequencing, and for telomerase complex and shelterin genes commercial panels are available. Mutations may be bi-allelic (especially in dyskeratosis congenita), but in most cases of aplastic anemia, myelodysplastic syndrome, acute myeloid leukemia, idiopathic pulmonary fibrosis, and hepatic cirrhosis, only one gene is mutated. Thus, it is crucial to appropriately interpret genetic screening results, as several rare singleton mutations of unknown significance have been identified in large cohorts of healthy individuals. In silico analysis, mutation location, and functional studies are utilized before declaring a mutation pathogenic.

Adequate genetic counseling is necessary after genetic screening, as the inheritance pattern may be autosomal dominant, mutation penetrance is highly variable, and phenotypes may be diverse even within a pedigree. Potential family stem cell donors must be screened before transplantation to ensure that they do not have mutations.

TREATMENT

TREATMENT

Telomere Disease

Patients with severe aplastic anemia due to telomere disease may undergo allogeneic hematopoietic stem cell transplant when a suitable donor is available. Treatment-related mortality may be increased due to pulmonary and hepatic complications. Lung transplant for pulmonary fibrosis is feasible but often not performed due to coexisting cytopenias, especially thrombocytopenia, and other comorbidities. Similarly, there is no specific treatment for the liver in telomere disease; liver transplant has been attempted in rare cases. Telomeropathy patients should be advised to avoid toxins (metal dust, busulfan, amiodarone), ionizing radiation, cigarette smoke, and alcohol as possibly harmful.

Sex hormones may mitigate telomere attrition and even elongate telomere length. In case reports and a small research trial, danazol improved blood counts in marrow failure patients, and it may even slow progression of lung and liver disease in patients with telomeropathies.

FURTHER READING

Blackburn EH et al: Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection. Science 350:1193, 2015. [PubMed: 26785477]

Calado RT, Young NS: Telomere diseases. N Engl J Med 361:2353, 2009. [PubMed: 20007561]

Collins J, Dokal I: Inherited bone marrow failure syndromes. Hematology 20:433, 2015. [PubMed: 26218202]

Devine MS, Garcia CK: Genetic interstitial lung disease. Clin Chest Med 33:95, 2012. [PubMed: 22365249]

Holohan B et al: Telomeropathies: An emerging spectrum disorder. J Cell Biol 205:289, 2014. [PubMed: 24821837]

Newton CA et al: Telomere-related lung fibrosis is diagnostically heterogeneous but uniformly progressive. Eur Respir J 48:1710, 2016. [PubMed: 27540018]

Savage SA, Bertuch AA: The genetics and clinical manifestations of telomere biology disorders. Genet Med 12:753, 2010. [PubMed: 21189492]

Townsley DM et al: Bone marrow failure and the telomeropathies. Blood 124:2775, 2014. [PubMed: 25237198]

Townsley DM et al: Danazol treatment for telomere diseases. N Engl J Med 374:1922, 2016. [PubMed: 27192671]

Chapter 471: The Role of Epigenetics in Disease and Treatment

Brian C. Capell; Shelley L. Berger

Figure 471-1

INTRODUCTION

The term *epigenetics* was first coined by Waddington in 1942, as he sought to explain how changes in phenotype could occur throughout development independent of any changes to genotype. Appending the prefix *epi-* (Greek, meaning “over, outside of, around”) to genetics aptly describes the numerous mechanisms by which gene expression and phenotypes are influenced, independent of any changes to the underlying DNA sequence. Today, epigenetics occupies one of the most exciting topics in biology and medicine, offering profound opportunities for discovery, as well as promise for the development of new therapies for disease. Interdisciplinary by nature, the field crosses virtually all areas of science and medicine: chemistry and genetics, development and differentiation, immunology, cancer, aging, and neuroscience.

The continuous introduction of ever more powerful techniques for interrogating the epigenome has led epigenetics to become one of the most innovative fields within the biomedical sciences. Given the vast expanse of the topic and limitations of space, in this chapter we provide a broad overview of the field and highlight key areas from across the landscape of biomedicine where epigenetics plays critical roles in disease, and importantly, where epigenetics-based therapies have demonstrated success in clinical medicine.

THE BIOCHEMICAL BASES OF EPIGENETICS

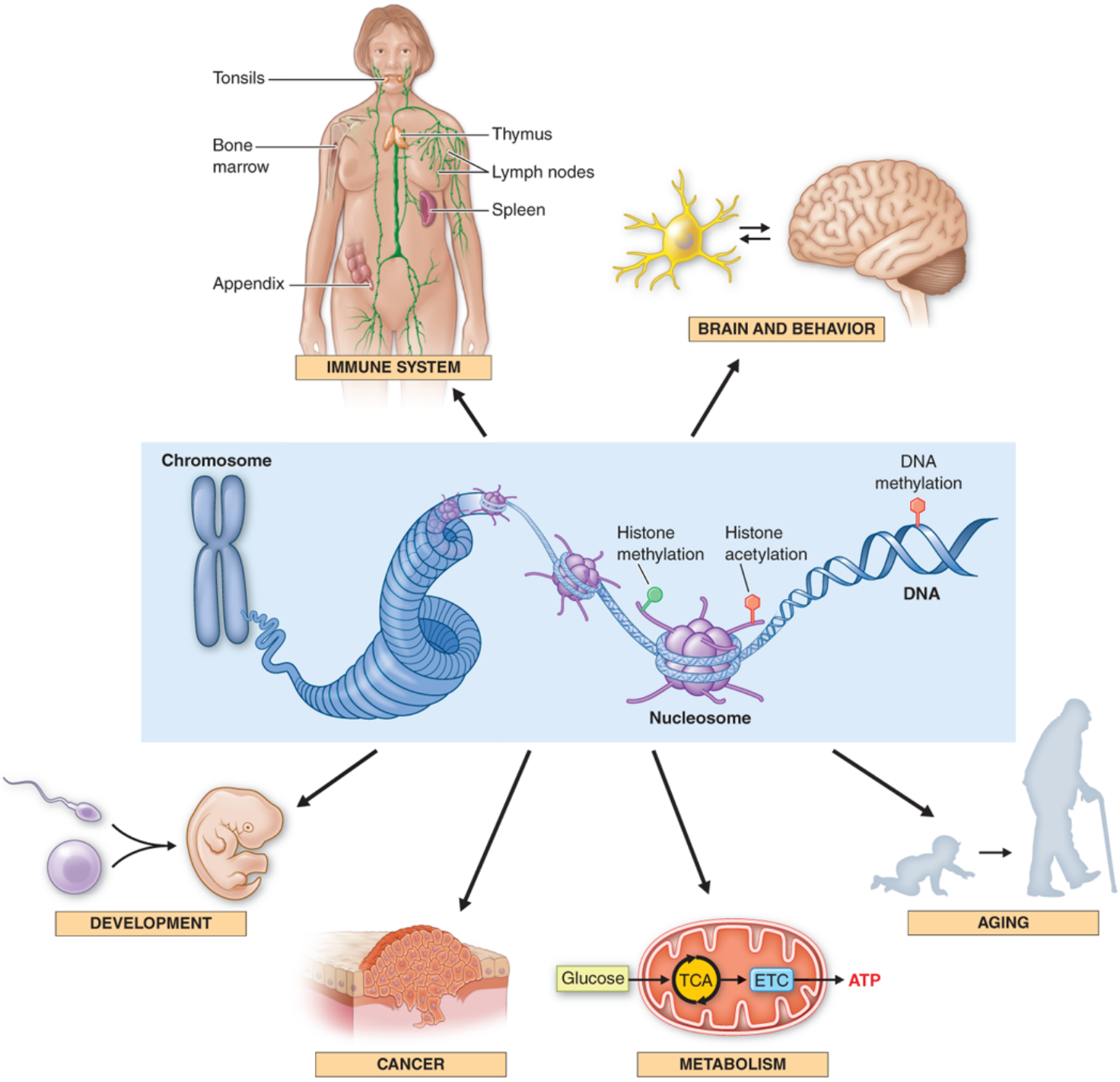
Fundamental to epigenetic regulation is the intricate organization of each cell’s genome into chromatin **(Chap. 456)**. The fundamental unit of the packaging into chromatin is the nucleosome, consisting of 147 base pairs of DNA wrapped around an octamer of 8 histone proteins (two copies of each of the four core histone proteins: H2A, H2B, H3, and H4). The level of compaction of this chromatin structure determines the accessibility of the DNA strand to transcription factors, the DNA repair machinery, and other DNA-binding entities. Thus, compaction has a profound influence on gene expression levels and on local DNA mutation rates. Open regions of chromatin (euchromatin) tend to be transcriptionally active, whereas compacted chromatin (heterochromatin) tends to be transcriptionally repressed.

Histones include the four core histones, which are the most abundant and most frequently found throughout the genome, and variant histones of H2A, H2B, and H3. The structure of core and variant histones include amino- and carboxyl-terminal “tails,” which are extended and unstructured, and highly conserved globular domains. The x-ray crystal structure of the nucleosome particle has illuminated the dynamic alterations of chromatin by an astounding range of regulatory mechanisms, summarized below.

The three main processes that regulate chromatin compaction, and thus access to the DNA template, include direct methylation modifications of the DNA strand itself, post-translational modifications of histones, and remodeling of nucleosomes to alter their location and composition with variant histones **(Fig. 471-1)**. The major modification of DNA is cytosine methylation of CpG dinucleotides (5-mC), associated with gene repression and catalyzed by the DNMT1, DNMT3A, and DNMT3B enzymes. DNMT3A and 3B catalyze the addition of methyl groups on unmethylated DNA de novo at CpG dinucleotides that are typically located throughout transcribed genes and in intergenic regions, but lacking at promoters, while DNMT1 is critical for the maintenance of the methylation state after DNA replication and after transcription during the S phase of the cell cycle. To further alter and reverse methylation, the TET enzymes (TET1–3) progressively oxidize 5-methylcytosine (5-mC) to 5-hydroxymethylcytosine (5-hmC), to 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC), which are unable to be recognized by DNMT1 but can be removed by additional enzymes. Hence, these are mechanisms to passively lose 5-mC following DNA replication, or to actively remove 5-mC, both returning to unmethylated cytosine.

Figure 471-1

Epigenetic pathways influence multiple physiologic and disease pathways. As depicted in the center of the illustration, epigenetics refers to the chemical modifications of DNA and histones, which influence chromatin structure, gene expression, and susceptibility to mutations. These molecular pathways, in turn, play important roles in development, cancer, metabolism, aging, neural function, and behavior, and in the immune system.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Histone post-translational modifications (hPTMs) are rich sources of diverse signaling to, and marking of, the chromatin template, including at least 60 different covalent chemical modifications on the histone N- and C-terminal tails and within the globular domains. The hPTMs are added (written) and removed (erased) by enzymes, and also serve as sequence- and PTM-specific binding surfaces for effector proteins and complexes (readers) to carry out a wide range of downstream actions including transcription, replication, and DNA repair and recombination. One key point is that the staggering numbers of writers, erasers, and readers provide unlimited potential for diagnostic and therapeutic pharmacologic discovery.

Throughout this chapter, we focus on histone methylation and acetylation, the most abundant and the most well-studied hPTMs (Fig. 471-1), although several others, such as serine/threonine/tyrosine phosphorylation, lysine ubiquitination, lysine SUMOylation, and lysine ADP-ribosylation also play important roles in epigenetic regulation. For instance, histone phosphorylation targets histone H2A at Ser139 (γH2A.X), which marks DNA double-strand breaks immediately following damage and is critical for the recruitment of the DNA repair machinery. Histone mono-ubiquitination functions similarly to other hPTMs, in signaling and marking the chromatin template, in particular serving to mark the initiation region or elongation of transcribed genes for future rounds of transcription, whereas histone SUMOylation plays a role in transcriptional repression. Poly-ubiquitination serves to tag proteins for degradation by the proteasome, and dysfunction in this system may play a role in the pathogenesis of neurodegenerative diseases such as Alzheimer's, Parkinson's, and Huntington's disease. ADP-ribosylation involves a class of enzymes, the poly-ADP-ribose polymerases (PARPs), which transfer ADP-ribose

units from NAD+ to a variety of nuclear proteins. This PARylation alters the chromatin environment through the recruitment and modification of chromatin-associated proteins. In general, future studies of the wealth of types and functions of hPTMs will enhance our understanding of these chromatin-based mechanisms and processes and will illuminate new opportunities and targets for therapies.

In contrast, there is extensive understanding of histone lysine acetyltransferases (KATs) and methyltransferases (KMTs). KATs, previously known as HATs, were among the first identified histone modification enzymes. They attach acetyl groups on the lysine residues of histone tails and other proteins, resulting in both a novel side chain (acetyl-lysine) and an increase in negative charge (from positive charged lysine to neutral acetyl-lysine). This alteration results in loosening of chromatin structure to become more permissive to the binding of transcription factors, and it also creates a novel binding surface for the association of reader proteins. Acetylation on core histones, such as lysine 9 on histone H3 (H3K9ac) or lysine 27 (H3K27ac), is typically associated with transcriptional activation. Acetylation is very dynamic and can be rapidly removed by histone deacetylases (HDACs), of which there are multiple classes, including HDACs and sirtuins (NAD-dependent deacetylases), which return the lysine to unmodified ground state.

Methylation of histone tails by KMTs provides more nuanced regulation, in that particular methylated lysines are associated with transcriptional activation (e.g., H3K4me3, H3K36me3, H3K79me3), transcriptional repression (e.g., H3K27me3), or DNA repeat and centromeric silencing (e.g., H3K9me3). The output is strictly determined by effector protein binding, as methylation of lysine does not alter side chain electrostatic charge. Lysine methylation is a more stable chemical modification than is acetylation, and, while demethylases are identified for some of the specific methylated sites (H3K4, H3K9, H3K36, H3K27), it is provocative that H3K79 and H4K20 demethylases have not yet been discovered.

Frequently coordinating with histone modification enzymes are nucleosome remodeling enzymes, which use the energy derived from the hydrolysis of ATP to reposition and remove nucleosomes along the DNA template, and to exchange core histones and variant histones (including variants that are located at the transcriptional initiation sites [H2AZ] and over the transcribed genes [H3.3]). These complexes activate or repress transcription. The SWI/SNF family creates nucleosome-free regions for transcriptional activation, the ISWI family evenly spaces nucleosomes to repress transcription, and the INO80 family exchanges H2A with H2AZ at transcription start sites to poise transcriptional activation. Other remodeling complexes play key roles in the DNA damage response and apoptosis, among additional genomic processes.

Because multiple enzymes redundantly write, erase, and recognize many of the hPTMs, there is great complexity and the potential for fine-tuning of gene regulation. While extensive knowledge gaps remain to fully explicate mechanisms of chromatin regulation, epigenetics has become a fully established discipline within biomedical research. In the coming years, it is likely that the basic understanding of these processes will be harnessed for further betterment of human health.

EPIGENETICS IN DEVELOPMENT AND DIFFERENTIATION

Epigenetic processes are critical to organismal development and to cellular differentiation and reprogramming of cell fate (Fig. 471-1). Transcription factors establish the epigenomic landscape that enables and stabilizes cell type–specific gene expression while simultaneously ensuring stable repression of alternative cell fates. This results in chromatin profiles that display remarkable cell-type specificity in differentiated cells, particularly at the key regulatory nodes of gene enhancers. In fact, epigenome profiling of the chromatin landscape in tumors of unknown cell origin can provide a better index of the origin tissue than does sequencing the tumor itself to catalog genetic mutations.

The cell type–specific epigenetic program is first derived from the template of embryonic stem cells, where numerous genes required for differentiation exist in a “bivalent” state, marked by both the activating histone modification, H3K4me3, and the repressive modification, H3K27me3. From this state, the genes are thought to be “poised” and ready either for activation or for repression, depending on their subsequent cell fate. Critical genes directing toward a specific cell fate will be turned on, whereas genes leading toward alternative fates will be repressed. Once differentiated, an epigenetic barrier will prevent the cells from returning to the stem cell state. For example, heterochromatin in the form of H3K9me3 can serve as a barrier to cellular reprogramming when attempting to create induced pluripotent stem cells, and inhibiting the enzymes that catalyze H3K9me3, such as SUV39H1, can enhance reprogramming efficiency.

DNA methylation contributes to the specification of cell fate and to other developmental pathways. DNA methylation alterations are involved in critical processes ranging from sex chromosome dosage compensation to coordinating expression of imprinted genes. Disruption of this latter process can lead to imprinting disorders including Prader-Willi syndrome, Angelman syndrome, and Beckwith-Wiedemann syndrome.

Beyond embryonic development, epigenetics can provide the necessary coordination and balance between adult stem cell renewal and differentiation. This epigenetic control is critical, as impaired self-renewal can lead to stem cell exhaustion and premature aging, while excessive self-renewal may promote cancer. Key epigenetic regulators tend to play conserved roles across diverse tissue types. For instance, BMI1, a component of the polycomb repressive complex 1 (PRC1), is required for stem cell proliferation and self-renewal, and its ablation leads to stem cell depletion in hematopoietic, epidermal, muscle, intestinal, and mammary stem cells. Similarly, the DNA methyltransferase DNMT1 also is required for stem cell self-renewal in hematopoietic, epidermal, and mammary stem cells. HDACs 1 and 2 possess some overlapping functions and are required for normal epidermal differentiation. Likewise, a loss of these enzymes in hematopoietic stem cells can lead to failure of differentiation and severe anemia. These factors represent repressive chromatin regulation, leading to the concept that restraining specific transcription pathways related to differentiation are crucial to maintaining undifferentiated stem cell pools.

The epigenetic regulation of the tumor suppressor p16 (*CDKN2A*) locus during differentiation provides a prime example of this finely tuned system. For example, as mentioned above, DNMT1 is necessary for self-renewal in human epidermal stem cells. Levels of DNMT1 are high in the basal undifferentiated layer of the epidermis, decreasing progressively with epidermal stratification, leading to derepression of the tumor suppressors p16 and p15, thereby promoting cell cycle arrest and full differentiation. BMI1 displays a similar phenotype in both hematopoietic and epidermal stem cells, repressing key genes that promote differentiation, such as p16 and p19ARF. Consistently, a loss of BMI1 leads to premature differentiation and defective self-renewal. In addition to the repression provided by DNMT1 and BMI1, the p16 locus is highly decorated with the repressive H3K27me3 catalyzed by EZH2 in epidermal stem cells. Then, during epidermal differentiation, H3K27me3 is removed by the KDM6B (JMJD3) histone demethylase. Loss of this control over programmed p16 expression occurs in epithelial cancers, such as squamous cell carcinoma (SCC), where EZH2 is overexpressed and KDM6B expression is lost. In the breast, progesterone can lead to increased levels of EZH2 to promote mammary epithelial cell proliferation. However, excessive EZH2 expression occurs in breast cancers, exemplifying how epigenetics can integrate environmental signals and have a profound influence on the fine balance between stem cell maintenance and overt carcinogenesis. Indeed, loss of key chromatin regulation that promotes cell differentiation, and gain of activities that promote stemness, is a recurrent theme in cancer.

Chromatin modifying enzymes also play a major role in influencing cell type specificity. High levels of EZH2 that modify H3K27me3 promote adipogenesis while simultaneously inhibiting osteogenesis. In contrast, the H3K27me3 demethylases, KDM6A (UTX) and KDM6B (JMJD3), derepress those same genes, driving stem cells toward osteogenesis. Through interactions with tissue-specific master regulators, epigenetic modifiers also shape cell type specificity. In the epidermis, p63, the p53 family member that is a master regulator of the epidermal compartment, interacts with several chromatin regulators including HDAC1 and HDAC2, SATB1, and BRG1 to orchestrate epidermal differentiation. Similarly, the gene-activating H3K4 histone methyltransferases, MLL3 (KMT2C) and MLL4 (KMT2D), are required for adipogenesis by forming a complex with the transcriptional activator ASC2 and the transcription factor PPARγ to induce adipogenic genes.

EPIGENETICS OF METABOLISM

One of the fascinating aspects of epigenetics is that it represents a mechanism for direct connection between the environment and gene expression. Numerous studies in the field of metabolism have identified a complex interplay between diet, metabolism, and the epigenome (Fig. 471-1). Seminal findings in *Drosophila* and mice have shown that changes in diet, particularly the paternal diet, and other environmental factors, can influence the metabolism of offspring, ultimately promoting obesity in later generations. Epidemiologic studies in humans have supported these results, as the nutritional status of grandparents has been correlated with phenotypic effects in grandchildren. In fact, diet can directly affect the levels and activity of chromatin modifiers. For instance, HDAC9 is elevated by high-fat diets, and inhibition of HDAC9 can be protective against the deleterious effects of high-fat diets in mice, including weight gain, decreased glucose tolerance, and increased insulin resistance.

These connections are driven by metabolites that constitute key cofactors for post-translational histone modifications such as acetylation and methylation. These include acetyl-CoA for histone acetylation and S-adenosylmethionine (SAM) for histone and DNA methylation. Levels of acetyl-CoA, in comparison to all measured metabolites, are indeed the best predictor of histone acetylation levels. Consistent with this, increased acetyl-CoA correlates with rising levels of total histone acetylation, including at the promoters of growth-associated genes. This increase in nuclear acetylation is associated with cell cycle progression and proliferation, and it can have clinically relevant downstream effects. For example, high levels of acetyl-CoA can delay stem cell differentiation and suppress autophagy. The oncogenes MYC and AKT can both hijack metabolic networks to enhance nutrient uptake by cancer cells, thus promoting acetyl-CoA production and resulting in both the initiation and progression of tumorigenesis.

Methylation is also altered by metabolism. Dietary factors are estimated to explain 30% of the variation in human serum methionine concentration, and this can alter histone methylation. For example, dietary methionine availability and intracellular production of SAM affects the levels of histone H3K4me3 associated with transcriptional activation. Furthermore, these fluctuations can have critical physiologic consequences: DNA methylation levels in rectal mucosa and colonic polyps are increased by higher levels of dietary folate, and a diet low in methyl donors reduces the formation of gastrointestinal cancers in mice predisposed to these tumors. Methionine metabolism and the availability of SAM regulates stem cell differentiation and contributes to carcinogenesis. For instance, cancers that display hypermethylation, including those with *IDH* mutations, are associated with poorly differentiated gene expression profiles. In contrast, loss of the *MTAP* gene, which is part of the 9p21 locus containing p16, and one of the most frequent events in human cancer, disrupts normal methionine metabolism. While this lowers methylation levels, interestingly, it can also sensitize cancer cells to inhibitors of the PRMT5 methyltransferase, therefore opening a new therapeutic opportunity. These observations illustrate how connections between epigenetics and metabolism can generate unanticipated advances in medicine.

CANCER EPIGENETICS

Cancer is now understood to be a mixed genetic and epigenetic disease, as epigenetic dysregulation is pervasive in human cancers (Fig. 471-1). Beyond simple activation of oncogenes or reduced expression of tumor suppressors, epigenetic mechanisms can contribute to chemotherapy resistance and to failure of antitumor immunity. Accordingly, the development of drugs targeting epigenetic pathways is one of the most active areas of clinical and pharmaceutical development, with several compounds already approved for human use and shown to be effective in a variety of cancers. Epigenetic perturbations in cancer largely affect chromatin-regulating enzymes, which represent robust targets for development of novel small-molecule inhibitors, especially as compared with canonical oncogenic transcription factors (e.g., MYC) and tumor suppressors (e.g., p53).

Epigenetics can contribute to carcinogenesis in a variety of ways. First, on a global scale, chromatin organization is the single most influential factor in determining local mutation rate across the genome. Analysis of abundant tumor sequencing data has demonstrated that heterochromatic regions of the genome contain a higher frequency of mutations compared with more open euchromatic regions. This difference is due to the improved accessibility of the DNA repair machinery to less compact, more open regions of chromatin.

The first discovery of an epigenetic mutation was found in 1998 when the chromatin remodeler *SMARCB1* was shown to drive the formation of malignant rhabdoid tumors. Extensive sequencing of human tumors from the majority of cancer types has been performed by The Cancer Genome Atlas (TCGA) consortium, and, remarkably, 25–30% of identified cancer driver mutations occur in chromatin regulatory proteins. Similar to *SMARCB1*, numerous other chromatin modifiers (e.g., methyltransferases *MLL3* and *MLL4*, and acetyltransferases *EP300* and *CBP*) and nucleosome remodeling enzymes and associated complex components (e.g., *SMARCA4*, *SMARCA2*, *ARID1A*) are heavily mutated and inactivated in many cancers. The majority of these mutations are loss-of-function mutations, and, indeed, enzymes like MLL4 and demethylase KDM6A possess tumor-suppressive activity. In contrast, the H3K27me3 histone methyltransferase EZH2 is an oncogene, and accordingly, it is overexpressed in many advanced-stage or metastatic solid tumors such as breast cancer, prostate cancer, and melanoma. Mechanistically, EZH2 represses the p16 tumor suppressor and other cell cycle genes required for cell cycle exit via H3K27me3 deposition. Consistent with a broad growth regulatory role, EZH2 inhibitors are therapeutically successful for a number of cancers in preclinical models and are being actively studied for B cell lymphoma, melanoma, and other solid tumors.

Recently, provocative evidence has emerged for a direct tumorigenic role of histones based on the discovery of causative mutations, such as histone H3 mutations identified in pediatric high-grade gliomas. Specifically, the majority of these mutations are in the H3 variant H3.3, where lysine 27 is replaced by methionine (K27M). Similarly, over 90% of chondroblastomas replace lysine 36 with methionine (K36M) in histone H3.3. These effects appear to be dominant negative because (1) in H3.3 these are heterozygous mutations, and (2) the mutations also occur in the canonical H3, which exists in approximately 30 orthologous genes in the human genome. Thus, a minority of H3/H3.3 mutant protein leads to global defects in the associated histone modifications (K27 or K36 methylation), possibly via irreversible inhibition of the cognate enzymes by the mutant histones. These histone mutations promote resistance to apoptosis and failure of normal differentiation in a number of pediatric cancers.

Beyond mutations, genetic translocations involving chromatin modifiers also implicate chromatin pathways as direct drivers in cancer. MLL1, the H3K4 histone methyltransferase, is a frequent translocation partner occurring in adult and pediatric acute myeloid leukemia (AML), and in approximately 80% of infant acute lymphoid leukemia (ALL) cases. MLL1 can fuse with more than 70 translocation partners, and these mutant proteins prevent normal hematopoietic differentiation. Consistent with a causative role of MLL1 in these gene fusions, drugs inhibiting the catalytic activity of MLL1 are effective in preclinical models of AML.

Given the abundance of epigenetic abnormalities in cancer combined with the inherent reversibility of epigenetic changes, extensive efforts are underway to develop epigenetic drugs. The first epigenetic therapeutic involved the use of DNA methylation inhibitors (DNMTi) to reactivate tumor suppressor genes. Interestingly, the mechanism of traditional chemotherapeutics, such as azacitidine and decitabine, is to inhibit DNMT1, thereby promoting global hypomethylation; these are currently in clinical use for myelodysplastic syndrome (MDS) and AML. In a second broad mechanism, loss of acetylation occurs in many cancers, and thus HDAC inhibitors (HDACi) are under intensive development. HDACi are effective and approved for treatment in cutaneous T cell lymphoma and multiple myeloma. Bromodomain (BRD)-containing proteins bind to lysine acetylated target proteins, including histones, and rationally designed BET inhibitors (BETi) block their binding. BETi reduce the amplified expression of oncogenes such as *MYC* in hematologic cancers. Current studies are focused on optimizing combinatorial therapies of these pan-DNMTi and pan-HDAC inhibitors with conventional chemotherapies and immunotherapies, particularly given the ability of epigenetic therapeutics to promote re-expression of tumor antigens and interferon (IFN)-mediated antitumor immunity.

There are several hundred chromatin enzymes and binding proteins in the human genome, and the current focus is on the identification of more specific inhibitors. Indeed, targeted inhibitors of numerous mutated chromatin regulators have been developed, with more than 30 compounds currently in various stages of development and preclinical trials. Some notable examples showing early clinical success include EZH2 inhibitors for lymphomas, sarcomas, and melanoma, IDH inhibitors for AML and gliomas carrying mutant *IDH1* or *IDH2* genes, LSD1 inhibitors for AML and small cell lung cancer, and DOT1L and MLL1 inhibitors for leukemias with activated MLL1.

EPIGENETICS OF AGING

Like many diseases of aging, human aging itself results from the complex interplay between genes and the environment. Evidence that the epigenome may be the key link between these processes derives from observations that numerous environmental stimuli and stressors—ranging from diet and exercise to hormones and circadian rhythms—contribute to both aging and epigenetic alterations (Fig. 471-1). Thus, a lifetime of exposures progressively disrupts the chromatin landscape. These age-dependent changes in chromatin organization increase the susceptibility of the genome to mutations and also reduce transcriptional fidelity. Further, provocative findings in model systems demonstrate that stress-induced epigenetic changes can be transmitted over several generations and can even affect the lifespan of later generations. Among these global epigenetic alterations, there is dysregulation of histone modifications and a general loss of histone proteins with aging across taxa. Amazingly, experimental increases in histone levels, particularly histones H3 and H4, but not H2A or H2B, can reverse these age-related changes in mammalian cells and yeast.

Thus, the sum of current evidence suggests a model of aging via a general increase in activating epigenetic modifications along with a loss of repressive modifications. Together these changes create a state of transcriptional instability and “noise” that inhibit accurate transcription. Cells from patients with Hutchinson-Gilford progeria syndrome (HGPS), the most severe form of human premature aging, display reduced levels of both H3K9me3 and H3K27me3

repressive chromatin. In another premature aging disease, Werner syndrome, DNA damage induces global loss of H3K9me3 and H3K27me3 due to the inherent absence of the Werner syndrome ATP-dependent DNA helicase, which is critical for DNA repair. Such heterochromatin loss is not limited to premature aging conditions, as aged cells derived from healthy older humans display age-dependent loss of H3K9me3 leading to aberrant expression of normally repressed transposable elements. Activation of these mobile elements correlates with neurodegenerative disorders and may also promote other aging-related phenotypes such as cancer. Human fibroblasts undergoing cellular senescence (exit from cell cycle due to replicative or other stress) undergo destabilization of compact heterochromatin adjacent to the nuclear periphery, in so-called lamin-associated-domains (LADs). At LADs, in addition to a reduction of repressive histone modifications as discussed above, there are broad new regions of the euchromatic histone modification H3K4me3.

In addition to age-associated alterations of histone modifications, direct manipulation of chromatin modifying enzymes that control these marks affects the balance between heterochromatic and euchromatic regions, and it alters the lifespan of model organisms. Inhibiting the H3K27me3 histone demethylase UTX (KDM6A) results in increased repressive H3K27me3 and extended lifespan in *Caenorhabditis elegans*. Consistent with this, genetic reduction of enzymes (*ash-2*, *set-2*, *wdr-5*) that add the activating H3K4me3 histone modification also extends lifespan in *C. elegans*. The consequences of these genetic manipulations nicely correspond to the observed changes in histone modifications as described above. Beyond histone modifying enzymes, dysregulation of the levels or function of chromatin remodelers can also affect lifespan in model organisms. This dysregulation occurs in humans as well, as in the nucleosome remodeling deacetylase complex (NuRD), which is reduced in HGPS fibroblasts and in aged healthy donors.

In addition to age-related changes in histone methylation, histone acetylation also contributes to aging phenotypes. Dysregulation of histone acetyltransferases (HATs) and HDACs is associated with reduced longevity across model organisms. Further, sirtuin deacetylases (class III NAD^+-dependent HDACs) promote healthspan and lifespan across species as key mediators of pro-longevity effects of caloric restriction. Indeed, loss of Sirt6 results in premature aging in mice. As discussed previously, metabolism and acetylation are intricately linked, and the sirtuins, via NAD^+ levels, and other HDACs may play key roles connecting the environment, gene expression, and physiologic output. For instance, exercise in humans reduces activity of HDACs 4 and 5, leading to increased H3K36ac in skeletal muscle, which likely promotes beneficial gene expression.

Epigenetic alterations with aging are not limited to histone modifications and extend to DNA methylation. Consistent with the histone patterns, DNA methylation data support the model described above—that is, general decompaction of the epigenome with aging. Specifically, levels of 5-methylcytosine (5-mC) are reduced in senescent humans cells, and global DNA hypomethylation occurs across the human genome with aging. Concurrent with this overall hypomethylated state, there are local regions of hypermethylation focused near CpGs at gene promoters, particularly at genes that maintain cellular differentiation and cell identity. This epigenetic disruption during aging thus leads to profound changes in transcription. For example, in HSCs, DNA hypermethylation blocks proper binding of transcription factors, resulting in dysregulation of normal gene expression with aging. Importantly, these patterns are not merely stochastic alterations in response to environmental stressors throughout aging. Indeed, the methylation status of a defined number of CpG sites is a highly accurate predictor of chronologic age in human tissues. This work reveals that DNA methylation status with aging outperforms previous standard biomarkers of aging, such as p16 expression levels and telomere length, and will be highly valuable in the near future to gauge effects of treatment aiming to ameliorate diseases of aging.

EPIGENETICS OF THE BRAIN AND BEHAVIOR

Brain disorders are among the greatest clinical challenges to understand and to treat. Most neurologic and psychiatric disorders result from complex dysregulation of numerous genes and pathways. In this interplay between underlying genetic predisposition and external environmental factors, aberrant epigenetic regulation is increasingly recognized as a potentially key modulator (Fig. 471-1).

More directly, however, several progressive neurodevelopmental disorders are caused by germline mutations in chromatin regulators. Mutations in methyl CpG binding protein 2 (*MECP2*), a protein important for binding to methylated DNA and contributing to gene repression, is the major cause of Rett syndrome. MeCP2 loss leads to overactive gene transcription in neurons and impaired presynaptic excitatory functions. Similarly, Kabuki syndrome, another progressive neurodevelopmental disorder, is caused by germline mutations in either the H3K4me1 histone methyltransferase, *MLL4* (*KMT2D*), or the H3K27me3 demethylase, *UTX* (*KDM6A*). This disorder may derive from dysregulation of transcriptional enhancers, a major class of gene regulatory elements, as both MLL4 and UTX play a key role in activation of enhancers. Finally, the acetyltransferase CBP (*CREBBP*) also is important for gene enhancer function, and when mutated can lead to Rubinstein-Taybi syndrome, a cause of intellectual disability.

Beyond germline mutations, altered methylation dynamics can drive disorders of neural development and of neurodegeneration. Fragile X syndrome, characterized by learning disabilities and cognitive impairment, is caused by mutations in the *FMR1* or *FMR2* genes, or by hypermethylation of the transcriptional promoters regulating *FMR1* or *FMR2*. Similarly, Prader-Willi syndrome and Angelman syndrome, neurodevelopmental conditions caused by abnormal imprinting of the paternal or maternal chromosomal region (15q11-13), respectively, are frequently caused by aberrant DNA methylation. Further, DNA hypomethylation is implicated in some neurodegenerative conditions. For instance, in Parkinson's disease, several genes involved in pathogenesis are hypomethylated due to DNMT1 depletion, including the α-synuclein gene (*SCNA*). In Alzheimer's disease (AD), DNA hypomethylation occurs at promoters of key pathogenic genes such as amyloid precursor protein (*APP*). Indeed, APP promoter methylation is responsive to environmental factors, including aging, a major risk factor for AD. Likewise, presenilin-1 (*PSEN1*) is implicated in AD and displays altered DNA methylation in response to variations in metabolic stimuli. Studies of Huntington's disease (HD) have demonstrated DNA hypomethylation and decreased histone acetylation, in part due to altered function of the acetyl transferase CBP, leading to transcriptional dysregulation. Together these observations underscore epigenetic regulation as a crucial feature of neurodegeneration.

Additional gene regulatory proteins in the nervous system interact with and are regulated by chromatin modifiers. REST (repressor element 1–silencing transcription factor) is important in neuronal homeostasis through its ability to recruit chromatin regulatory enzymes, such as histone deacetylases and histone methyltransferases, and via its control over gene expression. REST levels increase with aging and serve a protective function in neurons against age-associated stressors and loss of cognitive function associated with AD. Similar to REST, brain-derived neurotrophic factor (BDNF), another important mediator of neural development and homeostasis, is implicated in a variety of neurologic and psychiatric disorders including depression, schizophrenia, bipolar disorder, and autism. Knockdown of BDNF in the dentate gyrus leads to depression-like behavior in mouse models, and BDNF mediates effects of antidepressant therapies. Chromatin pathways, including DNA methylation/MeCP2 and H3K27me3, play a key role in BDNF regulation as observed in brains from patients with schizophrenia.

Finally, addiction medicine is another frontier where epigenetics holds great promise to reveal connections between environmental exposure and phenotypes. Although still in its early stages in terms of mechanistic understanding, emerging evidence demonstrates disruption of epigenetic homeostasis as a consequence of addictive substances ranging from alcohol to cocaine. For example, the acetylation of regulatory elements in the *FOSB* gene by the histone acetyltransferase CBP is associated with behavioral effects of cocaine. Ethanol also induces histone acetylation and a decompacted chromatin structure.

EPIGENETIC INFLUENCES ON INFECTION, IMMUNITY, AND INFLAMMATION

Alterations in gene expression patterns are important determinants of immune-mediated disease, and in turn, epigenetics regulates immunity and inflammation (Fig. 471-1). Treatment with immune-stimulating agents such as lipopolysaccharide (LPS) and TNF-α activate expression of numerous inflammatory genes within hours, with precise gene pathways and activation kinetics determined by the cellular epigenetic state. HATs and HDACs are critical components of this response, coordinating with pro-inflammatory transcription factors such as AP-1 and NF-κB, to activate (HATs) and to repress (HDACs). For example, corticosteroids recruit HDAC2 to promoters of NF-κB-stimulated inflammatory genes to prevent activation during asthma treatment.

Type 1 IFN responses are exceptional examples of regulatory complexity governed by epigenetic control. In an unstimulated state, the H3K9 methyltransferases G9a (*EHMT2*) and EHMT1 suppress expression of IFN and IFN-induced genes. Upon induction of IFN-stimulated genes, STAT transcription factors recruit chromatin remodeling complexes, such as BAF (*SMARCA4*), and recruit HATs including p300, CBP, and GCN5 (*KAT2A*). In turn, chromatin remodeling and acetylation recruit chromatin binding proteins including the bromodomain protein, BRD4, which promotes transcriptional elongation and full activation.

Major regulators of adaptive immunity pathways are similarly epigenetically regulated. CD4+ and CD8+ T cells undergo extensive changes in histone modification profiles during differentiation to distinct subsets of effector T cells. For example, genes associated with effector T cell functions in CD8+ memory T cells (such as *PRDM1*, *KLRG1*, *IFNG*) display enrichment of H3K4me3 and low levels of H3K27me3 compared with those genes in naïve T cells. DNA methylation also plays an important regulatory role and may contribute to disease. For example, CD4+ T cells from individuals with rheumatoid arthritis (RA), systemic scleroderma, and latent autoimmune diabetes in adults display hypermethylation of the *FOXP3* gene, which activates regulatory T cells that dampen immune responses. In addition, hypermethylation of the *CTLA4* locus occurs in regulatory T cells from RA patients, impairing their immunosuppressive abilities.

Although mechanistic studies remain limited, there are numerous examples of epigenetic-based therapies associated with extensive effects on the immune system, underscoring the potential hope for eventual treatment of immune-related conditions. For example, the DNA methylation inhibitors azacitidine and decitabine have immunosuppressive effects possibly mediated by enhanced expression of *FOXP3*, which generally suppresses immune responses. HDAC inhibitors upregulate and downregulate immune genes, and they inhibit cytokine production in macrophages from patients with RA. Further, the HDAC inhibitors vorinostat and panobinostat inhibit primary B cell responses and antibody production in vitro and in vivo. Given these broad effects, it is not surprising that the HDAC inhibitor trichostatin A (TSA) has efficacy in various model systems for treatment of RA, systemic lupus erythematosus (SLE), asthma, acute kidney injury, sepsis-induced lung and cardiac damage, and acute pancreatitis. Similarly, BET inhibitors also display broad effects. They block antigen presentation and T and B cell activation, and thus have beneficial protective effects in a variety of inflammatory settings including autoimmunity, sepsis, atherosclerosis, psoriasis, periodontitis, and arthritis. Beyond these “broad-spectrum” epigenetic inhibitors, a specific inhibitor of the H3K27me3 demethylases KDM6A and KDM6B, GSK-J4, has anti-inflammatory activity, presumably by preventing loss of H3K27me3 repression over inflammatory genes.

CONCLUSIONS

Due to the enormity and complexity of the chromatin and epigenetics fields, and their reach into all areas of biology and medicine, it is not possible to cover such a broad scope in a single chapter. Thus, here we provide a concise snapshot highlighting key areas of development in medicine. We hope to have conveyed the tremendous excitement and promise that pervades the discipline. Indeed, given the exponential growth in uncovering the interface between the epigenome and epigenetic therapies with the environment and disease, there is little doubt that the coming years will bring important additions to this field.

FURTHER READING

Allis CD, Jenuwein T: The molecular hallmarks of epigenetic control. Nat Rev Genet 17:487, 2016. [PubMed: 27346641]

Avgistinova A, Benitah SA: Epigenetic control of adult stem cell function. Nat Rev Mol Cell Biol 17:643, 2016. [PubMed: 27405257]

Benayoun BA et al: Epigenetic regulation of ageing: Linking environmental inputs to genomic stability. Nat Rev Mol Cell Biol 16:593, 2015. [PubMed: 26373265]

Brien GL et al: Exploiting the epigenome to control cancer-promoting gene-expression programs. Cancer Cell 29:464, 2016. [PubMed: 27070701]

Jones PA et al: Targeting the cancer epigenome for therapy. Nat Rev Genet 17:630, 2016. [PubMed: 27629931]

Kinnaird A et al: Metabolic control of epigenetics in cancer. Nat Rev Cancer 16:694, 2016. [PubMed: 27634449]

Sen P et al: Epigenetic mechanisms of longevity and aging. Cell 166:822, 2016. [PubMed: 27518561]

Tough DF et al: Epigenetic drug discovery: Breaking through the immune barrier. Nat Rev Drug Discov 15:835, 2016. [PubMed: 27765940]

Tsankova N et al: Epigenetic regulation in psychiatric disorders. Nat Rev Neurosci 8:355, 2007. [PubMed: 17453016]

Yao B et al: Epigenetic mechanisms in neurogenesis. Nat Rev Neurosci 17:537, 2016. [PubMed: 27334043]

Chapter 472: Mitochondrial DNA and Heritable Traits and Diseases

Karl L. Skorecki; Bruce H. Cohen

INTRODUCTION

Mitochondria are cytoplasmic organelles whose major function is to generate ATP by the process of oxidative phosphorylation under aerobic conditions. This process is mediated by the respiratory electron transport chain (ETC) multiprotein enzyme complexes I–V and the two electron carriers, coenzyme Q_{10} (CoQ_{10}) and cytochrome c. Other cellular processes to which mitochondria make a major contribution include apoptosis (programmed cell death) and additional cell type-specific functions **(Table 472-1)**. The efficiency of the mitochondrial ETC in ATP production is a major determinant of overall body energy balance and thermogenesis. In addition, mitochondria are the predominant source of reactive oxygen species (ROS), whose rate of production also relates to the coupling of ATP production to oxygen consumption. Given the centrality of oxidative phosphorylation to the normal activities of almost all cells, it is not surprising that mitochondrial dysfunction can affect almost any organ system **(Fig. 472-1)**. Until recently, it was thought that disruption of energy production was the source of the pathophysiology in those with mitochondrial dysfunction, but recent evidence suggests that free-radical production and the redox state of the mitochondria may play a role as well. Thus, physicians in many disciplines might encounter patients with mitochondrial diseases and should be aware of their existence and characteristics.

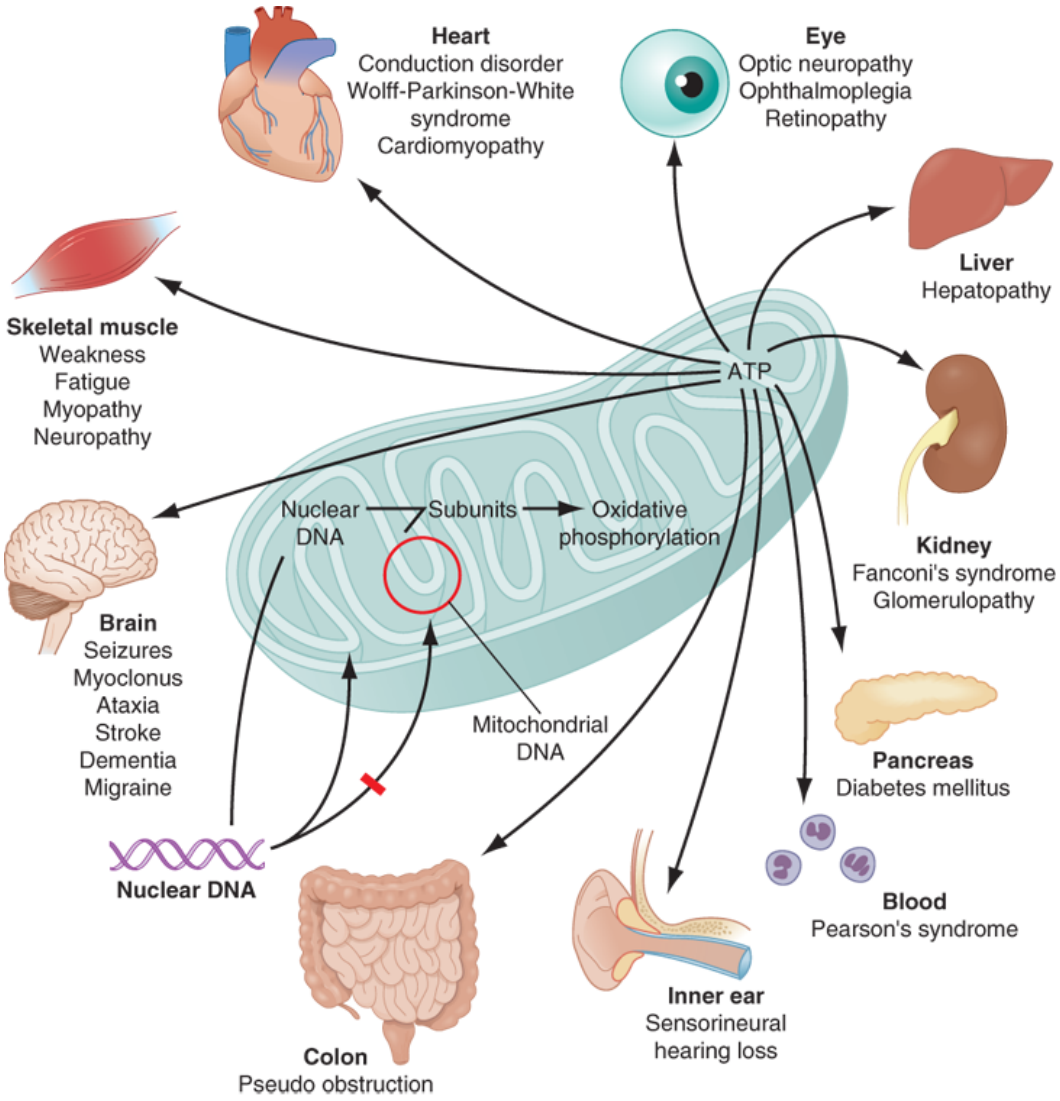
TABLE 472-1

Functions of Mitochondria

All Cells and Tissues
Oxidative phosphorylation Free radical production Calcium homeostasis Apoptosis (programmed cell death)
Tissue- or Cell-Specific
Cholesterol metabolism Amino and organic acid metabolism Fatty acid beta oxidation Sex steroid synthesis Heme synthesis Hepatic ammonia detoxification Neurotransmitter metabolism

FIGURE 472-1

Dual genetic control and multiple organ system manifestations of mitochondrial disease. *(Reproduced with permission from DR Johns: Mitochondrial DNA and disease. N Engl J Med 333:638, 1995.)*

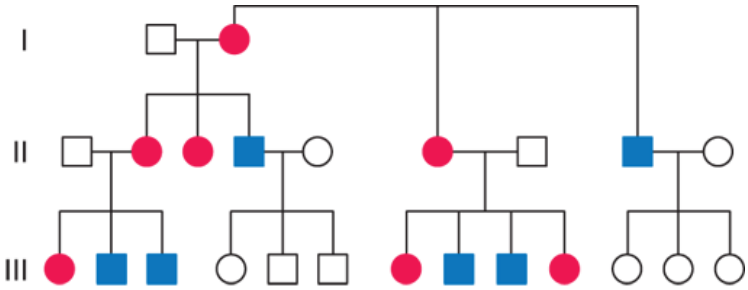


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The integrated activity of estimated 1500 gene products is required for normal mitochondrial biogenesis, function, and integrity. Aside from the 37 genes that comprise the mitochondrial DNA (mtDNA) molecule, the remaining 1400+ gene products are encoded by nuclear genes (referred to as nDNA) and thus follow the rules and patterns of nuclear genomic inheritance (Chap. 456). These nuclear-encoded proteins are synthesized in the cell cytoplasm and imported to their location of activity within the mitochondria through a complex biochemical process. This process includes unfolding of the nuclear-encoded protein, attachment to a chaperone protein that shuttles it through a specific channel to a specific mitochondrial location, and detachment from the chaperone followed by assembly with other mtDNA- and nDNA-encoded proteins. In addition, the mitochondria contain their own small genome consisting of numerous copies (polyploidy) per mitochondrion of a circular, double-strand mitochondrial DNA (mtDNA) molecule comprising 16,569 nucleotides. This mtDNA sequence (also known as the “mitogenome”) might represent the remnants of endosymbiotic prokaryotes from which mitochondria are thought to have originated. The mtDNA sequence contains a total of 37 genes, of which 13 encode mitochondrial protein components of the ETC (Fig. 472-2). The remaining 22 tRNA- and 2 rRNA-encoding genes are mitochondria-specific and dedicated to the process of translating the 13 mtDNA-encoded proteins. The mtDNA itself replicates constantly, independent of cell division, and requires its own unique polymerase, referred to as polymerase gamma (polγ), which is encoded by the nuclear protein *POLG*, disorders of which are discussed in Chaps. 457 and 441. However, mutations in *POLG* can disrupt the endonuclease function of polγ, resulting in somatic mutations in the mtDNA that endure with future replication. Unless this mutation occurs and repopulates in an oocyte, it is not heritable. This dual nuclear and mitochondrial genetic control of mitochondrial function results in unique and diagnostically challenging patterns of inheritance. The current chapter focuses on heritable traits and diseases related to the mtDNA component of the dual genetic control of mitochondrial function. The reader is referred to Chaps. 456 and 441 for consideration of mitochondrial disease originating from mutations in the nuclear genome. The latter include: (1) disorders due to mutations in nuclear genes directly encoding structural components or assembly factors of the oxidative phosphorylation complexes, (2) disorders due to mutations in nuclear genes encoding proteins indirectly related to oxidative phosphorylation, (3) mtDNA depletion syndromes (MDSs) characterized by a reduction of mtDNA copy number in affected tissues without mutations or rearrangements in the mtDNA, and (4) disorders due to mutations in nuclear genes that disrupt normal mitochondrial dynamics (biosynthesis, mitophagy, fission, and fusion).

FIGURE 472-2

Maternal inheritance of mitochondrial DNA (mtDNA) disorders and heritable traits. Affected women (*filled circles*) transmit the trait to their children. Affected men (*filled squares*) do not transmit the trait to any of their offspring.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The classic physical structure of the mitochondria is that of a thread-like organelle, which under fixed conditions, such as observed with immunohistochemical stains or electron microscopy, has a submarine-shape and measures about 1 µm in length. However, in the living state, mitochondrial shape is highly variable based on the cell type, and manifests a complex and ever-changing syncytial form, with continuous appearance and disappearance of budding structures (representing mitochondrial fission) and reorganization of separate mitochondria (representing mitochondrial fusion). Although we often think of the *mitochondrial number* in an individual cell, in fact the more accurate concept in a living cell is probably mitochondrial volume.

Although the presence of mitochondria have been known for >150 years, the first knowledge of their respiratory function was proposed about 100 years ago, and the initial description of an illness linked to mitochondrial dysfunction was only made in 1962. The presence of mtDNA was noted in the 1960s and it was not until 1988 when the first mutations in the mtDNA causing human illness were described. These included the demonstration of a large-scale mtDNA deletion causing Kearn-Sayre Syndrome (KSS) and the discovery of a point mutation in *ND4*, an mtDNA-encoded complex I gene causing Leber's Hereditary Optic Neuropathy (LHON). Following these two discoveries, >400 pathogenic mtDNA mutations or deletions have been reported to cause human disease.

MITOCHONDRIAL DNA STRUCTURE AND FUNCTION

As a result of its circular structure and extranuclear location, the replication and transcription mechanisms of mtDNA differ from the corresponding mechanisms in the nuclear genome, whose nucleosomal packaging and structure are more complex. Specifically, mitochondria have their own transcription system, and the mtDNA itself replicates independently of cellular replication. Because each cell contains many copies of mtDNA, and because the number of mitochondria can vary during the lifetime of each cell, mtDNA copy number is not directly coordinated with the cell cycle. Thus, vast differences in mtDNA copy number are observed between different cell types and tissues and during the lifetime of a cell. Another important feature of the mtDNA replication process is a reduced stringency of proofreading and replication error correction, leading to a greater degree of sequence variation compared to the nuclear genome. Some of these sequence variants are silent polymorphisms that do not have the potential for a phenotypic or pathogenic effect, whereas others may be considered pathogenic mutations. There are some mutations that may be considered ecogenetic, as they typically remain silent, meaning they do not cause disease, unless an external event occurs. One classic example is seen in a common (1:800) mutation in the mitochondrial 12S rRNA gene, m.A1555G., which is associated with hearing loss but is rapidly exacerbated by exposure to normal dosages of an aminoglycoside.

With respect to transcription, initiation can occur on both strands and proceeds through the production of an intronless polycistronic precursor RNA, which is then processed to produce the 13 individual mRNA and 24 individual tRNA and rRNA products. The 37 mtDNA genes comprise fully 93% of the 16,569 nucleotides of the mtDNA in what is known as the *coding region*. The *control region*, which is contained in the D-loop, consists of ~1.1 kilobases (kb) of noncoding DNA, which is thought to have an important role in replication and transcription initiation.

MATERNAL INHERITANCE AND LACK OF RECOMBINATION

In contrast to homologous pair recombination that takes place in the nucleus, mtDNA molecules do not undergo recombination, such that mutational events represent the only source of mtDNA genetic diversification. Moreover, it is only the maternal DNA that is transmitted to the offspring. The fertilized oocyte degrades mtDNA carried from the sperm in a complex process involving the ubiquitin proteasome system and autophagy which takes place on the inner membrane of the oocyte. Thus, although mothers transmit their mtDNA to both their sons and daughters, only the daughters are able to transmit the inherited mtDNA to future generations. Accordingly, mtDNA sequence variation and associated phenotypic traits and diseases are inherited exclusively along maternal lines, meaning both sons and daughters have equal chances of having symptomatic disease, with the only significant exception being LHON as described below.

The phenotypic expression, including age of onset and the exact pattern of organ dysfunction, of a pathogenic mtDNA mutation may vary greatly, even within families. Because of this complex relationship between mtDNA mutations and disease expression, sometimes it is difficult to recognize the maternal pattern of inheritance at the clinical or pedigree level. However, evidence of paternal transmission can almost certainly exclude an mtDNA genetic origin of phenotypic variation or disease; conversely, a disease affecting both sexes without evidence of paternal transmission strongly suggests a heritable mtDNA disorder (Fig. 472-2).

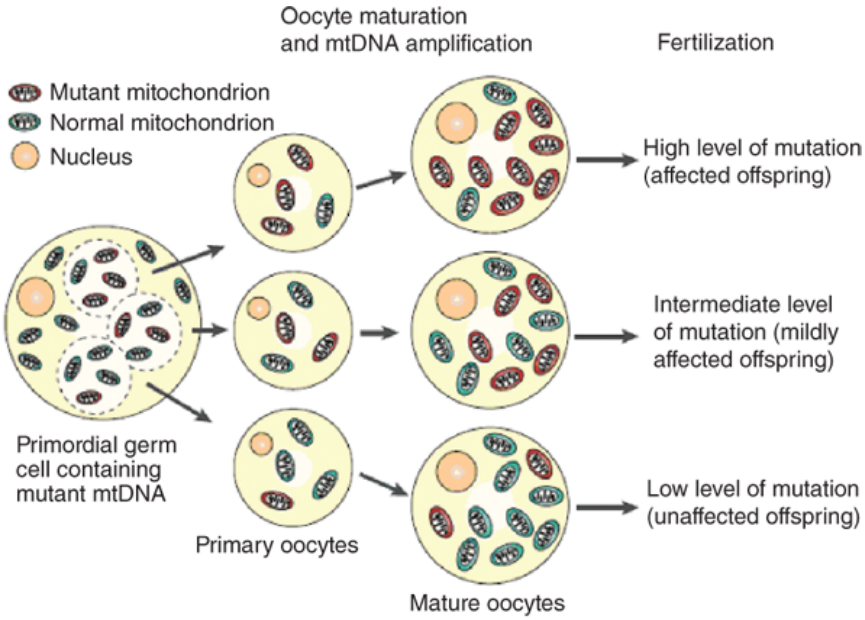
MULTIPLE COPY NUMBER (POLYPLOIDY), HIGH MUTATION RATE, HETEROPLASMY, AND MITOTIC SEGREGATION

Each aerobic cell in the body has multiple mitochondria, often numbering many hundreds or more in cells with extensive energy production requirements. Furthermore, the number of copies of mtDNA within each mitochondrion varies from several to hundreds; this is true of both somatic as well as germ cells, including oocytes in females. In the case of somatic cells, this means that the impact of most newly acquired somatic mtDNA mutations is likely to be very small in terms of total cellular or organ system function; however, because of the many fold higher mutation rate during mtDNA replication, numerous different mutations may accumulate with aging of the organism. It has been proposed that the total cumulative burden of acquired somatic mtDNA mutations with age may result in an overall perturbation of mitochondrial function, contributing to age-related reduction in the efficiency of oxidative phosphorylation and increased production of damaging ROS. Because mtDNA (and nDNA) mutations may result in electron leak within the ETC, the ROS damage may rise above the normal baseline in some with specific mutations, resulting in increased susceptibility to somatic mtDNA damage and disease expression. The accumulation of such acquired somatic mtDNA mutations with aging may contribute to age-related diseases, such as metabolic syndrome and diabetes, cancer, and neurodegenerative and cardiovascular disease in any given individual. However, somatic mutations are not carried forward to the next generation, and the hereditary impact of mtDNA mutagenesis requires separate consideration of events in the female germline.

The multiple mtDNA copy number within each cell, including the maternal germ cells, results in the phenomenon of heteroplasmy, in contrast to much greater uniformity (homoplasy) of somatic nuclear DNA sequence. Heteroplasmy for a given mtDNA sequence variant or mutation arises as a result of the coexistence within a cell, tissue, or individual of mtDNA molecules bearing more than one version of the sequence variant **(Fig. 472-3)**. The importance of the heteroplasmy phenomena to the understanding of mtDNA-related mitochondrial diseases is critical. The coexistence of mutant and nonmutant (wild-type) mtDNA and the variation of the mutant load, which can be thought of as the percentage of mutant mtDNA molecules within a specific cell, tissue, organ or organism, contributes to the expression of a phenotype among individuals from the same maternal sibship. At the level of the oocyte, the percentage of mtDNA molecules bearing each version of the polymorphic sequence variant or mutation depends on stochastic events related to partitioning of mtDNA molecules during the process of oogenesis itself. Thus, oocytes differ from each other in the degree of heteroplasmy for that sequence variant or mutation. In turn, the heteroplasmic state is carried forward to the zygote and to the organism as a whole, to varying degrees, depending on mitotic segregation of mtDNA molecules during organ system development and maintenance. For this reason, in vitro fertilization, followed by preimplantation genetic diagnosis (PGD), is not as predictive of the genetic health of the offspring in the case of mtDNA mutations as in the case of the nuclear genome. Similarly, the impact of somatic mtDNA mutations acquired during development and subsequently also shows an enormous spectrum of variability. In general, a higher mutant load will result in a more severe, and earlier phenotypic presentation. However, measuring heteroplamy in one tissue (lymphocytes from blood or urine sediment containing kidney and bladder epithelial cells for example) may not represent the percentage of mutant heteroplasmy in the tissue or organs most affected, such as the cardiac atrioventricular node or brain. Furthermore, the threshold of mutant heteroplasmy which results in clinical illness may vary depending on the specific mutation.

FIGURE 472-3

Heteroplasmy and the mitochondrial genetic bottleneck. During the production of primary oocytes, a selected number of mitochondrial DNA (mtDNA) molecules are transferred into each oocyte. Oocyte maturation is associated with the rapid replication of this mtDNA population. This restriction-amplification event can lead to a random shift of mtDNA mutational load between generations and is responsible for the variable levels of mutated mtDNA observed in affected offspring from mothers with pathogenic mtDNA mutations. Mitochondria that contain mutated mtDNA are shown in *red*, and those with normal mtDNA are shown in *green. (Reproduced with permission from R Taylor, D Turnbull: Mitochondrial DNA mutations in human disease. Nat Rev Genetics 6:389, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Mitotic segregation refers to the unequal distribution of wild-type and mutant versions of mtDNA molecules during all cell divisions that occur during prenatal development and subsequently throughout the lifetime of an individual. The phenotypic effect or disease impact will be a function not only of the inherent disruptive effect (pathogenicity) on the mtDNA-encoded gene (coding region mutations) or integrity of the mtDNA molecule (control region mutations), but also of its distribution among the multiple copies of mtDNA in the various mitochondria, cells, and tissues of the affected individual. Thus, one consequence can be the generation of a bottleneck due to the marked decline in given sets of mtDNA variants, pathogenic and nonpathogenic, consequent to such mitotic segregation. It is postulated that the main effects of this bottleneck occur between the primordial germ cell state and the primary oocyte stage of development. Heterogeneity arises from differences in the degree of heteroplasmy among oocytes of the transmitting female, together with subsequent, probably random, mitotic segregation of the pathogenic mutation during tissue and organ development, and throughout the lifetime of the individual offspring. The actual expression of disease might then depend on a threshold percentage of mitochondria whose function is disrupted by mtDNA mutations. This in turn confounds hereditary transmission patterns and hence genetic diagnosis of pathogenic heteroplasmic mutations. Generally, if the proportion of mutant mtDNA is <60%, the individual is unlikely to be affected, whereas proportions exceeding 90% cause clinical disease. One notable exception is LHON, in which these mutations are present either in 100% mutant homoplasmy, which causes the disease expression, or 100% wild-type homoplasmy. It is not understood why this specific phenotype and the several known mtDNA alleles that result in LHON behave in this manner.

HOMOPLASMIC VARIANTS AND HUMAN mtDNA PHYLOGENY

In contrast to classic mtDNA diseases, most of which begin in childhood and are the result of heteroplasmic mutations as noted above, during the course of human evolution, certain mtDNA sequence variants have drifted to a state of homoplasmy, wherein all of the mtDNA molecules in the organism contain the new sequence variant. This arises due to a “bottleneck” effect followed by genetic drift during the very process of oogenesis itself (Fig. 472-3). In other words, during certain stages of oogenesis, the mtDNA copy number becomes so substantially reduced that the particular mtDNA species bearing the novel or derived sequence variant may become the increasingly predominant, and eventually exclusive, version of the mtDNA for that particular nucleotide site. All of the offspring of a woman bearing an mtDNA sequence variant or mutation that has become homoplasmic will also be homoplasmic for that variant and will transmit the sequence variant forward in subsequent generations.

Considerations of reproductive fitness limit the evolutionary or population emergence of pathogenic homoplasmic mutations that are lethal or cause severe disease in infancy or childhood. Thus, with a number of notable exceptions (e.g., mtDNA mutations causing LHON; see below), most homoplasmic mutations are considered to be neutral markers of human evolution, which are useful and interesting in the population genetics analysis of shared maternal ancestry but which have little significance in human phenotypic variation or disease predisposition.

More importantly is the understanding that this accumulation of homoplastic mutations occurs at a genetic locus that is transmitted only through the female germline, and that lacks recombination. In turn, this enables reconstruction of the sequential topology and radiating phylogeny of mutations accumulated through the course of human evolution since the time of the most recent common mtDNA ancestor of all contemporary mtDNA sequences, some 200,000 years ago. The term *haplogroup* is usually used to define major branching points in the human mtDNA phylogeny, nested one within the other, which often demonstrate striking continental geographic ancestral partitioning. At the level of the complete mtDNA sequence, the term haplotype is usually used to describe the sum of mutations observed for a given mtDNA sequence and as compared to a reference sequence, such

that all haplotypes falling within a given haplogroup share the total sum of mutations that have accumulated since the most recent common ancestor and the bifurcation point they mark. The remaining observed variants are private to each haplotype. Consequentially, human mtDNA sequence is an almost perfect molecular prototype for a nonrecombining locus, and its variation has been extensively used in phylogenetic studies. Moreover, the mtDNA mutation rate is considerably higher than the rate observed for the nuclear genome, especially in the control region, which contains the displacement, or D-loop, in turn comprising two adjacent hypervariable regions (HVR-I and HVR-II). Together with the absence of recombination, this amplifies drift to high frequencies of novel haplotypes. As a result, mtDNA haplotypes are more highly partitioned across geographically defined populations than sequence variants in other parts of the genome. Despite extensive research, it has not been well established that such haplotype-based partitioning has a significant influence on human health conditions. However, mtDNA-based phylogenetic analysis can be used both as a quality assurance tool and as a filter in distinguishing neutral mtDNA variants comprising human mtDNA phylogeny from potentially deleterious mutations.

MITOCHONDRIAL DNA DISEASE

The true prevalence of mtDNA disease is difficult to estimate because of the phenotypic heterogeneity that occurs as a function of heteroplasmy, the challenge of detecting and assessing heteroplasmy in different affected tissues, and the other unique features of mtDNA function and inheritance described above. It is estimated that at least 1 in 200 healthy humans harbors a pathogenic mtDNA mutation with the potential to causes disease, but that heteroplasmic germline pathogenic mtDNA mutations actually affect up to ~1 in 5000 individuals.

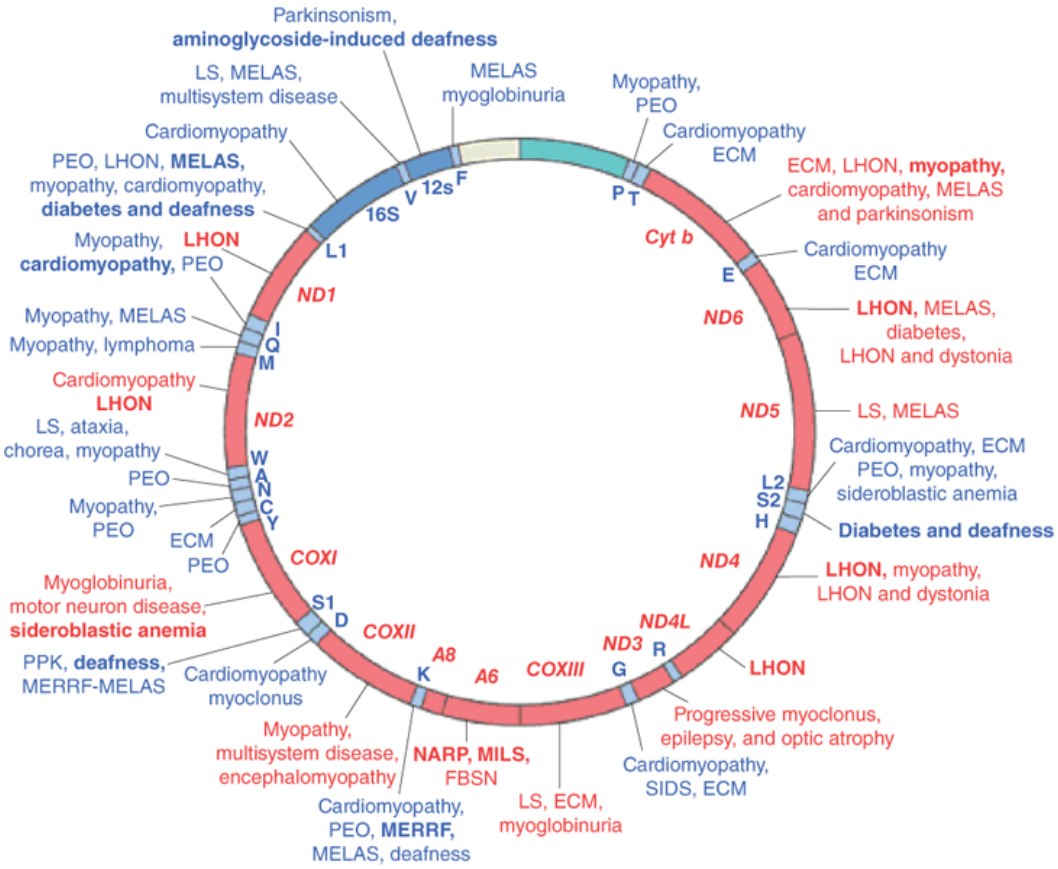
The true disease burden relating to mtDNA sequence variation will only be known when the following capabilities become available: (1) ability to distinguish a completely neutral sequence variant from a true phenotype-modifying or pathogenic mutation, (2) accurate assessment of heteroplasmy that can be determined with fidelity, and (3) a systems biology approach **(Chap. 476)** to determine the network of epistatic interactions of mtDNA sequence variations with mutations in the nuclear genome.

OVERVIEW OF CLINICAL AND PATHOLOGIC FEATURES OF HUMAN mtDNA DISEASE

Given the vital roles of mitochondria in all nucleated cells, it is not surprising that mtDNA mutations can affect numerous tissues with pleiotropic effects. More than 200 different disease-causing, mostly heteroplasmic mtDNA mutations have been described affecting ETC function. **Figure 472-4** provides a partial mtDNA map of some of the better characterized of these disorders. A number of clinical clues can increase the index of suspicion for a heteroplasmic mtDNA mutation as an etiology of a heritable trait or disease, including (1) familial clustering with absence of paternal transmission; (2) adherence to one of the classic syndromes (see below) or paradigmatic combinations of disease phenotypes involving several organ systems that normally do not fit together within a single nuclear genomic mutation category; (3) a complex of laboratory and pathologic abnormalities that reflect disruption in cellular energetics (e.g., lactic acidosis and neurodegenerative and myodegenerative symptoms with the finding of ragged red fibers, reflecting the accumulation of abnormal mitochondria under the muscle sarcolemmal membrane); or (4) a mosaic pattern reflecting a heteroplasmic state. There are no truly sensitive and specific biomarkers of disease, and the presence of a historically quintessential finding of ragged red fibers can be seen in numerous muscle disorders, so laboratory tests must always be interpreted in the context of their limitations and should not be used to define the disease.

FIGURE 472-4

Mutations in the human mitochondrial genome known to cause disease. Disorders that are frequently or prominently associated with mutations in a particular gene are shown in *boldface*. Diseases due to mutations that impair mitochondrial protein synthesis are shown in *blue*. Diseases due to mutations in protein-coding genes are shown in *red*. ECM, encephalomyopathy; FBSN, familial bilateral striatal necrosis; LHON, Leber's hereditary optic neuropathy; LS, Leigh syndrome; MELAS, mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes; MERRF, myoclonic epilepsy with ragged red fibers; MILS, maternally inherited Leigh syndrome; NARP, neuropathy, ataxia, and retinitis pigmentosa; PEO, progressive external ophthalmoplegia; PPK, palmoplantar keratoderma; and SIDS, sudden infant death syndrome. *(Reproduced with permission from S DiMauro, E Schon: Mitochondrial respiratory-chain diseases. N Engl J Med 348:2656, 2003.)*

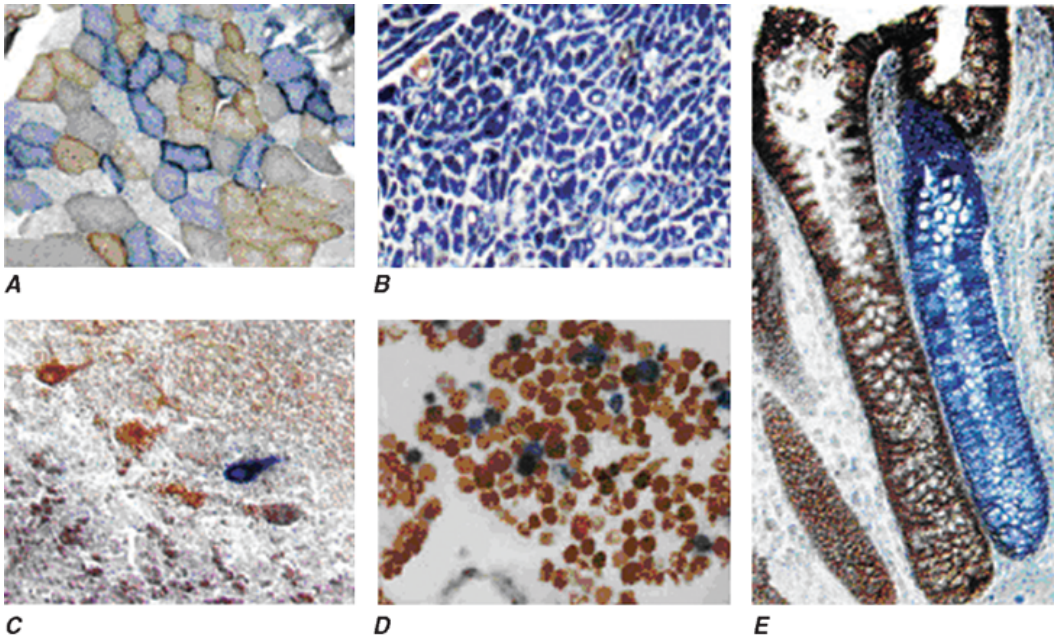


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Heteroplasmy can sometimes be elegantly demonstrated at the tissue level using histochemical staining for enzymes in the oxidative phosphorylation pathway, with a mosaic pattern indicating heterogeneity of the genotype for the coding region for the mtDNA-encoded enzyme. Complex II, CoQ, and cytochrome c are exclusively encoded by nuclear DNA. In contrast, complexes I, III, IV, and V contain at least some subunits encoded by mtDNA. Just 3 of the 13 subunits of the ETC complex IV enzyme, cytochrome c oxidase (COX), are encoded by mtDNA, and, therefore, this enzyme has the lowest threshold for dysfunction when a threshold level of mutated mtDNA is reached. Histochemical staining for COX activity in tissues of patients affected with heteroplasmic inherited mtDNA mutations (or with the somatic accumulation of mtDNA mutations, see below) can show a mosaic pattern of reduced histochemical staining in comparison with histochemical staining for the complex II enzyme, succinate dehydrogenase (SDH) (Fig. 472-5). Heteroplasmy can also be detected at the genetic level through direct Sanger-type mtDNA genotyping under special conditions, although clinically significant low levels of heteroplasmy can escape detection in genomic samples extracted from whole blood using conventional genotyping and sequencing techniques.

FIGURE 472-5

Cytochrome c oxidase (COX) deficiency in mitochondrial DNA (mtDNA)-associated disease. Transverse tissue sections that have been stained for COX and succinate dehydrogenase (SDH) activities sequentially, with COX-positive cells shown in *brown* and COX-deficient cells shown in *blue*. ***A.*** Skeletal muscle from a patient with a heteroplasmic mitochondrial tRNA point mutation. The section shows a typical "mosaic" pattern of COX activity, with many muscle fibers harboring levels of mutated mtDNA that are above the crucial threshold to produce a functional enzyme complex. ***B.*** Cardiac tissue (left ventricle) from a patient with a homoplasmic tRNA mutation that causes hypertrophic cardiomyopathy, which demonstrates an absence of COX in most cells. ***C.*** A section of cerebellum from a patient with mtDNA rearrangement that highlights the presence of COX-deficient neurons. ***D, E.*** Tissues that show COX deficiency due to clonal expansion of somatic mtDNA mutations within single cells—a phenomenon that is seen in both postmitotic cells (***D***; extraocular muscles) and rapidly dividing cells (***E***; colonic crypt) in aging humans. *(Reproduced with permission from R Taylor, D Turnbull: Mitochondrial DNA mutations in human disease. Nat Rev Genetics 6:389, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Next-generation sequencing (NGS) has dramatically improved the clinical genetic diagnostic evaluation of mitochondrial diseases at the level of both the nuclear genome and mtDNA. In the context of the larger nuclear genome, the ability of NGS techniques to dramatically increase the speed at which DNA can be sequenced at a fraction of the cost of conventional Sanger-type sequencing technology is particularly beneficial. Low sequencing costs and short turnaround time expedite "first-tier" screening of panels of hundreds of previously known or suspected mitochondrial disease genes or screening for the entire exome or genome in an attempt to identify novel genes and mutations affecting different patients or families. In the context of the mtDNA, NGS approaches now provide rapid and reliable detection of heteroplasmy in different affected tissues. Although Sanger sequencing allows for complete coverage of the mtDNA, it is limited by the lack of deep coverage and low sensitivity for heteroplasmy detection when levels are <50%. In contrast, NGS technology is an excellent tool for rapidly and accurately obtaining a patient's predominant mtDNA sequence and also lower frequency heteroplasmic variants and can typically detect mutant heteroplasmy <10%. Lower levels are often only clinically relevant if in the setting of a striking difference in heteroplasmy in different tissues. This capability to detect heteroplasmy at levels that assist in pointing to an mtDNA-based disease process emanates from deep coverage of the genome through multiple independent sequence reads. Accordingly, recent studies making use of NGS techniques have demonstrated sequence accuracy equivalent to Sanger-type sequencing, but also have uncovered heretofore unappreciated heteroplasmy rates ranging between 10 and 50% and detection of single-nucleotide heteroplasmy down to levels of <10%.

Clinically, the most striking overall characteristic of mitochondrial genetic disease is the phenotypic heterogeneity associated with mtDNA mutations. This extends to intrafamilial phenotypic heterogeneity for the same mtDNA pathogenic mutation and, conversely, to the overlap of phenotypic disease manifestations with distinct mutations. Thus, although fairly consistent and well-defined "classic" syndromes have been attributed to specific mutations, frequently "nonclassical" combinations of disease phenotypes ranging from isolated myopathy to extensive multisystem disease are often encountered, rendering genotype-phenotype correlation challenging. In both classical and nonclassical mtDNA disorders, there is often a clustering of some combination of abnormalities affecting the neurologic system (including optic nerve atrophy, pigment retinopathy, and sensorineural hearing loss), cardiac and skeletal muscle (including extraocular muscles), and endocrine and metabolic systems (including diabetes mellitus). Additional organ systems that may be affected include the hematopoietic, renal, hepatic, and gastrointestinal systems, although these are more frequently involved in infants and children. Disease-causing mtDNA coding region mutations can affect either one of the 13 protein encoding genes or one of the 24 protein synthetic genes. Clinical manifestations do not readily distinguish these two categories, although lactic acidosis and specific muscle pathologic findings (e.g., ragged-red and ragged-blue fibers, described immunohistochemical staining, paracrystalline inclusions on ultrastructure) tend to be more prominent in the latter. In all cases, either defective ATP production due to disturbances in the ETC or enhanced generation of ROS has been invoked as the mediating biochemical mechanism between mtDNA mutation and disease manifestation.

mtDNA DISEASE PRESENTATIONS

The clinical presentation of adult patients with mtDNA disease can be divided into three categories: (1) clinical features suggestive of mitochondrial disease (Table 472-2), but not a well-defined classic syndrome; (2) classic mtDNA syndromes; and (3) clinical presentation confined to one organ system (e.g., isolated sensorineural deafness, cardiomyopathy, or diabetes mellitus).

TABLE 472-2

Common Features of mtDNA-Associated Diseases in Adults

Neurologic: stroke, epilepsy, migraine headache, peripheral neuropathy, ataxia, dystonia, myoclonus, cranial neuropathy (optic atrophy, sensorineural deafness, dysphagia, dysphasia)
Skeletal myopathy: ophthalmoplegia, exercise intolerance, myalgia, weakness
Cardiac: conduction block, cardiomyopathy
Respiratory: hypoventilation, aspiration pneumonitis
Endocrine: diabetes mellitus, premature ovarian failure, hypothyroidism, hypoparathyroidism
Ophthalmologic: cataracts, pigment retinopathy, neurologic and myopathic (optic atrophy, ophthalmoplegia)

Table 472-3 provides a summary of eight illustrative classic mtDNA syndromes or disorders that affect adult patients and highlights some of the most interesting features of mtDNA disease in terms of molecular pathogenesis, inheritance, and clinical presentation. The first five of these syndromes result from heritable point mutations in either protein-encoding or protein synthetic mtDNA genes; the other three result from rearrangements or deletions that usually do not involve the germline.

TABLE 472-3

Mitochondrial Diseases Due to mtDNA Point Mutations and Large-Scale Rearrangements

Disease	Phenotype	Most Frequent mtDNA Mutations	Homoplasmic (Usually)	Maternal
NARP, Leigh disease	Loss of central vision leading to blindness in young adult life	m.1778G>A, m.14484T>C, m.3460G>A	Heteroplasmic	Maternal
MELAS	*M*itochondrial *e*ncephalomyopathy, *l*actic*a*cidosis, and *s*troke-like episodes; may manifest only as diabetes mellitus	Point mutation in $tRNA^{leu}$	Heteroplasmic	Maternal
MERRF	*M*yoclonic *e*pilepsy, *r*agged *r*ed *f*ibers in muscle, ataxia, increased CSF protein, sensorineural deafness, dementia	Point mutation in $tRNA^{lys}$	Heteroplasmic	Maternal
Deafness	Progressive sensorineural deafness, often induced by aminoglycoside antibiotics	m.1555A>G mutation in 12S rRNA	Homoplasmic	Maternal
	Nonsyndromic sensorineural deafness	m.7445A>G mutation in 12S rRNA	Homoplasmic	Maternal
Chronic progressive external ophthalmoplegia (PEO)	Late-onset bilateral ptosis and ophthalmoplegia, proximal muscle weakness, and exercise intolerance	Single deletions or duplications	Heteroplasmic	Mostly sporadic, somatic mutations
Pearson syndrome	Pancreatic insufficiency, pancytopenia, lactic acidosis	Large deletion	Heteroplasmic	Sporadic, somatic mutations
Kearns-Sayre syndrome (KSS)	External ophthalmoplegia, heart block, retinal pigmentation, ataxia	The 5-kb "common deletion"	Heteroplasmic	Sporadic, somatic mutations

Abbreviations: CSF, cerebrospinal fluid; NARP, neuropathy, ataxia, and retinitis pigmentosa.

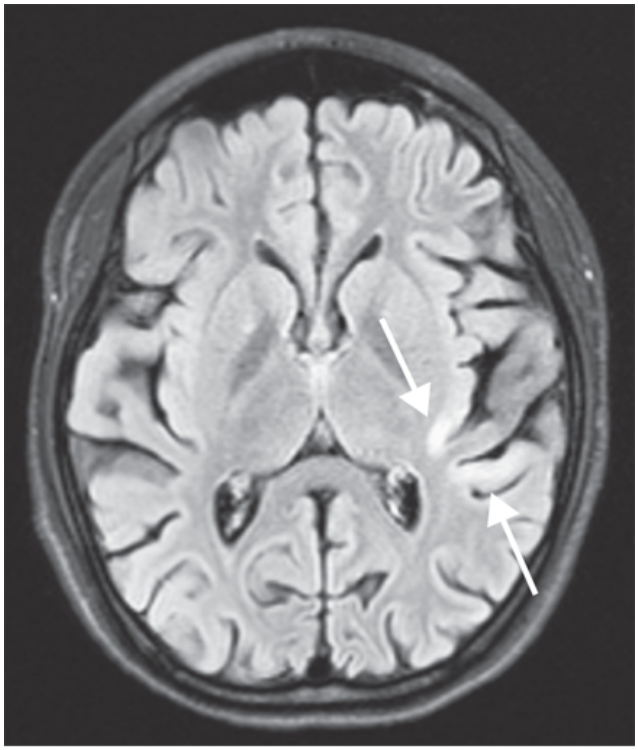
LHON is a common cause of maternally inherited visual failure. LHON typically presents during young adulthood with subacute painless loss of vision in one eye, with symptoms developing in the other eye 6–12 weeks after the initial onset. In some instances, cerebellar ataxia, peripheral neuropathy, and cardiac conduction defects are observed. In >95% of cases, LHON is due to one of the three homoplasmic point mutations of mtDNA that affect genes encoding different subunits of complex I of the mitochondrial ETC; however, not all individuals who inherit a primary LHON mtDNA mutation develop optic neuropathy, and males are four to five times more likely than females to be affected, indicating that additional environmental (e.g., tobacco exposure) or additional independent genetic factors are important in the etiology of the disorder. Both the nuclear and mitochondrial genomic backgrounds modify disease penetrance. Indeed, a region of the X chromosome containing a high-risk haplotype for LHON has been identified, supporting the formulation that nuclear genes act as modifiers and affording an explanation for the male prevalence of LHON. This haplotype can be used in predictive genomic testing and prenatal screening for this disease. In contrast to the other classic mtDNA disorders, it is of interest that patients with this syndrome are often homoplasmic for the disease-causing mutation. The somewhat later onset in young adulthood and modifying effect of protective background nuclear genomic haplotypes may have enabled homoplasmic pathogenic mutations to have escaped evolutionary censoring.

Mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes (MELAS) is a multisystem disorder with a typical onset between 2 and 10 years of age. Following normal early psychomotor development, the most common initial symptoms are seizures, recurrent headaches, anorexia, and recurrent vomiting. Exercise intolerance or proximal limb weakness can be the initial manifestation, followed by generalized tonic–clonic seizures. Short stature is common. Seizures are often associated with stroke-like episodes of transient hemiparesis or cortical blindness that may produce recurrent encephalopathy with impaired consciousness. It is often not possible to determine if the encephalopathy is due to refractory seizures or should be

attributed to an independent effect. The cumulative residual effects of the stroke-like episodes gradually impair motor abilities, vision, and cognition, often by adolescence or young adulthood. Sensorineural hearing loss adds to the progressive decline of these individuals. A plethora of less common symptoms have been described including myoclonus, ataxia, episodic coma, optic atrophy, cardiomyopathy, pigmentary retinopathy, ophthalmoplegia, diabetes mellitus, hirsutism, gastrointestinal dysmotility, and nephropathy. The typical age of death ranges from 10 to 35 years, but some individuals live into their sixth decade. Intercurrent infections or intestinal obstructions are often the terminal events. It is not atypical for some family members to have much less severe, or later onset illness, presumably because of a lessor mutation load, and "MELAS" is not used as a diagnosis for these restricted phenotypes. This creates somewhat of a disconnect between the genotype for MELAS (most commonly the m.3243A>G mutation), and a diverse phenotype, which includes the syndrome MELAS, as well as a syndrome of high-frequency hearing loss and diabetes with onset later in life, as well as many other phenotypes between these two extreme syndromes. Certain other mtDNA mutations can also cause such patterns of diverse phenotypic expression. Laboratory investigation commonly demonstrates elevated lactate concentrations at rest with excessive increase after moderate exercise. Brain imaging during stroke-like episodes shows areas of involvement on T2- or FLAIR sequences, with decreased signal on perfusion-weighted sequences, which typically involve the posterior cerebrum and not conforming to the distribution of major arteries. These MRI abnormalities may be temporary or evolve to subsequent atrophy **(Fig. 472-6)**. Electrocardiography (ECG) may show evidence of cardiomyopathy, preexcitation, or incomplete heart block. Electromyography and nerve conduction studies are consistent with a myopathic process, without or with coexisting axonal and sensory neuropathic findings. Muscle biopsy typically shows ragged red fibers with the modified Gomori trichrome stain or "ragged blue fibers" with the SDH histochemical stain, resulting from the hyperintense reaction. The diagnosis of MELAS is based on a combination of clinical findings and molecular genetic testing. Mutations in the mtDNA gene *MT-TL1* encoding $tRNA^{leu}$ are causative. The most common mutation, present in ~80% of individuals with typical clinical findings, is an A-to-G transition at nucleotide 3243 (m.3243A>G). Mutations can usually be detected in mtDNA from leukocytes in individuals with typical MELAS; however, the occurrence of heteroplasmy can result in varying tissue distribution of mutated mtDNA. In the absence of specific treatment, various manifestations of MELAS are treated according to standard modalities for prevention, surveillance, and treatment. Recent developments in therapy are described below.

FIGURE 472-6

A 15-year-old girl with MELAS due to m.A3243G ($tRNA^{Leu(UUR)}$), 85% mutant heteroplasmy, presenting at age 5 with focal motor seizures, ataxia and short stature, with episodes of acute language and motor dysfunction and progress cognitive impairment. The FLAIR MRI shows increased signal intensity (white arrows) in the left temporal-parietal region in addition to global mild volume loss (increased extra-axial CSF spaces).



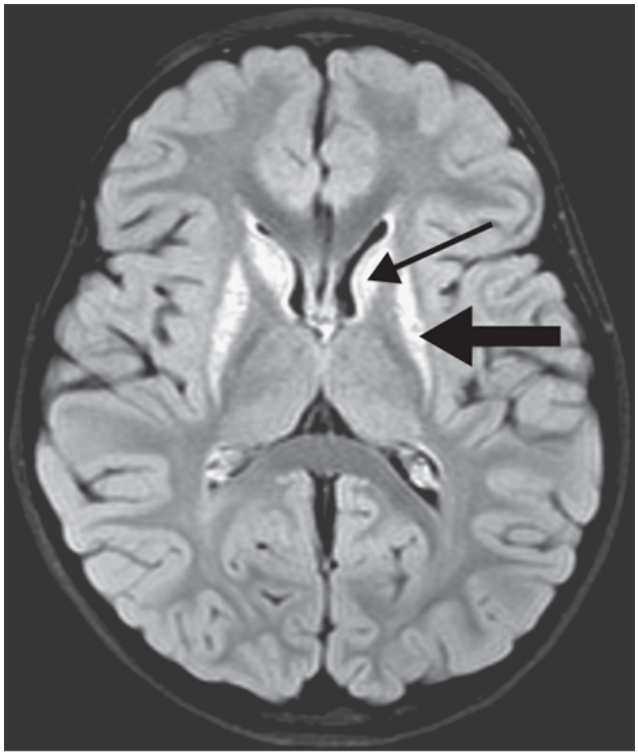
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Myoclonus epilepsy with ragged red fiber (MERRF) is a multisystem disorder characterized by myoclonus, seizures, ataxia, and myopathy with ragged red fibers. Hearing loss, exercise intolerance, neuropathy, and short stature are often present. Cerebrospinal fluid (CSF) analysis reveals an elevated protein content. Almost all MERRF patients have a mutation in the mtDNA $tRNA^{lys}$ gene, and the m.8344A>G mutation in the mtDNA gene encoding the lysine amino acid tRNA is responsible for 80–90% of MERRF cases.

Neuropathy, ataxia, and retinitis pigmentosa (NARP) is characterized by moderate diffuse cerebral and cerebellar atrophy and symmetric lesions of the basal ganglia on magnetic resonance imaging (MRI: **Figs. 472-7 and 472-8**). A heteroplasmic m.8993T>G mutation in the ATPase 6 subunit gene has been identified as causative. Ragged red fibers are not observed in muscle biopsy. When >95% of mtDNA molecules are mutant, a more severe clinical, neuroradiologic, and neuropathologic picture (Leigh syndrome) emerges. Not uncommonly, an infant is diagnosed with Leigh syndrome due to the m.8993T>G mutation and not until several years later will the mother present with symptoms of NARP; a situation that highlights the concept of a higher threshold for lower levels of tissue heteroplasmy.

FIGURE 472-7

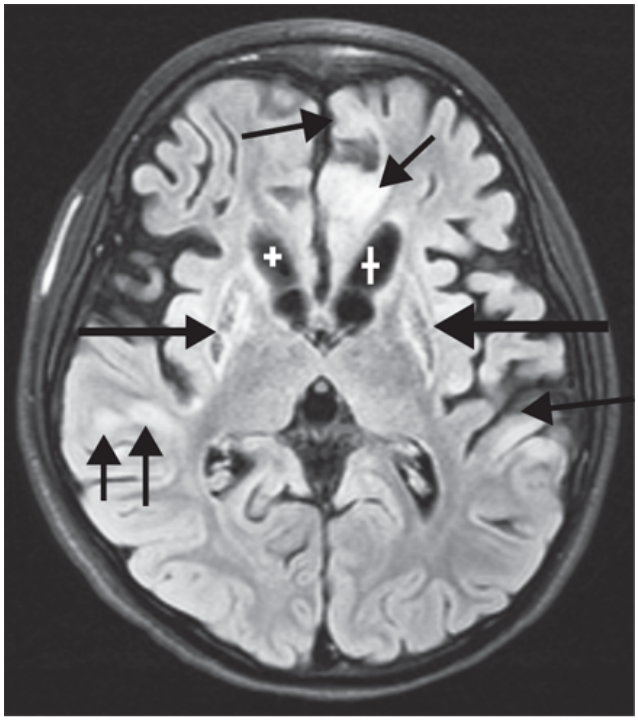
A 9-year-old girl with Leigh syndrome due to m.T8993G (ATPase subunit 6), 99% heteroplasmy, presenting at age 14 months with a motor delay and underwent an MRI at 24 months, at which time she had just begun to walk. She has moderate cognitive impairment, arm chorea, and distal leg dystonia. The FLAIR MRI shows symmetric bilateral increased signal in the caudate nuclei (*thin arrow*) and putamen (*thick arrow*); only left-sided lesions indicated with arrows.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 472-8

A 12-year-old boy with Leigh syndrome due to m.T10191C (ND3 gene, complex I), heteroplasmy percentage not determined, presenting with infantile spasms at 8 months of life. He responded well to Adrenocorticotropic Hormone (ACTH) and his MRI and development were normal until 30 months when he developed dystonia and progressive medically intractable epilepsy. The FLAIR MRI at 6 years of life shows global atrophy with large extraaxial CSF spaces, increased signal intensity in the cortex (*thin arrows*), necrotic bilaterally symmetric lesions in the putamina, and enlarged lateral ventricles due to loss of bilateral caudate nuclei volume (*stars*).



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Point mutations in the mtDNA gene encoding the 12S rRNA (m.A1555G) result in heritable nonsyndromic hearing loss. One such mutation causes heritable ototoxic susceptibility to aminoglycoside antibiotics, which opens a pathway for a simple pharmacogenetic test in the appropriate clinical settings. This is an example of an eco-genetic disorder in that most people with this mutation do not develop any symptoms until exposed to an external agent.

KSS, sporadic progressive external ophthalmoplegia (PEO), and Pearson syndrome are three disease phenotypes caused by large-scale mtDNA rearrangements including partial deletions or partial duplication. The majority of single large-scale rearrangements of mtDNA are thought to result from clonal amplification of a single sporadic mutational event, occurring in the maternal oocyte or during early embryonic development. The typical mtDNA deletion is specifically at nucleotide 4977, accounting for most KSS and PEO of mtDNA deletion origin. Because germline involvement is rare, most cases are sporadic rather than inherited. KSS is characterized by the triad of onset before age 20, chronic PEO, and pigmentary retinopathy. Cerebellar syndrome, heart block, increased CSF protein content, diabetes mellitus, and short stature are also part of the syndrome. Single deletions/duplication can also result in milder phenotypes such as PEO, characterized by late-onset PEO, proximal myopathy, and exercise intolerance. In both KSS and PEO, diabetes mellitus and hearing loss are frequent accompaniments. Pearson syndrome is also characterized by infantile onset of a sideroblastic anemia accompanied by lactic acidosis and failure to thrive caused in part by exocrine pancreatic insufficiency. If the child survives, the manifestations appear phenotypically similar to that of severe KSS with myopathy, PEO, encephalopathy, and cardiomyopathy. Pearson syndrome is generally caused by large-scale sporadic deletion of several mtDNA genes that differ from the common deletion seen in KSS. Typically, the deletion size is larger in Pearson syndrome than in KSS or PEO, but this is not always the case.

Two important dilemmas in classic mtDNA disease have benefited from recent important research insights. The first relates to the greater involvement of neuronal, muscular, renal, hepatic, and pancreatic manifestations in mtDNA disease in these syndromes. This observation has appropriately been mostly attributed to the high energy utilization of the involved tissues and organ systems and, hence, greater dependency on mitochondrial ETC integrity and health. However, because mutations are stochastic events, mitochondrial mutations should occur in any organ during embryogenesis and development. Recently, additional explanations have been suggested based on studies of the common m.3243A>G transition. The proportion of this mutation in peripheral blood cells was shown to decrease exponentially with age. A selective process acting at the stem cell level with a strong bias against the mutated form would have its greatest effect to reduce the mutant mtDNA only in highly proliferating cells, such as those derived from the hematopoietic system. Tissues and organs with lower cell turnover, such as those involved with mtDNA mutations, would not benefit from this effect and, thus, would be the most affected.

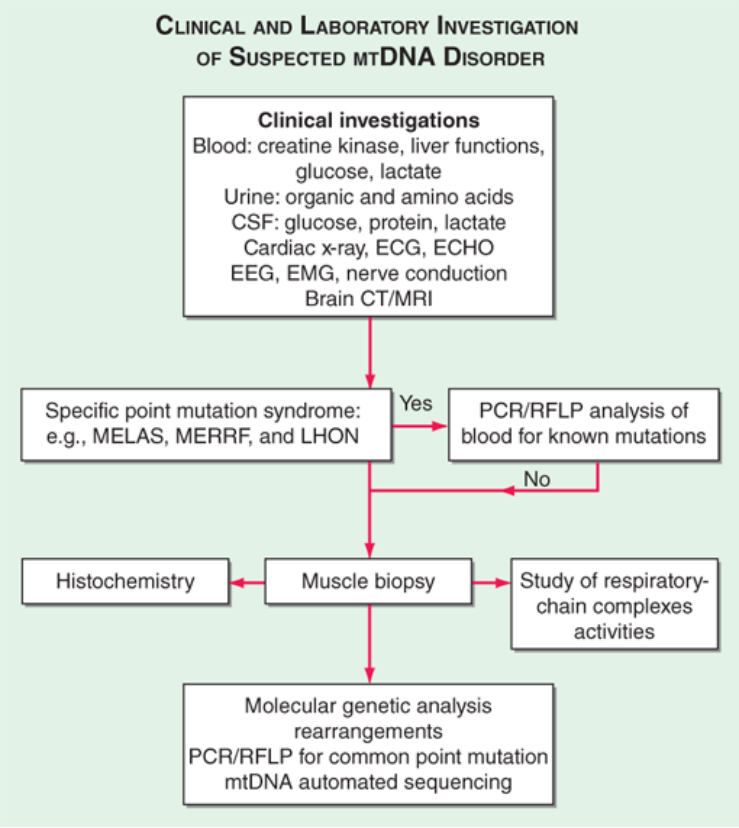
The other dilemma arises from the observation that only a subset of mtDNA mutations accounts for the majority of the familial mtDNA diseases. The random occurrence of mutations in the mtDNA sequence should yield a more uniform distribution of disease-causing mutations. However, recent studies using the introduction of one severe and one mild point mutation into the female germline of experimental animals demonstrated selective elimination during oogenesis of the severe mutation and selective retention of the milder mutation, with the emergence of mitochondrial disease in offspring after multiple generations. Thus, oogenesis itself can act as an “evolutionary” filter for mtDNA disease.

THE INVESTIGATION OF SUSPECTED mtDNA DISEASE

The clinical presentations of classic syndromes, groupings of disease manifestations in multiple organ systems, or unexplained isolated presentations of one of the disease features of a classic mtDNA syndrome should prompt a systematic clinical investigation as outlined in **Fig. 472-9**. Indeed, mitochondrial disease should be considered in the differential diagnosis of any progressive multisystem disorder. Despite the centrality of disruptive oxidative phosphorylation, an elevated blood lactate level is neither specific nor sensitive, because there are many causes of blood lactic acidosis, and many patients with mtDNA defects presenting in adulthood have normal blood lactate. An elevated CSF lactate is a more specific test for mitochondrial disease if there is central nervous system involvement. The serum creatine kinase may be elevated but is often normal, even in the presence of a proximal myopathy. Recently, testing for elevated levels of Growth Differentiating Factor 15 (GDF15) has shown a high degree of sensitivity and specificity in those with a mitochondrial myopathy, but it is not known yet if the degree of elevation for an individual patient reflects the severity of the illness or is in any way a marker of disease activity. Urinary organic and amino acids may also be abnormal, reflecting metabolic and kidney proximal tubule dysfunction. Every patient with seizures, episodes of confusion or atypical behavioral changes, or cognitive decline should have an electroencephalogram. A brain computed tomography (CT) scan may show calcified basal ganglia or bilateral hypodense regions with cortical atrophy. MRI is indicated in patients with brainstem signs or stroke-like episodes.

FIGURE 472-9

Clinical and laboratory investigation of a suspected mitochondrial DNA (mtDNA) disorder. CSF, cerebrospinal fluid; CT, computed tomography; ECG, electrocardiogram; ECHO, echocardiography; EEG, electroencephalogram; EMG, electromyogram; LHON, Leber's hereditary optic neuropathy; MELAS, mitochondrial encephalomyopathy, lactic acidosis, and stoke-like episodes; MERFF, myoclonic epilepsy with ragged red fibers; MRI, magnetic resonance imaging; PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

For some mitochondrial diseases, it is possible to obtain an accurate diagnosis with a simple molecular genetic screen. For examples, 95% of patients with LHON harbor one of the three mtDNA point mutations (m.11778A>G, m.A3460A>G, or m.14484T>C). These patients have very high levels of mutated mtDNA in peripheral blood cells, and, therefore, it is appropriate to send a blood sample for molecular genetic analysis by polymerase chain reaction (PCR) or restriction fragment length polymorphism (RFLP). The same is true for most MERRF patients who harbor a point mutation in the lysine tRNA gene at position 8344. In contrast, patients with the m.3243A>G MELAS mutation often have low levels of mutated mtDNA in blood. If clinical suspicion is strong enough to warrant peripheral blood testing, then patients with a negative result should be repeated in a saliva sample, or investigated further by performing a skeletal muscle biopsy.

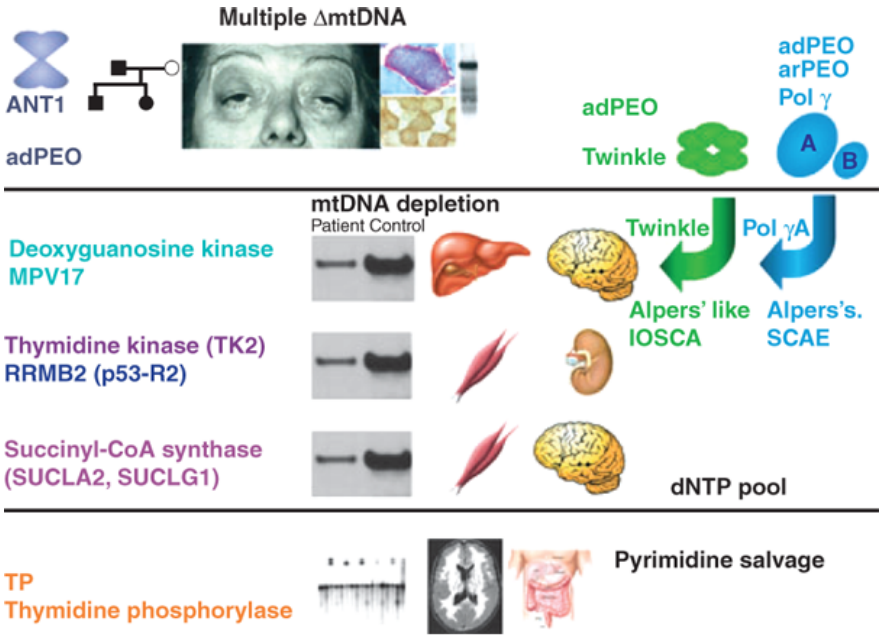
Muscle biopsy histochemical analysis had been the historical cornerstone for investigation of patients with suspected mitochondrial disease. Histochemical analysis may show subsarcolemmal accumulation of mitochondria with the appearance of ragged red fibers, especially in those with

mtDNA mutations affecting the tRNA and rRNA genes. Electron microscopy might show abnormal mitochondria with paracrystalline inclusions. Muscle histochemistry may show COX-deficient fibers, which indicate mitochondrial dysfunction (Fig. 472-5). Respiratory chain complex assays may also show reduced enzyme function. If enzymatic or polarographic data are used to aid in the confirmation of diagnosis, a standard method of analysis should be employed. Either of these two abnormalities, within the exact context of established peer-reviewed criteria may confirm the presence of a mitochondrial disease, to be followed by an in-depth molecular genetic analysis. In some centers, genetic testing may immediately follow establishment of a clinical phenotype and screening biochemical labs, which may obviate the need for invasive testing.

Recent evidence has provided important insights into the importance of nuclear-mtDNA genomic cross-talk and has provided a descriptive framework for classifying and understanding disorders that emanate from perturbations in this cross-talk. Although not strictly considered as mtDNA genetic disorders, manifestations do overlap those highlighted above **(Fig. 472-10)**.

FIGURE 472-10

Disorders associated with perturbations in nuclear-mitochondrial genomic cross-talk. Clinical features and genes associated with multiple mitochondrial DNA (mtDNA) deletions, mtDNA depletion, and mitochondrial neurogastrointestinal encephalomyopathy syndromes. ANT, adenine nucleotide translocators; adPEO, autosomal dominant progressive external ophthalmoplegia; arPEO, autosomal recessive progressive external ophthalmoplegia; IOSCA, infantile-onset spinocerebellar ataxia; SCAE, spinocerebellar ataxia and epilepsy. *(Reproduced with permission from A Spinazzola, M Zeviani: Disorders from perturbations of nuclear-mitochondrial intergenomic cross-talk. J Intern Med 265:174, 2009.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

IMPACT OF HOMOPLASMIC SEQUENCE VARIATION ON HERITABLE TRAITS AND DISEASE

The relationship among the degree of heteroplasmy, tissue distribution of the mutant mtDNA, and disease phenotype simplifies inference of a clear causative relationship between heteroplasmic mutation and disease. With the exception of certain mutations (e.g., those causing most cases of LHON), drift to homoplasmy of such mutations would be precluded normally by the severity of impaired oxidative phosphorylation and the consequent reduction in reproductive fitness. Therefore, sequence variants that have reached homoplasmy should be neutral in terms of human evolution and, hence, useful only for tracing human evolution, demography, and migration, as described above. One important exception is in the case of one or more of the homoplasmic population-level variants, which designate the mtDNA haplogroup J, and the interaction with the mtDNA mutations causing LHON. Reduced disease predilection suggests that one or more of the ancient sequence variants designating mtDNA haplogroup J appear to attenuate predisposition to degenerative disease, in the face of other risk factors. Whether or not additional epistatic interactions between population-level mtDNA haplotypes and common health conditions will be found remains to be determined. If such influences do exist, then they are more likely to be relevant to health conditions in the postreproductive age groups, wherein evolutionary filters would not have had the opportunity to censor deleterious effects and interactions and wherein the effects of oxidative stress may play a role. Although much has been written about the possible associations of population-level common mtDNA variants and human health and disease phenotypes or adaptation to different environmental influences (e.g., climate), a word of caution is in order.

Many studies that purport to show such associations with phenotypes such as longevity, athletic performance, and metabolic and neurodegenerative disease are limited by small sample sizes, possible genotyping inaccuracies, and the possibility of population stratification or ethnic ancestry bias.

Because mtDNA haplogroups are so prominently partitioned along phylogeographic lines, it is difficult to exclude the possibility that a haplogroup for which an association has been found is simply a marker for differences in populations with a societal or environmental difference or with different allele frequencies at other genomic loci, which are actually causally related to the heritable trait or disease of interest. The difficulty in generating cellular or animal models to test the functional influence of homoplasmic sequence variants (as a result of mtDNA polyploidy) further compounds the challenge. The most likely formulation is that the risk conferred by different mtDNA haplogroup-defining homoplasmic mutations for common diseases depends on the concomitant nuclear genomic background, together with environmental influences. Progress in minimizing potentially misleading associations in mtDNA heritable trait and disease studies should include ensuring adequate sample size taken from a large sample recruitment base, using carefully matched controls and population structure determination, and performing analysis that takes into account epistatic interactions with other genomic loci and environmental factors.

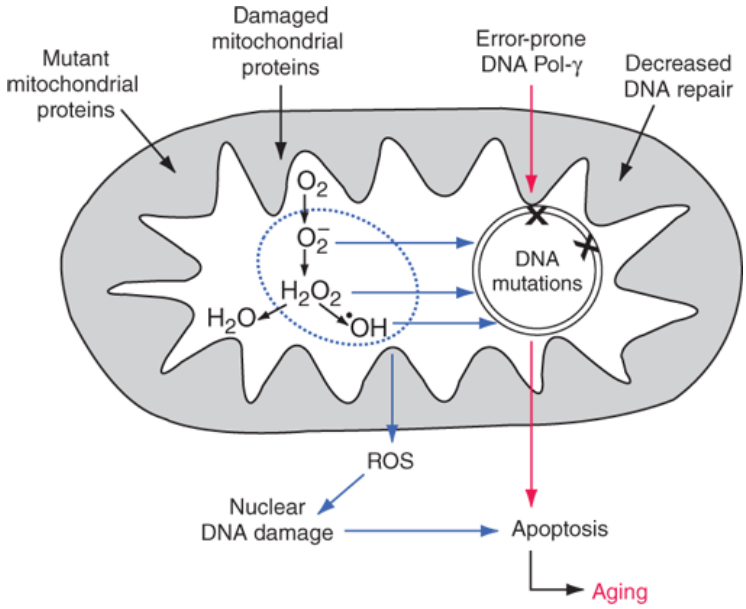
IMPACT OF ACQUIRED SOMATIC mtDNA MUTATION ON HUMAN HEALTH AND DISEASE

Studies on aging humans and animals have shown a potentially important correlation of age with the accumulation of heterogeneous mtDNA mutations, especially in those organ systems that undergo the most prominent age-related degenerative tissue phenotype. Sequencing of PCR-amplified single mtDNA molecules has demonstrated an average of two to three point mutations per molecule in elderly subjects when compared with younger ones. Point mutations observed include those responsible for known heritable heteroplasmic mtDNA disorders, such as the m.3344A>G and m.3243A>G mutations responsible for the MERRF and MELAS syndromes, respectively. However, the cumulative burden of these acquired somatic point mutations with age was observed to remain well below the threshold expected for phenotypic expression (<2%). Point mutations at other sites not normally involved in inherited mtDNA disorders have also been shown to accumulate to much higher levels in some tissues of elderly individuals, with the description of tissue-specific “hot spots” for acquired somatic mtDNA point mutations. Along the same lines, an age-associated and tissue-specific accumulation of mtDNA deletions has been observed, including deletions involved in known heritable mtDNA disorders, as well as others. The accumulation of functional mtDNA deletions in a given tissue is expected to be associated with mitochondrial dysfunction, as reflected in an age-associated patchy and reduced COX activity on histochemical staining, especially in skeletal and cardiac muscle and brain. A particularly well-studied and potentially important example is the accumulation of mtDNA deletions and COX deficiency observed in neurons of the substantia nigra in Parkinson’s disease patients.

The progressive accumulation of ROS has been proposed as the key factor connecting mtDNA mutations with aging and age-related disease pathogenesis **(Fig. 472-11)**. As noted above, ROS are a by-product of oxidative phosphorylation and are removed by detoxifying antioxidants into less harmful moieties; however, exaggerated production of ROS or impaired removal results in their accumulation. One of the main targets for ROS-mediated injury is DNA, and mtDNA is particularly vulnerable because of its proximity to the origin of free radical production, the lack of protective histones, and less efficient injury repair systems compared with nuclear DNA. In turn, accumulation of mtDNA mutations results in inefficient oxidative phosphorylation, with the potential for excessive production of ROS, generating a “vicious cycle” of cumulative mtDNA damage. Indeed, measurement of the oxidative stress biomarker 8-hydroxy-2-deoxyguanosine has been used to measure age-dependent increases in mtDNA oxidative damage at a rate exceeding that of nuclear DNA. It should be noted that mtDNA mutations can potentially occur in postmitotic cells as well, because mtDNA replication is not synchronized with the cell cycle. Two other proposed links between mtDNA mutation and aging, besides ROS-mediated tissue injury, are the perturbations in efficiency of oxidative phosphorylation with disturbed cellular aerobic function and perturbations in apoptotic pathways, whose execution steps involve mitochondrial activity.

FIGURE 472-11

Multiple pathways of mitochondrial DNA (mtDNA) damage and aging. Multiple factors may impinge on the integrity of mitochondria that lead to loss of cell function, apoptosis, and aging. The classic pathway is indicated with *blue arrows*; the generation of reactive oxygen species (ROS; superoxide anion, hydrogen peroxide, and hydroxyl radicals), as a by-product of mitochondrial oxidative phosphorylation, results in damage to mitochondrial macromolecules, including the mtDNA, with the latter leading to deleterious mutations. When these factors damage the mitochondrial energy-generating apparatus beyond a functional threshold, proteins are released from the mitochondria that activate the caspase pathway, leading to apoptosis, cell death, and aging. *(Reproduced with permission from L Loeb et al: The mitochondrial theory of aging and its relationship to reactive oxygen species damage and somatic mtDNA mutations. Proc Natl Acad Sci USA 102:18769, 2005.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Genetic intervention studies in animal models have sought to clarify the potential causative relationship between acquired somatic mtDNA mutation and the aging phenotype, and the role of ROS in particular. Replication of the mitochondrial genome is mediated by the activity of the nuclear-encoded *POLG*. A transgenic homozygous mouse knock-in mutation of this gene renders the polymerase enzyme deficient in proofreading and results in a threefold to fivefold increase in mtDNA mutation rate. Such mice develop a premature aging phenotype, which includes subcutaneous lipoatrophy, alopecia, kyphonia, and weight loss with premature death. Although the finding of increased mtDNA mutation and mitochondrial dysfunction with age has been solidly established, the causative role and specific contribution of mitochondrial ROS to aging and age-related disease in humans has yet to be proved. Similarly, although many tumors display higher levels of heterogeneous mtDNA mutations, a causal relationship to tumorigenesis has not been proved.

Besides the age-dependent acquired accumulation in somatic cells of heterogeneous point mutations and deletions, a quite different effect of nonheritable and acquired mtDNA mutation has been described affecting tissue stem cells. In particular, disease phenotypes attributed to acquired mtDNA mutation have been observed in sporadic and apparently nonfamilial cases involving a single individual or even tissue, usually skeletal muscle. The presentation consists of decreased exercise tolerance and myalgias, sometimes progressing to rhabdomyolysis. As in the case of the sporadic, heteroplasmic, large-scale deletion, classic syndromes of chronic PEO, Pearson syndrome, and KSS, the absence of a maternal inheritance pattern, together with the finding of limited tissue distribution, suggest a molecular pathogenic mechanism emanating from mutations arising de novo in muscle stem cells after germline differentiation (somatic mutations that are not sporadic and occur in tissue-specific stem cells during fetal development or in the postnatal maintenance or postinjury repair stage). Such mutations would be expected to be propagated only within the progeny of that stem cell and affect a particular tissue within a given individual, without evidence of heritability.

PROSPECTS FOR CLINICAL MANAGEMENT OF mtDNA DISEASE

TREATMENT OF mtDNA DISORDERS

No specific curative treatment for mtDNA disorders is currently available; therefore, the management of mitochondrial disease is largely supportive. Management issues may include early diagnosis and treatment of epilepsy, gastrointestinal dysfunction, diabetes mellitus, cardiac pacing, ptosis correction, and intraocular lens replacement for cataracts. Less specific interventions in the case of other disorders involve combined treatment strategies including dietary intervention and removal of toxic metabolites. Cofactors and vitamin supplements are widely used in the treatment of diseases of mitochondrial oxidative phosphorylation, although there is little evidence, apart from anecdotal reports, to support their use. This includes administration of artificial electron acceptors, including vitamin K_3, vitamin C, and ubiquinone (coenzyme Q10); administration of cofactors (coenzymes) including riboflavin, carnitine, and creatine; and use of oxygen radical scavengers, such as vitamin E, copper, selenium, ubiquinone, and idebenone. Drugs that could interfere with mitochondrial function, such as the anesthetic agent propofol, barbiturates, and high doses of valproate, should be avoided. The use of valproate in patients with pathogenic mutations in *POLG* and possibly other mutations affecting mtDNA stability and replication are especially contraindicated. Supplementation with the nitric oxide synthase substrate, L-arginine, and more recently L-citrulline has been advocated as a vasodilator treatment during stroke-like episodes as well as for chronic management in patients with MELAS. Open label studies demonstrate that levoarginine and levocitrulline may be helpful in reducing the stroke-like symptoms in MELAS, but may have serious side effects. As CSF folate deficiency has been reported in some cases of mitochondrial disease, this can be treated with folinic acid.

The physician should also be familiar with environmental interactions, such as the strong and consistent association between visual loss in LHON and smoking or ethanol consumption. A clinical penetrance of 93% was found in men who smoked. Asymptomatic carriers of an LHON mtDNA mutation should, therefore, be strongly advised not to smoke and to moderate their alcohol intake. Although not a cure, these interventions might stave off the devastating clinical manifestations of the LHON mutation. Another example is strict avoidance of aminoglycosides in the familial syndrome of ototoxic susceptibility to aminoglycosides in the presence of the mtDNA m.1555A>G mutation of the 12SrRNA encoding gene.

Clinical trials using novel agents are currently being conducted and analyzed. These include α-tocotrienol (EPI-743, BioElectron Technology Corporation), cysteamine bitartrate (RP-103, Horizon Pharma), omaveloxolone (RTA-408, Reata Pharma), and elamipretide (Stealth Biotheraputics). In an open label study of α-tocotrienol used to treat 10 children with Leigh syndrome, there were improvements in the primary endpoints including the Newcastle Pediatric Mitochondrial Diseases Scale, the Gross Motor Function Measure, and the PedsQL Neuromuscular Module. Ongoing studies continue for α-tocotrienol in children, and both omaveloxolone and elamipritide in adults with primary mitochondrial myopathy.

GENETIC COUNSELING, PRENATAL DIAGNOSIS, AND PGD IN mtDNA DISORDERS

The provision of accurate genetic counseling and reproductive options to families with mtDNA mutations is challenging due to the unique genetic features of mtDNA inheritance that distinguish it from Mendelian genetics. mtDNA defects are transmitted by maternal inheritance. mtDNA de novo mutations are often large deletions, affect one family member, and usually represent no significant risk to other members of the family. In contrast, mtDNA point mutations or duplications can be transmitted down the maternal line. Accordingly, the father of an affected individual has no risk of harboring the disease-causing mutation, and a male cannot transmit the mtDNA mutation to his offspring. In contrast, the mother of an affected individual usually harbors the same mutation but might be completely asymptomatic. This wide phenotypic variability is primarily related to the phenomena of heteroplasmy and the mutation load carried by different members of the same family. Consequently, a symptomatic or asymptomatic female harboring a disease-causing mutation in a heteroplasmic state will transmit to her offspring variable amounts of the mutant mtDNA molecules. The offspring will be symptomatic or asymptomatic primarily according to the mutant load transmitted via the oocyte and, to some extent, subsequent mitotic segregation during development. Interactions with the mtDNA haplotype background or nuclear human genome (as in the case of LHON) serve as an additional important determinant of disease penetrance. Because the severity of the disease phenotype associated with the heteroplasmic mutation load is a function of the stochastic differential segregation and copy number of mutant mtDNA during the oogenesis bottleneck and, subsequently, following tissue and organ development in the offspring, it is rarely predictable with any degree of accuracy. For this reason, prenatal diagnosis (PND) and PGD techniques that have evolved into integral and well-accepted standards of practice are severely hampered in the case of mtDNA-related diseases.

The value of PND and PGD is limited, partly due to the absence of data on the rules that govern the segregation of wild-type and mutant mtDNA species (heteroplasmy) among tissue in the developing embryo. Three factors are required to ensure the reliability of PND and PGD: (1) a close correlation between the mutant load and the disease severity, (2) a uniform distribution of mutant load among tissues, and (3) no major change in mutant load with time. These criteria are suggested to be fulfilled for the NARP m.8993T>G mutation but do not seem to apply to other mtDNA disorders. In fact, the level of mutant mtDNA in a chorionic villous or amniotic fluid sample may be very different from the level in the fetus, and it would be difficult to deduce whether the mutational load in the prenatal samples provides clinically useful information regarding the postnatal and adult state.

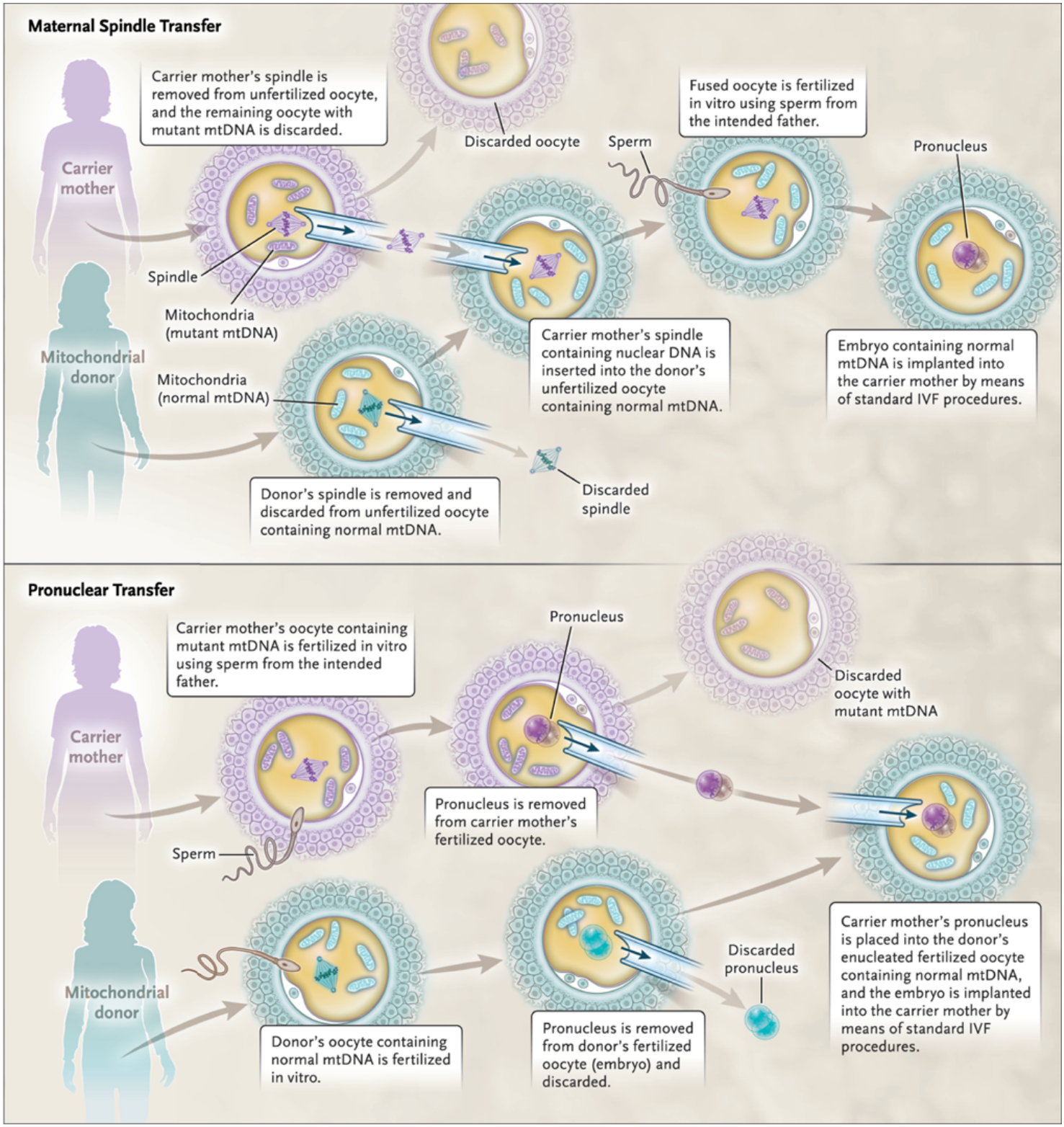
PREVENTION OF MITOCHONDRIAL DISEASE INHERITANCE BY ASSISTED REPRODUCTIVE TECHNOLOGIES

Because the treatment options for patients with mitochondrial disease are rather limited, with no current U.S. Food and Drug Administration (FDA) approved therapies for established mitochondrial DNA disease, preventive interventions that eliminate the likelihood of transmission of affected mtDNA into offspring are desirable. The poor reliability of prenatal and preimplantation approaches in predicting mitochondrial DNA disease has resulted in the search for alternative preventive approaches. The common purpose underlying various emerging approaches is to reduce mutant heteroplasmy levels to a level below a pathogenic threshold. This is based on the observed relationship between heteroplasmy and disease inheritance patterns, which indicates that even a small increase in copy number of nonmutant mtDNA molecules in the fertilized egg can exceed the threshold required to ameliorate serious clinical disease. Use of gene editing, with Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) technology for example, to shift the heteroplasmy load in affected tissues will require future development of corrective gene delivery techniques. Likewise, induced pluripotent cell technology has not yet met with widespread success in the preclinical research setting. This has prompted the application of Mitochondrial Replacement Therapy (MRT) approaches **(Fig. 472-12)**. Both of these approaches substitute in vitro the entire oocyte or zygote complement of mitochondria, together with their mtDNA from the carrier mother, with the unaffected complement of mitochondria and their unaffected mtDNA from a donor woman. This can be accomplished either by removing and transferring the carrier mother's spindle with her nuclear DNA into the unfertilized oocyte of the donor, or alternatively by transferring the pronucleus from the fertilized oocyte of the carrier mother to the unfertilized donor oocyte from which the pronucleus has been removed. Both of these approaches provide a "bulk" substitution and hence do not target the specific mtDNA mutation, and are potentially applicable to a wide variety of mtDNA disorders. This is a form of germline genetic therapy, and therefore projects onto future generations in the case of a female offspring. Accordingly, ethical and regulatory bodies have appropriately weighed in on the societal implications of such approaches, and have been tentatively supportive of human clinical investigation for situations of preventing great suffering and when the clinical need is clear and unambiguous, subject to specified conditions and principles and subject to ethical scrutiny. Several

such studies have been initiated, and careful examination and follow up are needed to determine developmental and longer-term health and fertility of children who had undergone genetic manipulation at the earliest stages of human development, and whose genomes comprise separate maternal origins of nuclear and mtDNA genomes. It has been recommended that such studies be limited to male offspring, who cannot then transmit the donor mtDNA to future generations, until such time as the health, ethical, and societal issues are well understood, and live up to the exciting promise of reducing the burden of clinical mtDNA disease in the future.

FIGURE 472-12

Mitochondrial replacement techniques—maternal spindle transfer and pronuclear transfer. In both procedures, some mutant mtDNA, estimated at 1–2%, might be carried over together with the spindle or pronucleus, but the levels are low enough to avoid disease risk. IVF denotes in vitro fertilization. *(Used with permission from MJ Falk et al: Mitochondrial Replacement Techniques–Implications for the Clinical Community. N Engl J Med 374: 1103, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FURTHER READING

Alston CL et al: The genetics and pathology of mitochondrial disease. J Pathol 241:236, 2017. [PubMed: 27659608]

Camp K et al: Nutritional interventions in primary mitochondrial disorders: Developing an evidence base. Mol Genet Metab 119:187, 2016. [PubMed: 27665271]

DiMauro S: Mitochondrial encephalomyopathies—Fifty years on: The Robert Wartenberg Lecture. Neurology 81:281, 2013. [PubMed: 23858410]

DiMauro S: et al: The clinical maze of mitochondrial neurology. Nat Rev Neurol 9:429, 2013. [PubMed: 23835535]

Haas RH, et al: Mitochondrial disease: A practical approach for primary care physicians. Pediatrics 129:1326, 2007.

Koopman WJ et al: Monogenic mitochondrial disorders. N Engl J Med 366:1132, 2012. [PubMed: 22435372]

Parikh S et al: Diagnosis and management of mitochondrial disease: A consensus statement from the Mitochondrial Medicine Society. Genet Med 17:689, 2015. [PubMed: 25503498]

Saneto RP et al (eds): *Mitochondrial Case Studies, Underlying Mechanisms and Diagnosis*. London, Academic Press/Elsevier, 2016.

Chapter 473: Applications of Stem Cell Biology in Clinical Medicine

John A. Kessler

FIGURE 473-1

INTRODUCTION

Damage to an organ initiates a series of events that lead to the reconstruction of the damaged tissue, including proliferation, differentiation, and migration of various cell types; release of cytokines and chemokines; and remodeling of the extracellular matrix. Endogenous stem and progenitor cells are among the cell populations that are involved in these injury responses. In normal steady-state conditions, an equilibrium is maintained in which endogenous stem cells intrinsic to the tissue replenish dying cells. After tissue injury, stem cells in organs such as the liver and skin have a remarkable ability to regenerate the organ, whereas other stem cell populations, such as those in the heart and brain, have a much more limited capability for self-repair. In rare circumstances, circulating stem cells may contribute to regenerative responses by migrating into a tissue and differentiating into organ-specific cell types. The goal of stem cell therapies is to promote cell replacement in organs that are damaged beyond their ability to self-repair.

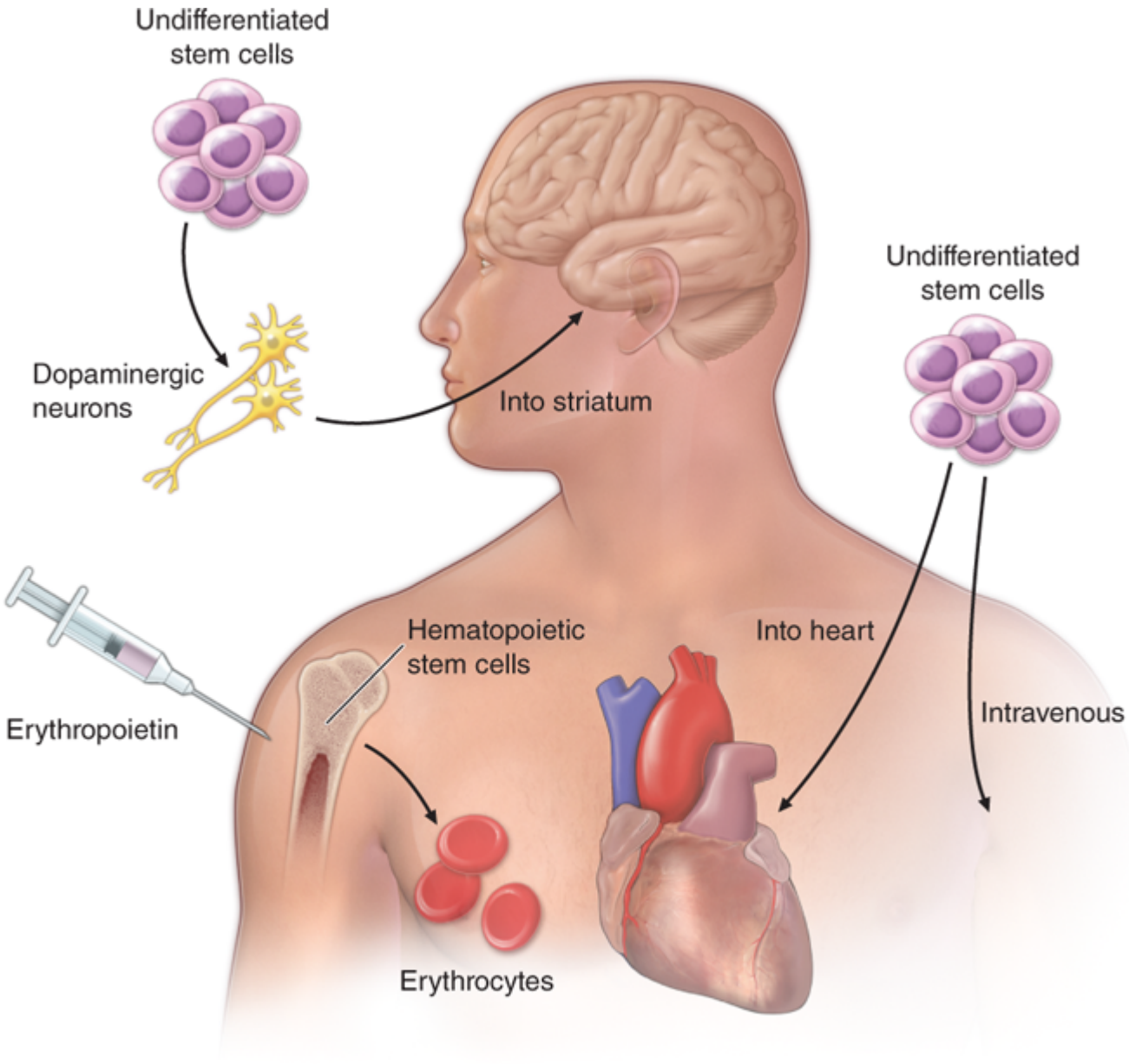
GENERAL STRATEGIES FOR STEM CELL REPLACEMENT

At least three different therapeutic concepts for cell replacement can be envisaged **(Fig. 473-1)**. One therapeutic approach involves direct administration of stem cells. The cells may be injected directly into the damaged organ, where they can differentiate into the desired cell type. Alternatively, stem cells may be injected systemically since they have the capacity to home in on damaged tissues by following gradients of cytokines and chemokines released by the diseased organ. A second approach involves transplantation of differentiated cells derived from stem cells. For example, pancreatic islet cells can be generated from stem cells before transplantation into diabetic patients, and cardiomyocytes can be generated to heart disease. A third approach involves stimulation of endogenous stem cells to facilitate repair. This goal might be accomplished by administration of appropriate growth factors and drugs that amplify the number of endogenous stem/progenitor cells and/or direct them to differentiate into the desired cell types. Therapeutic stimulation of precursor cells is already a clinical reality in the hematopoietic system, where factors such as erythropoietin, granulocyte colony-stimulating factor, and granulocyte-macrophage colony-stimulating factor are used to increase production of specific blood elements. In addition to these strategies for cell replacement, a number of other approaches could involve stem cells for ex vivo or in situ generation of

tissues, a process termed tissue engineering. Stem cells are excellent candidates as vehicles for cellular gene therapy (Chap. 458), and they also potentially can be used to modify immune responses. Finally, transplanted stem cells may exert paracrine effects to promote repair of damaged tissues without differentiating to replace lost cells.

FIGURE 473-1

Strategies for transplantation of stem cells. 1. Undifferentiated or partially differentiated stem cells may be injected directly into the target organ or intravenously. 2. Stem cells may be differentiated ex vivo before injection into the target organ. 3. Growth factors or other drugs may be injected to stimulate endogenous stem cell populations.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Stem cell transplantation is not a new concept but rather is already part of established medical practice. Hematopoietic stem cells (HSCs) (Chap. 92) are responsible for the long-term repopulation of all blood

elements in recipients of bone marrow transplants, and hematopoietic stem cell transplantation is the gold standard against which other stem cell transplantation therapies will be measured. Transplantation of differentiated cells is also a clinical reality, and donated organs and tissues are often used to replace damaged tissues. However, the need for transplantable tissues and organs far outweighs the available supply, and organ transplantation has limited potential for some tissues, such as the brain. Stem cells offer the possibility of a renewable source of replacement cells for virtually all organs.

SOURCES OF STEM CELLS FOR TISSUE REPAIR

A variety of different types of stem cells could be used in regenerative strategies, including embryonic stem (ES) cells, induced pluripotent stem (iPS) cells, umbilical-cord blood stem cells (USCs), organ-specific somatic stem cells (e.g., neural stem cells for treatment of the brain), and somatic stem cells that generate cell types specific for the target organ rather than the donor organ (e.g., bone marrow mesenchymal stem cells [MSCs] or CD34+ HSCs for cardiac repair). Although each cell type has potential advantages and disadvantages, there are a number of generic challenges associated with developing any of these cell types into a useful and reliable clinical tool.

Embryonic Stem Cells

ES cells have the potential to generate all of the cell types in the body; thus, in theory, there are no restrictions on the organs that could be regenerated. ES cells can self-renew endlessly, so that a single cell line with carefully characterized traits potentially could generate almost limitless numbers of cells. In the absence of moral or ethical constraints (see “Ethical Issues,” below), unused human blastocysts from fertility clinics could be used to derive new ES cell lines that are matched immunologically with potential transplant recipients. Alternatively, somatic cell nuclear transfer (“therapeutic cloning”) could be used to create ES cell lines that are genetically identical to those of the patient, although this endeavor has been technically refractory for human cells. However, human ES cells are difficult to culture and grow slowly. Techniques for differentiating them into specific cell types are just beginning to be developed. Cells tend to develop abnormal karyotypes and other abnormalities with increased time in culture, and ES cells have the potential to form teratomas if all cells are not committed to the desired cell types before transplantation. Further, human ES cells are ethically controversial and, on these grounds, their use would be unacceptable to some patients and physicians despite their therapeutic potential. Nevertheless, there have been limited clinical trials of ES-derived cells in a number of disorders, including macular degeneration, myopia, heart failure, diabetes, and spinal cord injury (SCI).

Induced Pluripotent Stem Cells

The field of stem cell biology was transformed by the discovery that adult somatic cells can be converted (“reprogrammed”) into pluripotent cells through the overexpression of four transcription factors normally expressed in pluripotent cells. These iPS cells share most properties with ES cells, although there are distinct differences in gene expression between ES and iPS cells. The initial use of viruses to insert the transcription

factors into somatic cells made the resulting cells unsuitable for clinical use. However, a number of strategies have since been developed to circumvent this problem, including the insertion of modified mRNAs, proteins, or microRNAs rather than cDNAs and treatment with small molecules; the use of non-integrating viruses such as Sendai virus; the insertion of transposons with the programming factors, followed by their subsequent removal; and the use of floxed viral constructs, followed by treatment with Cre recombinase to excise those constructs. iPS cells derived from patients with different disorders are currently being used extensively for disease modeling and drug discovery. However, the safety of iPS cells for use in regenerative strategies in humans remains to be demonstrated. The first clinical trial in macular degeneration was suspended after treatment of one patient because of discovery of a mutation in cells derived for the second patient, but it is scheduled to resume. Potential advantages of iPS cells are that somatic cells from patients would generate pluripotent cells genetically identical to those of the patient, and that these cells are not subject to the same ethical constraints as ES cells. It is not clear whether the differences in gene expression between ES and iPS cells will have any impact on their potential clinical utility, and studies of both cell types will be needed to resolve this issue.

Umbilical-Cord Stem Cells

USCs are widely available. These cells appear to be associated with less graft-versus-host disease than are some other cell types, such as marrow stem cells. They have less human leukocyte antigen restriction than adult marrow stem cells and are less likely to be contaminated with herpesvirus. However, it is unclear how many different cell types can be generated from USCs, and methods for differentiating these cells into non-hematopoietic phenotypes are largely lacking. Nevertheless, there are ongoing clinical trials of these cells in dozens of disorders, including cirrhosis, cardiomyopathies, multiple sclerosis, burns, stroke, autism, and critical limb ischemia.

Organ-Specific Multipotent Stem Cells

Organ-specific multipotent stem cells have the advantage of already being somewhat specialized so that the inducement of desired cell types may be easier. Cells potentially could be obtained from the patient and amplified in cell culture, circumventing the problems associated with immune rejection. Stem cells are relatively easy to harvest from some tissues, such as bone marrow and blood, but are difficult to harvest from other tissues, such as heart and brain. Moreover, these populations of cells are more limited in potentiality than are pluripotent ES or iPS cells, and they may be difficult to obtain in large quantities from many organs. Therefore, substantial efforts have been devoted to developing techniques for using more easily obtainable stem cell populations, such as bone marrow MSCs, CD34+ HSCs, cardiac mesenchymal cells, and adipose-derived stem cells (ASCs), for use in regenerative strategies. Tissue culture evidence suggests that these stem cell populations may be able to generate differentiated cell types unrelated to their organ source (including myocytes, chondrocytes, tendon cells, osteoblasts, cardiomyocytes, adipocytes, hepatocytes, and possibly neurons) in a process known as *transdifferentiation*. However, it is still unclear whether these stem cells are capable of generating differentiated cell types that integrate into organs, survive, and function after transplantation in vivo. A number of early studies of MSCs transplanted into heart,

liver, and other organs suggested that the cells had differentiated into organ-specific cell types with beneficial effects in animal models of disease. Unfortunately, subsequent studies revealed that the stem cells had simply fused with cells resident in the organs and that the observed beneficial effects were due to paracrine release of trophic and anti-inflammatory cytokines. Further studies will be necessary to determine whether transdifferentiation of MSCs, ASCs, or other stem cell populations occurs at a high enough frequency to make these cells useful for stem cell replacement therapy. Despite the remaining issues, clinical trials of MSCs, autologous HSCs, USCs, and ASCs are being performed in many disorders, including ischemic cardiac disease, cardiomyopathy, diabetes, stroke, cirrhosis, amyotrophic lateral sclerosis (ALS), and muscular dystrophy. Another approach has been to derive organ-specific stem cells by parthenogenesis, and a clinical trial of neural stem cells derived this way has begun for Parkinson's disease.

Regardless of the source of the stem cells used in regenerative strategies, a number of generic problems must be overcome for the development of successful clinical applications. Methods must be devised to reliably generate large numbers of specific cell types, to minimize the risk of tumor formation or proliferation of inappropriate cell types, to ensure the viability and function of the engrafted cells, to overcome immune rejection when autografts are not used, and to facilitate revascularization of regenerated tissue. Each organ system also will pose tissue-specific problems for stem cell therapies.

DISEASE-SPECIFIC APPLICATIONS OF STEM CELLS

Ischemic Heart Disease and Cardiomyocyte Regeneration

Because of the high prevalence of ischemic heart disease, extensive efforts have been devoted to the development of strategies for stem cell replacement of cardiomyocytes. Historically, the adult heart has been viewed as a terminally differentiated organ without the capacity for regeneration. However, recent studies have demonstrated that the heart has the capacity for low levels of cardiomyocyte regeneration (Chap. 232). This regeneration appears to be accomplished by cardiac stem cells resident in the heart and possibly also by cells originating in the bone marrow. The heart might be an ideal source of stem cells for therapeutic use, but techniques for isolating, characterizing, and amplifying large numbers of these cells have not yet been perfected. For successful myocardial repair, stem cell therapy must deliver cells either systemically or locally, and the cells must survive, engraft, and differentiate into functional cardiomyocytes that couple mechanically and electrically with the recipient myocardium. The optimal method for cell delivery is not clear, and various experimental and clinical studies have successfully employed intramyocardial, transendocardial, intravenous, intracoronary, and retrograde coronary venous injections. In experimental myocardial infarction, functional improvements have been achieved after transplantation of a variety of different cell types, including ES cells, HSCs, MSCs, USCs, and ASCs. Early studies suggested that each of these cell types might have the potential to engraft and generate cardiomyocytes. However, most investigators have found that the generation of new cardiomyocytes by these cells is at best a rare event and that graft survival over long periods is poor. The preponderance of evidence suggests that the observed beneficial effects of most experimental therapies were not derived from direct stem cell generation of cardiomyocytes but rather from indirect effects of the stem cells on resident cells. It is not clear whether

these effects reflect the release of soluble trophic factors, the induction of angiogenesis, the release of anti-inflammatory cytokines, or another mechanism. A wide variety of cell delivery methods, cell types, and cell doses have been used in a progressively enlarging series of clinical trials, but the fate of the cells and the mechanisms by which they alter cardiac function are still open questions. In aggregate, however, these studies have shown a small but measurable improvement in cardiac function and, in some cases, reduction in infarct size. Further, transplantation of bone marrow-derived stem cells improved outcome for patients in heart failure. In short, the available evidence suggests that the beneficial clinical impact reflects an indirect effect of the transplanted cells rather than cell replacement. However, genuine cell replacement may become possible as new protocols are being developed for generating cardiomyocytes from pluripotent and multipotent stem cells.

Diabetes

Successes with islet cell and pancreas transplantation have provided proof of concept for cell-based therapies for type 1 diabetes. However, the demand for donor pancreases far exceeds the number available, and maintenance of long-term graft survival remains a problem. The search for a renewable source of stem cells capable of regenerating pancreatic islets has therefore been intensive. Pancreatic beta cell turnover occurs even in the normal pancreas, although the source of the new beta cells remains controversial. This persistent turnover suggests that, in principle, it should be possible to develop strategies for reconstituting the beta cell population in diabetics. Attempts to devise techniques for promoting endogenous regenerative processes by using combinations of growth factors, drugs, and gene therapy have failed thus far, but this remains a potentially viable approach. A number of different cell types are candidates for use in stem cell replacement strategies, including iPS cells, ES cells, hepatic progenitor cells, pancreatic ductal progenitor cells, and MSCs. Successful therapy will depend on the development of a source of cells that can be amplified to produce large numbers of progeny with the ability to synthesize, store, and release insulin when it is required, primarily in response to changes in the ambient level of glucose. The proliferative capacity of the replacement cells must be tightly regulated to avoid excessive expansion of beta cell numbers and the consequent development of hyperinsulinemia/hypoglycemia; moreover, the cells must withstand immune rejection. Although ES and iPS cells can be differentiated into cells that produce insulin, these cells have a lower content of insulin and a higher rate of apoptosis than pancreatic beta cells, and generally lack the capacity to fully normalize blood glucose levels in diabetic animals. However, clinical trials of encapsulated ES cell-derived pancreatic progenitor cells are currently in progress.

During embryogenesis, the pancreas, liver, and gastrointestinal tract are all derived from the anterior endoderm, and transdifferentiation of pancreas to liver and vice versa has been observed in a number of pathologic conditions. There is also substantial evidence that multipotential stem cells reside within gastric glands and intestinal crypts. These observations suggest that hepatic, pancreatic, and/or gastrointestinal precursor cells may be reasonable candidates for cell-based therapy for diabetes, although it is unclear whether insulin-producing cells derived from pancreatic stem cells or liver progenitors can be expanded in vitro to clinically useful numbers. MSCs and neural stem cells both reportedly have the capacity to generate

insulin-producing cells, but there is no convincing evidence that either cell type will be clinically useful. Clinical trials of MSCs, USCs, HSCs, and ASCs in both type 1 and type 2 diabetes are ongoing.

Nervous System

Substantial progress has been made in the development of methodologies for generating neural cells from different stem cell populations. Human ES or iPS cells can be induced to generate cells with the properties of neural stem cells, and these cells in turn give rise to neurons, oligodendroglia, and astrocytes. Reasonably large numbers of these cells can be transplanted into the rodent brain with formation of appropriate cell types and no tumor formation. Multipotent stem cells present in the adult brain also can be easily amplified in number and used to generate all the major neural cell types, but the need for invasive procedures to obtain autologous cells is a major limitation. Fetal neural stem cells derived from miscarriages or abortions are an alternative but raise ethical concerns. Nevertheless, clinical trials of fetal neural stem cells have commenced in ALS, stroke, and several other disorders. Transdifferentiation of MSCs and ASCs into neural stem cells, and vice versa, has been reported by numerous investigators, and clinical trials of such cells have begun for a number of neurologic diseases. Clinical trials of a conditionally immortalized human cell line and of USCs in stroke are also in progress. Because of the incapacitating nature of neural disorders and the limited endogenous repair capacity of the nervous system, clinical trials of stem cells in neurologic disorders have been particularly numerous, including trials in SCI, multiple sclerosis, epilepsy, Alzheimer's disease, ALS, acute and chronic stroke, numerous genetic disorders, traumatic brain injury, Parkinson's disease, and others. In diseases such as ALS, possible benefits are more likely to be due to indirect trophic effects than to neuron replacement. In Parkinson's disease, the major motor features of the disorder result from the loss of a single cell population: dopaminergic neurons within the substantia nigra; this circumstance suggests that cell replacement should be relatively straightforward. However, two clinical trials of fetal nigral transplantation failed to meet their primary endpoint and were complicated by the development of dyskinesia. Transplantation of stem cell–derived dopamine-producing cells offers a number of potential advantages over the fetal transplants, including the ability of stem cells to migrate and disperse within tissue, the potential for engineering regulatable release of dopamine, and the ability to engineer cells to produce factors that will enhance cell survival. Nevertheless, the experience with fetal transplants points out the difficulties that may be encountered.

At least some of the neurologic dysfunction after SCI reflects demyelination, and both ES cells and MSCs can facilitate remyelination after experimental SCI. Clinical trials of MSCs in this disorder have commenced in a number of countries, and SCI was the first disorder targeted for the clinical use of ES cells. The first trial of ES cell derived oligodendroglial progenitor cells in SCI was terminated early for nonmedical reasons, but another trial has commenced. At present, no population of transplanted stem cells has been shown to have the capacity to generate neurons that extend axons over long distances to form synaptic connections (as would be necessary for replacement of upper motor neurons in ALS, stroke, or other disorders). For many injuries, including SCI, the balance between scar formation and tissue repair/regeneration may prove to be an important consideration. For example, it may ultimately prove necessary to limit scar formation so that axons can reestablish connections.

Liver

Liver transplantation is currently the only successful treatment for end-stage liver diseases, but the shortage of liver grafts limits its application. Clinical trials of hepatocyte transplantation demonstrate its potential as a substitute for organ transplantation, but this approach is limited by the paucity of available cells. Potential sources of stem cells for regenerative strategies include endogenous liver stem cells (such as oval cells), ES cells, MSCs, and USCs. Although a series of studies in humans as well as animals suggested that transplanted MSCs and HSCs can generate hepatocytes, fusion of the transplanted cells with endogenous liver cells, giving the erroneous appearance of new hepatocytes, appears to be the underlying event in most circumstances. The available evidence suggests that transplanted HSCs and MSCs can generate hepatocyte-like cells in the liver only at a very low frequency, but there are beneficial consequences presumably related to indirect paracrine effects. ES cells can be differentiated into hepatocytes and transplanted in animal models of liver failure without the formation of teratomas. Clinical trials are in progress in cirrhosis with numerous cell types, including MSCs, USCs, HSCs, and ASCs.

Other Organ Systems and the Future

The use of stem cells in regenerative strategies has been studied for many other organ systems and cell types, including skin, eye, cartilage, bone, kidney, lung, endometrium, vascular endothelium, smooth muscle, and striated muscle, and clinical trials in these and other organs are ongoing. In fact, the potential for stem cell regeneration of damaged organs and tissues is virtually limitless. However, numerous obstacles must be overcome before stem cell therapies can become a widespread clinical reality. Only HSCs have been adequately characterized by surface markers so that they can be unambiguously identified, a prerequisite for reliable clinical applications. The pathways for differentiating stem cells into specific cellular phenotypes are largely unknown, and the ability to control the migration of transplanted cells or predict the response of the cells to the environment of diseased organs is presently limited. Some strategies may employ the coadministration of scaffolding, artificial extracellular matrix, and/or growth factors to orchestrate differentiation of stem cells and their organization into appropriate constituents of the organ. There is currently no way to image stem cells in vivo after transplantation into humans, and it will be necessary to develop techniques to do so. Fortunately, stem cells can be engineered before transplantation to contain a contrast agent that may make in vivo imaging feasible. The potential for tumor formation and the problems associated with immune rejection are impediments, and it will also be necessary to develop techniques for ensuring vascularization of regenerated tissues. There already are many strategies for supporting cell replacement, including coadministration of vasoactive endothelial growth factor to foster vascularization of the transplant. Some strategies also include the genetic engineering of stem cells with an inducible suicide gene so that the cells can be easily eradicated in the event of tumor formation or another complication. The potential for stem cell therapies to revolutionize medical care is extraordinary, and disorders such as myocardial infarction, diabetes, and Parkinson’s disease, among many others, are potentially curable by such therapies. However, stem cell–based therapies are still at a very early stage of development, and perfection of techniques for clinical transplantation of predictable, well-characterized cells is going to be a difficult and lengthy undertaking.

ETHICAL ISSUES

Stem cell therapies raise ethical and socially contentious issues that must be addressed in parallel with the scientific and medical opportunities. Society has great diversity with respect to religious beliefs, concepts of individual rights, tolerance for uncertainty and risk, and boundaries for how scientific interventions should be used to alter the outcome of disease. In the United States, the federal government has authorized research using existing human ES cell lines but still restricts the use of federal funds for developing new human ES cell lines. Ongoing studies of existing lines have indicated that they develop abnormalities with time in culture and that they may be contaminated with mouse proteins. These findings highlight the need to develop new human ES cell lines. The development of iPS cell technology may lessen the need for deriving new ES cell lines, but it is still not clear whether the differences in gene expression by ES and iPS cells are important for potential clinical use.

In considering ethical issues associated with the use of stem cells, it is helpful to draw from experience with other scientific advances, such as organ transplantation, recombinant DNA technology, implantation of mechanical devices, neuroscience and cognitive research, in vitro fertilization, and prenatal genetic testing. These and other precedents have pointed to the importance of understanding and testing fundamental biology in the laboratory setting and in animal models before applying new techniques in carefully controlled clinical trials. When these trials occur, they must include full informed consent and careful oversight by external review groups.

Ultimately, there will be medical interventions that are scientifically feasible but ethically or socially unacceptable to some members of a society. Stem cell research raises fundamentally difficult questions about the definition of human life, and it has raised deep fears about the ability to balance issues of justice and safety with the needs of critically ill patients. Health care providers and experts with backgrounds in ethics, law, and sociology must help guard against the premature or inappropriate application of stem cell therapies and the inappropriate involvement of vulnerable population groups. However, these therapies offer important new strategies for the treatment of otherwise irreversible disorders. An open dialogue among the scientific community, physicians, patients and their advocates, lawmakers, and the lay population is critically important to raise and address important ethical issues and balance the benefits and risks associated with stem cell transfer.

FURTHER READING

Benthuysen JR, Carrano AC, Sander M: Advances in β cell replacement and regeneration strategies for treating diabetes. J Clin Invest 126:3651, 2016. [PubMed: 27694741]

Goodell MA, Rando TA: Stem cells and healthy aging. Science 350:1199, 2015. [PubMed: 26785478]

Guidelines for Stem Cell Research and Clinical Translation International Society for Stem Cell Research May 2016. Available from *http://www.isscr.org/docs/default-source/guidelines/isscr-guidelines-for-stem-cell-*

research-and-clinical-translation.pdf?sfvrsn=2. Accessed June 26, 2017.

Nguyen PK, Rhee J-W, Wu JC: Adult Stem Cell Therapy and Heart Failure, 2000 to 2016: A Systematic Review. JAMA Cardiol 1:831, 2016. [PubMed: 27557438]

Passier R, Orlova V, Mummery C: Complex tissue and disease modeling using hiPSCs. Cell Stem Cell 18:309, 2016. [PubMed: 26942851]

Rowe RG et al: Engineering hematopoietic stem cells: Lessons from development. Cell Stem Cell. 18:707, 2016. [PubMed: 27257760]

Steinbeck JA, Studer L: Moving stem cells to the clinic: Potential and limitations for brain repair. Neuron 86:187, 2015. [PubMed: 25856494]

Tapia N, Schöler HR: Molecular obstacles to clinical translation of iPSCs. Cell Stem Cell 19:298, 2016. [PubMed: 27452174]

Trounson A, DeWitt ND: Pluripotent stem cells progressing to the clinic. Nat Rev Mol Cell Biol 17:194, 2016. [PubMed: 26908143]

Chapter 474: Microbial Genomics and Infectious Disease

Roby P. Bhattacharyya; Yonatan H. Grad; Deborah T. Hung

INTRODUCTION

Just as microscopy opened up the worlds of microbiology by providing a tool with which to visualize microorganisms, technological advances in genomics are now providing microbiologists with powerful new methods to characterize the genetic map that underlies all microbes with unprecedented resolution, thereby illuminating their complex and dynamic interactions with each other, the environment, and human health. The field of infectious disease genomics encompasses a vast frontier of active research that is beginning to transform the clinical practice of infectious diseases. While genetics has long played a key role in elucidating the process of infection and impacting clinical infectious diseases, the ability to extend our thinking and our approaches beyond the study of single genes to an examination of the sequence, structure, and function of entire genomes is allowing us to identify new possibilities for research and opportunities to change clinical practice. From the development of diagnostics with unprecedented sensitivity, specificity, and speed to the design of novel public health interventions, technical and statistical genomic innovations are reshaping our understanding of the influence of the microbial world on human health and providing us with new tools to diagnose, track, and combat infection. In this chapter, we explore the application of genomics methods to microbial pathogens and the infections they cause. We discuss innovations that are driving the development of diagnostic approaches as well as the discovery of new pathogens, providing insight into novel therapeutic approaches and paradigms, and advancing methods in infectious disease epidemiology and the study of pathogen evolution that can inform infection control measures, public health responses to outbreaks, and vaccine development. We draw on examples in current practice and from the recent scientific literature as signposts that point toward ways in which the insights from pathogen genomics may influence infectious diseases in the short and long terms. **Table 474-1** provides definitions for a selection of important terms used in genomics.

TABLE 474-1

Glossary of Selected Terms in Genomics

Term	Definition
Contig	A DNA sequence representing a continuous fragment of a genome, assembled from overlapping sequences; relevant for de novo assembly of sequence data that do not align to previously sequenced genomes
Genome	The entire set of heritable genetic material within an organism
Horizontal gene transfer	The transfer of genes between organisms through mechanisms other than by clonal descent, such as through transformation, conjugation, or transduction
Metagenomics	Analysis of genetic material from multiple species directly from primary samples without requiring prior culture steps
Microarray	A collection of DNA oligonucleotides (“oligos”) spatially arranged on a solid surface and used to detect or quantify sequences in a sample of interest that are complementary (and therefore bind) to one or more of the arrayed oligos
Microbial genome-wide association study (GWAS)	An analytic framework to test statistical associations between microbial genotypes and phenotypes of interest, such as antibiotic resistance and virulence
Mobile genetic elements	DNA elements that can move within a genome and can be transferred between genomes through horizontal gene transfer (e.g., plasmids, bacteriophages, and transposons)
Multilocus sequence typing	A method for typing organisms based on DNA sequence fragments from a prespecified set of genes
Next-generation sequencing	High-throughput sequencing using a parallelized sequencing process that produces millions of sequences concurrently, far beyond the capacity of prior dye-terminator methods
Nucleic acid amplification test (NAAT)	A biochemical assay that evaluates for the presence of a particular string of nucleic acids through amplification by one of several methods, including polymerase and ligase chain reactions
Polymerase chain reaction (PCR)	A type of NAAT used to amplify a specific region of DNA by means of specific oligonucleotide primers and a DNA polymerase
Single-nucleotide polymorphism (SNP)	Point mutations, the number of which in different microbial isolates is a measure of their genetic distance from one another
Transcriptome	The catalog of the full set of messenger RNA (mRNA) transcripts from a cell or organism, which are typically measured by microarray or by next-generation sequencing of complementary DNA (cDNA) via a process called RNA-Seq
Whole-genome sequencing (WGS)	A process that determines the full DNA sequence of an organism’s genome; has been greatly facilitated by next-generation sequencing technology

MICROBIAL DIAGNOSTICS

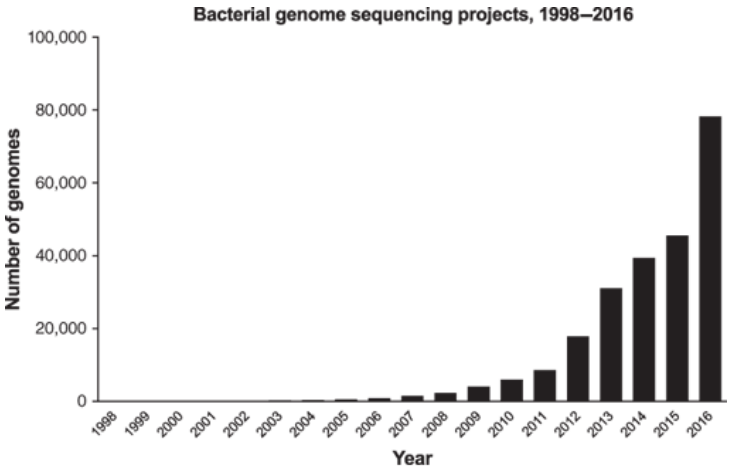
The basic goals of a clinical microbiology laboratory are to establish the presence of a pathogen in a clinical sample, to identify the pathogen, and, when possible, to provide other information that can help guide clinical management and even affect prognosis, such as antibiotic susceptibility profiles or the presence of virulence factors. To date, clinical microbiology laboratories have largely approached these goals phenotypically by growth-based assays and biochemical testing. Bacteria, for instance, are algorithmically grouped into species by their characteristic microscopic appearance, nutrient requirements for growth, and ability to catalyze certain reactions. Antibiotic susceptibility is determined in most cases by assessing bacterial growth in the presence of antibiotic.

With the sequencing revolution paving the way to easy access of complete pathogen genomes **(Fig. 474-1)**, we are now able to more systematically understand the genetic basis for these observable phenotypes. Compared with traditional growth-based methods for bacterial diagnostics that dominate

the clinical microbiology laboratory, nucleic acid–based diagnostics promise improved speed, sensitivity, specificity, and breadth of information. Bridging clinical and research laboratories, adaptations of genomic technologies have begun to deliver on this promise (Table 474-2).

FIGURE 474-1

Bacterial genome sequencing projects submitted to the Genomes Online Database, a manually curated repository for genome and metagenome sequencing data, from 1998 to 2016. *(Data compiled from https://gold.jgi.doe.gov/statistics, accessed July 17, 2017. See S Mukherjee et al: Genomes OnLine Database (GOLD) v.6: Data updates and feature enhancement. Nucleic Acids Res 45:D446, 2017.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 474-2

Selected Clinical Applications of Infectious Disease Genomics

Application	Technology	Notes/Examples
Organism Identification		
Viral detection	PCR	Identification of HIV, HBV, HCV, respiratory viruses including influenza, and others for diagnosis and response to therapy
TB detection	PCR	Amplification of the *rpoB* gene for species-specific identification of *Mycobacterium tuberculosis*
Pathogen detection	PCR, NAAT	Multiplexed identification of dozens of viruses, bacteria, yeasts, and parasites from a variety of clinical specimens
Bacterial detection	16S ribosomal gene sequencing	Targeted amplification and sequencing of regions of the 16S rRNA gene for identification of suspected bacterial infections undiagnosed by conventional methods
Pathogen Discovery		
Bacterial pathogens	Sequencing, metagenomic assembly	Unbiased "shotgun" sequencing of isolated nucleic acid from patient samples to identify associated pathogens; proofs-of-concept: new *Bradyrhizobium* species associated with cord colitis; *Escherichia coli* O104:H4 from 2011 diarrheal outbreak in Germany; *Leptospira* species from one patient's cerebrospinal fluid; research use only at this time
Viral pathogens	Microarray, sequencing	Hybridization of clinical samples to microarrays from phylogenetically diverse known viruses identified the SARS coronavirus and others. Direct sequencing has identified West Nile virus and the MERS coronavirus, among others. Use is primarily in research.
Antibiotic Resistance		
MRSA detection	PCR	Detection of the *mecA* gene, the genotypic cause of methicillin resistance in *Staphylococcus aureus*
VRE detection	PCR	Detection of the *vanA* or *vanB* gene, the main genotypic causes of vancomycin resistance in *Enterococcus*
MDR-TB detection	PCR, NAAT	Detection of polymorphisms in the *rpoB* gene from *M. tuberculosis*, which account for 95% of rifampin resistance. Other probes available for *inhA* and *katG* genes can detect up to 85% of isoniazid resistance.
Carbapenemase detection	PCR	Detection of genes encoding one of several types of enzymes (KPC, NDM, OXA-48, IMP-1, VIM) that hydrolyze carbapenems, accounting for much but not all carbapenemase resistance in Enterobacteriaceae
HIV resistance detection	Targeted sequencing	Targeted sequencing of specific genes with known resistance-conferring mutations; now the standard of care prior to initial therapy in the United States and Europe
Epidemiology		
Outbreak and epidemic tracking	Sequencing	Application to tracking outbreaks and epidemics on local and international scales, including spread of carbapenemase-producing *Klebsiella*, *S. aureus*, *M. tuberculosis*, *E. coli*, *Vibrio cholerae*, Ebola virus, Zika virus, and influenza virus
Evolution and spread of pathogens	Sequencing	Sequencing collections of pathogens to shed light on pathogen dissemination, virulence factors, and antibiotic resistance determinants; innumerable examples, including *V. cholerae*, influenza virus, Ebola virus, and Zika virus

Abbreviations: HBV, hepatitis B virus; HCV, hepatitis C virus; MDR, multidrug-resistant; MERS, Middle East respiratory syndrome; MRSA, methicillin-resistant *S. aureus*; NAAT, nucleic acid amplification test; PCR, polymerase chain reaction; SARS, severe acute respiratory syndrome; TB, tuberculosis; VRE, vancomycin-resistant enterococci.

HISTORICAL LIMITATIONS AND PROGRESS THROUGH GENETIC APPROACHES

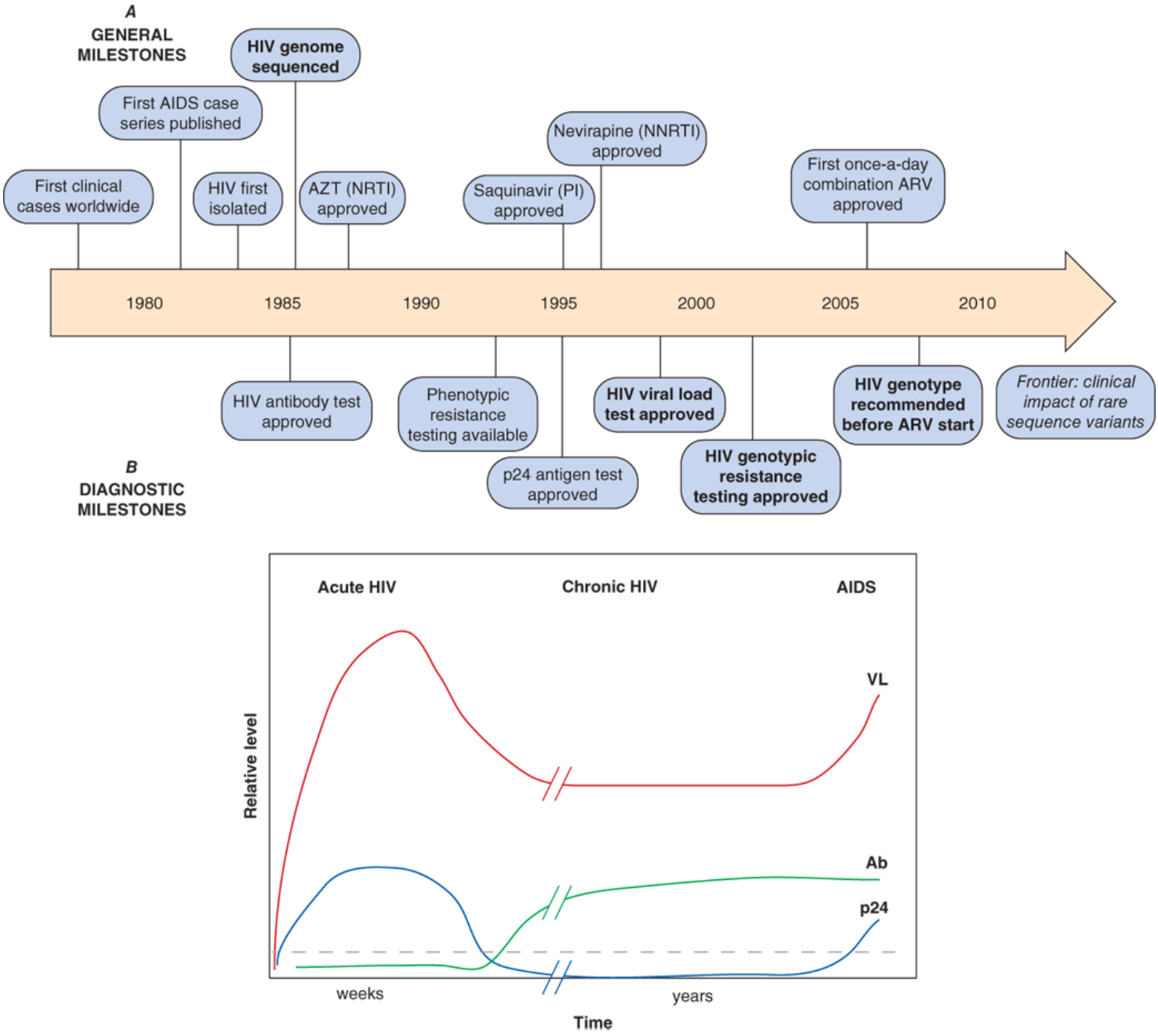
The molecular diagnostics revolution in the clinical microbiology laboratory is well under way, born of necessity in the effort to identify microbes that are refractory to traditional culture methods. Historically, diagnosis of many so-called unculturable pathogens has relied largely on serology and antigen detection. However, these methods provide only limited clinical information because of their suboptimal sensitivity and specificity as well as the long delays that diminish their utility for real-time patient management. Newer tests to detect pathogens, based on nucleic acid content, have already offered improvements in the select cases in which they have been applied.

Unlike direct pathogen detection, serologic diagnosis—measurement of the host's response to pathogen exposure—can typically be made only in retrospect, requiring both acute- and convalescent-phase serum samples. For chronic infections, distinguishing active from latent infection or identifying repeat exposure from serology alone can be difficult or impossible, depending on the syndrome. In addition, serologic diagnosis is variably sensitive, depending on the organism and the patient's immune status. For instance, tuberculosis is notoriously difficult to identify by serologic methods; tuberculin skin testing using purified protein derivative (PPD) is especially insensitive in active disease and possibly cross-reactive with vaccines or other mycobacteria. Even the newer interferon γ release assays (IGRAs), which measure cytokine release from T lymphocytes in response to *Mycobacterium tuberculosis*–specific antigens in vitro, have limited sensitivity in immunodeficient hosts. Neither PPD testing nor IGRAs can distinguish latent from active infection. Serologic Lyme disease diagnostics suffer similar limitations: in patients from endemic regions, the presence of IgG antibodies to *Borrelia burgdorferi* may reflect prior exposure rather than active disease, while IgM antibodies are imperfectly sensitive and specific (50% and 80%, respectively, in early disease). The complicated nature of these tests, particularly in view of the nonspecific symptoms that may accompany Lyme disease, has had substantial implications for public perception of Lyme disease and antibiotic misuse in endemic areas. Similarly, syphilis, a chronic infection caused by *Treponema pallidum*, is notoriously difficult to stage by serology alone, requiring multiple different nontreponemal and treponemal tests (e.g., rapid protein reagin and fluorescent treponemal antibody, respectively) in conjunction with clinical suspicion. Complementing serology, antigen detection can improve sensitivity and specificity in select cases, but has been validated only for a limited set of infections. Typically, structural elements of pathogens are detected, including components of viral envelopes (e.g., hepatitis B surface antigen, HIV p24 antigen), cell surface markers in certain bacteria (e.g., *Streptococcus pneumoniae*, *Legionella pneumophila* serotype 1) or fungi (e.g., *Cryptococcus*, *Histoplasma*), and less specific fungal cell-wall components such as galactomannan and β-glucan (e.g., *Aspergillus* and other dimorphic fungi).

Given the impracticality of culture and the lack of sensitivity or sufficient clinical information afforded by serologic and antigenic methods, the push toward nucleic acid–based diagnostics originated in pursuit of viruses and fastidious bacteria, becoming part of the standard of care for select organisms in U.S. hospitals. Such tests, including polymerase chain reaction (PCR) and other nucleic acid amplification tests (NAATs), are now widely used for many viral infections, both chronic (e.g., HIV infection, hepatitis C) and acute (e.g., influenza). NAATs provide essential information about both the initial diagnosis and the response to therapy and in some cases genotypically predict drug resistance. Indeed, progression from antigen detection to PCR transformed our understanding of the natural course of HIV infection, with profound implications for treatment **(Fig. 474-2A)**. In the early years of the AIDS pandemic, p24 antigenemia was detected in acute HIV infection but then disappeared for years before emerging again with progression to AIDS (Fig. 474-2B). Without a marker demonstrating viremia, the role of treatment during HIV infection prior to the development of clinical AIDS was uncertain, and assessing treatment efficacy was challenging. With the emergence of PCR as a progressively more sensitive test (now able to detect as few as 20 copies of virus per milliliter of blood), viremia was recognized as a near-universal feature of HIV infection. Given the challenges of phenotypic assays, genotypic antiviral resistance testing was also adopted early for HIV and is now the standard of care before the initiation of therapy in developed countries. These developments have been transformative in guiding therapy in early disease and, together with the development of less toxic therapies, have helped to shape policy that is moving toward ever-earlier introduction of antiretroviral therapy in HIV infection.

FIGURE 474-2

A. Timeline of select milestones in HIV management. Genomic advances are shown in bold type. The approvals and recommendations indicated apply to the United States. ARV, antiretroviral; AZT, zidovudine; NRTI, nucleoside reverse transcriptase (RT) inhibitor; NNRTI, nonnucleoside RT inhibitor; PI, protease inhibitor. ***B.*** Viral dynamics in the natural history of HIV infection. Three diagnostic markers are shown: HIV antibody (Ab), p24 antigen (p24), and viral load (VL). *Dashed gray line* represents limit of detection. *(Adapted from data in HH Fiebig et al: Dynamics of HIV viremia and antibody seroconversion in plasma donors: Implications for diagnosis and staging of primary HIV infection. AIDS 17:1871, 2003.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

As they are for viral testing, nucleic acid–based tests have become the diagnostic tests of choice for fastidious bacteria, including the common sexually transmitted bacterial pathogens *Neisseria gonorrhoeae* and *Chlamydia trachomatis* as well as the tick-borne *Ehrlichia chaffeensis* and *Anaplasma phagocytophilum*. More recently, nucleic acid amplification–based detection has offered improved sensitivity for diagnosis of the important nosocomial pathogen *Clostridium difficile*, and NAATs have provided clinically relevant information on the presence of cytotoxins A and B as well as molecular markers of hypervirulence, such as the North American pulsotype 1 (NAP1) strain that is enriched in severe illness. The importance of genomics in selecting loci for diagnostic assays and in monitoring test sensitivity was highlighted by the emergence in Sweden of a newly recognized variant of *C. trachomatis* with a deletion that includes the gene targeted by a set of commercial NAATs. By evading detection through this deletion (and thus avoiding treatment), this strain came to be highly prevalent in some areas of Sweden. While nucleic acid–based tests remain the diagnostic approach of choice for fastidious bacteria, this example serves as a reminder of the need for careful development and ongoing monitoring of molecular diagnostics.

In contrast, for typical bacterial pathogens for which culture methods are well established, growth-based assays followed by biochemical tests still dominate in the clinical laboratory. Informed by decades of clinical microbiology, these tests have served clinicians well, yet the limitations of growth-based tests—in particular, the delays associated with waiting for growth—have left opportunities for improvements. Driven by this need, mass spectrometry–based assays are already being adopted for highly accurate organism identification within a few hours of a positive blood culture. Looking ahead, molecular diagnostics, greatly informed by the vast quantity of microbial genome sequences generated in recent years, offers a way forward. First, sequencing studies can readily identify key genes (or noncoding nucleic acids) that can be developed into targets for clinical assays using PCR or

hybridization assay platforms. Second, sequencing itself may eventually become cheap and rapid enough to be performed routinely on clinical specimens, with consequent unbiased detection of pathogens.

One of the biggest drivers for the implementation of novel molecular technologies in the diagnosis of infectious diseases is the desire for more rapid—or even real-time—pathogen identification, ideally with antibiotic susceptibility information on those microbes for which resistance to the current anti-infective armamentarium is of concern. Such real-time tests have the potential to transform infectious disease management, impacting antibiotic stewardship in the outpatient setting, mortality risk in the critically ill (i.e., patients in whom early administration of effective antibiotics is the most significant factor in decreasing mortality risk), hospital admission, and length of hospital stay; the extent of this impact will depend on the economic forces that will help define the breadth of their deployment. On the public health level, such tests will likely play a role in improving antibiotic stewardship, thereby influencing the rise of antibiotic resistance and enabling surveillance of outbreaks by local, national, and international networks. In the United States and the United Kingdom, for example, public health agencies have shifted from pulsed-field gel electrophoresis to genome sequencing to track food-borne pathogens and identify outbreaks; in addition, these countries are rapidly expanding the routine use of genomics in identifying and characterizing other pathogens, from mycobacteria (both *M. tuberculosis* and nontuberculous mycobacteria) to *N. gonorrhoeae*. Further, international efforts to track the spread of viral diseases, including recent work on Ebola and Zika outbreaks and ongoing work on seasonal influenza, offer opportunities for improving interventions, surveillance, and prevention efforts, ranging from more accurate selection of the influenza virus strains to include in seasonal vaccine development to improved design of trials to evaluate novel vaccines and therapies.

Technological innovations are lowering several critical barriers to the widespread adoption of genomics and other molecular methods. Specifically, for clinical sequencing: (1) the cost and speed of sequencing and analysis methods continue to fall precipitously; (2) automation and miniaturization of the preparation of a sample for sequencing promise to reduce cost and minimize the expertise needed; and (3) direct sequencing technologies that eliminate the complex molecular biology required to prepare clinical samples for sequencing are improving in accuracy and robustness. Further barriers exist, including the need for standardized pipelines to process data and present clinicians with easily interpretable and readily actionable results. However, as these advances give rise to rapid, accurate diagnostic tests, the ultimate goal is to inform a clinician in real time whether antibiotics are indicated and, if so, which will be effective. Real-time diagnostics will allow more efficient deployment of our precious antibiotic arsenal, thus improving both societal and patient-specific outcomes in much the same way that a rapid, sensitive troponin assay has transformed bedside management of chest pain.

ORGANISM IDENTIFICATION

In order to adapt nucleic acid detection to diagnostic tests and thus to identify pathogens on a wide scale, sequences must be found that are conserved enough within a species to identify the diversity of strains that may be encountered in various clinical settings, but divergent enough to distinguish one species from another. Until recently, this problem has been solved for bacteria by targeting the element of a bacterial genome that is most highly conserved within a species, the 16S ribosomal RNA (rRNA) subunit. This method has now been used to confirm *Mycobacterium chimaera* infections in several patients after cardiothoracic surgery, leading ultimately to recognition of a widespread outbreak. At present, 16S PCR amplification from tissue specimens can be performed by specialty laboratories, though its sensitivity and clinical utility to date have remained somewhat limited, in part because of the rarity of pathogen nucleic acid in the sampled tissue, which necessitates reliable, sensitive nucleic acid amplification. As such barriers are reduced through technological advances and as the causes of culture-negative infection are clarified (perhaps in part through sequencing efforts), these tests may become both more accessible and more helpful.

With the wealth of sequencing data now available, other regions beyond 16S rRNA can be targeted for bacterial species identification. These other genomic loci can provide additional information about a clinical isolate that is relevant to patient management. For instance, detection of the presence, or potentially even the expression, of toxin genes such as *C. difficile* toxins A and B or Shiga toxin may provide clinicians with additional information that will help distinguish commensals or colonizing bacteria from pathogens and thus aid in prognostication and management as well as in diagnosis.

While amplification tests such as PCR exemplify one approach to nucleic acid detection, other approaches exist, including detection by hybridization. Although not currently used in the clinical realm, techniques for detection and identification of pathogens by hybridization to microarrays are being developed for other purposes. Of note, these different detection techniques require different degrees of conservation. Highly sensitive amplification methods require a high degree of sequence identity between PCR primer pairs and their short, specific target sequences; even a single base-pair mismatch (particularly near the 3’ end of the primer) may interfere with detection. In contrast, hybridization-based tests are more tolerant of mismatch and thus can be used to detect important regions that may be less precisely conserved within a species, thus potentially allowing detection of clinical isolates from a given species with greater diversity between isolates. Such assays take advantage of the predictable binding interactions of nucleic acids. The applicability of hybridization-based methods toward either DNA or RNA opens up the possibility of expression profiling, which can uncover phenotypic information from nucleic acid content.

Both PCR and hybridization methods target specific, known organisms. At the other extreme, as sequencing costs decline, metagenomic sequencing from patient samples is increasingly feasible. This shotgun sequencing approach is unbiased—i.e., is able to detect any microbial sequence, however divergent or unexpected. In one recent example, a clinical sample of cerebrospinal fluid from an immunocompromised patient with signs and symptoms of chronic meningitis was found through metagenomic sequencing and analysis to contain small amounts of *Leptospira* DNA. In light of this information, retrospective PCR testing confirmed the diagnosis of neuroleptospirosis, which had been missed prior to the sequencing result. The patient was treated with penicillin G and clinically recovered. Increasingly, efforts are under way to bring whole-genome sequencing to other clinical samples, including

sputum and blood, in order to more readily identify pathogens. This new approach brings its own set of challenges, however, including the need to recognize pathogenic sequences against a background of expected host and commensal sequences, and to distinguish true pathogens from either colonizers or laboratory contaminants. The burgeoning field of microbiome research is driving technology development for sequencing and analyzing complex microbial communities. Lessons from this field will inform diagnostic efforts.

PATHOGEN DISCOVERY

In addition to clinical diagnostic applications, novel genomic technologies, including whole-genome sequencing, are being applied to clinical research specimens with a goal of identifying new pathogens in a variety of circumstances. The tremendous sensitivity and unbiased nature of sequencing is also ideal in searching clinical samples for unknown or unsuspected pathogens.

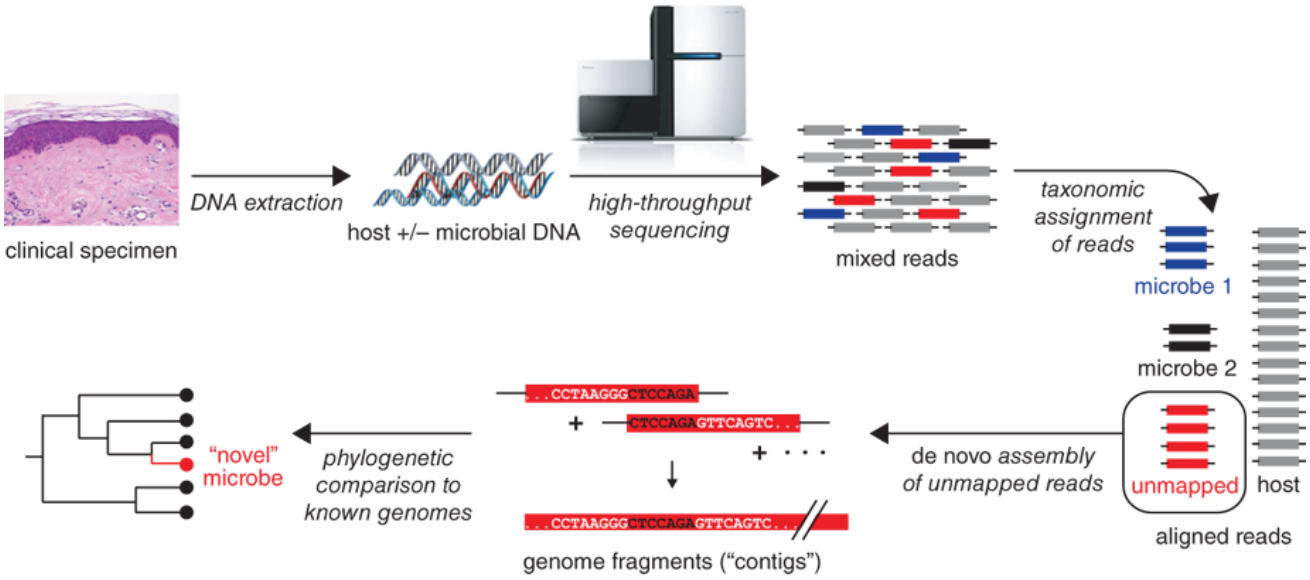
Causal inference in infectious diseases has progressed since the time of Koch, whose historical postulates provided a rigorous framework for attributing a disease to a microorganism. To modernize Koch's postulates: an organism, whether it can be cultured or not, should induce disease upon introduction into a healthy host if it is to be implicated as a causative pathogen. Current sequencing technologies are ideal for advancing this modern version of Koch's postulates because they can identify candidate causal pathogens with unprecedented sensitivity and in an unbiased way, unencumbered by limitations such as culturability. Yet, as direct sequencing on primary patient samples greatly expands our ability to recognize associations between microbes and disease states, critical thinking and experimentation will remain vital in establishing causality.

Virus discovery in particular has been greatly facilitated by new nucleic acid technology. These frontiers were first notably explored with high-density microarrays containing spatially arrayed sequences from a phylogenetically diverse collection of viruses. Despite bias toward those with homology to known viruses, novel viruses in clinical samples were successfully identified on the basis of their ability to hybridize to these prespecified sequences. This methodology famously contributed to identification of the coronavirus causing severe acute respiratory syndrome (SARS). Once discovered, the SARS coronavirus was rapidly sequenced: the full genome was assembled in April 2003, <6 months after recognition of the first case. This accomplishment illustrates the advancing power and speed of new diagnostic technologies.

With the advent of next-generation sequencing, unbiased pathogen discovery is now possible through a process known as *metagenomic assembly* **(Fig. 474-3)**. Sequences of random nucleotide fragments can be generated from clinical specimens with no a priori knowledge of pathogen identity through a process called *shotgun sequencing*. This collection of sequences can then be computationally aligned to host (i.e., human) sequences, with aligned sequences removed and remaining sequences compared with other known genomes to detect the presence of known microorganisms. Sequence fragments that remain unaligned suggest the presence of an additional organism that cannot be matched to a known, characterized genome; these reads can be assembled into contiguous nucleic acid stretches that can be compared with known sequences to construct the genome of a potentially novel organism. Assembled genomes (or parts of genomes) can then be compared with known genomes to infer the phylogeny of new organisms and identify related classes or traits. Thus, not only can this process identify unanticipated pathogens; it can even identify undiscovered organisms. Some early applications of sequencing on clinical samples have centered around the discovery of novel viruses, including such emerging pathogens as West Nile virus, SARS coronavirus, and the Middle East respiratory syndrome coronavirus (MERS-CoV) that has caused severe respiratory illnesses in healthy adults, as well as viral causes of myriad other conditions, from tropical hemorrhagic fevers to diarrhea in newborns.

FIGURE 474-3

Workflow of metagenomic assembly for pathogen discovery. DNA is isolated from a specimen of interest (e.g., tissue, body fluid) containing a mixture of host DNA and nucleic acids from coexisting microbes, either commensal or pathogenic. All DNA (and RNA, if a reverse transcription step is added) is then sequenced, yielding a mixture of DNA sequence fragments ("reads") from the organisms present. Except for reads that do not align ("map") to any known sequence, these reads are aligned to existing reference genomes for the host or any known microbes. The unmapped reads are computationally assembled de novo into the largest contiguous stretches of DNA possible ("contigs"), representing fragments of previously unsequenced genomes. These genome fragments (contigs) are then mapped onto a phylogenetic tree based on their sequence. Some may represent known but as-yet-unsequenced organisms, while others will represent novel species. *(Figure prepared with valuable input from Dr. Ami S. Bhatt, personal communication.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

As metagenomic sequencing and assembly techniques become more robust, this technology holds great promise for identifying microorganisms that are associated with clinical conditions of unknown etiology. Conventional methods already have unexpectedly linked numerous conditions with specific agents of infection—e.g., cervical and oropharyngeal cancers with human papillomavirus (HPV), Kaposi's sarcoma with human herpesvirus 8, and certain lymphomas with Epstein-Barr virus. Recently, Zika virus, first described in the 1940s, was found to be increasing in incidence as a cause of febrile syndromes, particularly in Central and South America. A concurrent increase in the incidence of microcephaly was noted that temporally and geographically matched the Zika epidemics. Zika was suspected to be neurotropic because of a previously recognized association with Guillain-Barré syndrome, but the strongest link between Zika virus and microcephaly came when the virus itself was detected by both quantitative reverse transcription PCR (RT-qPCR) and whole-genome sequencing in postmortem fetal brain tissue from microcephalic infants. An argument for causality was built on the foundation of epidemiologic evidence and direct viral detection, both of which were built on nucleic acid detection and genome sequencing. Sequencing techniques offer unprecedented sensitivity and specificity for identifying foreign nucleic acid sequences that may suggest other such pathogen-associated conditions—from malignancies to inflammatory conditions to unexplained fevers or other clinical syndromes—associated with organisms from viruses to bacteria to parasites. Caution is needed, though: in the absence of the ability to fulfill Koch's postulates, sequence-based identification of a microbe from patient specimens is not, on its own, sufficient to identify a novel pathogen. The increasing sensitivity of these methods warrants greater rigor and care in defining what is "noise" and what represents a pathogen.

As sequencing-based discovery expands, microbes may be found to be associated with conditions not classically thought of as infectious, such as the link between maternal Zika virus infection and fetal microcephaly. Studies of bowel flora in laboratory animals and even humans already suggest correlations between microbe composition and various aspects of metabolic and cardiovascular health. Improved methods for pathogen detection will continue to uncover unexpected correlations between microbes and disease states, but the mere presence of a microbe does not establish causality. Fortunately, once the relatively laborious and computationally intensive metagenomic sequencing and assembly efforts have identified a pathogen, further detection can more easily be undertaken with targeted methods such as PCR or hybridization, which may be more scalable and amenable to in situ confirmation. This capacity should facilitate the additional careful investigation that will be required to progress beyond correlation and to draw causal inference.

ANTIBIOTIC RESISTANCE

At present, antibiotic resistance in bacteria and fungi is conventionally determined by isolating a single colony from a cultured clinical specimen and testing its growth in the presence of drug. The requirement for multiple growth steps in these conventional assays has several consequences. First, only culturable pathogens can be readily processed. Second, this process requires considerable infrastructure to support the sterile environment needed for culture-based testing of diverse organisms. Finally, and perhaps most significantly, even the fastest-growing organisms require 1–2 days of processing for identification and 2–3 days for determination of susceptibilities. Some slow-growing organisms take even longer; for instance, weeks must pass before drug-resistant *M. tuberculosis* can be identified by growth phenotype. Given the clinical imperative in serious illness to begin effective therapy early, this inherent delay in susceptibility determination has obvious implications for empirical antibiotic use: broad-spectrum antibiotics often must be chosen up front in situations where it is later shown that preferred narrower-spectrum drugs would have been effective or even that no antibiotics were appropriate (i.e., in viral infections). Even with this strategy, as resistant organisms become more common, the empirical choice can be incorrect, often with devastating consequences. Real-time identification of the infecting organism and information on its susceptibility profile would guide initial therapy and support judicious antibiotic use, ideally improving patient outcomes while aiding in the ever-escalating fight against antibiotic resistance by reserving the use of broad-spectrum agents for cases in which they are truly needed.

Molecular diagnostics and sequencing offer a way to accelerate detection of a pathogen's antibiotic susceptibility profile. If a genotype that confers resistance can be identified, this genotype can be targeted for molecular detection. In infectious disease, this approach has most convincingly come to fruition for HIV (Fig. 474-2A). (In a conceptually parallel application of genomic analysis, molecular detection of certain resistance determinants in cancers informs selection of targeted chemotherapy.) Extensive sequencing of HIV strains and correlations drawn between viral genotypes and phenotypic resistance have delineated the majority of mutations in key HIV genes, such as reverse transcriptase, protease, and integrase, that confer resistance to the antiretroviral agents that target these proteins. For instance, the single amino acid substitution K103N in the HIV reverse transcriptase gene predicts resistance to the first-line nonnucleoside reverse transcriptase inhibitor efavirenz, and its detection informs a clinician to choose a different agent. The effects of these common mutations on HIV susceptibility to various drugs—as well as on viral fitness—are curated in publicly available databases. Thus, genotypes are now routinely used to predict drug resistance in HIV, as phenotypic resistance assays are far more cumbersome than targeted sequencing. Indeed, the current recommendation in the United States is to sequence virus from a patient's blood before initiating antiretroviral therapy, which is then tailored to the predicted resistance phenotype. As new targeted therapies are introduced, this targeted sequencing–based approach to drug resistance will likely prove important in other viral infections, such as hepatitis C.

The challenge of predicting drug susceptibility from genotype is more daunting for bacteria than for HIV, yet considerable progress has been made toward sequencing-based determination of bacterial antibiotic susceptibility. Bacteria have far more complex genomes than viruses, with thousands of genes on their chromosomes (many of which can functionally interact in ways that escape a priori prediction) and the capacity to acquire many more through horizontal gene transfer of plasmids and mobile genetic elements within and between species. Thus, the task of comprehensively defining all possible genetic resistance mechanisms is orders of magnitude more complex in bacteria than in viruses, which typically have far more limited genomes. Despite these challenges, considerable progress has been made in recent years. In select cases where biological factors appear to have constrained the genotypic basis for resistance to a small, well-defined set of mutations, genotypic assays for antibiotic resistance are already being introduced into clinical practice. One important example is the detection of methicillin-resistant *Staphylococcus aureus* (MRSA). *S. aureus* is one of the most common and serious bacterial pathogens of humans, particularly in health care settings. Resistance to methicillin, the most effective class of antistaphylococcal antibiotics, has become very common, even in community-acquired strains. Vancomycin—the alternative drug to methicillin—is effective against MRSA but is measurably inferior to methicillin against methicillin-susceptible *S. aureus* (MSSA). Analysis of clinical MRSA isolates has demonstrated that the molecular basis for resistance to methicillin in essentially all cases stems from the expression of an alternative penicillin-binding protein (PBP2A) encoded by the gene *mecA*, which is found within a transferable genetic element called *mec*. This mobile cassette has spread rapidly through the *S. aureus* population via horizontal gene transfer and selection from widespread antibiotic use. Because methicillin resistance is essentially always due to the presence of the *mec* cassette, MRSA is particularly amenable to molecular detection. In recent years, a PCR test for the *mec* cassette, which saves hours to days compared with standard culture-based methods, has been approved by the U.S. Food and Drug Administration (FDA). Similar to MRSA, vancomycin-resistant enterococci (VRE) harbor one of a limited number of *van* genes found to be responsible for resistance to this important antibiotic, which occurs through alteration of the mechanism for cell wall cross-linking that vancomycin inhibits. Detection of one of these genes by PCR indicates resistance. More recently, identification of carbapenemase-encoding plasmids responsible for a significant fraction of carbapenem resistance (though not all instances) has led to multiplexed PCR assays to detect this important resistance element to this crucial antibiotic class. Finally, a PCR assay targeting the highly conserved RNA polymerase gene serves not only to detect *M. tuberculosis* directly in sputum samples but also to detect resistance to rifampin, since the determinants of resistance to this RNA polymerase inhibitor map almost exclusively to a short region of this gene. Since rifampin resistance is epidemiologically associated with, though not causal for, multidrug resistance, this assay identifies strains at high risk for multidrug resistance, enhancing its value.

Although identification and rapid detection of monogenic resistance determinants have improved, bacteria have tended to evolve multiple, diverse resistance mechanisms to most antibiotics; therefore, these tasks often require probing for and integration of multiple genetic lesions, targets, or mechanisms. For instance, at least five distinct modes of resistance to fluoroquinolones are known: reduced import, increased efflux, target site mutation, drug modification, and shielding of the target sites by expression of another protein. These mechanisms are typically present in combination in clinically resistant isolates; thus the problem of detecting genetic resistance is often a combinatorial one. In another clinically important example, while carbapenem resistance in Enterobacteriaceae is often explained by the presence of carbapenemases, resistance may also develop when other, less broad-spectrum β-lactamases are found in combination with porin mutations or efflux pumps. Thus, while multiplexed PCR assays for the most common carbapenemases (e.g., those encoded by the KPC, NDM, OXA-48, IMP-1, and VIM genes) have become a valuable tool for rapid identification of the subset of carbapenem-resistant Enterobacteriaceae in which resistance is caused by carbapenemases, their sensitivity is limited by their inability to detect other mechanisms of carbapenem resistance. To further complicate genetic prediction, changes in gene expression (which may be detectable through mutations in promoter regions or regulatory genes without coding mutations in known resistance determinants) and even gene copy number (which may occur without changes in primary sequence) of resistance determinants play critical roles in some cases of genetic resistance. Thus, while predicting resistance when determinants are found is rapidly becoming feasible, the more clinically relevant task of predicting *susceptibility* when no known resistance determinants are found remains more difficult.

To build on early successes with the goal of advancing beyond binary detection of monogenic resistance determinants, the ultimate frontier for genetic prediction of bacterial antibiotic resistance lies in more comprehensive prediction of a resistance phenotype from sequence information—a task similar to HIV resistance prediction. Yet there is no comprehensive compendium of genetic elements conferring resistance and their pairwise and higher-order interactions with each other and with the genetic background of bacterial pathogens. Nonviral genomes are much larger than viral ones, and their abundance and diversity are such that thousands of genetic differences often exist between clinical isolates of the same species, of which perhaps only one or a few may contribute to resistance. In addition, new mechanisms may emerge in the face of antibiotic deployment or with the release of new drugs, and

genetic prediction of resistance will inevitably lag behind the emergence of unforeseen mechanisms. While confident prediction of bacterial antibiotic resistance from sequencing determinants may therefore seem daunting, the vast expansion of microbial sequencing capacity (Fig. 474-1), combined with analytic methods such as microbial genome-wide association studies and machine learning algorithms, offers powerful analytical approaches to this "needle in a haystack" problem and has permitted remarkable advances in the predictive power of sequence determinants to date. Particularly in *M. tuberculosis*, where horizontal gene transfer is minimal and the pathogen is essentially restricted to human hosts, a remarkably wide array of phenotypic resistance can be explained by known genetic determinants. Even in more highly variable pathogens, with sequencing of sufficient numbers of susceptible and resistant pathogens, sequence-based prediction methods are improving in predictive accuracy, at least within the geographic region from which the test samples have been sequenced.

It is important to note that genotype-based analytical methods largely identify correlates, not necessarily surrogates or determinants, of resistance. In HIV diagnostics, surrogates (i.e., causal determinants of resistance) were found to be more reliable predictors than mere correlates in expanding sequencing-based resistance prediction to the general population. Without a mechanistic understanding of genetic resistance, a correlative relationship may be lineage-specific and less generalizable. Especially with multiple possible mechanisms of resistance to a given antibiotic and ongoing evolutionary pressure resulting in the development and acquisition of new modes of resistance, a genotypic approach to diagnosing antibiotic resistance is likely to remain challenging and to require ongoing vigilance in constantly correlating genotypic with more traditional phenotypic methods. An important corollary benefit of a genomic approach to resistance prediction, anchored in phenotypic validation, could be the systematic identification of outliers with unexplained resistance. These strains can form the basis for understanding newly emerging resistance mechanisms, which can in turn inform new drug development endeavors. Understanding resistance mechanisms may also help direct infection control efforts. For instance, the first identification of the *mcr-1* (mobilized colistin resistance) gene on a plasmid, together with other antibiotic resistance determinants, heightened concern about colistin-resistant Enterobacteriaceae identified first in China and later elsewhere because it implied rapid transmissibility of multidrug resistance. Early recognition of these potentially dangerous strains elucidated the immediate need for strict containment protocols.

In parallel with advancing sequencing technologies, progress in computational techniques, bioinformatics and statistics, and data storage as well as experimental confirmatory testing of hypotheses will be needed to advance toward the ambitious goal of a comprehensive compendium of global antibiotic resistance determinants. Open sharing and careful curation of new sequence information will be of paramount importance, as will iterative or even continuous comparison of predictions with ongoing phenotypic testing in order to assess performance and allow prediction algorithms to keep up with newly evolving or emerging resistance mechanisms.

We continuously observe the accumulation of new or unanticipated modes of resistance from ongoing evolutionary pressure caused by the widespread clinical use of antibiotics. Even with MRSA, perhaps the best-studied case of antibiotic resistance and a model of relative simplicity with a single known monogenic resistance determinant (*mecA*), a genotype-based approach to resistance detection proved imperfect. One limitation was a recall of the initial commercial genotypic resistance assay that was deployed for the identification of MRSA. A clinical isolate of *S. aureus* that emerged in Belgium expressed a variant of the *mec* cassette not detected by the assay's PCR primers. New primers were added to detect this new variant, and the assay was re-approved for use. This example illustrates the need for ongoing monitoring of any genotypic resistance assay. A second limitation is that a contradiction can occur between genotypic and phenotypic evidence for resistance. Up to 5% of MSSA strains have been reported to carry a copy of the *mecA* gene that is either nonfunctional or not expressed. Thus, the erroneous identification of these strains as MRSA by genotypic detection would lead to administration of the inferior antibiotic vancomycin rather than the preferred β-lactam therapy.

These examples illustrate one of the prime challenges of moving beyond growth-based assays: genotype is merely a proxy for the resistance phenotype that directly informs patient care. Alternative approaches currently under development attempt to circumvent the limitations of genotypic resistance testing by returning to phenotypic assays, albeit more rapid ones. One such approach is informed by genomic methods: transcriptional profiles serve as a rapid phenotypic signature for antibiotic response. Conceptually, since dying cells are transcriptionally distinct from cells fated to survive, susceptible bacteria enact different transcriptional profiles after antibiotic exposure than resistant ones, independent of the mechanism of resistance. These differences can be measured and, since transcription is one of the most rapid responses to cell stress (minutes to hours), can be used to determine whether cells are resistant or susceptible much more rapidly than is possible if growth in the presence of antibiotics is awaited (days). Like DNA, RNA can be readily detected through predictable rules governing base pairing via either amplification or hybridization-based methods. Changes in a carefully selected set of transcripts form an expression signature that can represent the total cellular response to antibiotic without requiring full characterization of the entire transcriptome. Preliminary proof-of-concept studies suggest that this approach may identify antibiotic susceptibility on the basis of transcriptional phenotype much more quickly than is possible with growth-based assays.

Because of its sensitivity in detecting even very rare nucleic acid fragments, sequencing provides an unprecedented depth of study into complex populations of cells and tissues. The strength of this depth and sensitivity applies not only to the detection of rare, novel pathogens in a sea of host signal, but also to the identification of heterogeneous pathogen subpopulations in a single host that may differ, for example, in drug resistance profiles or pathogenesis determinants. For instance, recent studies have highlighted the diversification of pathogens in chronic bacterial infections, such as *Pseudomonas* in the lungs of patients with cystic fibrosis or *M. tuberculosis* in disseminated infection, perhaps allowing for niche specialization within the host. Such diversification has long been recognized in chronic viral populations, as exemplified by HIV. Future studies will be needed to elucidate the clinical significance of these variable subpopulations, even as deep sequencing is now providing unprecedented levels of detail about majority and minority members of this population.

HOST-BASED DIAGNOSTICS

While pathogen-based diagnostics continue to be the mainstay for confirming infection, serologic testing and nonspecific biomarkers—such as erythrocyte sedimentation rate, C-reactive protein level, and even total white blood cell and neutrophil counts—have long been the basis of a strategy for measuring host responses to aid in the diagnosis of infection. Even recently identified host biomarkers of bacterial infection, such as procalcitonin, have fallen short in their versatility, with positive and negative predictive values that are thus far adequate for only a few narrow applications but inadequate for generalized clinical use. Here, too, the application of genomics is now being explored to improve upon this approach, given the previously described limitations of serologic testing and the lack of specificity of protein biomarkers identified to date. Rather than using antibody responses as a retrospective biomarker for infection, recent efforts have focused on transcriptomic analysis of the host response as a new direction with diagnostic implications for human disease. For instance, while pathogen-based diagnostic tests to distinguish active from latent tuberculosis infection have proven elusive, recent work shows that the transcriptional profile of circulating white blood cells exhibits a differential pattern of expression of nearly 400 transcripts that distinguish active from latent tuberculosis; this expression pattern is driven in part by changes in interferon-inducible genes in the myeloid lineage. In a validation cohort, this transcriptional signature was able to distinguish patients with active versus latent disease, to distinguish tuberculosis infection from other pulmonary inflammatory states or infections, and to track responses to treatment in as little as 2 weeks, with normalization of expression toward that of patients without active disease over 6 months of effective therapy. Such a test could play an important role not only in the management of patients but also as a marker of efficacy in clinical trials of new therapeutic agents. Similarly, considerable progress has been made toward identifying host transcriptional signatures in circulating blood cells that distinguish viral from bacterial causes of upper respiratory infection, with better performance characteristics than current clinical parameters or available protein biomarkers. Additional host signatures have been reported that distinguish among bacterial infection, viral infection, and inflammatory states; identify Lyme disease; identify influenza; and even distinguish between gram-positive and gram-negative bacterial infections. In some cases, results have been extended to different host populations—including adults and children, and those with varying immune function—which obviously will be critical for generalizing such an approach. Thus, profiling of host transcriptional dynamics could augment the information obtained from studies of pathogens, both enhancing diagnosis and monitoring the progression of illness and the response to therapy.

In this era of genome-wide association studies and attempts to move toward personalized medicine, genomic approaches are also being applied to the identification of host genetic loci and factors that contribute to infection susceptibility. Such loci will have undergone strong selection among populations in which the disease is endemic. Through identification of the beneficial genetic alleles among individuals who survive in such settings, markers for susceptibility or resistance are being discovered; these markers can be translated to diagnostic tests to identify susceptible individuals in order to implement preventive or prophylactic interventions. Further, such studies may offer mechanistic insight into the pathogenesis of infection and inform new methods of therapeutic intervention. Such beneficial genetic associations were recognized long before the advent of genomics, as in the protective effects of the negative Duffy blood group or heterozygous hemoglobin abnormalities against *Plasmodium* infection. Genomic approaches allow more systematic and widespread application of this principle to identify not only people with increased susceptibility to prevalent diseases (e.g., HIV infection, tuberculosis, and cholera) but also host factors that contribute to and thus might predict the severity of disease.

THERAPEUTICS

Genomics has the potential to impact infectious disease therapeutics in two ways. By transforming the speed or type of diagnostic information that can be attained, it can influence therapeutic decision-making. Alternatively, by opening new avenues to a better understanding of pathogenesis, providing new ways to disrupt infection, and delineating new approaches to antibiotic discovery, it has the potential to facilitate the development of new therapeutic agents.

GENOMIC DIAGNOSTICS INFORMING THERAPEUTICS

Efforts at antibiotic discovery are declining, with few new agents in the pipeline and even fewer new drugs (in particular, few agents with new mechanisms of action) entering the market. This phenomenon is due in part to the lack of economic incentives for the private sector; however, it is also attributable in part to the enormous challenges involved in the discovery and development of antibiotics. Most recent efforts have focused on broad-spectrum antibiotics; the development of a chemical entity that works across an extremely diverse set of organisms (i.e., species more divergent from each other than a human is from an amoeba) is far more challenging than the development of an agent that is designed to target a single bacterial species. Nevertheless, the concept of narrow-spectrum antibiotics has heretofore been rejected because of the lack of early diagnostic information that would guide the selection of such agents. Thus, rapid diagnostics providing antibiotic susceptibility information that can guide antibiotic selection in real time has the potential to alter and simplify antibiotic strategies by allowing a paradigm shift away from broad-spectrum drugs and toward narrow-spectrum agents. Such a paradigm shift clearly would have additional implications for antibiotic resistance, helping to limit selective pressure applied to pathogens and commensal bacteria during therapy.

In yet another diagnostic paradigm with the potential to impact therapeutic interventions, genomics is opening new avenues to a better understanding not only of different host susceptibilities to infection but also of different host responses to therapy. For example, the role of glucocorticoids in tuberculous meningitis has long been debated. Recently, polymorphisms in the human genetic locus *LTA4H*, which encodes a leukotriene-modifying enzyme, were found to modulate the inflammatory response to tuberculosis. Patients with tuberculous meningitis who were homozygous for the proinflammatory *LTA4H* allele were most helped by adjunctive glucocorticoid treatment, while those who were homozygous for the anti-inflammatory allele were negatively affected by steroid treatment. Steroids have become part of the standard of care in tuberculous meningitis, but this study suggests that perhaps only a

subset of patients benefit from this anti-inflammatory adjunct (while others may be harmed) and further suggests a genetic means of prospectively identifying this subset. Thus, genomic diagnostic tests may eventually approach the goal of personalized medicine, informing diagnosis, prognosis, and treatment decisions by revealing the pathogenic potential of the microbe and by detecting individualized host responses to both infection and therapy.

GENOMICS IN DRUG AND VACCINE DEVELOPMENT

Genomic technologies are dramatically changing research on host–pathogen interactions, with a goal of increasingly influencing the process of therapeutic discovery and development. Sequencing offers several possible avenues into antimicrobial therapeutic discovery. First, genome-scale molecular methods have paved the way for comprehensive identification of all essential genes encoded by a pathogen, thereby systematically identifying critical vulnerabilities within a pathogen that could be targeted therapeutically. Second, genome-scale methodologies offer rapid ways to address the mechanism of action of newly identified hits from compound screens. Whole-genome sequencing offers a rapid, unbiased way to detect mutations arising in resistant mutants during selection. Similarly, transcriptional profiling can provide insights into mechanisms of action of new candidate drugs. For instance, the transcriptional signature of cell wall disruptors (e.g., β-lactams) is distinct from that of DNA-damaging agents (e.g., fluoroquinolones) or protein synthesis inhibitors (e.g., aminoglycosides). Either approach can thus suggest a mechanism of action or flag compounds for prioritization because of a potentially novel activity. In an alternative genomic strategy for determining mechanisms of action, an RNA interference approach followed by targeted sequencing was used to identify genes required for antitrypanosomal drug efficacy. This approach provided new insights into the mechanism of action of drugs that have been in use for decades for human African trypanosomiasis. Third, sequencing can readily identify the most conserved regions of a pathogen's genomes and corresponding gene products; this information is invaluable in narrowing antigen candidates in vaccine development. These surface proteins can be expressed recombinantly and tested for the ability to elicit a serologic response and protective immunity. This process, termed *reverse vaccinology*, has proved particularly useful for pathogens that are difficult to culture or poorly immunogenic.

Genomics has been employed in both developing vaccines and defining their impact on microbial epidemiology and ecology. Examples include recent studies of influenza, malaria, *S. pneumoniae*, and HPV following vaccine introduction. Extensive sequencing of influenza viruses has been valuable in understanding the modest efficacy of seasonal influenza vaccination, and the combination of genomics and antigenic cartography is proving helpful in the selection of strains to include in subsequent influenza vaccines. The new RTS,S/AS01 malaria vaccine was analyzed by targeted sequencing of parasites from vaccinated and control populations during a phase 3 trial conducted at 11 sites in Africa; these analyses revealed reduced vaccine efficacy against parasites with amino acid mutations in the circumsporozoite protein targeted by the vaccine. Similarly, studies of the more established pneumococcal vaccine (the 7-valent polysaccharide conjugate vaccine, PCV-7) documented serotype replacement: strains targeted by the vaccine have dramatically decreased in prevalence following widespread vaccination campaigns. Given that specific serotypes of HPV (e.g., types 16 and 18) clearly are more strongly associated than others with carcinogenesis, HPV vaccines have capitalized on serotype replacement, targeting vaccine strains to specifically prevent infection with the more dangerous serotypes. Such a strategy, informed by pathogen genomics, aims to protect individuals and ideally to decrease the circulating burden of more virulent strains within society.

Large-scale gene content analysis from sequencing or expression profiling enables new research directions that provide novel insights into the interplay of pathogen and host during infection or colonization. One important goal of such research is to suggest new therapeutic approaches to disrupt this interaction in favor of the host. Indeed, one of the most immediate applications of next-generation sequencing technology has come from simply characterizing human pathogens and related commensal or environmental strains and then finding genomic correlates for pathogenicity. For instance, as *Escherichia coli* varies from a simple nonpathogenic, lab-adapted strain (K-12) to a Shiga toxin–producing enterohemorrhagic gastrointestinal pathogen (O157:H7), it displays up to a 25% difference in gene content, though it is classified as the same species. Similarly, some isolates of *Enterococcus*—a genus notorious for its increasing incidence of resistance to common antibiotics such as ampicillin, vancomycin, and aminoglycosides—also contain recently acquired genetic material comprising up to 25% of the genome on mobile genetic elements. This fact suggests that horizontal gene transfer plays an important role in the organisms' adaptation as nosocomial pathogens. On closer study, this genome expansion is associated with loss of regulatory elements called CRISPRs (*c*lustered, *r*egularly *i*nterspaced *s*hort *p*alindromic *r*epeats). Loss of CRISPR elements, which protect the bacterial genome from invasion by certain foreign genetic material, may thus facilitate the acquisition of antibiotic resistance–conferring genetic elements. While loss of this regulation appears to impose a competitive disadvantage in antibiotic-free environments, these drug-resistant strains thrive in the presence of even some of the best antienterococcal therapies. In addition to insights gained from genome sequencing, extension of unbiased whole-transcriptome sequencing (RNA-Seq) efforts to bacteria is beginning to identify unexpected regulatory, noncoding RNAs in many diverse species. While the functional implications of these new transcripts are as yet largely unknown, the presence of such features—conserved across many bacterial species—implies evolutionary importance and suggests areas for future study and possible new therapeutic avenues. Transcriptomic and proteomic profiling of pathogens under various conditions that mimic colonization or infection, including existence as biofilms or in polymicrobial communities, intracellular infection models, antibiotic exposure, and nutrient starvation, has begun to reveal novel biological features that may be targeted by the next generation of therapies. At the cutting edge of the host–pathogen interface, single-cell transcriptomic methodologies are rapidly increasing in feasibility and extent, revealing previously unknown heterogeneity in the potential outcomes of intracellular infection.

Thus, genomic studies are transforming our understanding of infection, offering evidence of virulence factors or toxins and providing insight into ongoing evolution of pathogenicity and drug resistance. One goal of such studies is to identify therapeutic agents that can disrupt the pathogenic process. There is currently much interest in the theoretical concept of antivirulence drugs that inhibit virulence factors rather than killing the pathogen outright as a means to intervene in infection. Further, with sequencing ever more accessible and efficient, ongoing large-scale studies have unprecedented statistical power to associate clinical outcomes with pathogen and host genotypes and thus to further reveal vulnerabilities in the infection process that can be targeted for

disruption. Although this is just the beginning, such studies point to a tantalizing future in which the clinician is armed with genomic predictors of infection outcome and therapeutic response to guide clinical decision-making.

EPIDEMIOLOGY OF INFECTIOUS DISEASES

Epidemiologic studies of infectious diseases have several main goals: to identify and characterize outbreaks, to describe the pattern and dynamics of an infectious disease as it spreads through populations, and to identify interventions that can limit or reduce the burden of disease. One classic, paradigmatic example is John Snow's elucidation of the origin of the 1854 London cholera outbreak. Snow used careful geographic mapping of cases to determine that the likely source of the outbreak was contaminated water from the Broad Street pump, and, by removing the pump handle, he aborted the outbreak. Whereas that effort was undertaken without knowledge of the causative agent of cholera, advances in microbiology and genomics have expanded the purview of epidemiology to consider not just the disease but also the pathogen, its virulence factors, and the complex relationships between microbial and host populations.

Through use of genomic tools such as high-throughput sequencing, the diversity of a microbial population can be rapidly described with unprecedented resolution, with discrimination between isolates that have single-nucleotide differences across the entire genome and advancement beyond prior approaches that relied on phenotypes (such as antibiotic susceptibility profiles) or genetic markers (such as multilocus sequence typing). The development of statistical methods grounded in molecular genetics and evolutionary theory has established analytical approaches that translate descriptions of microbial population diversity and structure into descriptions of the origin and history of pathogen spread. By linking phylogenetic reconstruction with epidemiologic and demographic data, genomic epidemiology presents the opportunity to track transmission from person to person and across demographic and geographic boundaries, to infer transmission patterns of both pathogens and sequence elements that confer phenotypes of interest, and to estimate the transmission dynamics of outbreaks.

TRANSMISSION NETWORKS

Whole-genome sequencing of pathogen genomes can be used to infer transmission and identify point-source outbreaks. As reported in a seminal paper in 2010, a study of MRSA in a Thai hospital demonstrated the use of whole-genome sequencing in reconstructing the transmission of a pathogen from patient to patient by integrating the analysis of accumulation of mutations over time with the dates and hospital locations of the infected individuals. Since then, multiple instances of the use of whole-genome sequencing to define and motivate interventions aimed at interrupting transmission chains have been reported. In another MRSA outbreak in a special-care baby unit in Cambridge, United Kingdom, whole-genome sequencing extended the traditional infection control analysis, which relies on typing organisms by their antibiotic susceptibilities, to sequencing of isolates from clinical samples. This approach identified an otherwise unrecognized outbreak of a specific MRSA strain that was occurring against a background of the usual pattern of infection caused by a diverse circulating population of MRSA strains. The analysis showed evidence of transmission among mothers within the special-care baby unit and in the community and demonstrated the key role of MRSA carriage in a single health care provider in the persistence of the outbreak. In yet another example, in response to the observation of 18 cases of infection by carbapenemase-producing *Klebsiella pneumoniae* over 6 months at the National Institutes of Health Clinical Research Center, genome sequencing of the isolates was used to discriminate between the possibilities that these cases represented multiple, independent introductions into the health care system or a single introduction with subsequent transmission. On the basis of network and phylogenetic analysis of genomic and epidemiologic data, the authors reconstructed the likely relationships among the isolates from patient to patient, demonstrating that the spread of resistant *Klebsiella* infection was in fact due to nosocomial transmission of a single strain. Similar approaches have elucidated the extent to which presumed nosocomial *C. difficile*, VRE, and carbapenem-resistant Enterobacteriaceae represent within-hospital transmission rather than independent acquisitions. With these demonstrations of the potential contribution of genomics to hospital infection-control efforts, an important avenue of research seeks to develop statistical methods with which to ascertain when such tools are useful and their cost-effectiveness when compared with that of current nongenomic approaches.

The uncovering of unexpected transmission events by genomic epidemiology studies is motivating investigations into pathogen ecology and modes of transmission. For example, the rise in prevalence of infections with nontuberculous mycobacteria, including *Mycobacterium abscessus*, among patients with cystic fibrosis has led to speculation about the possible role of patient-to-patient transmission in the cystic fibrosis community; however, conventional typing approaches have lacked the resolution to define pathogen population structure accurately, a critical component of inferring transmission. Past infection-control guidelines discounted the possibility of acquisition of nontuberculous mycobacteria in health care settings, as no strong evidence for such transmission had been described. In whole-genome sequencing studies of *M. abscessus* isolates from patients with cystic fibrosis, an analytical approach using genome sequencing, epidemiology, and Bayesian modeling revealed that, contrary to the prior belief that infections with *M. abscessus* are independently acquired, the majority of infections appear to be transmitted. Because there are often no clear epidemiologic links that place the infected patients in the same place at the same time, this finding highlights a need to explore preexisting notions of circumstances required for transmission, including the roles of fomites and aerosols, and a reconsideration of *M. abscessus* infection-control guidelines. In a clear example of the utility of whole-genome sequencing for revealing unexpected transmission networks, isolates of *M. chimaera* causing infections after cardiothoracic surgery in patients in different locations were all found to be closely related. These isolates differed from one another by at most 38 pairwise single-nucleotide polymorphisms out of >5 million bases; in contrast, they differed by >2900 single-nucleotide polymorphisms from the nonclonally related reference isolate. Although a hospital source was initially suspected when the first of these cases were identified, this whole-genome sequencing analysis strongly supported a single point-source for these geographically dispersed isolates. A subsequent investigation ultimately implicated *M. chimaera* contamination in the

manufacturing chain of a temperature-control system used during cardiac bypass. Similar studies of other pathogens—particularly those that share human, other animal host, and environmental reservoirs—will continue to advance our understanding of the relative roles and prominence of sources of infection and the modes of spread through populations, thereby establishing evidence-based strategies for prevention and intervention.

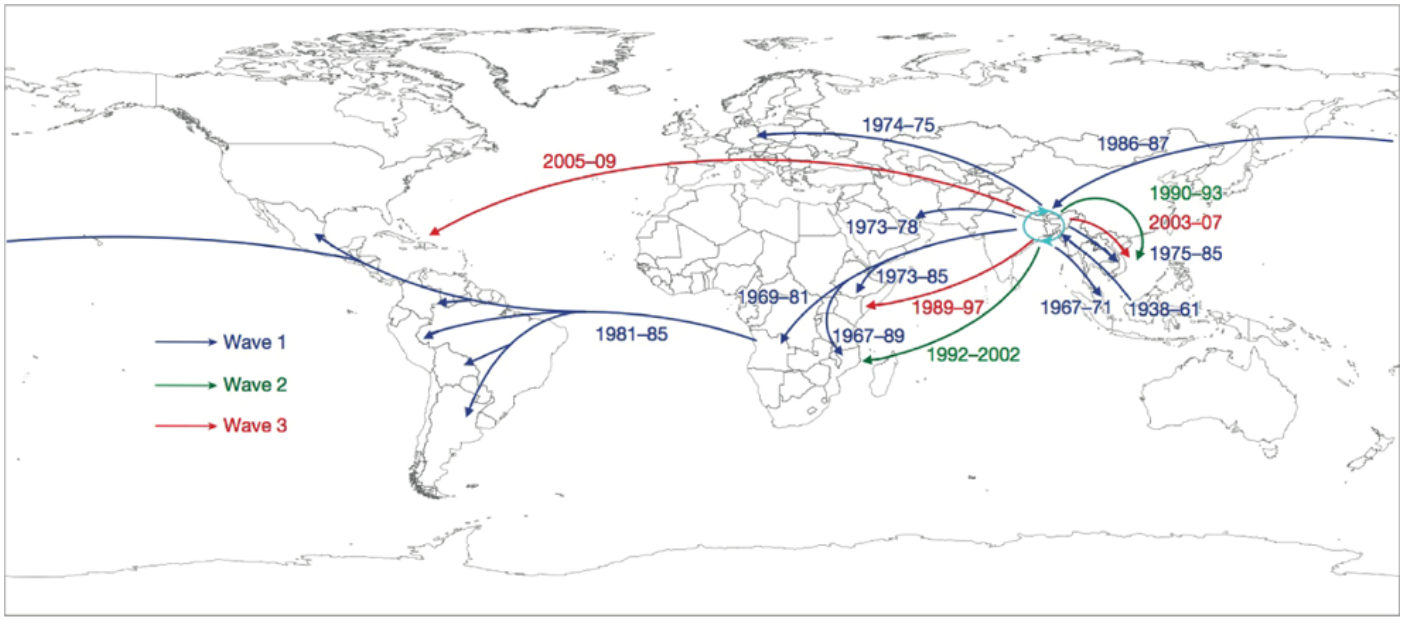
As more studies aim to carefully define the origins and spread of infectious agents using the high-resolution lens of whole-genome sequencing, fundamental questions arise about the diversity of infecting and colonizing microbial populations. Traditional microbiologic methods include taking a single colony from a growth plate as representative of the population. However, the more diverse the colonizing or infecting pathogen population, the less representative these individual isolates are and the greater the possibility for introducing error into whole-genome sequencing–based methods while reconstructing transmission. Sequencing studies of multiple colonies of an *S. aureus* strain colonizing a single individual showed a "cloud" of diversity. What is the clinical significance of this diversity? What are the processes that generate and limit it? What amount of diversity is transmitted under different conditions and routes of transmission? How do the answers to these questions vary by infectious organism, type of infection, host, and response to treatment? More comprehensive descriptions of diversity, population dynamics, transmission bottlenecks, and the forces that shape and influence the growth and spread of microbial populations will be a critically important focus of future investigations.

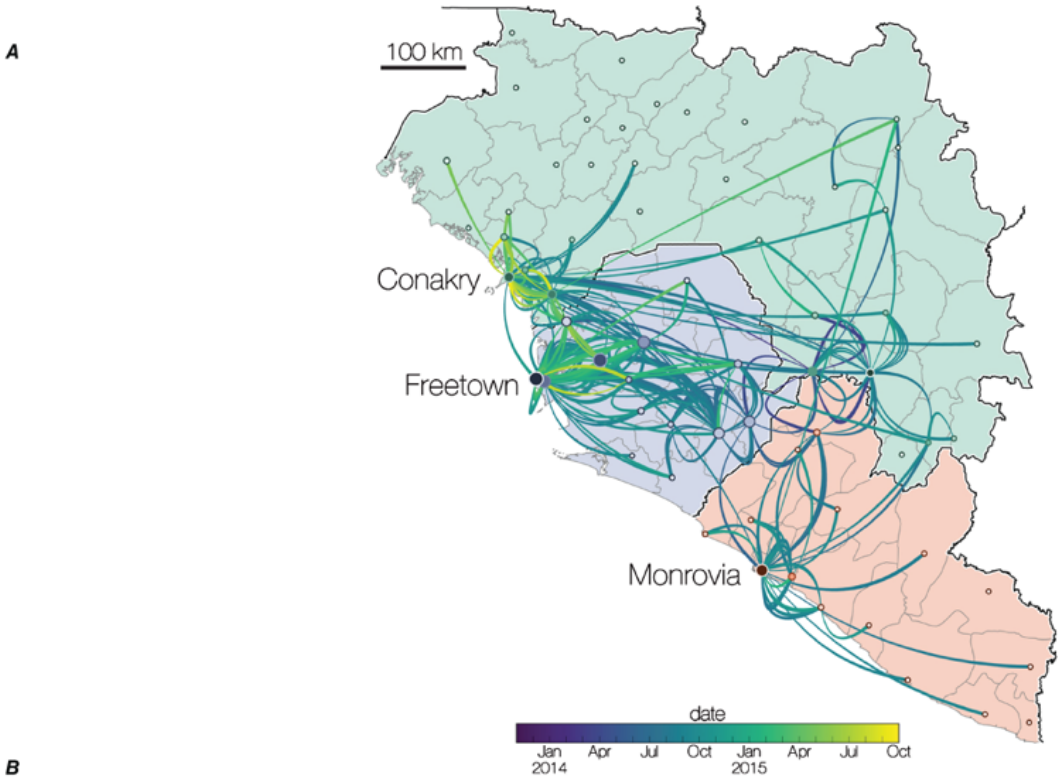
ORIGINS AND DYNAMICS OF PATHOGEN SPREAD

In addition to reconstructing the transmission chains of local outbreaks, genomics-based epidemiologic methods reveal broad-scale geographic and temporal spread of pathogens. Three recent examples include the origins of cholera in Haiti, the history of HIV-1 group M, and the spread of Ebola in West Africa. Cholera, a dehydrating diarrheal illness caused by infection with *Vibrio cholerae*, first spread worldwide from the Indian subcontinent in the 1800s and has since caused seven pandemics; the seventh pandemic has been ongoing since the 1960s. An investigation into the geographic patterns of cholera spread in the seventh pandemic used genome sequences from a global collection of 154 *V. cholerae* strains representing isolates from 1957 to 2010. This investigation revealed that the seventh pandemic has comprised at least three overlapping waves spreading out from the Indian subcontinent **(Fig. 474-4A)**. Further, analysis of the genome of an isolate of *V. cholerae* from the 2010 outbreak of cholera in Haiti showed it to be more closely related to isolates from South Asia than to isolates from neighboring Latin America, supporting the hypothesis that the outbreak was derived from *V. cholerae* introduced into Haiti by human travel (likely from Nepal) rather than by environmental or more geographically proximal sources. A subsequent study that dated the time to the most recent common ancestor of a population of *V. cholerae* isolates from Haiti provided further support for a single point-source introduction from Nepal. Application of similar methods that integrate pathogen genome sequences, mutation rates, geographic locations, and phylogenetic inference to HIV-1 group M dated the origin of the virus to the 1920s and the city of Kinshasa (then called Leopoldville), the capital of the Democratic Republic of the Congo (then called the Belgian Congo). This work established an understanding of how a boom in industry and a city with extensive railroad connections provide a scaffolding along which a virus can rapidly spread geographically.

FIGURE 474-4

A. Transmission events inferred from phylogenetic reconstruction of 154 *Vibrio cholerae* isolates from the seventh cholera pandemic. Date ranges represent estimated time to the most recent common ancestor for strains transmitted from source to destination locations, based on a Bayesian model of the phylogeny. *(Reprinted with permission from A Mutreja et al: Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462, 2011.)* ***B.*** Inferred Ebola virus spread in West Africa (Liberia, *red*; Guinea, *green*; and Sierra Leone, *blue*) by phylogeographic methods using virus genome sequences, dates, and an evolutionary model. The lines reflect spread between population centroids of each administrative region, going from the thin end to the thick end and colored by a time scale. *(Reprinted with permission from G Dudas et al: Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544:309, 2017.)*





Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

More recently, genome sequencing has proven invaluable in understanding the geographic, demographic, climatic, and administrative factors that drove, sustained, and limited the 2013–2016 Ebola outbreak that ravaged West Africa (Fig. 474-4B) as well as the factors and patterns of transmission of Zika virus in the Americas. These efforts illustrate the remarkable promise of genome sequencing in improving outbreak response strategy by elucidating previously hidden paths of disease spread and details of the forces that shape epidemics. The combination of in-the-field sequencing with portable sequencing platforms, rapid data sharing, and rapid open analysis through sites such as nextstrain.org offers a paradigm by which real-time genomic epidemiology may contribute to "weather maps," enabling prediction of epidemic patterns and thus providing guidance for public health interventions to slow or control their spread.

Increasing numbers of investigations into the spread of many pathogens are contributing to a growing atlas of maps describing routes, patterns, and tempos of microbial diversification and dissemination, not just for agents of emerging infectious diseases but for common pathogens as well. Such studies will create a vast amount of data that can be used to investigate the diversity and microbiologic links within distinct niches and the patterns of spread from one niche to another. The increasingly broad adoption of genome sequencing by health care and public health institutions will ensure that the available catalog of genome sequences and associated epidemiologic data will grow very rapidly. For example, updating from the pulsed-field gel electrophoresis techniques that have been used to define strains of food-borne pathogens since the late 1980s, PulseNet—the U.S. Centers for Disease Control and

Prevention network for monitoring these pathogens—is instituting routine genome sequencing. With higher-resolution description of microbial diversity and of the dynamics of that diversity over time and across epidemiologic and demographic boundaries and evolutionary niches, we will gain even greater insights into the relationships of transmission routes and patterns of historical spread.

EPIDEMIC POTENTIAL

Defining pathogen transmissibility is a critical step in the development of public health surveillance and intervention strategies as this information can help to predict the epidemic potential of an outbreak. Transmissibility can be estimated by a variety of methods, including inference from the growth rate of an epidemic and the generation time of an infection (the mean interval between infection of an index case and infection of the people infected by that index case). Genome sequencing and analysis of a well-sampled population provide another method by which to derive similar fundamental epidemiologic parameters. One key measure of transmissibility is the basic reproduction number, defined as the number of secondary infections generated from a single primary infectious case. When the basic reproduction number is >1, an outbreak has epidemic potential; when it is <1, the outbreak will become extinct. On the basis of sequences from influenza virus samples obtained from infected patients very early in the 2009 H1N1 influenza pandemic, the basic reproduction number was estimated through a population genomic analysis at 1.2; this result provided greater confidence to estimates derived by traditional epidemiologic data, which ranged from 1.4 to 1.6. In addition, with the assumption of a molecular clock model, sequences of H1N1 samples together with information about when and where the samples were obtained have been used to estimate the date and location of the pandemic's origin, providing insight into disease origins and dynamics. Because the magnitude and intensity of the public health response are guided by the predicted size of an outbreak, the ability of genomic methods to cast light on a pathogen's origin and epidemic potential adds an important dimension to the contributions of these methods to infectious disease epidemiology.

PATHOGEN EVOLUTION

Beyond describing transmission and dynamics, pathogen genomics can provide insight into the evolution of pathogens and the interactions of selective pressures, the host, and pathogen populations, which can have implications for clinical decision-making and the development of vaccines and therapeutics. From a clinical perspective, this process is central to the acquisition of antibiotic resistance, the generation of increasing pathogenicity or new virulence traits, the evasion of host immunity and clearance (leading to chronic infection), and vaccine efficacy.

Microbial genomes evolve through a variety of mechanisms, including mutation, duplication, insertion, deletion, recombination, and horizontal gene transfer. Segmented viruses (e.g., influenza virus) can reassort gene segments within multiply infected cells. The pandemic 2009 H1N1 influenza A virus, for example, appears to have been generated through reassortment of several avian, swine, and human influenza strains. Such potential for the evolution of novel pandemic strains has precipitated concern about the possible evolution to transmissibility of virulent strains that have been associated with high mortality rates but have not yet exhibited efficient human infectivity. Experiments with H5N1 avian influenza, for example, have defined five mutations that render it transmissible, at least in ferrets—the animal model system for human influenza. Studies that examine the genomes of pathogens collected longitudinally from individual infections have similarly demonstrated the evolution of bacteria as they adapt from colonization to invasion and to new host environments and new immune and therapeutic pressures.

The continuous antigenic evolution of seasonal influenza offers an example of how studies of pathogen evolution can impact surveillance and vaccine development. Frequent updates to the annual influenza vaccine are needed to ensure protection against the dominant strains. These updates are based on anticipating which viral populations from a pool of substantial locally and globally diverse circulating viruses will predominate in the upcoming season. Toward that end, sequencing-based studies of influenza virus dynamics have shed light on the global spread of influenza, providing concrete data on patterns of spread and helping to elucidate the origins, emergence, and circulation of novel strains. Through analysis of more than 1000 influenza A H3N2 virus isolates over the 2002–2007 influenza seasons, Southeast Asia was identified as the usual site from which diversity originates and spreads worldwide. Further studies of global isolate collections have shed further light on the diversity of circulating virus, showing that some strains persist and circulate outside of Asia for multiple seasons.

Not only do genomic epidemiology studies have the potential to help guide vaccine selection and development; they are also helping to track what happens to pathogens circulating in the population in response to vaccination. By describing pathogen evolution under the selective pressure of a vaccinated population, such studies can play a key role in surveillance and identification of virulence determinants and perhaps may even help to predict the future evolution of escape from vaccine protection. The seven-valent pneumococcal conjugate vaccine (PCV-7) targeted the seven serotypes of *S. pneumoniae* responsible for the majority of invasive disease at the time of its introduction in 2000; since then, PCV-7 has dramatically reduced the incidence of pneumococcal disease and mortality. However, sequencing of >600 Massachusetts pneumococcal isolates from 2001 to 2007 has shown that, in the pneumococcal population, previously rare nonvaccine serotypes are replacing vaccine serotypes and that some vaccine strains have persisted despite vaccination by recombining the vaccine-targeted capsule locus with a cassette of capsule genes from non-vaccine-targeted serotypes.

The large collections of pathogen genome sequences are driving development of tools to decipher the genetic basis for antibiotic resistance, virulence, and infection risk. Some pathogens have distinct types of clinical manifestations, the basis for which we are just beginning to unravel with the aid of genomics. For example, *Listeria* is a food-borne pathogen that can cause both central nervous system infections and maternal/neonatal infections. Although all *Listeria* isolates are treated the same from a public health perspective, variation in outcomes exists and appears to be linked to the strains' genomic background. Molecular analysis of a national reference laboratory's collections of well-characterized specimens, based on the fraction of

immunocompetent people in which they caused disease, revealed that some clonal complexes of *Listeria* appear to be more virulent than others. Linking epidemiology and comparative genomics then enabled enumeration of putative virulence factors that contribute to the clinical phenotypes as well as identification and confirmation of a novel gene cluster that mediates central nervous system tropism. This approach illustrates progress toward a future in which we can link pathogen identification with risk, thereby informing resource use and allocation.

GLOBAL CONSIDERATIONS

 While cutting-edge genomic technologies are largely implemented in the developed world, their application to infectious diseases perhaps offers the biggest potential impact in less developed regions where the burden of these infections is greatest. This globalization of genomic technology and its extensions has already begun in each of the areas of focus highlighted in this chapter; it has occurred both through the application of advanced technologies to samples collected in the developing world and through the adaptation and importation of technologies directly to the developing world for on-site implementation as they become more globally accessible.

Genomic characterization of the pathogens responsible for such important global illnesses such as tuberculosis, malaria, trypanosomiasis, and cholera has led to insights in diagnosis, treatment, and infection control. For instance, with the increasing burden of drug-resistant tuberculosis in the developing world, a molecular diagnostic test has been developed to detect rifampin-resistant tuberculosis. The genetic basis for rifampin resistance has been well defined by targeted sequencing: characteristic mutations in the molecular target of rifampin, RNA polymerase, account for the vast majority of instances of rifampin resistance. At least in areas that can afford to implement it, a rapid, automated PCR assay that can detect both *M. tuberculosis* and a rifampin-resistant allele of RNA polymerase directly in clinical samples has been implemented in parts of Africa and Asia, transforming the recognition and management of incident tuberculosis and multidrug resistance where they are most prevalent. Since rifampin resistance frequently accompanies resistance to other antibiotics, this test can suggest the presence of multidrug-resistant *M. tuberculosis* within hours instead of weeks, without the infrastructure required for culture.

High-resolution genomic tracking of the spread of epidemics—from cholera to Ebola to Zika—has yielded insights into which public health measures may prove most effective in controlling local epidemics. Many genomic tracking efforts have involved close collaborations with local scientists and public health officials, and considerable investment in sequencing infrastructure in sub-Saharan Africa has made on-location epidemic tracking in the event of another such outbreak feasible. Such investment can not only enable real-time outbreak recognition and tracking but also provide the infrastructure needed to capitalize on the many other benefits of high-throughput sequencing as they are developed. Overall, sequencing efforts have become cheaper and have moved closer to point-of-care with each passing year. As these technologies synergize with efforts to globalize information-technology resources, global implementation of genomic methods promises to spread state-of-the-art methods for diagnosis, treatment, and epidemic tracking of infections to areas that need these capabilities the most.

SUMMARY

By illuminating the genetic information that encodes the most fundamental processes of life, genomic technologies are transforming many aspects of medicine. In infectious diseases, methods such as next-generation sequencing and genome-scale expression analysis offer information of unprecedented depth about individual microbes as well as microbial communities. This information is expanding our understanding of the interactions of microorganisms with each other, their human hosts, and the environment. Despite technological and financial barriers that continue to slow the widespread adoption of large-scale pathogen sequencing in clinical and public health settings, genomic methodologies have utterly transformed the research landscape in infectious disease and are beginning to make meaningful inroads into clinical settings. As even vaster amounts of data are generated, innovations in data storage, development of bioinformatics tools to manipulate the data, standardization of methods, and training of end-users in both the research and clinical realms will be required. The cost-effectiveness and applicability of whole-genome sequencing, particularly in the clinic, remain to be studied, and studies of the impact of genome sequencing on patient outcomes will be needed to clarify the contexts in which these new methodologies can make the greatest contributions to patient well-being. The ongoing efforts to overcome limitations through collaboration, teaching, and reduction of financial obstacles should be applauded and expanded. With advances in genomic technologies and computational analysis, our ability to detect, characterize, treat, monitor, prevent, and control infections has advanced rapidly in recent years and will continue to do so, with the hope of heralding a new era where the clinician is better armed to combat infection and promote human health.

FURTHER READING

Berry MP et al: An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature 466:973, 2010. [PubMed: 20725040]

Bullman S et al: Emerging concepts and technologies for the discovery of microorganisms involved in human disease. Annu Rev Pathol 12:217, 2017. [PubMed: 27959634]

Burnham CD et al: Diagnosing antimicrobial resistance. Nat Rev Microbiol 15:697, 2017. [PubMed: 29021600]

Croucher NJ et al: Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet 45:656, 2013. [PubMed: 23644493]

Dudas G et al: Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544:309, 2017. [PubMed: 28405027]

Gardy JL, Loman NJ: Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 19:9, 2018. [PubMed: 29129921]

Grad YH, Lipsitch M: Epidemiologic data and pathogen genome sequences: A powerful synergy for public health. Genome Biol 15:538, 2014. [PubMed: 25418119]

Loman NJ, Pallen MJ: Twenty years of bacterial genome sequencing. Nat Rev Microbiol 13:787, 2015. [PubMed: 26548914]

Mutreja A et al: Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462, 2011. [PubMed: 21866102]

Snitkin ES et al: Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med 4:148ra16, 2012.

Chapter 475: The Role of Circadian Biology in Health and Disease

Jonathan Cedernaes; Kathryn Moynihan Ramsey; Joseph Bass

INTRODUCTION

Circadian rhythms are autonomous, systemic cycles of behavior and physiology that synchronize organismal function at the cell and tissue level in anticipation of the 24-h rotation of the Earth. A common feature of modern “24/7” life is the routine disruption in these evolutionarily conserved endogenous circadian cycles, due to the rise in shift work, jet travel across time zones, exposure to blue light–emitting devices at night, and disrupted sleep. A transformation in understanding the molecular basis of circadian disorders has generated a new wave of research on metabolic disease, inflammation, aging, and cancer. This chapter provides an overview of (1) the basic biology of the circadian system; (2) primary circadian rhythm and interrelated sleep disorders; and (3) the role of the circadian system in both normal human physiology and disease states. Lastly, we review the rapidly emerging field of chronobiology as a pathway for novel diagnostic and therapeutic activity. A glossary of terms used in circadian biology is summarized in **Table 475-1**.

TABLE 475-1

Glossary of Terms Used in Discussion of the Circadian System

Term	Description
ASPD	Advanced sleep phase disorder (see text for description).
CBT	Core body temperature. Often used as an indicator of the circadian rhythm, but can be masked by sleep and exercise.
CCGs	Clock-controlled genes; output of the molecular clock.
Chronotype	Internal circadian rhythm of an individual determined by phase of entrainment, determining sleep propensity and timing of maximum alertness over a 24-h period.
Circadian period	Time required for one complete cycle or oscillation. Calculated by the time distance between two consecutive peaks or troughs of a circadian variable.
Circadian phase	Timing of the circadian rhythm. Defined by comparing, e.g., the peak (acrophase) or trough (bathyphase) to a fixed event, e.g., to a point in time. Synonymous with phase angle.
Circadian rhythm	A biological process that exhibits an endogenous, entrainable oscillation of approximately 24 h.
Circadian rhythm sleep disorders	Disorders of multiple etiology that have in common that they result in maladjustment of the biological clock with respect to the environment.
Constant routine	An experimental paradigm designed to study endogenous circadian rhythms in humans, by keeping behavioral and environmental factors constant. These paradigms thereby entail constant dim lighting, evenly distributed isocaloric energy intake, semi-recumbent posture, and forced wakefulness.
Desynchrony	Loss of synchrony occurring either between a rhythm and its Zeitgeber (external), or between two or more rhythms within an organism (internal).
Diurnal rhythm	An oscillation synchronized with the day/night cycle that repeats itself with a 24-h period. The rhythm does not have to persist when time cues (e.g., light) are absent.
DLMO	Dim-light melatonin onset; a marker of melatonin rhythm.
DSPD	Delayed sleep phase disorder (see text for description).
Entrainment	Synchronization of a circadian rhythm or other self-sustaining oscillation by a factor—Zeitgeber—that enforces the oscillator. Constant entrainment between the Zeitgeber and the oscillator results in a stable phase relationship between these entities.
Infradian rhythm	A recurrent cycle or period with a period length significantly greater than 24 h.
Melatonin	Hormone produced by the pineal gland (chemical name N-acetyl-5-methoxytryptamine); derived from L-tryptophan. Various forms of melatonin can be prescribed for CRSDs or sleep disorders.
Non-24-h rhythm disorder	A syndrome in which there typically are chronic 1- to 2-h daily delays in sleep onset and wake times in an individual living in society, e.g., due to complete blindness.
Peripheral clocks	Clocks presiding outside of the suprachiasmatic nucleus, the circadian system's master pacemaker.
PRC	Phase response curve; visual representation of how a particular manipulation (e.g., light) produces phase shifts as a function of the phase (i.e., circadian time) at which the manipulation occurs. Defining the PRC to light has enabled researchers to understand and predict how entrainment to light cycles is accomplished.

Term	Description
SCN	The suprachiasmatic nucleus or nuclei, also known as the master pacemaker in mammalian species. A bilateral set of nuclei positioned in the anterior ventral hypothalamus. Essential for entraining extra-SCN central and peripheral oscillators to the prevailing light-dark cycle via photic input from the retina.
Shift work	Work scheduled so that it occurs outside of the traditional work schedule of 9:00 A.M. to 5:00 P.M., or 6:00 A.M. to 6:00 P.M., depending on definition. Various forms of shift work exist, such as early morning, evening, or night shifts, as well as rotating shifts.
Ultradian rhythm	A recurrent cycle or period with a period significantly shorter than 24 h—e.g., a 2-h rhythm would exhibit 12 cycles within a circadian (24-h) rhythm.

BASIC EVOLUTION AND STRUCTURE OF THE CIRCADIAN SYSTEM

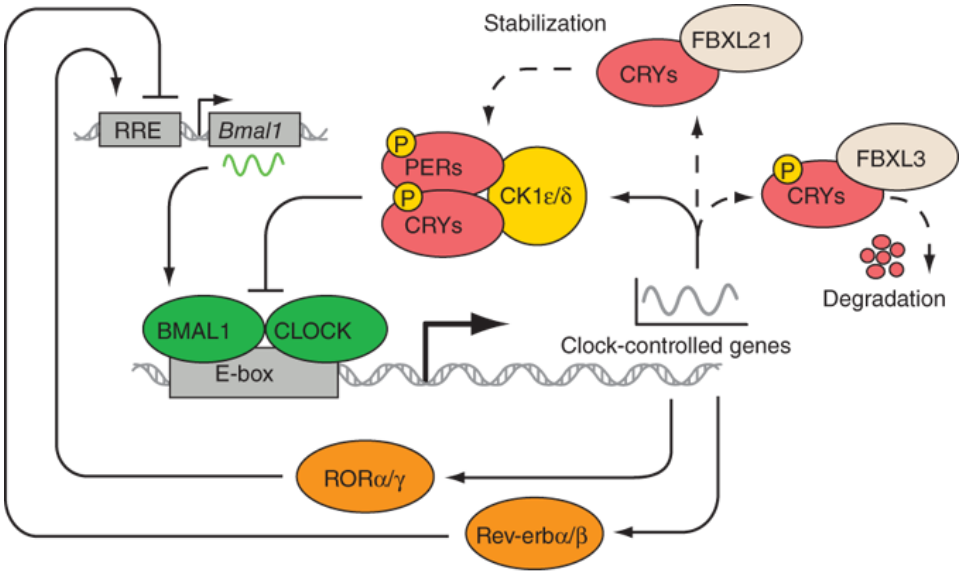
A daily cycle between light and darkness existed long before the emergence of multicellular life. Eubacteria, the most ancient photosynthetic prokaryote, emerged in the geologic record more than 2.5 billion years ago at the same time as the first endogenous molecular clock. The co-occurrence in molecular evolution of clocks and photosynthesis hints at an interrelated and selective advantage to the clock and energetic processes—indeed clocks coordinate oxygenic reactions with periods of sunlight each day, and perturbation of clock cycles reduces fitness, reproduction, and survival. Additionally, clocks protect photosynthetic organisms from the DNA-damaging effects of sunlight by timing the production of DNA repair processes, such as photolyase-mediated repair, to the nighttime. Across billions of years of evolution, highly conserved circadian clocks (from "*circa diem*," "about a day") have been found in all photosensitive organisms, governing a wide range of behavioral, physiologic, and biochemical processes. A defining property of the circadian clock system is that it enables organisms to anticipate, rather than simply react to, daily changes in the external environment that are tied to the day-night cycle. In mammals, circadian systems are organized hierarchically with a light-responsive "master" circadian pacemaker located within the suprachiasmatic nucleus (SCN) of the anterior hypothalamus, which in turn presides over a network of both extra-SCN and peripheral clocks (see "Anatomic Organization of the Circadian Clock Network," below). Daily light exposure signals to the SCN and entrains the circadian system to the 24-h day (see "Entrainment and Measurement of the Circadian System," below), and the SCN in turn maintains synchrony of a diverse network of peripheral clocks via a variety of signals that have as yet to be fully identified, but which involve direct physiologic rhythms (core body temperature), the autonomic nervous system, and neuroendocrine signals, including cortisol as part of the hypothalamic-pituitary-adrenal (HPA) axis.

MOLECULAR ORGANIZATION OF THE MAMMALIAN CIRCADIAN CLOCK

At the molecular level, mammalian circadian rhythms are generated by a transcription-translation autoregulatory feedback loop. The forward limb of the clock is composed of the basic helix-loop-helix transcription factors (TFs) CLOCK (or its paralogue, NPAS2) and BMAL1, which drive expression of their own repressors (PER and CRY) in the negative limb, in a cycle that repeats itself every 24 h **(Fig. 475-1)**. A second short feedback loop involves CLOCK/BMAL1-mediated transcription of the retinoic acid–related orphan nuclear receptors ROR and REV-ERB, which activate and repress *Bmal1* transcription, respectively. Additional posttranslational regulation of the stability and degradation of core clock TFs includes phosphorylation by casein kinase 1 epsilon (CK1ε) and casein kinase 1 delta (CK1δ) and ubiquitination by FBXL3 and FBXL21. In addition to the ~24-h oscillation of core clock genes, a wide array of downstream clock-controlled genes (CCGs) exhibit broad amplitude in expression, ultimately giving rise to rhythmic physiologic processes. The core clock feedback loop and the induction of CCG rhythms also involves epigenetic mechanisms such as acetylation and methylation. The molecular circadian feedback loop is synchronized with sunrise each day by photosensitive melanopsin-expressing neurons within the retina. The retinohypothalamic tract (RHT) represents the input circuit to the SCN that maintains coherent organismal rhythms. Of note, mutations in several of these clock genes are associated with impaired circadian rhythms and physiology in humans (see "Primary Pathologies of the Circadian System," below). The importance of clock genes in the brain has been demonstrated by genetic studies, which have found that deletion of *Bmal1* in the whole brain or regions that span the SCN cause behavioral arrhythmicity, even when genetic ablation occurs late in life. Conversely, restoring *Bmal1* expression specifically in brain in adult clock mutant mice can rescue behavioral locomotor rhythms. Of note, whereas CLOCK normally heterodimerizes with BMAL1, NPAS2 is able to functionally substitute for CLOCK within pacemaker neurons; thus while mice lacking either *Clock* or *Npas2* genes maintain rhythmicity, the double mutants lacking both CLOCK and NPAS2 have impaired circadian rhythms of locomotor activity.

Figure 475-1

Central clock molecular mechanism. The core molecular clock machinery in mammals is encoded by interlocking transcription-translation feedback loops that oscillate with ~24-h periodicity. The transcription factors CLOCK and BMAL1 heterodimerize to drive transcription of downstream clock-controlled target genes containing E-box enhancer elements. Among these, the PER and CRY proteins multimerize and inhibit CLOCK/BMAL1 while RORs and REV-ERBs activate and inhibit, respectively, *Bmal1* transcription, resulting in rhythmic oscillations of clock-controlled and downstream target genes.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

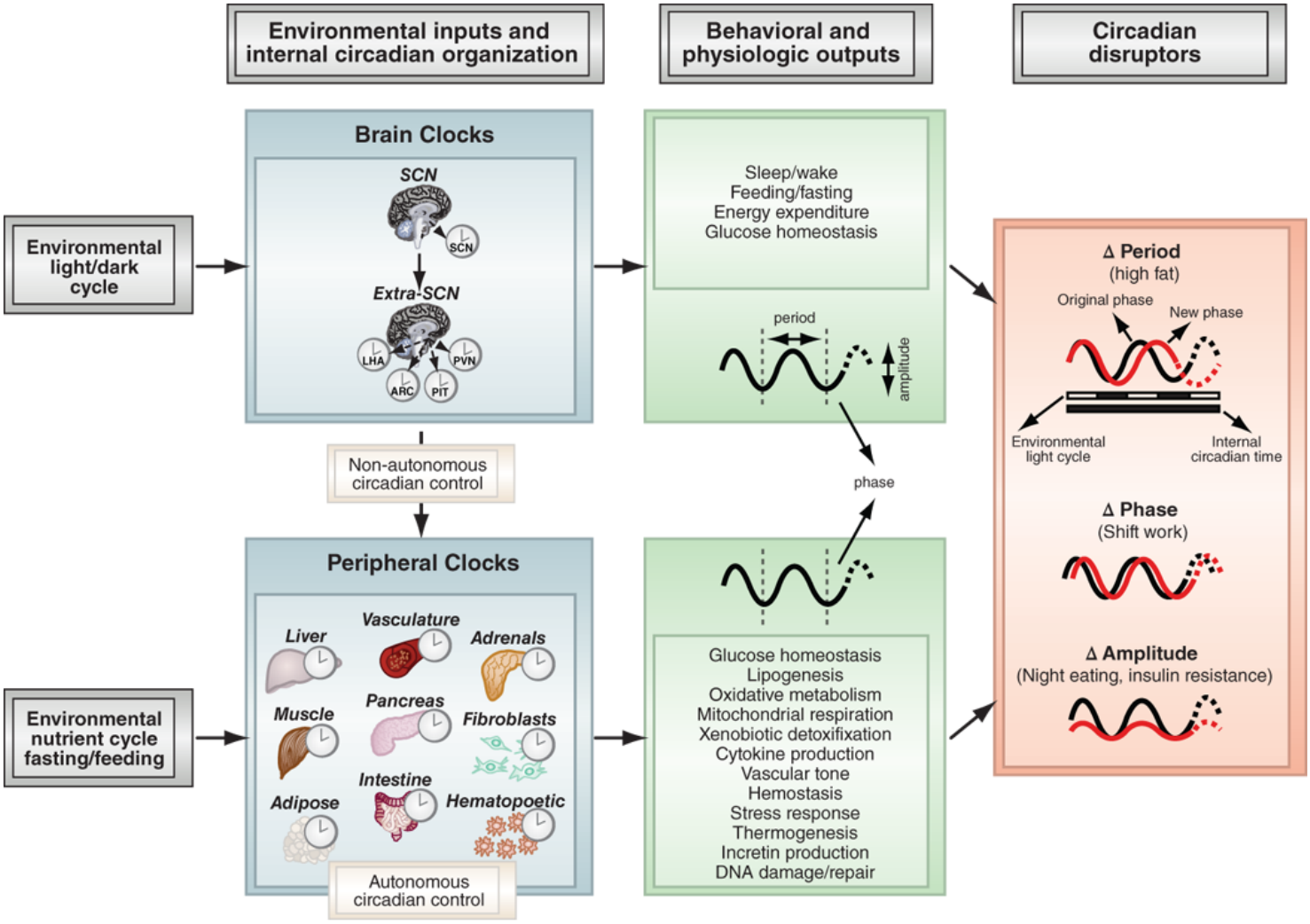
A major transformation in our understanding of circadian biology came with the discovery that the molecular clock network is present not only in the SCN, but also within most peripheral tissues, as well as in extra-SCN neurons in the brain. Interestingly, both the SCN and peripheral tissues exhibit a surprisingly large number of cycling transcripts at the genomic level, with approximately 3–16% of the mammalian transcriptome displaying a 24-h variation in mRNA expression levels across any given tissue, although the set of genes exhibiting circadian rhythmicity varies substantially between tissues depending upon tissue-specific functions. Posttranscriptional events such as RNA polyadenylation and mRNA translation also exhibit circadian variation, further increasing the repertoire of circadian processes at a cellular level. Thus, it is not surprising that among genes in the mammalian genome, almost half are considered to display circadian rhythms in at least one tissue.

ANATOMIC ORGANIZATION OF THE CIRCADIAN CLOCK NETWORK

Understanding the circuit organization of the circadian clock within the brain is increasingly relevant in understanding how the master circadian pacemaker center within the SCN regulates feeding, sleep-wake activity, energy expenditure, endocrine processes, and metabolism **(Fig. 475-2)**. Identification of the SCN as the master pacemaker was first established by the observation that SCN lesioning induced complete loss of rhythms of locomotor activity and endocrine hormone secretion. The ventral “core” region of the SCN, which is composed of neurons producing vasoactive intestinal polypeptide (VIP), receives photic information directly from the retina through the RHT. At the molecular level, circadian gene transcription is induced within the SCN through the initial activation of immediate early genes, such as *Per1*, *Per2*, *c-fos*, and *jun*. Cells within the “core” region of the SCN then signal primarily via GABA-ergic neurotransmitter release to synchronize the cells within the “shell” region of the SCN, which produce the neuropeptide arginine vasopressin (AVP).

Figure 475-2

Central and peripheral clocks coordinate environmental cues with behavior and physiologic outputs. Light entrains the master pacemaker neurons in the SCN, which subsequently synchronizes extra-SCN and peripheral clocks. Brain clock output includes sleep-wake, fasting-feeding, and energy expenditure cycles, while peripheral clock output includes a wide range of physiologic processes, including glucose homeostasis, oxidative metabolism, cytokine production, and stress response. The right column indicates different ways that circadian disruptors, such as diet, shift work, or other circadian rhythm sleep disorders, may impact the clock—i.e., by changing circadian period, phase, or amplitude.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

The SCN communicates to extra-SCN and peripheral clocks through both secreted factors and direct neuronal projections. The former was elegantly proven by the ability of SCN grafts to partially restore locomotor rhythms in an SCN-lesioned host. Efferent nerve outputs predominantly arise from the AVP-producing shell region of the SCN, although some output is also derived from the VIP-predominated core. The SCN projects to several hypothalamic relay regions, including the median preoptic nucleus, the subparaventricular zone (SPZ), the dorsomedial hypothalamus (DMH), and the paraventricular nucleus of the hypothalamus (PVH). These, in turn, regulate output to both sleep- and wake-promoting regions, as well as to regions involved in autonomic regulation and feeding. The SCN is thereby thought to promote sleep through the transmission of neural signals that terminate in the sleep-promoting ventrolateral preoptic nucleus (VLPO), i.e., one of the brain regions that is active during sleep. In contrast, the SCN promotes wakefulness during the active phase by transmission of neural signals that—by passing through regions such as the DMH—terminate in wake-promoting regions, including the locus coeruleus, lateral hypothalamic nucleus, ventral tegmental area, and dorsal raphe nucleus.

The SCN also signals via noradrenergic fibers to the pineal gland, thereby regulating the circadian production of the hormone melatonin. SCN control of the nighttime rise in pineal melatonin release (in both diurnal and nocturnal animals) is mediated through a pathway involving the PVH. Of note, artificial light at night is able to delay the secretion of melatonin, ultimately affecting sleep (see "Endocrine Systems Regulated by the Circadian Clock," below). Melatonin plays a complex role in the circadian system since the MT1 and MT2 melatonin receptors are expressed on the SCN itself; thus melatonin feeds back to modulate circadian outputs to other cells in the brain and body.

Neuronal output from the SCN also reaches the periphery, i.e., to the adrenal glands, the liver, and the pancreas. The SCN produces rhythmic variation in multiple neuroendocrine axes, producing daily rhythms of gonadotropin, thyrotropin, and somatotropin. Prominent hypothalamic-pituitary-adrenal (HPA) axis rhythms ultimately give rise to daily variation in diverse pathways essential for hemodynamic stability, metabolism, and inflammation. These rhythms originate with SCN control of corticotropin-releasing hormone (CRH)-producing cells in the PVH, which induce daily oscillations of both pituitary adrenocorticotropic hormone (ACTH) and adrenal cortisol. Highlighting the importance of SCN output for peripheral rhythms, there is a dramatic reduction in the number of transcripts that exhibit circadian rhythms in the liver following SCN ablation in mice. Nonetheless, when the autonomous clock in the liver is ablated in mice, some key clock transcripts such as *Per2* are still able to cycle as long as the core body temperature rhythm persists. Whereas the SCN is exclusively entrained by light, meal timing is able to signal circadian time directly to the liver. Thus, shifted meal timing as occurs during shift work or jetlag can uncouple peripheral clocks from the central pacemaker. In comparison to peripheral tissue clocks, the SCN is also resistant to phase shifts induced by

temperature. This is consistent with the concept that the SCN generates the core body temperature rhythm as one of the major mechanisms to signal circadian time to peripheral clocks.

ENTRAINMENT AND MEASUREMENT OF THE CIRCADIAN SYSTEM

Under normal light-dark cycles, the circadian system is corrected or "entrained" on a daily basis, producing diurnal rhythms of 24 h. Such signals of entrainment are called Zeitgebers ("time-giver" signals) and include light exposure, meal timing, and activity patterns. Light serves as the dominant Zeitgeber for the circadian system, and a breakthrough in understanding photoentrainment in mammals came with the discovery of the melanopsin system, which is composed of a specialized class of photosensitive retinal ganglion cells that expresses the blue-light sensitive photopigment melanopsin in the inner retina, separate from the photoreceptive rods and cones. Blue light around this wavelength (~480 nm) suppresses melatonin and promotes subjective and objective (electroencephalography-assessed) wakefulness.

The ability of light to entrain the circadian system functions according to a so-called phase response curve (PRC). When light exposure occurs prior to the critical phase of the core body temperature (CBT), defined by the CBT's minimum, light produces a phase delay in the circadian rhythm. Conversely, light exposure after this critical period causes phase advances. The circadian system can respond even to small changes in light intensity (e.g., dim light at ~100 lux can produce half of the phase delay compared with an almost 100-fold greater light exposure). This responsiveness is furthermore affected by our genetic makeup, as variants in clock genes can modulate the response of the human circadian system to light.

When an organism is placed in an environment without Zeitgebers, the circadian rhythm is said to free-run, as it will rely on the endogenous rhythm of the circadian system. In humans, the study of endogenous circadian rhythms can be achieved by using a so-called constant routine. In these paradigms, subjects are maintained in a constant semi-recumbent posture, meals are provided on an hourly basis, light is constantly kept below the level which phase shifts the SCN, and circadian rhythms are assessed by measuring CBT, melatonin, or peptidergic hormone rhythms. In animals, circadian rhythms are instead studied by examining behavior and physiologic responses after 30–36 h of complete darkness, and endogenous rhythms are assessed by measuring voluntary locomotor activity. From these measurements, key properties of the circadian system can be ascertained, such as period length (peak-to-peak or trough-to-trough time), amplitude (peak-to-trough difference) and phase (timing of peak or trough in relation to a reference point) (Fig. 475-2).

These studies have revealed that the endogenous human circadian clock runs with a period length of approximately 24.2 h (compared with mice that run at ~23.5 h, depending on the strain). Evidence indicates that human females may have a slightly shorter circadian clock (24.1 vs 24.2 h). Notably, interindividual variability in the endogenous circadian period length is further diversified by the existence of genetic polymorphisms in clock genes (see below). These gene variants can confer extremes in endogenous circadian period as well as phase; the latter can be advanced or delayed by about 3–4 h in each direction. This is due both to altered circadian rhythms at the cell level and to altered SCN responsiveness to entrainment by light. For instance, *PER3* exists in a variable-number, tandem-repeat polymorphism. Individuals homozygous for a *PER3 5/5* genotype have been reported to be more responsive than *PER3 4/4* homozygous individuals to the melatonin-suppressing effect of evening blue light exposure.

Using specifically developed questionnaires to establish preferred sleep-wake timing, individuals can be categorized into so-called morningness-eveningness types or chronotypes. The most commonly used questionnaires are the Horne-Östberg morningness-eveningness questionnaire (MEQ) and the Munich ChronoType Questionnaire (MCTQ). The MEQ is composed of 19 questions regarding traits distinguishing morning-type compared with evening-type individuals, such as preferred waking time. In contrast, the MCTQ centers on the midpoint of sleep as a circadian marker, queries age and sex across a range of geographical locations, and can be used to ascertain differences between socially imposed sleep patterns (e.g., on working days) and sleep patterns on free days (the difference constituting so-called social jetlag). According to MCTQs obtained from primarily European populations, ~1% of the general population goes to bed before 10:00 PM and about 8% after 03:00 AM. Differences in chronotype are linked to altered circadian timing—including peak levels of melatonin, which can vary by up to 4 h between extreme morning and evening types. Extreme chronotypes have also been shown to be linked to various traits; i.e., low morningness scores have been associated with greater tolerance to shift work.

Melatonin is one of the most commonly used peripheral markers of an individual's circadian rhythm, reflecting the rhythmic function of the SCN. Circadian rhythms of melatonin can be measured in saliva or plasma, while 6-sulphatoxymelatonin (aMT6S), a metabolite generated from the breakdown of melatonin, can also be measured in urine. Accurate estimations of melatonin rhythms are often obtained by analyzing the dim light melatonin onset (DLMO), which as the name implies does not require an entire 24-h sampling, making this marker useful in both the clinical and research setting. In normally entrained individuals, the DLMO can be used to ascertain whether an individual's circadian rhythm is phase advanced or delayed, and this onset typically occurs ~2 h before the onset of sleep. The midpoint of sleep—the main marker used by the MCTQ—also strongly correlates with melatonin onset.

The CBT is also often utilized as an indicator of the circadian rhythm, and even though the outcome is more variable when using the CBT, it usually correlates well with the phase obtained using the melatonin rhythm. The CBT, however, can be masked by factors such as sleep, food intake, and activity. CBT can be recorded and registered wirelessly with relative ease. In humans, CBT can for example be recorded via rectal thermometers or probes that are swallowed to pass through the gastrointestinal tract. When humans are studied under normal conditions with normal lighting and sleep duration from 2300 to 0700 h, the CBT reaches around 37.2°C by 0900 h, and from there continues to rise slowly until it reaches 37.4°C around 11 h later. The CBT then drops to the daily low of 36.5°C in the early morning (0400 h).

Given the interrelationship between the circadian system and sleep-wake systems, researchers have developed paradigms that uncouple the circadian system from sleep-wake states, enabling the study of the contribution of the circadian system to investigated parameters across the entire sleep-wake

cycle. These paradigms are known as “forced desynchrony” protocols and involve enforcing a significantly shortened (e.g., 20 h) or prolonged (e.g., 28 h) day length upon individuals. These protocols thus attempt to approximate what occurs during rotating shift work or “jetlag,” e.g., when travel across several time zones suddenly shifts the light-dark and behavioral cycle drastically away from the entrained 24-h rhythm. As described below, forced desynchrony protocols have contributed to uncovering how the circadian system regulates parameters such as cognitive performance, subjective alertness, and metabolic and cardiovascular health.

PRIMARY PATHOLOGIES OF THE CIRCADIAN SYSTEM

An overarching term for disorders of the circadian system is *circadian rhythm sleep disorders* (CRSD) (See Also Chap. 27). These disorders have become increasingly recognized as important factors in a number of conditions; a unifying feature of CRSD involves a mismatch between subjective behavioral and physiologic rhythm with the environmental light-dark or social activity-rest cycle (i.e., the body clock is out of sync with the external light-dark cycle). CRSDs can arise either due to misalignment of an exogenous environmental factor, such as light, with the intrinsic circadian cycle, or due to misalignment of activity/rest cycle in relation to endogenous circadian timing (e.g., shift work or jetlag). In addition to such environmental or exogenous conditions causing circadian disruption, in some cases intrinsic circadian timing is altered in relation to the external environment, as in the case of endogenous circadian disorders that include those caused by mutations in core clock genes. Under conditions of intrinsic circadian abnormalities, it is often exceedingly difficult for individuals suffering from CRSDs to try to properly realign themselves, and these disorders often result in adverse effects such as sleepiness or depressed mood. Societal and economic consequences also are common; these can result in the individual being unable to maintain a job or be unable to attend at regular school hours. The criteria for CRSDs based on the International Classification of Sleep Disorders (ICSD) is shown in **Table 475-2**.

TABLE 475-2

Criteria for Circadian Rhythm Sleep Disorders

Criteria	Description
A	A persistent or recurrent pattern of sleep disturbance due primarily to one of the following: • Alterations of the internal circadian timekeeping system. • Misalignment between endogenous circadian rhythms and exogenous factors that affect the timing or duration of sleep.
B	A circadian-related sleep disruption that leads to insomnia, excessive daytime sleepiness, or both.
C	A sleep disturbance that is associated with impairment of social, occupational, or other areas of functioning.

Animal models have greatly advanced our understanding of how core molecular clock components contribute to maintaining normal sleep-wake/rest-activity cycles **(Table 475-3)**. For example, *Clock*$^{\Delta 19/\Delta 19}$ mice have reduced total sleep duration and less induction of REM sleep in response to sleep deprivation. Further, mice that lack *Bmal1* have increased total sleep time, but it is more fragmented and lacks clear 24-h sleep-wake rhythms, and mice lacking the repressors *Cry1* and *Cry2* are not only arrhythmic but spend more time in non-REM sleep. Finally, while ablation of the circadian gene *Dbp* does not alter the specific duration of sleep stages, it does lead to an altered circadian sleep-wake distribution, with more sleep during their normal wake period and vice versa. Consistent with a key role of clock genes in regulating sleep-wake behavior, human genetic studies of twins have found that up to half of the variation in diurnal preference is heritable. Established genetic variants associated with diurnal preference and circadian sleep disorders are listed in **Table 475-4**.

TABLE 475-3

Animal Models of Genetic Circadian Disruption

Gene	Average Circadian Time of Peak Transcript Level		Allele	Mutant Phenotype
	SCN	Periphery		
Bmal1 (Arntl)	15–21	22–2	*Bmal1*$^{-/-}$	Arrhythmic
CK1δ (Csnk1δ)	No rhythm	No rhythm	*Csnk1δ*$^{+/-}$	0 to 0.5-h shorter period
CK1ε (Csnk1ε)	No rhythm	No rhythm	*CK1ε*^{tau}	4-h shorter period
CK1ε	—	—	*CK1ε*$^{-/-}$	0.2- to 0.4-h longer period
Clock	No rhythm	21–3	*Clock*$^{-/-}$	0.5-h shorter period
—	—	—	*Clock*$^{\Delta 19/\Delta 19}$	4-h longer period/arrhythmic
Clock/Npas2	—	—	*Clock*$^{-/-}$/*NPAS2*$^{-/-}$	Arrhythmic
Cry1	8–14	14–18	*Cry1*$^{-/-}$	1-h shorter period
Cry2	8–14	8–12	*Cry2*$^{-/-}$	1-h longer period
—	—	—	*Cry2*^{A260T}	0.2-h shorter period
Dbp	—	—	*Dbp*$^{-/-}$	0.5-h shorter period
Npas2	N/A	0–4	*Npas2*$^{-/-}$	0.2-h shorter period
Per1	4–8	10–16	*Per1*$^{-/-}$	0.7-h shorter period
—	—	—	*Per1*^{brdm1}	1-h shorter period
—	—	—	*Per1*^{ldc}	0.5-h shorter period/arrhythmic
Per2	6–12	14–18	*Per2*^{brdm1}	1.5-h shorter period/arrhythmic
—	—	—	*Per2*^{ldc}	Arrhythmic
Per3	4–9	10–14	*Per3*$^{-/-}$	0 to 0.5-h shorter period
Rev-erbα (Nr1d1)	2–6	4–10	*Rev-erbα*$^{-/-}$	0.5-h shorter period/disrupted photic entrainment
Rorα	6–10	Arrhythmic/various	*staggerer*	0.5-h shorter period/disrupted photic entrainment
Rorβ	4–8	18–22	*Rorβ*$^{-/-}$	0.5-h longer period
Rorγ	N/A	16–20/various	*Rorγ*$^{-/-}$	Normal behavior

Note: Normal circadian rhythms of circadian clock and related genes, with description of circadian phenotype in mutant mice.

Source: Adapted from Hum Mol Genet 15:R271, 2006, and Adv Genet 74:175, 2011.

TABLE 475-4

Mutations and Gene Variants Linked to Sleep-wake Disorders and Diurnal Preference

Gene	Position	Population	Syndrome/Sleep Preference
h*CKIε*	S408N	Japanese	Protection against DSPS
h*CKIγ*	T44A	Pedigree	FASPS
h*CKIδ*	H46R	Pedigree	FASPS
h*CLOCK*	T3111C (3'-UTR)	European	Eveningness
h*CRY2*	A260T	Pedigree	FASPS
h*PER2*	S662G (Missense mutations in CKIε binding region)	Pedigree	FASPS
h*PER2*	C111G (5'-UTR)	British	Extreme morningness
h*PER3*	P415A/H417R	Pedigree	FASPS and seasonal affective disorder
h*PER3*	G647	Swedish/Finish/Austrian/German	Morningness
h*PER3*	G647, P864, 4-repeat, T1037, R1158	Japanese	DSPS
h*PER3*	Increased repeats (exon 18, 54 bp)	Brazilian	DSPS
h*VIP*	rs9479402 (gene variant 54 kb upstream of VIP)	European (>97% European ancestry)	Morningness

Delayed Sleep Phase Disorder

Delayed sleep phase disorder (DSPD) is one of the more common circadian rhythm sleep disorders and is characterized by chronic and significant delays in both sleep onset and wake times compared to normal "socially acceptable" sleep-wake hours (i.e., scoring as "extreme night owls" on morningness-eveningness preference tests). Rhythms of CBT and melatonin levels in plasma and urine are likewise delayed and the circadian period (tau) is longer in DSPD. Onset of DSPD most commonly occurs during adolescence or early adulthood. While the precise etiology of DSPD is not well established, it has been associated with polymorphisms within the circadian clock genes *CLOCK* and *PER3*. An integrated behavioral and pharmaceutical therapeutic approach has been found to be most effective at treating individuals with DSPD. Such treatments include a combination of bright-light therapy soon after waking in the morning (and/or dark-room therapy in the evening) and melatonin administration in the evening several hours prior to the onset of sleep. These approaches aim to realign endogenous circadian rhythms with the desired sleep-wake schedule. As individuals suffering from DSPD also phase delay more rapidly, this explains why attempts to phase advance their sleep schedule can be difficult, as well as why relapse can easily occur after initial treatment.

Advanced Sleep Phase Disorder

Another circadian rhythm sleep disorder whereby one gets the correct amount and quality of sleep but at a shifted time is advanced sleep phase disorder (ASPD). Individuals with ASPD experience an advance in their major sleep episode in relation to the desired clock time. Thus, this disorder typically results both in very early evening bedtimes and morning awakenings (e.g., "extreme early birds"), resulting in reduced quality of life due to excessive sleepiness during early evening, even in social situations. Individuals with ASPD also have phase-advanced temperature and melatonin rhythms in parallel with their earlier sleep onsets. ASPD occurs more often in older individuals, although early-onset autosomal dominant familial variants (familial advanced sleep phase syndrome [FASPS]) have also been associated with mutations in either the *PER2* or the casein kinase 1δ (CK1δ) genes. PER2 is critical for SCN resetting by light, and the identification of PER2 mutations in familial ASPD was the first in which clock genes were tied to a CRSD. Such mutations have been found to be able to shorten the endogenous circadian period to about 23.3 h compared with the normal 24.2-h period length. Accordingly, ASPD can be distinguished from other non-circadian sleep disorders by an early onset of dim-light melatonin secretion, a reliable marker for the timing of endogenous circadian rhythms both in the research laboratory as well as in the clinical setting. Polysomnography (PSG) or actigraphy are not required for diagnosing ASPD, although actigraphy may be significantly more feasible for long-term analysis of circadian timing of sleep. Treatments for ASPD include bright light or blue-enriched phototherapy in the evening hours to delay the phase of the circadian clock to a later hour.

Non-24-h Sleep-Wake Rhythm Disorder

Individuals with non-24-h sleep-wake rhythm disorder ("non-24"), otherwise known as free-running disorder (FRD), have endogenous circadian rhythms that are not synchronized with the external 24-h day-night cycle due to an inability to readjust the circadian clock to the 24-h day on a daily basis. This most commonly occurs in individuals who are completely blind (i.e., lacking all photoreceptors) since they are unable to respond to daily light cues, which normally would reset the endogenous circadian clock on a daily basis (although the condition has also been reported in sighted individuals). Instead, the sleep-wake period length corresponds to the individual's endogenous circadian rhythms, which are typically slightly longer than 24 h, thereby shifting sleep and wake cycles over time in relation to the light-dark cycle. Instead of sleeping at the same time each day, their sleep time would gradually be delayed each day until their sleep period literally goes "around the clock." Depending on the individual's endogenous rhythm, the individual will take a given number of days to re-align his or her endogenous phase (in a 360° phase plot) with the zero time point in the exogenous 24-h light-dark cycle. For example, if an individual has an endogenous 24.5-h rhythm, it would take that individual 48 days to cycle from one cycle to the next. Because of this chronic cycling, prominent symptoms of non-24 include sleep-wake cycle disruption (insomnia and daytime sleepiness), impaired alertness and mood levels, and severe difficulties partaking in normally scheduled work, school, or social activities. Non-24 can be diagnosed following diurnal analysis of an individual's melatonin or cortisol rhythms, in combination with analyses of sleep diaries where the sleep onset and offset can be visualized over time to identify the free-running period. Treatments for sighted non-24-h patients include a combination of bright light therapy with appropriately timed melatonin administration, while melatonin and dual melatonin (MT1 and MT2) receptor agonist administration in completely blind non-24-h patients has been shown to entrain free-running rhythms and improve symptoms.

Shift Work Sleep Disorder

Given the increased prevalence of shift work in today's 24/7 society and the accumulating evidence for increased incidence of sleep and metabolic disorders, including obesity, diabetes, cardiovascular disease, and cancer, in shift workers, the need to develop effective treatments for shift work sleep disorder (SWSD) are increasingly important. Shift work sleep disorder is at its core defined by the primary symptom of either insomnia or excessive sleepiness, occurring as a transient phenomenon in relation to work that usually is scheduled during the habitual hours of sleep or comprises irregular work hours. The symptoms may result from recovery of sleep having to consume a large proportion of the individual's free time, which may produce negative social consequences such as difficulties maintaining social relationships. Older individuals are typically at an increased risk of SWSD due to age-associated decline in the ability to maintain sleep during normal waking hours. Since the symptoms likely arise from a misalignment of sleep-wake rhythms with the external light-dark cycle, therapeutic approaches aim to realign endogenous circadian rhythms with the sleep cycles dictated by work. In addition to optimizing the sleep environment at home to minimize disruptions, timed bright light therapy can help individuals with SWSD—i.e., for night workers, intermittent bright light exposure during the night and avoidance of bright light during the morning, even on days off, has been shown to improve sleep and feelings of alertness. Melatonin prior to bedtime may also help improve symptoms of SWSD. Genetic screening combined with chronotype questionnaires may become useful tools for determining whether a given individual is suited for shiftwork. For instance, a twin study indicated that a genetic variant of the circadian gene *DEC2* was associated with reduced sleep duration, and with shorter recovery sleep following extended sleep deprivation. More studies may reveal additional genetic variants that confer an advantage to repeated phase advances and phase delays as typically occurs in shift work.

Irregular Sleep-Wake Rhythm

Damage to the SCN can produce arrhythmicity in animals and is thought to be one of the possible underlying reasons for the temporally disorganized sleep-wake pattern that characterizes the disorder known as irregular sleep-wake rhythm (ISWR). Other contributing factors may be a reduced responsiveness to entraining signals such as light and physical activity, as well as a decreased exposure to such signals as often occurs with increasing age due to, for example, increased risk of poor health and impaired mobility. Whereas the total sleep time per 24 h may be comparable, there is a relative absence of a circadian pattern to the sleep-wake cycle. Sleep timing throughout the sleep-wake cycle can be shortened—sometimes close to randomly distributed—instead of occurring in several distinct bouts. ISWR is often associated with neurologic impairment, foremost Alzheimer's disease in older age; however, ISWR can also occur in individuals with poor sleep hygiene. The most effective treatments for ISWR involve not pharmacotherapy but rather multimodal interventions such as increased light exposure, improved sleep hygiene, and promotion of social and physical activities.

Jetlag

Most have experienced symptoms associated with jetlag, including insomnia, day-time sleepiness, and fatigue, when traveling from one time zone to another, as one's endogenous circadian rhythms are not yet aligned, or entrained, to the new external light-dark cycle. This is due to the slowness of the circadian system to adapt to the new time zone: typically, the human circadian system is able to shift ~1.5 h a day in the westward direction (i.e., a phase delay), whereas it shifts more slowly (about 1 h daily) with eastward direction of travel (i.e., achieving a phase advance). Importantly, the symptoms are distinguished from the more short-lived symptoms that can partially result from exposure to traditional airplane cabin conditions, including abdominal distention, dependent edema, muscle cramps, headaches, nausea, and, intermittent dizziness. Usually symptoms of jetlag abate within the first couple of days after traveling, and may present themselves after a first night of good sleep (which is more dependent on a high build-up of homeostatic sleep pressure). Older individuals (age >50) appear to be more at risk. While symptoms are transient, therapeutic approaches can alleviate or temper some of the side effects of travel by hastening synchronization of the internal and external circadian cycles. Behavioral treatments include appropriately timed bright-light exposure and avoidance of bright light during the night time in the new destination, while pharmacologic approaches include timed melatonin administration before bedtime both prior to and following travel, resulting in improved sleep quality and decreased night waking.

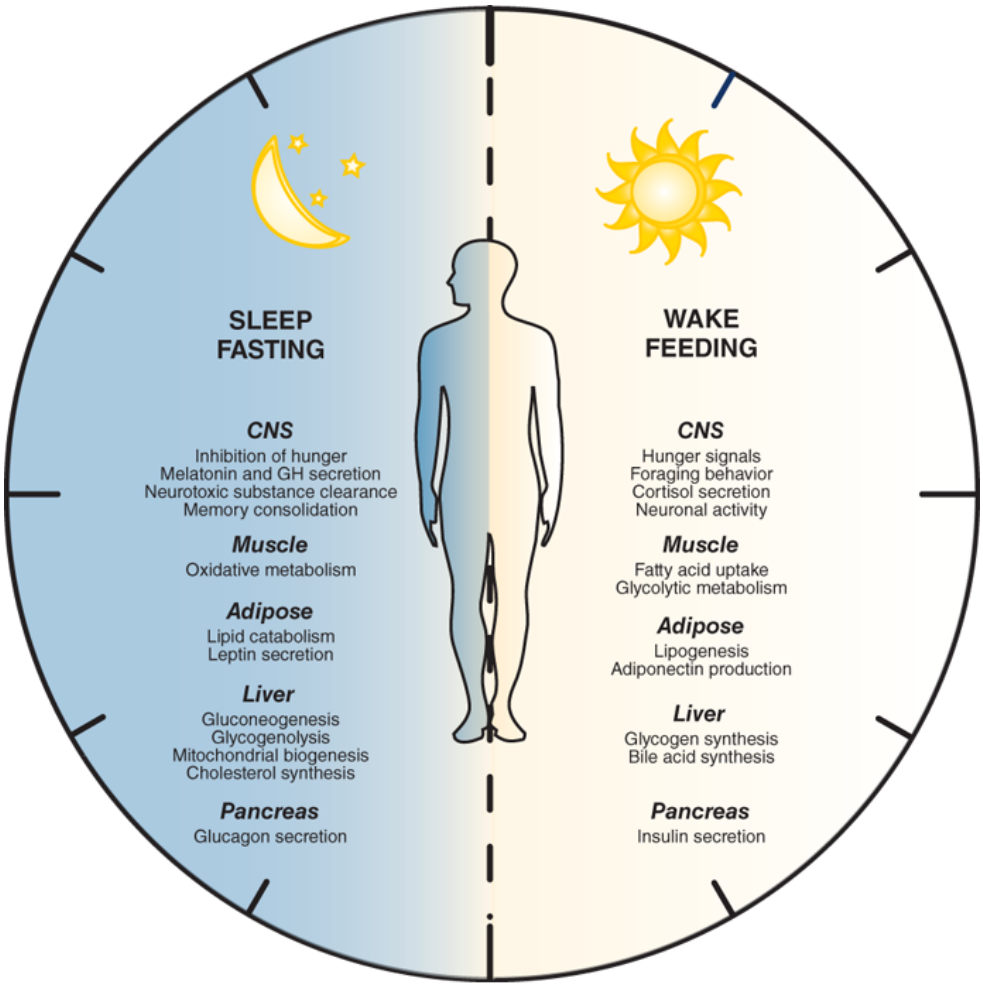
Social Jetlag

Individuals with a late chronotype are prone to suffer from “social jetlag,” a phenomenon in which individuals are forced to awaken at a point at which their bodies are entrained to be asleep due to discrepancy between alignment of social and biological time. Social jetlag can be estimated using questionnaires, such as the MCTQ, to compare sleep timing on non-free compared with free days. This has established that a large proportion of the European population suffers from 2 or more hours of social jetlag. Chronic social jetlag has been associated with an increased risk of developing obesity and the metabolic syndrome, as well as with greater alcohol consumption and smoking.

The aforementioned categories of defined clinical circadian disorders have been traditionally established based upon consideration of the endogenous behavioral and physiologic cycles (primarily of melatonin and temperature) with the external 24-h light-dark cycle. In the following sections, we build on the concepts of circadian behavioral disorders to consider new and emerging insight into the role of circadian disruption in organismal homeostasis (**Figs. 475-3 and 475-4**), and the availability of genetic strategies to dissect the interrelationship between clock function, health, and disease.

Figure 475-3

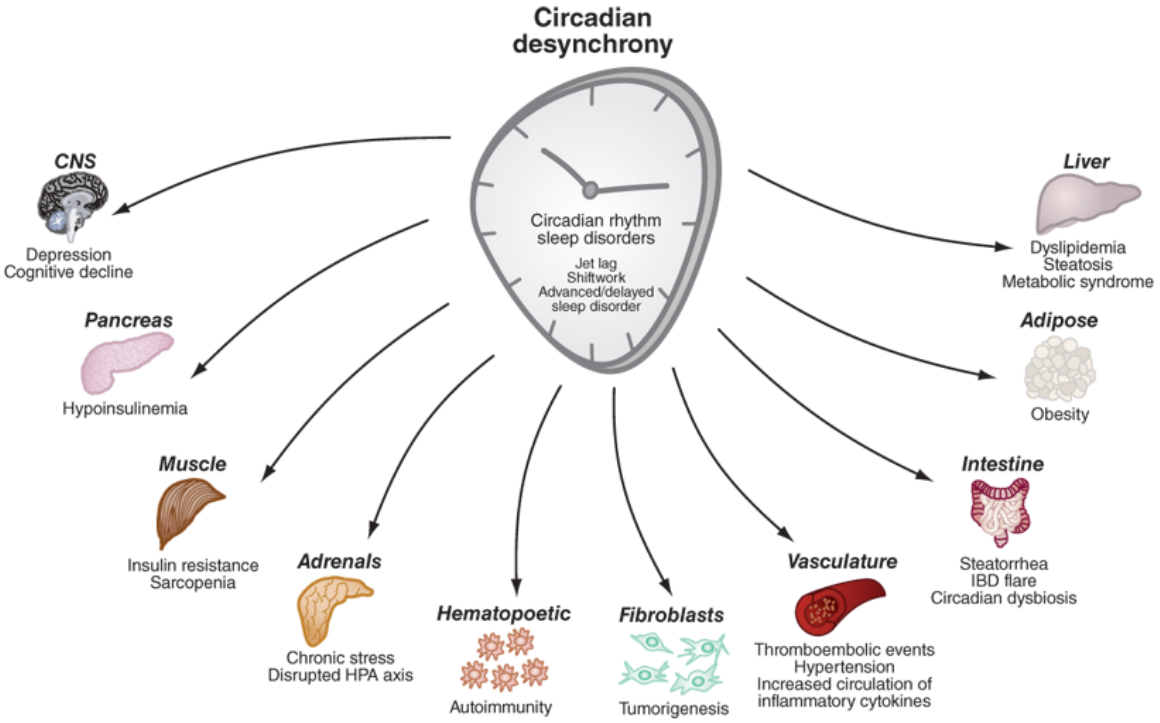
The circadian clock partitions behavioral, physiologic, and metabolic processes according to time of day. The partitioning of metabolic processes to appropriate times of day is critical for the maintenance of health from cellular to mammalian organisms. This figure highlights which processes peak within CNS, muscle, adipose, liver, and pancreas during either the sleep-fasting or wake-feeding cycle in humans.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Figure 475-4

Pathologies resulting from circadian desynchrony. Circadian rhythm sleep disorders, including advanced/delayed phase sleep disorder, jetlag, social jetlag, and shift work, result in a desynchrony between the environmental light-dark cycle “time” and the endogenous clock “time.” Pathologies can thus arise through misalignment imposed by exogenous factors (e.g., altered light cycle, feeding rhythm) as well as endogenous factors, such as mutations in core clock genes. Such desynchrony results in a host of wide-ranging pathologies across multiple tissues, including hypoinsulinemia (pancreas), disrupted HPA axis, autoimmunity, hypertension, obesity, and metabolic syndrome.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

ROLE OF THE CLOCK SYSTEM IN PHYSIOLOGY

Endocrine Systems Regulated by the Circadian Clock

In addition to regulation of behavioral rhythms such as sleep/wake and fasting/feeding cycles, the circadian clock also regulates rhythms of the endocrine system. Cortisol rhythms are regulated through a feedback loop known as the hypothalamic-pituitary (HPA) axis **(Chap. 379)**. Hypothalamic secretion of CRH and AVP promotes secretion of pituitary adrenocorticotropic hormone (ACTH), which in turn regulates rhythmic cortisol secretion from the adrenal cortex. Cortisol release increases towards the morning, and this increase is believed to prepare the brain and peripheral tissues for daytime activity and food intake. Daytime sleep can blunt circulating cortisol levels, presumably through the occurrence of non-REM sleep. AVP secretion in mice occurs prior to sleep to promote water intake, thereby preventing dehydration during the sleep period. Several hormonal systems are in fact influenced more by sleep than by circadian rhythms. For instance, secretion of growth hormone (GH) is profoundly blunted during acute overnight wakefulness. The secretion of this hormone is primarily dependent on the occurrence of slow-wave sleep, which is a homeostatically driven sleep stage that occurs primarily in the first part of the sleep period. Cortisol also exhibits a peak close after wakefulness: the cortisol awakening response (CAR). This peak seems to be independent of a circadian rhythm, as the CAR is severely blunted by acute overnight wakefulness. Curtailed sleep and overnight wakefulness increase the activity of the HPA axis and may increase diurnal cortisol levels. Sleep also influences melatonin amplitude, such that sleep deprivation can increase melatonin levels. In working environments, the effects of curtailed sleep are often confounded by mistimed exposure to light. Even low levels of light are able to potently suppress melatonin secretion. Together with altered timing in light exposure, perturbed hormonal levels may represent a mechanism through which altered timing and duration of sleep may impact central and peripheral circadian oscillators.

Centrally controlled rhythms of melatonin and cortisol are considered key regulators of extra-SCN and peripheral oscillators. Glucocorticoid receptors exist in both the central nervous system and in peripheral tissues such as skeletal muscle, liver, and adipose tissue. Following acute shifts in light-dark or feeding cycles, 24-h rhythms of circulating cortisol appear to shift more slowly than other rhythms and may thus contribute to adverse effects of circadian misalignment by hampering proper realignment of peripheral clocks. Glucocorticoids shift clock gene expression in muscle, kidney, and lung, while the powerful synthetic glucocorticoid dexamethasone is able to synchronize (e.g., reset) circadian rhythms of cells in culture, including liver cells. Consistent with a role for glucocorticoid regulation of the clock, both adrenalectomy, which results in a lack of cortisol, and exogenous corticosteroid supplementation significantly disrupt the circadian clock system.

Several peripherally produced hormones and peptides are not only produced rhythmically but can also feedback to central clocks, including the SCN. For instance, both cortisol and thyroid hormones regulate their own rhythmic synthesis by feedback to central brain regions, i.e., the hypothalamus (for cortisol) and pituitary (for both hormones). Several other peripherally produced factors have been proposed to influence the central clock, including fatty acids produced by the adipose tissue and fibroblast growth factor 21, a hormone primarily produced by the liver. Peripheral hormones that signal energy state and hunger also exhibit circadian rhythms. The most extensively studied hormones are leptin, which is released from white adipose tissue cells, and ghrelin, which is released from the upper fundus region of the stomach. Ghrelin also exhibits significant peaks related to anticipated meal timing, which persist for several days of fasting in humans. Circulating rhythms of leptin and ghrelin are disrupted in circadian mutant mice and are also perturbed in

humans subjected to circadian misalignment. For instance, *Per* and *Cry* mutant mice exhibit severely blunted leptin rhythms, and wild-type mice exposed to jetlag—through repeatedly altered light-dark cycles—show a reduced wake-associated decrease in leptin. Similarly, humans forced to live 28-h days exhibit increased 24-h profiles of ghrelin, and conversely decreased levels of leptin. Ghrelin and leptin signal to several regions of the brain, including integrative appetitive regions of the hypothalamus such as the arcuate and paraventricular region. Through actions in several such central sites, these hormones influence rhythms of food intake and energy homeostasis in a nutrient-dependent manner.

Role for the Clock in Metabolic Homeostasis

Circadian control of glucose homeostasis has long been recognized, as early studies demonstrated variation in glucose tolerance and insulin action across the day. For example, oral glucose tolerance is impaired in the evening and afternoon compared with the morning due to a combination of circadian control of both peripheral insulin sensitivity and pancreatic β-cell insulin secretion. Another example is the "dawn phenomenon," whereby glucose levels peak prior to the onset of activity. Further, destruction of the SCN has been shown to abolish circadian regulation of glucose metabolism in rats, and daily cycles of insulin secretion and glucose tolerance are often perturbed in patients with type 2 diabetes. Changes have also been observed in first-degree relatives of patients with type 2 diabetes, possibly highlighting a key hereditary component of the circadian clock in the pathogenesis of type 2 diabetes.

Ablating clock genes in mice has revealed a key function for both central and peripheral clocks in regulating energy homeostasis. The circadian system has been shown to regulate rhythmic insulin secretion from the pancreas via both neural signals and hormonal levels (e.g., cortisol and norepinephrine), as well as via cell autonomous clock regulation within the pancreatic β-cell itself. An early observation was that whole-body mutant $Clock^{\Delta 19/\Delta 19}$ mice developed obesity without displaying hyperinsulinemia, a phenomenon that indicated concurrent β-cell failure. This was later confirmed using pancreas- and β-cell–specific *Bmal1*-deficient mice, which exhibited glucose intolerance, hypoinsulinemia, and impaired glucose-stimulated insulin secretion. The molecular clock within other peripheral tissues such as liver, adipose tissue, and skeletal muscle also regulate circadian fluctuations in insulin sensitivity and glucose disposal, which are highest in the morning and decline towards the evening. Liver-specific *Bmal1* mutant studies have revealed liver clock promotion of gluconeogenesis, glycogenolysis, and mitochondrial oxidative metabolism in the sleep/fasting period while promoting glycogen synthesis in the wake/feeding period. Muscle-specific *Bmal1* deficient mice display reduced glucose tolerance, concomitant with lower levels of proteins involved in glucose uptake by muscle cells (e.g., the glucose transporter GLUT4). Ablation of the *Cry1* and *Cry2* repressors in the negative limb of the clock alters glucagon and glucocorticoid signaling in the liver, contributing to hyperglycemia and impaired glucose tolerance in these mutant mice. Together, these genetic studies in mice suggest a role for tissue-specific clocks in the partitioning of energy utilization across the sleep-wake cycle.

Importantly, peripheral clocks also interact with other environmental factors such as diet and time of feeding. For example, high-fat feeding leads not only to obesity and metabolic syndrome in mice, but also to perturbed clock gene expression across multiple peripheral tissues and a disrupted sleep-wake/fasting-feeding cycle, as revealed by increased activity and feeding during the daytime. Furthermore, mice that are fed a high-fat diet exclusively during the light phase gain significantly more weight than mice that are fed the same diet during the dark period—the active period for mice. Additionally, the metabolic phenotypes arising from ad lib high-fat feeding can be significantly ameliorated by restricting the time of high-fat feeding exclusively to the dark period. Time-restricted feeding can also increase the activity of brown adipose tissue in mice and reduce hepatic glucose production to instead promote beta oxidation of fatty acids. These findings have been corroborated in human interventional studies, which have demonstrated that time-restricted feeding can improve metabolic homeostasis and promote weight loss. Time-restricted feeding may also modulate central regulation of sleep and hunger, as a study found that humans who restricted their food intake to a shorter than ad lib period also consumed less daily calories and reported improved sleep.

Finally, animal studies have further shown that when the light-dark cycle is disrupted or animals are subjected to conditions mimicking "jetlag"—by artificially advancing or delaying the daily light period—there is desynchronization amongst circadian clocks and subsequent weight gain. Accumulating evidence in humans suggests that circadian misalignment both disrupts and desynchronizes circadian clocks across tissues. Prolonged circadian misalignment using forced desynchrony protocols reduces insulin sensitivity in the pre- and postprandial state. Under such conditions, insulin secretion fails to suppress glucose levels, suggesting inadequate β-cell compensation. Moreover, resting metabolic rate declines significantly both in the awake and sleeping state, altogether providing potential explanations why shift work can increase the risk of obesity, type 2 diabetes, and the metabolic syndrome.

Human genetic association studies also support a role for clock genes in metabolic homeostasis and beta cell function. Carriers of a certain *BMAL1* polymorphism have a greater risk of developing type 2 diabetes, while *CLOCK* variants have been found to interact with diet, such that variants can have a protective effect on insulin sensitivity in individuals with high monounsaturated fat intake or in individuals provided a low-fat diet. Instead, the minor allele of another *CLOCK* gene variant has been associated with increased waist circumference, but only in those with high saturated fat intake. Similarly, *NPAS2* and *BMAL1* variants have been associated with a greater risk of hypertension. Melatonin receptor *MTNR1B* gene variants, which result in increased expression of MTNR1B, have been associated with elevated fasting blood glucose levels and reduced insulin secretion irrespective of their level of glycemic control, consistent with the known effect of melatonin on insulin secretion and lower insulin secretion in the evening. These association studies highlight the role of the circadian system in metabolism, as well as potential for interactions of external perturbations—such as circadian misalignment—with a protective or adverse genetic profile.

A large proportion of society recurrently shifts sleep-wake times between working/non-free days and free days. This social jetlag has been increasingly tied to metabolic disruptions, including a greater risk of obesity and type 2 diabetes. As this involves recurrent phase advances and phase delays—like shift work but of smaller magnitude—it is possible that social jetlag also results in perturbed energy expenditure, in combination with disruptions to the circadian

rhythm of hunger drive, further increasing the risk of obesity. Repeated shifts in the food- and SCN-driven rhythm of insulin release may similarly over time increase the risk of type 2 diabetes. Shifted feeding rhythms in relation to the sleep-wake cycle and the timing of SCN activity may be causally involved in this pathogenesis. This is exemplified by the disorders known as night-eating syndrome and sleep-related eating disorder. In the former syndrome, a large part of daily calorie consumption occurs in the evening and nighttime hours, and this shifted meal pattern has been associated with a delayed timing of the internal clock. Some evidence exists that these syndromes are associated with obesity. Individuals who report sleeping fewer hours or who are subjected to restricted sleep for a few consecutive days have also been found to consume more calories later in the evening, perhaps explaining why sleep restriction increases the risk of obesity. These associations have also been observed in individuals with later onset of sleep, i.e., evening chronotypes. Night-eating syndrome and later chronotypes have also been linked to type 2 diabetes and may be more common than other eating disorders such as binge-eating disorder. Both conditions have also been found to be associated with impaired glycemic control—such as a greater likelihood of hemoglobin A1c values exceeding 7%—in patients already suffering from type 2 diabetes, emphasizing how proper alignment of internal circadian rhythms with external factors are key contributing factors for long-term metabolic homeostasis.

Circadian Clocks in Relation to Brain Health and Cognition

Molecular circadian clocks are present not only within the extra-SCN regions of the brain but also in neurons, astrocytes, and microglia. Emphasizing the functional significance of properly aligned clocks for brain health, shift workers have been found to have decreased grey matter in brain regions involved in memory and executive functions, with more notable effects in individuals who had shorter recovery periods between the onset of each shift work cycle. Adults performing rotating shift work for many years have also been shown to exhibit signs of accelerated cognitive aging. Notably, evidence suggests that these effects may be reversible, as those who have stopped carrying out shift work exhibit normal cognitive performance 5 or more years later.

Recent studies have also uncovered an important role for perturbed circadian and sleep-wake rhythms in neurodegenerative conditions such as Alzheimer's disease (AD), Huntington disease (HD), and Parkinson's disease (PD). Amyloid beta, a key pathognomonic component of AD, normally exhibits circadian fluctuations in the extracellular space in the brain, as well as in the cerebrospinal fluid and plasma in humans, peaking during the active period and falling during sleep. Of note, these daily rhythms of amyloid beta accumulation are dampened in mice that are prone to develop AD; reduced fluctuations in plasma amyloid beta fluctuations have also been noted in older compared with younger individuals. It is believed that removal of amyloid beta (and other neurotoxic substances) during the nighttime sleep period is facilitated by a lymphatic-like system that relies on glial cells (the "glymphatic" system). Whereas the function of this system has been shown to be important in mice, its circadian components and relevance to humans remain to be determined. Consistent with a role for circadian rhythms in the pathogenesis of AD, ablation of core clock genes throughout the brain or within subregions of the brain increases oxidative stress and neuronal cell death, while promoting scarring of brain tissue (astrogliosis). Furthermore, perturbed light-dark cycles increased pathology associated with oxidative stress, and single-nucleotide polymorphisms in *Clock* and *Bmal1* have been associated with increased risk of developing AD.

Evidence also indicates that the relationship between the circadian/sleep-wake system and AD is bidirectional. For example, patients suffering from AD exhibit several signs of perturbed circadian rhythms, the most prominent of such phenomena being "sundowning," whereby AD patients become more agitated and exhibit delirium-like symptoms in the afternoon or evening. Studies have furthermore indicated that in severe forms of AD, the circadian rhythm is phase delayed. Aged AD-prone mice also display perturbed sleep-wake patterns, which can be corrected by immunization against amyloid beta or by an orexin antagonist. Further research will help uncover the primary pathogenic contribution of the circadian system, and its independent contribution from perturbed sleep, in conditions like AD. Notably, evidence suggests that interventions that increase daytime light exposure and include melatonin supplementation are able to ameliorate symptoms of AD, presumably by counteracting disrupted circadian rhythms.

While the relation between shift work and depression has not been extensively studied, disruption of sleep and circadian rhythms and the pathogenesis of depression are intimately interlocked. Clock genes have also been implicated in depression and mood both in animal and human studies. Polymorphisms of genes that regulate sleep and circadian rhythms—for instance, a long gene variant of *PER3*—have also been linked to bipolar disorder and schizophrenia, while *CRY2* and *CLOCK* gene polymorphisms are associated with seasonal affective disorder, a type of depression arising in the fall and winter months when the levels of sunlight are lowest. Bipolar disorder is furthermore often triggered by circadian disruptions or curtailed sleep. Both bipolar disorder and schizophrenia have been linked to various forms of circadian disruption following disease onset, and a critical component of disease treatment often involves normalizing sleep and sleep-wake rhythms.

Sleep deprivation by itself is known to reduce alertness, impair decision-making, and increase risk for accidents—after 18–24 h of continuous wakefulness, several skills exhibit the same degree of decline as following mild alcohol intoxication. However, cognitive abilities may suffer even further when sleep restriction is combined with circadian misalignment as in shift work. In one study, participants were subjected to ~33-h long days in parallel with reduced sleep (equivalent to 5.6 h sleep in a 24-h period), yielding a forced desynchrony protocol coupled with sleep loss. When subjects were tested at the nadir of their circadian period, the subjects' reaction speed dropped almost by an order of magnitude compared with controls. In another study, researchers noted almost a 36% greater incidence of serious medical errors in resident interns who regularly worked 24-h or longer shifts compared with those who were randomly assigned to work up to 16-h long shifts. Furthermore, errors that resulted in patient death were three times more likely to occur in residents working extended hours compared with those who only worked up to 16-h long shifts.

Circadian Regulation of Gastrointestinal Homeostasis and the Microbiota

Physiologic aspects of the gastrointestinal (GI) tract exhibit day-night variations that anticipate and prepare for food intake and digestion during the active period. Gastric emptying, as well as colonic motility, are considerably greater during the active phase, as the phasic motor program supporting movement of digested material along the intestine is approximately twice as fast during the day compared with night. Bile acid secretion also exhibits circadian rhythmicity in the intestine, as does absorption and the expression of many nutrient uptake transporters in the intestinal wall, including the main glucose transporter protein SGLT1. The permeability of the intestinal wall also varies throughout the sleep-wake cycle, and mice exposed to chronic sleep fragmentation exhibit increased intestinal permeability, which may enable inflammatory molecules from bacteria to reach the systemic circulation.

The composition and function of the microbe population living in the intestine (i.e., the gut microbiota) also display circadian rhythmicity, orchestrated by both host circadian clock gene expression and food intake rhythms. Accordingly, circadian disruption, either by environmental or genetic means, perturbs these microbial rhythms, disrupting both bacterial levels and the metabolic functions of the gut microbiota. For example, alterations in the expression and functions of the gut microbiota have been noted in humans exposed to acute jetlag, and evidence suggests that curtailing sleep, which often accompanies shift work and jetlag, can alter the gut microbiota. By increasing local and systemic inflammation, circadian disruption of the gut microbiota may be causally involved in the increased risk of inflammatory bowel disease (Crohn's disease and ulcerative colitis) and colon cancer in shift workers. Gender-specific differences have also been reported, as female mice display more pronounced microbial rhythms. Interestingly, the gut microbiome has also been shown to influence the rhythms of host tissues, such as the intestine and liver, suggesting that a bidirectional relationship exists between tissues that regulate metabolic processes and the gut microbiome across the sleep-wake cycle. These findings may furthermore have clinical implications, given that the gut microbiome may both directly (in the gut lumen) and indirectly (through host-microbiota interactions) impact pharmacokinetic and pharmacodynamic properties of therapeutic drugs across the 24-h day-night cycle.

Cardiovascular Health and the Circadian Clock

An early epidemiologic observation was a greater incidence of myocardial infarction in the morning hours, with the lowest risk during the period preceding sleep. Other cardiovascular outcomes such as sudden cardiac death and syncope also exhibit a daily peak in the morning. Blood pressure (BP) typically peaks around 21:00 h and decreases later during sleep, partially due to a circadian nighttime dip of around 3–6 mmHg in systolic BP and 2–3 mmHg in diastolic BP. A dip in blood pressure of either less than 10% or greater than 20% during normal sleep has been associated with worse cardiovascular prognosis. Heart rate also typically decreases during sleep, while mistimed sleep leads to higher heart rate during their sleep time. Thus, a combination of reasons—which may also involve altered glucocorticoid levels and increased platelet aggregation—may contribute to a greater risk of cardiovascular disease in the morning. Subsequent epidemiologic studies also have demonstrated that shift work increases the risk of dyslipidemia and hypertension, as well as the risk of coronary heart disease, including myocardial infarction. These findings are in line with interventional findings in which circadian misalignment has been induced either by inverting the sleep-wake cycle or by imposing 28-h days on healthy human subjects. These studies have found that circadian misalignment elevates 24-h blood pressure, particularly during sleep. These changes may be causally related to how the autonomic system is regulated during sleep, as evidenced by reduced vagal cardiac control when the sleep-wake cycle is inverted.

Circadian Disruption and Cancer

In 2007, the International Agency for Research on Cancer declared that shift work that involves circadian disruption is likely carcinogenic to humans. While evidence for an association between shift work and general cancer incidence is mixed, accruing evidence supports a link between shift work and increased risk of developing colon and breast cancer, as well as having a poorer cancer prognosis. Telomere shortening, a phenomenon in aging that destabilizes the genome, has also been observed in shift workers as well as in individuals suffering from short sleep. This may reduce the ability of damaged or senescent cells to undergo apoptosis, and instead lead to uninhibited cell growth and cancer. An indirect role for the circadian clock has also come from retrospective studies on how cancer risk is related to food timing and duration of the nighttime fast in humans. These studies indicate that by portioning food intake to a restricted period of the day, circadian processes are optimized in a way that confers reduced risk of carcinogenic cell damage. Studies of recurring fasting have also been shown to lower the risk as well as to delay the onset of cancer.

Experimental genetic evidence has also implicated clock disruption as a factor in tumorigenesis. Genetic loss of *Per2* or *Bmal1* has been shown to promote lung tumorigenesis, while studies in *Per2* mutant mice have also revealed increased radiation-induced lymphoma associated with dysregulation of the cell cycle. However, disruption of the *Cry* gene in mice has also been implicated in tumor protection due to increased susceptibility to cell death. Thus, while both epidemiologic and experimental evidence suggests a link between circadian disruption and cancer, a full understanding of the role of circadian systems in tumorigenesis remains an area for investigation.

Circadian Regulation of the Immune System

Circadian misalignment and sleep restriction both alter population levels of immune cells and decrease the ability of immune cells to produce reactive radicals. Chronic circadian disruption may thereby impair the immune system's ability to conduct immunosurveillance at the proper time of day and reduce the ability to mount an appropriate response upon exposure of pathogens during the recovery/sleep phase, when the immune system is typically more active. Instead, circadian misalignment increases a range of clinically used inflammatory markers (e.g., C-reactive protein, tumor necrosis factor α, and interleukin 6), and such changes have been noted even when the sleep-wake cycle is only prolonged to a slightly longer than normal 24.6-h day. While similar effects are also observed following acute total sleep deprivation or recurrent partial sleep restriction, circadian misalignment has been found to promote an even more pronounced elevation of such markers. Genetic clock disruption in peritoneal macrophages has also revealed clock control of Toll-

like receptor 9, which is responsible for identifying molecules from foreign pathogens. *Clock* knockout mice also have reduced T cell antigen response, and mice immunized during the day had a stronger T cell response than mice immunized at night, supporting regulation of the immune system by the clock.

Aging and the Circadian Clock

Instability in the clock system is an often overlooked hallmark of aging. Aging is associated with a decline in the robustness of intrinsic rhythmic processes at the behavioral, physiologic, and molecular levels in both human and animal models. At the *behavioral* level, aging leads to reduced and fragmented sleep, dampened locomotor activity and feeding rhythms, and a reduced ability to entrain to light, as old rodents are 20 times less sensitive to the entraining effects of light relative to young animals. Even middle-aged individuals exposed to jetlag also exhibit more symptoms of circadian misalignment, such as more time awake and reduced alertness, compared with young individuals. On a *physiologic* level, some of the hallmarks of aging are a reduction in amplitude (e.g., flattening of circadian pattern) and a phase-advance (e.g., a shift in the timing of the peak or nadir) in rhythms of the endocrine and neuroendocrine systems, including sleep onset and offset. For example, cortisol, DHEA, and melatonin all have dampened rhythms and are phase-advanced in aging; the combination of such changes may, for instance, contribute to more fragmented sleep and lower levels of restorative slow-wave sleep in aged individuals. Aging also leads to alterations in peptide expression in the SCN (VIP and AVP) and reduced amplitude of rhythms of SCN electrical activity. Further, while the SCN-dependent body temperature rhythm—a generally accepted marker for the integrity of circadian rhythms—peaks in the evening and is lowest in the early morning in young individuals, aged healthy subjects display a phase advance and decrease in circadian amplitude in body temperature rhythms. Indeed, evidence suggests that internal desynchrony between core body temperature rhythms and the sleep-wake cycle may contribute to age-associated circadian alterations. On a *molecular* level, aging is associated with decreased expression and altered diurnal profiles of several of the core clock genes, including *Clock* and *Bmal1*, within both SCN and peripheral tissues such as heart and liver. Interestingly, the acute induction of *Per1* in response to light was markedly reduced in the SCN of aged mice compared with young mice, potentially contributing to their delayed response to light entrainment. Mice lacking *Bmal1* die prematurely compared with control mice, consistent with premature accumulation of reactive oxygen species. These mice have an accelerated onset of numerous age-related pathologies, including cataracts, sarcopenia, reduced organ size, and decreased hair growth. Instead, deficiency of *Cryptochrome*, a repressor of the internal clock repressor, has been associated with alterations in liver regeneration, while BMAL1 and PER2 may be important for proper neurogenesis in the hippocampus, a brain region in which adult mammals normally exhibit continuous cell division. Altogether, this suggests that the highly conserved circadian clock is important for regulating a wide range of homeostatic processes, including cell-cycle pathways, that when properly phased to each other promote organismal fitness.

Measurements of altered circadian rhythms with age may serve as a useful biomarker for aging. An intriguing question is whether the decline in amplitude of rhythms correlates with a decline in function, and importantly whether restoration of these rhythms with age, through either behavioral or pharmacologic intervention, would delay the aging process. Of note, transplantation of the SCN from a young rat into an old rat “rescued” the rhythms of both locomotor activity and corticotropin hormone (CRH), suggesting that the SCN is an important target for age-related changes in clocks. Treatments targeting the SCN may therefore ameliorate some of the deterioration in aged individuals.

CHRONOTHERAPY AND FUTURE DIRECTIONS

Chronopharmacology, the study of how the timing of drug administration may impact its effectiveness, is a rapidly emerging field. Since physiologic processes vary across the day, the timing of administration of medication may help optimize patient care. For example, since endogenous cholesterol synthesis is rhythmic in liver and peaks during the early morning hours, administration of statins (HMG-CoA reductase inhibitors) in the evening prior to bedtime has proven to be more effective than daytime administration at reducing low-density lipoprotein cholesterol (LDL-C) levels because the highest concentration of the medications coincides with the peak in the rhythmic endogenous cholesterol production. Given that blood pressure exhibits a 24-h rhythm—being lowest during sleep—angiotensin-converting enzyme (ACE) inhibitors have been shown to be most effective at night to normalize the blood pressure rhythms, restoring the nighttime dip in blood pressure. Administration of cancer treatments according to circadian rhythms has also been shown to increase chemotherapy effectiveness while decreasing toxicity in a wide range of drugs. For example, 5-FU works best to treat colorectal cancer when administered at night, a time when the cancerous cells are more vulnerable while normal cells are quiescent and therefore less sensitive. Doxorubicin administration early in the morning to treat ovarian cancer has also been shown to be less toxic, as white blood cells recover faster than if the drug is given in the evening. Finally, the more severe morning symptoms of rheumatoid arthritis are linked to increased inflammation during the evening; therefore, prevention of the night-time upregulation of the immune/inflammatory reaction is more effective when glucocorticoids are administered with a night-time release formulation.

Recognition of circadian rhythms is also critical for diagnoses and treatment of endocrine disorders. The diagnosis of Cushing’s syndrome, which is characterized by hypercortisolemia, might be missed if the patient’s cortisol levels were measured in the morning, as endogenous cortisol production is typically highest during morning and lowest at night; therefore, clinical diagnosis requires cortisol to be measured in the late evening when the levels of this hormone should typically be low. On the other hand, adrenal insufficiency is diagnosed by measuring cortisol in the morning when at its physiologic peak, and glucocorticoid therapy for these patients aims to mimic the endogenous rhythms of cortisol, as short-acting synthetic glucocorticoids are usually given several times a day in tapering doses, such that the largest amount is taken in the morning and the smallest in the evening. Diabetes is another endocrine disorder intimately tied to circadian rhythms. Oral glucose tolerance has repeatedly been shown to be impaired in the afternoon and evening compared to the morning, likely due to a combination of a circadian regulation of insulin sensitivity within peripheral tissues and reduced insulin secretion during the night. Similarly, due to a surge in hormone levels in the morning, diabetes patients may suffer from the dawn phenomenon (or dawn effect), an abnormally

high morning increase in blood glucose due to impaired response in insulin secretion. A related phenomenon that can be tied to evening timing of insulin doses is the “rebound” or Somogyi effect. In this scenario, the initially noted clinical sign, in the form of elevated glucose levels, may be noted in the morning. However, the underlying cause is hypoglycemia occurring during the night, which produces a counterregulatory hormonal response that subsequently results in morning hyperglycemia. As patients with type 2 diabetes often lack these daily cycles of insulin secretion and glucose tolerance, this further highlights that time of day is an important consideration for the diagnosis and treatment of metabolic disorders such as type 2 diabetes.

As our knowledge of the complexity of the interaction between circadian and physiologic processes deepens, further advances to rationally develop new strategies for treatments of disorders affected by circadian misalignment are essential. For example, novel compounds have begun to emerge from unbiased drug discovery screens that impact circadian clock components, either shortening or lengthening the period, including CRY stabilizers or various inhibitors of CKIδ, CKIε, and GSK-3. Pharmacologic control of the circadian cycle may be useful in the treatment of circadian disorders and metabolic disturbances with a circadian component. Understanding how the circadian clock controls biological functions will shed new light onto the pathogenesis of metabolic disorders with a circadian component, such as type 2 diabetes and metabolic syndrome, and will yield insight into how timing of drug delivery will impact patient care.

ACKNOWLEDGMENT

The authors would like to thank Billie Marcheva for her help with the figures and tables.

FURTHER READING

Bass J, Lazar MA: Circadian time signatures of fitness and disease. Science 354:994, 2016. [PubMed: 27885004]

Buxton OM et al: Adverse metabolic consequences in humans of prolonged sleep restriction combined with circadian disruption. Sci Transl Med 4:129ra43, 2012. [PubMed: 22496545]

Dibner C et al: The mammalian circadian timing system: Organization and coordination of central and peripheral clocks. Annu Rev Physiol 72:517, 2010. [PubMed: 20148687]

Hatori M et al: Time-restricted feeding without reducing caloric intake prevents metabolic diseases in mice fed a high-fat diet. Cell Metabolism 15:848, 2012. [PubMed: 22608008]

Nea FM et al: Dietary and lifestyle habits and the associated health risks in shift workers. Nutr Res Rev 28:143, 2015. [PubMed: 26650243]

Perelis M et al: Pancreatic β cell enhancers regulate rhythmic transcription of genes controlling insulin secretion. Science 350:aac4250, 2016.

Scheer FA et al: Adverse metabolic and cardiovascular consequences of circadian misalignment. Proc Natl Acad Sci USA 106:4453, 2009. [PubMed: 19255424]

Takahashi JS: Transcriptional architecture of the mammalian circadian clock. Nat Rev Genet 18:164, 2016. [PubMed: 27990019]

Turek FW et al: Obesity and metabolic syndrome in circadian clock mutant mice. Science 308:1043, 2005. [PubMed: 15845877]

Chapter 476: Network Medicine: Systems Biology in Health and Disease

Joseph Loscalzo

INTRODUCTION

The field of human biology has progressed over the last three centuries largely as a result of the reductionist approach to the scientific problems that challenge the discipline. Biologists study the experimental response of a variable of interest in a cell or organism while holding all other variables constant. In this way, it is possible to dissect the individual components of a biologic system and assume that a thorough understanding of a specific component (e.g., an enzyme or a transcription factor) will provide sufficient insight to explain the global behavior of that system (e.g., a metabolic pathway or a gene network, respectively). Biologic systems are, however, much more complex than this approach assumes and manifest behaviors that frequently (if not invariably) cannot be predicted from knowledge of their component parts characterized in isolation. Growing recognition of this shortcoming of conventional biologic research has led to the development of a new discipline, *systems biology, which is defined as the holistic study of living organisms or their cellular or molecular network components to predict their response to perturbations*. Concepts of systems biology can be applied readily to human disease and therapy and define the field of *systems pathobiology*, in which genetic or environmental perturbations produce disease and drug perturbations restore normal system behavior.

Systems biology evolved from the field of systems engineering in which a linked collection of component parts constitute a network whose output the engineer wishes to predict. The simple example of an electronic circuit can be used to illustrate some basic systems engineering concepts. All the individual elements of the circuit—resistors, capacitors, transistors—have well-defined properties that can be characterized precisely. However, they can be linked (wired or configured) in a variety of ways, each of which yields a circuit whose response to voltage applied across it is different from the response of every other configuration. To predict the circuit's (i.e., system's) behavior, the engineer must study its response to perturbation (e.g., voltage applied across it) holistically rather than its individual components' responses to that perturbation. Viewed another way, the resulting behavior of the system is greater than (or different from) the simple sum of its parts, and systems engineering utilizes rigorous mathematical approaches to predict these complex, often nonlinear, responses. By analogy to biologic systems, one can reason that detailed knowledge of a single enzyme in a metabolic pathway or of a single transcription factor in a gene network will not provide sufficient detail in context to predict the output of that metabolic pathway or transcriptional network, respectively. Only a systems-based approach will suffice.

It has taken biologists a long time to appreciate the importance of systems approaches to biomedical problems. Reductionism has reigned supreme for many decades, largely because it is experimentally and analytically simpler than holism, and because it has provided insights into biologic mechanisms and disease pathogenesis that have led to successful therapies. However, reductionism cannot solve all biomedical problems. For example, the so-called off-target effects of new drugs that frequently limit their adoption likely reflect the failure of a drug to be studied in holistic context, that is, the failure to explore all possible actions aside from the principal target action for which it was developed. Other approaches to understanding biology are, therefore, clearly needed. With the growing body of genomic, proteomic, and metabolomic data sets in which dynamic changes in the expression of many genes and many metabolites are recorded after a perturbation and with the growth of rigorous mathematical approaches to analyzing those changes, the stage has been set for applying systems engineering principles to modern biology.

Physiologists historically have had more of a (bio)engineering perspective on the conduct of their studies and have been among the first systems biologists. Yet, with few exceptions, they, too, have focused on comparatively simple physiologic systems that are tractable using conventional reductionist approaches. Efforts at integrative modeling of human physiologic systems, as first attempted by Guyton for blood pressure regulation, represent one application of systems engineering to human biology. These dynamic physiologic models often focus on the acute response of a measurable physiologic parameter to a system perturbation, and do so from a classic analytic perspective in which all the conventional physiologic determinants of the output parameter are known and can be modeled quantitatively.

Until recently, molecular systems analysis has been limited owing to inadequate knowledge of the molecular determinants of a biologic system of interest. Although biochemists have approached metabolic pathways from a systems perspective for over 50 years, their efforts have been limited by the inadequacy of key information for each enzyme (K_m, k_{cat}, and concentration) and substrate (concentration) in the pathway. With increasingly rich molecular data sets available for systems-based analyses, including genomic, transcriptomic, proteomic, and metabolomic data, molecular biologists and biochemists are now poised to use systems biology approaches to explore biologic and pathobiologic phenomena.

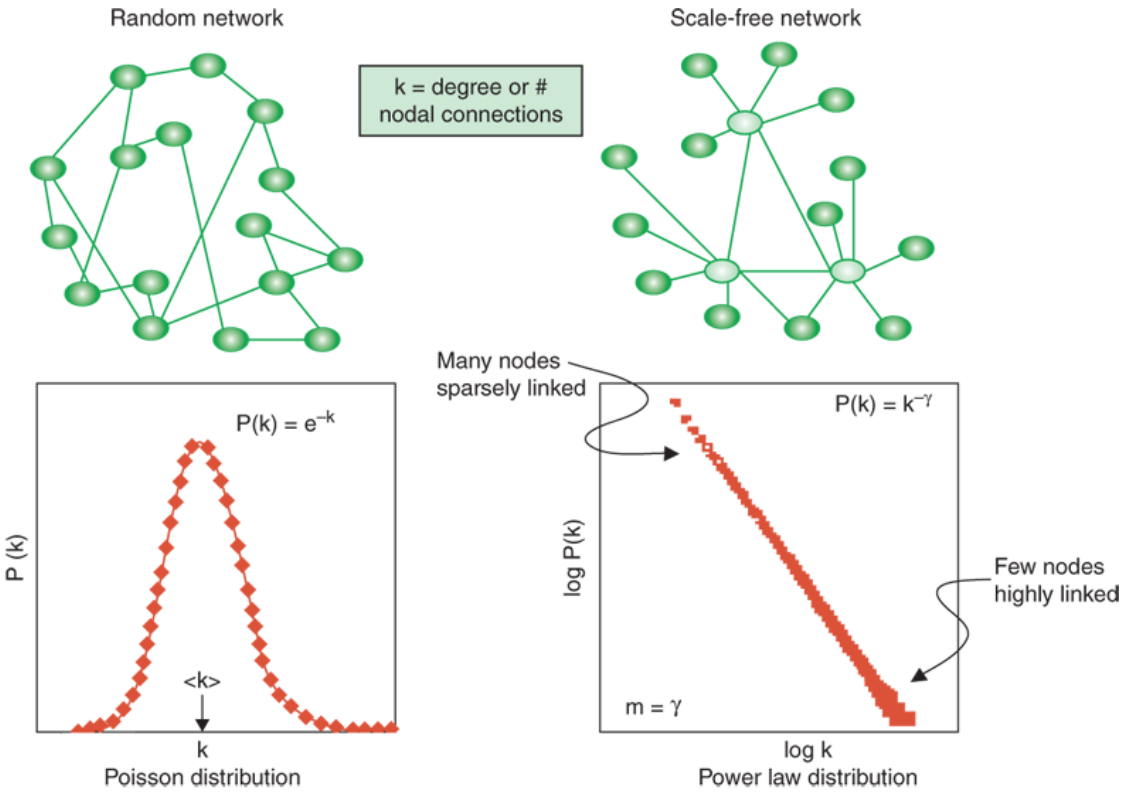
PROPERTIES OF COMPLEX BIOLOGIC SYSTEMS

To understand how best to apply the principles of systems biology to human biomedicine, it is necessary to review briefly the building blocks of any biologic system and the determinants of system complexity. All systems can be analyzed by defining their static topology (architecture) and their dynamic (i.e., time-dependent) response to perturbation. In the discussion that follows, system properties are described that derive from the consequences of topology (form) or dynamic response (function). Any system of interacting elements can be represented schematically as a *network* in which the individual elements are depicted as nodes and their connections are depicted as links. The nature of the links among nodes reflects the degree of complexity of the system. *Simple systems* are those in which the nodes are linearly linked with occasional feedback or feedforward loops modulating system throughput in highly predictable ways. By contrast, *complex systems* are nodes that are linked in more complicated, nonlinear networks; the behavior of these systems by definition is inherently more difficult to predict owing to the nature of the interacting links, the dependence of the system's behavior on its initial conditions, and the inability to measure the overall state of the system at any specific time with great precision. Complex systems can be depicted as a network of lower-complexity interacting components or modules, each of which can be reduced further to simpler analyzable canonical motifs (such as feedback and feedforward loops, or negative and positive autoregulation); however, a central property of complex systems is that simplifying their structures by identifying and characterizing the individual nodes and links or even simpler substructures does not necessarily yield a predictable understanding of a system's behavior. Thus, the functioning system is greater than (or different from) the sum of its individual, tractable parts.

Defined in this way, most biologic systems are complex systems that can be represented as networks whose behaviors are not readily predictable from simple reductionist principles. The nodes, for example, can be metabolites that are linked by the enzymes that cause their transformations, transcription factors that are linked by the genes whose expression they influence, or proteins in an interaction network that are linked by cofactors that facilitate interactions or by thermodynamic forces that facilitate their physical association. Biologic systems typically are organized as *scale-free*, rather than stochastic, networks of nodes. Scale-free networks are those in which a few nodes have many links to other nodes (highly linked nodes, or hubs), but most nodes have only a few links (weakly linked nodes). The term *scale-free* refers to the fact that the connectivity of nodes in the network is invariant with respect to the size of the network. This is quite different from two other common network architectures: random (Poisson) and exponential distributions. Scale-free networks can be mathematically described by a power law that defines the probability of the number of links per node ($P(k) = k^{-\gamma}$, where k is the number of links per node and γ is the slope of the log P(k) versus log(k) plot); this unique property of most biologic networks is a reflection of their self-similarity or fractal nature **(Fig. 476-1)**.

FIGURE 476-1

Network representations and their distributions. A random network is depicted on the left, and its Poisson distribution of the number of nodal connections (k) is shown in the graph below it. A scale-free network is depicted on the right, and its power law distribution of the number of nodal connections (k) is shown in the graph below it. Highly connected nodes (hubs) are lightly shaded.



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

There are unique properties of scale-free biologic systems that reflect their evolution and promote their adaptability and survival. Biologic networks likely evolved one node at a time in a process in which new nodes are more likely to link to a highly connected node than to a sparsely connected node. Furthermore, scale-free networks can become sparsely linked to one another, yielding more complex, *modular scale-free topologies*. This evolutionary

growth of biologic networks has three important properties that affect system function and survival. First, this scale-free addition of new nodes promotes *system redundancy*, which minimizes the consequences of errors and accommodates adverse perturbations to the system robustly with minimal effects on critical functions (unless the highly connected nodes are the focus of the perturbation). Second, this resulting network redundancy provides a survival advantage to the system. In complex gene networks, for example, mutations or polymorphisms in weakly linked genes account for biodiversity and biologic variability without disrupting the critical functions of the system; only mutations in highly linked (*essential*) genes (hubs) can shut down the system and cause embryonic lethality. Third, scale-free biologic systems facilitate the flow of information (e.g., metabolite flux) across the system compared with randomly organized biologic systems; this so-called "small-world" property of the system (in which the clustered nature of the highly linked hubs defines a local neighborhood within the network that communicates through weaker, less frequent links to other clusters) minimizes the energy cost for the dynamic action of the system (e.g., minimizes the transition time between states in a metabolic network).

These basic organizing principles of complex biologic systems lead to three unique properties that require emphasis. First, biologic systems are *robust*, which means that they are quite stable in response to most changes in external conditions or internal modification. Second, a corollary to the property of robustness is that complex biologic systems are *sloppy*, which means that they are insensitive to changes in external conditions or internal modification except under certain uncommon conditions (i.e., when a hub is involved in the change). Third, complex biologic systems exhibit *emergent properties*, which means that they manifest behaviors that cannot be predicted from the reductionist principles used to characterize their component parts. Examples of emergent behavior in biologic systems include spontaneous, self-sustained oscillations in glycolysis; spiral and scroll waves of depolarization in cardiac tissue that cause reentrant arrhythmias; and self-organizing patterns in biochemical systems governed by diffusion and chemical reaction.

APPLICATIONS OF SYSTEMS BIOLOGY TO PATHOBIOLOGY

The principles of systems biology have been applied to complex pathologic processes with some early successes. The key to these applications is the identification of emergent properties of the system under study in order to define novel, otherwise unpredictable (i.e., from the reductionist perspective) methods for regulating the system's response. Systems biology approaches have been used to characterize epidemics and ways to control them, taking advantage of the scale-free properties of the network of infected individuals that constitute the epidemic. Through the use of a systems analysis of a neural protein-protein interaction network, unique disease-modifying proteins have been identified that are common to a wide range of cerebellar neurodegenerative disorders causing inherited ataxias. Systems analysis and disease network construction of a pulmonary arterial hypertension network led to the identification of a unique disease module involving a pathway governed by microRNA-21. Systems biology models have been used to dissect the dynamics of the inflammatory response using oscillatory changes in the transcription factor nuclear factor (NF) κB as the system output. Systems biology principles also have been used to predict the development of an idiotypy–anti-idiotypy antibody network, describe the dynamics of species growth in microbial biofilms, and analyze the innate immune response. In each of these examples, a systems (patho) biology approach provided insights into the behavior of these complex systems that could not have been recognized with conventional scientific reductionism.

A unique application of systems biology to biomedicine is in the area of drug development. Conventional drug development involves identifying a potential target protein and then designing or screening compounds to identify those that inhibit the function of that target. This reductionist analysis has identified many potential drug targets and drugs, yet only when a drug is tested in animal models or humans are the systems consequences of the drug's action revealed; not uncommonly, so-called off-target effects may become apparent and be sufficiently adverse for researchers to cease development of the agent. A good example of this problem is the unexpected outcomes of the vitamin B–based regimens for lowering homocysteine levels. In these trials, plasma homocysteine levels were reduced effectively; however, there was no effect of this reduction on clinical vascular endpoints. One explanation for this outcome is that one of the B vitamins in the regimen, folate, has a panoply of effects on cell proliferation and metabolism that probably offset its homocysteine-lowering benefits, promoting progressive atherosclerotic plaque growth and its consequences for clinical events. In addition to these types of unexpected outcomes exerted through pathways that were not considered *ab initio*, conventional approaches to drug development typically do not take into consideration the possibility of emergent behaviors of the organism or the metabolic pathway or the transcriptional network of interest. Thus, a systems-based analysis of potential drugs (drug-target network analysis) can benefit the development paradigm both by enhancing the likelihood that a compound of interest will not manifest unforeseen adverse effects and by promoting novel analytic methods for identifying unique control points or pathways in metabolic or genetic networks that would benefit from drug-based modulation, including drug combinations.

SYSTEMS PATHOBIOLOGY AND HUMAN DISEASE CLASSIFICATION: NETWORK MEDICINE

Perhaps most important, systems pathobiology can be used to revise and refine the definition of human disease. The classification of human disease used in this and all medical textbooks derives from the correlation between pathologic analysis and clinical syndromes that began in the nineteenth century. Although this approach has been very successful, serving as the basis for the development of many effective therapies, it has major shortcomings. Those shortcomings include a lack of sensitivity in defining preclinical disease, a primary focus on overtly manifest disease, failure to recognize different and potentially differentiable causes of common late-stage pathophenotypes, and a limited ability to incorporate the growing body of molecular and genetic determinants of pathophenotype into the conventional classification scheme.

Two examples will illustrate the weakness of simple correlation analyses grounded in the reductionist principle of simplification (Occam's razor) in defining human disease. Sickle cell anemia, the "classic" Mendelian disorder, is caused by a Val6Gln substitution in the β chain of hemoglobin. If conventional genetic teaching holds, this single mutation should lead to a single phenotype in patients who harbor it (genotype-phenotype correlation). This

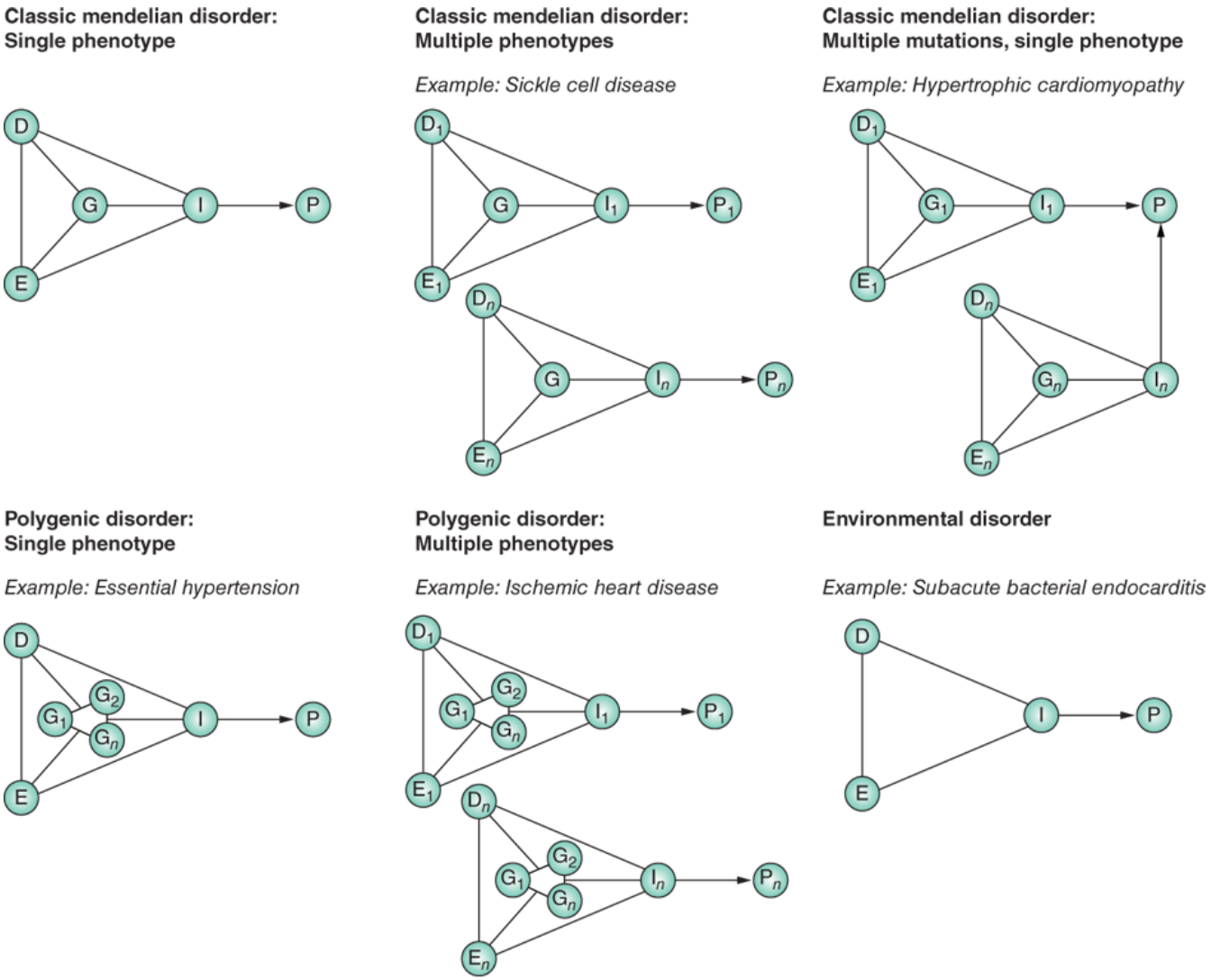
assumption is, however, false, as patients with sickle cell disease manifest a variety of pathophenotypes, including hemolytic anemia, stroke, acute chest syndrome, bony infarction, and painful crisis, as well as an overtly normal phenotype. The reasons for these different phenotypic presentations include the presence of disease-modifying genes or gene products (e.g., hemoglobin F, hemoglobin C, glucose-6-phosphate dehydrogenase), exposure to adverse environmental factors (e.g., hypoxia, dehydration), and the genetic and environmental determinants of common intermediate pathophenotypes or endophenotypes (i.e., variations in those generic pathologic mechanisms underlying all human disease—inflammation, thrombosis/hemorrhage, fibrosis, cell proliferation, apoptosis/necrosis, immune response).

A second example of note is familial pulmonary arterial hypertension. This disorder is associated with over 100 different mutations in three members of the transforming growth factor β (TGF-β) superfamily alone: bone morphogenetic protein receptor-2 (BMPR-2), activin receptor-like kinase-1 (Alk-1), and endoglin. All these different genotypes are associated with a common pathophenotype, and each leads to that pathophenotype by molecular mechanisms that range from haploinsufficiency to dominant negative effects. As only approximately one-fourth of individuals in families that harbor these mutations manifest the pathophenotype, other disease-modifying genes (e.g., the serotonin receptor 5-HT2B, the serotonin transporter 5-HTT), genomic and environmental determinants of common intermediate pathophenotypes, and environmental exposures (e.g., hypoxia, infective agents [HIV], anorexigens) probably account for the incomplete penetrance of the disorder.

On the basis of these and many other related examples, one can approach human disease from a systems pathobiology perspective in which each "disease" can be depicted as a network that includes the following modules: the primary disease-determining elements of the genome (or proteome, if posttranslationally modified), the disease-modifying elements of the genome or proteome, environmental determinants, and genomic and environmental determinants of the generic intermediate pathophenotypes. **Figure 476-2** graphically depicts these genotype-phenotype relationships as modules for the six common disease types with specific examples for each type. **Figure 476-3** shows a network-based depiction of sickle cell disease using this kind of modular approach.

FIGURE 476-2

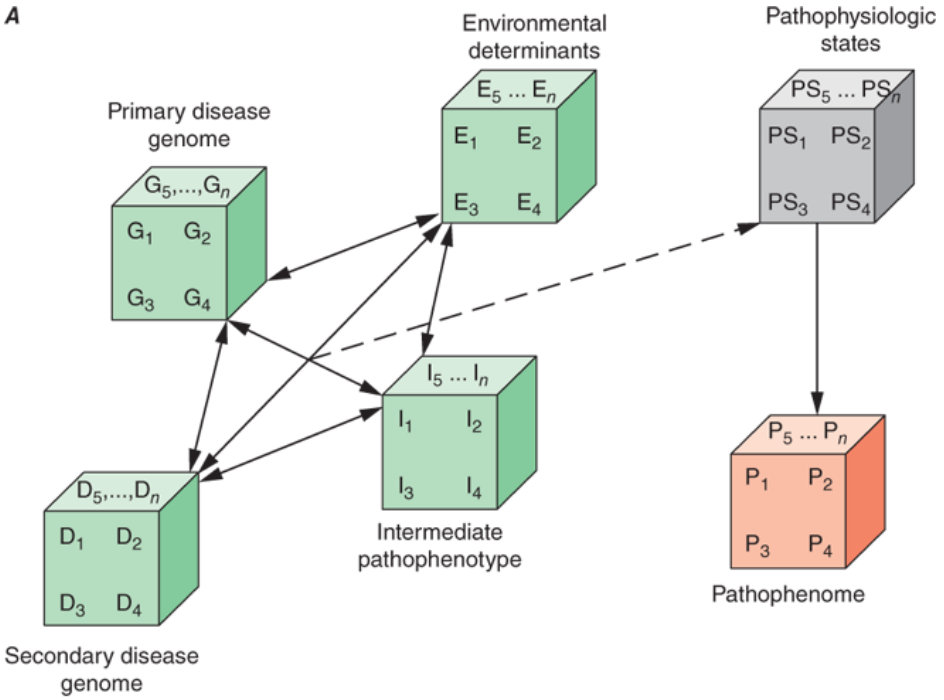
Examples of modular representations of human disease. D, secondary human disease genome or proteome; E, environmental determinants; G, primary human disease genome or proteome; I, intermediate phenotype; P, pathophenotype. *(Reproduced with permission from J Loscalzo et al: Molec Syst Biol 3:124, 2007.)*



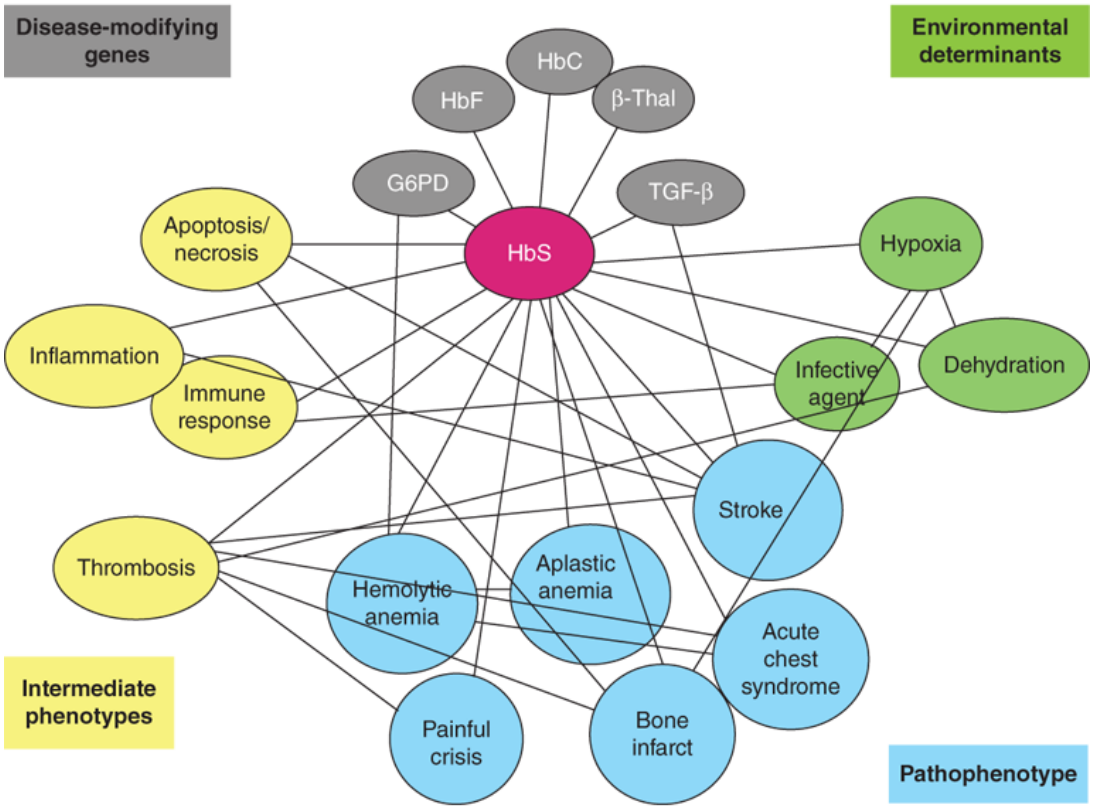
Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 476-3

A. Theoretical human disease network illustrating the relationships among genetic and environmental determinants of the pathophenotypes. Key: D, secondary disease genome or proteome; E, environmental determinants; G, primary disease genome or proteome; I, intermediate phenotype; PS, pathophysiologic states leading to P, pathophenotype. ***B.*** Example of this theoretical construct applied to sickle cell disease. Key: Red, primary molecular abnormality; gray, disease-modifying genes; yellow, intermediate phenotypes; green, environmental determinants; blue, pathophenotypes. *(Reproduced with permission from J Loscalzo et al: Molec Syst Biol 3:124, 2007.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

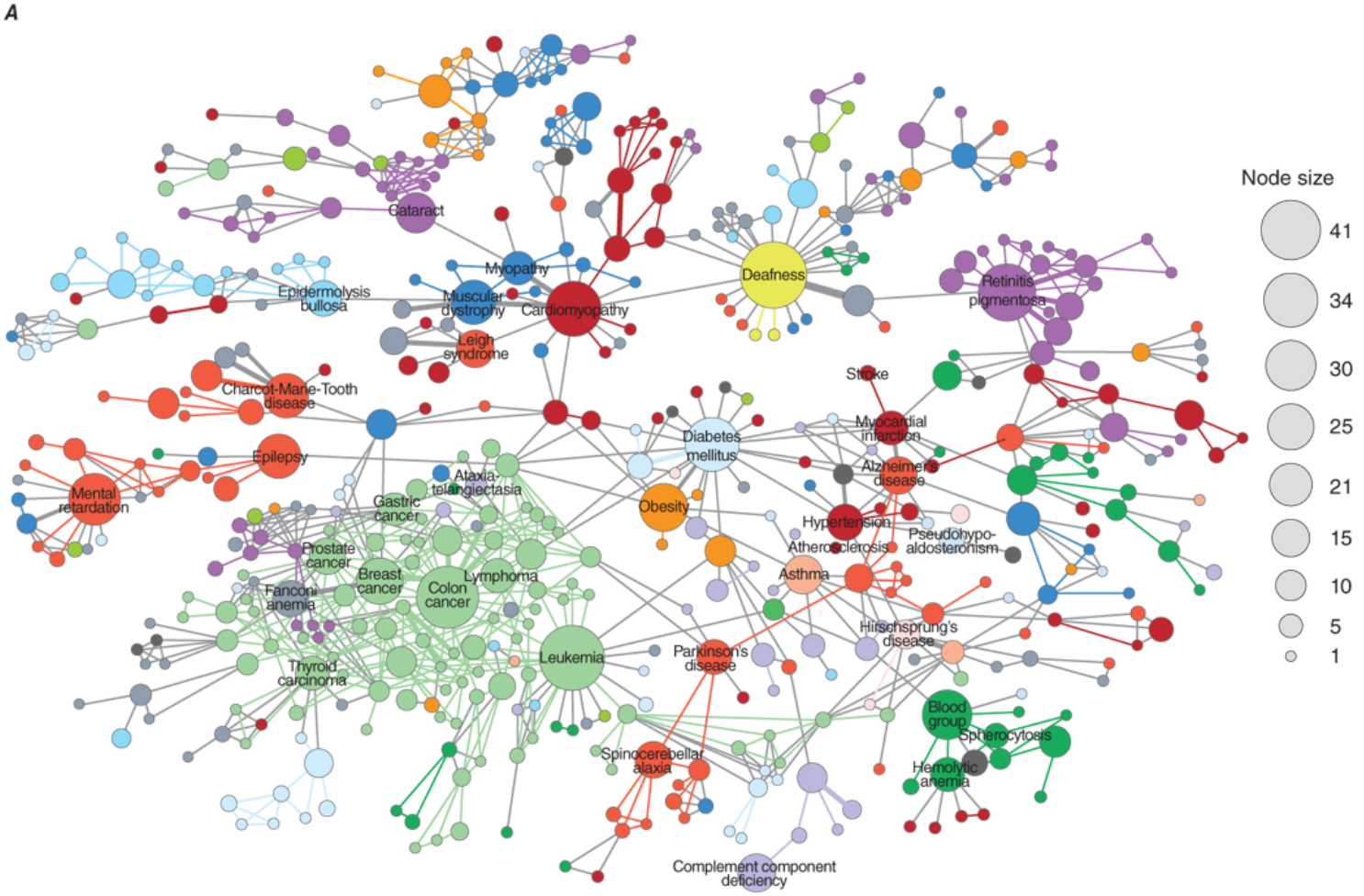


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

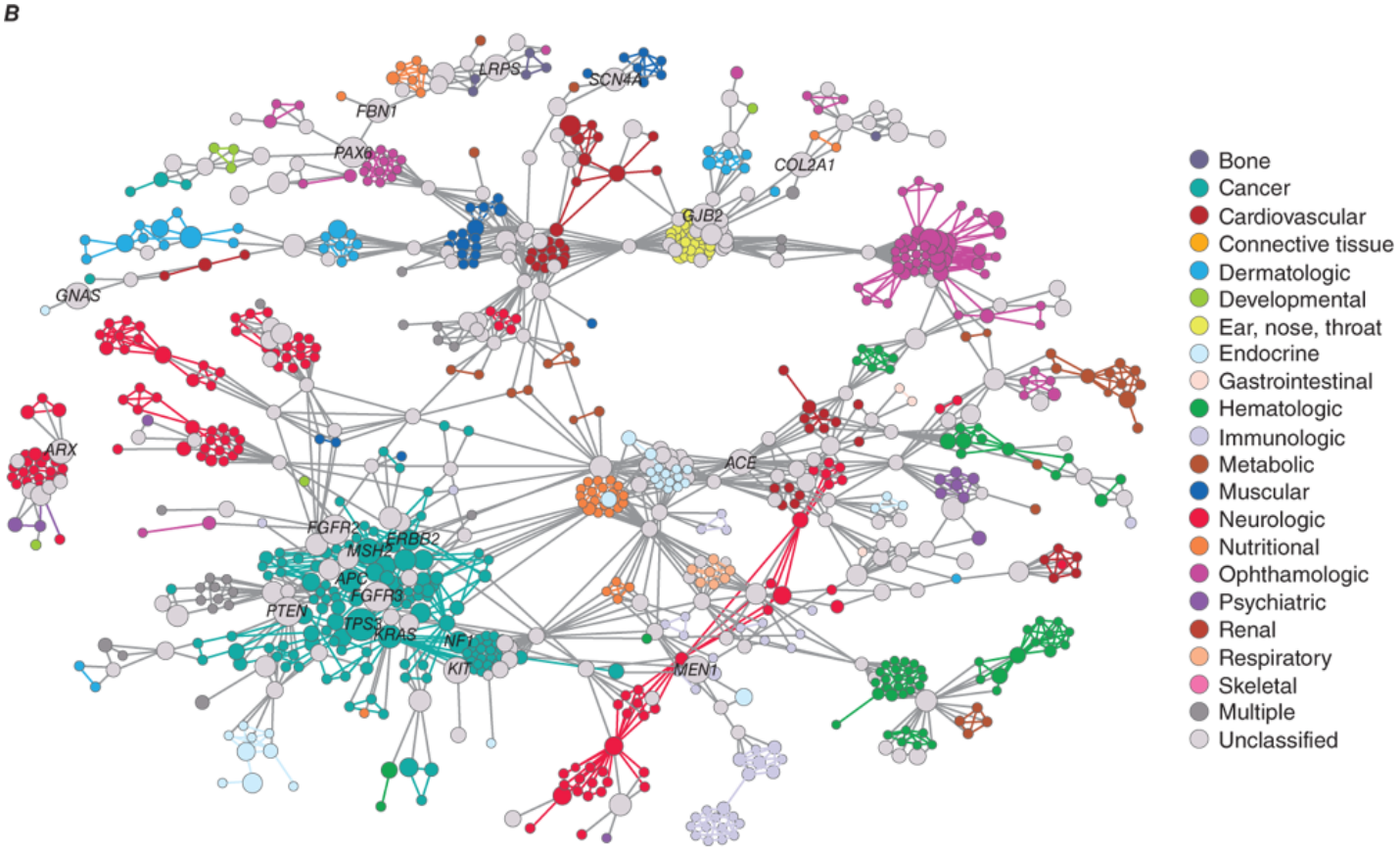
Goh and colleagues developed the concept of a human disease network (Fig. 476-4) in which they used a systems approach to characterize the disease-gene associations listed in the Online Mendelian Inheritance in Man database. Their analysis showed that genes linked to similar disorders are more likely to have products that physically associate and greater similarity between their transcription profiles than do genes not associated with similar disorders. In addition, proteins associated with the same pathophenotype are significantly more likely to interact with one another than with other proteins not associated with the pathophenotype. Finally, these authors showed that the great majority of disease-associated genes are not highly connected genes (i.e., not hubs) and are typically weakly linked nodes within the functional periphery of the network in which they operate.

FIGURE 476-4

A. Human disease network. Each node corresponds to a specific disorder colored by class (22 classes, shown in the key to **B**). The size of each node is proportional to the number of genes contributing to the disorder. Edges between disorders in the same disorder class are colored with the same (lighter) color, and edges connecting different disorder classes are colored gray, with the thickness of the edge proportional to the number of genes shared by the disorders connected by it. ***B.*** Disease gene network. Each node is a single gene, and any two genes are connected if implicated in the same disorder. In this network map, the size of each node is proportional to the number of specific disorders in which the gene is implicated. *(Reproduced with permission from KI Goh et al: Proc Natl Acad Sci USA 104:8685, 2007.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

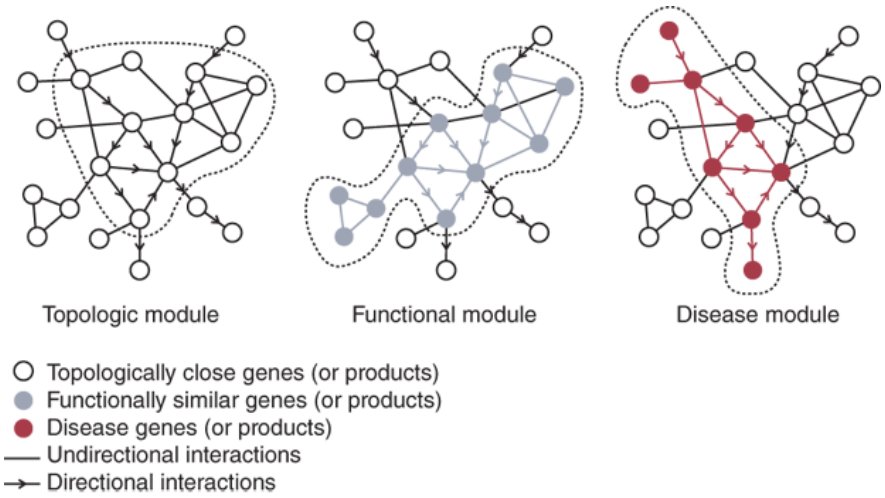


Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

This type of analysis validates the potential importance of defining disease on the basis of its systems pathobiologic determinants. Clearly, doing this will require a more careful dissection of the molecular elements in the relevant pathways (i.e., more precise molecular pathophenotyping), less reliance on overt manifestations of disease for their classification, and an understanding of the dynamics (not just the static architecture) of the pathobiologic networks that underlie pathophenotypes defined in this way. Figure 476-5 illustrates the elements of a molecular network within which a disease module is contained. This network is first identified by determining the interactions (physical or regulatory) among the proteins or genes that comprise it (the "interactome"). These interactions then define a topologic module within which exist functional modules (pathways) and disease modules. One approach to constructing this module is illustrated in Fig. 476-6. Examples of the use of this approach in defining novel determinants of disease are given in Table 476-1.

FIGURE 476-5

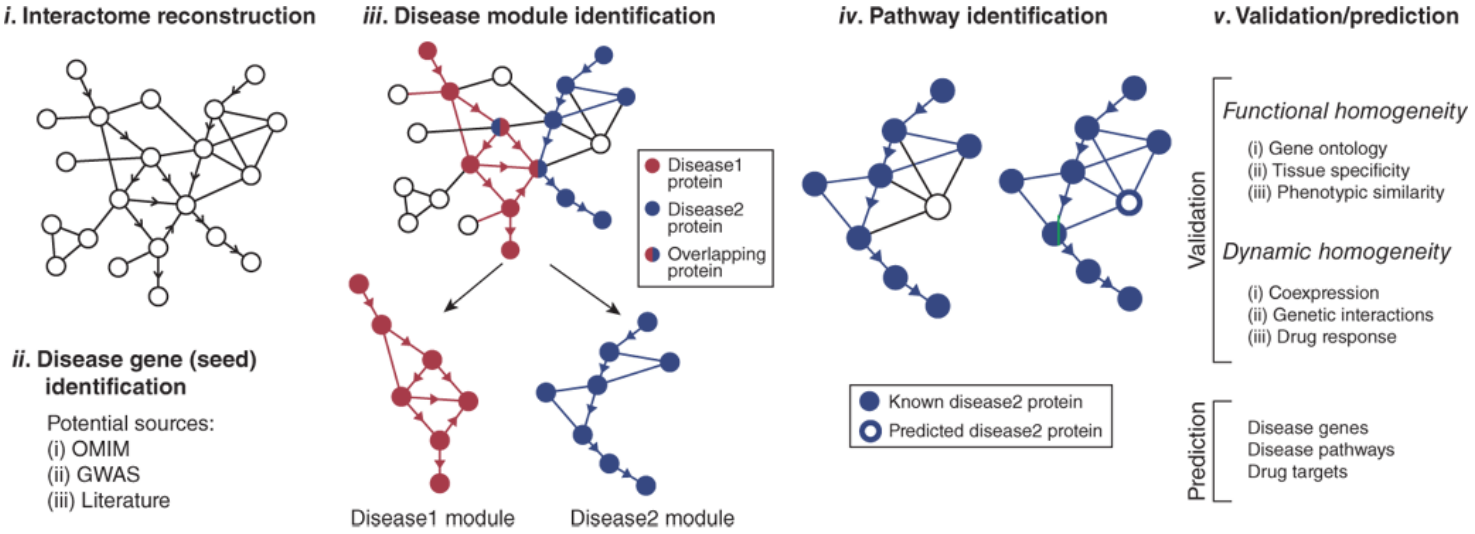
The elements of the interactome. The interactome includes topologic modules (genes or gene products that are closely associated with one another through direct interactions), functional modules (genes or gene products that work together to define a pathway), and disease modules (genes or gene products that interact to yield a pathophenotype). *(Reproduced with permission from AL Barabasi et al: Nat Rev Genet 12:56, 2011.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

FIGURE 476-6

Approaches to identifying disease modules in molecular networks. A strategy for defining disease modules involves (i) reconstructing the interactome; (ii) ascertaining potential seed (disease) genes from the curated literature, the Online Mendelian Inheritance in Man (OMIM) database, or genomic analyses (genome-wide association studies [GWAS] or transcriptional profiling); (iii) identifying the disease module using different modeling or statistical approaches; (iv) identifying pathways and the role of disease genes or modules in those pathways; and (v) disease module validation and prediction. *(Reproduced with permission from AL Barabasi et al: Nat Rev Genet 12:56, 2011.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

TABLE 476-1

Examples of Systems Biology Application to Disease

Disease	Analysis	Reference
Hereditary ataxias	Many ataxia-causing proteins share interacting partners that affect neurodegeneration	Lim et al: Cell 125:801–814, 2006
Diabetes mellitus	Metabolite-protein network analysis links three unique metabolite abnormalities in prediabetics to seven type 2 diabetes genes through four enzymes	Wang-Sattler et al: Mol Syst Biol 8:615, 2012
Ebstein-Barr virus infection	Viral proteome exerts its effects through linking to host interactome	Gulbahce et al: PLoS One 8:e1002531, 2012
Pulmonary arterial hypertension	Network analysis indicates adaptive role for microRNA 21 in suppressing rho kinase pathway	Parikh et al: Circulation 125:1520–1532, 2012

As yet another potential consideration, one can argue that disease reflects the later-stage consequences of the predilection of an organ system to manifest a particular intermediate pathophenotype in response to injury. This paradigm reflects a reverse causality view in which a disease is defined as a tendency to heightened inflammation, thrombosis, or fibrosis after an injurious perturbation. Where the process is manifest (i.e., the organ in which it occurs) is less important than that it occurs (with the exception of the organ-specific pathophysiologic consequences that may require acute attention). For example, from this perspective, acute myocardial infarction (AMI) and its consequences are a reflection of thrombosis (in the coronary artery), inflammation (in the acutely injured myocardium), and fibrosis (at the site or sites of cardiomyocyte death). In effect, the major therapies for AMI address these intermediate pathophenotypes (e.g., antithrombotics, statins) rather than any organ-specific disease-determining process. This paradigm would argue for a systems-based analysis that would first identify the intermediate pathophenotypes to which a person is predisposed, then determine how and when to intervene to attenuate that adverse predisposition, and finally limit the likelihood that a major organ-specific event will occur. Evidence for the validity of this approach is found in the work of Rzhetsky and colleagues, who reviewed 1.5 million patient records and 161 diseases and found that these disease phenotypes form a network of strong pairwise correlations. This result is consistent with the notion that underlying genetic predispositions to intermediate pathophenotypes form the predicate basis for conventionally defined end organ diseases.

Regardless of the specific nature of the systems pathobiologic approach used, these analyses will lead to a drastic revision of the way human disease is defined and treated, establishing the discipline of *network medicine, which reflects a fusion of the fields of systems biology and network science in the*

study of disease. This will be a lengthy and complicated process, but ultimately will lead to better disease prevention and therapy and probably do so from an increasingly personalized perspective. The analysis of pathobiology from a systems-based perspective is likely to help define specific subsets of patients more likely to respond to particular interventions based on shared disease mechanisms. Although it is unlikely that the extreme of "individualized medicine" will ever be practical (or even desirable), complex diseases can be mechanistically subclassified and interventions may be tailored to those settings in which they are more likely to work. This approach serves as a basis for the development of precision medicine.

FURTHER READING

Barabasi A-L, Gulbahce N, Loscalzo J: Network medicine: A network-based approach to human disease. Nat Rev Genet 12:56, 2011. [PubMed: 21164525]

Loscalzo J, Barabási A-L, Silverman EK (eds): *Network Medicine: Complex Systems in Human Disease and Therapeutics*. Cambridge, Massachusetts, Harvard University Press. Copyright © 2017 by the President and Fellows of Harvard College. All rights reserved.

Loscalzo J, Kohane I, Barabas A-L: Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. Mol Syst Biol 3:124, 2007. [PubMed: 17625512]

Menche J et al: Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science 347:1257601, 2015. [PubMed: 25700523]

Parikh VN et al: MicroRNA-21 integrates pathogenic signaling to control pulmonary hypertension: Results of a network bioinformatics approach. Circulation 125:1520, 2012. [PubMed: 22371328]

Chapter 477: Emerging Neurotherapeutic Technologies

Jyoti Mishra; Karunesh Ganguly

INTRODUCTION

Neurotherapeutic technologies represent a diverse group of very promising treatment approaches with a common purpose of improving neurological function. Decades of basic science research has paved the path for these novel technologies that have the potential to transform the lives of patients with neurological diseases. A key goal is to minimize the consequences of lost abilities, whether it is motor, sensory, or cognitive. A common objective is to also harness the inherent plasticity of the nervous system, regardless of age, and even in the face of a degenerative process.

The technologies described below are the culmination of both an increased understanding of neural plasticity mechanisms in both the intact and the injured nervous system as well as advances in technology and computational power. There has been important progress in understanding neural plasticity at the level of the microscale (e.g., cellular and molecular processes), the mesoscale (e.g., between distinct cortical and subcortical areas), and the macroscale (e.g., at the level of brain networks). While it is also clear that there may be fundamental limits on plasticity (e.g., the closing of developmental windows) and repair mechanisms, the brain remains highly plastic regardless of age and even in the face of ongoing injury and/or degenerative processes. Collectively, there is now growing evidence to support neurological restorative efforts for both “static” (e.g., stroke) and progressive neurological disorders.

Importantly, while these technologies may not appear, at first glance, directly relevant to traditional medical care, it is worth noting that clinicians have the most knowledge and experience about the specific disease process, the treatments available, and the expected course of illnesses affecting the nervous system. It is thus critical that neurologic specialists and other clinicians can and should play an important role in the future adoption of these technologies for neurological rehabilitation. The sections below outline emerging diagnostic and therapeutic approaches that have the potential to transform the lives of patients with neurological disorders. These include technologies to harness plasticity, neuroimaging, neurostimulation, and brain–machine interfaces (BMIs).

TECHNOLOGIES TO HARNESS PLASTICITY

Neurological rehabilitation aims to harness activity-dependent plasticity mechanisms to maximize functional restoration. This principle can be applied to a diverse range of functional domains such as movement control, sensory processing, language, pain, and cognition. For example, recent randomized controlled clinical trials for motor recovery after stroke have suggested that intensity of training may be particularly important for sustained long-term improvements. Moreover, studies of the effects of such training in rodent and nonhuman primate models further suggest that plasticity of cortical “motor maps” might underlie the observed functional improvements. The incorporation of technology for neurological rehabilitation has the great potential to revolutionize the delivery of care by significantly increasing access, reducing the burden for adherence to high intensity regimens, and by maximizing engagement. Below are three examples of how emerging technology can be used to harness neural plasticity and to maximize functional restoration.

ROBOTICS

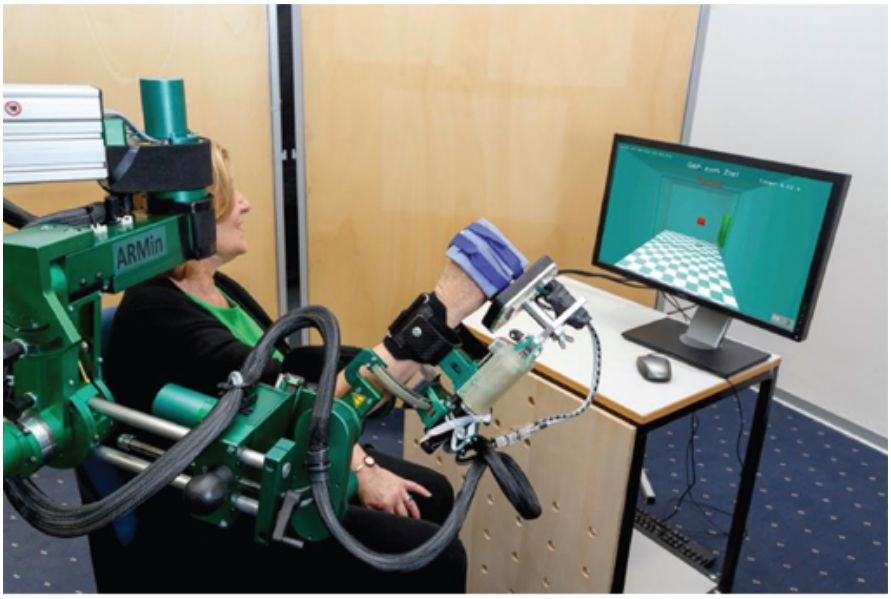
Rehabilitation robotics for both the upper and the lower limb have the potential to improve motor outcomes after stroke or other forms of brain injury. There is a growing recognition that focused training involving a range of tasks might be important for improved functional outcomes. While it remains unclear exactly when such training might be optimal after the initial stroke and during the early recovery period, such training likely has a role in both the acute and the chronic periods after stroke; maintenance therapy may also provide a guard against observed declines in function over time. Notably, the delivery of intensive training is a great challenge from both the perspective of the health care system and each patient. Outside of clinical trials, such a training program can be quite difficult to implement and maintain. It can also be costly and require significant effort.

Motor rehabilitation using robotics has been developed and tested for both the upper-limb and the lower-limb. Such robotic therapies have often focused on the delivery of high-intensity movement practice that can surpass what is possible via existing standards of care. Moreover, the robotic systems are capable of precisely measuring movement parameters (e.g., the kinematics of the movements) and providing quantitative feedback regarding the changes in performance during the training period. A particular focus has been on maximal patient engagement and recruitment of attentional and reward pathways, both of which are increasingly recognized to drive neural plasticity. Continued advances in design and the user interface will ensure maximal comfort and sustained effort. For example, via close monitoring of performance and movement parameters, the system can provide assistance at key points in order to minimize fatigue and to ensure maximal engagement. Moreover, antigravity support of the upper-limb can allow practice and task engagement even in the presence of severe weakness; this would be extremely challenging and labor intensive under current standards of care. Recent analysis also suggests that robotic devices may at least match outcomes realized with existing standards of care. However, rehabilitation robotics may also provide more precise feedback and permit novel quantitative rehabilitation approaches.

Figure 477-1 shows one example of an upper-limb robotic exoskeleton device that is currently being evaluated for training after stroke. A recent randomized, multicenter trial compared treatment with this exoskeleton system against conventional therapy provided by physical and occupational therapists. Participants were enrolled in the chronic phase and all had moderate-to-severe deficits; the groups underwent three sessions per week over an 8-week period. For robotic training, subjects trained with games to improve mobilization and to practice activities of daily living. This study provided evidence that both conventional and robotic therapy could improve function in patients with chronic stroke. Multiple studies have also found similar gains when using either conventional or traditional approaches. Thus, a growing body of research supports the idea that such devices might complement current conventional approaches to rehabilitation. Future work will need to define how rehabilitation robotics can optimally use adaptive and quantitative methods to further augment the recovery process.

FIGURE 477-1

Photograph of a subject interacting with a complex upper-limb exoskeleton and a virtual reality system. *(From U Keller et al: Robot-Assisted Arm Assessments in Spinal Cord Injured Patients: A Consideration of Concept Study. PLoS One 10:e0126948, 2015.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

VIRTUAL AND AUGMENTED REALITY

Therapeutic approaches using Virtual Reality (VR) and Augmented Reality (AR) aim to treat neurological illnesses by specifically and quantitatively altering a patient's subjective experiences and interactions with the environment. Core components of both are advanced hardware and computational methods to generate simulated, yet realistic, perceptions. While some applications permit users to dynamically change the viewed perspective, other applications are designed to allow interactions among multiple users. Visual feedback is often a key component; this can include simple computer monitors or more immersive "head mounted" viewers that modify the simulation based on changes in perspective. Tracking of movements (e.g., hand and head position) are often included. Multiple methods are used to allow a user to interact with the environment. For example, interactions can be guided by straightforward means such as a keyboard, mouse or even a joystick. More immersive methods are also frequently used. For example, gloves with embedded sensors and haptic inputs can allow the user's hand to be represented in real time in the simulated environment. Moreover, haptic interfaces can provide sensory feedback, allowing patients to interact and "feel" virtual objects through multiple sensory modalities. A particular strength of these approaches is that any therapeutic intervention can be studied in very controlled environments.

VR enables a user to interact with a simulated reality that can be precisely and quantitatively controlled. In addition to allowing patients to dynamically experience an altered reality, it can simultaneously monitor a subject's behaviors and responses. Such monitoring can allow both precise measurements of clinically relevant parameters (e.g., motor actions, perception, cognitive processing) and for specific rehabilitation training that can attain clinically relevant goals. A growing body of literature indicates that VR environments can be tailored to individual needs and preferences, thereby maximizing engagement, motivation, and adaptation to ensure sufficient difficulty of tasks. VR environments can be designed to create powerful "gaming" platforms that are actually targeting clinically relevant parameters. For example, the upper-limb robotic systems described previously are frequently combined with VR environments that allow interaction with virtual objects.

In contrast to VR, AR overlays an artificial filter over a subject's view of the actual physical world, thus providing an "augmented" or enhanced view of the world around. AR is being tested in a diverse group of patients with neurological impairments in either the motor, sensory, or cognitive domains. AR may offer a particularly unique rehabilitation intervention for stroke patients. It is widely known that brain injuries limit each patient's physical interaction with their environment. Furthermore, physical and cognitive impairments may limit social interactions. Such impoverished experiences are likely to be present during both the acute and the chronic phases. Importantly, there is clear basic science evidence that environmental

enrichment can be a key component of rehabilitation; such enrichment may offer additive benefits to the often limited formal rehabilitation sessions per week. Consistent with this are clinical studies suggesting that motor and cognitive outcomes may suffer when interactions with the environment are reduced. AR may be capable of increasing enrichment. For example, in the case of spatial neglect after stroke, the impaired modality may be accounted for using AR methods. Similarly, physical impairments that limit walking speeds can also limit visual feedback; both AR and VR can be used to enhance visual feedback during gait training.

Figure 477-2 shows a recent innovative application of AR for the treatment of "phantom limb" pain. A subset of both upper-limb and lower-limb amputees experience painful sensations that appear to originate from the missing limb. Past research has suggested that mirror therapy can be an effective treatment for phantom limb pain. During mirror therapy treatments, patients move their healthy arm in front of a mirror in order to produce a perception of movements of the missing limb. Previous studies have suggested that maladaptive plasticity of affected sensory cortices may be treated with mirror therapy. Importantly, in comparison to mirror therapy, AR-based therapy for phantom limb pain can be based on movements of the affected limb, i.e., the remaining limb portion as opposed to using the unaffected contralateral limb. This study demonstrated a novel treatment in which "phantom motor execution" is enabled using sophisticated machine learning algorithms. More specifically, the study "decoded" phantom limb movements by measuring electromyogram (EMG) activity at the stump. Importantly, while the distal muscles responsible for movements were lost as a result of amputation, the remaining EMG activity could be used to predict presumed distal limb movements. As shown in the figure, these inferred movements were projected onto an AR screen to create the perception of limb movements. The study showed that a subset of patients with long-term refractory phantom limb pain could experience a significant reduction in pain levels after using the AR system.

FIGURE 477-2

Augmented reality (AR) for phantom limb pain. ***A.*** A patient is shown live Augmented Reality (AR) video. ***B.*** Electromyogram (EMG) electrodes placed over the stump record muscle activation during training. ***C.*** The patient matches target postures during rehabilitation. ***D.*** Patient playing a game in which a car is controlled by "phantom movements." *(From M Ortiz-Catalan et al: Phantom motor execution facilitated by machine learning and augmented reality as treatment for phantom limb pain: A single group, clinical trial in patients with chronic intractable phantom limb pain. Lancet 388:2885, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

NEUROGAMING

Computerized programs that harness the power of “video games” have shown some evidence for ameliorating deficits in visual perception, age-related degeneration, and neuropsychiatric disorders. An essential feature of effective video game training is the progressive adjustment of the level of difficulty in line with the cognitive improvement of the patient. Important areas of active research include ways to enhance sustainability of neurogame training over long time periods and improving training *transfer*, i.e., the generalizability of task-specific training in one cognitive domain to more broad-based functional improvements. By leveraging video game technology, neurogames allow for dynamic user interaction and maintain user-engagement over multiple sessions over several days of training. Important game mechanics include repetitive practice, performance-adaptive challenges, and several layers of reward feedback—from moment-to-moment point rewards to reward milestones over multiple sessions.

Notably, neurogames have therapeutic potential as they can be targeted to specific neurocognitive deficits. For instance, games have shown significant benefits in aging, by targeting speed of processing and training the abilities to multitask and suppress distractions. In each case, selective targeting is achieved by focusing the adaptive challenges to the neurocognitive domain of interest. Duration of response time windows available to the user or the level of interference is selectively targeted in the case of speed of processing training and interference training, respectively. In more recent research, it was demonstrated that it is possible to engender focused circuit neuroplasticity using such selective targeting in neurogaming. For example, older adults learned to adaptively perform within progressively more challenging distractor environments. Neuroplasticity selective to distractor processing was evidenced in this study at both the microscale, i.e., at the resolution of single neuron spiking in sensory cortex, as well as macroscale, i.e., EEG (electroencephalography) based event-related potential recordings.

Video games have also shown promise in the treatment of visual deficits such as amblyopia and in cognitive remediation in neuropsychiatric disorders such as schizophrenia. However, while the evidence base has been encouraging in small sample randomized controlled studies (RCTs), larger RCTs are needed to demonstrate definitive therapeutic benefit. This is especially necessary as the commercial *brain training* industry continues to make unsubstantiated claims of the benefits of neurogaming; such claims have been formally dismissed by the scientific community. Like any other pharmacological or device-based therapy, neurogames need to be systematically validated in multi-phase RCTs establishing neural target engagement, and documenting cognitive and behavioral outcomes in specific disorder populations.

Generalizability of training benefits from task-specific cognitive outcomes to more broad-based functional improvements remains the holy grail of neurogaming. Next-generation neurogames will aim to integrate physiological measures such as heart rate variability (an index of physical exertion), galvanic skin responses and respiration rate (indices of stress response), and even EEG-based neural measures. The objectives of such multimodal biosensor integration are to enhance the “closed-loop mechanics” that drive game adaptation and hence improve therapeutic outcomes and perhaps result in greater generalizability. These complex, yet potentially more effective, neurogames of the future will need rigorous clinical study for demonstration of validity and efficacy.

NEUROIMAGING

NEUROIMAGING OF CONNECTIVITY

Multimodal neuroimaging methods including fMRI (functional magnetic resonance imaging), EEG, and MEG (magnetoencephalography) are now being investigated as tools to study functional connectivity between brain regions, i.e., extent of correlated activity between brain regions of interest. Snapshots of functional connectivity can be analyzed while an individual is engaged in specific cognitive tasks or during rest. Resting state functional connectivity (rsFC) is especially attractive as a robust, task-independent measure of brain function that can be evaluated in diverse neurological and neuropsychiatric disorders. In fact, methodological research has shown that rs-fMRI can provide more reliable brain signals of energy consumption than specific task-based fMRI approaches.

In recent years, there has been a surge of research to identify robust rsFC-based biomarkers for specific neurological and neuropsychiatric disorders and thereby inform diagnoses, and even predict specific treatment outcomes. For many such disorders, the network level neurobiological substrates that correspond to the clinical symptoms are not known. Furthermore, many are not unitary diseases, but rather heterogeneous syndromes composed of varied co-occurring symptoms. Hence, the research quest for robust network biomarkers for complex neuropsychological disorders is challenging and still in its infancy. However, some studies have made significant headway in this domain. For example, in a large multisite cohort of ~1000 depressed patients, Drysdale et al. (2017) recently showed that rsFC measures can subdivide patients into four neurophysiological “biotypes” with distinct patterns of dysfunctional connectivity in limbic and frontostriatal networks. These biotypes were associated with different clinical-symptom profiles (combinations of anhedonia, anxiety, insomnia, anergia,

etc.) and had high (>80%) diagnostic sensitivity and specificity. Moreover, these biotypes could also predict responsiveness to TMS (transcranial magnetic stimulation) therapy. Another recent study demonstrated utility of rsFC measures to predict diagnosis of mTBI (mild traumatic brain injury), which is clinically challenging by conventional means.

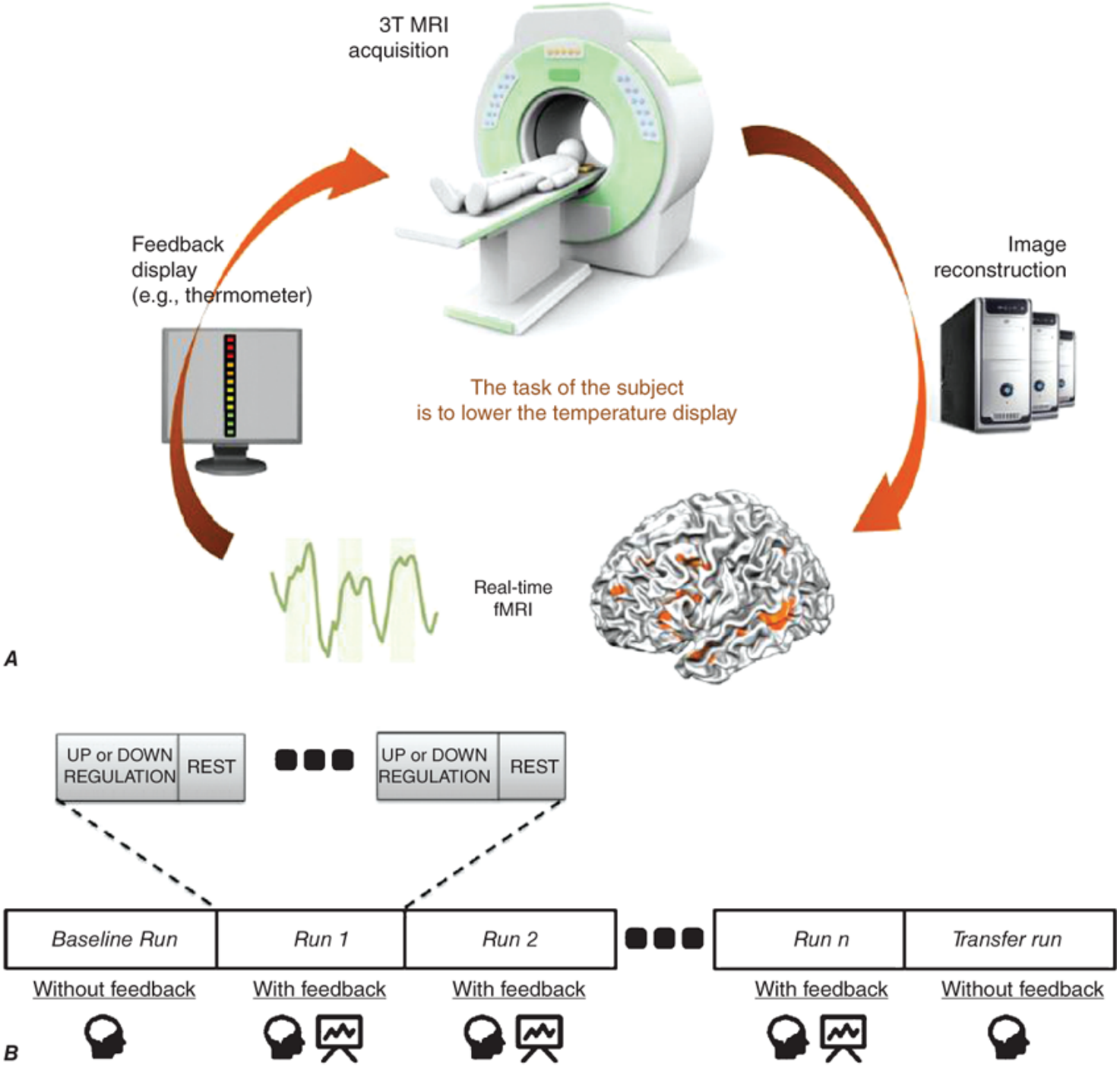
Apart from fMRI-based measures of rsFC, EEG- and MEG-based rsFC measures are also being actively investigated, as these provide a relatively lower cost alternative to fMRI. While EEG is of lowest cost, it compromises on spatial resolution. The major strength of MEG is its ability to provide more accurate source-space estimates of functional oscillatory coupling than EEG as well as provide measures at various physiologically relevant frequencies (up to 50 Hz shown to be clinically useful). In this regard, EEG/MEG are complementary to fMRI, which can only be used to study slow activity fluctuations (i.e., <0.1 Hz); the potential for EEG/MEG modalities to provide valid diagnostic biomarkers is currently underexploited and requires further study.

CLOSED-LOOP NEUROIMAGING

Neuroscientific studies to date are predominantly designed as “open-loop experiments,” interpreting the neurobiological substrates of human behavior via correlation with simultaneously occurring neural activity. In recent years, advances in real-time signal processing have paved the way for “closed-loop neuroimaging,” wherein humans can directly manipulate experiment parameters in real-time based on specific brain signals **(Fig. 477-3)**. Closed-loop imaging methods can not only advance our understanding of dynamic brain function but also have therapeutic potential. Humans can learn to modulate their neural dynamics in specific ways when they are able to perceive (i.e., see/hear) their brain signals in real-time using closed-loop neuroimaging-based neurofeedback. Early studies showed that such neurofeedback learning and resulting neuromodulation could be applied as therapy for patients suffering from chronic pain, motor rehabilitation in Parkinson’s and stroke patients, modulation of aberrant oscillatory activity in epilepsy, and improvement of cognitive abilities such as sustained attention in healthy individuals and patients with attention deficit hyperactivity disorder (ADHD). It has also shown potential for deciphering state of consciousness in comatose patients, wherein a proportion of vegetative/minimally conscious patients could communicate awareness via neuroimaging-based mental imagery.

FIGURE 477-3

Neurofeedback using functional MRI. *(From T Fovet et al: Translating neurocognitive models of auditory-visual hallucinations into therapy. Front Psychiatry 7:103, 2016.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

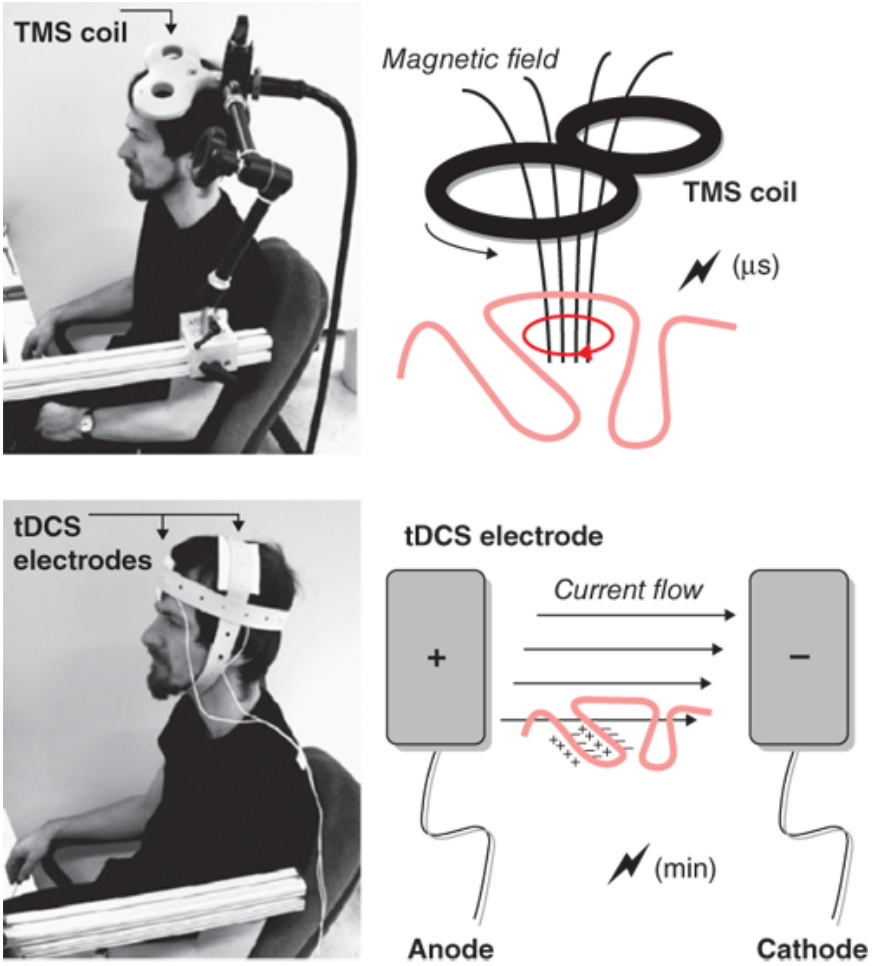
Closed-loop neuroimaging therapeutic studies have utilized real-time fMRI, EEG, and MEG methods. It is common for neural signals to be extracted from specific target brain regions for neuromodulation. However, given that distributed neural networks underlie behavioral deficits, new studies have also explored neurofeedback on combinatorial brain signals from multiple brain regions extracted using multivariate pattern analysis (MVPA). While early studies indicate therapeutic potential, clinical RCTs of closed-loop neuroimaging neurofeedback have shown mixed results. This may largely be because of the individual heterogeneity in neuropsychiatric disorders such that there is no one-size-fits-all therapy. Closed-loop neuroimaging–based therapies need to be better personalized to the preintervention cognitive and neurophysiological states of the individual, and a better understanding needs to be developed regarding learning principles and mechanisms of self-regulation underlying neurofeedback. Clinical practitioners applying these methods also need better education on the hardware/software capabilities of these brain–computer interfaces to maximize patient outcomes.

NONINVASIVE BRAIN STIMULATION (NIBS)

NIBS is widely recognized as having great potential to modulate brain networks in a range of neurological and psychiatric diseases; it is currently approved by the U.S. Food and Drug Administration (FDA) as a treatment for depression. Importantly, there is a very large body of basic research indicating that neuromodulation of the nervous system with electrical stimulation can have both short-term and long-term effects. While TMS uses magnetic fields to generate electrical currents, transcranial direct current stimulation (tDCS), in contrast, is based on direct stimulation using electrical currents applied at the scalp **(Fig. 477-4)**. TMS induces small electrical currents in the brain by magnetic fields which pass through the skull; it is known to be painless and therefore widely used for noninvasive stimulation. Animal research suggests that anodal tDCS causes a generalized reduction in resting membrane potential over large cortical areas, whereas cathodal stimulation causes hyperpolarization. Prolonged stimulation with tDCS can cause an enduring change in cortical excitability under the stimulated regions. Further, changes in resting state fMRI-based activity and functional connectivity have also been observed post-tDCS. Notably, there is uncertainty regarding precisely how much electrical current is able to penetrate through the skull and modulate neural networks.

FIGURE 477-4

Illustration of TMS and tDCS setups. The upper panels show a TMS setup. Coils generate magnetic fields that can in turn generate electrical fields in the cortical tissue. The lower panels show a tDCS setup. The electrical current is believed to flow from the anode (+) to the cathode (−) through the superficial cortical areas leading to polarization. *(From R Sparing, FM Mottaghy: Noninvasive brain stimulation with transcranial magnetic or direct current stimulation [TMS/tDCS]—From insights into human memory to therapy of its dysfunction. Methods 44:329, 2008.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Neuromodulation via stimulation techniques such as tDCS and TMS have shown promise as methods to improve motor function after stroke; there are a growing number of studies demonstrating functional benefits of combining physical therapy with brain stimulation. Two commonly utilized TMS paradigms include low-frequency “inhibitory” stimulation of the healthy cortex or high frequency “excitatory” stimulation of the injured hemisphere. Each of these two approaches aims to modify the balance of reciprocal inhibition between the two hemispheres after stroke. A recent meta-analysis of randomized controlled trials published over the past decade found a significant beneficial effect on motor outcomes. Planned large multicenter trials to assess the long-term benefits of TMS on motor recovery after stroke should provide additional important information on this topic.

TMS and tDCS interventions are also being applied in psychiatric disorders. A substantial body of evidence supports the use of TMS as an antidepressant in MDD (major depressive disorder). TMS is also being investigated for its potential efficacy in posttraumatic stress disorder (PTSD), obsessive compulsive disorder (OCD), and treatment of auditory hallucinations in schizophrenia. Various repetitive TMS (rTMS) protocols have shown efficacy in major depression. These include both low frequency (≤1 Hz) and high frequency rTMS (10–20 Hz) stimulation over dorsolateral prefrontal cortex (DLPFC). Mechanistically, low frequency rTMS is associated with decreased regional cerebral blood flow while high frequency rTMS elicits increased blood flow, not only over the prefrontal region where the TMS is applied, but also in associated basal ganglia and amygdala circuits. Notably, the differential mechanisms of the low versus high frequency rTMS protocols are associated with mood improvements in different sets of MDD patients, and patients showing benefits with one protocol may even show worsening with the other, again pointing to individual heterogeneity in network function. EEG-guided TMS is also being investigated in psychiatric disorders, for instance, individual resting alpha-band (8–12 Hz) peak frequency to determine TMS stimulation rates. With respect to transcranial electrical stimulation in psychiatry, tDCS is the most commonly used protocol. In major depression, there is a documented imbalance in left versus right DLPFC activity, hence, differential anodal versus cathodal tDCS in the left versus right prefrontal cortex may be a potentially efficacious approach. Interestingly, while meta-analysis shows promise for noninvasive brain stimulation methods in psychiatric illness, large RCTs have failed to generate effects compared to placebo treatment. Future success may require careful personalized targeting based on network dynamics and refinement of protocols to accommodate combinatorial treatments.

IMPLANTABLE NEURAL INTERFACES INCLUDING BMIs

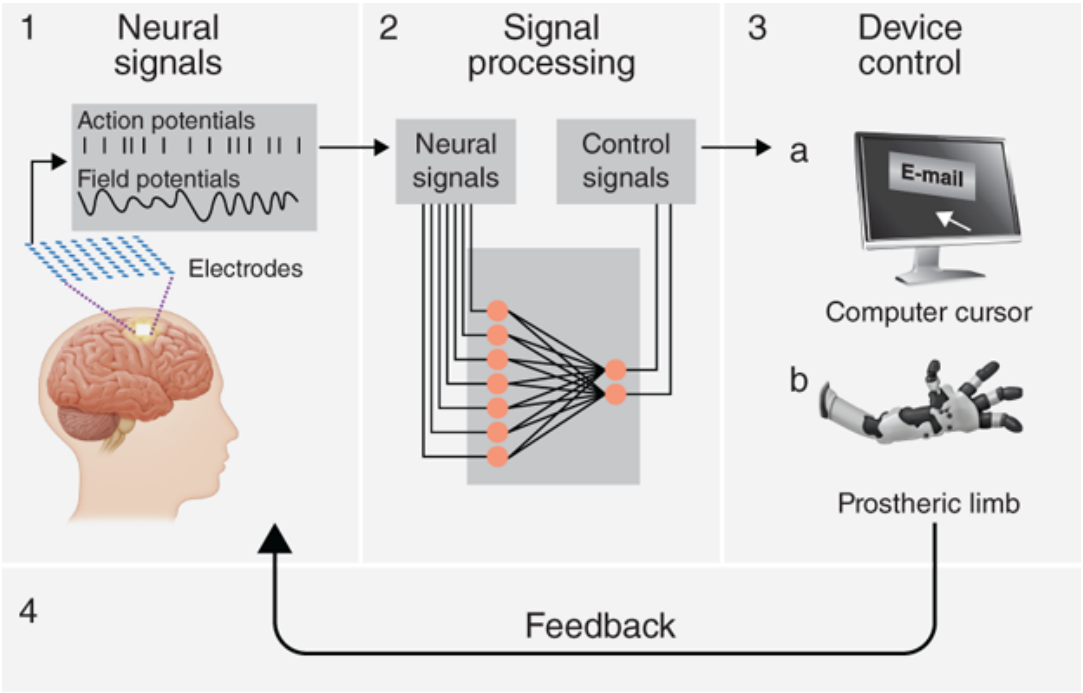
Fully implantable clinically relevant neural interfaces that can improve function already exist. Cochlear implants, for example, are sensory prostheses that can restore hearing in deaf patients. Real-time processing of environmental sounds is converted into patterned stimulation delivered to the cochlear nerve. Importantly, even while the patterned stimulation remains the same, there are gradual improvements in the perception of speech and other complex sounds over a period of several months after device implantation. Activity-dependent sculpting of neural circuits is hypothesized to underlie the observed perceptual improvements. Similarly, the development of deep-brain stimulation (DBS) was based on decades of work showing that surgical lesions to specific nuclei could alleviate tremor and bradykinesia symptoms in animal models. DBS involves chronic implantation of a stimulating electrode that targets specific neural structures (e.g., subthalamic nuclei or the globus pallidus in Parkinson’s disease). At least for movement disorders, it is commonly thought that targeted areas are functionally inhibited by the chronic electrical stimulation.

BMIs represent a more advanced neural interface that aims to restore motor function. Multiple neurological disorders (e.g., traumatic and nontraumatic spinal cord injury, motor-neuron disease, neuromuscular disorders, and strokes) can result in severe and devastating paralysis. Patients cannot perform simple activities and remain fully dependent for care. Especially in patients with high cervical injuries, advanced amyotrophic lateral sclerosis (ALS) or brain-stem strokes, the effects are especially devastating and often leave patients unable to communicate. While there has been extensive research into each disorder, little has proven to be clinically effective for rehabilitation of long-term disability. BMIs offer a promising means to restore function. In the patients described above, while the pathways for transmission of signals to muscles are disrupted, the

brain itself is largely functional. Thus, BMIs can restore function by communicating directly with the brain. For example, in a "motor" BMI, a subject's intention to move is translated in real-time to control a device. As illustrated in **Fig. 477-5**, the components of a motor BMI include: (1) recordings of neural activity, (2) algorithms to transform the neural activity into control signals, (3) an external device driven by these control signals, and (4) feedback regarding the current state of the device.

FIGURE 477-5

Components of a brain–machine interface (BMI). *(From A Tsu et al: Cortical neuroprosthetics from a clinical perspective. Neurobiol Dis 83:154, 2015.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

Many sources of neural signals can be used in a BMI. While EEG signals can be obtained noninvasively, other neural signals require invasive placement of electrodes. Three invasive sources of neural signals include electrocorticography (ECoG), action potentials or spikes, and local field potential (LFP). Spikes and LFP are recorded with electrodes that penetrate the cortex. Spikes represent high-bandwidth signals (300–25,000 Hz) that are recorded from either single neurons ("single unit") or multiple neurons ("multiunit" or MUA). LFPs are the low frequency (~0.1–300 Hz) components. In contract, ECoG is recorded from electrodes that are placed on the cortical surface. ECoG signals may be viewed as an intermediate resolution signal in comparison to spikes/LFP and EEG. It is worth noting that there is still considerable research into the specific neural underpinning of each signal source and what information can be ultimately extracted regarding neural processes.

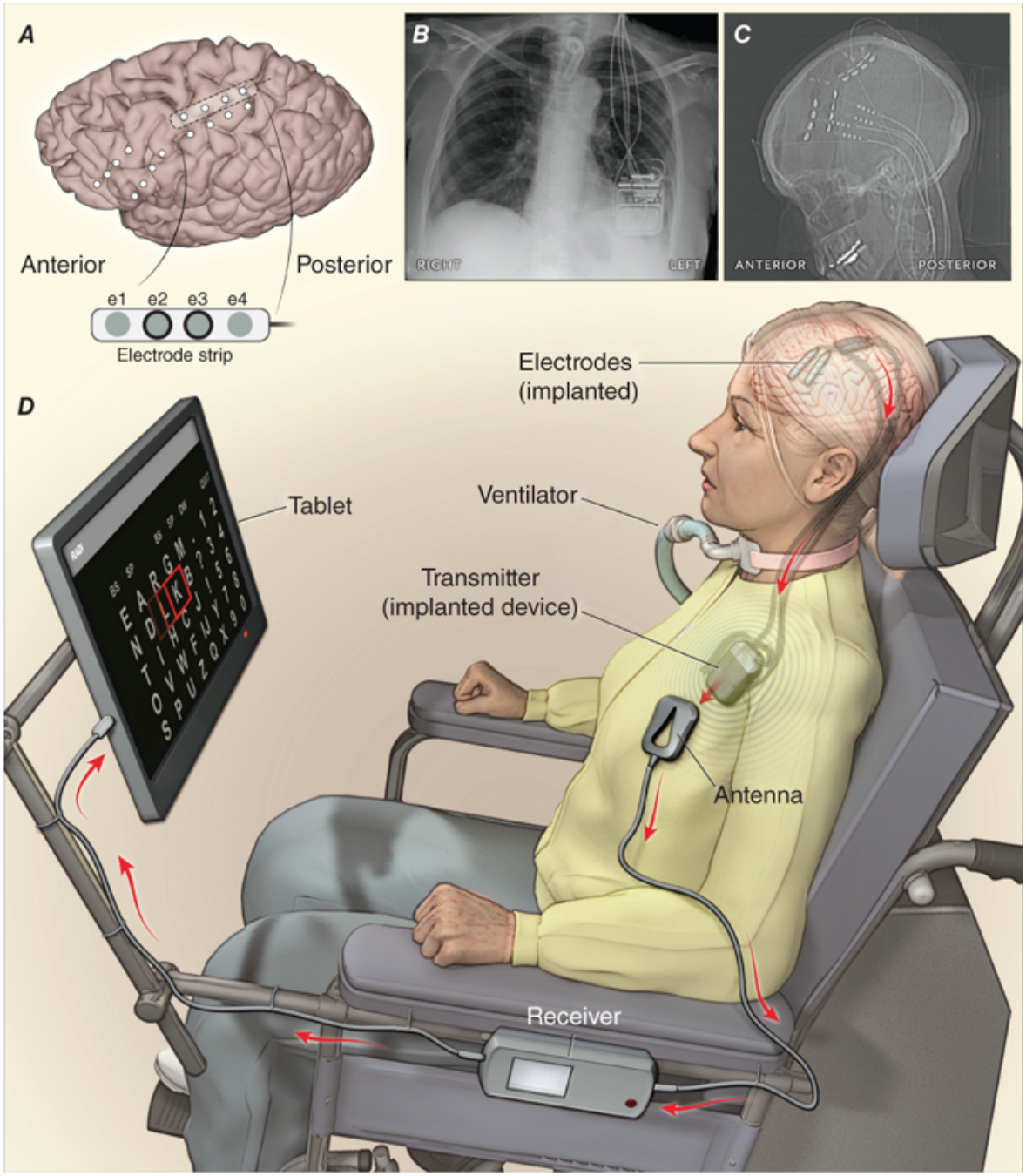
A critical component of a BMI is the transform of neural activity into a reliable control signal. The decoder is an algorithm that converts the neural signals into control signals. One important distinction between classes of decoders is biomimetic versus nonbiomimetic. In the case of biomimetic decoders, the transform attempts to capture the natural relationship between neural activity and a movement parameter. In contrast, nonbiomimetic decoders can be more arbitrary transforms between neural activity and prosthetic control. It had been hypothesized that learning prosthetic control with a biomimetic decoder is more intuitive. Recent evidence, however, reveals that learning may be important for achieving improvements in the level of control over an external device (e.g., a computer cursor, a robotic limb) for either type of decoder. This may be like learning a new motor skill.

A central goal of the field of BMIs is to improve function in patients with permanent disability. This can consist of a range of communication and assistive devices such as a computer cursor, keyboard control, wheelchairs, or a robotic limb. In the ideal scenario, the least invasive method of recording neural signals would allow the most complex level of control. Moreover, control should be allowed in an intuitive manner that resembles the neural control of our natural limbs. There is currently active research into developing and refining techniques to achieve the most complex control possible using each signal source. One measure of complexity is the degrees of freedom that are controlled. For example, control of a computer cursor on the screen (i.e., on the "x" and "y" axis) represents 2 degrees of freedom (DOF). Control of a fully functional prosthetic upper arm that approaches our natural range of motion would require >7 DOF. If the functionality of the hand and fingers are included, then an even more complex level of control would be required. There has been a large body of research on the use of noninvasive recording of EEG signals. Studies suggest that 2 DOF control using EEG is feasible. There are also promising reports of patients with advanced ALS communicating via email using EEG–based BMI. Known limitations of EEG-based BMIs include its "signal-to-noise" ratio (due to filtering of neural signals by bone and skin) and contamination by muscle activity. Ongoing research aims to test usability in a more general nonresearch setting as well as targeted use in patients with disability.

Numerous studies now also indicate that BMIs using invasive recording of neural signals can allow rapid control over devices with multiple DOF. The clear majority of this research has been conducted using recordings of spiking activity via implanted microelectrode arrays. Initial preclinical studies were performed in able-bodied nonhuman primates. More recently, there have been numerous examples of human subjects with a range of neurological illnesses (e.g., brainstem stroke, ALS, spinal cord injury) who have demonstrated the actual use of implantable neural interfaces. This includes demonstrations of both the control of communication interfaces as well as robotic limbs. Pilot clinical trials of BMIs based on invasive recordings of neural signals have further shown that significantly greater rates of communication are possible (e.g., 15–30 characters per minute). Notably, these BMI devices required a percutaneous connection and were always tested in the presence of research staff. A recent case study additionally demonstrated that a fully implantable BMI system could allow communication in a locked-in ALS patient **(Fig. 477-6)**. At the time of the study, the patient required mechanical ventilation and could only communicate using eye movements. She was implanted with multiple subdural cortical electrodes; the neural signals were then processed and sent wirelessly to an external AAC (augmentative and alternative communication) system. Importantly, she could use the interface with no supervision from research staff.

FIGURE 477-6

Illustration of an amyotrophic lateral sclerosis (ALS) patient with a fully implanted communication interface. *(From MJ Vansteensel et al: Fully implanted brain–computer interface in a locked-in patient with ALS. N Engl J Med 375:2060, 2016, Figure 1.)*



Source: J.L. Jameson, A.S. Fauci, D.L. Kasper, S.L. Hauser, D.L. Longo, J. Loscalzo: Harrison's Principles of Internal Medicine, 20th Edition
Copyright © McGraw-Hill Education. All rights reserved.

BMIs have the potential to revolutionize the care of neurologically impaired patients. While in its infancy, there have been multiple proof-of-principle studies that highlight possibilities. Combined basic and clinical efforts will ultimately lead to the development of products that are designed for patients with specific disabilities. As outlined earlier, each signal source has strengths (e.g., noninvasive vs invasive, recording stability) and weaknesses (e.g., bandwidth or the amount of information that can be extracted). With additional research a more precise delineation of these strengths and weakness should occur. For example, one hypothesis is that control of complex devices with high DOF will only be possible using invasive recordings of high-resolution neural activity such as spikes from small clusters of neurons. As these limits become increasingly clear it should allow targeted clinical translational efforts that are geared to specific patient needs and preferences (e.g., extent of disability, medical condition, noninvasive vs invasive). For example, patients with high cervical injuries (i.e., above C4, where the arm and the hand are affected) have different rehabilitation needs than patients with lower cervical injuries (i.e., below C5–C6, where the primary deficits is the hand and fingers).

FURTHER READING

Bassett DS et al: Emerging frontiers of neuroengineering: A network science of brain connectivity. Annu Rev Biomed Eng 19:327, 2017. [PubMed: 28375650]

Drysdale AT et al: Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23:28, 2016. [PubMed: 27918562]

Ganguly K, Poo MM: Activity-dependent neural plasticity from bench to bedside. Neuron 80:729, 2013. [PubMed: 24183023]

Mishra J et al: Video games for neuro-cognitive optimization. Neuron 90:214, 2016. [PubMed: 27100194]

Reinkensmeyer DJ et al: Computational neurorehabilitation: Modeling plasticity and learning to predict recovery. J Neuroeng Rehabil 13:42, 2016. [PubMed: 27130577]

Tsu A et al: Cortical neuroprosthetics from a clinical perspective. Neurobiol Dis 83:154, 2015. [PubMed: 26253606]

Wessel MJ et al: Non-invasive brain stimulation: An interventional tool for enhancing behavioral training after stroke. Front Hum Neurosci 9:265, 2015. [PubMed: 26029083]

McGraw Hill
Copyright © McGraw-Hill Global Education Holdings, LLC.
All rights reserved.
Your IP address is **129.78.139.29**

Access Provided by: University of Sydney Library
Silverchair